OPT2024: 16th Annual Workshop on Optimization for Machine Learning 1–15

Solving hidden monotone variational inequalities with surrogate losses

Ryan D'Orazio RYAN.DORAZIO@MILA.QUEBEC
Danilo Vucetic DANILO.VUCETIC@MILA.QUEBEC
Zichu Liu ZICHU.LIU@MILA.QUEBEC
Mila Québec AI Institute, Université de Montréal

Junhyung Lyle Kim JLYLEKIM@RICE.EDU
Department of Computer Science, Rice University

Ioannis Mitliagkas IOANNIS@MILA.QUEBEC
Gauthier Gidel GIDELGAU@MILA.QUEBEC
Mila Québec AI Institute, Université de Montréal, CIFAR AI Chair

Abstract

Deep learning has proven to be effective in a wide variety of loss minimization problems. However, many applications of interest, like minimizing projected Bellman error and min-max optimization, cannot be modelled as minimizing a scalar loss function but instead correspond to solving a variational inequality (VI) problem. This difference in setting has caused many practical challenges as naive gradient-based approaches from supervised learning tend to diverge and cycle in the VI case. In this work, we propose a principled surrogate-based approach compatible with deep learning to solve VIs. We propose a surrogate-based approach that is principled in the VI setting and compatible with deep learning. We show that our approach has three main benefits: (1) it guarantees linear convergence under sufficient descent in the surrogate when hidden monotone structure is present (e.g. convex-concave in with respect to model predictions), (2) it provides a unifying perspective of existing methods, and (3) is amenable to existing deep learning optimizers like ADAM.

Keywords: Variational Inequality Problems, Hidden Monotone, Surrogate Loss.

1. Introduction

Most machine learning approaches learn from data by minimizing a loss function with respect to model parameters. Despite the non-convexity of such losses and lack of global convergence guarantees due to deep learning, they can often still be approximately minimized with an appropriately tuned first-order adaptive method such as ADAM (Kingma, 2014). Unfortunately, outside of scalar loss minimization, the challenges of using deep learning are exacerbated: the dynamics of variational inequality (VI) problems (*e.g.*, min-max) are often plagued with rotations and posses no efficient stationary point guarantees (Daskalakis et al., 2021). Thus, the additional challenges posed by VI problems do not allow one to easily plug in existing techniques from supervised learning.

Despite these optimization challenges, loss functions are typically well-grounded and chosen carefully, admitting monotonicity (e.g. a convex-concave min-max objective) and

smoothness with respect to model outputs. More precisely, such problems admit a hidden structure corresponding to the following VI problem: find z_* such that

$$\langle F(z_*), z - z_* \rangle \geq 0 \quad \forall z \in \mathcal{Z} = \text{cl}\{g(\theta) : \theta \in \mathbb{R}^d\}. \tag{1}$$

Where the mapping $g : \mathbb{R}^d \to \mathcal{Z} \subseteq \mathbb{R}^n$ maps model parameters to model outputs, and the set $\mathcal{Z}$ encodes the closure of the set of realizable outputs from the chosen model. Since F is defined over model outputs, it will often be monotone and smooth.

In this work, we leverage the structure in model outputs in VI problems by extending the use of surrogate losses (Johnson and Zhang, 2020; Vaswani et al., 2021) from scalar minimization. The surrogate loss approach has been shown to be scalable with deep learning and has been used in modern reinforcement learning policy gradient methods (Schulman et al., 2015, 2017; Abdolmaleki et al., 2018).

Our approach reduces the VI (1) to the approximate minimization of a sequence of surrogate losses $\{\ell_t\}_{t\in\mathbb{N}}$ for which are then used to produce a sequence of parameters $\{\theta_t\}_{t\in\mathbb{N}}$. To ensure convergence we propose a new α-descent condition on ℓ_t, allowing for a dynamic inner-loop that makes no assumption on how the surrogate losses are minimized, thereby allowing for any deep-learning optimizer to minimize the scalar loss ℓ_t. With our α-descent condition we provide convergence guarantees to a solution in the space of model predictions $\{z_t = g(\theta_t)\}_{t\in\mathbb{N}} \to z_*$ for a sufficiently small α. Our general method as described above is summarized in Algorithm 1. Our contributions can be summarized as follows:

- **Extension of surrogate losses to VI problems and challenges.** This work provides the first extension of surrogate losses to VI problems. Although scalar minimization is as a special case, we show there is a clear separation in problem difficulty. In the scalar minimization case any progress on the surrogate loss is sufficient for convergence (Vaswani et al., 2021; Lavington et al., 2023); meanwhile, we show that it is possible to diverge in the VI case with consistent progress on the surrogate losses even if F is strongly monotone.
- **Inner-loop condition and convergence guarantees.** In contrast to existing approaches, we control our inner-loop via an α-descent condition. This condition allows for global convergence without forcing a global upper bound on all losses $\{\ell_t\}_{t\in\mathbb{N}}$.
- **Unifying perspective of pre-conditioning methods.** With our surrogate loss approach we are able to unify existing pre-conditioning methods (Bertsekas, 2009; Mladenovic et al., 2022; Sakos et al., 2024). In Section 4 we show they are equivalent to picking the Gauss-Newton method as the optimizer $\mathcal{A}$ in Algorithm 1.

2. Background and Related Work

We use standard notation from optimization, for more details please see Appendix A.

Surrogate Loss Background. In scalar minimization, such as supervised learning, a non-convex loss function of the form $f(g(\theta))$ is minimized where the non-convexity is due to the model parametrization. In this case $g(\theta) \in \mathbb{R}^n$ where n is the number of samples, each prediction is $z^i = g^i(\theta) = h(x_i, \theta)$, for some feature vector x_i and fixed model architecture h. Despite the non-convexity with respect to θ, the loss function is often convex and smooth

Algorithm 1: α-descent on surrogate

Input: Outer loop interations T, initial parameters $\theta_1 \in \mathbb{R}^d$, η stepsize for surrogate loss, $\alpha \in (0, 1)$, optimizer update $\mathcal{A} : \mathcal{L} \times \mathbb{R}^d \to \mathbb{R}^d$.

for $t = 1 \leftarrow$ **to** T **do**

  Compute VI operator: $F(g(\theta_t))$

  Set loss: $\ell_t(\theta) = \frac{1}{2}\|g(\theta) - (g(\theta_t) - \eta F(g(\theta_t))\|^2$

  $\theta_s \leftarrow \theta_t$

  while $\ell_t(\theta_s) - \ell_t^* > \alpha^2(\ell_t(\theta_t) - \ell_t^*)$ **do**

    Update parameters with optimizer: $\theta_s \leftarrow \mathcal{A}(\ell_t, \theta_s)$

  $\theta_{t+1} \leftarrow \theta_s$

return θ_{T+1}

with respect to the closure of the predictions $\mathcal{Z} = \text{cl}\{g(\theta) : \theta \in \mathbb{R}^d\}$. The optimization problem can then be often reframed as the constrained convex optimization problem

$$\min_{\theta} f(g(\theta)) = \min_{z \in \mathcal{Z}} f(z). \tag{2}$$

If $\mathcal{Z}$ is convex then projected gradient descent $z_{t+1} = \Pi(z_t - \eta \nabla f(z_t))$ is guaranteed to converge (Beck, 2017). However, the projection Π is expensive since it is with respect to the set $\mathcal{Z} \subseteq \mathbb{R}^n$. In general, the model and data dependent constraint $\mathcal{Z}$ may not be convex, and is a limitation of our analysis. However, this assumption is satisfied in two extreme cases, when a model is linear or with large capacity neural networks where $\mathcal{Z} = \mathbb{R}^n$.

Beyond supervised learning, similar hidden structure exists in losses within machine learning such as: generative models and min max optimization (Gidel et al., 2021). However, these applications cannot be written as minimizing a loss and we must instead consider the VI problem (1). In the min-max case, the min and max players' strategies may be given by two separate networks $h^1(\theta^1), h^2(\theta^2)$, respectively, with the following objective:

$$\min_{\theta^1} \max_{\theta^2} f(h^1(\theta^1), h^2(\theta^2)), \tag{3}$$

where f is convex-concave. Similar to the scalar minimization case we can rewrite the min-max problem in parameter space as a constrained problem with respect to model predictions but instead within the the VI (1); where $\theta = (\theta^1, \theta^2)$, and $g(\theta) = (h^1(\theta^1), h^2(\theta^2))$, with operator $F(z) = F(g(\theta)) = [\nabla_{z^1} f(z^1, z^2), -\nabla_{z^2} f(z^1, z^2)]^\top$ where $z^1 = h^1(\theta^1)$ and $z^2 = h^2(\theta^2)$. If F is well-conditioned and $\mathcal{Z}$ is closed and convex, then the projected gradient method $z_{t+1} = \Pi(z_t - \eta F(z_t))$ converges to a solution z_* with an appropriate stepsize η (Facchinei and Pang, 2003; Bauschke and Combettes, 2017).

To solve problems of the form (1) and take advantage of the structure given by g and F, we extend surrogate losses (Johnson and Zhang, 2020; Vaswani et al., 2021). At iteration t, θ_{t+1} is selected by descending the surrogate loss

$$\ell_t(\theta) = \frac{1}{2}\|g(\theta) - [g(\theta_t) - \eta F(g(\theta_t))]\|^2. \tag{4}$$

Denoting z_t^* as the exact projected gradient step $\Pi(z_t - \eta F(z_t))$, minimizing the surrogate exactly would ensure that $z_t = g(\theta_t) = z_t^*$ and therefore guarantee convergence of $\{z_t\}_{t \in \mathbb{N}}$.

Related work. The surrogate losses proposed by Johnson and Zhang (2020) and Vaswani et al. (2021), apply to supervised learning and RL respectively.[1] They did not study the VI case, nor do they exploit any convexity properties that may be present in the scalar minimization case. Lavington et al. (2023) also study the scalar minimization case and provide convergence to a neighbourhood of a global minimum of (2) and allow for stochasticity. The neighbourhood of convergence depends both on an upperbound on the errors $\epsilon_t = \|z_{t+1} - z_t^*\|$ and a variance term. Therefore, the neighbourhood of convergence scales with the worst error ϵ_t across the trajectory $\{z_t = g(\theta_t)\}_{t\in\mathbb{N}}$. Shrinking the neighbourhood necessitates a double loop algorithm that might unecessarily spend too much time optimizing the surrogate.

In contrast to the existing surrogate loss approaches, we propose a simple α-descent condition on ℓ_t (Definition 1) that does not require all errors to be bounded or summable apriori. This condition allows for convergence without fully minimizing ℓ_t and better models implementations in practice where a fixed number of gradient descent steps are used.

Definition 1 (α-descent) *Let ℓ_t be the surrogate defined in (4) and $\ell_t^* = \inf_{\theta\in\mathbb{R}^d} \ell_t(\theta)$. The trajectory $\{\theta_t\}_{t\in\mathbb{N}}$ satisfies the α-descent condition if at each step t the following holds*

$$\ell_t(\theta_{t+1}) - \ell_t^* \leq \alpha^2 \left(\ell_t(\theta_t) - \ell_t^*\right), \quad \alpha \in [0, 1), \tag{5}$$

Given the α-descent condition we can define a general purpose algorithm (Algorithm 1). With any black-box optimizer update $\mathcal{A}$ and a double-loop structure, we can construct a trajectory that satisfies the condition so long as $\mathcal{A}$ can effectively descend the least-squares loss ℓ_t. In general ℓ_t^* may not be zero and so this condition cannot be verified directly, however, this condition can often be met via first-order methods for a fixed number of steps or can be approximated with $\ell_t^* = 0$.

Hidden monotone problems and preconditioning methods. Gidel et al. (2021) study the existence of equilibria in zero-sum games with hidden monotonicity but do not propose an algorithm to take advantage of the hidden structure. For games that admit a hidden strictly convex-concave structure, Vlatakis-Gkaragkounis et al. (2021) prove global convergence of continuous time gradient descent-ascent. Similarly, Mladenovic et al. (2022) propose natural hidden gradient dynamics (NHGD) with continuous time convergence guarantees. A more general and discretized version of NHGD was studied by Sakos et al. (2024), the preconditioned hidden gradient descent method (PHGD), to solve VIs of the form (1). PHGD and stochastic variants were also studied in the linear case by Bertsekas (2009). In Section 4 we show that PHGD is equivalent to taking one step of the Gauss-Newton method (GN) (Björck, 1996) on the surrogate loss.

3. Convergence and divergence under α-descent on surrogate losses

Under the α-descent condition (1) with a sufficiently small α, and Assumption 2, Theorem 3 guarantees linear convergence of $\{z_t = g(\theta_t)\}_{t\in\mathbb{N}}$ to the solution z_* of the VI (1).

1. We note that Johnson and Zhang (2020) and Vaswani et al. (2021) provide a more general loss using Bregman divergences.

Assumption 2 *In the VI (1), $\mathcal{Z}$ is closed and convex. There exists a solution within the relative interior, $z_* \in \operatorname{ri} \mathcal{Z}$. F is both L-smooth $\|F(x) - F(y)\| \leq L\|x - y\|$ and μ-strongly monotone $\langle F(x) - F(y), x - y\rangle \geq \mu\|x - y\|^2$ for any $x, y \in \mathcal{Z}$ and some $\mu > 0$.*

Theorem 3 *Let Assumption 2 hold and let $\{z_t = g(\theta_t)\}_{t\in\mathbb{N}}$ be the iterates produced by Algorithm 1. If α and η are picked such that $\rho := 1 - 2\eta(\mu - \alpha L) + (1 + \alpha^2)\eta^2 L^2 < 1$ then, z_t converge linearly to the solution z_* at the following linear rate:*

$$\|z_{t+1} - z_*\|^2 \leq \rho^t \|z_1 - z_*\|^2. \tag{6}$$

Particularly, if $\alpha < \frac{\mu}{L}$ and $\eta < \frac{2(\mu - \alpha L)}{(1+\alpha^2)L^2}$ then $\rho < 1$ and if $\alpha \leq \frac{\mu}{2L}$ and $\eta = \frac{2\mu}{5L^2}$ then $\rho \leq 1 - \frac{\mu^2}{5L^2}$.

Divergence with $\alpha < 1$. In the scalar minimizing case, $\alpha < 1$ guarantees convergence to a stationary point with a smooth *non-convex* loss function (see Proposition 7). However, for the VI case, we show that α small enough is *necessary* for convergence. Our construction uses the strongly convex-concave min-max problem $\min_x \max_y \frac{1}{2}x^2 + xy - \frac{1}{2}y^2$ to show that gradient descent-ascent on $f(x, y) = xy$ satisfies the α-descent condition with $\alpha = \frac{1}{\sqrt{2}}$.

Proposition 4 *There exists an L-smooth and μ-strongly monotone F, and a sequence of iterates $\{z_t\}_{t\in\mathbb{N}}$ verifying the alpha descent condition with $\alpha < 1$ yet z_t diverges for all η.*

4. A Nonlinear Least Squares Perspective on Surrogate Minimization

The surrogate loss perspective and our α-descent condition allows for convergence so long as the surrogate losses $\{\ell_t\}_{t\in\mathbb{N}}$ are sufficiently minimized. One approach to minimizing ℓ_t is to view it as the following non-linear least-squares problem

$$\min_\theta f(\theta) = \min_\theta \frac{1}{2}\|r(\theta)\|^2, \tag{7}$$

with a residual function $r : \mathbb{R}^d \to \mathbb{R}^n$, where $\ell_t(\theta) = f(\theta)$ if $r(\theta) = g(\theta) - g(\theta_t) + \eta F(g(\theta_t))$. Due to the specific structure of f we can consider specialized methods such as Gauss-Newton (GN), Damped Gauss-Newton (DGN), and Levenbergh-Marquardt (LM) (Björck, 1996; Nocedal and Wright, 1999). These methods can be viewed as quasi-Newton methods that use a linear approximation of r, $r(\theta) \approx r(\theta_t) + Dr(\theta_t)(\theta - \theta_t)$.

The GN method is defined by the update rule $\theta_{t+1} = \theta_t - (Dr(\theta_t)^\top Dr(\theta_t))^\dagger Dr(\theta_t)^\top r(\theta_t)$. GN inherits the same local quadratic convergence properties as Newton's method when the Hessian at the minimum $\nabla^2 f(\theta_*) \approx Dr(\theta_*)^\top Dr(\theta_*)$. However, GN is known to struggle with highly non-linear problems, those with large residuals, or if $Dr(\theta_t)$ is nearly rank-deficient (Björck, 1996; Nocedal and Wright, 1999). Fortunately, the GN direction is a descent direction of f, the DGN method takes steps in the GN direction with a stepsize parameter η_{GN} and converges for a sufficiently small stepsize or with line search (Björck, 1996). In cases where $Dr(\theta_t)$ is nearly rank deficient the LM method can be used instead.

To minimize the surrogate we can therefore consider taking multiple steps of gradient descent (GD), DGN or LM. Denoting θ_t^s as s^{th} intermediate step between θ_{t+1} and θ_t we

consider the following updates:

$$\theta_t^{s+1} = \theta_t^s - \eta_{GD} Dg(\theta_t^s)^\top (g(\theta_t^s) - g(\theta_t) + \eta F(g(\theta_t)) \tag{GD}$$

$$\theta_t^{s+1} = \theta_t^s - \eta_{GN} (Dg(\theta_t^s)^\top Dg(\theta_t^s))^\dagger Dg(\theta_t^s)^\top (g(\theta_t^s) - g(\theta_t) + \eta F(g(\theta_t)) \tag{DGN}$$

$$\theta_t^{s+1} = \theta_t^s - (Dg(\theta_t^s)^\top Dg(\theta_t^s) + \lambda \,\mathrm{Id})^{-1} Dg(\theta_t^s)^\top (g(\theta_t^s) - g(\theta_t) + \eta F(g(\theta_t)). \tag{LM}$$

Note that we used the fact that $Dr(\theta) = Dg(\theta)$, and if $\eta_{GN} = 1$ then DGN is the same as GN. Also, note that one step of GN recovers the PHGD method from Sakos et al. (2024), $\theta_{t+1} = \theta_t - \eta(Dg(\theta_t)^\top Dg(\theta_t))^\dagger Dg(\theta_t)^\top F(g(\theta_t))$.

In Appendix C we compare GD, PHGD, DGN, LM, in two domains from Sakos et al. (2024), the hidden matching pennies game and hidden rock-paper-scissors.

References

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018.

Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Dimitri P Bertsekas. Projected equations, variational inequalities, and temporal difference methods. *Lab. for Information and Decision Systems Report LIDS-P-2808, MIT*, 2009.

Åke Björck. *Numerical methods for least squares problems*. SIAM, 1996.

Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1466–1478, 2021.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Barbara Franci and Sergio Grammatico. Convergence of sequences: A survey. *Annual Reviews in Control*, 53:161–186, 2022. ISSN 1367-5788. doi: https://doi.org/10.1016/j.arcontrol.2022.01.003. URL https://www.sciencedirect.com/science/article/pii/S1367578822000037.

Gauthier Gidel, David Balduzzi, Wojciech Czarnecki, Marta Garnelo, and Yoram Bachrach. A limited-capacity minimax theorem for non-convex games or: How i learned to stop worrying about mixed-nash and love neural nets. In *International Conference on Artificial Intelligence and Statistics*, pages 2548–2556. PMLR, 2021.

Rie Johnson and Tong Zhang. Guided learning of nonconvex models through successive functional gradient optimization. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4921–4930. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/johnson20b.html.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jonathan Wilder Lavington, Sharan Vaswani, Reza Babanezhad Harikandeh, Mark Schmidt, and Nicolas Le Roux. Target-based surrogates for stochastic optimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18614–18651. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lavington23a.html.

Andjela Mladenovic, Iosif Sakos, Gauthier Gidel, and Georgios Piliouras. Generalized natural gradient flows in hidden convex-concave games and GANs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=bsycpMi0OR1.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

R.T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1997. ISBN 9780691015866. URL https://books.google.ca/books?id=wj4Fh4h_V7QC.

Iosif Sakos, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Panayotis Mertikopoulos, and Georgios Piliouras. Exploiting hidden structures in non-convex games for convergence to nash equilibrium. *Advances in Neural Information Processing Systems*, 36, 2024.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Müller, Shivam Garg, Matthieu Geist, Marlos C Machado, Pablo Samuel Castro, and Nicolas Le Roux. A general class of surrogate functions for stable and efficient reinforcement learning. *arXiv preprint arXiv:2108.05828*, 2021.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Solving min-max optimization with hidden structure via gradient descent ascent. *Advances in Neural Information Processing Systems*, 34:2373–2386, 2021.

Appendix A. Notation

We use $\langle x, y\rangle = \sum_{i=1}^n x^i y^i$ to denote the standard inner product over $\mathbb{R}^n$ and $\|x\| = \sqrt{\langle x, x\rangle}$ to be the Euclidean norm. We write $\|x\|_\Xi^2$ to mean $\langle x, \Xi x\rangle$, and $\|x\|_\Xi$ is a norm if and only if Ξ is positive definite. For a set $\mathcal{X}$ we denote $\operatorname{cl}\mathcal{X}$ its closure and $\operatorname{ri}\mathcal{X}$ its relative interior. For a given set, which will be clear from context, we denote $\Pi(x)$ as the Euclidean projection of x onto the set and similarly $\Pi_\Xi(x)$ the projection with respect to the norm $\|x\|_\Xi$. We use Id to denote the identity matrix. A matrix A has lower and upper bounded singular values if there exists $\sigma_{\min}, \sigma_{\max} \in (0, \infty)$ such that for any x we have $\sigma_{\min}^2\|x\|^2 \leq \langle x, A^\top A x\rangle \leq \sigma_{\max}^2\|x\|^2$. If a matrix A is invertible we write A^{-1} otherwise we denote the pseudo-inverse as $A^\dagger$. For a given function $g : \mathbb{R}^d \to \mathbb{R}^n$ we write $Dg(\theta)$ as its Jacobian evaluated at θ. We say $Dg^\top$ has uniformly lower and upper bounded singular values if there is a constant upper and lower bound to the singular values of $Dg(\theta)^\top$ for all $\theta \in \mathbb{R}^d$. We use $\Delta_n = \{x \in \mathbb{R}^n : \sum_i x^i = 1, x^i \geq 0\}$ to denote the $(n-1)$-dimensional simplex.

Appendix B. Proofs

Remark 5 *If g is continuous, and $\{g(\theta) : \theta \in \mathbb{R}^d\}$ is convex with $\mathcal{Z}$ as its closure, then the least-squares surrogate loss $\ell_t(\theta) = \frac{1}{2}\|g(\theta) - z_t + \eta F(z_t)\|^2$ admits a unique point $z_t^* \in \mathcal{Z}$ such that*

$$\ell_t^* = \frac{1}{2}\|z_t^* - z_t + \eta F(z_t)\|^2,$$

and for any θ

$$\frac{1}{2}\|g(\theta) - z_t^*\|^2 \leq \ell_t(\theta) - \ell_t^*,$$

where $\ell_t^* = \inf_{\theta\in\mathbb{R}^d} \ell_t(\theta)$.

Proof Let $f(z) = \frac{1}{2}\|z - z_t + \eta F(z_t)\|^2$ be the surrogate loss with respect to the predictions $z = g(\theta)$. We have that $f(z) = \ell_t(\theta)$ for all $\theta \in \mathbb{R}^d$. Now consider the set $\mathcal{Z} = \operatorname{cl}\{g(\theta) : \theta \in \mathbb{R}^d\}$, since it is closed and convex, we have that f has a unique minimum z_t^* because it is 1-strongly convex. Furthermore, we have

$$\frac{1}{2}\|z - z_t^*\|^2 \leq f(z) - f(z_t^*).$$

Now since $z_t^* \in \mathcal{Z}$ and $\mathcal{Z}$ is the closure of $\{g(\theta) : \theta \in \mathbb{R}^d\}$, there exists a sequence of parameters $\{\theta_t\}_{t\in\mathbb{N}}$ such that $\{z_t = g(\theta_t)\}_{t\in\mathbb{N}} \to z_t^*$. Therefore we have that,

$$f(z_t^*) = \lim_{t\to\infty} f(z_t) = \lim_{t\to\infty} \ell_t(\theta_t) \geq \ell_t^* \geq f(z_t^*).$$

Where we have used the continuity of f and g. The last inequality follows because $\{g(\theta) : \theta \in \mathbb{R}^d\} \subseteq \mathcal{Z}$. Therefore $\ell_t^* = f(z_t^*)$ and the result follows. ■

Lemma 6 *If F is monotone and z_* is in the relative interior of the constraint $\mathcal{Z}$, then*

$$\ell_t(\theta_t) - \ell_t^* \leq \frac{\eta^2}{2}\|F(z_t) - F(z_*)\|^2.$$

Furthermore, under the α-descent condition we have

$$\|z_{t+1} - z_t^*\| \leq \alpha\eta\|F(z_t) - F(z_*)\|.$$

Proof If z_* is a solution then we have

$$\langle F(z_*), z - z_* \rangle \geq 0\,,\ \forall z \in \mathcal{Z}.$$

If z_* is in the relative interior then for any $z \in \mathcal{Z}$ there exists a $\lambda > 1$ such that $z' = (1-\lambda)z + \lambda z_* \in \mathcal{Z}$ (Rockafellar, 1997)[Theorem 6.4]. Therefore by optimality we have

$$0 \leq \langle F(z_*), z' - z_* \rangle = \langle F(z_*), (1-\lambda)z + \lambda z_* - z_* \rangle = (1-\lambda)\langle F(z_*), z - z_* \rangle \leq 0.$$

Where the last inequality follows from $\lambda > 1$. Altogether we have

$$\langle F(z_*), z - z_* \rangle = 0\,,\ \forall z \in \mathcal{Z}.$$

Moreover, as a consequence we have that for any two points $z, z' \in \mathcal{Z}$

$$\langle F(z_*), z - z' \rangle = \langle F(z_*), z - z_* \rangle + \langle F(z_*), z_* - z' \rangle = 0.$$

Letting z_t^* be the exact projected gradient step and $z_t = g(\theta_t)$ the current iterate we have

$$\begin{aligned}
\ell_t(\theta_t) - \ell_t^* &= \frac{1}{2}\|\eta F(z_t)\|^2 - \left(\frac{1}{2}\|z_t^* - z_t + \eta F(z_t)\|^2\right) \\
&= \eta\langle F(z_t), z_t - z_t^* \rangle - \frac{1}{2}\|z_t^* - z_t\|^2 \\
&= \eta\langle F(z_t) - F(z_*), z_t - z_t^* \rangle + \eta\langle F(z_*), z_t - z_t^* \rangle - \frac{1}{2}\|z_t^* - z_t\|^2 \\
&= \eta\langle F(z_t) - F(z_*), z_t - z_t^* \rangle - \frac{1}{2}\|z_t^* - z_t\|^2 \\
&\leq \frac{\eta^2}{2}\|F(z_t) - F(z_*)\|^2.
\end{aligned}$$

Where the last two inequalities follow by: z_* being within the relative interior, and the Fenchel-Young inequality.

By Remark 5 and the α-decent condition,

$$\frac{1}{2}\|z_{t+1} - z_t^*\|^2 \leq \ell_t(\theta_{t+1}) - \ell_t^* \leq \alpha^2(\ell_t(\theta_t) - \ell_t^*) \leq \alpha^2\frac{\eta^2}{2}\|F(z_t) - F(z_*)\|^2.$$

■

Proposition 7 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. For some $g : \mathbb{R}^d \to \mathbb{R}^n$ define the surrogate loss $\ell_t(\theta) = \frac{1}{2}\|g(\theta) - z_t + \frac{1}{L}\nabla f(z_t)\|^2$, where $z_t = g(\theta_t)$. Then the α-descent condition guarantees*

$$f(z_{t+1}) \leq f(z_t) - L(1-\alpha^2)(\ell_t(\theta_t) - \ell_t^*).$$

Proof Let $\hat{z}_t = z_t - \frac{1}{L}\nabla f(z_t)$.

$$\begin{aligned} f(z_{t+1}) - f(z_t) &\leq \langle \nabla f(z_t), z_{t+1} - z_t \rangle + \frac{L}{2}\|z_{t+1} - z_t\|^2 \\ &= L\langle z_t - \hat{z}_t, z_{t+1} - z_t \rangle + \frac{L}{2}\|z_{t+1} - z_t\|^2 \\ &= L\left(-\|\hat{z}_t - z_t\|^2 + \langle z_t - \hat{z}_t, z_{t+1} - \hat{z}_t \rangle + \frac{1}{2}\|z_{t+1} - z_t\|^2\right) \\ &= L\left(\frac{1}{2}\|\hat{z}_t - z_{t+1}\|^2 - \frac{1}{2}\|\hat{z}_t - z_t\|^2\right) \\ &= L\left(\ell_t(\theta_{t+1}) - \ell_t(\theta_t)\right). \end{aligned}$$

The second to last equality follows from exanding $\frac{1}{2}\|z_{t+1} - z_t\|^2 = \frac{1}{2}\|z_{t+1} - \hat{z}_t + \hat{z}_t - z_t\|^2$. Using the α-descent condition we have

$$\ell_t(\theta_{t+1}) - \ell_t(\theta_t) \leq (\alpha^2 - 1)(\ell_t(\theta_t) - \ell_t^*),$$

yielding the result. ■

Theorem 3 *Let Assumption 2 hold and let $\{z_t = g(\theta_t)\}_{t\in\mathbb{N}}$ be the iterates produced by Algorithm 1. If α and η are picked such that $\rho := 1 - 2\eta(\mu - \alpha L) + (1+\alpha^2)\eta^2 L^2 < 1$ then, z_t converge linearly to the solution z_* at the following linear rate:*

$$\|z_{t+1} - z_*\|^2 \leq \rho^t \|z_1 - z_*\|^2. \tag{8}$$

Particularly, if $\alpha < \frac{\mu}{L}$ and $\eta < \frac{2(\mu - \alpha L)}{(1+\alpha^2)L^2}$ then $\rho < 1$ and if $\alpha \leq \frac{\mu}{2L}$ and $\eta = \frac{2\mu}{5L^2}$ then $\rho \leq 1 - \frac{\mu^2}{5L^2}$.

Proof First note that by definition of Algorithm 1, the iterates $z_t = g(\theta_t)_{t\in\mathbb{N}}$ satisfy the α-descent property (Definition 1), therefore Lemma 6 holds. Recall that z_t^*, the exact projection update, is a contraction if $\eta < \frac{2\mu}{L}$ since

$$\|z_t^* - z_*\|^2 \leq \kappa^2 \|z_t - z_*\|^2 \tag{9}$$

where $\kappa^2 = 1 - 2\eta\mu + \eta^2 L^2$(Facchinei and Pang, 2003, Theorem 12.1.2). For the remainder, assume that $\eta < \frac{2\mu}{L}$ so that $\kappa \in [0, 1)$. Denoting $F_t = F(z_t)$ and $F_* = F(z_*)$, we have

$$\begin{aligned} \|z_{t+1} - z_*\|^2 &= \|z_t^* - z_* + z_{t+1} - z_t^*\|^2 \\ &= \|z_t^* - z_*\|^2 + 2\langle z_{t+1} - z_t^*, z_t^* - z_* \rangle + \|z_{t+1} - z_t^*\|^2 \\ &\leq \|z_t^* - z_*\|^2 + 2\|z_{t+1} - z_t^*\|\|z_t^* - z_*\| + \|z_{t+1} - z_t^*\|^2 && \text{(Cauchy–Schwarz)} \\ &\leq \|z_t^* - z_*\|^2 + 2\alpha\eta\kappa\|F_t - F_*\|\|z_t - z_*\| + \alpha^2\eta^2\|F_t - F_*\|^2 && \text{(Lemma 6)} \\ &\leq \|z_t^* - z_*\|^2 + 2\alpha\eta L\|z_t - z_*\|^2 + \alpha^2\eta^2 L^2\|z_t - z_*\|^2 && \text{(Smoothness of } F \text{ and } \kappa < 1) \\ &\leq \kappa^2\|z_t - z_*\|^2 + 2\alpha\eta L\|z_t - z_*\|^2 + \alpha^2\eta^2 L^2\|z_t - z_*\|^2 && \text{(Eq. 9)} \\ &= \|z_t - z_*\|^2\left(1 - 2\eta\mu + 2\alpha\eta L + (1+\alpha^2)\eta^2 L^2\right). \end{aligned}$$

If $\alpha < \frac{\mu}{L}$ then

$$\|z_{t+1} - z_*\|^2 \leq \|z_t - z_*\|^2 \left(1 - 2\eta \underbrace{(\mu - \alpha L)}_{>0} + (1+\alpha^2)\eta^2 L^2\right).$$

Taking $\eta < \frac{2(\mu - \alpha L)}{(1+\alpha^2)L^2}$ would guarantee a contraction.

If $\alpha \leq \frac{\mu}{2L}$ and taking $\eta = \frac{2\mu}{5L^2}$ we have:

$$\begin{aligned} 1 - 2\eta\mu + 2\alpha\eta L + (1+\alpha^2)\eta^2 L^2 &\overset{\alpha \leq \mu/2L \leq 1/2}{\leq} 1 - \eta\mu + \left(1 + \frac{1}{4}\right)\eta^2 L^2 \\ &\overset{\eta = \frac{2\mu}{L^2}}{=} 1 - \frac{2\mu^2}{5L^2} + \frac{\mu^2}{5L^2} = 1 - \frac{\mu^2}{5L^2}. \end{aligned}$$

■

Proposition 9 *Take θ_{t+1} to be an approximate minima of $\ell_t(\theta)$. Suppose $T = \Pi \circ (\mathrm{Id} - \eta F)$ is a contraction, if $\ell_t(\theta_{t+1}) - \ell_t^* \to 0$ then the induced sequence $\{z_t = g(\theta_t)\}_{t \in \mathbb{N}}$ converges, $z_t \to z_*$ where z_* is the unique solution to $\mathrm{VI}(\mathcal{Z}, F)$.*

Proof Let $\epsilon_t = z_{t+1} - z_t^*$ be the approximation error between z_{t+1} and z_t^* the minimum of the surrogate ℓ_t (exact projected gradient step).

$$\begin{aligned} \frac{1}{2}\|z_{t+1} - z_*\|^2 &= \frac{1}{2}\|z_t^* - z_* + \epsilon_t\|^2 \\ &= \frac{1}{2}\|z_t^* - z_*\|^2 + \langle z_t^* - z_*, \epsilon_t \rangle + \frac{1}{2}\epsilon_t^2 \\ &\overset{\rho > 0}{\leq} \frac{1}{2}\|z_t^* - z_*\|^2 + \frac{\rho}{2}\|z_t^* - z_*\|^2 + \frac{1}{2\rho}\epsilon_t^2 + \frac{1}{2}\epsilon_t^2 \\ &\overset{\kappa \in [0,1)}{\leq} \frac{\kappa(1+\rho)}{2}\|z_t - z_*\|^2 + (1 + \frac{1}{\rho})(\ell_t(\theta_{t+1}) - \ell_t^*). \end{aligned}$$

Where we use the fact that $\frac{1}{2}\epsilon_t^2 = \frac{1}{2}\|z_{t+1} - z_t^*\|^2 \leq \ell_t(\theta_{t+1}) - \ell_t^*$ from Remark 5. Take any ρ such that $\kappa(1+\rho) < 1$ then apply Lemma 3.9 in Franci and Grammatico (2022). ■

Proposition 4 *There exists an L-smooth and μ-strongly monotone F, and a sequence of iterates $\{z_t\}_{t \in \mathbb{N}}$ verifying the alpha descent condition with $\alpha < 1$ such that z_t diverges for any η.*

Proof

Let us consider the simple min max example with loss

$$\min_x \max_y \frac{1}{2}x^2 + xy - \frac{1}{2}y^2.$$

This problem can be written as VIP where $z = (x, y)$ and the operator F is linear given by the matrix

$$F = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

It is well-known that F is both smooth and strongly monotone with $z_{t+1} = z_t - \eta F(z_t)$ being a contraction for small enough η (Facchinei and Pang, 2003). Now if we consider a biased direction given by a matrix P, that is $z_{t+1} = z_t - Pz_t$, then using the fact that $\ell_t^* = 0$ the α-descent on the surrogate corresponds to the following bound

$$\|z_{t+1} - z_t + \eta F z_t\| = \|(\eta F - P)z_t\| \leq \alpha \|\eta F z_t\|.$$

Despite being a contraction when we follow the true gradient F, the above min max loss causes rotations in the dynamics that are inherent to the adversarial nature of the problem. These rotations are carefully controlled by the stepsize and strong convexity/concavity of the loss. Our counterexample simply adds a bit of rotation that ensures $\alpha < 1$ but yet is detrimental to the convergence. Mathematically, we take $P = (\mathrm{Id} - \alpha Q)\eta F$ where Q is the rotation matrix

$$Q = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}. \end{bmatrix}$$

With this construction we are guaranteed that $\alpha < 1$ since

$$\|(\eta F - P)z_t\| = \|\alpha Q \eta F z_t\| = \alpha\eta \|F z_t\|,$$

where the last equality is due the the fact that Q is an orthogonal matrix and therefore does not change the Euclidean norm of a vector.

Now taking $\alpha = 1/\sqrt{2}$ gives

$$\begin{aligned} P &= \left(\mathrm{Id} - \alpha \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \right) \begin{bmatrix} \eta & \eta \\ -\eta & \eta \end{bmatrix}. \\ &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \eta & \eta \\ -\eta & \eta \end{bmatrix} = \begin{bmatrix} 0 & \eta \\ -\eta & 0. \end{bmatrix} \end{aligned}$$

We have that $z_{t+1} = z_t - Pz_t = \begin{bmatrix} 1 & -\eta \\ \eta & 1 \end{bmatrix} z_t$ has an $\alpha = 1/\sqrt{2}$ but yet diverges for any $\eta > 0$ since the Eigen values of the linear system are $\lambda = 1 \pm i\eta$ therefore the spectral radius is strictly greater than one. Note that these dynamics are equivalent to gradient descent on the bilinear loss $f(x, y) = xy$, which is known to diverge for any stepsize.

■

Appendix C. Min-max Experiments

To demonstrate our surrogate loss approach we compare different approaches from Section 4, namely: GN, DGN, LM, and GD. Taking only one inner step for GN and GD gives PHGD and gradient descent-ascent (GDA) as special cases respectively.

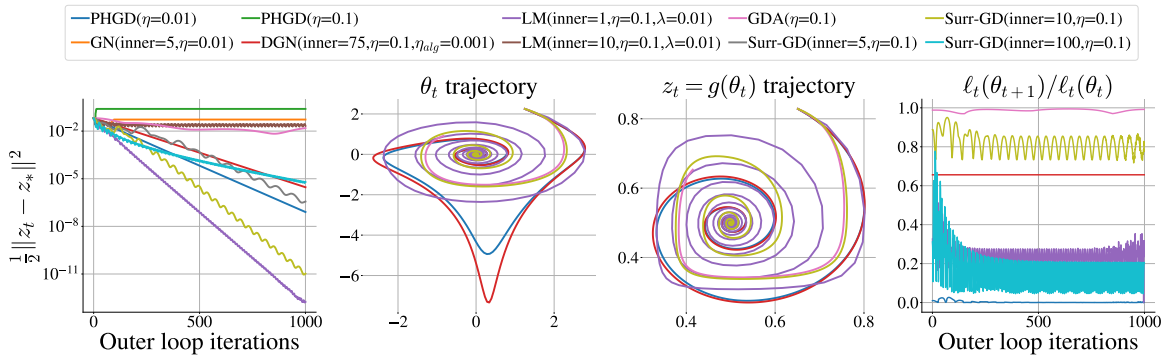

Figure 1: Convergence of various algorithms from Section 4 on the hidden matching pennies game. PHGD and GDA as presented in Sakos et al. (2024) are compared against GN, DGN, LM, and GD. (left) Linear convergence to the equilbrium is observed for several methods with LM and GD outperforming the rest. (middle) Trajectories for some methods are plotted in both the parameter and prediction space. (right) The loss ratio $\ell_t(\theta_{t+1})/\ell_t(\theta_t)$ is observed for select methods.

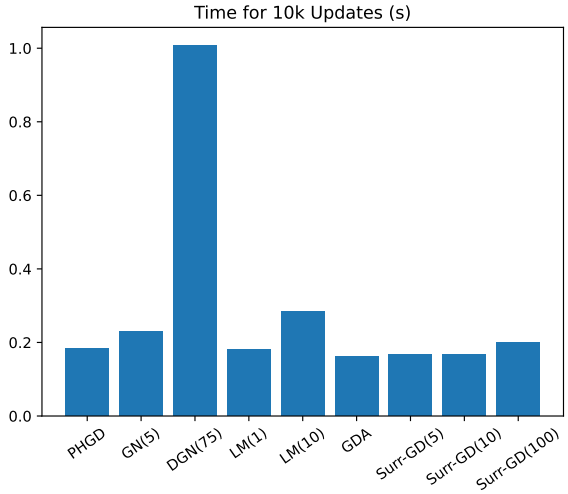

Figure 2: Total runtime in seconds for executing 10,000 updates for different algorithms, where each update may include several inner steps.

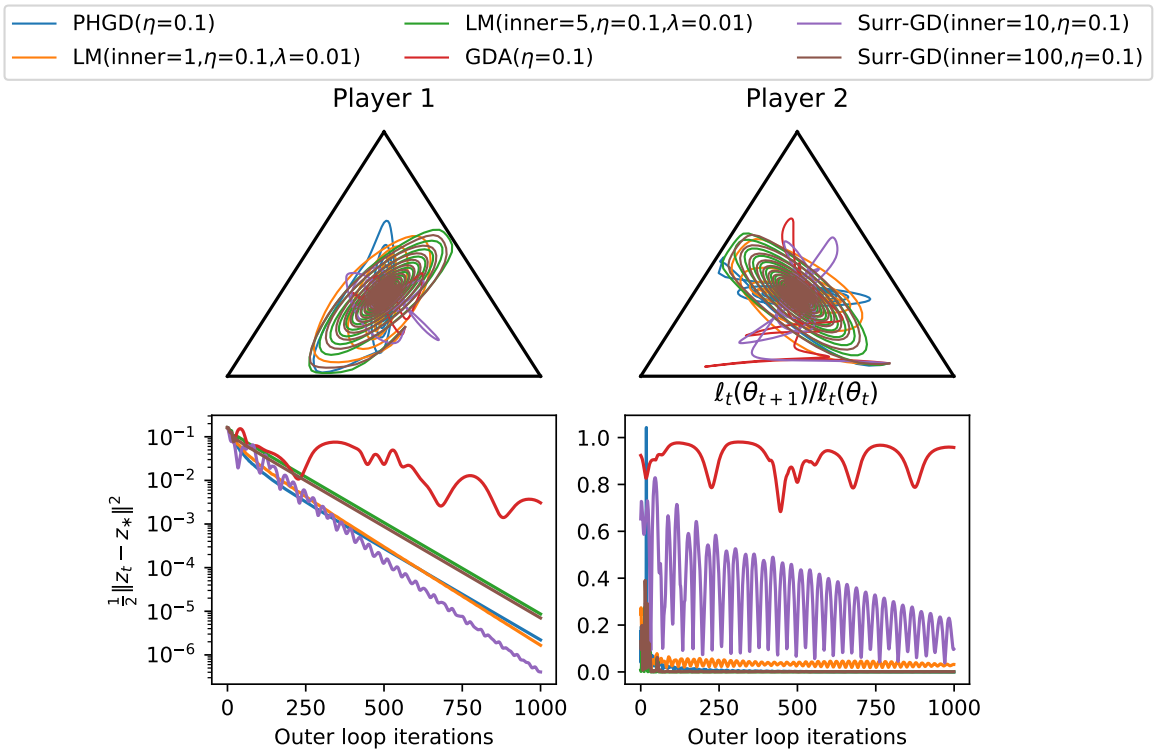

Figure 3: Convergence in the hidden rock-paper-scissors game.

Hidden Matching Pennies. The hidden matching pennies game corresponds to a zero-sum game of the form (3), where each player has the parameterization $h^i(\theta) = \text{sigmoid}(\alpha_2^i \,\text{CELU}(\alpha_1^i \theta))$. The convex-concave objective is given by $f(z^1, z^2) = -(2z^1-1)(2z^2-1) + \frac{0.75}{2}\left((z^1 - 1/2)^2 + (z^2 - 1/2)^2\right)$. The parameters for player i, α_1^i, α_2^i, are chosen to approximately replicate the trajectory of PHGD presented in Sakos et al. (2024, Figure 4). In Figure 1 we observe that PHGD converges with linear convergence in the squared distance to the equilibrium $z_* = (1/2, 1/2)$, however, performs poorly if multiple inner steps are taken (GN with 5 inner steps). If η is increased by an order of magnitude to 0.1 PHGD is observed to diverge, however, convergence is possible via DGN under the same η, with multiple iterations and an appropriate choice of η_{DGN}. In contrast to GN, LM is more stable with a larger η, and converges faster than PHGD/GN. Finally we tested GD for a different number of inner steps with $\eta = 0.1$. Convergence is observed for GDA albeit slow. The benefit of multiple steps is clear, with 10 inner steps outperforming PHGD and only surpassed by LM. Although more inner steps increases the computation cost, it is marginal when compared to the cost of evaluating F(see Figure 2). Interestingly, Figure 1 (right) shows that spending more compute to minimize the surrogate at each iteration does not necessarily translate to faster overall convergence with respect to the outer loop (left). GD with 10 inner steps has a larger loss ratio than GD with 100 steps but converges faster to the equilibrium.

Hidden rock-paper-scissors. In the hidden rps game each player's mixed strategy in rock-paper-scissors is parameterized. Where player i's strategy z^i is given by the function

$h^i(\theta) = \text{softmax}(A_2^i \,\text{CELU}(A_1^i \theta^i))$, with $\theta^i \in \mathbb{R}^5$ and randomly initialized matrices: $A_1^i \in \mathbb{R}^{4\times 5}$, and $A_2^i \in \mathbb{R}^{3\times 4}$. Figure 3 demonstrates the behaviour of various algorithms for a fixed initialization of $\theta = (\theta^1, \theta^2)$ and the matrices A_j^i. We observe that PHGD and LM with one inner step achieve linear convergence while GDA performs poorly with an unstable behaviour. Like in the hidden matching pennies game, increasing the number of inner steps for GD improves stability and performance, with the best performance not necessarily corresponding to the methods with the lowest loss ratio. Both LM and GD degrade in performance if too many inner steps are taken, with one and 10 inner steps outperforming 5 and 100 steps for LM and GD respectively.