ML Project: Generative modeling

Ricardo Baptista, Suvedei Soyolerdene, Giulio Trigila, and Tanya Wang

Caltech, Baruch College CUNY, and NYU

PolyMathJr Summer Program

June 20, 2025

Probabilistic modeling

- Given data $\{\mathbf{z}^i\}$ sampled from an unknown probability density ρ, we would like to estimate $\rho(\mathbf{z})$
- Example: what is the distribution of the heights among the high school students in NYC?
- A more complicated example: given samples $\{(\mathbf{z}^i, \mathbf{y}^i)\}$ we would like to estimate the conditional probability density $\rho(\mathbf{z}|\mathbf{y})$
- Example: given that New York is located at 40.7128° N, 74.0060° W on the sea level and that tomorrow is June 22 (these are the factors $\mathbf{y}_i \in \mathbb{R}^m$) what is the probability that the temperature (i.e., the outcome $z \in \mathbb{R}$) is going to be 25 degrees Celsius?

We use probabilistic models to describe certain outputs (e.g. from a physical system) and make predictions about the future.

- **Applications**: Numerical weather prediction, GPS tracking, infectious disease spread, financial market analysis, etc.

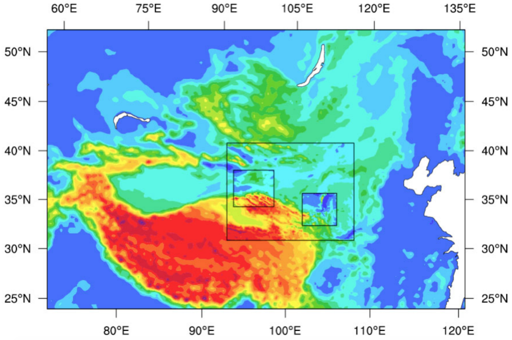

Figure: Source: NCAR ensemble wind forecast

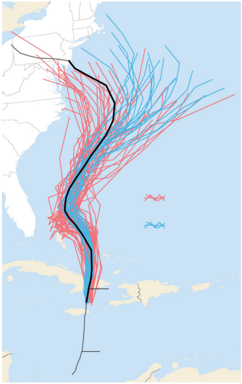
Figure: Source: Wall Street Journal

Main idea behind Mapping Methods

Most of the work related to Mapping methods is based on KL minimization

$$\boxed{KL[\rho||\mu] = \int \log(\rho(x)/\mu(x))\rho(x)dx}$$

- We want to find a map from ρ, known only through samples $\{x_i\}$, to a reference density $\mu = \mathcal{N}(0, I)$.
- Let $\mu = T_{\#}\tilde{\rho}$ and consider $KL[\rho||\tilde{\rho}] = KL[T] = \int \rho \log(\rho/\tilde{\rho})$
- Parametrize $T = T_\beta$, minimize $KL[T_\beta] = \text{const} - \int \rho \log(\tilde{\rho})$ over the parameters $\beta \Rightarrow \bar{\beta}$
- Use the change of variable formula to estimate ρ:

$$\rho(x) = J^G(x)\mu(G(x)) \;\text{ where }\; G = T_{\bar{\beta}}$$

∗Tabak, Esteban G., and Eric Vanden-Eijnden. "Density estimation by dual ascent of the log-likelihood." Communications in Mathematical Sciences 8.1 (2010): 217-233.
∗∗El Moselhy, Tarek A., and Youssef M. Marzouk. "Bayesian inference with optimal maps." Journal of Computational Physics 231.23 (2012): 7815-7850.

Mapping method

- So far we need an a-priori parametrization T_β rich enough to map ρ to a reference pdf (standard normal, say)
- Coming up with T_β is not an easy task $\rightarrow$ deep neural network
- The task becomes even harder if we want to impose specific characteristics on T_β like being optimal (theory of optimal transport) or triangular

$$\Downarrow$$

Normalizing Flows

Normalizing Flows (NF)

Main idea

Find $T = T_n \circ ... \circ T_1$ as a composition of elementary maps, easier to parametrize $T_k = T_{\beta^k}$

- Find T descending $KL[T(.,t)] = \int \rho \log(\rho/\tilde{\rho}_t)$:

$$\frac{dT(x,t)}{dt} = -\frac{\delta KL[\phi \circ T(.,t)]}{\delta \phi}\bigg|_{\phi=id} \tag{1}$$

where $T(x, t = 0) = x$ and $\tilde{\rho}_t = J^{T(x,t)}\mu(T(x,t))$

In the end we have that $\tilde{\rho}_{t=0} = \mu$ and $\tilde{\rho}_\infty = \rho$

* Tabak, Esteban G., and Cristina V. Turner. "A family of nonparametric density estimation algorithms." Communications on Pure and Applied Mathematics 66.2 (2013): 145-164.
**Rezende, Danilo, and Shakir Mohamed. "Variational inference with normalizing flows." International conference on machine learning. PMLR, 2015.

Introduction: Normalizing Flows (NF)

In practice:

- The flow
$$\frac{dT(x,t)}{dt} = -\frac{\delta KL[\phi \circ T(.,t)]}{\delta \phi}\bigg|_{\phi=id} \tag{2}$$
is time discretized: $y^{n+1} = y^n + \phi(y^n, \theta^n)$ where ϕ is a perturbation of the identity map
- The resulting map $T(\cdot, t^n) = \phi^n \circ \phi^{n-1} \circ ... \circ \phi^1$ is the composition of elementary maps $\phi^k = \phi(\cdot, \theta^k)$

Why it is useful?

- Normalization, positivity constraints, curse of dimensionality and over-resolution make the parametrization of ρ a hard task.
- We don't need to parametrize ρ, we rather parametrize elementary maps that, through composition, form a rich family of one-to-one functions we use to recover ρ.

Optimal transport framework to map PDFs

Problem

How to move a pile of sand to fill a pit of the same volume minimizing a given cost?

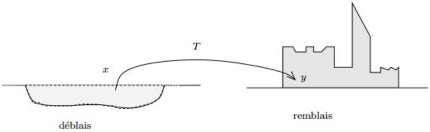

Fig. 3.1. Monge's problem of déblais and remblais

<u>Math formulation</u>

Given two probability densities, $\rho(\mathbf{z})$ and $\mu(\mathbf{z})$, $\mathbf{z} \in \mathbb{R}^d$, find a 1-to-1 map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_{\#\rho} = \mu$ and that minimizes the functional

$$M[T] = \int_{\mathbb{R}^d} c(\mathbf{z}, T(\mathbf{z}))\rho(\mathbf{z})d\mathbf{z}$$

The minimizer is called an *optimal transport map*.

Quadratic cost

Consider quadratic cost:

$$M[T] = \int_{\mathbb{R}^d} |\mathbf{z} - T(\mathbf{z})|^2 \rho(\mathbf{z}) d\mathbf{z}$$

If ρ and μ are smooth enough and have finite second moment, the optimal (Brenier) map exists, is unique and is given by the gradient of a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfying the Monge-Ampere (MA) equation:

$$\rho(\mathbf{z}) = \mu(\nabla\phi(\mathbf{z})) \det(D^2\phi(\mathbf{z}))$$

One way of finding the optimal map from ρ to μ is to solve the MA equation enforcing convexity of the solution

One idea on OT

We need to impose the convexity of the potential $\rightarrow$ we want the map to be the gradient of a convex function. If we build the flow

$$z^{k+1} = z^k + \beta^k \nabla_x F(z^k) = z^0 + \sum_{i=1}^{k} \nabla_x F(z^i)$$

with $z^0 = x$ the starting position, $\beta \in R$, and F convex, then the convexity of the potential depends on the sign of β.

- When β^i is positive there is no problem (sum of convex functions is still convex)
- If β is negative we need to pay attention
- We only have the value $\sum_{i=1}^{k} \nabla_x F(z^i)$ at the sample points z^i
- A condition that $\sum_{i=1}^{k} \nabla_x F(z^i)$ should satisfy is that, at least, should interpolate a convex function

Research opportunity

Main goal of this project

Given samples $(\mathbf{z}^i)$ drawn from the unknown $\rho(\mathbf{z})$ we want to build a new generative model that map samples from $\rho_0(\mathbf{z})$ to $\rho(\mathbf{z})$.

Avenues for numerical exploration:

- Implement the data-driven OT flow algorithm with adaptive bandwidth and kernel
- Evaluate convergence when using an improved back-and-forth procedure

Avenues for theoretical exploration:

- Show the equivalence between the data-driven and maximum mean discrepancy flow
- Mathematical understanding of back-and-forth procedure

Delve into applications:

- Evaluation on 2D benchmark problems (banana, checkerboard)
- Image generation and solving Bayesian inference problems

Thank you!
Questions?

References

- Marzouk et. al, An introduction to sampling via measure transport, Handbook of UQ, 2016
- Papamakarios et. al, Normalizing Flows for Probabilistic Modeling and Inference, JMLR, 2021
- Santambrogio, Optimal Transport for Applied Mathematicians, Springer, 2015
- Tabak E.G., Trigila G. Data-driven optimal transport, Communications on Pure and Applied Math, 2016
- Galashov, de Bortoli, Gretton, Deep MMD Gradient Flow without adversarial training, 2024