Double-Step Alternating Extragradient with Increasing Timescale Separation for Finding Local Minimax Points: Provable Improvements

Kyuwon Kim [1] **Donghwan Kim** [1]

Abstract

In nonconvex-nonconcave minimax optimization, *two-timescale* gradient methods have shown their potential to find local minimax (optimal) points, provided that the timescale separation between the min and the max player is sufficiently large. However, existing two-timescale variants of gradient descent ascent and extragradient methods face two shortcomings, especially when we search for *non-strict* local minimax points that are prevalent in modern overparameterized setting. In specific, (i) these methods can be unstable at some *non-strict* local minimax points even with sufficiently large timescale separation, and even (ii) computing a proper amount of timescale separation is infeasible in practice. To remedy these two issues, we propose to incorporate two simple but provably effective schemes, *double-step alternating* update and *increasing* timescale separation, into the two-timescale extragradient method, respectively. Under mild conditions, we show that the proposed methods converge to *non-strict* local minimax points that all existing two-timescale methods fail to converge.

1. Introduction

The significance of minimax problems in the machine learning community has grown considerably, since generative adversarial network (GAN) (Goodfellow et al., 2014), adversarial training (Madry et al., 2018), multi-agent reinforcement learning (Wai et al., 2018), fair classification (Martinez et al., 2020) and sharpness-aware minimization (Foret et al., 2021), are formulated as

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}).$$

[1]Department of Mathematical Sciences, KAIST, Daejeon, Republic of Korea. Correspondence to: Kyuwon Kim <kkw4053@kaist.ac.kr>, Donghwan Kim <donghwankim@kaist.ac.kr>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

However, solving a nonconvex-nonconcave minimax problem is known to be problematic, even when using a gradient descent ascent (GDA) method that is a natural extension of a gradient descent method in minimization. This stands in stark contrast to the remarkable success of the gradient descent method in machine learning, a success underpinned by theoretical results; under mild assumptions, the gradient descent finds local minimum and escapes strict saddle points with probability one (Lee et al., 2016; 2019). Therefore, the goal of this paper is to establish a comparable theory in nonconvex-nonconcave minimax optimization.

Nonconvex-nonconcave minimax problems in most machine learning applications are *sequential* games and thus have an intrinsic order between the min-player $\boldsymbol{x}$ and max-player $\boldsymbol{y}$ (Fiez et al., 2020; Jin et al., 2020). While such order is negligible in a convex-concave setting, if we disregard this order in a nonconvex-nonconcave setting, for example in GAN, the undesirable mode collapse phenomenon can arise (Goodfellow, 2016). Nevertheless, a proper definition of local optimal point in minimax problems that takes account of the intrinsic order between the players was not widely recognized until (Fiez et al., 2020; Jin et al., 2020). Accordingly, Jin et al. (2020) proposed the first proper notion of local optimal points, named *local minimax points*, for the *sequential* games, which differs from and includes the commonly employed notion of Nash equilibrium in *simultaneous* games.

To find such local minimax (optimal) points, Jin et al. (2020) considered a *two-timescale* GDA (Heusel et al., 2017) that updates with different step sizes (timescales) for each variable $\boldsymbol{x}$ and $\boldsymbol{y}$. In specific, Jin et al. (2020) analyzed that the GDA with a sufficiently large timescale separation can find local minimax points, using dynamical system theory. However, it faces two challenges. The first limitation arises from the *non-degeneracy* assumption on $\nabla^2_{\boldsymbol{yy}} f$ required in (Jin et al., 2020), which neglects the *non-strict* local minimax points that are ubiquitous in the modern overparameterized setting (Cooper, 2021; Liu et al., 2022). The second is that Jin et al. (2020) have not specified how large one should choose an appropriate timescale separation to guarantee the convergence to the local minimax points.

The partial answers to resolve the aforementioned two limitations were provided by Chae et al. (2024b) and Fiez & Ratliff (2021); Li et al. (2022), respectively. First, Chae et al. (2024b) removed the non-degeneracy assumption of $\nabla^2_{\boldsymbol{yy}} f$, and demonstrated that, under mild assumptions, the two-timescale extragradient (EG) method can converge to a set of local minimax points, especially including the *non-strict* local minimax points, that is larger than that of the two-timescale GDA method. Nevertheless, Chae et al. (2024b) state that there still exist some *non-strict* local minimax points that two-timescale EG cannot converge to, due to its insufficient stability. Second, Fiez & Ratliff (2021) and Li et al. (2022) specified an appropriate timescale separation needed to guarantee a local convergence to *strict* local minimax points. However, computing it is infeasible in general as it requires the second-order derivative information of the function f at the local minimax point, and even how one can generalize their result to *non-strict* local minimax points remains open.

This paper thus focuses on addressing these two issues of existing two-timescale gradient methods, which are further detailed in Section 4. This paper then proposes to integrate two simple but provably effective schemes namely *double-step alternating* update and *increasing* timescale separation, in Sections 5 and 6, respectively. Sections 5 and 6 provide improved local convergence analyses in terms of the local minimax points, and these are generalized to a global statement in Section 7. Our contributions via dynamical system theory can be summarized as below.

- **Local Convergence to Local Minimax Points**

(a) **Double-Step Alternating Update:** In Section 5, we present that updating the min and the max player alternatingly, which enhances stability from a dynamical system perspective, proves particularly helpful in addressing the first issue of the (simultaneous-update) two-timescale EG. We would like to highlight that the resulting *double-step alternating* EG (Alt2-EG) method, in Algorithm 1, entails a slight yet essential deviation from the usual alternating scheme, which will be detailed later. Built upon its spectral analysis, we show that, under mild conditions, the Alt2-EG with sufficiently large timescale separation is stable at non-strict local minimax points, for those that are unstable for other existing two-timescale methods.

(b) **Increasing Timescale Separation:** In Section 6, to remedy the second issue, we suggest to simply increase the timescale separation indefinitely as iteration goes. In particular, we demonstrate that the Alt2-EG with *increasing* timescale separation (Alt2-EG-ITS) is stable at non-strict local minimax points that are stable for Alt2-EG with (sufficiently large) fixed timescale separation (Alt2-EG-FTS) in Section 5. We would like to emphasize here that, unlike the technique being simple and straightforward, analyzing its stability via dynamical system theory is rather complicated, since the resulting system is *non-autonomous*.

- **Global Convergence to Local Minimax Points**

  In Section 7, we show that both Alt2-EG-FTS and Alt2-EG-ITS globally find first-order stationary points under a star-convex-star-concave setting. Combined with the aforementioned local convergence analyses, we claim that, under mild conditions, both Alt2-EG-FTS and Alt2-EG-ITS can globally converge to local minimax points under the aforementioned nonconvex-nonconcave setting, while our current analysis for the latter has some limitation due to its *non-autonomous* property.

2. Related Work

Two-Timescale Gradient Methods The vanilla gradient descent ascent (GDA) method may not converge to local minimax points (Daskalakis & Panageas, 2018; Jin et al., 2020). To resolve this non-convergence, the *two-timescale* GDA (Heusel et al., 2017) has been widely studied. For example, under a nonconvex-strongly-concave setting, Lin et al. (2020) established that the GDA with a timescale separation of the order $\Theta(\kappa_{\boldsymbol{y}}^2)$, where $\kappa_{\boldsymbol{y}}$ is a global condition number for $\boldsymbol{y}$, globally finds a first-order stationary point. However, since the set of first-order stationary points is considerably larger than that of the local minimax (optimal) points, it is crucial for a method to only converge to local minimax points.

After introducing the definition of the local minimax point in their paper, Jin et al. (2020) showed that, under the non-degeneracy assumption on $\nabla^2_{\boldsymbol{yy}} f$, the two-timescale GDA locally converges to *strict* local minimax points. Fiez & Ratliff (2021) then specified a proper value of timescale separation needed for such guarantee, which was missing in (Jin et al., 2020). This was further refined in (Li et al., 2022), which demonstrated that GDA with a timescale separation of the order $\Theta(\tilde{\kappa}_{\boldsymbol{y}})$ locally converges to *strict* local minimax points. Here, $\tilde{\kappa}_{\boldsymbol{y}}$ is a local condition number for $\boldsymbol{y}$ at the local minimax point, which is not available in practice.

Recently, Chae et al. (2024b) generalized the local convergence result of (Jin et al., 2020) by removing the crucial non-degeneracy assumption needed in (Jin et al., 2020; Fiez & Ratliff, 2021; Li et al., 2022). This forward step is essential, given that *non-strict* optimal points are everywhere in the modern overparameterized setting (Cooper, 2021; Liu et al., 2022). In specific, Chae et al. (2024b) proved that there exist non-strict local minimax points that the two-timescale EG converges to, while the two-timescale GDA cannot. However, as pointed out in (Chae et al., 2024b) there still exist non-strict local minimax points that two-timescale EG does not converge to, due to its insufficient stability from dynamical system perspective.

Alternating GDA The *alternating* gradient descent ascent (Alt-GDA), which updates each variables $\boldsymbol{x}$ and $\boldsymbol{y}$ sequentially, has been found to be more stable and sometimes faster than the plain *simultaneous* update GDA (Sim-GDA) under various settings (Mescheder et al., 2018; Zhang et al., 2022; Lee et al., 2024). For instance, in a bilinear setting, Sim-GDA moves far away from the optimal point indefinitely, whereas Alt-GDA also does not converge but oscillates in a stable cycle around the optimal point (Mescheder et al., 2018). Moreover, under a strongly-convex-strongly-concave setting, the Sim-GDA locally converges to the optimal point with an iteration complexity of $O(\kappa^2)$, where κ is a condition number, while the Alt-GDA achieves a better near-optimal bound $O(\kappa)$ (Zhang et al., 2022). Very recently and concurrently, it is further demonstrated that the Alt-GDA and its variant (called Alex-GDA) achieve faster rates even globally, compared to that of Sim-GDA (Lee et al., 2024). This paper utilizes these stabilizing effect of the alternating update, from a dynamical system perspective, to improve the stability guarantee of the two-timescale EG.

GDA with Increasing Timescale Separation Under a nonconvex-strongly-concave setting, Li et al. (2023) studied an AdaGrad-like *parameter-free* GDA method that adapts step sizes for each min and max player, which has a practical importance in deep neural network training. Here, the step sizes are determined by past gradient information, *without requiring problem parameters* such as the condition number needed to determine the appropriate amount of timescale separation in (Lin et al., 2020). In regard of (Lin et al., 2020), the resulting adaptive GDA will likely suffer from the non-convergence, since the adaptive step sizes may not have sufficiently large timescale separation. Therefore, Li et al. (2023) suggested to modify the adaptive step size rule so that the timescale separation increases indefinitely as iteration goes. Similar to (Lin et al., 2020), Li et al. (2023) analyzed the global convergence to a first-order stationary point under a nonconvex-strongly-concave setting. Under a similar *parameter-free* but a more general nonconvex-nonconcave setting, this paper establishes a local convergence to the local minimax points, especially for the Alt2-EG with increasing timescale separation, from a *non-autonomous* dynamical system perspective.

3. Preliminaries

3.1. Notations and Problem Setting

We use the notation $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{y})$ to represent a concatenation of the minimization variable $\boldsymbol{x} \in \mathbb{R}^{d_1}$ and the maximization variable $\boldsymbol{y} \in \mathbb{R}^{d_2}$. The saddle-gradient operator of the objective function f will be denoted by $\boldsymbol{F} := (\nabla_{\boldsymbol{x}} f, -\nabla_{\boldsymbol{y}} f)$. For convenience, we denote the second derivatives of f by $\boldsymbol{A} = \nabla^2_{\boldsymbol{xx}} f$, $\boldsymbol{B} = \nabla^2_{\boldsymbol{yy}} f$, and $\boldsymbol{C} = \nabla^2_{\boldsymbol{xy}} f$. Then, the Jacobian of the saddle-gradient $\boldsymbol{F}$ can be expressed as

$$\boldsymbol{H} := D\boldsymbol{F} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ -\boldsymbol{C}^\top & -\boldsymbol{B} \end{bmatrix} \in \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}. \quad (1)$$

Note that the second derivatives $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are matrix valued functions of a $(d_1 + d_2)$-dimensional input point. However, since the input vector will be clear from the context, we simply use $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ to denote the function values. We denote the set of all eigenvalues of a square matrix $\boldsymbol{A}$ by $\operatorname{spec}(\boldsymbol{A})$, the smallest eigenvalue of $\boldsymbol{A}$ by $\lambda_{\min}(\boldsymbol{A})$, and the spectral radius of $\boldsymbol{A}$ by $\rho(\boldsymbol{A})$.

Most of the time, we will impose the following standard smoothness assumption on f.

Assumption 1 (Smoothness of f)**.** *Let $f \in C^1$, and there exist positive constants $L_{\boldsymbol{x}}$ and $L_{\boldsymbol{y}}$ such that, for all $(\boldsymbol{u}, \boldsymbol{v}), (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{d_1+d_2}$,*

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}, \boldsymbol{v}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})\| \le L_{\boldsymbol{x}} \|(\boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{x}, \boldsymbol{y})\|,$$
$$\|\nabla_{\boldsymbol{y}} f(\boldsymbol{u}, \boldsymbol{v}) - \nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})\| \le L_{\boldsymbol{y}} \|(\boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{x}, \boldsymbol{y})\|.$$

This implies that $\|\boldsymbol{F}(\boldsymbol{z}) - \boldsymbol{F}(\boldsymbol{z}')\| \le L\|\boldsymbol{z} - \boldsymbol{z}'\|$ for $L := \sqrt{L_{\boldsymbol{x}}^2 + L_{\boldsymbol{y}}^2}$ and for all $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^{d_1+d_2}$.

For the global analysis in Section 7, we will further impose the following nonconvex-nonconcave condition.

Assumption 2 (Star-convex-star-concave property of f)**.** *Let $f \in C^1$, and f is star-convex-star-concave, i.e., there exists a stationary point $\boldsymbol{z}^* := (\boldsymbol{x}^*, \boldsymbol{y}^*)$ that satisfies*

$$f(\boldsymbol{x}^*, \boldsymbol{y}) \ge f(\boldsymbol{x}, \boldsymbol{y}) + \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{x}^* - \boldsymbol{x} \rangle,$$
$$f(\boldsymbol{x}, \boldsymbol{y}) \ge f(\boldsymbol{x}^*, \boldsymbol{y}) + \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{x}^* \rangle,$$
$$f(\boldsymbol{x}, \boldsymbol{y}^*) \le f(\boldsymbol{x}, \boldsymbol{y}) + \langle \nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y}^* - \boldsymbol{y} \rangle,$$
$$f(\boldsymbol{x}, \boldsymbol{y}) \le f(\boldsymbol{x}, \boldsymbol{y}^*) + \langle \nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}^*), \boldsymbol{y} - \boldsymbol{y}^* \rangle.$$

for all $\boldsymbol{x} \in \mathbb{R}^{d_1}$ and $\boldsymbol{y} \in \mathbb{R}^{d_2}$.

This implies that the Minty variational inequality (MVI) condition (Minty, 1967), *i.e.,* $\langle \boldsymbol{F}(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{z}^* \rangle \ge 0$ for all $\boldsymbol{z} \in \mathbb{R}^{d_1+d_2}$, holds.

3.2. Restricted Schur Complement

Chae et al. (2024b) defined the following matrix that reduces to the standard Schur complement $\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top$ if $\boldsymbol{B}$ is non-degenerate. This is needed in next subsection for expressing the property of the local minimax point.

Definition 1 (Chae et al. (2024b, Definition 4))**.** *For $f \in C^2$, the* **restricted Schur complement** *is defined as $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{H}) := \boldsymbol{U}^\top(\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^\dagger\boldsymbol{C}^\top)\boldsymbol{U}$, with the matrix $\boldsymbol{U}$ defined below.*

Let $r := \operatorname{rank}(\boldsymbol{B})$. As $\boldsymbol{B}$ is symmetric, it is orthogonally diagonalizable into $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{P}^\top$, where $\boldsymbol{\Delta} = \operatorname{diag}\{\delta_1, \dots, \delta_r, 0, \dots, 0\}$ and $\boldsymbol{P}$ is an orthogonal matrix. Define $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ to be submatrices of $\boldsymbol{CP}$, corresponding to the r leftmost columns and $d_2 - r$ rightmost columns of $\boldsymbol{CP}$, respectively. Then, $\boldsymbol{U}$ is defined to be a matrix whose columns form an orthonormal basis for $\mathcal{R}(\boldsymbol{C}_2)^\perp$. The matrix $\boldsymbol{U}$ is not unique in general, but since the spectrum of $\boldsymbol{S}_{\text{res}}(\boldsymbol{H})$ only matters later, $\boldsymbol{U}$ is selected to be any one of possible choices.

3.3. Local Minimax Point

Jin et al. (2020) introduced the following new notion of local optimality for minimax problems.

Definition 2 (Jin et al. (2020, Definition 14))**.** *A point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is said to be a* **local minimax point** *if there exists $\delta_0 > 0$ and a function h satisfying $h(\delta) \to 0$ as $\delta \to 0$ such that, for any $\delta \in (0, \delta_0]$ and any $(\boldsymbol{x}, \boldsymbol{y})$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}^*\| \le \delta$ and $\|\boldsymbol{y} - \boldsymbol{y}^*\| \le \delta$, we have*

$$f(\boldsymbol{x}^*, \boldsymbol{y}) \le f(\boldsymbol{x}^*, \boldsymbol{y}^*) \le \max_{\boldsymbol{y}' \,:\, \|\boldsymbol{y}' - \boldsymbol{y}^*\| \le h(\delta)} f(\boldsymbol{x}, \boldsymbol{y}').$$

Nevertheless, most of the existing literature, such as (Fiez & Ratliff, 2021; Jin et al., 2020; Wang et al., 2020), only focused on finding a *strict* local minimax point that is a stationary point satisfying the second-order *sufficient* condition of local minimax point (Jin et al., 2020):

$$[\nabla^2_{\boldsymbol{xx}} f - \nabla^2_{\boldsymbol{xy}} f (\nabla^2_{\boldsymbol{yy}} f)^{-1} \nabla^2_{\boldsymbol{yx}} f](\boldsymbol{x}^*, \boldsymbol{y}^*) \succ \boldsymbol{0},$$
$$\text{and} \quad \nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) \prec \boldsymbol{0}.$$

Considering that *non-strict* solutions are prevalent in modern overparameterized model training (Cooper, 2021; Liu et al., 2022), Chae et al. (2024a) recently studied constructing a gradient method that can find a *non-strict* (local) minimax point. More precisely, Chae et al. (2024a) (and also this paper) modestly[1] aim to find a stationary point that satisfies the following second-order necessary condition given in (Chae et al., 2024b, Proposition 3.2):

- (Second-order necessary) For $f \in C^2$, any local minimax point $\boldsymbol{z}^*$ satisfies $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) \preceq \boldsymbol{0}$. If the function $h(\delta)$ in Definition 2 satisfies $\limsup_{\delta \to 0+} {h(\delta)}/{\delta} < \infty$, then $\boldsymbol{S}_{\text{res}}(D\boldsymbol{F}(\boldsymbol{x}^*, \boldsymbol{y}^*)) \succeq \boldsymbol{0}$.

Note that this condition does not require a restrictive non-degeneracy condition on $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*)$, unlike that in (Jin et al., 2020), although it adds a mild condition on the function $h(\delta)$; see (Chae et al., 2024b, Remark 3.3) on this matter.

Built upon their necessary condition, Chae et al. (2024b) presented the concept of *strict non-minimax* point that one hopes to escape, which is analogous to the *strict saddle* point in minimization (Lee et al., 2016).

Definition 3 (Chae et al. (2024b, Definition 5))**.** *For $f \in C^2$, a stationary point $\boldsymbol{z}^*$ is said to be a* **strict non-minimax point** *of f if $\lambda_{\min}(\boldsymbol{S}_{\text{res}}(D\boldsymbol{F}(\boldsymbol{z}^*))) < 0$ or $\lambda_{\min}(-\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}^*)) < 0$. We denote the set of strict non-minimax points by $\mathcal{T}^*$.*

When necessary, we will impose the following Assumption 3 (or its stronger version below) at stationary points $\boldsymbol{z}^*$ as (Chae et al., 2024b). This is weaker than the non-degeneracy condition on $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}^*)$, which was crucial in (Jin et al., 2020; Fiez & Ratliff, 2021; Li et al., 2022).

Assumption 3. *Let $f \in C^2$, and for a stationary point $\boldsymbol{z}^*$ in consideration, at least one of the matrices $\boldsymbol{S}_{\text{res}}(D\boldsymbol{F}(\boldsymbol{z}^*))$ and $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}^*)$ is non-degenerate.*

Assumption 3′. *Assumption 3 holds, and $D\boldsymbol{F}(\boldsymbol{z}^*)$ is non-degenerate.*

To aid understanding, a simple example with a non-strict local minimax point is provided in Section 8, accompanied with numerical experiment.

3.4. Stability of Dynamical Systems

From a dynamical system perspective, this paper analyzes the *asymptotic (or exponential) stability* and the instability of the method in a form

$$\boldsymbol{z}_{k+1} = \boldsymbol{w}_k(\boldsymbol{z}_k), \tag{2}$$

at an equilibrium $\boldsymbol{z}^*$, where k denotes the iteration number.

Definition 4. *The equilibrium point $\boldsymbol{z}^*$ of (2) is*

- **(Lyapunov) stable** *if, for each $\epsilon > 0$, there is $\delta = \delta(\epsilon, k_0) > 0$ such that $\|\boldsymbol{z}_{k_0} - \boldsymbol{z}^*\| < \delta$ implies $\|\boldsymbol{z}_k - \boldsymbol{z}^*\| < \epsilon, \ \forall k \ge k_0 \ge 0$,*
- **asymptotically stable** *if it is stable and there is a positive constant $c = c(k_0)$ such that $\boldsymbol{z}_k \to \boldsymbol{z}^*$ as $k \to \infty$, for all $\|\boldsymbol{z}_{k_0} - \boldsymbol{z}^*\| < c$,*
- **exponentially stable** *if it is asymptotically stable and there are constants $\beta, c > 0$ such that $\|\boldsymbol{z}_k - \boldsymbol{z}^*\| \le e^{-\beta k} \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|$, for all $\|\boldsymbol{z}_0 - \boldsymbol{z}^*\| < c, \ \forall k \ge 1$,*
- **unstable***, if it is not stable.*

If a dynamical system is *autonomous*, *i.e.*, $\boldsymbol{w}_k$ is invariant with respect to the iteration number k, we can characterize

[1] Already in minimization, finding a local minimizer is NP-Hard in the worst-case (Murty & Kabadi, 1987). So instead, a gradient descent method is similarly shown to find a stationary point that satisfies the second-order necessary condition of a local minimizer (Lee et al., 2016; 2019).

its exponential stability and instability at an equilibrium point $\boldsymbol{z}^*$ by just examining the spectrum of the Jacobian of $\boldsymbol{w}_k$ at $\boldsymbol{z}^*$ as below.

Proposition 3.1 (Galor (2007), Polyak (1987)). *Let $\boldsymbol{w} \in C^1$, and $\boldsymbol{z}^*$ be an equilibrium of $\boldsymbol{z}_{k+1} = \boldsymbol{w}(\boldsymbol{z}_k)$. Then,*

- *$\boldsymbol{z}^*$ is exponentially stable iff $\rho(D\boldsymbol{w}(\boldsymbol{z}^*)) < 1$*
- *$\boldsymbol{z}^*$ is unstable if $\rho(D\boldsymbol{w}(\boldsymbol{z}^*)) > 1$.*

This paper also analyzes a *non-autonomous* system, *i.e.,* a system $\boldsymbol{w}_k$ that is not invariant with respect to k, in Section 6. Proposition 3.1 cannot be applied here, so we directly analyze the asymptotic stability and the convergence rate of the considered method.

4. Stability and Limitations of the Two-Timescale EG

In this section, we review the stability of the two-timescale EG and its two limitations in (Chae et al., 2024b).

4.1. Stability of the Two-Timescale EG

Chae et al. (2024b) studied the two-timescale EG:

$$\boldsymbol{z}_{k+1} = \boldsymbol{w}_\tau(\boldsymbol{z}_k) := \boldsymbol{z}_k - \eta\boldsymbol{\Lambda}_\tau \boldsymbol{F}(\boldsymbol{z}_k - \eta\boldsymbol{\Lambda}_\tau \boldsymbol{F}(\boldsymbol{z}_k)), \quad (3)$$

where $\eta > 0$ is a step size, $\tau \geq 1$ is a fixed timescale separation parameter and $\boldsymbol{\Lambda}_\tau := \mathrm{diag}\{(1/\tau)\boldsymbol{I}, \boldsymbol{I}\}$. Based on Proposition 3.1, the stability of the two-timescale EG (3) at an equilibrium $\boldsymbol{z}^*$ depends on whether or not the spectral radius of its Jacobian matrix $D\boldsymbol{w}_\tau$ at $\boldsymbol{z}^*$ is smaller than 1. However, since directly computing $\rho(D\boldsymbol{w}_\tau)$ is complicated, Chae et al. (2024b) demonstrated that it is equivalent to examine the relationship between the spectrum of

$$\boldsymbol{H}_\tau := \boldsymbol{\Lambda}_\tau D\boldsymbol{F} = \begin{bmatrix} \frac{1}{\tau}\boldsymbol{A} & \frac{1}{\tau}\boldsymbol{C} \\ -\boldsymbol{C}^\top & -\boldsymbol{B} \end{bmatrix}. \quad (4)$$

and the set $\mathcal{P}_\eta := \{(x,y) \in \mathbb{C} \mid \left(\eta x - \frac{1}{2}\right)^2 + \eta^2 y^2 + \frac{3}{4} < \sqrt{1 + 3\eta^2 y^2}\}$ (see Figure 1) as below.

Proposition 4.1 (Chae et al. (2024b, Proposition 5.5)). *Let $f \in C^2$. An equilibrium point $\boldsymbol{z}^*$ is exponentially stable, i.e., $\rho(D\boldsymbol{w}_\tau(\boldsymbol{z}^*)) < 1$, if and only if $\mathrm{spec}(\boldsymbol{H}_\tau(\boldsymbol{z}^*)) \subset \mathcal{P}_\eta$.*

This does not necessarily imply that the two-timescale EG is stable at the desired local minimax point, so Chae et al. (2024b) next relate the stability condition and the second-order necessary condition when τ is sufficiently large.

4.2. Eigenvalue Characteristics of $\boldsymbol{H}_\tau$ and its Relation to Second-Order Necessary Condition

Chae et al. (2024b) analyzed the behavior of the eigenvalues of $\boldsymbol{H}_\tau$ in terms of τ as below, under Assumption 3. This reduces to (Jin et al., 2020, Lemma 40) when we strictly impose the non-degeneracy condition on $\nabla^2_{\boldsymbol{yy}} f$.

Theorem 4.2 (Chae et al. (2024b, Theorem 4.3)). *Under Assumption 3, for $\tau \geq 1$ and $\epsilon := 1/\tau$, it is possible to construct continuous functions $\lambda_j(\epsilon)$, $j = 1, \ldots, d_1 + d_2$ so that they are the $d_1 + d_2$ complex eigenvalues λ_j of $\boldsymbol{H}_\tau$ in (4) with the following asymptotics as $\epsilon \to 0+$:*

$$\begin{array}{lll} (i) & |\lambda_j - i\sigma_j\sqrt{\epsilon}| = o(\sqrt{\epsilon}), & j = 1, \ldots, q \\ & |\lambda_{j+d_1} + i\sigma_j\sqrt{\epsilon}| = o(\sqrt{\epsilon}), & \\ (ii) & |\lambda_{j+q} - \epsilon\mu_j| = o(\epsilon), & j = 1, \ldots, d_1 - q, \\ (iii) & |\lambda_{j+d_1+q} - \nu_j| = o(1), & j = 1, \ldots, r, \end{array}$$

where $q := \mathrm{rank}(\boldsymbol{C}_2)$, which are nonzero for all $\epsilon > 0$, while the $(d_2 - r - q)$ remaining λ_j's are 0. Here, μ_j are the eigenvalues of $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{H})$[2], ν_j are the nonzero eigenvalues of $-\boldsymbol{B}$, and σ_j are the singular values of $\boldsymbol{C}_2$.

The key distinction between the spectral analysis of $\boldsymbol{H}_\tau$ in (Jin et al., 2020, Lemma 40) and Theorem 4.2 is the existence of additional type (i) eigenvalues (and the $d_2 - r - q$ number of zero eigenvalues), which are irrelevant to the second-order necessary condition. On the other hand, the type (ii) and type (iii) eigenvalues are associated with $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{H})$ and $-\boldsymbol{B}$, respectively, so they are directly connected to the second-order necessary condition.

4.3. Stability of the Two-Timescale EG at Local Minimax Points

When the point $\boldsymbol{z}^*$ satisfies the second-order necessary condition of local minimax points, by Theorem 4.2, the type (ii) and type (iii) eigenvalues lie in the set $\mathcal{P}_\eta$, for a sufficiently large τ (or equivalently, for a sufficiently small ϵ), as desired. On the other hand, consider the *strict non-minimax* point $\boldsymbol{z}^*$ that, by definition, does not satisfy the second-order necessary condition. Then, at least one of the type (ii) and type (iii) eigenvalues lie outside the closure of $\mathcal{P}_\eta$, even with a large τ, as one wishes.

Let us now focus particularly on the eigenvalues of type (i) in Theorem 4.2. As they are irrelevant to the second-order necessary condition, ideally we want these to be contained in the set $\mathcal{P}_\eta$, for a sufficiently large τ, and do not cause instability of the two-timescale EG.

What we know from Theorem 4.2 is that the type (i) eigenvalue $\lambda_j(\epsilon)$ converges to 0 as $\epsilon \to 0^+$, and asymptotically approaches the imaginary axis. Since the target set $\mathcal{P}_\eta$, illustrated in Figure 1, has the imaginary axis as its tangential line at the origin, Theorem 4.2 alone is not sufficient to determine the stability of the two-timescale EG.

[2]Although $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{H})$ is not unique due to its non-unique choice of matrix $\boldsymbol{U}$, its eigenvalues remain the same regardless of the choice of $\boldsymbol{U}$ due to matrix similarity.

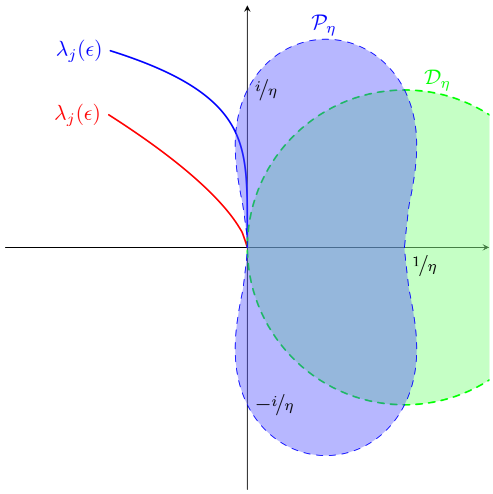

Figure 1. A target set $\mathcal{P}_\eta$ of the two-timescale EG, and two representative scenarios of type (i) eigenvalues, approaching 0 from the left-half plane, in Theorem 4.2. (As a comparison, we added a similarly derived target set $\mathcal{D}_\eta := \{(x, y) \in \mathbb{C} \mid (\eta x - 1)^2 + \eta^2 y^2 < 1\}$ of the two-timescale GDA $\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta \boldsymbol{\Lambda}_\tau \boldsymbol{F}(\boldsymbol{z}_k)$.)

Figure 1 illustrates two representative examples of the type (i) eigenvalue asymptotics, where the blue-colored one converges to 0 from inside the set $\mathcal{P}_\eta$, while the other (colored red) converges to 0 from outside the set.[3] This observation necessitated Chae et al. (2024b) to further investigate the curvature of both the type (i) eigenvalue $\lambda_j(\epsilon)$ and the target set $\mathcal{P}_\eta$ around the origin, for completing the stability analysis of the two-timescale EG. This, however, only revealed the fact that the two-timescale EG cannot avoid the red-colored case, resulting in instability for such corresponding non-strict local minimax points.

4.4. Two Limitations of the Two-Timescale EG

We summarize the two limitations of the two-timescale EG, which we address in Sections 5 and 6, respectively.

1) There are type (i) eigenvalues $\lambda_j(\epsilon)$ that are not related to the second-order necessary condition but not contained in the set $\mathcal{P}_\eta$, even for a large τ, which consequently leads to instability of the two-timescale EG at certain non-strict local minimax points.
2) Type (ii) and (iii) eigenvalues become associated with the second-order necessary condition for sufficiently large τ, but its specific value is not available in practice.

[3]The GDA is unstable even for the blue-colored eigenvalue asymptotic, as the target set $\mathcal{D}_\eta$ of GDA does not cover a region nearby the imaginary axis as much as the set $\mathcal{P}_\eta$ of EG. This is illustrated in Figure 1, which shows the superiority of EG over GDA in this context (Chae et al., 2024b).

5. Addressing the First Limitation: Double-Step Alternating Update

In this section, we investigate two stabilizing approaches, namely explicit regularization and implicit regularization (by alternating update), that shift the type (i) eigenvalues towards the set $\mathcal{P}_\eta$ without affecting the eigenvalue asymptotics of the type (ii) and (iii). In specific, we look for ways to shift the type (i) eigenvalues to the right in a complex plane by the order of $\Omega(\sqrt{\epsilon})$, needed by Theorem 4.2.

5.1. Explicit Regularization for Two-Timescale EG

Since the type (i) eigenvalues are critically related to $\nabla^2_{\boldsymbol{yy}} f$, we consider explicitly adding a regularization term $-\frac{c}{\gamma}\|\boldsymbol{y}\|^2$ to an objective function $f(\boldsymbol{x}, \boldsymbol{y})$, where c and γ are positive constants, while the latter depends on τ. The saddle gradient operator of the resulting regularized function is

$$\boldsymbol{F}_{\text{reg},\gamma}(\boldsymbol{x}, \boldsymbol{y}) := \Big(\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) + \frac{2c}{\gamma}\boldsymbol{y}\Big).$$

Although this explicit regularization has a drawback that the stationary points of the regularized function differ from those of the original function f, we proceed as it is straightforward and provides an insight on how we should choose γ in terms of τ.

We first consider the case "$\gamma = \tau$", where the Jacobian of the timescaled $\boldsymbol{\Lambda}_\tau \boldsymbol{F}_{\text{reg},\tau}$ at its stationary point is

$$\boldsymbol{H}_{\text{reg1},\tau} := \boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\text{reg},\tau} = \begin{bmatrix} \epsilon \boldsymbol{A} & \epsilon \boldsymbol{C} \\ -\boldsymbol{C}^\top & -\boldsymbol{B} + 2c\epsilon \boldsymbol{I} \end{bmatrix} \quad (5)$$

where $\epsilon = 1/\tau$. The eigenvalues of $\boldsymbol{H}_{\text{reg1},\tau}$ have the following asymptotic behaviors.

Proposition 5.1. *Under the same setting as Theorem 4.2, the eigenvalues λ_j of $\boldsymbol{H}_{\text{reg1},\tau}$ behave the same as those of $\boldsymbol{H}_\tau$ in Theorem 4.2, which are nonzero for all but finitely many $\epsilon > 0$, except for the eigenvalues λ_{j+d_1+r+q} being $2c\epsilon$ for $j = 1, \ldots, d_2 - r - q$ (instead of 0 in Theorem 4.2).*

The result is what we anticipated as we perturbed only in the order of $o(\sqrt{\epsilon})$. So, we next consider the choice "$\gamma = \sqrt{\tau}$", where the eigenvalues of the corresponding Jacobian

$$\boldsymbol{H}_{\text{reg2},\tau} := \boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\text{reg},\sqrt{\tau}} = \begin{bmatrix} \epsilon \boldsymbol{A} & \epsilon \boldsymbol{C} \\ -\boldsymbol{C}^\top & -\boldsymbol{B} + 2c\sqrt{\epsilon} \boldsymbol{I} \end{bmatrix}$$

have the following asymptotic behaviors.

Theorem 5.2. *Under the same setting as Theorem 4.2, the eigenvalues λ_j of $\boldsymbol{H}_{\text{reg2},\tau}$ behave the same as those of $\boldsymbol{H}_\tau$ in Theorem 4.2, except for the type (i) eigenvalues being*

$$\text{(i)} \quad \Big|\lambda_j - \Big(c + \sqrt{c^2 - \sigma_j^2}\Big)\sqrt{\epsilon}\Big| = o(\sqrt{\epsilon}).$$

$$\left|\lambda_{j+d_1} - \left(c - \sqrt{c^2 - \sigma_j^2}\right)\sqrt{\epsilon}\right| = o(\sqrt{\epsilon}),$$

for $j = 1, \ldots, q$, which are nonzero for all but finitely many $\epsilon > 0$, and the eigenvalues λ_{j+d_1+r+q} being $2c\sqrt{\epsilon}$ for $j = 1, \ldots, d_2 - r - q$.

Here, regardless of the radicand of $\sqrt{c^2 - \sigma_j^2}$ being positive or negative, it is straightforward that the type (i) eigenvalues converge to 0 as $\epsilon \to 0^+$ from the right-half plane (but not necessarily approaches the positive real axis asymptotically). Therefore, they now lie in the set $\mathcal{P}_\eta$, unlike the previous results in Theorem 4.2 and Proposition 5.1. This implies that when an appropriate explicit regularization of the order $\Theta(\sqrt{\epsilon})$ is provided, we will not encounter the eigenvalue examples in Figure 1, approaching 0 from the left-half plane. Nevertheless, as mentioned before, stationary points of $\boldsymbol{F}_{\mathrm{reg2},\sqrt{\tau}}$ differ from those of $\boldsymbol{F}$, which is not usually desirable, so we next investigate an approach that provides a similar regularization effect without changing stationary points.

5.2. Double-Step Alternating Update in Two-Timescale EG

Consider an operator for an alternating update:

$$\boldsymbol{F}_{\mathrm{alt},\gamma}(\boldsymbol{x},\boldsymbol{y}) := \left(\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{y}), -\nabla_{\boldsymbol{y}} f\left(\boldsymbol{x} - \frac{\eta}{\gamma}\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{y}), \boldsymbol{y}\right)\right),$$

where $\nabla_{\boldsymbol{y}} f$ is computed after the update of $\boldsymbol{x}$. Here, $\eta > 0$ is a step size and γ is a positive constant that depends on τ. Note that $\boldsymbol{F}_{\mathrm{alt},\gamma}$ reduces to $\boldsymbol{F}$ as $\gamma \to \infty$. In addition, it is obvious that $\boldsymbol{F}$ and $\boldsymbol{F}_{\mathrm{alt},\gamma}$ share same stationary points, unlike $\boldsymbol{F}_{\mathrm{reg},\gamma}$.

Before we proceed to analyze the eigenvalue asymptotics of $\boldsymbol{F}_{\mathrm{alt},\gamma}$, we illustrate the corresponding alternating method:

$$\begin{aligned} \boldsymbol{z}_{k+1} &= \boldsymbol{w}_{\mathrm{alt},\tau_k,\gamma_k}(\boldsymbol{z}_k) \\ &:= \boldsymbol{z}_k - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\mathrm{alt},\gamma_k}(\boldsymbol{z}_k - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\mathrm{alt},\gamma_k}(\boldsymbol{z}_k)), \end{aligned} \quad (6)$$

which is constructed by replacing $\boldsymbol{F}$ by $\boldsymbol{F}_{\mathrm{alt},\gamma}$ in (3). Here, we use τ_k and γ_k, instead of τ and γ, so that they can change over iteration, which will be useful in Section 6. The proposed update (6) takes a stationary point of $\boldsymbol{F}$ as its equilibrium, see Appendix B.4. We further write down this method in terms of f, in Algorithm 1. Note that this method reduces to the (simultaneous-update) two-timescale EG in (Chae et al., 2024b) if we choose $(\tau_k, \gamma_k) = (\tau, \infty)$ for all k.

We would like to emphasize here that Algorithm 1 deviates from the usual alternating update scheme, as we allow τ_k and γ_k to take different values. In alignment with the spectral analysis in Section 5.1, we will soon show that allowing $\tau_k \neq \gamma_k$ is essential for our purpose. To highlight this deviation, we name this method in Algorithm 1 as a *double-step* alternating extragradient method with timescale separation (Alt2-EG-TS). In the rest of this section, we consider fixed constants $(\tau_k, \gamma_k) = (\tau, \gamma)$ for all k, and we name the corresponding method as a double-step alternating extragradient method with *fixed* timescale separation (Alt2-EG-FTS).

Algorithm 1 Double-step alternating extragradient method with timescale separation (Alt2-EG-TS)

Input: $\boldsymbol{x}_0 \in \mathbb{R}^{d_1}, \boldsymbol{y}_0 \in \mathbb{R}^{d_2}, \tau_k \in (1,\infty), \gamma_k \in (1,\infty)$
for $k = 0, 1, \ldots$ **do**

$$\boldsymbol{u}_k = \boldsymbol{x}_k - \frac{\eta}{\tau_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k,\boldsymbol{y}_k)$$

$$\boldsymbol{v}_k = \boldsymbol{y}_k + \eta\nabla_{\boldsymbol{y}} f\left(\boldsymbol{x}_k - \frac{\eta}{\gamma_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k,\boldsymbol{y}_k), \boldsymbol{y}_k\right)$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\eta}{\tau_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k,\boldsymbol{v}_k)$$

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \eta\nabla_{\boldsymbol{y}} f\left(\boldsymbol{u}_k - \frac{\eta}{\gamma_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k,\boldsymbol{v}_k), \boldsymbol{v}_k\right)$$

end for

Remark 5.1. A concurrent work by Lee et al. (2024) also studied and improved the alternating GDA, especially by adding extrapolation steps, which provided an accelerated rate of convergence in a strongly-convex-strongly-concave problem. Interestingly, our double-step alternating scheme can be also viewed as taking an extrapolation step, like (Lee et al., 2024), since we can rewrite our step as

$$\boldsymbol{x}_k - \frac{\eta}{\gamma_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k,\boldsymbol{y}_k) = \left(1 - \frac{\tau_k}{\gamma_k}\right)\boldsymbol{x}_k + \frac{\tau_k}{\gamma_k}\boldsymbol{u}_k,$$

where the choice of $\frac{\tau_k}{\gamma_k}$ being larger than 1, in the next subsection, yields extrapolation.

5.3. Implicit Regularization via Double-Step Alternating Update

This section investigates an implicit regularization provided by the double-step alternating update. Similar to Proposition 5.1, the choice "$\gamma = \tau$" in $\boldsymbol{F}_{\mathrm{alt},\gamma}$ does not lead to any change in the eigenvalue asymptotics; see Appendix B.5.

We now focus on the choice "$\gamma = \sqrt{\tau}$", hoping for an appropriate shift of the type (i) eigenvalues. The Jacobian of the timescaled $\boldsymbol{H}_{\mathrm{alt2},\tau} := \boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\mathrm{alt},\sqrt{\tau}}$, at its stationary point is, where $\epsilon = 1/\tau$,

$$\boldsymbol{H}_{\mathrm{alt2},\tau} = \begin{bmatrix} \epsilon\boldsymbol{A} & \epsilon\boldsymbol{C} \\ -\boldsymbol{C}^\top + \eta\sqrt{\epsilon}\boldsymbol{C}^\top\boldsymbol{A} & -\boldsymbol{B} + \eta\sqrt{\epsilon}\boldsymbol{C}^\top\boldsymbol{C} \end{bmatrix},$$

whose eigenvalues have the following asymptotic behaviors.

Theorem 5.3. *Under the same setting as Theorem 4.2, the eigenvalues λ_j of $\boldsymbol{H}_{\mathrm{alt2},\tau}$ behave the same as those of $\boldsymbol{H}_\tau$ in Theorem 4.2, except for the type (i) eigenvalues being* *(i)* $\left|\lambda_j - \left(\frac{\eta\sigma_j^2}{2} + \frac{\sqrt{\eta^2\sigma_j^4 - 4\sigma_j^2}}{2}\right)\sqrt{\epsilon}\right| = o(\sqrt{\epsilon})$,

$\left|\lambda_{j+d_1} - \left(\frac{\eta\sigma_j^2}{2} - \frac{\sqrt{\eta^2\sigma_j^4 - 4\sigma_j^2}}{2}\right)\sqrt{\epsilon}\right| = o(\sqrt{\epsilon})$,

for $j = 1, \ldots, q$.

Similar to Theorem 5.2 for the explicit regularization with $\gamma = \sqrt{\tau}$, the type (i) eigenvalues here approach 0 from the right-half plane and thus from inside the set $\mathcal{P}_\eta$. This implies that the double-step alternating update with $\gamma = \sqrt{\tau}$ implicitly induces a regularizing behavior similar to the explicit regularization with $\gamma = \sqrt{\tau}$. We would like to emphasize again that a clear advantage of the double-step alternating, over the explicit regularization, is that the resulting method takes a stationary point of $\boldsymbol{F}$ as its equilibrium.

5.4. Stability Analysis of Alt2-EG-FTS

Built upon the previous spectral analysis, we show that, under Assumption $3'$, a stationary point that satisfies the second-order necessary condition is *exponentially stable* for the Alt2-EG-FTS with $(\tau_k, \gamma_k) = (\tau, \sqrt{\tau})$, for a sufficiently large τ, under mild conditions.

Theorem 5.4. *Suppose Assumptions 1 and $3'$ hold and $f \in C^2$. Then, an equilibrium point $\boldsymbol{z}^*$ satisfies $\boldsymbol{S}_{\text{res}} \succeq \boldsymbol{0}$ and $\boldsymbol{B} \preceq \boldsymbol{0}$ if and only if there exists some τ^* such that, for any $\tau > \tau^*$, $\boldsymbol{z}^*$ is an exponentially stable point of Alt2-EG-FTS with $(\tau_k, \gamma_k) = (\tau, \sqrt{\tau})$, for any step size $0 < \eta < 1/L$.*

In addition, we show that, under Assumption 3, the Alt2-EG-FTS with $(\tau_k, \gamma_k) = (\tau, \sqrt{\tau})$ almost surely escapes the strict non-minimax points, which we hope to avoid.

Theorem 5.5. *Let $\boldsymbol{z}^*$ be a strict non-minimax point, i.e., $\boldsymbol{z}^* \in \mathcal{T}^*$. Under Assumptions 1, 3, and $0 < \eta < (\sqrt{5}-1)/2\sqrt{2}L$, there exists $\tau^\star > 0$ such that for any $\tau > \tau^\star$, the set of initial points that converge to $\boldsymbol{z}^*$ by Alt2-EG-FTS with $(\tau_k, \gamma_k) = (\tau, \sqrt{\tau})$ has measure zero. Moreover, if $\mathcal{T}^*$ is finite, then there exists $\tau^\star > 0$ such that for any $\tau > \tau^\star$, $\mu(\{\boldsymbol{z}_0 : \lim_{k\to\infty} \boldsymbol{w}^k_{\text{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}_0) \in \mathcal{T}^*\}) = 0$.*

We ultimately want that there are no points, other than strict non-minimax points, that the method escapes almost surely. Fortunately, the stability results above imply that such can be obtained if we choose step size $0 < \eta < (\sqrt{5}-1)/2\sqrt{2}L$ for the Alt2-EG-FTS with $(\tau_k, \gamma_k) = (\tau, \sqrt{\tau})$ and a sufficiently large τ. This statement will be further generalized to a global statement in Section 7.

6. Addressing the Second Limitation: Increasing Timescale Separation

In this section, we consider increasing the timescale separation, and specifically, we analyze the stability of the double-step alternating extragradient with *increasing* timescale separation (Alt2-EG-ITS) from a *non-autonomous* dynamical system perspective.

6.1. Increasing Timescale Separation for Alt2-EG

Choosing a properly large value of τ is practically infeasible, so we consider increasing the coefficients (τ_k, γ_k) of Alt2-EG-ITS indefinitely as k increases. We first let $\gamma_k = \sqrt{\tau_k}$, based on the arguments in Section 5, and the only thing left to determine is the rate at which we should increase τ_k.

Our global convergence analysis in Section 7 requires the sequence τ_k to not increase faster than $\sqrt{k}$ to warrant global convergence; see Theorem 7.1. Therefore, we focused on the choice $\tau_k = k^{1/(2+2c)} \approx \sqrt{k}$ for any positive constant c, and we leave further investigation on τ_k as future work.

6.2. Stability Analysis of Alt2-EG-ITS

Since Proposition 3.1 cannot be applied to the *non-autonomous* Alt2-EG-ITS, we derive an analogous *asymptotic* stability result that is specifically designed for the Alt2-EG-ITS ($\boldsymbol{w}_{\text{alt},\tau_k,\gamma_k}$), under Assumption $3'$.

Theorem 6.1. *Suppose Assumptions 1 and $3'$ hold and $f \in C^3$. Let $\boldsymbol{z}^*$ be an equilibrium point satisfies $\boldsymbol{S}_{\text{res}} \succeq \boldsymbol{0}$ and $\boldsymbol{B} \preceq \boldsymbol{0}$. Then, $\boldsymbol{z}^*$ is an asymptotically stable point of Alt2-EG-ITS with $\left(k^{1/(2+2c)}, k^{1/(4+4c)}\right)$ for any $c > 0$ and $0 < \eta < 1/L$, with a rate $O\left(\frac{1}{\sqrt{k}} e^{-2\sqrt{k}}\right)$.*

Note that, due to decreasing step sizes (as we iteratively increase both τ_k and γ_k), we have a rate of convergence slower than the *exponential* rate of Alt2-EG-FTS in Theorem 5.4, while not requiring one to choose a properly large value of τ as in Theorem 5.4. Moreover, we needed a slightly stronger condition $\boldsymbol{w}_{\text{alt},\tau_k,\gamma_k} \in C^2$ (and thus $f \in C^3$), compared to $\boldsymbol{w} \in C^1$ in Proposition 3.1. We leave relaxing such condition as a future work.

Our next step would be to investigate the instability of Alt2-EG-ITS at strict non-minimax points that we would like to avoid, as for the Alt2-EG-FTS in the previous section. We, however, leave this as a future work. This is because Theorem 5.5 for the Alt2-EG-FTS uses the stable manifold theorem (Shub, 1987, Theorem III.7) for an *autonomous* system, and to the best of our knowledge, how one can generalize it to a *non-autonomous* system is not known yet.

7. Global Convergence of Alt2-EG-TS

So far, our convergence and avoidance results remained local. To generalize these statements globally, we first show that both Alt2-EG-FTS and Alt2-EG-ITS globally find first-order stationary points, under Assumption 2.

Theorem 7.1. *Under Assumptions 1 and 2, consider the*

Alt2-EG-TS methods with $\tau_k \geq 1$ and $\gamma_k \geq 1$. Then,

- *Alt2-EG-FTS with (τ, γ) and $0 < \eta < \frac{\sqrt{\gamma}}{\sqrt{2+\gamma}L}$ satisfies* $\lim_{k\to\infty} \|\boldsymbol{F}_{\text{alt},\gamma}(\boldsymbol{x}_k, \boldsymbol{y}_k)\| = 0$,
- *Alt2-EG-ITS with (τ_k, γ_k) and $0 < \eta < \frac{1}{L}$, for any sequence τ_k satisfying $\sum_{k=0}^{\infty} \frac{1}{\tau_k^2} = \infty$, satisfies* $\liminf_{k\to\infty} \|\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{x}_k, \boldsymbol{y}_k)\| = 0$.

Although $\lim_{k\to\infty} \|\boldsymbol{F}_{\text{alt},\gamma}(\boldsymbol{x}_k, \boldsymbol{y}_k)\| = 0$ only implies that any accumulation point of the iterates $\{\boldsymbol{x}_k, \boldsymbol{y}_k\}$ is a stationary point (see Appendix D.2), it is possible that the iterates of Alt2-EG-FTS converge to a stationary point. For such convergent case, based on Theorems 5.5 and 7.1 and under their settings (such as Assumptions 1, 2 and 3), we have the following global statement.

- The iterates of the Alt2-EG-FTS with sufficiently large τ globally and almost surely converges to a stationary point that satisfies the second-order necessary condition of local minimax point.

  (Note that the invertibility of $D\boldsymbol{F}$, in Assumption 3′, is not needed here, which was needed in analyzing the exponential stability in Theorem 5.4. This implies that Alt2-EG-FTS can even converge to *non-strict* local minimax points even with *degenerate* $D\boldsymbol{F}$.)

The two-timescale EG (Chae et al., 2024b) does not achieve this statement for some *non-strict* local minimax points, as discussed in Section 4.2. Therefore, our result is a lot closer to being analogous to the well-known result of (Lee et al., 2016; 2019), in minimization, that the iterates of the gradient descent method (if they converge) globally and almost surely converges to a stationary point that satisfies the second-order necessary condition of local minimizer.

On the other hand, since we do not have a result of avoiding a strict non-minimax point for the Alt2-EG-ITS, unlike Theorem 5.5 for Alt2-EG-FTS, we have a relatively weaker global statement, based on Theorems 6.1 and 7.1 and under their settings (such as Assumptions 1, 2, 3′ and $f \in C^3$).

- Alt2-EG-ITS globally finds a stationary point, and it also locally converges to a stationary point that satisfies the second-order necessary condition of the local minimax point with non-degenerate $D\boldsymbol{F}$.

8. Example and Experiment

We consider a simple example of a non-strict local minimax point to verify our theory.

Example 1. *Consider the function $f(x, y) = -x^2 + 2xy$. This has a unique stationary point $(0, 0)$, which is a* non-strict *local minimax point. Since this optimal point further satisfies Assumption 3′, by Theorems 5.4 and 6.1, it is asymptotically stable for both Alt2-EG-FTS with $(\tau, \sqrt{\tau})$ and sufficiently large τ, and Alt2-EG-ITS with $(k^{1/(2+2c)}, k^{1/(4+4c)})$ for any $c > 0$, whereas it is unstable for the (vanilla) two-timescale EG; see Section E.1 for the proof.*

We ran our proposed Alt2-EG-FTS with $(\tau, \sqrt{\tau})$ and Alt2-EG-ITS with $(k^{1/(2+2c)}, k^{1/(4+4c)})$, and compared them with existing two-timescale gradient methods (GDA-FTS and EG-FTS). We also performed Alt2-EG-TS with (τ, τ), named Alt1-EG-FTS here, which uses a standard alternating scheme that differs from our double-step alternating scheme. We consider two choices of $\tau = 2$ and 20, where $\tau = 20$ corresponds to a sufficiently large τ, whereas $\tau = 2$ leads to an insufficient timescale separation. As expected, our Alt2-EG-FTS with $\tau = 20$ and Alt2-EG-ITS converge to the local minimax point, while the others do not.

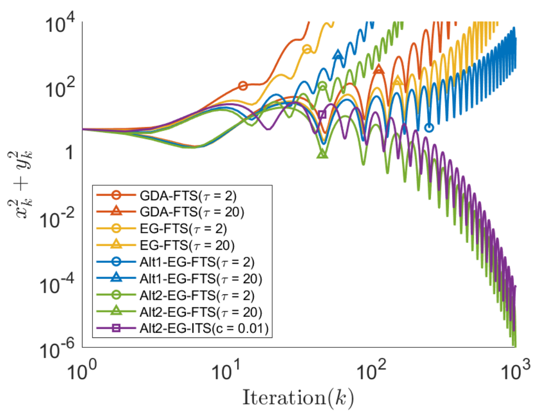

Figure 2. Numerical results with $f(x, y) = -x^2 + 2xy$.

9. Conclusion

We proposed to incorporate double-step alternating update and increasing timescale separation schemes into two-timescale extragradient method—the first method developed for finding *non-strict* local minimax points—to address its two limitations (summarized in Section 4.4). We have then demonstrated that the proposed Alt2-EG-TS method is capable of finding non-strict local minimax points, which cannot be found by existing methods. Therefore, our work, built upon the initial step of (Chae et al., 2024b), is a step closer to establishing a convergence theory in minimax optimization, comparable to the well established theory of gradient descent in minimization problems (Lee et al., 2016; 2019).

Yet, our analysis requires mild but somewhat restrictive conditions, such as Assumption 3, and the escaping behavior around strict non-minimax points for the increasing timescale separation scheme remains unknown. We leave investigating these issues as an interesting future work.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2019R1A5A1028324, 2022R1C1C1003940), and the Samsung Science & Technology Foundation grant (No. SSTF-BA2101-02).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Apostol, T. M. *Calculus, Volume 1*. John Wiley & Sons, 1991.

Chae, J., Kim, K., and Kim, D. Two-timescale extragradient for finding local minimax points. In *Proc. Intl. Conf. on Learning Representations*, 2024a.

Chae, J., Kim, K., and Kim, D. Two-timescale extragradient for finding local minimax points. In *Proc. Intl. Conf. on Learning Representations*, 2024b.

Cooper, Y. Global minima of overparameterized neural networks. *SIAM Journal on Mathematics of Data Science*, 3(2):676–691, 2021.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Fiez, T. and Ratliff, L. Local convergence analysis of gradient descent ascent with finite timescale separation. In *Proc. Intl. Conf. on Learning Representations*, 2021.

Fiez, T., Chasnov, B., and Ratliff, L. Implicit learning dynamics in Stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *International Conference on Machine Learning*, pp. 3133–3144. PMLR, 2020.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *Proc. Intl. Conf. on Learning Representations*, 2021.

Galor, O. *Discrete dynamical systems*. Springer Science & Business Media, 2007.

Goodfellow, I. NIPS 2016 Tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. In *Neural Info. Proc. Sys.*, 2014.

Heusel, M., Ramsauer, H., Nessler, T. U. B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Neural Info. Proc. Sys.*, 2017.

Jin, C., Netrapalli, P., and Jordan, M. I. What is local optimality in nonconvex-nonconcave minimax optimization? In *Proc. Intl. Conf. Mach. Learn*, 2020.

Lee, J., Cho, H., and Yun, C. Fundamental benefit of alternating updates in minimax optimization. In *International Conference on Machine Learning*, 2024.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257. PMLR, 2016.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176:311–337, 2019.

Li, H., Farnia, F., Das, S., and Jadbabaie, A. On convergence of gradient descent ascent: A tight local analysis. In *International Conference on Machine Learning*, pp. 12717–12740. PMLR, 2022.

Li, X., Yang, J., and He, N. Tiada: A time-scale adaptive algorithm for nonconvex minimax optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Liu, C., Zhu, L., and Belkin, M. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Madry, A., Makelov, A., Schmdit, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *Proc. Intl. Conf. on Learning Representations*, 2018.

Martinez, N., Bertran, M., and Sapiro, G. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pp. 6755–6764. PMLR, 2020.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually converge. In *Proc. Intl. Conf. Mach. Learn*, 2018.

Minty, G. J. On the generalization of a direct method of the calculus of variations. *Bull. Amer. Math. Soc*, 73: 314–321, 1967.

Murty, K. G. and Kabadi, S. N. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

Nesterov, Y. *Lectures on convex optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, second edition, 2018.

Polyak, B. T. *Introduction to optimization*. New York, Optimization Software,, 1987.

Shub, M. *Global Stability of Dynamical Systems*. Springer, 1987.

Tao, T. *Analysis II*. Springer, third edition, 2016.

Wai, H.-T., Yang, Z., Wang, Z., and Hong, M. Multi-agent reinforcement learning via double averaging primal-dual optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Wang, Y., Zhang, G., and Ba, J. On solving minimax optimization locally: A follow-the-ridge approach. In *Proc. Intl. Conf. on Learning Representations*, 2020.

Zedek, M. Continuity and location of zeros of linear combinations of polynomials. *Proceedings of the American Mathematical Society*, 16(1):78–84, 1965.

Zhang, G., Wang, Y., Lessard, L., and Grosse, R. B. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 7659–7679. PMLR, 2022.

A. Proofs for Section 3

A.1. Proofs for Proposition 3.1

In proving Proposition 3.1, we need the following stable manifold theorem.

Lemma A.1 (Shub (1987, Theorem III.7))**.** *Let $\boldsymbol{z}^*$ be a fixed point for the C^r local diffeomorphism $\phi : U \to E$ where U is a neighborhood of $\boldsymbol{z}^*$ in the Banach space E. Suppose that $E = E_{\mathrm{cs}} \oplus E_{\mathrm{u}}$, where E_{cs} is the invariant subspace corresponding to the eigenvalues of $D\phi(\boldsymbol{z}^*)$ whose magnitude is less than or equal to 1, and E_{u} is the invariant subspace corresponding to eigenvalues of $D\phi(\boldsymbol{z}^*)$ whose magnitude is greater than 1. Then there exists a C^r embedded disc $W^{\mathrm{u}}_{\mathrm{loc}}$ tangent to the E_{u} at $\boldsymbol{z}^*$ called the local unstable center manifold. Additionally, there exists a neighborhood B of $\boldsymbol{z}^*$ such that $W^{\mathrm{u}}_{\mathrm{loc}} = \{\boldsymbol{z} \in B \mid \phi^k(\boldsymbol{z}) \in B$ for all $k \le 0$ and $d(\phi^k(\boldsymbol{z}), \boldsymbol{z}^*)$ tends to zero exponentially$\}$ where $\phi^{-1} : W^{\mathrm{u}}_{\mathrm{loc}} \to W^{\mathrm{u}}_{\mathrm{loc}}$ is a contraction mapping.*

Proof of Proposition 3.1. The proof of the first statement can be found in both Galor (2007, Theorem 4.8) and Polyak (1987, Theorem 2.1.2.1), and we are only left to prove the second statement.

Suppose that $\rho(D\boldsymbol{w}(\boldsymbol{z}^*)) > 1$. Then, by Lemma A.1, there exists a disk $W^{\mathrm{u}}_{\mathrm{loc}}$. Clearly, $\boldsymbol{z}^*$ is contained in $W^{\mathrm{u}}_{\mathrm{loc}}$, since $\boldsymbol{w}^k(\boldsymbol{z}^*) = \boldsymbol{z}^*$ for any $k \le 0$. Take $\epsilon = \frac{1}{2}\max_{\boldsymbol{z} \in W^{\mathrm{u}}_{\mathrm{loc}}} d(\boldsymbol{z}, \boldsymbol{z}^*)$. Then, there exists $\boldsymbol{z} \in W^{\mathrm{u}}_{\mathrm{loc}}$ such that $\boldsymbol{z} \notin D_\epsilon(\boldsymbol{z}^*)$, where $D_\epsilon(\boldsymbol{z}^*)$ is a disc centered at $\boldsymbol{z}^*$ with radius ϵ.

For the sake of contradiction, suppose that $\boldsymbol{z}^*$ of dynamics $\boldsymbol{w}(\cdot)$ is (Lyapunov) stable. Then, for ϵ discussed above and any given $k_0 \in \mathbb{N}$, there exists $\delta > 0$ such that $\|\boldsymbol{z}_{k_0} - \boldsymbol{z}^*\| < \delta$ implies $\|\boldsymbol{z}_k - \boldsymbol{z}^*\| < \epsilon$ for all $k \ge k_0$. Then, by the definition of $W^{\mathrm{u}}_{\mathrm{loc}}$, there exists $n_0 \in \mathbb{N}$ such that $\boldsymbol{w}^{-n}(\boldsymbol{z}) \in D_\delta(\boldsymbol{z}^*)$ for all $n \ge n_0$. Then, for $m := \max\{k_0, n_0\} + 1$, we have $\boldsymbol{z} = \boldsymbol{w}^m(\boldsymbol{w}^{-m}(\boldsymbol{z})) \in \boldsymbol{w}^m(D_\delta(\boldsymbol{z}^*)) \subset D_\epsilon(\boldsymbol{z}^*)$, which is absurd. Therefore, we conclude that $\rho(D\boldsymbol{w}(\boldsymbol{z}^*)) > 1$ implies that $\boldsymbol{z}^*$ is not stable.

□

B. Proofs for Section 5

In proving theorems in Section 5, we need the following lemmas.

Lemma B.1. *Under Assumption 1, the operator $\boldsymbol{F}_{\mathrm{alt},\gamma}$ with $\gamma > 0$ satisfies, for all $(\boldsymbol{u}, \boldsymbol{v}), (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{d_1+d_2}$,*

$$\|\boldsymbol{F}_{\mathrm{alt},\gamma}(\boldsymbol{u}, \boldsymbol{v}) - \boldsymbol{F}_{\mathrm{alt},\gamma}(\boldsymbol{x}, \boldsymbol{y})\| \le \sqrt{L_{\boldsymbol{x}}^2 + L_{\boldsymbol{y}}^2\left(1 + \frac{1}{\gamma}\right)\left(1 + L_{\boldsymbol{x}}^2\frac{\eta^2}{\gamma}\right)}\|(\boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{x}, \boldsymbol{y})\|.$$

Proof. For simplicity, let us denote $\bar{\boldsymbol{u}} := \boldsymbol{u} - \frac{\eta}{\gamma}\nabla_{\boldsymbol{x}} f(\boldsymbol{u}, \boldsymbol{v})$ and $\bar{\boldsymbol{x}} := \boldsymbol{x} - \frac{\eta}{\gamma}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})$. Then, we have

$$\begin{aligned}
\|\boldsymbol{F}_{\mathrm{alt},\gamma}(\boldsymbol{u}, \boldsymbol{v}) - \boldsymbol{F}_{\mathrm{alt},\gamma}(\boldsymbol{x}, \boldsymbol{y})\|^2 &= \|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}, \boldsymbol{v}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})\|^2 + \|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{u}}, \boldsymbol{v}) - \nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}, \boldsymbol{y})\|^2 \\
&\le L_{\boldsymbol{x}}^2\|(\boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{x}, \boldsymbol{y})\|^2 + L_{\boldsymbol{y}}^2\|(\bar{\boldsymbol{u}}, \boldsymbol{v}) - (\bar{\boldsymbol{x}}, \boldsymbol{y})\|^2 \\
&\le L_{\boldsymbol{x}}^2\|(\boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{x}, \boldsymbol{y})\|^2 + L_{\boldsymbol{y}}^2\|\bar{\boldsymbol{u}} - \bar{\boldsymbol{x}}\|^2 + L_{\boldsymbol{y}}^2\|\boldsymbol{v} - \boldsymbol{y}\|^2
\end{aligned}$$

Since $\bar{\boldsymbol{u}} - \bar{\boldsymbol{x}} = \boldsymbol{u} - \frac{\eta}{\gamma}\nabla_{\boldsymbol{x}} f(\boldsymbol{u}, \boldsymbol{v}) - \boldsymbol{x} + \frac{\eta}{\gamma}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})$ holds, and using the Young's inequality

$$\|\boldsymbol{a} + \boldsymbol{b}\|^2 \le \left(1 + \frac{1}{\gamma}\right)\|\boldsymbol{a}\|^2 + (1 + \gamma)\|\boldsymbol{b}\|^2,$$

we have

$$\begin{aligned}
\|\bar{\boldsymbol{u}} - \bar{\boldsymbol{x}}\|^2 &\le \left(1 + \frac{1}{\gamma}\right)\|\boldsymbol{u} - \boldsymbol{x}\|^2 + (1 + \gamma)\frac{\eta^2}{\gamma^2}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}, \boldsymbol{v}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})\|^2 \\
&\le \left(1 + \frac{1}{\gamma}\right)\|\boldsymbol{u} - \boldsymbol{x}\|^2 + L_{\boldsymbol{x}}^2\frac{\eta^2}{\gamma}\left(1 + \frac{1}{\gamma}\right)\|(\boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{x}, \boldsymbol{y})\|^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\|\boldsymbol{F}_{\mathrm{alt},\gamma}(\boldsymbol{u},\boldsymbol{v}) - \boldsymbol{F}_{\mathrm{alt},\gamma}(\boldsymbol{x},\boldsymbol{y})\|^2 \\
&\le L_{\boldsymbol{x}}^2\|(\boldsymbol{u},\boldsymbol{v}) - (\boldsymbol{x},\boldsymbol{y})\|^2 + L_{\boldsymbol{y}}^2\left(1+\frac{1}{\gamma}\right)\|\boldsymbol{u}-\boldsymbol{x}\|^2 + L_{\boldsymbol{x}}^2 L_{\boldsymbol{y}}^2\frac{\eta^2}{\gamma}\left(1+\frac{1}{\gamma}\right)\|(\boldsymbol{u},\boldsymbol{v}) - (\boldsymbol{x},\boldsymbol{y})\|^2 + L_{\boldsymbol{y}}^2\|\boldsymbol{v}-\boldsymbol{y}\|^2 \\
&= \left(L_{\boldsymbol{x}}^2 + L_{\boldsymbol{y}}^2 + L_{\boldsymbol{x}}^2 L_{\boldsymbol{y}}^2\frac{\eta^2}{\gamma}\left(1+\frac{1}{\gamma}\right)\right)\|(\boldsymbol{u},\boldsymbol{v}) - (\boldsymbol{x},\boldsymbol{y})\|^2 + \frac{L_{\boldsymbol{y}}^2}{\gamma}\|\boldsymbol{u}-\boldsymbol{x}\|^2 \\
&\le \left(L_{\boldsymbol{x}}^2 + L_{\boldsymbol{y}}^2\left(1+\frac{1}{\gamma}\right)\left(1+L_{\boldsymbol{x}}^2\frac{\eta^2}{\gamma}\right)\right)\|(\boldsymbol{u},\boldsymbol{v}) - (\boldsymbol{x},\boldsymbol{y})\|^2,
\end{aligned}$$

and this completes the proof. □

Lemma B.2 (Zedek (1965, Theorem 1))**.** *Given a polynomial $p_n(z) \coloneqq \sum_{k=0}^n a_k z^k$, $a_n \neq 0$, an integer $m \ge n$ and a number $\epsilon > 0$, there exists a number $\delta > 0$ such that whenever the $m+1$ complex numbers b_k, $0 \le k \le m$, satisfy the inequalities*

$$|b_k - a_k| < \delta \quad \textit{for} \quad 0 \le k \le n, \quad \textit{and} \quad |b_k| < \delta \quad \textit{for} \quad n+1 \le k \le m,$$

then the roots β_k, $1 \le k \le m$, of the polynomial $q_m(z) \coloneqq \sum_{k=0}^m b_k z^k$ can be labeled in such a way as to satisfy, with respect to the zeros α_k, $1 \le k \le n$, of $p_n(z)$, the inequalities

$$|\beta_k - \alpha_k| < \epsilon \quad \textit{for} \quad 1 \le k \le n, \quad \textit{and} \quad |\beta_k| > \frac{1}{\epsilon} \quad \textit{for} \quad n+1 \le k \le m.$$

Lemma B.3 (Chae et al. (2024b, Lemma E.1 and Corollary E.2))**.** *A (possibly complex) number μ is an eigenvalue of the restricted Schur complement $\boldsymbol{S}_{\mathrm{res}}$ if and only if it is a root of the equation*

$$\det\begin{bmatrix} \mu\boldsymbol{I} - \boldsymbol{A} & -\boldsymbol{C}_1 & -\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & -\boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{L}_q^\top & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} = 0, \tag{7}$$

where the definitions of $\boldsymbol{D}$ and $\boldsymbol{L}_q$ are given in the proof of Proposition 5.1 below. Moreover, if $\boldsymbol{S}_{\mathrm{res}}$ is invertible. the equation (7) does not have $\mu = 0$ as a solution.

Lemma B.4 (Jin et al. (2020, Lemma 40))**.** *Let $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ respectively be $d_1 \times d_1$ symmetric, $d_2 \times d_2$ non-degenerate symmetric, and $d_1 \times d_2$ matrices. Then, the $d_1 + d_2$ complex eigenvalues of $\boldsymbol{H}_\tau$ have the following asymptotics as $\epsilon = \frac{1}{\tau} \to 0+$:*

$$|\lambda_j - \epsilon\mu_j| = o(\epsilon), \quad j = 1, \dots, d_1, \qquad |\lambda_{j+d_1} - \nu_j| = o(1), \quad j = 1, \dots, d_2,$$

where $\{\mu_j\}_{j=1,\dots,d_1}$ and $\{\nu_j\}_{j=1,\dots,d_2}$ are the eigenvalues of $\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top$ and $-\boldsymbol{B}$, respectively.

B.1. Proof of Proposition 5.1

Proof of Proposition 5.1. We first consider the case where $\boldsymbol{S}_{\mathrm{res}}$ is non-degenerate in Assumption 3, and we begin with the following observation. Consider a block matrix $\boldsymbol{Q} = \mathrm{diag}(\boldsymbol{I}, \boldsymbol{P}^\top)$ where $\boldsymbol{P}$ is orthogonal matrix such that $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{P}^\top$ for some diagonal matrix $\boldsymbol{\Delta} = \mathrm{diag}\{\delta_1, \dots, \delta_r, 0, \dots, 0\}$. Then, one can show that $D\boldsymbol{F}_{\mathrm{reg},\tau}$ is similar to

$$\boldsymbol{G}_{\mathrm{reg1},\tau} \coloneqq \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{C}_2 \\ -\boldsymbol{C}_1^\top & \boldsymbol{D} + 2c\epsilon\boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{C}_2^\top & \boldsymbol{0} & 2c\epsilon\boldsymbol{I} \end{bmatrix}$$

where $\boldsymbol{D} = \mathrm{diag}(-\delta_1, \dots, -\delta_r)$ and $\epsilon = \frac{1}{\tau}$, since $\boldsymbol{G}_{\mathrm{reg1},\tau} = \boldsymbol{Q}D\boldsymbol{F}_{\mathrm{reg},\tau}\boldsymbol{Q}^\top$. Here, the matrix $\boldsymbol{C}_2$ may not be of full column rank matrix, and we further refine the similarity statement below.

Let $q \coloneqq \mathrm{rank}(\boldsymbol{C}_2)$, and let $\boldsymbol{C}_2 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the (full) singular value decomposition of $\boldsymbol{C}_2$, where for some invertible diagonal matrix $\boldsymbol{\Sigma}_q$, it holds that

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_q & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}.$$

Then, by defining $\boldsymbol{L}_q \coloneqq \boldsymbol{U}\begin{bmatrix}\boldsymbol{\Sigma}_q \\ \mathbf{0}\end{bmatrix}$, we have $\boldsymbol{U}\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{L}_q & \mathbf{0}\end{bmatrix}$. Thus, for $\tilde{\boldsymbol{Q}} = \mathrm{diag}\{\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{V}^\top\}$ we have

$$\tilde{\boldsymbol{Q}}\boldsymbol{G}_{\mathrm{reg1},\tau}\tilde{\boldsymbol{Q}}^\top = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{C}_2\boldsymbol{V} \\ -\boldsymbol{C}_1^\top & \boldsymbol{D}+2c\epsilon\boldsymbol{I} & \mathbf{0} \\ -\boldsymbol{V}^\top\boldsymbol{C}_2^\top & \mathbf{0} & 2c\epsilon\boldsymbol{I}\end{bmatrix} = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q & \mathbf{0} \\ -\boldsymbol{C}_1^\top & \boldsymbol{D}+2c\epsilon\boldsymbol{I} & \mathbf{0} & \mathbf{0} \\ -\boldsymbol{L}_q^\top & \mathbf{0} & 2c\epsilon\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 2c\epsilon\boldsymbol{I}\end{bmatrix}$$

and therefore

$$\det(\lambda\boldsymbol{I} - D\boldsymbol{F}_{\mathrm{reg},\tau}) = \det(\lambda\boldsymbol{I} - \tilde{\boldsymbol{Q}}\boldsymbol{G}_{\mathrm{reg},\tau}\tilde{\boldsymbol{Q}}^\top) = (\lambda - 2c\epsilon)^{d_2-r-q}\det\left(\lambda\boldsymbol{I} - \begin{bmatrix}\boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & \boldsymbol{D}+2c\epsilon\boldsymbol{I} & \mathbf{0} \\ -\boldsymbol{L}_q^\top & \mathbf{0} & 2c\epsilon\boldsymbol{I}\end{bmatrix}\right).$$

Hence, we notice that the eigenvalues of $D\boldsymbol{F}_{\mathrm{reg},\tau}$ are either $2c\epsilon$ or the eigenvalues of

$$\boldsymbol{\Phi}_{\mathrm{reg},\tau} \coloneqq \begin{bmatrix}\boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & \boldsymbol{D}+2c\epsilon\boldsymbol{I} & \mathbf{0} \\ -\boldsymbol{L}_q^\top & \mathbf{0} & 2c\epsilon\boldsymbol{I}\end{bmatrix},$$

and analogously the eigenvalues of $\boldsymbol{H}_{\mathrm{reg1},\tau} = \boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\mathrm{reg},\tau}$ (5) are either $2c\epsilon$ or the eigenvalues of $\boldsymbol{\Phi}_{\mathrm{reg1},\tau} \coloneqq \boldsymbol{\Lambda}_\tau\boldsymbol{\Phi}_{\mathrm{reg},\tau}$. Therefore, characterizing the eigenvalues of $\boldsymbol{\Phi}_{\mathrm{reg1},\tau}$ is equivalent to characterizing the nonzero eigenvalues of $\boldsymbol{H}_{\mathrm{reg1},\tau}$.

Since the eigenvalues of $\boldsymbol{\Phi}_{\mathrm{reg1},\tau}$ are the solutions of the equation

$$0 = p_\epsilon(\lambda) \coloneqq \det\begin{bmatrix}\lambda\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C}_1 & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & \lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I} & \mathbf{0} \\ \boldsymbol{L}_q^\top & \mathbf{0} & (\lambda - 2c\epsilon)\boldsymbol{I}\end{bmatrix} = \det(\lambda\boldsymbol{I} - \boldsymbol{\Phi}_{\mathrm{reg1},\tau}), \tag{8}$$

we need to investigate the solutions of the equation. By Lemma B.2, constructing the functions $\lambda_j(\epsilon)$ so that they are continuous is possible, and the eigenvalues converge to the solutions of the equation

$$p_0(\lambda) = \det\begin{bmatrix}\lambda\boldsymbol{I} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{C}_1^\top & \lambda\boldsymbol{I} - \boldsymbol{D} & \mathbf{0} \\ \boldsymbol{L}_q^\top & \mathbf{0} & \lambda\boldsymbol{I}\end{bmatrix} = 0$$

as $\epsilon \to 0$. Hence, the r eigenvalues of $\boldsymbol{\Phi}_{\mathrm{reg1},\tau}$ converge to the r nonzero eigenvalues of $-\boldsymbol{B}$, and the other $d_1 + q$ eigenvalues converge to zero, as $\epsilon \to 0$.

To investigate the order of eigenvalues that converges to zero further, we begin by observing that, whenever $|\lambda|$ and ϵ are small enough so that $\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I}$ is invertible, it holds that

$$\begin{aligned}\det(\lambda\boldsymbol{I} - \boldsymbol{\Phi}_{\mathrm{reg1},\tau}) &= \det\begin{bmatrix}\lambda\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C}_1 & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & \lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I} & \mathbf{0} \\ \boldsymbol{L}_q^\top & \mathbf{0} & \lambda\boldsymbol{I} - 2c\epsilon\boldsymbol{I}\end{bmatrix} \\ &= \det\begin{bmatrix}\lambda\boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & \mathbf{0} & -\epsilon\boldsymbol{L}_q \\ \mathbf{0} & \lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I} & \mathbf{0} \\ \boldsymbol{L}_q^\top & \mathbf{0} & \lambda\boldsymbol{I} - 2c\epsilon\boldsymbol{I}\end{bmatrix} \\ &= \det(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I})\det\begin{bmatrix}\lambda\boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \lambda\boldsymbol{I} - 2c\epsilon\boldsymbol{I}\end{bmatrix}.\end{aligned}$$

This implies that the λ_j, which converges to zero as $\epsilon \to 0$, is a solution of the following equation

$$0 = \det\begin{bmatrix}\lambda\boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \lambda\boldsymbol{I} - 2c\epsilon\boldsymbol{I}\end{bmatrix}. \tag{9}$$

Now let us reparametrize (9) by $\lambda = \kappa\sqrt{\epsilon}$ to get

$$
\begin{aligned}
0 &= \det \begin{bmatrix} \kappa\sqrt{\epsilon}\boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \kappa\sqrt{\epsilon}\boldsymbol{I} - 2c\epsilon\boldsymbol{I} \end{bmatrix} \\
&= \sqrt{\epsilon}^{d_1} \det \begin{bmatrix} \kappa\boldsymbol{I} - \sqrt{\epsilon}\boldsymbol{A} + \sqrt{\epsilon}\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\sqrt{\epsilon}\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \kappa\sqrt{\epsilon}\boldsymbol{I} - 2c\epsilon\boldsymbol{I} \end{bmatrix} \\
&= \sqrt{\epsilon}^{d_1+q} \det \begin{bmatrix} \kappa\boldsymbol{I} - \sqrt{\epsilon}\boldsymbol{A} + \sqrt{\epsilon}\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \kappa\boldsymbol{I} - 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix}.
\end{aligned}
$$

Since $(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I})^{-1}$ converges to $\boldsymbol{D}^{-1}$ as $\epsilon \to 0$, we have that if $\lambda_j \to 0$ as $\epsilon \to 0$ then λ_j should be a solution of the equation

$$
0 = \det \begin{bmatrix} \kappa\boldsymbol{I} - \sqrt{\epsilon}\boldsymbol{A} + \sqrt{\epsilon}\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & (\kappa - 2c\sqrt{\epsilon})\boldsymbol{I} \end{bmatrix} \tag{10}
$$

By Lemma B.2, notice that eigenvalues divided by $\sqrt{\epsilon}$ converge to the solutions of

$$
0 = \det \begin{bmatrix} \kappa\boldsymbol{I} & -\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \kappa\boldsymbol{I} \end{bmatrix}. \tag{11}
$$

From the fact that $\boldsymbol{L}_q$ is of full column rank matrix, $\boldsymbol{L}_q$ has exactly q singular values. Therefore, solutions of (11) are nonzero, and those are exactly $i\sigma_k$, $k = 1, \ldots, q$ where σ_k are the nonzero singular values of $\boldsymbol{C}_2$, or equivalently singular values of $\boldsymbol{L}_q$, and therefore there are $2q$ instances among λ_j such that each λ_j has order of $\sqrt{\epsilon}$, and has asymptotic $+i\sigma_k\sqrt{\epsilon}$ or $-i\sigma_k\sqrt{\epsilon}$.

So far, we have shown that r eigenvalues have magnitude $\Theta(1)$, and $2q$ eigenvalues have magnitude $\Theta(\sqrt{\epsilon})$. On the other hand, we have

$$
\begin{aligned}
\det(\boldsymbol{\Phi}_{\mathrm{reg1},\tau}) &= \det \begin{bmatrix} \epsilon\boldsymbol{A} & \epsilon\boldsymbol{C}_1 & \epsilon\boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & \boldsymbol{D} + 2c\epsilon\boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{L}_q^\top & \boldsymbol{0} & 2c\epsilon\boldsymbol{I} \end{bmatrix} \\
&= \epsilon^{d_1} \det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & \boldsymbol{D} + 2c\epsilon\boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{L}_q^\top & \boldsymbol{0} & 2c\epsilon\boldsymbol{I} \end{bmatrix}.
\end{aligned} \tag{12}
$$

Here, since we assumed that $\boldsymbol{S}_{\mathrm{res}}$ is non-degenerate, by Lemma B.3, the RHS of (12) for $\epsilon = 0$ is not zero. Moreover, RHS of (12) is nonzero for sufficiently small ϵ by Lemma B.2. Therefore, the product of all λ_j of $\boldsymbol{\Phi}_{\mathrm{reg1},\tau}$ should be of order $\Theta(\epsilon^{d_1})$. From these two observations, we know that the product of the remaining $d_1 - q$ eigenvalues should be of order $\Theta(\epsilon^{d_1-q})$. And we claim that each of these $d_1 - q$ eigenvalues is exactly of order $\Theta(\epsilon)$. To this end, let us examine what properties would the eigenvalues of order $O(\epsilon)$ have. By reparametrizing $\lambda = \mu\epsilon$ in (8), we have

$$
\begin{aligned}
0 &= \det \begin{bmatrix} \mu\epsilon\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C}_1 & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & \mu\epsilon\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{L}_q^\top & \boldsymbol{0} & \mu\epsilon\boldsymbol{I} - 2c\epsilon\boldsymbol{I} \end{bmatrix} \\
&= \epsilon^{d_1} \det \begin{bmatrix} \mu\boldsymbol{I} - \boldsymbol{A} & -\boldsymbol{C}_1 & -\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & \mu\epsilon\boldsymbol{I} - \boldsymbol{D} - 2c\epsilon\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{L}_q & \boldsymbol{0} & \mu\epsilon\boldsymbol{I} - 2c\epsilon\boldsymbol{I} \end{bmatrix}.
\end{aligned}
$$

Then, by Lemma B.2, μ converges to a root of the equation

$$
0 = \det \begin{bmatrix} \mu\boldsymbol{I} - \boldsymbol{A} & -\boldsymbol{C}_1 & -\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & -\boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{L}_q & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \tag{13}
$$

as $\epsilon \to 0$.

Then, $\mu = 0$ cannot be a root of (13) by Lemma B.3. This implies that there is no λ_j of order $o(\epsilon)$, or equivalently, all eigenvalues of $\mathbf{\Phi}_{\mathrm{reg1},\tau}$ are of order $\Omega(\epsilon)$. Therefore, from the fact that a product of $d_1 - q$ eigenvalues of $\mathbf{\Phi}_{\mathrm{reg1},\tau}$ is of order $\Theta(\epsilon^{d_1-q})$, each of those eigenvalues is exactly order of $\Theta(\epsilon)$, then the claim follows.

Here, we summarize the following two facts implied by the previous discussions:

- If λ is an eigenvalue of order $O(\epsilon)$, then λ/ϵ converges to a solution of (13) as $\epsilon \to 0$.
- The right-hand side of (13) is a polynomial of degree $d_1 - q$ in μ, whose solutions are nonzero.

By Lemma B.3, it is now immediate that the $d_1 - q$ eigenvalues of $\mathbf{\Phi}_{\mathrm{reg1},\tau}$ that are of order $\Theta(\epsilon)$ is of the form $\lambda(\epsilon) = \mu\epsilon + o(\epsilon)$ for μ that is a solution of (13).

The final claim, asserting that $\lambda_j(\epsilon) \neq 0$ for any j can be deduced from the invertibility of $\boldsymbol{S}_{\mathrm{res}}$ by Lemmas B.2 and B.3. More precisely, we have $\det(\mathbf{\Phi}_{\mathrm{reg1},\tau}) = \det(\mathbf{\Lambda}_\tau)\det(D\mathbf{\Phi}_{\mathrm{reg},\tau})$. By Lemma B.2,

$$\det(D\mathbf{\Phi}_{\mathrm{reg},\tau}) \to \det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & \boldsymbol{D} & \mathbf{0} \\ -\boldsymbol{L}_q^\top & \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \text{as} \quad \epsilon \to 0.$$

Therefore, the invertibility of $\boldsymbol{S}_{\mathrm{res}}$ and Lemma B.3 imply that the right-hand side is not zero. Then for sufficiently large τ, the assertion $\lambda_j(\epsilon) \neq 0$ follows.

For the case where $\nabla^2_{\boldsymbol{yy}} f$ is non-degenerate in Assumption 3, the following Lemma completes the proof.

Lemma B.5. *Let $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ respectively be $d_1 \times d_1$ symmetric, $d_2 \times d_2$ non-degenerate symmetric, and $d_1 \times d_2$ matrices. Then, the $d_1 + d_2$ complex eigenvalues of $\begin{bmatrix} \epsilon\boldsymbol{A} & \epsilon\boldsymbol{C} \\ -\boldsymbol{C}^\top & -\boldsymbol{B} + 2c\epsilon\boldsymbol{I} \end{bmatrix}$ have the following asymptotics as $\epsilon = \frac{1}{\tau} \to 0+$:*

$$|\lambda_j - \epsilon\mu_j| = o(\epsilon), \quad j = 1, \dots, d_1, \qquad |\lambda_{j+d_1} - \nu_j| = o(1), \quad j = 1, \dots, d_2,$$

where $\{\mu_j\}_{j=1,\dots,d_1}$ and $\{\nu_j\}_{j=1,\dots,d_2}$ are the eigenvalues of $\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top$ and $-\boldsymbol{B}$, respectively

Proof. Mimicking the proof of (Jin et al., 2020, Lemma 40), we can demonstrate the statement as follows. By definition of eigenvalues, any eigenvalue λ of $\begin{bmatrix} \epsilon\boldsymbol{A} & \epsilon\boldsymbol{C} \\ -\boldsymbol{C}^\top & -\boldsymbol{B} + 2c\epsilon\boldsymbol{I} \end{bmatrix}$ is the roots of the characteristic equation

$$p_\epsilon(\lambda) := \det \begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C} \\ \boldsymbol{C}^\top & \lambda\boldsymbol{I} + \boldsymbol{B} - 2c\epsilon\boldsymbol{I} \end{bmatrix}$$

We can express the equation $p_\epsilon(\lambda)$ as follows.

$$p_\epsilon(\lambda) = p_0(\lambda) + \sum_{i=1}^{d_1+d_2} \epsilon^i p_i(\lambda)$$

where $p_0(\lambda) = \det \begin{bmatrix} \lambda\boldsymbol{I} & \mathbf{0} \\ \boldsymbol{C}^\top & \lambda\boldsymbol{I} + \boldsymbol{B} \end{bmatrix} = \lambda^{d_1}\det(\lambda\boldsymbol{I} + \boldsymbol{B})$ and $p_i(\lambda)$ for $i \geq 1$ are polynomials of order equal to or smaller than $d_1 + d_2$. Then, by Lemma B.2, the roots of $p_\epsilon(\lambda)$ are

$$\begin{aligned} |\lambda_j| &= o(1), \quad 1 \leq j \leq d_1, \\ |\lambda_{j+d_1} - \nu_j| &= o(1), \quad 1 \leq j \leq d_2, \end{aligned}$$

Since $\boldsymbol{B}$ is non-degenerate, λ_{j+d_1} for $1 \leq j \leq d_2$ are of $\Omega(1)$, and therefore only λ_j for $1 \leq j \leq d_1$ converge to zero as $\epsilon \to 0$. To investigate the eigenvalues that converge to zero further, we reparametrize $\lambda = \kappa\epsilon$, then we have

$$\begin{aligned} p_\epsilon(\kappa\epsilon) &= \det \begin{bmatrix} \kappa\epsilon\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C} \\ \boldsymbol{C}^\top & \kappa\epsilon\boldsymbol{I} + \boldsymbol{B} - 2c\epsilon\boldsymbol{I} \end{bmatrix} \\ &= \epsilon^{d_1} \det \begin{bmatrix} \kappa\boldsymbol{I} - \boldsymbol{A} & -\boldsymbol{C} \\ \boldsymbol{C}^\top & \kappa\epsilon\boldsymbol{I} + \boldsymbol{B} - 2c\epsilon\boldsymbol{I} \end{bmatrix}. \end{aligned}$$

This implies that the λ_j, for $1 \le j \le d_1$, divided by ϵ is a solution of the following equation

$$0 = \det \begin{bmatrix} \kappa \boldsymbol{I} - \boldsymbol{A} & -\boldsymbol{C} \\ \boldsymbol{C}^\top & \kappa\epsilon \boldsymbol{I} + \boldsymbol{B} - 2c\epsilon \boldsymbol{I} \end{bmatrix}.$$

and by Lemma B.2, it converge to a root of

$$0 = \det \begin{bmatrix} \kappa \boldsymbol{I} - \boldsymbol{A} & -\boldsymbol{C} \\ \boldsymbol{C}^\top & \boldsymbol{B} \end{bmatrix} \quad \text{as} \quad \epsilon \to 0,$$

which is an eigenvalue of $\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top$ by Lemma B.3. These arguments complete the proof. □

Therefore, there is no difference between the eigenvalues asymptotics of both the (vanilla) two-timescale EG (in Theorem 4.2) and the two-timescale EG with the "$\gamma = \tau$" explicit regularization. □

B.2. Proof of Theorem 5.2

Proof of Theorem 5.2. We first consider the case where $\boldsymbol{S}_{\text{res}}$ is non-degenerate in Assumption 3. Analogous to the observations in proof of Proposition 5.1, by replacing ϵ with $\sqrt{\epsilon}$, one can deduce that the eigenvalues of $\boldsymbol{H}_{\text{reg2},\tau}$ are either $2c\sqrt{\epsilon}$ or the nonzero eigenvalues of $\boldsymbol{\Phi}_{\text{reg2},\tau} \coloneqq \boldsymbol{\Lambda}_\tau \boldsymbol{\Phi}_{\text{reg},\sqrt{\tau}}$ where

$$\boldsymbol{\Phi}_{\text{reg},\sqrt{\tau}} \coloneqq \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & \boldsymbol{D} + 2c\sqrt{\epsilon}\boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{L}_q^\top & \boldsymbol{0} & 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix}.$$

Therefore, our next step is to characterize the eigenvalues of $\boldsymbol{\Phi}_{\text{reg2},\tau}$, and such eigenvalues are the solutions of the equation

$$0 = p_\epsilon(\lambda) \coloneqq \det \begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C}_1 & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & \lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{L}_q^\top & \boldsymbol{0} & (\lambda - 2c\sqrt{\epsilon})\boldsymbol{I} \end{bmatrix} = \det(\lambda\boldsymbol{I} - \boldsymbol{\Phi}_{\text{reg2},\tau}). \tag{14}$$

By Lemma B.2, constructing the functions $\lambda_j(\epsilon)$ so that they are continuous is possible, and the eigenvalues converge to the solutions of the equation

$$p_0(\lambda) = \det \begin{bmatrix} \lambda\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{C}_1^\top & \lambda\boldsymbol{I} - \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{L}_q^\top & \boldsymbol{0} & \lambda\boldsymbol{I} \end{bmatrix} = 0$$

as $\epsilon \to 0$. Hence, the r eigenvalues of $\boldsymbol{\Phi}_{\text{reg2},\tau}$ converge to the r nonzero eigenvalues of $-\boldsymbol{B}$, and the other $d_1 + q$ eigenvalues converge to zero, as $\epsilon \to 0$.

To investigate the order of eigenvalues that converges to zero further, we begin by observing that, whenever $|\lambda|$ and ϵ are small enough so that $\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I}$ is invertible, it holds that

$$\begin{aligned} \det(\lambda\boldsymbol{I} - \boldsymbol{\Phi}_{\text{reg2},\tau}) &= \det \begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C}_1 & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & \lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{L}_q^\top & \boldsymbol{0} & \lambda\boldsymbol{I} - 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix} \\ &= \det \begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & \boldsymbol{0} & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{0} & \lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{L}_q^\top & \boldsymbol{0} & \lambda\boldsymbol{I} - 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix} \\ &= \det(\lambda\boldsymbol{I} - \boldsymbol{D} + 2c\sqrt{\epsilon}\boldsymbol{I}) \det \begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} + 2c\sqrt{\epsilon}\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \lambda\boldsymbol{I} - 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix}. \end{aligned}$$

This implies that the λ_j, which converges to zero as $\epsilon \to 0$, is a solution of the following equation

$$0 = \det \begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \lambda\boldsymbol{I} - 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix}. \tag{15}$$

Now let us reparametrize (15) by $\lambda = \kappa\sqrt{\epsilon}$ to get

$$\begin{aligned}
0 &= \begin{bmatrix} \kappa\sqrt{\epsilon}\boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \kappa\sqrt{\epsilon}\boldsymbol{I} - 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix} \\
&= \sqrt{\epsilon}^{d_1} \det \begin{bmatrix} \kappa\boldsymbol{I} - \sqrt{\epsilon}\boldsymbol{A} + \sqrt{\epsilon}\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\sqrt{\epsilon}\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \kappa\sqrt{\epsilon}\boldsymbol{I} - 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix} \\
&= \sqrt{\epsilon}^{d_1+q} \det \begin{bmatrix} \kappa\boldsymbol{I} - \sqrt{\epsilon}\boldsymbol{A} + \sqrt{\epsilon}\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & \kappa\boldsymbol{I} - 2c\boldsymbol{I} \end{bmatrix}.
\end{aligned}$$

Since $(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I})^{-1}$ converges to $\boldsymbol{D}^{-1}$ as $\epsilon \to 0$, we have that if $\lambda_j \to 0$ as $\epsilon \to 0$ then λ_j should be a solution of the equation

$$0 = \det \begin{bmatrix} \kappa\boldsymbol{I} - \sqrt{\epsilon}\boldsymbol{A} + \sqrt{\epsilon}\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I})^{-1}\boldsymbol{C}_1^\top & -\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & (\kappa - 2c)\boldsymbol{I} \end{bmatrix} \tag{16}$$

By Lemma B.2, notice that eigenvalues divided by $\sqrt{\epsilon}$ converge to the solutions of

$$0 = \det \begin{bmatrix} \kappa\boldsymbol{I} & -\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top & (\kappa - 2c)\boldsymbol{I} \end{bmatrix}. \tag{17}$$

From the fact that $\boldsymbol{L}_q$ is of full column rank matrix, $\boldsymbol{L}_q$ has exactly q singular values. Therefore, solutions of (17) are nonzero, and those are exactly $c \pm \sqrt{c^2 - \sigma_k^2}$, $k = 1, \ldots, q$ where σ_k are the nonzero singular values of $\boldsymbol{C}_2$, or equivalently singular values of $\boldsymbol{L}_q$. Note that the $c \pm \sqrt{c^2 - \sigma_k^2}$ can be written as $c \pm i\sqrt{\sigma_k^2 - c^2}$ when $\sigma_k > c$, however, we will denote it as $c \pm \sqrt{c^2 - \sigma_k^2}$, to cover both cases. Therefore, there are $2q$ instances among λ_j such that each λ_j has order of $\sqrt{\epsilon}$, and has asymptotic $\sqrt{\epsilon}(c + \sqrt{c^2 - \sigma_k^2})$ or $\sqrt{\epsilon}(c - \sqrt{c^2 - \sigma_k^2})$.

So far, we have shown that r eigenvalues have magnitude $\Theta(1)$, and $2q$ eigenvalues have magnitude $\Theta(\sqrt{\epsilon}(c \pm \sqrt{c^2 - \sigma_k^2}))$. On the other hand, we have

$$\begin{aligned}
\det(\boldsymbol{\Phi}_{\text{reg}2,\tau}) &= \det \begin{bmatrix} \epsilon\boldsymbol{A} & \epsilon\boldsymbol{C}_1 & \epsilon\boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & -\boldsymbol{D} + 2c\sqrt{\epsilon}\boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{L}_q^\top & \boldsymbol{0} & 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix} \\
&= \epsilon^{d_1} \det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & -\boldsymbol{D} + 2c\sqrt{\epsilon}\boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{L}_q^\top & \boldsymbol{0} & 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix}.
\end{aligned} \tag{18}$$

Here, since we assumed that $\boldsymbol{S}_{\text{res}}$ is non-degenerate, by Lemma B.3, the RHS of (18) for $\epsilon = 0$ is not zero. Moreover, RHS of (18) is nonzero for sufficiently small ϵ by Lemma B.2. Therefore, the product of all λ_j of $\boldsymbol{\Phi}_{\text{reg}2,\tau}$ should be of order $\Theta(\epsilon^{d_1})$. From these two observations, we know that the product of the remaining $d_1 - q$ eigenvalues should be of order $\Theta(\epsilon^{d_1 - q})$. And we claim that each of these $d_1 - q$ eigenvalues is exactly of order $\Theta(\epsilon)$. To this end, let us examine what properties would the eigenvalues of order $O(\epsilon)$ have. By reparametrizing $\lambda = \mu\epsilon$ in (14), we have

$$\begin{aligned}
0 &= \det \begin{bmatrix} \mu\epsilon\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C}_1 & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & \mu\epsilon\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{L}_q^\top & \boldsymbol{0} & \mu\epsilon\boldsymbol{I} - 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix} \\
&= \epsilon^{d_1} \det \begin{bmatrix} \mu\boldsymbol{I} - \boldsymbol{A} & -\boldsymbol{C}_1 & -\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & \mu\epsilon\boldsymbol{I} - \boldsymbol{D} - 2c\sqrt{\epsilon}\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{L}_q & \boldsymbol{0} & \mu\epsilon\boldsymbol{I} - 2c\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix}.
\end{aligned}$$

Then, by Lemma B.2, μ converges to a root of the equation

$$0 = \det \begin{bmatrix} \mu\boldsymbol{I} - \boldsymbol{A} & -\boldsymbol{C}_1 & -\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & -\boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{L}_q & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \tag{19}$$

as $\epsilon \to 0$.

Since the $\boldsymbol{S}_{\text{res}}$ is non-degenerate, the $\mu = 0$ cannot be a root of (19) by Lemma B.3. This particularly showing that there is no λ_j of order $o(\epsilon)$, or equivalently, all eigenvalues of $\boldsymbol{\Phi}_{\text{reg}2,\tau}$ are of order $\Omega(\epsilon)$. Therefore, from the fact that a product of $d_1 - q$ eigenvalues of $\boldsymbol{\Phi}_{\text{reg}2,\tau}$ is of order $\Theta(\epsilon^{d_1-q})$, each of those eigenvalues is exactly order of $\Theta(\epsilon)$, then the claim follows.

Here, we want to summarize the following two facts implied by the previous discussions:

- If λ is an eigenvalue of order $O(\epsilon)$, then λ divided by ϵ converges to a solution of (19) as $\epsilon \to 0$.
- The right-hand side of (19) is a polynomial of degree $d_1 - q$ in μ, whose solutions are nonzero.

By Lemma B.3, it is now immediate that the $d_1 - q$ eigenvalues of $\boldsymbol{\Phi}_{\text{reg}2,\tau}$ that are of order $\Theta(\epsilon)$ is of the form $\lambda(\epsilon) = \mu\epsilon + o(\epsilon)$ for μ that is a solution of (19).

The final claim, asserting that $\lambda_j(\epsilon) \neq 0$ for any j, can be deduced from the invertibility of $\boldsymbol{S}_{\text{res}}$ with Lemmas B.2 and B.3. More precisely, we have $\det(\boldsymbol{\Phi}_{\text{reg}2,\tau}) = \det(\boldsymbol{\Lambda}_\tau)\det(D\boldsymbol{\Phi}_{\text{reg},\sqrt{\tau}})$. By Lemma B.2,

$$\det(D\boldsymbol{\Phi}_{\text{reg},\sqrt{\tau}}) \to \det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & \boldsymbol{D} & \boldsymbol{0} \\ -\boldsymbol{L}_q^\top & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \quad \text{as} \quad \epsilon \to 0.$$

Therefore, the invertibility of $\boldsymbol{S}_{\text{res}}$ and Lemma B.3 imply that the RHS is not zero. Then, the assertion $\lambda_j(\epsilon) \neq 0$ follows.

For the case where $\nabla^2_{\boldsymbol{yy}} f$ is non-degenerate in Assumption 3, mimicking the proof of Lemma B.5 completes the proof. □

B.3. Proof of Theorem 5.3

Proof of Theorem 5.3. We first consider the case where $\boldsymbol{S}_{\text{res}}$ is non-degenerate. Recall that, alternating saddle gradient operator is as follow

$$\boldsymbol{F}_{\text{alt},\gamma}(\boldsymbol{x},\boldsymbol{y}) \coloneqq (\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{y}), -\nabla_{\boldsymbol{y}} f(\boldsymbol{x} - \frac{\eta}{\gamma}\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{y}))$$

where the η is given step size.

Then the proposed alternating extragradient can be formulated as follows

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(z_k - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}_k)).$$

By the matrix version of chain rule, the Jacobian of the $\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z})$ at $\boldsymbol{z}^*$ is

$$\begin{aligned}\boldsymbol{H}_{\text{alt},\tau_k,\gamma_k} &= \begin{bmatrix} \epsilon_1\boldsymbol{A} & \epsilon_1\boldsymbol{C} \\ \begin{bmatrix} -\boldsymbol{C}^\top & -\boldsymbol{B}\end{bmatrix}\begin{bmatrix}\boldsymbol{I} - \eta\epsilon_2\boldsymbol{A} \\ 0\end{bmatrix} & \begin{bmatrix} -\boldsymbol{C}^\top & -\boldsymbol{B}\end{bmatrix}\begin{bmatrix}-\eta\epsilon_2\boldsymbol{C} \\ \boldsymbol{I}\end{bmatrix}\end{bmatrix} \\ &= \begin{bmatrix} \epsilon_1\boldsymbol{A} & \epsilon_1\boldsymbol{C} \\ -\boldsymbol{C}^\top + \eta\epsilon_2\boldsymbol{C}^\top\boldsymbol{A} & -\boldsymbol{B} + \eta\epsilon_2\boldsymbol{C}^\top\boldsymbol{C}\end{bmatrix}.\end{aligned}$$

where $\epsilon_1 = \dfrac{1}{\tau_k}$, $\epsilon_2 = \dfrac{1}{\gamma_k}$ and $D\boldsymbol{F} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ -\boldsymbol{C}^\top & -\boldsymbol{B}\end{bmatrix}$ for saddle gradient $\boldsymbol{F} = (\nabla_{\boldsymbol{x}} f, -\nabla_{\boldsymbol{y}} f)$.

Analogous to the proof of Proposition 5.1, one can show that $\boldsymbol{H}_{\text{alt},\tau_k,\gamma_k}$ is similar to

$$\boldsymbol{G}_{\text{alt},\tau_k,\gamma_k} = \begin{bmatrix} \epsilon_1\boldsymbol{A} & \epsilon_1\boldsymbol{C}_1 & \epsilon_1\boldsymbol{C}_2 \\ -\boldsymbol{C}_1^\top + \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{A} & \boldsymbol{D} + \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{C}_1 & \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{C}_2 \\ -\boldsymbol{C}_2^\top + \eta\epsilon_2\boldsymbol{C}_2^\top\boldsymbol{A} & \eta\epsilon_2\boldsymbol{C}_2^\top\boldsymbol{C}_1 & \eta\epsilon_2\boldsymbol{C}_2^\top\boldsymbol{C}_2\end{bmatrix},$$

under the same settings and notations.

Then, for $\tilde{\boldsymbol{Q}} = \mathrm{diag}\{\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{V}^\top\}$ we have

$$\begin{aligned}\tilde{\boldsymbol{Q}}\boldsymbol{G}_{\mathrm{alt},\tau_k,\gamma_k}\tilde{\boldsymbol{Q}}^\top &= \begin{bmatrix} \epsilon_1\boldsymbol{A} & \epsilon_1\boldsymbol{C}_1 & \epsilon_1\boldsymbol{C}_2\boldsymbol{V} \\ -\boldsymbol{C}_1^\top + \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{A} & \boldsymbol{D} + \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{C}_1 & \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{C}_2\boldsymbol{V} \\ -\boldsymbol{V}^\top\boldsymbol{C}_2^\top + \eta\epsilon_2\boldsymbol{V}^\top\boldsymbol{C}_2^\top\boldsymbol{A} & \eta\epsilon_2\boldsymbol{V}^\top\boldsymbol{C}_2^\top\boldsymbol{C}_1 & \eta\epsilon_2\boldsymbol{V}^\top\boldsymbol{C}_2^\top\boldsymbol{C}_2\boldsymbol{V}\end{bmatrix} \\ &= \begin{bmatrix} \epsilon_1\boldsymbol{A} & \epsilon_1\boldsymbol{C}_1 & \epsilon_1\boldsymbol{L}_q & \boldsymbol{0} \\ -\boldsymbol{C}_1^\top + \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{A} & \boldsymbol{D} + \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{C}_1 & \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{L}_q & \boldsymbol{0} \\ -\boldsymbol{L}_q^\top + \eta\epsilon_2\boldsymbol{L}_q^\top\boldsymbol{A} & \eta\epsilon_2\boldsymbol{L}_q^\top\boldsymbol{C}_1 & \eta\epsilon_2\boldsymbol{L}_q^\top\boldsymbol{L}_q & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\end{aligned}$$

and therefore

$$\begin{aligned}\det(\lambda\boldsymbol{I} - \boldsymbol{H}_{\mathrm{alt},\tau_k,\gamma_k}) &= \det(\lambda\boldsymbol{I} - \tilde{\boldsymbol{Q}}\boldsymbol{G}_{\mathrm{alt},\tau_k,\gamma_k}\tilde{\boldsymbol{Q}}^\top) \\ &= \lambda^{d_2-r-q}\det\left(\lambda\boldsymbol{I} - \begin{bmatrix} \epsilon_1\boldsymbol{A} & \epsilon_1\boldsymbol{C}_1 & \epsilon_1\boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top + \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{A} & \boldsymbol{D} + \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{C}_1 & \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{L}_q \\ -\boldsymbol{L}_q^\top + \eta\epsilon_2\boldsymbol{L}_q^\top\boldsymbol{A} & \eta\epsilon_2\boldsymbol{L}_q^\top\boldsymbol{C}_1 & \eta\epsilon_2\boldsymbol{L}_q^\top\boldsymbol{L}_q\end{bmatrix}\right).\end{aligned}$$

So, the eigenvalues of $\boldsymbol{H}_{\mathrm{alt},\tau_k,\gamma_k}$ are either zero or the eigenvalues of

$$\boldsymbol{\Phi}_{\mathrm{alt},\tau_k,\gamma_k} := \begin{bmatrix} \epsilon_1\boldsymbol{A} & \epsilon_1\boldsymbol{C}_1 & \epsilon_1\boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top + \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{A} & \boldsymbol{D} + \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{C}_1 & \eta\epsilon_2\boldsymbol{C}_1^\top\boldsymbol{L}_q \\ -\boldsymbol{L}_q^\top + \eta\epsilon_2\boldsymbol{L}_q^\top\boldsymbol{A} & \eta\epsilon_2\boldsymbol{L}_q^\top\boldsymbol{C}_1 & \eta\epsilon_2\boldsymbol{L}_q^\top\boldsymbol{L}_q\end{bmatrix}.$$

Therefore, characterizing the eigenvalues of $\boldsymbol{\Phi}_{\mathrm{alt},\tau_k,\gamma_k}$ is equivalent to characterizing the nonzero eigenvalues of $\boldsymbol{H}_{\mathrm{alt},\tau_k,\gamma_k}$. From now on, let $\tau_k = \tau$ and $\gamma_k = \sqrt{\tau}$. Then, the eigenvalues of $\boldsymbol{\Phi}_{\mathrm{alt},\tau,\sqrt{\tau}}$ are the solutions of the equation

$$0 = p_\epsilon(\lambda) := \det\begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C}_1 & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top - \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top\boldsymbol{A} & \lambda\boldsymbol{I} - \boldsymbol{D} - \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top\boldsymbol{C}_1 & -\eta\sqrt{\epsilon}\boldsymbol{C}_1^\top\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top - \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top\boldsymbol{A} & -\eta\sqrt{\epsilon}\boldsymbol{L}_q^\top\boldsymbol{C}_1 & \lambda\boldsymbol{I} - \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top\boldsymbol{L}_q\end{bmatrix} = \det(\lambda\boldsymbol{I} - \boldsymbol{\Phi}_{\mathrm{alt},\tau,\sqrt{\tau}}). \tag{20}$$

By Lemma B.2, constructing the functions $\lambda_j(\epsilon)$ so that they are continuous is possible, and the eigenvalues converge to the solutions of the equation

$$p_0(\lambda) = \det\begin{bmatrix} \lambda\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{C}_1^\top & \lambda\boldsymbol{I} - \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{L}_q & \boldsymbol{0} & \lambda\boldsymbol{I}\end{bmatrix} = 0.$$

as $\epsilon \to 0$. Hence, the r eigenvalues of $\boldsymbol{\Phi}_{\mathrm{alt},\tau,\sqrt{\tau}}$ converge to the r nonzero eigenvalues of $-\boldsymbol{B}$, and the other $d_1 + q$ eigenvalues converge to zero, as $\epsilon \to 0$.

To investigate the order of eigenvalues that converges to zero further, we begin by observing that, whenever $|\lambda|$ and ϵ are small enough so that $\lambda\boldsymbol{I} - \boldsymbol{D} - \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top\boldsymbol{C}_1$ is invertible, it holds that

$$\begin{aligned}\det(\lambda\boldsymbol{I} - \boldsymbol{\Phi}_{\mathrm{alt},\tau,\sqrt{\tau}}) &= \det\begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C}_1 & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top - \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top\boldsymbol{A} & \lambda\boldsymbol{I} - \boldsymbol{D} - \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top\boldsymbol{C}_1 & -\eta\sqrt{\epsilon}\boldsymbol{C}_1^\top\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top - \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top\boldsymbol{A} & -\eta\sqrt{\epsilon}\boldsymbol{L}_q^\top\boldsymbol{C}_1 & \lambda\boldsymbol{I} - \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top\boldsymbol{L}_q\end{bmatrix} \\ &= \det\begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} & -\epsilon\boldsymbol{C}_1 & -\epsilon\boldsymbol{L}_q \\ (1 - \frac{\lambda\eta}{\sqrt{\epsilon}})\boldsymbol{C}_1^\top & \lambda\boldsymbol{I} - \boldsymbol{D} & \boldsymbol{0} \\ (1 - \frac{\lambda\eta}{\sqrt{\epsilon}})\boldsymbol{L}_q^\top & \boldsymbol{0} & \lambda\boldsymbol{I}\end{bmatrix} \\ &= \det\begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon(1 - \frac{\lambda\eta}{\sqrt{\epsilon}})\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D})^{-1}\boldsymbol{C}_1^\top & \boldsymbol{0} & -\epsilon\boldsymbol{L}_q \\ \boldsymbol{0} & \lambda\boldsymbol{I} - \boldsymbol{D} & \boldsymbol{0} \\ (1 - \frac{\lambda\eta}{\sqrt{\epsilon}})\boldsymbol{L}_q^\top & \boldsymbol{0} & \lambda\boldsymbol{I}\end{bmatrix} \\ &= \det(\lambda\boldsymbol{I} - \boldsymbol{D})\det\begin{bmatrix} \lambda\boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon(1 - \frac{\lambda\eta}{\sqrt{\epsilon}})\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D})^{-1}\boldsymbol{C}_1^\top & -\epsilon\boldsymbol{L}_q \\ (1 - \frac{\lambda\eta}{\sqrt{\epsilon}})\boldsymbol{L}_q^\top & \lambda\boldsymbol{I}\end{bmatrix}.\end{aligned}$$

This implies that the λ_j, which converges to zero as $\epsilon \to 0$, is a solution of the following equation

$$0 = \det \begin{bmatrix} \lambda \boldsymbol{I} - \epsilon \boldsymbol{A} + \epsilon(1 - \frac{\lambda\eta}{\sqrt{\epsilon}})\boldsymbol{C}_1(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}\boldsymbol{C}_1^\top & -\epsilon \boldsymbol{L}_q \\ (1 - \frac{\lambda\eta}{\sqrt{\epsilon}})\boldsymbol{L}_q^\top & \lambda \boldsymbol{I} \end{bmatrix}. \tag{21}$$

Now let us reparametrize (21) by $\lambda = \kappa\sqrt{\epsilon}$ to get

$$\begin{aligned} 0 &= \begin{bmatrix} \kappa\sqrt{\epsilon}\boldsymbol{I} - \epsilon \boldsymbol{A} + \epsilon(1 - \kappa\eta)\boldsymbol{C}_1(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}\boldsymbol{C}_1^\top & -\epsilon \boldsymbol{L}_q \\ (1 - \kappa\eta)\boldsymbol{L}_q^\top & \kappa\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix} \\ &= \sqrt{\epsilon}^{d_1} \det \begin{bmatrix} \kappa \boldsymbol{I} - \sqrt{\epsilon}\boldsymbol{A} + \sqrt{\epsilon}(1 - \kappa\eta)\boldsymbol{C}_1(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}\boldsymbol{C}_1^\top & -\sqrt{\epsilon}\boldsymbol{L}_q \\ (1 - \kappa\eta)\boldsymbol{L}_q^\top & \kappa\sqrt{\epsilon}\boldsymbol{I} \end{bmatrix} \\ &= \sqrt{\epsilon}^{d_1+q} \det \begin{bmatrix} \kappa \boldsymbol{I} - \sqrt{\epsilon}\boldsymbol{A} + \sqrt{\epsilon}(1 - \kappa\eta)\boldsymbol{C}_1(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}\boldsymbol{C}_1^\top & -\boldsymbol{L}_q \\ (1 - \kappa\eta)\boldsymbol{L}_q^\top & \kappa\boldsymbol{I} \end{bmatrix}. \end{aligned}$$

Since $(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}$ converges as $\epsilon \to 0$, we have that if $\lambda_j \to 0$ as $\epsilon \to 0$ then λ_j should be a solution of the equation

$$0 = \det \begin{bmatrix} \kappa \boldsymbol{I} - \sqrt{\epsilon}\boldsymbol{A} + \sqrt{\epsilon}(1 - \kappa\eta)\boldsymbol{C}_1(\lambda \boldsymbol{I} - \boldsymbol{D})^{-1}\boldsymbol{C}_1^\top & -\boldsymbol{L}_q \\ (1 - \kappa\eta)\boldsymbol{L}_q^\top & \kappa\boldsymbol{I} \end{bmatrix} \tag{22}$$

By Lemma B.2, notice that eigenvalues divided by $\sqrt{\epsilon}$ converge to the solutions of

$$0 = \det \begin{bmatrix} \kappa \boldsymbol{I} & -\boldsymbol{L}_q \\ (1 - \kappa\eta)\boldsymbol{L}_q^\top & \kappa\boldsymbol{I} \end{bmatrix}. \tag{23}$$

From the fact that $\boldsymbol{L}_q$ is of full column rank matrix, $\boldsymbol{L}_q$ has exactly q singular values. Therefore, solutions of (23) are nonzero, and those are exactly $\frac{\eta\sigma_k^2}{2} \pm \frac{\sigma_k\sqrt{\eta^2\sigma_k^2 - 4}}{2}$, $k = 1, \ldots, q$ where σ_k are the nonzero singular values of $\boldsymbol{C}_2$, or equivalently singular values of $\boldsymbol{L}_q$. Note that the $\frac{\eta\sigma_k^2}{2} \pm \frac{\sigma_k\sqrt{\eta^2\sigma_k^2 - 4}}{2}$ can be written as $\frac{\eta\sigma_k^2}{2} \pm i\frac{\sigma_k\sqrt{4 - \eta^2\sigma_k^2}}{2}$ when $\eta\sigma_k < 2$, however, we will denote it as $\frac{\eta\sigma_k^2}{2} \pm \frac{\sigma_k\sqrt{\eta^2\sigma_k^2 - 4}}{2}$, to cover both cases. Therefore, there are $2q$ instances among λ_j such that each λ_j has order of $\sqrt{\epsilon}$, and has asymptotic, and has asymptotic $\frac{\eta\sigma_k^2}{2} + \frac{\sigma_k\sqrt{\eta^2\sigma_k^2 - 4}}{2}$ or $\frac{\eta\sigma_k^2}{2} - \frac{\sigma_k\sqrt{\eta^2\sigma_k^2 - 4}}{2}$.

So far, we have shown that r eigenvalues have magnitude $\Theta(1)$, and $2q$ eigenvalues have magnitude $\Theta(\sqrt{\epsilon}(\frac{\eta\sigma_k^2}{2} \pm \frac{\sigma_k\sqrt{\eta^2\sigma_k^2 - 4}}{2}))$. On the other hand, we have

$$\begin{aligned} \det(\boldsymbol{\Phi}_{\mathrm{alt},\tau,\sqrt{\tau}}) &= \epsilon^{d_1} \det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top + \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top\boldsymbol{A} & \boldsymbol{D} + \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top\boldsymbol{C}_1 & \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top\boldsymbol{L}_q \\ -\boldsymbol{L}_q^\top + \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top\boldsymbol{A} & \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top\boldsymbol{C}_1 & \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top\boldsymbol{L}_q \end{bmatrix} \\ &= \epsilon^{d_1} \det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & \boldsymbol{D} & \boldsymbol{0} \\ -\boldsymbol{L}_q & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \end{aligned} \tag{24}$$

Here, since we assumed that $\boldsymbol{S}_{\mathrm{res}}$ is non-degenerate, by Lemma B.3, the RHS of (24) is not zero, therefore, the product of all λ_j of $\boldsymbol{\Phi}_{\mathrm{alt},\tau,\sqrt{\tau}}$ should be of order $\Theta(\epsilon^{d_1})$. From these two observations, we know that product of the remaining $d_1 - q$ eigenvalues should be of order $\Theta(\epsilon^{d_1-q})$. And we claim that each of these $d_1 - q$ eigenvalues is exactly of order $\Theta(\epsilon)$. To this end, let us examine what properties would the eigenvalues of order $O(\epsilon)$ have. By reparametrizing $\lambda = \mu\epsilon$ in (20), we have

$$
\begin{aligned}
0 &= \det \begin{bmatrix} \mu\epsilon \boldsymbol{I} - \epsilon \boldsymbol{A} & -\epsilon \boldsymbol{C}_1 & -\epsilon \boldsymbol{L}_q \\ \boldsymbol{C}_1^\top - \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top \boldsymbol{A} & \mu\epsilon \boldsymbol{I} - \boldsymbol{D} - \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top \boldsymbol{C}_1 & -\eta\sqrt{\epsilon}\boldsymbol{C}_1^\top \boldsymbol{L}_q \\ \boldsymbol{L}_q^\top - \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top \boldsymbol{A} & -\eta\sqrt{\epsilon}\boldsymbol{L}_q^\top \boldsymbol{C}_1 & \mu\epsilon \boldsymbol{I} - \eta\sqrt{\epsilon}\boldsymbol{C}_q^\top \boldsymbol{L}_q \end{bmatrix} \\
&= \epsilon^{d_1} \det \begin{bmatrix} \mu \boldsymbol{I} - \boldsymbol{A} & -\boldsymbol{C}_1 & -\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top - \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top \boldsymbol{A} & \mu\epsilon \boldsymbol{I} - \boldsymbol{D} - \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top \boldsymbol{C}_1 & -\eta\sqrt{\epsilon}\boldsymbol{L}_q \\ \boldsymbol{L}_q^\top - \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top \boldsymbol{A} & -\eta\sqrt{\epsilon}\boldsymbol{L}_q^\top \boldsymbol{C}_1 & \mu\epsilon \boldsymbol{I} - \eta\sqrt{\epsilon}\boldsymbol{C}_q^\top \boldsymbol{L}_q \end{bmatrix}.
\end{aligned}
$$

Then, by Lemma B.2, μ converges to a root of the equation

$$
0 = \det \begin{bmatrix} \mu \boldsymbol{I} - \boldsymbol{A} & -\boldsymbol{C}_1 & -\boldsymbol{L}_q \\ \boldsymbol{C}_1^\top & -\boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{L}_q^\top & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \tag{25}
$$

as $\epsilon \to 0$.

Then, again $\mu = 0$ cannot be a root of (25) by Lemma B.3. This particularly showing that there is no λ_j of order $o(\epsilon)$, or equivalently, all eigenvalues of $\boldsymbol{\Phi}_{\mathrm{alt},\tau,\sqrt{\tau}}$ are of order $\Omega(\epsilon)$. Therefore, from the fact that a product of $d_1 - q$ eigenvalues of $\boldsymbol{\Phi}_{\mathrm{alt},\tau,\sqrt{\tau}}$ is of order $\Theta(\epsilon^{d_1-q})$, each of those eigenvalues is exactly order of $\Theta(\epsilon)$, then the claim follows.

Here, we want to summarize the previous discussions imply the following two facts:

- If λ is an eigenvalue of order $O(\epsilon)$, then λ divided by ϵ converges to a solution of (25) as $\epsilon \to 0$.
- The right-hand side of (25) is a polynomial of degree $d_1 - q$ in μ, whose solutions are nonzero.

By Lemma B.3, it is now immediate that the $d_1 - q$ eigenvalues of $\boldsymbol{\Phi}_{\mathrm{alt},\tau,\sqrt{\tau}}$ that are of order $\Theta(\epsilon)$, and is of the form $\lambda(\epsilon) = \mu\epsilon + o(\epsilon)$ for a μ which is solution of (25).

The final claim, asserting that $\lambda_j(\epsilon) \neq 0$ for any j, can be deduced from the invertibility of $\boldsymbol{S}_{\mathrm{res}}$ with Lemmas B.2 and B.3. More precisely, we have $\det(\boldsymbol{\Phi}_{\mathrm{alt},\tau,\sqrt{\tau}}) = \det(\boldsymbol{\Lambda}_\tau)\det(D\boldsymbol{\Phi}_{\mathrm{alt},\sqrt{\tau}})$ where

$$
D\boldsymbol{\Phi}_{\mathrm{alt},\sqrt{\tau}} := \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top + \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top \boldsymbol{A} & \boldsymbol{D} + \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top \boldsymbol{C}_1 & \eta\sqrt{\epsilon}\boldsymbol{C}_1^\top \boldsymbol{L}_q \\ -\boldsymbol{L}_q^\top + \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top \boldsymbol{A} & \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top \boldsymbol{C}_1 & \eta\sqrt{\epsilon}\boldsymbol{L}_q^\top \boldsymbol{L}_q \end{bmatrix}.
$$

Since

$$
\det(D\boldsymbol{\Phi}_{\mathrm{alt},\sqrt{\tau}}) = \det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C}_1 & \boldsymbol{L}_q \\ -\boldsymbol{C}_1^\top & \boldsymbol{D} & \boldsymbol{0} \\ -\boldsymbol{L}_q^\top & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \tag{26}
$$

the assumption and lemma B.3 implies that the RHS of (26) is not zero. Then, for any $\tau \geq 1$, the assertion $\lambda_j(\epsilon) \neq 0$ follows.

For the case where $\nabla^2_{\boldsymbol{yy}} f$ is non-degenerate in Assumption 3, following the proof of Lemma B.4 with the fact that $\det(\boldsymbol{H}_{\mathrm{alt},\tau,\sqrt{\tau}}) = \det \begin{bmatrix} \epsilon \boldsymbol{A} & \epsilon \boldsymbol{C} \\ -\boldsymbol{C}^\top & -\boldsymbol{B} \end{bmatrix} = \det(\boldsymbol{H}_\tau)$ completes the proof.

□

B.4. Relationship between the equilibrium of Alt2-EG-TS and the stationary point of F

Proposition B.6. *Under Assumption 1, a point $\boldsymbol{z}^*$ is an equilibrium point of Alt2-EG-TS* (6)

$$
\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta \boldsymbol{\Lambda}_{\tau_k} \boldsymbol{F}_{\mathrm{alt},\gamma_k}(\boldsymbol{z}_k - \eta \boldsymbol{\Lambda}_{\tau_k} \boldsymbol{F}_{\mathrm{alt},\gamma_k}(\boldsymbol{z}_k))
$$

if and only if $\boldsymbol{F}(\boldsymbol{z}^*) = \boldsymbol{0}$, for $0 < \eta < \frac{1}{\sqrt{2}L}$, $\tau_k \geq 1$ and $\gamma_k \geq 2$.

Proof. It is obvious that $\boldsymbol{F}(\boldsymbol{z}^*) = \mathbf{0}$ if and only if $\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*) = \mathbf{0}$. It is also straightforward that $\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*) = \mathbf{0}$ implies $\boldsymbol{z}^* - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^* - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*)) = \boldsymbol{z}^*$. We are now left to prove the "only if" statement.

Suppose that $\boldsymbol{z}^*$ is an equilibrium point of Alt2-EG-TS. Then, we have

$$\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^* - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*)) = \mathbf{0}.$$

For the sake of contradiction, suppose that $\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*) \neq \mathbf{0}$ and let $\boldsymbol{w}^* := \boldsymbol{z}^* - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*)$. Note that, by the assumption, $\boldsymbol{w}^* \neq \boldsymbol{z}^*$. Then, we have

$$\begin{aligned}\boldsymbol{z}^* - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*) &= \boldsymbol{z}^* - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^* - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*)) - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*)\\ &= \boldsymbol{w}^* - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{w}^*),\end{aligned}$$

hence we have $\boldsymbol{z}^* - \boldsymbol{w}^* = \eta\boldsymbol{\Lambda}_{\tau_k}(\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*) - \boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{w}^*))$.

Meanwhile, under Assumption 1, one can deduce that $\boldsymbol{F}_{\text{alt},\gamma_k}$ is $\sqrt{2}L$-Lipschitz for $0 < \eta < \frac{1}{\sqrt{2}L}$ and $\gamma_k \geq 2$, since

$$\begin{aligned}\|\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{u},\boldsymbol{v}) - \boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{x},\boldsymbol{y})\|^2 &\leq \left(L_{\boldsymbol{x}}^2 + L_{\boldsymbol{y}}^2\left(1+\frac{1}{\gamma_k}\right)\left(1+L_{\boldsymbol{x}}^2\frac{\eta^2}{\gamma_k}\right)\right)\|(\boldsymbol{u},\boldsymbol{v}) - (\boldsymbol{x},\boldsymbol{y})\|^2 \quad \text{(by Lemma B.1)}\\ &\leq \left(L_{\boldsymbol{x}}^2 + L_{\boldsymbol{y}}^2\left(1+\frac{1}{2}\right)\left(1+\frac{1}{4}\right)\right)\|(\boldsymbol{u},\boldsymbol{v}) - (\boldsymbol{x},\boldsymbol{y})\|^2\\ &\leq 2\left(L_{\boldsymbol{x}}^2 + L_{\boldsymbol{y}}^2\right)\|(\boldsymbol{u},\boldsymbol{v}) - (\boldsymbol{x},\boldsymbol{y})\|^2.\end{aligned} \tag{27}$$

Therefore, for $0 < \eta < \frac{1}{\sqrt{2}L}$, $\tau_k \geq 1$, and $\gamma_k \geq 2$, we have

$$\begin{aligned}\|\boldsymbol{z}^* - \boldsymbol{w}^*\| &= \eta\,\|\boldsymbol{\Lambda}_{\tau_k}(\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*) - \boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{w}^*))\|\\ &\leq \sqrt{2}\eta L\,\|\boldsymbol{\Lambda}_{\tau_k}\|\,\|\boldsymbol{z}^* - \boldsymbol{w}^*\|\\ &< \|\boldsymbol{z}^* - \boldsymbol{w}^*\|\end{aligned}$$

which is absurd. Therefore, we can deduce that $\boldsymbol{F}_{\text{alt},\gamma_k}(\boldsymbol{z}^*) = \mathbf{0}$. $\square$

B.5. Eigenvalue Asymptotic of $\boldsymbol{\Lambda}_\tau\boldsymbol{F}_{\text{alt},\gamma}$

We present the eigenvalues asymptotic of $\boldsymbol{\Lambda}_\tau\boldsymbol{F}_{\text{alt},\gamma}$. Following few statements in proof of Theorem 5.3, one can deduce that the eigenvalues of order $\Theta(\sqrt{\epsilon})$, divided by $\sqrt{\epsilon}$, are solutions of the equation

$$0 = \det\begin{bmatrix}\kappa\boldsymbol{I} - \sqrt{\epsilon}\boldsymbol{A} + \sqrt{\epsilon}(1-\kappa\sqrt{\epsilon}\eta)\boldsymbol{C}_1(\lambda\boldsymbol{I} - \boldsymbol{D})^{-1}\boldsymbol{C}_1^\top & -\boldsymbol{L}_q\\ (1-\kappa\sqrt{\epsilon}\eta)\boldsymbol{L}_q^\top & \kappa\boldsymbol{I}\end{bmatrix}$$

By Lemma B.2, notice that eigenvalues of order $\Theta(\sqrt{\epsilon})$, divided by $\sqrt{\epsilon}$, converge to the solutions of

$$0 = \det\begin{bmatrix}\kappa\boldsymbol{I} & -\boldsymbol{L}_q\\ \boldsymbol{L}_q^\top & \kappa\boldsymbol{I}\end{bmatrix},$$

as $\epsilon \to 0$. Therefore, these eigenvalues have asymptotics that are equivalent to those of the type (i) eigenvalues in Theorem 4.2.

B.6. Proof of Theorem 5.4

Proof of Proposition 5.4. We first investigate the necessary and/or sufficient condition for each type of eigenvalues λ_j of $\boldsymbol{H}_{\text{alt2},\tau} = \boldsymbol{H}_{\text{alt},\tau,\sqrt{\tau}}$, which is categorized in Theorem 5.3, to lie in $\mathcal{P}_\eta$.

(i) First, consider the type (i) eigenvalues $\lambda_j = \left(\frac{\eta\sigma_k^2}{2} \pm \frac{\sqrt{\eta^2\sigma_k^4 - 4\sigma_k^2}}{2}\right)\sqrt{\epsilon} + o(\sqrt{\epsilon})$ for some k. The radicand of $\sqrt{\eta^2\sigma_k^4 - 4\sigma_k^2}$ is negative for $0 < \eta < \frac{1}{L}$, since $\eta\sigma_k < \frac{1}{L}\cdot L = 1$. Therefore, the type (i) eigenvalues is in a form $\lambda_j = \left(\frac{\eta\sigma_k^2}{2} \pm i\frac{\sqrt{4\sigma_k^2 - \eta^2\sigma_k^4}}{2}\right)\sqrt{\epsilon} + o(\sqrt{\epsilon})$. Here, the leading term of real part of λ_j is $\frac{\eta\sigma_k^2}{2}\sqrt{\epsilon}$ that is positive, so $\lambda_j(\epsilon) \in \mathcal{P}_\eta$ for sufficiently large τ.

(ii) Secondly, consider the type (ii) eigenvalues, $\lambda_j = \epsilon\mu_k + o(\epsilon)$ for some k. Since the coefficient of leading term is μ_k, for sufficiently small ϵ, $\mu_k > 0$ implies that $\lambda_j(\epsilon) \in \mathcal{P}_\eta$, and $\mu_k < 0$ implies that $\lambda_j(\epsilon) \notin \mathcal{P}_\eta$. Recall that the μ_k are the eigenvalues of the restricted Schur complement $\boldsymbol{S}_{\text{res}}$ in Theorem 5.3, which are nonzero due to Assumption 3′. Therefore, $\boldsymbol{S}_{\text{res}} \succeq \mathbf{0}$ if and only if there exists some $\tau^\star$ such that $\tau > \tau^\star$ implies that every λ_j of order $\Theta(\epsilon)$ satisfies $\lambda_j(\epsilon) \in \mathcal{P}_\eta$.

(iii) Finally, consider the type (iii) eigenvalues, $\lambda_j = \nu_k + o(1)$ for some k. By the inequality

$$\|\boldsymbol{B}\| = \left\| \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B} \end{bmatrix} \right\| = \left\| \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ -\boldsymbol{C}^\top & \boldsymbol{B} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{bmatrix} \right\| \leq \|D\boldsymbol{F}(\boldsymbol{z}^*)\| = L,$$

we have $\|\nu_k\| \leq L$ for the eigenvalues of $-\boldsymbol{B}$. Now, let $0 < \eta < \frac{1}{L}$, and suppose that $\boldsymbol{B} \not\preceq \mathbf{0}$. Then, there exists some k such that $-L \leq \nu_k < 0$. Since the half-open interval $[-L, 0)$ is contained in the complement of $\bar{\mathcal{P}}_\eta$, for sufficiently large τ, we have $\lambda_j(\epsilon) \notin \mathcal{P}_\eta$ clearly. On the other hand, suppose that $\boldsymbol{B} \preceq \mathbf{0}$. Then $\nu_k > 0$ for all k, and it implies that for $0 < \eta < \frac{1}{L}$ and sufficiently large τ, we have $\lambda_j(\epsilon) \in \mathcal{P}_\eta$ for all j such that $\lambda_j(\epsilon) = \nu_k + o(1)$. Therefore, $\boldsymbol{B} \preceq \mathbf{0}$ if and only if there exists some $0 < \eta < \frac{1}{L}$ such that τ being sufficiently large implies that every λ_j of order $\Theta(1)$ satisfies $\lambda_j(\epsilon) \in \mathcal{P}_\eta$.

Combining all the previous discussions, we can conclude that $\boldsymbol{S}_{\text{res}} \succeq \mathbf{0}$ and $\boldsymbol{B} \preceq \mathbf{0}$ if and only if, for any η satisfying $0 < \eta < \frac{1}{L}$, there exists sufficiently large τ such that $\lambda_j(\epsilon) \in \mathcal{P}_\eta$ for all j. The conclusion then follows from the Proposition 4.1.

□

B.7. Proof of Theorem 5.5

In proving Theorem 5.5, we need the following results.

Proposition B.7. *Under Assumption 1, $f \in C^2$, $0 < \eta < \frac{\sqrt{5}-1}{2\sqrt{2}L}$ and $\tau \geq 4$, we have $\det(D\boldsymbol{w}_{\text{alt},\tau,\sqrt{\tau}}(\boldsymbol{z})) \neq 0$ for all $\boldsymbol{z}$.*

Proof. We begin by observing that

$$D\boldsymbol{w}_{\text{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}) = \boldsymbol{I} - \eta\boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z} - \eta\boldsymbol{\Lambda}_\tau\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z}))(\boldsymbol{I} - \eta\boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z})).$$

Under Assumption 1 and by (27) with $\frac{\sqrt{5}-1}{2\sqrt{2}L} < \frac{1}{\sqrt{2}L}$, we have $\|D\boldsymbol{F}_{\text{alt},\sqrt{\tau}}\| \leq \sqrt{2}L$. Hence, whenever $0 < \eta < \frac{\sqrt{5}-1}{2\sqrt{2}L}$ and $\tau \geq 4$, we obtain the bound

$$\begin{aligned}
&\left\|\eta\boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z} - \eta\boldsymbol{\Lambda}_\tau\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z}))(\boldsymbol{I} - \eta\boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z}))\right\| \\
&\leq \left\|\eta\boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z} - \eta\boldsymbol{\Lambda}_\tau\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z}))\right\| \left\|\boldsymbol{I} - \eta\boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z})\right\| \\
&\leq \sqrt{2}\eta L(1 + \sqrt{2}\eta L) \\
&< 1.
\end{aligned}$$

It follows that any eigenvalue of $\eta\boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z} - \eta\boldsymbol{\Lambda}_\tau\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z}))(\boldsymbol{I} - \eta\boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\text{alt},\sqrt{\tau}}(\boldsymbol{z}))$ has its magnitude strictly less than 1, and hence, $D\boldsymbol{w}_{\text{alt},\tau,\sqrt{\tau}}(\boldsymbol{z})$ cannot have zero eigenvalue. Therefore, $\det(D\boldsymbol{w}_{\text{alt},\tau,\sqrt{\tau}}(\boldsymbol{z})) \neq 0$ holds. □

For convenience, let us define a subset of the complex plane, for a real negative constant $a < 0$,

$$\mathcal{O}_a^\sharp := \left\{ z \in \mathbb{C} \ : \ |z - a| < \frac{|a|}{2} \right\},$$

which an open disk centered at a with radius $\frac{|a|}{2}$.

Lemma B.8. *$\mathcal{O}_a^\sharp \cap \mathcal{P}_\eta = \varnothing$ for any real negative constant $a < 0$ and real positive constant $\eta > 0$.*

Proof. Noticing that $\mathcal{O}_a$ lies in the left half plane, the only region to care about is $\mathbb{C}_-^\circ$. Thus, if the disk $\mathcal{O}_a$ and the peanut-shaped $\mathcal{P}_\eta$ do not intersect on that region, then the assertion follows immediately.

Consider a circle centered at origin with radius R, denoted by $\mathcal{O}^*$. Then, if the circle $\mathcal{O}^*$ and the boundary of $\mathcal{O}_a$ intersects, then it intersects at a point z_1 with a real part $\operatorname{Re} z_1 = \frac{R^2}{2a} + \frac{3a}{8}$. Similarly, if the circle $\mathcal{O}^*$ and the boundary of $\mathcal{P}_\eta$ intersects, then it intersects at a point z_2 with a real part $\operatorname{Re} z_2 = \frac{1}{4\eta} + \frac{\eta R^2}{4} - \sqrt{\frac{1}{16\eta^2} - \frac{3\eta^2 R^4}{16} + \frac{3R^2}{8}}$. For $\mathcal{O}_a$ and $\mathcal{P}_\eta$ to have an overlap, there must exist some R such that $\operatorname{Re} z_1 = \operatorname{Re} z_2$. We show that such R does not exist for any $a < 0$ and $\eta > 0$, by proving the following statement

$$\operatorname{Re} z_2 - \operatorname{Re} z_1 = \frac{1}{4\eta} + \frac{\eta R^2}{4} - (\frac{R^2}{2a} + \frac{3a}{8}) - \sqrt{\frac{1}{16\eta^2} - \frac{3\eta^2 R^4}{16} + \frac{3R^2}{8}} > 0$$

for any $a < 0$, $\eta > 0$ and $R > 0$. This is done by showing that the following

$$\begin{aligned}&\left(\frac{1}{4\eta} + \frac{\eta R^2}{4} - \frac{R^2}{2a} - \frac{3a}{8}\right)^2 - \left(\frac{1}{16\eta^2} - \frac{3\eta^2 R^4}{16} + \frac{3R^2}{8}\right) \\ &= \frac{9a^2}{64} - \frac{3a}{16\eta} + R^2\left(\frac{1}{8} - \frac{3a\eta}{16} - \frac{1}{4a\eta}\right) + R^4\left(\frac{1}{4a^2} - \frac{\eta}{4a} + \frac{\eta^2}{4}\right) > 0\end{aligned}$$

holds for any negative a with positive η and R, and this concludes the proof. □

Definition 5. *Given a C^1 mapping $\boldsymbol{w}$, the set $\mathcal{A}^*(\boldsymbol{w}) \coloneqq \{\boldsymbol{z}^* : \boldsymbol{z}^* = \boldsymbol{w}(\boldsymbol{z}^*),\ \rho(D\boldsymbol{w}(\boldsymbol{z}^*)) > 1\}$ is the set of* **strictly unstable** *equilibrium points.*

Theorem B.9 (Lee et al. (2019, Theorem 2))**.** *Let $\boldsymbol{w}$ be a C^1 mapping such that $\det(D\boldsymbol{w}(\boldsymbol{z})) \neq \mathbf{0}$ for all $\boldsymbol{z}$. Then the set of initial points that converge to a unstable equilibrium point has (Lebesgue) measure zero, i.e., $\mu(\{\boldsymbol{z}_0 : \lim_{k\to\infty} \boldsymbol{w}^k(\boldsymbol{z}_0) \in \mathcal{A}^*(\boldsymbol{w})\}) = 0$.*

Proposition B.10. *Let $\boldsymbol{z}^*$ be a strict non-minimax point i.e., $\boldsymbol{z}^* \in \mathcal{T}^*$. Under Assumptions 3, there exists a positive constant $\tau^\star > 0$ such that $\boldsymbol{z}^* \in \mathcal{A}^*(\boldsymbol{w}_{\mathrm{alt},\tau,\sqrt{\tau}})$ for any $\tau > \tau^\star$.*

Proof. By Proposition 4.1, we have

$$\begin{aligned}\mathcal{A}^*(\boldsymbol{w}_{\mathrm{alt},\tau,\sqrt{\tau}}) &= \{\boldsymbol{z}^* \ : \ \boldsymbol{z}^* = \boldsymbol{w}_{\mathrm{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}^*),\ \rho(D\boldsymbol{w}_{\mathrm{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}^*)) > 1\} \\ &= \{\boldsymbol{z}^* \ : \ \boldsymbol{z}^* = \boldsymbol{w}_{\mathrm{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}^*),\ \exists\lambda \in \operatorname{spec}(\boldsymbol{H}_{\mathrm{alt2},\tau}(\boldsymbol{z}^*)) \ \text{ s.t. } \ \lambda \notin \bar{\mathcal{P}}_\eta\}.\end{aligned}$$

For any strict non-minimax point $\boldsymbol{z}^* \in \mathcal{T}^*$, either $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{z}^*)$ or $-\boldsymbol{B}(\boldsymbol{z}^*)$ has at least one strictly negative eigenvalue. First, suppose that $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{z}^*)$ has a strictly negative eigenvalue $\mu < 0$. By Theorem 5.3, there exists a constant $\tau^\star$ such that at least one eigenvalue of $\boldsymbol{H}_{\mathrm{alt2},\tau}(\boldsymbol{z}^*)$ lies in a disk $\mathcal{O}^\sharp_{\mu\epsilon}$ for any $\tau > \tau^\star$. So by Lemma B.8, we would have $\mathcal{O}^\sharp_{\mu\epsilon} \cap \mathcal{P}_\eta = \varnothing$. On the other hand, suppose that $-\boldsymbol{B}(\boldsymbol{z}^*)$ has a strictly negative eigenvalue $\nu < 0$. Similarly, by Theorem 5.3, there exists a constant $\tau^\star$ such that at least one eigenvalue of $\boldsymbol{H}_{\mathrm{alt2},\tau}(\boldsymbol{z}^*)$ lies in a disk $\mathcal{O}^\sharp_\nu$ for any $\tau > \tau^\star$. So by Lemma B.8, we would have $\mathcal{O}^\sharp_\nu \cap \mathcal{P}_\eta = \varnothing$. Therefore, we can conclude that for any $\boldsymbol{z}^* \in \mathcal{T}^*$, there exists a constant $\tau^\star$ such that $\boldsymbol{z}^* \in \mathcal{A}^*(\boldsymbol{w}_{\mathrm{alt},\tau,\sqrt{\tau}})$ for any $\tau > \tau^\star$. □

Proof of Theorem 5.5. Because $\boldsymbol{z}^* \in \mathcal{A}^*(\boldsymbol{w}_{\mathrm{alt},\tau,\sqrt{\tau}})$ implies

$$\left\{\boldsymbol{z}_0 : \lim_{k\to\infty} \boldsymbol{w}^k_{\mathrm{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}_0) = \boldsymbol{z}^*\right\} \subset \left\{\boldsymbol{z}_0 : \lim_{k\to\infty} \boldsymbol{w}^k_{\mathrm{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}_0) \in \mathcal{A}^*(\boldsymbol{w}_{\mathrm{alt},\tau,\sqrt{\tau}})\right\},$$

by Theorem B.9, there exists a positive constant $\tau^\star > 0$ such that

$$\mu\left(\left\{\boldsymbol{z}_0 : \lim_{k\to\infty} \boldsymbol{w}^k_{\mathrm{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}_0) = \boldsymbol{z}^*\right\}\right) = 0$$

for any $\tau > \tau^\star$.

Moreover, if $\mathcal{T}^*$ is finite, then a maximum of $\tau^\star$ for all $\boldsymbol{z}^* \in \mathcal{T}^*$ is also finite. Let us denote such maximum by $\tau^*_{\max}$. Then, for any $\tau > \tau^\star_{\max}$ we have $\mathcal{T}^* \subset \mathcal{A}^*(\boldsymbol{w}_{\mathrm{alt},\tau,\sqrt{\tau}})$. This implies that

$$\left\{\boldsymbol{z}_0 : \lim_{k\to\infty} \boldsymbol{w}^k_{\mathrm{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}_0) \in \mathcal{T}^*\right\} \subset \left\{\boldsymbol{z}_0 : \lim_{k\to\infty} \boldsymbol{w}^k_{\mathrm{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}_0) \in \mathcal{A}^*(\boldsymbol{w}_{\mathrm{alt},\tau,\sqrt{\tau}})\right\},$$

for any $\tau > \tau^\star_{\max}$, and by Theorem B.9, we can conclude that

$$\mu\left(\left\{\boldsymbol{z}_0 : \lim_{k\to\infty} \boldsymbol{w}^k_{\mathrm{alt},\tau,\sqrt{\tau}}(\boldsymbol{z}_0) \in \mathcal{T}^*\right\}\right) = 0 \qquad \square.$$

C. Proofs for Section 6

C.1. Proof of Theorem 6.1

In proving the theorem, we begin with the following observations. Recall that, the dynamical system of the Alt2-EG-ITS method at time k is as follows

$$\boldsymbol{w}_{\mathrm{alt},\tau_k,\sqrt{\tau_k}}(\boldsymbol{z}) = \boldsymbol{z} - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\mathrm{alt},\sqrt{\tau_k}}(\boldsymbol{z} - \eta\boldsymbol{\Lambda}_{\tau_k}\boldsymbol{F}_{\mathrm{alt},\sqrt{\tau_k}}(\boldsymbol{z})).$$

Then, by the Taylor expansion around $\boldsymbol{z}^*$, we have

$$\begin{aligned}\boldsymbol{w}_{\mathrm{alt},\tau_k,\sqrt{\tau_k}}(\boldsymbol{z}) &= \boldsymbol{w}_{\mathrm{alt},\tau_k,\sqrt{\tau_k}}(\boldsymbol{z}^*) + D\boldsymbol{w}_{\mathrm{alt},\tau_k,\sqrt{\tau_k}}(\boldsymbol{z}^*)(\boldsymbol{z}-\boldsymbol{z}^*) + o(\boldsymbol{z}-\boldsymbol{z}^*)\\ &= \boldsymbol{z}^* + \left(\boldsymbol{I} - \eta\boldsymbol{\Lambda}_{\tau_k}D\boldsymbol{F}_{\mathrm{alt},\sqrt{\tau_k}}(\boldsymbol{z}^*) + \eta^2(\boldsymbol{\Lambda}_{\tau_k}D\boldsymbol{F}_{\mathrm{alt},\sqrt{\tau_k}}(\boldsymbol{z}^*))^2\right)(\boldsymbol{z}-\boldsymbol{z}^*) + o(\boldsymbol{z}-\boldsymbol{z}^*) \end{aligned} \tag{28}$$

For convenience, let us denote $\boldsymbol{x}_k \coloneqq \boldsymbol{z}_k - \boldsymbol{z}^*$, $\boldsymbol{y}_k \coloneqq o(\boldsymbol{z}_k - \boldsymbol{z}^*)$ and

$$\boldsymbol{A}_k \coloneqq \boldsymbol{I} - \eta\boldsymbol{\Lambda}_{\tau_k}D\boldsymbol{F}_{\mathrm{alt},\sqrt{\tau_k}}(\boldsymbol{z}^*) + \eta^2(\boldsymbol{\Lambda}_{\tau_k}D\boldsymbol{F}_{\mathrm{alt},\sqrt{\tau_k}}(\boldsymbol{z}^*))^2. \tag{29}$$

Then, for arbitrary k_0 such that $0 \le k_0 < k$, we have

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \boldsymbol{w}_{\mathrm{alt},\tau_k,\sqrt{\tau_k}}(\boldsymbol{z}_k) - \boldsymbol{z}^*\\ &= \boldsymbol{A}_k\boldsymbol{x}_k + \boldsymbol{y}_k\\ &= \boldsymbol{A}_k(\boldsymbol{A}_{k-1}\boldsymbol{x}_{k-1} + \boldsymbol{y}_{k-1}) + \boldsymbol{y}_k\\ &= \cdots\\ &= \left(\prod_{j=k_0}^{k}\boldsymbol{A}_j\right)\boldsymbol{x}_{k_0} + \sum_{i=k_0}^{k}\left(\prod_{j=i+1}^{k}\boldsymbol{A}_j\right)\boldsymbol{y}_i\end{aligned}$$

which implies that

$$\|\boldsymbol{x}_{k+1}\| \le \left(\prod_{j=k_0}^{k}\|\boldsymbol{A}_j\|\right)\|\boldsymbol{x}_{k_0}\| + \sum_{i=k_0}^{k}\left(\prod_{j=i+1}^{k}\|\boldsymbol{A}_j\|\right)\|\boldsymbol{y}_i\|. \tag{30}$$

Then, the rest of the proof is to show that, for a certain increasing sequence of τ_k, the RHS of (30) (and thus $\|\boldsymbol{x}_k\|$) decreases at a certain rate. Note that, for a simpler setting, where $\boldsymbol{A}_k$ is fixed for all k with $\|\boldsymbol{A}_k\| < 1$, Polyak (1987, Theorem 2.1.2.1) showed in a few lines that the RHS of (30) (and thus $\|\boldsymbol{x}_k\|$) decreases at an exponential rate. However, our proof is not as straightforward as that of (Polyak, 1987, Theorem 2.1.2.1), since here we not only consider $\boldsymbol{A}_k$ that varies over time k, but also satisfies $\lim_{k\to\infty}\|\boldsymbol{A}_k\| = 1$ as shown next.

Lemma C.1. *Suppose Assumptions 1 and 3′ hold, and let $\boldsymbol{z}^*$ be an equilibrium point that satisfies the second-order necessary condition of local minimax points. Then, there exists $K > 0$ such that the matrix $\boldsymbol{A}_k$ (29) of the Alt2-EG-ITS with $(\tau_k, \sqrt{\tau_k})$, where τ_k is increasing and $\lim_{k\to\infty}\tau_k = \infty$, for any $0 < \eta < 1/L$, satisfies*

$$\|\boldsymbol{A}_k\| \le 1 - \frac{2}{\tau_k^{1+c}}$$

for all $k \ge K$ and for any $c > 0$.

Proof. Recall that, under Assumption 3′, Theorem 5.3 implies that there are three types of eigenvalue asymptotics of $\boldsymbol{H}_{\mathrm{alt2},\tau} \coloneqq \boldsymbol{\Lambda}_\tau D\boldsymbol{F}_{\mathrm{alt},\sqrt{\tau}}$. This directly implies that the matrix $\boldsymbol{A}_k$ has the following three types of eigenvalue asymptotics

$$
\begin{aligned}
&\text{(i)}\ \ 1-\eta\left(\left(\frac{\eta\sigma_j^2}{2}\pm i\frac{\sqrt{4\sigma_j^2-\eta^2\sigma_j^4}}{2}\right)\frac{1}{\sqrt{\tau_k}}+o\left(\frac{1}{\sqrt{\tau_k}}\right)\right)+\eta^2\left(\left(\frac{\eta\sigma_j^2}{2}\pm i\frac{\sqrt{4\sigma_j^2-\eta^2\sigma_j^4}}{2}\right)\frac{1}{\sqrt{\tau_k}}+o\left(\frac{1}{\sqrt{\tau_k}}\right)\right)^2\\
&\quad=1-\frac{\eta^2\sigma_j^2}{2\sqrt{\tau_k}}+\frac{\eta^4\sigma_j^4}{4\tau_k}-\eta^2\frac{4\sigma_j^2-\eta^2\sigma_j^4}{4\tau_k}+i\left(\mp\eta\frac{\sqrt{4\sigma_j^2-\eta^2\sigma_j^4}}{2\sqrt{\tau_k}}\pm\frac{\eta^3\sigma_j^2\sqrt{4\sigma_j^2-\eta^2\sigma_j^4}}{2\tau_k}\right)\\
&\qquad-\eta o\left(\frac{1}{\sqrt{\tau_k}}\right)+\eta^2 o\left(\frac{1}{\tau_k}\right)+\eta^3\sigma_m^2\frac{1}{\sqrt{\tau_k}}o\left(\frac{1}{\sqrt{\tau_k}}\right)\pm i\frac{\eta^2\sqrt{4\sigma_m^2-\eta^2\sigma_m^4}}{2\sqrt{\tau_k}}o\left(\frac{1}{\sqrt{\tau_k}}\right),
\end{aligned}
$$

$$
\text{(ii)}\ \ 1-\eta\left(\frac{\mu_j}{\tau_k}+o\left(\frac{1}{\tau_k}\right)\right)+\eta^2\left(\frac{\mu_j}{\tau_k}+o\left(\frac{1}{\tau_k}\right)\right)^2=1-\frac{1}{\tau_k}(\eta\mu_j+o(1))+\frac{1}{\tau_k^2}(\eta^2\mu_j^2+o(1)),
$$

$$
\text{(iii)}\ \ 1-\eta(\nu_j+o(1))+\eta^2(\nu_j+o(1))^2,
$$

where μ_j are the eigenvalues of the restricted Schur complement $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{H})$, ν_j are the nonzero eigenvalues of $-\boldsymbol{B}$, and σ_j are the singular values of $\boldsymbol{C}_2$. Note that, under Assumption 3′, $\boldsymbol{S}_{\mathrm{res}}(\boldsymbol{H})$ is invertible.

Since the sequence τ_k is increasing and $\lim_{k\to\infty}\tau_k=\infty$, for any $c>0$, there exists sufficiently large K such that the followings satisfy

$$
\left|o\left(\frac{1}{\sqrt{\tau_k}}\right)\right|\le\min_j\left(\frac{\eta\sigma_j^2}{8\sqrt{\tau_k}},\frac{\sqrt{4\sigma_j^2-\eta^2\sigma_j^4}}{8}\right) \tag{31}
$$

$$
\left|o\left(\frac{1}{\tau_k}\right)\right|\le\frac{\mu_j}{4\tau_k}, \tag{32}
$$

$$
6.25\eta\mu_j\le\tau_k, \tag{33}
$$

$$
\max_j\left(\frac{4}{\eta^2\sigma_j^2},\frac{8}{\eta\mu_j}\right)\le\tau_k^c, \tag{34}
$$

$$
\max_j\left(6.5\eta^2\sigma_j^2+\eta^3\sigma_j^2+1,5\eta^2\sigma_j^2+\frac{\eta^3\sigma_j^2}{2}+\frac{\eta\sqrt{4\sigma_j^2-\eta^2\sigma_j^4}}{8}\right)\le\sqrt{\tau_k}, \tag{35}
$$

$$
1-\frac{1}{2}\eta\nu_j(1-\eta\nu_j)\le1-\frac{2}{\tau_k^{1+c}}, \tag{36}
$$

$$
|1-\eta(\nu_j+o(1))+\eta^2(\nu_j+o(1))^2|\le1-\frac{\eta\nu_j(1-\eta\nu_j)}{2}, \tag{37}
$$

for all $k>K$, and for all μ_j and σ_j. Using the above inequalities for any $k>K$, we characterize the type (i), (ii) and (iii) eigenvalues as follows. First, for the type (i) eigenvalue, we have

$$
\begin{aligned}
|\lambda| &\leq \sqrt{\left(1-\frac{\eta^2\sigma_j^2}{4\sqrt{\tau_k}}\right)^2+\left(\eta\frac{\sqrt{4\sigma_j^2-\eta^2\sigma_j^4}}{4\sqrt{\tau_k}}\right)^2} \quad \text{by (31) and (35)} \\
&= \sqrt{1-\frac{\eta^2\sigma_j^2}{2\sqrt{\tau_k}}+\frac{\eta^2\sigma_j^2}{4\tau_k}} \\
&\leq \sqrt{1-\frac{\eta^2\sigma_j^2}{4\sqrt{\tau_k}}} \\
&\leq 1-\frac{\eta^2\sigma_j^2}{2\sqrt{\tau_k}} \\
&\leq 1-\frac{2}{\tau_k^{(1+c)/2}} \quad \text{by (34).}
\end{aligned}
$$

Next, for the type (ii) eigenvalue, we have we have

$$
\begin{aligned}
|\lambda| &\leq \sqrt{\left(1-\frac{\eta\mu_j}{2\tau_k}\right)^2+\left(\frac{\eta\mu_j}{4\tau_k}\right)^2} \quad \text{by (32) and (33)} \\
&= \sqrt{1-\frac{\eta\mu_j}{\tau_k}+\frac{5\eta^2\mu_j^2}{16\tau_k^2}} \\
&\leq \sqrt{1-\frac{\eta\mu_j}{2\tau_k}} \\
&\leq 1-\frac{\eta\mu_j}{4\tau_k} \\
&\leq 1-\frac{2}{\tau_k^{1+c}} \quad \text{by (34).}
\end{aligned}
$$

Finally, for the type (iii) eigenvalue, we have

$$
\begin{aligned}
|\lambda| &\leq 1-\frac{1}{2}\eta\nu_j(1-\eta\nu_j) \quad \text{by (37)} \\
&\leq 1-\frac{2}{\tau_k^{1+c}} \quad \text{by (36).}
\end{aligned}
$$

Combining all the previous arguments, the assertion then follows. □

From now on, let $\tau_k = k^{1/(2+2c)}$. Then, by Lemma C.1, the upper bound of $\|\boldsymbol{A}_k\|$ is $\|\boldsymbol{A}_k\| \leq 1-\frac{2}{\sqrt{k}}$. Our next step is to find a bound of RHS in (30). To accomplish this, we require the following technical Lemma, which will be used later.

Lemma C.2. *For any k and $n \geq k+1$, there exists a constant M_1 such that $\frac{1}{\sqrt{k}}\prod_{j=k}^{n}\left(1-\frac{1}{\sqrt{j}}\right) \leq M_1\frac{1}{\sqrt{n}}$.*

Proof. We will find upper bound of $\prod_{j=i+1}^{n}\left(1-\frac{1}{\sqrt{j}}\right)$ via finding tight upper and lower bound of $\prod_{j=2}^{n}\left(1-\frac{1}{\sqrt{j}}\right)$.

From the fact that logarithm function is an increasing function on $\mathbb{R}^+$, we have

$$\begin{aligned}
&\int_2^{n+1} \log\left(1-\frac{1}{\sqrt{x}}\right)dx \\
&= -\sqrt{n+1} - \frac{1}{2}\log(n+1) + n\log\left(1-\frac{1}{\sqrt{n+1}}\right) + \sqrt{2} + \log 2 + \sinh^{-1}(1) \\
&\geq \sum_{j=2}^{n} \log\left(1-\frac{1}{\sqrt{j}}\right) \\
&\geq \int_1^{n} \log\left(1-\frac{1}{\sqrt{x}}\right)dx \\
&= -\sqrt{n} - \frac{\log(n)}{2} + (n-1)\log\left(1-\frac{1}{\sqrt{n}}\right) + 1.
\end{aligned}$$

Therefore, for sufficiently large n, we have upper bound

$$\begin{aligned}
\prod_{j=2}^{n}\left(1-\frac{1}{\sqrt{j}}\right) &\leq C_1 e^{-\sqrt{n+1}} \cdot \frac{1}{\sqrt{n+1}} \cdot \left(1-\frac{1}{\sqrt{n+1}}\right)^n \\
&\leq C_2 \frac{1}{\sqrt{n+1}} e^{-(\sqrt{n}+\sqrt{n+1})} \\
&\quad \text{(From the fact that } \left(1-\frac{1}{\sqrt{n+1}}\right)^n \times e^{\sqrt{n}} \text{ converges to } 1/\sqrt{e}) \\
&\leq C_2 \frac{1}{\sqrt{n}} e^{-(\sqrt{n-1}+\sqrt{n})} \qquad (38)
\end{aligned}$$

for some positive constant C_1 and C_2. Simiarly, we have lower bound

$$\begin{aligned}
\prod_{j=2}^{n}\left(1-\frac{1}{\sqrt{j}}\right) &\geq C_3 e^{-\sqrt{n}} \cdot \frac{1}{\sqrt{n}} \cdot \left(1-\frac{1}{\sqrt{n}}\right)^n \\
&\geq C_4 \frac{1}{\sqrt{n}} e^{-(\sqrt{n-1}+\sqrt{n})} \\
&\quad \text{(From the fact that } \left(1-\frac{1}{\sqrt{n}}\right)^n \times e^{\sqrt{n-1}} \text{ converges to } 1/\sqrt{e}) \\
&\geq C_4 \frac{1}{\sqrt{n+1}} e^{-2\sqrt{n+1}} \qquad (39)
\end{aligned}$$

for some positive constant C_3 and C_4. Using (38) and (39), we can bound the product $\frac{1}{\sqrt{k}}\prod_{i=k}^{n}\left(1-\frac{1}{\sqrt{i}}\right)$ as follows.

$$\begin{aligned}
\frac{1}{\sqrt{k}}\prod_{i=k}^{n}\left(1-\frac{1}{\sqrt{i}}\right) &\leq \frac{1}{\sqrt{k}} \cdot \frac{\frac{C_2}{\sqrt{n}} e^{-(\sqrt{n-1}+\sqrt{n})}}{\frac{C_4}{\sqrt{k}} e^{-2\sqrt{k}}} \\
&\leq \frac{1}{\sqrt{n}} \cdot \frac{C_2 e^{-(\sqrt{n-1}+\sqrt{n})}}{C_4 e^{-2\sqrt{k}}}
\end{aligned}$$

Here, for any $n \geq k+1$, the $e^{-(\sqrt{n-1}+\sqrt{n}-2\sqrt{k})} \leq 1$. Therefore, $\frac{1}{\sqrt{k}}\prod_{i=k}^{n}\left(1-\frac{1}{\sqrt{i}}\right) \leq M_1 \frac{1}{\sqrt{n}}$ for $M_1 = \frac{C_2}{C_4}$, and this completes the proof. $\square$

By utilizing the previous Lemma C.2, we can bound the RHS in (30) via mathematical induction. The precise statement is as follows.

Proposition C.3. *Let $f \in C^3$ and $(\tau_k, \gamma_k) = (k^{1/(2+2c)}, k^{1/(4+4c)})$ for $c > 0$. Let $\boldsymbol{z}^*$ be an equilibrium point that satisfies the necessary condition of local minimax points. Then, under Assumptions 1 and 3′, $\boldsymbol{z}^*$ is asymptotically stable point of Alt2-EG-ITS.*

Proof. Under the assumption $f \in C^3$, the Lagrange's form of the remainder (Apostol, 1991, §7.7) is $o(\boldsymbol{z} - \boldsymbol{z}^*) = \frac{\boldsymbol{w}_k^{(2)}(\xi)}{2!}(\boldsymbol{z} - \boldsymbol{z}^*)^2$ for some ξ lies in the closed interval between $\boldsymbol{z}$ and $\boldsymbol{z}^*$. Moreover, $\boldsymbol{w}_k^{(2)}/2$ can be bounded by some positive constant M_2 on small neighborhood $B_{\delta_1}(\boldsymbol{z}^*)$. Let $M \coloneqq \max(1, M_1)$. Then, we will prove that, if for some k-th step, the iterate lies in $B_{\delta_k}(\boldsymbol{z}^*)$ where $\delta_k \coloneqq \min\left(\delta_1, \frac{1}{M_2 \max(1, M_1)\sqrt{k}}\right)$, the future step $\boldsymbol{z}_n$ converge to $\boldsymbol{z}^*$ as n goes to infinity.

Under aforementioned settings, suppose that z_k lies in $B_{\delta_k}(\boldsymbol{z}^*)$ for some k. Then

$$
\begin{aligned}
\|o(\boldsymbol{z}_k - \boldsymbol{z}^*)\| &\le M_2 \|\boldsymbol{z}_k - \boldsymbol{z}^*\|^2 \\
&\le M_2 \delta_k \|\boldsymbol{z}_k - \boldsymbol{z}^*\| \\
&\le \frac{1}{\max(1, M_1)\sqrt{k}} \|\boldsymbol{z}_k - \boldsymbol{z}^*\| \\
&= \frac{1}{M\sqrt{k}} \|\boldsymbol{z}_k - \boldsymbol{z}^*\| .
\end{aligned}
$$

Hence, the first iterate satisfies the following iterates

$$
\begin{aligned}
\|\boldsymbol{x}_{k+1}\| &= \|\boldsymbol{A}_k \boldsymbol{x}_k + o(\boldsymbol{x}_k)\| \\
&\le \|\boldsymbol{A}_k\| \|\boldsymbol{x}_k\| + \|o(\boldsymbol{x}_k)\| \\
&\le \left(1 - \frac{2}{\sqrt{k}}\right) \|\boldsymbol{x}_k\| + \frac{1}{M\sqrt{k}} \|\boldsymbol{x}_k\| \quad \text{(By Lemma C.1)} \\
&\le \left(1 - \frac{1}{\sqrt{k}}\right) \|\boldsymbol{x}_k\| .
\end{aligned}
$$

We use induction on n to prove that $\|\boldsymbol{x}_n\| \le \prod_{j=k}^{n-1}\left(1 - \frac{1}{\sqrt{j}}\right)\|\boldsymbol{x}_k\|$. The $n = k$ case is trivial. Suppose that $\|\boldsymbol{x}_n\| \le \prod_{j=k}^{n-1}\left(1 - \frac{1}{\sqrt{j}}\right)\|\boldsymbol{x}_k\|$ for some n such that $n \ge k+1$. Then, we have

$$
\begin{aligned}
\|\boldsymbol{x}_{n+1}\| &\le \|\boldsymbol{A}_n\| \|\boldsymbol{x}_n\| + \|o(\boldsymbol{x}_n)\| \\
&\le \left(1 - \frac{2}{\sqrt{n}}\right) \prod_{j=k}^{n-1}\left(1 - \frac{1}{\sqrt{j}}\right) \|\boldsymbol{x}_k\| + M_2 \left(\prod_{j=k}^{n-1}\left(1 - \frac{1}{\sqrt{j}}\right)\right)^2 \|\boldsymbol{x}_k\|^2 \quad \text{(By Lemma C.1)} \\
&\le \left(1 - \frac{2}{\sqrt{n}}\right) \prod_{j=k}^{n-1}\left(1 - \frac{1}{\sqrt{j}}\right) \|\boldsymbol{x}_k\| + \left(\prod_{j=k}^{n-1}\left(1 - \frac{1}{\sqrt{j}}\right)\right)^2 \frac{1}{M\sqrt{k}} \|\boldsymbol{x}_k\| \\
&\le \left(1 - \frac{2}{\sqrt{n}}\right) \prod_{j=k}^{n-1}\left(1 - \frac{1}{\sqrt{j}}\right) \|\boldsymbol{x}_k\| + \left(\prod_{j=k}^{n-1}\left(1 - \frac{1}{\sqrt{j}}\right)\right)^2 \frac{1}{M_1\sqrt{k}} \|\boldsymbol{x}_k\|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left(1 - \frac{2}{\sqrt{n}}\right) \prod_{j=k}^{n-1} \left(1 - \frac{1}{\sqrt{j}}\right) \|\boldsymbol{x}_k\| + \frac{1}{\sqrt{n}} \prod_{j=k}^{n-1} \left(1 - \frac{1}{\sqrt{j}}\right) \|\boldsymbol{x}_k\| \quad \text{(By Lemma C.2)} \\
&\leq \prod_{j=k}^{n} \left(1 - \frac{1}{\sqrt{j}}\right) \|\boldsymbol{x}_k\| .
\end{aligned}
$$

Then the assertion follows from the fact that $\lim_{n\to\infty} \prod_{j=k}^{n} \left(1 - \frac{1}{\sqrt{j}}\right) = 0.$ □

We are now ready to prove the asymptotic stability of the Alt2-EG-ITS.

Proof of Theorem 6.1. Suppose that the stationary point $\boldsymbol{z}^*$ satisfies the second-order necessary condition of local minimax points. Then, by Proposition C.3, Alt2-EG-ITS can converge to the $\boldsymbol{z}^*$. Moreover, following a few statements in proof of Proposition C.3, the convergence rate is upper bounded by $\prod_{j=i}^{k} \left(1 - \frac{1}{\sqrt{j}}\right) \|\boldsymbol{x}_i\|$, and the product has (tight) upper bound $O\left(\frac{1}{\sqrt{k}} e^{-2\sqrt{k}}\right)$. These arguments complete the proof. □

Intuitively, as the value of k increases, the neighborhood ensuring local convergence gradually shrinks. However, once it lies inside of the neighborhood, the future iterates converge to the $\boldsymbol{z}^*$.

D. Proof for Section 7

D.1. Proof of Theorem 7.1

We need the following lemma for proving Theorem 7.1.

Lemma D.1. *Suppose Assumptions 1 and 2 hold. Then, there exists a stationary point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ that satisfies*

$$
\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{x}^* \rangle - \langle \nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}, \boldsymbol{y}), \boldsymbol{y} - \boldsymbol{y}^* \rangle \geq f(\boldsymbol{x}, \boldsymbol{y}) - f(\bar{\boldsymbol{x}}, \boldsymbol{y}) \geq \frac{\eta}{\gamma} \left(1 - \frac{L_{\boldsymbol{x}}}{2} \frac{\eta}{\gamma}\right) \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})\|^2, \tag{40}
$$

for all $\boldsymbol{x} \in \mathbb{R}^{d_1}$ and $\boldsymbol{y} \in \mathbb{R}^{d_2}$, where we let $\bar{\boldsymbol{x}} \coloneqq \boldsymbol{x} - \frac{\eta}{\gamma} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})$.

Proof. Assumption 2 implies that the following four inequalities hold, for any $\boldsymbol{x} \in \mathbb{R}^{d_1}$ and $\boldsymbol{y} \in \mathbb{R}^{d_2}$:

$$
\begin{aligned}
&f(\boldsymbol{x}^*, \boldsymbol{y}) \geq f(\boldsymbol{x}, \boldsymbol{y}) + \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{x}^* - \boldsymbol{x} \rangle , \\
&f(\bar{\boldsymbol{x}}, \boldsymbol{y}^*) \geq f(\boldsymbol{x}^*, \boldsymbol{y}^*), \\
&f(\bar{\boldsymbol{x}}, \boldsymbol{y}^*) \leq f(\bar{\boldsymbol{x}}, \boldsymbol{y}) + \langle \nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}, \boldsymbol{y}), \boldsymbol{y}^* - \boldsymbol{y} \rangle , \\
&f(\boldsymbol{x}^*, \boldsymbol{y}) \leq f(\boldsymbol{x}^*, \boldsymbol{y}^*).
\end{aligned}
$$

By summing over the above four inequalities, we have the first inequality of (40). Moreover, the second inequality of (40) can be shown as

$$
\begin{aligned}
f(\bar{\boldsymbol{x}}, \boldsymbol{y}) &\leq f(\boldsymbol{x}, \boldsymbol{y}) + \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), \bar{\boldsymbol{x}} - \boldsymbol{x} \rangle + \frac{L_{\boldsymbol{x}}}{2} \|\bar{\boldsymbol{x}} - \boldsymbol{x}\|^2 \\
&= f(\boldsymbol{x}, \boldsymbol{y}) + \left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), -\frac{\eta}{\gamma} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \right\rangle + \frac{L_{\boldsymbol{x}}}{2} \left\| \frac{\eta}{\gamma} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) \right\|^2 \\
&= f(\boldsymbol{x}, \boldsymbol{y}) - \frac{\eta}{\gamma} \left(1 - \frac{L_{\boldsymbol{x}}}{2} \frac{\eta}{\gamma}\right) \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y})\|^2 .,
\end{aligned}
$$

where the first inequality uses Assumption 1 and (Nesterov, 2018, Theorem 2.1.5). □

We are now ready to show that the Alt2-EG-TS

$$\begin{cases} \boldsymbol{u}_k = \boldsymbol{x}_k - \dfrac{\eta}{\tau_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k), \\ \boldsymbol{v}_k = \boldsymbol{y}_k + \eta\nabla_{\boldsymbol{y}} f\left(\boldsymbol{x}_k - \dfrac{\eta}{\gamma_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k), \boldsymbol{y}_k\right), \end{cases} \qquad \begin{cases} \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \dfrac{\eta}{\tau_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k), \\ \boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \eta\nabla_{\boldsymbol{y}} f\left(\boldsymbol{u}_k - \dfrac{\eta}{\gamma_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k), \boldsymbol{v}_k\right). \end{cases}$$

finds a stationary point under Assumptions 1 and 2, *i.e.,* the smoothness and star-convex-star-concave assumptions on f.

Proof of Theorem 7.1. To prove Theorem 7.1, we begin with the following observation. For simplicity, let us denote $\bar{\boldsymbol{x}}_k \coloneqq \boldsymbol{x}_k - \frac{\eta}{\gamma_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)$ and $\bar{\boldsymbol{u}}_k \coloneqq \boldsymbol{u}_k - \frac{\eta}{\gamma_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k)$. Then, we have the inequality

$$\begin{aligned}
&\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_k}\|\boldsymbol{y}_{k+1} - \boldsymbol{y}^*\|^2 \\
&= \left\|\boldsymbol{x}_k - \frac{\eta}{\tau_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k) - \boldsymbol{x}^*\right\|^2 + \frac{1}{\tau_k}\|\boldsymbol{y}_k + \eta\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{u}}_k, \boldsymbol{v}_k) - \boldsymbol{y}^*\|^2 \\
&= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_k}\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 - \frac{2\eta}{\tau_k}(\langle\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle - \langle\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{u}}_k, \boldsymbol{v}_k), \boldsymbol{y}_k - \boldsymbol{y}^*\rangle) \\
&\quad + \frac{\eta^2}{\tau_k^2}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k)\|^2 + \frac{\eta^2}{\tau_k}\|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{u}}_k, \boldsymbol{v}_k)\|^2 \\
&\le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_k}\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 - \frac{2\eta}{\tau_k}(\langle\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k), \boldsymbol{x}_k - \boldsymbol{u}_k\rangle - \langle\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{u}}_k, \boldsymbol{v}_k), \boldsymbol{y}_k - \boldsymbol{v}_k\rangle) \\
&\quad - \frac{\eta}{\gamma_k}\left(1 - \frac{\eta L_{\boldsymbol{x}}}{2\gamma_k}\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k)\|^2 + \frac{\eta^2}{\tau_k^2}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k)\|^2 + \frac{\eta^2}{\tau_k}\|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{u}}_k, \boldsymbol{v}_k)\|^2 \\
&= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_k}\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 - \frac{2\eta}{\tau_k}\left(\left\langle\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k), \frac{\eta}{\tau_k}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\right\rangle - \langle\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{u}}_k, \boldsymbol{v}_k), -\eta\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_k, \boldsymbol{y}_k)\rangle\right) \\
&\quad - \frac{\eta}{\gamma_k}\left(1 - \frac{\eta L_{\boldsymbol{x}}}{2\gamma_k}\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k)\|^2 + \frac{\eta^2}{\tau_k^2}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k)\|^2 + \frac{\eta^2}{\tau_k}\|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{u}}_k, \boldsymbol{v}_k)\|^2 \\
&= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_k}\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 + \frac{\eta^2}{\tau_k^2}(\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2 - \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2) \\
&\quad + \frac{\eta^2}{\tau_k}(\|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{u}}_k, \boldsymbol{v}_k) - \nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_k, \boldsymbol{y}_k)\|^2 - \|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_k, \boldsymbol{y}_k)\|^2) - \frac{\eta}{\gamma_k}\left(1 - \frac{\eta L_{\boldsymbol{x}}}{2\gamma_k}\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k)\|^2 \\
&\le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_k}\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 + \frac{\eta^2}{\tau_k}(\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2 + \|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{u}}_k, \boldsymbol{v}_k) - \nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_k, \boldsymbol{y}_k)\|^2) \\
&\quad - \frac{\eta^2}{\tau_k^2}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2 - \frac{\eta^2}{\tau_k}\|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_k, \boldsymbol{y}_k)\|^2 - \frac{\eta}{\gamma_k}\left(1 - \frac{\eta L_{\boldsymbol{x}}}{2\gamma_k}\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k)\|^2 \\
&\le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_k}\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 \\
&\quad + \frac{\eta^2}{\tau_k}\left(L_{\boldsymbol{x}}^2 + L_{\boldsymbol{y}}^2 + L_{\boldsymbol{x}}^2 L_{\boldsymbol{y}}^2\frac{\eta^2}{\gamma_k}\left(1 + \frac{1}{\gamma_k}\right)\right)\|(\boldsymbol{u}_k, \boldsymbol{v}_k) - (\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2 + \frac{\eta^2}{\tau_k}\frac{L_{\boldsymbol{y}}^2}{\gamma_k}\|\boldsymbol{u}_k - \boldsymbol{x}_k\|^2 \\
&\quad - \frac{\eta^2}{\tau_k^2}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2 - \frac{\eta^2}{\tau_k}\|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_k, \boldsymbol{y}_k)\|^2 - \frac{\eta}{\gamma_k}\left(1 - \frac{\eta L_{\boldsymbol{x}}}{2\gamma_k}\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k)\|^2 \\
&\le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_k}\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2 \\
&\quad - \frac{\eta^2}{\tau_k^2}\left(1 - \frac{\eta^2 L^2}{\tau_k}(1 + \frac{1}{2\gamma_k}(1 + \frac{1}{\gamma_k})) - \frac{\eta^2 L^2}{\tau_k\gamma_k}\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2 \\
&\quad - \frac{\eta^2}{\tau_k}\left(1 - \eta^2 L^2(1 + \frac{1}{2\gamma_k}(1 + \frac{1}{\gamma_k}))\right)\|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_k, \boldsymbol{y}_k)\|^2 - \frac{\eta}{\gamma_k}\left(1 - \frac{\eta L}{2\gamma_k}\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{u}_k, \boldsymbol{v}_k)\|^2 \\
&\le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_k}\|\boldsymbol{y}_k - \boldsymbol{y}^*\|^2
\end{aligned}$$

$$- \frac{\eta^2}{\tau_k^2}\left(1 - \frac{\eta^2 L^2}{\tau_k}(1 + \frac{1}{2\gamma_k}(3 + \frac{1}{\gamma_k}))\right) \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|^2 - \frac{\eta^2}{\tau_k}\left(1 - \eta^2 L^2(1 + \frac{1}{2\gamma_k}(1 + \frac{1}{\gamma_k}))\right) \|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_k, \boldsymbol{y}_k)\|^2$$

where the first inequality uses Lemma D.1, and the second inequality uses $\tau_k \geq 1$, and the fourth inequality uses the update rules and Lemma B.1. These inequalities lead to the following lemma which is essential in proving the convergence of $\|\boldsymbol{F}_{\mathrm{alt},\gamma_k}(\boldsymbol{x}_k, \boldsymbol{y}_k)\|$.

Lemma D.2. *Let $\tau_i \geq 1$. Then, the series*

$$\sum \frac{\eta^2}{\tau_i^2}\left(1 - \frac{\eta^2 L^2}{\tau_i}\left(1 + \frac{1}{2\gamma_i}\left(3 + \frac{1}{\gamma_i}\right)\right)\right) \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_i, \boldsymbol{y}_i)\|^2 + \sum \frac{\eta^2}{\tau_i}\left(1 - \eta^2 L^2(1 + \frac{1}{2\gamma_i}(1 + \frac{1}{\gamma_i}))\right) \|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_i, \boldsymbol{y}_i)\|^2 \tag{41}$$

is bounded.

Proof. By taking a telescoping summation, then we have

$$\begin{aligned}
&\sum_{i=1}^{k} \frac{\eta^2}{\tau_i^2}\left(1 - \frac{\eta^2 L^2}{\tau_i}(1 + \frac{1}{2\gamma_i}(3 + \frac{1}{\gamma_i}))\right) \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_i, \boldsymbol{y}_i)\|^2 + \sum_{i=1}^{k} \frac{\eta^2}{\tau_i}\left(1 - \eta^2 L^2(1 + \frac{1}{2\gamma_i}(1 + \frac{1}{\gamma_i}))\right) \|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_i, \boldsymbol{y}_i)\|^2 \\
&\leq \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_1}\|\boldsymbol{y}_1 - \boldsymbol{y}^*\|^2 - \|\boldsymbol{x}_{i+1} - \boldsymbol{x}^*\|^2 - \frac{1}{\tau_{i+1}}\|\boldsymbol{y}_{i+1} - \boldsymbol{y}^*\|^2 \\
&\leq \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|^2 + \frac{1}{\tau_1}\|\boldsymbol{y}_1 - \boldsymbol{y}^*\|^2.
\end{aligned}$$

Therefore, the series is bounded. □

We will then show that the series increases monotonically as k increases. Then combining monotonicity and boundedness, one can deduce that both summands of (41) converge to zero as $i \to \infty$.

Lemma D.3. *Let $\tau_i \geq 1$ and $\gamma_i \geq 1$. Then, the series*

$$\sum_{i=j}^{k} \frac{\eta^2}{\tau_i^2}\left(1 - \frac{\eta^2 L^2}{\tau_i}(1 + \frac{1}{2\gamma_i}(3 + \frac{1}{\gamma_i}))\right) \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_i, \boldsymbol{y}_i)\|^2 + \sum_{i=j}^{k} \frac{\eta^2}{\tau_i}\left(1 - \eta^2 L^2(1 + \frac{1}{2\gamma_i}(1 + \frac{1}{\gamma_i}))\right) \|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_i, \boldsymbol{y}_i)\|^2 \tag{42}$$

for some fixed j is monotonically increasing.

Proof. To verify the monotonicity of the series, we need to check the positivity of the summands. The coefficient of the norms are positive, when $\gamma_i > \frac{2\eta^2 L^2}{(1-\eta^2 L^2)}$ satisfied, because

$$\begin{aligned}
&\gamma_i > \frac{2\eta^2 L^2}{(1 - \eta^2 L^2)} \\
&\Rightarrow \gamma_i > \frac{2\eta^2 L^2}{(\tau_i - \eta^2 L^2)} \quad \text{(Since } \tau_i \geq 1\text{)} \\
&\Rightarrow \frac{\tau_i - \eta^2 L^2}{\eta^2 L^2} > \frac{1}{2\gamma_i}(3 + \frac{1}{\gamma_i}) \\
&\Rightarrow \frac{1}{\eta^2 L^2} > \frac{1}{\tau_i}(1 + \frac{1}{2\gamma_i}(3 + \frac{1}{\gamma_i})) \\
&\Rightarrow 1 - \frac{\eta^2 L^2}{\tau_i}(1 + \frac{1}{2\gamma_i}(3 + \frac{1}{\gamma_i})) > 0
\end{aligned}$$

and

$$\begin{aligned}
&\gamma_i > \frac{\eta^2 L^2}{(1-\eta^2 L^2)} \\
&\Rightarrow \frac{1-\eta^2 L^2}{\eta^2 L^2} > \frac{1}{2\gamma_i}(1+\frac{1}{\gamma_i}) \\
&\Rightarrow 1-\eta^2 L^2(1+\frac{1}{2\gamma_i}(1+\frac{1}{\gamma_i})) > 0.
\end{aligned}$$

For fixed (τ, γ), since the $0 < \eta < \frac{\sqrt{\gamma}}{\sqrt{2+\gamma}L}$ implies the condition $\gamma > \frac{2\eta^2 L^2}{(1-\eta^2 L^2)}$, the both summands are positive. For increasing (τ_i, γ_i), since the γ_i increases without any upper bound, for sufficiently large (fixed) i, the condition $\gamma_i > \frac{2\eta^2 L^2}{(1-\eta^2 L^2)}$ holds for any $0 < \eta < \frac{1}{L}$. Therefore, the summation is monotically increasing. □

We are now ready to prove the convergence of each norms $\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k)\|$ and $\|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_k, \boldsymbol{y}_k)\|$. For the case of Alt2-EG-FTS, by Lemmas D.2 and D.3, for $\tau \geq 1$, $\gamma \geq 1$ and $\eta < \frac{\sqrt{\gamma}}{\sqrt{2+\gamma}L}$, the both summands $\frac{\eta^2}{\tau^2}\left(1-\frac{\eta^2 L^2}{\tau}(1+\frac{1}{2\gamma}(3+\frac{1}{\gamma}))\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_i, \boldsymbol{y}_i)\|^2$ and $\frac{\eta^2}{\tau}\left(1-\eta^2 L^2(1+\frac{1}{2\gamma}(1+\frac{1}{\gamma}))\right)\|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_i, \boldsymbol{y}_i)\|^2$ converge to zero as $i \to \infty$. Then the assertion follows from the fact that the coefficients of the both terms are invariant as i varies.

For the case of Alt2-EG-ITS, since the conditions $\tau_i \geq 1$, $\gamma > \frac{2\eta^2 L^2}{1-\eta^2 L^2}$ for $\eta < \frac{1}{L}$ are satisfied as $i \to \infty$, both summands $\frac{\eta^2}{\tau^2}\left(1-\frac{\eta^2 L^2}{\tau}(1+\frac{1}{2\gamma}(3+\frac{1}{\gamma}))\right)\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_i, \boldsymbol{y}_i)\|^2$ and $\frac{\eta^2}{\tau}\left(1-\eta^2 L^2(1+\frac{1}{2\gamma}(1+\frac{1}{\gamma}))\right)\|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_i, \boldsymbol{y}_i)\|^2$ converge to zero as $i \to \infty$ by Lemmas D.2 and D.3. However, the coefficient of the terms also diminishes as $i \to \infty$, therefore, the only thing we can say about the terms is the limit inferior of the both terms converge to zero.

Therefore, the rest of the proof is to demonstrate that, the summation of the coefficients in (42) are infinite. Since, $1-\frac{\eta^2 L^2}{\tau_i}(1+\frac{1}{2\gamma_i}(3+\frac{1}{\gamma_i})) \to 1$ as $i \to \infty$, there exists i_0 such that any $i \geq i_0$ implies the $1-\frac{\eta^2 L^2}{\tau_i}(1+\frac{1}{2\gamma_i}(3+\frac{1}{\gamma_i})) > \frac{1}{2}$. Therefore, we have

$$\begin{aligned}
\sum_{i=i_0}^{\infty} \frac{\eta^2}{\tau_i^2}\left(1-\eta^2 L^2(1+\frac{1}{2\gamma_i}(3+\frac{1}{\gamma_i}))\right) &\geq \sum_{i=i_0}^{\infty} \frac{\eta^2}{2\tau_i^2} \\
&= \frac{\eta^2}{2}\sum_{i=i_0}^{\infty} \frac{1}{\tau_i^2} \\
&= \infty
\end{aligned}$$

Therefore, $\liminf_{i\to\infty} \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_i, \boldsymbol{y}_i)\| = 0$ holds.

Similarly, we have

$$\begin{aligned}
\sum_{i=i_1}^{\infty} \frac{\eta^2}{\tau_i}\left(1-\eta^2 L^2(1+\frac{1}{2\gamma_i}(1+\frac{1}{\gamma_i}))\right) &\geq \sum_{i=i_1}^{\infty} \frac{\eta^2}{2\tau_i} \\
&= \frac{\eta^2}{2}\sum_{i=i_0}^{\infty} \frac{1}{\tau_i} \\
&= \infty
\end{aligned}$$

Therefore, $\liminf_{i\to\infty} \|\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{x}}_i, \boldsymbol{y}_i)\| = 0$ holds, and these arguments complete the proof. □

D.2. Accumulation Point of the Alt2-EG-FTS is a Stationary Point

Proof. Let $\tilde{z}$ be an accumulation point of the sequence $\{z_k\}_{k\geq 0}$. Then, there exists a subsequence $\{z_{k_j}\}_{j\geq 0}$ of original sequence such that $z_{k_j} \to \tilde{z}$ as $j \to \infty$ (Tao, 2016, Proposition 1.4.5). By triangle inequality and Lemma B.1, we have

$$\begin{aligned} 0 &\leq \|F_{\text{alt},\gamma}(\tilde{z})\| \\ &\leq \left\|F_{\text{alt},\gamma}(\tilde{z}) - F_{\text{alt},\gamma}(z_{k_j})\right\| + \left\|F_{\text{alt},\gamma}(z_{k_j})\right\| \\ &\leq \sqrt{L_x^2 + L_y^2\left(1+\frac{1}{\gamma}\right)\left(1+L_x^2\frac{\eta^2}{\gamma}\right)}\left\|\tilde{z} - z_{k_j}\right\| + \left\|F_{\text{alt},\gamma}(z_{k_j})\right\|. \end{aligned}$$

From the fact that both $\left\|\tilde{z} - z_{k_j}\right\|$ and $\left\|F_{\text{alt},\gamma}(z_{k_j})\right\|$ converge to zero as $j \to \infty$, we have $\|F_{\text{alt},\gamma}(\tilde{z})\| = 0$, and this completes the proof. □

E. Proofs for Section 8

E.1. Proof of Example 1

Proof. Consider the function $f(x,y) = -x^2 + 2xy$. Its saddle-gradient is $F(x,y) = (-2x+2y, -2x)$, and it has a unique stationary point $(0,0)$. Moreover, f satisfies Assumption 1 with $L_x = L_y = 2$, so $L := \sqrt{L_x^2 + L_y^2} = 2\sqrt{2}$.

For any $\delta > 0$ and any (x,y) satisfying $|x-0| \leq \delta$ and $|y-0| \leq \delta$, the inequality

$$f(0,y) = 0 \leq f(0,0) = 0 \leq \max_{y' \,:\, |y'-0|\leq\delta} f(x,y') = -x^2 + 2|x\delta|$$

holds, so the stationary point $(0,0)$ is a local minimax point. In addition, since $\nabla_{yy} f(x,y)$ is degenerate, it is a non-strict local minimax point.

Since

$$DF = \begin{bmatrix} A & C \\ -C^\top & -B \end{bmatrix} = \begin{bmatrix} -2 & 2 \\ -2 & 0 \end{bmatrix},$$

we have $C_2 = [2] \in \mathbb{R}^{1\times 1}$ and $q = \operatorname{rank}(C_2) = 1$. Then, the matrix U is of size 0×0, and so is $S_{\text{res}}(DF) = U^\top(A - CB^\dagger C^\top)U$, which is vacuously positive definite. Therefore, Assumption 3′ is satisfied, and thus, by Theorems 5.4 and 6.1, both Alt2-EG-FTS and Alt2-EG-ITS are asymptotically stable at $(0,0)$, respectively.

Let us now show that the (vanilla) two-timescale EG is unstable at $(0,0)$ for any choice of timescale separation τ. In particular, we show that the eigenvalues of its Jacobian

$$H_\tau = \Lambda_\tau DF = \begin{bmatrix} \frac{1}{\tau} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -2 & 2 \\ -2 & 0 \end{bmatrix} = \begin{bmatrix} -2\epsilon & 2\epsilon \\ -2 & 0 \end{bmatrix},$$

which are $-\epsilon \pm i\sqrt{4\epsilon - \epsilon^2}$, are outside $\mathcal{P}_\eta$ for any choice of τ, based on Propositions 3.1 and 4.1. By the definition of $\mathcal{P}_\eta$, it is enough to show that all $\epsilon > 0$ satisfy

$$\begin{aligned} &(\eta\epsilon + \frac{1}{2})^2 + \eta^2(4\epsilon - \epsilon^2) + \frac{3}{4} > \sqrt{1 + 3\eta^2(4\epsilon - \epsilon^2)} \\ &\Leftrightarrow 1 + \eta\epsilon + 4\eta^2\epsilon > \sqrt{1 + 3\eta^2(4\epsilon - \epsilon^2)} \\ &\Leftrightarrow 1 + \eta^2\epsilon^2 + 16\eta^4\epsilon^2 + 2\eta\epsilon + 8\eta^2\epsilon + 8\eta^3\epsilon^2 > 1 + 12\eta^2\epsilon - 3\eta^2\epsilon^2 \\ &\Leftrightarrow 4\eta^2\epsilon^2 + 16\eta^4\epsilon^2 + 2\eta\epsilon - 4\eta^2\epsilon + 8\eta^3\epsilon^2 > 0 \\ &\Leftrightarrow 4\eta^2\epsilon + 16\eta^4\epsilon + 2\eta - 4\eta^2 + 8\eta^3\epsilon > 0 \\ &\Leftrightarrow \epsilon(4\eta^2 + 16\eta^4 + 8\eta^3) > 4\eta^2 - 2\eta \end{aligned}$$

where the fifth equivalent comes from the $\epsilon > 0$. Since $4\eta^2 - 2\eta \leq 0$ for any $0 < \eta < \frac{1}{L} = \frac{1}{2\sqrt{2}}$, the above inequality indeed holds for any $\epsilon > 0$, which completes the proof.

□