Provable benefits of annealing for estimating normalizing constants

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent research has developed several Monte Carlo methods for estimating the normalization constant (partition function) based on the idea of annealing. This means sampling successively from a path of distributions which interpolate between a tractable "proposal" distribution and the unnormalized "target" distribution. Prominent estimators in this family include annealed importance sampling and annealed noise-contrastive estimation (NCE). Such methods hinge on a number of design choices: which estimator to use, which path of distributions to use and whether to use a path at all; so far, there is no definitive theory on which choices are efficient. Here, we evaluate each design choice by the asymptotic estimation error it produces. First, we show that using NCE is more efficient than the importance sampling estimator, but in the limit of infinitesimal path steps, the difference vanishes. Second, we find that using the geometric path brings down the estimation error from an exponential to a polynomial function of the parameter distance between the target and proposal distributions. Third, we find that the arithmetic path, while rarely used, can offer optimality properties over the universally-used geometric path. In fact, in a particular limit, the optimal path is arithmetic. Based on this theory, we finally propose a two-step estimator to approximate the optimal path in an efficient way.

1 Introduction

Recent progress in generative modeling has sparked renewed interest in models of data that are defined by an unnormalized distribution. One prominent example is energy-based (or score-based) models, which are increasingly used in deep learning [1], and for which there are a variety of parameter estimation procedures [2–5]. Another example comes from Bayesian statistics, where the posterior model of parameters given data is frequently known only up to a proportionality constant. Such models can be evaluated and compared by the probability they assign to a dataset, yet this requires computing their normalization constants (partition functions) which are typically high-dimensional, intractable integrals.

Monte-Carlo methods have been successful at computing these integrals using sampling methods [6]. The most common is importance sampling [6] which draws a sample from a tractable, "proposal" distribution to integrate the unnormalized "target" density. Noise-contrastive estimation (NCE) [3] uses a sample from *both* the proposal and the target, to compute the integral. Yet such methods suffer from high variance, especially when the "gap" between the proposal and target densities is large [7–9]. This has motivated various approaches to gradually bridge the gap with intermediate distributions, which is loosely referred to as "annealing". Among them, annealed importance sampling (AIS) [10–12] is widely adopted: it has been used to compute the normalization constants of deep stochastic models [13, 14] or to motivate a lower-bound for learning objectives [15, 16]. To integrate

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Name	Loss identified by $\phi(\boldsymbol{x})$	Estimator $\hat{Z}_1$	MSE
IS	$x \log x$	$\mathbb{E}_{p_0} \frac{f_1}{f_0}$	$\frac{1+\nu}{\nu N} \mathcal{D}_{\chi^2}(p_1, p_0)$
RevIS	$-\log x$	$\left(\mathbb{E}_{p_1} \frac{f_0}{f_1}\right)^{-1}$	$\frac{1+\nu}{N} \mathcal{D}_{\chi^2}(p_0, p_1)$
NCE	$x \log x - (1+x) \log(\frac{1+x}{2})$	implicit	$\frac{(1+\nu)^2}{\nu N} \frac{\mathcal{D}_{\mathrm{HM}}(p_1, p_0)}{1-\mathcal{D}_{\mathrm{HM}}(p_1, p_0)}$

Table 1: Some estimators of the normalization obtained by minimizing a classification loss, and their estimation error in terms of well-known divergences [22]. For details and definitions, see Appendix A.

the unnormalized "target" density, it draws a sample from an entire path of distributions between the proposal and the target. While annealed importance sampling has been shown to be effective empirically, its theoretical understanding remains limited [17, 18]: it is yet unclear when annealing is effective, for which annealing paths, and whether AIS is a statistically efficient way to do it.

In this paper, we define a family of annealed Bregman estimators (ABE) for the normalization constant. We show that it is general enough to recover many classical estimators as a special case, including importance sampling, noise-contrastive estimation, umbrella sampling [19], bridge sampling [20] and annealed importance sampling. We provide a statistical analysis of its hyperparameters such as the choice of paths, and show the following. First, we show that using NCE is more efficient than the importance sampling estimator, but in the limit of infinitesimal path steps, the difference vanishes. Second, we find that the near-universally used geometric path brings down the estimation error from an exponential to a polynomial function of the parameter distance between the target and proposal distributions. Third, we find that using the recently introduced arithmetic path [21] is exponentially inefficient in its basic form, yet it can be reparameterized to be in some sense optimal. Based on this optimality result, we finally propose a two-stage estimation procedure which first finds an approximation of the optimal (arithmetic) path, then uses it to estimate the normalization constant.

2 Background

Importance sampling and NCE The problem considered here is how to compute the normalization constant [1], *i.e.* the integral of some unnormalized density $f_1(\boldsymbol{x})$ called "target". Importance sampling and noise-contrastive estimation are two common estimators which integrate the unnormalized target over a random sample drawn from a tractable density $p_0(\boldsymbol{x})$ called "proposal"(Table 1, column 3). In fact, they are part of a larger family of Monte-Carlo estimators which can be interpreted as solving a binary classification task, aiming to distinguish between a sample drawn from the proposal and another from the target [22]. Each estimator is obtained by minimizing a specific binary classification loss that is identified by a convex function $\phi(\boldsymbol{x})$. For example, minimizing the classification loss identified by $\phi_{\mathrm{IS}}(x) = x \log x$ yields the importance sampling estimator. Similarly, $\phi_{\mathrm{RevIS}}(x) = -\log x$ leads to the reverse importance sampling estimator [23], and $\phi_{\mathrm{NCE}}(x) = x \log x - (1+x) \log((1+x)/2)$ to the noise-contrastive estimator. These estimators are summarized in Table 1.

Annealed estimators Annealing extends the above "binary" setup, by introducing a sequence of $K+1$ distributions from the proposal to the target (included). The idea will be to draw a sample from *all* these distributions to integrate the target $f_1(\boldsymbol{x})$. These intermediate distributions are obtained from a path $(f_t)_{t\in[0,1]}$, defined by interpolating between the proposal p_0 and unnormalized target f_1: this path is therefore unnormalized. Different interpolation schemes can be chosen. A general one, explained in Figure 1, is to take the q-mean of the proposal and target [21]. Two values of q are of particular interest: $q \to 0$ defines a a near-universal path [18], obtained by taking the geometric mean of the target and proposal, while $q = 1$ defines a path obtained by the arithmetic mean.

Once a path is chosen, it can be uniformly[2] discretized into a sequence of $K+1$ unnormalized densities, denoted by $(f_{k/K})_{k\in[0,K]}$ with corresponding normalizations $(Z_{k/K})_{k\in[0,K]}$. In practice, samples are drawn from the corresponding normalized densities $(p_{k/K})_{k\in[0,K]}$ using Markov Chain

[1] in this paper we also say we "estimate" the the normalization constant, though this terminology is unconventional as estimation traditionally refers to the *parameters* of a statistical model

[2] other discretization schemes can be equivalently achieved by re-parameterizing the path [17]

General ($q \in]0, 1]$)	$f_t(\boldsymbol{x}) = \left((1-t)p_0(\boldsymbol{x})^q + tf_1(\boldsymbol{x})^q\right)^{\frac{1}{q}}$
Geometric ($q \to 0$)	$f_t(\boldsymbol{x}) = p_0(\boldsymbol{x})^{1-t} \times f_1(\boldsymbol{x})^t$
Arithmetic ($q = 1$)	$f_t(\boldsymbol{x}) = (1-t)p_0(\boldsymbol{x}) + tf_1(\boldsymbol{x})$

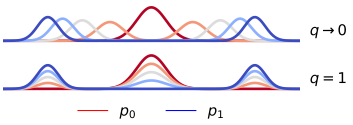


Figure 1: Geometric and arithmetic path between the proposal and target distributions. Here, the proposal (red) is a standard gaussian. The target (blue) is a gaussian mixture with two modes, and same first and second moments as the proposal.

Monte Carlo (MCMC). This sampling step incurs a computational cost, which is paid in the hope of reducing the variance of the estimation. It is common in the literature [17, 18] to assume *perfect sampling*, meaning the MCMC has converged and produced exact and independent samples from the distributions along the path, which simplifies the analysis.

Estimation error A measure of "quality" is required to compare different estimation choices, such as whether to anneal and which path to use. Such a measure is given by the Mean Squared Error (MSE), which is generally tractable when written at the first order in the asymptotic limit of a large sample size [24, Eq. 5.20]. These expressions have been derived for estimators obtained by minimimizing a classification loss [22] and are included in table 1. They measure the "gap" between the proposal and target distributions using statistical divergences. Note also that the estimation error depends on the *normalized* target density (column 4), while the estimators are computed using the *unnormalized* target density (column 3). Further details are available in Appendix A.

3 Annealed Bregman Estimators of the normalization constant

The question that we try to answer in this paper is: *How should we choose the $K+1$ distributions in annealing, and how are their samples best used?* To answer this, we will study the error produced by different estimation choices. But first we define the set of estimators for which the analysis is done.

Definition of Annealed Bregman Estimators We now define a new family of estimators, which we call annealed Bregman estimators (ABE); this terminology is explained later. We will show that this is a general class of estimators for computing the normalization using a sample drawn from the sequence of $K+1$ distributions. For ABE, the log normalization $\log Z_1$ is estimated additively along the sequence of distributions

$$\widehat{\log Z_1} = \sum_{k=0}^{K-1} \widehat{\log\left(\frac{Z_{(k+1)/K}}{Z_{k/K}}\right)} + \log Z_0 \ . \tag{1}$$

Defining the estimation in log-space is analytically convenient, as it is easier to analyze a sum of estimators than a product. Exponentiating the result leads to an estimator of Z_1. We naturally extend the binary setup ($K = 1$) of Chehab et al. [22] and propose to compute each of the intermediate log-ratios, by solving a classification task between samples drawn from their corresponding densities $p_{k/K}$ and $p_{(k+1)/K}$. Each binary classification loss is now identified by a convex function $\phi_k(\boldsymbol{x})$ and defined as

$$\mathcal{L}_k(\beta) := \mathbb{E}_{\boldsymbol{x}\sim p_{k/K}}[\phi_k^{'}(r(\boldsymbol{x};\beta)) \times r(\boldsymbol{x};\beta) - \phi_k(r(\boldsymbol{x}))] - \mathbb{E}_{\boldsymbol{x}\sim p_{(k+1)/K}}[\phi_k^{'}(r(\boldsymbol{x};\beta))] \tag{2}$$

where $r(\boldsymbol{x};\beta)$ is parameterized by the unknown log-ratio $\beta^* = \log(Z_{(k+1)/K}/Z_{k/K})$

$$r(\boldsymbol{x};\beta) = \exp(-\beta) \times f_{(k+1)/K}(\boldsymbol{x})/f_{k/K}(\boldsymbol{x}) \ . \tag{3}$$

The convex functions $(\phi_k)_{k\in[0,K-1]}$ which identify the classification losses are called "Bregman" generators, hence ABE. As mentioned above, we assume perfect sampling and allocate the total sample size N equally among the K estimators in the sum.

Hyperparameters The annealed Bregman estimator depends on the following hyperparameters: (1) the choice of path q; (2) the number of distributions along that path $K+1$ (including the proposal and the target); (3) the classification losses identified by the convex functions $(\phi_k)_{k\in[0,K-1]}$.

Different combinations of these hyperparameters recover several common estimators of the log-partition function. In binary case of $K=1$ this includes importance sampling, reverse importance sampling, and noise-contrastive estimation, each obtained for a different choice of the classification loss [22]. To build intuition, consider $K=2$ so that we add a single intermediate distribution $p_{1/2}$ to the sequence. Using the importance sampling loss ($\phi_0 = x\log x$) for the first ratio, and reverse importance sampling ($\phi_1 = -\log x$) for the second ratio, recovers the *bridge sampling estimator* as a special case [20]

$$\widehat{\log Z_1} = -\log \mathbb{E}_{p_1}\frac{f_{1/2}}{f_1} + \log \mathbb{E}_{p_0}\frac{f_{1/2}}{f_0} \log Z_0 \ . \tag{4}$$

Alternatively, we can use these classification losses in reverse order: reverse importance sampling ($\phi_0 = -\log x$) for the first ratio, and importance sampling ($\phi_1 = x\log x$) for the second ratio, and recover the *umbrella sampling estimator* [19] also known as the *ratio sampling estimator* [25]

$$\widehat{\log Z_1} = \log \mathbb{E}_{p_{1/2}}\frac{f_1}{f_{1/2}} - \log \mathbb{E}_{p_{1/2}}\frac{f_0}{f_{1/2}} \log Z_0 \ . \tag{5}$$

Another option yet, is to use the same classification loss for all ratios. With importance sampling ($\phi_k = x\log x,\ \forall k \in [\![0, K-1]\!]$), we recover the *annealed importance sampling* estimator [10–12]

$$\widehat{\log Z_1} = \sum_{k=1}^{K} \log \mathbb{E}_{\boldsymbol{x}\sim p_{k-1}}\left[\frac{f_k}{f_{k-1}}(\boldsymbol{x})\right] + \log Z_0 \ . \tag{6}$$

The family of annealed Bregman estimators is visibly large enough to include many existing estimators, obtained for different hyperparmeter choices. This raises the fundamental question of how these hyperparameters should be chosen, in particular in the challenging case where the *target and proposal have little overlap and the data is high dimensional*. To answer this question, we will study the estimation error produced by different hyperparameter choices.

4 Statistical analysis of the hyperparameters

We consider a fixed data budget N and investigate how the remaining hyperparameters are best chosen for statistical efficiency. The starting point for the analysis is that as ABE estimates the normalization in log-space, the estimator is obtained by a sum of independent and asymptotically unbiased estimators [26] given in Eq. 1 and thus the mean squared errors written in table 1 are additive. Each individual error actually measures an overlap between two consecutive distributions along the path, and annealing integrates these overlaps.

4.1 Classification losses, ϕ_k

Given the popularity of annealed importance sampling, we should first ask if the importance sampling loss is really an acceptable default. We recall an important limitation of annealed importance sampling [27]: its estimation error is notoriously sensitive to distribution tails. Without annealing, it is infinite when the target p_1 has a heavier tail than the proposal p_0. When annealing with a geometric path, for example between two gaussians with different covariances $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{Id})$ and $p_1 = \mathcal{N}(\mathbf{0}, 2\,\mathbf{Id})$, the geometric path produces gaussians with increasing variances $\Sigma_t = (1-t/2)^{-1}\,\mathbf{Id}$ and therefore increasing tails. Hence, the same tail mismatch holds along the path. Note that this concern is a realistic one for natural image data, as the target distribution over images is typically super-gaussian [28] while the proposal is usually chosen as gaussian.

This warrants a better choice for the loss: In the binary setup ($K=1$), the NCE loss is optimal [20, 22] and its error can be orders of magnitude less than importance sampling [20]. We extend this optimality result over a sequence of distributions $K>1$ and also show that the gap between annealed IS and annealed NCE is closed in the limit of a continuous path:

Theorem 1 (Estimation error and the Fisher-Rao path length) *For a finite value of K, the optimal loss is NCE*

$$\mathrm{MSE}(p_0, p_1; q, K, N, \phi_{\mathrm{NCE}}) \leq \mathrm{MSE}(p_0, p_1; q, K, N, \phi), \qquad \forall q, K, N, \phi \ . \tag{7}$$

In the limit of $K \to \infty$, NCE, IS, and revIS converge to the same estimation error, given by the Fisher-Rao path length from the proposal to the target

$$\mathrm{MSE}(p_0, p_1; q, K, N, \phi) \to \frac{1}{N} \int_0^1 I(t) dt, \qquad \text{for } K \to \infty, \phi \in \{\phi_{\mathrm{NCE}}, \phi_{\mathrm{IS}}, \phi_{\mathrm{RevIS}}\} \tag{8}$$

where the Fisher-Rao metric $I(t) := \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}, t)}[(\frac{d}{dt} \log p_t(\boldsymbol{x}))^2]$ defined as the Fisher information over the path, using time t as the parameter.

This is proven in Appendix B. While the NCE estimator requires solving a (potentially non-convex) scalar optimization problem in Eq. 2 and IS does not, this is the price to pay for statistical optimality. In the following we will keep the optimal NCE loss and will indicate the dependency of the estimation error on ϕ_{NCE} with a subscript, instead. We highlight that our theorems in this paper apply to the MSE in the limit of $K \to \infty$: their results hold the same for the IS and RevIS losses by virtue of theorem 1. Just as in the binary case, while the estimator is computed with the *unnormalized* path of densities (Eq. 2), the estimation error depends on the *normalized* path of densities (Eq. 8).

4.2 Number of distributions, $K + 1$

It is known that estimating the normalization constant using plain importance sampling ($K = 1$) can produce a statistical error than is exponential in the dimension [6, Example 9.1]. We show that in the binary case, NCE also suffers from an estimation error that scales exponentially with the dimension.

In the following, we consider a proposal p_0 and target p_1 that are in an exponential family; note that certain exponential families have universal approximation capabilities [29]. The exponential family is defined as

$$p(\boldsymbol{x}; \boldsymbol{\theta}) := \exp(\langle \boldsymbol{\theta}, \boldsymbol{t}(\boldsymbol{x}) \rangle - \log Z(\boldsymbol{\theta})) \tag{9}$$

where $Z(\boldsymbol{\theta}) = \int \exp(\langle \boldsymbol{\theta}_1, \boldsymbol{t}(\boldsymbol{x}) \rangle)$. We will consider that the (unnormalized) target density f_1 is what we call a *simply unnormalized model* defined as

$$f_1(\boldsymbol{x}) = \exp(\langle \boldsymbol{\theta}_1, \boldsymbol{t}(\boldsymbol{x}) \rangle) \tag{10}$$

Note that in general, a pdf can be unnormalized in many ways: one can multiply an unnormalized density by any positive function of $\boldsymbol{\theta}$ and it will still be unnormalized. However, the simple and intuitive case defined above is what we encounter in the analysis below.

For exponential families, the log-normalization $\log Z(\boldsymbol{\theta})$ is a convex function ("log-sum-exp") of the parameter $\boldsymbol{\theta}$ [30], which implies $0 \preccurlyeq \nabla^2_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta})$. In our theorems we use the further assumptions of strong convexity with constant M, and/or smoothness with constant L (gradient is L-Liptschitz):

$$\nabla^2_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) \succcurlyeq M \, \mathbf{Id} \tag{11}$$

$$\nabla^2_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) \preccurlyeq L \, \mathbf{Id} \tag{12}$$

For exponential families, the derivatives of the log partition function yield moments of the sufficient statistics, so we are effectively assuming that the eigenvalues of $\nabla^2_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{x} \sim p}[\boldsymbol{t}(\boldsymbol{x})]$ are bounded, which will be the case for parameters in a bounded domain $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. An example along with the proofs of the following theorems 2 and 3, are provided in Appendix B.

Theorem 2 (Exponential error of binary NCE) *Assume the proposal p_0 is from the normalized exponential family, while the (unnormalized) target f_1 is from the simply unnormalized exponential family (Eq. 10). The log-partition function $\log Z(\boldsymbol{\theta})$ is assumed to be strongly convex (Eq. 11). Then in the binary case $K = 1$, the estimation error of NCE is (at least) exponential in the parameter-distance between the proposal and the target*

$$\mathrm{MSE}_{\mathrm{NCE}}(p_0, p_1; q, K, N) \geq \frac{4}{N} \exp\left(\frac{1}{8} M \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2\right) - 1, \qquad \text{when } K = 1 \tag{13}$$

where M is the strong convexity constant of $\log Z(\boldsymbol{\theta})$.

Annealing the importance sampling estimator (increasing K) was proposed in hope that we can trade the statistical cost in the dimension for a computational cost (number of classification tasks) which is more acceptable. Yet, there is no definitive theory on the ability of annealing to reduce the statistical cost in a general setup [18, 21]. For both importance sampling and noise-contrastive estimation, we prove that annealing with the near-universal geometric path brings down the estimation error, from exponential to polynomial in the parameter-distance between the proposal and target. Given that the parameter-distance itself scales as $\sqrt{D}$ with the dimension, using these paths effectively makes annealed estimation amenable to high-dimensional problems. This corroborates empirical [31] and theoretical [17] results which suggested in simple cases that annealing with an appropriate path can reduce the estimator error up to several orders of magnitude.

Theorem 3 (Polynomial error of annealed NCE with a geometric path) *Assume the proposal p_0 is from the normalized exponential family, while the (unnormalized) target f_1 is from the simply unnormalized exponential family (Eq. 10). The log-partition function $\log Z(\boldsymbol{\theta})$ is assumed to be strongly convex and smooth (Eq. 11, Eq. 12).*
Then in the annealing limit of a continuous path $K \to \infty$, the estimation error of annealed NCE with the geometric path is (at most) polynomial in the parameter-distance between the proposal and the target

$$\mathrm{MSE}_{\mathrm{NCE}}(p_0, p_1; q, K, N) \leq \frac{L^2}{MN}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2, \qquad \text{when } K \to \infty, q = 0 \tag{14}$$

where M and L are respectively the strong convexity and smoothness constants of $\log Z(\boldsymbol{\theta})$.

To our knowledge, this is the first result building on Gelman and Meng [17, Table 1] and Grosse et al. [18] which showcases the benefits of annealed estimation for a general target distribution.

We conclude that annealing with the near-universal geometric path provably benefits noise-contrastive estimation, as well as importance sampling and reverse importance sampling, when the proposal and target distributions have little overlap.

4.3 Path parameter, q — geometric vs. arithmetic

Despite the near-universal popularity of the geometric path ($q \to 0$), it is worth asking if there are other simple paths that are more optimal. Interpolating moments of exponential families was shown to outperform the geometric path by Grosse et al. [18], yet building such a path requires knowing the exponential family of the target. Other paths based on the arithmetic mean (and generalizations) of the target and proposal, were proposed in Masrani et al. [21], without a definitive theory of the estimation error.

Next, we analyze the error of the arithmetic path. We prove that the arithmetic path ($q = 1$) does *not* exhibit the same benefits as the geometric path: in general, its estimation error grows exponentially in the parameter-distance between the target and proposal distributions. However, in the case where an oracle gives us the normalization Z_1 to be used only in the construction of the path (we will discuss what this means in practice below), the arithmetic path can be reparameterized so as to bring down the estimation error to polynomial, even constant, in the parameter-distance. We start by the negative result.

Theorem 4 (Exponential error of annealed NCE with an arithmetic path) *Assume the proposal p_0 is from the normalized exponential family, while the (unnormalized) target f_1 is from the simply unnormalized exponential family (Eq. 10). The log-partition function $\log Z(\boldsymbol{\theta})$ is assumed to be strongly convex (Eq. 11).*
Consider the annealing limit of a continuous path $K \to \infty$ path, and of a far-away target $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\| \to \infty$. For estimating the log normalization of the (unnormalized) target density f_1, the estimation error of annealed NCE with the arithmetic path is (at least) exponential in the parameter-distance between the proposal and the target.

$$\mathrm{MSE}_{\mathrm{NCE}}(p_0, p_1; q, K, N) = O\Big(f\Big(\frac{1}{N}\exp\Big(\frac{M}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2\Big)\Big)\Big), \qquad \text{when } K \to \infty, q = 1 \tag{15}$$

where f is an increasing function defined in Appendix B.3.

Path name	Unnormalized density	Normalized density	Error
Geometric	$f_t(\boldsymbol{x}) = p_0(\boldsymbol{x})^{1-t} f_1(\boldsymbol{x})^t$	$p_t(\boldsymbol{x}) \propto p_0(\boldsymbol{x})^{1-t} p_1(\boldsymbol{x})^t$	poly
Arithmetic	$f_t(\boldsymbol{x}) = (1-w_t)p_0(\boldsymbol{x}) + w_t f_1(\boldsymbol{x})$	$p_t(\boldsymbol{x}) = (1-\tilde{w}_t)p_0(\boldsymbol{x}) + \tilde{w}_t p_1(\boldsymbol{x})$	
vanilla	$w_t = t$	$\tilde{w}_t = \frac{tZ_1}{(1-t)+tZ_1}$	exp
oracle	$w_t = \frac{t}{t+Z_1(1-t)}$	$\tilde{w}_t = t$	poly
oracle-trig	$w_t = \frac{\sin^2\left(\frac{\pi t}{2}\right)}{\sin^2\left(\frac{\pi t}{2}\right)+Z_1\cos^2\left(\frac{\pi t}{2}\right)}$	$\tilde{w}_t = \sin^2\left(\frac{\pi t}{2}\right)$	const

Table 2: Geometric and arithmetic paths, defined in the space of unnormalized densities (second column); "oracle" and "oracle-trig" are reparameterizations of the arithmetic path which depend on the true normalization Z_1. The corresponding normalized densities (third column) produce an estimation error (fourth column) which we quantify.

We suggest an intuitive explanation for this negative result. We begin with the observation that the estimation error (Eq. 8) depends on the *normalized* path of densities. Suppose the target model is rescaled by a constant 100, so that the new unnormalized target density is $f_1(\boldsymbol{x}) \times 100$ and its new normalization is $Z_1 \times 100$. Looking at table 2, this rescaling does not modify the geometric path of normalized densities, while it does the arithmetic path of normalized densities. Because the estimation error depends on path of normalized densities, this makes the arithmetic choice sensitive to target normalization, even more so as the parameter distance grows and the log-normalization with it, as a strongly convex function of it (Appendix, Eq. 85). This suggests making the arithmetic path of normalized distributions "robust" to the choice of Z_1. We will show this can be achieved by re-parameterizing the path in terms of Z_1.

We next prove that certain reparameterizations can bring down the error to a polynomial and even constant function of the parameter-distance between the target and proposal. The following theorems may seem purely theoretical, as if necessitating an oracle for Z_1, but they will actually lead to an efficient estimation algorithm later.

Theorem 5 (Polynomial error of annealed NCE with an arithmetic path and oracle) *Assume the same as in Theorem 4, replacing the strong convexity of the log-partition by smoothness (Eq. 12). Additionally, suppose an oracle gives the normalization constant Z_1 to be used only in the reparameterization of the arithmetic path with $t \to \frac{t}{t+Z_1(1-t)}$ (see Table 2). This brings down the estimation error of annealed NCE to (at most) polynomial in the parameter-distance*

$$\mathrm{MSE}_{\mathrm{NCE}}(p_0, p_1; q, K, N) \leq \frac{1}{N}(2 + L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2), \qquad \text{when } K \to \infty, q = 1 \tag{16}$$

where L is the smoothness constant of $\log Z(\boldsymbol{\theta})$.

In fact, supposing we have (oracle) access to the normalizing constant Z_1, the arithmetic path can even be reparameterized such that it is the optimal path in a certain limit. We next prove such optimality in the limits of a continuous path $K \to \infty$ and "far-away" target and proposal:

Theorem 6 (Constant error of annealed NCE with an arithmetic path and oracle) *Suppose we can successively take the limit of a continuous annealing path $K \to \infty$, then the limit of a target distribution that tends toward no overlap with the proposal $p_1(\boldsymbol{x})p_0(\boldsymbol{x}) \to 0$ pointwise (and assuming domination by an integrable function)* [3]*. Then the optimal annealing path convergences pointwise to an arithmetic path reparameterized trigonometrically with $t \to \frac{t}{\sin^2(\frac{\pi t}{2})+Z_1(1-\sin^2(\frac{\pi t}{2}))}$. In that limit, the estimation error is constant with respect to the parameter-distance*

$$\mathrm{MSE}_{\mathrm{NCE}}(p_0, p_1; q, K, N) \sim \frac{2\pi^2}{N}, \qquad \text{when } K \to \infty, q = 1 \tag{17}$$

Two-step estimation Thus, we see that, perhaps unsurprisingly, the "optimal" mixture weights in the space of unnormalized densities depends on the true Z_1: however, this dependency is simple. We

[3] this effectively assumes that $K \to \infty$ faster than $p_1(\boldsymbol{x})p_0(\boldsymbol{x}) \to 0$, so that the error in the first limit is dominated by the error in the second limit.

propose a two-step estimation method: first, Z_1 is pre-estimated, for example using the geometric path; second, the estimate of Z_1 is plugged into the "oracle" or "oracle-trig" weight of the arithmetic path (table 2, column 2), and which is used to obtain a second estimation of Z_1. Note that pre-estimating a problematic (hyper)parameter, here Z_1, has proved beneficial to reduce the estimation error of NCE in a related context [32].

5 Numerical results

We now present numerical evidence for our theory and validate our two-step estimators. Importantly, we do *not* claim to achieve state of the art in terms of practical evaluation of the normalization constants; our goal is to support our theoretical analysis. We follow the evaluation methods of importance sampling literature [18] and evaluate our methods on synthetic gaussians. This setup is specially convenient for validating our theory: the optimal estimation error can conveniently be computed in closed-form, so too can the geometric and arithmetic paths which avoids a sampling error from MCMC algorithms. These derivations are included in the Appendix B. We specifically consider the high-dimensional setting, where the computation of the determinant of a high-dimensional (covariance) matrix which appears in the normalization of a gaussian, can in fact be challenging [33].

Numerical Methods The proposal distribution is always a standard gaussian, while the target differs by the second moment: $p_1 = \mathcal{N}(\mathbf{0}, 2\,\mathbf{Id})$ in Figure 2, $p_1 = \mathcal{N}(\mathbf{0}, 0.25\,\mathbf{Id})$ in Figure 3b and $p_1 = \mathcal{N}(\mathbf{0}, \sigma^2\,\mathbf{Id})$ in Figure 3a, where the target variance decreases as $\sigma(i) = i^{-1}$ so that the (natural) parameter distance grows linearly [30, Part II-4]. We use a sample size of $N = 50000$ points, and, unless otherwise mentioned, $K + 1 = 10$ distributions from the annealing paths and the dimensionality is 50. To compute an estimator of the normalization constant using the non-convex NCE loss, we used a non-linear conjugate gradient scheme implemented in Scipy [34]. We chose the conjugate-gradient algorithm as it is deterministic (no residual variance like in SGD). The empirical Mean-Squared Error was computed over 100 random seeds, parallelized over 100 CPUs. The longest experiment took 7 wall-clock hours to complete. For the two-step estimators ("two-step" and "two-step (trig)"), a pre-estimate of the normalization was first computed using the geometric path with 10 distributions. Then, this estimate was used to to re-parameterize an arithmetic path with 10 distributions which produced the second estimate.

Results Figure 2 numerically supports the optimality of the NCE loss for a finite K (here, $K = 2$ so three distributions are used) proven in Theorem 1. Figure 3 validates our main results for annealing paths. It shows how the estimation error scales with the proposal and target distributions growing apart, either with the parameter-distance in Figure 3a or with the dimensionality in Figure 3b. Using no annealing path ($K = 1$) produces an estimation error which grows linearly in log space; this numerically supports the exponential growth predicted by Theorem 2. Meanwhile, annealing ($K \to \infty$) sets the estimation error on different trends, depending on the choice of path. Choosing the geometric path brings the growth down to sub-exponential, as predicted by Theorem 3, while choosing the (basic) arithmetic path does not as in Theorem 4. To alleviate this, our two-step estimation methods consist in reparameterizing the arithmetic path so that it actually does bring down the estimation error. In fact, our two-step estimators in table 2 empirically approach the optimal estimation error in Figure 3. While this requires more computation, it has the appeal of making the estimation error *constant* with respect to the parameter-distance between the target and proposal distributions. Practically, this means that in Figure 3a, regular Noise-Contrastive Estimation (black, full line) fails when the parameter-distance between the target and proposal distributions is higher than 20, while our two-step estimators remain optimal.

We next explain interesting observations in Figure 3 which are actually coherent with our theory. First, in Figure 3a, the "two-step (trig)" estimator is only optimal when the parameter-distance between the target and proposal distributions is larger than 10. This is because the optimality of this two-step estimator was derived in Theorem 6 conditionally on non-overlapping distributions, here achieved by a large parameter-distance. Second, in both Figures 3a and 3b, the "two-step" estimator empirically achieves the optimal estimation error that was predicted for the "two-step (trig)" estimator. This suggests our polynomial upper bound from Theorem 5 may be loose in certain cases. This further explains why, in Figure 3a, the arithmetic path is near-optimal for a single value of the parameter-distance. At this value of 20, the partition function happens to be equal to one $Z(\theta_1) = 1$, so that the arithmetic path is effectively the same as the "two-step" estimator.

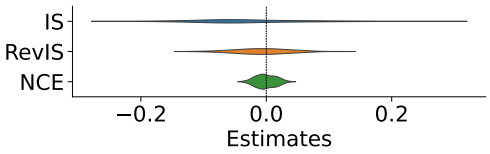
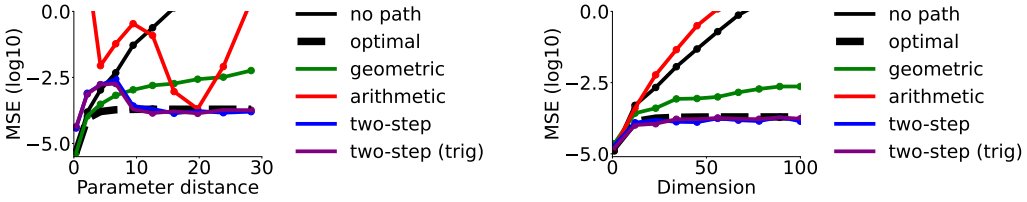


Figure 2: Optimality of the NCE loss, using the geometric path with $K = 2$.



(a) MSE against parameter distance.

(b) MSE against dimensionality.

Figure 3: Estimation error as the target and proposal distributions grow apart. Without annealing, the error is exponential in the parameter distance (diagonal in log-scale). Annealing with the geometric path and our two-step methods brings down the error to slower growth, as predicted by our theorems.

6 Related work and Limitations

Previous work has mainly focused on annealed importance sampling [18, 35], which is a special case of our annealed Bregman estimator. They have evaluated the merits of different paths empirically, using an approximation of the estimation error called Effective Sample Size (ESS) and the consistency-gap. In our analysis, we consider consistent estimators and derive and optimize the exact estimation error of the optimal Noise-Contrastive Estimation. Liu et al. [27] considered the NCE estimate for Z (not $\log Z$) with the name "discriminance sampling", and annealed the estimator using an extended state-space construction similar to Neal [10]. Their analysis of the estimation error is relevant but does not deal with hyperparameters other than the classification loss.

We made the common assumption of perfect sampling [18] in order to study the estimation error only and obtain practical guidelines to reduce it. We note however, that this leaves a gap to bridge with a practical setup where the sampling error cannot be ignored; in fact, annealed importance sampling [10] was originally proposed such that the samples can be obtained from a Markov Chain that has not converged. It might also be argued that the limit of almost no overlap between proposal and target, which we use a lot, is unrealistic. To see why it can be realistic, consider the case of natural image data. A typical proposal is Gaussian, since nothing much more sophisticated is tractable in high dimensions. However, there is almost no overlap between Gaussian data and natural images, which is seen in the fact that a human observer can effortlessly discriminate between the two.

7 Conclusion

We defined a class of estimators of the normalization constant, annealed Bregman estimation, which relies on a sampling phase from a path of distributions, and an estimation phase where these samples are used to estimate the log-normalization of the target distribution. Our results suggest a number of simple recommendations regarding hyperparameter choices in annealing. First, if the path has very few intermediate distributions, it is better to choose NCE due to its statistical optimality (Theorem 1). If however, the path has many intermediate distributions and approaches the annealing limit, it is better to use IS due to its computational simplicity (since the statistical efficiency is the same as NCE). Annealing can always provide substantial benefits (Theorem 2). Moreover, if we have a reasonable a priori estimate of Z_1, the arithmetic path achieves very low error (Theorem 5) — sometimes even approaching optimal (Theorem 6). On the other hand, even absent an initial estimate of Z_1, the geometric path can exponentially reduce the estimation error compared with no annealing (Theorems 2 and 3).

References

[1] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2021.

[2] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

[3] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(11): 307–361, 2012.

[4] G.E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[5] R. Gao, E. Nijkamp, D.P. Kingma, Z. Xu, A.M. Dai, and Y. Nian Wu. Flow contrastive estimation of energy-based models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7515–7525, 2020.

[6] A.B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[7] B. Liu, E. Rosenfeld, P. Ravikumar, and A. Risteski. Analyzing and improving the optimization landscape of noise-contrastive estimation. In *International Conference on Learning Representations (ICLR)*, 2022.

[8] H. Lee, C. Pabbaraju, A.P. Sevekari, and A. Risteski. Pitfalls of gaussians as a noise distribution in NCE. In *International Conference on Learning Representations (ICLR)*, 2023.

[9] O. Chehab, A. Gramfort, and A. Hyvärinen. The optimal noise in noise-contrastive learning is not what you think. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 180, pages 307–316. PMLR, 2022.

[10] R.M. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 1998.

[11] C. Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78: 2690–2693, 1996.

[12] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56:5018–5035, 1997.

[13] R. Salakhutdinov and G. Hinton. Replicated softmax: an undirected topic model. In *Neural Information Processing Systems (NIPS)*, 2009.

[14] Y. Dauphin and Y. Bengio. Stochastic ratio matching of rbms for sparse high-dimensional inputs. In *Neural Information Processing Systems (NIPS)*, volume 26. Curran Associates, Inc., 2013.

[15] V. Masrani, T.A. Le, and F.D. Wood. The thermodynamic variational objective. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[16] R. Brekelmans, S. Huang, M. Ghassemi, G. Ver Steeg, R.B. Grosse, and A. Makhzani. Improving mutual information estimation with annealed and energy-based bounds. In *International Conference on Learning Representations (ICLR)*, 2022.

[17] A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998.

[18] R. Grosse, C. Maddison, and R. Salakhutdinov. Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26. Curran Associates, Inc., 2013.

[19] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[20] X.-L. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

[21] V. Masrani, R. Brekelmans, T. Bui, F. Nielsen, A. Galstyan, G. Ver Steeg, and F. Wood. q-paths: Generalizing the geometric annealing path using power means. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 161, pages 1938–1947. PMLR, 27–30 Jul 2021.

[22] O. Chehab, A. Gramfort, and A. Hyvärinen. Optimizing the noise in self-supervised learning: from importance sampling to noise-contrastive estimation, 2023.

[23] M.A. Newton. Approximate bayesian-inference with the weighted likelihood bootstrap. *Journal of the royal statistical society series b-methodological*, 1994.

[24] A.W. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000.

[25] M.-H. Chen and Q.-M. Shao. On monte carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4):1563–1594, 1997.

[26] M. Uehara, T. Matsuda, and F. Komaki. Analysis of noise contrastive estimation from the perspective of asymptotic variance. *ArXiv*, 2018. doi: 10.48550/ARXIV.1808.07983.

[27] Q. Liu, J. Peng, A.T. Ihler, and J.W. Fisher III. Estimating the partition function by discriminance sampling. In *Uncertainty in Artificial Intelligence (UAI)*, 2015.

[28] A. Hyvärinen, J. Hurri, and P.O. Hoyer. Natural image statistics - a probabilistic approach to early computational vision. In *Computational Imaging and Vision*, 2009.

[29] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57): 1–59, 2017.

[30] F. Nielsen and V. Garcia. Statistical exponential families: A digest with flash cards, 2011.

[31] X.-L. Meng and S. Schilling. Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, 91(435): 1254–1267, 1996.

[32] M. Uehara, T. Kanamori, T. Takenouchi, and T. Matsuda. A unified statistically efficient estimation framework for unnormalized models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pages 809–819. PMLR, 2020.

[33] L. Ellam, H. Strathmann, M.A. Girolami, and I. Murray. A determinant-free method to simulate the parameters of large gaussian fields. *Stat*, 6:271–281, 2017.

[34] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, J.K. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, I. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272, 2020.

[35] T. Kiwaki. Variational optimization of annealing schedules, 2015.

[36] M. Pihlaja, M. Gutmann, and A. Hyvärinen. A family of computationally efficient and simple estimators for unnormalized statistical models. In *Uncertainty in Artificial Intelligence (UAI)*, 2010.

[37] L. Wang, D.E. Jones, and X.-L. Meng. Warp bridge sampling: The next generation. *Journal of the American Statistical Association*, 117(538):835–851, 2022.

[38] H. Xing. Improving bridge estimators via f-GAN. *Statistics and Computing*, 32, 2022.

[39] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2022.

[40] S.-I. Amari and H. Nagaoka. *Methods of information geometry*. AMS, 2000.

In the following, we will study the estimation error of of annealed Bregman estimation (ABE) in two importants setups: the log-normalization is computed using two distributions ($K = 1$), the proposal and the target, or else using a path of distributions ($K \to \infty$).

The anonymized code used for the experiments is available at `https://anonymous.4open.science/r/annealed-nce-submission-B8A0`.

A No annealing, $K = 1$

We use [22, Eq.21] for the estimation error of any suitably parameterized [4] classifier $F(\boldsymbol{x}; \boldsymbol{\beta})$ between two distributions p_1 and p_0. The estimation error is measured by the Mean-Squared Error (MSE)

$$\mathrm{MSE}_{\hat{\boldsymbol{\beta}}}(p_n, \nu, \phi, N) = \frac{\nu + 1}{N} \mathrm{tr}(\boldsymbol{\Sigma}) \tag{18}$$

which depends on the sample sizes $N = N_1 + N_0$, their ratio $\nu = N_1/N_0$, the Bregman classification loss indexed by the convex function $\phi(x)$, and the asymptotic variance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{I}_w^{-1} \big(\boldsymbol{I}_v - (1 + \frac{1}{\nu}) \boldsymbol{m}_w \boldsymbol{m}_w^\top \big) \boldsymbol{I}_w^{-1} \ . \tag{19}$$

Here, $\boldsymbol{m}_w(\boldsymbol{\beta}^*)$, $\boldsymbol{I}_w(\boldsymbol{\beta}^*)$ and $\boldsymbol{I}_v(\boldsymbol{\beta}^*)$ are the reweighted mean and covariances of the paramete-gradient of the classifier, also known as the "relative" Fisher score $\nabla_{\boldsymbol{\beta}} F(\boldsymbol{x}; \beta^*)$,

$$\boldsymbol{m}_w(\boldsymbol{\beta}^*) = \mathbb{E}_{\boldsymbol{x} \sim p_d} \big[w(\boldsymbol{x}) \nabla_{\boldsymbol{\beta}} F(\boldsymbol{x}; \beta^*) \big] \tag{20}$$

$$\boldsymbol{I}_w(\boldsymbol{\beta}^*) = \mathbb{E}_{\boldsymbol{x} \sim p_d} \big[w(\boldsymbol{x}) \nabla_{\boldsymbol{\beta}} F(\boldsymbol{x}; \beta^*) \, \nabla_{\boldsymbol{\beta}} F(\boldsymbol{x}; \beta^*)^\top \big] \tag{21}$$

$$\boldsymbol{I}_v(\boldsymbol{\beta}^*) = \mathbb{E}_{\boldsymbol{x} \sim p_d} \big[v(\boldsymbol{x}) \nabla_{\boldsymbol{\beta}} F(\boldsymbol{x}; \beta^*) \, \nabla_{\boldsymbol{\beta}} F(\boldsymbol{x}; \beta^*)^\top \big] \tag{22}$$

where the reweighting of data points is by $w(\boldsymbol{x}) := \frac{p_1}{\nu p_0}(\boldsymbol{x}) \phi''\big(\frac{p_1}{\nu p_0}(\boldsymbol{x})\big)$ and by $v(\boldsymbol{x}) = w(\boldsymbol{x})^2 \frac{\nu p_0(\boldsymbol{x}) + p_1(\boldsymbol{x})}{\nu p_0(\boldsymbol{x})}$, which are all evaluated at the true parameter value $\boldsymbol{\beta}^*$.

Scalar parameterization We now consider a specific parameterization of the classifier:

$$F(\boldsymbol{x}; \beta) = \log \left(\frac{f_1(\boldsymbol{x})}{\nu f_0(\boldsymbol{x})} \right) - \beta \tag{23}$$

where the optimal parameter is the log-ratio of normalizations $\beta^* = \log(Z_1/Z_0)$. Consequently, we have $\nabla_{\boldsymbol{\beta}} F(\boldsymbol{x}; \beta^*) = -1$ and plugging this into the above quantities yields

$$\mathrm{MSE} = \frac{1 + \nu}{T} \Big(\frac{\mathbb{E}_{\boldsymbol{x} \sim p_1} \big[w^2(\boldsymbol{x}) \frac{\nu p_0(\boldsymbol{x}) + p_1(\boldsymbol{x})}{\nu p_0(\boldsymbol{x})} \big]}{\mathbb{E}_{\boldsymbol{x} \sim p_1} [w(\boldsymbol{x})]^2} - (1 + \frac{1}{\nu}) \Big)$$

which matches the formula found in [20, Eq 3.2]. For different choices of the Bregman classification loss, the estimation error is written using a divergence between the two distributions

Name	Loss identified by $\phi(\boldsymbol{x})$	Estimator	MSE
IS	$x \log x$	$\log \mathbb{E}_{p_0} \frac{f_1}{f_0}$	$\frac{1+\nu}{\nu N} \mathcal{D}_{\chi^2}(p_1, p_0)$
RevIS	$-\log x$	$-\log \mathbb{E}_{p_1} \frac{f_0}{f_1}$	$\frac{1+\nu}{N} \mathcal{D}_{\chi^2}(p_0, p_1)$
NCE	$x \log x - (1+x) \log(\frac{1+x}{2})$	implicit	$\frac{(1+\nu)^2}{\nu N} \frac{\mathcal{D}_{\mathrm{HM}}(p_1, p_0)}{1 - \mathcal{D}_{\mathrm{HM}}(p_1, p_0)}$
IS-RevIS	$(1 - \sqrt{x})^2$	$\log \mathbb{E}_{p_0} \frac{f_1}{f_0} - \log \mathbb{E}_{p_1} \frac{f_0}{f_1}$	$\frac{(1+\nu)^2}{\nu N} \frac{1 - (1 - \mathcal{D}_{H^2}(p_d, p_n))^2}{(1 - \mathcal{D}_{H^2}(p_d, p_n))^2}$

where

$\mathcal{D}_{\chi^2}(p_1, p_0) := \big(\int \frac{p_1^2}{p_0} \big) - 1$ is the chi-squared divergence

[4] technically, the formula was derived in [36, 3] assuming the classifier was parameterized as $F(\boldsymbol{x}; \boldsymbol{\beta}) = \log p_1(\boldsymbol{x}; \boldsymbol{\beta}) / \nu p_0(\boldsymbol{x})$ but the proof seems to generalize to any well-defined parameterization $F(\boldsymbol{x}; \boldsymbol{\beta})$.

$D_{H^2}(p_1, p_0) := 1 - \left(\int \sqrt{p_1 p_0}\right) \in [0, 1]$ is the squared Hellinger distance

$\mathcal{D}_{\mathrm{HM}}(p_1, p_0) := 1 - \int \left(\pi p_1^{-1} + (1-\pi) p_0^{-1}\right)^{-1} = 1 - \frac{1}{\pi}\mathbb{E}_{p_1}\frac{\pi p_0}{(1-\pi)p_1 + \pi p_0} \in [0, 1]$ is the harmonic divergence with weight $\pi \in [0, 1]$.

Here, the weight $\pi = P(Y = 0) = \frac{T_n}{T} = \frac{\nu}{1+\nu}$.

Proof of Theorem 2 *Exponential error of binary NCE*

The estimation error of binary NCE is expressed in terms of the harmonic divergence

$$\mathrm{MSE} = \frac{4}{N}\frac{\mathcal{D}_{\mathrm{HM}}(p_1, p_0)}{1 - \mathcal{D}_{\mathrm{HM}}(p_1, p_0)} \tag{24}$$

which is intractable for general exponential families. Instead, we can lower-bound the estimation error. To do so, we lower-bound the harmonic divergence using the inequality of means (harmonic vs. geometric)

$$\mathcal{D}_{\mathrm{HM}}(p_1, p_0) = 1 - \int \frac{2p_0 p_1}{p_0 + p_1} \geq 1 - \int \sqrt{p_0 p_1} = \mathcal{D}_{H^2}(p_0, p_1) \tag{25}$$

and therefore

$$\mathrm{MSE}_{\mathrm{LB}} = \frac{4}{N}\frac{\mathcal{D}_{H^2}(p_1, p_0)}{1 - \mathcal{D}_{H^2}(p_1, p_0)} \ . \tag{26}$$

This lower bound is expressed in terms of the squared Hellinger distance, that is tractable for exponential families:

$$\begin{aligned}
\mathcal{D}_{H^2}(p_1, p_0) &:= 1 - \int_{\boldsymbol{x} \in \mathbb{R}^D} \sqrt{p_1 p_0} d\boldsymbol{x} && (27)\\
&= 1 - \int_{\boldsymbol{x} \in \mathbb{R}^D} \frac{1}{Z(\boldsymbol{\theta}_1)^{\frac{1}{2}} Z(\boldsymbol{\theta}_0)^{\frac{1}{2}}} \exp\left(\frac{1}{2}(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_0)^\top \boldsymbol{t}(\boldsymbol{x})\right) d\boldsymbol{x} && (28)\\
&= 1 - \frac{Z(\frac{1}{2}\boldsymbol{\theta}_1 + \frac{1}{2}\boldsymbol{\theta}_0)}{Z(\boldsymbol{\theta}_1)^{\frac{1}{2}} Z(\boldsymbol{\theta}_0)^{\frac{1}{2}}} && (29)\\
&= 1 - \exp\left(\log Z\left(\frac{1}{2}\boldsymbol{\theta}_1 + \frac{1}{2}\boldsymbol{\theta}_0\right) - \frac{1}{2}\log Z(\boldsymbol{\theta}_1) - \frac{1}{2}\log Z(\boldsymbol{\theta}_0)\right) \ . && (30)
\end{aligned}$$

We now wish to lower bound $\mathrm{MSE}_{\mathrm{LB}}$, and therefore $\mathcal{D}_{H^2}(p_1, p_0)$, by an expression which is exponential in the parameter distance $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|$. To do so, we note that for exponential families, the log-normalization is convex in the parameters. Here, we further assume strong convexity, so that

$$\log Z\left(\frac{1}{2}\boldsymbol{\theta}_1 + \frac{1}{2}\boldsymbol{\theta}_0\right) \leq \frac{1}{2}\log Z(\boldsymbol{\theta}_1) + \frac{1}{2}\log Z(\boldsymbol{\theta}_0) - \frac{1}{8}M\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2 \tag{31}$$

where M is the strong convexity constant. Plugging this back into the squared Hellinger distance, we obtain

$$\mathcal{D}_{H^2}(p_1, p_0) \geq 1 - \exp\left(-\frac{1}{8}M\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2\right) \tag{32}$$

so that the MSE

$$\mathrm{MSE} \geq \frac{4}{N}\frac{\mathcal{D}_{H^2}(p_1, p_0)}{1 - \mathcal{D}_{H^2}(p_1, p_0)} \geq \frac{4}{N}\exp\left(\frac{1}{8}M\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2\right) - 1 \tag{33}$$

grows exponentially with the euclidean distance between the parameters.

B Annealing limit, $K \to \infty$

We now consider annealing paths $(p_t)_{t\in[0,1]}$ that interpolate between between the proposal p_0 and the target p_1.

B.1 Estimation error

We first show the optimality of the NCE loss within the family of Annealed Bregman Estimators, in the sense that it produces the smallest estimation error. We then study the estimation error of different annealed Bregman estimators in the annealing limit of a continuous path ($K \to \infty$).

Proof of Theorem 1 *Optimality of the NCE loss and the estimation error in the annealing limit* $K \to \infty$

- Optimality of the NCE loss
  Because the annealed Bregman estimator is built by adding independent estimators
  $$\widehat{\log Z_1} = \sum_{k=0}^{K-1} \widehat{\log\left(\frac{Z_{(k+1)/K}}{Z_{k/K}}\right)} + \log Z_0 \ . \tag{34}$$
  the total Mean Squared Error (MSE) is the sum of each MSEs for each estimator (indexed by $k \in [\![0, K-1]\!]$)
  $$\mathrm{MSE}((\phi_k)_{k\in[\![0,K]\!]}) = \sum_{k=0}^{K-1} \mathrm{MSE}_k(\phi_k) \tag{35}$$
  where we highlighted the dependency on the classification losses identified by $(\phi_k)_{k\in[\![0,K]\!]}$. The MSEs follow Eq. 24. It was shown by Meng and Wong [20] that for any of these MSEs, the optimal loss is identified by $\phi_k(x) = x\log x - (1+x)\log(\frac{1+x}{2})$ and is in fact the NCE loss [22]. Thus the sum of MSEs is minimized for the same loss.
- Annealed Noise-Contrastive Estimation (NCE)
  We are interested in the estimation error (asymptotic MSE) obtained for the NCE loss. Based off table 1, it is written as
  $$\mathrm{MSE} = \frac{4K}{N}\sum_{k=0}^{K-1} \frac{\mathcal{D}_{\mathrm{HM}}(p_{k/K}, p_{(k+1)/K})}{1-\mathcal{D}_{\mathrm{HM}}(p_{k/K}, p_{(k+1)/K})} \ . \tag{36}$$
  The estimation error of balanced ($\nu = 1$) NCE-JS between two "close" distributions p_t and p_{t+h}, is
  $$\mathrm{MSE}(p_t, p_{t+h}) \propto \frac{\mathcal{D}_{\mathrm{HM}}(p_t, p_{t+h})}{1-\mathcal{D}_{\mathrm{HM}}(p_t, p_{t+h})} \tag{37}$$
  The estimation error can be simplified using a Taylor expansion. To do so, we recall that D_{HM} is an f-divergence generated by $\phi(x) = 1 - \frac{x}{\pi+(1-\pi)x}$ [37, 38] ($\pi = \frac{1}{2}$ here) and its expansion is therefore [39, Eq.7.64]
  $$\mathcal{D}_{\mathrm{HM}}(p_t, p_{t+h}) = \frac{1}{2}h^2\nabla_t^2\mathcal{D}_{\mathrm{HM}}(p_t, p_{t+h}) + o(h^2) \tag{38}$$
  $$= \frac{1}{2}\phi''(1)h^2 I(t) + o(h^2) = \frac{1}{4}h^2 I(t) + o(h^2) \ . \tag{39}$$
  It follows that
  $$\frac{\mathcal{D}_{\mathrm{HM}}(p_t, p_{t+h})}{1-\mathcal{D}_{\mathrm{HM}}(p_t, p_{t+h})} = \frac{1}{4}I(t)h^2 + o(h^2) \ . \tag{40}$$
  Summing these estimation errors along the path of distributions with $h = 1/K$,
  $$\mathrm{MSE} = \frac{4K}{N}\sum_{k=0}^{K-1}\left(\frac{1}{4}I(t)\frac{1}{K^2} + o\Big(\frac{1}{K^2}\Big)\right) \tag{41}$$
  $$= \left(\frac{1}{NK}\sum_{k=0}^{K-1} I(t)\right) + o(1) \underset{K\to\infty}{\sim} \frac{1}{N}\int_0^1 I(t)dt \ . \tag{42}$$

In the case of a parametric path $p(\boldsymbol{x}|\boldsymbol{\theta}(t))_{t\in[0,1]}$, the proof is the same. Simply, the second-order term in the Taylor expansion of Eq. 39 is computed using the chain rule

$$\nabla_t^2 \mathrm{MSE}(p_{\boldsymbol{\theta}(t)}, p_{\boldsymbol{\theta}(t+h)}) \tag{43}$$

$$= \dot{\boldsymbol{\theta}}(t)^\top \nabla_{\boldsymbol{\theta}}^2 \mathrm{MSE}(p_{\boldsymbol{\theta}(t)}, p_{\boldsymbol{\theta}(t+h)}) \dot{\boldsymbol{\theta}}(t) + \ddot{\boldsymbol{\theta}}(t)^\top \nabla_{\boldsymbol{\theta}} \mathrm{MSE}(p_{\boldsymbol{\theta}(t)}, p_{\boldsymbol{\theta}(t+h)}) \tag{44}$$

$$= \dot{\boldsymbol{\theta}}(t)^\top \nabla_{\boldsymbol{\theta}}^2 \mathrm{MSE}(p_{\boldsymbol{\theta}(t)}, p_{\boldsymbol{\theta}(t+h)}) \dot{\boldsymbol{\theta}}(t)^\top + 0 \tag{45}$$

$$= \dot{\boldsymbol{\theta}}(t)^\top \boldsymbol{I}(\theta(t)) \dot{\boldsymbol{\theta}}(t) \tag{46}$$

- Annealed importance sampling (IS)
  Similarly, for the choice of the importance sampling base estimator,

$$\mathrm{MSE} = \frac{2K}{N} \sum_{k=0}^{K-1} \mathcal{D}_{\chi^2}(p_{(k+1)/K}, p_{k/K}) \tag{47}$$

$$= \frac{2K}{N} \sum_{k=0}^{K-1} \mathcal{D}_{\mathrm{rev}\chi^2}(p_{k/K}, p_{(k+1)/K}) = \frac{2K}{N} \sum_{k=0}^{K-1} \left(\frac{1}{2} \phi^{''}(1) I(t) \frac{1}{K^2} + o\Big(\frac{1}{K^2}\Big) \right) \tag{48}$$

$$= \left(\frac{1}{NK} \sum_{k=0}^{K-1} I(t) \right) + o(1) \underset{K\to\infty}{\sim} \frac{1}{N} \int_0^1 I(t) dt \ . \tag{49}$$

  given that $\phi(x) = -\log(x)$ and therefore $\phi^{''}(1) = 1$ for the reverse χ^2 divergence.
- Annealed reverse importance sampling (RevIS)
  Similarly, for the choice of the reverse importance sampling base estimator,

$$\mathrm{MSE} = \frac{2K}{N} \sum_{k=0}^{K-1} \mathcal{D}_{\chi^2}(p_{k/K}, p_{(k+1)/K}) \tag{50}$$

$$= \frac{2K}{N} \sum_{k=0}^{K-1} \left(\frac{1}{2} \phi^{''}(1) I(t) \frac{1}{K^2} + o\Big(\frac{1}{K^2}\Big) \right) \tag{51}$$

$$= \left(\frac{1}{NK} \sum_{k=0}^{K-1} I(t) \right) + o(1) \underset{K\to\infty}{\sim} \frac{1}{N} \int_0^1 I(t) dt \ . \tag{52}$$

  given that $\phi(x) = x\log(x)$ and therefore $\phi^{''}(1) = 1$ for the χ^2 divergence.

B.2 Examples of paths

Geometric path The geometric path is defined in the space of unnormalized densities by

$$f_t(\boldsymbol{x}) := p_0(\boldsymbol{x})^{1-t} f_1(\boldsymbol{x})^t = p_0(\boldsymbol{x})^{1-t} p_1(\boldsymbol{x})^t Z_1^t \propto p_0(\boldsymbol{x})^{1-t} p_1(\boldsymbol{x})^t \tag{53}$$

so in the space of normalized densities, the path is

$$p_t := \frac{p_0(\boldsymbol{x})^{1-t} p_1(\boldsymbol{x})^t}{Z_t} \tag{54}$$

where the normalization is

$$Z_t := \int_{\boldsymbol{x}\in\mathbb{R}^d} p_0(\boldsymbol{x})^{1-t} p_1(\boldsymbol{x})^t d\boldsymbol{x} = \mathbb{E}_{\boldsymbol{x}\sim p_1}\left[\left(\frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})} \right)^t \right] = \mathbb{E}_{\boldsymbol{x}\sim p_0}\left[\left(\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} \right)^{1-t} \right] \ . \tag{55}$$

Arithmetic path The arithmetic path is defined in the space of unnormalized densities by

$$f_t(\boldsymbol{x}) := (1-t)p_0(\boldsymbol{x}) + t f_1(\boldsymbol{x}) = (1-t)p_0 + t Z_1 p_1 \tag{56}$$

$$\propto \frac{(1-t)}{(1-t)+tZ_1} p_0 + \frac{tZ_1}{(1-t)+tZ_1} p_1 \tag{57}$$

so in the space of normalized densities, the path is actually a mixture between the target and the proposal, where the weight of the mixture is a nonlinear function of the target normalization

$$p_t := (1-\tilde{w}_t)p_0 + \tilde{w}_t p_1, \quad \tilde{w}_t = \frac{tZ_1}{(1-t)+tZ_1} \ . \tag{58}$$

Optimal path We know (*e.g.* from Gelman and Meng [17, Eq. 49]) that the optimal path is

$$p_t(\boldsymbol{x}) = \big(a(t)\sqrt{p_0(\boldsymbol{x})} + b(t)\sqrt{p_1(\boldsymbol{x})}\big)^2 \tag{59}$$

where the coefficients $a(t)$ and $b(t)$

$$a(t) = \frac{\cos((2t-1)\alpha_H)}{2\cos(\alpha_H)} - \frac{\sin((2t-1)\alpha_H)}{2\sin(\alpha_H)} \tag{60}$$

$$b(t) = \frac{\cos((2t-1)\alpha_H)}{2\cos(\alpha_H)} + \frac{\sin((2t-1)\alpha_H)}{2\sin(\alpha_H)} \tag{61}$$

are simple functions of the squared Hellinger distance $\mathcal{D}_{H^2}(p_0, p_1)$ between the proposal and the target [5]

$$\alpha_H = \arctan\left(\sqrt{\frac{\mathcal{D}_{H^2}(p_0,p_1)}{2-\mathcal{D}_{H^2}(p_0,p_1)}}\right) \in [0, \frac{\pi}{4}] \ . \tag{62}$$

The estimation error produced by that optimal path is [17, Eq. 48]

$$\text{MSE} = \frac{1}{N}\int_0^1 I(t)dt = \frac{1}{N} 16\alpha_H^2 \ . \tag{63}$$

For two gaussians

$$p_0 := \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \tag{64}$$

$$p_1 := \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \tag{65}$$

the squared Hellinger distance can be written in closed-form

$$\mathcal{D}_{H^2}(p_0,p_1) = 1 - \frac{|\boldsymbol{\Sigma}_0|^{\frac{1}{4}}|\boldsymbol{\Sigma}_1|^{\frac{1}{4}}}{|\frac{1}{2}\boldsymbol{\Sigma}_0 + \frac{1}{2}\boldsymbol{\Sigma}_1|^{\frac{1}{2}}} \exp\Big(-\frac{1}{8}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^\top\big(\frac{1}{2}\boldsymbol{\Sigma}_0 + \frac{1}{2}\boldsymbol{\Sigma}_1\big)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)\Big) \tag{66}$$

and plugs into the optimal path formula, which is also obtained in closed-form.

B.3 Estimation error from taking different paths

Proof of Theorem 3 *Polynomial error of annealed NCE with the geometric path*

We next study the estimation error produced by the geometric path (Figure 1). In the annealing limit $K \to \infty$, the MSE is written as

$$\text{MSE} = \frac{1}{N}\int_0^1 I(t)dt \ . \tag{67}$$

We recall from Grosse et al. [18] that the geometric path is closed for distributions in the exponential family: all distributions along the path remain in the exponential family. Furthermore, their Fisher information can be written in terms of the terms parameters [18, Eq. 17]; this is based off a a result of exponential families from [40, Section 3.3]

$$I(t) = \dot{\boldsymbol{\theta}}(t)^\top \boldsymbol{I}(\boldsymbol{\theta}(t))\dot{\boldsymbol{\theta}}(t) = \dot{\boldsymbol{\theta}}(t)^\top \dot{\boldsymbol{\mu}}(t) \tag{68}$$

where $\boldsymbol{\mu}(t)$ are the generalized moments, defined as $\boldsymbol{\mu}(t) = \mathbb{E}_{\boldsymbol{x}\sim p_t(\boldsymbol{x})}[\boldsymbol{t}(\boldsymbol{x})] = \nabla_{\boldsymbol{\theta}} \log Z_t(\boldsymbol{\theta})$. It follows,

$$\text{MSE} = \frac{1}{N}\int_0^1 \dot{\boldsymbol{\theta}}(t)^\top \dot{\boldsymbol{\mu}}(t)dt \ . \tag{69}$$

The geometric path is defined in parameter space by $\boldsymbol{\theta}_t = t\boldsymbol{\theta}_1 + (1-t)\boldsymbol{\theta}_0$, therefore

$$\text{MSE} = \frac{1}{N}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_0)^\top \int_0^1 \dot{\boldsymbol{\mu}}(t)dt = \frac{1}{N}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_0)^\top(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) \tag{70}$$

[5] In Gelman and Meng [17, Eq. 49], the Hellinger distance is defined such that it is in $[0, \sqrt{2}]$. We here instead use the conventional definition of the squared Hellinger distance which is normalized so that it is in $[0, 1]$.

as in [18, Eq. 17]. For exponential families, $\log Z(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$. Here, we further assume strong convexity (with constant M) and smoothness (with constant L) so that

$$(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^\top (\nabla_{\boldsymbol{\theta}} \log Z_t(\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} \log Z_t(\boldsymbol{\theta}_0)) \tag{71}$$

$$\leq \frac{1}{M} \|\nabla_{\boldsymbol{\theta}} \log Z_t(\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} \log Z_t(\boldsymbol{\theta}_0)\|^2 \leq \frac{L^2}{M} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2 \tag{72}$$

so that the MSE

$$\text{MSE} \leq \frac{L^2}{MN} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2 \tag{73}$$

is polynomial in the euclidean distance between the parameters.

Proof of Theorem 4 *Exponential error of annealed NCE with the arithmetic path and "vanilla" schedule*

We now study the estimation error produced by the arithmetic path with "vanilla" schedule (table 2, line 3). Similarly, we start with the formula of the estimation error of NCE in the limit of a continuous path

$$\text{MSE} = \frac{1}{N} \int_0^1 I(t) dt \ . \tag{74}$$

where $I(t) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x},t)}[(\frac{d}{dt} \log p(\boldsymbol{x}, t))^2]$ is the Fisher information over the path, using time t as the parameter. The arithmetic path is a gaussian mixture (see table 2) so we will conveniently use the parametric form of the path to compute the Fisher information

$$I(t) = \dot{\tilde{w}}_t^\top I(\tilde{w}_t) \dot{\tilde{w}}_t \tag{75}$$

where the parameter here is the weight of the Gaussian mixture $\tilde{w}_t = tZ_1/(tZ_1 + 1 - t)$. We will need to compute two quantites: the Fisher information to that mixture parameter (not the time), and the parameter speed $\dot{\tilde{w}}_t$.

$$I(\tilde{w}_t) := \mathbb{E}_{\boldsymbol{x} \sim p_{\tilde{w}_t}} \left[\left(\frac{\partial \log p_{\tilde{w}_t}}{\partial \tilde{w}_t}(\boldsymbol{x}) \right)^2 \right] = \mathbb{E}_{\boldsymbol{x} \sim p_{\tilde{w}_t}} \left[\left(\frac{1}{p_{\tilde{w}_t}(\boldsymbol{x})} \frac{\partial p_{\tilde{w}_t}}{\partial \tilde{w}_t}(\boldsymbol{x}) \right)^2 \right] \tag{76}$$

$$= \int_{\boldsymbol{x} \in \mathbb{R}^D} \frac{(p_1(\boldsymbol{x}) - p_0(\boldsymbol{x}))^2}{p_{\tilde{w}_t}(\boldsymbol{x})} d\boldsymbol{x} = \int_{\boldsymbol{x} \in \mathbb{R}^D} \frac{(p_1(\boldsymbol{x}) - p_0(\boldsymbol{x}))^2}{(1 - \tilde{w}_t) p_0(\boldsymbol{x}) + \tilde{w}_t p_1(\boldsymbol{x})} d\boldsymbol{x} \tag{77}$$

$$\geq \int_{\boldsymbol{x} \in \mathbb{R}^D} \frac{(p_1(\boldsymbol{x}) - p_0(\boldsymbol{x}))^2}{p_0(\boldsymbol{x}) + p_1(\boldsymbol{x})} d\boldsymbol{x} = \int p_0(\boldsymbol{x}) \frac{\left(1 - \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}\right)^2}{1 + \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}} = \mathcal{D}_\phi(p_1, p_0) \tag{78}$$

which is an f-divergence with generator $\phi(\boldsymbol{x}) = (1 - x)^2/(1 + x)$ that provides a t-independent lower bound. This will allow us to factor this quantity out of the integral defining the MSE, and simplify computations. We also have

$$\dot{\tilde{w}}_t := \frac{\partial}{\partial t} \tilde{w}_t = \frac{1}{t(1-t)} \times \sigma\left(\log \frac{tZ_1}{1-t} \right) \times \left(1 - \sigma\left(\log \frac{tZ_1}{1-t} \right) \right) \tag{79}$$

$$= \frac{1}{t(1-t)} \times \frac{tZ_1}{(1-t) + tZ_1} \times \frac{(1-t)}{(1-t) + tZ_1} = \frac{Z_1}{((1-t) + tZ_1)^2} \ . \tag{80}$$

where we choose to keep the dependency on t. The intuition is that integrating this quantity will yield a function of Z_1, which will drive the MSE toward high values. We next show this rigorously and finally compute the estimation error.

$$\text{MSE} = \frac{1}{N} \int_0^1 I(t) dt = \frac{1}{N} \int_0^1 \dot{\tilde{w}}(t) I(\tilde{w}(t)) \dot{\tilde{w}}(t) dt \tag{81}$$

$$\geq \frac{1}{N} \times \mathcal{D}_\phi(p_1, p_0) \times \int_0^1 \dot{\tilde{w}}(t)^2 dt = \frac{1}{N} \times \mathcal{D}_\phi(p_1, p_0) \times Z_1^2 \times \int_0^1 \frac{1}{(t(Z_1 - 1) + 1)^4} dt \tag{82}$$

$$= \frac{1}{N} \times \mathcal{D}_\phi(p_1, p_0) \times Z_1^2 \times \frac{Z_1^2 + Z_1 + 1}{3 Z_1^3} = \frac{1}{3N} \times \mathcal{D}_\phi(p_1, p_0) \times (Z_1^{-1} + 1 + Z_1) \ . \tag{83}$$

We would like to write Z_1 in terms of the parameters. To do so, we now suppose the unnormalized target is in a simply unnormalized exponential family. Consequently,

$$Z_1 := \exp(\log Z(\boldsymbol{\theta}_1) - \log Z(\boldsymbol{\theta}_0) + \log Z(\boldsymbol{\theta}_0)) \tag{84}$$

$$\geq \exp\left(\nabla \log Z(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) + \frac{M}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2 + \log Z(\boldsymbol{\theta}_0)\right) \ . \tag{85}$$

using the strong convexity of the log-partition function. It follows that in the limit of parameter-distance $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\| \to \infty$, the MSE grows (at least) exponentially with the parameter-distance

$$\text{MSE} = O\Big(\frac{1}{3N} \times \mathcal{D}_\phi(p_1, p_0) \times \exp\left(\frac{M}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2\right)\Big), \quad \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2 \to \infty \tag{86}$$

Proof of Theorem 5 *Polynomial error of annealed NCE with the arithmetic path and "oracle" schedule*

We now study the estimation error produced by the arithmetic path with "oracle" schedule (table 2, line 4). Similarly, we start with the formula of the estimation error of NCE annealed over a continuous path

$$\text{MSE} = \frac{1}{N}\int_0^1 I(t)dt \ . \tag{87}$$

where $I(t) = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x},t)}[(\frac{d}{dt}\log p(\boldsymbol{x},t))^2]$ is the Fisher information over the path, using time t as the parameter. The arithmetic path is the gaussian mixture $p_t(\boldsymbol{x}) = tp_1(\boldsymbol{x}) + (1-t)p_0(\boldsymbol{x})$ (see table 2). The Fisher information is therefore

$$I(t) := \mathbb{E}_{\boldsymbol{x}\sim p_t}\left[\left(\frac{\partial \log p_t}{\partial t}(\boldsymbol{x})\right)^2\right] = \mathbb{E}_{\boldsymbol{x}\sim p_t}\left[\left(\frac{1}{p_t(\boldsymbol{x})}\frac{\partial p_t}{\partial t}(\boldsymbol{x})\right)^2\right] \tag{88}$$

$$= \int_{\boldsymbol{x}\in\mathbb{R}^D} \frac{(p_1(\boldsymbol{x}) - p_0(\boldsymbol{x}))^2}{p_t(\boldsymbol{x})}d\boldsymbol{x} = \int_{\boldsymbol{x}\in\mathbb{R}^D} \frac{(p_1(\boldsymbol{x}) - p_0(\boldsymbol{x}))^2}{(1-t)p_0(\boldsymbol{x}) + tp_1(\boldsymbol{x})}d\boldsymbol{x} \tag{89}$$

$$\leq \int_{\boldsymbol{x}\in\mathbb{R}^D} \frac{p_1(\boldsymbol{x})^2 + p_0(\boldsymbol{x})^2}{(1-t)p_0(\boldsymbol{x}) + tp_1(\boldsymbol{x})}d\boldsymbol{x} \tag{90}$$

where we choose to keep the dependency on t in the bound.

We briefly justify this choice. We had first tried a t-independent bound, which led to an upper bound of the MSE that was too loose. We share insight as to why: first, recognize that the fraction can be broken in two terms, each of them a chi-square divergence between an endpoint of the path (p_0 or p_1) and the mixture p_t. Each of them admits a t-independent upper bound given by the chi-square divergence between the endpoints p_0 and p_1, using lemma 1. However, the chi-square divergence between two gaussians, for example, is exponential (not polynomial) in the natural parameters [39, eq 7.41]. In fact, plotting $I(t)$ for a univariate gaussian model revealed that it took high values at the endpoints $t = 0$ and $t = 1$, and was near zero almost everywhere else in the interval $t \in [0, 1]$, which again suggested that dropping the dependency on t was unreasonable.

Now we can compute the estimation error, as

$$\text{MSE} = \frac{1}{N}\int_0^1 I(t)dt \leq \frac{1}{N}\int_{\mathbb{R}^d}\int_0^1 \frac{p_1(\boldsymbol{x})^2 + p_0(\boldsymbol{x})^2}{(1-t)p_0(\boldsymbol{x}) + tp_1(\boldsymbol{x})}dtd\boldsymbol{x} = \frac{1}{N}(J_1 + J_2) \ . \tag{91}$$

Let us try to solve one of these integrals, say J_1.

$$J_1 = \int_{\mathbb{R}^d}\int_0^1 \frac{p_1(\boldsymbol{x})^2}{(1-t)p_0(\boldsymbol{x}) + tp_1(\boldsymbol{x})}dtd\boldsymbol{x} = \int_{\mathbb{R}^d}\frac{p_1(\boldsymbol{x})^2}{p_0(\boldsymbol{x})}\Big(\int_0^1 \frac{1}{1 + t(\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} - 1)}dt\Big)d\boldsymbol{x} \tag{92}$$

$$= \int_{\mathbb{R}^d}\frac{p_1(\boldsymbol{x})^2}{p_0(\boldsymbol{x})}\Big(\frac{1}{\frac{p_1}{p_0} - 1}\log\frac{p_1}{p_0}\Big)d\boldsymbol{x} = 1 + \mathbb{E}_{p_1}\left[\frac{1}{1 - \frac{p_0}{p_1}}\log\frac{p_1}{p_0} - 1\right] = 1 + \mathcal{D}_\phi(p_0, p_1) \ . \tag{93}$$

which we rewrote using an f-divergence defined by $\phi(x) = \frac{-\log(x)}{1-x} - 1$. Similarly, we obtain

$$J_2 = \int_{\mathbb{R}^d}\int_0^1 \frac{p_0(\boldsymbol{x})^2}{(1-t)p_0(\boldsymbol{x}) + tp_1(\boldsymbol{x})}dtd\boldsymbol{x} = \int_{\mathbb{R}^d}\frac{p_0(\boldsymbol{x})^2}{p_1(\boldsymbol{x})}\Big(\int_0^1 \frac{1}{\frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})} + t(1 - \frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})})}dt\Big)d\boldsymbol{x} \tag{94}$$

$$= \int_{\mathbb{R}^d}\frac{p_0(\boldsymbol{x})^2}{p_1(\boldsymbol{x})}\Big(\frac{1}{\frac{p_0}{p_1} - 1}\log\frac{p_0}{p_1}\Big)d\boldsymbol{x} = 1 + \mathbb{E}_{p_0}\left[\frac{1}{1 - \frac{p_1}{p_0}}\log\frac{p_0}{p_1} - 1\right] = 1 + \mathcal{D}_\phi(p_1, p_0) \ . \tag{95}$$

Putting this together, we get

$$\text{MSE} \leq \frac{1}{N}(2 + \mathcal{D}_\phi(p_0, p_1) + \mathcal{D}_\phi(p_1, p_0)) \ . \tag{96}$$

How does this divergence depend on the parameter-distance $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|$? Does it bring down the dependency from exponential to something lower? We next analyze this:

$$\mathcal{D}_\phi(p_0, p_1) + 1 = \mathbb{E}_{p_1} \frac{1}{1 - \frac{p_0}{p_1}} \log \frac{p_1}{p_0}$$

which looks like a Kullback-Leibler divergence, where the integrand is reweighted by $\frac{1}{1-\frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})}}$. Note that

$$\begin{cases} \frac{1}{1-\frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})}} \geq 1 & p_0(\boldsymbol{x}) \leq p_1(\boldsymbol{x}) \\ \frac{1}{1-\frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})}} < 1 & p_0(\boldsymbol{x}) > p_1(\boldsymbol{x}) \end{cases} \tag{97}$$

which motivates separating the integral over both domains

$$1 + \mathcal{D}_\phi(p_0, p_1) = \int_{\{\boldsymbol{x} \in \mathbb{R}^D | p_0(\boldsymbol{x}) \leq p_1(\boldsymbol{x})\}} p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} \frac{1}{1 - \frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})}} \tag{98}$$

$$+ \int_{\{\boldsymbol{x} \in \mathbb{R}^D | p_0(\boldsymbol{x}) > p_1(\boldsymbol{x})\}} p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} \frac{1}{1 - \frac{p_0(\boldsymbol{x})}{p_1(\boldsymbol{x})}} \tag{99}$$

$$\leq \int_{\{\boldsymbol{x} \in \mathbb{R}^D | p_0(\boldsymbol{x}) \leq p_1(\boldsymbol{x})\}} p_1(\boldsymbol{x}) + \int_{\{\boldsymbol{x} \in \mathbb{R}^D | p_0(\boldsymbol{x}) > p_1(\boldsymbol{x})\}} p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} \tag{100}$$

$$\leq 1 + D_{\text{KL}}(p_1, p_0) \tag{101}$$

Hence we get

$$\text{MSE} \leq \frac{1}{N} \times (2 + \mathcal{D}_{\text{KL}}(p_0, p_1) + \mathcal{D}_{\text{KL}}(p_1, p_0)) \ . \tag{102}$$

We now suppose the proposal and target are distributions in an exponential family. The KL divergence between exponential distributions with parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$, is given by the Bregman divergence of the log-partition on the swapped parameters [30, Eq. 29]

$$\mathcal{D}_{\text{KL}}(p_0, p_1) = \mathcal{D}^{\text{Bregman}}_{\log Z}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0) := \log Z(\boldsymbol{\theta}_1) - \log Z(\boldsymbol{\theta}_0) - \nabla \log Z(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0) \tag{103}$$

$$\leq \frac{L}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2 \tag{104}$$

Hence

$$\text{MSE} \leq \frac{1}{N} \times (2 + L \|\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|^2) \tag{105}$$

using the L-smoothness of the log-partition function $\log Z(\boldsymbol{\theta})$.

Discussion on the assumptions for theorems 2, 3, 4, 5 For these theorems, we have supposed that the target and proposal distributions are in an exponential family with a log parition that verifies

$$M \, \mathbf{Id} \preccurlyeq \nabla^2_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) \preccurlyeq L \, \mathbf{Id} \ . \tag{106}$$

We now look at the validity of this assumption for a simple example: the univariate gaussian, which is in an exponential family. The canonical parameters are its mean and variance (μ, v). Written as an exponential family,

$$p(\boldsymbol{x}) := \exp(\langle \boldsymbol{\theta}, \boldsymbol{t}(\boldsymbol{x}) \rangle - \log Z(\boldsymbol{\theta})) \tag{107}$$

the natural parameters are $\boldsymbol{\theta} = (\mu/v, -1/(2v))$, associated with the sufficient statistics $\boldsymbol{t}(x) = (x, x^2)$ [30]. The log-partition function and its derivatives are

$$\log Z(\boldsymbol{\theta}) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2} \log(-2\theta_2) \tag{108}$$

$$\nabla \log Z(\boldsymbol{\theta}) = \mathbb{E}_{x \sim p}[t(x)] = \left(-\frac{\theta_1}{2\theta_2}, -\frac{1}{2\theta_2} + \frac{\theta_1^2}{4\theta_2^2} \right) \tag{109}$$

$$\nabla^2 \log Z(\boldsymbol{\theta}) = \text{Var}_{x \sim p}[t(x)] = \frac{1}{2\theta_2} \begin{pmatrix} -1 & \frac{\theta_1}{\theta_2} \\ \frac{\theta_1}{\theta_2} & \frac{1}{\theta_2} - \frac{1}{2}\frac{\theta_1^2}{\theta_2} \end{pmatrix} = \begin{pmatrix} v & 2\mu v \\ 2\mu v & 2v^2 - \mu^2 \end{pmatrix} \tag{110}$$

When the mean is zero, the eigenvalues of the hessian are in fact the diagonal values $(v, 2v^2)$, and they are bounded if and only if the variance v is bounded.

Proof of Theorem 6 *Constant error of annealed NCE with the arithmetic path and "oracle-trig" schedule*

We now study the estimation error produced by the arithmetic path with "oracle-trig" schedule (table 2, line 5). We write the optimal path of Eq. 59 in the limit where the distributions do not overlap: $p_0(\boldsymbol{x})p_1(\boldsymbol{x}) \to 0$ pointwise and is bounded by an integrable function. In that limit, many quantities involved in the optimal distribution simplify

$$\mathcal{D}_{H^2}(p_0, p_1) = 1 - \int \sqrt{p_0 p_1} \to 1 \tag{111}$$

$$\alpha_H = \arctan\left(\sqrt{\frac{\mathcal{D}_{H^2}(p_0, p_1)}{2 - \mathcal{D}_{H^2}(p_0, p_1)}}\right) \to \frac{\pi}{4} \tag{112}$$

$$a(t) = \frac{\cos((2t-1)\alpha_H)}{2\cos(\alpha_H)} - \frac{\sin((2t-1)\alpha_H)}{2\sin(\alpha_H)} \to \cos\left(\frac{\pi t}{2}\right) \tag{113}$$

$$b(t) = \frac{\cos((2t-1)\alpha_H)}{2\cos(\alpha_H)} + \frac{\sin((2t-1)\alpha_H)}{2\sin(\alpha_H)} \to \sin\left(\frac{\pi t}{2}\right) \ . \tag{114}$$

All these limits are pointwise: for the first line, the dominated convergence theorem is used to justify the pointwise convergence of the integral $\int p_0 p_1 \to 0$ (L2 convergence of $\sqrt{p_0 p_1}$) and consequently the pointwise convergence of the integral $\int \sqrt{p_0 p_1} \to 0$ (L1 convergence of $\sqrt{p_0 p_1}$). This leads to the following simplification of the optimal path

$$p_t(\boldsymbol{x}) = \left(a(t)\sqrt{p_0(\boldsymbol{x})} + b(t)\sqrt{p_1(\boldsymbol{x})}\right)^2 = a(t)^2 p_0(\boldsymbol{x}) + b(t)^2 p_1(\boldsymbol{x}) + 2a(t)b(t)\sqrt{p_0 p_1} \tag{115}$$

$$\to \cos^2\left(\frac{\pi t}{2}\right) p_0(\boldsymbol{x}) + \sin^2\left(\frac{\pi t}{2}\right) p_1(\boldsymbol{x}) \tag{116}$$

which is the arithmetic path with "oracle-trig" schedule defined in table 2 (line 5). The trigonometric weights evolve slowly at the end points $t = 0$ and $t = 1$. The estimation error in Eq. 63 produced by this path converges to

$$\text{MSE} = \frac{1}{N}\int_0^1 I(t)dt = \frac{1}{N} 16\alpha_H^2 \sim \frac{1}{N} 16 \frac{\pi^2}{8} = \frac{1}{N} 2\pi^2 \ . \tag{117}$$

which is a constant function of the parameter-distance.

C Useful Lemma

Lemma 1 *Chi-square divergence of between a density and a mixture We wish to upper bound the chi-square divergence between a distribution $p(\boldsymbol{x})$ and a mixture $wp(\boldsymbol{x}) + (1-w)q(\boldsymbol{x})$, where $0 < w < 1$.*

$$\mathcal{D}_{\chi^2}(p, wp + (1-w)q) = \int_{\boldsymbol{x}\in\mathbb{R}^D} \frac{p(\boldsymbol{x})^2}{wp(\boldsymbol{x}) + (1-w)q(\boldsymbol{x})} d\boldsymbol{x} - 1 \tag{118}$$

$$= \int_{\{\boldsymbol{x}\in\mathbb{R}^D | p(\boldsymbol{x}) < q(\boldsymbol{x})\}} \frac{p(\boldsymbol{x})^2}{wp(\boldsymbol{x}) + (1-w)q(\boldsymbol{x})} d\boldsymbol{x} \tag{119}$$

$$+ \int_{\{\boldsymbol{x}\in\mathbb{R}^D | p(\boldsymbol{x}) > q(\boldsymbol{x})\}} \frac{p(\boldsymbol{x})^2}{wp(\boldsymbol{x}) + (1-w)q(\boldsymbol{x})} d\boldsymbol{x} - 1 \tag{120}$$

$$\leq \int_{\{\boldsymbol{x}\in\mathbb{R}^D | p(\boldsymbol{x}) < q(\boldsymbol{x})\}} \frac{p(\boldsymbol{x})^2}{wp(\boldsymbol{x}) + (1-w)p(\boldsymbol{x})} d\boldsymbol{x} \tag{121}$$

$$+ \int_{\{\boldsymbol{x}\in\mathbb{R}^D | p(\boldsymbol{x}) > q(\boldsymbol{x})\}} \frac{p(\boldsymbol{x})^2}{wq(\boldsymbol{x}) + (1-w)q(\boldsymbol{x})} d\boldsymbol{x} - 1 \tag{122}$$

$$\leq \int_{\boldsymbol{x}\in\mathbb{R}^D} p(\boldsymbol{x}) d\boldsymbol{x} + \int_{\boldsymbol{x}\in\mathbb{R}^D} \frac{p(\boldsymbol{x})^2}{q(\boldsymbol{x})} d\boldsymbol{x} - 1 \tag{123}$$

$$= \int_{\boldsymbol{x}\in\mathbb{R}^D} \frac{p(\boldsymbol{x})^2}{q(\boldsymbol{x})} d\boldsymbol{x} = \mathcal{D}_{\chi^2}(p, q) + 1 \tag{124}$$