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 $\dfrac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}$.

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: $\log \dfrac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \displaystyle\int_0^1 \partial_t \log p_t(\boldsymbol{x}) \mathrm{d}t$

Learning objectives for the time score

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

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

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

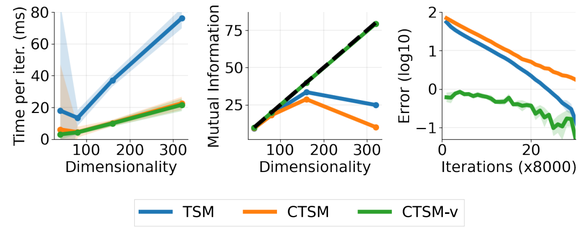
$$\mathbb{E}_{\hat{p}_1}\left\| \log \frac{p_1}{p_0} - \widehat{\log \frac{p_1}{p_0}} \right\|^2_{L^2(p_1)} \leq \underbrace{\frac{1}{2K^2}\mathbb{E}_{p_1(x)}[L(x)^2]}_{\text{integral discretization error}} + \underbrace{\frac{2}{N}e(\theta*, \lambda, p_t)}_{\text{score estimation error}} + o\left(\frac{1}{N}\right)$$

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 $\partial_t \log p_t(\boldsymbol{x}\,|\,z) = \partial_t \log p_t(\boldsymbol{x})$*

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

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

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.

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

Annealed Langevin sampler

Probability flow ODE sampler

Theoretical guarantees: for K integration steps and N samples,

$$\mathbb{E}_{\hat{p}_1}[\mathrm{KL}(p_1, \hat{p}_1)^2] \leq \underbrace{\frac{1}{2K^2}}_{\text{integral discretization error}} \mathbb{E}_{p_1(x)}[L(x)^2] + \underbrace{\frac{2}{N} e(\theta^*, \lambda, p)}_{\text{score estimation error}} +$$

integral discretization error
null if $\partial_t \log p_t(\boldsymbol{x})$ constant wrt t

score estimation error
null if $\partial_t \log p_t(\boldsymbol{x} \mid z) = \partial_t \log p_t(\boldsymbol{x})$

NCE

	Binary	Multi-class	Continuous
NCE	Yes	No	No
TRE	No	Consecutive	No
DRE-infinity	No	Consecutive	Yes
Michael	No	Full	No
Hanlin	No	Full	Yes

The literature suggests that multi-class (full) and continuous would be the thing to aim for.

CNCE

	Binary	Multi-class	Continuous
CNCE	Yes	No	No
No-one	No	Consecutive	No
No-one	No	Consecutive	Yes
Us	No	Full	No
Us	No	Full	Yes

$$p(Y=1\,|\,x) = \frac{p_1(x)}{p_0(x)+p_1(x)} = \frac{1}{1+\frac{p_0}{p_1}(x)}$$

$$p(Y=k\,|\,x) = \frac{p_k(x)}{\sum_{j=1}^{M} p_j(x)} = \frac{1}{1+\sum_{j\neq k}^{M} \frac{p_j}{p_k}(x)}$$

Similar to mean flow

$$= \frac{1}{1+\sum_{j\neq k}^{M} \frac{\frac{p_j}{p_1}(x)}{\frac{p_k}{p_1}(x)}}$$

Similar to flow map

$$\log \frac{p(x', t')}{p(x, t)} = \log \frac{p(x' \mid t')}{p(x \mid t)} = \log \frac{p(x_{t'})}{p(x_t)} = -\int_t^{t'} \nabla \cdot v(s)\, ds$$

OLDER VERSIONS

$$\mathscr{L}_{\text{TSM}}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x},t)}\left[\lambda(t)\|\partial_t \log p_t(\boldsymbol{x}) - s_{\boldsymbol{\theta}}(\boldsymbol{x},t)\|^2\right]$$ Hard to compute.

$$\mathscr{L}_{\text{CTSM}}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{z},t)}\left[\lambda(t)\|\partial_t \log p_t(\boldsymbol{x}\,|\,\boldsymbol{z}) - s_{\boldsymbol{\theta}}(\boldsymbol{x},t)\|^2\right]$$ Easy to compute.

Possible sellings points / research directions

1. Link between multi-class NCE variants and modern papers (mean flows, flow map matching) that parameterize the jump between any two times t, t'. Link with ST-NCE.

2. Focusing the "spatial" part of ST-NCE.

NCE has in very recent years benefitted from a lot of modernization (TRE, DRE-infinity, Srivastava). Applying this modernization to CNCE is very unexplored yet!!

Presumably, binary CNCE estimates decently the EBM on the data manifold, but maybe the multi-class estimates the EBM much better *outside* of the data manifold. This OOD is of interest to many (Yilun, Florentin).

Psychologically, people love scores. The link between CNCE and **space score** matching seems very under-explored yet.

More generally, solving OOD estimating of the space score is very relevant now for composing EBMs.

3. ST-NCE.

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

We have samples from two distributions, $X_0 \sim p_0$ and $X_1 \sim p_1$ and want to estimate their density ratio $\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}$ [1][2][3].

Time Score Matching (TSM) [3]

1. Interpolate samples: $X_t = \sqrt{1-t^2}X_0(\boldsymbol{x}) + \sqrt{t^2}X_1(\boldsymbol{x})$. Its probability 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: $\log \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \int_0^1 \partial_t \log p_t(\boldsymbol{x}) \mathrm{d}t$

Step 2 is currently **slow**: it involves minimizing a loss with higher order gradients (against time and parameters). Can we obtain a faster variant?

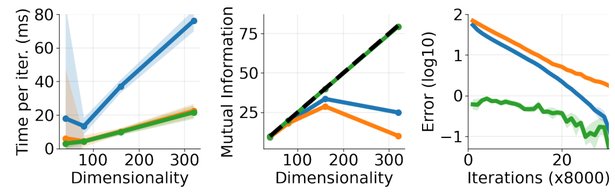
Conditional Time Score Matching (CTSM)

We propose a cheaper version of step 2, using a conditioning variable z to make the intermediate $p_t(\boldsymbol{x}\,|\,z)$ *known in closed-form,* similar to diffusion [4] and flow matching [5]: $\mathscr{L}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x},z,t)}\left[\lambda(t)\left(\partial_t \log p_t(\boldsymbol{x}\,|\,z) - s_{\boldsymbol{\theta}}(\boldsymbol{x},t)\right)^2\right]$, $\lambda(t)$ are positive weights.

We also propose an efficient, **vectorized** version of this loss (CTSM-v), using instead vector of $\partial_t \log p_t(x^i\,|\,\boldsymbol{x}^{<i}, z)$, as well as **theoretical guarantees** on the estimation errors.

Mutual information estimation

Can be reframed as a density ratio estimation problem. CTSM-v is faster and outperforms others especially in high dimensions.



MNIST

Likelihood estimation (in bits per dimension, BPD). Our method is faster and more accurate, achieving good results without relying on pre-trained normalizing flows.
Sample generation. Our trained score network generates reasonable samples using popular diffusion-based samplers with ambient space scores induced by time scores..

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

Annealed Langevin sampler

PF-ODE sampler

References

1. Gutmann, M.U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics, JMLR 2012
2. Rhodes, B., Xu., K. and Gutmann, M.U., Telescoping Density-Ratio Estimation, NeurIPS 2020
3. Choi, K., Meng, C., Song, Y., and Ermon, S., Density ratio estimation via infinitesimal classification, AISTATS 2022
4. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Eamon, S., and Poole, B., Score-based generative modeling through stochastic differential equations, ICLR 2021
5. Lipman, Y., Chen, R. T.Q., Ben-Hamu, H., Nickel, M., and Le, M., Flow matching for generative modeling, ICLR 2023

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

We have samples from two distributions, $X_0 \sim p_0$ and $X_1 \sim p_1$ and want to estimate their density ratio $\dfrac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}$.

Can be solved using probabilistic classification [1][2][3].

Time Score Matching (TSM) [3]

1. Interpolate samples: $X_t = \sqrt{1-t^2}X_0(\boldsymbol{x}) + tX_1(\boldsymbol{x})$. Its probability law $p_t(\boldsymbol{x})$ is *implicit*.

2. Estimate the time score $\partial_t \log p_t(\boldsymbol{x}) \approx s_\theta(\boldsymbol{x}, t)$ using the TSM loss:
$\mathscr{L}(\boldsymbol{\theta}) = 2\mathbb{E}_{p_0(\boldsymbol{x})}[s_{\boldsymbol{\theta}}(\boldsymbol{x},0)] - 2\mathbb{E}_{p_1(\boldsymbol{x})}[s_{\boldsymbol{\theta}}(\boldsymbol{x},1)] + \mathbb{E}_{p(t,\boldsymbol{x})}[2\partial_t s_{\boldsymbol{\theta}}(\boldsymbol{x},t) + 2\dot{\lambda}(t)s_{\boldsymbol{\theta}}(\boldsymbol{x},t) + \lambda(t)s_{\boldsymbol{\theta}}(\boldsymbol{x},t)^2]$,
Where $\lambda(t)$ are positive weights.

3. Obtain the ratio through numerical integration: $\dfrac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \exp\left(\int_0^1 s_{\boldsymbol{\theta}}(\boldsymbol{x},t)\mathrm{d}t\right)$

Step 2 involves **higher order gradients** and thus is expensive. Can we obtain a cheaper variant with higher accuracies?

References

1. Gutmann, M.U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics, JMLR 2012
2. Rhodes, B., Xu., K. and Gutmann, M.U., Telescoping Density-Ratio Estimation, NeurIPS 2020
3. Choi, K., Meng, C., Song, Y., and Ermon, S., Density ratio estimation via infinitesimal classification, AISTATS 2022
4. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Eamon, S., and Poole, B., Score-based generative modeling through stochastic differential equations, ICLR 2021
5. Lipman, Y., Chen, R. T.Q., Ben-Hamu, H., Nickel, M., and Le, M., Flow matching for generative modeling, ICLR 2023

Conditional Time Score Matching (CTSM)

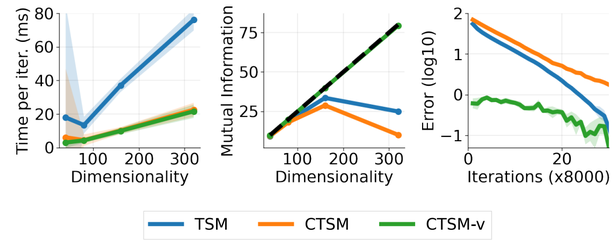
We propose a cheaper version of step 2. The idea is to introduce a conditioning variable z to make the intermediate $p_t(\boldsymbol{x}\,|\,z)$ *known in closed-form,* similar to diffusion [4] and flow matching [5]:

$$\mathscr{L}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x},z,t)}\left[\lambda(t)\left(\partial_t \log p_t(\boldsymbol{x}\,|\,z) - s_{\boldsymbol{\theta}}(\boldsymbol{x},t)\right)^2\right].$$

We also propose an efficient, **vectorized** version of this loss (CTSM-v), using instead vector of $\partial_t \log p_t(x^i\,|\,\boldsymbol{x}^{<i}, z)$, as well as **theoretical guarantees**.

Mutual information estimation

Can be reframed as a density ratio estimation problem. CTSM-v is faster and outperforms others especially in high dimensions.



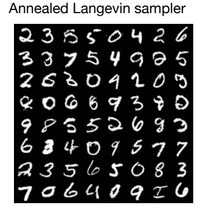
MNIST

Likelihood estimation (BPD, lower is better). Our method is faster and more accurate, achieving good results without relying on pre-trained normalizing flows.

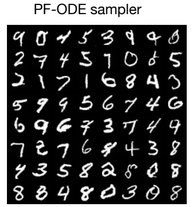
Sample generation. Our trained score network generates reasonable samples using popular diffusion-based samplers with ambient space scores.

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

Annealed Langevin sampler



PF-ODE sampler



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 their density ratio $\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}$.

Density Ratio Estimation using a time score [Choi et al., AISTATS 2022]

1. Interpolate samples: $X_t = \sqrt{1-t^2}X_0(\boldsymbol{x}) + \sqrt{t^2}X_1(\boldsymbol{x})$. Its probability 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: $\log \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \int_0^1 \partial_t \log p_t(\boldsymbol{x}) \mathrm{d}t$

Our loss for learning the time score using positive weights $\lambda(t)$

Classical regression loss (**Choi et al.**)
$\mathscr{L}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x},t)}\left[\lambda(t)\|\partial_t \log p_t(\boldsymbol{x}) - s_t(\boldsymbol{x};\boldsymbol{\theta})\|^2\right]$

No explicit formula for $p_t(\boldsymbol{x})$.

Integration by parts (**Choi et al.**)
$\mathscr{L}(\boldsymbol{\theta}) = 2\mathbb{E}_{p_0(\boldsymbol{x})}[s_{\boldsymbol{\theta}}(\boldsymbol{x},0)] - 2\mathbb{E}_{p_1(\boldsymbol{x})}[s_{\boldsymbol{\theta}}(\boldsymbol{x},1)]$
$+\mathbb{E}_{p(t,\boldsymbol{x})}[2\partial_t s_{\boldsymbol{\theta}}(\boldsymbol{x},t) + 2\dot{\lambda}(t)s_{\boldsymbol{\theta}}(\boldsymbol{x},t) + \lambda(t)s_{\boldsymbol{\theta}}(\boldsymbol{x},t)^2]$

The loss gradient computes $\partial_{\boldsymbol{\theta}}\partial_t s_{\boldsymbol{\theta}}(\boldsymbol{x},t)$ which is slow.

Conditioning variable (**ours**): like diffusion / flow
$\mathscr{L}(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{x},z,t)}\left[\lambda(t)\|\partial_t \log p_t(\boldsymbol{x}|z) - s_{\boldsymbol{\theta}}(\boldsymbol{x},t)\|^2\right]$

Conditioning variable z makes $p_t(\boldsymbol{x}|z)$ known in closed-form.

Efficient implementation of our loss

Reweigh the loss: $\lambda(t) \propto 1/\|\partial_t \log p_t(\boldsymbol{x}|z)\|$

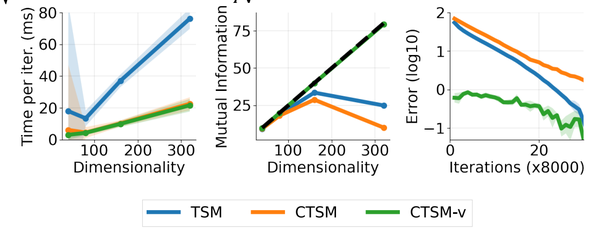
Vectorize the loss: replace $\partial_t \log p_t(\boldsymbol{x}|z) \in \mathbb{R}$ by the vector of $\partial_t \log p_t(x^i|\boldsymbol{x}^{<i}, z)$

Our theoretical guarantees control the expected error as time discretization steps and number of samples increase.

$$\mathbb{E}_{\hat{p}_1}[\mathrm{KL}(p_1,\hat{p}_1)^2] \leq \frac{1}{2K^2}\mathbb{E}_{p_1(x)}[L(x)^2] + \frac{2}{N}e(\theta*,\lambda,p) + o(\frac{1}{N})$$

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)*. Higher likelihoods and faster training directly in pixel space.

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

Sample generation. We convert the estimated time scores $\partial_t \log p_t(\boldsymbol{x})$ into space scores $\nabla \log p_t(\boldsymbol{x})$ and plug them into popular score-based samplers.

Annealed Langevin sampler

PF-ODE sampler