Generalization and Memorization: The Bias-potential Model

Hongkang Yang and Weinan E

Princeton University

MSML 2021

Theme

Distribution learning

Understand generalization ability

Reconcile with memorization and curse of dimensionality

Simple setting: Bias-potential model

Challenges

Notation:

Target distribution Q_* and empirical distribution $Q_*^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}$

1. Curse of dimensionality

$$W_2(Q_*, Q_*^{(n)}) = n^{-O(1/d)}$$

   Worst case lower bound for all models

2. Memorization

$$\lim_{t\to\infty} Q_t \to Q_*^{(n)}$$

   Model becomes trivial

Need a dimension-independent α

$$W_2(Q_*, Q_t) \text{ or } \mathsf{KL}(Q_* \| Q_t) = O(n^{-\alpha})$$

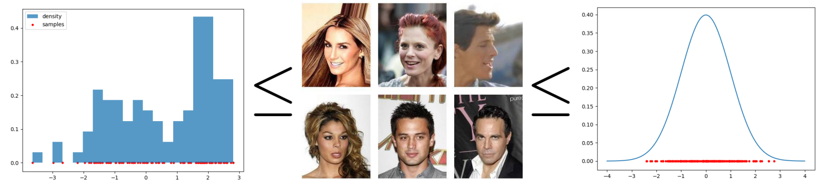


Figure 1: [1] Between universal approximation and strong regularity

[1]Source: the CelebA dataset.

Solution

Framework

Continuous perspective

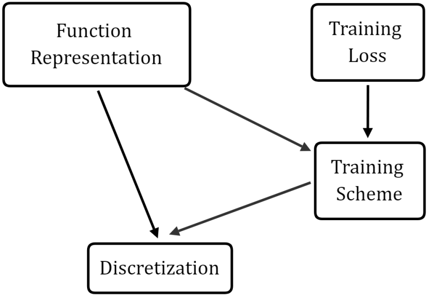
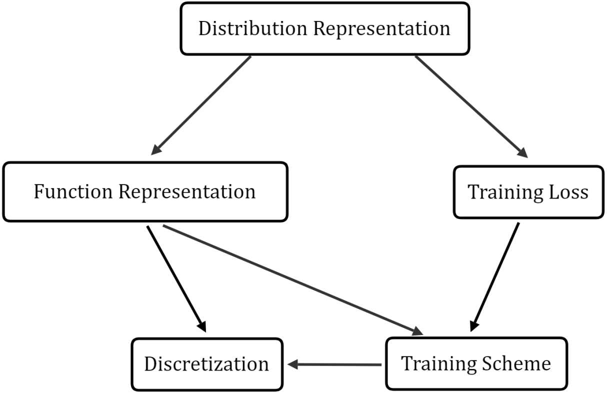


Figure 2: Supervised learning [E, Ma & Wu, 2020]

Distribution learning



1. Distribution representation:
   Bias-potential model

$$Q = \frac{1}{Z} e^{-V} P, \quad Z = \mathbb{E}_P[e^{-V}]$$

2. Training loss:
   Relative entropy

$$\mathrm{KL}(Q_* \| Q) = \mathbb{E}_{Q_*}[V] + \log \mathbb{E}_P[e^{-V}] + \text{constant}$$

3. Function representation:

- 2-layer network (integral transform)

$$V(\mathbf{x}) = \mathbb{E}_{\rho(a,\mathbf{w},b)}[a \ \sigma(\mathbf{w} \cdot \mathbf{x} + b)]$$

- Residual network (flow)

$$V(\mathbf{x}_0) = l(\mathbf{x}_1), \quad \dot{\mathbf{x}}_t = \mathbb{E}_{\rho_t(a,\mathbf{w},b)}[a \ \sigma(\mathbf{w} \cdot \mathbf{x} + b)]$$

- Random feature function (or kernel function)

$$V(\mathbf{x}) = \mathbb{E}_{\rho_0(\mathbf{w},b)}[a(\mathbf{w},b)\sigma(\mathbf{w} \cdot \mathbf{x} + b)]$$

RKHS norm

$$\|V\|_{\mathcal{H}} := \|a\|_{L^2(\rho_0)}$$

Rademacher complexity

$$Rad_n(\{\|V\|_{\mathcal{H}} \leq R\}) \leq 2R\frac{\sqrt{2\log 2d}}{\sqrt{n}}$$

4. Training rule

Parameter a_t and distribution Q_t

Gradient flow

$$\frac{d}{dt}a_t = -\frac{\delta L}{\delta a} = \int \sigma(\mathbf{w} \cdot \mathbf{x} + b)\ d(Q_t - Q_*)(\mathbf{x})$$
$$L(a) = \mathbb{E}_{Q_*}[V] + \log \mathbb{E}_P[e^{-V}]$$

Empirical loss

$$L^{(n)}(a) = \mathbb{E}_{Q_*^{(n)}}[V] + \log \mathbb{E}_P[e^{-V}]$$

Empirical training trajectory: $a_t^{(n)}$ and $Q_t^{(n)}$.

Universal Approximation

Proposition

(Under technical conditions), if $\mathcal{V}$ has universal approximation property among continuous functions, then the family

$$\mathcal{Q} = \left\{ \frac{1}{Z} e^{-V} P \;\middle|\; V \in \mathcal{V} \right\}$$

satisfies universal approximation among probability distributions, under KL-divergence, TV norm and Wasserstein metric.

Generalization Error

Theorem

(Under technical conditions), suppose the target distribution is given by $Q_* \propto e^{-V_*} P$. *Then, with probability* $1-\delta$,

$$KL(Q_* \| Q_t^{(n)}) \leq \frac{\|V_* - V_0\|_{\mathcal{H}}^2}{2t} + \frac{8\sqrt{2\log 2d} + 2\sqrt{2\log 2/\delta}}{\sqrt{n}} t$$

Generalization error $\leq$ *Training error* $+$ *Generalization gap*

Corollary

Early-stopping at $T \asymp \|V_* - V_0\|_{\mathcal{H}} \left(\frac{n}{\log d}\right)^{1/4}$ *achieves error*

$$KL(Q_* \| Q_t^{(n)}) \lesssim \frac{\|V_* - V_0\|_{\mathcal{H}} (\log d)^{1/4}}{n^{1/4}}$$

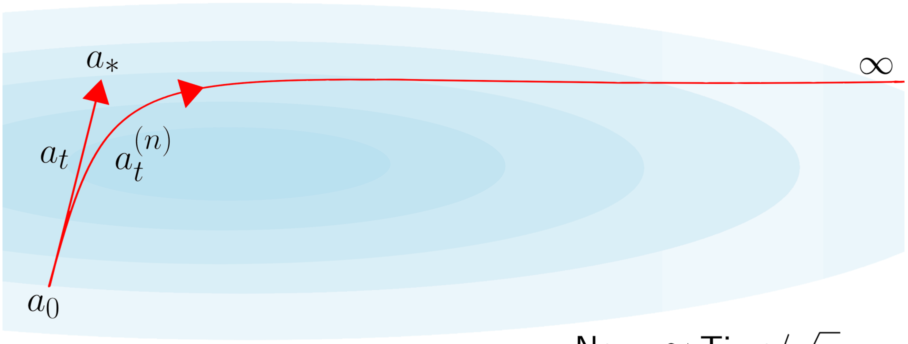
Mechanism for Generalization

The sampling gap $Q_* - Q_*^{(n)}$ is hidden by function representation

$$\frac{\delta(L - L^{(n)})}{\delta a} = \int \frac{\delta(L - L^{(n)})}{\delta V} \frac{\delta V}{\delta a} d\mathbf{x}$$
$$= \langle Q_* - Q_*^{(n)}, \sigma(\mathbf{w} \cdot \mathbf{x} + b) \rangle$$

So the trajectories remain close

$$\|a_t - a_t^{(n)}\|_{L^2(\rho_0)} \lesssim \frac{t}{\sqrt{n}}$$



Norm $\approx$ Time$/\sqrt{n}$

Memorization

Proposition

If $Q_t^{(n)}$ converges weakly to some limit, then the limit must be $Q_*^{(n)}$. The generalization error always blows-up

$$\lim_{t\to\infty} KL(Q_* \| Q_t^{(n)}) = \lim_{t\to\infty} \|a_t^{(n)}\|_{L^2(\rho_0)} = \infty$$

Memorization seems inevitable.

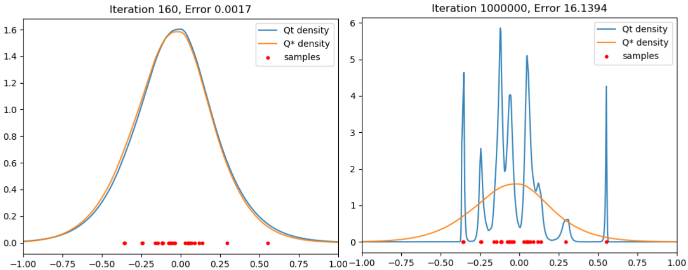


Figure 3: Left: Early stopping. Right: Memorization. (Training accelerated by Adam)

Difference from Supervised Learning

Regression with implicit regularization:

$$\min_{f \in \mathcal{H}} \|f_* - f\|^2_{L^2(P^{(n)})}, \quad P^{(n)} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{x}_i}$$

Generalization error bound [E, Ma, Wu, 2019]

$$\|f_* - f_t^{(n)}\|^2_{L^2(P)} \leq \frac{\|f_*\|^2_{\mathcal{H}}}{2t} + \frac{(1 + \sqrt{\log 1/\delta})\|f_*\|_{\mathcal{H}}}{\sqrt{n}}$$

Early stopping achieves error $O(n^{-1/2})$.

Memorization vs interpolation:

Regression: $\|a_t^{(n)}\|_{L^2(\rho_0)} = O(\|a_*\|)$

Bias-potential: $\|a_t^{(n)}\| = O(\|a_*\| + t/\sqrt{n})$

Analogous to regression with noise.

Contribution

- Reconcile generalization and memorization:
  Time scales and early stopping
- Mechanism of generalization:
  Complexity of function representation overcomes the curse of dimensionality
- Implication to distribution learning:
  How function representation influences training
  Our new paper "Generalization Error of GAN from the Discriminator's Perspective" [arXiv 2107.03633]