A unifying view of contrastive learning, importance sampling, and bridge sampling for energy-based models

Luca Martino,
University of Catania, Italy.

April 9, 2026

Abstract

In the last decades, energy-based models (EBMs) have become an important class of probabilistic models in which a component of the likelihood is intractable and therefore cannot be evaluated explicitly. Consequently, parameter estimation in EBMs is challenging for conventional inference methods. In this work, we provide a unified framework that connects noise contrastive estimation (NCE), reverse logistic regression (RLR), multiple importance sampling (MIS), and bridge sampling within the context of EBMs. We further show that these methods are equivalent under specific conditions. This unified perspective clarifies relationships among existing methods and enables the development of new estimators, with the potential to improve statistical and computational efficiency. Furthermore, this study helps elucidate the success of NCE in terms of its flexibility and robustness, while also identifying scenarios in which its performance can be further improved. Hence, rather than being a purely descriptive review, this work offers a unifying perspective and additional methodological contributions. The MATLAB code used in the numerical experiments is also made freely available to support the reproducibility of the results.

Keyword: Contrastive learning; bridge sampling; reverse logistic regression; multiple importance sampling; binary classification.

1 Introduction

Energy-based models (EBMs), denoted as $\bar{\phi}(\mathbf{y}|\boldsymbol{\theta}) = \frac{\phi(\mathbf{y}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$, provide a flexible and powerful framework for probabilistic modeling. Here, $Z(\boldsymbol{\theta})$ is an intractable partition function, and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{d_\theta}$ is the object of interest for inference [1, 2, 3, 4, 5]. Despite their flexibility and expressive capability, inference and learning in EBMs are inherently challenging due to the intractability of the normalizing constant $Z(\boldsymbol{\theta}) \in \mathbb{R}$, which is typically unknown. As a result, EBMs are often referred to as unnormalized models, since the numerator is $\phi(\mathbf{y}|\boldsymbol{\theta})$ can be evaluated pointwise, whereas $Z(\boldsymbol{\theta})$ cannot. In a Bayesian framework, such likelihood functions give rise to so-called doubly intractable posteriors [6, 7, 8, 9]. The intractability of the partition function $Z(\boldsymbol{\theta})$, especially in high-dimensional settings, severely hinders likelihood-based inference, complicating model comparison and parameter estimation.

Several strategies have been proposed to enable practical inference in these models [10, 11, 12, 13]. In this work, we focus on the contrastive learning (CL) paradigm, and in particular on noise-contrastive estimation (NCE), which recasts parameter estimation as a classification problem between observed data and artificially generated samples [14, 15, 16]. NCE builds a cost function $J(\boldsymbol{\theta}, Z)$ over the augmented parameter space $\boldsymbol{\Theta} \times \mathbb{R}$. BY minimizing $J(\boldsymbol{\theta}, Z)$, one obtains estimates of both the model parameters $\boldsymbol{\theta}_{\mathtt{tr}}$, such that $\mathbf{y}_n \sim \bar{\phi}(\mathbf{y}|\boldsymbol{\theta}_{\mathtt{tr}})$ is an observed vector, and the corresponding normalizing constant $Z_{\mathtt{tr}} = Z(\boldsymbol{\theta}_{\mathtt{tr}})$. Owing to its effectiveness and flexibility, NCE has been widely studied and applied in a variety of settings [17, 18, 19]. Recently, in [20], the authors study the NCE performance focusing mainly on the estimation in the $\boldsymbol{\theta}$-space.

In this work, unlike in [20], we mainly focus on the estimation of the normalizing constant $Z_{\mathtt{tr}} = Z(\boldsymbol{\theta}_{\mathtt{tr}})$ by NCE-type approaches. More specifically, we provide a unifying view that connects NCE, reverse logistic regression (RLR), multiple importance sampling (MIS), and bridge sampling within a common framework for EBMs. We show their equivalence under some specific conditions. Although these methods originate from different communities and are often presented from distinct perspectives, they clearly share a common underlying structure: all rely on comparing samples drawn from the model of interest $\bar{\phi}(\mathbf{y}|\boldsymbol{\theta}_{\mathtt{tr}})$, with samples generated from an auxiliary proposal/reference distribution, denoted as $q(\mathbf{y})$. In particular, contrastive learning methods frame the problem as a classification

task between data and noise, while importance sampling and bridge sampling construct estimators of normalizing constants through weighted combinations of samples from multiple distributions.

This unified view not only clarifies the relationships among existing methods, but also enables the design of new estimators that interpolate between NCE, multiple importance sampling [21, 22] and bridge sampling [23, 24], potentially offering improved statistical and computational properties (see Figure 1). Thus, we also extend the presented frameworks to encompass a broader class of importance sampling schemes that jointly exploit samples from both the data distributed as the given model and artificial data from a proposal/contrastive density. Moreover, the proposed unified formulation naturally enables the development of new estimation schemes for $\boldsymbol{\theta}$, which are also introduced and empirically evaluated. Figure 1 summarizes the main relationships studied.

Thus, in line with other works in the literature of a similar spirit [5, 25, 26], the connections established in this work offer a twofold contribution: they provide a unifying perspective on existing methods and a principled framework for designing novel estimation schemes. Furthermore, this study helps to elucidate the success of the NCE method in terms of its flexibility and robustness, while also highlighting scenarios in which its performance may be further improved. Additionally, some of the proposed schemes may admit a more tractable theoretical analysis, which in turn can simplify the characterization of the *optimal* proposal/reference density, an aspect that is not straightforward in standard NCE [27, 28]. Thus, through theoretical analysis and empirical evaluation, we demonstrate how these connections provide insight into the behavior of existing estimators and can guide the construction of more effective learning and inference procedures for EBMs. The Matlab code related to the experiments is also provided.[1]

2 Preliminaries and main notation

In this work, we mainly focus on the so-called energy-based models (EBMs). Let us define $\phi(\mathbf{y}|\boldsymbol{\theta}) \geq 0$ a function parametrized by a vector of parameters $\boldsymbol{\theta}$ taking values in $\boldsymbol{\Theta} \subseteq \mathbb{R}^{d_\theta}$, and $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$. We assume that $\phi(\mathbf{y}|\boldsymbol{\theta})$ is analytically known and we can evaluate it. An energy-based model is represented by the

[1] The code is publicly available at http://www.lucamartino.altervista.org/PUBLIC_CODE_NCE_BRIDGE.zip.

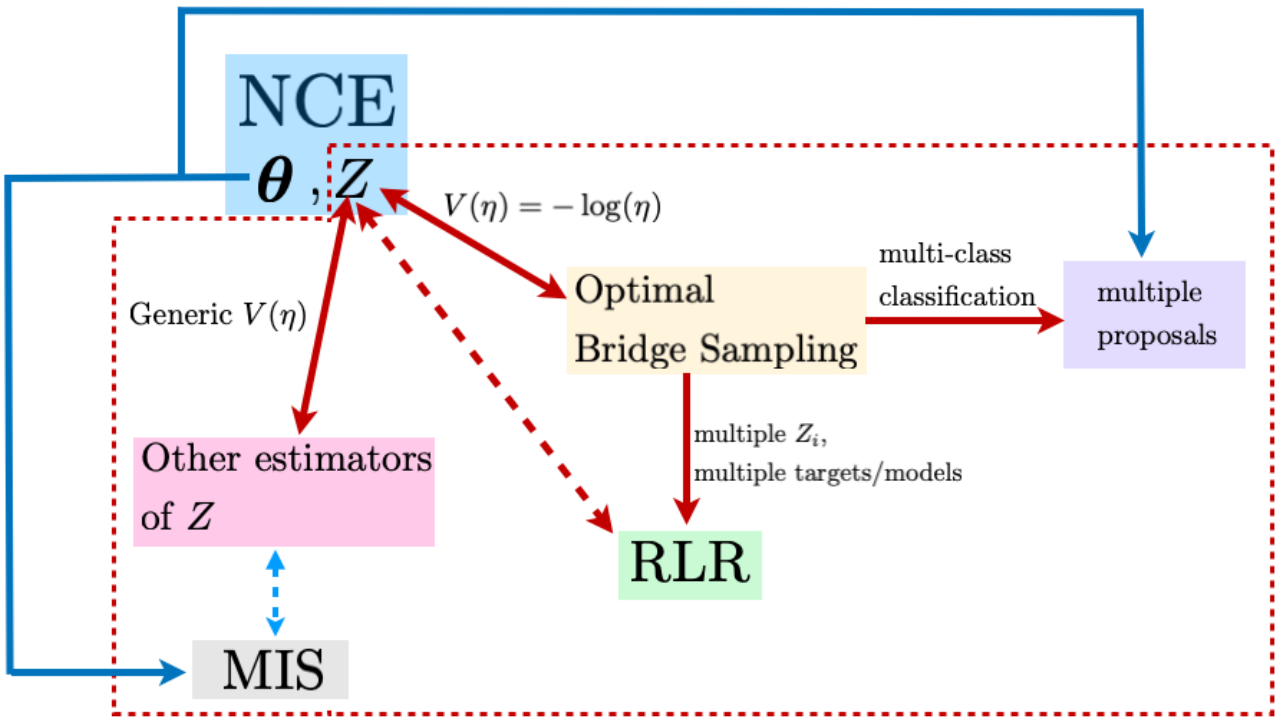

Figure 1: Graphical summary of the connections and extensions described in this work. The noise contrastive estimation (NCE) method provides estimators of $\boldsymbol{\theta}_{\mathtt{tr}}$ and $Z_{\mathtt{tr}} = Z(\boldsymbol{\theta}_{\mathtt{tr}})$ designing a binary classification problem. Setting $V(\eta) = -\log(\eta)$ as a scoring rule, we show that NCE operates as an optimal bridge estimator in the Z-domain. The reverse logistic regression (RLR) coincides with NCE in the Z-domain, and as an extension of bridge sampling, when several models/targets are considered. Several other generalizations (even for the estimation of $\boldsymbol{\theta}$) can be studied considering different scoring rules $V(\eta)$ and multiple importance sampling (MIS) procedures [21, 22] (see Section 7).

probability density function (pdf),

$$\bar{\phi}(\mathbf{y}|\boldsymbol{\theta}) = \frac{\phi(\mathbf{y}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \propto \phi(\mathbf{y}|\boldsymbol{\theta}), \tag{1}$$

parametrized by the vector $\boldsymbol{\theta}$. In many applications, the following integral cannot be evaluated analytically:

$$Z(\boldsymbol{\theta}) = \int_{\mathcal{Y}} \phi(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y}. \tag{2}$$

Namely, $Z(\boldsymbol{\theta}) : \boldsymbol{\Theta} \to \mathbb{R}^+$ is positive function that is unknown since the integral above cannot be solved analytically in closed form, i.e., is intractable.[2] Hence, the normalizing constant $Z(\boldsymbol{\theta})$, often called *partition function*, cannot be evaluated point-wise. For this reason, sometimes they are also known as *non-normalized models*. This represents a challenge for making inference on $\boldsymbol{\theta}$. Note that fixing $\boldsymbol{\theta}$, $Z(\boldsymbol{\theta})$ is a positive (unknown) normalizing constant.

Ovserved data. Let us assume that we have an observed dataset $\mathbf{y}_{1:N} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\} \in \mathcal{Y}^N$, that contains i.i.d. realizations distributed as the the EBM in Eq. (1) for a specific unknown vector of parameters $\boldsymbol{\theta}_{\mathtt{tr}}$ (true vector of parameters), i.e.,

$$\mathbf{y}_n \sim \bar{\phi}(\mathbf{y}|\boldsymbol{\theta}_{\mathtt{tr}}) = \frac{\phi(\mathbf{y}|\boldsymbol{\theta}_{\mathtt{tr}})}{Z(\boldsymbol{\theta}_{\mathtt{tr}})}, \qquad n = 1, ..., N. \tag{3}$$

Note that $Z(\boldsymbol{\theta}_{\mathtt{tr}})$ is a scalar normalizing constant, i.e., the true partition function evaluated at $\boldsymbol{\theta}_{\mathtt{tr}}$.

Goal. Given the observed data $\mathbf{y}_{1:N}$, the goal is to infer the parameter vector $\boldsymbol{\theta}_{\mathtt{tr}}$ and the scalar value $Z_{\mathtt{tr}} = Z(\boldsymbol{\theta}_{\mathtt{tr}})$ (or related to other generic $\boldsymbol{\theta}$). For this reason, in many sections, we will simplify the notation as

$$\bar{\phi}(\mathbf{y}) = \bar{\phi}(\mathbf{y}|\boldsymbol{\theta}_{\mathtt{tr}}), \quad \phi(\mathbf{y}) = \phi(\mathbf{y}|\boldsymbol{\theta}_{\mathtt{tr}}), \quad Z_{\mathtt{tr}} = Z(\boldsymbol{\theta}_{\mathtt{tr}}). \tag{4}$$

3 Noise contrastive estimation (NCE)

In this section, we present one of the most prominent methods for performing inference in EBMs, i.e., the noise-contrastive estimation (NCE). NCE is a

[2] We assume that $\mathbf{y}$ is a continuous vector, although several considerations are also valid for the discrete case.

contrastive learning (CL) approach applied in EBMs. The inference is driven by comparing samples from the observed data distribution against samples from a reference/noise distribution. More specifically, the idea in NCE is to learn $\boldsymbol{\theta}$, and a pointwise estimation of $Z(\boldsymbol{\theta})$, by designing a suitable binary classification problem. Let us define a generic input vector $\mathbf{u} \in \mathbb{R}^d$ and a binary label $a \in \{0, 1\}$, more specifically, $\mathbf{y}_n \sim p(\mathbf{u}|a = 1)$ and $\mathbf{x}_m \sim p(\mathbf{u}|a = 0)$, where $\mathbf{x}_m \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}$, i.e., each $\mathbf{y}_n$ and $\mathbf{x}_m$ live in the same space. This framework can be rewritten as

$$\mathbf{y}_n \sim \bar{\phi}(\mathbf{u}|\boldsymbol{\theta}_{\mathtt{tr}}) = \frac{\phi(\mathbf{u}, \boldsymbol{\theta}_{\mathtt{tr}})}{Z(\boldsymbol{\theta}_{\mathtt{tr}})}, \qquad n = 1, ..., N,$$

and

$$\mathbf{x}_m \sim q(\mathbf{u}), \qquad m = 1, ..., M,$$

i.e., $p(\mathbf{u}|a = 1) = \bar{\phi}(\mathbf{u}|\boldsymbol{\theta}_{\mathtt{tr}})$ and again $p(\mathbf{u}|a = 0) = q(\mathbf{u})$ is a density chosen by the user.[3] Thus, we have $M + N$ labelled inputs $\mathbf{u}_i$, i.e., $\{\mathbf{u}_i, a_i\}_{i=1}^{M+N}$, set as

$$\underbrace{\mathbf{u}_1 = \mathbf{y}_1, \ldots, \mathbf{u}_N = \mathbf{y}_N}_{a=1}, \underbrace{\mathbf{u}_{N+1} = \mathbf{x}_1, \ldots, \mathbf{u}_{N+M} = \mathbf{x}_M}_{a=0}. \tag{5}$$

Namely, the first N inputs are labelled with $a = 1$, and the rest M inputs are labelled with $a = 0$. In the CL context, the samples $\mathbf{x}_1, ..., \mathbf{x}_M$ are usually called reference/noise data and q is often referred as *reference density*. In this work, we will call it *proposal density*, to clarify the link with the importance sampling framework.

Thus, we can consider a binary classification problem with the entire dataset $\{\mathbf{u}_i, a_i\}_{i=1}^{M+N}$, formed by the union of the two sets of vectors of $\mathbf{y}$'s and $\mathbf{x}$'s. Then, we can apply a binary classifier in order to estimate the unknown variables $\boldsymbol{\theta}_{\mathtt{tr}}$ and $Z(\boldsymbol{\theta}_{\mathtt{tr}})$, comparing the two sets of data. The marginal (prior) probabilities of the labels can be approximated as $p(a = 1) \approx \alpha_1 = \frac{N}{M+N}$, $p(a = 0) \approx \alpha_2 = \frac{M}{M+N}$.

[3] We assume that q is normalized (i.e., $\int_{\mathcal{Y}} q(\mathbf{y}) d\mathbf{y} = 1$)

Setting $\nu = \frac{p(a=0)}{p(a=1)} \approx \frac{M}{N}$ and $\boldsymbol{\xi} = [\boldsymbol{\theta}, Z]$, the posterior probabilities are

$$p(a=1|\mathbf{u}) = \eta(\mathbf{u}, \boldsymbol{\xi}) = \eta(\mathbf{u}, \boldsymbol{\theta}, Z) = \frac{p(\mathbf{u}|a=1)p(a=1)}{p(\mathbf{u}|a=1)p(a=1) + p(\mathbf{u}|a=0)p(a=0)}$$

$$= \frac{\bar{\phi}(\mathbf{u}|\boldsymbol{\theta})}{\bar{\phi}(\mathbf{u}|\boldsymbol{\theta}) + \nu q(\mathbf{u})}, \tag{6}$$

$$= \frac{\phi(\mathbf{u}, \boldsymbol{\theta})}{\phi(\mathbf{u}, \boldsymbol{\theta}) + \nu Z(\boldsymbol{\theta}) q(\mathbf{u})}, \tag{7}$$

Clearly, we also have $p(a=0|\mathbf{u}) = 1 - \eta(\mathbf{u}, \boldsymbol{\theta}, Z)$. Note that η depends on the analytic form of ϕ and q and on the unknown values of $\boldsymbol{\theta}$ and $Z(\boldsymbol{\theta})$, i.e., the parameter vector $\boldsymbol{\xi} = [\boldsymbol{\theta}, Z]$. Note that here we are considering a generic vector $\boldsymbol{\theta}$ and a generic function $Z(\boldsymbol{\theta})$.

Moreover, a Bernoulli model can be considered with parameter $p(a=1|\mathbf{u}) = \eta(\mathbf{u}, \boldsymbol{\theta}, Z)$ and build a likelihood function (according to the data) exactly as in a logistic regression. Thus, the corresponding negative log-likelihood functions is:

$$\begin{cases} J_{\texttt{NCE}}(\boldsymbol{\xi}) = -\sum_{n=1}^{N} \log\left(\eta\left(\mathbf{y}_n, \boldsymbol{\theta}, Z\right)\right) - \sum_{m=1}^{M} \log\left(1 - \eta\left(\mathbf{x}_m, \boldsymbol{\theta}, Z\right)\right), \quad \text{with} \\ \eta(\mathbf{u}, \boldsymbol{\theta}, Z) = \dfrac{\bar{\phi}(\mathbf{u}|\boldsymbol{\theta})}{\bar{\phi}(\mathbf{u}|\boldsymbol{\theta}) + \nu q(\mathbf{u})}, \qquad 1 - \eta(\mathbf{u}, \boldsymbol{\theta}, Z) = \dfrac{\nu q(\mathbf{u})}{\bar{\phi}(\mathbf{u}|\boldsymbol{\theta}) + \nu q(\mathbf{u})}. \end{cases} \tag{8}$$

Recalling $\nu = \dfrac{M}{N}$, the final cost function to minimize is

$$J_{\texttt{NCE}}(\boldsymbol{\theta}, Z) = -\sum_{n=1}^{N} \log\left[\frac{\bar{\phi}(\mathbf{y}_n|\boldsymbol{\theta})}{\bar{\phi}(\mathbf{y}_n|\boldsymbol{\theta}) + \nu q(\mathbf{y}_n)}\right] - \sum_{m=1}^{M} \log\left[\frac{\nu q(\mathbf{x}_m)}{\bar{\phi}(\mathbf{x}_m|\boldsymbol{\theta}) + \nu q(\mathbf{x}_m)}\right], \tag{9}$$

$$= -\sum_{n=1}^{N} \log\left[\frac{\phi(\mathbf{y}_n, \boldsymbol{\theta})}{\phi(\mathbf{y}_n, \boldsymbol{\theta}) + \nu Z(\boldsymbol{\theta}) q(\mathbf{y}_n)}\right] - \sum_{m=1}^{M} \log\left[\frac{\nu Z(\boldsymbol{\theta}) q(\mathbf{x}_m)}{\phi(\mathbf{x}_m, \boldsymbol{\theta}) + \nu Z(\boldsymbol{\theta}) q(\mathbf{x}_m)}\right]. \tag{10}$$

We can minimize $J_{\texttt{NCE}}(\boldsymbol{\theta}, Z)$ with respect to $\boldsymbol{\theta}$ and Z, i.e.,

$$[\widehat{\boldsymbol{\theta}}_{\texttt{NCE}}, \widehat{Z}_{\texttt{NCE}}] = \arg\min J_{\texttt{NCE}}(\boldsymbol{\theta}, Z), \tag{11}$$

where $\widehat{\boldsymbol{\theta}}_{\texttt{NCE}} \longrightarrow \boldsymbol{\theta}_{\texttt{tr}}$ and

$$\widehat{Z}_{\texttt{NCE}} \longrightarrow Z_{\texttt{tr}} = Z(\boldsymbol{\theta}_{\texttt{tr}}), \tag{12}$$

is a scalar value, that is the approximation of function $Z(\boldsymbol{\theta})$ in one specific point, $\boldsymbol{\theta}_{\texttt{tr}}$. For considerations about the optimality of proposal/reference density in NCE see [27, 28].

4 From NCE to reverse logistic regression

We can rewrite Eq. (10) as

$$\begin{aligned} J_{\mathrm{NCE}}(\boldsymbol{\theta}, Z) &= -\sum_{n=1}^{N} \log \left[\frac{N\bar{\phi}\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}{N\bar{\phi}\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + Mq\left(\mathbf{y}_n\right)} \right] - \sum_{m=1}^{M} \log \left[\frac{Mq\left(\mathbf{x}_m\right)}{N\bar{\phi}\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + Mq\left(\mathbf{x}_m\right)} \right], \\ &= -\sum_{n=1}^{N} \log \left[\frac{\alpha_1\bar{\phi}\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}{\alpha_1\bar{\phi}\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + \alpha_2 q\left(\mathbf{y}_n\right)} \right] - \sum_{m=1}^{M} \log \left[\frac{\alpha_2 q\left(\mathbf{x}_m\right)}{\alpha_1\bar{\phi}\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + \alpha_2 q\left(\mathbf{x}_m\right)} \right], \end{aligned}$$

where we have multiplied numerators and denominators of the fractions (inside the log) by $\frac{1}{M+N}$, and we have also defined

$$\alpha_1 = \frac{N}{M+N} \quad \text{and} \quad \alpha_2 = \frac{M}{M+N}.$$

Note that $\alpha_1 + \alpha_2 = 1$. Furthermore, using the property $\log(ab) = \log(a) + \log(b)$, we obtain:

$$\begin{aligned} J_{\mathrm{NCE}}(\boldsymbol{\theta}, Z) &= -\sum_{n=1}^{N} \log \left[\frac{\bar{\phi}\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}{\alpha_1\bar{\phi}\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + \alpha_2 q\left(\mathbf{y}_n\right)} \right] - \sum_{m=1}^{M} \log \left[\frac{q\left(\mathbf{x}_m\right)}{\alpha_1\bar{\phi}\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + \alpha_2 q\left(\mathbf{x}_m\right)} \right] - N\log\alpha_1 - M\log\alpha_2, \\ &= -\sum_{n=1}^{N} \log \left[\frac{\bar{\phi}\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}{\alpha_1\bar{\phi}\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + \alpha_2 q\left(\mathbf{y}_n\right)} \right] - \sum_{m=1}^{M} \log \left[\frac{q\left(\mathbf{x}_m\right)}{\alpha_1\bar{\phi}\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + \alpha_2 q\left(\mathbf{x}_m\right)} \right] + C_0. \end{aligned}$$

Taking the minus-expectation of the last expression above, we finally have:

$$\begin{aligned} \exp\left(-J_{\mathrm{NCE}}(\boldsymbol{\theta}, Z)\right) &\propto \prod_{n=1}^{N} \frac{\bar{\phi}\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}{\alpha_1\bar{\phi}\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + \alpha_2 q\left(\mathbf{y}_n\right)} \prod_{m=1}^{M} \frac{q\left(\mathbf{x}_m\right)}{\alpha_1\bar{\phi}\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + \alpha_2 q\left(\mathbf{x}_m\right)}, \\ &\propto \prod_{n=1}^{N} \frac{\frac{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}{Z(\boldsymbol{\theta})}}{\alpha_1\frac{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}{Z(\boldsymbol{\theta})} + \alpha_2 q\left(\mathbf{y}_n\right)} \prod_{m=1}^{M} \frac{q\left(\mathbf{x}_m\right)}{\alpha_1\frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)}{Z(\boldsymbol{\theta})} + \alpha_2 q\left(\mathbf{x}_m\right)}. \end{aligned} \tag{13}$$

We now fix $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{tr}}$ and focus on the computation of $Z = Z(\boldsymbol{\theta}_{\text{tr}})$. The resulting (pseudo-)likelihood can be written as

$$L(Z) = L(\mathbf{y}_{1:N}, \mathbf{x}_{1:M}|Z) = \exp(-J_{\text{NCE}}(Z)) \tag{14}$$

$$\propto \prod_{n=1}^{N} \frac{\frac{\phi(\mathbf{y}_n)}{Z}}{\alpha_1 \frac{\phi(\mathbf{y}_n)}{Z} + \alpha_2 q(\mathbf{y}_n)} \prod_{m=1}^{M} \frac{q(\mathbf{x}_m)}{\alpha_1 \frac{\phi(\mathbf{x}_m)}{Z} + \alpha_2 q(\mathbf{x}_m)}. \tag{15}$$

Here, $L(Z) = L(\mathbf{y}_{1:N}, \mathbf{x}_{1:M}|Z)$ denotes a (pseudo-)likelihood function used to obtain an estimate $\widehat{Z}$ of Z by maximization. This likelihood can be obtained knowing that $\mathbf{y}_{1:N} \sim \bar{\phi}(\mathbf{y})$, $\mathbf{x}_{1:M} \sim q(\mathbf{y})$ are data generated from, respectively, a first and second component of the mixture,

$$q_{\texttt{mix}}(\mathbf{y}) = \alpha_1 \bar{\phi}(\mathbf{y}) + \alpha_2 q(\mathbf{y}) = \alpha_1 \bar{\phi}(\mathbf{y}) + \alpha_2 q(\mathbf{y}), \tag{16}$$

that is the denominator of the ratios above. This approach, equivalent to the NCE, is also called *reverse logistic regression* (RLR) [29, 30, 31]. The RLR scheme was proposed in a more generic scenario with more than one normalizing constant to estimate: let $\{\bar{\phi}_k(\mathbf{y})\}_{k=1}^{K}$ be a collection of nonnegative functions on a common space $\mathcal{Y}$, and define the corresponding normalized densities

$$\bar{\phi}_k(\mathbf{y}) = \frac{\phi_k(\mathbf{y})}{Z_k}, \quad Z_k = \int_{\mathcal{Y}} \phi_k(\mathbf{y}) d\mathbf{y}.$$

where the normalizing constants Z_k are unknown. Assuming that, we have access to different sets of samples $\mathbf{y}_{k,1}, \ldots, \mathbf{y}_{k,N_k} \sim \bar{\phi}_k(\mathbf{y})$ for each $k = 1, \ldots, K$, the objective of RLR is to estimate Z_k's values up to an additive constant. RLR models the conditional probability

$$p(a = k|\mathbf{y}) = \frac{N_k \phi_k(\mathbf{y})/Z_k}{\sum_{j=1}^{K} N_j \phi_j(\mathbf{y})/Z_j}.$$

This expression has the form of a multinomial logistic regression model, where the parameters $\{Z_k\}$ (that can expressed as $Z_k = e^{\lambda_k}$, if desired) play the role of regression coefficients. The parameters Z_k (or λ_k) are estimated by maximizing the log-likelihood $L(Z_{1:k}) = \prod_{k=1}^{K} \prod_{n=1}^{N_k} p(a = k|\mathbf{y})$. Identifiability is ensured by fixing one parameter, typically, e.g., one $Z_k = 1$ for some k.

Remark 1 *Hence, when focusing exclusively on the estimation of Z and setting $K = 2$, with $p_1(\mathbf{y}) = p(\mathbf{y})$, $p_2(\mathbf{y}) = q(\mathbf{y})$, and $Z_2 = 1$, we can conclude that the two methods, NCE and RLR, coincide.*

5 From NCE and RLR to bridge sampling

In the next sections, we fix $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{tr}}$ and use the simplified notation $\bar{\phi}(\mathbf{y}) = \bar{\phi}(\mathbf{y}|\boldsymbol{\theta}_{\texttt{tr}})$, $\phi(\mathbf{y}) = \phi(\mathbf{y}|\boldsymbol{\theta}_{\texttt{tr}})$, and $Z = Z(\boldsymbol{\theta}_{\texttt{tr}})$. We first show how the optimal bridge sampling formula can be obtained by deriving the NCE cost function (or, equivalently, the negative log-likelihood of reverse logistic regression). We then recall the standard derivation of bridge sampling.

5.1 Equivalence to optimal bridge sampling

Let consider the negative log-likelihood, $-\log L(Z) = J_{\text{NCE}}(Z)$ or Eq. (10), i.e.,

$$J_{\text{NCE}}(Z) = \sum_{n=1}^{N} \log \frac{\phi(\mathbf{y}_n)}{\alpha_1 \phi(\mathbf{y}_n) + \alpha_2 Z q(\mathbf{y}_n)} + \sum_{m=1}^{M} \log \frac{Z q(\mathbf{x}_m)}{\alpha_1 \phi(\mathbf{x}_m) + \alpha_2 Z q(\mathbf{x}_m)}. \quad (17)$$

For minimizing $J_{\text{NCE}}(Z)$, we can take the derivative with respect to Z and equaling to zero. Using the following rules and properties,

$$\frac{d \log(\frac{c}{a+bZ})}{dZ} = -\frac{b}{a+bZ}, \qquad \log\left(\frac{cZ}{a+bZ}\right) = \log(cZ) - \log(a+bZ),$$

and hence

$$\frac{d \log(\frac{cZ}{a+bZ})}{dZ} = \frac{1}{Z} - \frac{b}{a+bZ},$$

we can write:

$$\boxed{\frac{dJ_{\text{NCE}}}{dZ} = -\sum_{n=1}^{N} \frac{\alpha_2 q(\mathbf{y}_n)}{\alpha_1 \phi(\mathbf{y}_n) + \alpha_2 Z q(\mathbf{y}_n)} + \sum_{m=1}^{M} \frac{1}{Z} - \sum_{m=1}^{M} \frac{\alpha_2 q(\mathbf{x}_m)}{\alpha_1 \phi(\mathbf{x}_m) + \alpha_2 q(\mathbf{x}_m) Z} = 0.}$$

With some additional algebra, we obtain

$$\frac{dJ_{\text{NCE}}}{dZ} = -\sum_{n=1}^{N} \frac{\alpha_2 q(\mathbf{y}_n)}{\alpha_1 \phi(\mathbf{y}_n) + \alpha_2 Z q(\mathbf{y}_n)} + \sum_{m=1}^{M} \frac{\alpha_1 \phi(\mathbf{x}_m) + \cancel{\alpha_2 Z q(\mathbf{x}_m)} - \cancel{\alpha_2 Z q(\mathbf{x}_m)}}{Z\left(\alpha_1 \phi(\mathbf{x}_m) + \alpha_2 q(\mathbf{x}_m) Z\right)} = 0.$$

so finally we get

$$-\sum_{n=1}^{N}\frac{\alpha_2 q(\mathbf{y}_n)}{\alpha_1\phi(\mathbf{y}_n)+\alpha_2 Z q(\mathbf{y}_n)}+\sum_{m=1}^{M}\frac{\alpha_1\phi(\mathbf{x}_m)}{Z\left(\alpha_1\phi(\mathbf{x}_m)+\alpha_2 Z q(\mathbf{x}_m)\right)}=0, \tag{18}$$

$$\sum_{m=1}^{M}\frac{\alpha_1\phi(\mathbf{x}_m)}{\alpha_1\phi(\mathbf{x}_m)+\alpha_2 Z q(\mathbf{x}_m)}=Z\sum_{n=1}^{N}\frac{\alpha_2 q(\mathbf{y}_n)}{\alpha_1\phi(\mathbf{y}_n)+\alpha_2 Z q(\mathbf{y}_n)}. \tag{19}$$

The expression above can be rewritten as fixed-point equation:

$$Z=\frac{\alpha_1\displaystyle\sum_{m=1}^{M}\frac{\phi(\mathbf{x}_m)}{\alpha_1\phi(\mathbf{x}_m)+\alpha_2 Z q(\mathbf{x}_m)}}{\alpha_2\displaystyle\sum_{n=1}^{N}\frac{q(\mathbf{y}_n)}{\alpha_1\phi(\mathbf{y}_n)+\alpha_2 Z q(\mathbf{y}_n)}},\quad \mathbf{y}_{1:N}\sim\bar{\phi}(\mathbf{y}),\quad \mathbf{x}_{1:M}\sim q(\mathbf{y}), \tag{20}$$

where Z appears in the two sides of the equation. Recall that $\bar{\phi}(\mathbf{y})=\frac{\phi(\mathbf{y})}{Z}$ and $\frac{\alpha_1}{\alpha_2}=\frac{N}{M}$.

Remark 2 *Considering the asymptotic case, i.e., $M\to\infty$, $N\to\infty$, the expression above represents a fixed point equation, that is Eq.* (26) *below.*

Thus, assuming great values of N, M, the expression above suggests the iterative procedure (with iteration index $t\in\mathbb{N}$) for obtaining an estimator $\widehat{Z}$:

$$\boxed{\widehat{Z}_{t+1}=\frac{\frac{1}{M}\displaystyle\sum_{m=1}^{M}\frac{\phi(\mathbf{x}_m)}{\alpha_1\phi(\mathbf{x}_m)+\alpha_2\widehat{Z}_t q(\mathbf{x}_m)}}{\frac{1}{N}\displaystyle\sum_{n=1}^{N}\frac{q(\mathbf{y}_n)}{\alpha_1\phi(\mathbf{y}_n)+\alpha_2\widehat{Z}_t q(\mathbf{y}_n)}},\quad \mathbf{y}_n\sim\bar{\phi}(\mathbf{y}),\quad \mathbf{x}_m\sim q(\mathbf{y}),} \tag{21}$$

that coincides exactly with iteration procedure of the **optimal bridge sampling** [23, 24].

Remark 3 *With respect to the estimation of the normalizing constant Z (with $\theta=\theta_{\mathrm{tr}}$ fixed), the three methodologies, (a) NCE, (b) reverse logistic regression, and (c) optimal bridge sampling, coincide.*

Remark 4 *Note that, in this work, we are not assuming to be able to draw samples from the model $\bar{\phi}(\mathbf{y})$. The N samples*

$$\mathbf{y}_1, ..., \mathbf{y}_N \sim \bar{\phi}(\mathbf{y}),$$

are the observed data. Moreover, the posterior density $\bar{\phi}(\mathbf{y})$ cannot be completely evaluated because the normalizing constant Z is unknown. This difficulty is usually addressed by employing recursive procedures in most of the estimators discussed above.

The considerations in Remark 4 are also relevant for the estimators described in Section 6.

5.2 Classical derivation of bridge sampling

Let us define with $b(\mathbf{y}) > 0$ an arbitrary, positive, generic function defined on the support of $\bar{\phi}(\mathbf{y})$ i.e., $\mathcal{Y}$. Moreover, $b(\mathbf{y})$ must be such that $b(\mathbf{y})q(\mathbf{y})$ and $b(\mathbf{y})\bar{\phi}(\mathbf{y})$ are both integrable. Bridge sampling can be derived from the following identity [23, 24]:

$$\frac{\int_{\mathcal{Y}} b(\mathbf{y})\bar{\phi}(\mathbf{y})q(\mathbf{y})d\mathbf{y}}{\int_{\mathcal{Y}} b(\mathbf{y})\bar{\phi}(\mathbf{y})q(\mathbf{y})d\mathbf{y}} = 1, \tag{22}$$

that is true since numerator and denominator are the exactly the same integral. This integral can be expressed as expectation with respect to q, i.e., $\mathbb{E}_q[\alpha(\mathbf{y})\bar{\phi}(\mathbf{y})]$, or as expectation with respect to $\bar{\phi}$, i.e. $\mathbb{E}_{\bar{\phi}}[\alpha(\mathbf{y})q(\mathbf{y})]$, hence

$$\frac{\mathbb{E}_q[b(\mathbf{y})\bar{\phi}(\mathbf{y})]}{\mathbb{E}_{\bar{\phi}}[b(\mathbf{y})q(\mathbf{y})]} = \frac{\frac{1}{Z}\mathbb{E}_q[b(\mathbf{y})\phi(\mathbf{y})]}{\mathbb{E}_{\bar{\phi}}[b(\mathbf{y})q(\mathbf{y})]} = 1. \tag{23}$$

Then, we arrive to the main bridge sampling identity:

$$\boxed{\frac{\mathbb{E}_q[b(\mathbf{y})\phi(\mathbf{y})]}{\mathbb{E}_{\bar{\phi}}[b(\mathbf{y})q(\mathbf{y})]} = Z} \tag{24}$$

It is possible to show that the choice

$$b(\mathbf{y}) = \frac{1}{\alpha_1\bar{\phi}(\mathbf{y}) + \alpha_2 q(\mathbf{y})} = \frac{1}{\alpha_1\frac{1}{Z}\phi(\mathbf{y}) + \alpha_2 q(\mathbf{y})}, \tag{25}$$

is optimal [23, 24]. It yields the optimal bridge sampling scheme,

$$\frac{\mathbb{E}_q\left[\frac{\phi(\mathbf{y})}{\alpha_1\bar{\phi}(\mathbf{y})+\alpha_2 q(\mathbf{y})}\right]}{\mathbb{E}_{\bar{\phi}}\left[\frac{q(\mathbf{y})}{\alpha_1\bar{\phi}(\mathbf{y})+\alpha_2 q(\mathbf{y})}\right]} = Z, \tag{26}$$

by replacing the expectations above with empirical estimators as in Eq. (20).

6 Related importance sampling (IS) estimators

6.1 Samples from two densities

In this section, we introduce other schemes for estimating of $Z = Z(\boldsymbol{\theta}_{\mathtt{tr}})$ where $\bar{q}(\mathbf{y})$ *and* $\bar{\phi}(\mathbf{y})$ are employed separately or jointly. We begin by describing estimators that leverage both densities jointly. In this setting, the model $\bar{\phi}(\mathbf{y})$ is also used as a proposal distribution. Note that drawing N samples from $\bar{\phi}(\mathbf{y})$ and M samples from $q(\mathbf{y})$ is equivalent to sampling by a *deterministic* approach from the mixture [21, 22],

$$\begin{aligned} q_{\mathtt{mix}}(\mathbf{y}) &= \alpha_1\bar{\phi}(\mathbf{y}) + \alpha_2 q(\mathbf{y}), \\ &= \alpha_1\frac{1}{Z}\phi(\mathbf{y}) + \alpha_2 q(\mathbf{y}), \end{aligned}$$

i.e., a single density defined as mixture of the two densities [22, 24]. The first estimator is based on the following classical equality:

$$Z = \int \phi(\mathbf{y})d\mathbf{y} = \mathbb{E}_{q_{\mathtt{mix}}}\left[\frac{\phi(\mathbf{y})}{q_{\mathtt{mix}}(\mathbf{y})}\right] = \int \frac{\phi(\mathbf{y})}{q_{\mathtt{mix}}(\mathbf{y})} q_{\mathtt{mix}}(\mathbf{y})d\mathbf{y}. \tag{27}$$

Hence, applying a deterministic mixture sampling approach from $q_{\mathtt{mix}}(\mathbf{y})$,

$$\mathbf{y}_1, ..., \mathbf{y}_N \sim \bar{\phi}(\mathbf{y}), \quad \mathbf{x}_1, ..., \mathbf{x}_M \sim q(\mathbf{y}), \tag{28}$$

and denoting

$$\mathbf{u}_1 = \mathbf{y}_1, \ldots, \mathbf{u}_N = \mathbf{y}_N, \quad \mathbf{u}_{N+1} = \mathbf{x}_1, \ldots, \mathbf{u}_{N+M} = \mathbf{x}_M, \tag{29}$$

we can consider $\mathbf{u}_i \sim q_{\texttt{mix}}(\mathbf{u}_i)$ [21, 22]. we have the IS estimator

$$\widehat{Z} = \frac{1}{N+M} \sum_{i=1}^{N+M} \frac{\phi(\mathbf{u}_i)}{q_{\texttt{mix}}(\mathbf{u}_i)}, \tag{30}$$

that can be rewritten expressed with a recursive procedure as in the bride sampling:

$$\boxed{\widehat{Z}_{t+1} = \frac{1}{N+M} \sum_{i=1}^{N+M} \frac{\widehat{Z}_t \phi(\mathbf{u}_i)}{\alpha_1 \phi(\mathbf{u}_i) + \alpha_2 \widehat{Z}_t q(\mathbf{u}_i)}, \qquad \{\mathbf{u}_i\} = \{\mathbf{y}_n\} \cup \{\mathbf{x}_m\}\,.} \tag{31}$$

We call it as **MIS** estimator. Note that this estimator can be rewritten as

$$\widehat{Z}_{t+1} = \frac{1}{N} \sum_{n=1}^{N} \frac{\widehat{Z}_t \phi(\mathbf{y}_n)}{\alpha_1 \phi(\mathbf{y}_n) + \alpha_2 \widehat{Z}_t q(\mathbf{y}_n)} + \frac{1}{M} \sum_{m=1}^{M} \frac{\widehat{Z}_t \phi(\mathbf{x}_m)}{\alpha_1 \phi(\mathbf{x}_m) + \alpha_2 \widehat{Z}_t q(\mathbf{x}_m)}, \tag{32}$$

where the expression separates into two components, one involving $\mathbf{y}_n$ and the other $\mathbf{x}_m$, similarly to the bridge sampling estimator. However, in the bridge sampling, the estimator is given by the ratio of two sums. It is possible here to construct an alternative estimator that more closely mirrors that structure. Indeed, *estimating* also the constant value $N + M$ in Eq. (31), i.e.,

$$N + M \approx \sum_{i=1}^{N+M} \frac{q(\mathbf{u})}{q_{\texttt{mix}}(\mathbf{u})}, \tag{33}$$

since we assume that q is normalized (i.e., $\int_{\mathcal{Y}} q(\mathbf{y}) d\mathbf{y} = 1$), by using the previous IS arguments,

$$\frac{1}{N+M} \sum_{i=1}^{N+M} \frac{q(\mathbf{u})}{q_{\texttt{mix}}(\mathbf{u})} \approx 1.$$

Replacing (33) into Eq. (31),

$$\widehat{Z}_{t+1} = \frac{1}{\sum_{k=1}^{N+M} \frac{q(\mathbf{u}_k)}{q_{\texttt{mix}}(\mathbf{u}_k)}} \sum_{i=1}^{N+M} \frac{\phi(\mathbf{u}_i)}{q_{\texttt{mix}}(\mathbf{u}_i)}, \tag{34}$$

$$= \frac{\sum_{i=1}^{N+M} \frac{\phi(\mathbf{u}_i)}{q_{\texttt{mix}}(\mathbf{u}_i)}}{\sum_{k=1}^{N+M} \frac{q(\mathbf{u}_k)}{q_{\texttt{mix}}(\mathbf{u}_k)}}, \tag{35}$$

and replacing inside the expression of the mixture $q_{\texttt{mix}}(\mathbf{y}) = \alpha_1\bar{\phi}(\mathbf{y}) + \alpha_2 q(\mathbf{y})$, we obtain the iterative procedure:[4]

$$\boxed{\widehat{Z}_{t+1} = \frac{\sum_{i=1}^{N+M} \dfrac{\phi(\mathbf{u}_i)}{\alpha_1\phi(\mathbf{u}_i) + \alpha_2\widehat{Z}_t q(\mathbf{u}_i)}}{\sum_{k=1}^{N+M} \dfrac{q(\mathbf{u}_k)}{\alpha_1\phi(\mathbf{u}_k) + \alpha_2\widehat{Z}_t q(\mathbf{u}_k)}}, \qquad \{\mathbf{u}_i\} = \{\mathbf{y}_n\} \cup \{\mathbf{x}_m\}\,.} \tag{36}$$

The expression above is very similar to Eq. (21) with the difference that both summations consider all the data $\{\mathbf{u}_i\}_{i=1}^{N+M}$ in Eq, (29), instead of just $\mathbf{y}_n$ or $\mathbf{x}_m$ in Eq. (28). We name this estimator as **Self-IS-with-mix**.

Remark 5 *In [23], the authors assert that both estimators in Eqs.* (31)(36) *converge to the solution given by optimal bridge sampling estimator, expressed as* (21)*. As demonstrated in the simulation study in Section 8, however, the convergence rates of the corresponding iterative methods differ depending also on the starting point.*

Remark 6 *Within the EBM framework, the observed data* $\{\mathbf{y}_n\}_{n=1}^{N}$ *are assumed to be generated directly by the model itself; consequently, the issue of sampling from a posterior distribution in Bayesian inference, that is central in standard bridge sampling applications, does not arise here, i.e., in the frequentist inference for EBMs.*

6.2 Samples from one density or combinations of estimators

Considering only $q(\mathbf{y})$ or only $\bar{\phi}(\mathbf{y})$, we have the standard IS estimator and the reverse IS estimator, respectively [24, 32]. The first one is derived from the following equality,

$$E_q\left[\frac{\phi(\mathbf{y})}{q(\mathbf{y})}\right] = \int_{\mathcal{Y}} \frac{\phi(\mathbf{y})}{q(\mathbf{y})} q(\mathbf{y}) d\mathbf{y} = \int_{\mathcal{Y}} \phi(\mathbf{y}) d\mathbf{y} = Z. \tag{37}$$

$$\tag{38}$$

[4]Note that the two $\widehat{Z}_t$ terms that should appear in the numerators cancel each other out, as in the bridge sampling expression.

and the **standard IS estimator (Stand-IS)** has the form:

$$\boxed{\widehat{Z} = \frac{1}{M}\sum_{m=1}^{M}\frac{\phi(\mathbf{x}_m)}{q(\mathbf{x}_m)}, \qquad \mathbf{x}_m \sim q(\mathbf{y}).} \tag{39}$$

The reverse IS estimator is based on the following equality,

$$\begin{aligned} E_{\bar{\phi}}\left[\frac{q(\mathbf{y})}{\bar{\phi}(\mathbf{y})}\right] = \int_{\mathcal{Y}} \frac{q(\mathbf{y})}{\bar{\phi}(\mathbf{y})}\bar{\phi}(\mathbf{y})d\mathbf{y} = 1, \\ Z\int_{\mathcal{Y}} \frac{q(\mathbf{y})}{\phi(\mathbf{y})}\bar{\phi}(\mathbf{y})d\mathbf{y} = 1, \\ Z\mathbb{E}_{\bar{\phi}}\left[\frac{q(\mathbf{y})}{\phi(\mathbf{y})}\right] = 1, \\ \mathbb{E}_{\bar{\phi}}\left[\frac{q(\mathbf{y})}{\phi(\mathbf{y})}\right] = \frac{1}{Z}. \end{aligned}$$

where we have used the fact that $q(\mathbf{y})$ is normalized, i.e., $\int_{\mathcal{Y}} q(\mathbf{y})d\mathbf{y} = 1$. Therefore, the **reverse IS (RIS) estimator** has the form:

$$\boxed{\widehat{Z} = \left(\frac{1}{N}\sum_{n=1}^{N}\frac{q(\mathbf{y}_n)}{\phi(\mathbf{y}_n)}\right)^{-1}, \quad \mathbf{y}_n \sim \bar{\phi}(\mathbf{y}) = \frac{1}{Z}\phi(\mathbf{y})}, \tag{40}$$

Note that the quantity $\widehat{A} = \frac{1}{N}\sum_{n=1}^{N}\frac{q(\mathbf{y}_n)}{\phi(\mathbf{y}_n)}$ is an unbiased estimate of $1/Z$, i.e., $\mathbb{E}[\widehat{A}] = 1/Z$. However, by Jensen's inequality, we have $\mathbb{E}\left[\frac{1}{\widehat{A}}\right] \geq \frac{1}{\mathbb{E}[\widehat{A}]} = Z$. Hence, the RIS estimator is positively biased, i.e., overestimates Z.
Both estimators above do not require recursion. Finally, another related estimator is the so-called optimal umbrella estimator [33, 30, 24]. In this case, we draw samples from a single density

$$\bar{r}(\mathbf{y}) \propto r(\mathbf{y}) = \left|\bar{\phi}(\mathbf{y}) - q(\mathbf{y})\right|, \tag{41}$$

$$= \frac{1}{c}\left|\frac{1}{Z}\phi(\mathbf{y}) - q(\mathbf{y})\right|, \tag{42}$$

where $c = \int_{\mathcal{Y}} \left| \frac{1}{Z}\phi(\mathbf{y}) - q(\mathbf{y}) \right| d\mathbf{y}$ is generally unknown and intractable. Hence, drawing $\widetilde{\mathbf{x}}_1, ..., \widetilde{\mathbf{x}}_{M+N}$ samples from $\bar{d}(\mathbf{y})$, we have

$$Z = \frac{c}{(M+N)} \sum_{i=1}^{M+N} \frac{\phi(\widetilde{\mathbf{x}}_i)}{|\frac{1}{Z}\phi(\widetilde{\mathbf{x}}_i) - q(\widetilde{\mathbf{x}}_i)|}, \tag{43}$$

and

$$1 = \frac{c}{(M+N)} \sum_{i=1}^{M+N} \frac{q(\widetilde{\mathbf{x}}_i)}{|\frac{1}{Z}\phi(\widetilde{\mathbf{x}}_i) - q(\widetilde{\mathbf{x}}_i)|}, \tag{44}$$

$$\frac{1}{c} = \frac{1}{M+N} \sum_{i=1}^{M+N} \frac{q(\widetilde{\mathbf{x}}_i)}{|\frac{1}{Z}\phi(\widetilde{\mathbf{x}}_i) - q(\widetilde{\mathbf{x}}_i)|}, \tag{45}$$

$$c = \left(\frac{1}{M+N} \sum_{i=1}^{M+N} \frac{q(\widetilde{\mathbf{x}}_i)}{|\frac{1}{Z}\phi(\widetilde{\mathbf{x}}_i) - q(\widetilde{\mathbf{x}}_i)|} \right)^{-1} \tag{46}$$

where we have used again that $q(\mathbf{y})$ is normalized, i.e., $\int_{\mathcal{Y}} q(\mathbf{y}) d\mathbf{y} = 1$. Replacing the expression of c in Eq. (46) into (43), we obtain (after some simple algebra) the final fixed point and consequently recursive equation,

$$\boxed{\widehat{Z}_{t+1} = \frac{\sum_{i=1}^{M+N} \frac{\phi(\widetilde{\mathbf{x}}_i)}{|\phi(\widetilde{\mathbf{x}}_i) - \widehat{Z}_t q(\widetilde{\mathbf{x}}_i)|}}{\sum_{k=1}^{M+N} \frac{q(\widetilde{\mathbf{x}}_k)}{|\phi(\widetilde{\mathbf{x}}_k) - \widehat{Z}_t q(\widetilde{\mathbf{x}}_k)|}}, \qquad \widetilde{\mathbf{x}}_i \sim \bar{r}(\mathbf{y}).} \tag{47}$$

that is the the optimal umbrella sampling estimator (**Opt-Umb**) [33, 30, 24]. However, we need another Monte Carlo method to draw samples from $\bar{r}(\mathbf{y}) \propto \left| \bar{\phi}(\mathbf{y}) - q(\mathbf{y}) \right|$ (it is not a straightforward task). See Table 1 for a summary of the described estimators.

7 Novel possible schemes and estimators

7.1 MIS arguments in NCE

Building on observations from prior works [21, 22], one can argue that treating $\mathbf{u}i = \mathbf{y}_n \cup \mathbf{x}_m$ jointly as samples drawn from the mixture distribution $q\texttt{mix}(\mathbf{u})$ may

lead to improved performance. Thus, one could design a cost function of type:

$$J_{\texttt{MIS}}(\boldsymbol{\theta}, Z) = \\ -\sum_{k=1}^{M+N} \log \frac{\phi(\mathbf{u}_k|\boldsymbol{\theta})}{\alpha_1\phi(\mathbf{u}_k|\boldsymbol{\theta}) + \alpha_2 Z q(\mathbf{u}_k)} - \sum_{k=1}^{M+N} \log \frac{Z q(\mathbf{u}_k)}{\alpha_1\phi(\mathbf{u}_k|\boldsymbol{\theta}) + \alpha_2 Z q(\mathbf{u}_k)}. \tag{48}$$

Remark 7 *Fixing $\boldsymbol{\theta}$, differentiating the above expression with respect to Z and setting the result equal to zero yields the self-IS-with-mixture estimator given in Eq.* (36)*.*

Remark 8 *Given the results on prior MIS works (e.g., [22]), we could expect that $J_{\texttt{MIS}}(\boldsymbol{\theta}, Z)$ and Eq.* (36) *provide better results in the estimation of Z. For the other side, in terms of binary classification, $J_{\texttt{MIS}}(\boldsymbol{\theta}, Z)$ is expected to perform worse than $J_{\texttt{NCE}}(\boldsymbol{\theta}, Z)$, at least for estimating $\boldsymbol{\theta}$. Indeed, $J_{\texttt{NCE}}(Z)$ leverages class label information, whereas $J_{\texttt{MIS}}(\boldsymbol{\theta}, Z)$ does not. The numerical simulations in Section 8 partially support this intuition: the performance minimizing $J_{\texttt{MIS}}$ in the $\boldsymbol{\theta}$-space depends strongly on the choice of the proposal parameters. While, under certain ideal conditions, minimizing $J_{\texttt{MIS}}$ in the Z-space provides the best performance.*

7.2 Deriving other estimators of Z from binary classifiers

We can consider other loss in the binary classification problem described in Section 3. Let us consider a positive, decreasing, concave function V defined in [0,1], that is also a *strictly proper scoring rule* [34]. The NCE procedure described above is also valid considering the cost function:

$$J\big(\boldsymbol{\theta}, Z\big) = \sum_{n=1}^{N} V\left(\eta\left(\mathbf{y}_n, \boldsymbol{\theta}, Z\right)\right) + \sum_{m=1}^{M} V\left(1 - \eta\left(\mathbf{x}_m, \boldsymbol{\theta}, Z\right)\right), \tag{49}$$

that can be minimize with respect to $\boldsymbol{\xi} = [\boldsymbol{\theta}, Z]$ for obtaining an estimators of $\boldsymbol{\theta}_{\texttt{tr}}$ and $Z(\boldsymbol{\theta}_{\texttt{tr}})$, since this is a solution of a binary classification problem. Repeating the procedure done in Section 5.1, we can derive the cost function above $J\big(\boldsymbol{\theta}, Z\big)$ with respect to Z,

$$\boxed{\frac{\partial J}{\partial Z} = \sum_{n=1}^{N} \frac{dV}{d\eta}\dot{\eta}(\mathbf{y}_n, \boldsymbol{\theta}, Z) - \sum_{m=1}^{M} \left.\frac{dV}{d\eta}\right|_{1-\eta} \dot{\eta}(\mathbf{x}_m, \boldsymbol{\theta}, Z),} \tag{50}$$

where we have denoted $\dot{\eta} = \frac{d\eta}{dZ}$ we have used $\frac{dV(1-\eta)}{d\eta} = -\left.\frac{dV(\eta)}{d\eta}\right|_{1-\eta}$. Recalling

$$\eta\left(\mathbf{u}, \boldsymbol{\theta}, Z\right) = \frac{\bar{\phi}\left(\mathbf{u}|\boldsymbol{\theta}\right)}{\bar{\phi}\left(\cdot|\boldsymbol{\theta}\right) + \nu q\left(\mathbf{u}\right)} = \frac{\phi\left(\mathbf{u}|\boldsymbol{\theta}\right)}{\phi\left(\mathbf{u}|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{u}\right)}, \tag{51}$$

hence we can write

$$\boxed{\dot{\eta}(\mathbf{u}, \boldsymbol{\theta}, Z) = -\frac{\nu\phi\left(\mathbf{u}|\boldsymbol{\theta}\right) q\left(\mathbf{u}\right)}{\left(\phi\left(\mathbf{u}|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{u}\right)\right)^2},} \tag{52}$$

$$\boxed{\dot{\eta}(\mathbf{u}, \boldsymbol{\theta}, Z) = -\eta(\mathbf{u}, \boldsymbol{\theta}, Z)(1 - \eta(\mathbf{u}, \boldsymbol{\theta}, Z)).} \tag{53}$$

Fixing $\boldsymbol{\theta}$, one could derive other estimators and/other iterative procedures.

Remark 9 *These derivations are valuable for developing alternative estimators of normalizing constants. Furthermore, the resulting estimator (or its associated iterative procedure) can be naturally integrated into the NCE framework, for instance through an alternating optimization scheme.*

7.2.1 Example 1 with a proper scoring rule

Let us consider a proper scoring rule, $V(\eta) = (1-\eta)^2$. In this case, we have

$$\frac{dV(\eta)}{d\eta} = -2(1-\eta) = -\frac{2\nu Z q\left(\mathbf{u}\right)}{\phi\left(\mathbf{u}|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{u}\right)}, \tag{54}$$

$$\frac{dV(1-\eta)}{d\eta} = -\left.\frac{dV(\eta)}{d\eta}\right|_{1-\eta} = 2\eta = \frac{2\phi\left(\mathbf{u}|\boldsymbol{\theta}\right)}{\phi\left(\mathbf{u}|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{u}\right)}. \tag{55}$$

where we have also used the definition η recalled in Eq. (51). Replacing (54)-(55) and (52) into Eq. (50), we obtain:

$$2\nu^2 Z^2 \sum_{n=1}^{N} \frac{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right) q\left(\mathbf{y}_n\right)^2}{\left(\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{y}_n\right)\right)^3} - 2\nu Z \sum_{m=1}^{M} \frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)^2 q\left(\mathbf{y}_n\right)}{\left(\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{x}_m\right)\right)^3} = 0$$

$$\nu Z \sum_{n=1}^{N} \frac{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right) q\left(\mathbf{y}_n\right)^2}{\left(\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{y}_n\right)\right)^3} - \sum_{m=1}^{M} \frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)^2 q\left(\mathbf{x}_m\right)}{\left(\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{x}_m\right)\right)^3} = 0.$$

Isolating the first Z in one side, we find a fixed point equation over Z and can write the final iterative procedure:

$$\boxed{\widehat{Z}_{t+1} = \frac{\frac{1}{M}\sum_{m=1}^{M} \frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)^2 q\left(\mathbf{x}_m\right)}{\left(\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + \nu\widehat{Z}_t q\left(\mathbf{x}_m\right)\right)^3}}{\frac{1}{N}\sum_{n=1}^{N} \frac{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right) q\left(\mathbf{y}_n\right)^2}{\left(\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + \nu\widehat{Z}_t q\left(\mathbf{y}_n\right)\right)^3}}.} \tag{56}$$

we could also obtain the estimator above from Eq. (23), setting as bridge function:

$$b(\mathbf{y}) = \frac{\phi\left(\mathbf{y}|\boldsymbol{\theta}\right) q\left(\mathbf{y}\right)}{\left(\phi\left(\mathbf{y}|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{y}\right)\right)^3}. \tag{57}$$

Remark 10 *From this result, we could speculate that there is a correspondence between proper scoring rules $V(\eta)$ and bridge functions $b(\mathbf{y})$ in Eq.* (23)*.*

7.2.2 Example 2 with a non-proper scoring rule

Let us consider now a non-proper scoring rule. In this scenario, we could obtain highly-biased estimators that require some corrections. For instance, let assume $V(\eta) = 1/\eta$. Hence, we have

$$\frac{dV(\eta)}{d\eta} = -\frac{1}{\eta^2} = -\frac{\left[\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{y}_n\right)\right]^2}{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)^2}, \tag{58}$$

$$\frac{dV(1-\eta)}{d\eta} = -\left.\frac{dV(\eta)}{d\eta}\right|_{1-\eta} = \frac{1}{(1-\eta)^2} = \frac{a\left[\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{x}_m\right)\right]^2}{\left[\nu Z q\left(\mathbf{x}_m\right)\right]^2}, \tag{59}$$

where we have substituted the definition of η in Eq. (51). Replacing (58)-(59) and (52) into Eq. (50), we obtain

$$\sum_{n=1}^{N}\left[-\frac{\left(\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{y}_n\right)\right)^2}{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)^2}\right]\left[-\frac{\nu\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right) q\left(\mathbf{y}_n\right)}{\left(\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{y}_n\right)\right)^2}\right] +$$
$$\sum_{m=1}^{M}\left[\frac{\left(\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{x}_m\right)\right)^2}{\left[\nu Z q\left(\mathbf{x}_m\right)\right]^2}\right]\left[-\frac{\nu\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right) q\left(\mathbf{x}_m\right)}{\left(\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right) + \nu Z q\left(\mathbf{x}_m\right)\right)^2}\right] = 0,$$

so that

$$\nu \sum_{n=1}^{N} \frac{q\left(\mathbf{y}_n\right)}{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)} - \frac{1}{\nu Z^2} \sum_{m=1}^{M} \frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)}{q\left(\mathbf{x}_m\right)} = 0,$$
$$\nu^2 Z^2 \sum_{n=1}^{N} \frac{q\left(\mathbf{y}_n\right)}{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)} - \sum_{m=1}^{M} \frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)}{q\left(\mathbf{x}_m\right)} = 0,$$

and isolating Z^2 in one side, we get

$$Z^2 = \frac{1}{\nu^2} \frac{\sum_{m=1}^{M} \frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)}{q\left(\mathbf{x}_m\right)}}{\sum_{n=1}^{N} \frac{q\left(\mathbf{y}_n\right)}{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}} = \frac{N}{M} \frac{\frac{1}{M} \sum_{m=1}^{M} \frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)}{q\left(\mathbf{x}_m\right)}}{\frac{1}{N} \sum_{n=1}^{N} \frac{q\left(\mathbf{y}_n\right)}{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}}.$$

Finally, we obtain a "bad" estimator

$$\widehat{Z}_{\mathtt{bad}} = \sqrt{\frac{N}{M}} \sqrt{\frac{\frac{1}{M} \sum_{m=1}^{M} \frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)}{q\left(\mathbf{x}_m\right)}}{\frac{1}{N} \sum_{n=1}^{N} \frac{q\left(\mathbf{y}_n\right)}{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}}}, \tag{60}$$

that can have highly biased with $M \neq N$, for finite values M and N. Indeed, note that the numerator is the stand-IS estimator and the denominator is the RIS estimator, i.e.,

$$\frac{1}{M} \sum_{m=1}^{M} \frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)}{q\left(\mathbf{x}_m\right)} \approx Z, \qquad \left(\frac{1}{N} \sum_{n=1}^{N} \frac{q\left(\mathbf{y}_n\right)}{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)} \right)^{-1} \approx Z,$$

so that $\widehat{Z}_{\mathtt{bad}} \approx \sqrt{\frac{N}{M}} Z$. Therefore, we can easily improve this estimator defining a scaled version, i.e.,

$$\boxed{\widehat{Z}_{\mathtt{geo}} = \sqrt{\frac{M}{N}} \widehat{Z}_{\mathtt{bad}} = \sqrt{\frac{\frac{1}{M} \sum_{m=1}^{M} \frac{\phi\left(\mathbf{x}_m|\boldsymbol{\theta}\right)}{q\left(\mathbf{x}_m\right)}}{\frac{1}{N} \sum_{n=1}^{N} \frac{q\left(\mathbf{y}_n\right)}{\phi\left(\mathbf{y}_n|\boldsymbol{\theta}\right)}}},} \tag{61}$$

that also represents the *geometric mean* between the stand-IS and the RIS estimators. Table 1 summarizes the main described estimators.

7.3 Multiple proposal densities in bridge sampling

All the previous considerations and connections highlighted above allow us to extend the optimal bridge sampling using multiple proposal densities. Let us consider K proposal densities $\{q_k(\mathbf{y})\}_{k=1}^K$, and we draw M_k samples from each of them, i.e.,

$$\mathbf{x}_{k,1}, ..., \mathbf{x}_{k,M_k} \sim q_k(\mathbf{y}).$$

We also recall that we have N observed data from the model, i.e., $\mathbf{y}_1, ..., \mathbf{y}_N \sim \bar{\phi}(\mathbf{y}_n|\boldsymbol{\theta})$. Thus, similarly as in Section 3, we can design a classification problem *with* $K+1$ *classes*, with cost function:

$$\begin{aligned} J(\boldsymbol{\theta}, Z) = &- \sum_{n=1}^{N} \log \left[\frac{N\bar{\phi}(\mathbf{y}_n|\boldsymbol{\theta})}{N\bar{\phi}(\mathbf{y}_n|\boldsymbol{\theta}) + \sum_{j=1}^{K} M_j q_j(\mathbf{y}_n)} \right] + \\ &- \sum_{k=1}^{K} \sum_{m=1}^{M_k} \log \left[\frac{M_k q_k(\mathbf{x}_{k,m})}{N\bar{\phi}(\mathbf{x}_{k,m}|\boldsymbol{\theta}) + \sum_{j=1}^{K} M_j q_j(\mathbf{x}_{k,m})} \right], \end{aligned} \tag{62}$$

Deriving the expression above with respect to Z as in Section 5.1, we obtain:

$$\boxed{\widehat{Z}_{t+1} = \frac{\sum_{k=1}^{K} \frac{1}{M_k} \sum_{m=1}^{M_k} \frac{N\phi(\mathbf{x}_{k,m}|\boldsymbol{\theta})}{N\phi(\mathbf{x}_{k,m}|\boldsymbol{\theta}) + \widehat{Z}_t \sum_{j=1}^{K} M_j q_j(\mathbf{x}_{k,m})}}{\frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{M_k q_k(\mathbf{y}_n)}{N\phi(\mathbf{y}_n|\boldsymbol{\theta}) + \widehat{Z}_t \sum_{j=1}^{K} M_j q_j(\mathbf{y}_n)}},} \tag{63}$$

This iterative procedure could be easily integrated into the NCE optimization through an alternating optimization scheme with respect to $\boldsymbol{\theta}$ and Z. The use of multiple proposal densities is particularly interesting for designing adaptive schemes, as suggested in [35, 36]. Furthermore, the use of different proposal densities can be combined with the idea of including tempered models in bridge sampling to help the exploration of the state-space. However, in this case we have one than more unknown normalizing constants to be estimated as in RLR.

8 Numerical Simulations

In this section, we provide some numerical results comparing different estimators of Z and $\boldsymbol{\theta}$. We assume finite values of N and M, instead of asymptotical

performance as in other studies [17]. The purpose of this section is not to show performance on a complex model, but rather to illustrate the behavior of the estimators computing the mean square error (MSE), under controlled scenarios, helping the reproducibility as well.[5] For this reason, we consider a univariate Gaussian target distribution as model,

$$\bar{\phi}\left(y|\theta\right) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left(-\frac{y^2}{2\theta^2}\right), \quad \text{hence} \quad \phi(y|\theta) = \exp\left(-\frac{y^2}{2\theta^2}\right), \tag{64}$$

$$\text{and} \qquad Z(\theta) = \sqrt{2\pi\theta^2}, \tag{65}$$

so that we also know the ground-truth $Z(\theta) = \sqrt{2\pi\theta^2}$. Thus, given $\theta_{\text{tr}} = 1$,we also observe the data $y_1, \ldots, y_N$ are generated from the model above, i.e.,

$$y_n \sim \bar{\phi}(y) = \bar{\phi}\left(y|\theta_{\text{tr}}\right) = \frac{1}{Z\left(\theta_{\text{tr}}\right)} \exp\left(-\frac{y^2}{2\theta_{\text{tr}}^2}\right), \quad Z_{\text{tr}} = Z\left(\theta_{\text{tr}}\right) = \sqrt{2\pi\theta_{\text{tr}}^2},$$

with $n = 1, \ldots, N$. We also consider a Gaussian proposal/reference density,

$$q(y) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{\left(y-\mu_p\right)^2}{2\sigma_p^2}\right), \tag{66}$$

where we set $\mu_p = 0$ and vary the value of σ_p.

8.1 Estimation of the normalizing constant $Z\left(\theta_{\text{tr}}\right)$

Given $\{y_n\}_{n=1}^N$, the goal is to estimate $Z_{\text{tr}} = Z\left(\theta_{\text{tr}}\right)$ employing three estimators that use sets of samples from both densities, $x_m \sim q(y)$ and $y_n \sim \bar{\phi}(y)$, and require recursion. They are (a) the optimal bridge sampling, (b) the MIS and (c) the Self-IS-with-mix estimators, which are summarized in Table 1. The comparison is done in terms of mean square error (MSE) versus different values of σ_p. The results are averaged over 10^6 independent runs. We set $M + N = 40$, considering the three cases (a) $M = 20$, $N = 20$, (b) $M = 5$, $N = 35$ and (c) $M = 35$, $N = 5$. Furthermore, we consider four scenarios, one ideal and three more realistic scenarios, corresponding to whether we can evaluate $\bar{\phi}(y) = \frac{1}{Z}\phi(y)$ in the right side of the estimators (ideal and impossible scenario) or we can only evaluate $\phi(y)$ (realistic scenarios):

[5]The code used is publicly available at http://www.lucamartino.altervista.org/PUBLIC_CODE_NCE_BRIDGE.zip.

- **Ideal scenario.** We replace $Z = Z_{\mathtt{tr}}$ on the right side of Eqs. (21), (31), and (36), so that the resulting estimators do not require recursion. This setting can also be interpreted as initializing the iterative procedure at the true value, $Z_0 = Z_{\mathrm{tr}}$ (i.e., a very good initialization), and performing a single iteration step, i.e., $T = 1$. The first scenario is for illustration purposes. The results are given in Figure 2.

- **Almost-ideal scenario.** This is a realistic scenario since we apply the recursion using with $T = 10$ iterative steps. However, we start $Z_0 \approx Z_{\mathtt{tr}}$ very close to the true value. The corresponding results are given in Figure 3.

- **Realistic scenario 1.** We set again $T = 10$, but the initializing point is $Z_0 = 0.1$. The corresponding results are given in Figure 4.

- **Realistic scenario 2.** We set again $T = 10$, but the initializing point is $Z_0 = 5$. The corresponding results are given in Figure 5.

Results in ideal scenario. As shown in Figure 2, the optimal bridge estimator provides the worst results in terms of MSE, whereas the MIS estimator provides the best results in line with the studies [21, 22] that consider estimators where the proposal density (hence the denominators of the weights) can be completely evaluated. However, this is not a realistic case in our framework.
Results in the rest of scenarios. As shown in Figures 3, 4, and 5, the optimal bridge sampling gives the best results in the realistic scenarios, but the results of the Self-IS-with-mix estimator (36) are very close and tends to be better for small values of σ_p (smaller than $\theta_{\mathrm{tr}} = 1$ that is the true standard deviation of the model). The MIS estimator provides the worst results except in Figure 3 where we use a very good initialization, where provides the best results.

8.2 Different cost functions for estimating $\theta_{\mathtt{tr}}$

In this section, we focus on the estimation of $\theta_{\mathtt{tr}} = 1$ in EBM, fixing the true normalizing constant $Z_{\mathtt{tr}} = Z(\theta_{\mathtt{tr}})$ in the cost functions to minimize. For the sake of simplicity, we assume again the model in Eq. (64) and the same proposal density in Eq. (66).
We test different cost functions. We consider the cost function $J(\theta) = J(\theta, Z_{\mathtt{tr}})$ in Eq. (49) with different choices of $V(\eta)$, more specifically:

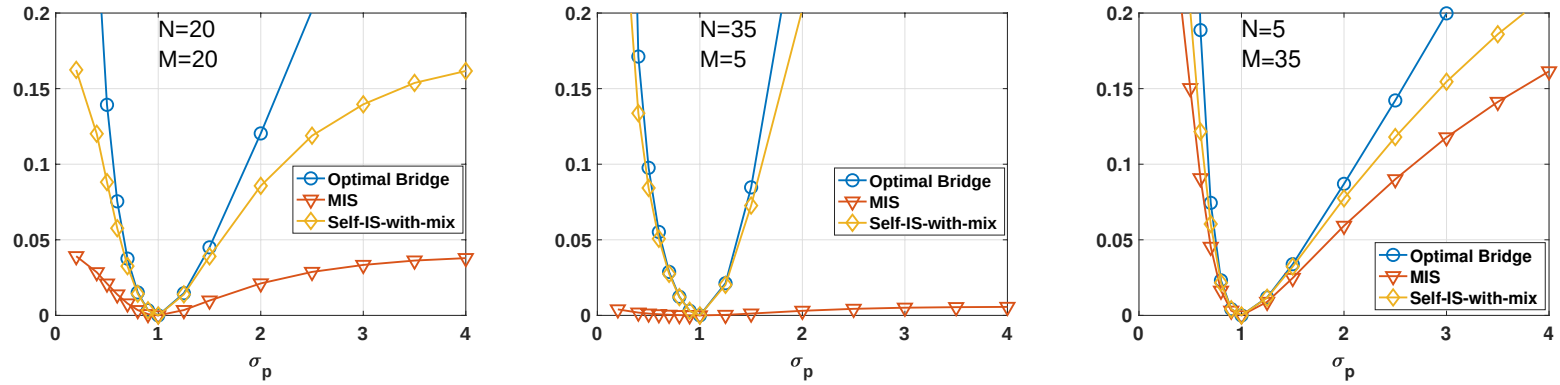

Figure 2: (Ideal scenario) MSE in the estimation of $Z_{\mathtt{tr}}$ versus σ_p. We set $Z = Z_{\mathtt{tr}}$ on the right side of Eqs. (21), (31), and (36), so that the resulting estimators do not require recursion. It can be interpreted as $Z_0 = Z_{\mathrm{tr}}$ and $T = 1$. The figures differ for the numbers of $N \in \{5, 20, 35\}$ and $M \in \{5, 20, 35\}$ such that $N + M = 40$. Surprisingly, the optimal bridge estimator provides the highest MSE values.

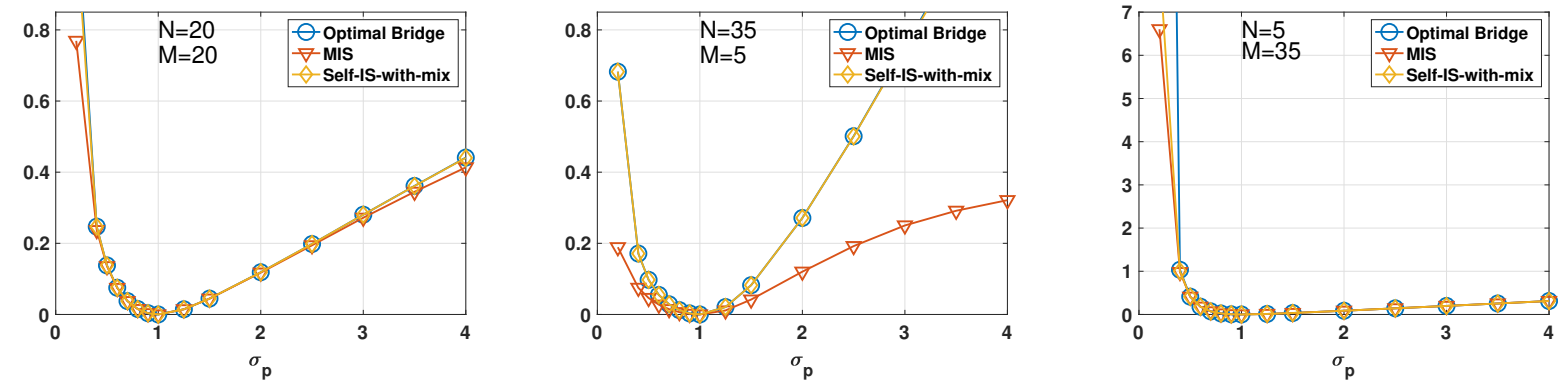

Figure 3: (Almost-ideal scenario) MSE in the estimation of $Z_{\mathtt{tr}}$ versus σ_p. In this figure, we use $Z_0 \approx Z_{\mathrm{tr}}$ and $T = 10$. The figures differ for the numbers of $N \in \{5, 20, 35\}$ and $M \in \{5, 20, 35\}$ such that $N + M = 40$.

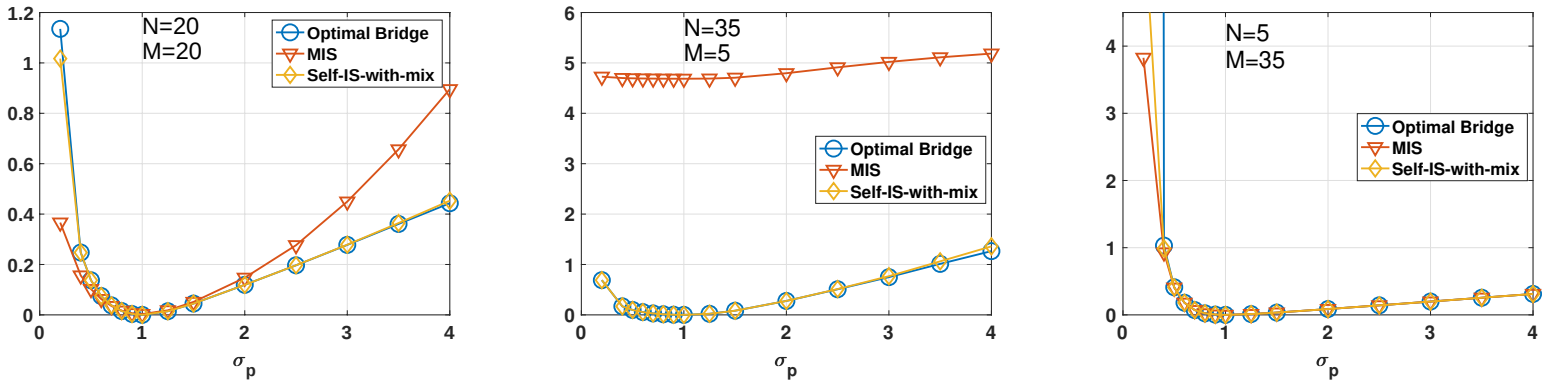

Figure 4: (Realistic scenario 1) MSE in the estimation of $Z_{\mathtt{tr}}$ versus σ_p. In this figure, we use $Z_0 = 0.1$ and $T = 10$. The figures differ for the numbers of $N \in \{5, 20, 35\}$ and $M \in \{5, 20, 35\}$ such that $N + M = 40$.

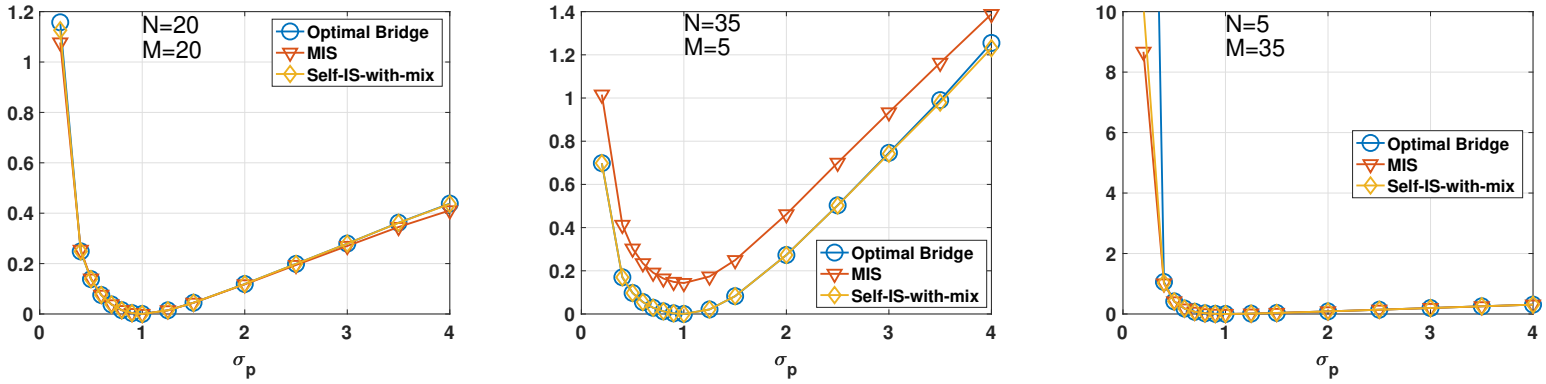

Figure 5: (Realistic scenario 2) MSE in the estimation of $Z_{\mathtt{tr}}$ versus σ_p. In this figure, we use $Z_0 = 5$ and $T = 10$. The figures differ for the numbers of $N \in \{5, 20, 35\}$ and $M \in \{5, 20, 35\}$ such that $N + M = 40$.

- $V(\eta) = -\log(\eta)$ as in Eq. (9),
- $V(\eta) = (1-\eta)^2$,
- $V(\eta) = 1/\eta$ and
- $J_{\texttt{MIS}}(\theta) = J_{\texttt{MIS}}(\theta, Z_{\texttt{tr}})$ in Eq. (48).

Moreover, since $Z_{\texttt{tr}}$ is assumed to be known, we can also compare with the maximum likelihood (ML) estimator [11, 37], which relies solely on $\{y_n\}_{n=1}^N$ and does not depend on the proposal density or on $\{x_m\}_{m=1}^M$.
We compute the MSE in estimation of $\theta_{\texttt{tr}} = 1$ averaged over 5000 independent runs. We vary the standard deviation σ_p of the proposal density. Since the ML solution does not depend on the proposal density, its MSE remains constant with respect to variations in σ_p. We also consider different pairs of N and M values, $\{N = 5, M = 5\}$, $\{N = 5, M = 15\}$, $\{N = 1, M = 20\}$ and $\{N = 1, M = 100\}$.

Results. The curves MSE versus σ_p as depicted in Figure 6. Each figure corresponds to a pair of values of N and M. We can observe that the classical NCE with $V(\eta) = -\log(\eta)$ generally yields good performance, particularly for larger values of σ_p, where its MSE approaches that of the ML solution. However, for certain values of σ_p, other cost functions seem to perform better specially for values of σ_p around the true value $\theta_{\texttt{tr}}$ (that is the standard deviation of the model). Moreover, as M grows and the classes are more unbalanced (having less true data N and more artificial data M), other options of $V(\eta)$ seem to work better than $V(\eta) = -\log \eta$. The cost function $J_{\texttt{MIS}}$ depends strongly on the choice of σ_p. Generally, the choice of the proposal is also a relevant topic. The optimal proposal seems to be different for each cost functions [27, 28, 38]. The analysis of these results suggests that, for $J_{\texttt{MIS}}$, the optimal proposal may be $q_{\texttt{opt}}(y) = \bar{\phi}(y|\theta_{\texttt{tr}})$.

9 Conclusions

In this work, we provide a unified perspective on several techniques that have been developed independently across the literature and different fields. We show the relationships among existing methods as the noise contrastive estimation (NCE), multiple importance sampling, reverse logistic regression (RLR), and bridge sampling. This unified framework not only elucidates the relationships

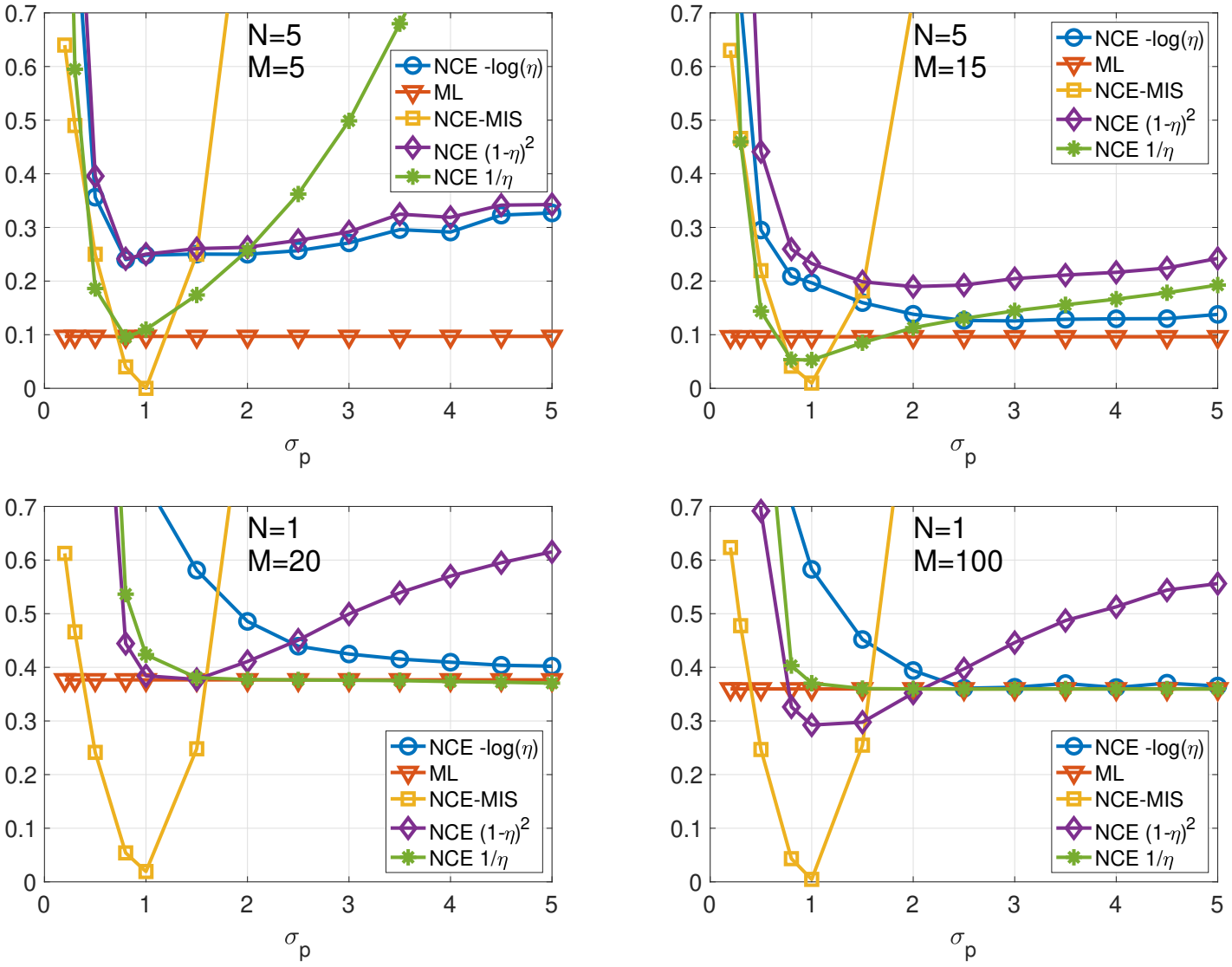

Figure 6: MSE in the estimation of $\theta_{\mathtt{tr}} = 1$ versus σ_p (standard deviation of the proposal/reference density), for different values of N and M.

among existing methods, but also enables the principled design of novel estimators with potentially superior statistical and computational performance.
Contrastive learning, and in particular the NCE method [14, 15], has become a widely adopted and highly successful approach, often regarded as a benchmark method. NCE is asymptotically equivalent to maximum likelihood estimation in the $\boldsymbol{\theta}$-space, as demonstrated in [17, 20], and, as highlighted in this work, it is also equivalent to the optimal bridge sampling solution in the Z-space. This equivalence explains NCEs ability to estimate the normalizing constant and its success in the literature for inference in EBMs. Accordingly, NCE serves as a standard benchmark for frequentist inference in energy-based models.
However, as shown in this work, for specific choices of the proposal (or reference) density and for finite values of N and M, alternative estimation schemes for $\boldsymbol{\theta}$ and Z may yield improved performance. The related code has been made freely available to support reproducibility. This effect has been also highlighted in [20] regarding the inference in the $\boldsymbol{\theta}$-space.
Recursive procedures commonly used for estimating normalizing constants Z (as for the optimal bridge sampling) can also be incorporated into NCE optimization frameworks. Moreover, the joint selection of a specific scoring rule $V(\eta)$ and a proposal density $q(\mathbf{y})$ represents a promising direction for future research. Moreover, the use of alternative scoring rules could lead to the analytical design of novel estimators for Z. In addition, the use of multiple proposal densities, for instance defined through tempered-versions of the EBM, warrants further investigation.

Acknowledgements

L. Martino acknowledges financial support by the PIACERI Starting Grant BA-GRAPH (UPB 28722052144) of the University of Catania.

References

[1] Y. Du and I. Mordatch, "Implicit generation and modeling with energy based models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[2] A. Dawid and Y. LeCun, "Introduction to latent variable energy-based models: a path toward autonomous machine intelligence," *Journal of*

Statistical Mechanics: Theory and Experiment, vol. 2024, no. 10, p. 104011, 2024.

[3] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang, "A tutorial on energy-based learning," *Predicting Structured Data*, pp. 1–59, 2006.

[4] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers, 2008.

[5] F. Llorente, L. Martino, J. Read, and D. Delgado, "A survey of Monte Carlo methods for noisy and costly densities with application to reinforcement learning and ABC," *International Statistical Review*, vol. 93, no. 1, pp. 18–61, 2025.

[6] A. Caimo and A. Mira, "Efficient computational strategies for doubly intractable problems with applications to bayesian social networks," *Statistics and Computing*, vol. 25, pp. 113–125, 2015.

[7] F. Liang, "A double metropolis-hastings sampler for spatial models with intractable normalizing constants," *Journal of Statistical Computation and Simulation*, vol. 80, no. 9, pp. 1007–1022, 2010.

[8] I. Murray, Z. Ghahramani, and D. MacKay, "Mcmc for doubly-intractable distributions," *arXiv preprint arXiv:1206.6848*, 2012.

[9] J. Park and M. Haran, "Bayesian inference in the presence of intractable normalizing functions," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1372–1390, 2018.

[10] C. J. Geyer, "Markov chain Monte Carlo maximum likelihood," *Computing Science and Statistics*, vol. 23, pp. 156–163, 1991.

[11] ——, "On the convergence of Monte Carlo maximum likelihood calculations," *Journal of the Royal Statistical Society, Series B*, vol. 56, no. 2, pp. 261–274, 1994.

[12] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, pp. 695–709, 2005.

[13] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society, Series B*, vol. 36, no. 2, pp. 192–236, 1974.

[14] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," *Journal of Machine Learning Research*, vol. 13, pp. 307–361, 2012.

[15] M. U. Gutmann, S. Kleinegesse, and B. Rhodes, "Statistical applications of contrastive learning," *Behaviormetrika*, vol. 49, pp. 277–301, 2022.

[16] L. Martino, S. Ingrassia, S. Mangano, and L. Scaffidi, "A note on gradient-based parameter estimation for energy-based models," *proceedings of 15th conference of Scientific Meeting of the Classification and Data Analysis Group (CLADAG) — https://vixra.org/abs/2503.0117*, pp. 1–10, 2025.

[17] L. Riou-Durand and N. Chopin, "Noise contrastive estimation: Asymptotics and comparison with MC-MLE," *arXiv:1801.10381*, 2019.

[18] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, p. 193907193934, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.3031549

[19] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 8547–8555.

[20] L. Martino, L. Scaffidi-Domianello, and S. Mangano, "Importance sampling and contrastive learning schemes for parameter estimation in non-normalized models," *viXra:2601.0065*, pp. 1–30, 2026.

[21] A. B. Owen and Y. Zhou, "Safe and effective importance sampling," *Journal of the American Statistical Association*, vol. 95, no. 449, pp. 135–143, 2000.

[22] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Generalized multiple importance sampling," *Statistical Science*, vol. 34, no. 1, pp. 129–155, 2019.

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

[24] F. Llorente, L. Martino, D. Delgado, and J. López-Santiago, "Marginal likelihood computation for model selection and hypothesis testing: An extensive review," *SIAM Review*, vol. 65, no. 1, pp. 3–58, 2023.

[25] G. Storvik, "On the flexibility of Metropolis-Hastings acceptance probabilities in auxiliary variable proposal generation," *Scandinavian Journal of Statistics*, vol. 38, no. 2, pp. 342–358, 2011.

[26] L. Martino and J. Read, "On the flexibility of the design of multiple try Metropolis schemes," *Computational Statistics*, vol. 28, no. 6, pp. 2797–2823, 2013.

[27] O. Chehab, A. Gramfort, and A. Hyvärinen, "The optimal noise in noise-contrastive learning is not what you think," in *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, vol. 180, 2022, pp. 307–316.

[28] ——, "Optimizing the noise in self-supervised learning: From importance sampling to noise-contrastive estimation," *arXiv:2301.09696*, 2023.

[29] C. J. Geyer, "Estimating normalizing constants and reweighting mixtures," *Technical Report, number 568 - School of Statistics, University of Minnesota*, 1994.

[30] M. H. Chen, Q.-M. Shao *et al.*, "On Monte Carlo methods for estimating ratios of normalizing constants," *The Annals of Statistics*, vol. 25, no. 4, pp. 1563–1594, 1997.

[31] E. Cameron and A. Pettitt, "Recursive pathways to marginal likelihood estimation with prior-sensitivity analysis," *Statistical Science*, vol. 29, no. 3, pp. 397–419, 2014.

[32] R. Neal, "The harmonic mean of the likelihood: worst Monte Carlo method ever," *https://radfordneal.wordpress.com/*, 2008.

[33] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," *Journal of Computational Physics*, vol. 23, no. 2, pp. 187–199, 1977.

[34] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[35] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert, "Population Monte Carlo," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 907–929, 2004.

[36] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric, "Adaptive importance sampling: the past, the present, and the future," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 60–79, 2017.

[37] C. J. Geyer and E. A. Thompson, "Likelihood inference for spatial point processes," *Journal of the Royal Statistical Society, Series B*, vol. 61, no. 3, pp. 657–689, 1999.

[38] F. Llorente and L. Martino, "Optimality in importance sampling: A gentle survey," *arXiv:2502.07396*, 2025.

Table 1: Summary of the estimators of Z using $q(\mathbf{y})$ and/or $\bar{\phi}(\mathbf{y})$. The last column shows if a recursive procedure is required. The first four rows correspond to estimators using samples from $\bar{\phi}(\mathbf{y})$ and $q(\mathbf{y})$. The last four rows correspond to estimators using samples from a single proposal density.

Name	**Estimator**	**Samples**	**Rec.**
Opt-Bridge	$\widehat{Z}_{t+1} = \dfrac{\frac{1}{M}\sum_{m=1}^{M} \dfrac{\phi(\mathbf{x}_m)}{\alpha_1\phi(\mathbf{x}_m) + \alpha_2\widehat{Z}_t q(\mathbf{x}_m)}}{\frac{1}{N}\sum_{n=1}^{N} \dfrac{q(\mathbf{y}_n)}{\alpha_1\phi(\mathbf{y}_n) + \alpha_2\widehat{Z}_t q(\mathbf{y}_n)}}$	$\mathbf{y}_n \sim \bar{\phi}(\mathbf{y}), \quad \mathbf{x}_m \sim q(\mathbf{y})$	✔
MIS	$\widehat{Z}_{t+1} = \dfrac{1}{N+M}\sum_{i=1}^{N+M} \dfrac{\widehat{Z}_t\phi(\mathbf{u}_i)}{\alpha_1\phi(\mathbf{u}_i) + \alpha_2\widehat{Z}_t q(\mathbf{u}_i)}$	$\{\mathbf{u}_i\} = \{\mathbf{y}_n\} \cup \{\mathbf{x}_m\}$	✔
Self-IS-with-mix	$\widehat{Z}_{t+1} = \dfrac{\sum_{i=1}^{N+M} \dfrac{\phi(\mathbf{u}_i)}{\alpha_1\phi(\mathbf{u}_i) + \alpha_2\widehat{Z}_t q(\mathbf{u}_i)}}{\sum_{k=1}^{N+M} \dfrac{q(\mathbf{u}_k)}{\alpha_1\phi(\mathbf{u}_k) + \alpha_2\widehat{Z}_t q(\mathbf{u}_k)}}$	$\{\mathbf{u}_i\} = \{\mathbf{y}_n\} \cup \{\mathbf{x}_m\}$	✔
Geo	$\widehat{Z}_{\texttt{geo}} = \sqrt{\dfrac{\frac{1}{M}\sum_{m=1}^{M} \frac{\phi(\mathbf{x}_m)}{q(\mathbf{x}_m)}}{\frac{1}{N}\sum_{n=1}^{N} \frac{q(\mathbf{y}_n)}{\phi(\mathbf{y}_n)}}}$	$\mathbf{y}_n \sim \bar{\phi}(\mathbf{y}), \quad \mathbf{x}_m \sim q(\mathbf{y})$	✗
Stand-IS	$\widehat{Z}_{\texttt{IS}} = \dfrac{1}{M}\sum_{m=1}^{M} \dfrac{\phi(\mathbf{x}_m)}{q(\mathbf{x}_m)}$	$\mathbf{x}_m \sim q(\mathbf{y})$	✗
RIS	$\widehat{Z}_{\texttt{RIS}} = \left(\dfrac{1}{N}\sum_{n=1}^{N} \dfrac{q(\mathbf{y}_n)}{\phi(\mathbf{y}_n)}\right)^{-1}$	$\mathbf{y}_n \sim \bar{\phi}(\mathbf{y})$	✗
Opt-Umb	$\widehat{Z}_{t+1} = \dfrac{\sum_{i=1}^{N+M} \dfrac{\phi(\widetilde{\mathbf{x}}_i)}{\lvert\phi(\widetilde{\mathbf{x}}_i) - \widehat{Z}_t q(\widetilde{\mathbf{x}}_i)\rvert}}{\sum_{k=1}^{N+M} \dfrac{q(\widetilde{\mathbf{x}}_k)}{\lvert\phi(\widetilde{\mathbf{x}}_k) - \widehat{Z}_t q(\widetilde{\mathbf{x}}_k)\rvert}}$	$\widetilde{\mathbf{x}}_i \sim \bar{r}(\mathbf{y}) \propto \lvert\bar{\phi}(\mathbf{y}) - q(\mathbf{y})\rvert$	✔