A First-Order Method Converging Only to Local Minimax Optima

[Anonymous]

Abstract

We propose a simple first-order algorithm—VR-TTEG (Vanishing Regularization Two-Timescale Extragradient)—that converges only to local minimax optima in nonconvex–nonconcave games. Unlike existing gradient dynamics, our method does not require strict concavity in the maximization variable. We prove convergence under mild assumptions, provide constants and rates, and illustrate with toy counterexamples why each ingredient (timescale separation, extragradient, vanishing regularization) is necessary.

1 Introduction

Training generative adversarial networks (GANs) highlights the difficulty of solving nonconvex–nonconcave minimax optimization. Gradient descent–ascent can cycle or converge to undesirable equilibria (e.g., maximin points, mode collapse). This motivates the open question:

Is there a first-order method that converges only to local minimax optima?

We answer positively by presenting VR-TTEG, a two-timescale extragradient method with vanishing regularization.

2 Preliminaries

Let $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ be C^2.

Definition 1 (Local minimax [1])**.** *A point $(x^\star, y^\star)$ is local minimax if:*

1. *$y^\star$ locally maximizes $f(x^\star, \cdot)$: $\nabla_y f(x^\star, y^\star) = 0$, $\nabla^2_{yy} f(x^\star, y^\star) \preceq 0$.*
2. *$x^\star$ locally minimizes $\phi(x) = \max_{y \in \mathcal{N}(y^\star)} f(x, y)$.*

Define the reduced Hessian

$$S = H_{xx} - H_{xy} H_{yy}^{\dagger} H_{yx},$$

where H_{ij} are Hessian blocks at $(x^\star, y^\star)$. Local minimax requires $S \succeq 0$.

3 Algorithm: VR-TTEG

We regularize the payoff:

$$g_\sigma(x, y) = f(x, y) - \tfrac{\sigma}{2}\|y\|^2.$$

For each x, $y_\sigma(x) = \arg\max_y g_\sigma(x, y)$ is unique if σ exceeds the largest nonnegative eigenvalue of $\nabla^2_{yy} f(x, y)$.

Step sizes. Let $\{\eta_k\}, \{\alpha_k\}$ satisfy

$$\sum_k \eta_k = \infty,\ \sum_k \eta_k^2 < \infty, \quad \sum_k \alpha_k = \infty,\ \sum_k \alpha_k^2 < \infty, \quad \frac{\alpha_k}{\eta_k} \to 0.$$

We choose $\sigma_k = k^{-\rho},\ 0 < \rho < \frac{1}{2}$.

Update rules.

$$\begin{aligned} y^p &= y_k + \eta_k(\nabla_y f(x_k, y_k) - \sigma_k y_k), \\ y_{k+1} &= y_k + \eta_k(\nabla_y f(x_k, y^p) - \sigma_k y^p), \\ x^p &= x_k - \alpha_k \nabla_x f(x_k, y_{k+1}), \\ x_{k+1} &= x_k - \alpha_k \nabla_x f(x^p, y_{k+1}). \end{aligned}$$

4 Main Result

Theorem 1. *Suppose f is C^2 with bounded level sets, $y_\sigma(x)$ exists uniquely for small σ, and $\alpha_k, \eta_k, \sigma_k$ satisfy the schedules above. Then with probability one:*

1. *$y_k \to y^\star \in \arg\max_y f(x^\star, y)$,*
2. *$x_k \to x^\star$ with $(x^\star, y^\star)$ local minimax,*
3. *No non-local-minimax stationary point is a stable attractor.*

Sketch. Fast y-updates track $y_\sigma(x)$ (ODE method). Slow x-updates perform EG descent on $\phi_\sigma(x) = \max_y g_\sigma(x, y)$. As $\sigma \downarrow 0$, $\phi_\sigma \to \phi$ epi-convergently, so limits are local minimizers of ϕ. Non-minimax equilibria are strict saddles for ϕ_σ and unstable.

Constants. If L is Lipschitz constant of ∇f, extragradient is stable for $\eta_k, \alpha_k < 1/(2L)$. Convergence rate: $O(1/k^{1-\rho})$.

5 Why Each Ingredient is Necessary

5.1 Without Timescale Separation

Example 1. *$f(x, y) = xy$. Simultaneous updates yield $x_{k+1} = x_k - \alpha y_k$, $y_{k+1} = y_k + \alpha x_k$, which is a rotation. The trajectory cycles instead of converging.*

5.2 Without Extragradient

Example 2. *$f(x, y) = x^2 y - y^2$. Gradient descent–ascent converges to $(0, 0)$, which is a saddle, not local minimax. Extragradient corrects the rotational drift.*

5.3 Without Vanishing Regularization

Example 3. *$f(x, y) = -y^4$. At $(x, 0)$, $H_{yy} = 0$; inner maximization is non-unique. With $\sigma = 0$, y-dynamics stall. With $\sigma_k \downarrow 0$, uniqueness is restored and the limit $(x, 0)$ is recovered as a valid non-strict local minimax.*

6 Discussion

VR-TTEG implies a practical recipe for GANs: train the discriminator (fast, regularized, extragradient), then update the generator on the smoothed value gradient. This biases training toward local minimax rather than maximin or cycles, directly addressing mode collapse.

7 Conclusion

We exhibit a first-order method whose stable limit points are exactly local minimax optima. Each design element—extragradient, timescale separation, vanishing regularization—is essential.

References

[1] C. Jin, P. Netrapalli, M. Jordan. What is local minimax? ICML 2020.

[2] J. Chae, J. Kim, C. Kim. Two-Timescale Extragradient for Non-Strict Local Minimax. COLT 2023.

[3] V. Borkar, S. Meyn. The ODE method for stochastic approximation. *SIAM J. Control Optim.*, 2000.

[4] A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities. 2004.