Semi-Implicit Hybrid Gradient Methods with Application to Adversarial Robustness

Beomsu Kim
Dept. of Mathematical Sciences, KAIST

Junghoon Seo
SI Analytics, Inc.

Abstract

Adversarial examples, crafted by adding imperceptible perturbations to natural inputs, can easily fool deep neural networks (DNNs). One of the most successful methods for training adversarially robust DNNs is solving a nonconvex-nonconcave minimax problem with an adversarial training (AT) algorithm. However, among the many AT algorithms, only Dynamic AT (DAT) and You Only Propagate Once (YOPO) is guaranteed to converge to a stationary point with rate $O(1/K^{1/2})$. In this work, we generalize the stochastic primal-dual hybrid gradient algorithm to develop semi-implicit hybrid gradient methods (SI-HGs) for finding stationary points of nonconvex-nonconcave minimax problems. SI-HGs have the convergence rate $O(1/K)$, which improves upon the rate $O(1/K^{1/2})$ of DAT and YOPO. We devise a practical variant of SI-HGs, and show that it outperforms other AT algorithms in terms of convergence speed and robustness.

1 INTRODUCTION

Adversarial examples, crafted by adding imperceptible adversarial perturbations to natural inputs, can easily fool deep neural networks (DNNs) (Goodfellow et al., 2015). One of the most successful methods for learning adversarially robust DNNs is adversarial training (AT) (Athalye et al., 2018). Given a dataset $\{(x_i, y_i)\}_{i=1}^n$ comprised of n input-label pairs or batches, a vector of perturbations $\delta = (\delta_1, \dots, \delta_n)$, perturbation radius $\epsilon > 0$, DNN parameters w, and a loss function ℓ, AT solves the nonconvex-nonconcave minimax problem

$$\min_w \max_{\|\delta\|_\infty \le \epsilon} \frac{1}{n} \sum_{i=1}^n \ell(x_i + \delta_i, y_i, w). \tag{1}$$

However, among the large number of AT algorithms (Silva & Najafirad, 2020), many do not have theoretical convergence guarantees. Some such algorithms even exhibit a failure mode called catastrophic overfitting, where the DNN accuracy on adversarial examples generated by multiple steps of projected gradient ascent (PGD) drops to a low value in the middle of training (Andriushchenko & Flammarion, 2020).

To the best of our knowledge, there are only three AT algorithms guaranteed to converge to a stationary point of (1): PGD AT (Mądry et al., 2018), Dynamic AT (DAT) (Wang et al., 2019), and You Only Propagate Once (YOPO) (Zhang et al., 2019a; Seidman et al., 2020). PGD AT only has a global convergence guarantee. Under certain assumptions on the loss function and stochastic gradients, DAT and YOPO decrease the squared saddle subdifferential norm with rate $O(1/K^{1/2})$ up to an additive constant.

In this work, we take a step towards AT algorithms with better convergence guarantees. To this end, we consider minimax optimization problems of the form

$$\min_w \max_\delta f(w) + \phi(w, \delta) - g(\delta) \tag{2}$$

where

$$\phi(w, \delta) = \sum_{i=1}^n \phi_i(w, \delta_i), \qquad g(\delta) = \sum_{i=1}^n g_i(\delta_i) \tag{3}$$

and $w \in \mathbb{R}^m$, $\delta = (\delta_1, \dots, \delta_n) \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$, f, g_i are convex, and ϕ is nonconvex-nonconcave. The AT problem (1) is a special case of this template (c.f. Equation (5)). We propose two semi-implicit hybrid gradient methods (SI-HGs) which solve (2) by alternating between a hybrid gradient descent step on w and an implicit ascent step on δ. The

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

Table 1: Convergence rates w.r.t. squared saddle subdifferential norm for AT algorithms. "—" means N/A. "DT" means deterministic, "S" means stochastic, and "DS" means doubly-stochastic. K is the number of iterations. We omit additive constants for the convergence rates. For SSI-HG, smoothness of $\ell(\cdot, y_i, \cdot)$ is actually more than what we need to prove the $O(1/K)$ rate (c.f. Assumption 1 (b) and Theorem 1). Yet, we have written the stronger assumption in this table since we use the smoothness of $\ell(\cdot, y_i, \cdot)$ in the process of interpreting SSI-HG as MGDA (see Section 4).

AT Algorithm	Algorithm Type	Assumption	Setting	Convergence
FGSM AT (Goodfellow et al., 2015)	GDA	—	DS	—
PGD AT (Mądry et al., 2018)	MGDA	ℓ continuously diff. in w	DT	Global Convergence
DAT (Wang et al., 2019)	MGDA	$\ell(\cdot, y_i, \cdot)$ smooth $\ell(\cdot, y_i, w)$ locally strongly concave Stochastic gradient bounded variance	DS	$O(1/K^{1/2})$
YOPO (Seidman et al., 2020)	MGDA	$\ell(\cdot, y_i, \cdot)$ smooth $\ell(\cdot, y_i, w)$ locally strongly concave Stochastic gradient bounded variance	DS	$O(1/K^{1/2})$
SSI-HG (Ours)	MGDA	$\ell(\cdot, y_i, \cdot)$ smooth Weak MVI has solution	S	$O(1/K)$

first is stochastic[1] SI-HG (SSI-HG), which generalizes stochastic primal-dual hybrid gradient (SPDHG) (Chambolle et al., 2018). The second is deterministic SI-HG (DSI-HG), which is the $n = 1$ case of SSI-HG.

SI-HGs decrease the squared saddle subdifferential norm with rate $O(1/K)$ under the weak Minty variational inequality (MVI) condition (Diakonikolas & Daskalakis, 2021). This improves upon the aforementioned $O(1/K^{1/2})$ convergence rate of DAT and YOPO. We also prove that SI-HGs achieve a linear rate under the strong MVI condition (Zhou et al., 2017; Song et al., 2020). Our work is a solid step towards AT algorithms with better convergence properties. A variety of experiments substantiate this claim.

We have the following key contributions.

- **SSI-HG.** We propose SSI-HG, which generalizes SPDHG to the nonconvex-nonconcave minimax optimization setting. We show that SSI-HG can be interpreted as a stochastic multi-step gradient descent ascent method (MGDA). Under the weak MVI assumption, we prove that SSI-HG decreases the squared saddle subdifferential norm with rate $O(1/K)$. We also prove linear convergence under the strong MVI assumption. Our work improves upon the convergence rates for other AT algorithms (Table 1).
- **DSI-HG.** When $n = 1$, SSI-HG becomes DSI-HG. DSI-HG inherits all the convergence results for SSI-HG. Hence, if we interpret DSI-HG as MGDA, our work improves upon previous convergence results for MGDA in the nonconvex-nonconcave setting (Table 2).
- **Application of SI-HGs to AT.** We extend the theoretical development behind SI-HGs to the AT setting. Specifically, we develop a minibatch version of SI-HG (MSI-HG). We also propose a heuristic for solving the implicit step in the AT setting. In experiments, we demonstrate MSI-HG indeed converges faster and achieves better robustness than other popular AT methods on multiple datasets.

[1]We will refer to methods which use full gradients to update both w and δ as being *deterministic*, methods which use full gradients to update w and stochastic gradients to update δ as being *stochastic*, and methods which use stochastic gradients to update both w and δ as being *doubly-stochastic*.

2 RELATED WORK

Stochastic Primal-Dual Coordinate Methods (SPDCMs). When ϕ is bilinear, the template (2) encompasses problems such as total variation regularized imaging (Chambolle et al., 2018) and regularized empirical risk minimization (Shalev-Shwartz & Zhang, 2013). SPDCMs are often used to solve such problems (Chambolle et al., 2018; Fercoq & Bianchi, 2019; Latafat et al., 2019; Alacaoglu et al., 2020). Due to stochastic coordinate updates in the variable δ, these methods have lower per-iteration costs than deterministic primal-dual hybrid gradient methods (Chambolle

Table 2: Convergence rates for MGDA for (2) with $n = 1$ and nonconvex-nonconcave ϕ . "—" means no additional assumptions. Here, K is the number of iterations. For MGDA, one iteration consists of one x update and multiple y updates. The optimality metric for the first and third rows is the squared saddle subdifferential norm whereas the optimality metric for the second row is the squared gradient norm of the Moreau envelope of $\max_\delta \phi(\cdot, \delta)$.

$f(w)$	$\phi(w, \delta)$	$g(\delta)$	**Assumption**	**Convergence Rate**
Indicator function of a convex compact set	Smooth $-\phi(w, \cdot)$ is μ-PL	0	—	$O(1/K)$ (Nouiehed et al., 2019)
0	Lipschitz and smooth	0	—	$O(1/K^{1/2})$ (Jin et al., 2020)
Convex	Smooth	Convex	Weak MVI has solution	$O(1/K)$ (Ours)

& Pock, 2011; Condat, 2013; Vũ, 2013). SSI-HG is inspired by a SPDCM called SPDHG, and it generalizes SPDHG to the setting where the coupling function ϕ is nonconvex-nonconcave.

MGDA. MGDA (Nouiehed et al., 2019; Thekumparampil et al., 2019; Barazandeh & Razaviyayn, 2020; Jin et al., 2020) alternates between one gradient descent step for x and multiple gradient ascent steps for y to solve minimax problems. There are two convergence rates for MGDA in the nonconvex-nonconcave setting. Nouiehed et al. (2019) proves the rate $O(1/K)$ with respect to the squared saddle subdifferential norm assuming ϕ is smooth, $-\phi(w, \cdot)$ is μ-Polyak-Łojasiewicz (μ-PL), f is an indicator function of some convex compact set, and $g \equiv 0$. Jin et al. (2020) proves the rate $O(1/K^{1/2})$ with respect to the squared gradient norm of the Moreau envelope of $\max_\delta \phi(\cdot, \delta)$ assuming $f \equiv g \equiv 0$ and ϕ is Lipschitz and smooth. See Section 4 of the work by Jin et al. (2020) for an explanation on the relation between a function and its Moreau envelope. SI-HGs can be interpreted as variants of MGDA, and our work improves upon previous convergence results for MGDA in the nonconvex-nonconcave setting (Table 2).

Convergence Guarantees for AT Methods. PGD AT (Mądry et al., 2018) alternates between one gradient descent step for w and multiple projected (sign) gradient ascent steps for δ. Hence, PGD AT is essentially MGDA applied to (1). However, we cannot apply the convergence results by Nouiehed et al. (2019) or Jin et al. (2020) for MGDA to PGD AT. This is because g is an indicator function in the AT setting (c.f. Equation (5)). Mądry et al. (2018) shows the global convergence of PGD AT in the deterministic setting via Danskin's Theorem. There are also works which show the convergence of PGD AT in terms of the loss function value (Xing et al., 2020; Gao et al., 2019; Zhang et al., 2020b).

There are two other AT methods which guarantee convergence to a stationary point of (1). They are variants of PGD AT. The first method, DAT, uses a criterion called first-order stationary condition to adaptively control the number of ascent steps in the inner loop (Wang et al., 2019). The second method, YOPO, exploits the compositional structure of DNNs to reduce the computational cost of the inner loop (Zhang et al., 2019a; Seidman et al., 2020).

Both DAT and YOPO possess a $O(1/K^{1/2})$ rate of convergence under smoothness and locally strongly concave assumptions. In this work, we prove a $O(1/K)$ convergence rate for SSI-HG under smoothness and weak MVI assumptions.

3 PRELIMINARIES

Notations. We define the saddle subdifferential operator of (2) as

$$F(w, \delta) = \begin{bmatrix} \partial f(w) + \nabla_w \phi(w, \delta) \\ \partial g(\delta) - \nabla_\delta \phi(w, \delta) \end{bmatrix}$$

and its norm as

$$\|F(w, \delta)\| = \inf_{\gamma_f, \gamma_g} \left\| \begin{bmatrix} \gamma_f + \nabla_w \phi(w, \delta) \\ \gamma_g - \nabla_\delta \phi(w, \delta) \end{bmatrix} \right\|$$

where the infimum is taken over $\gamma_f \in \partial f(w)$ and $\gamma_g \in \partial g(\delta)$. For a positive scalar η, we denote $\|\cdot\|_\eta^2 = \eta \|\cdot\|^2$. Proximal operator with some function $h(z)$ and $\eta > 0$ is defined as

$$\text{prox}_h^\eta(z) = \arg\min_u h(u) + \frac{1}{2}\|u - z\|_{\eta^{-1}}^2.$$

We also use the notation $[n] = \{1, \dots, n\}$. Given a set $\mathcal{S}$, $\mathbb{I}_\mathcal{S}(z)$ is the indicator function which is 0 on $z \in \mathcal{S}$ and ∞ on $z \notin \mathcal{S}$. $\Pi_\mathcal{S}[z]$ denotes the projection of z onto $\mathcal{S}$ w.r.t. the Euclidean norm.

Nonconvex-nonconcave Minimax Optimization. We are interested in finding a first-order stationary point of (2), i.e., a point (w, δ) which satisfies

$$\mathbf{0} \in F(w, \delta) \quad \text{or equivalently,} \quad \|F(w, \delta)\| = 0.$$

We first make the following common assumption.

Assumption 1. *(a) f and g_i are closed, convex, and proper,*
(b) $\nabla\phi$ is Lipschitz continuous, i.e., there are $L_{11}, L_{12}, L_{22} > 0$ such that

$$\begin{aligned}\|\nabla_w\phi(w,\delta) - \nabla_w\phi(\bar{w},\delta)\| &\le L_{11}\|w-\bar{w}\|,\\ \|\nabla_\delta\phi(w,\delta) - \nabla_\delta\phi(\bar{w},\delta)\| &\le L_{12}\|w-\bar{w}\|,\\ \|\nabla_w\phi(w,\delta) - \nabla_w\phi(w,\bar{\delta})\| &\le L_{12}\|\delta-\bar{\delta}\|,\\ \|\nabla_\delta\phi(w,\delta) - \nabla_\delta\phi(w,\bar{\delta})\| &\le L_{22}\|\delta-\bar{\delta}\|.\end{aligned}$$

However, even with Assumption 1, finding a stationary point of (2) is, in general, intractable (Diakonikolas & Daskalakis, 2021). Thus, we need to impose additional structures onto the problem. One possible structure is the assumption that there is a solution (w^*, δ^*) to the MVI problem

$$\begin{bmatrix}\gamma_f + \nabla_w\phi(w,\delta)\\ \gamma_g - \nabla_\delta\phi(w,\delta)\end{bmatrix}^\top \begin{bmatrix}w-w^*\\ \delta-\delta^*\end{bmatrix} \ge 0$$

$\forall(w,\delta)$, $\forall\gamma_f \in \partial f(w)$, $\forall\gamma_g \in \partial g(\delta)$. The MVI assumption has been studied by numerous works (Dang & Lan, 2014; Malitsky, 2019; Liu et al., 2020b), and algorithms with good convergence guarantees under the MVI assumption have shown better performance than previous algorithms in training Generative Adversarial Nets as well (Gidel et al., 2019; Mertikopoulos et al., 2019; Liu et al., 2020a). In this work, we consider the following variants of the MVI assumption:

Assumption 2. *There is a solution (w^*, δ^*) to the weak MVI problem*

$$\begin{aligned}&\begin{bmatrix}\gamma_f + \nabla_w\phi(w,\delta)\\ \gamma_g - \nabla_\delta\phi(w,\delta)\end{bmatrix}^\top \begin{bmatrix}w-w^*\\ \delta-\delta^*\end{bmatrix}\\ &\ge -\frac{\rho}{2}\left\|\begin{bmatrix}\gamma_f + \nabla_w\phi(w,\delta)\\ \gamma_g - \nabla_\delta\phi(w,\delta)\end{bmatrix}\right\|^2\end{aligned}$$

for some $\rho > 0$, $\forall(w,\delta)$, $\forall\gamma_f \in \partial f(w)$, $\forall\gamma_g \in \partial g(\delta)$.

Assumption 3. *There is a solution (w^*, δ^*) to the strong MVI problem*

$$\begin{bmatrix}\gamma_f + \nabla_w\phi(w,\delta)\\ \gamma_g - \nabla_\delta\phi(w,\delta)\end{bmatrix}^\top \begin{bmatrix}w-w^*\\ \delta-\delta^*\end{bmatrix} \ge \frac{\mu}{2}\left\|\begin{bmatrix}w-w^*\\ \delta-\delta^*\end{bmatrix}\right\|^2$$

for some $\mu > 0$, $\forall(w,\delta)$, $\forall\gamma_f \in \partial f(w)$, $\forall\gamma_g \in \partial g(\delta)$.

We will discuss the justification of these theoretical assumptions in Section 6.

4 SEMI-IMPLICIT HYBRID GRADIENT METHODS (SI-HGs)

Stochastic SI-HG (SSI-HG). We propose SSI-HG (Algorithm 1) to find a stationary point of (2). SSI-HG alternates between a hybrid gradient ascent step on w (line 4) and a stochastic implicit step on δ (line 6). We remark that when $\phi(w,\delta)$ is bilinear, the implicit step becomes explicit, and SSI-HG reduces to SPDHG with uniform sampling probability[2]. Hence, SSI-HG generalizes SPDHG to the nonconvex-nonconcave minimax optimization setting.

Algorithm 1 `SSI-HG`

1: **Input:** $(w^{-1}, \delta^{-1}) = (w^0, \delta^0)$, σ, τ, θ.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad q^k = \nabla_w\phi(w^{k-1}, \delta^{k-1}) - (n-1)\{\nabla_w\phi(w^k, \delta^k) - \nabla_w\phi(w^k, \delta^{k-1})\}$
4: $\quad w^{k+1} = \mathrm{prox}_f^\sigma[w^k - \sigma\{\nabla_w\phi(w^k, \delta^k) + \theta(\nabla_w\phi(w^k, \delta^k) - q^k)\}]$
5: $\quad$ Draw $i_k \in [n]$ uniformly at random.
6: $\quad \delta_{i_k}^{k+1} = \mathrm{prox}_{g_{i_k}}^\tau[\delta_{i_k}^k + \tau\nabla_{\delta_{i_k}}\phi_{i_k}(w^{k+1}, \delta_{i_k}^{k+1})]$ and $\delta_i^{k+1} = \delta_i^k$ for all $i \ne i_k$
7: **end for**

We can also interpret SSI-HG as stochastic MGDA. The implicit step is equivalent to

$$\begin{aligned}\delta_{i_k}^{k+1} = \arg\min_{\delta_{i_k}} g_{i_k}(\delta_{i_k}) - \phi_{i_k}(w^{k+1}, \delta_{i_k})\\ + \frac{1}{2\tau}\|\delta_{i_k} - \delta_{i_k}^k\|^2. \quad (4)\end{aligned}$$

The equivalence is proven in Appendix B.1. If we assume each ϕ_i is smooth and use an iterative proximal method such as FISTA (Beck & Teboulle, 2009) to solve (4), SSI-HG alternates between gradient descent on w and multi-step stochastic gradient ascent on δ. Hence, SSI-HG becomes stochastic MGDA. We now present the main results of our paper. The proofs are deferred to Appendices B.3 and B.4.

Theorem 1. *Suppose Assumptions 1 and 2 are true. Let $\{(w^k, \delta^k)\}$ be the sequence generated by SSI-HG, and define the full-dimensional update (which only depends on w^k and δ^{k-1})*

$$\hat{\delta}^k = \mathrm{prox}_g^\tau[\delta^{k-1} + \tau\nabla_\delta\phi(w^k, \hat{\delta}^k)].$$

Let $L = \max\{L_{11}, L_{12}, L_{22}\}$. If $\theta = 1$ and

$$0 < \sigma \le \frac{1}{6L}, \qquad 0 < \tau \le \frac{1}{6nL},$$
$$0 < \rho < \frac{1}{6\max\{\sigma^{-1} + 4L^2\sigma, \tau^{-1} + 12nL^2\tau\}},$$

we have

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\|F(w^k, \hat{\delta}^k)\|^2 = O(1/K).$$

[2]Let $\phi(w,\delta) = \langle Aw, \delta\rangle$ for some matrix A. Then $\nabla_w\phi(w,\delta) = A^\top\delta$ and $\nabla_\delta\phi(w,\delta) = Aw$. Plug these relations into Algorithm 1, and compare with Algorithm 1 in the paper for SPDHG (Chambolle et al., 2018).

Theorem 2. *Suppose Assumptions 1 and 3 are true. Let $\{(w^k, \delta^k)\}$ be the sequence generated by SSI-HG. Let $L = \max\{L_{11}, L_{12}, L_{22}\}$. If θ, σ, τ satisfy*

$$0 < \sigma \leq \frac{1}{3L}, \qquad 0 < \tau \leq \frac{1}{3nL},$$
$$\theta = \max\left\{\frac{1}{1+\mu\sigma}, \frac{1+(n-1)\mu\tau/n}{1+\mu\tau}\right\},$$

we have

$$\mathbb{E}\|w^* - w^K\|^2 = O(\theta^K), \qquad \mathbb{E}\|\delta^* - \delta^K\|^2 = O(\theta^K).$$

Our proofs for SSI-HG are inspired by those in the work of Alacaoglu et al. (2020). Specifically, Assumption 2 or 3 allows us to characterize the one-iteration behavior of SSI-HG in the nonconvex-nonconcave scenario (Lemma 7 in Appendix B). We then use telescoping or induction to establish Theorems 1 and 2.

Since the AT problem (1) is a special case of (2) (c.f. Equation (5)), Theorem 1 improves upon the convergence rates for DAT and YOPO (Table 1). We have simplified the parameter conditions in Theorems 1 and 2 for readability. The general forms are written in Appendices B.3 and B.4.

Deterministic SI-HG (DSI-HG). When $n = 1$, SSI-HG becomes DSI-HG (Algorithm 2). All the convergence results in Theorems 1 and 2 hold for DSI-HG with expectation removed (Corollaries 3 and 4). In particular, if we interpret DSI-HG as MGDA, Corollary 3 improves upon previous guarantees for MGDA in the nonconvex-nonconcave setting (Table 2).

Algorithm 2 `DSI-HG`

1: **Input:** $(w^{-1}, \delta^{-1}) = (w^0, \delta^0)$, σ, τ, θ.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad w^{k+1} = \text{prox}_f^\sigma[w^k - \sigma\{\nabla_w \phi(w^k, \delta^k) + \theta(\nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^{k-1}, \delta^{k-1}))\}]$
4: $\quad \delta^{k+1} = \text{prox}_g^\tau[\delta^k + \tau \nabla_\delta \phi(w^{k+1}, \delta^{k+1})]$
5: **end for**

Corollary 3. *Suppose Assumptions 1 and 2 are true. Let $\{(w^k, \delta^k)\}$ be the sequence generated by DSI-HG. Let $L = \max\{L_{11}, L_{12}, L_{22}\}$. If $\theta = 1$ and*

$$0 < \sigma, \tau \leq \frac{1}{6L},$$
$$0 < \rho < \frac{1}{6\max\{\sigma^{-1} + 4L^2\sigma, \tau^{-1} + 12L^2\tau\}},$$

we have

$$\frac{1}{K}\sum_{k=1}^{K} \|F(w^k, \delta^k)\|^2 = O(1/K).$$

Corollary 4. *Suppose Assumptions 1 and 3 are true. Let $\{(w^k, \delta^k)\}$ be the sequence generated by DSI-HG. Let $L = \max\{L_{11}, L_{12}, L_{22}\}$. If θ, σ, τ satisfy*

$$0 < \sigma, \tau \leq \frac{1}{3L}, \qquad \theta = \max\left\{\frac{1}{1+\mu\sigma}, \frac{1}{1+\mu\tau}\right\},$$

we have

$$\|w^* - w^K\|^2 = O(\theta^K), \qquad \|\delta^* - \delta^K\|^2 = O(\theta^K).$$

We again remark that we have simplified the parameter conditions for readability of the Corollaries. The general forms are written in Appendix B.5.

4.1 Applying SI-HGs to AT

We now extend our theoretical intuition to the AT setting. This section is inspired by works which combine theoretically established algorithms with practical algorithms or use heuristics to extend algorithms to deep learning settings (Daskalakis et al., 2018; Gidel et al., 2019; Mertikopoulos et al., 2019; Nouiehed et al., 2019; Chavdarova et al., 2020; Sinha et al., 2018; Wang et al., 2019).

Denote $\mathbb{B}_i := \{\delta_i' : \|\delta_i'\|_\infty \leq \epsilon\}$. The AT problem (1) can be written as

$$\min_w \max_\delta \sum_{i=1}^{n} \phi_i(w, \delta_i) - \mathbb{I}_{\mathbb{B}_i}(\delta_i) \tag{5}$$

where

$$\phi_i(w, \delta_i) = \frac{1}{n}\ell(x_i + \delta_i, y_i, w).$$

When the dimension of δ or w is large, i.e., the dataset is large or the DNN has a lot of parameters, it may be difficult to directly apply SSI-HG or DSI-HG to the AT problem. Hence, we propose minibatch SI-HG (MSI-HG, Algorithm 3 in Appendix A), which is a combination of the minibatch gradient method (Bottou et al., 2016) and DSI-HG. We also develop a heuristic for solving the implicit step in SI-HGs.

Under the AT setting (5), the minimization form of the implicit step (4) becomes

$$\delta_{i_k}^{k+1} = \arg\min_{\delta_{i_k} \in \mathbb{B}_{i_k}} -\phi_{i_k}(w^{k+1}, \delta_{i_k}) + \frac{1}{2\tau}\|\delta_{i_k} - \delta_{i_k}^k\|^2. \tag{6}$$

Following the intuition behind Section 4 of the work by Goodfellow et al. (Goodfellow et al., 2015), we would like to use sign gradient to update δ. Sign gradient is used in many other AT methods as well (Wang et al., 2019; Zhang et al., 2019a; Mądry et al., 2018; Zhang et al., 2019b). However, the quadratic penalty

Table 3: Accuracy (%) on natural and adversarial examples at the final iteration.

Method	MNIST			SVHN			CIFAR-10		
	Natural	PGD-20	PGD-50-10	Natural	PGD-20	PGD-50-10	Natural	PGD-20	PGD-50-10
PGD AT	$94.68_{\pm 0.21}$	$54.24_{\pm 4.27}$	$44.30_{\pm 4.60}$	$91.31_{\pm 0.55}$	$66.94_{\pm 0.80}$	$66.40_{\pm 0.83}$	$76.19_{\pm 0.25}$	$46.35_{\pm 0.49}$	$45.84_{\pm 0.46}$
DAT	$93.27_{\pm 0.71}$	$16.73_{\pm 6.17}$	$10.00_{\pm 3.87}$	$91.23_{\pm 0.38}$	$61.68_{\pm 0.46}$	$61.04_{\pm 0.52}$	$67.93_{\pm 0.39}$	$34.72_{\pm 0.18}$	$34.18_{\pm 0.18}$
YOPO	$\mathbf{97.73}_{\pm 0.09}$	$24.92_{\pm 12.29}$	$14.63_{\pm 10.09}$	$89.87_{\pm 0.52}$	$46.35_{\pm 2.20}$	$44.71_{\pm 2.24}$	**83.99**	44.72	—
MSI-HG (Ours)	$94.89_{\pm 0.63}$	$\mathbf{62.34}_{\pm \mathbf{3.33}}$	$\mathbf{44.51}_{\pm \mathbf{2.04}}$	$\mathbf{92.52}_{\pm \mathbf{0.11}}$	$\mathbf{68.44}_{\pm \mathbf{0.15}}$	$\mathbf{67.86}_{\pm \mathbf{0.15}}$	$82.04_{\pm 0.13}$	$\mathbf{48.91}_{\pm \mathbf{0.23}}$	$\mathbf{48.27}_{\pm \mathbf{0.20}}$

Table 4: Accuracy (%) on natural and adversarial examples at the moment of best robust accuracy. We do not report accuracies on adversarial examples generated by PGD-50-10, as it was too expensive to run PGD-50-10 at every iteration of training.

Method	MNIST		SVHN		CIFAR-10	
	Natural	PGD-20	Natural	PGD-20	Natural	PGD-20
PGD AT	$94.65_{\pm 0.22}$	$54.71_{\pm 4.23}$	$91.09_{\pm 0.50}$	$67.21_{\pm 0.88}$	$75.83_{\pm 0.36}$	$46.73_{\pm 0.21}$
DAT	$90.96_{\pm 4.88}$	$28.58_{\pm 4.04}$	$91.08_{\pm 0.21}$	$61.94_{\pm 0.26}$	$67.83_{\pm 0.35}$	$34.85_{\pm 0.11}$
YOPO	$\mathbf{97.66}_{\pm 0.05}$	$37.23_{\pm 2.48}$	$89.54_{\pm 0.47}$	$48.19_{\pm 0.84}$	—	—
MSI-HG (Ours)	$94.95_{\pm 0.44}$	$\mathbf{66.31}_{\pm \mathbf{3.41}}$	$\mathbf{92.52}_{\pm \mathbf{0.11}}$	$\mathbf{68.44}_{\pm \mathbf{0.15}}$	$\mathbf{81.63}_{\pm \mathbf{0.23}}$	$\mathbf{49.30}_{\pm \mathbf{0.11}}$

in (6) may be incompatible with sign gradient. To circumvent this problem, we interpret (6) as searching for an adversarial perturbation within the proximity of $\delta_{i_k}^k$. Previous works use the infinity norm to measure the distance between perturbation vectors (Yao et al., 2019; Pooladian et al., 2020). Hence, we solve the surrogate problem

$$\delta_{i_k}^{k+1} = \underset{\delta_{i_k} \in \mathbb{S}_{i_k}}{\arg\min} -\phi_{i_k}(w^{k+1}, \delta_{i_k}) \tag{7}$$

where

$$\mathbb{S}_{i_k} = \mathbb{B}_{i_k} \cap \{\delta_{i_k} : \|\delta_{i_k} - \delta_{i_k}^k\|_\infty \leq \tau\} \tag{8}$$

with T steps of PGD:

$$\delta_{i_k}^{k,0} = \delta_{i_k}^k + \text{unif}[-\tau, \tau]^{d_{i_k}} \tag{9}$$

and

$$\delta_{i_k}^{k,t+1} = \Pi_{\mathbb{S}_{i_k}}[\delta_{i_k}^{k,t} + \eta \cdot \text{sign}(\nabla_{\delta_{i_k}} \phi_{i_k}(w^{k+1}, \delta_{i_k}^{k,t}))] \tag{10}$$

for $t = 0, \ldots, T-1$, and $\delta_{i_k}^{k+1} = \delta_{i_k}^{k,T}$. τ and η are hyperparameters.

5 EXPERIMENTS

We run AT on MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), and CIFAR-10 (Krizhevsky, 2009). We denote the T-step PGD attack with R restarts by PGD-T-R. If $R = 0$, we omit R from the name. The baseline methods are AT methods which have convergence guarantees: PGD AT, DAT, and YOPO. Following common practice, we set the number of inner iterations in PGD AT to 10 (Andriushchenko & Flammarion, 2020; Wang et al., 2019; Zhang et al., 2019a). For SI-HGs, we set $\theta = 1$ and use the smallest choices of $\tau \in (0, 2\epsilon)$ and $T \in \{2, 3, 5, 10\}$ which lead to convergence without harming the robustness. For PGD, DAT, and MSI-HG, the δ update step size η is always set to $2.5\epsilon/T$ so δ may reach and move around on the boundary of the constraint set. For other hyperparameters, we use the recommended settings. All learning curves and statistics are produced by averaging over five random trials.

Training settings. We use a four-layer DNN (3 conv. layers with channels 16, 32, 64, and a final linear layer) on MNIST with $\epsilon = 0.4$, PreActResNet-8 (He et al., 2020) for SVHN with $\epsilon = 4/255$, and PreActResNet-18 on CIFAR-10 with $\epsilon = 8/255$. We use a single A100-SXM4-40GB to train each model. We use batch size 150 on MNIST and batch size 100 on SVHN and CIFAR-10. On each dataset, we combine each AT algorithm with SGD with momentum 0.9 plot its learning curve[3]. On MNIST, each methods is run for 50 epochs. SGD uses a constant learning rate 0.01. On SVHN, each method is run for 15 epochs. SGD uses a triangular learning rate (Smith, 2017) which starts at zero, peaks at epoch 5 with value 0.2, and decays back to zero[4]. On CIFAR-10, each method is run for 30 epochs. SGD uses a cyclic learning rate which starts at zero, peaks at epoch 5 with value 0.2, decays to zero

[3]In Appendix A, we show how MSI-HG is combined with GD or SGD with momentum.

[4]We remark that triangular learning rates have been used in recent works to reduce the training time of AT by up to hundred orders of magnitude (Andriushchenko & Flammarion, 2020; Wong et al., 2020).

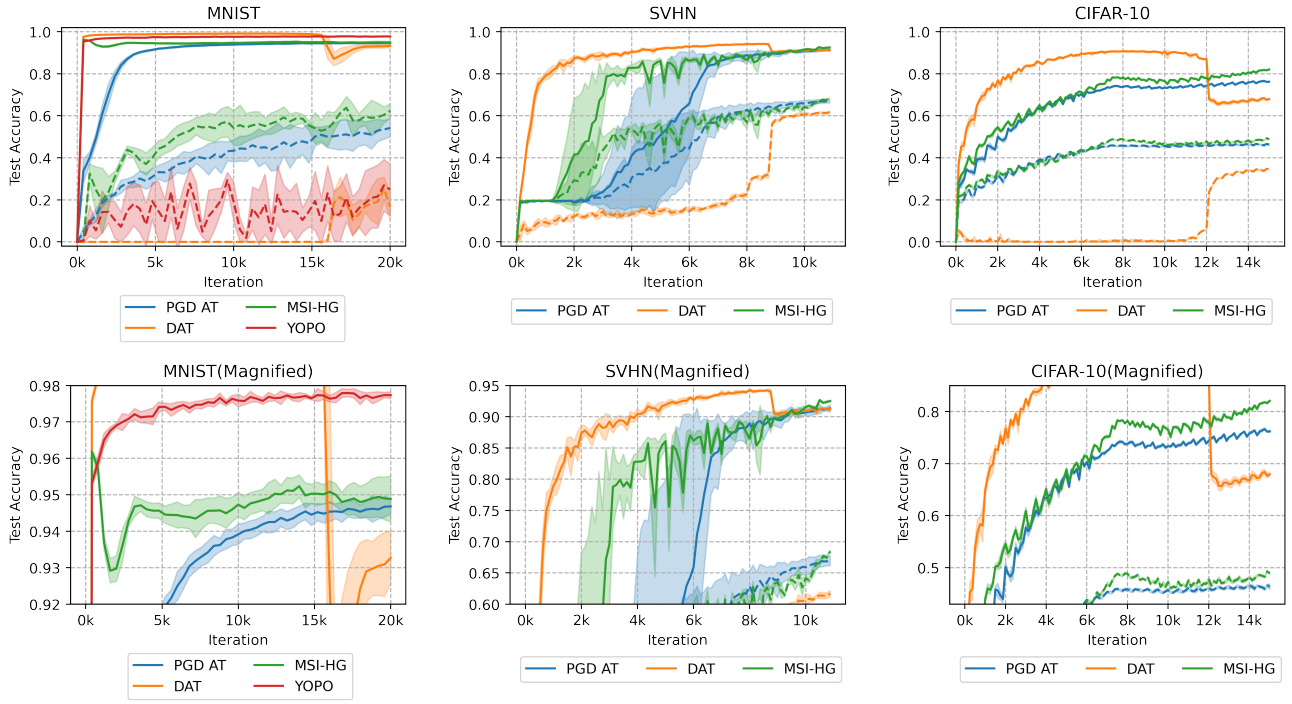

Figure 1: Learning curves of AT algorithms. Solid line denotes natural accuracy and dotted line denotes accuracy on adversarial examples generated by PGD-20. On SVHN, the natural accuracy curves for PGD AT and DAT overlap near the end.

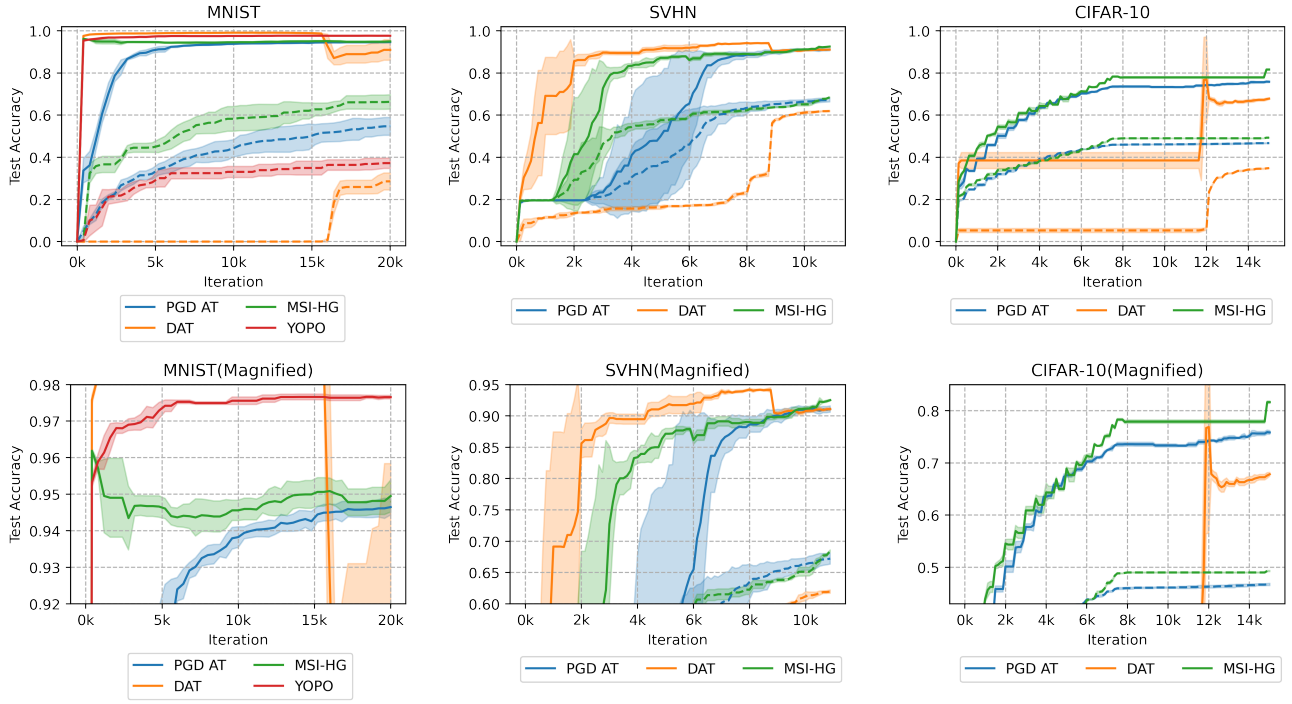

Figure 2: Processed learning curves of AT algorithms. At iteration K, we plot the natural and robust accuracies of the iteration $\leq K$ which had the best robust accuracy. Solid line denotes natural accuracy and dotted line denotes accuracy on adversarial examples generated by PGD-20. On SVHN, the natural accuracy curves for PGD AT and DAT overlap near the end.

at epoch 15, peaks at epoch 20 with value 0.02, and decays to zero at epoch 30. We found that YOPO did not converge under the triangular and cyclic learning rate schedules. Hence, we follow the settings recommended by the authors for YOPO on SVHN experiments. On CIFAR-10, we reuse the best accuracies reported by the authors of YOPO (c.f. Table 2 in (Zhang et al., 2019a)). Further training settings are described in Appendix C.

Experiment results. In Figure 1, we plot the learning curves of AT algorithms. On all datasets, we see acceleration for MSI-HG. For YOPO and DAT, the natural accuracies rise faster than those of MSI-HG, but robust accuracies grow extremely slowly. This is undesirable, since our goal is *faster convergence to adversarially robust DNNs*. On MNIST and SVHN, MSI-HG shows acceleration during the first half of the training process. If we examine the magnified learning curves on CIFAR-10, MSI-HG shows faster convergence than PGD AT at iterations 6k to 8k and 10k to 15k. As a result, the accuracies of PGD AT at epoch 30 is lower than those of MSI-HG at epoch 15. If we process the learning curves to reduce oscillations (Figure 2), acceleration for MSI-HG is even more evident. We note that MSI-HG uses at most ten steps in the inner loop, so its computational cost is similar to or better than that of PGD AT.

We also report natural and robust accuracies in Tables 3 and 4. MSI-HG achieve better robustness than the baseline methods on all datasets[5]. YOPO achieves higher natural accuracy at the cost of lower robustness. On MNIST, robustness for YOPO is especially poor. MSI-HG beat PGD AT and DAT in terms of both natural and robust accuracies.

6 A DISCUSSION ON THE LIMITATIONS OF OUR WORK

6.1 Theoretical Assumptions

Here, we discuss and attempt to justify the theoretical assumptions used in our work.

Assumption 1. Assumption 1 is generally false in the deep learning setting. For instance, the assumption that the coupling function ϕ has Lipschitz continuous gradients is false when we train DNNs which use non-differentiable operations such as ReLU or max-pooling. Still, many works show that algorithms with such theoretical results perform surprisingly well in the deep learning setting (Daskalakis et al., 2018; Gidel et al., 2019; Mertikopoulos et al., 2019; Nouiehed et al., 2019; Chavdarova et al., 2020; Wang et al., 2019; Zhang et al., 2020a). We speculate that this is because the assumptions hold approximately or locally when we train DNNs[6]. For example, a recent work has shown that semi-smoothness, an approximate version of smoothness, holds for overparametrized ReLU DNNs (Allen-Zhu et al., 2018). Thus, even if we have to work under some restrictive assumptions, it is crucial that we continue to develop theoretical grounds for AT methods.

Assumption 2. The relation between MVI and AT is a non-trivial research topic by itself, but we try our best to justify the MVI condition here. First, the MVI condition is already weaker than other assumptions such as pseudo-monotonicity, monotonicity, or coherence (Mertikopoulos et al., 2019). In fact, we use the even weaker weak MVI condition. Second, algorithms developed under the MVI condition have shown good performance when applied to deep learning (Gidel et al., 2019; Mertikopoulos et al., 2019; Liu et al., 2020a). Finally, Liu et al. (2020a) has pointed out that the MVI condition holds while using SGD to learn neural nets for minimization. As nonconvex-nonconcave minimax optimization is in general intractable (Diakonikolas & Daskalakis, 2021), we believe the MVI condition is an adequate choice to develop new AT algorithms in a principled manner.

6.2 Per-Iteration Costs

Due to the full-gradient update in w, the per-iteration cost for SSI-HG is larger than per-iteration costs of doubly-stochastic methods. Hence, it is necessary to establish theoretical results for methods such as MSI-HG. Based on the experiments, we cautiously conjecture that MSI-HG also has a rate better than $O(1/K^{1/2})$. It may be possible to use ideas from our proofs for SSI-HG and DSI-HG to prove the convergence of MSI-HG, but we leave this for future work.

6.3 Memorization of δ

Although we introduced hybrid gradient methods and PGD AT / DAT / YOPO as the same class of algorithm (MGDA), they are different in the aspect that PGD AT / DAT / YOPO does not memorize δ but hybrid gradient methods do. This is because PGD AT / DAT / YOPO randomly initializes δ at every iteration. This may seem like a drawback of the hybrid gradient methods, yet this memory of δ is what allows us to ap-

[5]On CIFAR-10, to generate PGD-20 adversarial examples, the authors of YOPO use $\eta = 2/255$ while we use $\eta = 1/255$. However, PGD-50-10 accuracy for MSI-HG is higher than PGD-20 accuracy for YOPO. Thus, PreActResNet-18 trained by MSI-HG is indeed more robust than those trained by YOPO.

[6]Wang et al. (Wang et al., 2019) also considers this perspective.

ply the momentum-like update (line 3 of Algorithms 1 and 2) and thus obtain better convergence rates. This intuition is reflected in the faster convergence of MSI-HG in the adversarial training (AT) experiments.

7 CONCLUSION

In this work, we introduced SI-HGs to solve nonconvex-nonconcave minimax problems separable in the maximization variable. We proved that SI-HGs achieve the convergence rate $O(1/K)$ which improves upon the convergence rate $O(1/K^{1/2})$ of YOPO and DAT. Our work also improved upon previous convergence results for MGDA. Experiments showed that a practical variant of SI-HGs indeed converges faster and achieves better robustness than other AT methods. Finally, we discussed the limitations of our work, and proposed future directions of research.

We generally expect positive outcomes from this work, since robustness and efficiency are desirable properties of machine learning systems. Adversarially robust DNNs are more likely to be resistant to malicious input manipulations than their naturally trained counterparts. Algorithms which converge fast consume less resource than other algorithms with similar per-iteration costs.

References

Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. Random extrapolation for primal-dual coordinate descent. In Proc. Int. Conf. Mach. Learn., 2020.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parametrization. arXiv preprint arXiv:1811.03962, 2018.

Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. In Neural Info. Proc. Sys., 2020.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Proc. Int. Conf. Mach. Learn., 2018.

Babak Barazandeh and Meisam Razaviyayn. Solving non-convex non-differentiable min-max games using proximal gradient method. In Proc. Int. Conf. Acoust. Speech Sig. Proc., 2020.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci., 2(1):183–202, 2009.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. arXiv preprint arXiv:1606.04838, 2016.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imag. and Vis., 40(1): 120–145, 2011.

Antonin Chambolle, Matthias J. Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. SIAM J. Optim., 28(4):2783–2808, 2018.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. In Neural Info. Proc. Sys., 2020.

Laurent Condat. A primal-dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. J. Optim. Theory and Appl., 158(2):460–479, 2013.

Cong D. Dang and Guanghui Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. Comput. Optim. and Appl., 60(2):277–310, 2014.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In Proc. Int. Conf. Learn. Representations, 2018.

Jelena Diakonikolas and Constantinos Daskalakis. Efficient methods for structured nonconvex-nonconcave min-max optimization. arXiv preprint arXiv:2011.00364, 2021.

Olivier Fercoq and Pascal Bianchi. A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. SIAM J. Optim., 20(1):100–134, 2019.

Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D. Lee. Convergence of adversarial training in overparametrized neural networks. In Neural Info. Proc. Sys., 2019.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, and Pascal Vincent. A variational inequality perspective on generative adversarial networks. In Proc. Int. Conf. Learn. Representations, 2019.

Ian J. Goodfellow, Jonathan Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Proc. Int. Conf. Learn. Representations, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Proc. Eur. Conf. on Comput. Vision, 2020.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In Proc. Int. Conf. Mach. Learn., 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Puya Latafat, Nikolaos M. Freris, and Panagiotis Patrinos. A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. arXiv preprint arXiv:1706:02882v4, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proc. of the IEEE, 86(11):2278–2324, 1998.

Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative advesarial nets. In Proc. Int. Conf. Learn. Representations, 2020a.

Mingrui Liu, Hassan Rafique, Qihang Lin, and Tianbao Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. arXiv preprint arXiv:1810.10207v3, 2020b.

Yura Malitsky. Golden ratio algorithms for variational inequalities. Math. Program., 2019.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In Proc. Int. Conf. Learn. Representations, 2019.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In Proc. Int. Conf. Learn. Representations, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In NIPS Workshop on Deep Learn. and Unsupervised Feature Learn., 2011.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, and Jason D. Lee. Solving a class of non-convex min-max games using iterative first order methods. In Neural Info. Proc. Sys., 2019.

Aram-Alexandre Pooladian, Chris Finlay, Tim Hoheisel, and Adam M. Oberman. A principled approach for generating adversarial images under non-smooth dissimilarity metrics. In Proc. Int. Conf. on Artif. Intell. and Statist., 2020.

Jacob H. Seidman, Mahyar Fazlyab, Victor M. Preciado, and George J. Pappas. Robust deep learning as optimal control: Insights and convergence guarantees. Proc. Mach. Learn. Res., 120:1–10, 2020.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. J. Mach. Learn. Res., 14(1):567–599, 2013.

Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. arXiv preprint arXiv:2007.00753, 2020.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In Proc. Int. Conf. Learn. Representations, 2018.

Leslie N. Smith. Cyclic learning rates for training neural networks. In Proc. IEEE Winter Conf. on Appl. of Comp. Vision, 2017.

Chaobing Song, Zhengyuan Zhou, Yichao Zhou, Yong Jiang, and Yi Ma. Optimistic dual extrapolation for coherent non-monotone variational inequalities. In Neural Info. Proc. Sys., 2020.

Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In Neural Info. Proc. Sys., 2019.

Bảng Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. Advances in Comput. Math., 38(3):667–681, 2013.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In Proc. Int. Conf. Mach. Learn., 2019.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In Proc. Int. Conf. Mach. Learn., 2020.

Yue Xing, Qifan Song, and Guang Cheng. On the generalization properties of adversarial training. In Proc. Int. Conf. Artif. Intell. and Statist., 2020.

Zhewei Yao, Amir Gholami, Peng Xu, Kurt Keutzer, and Michael W. Mahoney. Trust region based adversarial attack on neural networks. In Proc. IEEE Conf. on Comp. Vision and Pattern Recognit., 2019.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In Neural Info. Proc. Sys., 2019a.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Larent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In Proc. Int. Conf. Mach. Learn., 2019b.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In Neural Info. Proc. Sys., 2020a.

Yi Zhang, Orestis Plevrakis, Simon S. Du, Xingguo Li, Zhao Song, and Sanjeev Arora. Over-parametrized adversarial training: An analysis overcoming the curse of dimensionality. In Neural Info. Proc. Sys., 2020b.

Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephan Boyd, and Peter Glynn. Stochastic mirror descent in variationally coherent optimization problems. In Neural Info. Proc. Sys., 2017.

Supplementary Material: Semi-Implicit Hybrid Gradient Methods with Application to Adversarial Robustness

A PSEUDOCODES

Algorithm 3 `MSI-HG`

Input: $(w^{-1}, \delta^{-1}) = (w^0, \delta^0)$, σ, τ.
for $k = 0, 1, 2, \ldots$ **do**
  Jointly shuffle the entries of δ and ϕ.
  for $i = 1, \ldots, n$ **do**
    $\delta_i^{nk+i} = \Pi_{\mathbb{B}_i}[\delta_i^{nk+i-1} + \tau \nabla_{\delta_i} \phi_i(w^{nk+i-1}, \delta_i^{nk+i})]$
    $\delta_j^{nk+i} = \delta_j^{nk+i-1}$ for all $j \neq i$
    $w^{nk+i} = w^{nk+i-1} - \sigma\{2\nabla_w \phi_i(w^{nk+i-1}, \delta_i^{nk+i}) - \nabla_w \phi_{i-1}(w^{nk+i-2}, \delta_{i-1}^{nk+i-1})\}$
  end for
end for

Algorithm 4 `MSI-HG+GD`

Input: $(w^{-1}, \delta^{-1}) = (w^0, \delta^0)$, σ, τ, ρ.
for $k = 0, 1, 2, \ldots$ **do**
  Jointly shuffle the entries of δ and ϕ.
  for $i = 1, \ldots, n$ **do**
    $\delta_i^{nk+i} = \Pi_{\mathbb{B}_i}[\delta_i^{nk+i-1} + \tau \nabla_{\delta_i} \phi_i(w^{nk+i-1}, \delta_i^{nk+i})]$
    $\delta_j^{nk+i} = \delta_j^{nk+i-1}$ for all $j \neq i$
    $\nabla_w^{nk+i-1} = 2\nabla_w \phi_i(w^{nk+i-1}, \delta_i^{nk+i}) - \nabla_w \phi_{i-1}(w^{nk+i-2}, \delta_{i-1}^{nk+i-1})$
    $w^{nk+i} = w^{nk+i-1} - \sigma \cdot \texttt{GD}[\nabla_w^{nk+i-1}, \rho, nk+i]$
  end for
end for

We also use T steps of PGD to approximate line 5 in Algorithms 3 and 4.

B MISSING PROOFS

B.1 Proof of the equivalence between the implicit step and Equation (4)

By the definition of the proximal operator, line 6 of SSI-HG (Algorithm 1) is

$$\delta_{i_k}^{k+1} = \arg\min_{\delta_{i_k}} g_{i_k}(\delta_{i_k}) + \frac{1}{2}\|\delta_{i_k} - \{\delta_{i_k}^k + \tau_k \nabla_{\delta_{i_k}} \phi_{i_k}(w^{k+1}, \delta_{i_k}^{k+1})\}\|_{\tau^{-1}}^2 \tag{11}$$

The optimality condition of (11) is

$$0 \in \partial g_{i_k}(\delta_{i_k}^{k+1}) + \tau^{-1}[\delta_{i_k}^{k+1} - \{\delta_{i_k}^k + \tau_k \nabla_{\delta_{i_k}} \phi_{i_k}(w^{k+1}, \delta_{i_k}^{k+1})\}] \tag{12}$$

which is equivalent to

$$0 \in \partial g_{i_k}(\delta_{i_k}^{k+1}) - \nabla_{\delta_{i_k}} \phi_{i_k}(w^{k+1}, \delta_{i_k}^{k+1}) + \tau^{-1}(\delta_{i_k}^{k+1} - \delta_{i_k}^k). \tag{13}$$

This is the optimality condition for (4).

B.2 One-iteration result for SSI-HG

Define for $k \geq 0$, the filtration and the conditional expectation

$$\mathcal{F}_0 = \emptyset, \qquad \mathcal{F}_k = \{i_0, \ldots, i_{k-1}\}, \qquad \mathbb{E}_k = \mathbb{E}[\cdot \mid \mathcal{F}_k].$$

We also define the following representation of SSI-HG with full dimensional updates

$$\begin{aligned}
q^k &= \nabla_w \phi(w^{k-1}, \delta^{k-1}) - (n-1)\{\nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^k, \delta^{k-1})\},\\
w^{k+1} &= \operatorname{prox}_f^{\sigma_k}[w^k - \sigma\{\nabla_w \phi(w^k, \delta^k) + \theta_k(\nabla_w \phi(w^k, \delta^k) - q^k)\}],\\
\hat{\delta}_i^{k+1} &= \operatorname{prox}_{g_i}^{\tau}[\delta_i^k + \tau \nabla_{\delta_i} \phi_i(w^{k+1}, \hat{\delta}_i^{k+1})].
\end{aligned}$$

The optimality conditions for w^{k+1} and $\hat{\delta}_i^{k+1}$ are

$$0 = \gamma_f^{k+1} + \sigma^{-1}[w^{k+1} - w^k + \sigma\{\nabla_w \phi(w^k, \delta^k) + \theta(\nabla_w \phi(w^k, \delta^k) - q^k)\}] \tag{14}$$

$$0 = \hat{\gamma}_{g_i}^{k+1} + \tau^{-1}(\hat{\delta}_i^{k+1} - \delta_i^k - \tau \nabla_{\delta_i} \phi_i(w^{k+1}, \hat{\delta}_i^{k+1})) \tag{15}$$

for some $\gamma_f^{k+1} \in \partial f(w^{k+1})$ and $\hat{\gamma}_{g_i}^{k+1} \in \partial g_i(\hat{\delta}_i^{k+1})$. Note that

$$\hat{\gamma}_g^{k+1} \coloneqq (\hat{\gamma}_{g_1}^{k+1}, \ldots, \hat{\gamma}_{g_n}^{k+1}) \in \partial g(\hat{\delta}^{k+1})$$

so from (15),

$$0 = \hat{\gamma}_g^{k+1} + \tau^{-1}(\hat{\delta}^{k+1} - \delta^k - \tau \nabla_\delta \phi(w^{k+1}, \hat{\delta}^{k+1})). \tag{16}$$

We also note that the optimality condition for $\delta_{i_k}^{k+1}$ is

$$0 = \gamma_{g_{i_k}}^{k+1} + \tau^{-1}(\delta_{i_k}^{k+1} - \delta_{i_k}^k - \tau \nabla_{\delta_{i_k}} \phi_{i_k}(w^{k+1}, \delta_{i_k}^{k+1})) \tag{17}$$

for some $\gamma_{g_{i_k}}^{k+1} \in \partial g_{i_k}(\delta_{i_k}^{k+1})$. We start with some technical Lemmas.

Lemma 5. *For any $\delta \in \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_n}$ and any function $r(\delta) = \sum_{i=1}^n r_i(\delta_i)$,*

$$r(\hat{\delta}^{k+1}) = \mathbb{E}_k r(\delta^{k+1}) + (n-1)\{\mathbb{E}_k r(\delta^{k+1}) - r(\delta^k)\} \tag{18}$$

$$\|\delta - \hat{\delta}^{k+1}\|_{\tau^{-1}}^2 = n\mathbb{E}_k\|\delta - \delta^{k+1}\|_{\tau^{-1}}^2 - n\|\delta - \delta^k\|_{\tau^{-1}}^2 + \|\delta - \delta^k\|_{\tau^{-1}}^2, \tag{19}$$

$$\|\hat{\delta}^{k+1} - \delta^k\|_{\tau^{-1}}^2 = n\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2. \tag{20}$$

Proof. Let us observe that

$$\mathbb{E}_k r_i(\delta_i^{k+1}) = \frac{1}{n} r_i(\hat{\delta}_i^{k+1}) + \left(1 - \frac{1}{n}\right) r_i(\delta_i^k).$$

Summing the above equation over i, multiplying both sides by n, and rearranging the terms yields (18). Using $r(\delta) = \|\delta - \hat{\delta}^{k+1}\|_{\tau^{-1}}^2$ yields (19). Finally, plugging in $\delta = \delta^k$ into (19) yields (20). □

Lemma 6. *Assume $\sigma > 0$, $\tau > 0$, $\theta \in (0, 1]$, and define*

$$\kappa = \max\{L_{12}(\sigma\tau n)^{1/2}, L_{11}\sigma\}.$$

We then have

$$\begin{aligned}
&|\langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - q^k\rangle| \\
&\leq \kappa\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + \frac{\kappa}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 + \frac{n\kappa}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2 \qquad (21)\\
&\leq \frac{\kappa}{\theta}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + \frac{\kappa}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 + \frac{n\kappa}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2. \qquad (22)
\end{aligned}$$

Proof. We note that since $\theta \in (0, 1]$, it suffices to prove (21). By the definition of q^k,

$$\begin{aligned}
&\langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - q^k\rangle \\
&= \langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) + (n-1)\{\nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^k, \delta^{k-1})\} - \nabla_w \phi(w^{k-1}, \delta^{k-1})\rangle \\
&= n\langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^k, \delta^{k-1})\rangle + \langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^{k-1}) - \nabla_w \phi(w^{k-1}, \delta^{k-1})\rangle. \qquad (23)
\end{aligned}$$

We now bound the two inner products in (23). The first inner product can be bounded as

$$\begin{aligned}
&\kappa^{-1}|\langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^k, \delta^{k-1})\rangle| \\
&\leq (L_{12}^2\sigma\tau n)^{-1/2}|\langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^k, \delta^{k-1})\rangle| \\
&\leq (L_{12}^2\sigma\tau n)^{-1/2}\|w^{k+1} - w^k\|\|\nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^k, \delta^{k-1})\| \\
&\leq (\sigma\tau n)^{-1/2}\|w^{k+1} - w^k\|\|\delta^k - \delta^{k-1}\| \\
&\leq \frac{1}{2n}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + \frac{1}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2 \qquad (24)
\end{aligned}$$

where we have used the definition of κ at the first inequality, Cauchy-Schwarz inequality at the second inequality, Lipschitz continuity of $\nabla_w \phi$ at the third inequality, and Young's inequality at the last inequality.

The second inner product can be bounded as

$$\begin{aligned}
&\kappa^{-1}|\langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^{k-1}) - \nabla_w \phi(w^{k-1}, \delta^{k-1})\rangle| \\
&\leq (L_{11}^2\sigma^2)^{-1/2}|\langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^{k-1}) - \nabla_w \phi(w^{k-1}, \delta^{k-1})\rangle| \\
&\leq (L_{11}^2\sigma^2)^{-1/2}\|w^{k+1} - w^k\|\|\nabla_w \phi(w^k, \delta^{k-1}) - \nabla_w \phi(w^{k-1}, \delta^{k-1})\| \\
&\leq (\sigma^2)^{-1/2}\|w^{k+1} - w^k\|\|w^k - w^{k-1}\| \\
&\leq \frac{1}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + \frac{1}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \qquad (25)
\end{aligned}$$

where we have used the definition of κ at the first inequality, Cauchy-Schwarz inequality at the second inequality, Lipschitz continuity of $\nabla_w \phi$ at the third inequality, and Young's inequality at the last inequality.

It follows that

$$\begin{aligned}
&|\langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - q^k\rangle| \\
&\leq n|\langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - \nabla_x \phi(w^k, \delta^{k-1})\rangle| \\
&\quad + |\langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^{k-1}) - \nabla_w \phi(w^{k-1}, \delta^{k-1})\rangle| \\
&\leq \kappa\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + \frac{\kappa}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 + \frac{n\kappa}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2
\end{aligned}$$

where we have applied the triangle inequality to (23) at the first inequality and have used (24) and (25) at the second inequality. □

Lemma 7 (One-Iteration Result)**.** *Assume $\sigma > 0$, $\tau > 0$, $\theta > 0$. We then have for any (w, δ),*

$$
\begin{aligned}
0 \geq & \begin{bmatrix} \gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \\ \hat{\gamma}_g^{k+1} - \nabla_\delta \phi(w^{k+1}, \hat{\delta}^{k+1}) \end{bmatrix}^\top \begin{bmatrix} w^{k+1} - w \\ \hat{\delta}^{k+1} - \delta \end{bmatrix} \\
& + \mathbb{E}_k \langle w - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1} \rangle - \theta \langle w - w^k, \nabla_w \phi(w^k, \delta^k) - q^k \rangle \\
& + \frac{1}{2} \|w - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{(1-2\kappa)}{2} \|w^{k+1} - w^k\|_{\sigma^{-1}}^2 - \frac{1}{2} \|w - w^k\|_{\sigma^{-1}}^2 - \frac{\kappa\theta}{2} \|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \\
& + \frac{n}{2} \mathbb{E}_k \|\delta - \delta^{k+1}\|_{\tau^{-1}}^2 + \frac{n}{2} \mathbb{E}_k \|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 - \frac{n}{2} \|\delta - \delta^k\|_{\tau^{-1}}^2 - \frac{n\kappa\theta}{2} \|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2.
\end{aligned}
$$

Proof. Optimality condition (14) implies that

$$
\begin{aligned}
0 &= \langle w^{k+1} - w, \gamma_f^{k+1} + \sigma^{-1}[w^{k+1} - w^k + \sigma\{\nabla_w \phi(w^k, \delta^k) + \theta(\nabla_w \phi(w^k, \delta^k) - q^k)\}] \rangle \\
&= \langle w^{k+1} - w, \gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \rangle - \langle w - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \rangle \\
&\quad - \theta \langle w - w^k, \nabla_w \phi(w^k, \delta^k) - q^k \rangle - \theta \langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - q^k \rangle \\
&\quad + \sigma^{-1} \langle w^{k+1} - w, w^{k+1} - w^k \rangle \\
&= \langle w^{k+1} - w, \gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \rangle - \langle w - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \rangle \\
&\quad - \theta \langle w - w^k, \nabla_w \phi(w^k, \delta^k) - q^k \rangle - \theta \langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - q^k \rangle \\
&\quad + \frac{1}{2} \|w - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{1}{2} \|w^{k+1} - w^k\|_{\sigma^{-1}}^2 - \frac{1}{2} \|w - w^k\|_{\sigma^{-1}}^2.
\end{aligned} \tag{26}
$$

By (18) with $r(\delta) = \nabla_w \phi(w^{k+1}, \delta)$, we have

$$
\begin{aligned}
&\nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \\
&= \nabla_w \phi(w^k, \delta^k) - (n-1)\{\mathbb{E}_k \nabla_w \phi(w^{k+1}, \delta^{k+1}) - \nabla_w \phi(w^{k+1}, \delta^k)\} - \mathbb{E}_k \nabla_w \phi(w^{k+1}, \delta^{k+1}) \\
&= \mathbb{E}_k q^{k+1} - \mathbb{E}_k \nabla_w \phi(w^{k+1}, \delta^{k+1})
\end{aligned} \tag{27}
$$

and by plugging this into (26) and using linearity of expectation, we obtain

$$
\begin{aligned}
0 &= \langle w^{k+1} - w, \gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \rangle + \mathbb{E}_k \langle w - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1} \rangle \\
&\quad - \theta \langle w - w^k, \nabla_w \phi(w^k, \delta^k) - q^k \rangle - \theta \langle w^k - w^{k+1}, \nabla_w \phi(w^k, \delta^k) - q^k \rangle \\
&\quad + \frac{1}{2} \|w - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{1}{2} \|w^{k+1} - w^k\|_{\sigma^{-1}}^2 - \frac{1}{2} \|w - w^k\|_{\sigma^{-1}}^2.
\end{aligned} \tag{28}
$$

Applying (22) of Lemma 6 to (28), we have

$$
\begin{aligned}
0 \geq & \langle w^{k+1} - w, \gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \rangle + \mathbb{E}_k \langle w - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1} \rangle \\
& - \theta \langle w - w^k, \nabla_w \phi(w^k, \delta^k) - q^k \rangle + \frac{1}{2} \|w - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{(1-2\kappa)}{2} \|w^{k+1} - w^k\|_{\sigma^{-1}}^2 - \frac{1}{2} \|w - w^k\|_{\sigma^{-1}}^2 \\
& - \frac{\kappa\theta}{2} \|w^k - w^{k-1}\|_{\sigma^{-1}}^2 - \frac{n\kappa\theta}{2} \|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2.
\end{aligned} \tag{29}
$$

Optimality condition (16) implies that

$$
\begin{aligned}
0 &= \langle \hat{\delta}^{k+1} - \delta, \hat{\gamma}_g^{k+1} + \tau^{-1}(\hat{\delta}^{k+1} - \delta^k - \tau \nabla_\delta \phi(w^{k+1}, \hat{\delta}^{k+1})) \rangle \\
&= \langle \hat{\delta}^{k+1} - \delta, \hat{\gamma}_g^{k+1} - \nabla_\delta \phi(w^{k+1}, \hat{\delta}^{k+1}) \rangle + \tau^{-1} \langle \hat{\delta}^{k+1} - \delta, \hat{\delta}^{k+1} - \delta^k \rangle \\
&= \langle \hat{\delta}^{k+1} - \delta, \hat{\gamma}_g^{k+1} - \nabla_\delta \phi(w^{k+1}, \hat{\delta}^{k+1}) \rangle + \frac{1}{2} \|\delta - \hat{\delta}^{k+1}\|_{\tau^{-1}}^2 + \frac{1}{2} \|\hat{\delta}^{k+1} - \delta^k\|_{\tau^{-1}}^2 - \frac{1}{2} \|\delta - \delta^k\|_{\tau^{-1}}^2 \\
&= \langle \hat{\delta}^{k+1} - \delta, \hat{\gamma}_g^{k+1} - \nabla_\delta \phi(w^{k+1}, \hat{\delta}^{k+1}) \rangle \\
&\quad + \frac{n}{2} \mathbb{E}_k \|\delta - \delta^{k+1}\|_{\tau^{-1}}^2 + \frac{n}{2} \mathbb{E}_k \|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 - \frac{n}{2} \|\delta - \delta^k\|_{\tau^{-1}}^2
\end{aligned} \tag{30}
$$

where we have used (19) and (20) at the last equality. Adding (29) and (30) concludes the proof. □

B.3 Proof of Theorem 1

Lemma 8. *Let $\theta = 1$. We then have*

$$\left\| \begin{bmatrix} \gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \\ \hat{\gamma}_g^{k+1} - \nabla_\delta \phi(x^{k+1}, \hat{\delta}^{k+1}) \end{bmatrix} \right\|^2 \leq 3(\sigma^{-1} + 2L_{11}^2\sigma)\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + 6L_{11}^2\sigma\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 + n(\tau^{-1} + 6nL_{12}^2\tau)\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 + 6n^2L_{12}^2\tau\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2.$$

Proof. First, observe that

$$\begin{aligned} &\|\gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1})\|^2 \\ &= \|\sigma^{-1}(-w^{k+1} + w^k) + \mathbb{E}_k\{\nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1}\} - \{\nabla_w \phi(w^k, \delta^k) - q^k\}\|^2 \\ &\leq 3\sigma^{-1}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + 3\|\mathbb{E}_k\{\nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1}\}\|^2 + 3\|\nabla_w \phi(w^k, \delta^k) - q^k\|^2 \\ &\leq 3\sigma^{-1}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + 3\mathbb{E}_k\|\nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1}\|^2 + 3\|\nabla_w \phi(w^k, \delta^k) - q^k\|^2 \end{aligned} \tag{31}$$

where we have used the combination of (14) and (27) at the first equality and Jensen's inequality at the first and second inequalities. We have

$$\begin{aligned} &\|\nabla_w \phi(w^k, \delta^k) - q^k\|^2 \\ &= \|\nabla_w \phi(w^k, \delta^k) + (n-1)\{\nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^k, \delta^{k-1})\} - \nabla_w \phi(w^{k-1}, \delta^{k-1})\|^2 \\ &= \|n\{\nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^k, \delta^{k-1})\} + \{\nabla_w \phi(w^k, \delta^{k-1}) - \nabla_w \phi(w^{k-1}, \delta^{k-1})\}\|^2 \\ &\leq 2n^2\|\nabla_w \phi(w^k, \delta^k) - \nabla_w \phi(w^k, \delta^{k-1})\|^2 + 2\|\nabla_w \phi(w^k, \delta^{k-1}) - \nabla_w \phi(w^{k-1}, \delta^{k-1})\|^2 \\ &\leq 2n^2L_{12}^2\|\delta^k - \delta^{k-1}\|^2 + 2L_{11}^2\|w^k - w^{k-1}\|^2 \\ &= 2n^2L_{12}^2\tau\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2 + 2L_{11}^2\sigma\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \end{aligned} \tag{32}$$

where we have used the definition of q^k at the first equality, Jensen's inequality at the first inequality, and Lipschitz continuity of $\nabla_w \phi(w, \delta)$ at the second inequality. Similarly,

$$\|\nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1}\|^2 \leq 2n^2L_{12}^2\tau\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 + 2L_{11}^2\sigma\|w^{k+1} - w^k\|_{\sigma^{-1}}^2. \tag{33}$$

Applying (32) and (33) to (31), we obtain

$$\begin{aligned} \|\gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1})\|^2 \leq 3(\sigma^{-1} + 2L_{11}^2\sigma)\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + 6L_{11}^2\sigma\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \\ + 6n^2L_{12}^2\tau\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 + 6n^2L_{12}^2\tau\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2. \end{aligned} \tag{34}$$

We also have

$$\begin{aligned} \|\hat{\gamma}_g^{k+1} - \nabla_\delta \phi(w^{k+1}, \hat{\delta}^{k+1})\|^2 &= \|\tau^{-1}(-\hat{\delta}^{k+1} + \delta^k)\|^2 \\ &= \tau^{-1}\|\hat{\delta}^{k+1} - \delta^k\|_{\tau^{-1}}^2 \\ &= n\tau^{-1}\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 \end{aligned} \tag{35}$$

where we have used (16) at the first equality and (20) at the third equality. Combining (34) and (35) concludes the proof. □

Theorem 1. *Suppose Assumptions 1 and 2 are true. Let $\{(w^k, \delta^k)\}$ be the sequence generated by SSI-HG, and define the full-dimensional update (which only depends on w^k and δ^{k-1})*

$$\hat{\delta}^k = \mathrm{prox}_g^\tau[\delta^{k-1} + \tau \nabla_\delta \phi(w^k, \hat{\delta}^k)].$$

If $\theta = 1$ and $\rho > 0$, $\sigma > 0$, $\tau > 0$ satisfy

$$\max\{L_{12}(\sigma\tau n)^{1/2}, L_{11}\sigma\} < \min\{1/3 - \rho(\sigma^{-1} + 4L_{11}^2\sigma), 1 - \rho(\tau^{-1} + 12nL_{12}^2\tau)\},$$

we have

$$\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\|F(w^k, \hat{\delta}^k)\|^2 = O(1/K).$$

Proof. Let (w^*, δ^*) be a solution of the weak MVI problem (which exists by Assumption 2). Then

$$\begin{aligned}
0 \geq & \begin{bmatrix} \gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \\ \hat{\gamma}_g^{k+1} - \nabla_\delta \phi(w^{k+1}, \hat{\delta}^{k+1}) \end{bmatrix}^\top \begin{bmatrix} w^{k+1} - w^* \\ \hat{\delta}^{k+1} - \delta^* \end{bmatrix} \\
& + \mathbb{E}_k \langle w^* - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1} \rangle - \langle w^* - w^k, \nabla_w \phi(w^k, \delta^k) - q^k \rangle \\
& + \frac{1}{2}\|w^* - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{(1-2\kappa)}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 - \frac{1}{2}\|w^* - w^k\|_{\sigma^{-1}}^2 - \frac{\kappa}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \\
& + \frac{n}{2}\mathbb{E}_k\|\delta^* - \delta^{k+1}\|_{\tau^{-1}}^2 + \frac{n}{2}\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 - \frac{n}{2}\|\delta^* - \delta^k\|_{\tau^{-1}}^2 - \frac{n\kappa}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2 \\
\geq & -\frac{\rho}{2}\left\| \begin{bmatrix} \gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \\ \hat{\gamma}_g^{k+1} - \nabla_y \phi(w^{k+1}, \hat{\delta}^{k+1}) \end{bmatrix} \right\|^2 \\
& + \mathbb{E}_k \langle w^* - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1} \rangle - \langle w^* - w^k, \nabla_w \phi(w^k, \delta^k) - q^k \rangle \\
& + \frac{1}{2}\|w^* - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{(1-2\kappa)}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 - \frac{1}{2}\|w^* - w^k\|_{\sigma^{-1}}^2 - \frac{\kappa}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \\
& + \frac{n}{2}\mathbb{E}_k\|\delta^* - \delta^{k+1}\|_{\tau^{-1}}^2 + \frac{n}{2}\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 - \frac{n}{2}\|\delta^* - \delta^k\|_{\tau^{-1}}^2 - \frac{n\kappa}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2 \\
\geq & \mathbb{E}_k \langle w^* - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1} \rangle - \langle w^* - w^k, \nabla_w \phi(w^k, \delta^k) - q^k \rangle \\
& + \frac{1}{2}\|w^* - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{\{1 - 2\kappa - 3\rho(\sigma^{-1} + 2L_{11}^2\sigma)\}}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 \\
& - \frac{1}{2}\|w^* - w^k\|_{\sigma^{-1}}^2 - \frac{\{\kappa + \rho(6L_{11}^2\sigma)\}}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \\
& + \frac{n}{2}\mathbb{E}_k\|\delta^* - \delta^{k+1}\|_{\tau^{-1}}^2 + \frac{n\{1 - \rho(\tau^{-1} + 6nL_{12}^2\tau)\}}{2}\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 \\
& - \frac{n}{2}\|\delta^* - \delta^k\|_{\tau^{-1}}^2 - \frac{n\{\kappa + \rho(6nL_{12}^2\tau)\}}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2
\end{aligned} \tag{36}$$

where we have used Lemma 7 with $\theta = 1$ at the first inequality, Assumption 2 at the second inequality, and Lemma 8 at the third inequality. Rearranging and adding and subtracting some terms in (36), we obtain

$$\begin{aligned}
& \langle w^* - w^k, \nabla_w \phi(w^k, \delta^k) - q^k \rangle + \frac{1}{2}\|w^* - w^k\|_{\sigma^{-1}}^2 + \frac{n}{2}\|\delta^* - \delta^k\|_{\tau^{-1}}^2 \\
& \quad + \frac{\{\kappa + \rho(6L_{11}^2\sigma)\}}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 + \frac{n\{\kappa + \rho(6nL_{12}^2\tau)\}}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2 \\
& \geq \mathbb{E}_k \langle w^* - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1} \rangle + \frac{1}{2}\|w^* - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{n}{2}\mathbb{E}_k\|\delta^* - \delta^{k+1}\|_{\tau^{-1}}^2 \\
& \quad + \frac{\{\kappa + \rho(6L_{11}^2\sigma)\}}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + \frac{n\{\kappa + \rho(6nL_{12}^2\tau)\}}{2}\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 \\
& \quad + \frac{\{1 - 3\kappa - 3\rho(\sigma^{-1} + 4L_{11}^2\sigma)\}}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 \\
& \quad + \frac{n\{1 - \kappa - \rho(\tau^{-1} + 12nL_{12}^2\tau)\}}{2}\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2.
\end{aligned} \tag{37}$$

Taking full expectation over (37), summing both sides over $k = 0, \dots, K-1$, and using $(w^{-1}, \delta^{-1}) = (w^0, \delta^0)$, we obtain

$$\begin{aligned}
&\frac{1}{2}\|w^* - w^0\|_{\sigma^{-1}}^2 + \frac{n}{2}\|\delta^* - \delta^0\|_{\tau^{-1}}^2 \\
&\geq \mathbb{E}\langle w^* - w^K, \nabla_w \phi(w^K, \delta^K) - q^K\rangle + \frac{1}{2}\|w^* - w^K\|_{\sigma^{-1}}^2 + \frac{n}{2}\mathbb{E}\|\delta^* - \delta^K\|_{\tau^{-1}}^2 \\
&\quad + \frac{\{\kappa + \rho(6L_{11}^2\sigma)\}}{2}\mathbb{E}\|w^K - w^{K-1}\|_{\sigma^{-1}}^2 + \frac{n\{\kappa + \rho(6nL_{12}^2\tau)\}}{2}\mathbb{E}\|\delta^K - \delta^{K-1}\|_{\tau^{-1}}^2 \\
&\quad + \frac{\{1 - 3\kappa - 3\rho(\sigma^{-1} + 4L_{11}^2\sigma)\}}{2}\sum_{k=0}^{K-1}\mathbb{E}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 \\
&\quad + \frac{n\{1 - \kappa - \rho(\tau^{-1} + 12nL_{12}^2\tau)\}}{2}\sum_{k=0}^{K-1}\mathbb{E}\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2. \qquad (38)
\end{aligned}$$

By reasoning similarly as the proof of Lemma 6, we have

$$|\langle w^* - w^K, \nabla_w \phi(w^K, \delta^K) - q^K\rangle| \leq \kappa\|w^* - w^K\|_{\sigma^{-1}}^2 + \frac{\kappa}{2}\|w^K - w^{K-1}\|_{\sigma^{-1}}^2 + \frac{n\kappa}{2}\|\delta^K - \delta^{K-1}\|_{\tau^{-1}}^2 \qquad (39)$$

and so (38) implies

$$\begin{aligned}
&\frac{1}{2}\|w^* - w^0\|_{\sigma^{-1}}^2 + \frac{n}{2}\|\delta^* - \delta^0\|_{\tau^{-1}}^2 \\
&\geq \frac{(1 - 2\kappa)}{2}\|w^* - w^K\|_{\sigma^{-1}}^2 + \frac{n}{2}\mathbb{E}\|\delta^* - \delta^K\|_{\tau^{-1}}^2 \\
&\quad + \frac{\rho(6L_{11}^2\sigma)}{2}\mathbb{E}\|w^K - w^{K-1}\|_{\sigma^{-1}}^2 + \frac{\rho(6n^2L_{12}^2\tau)}{2}\mathbb{E}\|\delta^K - \delta^{K-1}\|_{\tau^{-1}}^2 \\
&\quad + \frac{\{1 - 3\kappa - 3\rho(\sigma^{-1} + 4L_{11}^2\sigma)\}}{2}\sum_{k=0}^{K-1}\mathbb{E}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 \\
&\quad + \frac{n\{1 - \kappa - \rho(\tau^{-1} + 12nL_{12}^2\tau)\}}{2}\sum_{k=0}^{K-1}\mathbb{E}\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 \qquad (40)
\end{aligned}$$

All the coefficients for the quadratic terms in (40) are positive by the definition of κ (in Lemma 6) and the step-size conditions. Hence, we may remove the first four quadratic terms at the RHS of (40) to obtain

$$\begin{aligned}
&\frac{1}{2}\|w^* - w^0\|_{\sigma^{-1}}^2 + \frac{n}{2}\|\delta^* - \delta^0\|_{\tau^{-1}}^2 \\
&\geq \frac{\{1 - 3\kappa - 3\rho(\sigma^{-1} + 4L_{11}^2\sigma)\}}{2}\sum_{k=0}^{K-1}\mathbb{E}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 \\
&\quad + \frac{n\{1 - \kappa - \rho(\tau^{-1} + 12nL_{12}^2\tau)\}}{2}\sum_{k=0}^{K-1}\mathbb{E}\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2. \qquad (41)
\end{aligned}$$

This proves that

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|w^{k+1} - w^k\|^2 = O(1/K), \qquad \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\delta^{k+1} - \delta^k\|^2 = O(1/K). \qquad (42)$$

By Lemma 8 and the definition of the saddle subdifferential norm,

$$\begin{aligned}
\mathbb{E}\|F(w^{k+1}, \hat{\delta}^{k+1})\|^2 &\leq 3(\sigma^{-1} + 2L_{11}^2\sigma)\mathbb{E}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + 6L_{11}^2\sigma\mathbb{E}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \\
&\quad + n(\tau^{-1} + 6nL_{12}^2\tau)\mathbb{E}\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 + 6n^2L_{12}^2\tau\mathbb{E}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2. \qquad (43)
\end{aligned}$$

Averaging (43) over $k = 0, \dots, K-1$ and using (42) concludes the proof. □

B.4 Proof of Theorem 2

Theorem 2. *Suppose Assumptions 1 and 3 are true. Let $\{(w^k, \delta^k)\}$ be the sequence generated by SSI-HG. If $\theta > 0$, $\sigma > 0$, $\tau > 0$ satisfy*

$$\max\{L_{12}(\sigma\tau n)^{1/2}, L_{11}\sigma\} \le 1/3, \qquad \theta = \max\left\{\frac{1}{1+\mu\sigma}, \frac{1+(n-1)\mu\tau/n}{1+\mu\tau}\right\},$$

we have

$$\mathbb{E}\|w^* - w^K\|^2 = O(\theta^K), \qquad \mathbb{E}\|\delta^* - \delta^K\|^2 = O(\theta^K).$$

Proof. Let (w^*, δ^*) be a solution of the strong MVI problem (which exists by Assumption 3). Then

$$\begin{aligned}
0 \ge & \begin{bmatrix} \gamma_f^{k+1} + \nabla_w \phi(w^{k+1}, \hat{\delta}^{k+1}) \\ \hat{\gamma}_g^{k+1} - \nabla_y \phi(w^{k+1}, \hat{\delta}^{k+1}) \end{bmatrix}^\top \begin{bmatrix} w^{k+1} - w^* \\ \hat{\delta}^{k+1} - \delta^* \end{bmatrix} \\
& + \mathbb{E}_k\langle w^* - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1}\rangle - \theta\langle w^* - w^k, \nabla_w \phi(w^k, \delta^k) - q^k\rangle \\
& + \frac{1}{2}\|w^* - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{(1-2\kappa)}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 - \frac{1}{2}\|w^* - w^k\|_{\sigma^{-1}}^2 - \frac{\kappa\theta}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \\
& + \frac{n}{2}\mathbb{E}_k\|\delta^* - \delta^{k+1}\|_{\tau^{-1}}^2 + \frac{n}{2}\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 - \frac{n}{2}\|\delta^* - \delta^k\|_{\tau^{-1}}^2 - \frac{n\kappa\theta}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2 \\
\ge & \frac{\mu}{2}\left\|\begin{bmatrix} w^{k+1} - w^* \\ \hat{\delta}^{k+1} - \delta^* \end{bmatrix}\right\|^2 \\
& + \mathbb{E}_k\langle w^* - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1}\rangle - \theta\langle w^* - w^k, \nabla_w \phi(w^k, \delta^k) - q^k\rangle \\
& + \frac{1}{2}\|w^* - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{(1-2\kappa)}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 - \frac{1}{2}\|w^* - w^k\|_{\sigma^{-1}}^2 - \frac{\kappa\theta}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \\
& + \frac{n}{2}\mathbb{E}_k\|\delta^* - \delta^{k+1}\|_{\tau^{-1}}^2 + \frac{n}{2}\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 - \frac{n}{2}\|\delta^* - \delta^k\|_{\tau^{-1}}^2 - \frac{n\kappa\theta}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2 \\
= & \mathbb{E}_k\langle w^* - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1}\rangle - \theta\langle w^* - w^k, \nabla_w \phi(w^k, \delta^k) - q^k\rangle \\
& + \frac{(1+\mu\sigma)}{2}\|w^* - w^{k+1}\|_{\sigma^{-1}}^2 + \frac{(1-2\kappa)}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 - \frac{1}{2}\|w^* - w^k\|_{\sigma^{-1}}^2 - \frac{\kappa\theta}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 \\
& + \frac{n(1+\mu\tau)}{2}\mathbb{E}_k\|\delta^* - \delta^{k+1}\|_{\tau^{-1}}^2 + \frac{n}{2}\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 \\
& - \frac{n\{1+(n-1)\mu\tau/n\}}{2}\|\delta^* - \delta^k\|_{\tau^{-1}}^2 - \frac{n\kappa\theta}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2
\end{aligned} \tag{44}$$

where we have used Lemma 7 at the first inequality, Assumption 3 at the second inequality, and (19) at the first equality. Rearranging and adding and subtracting some terms in (44), we obtain

$$\begin{aligned}
& \theta\langle w^* - w^k, \nabla_w \phi(w^k, \delta^k) - q^k\rangle + \frac{1}{2}\|w^* - w^k\|_{\sigma^{-1}}^2 + \frac{n\{1+(n-1)\mu\tau/n\}}{2}\|\delta^* - \delta^k\|_{\tau^{-1}}^2 \\
& \quad + \frac{\kappa\theta}{2}\|w^k - w^{k-1}\|_{\sigma^{-1}}^2 + \frac{n\kappa\theta}{2}\|\delta^k - \delta^{k-1}\|_{\tau^{-1}}^2 \\
& \ge \mathbb{E}_k\langle w^* - w^{k+1}, \nabla_w \phi(w^{k+1}, \delta^{k+1}) - q^{k+1}\rangle + \frac{(1+\mu\sigma)}{2}\|w^* - w^{k+1}\|_{\sigma^{-1}}^2 \\
& \quad + \frac{n(1+\mu\tau)}{2}\mathbb{E}_k\|\delta^* - \delta^{k+1}\|_{\tau^{-1}}^2 + \frac{\kappa}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + \frac{n\kappa}{2}\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2 \\
& \quad + \frac{(1-3\kappa)}{2}\|w^{k+1} - w^k\|_{\sigma^{-1}}^2 + \frac{n(1-\kappa)}{2}\mathbb{E}_k\|\delta^{k+1} - \delta^k\|_{\tau^{-1}}^2.
\end{aligned} \tag{45}$$

All the coefficients for the quadratic terms in (45) are non-negative by the definition of κ (in Lemma 6) and the step-size conditions. Hence, we may remove the last two quadratic terms at the RHS of (45) and take full

expectation to obtain

$$\begin{aligned}
&\theta\mathbb{E}\langle w^* - w^k, \nabla_w\phi(w^k,\delta^k) - q^k\rangle + \frac{1}{2}\mathbb{E}\|w^* - w^k\|^2_{\sigma^{-1}} + \frac{n\{1+(n-1)\mu\tau/n\}}{2}\mathbb{E}\|\delta^* - \delta^k\|^2_{\tau^{-1}} \\
&\quad + \frac{\kappa\theta}{2}\mathbb{E}\|w^k - w^{k-1}\|^2_{\sigma^{-1}} + \frac{n\kappa\theta}{2}\mathbb{E}\|\delta^k - \delta^{k-1}\|^2_{\tau^{-1}} \\
&\geq \mathbb{E}\langle w^* - w^{k+1}, \nabla_w\phi(w^{k+1},\delta^{k+1}) - q^{k+1}\rangle + \frac{(1+\mu\sigma)}{2}\mathbb{E}\|w^* - w^{k+1}\|^2_{\sigma^{-1}} \\
&\quad + \frac{n(1+\mu\tau)}{2}\mathbb{E}\|\delta^* - \delta^{k+1}\|^2_{\tau^{-1}} + \frac{\kappa}{2}\mathbb{E}\|w^{k+1} - w^k\|^2_{\sigma^{-1}} + \frac{n\kappa}{2}\mathbb{E}\|\delta^{k+1} - \delta^k\|^2_{\tau^{-1}}. \qquad (46)
\end{aligned}$$

Combining (46) and the definition of θ, we have

$$\begin{aligned}
&\theta\Bigg[\mathbb{E}\langle w^* - w^k, \nabla_w\phi(w^k,\delta^k) - q^k\rangle + \frac{(1+\mu\sigma)}{2}\mathbb{E}\|w^* - w^k\|^2_{\sigma^{-1}} \\
&\quad + \frac{n(1+\mu\tau)}{2}\mathbb{E}\|\delta^* - \delta^k\|^2_{\tau^{-1}} + \frac{\kappa}{2}\mathbb{E}\|w^k - w^{k-1}\|^2_{\sigma^{-1}} + \frac{n\kappa}{2}\mathbb{E}\|\delta^k - \delta^{k-1}\|^2_{\tau^{-1}}\Bigg] \\
&\geq \mathbb{E}\langle w^* - w^{k+1}, \nabla_w\phi(w^{k+1},\delta^{k+1}) - q^{k+1}\rangle + \frac{(1+\mu\sigma)}{2}\mathbb{E}\|w^* - w^{k+1}\|^2_{\sigma^{-1}} \\
&\quad + \frac{n(1+\mu\tau)}{2}\mathbb{E}\|\delta^* - \delta^{k+1}\|^2_{\tau^{-1}} + \frac{\kappa}{2}\mathbb{E}\|w^{k+1} - w^k\|^2_{\sigma^{-1}} + \frac{n\kappa}{2}\mathbb{E}\|\delta^{k+1} - \delta^k\|^2_{\tau^{-1}}
\end{aligned}$$

which, together with the fact that $(w^{-1},\delta^{-1}) = (w^0,\delta^0)$, establishes

$$\begin{aligned}
&\theta^K\left[\frac{(1+\mu\sigma)}{2}\|w^* - w^0\|^2_{\sigma^{-1}} + \frac{n(1+\mu\tau)}{2}\|\delta^* - \delta^0\|^2_{\tau^{-1}}\right] \\
&\geq \mathbb{E}\langle w^* - w^K, \nabla_w\phi(w^K,\delta^K) - q^K\rangle + \frac{(1+\mu\sigma)}{2}\mathbb{E}\|w^* - w^K\|^2_{\sigma^{-1}} \\
&\quad + \frac{n(1+\mu\tau)}{2}\mathbb{E}\|\delta^* - \delta^K\|^2_{\tau^{-1}} + \frac{\kappa}{2}\mathbb{E}\|w^K - w^{K-1}\|^2_{\sigma^{-1}} + \frac{n\kappa}{2}\mathbb{E}\|\delta^K - \delta^{K-1}\|^2_{\tau^{-1}} \\
&\geq \frac{(1+\mu\sigma-2\kappa)}{2}\mathbb{E}\|w^* - w^K\|^2_{\sigma^{-1}} + \frac{n(1+\mu\tau)}{2}\mathbb{E}\|\delta^* - \delta^K\|^2_{\tau^{-1}} \qquad (47)
\end{aligned}$$

where we have used (39) at the last inequality. By the definition of κ and the step-size conditions, all the coefficients for the quadratic terms at the RHS of (47) are positive. This concludes the proof. □

B.5 Corollaries 3 and 4

Corollary 3. *Suppose Assumptions 1 and 2 are true. Let $\{(w^k,\delta^k)\}$ be the sequence generated by DSI-HG. If $\theta = 1$ and $\rho > 0$, $\sigma > 0$, $\tau > 0$ satisfy*

$$\max\{L_{12}(\sigma\tau)^{1/2}, L_{11}\sigma\} < \min\{1/3 - \rho(\sigma^{-1} + 4L_{11}^2\sigma), 1 - \rho(\tau^{-1} + 12L_{12}^2\tau)\},$$

we have

$$\frac{1}{K}\sum_{k=1}^{K}\|F(w^k,\delta^k)\|^2 = O(1/K).$$

Corollary 4. *Suppose Assumptions 1 and 3 are true. Let $\{(w^k,\delta^k)\}$ be the sequence generated by DSI-HG. If $\theta > 0$, $\sigma > 0$, $\tau > 0$ satisfy*

$$\max\{L_{12}(\sigma\tau)^{1/2}, L_{11}\sigma\} \leq 1/3, \qquad \theta = \max\left\{\frac{1}{1+\mu\sigma}, \frac{1}{1+\mu\tau}\right\},$$

we have

$$\|w^* - w^K\|^2 = O(\theta^K), \qquad \|\delta^* - \delta^K\|^2 = O(\theta^K).$$

C OMITTED EXPERIMENTS SETTINGS

All images are normalized into the range $[0, 1]$.

YOPO. For YOPO-M-N, we use $M = 10$ and $N = 5$. YOPO-5-10, which is compared with PGD-40 AT on MNIST in the paper for YOPO (Zhang et al., 2019a), performed worse than YOPO-10-5. Other than the learning rate schedule and (M, N), we use the code and the exact hyperparameter choices released by the authors of YOPO.

DAT. For DAT training, maximum first-order stationary condition (FOSC) value is set to 0.5, and FOSC control epoch is set to 0.8 times the number of training epochs. Both are the parameter settings used by the authors of DAT (Wang et al., 2019).

Other settings. On SVHN and CIFAR-10, we use random horizontal flipping and random cropping augmentations. For MSI-HG, to promote exploration of the constraint set, we apply augmentation to δ as well. All other settings are described in Section 5.

Table 5: Hyperparameter choices for MSI-HG in Section 5.

Dataset	ϵ	τ	T
MNIST	0.4	0.2	5
SVHN	4/255	6/255	10
CIFAR-10	8/255	14/255	10