Bilevel Optimization for Adversarial Learning Problems: Sharpness, Generation, and Beyond

Risheng Liu[†], Zhu Liu[†], Weihao Mao[‡], Wei Yao[‡§], Jin Zhang[‡§¶*]
[†]School of Software Technology, Dalian University of Technology
[‡]Mathematical Department, Southern University of Science and Technology
[§]National Center for Applied Mathematics Shenzhen
[¶]Detection Institute for Advanced Technology Longhua-Shenzhen (DIATLHSZ)
rsliu@dlut.edu.cn, liuzhu@mail.dlut.edu.cn, mwhaea123456@google.com,
yaow@sustech.edu.cn, zhangj9@sustech.edu.cn

Abstract

Adversarial learning is a widely used paradigm in machine learning, often formulated as a min-max optimization problem where the inner maximization imposes adversarial constraints to guide the outer learner toward more robust solutions. This framework underlies methods such as Sharpness-Aware Minimization (SAM) and Generative Adversarial Networks (GANs). However, traditional gradient-based approaches to such problems often face challenges in balancing accuracy and efficiency due to second-order complexities. In this paper, we propose a bilevel optimization framework that reformulates these adversarial learning problems by leveraging the tractability of the lower-level problem. The bilevel framework introduces no additional complexity and enables the use of advanced bilevel tools. We further develop a provably convergent single-loop stochastic algorithm that effectively balances learning accuracy and computational cost. Extensive experiments show that our method improves generation quality of GANs, and consistently achieves higher accuracy for SAM under label noise and across various backbones, while promoting flatter loss landscapes. Overall, this work provides a practical and theoretically grounded framework for solving adversarial learning tasks through bilevel optimization.

1 Introduction

Nowadays, adversarial learning has attracted considerable attention due to its broad applicability in machine learning, particularly in addressing fairness and robustness. Typically formulated as a min-max optimization problem [46, 80, 16], adversarial learning involves maximizing an inner objective to impose adversarial constraints, thereby guiding the outer learner toward more robust solutions. This framework underpins a wide range of applications. In Generative Adversarial Networks (GANs) [8, 33, 41, 66, 77, 79], a generator (minimizer) and discriminator (maximizer) compete to improve the quality of synthetic data. In Sharpness-Aware Minimization (SAM) [2, 68, 69], adversarial perturbations on model parameters encourage optimization in flatter loss regions. In adversarial training [62, 46, 1, 80], input-space attacks are maximized to enhance model robustness against perturbations. These examples highlight the foundational role of adversarial learning across various domains.

Despite the widespread adoption and their empirical success, the traditional min-max framework introduces significant challenges for scalability, primarily because each training step involves

[*]Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

multiple nested maximization steps. This results in prohibitively high computational costs, rendering these methods impractical for many real-world applications and further understandings [32, 80, 61, 64, 82].

Among the various applications of adversarial learning, we use Sharpness-Aware Minimization (SAM) as a representative example to highlight the challenges inherent in the min-max formulation and to motivate the need for a new reformulation.

Sharpness-Aware Minimization. Due to the influence of noise and overparameterization, modern neural networks are often prone to overfitting, resulting in poor generalization performance [78, 58, 68]. This challenge remains a significant barrier to further progress in machine learning. Recent numerical and theoretical studies have revealed a strong correlation between a model's generalization ability and the sharpness of the loss function landscape around its optimized parameters [28, 12, 14, 50, 22]. Classical optimization methods, which aim solely to minimize the value of the loss function, often fail to account for this geometric aspect and are thus insufficient for improving generalization.

To overcome this, researchers have turned their attention to the geometry of the loss surface, particularly the sharpness near minimizers, leading to the formulation of a min-max problem. This direction has culminated in the development of *Sharpness-Aware Minimization* (SAM), which explicitly incorporates sharpness into the optimization process [71, 83, 16, 51]. Recent work by [59] further suggests that SAMs effectiveness lies in its ability to balance the quality of diverse features, rather than merely seeking flatter minima.

Difficulties in SAM: Min-Max Perspective. Despite the rapid development and empirical success of SAM, algorithms based on SAMs original min-max formulation require implicit differentiation, which is computationally demanding. To reduce cost, practical implementations often avoid this step, but at the expense of accuracy, even with various refinements [16, 13, 5, 68]. For example, [72] incorporated second-order information to improve efficiency. Most theoretical analyses, however, still rely on the full SAM formulation without simplifications [5, 68, 69, 11]. For instance, [16] derived a PAC-Bayes bound for an averaged direction model, while [68] showed SAMs local regularization effect on the Hessian eigenvalues. These results point to a gap between simplified practical algorithms and the theoretical models that justify them.

Difficulties in Adversarial Learning: the Min-Max Perspective. The challenges encountered in SAM are not isolated but reflect a broader set of issues faced in many adversarial learning tasks. Similar difficulties appear in fast adversarial training [80, 1], generative adversarial learning [7, 64, 61, 82]. These methods frequently exhibit the following limitations [32, 80]: (i) reduced computational cost at the expense of learning accuracy, and (ii) a lack of sufficient theoretical guarantees.

These recurring issues raise a fundamental question:

Is there an efficient framework that balances computational cost and learning accuracy in such adversarial learning tasks?

To address this, we propose a unified bilevel optimization perspective, where the lower-level problem is chosen to be simple and, in many cases, admits a closed-form solution. This design allows us to apply advanced bilevel optimization theory and algorithms without increasing computational burden, as the lower-level remains analytically tractable. As a result, the proposed framework offers a principled and efficient approach applicable to various tasks, including SAM and generative adversarial networks, as discussed in this paper.

Contributions. We summarize our contributions as follows:

Formulation-wise. We propose a unified bilevel optimization framework to reformulate a range of adversarial learning tasks, including SAM and GANs, by introducing an efficient lower-level problem. This formulation enables the use of advanced bilevel optimization algorithms, achieving a better balance between computational efficiency and learning accuracy.

Experiment-wise. We conduct comprehensive experiments to evaluate the proposed bilevel framework, focusing on Sharpness-Aware Minimization (SAM) and generative adversarial learning (GAN). The results demonstrate that our approach effectively balances computational cost and numerical accuracy, highlighting its practicality and robustness in real-world adversarial learning tasks.

2 Related work

Examples from the Traditional Perspective. In recent years, mathematical formulations and corresponding optimization algorithms have played an increasingly prominent role in adversarial learning. Prominent examples of min-max formulations include adversarial training [78, 70, 1] and generative adversarial learning [55, 24, 49, 27, 26, 48, 25, 9, 23, 76]. While these approaches often reduce computational complexity, they tend to compromise model accuracy and robustness.

As an illustrate example, [16] leveraged the relationship between flat minima and generalization error to train deep neural networks (DNNs) with improved generalization across natural distributions using SAM. Similarly, [71] proposed a method that regularizes the sharpness term in adversarial training, achieving significantly more robust generalization against adversarial attacks. A substantial body of research has explored combining SAM with other training strategies or neural architectures [10, 67, 63]. For instance, [31] improved SAM by scaling the sharpness adjustment relative to the parameter size, while [42] optimized computational efficiency by reusing previously computed weight perturbations. Numerous efforts have been made to refine and adapt the SAM algorithm. For example, [13] introduced the Efficient Sharpness-Aware Minimizer (ESAM), which incorporates two key training strategies: Stochastic Weight Perturbation (SWP) and Sharpness-Sensitive Data Selection (SDS). [52] provides a unified framework of analysis on SAM.

Bilevel Optimization. Bilevel optimization is a hierarchical framework involving an upper-level and a lower-level problem. It is a highly challenging yet impactful area in both theory and practice. This framework has led to significant progress in machine learning domains such as hyperparameter selection [53, 17, 45] and meta-learning [6, 30, 56, 17, 36, 21, 75, 81]. [74] proposed a bilevel algorithm for the 0-1 classification problem, and [80] introduced a bilevel reformulation of fast adversarial training originally modeled as a traditional min-max problem in [46]. In this work, we take the step toward establishing a bilevel perspective for adversarial learning tasks, and propose corresponding algorithms and theoretical analyses to overcome the limitations of traditional perspectives.

3 The Bilevel Perspective Beyond the Traditional Framework

In this section, we first revisit the traditional optimization approach commonly used in adversarial learning tasks, using Sharpness-Aware Minimization (SAM) as a representative example, and point out their key limitations. We then introduce a bilevel optimization perspective featuring an efficiently solvable lower-level problem. This reformulation enables the application of advanced bilevel methods without incurring additional computational complexity, offering a principled and practical framework for a broad class of adversarial learning problems.

3.1 Preliminary on The Traditional Perspective: Min-Max Formulations

We begin by introducing a widely used min-max optimization formulation that underlies several adversarial learning applications, such as adversarial training [78, 70, 1], generative adversarial learning [55, 24, 49, 27, 26, 48, 25, 9, 23, 76], and, as discussed in the next subsection, SAM.

Consider the standard min-max formulation for many adversarial learning tasks:

$$\min_{\omega} \max_{\delta \in \mathcal{C}} L(\omega, \delta), \tag{1}$$

where the parameter $\omega \in \mathbb{R}^n$, $\delta \in \mathcal{C} \subset \mathbb{R}^m$, $\mathcal{C}$ represents the regular closed constraint set. Let $\mathcal{S} = \bigcup_{i=1}^{N} \{(u_i, v_i)\}$ represents the training dataset sampled from the data space $\mathcal{U} \times \mathcal{V}$. The function $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the empirical loss defined as

$$L(\omega, \delta) = \frac{1}{N} \sum_{i=1}^{N} L_S(\omega, \delta, u_i, v_i),$$

where $L_S : \mathbb{R}^n \times \mathbb{R}^m \times \mathcal{U} \times \mathcal{V} \to \mathbb{R}_+$ is the per-sample loss. The goal is to learn optimal parameters ω and δ that solve the $\min\text{-}\max$ objective (1).

The loss function $L(\omega, \delta)$ is often nonconvex with respect to both ω and δ. This nonconvexity implies that local or global minimizers with identical loss values may exhibit vastly different neighborhood behaviors and theoretical capabilities.

An example of the computational challenges in $\mathrm{min\text{-}max}$ problems is provided by [73], where 128 GPUs were used to run Adversarial Training (AT) on the ImageNet dataset. Such computational requirements highlight the prohibitive costs of solving large-scale $\mathrm{min\text{-}max}$ problems directly. Several computational techniques have been proposed to improve the efficiency of $\mathrm{min\text{-}max}$ optimization. For instance, gradient alignment (GA) regularization [1] and fast adversarial training methods [70] have been developed to address these challenges. However, these methods often suffer from practical issues such as instability, catastrophic overfitting, and degraded robustness [32, 80]. These limitations underscore the need for more robust and scalable approaches, such as the bilevel perspective.

3.2 SAM: An Example of The Traditional Perspective

In this subsection, we present an example of Sharpness-Aware Minimization (SAM) [16], which is based on the observation that the generalization ability of a model is closely related to the sharpness of its loss function landscape.

Consider a local neighborhood around a parameter ω with radius $r > 0$ and a unit direction δ. SAM introduces the following min-max formulation:

$$\min_{\omega} \left[L(\omega) + \max_{\|\delta\| \leq 1} [L(\omega + r\delta) - L(\omega)] \right], \tag{2}$$

which can be equivalently written as:

$$\min_{\omega} \max_{\|\delta\| \leq 1} L(\omega + r\delta), \tag{3}$$

where $\mathcal{C} = \{\delta \in \mathbb{R}^n : \|\delta\| \leq 1\}$ is the set of feasible perturbations, and r denotes the radius. This formulation augments the original loss function $L(\omega)$ by incorporating a sharpness term $L(\omega + r\delta) - L(\omega)$, which captures the worst-case increase in the loss within the neighborhood defined by $r\delta$. The goal of SAM is thus twofold: minimize the loss function while simultaneously reducing sharpness in the worst-case direction.

The common strategy to address these computational challenges is to adopt approximation techniques. In the context of (2), the maximum function $L_{\max} : \mathbb{R}^n \to \mathbb{R}$ is defined as:

$$L_{\max}(\omega) := \max_{\|\delta\| \leq 1} L(\omega + r\delta), \tag{4}$$

where r denotes the neighborhood radius. The traditional SAM algorithm minimizes $L_{\max}(\omega)$ by employing an approximation technique and a subsequent *discard* process.

Numerical Technique The approximation begins with a first-order Taylor expansion around ω:

$$L(\omega + r\delta) \approx L(\omega) + r\delta^\top \nabla_\omega L(\omega). \tag{5}$$

Using this approximation, the maximizer $\delta^*(\omega)$ is computed as:

$$\delta^*(\omega) = \arg \max_{\|\delta\| \leq 1} \left[L(\omega) + r\delta^\top \nabla_\omega L(\omega) \right] = \frac{\nabla_\omega L(\omega)}{\|\nabla_\omega L(\omega)\|}.$$

Substituting $\delta^*(\omega)$ back into the SAM formulation, it becomes:

$$\min_{\omega} L_{\mathrm{ASC}}(\omega) := L\left(\omega + r \frac{\nabla_\omega L(\omega)}{\|\nabla_\omega L(\omega)\|} \right). \tag{6}$$

To efficiently compute the gradient, traditional methods calculate:

$$\begin{aligned} \nabla_\omega L\big(\omega + r\delta^*(\omega)\big) &= \frac{d\big(\omega + r\delta^*(\omega)\big)}{d\omega} \nabla_\omega L\big(\omega + r\delta^*(\omega)\big) \\ &= \nabla_\omega L\big(\omega + r\delta^*(\omega)\big) + r\frac{d\delta^*(\omega)}{d\omega} \nabla_\omega L\big(\omega + r\delta^*(\omega)\big). \end{aligned} \tag{7}$$

The second term, $r\frac{d\delta^*(\omega)}{d\omega}\nabla_\omega L\big(\omega + r\delta^*(\omega)\big)$, involves the derivative of the composition function $\delta^*(\omega)$, which is computationally expensive to evaluate. Therefore, traditional methods often discard this term to simplify the gradient computation [16, 68, 29]. Thus, the standard iterative process can be expressed as:

$$\omega^{k+1} = \omega^k - t\nabla L\left(\omega^k + r\frac{\nabla L(\omega^k)}{\|\nabla L(\omega^k)\|}\right), \tag{8}$$

where $t > 0$ is the step size. Convergence analysis for this iterative scheme and its extensions has been extensively studied (e.g., [16, 29, 52]).

Importance of Hessian. To better understand the behavior of SAM, consider Table 1. In the table, λ_1 denotes the largest eigenvalue, $\lambda_{\min}$ the smallest non-zero eigenvalue, Tr the trace operator, and $H_L(\omega)$ the Hessian of the loss function L at ω.

Table 1: Definitions and biases of different SAM loss formulations [68].

Type of Loss	Definition	Biases
Worst	$\max_{\|\delta\|\le 1} L(\omega + r\delta)$	$\min_{\omega} \lambda_1(H_L(\omega))$
Ascent	$L\left(\omega + r\frac{\nabla L(\omega)}{\|\nabla L(\omega)\|}\right)$	$\min_{\omega} \lambda_{\min}(H_L(\omega))$
Average	$\mathbb{E}_{(u,v)\sim\mathcal{D}}[L_S(\omega, u, v)]$	$\min_{\omega} \mathrm{Tr}(H_L(\omega))$

In Table 1, the *worst-direction loss* captures the sharpest curvature of the loss landscape and is associated with λ_1. The *ascent-direction loss* reflects the local curvature along the most favorable direction of increase and corresponds to $\lambda_{\min}$. Finally, the *average-direction loss* characterizes the overall sharpness of the loss surface and is measured by $\mathrm{Tr}(H_L(\omega))$. From this table, it is clear that the effectiveness of the SAM technique in minimizing sharpness arises from its incorporation of second-order information via the Hessian matrix $H_L(\omega)$. Specifically, the largest eigenvalue $\lambda_1(H_L(\omega))$ directly measures the sharpness of the loss function. This connection explains why SAM improves generalization by reducing sharpness. In fact, prior studies (e.g., [44, 4, 43]) have used $\lambda_1(H_L(\omega))$ as a key metric to evaluate SAM's efficiency in enhancing generalization performance.

Loss of Accuracy. While the discarding step reduces computational costs, it introduces inaccuracies that may compromise learning performance. Specifically, in (7), we have

$$\frac{d\delta^*(\omega)}{d\omega} = \nabla_\omega\left(\frac{\nabla L(\omega)}{\|\nabla L(\omega)\|}\right) = \frac{1}{\|\nabla L(\omega)\|}\left(H_L(\omega) - \frac{\nabla L(\omega)\nabla L(\omega)^\top H_L(\omega)}{\|\nabla L(\omega)\|^2}\right), \tag{9}$$

where $H_L(\omega)$ denotes the Hessian of the loss function L (details are in Subsection 3.1). The term discarded in (7) contains valuable geometric information about the loss landscape, which plays a critical role in establishing theoretical guarantees and ensuring robust optimization [68, 69]. This discrepancy between the theoretical formulation and its practical implementation highlights the need for a more principled and efficient framework. Experiments in Subsection 5.1 also shows that its loses accuracy.

3.3 The Bilevel Perspective

We introduce the bilevel perspective to address the limitations of traditional approaches, offering a simple lower-level problem without introducing additional complexities.

Bilevel Perspective. The bilevel optimization framework provides a natural and effective solution to the challenges in traditional perspectives. This formulation facilitates the application of advanced methods, such as those based on the Moreau envelope. Crucially, the lower-level problem remains simple and tractable, avoiding unnecessary complexity. This structure allows for the preservation of learning accuracy while managing computational costs, making it a robust alternative to traditional formulations.

Specifically, by selecting a suitable lower-level function $L_\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, we reformulate traditional models - such as the min-max problem (1) – into a bilevel optimization problem:

$$\begin{aligned} &\min_{\omega} L(\omega, \tilde{\delta}) \\ &\text{subject to } \tilde{\delta} = \arg\min_{\delta\in\mathcal{C}}\{L_\ell(\omega, \delta)\}, \end{aligned} \tag{10}$$

where the lower-level problem allows efficient computation of the optimal solution $\tilde{\delta}$ for a given ω, and $\mathcal{C}$ denotes the feasible set for δ.

As an example, consider the SAM formulation in (2) and (6). By setting $L_\ell(\omega, \delta) := -\delta^T \nabla L(\omega)$, the lower-level problem becomes a linear program over the unit ball $\mathcal{C}$. This admits a closed-form solution: $\tilde{\delta} = \frac{\nabla L(\omega)}{\|\nabla L(\omega)\|}$. Hence, this bilevel reformulation introduces no additional complexity compared to the original model. In this case, the bilevel formulation (10) is equivalent to the traditional SAM model (2). The bilevel perspective allows both theoretical analysis and numerical optimization tools from bilevel programming to be applied effectively to such tasks.

4 Algorithm and Theoretical Investigation

Given the bilevel perspective, we now investigate algorithms tailored for this type of bilevel optimization. Solving large-scale bilevel optimization (BLO) problems for complex learning tasks presents two significant challenges: ensuring computational efficiency and guarantees of accuracy. To address these challenges, we apply a stochastic single-loop algorithm adopted from [40], inspired by [18], based on the Moreau envelope.

Value Function. One of the main challenges in solving bilevel programming problems (BLPPs) lies in their nested structure: one must first solve the lower-level problem and then optimize the upper-level objective based on that solution.

Let the value function be defined as

$$V(\omega) := \min_{\delta \in \mathcal{C}} L_\ell(\omega, \delta).$$

Then, the bilevel problem (10) can be equivalently reformulated as:

$$\begin{aligned} &\min_{\omega} \; L(\omega, \delta) \\ &\text{subject to} \quad L_\ell(\omega, \delta) - V(\omega) \leq 0, \quad \delta \in \mathcal{C}. \end{aligned} \tag{11}$$

However, the nonsmooth and nonconvex nature of $V(\omega)$ poses significant computational challenges.

Moreau Envelope. To address this issue, we reformulate the problem using the Moreau envelope of $L_\ell(\omega, \delta)$ with parameter $\gamma > 0$:

$$v_\gamma(\omega, \delta) := \inf_{\theta \in \mathcal{C}} \left\{ L_\ell(\omega, \theta) + \frac{1}{2\gamma} \|\theta - \delta\|^2 \right\}. \tag{12}$$

The Moreau envelope provides a smooth approximation to the value function. Geometrically, it defines a family of smooth surfaces that approximate the graph of V. Using this, we reformulate problem (10) as:

$$\begin{aligned} &\min_{\omega} \; L(\omega, \delta) \\ &\text{subject to} \quad L_\ell(\omega, \delta) - v_\gamma(\omega, \delta) \leq 0, \quad \delta \in \mathcal{C}. \end{aligned} \tag{13}$$

If $L_\ell(\omega, \cdot)$ is convex in δ for each fixed ω, then (10) and (13) are equivalent [18, Theorem 2.1]. If $L_\ell(\omega, \cdot)$ is only weakly convex, then the latter is a relaxation of the former [40, Theorem A.1].

Note that $L_\ell(\omega, \delta) - v_\gamma(\omega, \delta) \geq 0$ for all $(\omega, \delta) \in \mathbb{R}^n \times \mathcal{C}$, so the constraint is effectively an equality constraint. This violates standard constraint qualifications and complicates the derivation of optimality conditions.

To circumvent this, we introduce a penalty parameter $\mu > 0$ and consider the penalized problem:

$$\min_{\omega} \; F_\mu(\omega, \delta) := L(\omega, \delta) + \mu \left[L_\ell(\omega, \delta) - v_\gamma(\omega, \delta) \right], \quad \text{subject to} \quad \delta \in \mathcal{C}. \tag{14}$$

Algorithm. Based on the Moreau envelope, we propose a single-loop stochastic algorithm. At iteration k, given a mini-batch $b_k \subset \{1, \ldots, N\}$, define:

$$L^{b_k}(\omega, \delta) := \frac{1}{|b_k|} \sum_{i \in b_k} L_S(\omega, \delta, u_i, v_i), \tag{15}$$

where L_S denotes the sample-wise loss. The batch-based lower-level loss is denoted as $L_\ell^{b_k}(\omega, \delta)$. The update is computed using gradients of $L^{b_k}(\omega, \delta)$, as described in (18).

Unlike SAM [17, 40] or fast adversarial training [80], our method avoids computing gradients of composite functions, thus significantly reducing computational cost while preserving robustness and accuracy.

Convergence Analysis. We define the residual function using the selected mini-batch b_k:

$$R_k^{b_k}(\omega, \delta) := \text{dist}\left(0, \nabla L^{b_k}(\omega, \delta) + \mu_k \left(\nabla L_\ell^{b_k}(\omega, \delta) - \nabla v_\gamma^{b_k}(\omega, \delta)\right) + N_{\mathbb{R}^n \times \mathcal{C}}(\omega, \delta)\right),$$

where $N_{\mathbb{R}^n \times \mathcal{C}}(\omega, \delta)$ is the normal cone to $\mathbb{R}^n \times \mathcal{C}$ at (ω, δ).

When $b_k = \{1, \ldots, N\}$, denote

$$R_k(\omega, \delta) := \text{dist}\left(0, \nabla L(\omega, \delta) + \mu_k \left(\nabla L_\ell(\omega, \delta) - \nabla v_\gamma(\omega, \delta)\right) + N_{\mathbb{R}^n \times \mathcal{C}}(\omega, \delta)\right).$$

This residual serves as a stationarity measure for the penalized problem:

$$\min_{\omega} F_{\mu_k}(\omega, \delta) := L(\omega, \delta) + \mu_k \left[L_\ell(\omega, \delta) - v_\gamma(\omega, \delta)\right], \quad \text{subject to} \quad \delta \in \mathcal{C}. \tag{16}$$

A point (ω, δ) is stationary for (16) if and only if $R_k(\omega, \delta) = 0$.

Special Case: SAM. In the SAM setting where $L(\omega, \delta) = L(\omega + r\delta)$ and $L_\ell(\omega, \delta) = -\delta^\top \nabla_\omega L(\omega)$, the lower-level solution is given by:

$$\delta = \theta_\gamma^*(\omega, \delta) = \frac{\nabla L(\omega)}{\|\nabla L(\omega)\|}.$$

Then the gradient of the penalized objective becomes:

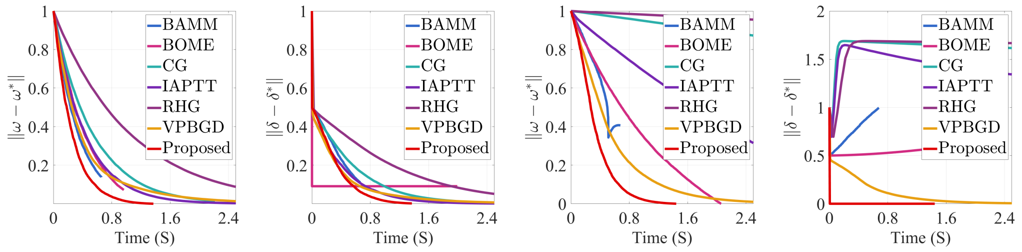
$$\nabla F_\mu(\omega, \delta) = \left(\nabla L(\omega + r\delta),\ r\nabla L(\omega + r\delta) - \mu \nabla L(\omega)\right).$$

Thus $R_k(\omega, \delta)$ captures the sharpness in the ascent direction model (6). We then present the following convergence result for the proposed algorithm:

Theorem 4.1. *Assume the algorithm satisfies some standard conditions. Then for any $p \in (0, \frac{1}{2})$, we have:*

$$\mathbb{E}\left[\min_{0 \le k \le K} R_k^{b_k}(\omega^{k+1}, \delta^{k+1})\right] = O\left(\frac{1}{K^{(1-2p)/2}}\right). \tag{17}$$

Figure 1 compares the convergence performance of our algorithm with existing bilevel methods, demonstrating its superior efficiency.



(a) $\|\omega - \omega^*\|$ (b) $\|\delta - \delta^*\|$ (c) $\|\omega - \omega^*\|$ (d) $\|\delta - \delta^*\|$

Figure 1: Illustration of convergence behaviors compared with representative BLO schemes.

Future Generalization. For various types of min-max problems, modern methods have been proposed (e.g., [1, 2, 32, 80, 65]). For instance, [65] investigates min-max problems of the form $\min_\theta \max_\phi L(F(\theta), G(\phi))$ in the convex-concave setting. Our method can be applied to some of these problems, particularly when the inner concave subproblem $\max_\phi L(F(\theta), G(\phi))$ is tractable, but solving the overall min-max problem requires sacrificing learning accuracy for computational efficiency. Some of these problems suffer from cyclic behavior of level sets (see Figure 4 in [65]). Unlike the approach in [65], we address this issue through a bilevel reformulation, whose convergence is established in Theorem A.8. In such cases, our bilevel framework provides a compelling alternative, offering a more favorable balance between learning accuracy and computational cost.

5 Experiments

In this section, we validate the performance under two real-world learning applications, including generative adversarial network and sharpness-ware minimization. Then we demonstrate the superiority of proposed scheme, illustrating the convergence behaviors and computation efficiency on synthetic numerical problems compared with existing BLO schemes. Details of implementation configurations and parameters selection are provided in the Appendix B. The source codes will be released at `https://github.com/LiuZhu-CV/BLOAL`.

Table 2: Evaluation of the robustness for SAM under varying noise labels and diverse backbones.

Noise Label	SGD	SAM	Ours	Backbone	SGD	SAM	Ours
Clean	95.47 ± 0.12	96.27±0.03	**96.34±0.11**	ResNet34	95.58 ± 0.15	96.61 ± 0.08	**97.75 ± 0.06**
10%	89.58 ± 0.19	92.84±0.55	**93.41±0.18**	ResNet50	95.40 ± 0.44	96.48 ± 0.05	**96.50 ± 0.09**
20%	82.64 ± 0.31	90.80±0.73	**91.48±0.06**	ResNet101	95.66 ± 0.18	96.60 ± 0.13	**96.69 ± 0.05**

Table 3: Evaluation of the robustness based on sharpness metrics and diverse perturbation rates.

Sharpness Metric	SGD	SAM	Ours	Perturbation	SAM	Ours
Hessian Norm	8.869 ± 3.31	4.412±1.84	**3.840±1.63**	r=0.05	96.22 ± 0.13	**96.23 ± 0.10**
Trace	13756 ± 1710	6650±1828	**5023±1101**	r=0.15	96.25 ± 0.05	**96.38 ± 0.26**
Top Eigenvalues	1048.347	621.349	**520.889**	r=0.2	96.26 ± 0.05	**96.28 ± 0.17**

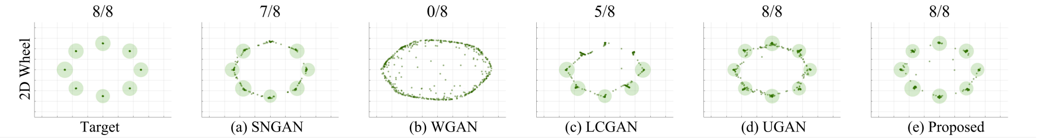


Figure 2: Comparison of generative adversarial learning under synthesized 2D wheels.

5.1 Real-world Applications

Generative Adversarial Learning. We conduct comparison with Stacked MNIST, a challenging dataset with 1000 modes and two-dimensional simulation experiments based on Gaussian distribution, generating eight distribution of 2D wheels. Table 4 reports the numerical performances compared with specialized methods, in terms of KL divergence and the maximum of modes. The visualization of data generation is shown in Figure 2. It can be observed that the proposed scheme can effectively generate all modes, compared with SNGAN [49], WGAN [3], LCGAN [15] and UGAN [47].

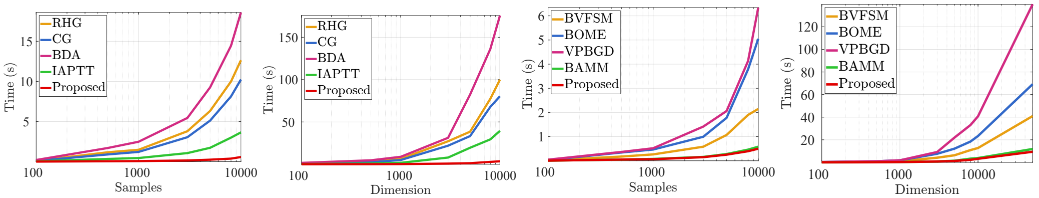
Table 4: Performance comparison of GAN on StackedMNIST.

Methods	Modes	D_{KL}
DcGAN [54]	99	3.40
VEEGAN [60]	150	2.95
WGAN [3]	959	0.73
PacGAN [34]	992	0.28
R3GAN [20]	1000	0.12
Ours	1000	0.08

Sharpness-ware Minimization. As aforementioned, we propose a BLO perspective for SAM with computation accuracy and efficiency. Table 2 illustrates the robustness of different optimizer under varying levels of label noise and backbones on the Cifar-10 benchmark. Compared to standard SGD and SAM, our method consistently achieves superior accuracy across all noise labels and backbone, demonstrating its adaptability and robustness. Table 3 presents the analyses of model robustness based on sharpness-aware metrics, specifically analyzing the Hessian properties and the impact of varying perturbation rates r. The Hessian Norm quantifies the overall curvature of the loss landscape. The trace of the Hessian matrix reflects the sum of its eigenvalues. Lower top eigenvalues correspond to smoother loss surfaces. These metrics indicates our method encourages convergence toward flatter, more robust regions in the parameter space. Furthermore, Ours achieves consistently higher accuracy than SAM, demonstrating the practical robustness against diverse perturbation radius.

5.2 Synthetic Numerical Evaluation

General Min-Max Numerical Cases. Here, we compare the convergence behavior of our method with representative bilevel optimization (BLO) schemes, including Hessian-based algorithms (*e.g.*, RHG [17], CG [53], IAPTT [37]), first-order methods (*e.g.*, BOME [35], VPBGD [57]), and the single-loop approach BAMM [39].



(a) Comparison with Hessian-based algorithms (b) Comparison with Hessian-free algorithms

Figure 3: Computation efficiency comparison as the dimension of samples and features increase.

The first min-max numerical example is $\min_{\omega\in\mathbb{R}}\max_{\delta\in\mathbb{R}}\omega^2-\delta^2+\alpha\sin(\omega\delta)$, where α controls the nonlinearity. This problem features bilinear coupling and oscillatory components, resulting in multiple saddle points that can hinder convergence. For simplicity, we set $\alpha = 1$ and constrain $\omega, \delta \in [0, 1]$, with the optimal solution at $(0, 0)$. Figure 1 (a) and (b) show the convergence of various BLO methods in terms of $\|\omega-\omega^*\|$ and $\|\delta-\delta^*\|$. Our method achieves the fastest convergence across both metrics, while BOME and VPBGD exhibit slower decay and higher residuals. We also consider a more challenging case: $\min_\omega\max_\delta\omega^T A\delta+\sin(\|\omega\|^2-\|\delta\|^2)$, where $\omega, \delta, A \in \mathbb{R}$. The nonlinear squared coupling introduces varying oscillation frequencies and degenerate critical points. With $\omega, \delta \in [0, 1]$ and $A = 1$, the solution remains $(0, 0)$. Figure 1 (c) and (d) illustrate that our method maintains superior convergence speed and stability, while other approaches, including BAMM, BOME, CG, and VPBGD, suffer from oscillations in the lower-level updates.

Large-scale Computation Efficiency. We construct a synthetic numerical case to perform the large-scale computation, which is formulated as $\min_\omega\max_\delta\sum_{i=1}^m\sum_{j=1}^n(\omega_{i,j}^2-\delta_{i,j}^2+\sin(\omega_{i,j}\delta_{i,j}))$, where m and n denote the numbers of samples and dimensions. Figure 3 illustrates the computation time of single step under varying numbers of samples and feature dimensions. Compared with Hessian-free algorithms, value function based schemes (*e.g.,* BVFSM [38], BOME and VPBGD) suffer from steep computational increases, especially with larger feature dimensions. Moreover, compared with single-loop scheme BAMM, the proposed method remains highly efficient, improving 16.8% at 10000 samples.

6 Conclusion

This paper presents a unified bilevel optimization framework for solving adversarial learning tasks, aiming to balance computational efficiency and learning accuracy. Motivated by the limitations of traditional methods, such as high cost and limited theoretical grounding, our approach offers an efficient and interpretable alternative. In particular, it effectively addresses Sharpness-Aware Minimization (SAM) and generative adversarial learning tasks. Extensive experiments show that our method captures all data modes in generative modeling and improves FID and JS scores. For SAM, it consistently achieves higher accuracy under label noise and across different backbones, while encouraging flatter loss landscapes. Overall, this work provides a practical and theoretically sound foundation for addressing adversarial learning tasks via bilevel optimization.

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China (Nos.62450072, U22B2052, 624B2033, 12371305, 12222106, 12326605), the Distinguished Youth Funds of the Liaoning Natural Science Foundation (No.2025JH6/101100001), the Distinguished Young Scholars Funds of Dalian (No.2024RJ002), the Fundamental Research Funds for the Central Universities, Guangdong Basic and Applied Basic Research Foundation (No.2022B1515020082) and the Longhua District Science and Innovation Commission Project Grants of Shenzhen (No.20250113G43468522). We thank the anonymous reviewers for their valuable comments and constructive suggestions on this work. Authors listed in alphabetical order.

References

[1] Andriushchenko, M. and Flammarion, N. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

[2] Andriushchenko, M. and Flammarion, N. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, pp. 639–668, 2022.

[3] Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223, 2017.

[4] Arora, S., Li, Z., and Panigrahi, A. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pp. 948–1024, 2022.

[5] Bartlett, P. L., Long, P. M., and Bousquet, O. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *Journal of Machine Learning Research*, 24:1–36, 2023.

[6] Bennett, K. P., Kunapuli, G., Hu, J., and Pang, J.-S. Bilevel optimization and machine learning. In *IEEE world congress on computational intelligence*, pp. 25–47. Springer, 2008.

[7] Cai, Q., Hong, M., Chen, Y., and Wang, Z. On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*, 2019.

[8] Chan, E. R., Monteiro, M., Kellnhofer, P., Wu, J., and Wetzstein, G. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5799 – 5809, 2021.

[9] Chen, T., Cheng, Y., Gan, Z., Liu, J., and Wang, Z. Data-efficient gan training beyond (just) augmentations: A lottery ticket perspective. *Conference on Neural Information Processing Systems*, pp. 20941–20955, 2021.

[10] Chen, X., Hsieh, C.-J., and Gong, B. When vision transformers outperform resnets without pre-training or strong data augmentations. In *International Conference on Learning Representations*, 2022.

[11] Chen, Z., Zhang, J., Kou, Y., Chen, X., Hsieh, C.-J., and Gu, Q. Why does sharpness-aware minimization generalize better than sgd? In *Advances in Neural Information Processing Systems*, volume 36, pp. 72325–72376, 2023.

[12] Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1019–1028, 2017.

[13] Du, J., Yan, H., Feng, J., Zhou, J., Zhen, L., Goh, R., and Tan, V. Efficient sharpness-aware minimization for improved training of neural networks. In *Proceedings of International Conference on Learning Representations*, 2022.

[14] Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Uncertainty in Artificial Intelligence*, 2017.

[15] Engel, J., Hoffman, M., and Roberts, A. Latent constraints: Learning to generate conditionally from unconditional generative models. In *International Conference on Learning Representations*, 2018.

[16] Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[17] Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pp. 1568–1577, 2018.

[18] Gao, L. L., Ye, J. J., Yin, H., Zeng, S., and Zhang, J. Moreau envelope based difference-of-weakly-convex reformulation and algorithm for bilevel programs. *arXiv preprint arXiv:2306.16761*, 2023.

[19] Huang, F. Optimal hessian/jacobian-free nonconvex-pl bilevel optimization. In *International Conference on Machine Learning*, 2024.

[20] Huang, N., Gokaslan, A., Kuleshov, V., and Tompkin, J. The gan is dead; long live the gan! a modern gan baseline. *Advances in Neural Information Processing Systems*, 37:44177–44215, 2024.

[21] Ji, K., Lee, J. D., Liang, Y., and Poor, H. V. Convergence of meta-learning with task-specific adaptation over partial parameters. In *Conference on Neural Information Processing Systems*, 2020.

[22] Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

[23] Jiang, Y., Chang, S., and Wang, Z. Transgan: Two transformers can make one strong gan. In *Conference on Neural Information Processing Systems*, 2021.

[24] Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[25] Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[26] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. Training generative adversarial networks with limited data. In *Conference on Neural Information Processing Systems*, 2020.

[27] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[28] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[29] Khanh, P. D., Luong, H.-C., Mordukhovich, B., and Tran, D. B. Fundamental convergence analysis of sharpness-aware minimization. In *Conference on Neural Information Processing Systems*, 2024.

[30] Kunapuli, G., Bennett, K. P., Hu, J., and Pang, J. S. Classification model selection via bilevel programming. *Optimization Methods and Software*, 23(4):475–489, 2008.

[31] Kwon, J., Kim, J., Park, H., and Choi, I. K. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914, 2021.

[32] Li, B., Wang, S., Jana, S., and Carin, L. Towards understanding fast adversarial training. *arXiv preprint arXiv:2006.03089*, 2020.

[33] Li, R., Wang, C., Wang, J., Liu, G., Zhang, H.-Y., Zeng, B., and Liu, S. Uphdr-gan: Generative adversarial network for high dynamic range imaging with unpaired data. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7532–7546, 2022.

[34] Lin, Z., Khetan, A., Fanti, G., and Oh, S. Pacgan: The power of two samples in generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

[35] Liu, B., Ye, M., Wright, S., Stone, P., and Liu, Q. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in neural information processing systems*, 35:17248–17262, 2022.

[36] Liu, R., Mu, P., Yuan, X., Zeng, S., and Zhang, J. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International Conference on Machine Learning*, pp. 6305–6315, 2020.

[37] Liu, R., Liu, Y., Zeng, S., and Zhang, J. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34: 8662–8675, 2021.

[38] Liu, R., Liu, X., Zeng, S., Zhang, J., and Zhang, Y. Value-function-based sequential minimization for bi-level optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15930–15948, 2023.

[39] Liu, R., Liu, Y., Yao, W., Zeng, S., and Zhang, J. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. In *International Conference on Machine Learning*, pp. 21839–21866, 2023.

[40] Liu, R., Liu, Z., Yao, W., Zeng, S., and Zhang, J. Rethinking moreau envelope for nonconvex bi-level optimization: A single-loop and hessian-free solution strategy. In *International Conference on Learning Representations*, 2024.

[41] Liu, S., Li, S., and Cheng, H. Towards an end-to-end visual-to-raw-audio generation with gan. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1299–1312, 2021.

[42] Liu, Y., Mai, S., Chen, X., Hsieh, C.-J., and You, Y. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[43] Luo, H., Truong, T., Pham, T., Harandi, M., Phung, D., and Le, T. Explicit eigenvalue regularization improves sharpness-aware minimization. In *Conference on Neural Information Processing Systems*, 2024.

[44] Lyu, K., Li, Z., and Arora, S. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35:34689–34708, 2022.

[45] Mackay, M., Vicol, P., Lorraine, J., Duvenaud, D., and Grosse, R. Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions. In *International Conference on Learning Representations*, 2019.

[46] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[47] Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

[48] Miyato, T. and Koyama, M. cgans with projection discriminator. In *International Conference on Learning Representations*, 2018.

[49] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[50] Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

[51] Norton, M. D. and Royset, J. O. Diametrical risk minimization: Theory and computations. In *Machine Learning*, pp. 1–19, 2021.

[52] Oikonomou, D. and Loizou, N. Sharpness-aware minimization: General analysis and improved rates. In *International Conference on Learning Representations*, 2025.

[53] Pedregosa, F. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746, 2016.

[54] Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[55] Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[56] Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. In *Conference on Neural Information Processing Systems*, 2019.

[57] Shen, H. and Chen, T. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pp. 30992–31015, 2023.

[58] Si, D. and Yun, C. Practical sharpness-aware minimization cannot converge all the way to optima. In *Conference on Neural Information Processing Systems*, 2023.

[59] Springer, J. M., Nagarajan, V., and Raghunathan, A. Sharpness-aware minimization enhances feature quality via balanced learning. In *International Conference on Learning Representations*, 2024.

[60] Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., and Sutton, C. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017.

[61] Sun, R., Fang, T., and Schwing, A. Towards a better global loss landscape of gans. In *Conference on Neural Information Processing Systems*, 2020.

[62] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[63] Tseng, C.-H., Lee, S.-J., Feng, J.-N., Mao, S., Wu, Y.-P., Shang, J.-Y., and Zeng, X.-J. Upanets: Learning from the universal pixel attention networks. *Entropy*, 24(9):1243, 2022.

[64] Uppal, A., Singh, S., and Poczos, B. Non-parametric density estimation & convergence rates for gans under besov ipm losses. In *Conference on Neural Information Processing Systems*, 2019.

[65] Vlatakis-Gkaragkounis, E.-V., Flokas, L., and Piliouras, G. Solving min-max optimization with hidden structure via gradient descent ascent. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2373–2386, 2021.

[66] Wang, P., Zhu, H., Huang, H., Zhang, H., and Wang, N. Tms-gan: A twofold multi-scale generative adversarial network for single image dehazing. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2760–2772, 2021.

[67] Wang, P., Wang, X., Luo, H., Zhou, J., Zhou, Z., Wang, F., Li, H., and Jin, R. Scaled relu matters for training vision transformers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2495–2503, 2022.

[68] Wen, K., Ma, T., and Li, Z. How sharpness-aware minimization minimizes sharpness? In *International Conference on Learning Representations*, 2023.

[69] Wen, K., Li, Z., and Ma, T. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

[70] Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

[71] Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

[72] Wu, T., Luo, T., and Wunsch II, D. C. Cr-sam: Curvature regularized sharpness-aware minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6144–6152, 2024.

[73] Xie, C., Wu, Y., Maaten, L. v. d., Yuille, A. L., and He, K. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 501–509, 2019.

[74] Xie, W., Latorre, F., Antonakopoulos, K., Pethick, T. M., and Cevher, V. Improving sam requires rethinking its optimization formulation. In *International Conference on Machine Learning*, 2024.

[75] Ye, J. J., Yuan, X., Zeng, S., and Zhang, J. Difference of convex algorithms for bilevel programs with applications in hyperparameter selection. *Mathematical Programming*, 198(2): 1583–1616, 2023.

[76] Yu, N., Liu, G., Dundar, A., Tao, A., Catanzaro, B., Davis, L. S., and Fritz, M. Dual contrastive loss and attention for gans. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6731–6742, 2021.

[77] Yuan, M. and Peng, Y. Bridge-gan: Interpretable representation learning for text-to-image synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):4258–4268, 2019.

[78] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[79] Zhang, G., Cui, K., Hung, T.-Y., and Lu, S. Defect-gan: High-fidelity defect synthesis for automated defect inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2524–2534, 2021.

[80] Zhang, Y., Zhang, G., Khanduri, P., Hong, M., Chang, S., and Liu, S. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *International Conference on Machine Learning*, pp. 26693–26712, 2022.

[81] Zhang, Y., Khanduri, P., Tsaknakis, I., Yao, Y., Hong, M., and Liu, S. An introduction to bilevel optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine*, 41(1):38–59, 2024.

[82] Zhang, Z., Hua, Y., Wang, H., and McLoone, S. Improving the fairness of the min-max game in gans training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2910–2919, 2024.

[83] Zheng, Y., Zhang, R., and Mao, Y. Regularizing neural networks via adversarial model perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8156–8165, 2021.

A Convergence Analysis

This part will present some classical and basic properties in the convergence analysis for this paper.

Preliminaries for Convergence Analysis

In this subsection, we present essential properties of the Moreau envelope and its gradient, which form the foundation for our convergence analysis. Detailed proofs of the following results can be found in [18, 40].

We now recall a regularity result showing that the Moreau envelope inherits convexity under weak convexity of the original function.

Lemma A.1. *In (13), suppose $L_\ell(\omega, \delta)$ is weakly convex in ω and δ with constants $\rho_1, \rho_2 \geq 0$, respectively, and $\gamma \in (0, \frac{1}{2\rho_2})$. Then, for any $\rho_{v_1} \geq \rho_1$ and $\rho_{v_2} \geq \frac{1}{\gamma}$, the function*

$$v_\gamma(\omega, \delta) + \frac{\rho_{v_1}}{2}\|\omega\|^2 + \frac{\rho_{v_2}}{2}\|\delta\|^2$$

is convex over $\mathbb{R}^n \times \mathbb{R}^n$.

Next, we state a key result on the differentiability of the Moreau envelope, which allows its gradient to be expressed via the minimizer of the inner problem.

Lemma A.2. *Under the assumptions of Lemma A.1, the solution of the minimization problem*

$$S_\gamma(\omega, \delta) := \arg\min_{\theta \in \mathcal{C}} \left\{ L_\ell(\omega, \theta) + \frac{1}{2\gamma}\|\theta - \delta\|^2 \right\}$$

is unique for all (ω, δ). Let $\theta^*_\gamma(\omega, \delta)$ denote this unique minimizer. Then the Moreau envelope v_γ is differentiable and satisfies

$$\nabla v_\gamma(\omega, \delta) = \left(\nabla_\omega L_\ell(\omega, \theta^*_\gamma(\omega, \delta)), \frac{1}{\gamma}(\delta - \theta^*_\gamma(\omega, \delta)) \right).$$

Finally, we establish a Lipschitz-type bound for the smoothed minimizer $\theta^*_\gamma(\omega, \delta)$, which is critical for bounding gradient approximation errors.

Lemma A.3. *Let $\gamma \in \left(0, \frac{1}{2\rho_2}\right)$ and define $L_\theta := \max\{1, \gamma\ell\}$. Then for any $(\omega, \delta), (\bar{\omega}, \bar{\delta}) \in \mathbb{R}^n \times \mathbb{R}^m$, we have*

$$\|\theta^*_\gamma(\omega, \delta) - \theta^*_\gamma(\bar{\omega}, \bar{\delta})\| \leq L_\theta \left(\|\omega - \bar{\omega}\| + \|\delta - \bar{\delta}\| \right).$$

These results provide the theoretical basis for analyzing the approximation error and stability of our proposed algorithm.

Gradient of the Normalized Gradient

In this subsection, we calculate the gradient of the discarded term $\frac{d\delta^*(\omega)}{d\omega}$ in (7) in details. Recall that $\omega \in \mathbb{R}^n$ and $L : \mathbb{R}^n \to \mathbb{R}$ be a smooth function. Define the normalized gradient mapping as:

$$f(\omega) := \frac{\nabla L(\omega)}{\|\nabla L(\omega)\|}.$$

We aim to compute the Jacobian $\nabla_\omega f(\omega) \in \mathbb{R}^{n \times n}$. Denote:

$$g(\omega) := \nabla L(\omega), \quad h(\omega) := \|g(\omega)\| = \|\nabla L(\omega)\|.$$

Then we have:

$$f(\omega) = \frac{g(\omega)}{h(\omega)}.$$

Using the quotient rule for vector-valued functions:

$$\nabla_\omega \left(\frac{g(\omega)}{h(\omega)} \right) = \frac{1}{h(\omega)} \nabla_\omega g(\omega) - \frac{g(\omega)}{h(\omega)^2} \nabla_\omega h(\omega)^T.$$

We now compute the two components separately:

- The Jacobian of $g(\omega)$ is the Hessian matrix:

$$\nabla_\omega g(\omega) = \nabla^2 L(\omega) =: H_L(\omega).$$

- The gradient of the scalar function $h(\omega)$ is given by:

$$\nabla_\omega h(\omega) = \nabla_\omega \|g(\omega)\| = \frac{H_L(\omega)\nabla L(\omega)}{\|\nabla L(\omega)\|}.$$

Substituting into the quotient rule, we obtain:

$$\begin{aligned}\nabla_\omega \left(\frac{\nabla L(\omega)}{\|\nabla L(\omega)\|}\right) &= \frac{1}{\|\nabla L(\omega)\|} H_L(\omega) - \frac{\nabla L(\omega)}{\|\nabla L(\omega)\|^2}\left(\frac{H_L(\omega)\nabla L(\omega)}{\|\nabla L(\omega)\|}\right)^T \\ &= \frac{1}{\|\nabla L(\omega)\|}\left(H_L(\omega) - \frac{\nabla L(\omega)\nabla L(\omega)^\top H_L(\omega)}{\|\nabla L(\omega)\|^2}\right).\end{aligned}$$

This expression gives the Jacobian matrix of the normalized gradient $f(\omega)$, which appears frequently in sharpness-aware optimization and directional smoothing.

Algorithms

In this subsection, we underly our algorithm. The algorithm follows a standard stochastic process. Given batches $b_k \subset \{1, \ldots, N\}$ at each iteration, the procedure is summarized as follows:

- **Initialization:** Set step sizes $\{\alpha_k\}$, $\{\beta_k\}$, $\{\eta_k\}$, and penalty parameters $\{\mu_k\}$, and initialize the variables ω^0, δ^0, and θ^0.
- **Update rules:**

$$\begin{aligned}\theta^{k+1} &= \text{Proj}_{\mathcal{C}}\left(\theta^k - \eta_k[\nabla_\delta L_\ell^{b_k}(\omega^k, \theta^k) + \frac{1}{\gamma}(\theta^k - \delta^k)]\right), \\ \delta^{k+1} &= \text{Proj}_{\mathcal{C}}\left(\delta^k - \beta_k\left[\frac{1}{\mu_k}\nabla_\delta L^{b_k}(\omega^k, \delta^k) + \nabla_\delta L_\ell^{b_k}(\omega^k, \delta^k) - \frac{1}{\gamma}(\delta^k - \theta^{k+1})\right]\right), \\ \omega^{k+1} &= \omega^k - \alpha_k\left[\frac{1}{\mu_k}\nabla_\omega L^{b_k}(\omega^k, \delta^{k+1}) + \nabla_\omega L_\ell^{b_k}(\omega^k, \delta^{k+1}) - \nabla_\omega L_\ell^{b_k}(\omega^k, \theta^{k+1})\right].\end{aligned} \tag{18}$$

Algorithm for SAM

In particular, we consider the SAM problem given in (3). For a fixed pair (ω, δ), the Moreau envelope reformulates the lower-level problem as the following smooth optimization problem:

$$\min_{\theta \in \mathcal{C}} L_\ell(\omega, \theta) + \frac{1}{2\gamma}\|\theta - \delta\|^2, \tag{19}$$

which is typically convex in θ. In this case, any Karush-Kuhn-Tucker (KKT) point corresponds to a global minimizer. The global minimizer $\theta^*_\gamma(\omega, \delta)$ of (19) satisfies the optimality condition:

$$0 \in \nabla_\delta L_\ell(\omega, \theta^*) + \frac{1}{\gamma}(\theta^* - \delta) + N(\theta^*, \mathcal{C}),$$

where $\mathcal{C} = \{\delta \in \mathbb{R}^n : \|\delta\| \leq 1\}$, and $N(\theta^*, \mathcal{C})$ denotes the normal cone to $\mathcal{C}$ at θ^*, defined by:

$$N(\theta^*, \mathcal{C}) = \begin{cases}\{0\}, & \|\theta^*\| < 1; \\ \{\lambda\theta^* : \lambda \geq 0\}, & \|\theta^*\| = 1.\end{cases} \tag{20}$$

Consequently, the closed-form expression for the global minimizer $\theta^*_\gamma(\omega, \delta)$ is:

$$\theta^*_\gamma(\omega, \delta) = \begin{cases}\delta + \gamma\nabla L(\omega), & \|\delta + \gamma\nabla L(\omega)\| < 1; \\ \frac{\delta + \gamma\nabla L(\omega)}{\|\delta + \gamma\nabla L(\omega)\|}, & \|\delta + \gamma\nabla L(\omega)\| \geq 1.\end{cases} \tag{21}$$

Then we adopt the following algorithm:

- **Initialization:** Set step sizes $\{\alpha_k\}$, $\{\beta_k\}$, $\{\eta_k\}$, and penalty parameters $\{\mu_k\}$, and initialize the variables ω^0, δ^0, and θ^0.
- **Update:**

$$\theta^{k+1} = \begin{cases} \delta^k + \gamma \nabla L^{b_k}(\omega^k), & \|\delta^k + \gamma \nabla L^{b_k}(\omega^k)\| < 1, \\ \frac{\delta^k + \gamma \nabla L^{b_k}(\omega^k)}{\|\delta^k + \gamma \nabla L^{b_k}(\omega^k)\|}, & \|\delta^k + \gamma \nabla L^{b_k}(\omega^k)\| \geq 1. \end{cases}$$

- **Generate i.i.d. Gaussian vectors:** Generate a sequence of independent and identically distributed (i.i.d.) Gaussian vectors $\{u_{k,j} \in \mathbb{R}^n\}_{j=1}^Q$, where each $u_{k,j}$ is sampled from a standard normal distribution. Update the parameters as follows:

$$\begin{aligned} \delta^{k+1} &= \frac{\nabla L^{b_k}(\omega^k)}{\|\nabla L^{b_k}(\omega^k)\|}, \\ \omega^{k+1} &= \omega^k - \alpha_k \left[\frac{1}{\mu_k} \nabla_\omega L^{b_k}(\omega^k + r\delta^{k+1}) + H_{L^{b_k}}(\omega^k)(\theta^{k+1} - \delta^{k+1}) \right] \\ &\approx \omega^k - \alpha_k \left[\frac{1}{\mu_k} \nabla_\omega L^{b_k}(\omega^k + r\delta^{k+1}) + J^*(\omega^k, \theta^{k+1} - \delta^{k+1}) \right] \\ &= \omega^k - \alpha_k \left[\frac{1}{\mu_k} \nabla_\omega L^{b_k}(\omega^k + r\delta^{k+1}) \right. \\ &\quad \left. + \frac{1}{Q} \sum_{j=1}^{Q} \left\langle \frac{\nabla L^{b_k}(\omega^k + \mu u_{k,j}) - \nabla L^{b_k}(\omega^k)}{\mu}, \theta^{k+1} - \delta^{k+1} \right\rangle u_{k,j} \right]. \end{aligned}$$

Here, H_L denotes the Hessian matrix, and J^* is an approximation matrix used for numerical purposes, as described in [19].

Convergence Result with Biased Estimator

In this section, we present the convergence analysis of the proposed algorithm. For convergence analysis, we impose the following assumptions on the bilevel formulation (10):

Assumption A.4. The lower-level function $L_\ell(\omega, \delta)$ is locally Lipschitz continuous with constant ℓ.

Assumption A.5. The functions $L(\omega, \delta)$ and $L_\ell(\omega, \delta)$ in (10) are weakly convex with respect to both ω and δ, with constants $\rho_{L,1}, \rho_{L,2}, \rho_1, \rho_2 \geq 0$, respectively. That is, the functions

$$L(\omega, \delta) + \rho_{L,1}\|\omega\|^2 + \rho_{L,2}\|\delta\|^2 \quad \text{and} \quad L_\ell(\omega, \delta) + \rho_1\|\omega\|^2 + \rho_2\|\delta\|^2$$

are convex.

The analysis relies on the following assumption regarding the variance of the biased gradient estimators.

Assumption A.6 (Biased Estimator)**.** For any batch $b \subset \{1, \dots, N\}$, the following bounds on the expected squared bias hold for $\sigma_\omega, \sigma_\delta, \sigma_{\ell,\omega}, \sigma_{\ell,\delta} \geq 0$:

$$\begin{aligned} &\mathbb{E}\left[\|\nabla_\omega L^b(\omega, \delta) - \nabla_\omega L(\omega, \delta)\|^2\right] \leq \sigma_\omega^2, \\ &\mathbb{E}\left[\|\nabla_\delta L^b(\omega, \delta) - \nabla_\delta L(\omega, \delta)\|^2\right] \leq \sigma_\delta^2, \\ &\mathbb{E}\left[\|\nabla_\omega L_\ell^b(\omega, \delta) - \nabla_\omega L_\ell(\omega, \delta)\|^2\right] \leq \sigma_{\ell,\omega}^2, \\ &\mathbb{E}\left[\|\nabla_\delta L_\ell^b(\omega, \delta) - \nabla_\delta L_\ell(\omega, \delta)\|^2\right] \leq \sigma_{\ell,\delta}^2. \end{aligned} \tag{22}$$

Given a penalty parameter $\mu > 0$, we define:

$$\begin{aligned} f_\mu(\omega, \delta) &:= \frac{1}{\mu} L(\omega, \delta) + L_\ell(\omega, \delta) - v_\gamma(\omega, \delta), \\ F_\mu(\omega, \delta) &:= L(\omega, \delta) + \mu[L_\ell(\omega, \delta) - v_\gamma(\omega, \delta)]. \end{aligned}$$

The convergence analysis will be conducted primarily based on these two functions. Set

$$\begin{aligned} d_\delta^k &:= \frac{1}{\mu_k}\nabla_\delta L^{b_k}(\omega^k,\delta^k) + \nabla_\delta L_\ell^{b_k}(\omega^k,\delta^k) - \frac{1}{\gamma}(\delta^k - \theta^{k+1}); \\ d_\omega^k &:= \frac{1}{\mu_k}\nabla_\omega L^{b_k}(\omega^k,\delta^{k+1}) + \nabla_\omega L_\ell^{b_k}(\omega^k,\delta^{k+1}) - \nabla_\omega L_\ell^{b_k}(\omega^k,\theta^{k+1}). \end{aligned} \tag{23}$$

The following lemma shows that the expected value of f_{μ_k} exhibits a monotonic decreasing behavior, which plays a critical role in the convergence analysis.

Lemma A.7. *Under Assumptions A.4, A.5, A.6 and assumptions in Lemma A.2, the sequence (ω^k, δ^k) generated by the algorithm satisfies*

$$\begin{aligned} \mathbb{E}[f_{\mu_k}(\omega^{k+1},\delta^{k+1}) - f_{\mu_k}(\omega^k,\delta^k)] \le & \left[\frac{L_{f_k}}{2} - \frac{1}{8\alpha_k}\right]\|\omega^k - \omega^{k+1}\|^2 \\ & + \left[\frac{1}{2}L_{f_k} - \frac{1}{4\beta_k} + \frac{\alpha_k \ell^2 L_\theta^2}{2}\right]\|\delta^k - \delta^{k+1}\|^2 \\ & + \left(\frac{\alpha_k\ell^2}{2} + \frac{\beta_k}{2\gamma^2}\right)\|\theta^{k+1} - \theta_\gamma^*(\omega^k,\delta^k)\|^2 \\ & + \frac{2\alpha_k}{\mu_k^2}\sigma_\omega^2 + 4\alpha_k\sigma_{\ell,\omega}^2 + \frac{2\beta_k}{\mu_k^2}\sigma_\delta^2 + 2\beta_k\sigma_{\ell,\delta}^2. \end{aligned} \tag{24}$$

where L_{f_k} is the Lipschitz constant of the gradient of f_{μ_k} in its second-order Taylor expansion.

Proof. Assuming that f_{μ_k} is smooth, we have the inequality with $L_{f_k} > 0$:

$$\begin{aligned} & f_{\mu_k}(\omega^{k+1},\delta^{k+1}) \le f_{\mu_k}(\omega^k,\delta^{k+1}) + \langle \nabla_\omega f_{\mu_k}(\omega^k,\delta^{k+1}), \omega^{k+1} - \omega^k\rangle + \frac{L_{f_k}}{2}\|\omega^{k+1} - \omega^k\|^2 \\ & = f_{\mu_k}(\omega^k,\delta^{k+1}) + \langle \nabla_\omega f_{\mu_k}(\omega^k,\delta^{k+1}) - d_\omega^k + d_\omega^k, \omega^{k+1} - \omega^k\rangle + \frac{L_{f_k}}{2}\|\omega^{k+1} - \omega^k\|^2 \\ & = f_{\mu_k}(\omega^k,\delta^{k+1}) + \langle \nabla_\omega f_{\mu_k}(\omega^k,\delta^{k+1}) - d_\omega^k, \omega^{k+1} - \omega^k\rangle + \left(\frac{L_{f_k}}{2} - \frac{1}{\alpha_k}\right)\|\omega^{k+1} - \omega^k\|^2. \end{aligned} \tag{25}$$

By Assumption A.6, the expectation of the first inner product term satisfies:

$$\begin{aligned} & E\left[\langle \nabla_\omega f_{\mu_k}(\omega^k,\delta^{k+1}) - d_\omega^k, \omega^{k+1} - \omega^k\rangle\right] \\ & \le E[\langle \frac{1}{\mu_k}(\nabla_\omega L(\omega^k,\delta^{k+1}) - \nabla_\omega L^{b_k}(\omega^k,\delta^{k+1})) + (\nabla_\omega L_\ell(\omega^k,\delta^{k+1}) - \nabla_\omega L_\ell^{b_k}(\omega^k,\delta^{k+1})) \\ & \quad - (\nabla_\omega L_\ell(\omega^k,\theta_\gamma^*(\omega^k,\delta^{k+1})) - \nabla_\omega L_\ell^{b_k}(\omega^k,\theta^{k+1})), \omega^{k+1} - \omega^k\rangle] \\ & \le \frac{2\alpha_k}{\mu_k^2}\sigma_\omega^2 + \frac{1}{8\alpha_k}\|\omega^{k+1} - \omega^k\|^2 + 2\alpha_k\sigma_{\ell,\omega}^2 + \frac{1}{8\alpha_k}\|\omega^{k+1} - \omega^k\|^2 \\ & \quad + 2\alpha_k\sigma_{\ell,\omega}^2 + \frac{1}{8\alpha_k}\|\omega^{k+1} - \omega^k\|^2 + \ell\|\theta^{k+1} - \theta_\gamma^*(\omega^k,\delta^{k+1})\|\|\omega^{k+1} - \omega^k\| \\ & \le \frac{7}{8\alpha_k}\|\omega^{k+1} - \omega^k\|^2 + \frac{\alpha_k\ell^2}{2}\|\theta^{k+1} - \theta_\gamma^*(\omega^k,\delta^{k+1})\|^2 + \frac{2\alpha_k}{\mu_k^2}\sigma_\omega^2 + 4\alpha_k\sigma_{\ell,\omega}^2. \end{aligned} \tag{26}$$

Similarly, we apply the same approach for the update of δ^{k+1} and obtain:

$$\begin{aligned} & f_{\mu_k}(\omega^k,\delta^{k+1}) \le f_{\mu_k}(\omega^k,\delta^k) + \langle \nabla_\delta f_{\mu_k}(\omega^k,\delta^k), \delta^{k+1} - \delta^k\rangle + \frac{L_{f_k}}{2}\|\delta^{k+1} - \delta^k\|^2 \\ & = f_{\mu_k}(\omega^k,\delta^k) + \langle \nabla_\delta f_{\mu_k}(\omega^k,\delta^k) - d_\delta^k + d_\delta^k, \delta^{k+1} - \delta^k\rangle + \frac{L_{f_k}}{2}\|\delta^{k+1} - \delta^k\|^2 \\ & = f_{\mu_k}(\omega^k,\delta^k) + \langle \nabla_\delta f_{\mu_k}(\omega^k,\delta^k) - d_\delta^k, \delta^{k+1} - \delta^k\rangle + \left(\frac{L_{f_k}}{2} - \frac{1}{\beta_k}\right)\|\delta^{k+1} - \delta^k\|^2, \end{aligned} \tag{27}$$

and

$$
\begin{aligned}
&E\left[\left\langle \nabla_\delta f_{\mu_k}(\omega^k,\delta^k) - d_\delta^k, \delta^{k+1}-\delta^k\right\rangle\right] \\
&\le \frac{2\beta_k}{\mu_k^2}\sigma_\delta^2 + \frac{1}{8\beta_k}\|\delta^{k+1}-\delta^k\|^2 + 2\beta_k\sigma_{\ell,\delta}^2 + \frac{1}{8\beta_k}\|\delta^{k+1}-\delta^k\|^2 + \frac{1}{2\beta_k}\|\delta^{k+1}-\delta^k\|^2 \\
&\quad + \frac{\beta_k}{2\gamma^2}\|\theta^{k+1} - \theta_\gamma^*(\omega^k,\delta^k)\|^2 \\
&\le \frac{2\beta_k}{\mu_k^2}\sigma_\delta^2 + 2\beta_k\sigma_{\ell,\delta}^2 + \frac{3}{4\beta_k}\|\delta^{k+1}-\delta^k\|^2 + \frac{\beta_k}{2\gamma^2}\|\theta^{k+1} - \theta_\gamma^*(\omega^k,\delta^k)\|^2.
\end{aligned}
\tag{28}
$$

Combining all parts, we obtain the claimed result. □

Using Lemma A.7, we obtain the following convergence result for the proposed algorithm:

Theorem A.8. *Suppose $\gamma \in (0, \frac{1}{2\rho_2})$ and penalty parameters satisfy $\bar{\mu}(k+1)^p \ge \mu_{k+1} \ge \mu_k \ge \underline{\mu} > 0$. Then, for any $0 < \underline{\alpha} < \alpha_k < \frac{1}{8L_{f_k}}$, $0 < \underline{\beta} < \beta_k < \frac{1}{4L_{f_k}}$, and $0 < \eta_k < \frac{1}{\rho_{L,2}+2\rho_2}$ the iterates $\{(\omega^k,\delta^k)\}$ generated by the algorithm satisfy*

$$
\mathbb{E}\left[\frac{1}{K^{1/2}}\min_{0\le k\le K} R_k^{b_k}(\omega^{k+1},\delta^{k+1})\right] = O\left(\frac{1}{K^{1/2}}\right). \tag{29}
$$

Furthermore, if the step sizes satisfy $\sum_{k=0}^{\infty}(\alpha_k+\beta_k) < \infty$, then for any $p \in (0, \frac{1}{2})$, it holds that

$$
\mathbb{E}\left[\min_{0\le k\le K} R_k^{b_k}(\omega^{k+1},\delta^{k+1})\right] = O\left(\frac{1}{K^{(1-2p)/2}}\right). \tag{30}
$$

Proof. By [18, Proposition 2.1,],[40, Lemma A.7], we deduce that $(e_\omega^k, e_\delta^k) \in \nabla R_k^{b_k}(\omega^{k+1},\delta^{k+1})$ with

$$
\begin{aligned}
e_\omega^k &= \nabla_\omega F_{\mu_k}^{b_k}(\omega^{k+1},\delta^{k+1}) - \mu_k d_\omega^k - \frac{\mu_k}{\alpha_k}(\omega^{k+1}-\omega^k) \\
e_\delta^k &= \nabla_\delta F_{\mu_k}^{b_k}(\omega^{k+1},\delta^{k+1}) - \mu_k d_\delta^k - \frac{\mu_k}{\beta_k}(\delta^{k+1}-\delta^k).
\end{aligned}
\tag{31}
$$

Then by Lipschitz continuity of F_{μ_k} and L_ℓ with constants L_{F_k}, ℓ, we have

$$
\begin{aligned}
\|e_\omega^k\| &\le \|\nabla_\omega F_{\mu_k}^{b_k}(\omega^{k+1},\delta^{k+1}) - \nabla_\omega F_{\mu_k}^{b_k}(\omega^k,\delta^{k+1})\| + \|\nabla_\omega F_{\mu_k}^{b_k}(\omega^k,\delta^{k+1}) - \mu_k d_\omega^k\| \\
&\quad + \|\frac{\mu_k}{\alpha_k}(\omega^{k+1}-\omega^k)\| \\
&\le \mu_k L_{F_k}\|\omega^{k+1}-\omega^k\| + \frac{\mu_k}{\alpha_k}\|\omega^{k+1}-\omega^k\| + \mu_k\ell\|\theta^{k+1}-\theta_\gamma^*(\omega^k,\delta^{k+1})\|,
\end{aligned}
\tag{32}
$$

$$
\begin{aligned}
\|e_\delta^k\| &\le \|\nabla_\delta F_{\mu_k}^{b_k}(\omega^{k+1},\delta^{k+1}) - \nabla_\delta F_{\mu_k}^{b_k}(\omega^k,\delta^k)\| + \|\nabla_\delta F_{\mu_k}^{b_k}(\omega^k,\delta^k) - \mu_k d_\delta^k\| \\
&\quad + \|\frac{\mu_k}{\beta_k}(\delta^{k+1}-\delta^k)\| \\
&\le \mu_k L_{F_k}\|(\omega^{k+1}-\omega^k, \delta^{k+1}-\delta^k)\| + \frac{\mu_k}{\beta_k}\|\delta^{k+1}-\delta^k\| \\
&\quad + \frac{\mu_k}{\gamma}(\|\theta^{k+1}-\theta_\gamma^*(\omega^k,\delta^k)\| + L_\theta\|\omega^{k+1}-\omega^k\|).
\end{aligned}
\tag{33}
$$

Thus,

$$
\begin{aligned}
R_k^{b_k}(\omega^{k+1},\delta^{k+1}) \le &\mu_k(2L_{F_k} + \frac{1}{\alpha_k} + \frac{L_\theta}{\gamma})\|\omega^{k+1}-\omega^k\| + \mu_k(L_{F_k} + \frac{1}{\beta_k})\|\delta^{k+1}-\delta^k\| \\
&+ \mu_k(\ell\|\theta^{k+1}-\theta^*(\omega^k,\delta^{k+1})\| + \frac{1}{\gamma}\|\theta^{k+1}-\theta^*(\omega^k,\delta^k)\|).
\end{aligned}
\tag{34}
$$

By [40][Lemma A.6], there exists $\sigma_k \in (0,1)$ such that

$$
\|\theta^{k+1} - \theta_\gamma^*(\omega^k,\delta^k)\| \le \sigma_k\|\theta^k - \theta_\gamma^*(\omega^k,\delta^k)\|.
$$

Then we have

$$\begin{aligned}
&R_k^{b_k}(\omega^{k+1},\delta^{k+1})\\
&\le \mu_k(2L_{F_k}+\frac{1}{\alpha_k}+\frac{L_\theta}{\gamma})\|\omega^{k+1}-\omega^k\|+\mu_k(L_{F_k}+\frac{1}{\beta_k}+L_\theta\ell)\|\delta^{k+1}-\delta^k\|\\
&\quad+\mu_k(\ell+\frac{1}{\gamma})\|\theta^{k+1}-\theta_\gamma^*(\omega^k,\delta^k)\|\\
&\le \mu_k(2L_{F_k}+\frac{1}{\alpha_k}+\frac{L_\theta}{\gamma})\|\omega^{k+1}-\omega^k\|+\mu_k(L_{F_k}+\frac{1}{\beta_k}+L_\theta\ell)\|\delta^{k+1}-\delta^k\|\\
&\quad+\mu_k\sigma_k(\ell+\frac{1}{\gamma})\|\theta^k-\theta_\gamma^*(\omega^k,\delta^k)\|.
\end{aligned}\tag{35}$$

Because $\alpha_k>\underline{\alpha}>0,\beta_k>\underline{\beta}>0$, there exists $C_R>0,C_\theta>0$ such that

$$\begin{aligned}
&\frac{1}{\mu_k^2}[R_k^{b_k}(\omega^{k+1},\delta^{k+1})]^2\\
&\le C_R(\frac{1}{8\alpha_k}\|\omega^{k+1}-\omega^k\|^2+\frac{1}{16\beta_k}\|\delta^{k+1}-\delta^k\|^2+C_\theta\|\theta^k-\theta_\gamma^*(\omega^k,\delta^k)\|^2).
\end{aligned}\tag{36}$$

Since $\mu_k\le\mu_{k+1}$, then

$$f_{\mu_{k+1}}(\omega^{k+1},\delta^{k+1})-f_{\mu_k}(\omega^{k+1},\delta^{k+1})=(\frac{1}{\mu_{k+1}}-\frac{1}{\mu_k})L(\omega^{k+1},\delta^{k+1})<0.\tag{37}$$

By inserting (37) into Lemma A.7, we know

$$\begin{aligned}
&\mathbb{E}[f_{\mu_{k+1}}(\omega^{k+1},\delta^{k+1})+(\ell^2+\frac{1}{\gamma^2})\|\theta^{k+1}-\theta_\gamma^*(\omega^{k+1},\delta^{k+1})\|]\\
&\quad-\mathbb{E}[f_{\mu_k}(\omega^k,\delta^k)+(\ell^2+\frac{1}{\gamma^2})\|\theta^k-\theta_\gamma^*(\omega^k,\delta^k)\|]\\
&\le\mathbb{E}[f_{\mu_k}(\omega^{k+1},\delta^{k+1})-f_{\mu_k}(\omega^k,\delta^k)]+(\ell^2+\frac{1}{\gamma^2})[\|\theta^{k+1}-\theta_\gamma^*(\omega^{k+1},\delta^{k+1})\|\\
&\quad-\|\theta^k-\theta_\gamma^*(\omega^k,\delta^k)\|]\\
&\le\left[\frac{L_{f_k}}{2}-\frac{1}{8\alpha_k}\right]\|\omega^k-\omega^{k+1}\|^2+\left[\frac{1}{2}L_{f_k}-\frac{1}{4\beta_k}+\frac{\alpha_k\ell^2L_\theta^2}{2}\right]\|\delta^k-\delta^{k+1}\|^2\\
&\quad+(\frac{\alpha_k\ell^2}{2}+\frac{\beta_k}{2\gamma^2})\|\theta^{k+1}-\theta_\gamma^*(\omega^k,\delta^k)\|^2+(\ell^2+\frac{1}{\gamma^2})[\|\theta^{k+1}-\theta_\gamma^*(\omega^{k+1},\delta^{k+1})\|^2\\
&\quad-\|\theta^k-\theta_\gamma^*(\omega^k,\delta^k)\|^2]+\frac{2\alpha_k}{\mu_k^2}\sigma_\omega^2+4\alpha_k\sigma_{\ell,\omega}^2+\frac{2\beta_k}{\mu_k^2}\sigma_\delta^2+2\beta_k\sigma_{\ell,\delta}^2.
\end{aligned}\tag{38}$$

Besides, recall that by [40][Lemma A.6], there exists $\sigma_k\in(0,1)$ such that

$$\|\theta^{k+1}-\theta_\gamma^*(\omega^k,\delta^k)\|\le\sigma_k\|\theta^k-\theta_\gamma^*(\omega^k,\delta^k)\|.$$

Then for any $\epsilon_k>0$, we have

$$\begin{aligned}
&\frac{1}{2}\alpha_k\|\theta^{k+1}-\theta_\gamma^*(\omega^k,\delta^k)\|^2+\|\theta^{k+1}-\theta_\gamma^*(\omega^{k+1},\delta^{k+1})\|^2-\|\theta^k-\theta_\gamma^*(\omega^k,\delta^k)\|^2\\
&\le(\epsilon_k+\frac{1}{2}\alpha_k)\|\theta^{k+1}-\theta_\gamma^*(\omega^k,\delta^k)\|^2-\|\theta^k-\theta_\gamma^*(\omega^k,\delta^k)\|^2\\
&\quad+\frac{1}{\epsilon_k}\|\theta_\gamma^*(\omega^{k+1},\delta^{k+1})-\theta_\gamma^*(\omega^k,\delta^k)\|^2\\
&\le[\sigma_k^2(\epsilon_k+\frac{1}{2}\alpha_k)-1]\|\theta^k-\theta_\gamma^*(\omega^k,\delta^k)\|^2+\frac{L_\theta^2}{\epsilon_k}(\|\omega^{k+1}-\omega^k\|^2+\|\delta^{k+1}-\delta^k\|^2)
\end{aligned}\tag{39}$$

Similarily, we have

$$\begin{aligned}
&\frac{1}{2}\beta_k\|\theta^{k+1}-\theta^*_\gamma(\omega^k,\delta^k)\|^2+\|\theta^{k+1}-\theta^*_\gamma(\omega^{k+1},\delta^{k+1})\|^2-\|\theta^k-\theta^*_\gamma(\omega^k,\delta^k)\|^2\\
&\leq(\epsilon_k+\frac{1}{2}\beta_k)\|\theta^{k+1}-\theta^*_\gamma(\omega^k,\delta^k)\|^2-\|\theta^k-\theta^*_\gamma(\omega^k,\delta^k)\|^2\\
&\quad+\frac{1}{\epsilon_k}\|\theta^*_\gamma(\omega^{k+1},\delta^{k+1})-\theta^*_\gamma(\omega^k,\delta^k)\|^2\\
&\leq[\sigma_k^2(\epsilon_k+\frac{1}{2}\beta_k)-1]\|\theta^k-\theta^*_\gamma(\omega^k,\delta^k)\|^2+\frac{L_\theta^2}{\epsilon_k}(\|\omega^{k+1}-\omega^k\|^2+\|\delta^{k+1}-\delta^k\|^2)
\end{aligned}\tag{40}$$

Thus, when $\frac{1}{2}L_{f_k}+\frac{L_\theta^2}{\epsilon_k}\leq\frac{1}{16\alpha_k}$ and $\frac{1}{2}L_{f_k}+\frac{L_\theta^2}{\epsilon_k}\leq\frac{1}{8\beta_k}$ (by $\alpha_k<\frac{1}{8L_{f_k}},\beta_k<\frac{1}{4L_{f_k}}$), we have

$$\begin{aligned}
&\mathbb{E}[f_{\mu_{k+1}}(\omega^{k+1},\delta^{k+1})+(\ell^2+\frac{1}{\gamma^2})\|\theta^{k+1}-\theta^*_\gamma(\omega^{k+1},\delta^{k+1})\|]\\
&\quad-\mathbb{E}[f_{\mu_k}(\omega^k,\delta^k)+(\ell^2+\frac{1}{\gamma^2})\|\theta^k-\theta^*_\gamma(\omega^k,\delta^k)\|]\\
&\leq\ell^2[\sigma_k^2(\epsilon_k+\frac{1}{2}\alpha_k)+\frac{1}{\gamma^2}\sigma_k^2(\epsilon_k+\frac{1}{2}\beta_k)-1]\|\theta^k-\theta^*_\gamma(\omega^k,\delta^k)\|^2\\
&\quad+[\frac{1}{2}L_{f_k}-\frac{1}{8\alpha_k}+\frac{L_\theta^2}{\epsilon_k}]\|\omega^{k+1}-\omega^k\|^2+[\frac{1}{2}L_{f_k}-\frac{1}{4\beta_k}+\frac{L_\theta^2}{\epsilon_k}]\|\delta^{k+1}-\delta^k\|^2\\
&\quad+\frac{2\alpha_k}{\mu_k^2}\sigma_\omega^2+4\alpha_k\sigma_{\ell,\omega}^2+\frac{2\beta_k}{\mu_k^2}\sigma_\delta^2+2\beta_k\sigma_{\ell,\delta}^2\\
&\leq-C_R(C_\theta\|\theta^k-\theta^*_\gamma(\omega^k,\delta^k)\|^2+\frac{1}{16\alpha_k}\|\omega^{k+1}-\omega^k\|^2+\frac{1}{8\beta_k}\|\delta^{k+1}-\delta^k\|^2)\\
&\quad+\frac{2\alpha_k}{\mu_k^2}\sigma_\omega^2+4\alpha_k\sigma_{\ell,\omega}^2+\frac{2\beta_k}{\mu_k^2}\sigma_\delta^2+2\beta_k\sigma_{\ell,\delta}^2.
\end{aligned}\tag{41}$$

As a result, by take the sum of iteratives k in (36), when $\max\{\sigma_\omega^2,\sigma_\delta^2,\sigma_{\ell,\omega}^2,\sigma_{\ell,\delta}^2\}\leq\sigma^2$, $\mu_k\geq\underline{\mu}>0$, there exists a constant $C>0$ such that

$$\begin{aligned}
&C\sum_{k=0}^K\frac{1}{\mu_k^2}\mathbb{E}[R_k^{b_k}(\omega^{k+1},\delta^{k+1})]^2\\
&\leq f_{\mu_0}(\omega^0,\delta^0)+(\ell^2+\frac{1}{\gamma^2})\|\theta^0-\theta^*_\gamma(\omega^0,\delta^0)\|+\sum_{k=0}^K[\frac{2\alpha_k}{\mu_k^2}\sigma_\omega^2+4\alpha_k\sigma_{\ell,\omega}^2+\frac{2\beta_k}{\mu_k^2}\sigma_\delta^2+2\beta_k\sigma_{\ell,\delta}^2]\\
&\leq f_{\mu_0}(\omega^0,\delta^0)+(\ell^2+\frac{1}{\gamma^2})\|\theta^0-\theta^*_\gamma(\omega^0,\delta^0)\|+2\sigma^2(\frac{1}{\underline{\mu}^2}+1)\sum_{k=0}^K(2\alpha_k+\beta_k).
\end{aligned}\tag{42}$$

Because $\alpha_k<\frac{1}{8L_{f_k}}$ and $\beta_k<\frac{1}{4L_{f_k}}$, we have

$$\begin{aligned}
&\frac{C}{K}\sum_{k=0}^K\frac{1}{\mu_k^2}\mathbb{E}[R_k^{b_k}(\omega^{k+1},\delta^{k+1})]^2\\
&\leq\frac{1}{K}[f_{\mu_0}(\omega^0,\delta^0)+(\ell^2+\frac{1}{\gamma^2})\|\theta^0-\theta^*_\gamma(\omega^0,\delta^0)\|]+2\sigma^2(\frac{1}{\underline{\mu}^2}+1)(\frac{1}{2L_{f_k}}+\frac{1}{4L_{f_k}}).
\end{aligned}\tag{43}$$

Then

$$\frac{1}{K^{1/2}}\mathbb{E}[R_k^{b_k}(\omega^{k+1},\delta^{k+1})]=O(\frac{1}{K^{1/2}}).\tag{44}$$

Furthermore, when $\sum\limits_{k=0}^\infty(\alpha_k+\beta_k)<\infty$, we have

$$\sum_{k=0}^\infty\frac{1}{\mu_k^2}\mathbb{E}[R_k^{b_k}(\omega^{k+1},\delta^{k+1})]^2<\infty.\tag{45}$$

For $p \in (0, \frac{1}{2})$, because $\mu_k \leq \bar{\mu}(k+1)^p$

$$\sum_{k=0}^{K} \frac{1}{\mu_k^2} \geq \frac{1}{\bar{\mu}^2} \sum_{k=0}^{K} (\frac{1}{k+1})^{2p} \geq \frac{1}{\bar{\mu}^2} \int_1^{K+2} \frac{1}{t^{2p}} dt \geq \frac{(K+2)^{1-2p} - 1}{(1-2p)\bar{\mu}^2}. \tag{46}$$

then

$$\mathbb{E}[\min_{0 \leq k \leq K} R_k^{b_k}(\omega^{k+1}, \delta^{k+1})] = O(\frac{1}{K^{(1-2p)/2}}). \tag{47}$$

□

B Experimental Details

We conducted the experiments on a PC with Intel i5-13600KF CPU (3.5 GHz), 32GB RAM and NVIDIA RTX 4090 GPU. We leveraged the PyTorch framework on the 64-bit Linux system. In the following, we elaborately introduce the implementation details and parameter configurations.

Numerical Cases

As for the first case, we set η, β, α, γ, μ, and p as 0.001, 0.01, 0.0001, 20, and 0.1 and leverage $\|\omega - \omega^*\| \leq 1e^{-4}$ as the stop criterion. SGD optimizer is used for the update of ω. We set the the maximum steps of optimization as 1000 uniformly. As for the second case, η, β, α, γ, μ, and p are 0.001, 0.05, 0.0001, 20, and 0.1, respectively.

Generative Adversarial Learning

The goal of generative adversarial learning is to build a min-max game between the generator $\texttt{Gen}(\omega; \cdot)$ and discriminator $\texttt{Dis}(\delta; \cdot)$, which can be formulated as $\min_\omega \max_\delta \log(\texttt{Dis}(\delta; \mathbf{u})) + \log(1 - \texttt{Dis}(\delta; \texttt{Gen}(\omega; \mathbf{v})))$, where $\mathbf{u} \sim p_{\texttt{data}}$ represents the real data distribution and $\mathbf{v}$ denotes the random latent vector. The hyperparameters η, β, α, γ, μ, and p are set to 0.005, 0.005, 0.01, 100, 5 and 0.1, respectively. The generator is a three-layer fully connected neural network that maps an input vector to a target output through two hidden layers. Each hidden layer is followed by a non-linear activation function. The final layer outputs the generated vector without activation. The discriminator consists of two hidden layers and a final classification layer. All hidden layers use the ReLU activation function. The output is passed through a sigmoid activation. We conduct the experiments under 8 distribution of 2D wheel, sampling 500 data points from 2D surfaces. All experiments are repeated three times with diverse random seeds.

Sharpness-aware Minimization

We conduct image classification experiments using the standard open-source CIFAR-10 benchmark, which consists of 50,000 training and 10,000 testing image-label pairs. ResNet-18 is employed as the backbone model to evaluate performance under noisy labels and various perturbation radius r. For fair comparison in experiments involving both noisy labels and different backbone architectures, we fix $r = 0.1$. The hyperparameters α, γ, μ, Q, and p are set to 0.05, 1×10^{-4}, 0.75, 1 and 0.01, respectively. Following the setup in [2], we apply basic augmentation during training, including horizontal flipping, four-pixel padding, and cropping. Models are trained from scratch for 200 epochs using a batch size of 128 and a cosine learning rate schedule. To ensure fair comparison, all experiments are repeated three times with diverse random seeds.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: Our theoretical results (assumptions, lemmas, theorems) are formulated under general problem settings, with limitations naturally addressed within this framework, eliminating the need for a separate limitations section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This study focuses on advancements in machine learning. While potential societal implications may exist, they fall beyond the scope of the current technical contribution and thus are not analyzed herein.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve such researches.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This Paper does not involve such researches.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs only for editing and formatting.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.