Convolutional networks

- Example: classification problem
- Output: $p(y|x, w)$ for each class $y$. 
Uncertainty estimation

If input is far away from training data, instead of saying "I don’t know", the network can be overconfident due to overfitting.

- We want to know the uncertainty of the predictions.
Bayesian deep learning

- To account for uncertainty, the output of the NN is a probability distribution of well, probability distributions.
- Random initialisations for $w \sim p$ some prior.

Instead of point prediction $p(y|x, w)$, Bayesian $p(y|x, w)$ is a random variable, $w$-measurable.

$\leadsto$ Take the expectancy $\mathbb{E}_w[p(y|x, w)]$ for a prediction.

1 inspired from Stochastic variational inference and Bayesian neural networks by Nadezhda Chirkova, HSE Moscow, Russia
BNN as a family of NN

- To estimate $\mathbb{E}_w[p(y|x, w)]$, take the average over several NN with different initial weights, sampled from the prior $p(w|X, Y) \sim$ more calculations.
Why go Bayesian?

This method allows to:

1. **Limit overfitting**, as an average of - opposite - overconfident guesses is closer to uniform = no guess.
2. Estimate **uncertainty** and reduce it by taking more averages.
3. Prior can come with properties we want to infer into the network. E.g. after one training, we can use the posterior distribution as a new prior to train again on another data set \( \leadsto \) **continual learning.**
Bayesian considerations for training BNN

- The framework "prior x data $\sim$ posterior" is the classical Bayesian one.
- We need to access the posterior (renormalized Likelihood x Prior)

$$p(w|X, Y) = \frac{p(Y|X, w)p(w)}{\int p(Y|X, \tilde{w})p(\tilde{w})d\tilde{w}}.$$  

- In practice, the integral is intractable:
  - $\sim$ Usually we approximate the posterior with a well known parametric family with stochastic variational inference$^2$.
  - $\sim$ We can also approximate this integral with MCMC methods.

---

$^2$see *Bayesian neural networks* from Dmitry Molchanov, Samsung AI, Moscow Russia
Bayesian considerations for training BNN

- The framework "prior x data ⇝ posterior" is the classical Bayesian one.
- We need to access the posterior (renormalized Likelihood x Prior)
  \[ p(w|X,Y) = \frac{p(Y|X,w)p(w)}{\int p(Y|X,\tilde{w})p(\tilde{w})d\tilde{w}}. \]
- In practice, the integral is intractable:
  - Usually we approximate the posterior with a well known parametric family with stochastic variational inference\(^2\).
  - We can also approximate this integral with MCMC methods.
    - This is where my study can be useful.

\(^2\)see Bayesian neural networks from Dmitry Molchanov, Samsung AI, Moscow Russia
Optimal Scaling of Metropolis adjusted Langevin approximations with Laplace target

Pablo Jiménez
under the supervision of Alain Durmus
December 11 2019

ENS Paris Saclay, CMLA
Outline

Motivation

Metropolis-Hastings algorithm

The optimal scaling problem

Main results
Motivation
Sampling from a probability distribution

- **Input:** target density $\pi$ over $\mathbb{R}^d$.
- **Output:** $(X_1^d, \ldots, X_n^d)_n$ i.i.d. sampling $\pi$.
- **Goal:** approximate integrals of the form
  \[ \int_{\mathbb{R}^d} h(x^d) \pi(x^d) \, dx^d. \]

See an example
Sampling from a probability distribution

- Input: target density $\pi$ over $\mathbb{R}^d$.
- Output: $(X_1^d, \ldots, X_n^d)_n$ i.i.d. sampling $\pi$.
- Goal: approximate integrals of the form
  \[
  \int_{\mathbb{R}^d} h(x^d) \pi(x^d) dx^d.
  \]

See an example
Sampling from a probability distribution

• Input: target density $\pi$ over $\mathbb{R}^d$.
• Output: $(X_1^d, \ldots, X_n^d)_n$ i.i.d. sampling $\pi$.
• Goal: approximate integrals of the form
  \[ \int_{\mathbb{R}^d} h(x^d)\pi(x^d)dx^d. \]

See an example
Monte Carlo method: summary

- Goal: approximate integrals of the form \( \int_{\mathbb{R}^d} h(x^d)\pi(x^d)dx^d \).
- Estimate with \( n^{-1} \sum_{k=1}^{n} h(X_k^d) \).
- Useful in high dimension \( \rightarrow \) Bayesian inference.
- Problem: \( (X_n^d)_{n \in \mathbb{N}^*} \) i.i.d. of law \( \pi \) \( \rightarrow \) costly/intractable.
Monte Carlo method: summary

- Goal: approximate integrals of the form $\int_{\mathbb{R}^d} h(x^d)\pi(x^d)dx^d$.
- Estimate with $n^{-1}\sum_{k=1}^{n} h(X_k^d)$.
- Useful in high dimension $\rightarrow$ Bayesian inference.
- Problem: $(X_n^d)_{n \in \mathbb{N}^*}$ i.i.d. of law $\pi$ $\rightarrow$ costly/intractable.
- Solution: $(X_n^d)_{n \in \mathbb{N}^*}$ Markov chain approximating $\pi$.

Markov chain Monte Carlo (MCMC)

- Metropolis et al. (1953)
- Hastings (1970)

N. Metropolis (1915-1999)
W.K. Hastings (1930-2016)
Metropolis-Hastings algorithm
1. Given $X_n^d \in \mathbb{R}^d$, sample from the proposal $Y_{n+1}^d$ under $q^d(X_n^d, \cdot)$.

2. Calculate the acceptance rate $\alpha(X_n^d, Y_{n+1}^d)$, where

$$\alpha(x^d, y^d) = 1 \wedge \frac{\pi(y^d) q^d(y^d, x^d)}{\pi(x^d) q^d(x^d, y^d)}.$$  

3. Sample $U_{n+1}$, uniform over $[0, 1]$.

4. Evaluate the acceptance event $A_{n+1}^d$, given by

$$A_{n+1}^d = \left\{ U_{n+1} \leq \alpha(X_n^d, Y_{n+1}^d) \right\}.$$  

5. Summary: $X_{n+1}^d$ is given by

$$X_{n+1}^d = Y_{n+1}^d \mathbb{1}_{A_{n+1}^d} + X_n^d \left( 1 - \mathbb{1}_{A_{n+1}^d} \right).$$
Mise en pratique de l’algorithme

1. Sample $X_0^d$.
2. Sample $Y_1^d$.
3. Compute the acceptance probability $\alpha(X_n^d, Y_{n+1}^d)$.
4. Accept or reject $Y_1^d$. 
Mise en pratique de l’algorithme

1. Sample $X_0^d$.

2. Sample $Y_1^d$.

3. Compute the acceptance probability $\alpha(X_n^d, Y_{n+1}^d)$.

4. Accept or reject $Y_1^d$. 
Mise en pratique de l’algorithme

1. Sample $X_0^d$.
2. Sample $Y_1^d$.
3. Compute the acceptance probability $\alpha(X_n^d, Y_{n+1}^d)$.
4. Accept or reject $Y_1^d$. 

Accept ratio 0.7659
Theorem: Tierney (1994)

Under reasonable assumptions on $q^d$, $(X_n^d)_n$ is an **ergodic** Markov chain, with **invariant law** $\pi$.  

I.i.d. sample

Sample from a Markov chain
Theorem: Tierney (1994)

Under reasonable assumptions on $q^d$, $(X_n^d)_n$ is an ergodic Markov chain, with invariant law $\pi$.

Assumptions: $q^d$ and $\pi$ positives over $\mathbb{R}^d$.

How to choose the proposal $q^d$?
Example: random walk Metropolis (RWM)

Step of the random walk Metropolis algorithm.

- \( Y_{n+1}^d = X_n^d + Z_{n+1}^d \), where \( Z_{n+1}^d \) is a zero-mean gaussian.
- Proposal \( q^d(x^d, \cdot) \) zero-mean gaussian \( N(x^d, \sigma_d^2 \text{Id}_d) \).
- Tune \( \sigma_d^2 \) w.r.t. the target \( \pi \):
  - large enough to limit correlations,
  - small enough to have a high acceptance rate.
Example: Langevin dynamics

- Assume $\pi : x^d \mapsto \exp (-V(x^d))/Z$, with normalization constant $Z$, and smooth $V : \mathbb{R}^d \to \mathbb{R}$.

- SDE of the diffusion process $(X_t^d)_{t \geq 0}$:

$$dX_t^d = -\nabla V(X_t^d) \, dt + \sqrt{2} \, dB_t^d,$$

where $B_t^d$ is a standard brownian motion over $\mathbb{R}^d$.

- Roberts and Tweedie (1996): $(X_t^d)_{t \geq 0}$ is ergodic with invariant distribution $\pi$.

- Problem: intractable in practice.

- Idea: use Euler-Murayama discretization.
Example: unadjusted Langevin approximation

- SDE of the diffusion process \((X^d_t)_{t \geq 0}\):
  \[
  dX^d_t = -\nabla V(X^d_t) \, dt + \sqrt{2} dB^d_t.
  \]

- Euler–Maruyama discretization \((X^d_k)_{k \geq 0}\), with step-size \(\sigma^2_d > 0\) defines a Markov chain:
  \[
  X^d_{k+1} = X^d_k - \sigma^2_d \nabla V(X^d_k) + \sqrt{2\sigma^2_d} Z^d_{k+1},
  \]
  with independent sequences \(\{(Z^d_k)_{k \geq 1} \mid d \in \mathbb{N}^*\}\), s.t. \((Z^d_k)_{k \geq 1}\) i.i.d. sampled from \(\mathcal{N}(0, \text{Id}_d)\).

- Problem: \(\pi\) is not invariant for \((X^d_k)_{k \geq 1}\).

- Idea: introduce this dynamic as a proposal in M-H.
Metropolis adjusted Langevin approximations (MALA)

- Proposals \((Y_n^d)_n\) given by discretization:
  \[
  Y_{n+1}^d = X_n^d - \sigma_d^2 \nabla V (X_n^d) + \sqrt{2\sigma_d^2} Z_{n+1}^d,
  \]
  with \(Z_{n+1}^d\) sampled from \(N(0, \text{Id}_d)\).

- Biased gaussian proposal towards high density.

- Tune \(\sigma_d^2\) w.r.t. the target \(\pi\).
How to optimize $\sigma_d^2$ w.r.t. dimension $d$ and target $\pi$?

Choice of $\sigma_d^2 = 2.25$

Choice of $\sigma_d^2 = 0.5625$
How to optimize $\sigma_d^2$ w.r.t. dimension $d$ and target $\pi$?

Before answering, explain what an "optimal choice" is.
How to tune the step-size of the proposal?

In the case of RWM and MALA: finding a "good" choice of $\sigma_d^2$ and quantify the efficiency of the methods.

Objectives:

1. approximate the behavior in high dimension with a simpler object,
2. find an optimization criterion for $\sigma_d^2$,
3. compare different methods.
An asymptotical frame

- From target $\pi$ over $\mathbb{R}$, consider a sequence $(\pi^d)_{d \geq 1}$
- $\pi^d(x^d) = \prod_{i=1}^{d} \pi(x^d_i)$
- Consider the sequence of Markov chains $\{(X^n_d)_{n \geq 0} \mid d \geq 1\}$, started at stationnarity, over the same probabilistic space.
- Goal: study the limit of $\mathbb{P}(A^d_1)$ when $d \to +\infty$.
- $A^d_1 = \left\{ U_1 \leq \alpha(X^d_0, Y^d_1) \right\}$
- Problem: if $\sigma^2_d$ constant w.r.t. $d$, $\lim_{d \to +\infty} \mathbb{P}(A^d_1) = 0$. 

MALA with Laplace target

- Dimension vs. acceptance rate for different values of $\sigma^2_d$. 

- $\sigma^2_d = 4$, $\sigma^2_d = 1$, $\sigma^2_d = 0.25$.
First criterion for $\sigma_d^2$: scaling

- Choice: $\sigma_d^2 = \ell^2 / d^\alpha$ with $\ell > 0$ and $\alpha > 0$.
- Goal: find $\alpha$ s.t.

$$a(\ell) = \lim_{d \to +\infty} \mathbb{P}(A_1^d) \in (0, 1)$$

for any $\ell > 0$.
- We say that the method scales as $1/d^\alpha$.

Example of scaling in $\alpha = 2/3$. 

MALA with Laplace target
A scaling result

- Interpolate and scale $X^d$ into $Y^d$, with

  \[ Y_t^d = (\lceil d^\alpha t \rceil - d^\alpha t) X^d_{\lfloor d^\alpha t \rfloor} + (d^\alpha t - \lfloor d^\alpha t \rfloor) X^d_{\lceil d^\alpha t \rceil}. \]

- Roberts et al. (1997): first component of $Y^d$ converges in distribution towards $Y$, solution of the SDE:

  \[ dY_t = h(\ell)V'(Y_t)dt + \sqrt{2h(\ell)}dB_t. \]
A scaling result

- Interpolate and scale $X^d$ into $Y^d$, with

$$Y_t^d = ([d^\alpha t] - d^\alpha t) X_{[d^\alpha t]}^d + (d^\alpha t - [d^\alpha t]) X_{[d^\alpha t]}^d.$$ 

- Roberts et al. (1997): first component of $Y^d$ converges in distribution towards $Y$, solution of the SDE:

$$dY_t = h(\ell)V'(Y_t)dt + \sqrt{2h(\ell)}dB_t.$$
• Interpolate and scale $X^d$ into $Y^d$, with

$$Y^d_t = ([d^\alpha t] - d^\alpha t) X^d_{[d^\alpha t]} + (d^\alpha t - [d^\alpha t]) X^d_{[d^\alpha t]}.$$ 

• Roberts et al. (1997): first component of $Y^d$ converges in distribution towards $Y$, solution of the SDE:

$$dY_t = h(\ell)V'(Y_t)dt + \sqrt{2h(\ell)}dB_t.$$
A scaling result

- Interpolate and scale $X^d$ into $Y^d$, with

\[ Y_t^d = (\lceil d^\alpha t \rceil - d^\alpha t) X_{\lceil d^\alpha t \rceil}^d + (d^\alpha t - \lfloor d^\alpha t \rfloor) X_{\lfloor d^\alpha t \rfloor}^d. \]

- Roberts et al. (1997): first component of $Y^d$ converges in distribution towards $Y$, solution of the SDE:

\[ dY_t = h(\ell) V'(Y_t) dt + \sqrt{2h(\ell)} dB_t. \]
A scaling result

- Interpolate and scale $X^d$ into $Y^d$, with

\[
Y^d_t = ([d^\alpha t] - d^\alpha t) X^d_{[d^\alpha t]}
+ (d^\alpha t - [d^\alpha t]) X^d_{[d^\alpha t]}. 
\]

- Roberts et al. (1997): first component of $Y^d$ converges in distribution towards $Y$, solution of the SDE:

\[
dY_t = h(\ell)V'(Y_t)dt \\
+ \sqrt{2h(\ell)} dB_t. 
\]
A scaling result

- Interpolate and scale $X^d$ into $Y^d$, with

$$Y^d_t = (\lceil d^\alpha t \rceil - d^\alpha t) X^d_{\lceil d^\alpha t \rceil} + (d^\alpha t - \lfloor d^\alpha t \rfloor) X^d_{\lfloor d^\alpha t \rfloor}.$$  

- Roberts et al. (1997): first component of $Y^d$ converges in distribution towards $Y$, solution of the SDE:

$$dY_t = h(\ell) V'(Y_t) dt + \sqrt{2} h(\ell) dB_t.$$
A scaling result

- Interpolate and scale $X^d$ into $Y^d$, with

$$Y^d_t = (\lceil d^\alpha t \rceil - d^\alpha t) X^d_{\lfloor d^\alpha t \rfloor} + (d^\alpha t - \lfloor d^\alpha t \rfloor) X^d_{\lceil d^\alpha t \rceil}.$$  

- Roberts et al. (1997): first component of $Y^d$ converges in distribution towards $Y$, solution of the SDE:

$$dY_t = h(\ell)V'(Y_t)dt + \sqrt{2h(\ell)}dB_t.$$
Objectives:

1. approximate the behavior in high dimension with a simpler object,
2. find an optimization criterion for $\sigma^2_d$,
3. compare different methods.
How to tune the step-size of the proposal?

Objectives:

1. approximate the behavior in high dimension with a simpler object, [diffusive limit](#).
2. find an [optimization criterion](#) for $\sigma_d^2$.
3. compare different methods.
Consequences of the diffusive limit

- \( Y \) is solution of the SDE:

\[
dY_t = h(\ell) \left( \log \pi \right)'(Y_t) dt + \sqrt{2h(\ell)} dB_t.
\]

(1)

- The diffusion speed \( h \) is the limit of the first order efficiency:

\[
2h(\ell) = \lim_{d \to +\infty} d^{\alpha} \mathbb{E} \left[ (X_{0,1}^d - X_{1,1}^d)^2 \right].
\]

- If \( Y^{(1)} \) satisfies SDE (1) with \( h = 1 \), then

\[
\left( Y^{(1)}_{h(\ell)t} \right)_t \xrightarrow{\text{distr.}} (Y_t)_t.
\]
Consequences of the diffusive limit

- If $Y^{(1)}$ satisfies SDE (1) with $h = 1$, then
  \[
  \left( Y^{(1)}_{\ell t} \right) \stackrel{\text{distr.}}{=} \left( Y_t \right)_t.
  \]

\[\neq\] diffusion speeds $h_0 < h_1 \nRightarrow \neq$ convergence speeds.
Consequences of the diffusive limit

• Roberts and Tweedie (1996): $Y^{(1)}$ is ergodic and $\pi$ is an invariant distribution.

$\Rightarrow$ Maximizing $h$ yields the best convergence.

• Scaling of $Y^d$ in $d^\alpha$:

\[
Y^d_t = (\lceil d^\alpha t \rceil - d^\alpha t) X^d_{d^\alpha t} + (d^\alpha t - \lfloor d^\alpha t \rfloor) X^d_{d^\alpha t}.
\]

$\Rightarrow$ $d^\alpha$ is an estimate of the method’s complexity.
How to tune the step-size of the proposal?

Objectives:

1. approximate the behavior in high dimension with a simpler object, \( \text{diffusive limit} \),
2. find an optimization criterion for \( \sigma^2_d \),
3. compare different methods.
How to tune the step-size of the proposal?

Objectives:

1. approximate the behavior in high dimension with a simpler object, \( \Box \text{diffusive limit} \),

2. find an optimization criterion for \( \sigma_d^2 \), \( \checkmark \sigma_d^2 = \ell^2/d^\alpha \) where \( \ell \) maximises \( h \),

3. compare different methods.
Objectives:

1. approximate the behavior in high dimension with a simpler object, \( \text{diffusive limit} \),
2. find an optimization criterion for \( \sigma_d^2 \), \( \sqrt{\sigma_d^2} = \ell^2 / d^\alpha \) where \( \ell \) maximises \( h \),
3. compare different methods, \( \text{comparing the scaling value } \alpha \) \( \sim \) better if \( \alpha \) is small.
### State of the art: comparing methods

<table>
<thead>
<tr>
<th>Target</th>
<th>MALA</th>
<th>RWM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smooth</td>
<td>Roberts and Rosenthal (1997) $\alpha = 1/3$, $a = 0.574$</td>
<td>Roberts et al. (1997) $\alpha = 1$, $a = 0.234$</td>
</tr>
<tr>
<td>Non-smooth</td>
<td>no general result</td>
<td>Durmus et al. (2017) $\alpha = 1$, $a = 0.234$</td>
</tr>
</tbody>
</table>

- $a$ is the optimal acceptance rate $\Rightarrow$ equivalent to optimize $\ell > 0$.  

Question: MALA for non-smooth targets?  

Special case $\pi: x \mapsto e^{-|x|}/2$.  


State of the art: comparing methods

<table>
<thead>
<tr>
<th>Target</th>
<th>MALA</th>
<th>RWM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smooth</td>
<td>Roberts and Rosenthal (1997) $\alpha = 1/3$, $a = 0.574$</td>
<td>Roberts et al. (1997) $\alpha = 1$, $a = 0.234$</td>
</tr>
<tr>
<td>Non-smooth</td>
<td>no general result</td>
<td>Durmus et al. (2017) $\alpha = 1$, $a = 0.234$</td>
</tr>
</tbody>
</table>

- $a$ is the optimal acceptance rate $\mapsto$ equivalent to optimize $\ell > 0$.
- Question: MALA for non-smooth targets?
- Special case $\pi : x \mapsto e^{-|x|}/2$. 
Main results
Optimal scaling of MALA with Laplace target

- **Target** $\pi : x \mapsto e^{-|x|}/2$.
- **Markov chain** given by:
  \[ X_{n+1}^d = X_n^d + \left( -\frac{\ell^2}{d^\alpha} \text{sgn}(X_n^d) + \sqrt{2} \frac{\ell}{d^{\alpha/2}} Z_{n+1}^d \right) 1_{A_{n+1}^d}. \]
- **Reminder**: acceptance rate $P(A_1^d) = \mathbb{E}[\alpha(X_0^d, Y_1^d)]$.

**Theorem (scaling value)**

Let $\sigma_d^2 = \ell^2/d^\alpha$, with $\alpha = 2/3$, and $\Phi$ the $N(0,1)$ c.d.f. Then
\[
\lim_{d \to +\infty} P[A_1^d] = a(\ell), \text{ where } a(\ell) = 2\Phi(-\ell^{3/2}/(3\pi^{1/2})^{1/2}).
\]
Acceptance rate w.r.t. $\ell$ for MALA with Laplace target.
Optimal scaling of MALA with Laplace target

**Theorem (diffusive limit and optimal scaling)**

- $\alpha = 2/3$ and $(Y^d_t)_{t \geq 0}$ given by

$$Y^d_t = \left(\lceil d^{2/3}t \rceil - d^{2/3}t\right)X^d_{\lfloor d^{2/3}t \rfloor} + \left(d^{2/3}t - \lfloor d^{2/3}t \rfloor\right)X^d_{\lceil d^{2/3}t \rceil}.$$

- Let $(Y_t)_{t \geq 0}$ be a solution of the SDE

$$dY_t = h(\ell) \text{sgn}(Y_t) dt + \sqrt{2h(\ell)} dB_t.$$

- Then $\{(Y^d_t,1)_{t \geq 0} \mid d \in \mathbb{N}^*\}$ converges in distribution to $(Y_t)_{t \geq 0}$.

- Moreover $h(\ell) = \ell^2 \Phi(-\ell^{3/2}/(3\pi^{1/2})^{1/2})$,

- $h$ is maximized for $\ell$ s.t. $a(\ell) = a = 0.360$. 
First order efficiency times $d^{2/3}$ w.r.t. acceptance rate, for MALA with Laplace target.
Scaling results for MALA with Laplace target:

- New scaling in $\alpha = 2/3$.
- Weakly convergence of $Y^d$ to a diffusive limit.
- New optimal acceptance rate $a(\ell) = 0.360$.

Perspective: generalize to more non-smooth densities e.g. continuous but not $C^1$. 
Thank you for your time!


Monte Carlo method: example

- On veut estimer $\pi$.
- Alors
  \[ \pi = U([-1, 1]^2) \text{ et } h = 1_{B(0,1)}. \]
- \[ \int_{\mathbb{R}^2} h(x^2)\pi(x^2)dx^2 = \pi/4. \]
Monte Carlo method: example

On veut estimer $\pi$.

Alors

$$\pi = U([-1,1]^2) \text{ et } h = 1_{B(0,1)}.$$ 

$$\int_{\mathbb{R}^2} h(x^2) \pi(x^2) dx^2 = \pi/4.$$
Monte Carlo method: example

On veut estimer $\pi$.

Alors

$$\pi = U([-1, 1]^2) \text{ et } h = 1_{B(0,1)}.$$ 

$$\int_{\mathbb{R}^2} h(x^2)\pi(x^2)dx^2 = \pi/4.$$ 

Continue presentation