Generative Models : high-dimensional sampling revisited.

Eric Moulines
École Polytechnique, Département de Mathématiques Appliquées,
French Academy of Sciences, Deputy of the Computer Sciences and Mechanics Section,
HSE, HDI Lab - Stochastic Algorithms
Many co-authors: D. Belomestny (HSE), A. Durmus (ENS Paris-Saclay), Hoi To Wai (Chinese U. HK), A. Naumov (HSE), M. Panov (Skoltech), M. Rasonyi (Budapest), S. Sabanis (U. Edinburgh)
and a lot of current, past and future PhD Students, N. Brosse (E. Polytechnique), A. Thin (E. Polytechnique), Pablo Jimenez-Moreno (E. Polytechnique), L. Iosipoi (HSE), S. Samsonov (HSE)

Deep Learning / Probabilistic Reasoning

"classical" Deep Learning

+ Rich non-linear models for classification and sequence prediction.
+ Scalable learning using stochastic approximation and conceptually simple.
+ Easily composable and scalable.
- Only point estimates
- Hard to score models, implement model selection and complexity penalisation.

Probabilistic reasoning

+ Unified framework for model building, inference, prediction and decision making.
+ Explicit accounting for uncertainty and variability of outcomes.
+ Robust to overfitting; tools for model selection and composition.
- Potentially intractable inference, computationally expensive or long simulation time.
- Probabilistic models are often 'too simple' and do not have the discriminative power of DL

1 Probabilistic Reasoning with DL: three examples
- Deep Variational Autoencoders
- Uncertainty quantification
- Macrocanonical sampling

2 High-Dimensional Sampling

3 Variance reduction for Makov Chains

4 Bibliography

Move beyond associating inputs to outputs

- Improve supervised learning from few samples
  - Unlabeled data often abundant
  - Learn representations / concepts / features from unlabeled data
- A lot of applications: generate new patterns, learn from few examples, detect unusual behaviors, ...

Joint work with A. Doucet (Oxford), G. Fort (U. Toulouse), Hoi To Wai (CHUHK), M. Panov (NRU-HSE Moscow), with N. Kotolevski (Skoltech), A. Thin (Ecole Polytechnique and Oxford)

(Un)supervised learning and (un)conditional models

- **Supervised** learning: model **conditional distribution** $p_\theta(y|x)$
  - for example x is an image, y is a class label

$$\max_\theta \sum_{\mathsf{D}_N} \log p_\theta(Y_i|X_i)$$

  - $\mathsf{D}_N = \{(X_i, Y_i)\}_{i=1}^N$ is the training sequence (sample from the unknown data generating distribution)
  - θ is the model parameter
- **Unsupervised** learning: model **unconditional distribution** $p_\theta(x)$
  - For example x is an image and the parameters can be estimated by maximizing the likelihood

$$\max_\theta \sum_{\mathsf{D}_N} \log p_\theta(X_i)$$

  - Possible to draw from p_θ [generate new samples]

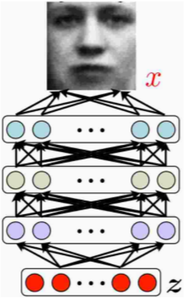
Deep Latent Models

Generative model: unconditional density model $p_\theta(x)$

- Deep Latent Gaussian Models (DLGM) are powerful generative models that can be effectively fit to very large datasets of complicated high-dimensional data Kingma and Welling, 2014; Rezende et al, 2014
- **Assumptions:** the observations are generated by
  - sampling some **latent variables**
  - feeding them into a deep neural network
  - adding some **structured noise** to the network output
- Non-linear extensions of Probabilistic Principal Component Analysis, Gaussian Linear State-Space models, Hidden Markov Models, etc... Roweis, 1997; Bishop, Tipping, 1999

Deep Latent Gaussian Models - DLGM

- Sample the latent variables $Z \in \mathbb{R}^p$ from a normal distribution
- Compute a vector-valued non-linear function g_θ, the **decoder**, typically a deep neural network with some parameters θ. The decoder **maps** latent **code** to the **observations**.
- Sample the observation $X \in \mathsf{X} \subset \mathbb{R}^d$ independently conditionally to $Z \in \mathbb{R}^p$ from a distribution parameterized by $g_\theta(Z)$



DLGM for MNIST

- **Observations:** binarized MNIST handwritten digit data set : $\mathsf{X} = \{0, 1\}^d$ with $d = 28^2 = 784$
- **Generative model:**

$$Z \sim \mathcal{N}(0, \mathrm{Id}_p)$$

$$X|Z \sim \prod_{j=1}^{d} \mathrm{Ber}(X^{(j)}, [g_\theta(Z)]^{(j)}) \quad X = (X^{(1)}, \ldots, X^{(d)})$$

- The pixels in the binarized image $X^{(j)}$ are conditionally independent.
- Each pixel $X^{(j)}$ is a Bernoulli random variable with a success probability given by the the j-th component of the **decoder** output.

Inference problem

- The DLGM model specifies the **likelihood** the observation x

$$p_\theta(x) = \int_{\mathbb{R}^p} p_\theta(x,z)\mathrm{d}z \quad \text{where} \quad p_\theta(x,z) = p_\theta(x|z)\phi(z)$$

- Given a **training** data set $X_1,\ldots,X_N$, a natural idea is to fit the parameter θ using a **maximum likelihood**

$$\hat{\theta} = \arg\max_{\theta\in\theta} N^{-1}\sum_{n=1}^{N} \log p_\theta(X_n)$$

- **Problems**:
  1. The integral is intractable (no **closed-form** expression)
  2. The dimension of the parameters p is large
  3. The number of data point N is also large

The Fisher identity

$$\begin{aligned}
\nabla_\theta \log p_\theta(x) &= \int_{\mathbb{R}^p} \frac{\nabla_\theta p_\theta(x,z)}{p_\theta(x)} \mathrm{d}z \\
&= \int_{\mathbb{R}^p} \nabla \log p_\theta(x,z) \frac{p_\theta(x,z)}{p_\theta(x)} \mathrm{d}z \\
&= \int_{\mathbb{R}^p} \nabla \log p_\theta(x,z) p_\theta(z|x) \mathrm{d}z
\end{aligned}$$

- In words, the gradient of the **incomplete** likelihood is the conditional expectation of the gradient of the **complete data** likelihood (the joint likelihood of the observations and the latent data).
- The key behind many successful algorithms: EM algorithm, variational EM, etc...

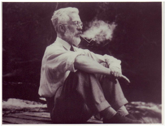


Figure: Fisher, 1946

A direct implementation

- **minibatch gradient descent:** intractable

$$\theta_{k+1} = \theta_k + \gamma_{k+1} \frac{N}{M} \sum_{j \in I_{k+1}} \nabla \log p_{\theta_k}(X_j)$$

here M is the batch size, I_{k+1} is a minibatch and γ_{k+1} is the stepsize.

- **Idea:** Using the **Fisher Identity**, replace the gradient by a **Monte-Carlo** approximation

$$\theta_{k+1} = \theta_k + \gamma_{k+1} \frac{N}{M} \sum_{j \in I_{k+1}} L^{-1} \sum_{\ell=1}^{L} \nabla \log p_{\theta_k}(X_j, Z_{j,i})$$

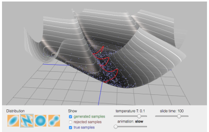
where $Z_{j,1}, \ldots, Z_{j,L}$ is a sample from the $p_{\theta_k}(z|X_j)$.

- **Problem:** direct sampling from $p_\theta(z|X_j)$ is not feasible.

this is a special case of **incremental optimization**. Better algorithm are available -but the memory imprint can be huge-: (Karimi and Wai, 2019).

Markov Chain Monte Carlo I

- Markov Chain Monte Carlo (MCMC) defines a class of methods to draw sample from an **arbitrary target distribution**
  - in this application, the posterior distribution of the latent state $\pi(z) = p_{\theta_k}(z|X_j)$.
- **Idea:** Construct an ergodic Markov chain whose invariant distribution is the target distribution π.
  - Take its roots in statistical physics.
- In most implementations, π need to be known up to a normalizing constant.
  - Since $p_\theta(z|x) \propto p_\theta(x,z)$ the knowledge of the complete data likelihood is all what we need !



Distribution

Show
☑ generated samples
☐ rejected samples
☑ true samples

temperature T: 0.1

animation: slow

slide time: 100

Markov Chain Monte Carlo II

- Many recent progresses have been achieved: Langevin Dynamics, Hamiltonian Monte Carlo,...
  - Complexity results are now available (Durmus and Moulines, 2017b), (Durmus and Moulines, 2017a)... [end of this talk]
- **Advantage:** estimate any expectation w.r.t. the posterior arbitrarily precisely
- **Drawbacks:**
  - MCMC may require many "burn-in" iterations to forget their initial state,
  - Successive samples from the chain may be highly correlated
  - Diagnosing when the chain has reached its stationary regime is challenging.

Variational Expectation-Maximization

- The Variational Expectation-Maximization (VEM) approximates the intractable $\log p_\theta(x)$ with the **evidence lower bound** (ELBO)
- **Idea:** Choose $q_\phi(z|x)$ a family of distributions indexed by parameters ϕ.
- **ELBO**

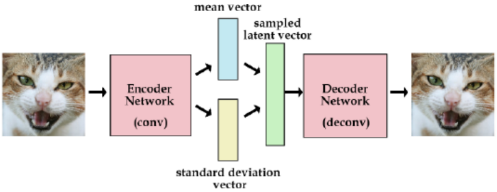
$$\begin{aligned}\mathrm{ELBO}(\theta,\phi,x) &\overset{\text{def}}{=} \int \log\left(\frac{p_\theta(x,z)}{q_\phi(z|x)}\right) q_\phi(z|x)\mathrm{d}z \\ &= \log p_\theta(x) - \int \log\left(\frac{q_\phi(z|x)}{p_\theta(z|x)}\right) p_\theta(z|x)\mathrm{d}z \\ &= \log p_\theta(x) - \mathrm{KL}(q_\phi(\cdot|x)\|p_\theta(\cdot|x)) \leq \log p_\theta(x)\ .\end{aligned}$$

  - Use the ELBO as a proxy to the incomplete data log-likelihood ! Bound tight if variational autoencoder matches real posterior.
  - Of course $\mathrm{ELBO}(\theta,\phi,x)$ is intractable but sampling from $q_\phi(z|x)$ is easy, cheap MC implementation can be obtained.

VEM in DGLM

- Standard choice: $q_\phi(z|x) = h(z; r_\phi(x))$ where $r_\phi(x)$, the **encoder** is the a output an neural network and $h(z;r)$ is a tractable distribution with parameters r.
- ELBO becomes a function of the **encoder** (inference net) and **decoder** (generative net)

$$\mathrm{ELBO}(\theta, \phi, x) = \int \log p_\theta(x|z) q_\phi(\cdot|x) \mathrm{d}z - \mathrm{KL}(q_\phi(\cdot|x)|\phi(z))$$



VAE vs. GAN

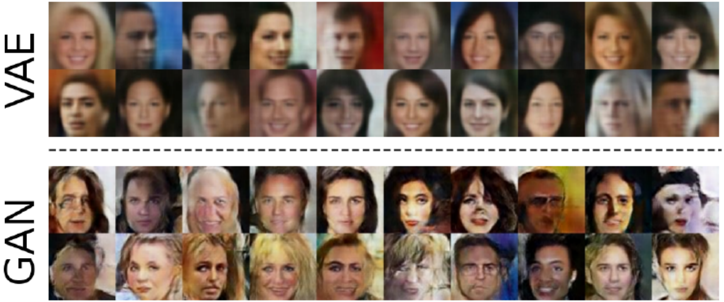


Figure from [Hou et al., 2016]

Making Progresses and Open problems

Many works in progress (and the competition is fierce !)

- More accurate bound for a given posterior
  - using importance sampling (Burda et al., 2016)
- Enlarge the family of variational posteriors
  - Improves posterior with a series of invertible transformations (VAE with normalizing flows)
  - Combines VAE with MCMC [Hamiltonian Variational Inference]
- Avoid (whenever possible) approximate inference

Uncertainty quantification

- ML is incorporated into many systems affecting the quality of human life, it is crucial to know how confident the model is when making decisions.
- Most ML methods are focused on point estimates and are poor at **representing uncertainty**...
  - Model uncertainty about the predicted class for data not trained to be recognized
  - Determine which examples are hard to recognize and require further inspection,
  - Sort out unusable data.
- In addition, most ML techniques are very sensitive to **adversarial attacks**.

Making informed decisions needs risk measures

A. Durmus (ENS Paris-Saclay), D. Belomestny (HSE, Moscow), A. Naumov (HSE, Moscow); PhD N. Brosse (Ecole Polytechnique), S. Samsonov (HSE, Moscow), L. Iosipoi (HSE, Moscow)

What should we care I ?

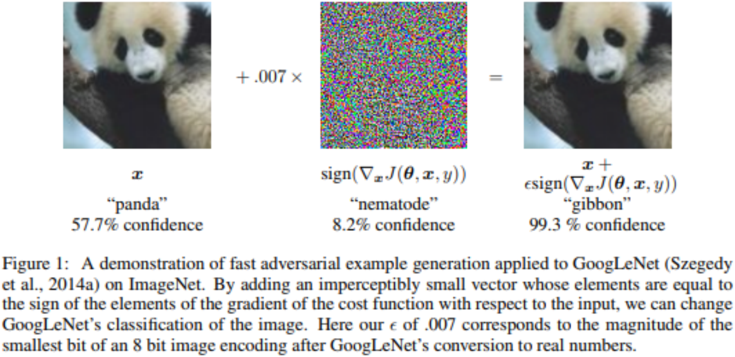
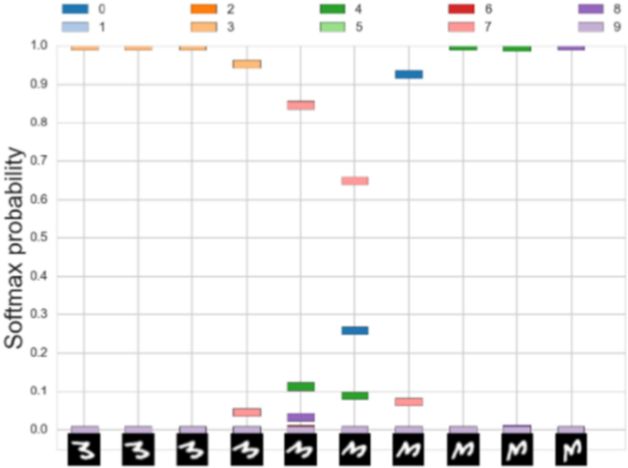

$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

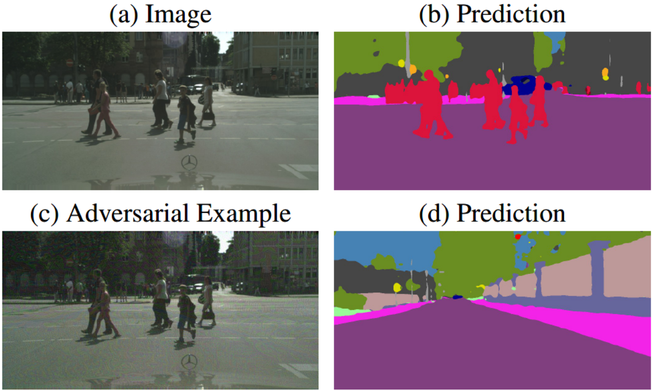
Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

Figure: From *Goodfellow et al*, Explaining and Harnessing adversarial examples (2014)

What should we care II ?



What should we care III ?



(a) Image

(b) Prediction

(c) Adversarial Example

(d) Prediction

Figure 1. The upper row shows an image from the validation set of Cityscapes and its prediction. The lower row shows the image perturbed with universal adversarial noise and the resulting prediction. Note that the prediction would look very similar for other

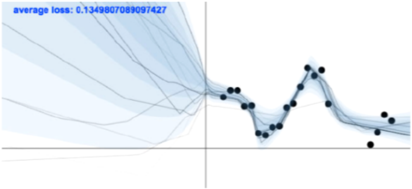
Sources of uncertainty in ML

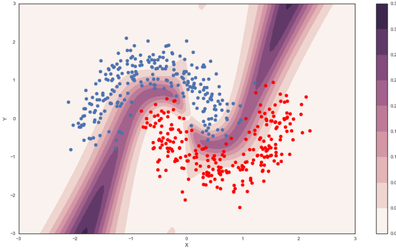
- **Random errors**
  - Noisy observations, missing data, wrong labels, latent data
- **Epistemic errors**
  - Uncertainty in the parameters (weights in a deep net may be poorly specified),
  - Uncertainty in the model (it is common practice to use a very large NN to flexibly fit data, and then reign in overfitting using regularization terms etc...)

Why being bayesian ?

- To address the issue of **overconfident prediction**, recent works have proposed approaches like
  - **calibration methods**,
  - **ensemble methods**,
  - **sampling techniques (DropOut, Dropconnect, bootstrap)**
- **Bayesian** inference offers a simple and principled approach to enable uncertainty estimates as it aims to **marginalize** the model parameters.

The holy grail





First results

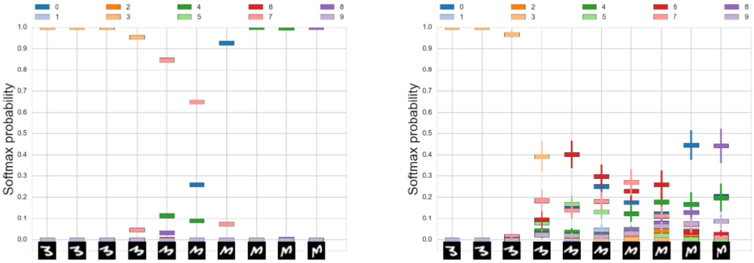


Figure: **Deterministic NN**: a **point estimate** of the output **overconfident**. **Bayesian framework** allows us to obtain a **distribution** over the outputs

Bayesian approach in a nutshell

- **Learning:** Compute the **posterior distribution** over the **model parameter**

$$p(\theta \mid \mathsf{D}_N) \propto \pi(\theta) \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \theta) \quad \mathsf{D}_N = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

  where

  - $\pi(\theta|)$ is the **prior** distribution of the parameter given the model,
  - $\mathsf{L}(y \mid \mathbf{x}, \theta)$ is the conditional distribution of the label y given the **features** $\mathbf{x}$ and the **parameters** θ.

- **Prediction** Compute the **posterior predictive distribution**

$$p(y \mid \mathbf{x}, \mathsf{D}_N) = \int p(y|\mathbf{x}, \theta) p(\theta \mid \mathsf{D}_N) \mathrm{d}\theta$$

  Bayesian methods outputs a **distribution on the labels** : it is a **probabilistic classifier**

Challenges in Bayesian machine learning

- One of the core problems of Bayesian machine learning is to sample / represent the **posterior distribution**
  - Fundamental for Bayesian inference which **frames all inference** about unknown quantities as a **calculation about the posterior**.
- Most of the Bayesian computational methods developed so far do not adapt when new needs arose such as
  - scalability to **massive data collections**,
  - number of free parameters.
- **Objective:** Develop new Bayesian computation paradigms ! Here again two options
  - Develop a new generation of sampling methods which scale (in the dimension, number of parameters)
  - Use variational inference (here again, use deep net to encode the distributions, combine with invertible flows, etc..)

Models for texture synthesis

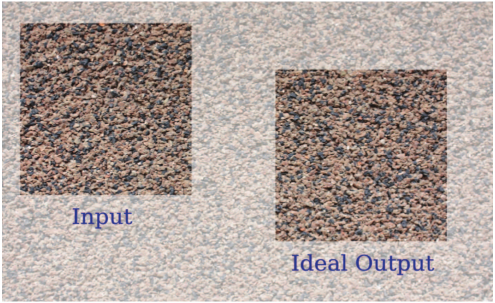


Figure: Exemplar-based texture synthesis

A. Desolneux (ENS Paris-Saclay), B. Galerne, (U. Orleans) and **V. De Bortoli** (PhD, ENS Paris-Saclay)

Work in Progress... A. Durmus (ENS Paris-Saclay), A. Doucet (U. Oxford), S. Mallat (ENS Ulm), A. Thin (PhD, Polytechnique)

Models for texture synthesis

- **Goal:** **sample** textures $X \sim \Pi^\star$ which **look like** an original texture x_0 but are not verbatim copies of x_0.
- **Challenge:** How to combine **randomness** and geometric **structure** in an image model?
- **A possible answer:** Maximize the **entropy** Ent under geometrical constraints specified by a vector-valued non-linear functions f.

Microcanonical and Macrocanonical models

Links with statistical physics (Bruna & Mallat, 2018)...

Microcanonical model

The probability distribution function $\Pi^\star \in \mathscr{P}$ is a *microcanonical model* associated with the exemplar texture $x_0 \in \mathbb{R}^d$, statistics $f : \mathbb{R}^d \to \mathbb{R}^p$ if

$$\mathrm{Ent}(\Pi^\star) = \max\left\{\mathrm{Ent}(\Pi),\ \Pi \in \mathscr{P},\ \Pi(\{x \in \mathbb{R}^d\colon \|f(x) - f(x_0)\| \leq \epsilon\}) = 1\right\}\ .$$

Macrocanonical model

The probability distribution function $\Pi^\star \in \mathscr{P}$ is a *macrocanonical model* associated with the exemplar texture $x_0 \in \mathbb{R}^d$, statistics $f : \mathbb{R}^d \to \mathbb{R}^p$ if

$$\mathrm{Ent}(\Pi^\star) = \max\left\{\mathrm{Ent}(\Pi),\ \Pi \in \mathscr{P},\ \Pi(f) = f(x_0)\right\}\ .$$

Notations: $\mathbb{E}_\Pi\left[f(X)\right] := \Pi(f)$.

Microcanonical and Macrocanonical models

Links with statistical physics (Bruna & Mallat, 2018)...

Microcanonical model

The probability distribution function $\Pi^\star \in \mathscr{P}$ is a *microcanonical model* associated with the exemplar texture $x_0 \in \mathbb{R}^d$, statistics $f : \mathbb{R}^d \to \mathbb{R}^p$ if

$$\mathrm{Ent}(\Pi^\star) = \max \left\{ \mathrm{Ent}(\Pi),\ \Pi \in \mathscr{P},\ \Pi(\{x \in \mathbb{R}^d \colon \|f(x) - f(x_0)\| \leq \epsilon\}) = 1 \right\} .$$

Macrocanonical model

The probability distribution function $\Pi^\star \in \mathscr{P}$ is a *macrocanonical model* associated with the exemplar texture $x_0 \in \mathbb{R}^d$, statistics $f : \mathbb{R}^d \to \mathbb{R}^p$ if

$$\mathrm{Ent}_\mu(\Pi^\star) = \min \left\{ \mathrm{KL}(\Pi, \mu),\ \Pi \in \mathscr{P},\ \Pi(f) = f(x_0) \right\} .$$

Notations: $\mathbb{E}_\Pi [f(X)] := \Pi(f)$, $\mu =$ reference probability measure

Fix $x_0 \in \mathbb{R}^d$ (exemplar texture), $f: \mathbb{R}^d \to \mathbb{R}^p$ (constraints) with $p \ll d$ and

- $\mathbb{R}^d$ - image space,
- $\mathbb{R}^p$ - parameter space.

Macrocanonical models are Gibbs measures

Under **mild** technical conditions, there exists $\theta^\star \in \mathbb{R}^p$ such that

$$\frac{\mathrm{d}\Pi_{\theta^\star}}{\mathrm{d}\mu}(x) \propto \exp(-\langle \theta^\star, f(x) - f(x_0) \rangle)$$

is a **macrocanonical model** associated with x_0 and f.
The **optimal parameter** $\theta^\star$ solves the following optimization problem:

$$\theta^\star \in \arg\min \left\{ \log \left[\int_{\mathbb{R}^d} \exp(-\langle \theta, f(x) - f(x_0) \rangle) \mathrm{d}\mu(x) \right], \theta \in \mathbb{R}^p \right\} .$$

1. How to find the optimal parameters $\theta^\star$?
2. How to sample from the model $\frac{\mathrm{d}\Pi_\theta}{\mathrm{d}\mu}(x) \propto \exp(-\langle \theta, f(x) - f(x_0) \rangle)$, *i.e.* how to sample from a Gibbs measure?

Finding the optimal parameters

The optimal parameters $\theta^\star$ minimize the **log-partition function**

$$L(\theta) = \log \left[\int_{\mathbb{R}^d} \exp(-\langle \theta, f(x) - f(x_0) \rangle) \mathrm{d}\mu(x) \right] \ .$$

Properties of the log-partition function

- $\nabla_\theta L(\theta) = -(\Pi_\theta(f) - f(x_0))$,
- $\nabla_\theta^2 L(\theta) = \mathrm{Cov}_{\Pi_\theta}(f)$ ⇒ convexity

$$\boxed{\theta_{n+1} = \theta_n + \delta_{n+1} \Pi_{\theta_n}(f - f(x_0)) \ ,}$$

Compute $\Pi_\theta(f)$ ⇒ Compute $\nabla L(\theta)$ ⇒ Gradient descent ⇒ Find $\theta^\star$.
Monte Carlo approximation of $\Pi_\theta(f)$ ⇒ how to sample from Π_θ?

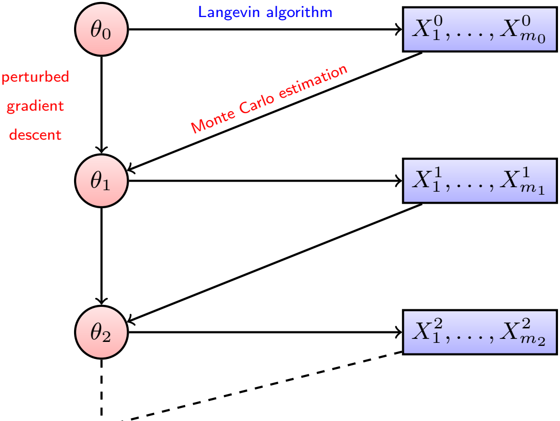
Sampling from Π_θ

- Usually it is not possible to sample from $\Pi(\mathrm{d}x) \propto \exp[-U(x)]\mathrm{d}x$, **but**,
- Approximate sampling is available using Markov Chain Monte Carlo, for example **(overdamped) Langevin Dynamic** (Durmus and Moulines, 2017b), (Durmus and Moulines, 2019)

$$\boxed{X_{n+1} = X_n - \gamma_{n+1} \nabla U(X_n) + \sqrt{2\gamma_{n+1}} Z_{n+1} \ ,}$$

where $Z_{n+1} \sim \mathcal{N}(0, \mathrm{Id})$, i.i.d., $\gamma_n > 0$. $\rightarrow$ **sample** $\approx \Pi(\mathrm{d}x) \propto \exp[-U(x)]\mathrm{d}x$.

Combining Optimization and Sampling



🟥 – parameter sequence $\in \mathbb{R}^p$ (optimization)

🟦 – image sequence $\in \mathbb{R}^d$ (sampling)

Let $U_\theta(x) = \langle \theta, f(x) - f(x_0) \rangle + r(x)$ (assuming that $\frac{\mathrm{d}\mu}{\mathrm{dLeb}}(x) \propto \exp(-r(x))$).

Finding optimal parameters

$\theta^\star$ is the minimum of the log-partition function which is a *convex problem*.

Gradient descent dynamics

$$\theta_{n+1} = \theta_n + \delta_{n+1} \Pi_{\theta_n}(f - f(x_0))$$

Sampling from a Gibbs measure

The potential $x \mapsto U_\theta(x)$ is usually *non-convex* but has *curvature at infinity*.

Langevin dynamics

$$X_{n+1} = X_n - \gamma_{n+1} \nabla U_\theta(X_n) + \sqrt{2\gamma_{n+1}} Z_{n+1}$$

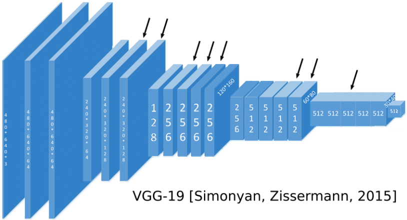
$$X_{k+1}^n = X_k^n - \gamma_n \nabla U_{\theta_n}(X_k^n) + \sqrt{2\gamma_n} Z_{k+1}^n \ , \text{ with } X_0^n = X_{m_{n-1}}^{n-1} \ ,$$

$$\theta_{n+1} = \theta_n + \delta_{n+1} m_n^{-1} \sum_{k=1}^{m_n} \{ f(X_k^n) - f(x_0) \} \ ,$$

where $Z_k^n \sim \mathcal{N}(0, \mathrm{Id})$, i.i.d.

Max entropy with Deep features

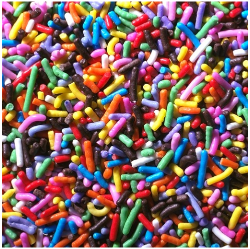
$\mathbf{j}$ = family of layers
c_ℓ = channels of layer ℓ
$\mathscr{G}_{\ell,c}$ = CNN feature at layer ℓ and channel c
$n_{\ell,c}$ = number of pixels at layer ℓ and channel c



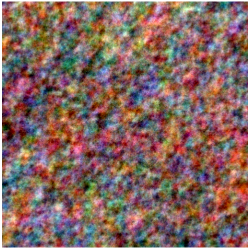
VGG-19 [Simonyan, Zissermann, 2015]

Figure: Structure of the neural network VGG-19

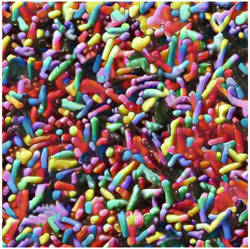
Choice of features: mean of each *channel* for selected *layers*, $p \approx 10^3$, *i.e.* $f(x) = (\sum_{i=1}^{n_{\ell,c}} \mathscr{G}_{\ell,c}(x)_i / n_{\ell,c})_{\ell \in \mathbf{j}, c \in c_\ell}$.



Input (x_0)



Initialization (Gaussian)



After 10000 iterations

(Some) open questions

- Interpretation of the macrocanonical parameters $\theta^\star$? Tight bounds? Concentration of the (non-convex) measure in high dimensional settings?
- (Markov chain) the mixing of the Langevin dynamic might be slow for some value of the macrocanonical parameters → replace MCMC by **generative networks** or **transform** the distribution with invertible flows ?
- (model) VGG features are arbitrary → which features for which problem?

1 Probabilistic Reasoning with DL: three examples

2 High-Dimensional Sampling

- Langevin Diffusion and Unadjusted Langevin Algorithm
- Strongly log-concave distribution

3 Variance reduction for Makov Chains

4 Bibliography

Framework

- Denote by π a target density w.r.t. the Lebesgue measure on $\mathbb{R}^d$, known up to a normalisation factor

$$x \mapsto \pi(x) \overset{\text{def}}{=} \mathrm{e}^{-U(x)} / \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y \ ,$$

  Implicitly, $d \gg 1$.

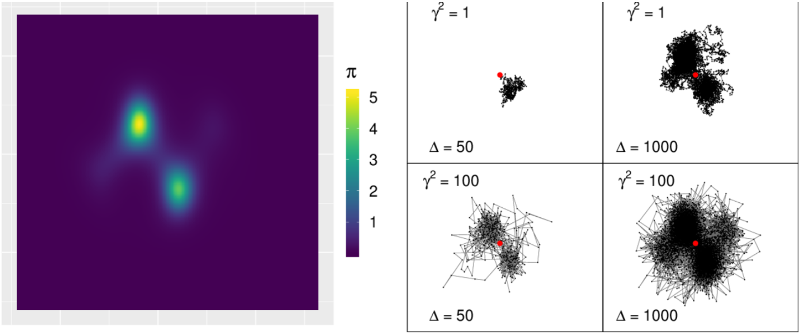
- **Assumption**: U is L-smooth : twice continuously differentiable and there exists a constant L such that for all $x, y \in \mathbb{R}^d$,

$$\|\nabla U(x) - \nabla U(y)\| \leq L \|x - y\| \ .$$

(Overdamped) Langevin diffusion

- **Langevin SDE**:
$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ ,$$
where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian Motion.
- **Notation**: $(P_t)_{t\geq 0}$ the Markov semigroup associated to the Langevin diffusion:
$$P_t(x, A) = \mathbb{P}(X_t \in A | X_0 = x) \ , \quad x \in \mathbb{R}^d, A \in \mathcal{B}(\mathbb{R}^d) \ .$$
- $\pi(x) \propto \exp(-U(x))$ is the unique **invariant probability** measure.

Langevin diffusion



Ergodicity

- **Key property 1**: For all $x \in \mathbb{R}^d$,

$$\lim_{t \to +\infty} \|\delta_x P_t - \pi\|_{\mathrm{TV}} = 0 \ .$$

- **Key property 2**: for "nice" functions

$$\frac{1}{T}\int_0^T f(X_t)\mathrm{d}t \xrightarrow{\mathbb{P}_x - \text{a.s.}} \pi(f) = \int \pi(\mathrm{d}x) f(x)$$

$$\frac{1}{\sqrt{T}}\int_0^T \{f(X_t) - \pi(f)\}\mathrm{d}t \overset{\mathbb{P}_x}{\Longrightarrow} \mathcal{N}(0, \sigma^2(\pi, f)) \ .$$

- The Langevin diffusion provides a mean to sample **any** smooth distribution... Of course, this is a highly theoretical solution...

Discretized Langevin diffusion

- **Idea:** Sample the diffusion paths, using the **Euler-Maruyama (EM)** scheme:

$$X_{k+1} = X_k - \gamma_{k+1} \nabla U(X_k) + \sqrt{2\gamma_{k+1}} Z_{k+1}$$

  where
  - $(Z_k)_{k \geq 1}$ is i.i.d. $\mathcal{N}(0, \mathrm{I}_d)$
  - $(\gamma_k)_{k \geq 1}$ is a sequence of stepsizes, which can either be held constant or be chosen to decrease to 0 at a certain rate.
- Closely related to the **(stochastic) gradient descent algorithm**.

Discretized Langevin diffusion: constant stepsize

- When the stepsize is held **constant**, *i.e.* $\gamma_k = \gamma$, then $(X_k)_{k\geq 1}$ is an **homogeneous Markov chain** with Markov kernel R_γ
- Under some appropriate conditions, this Markov chain is irreducible, positive recurrent $\rightsquigarrow$ unique invariant distribution π_γ which **does not coincide** with the target distribution π.
- **Questions**:
  - For a given precision $\epsilon > 0$, how should I choose the stepsize $\gamma > 0$ and the number of iterations n so that : $\|\delta_x R_\gamma^n - \pi\|_{\mathrm{TV}} \leq \epsilon$
  - Is there a way to choose the starting point x cleverly ?
  - Auxiliary question: quantify the distance between π_γ and π.

Discretized Langevin diffusion: decreasing stepsize

- When $(\gamma_k)_{k\geq 1}$ is nonincreasing and non constant, $(X_k)_{k\geq 1}$ is an **inhomogeneous Markov chain** associated with the kernels $(R_{\gamma_k})_{k\geq 1}$.
- **Notation**: Q^p_γ is the composition of Markov kernels

$$Q^p_\gamma = R_{\gamma_1} R_{\gamma_2} \dots R_{\gamma_p}$$

With this notation, $\mathbb{E}_x[f(X_p)] = \delta_x Q^p_\gamma f$.

- **Questions:**
  - **Convergence** : is there a way to choose the step sizes so that $\|\delta_x Q^p_\gamma - \pi\|_{\mathrm{TV}} \to 0$ and if yes, what is the optimal way of choosing the stepsizes ?...
  - **Optimal choice of simulation parameters** : What is the number of iterations required to reach a neighborhood of the target: $\|\delta_x Q^p_\gamma - \pi\|_{\mathrm{TV}} \leq \epsilon$ starting from a given point x
  - Should we use **fixed** or **decreasing** step sizes ?

Strongly convex potential

- **Assumption**: U is L-smooth and m-strongly convex

$$\|\nabla U(x) - \nabla U(y)\|^2 \leq L \|x-y\|^2$$
$$\langle \nabla U(x) - \nabla U(y), x-y \rangle \geq m \|x-y\|^2 \ .$$

- **Outline of the proof**
  1. Control in W_2 the distance of the laws of the Langevin diffusion and its discretized version.
  2. Relate W_2 control to total variation.
- **Key technique**: **(Synchronous and Reflection) coupling !**; see (Durmus and Moulines, 2019)

Coupling of probability measures

Definition

- A coupling of two probability measures $(\xi, \xi') \in \mathbb{M}_1(\mathcal{X}) \times \mathbb{M}_1(\mathcal{X})$ is a probability measure γ on the product space $(\mathsf{X} \times \mathsf{X}, \mathcal{X} \otimes \mathcal{X})$ whose marginals are ξ and ξ', *i.e.* $\gamma(A \times \mathsf{X}) = \xi(A)$ and $\gamma(\mathsf{X} \times A) = \xi'(A)$ for all $A \in \mathcal{X}$.
- The set of all couplings of ξ and ξ' is denoted by $\mathcal{C}(\xi, \xi')$.
- A coupling $\gamma \in \mathcal{C}(\xi, \xi')$ is said to be optimal for the Hamming distance if $\gamma(\Delta^c) = d_{\mathrm{TV}}(\xi, \xi')$ where $\Delta = \left\{(x, x') \in \mathsf{X}^2 \;:\; x = x'\right\}$ is the **diagonal** of $\mathsf{X} \times \mathsf{X}$

Wasserstein distance

Definition

For $p \geq 1$ and $\xi, \xi' \in \mathbb{M}_1(\mathcal{X})$, the **Wasserstein distance** of order p between ξ and ξ' denoted by $\mathbf{W}_{\mathrm{d},p}(\xi, \xi')$, is defined by

$$\mathbf{W}^p_{\mathrm{d},p}\left(\xi, \xi'\right) = \inf_{\gamma \in \mathcal{C}(\xi,\xi')} \int_{\mathsf{X}\times\mathsf{X}} \mathrm{d}^p(x, x')\gamma(\mathrm{d}x\mathrm{d}x')\ ,$$

where $\mathcal{C}(\xi, \xi')$ is the set of coupling of ξ and ξ'. For $p = 1$, we simply write $\mathbf{W}_{\mathrm{d}}$.

Properties of the Wasserstein distance

- The Wasserstein distance can be expressed in terms of random variables as:

$$\mathbf{W}_{\mathrm{d},p}\left(\xi,\xi'\right) = \inf_{(X,X')\in\mathcal{C}(\xi,\xi')} \left\{\mathbb{E}[\mathrm{d}^p(X,X')]\right\}^{1/p} \ ,$$

  where $(X,X') \in \mathcal{C}(\xi,\xi')$ that the distribution of the pair of random elements (X,X') is a coupling of ξ and ξ'.
- Any particular coupling therefore provides an upper bound of the Wasserstein distance.
- By Hölder's inequality, it obviously holds that if $p \leq q$, then for all $\xi,\xi' \in \mathbb{M}_1(\mathcal{X})$,

$$\mathbf{W}_{\mathrm{d},p}\left(\xi,\xi'\right) \leq \mathbf{W}_{\mathrm{d},q}\left(\xi,\xi'\right) \ .$$

Wasserstein distance convergence

Theorem

*Assume that U is L-**smooth** and m-**strongly convex**. Then, for all $x, y \in \mathbb{R}^d$ and $t \geq 0$,*

$$\mathbf{W}_2 \left(\delta_x P_t, \delta_y P_t \right) \leq \mathrm{e}^{-mt} \left\| x - y \right\|$$

The **contraction** depends only on the **strong convexity** constant.

Synchronous Coupling

$$\begin{cases} \mathrm{d}Y_t &= -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ , \\ \mathrm{d}\tilde{Y}_t &= -\nabla U(\tilde{Y}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ , \end{cases} \quad \text{where } (Y_0, \tilde{Y}_0) = (x, y).$$

This SDE has a unique strong solution $(Y_t, \tilde{Y}_t)_{t \geq 0}$. Since

$$\mathrm{d}\{Y_t - \tilde{Y}_t\} = -\left\{\nabla U(Y_t) - \nabla U(\tilde{Y}_t)\right\}\mathrm{d}t$$

The product rule for semimartingales imply

$$\mathrm{d}\left\|Y_t - \tilde{Y}_t\right\|^2 = -2\left\langle \nabla U(Y_t) - \nabla U(\tilde{Y}_t), Y_t - \tilde{Y}_t \right\rangle \mathrm{d}t \ .$$

Synchronous Coupling

$$\left\|Y_t - \tilde{Y}_t\right\|^2 = \left\|Y_0 - \tilde{Y}_0\right\|^2 - 2\int_0^t \left\langle (\nabla U(Y_s) - \nabla U(\tilde{Y}_s)), Y_s - \tilde{Y}_s \right\rangle \mathrm{d}s \ ,$$

Since U is strongly convex $\langle \nabla U(y) - \nabla U(y'), y - y' \rangle \geq m \left\|y - y'\right\|^2$ which implies

$$\left\|Y_t - \tilde{Y}_t\right\|^2 \leq \left\|Y_0 - \tilde{Y}_0\right\|^2 - 2m\int_0^t \left\|Y_s - \tilde{Y}_s\right\|^2 \mathrm{d}s \ .$$

Grömwall inequality:

$$\left\|Y_t - \tilde{Y}_t\right\|^2 \leq \left\|Y_0 - \tilde{Y}_0\right\|^2 \mathrm{e}^{-2mt}$$

Theorem

*Assume that U is L-**smooth** and m-**strongly convex**. Then, for any $x \in \mathbb{R}^d$ and $t \geq 0$*

$$\mathbb{E}_x \left[\|Y_t - x^\star\|^2 \right] \leq \|x - x^\star\|^2 \, \mathrm{e}^{-2mt} + \frac{d}{m}(1 - \mathrm{e}^{-2mt}) \; .$$

where

$$x^\star = \arg\min_{x \in \mathbb{R}^d} U(x) \; .$$

The stationary distribution π satisfies

$$\int_{\mathbb{R}^d} \|x - x^\star\|^2 \, \pi(\mathrm{d}x) \leq d/m.$$

The constant depends only **linearly** in the **dimension** d.

Contraction property of the discretization

Theorem

Assume that U *is* L*-smooth and* m*-strongly convex. Then,*

(i) *Let* $(\gamma_k)_{k\geq 1}$ *be a nonincreasing sequence with* $\gamma_1 \leq 2/(m+L)$*. For all* $x, y \in \mathbb{R}^d$ *and* $\ell \geq n \geq 1$*,*

$$W_2(\delta_x Q_\gamma^{n,\ell}, \delta_y Q_\gamma^{n,\ell}) \leq \left\{ \prod_{k=n}^{\ell} (1-\kappa\gamma_k) \|x-y\|^2 \right\}^{1/2} .$$

where $\kappa = 2mL/(m+L)$*.*

(ii) *For any* $\gamma \in (0, 2/(m+L))$*, for all* $x \in \mathbb{R}^d$ *and* $n \geq 1$*,*

$$W_2(\delta_x R_\gamma^n, \pi_\gamma) \leq (1-\kappa\gamma)^{n/2} \left\{ \|x-x^\star\|^2 + 2\kappa^{-1} d \right\}^{1/2} .$$

A coupling proof (I)

- **Objective** compute bound for $W_2(\delta_x Q_\gamma^n, \pi)$
- Since $\pi P_t = \pi$ for all $t \geq 0$, it suffices to get bounds of the Wasserstein distance

$$\mathbf{W}_2\left(\delta_x Q_\gamma^n, \pi P_{\Gamma_n}\right)$$

where

$$\Gamma_n = \sum_{k=1}^n \gamma_k \ .$$

  - $\delta_x Q_\gamma^n$: law of the discretized diffusion
  - $\pi P_{\gamma_n} = \pi$, where $(P_t)_{t \geq 0}$ is the semi group of the diffusion
- **Idea ! synchronous coupling** between the **diffusion** and the interpolation of the **Euler discretization**.

Explicit bound in Wasserstein distance

Theorem

Assume that U is m-strongly convex and L-smooth. Let $(\gamma_k)_{k\geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq 1/(m+L)$. Then

$$W_2^2(\delta_x Q_\gamma^n, \pi) \leq u_n^{(1)}(\gamma)\left\{\|x - x^\star\|^2 + d/m\right\} + u_n^{(2)}(\gamma) ,$$

where $u_n^{(1)}(\gamma) = 2\prod_{k=1}^{n}(1-\kappa\gamma_k)$ *with* $\kappa = mL/(m+L)$ *and*

$$u_n^{(2)}(\gamma) = 2\frac{dL^2}{m}\sum_{i=1}^{n}\left[\gamma_i^2 c(m, L, \gamma_i)\prod_{k=i+1}^{n}(1-\kappa\gamma_k)\right] .$$

Can be sharpened if U is three times continuously differentiable and there exists $\tilde{L}$ such that for all $x, y \in \mathbb{R}^d$, $\left\|\nabla^2 U(x) - \nabla^2 U(y)\right\| \leq \tilde{L}\,\|x - y\|$.

Results

- **Fixed step size** For any $\epsilon > 0$, one may choose γ so that

$$\mathbf{W}_2\left(\delta_{x_*} R_\gamma^p, \pi\right) \leq \epsilon \quad \text{in } p = \mathcal{O}(\sqrt{d}\epsilon^{-1}) \text{ iterations}$$

  where x_* is the unique maximum of π

- **Decreasing step size** with $\gamma_k = \gamma_1 k^{-\alpha}$, $\alpha \in (0,1)$,

$$\mathbf{W}_2\left(\delta_{x_*} Q_\gamma^n, \pi\right) = \sqrt{d}\mathcal{O}(n^{-\alpha}) \; .$$

- These results are tight (check with $U(x) = 1/2\|x\|^2$).

Total Variation

Definition

For μ, ν two probabilities measure on $\mathbb{R}^d$, define

$$d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \sup_{\|f\|_\infty \leq 1} |\mu(f) - \nu(f)| = \inf_{(X,Y) \in \mathcal{C}(\mu,\nu)} \mathbb{P}(X \neq Y),$$

where $(X, Y) \in \mathcal{C}(\mu, \nu)$ if $X \sim \mu$ and $Y \sim \nu$.

$$\begin{aligned} |\mu(f) - \nu(f)| &= \mathbb{E}[f(X) - f(Y)] \\ &= \mathbb{E}[\{f(X) - f(Y)\}\mathbb{1}_{\{X \neq Y\}}] \leq \operatorname{osc}(f)\mathbb{P}(X \neq Y) \ . \end{aligned}$$

From the Wasserstein distance to the TV

Theorem

If U is strongly convex, then for all $x, y \in \mathbb{R}^d$,

$$\|P_t(x,\cdot) - P_t(y,\cdot)\|_{\mathrm{TV}} \leq 1 - 2\boldsymbol{\Phi}\left\{-\frac{\|x-y\|}{\sqrt{(4/m)(\mathrm{e}^{2mt}-1)}}\right\}$$

Use **reflection coupling** (Lindvall and Rogers, 1986)

Explicit bound in total variation

Theorem

- *Assume U is* **L-smooth and strongly convex**. *Let $(\gamma_k)_{k\geq 1}$ be a nonincreasing sequence with $\gamma_1 \leq 1/(m+L)$.*
- **(Optional assumption)** $U \in C^3(\mathbb{R}^d)$ *and there exists $\tilde{L}$ such that for all $x, y \in \mathbb{R}^d$: $\left\|\nabla^2 U(x) - \nabla^2 U(y)\right\| \leq \tilde{L}\,\|x-y\|$.*

Then there exist sequences $\{\tilde{u}_n^{(1)}(\gamma), n \in \mathbb{N}\}$ and $\{\tilde{u}_n^{(1)}(\gamma), n \in \mathbb{N}\}$ such that for all $x \in \mathbb{R}^d$ and $n \geq 1$,

$$\|\delta_x Q_\gamma^n - \pi\|_{\mathrm{TV}} \leq \tilde{u}_n^{(1)}(\gamma)\left\{\|x - x^\star\|^2 + d/m\right\} + \tilde{u}_n^{(2)}(\gamma)\ .$$

Constant step sizes

- For any $\epsilon > 0$, the minimal number of iterations to achieve $\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \epsilon$ is

$$p = \mathcal{O}(\sqrt{d}\log(d)\epsilon^{-1}\left|\log(\epsilon)\right|) \ .$$

- For a given stepsize γ, letting $p \to +\infty$, we get:

$$\|\pi_\gamma - \pi\|_{\mathrm{TV}} \leq C\gamma\left|\log(\gamma)\right| \ .$$

1 Probabilistic Reasoning with DL: three examples

2 High-Dimensional Sampling

3 Variance reduction for Makov Chains
- Control variates for Markov Chains
- Control variates for Langevin diffusion

4 Bibliography

Control variates methodology

- A "naive" Monte Carlo estimator of $\pi(f)$ is

$$\hat{\pi}_n(f) = \frac{1}{n} \sum_{k=0}^{n-1} f(X_k)$$

where $(X_k)_{k \in \mathbb{N}}$ is a MC of kernel R with invariant distribution π.

- Denote by $\hat{f}_{\mathrm{d}}$ a solution of the Poisson equation

$$\hat{f}_{\mathrm{d}} - R\hat{f}_{\mathrm{d}} = \tilde{f}\,, \quad \tilde{f} = f - \pi(f)\,,$$

- Using $\tilde{f}(X_k) = \hat{f}_{\mathrm{d}}(X_k) - R\hat{f}_{\mathrm{d}}(X_k)$ we get

$$\begin{aligned} n^{-1/2} \sum_{k=0}^{n-1} \tilde{f}(X_k) &= n^{-1/2} \sum_{k=0}^{n-1} \{\hat{f}_{\mathrm{d}}(X_k) - R\hat{f}_{\mathrm{d}}(X_k)\} \\ &= n^{-1/2} \sum_{k=1}^{n} \left\{\hat{f}_{\mathrm{d}}(X_k) - R\hat{f}_{\mathrm{d}}(X_{k-1})\right\} + \hat{f}_{\mathrm{d}}(X_0) - \hat{f}_{\mathrm{d}}(X_n)\,. \end{aligned}$$

CLT for Markov Chains

Theorem

Let P *be a Markov kernel with a unique invariant probability measure* π*. Let* $f \in \mathrm{L}^2(\pi)$*. Assume that there exists a solution* $\hat{f}_{\mathrm{d}} \in \mathrm{L}^2(\pi)$ *to the Poisson equation* $\hat{f}_{\mathrm{d}} - R\hat{f}_{\mathrm{d}} = \tilde{f}$ *where* $\tilde{f} = f - \pi(f)$*. Then*

$$n^{-1/2} \sum_{k=0}^{n-1} \tilde{f}(X_k) \xrightarrow{\mathbb{P}_\pi} \mathcal{N}(0, \sigma^2_\pi(\tilde{f})) ,$$

where

$$\begin{aligned} \sigma^2_\pi(\tilde{f}) &= \mathbb{E}_\pi[\{\hat{f}_{\mathrm{d}}(X_1) - R\hat{f}_{\mathrm{d}}(X_0)\}^2] = \pi(\hat{f}^2_{\mathrm{d}}) - \pi((R\hat{f}_{\mathrm{d}})^2) \\ &= 2\pi(\tilde{f}\hat{f}_{\mathrm{d}}) - \pi(\tilde{f}^2) \end{aligned}$$

Control variate

- **Control variates** *i.e.* π-integrable functions $\mathcal{H} \subset \left\{h : \mathbb{R}^d \to \mathbb{R} \ : \ \pi(h) = 0\right\}$ and then choose $h \in \mathcal{H}$ such that

$$\boxed{\sigma^2_{\infty,\mathrm{d}}(f+h) \leq \sigma^2_{\infty,\mathrm{d}}(f).}$$

- **Idea:** Consider control variates of the form $h = (R - \mathrm{Id})g$
- **Key remark**:

$$(R - \mathrm{Id})(\hat{f}_{\mathrm{d}} - g) = -\{\tilde{f} + (R - \mathrm{Id})g\}$$

- **Optimization problem**

$$\sigma^2_{\infty,\mathrm{d}}(f + (R - \mathrm{Id})g) \\ = \min_g 2\pi(\{\tilde{f} + (R - \mathrm{Id})g\}\{\hat{f}_{\mathrm{d}} - g\}) - \pi(\{\tilde{f} + (R - \mathrm{Id})g\}^2) \ .$$

Control variates

- **Computational bottleneck**: $\hat{f}_{\mathrm{d}}$ is intractable !
- **Idea**: This computation can be bypassed if R is **reversible** with respect to π !

$$
\begin{aligned}
&\pi(\{\tilde{f} + (R - \mathrm{Id})g\}\{\hat{f}_{\mathrm{d}} - g\}) \\
&\qquad\qquad = \pi(\tilde{f}\hat{f}_{\mathrm{d}}) + \pi((R - \mathrm{Id})g\hat{f}_{\mathrm{d}}) - \pi(g(R - \mathrm{Id})g) - \pi(\tilde{f}g)
\end{aligned}
$$

Since $R - \mathrm{Id}$ is **reversible** and $(R - \mathrm{Id})\hat{f}_{\mathrm{d}} = -\tilde{f}$,

$$
\boxed{\pi((R - \mathrm{Id})g\hat{f}_{\mathrm{d}}) = -\pi(g\tilde{f})}
$$

- **Good news** ☺ we can minimize the asymptotic variance without computing $\hat{f}_{\mathrm{d}}$.
- **Bad news** ☹ It is still required to compute Rg, which is in many instances overwhelming !

Langevin diffusion

- Langevin SDE:

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ .$$

- Generator of the Langevin semigroup: for any smooth function φ by

$$\mathscr{A}\varphi = \lim_{t\downarrow 0}(P_t\varphi - \varphi)/t = -\langle \nabla U, \nabla\varphi\rangle + \Delta\varphi \ .$$

- **Central Limit Theorem**:

$$t^{-1/2}\int_0^t \tilde{f}(Y_s)\mathrm{d}s \overset{\text{weakly}}{\underset{t\to+\infty}{\Longrightarrow}} \mathcal{N}(0, \sigma^2_\infty(f)) \ , \quad \sigma^2_\infty(f) = 2\pi\left(\hat{f}\tilde{f}\right) \ .$$

where $\hat{f}$ is a solution of the (continuous-time) **Poisson equation**

$$\mathscr{A}\hat{f} = -\tilde{f} \quad \text{with} \quad \tilde{f} = f - \pi(f) \ .$$

"Carré du champ"

- **"Carré du champ"** for all $g, h \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$,

$$\pi\left(g \mathscr{A} h\right) = \pi\left(h \mathscr{A} g\right) = -\pi\left(\langle \nabla g, \nabla h \rangle\right) \ ,$$

  Hence $\mathscr{A}$ is a self-adjoint operator on a dense subspace of $\mathrm{L}^2(\pi)$.

- A straightforward consequence of the "carré du champ" property (setting $h = 1$) is that

$$\pi(\mathscr{A} g) = 0 \quad \text{for any function } g \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R}).$$

- This observation implies that f and $f + \mathscr{A} g$ have the same expectation with respect to π for any $f \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $g \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$.

Control variate for Langevin diffusion

- f and $f + \mathscr{A} g$ have the same expectation with respect to π for any $f \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ and $g \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$.
- **Idea:** minimize the asymptotic variance

$$g \mapsto \sigma^2_\infty(f + \mathscr{A} g) \ .$$

- **Good news** ☺**!** This minimization problem can be solve without knowing the Poisson solution $\hat{f}$!

Another expression for the variance

- The asymptotic variance of $t^{-1/2} \int_0^t \tilde{f}(Y_s) \mathrm{d}s$ is given by

$$\sigma^2_\infty(f) = 2\pi(\hat{f}\tilde{f}) \; ,$$

where $\hat{f} \in \mathrm{C}^2_{\mathrm{poly}}(\mathbb{R}^d, \mathbb{R})$ satisfies Poisson's equation: $\mathscr{A} \hat{f} = -\tilde{f}$

- Using the "carré du champ" property, the asymptotic variance may also be expressed as

$$\sigma^2_\infty(f) = 2\pi\left(\hat{f}\tilde{f}\right) = -2\pi(\hat{f}\mathscr{A}\hat{f}) = 2\pi(\|\nabla \hat{f}\|^2) \; .$$

Control variates for the Langevin diffusion

- **Control variate:** $h^\star = \mathscr{A} g^\star$, where $g^\star$ is a minimizer of

$$g \mapsto \sigma^2_\infty(f + \mathscr{A} g) = 2\pi(\{\widehat{f + \mathscr{A} g}\}\{\tilde{f} + \mathscr{A} g\}) \ .$$

- **Key remark**:

$$\mathscr{A}(\hat{f} - g) = -(\tilde{f} + \mathscr{A} g) \ .$$

- **Consequence**:

$$\sigma^2_\infty(f + \mathscr{A} g) = 2\pi\left((\hat{f} - g)\left\{\tilde{f} + \mathscr{A} g\right\}\right) \ .$$

Control variates for the Langevin diffusion

- $\mathscr{A}$ **self-adjoint** in $\mathrm{L}^2(\pi)$ implies

$$\pi(\hat{f}\mathscr{A}g) = \pi(g\mathscr{A}\hat{f}) = -\pi(\tilde{f}g)$$

- An equivalent expression of the variance:

$$\begin{aligned}\sigma^2_\infty(f+\mathscr{A}g) &= 2\pi(\hat{f}\tilde{f}) - 2\pi(g\tilde{f}) + 2\pi(\hat{f}\mathscr{A}g) - 2\pi(g\mathscr{A}g)\\ &= 2\pi(\hat{f}\tilde{f}) - 4\pi(g\tilde{f}) + 2\pi(\|\nabla g\|^2)\ .\end{aligned}$$

- We can solve the minimization problem without computing $\hat{f}$!

$$\min_g \sigma^2_\infty(f+\mathscr{A}g) \ \Leftrightarrow\ \min_g -4\pi(g\tilde{f}) + 2\pi(\|\nabla g\|^2)$$

Linear parameterized family

- Linear parameterized family of smooth functions

$$g_\theta = \langle \theta, \psi \rangle \ , \quad \text{where } \theta \in \mathbb{R}^p \ , \ \psi = \{\psi_i\}_{i=1}^p \ .$$

- Minimization of the asymptotic variance:

$$\sigma_\infty^2(f + \mathscr{A} g_\theta) = 2\theta^{\mathrm{Tr}} H\theta - 4 \langle \theta, b \rangle + \sigma_\infty^2(f) \ ,$$
$$H_{ij} = \pi(\langle \nabla\psi_i, \nabla\psi_j \rangle) \quad \text{and} \quad b_i = \pi(\psi_i \tilde{f}) \ .$$

which is a quadratic minimization problem, which admits a closed form solution $\theta^* = H^{-1}b$ (beware that in practice H can be badly conditioned).

- **Good news !** $\mathscr{A} g$ is generally easy to compute contrary to Rg.

Problem

- **Minimizing the asymptotic variance** of the Langevin diffusion of control variates is **easy** (e.g. for "linear" control variates).
- **Question**: is this useful to compute control variates for discrete-time MC ?
  - Yes ! provided that the asymptotic variance of the Markov chain is in some sense well approximated by the asymptotic variance of the Langevin diffusion !
  - Of course, this is not always true !... but at least, this holds for several algorithms of interest ...

1 Probabilistic Reasoning with DL: three examples

2 High-Dimensional Sampling

3 Variance reduction for Makov Chains

4 Bibliography

Karimi, Belhal and Hoi-To and Moulines Wai Éric and Lavielle. 2019. *On the global convergence of (fast) incremental expectation maximization methods*.

Durmus, Alain and Eric and Saksman Moulines Eero. 2017a. *On the convergence of Hamiltonian Monte Carlo*, arXiv preprint arXiv:1705.00166.

Durmus, Alain and Eric Moulines. 2017b. *Nonasymptotic convergence analysis for the unadjusted Langevin algorithm*, The Annals of Applied Probability **27**, no. 3, 1551–1587.

______. 2019. *High-dimensional Bayesian inference via the unadjusted Langevin algorithm*, Bernoulli **25**, no. 4A, 2854–2882.

U. Simsekli and A. Durmus and R. Badeau and G. Richard and É. Moulines and A. T. Cemgil. 2017. *Parallelized Stochastic Gradient Markov Chain Monte Carlo algorithms for non-negative matrix factorization*, 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Randal Douc · Eric Moulines · Pierre Priouret · Philippe Soulier

Markov Chains

This book covers the classical theory of Markov chains on general state-spaces as well as many recent developments. The theoretical results are illustrated by simple examples, many of which are taken from Markov Chain Monte Carlo methods. The book is self-contained, while all the results are carefully and concisely proven. Bibliographical notes are added at the end of each chapter to provide an overview of the literature.

Part I lays the foundations of the theory of Markov chain on general states-space. Part II covers the basic theory of irreducible Markov chains on general states-space, relying heavily on regeneration techniques. These two parts can serve as a text on general state-space applied Markov chain theory. Although the choice of topics is quite different from what is usually covered, where most of the emphasis is put on countable state space, a graduate student should be able to read almost all these developments without any mathematical background deeper than that needed to study countable state space (very little measure theory is required).

Part III covers advanced topics on the theory of irreducible Markov chains. The emphasis is on geometric and subgeometric convergence rates and also on computable bounds. Some results appeared for a first time in a book and others are original. Part IV are selected topics on Markov chains, covering mostly hot recent developments.

Mathematics

ISBN 978-3-319-97703-4

9 783319 977034

▸ springer.com

ORFE

Douc · Moulines · Priouret · Soulier

Markov Chains

Springer Series in Operations Research and Financial Engineering

Randal Douc · Eric Moulines
Pierre Priouret · Philippe Soulier

Markov Chains

Springer

Figure: A beach reading or a nice Christmas present for a loved one... This new book covers the classical theory of Markov chains on general state-spaces as well as many recent developments. The theoretical results are illustrated by simple examples, many of which are taken from Markov Chain Monte Carlo methods.