Improved Autoregressive Modeling with Distribution Smoothing

Mikhail Kuznetsov

CMC MSU

February 8, 2021

Let $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ — D-dimensional i.i.d samples from a continuous data distribution $p_{\text{data}}(\boldsymbol{x})$

Property

An autoregressive model decomposes a joint distribution into univariate conditionals [2]:

$$p(\boldsymbol{x}) = \prod_{i=1}^{D} p\left(x_i \mid \boldsymbol{x}_{<i}\right)$$

Goal

Find θ — parameters of the model such that $p_\theta(\boldsymbol{x}) \approx p_{\text{data}}(\boldsymbol{x})$

A commonly used approach for density estimation is maximum likelihood estimation (MLE), i.e., by maximizing

$$L(\theta) \triangleq \frac{1}{N} \sum_{i=1}^{N} \log p_\theta\left(\boldsymbol{x}_i\right)$$

General idea

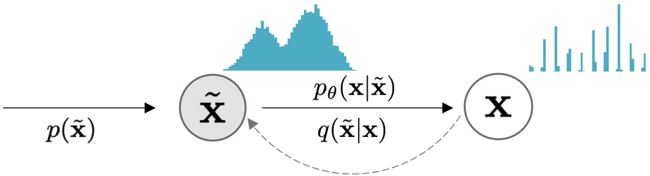


Figure: Overview of the method
($\tilde{x}$ — the smoothed data, $q(\tilde{x} \mid x)$ — the smoothing distribution) [1]

Problem 1: Manifold hypothesis

Many real world data distributions (e.g.natural images) **may lie in the vicinity of a low-dimensional manifold** and can often **have complicated densities with sharp transitions** (i.e. high Lipschitz constants) that are difficult to model.

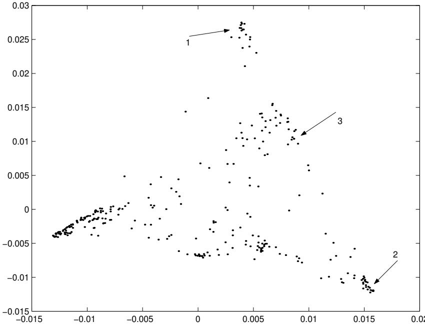


Figure (a): The 300 most frequent words of the Brown corpus represented in the spectral domain.

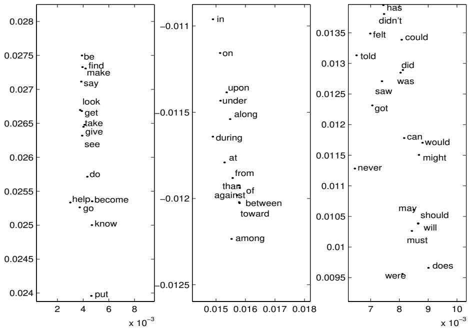


Figure (b): Fragments labeled by arrows: (left) infnitives of verbs, (middle) prepositions, and (right) mostly modal and auxiliary verbs.

Source: (Belkin M., Niyogi P., 2003) [3]

Problem 1: Manifold hypothesis

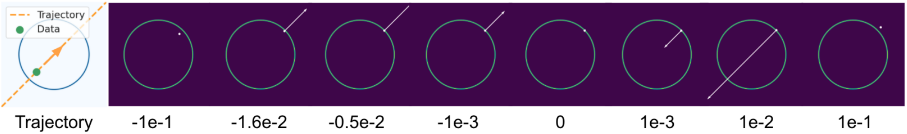


Figure: Ring distribution (almost a unit circle) formed by rotating the 1-d Gaussian distribution $\mathcal{N}\left(1, 0.01^2\right)$ around the origin. The data location for each figure is $(\sqrt{0.5}+c, \sqrt{0.5}+c)$, where c is the number below each figure and $(\sqrt{0.5}, \sqrt{0.5})$ is the upper right intersection of the trajectory with the unit circle. [1]

Problem 2: Compounding errors

$$p(\boldsymbol{x}) = \prod_{i=1}^{D} p\left(x_i \mid \boldsymbol{x}_{<i}\right)$$

Compounding errors **comes from the inaccurate approximation of the conditional distributions.**

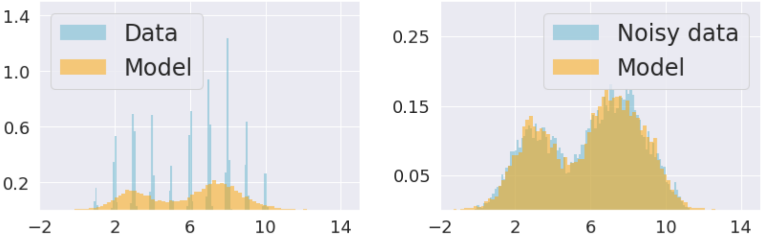
The reasons for this may be:

- Curse of dimensionality + very limited amount of training data (in some cases)
- The current state is based on the values of the previous states
- Adversarial attacks

Theorem 1 Let:

1. $p(x)$ — a continuous 1-d distribution that is supported on $\boldsymbol{R}$
2. $q(\tilde{x} \mid x)$ — 1-d distribution that is:
   - symmetric (i.e. $q(\tilde{x} \mid x) = q(x \mid \tilde{x})$)
   - stationary (i.e. translation invariant)
   - $\lim_{x\to\infty} p(x)q(x \mid \tilde{x}) = 0$ for any given $\tilde{x}$

Then: $\mathrm{Lip}(q(\tilde{x})) \leq \mathrm{Lip}\left(p_{\text{data}}\ (x)\right)$,
where $q(\tilde{x}) \triangleq \int q(\tilde{x} \mid x)p(x)dx$ and $\mathrm{Lip}(\cdot)$ denotes the Lipschitz constant of the given $1 - d$ function.



(a) Data (b) Smoothed data

Figure: Illustration of the Theorem 1 [1]

Proposition 1 (Informal) Let:

1. $q(\tilde{\boldsymbol{x}} \mid \boldsymbol{x})$ is such that:
    - symmetric
    - stationary
    - has small variance
    - has negligible higher order moments (i.e. very small)

Then:

$$\boldsymbol{E}_{p_{\text{data}}(\boldsymbol{x})} \boldsymbol{E}_{q(\tilde{\boldsymbol{x}}|\boldsymbol{x})} [\log p_\theta(\tilde{\boldsymbol{x}})] \approx \boldsymbol{E}_{p_{\text{data}}(\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}) + \frac{\eta}{2} \sum_i \frac{\partial^2 \log p_\theta}{\partial x_i^2} \right]$$

for some constant η.

Since the samples from p_{data} should be close to a local maximum of the model, this encourages the second order gradients computed at a data point $\boldsymbol{x}$ to become closer to zero (if it were positive then $\boldsymbol{x}$ will not be a local maximum), creating a smoothing effect.

1. Sing-step recovering
2. Obtaining ELBO

[1] Improved autoregressive modeling with distribution smoothing.
https://openreview.net/pdf?id=rJA5Pz7lHKb, 2021.

[2] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma.
Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications.
arXiv:1701.05517, 2017.

[3] Mikhail Belkin, Partha Niyogi.
Laplacian eigenmaps for dimensionality reduction and data representation.
https://www2.imm.dtu.dk/projects/manifold/Papers/Laplacian.pdf, 2003.