1908.09257: Normalising Flow: An Introduction and Review of Current Methods

INTRODUCTION

- model a probability distribution given samples drawn from that distribution.

- methods for generative modeling
  - introduce latent variables to explain the observed data
    - provide additional flexibility
    - can increase the complexity of learning and inference
    - Methods
      - Variational approaches
      - expectation maximization
  - Direct analytic approaches
  - Graphical models
    - explicitly model the conditional dependence between random variables
  - generative neural approaches
    - GAN and VAE
      - GANs
      - VAEs
      - impressive performance
      - issues
        - Neither allows for exact evaluation of the probability density of new points
        - training can be challenging
          - mode collapse
          - posterior collapse
          - vanishing gradients
          - training instability
  - Normalizing Flows
    - distribution learning
    - both sampling and density evaluation can be efficient and exact

BACKGROUND

- What is a Normalizing Flow
  - transformation of a simple probability distribution (e.g., a standard normal) into a more complex distribution by a sequence of invertible and differentiable mappings.

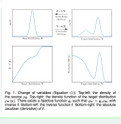
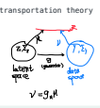


Fig. 1. Change of variables (Equation (1)). Top left: the density of an RV Z is given by the density function of the layer distribution $p_Z(z)$. Top right: the density function of $Y = 2Z + 3$ is $p_Y(y) = p_Z(\frac{y-3}{2})/2$. Bottom left: the inverse function f^{-1}. Bottom right: the absolute Jacobian (derivative) of f.

  - choosing an initial density and then chaining together some number of parameterized, invertible and differentiable transformations
- Basics
  - $p_Y(\mathbf{y}) = p_Z(\mathbf{f}(\mathbf{y})) \, |\det \mathbf{D}\mathbf{f}(\mathbf{y})|$
    $= p_Z(\mathbf{f}(\mathbf{y})) \, |\det \mathbf{D}\mathbf{g}(\mathbf{f}(\mathbf{y}))|^{-1}$
  - normalizing direction
    - from a complicated and irregular data distribution towards the simpler, more regular or "nor- mal" form, of the base measure pZ
  - generate any distribution pY from any base distribution pZunder reasonable assumptions on the two distributions.
    - if the transformation g can be arbitrarily complex
  - More formal construction
    - $(\mathcal{Z}, \Sigma_{\mathcal{Z}})$, $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ are measurable spaces
      - Or Borel space
      - Z: set
      - Σ: σ-algebra/σ-field, e.g., power
    - $\mathbf{g}_* \mu(U) = \mu(\mathbf{g}^{-1}(U)), \quad \text{for all } U \in \Sigma_{\mathcal{Y}}.$

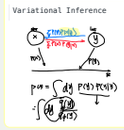
    - Generative Model
      - transportation theory



      - "data" space $(\mathcal{Y}, \Sigma_{\mathcal{Y}}, \nu)$.
        - Training data is a sample of the data space.
      - introduce a simpler measured space
        - $(\mathcal{Z}, \Sigma_{\mathcal{Z}}, \mu)$
      - find a function $\mathbf{g} : \mathcal{Z} \to \mathcal{Y}$,
        - $\nu = \mathbf{g}_* \mu.$
        - g can be interpreted as a "generator"
        - Z as a latent space
    - Z = RD,
    - ***Definition 2.*** A function $\mathbf{g} : \mathbb{R}^D \to \mathbb{R}^D$ is called a diffeomorphism, if it is bijective, differentiable, and its inverse is differentiable as well.

Applications

- Density estimation and sampling
  - Maximum Likelihood Estimation
    - base measure, pZ
      - parameterized by the vector φ.
    - data
      - $\mathcal{D} = \{\mathbf{y}^{(i)}\}_{i=1}^{M},$
    - perform likelihood-based estimation of the parameters
      - $\Theta = (\theta, \phi).$
        - $$\begin{aligned} \log p(\mathcal{D}|\Theta) &= \sum_{i=1}^{M} \log p_{\mathbf{Y}}(\mathbf{y}^{(i)}|\Theta) \qquad (5) \\ &= \sum_{i=1}^{M} \log p_{\mathbf{Z}}(\mathbf{f}(\mathbf{y}^{(i)}|\theta)|\phi) + \log \left|\det \mathbf{D}\mathbf{f}(\mathbf{y}^{(i)}|\theta)\right| \end{aligned}$$
    - During training, the parameters of the flow (θ) and of the base distribution (φ) are adjusted to maximize the log-likelihood.
    - for density estimation it is common to model a flow in the normalizing direction
  - adversarial losses can be used with normalizing flow models
    - Flow-GAN
- Variational Inference



  - latent variable model $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y}$
    - $\mathbf{x}$ is an observed variable $\mathbf{y}$ the latent variable.
  - posterior distribution $p(\mathbf{y}|\mathbf{x})$
    - estimating the parameters of the model
    - its computation is usually intractable in practice
  - use variational inference and introduce the approximate posterior
    - $q(\mathbf{y}|\mathbf{x}, \theta)$
      - θ are parameters of the variational distribution
      - this distribution should be as close to the real posterior as possible.
    - First thoughts
      - minimizing the KL divergence
      - maximizing the evidence lower bound
        - with gradient descent
  - reparametrize $q(\mathbf{y}|\mathbf{x}, \theta) = p_{\mathbf{Y}}(\mathbf{y}|\theta)$
    - with normalizing flows

METHODS

- to be practical
  - be invertible;
  - be sufficiently expressive to model the distribution of interest
  - be computationally efficient,
- different types of flows
  - Elementwise Flows
    - bijective scalar function h(x)

    - $\mathbf{g}(\mathbf{x}) = (h(x_1), h(x_2), \ldots, h(x_D))^T$
    - h_τ could be viewed as an "activation function".
  - Linear Flows
    - $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$
    - limited in their expressiveness
      - $y = Ax + b \sim \mathcal{N}(\mu, \Sigma)$
      - $(x-\mu)^T \Sigma^{-1} (x-\mu)$
    - $p_{\mathbf{Z}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}, \mu, \Sigma)$
      - $p_{\mathbf{Y}} = \mathcal{N}(\mathbf{y}, \mathbf{A}\mu + \mathbf{b}, \mathbf{A}^T \Sigma \mathbf{A}).$
    - Factorizations
      - $\mathbf{g}(\mathbf{x}) = \mathbf{PLUx} + \mathbf{b}$
    - Convolution
  - Planar and Radial Flows
    - Planar Flows
      - $\mathbf{g}(\mathbf{x}) = \mathbf{x} + \mathbf{u}h(\mathbf{w}^T\mathbf{x} + b).$
    - Radial Flows
      - $\mathbf{g}(\mathbf{x}) = \mathbf{x} + \frac{\beta}{\alpha + ||\mathbf{x} - \mathbf{x}_0||}(\mathbf{x} - \mathbf{x}_0)$

  - Coupling and Autoregressive Flows
    - Coupling Flows
      - $\mathbf{y}^A = \mathbf{h}(\mathbf{x}^A; \Theta(\mathbf{x}^B))$
        $\mathbf{y}^B = \mathbf{x}^B,$
        - conditioner
          - $\Theta(\mathbf{x}^B)$
    - Autoregressive Flows
      - $y_t = h(x_t; \Theta_t(\mathbf{x}_{1:t-1})),$
        - $y_t = x_t + \sigma_t(x_{1:t-1}) + \mu_t(x_{1:t-1})$
      - IAF
        - $y_t = h(x_t; \theta_t(\mathbf{y}_{1:t-1})).$
      - several autoregressive flows the universality property has been proven
    - Coupling Functions
  - Residual Flows
  - Infinitesimal (Continuous) Flows
    - ODE-based methods
    - Langevin flows