Towards Better Understanding Of Adaptive Gradient Algorithms In Generative Adversarial Nets

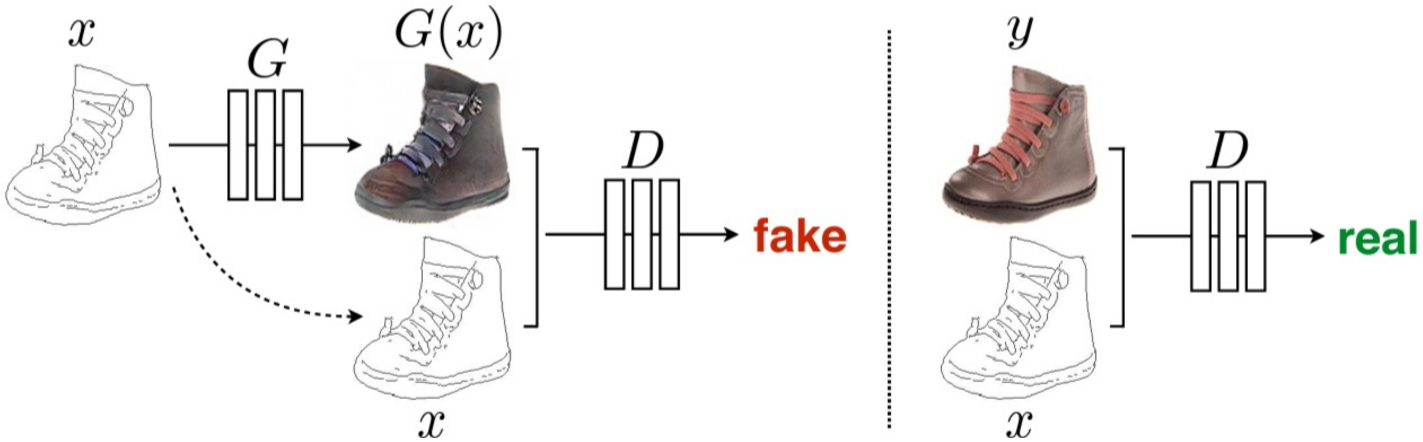
Yuting Hu yhu54@buffalo.edu

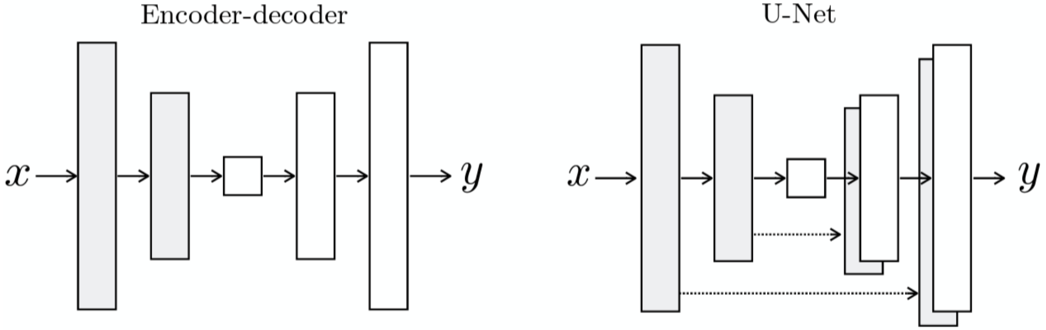
CONTENT

1. Background
2. Theoretical basics
3. OSG & OAGrad
4. Experiments
5. Novelty

Recap: Generative Adversarial Network

Def: *GAN is composed by a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game.*





$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x,z))]$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1] x, G(x,z))]$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G)$$

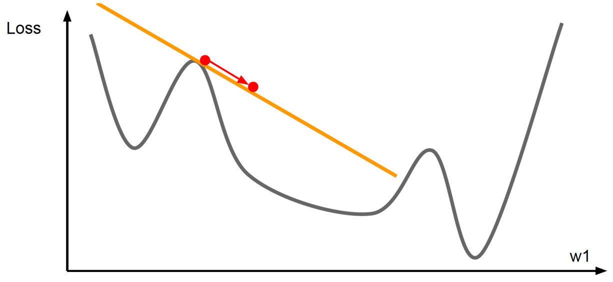
AdamGrad

Recap: Adaptive Gradient Descent

Def: *Using observed gradients to help optimization process adapt to local or global smoothness and convexity and automatically learn the step size.*

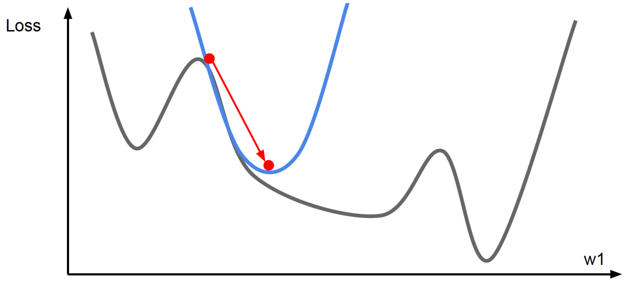
Adam -> Adaptive + Momentum

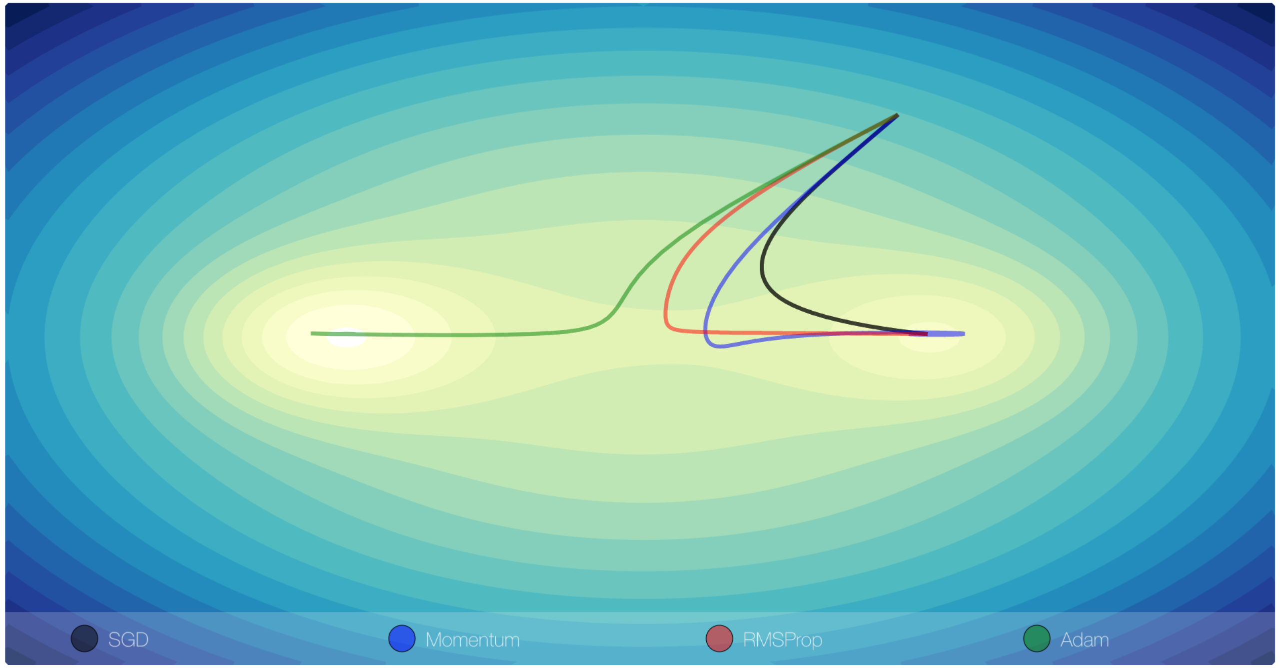
$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$$



$$\eta_t = \alpha \cdot m_t / \sqrt{V_t}$$

$$V_t = \beta_2 * V_{t-1} + (1 - \beta_2) g_t^2$$





MinMax Optimization

$$\min_{\mathbf{u}\in\mathcal{U}} \max_{\mathbf{v}\in\mathcal{V}} F(\mathbf{u},\mathbf{v}) := \mathbb{E}_{\xi\sim\mathcal{D}}\left[f(\mathbf{u},\mathbf{v};\xi)\right]$$

where $\mathcal{U}$, $\mathcal{V}$ are closed and convex sets, $F(\mathbf{u},\mathbf{v})$ is possibly non-convex in $\mathbf{u}$ and non-concave in $\mathbf{v}$.

Idea Goal: find a saddle point $(\boldsymbol{u}_*, \boldsymbol{v}_*)$ $\rightarrow$ $F(\boldsymbol{u}_*, \boldsymbol{v}) \leq F(\boldsymbol{u}_*, \boldsymbol{v}_*) \leq F(\boldsymbol{u}, \boldsymbol{v}_*)$ (*NP Hard*)

Final Goal: find the first-order stationary point $\rightarrow$ $\nabla_{\boldsymbol{u}} F(\boldsymbol{u}, \boldsymbol{v}) = 0, \nabla_{\boldsymbol{v}} F(\boldsymbol{u}, \boldsymbol{v}) = 0$ (*Necessary Cond*)

Def: $x = (\boldsymbol{u}, \boldsymbol{v}), T(x;\xi) = [\nabla_{\boldsymbol{u}} F(\boldsymbol{u}, \boldsymbol{v}; \xi), -\nabla_{\boldsymbol{v}} F(\boldsymbol{u}, \boldsymbol{v}; \xi)]^T$ (min min)

MinMax Optimization & SVI/MVI

Def: $x = (\boldsymbol{u}, \boldsymbol{v}), T(x;\xi) = [\nabla_{\boldsymbol{u}} F(\boldsymbol{u}, \boldsymbol{v}; \xi), -\nabla_{\boldsymbol{v}} F(\boldsymbol{u}, \boldsymbol{v}; \xi)]^T$

Goal: *solve* $\|T(x;\xi\| \leq \varepsilon$

Tool: ***variational inequality*** **SVI/MVI**

SVI: *Stampacchia Variational Inequalityinequality*

find x_* *such that* $\langle T(x_*),\ x -_* \rangle \geq 0$ *for* $\forall\ x \epsilon X$

MVI: *Minty Variational Inequalityinequality*

find x_* *such that* $\langle T(x),\ x - x_* \rangle \geq 0$ *for* $\forall\ x \epsilon X$

Note: *ε-first-order stationary point means* ||T(x;ξ)|| ≤ ε.

MinMax Optimization & SVI/MVI

Definition 1 (Monotonicity). *An operator T is monotone if $\langle T(\mathbf{x}) - T(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0$ for $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$. An operator T is pseudo-monotone if $\langle T(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq 0 \Rightarrow \langle T(\mathbf{y}), \mathbf{y} - \mathbf{x} \rangle \geq 0$ for $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$. An operator T is γ-strongly-monotone if $\langle T(\mathbf{x}) - T(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{\gamma}{2} \|\mathbf{x} - \mathbf{y}\|^2$ for $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$.*

Strong-monotonicity => **monotonicity** => **pseudo-monotonicity**

Conslusion: 1. SVI has a solution, MVI must has a resolution.

2. When F is convex in u and concave in v, T is monotone, the SVI solution is our target;

When F is non-convex in u and non-concave in v, If assuming T is Lipschitz continuous, our target is a subset of SVI solution;

MinMax Optimization & SVI/MVI

How to solve SVI? **Stochastic Approximation(SA)**

$$x^{k+1} = \Pi[x^k - \alpha_k F(\xi^k, x^k)],$$

where Π is the Euclidean projection onto X, $\{\xi^k\}$ is a sample of ξ and $\{\alpha_k\}$ is a sequence of positive steps. In [18], the almost sure (a.s.) convergence is proved assuming L-Lipschitz continuity of T, strong monotonicity or strict monotonicity of T, stepsizes satisfying $\sum_k \alpha_k = \infty, \sum_k \alpha_k^2 < \infty$ (with $0 < \alpha_k < 2\rho/L^2$, assuming that T is ρ-strongly monotone), and an unbiased oracle with uniform variance, i.e., there exists $\sigma > 0$ such that for all $x \in X$,

$$z^k = \Pi\left[x^k - \frac{\alpha_k}{N_k}\sum_{j=1}^{N_k} F(\xi_j^k, x^k)\right].$$

$$x^{k+1} = \Pi\left[x^k - \frac{\alpha_k}{N_k}\sum_{j=1}^{N_k} F(\eta_j^k, z^k)\right].$$

Ref: Iusem, Alfredo N., et al. "Extragradient method with variance reduction for stochastic variational inequalities." SIAM Journal on Optimization 27.2 (2017): 686-724.

Optimistic Stochastic Gradient

Algorithm 1 Optimistic Stochastic Gradient (OSG)

1: **Input:** $\mathbf{z}_0 = \mathbf{x}_0 = 0$
2: **for** $k = 1, \ldots, N$ **do**
3: $\quad \mathbf{z}_k = \Pi_{\mathcal{X}} \left[\mathbf{x}_{k-1} - \eta \cdot \frac{1}{m_{k-1}} \sum_{i=1}^{m_{k-1}} T(\mathbf{z}_{k-1}; \xi_{k-1}^i) \right]$
4: $\quad \mathbf{x}_k = \Pi_{\mathcal{X}} \left[\mathbf{x}_{k-1} - \eta \cdot \frac{1}{m_k} \sum_{i=1}^{m_k} T(\mathbf{z}_k; \xi_k^i) \right]$
5: **end for**

Define $\hat{\mathbf{g}}_k = \frac{1}{m_k} \sum_{i=1}^{m_k} T(\mathbf{z}_k; \xi_k^i)$, then the update rule of Algorithm 1 becomes

$$\mathbf{z}_k = \mathbf{x}_{k-1} - \eta \hat{\mathbf{g}}_{k-1}$$

and

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \eta \hat{\mathbf{g}}_k.$$

These two equalities together imply that

$$\mathbf{z}_{k+1} = \mathbf{x}_k - \eta \hat{\mathbf{g}}_k = \mathbf{x}_{k-1} - 2\eta \hat{\mathbf{g}}_k = \boxed{\mathbf{z}_k + \eta \hat{\mathbf{g}}_{k-1} - 2\eta \hat{\mathbf{g}}_k},$$

$$\mathbf{z}_{k+1} = \mathbf{z}_k - 2\eta \cdot \frac{1}{m_{k-1}} \sum_{i=1}^{m_k} T(\mathbf{z}_k; \xi_k^i) + \eta \cdot \frac{1}{m_{k-1}} \sum_{i=1}^{m_{k-1}} T(\mathbf{z}_{k-1}; \xi_{k-1}^i)$$

fixed gradient at step k，k-1

Theorem 1. *Suppose that Assumption 1 holds. Let $r_\alpha(\mathbf{z}_k) = \|\mathbf{z}_k - \Pi_{\mathcal{X}}(\mathbf{z}_k - \alpha T(\mathbf{z}_k))\|$. Let $\eta \leq 1/9L$ and run Algorithm 1 for N iterations. Then we have*

$$\frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left[r_\eta^2(\mathbf{z}_k) \right] \leq \frac{8\|\mathbf{x}_0 - \mathbf{x}_*\|^2}{N} + \frac{100\eta^2}{N} \sum_{k=0}^{N} \frac{\sigma^2}{m_k},$$

Corollary 1. *Consider the unconstrained case where $\mathcal{X} = \mathbb{R}^d$. Let $\eta \leq 1/9L$, and we have*

$$\frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\|T(\mathbf{z}_k)\|_2^2 \leq \frac{8\|\mathbf{x}_0 - \mathbf{x}_*\|^2}{\eta^2 N} + \frac{100}{N} \sum_{k=0}^{N} \frac{\sigma^2}{m_k},$$

Optimistic Stochastic Gradient

Corollary 1. *Consider the unconstrained case where $\mathcal{X} = \mathbb{R}^d$. Let $\eta \leq 1/9L$, and we have*

$$\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\|T(\mathbf{z}_k)\|_2^2 \leq \frac{8\|\mathbf{x}_0 - \mathbf{x}_*\|^2}{\eta^2 N} + \frac{100}{N}\sum_{k=0}^{N}\frac{\sigma^2}{m_k},$$

Conclusion

- (Increasing Minibatch Size) Let $\eta = \frac{1}{9L}$, $m_k = k + 1$. To guarantee $\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\|T(\mathbf{z}_k)\|_2^2 \leq \epsilon^2$, the total number of iterations is $N = \widetilde{O}(\epsilon^{-2})$, and the total complexity is $\sum_{k=1}^{N} m_k = \widetilde{O}(\epsilon^{-4})$, where $\widetilde{O}(\cdot)$ hides a logarithmic factor of ϵ.

- (Constant Minibatch Size) Let $\eta = \frac{1}{9L}$, $m_k = 1/\epsilon^2$. To guarantee $\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\|T(\mathbf{z}_k)\|_2^2 \leq \epsilon^2$, the total number of iterations is $N = O(\epsilon^{-2})$, and the total complexity is $\sum_{k=0}^{N} m_k = O(\epsilon^{-4})$.

Optimistic AdaGrad

***Recap** AdaGrad in Minimization Probelm:*

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \mathbb{E}_{\zeta\sim\mathcal{P}} f(\mathbf{w};\zeta) \qquad \mathbf{w}_{t+1} = \mathbf{w}_t - \eta H_t^{-1}\hat{\mathbf{g}}_t.$$

$$\text{where } \eta > 0,\ \hat{\mathbf{g}}_t = \nabla f(\mathbf{w}_t;\zeta_t),\ H_t = \operatorname{diag}\left(\left(\textstyle\sum_{i=1}^t \hat{\mathbf{g}}_i \circ \hat{\mathbf{g}}_i\right)^{\frac{1}{2}}\right)$$

***Optimistic AdaGrad** in MinMax Probelm:*

Algorithm 2 Optimistic AdaGrad (OAdagrad)

1: **Input:** $\mathbf{z}_0 = \mathbf{x}_0 = 0$, $H_0 = \delta I$
2: **for** $k = 1, \ldots, N$ **do**
3: $\quad \mathbf{z}_k = \mathbf{x}_{k-1} - \eta H_{k-1}^{-1}\widehat{\mathbf{g}}_{k-1}$
4: $\quad \mathbf{x}_k = \mathbf{x}_{k-1} - \eta H_{k-1}^{-1}\widehat{\mathbf{g}}_k$
5: $\quad$ Update $\widehat{\mathbf{g}}_{0:k} = [\widehat{\mathbf{g}}_{0:k-1}\ \widehat{\mathbf{g}}_k]$, $s_{k,i} = \|\widehat{\mathbf{g}}_{0:k,i}\|$, $i = 1, \ldots, d$ and set $H_k = \delta I + \operatorname{diag}(s_{k-1})$
6: **end for**

Optimistic AdaGrad

Optimistic AdaGrad *in MinMax Probelm:*

Algorithm 2 Optimistic AdaGrad (OAdagrad)

1: **Input:** $\mathbf{z}_0 = \mathbf{x}_0 = 0$, $H_0 = \delta I$
2: **for** $k = 1, \ldots, N$ **do**
3: $\quad \mathbf{z}_k = \mathbf{x}_{k-1} - \eta H_{k-1}^{-1} \widehat{\mathbf{g}}_{k-1}$
4: $\quad \mathbf{x}_k = \mathbf{x}_{k-1} - \eta H_{k-1}^{-1} \widehat{\mathbf{g}}_k$
5: $\quad$ Update $\widehat{\mathbf{g}}_{0:k} = [\widehat{\mathbf{g}}_{0:k-1}\ \widehat{\mathbf{g}}_k]$, $s_{k,i} = \|\widehat{\mathbf{g}}_{0:k,i}\|$, $i = 1, \ldots, d$ and set $H_k = \delta I + \mathrm{diag}(s_{k-1})$
6: **end for**

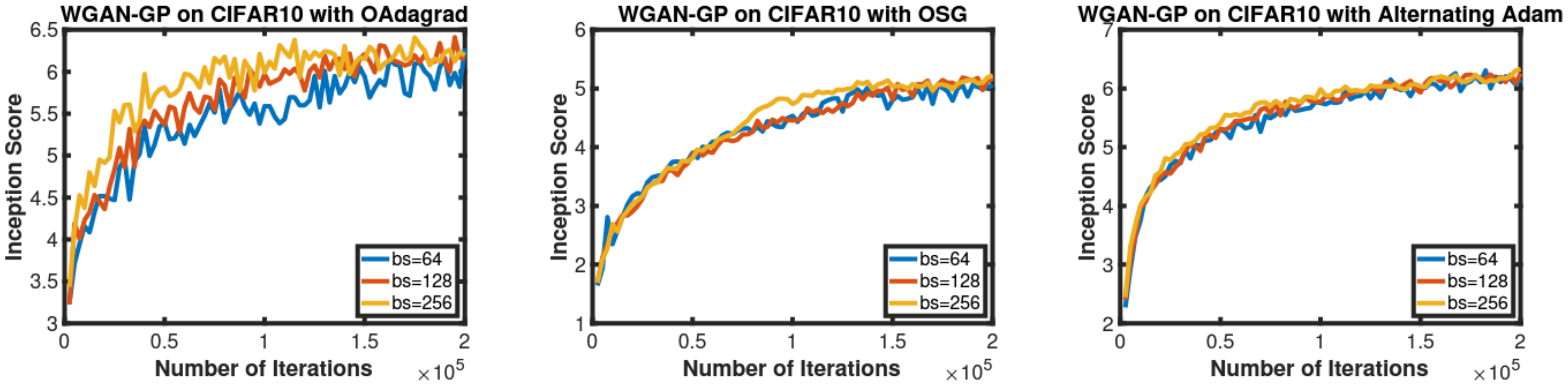
Theorem 2. *Suppose Assumption 1 and 2 hold. Suppose $\|\widehat{\mathbf{g}}_{1:k,i}\|_2 \leq \delta k^\alpha$ with $0 \leq \alpha \leq 1/2$ for every $i = 1, \ldots, d$ and every $k = 1, \ldots, N$. When $\eta \leq \frac{\delta}{9L}$, after running Algorithm 2 for N iterations, we have*

$$\frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\|T(\mathbf{z}_k)\|^2_{H_{k-1}^{-1}} \leq \frac{8D^2\delta^2(1 + d(N-1)^\alpha)}{\eta^2 N} + \frac{100\left(\sigma^2/m + d\left(2\delta^2 N^\alpha + G^2\right)\right)}{N}. \tag{6}$$

To make sure $\frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\|T(\mathbf{z}_k)\|^2_{H_{k-1}^{-1}} \leq \epsilon^2$, the number of iterations is $N = O\left(\epsilon^{-\frac{2}{1-\alpha}}\right)$.

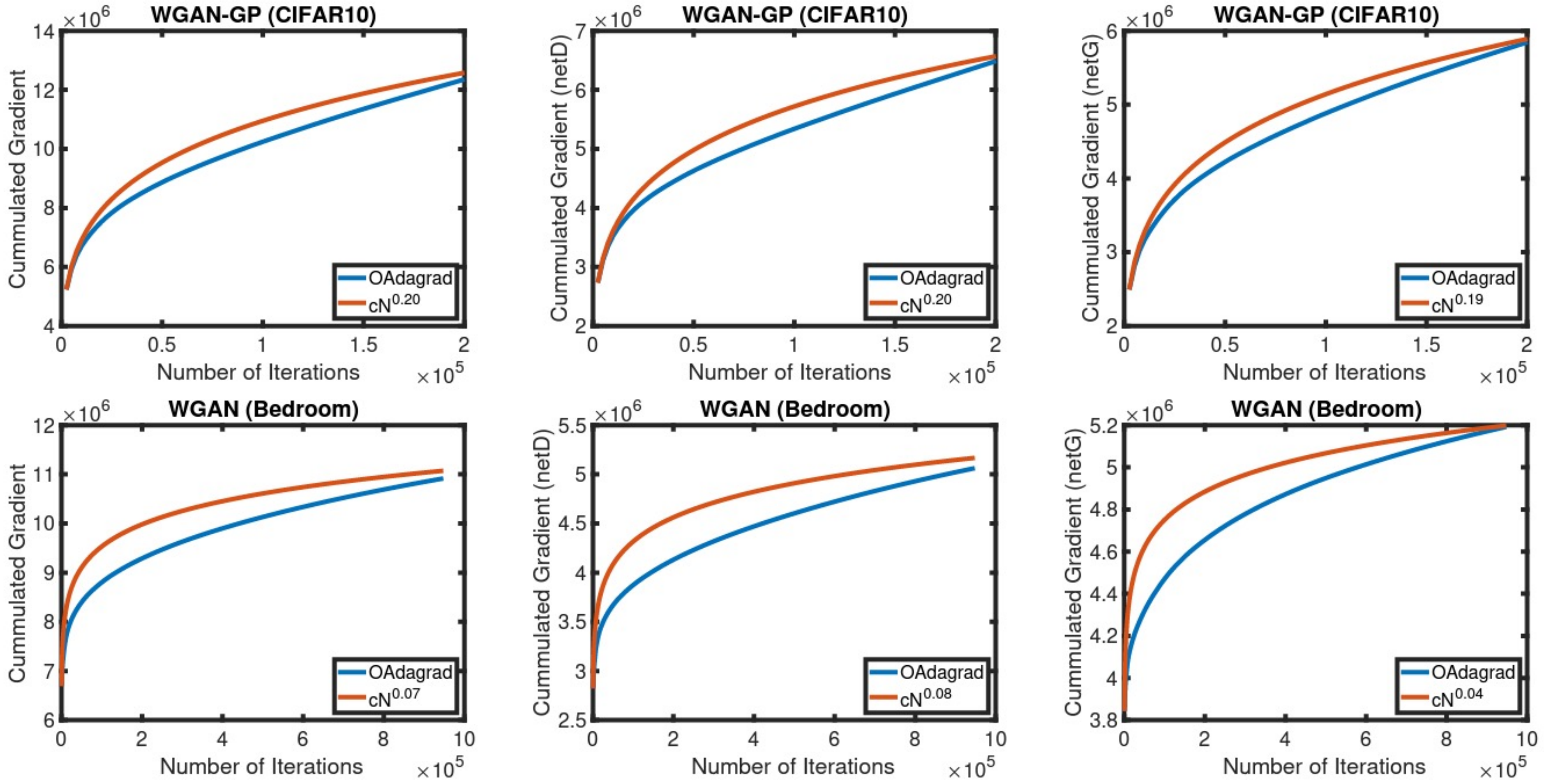
Experiments

Wasserstein GAN with Gradient Penalty on CIFAR10



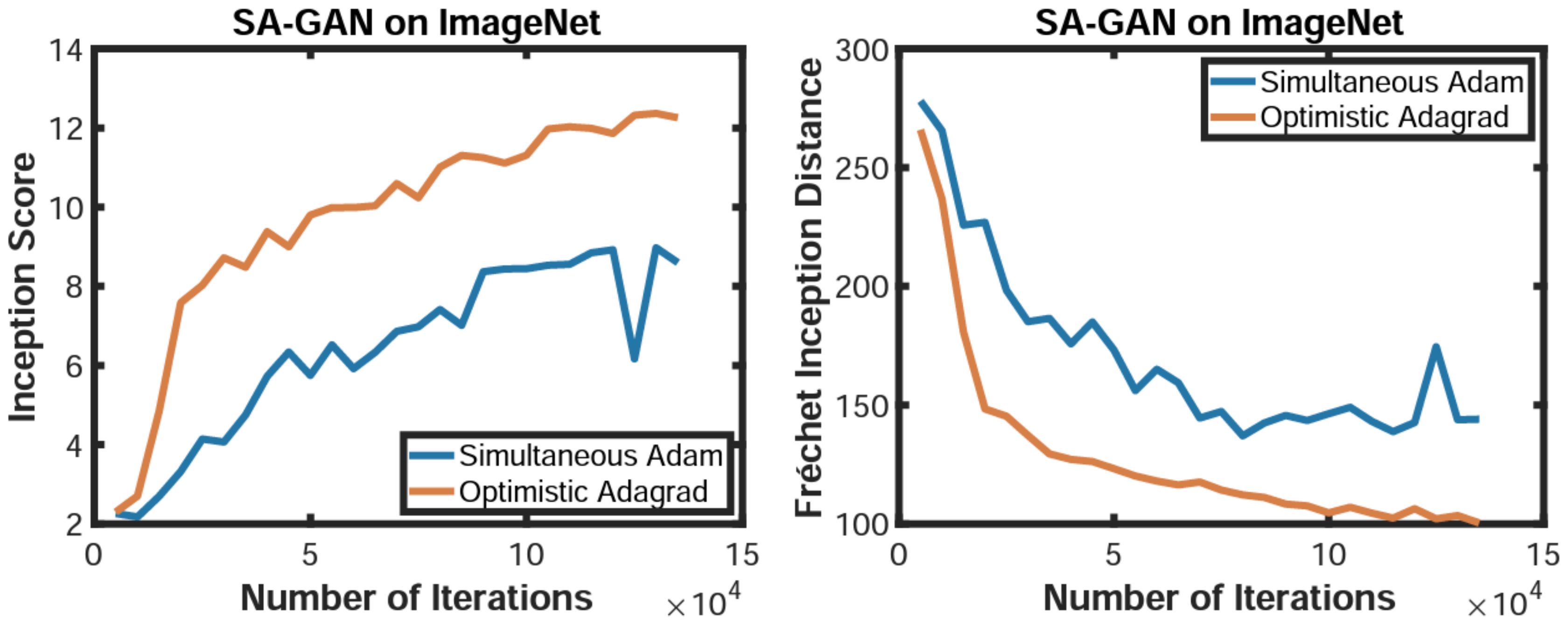
Experiments

Growth Rate Analysis of Cumulative Stochastic Gradient



Experiments

Self-attention GAN on ImageNet



(a) Inception Score (b) FID

Novelty

1. formulate the problem of first-order stationary point of minmax optimization as a variational inequality problem, and use stochastic approximation(SA) method to solve SVI.

2. provided a variant OSG for solving a class of nonconvex non-concave min-max problem and establish $O(\varepsilon^{-4})$ complexity for finding-first-order stationary point.

3.provided an adaptive variant of OSG called OAdagrad and reveal an improved adaptive complexity $O\left(\epsilon^{-\frac{2}{1-\alpha}}\right)$ *, where α characterizes the growth rate of the cumulative stochastic gradient and $0 \leq \alpha \leq 1/2$.*

Thank you!

Any questions?