Provable benefits of annealing for estimating normalizing constants: Importance Sampling, Noise-Contrastive Estimation, and beyond.

Inria   université PARIS-SACLAY

Omar Chehab, Aapo Hyvärinen, Andrej Risteski

Carnegie Mellon University   UNIVERSITY OF HELSINKI

Question: when does annealing reduce the estimation error (MSE) of a log-normalizing constant?

Approach: write the estimation error (MSE) produced by different annealing choices.

Contributions: compare the estimation error (MSE) produced by different annealing paths.

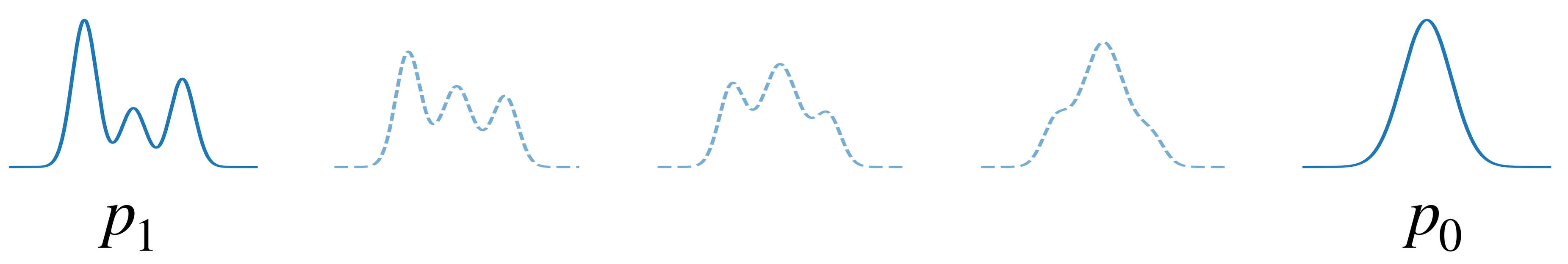
Annealed Estimation of a Normalizing Constant

In many areas of statistics, a target distribution is specified by an **unnormalized** density $f_1(x)$. Evaluating the probability

$$p_1(x) = \frac{f_1(x)}{Z_1} \qquad Z_1 = \int f_1(x)dx$$

requires computing the normalization Z_1 defined by an often intractable integral.

The (log) normalization can be **estimated** using a random **sample** from $K+1$ **distributions** that link the intractable target p_1 to a tractable proposal p_0



Using the identity,

$$\log Z_1 = \underbrace{\sum_{k=1}^{K} \log\left(\frac{Z_{k/K}}{Z_{(k-1)/K}}\right)}_{\textit{unknown}} + \underbrace{\log Z_0}_{\textit{known}}$$

each log-ratio is estimated by solving a binary classification task between samples from $p_{(k-1)/K}$ and $p_{k/K}$. Different classification losses lead to the noise-contrastive or importance sampling estimators.

Estimation error produced by different estimators

In the limit of many distributions $K \to \infty$, we prove that the annealed importance sampling and noise-contrastive estimators produce the same estimation error (MSE) of the target log-normalization:

$$\boxed{\mathrm{MSE} = \frac{1}{N}\int_0^1 I(t)dt + o\left(\frac{1}{N}\right) + o\left(\frac{K^2}{N}\right), \quad \text{with } I(t) = \mathbb{E}_{x\sim p_t}[\partial_t \log p_t(x)^2] \text{ Fisher Information of } p_t(x)}$$

To reduce the error: increase the **sample size** N or reduce the **path length between the target and proposal** $\int_0^1 I(t)dt$ measured by summing the Fisher Information of distributions along the path.

Estimation error produced by different paths

Assume the target and proposal are in exponential family with parameters θ_1 and θ_0. We study the length $\int_0^1 I(t)dt$ of common paths as a func. of the **gap between the target and proposal** $\|\theta_1 - \theta_0\|^2$:

No path

$$\mathrm{MSE} = O\left(\frac{1}{N}\exp\left(\|\theta_1 - \theta_0\|^2\right)\right)$$

Arithmetic path:

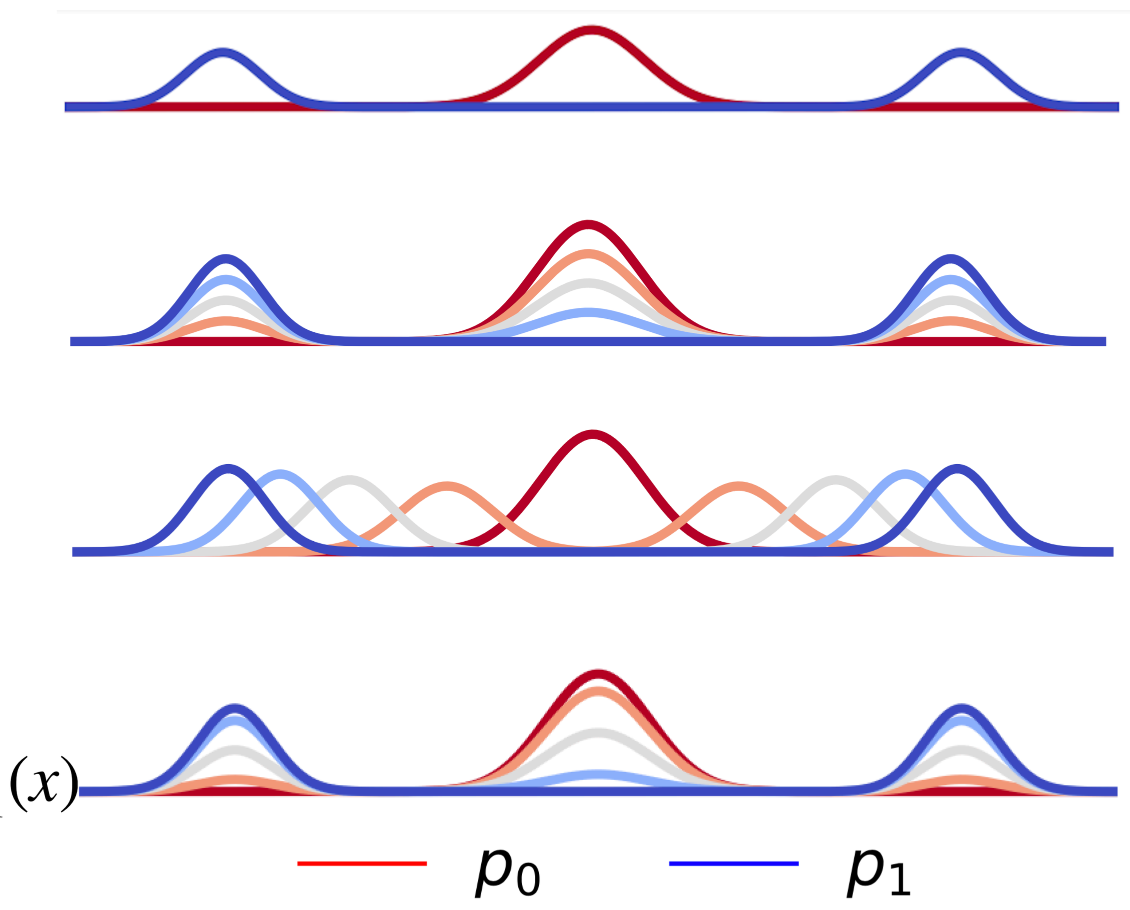
$$p_t(x) \propto (1-t)p_0(x) + tf_1(x) \qquad \mathrm{MSE} = O\left(\frac{1}{N}\exp\left(\|\theta_1 - \theta_0\|^2\right)\right)$$

Geometric path:

$$p_t(x) \propto p_0(x)^{1-t} \times f_1(x)^t \qquad \mathrm{MSE} = O\left(\frac{1}{N}\mathrm{poly}\left(\|\theta_1 - \theta_0\|^2\right)\right)$$

Optimal path:

$$p_t(x) = \cos^2\left(\frac{\pi}{2}t\right)p_0(x) + \sin^2\left(\frac{\pi}{2}t\right)p_1(x) \qquad \mathrm{MSE} = O\left(\frac{1}{N}\pi^2\right)$$



A same estimation error requires a sample size that is **exponential** using no path, **polynomial** using the geometric path, and **constant** using the optimal path — all relative to the target-proposal gap.

But there is no free lunch: the optimal path is an arithmetic path that is **reparameterized** using the unknown Z_1. We pre-estimate it in a two-step procedure.