Density Ratio Estimation with Conditional Probability Paths

Hanlin Yu[1], Arto Klami[1], Aapo Hyvärinen[1], Anna Korba[2], Omar Chehab[2]

1. University of Helsinki, Finland 2. ENSAE, CREST, IP Paris, France

Problem statement

Given samples from two distributions, $X_0 \sim p_0$ and $X_1 \sim p_1$, estimate the ratio $\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}$.

Algorithm [Choi et al., AISTATS 2022]

1. Interpolate samples: $X_t = \sqrt{1-t^2} X_0(\boldsymbol{x}) + \sqrt{t^2} X_1(\boldsymbol{x})$. The law $p_t(\boldsymbol{x})$ is implicit.
2. Estimate the time score $\partial_t \log p_t(\boldsymbol{x})$.
3. Obtain the log ratio through numerical integration: $\log \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \int_0^1 \partial_t \log p_t(\boldsymbol{x}) \mathrm{d}t$

Learning objectives for the time score

Original regression	$\mathscr{L}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x},t)}\left[\lambda(t)\left(\partial_t \log \boxed{p_t(\boldsymbol{x})} - s_{\boldsymbol{\theta}}(\boldsymbol{x},t)\right)^2\right]$	not explicit
Integrate by parts *TSM*	$\mathscr{L}(\boldsymbol{\theta}) = 2\mathbb{E}_{p_0(\boldsymbol{x})}[s_{\boldsymbol{\theta}}(\boldsymbol{x},0)] - 2\mathbb{E}_{p_1(\boldsymbol{x})}[s_{\boldsymbol{\theta}}(\boldsymbol{x},1)]$ $+\mathbb{E}_{p(t,\boldsymbol{x})}[\boxed{2\partial_t s_{\boldsymbol{\theta}}(\boldsymbol{x},t)} + 2\dot{\lambda}(t) s_{\boldsymbol{\theta}}(\boldsymbol{x},t) + \lambda(t) s_{\boldsymbol{\theta}}(\boldsymbol{x},t)^2]$	*slow to differentiate*
Condition **(ours)** *CTSM*	$\mathscr{L}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x},z,t)}\left[\lambda(t)\left(\partial_t \log \boxed{p_t(\boldsymbol{x}\,\vert\,z)} - s_{\boldsymbol{\theta}}(\boldsymbol{x},t)\right)^2\right]$	*explicit*
Factorize **(ours)** ***CTSM-v***	$\mathscr{L}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x},z,t)}\left[\lambda(t)\sum_{i=1}^{D}\left(\partial_t \log p_t(x^i\,\vert\,\boldsymbol{x}^{<i},z) - s^i_{\boldsymbol{\theta}}(\boldsymbol{x},t)\right)^2\right]$	

We also introduce the weighting function $\lambda(t) \propto 1/\lvert \partial_t \log p_t(\boldsymbol{x}\,\vert\,z)\rvert$.

Theoretical guarantees (modified): for K integration steps and N samples,

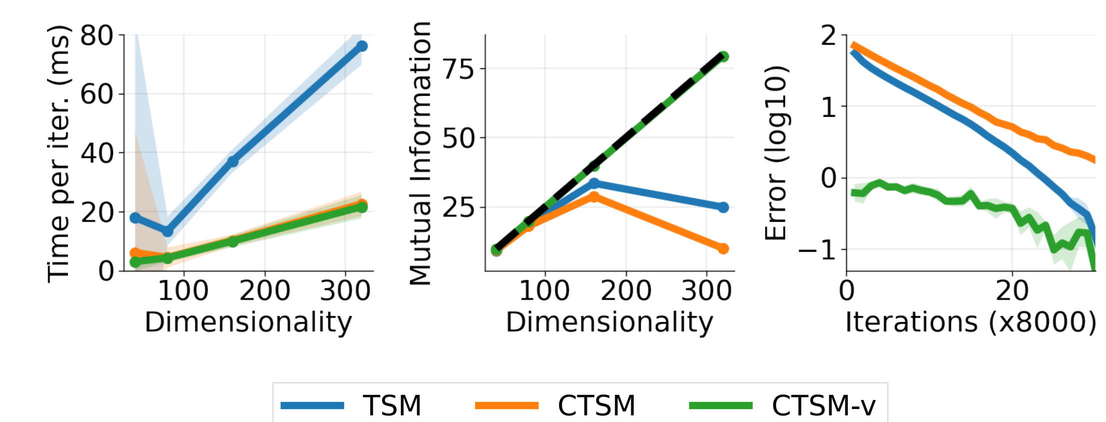
$$\mathbb{E}_{\hat{p}_1}\left\| \log\frac{p_1}{p_0} - \widehat{\log\frac{p_1}{p_0}}\right\|^2_{L^2(p_1)} \leq \underbrace{\frac{1}{2K^2}\mathbb{E}_{p_1(x)}[L(x)^2]}_{\text{integral discretization error}} + \underbrace{\frac{2}{N} e(\theta^*, \lambda, p_t)}_{\text{score estimation error}} + o\left(\frac{1}{N}\right)$$

integral discretization error: *null if $t \to \partial_t \log p_t(\boldsymbol{x})$ constant, i.e. Lipschitz constant L(x) is null*

score estimation error: *null if $\partial_t \log p_t(\boldsymbol{x}\,\vert\,z) = \partial_t \log p_t(\boldsymbol{x})$*

Applications of density-ratio estimation

Mutual information estimation
CTSM-v is faster and outperforms others especially in high dimensions.



Likelihood estimation *(in bits per dimension, BPD)*. We use

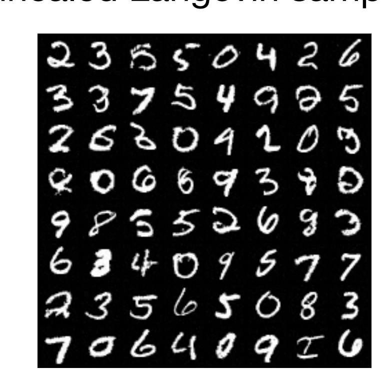
$$\log p_1(\boldsymbol{x}) = \underbrace{\log p_0(\boldsymbol{x})}_{\text{Known}} + \int_0^1 \underbrace{\partial_t \log p_t(\boldsymbol{x})\mathrm{d}t}_{\text{Estimated}}$$

Space	Methods	Approx. BPD	Time per step
Latent space	TSM	1.30	347 ms
	Ours	**1.26**	**58 ms**
Pixel space	TSM	unstable	1103 ms
	Ours	**1.03**	**142 ms**

Sample generation. We convert the estimated time scores into space scores and plug them into popular score-based samplers.

$$\nabla \log p_t(\boldsymbol{x}) = \nabla\left(\underbrace{\log p_0(\boldsymbol{x})}_{\text{Known}} + \int_0^t \underbrace{\partial_s \log p_s(\boldsymbol{x}) ds}_{\text{Estimated}}\right)$$

Annealed Langevin sampler



Probability flow ODE sampler