Open Problem: Is There a First-Order Method that Only Converges to Local Minimax Optima?

Jiseok Chae JSCH@KAIST.AC.KR
Kyuwon Kim KKW4053@KAIST.AC.KR
Donghwan Kim DONGHWANKIM@KAIST.AC.KR
Korea Advanced Institute of Science and Technology

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Can we effectively train a generative adversarial network (GAN) (or equivalently, optimize a minimax problem), similar to how we successfully learn a classification neural network (or equivalently, minimize a function) by gradient methods? The answer to this question at the moment is “No”. Despite extensive studies over the past ten years, training GANs still remains challenging. As a result, diffusion-based generative models are largely replacing GANs. When training GANs, we not only struggle with finding stationary points, but also, from a practical view, suffer from the so-called mode-collapse phenomenon, generating samples that lack diversity compared to the training data. Due to the nature of GAN, the mode-collapse is likely to occur when we accidentally find an optimal point for the maximin problem, rather than the original minimax problem (Goodfellow, 2016).

This consequently suggests that addressing a long-standing open question of whether there exists a first-order method that only converges to (local) optimum of minimax problems can resolve the aforementioned shortcomings. Apparently, none of the existing methods possess such a property, neither theoretically nor practically. This is in contrast to the fact that a standard gradient descent method successfully finds (local) minima (Lee et al., 2016). Surprisingly, in nonconvex-nonconcave minimax optimization, Jin et al. (2020) are the first to suggest an appropriate notion of local optimality, especially taking account of the order of minimization and maximization. Jin et al. (2020) also presented a partial answer to the above open question, by demonstrating that a *two-timescale* gradient descent ascent only converges to a *strict* local minimax optimum, under a certain condition. However, the convergence to general local minimax optimum was left mostly unexplored, even though such a *non-strict* local minimax optimum is prevalent in practice. Our recent findings in (Chae et al., 2023) illustrate that it is indeed possible to find some *non-strict* local minimax optimum by a *two-timescale* variant of *extragradient* method.

This positive result brings new attention to the aforementioned open question. In this paper, we detail it in regard to our recent findings. Furthermore, we are writing to revive discussion on the appropriate notion of local minimax optimum. This was initially discussed by Jin et al. (2020), but not much thereafter, which we believe is an important piece of answering the open question.

1. Introduction

Minimax optimization, $\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$, has become an important part in the machine learning community, especially since the appearance of GANs (Goodfellow et al., 2014). However, optimizing a minimax problem yet remains challenging, and this drawback has been critical for GANs to incrementally lose its role in the field of generative models by its strong contender, diffusion-based generative models (Song et al., 2021; Dhariwal and Nichol, 2021). In particular, training GANs is known to require extensive handcrafted tuning of hyperparameters and regularizers. Most of the time, it fails to find stationary points, and more importantly, suffers from the so-called mode-collapse phenomenon, not sufficiently capturing the diversity in the training data. Regarding the former, there has been an increased recent interest in developing first-order methods that find stationary points of certain nonconvex-nonconcave problems, see *e.g.*, (Diakonikolas et al., 2021; Pethick et al., 2022). This, however, does not resolve the more important mode-collapse issue, which is known to happen when we accidentally find an optimal (stationary) point for the maximin problem, rather than the original minimax problem (Goodfellow, 2016). This consequently has asked for years whether there exists a first-order method that only converges to a local minimax optimum. However, no existing methods possess such a property, and we are writing to renew the interest on the aforementioned open question in regard to our recent findings in (Chae et al., 2023). The rest of this section reviews the local minimax points defined in (Jin et al., 2020), and Section 2 presents the latest work on this open question in (Jin et al., 2020), new findings in (Chae et al., 2023), and remaining challenges in answering the open question.

1.1. Preliminaries: Local minimax optimum

Considering the sequential nature of nonconvex-nonconcave problems, Jin et al. (2020) proposed the following local version of the Stackelberg equilibrium (von Stackelberg, 2011), rather than the Nash equilibrium for simultaneous games. Evtushenko (1974) originally proposed a concept of local Stackelberg equilibrium, but Jin et al. (2020, Proposition 37) showed that it is not a truly local notion. Fiez et al. (2020) have also recognized the importance of the sequential nature, but their notion, which is exactly equivalent to *strict* local minimax points defined below, is more restrictive as it implicitly assumes that the Hessian for the maximization player is nondegenerate.

Definition 1 (Jin et al. (2020)) *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*

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

This definition especially introduces a function $h(\delta)$ to allow taking the radii of neighborhoods for minimization and maximization differently. Jin et al. (2020, Remark 14) showed that the definition remains equivalent even when we further assume that $h(\delta)$ is monotonic or continuous. Still, whether imposing only such weak restrictions on $h(\delta)$, or even further, introducing $h(\delta)$, is plausible or not, has not been carefully discussed anywhere. Later in this paper we comment on possible limitations of Definition 1 regarding the choice of the neighborhood, and suggest some alternatives.

Local minimax points are (first-order) stationary points, and can be further characterized by the following second-order conditions (Jin et al., 2020).

- (Second-order necessary condition) Any local minimax $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfies $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) \preceq \mathbf{0}$. In addition, if $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) \prec \mathbf{0}$, then $[\nabla^2_{\boldsymbol{xx}} f - \nabla^2_{\boldsymbol{xy}} f (\nabla^2_{\boldsymbol{yy}} f)^{-1} \nabla^2_{\boldsymbol{yx}} f](\boldsymbol{x}^*, \boldsymbol{y}^*) \succeq \mathbf{0}$.
- (Second-order sufficient condition) Any stationary point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfying $[\nabla^2_{\boldsymbol{xx}} f - \nabla^2_{\boldsymbol{xy}} f (\nabla^2_{\boldsymbol{yy}} f)^{-1} \nabla^2_{\boldsymbol{yx}} f](\boldsymbol{x}^*, \boldsymbol{y}^*) \succ \mathbf{0}$ and $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) \prec \mathbf{0}$ is local minimax.

A possible approach to the solution of the open question would be to construct a first-order method that only converges to second-order stationary points, *i.e.,* points satisfying the second-order necessary condition. However, this condition is loose unless $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) \prec \mathbf{0}$, which impeded further studies. A more feasible answer to the open question was to consider *strict* local minimax points, the points satisfying the second-order sufficient condition. To the best of our knowledge, all existing studies on local minimax points have focused on methods finding *strict* local minimax points.

2. Existing works and remaining challenges

2.1. First-order methods in minimax optimization

Why is it so challenging to find first-order methods that converge to local optima in minimax problems? One possible explanation is the lack of interest in analyzing first-order methods in terms of *non-strict* local minimax points. This section focuses what is known so far about first-order methods, from this point of view. For simplicity, let $\boldsymbol{F} := (\nabla_{\boldsymbol{x}} f, -\nabla_{\boldsymbol{y}} f)$ denote the saddle gradient of f, and $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{y})$. We assume that $f \in C^2$, and there exists $L > 0$ such that $||D\boldsymbol{F}|| \leq L$.

2.1.1. First-order method that only converges to strict local minimax points

In view of the sequential nature of minimax problems, the *two-timescale* gradient descent ascent (GDA) method, $\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta \boldsymbol{\Lambda}_\tau \boldsymbol{F}(\boldsymbol{z}_k)$ with $\boldsymbol{\Lambda}_\tau := \mathrm{diag}\{{}^1/_\tau \boldsymbol{I}, \boldsymbol{I}\}$, which puts more emphasis on the maximization when $\tau \geq 1$, has been studied and widely used in practice (Heusel et al., 2017). In the theoretical perspective, Jin et al. (2020) showed that the two-timescale GDA converges to *strict* local minimax points for $\tau \gg 1$, and Fiez and Ratliff (2021) established exactly how large should τ be in order to ensure such convergence.

2.1.2. First-order methods and nonstrict local minimax points

Despite the success of timescale separation on GDA, not much further progress has been made until recently, and thus nothing was known about first-order methods in finding non-strict local minimax points. One major hurdle in studying non-strict local minimax points was that a tight yet simple second-order characterization of a local minimax point was not known.

By restricting the choice of $h(\delta)$ in Definition 1, Chae et al. (2023) proposed a refined second-order necessary condition, which is simply written in terms of the restricted Schur complement[1] $\boldsymbol{S}_{\mathrm{res}}$ of $D\boldsymbol{F}$. This yielded a simple[2] notion of second-order stationary points of local minimax points, and lead us to come up with a method that only converges to second-order stationary points under some mild condition.

Proposition 2 (Refined second-order necessary condition (Chae et al., 2023)) *Let $f \in C^2$, then any local minimax point $\boldsymbol{z}^*$ satisfies $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}^*) \preceq \boldsymbol{0}$. In addition, if the function $h(\delta)$ in Definition 1 satisfies* $\operatorname{limsup}_{\delta \to 0+} {}^{h(\delta)}/_{\delta} < \infty$, *then* $\boldsymbol{S}_{\mathrm{res}}(D\boldsymbol{F}(\boldsymbol{z}^*)) \succeq \boldsymbol{0}$.

Chae et al. (2023) also showed that the two-timescale extragradient (EG) method, $\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta \boldsymbol{\Lambda}_\tau \boldsymbol{F}(\boldsymbol{z}_k - \eta \boldsymbol{\Lambda}_\tau \boldsymbol{F}(\boldsymbol{z}_k))$, with sufficiently large τ converges to the second-order stationary points of local minimax points under some mild condition. This supersedes the two-timescale GDA, but is not powerful enough to completely solve the open problem, which we detail below. The

[1] The precise definition of the restricted Schur complement $\boldsymbol{S}_{\mathrm{res}}(D\boldsymbol{F})$ can be found in (Chae et al., 2023), which reduces to a standard Schur complement $\nabla^2_{\boldsymbol{xx}} f - \nabla^2_{\boldsymbol{xy}} f (\nabla^2_{\boldsymbol{yy}} f)^{-1} \nabla^2_{\boldsymbol{yx}} f$ when $\nabla^2_{\boldsymbol{yy}} f$ is nondegenerate.

[2] (Zhang et al., 2022) also refined the second-order necessary condition for the case when $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}^*)$ is nonsingular, but it is more complicated than ours.

following is the first convergence result to second-order stationary points. The value s_0 represents a certain property of the local minimax points, whose precise definition is given in the cited work.

Theorem 3 (Informal version of Theorem 6.8 (Chae et al., 2023)) *A stationary point* $\boldsymbol{z}^*$ *satisfies* $\boldsymbol{S}_{\mathrm{res}}(D\boldsymbol{F}(\boldsymbol{z}^*)) \succeq \mathbf{0}$*,* $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}^*) \preceq \mathbf{0}$*, and* $s_0(\boldsymbol{z}^*) < {}^{1}\!/_{2L}$ *if and only if there exists* $0 < \eta^\star < {}^{1}\!/_{L}$ *such that two-timescale EG with sufficiently large* τ *locally converges to* $\boldsymbol{z}^*$ *for any* $\eta \in (\eta^\star, {}^{1}\!/_{L})$*.*

This is an important finding, but it rules out the points with $s_0 \geq {}^{1}\!/_{2L}$, and depends on the choice of η. We thus considered removing both, but then we ended up identifying a gap between sufficient and necessary conditions. For further details, see (Chae et al., 2023). Although the two-timescale EG showed potential for closing the open problem, our attempts were not enough to remove the gap between its limit points and local minimax optima. We thus invite everyone to resolve this last remaining piece, possibly coming up with a new method or a new notion of local minimax optimum.

2.2. Local optimality in minimax problems

It is also true that not much discussion has been made on the "correct" notion of local optimality for minimax problems. In this section, we make further discussions on the definition given by Jin et al. (2020), and appeal the necessity of such a discussion by suggesting an alternative which could still be a plausible local optimum, while being more suitable for first-order methods.

2.2.1. Another candidate definition of local minimax optimum

The choice of the neighborhood which maximization is taken over has not yet reached an agreement. So, one may consider choosing, for example, $\{\boldsymbol{y}' \in \boldsymbol{y}^* + \mathcal{R}(\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*)) \;:\; \|\boldsymbol{y}' - \boldsymbol{y}^*\| \leq h(\delta)\}$ rather than $\{\boldsymbol{y}' \;:\; \|\boldsymbol{y}' - \boldsymbol{y}^*\| \leq h(\delta)\}$ in the original definition. We call the resulting notion of local optimum the local *restricted* minimax point, as we restricted the neighborhood in the range of the Hessian of the maximization player. This new notion is comparable to the original local minimax point in several aspects. First of all, the second-order necessary condition reduces to using a generalized Schur complement $\boldsymbol{S}(D\boldsymbol{F}) := \nabla^2_{\boldsymbol{xx}} f - \nabla^2_{\boldsymbol{xy}} f (\nabla^2_{\boldsymbol{yy}} f)^\dagger \nabla^2_{\boldsymbol{yx}} f$, without any restriction on $h(\delta)$ as in Definition 1 (see Appendix A). Also, other necessary and sufficient conditions remain equivalent to those of local minimax points (see Appendix B).

2.2.2. What is really an appropriate notion of local minimax optimum?

There also is an arguable local minimax point, which we would like to share for discussion. Although a global[3] minimax point (global Stackelberg equilibrium) might not be locally optimal (Jin et al., 2020, Proposition 21), we believe it is reasonable to expect a *quadratic* function without a global minimax point to not have any local minimax point. However, this fails for the following *concave-linear quadratic* function having a (possibly undesirable) *non-strict* local minimax point (with $h(\delta)$ satisfying $\limsup_{\delta \to 0+} {}^{h(\delta)}\!/_{\delta} < \infty$).

Example 1 *A function* $f(x, y) = -\frac{a}{2}x^2 + cxy$ *with constants* $a > 0$ *and* $c \neq 0$ *does not have a global minimax point in general, but has a local minimax point* $(0, 0)$ *for any* $a < 0$*. On the other hand, the unique stationary point* $(0, 0)$ *is not a local restricted minimax point. (See Appendix C.)*

This again questions the validity of Definition 1, and we encourage active discussion on the proper notion of local minimax optimality, possibly considering the practical mode-collapse issue.

[3] $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a global minimax point, if for any $(\boldsymbol{x}, \boldsymbol{y})$ in $\mathcal{X} \times \mathcal{Y}$ we have $f(\boldsymbol{x}^*, \boldsymbol{y}) \leq f(\boldsymbol{x}^*, \boldsymbol{y}^*) \leq \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}')$.

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).

References

J. Chae, K. Kim, and D. Kim. Two-timescale extragradient for finding local minimax points, 2023. arxiv 2305.16242.

P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *Neural Info. Proc. Sys.*, 2021.

J. Diakonikolas, C. Daskalakis, and M. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

Y. G. Evtushenko. Some local properties of minimax problems. *USSR Comp. Math. and Math. Phys.*, 14(3):129–138, 1974.

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

T. Fiez, B. Chasnov, and L. Ratliff. Implicit learning dynamics in Stackelberg games: equilibria characterization, convergence analysis, and empirical study. In *Proc. Intl. Conf. Mach. Learn*, 2020.

I. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks, 2016. arxiv 1701.00160.

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

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

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

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

T. M. Pethick, P. Latafat, P. Patrinos, O. Fercoq, and V. Cevher. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *Proc. Intl. Conf. on Learning Representations*, 2022.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *Proc. Intl. Conf. on Learning Representations*, 2021.

H. von Stackelberg. *Market structure and equilibrium.* Springer-Verlag Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-12586-7.

G. Zhang, P. Poupart, and Y. Yu. Optimality and stability in non-convex smooth games. *J. Mach. Learn. Res.*, 23:35–1, 2022.

Appendix A. Second-order necessary condition of local restricted minimax point

Proposition 4 (Second-order necessary condition) *Any local restricted minimax point $\boldsymbol{z}^*$ satisfies $\boldsymbol{S}(D\boldsymbol{F}(\boldsymbol{z}^*)) \succeq \mathbf{0}$ and $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{z}^*) \preceq \mathbf{0}$.*

Proof Let $\boldsymbol{A} := \nabla^2_{\boldsymbol{xx}} f(\boldsymbol{x}^*, \boldsymbol{y}^*)$, $\boldsymbol{B} := \nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*)$, and $\boldsymbol{C} := \nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*)$. Since $\boldsymbol{y}^*$ is a local maximum of $f(\boldsymbol{x}^*, \cdot)$ in the range of $\boldsymbol{B}$, we have $\boldsymbol{B} \preceq \mathbf{0}$.

Since $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a local restricted minimax point, there exists a function h such that

$$f(\boldsymbol{x}^*, \boldsymbol{y}) \leq f(\boldsymbol{x}^*, \boldsymbol{y}^*) \leq \max_{\substack{\boldsymbol{y}'-\boldsymbol{y}^* \in \mathcal{R}(\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*)) \\ \|\boldsymbol{y}'-\boldsymbol{y}^*\| \leq h(\delta)}} f(\boldsymbol{x}, \boldsymbol{y}').$$

holds. Denote $\tilde{h}(\delta) = 2\|\boldsymbol{B}^\dagger \boldsymbol{C}^\top\|\delta$. Using the first-order necessary condition (Proposition 5), we have

$$f(\boldsymbol{x}^* + \boldsymbol{\delta_x}, \boldsymbol{y}^* + \boldsymbol{\delta_y}) = f(\boldsymbol{x}^*, \boldsymbol{y}^*) + \frac{1}{2}\boldsymbol{\delta_x}^\top \boldsymbol{A} \boldsymbol{\delta_x} + \boldsymbol{\delta_x}^\top \boldsymbol{C} \boldsymbol{\delta_y} + \frac{1}{2}\boldsymbol{\delta_y}^\top \boldsymbol{B} \boldsymbol{\delta_y} + o(\|\boldsymbol{\delta_x}\|^2 + \|\boldsymbol{\delta_y}\|^2)$$

For any $\boldsymbol{\delta_x}$ and $\boldsymbol{\delta_y} \in \mathcal{R}(\boldsymbol{B})$, we have

$$(\boldsymbol{C}^\top \boldsymbol{\delta_x})^\top \boldsymbol{\delta_y} = (P_{\mathcal{R}(\boldsymbol{B})}[\boldsymbol{C}^\top \boldsymbol{\delta_x}])^\top \boldsymbol{\delta_y} = (\boldsymbol{B}\boldsymbol{B}^\dagger \boldsymbol{C}^\top \boldsymbol{\delta_x})^\top \boldsymbol{\delta_y},$$

where $P_{\mathcal{R}(\boldsymbol{B})}$ is a projection onto $\mathcal{R}(\boldsymbol{B})$. By (local) concavity of $f(\boldsymbol{x}^* + \boldsymbol{\delta_x}, \boldsymbol{y}^* + \cdot)$ within $\mathcal{R}(\boldsymbol{B})$, if some $\boldsymbol{\delta_y} \in \mathcal{R}(\boldsymbol{B})$ satisfies $\nabla_{\boldsymbol{\delta_y}} f(\boldsymbol{x}^* + \boldsymbol{\delta_x}, \boldsymbol{y}^* + \boldsymbol{\delta_y}) = \mathbf{0}$ within $||\boldsymbol{\delta_y}|| \leq \max\{h(\delta), \tilde{h}(\delta)\}$, then it will be a local maximizer of $f(\boldsymbol{x}^* + \boldsymbol{\delta_x}, \boldsymbol{y}^* + \cdot)$ within $\mathcal{R}(\boldsymbol{B})$ and $||\boldsymbol{\delta_y}|| \leq \max\{h(\delta), \tilde{h}(\delta)\}$. As $\boldsymbol{\delta_y}$ is restricted to $\mathcal{R}(\boldsymbol{B})$, it is not hard to verify that

$$\exists\, \boldsymbol{\delta_y} = -\boldsymbol{B}^\dagger \boldsymbol{C}^\top \boldsymbol{\delta_x} + o(\|\boldsymbol{\delta_x}\|) \quad \text{s.t.} \quad \nabla_{\boldsymbol{\delta_y}} f(\boldsymbol{x}^* + \boldsymbol{\delta_x}, \boldsymbol{y}^* + \boldsymbol{\delta_y}) = \mathbf{0}.$$

Since $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a local restricted minimax point, we have

$$\begin{aligned} 0 \leq \max_{\substack{\boldsymbol{\delta_y} \in \mathcal{R}(\boldsymbol{B}): \\ \|\boldsymbol{\delta_y}\| \leq h(\delta)}} f(\boldsymbol{x}^* + \boldsymbol{\delta_x}, \boldsymbol{y}^* + \boldsymbol{\delta_y}) - f(\boldsymbol{x}^*, \boldsymbol{y}^*) &\leq \max_{\substack{\boldsymbol{\delta_y} \in \mathcal{R}(\boldsymbol{B}): \\ \|\boldsymbol{\delta_y}\| \leq \max\{h(\delta), \tilde{h}(\delta)\}}} f(\boldsymbol{x}^* + \boldsymbol{\delta_x}, \boldsymbol{y}^* + \boldsymbol{\delta_y}) - f(\boldsymbol{x}^*, \boldsymbol{y}^*) \\ &= \frac{1}{2}\boldsymbol{\delta_x}^\top (\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^\dagger \boldsymbol{C}^\top)\boldsymbol{\delta_x} + o(\|\boldsymbol{\delta_x}\|^2) \end{aligned}$$

for any $\boldsymbol{\delta_x}$, which concludes the proof. ■

Appendix B. Additional properties of local restricted minimax point

Proposition 5 (First-order necessary condition) *For $f \in C^1$, any local restricted minimax point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfies $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) = 0$ and $\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) = \mathbf{0}$.*

Proof Since $\boldsymbol{y}^*$ is a local maximum of $f(\boldsymbol{x}^*, \cdot)$ we have $\nabla_{\boldsymbol{y}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) = \mathbf{0}$. Let $\boldsymbol{\delta}'_{\boldsymbol{y}}(\boldsymbol{\delta}_{\boldsymbol{x}}) \coloneqq \operatorname{argmax}_{\substack{\boldsymbol{\delta}_{\boldsymbol{y}} \in \mathcal{R}(\boldsymbol{B}) \\ ||\boldsymbol{\delta}_{\boldsymbol{y}}|| \le h(\delta)}} f(\boldsymbol{x}^* + \boldsymbol{\delta}_{\boldsymbol{x}}, \boldsymbol{y}^* + \boldsymbol{\delta}_{\boldsymbol{y}})$. By definition we have $\|\boldsymbol{\delta}'_{\boldsymbol{y}}(\boldsymbol{\delta}_{\boldsymbol{x}})\| \le h(\delta) \to 0$ as $\delta \to 0$. We then have

$$\begin{aligned}
0 &\le f(\boldsymbol{x}^* + \boldsymbol{\delta}_{\boldsymbol{x}}, \boldsymbol{y}^* + \boldsymbol{\delta}'_{\boldsymbol{y}}(\boldsymbol{\delta}_{\boldsymbol{x}})) - f(\boldsymbol{x}^*, \boldsymbol{y}^*) \\
&= f(\boldsymbol{x}^* + \boldsymbol{\delta}_{\boldsymbol{x}}, \boldsymbol{y}^* + \boldsymbol{\delta}'_{\boldsymbol{y}}(\boldsymbol{\delta}_{\boldsymbol{x}})) - f(\boldsymbol{x}^*, \boldsymbol{y}^* + \boldsymbol{\delta}'_{\boldsymbol{y}}(\boldsymbol{\delta}_{\boldsymbol{x}})) + f(\boldsymbol{x}^*, \boldsymbol{y}^* + \boldsymbol{\delta}'_{\boldsymbol{y}}(\boldsymbol{\delta}_{\boldsymbol{x}})) - f(\boldsymbol{x}^*, \boldsymbol{y}^*) \\
&\le f(\boldsymbol{x}^* + \boldsymbol{\delta}_{\boldsymbol{x}}, \boldsymbol{y}^* + \boldsymbol{\delta}'_{\boldsymbol{y}}(\boldsymbol{\delta}_{\boldsymbol{x}})) - f(\boldsymbol{x}^*, \boldsymbol{y}^* + \boldsymbol{\delta}'_{\boldsymbol{y}}(\boldsymbol{\delta}_{\boldsymbol{x}})) \\
&= \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{y}^* + \boldsymbol{\delta}'_{\boldsymbol{y}}(\boldsymbol{\delta}_{\boldsymbol{x}}))^\top \boldsymbol{\delta}_{\boldsymbol{x}} + o(\|\boldsymbol{\delta}_{\boldsymbol{x}}\|) \\
&= \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{y}^*)^\top \boldsymbol{\delta}_{\boldsymbol{x}} + o(\|\boldsymbol{\delta}_{\boldsymbol{x}}\|),
\end{aligned}$$

where the second inequality uses the fact that $\boldsymbol{y}^*$ is a local maximum of $f(\boldsymbol{x}^*, \cdot)$. This inequality holds for any small $\boldsymbol{\delta}_{\boldsymbol{x}}$, which implies $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) = \mathbf{0}$. ■

Proposition 6 (Second-order sufficient condition) *For $f \in C^2$, any stationary point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfying $[\nabla^2_{\boldsymbol{xx}} f - \nabla^2_{\boldsymbol{xy}} f (\nabla^2_{\boldsymbol{yy}} f)^{-1} \nabla^2_{\boldsymbol{yx}} f](\boldsymbol{x}^*, \boldsymbol{y}^*) \succ \mathbf{0}$ and $\nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) \prec \mathbf{0}$ is a local restricted minimax point.*

Proof Let $\boldsymbol{A} \coloneqq \nabla^2_{\boldsymbol{xx}} f(\boldsymbol{x}^*, \boldsymbol{y}^*)$, $\boldsymbol{B} \coloneqq \nabla^2_{\boldsymbol{yy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*)$, and $\boldsymbol{C} \coloneqq \nabla^2_{\boldsymbol{xy}} f(\boldsymbol{x}^*, \boldsymbol{y}^*)$. Since $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a stationary point, and $\boldsymbol{B} \prec \mathbf{0}$, $\boldsymbol{y}^*$ is a local maximum of $f(\boldsymbol{x}^*, \cdot)$. Let $h(\delta) = \|\boldsymbol{B}^{-1}\boldsymbol{C}^\top\|\delta$, and choose $\boldsymbol{\delta}'_{\boldsymbol{y}} = -\boldsymbol{B}^{-1}\boldsymbol{C}^\top \boldsymbol{\delta}_{\boldsymbol{x}}$ that satisfies $\|\boldsymbol{\delta}'_{\boldsymbol{y}}\| \le h(\delta)$ whenever $||\boldsymbol{\delta}_{\boldsymbol{x}}|| \le \delta$. Then, we have

$$\begin{aligned}
\max_{\substack{\boldsymbol{\delta}_{\boldsymbol{y}} \in \mathcal{R}(\boldsymbol{B}) \\ ||\boldsymbol{\delta}_{\boldsymbol{y}}|| \le h(\delta)}} f(\boldsymbol{x}^* + \boldsymbol{\delta}_{\boldsymbol{x}}, \boldsymbol{y}^* + \boldsymbol{\delta}_{\boldsymbol{y}}) - f(\boldsymbol{x}^*, \boldsymbol{y}^*) &\ge f(\boldsymbol{x}^* + \boldsymbol{\delta}_{\boldsymbol{x}}, \boldsymbol{y}^* + \boldsymbol{\delta}'_{\boldsymbol{y}}) - f(\boldsymbol{x}^*, \boldsymbol{y}^*) \\
&= \frac{1}{2} \boldsymbol{\delta}_{\boldsymbol{x}}^\top (\boldsymbol{A} - \boldsymbol{C}\boldsymbol{B}^{-1}\boldsymbol{C}^\top) \boldsymbol{\delta}_{\boldsymbol{x}} + o(\|\boldsymbol{\delta}_{\boldsymbol{x}}\|^2) > 0,
\end{aligned}$$

which concludes the proof. ■

Appendix C. Proof of Example 1

Take $h(\delta) = \frac{a}{2|c|}\delta$, then because f is linear on y, for any $|x| \le \delta$ and $|y| \le \delta$ we have

$$\max_{y' \,:\, |y'| \le \frac{a}{2|c|}\delta} -\frac{a}{2}x^2 + cxy' = -\frac{a}{2}x^2 + |cx|\frac{a}{2|c|}\delta = \frac{a|x|}{2}(\delta - |x|) \ge f(0,0) = 0 = f(0, y).$$

Hence, by definition, the point $(0, 0)$ is a local minimax point for any $a > 0$. However, in general, the point $(0, 0)$ is not a global minimax point. For example, if we let $\mathcal{Y} = \{y \;:\; |y| \le 1\}$ and $\mathcal{X}$ to contain a point x such that $|x| > \frac{2|c|}{a}$, the point $(0,0)$ is not a global minimax point for any $a < 0$, since for any $|x| > \frac{2|c|}{a}$ we have

$$\max_{y' \in \mathcal{Y}} -\frac{a}{2}x^2 + cxy' = -\frac{a}{2}x^2 + |cx| = \frac{a|x|}{2}\left(\frac{2|c|}{a} - |x|\right) < 0 = f(0,0).$$

Meanwhile, since $\nabla^2_{yy} f(0,0) = 0$, for any nonzero x and positive a it holds that

$$\max_{\substack{y' \in 0 + \mathcal{R}(\nabla^2_{yy} f(0,0)): \\ |y'-0| \leq h(\delta)}} f(x,y') = f(x,0) = -\frac{a}{2}x^2 < 0 = f(0,0).$$

So, by definition, $(0,0)$ is not a local restricted minimax point for any $a > 0$.