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. 2012; Jordan et al. 2013), 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. 2018; 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) (Feitman et al. 2020; Scaman & Malherbe 2020; Chen et al. 2020); let \( \theta_n \) be the point generated by the generator at iteration \( n \) and \( w_n \) be the point generated by the discriminator at iteration \( n \). The loss function of the generator \( L_G(\cdot, w_n) \) is minimized by using SGD with a learning rate \( \alpha_n^G \) and a mini-batch stochastic gradient at \( \theta_n \) with a batch of size \( b \). The loss function of the discriminator \( L_D(\theta_n, \cdot) \) is minimized with a learning rate \( \alpha_n^D \) and a mini-batch stochastic gradient at \( w_n \) with the same batch size \( b \). If \( \alpha_n^G \) and \( \alpha_n^D \) 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 \( \alpha_n^G \) and \( \alpha_n^D \) 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 \( \alpha^G \) and \( \alpha^D \) 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>.

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 $\alpha^G$ and $\alpha^D$ 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 (11) of the gradients of both loss functions.

Previous results (Shalug et al. 2019; Zhang et al. 2019; Indira, 2022a; Koval 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 $(\theta^*, w^*)$ of stationary points of the loss function $L_D(\theta^*, \cdot)$ of the discriminator and the loss function $L_G(\cdot, w^*)$ of the generator.

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

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

which are equivalent to the following variational inequalities (see Appendix A.1): for all $\theta \in \mathbb{R}^\Theta$ and all $w \in \mathbb{R}^W$,

$$\langle \theta^* - \theta, \nabla_{\theta} L_G(\theta^*, w^*) \rangle \leq 0 \quad \text{and} \quad \langle w^* - w, \nabla_{w} L_D(\theta^*, w^*) \rangle \leq 0. \quad (1)$$

Let us examine TTURs based on adaptive methods (see Algorithm 11 and Table 4 in Appendix A.1) such as Adam (Kingma & Ba 2015), AdaBelief (Zhuang et al. 2020), and RMSProp (Tieleman & Hinton, 2012). Let $((\theta_n, w_n))_{n \in \mathbb{N}} \subset \mathbb{R}^\Theta \times \mathbb{R}^W$ be the sequence generated by TTUR with constant learning rates $\alpha^G$ and $\alpha^D$. We will show that, under certain assumptions, for all $\theta \in \mathbb{R}^\Theta$ and all $w \in \mathbb{R}^W$,

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{E} [\langle \theta_n - \theta, \nabla_{\theta} L_G(\theta_n, w_n) \rangle] \leq \Theta Dist(\theta) H^G \left( \frac{\sigma^2_G \alpha^G}{2 \beta_1^G \beta_2^G} \frac{1}{b} \right) + \mathbb{E} [\langle w_n - w, \nabla_{w} L_D(\theta_n, w_n) \rangle]$$

$$\leq W Dist(w) H^D \left( \frac{\sigma^2_D \alpha^D}{2 \beta_1^D \beta_2^D} \frac{1}{b} \right) + \mathbb{E} [\langle w_n - w, \nabla_{w} L_D(\theta_n, w_n) \rangle]$$

$$\leq \frac{1}{N} \sum_{n=1}^{N} \mathbb{E} [\langle \theta_n - \theta, \nabla_{\theta} L_G(\theta_n, w_n) \rangle] \leq \frac{1}{N} \sum_{n=1}^{N} \mathbb{E} [\langle w_n - w, \nabla_{w} L_D(\theta_n, w_n) \rangle]$$

$$\leq \frac{W Dist(w) H^D}{2 \beta_1^D \beta_2^D} \frac{1}{b} \left( \frac{\sigma^2_D \alpha^D}{2 \beta_1^D \beta_2^D} \frac{1}{b} \right) + \mathbb{E} [\langle w_n - w, \nabla_{w} L_D(\theta_n, w_n) \rangle]$$

$$\leq \frac{1}{N} \sum_{n=1}^{N} \mathbb{E} [\langle \theta_n - \theta, \nabla_{\theta} L_G(\theta_n, w_n) \rangle] \leq \frac{1}{N} \sum_{n=1}^{N} \mathbb{E} [\langle w_n - w, \nabla_{w} L_D(\theta_n, w_n) \rangle]$$

$$\leq \frac{W Dist(w) H^D}{2 \beta_1^D \beta_2^D} \frac{1}{b} \left( \frac{\sigma^2_D \alpha^D}{2 \beta_1^D \beta_2^D} \frac{1}{b} \right) + \mathbb{E} [\langle w_n - w, \nabla_{w} L_D(\theta_n, w_n) \rangle]$$

(2)
makes number of steps $N$ of $b$. Moreover, steps are useful for training GANs. Meanwhile, using small constant learning rates $\alpha$ 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 $\epsilon$-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 $b$,

$$\begin{align*}
N_G(b) &:= \frac{A_Gb}{\epsilon_G^2 - C_G} - B_G, \\
N_D(b) &:= \frac{A_Db}{\epsilon_D^2 - C_D} - B_D,
\end{align*}$$

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

$$\frac{1}{N_G} \sum_{n=1}^{N_G} \mathbb{E}[|\theta_n - \theta, \nabla_\theta L_G(\theta_n, w_n)|] \leq \epsilon_G$$

and

$$\frac{1}{N_D} \sum_{n=1}^{N_D} \mathbb{E}[|w_n - w, \nabla_w L_D(\theta_n, w_n)|] \leq \epsilon_D.$$

Moreover, $N_G$ and $N_D$ are monotonically decreasing and convex functions of the batch size (Theorem 5.2). This result implies that large batch sizes are desirable in the sense of minimizing the number of steps for training GANs. 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 (Hida et al. 2022) 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 $b^*_G$ and $b^*_D$ of $N_G(b)b$ and $N_D(b)b$ (Theorem 5.3). Accordingly, $b^*_G$ and $b^*_D$ are given by

$$\begin{align*}
b^*_G &:= \frac{2B_G}{\epsilon_G^2 - C_G}, \\
b^*_D &:= \frac{2B_D}{\epsilon_D^2 - C_D}.
\end{align*}$$

It would be desirable to use critical batch sizes $b^*_G$ and $b^*_D$ as it minimizes the SFO complexity, which is the computational cost of the stochastic gradient. Furthermore, we show that lower bounds for $b^*_G$ and $b^*_D$ can be estimated from some hyperparameters, the total number of datasets, and the number of dimensions of the model (Proposition 5.4).

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. 2018), 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 $b^*_G$ and $b^*_D$ in (3) and compare them with ones measured in numerical experiments. We find that the estimated sizes are close to the measured ones (see Section 4.3).

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) $L^{(1)}_G(\cdot, w) : \mathbb{R}^\Theta \to \mathbb{R}$ and $L^{(1)}_D(\theta, \cdot) : \mathbb{R}^W \to \mathbb{R}$ are continuously differentiable.

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

(i) For each iteration $n$,

$$\mathbb{E}_{\xi_G} [G_{\xi_G}(\theta_n)] = \nabla_\theta L_G(\theta_n, w_n)$$

and

$$\mathbb{E}_{\xi_D} [G_{\xi_D}(w_n)] = \nabla_w L_D(\theta_n, w_n).$$

(ii) There exist nonnegative constants $\sigma^2_G$ and $\sigma^2_D$ such that

$$\mathbb{E}_{\xi_G} [||G_{\xi_G}(\theta_n) - \nabla_\theta L_G(\theta_n, w_n)||^2] \leq \sigma^2_G$$

and

$$\mathbb{E}_{\xi_D} [|G_{\xi_D}(w_n) - \nabla_w L_D(\theta_n, w_n)|^2] \leq \sigma^2_D.$$
Table 1. Notation List (The parameters and functions of the discriminator are defined by replacing \( G, S_n, \theta \) and \( w \) in the generator with \( D, R_n, w \) and \( \theta \))

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>The set of all nonnegative integers</td>
</tr>
<tr>
<td>([N])</td>
<td>([N] := {1, 2, \ldots, N} (N \in N \setminus {0}))</td>
</tr>
<tr>
<td>(</td>
<td>A</td>
</tr>
<tr>
<td>( \mathbb{R}^d )</td>
<td>A ( d )-dimensional Euclidean space with inner product ((\cdot, \cdot)), which induces the norm ( | \cdot | )</td>
</tr>
<tr>
<td>( \mathbb{R}_+^d )</td>
<td>A random variable supported on ( \mathbb{R}_+^d )</td>
</tr>
<tr>
<td>( \mathbb{R}_+^{d,+} )</td>
<td>( {x \in \mathbb{R}^d: x_i \geq 0 \ (i \in [d])} )</td>
</tr>
<tr>
<td>( \mathbb{R}_+^{d,+} )</td>
<td>( {x \in \mathbb{R}^d: x_i &gt; 0 \ (i \in [d])} )</td>
</tr>
<tr>
<td>( S_{n,i} )</td>
<td>The set of ( d \times d ) symmetric positive-definite matrices</td>
</tr>
<tr>
<td>( \mathbb{D}^d )</td>
<td>The set of ( d \times d ) diagonal matrices, i.e., ( \mathbb{D}^d = {M \in \mathbb{R}^{d \times d}: M = \text{diag}(x_i), x_i \in \mathbb{R} \ (i \in [d])} )</td>
</tr>
<tr>
<td>( \mathbb{E}[X] )</td>
<td>The expectation with respect to ( \xi ) of a random variable ( X )</td>
</tr>
<tr>
<td>( S )</td>
<td>A set of synthetic samples ( z^{(i)} )</td>
</tr>
<tr>
<td>( R )</td>
<td>A set of real-world samples ( x^{(i)} )</td>
</tr>
<tr>
<td>( S_m )</td>
<td>Mini-batch of ( b ) synthetic samples ( z^{(i)} ) at time ( n )</td>
</tr>
<tr>
<td>( R_m )</td>
<td>Mini-batch of ( b ) real-world samples ( x^{(i)} ) at time ( n )</td>
</tr>
<tr>
<td>( L_{G,\xi}^{(i)}(\cdot, w) )</td>
<td>A loss function of the generator for ( w \in \mathbb{R}^W ) and ( z^{(i)} )</td>
</tr>
<tr>
<td>( L_G(\cdot, w) )</td>
<td>The total loss function of the generator for ( w \in \mathbb{R}^W ), i.e., ( L_G(\cdot, w) :=</td>
</tr>
<tr>
<td>( \xi_G^0 )</td>
<td>A random variable supported on ( \Xi_G ) that does not depend on ( w \in \mathbb{R}^W ) and ( \theta \in \mathbb{R}^\Theta )</td>
</tr>
<tr>
<td>( \xi_G^0, \xi_G^1, \ldots ) are independent samples and ( \xi_G^n ) is independent of ( \theta_k ) for ( k = 0 \subset \mathbb{R}^\Theta ) and ( w \in \mathbb{R}^W )</td>
<td></td>
</tr>
<tr>
<td>( \xi_{n,i}^{(i)} )</td>
<td>A random variable generated from the ( i )-th sampling at time ( n )</td>
</tr>
<tr>
<td>( G_{n,i}^{(i)}(\theta) )</td>
<td>The history of process ( \xi_G^{(i)} ) at time ( n ), i.e., ( \xi_{n,i}^{(i)} := (\xi_G^0, \xi_G^1, \ldots, \xi_G^{n-1}) )</td>
</tr>
<tr>
<td>( \nabla L_{G,S_n}(\theta_n) )</td>
<td>The stochastic gradient of ( L_G(\cdot, w) ) at ( \theta \in \mathbb{R}^\Theta )</td>
</tr>
<tr>
<td>( \nabla L_{G,S_n}(\theta_n) )</td>
<td>The mini-batch stochastic gradient of ( L_G(\cdot, w_n) ) for ( S_n ), i.e., ( \nabla L_{G,S_n}(\theta_n) := b^{-1} \sum_{i \in [b]} G_{n,i}^{(i)}(\theta_n) )</td>
</tr>
</tbody>
</table>

2.2. Adaptive method

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

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

**Assumption 2.2.**

(A1) \( H_{n,j}^G = \text{diag}(h_{n,j}^G) \) depends on \( \xi_{n,j}^G \) and \( H_{n,j}^D = \text{diag}(h_{n,j}^D) \) depends on \( \xi_{n,j}^D \). Moreover, \( h_{n,i,j}^G \geq h_{n,i,j}^D \) and \( h_{n+1,i,j} \geq h_{n,i,j}^D \) 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 \( H_{n,i}^G \) such that \( \sup_{n \in \mathbb{N}} \mathbb{E}[h_{n,i}^G] \leq H_{n,i}^G \). For all \( j \in \mathbb{N} \), there exists a positive number \( H_{n,j}^D \) such that \( \sup_{n \in \mathbb{N}} \mathbb{E}[h_{n,j}^D] \leq H_{n,j}^D \).
Examples of $H^G_n \in S^G_{\text{pos}} \cap \mathbb{D}^\Theta$ and $H^D_n \in S^D_{\text{pos}} \cap \mathbb{D}^W$ satisfying Assumption 2.1 are listed in Table 1 (see Appendix A.1). By referring to the results in (Chen et al., 2019; Iiduka, 2022a; Zhuang et al., 2020), we can check that $H^G_n$ and $H^D_n$ in Table 1 satisfy (A1) and (A2).

3. Main Results

3.1. Convergence analysis of Algorithm II 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 $\mathbb{E}[\|\nabla_\theta L_G(\theta_n, w_n)\|^2] \leq M^G_2$ and $\mathbb{E}[\|\nabla_w L_D(\theta_n, w_n)\|^2] \leq M^D_2$.

(C3) For all $\theta = (\theta_i) \in \mathbb{R}^\Theta$ and all $w = (w_i) \in \mathbb{R}^W$, there exist positive numbers Dist($\theta$) and Dist($\theta$) such that $\max_{i \in [\Theta]} \sup \{\theta_n - \theta_i : n \in \mathbb{N}\} \leq \text{Dist}(\theta)$ and $\max_{i \in [W]} \sup \{w_n - w_i : n \in \mathbb{N}\} \leq \text{Dist}(w)$.

A previous study (Heusel et al., 2017) used a decaying learning rate $O(n^{-\tau})$, where $\tau \in (0, 1]$, for TTUR based on Adam to train GANs. This paper investigates the analogies of the performance of TTUR based on adaptive methods using constant learning rates defined by (C1). Condition (C2) provides upper bounds on the performance measures $(1/N) \sum_{n=1}^N \mathbb{E}[\|\theta_n - \theta, \nabla_\theta L_G(\theta_n, w_n)\|^2]$ and $(1/N) \sum_{n=1}^N \mathbb{E}[\|w_n - w, \nabla_w L_D(\theta_n, w_n)\|^2]$ (see (2) and Theorem A.4 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 & Bai, 2013; Reddi et al., 2018; Zhuang et al., 2020)). See Appendix A.12 for remarks regarding (C3).

The following is a convergence analysis of Algorithm II (The proof of Theorem 3.1 is in Appendices A.3 and A.17).

**Theorem 3.1.** Suppose that Assumptions 2.1, 2.2, 3.1 and 3.2 hold and consider the sequence $((\theta_n, w_n))_{n \in \mathbb{N}}$ generated by Algorithm II. Then, the following hold:

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

$$\liminf_{n \to +\infty} \mathbb{E}[\|\theta_n - \theta, \nabla_\theta L_G(\theta_n, w_n)\|^2] \leq \frac{\alpha^G}{\beta^G} \left( \frac{\sigma_n^2}{b} + M^G_2 \right) + \frac{1}{\beta^G} \sqrt{\text{Dist}(\theta) \left( \frac{\sigma_n^2}{b} + M^G_2 \right)}.$$ 

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

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E}[\|\theta_n - \theta, \nabla_\theta L_G(\theta_n, w_n)\|^2] \leq \frac{\alpha^G}{\beta^G} \left( \frac{\sigma_n^2}{b} + M^G_2 \right) + \frac{1}{\beta^G} \sqrt{\text{Dist}(\theta) \left( \frac{\sigma_n^2}{b} + M^G_2 \right)},$$

where $\frac{\sigma_n^2}{b} = \frac{\sigma_n^2}{\sqrt{\text{Dist}(\theta)}}$ and $\beta^G_D := 1 - \beta^G_D$.

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 $\alpha^G$ and $\alpha^D$. Meanwhile, Theorem 3.1(ii) indicates that there is no evidence that using sufficiently small $\alpha^G$ and $\alpha^D$ is good for training GANs.
3.2. Relationship between batch size and number of steps for Algorithm G

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

**Theorem 3.2.** Suppose that Assumptions A.1, A.2, and A.3 hold and consider Algorithm G. Then, $N_G$ and $N_D$ defined by

$$N_G(b) := \frac{A_G b}{(\epsilon_G^2 - C_G)b - B_G} \leq N_G \text{ for } b \geq \frac{B_G}{\epsilon_G^2 - C_G},$$

$$N_D(b) := \frac{A_D b}{(\epsilon_D^2 - C_D)b - B_D} \leq N_D \text{ for } b \geq \frac{B_D}{\epsilon_D^2 - C_D}$$

satisfy

$$\frac{1}{N_G} \sum_{n=1}^{N_G} E[(\theta_n - \theta, \nabla_\theta L_G(\theta_n, w_n))] \leq \epsilon_G,$$

$$\frac{1}{N_D} \sum_{n=1}^{N_D} E[(w_n - w, \nabla_w L_D(\theta_n, w_n))] \leq \epsilon_D,$$

where $A_G, B_G, C_G, A_D, B_D, C_D$ are defined as in (2). Moreover, the functions $N_G(b)$ and $N_D(b)$ defined by (3) 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 (3) decrease as the batch size increases. Accordingly, it is useful to choose a sufficiently large $b$ in the sense of minimizing the number of steps $N$ needed for an $\epsilon$-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 (3) and for the batch size by

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

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

**Theorem 3.3.** Suppose that Assumptions A.1, A.2, and A.3 hold and consider Algorithm G. Then, there exist

$$b^*_G := \frac{2B_G}{\epsilon_G^2 - C_G} \text{ and } b^*_D := \frac{2B_D}{\epsilon_D^2 - C_D}$$

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

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 G. Then, $b^*_G$ and $b^*_D$ defined by (3) satisfy the following

(i) for Adam,

$$b^*_G \geq \frac{\sigma^2 G \alpha_G}{\epsilon_G^2(1 - \beta^G_1)^2} \sqrt{\frac{9}{1 - \beta^G_2}} \left| S \right|^2,$$

(ii) for AdaBelief,

$$b^*_G \geq \frac{\sigma^2 G \alpha_G}{\epsilon_G^2(1 - \beta^G_1)^2} \sqrt{\frac{169}{1 - \beta^G_2}} \left| S \right|^2,$$

(iii) for RMSProp,

$$b^*_G \geq \frac{\sigma^2 G \alpha_G}{\epsilon_G^2} \sqrt{\frac{4}{9} \left| S \right|^2} \text{ and } b^*_D \geq \frac{\sigma^2 D \alpha_D}{\epsilon_D^2} \sqrt{\frac{4W}{1 - \beta^D_2}} \left| S \right|^2,$$

where $\sigma^2 G, \sigma^2 D \geq 0$, $\alpha_G, \alpha_D, \epsilon_G, \epsilon_D > 0$, $\beta^G_1, \beta^G_2, \beta^D_1, \beta^D_2 \in (0, 1)$.

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 (Shallic et al. 2013, Zhang et al. 2015, Indukt, 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 $b^*$ at which $N(b)$ 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 $\xi$ 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 $H^G_n$ and $H^D_n$ defined in Table 2. 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 $\alpha^G$ and discriminator learning rate $\alpha^D$ (see Figure A in Appendix A2). Appendix A2 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 a 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 4 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 5 plots the number of steps $N$ needed to achieve a 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 7 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)
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 2.2 and 2.3).

4.4. Estimation of lower bound of critical batch size

Proposition 4.3 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 4, see Section 4.2 for the settings of $\alpha^G$, $\beta^G$, and so on. Figure 4 indicates that the measured critical batch sizes minimizing the SFO complexities of Adam, AdaBelief, and RMSProp are $2^5 = 32$, $2^3 = 32$, 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 (Shalir 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 5, in the DCGAN on the LSUN-Bedroom dataset setting, the measured critical batch size for Adam is $2^5$; using this and Proposition 4.3(i) to back-calculate $\sigma_2^G/\sigma_2^{G(i)}$ gives $\sigma_2^G/\sigma_2^{G(i)} = 788.7$. We can use this ratio and Proposition 4.3(ii), (iii) to estimate a lower bound of 47.9 for AdaBelief and 126.5 for RMSProps (see Table 5).

Figure 6 indicates that the measured critical batch sizes for Adam, AdaBelief, and RMSProp are $2^2 = 4$, and $2^6 = 64$, respectively. As in Figure 4 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 8, in the BigGAN on the ImageNet dataset setting, the measured critical batch size for Adam is $2^6$; using this and Proposition 4.3(i) to back-calculate

<table>
<thead>
<tr>
<th>Table 3. Measured and estimated critical batch sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 4.1</td>
</tr>
<tr>
<td>Adam</td>
</tr>
<tr>
<td>AdaBelief</td>
</tr>
<tr>
<td>RMSProp</td>
</tr>
</tbody>
</table>

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.
\[ \sigma^2_G \beta \] gives \( \sigma^2_H / \epsilon^2 \) = 530303.8. We can use this ratio and Proposition 3.4(iii) 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 \( \beta_1 \) and \( \beta_2 \). 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 \( \beta_1 \) and \( \beta_2 \). In particular, \( \beta_1 \) in RMSProp is always 0 and \( \beta_2 \) 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 \( \beta_1 \) and \( \beta_2 \) 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


Existence and Estimation of Critical Batch Size for Training GANs with TTUR


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 \( x, y \in \mathbb{R}^d \) by \( \langle x, y \rangle_S := \langle x, Sy \rangle \) and the \( S \)-norm is defined by \( ||x||_S := \sqrt{\langle x, x \rangle_S} \).

The Hadamard product of \( \mathbb{R}^d \) is defined for all \( x = (x_i)_{i=1}^d \in \mathbb{R}^d \) by \( x \odot x := (x_i^2)_{i=1}^d \in \mathbb{R}^d \).

A.1. Examples of diagonal matrix in Algorithm \( \mathbb{R} \)

<table>
<thead>
<tr>
<th>Optimizer</th>
<th>Example (( \gamma^G = \gamma^D = \beta_1^G = \beta_1^D ))</th>
<th>Example (( \gamma^G = \gamma^D = \beta_1^G = \beta_1^D ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSProp</td>
<td>( p_n^G = \nabla L_G.S_n(\theta_n) \odot \nabla L_G.S_n(\theta_n) )</td>
<td>( p_n^D = \nabla L_{D.R_n}(w_n) \odot \nabla L_{D.R_n}(w_n) )</td>
</tr>
<tr>
<td></td>
<td>( v_n^G = \beta_2^G v_{n-1}^G + (1 - \beta_2^G) p_n^G )</td>
<td>( v_n^D = \beta_2^D v_{n-1}^D + (1 - \beta_2^D) p_n^D )</td>
</tr>
<tr>
<td></td>
<td>( \tilde{v}<em>n^G = \text{diag}(\sqrt{v</em>{n,i}^G}) )</td>
<td>( \tilde{v}<em>n^D = \text{diag}(\sqrt{v</em>{n,i}^D}) )</td>
</tr>
<tr>
<td>Adam</td>
<td>( p_n^G = \nabla L_G.S_n(\theta_n) \odot \nabla L_G.S_n(\theta_n) )</td>
<td>( p_n^D = \nabla L_{D.R_n}(w_n) \odot \nabla L_{D.R_n}(w_n) )</td>
</tr>
<tr>
<td></td>
<td>( v_n^G = \beta_2^G v_{n-1}^G + (1 - \beta_2^G) p_n^G )</td>
<td>( v_n^D = \beta_2^D v_{n-1}^D + (1 - \beta_2^D) p_n^D )</td>
</tr>
<tr>
<td></td>
<td>( \tilde{v}<em>n^G = \text{diag}(\sqrt{v</em>{n,i}^G}) )</td>
<td>( \tilde{v}<em>n^D = \text{diag}(\sqrt{v</em>{n,i}^D}) )</td>
</tr>
<tr>
<td>AMSGrad</td>
<td>( p_n^G = \nabla L_G.S_n(\theta_n) \odot \nabla L_G.S_n(\theta_n) )</td>
<td>( p_n^D = \nabla L_{D.R_n}(w_n) \odot \nabla L_{D.R_n}(w_n) )</td>
</tr>
<tr>
<td></td>
<td>( v_n^G = \beta_2^G v_{n-1}^G + (1 - \beta_2^G) p_n^G )</td>
<td>( v_n^D = \beta_2^D v_{n-1}^D + (1 - \beta_2^D) p_n^D )</td>
</tr>
<tr>
<td></td>
<td>( \tilde{v}<em>n^G = \text{diag}(\sqrt{v</em>{n,i}^G}) )</td>
<td>( \tilde{v}<em>n^D = \text{diag}(\sqrt{v</em>{n,i}^D}) )</td>
</tr>
<tr>
<td>AdaBelief</td>
<td>( p_n^G = \nabla L_G.S_n(\theta_n) - m_n^G )</td>
<td>( p_n^D = \nabla L_{D.R_n}(w_n) - m_n^D )</td>
</tr>
<tr>
<td></td>
<td>( s_n^G = p_n^G \odot \tilde{p}_n^G )</td>
<td>( s_n^D = p_n^D \odot \tilde{p}_n^D )</td>
</tr>
<tr>
<td></td>
<td>( s_n^G = \beta_2^G s_{n-1}^G + (1 - \beta_2^G) s_n^G )</td>
<td>( s_n^D = \beta_2^D s_{n-1}^D + (1 - \beta_2^D) s_n^D )</td>
</tr>
<tr>
<td></td>
<td>( \tilde{s}<em>n^G = \text{diag}(\sqrt{s</em>{n,i}^G}) )</td>
<td>( \tilde{s}<em>n^D = \text{diag}(\sqrt{s</em>{n,i}^D}) )</td>
</tr>
</tbody>
</table>

A.2. Hyperparameters of optimizers

<table>
<thead>
<tr>
<th>Optimizer</th>
<th>Example (( \gamma^G = \gamma^D = \beta_1^G = \beta_1^D ))</th>
<th>Example (( \gamma^G = \gamma^D = \beta_1^G = \beta_1^D ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam</td>
<td>0.0003</td>
<td>0.0001</td>
</tr>
<tr>
<td>AdaBelief</td>
<td>0.00003</td>
<td>0.0003</td>
</tr>
<tr>
<td>RMSProp</td>
<td>0.00003</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

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.
Lemma A.1. Suppose that (S1), (S2)(i), and (S3) hold and consider Algorithm 1. Then, for all $\theta \in \mathbb{R}^\Theta$ and all $n \in \mathbb{N}$,
\[
E \left[ \|\theta_{n+1} - \theta_H^2 \|_{H^2} \right] = E \left[ \|\theta_n - \theta_H^2 \|_{H^2} \right] + \alpha_n^G E \left[ \|\theta_n - \theta_H^2 \|_{H^2} \right] + 2\alpha_n^G \left\{ \frac{\beta_n^G}{\gamma^2} E \left[ \langle \theta - \theta_n, m_{n-1}^G \rangle \right] + \frac{\beta_n^G}{\gamma^2} E \left[ \langle \theta - \theta_n, \nabla_G \theta \rangle \right] \right\},
\]

![Figure 7](image_url)

Figure 7. Analysis of the relationship between combination of learning rates and FID score: discriminator learning rate $\alpha^D$ on the vertical axis and generator learning rate $\alpha^G$ 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](image_url)

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 4.2 and 4.3.

A.5. Lemmas

Lemma A.1. Suppose that (S1), (S2)(i), and (S3) hold and consider Algorithm 1. Then, for all $\theta \in \mathbb{R}^\Theta$ and all $n \in \mathbb{N}$,
where $\beta_1^G := 1 - \beta_1^D$ and $\gamma_n^G := 1 - \gamma^{G,n+1}$.

**Proof.** Let $\theta \in \mathbb{R}^d$ and $n \in \mathbb{N}$. The definition of $\theta_{n+1}$ implies that
\[
\|\theta_{n+1} - \theta\|^2_{H_n^G} = \|\theta_n - \theta\|^2_{H_n^G} + 2\alpha_n^G \langle \theta_n - \theta, d_n^G \rangle_{H_n^G} + \alpha_n^G \|d_n^G\|^2_{H_n^G}.
\]
Moreover, the definitions of $d_n^G$, $m_n^G$, and $m_n^G$ ensure that
\[
\langle \theta_n - \theta, d_n^G \rangle_{H_n^G} = \frac{1}{\gamma_n^G} \langle \theta_n - \theta, m_n^G \rangle = \frac{\beta_1^G}{\gamma_n^G} \langle \theta_n - \theta, m_{n-1}^G \rangle + \frac{\tilde{\gamma}_n^G}{\gamma_n^G} \langle \theta_n - \theta, \nabla L_{G,S_n}(\theta_n) \rangle.
\]
Hence,
\[
\|\theta_{n+1} - \theta\|^2_{H_n^G} = \|\theta_n - \theta\|^2_{H_n^G} + \alpha_n^G \|d_n^G\|^2_{H_n^G}
+ 2\alpha_n^G \left\{ \frac{\beta_1^G}{\gamma_n^G} \langle \theta_n - \theta, m_{n-1}^G \rangle + \frac{\tilde{\gamma}_n^G}{\gamma_n^G} \langle \theta_n - \theta, \nabla L_{G,S_n}(\theta_n) \rangle \right\}.
\] (9)

Conditions (S2)(i) and (S3) guarantee that
\[
\mathbb{E} \left[ \mathbb{E} \left[ \langle \theta_n - \theta, \nabla L_{G,S_n}(\theta_n) \rangle \right| \theta_n \right] = \mathbb{E} \left[ \langle \theta - \theta_n, \nabla L_{G,S_n}(\theta_n) \rangle \right| \theta_n \right] = \mathbb{E} \left[ \langle \theta - \theta_n, \nabla \alpha L_G(\theta_n, w_n) \rangle \right].
\]
The lemma follows by taking the expectation with respect to $\xi_n^G$ on both sides of (9). □

A discussion similar to the one proving Lemma 5.3 leads to the following lemma.

**Lemma A.2.** Suppose that (S1), (S2)(i), and (S3) hold and consider Algorithm 1. Then, for all $w \in \mathbb{R}^W$ and all $n \in \mathbb{N}$,
\[
\mathbb{E} \left[ \|w_{n+1} - w\|^2_{H_n^G} \right] = \mathbb{E} \left[ \|w_n - w\|^2_{H_n^G} \right] + \alpha_n^D \mathbb{E} \left[ \|d_n^D\|^2_{H_n^G} \right]
+ 2\alpha_n^D \left\{ \frac{\beta_1^D}{\gamma_n^D} \mathbb{E} \left[ \langle w - w_n, m_{n-1}^D \rangle \right] + \frac{\tilde{\gamma}_n^D}{\gamma_n^D} \mathbb{E} \left[ \langle w - w_n, \nabla \psi_{L_D}(\theta_n, w_n) \rangle \right] \right\},
\] where $\beta_1^D := 1 - \beta_1^D$ and $\gamma_n^D := 1 - \gamma^{D,n+1}$.

**Lemma A.3.** Algorithm 1 satisfies that, under (S2)(i), (ii) and (C2), for all $n \in \mathbb{N}$,
\[
\mathbb{E} \left[ \|m_n^G\|^2 \right] \leq \frac{\sigma_n^G}{b} + M_n^G.
\]
Under (A1) and (C2), for all $k \in \mathbb{N}$,
\[
\mathbb{E} \left[ \|d_n^G\|^2_{H_n^G} \right] \leq \frac{1}{(1 - \gamma^G)^2 h_{0,n}^G} \left( \frac{\sigma_n^G}{b} + M_n^G \right),
\] where $h_{0,n}^G := \min_{i \in [\theta]} h_{0,i}^G$.

**Proof.** Let $n \in \mathbb{N}$. From (S2)(i), we have
\[
\mathbb{E} \left[ \|\nabla L_{G,S_n}(\theta_n)\|^2 | \theta_n \right] = \mathbb{E} \left[ \|\nabla L_{G,S_n}(\theta_n) - \nabla L_{G}(\theta_n, w_n) + \nabla L_{G}(\theta_n, w_n)\|^2 | \theta_n \right]
= \mathbb{E} \left[ \|\nabla L_{G}(\theta_n, w_n)\|^2 | \theta_n \right] + \mathbb{E} \left[ \|\nabla L_{G}(\theta_n, w_n)\|^2 | \theta_n \right]
+ 2\mathbb{E} \left[ \langle \nabla L_{G}(\theta_n, w_n), \nabla L_{G}(\theta_n, w_n) \rangle | \theta_n \right]
= \mathbb{E} \left[ \|\nabla L_{G}(\theta_n, w_n)\|^2 | \theta_n \right] + \mathbb{E} \left[ \|\nabla L_{G}(\theta_n, w_n)\|^2 | \theta_n \right],
\]
which, together with (S2)(ii) and (C2), implies that

$$\mathbb{E} \left[ \| \nabla L_{G,S_\epsilon}(\theta_n) \|^2 \right] \leq \frac{\sigma^2_G}{b} + M^2_G. \quad (10)$$

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

$$\mathbb{E} \left[ \| m_n^G \|^2 \right] \leq \beta_1 G \mathbb{E} \left[ \| m_{n-1}^G \|^2 \right] + (1 - \beta_1 G) \mathbb{E} \left[ \| \nabla L_{G,S_\epsilon}(\theta_n) \|^2 \right]$$

$$\leq \beta_1 G \mathbb{E} \left[ \| m_{n-1}^G \|^2 \right] + (1 - \beta_1 G) \left( \frac{\sigma^2_G}{b} + M^2_G \right).$$

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

$$\mathbb{E} \left[ \| m_n^G \|^2 \right] \leq \max \left\{ \| m_{-1}^G \|^2, \frac{\sigma^2_G}{b} + M^2_G \right\} = \frac{\sigma^2_G}{b} + M^2_G, \quad (11)$$

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

$$\mathbb{E} \left[ \| d_n^G \|^2_{H^G_n} \right] = \mathbb{E} \left[ \| \overline{H}^G_n^{-1} H^G_n d_n^G \|^2 \right] \leq \frac{1}{\tilde{\gamma}^G_n} \mathbb{E} \left[ \left\| \overline{H}^G_n^{-1} \right\| \| m_n^G \|^2 \right]$$

$$\leq \frac{1}{(1 - \gamma^G)^2} \mathbb{E} \left[ \left\| \overline{H}^G_n^{-1} \right\| \| m_n^G \|^2 \right],$$

where

$$\left\| \overline{H}^G_n^{-1} \right\| = \| \text{diag} \left( h_{n,i}^{-\frac{1}{2}} \right) \| = \max_{i \in [b]} h_{n,i}^{-\frac{1}{2}}$$

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

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

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

$$\mathbb{E} \left[ \| d_n^G \|^2_{H^G_n} \right] \leq \frac{1}{(1 - \gamma^G)^2 h_{0,*}^G} \left( \frac{\sigma^2_G}{b} + M^2_G \right),$$

completing the proof. \(\square\)

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

**Lemma A.4.** Algorithm 0 satisfies that, under (S2)(i), (ii) and (C2), for all $n \in \mathbb{N}$,

$$\mathbb{E} \left[ \| m_n^D \|^2 \right] \leq \frac{\sigma^2_D}{b} + M^2_D. \quad \text{Under (A1) and (C2), for all } k \in \mathbb{N},$$

$$\mathbb{E} \left[ \| d_n^D \|^2_{H^D_n} \right] \leq \frac{1}{(1 - \gamma^D)^2 h_{0,*}^D} \left( \frac{\sigma^2_D}{b} + M^2_D \right),$$

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

Lemmas A.4 and A.3 lead to the following:
Lemma A.5. Suppose that (S1)–(S3), (A1), and (C2)–(C3) hold and define \( X_n^G(\theta) := \mathbb{E} [\| \theta_n - \theta \|^2_{H_n^G}] \) for all \( n \in \mathbb{N} \) and all \( \theta \in \mathbb{R}^p \). Then, for all \( \theta \in \mathbb{R}^p \) and all \( n \in \mathbb{N} \),

\[
X_{n+1}^G(\theta) \leq X_n^G(\theta) + \text{Dist}(\theta) \mathbb{E} \left[ \sum_{i \in [\theta]} (h_{n+1,i}^G - h_{n,i}^G) \right] + \frac{\alpha_n^G (1 - \gamma^G)^2 h_{0,*}^G}{(1 - \gamma^G)^2 h_{0,*}^G} \left( \frac{\sigma_n^G}{b} + M_n^G \right) \\
+ 2\alpha_n^G \left\{ \frac{\beta_n^G}{\gamma_n^G} \sqrt{\text{Dist}(\theta)} \left( \frac{\sigma_n^G}{b} + M_n^G \right) + \frac{\tilde{\beta}_n^G}{\gamma_n^G} \mathbb{E} [\langle \theta - \theta_n, \nabla_{\theta} L_G(\theta_n, w_n) \rangle] \right\}.
\]

Proof. Lemma A.2 and Jensen’s inequality guarantee that

\[
\mathbb{E} [\| m_{n-1}^G \|] \leq \sqrt{\frac{\sigma_n^G}{b} + M_n^G}.
\]

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

\[
\| \theta_n - \theta \|^2 = \sum_{i \in [\theta]} (\theta_{n,i} - \theta_i)^2 \leq \text{Dist}(\theta).
\]

The Cauchy-Schwarz inequality thus ensures that

\[
\mathbb{E} [\| \theta - \theta_n, m_{n-1}^G \|] \leq \mathbb{E} [\| \theta - \theta_n \| m_{n-1}^G] \leq \sqrt{\text{Dist}(\theta)} \left( \frac{\sigma_n^G}{b} + M_n^G \right).
\]

We define \( X_n^G(\theta) := \mathbb{E} [\| \theta_n - \theta \|^2_{H_n^G}] \) for all \( n \in \mathbb{N} \) and all \( \theta \in \mathbb{R}^p \). Then, we have

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

which, together with (C3), implies that

\[
X_{n+1}^G(\theta) - \mathbb{E} [\| \theta_{n+1} - \theta \|^2_{H_n^G}] \leq \text{Dist}(\theta) \mathbb{E} \left[ \sum_{i \in [\theta]} (h_{n+1,i}^G - h_{n,i}^G) \right].
\]

Hence, Lemmas A.2 and A.3 lead to the assertion in Lemma A.5. \( \square \)

A discussion similar to the one proving Lemma A.3, 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 \( X_n^D(w) := \mathbb{E} [\| w_n - w \|^2_{H_n^D}] \) for all \( n \in \mathbb{N} \) and all \( w \in \mathbb{R}^w \). Then, for all \( w \in \mathbb{R}^w \) and all \( n \in \mathbb{N} \),

\[
X_{n+1}^D(w) \leq X_n^D(w) + \text{Dist}(w) \mathbb{E} \left[ \sum_{i \in [w]} (h_{n+1,i}^D - h_{n,i}^D) \right] + \frac{\alpha_n^D (1 - \gamma^D)^2 h_{0,*}^D}{(1 - \gamma^D)^2 h_{0,*}^D} \left( \frac{\sigma_n^D}{b} + M_n^D \right) \\
+ 2\alpha_n^D \left\{ \frac{\beta_n^D}{\gamma_n^D} \sqrt{W \text{Dist}(w)} \left( \frac{\sigma_n^D}{b} + M_n^D \right) + \frac{\tilde{\beta}_n^D}{\gamma_n^D} \mathbb{E} [\langle w - w_n, \nabla_w L_D(\theta_n, w_n) \rangle] \right\}.
\]

A.6. Proof of Theorem 3.1(i)

The following is a convergence analysis of Algorithm II.
Let us show that, for all $\theta$, Theorem A.1.

Furthermore, there exist accumulation points $(\theta^*, w^*)$ for all $w := 1 - \beta_1^D \gamma^G : = 1 - \gamma^G, \gamma^D : = 1 - \gamma^D, h_{\theta,0}^G := \min_{i \in [\theta]} h_{\theta,0}^G, and h_{\theta,0}^D := \min_{j \in [W]} h_{\theta,0}^D$. Furthermore, there exist accumulation points $(\theta^*, w^*)$ and $(\theta_*, w_*)$ of $((\theta_n, w_n))_{n \in \mathbb{N}}$ such that

$$
E \left[ \| \nabla_{\theta} L_G(\theta^*, w^*) \|^2 \right] \leq \frac{\alpha^G (\sigma_b^2 b^{-1} + M_0^2)}{2 \beta^G \gamma^G h_{\theta,0}^G} + \frac{\sqrt{\ThetaDist(\theta) \left( \frac{\sigma_b^2 b^{-1} + M_0^2}{\beta^G} \right)}}{\beta^G_1},
$$

$$
E \left[ \| \nabla_{\theta} L_D(\theta_*, w_*) \|^2 \right] \leq \frac{\alpha^D (\sigma^2 b^{-1} + M_0^2)}{2 \beta^D \gamma^D h_{\theta,0}^D} + \frac{\sqrt{WDist(w) \left( \frac{\sigma^2 b^{-1} + M_0^2}{\beta^D} \right)}}{\beta^D_1},
$$

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

$$
X_{n+1}^G(\theta) \leq X_n^G(\theta) + \text{Dist} (\theta) \left[ \sum_{i \in [\theta]} (h_{n+1,i} - h_n^G) \right] + \frac{\alpha^G (\sigma_b^2 b^{-1} + M_0^2)}{1 - \gamma^G h_{\theta,0}^G} + \frac{\alpha^G (\sigma^2 b^{-1} + M_0^2)}{1 - \gamma^D h_{\theta,0}^D} + 2 \alpha G \left\{ \frac{\gamma^G}{\gamma} \sqrt{\ThetaDist(\theta) \left( \frac{\sigma^2 b^{-1} + M_0^2}{\beta^D} \right)} + \frac{\beta^D_1}{\gamma^D} E \left[ \left( \theta - \theta_n, \nabla_{\theta} L_G(\theta_n, w_n) \right) \right] \right\}.
$$

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

$$
X_{n+1}^G(\theta) \leq X_n^G(\theta) + \gamma^G_{n+1} X_{n+1}^G(\theta) + \text{Dist} (\theta) \left[ \sum_{i \in [\theta]} (h_{n+1,i} - h_n^G) \right] + \frac{\alpha^G (\sigma_b^2 b^{-1} + M_0^2)}{1 - \gamma^G h_{\theta,0}^G} + \frac{\alpha^G (\sigma^2 b^{-1} + M_0^2)}{1 - \gamma^D h_{\theta,0}^D} + 2 \alpha G \left\{ \beta_1^G \sqrt{\ThetaDist(\theta) \left( \frac{\sigma^2 b^{-1} + M_0^2}{\beta^D} \right)} + \frac{\beta^D_1}{\gamma^D} E \left[ \left( \theta - \theta_n, \nabla_{\theta} L_G(\theta_n, w_n) \right) \right] \right\}.
$$

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

$$
\liminf_{n \rightarrow +\infty} E \left[ \| \nabla_{\theta} L_G(\theta_n, w_n) \| \right] \leq \frac{\alpha^G}{2 \beta^G_1 (1 - \gamma^G h_{\theta,0}^G)} \left( \frac{\sigma^2 b^{-1} + M_0^2}{\beta^D} \right) + \sqrt{\ThetaDist(\theta) \left( \frac{\sigma^2 b^{-1} + M_0^2}{\beta^D} \right)} + \epsilon.
$$

If (13) does not hold, then there exist $\tilde{\theta} \in \mathbb{R}^\theta$ and $\epsilon_0 > 0$ such that

$$
\liminf_{n \rightarrow +\infty} E \left[ \| \nabla_{\theta} L_G(\theta_n, w_n) \| \right] > \frac{\alpha^G}{2 \beta^G_1 (1 - \gamma^G h_{\theta,0}^G)} \left( \frac{\sigma^2 b^{-1} + M_0^2}{\beta^D} \right) + \sqrt{\ThetaDist(\theta) \left( \frac{\sigma^2 b^{-1} + M_0^2}{\beta^D} \right)} + \epsilon_0.
$$

Existence and Estimation of Critical Batch Size for Training GANs with TTUR
Then, there exists \( n_0 \in \mathbb{N} \) such that, for all \( n \geq n_0 \),
\[
\mathbb{E} \left[ (\theta_n - \hat{\theta}, \nabla_\theta L_G(\theta_n, w_n)) \right] \\
> \frac{\alpha^G}{2 \beta_1^G (1 - \gamma^G)^2 h_{0,*}^G} \left( \frac{\sigma_0^2}{b} + M_G^2 \right) + \sqrt{\Theta \text{Dist}(\theta) \left( \frac{\sigma_2^2}{b} + M_{\theta}^2 \right) \beta_1^G} + \epsilon_0 \frac{\beta_1^G}{2}.
\]

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 \( n \geq n_1 \),
\[
\gamma^{n+1} X^G_{n+1}(\hat{\theta}) + \text{Dist}(\hat{\theta}) \mathbb{E} \left[ \sum_{i \in [\theta]} \left( h_{n+1,i} - h_{n,i}^G \right) \right] \leq \frac{\alpha^G \tilde{\beta}^G \epsilon_0}{2}.
\]

Accordingly, from (14), for all \( n \geq n_2 := \max\{n_0, n_1\} \),
\[
X^G_{n+1}(\hat{\theta}) < X^G_n(\hat{\theta}) + \frac{\alpha^G \tilde{\beta}^G \epsilon_0}{2} + \frac{\alpha^G}{(1 - \gamma^G)^2 h_{0,*}^G} \left( \frac{\sigma_0^2}{b} + M_G^2 \right)
+ 2 \alpha^G \beta_1^G \sqrt{\Theta \text{Dist}(\theta) \left( \frac{\sigma_2^2}{b} + M_{\theta}^2 \right)} - 2 \alpha^G \tilde{\beta}_1^G \left\{ \frac{\alpha^G}{2 \beta_1^G (1 - \gamma^G)^2 h_{0,*}^G} \left( \frac{\sigma_0^2}{b} + M_G^2 \right) \right. \\
+ \left. \sqrt{\Theta \text{Dist}(\theta) \left( \frac{\sigma_2^2}{b} + M_{\theta}^2 \right) \beta_1^G} + \epsilon_0 \frac{\beta_1^G}{2} \right\}
= X^G_n(\hat{\theta}) - \frac{\alpha^G \tilde{\beta}^G \epsilon_0}{2}
< X^G_n(\hat{\theta}) - \frac{\alpha^G \tilde{\beta}^G \epsilon_0}{2} (n + 1 - n_2).
\]

Note that the right-hand side of the above inequality approaches minus infinity as \( n \) approaches positive infinity, producing a contradiction. Therefore, (15) holds, which implies that, for all \( \theta \in \mathbb{R}^\Theta \),
\[
\liminf_{n \to +\infty} \mathbb{E} \left[ (\theta_n - \theta, \nabla_\theta L_G(\theta_n, w_n)) \right] \\
\leq \frac{\alpha^G}{2 \beta_1^G (1 - \gamma^G)^2 h_{0,*}^G} \left( \frac{\sigma_0^2}{b} + M_G^2 \right) + \sqrt{\Theta \text{Dist}(\theta) \left( \frac{\sigma_2^2}{b} + M_{\theta}^2 \right) \beta_1^G}.
\]

A discussion similar to the one showing (15), together with Lemma A.6, leads to the finding that, for all \( w \in \mathbb{R}^W \),
\[
\liminf_{n \to +\infty} \mathbb{E} \left[ (w_n - w, \nabla_w L_D(\theta_n, w_n)) \right] \\
\leq \frac{\alpha^D}{2 \beta_1^D (1 - \gamma^D)^2 h_{0,*}^D} \left( \frac{\sigma_0^2}{b} + M_D^2 \right) + \sqrt{W \text{Dist}(w) \left( \frac{\sigma_2^2}{b} + M_{\theta}^2 \right) \beta_1^D}.
\]

Let \( \theta \in \mathbb{R}^\Theta \). From (15), there exists a subsequence \(( (\theta_n, w_n) ) \) of \(( (\theta_n, w_n) ) \) such that
\[
\lim_{i \to +\infty} \mathbb{E} \left[ (\theta_n - \theta, \nabla_\theta L_G(\theta_n, w_n)) \right] \\
\leq \frac{\alpha^G}{2 \beta_1^G (1 - \gamma^G)^2 h_{0,*}^G} \left( \frac{\sigma_0^2}{b} + M_G^2 \right) + \sqrt{\Theta \text{Dist}(\theta) \left( \frac{\sigma_2^2}{b} + M_{\theta}^2 \right) \beta_1^G}.
\]

Conditions (S1) and (C3) guarantee that there exists \(( (\theta_n, w_n) ) \) of \(( (\theta_n, w_n) ) \) such that \(( (\theta_n, w_n) ) \) converges almost surely to \(( \theta^*, w^* ) \) in \( \mathbb{R}^\Theta \times \mathbb{R}^W \). Therefore, for all \( \theta \in \mathbb{R}^\Theta \),
\[
\mathbb{E} \left[ (\theta - \theta, \nabla_\theta L_G(\theta^*, w^*)) \right] \\
\leq \frac{\alpha^G}{2 \beta_1^G (1 - \gamma^G)^2 h_{0,*}^G} \left( \frac{\sigma_0^2}{b} + M_G^2 \right) + \sqrt{\Theta \text{Dist}(\theta) \left( \frac{\sigma_2^2}{b} + M_{\theta}^2 \right) \beta_1^G}.
\]
Hence, letting \( \theta = \tilde{\theta} := \theta^* - \nabla_\theta L_G(\theta^*, w^*) \) implies that
\[
\mathbb{E} \left[ \| \nabla_\theta L_G(\theta^*, w^*) \|^2 \right] 
\leq \frac{\alpha^G}{2\beta^G_1} \left( \frac{\sigma^2_G}{b} + M^2_G \right) + \sqrt{\Theta \text{Dist}(\tilde{\theta}) \left( \frac{\sigma^2_G}{b} + M^2_G \right)} \beta^G_1, \tag{17}
\]

A discussion similar to the one showing (17), together with (16), implies that there exists \((\theta_*, w_*) \in \mathbb{R}^\Theta \times \mathbb{R}^W\) such that
\[
\mathbb{E} \left[ \| \nabla_w L_D(\theta_*, w_*) \|^2 \right] 
\leq \frac{\alpha^D}{2\beta^D_1} \left( \frac{\sigma^2_D}{b} + M^2_D \right) + \sqrt{W \text{Dist}(\tilde{w}) \left( \frac{\sigma^2_D}{b} + M^2_D \right)} \beta^D_1,
\]

where \( \tilde{w} = w_* - \nabla_w L_D(\theta_*, w_*) \). This completes the proof. \(\square\)

A.7. Proof of Theorem 5.1(ii)

Lemmas 5.1 and 5.2 lead to the following lemma:

**Lemma A.7.** Suppose that \((S1)-(S3), (A1)-(A2), and (C2)-(C3) hold and consider Algorithm 11, where \((\alpha^G_n)_{n \in \mathbb{N}}\) is monotone decreasing. Then, for all \( \theta \in \mathbb{R}^\Theta \) and all \( N \geq 1 \),
\[
\frac{1}{N} \sum_{n \in [N]} \mathbb{E} \left[ (\theta_n - \theta, \nabla_\theta L_G(\theta_n, w_n)) \right]
\leq \frac{\Theta \text{Dist}(\theta)}{2\alpha^G_1 \beta^G_1 N} + \frac{1}{N} \sum_{n \in [N]} \frac{\alpha^G_n}{2\beta^G_1 (1 - \gamma^G)^2 \beta^G_1} \left( \frac{\sigma^2_G}{b} + M^2_G \right) + \frac{\beta^G_1}{\beta^G_1} \sqrt{\Theta \text{Dist}(\theta) \left( \frac{\sigma^2_G}{b} + M^2_G \right)},
\]

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

**Proof.** Lemma 5.1 implies that, for all \( \theta \in \mathbb{R}^\Theta \) and all \( n \in \mathbb{N} \),
\[
\mathbb{E} \left[ (\theta_n - \theta, \nabla_\theta L_G(\theta_n, w_n)) \right]
= \frac{\alpha^G_n}{2\alpha^G_1 \beta^G_1} \left\{ \mathbb{E} \left[ \| \theta_n - \theta \|_{H^G}^2 \right] - \mathbb{E} \left[ \| \theta_{n+1} - \theta \|_{H^G}^2 \right] \right\}
+ \frac{\alpha^G_n}{2\beta^G} \mathbb{E} \left[ \| d^G_n \|_{H^G}^2 \right] + \frac{\beta^G_1}{\beta^G_1} \mathbb{E} \left[ (\theta_n, m^G_{n-1}) \right],
\]

which implies that, for all \( \theta \in \mathbb{R}^\Theta \) and all \( N \geq 1 \),
\[
\frac{1}{N} \sum_{n \in [N]} \mathbb{E} \left[ (\theta_n - \theta, \nabla_\theta L_G(\theta_n, w_n)) \right]
= \frac{\alpha^G_n}{2\alpha^G_1 \beta^G_1} \left\{ \mathbb{E} \left[ \| \theta_n - \theta \|_{H^G}^2 \right] - \mathbb{E} \left[ \| \theta_{n+1} - \theta \|_{H^G}^2 \right] \right\}
+ \frac{1}{N} \sum_{n \in [N]} \frac{\alpha^G_n}{2\beta^G} \mathbb{E} \left[ \| d^G_n \|_{H^G}^2 \right]
+ \frac{\beta^G_1}{\beta^G_1} \mathbb{E} \left[ (\theta_n, m^G_{n-1}) \right].
\]

Let \( \delta_n^G := \gamma_n^G/(2\alpha^G_n \beta^G_1) \) for all \( n \in \mathbb{N} \). Then, we have that
\[
\Theta_N := \delta^G \mathbb{E} \left[ \| \theta_1 - \theta \|_{H^G}^2 \right] + \sum_{n=2}^{N} \left\{ \delta_n^G \mathbb{E} \left[ \| \theta_n - \theta \|_{H^G}^2 \right] - \delta_{n-1}^G \mathbb{E} \left[ \| \theta_{n-1} - \theta \|_{H^G}^2 \right] \right\}
- \delta_N^G \mathbb{E} \left[ \| \theta_N - \theta \|_{H^G}^2 \right].
\]
Hence, for all $N \geq 2$,

$$\hat{\Theta}_N = \mathbb{E} \left[ \sum_{n=2}^{N} \left( \delta_n^G \left\| H_n^G (\theta_n - \theta) \right\|^2 - \delta_{n-1}^G \left\| H_{n-1}^G (\theta_n - \theta) \right\|^2 \right) \right].$$

Since $(\sigma_n^G)_{n \in \mathbb{N}}$ is monotone decreasing, we have that $\sigma_{n+1}^G \geq \sigma_n^G$ ($n \in \mathbb{N}$). 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), $\max_{i \in [\Theta]} \sup_{n \in \mathbb{N}} (\theta_n - \theta_i)^2 \leq \text{Dist}(\theta)$. Accordingly, for all $N \geq 2$,

$$\hat{\Theta}_N \leq \text{Dist}(\theta) \mathbb{E} \left[ \sum_{n=2}^{N} \sum_{i=1}^{\Theta} (\delta_n^G h_{n,i}^G - \delta_{n-1}^G h_{n-1,i}^G) (\theta_n - \theta_i)^2 \right].$$

Therefore, $\delta_1^G \mathbb{E}[\| \theta_1 - \theta \|^2_{H_1^G}] \leq \text{Dist}(\theta) \delta_1^G \mathbb{E} \left[ \sum_{i=1}^{\Theta} h_{1,i}^G \right]$, and (A2) imply that, for all $N \geq 1$,

$$\Theta_N \leq \delta_1^G \text{Dist}(\theta) \mathbb{E} \left[ \sum_{i=1}^{\Theta} h_{1,i}^G \right] + \text{Dist}(\theta) \mathbb{E} \left[ \sum_{i=1}^{\Theta} (\delta_n^G h_{n,i}^G - \delta_1^G h_{1,i}^G) \right]$$

$$\leq \delta_1^G \text{Dist}(\theta) \sum_{i=1}^{\Theta} H_i^G$$

$$\leq \delta_1^G \Theta \text{Dist}(\theta) H_i^G,$$

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

$$\Theta_N \leq \frac{\Theta \text{Dist}(\theta) H_i^G}{2 \alpha_n^G \beta_1^G}. \tag{19}$$

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

$$A_N := \sum_{n \in [N]} \frac{\alpha_n^G \gamma_n^G}{2 \beta_1^G} \mathbb{E} \left[ \left\| d_n^G \right\|_{H_n^G}^2 \right] \leq \sum_{n \in [N]} \frac{\alpha_n^G}{2 \beta_1^G (1 - \gamma_n^G)^2 h_{n,i}^G} \left( \frac{\sigma_n^G}{b} + M_i^2 \right). \tag{20}$$

From (12), we have

$$B_N := \sum_{n \in [N]} \frac{\beta_1^G}{\beta_1^G} \mathbb{E} \left[ (\theta - \theta_n, m_{n-1}^G) \right] \leq \frac{\beta_1^G N}{\beta_1^G} \sqrt{\text{Dist}(\theta) \left( \frac{\sigma_1^G}{b} + M_i^2 \right)}. \tag{21}$$

Therefore, (13), (20), and (21) lead to the assertion in Lemma 3. This completes the proof. □

A discussion similar to the one proving Lemma 3, together with Lemmas 2 and 4, leads to the following lemma:
Lemma A.8. Suppose that (S1)–(S3), (A1)–(A2), and (C2)–(C3) hold and consider Algorithm 3.1 where \((\alpha^D_n)_{n \in \mathbb{N}}\) is monotone decreasing. Then, for all \(w \in \mathbb{R}^W\) and all \(N \geq 1\),

\[
\frac{1}{N} \sum_{n \in [N]} \mathbb{E} \left[ \langle w_n - w, \nabla_w L_D(\theta_n, w_n) \rangle \right] \leq \frac{WDist(w)H^D}{2\alpha^D_N \beta^D_1 N} + \frac{1}{N} \sum_{n \in [N]} \frac{\alpha^D_n}{2\beta^D_1 (1 - \gamma^D)^2 h_0^D,} \left( \frac{\sigma^2_D}{b} + M_D^2 \right) + \frac{\beta^D_1}{\beta^D_1} \sqrt{WDist(w) \left( \frac{\sigma^2_D}{b} + M_D^2 \right)},
\]

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

Proof of Theorem 3.2(ii). Let \(\alpha^G_n := \alpha^G\) and \(\alpha^D_n := \alpha^D\) for all \(n \in \mathbb{N}\). Lemmas A.7 and A.8 thus guarantee that, for all \(\theta \in \mathbb{R}^B\) and all \(N \geq 1\),

\[
\frac{1}{N} \sum_{n \in [N]} \mathbb{E} \left[ \langle \theta_n - \theta, \nabla_{\theta} L_G(\theta_n, w_n) \rangle \right] \leq \frac{\Theta Dist(\theta)H^G}{2\alpha^G \beta^G N} + \frac{\alpha^G}{2\beta^G (1 - \gamma^G)^2 h_0^G,} \left( \frac{\sigma^2_G}{b} + M_G^2 \right) + \frac{\beta^G_1}{\beta^G_1} \sqrt{\Theta Dist(\theta) \left( \frac{\sigma^2_G}{b} + M_G^2 \right)},
\]

and

\[
\frac{1}{N} \sum_{n \in [N]} \mathbb{E} \left[ \langle w_n - w, \nabla_w L_D(\theta_n, w_n) \rangle \right] \leq \frac{WDist(w)H^D}{2\alpha^D \beta^D_1 N} + \frac{\alpha^D}{2\beta^D_1 (1 - \gamma^D)^2 h_0^D,} \left( \frac{\sigma^2_D}{b} + M_D^2 \right) + \frac{\beta^D_1}{\beta^D_1} \sqrt{WDist(w) \left( \frac{\sigma^2_D}{b} + M_D^2 \right)},
\]

which completes the proof.

A.8. Proof of Theorem 3.5.

Proof of Theorem 3.5. Condition 3.5(i) implies that TTUR achieves an \(\epsilon\)-approximation. We have that, for \(b > B_G/(\epsilon^2_G - C_G)\),

\[
\frac{dN_G(b)}{db} = \frac{-A_G B_G}{\{(\epsilon^2_G - C_G)b - B_G\}^2} \leq 0,
\]

\[
\frac{d^2 N_G(b)}{db^2} = \frac{2A_G B_G(\epsilon^2_G - C_G)}{\{(\epsilon^2_G - C_G)b - B_G\}^3} \geq 0.
\]

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


Proof of Theorem 3.6. We have that, for \(b > B_G/(\epsilon^2_G - C_G)\),

\[
N_G(b) = \frac{A_G b^2}{(\epsilon^2_G - C_G)b - B_G}.
\]

Hence,

\[
\frac{dN_G(b)}{db} = \frac{A_G b((\epsilon^2_G - C_G)b - 2B_G)}{\{(\epsilon^2_G - C_G)b - B_G\}^2},
\]

\[
\frac{d^2 N_G(b)}{db^2} = \frac{2A_G B_G^2}{\{(\epsilon^2_G - C_G)b - B_G\}^3} \geq 0,
\]

21
which implies that \( N_G(b)b \) is convex for \( b > B_G/(\epsilon_G^2 - C_G) \) and

\[
\frac{dN_G(b)b}{db} \begin{cases} < 0 & \text{if } b < b_G^*; \\ = 0 & \text{if } b = b_G^* = \frac{2B_G}{\epsilon_G^2 - C_G}; \\ > 0 & \text{if } b > b_G^*. \end{cases}
\]

The point \( b_G^* \) attains the minimum value \( N_G(b_G^*)b_G^* \) 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 \( B_G, C_G, \beta_1^G, \) and \( \gamma^G \) ensure that

\[
b_G^* = \frac{2B_G}{\epsilon_G^2 - C_G} > \frac{2B_G}{\epsilon_G^2}, \quad \frac{\sigma_G^2\alpha_G^G}{2\beta_1^G\gamma^G\bar{h}_{0,s}} = \frac{\sigma_G^2\alpha_G^G}{\epsilon_G^2(1 - \beta_1^G)(1 - \gamma^G)^2h_{0,s}^G}.
\]

Moreover, (3.4) and the definition of \( L_G(\theta_n, w_n) \) ensure that

\[
\|\nabla L_G, s_n(\theta_n)\|^2 \leq \frac{\sigma_G^2}{b} + \frac{\|\nabla L_G(\theta_n, w_n)\|^2}{2} = \left( -\frac{i}{|S|^2}\right) \sum_{i=1}^{|S|} \left| \frac{\partial L_G^{\Theta}(\theta_n, w_n)}{\partial \theta_n}\right|^2 \leq \frac{\Theta}{|S|^2} \sum_{i=1}^{|S|} \max_{i \in [\Theta]} G_{n,i}^2.
\]

**Proof of Proposition 3.4(i).** The definition of \( v_{n,i}^G \) implies that

\[
v_{n,i}^G = \beta_2^G v_{n-1,i}^G + (1 - \beta_2^G)g_{n,i}^2 \leq \max_{n,i} g_{n,i}^2 =: g_{n,i}^2, \quad \sum_{i=1}^{|\Theta|} g_{n,i}^2 = \|\nabla L_G, s_n^G(\theta_n^s)\|^2,
\]

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 \( v_{n,i}^G \) and \( \beta_2 \in [0, 1) \), we have

\[
b_{0,s}^G := \min_{i \in [\Theta]} b_{0,s}^G \leq \sqrt{v_{n,i}^G} = \sqrt{\frac{v_{n,i}^G}{1 - \beta_2^G}} \leq \sqrt{\frac{v_{n,i}^G}{1 - \beta_2^G}}.
\]

Hence,

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

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

\[
b_G^* \geq \frac{\sigma_G^2\alpha_G^G}{\epsilon_G^2(1 - \beta_1^G)^3 \sqrt{\frac{\Theta}{|S|^2}}}
\]

The discussion for \( b_D^* \) is similar to the one for \( b_G^* \). This completes the proof.
existence and estimation of critical batch size for training GANs with TTUR

Proof of Proposition 5(ii). The definition of $s^G_n$ implies that
\[
s^G_{n,i} = \beta^G v^G_n(i) + (1 - \beta^G) (g_n(i) - m_n(i))^2 \leq \max_{n,i} (g_n(i) - m_n(i))^2 =: (g^*,i - m^*,i)^2
\]
\[
\leq \sum_{i=1}^\Theta (g^*,i - m^*,i)^2 = \|\nabla L_{G,s^*} (\theta^*) - m^*\|^2 ,
\]
where the first inequality can be shown by induction. Hence, we have
\[
\|\nabla L_{G,s^*} (\theta^*) - m^*\|^2 \leq 2 \|\nabla L_{G,s^*} (\theta^*)\|^2 + 2 \|m^*\|^2 \leq \frac{4 \sigma^G}{b} + 4 \|\nabla L_G (\theta^*, w^*)\|^2
\]
Therefore, for all $n \in \mathbb{N}$ and all $i \in \{\Theta\}$,
\[
s^G_{n,i} \leq \frac{4 \Theta}{|S|^2} \max_{i \in \Theta} G^2_{n^*,i}.
\]
From the definition of $s^G_{n,i}$ and $\beta^G \in [0, 1)$, we have
\[
h^G_{0,*} := \min_{i \in \Theta} h^G_{0,i} \leq \sqrt{s^G_{n,i}} = \sqrt{\frac{s^G_{n,i}}{1 - \beta^G}} \leq \sqrt{\frac{s^G_{n,i}}{1 - \beta^G}}.
\]
Hence,
\[
b^*_G \geq \frac{\sigma^G \alpha^G}{\epsilon^G (1 - \beta^G)(1 - \gamma^G)^2} \frac{\Theta}{|S|^2} \max_{i \in \Theta} G^2_{n^*,i}.
\]
Accordingly, using $\gamma^G = \beta^G_1$ and $\max_{i \in \Theta} G^2_{n^*,i} \approx \epsilon^G_2$ implies that
\[
b^*_G \geq \frac{\sigma^G \alpha^G}{\epsilon^G (1 - \beta^G)^3} \frac{\Theta}{|S|^2} \max_{i \in \Theta} G^2_{n^*,i}.
\]
The discussion for $b^*_D$ is similar to the one for $b^*_G$. This completes the proof.

Proof of Proposition 5(iii). From the definition of $v^G_n$ and a discussion similar to the proof of Proposition 3.4(i), we have
\[
v^G_{n,i} \leq \frac{\Theta}{|S|^2} \max_{i \in \Theta} G^2_{n^*,i}.
\]
From the definition of $v^G_{n,i}$, we have
\[
h^G_{0,*} := \min_{i \in \Theta} h^G_{0,i} \leq \sqrt{v^G_{n,i}}.
\]
Hence,
\[
b^*_G \geq \frac{\sigma^G \alpha^G}{\epsilon^G (1 - \beta^G)(1 - \gamma^G)^2} \frac{\Theta}{|S|^2} \max_{i \in \Theta} G^2_{n^*,i}.
\]
Accordingly, using $\gamma^G = \beta^G_1 = 0$ and $\max_{i \in \Theta} G^2_{n^*,i} \approx \epsilon^G_2$ implies that
\[
b^*_G \geq \frac{\sigma^G \alpha^G}{\epsilon^G_2 \sqrt{|S|^2}}.
\]
The discussion for $b^*_D$ is similar to the one for $b^*_G$. 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 \( x^* \) is a stationary point of \( f \). Then, \( \nabla f(x^*) = 0 \) is equivalent to the following variational inequality: for all \( x \in \mathbb{R}^d \),

\[
\langle \nabla f(x^*), x - x^* \rangle \geq 0.
\]

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

\[
\langle \nabla f(x), y - x \rangle \geq 0.
\]

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

\[
0 \leq \langle \nabla f(x), y - x \rangle = -\|\nabla f(x)\|^2.
\]

Hence,

\[
\nabla f(x) = 0.
\]

A.12. Remarks regarding (C3)

We make the following remarks regarding (C3).

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

(C3)(ii) Let \( L_G^{(i)}(\cdot, w) : \mathbb{R}^\Theta \to \mathbb{R} \) be convex and \( \nabla_\theta L_G^{(i)}(\cdot, w) : \mathbb{R}^\Theta \to \mathbb{R}^{\Theta} \) be Lipschitz continuous with the Lipschitz constant \( l_G \). Let \( L_D^{(i)}(\theta, \cdot) : \mathbb{R}^W \to \mathbb{R} \) be convex and \( \nabla_w L_D^{(i)}(\theta, \cdot) : \mathbb{R}^W \to \mathbb{R}^W \) be Lipschitz continuous with the Lipschitz constant \( l_D \). Then, the sequences \((\theta_n)_{n \in \mathbb{N}}\) and \((w_n)_{n \in \mathbb{N}}\) generated by SGD guarantee that Algorithm II using (22) is bounded. Thus, the nonexpansivity conditions of \( P_G \) and \( P_D \) guarantee that Algorithm II using (22) satisfies (3) without assuming (C3).

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

\[
\phi_n(\theta) := L_{G,S_n}(\theta, w_n) + \langle \nabla L_{G,S_n}(\theta_n), \theta - \theta_n \rangle + \frac{1}{2\alpha^G} \| \theta - \theta_n \|^2.
\]

The convexity of \( L_{G,S_n}(\cdot, w_n) \) implies that, for all \( \theta \in \mathbb{R}^\Theta \), \( L_{G,S_n}(\theta, w_n) \geq L_{G,S_n}(\theta_n, w_n) + \langle \nabla L_{G,S_n}(\theta_n), \theta - \theta_n \rangle \). Hence, we have that, for all \( \theta \in \mathbb{R}^\Theta \),

\[
\phi_n(\theta) \leq L_{G,S_n}(\theta, w_n) + \frac{1}{2\alpha^G} \| \theta - \theta_n \|^2.
\]

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

\[
0 = \nabla_\theta \phi_n(\hat{\theta}_n) = \nabla L_{G,S_n}(\theta_n) + \frac{1}{\alpha^G}(\theta - \theta_n),
\]
The strong convexity of $\phi$, and $\nabla \phi_n(\theta_{n+1}) = 0$ imply that

$$\phi_n(\theta) \geq \phi_n(\theta_{n+1}) + \langle \nabla \phi_n(\theta_{n+1}), \theta - \theta_{n+1} \rangle + \frac{1}{2\alpha^G} \| \theta_{n+1} - \theta \|^2$$

$$= \phi_n(\theta_{n+1}) + \frac{1}{2\alpha^G} \| \theta_{n+1} - \theta \|^2.$$

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

$$\phi_n(\theta_{n+1}) := L_{G,S_n}(\theta_n, w_n) + \langle \nabla L_{G,S_n}(\theta_n), \theta_{n+1} - \theta_n \rangle + \frac{1}{2\alpha^G} \| \theta_{n+1} - \theta_n \|^2$$

$$\geq L_{G,S_n}(\theta_n, w_n) + \langle \nabla L_{G,S_n}(\theta_n), \theta_{n+1} - \theta_n \rangle + \frac{l_G}{2} \| \theta_{n+1} - \theta_n \|^2$$

$$\geq L_{G,S_n}(\theta_{n+1}, w_n).$$

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

$$L_{G,S_n}(\theta, w_n) + \frac{1}{2\alpha^G} \| \theta - \theta \|^2 \geq L_{G,S_n}(\theta_{n+1}, w_n) + \frac{1}{2\alpha^G} \| \theta_{n+1} - \theta \|^2,$$

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

$$L_G(\theta, w_n) + \frac{1}{2\alpha^G} \| \theta - \theta \|^2 \geq L_G(\theta_{n+1}, w_n) + \frac{1}{2\alpha^G} \| \theta_{n+1} - \theta \|^2.$$

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

$$\frac{1}{2\alpha^G} (\| \theta_{n+1} - \theta^* \|^2 - \| \theta_n - \theta^* \|^2) \leq L_G(\theta^*, w_n) - L_G(\theta_{n+1}, w_n) \leq 0.$$

Since $(\| \theta_n - \theta^* \|^2)_{n \in \mathbb{N}}$ is monotone decreasing, the sequence $(\| \theta_n - \theta^* \|^2)_{n \in \mathbb{N}}$ is bounded. The discriminator can be defined by replacing $G$, $S_n$, $\theta$, and $w$ in the generator with $D$, $R_n$, $\omega$, and $\theta$. 

Details of (C3(ii)): Let $\theta \in B^G \subset \mathbb{R}^G$ and $n \in \mathbb{N}$. The definition of $\theta_{n+1}$ and the nonexpansivity of $P_G$ (i.e., $\| P_G(\theta_1) - P_G(\theta_2) \|_H \leq \| \theta_1 - \theta_2 \|_H$), $\theta_1, \theta_2 \in \mathbb{R}^G, H \in S_+^G$) imply that

$$\| \theta_{n+1} - \theta \|_H^2 = \| P_G(\theta_n + \alpha^G_n d_n^G) - P_G(\theta) \|_H^2$$

$$\leq \| (\theta_n - \theta) + \alpha^G_n d_n^G \|_H^2 = \| \theta_n - \theta \|_H^2 + 2\alpha^G_n \langle \theta_n - \theta, d_n^G \rangle_H + \alpha^G_n \| d_n^G \|_H^2.$$

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

i.e.,

$$\tilde{\theta}_n = \theta_n - \alpha^G \nabla L_{G,S_n}(\theta_n) =: \theta_{n+1}.$$