Regularized Optimal Transport

From Computational and Statistical Perspectives

Tao Li, Yuan Ni, Michael Stanley, Nikhil Supekar

New York University

December 23, 2020

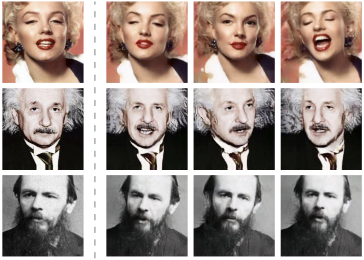

Figure 1: Deepfake images (right) generated from single, original (left)

- **Deepfakes** generates new facial expressions on existing images.
- **Generative adversarial networks (GANs)** learn a distribution from training data

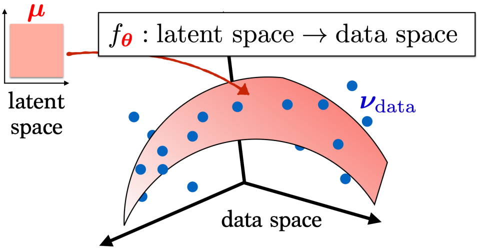

Figure 2: A generative model is a mapping from the latent space to the data space (Peyré and Cuturi 2019)

- ν_{data}: real faces
- f_θ: a parameterized mapping, e.g., deep neural networks
- $f_\theta(\mu)$: deepfakes

Discrepancy Function

Measuring how close $f_\theta(\boldsymbol{\mu})$ is to $\boldsymbol{\nu}_{\text{data}}$

- Total Variation
- Kullback-Leibler (KL) divergence, Jensen-Shannon (JS) divergence
- Wasserstein Distance (Arjovsky, Chintala, and Bottou 2017)

$$W(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) \tag{OT}$$

Training problem:

$$\min_\theta W(f_\theta(\boldsymbol{\mu}), \boldsymbol{\nu})$$

- $W(f_\theta(\boldsymbol{\mu}), \boldsymbol{\nu})$ is not smooth with respect to θ.
- Though $\nabla_\theta W(f_\theta(\boldsymbol{\mu}), \boldsymbol{\nu})$ can be approximated, such methods fail to converge sometimes.(Gulrajani et al. 2017)(Bousquet et al. 2017)
- Adding regularization to make it smooth (Sanjabi et al. 2018)

$$W_\varepsilon(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\pi \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \underbrace{\int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y)}_{\text{original } eq. \text{ (OT)}} + \overbrace{\varepsilon \mathrm{KL}(\pi \mid \boldsymbol{\mu} \otimes \boldsymbol{\nu})}^{\text{entropy regulzrizer}} \tag{ROT}$$

Computational Aspect

Optimal Transport

$$\min_{\pi \in \Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) \qquad \text{(Primal)}$$

$$\max_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x)\mathrm{d}\boldsymbol{\mu} + \int_{\mathcal{Y}} v(y)\mathrm{d}\boldsymbol{\nu} \qquad \text{(Dual)}$$

$$\text{subject to } u(x)+v(y) \leq c(x,y)$$

Regularized Optimal Transport

$$\min_{\pi \in \Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi \mid \boldsymbol{\mu}\otimes\boldsymbol{\nu}) \qquad \text{(Primal)}$$

$$\max_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x)\mathrm{d}\boldsymbol{\mu} + \int_{\mathcal{Y}} v(y)\mathrm{d}\boldsymbol{\nu} \qquad \text{(Dual)}$$

$$-\varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\nu}$$

Define

$$f_\varepsilon(x, y, u, v) := u(x) + v(y) - \varepsilon \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right)$$

Dual Formulation

$$\max_{u \in \mathcal{C}(\mathcal{X}), v \in \mathcal{C}(\mathcal{Y})} \mathbb{E}_{X \sim \mu, Y \sim \nu}[f_\varepsilon(X, Y, u, v)]$$

- Alternating maximization: Sinkhorn algorithm (Cuturi 2013)
- Stochastic Gradient Descent (Aude et al. 2016)
- More methods from stochastic programming

Wasserstein distance on empirical measures:

$$W(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}}), \quad \hat{\boldsymbol{\mu}} = 1/n \sum_i^n \delta_{x_i}, \hat{\boldsymbol{\nu}} = 1/m \sum_j^m \delta_{y_j},$$

Curse of dimensionality

For $\mathbb{R}^d, d \geq 3$, $\mathbb{E}[|W(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}}) - W(\boldsymbol{\mu}, \boldsymbol{\nu})|] = O(n^{-1/d})$

Regularization for breaking the curse

$$\mathbb{E}\left|W_\varepsilon(\alpha, \beta) - W_\varepsilon\left(\hat{\alpha}_n, \hat{\beta}_n\right)\right| = \quad O\left(\frac{e^{\frac{\kappa}{\varepsilon}}}{\varepsilon^{\lfloor d/2 \rfloor}\sqrt{n}}\right) \text{ as } \varepsilon \to 0$$

$$\mathbb{E}\left|W_\varepsilon(\alpha, \beta) - W_\varepsilon\left(\hat{\alpha}_n, \hat{\beta}_n\right)\right| = O\left(\frac{1}{\sqrt{n}}\right) \quad \text{ as } \varepsilon \to +\infty$$

[Klatt, Tameling, and Munk 2019]

As $m, n \to \infty$

$$\sqrt{\frac{nm}{n+m}}\{W_{p,\varepsilon}(\hat{\mu}_n, \hat{\nu}_m) - W_{p,\varepsilon}(\mu, \nu)\} \xrightarrow{D} \mathcal{N}_1(0, \sigma^2_{p,\varepsilon}(\mu, \nu))$$

- Asymptotic to a Gaussian
- Empirical Sinkhorn Divergence

[Sommerfeld and Munk 2017]

Under the null hypothesis $\boldsymbol{\mu} = \boldsymbol{\nu}$, as $m, n \to \infty$

$$\sqrt{\frac{mn}{m+n}}^{\frac{1}{p}} W_p(\hat{\mu}_n, \hat{\nu}_m) \xrightarrow{D} \{\max_u \langle G, u\rangle\}^{1/p}$$