Non Equilibrium Sampling on the Orbit of a Deterministic Transform: NEO estimator

Achille Thin[1], Yazid Janati[2], Sylvain Le Corff[2], Charles Ollion[1], Arnaud Doucet[3], Alain Durmus[4], Eric Moulines[1], Christian Robert[5]

[1] Ecole Polytechnique, [2] Télécom SudParis, [3] Oxford University, [4] ENS Paris-Saclay, [5] Université Paris Dauphine

December 2, 2021

Summary

- Setting: a target distribution

$$\pi(x) = \mathrm{L}(x)\rho(x)/\,\mathrm{Z}$$

  with unknown normalizing constant Z.
  - Bayesian setting: ρ is the prior distribution, L is the likelihood.
  - Generative Adversarial Networks: ρ is the generator and $\mathrm{L}(x)$ is derived from discriminator.
  - Variational autoencoders: here $\rho(x)\mathrm{L}(x) = p_\theta(x, y)$, $\mathrm{Z} = p_\theta(y)$ and we need to compute a lower bound of $\log(\mathrm{Z})$.
- Objectives:
  - Estimate the normalizing constant $\mathrm{Z} = \int \mathrm{L}(x)\rho(x)\mathrm{d}x$
  - Sample from π.

Basic idea

- An Importance Sampling (IS) estimator of $\mathrm{Z} = \int \mathrm{L}(x)\rho(\mathrm{d}x)$ is

$$\hat{\mathrm{Z}} = N^{-1} \sum_{i=1}^{N} \mathrm{L}(X^i)\,, \qquad (X^i)_{1 \le i \le N} \overset{\text{iid}}{\sim} \rho\,.$$

  $\hat{\mathrm{Z}}$ is unbiased:

$$\mathbb{E}_{X^i \overset{\text{iid}}{\sim} \rho}[\widehat{\mathrm{Z}}^{\varpi}] = \mathrm{Z}\,.$$

- Idea: Build estimator refining proposal distribution ρ using transport map T.
- T does not necessarily leaves π or ρ invariant (hence the name non-equilibrium).

Construction

- Define ρ_k the pushforward of ρ by T^k:

$$\rho_k(x) = \rho(\mathrm{T}^{-k}(x))\mathbf{J}_{\mathrm{T}^{-k}}(x)$$

- For any probability distribution $(\varpi_k)_{k\in\mathbb{Z}}$, define:

$$\rho_{\mathrm{T}}(x) = \sum_{k\in\mathbb{Z}} \varpi_k \rho_k(x) \ .$$

  In practice, we choose $\varpi_k \propto \mathbb{1}_{[0,K]}(k)$ for some $K \in \mathbb{N}^*$.

- Key identity: For any nonnegative function f,

$$\int f(x)\rho(x)\mathrm{d}x = \int f(x)\frac{\rho(x)}{\rho_{\mathrm{T}}(x)}\rho_{\mathrm{T}}(x)\mathrm{d}x = \int \left(\sum_{k\in\mathbb{Z}} f(\mathrm{T}^k(x))w_k(x)\right)\rho(x)\mathrm{d}x$$

  where $w_k(x)$ are the 'importance' weights

$$w_k(x) = \varpi_k \frac{\rho(\mathrm{T}^k(x))}{\rho_{\mathrm{T}}(\mathrm{T}^k(x))} = \frac{\varpi_k \rho_{-k}(x)}{\sum_{j\in\mathbb{Z}} \varpi_j \rho_{j-k}(x)} \ .$$

NEO estimator

- For $f(x) = \mathrm{L}(x)$, we obtain the NEO estimator of the normalizing constant Z:

$$\widehat{\mathrm{Z}}^{\varpi}_{X^{1:N}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k \in \mathbb{Z}} w_k(X^i) \mathrm{L}(\mathrm{T}^k(X^i)) \,, \quad X^i \overset{\text{iid}}{\sim} \rho$$

- NEO estimators are unbiased: $\mathbb{E}_{X^{1:N} \overset{\text{iid}}{\sim} \rho}[\widehat{\mathrm{Z}}^{\varpi}_{X^{1:N}}] = \mathrm{Z}$.
- Algorithm:
  1. Sample $X^{1:N} \overset{\text{iid}}{\sim} \rho$ for $i \in [N]$.
  2. For $i \in [N]$, compute the path $(\mathrm{T}^j(X^i))_{j \in \mathbb{Z}}$ and weights $(w_j(X^i))_{j \in \mathbb{Z}}$.
  3. $I^{\mathrm{NEO}}_{\varpi,N}(f) = N^{-1} \sum_{i=1}^{N} \sum_{k \in \mathbb{Z}} w_k(X^i) f(\mathrm{T}^k(X^i))$.

Self Normalized Importance Sampling

- Build the self normalized importance sampling estimator of $\pi(g) = \mathbb{E}_\pi[g(X)] = \int \pi(x)g(x)\mathrm{d}x$:

$$J^{\mathrm{NEO}}_{\varpi,N}(f) = N^{-1}\sum_{i=1}^{N} \frac{\widehat{\mathrm{Z}}^{\varpi}_{X^i}}{\widehat{\mathrm{Z}}^{\varpi}_{X^{1:N}}} \sum_{k\in\mathbb{Z}} \frac{\mathrm{L}(\mathrm{T}^k(X^i))w_k(X^i)}{\widehat{\mathrm{Z}}^{\varpi}_{X^i}} f(\mathrm{T}^k(X^i)) \ .$$

- Non asymptotic properties of the estimator: Define

$$E^{\varpi}_{\mathrm{T}} = \mathbb{E}_{X\sim\rho}\left[\left(\textstyle\sum_{k\in\mathbb{Z}} w_k(X)\mathrm{L}(\mathrm{T}^k(X))/\,\mathrm{Z}\right)^2\right] .$$

- Assume that $E^{\varpi}_{\mathrm{T}} < \infty$. Then, for any function g satisfying $\sup_{x\in\mathbb{R}^d}|g(x)| \leq 1$ on $\mathbb{R}^d$, and $N \in \mathbb{N}$

$$\mathbb{E}_{X^{1:N}\overset{\mathrm{iid}}{\sim}\rho}\left[\left|J^{\mathrm{NEO}}_{\varpi,N}(g) - \pi(g)\right|^2\right] \leq 4\cdot N^{-1}E^{\varpi}_{\mathrm{T}} \ ,$$
$$\left|\mathbb{E}_{X^{1:N}\overset{\mathrm{iid}}{\sim}\rho}\left[J^{\mathrm{NEO}}_{\varpi,N}(g) - \pi(g)\right]\right| \leq 2\cdot N^{-1}E^{\varpi}_{\mathrm{T}} \ .$$

Choice of T

Lemma

For any nonnegative sequence $(\varpi_k)_{k\in\mathbb{Z}}$, we have

$$E_{\mathrm{T}}^{\varpi} \leq \mathrm{D}_{\chi^2}(\pi\|\rho_{\mathrm{T}}) + 1 \,.$$

- Desired properties
  1. T should drive samples from ρ to the sets where L is large.
  2. The Jacobian of T should be easy to compute
- Many normalizing flows have been introduced recently in the literature. But learning the flows introduces an additional layer of complexity
- Idea: Use a conformal Hamiltonian transform (linked with the momentum estimator).

Conformal Hamiltonian systems

- Define
  - Potential energy: $U(q) = \log[\mathrm{L}(q)\rho(q)]$, where q is the position
  - Kinetic energy: $K(p) = p^T \mathrm{M}^{-1} p/2$, where p is the momentum, M is the mass matrix
  - Hamiltonian (total energy) of the system $H(q,p) = U(q) + K(p)$.
- Transformation $\mathrm{T}(q_0, p_0) = (q_1, p_1)$ where

$$p_1 = \mathrm{e}^{-h\gamma} p_0 - h\nabla U(q_0)\ , \quad q_1 = q_0 + h p_1\ .$$

- Euler discretization of conformal Hamiltonian dynamics

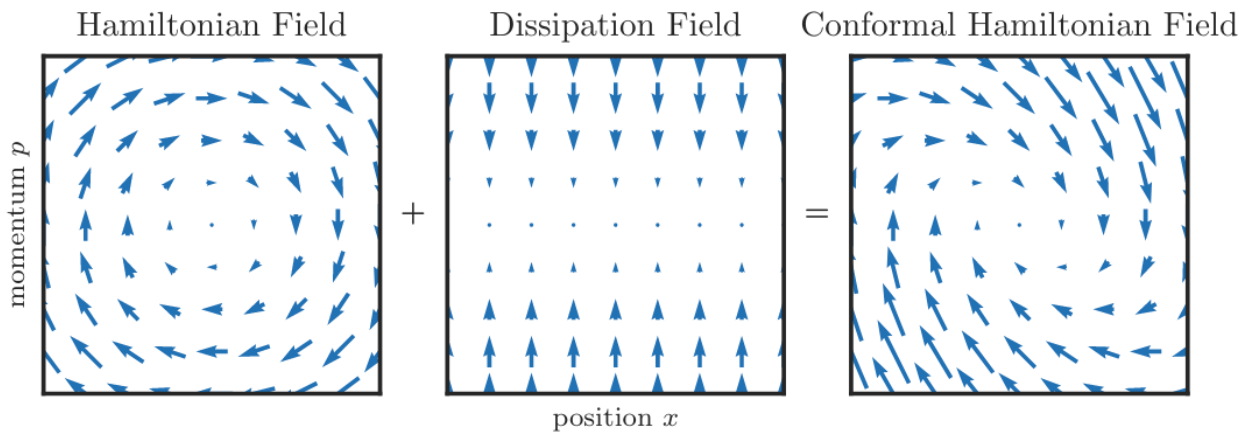

Figure: Vector field of a conformal Hamiltonian system.

Conformal Hamiltonian vector field

Compare $E_{\mathrm{T}}^{\varpi}(K)$ for $\varpi_k \propto \mathbb{1}_{[0,K]}(k)$ to $E^{\mathrm{IS}}(K) - 1 = (K+1)^{-1}\mathbb{E}_{X\sim\rho}[\mathrm{L}(X)^2]$ the IS equivalent.

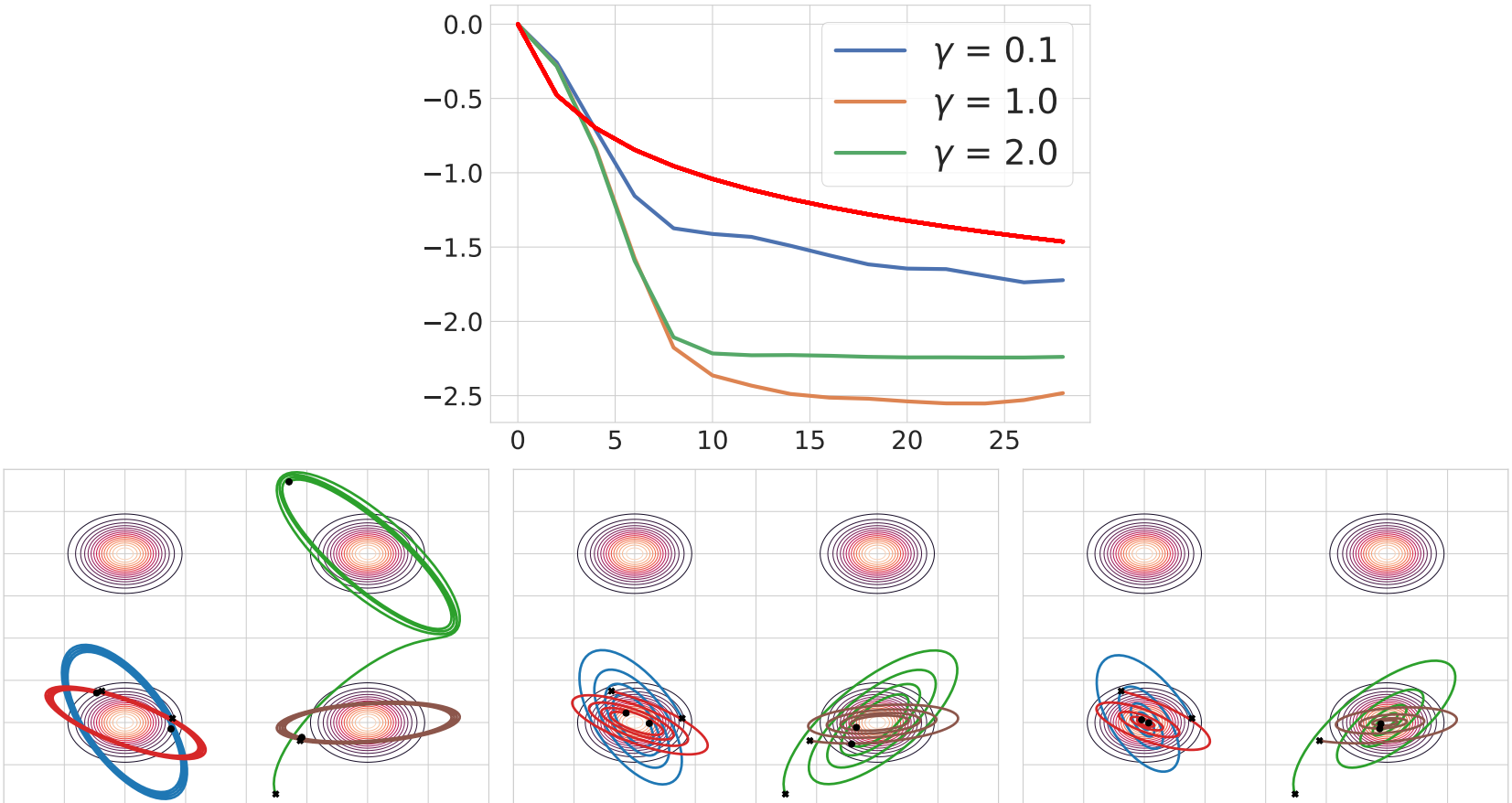

Figure: Top: $E_{\mathrm{T}_h}^{\mathbb{1}_{[K]}}(K)$ vs $E^{\mathrm{IS}}(K)$ (red) in $\log_{10}$-scale as a function of optimization step K. Bottom, left to right: Corresponding orbits for $\gamma = 0.1, 1, 2$.

Sampling Importance Resampling (SIR)

- Algorithm
  1. Draw $X^{1:N} \overset{\text{iid}}{\sim} \rho$,
  2. Draw $I^* \sim \mathrm{Cat}(\{\tilde{\mathrm{L}}(X^i)\}_{i=1}^N)$ and set $X^* = X^{I^*}$.
- Main result When $N \to \infty$ the distribution of X^* converges weakly to π.
- Caveats: the number N of proposals typically grows exponentially with the dimension d to maintain a given accuracy.

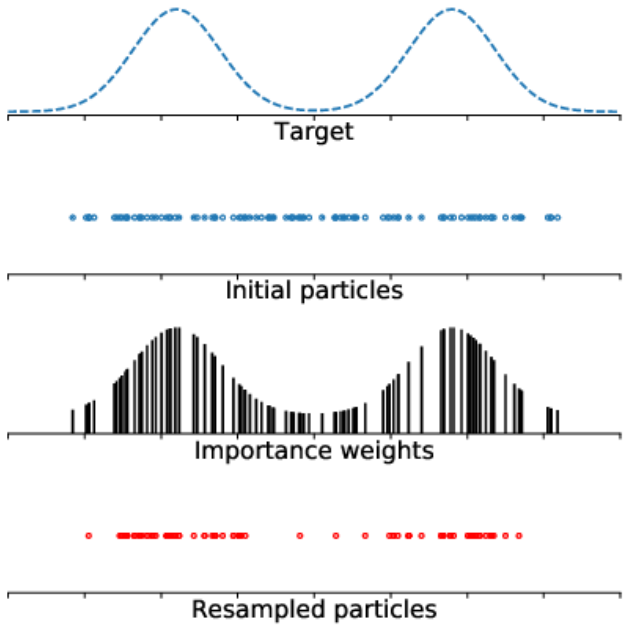

Figure: SIR scheme

NEO-MCMC sampler

At step $n \in \mathbb{N}^*$, given the conditioning orbit point Y_{n-1}.

Step 1: Update the conditioning point

1. Set $X_n^1 = Y_{n-1}$ and for any $i \in \{2, \dots, N\}$, sample $X_n^i \overset{\text{iid}}{\sim} \rho$.
2. Sample the orbit index I_n with probability proportional to $(\widehat{\mathrm{Z}}^{\varpi}_{X_n^i})_{i \in [N]}$.
3. Set $Y_n = X_n^{I_n}$.

Step 2: Output a sample

1. Sample index K_n with probability proportional to $\{w_k(Y_n)\mathrm{L}(\mathrm{T}^k(Y_n))/\widehat{\mathrm{Z}}^{\varpi}_{Y_n}\}_{k \in \mathbb{Z}}$
2. Output $U_n = \mathrm{T}^{K_n}(Y_n)$.
   Geometric ergodicity and invariance results can be achieved on this sampler !

Estimation of normalizing constant

- Consider now the 25 Gaussian distribution `MG25` mixture of 25 d-dimensional Gaussian distributions in dimension $d = 10, 20, 45$.
- Parameters
  - Diagonal covariances with diagonal elements equal to $(0.01, 0.01, 0.1, \ldots, 0.1)$.
  - Means given by $(i, j, 0, \ldots, 0)$ with $i, j \in \{-2, \ldots, 2\}$
  - Consider also the complex Funnel distribution `(Fun)` $\pi(x) = \mathrm{N}(x_1; 0, a^2) \prod_{i=1}^{d} \mathrm{N}(x_i; 0, \mathrm{e}^{2bx_1})$ with $d \in \{10, 20, 45\}$, $a = 1$, and $b = 0.5$.
  - Proposal is $\rho = \mathrm{N}(0, \sigma_\rho^2 \, \mathrm{Id}_d)$ with $\sigma_\rho^2 = 5$.

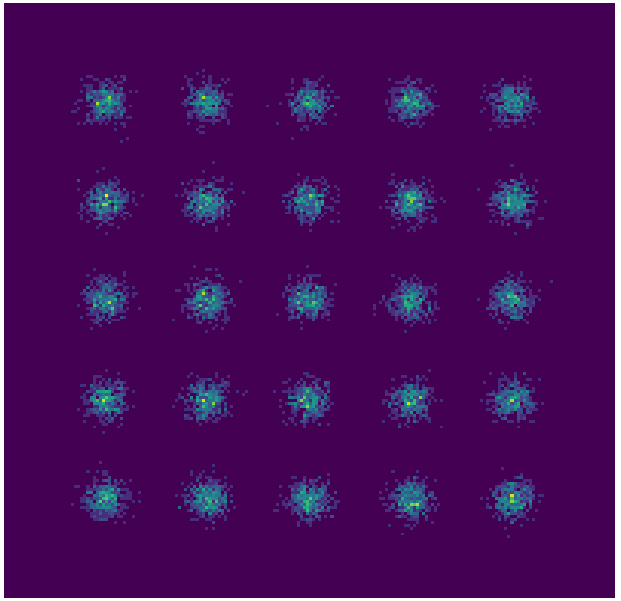

Figure: First two dimensions of the 25 Gaussian dataset.

Estimation of Normalizing constants

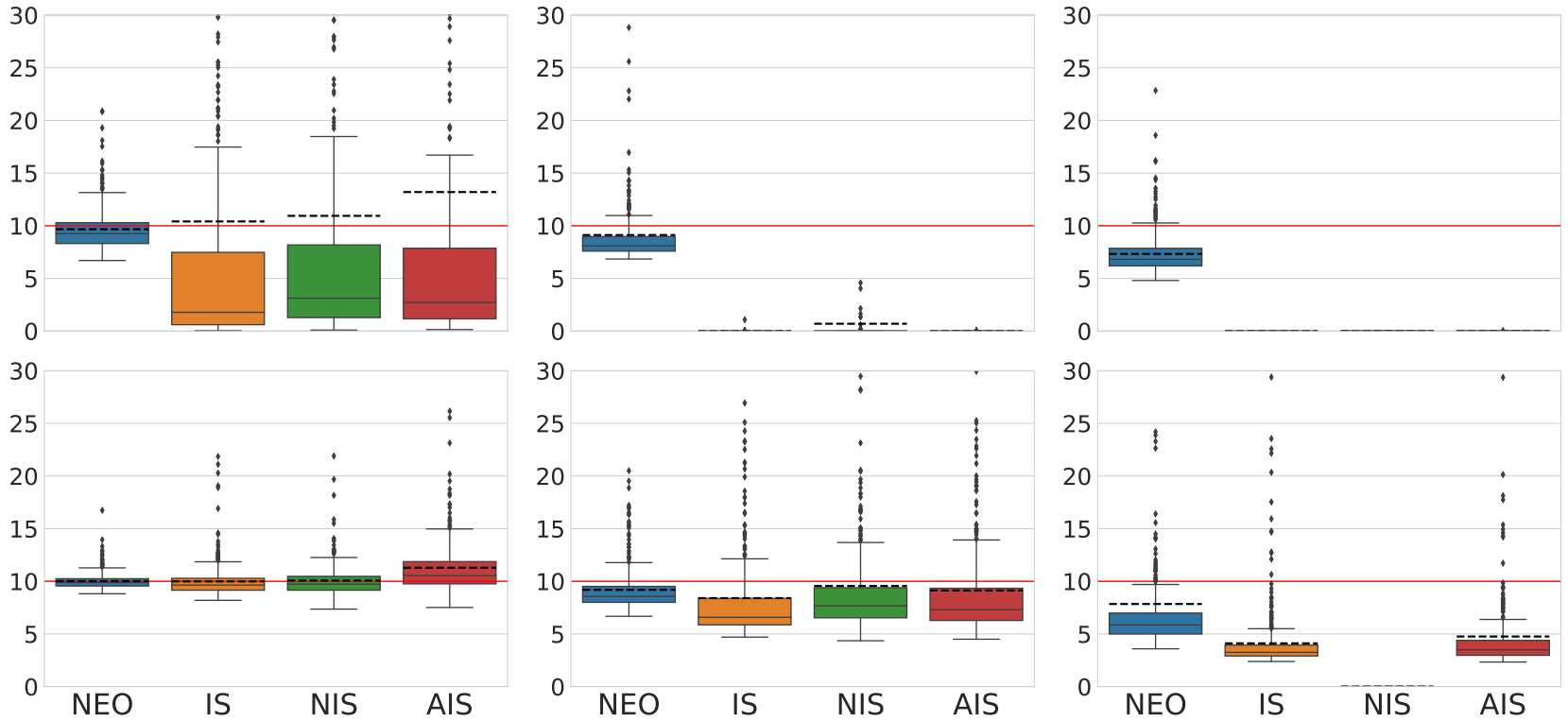

Figure: Boxplots of 500 independent estimations of the normalizing constant in dimension $d = \{10, 20, 45\}$ (from left to right) for `MG25` (top) and `Fun` (bottom). The true value is given by the red line. The figure displays the median (solid lines), the interquartile range, and the mean (dashed lines) over the 500 runs.

MCMC sampler results

- First focus on 25 Gaussians dataset, with $d = 40$.
- We compare NEO with Correlated ISIR and the No U-Turn Sampler (NUTS).
- All algorithms are run during the same computational time.

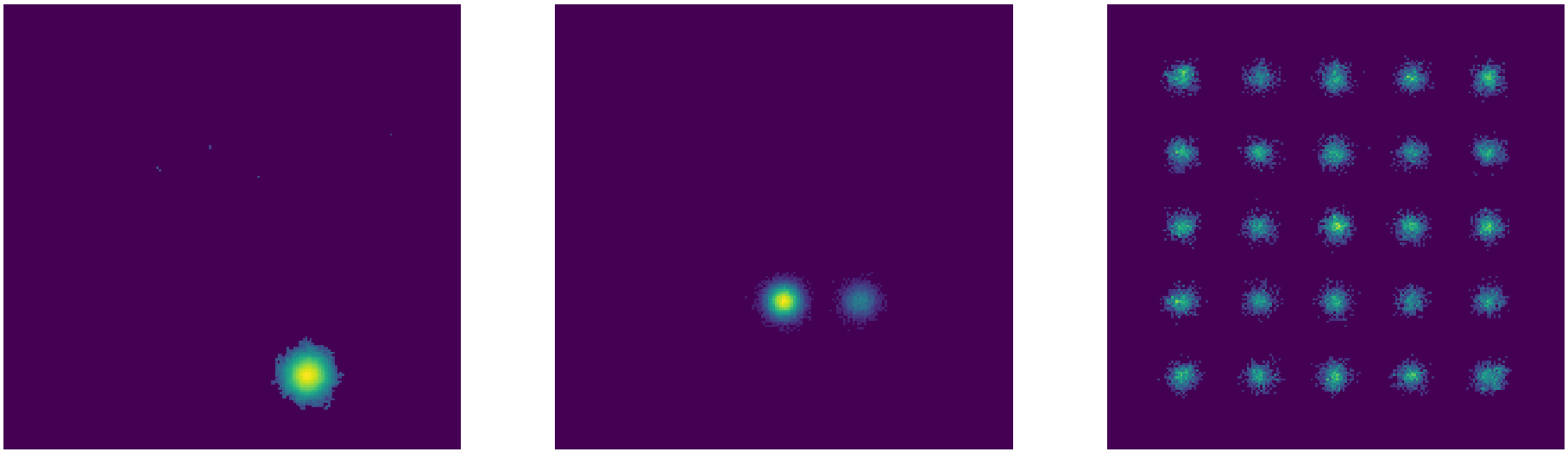

Figure: Histograms of the MCMC samples. From left to right, ISIR, NUTS and NEO.

Gibbs inpainting on images

- Consider a more striking examples to illustrate mixing time.
- Focus on some trained VAE on CelebA dataset (no optimization objective here)
- Given an image y, denote by $[y^t, y^b]$ the top and the bottom half pixels.
- Two-stage Gibbs sampler
  1. Sample $p_{\theta^*}(z|y^t, y^b)$
  2. Sample $p_{\theta^*}(y^b|z, y^t) = p_{\theta^*}(y^b|x)$.
- Perform MCMC-within-Gibbs for stage 1 using i-SIR, NEO-MCMC and HMC with same computational complexity.

Gibbs inpainting

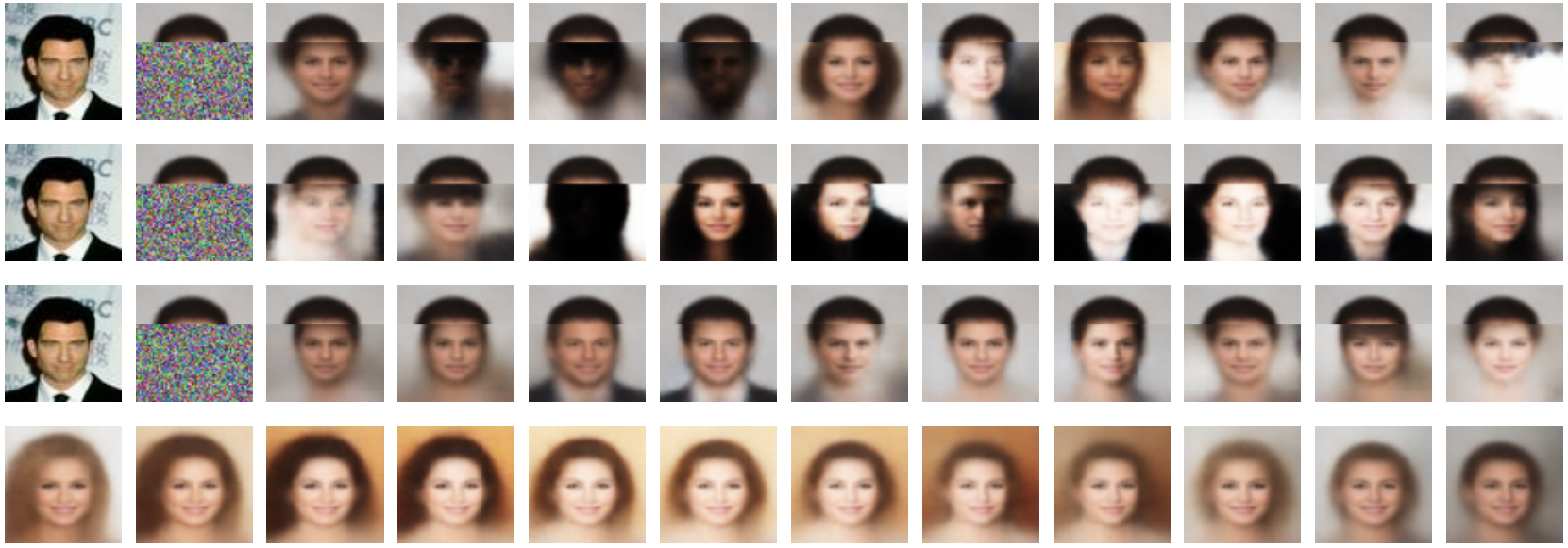

Figure: Gibbs inpainting for CelebA dataset. From top to bottom: i-SIR, HMC and NEO-MCMC: From left to right, original image, blurred image to reconstruct, and output every 5 iterations of the Markov chain. Last line: a forward orbit used in NEO-MCMC.

NEO-MCMC mixes faster than i-SIR and HMC ! Last line illustrates the effect of the trajectory.

Conclusion

- Powerful estimator building on optimization paths
- Extends easily to an efficient MCMC sampler and other types of deep generative models
- Difficulty however to find good parameters automatically
- We could consider other transformations (the framework easily extends to non homogeneous mappings, such as normalizing flows !)
- Natural connection to Nested Sampling in the approach. Could apply to microcanonical sampling.

Thank you for your attention !

Bibliography I