The Implicit Metropolis-Hastings Algorithm

Seminar on BMML, HSE

Kirill Neklyudov[1,2,3] Evgenii Egorov[4] Dmitry Vetrov[1,2,3]

Samsung AI Center Moscow[1],

Samsung-HSE Laboratory[2],

HSE[3], *Moscow*, *Russia*

Skoltech[4], *Moscow*, *Russia*

October 11, 2019

Outline

Markov Chains and The Metropolis-Hastings Algorithm: Revisited (a bit)

Motivation

- We should revised a bit
  The Metropolis-Hastings Algorithm
- As I want to refer some books, I want to introduce specific notation
- I am sorry about that

Books/Papers:

- Roberts, G. O., Rosenthal, J. S., et al. General state space markov chains and mcmc algorithms.Probability surveys, 1:20–71, 2004
- Peskun P. H. Optimum monte-carlo sampling using markov chains //Biometrika. – 1973. – . 60. – . 3. – . 607-612.
- Mira A., Geyer C. J. Ordering Monte Carlo Markov chains. – University of Minnesota, 1999

What is Markov Chain?

What is time homogeneous, discrete time Markov Chain?

Intuition

When considering a Markov chain on a general state space X (comparing to the discrete state space), we should think about **sets**, rather than points

The transition probabilities $Pr\{X_n \in B | X_{n-1} = x\}$ are specified by a kernel $P(x, B)$:

- for each fixed x the function $B \to P(x, B)$ is a probability measure
- for each fixed B the function $x \to P(x, B)$ is a measurable function

And we could write n-step kernel:

$$P^n(x, B) = \int_X P(z, B) P^{n-1}(x, dz), \ n \geq 2.$$

Our case is nice

If the distribution of X_n given X_{n-1} is a continuous distribution on $\mathbb{R}^d$ with some density $q(y|x)$, then the kernel could be written in the following way:

$$P(x, B) = \int_B q(y|x) dy.$$

Special kernel: The Identity Kernel

All stay in place kernel

Consider special Markov kernel, which produce its self very uninteresting chain, still it is very useful.

- for fixed x, the measure $I(x, \cdot)$ is the probability measure concentrated at x, $\delta(x)$
- for fixed B, the function $I(\cdot, B)$ is the indicator of the set B.

Also it is clear example that kernels are non-commutative (in discrete states it was matrix multiplication):

$$(IP)(x, B) = \int_X I(x, dy)P(y, B) = \int_X \delta_x(dy)P(y, B) = P(x, B)$$

$$(PI)(x, B) = \int_X P(x, dy)I(y, B) = \int_B P(x, dy) = P(x, B)$$

States classification

Discrete State Space
Define the occupation time of the **state** i to be

$$\eta_i := \sum_{n=1}^{\infty} I\{X_n = i\}.$$

Def.

The state is recurrent if $\mathbb{E}\eta_i = \infty$.

General State
Define the occupation time of a **set** R to be

$$\eta_R := \sum_{n=1}^{\infty} I\{X_n \in R\}.$$

Def.

The set R is recurrent if $\mathbb{E}\eta_R = \infty$ for all $x \in R$.

Invariant measures and Reversible Markov chains

Def.

A measure p on $\mathbb{B}(X)$ with the property:

$$p(A) = \int_X p(dx)P(x, A),\ \forall A \in \mathbb{B}(X),$$

will be called **invariant** and the chain **p-invariant**.

Def.

A **p-reversible** Markov chain is a **p-invariant** Markov chain satisfying

$$P_\pi(X_0 \in A_0, X_1 \in A_1) = P_\pi(X_0 \in A_1, X_1 \in A_0),$$

i.e $\int_{A_0} p(dx)P(x, A_1) = \int_{A_1} p(dx)P(x, A_0)$.

"Detailed balance"

As soon as we have densities for kernel, we could write it point-wise:

$$p(x)q(y|x) = p(y)q(x|y)$$

Accept-Reject Kernel

- Consider $p_{MH}(x,y) = q(y|x)\alpha(x,y)$ and the detailed balance holds:

$$p(x)p_{MH}(x,y) = p(x)q(y|x)\alpha(x,y) = p(y)q(x|y)\alpha(y,x) = p(y)p_{MH}(y,x).$$

- Consider the kernel: $P(x,A) = \int_A p_{MH}(x,y)dy + r(x)I(x,A)$

Then we could check the p-reversibility:

$$\int_{A_0} p(dx)P(x,A_1) = \int_{A_0} p(x)dx \left[\int_{A_1} p_{MH}(x,y)dy + r(x)I(x,A_1)\right] =$$

$$= \int_{A_0}\int_{A_1} p(x)p_{MH}(x,y)dxdy + \int_{A_0\cap A_1} p(x)r(x)dx =$$

$$= \int_{A_1}\int_{A_0} p(x)p_{MH}(x,y)dxdy + \int_{A_0\cap A_1} p(x)r(x)dx = \int_{A_1} p(dx)P(x,A_0)$$

Metropolis-Hastings, Peskun

The question is how to choose the function $\alpha(x,y)$? With different choice the algorithm still valid!

The thing to keep in mind

$\int |p(x)q(y|x)\alpha(x,y) - p(y)q(x|y)\alpha(y,x)|dxdy$

Metropolis-Hastings

- Metropolis [1953], Hastings [1970], Manhattan project: taking expectations on correlated samples
- The most Markov chains used in statistics are constructed using reversible Markov transition kernels.
- The kernel based on the proposals and rejections with the test

$$\alpha(x, y) = \min\left(1, \frac{p(y)q(y|x)}{p(x)q(x|y)}\right)$$

Limitation

We need to have the non-normalized densities both for the **proposal** and the **target** distributions.

Cruel reality

- The distributions of the interest at the ML are empirical
- The most powerful generative models are implicit (VAE, GAN)

The Implicit Metropolis-Hastings Algorithm

Motivation

Given:

- (implicit) generative model $q(x, y)$ (VAE, GAN, .etc)
- empirical data distribution $p_e(x)$ (a target distribution represented by a set of samples)

We would like to:

Introduce the Markov chain based on the proposal $q(x, y)$, with bounded distance between its stationary distribution and $p_e(x)$

As result we obtain:

1. Loss functions to upper bound on the distance between the target distribution and the stationary distribution of the proposed chain
2. Empirical validation the obtained theoretical result on real-world datasets (CIFAR-10, CelebA)
3. We also demonstrate empirical gains by applying our algorithm for Markov proposals.

Proposition

Proposed kernel transition kernel of the Implicit Metropolis-Hastings algorithm:

$$t(x|y) = q(x|y) \min\left(1, \frac{d(x,y)}{d(y,x)}\right) + \delta(x-y) \int dx' q(x' \mid y) \left(1 - \min\left(1, \frac{d(x',y)}{d(y,x')}\right)\right).$$

We want to get the answer for questions:

- Does the kernel is the valid markov kernel?
- Does the distribution t_∞ exists?
- How we should choose (train) the test function to bound $\|t_\infty - p\|_{TV} = \frac{1}{2} \int |t_\infty(x) - p(x)| dx$ for given proposal?

Assumptions

We require:

- The proposal distribution $q(x|y)$ and the discriminator $d(x, y)$ to be *continuous* and *positive* on $\mathbb{R}^D \times \mathbb{R}^D$.
- Limit the range of the discriminator as $d(x, y) \in [b, 1]\ \forall x, y$, where b is some positive constant that can be treated as a hyperparameter of the algorithm.
- The *minorization condition* (Robertset al., 2004) the proposal $q(x \mid y)$ to satisfy minorization condition with some constant and distribution ν (note that for an independent proposal, the minorization condition holds automatically with $= 1$).

Minorization condition:

$$q(x|y) > \varepsilon\nu(x) \quad \forall (x, y) \in \mathbb{R}^D \times \mathbb{R}^D.$$

Minorization condition is the interesting one.

Intuition for the Minorization condition

First, let's note that with minorization condition for proposal, we also have one for the whole transitional kernel:

$$t(x|y) \geq bq(x|y) > b\varepsilon\nu(x) = \varepsilon\nu(x)$$

Hence, we could consider the following **residual kernel** (easy to see that it is valid kernel iff t(x|y) is the valid kernel):

$$r(x|y) = \frac{t(x|y) - \varepsilon\nu(x)}{1 - \varepsilon}$$

Therefore, we could consider our **markov kernel** as the mixture of the **independent sampler** and residual markov kernel:

$$t(x|y) = \underbrace{\varepsilon q(x)}_{\text{Independent}} + \underbrace{(1-\varepsilon) r(x|y)}_{\text{Markov Guy}}.$$

- The Independent Guy makes chain "good": ν-recurrent, irreducible
- We could go further and restrict our minorization condition to be valid only on the some set. But then we need to introduce and formalize the "drift" of the chain to that set. Police rigor, the (Roberts, 2004) waits for your with more formal analysis.

Bound, the First of Its Name

Proposition 1

Consider a transition kernel $t(x|y)$ that satisfies the minorization condition $t(x|y) > \nu(x)$ for some > 0, and distribution ν. Then the distance between two consequent steps decreases as:

$$\|t_{n+2} - t_{n+1}\|_{TV} \leq (1 - \varepsilon)\|t_{n+1} - t_n\|_{TV},$$

where distribution $t_{k+1}(x) = \int t(x|y)t_k(y)dy$.

Now we can upper bound the TV-distance between an initial distribution t_0 and the stationary distribution t_∞ of the Implicit Metropolis-Hastings.

$$\|t_\infty - t_0\|_{TV} \leq \sum_{i=0}^{\infty} \|t_{i+1} - t_i\|_{TV} \leq \sum_{i=0}^{\infty} (1 - b)^i \|t_1 - t_0\|_{TV} = \frac{1}{b}\|t_1 - t_0\|_{TV}$$

Also we could start at the sample from the target distribution of our chain $t(x|y)$:

$$\|t_\infty - p\|_{TV} \leq \frac{1}{b}\|t_1 - p\|_{TV} = \frac{1}{2b}\int dx \left| \int t(x|y)p(y)dy - p(x) \right|.$$

Bound, the Bound of the Bound

Proposition 2

For the kernel $t(x \mid y)$ of the implicit Metropolis-Hastings algorithm, the distance between initial distribution $p(x)$ and the distribution $t_1(x)$ has the following upper bound

$$\|t_1 - p\|_{TV} \leq 2 \left\| q(y \mid x)p(x) - q(x \mid y)p(y)\frac{d(x,y)}{d(y,x)} \right\|_{TV},$$

$$\left\| q(y \mid x)p(x) - q(x \mid y)p(y)\frac{d(x,y)}{d(y,x)} \right\|_{TV} = \frac{1}{2}\int p(y)q(x \mid y)\left| \frac{q(y \mid x)p(x)}{q(x \mid y)p(y)} - \frac{d(x,y)}{d(y,x)} \right| dxdy$$

Also, the same as:

The thing to keep in mind

$\int |p(x)q(y|x)\alpha(x,y) - p(y)q(x|y)\alpha(y,x)|dxdy$

So, we force to satisfy the "detailed balance" condition.

Bound, the Third of Its Name (and 4th also)

Still, we have the ratio of the densities. Hence, we used the Pinsker inequality:

Proposition 3

For a distribution $\alpha(x)$ and some positive function $f(x) > 0\ \forall x$ the following inequality holds:

$$\|\alpha - f\|_{TV}^2 \leq \left(\frac{2C_f + 1}{6}\right)(\widehat{\mathrm{KL}}(\alpha\|f) + C_f - 1),$$

where C_f is the normalization constant of function f: $C_f = \int f(x)dx$, and $\widehat{\mathrm{KL}}(\alpha\|f)$ is the formal evaluation of the KL divergence

$$\widehat{\mathrm{KL}}(\alpha\|f) = \int \alpha(x) \log \frac{\alpha(x)}{f(x)} dx.$$

Let's apply it to the our bound:

$$\left\| q(y \mid x)p(x) - q(x \mid y)p(y)\frac{d(x,y)}{d(y,x)} \right\|_{TV} = \frac{1}{2}\int p(y)q(x \mid y)\left|\frac{q(y \mid x)p(x)}{q(x \mid y)p(y)} - \frac{d(x,y)}{d(y,x)}\right| dxdy$$

Bound, the Third of Its Name (and 4th also)

Applying the Pinsker:

$$\left\| q(y \mid x)p(x) - q(x \mid y)p(y)\frac{d(x,y)}{d(y,x)} \right\|^2 \leq \frac{2C+1}{6}\left(\widehat{\mathrm{KL}}\left(q(y \mid x)p(x) \middle| q(x \mid y)p(y)\frac{d(x,y)}{d(y,x)} \right) + C - 1 \right).$$

Here C is the normalization constant of $q(x \mid y)p(y)\frac{d(x,y)}{d(y,x)}$. As we bound our $\frac{d(x,y)}{d(y,x)} \in [b, \frac{1}{b}]$, we also have the upper bound on the C as $\frac{1}{b}$.

The Loss:

$$\|t_\infty - p\|_{TV}^2 \leq \frac{1}{b^2\varepsilon^2}\|t_1 - p\|_{TV}^2 \leq \frac{4}{b^2\varepsilon^2}\left\| q(y \mid x)p(x) - q(x \mid y)p(y)\frac{d(x,y)}{d(y,x)} \right\|_{TV}^2 \leq$$

$$\leq \left(\frac{4+2b}{3\varepsilon^2 b^3}\right)\Bigg(\underbrace{\mathbb{E}_{\substack{x \sim p(x) \\ y \sim q(y \mid x)}}\left[\log\frac{d(y,x)}{d(x,y)} + \frac{d(y,x)}{d(x,y)} \right]}_{\text{loss for the discriminator}} - 1 + \mathrm{KL}\left(q(y \mid x)p(x) \middle\| q(x \mid y)p(y) \right) \Bigg).$$

Minimization of the resulting upper bound w.r.t. the discriminator $d(x, y)$ is equivalent to the following optimization problem:

$$\min_d \mathbb{E}_{\substack{x \sim p(x) \\ y \sim q(y \mid x)}}\left[\log\frac{d(y,x)}{d(x,y)} + \frac{d(y,x)}{d(x,y)} \right].$$

The Loss

$$\min_d \mathbb{E}_{\substack{x \sim p(x) \\ y \sim q(y \mid x)}} \left[\log \frac{d(y,x)}{d(x,y)} + \frac{d(y,x)}{d(x,y)} \right].$$

We could take the point-wise gradient $d(x, y)$ in a single point (x, y), and find out optimal d.

$$\nabla_{d(x,y)} \left(p(x)q(y \mid x) \left[\log \frac{d(y,x)}{d(x,y)} + \frac{d(y,x)}{d(x,y)} \right] + p(y)q(x \mid y) \left[\log \frac{d(x,y)}{d(y,x)} + \frac{d(x,y)}{d(y,x)} \right] \right) = 0$$

% some equations% and optimal answer:

$$\frac{p(x)q(y \mid x)}{p(y)q(x \mid y)} = \frac{d(x,y)}{d(y,x)}$$

What is nice:

- It is the DRE, and the original TV is 0 at the optimal point.
- The "discriminator" takes 2 samples.

Question: relation to the cross entropy?

Relation to the cross entropy (and more upper bounds)

It is possible to upper bound the loss by the binary cross-entropy. For a Markov proposal, it is

$$\mathbb{E}_{\substack{x \sim p(x) \\ y \sim q(y \mid x)}} \left[\log \frac{d(y,x)}{d(x,y)} + \frac{d(y,x)}{d(x,y)} \right] \leq \mathbb{E}_{\substack{x \sim p(x) \\ y \sim q(y \mid x)}} \left[-\log d(x,y) - \log(1 - d(y,x)) + \frac{1}{b} \right].$$

In the case of an independent proposal, we factorize the discriminator as $d(x,y) = d(x)(1 - d(y))$ and obtain the following inequality

$$\mathbb{E}_{\substack{x \sim p(x) \\ y \sim q(y \mid x)}} \left[\log \frac{d(y,x)}{d(x,y)} + \frac{d(y,x)}{d(x,y)} \right] \leq -\mathbb{E}_{x \sim p(x)} \log d(x) - \mathbb{E}_{y \sim q(y)} \log(1 - d(y)) + \frac{1}{b}$$

Thus, learning a discriminator via the binary cross-entropy, we also minimize the distance $\|t_\infty - p\|_{TV}$.

The Algorithm

Algorithm 1
The implicit Metropolis-Hastings algorithm

Input: target dataset $\mathcal{D}$
Input: implicit model $q(x \mid y)$
Input: learned discriminator $d(\cdot, \cdot)$
$y \sim \mathcal{D}$ initialize from the dataset
for $i = 0 \ldots n$ **do**
sample proposal point $x \sim q(x \mid y)$
$P = \min\{1, \frac{d(x,y)}{d(y,x)}\}$

$$x_i = \begin{cases} x, & \text{with probability } P \\ y, & \text{with probability } (1 - P) \end{cases}$$

$y \leftarrow x_i$
end for $\{x_0, \ldots, x_n\}$

Experiments

General Set-up

For both independent and markov proposal, we have the following routine Pipe-line:

1. Given some trained generator (WPGAN, VAE), we train the discriminator (a bit)
2. Then we perform The Implicit MH sampling
3. Stop on the reasonable empirical Acceptance Rate

Data/Metrics:

1. Datasets: CIFAR10, CelebA
2. Metrics: Inception score, FID

$$IS(G) = \exp(E_g D_{KL}(p(y|x)||E_g p(y|x))$$

$$FID(p_e, G) = \|\mu_g - \mu_{p_e}\|_2^2 + \mathrm{Tr}\left(\Sigma_{p_e} + \Sigma_g - 2(\Sigma_{p_e}\Sigma_g)^{0.5}\right)$$

Independent proposal: WPGAN

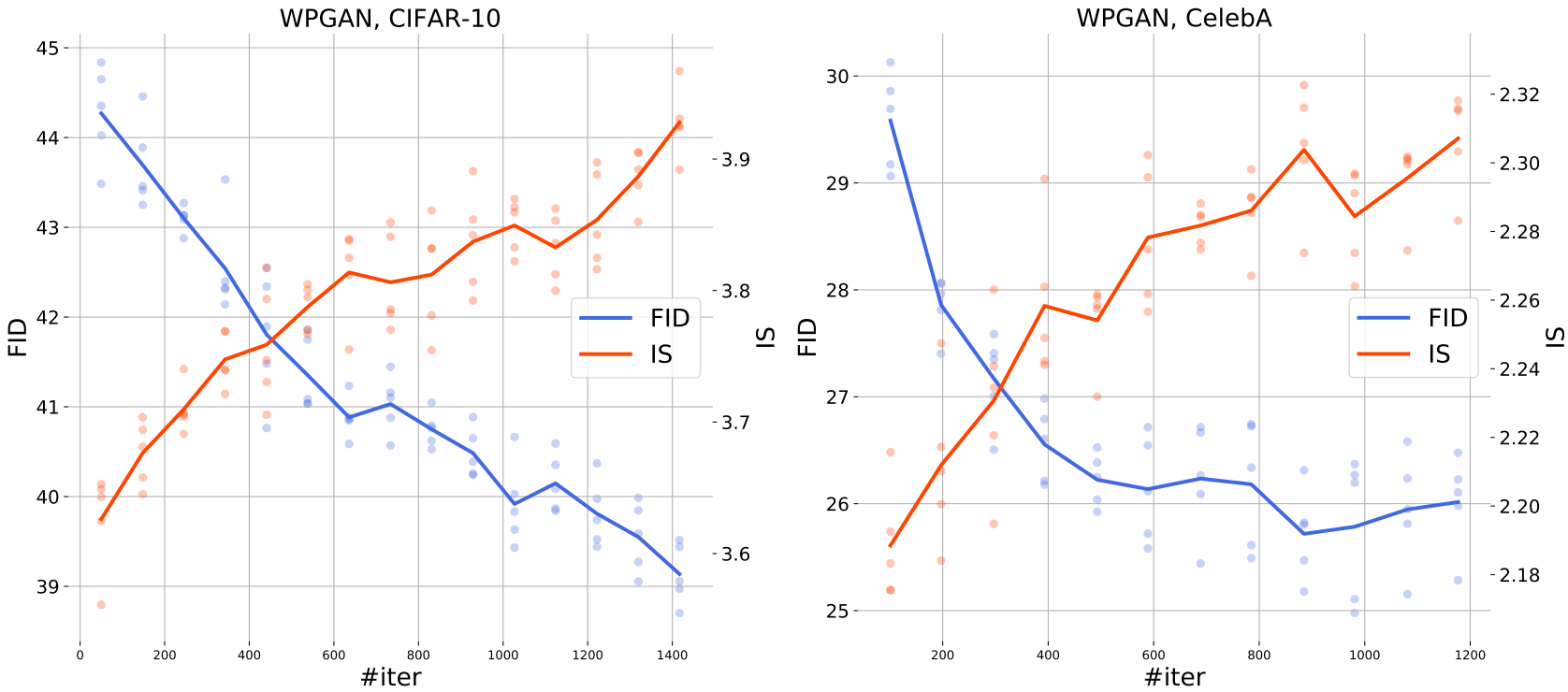

Figure: Monotonous improvements in terms of FID and IS for the learning of discriminator by CCE. During iterations, we evaluate metrics 5 times (scatter) and then average them (solid lines). For a single metric evaluation, we use 10k samples. Higher values of IS and lower values of FID are better. Performance for the original generator corresponds to 0th iteration of a discriminator.

Independent proposal: VAE

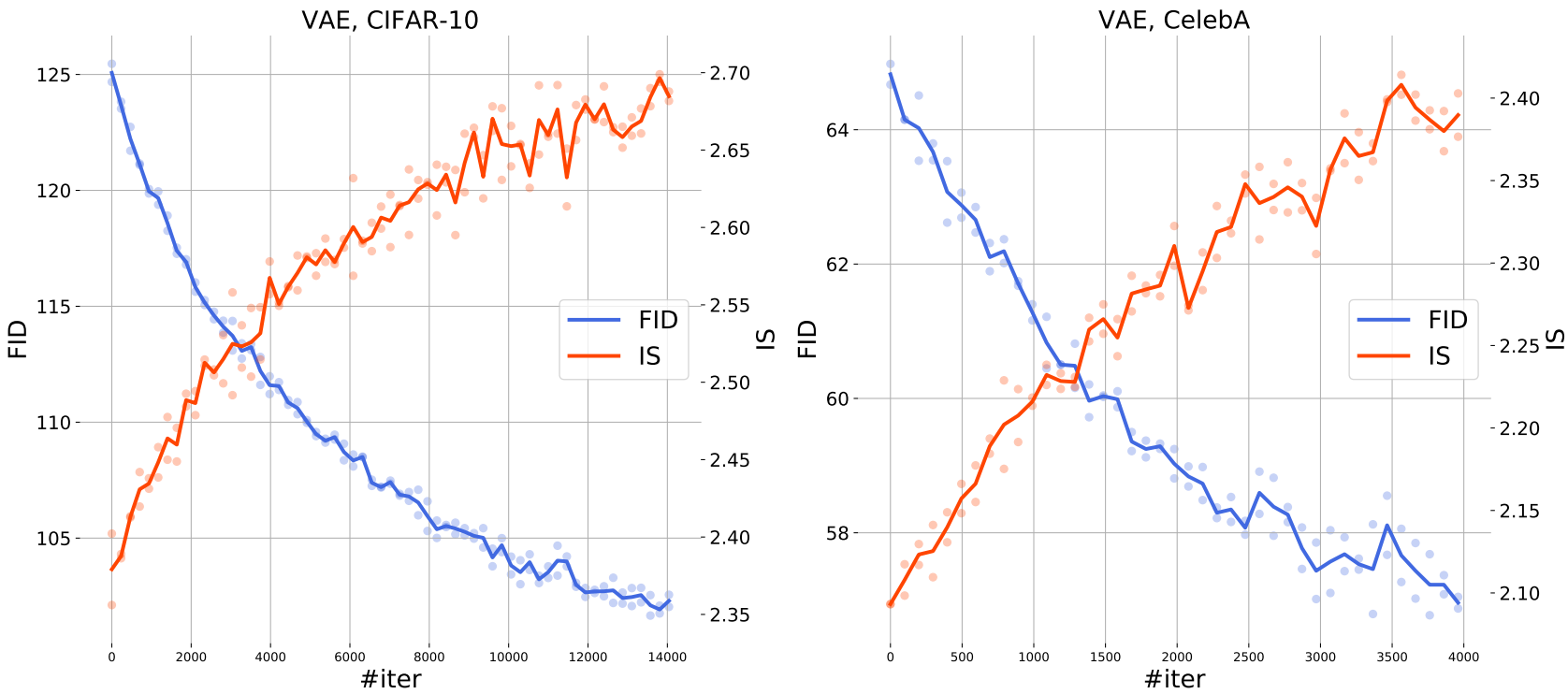

Figure: Monotonous improvements in terms of FID and IS for the learning of discriminator by CCE. During iterations, we evaluate metrics 5 times (scatter) and then average them (solid lines). For a single metric evaluation, we use 10k samples. Higher values of IS and lower values of FID are better. Performance for the original generator corresponds to 0th iteration of a discriminator.

Markov proposal

Proposal Markovisation:

To simulate Markov proposals we take the same WPGAN as in the independent case and traverse its latent space by a Markov chain.

$$z_x = \cos(t) z_y + \sin(t) v, \quad v \sim \mathcal{N}(0, I).$$

Loss estimation require samples from the dataset $x \sim q(x \mid y), y \sim p(y)$. To sample an image $x \sim q(x \mid y)$ we need to know the latent vector z_y for an image y from the dataset. We find such vectors by optimization in the latent space.

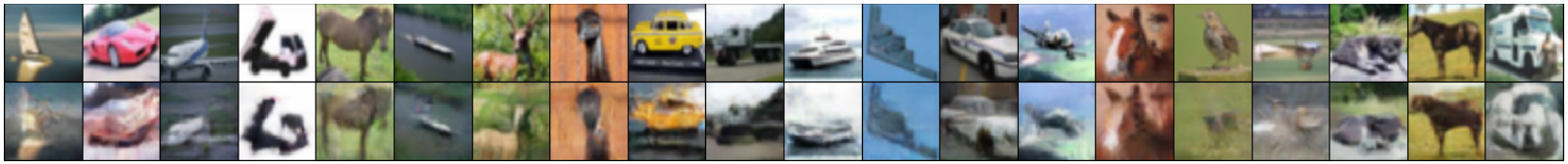

Figure: Samples from CIFAR-10 (top line) and their reconstructions (bottom line)

Discriminator:

$$d(x, y) = \frac{1}{1 + \exp(\text{net}(y) - \text{net}(x))}$$

Markov proposal: WPGAN

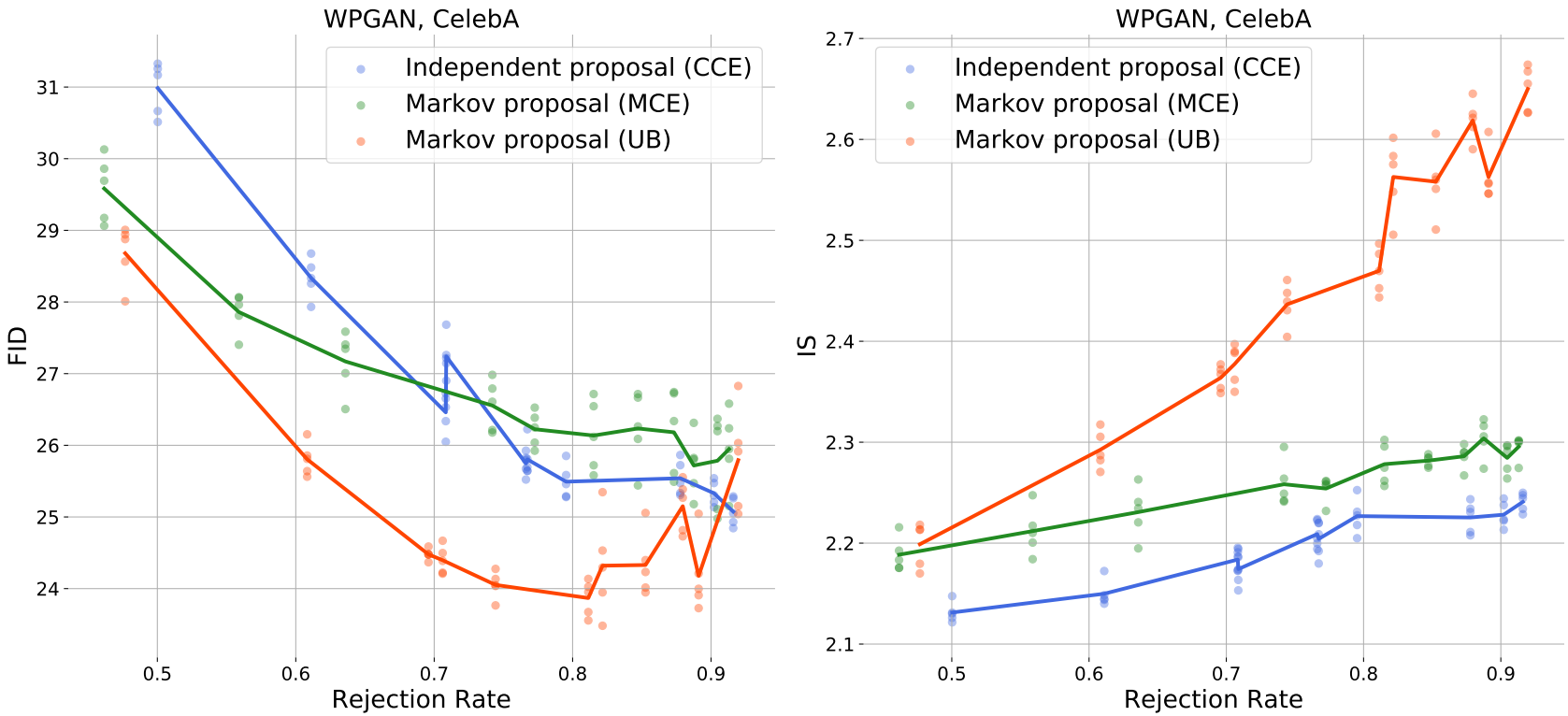

Figure: Monotonous improvements in terms of FID and IS

Markov proposal: WPGAN

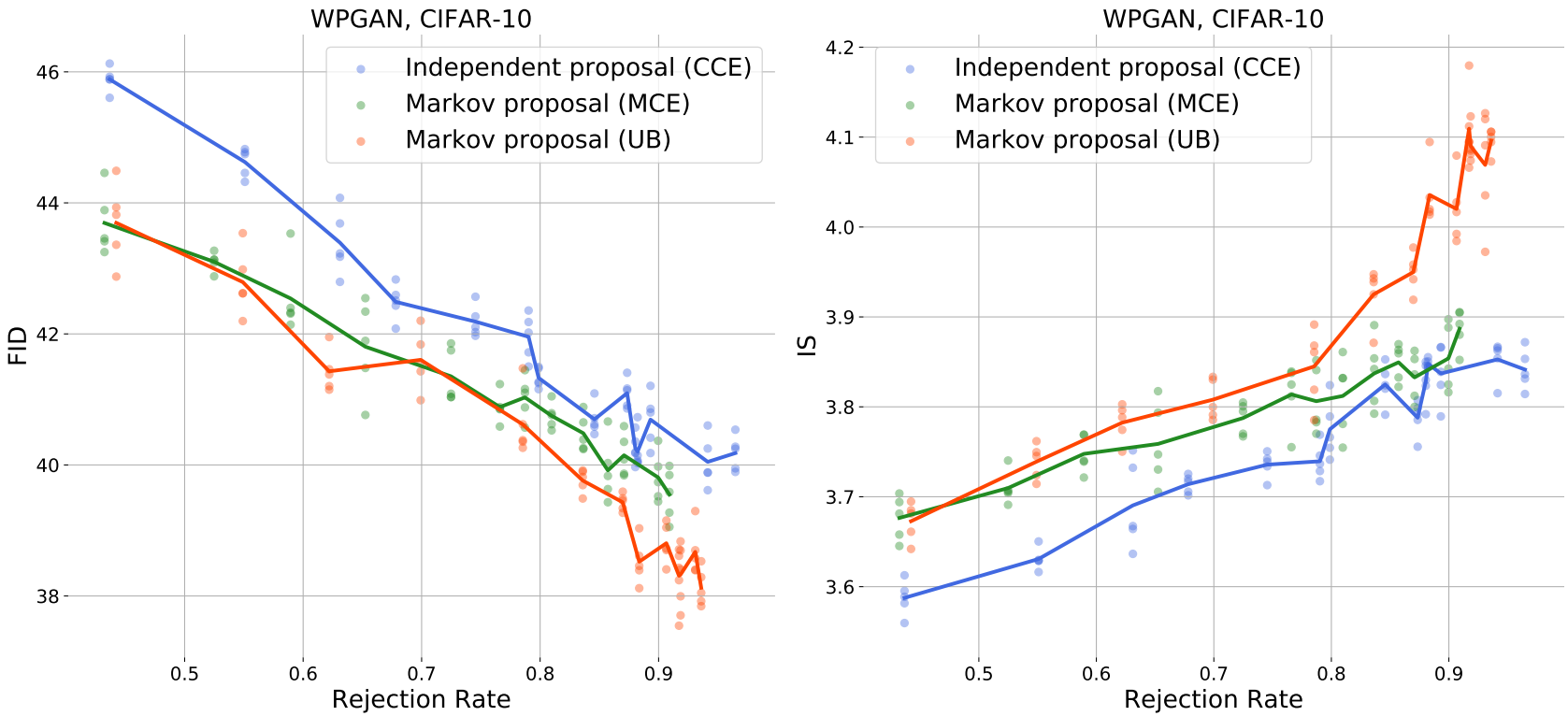

Figure: Monotonous improvements in terms of FID and IS

Questions

?Questions, Comments, Remarks?