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

Naoki Sato[1], Hideaki Iiduka[1]

[1]Meiji University

MEIJI UNIVERSITY

ICML International Conference On Machine Learning

Introduction

【Motivation】

- ▷ There is no convergence proof for TTUR with constant learning rates.
- ▷ For DNN training, there is a critical batch size that is optimal for training in terms of computational complexity. [Sha+19]

【Contribution】

- ▷ We provide convergence for TTUR with constant learning rates.
- ▷ We showed that there is a critical batch size for training GANs.
- ▷ We showed that the critical batch size can be estimated.

Problem Setting

This paper considers the following LNE problem with two players, a generator and a discriminator [Heu+17]:

Problem *Find a pair* $(\boldsymbol{\theta}^\star, \boldsymbol{w}^\star) \in \mathbb{R}^\Theta \times \mathbb{R}^W$ *satisfying*

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

L_G : The loss function of the Generator for a fixed $\boldsymbol{w} \in \mathbb{R}^W$

L_D : The loss function of the Discriminator for a fixed $\boldsymbol{\theta} \in \mathbb{R}^\Theta$

For the convergence analysis, we use the following variational inequality equivalent to the above equation.

$$\forall \boldsymbol{\theta} \in \mathbb{R}^\Theta, \forall \boldsymbol{w} \in \mathbb{R}^W :$$

$$\langle \boldsymbol{\theta}^\star - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} L_G(\boldsymbol{\theta}^\star, \boldsymbol{w}^\star) \rangle \leq 0,$$

$$\langle \boldsymbol{w}^\star - \boldsymbol{w}, \nabla_{\boldsymbol{w}} L_D(\boldsymbol{\theta}^\star, \boldsymbol{w}^\star) \rangle \leq 0.$$

Theoretical Analysis

【Convergence Analysis】(Theorem 3.1(ii) / Proof is in Appendix A.7)

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

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

【Relationship between b and N】

(Theorem 3.2 / Proof is in Appendix A.8)

Suppose we can set parameters such as the ideal number of steps N, batch size b, and learning rate α^G, α^D that can approximate the local Nash equilibrium. Let $\epsilon^G, \epsilon^D > 0$ be sufficiently small positive numbers such that,

$$\frac{A_G}{N_G} + \frac{B_G}{b} + C_G = \epsilon_G^2, \quad \frac{A_D}{N_D} + \frac{B_D}{b} + C_D = \epsilon_D^2$$

i.e.

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

We see that $N_G(b)$ and $N_D(b)$ are monotone decreasing and convex with respect to b.

【Existence of a Critical Batch Size】

(Theorem 3.3 / Proof is in Appendix A.9)

Since b stochastic gradients are computed in one iteration, SFO can be defined by $N(b)b$.

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

We see that $N_G(b)b$ and $N_D(b)b$ are convex functions. Also, there exists a $b^\star$ minimizing $N(b)b$,

$$b_G^\star := \frac{2B_G}{\epsilon_G^2 - C_G}, \quad b_D^\star := \frac{2B_D}{\epsilon_D^2 - C_D}$$

【Estimation of the Critical Batch Size】

(Proposition 3.4 / Proofs are in Appendix A.10)

From the definition of B_G and C_G , the lower bound of $b_G^\star$ can be expressed for each optimizer as follows

(i) for Adam,

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

(ii) for AdaBelief,

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

(iii) for RMSProp,

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

Proposition 3.4 indicates that it is possible to estimate the critical batch size specific to the model-dataset-optimizer combination.

【Notation】

α^G : the learning rate for the Generator

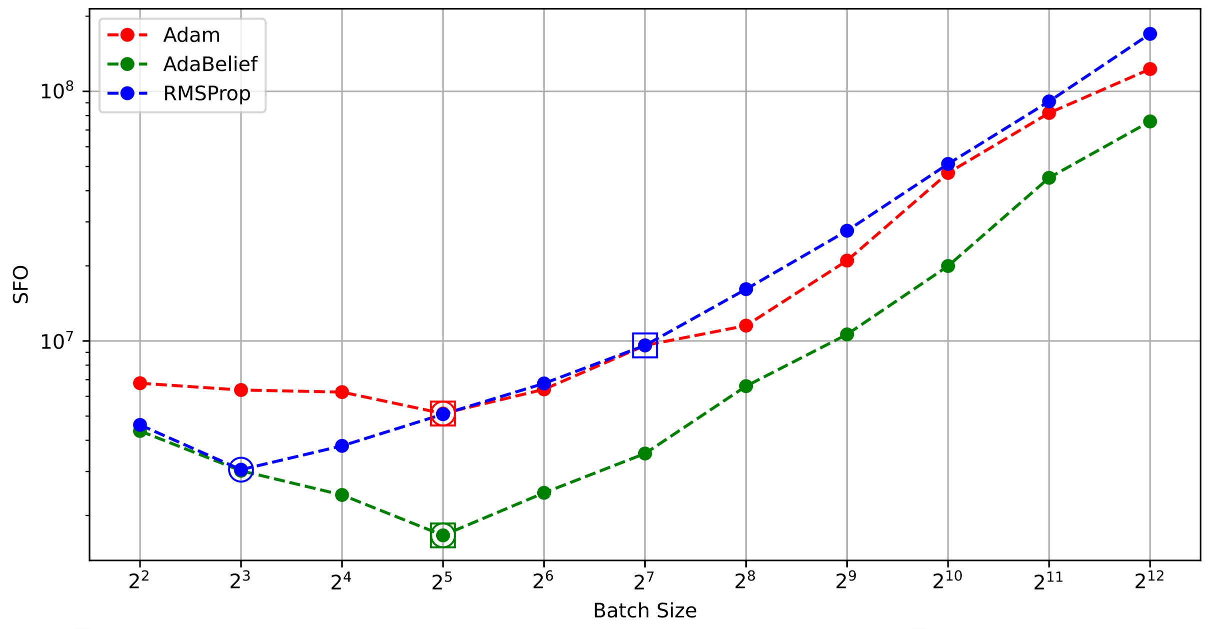
β_1^G, β_2^G : parameters for the adaptive optimizer

Θ : the dimensions of the Generator model

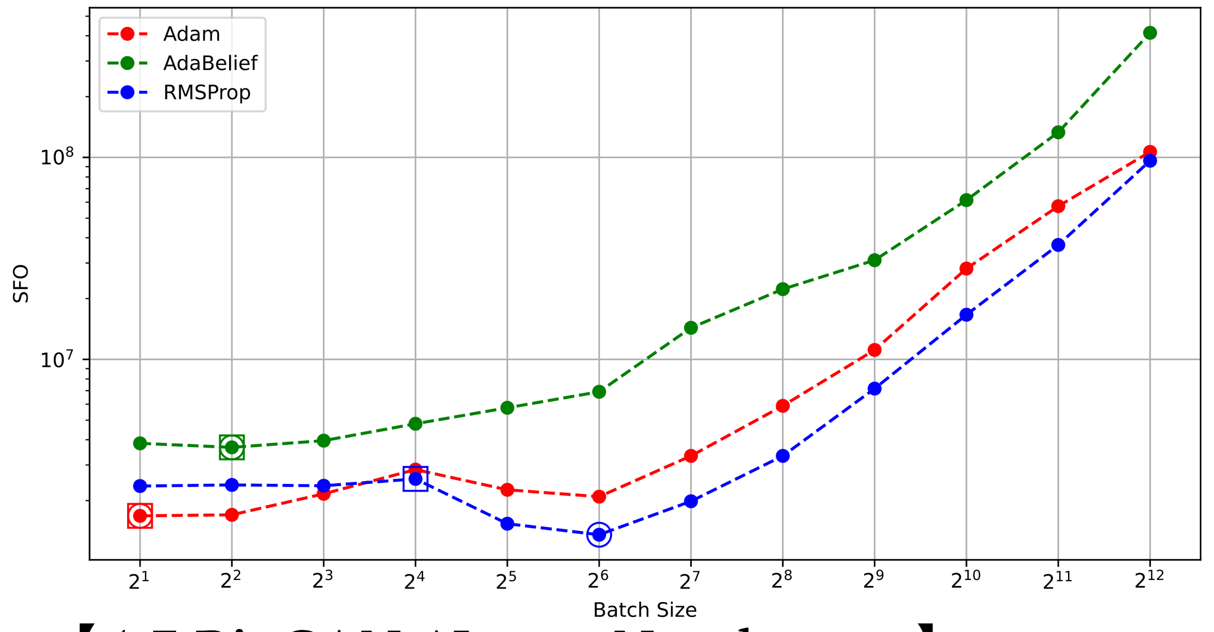
$|S|$: the number of items in the datasets

Numerical Results

【4.1 DCGAN / LSUN Bedroom dataset】



【4.2 WGAN-GP / CelebA dataset】



【4.3 BigGAN / ImageNet dataset】

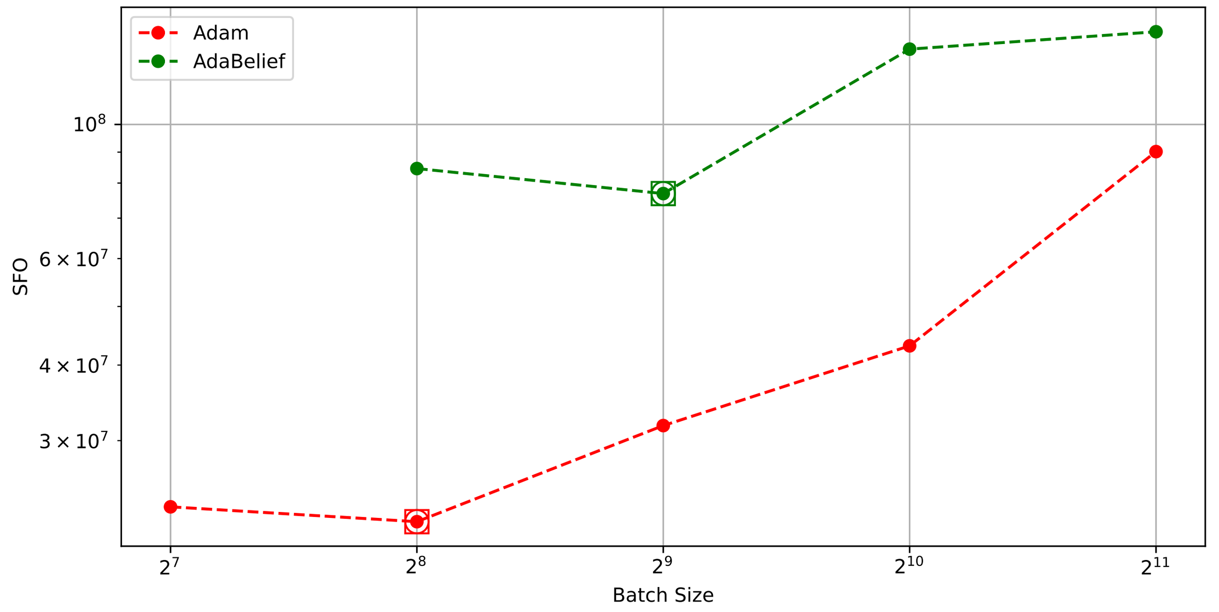


Table 3. Measured and estimated critical batch sizes

	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	-	-

【How to estimate critical batch size】

First, we back calculate σ_G^2/ϵ_G^3, the only unknown, from the measured critical batch size. According to the left figure, Adam's measured critical batch size is 2^5, so from Proposition 3.4(i), we can calculate

$$\sigma_G^2/\epsilon_G^3 \leq 788.7.$$

Using this ratio, we can estimate other optimizers' critical batch size. Moreover, this ratio can also be appropriated for another GAN's training, as long as the model adopted is the same. In fact, since both models used in Sections 4.1 and 4.2 have a DCGAN architecture, the ratios obtained from DCGAN training can be used to estimate the critical batch size for WGAN-GP training.

However, BigGAN training cannot use the ratios obtained in DCGAN training because the model is different. Calculating the ratio in the same way for BigGAN training, we find that

$$\sigma_G^2/\epsilon_G^3 \leq 530303.8.$$

Table 3 shows a comparison of the estimated and measured critical batch sizes.

References

[Sha+19] C.J. Shallue et al. "Measuring the effects of data parallelism on neural network training." *Journal of Machine Learning Research*, 20:1–49, 2019.

[Heu+17] M. Heusel et al. "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.