Generative modelling challenge
Second session

Alain Durmus
École polytechnique

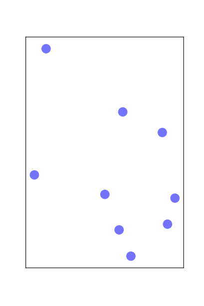
November 15, 2022

Variational Autoencoders

Introduction to generative models

Let $\mathsf{X} \subset \mathbb{R}^d$.

- **Input**
  Data $\{x_i\}_{i=1}^N$: N i.i.d. observations from an unknown distribution $\mu^\star \in \mathcal{P}(\mathsf{X})$.
  Notation: $\mathcal{P}(\mathsf{X})$ space of probability measures on $(\mathsf{X}, \mathcal{X})$.
- **Output**
  New samples from $\mu^\star$



thiscatdoesnotexist.com

Minimum Distance Estimation Wolfowitz (1957)

- **Input** For $X \subset \mathbb{R}^d$.
  Data $\{x_i\}_{i=1}^N$: N i.i.d. observations from $\mu^\star \in \mathcal{P}(X)$ unknown
- **Output**
  New samples from $\mu^\star$

- Consider a parametric family of distributions $\{\mu_\theta \ : \ \theta \in \Theta\}$.
- Minimum distance estimation :
  - minimize $\theta \mapsto \mathbf{D}(\mu_\theta | \mu^\star)$ where $\mathbf{D}$ is a divergence over the space of probability measure on X.
  - Sample a new observation from $\mu_{\theta^\star}$.

Divergence over $\mathcal{P}(\mathsf{X})$

- A divergence on $\mathcal{P}(\mathsf{X})$, is a function $\mathbf{D} : \mathcal{P}(\mathsf{X})^2 \to \mathbb{R}_+$ which satisfies *the most important* axiom of a distance: for μ, ν two probability measures

$$\mathbf{D}(\mu|\nu) = 0 \text{ if and only if } \mu = \nu \quad .$$

- Any distance on $\mathcal{P}(\mathsf{X})$ is a divergence.
- Do not satisfy the triangle inequality in general except if it is a distance...
- It is not symmetric in general.
- Important example:

$$\mathrm{KL}\left(\mu \,\|\, \nu\right) = \begin{cases} \int \log(\mathrm{d}\mu/\mathrm{d}\nu)\mathrm{d}\nu & \text{if } \mu \ll \nu \\ +\infty & \text{otherwise} \quad . \end{cases}$$

Minimum Distance Estimation: an ideal?

- **Input** For $X \subset \mathbb{R}^d$.
  Data $\{x_i\}_{i=1}^N$: N i.i.d. observations from $\mu^\star \in \mathcal{P}(X)$ unknown
- **Output**
  New samples from $\mu^\star$

- Consider a parametric family of distributions $\{\mu_\theta \;:\; \theta \in \Theta\}$.
- Minimum distance estimation :
  - minimize $\theta \mapsto \mathbf{D}(\mu_\theta|\mu^\star)$ where $\mathbf{D}$ is a divergence over the space of probability measure on X.
  - Sample a new observation from $\mu_{\theta^\star}$.
- Problem: $\mu^\star$ is not known and in general $\mathbf{D}(\mu_\theta|\mu^\star)$ is not tractable.
- One solution presented here: likelihood estimation.

Maximum likelihood estimation

- Consider the case $\mathsf{X} = \mathbb{R}^p$ and $\Theta \subset \mathbb{R}^d$.
- Choice for the family $\{\mu_\theta \ : \ \theta \in \Theta\}$?

$$\{\mu_\theta \ : \ \mu_\theta \ll \mathrm{Leb} \ , \quad p_\theta = \mathrm{d}\mu_\theta / \mathrm{dLeb}\} \ .$$

  Example:

$$p_\theta : \text{ density w.r.t. Leb of } \mathcal{N}(\mathrm{m}, \sigma^2) \ .$$

- We should be able to sample from p_θ for any θ...
- Choice for the divergence $\mathbf{D}$?

Maximum likelihood estimation

- Consider the case $\mathsf{X} = \mathbb{R}^p$ and $\Theta \subset \mathbb{R}^d$.
- Choice for the family $\{\mu_\theta \,:\, \theta \in \Theta\}$?

$$\{\mu_\theta \,:\, \mu_\theta \ll \mathrm{Leb}\,, \quad p_\theta = \mathrm{d}\mu_\theta / \mathrm{dLeb}\}\ .$$

- Choice for the divergence $\mathbf{D}$?
- $\mathbf{D} = \mathsf{KL}$: problem $\mathrm{KL}\left(\mu_\theta \,\|\, \hat{\mu}_N\right) = \infty$...
- Recall that ideally if $\mathbf{D} = \mathsf{KL}$, we would like to minimize

$$\mathrm{KL}\left(\mu^\star \,\|\, \mu_\theta\right) = -\int \mathrm{d}\mu^\star \log\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu^\star}\right)\ .$$

- This is equivalent to maximize if $\mu^\star \ll \mathrm{Leb}$,

$$\theta \mapsto \int \mathrm{d}\mu^\star \log\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{dLeb}}\right)\ .$$

- Solution: replace the integral by an empirical version

$$\theta \mapsto N^{-1} \sum_{i=1}^{N} \log p_\theta(x_i)\ .$$

Density estimation

- Assume now

$$\mu^\star \ll \mathrm{Leb}\ ,\ \text{with density } p^\star\ .$$

- Choice for the family $\{\mu_\theta\ :\ \theta \in \Theta\}$?

$$\{\mu_\theta\ :\ \mu_\theta \ll \mathrm{Leb}\ ,\quad p_\theta = \mathrm{d}\mu_\theta/\mathrm{dLeb}\}\ .$$

- We should be able to sample from μ_θ for any θ and in the same time the family has to be sufficiently rich/large.
- First solution:

$$\mu_\theta = (\mathrm{T}_\theta)_\sharp \nu_0\ ,$$

where

$$\nu_0 \in \mathcal{P}(\mathbb{R}^p) \text{ with density } q_0\ ,\quad \mathrm{T}_\theta : \mathbb{R}^p \to \mathbb{R}^p \text{ is a } \mathrm{C}^1 \text{ diffeomorphism}\ .$$

- In that case, we have

$$p_\theta(x) = q_0(\mathrm{T}_\theta^{\leftarrow}(x))\mathbf{Jac}_x[\mathrm{T}_\theta^{\leftarrow}](x)\ .$$

- Example: $\mathrm{T}_\theta(x) = \mathrm{m} + \mathbf{\Sigma}x$, $\theta = (\mathrm{m}, \mathbf{\Sigma})$.

- Approximating the function $p^\star(x)$ is called the density estimation problem.
- Suppose we (somehow) optimize a generator p_θ

$$\theta \mapsto \int \mathrm{d}\mu^\star \log\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{dLeb}}\right) \text{ or } \theta \mapsto \sum_{i=1}^{N} \log p_\theta(x_i) \ ,$$

  such that $p_\theta \approx p^\star$.
- If we can compute $\mathrm{T}_\theta^{\leftarrow}(x)$ and $\mathbf{Jac}_x[\mathrm{T}_\theta^{\leftarrow}](x)$ then we have a density estimator.
- In many cases, generation (sampling) is easier than density estimation.

Density estimation

- How to optimize?

$$\theta \mapsto \int \mathrm{d}\mu^\star \log\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{dLeb}}\right) \text{ or } \theta \mapsto \sum_{i=1}^{N} \log p_\theta(x_i)\ . \tag{1}$$

- Stochastic gradient descent:

$$\theta_{k+1} = \theta_k + \gamma_{k+1} \sum_{x_i \in \mathrm{B}_{k+1}} \nabla_\theta[\log p_{(\cdot)}(x_i)](\theta_k)\ ,$$

where

$$(\mathrm{B}_k)_k \text{ is a sequence of random batch of data points },$$
$$(\gamma_k)_k \text{ is a sequence of stepsizes/learning rates }.$$

- Under some assumptions, it can be shown that almost surely $(\theta_k)_{k\in\mathbb{N}}$ converges to some minimizers of (1).

Density estimation

- Choice for the family $\{\mu_\theta \,:\, \theta \in \Theta\}$? $\{\mu_\theta \,:\, \mu_\theta \ll \mathrm{Leb}\;, \quad p_\theta = \mathrm{d}\mu_\theta / \mathrm{dLeb}\}$.
- First solution:

$$\mu_\theta = (\mathrm{T}_\theta)_\sharp \nu_0 \;,$$

  where

$$\nu_0 \in \mathcal{P}(\mathbb{R}^p) \text{ with density } q_0 \;, \quad \mathrm{T}_\theta : \mathbb{R}^p \to \mathbb{R}^p \text{ is a } \mathrm{C}^1 \text{ diffeomorphism} \;.$$

- In that case, we have

$$p_\theta(x) = q_0(\mathrm{T}_\theta^{\leftarrow}(x))\mathbf{Jac}_{\mathrm{T}_\theta^{\leftarrow}}(x) \;.$$

- What people thought it was hard:

$$\text{find } \{\mathrm{T}_\theta \,:\, \theta \in \Theta\} \text{ such that } \theta \mapsto \sum_{i=1}^{N} \log p_\theta(x_i) \text{ easy to optimize...}$$

- It turns out that such constructions are now possible using neural networks; see normalizing flows Rezende and Mohamed (2015)!
- Here we present a first alternative using MLE: latent variable models.

- Here we aim to construct a family $\{p_\theta : \theta \in \Theta\}$ which is expressive enough.
- Given some samples $\{x_i\}_{i=1}^N$ and a samplable prior r.
- Generative latent variable model:
  1. $z \sim r$;
  2. $x \sim p_\theta(x|z)$
- Typically, $\log p_\theta$ is the error function of a Neural Network:

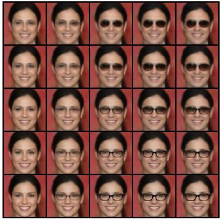
$$\log p_\theta(x, z) = -\ell(\mathrm{T}_\theta(z), x) \ .$$

- This corresponds to the marginal likelihood:

$$\begin{aligned} p_\theta(x) &= \int p_\theta(x|z) r(z) \mathrm{d}z \\ &= \int p_\theta(x, z) \mathrm{d}z \ . \end{aligned}$$

Motivation for latent variables

- Conditioning on the latent variable z may give to the samples x more global coherence.
- Learned latent distribution might reveal structure in the data distribution.
- The latent variables could be useful for downstream tasks or interpretability.



Klys, Snell, and Zemel, Neurips 2018

MLE for latent variable models?

- Marginal likelihood:

$$p_\theta(x) = \int p_\theta(x|z)r(z)\mathrm{d}z = \int p_\theta(x,z)\mathrm{d}z \ .$$

- Based on samples $\{x_i\}_{i=1}^N$:

$$\text{fit the MLE: } \theta^\star \in \arg\max \left[N^{-1} \sum_{i=1}^{N} \log p_\theta(x_i) \right] \ .$$

- This doesn't look promising since $p_\theta(x_i)$ are intractable...

The evidence lower bound (ELBO)

- A first option to approximate $p_\theta(x_i)$ is to use importance sampling.
- Using Jensen inequality, we can show that for any condition distribution q:

$$\log p_\theta(x) = \log \int \frac{p_\theta(x,z)}{q(z|x)} q(z|x)\mathrm{d}z \geq \int \log \left[\frac{p_\theta(x,z)}{q(z|x)}\right] q(z|x)\mathrm{d}z \; .$$

- The RHS is a lower-bound on the marginal log-likehood, referred to as ELBO MacKay (1992).
- This is this quantity that we maximize through an empirical version:

$$\int \log \left[\frac{p_\theta(x,z)}{q(z|x)}\right] q(z|x)\mathrm{d}z \approx M^{-1} \sum_{i=1}^{M} \log \left[\frac{p_\theta(x_i,z)}{q(z_i|x)}\right] \; ,$$

with $z_i \overset{\text{iid}}{\sim} q(\cdot|x)$.

Do we lose something compared to usual MLE?

- Define the ELBO:

$$\mathrm{ELBO}(x_{1:N};\theta,q) = N^{-1}\sum_{i=1}^{N}\int \log\left[\frac{p_\theta(x_i,z)}{q(z|x_i)}\right] q(z|x_i)\mathrm{d}z\ .$$

- This satisfies:

$$N^{-1}\sum_{i=1}^{N}\log p_\theta(x_i) \geq \mathrm{ELBO}(x_{1:N};\theta,q)\ ,$$

  with equality iif:

$$q(z|x) = p_\theta(z|x)\ .$$

- Therefore, the MLE can be re-written:

$$\begin{aligned}\theta^\star &\in \arg\max\left[N^{-1}\sum_{i=1}^{N}\log p_\theta(x_i)\right] \\ &\in \arg\max_{\theta,q}\mathrm{ELBO}(x_{1:N};\theta,q)\ .\end{aligned}$$

- How to estimate/learn $q(z|x) \approx p_\theta(z|x)$?

Posterior inference

- How to estimate/learn $q(z|x) \approx p_\theta(z|x)$?
- Let $q_\phi(z|x)$ be a family of density estimators with parameters ϕ.
- People sometimes call this amortized inference.
- The approximate problem, we consider then

$$\text{maximize over } \theta, \phi \text{ the function } \mathrm{ELBO}(x_{1:N}; \theta, \phi) \ ,$$

  writing $\mathrm{ELBO}(x_{1:N}; \theta, \phi) = \mathrm{ELBO}(x_{1:N}; \theta, q_\phi)$.
- This option has been popularized in Kingma and Welling (2013).

Stochastic optimization

- Recall that

$$\mathrm{ELBO}(x_{1:N}; \theta, \phi) = N^{-1} \sum_{i=1}^{N} f_i(\theta, \phi) \ , \quad f_i = \int \log \left[\frac{p_\theta(x_i, z)}{q_\phi(z|x_i)} \right] q_\phi(z|x_i) \mathrm{d}z \ .$$

- How to optimize this function over θ, ϕ ?
- Solution: SGD
- For any i, we have the Monte Carlo approximation:

$$\nabla_\theta f_i(\theta, \phi) \approx M_i \sum_{k=1}^{M_i} \nabla_\theta \log \left[\frac{p_\theta(x_i, z_k)}{q_\phi(z_k|x_i)} \right]$$

with $z_k \overset{\text{iid}}{\sim} q_\phi(\cdot|x_i)$.

- However, the computation/approximation of $\nabla_\phi f_i(\theta, \phi)$ is not that easy.

The reparametrization Trick

- Recall that

$$\mathrm{ELBO}(x_{1:N};\theta,\phi) = N^{-1}\sum_{i=1}^{N} f_i(\theta,\phi)\ , \quad f_i = \int \log\left[\frac{p_\theta(x_i,z)}{q_\phi(z|x_i)}\right] q_\phi(z|x_i)\mathrm{d}z\ .$$

- To estimate $\nabla_\phi f_i(\theta,\phi)$, we always assume that $q_\phi(\cdot|x)$ is the distribution of some random variable

$$\mathrm{T}_\phi(x,\varepsilon)\ , \quad \varepsilon \sim r_0\ ,$$

  for some samplable distribution r_0.
- Example: $\mu_\phi(x) + \sigma_\phi(x)\varepsilon$, $\varepsilon \sim \mathcal{N}(0,\mathrm{Id})$
- This is called the reparametrization Trick.
- In that case:

$$\nabla_\phi f_i(\theta,\phi) \approx K_i^{-1}\sum_{k=1}^{K_i}\nabla_\phi \log\left[\frac{p_\theta(x_i,\mathrm{T}_\phi(x,\varepsilon_k))}{q_\phi(\mathrm{T}_\phi(x,\varepsilon_k)|x_i)}\right]\ ,$$

  with $\varepsilon_k \overset{\text{iid}}{\sim} r_0$.

References

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

J. Wolfowitz. The minimum distance method. *Ann. Math. Statist.*, 28(1):75–88, 03 1957. doi: 10.1214/aoms/1177707038.