GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter

LIT AI Lab & Institute of Bioinformatics, Johannes Kepler Universität Linz, Austria

JOHANNES KEPLER UNIVERSITY LINZ
LIT AI LAB

Abstract

Generative Adversarial Networks (GANs) excel at creating realistic images with complex models for which maximum likelihood is infeasible. However, the convergence of GAN training has still not been proved. We propose a two time-scale update rule (TTUR) for training GANs with stochastic gradient descent on arbitrary GAN loss functions. TTUR has an individual learning rate for both the discriminator and the generator. Using the theory of stochastic approximation, we prove that the TTUR converges under mild assumptions to a stationary local Nash equilibrium. The convergence carries over to the popular Adam optimization, for which we prove that it follows the dynamics of a heavy ball with friction and thus prefers flat minima in the objective landscape. For the evaluation of the performance of GANs at image generation, we introduce the `Fréchet Inception Distance'' (FID) which captures the similarity of generated images to real ones better than the Inception Score. In experiments, TTUR improves learning for DCGANs and Improved Wasserstein GANs (WGAN-GP) outperforming conventional GAN training on CelebA, CIFAR-10, SVHN, LSUN Bedrooms, and the One Billion Word Benchmark.

General GAN Update Settings

Generator Updates:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + a(n)\left(\boldsymbol{h}(\boldsymbol{\theta}_n, \boldsymbol{w}_n) + \boldsymbol{M}_n^{(\theta)}\right)$$

learning rate: $a(n)$
true gradient: $h(\boldsymbol{\theta}, \boldsymbol{w})$
gradient noise: $\boldsymbol{M}^{(\theta)}$
stochastic gradient: $\tilde{\boldsymbol{h}}(\boldsymbol{\theta}, \boldsymbol{w}) = \boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{w}) + \boldsymbol{M}^{(\theta)}$

Discriminator Updates:

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n + b(n)\left(\boldsymbol{g}(\boldsymbol{\theta}_n, \boldsymbol{w}_n) + \boldsymbol{M}_n^{(w)}\right)$$

learning rate: $b(n)$
true gradient: $g(\boldsymbol{\theta}, \boldsymbol{w})$
gradient noise: $\boldsymbol{M}^{(w)}$
stochastic gradient: $\tilde{\boldsymbol{g}}(\boldsymbol{\theta}, \boldsymbol{w}) = \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}) + \boldsymbol{M}^{(w)}$

Two Time-Scale Update Rule Converges

Theorem 1 (Borkar 1997)**.** *If the assumptions are satisfied, then the updates converge to* $(\boldsymbol{\theta}^*, \boldsymbol{\lambda}(\boldsymbol{\theta}^*))$ *a.s.*

$\boldsymbol{\theta}^*$, $\boldsymbol{\lambda}(\boldsymbol{\theta}^*)$ are local asymptotically stable attractors, $\boldsymbol{g}(\boldsymbol{\theta}^*, \boldsymbol{\lambda}(\boldsymbol{\theta}^*)) = \boldsymbol{0}$, $\boldsymbol{h}(\boldsymbol{\theta}^*, \boldsymbol{\lambda}(\boldsymbol{\theta}^*)) = \boldsymbol{0}$, the solution $(\boldsymbol{\theta}^*, \boldsymbol{\lambda}(\boldsymbol{\theta}^*))$ is a **stationary local Nash equilibrium.** The assumptions are:

(A1) The gradients $\boldsymbol{h}$ and $\boldsymbol{g}$ are Lipschitz.

(A2) $\sum_n a(n) = \infty$, $\sum_n a^2(n) < \infty$, $\sum_n b(n) = \infty$, $\sum_n b^2(n) < \infty$, $a(n) = o(b(n))$.

(A3) The stochastic gradient errors $\{\boldsymbol{M}_n^{(\theta)}\}$ and $\{\boldsymbol{M}_n^{(w)}\}$ are martingale difference sequences w.r.t. the increasing σ-field $\mathcal{F}_n = \sigma(\boldsymbol{\theta}_l, \boldsymbol{w}_l, \boldsymbol{M}_l^{(\theta)}, \boldsymbol{M}_l^{(w)}, l \leqslant n), n \geqslant 0$ with $\mathrm{E}\left[\|\boldsymbol{M}_n^{(\theta)}\|^2 \mid \mathcal{F}_n^{(\theta)}\right] \leqslant B_1$ and $\mathrm{E}\left[\|\boldsymbol{M}_n^{(w)}\|^2 \mid \mathcal{F}_n^{(w)}\right] \leqslant B_2$, where B_1 and B_2 are positive deterministic constants.

(A4) For each $\boldsymbol{\theta}$, the ODE $\dot{\boldsymbol{w}}(t) = \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}(t))$ has a local asymptotically stable attractor $\boldsymbol{\lambda}(\boldsymbol{\theta})$ within a domain of attraction G_θ such that $\boldsymbol{\lambda}$ is Lipschitz. The ODE $\dot{\boldsymbol{\theta}}(t) = \boldsymbol{h}(\boldsymbol{\theta}(t), \boldsymbol{\lambda}(\boldsymbol{\theta}(t)))$ has a local asymptotically stable attractor $\boldsymbol{\theta}^*$ within a domain of attraction.

(A5) $\sup_n \|\boldsymbol{\theta}_n\| < \infty$ and $\sup_n \|\boldsymbol{w}_n\| < \infty$.

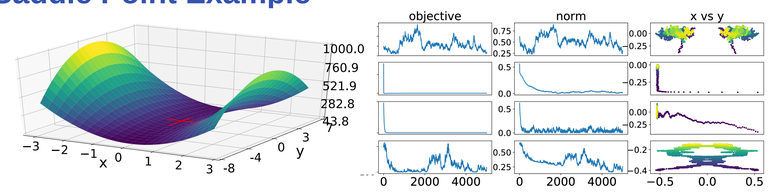
Fréchet Inception Distance to Evaluate GANs

For $p(.)$ the distribution of model samples and $p_w(.)$ the distribution of the real world samples the equality $p(.) = p_w(.)$ holds if and only if $\int p(.) f(x) dx = \int p_w(.) f(x) dx$ where $f(x)$ is the first and second order polynomial of the coding layer of an Inception model. From the resulting moments the vector of mean and covariance $(\boldsymbol{m}, \boldsymbol{C})$ can be computed.

The Fréchet distance $d(.,.)$ between the Gaussian with mean and covariance $(\boldsymbol{m}, \boldsymbol{C})$ obtained from $p(.)$ and the Gaussian $(\boldsymbol{m}_w, \boldsymbol{C}_w)$ obtained from $p_w(.)$ is called the "Fréchet Inception Distance" (FID), given by:

$$d^2((\boldsymbol{m}, \boldsymbol{C}), (\boldsymbol{m}_w, \boldsymbol{C}_w)) = \|\boldsymbol{m} - \boldsymbol{m}_w\|_2^2 + \mathrm{Tr}\left(\boldsymbol{C} + \boldsymbol{C}_w - 2\left(\boldsymbol{C}\boldsymbol{C}_w\right)^{1/2}\right) .$$

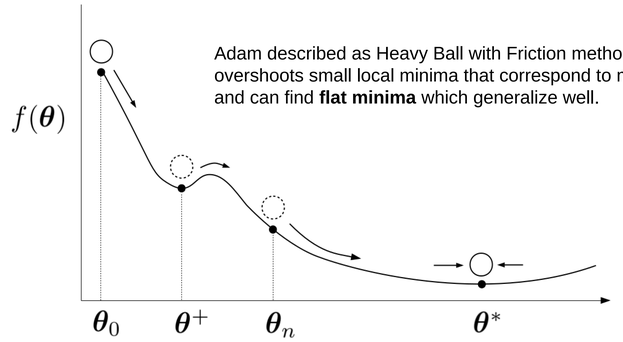
Saddle Point Example



TTUR on a simple toy min/max saddle point problem. The objective $\min_x \max_y f(x, y) = (1 + x^2)(100 - y^2)$ with a saddle point at $(\mathrm{x},\mathrm{y}) = (0,0)$ fulfills assumption A4. The norm $\|(x, y)\|$ measures the distance of the parameter vector to the saddle point. (x, y) is updated by *gradient descent* in x and *gradient ascent* in y with additive Gaussian noise to simulate a stochastic update. The updates should converge to the saddle point $(\mathrm{x},\mathrm{y}) = (0,0)$ with objective value $\mathrm{f}(0,0) = 100$ and norm 0.

- First row: One time-scale update rule with learning rate 0.01 diverges.
- Second row: One time scale update rule with learning rate 0.001 converges.
- Third row: TTUR with learning rates 0.0001 for x and 0.01 for y converges faster than with equal time-scale updates. Moves directly to the saddle point.
- Fourth row: TTUR with learning rates 0.01 and 0.0001 converges but slower.

Adam Follows a Heavy Ball with Friction ODE



Adam described as Heavy Ball with Friction method typically overshoots small local minima that correspond to mode collapse and can find **flat minima** which generalize well.

Theorem 2. *If Adam is used with* $\beta_1 = 1 - a(n+1)r(n)$, $\beta_2 = 1 - \alpha a(n+1)r(n)$ *and with* ∇f *as the full gradient of the lower bounded, continuously differentiable objective* f*, then for stationary second moments of the gradient, Adam follows the differential equation for Heavy Ball with Friction (HBF):*

$$\ddot{\boldsymbol{\theta}}_t \;+\; a(t)\;\dot{\boldsymbol{\theta}}_t \;+\; \nabla f(\boldsymbol{\theta}_t) \;=\; \boldsymbol{0}\;.$$

Adam converges for gradients ∇f *that are L-Lipschitz.*

Adam described by a differential equation with the Lyapunov function:

$$E(t) = \frac{1}{2}|\dot{\boldsymbol{\theta}}(t)|^2 \;+\; f(\boldsymbol{\theta}(t)) \qquad \text{with}$$

$$\dot{E}(t) = -\,a(t)|\dot{\boldsymbol{\theta}}(t)|^2 \;<\; 0\;.$$

Gadat et al. derived a discrete and stochastic version of the HBF:

$$\boldsymbol{\theta}_{n+1} \;=\; \boldsymbol{\theta}_n \;-\; a(n+1)\;\boldsymbol{m}_n$$

$$\boldsymbol{m}_{n+1} \;=\; \boldsymbol{m}_n \;+\; a(n+1)\;r(n)\;\left(\nabla f(\boldsymbol{\theta}_n) \;-\; \boldsymbol{m}_n\right) \;+\; a(n+1)\;r(n)\;\boldsymbol{M}_{n+1}\;.$$

The recursion can be rewritten as

$$\boldsymbol{\theta}_{n+1} \;=\; \boldsymbol{\theta}_n \;-\; a(n+1)\;\boldsymbol{m}_n$$

$$\boldsymbol{m}_{n+1} \;=\; \left(1 \;-\; a(n+1)\;r(n)\right)\;\boldsymbol{m}_n \;+\; a(n+1)\;r(n)\;\left(\nabla f(\boldsymbol{\theta}_n) \;+\; \boldsymbol{M}_{n+1}\right)\;.$$

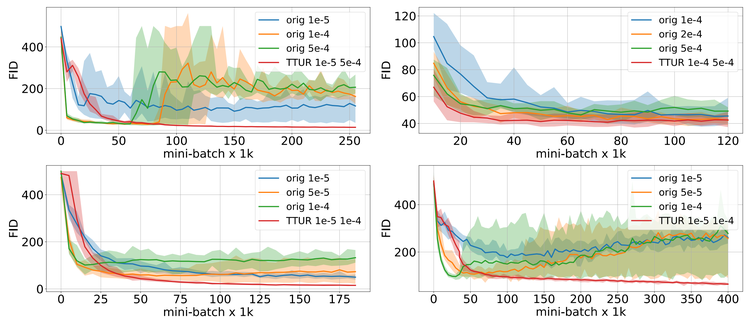
The recursion is the first moment update of Adam.

Learning GANs with TTUR and Adam converges.

Experiments

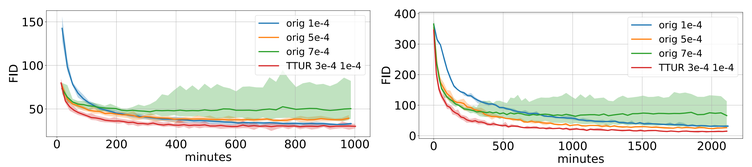
Model Selection and Evaluation. Learning rates are optimized to be small to ensure convergence but large enough to allow fast learning. Evaluation metric is the FID for image and the Jensen-Shannon divergence (JSD) for language data. WGAN-GP updates the critic five times for image and ten times for language data per iteration, TTUR updates the critic only once, therefore the training progress is aligned to wall-clock time for better comparison. TTUR learning rates are given for the discriminator b and generator a as: "TTUR b a".

DCGAN on Image Data:



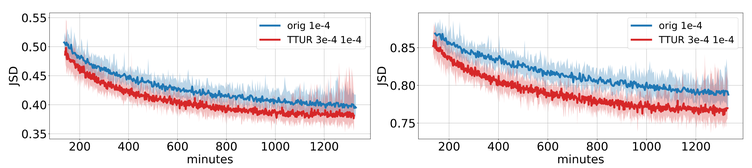
Mean FID (solid line) surrounded by a shaded area bounded by the maximum and minimum over 8 runs. **Top left:** CelebA. **Top right:** CIFAR-10, starting at mini-batch 10k. **Bottom left:** SVHN. **Bottom right:** LSUN Bedrooms.

WGAN-GP on Image Data:



Mean FID (solid line) surrounded by a shaded area bounded by the maximum and minimum over 8 runs, aligned to wall-clock time. **Left:** CIFAR-10, starting at minute 20. **Right:** LSUN Bedrooms.

WGAN-GP on Language Data (One Billion Word Benchmark):



Mean normalized JSD (solid line) surrounded by a shaded area bounded by the maximum and minimum over 10 runs, aligned to wall-clock time, starting at minute 150. **Left:** 4-gram statistics. **Right:** 6-gram statistics.

Results:

DCGAN Image								
dataset	*method*	*b, a*	*updates*	*FID*	*method*	*b = a*	*updates*	*FID*
CelebA	TTUR	1e-5, 5e-4	225k	**12.5**	orig	5e-4	70k	21.4
CIFAR-10	TTUR	1e-4, 5e-4	75k	**36.9**	orig	1e-4	100k	37.7
SVHN	TTUR	1e-5, 1e-4	165k	**12.5**	orig	5e-5	185k	21.4
LSUN	TTUR	1e-5, 1e-4	340k	**57.5**	orig	5e-5	70k	70.4
WGAN-GP Image								
dataset	*method*	*b, a*	*time(m)*	*FID*	*method*	*b = a*	*time(m)*	*FID*
CIFAR-10	TTUR	3e-4, 1e-4	700	**24.8**	orig	1e-4	800	29.3
LSUN	TTUR	3e-4, 1e-4	1900	**9.5**	orig	1e-4	2010	20.5
WGAN-GP Language								
n-gram	*method*	*b, a*	*time(m)*	*JSD*	*method*	*b = a*	*time(m)*	*JSD*
4-gram	TTUR	3e-4, 1e-4	1150	**0.35**	orig	1e-4	1040	0.38
6-gram	TTUR	3e-4, 1e-4	1120	**0.74**	orig	1e-4	1070	0.77

Performance with respect to the FID and JSD for optimized number of updates.