Resampling Based Estimation of Generative Models

Benedek Bálint Novák

Supervisor:
Balázs Csanád Csáji

Eötvös Loránd Tudományegyetem

2025 December 5.

Motivation

Generative models

Models that create i.i.d. samples from a distribution given some initial parameters:

- Diffusion models, ChatGPT
- Inverse cumulative distribution functions
- ARIMA models (for multiple trajectories)
- Scientific experiments

The Goal:

Given an original sample $S^{(0)}$, estimate the "parameters" that could have generated it.

Motivation

Standard methods for parameter estimation

- Maximum likelihood estimation
- Stochastic gradient descent
- Assumes knowledge about the model structure
- Provides only point estimates
- Asymptotic guarantees for confidence regions

Why bother?

- What if the data is generated from a black box?
  Instead: Distribution-free methods
- Different point estimates could lead to different interpretations (Rashomon-effect)
  Instead: Confidence regions
- Asymptotic guarantees don't work well for small sample sizes
  Instead: Exact guarantees for finite samples

Motivation

Assumptions

- There exists a "parameter" space Θ containing θ^*, that parametrizes the distribution from which the original sample $S^{(0)}$ is obtained from
- Instead of assuming the parametrization of the data generation process (e.g., knowing the likelihood function)
- There is a black box, that can generate i.i.d. samples $S_\theta^{(1)}, \ldots, S_\theta^{(m)}$ from $\mathbb{P}_\theta$ given any parameter θ.
- The seed for the black box can be fixed.

Estimating confidence regions

The Resampling framework

- Given a parameter θ
- 1. Generate $m-1$ alternative samples from $\mathbb{P}_\theta$
- 2. Assign a real number to each sample based on θ called its reference variable: $Z_\theta^{(i)} := T(S_\theta^{(i)}, \theta)$
- 3. Rank the samples based on the reference variables, and denote the rank of the original sample with

$$\mathcal{R}_\theta^{(m)} = 1 + \sum_{i=1}^{m-1} \mathbb{I}_{\{Z_\theta^{(i)} < Z_\theta^{(0)}\}}$$

Theorem (Csáji and Tamás, 2019)

$\mathbb{P}(\theta^* \in \{\theta \in \Theta \mid p \leq \mathcal{R}_\theta^{(m)} \leq q\}) = \frac{q-p+1}{m}$ if there is almost surely a strict ordering of the reference variables

Reference variables

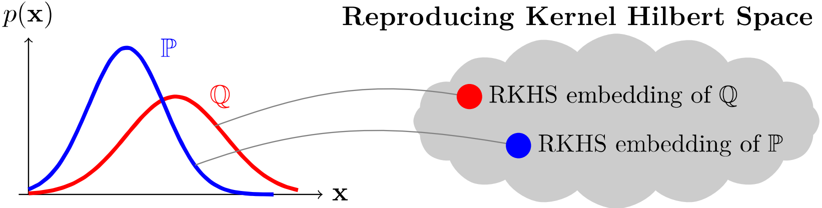
Reference variable $Z_\theta^{(i)} := T(S_\theta^{(i)}, \theta)$ depends on θ explicitly;
e.g., Maximum Likelihood based reference variable:

$$Z_\theta^{(i)} = ||\nabla_\theta \mathcal{L}(\theta, S_\theta^{(i)})||^2$$

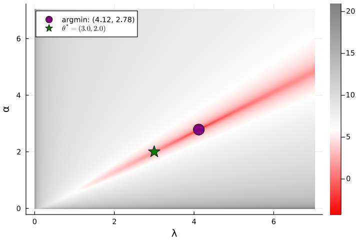
Introducing a seed component $Z_\theta^{(i)} := T(S_\theta^{(i)}, \theta, \xi_i)$;
e.g., MMD-based reference variable:

$$Z_\theta^{(i)} = \widehat{\mathrm{MMD}}^2[S_\theta^{(i)}, S_{\theta,\xi_i}^{(i+m)}]$$

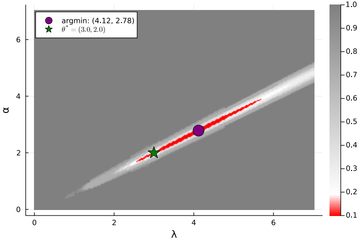
Where $\widehat{\mathrm{MMD}}^2$ is an unbiased for the Maximum Mean Discrepancy



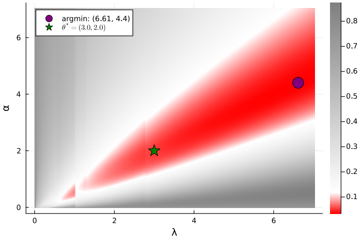
Examples for confidence regions (Gamma distribution)



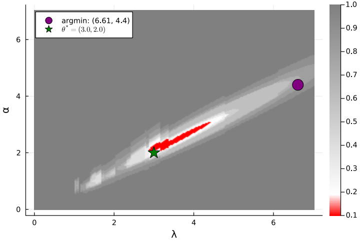
(a) ML reference variable



(b) ML Rank



(c) MMD reference variable



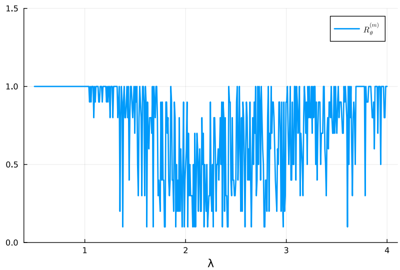
(d) MMD Rank

Point estimation: The smoothed rank

Idea:

$$\hat{\theta} \in \underset{\theta \in \Theta}{\text{argmin}}\ \mathcal{R}_\theta^{(m)}$$

Problem:



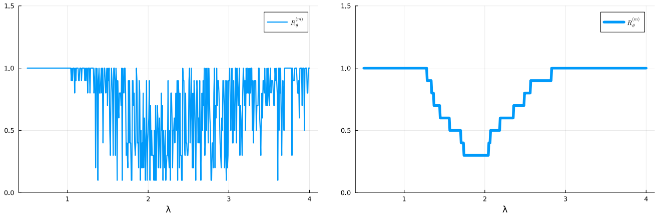
(a) Not fixing the seed

Point estimation: The smoothed rank

Idea:

$$\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \, \mathcal{R}_\theta^{(m)}$$

Problem:



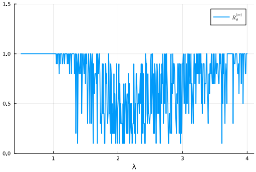
(a) Not fixing the seed (b) Fixing the seed

Point estimation: The smoothed rank

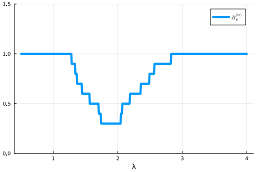
Idea:

$$\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \ \mathcal{R}_{\theta}^{(m)}$$

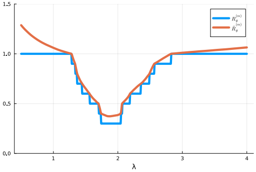
Solution:



(a) Not fixing the seed



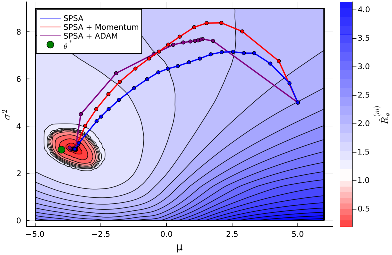
(b) Fixing the seed



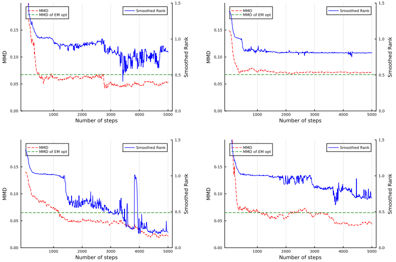
(c) Smoothed rank

Optimization

Stochastic optimization of the smoothed rank (SPSA)



(a) Gaussian distribution



(a) Gaussian Mixture Model

Asymptotic properties in the number of resamplings

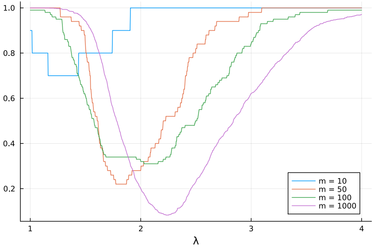
Definition

The normalised rank can be defined as

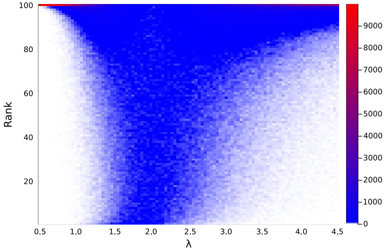
$$\mathcal{R}_\theta^{(m)} = \frac{1}{m}\left(1 + \sum_{i=1}^{m-1} \mathbb{I}_{\{Z_\theta^{(i)} < Z_\theta^{(0)}\}}\right)$$

Proposition

$\lim_{m\to\infty} \mathcal{R}_\theta^{(m)} = F_{Z_\theta}(Z_\theta^{(0)})$ where F_{Z_θ} denotes the CDF of $Z_\theta^{(i)}$



(a) Increasing the n.o. resamplings



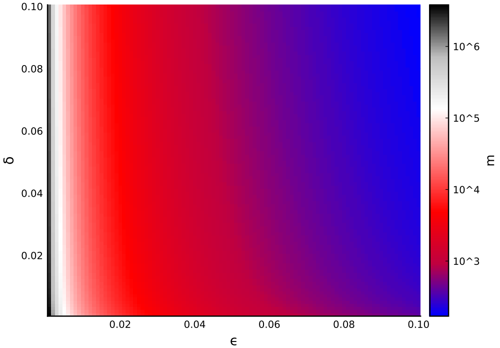
(b) Distribution of Rank

Convergence rate

Proposition

A finite sample bound based on the concentration inequality of Dvoretzky-Kiefer-Wolfowitz (quantitative Glivenko-Cantelli):

$$\mathbb{P}\left(\left|\mathcal{R}_\theta^{(m)} - F_{Z_\theta}\left(Z_\theta^{(0)}\right)\right| > \varepsilon\right) \leq 2\exp\left(-2m\varepsilon^2 + 4\varepsilon\right)$$



Uniform convergence

Question:

Is it possible to guarantee a uniform convergence over Θ?

Definition (Alon et al., 1997)

A set of $\mathcal{X} \to \mathbb{R}$ functions $\mathcal{H}$ is Uniform Glivenco-Cantelli (UGC) if for every $\varepsilon > 0$ it holds that

$$\lim_{l \to \infty} \sup_{\mu \in M_{\mathcal{X}}} \mathbb{P}\left(\sup_{m \geq l} \sup_{f \in \mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^{m} f(x_i) - \int_{\mathcal{X}} f(x)\, d\mu \right| \geq \varepsilon \right) = 0$$

where $\{x_i\}$ are sampled i.i.d. from the distribution μ.

Theorem (Alon et al., 1997)

Let $\mathcal{H}$ be a set of $\mathcal{X} \to [0,1]$ functions. Then $\mathcal{H}$ is UGC if and only if the V_γ-dimension of $\mathcal{H}$ is finite for every $\gamma > 0$

Uniform convergence

Theorem

Let $S^{(0)}$ be a fixed original sample. If there exists a set of $\Theta \to \mathbb{R}$ functions $\mathcal{G}$ such that $Z^{(0)}_{\theta,\xi} \in \mathcal{G}$ for every $\xi \in \mathcal{Q}$, and $Pdim(\mathcal{G}) < \infty$, then

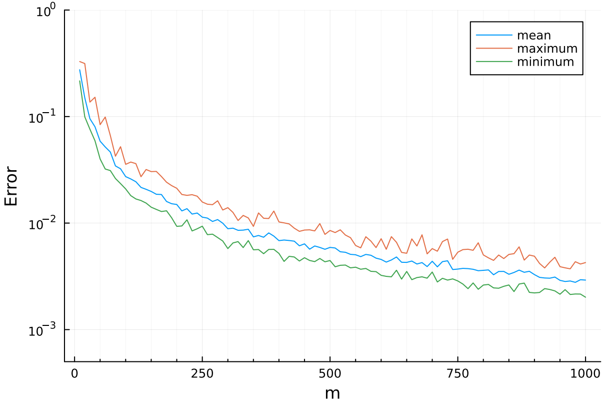
$$\lim_{l \to \infty} \sup_{\theta \in \Theta} \mathbb{P}\left(\sup_{m \geq l} \sup_{\xi \in \mathcal{Q}} \left| \mathcal{R}^{(m)}_{\theta,\xi} - F_{Z^{(1)}_{\theta}}\left(Z^{(0)}_{\theta,\xi} \right) \right| \geq \varepsilon \,\middle|\, S^{(0)} \right) = 0$$

Examples

- Non-randomized reference variables
- ML-based reference variable from the exponential distribution family (across the sample space)
- MMD-based reference variable using a finite dimensional RKHS and a fixed seed (across the sample space)

Uniform convergence

Uniform convergence for ML-based reference variables, in a single parameter exponential family, across the sample space



Uniform convergence

Some useful tools to prove finite pseudo-dimension

Theorem (Anthony and Bartlett, 1999)

If $\mathcal{H}$ is a vector space of $\mathcal{X} \rightarrow \mathbb{R}$ functions, then $\mathrm{Pdim}(\mathcal{H}) = \dim(\mathcal{H})$

Lemma (Anthony and Bartlett, 1999)

If $g : \mathbb{R} \rightarrow \mathbb{R}$ is a non-decreasing function, then for $\mathcal{G} = \{g(f(x)) | f \in \mathcal{H}\}$ it holds that $\mathrm{Pdim}(\mathcal{G}) \leq \mathrm{Pdim}(\mathcal{H})$

Lemma

Let g be any $\Psi \rightarrow \mathcal{X}$ function. Then for $\mathcal{G} = \{h(g(\psi)) | h \in \mathcal{H}\}$ it holds that $Pdim(\mathcal{G}) \leq Pdim(\mathcal{H})$.

Asymptotic behaviour in sample size

Assumption: $S^{(0)} = (x_1, ..., x_n)$ contains i.i.d. instances from $\mathbb{P}_\theta$.

Definition

A reference variable is *consistent*, if it holds that

$$\lim_{n\to\infty} Z_\theta^{(i)} = \begin{cases} 0 & \text{if } x_j \sim \mathbb{P}_\theta \text{ i.i.d.} \\ c \in \mathbb{R}_+ \cup \{\infty\} & \text{else} \end{cases}$$

almost surely for any $\theta \in \Theta$ parameter and $i = 0, ..., m-1$.

Proposition

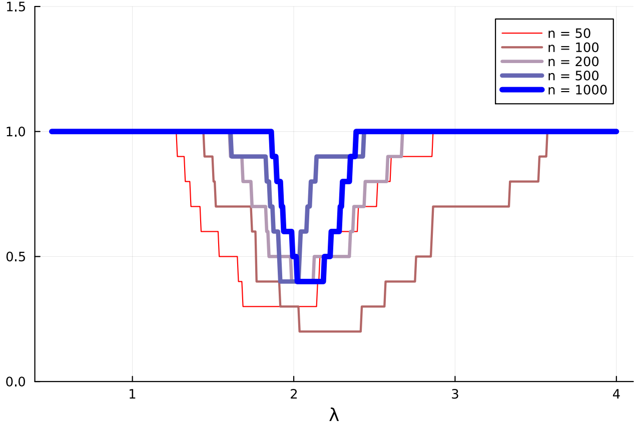
If $\{Z_\theta^{(i)}\}$ are *consistent*, then the normalised rank $\mathcal{R}_\theta^{(m)}$ constructed from it has the following properties:

- I.) $\mathcal{R}_\theta^{(m)} \to 1$ a.s. as $n \to \infty$ if $\mathbb{P}_{\theta^*} \neq \mathbb{P}_\theta$.
- II.) $\mathcal{R}_\theta^{(m)} \xrightarrow{d} U_m[0,1]$ as $n \to \infty$ if $\mathbb{P}_{\theta^*} = \mathbb{P}_\theta$ where $U_m[0,1]$ is the discrete uniform distribution over $\left\{\frac{1}{m}, ..., \frac{m-1}{m}, 1\right\}$.

Asymptotic behaviour in sample size

Proposition

MMD-based reference variables using *characteristic kernels* are *consistent* (e.g., Gaussian, Laplace kernels)



Conclusions

Conclusions

Benefits of the proposed framework:

- Exact confidence regions for finite samples
- Distribution free
- Point estimates
- Asymptotic guarantees

Future directions

- Rashomon sets
- Fine-tuning diffusion models

References

Alon, Noga et al. (1997). "Scale-sensitive dimensions, uniform convergence, and learnability". In: *Journal of the ACM (JACM)* 44.4, pp. 615–631.

Anthony, Martin and Peter L. Bartlett (1999). *Neural Network Learning: Theoretical Foundations*. 1st. Cambridge University Press. ISBN: 052111862X.

Csáji, Balázs Csanád and Ambrus Tamás (2019). "Semi-Parametric Uncertainty Bounds for Binary Classification". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 4427–4432. DOI: 10.1109/CDC40024.2019.9029477.