Conjugate Gradient Method for Generative Adversarial Networks

Hiroki Naganuma *[1,2], Hideaki Iiduka [3]
*: naganuma.hiroki@mila.quebec
[1] Mila [2] Université de Montréal [3] 明治大学 MEIJI UNIVERSITY

Introduction

【Motivation】

- **GANs applied in various fields:** image generation, style transfer, data augmentation, image inpainting, NLP, medical image analysis
- **GAN training as Local Nash Equilibria (LNE) problem;** Two time-scale update rules (TTUR) as effective algorithm; no convergence proof for constant learning rate schedule
- **LNE problems more complex than NN minimization problems;** Study aims to verify Conjugate Gradient method effectiveness considering current and past gradient directions

【Contribution】

- **Proposed CG-type algorithm** for solving LNE problems in GANs, using efficient CG parameters (FR, PRP, HS, DY formulas)
- **Convergence guarantee and rate analyses** for constant and diminishing learning rate rules (Theorem 3.1 & 3.2), summarized in Table 1
- **Extensive empirical study on convergence to LNE for SGD, momentum SGD, and CG-type algorithms;** CG-type outperforms SGD, momentum SGD, and Adam in GAN training (Figure 2, Table 3)

Problem formulation

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

Problem 2.1 *Find a pair* $(\boldsymbol{\theta}^*, \boldsymbol{w}^*) \in \mathbb{R}^\Theta \times \mathbb{R}^W$ *satisfying*

$$\nabla_{\boldsymbol{w}} \mathcal{L}_D(\boldsymbol{\theta}^*, \boldsymbol{w}^*) = \mathbf{0} \text{ and } \nabla_{\boldsymbol{\theta}} \mathcal{L}_G(\boldsymbol{\theta}^*, \boldsymbol{w}^*) = \mathbf{0}. \quad (1)$$

$\mathscr{L}_D$: A loss function of discriminator for a fixed $\theta \in \mathbb{R}^\Theta$
$\mathscr{L}_G$: A loss function of generator for a fixed $w \in \mathbb{R}^W$

Conjugate gradient methods

We consider a stationary point problem associated with unconstrained nonconvex optimization

$$\text{find a point } \boldsymbol{x}^* \in \mathbb{R}^d \text{ such that } \nabla f(\boldsymbol{x}^*) = \mathbf{0}, \quad (3)$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$

Newton method, quasi-Newton methods, and CG method. The CG method [NW06, Chapter 5], [HZ06] is defined as follows: given $x_0 \in \mathbb{R}^d$ and $d_0 := -\nabla f(x_0)$

$$\boldsymbol{x}_{n+1} := \boldsymbol{x}_n + \alpha_n \boldsymbol{d}_n,\ \boldsymbol{d}_{n+1} := -\nabla f(\boldsymbol{x}_{n+1}) + \beta_{n+1} \boldsymbol{d}_n, \quad (4)$$

where $(\alpha_n)_{n\in\mathbb{N}}$ is the sequence of step sizes (referred to as learning rates in the machine learning field), $\beta_{n+1} \in \mathbb{R}_+$, and d_n denotes the search direction called the CG direction.

CG for Local Nash Equilibrium Problem

【Theoretical Assumptions】

- Assumption 2.1
  - (A1) $\mathcal{L}_D^{(i)}\colon \mathbb{R}^\Theta \times \mathbb{R}^W \to \mathbb{R}$ $(i \in \mathcal{R})$ *and* $\mathcal{L}_G^{(i)}\colon \mathbb{R}^\Theta \times \mathbb{R}^W \to \mathbb{R}$ $(i \in \mathcal{S})$ *are continuously differentiable;*
  - (A2) *For* $\boldsymbol{w} \in \mathbb{R}^W$, $\mathcal{G}_{\boldsymbol{w}}\colon \mathbb{R}^\Theta \to \mathbb{R}^\Theta$ *is* L_1*-Lipschitz continuous.*[1] *For* $\boldsymbol{\theta} \in \mathbb{R}^\Theta$, $\mathcal{D}_{\boldsymbol{\theta}}\colon \mathbb{R}^W \to \mathbb{R}^W$ *is* L_2*-Lipschitz continuous.*

  (A1) the objective function is bounded, and (A2) the objective function is differentiable and its gradient vector is Lipschitz continuous.
- Assumption 3.1
  - (C1) $a_n := a \in \mathbb{R}_{++}$ *and* $b_n := b \in \mathbb{R}_{++}$ *for all* $n \in \mathbb{N}$.
  - (C2) $\beta^D := \sup\{\beta_n^D\colon n \in \mathbb{N}\} \in [0, 1/2)$, $\beta^G := \sup\{\beta_n^G\colon n \in \mathbb{N}\} \in [0, 1/2]$.
  - (C3) $(\boldsymbol{\theta}_n)_{n\in\mathbb{N}}$ *and* $(\boldsymbol{w}_n)_{n\in\mathbb{N}}$ *are almost surely bounded.*

  (C1) the step size sequence for stochastic gradient descent is chosen appropriately and (C2)+(C3) the gradients of the objective function are bounded.

【Theorem / Convergence Rate for Constant Learning Rate Rule】

- Theorem 3.1
  - provides analytical results for the convergence of the generator and discriminator parameters in the GAN training algorithm.
  - Proofs for the discriminator can also be similarly demonstrated.
  - Under Assumptions 2.1 and 3.1 for constant LR schedule (2.1 and 3.2 for diminishing LR schedule), the following hold:

(i) [Convergence] *For all* $\boldsymbol{\theta} \in \mathbb{R}^\Theta$,

$$\liminf_{n\to+\infty} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} \mathcal{L}_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq 2\tilde{K}_1^2 a + 2C_1 \tilde{K}_1 \beta^G. \quad (6)$$

In particular, 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* $((\boldsymbol{\theta}_{n_i}, \boldsymbol{w}_{n_i}))_{i\in\mathbb{N}}$ *converges almost surely to* $(\boldsymbol{\theta}^*, \boldsymbol{w}^*)$ *satisfying*

$$\mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_G(\boldsymbol{\theta}^*, \boldsymbol{w}^*)\|^2\right] \leq 2\tilde{K}_1^2 a + 2C_1 \tilde{K}_1 \beta^G. \quad (7)$$

(ii) [Convergence Rate] *For all* $\boldsymbol{\theta} \in \mathbb{R}^\Theta$ *and all* $N \geq 1$,

$$\frac{1}{N} \sum_{n\in[N]} \mathbb{E}\left[\langle \boldsymbol{\theta}_n - \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} \mathcal{L}_G(\boldsymbol{\theta}_n, \boldsymbol{w}_n)\rangle\right] \leq \frac{\mathbb{E}[\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}\|^2]}{2aN} + 2\tilde{K}_1^2 a + 2C_1 \tilde{K}_1 \beta^G. \quad (10)$$

where C_i and $\tilde{K}_i$ $(i = 1,2)$ be positive constants independent of n

(i) if the gradient of objective function vector is Lipschitz continuous, then **CG methods converges to a LNE**;
(ii) if the step size sequence is chosen appropriately, then **CG methods converges at a linear rate**.

【Summary of Theoretical Results】

Similarly to the above, we also **obtained convergence rates** for diminishing learning rates in the case of SGD, momentum, and CG methods under mild Assumption 3.2.

Table 1: Convergence rates of our algorithms with constant and diminishing learning rates.

		Constant learning rate $(a_n = a, b_n = b)$	Diminishing learning rate $(a_n = n^{-1/2}, b_n = n^{-1/2})$	Diminishing learning rate $(a_n = n^{-\eta_a}, b_n = n^{-\eta_b})$
SGD	Generator	$\mathcal{O}(N^{-1}) + C_1^G a$	$\mathcal{O}(N^{-1/2})$	$\mathcal{O}(N^{-\min\{\eta_a, 1-\eta_a\}})$
	Discriminator	$\mathcal{O}(N^{-1}) + C_1^D b$	$\mathcal{O}(N^{-1/2})$	$\mathcal{O}(N^{-\min\{\eta_b, 1-\eta_b\}})$
Momentum	Generator	$\mathcal{O}(N^{-1}) + C_1^G a + C_2^G \beta^G$	$\mathcal{O}(N^{-1/2})$	$\mathcal{O}(N^{-\min\{\eta_a, 1-\eta_a\}})$
	Discriminator	$\mathcal{O}(N^{-1}) + C_1^D b + C_2^D \beta^D$	$\mathcal{O}(N^{-1/2})$	$\mathcal{O}(N^{-\min\{\eta_b, 1-\eta_b\}})$
CG	Generator	$\mathcal{O}(N^{-1}) + C_1^G a + C_2^G \beta^G$	$\mathcal{O}(N^{-1/2})$	$\mathcal{O}(N^{-\min\{\eta_a, 1-\eta_a\}})$
	Discriminator	$\mathcal{O}(N^{-1}) + C_1^D b + C_2^D \beta^D$	$\mathcal{O}(N^{-1/2})$	$\mathcal{O}(N^{-\min\{\eta_b, 1-\eta_b\}})$

C_i^G and C_i^D $(i = 1, 2)$ are positive constants independent of learning rates a and b and number of iterations N. β^G and β^D are upper bounds of the CG parameter β_n^G and β_n^D, respectively (see Assumption 3.1(C2) for details). η_a and η_b satisfy $1/2 < \eta_b < \eta_a < 1$. The convergence rate is measured as the expectation of the variational inequality (see, e.g., (10) and (11)). See Theorems 3.1 and 3.2 for detailed convergence analyses.

Numerical Experiments

【Experimental Protocol】

- Toy-Example
  - Metric: Objective, Norm of Gradient
- Experiments on GANs
  - Model: SNGAN with ResNet Generator
  - Dataset: CIFAR10
  - Metric: FID, Loss, Norm of Gradient
  - Hyperparameter: BS=64, Initial LR = {5e-5,...5e-3}, LR Schedule={Constant, Diminishing}

【Toy Example】

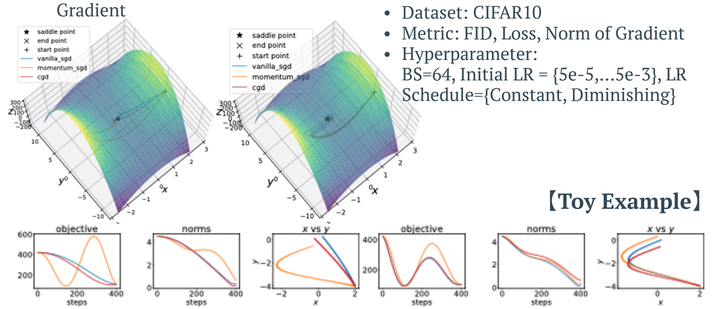

Figure 1. Toy example of minimax optimization. Objective function: $f(x, y) = (1 + x^2) \cdot (100 - y^2)$. **Top Left**: trajectory with a *constant* learning rate, **Top Right**: trajectory with a *diminishing* learning rate, **Bottom Left**: objective, norms, and x vs y with a *constant* learning rate schedule, **Bottom Right**: same as the bottom left but with a *diminishing* learning rate schedule. We tuned the initial learning rate (See Appendix D.3).

- **Toy example confirms convergence to LNE for appropriate learning rate**

【Experiments on GANs】

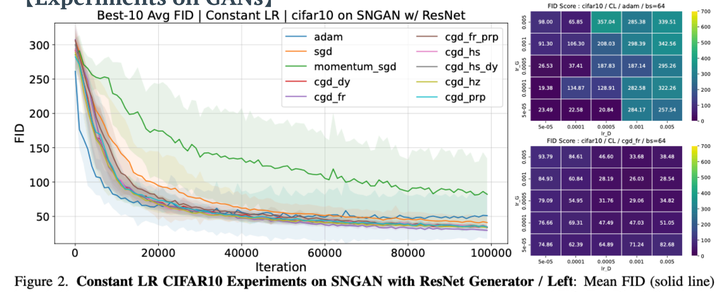

Figure 2. **Constant LR CIFAR10 Experiments on SNGAN with ResNet Generator / Left**: Mean FID (solid line) bounded by the maximum and minimum over the best ten runs (the shaded areas) in the sense of FID. **Top Right**: Sensitivity of learning rate to FID scores of Adam. **Bottom Right**: Sensitivity of learning rate to FID scores of CG method (FR β update rule) .

	adam	sgd	momentum_sgd	cgd_dy	cgd_fr	cgd_fr_prp	cgd_hs	cgd_hs_dy	cgd_hz	cgd_prp
Miyato+2017 (CL[1])	21.7	-	-	-	-	-	-	-	-	-
Ours (CL) / Best	19.38	30.34	34.42	30.13	**26.03**	30.31	29.29	29.54	29.41	28.94
Ours (DL[2]) / Best	51.96	40.55	73.66	35.29	32.17	31.06	30.03	29.09	**29.64**	30.47
Ours (CL) / Average	51.16±33.71	41.09±8.13	82.15±51.82	33.70±2.04	**29.63±2.26**	34.78±2.17	34.82±1.07	34.12±1.31	34.28±0.96	34.53±2.01
Ours (DL) / Average	135.86±32.65	51.80±7.84	205.29±59.333	42.39±12.89	38.20±8.82	37.39±2.74	**36.96±2.77**	39.90±3.06	38.24±2.97	37.94±3.62

Table 3: Best and Average FID scores for CIFAR10 on SNGAN with ResNet generator for *constant* and *dimimishing* learning rate scheduling.

- Experiments employ techniques such as Spectral Normalization to satisfy assumptions in the theory
- **Achieved competitive FID scores even with diminishing learning rates**
- CG method demonstrated stable high performance with lower sensitivity to hyperparameters compared to other optiomizer

References

[Heu+17] M. Heusel et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: International Confer- ence on Neural Information Processing Sys- tems (2017), pp. 1–12.

[NW06] Jorge Nocedal and Stephen J. Wright. Numeri- cal Optimization. second. New York, NY, USA: Springer, 2006.

[HZ06] W. W. Hager and H. Zhang. "A survey of non- linear conjugate gradient methods". In: Pacific Journal of Optimization 2.1 (2006), pp. 35–58.