Existence and Estimation of Critical Batch Size for Training Generative Adversarial Networks with Two Time-Scale Update Rule

Naoki Sato [*1] **Hideaki Iiduka** [*1]

Abstract

Previous results have shown that a two time-scale update rule (TTUR) using different learning rates, such as different constant rates or different decaying rates, is useful for training generative adversarial networks (GANs) in theory and in practice. Moreover, not only the learning rate but also the batch size is important for training GANs with TTURs and they both affect the number of steps needed for training. This paper studies the relationship between batch size and the number of steps needed for training GANs with TTURs based on constant learning rates. We theoretically show that, for a TTUR with constant learning rates, the number of steps needed to find stationary points of the loss functions of both the discriminator and generator decreases as the batch size increases and that there exists a critical batch size minimizing the stochastic first-order oracle (SFO) complexity. Then, we use the Fréchet inception distance (FID) as the performance measure for training and provide numerical results indicating that the number of steps needed to achieve a low FID score decreases as the batch size increases and that the SFO complexity increases once the batch size exceeds the measured critical batch size. Moreover, we show that measured critical batch sizes are close to the sizes estimated from our theoretical results.

1. Introduction

1.1. Background

Generative adversarial networks (GANs) have attracted attention (see, e.g., (Thekumparampil et al., 2019; Jordon et al., 2019; Xu et al., 2020; Zhu et al., 2021)) with the development of real-world applications (Arjovsky et al., 2017; Gulrajani et al., 2017; Brock et al., 2019; Zhang & Khoreva, 2019). The generator network in a GAN constructs synthetic data from random variables, and the discriminator network separates the synthetic data from the real-world data.

Many optimizers have been presented for training GANs (see, e.g., (Goodfellow et al., 2014; Nagarajan & Kolter, 2017; Heusel et al., 2017; Chavdarova et al., 2019; Xu et al., 2020; Sauer et al., 2021)). In this paper, we focus on a two time scale update rule (TTUR) (Heusel et al., 2017) for finding a pair of stationary points of the loss functions of the discriminator and generator. TTUR uses sequences generated by each of the generator and the discriminator.

For example, let us consider a TTUR based on stochastic gradient descent (SGD) (Fehrman et al., 2020; Scaman & Malherbe, 2020; Chen et al., 2020); let $\boldsymbol{\theta}_n$ be the point generated by the generator at iteration n and $\boldsymbol{w}_n$ be the point generated by the discriminator at iteration n. The loss function of the generator $L_G(\cdot, \boldsymbol{w}_n)$ is minimized by using SGD with a learning rate α_n^G and a mini-batch stochastic gradient at $\boldsymbol{\theta}_n$ with a batch of size b. The loss function of the discriminator $L_D(\boldsymbol{\theta}_n, \cdot)$ is minimized with a learning rate α_n^D and a mini-batch stochastic gradient at $\boldsymbol{w}_n$ with the same batch size b. If α_n^G and α_n^D are decaying learning rates, then TTUR based on SGD converges almost surely to a pair of stationary points of L_G and L_D (Heusel et al., 2017, Theorem 1).

TTURs based on adaptive methods can be defined by replacing SGD with adaptive methods for training deep neural networks. For example, TTUR based on adaptive moment estimation (Adam) (Kingma & Ba, 2015) with decaying learning rates α_n^G and α_n^D converges almost surely to a pair of stationary points of the loss functions of the generator and the discriminator (Heusel et al., 2017, Theorem 2). It was shown numerically that TTUR based on AdaBelief (short for adapting step sizes by the belief in observed gradients) (Zhuang et al., 2020) with constant learning rates α^G and α^D has good scores in terms of the Fréchet inception distance (FID) (Heusel et al., 2017), which is a performance measure of optimizers for training GANs. This

[*]Equal contribution [1]Department of Computer Science, Meiji University, Japan. Correspondence to: Naoki Sato <ce235017@meiji.ac.jp>, Hideaki Iiduka <iiduka@cs.meiji.ac.jp>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

implies that TTUR based on AdaBelief is a powerful way to train GANs.

1.2. Motivation

The above subsection described that TTUR with decaying or constant learning rates can be used in both theory and practice to train GANs. In particular, the numerical results in (Heusel et al., 2017; Zhuang et al., 2020) show that TTUR performs well with constant learning rates. Hence, we will set constant learning rates α^G and α^D for the generator and the discriminator.

Meanwhile, the batch size also affects the performance of TTUR. Previous numerical evaluations (Heusel et al., 2017, Figure 5) have indicated that increasing the batch size used in TTUR tends to decrease the FID. This implies that large batch sizes are desirable for training GANs. The motivation behind this work is thus to identify the theoretical relationship between the performance of TTUR with constant learning rates and the batch size. In so doing, we may bridge the gap between theory and practice in regard to the performance of TTUR based on the batch size.

We will use the number of steps N needed to train the GAN as the performance metric and examine the relationship between the batch size b and number of steps N. Motivated by the previously reported results (Heusel et al., 2017), we tried to find a pair of stationary points of the loss functions of the discriminator and generator that satisfy the variational inequalities (1) of the gradients of both loss functions.

Previous results (Shallue et al., 2019; Zhang et al., 2019; Iiduka, 2022b; Goyal et al., 2017; Hoffer et al., 2017; You et al., 2017) on training deep neural networks in practical tasks have shown that, for each deep learning optimizer, the number of training steps is halved by each doubling of the batch size and that diminishing returns exist beyond a critical batch size. On the other hand, we are interested in verifying whether a critical batch size for GANs exists in theory and in practice.

1.3. Contribution

1.3.1. Theoretical result on TTUR with small constant learning rates and large batch sizes

The theoretical contribution of this paper is to show that, for TTURs with constant learning rates, the number of steps N needed to find a pair of stationary points of the loss functions of the discriminator and the generator decreases as the batch size b increases (see also (3) for the explicit forms of N). To show this, we need to clarify that TTUR with constant learning rates can approximate a pair $(\boldsymbol{\theta}^\star, \boldsymbol{w}^\star)$ of stationary points of the loss function $L_D(\boldsymbol{\theta}^\star, \cdot)$ of the discriminator and the loss function $L_G(\cdot, \boldsymbol{w}^\star)$ of the generator.

We define the inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^\top \boldsymbol{y}$ and the norm of $\boldsymbol{x} \in \mathbb{R}^d$ by $\|\boldsymbol{x}\| := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$. Let $\nabla_{\boldsymbol{\theta}} L_G(\cdot, \boldsymbol{w})$ be the gradient of $L_G(\cdot, \boldsymbol{w})$ ($\boldsymbol{w} \in \mathbb{R}^W$) and $\nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}, \cdot)$ be the gradient of $L_D(\boldsymbol{\theta}, \cdot)$ ($\boldsymbol{\theta} \in \mathbb{R}^\Theta$). A pair $(\boldsymbol{\theta}^\star, \boldsymbol{w}^\star)$ of stationary points of $L_D(\boldsymbol{\theta}^\star, \cdot)$ and $L_G(\cdot, \boldsymbol{w}^\star)$ is such that

$$\|\nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}^\star, \boldsymbol{w}^\star)\| = 0 \text{ and } \|\nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}^\star, \boldsymbol{w}^\star)\| = 0,$$

which are equivalent to the following variational inequalities (see Appendix A.11): for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$ and all $\boldsymbol{w} \in \mathbb{R}^W$,

$$\begin{aligned} &\langle \boldsymbol{\theta}^\star - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}^\star, \boldsymbol{w}^\star) \rangle \leq 0 \text{ and} \\ &\langle \boldsymbol{w}^\star - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}^\star, \boldsymbol{w}^\star) \rangle \leq 0. \end{aligned} \tag{1}$$

Let us examine TTURs based on adaptive methods (see Algorithm 1 and Table 4 in Appendix A.1) such as Adam (Kingma & Ba, 2015), AdaBelief (Zhuang et al., 2020), and RMSProp (Tieleman & Hinton, 2012). Let $((\boldsymbol{\theta}_n, \boldsymbol{w}_n))_{n \in \mathbb{N}} \subset \mathbb{R}^\Theta \times \mathbb{R}^W$ be the sequence generated by TTUR with constant learning rates α^G and α^D. We will show that, under certain assumptions, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$ and all $\boldsymbol{w} \in \mathbb{R}^W$,

$$\begin{aligned} &\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n) \rangle\right] \\ &\quad \leq \underbrace{\frac{\Theta \mathrm{Dist}(\boldsymbol{\theta}) H^G}{2\alpha^G \tilde{\beta}_1^G}}_{A_G} \frac{1}{N} + \underbrace{\frac{\sigma_G^2 \alpha^G}{2\tilde{\beta}_1^G \tilde{\gamma}^{G^2} h_{0,*}^G}}_{B_G} \frac{1}{b} \\ &\qquad + \underbrace{\frac{M_G^2 \alpha^G}{2\tilde{\beta}_1^G \tilde{\gamma}^{G^2} h_{0,*}^G} + \frac{\beta_1^G}{\tilde{\beta}_1^G} \sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta})(\sigma_G^2 + M_G^2)}}_{C_G}, \\ &\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\left[\langle \boldsymbol{w}_n - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n) \rangle\right] \\ &\quad \leq \underbrace{\frac{W \mathrm{Dist}(\boldsymbol{w}) H^D}{2\alpha^D \tilde{\beta}_1^D}}_{A_D} \frac{1}{N} + \underbrace{\frac{\sigma_D^2 \alpha^D}{2\tilde{\beta}_1^D \tilde{\gamma}^{D^2} h_{0,*}^D}}_{B_D} \frac{1}{b} \\ &\qquad + \underbrace{\frac{M_D^2 \alpha^D}{2\tilde{\beta}_1^D \tilde{\gamma}^{D^2} h_{0,*}^D} + \frac{\beta_1^D}{\tilde{\beta}_1^D} \sqrt{W \mathrm{Dist}(\boldsymbol{w})(\sigma_D^2 + M_D^2)}}_{C_D}, \end{aligned} \tag{2}$$

where $\sigma_G^2, \sigma_D^2 \geq 0$, $M_G, M_D, \mathrm{Dist}(\boldsymbol{\theta}), \mathrm{Dist}(\boldsymbol{w}) > 0$, $\beta_1^G, \beta_1^D, \gamma^G, \gamma^D \in [0, 1)$, $\tilde{\beta}_1^G := 1 - \beta_1^G$, $\tilde{\gamma}^G := 1 - \gamma^G$, $\tilde{\beta}_1^D := 1 - \beta_1^D$, $\tilde{\gamma}^D := 1 - \gamma^D$, $h_{0,*}^G, h_{0,*}^D > 0$, $H^G := \max_{i \in [\Theta]} H_i^G$, and $H^D := \max_{j \in [W]} H_j^D$ (see Theorem

3.1 and Section 3.1 for the definitions of the parameters). This result implies that using small constant learning rates α^G and α^D makes B_G, C_G, B_D, and C_D small and that using a large batch size b makes B_G/b and B_D/b small. Meanwhile, using small constant learning rates α^G and α^D makes A_G and A_D large. Hence, we need to use a large number of steps N to make A_G and A_D small when small constant learning rates are used. Therefore, small constant learning rates, a large batch size, and a large number of steps are useful for training GANs.

1.3.2. Relationship between batch size and the number of steps needed for an ϵ−approximation of TTUR

Let us suppose that the generator and the discriminator run in at most N_G and N_D steps for a certain batch size b,

$$N_G(b) := \frac{A_G b}{(\epsilon_G^2 - C_G)b - B_G} \text{ and } \quad N_D(b) := \frac{A_D b}{(\epsilon_D^2 - C_D)b - B_D}, \tag{3}$$

where $\epsilon_G, \epsilon_D > 0$. Then, we can show that TTUR is an ϵ–approximation in the sense that

$$\frac{1}{N_G}\sum_{n=1}^{N_G} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq \epsilon_G^2 \text{ and}$$
$$\frac{1}{N_D}\sum_{n=1}^{N_D} \mathbb{E}\left[\langle \boldsymbol{w}_n - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq \epsilon_D^2.$$

Moreover, N_G and N_D are monotone decreasing and convex functions of the batch size (Theorem 3.2). This result implies that large batch sizes are desirable in the sense of minimizing the number of steps for training GANs. A particularly interesting concern is how large the batch size b should be.

1.3.3. Existence of a critical batch size minimizing SFO complexity

Here, we consider stochastic first-order oracle (SFO) complexities (Iiduka, 2022b) defined by $N_G(b)b$ and $N_D(b)b$. We show that $N_G(b)b$ and $N_D(b)b$ are convex functions of b and that there exist global minimizers $b_G^\star$ and $b_D^\star$ of $N_G(b)b$ and $N_D(b)b$ (Theorem 3.3). Accordingly, $b_G^\star$ and $b_D^\star$ are given by

$$b_G^\star := \frac{2B_G}{\epsilon_G^2 - C_G} \text{ and } b_D^\star := \frac{2B_D}{\epsilon_D^2 - C_D}. \tag{4}$$

It would be desirable to use critical batch sizes $b_G^\star$ and $b_D^\star$ as it minimizes the SFO complexity, which is the computation cost of the stochastic gradient. Furthermore, we show that lower bounds for $b_G^\star$ and $b_D^\star$ can be estimated from some hyperparameters, the total number of datasets, and the number of dimensions of the model (Proposition 3.4).

1.3.4. Numerical results supporting our theoretical results: Estimation of critical batch size minimizing SFO complexity

We use FID as a performance measure for training a deep convolutional GAN (DCGAN) (Radford et al., 2016) on the LSUN-Bedroom dataset (Yu et al., 2015), a Wasserstein GAN with Gradient Penalty (WGAN-GP) (Gulrajani et al., 2017) on the CelebA dataset (Liu et al., 2015), and a BigGAN (Brock et al., 2019) on the ImageNet dataset (Deng et al., 2009). We numerically show that increasing the batch size decreases the number of steps needed to achieve a low FID score and that there exist critical batch sizes minimizing the SFO complexities. The numerical results match our theoretical results in Sections 1.3.2 and 1.3.3. We are also interested in estimating appropriate batch sizes before implementing TTURs. Hence, we estimate batch sizes using $b_G^\star$ and $b_D^\star$ in (4) and compare them with ones measured in numerical experiments. We find that the estimated sizes are close to the measured ones (see Section 4.4).

2. Mathematical Preliminaries

2.1. Assumptions

The notation used in this paper is summarized in Table 1.

We assume the following standard conditions:

Assumption 2.1.

(S1) $L_G^{(i)}(\cdot, \boldsymbol{w})\colon \mathbb{R}^\Theta \to \mathbb{R}$ *and* $L_D^{(i)}(\boldsymbol{\theta}, \cdot)\colon \mathbb{R}^W \to \mathbb{R}$ *are continuously differentiable.*

(S2) *Let* $((\boldsymbol{\theta}_n, \boldsymbol{w}_n))_{n\in\mathbb{N}} \subset \mathbb{R}^\Theta \times \mathbb{R}^W$ *be the sequence generated by an optimizer.*

(i) *For each iteration* n,

$$\mathbb{E}_{\xi_n^G}\left[\mathsf{G}_{\xi_n^G}(\boldsymbol{\theta}_n)\right] = \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n) \text{ and}$$
$$\mathbb{E}_{\xi_n^D}\left[\mathsf{G}_{\xi_n^D}(\boldsymbol{w}_n)\right] = \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n).$$

(ii) *There exist nonnegative constants* σ_G^2 *and* σ_D^2 *such that*

$$\mathbb{E}_{\xi_n^G}\left[\|\mathsf{G}_{\xi_n^G}(\boldsymbol{\theta}_n) - \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2\right] \leq \sigma_G^2 \text{ and}$$
$$\mathbb{E}_{\xi_n^D}\left[\|\mathsf{G}_{\xi_n^D}(\boldsymbol{w}_n) - \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2\right] \leq \sigma_D^2.$$

(S3) *For each iteration* n, *the optimizer samples mini-batches* $\mathcal{S}_n \subset \mathcal{S}$ *and* $\mathcal{R}_n \subset \mathcal{R}$ *and estimates the full gradi-*

Table 1. Notation List (The parameters and functions of the discriminator are defined by replacing G, $\mathcal{S}_n$, $\boldsymbol{\theta}$, and $\boldsymbol{w}$ in the generator with D, $\mathcal{R}_n$, $\boldsymbol{w}$, and $\boldsymbol{\theta}$)

Notation	Description
$\mathbb{N}$	The set of all nonnegative integers
$[N]$	$[N] := \{1, 2, \ldots, N\}$ $(N \in \mathbb{N}\backslash\{0\})$
$\lvert A \rvert$	The number of elements of a set A
$\mathbb{R}^d$	A d-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$, which induces the norm $\lVert \cdot \rVert$
$\mathbb{R}^d_+$	$\mathbb{R}^d_+ := \{\boldsymbol{x} \in \mathbb{R}^d \colon x_i \geq 0 \ (i \in [d])\}$
$\mathbb{R}^d_{++}$	$\mathbb{R}^d_{++} := \{\boldsymbol{x} \in \mathbb{R}^d \colon x_i > 0 \ (i \in [d])\}$
$\mathbb{S}^d_{++}$	The set of $d \times d$ symmetric positive-definite matrices
$\mathbb{D}^d$	The set of $d \times d$ diagonal matrices, i.e., $\mathbb{D}^d = \{M \in \mathbb{R}^{d \times d} \colon M = \mathsf{diag}(x_i),\ x_i \in \mathbb{R} \ (i \in [d])\}$
$\mathbb{E}_\xi[X]$	The expectation with respect to ξ of a random variable X
$\mathcal{S}$	A set of synthetic samples $\boldsymbol{z}^{(i)}$
$\mathcal{R}$	A set of real-world samples $\boldsymbol{x}^{(i)}$
$\mathcal{S}_n$	Mini-batch of b synthetic samples $\boldsymbol{z}^{(i)}$ at time n
$\mathcal{R}_n$	Mini-batch of b real world samples $\boldsymbol{x}^{(i)}$ at time n
$L_G^{(i)}(\cdot, \boldsymbol{w})$	A loss function of the generator for $\boldsymbol{w} \in \mathbb{R}^W$ and $\boldsymbol{z}^{(i)}$
$L_G(\cdot, \boldsymbol{w})$	The total loss function of the generator for $\boldsymbol{w} \in \mathbb{R}^W$, i.e., $L_G(\cdot, \boldsymbol{w}) := \lvert\mathcal{S}\rvert^{-1} \sum_{i \in \mathcal{S}} L_G^{(i)}(\cdot, \boldsymbol{w})$
ξ^G	A random variable supported on Ξ^G that does not depend on $\boldsymbol{w} \in \mathbb{R}^W$ and $\boldsymbol{\theta} \in \mathbb{R}^\Theta$
ξ_n^G	$\xi_0^G, \xi_1^G, \ldots$ are independent samples and ξ_n^G is independent of $(\boldsymbol{\theta}_k)_{k=0}^n \subset \mathbb{R}^\Theta$ and $\boldsymbol{w} \in \mathbb{R}^W$
$\xi_{n,i}^G$	A random variable generated from the i-th sampling at time n
$\xi_{[n]}^G$	The history of process $\xi_0^G, \xi_1^G, \ldots$ to time step n, i.e., $\xi_{[n]}^G := (\xi_0^G, \xi_1^G, \ldots, \xi_n^G)$
$\mathsf{G}_{\xi^G}(\boldsymbol{\theta})$	The stochastic gradient of $L_G(\cdot, \boldsymbol{w})$ at $\boldsymbol{\theta} \in \mathbb{R}^\Theta$
$\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n)$	The mini-batch stochastic gradient of $L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)$ for $\mathcal{S}_n$, i.e., $\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) := b^{-1} \sum_{i \in [b]} \mathsf{G}_{\xi_{n,i}^G}(\boldsymbol{\theta}_n)$

ents ∇L_G and ∇L_D as

$$\begin{aligned}
\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) &:= \frac{1}{b} \sum_{i \in [b]} \mathsf{G}_{\xi_{n,i}^G}(\boldsymbol{\theta}_n) \\
&= \frac{1}{b} \sum_{\{i \colon \boldsymbol{z}^{(i)} \in \mathcal{S}_n\}} \nabla_{\boldsymbol{\theta}} L_G^{(i)}(\boldsymbol{\theta}_n, \boldsymbol{w}_n) \textit{ and} \\
\nabla L_{D,\mathcal{R}_n}(\boldsymbol{w}_n) &:= \frac{1}{b} \sum_{i \in [b]} \mathsf{G}_{\xi_{n,i}^D}(\boldsymbol{w}_n) \\
&= \frac{1}{b} \sum_{\{i \colon \boldsymbol{x}^{(i)} \in \mathcal{R}_n\}} \nabla_{\boldsymbol{w}} L_D^{(i)}(\boldsymbol{\theta}_n, \boldsymbol{w}_n).
\end{aligned}$$

2.2. Adaptive method

We will consider the following TTUR-type optimizer, described by Algorithm 1, for solving Problem (1).

In order to analyze Algorithm 1, we will assume the following conditions:

Assumption 2.2.

(A1) $\mathsf{H}_n^G = \mathsf{diag}(h_{n,i}^G)$ depends on $\xi_{[n]}^G$ and $\mathsf{H}_n^D = \mathsf{diag}(h_{n,i}^D)$ depends on $\xi_{[n]}^D$. Moreover, $h_{n+1,i}^G \geq h_{n,i}^G$ and $h_{n+1,j}^D \geq h_{n,j}^D$ hold for all $n \in \mathbb{N}$, all $i \in [\Theta]$, and all $j \in [W]$.

(A2) For all $i \in [\Theta]$, there exists a positive number H_i^G such that $\sup_{n \in \mathbb{N}} \mathbb{E}[h_{n,i}^G] \leq H_i^G$. For all $j \in$

Algorithm 1 Adaptive Method for Solving Problem (1)

Require: $(\alpha_n^G)_{n \in \mathbb{N}}, (\alpha_n^D)_{n \in \mathbb{N}} \subset \mathbb{R}_{++}$, $\beta_1^G, \beta_1^D \in [0, 1)$, $\gamma^G, \gamma^D \in [0, 1)$

$n \leftarrow 0$, $(\boldsymbol{\theta}_0, \boldsymbol{w}_0) \in \mathbb{R}^\Theta \times \mathbb{R}^W$, $\boldsymbol{m}_{-1}^G = \boldsymbol{0} \in \mathbb{R}^\Theta$, $\boldsymbol{m}_{-1}^D = \boldsymbol{0} \in \mathbb{R}^W$
loop
  loop
    $\boldsymbol{m}_n^G := \beta_1^G \boldsymbol{m}_{n-1}^G + (1 - \beta_1^G) \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n)$
    $\hat{\boldsymbol{m}}_n^G := (1 - \gamma^{G^{n+1}})^{-1} \boldsymbol{m}_n^G$
    $\mathsf{H}_n^G \in \mathbb{S}_{++}^\Theta \cap \mathbb{D}^\Theta$
    Find $\boldsymbol{d}_n^G \in \mathbb{R}^\Theta$ that solves $\mathsf{H}_n^G \boldsymbol{d} = -\hat{\boldsymbol{m}}_n^G$
    $\boldsymbol{\theta}_{n+1} := \boldsymbol{\theta}_n + \alpha_n^G \boldsymbol{d}_n^G$
  end loop
  loop
    $\boldsymbol{m}_n^D := \beta_1^D \boldsymbol{m}_{n-1}^D + (1 - \beta_1^D) \nabla L_{D,\mathcal{R}_n}(\boldsymbol{w}_n)$
    $\hat{\boldsymbol{m}}_n^D := (1 - \gamma^{D^{n+1}})^{-1} \boldsymbol{m}_n^D$
    $\mathsf{H}_n^D \in \mathbb{S}_{++}^W \cap \mathbb{D}^W$
    Find $\boldsymbol{d}_n^D \in \mathbb{R}^W$ that solves $\mathsf{H}_n^D \boldsymbol{d} = -\hat{\boldsymbol{m}}_n^D$
    $\boldsymbol{w}_{n+1} := \boldsymbol{w}_n + \alpha_n^D \boldsymbol{d}_n^D$
  end loop
  $n \leftarrow n + 1$
end loop

$[W]$, there exists a positive number H_j^D such that $\sup_{n\in\mathbb{N}} \mathbb{E}[h_{n,j}^D] \leq H_j^D$.

Examples of $\mathsf{H}_n^G \in \mathbb{S}_{++}^{\Theta} \cap \mathbb{D}^{\Theta}$ and $\mathsf{H}_n^D \in \mathbb{S}_{++}^{W} \cap \mathbb{D}^{W}$ satisfying Assumption 2.2 are listed in Table 4 (see Appendix A.1). By referring to the results in (Chen et al., 2019; Iiduka, 2022a; Zhuang et al., 2020), we can check that H_n^G and H_n^D in Table 4 satisfy (A1) and (A2).

3. Main Results

3.1. Convergence analysis of Algorithm 1 using constant learning rates

We will assume the following conditions:

Assumption 3.1.

(C1) $\alpha_n^G := \alpha^G$ and $\alpha_n^D := \alpha^D$ for all $n \in \mathbb{N}$.

(C2) There exist positive numbers M_G and M_D such that $\mathbb{E}[\|\nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2] \leq M_G^2$ and $\mathbb{E}[\|\nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2] \leq M_D^2$.

(C3) For all $\boldsymbol{\theta} = (\theta_i) \in \mathbb{R}^{\Theta}$ and all $\boldsymbol{w} = (w_i) \in \mathbb{R}^W$, there exist positive numbers $\mathrm{Dist}(\boldsymbol{\theta})$ and $\mathrm{Dist}(\boldsymbol{w})$ such that $\max_{i\in[\Theta]} \sup\{(\theta_{n,i} - \theta_i)^2 \colon n \in \mathbb{N}\} \leq \mathrm{Dist}(\boldsymbol{\theta})$ and $\max_{i\in[W]} \sup\{(w_{n,i} - w_i)^2 \colon n \in \mathbb{N}\} \leq \mathrm{Dist}(\boldsymbol{w})$.

A previous study (Heusel et al., 2017) used a decaying learning rate $\mathcal{O}(n^{-\tau})$, where $\tau \in (0, 1]$, for TTUR based on Adam to train GANs. This paper investigates the performance of TTUR based on adaptive methods using *constant* learning rates defined by (C1). Condition (C2) provides upper bounds on the performance measures $(1/N)\sum_{n=1}^N \mathbb{E}[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle]$ and $(1/N)\sum_{n=1}^N \mathbb{E}[\langle \boldsymbol{w}_n - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle]$ (see (2) and Theorem 3.1 for details). (C2) has also been used to analyze adaptive methods for training deep neural networks (see, e.g., (Chen et al., 2019; Zhuang et al., 2020)). Condition (C3) has been used to provide upper bounds on the performance measures and for analyzing both convex and nonconvex optimization in deep neural networks (see, e.g., (Kingma & Ba, 2015; Reddi et al., 2018; Zhuang et al., 2020)). See Appendix A.12 for remarks regarding (C3).

The following is a convergence analysis of Algorithm 1 (The proof of Theorem 3.1 is in Appendices A.6 and A.7).

Theorem 3.1. *Suppose that Assumptions 2.1, 2.2, and 3.1 hold and consider the sequence $((\boldsymbol{\theta}_n, \boldsymbol{w}_n))_{n\in\mathbb{N}}$ generated by Algorithm 1. Then, the following hold:*

(i) *For all $\boldsymbol{\theta} \in \mathbb{R}^{\Theta}$ and all $\boldsymbol{w} \in \mathbb{R}^W$,*

$$\liminf_{n\to+\infty} \mathbb{E}[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle] \leq \frac{\alpha^G(\sigma_G^2 b^{-1} + M_G^2)}{2\tilde{\beta_1^G}\tilde{\gamma}^{G^2} h_{0,*}^G} + \sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta}) \left(\frac{\sigma_G^2}{b} + M_G^2\right)} \frac{\beta^G}{\tilde{\beta_1^G}},$$

$$\liminf_{n\to+\infty} \mathbb{E}[\langle \boldsymbol{w}_n - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle] \leq \frac{\alpha^D(\sigma_D^2 b^{-1} + M_D^2)}{2\tilde{\beta_1^D}\tilde{\gamma}^{D^2} h_{0,*}^D} + \sqrt{W \mathrm{Dist}(\boldsymbol{w}) \left(\frac{\sigma_D^2}{b} + M_D^2\right)} \frac{\beta_1^D}{\tilde{\beta_1^D}},$$

where $\tilde{\beta_1^G} := 1 - \beta_1^G$, $\tilde{\beta_1^D} := 1 - \beta_1^D$, $\tilde{\gamma}^G := 1 - \gamma^G$, $\tilde{\gamma}^D := 1 - \gamma^D$, $h_{0,*}^G := \min_{i\in[\Theta]} h_{0,i}^G$, and $h_{0,*}^D := \min_{j\in[W]} h_{0,j}^D$. Furthermore, there exist accumulation points $(\boldsymbol{\theta}^*, \boldsymbol{w}^*)$ and $(\boldsymbol{\theta}_*, \boldsymbol{w}_*)$ of $((\boldsymbol{\theta}_n, \boldsymbol{w}_n))_{n\in\mathbb{N}}$ such that

$$\mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}^*, \boldsymbol{w}^*)\|^2\right] \leq \frac{\alpha^G(\sigma_G^2 b^{-1} + M_G^2)}{2\tilde{\beta_1^G}\tilde{\gamma}^{G^2} h_{0,*}^G} + \sqrt{\Theta \mathrm{Dist}(\tilde{\boldsymbol{\theta}}) \left(\frac{\sigma_G^2}{b} + M_G^2\right)} \frac{\beta_1^G}{\tilde{\beta_1^G}},$$

$$\mathbb{E}\left[\|\nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_*, \boldsymbol{w}_*)\|^2\right] \leq \frac{\alpha^D(\sigma_D^2 b^{-1} + M_D^2)}{2\tilde{\beta_1^D}\tilde{\gamma}^{D^2} h_{0,*}^D} + \sqrt{W \mathrm{Dist}(\tilde{\boldsymbol{w}}) \left(\frac{\sigma_D^2}{b} + M_D^2\right)} \frac{\beta_1^D}{\tilde{\beta_1^D}},$$

where $\tilde{\boldsymbol{\theta}} := \boldsymbol{\theta}^* - \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}^*, \boldsymbol{w}^*)$ and $\tilde{\boldsymbol{w}} := \boldsymbol{w}_* - \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_*, \boldsymbol{w}_*)$.

(ii) *For all $\boldsymbol{\theta} \in \mathbb{R}^{\Theta}$, all $\boldsymbol{w} \in \mathbb{R}^W$, and all $N \geq 1$,*

$$\frac{1}{N}\sum_{n\in[N]} \mathbb{E}[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle] \leq \frac{\Theta \mathrm{Dist}(\boldsymbol{\theta}) H^G}{2\alpha^G \tilde{\beta_1^G} N} + \frac{\alpha^G}{2\tilde{\beta_1^G}\tilde{\gamma}^{G^2} h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) + \frac{\beta_1^G}{\tilde{\beta_1^G}}\sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta}) \left(\frac{\sigma_G^2}{b} + M_G^2\right)},$$

$$\frac{1}{N}\sum_{n\in[N]} \mathbb{E}[\langle \boldsymbol{w}_n - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle] \leq \frac{W \mathrm{Dist}(\boldsymbol{w}) H^D}{2\alpha^D \tilde{\beta_1^D} N} + \frac{\alpha^D}{2\tilde{\beta_1^D}\tilde{\gamma}^{D^2} h_{0,*}^D}\left(\frac{\sigma_D^2}{b} + M_D^2\right) + \frac{\beta_1^D}{\tilde{\beta_1^D}}\sqrt{W \mathrm{Dist}(\boldsymbol{w}) \left(\frac{\sigma_D^2}{b} + M_D^2\right)},$$

where $H^G := \max_{i\in[\Theta]} H_i^G$ and $H^D := \max_{j\in[W]} H_j^D$.

Theorem 3.1(i) and (ii) indicate that the larger the batch size b is, the smaller the upper bounds of the performance measures become. From Theorem 3.1(i), it would be desirable to use small learning rates α^G and α^D. Meanwhile, Theorem 3.1(ii) indicates that there is no evidence that using sufficiently small α^G and α^D is good for training GANs

since the upper bounds in Theorem 3.1 depend on α^G, α^D, $1/\alpha^G$, and $1/\alpha^D$. Indeed, the results reported in (Heusel et al., 2017) used small $\alpha^G, \alpha^D = 10^{-4}, 10^{-5}$.

3.2. Relationship between batch size and number of steps for Algorithm 1

The relationship between b and the number of steps N satisfying an ϵ–approximation of TTUR is as follows (The proof of Theorem 3.2 is in Appendix A.8):

Theorem 3.2. *Suppose that Assumptions 2.1, 2.2, and 3.1 hold and consider Algorithm 1. Then, N_G and N_D defined by*

$$N_G(b) := \frac{A_G b}{(\epsilon_G^2 - C_G)b - B_G} \leq N_G \text{ for } b > \frac{B_G}{\epsilon_G^2 - C_G},$$
$$N_D(b) := \frac{A_D b}{(\epsilon_D^2 - C_D)b - B_D} \leq N_D \text{ for } b > \frac{B_D}{\epsilon_D^2 - C_D} \tag{5}$$

satisfy

$$\frac{1}{N_G}\sum_{n=1}^{N_G} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq \epsilon_G^2,$$
$$\frac{1}{N_D}\sum_{n=1}^{N_D} \mathbb{E}\left[\langle \boldsymbol{w}_n - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq \epsilon_D^2, \tag{6}$$

where A_G, B_G, C_G, A_D, B_D, and C_D are defined as in (2). Moreover, the functions $N_G(b)$ and $N_D(b)$ defined by (5) are monotone decreasing and convex for $b > B_G/(\epsilon_G^2 - C_G)$ and $b > B_D/(\epsilon_D^2 - C_D)$.

Theorem 3.2 indicates that N_G and N_D defined by (5) decrease as the batch size increases. Accordingly, it is useful to choose a sufficiently large b in the sense of minimizing of the number of steps N needed for an ϵ–approximation of TTUR.

3.3. Existence of a critical batch size

A particular concern is how large b should be. Here, we consider SFO complexities defined for the number of steps needed for (6) and for the batch size by

$$N_G(b)b = \frac{A_G b^2}{(\epsilon_G^2 - C_G)b - B_G} \text{ and}$$
$$N_D(b)b = \frac{A_D b^2}{(\epsilon_D^2 - C_D)b - B_D}. \tag{7}$$

The following theorem guarantees the existence of critical batch sizes that are global minimizers of $N_G(b)b$ and $N_D(b)b$ defined by (7) (The proof of Theorem 3.3 is in Appendix A.9).

Theorem 3.3. *Suppose that Assumptions 2.1, 2.2, and 3.1 hold and consider Algorithm 1. Then, there exist*

$$b_G^\star := \frac{2B_G}{\epsilon_G^2 - C_G} \text{ and } b_D^\star := \frac{2B_D}{\epsilon_D^2 - C_D} \tag{8}$$

such that $b_G^\star$ minimizes the convex function $N_G(b)b$ ($b > B_G/(\epsilon_G^2 - C_G)$) and $b_D^\star$ minimizes the convex function $N_D(b)b$ ($b > B_D/(\epsilon_D^2 - C_D)$).

Theorem 3.3 leads to the following proposition that gives lower bounds for the critical batch sizes.

Proposition 3.4. *Suppose that the assumptions in Theorem 3.3 hold and consider Algorithm 1. Then, $b_G^\star$ and $b_D^\star$ defined by (8) satisfy the following that*
(i) *for Adam,*

$$b_G^\star \geq \frac{\sigma_G^2 \alpha^G}{\epsilon_G^3 (1-\beta_1^G)^3 \sqrt{\frac{\Theta}{1-\beta_2^G}\frac{1}{|S|^2}}} \text{ and}$$
$$b_D^\star \geq \frac{\sigma_D^2 \alpha^D}{\epsilon_D^3 (1-\beta_1^D)^3 \sqrt{\frac{W}{1-\beta_2^D}\frac{1}{|S|^2}}},$$

(ii) *for AdaBelief,*

$$b_G^\star \geq \frac{\sigma_G^2 \alpha^G}{\epsilon_G^3 (1-\beta_1^G)^3 \sqrt{\frac{4\Theta}{1-\beta_2^G}\frac{1}{|S|^2}}} \text{ and}$$
$$b_D^\star \geq \frac{\sigma_D^2 \alpha^D}{\epsilon_D^3 (1-\beta_1^D)^3 \sqrt{\frac{4W}{1-\beta_2^D}\frac{1}{|S|^2}}},$$

(iii) *for RMSProp,*

$$b_G^\star \geq \frac{\sigma_G^2 \alpha^G}{\epsilon_G^3 \sqrt{\frac{\Theta}{|S|^2}}} \text{ and } b_D^\star \geq \frac{\sigma_D^2 \alpha^D}{\epsilon_D^3 \sqrt{\frac{W}{|S|^2}}},$$

where $\sigma_G^2, \sigma_D^2 \geq 0$, $\alpha^G, \alpha^D, \epsilon_G, \epsilon_D > 0$, $\beta_1^G, \beta_1^D \in [0, 1)$, and $\beta_2^G, \beta_2^D \in [0, 1)$.

Theorem 3.3 indicates that critical batch sizes exist in the sense of minimizing the SFO complexities $N_G(b)b$ and $N_D(b)b$. We are interested in verifying whether or not a critical batch size exists for training GANs as is the case of training deep neural networks (Shallue et al., 2019; Zhang et al., 2019; Iiduka, 2022b). The next section numerically examines the relationship between the batch size b and the number of steps N and also that between b and the SFO complexity Nb to see if there is a critical batch size $b^\star$ at which $N(b)b$ is minimized. Proposition 3.4 indicates that a lower bound for the critical batch size can be estimated from some hyperparameters. Hence, we would like to check whether the estimated sizes are close to the measured ones (see Section 4.4).

4. Numerical Results

We measured the number of steps required to achieve a low FID (Heusel et al., 2017) score for different batch sizes in several GAN trainings. The experimental environment consisted of NVIDIA DGX A100×8GPU and Dual AMD Rome7742 2.25-GHz, 128 Cores×2CPU. The software environment was Python 3.8.2, Pytorch 1.6.0, and CUDA 11.6. The distribution of ξ_n^G was a uniform one. The clean-fid package (Parmar et al., 2022) was used to calculate the FID. The code is available at `https://github.com/iiduka-researches/GANs`. The TTURs based on Adam, AdaBelief, and RMSProp used H_n^G and H_n^D defined in Table 4. See Table 5 for the optimizer hyperparameters used in the experiment. The learning rate used in each experiment was determined on the basis of a grid search of 36 combinations of the generator learning rate α^G and discriminator learning rate α^D (see Figure 7 in Appendix A.3). Appendix A.4 indicates that, when a fixed batch size is used, the FID scores of the TTURs used in the experiments decrease sufficiently as the number of steps increases.

4.1. Training DCGAN on the LSUN-Bedroom dataset

First, we evaluated the performance of TTURs based on RMSProp, AdaBelief, and Adam in training DCGAN (Radford et al., 2016) on the LSUN-Bedroom dataset. Figure 1 plots the number of steps N needed to achieve an FID score lower than 70 versus the batch size b. The figure indicates that the number of steps for TTUR based on any optimizer is monotone decreasing and convex with respect to b. Figure 2 plots the SFO complexity Nb versus b. The figure indicates that Nb for TTUR based on any optimizer is convex with respect to b.

4.2. Training WGAN-GP on the CelebA dataset

Next, we evaluated the performance of TTURs based on RMSProp, AdaBelief, and Adam in training WGAN-GP on the CelebA dataset. The original WGAN-GP code updates the discriminator five times for each generator update, whereas the discriminator is updated only once when using TTUR. Figure 3 plots the number of steps N needed to achieve an FID score lower than 50 versus the batch size b. The figure indicates that the number of steps for TTUR based on any optimizer is a monotone decreasing and convex function of b. Figure 4 plots Nb versus b. The figure indicates that Nb for TTUR based on any optimizer is a convex function of b.

4.3. Training BigGAN on the ImageNet dataset

We evaluated the performance of TTURs based on AdaBelief and Adam in training BigGAN (Brock et al., 2019)

Table 2. Parameters used to train GANs

	Section 4.1	Section 4.2	Section 4.3
Θ	$3,576,704$	$3,576,704$	$70,433,795$
W	$2,765,568$	$2,765,568$	$87,982,369$
$\lvert S\rvert$	$3,033,042$	$162,770$	$1,281,167$

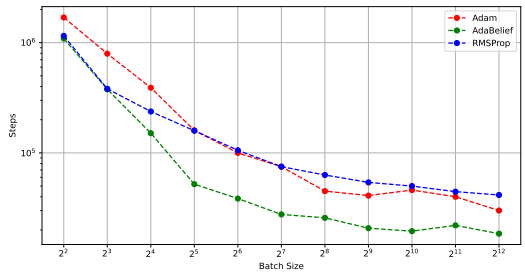

Figure 1. Number of steps for TTURs based on Adam, AdaBelief, and RMSProp versus batch size needed to train DCGAN on the LSUN-Bedroom dataset. The average of multiple runs is plotted.

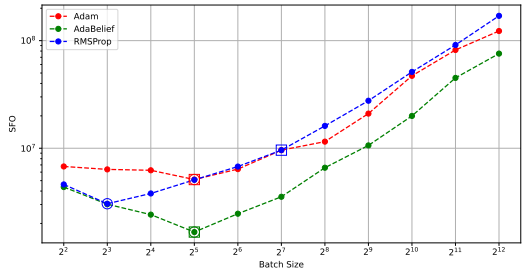

Figure 2. SFO complexities for TTURs based on Adam, AdaBelief, and RMSProp versus batch size needed to train DCGAN on the LSUN-Bedroom dataset. The double circle symbol denotes the measured critical batch size that minimizes SFO complexity. The square symbol denotes the estimated critical batch size.

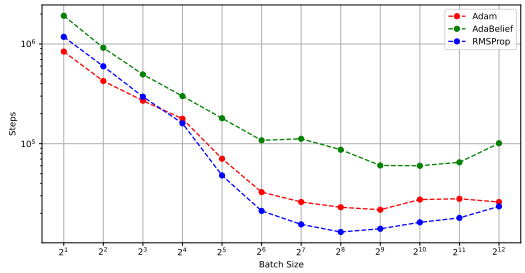

Figure 3. Number of steps for TTURs based on Adam, AdaBelief, and RMSProp versus batch size needed to train WGAN-GP on the CelebA dataset. The average of multiple runs is plotted.

Table 3. Measured and estimated critical batch sizes

	Section 4.1		Section 4.2		Section 4.3	
	measured	estimated	measured	estimated	measured	estimated
Adam	2^5	2^5	2^1	2^1	2^8	2^8
AdaBelief	2^5	2^5	2^2	2^2	2^9	2^9
RMSProp	2^3	2^7	2^6	2^4	-	-

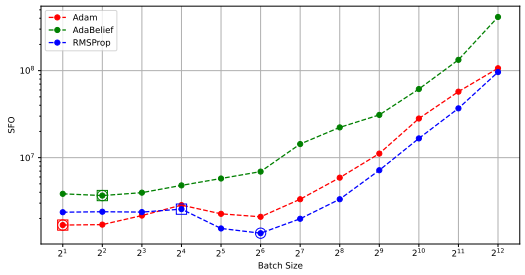

Figure 4. SFO complexities for TTURs based on Adam, AdaBelief, and RMSProp versus batch size needed to train WGAN-GP on the CelebA dataset. The double circle symbol denotes the measured critical batch size that minimizes SFO complexity. The square symbol denotes the estimated critical batch size.

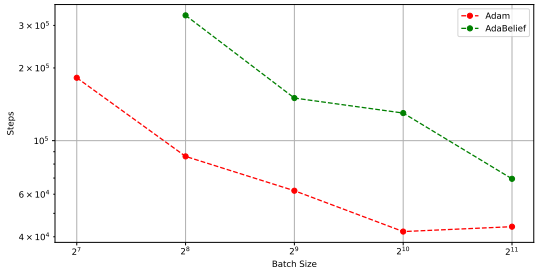

Figure 5. Number of steps for TTURs based on Adam and AdaBelief versus batch size needed to train BigGAN on the ImageNet dataset. The average of multiple runs is plotted.

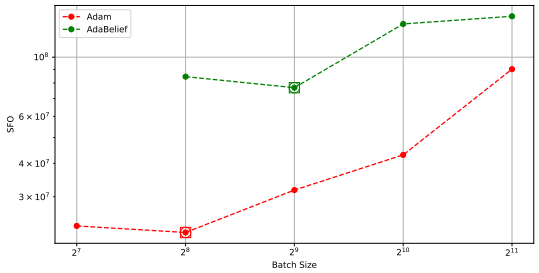

Figure 6. SFO complexities for TTURs based on Adam and AdaBelief versus batch size needed to train BigGAN on the ImageNet dataset. The double circle symbol denotes the measured critical batch size that minimizes SFO complexity. The square symbol denotes the estimated critical batch size.

on the ImageNet dataset. Figure 5 plots the number of steps N needed to achieve an FID score lower than 25 versus the batch size b. The figure indicates that the number of steps for TTUR based on AdaBelief and Adam is monotone decreasing and convex with respect to b. Figure 6 plots the SFO complexity Nb versus b. The figure indicates that Nb for TTUR based on AdaBelief and Adam is convex with respect to b.

We can conclude that the results in Sections 4.1, 4.2, and 4.3 support the theoretical results (Theorems 3.2 and 3.3).

4.4. Estimation of lower bound of critical batch size

Proposition 3.4 indicates that a lower bound on the critical batch size can be estimated with some parameters. The parameters used in the experiments are shown in Table 2; see Section A.2 for the settings of α^G, β_1^G, and so on. Figure 2 indicates that the measured critical batch sizes minimizing the SFO complexities of Adam, AdaBelief, and RMSProp are $2^5 = 32$, $2^5 = 32$, and $2^3 = 8$, respectively. For batch sizes below $2^7 = 128$, there is no significant difference in SFO complexity, but it is clear that the measured critical batch size is less than 2^7. In the previous studies (Shallue et al., 2019; Zhang et al., 2019), the critical batch sizes for training deep neural networks are large, such as 2^{12}, while the critical batch sizes for training GANs are small. According to Figure 2, in the DCGAN on the LSUN-Bedroom dataset setting, the measured critical batch size for Adam is 2^5; using this and Proposition 3.4(i) to back-calculate σ_G^2/ϵ_G^3 gives $\sigma_G^2/\epsilon_G^3 = 788.7$. We can use this ratio and Proposition 3.4(ii), (iii) to estimate a lower bound of 47.9 for AdaBelief and 126.5 for RMSProps (see Table 3).

Figure 4 indicates that the measured critical batch sizes for Adam, AdaBelief, and RMSProp are 2, $2^2 = 4$, and $2^6 = 64$, respectively. As in Figure 2, there is no significant difference in SFO complexity for batch sizes less than 2^6, but it is clear that the critical batch size is less than 2^6. As expected, it is smaller than the critical batch sizes of deep neural networks. In the same way as above, the estimated lower bounds on the WGAN-GP on the CelebA dataset are 1.7 for Adam, 4.25 for AdaBelief, and 20.3 for RMSProp (see Table 3).

According to Figure 6, in the BigGAN on the ImageNet dataset setting, the measured critical batch size for Adam is 2^8; using this and Proposition 3.4(i) to back calculate

σ_G^2/ϵ_G^3 gives $\sigma_G^2/\epsilon_G^3 = 530303.8$. We can use this ratio and Proposition 3.4(ii) to estimate a lower bound of 511.99 for AdaBelief (see Table 3).

Proposition 3.4(i) and (ii) indicate that the estimated critical batch sizes of Adam and AdaBelief strongly depend on the values of β_1 and β_2. Accordingly, the estimated critical batch sizes of Adam and AdaBelief were found to be the same as the measured ones. Meanwhile, Proposition 3.4(iii) indicates that the critical batch size of RMSProp is completely independent of the values of β_1 and β_2. In particular, β_1 in RMSProp is always 0 and β_2 in RMSProp is not used to estimate the critical batch size (see the proof of Proposition 3.4 in Appendix A.10 for details). The independence of β_1 and β_2 may have caused the failure in estimating the critical batch size of RMSProp.

5. Conclusion

We considered a stationary point problem in a GAN and performed a theoretical analysis of TTUR with constant learning rates to find a solution. We evaluated the upper bound of the expectation of the gradient of the loss function of the discriminator and the generator and showed that it is small when small constant learning rates and a large batch size are used. Next, we examined the relationship between the number of steps needed for solving the problem and batch size and showed that the number of steps decreases as the batch size increases. Moreover, we evaluated the SFO complexity of TTUR to check how large the batch size should be and showed that there is a critical batch size minimizing the SFO complexity, which is a convex function of the batch size. We also showed that it is possible to estimate the critical batch size specific to the model-dataset-optimizer combination. Finally, we provided numerical examples to support our theoretical analyzes. In particular, the numerical results showed that TTUR with small constant learning rates can be used to train DCGAN, WGAN-GP, and BigGAN, the number of steps needed to train them is monotone decreasing with the batch size, a critical batch size that minimizes the SFO complexity exists, and the estimated critical batch size is close to the experimentally measured value for DCGAN, WGAN-GP, and BigGAN.

Acknowledgements

We are sincerely grateful to Program Chairs, Area Chairs, and the three anonymous reviewers for helping us improve the original manuscript. We would like to thank Hiroki Naganuma (Mila, UdeM) for his help on the PyTorch implementation. This research is partly supported by the computational resources of the DGX A100 named TAIHO at Meiji University. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 21K11773 awarded to Hideaki Iiduka.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 214–223, 2017.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *Proceedings of The International Conference on Learning Representations*, 2019.

Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems*, volume 32, pp. 393–403, 2019.

Chen, H., Zheng, L., AL Kontar, R., and Raskutti, G. Stochastic gradient descent in correlated settings: A study on Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2722–2733, 2020.

Chen, X., Liu, S., Sun, R., and Hong, M. On the convergence of a class of Adam-type algorithms for non-convex optimization. In *Proceedings of The International Conference on Learning Representations*, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Fehrman, B., Gess, B., and Jentzen, A. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21:1–48, 2020.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: Training ImageNet in 1 hour. `https://arxiv.org/abs/1706.02677`, 2017.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5769–5779, 2017.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6629–6640, 2017.

Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, volume 11, pp. 1729–1739, 2017.

Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

Iiduka, H. Appropriate learning rates of adaptive learning rate optimization algorithms for training deep neural networks. *IEEE Transactions on Cybernetics*, 52(12):13250-13261, 2022a.

Iiduka, H. Critical bach size minimizes stochastic first-order oracle complexity of deep learning optimizer using hyperparameters close to one. `https://arxiv.org/abs/2208.09814`, 2022b.

Jordon, J., Yoon, J., and van der Schaar, M. Knockoff-GAN: Generating knockoffs for feature selection using generative adversarial networks. In *Proceedings of The International Conference on Learning Representations*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of The International Conference on Learning Representations*, 2015.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, 2015.

Nagarajan, V. and Kolter, J. Z. Gradient descent GAN optimization is locally stable. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5591–5600, 2017.

Parmar, G., Zhang, R., and Zhu, J.-Y. On aliased resizing and surprising subtleties in GAN evaluation. In *2022 IEEE/CVF Computer Vision and Pattern Recognition*, pp. 11400–11410, 2022.

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

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of Adam and beyond. In *Proceedings of The International Conference on Learning Representations*, 2018.

Sauer, A., Chitta, K., Müller, J., and Geiger, A. Projected GANs converge faster. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17480–17492, 2021.

Scaman, K. and Malherbe, C. Robustness analysis of non-convex stochastic gradient descent using biased expectations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16377–16387, 2020.

Shallue, C. J., Lee, J., Antognini, J., Sohl-Dickstein, J., Frostig, R., and Dahl, G. E. Measuring the effects of data parallelism on neural network training. *Journal of Machine Learning Research*, 20:1–49, 2019.

Thekumparampil, K. K., Oh, S., and Khetan, A. Robust conditional GANs under missing or uncertain labels. In *The ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning*, volume 1–7, 2019.

Tieleman, T. and Hinton, G. RMSProp: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Xu, T., Wenliang, L. K., Munn, M., and Acciaio, B. COT-GAN: Generating sequential data via causal optimal transport. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8798–8809, 2020.

You, Y., Gitman, I., and Ginsburg, B. Large batch training of convolutional networks. `https://arxiv.org/abs/1708.03888`, 2017.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. `https://arxiv.org/abs/1506.03365`, 2015.

Zhang, D. and Khoreva, A. Progressive augmentation of GANs. In *Advances in Neural Information Processing Systems*, volume 32, pp. 6246–6256, 2019.

Zhang, G., Li, L., Nado, Z., Martens, J., Sachdeva, S., Dahl, G. E., Shallue, C. J., and Grosse, R. Which algorithmic choices matter at which batch sizes? Insights from a noisy quadratic model. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8196–8207, 2019.

Zhu, J., Feng, R., Shen, Y., Zhao, D., Zha, Z.-J., Zhou, J., and Chen, Q. Low-rank subspaces in GANs. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16648–16658, 2021.

Zhuang, J., Tang, T., Ding, Y., Tatikonda, S., Dvornek, N., Papademetris, X., and Duncan, J. S. AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18795–18806, 2020.

A. Appendix

Unless stated otherwise, all relations between random variables are supported to hold almost surely. Let $S \in \mathbb{S}^d_{++}$. The S-inner product of $\mathbb{R}^d$ is defined for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_S := \langle \boldsymbol{x}, S\boldsymbol{y} \rangle$ and the S-norm is defined by $\|\boldsymbol{x}\|_S := \sqrt{\langle \boldsymbol{x}, S\boldsymbol{x} \rangle}$. The Hadamard product of $\mathbb{R}^d$ is defined for all $\boldsymbol{x} = (x_i)_{i=1}^d \in \mathbb{R}^d$ by $\boldsymbol{x} \odot \boldsymbol{x} := (x_i^2)_{i=1}^d \in \mathbb{R}^d$.

A.1. Examples of diagonal matrix in Algorithm 1

Table 4. Examples of $\mathsf{H}_n^G \in \mathbb{S}^\Theta_{++} \cap \mathbb{D}^\Theta$ and $\mathsf{H}_n^D \in \mathbb{S}^W_{++} \cap \mathbb{D}^W$ (steps 6 and 13) in Algorithm 1 ($\beta_2^G, \beta_2^D \in [0, 1)$)

	H_n^G	H_n^D
RMSProp (Tieleman & Hinton, 2012) $(\gamma^G = \gamma^D = \beta_1^G = \beta_1^D)$	$\boldsymbol{p}_n^G = \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) \odot \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n)$ $\boldsymbol{v}_n^G = \beta_2^G \boldsymbol{v}_{n-1}^G + (1-\beta_2^G)\boldsymbol{p}_n^G$ $\mathsf{H}_n^G = \mathsf{diag}(\sqrt{v_{n,i}^G})$	$\boldsymbol{p}_n^D = \nabla L_{D,\mathcal{R}_n}(\boldsymbol{w}_n) \odot \nabla L_{D,\mathcal{R}_n}(\boldsymbol{w}_n)$ $\boldsymbol{v}_n^D = \beta_2^D \boldsymbol{v}_{n-1}^D + (1-\beta_2^D)\boldsymbol{p}_n^D$ $\mathsf{H}_n^D = \mathsf{diag}(\sqrt{v_{n,i}^D})$
Adam (Heusel et al., 2017) (Kingma & Ba, 2015) $(v_{n,i}^G \leq v_{n+1,i}^G)$ $(v_{n,i}^D \leq v_{n+1,i}^D)$ $(\gamma^G = \gamma^D = \beta_1^G = \beta_1^D)$	$\boldsymbol{p}_n^G = \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) \odot \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n)$ $\boldsymbol{v}_n^G = \beta_2^G \boldsymbol{v}_{n-1}^G + (1-\beta_2^G)\boldsymbol{p}_n^G$ $\bar{\boldsymbol{v}}_n^G = \frac{\boldsymbol{v}_n^G}{1-\beta_2^{Gn}}$ $\mathsf{H}_n^G = \mathsf{diag}(\sqrt{\bar{v}_{n,i}^G})$	$\boldsymbol{p}_n^D = \nabla L_{D,\mathcal{R}_n}(\boldsymbol{w}_n) \odot \nabla L_{D,\mathcal{R}_n}(\boldsymbol{w}_n)$ $\boldsymbol{v}_n^D = \beta_2^D \boldsymbol{v}_{n-1}^D + (1-\beta_2^D)\boldsymbol{p}_n^D$ $\bar{\boldsymbol{v}}_n^D = \frac{\boldsymbol{v}_n^D}{1-\beta_2^{Dn}}$ $\mathsf{H}_n^D = \mathsf{diag}(\sqrt{\bar{v}_{n,i}^D})$
AMSGrad (Reddi et al., 2018) (Chen et al., 2019) $(\gamma^G = \gamma^D = 0)$	$\boldsymbol{p}_n^G = \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) \odot \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n)$ $\boldsymbol{v}_n^G = \beta_2^G \boldsymbol{v}_{n-1}^G + (1-\beta_2^G)\boldsymbol{p}_n^G$ $\hat{\boldsymbol{v}}_n^G = (\max\{\hat{v}_{n-1,i}^G, v_{n,i}^G\})_{i=1}^\Theta$ $\mathsf{H}_n^G = \mathsf{diag}(\sqrt{\hat{v}_{n,i}^G})$	$\boldsymbol{p}_n^D = \nabla L_{D,\mathcal{R}_n}(\boldsymbol{w}_n) \odot \nabla L_{D,\mathcal{R}_n}(\boldsymbol{w}_n)$ $\boldsymbol{v}_n^D = \beta_2^D \boldsymbol{v}_{n-1}^D + (1-\beta_2^D)\boldsymbol{p}_n^D$ $\hat{\boldsymbol{v}}_n^D = (\max\{\hat{v}_{n-1,i}^D, v_{n,i}^D\})_{i=1}^W$ $\mathsf{H}_n^D = \mathsf{diag}(\sqrt{\hat{v}_{n,i}^D})$
AdaBelief (Zhuang et al., 2020) $(s_{n,i}^G \leq s_{n+1,i}^G)$ $(s_{n,i}^D \leq s_{n+1,i}^D)$ $(\gamma^G = \gamma^D = \beta_1^G = \beta_1^D)$	$\tilde{\boldsymbol{p}}_n^G = \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) - \boldsymbol{m}_n^G$ $\tilde{\boldsymbol{s}}_n^G = \tilde{\boldsymbol{p}}_n^G \odot \tilde{\boldsymbol{p}}_n^G$ $\boldsymbol{s}_n^G = \beta_2^G \boldsymbol{v}_{n-1}^G + (1-\beta_2^G)\tilde{\boldsymbol{s}}_n^G$ $\hat{\boldsymbol{s}}_n^G = \frac{\boldsymbol{s}_n^G}{1-\beta_2^{Gn}}$ $\mathsf{H}_n^G = \mathsf{diag}(\sqrt{\hat{s}_{n,i}^G})$	$\tilde{\boldsymbol{p}}_n^D = \nabla L_{D,\mathcal{R}_n}(\boldsymbol{w}_n) - \boldsymbol{m}_n^D$ $\tilde{\boldsymbol{s}}_n^D = \tilde{\boldsymbol{p}}_n^D \odot \tilde{\boldsymbol{p}}_n^D$ $\boldsymbol{s}_n^D = \beta_2^D \boldsymbol{v}_{n-1}^D + (1-\beta_2^D)\tilde{\boldsymbol{s}}_n^D$ $\hat{\boldsymbol{s}}_n^D = \frac{\boldsymbol{s}_n^D}{1-\beta_2^{Dn}}$ $\mathsf{H}_n^D = \mathsf{diag}(\sqrt{\hat{s}_{n,i}^D})$

A.2. Hyperparameters of optimizers

Table 5. Hyperparameters of the optimizer used in the experiments in Sections 4.1, 4.2, and 4.3.

	optimizer	α^D	α^G	$\beta_1^G = \beta_1^D$	$\beta_2^G = \beta_2^D$	β_1 and β_2's reference
Section 4.1	Adam	0.0003	0.0001	0.5	0.999	(Radford et al., 2016)
	AdaBelief	0.00003	0.0003	0.5	0.999	(Zhuang et al., 2020)
	RMSProp	0.00003	0.0001	0	0.99	
Section 4.2	Adam	0.0003	0.0001	0.5	0.999	(Gulrajani et al., 2017)
	AdaBelief	0.00003	0.0005	0.5	0.999	(Zhuang et al., 2020)
	RMSProp	0.0001	0.0003	0	0.99	
Section 4.3	Adam	0.0004	0.0001	0	0.999	(Brock et al., 2019)
	AdaBelief	0.0005	0.00005	0.5	0.999	

A.3. Grid search

The combinations of learning rates used in the experiments in Sections 4.1 and 4.2 are determined using a grid search. Figure 7 shows the results of the grid search.

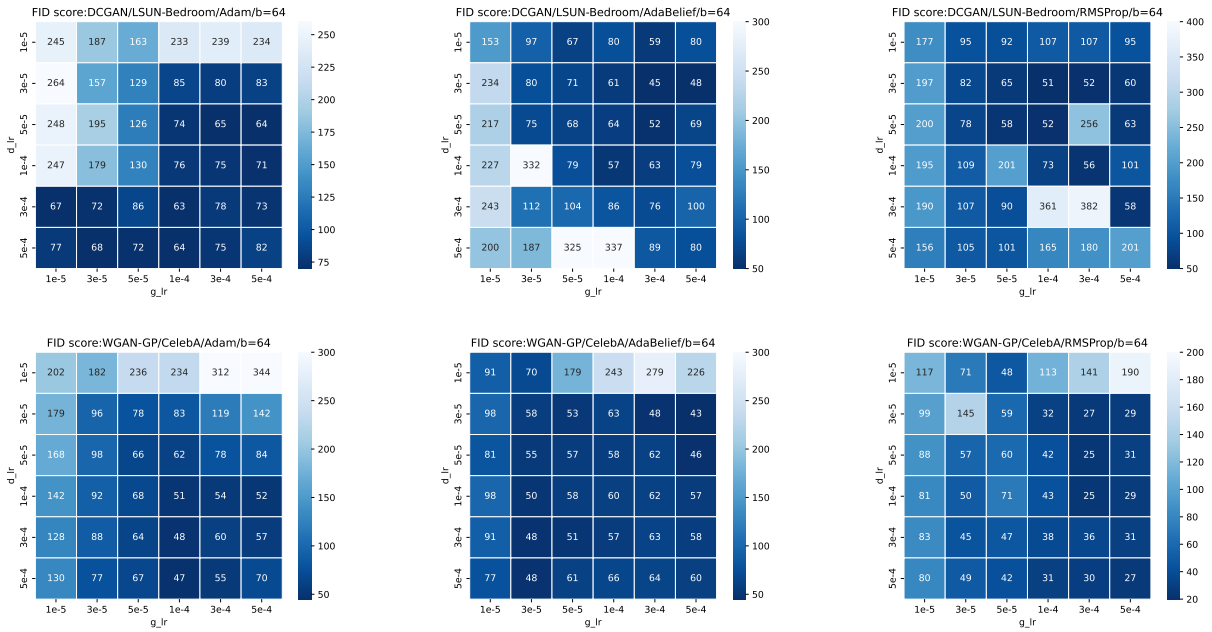

Figure 7. Analysis of the relationship between combination of learning rates and FID score: discriminator learning rate α^D on the vertical axis and generator learning rate α^G on the horizontal axis. The heatmap colors denote the FID scores; the darker the blue, the lower the FID, meaning that the training of the generator succeeded.

A.4. FID decreases sufficiently

We measured the number of steps required to achieve a good FID with different batch sizes. To demonstrate the soundness of the model used in the experiments, Figure 8 shows the decrease in FID. We find that, with DCGAN on the LSUN-Bedroom dataset, the FID score decreases to 41.8, and with WGAN-GP on the CelebA dataset, it decreases to 24.8.

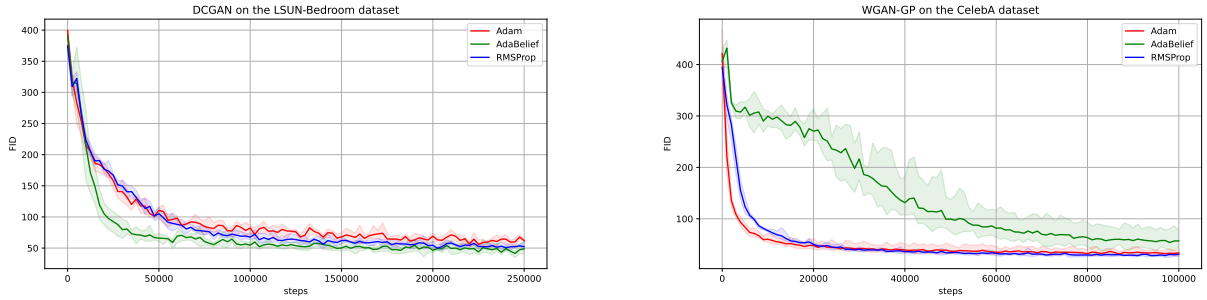

Figure 8. Mean FID (solid line) bounded by the maximum and the minimum over 5 runs (shaded area) for DCGAN on the LSUN-Bedroom dataset and WGAN-GP on the CelebA dataset for three optimizers. For all runs, the batch size is 64 and the learning rate combinations are determined with the same grid search (see Figure 7) used in Sections 4.1 and 4.2.

A.5. Lemmas

Lemma A.1. *Suppose that (S1), (S2)(i), and (S3) hold and consider Algorithm 1. Then, for all $\boldsymbol{\theta} \in \mathbb{R}^{\Theta}$ and all $n \in \mathbb{N}$,*

$$\mathbb{E}\left[\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|^2_{\mathsf{H}^G_n}\right] = \mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|^2_{\mathsf{H}^G_n}\right] + {\alpha^G_n}^2 \mathbb{E}\left[\|\boldsymbol{d}^G_n\|^2_{\mathsf{H}^G_n}\right]$$
$$+ 2\alpha^G_n \left\{ \frac{\beta^G_1}{\tilde{\gamma}^G_n} \mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \boldsymbol{m}^G_{n-1} \rangle\right] + \frac{\tilde{\beta}^G_1}{\tilde{\gamma}^G_n} \mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n) \rangle\right] \right\},$$

where $\tilde{\beta_1^G} := 1 - \beta_1^G$ *and* $\tilde{\gamma_n^G} := 1 - \gamma^{G^{n+1}}$.

Proof. Let $\boldsymbol{\theta} \in \mathbb{R}^\Theta$ and $n \in \mathbb{N}$. The definition of $\boldsymbol{\theta}_{n+1}$ implies that

$$\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|_{\mathsf{H}_n^G}^2 = \|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|_{\mathsf{H}_n^G}^2 + 2\alpha_n^G \langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \boldsymbol{d}_n^G \rangle_{\mathsf{H}_n^G} + \alpha_n^{G^2} \|\boldsymbol{d}_n^G\|_{\mathsf{H}_n^G}^2.$$

Moreover, the definitions of $\boldsymbol{d}_n^G$, $\boldsymbol{m}_n^G$, and $\hat{\boldsymbol{m}}_n^G$ ensure that

$$\begin{aligned}\left\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \boldsymbol{d}_n^G \right\rangle_{\mathsf{H}_n^G} &= \frac{1}{\tilde{\gamma_n^G}} \langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \boldsymbol{m}_n^G \rangle \\ &= \frac{\beta_1^G}{\tilde{\gamma_n^G}} \langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \boldsymbol{m}_{n-1}^G \rangle + \frac{\tilde{\beta_1^G}}{\tilde{\gamma_n^G}} \langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) \rangle.\end{aligned}$$

Hence,

$$\begin{aligned}\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|_{\mathsf{H}_n^G}^2 &= \|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|_{\mathsf{H}_n^G}^2 + \alpha_n^{G^2} \|\boldsymbol{d}_n^G\|_{\mathsf{H}_n^G}^2 \\ &\quad + 2\alpha_n^G \left\{ \frac{\beta_1^G}{\tilde{\gamma_n^G}} \langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \boldsymbol{m}_{n-1}^G \rangle + \frac{\tilde{\beta_1^G}}{\tilde{\gamma_n^G}} \langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) \rangle \right\}.\end{aligned} \tag{9}$$

Conditions (S2)(i) and (S3) guarantee that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) \rangle \middle| \boldsymbol{\theta}_n\right]\right] &= \mathbb{E}\left[\left\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \mathbb{E}\left[\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) \middle| \boldsymbol{\theta}_n\right] \right\rangle\right] \\ &= \mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n) \rangle\right].\end{aligned}$$

The lemma follows by taking the expectation with respect to ξ_n^G on both sides of (9). □

A discussion similar to the one proving Lemma A.1 leads to the following lemma.

Lemma A.2. *Suppose that (S1), (S2)(i), and (S3) hold and consider Algorithm 1. Then, for all* $\boldsymbol{w} \in \mathbb{R}^W$ *and all* $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{w}_{n+1} - \boldsymbol{w}\|_{\mathsf{H}_n^D}^2\right] &= \mathbb{E}\left[\|\boldsymbol{w}_n - \boldsymbol{w}\|_{\mathsf{H}_n^D}^2\right] + \alpha_n^{D^2} \mathbb{E}\left[\|\boldsymbol{d}_n^D\|_{\mathsf{H}_n^D}^2\right] \\ &\quad + 2\alpha_n^D \left\{ \frac{\beta_1^D}{\tilde{\gamma_n^D}} \mathbb{E}\left[\langle \boldsymbol{w} - \boldsymbol{w}_n, \boldsymbol{m}_{n-1}^D \rangle\right] + \frac{\tilde{\beta_1^D}}{\tilde{\gamma_n^D}} \mathbb{E}\left[\langle \boldsymbol{w} - \boldsymbol{w}_n, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n) \rangle\right] \right\},\end{aligned}$$

where $\tilde{\beta_1^D} := 1 - \beta_1^D$ *and* $\tilde{\gamma_n^D} := 1 - \gamma^{D^{n+1}}$.

Lemma A.3. *Algorithm 1 satisfies that, under (S2)(i), (ii) and (C2), for all* $n \in \mathbb{N}$,

$$\mathbb{E}\left[\|\boldsymbol{m}_n^G\|^2\right] \leq \frac{\sigma_G^2}{b} + M_G^2.$$

Under (A1) and (C2), for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\boldsymbol{d}_n^G\|_{\mathsf{H}_n^G}^2\right] \leq \frac{1}{(1-\gamma^G)^2 h_{0,*}^G} \left(\frac{\sigma_G^2}{b} + M_G^2 \right),$$

where $h_{0,*}^G := \min_{i \in [\Theta]} h_{0,i}^G$.

Proof. Let $n \in \mathbb{N}$. From (S2)(i), we have

$$\begin{aligned}\mathbb{E}\left[\|\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n)\|^2 \middle| \boldsymbol{\theta}_n\right] &= \mathbb{E}\left[\|\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) - \nabla L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n) + \nabla L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2 \middle| \boldsymbol{\theta}_n\right] \\ &= \mathbb{E}\left[\|\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) - \nabla L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2 \middle| \boldsymbol{\theta}_n\right] + \mathbb{E}\left[\|\nabla L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2 \middle| \boldsymbol{\theta}_n\right] \\ &\quad + 2\mathbb{E}\left[\langle \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) - \nabla L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n), \nabla L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n) \rangle \middle| \boldsymbol{\theta}_n\right] \\ &= \mathbb{E}\left[\|\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) - \nabla L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2 \middle| \boldsymbol{\theta}_n\right] + \|\nabla L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2,\end{aligned}$$

which, together with (S2)(ii) and (C2), implies that

$$\mathbb{E}\left[\|\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n)\|^2\right] \leq \frac{\sigma_G^2}{b} + M_G^2. \tag{10}$$

The convexity of $\|\cdot\|^2$, together with the definition of $\boldsymbol{m}_n^G$ and (10), guarantees that, for all $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{m}_n^G\|^2\right] &\leq \beta_1^G \mathbb{E}\left[\|\boldsymbol{m}_{n-1}^G\|^2\right] + (1-\beta_1^G)\mathbb{E}\left[\|\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n)\|^2\right] \\ &\leq \beta_1^G \mathbb{E}\left[\|\boldsymbol{m}_{n-1}^G\|^2\right] + (1-\beta_1^G)\left(\frac{\sigma_G^2}{b} + M_G^2\right).\end{aligned}$$

Induction thus ensures that, for all $n \in \mathbb{N}$,

$$\mathbb{E}\left[\|\boldsymbol{m}_n^G\|^2\right] \leq \max\left\{\|\boldsymbol{m}_{-1}^G\|^2, \frac{\sigma_G^2}{b} + M_G^2\right\} = \frac{\sigma_G^2}{b} + M_G^2, \tag{11}$$

where $\boldsymbol{m}_{-1}^G = \mathbf{0}$ is used. For $n \in \mathbb{N}$, $\mathsf{H}_n^G \in \mathbb{S}_{++}^{\Theta}$ guarantees the existence of a unique matrix $\overline{\mathsf{H}}_n^G \in \mathbb{S}_{++}^{\Theta}$ such that $\mathsf{H}_n^G = \overline{\mathsf{H}}_n^{G^2}$ (Horn & Johnson, 1985, Theorem 7.2.6). We have that, for all $\boldsymbol{x} \in \mathbb{R}^{\Theta}$, $\|\boldsymbol{x}\|_{\mathsf{H}_n^G}^2 = \|\overline{\mathsf{H}}_n^G \boldsymbol{x}\|^2$. Accordingly, the definitions of $\boldsymbol{d}_n^G$ and $\hat{\boldsymbol{m}}_n^G$ imply that, for all $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{d}_n^G\|_{\mathsf{H}_n^G}^2\right] &= \mathbb{E}\left[\left\|\overline{\mathsf{H}}_n^{G^{-1}} \mathsf{H}_n^G \boldsymbol{d}_n^G\right\|^2\right] \leq \frac{1}{\tilde{\gamma}_n^{G^2}} \mathbb{E}\left[\left\|\overline{\mathsf{H}}_n^{G^{-1}}\right\|^2 \|\boldsymbol{m}_n^G\|^2\right] \\ &\leq \frac{1}{(1-\gamma^G)^2} \mathbb{E}\left[\left\|\overline{\mathsf{H}}_n^{G^{-1}}\right\|^2 \|\boldsymbol{m}_n^G\|^2\right],\end{aligned}$$

where

$$\left\|\overline{\mathsf{H}}_n^{G^{-1}}\right\| = \left\|\mathsf{diag}\left(h_{n,i}^{G^{-\frac{1}{2}}}\right)\right\| = \max_{i \in [\Theta]} h_{n,i}^{G^{-\frac{1}{2}}}$$

and $\tilde{\gamma}_n^G := 1 - \gamma^{G^{n+1}} \geq 1 - \gamma^G$. Moreover, (A1) ensures that, for all $n \in \mathbb{N}$,

$$h_{n,i}^G \geq h_{0,i}^G \geq h_{0,*}^G := \min_{i \in [\Theta]} h_{0,i}^G.$$

Hence, (11) implies that, for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\boldsymbol{d}_n^G\|_{\mathsf{H}_n^G}^2\right] \leq \frac{1}{(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right),$$

completing the proof. □

A discussion similar to the one proving Lemma A.3 leads to the following lemma.

Lemma A.4. *Algorithm 1 satisfies that, under (S2)(i), (ii) and (C2), for all $n \in \mathbb{N}$,*

$$\mathbb{E}\left[\|\boldsymbol{m}_n^D\|^2\right] \leq \frac{\sigma_D^2}{b} + M_D^2.$$

Under (A1) and (C2), for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\boldsymbol{d}_n^D\|_{\mathsf{H}_n^D}^2\right] \leq \frac{1}{(1-\gamma^D)^2 h_{0,*}^D}\left(\frac{\sigma_D^2}{b} + M_D^2\right),$$

where $h_{0,*}^D := \min_{i \in [W]} h_{0,i}^D$.

Lemmas A.1 and A.3 lead to the following:

Lemma A.5. *Suppose that (S1)–(S3), (A1), and (C2)–(C3) hold and define* $X_n^G(\boldsymbol{\theta}) := \mathbb{E}[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|^2_{\mathsf{H}_n^G}]$ *for all* $n \in \mathbb{N}$ *and all* $\boldsymbol{\theta} \in \mathbb{R}^\Theta$*. Then, for all* $\boldsymbol{\theta} \in \mathbb{R}^\Theta$ *and all* $n \in \mathbb{N}$*,*

$$X_{n+1}^G(\boldsymbol{\theta}) \leq X_n^G(\boldsymbol{\theta}) + \mathrm{Dist}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i\in[\Theta]}(h_{n+1,i}^G - h_{n,i}^G)\right] + \frac{{\alpha_n^G}^2}{(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right)$$
$$+ 2\alpha_n^G\left\{\frac{\beta_1^G}{\tilde{\gamma}_n^G}\sqrt{\Theta\mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)} + \frac{\tilde{\beta_1^G}}{\tilde{\gamma}_n^G}\mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right]\right\}.$$

Proof. Lemma A.3 and Jensen's inequality guarantee that

$$\mathbb{E}\left[\|\boldsymbol{m}_n^G\|\right] \leq \sqrt{\frac{\sigma_G^2}{b} + M_G^2}.$$

Condition (C3) implies that, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$,

$$\|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|^2 = \sum_{i\in[\Theta]}(\theta_{n,i} - \theta_i)^2 \leq \Theta\mathrm{Dist}(\boldsymbol{\theta}).$$

The Cauchy-Schwarz inequality thus ensures that

$$\mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \boldsymbol{m}_{n-1}^G\rangle\right] \leq \mathbb{E}\left[\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|\|\boldsymbol{m}_{n-1}^G\|\right] \leq \sqrt{\Theta\mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}. \tag{12}$$

We define $X_n^G(\boldsymbol{\theta}) := \mathbb{E}[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|^2_{\mathsf{H}_n^G}]$ for all $n \in \mathbb{N}$ and all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$. Then, we have

$$X_{n+1}^G(\boldsymbol{\theta}) - \mathbb{E}\left[\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|^2_{\mathsf{H}_n^G}\right] = \mathbb{E}\left[\sum_{i\in[\Theta]}(h_{n+1,i}^G - h_{n,i}^G)(\theta_{n+1,i} - \theta_i)^2\right],$$

which, together with (C3), implies that

$$X_{n+1}^G(\boldsymbol{\theta}) - \mathbb{E}\left[\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|^2_{\mathsf{H}_n^G}\right] \leq \mathrm{Dist}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i\in[\Theta]}(h_{n+1,i}^G - h_{n,i}^G)\right].$$

Hence, Lemmas A.1 and A.3 lead to the assertion in Lemma A.5. □

A discussion similar to the one proving Lemma A.5, together with Lemmas A.2 and A.4, leads to the following lemma.

Lemma A.6. *Suppose that (S1)–(S3), (A1), and (C2)–(C3) hold and define* $X_n^D(\boldsymbol{w}) := \mathbb{E}[\|\boldsymbol{w}_n - \boldsymbol{w}\|^2_{\mathsf{H}_n^D}]$ *for all* $n \in \mathbb{N}$ *and all* $\boldsymbol{w} \in \mathbb{R}^W$*. Then, for all* $\boldsymbol{w} \in \mathbb{R}^W$ *and all* $n \in \mathbb{N}$*,*

$$X_{n+1}^D(\boldsymbol{w}) \leq X_n^D(\boldsymbol{w}) + \mathrm{Dist}(\boldsymbol{w})\mathbb{E}\left[\sum_{i\in[W]}(h_{n+1,i}^D - h_{n,i}^D)\right] + \frac{{\alpha_n^D}^2}{(1-\gamma^D)^2 h_{0,*}^D}\left(\frac{\sigma_D^2}{b} + M_D^2\right)$$
$$+ 2\alpha_n^D\left\{\frac{\beta_1^D}{\tilde{\gamma}_n^D}\sqrt{W\mathrm{Dist}(\boldsymbol{w})\left(\frac{\sigma_D^2}{b} + M_D^2\right)} + \frac{\tilde{\beta_1^D}}{\tilde{\gamma}_n^D}\mathbb{E}\left[\langle \boldsymbol{w} - \boldsymbol{w}_n, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right]\right\}.$$

A.6. Proof of Theorem 3.1(i)

The following is a convergence analysis of Algorithm 1.

Theorem A.1. *Suppose that Assumptions 2.1, 2.2, and 3.1 hold and consider the sequence $((\boldsymbol{\theta}_n, \boldsymbol{w}_n))_{n\in\mathbb{N}}$ generated by Algorithm 1. Then, the following hold: For all $\boldsymbol{\theta} \in \mathbb{R}^{\Theta}$ and all $\boldsymbol{w} \in \mathbb{R}^{W}$,*

$$\liminf_{n\to+\infty} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq \frac{\alpha^G(\sigma_G^2 b^{-1} + M_G^2)}{2\tilde{\beta_1^G}\tilde{\gamma}^{G^2} h_{0,*}^G} + \sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}\frac{\beta_1^G}{\tilde{\beta_1^G}},$$

$$\liminf_{n\to+\infty} \mathbb{E}\left[\langle \boldsymbol{w}_n - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq \frac{\alpha^D(\sigma_D^2 b^{-1} + M_D^2)}{2\tilde{\beta_1^D}\tilde{\gamma}^{D^2} h_{0,*}^D} + \sqrt{W \mathrm{Dist}(\boldsymbol{w})\left(\frac{\sigma_D^2}{b} + M_D^2\right)}\frac{\beta_1^D}{\tilde{\beta_1^D}},$$

where $\tilde{\beta_1^G} := 1 - \beta_1^G$, $\tilde{\beta_1^D} := 1 - \beta_1^D$, $\tilde{\gamma}^G := 1 - \gamma^G$, $\tilde{\gamma}^D := 1 - \gamma^D$, $h_{0,*}^G := \min_{i\in[\Theta]} h_{0,i}^G$, and $h_{0,*}^D := \min_{j\in[W]} h_{0,j}^D$. Furthermore, there exist accumulation points $(\boldsymbol{\theta}^*, \boldsymbol{w}^*)$ and $(\boldsymbol{\theta}_*, \boldsymbol{w}_*)$ of $((\boldsymbol{\theta}_n, \boldsymbol{w}_n))_{n\in\mathbb{N}}$ such that

$$\mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}^*, \boldsymbol{w}^*)\|^2\right] \leq \frac{\alpha^G(\sigma_G^2 b^{-1} + M_G^2)}{2\tilde{\beta_1^G}\tilde{\gamma}^{G^2} h_{0,*}^G} + \sqrt{\Theta \mathrm{Dist}(\tilde{\boldsymbol{\theta}})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}\frac{\beta_1^G}{\tilde{\beta_1^G}},$$

$$\mathbb{E}\left[\|\nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_*, \boldsymbol{w}_*)\|^2\right] \leq \frac{\alpha^D(\sigma_D^2 b^{-1} + M_D^2)}{2\tilde{\beta_1^D}\tilde{\gamma}^{D^2} h_{0,*}^D} + \sqrt{W \mathrm{Dist}(\tilde{\boldsymbol{w}})\left(\frac{\sigma_D^2}{b} + M_D^2\right)}\frac{\beta_1^D}{\tilde{\beta_1^D}},$$

where $\tilde{\boldsymbol{\theta}} := \boldsymbol{\theta}^* - \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}^*, \boldsymbol{w}^*)$ and $\tilde{\boldsymbol{w}} := \boldsymbol{w}_* - \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_*, \boldsymbol{w}_*)$.

Proof. Let us assume (C1), i.e., $\alpha_n^G := \alpha^G$ for all $n \in \mathbb{N}$. Then, Lemma A.5 ensures that, for all $\boldsymbol{\theta} \in \mathbb{R}^{\Theta}$ and all $n \in \mathbb{N}$,

$$\begin{aligned} X_{n+1}^G(\boldsymbol{\theta}) \leq X_n^G(\boldsymbol{\theta}) &+ \mathrm{Dist}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i\in[\Theta]}(h_{n+1,i}^G - h_{n,i}^G)\right] + \frac{\alpha^{G^2}}{(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) \\ &+ 2\alpha^G\left\{\frac{\beta_1^G}{\tilde{\gamma}_n^G}\sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)} + \frac{\tilde{\beta_1^G}}{\tilde{\gamma}_n^G}\mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right]\right\}. \end{aligned}$$

Since we have that $\tilde{\gamma}_n^G = 1 - \gamma^{G^{n+1}} \leq 1$, $\gamma^{G^{n+1}}(X_{n+1}^G(\boldsymbol{\theta}) - X_n^G(\boldsymbol{\theta})) \leq \gamma^{G^{n+1}} X_{n+1}^G(\boldsymbol{\theta})$, and $h_{n+1,i}^G \geq h_{n,i}^G$ (by (A1)) for all $n \in \mathbb{N}$, we also have that

$$\begin{aligned} X_{n+1}^G(\boldsymbol{\theta}) \leq X_n^G(\boldsymbol{\theta}) &+ \gamma^{G^{n+1}} X_{n+1}^G(\boldsymbol{\theta}) + \mathrm{Dist}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i\in[\Theta]}(h_{n+1,i}^G - h_{n,i}^G)\right] \\ &+ \frac{\alpha^{G^2}}{(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) \\ &+ 2\alpha^G\left\{\beta_1^G\sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)} + \tilde{\beta_1^G}\mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right]\right\}. \end{aligned} \tag{13}$$

Let us show that, for all $\boldsymbol{\theta} \in \mathbb{R}^{\Theta}$ and all $\epsilon > 0$,

$$\begin{aligned} &\liminf_{n\to+\infty} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \\ &\leq \frac{\alpha^G}{2\tilde{\beta_1^G}(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) + \sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}\frac{\beta_1^G}{\tilde{\beta_1^G}} + \epsilon. \end{aligned} \tag{14}$$

If (14) does not hold, then there exist $\hat{\boldsymbol{\theta}} \in \mathbb{R}^{\Theta}$ and $\epsilon_0 > 0$ such that

$$\begin{aligned} &\liminf_{n\to+\infty} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \\ &> \frac{\alpha^G}{2\tilde{\beta_1^G}(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) + \sqrt{\Theta \mathrm{Dist}(\hat{\boldsymbol{\theta}})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}\frac{\beta_1^G}{\tilde{\beta_1^G}} + \epsilon_0. \end{aligned}$$

Then, there exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$\begin{aligned}&\mathbb{E}\left[\langle \boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \\ &> \frac{\alpha^G}{2\tilde{\beta_1^G}(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) + \sqrt{\Theta \mathrm{Dist}(\hat{\boldsymbol{\theta}})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}\frac{\beta_1^G}{\tilde{\beta_1^G}} + \frac{\epsilon_0}{2}.\end{aligned}$$

Meanwhile, the conditions $\gamma^G \in [0,1)$, (C3), and (A1)–(A2) guarantee that there exists $n_1 \in \mathbb{N}$ such that, for all $n \geq n_1$,

$$\gamma^{G^{n+1}} X_{n+1}^G(\hat{\boldsymbol{\theta}}) + \mathrm{Dist}(\hat{\boldsymbol{\theta}})\mathbb{E}\left[\sum_{i\in[\Theta]}(h_{n+1,i}^G - h_{n,i}^G)\right] \leq \frac{\alpha^G \tilde{\beta_1^G}\epsilon_0}{2}.$$

Accordingly, from (13), for all $n \geq n_2 := \max\{n_0, n_1\}$,

$$\begin{aligned}X_{n+1}^G(\hat{\boldsymbol{\theta}}) &< X_n^G(\hat{\boldsymbol{\theta}}) + \frac{\alpha^G \tilde{\beta_1^G}\epsilon_0}{2} + \frac{\alpha^{G^2}}{(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) \\ &\quad + 2\alpha^G \beta_1^G \sqrt{\Theta \mathrm{Dist}(\hat{\boldsymbol{\theta}})\left(\frac{\sigma_G^2}{b} + M_G^2\right)} - 2\alpha^G \tilde{\beta_1^G}\left\{\frac{\alpha^G}{2\tilde{\beta_1^G}(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right)\right. \\ &\quad \left. + \sqrt{\Theta \mathrm{Dist}(\hat{\boldsymbol{\theta}})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}\frac{\beta_1^G}{\tilde{\beta_1^G}} + \frac{\epsilon_0}{2}\right\} \\ &= X_n^G(\hat{\boldsymbol{\theta}}) - \frac{\alpha^G \tilde{\beta_1^G}\epsilon_0}{2} \\ &< X_{n_2}^G(\hat{\boldsymbol{\theta}}) - \frac{\alpha^G \tilde{\beta_1^G}\epsilon_0}{2}(n+1-n_2).\end{aligned}$$

Note that the right-hand side of the above inequality approaches minus infinity as n approaches positive infinity, producing a contradiction. Therefore, (14) holds, which implies that, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$,

$$\begin{aligned}&\liminf_{n\to+\infty} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \\ &\quad \leq \frac{\alpha^G}{2\tilde{\beta_1^G}(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) + \sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}\frac{\beta_1^G}{\tilde{\beta_1^G}}.\end{aligned} \tag{15}$$

A discussion similar to the one showing (15), together with Lemma A.6, leads to the finding that, for all $\boldsymbol{w} \in \mathbb{R}^W$,

$$\begin{aligned}&\liminf_{n\to+\infty} \mathbb{E}\left[\langle \boldsymbol{w}_n - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \\ &\quad \leq \frac{\alpha^D}{2\tilde{\beta_1^D}(1-\gamma^D)^2 h_{0,*}^D}\left(\frac{\sigma_D^2}{b} + M_D^2\right) + \sqrt{W \mathrm{Dist}(\boldsymbol{w})\left(\frac{\sigma_D^2}{b} + M_D^2\right)}\frac{\beta_1^D}{\tilde{\beta_1^D}}.\end{aligned} \tag{16}$$

Let $\boldsymbol{\theta} \in \mathbb{R}^\Theta$. From (15), there exists a subsequence $((\boldsymbol{\theta}_{n_i}, \boldsymbol{w}_{n_i}))_{i\in\mathbb{N}}$ of $((\boldsymbol{\theta}_n, \boldsymbol{w}_n))_{n\in\mathbb{N}}$ such that

$$\begin{aligned}&\lim_{i\to+\infty} \mathbb{E}\left[\langle \boldsymbol{\theta}_{n_i} - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_{n_i}, \boldsymbol{w}_{n_i})\rangle\right] \\ &\quad \leq \frac{\alpha^G}{2\tilde{\beta_1^G}(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) + \sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}\frac{\beta_1^G}{\tilde{\beta_1^G}}.\end{aligned}$$

Conditions (S1) and (C3) guarantee that there exists $((\boldsymbol{\theta}_{n_{i_j}}, \boldsymbol{w}_{n_{i_j}}))_{j\in\mathbb{N}}$ of $((\boldsymbol{\theta}_{n_i}, \boldsymbol{w}_{n_i}))_{i\in\mathbb{N}}$ such that $((\boldsymbol{\theta}_{n_{i_j}}, \boldsymbol{w}_{n_{i_j}}))_{j\in\mathbb{N}}$ converges almost surely to $(\boldsymbol{\theta}^*, \boldsymbol{w}^*) \in \mathbb{R}^\Theta \times \mathbb{R}^W$. Therefore, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$,

$$\begin{aligned}&\mathbb{E}\left[\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}^*, \boldsymbol{w}^*)\rangle\right] \\ &\quad \leq \frac{\alpha^G}{2\tilde{\beta_1^G}(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) + \sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}\frac{\beta_1^G}{\tilde{\beta_1^G}}.\end{aligned}$$

Hence, letting $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} := \boldsymbol{\theta}^* - \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}^*, \boldsymbol{w}^*)$ implies that

$$\begin{aligned}&\mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}^*, \boldsymbol{w}^*)\|^2\right] \\ &\quad \leq \frac{\alpha^G}{2\tilde{\beta_1^G}(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) + \sqrt{\Theta \mathrm{Dist}(\tilde{\boldsymbol{\theta}})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}\frac{\beta_1^G}{\tilde{\beta_1^G}}.\end{aligned} \tag{17}$$

A discussion similar to the one showing (17), together with (16), implies that there exists $(\boldsymbol{\theta}_*, \boldsymbol{w}_*) \in \mathbb{R}^\Theta \times \mathbb{R}^W$ such that

$$\begin{aligned}&\mathbb{E}\left[\|\nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_*, \boldsymbol{w}_*)\|^2\right] \\ &\quad \leq \frac{\alpha^D}{2\tilde{\beta_1^D}(1-\gamma^D)^2 h_{0,*}^D}\left(\frac{\sigma_D^2}{b} + M_D^2\right) + \sqrt{W \mathrm{Dist}(\tilde{\boldsymbol{w}})\left(\frac{\sigma_D^2}{b} + M_D^2\right)}\frac{\beta_1^D}{\tilde{\beta_1^D}},\end{aligned}$$

where $\tilde{\boldsymbol{w}} = \boldsymbol{w}_* - \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_*, \boldsymbol{w}_*)$. This completes the proof. □

A.7. Proof of Theorem 3.1(ii)

Lemmas A.1 and A.3 lead to the following lemma:

Lemma A.7. *Suppose that (S1)–(S3), (A1)–(A2), and (C2)–(C3) hold and consider Algorithm 1, where $(\alpha_n^G)_{n\in\mathbb{N}}$ is monotone decreasing. Then, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$ and all $N \geq 1$,*

$$\begin{aligned}&\frac{1}{N}\sum_{n\in[N]} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \\ &\quad \leq \frac{\Theta \mathrm{Dist}(\boldsymbol{\theta}) H^G}{2\alpha_N^G \tilde{\beta_1^G} N} + \frac{1}{N}\sum_{n\in[N]} \frac{\alpha_n^G}{2\tilde{\beta_1^G}(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) + \frac{\beta_1^G}{\tilde{\beta_1^G}}\sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)},\end{aligned}$$

where $H^G := \max_{i\in[\Theta]} H_i^G$.

Proof. Lemma A.1 implies that, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$ and all $n \in \mathbb{N}$,

$$\begin{aligned}&\mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \\ &= \frac{\tilde{\gamma}_n^G}{2\alpha_n^G \tilde{\beta_1^G}}\left\{\mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|_{\mathsf{H}_n^G}^2\right] - \mathbb{E}\left[\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|_{\mathsf{H}_n^G}^2\right]\right\} + \frac{\alpha_n^G \tilde{\gamma}_n^G}{2\tilde{\beta_1^G}}\mathbb{E}\left[\|\boldsymbol{d}_n^G\|_{\mathsf{H}_n^G}^2\right] + \frac{\beta_1^G}{\tilde{\beta_1^G}}\mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \boldsymbol{m}_{n-1}^G\rangle\right],\end{aligned}$$

which implies that, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$ and all $N \geq 1$,

$$\begin{aligned}&\frac{1}{N}\sum_{n\in[N]} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \\ &= \underbrace{\frac{1}{N}\sum_{n\in[N]} \frac{\tilde{\gamma}_n^G}{2\alpha_n^G \tilde{\beta_1^G}}\left\{\mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|_{\mathsf{H}_n^G}^2\right] - \mathbb{E}\left[\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|_{\mathsf{H}_n^G}^2\right]\right\}}_{\Theta_N} + \underbrace{\frac{1}{N}\sum_{n\in[N]} \frac{\alpha_n^G \tilde{\gamma}_n^G}{2\tilde{\beta_1^G}}\mathbb{E}\left[\|\boldsymbol{d}_n^G\|_{\mathsf{H}_n^G}^2\right]}_{A_N} \\ &\quad + \underbrace{\frac{1}{N}\sum_{n\in[N]} \frac{\beta_1^G}{\tilde{\beta_1^G}}\mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \boldsymbol{m}_{n-1}^G\rangle\right]}_{B_N}.\end{aligned}$$

Let $\delta_n^G := \tilde{\gamma}_n^G/(2\alpha_n^G \tilde{\beta_1^G})$ for all $n \in \mathbb{N}$. Then, we have that

$$\begin{aligned}\Theta_N &:= \delta_1^G \mathbb{E}\left[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}\|_{\mathsf{H}_1^G}^2\right] + \underbrace{\sum_{n=2}^N \left\{\delta_n^G \mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|_{\mathsf{H}_n^G}^2\right] - \delta_{n-1}^G \mathbb{E}\left[\|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|_{\mathsf{H}_{n-1}^G}^2\right]\right\}}_{\tilde{\Theta}_N} \\ &\quad - \delta_N^G \mathbb{E}\left[\|\boldsymbol{\theta}_{N+1} - \boldsymbol{\theta}\|_{\mathsf{H}_N^G}^2\right].\end{aligned}$$

Since $\overline{\mathsf{H}}_n^G \in \mathbb{S}_{++}^{\Theta}$ exists such that $\mathsf{H}_n^G = \overline{\mathsf{H}}_n^{G^2}$, we have $\|\boldsymbol{x}\|_{\mathsf{H}_n^G}^2 = \|\overline{\mathsf{H}}_n^G \boldsymbol{x}\|^2$ for all $\boldsymbol{x} \in \mathbb{R}^{\Theta}$. Accordingly, we have

$$\tilde{\Theta}_N = \mathbb{E}\left[\sum_{n=2}^{N}\left\{\delta_n^G \left\|\overline{\mathsf{H}}_n^G(\boldsymbol{\theta}_n - \boldsymbol{\theta})\right\|^2 - \delta_{n-1}^G \left\|\overline{\mathsf{H}}_{n-1}^G(\boldsymbol{\theta}_n - \boldsymbol{\theta})\right\|^2\right\}\right].$$

Hence, for all $N \geq 2$,

$$\tilde{\Theta}_N = \mathbb{E}\left[\sum_{n=2}^{N}\sum_{i=1}^{\Theta}\left(\delta_n^G h_{n,i}^G - \delta_{n-1}^G h_{n-1,i}^G\right)(\theta_{n,i} - \theta_i)^2\right]. \tag{18}$$

Since $(\alpha_n^G)_{n\in\mathbb{N}}$ is monotone decreasing, we have that $\delta_{n+1}^G \geq \delta_n^G$ $(n \in \mathbb{N})$. Hence, from (A1), we have that, for all $n \geq 1$ and all $i \in [\Theta]$,

$$\delta_n^G h_{n,i}^G - \delta_{n-1}^G h_{n-1,i}^G \geq 0.$$

Moreover, from (C3), $\max_{i\in[\Theta]} \sup_{n\in\mathbb{N}}(\theta_{n,i} - \theta_i)^2 \leq \mathrm{Dist}(\boldsymbol{\theta})$. Accordingly, for all $N \geq 2$,

$$\begin{aligned}\tilde{\Theta}_N &\leq \mathrm{Dist}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{n=2}^{N}\sum_{i=1}^{\Theta}\left(\delta_n^G h_{n,i}^G - \delta_{n-1}^G h_{n-1,i}^G\right)\right] \\ &= \mathrm{Dist}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i=1}^{\Theta}\left(\delta_N^G h_{N,i}^G - \delta_1^G h_{1,i}^G\right)\right].\end{aligned}$$

Therefore, $\delta_1^G \mathbb{E}[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}\|_{\mathsf{H}_1^G}^2] \leq \mathrm{Dist}(\boldsymbol{\theta})\delta_1^G \mathbb{E}[\sum_{i=1}^{\Theta} h_{1,i}^G]$, and (A2) imply that, for all $N \geq 1$,

$$\begin{aligned}\Theta_N &\leq \delta_1^G \mathrm{Dist}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i=1}^{\Theta} h_{1,i}^G\right] + \mathrm{Dist}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i=1}^{\Theta}\left(\delta_N^G h_{N,i}^G - \delta_1^G h_{1,i}^G\right)\right] \\ &= \delta_N^G \mathrm{Dist}(\boldsymbol{\theta})\mathbb{E}\left[\sum_{i=1}^{\Theta} h_{N,i}^G\right] \\ &\leq \delta_N^G \mathrm{Dist}(\boldsymbol{\theta})\sum_{i=1}^{\Theta} H_i^G \\ &\leq \delta_N^G \Theta \mathrm{Dist}(\boldsymbol{\theta}) H^G,\end{aligned}$$

where $H^G = \max_{i\in[\Theta]} H_i^G$. From $\delta_n^G := \tilde{\gamma}_n^G/(2\alpha_n^G \tilde{\beta_1^G})$ and $\tilde{\gamma}_n^G = 1 - \gamma^{G^{n+1}} \leq 1$, we have

$$\Theta_N \leq \frac{\Theta \mathrm{Dist}(\boldsymbol{\theta}) H^G}{2\alpha_N^G \tilde{\beta_1^G}}. \tag{19}$$

Lemma A.3 implies that, for all $N \geq 1$,

$$A_N := \sum_{n\in[N]} \frac{\alpha_n^G \tilde{\gamma}_n^G}{2\tilde{\beta_1^G}} \mathbb{E}\left[\|\boldsymbol{d}_n^G\|_{\mathsf{H}_n^G}^2\right] \leq \sum_{n\in[N]} \frac{\alpha_n^G}{2\tilde{\beta_1^G}(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right). \tag{20}$$

From (12), we have

$$B_N := \sum_{n\in[N]} \frac{\beta_1^G}{\tilde{\beta_1^G}} \mathbb{E}\left[\langle \boldsymbol{\theta} - \boldsymbol{\theta}_n, \boldsymbol{m}_{n-1}^G\rangle\right] \leq \frac{\beta_1^G N}{\tilde{\beta_1^G}}\sqrt{\Theta \mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)}. \tag{21}$$

Therefore, (19), (20), and (21) lead to the assertion in Lemma A.7. This completes the proof. □

A discussion similar to the one proving Lemma A.7, together with Lemmas A.2 and A.4, leads to the following lemma:

Lemma A.8. *Suppose that (S1)–(S3), (A1)–(A2), and (C2)–(C3) hold and consider Algorithm 1, where $(\alpha_n^D)_{n\in\mathbb{N}}$ is monotone decreasing. Then, for all $\boldsymbol{w} \in \mathbb{R}^W$ and all $N \geq 1$,*

$$\frac{1}{N}\sum_{n\in[N]} \mathbb{E}\left[\langle \boldsymbol{w}_n - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq \frac{W\mathrm{Dist}(\boldsymbol{w})H^D}{2\alpha_N^D \tilde{\beta}_1^D N} + \frac{1}{N}\sum_{n\in[N]} \frac{\alpha_n^D}{2\tilde{\beta}_1^D(1-\gamma^D)^2 h_{0,*}^D}\left(\frac{\sigma_D^2}{b} + M_D^2\right) + \frac{\beta_1^D}{\tilde{\beta}_1^D}\sqrt{W\mathrm{Dist}(\boldsymbol{w})\left(\frac{\sigma_D^2}{b} + M_D^2\right)},$$

where $H^D := \max_{j\in[W]} H_j^D$.

Proof of Theorem 3.1(ii). Let $\alpha_n^G := \alpha^G$ and $\alpha_n^D := \alpha^D$ for all $n \in \mathbb{N}$. Lemmas A.7 and A.8 thus guarantee that, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$ and all $N \geq 1$,

$$\frac{1}{N}\sum_{n\in[N]} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq \frac{\Theta\mathrm{Dist}(\boldsymbol{\theta})H^G}{2\alpha^G \tilde{\beta}_1^G N} + \frac{\alpha^G}{2\tilde{\beta}_1^G(1-\gamma^G)^2 h_{0,*}^G}\left(\frac{\sigma_G^2}{b} + M_G^2\right) + \frac{\beta_1^G}{\tilde{\beta}_1^G}\sqrt{\Theta\mathrm{Dist}(\boldsymbol{\theta})\left(\frac{\sigma_G^2}{b} + M_G^2\right)},$$

$$\frac{1}{N}\sum_{n\in[N]} \mathbb{E}\left[\langle \boldsymbol{w}_n - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq \frac{W\mathrm{Dist}(\boldsymbol{w})H^D}{2\alpha^D \tilde{\beta}_1^D N} + \frac{\alpha^D}{2\tilde{\beta}_1^D(1-\gamma^D)^2 h_{0,*}^D}\left(\frac{\sigma_D^2}{b} + M_D^2\right) + \frac{\beta_1^D}{\tilde{\beta}_1^D}\sqrt{W\mathrm{Dist}(\boldsymbol{w})\left(\frac{\sigma_D^2}{b} + M_D^2\right)},$$

which completes the proof. □

A.8. Proof of Theorem 3.2

Proof of Theorem 3.2. Condition (5) implies that TTUR achieves an ϵ–approximation. We have that, for $b > B_G/(\epsilon_G^2 - C_G)$,

$$\frac{\mathrm{d}N_G(b)}{\mathrm{d}b} = \frac{-A_G B_G}{\{(\epsilon_G^2 - C_G)b - B_G\}^2} \leq 0,$$
$$\frac{\mathrm{d}^2 N_G(b)}{\mathrm{d}b^2} = \frac{2A_G B_G(\epsilon_G^2 - C_G)}{\{(\epsilon_G^2 - C_G)b - B_G\}^3} \geq 0.$$

Hence, N_G is monotone decreasing and convex for $b > B_G/(\epsilon_G^2 - C_G)$. We also have that N_D is monotone decreasing and convex for $b > B_D/(\epsilon_D^2 - C_D)$. This completes the proof. □

A.9. Proof of Theorem 3.3

Proof of Theorem 3.3. We have that, for $b > B_G/(\epsilon_G^2 - C_G)$,

$$N_G(b)b = \frac{A_G b^2}{(\epsilon_G^2 - C_G)b - B_G}.$$

Hence,

$$\frac{\mathrm{d}N_G(b)b}{\mathrm{d}b} = \frac{A_G b\{(\epsilon_G^2 - C_G)b - 2B_G\}}{\{(\epsilon_G^2 - C_G)b - B_G\}^2},$$
$$\frac{\mathrm{d}^2 N_G(b)b}{\mathrm{d}b^2} = \frac{2A_G B_G^2}{\{(\epsilon_G^2 - C_G)b - B_G\}^3} \geq 0,$$

which implies that $N_G(b)b$ is convex for $b > B_G/(\epsilon_G^2 - C_G)$ and

$$\frac{\mathrm{d}N_G(b)b}{\mathrm{d}b} \begin{cases} < 0 & \text{if } b < b_G^\star, \\ = 0 & \text{if } b = b_G^\star = \frac{2B_G}{\epsilon_G^2 - C_G}, \\ > 0 & \text{if } b > b_G^\star. \end{cases}$$

The point $b_G^\star$ attains the minimum value $N_G(b_G^\star)b_G^\star$ of $N_G(b)b$. The discussion for N_D is similar to the one for N_G. This completes the proof. □

A.10. Proof of Proposition 3.4

Theorem 3.3 and the definitions of B_G, C_G, $\tilde{\beta_1^G}$ and $\tilde{\gamma}^G$ ensure that

$$b_G^\star = \frac{2B_G}{\epsilon_G^2 - C_G} > \frac{2B_G}{\epsilon_G^2} = \frac{2}{\epsilon_G^2} \cdot \frac{\sigma_G^2 \alpha^G}{2\tilde{\beta_1^G}\tilde{\gamma}^{G2} h_{0,*}^G} = \frac{\sigma_G^2 \alpha^G}{\epsilon_G^2 (1-\beta_1^G)(1-\gamma^G)^2 h_{0,*}^G}.$$

Moreover, (10) and the definition of $L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)$ ensure that

$$\|\nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n)\|^2 \leq \frac{\sigma_G^2}{b} + \|\nabla L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2 \approx \|\nabla L_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\|^2 = \frac{1}{|S|^2} \left\| \sum_{i=1}^{|S|} \nabla L_G^{(i)}(\boldsymbol{\theta}_n, \boldsymbol{w}_n) \right\|^2,$$

where $\sigma_G^2/b \approx 0$ holds when b is sufficiently large. We define $\sum_{i=1}^{|S|} \nabla L_G^{(i)}(\boldsymbol{\theta}_n, \boldsymbol{w}_n) := \boldsymbol{G}_n$ for all $n \in \mathbb{N}$. Then, we have

$$\frac{1}{|S|^2} \left\| \sum_{i=1}^{|S|} \nabla L_G^{(i)}(\boldsymbol{\theta}_n, \boldsymbol{w}_n) \right\|^2 = \frac{1}{|S|^2} \sum_{i=1}^{\Theta} G_{n,i}^2 \leq \frac{\Theta}{|S|^2} \max_{i \in [\Theta]} G_{n,i}^2.$$

Proof of Proposition 3.4(i). The definition of $\boldsymbol{v}_n^G$ implies that

$$v_{n,i}^G = \beta_2^G v_{n-1,i}^G + (1-\beta_2^G) g_{n,i}^2 \leq \max_{n,i} g_{n,i}^2 =: g_{n^*,i^*}^2 \leq \sum_{i=1}^{\Theta} g_{n^*,i}^2 = \|\nabla L_{G,\mathcal{S}_{n^*}}(\boldsymbol{\theta}_{n^*})\|^2,$$

where the first inequality can be shown by induction. Therefore, for all $n \in \mathbb{N}$ and all $i \in [\Theta]$,

$$v_{n,i}^G \leq \frac{\Theta}{|S|^2} \max_{i \in [\Theta]} G_{n^*,i}^2.$$

From the definition of $\bar{v}_{n,i}^G$ and $\beta_2 \in [0,1)$, we have

$$h_{0,*}^G := \min_{i \in [\Theta]} h_{0,i}^G \leq \sqrt{\bar{v}_{n,i}^G} = \sqrt{\frac{v_{n,i}^G}{1-\beta_2^{G^n}}} \leq \sqrt{\frac{v_{n,i}^G}{1-\beta_2^G}}.$$

Hence,

$$b_G^\star \geq \frac{\sigma_G^2 \alpha^G}{\epsilon_G^2 (1-\beta_1^G)(1-\gamma^G)^2 \sqrt{\frac{\Theta}{1-\beta_2^G} \frac{1}{|S|^2} \max_{i \in [\Theta]} G_{n^*,i}^2}}.$$

Accordingly, using $\gamma^G = \beta_1^G$ and $\max_{i \in [\Theta]} G_{n^*,i}^2 \approx \epsilon_G^2$ implies that

$$b_G^\star \geq \frac{\sigma_G^2 \alpha^G}{\epsilon_G^3 (1-\beta_1^G)^3 \sqrt{\frac{\Theta}{1-\beta_2^G} \frac{1}{|S|^2}}}.$$

The discussion for $b_D^\star$ is similar to the one for $b_G^\star$. This completes the proof. □

Proof of Proposition3.4(ii). The definition of $\boldsymbol{s}_n^G$ implies that

$$s_{n,i}^G = \beta_2^G v_{n-1,i}^G + (1-\beta_2^G)(g_{n,i} - m_{n,i})^2 \leq \max_{n,i}(g_{n,i} - m_{n,i})^2 =: (g_{n^*,i^*} - m_{n^*,i^*})^2$$
$$\leq \sum_{i=1}^{\Theta}(g_{n^*,i} - m_{n^*,i})^2 = \left\|\nabla L_{G,\mathcal{S}_{n^*}}(\boldsymbol{\theta}_{n^*}) - \boldsymbol{m}_{n^*}^G\right\|^2,$$

where the first inequality can be shown by induction. Hence, we have

$$\left\|\nabla L_{G,\mathcal{S}_{n^*}}(\boldsymbol{\theta}_{n^*}) - \boldsymbol{m}_{n^*}^G\right\|^2 \leq 2\left\|\nabla L_{G,\mathcal{S}_{n^*}}(\boldsymbol{\theta}_{n^*})\right\|^2 + 2\left\|\boldsymbol{m}_{n^*}^G\right\|^2 \leq \frac{4\sigma_G^2}{b} + 4\left\|\nabla L_G(\boldsymbol{\theta}_{n^*}, \boldsymbol{w}_{n^*})\right\|^2$$

Therefore, for all $n \in \mathbb{N}$ and all $i \in [\Theta]$,

$$s_{n,i}^G \leq \frac{4\Theta}{|S|^2}\max_{i\in[\Theta]} G_{n^*,i}^2.$$

From the definition of $\hat{s}_{n,i}^G$ and $\beta_2 \in [0,1)$, we have

$$h_{0,*}^G := \min_{i\in[\Theta]} h_{0,i}^G \leq \sqrt{\hat{s}_{n,i}^G} = \sqrt{\frac{s_{n,i}^G}{1-\beta_2^{G\,n}}} \leq \sqrt{\frac{s_{n,i}^G}{1-\beta_2^G}}.$$

Hence,

$$b_G^\star \geq \frac{\sigma_G^2\alpha^G}{\epsilon_G^2(1-\beta_1^G)(1-\gamma^G)^2\sqrt{\frac{4\Theta}{1-\beta_2^G}\frac{1}{|S|^2}\max_{i\in[\Theta]} G_{n^*,i}^2}}.$$

Accordingly, using $\gamma^G = \beta_1^G$ and $\max_{i\in[\Theta]} G_{n^*,i}^2 \approx \epsilon_G^2$ implies that

$$b_G^\star \geq \frac{\sigma_G^2\alpha^G}{\epsilon_G^3(1-\beta_1^G)^3\sqrt{\frac{4\Theta}{1-\beta_2^G}\frac{1}{|S|^2}}}.$$

The discussion for $b_D^\star$ is similar to the one for $b_G^\star$. This completes the proof. □

Proof of Proposition 3.4(iii). From the definition of $\boldsymbol{v}_n^G$ and a discussion similar to the proof of Proposition 3.4(i), we have

$$v_{n,i}^G \leq \frac{\Theta}{|S|^2}\max_{i\in[\Theta]} G_{n^*,i}^2.$$

From the definition of $v_{n,i}^G$, we have

$$h_{0,*}^G := \min_{i\in[\Theta]} h_{0,i}^G \leq \sqrt{v_{n,i}^G}.$$

Hence,

$$b_G^\star \geq \frac{\sigma_G^2\alpha^G}{\epsilon_G^2(1-\beta_1^G)(1-\gamma^G)^2\sqrt{\frac{\Theta}{|S|^2}\max_{i\in[\Theta]} G_{n^*,i}^2}}.$$

Accordingly, using $\gamma^G = \beta_1^G = 0$ and $\max_{i\in[\Theta]} G_{n^*,i}^2 \approx \epsilon_G^2$ implies that

$$b_G^\star \geq \frac{\sigma_G^2\alpha^G}{\epsilon_G^3\sqrt{\frac{\Theta}{|S|^2}}}.$$

The discussion for $b_D^\star$ is similar to the one for $b_G^\star$. This completes the proof. □

A.11. Relationship between stationary point problem and variational inequality

Proposition A.1. *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and $\boldsymbol{x}^*$ is a stationary point of f. Then, $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ is equivalent to the following variational inequality: for all $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0.$$

Proof of Proposition A.1. Suppose that $\boldsymbol{x} \in \mathbb{R}^d$ satisfies $\nabla f(\boldsymbol{x}) = \mathbf{0}$. Then, for all $\boldsymbol{y} \in \mathbb{R}^d$,

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \geq 0.$$

Suppose that $\boldsymbol{x} \in \mathbb{R}^d$ satisfies $\langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \geq 0$ for all $\boldsymbol{y} \in \mathbb{R}^d$. Let $\boldsymbol{y} := \boldsymbol{x} - \nabla f(\boldsymbol{x})$. Then we have

$$0 \leq \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle = -\|\nabla f(\boldsymbol{x})\|^2.$$

Hence,

$$\nabla f(\boldsymbol{x}) = \mathbf{0}.$$

□

A.12. Remarks regarding (C3)

We make the following remarks regarding (C3).

(C3)(i) Let $L_G^{(i)}(\cdot, \boldsymbol{w}) \colon \mathbb{R}^\Theta \to \mathbb{R}$ $(i \in \mathcal{S})$ be convex and $\nabla_{\boldsymbol{\theta}} L_G^{(i)}(\cdot, \boldsymbol{w}) \colon \mathbb{R}^\Theta \to \mathbb{R}^\Theta$ be Lipschitz continuous with the Lipschitz constant l_G. Let $L_D^{(i)}(\boldsymbol{\theta}, \cdot) \colon \mathbb{R}^W \to \mathbb{R}$ $(i \in \mathcal{R})$ be convex and $\nabla_{\boldsymbol{w}} L_D^{(i)}(\boldsymbol{\theta}, \cdot) \colon \mathbb{R}^W \to \mathbb{R}^W$ be Lipschitz continuous with the Lipschitz constant l_D. Then, the sequences $(\boldsymbol{\theta}_n)_{n \in \mathbb{N}}$ and $(\boldsymbol{w}_n)_{n \in \mathbb{N}}$ generated by SGD with $\alpha^G \leq 1/l_G$ and $\alpha_D \leq 1/l_D$ satisfy (C3).

(C3)(ii) Let $L_G^{(i)}(\cdot, \boldsymbol{w}) \colon \mathbb{R}^\Theta \to \mathbb{R}$ $(i \in \mathcal{S})$ and $L_D^{(i)}(\boldsymbol{\theta}, \cdot) \colon \mathbb{R}^W \to \mathbb{R}$ $(i \in \mathcal{R})$ be nonconvex. Let $B^G \subset \mathbb{R}^\Theta$ and $B^D \subset \mathbb{R}^W$ be closed balls defined by $B^G := \{\boldsymbol{\theta} \in \mathbb{R}^\Theta \colon \|\boldsymbol{\theta} - \boldsymbol{c}^G\| \leq r^G\}$ and $B^D := \{\boldsymbol{w} \in \mathbb{R}^W \colon \|\boldsymbol{w} - \boldsymbol{c}^D\| \leq r^D\}$, where $\boldsymbol{c}^G \in \mathbb{R}^\Theta$, $\boldsymbol{c}^D \in \mathbb{R}^W$, and $r^G, r^D > 0$ are large radii of B^G and B^D. We replace $\boldsymbol{\theta}_{n+1} := \boldsymbol{\theta}_n + \alpha_n^G \boldsymbol{d}_n^G$ and $\boldsymbol{w}_{n+1} := \boldsymbol{w}_n + \alpha_n^D \boldsymbol{d}_n^D$ in Algorithm 1 with

$$\boldsymbol{\theta}_{n+1} := P_G\left(\boldsymbol{\theta}_n + \alpha_n^G \boldsymbol{d}_n^G\right) \text{ and } \boldsymbol{w}_{n+1} := P_D\left(\boldsymbol{w}_n + \alpha_n^D \boldsymbol{d}_n^D\right), \tag{22}$$

where P_G is the projection onto B^G and P_D is the projection onto B^D. Then, the sequences $(\boldsymbol{\theta}_n)_{n \in \mathbb{N}} \subset B^G$ and $(\boldsymbol{w}_n)_{n \in \mathbb{N}} \subset B^D$ generated by (22) are bounded. Thus, the nonexpansivity conditions of P_G and P_D guarantee that Algorithm 1 using (22) satisfies (2) without assuming (C3).

Proof of (C3)(i). Let $n \in \mathbb{N}$. We define $\phi_n \colon \mathbb{R}^\Theta \to \mathbb{R}$ for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$ by

$$\phi_n(\boldsymbol{\theta}) := L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n, \boldsymbol{w}_n) + \langle \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle + \frac{1}{2\alpha^G}\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|^2.$$

The convexity of $L_{G,\mathcal{S}_n}(\cdot, \boldsymbol{w}_n)$ implies that, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$, $L_{G,\mathcal{S}_n}(\boldsymbol{\theta}, \boldsymbol{w}_n) \geq L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n, \boldsymbol{w}_n) + \langle \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n), \boldsymbol{\theta} - \boldsymbol{\theta}_n \rangle$. Hence, we have that, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$,

$$\phi_n(\boldsymbol{\theta}) \leq L_{G,\mathcal{S}_n}(\boldsymbol{\theta}, \boldsymbol{w}_n) + \frac{1}{2\alpha^G}\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|^2. \tag{23}$$

Moreover, from $\nabla_{\boldsymbol{\theta}} \phi_n(\boldsymbol{\theta}) = \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) + (1/\alpha^G)(\boldsymbol{\theta} - \boldsymbol{\theta}_n)$, ϕ_n is strongly convex with a constant $1/\alpha^G$. Accordingly, there exists a unique minimizer $\tilde{\boldsymbol{\theta}}_n$ of ϕ_n such that

$$\mathbf{0} = \nabla_{\boldsymbol{\theta}} \phi_n(\tilde{\boldsymbol{\theta}}_n) = \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) + \frac{1}{\alpha^G}(\boldsymbol{\theta} - \boldsymbol{\theta}_n),$$

i.e.,

$$\tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_n - \alpha^G \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n) =: \boldsymbol{\theta}_{n+1}.$$

The strong convexity of ϕ_n and $\nabla_{\boldsymbol{\theta}} \phi_n(\boldsymbol{\theta}_{n+1}) = \mathbf{0}$ imply that

$$\begin{aligned} \phi_n(\boldsymbol{\theta}) &\geq \phi_n(\boldsymbol{\theta}_{n+1}) + \langle \nabla_{\boldsymbol{\theta}} \phi_n(\boldsymbol{\theta}_{n+1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n+1} \rangle + \frac{1}{2\alpha^G} \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|^2 \\ &= \phi_n(\boldsymbol{\theta}_{n+1}) + \frac{1}{2\alpha^G} \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|^2. \end{aligned} \tag{24}$$

The condition $\alpha^G \leq 1/l_G$ and the Lipschitz continuity of $\nabla L_{G,\mathcal{S}_n}(\cdot)$ ensure that

$$\begin{aligned} \phi_n(\boldsymbol{\theta}_{n+1}) &:= L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n, \boldsymbol{w}_n) + \langle \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle + \frac{1}{2\alpha^G} \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 \\ &\geq L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n, \boldsymbol{w}_n) + \langle \nabla L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_n), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle + \frac{l_G}{2} \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 \\ &\geq L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_{n+1}, \boldsymbol{w}_n). \end{aligned} \tag{25}$$

From (23), (24), and (25),

$$L_{G,\mathcal{S}_n}(\boldsymbol{\theta}, \boldsymbol{w}_n) + \frac{1}{2\alpha^G} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|^2 \geq L_{G,\mathcal{S}_n}(\boldsymbol{\theta}_{n+1}, \boldsymbol{w}_n) + \frac{1}{2\alpha^G} \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|^2,$$

which implies that, for all $\boldsymbol{\theta} \in \mathbb{R}^\Theta$,

$$L_G(\boldsymbol{\theta}, \boldsymbol{w}_n) + \frac{1}{2\alpha^G} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|^2 \geq L_G(\boldsymbol{\theta}_{n+1}, \boldsymbol{w}_n) + \frac{1}{2\alpha^G} \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|^2.$$

Let $\boldsymbol{\theta}^* \in \mathbb{R}^\Theta$ be a minimizer of $L_G(\boldsymbol{\theta}, \boldsymbol{w}_n)$. Then, we have

$$\frac{1}{2\alpha^G} \left(\|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}^*\|^2 - \|\boldsymbol{\theta}_n - \boldsymbol{\theta}^*\|^2 \right) \leq L_G(\boldsymbol{\theta}^*, \boldsymbol{w}_n) - L_G(\boldsymbol{\theta}_{n+1}, \boldsymbol{w}_n) \leq 0.$$

Since $(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^*\|^2)_{n \in \mathbb{N}}$ is monotone decreasing, the sequence $(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}^*\|^2)_{n \in \mathbb{N}}$ is bounded. The discriminator can be defined by replacing G, $\mathcal{S}_n$, $\boldsymbol{\theta}$, and $\boldsymbol{w}$ in the generator with D, $\mathcal{R}_n$, $\boldsymbol{w}$, and $\boldsymbol{\theta}$. □

Details of (C3)(ii): Let $\boldsymbol{\theta} \in B^G \subset \mathbb{R}^\Theta$ and $n \in \mathbb{N}$. The definition of $\boldsymbol{\theta}_{n+1}$ and the nonexpansivity of P_G (i.e., $\|P_G(\boldsymbol{\theta}_1) - P_G(\boldsymbol{\theta}_2)\|_{\mathsf{H}} \leq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{\mathsf{H}}$ $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^\Theta, \mathsf{H} \in \mathbb{S}^\Theta_{++})$) imply that

$$\begin{aligned} \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}\|^2_{\mathsf{H}^G_n} &= \|P_G(\boldsymbol{\theta}_n + \alpha^G_n \boldsymbol{d}^G_n) - P_G(\boldsymbol{\theta})\|^2_{\mathsf{H}^G_n} \\ &\leq \|(\boldsymbol{\theta}_n - \boldsymbol{\theta}) + \alpha^G_n \boldsymbol{d}^G_n\|^2_{\mathsf{H}^G_n} = \|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|^2_{\mathsf{H}^G_n} + 2\alpha^G_n \langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \boldsymbol{d}^G_n \rangle_{\mathsf{H}^G_n} + {\alpha^G_n}^2 \|\boldsymbol{d}^G_n\|^2_{\mathsf{H}^G_n}. \end{aligned}$$

Hence, a discussion similar to the one proving Theorem 3.1 ensures that Algorithm 1 using (22) satisfies (2) without assuming (C3).