Importance sampling and contrastive learning schemes for parameter estimation in non-normalized models

L. Martino*, L. Scaffidi*, S. Mangano*

* Universitá degli Studi di Catania, Italy.

January 13, 2026

Abstract

Likelihood-approximation methods and contrastive learning (CL) are two prominent approaches for inference in models with unknown partition function. In this work, we provide a detailed comparison between the likelihood approximation by Geyer's approach (GA) and CL. Rather than increasing the complexity of Geyer's method to enable comparison, as proposed in [1], we adopt the opposite strategy by simplifying CL. We introduce a class of IS-within-CL schemes that estimate the partition function via importance sampling (IS) and reduce the optimization problem to the original parameter space. This perspective motivates the development of novel variants, whose theoretical properties are analyzed and empirically compared in a replicable experimental study. The described IS-within-CL schemes yield an entire approximation of the partition function, so enabling a possible efficient Bayesian inference. An optimal independent proposal density for IS-within-CL methods and the GA is also introduced. Overall, this work contributes to a clearer unification of likelihood-approximation and CL approaches, offering both theoretical understanding and practical tools for inference in energy-based and non-normalized models. Related MATLAB and R codes are also made freely available to help the reproducibility of the results.

1 Introduction

Non-normalized models, also known as energy-based models, define probability distributions only up to an unknown normalizing function $Z_{\texttt{tr}}(\boldsymbol{\theta})$, commonly referred to as the partition function, where the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{d_\theta}$ is the object of interest for inference. In this framework, the model is defined as $\psi(\mathbf{y}|\boldsymbol{\theta}, Z_{\mathtt{tr}}) = \frac{\phi(\mathbf{y}|\boldsymbol{\theta})}{Z_{\mathtt{tr}}(\boldsymbol{\theta})}$ where $\mathbf{y}$ are the observed data [2, 3, 4, 5, 6]. Hence, the probability of an observation is specified through the numerator function $\phi(\mathbf{y}|\boldsymbol{\theta})$, that captures the compatibility between the data and the model parameters, while normalization is implicitly handled by integrating (or summing) over the entire sample space. This formulation provides remarkable modeling flexibility and has been successfully applied across a wide range of domains, including statistical physics, spatial statistics, computer vision, and modern machine learning [7, 8, 9, 10, 11]. In Bayesian inference, these models generate the so-called doubly intractable posteriors [12, 13, 14, 15]. However, the intractability of the partition function in high-dimensional settings poses significant challenges for likelihood-based inference, parameter estimation, and model comparison.

A variety of alternative estimation strategies have been developed to enable practical inference for this kind of models. The main approaches can be broadly organized into four major families. A first class consists of likelihood approximation-based approaches, which approximate the intractable likelihood or partition function, as originally proposed by C. J. Geyer [16, 17, 18]. The main idea is to use importance sampling (IS) estimators to approximate $Z(\boldsymbol{\theta})$ [19, 20]. Hereafter, we refer to it as *Geyer's approach* (GA). A second family is formed by the *score matching* method, which bypasses the computation of the unknown normalization function designing a proper cost function involving the first and second derivatives of the numerator $\phi(\mathbf{y}|\boldsymbol{\theta})$ [21, 22]. A third class includes the *contrastive learning* (CL) approach, more precisely noise-contrastive estimation, which reformulates parameter estimation as a classification or discrimination problem between observed data and artificially generated samples [23, 24]. Finally, the fourth class is formed by the *pseudo-likelihood* methods which replace the full likelihood with products of conditional distributions, thereby avoiding direct evaluation of the partition function at the cost of introducing approximation bias [7, 25].

In this work, we focus on the Geyer's and CL approaches. First of all, we provide an exhaustive description and comparison between the two methodologies. In the literature, the authors in [1] compare GA and CL within an asymptotic framework, considering the joint increase of the number of observed and artificial data. Since

Geyer's approach optimizes only over the parameter space $\boldsymbol{\Theta}$, whereas CL operates in the higher-dimensional space $\boldsymbol{\Theta}\times\mathbb{R}$, the authors reformulate Geyer's method by introducing an additional parameter. This modification increases the complexity of the GA enabling the comparison between the two methods in the common (greater) space $\boldsymbol{\Theta} \times \mathbb{R}$. However, this strategy may be regarded as questionable, since it renders Geyer's approach more complex than its original formulation. In this work, we adopt the opposite strategy by simplifying the CL approach: we estimate the partition function using IS estimators within CL schemes and reduce the optimization problem to the space $\boldsymbol{\Theta}$ (lower-dimensional with respect to $\boldsymbol{\Theta} \times \mathbb{R}$). We refer to them as IS-within-CL schemes. Moreover, the theoretical comparison of Geyer's and CL cost functions motivates the development of novel variants and schemes. We discuss their potential benefits and compare all the proposed approaches through a replicable experimental study. The associated MATLAB and R code is provided to facilitate the reproducibility of the results.[1] Furthermore, we theoretically derive the optimal proposal density for the GA and all IS-within-CL schemes, and we discuss and design an algorithm for its practical implementation. Both GA and the IS-within-CL schemes yield an approximation $\widehat{Z}(\boldsymbol{\theta})$ of the entire partition function $Z_{\texttt{tr}}(\boldsymbol{\theta})$. In this sense, they can be employed for a pre-Bayesian analysis, providing information about high-probability regions of the parameter space as well as a full approximation of $Z_{\texttt{tr}}(\boldsymbol{\theta})$.

2 Problem statement

2.1 Energy-based models (EBMs)

Let $\phi(\mathbf{y}|\boldsymbol{\theta}) = e^{-E(\mathbf{y}|\boldsymbol{\theta})} \geq 0$ is a function parametrized by a vector of parameters $\boldsymbol{\theta}$ taking values in $\boldsymbol{\Theta} \subseteq \mathbb{R}^{d_\theta}$. The non-negative function $E(\mathbf{y}|\boldsymbol{\theta}) = -\log\phi(\mathbf{y}|\boldsymbol{\theta})$ defined is often called *energy function*. We assume that $\phi(\mathbf{y}|\boldsymbol{\theta})$ is analytically known and we can evaluate it. Thus, an energy-based model is represented by a parametrized family of density functions $\psi(\mathbf{y}|\boldsymbol{\theta}, Z_{\texttt{tr}})$, defined for each $\boldsymbol{\theta}$ as

$$\psi(\mathbf{y}|\boldsymbol{\theta}, Z_{\texttt{tr}}) = \frac{\phi(\mathbf{y}|\boldsymbol{\theta})}{Z_{\texttt{tr}}(\boldsymbol{\theta})}, \tag{1}$$

[1] Related code is given at `http://www.lucamartino.altervista.org/PUBLIC_CODE_ISCL.zip`

Generally, for given $\boldsymbol{\theta}$, the following integral cannot be evaluated analytically (or its computational cost is prohibitively high):

$$Z_{\mathtt{tr}}(\boldsymbol{\theta}) = \int_{\mathcal{D}} \phi(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y}. \tag{2}$$

Namely, $Z_{\mathtt{tr}}(\boldsymbol{\theta})$ is unknown because the integral above cannot be solved analytically in closed form, i.e., is intractable.[2] The subindex $\mathtt{tr}$ stands for "*true*" function in this case (and *true* value when referred to $\boldsymbol{\theta}$, as shown in Table 1). Hence, the normalizing constant $Z_{\mathtt{tr}}(\boldsymbol{\theta})$, often called *partition function*, cannot be evaluated point-wise. This represents a challenge for making inference on $\boldsymbol{\theta}$, as we discuss in the rest of the work.

2.2 Observed data and goal

Let us assume that we have an observed dataset $\underline{\mathbf{y}} = \mathbf{y}_{1:N} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \in \mathcal{D}^N$, 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.,

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

Note that $Z_{\mathtt{tr}}(\boldsymbol{\theta}_{\mathtt{tr}})$ is a scalar normalizing constant, i.e., the true partition function evaluated at $\boldsymbol{\theta}_{\mathtt{tr}}$. As a consequence, $\psi(\mathbf{y}|\boldsymbol{\theta}_{\mathtt{tr}}, Z_{\mathtt{tr}})$ can be regarded as the true model, in the sense that it represents the model generating the observed data. Table 1 summarizes the main notation of the work. The aim of this work is to infer the parameter vector $\boldsymbol{\theta}$ and, possibly, to approximate the function $Z(\boldsymbol{\theta})$. We focus on a frequentist approach. However, all the algorithms in this work can be employed for a pre-Bayesian analysis.

3 Likelihood function in EBMs

The most straightforward approach to perform inference in EBMs is through maximum likelihood estimation (MLE). In order to estimate the parameter of the

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

Table 1: Main notation of the work and energy-based models (EBMs).

Notation	**Description**
$\boldsymbol{\theta}_{\mathtt{tr}}$	True vector of parameters
$Z_{\mathtt{tr}}(\boldsymbol{\theta}) = \int_{\mathcal{D}} \phi(\mathbf{y}\|\boldsymbol{\theta})d\mathbf{y}$	True partition function
$\psi(\mathbf{y}\|\boldsymbol{\theta}_{\mathtt{tr}}, Z_{\mathtt{tr}}) = \frac{\phi(\mathbf{y}\|\boldsymbol{\theta}_{\mathtt{tr}})}{Z_{\mathtt{tr}}(\boldsymbol{\theta}_{\mathtt{tr}})}$	Generating EBM (true model that generates the observed data $\mathbf{y}_{1:N}$)
$\mathbf{y}_n \sim \psi(\mathbf{y}\|\boldsymbol{\theta}_{\mathtt{tr}}, Z_{\mathtt{tr}})$	n-th observed data
$Z_{\mathtt{tr}}(\boldsymbol{\theta}_{\mathtt{tr}})$	Normalizing constant of the true model
$\psi(\mathbf{y}\|\boldsymbol{\theta}, Z_{\mathtt{tr}}) = \frac{\phi(\mathbf{y}\|\boldsymbol{\theta})}{Z_{\mathtt{tr}}(\boldsymbol{\theta})}$	EBM with true partition function but with a generic $\boldsymbol{\theta}$
$Z(\boldsymbol{\theta})$	Generic partition function
$\widehat{Z}(\boldsymbol{\theta})$	Approximated/estimated partition function

	IS-based	**CL**
$q(\mathbf{x})$	Proposal density	Reference density
$\mathbf{x}_m \sim q(\mathbf{x})$	Auxiliary data	Reference data
$\underline{\mathbf{y}} = \mathbf{y}_{1:N} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$	Set of true observed data	
$\underline{\mathbf{x}} = \mathbf{x}_{1:N} = \{\mathbf{x}_1, ..., \mathbf{x}_M\}$	Set of auxiliary/reference data	
$\underline{\mathbf{u}} = \underline{\mathbf{y}} \cup \underline{\mathbf{x}}$	Union of the two sets above: observed and auxiliary/reference data	

distribution, the likelihood function of $\boldsymbol{\theta}$ given $\underline{\mathbf{y}}$ is given by

$$L(\underline{\mathbf{y}}|\boldsymbol{\theta}, Z_{\mathtt{tr}}) = p(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N|\boldsymbol{\theta}) = \prod_{n=1}^{N} \psi(\mathbf{y}_n|\boldsymbol{\theta}, Z_{\mathtt{tr}}) = \frac{1}{Z_{\mathtt{tr}}(\boldsymbol{\theta})^N} \prod_{n=1}^{N} \phi(\mathbf{y}_n|\boldsymbol{\theta}), \quad (4)$$

and the corresponding log-likelihood is

$$\log L(\underline{\mathbf{y}}|\boldsymbol{\theta}, Z_{\mathtt{tr}}) = \sum_{n=1}^{N} \log \psi(\mathbf{y}_n|\boldsymbol{\theta}, Z_{\mathtt{tr}}), \quad (5)$$

$$= -\sum_{n=1}^{N} E(\mathbf{y}_n|\boldsymbol{\theta}) - N \log Z_{\mathtt{tr}}(\boldsymbol{\theta}), \quad (6)$$

where $E(\mathbf{y}_n|\boldsymbol{\theta}) = -\log \phi(\mathbf{y}_n|\boldsymbol{\theta})$. The ML estimation of $\boldsymbol{\theta}$ is often reformulated as the minimization of a cost function defined as *the negative log-likelihood function* (NLL), i.e.,

$$J_{\mathtt{NLL}}(\boldsymbol{\theta}|Z_{\mathtt{tr}}, \underline{\mathbf{y}}) = -\log L(\underline{\mathbf{y}}|\boldsymbol{\theta}, Z_{\mathtt{tr}}) = \sum_{n=1}^{N} E(\mathbf{y}_n|\boldsymbol{\theta}) + N \log Z_{\mathtt{tr}}(\boldsymbol{\theta}), \quad (7)$$

so that $\widehat{\boldsymbol{\theta}}_{\mathtt{ML}} = \arg_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max L(\underline{\mathbf{y}}|\boldsymbol{\theta}, Z_{\mathtt{tr}}) = \arg_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \min J_{\mathtt{NLL}}(\boldsymbol{\theta}|Z_{\mathtt{tr}}, \underline{\mathbf{y}})$. Generally, a standard widely-used optimization approach is based on the so called *gradient-descent*, where the information of the gradient $\nabla_{\boldsymbol{\theta}} J_{\mathtt{NLL}}(\widehat{\boldsymbol{\theta}}_{t-1}|Z_{\mathtt{tr}}, \underline{\mathbf{y}})$ is required.

Remark 1. However, $Z_{\mathtt{tr}}(\boldsymbol{\theta})$ is unknown for any $\boldsymbol{\theta}$, and cannot be computed in closed form. As a consequence, we cannot compute $\nabla_{\boldsymbol{\theta}} J_{\mathtt{NLL}}(\boldsymbol{\theta}|Z_{\mathtt{tr}}, \underline{\mathbf{y}})$ neither. Therefore, suitable strategies must be adopted for approximating $Z_{\mathtt{tr}}(\boldsymbol{\theta})$ (or, at least, its gradient).

4 IS estimators for $Z_{\mathtt{tr}}(\boldsymbol{\theta})$

Theoretically speaking, the simplest approaches rely on applications of some numerical method for computing $Z_{\mathtt{tr}}(\boldsymbol{\theta})$. One way to obtain this approximation is based on importance sampling (IS) [20]. Let $q(\mathbf{x})$ be a known normalized density ($\int_{\mathcal{D}} q(\mathbf{x})d\mathbf{x} = 1$) that we are able to draw from. This density is chosen by the user and that does not depend on $\boldsymbol{\theta}$. Let $\underline{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$, be a sample of size M generated from $q(\mathbf{x})$, i.e.,

$$\mathbf{x}_m \sim q(\mathbf{x}), \qquad \mathbf{x}_m \in \mathcal{D} \subseteq \mathbb{R}^d \qquad m = 1, ..., M.$$

In this context, $q(\mathbf{x})$ is referred to as the *proposal density*, and the samples $\mathbf{x}_m$ are often called *auxiliary* or *artificial data* [17]. In this case, the IS estimator is given by

$$\widehat{Z}(\boldsymbol{\theta}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\phi(\mathbf{x}_m \mid \boldsymbol{\theta})}{q(\mathbf{x}_m)}, \qquad \mathbf{x}_m \sim q(\mathbf{x}), \tag{8}$$

which is unbiased and consistent for every $\boldsymbol{\theta}$, even when the same set of samples $\mathbf{x}_1, \ldots, \mathbf{x}_M$ is employed. A potentially more efficient estimator is

$$\widehat{Z}_\theta(\boldsymbol{\theta}) = \frac{1}{M} \sum_{m=1}^{M} \frac{\phi(\mathbf{x}_m^{(\boldsymbol{\theta})} \mid \boldsymbol{\theta})}{q(\mathbf{x}_m^{(\boldsymbol{\theta})} \mid \boldsymbol{\theta})}, \qquad \mathbf{x}_m^{(\boldsymbol{\theta})} \sim q(\mathbf{x}|\boldsymbol{\theta}), \tag{9}$$

where the proposal depends on $\boldsymbol{\theta}$, and a different set $\{\mathbf{x}_1^{(\boldsymbol{\theta})}, \ldots, \mathbf{x}_M^{(\boldsymbol{\theta})}\}$ is generated for each value of $\boldsymbol{\theta}$. In this scenario, the optimal proposal is $q(\mathbf{x}|\boldsymbol{\theta}) \propto \phi(\mathbf{x}|\boldsymbol{\theta})$ [20]. Hence, for a fixed $\boldsymbol{\theta}$, the estimator $\widehat{Z}_\theta(\boldsymbol{\theta})$ can be generally more efficient than $\widehat{Z}(\boldsymbol{\theta})$. The quality (i.e., the variance) of $\widehat{Z}(\boldsymbol{\theta})$ changes with $\boldsymbol{\theta}$: the proposal $q(\mathbf{x})$ may be well suited for some values of $\boldsymbol{\theta}$ and inadequate for others [20]. Nevertheless, $\widehat{Z}(\boldsymbol{\theta})$ provides an approximation that is valid for all $\boldsymbol{\theta}$. Moreover, note that once the known part ϕ, the proposal q, and the auxiliary samples $\mathbf{x}_m$ are fixed, $\widehat{Z}(\boldsymbol{\theta})$ becomes a deterministic, analytically known, and easily evaluable function of $\boldsymbol{\theta}$.

4.1 Geyer's approach (GA)

In [17], the first IS approximation (8) is substituted directly into Eq. (7):

$$\begin{aligned} J_{\texttt{GA}}(\boldsymbol{\theta}) = J_{\texttt{NLL}}(\boldsymbol{\theta}|\widehat{Z}, \underline{\mathbf{y}}) &= \sum_{n=1}^{N} E(\mathbf{y}_n|\boldsymbol{\theta}) + N \log \widehat{Z}(\boldsymbol{\theta}), \\ &= \sum_{n=1}^{N} E(\mathbf{y}_n|\boldsymbol{\theta}) + N \log \left[\frac{1}{M} \sum_{m=1}^{M} \frac{\phi(\mathbf{x}_m|\boldsymbol{\theta})}{q(\mathbf{x}_m)} \right], \quad \mathbf{x}_m \sim q(\mathbf{x}). \end{aligned} \tag{10}$$

By IS arguments, we have $J_{\texttt{GA}}(\boldsymbol{\theta}) \approx J_{\texttt{NLL}}(\boldsymbol{\theta}|Z_{\texttt{tr}}, \underline{\mathbf{y}})$. The idea is then to minimize this function [17], i.e., $\widehat{\boldsymbol{\theta}}_{\texttt{GA}} = \arg\min J_{\texttt{GA}}(\boldsymbol{\theta})$. Clearly, as $M \to \infty$,

$$\begin{aligned} \widehat{Z}(\boldsymbol{\theta}) &\longrightarrow Z_{\texttt{tr}}(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta}, \\ J_{\texttt{GA}}(\boldsymbol{\theta}) = J_{\texttt{NLL}}(\boldsymbol{\theta}|\widehat{Z}, \underline{\mathbf{y}}) &\longrightarrow J_{\texttt{NLL}}(\boldsymbol{\theta}|Z_{\texttt{tr}}, \underline{\mathbf{y}}) \qquad \text{and, as a consequence,} \\ \widehat{\boldsymbol{\theta}}_{\texttt{GA}} &\longrightarrow \widehat{\boldsymbol{\theta}}_{\texttt{ML}}. \end{aligned} \tag{11}$$

Furthermore, an approximation of the gradient can be also directly computed as

$$\frac{1}{N}\nabla_{\boldsymbol{\theta}} J_{\texttt{GA}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} E(\mathbf{y}_n|\boldsymbol{\theta}) - \sum_{m=1}^{M} \bar{w}_m^{(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_m|\boldsymbol{\theta}), \tag{12}$$

where we have defined the normalized importance weights, $\bar{w}_m^{(\boldsymbol{\theta})} = \frac{w_m^{(\boldsymbol{\theta})}}{\sum_{j=1}^{M} w_j^{(\boldsymbol{\theta})}}$ where $w_m^{(\boldsymbol{\theta})} = \frac{\phi(\mathbf{x}_m|\boldsymbol{\theta})}{q(\mathbf{x}_m)}$. See Appendix A for the derivation.

Remark 2. $J_{\texttt{GA}}(\boldsymbol{\theta})$ is a random function depending on the generated samples $\mathbf{x}_m$'s. However, given the known part ϕ of the model, the proposal q and fixing the auxiliary samples $\mathbf{x}_m$, then $J_{\texttt{GA}}(\boldsymbol{\theta})$ becomes a deterministic, analytically known, and evaluable function of $\boldsymbol{\theta}$.

Remark 3. In Geyer's approach, the same set of auxiliary samples $\underline{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\} \sim q(\mathbf{x})$ is used for every $\boldsymbol{\theta}$. As a consequence, $\widehat{Z}(\boldsymbol{\theta})$ may provide an accurate estimate for some values of $\boldsymbol{\theta}$, while performing poorly for others. Namely, the quality of the estimator (i.e., its variance, since it is unbiased) varies with $\boldsymbol{\theta}$ [20]. There are two main strategies to reduce the variance of $J_{\texttt{GA}}(\boldsymbol{\theta})$: (a) increasing the sample size M; or (b) for a fixed M, selecting a proposal $q(\mathbf{x})$ that performs well for a broad range of $\boldsymbol{\theta}$ values.

5 Contrastive learning (CL)

From a statistical perspective, contrastive learning (CL) is an alternative framework where the inference is driven by comparing samples from the observed data distribution against samples from a reference/noise distribution. More specifically, the idea in CL is to learn $\boldsymbol{\theta}$ 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 $\mathbf{x}_m \sim p(\mathbf{u}|a=0)$. Note that the framework considered so far can be rewritten as

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

and

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

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

$$\underbrace{\mathbf{u}_1 = \mathbf{y}_1, \ldots, \mathbf{u}_N = \mathbf{y}_N,}_{a=1} \underbrace{\mathbf{u}_{N+1} = \mathbf{x}_1, \ldots, \mathbf{u}_{N+M} = \mathbf{x}_M,}_{a=0}. \tag{13}$$

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 $\mathbf{x}_m$'s samples are usually called *reference data* and q is often referred as *reference density.* The vectors $\mathbf{x}_1, ..., \mathbf{x}_M$ are called reference/noise data.

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 $\boldsymbol{\theta}_{\tt tr}$ and $Z_{\tt tr}(\boldsymbol{\theta}_{\tt tr})$, comparing the two sets of data. The marginal (prior) probabilities of the labels can be approximated as $p(a=1) \approx \frac{N}{N+M}$, $p(a=0) \approx \frac{M}{N+M}$. Setting $\nu = \frac{p(a=0)}{p(a=1)} \approx \frac{M}{N}$ and $\boldsymbol{\xi} = [\boldsymbol{\theta}, Z]$, the posterior probabilities are

$$\begin{aligned} p(a=1|\mathbf{u}) = \eta(\mathbf{u}, \boldsymbol{\xi}) = \eta(\mathbf{u}, \boldsymbol{\theta}, Z) &= \frac{p(\mathbf{u}|a=1)p(a=1)}{p(\mathbf{u}|a=1)p(a=1) + p(\mathbf{u}|a=0)p(a=0)} \\ &= \frac{\psi(\mathbf{u}|\boldsymbol{\theta}, Z)}{\psi(\mathbf{u}|\boldsymbol{\theta}, Z) + \nu q(\mathbf{u})}, \qquad (14) \\ &= \frac{\phi(\mathbf{u}, \boldsymbol{\theta})}{\phi(\mathbf{u}, \boldsymbol{\theta}) + \nu Z(\boldsymbol{\theta}) q(\mathbf{u})}, \qquad (15) \end{aligned}$$

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 $\boldsymbol{\xi} = [\boldsymbol{\theta}, Z]$.

Remark 4. Note that here we are considering a generic vector $\boldsymbol{\theta}$ and a generic function $Z(\boldsymbol{\theta})$.

The goal is to learn a *classification rule* $c(\mathbf{u})$ that, for a generic input $\mathbf{u}$, yields a label prediction/estimation $\widehat{a} = c(\mathbf{u}) \in \{0, 1\}$. It is possible to show that the optimal rule $c^*(\mathbf{u})$ that achieves the largest classification accuracy is the *Bayes classification rule*,

$$\widehat{a} = c^*(\mathbf{u}) = \begin{cases} 1 & \text{if} \quad \eta(\mathbf{u}) \geq 0.5, \\ 0 & \text{if} \quad \eta(\mathbf{u}) < 0.5, \end{cases} . \tag{16}$$

The classification literature provides a wealth of methods to learn *(directly or indirectly)* an approximation of $c^*(\mathbf{u})$. Considering the parametric family $\eta(\mathbf{u}, \boldsymbol{\xi})$, a common approach to find a good approximation $\widehat{\eta}(\mathbf{u})$ consists in minimizing an empirical risk:

$$J_{\texttt{CL}}(\boldsymbol{\xi}) = \sum_{n=1}^{N} V(\eta(\mathbf{y}_n; \boldsymbol{\xi})) + \sum_{m=1}^{M} V(1 - \eta(\mathbf{x}_m; \boldsymbol{\xi})), \tag{17}$$

where $V(\eta) : [0, 1] \to [0, 1]$ is a given is a *non-increasing function* in $[0, 1]$ (called *loss function*).[3] Therefore, for each possible input $\mathbf{u}$, the procedure is to obtain

$$\widehat{\boldsymbol{\xi}} = \arg\min_{\boldsymbol{\xi}} J_{\texttt{CL}}\big(\boldsymbol{\xi}\big), \quad \text{and set:} \quad \widehat{\eta}(\mathbf{u}) = \eta(\mathbf{u}, \widehat{\boldsymbol{\xi}}), \qquad \forall \mathbf{u}. \tag{18}$$

CL cost function. Perhaps, the most common loss function is log-loss,

$$V(\eta) = -\log(\eta). \tag{19}$$

With this loss, the risk $V(\eta)$ above coincides with the negative log-likelihood function assuming a Bernoulli model, and with the so-called cross-entropy. Namely, the empirical risk becomes

$$\begin{cases} J_{\texttt{CL}}\big(\boldsymbol{\xi}\big) = J_{\texttt{CL}}(\boldsymbol{\theta}, Z) = -\sum_{n=1}^{N} \log\left(\eta\left(\mathbf{y}_n, \boldsymbol{\theta}, Z\right)\right) - \sum_{m=1}^{M} \log\left(1 - \eta\left(\mathbf{x}_m, \boldsymbol{\theta}, Z\right)\right), \quad \text{with} \\ \qquad \eta(\mathbf{u}, \boldsymbol{\theta}, Z) = \dfrac{\psi(\mathbf{u}|\boldsymbol{\theta}, Z)}{\psi(\mathbf{u}|\boldsymbol{\theta}, Z) + \nu q(\mathbf{u})}, \qquad 1 - \eta(\mathbf{u}, \boldsymbol{\theta}, Z) = \dfrac{\nu q(\mathbf{u})}{\psi(\mathbf{u}|\boldsymbol{\theta}, Z) + \nu q(\mathbf{u})}. \end{cases} \tag{20}$$

[3] The idea above is to penalize when $\eta(\mathbf{y}_n, \boldsymbol{\xi}) = p(a = 1|\mathbf{u} = \mathbf{y}_n)$ is small (i.e., close to zero), since $\mathbf{y}_n$ has label $a = 1$ and $\eta(\mathbf{y}_n, \boldsymbol{\xi})$ should be close to 1. Whereas, we should penalize the small values of $1 - \eta(\mathbf{x}_m, \boldsymbol{\xi})$ since $\mathbf{x}_m$ has label $a = 0$ and $\eta(\mathbf{x}_m, \boldsymbol{\xi})$ should be close to 0 (so that $1 - \eta(\mathbf{x}_m, \boldsymbol{\xi})$ should be close to 1).

Recall that $\nu = \frac{M}{N}$. Hence, the final cost function to minimize is

$$J_{\texttt{CL}}(\boldsymbol{\theta}, Z) = -\sum_{n=1}^{N} \log\left[\frac{\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z)}{\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z) + \nu q(\mathbf{y}_n)}\right] - \sum_{m=1}^{M} \log\left[\frac{\nu q(\mathbf{x}_m)}{\psi(\mathbf{x}_m|\boldsymbol{\theta}, Z) + \nu q(\mathbf{x}_m)}\right], \tag{21}$$

$$= -\sum_{n=1}^{N} \log\left[\frac{\phi(\mathbf{y}_n, \boldsymbol{\theta})}{\phi(\mathbf{y}_n, \boldsymbol{\theta}) + \nu Z(\boldsymbol{\theta}) q(\mathbf{y}_n)}\right] - \sum_{m=1}^{M} \log\left[\frac{\nu Z(\boldsymbol{\theta}) q(\mathbf{x}_m)}{\phi(\mathbf{x}_m, \boldsymbol{\theta}) + \nu Z(\boldsymbol{\theta}) q(\mathbf{x}_m)}\right]. \tag{22}$$

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

$$[\widehat{\boldsymbol{\theta}}_{\texttt{CL}}, \widehat{Z}_{\texttt{CL}}] = \arg\min J_{\texttt{CL}}(\boldsymbol{\theta}, Z), \tag{23}$$

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

$$\widehat{Z}_{\texttt{CL}} \longrightarrow Z_{\texttt{tr}}(\boldsymbol{\theta}_{\texttt{tr}}), \tag{24}$$

is a scalar value, that is the approximation of $Z_{\texttt{tr}}$ in one specific point, $\boldsymbol{\theta}_{\texttt{tr}}$. The method above is also called *noise contrastive estimation* (NCE).

Remark 5. A CL scheme is completely defined by the choice of: (a) the loss function V, (b) the number M of reference points and (c) the choice of reference density q. Regarding the optimal choice if the reference density in CL, see [26, 27].

Remark 6. Let us consider Eq. (22). The cost function $J_{\texttt{CL}}$ is a random function, but fixing q and the reference samples $\mathbf{x}_m$, $J_{\texttt{CL}}$ becomes a deterministic function of $\boldsymbol{\theta}$ and Z, like in the Geyer's approach.

6 On the relationship between GA and CL

In this section, we emphasize similarities, connections but also differences between GA and CL. More precisely, the main aspects comparing GA and CL are summarized below:

1. In both techniques, we use two datasets, (a) the observed data and (b) the auxiliary/reference data generated by a proposal/reference density q.

2. In both cases, we obtain two cost functions $J_{\texttt{GA}}$ and $J_{\texttt{CL}}$ that, fixing the M auxiliary/reference samples $\mathbf{x}_m$, become deterministic functions.

3. In both cases, the choice of the proposal/reference density q strongly affects the overall performance. However, the optimal density q generally differs across the two schemes. Regarding the optimal proposal q in IS, see [20]; whereas, regarding the optimal reference density, see [26, 27].

4. The cost function $J_{\texttt{CL}}$ takes values in a higher-dimensional space, i.e., $J_{\texttt{CL}} : \mathbb{R}^{d_\theta+1} \to \mathbb{R}$, whereas $J_{\texttt{GA}} : \mathbb{R}^{d_\theta} \to \mathbb{R}$. In other words, $J_{\texttt{CL}}$ depends on one additional input, so its support domain has one higher dimension than that of $J_{\texttt{GA}}$. This difference arises because, in the GA scheme, $Z(\boldsymbol{\theta})$ is pre-approximated using importance sampling. Furthermore, in standard CL, the scalar value $Z_{\texttt{tr}}(\boldsymbol{\theta}_{\texttt{tr}})$ is considered as an additional parameter to be estimated.

5. Using GA, we obtain an approximation of the full partition function $Z_{\texttt{tr}}(\boldsymbol{\theta})$, whereas with CL we only obtain a point-wise approximation of the specific value $Z_{\texttt{tr}}(\boldsymbol{\theta}_{\texttt{tr}})$: this represents just a value, the function form remains unknown.

Moreover, recalling that $\nu = \frac{M}{N}$, we can rewrite Eq. (21) as

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

where we have multiplied numerators and denominators of the fractions (inside the log) by $\frac{1}{M+N}$, and we have also defined $\alpha_1 = \frac{N}{M+N}$ and $\alpha_2 = \frac{M}{M+N}$ (note that $\alpha_1 + \alpha_2 = 1$). Furthermore, using the property $\log(ab) = \log(a) + \log(b)$, we obtain

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

where we have set $C_0 = -N \log \alpha_1 - M \log \alpha_2$, that is a constant value independent from $\boldsymbol{\theta}$ and Z. Denoting the mixture in the denominators as

$$\mathtt{MIX}(\mathbf{u}|\boldsymbol{\theta}, Z) = \alpha_1 \psi(\mathbf{u}|\boldsymbol{\theta}, Z) + \alpha_2 q(\mathbf{u}),$$

and, using again the properties of the logarithm, then we have

$$\begin{aligned} J_{\mathtt{CL}}(\boldsymbol{\theta}, Z) = & -\sum_{n=1}^{N} \log\left[\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z)\right] + \sum_{n=1}^{N} \log\left[\mathtt{MIX}(\mathbf{y}_n|\boldsymbol{\theta}, Z)\right] + \\ & \qquad - \sum_{m=1}^{M} \log\left[q(\mathbf{x}_m)\right] + \sum_{m=1}^{M} \log\left[\mathtt{MIX}(\mathbf{x}_m|\boldsymbol{\theta}, Z)\right] + C_0, \\ = & \ J_{\mathtt{NLL}}(\boldsymbol{\theta}|Z, \underline{\mathbf{y}}) + \sum_{n=1}^{N} \log\left[\mathtt{MIX}(\mathbf{y}_n|\boldsymbol{\theta}, Z)\right] + \sum_{m=1}^{M} \log\left[\mathtt{MIX}(\mathbf{x}_m|\boldsymbol{\theta}, Z)\right] + C_1, \\ = & \ \underbrace{J_{\mathtt{NLL}}(\boldsymbol{\theta}|Z, \underline{\mathbf{y}})}_{\text{term 1}} + \underbrace{\sum_{i=1}^{N+M} \log\left[\mathtt{MIX}(\mathbf{u}_i|\boldsymbol{\theta}, Z)\right]}_{\text{term 2}} + C_1, \end{aligned} \tag{25}$$

where we have set $C_1 = C_0 - \sum_{m=1}^{M} \log\left[q(\mathbf{x}_m)\right]$ and used Eq. (13).

Remark 7. The first term in Eq. (25), i.e., $J_{\mathtt{NLL}}(\boldsymbol{\theta}|Z, \underline{\mathbf{y}})$, coincides with ML approach when $Z_{\mathtt{tr}}(\boldsymbol{\theta})$ is known, and/or with the Geyer's scheme when the estimator $\widehat{Z}(\boldsymbol{\theta})$ is employed. Therefore, focusing on the inference of $\boldsymbol{\theta}$, the essential difference between $J_{\mathtt{GA}}$ and $J_{\mathtt{CL}}$ lies in the second term of Eq. (25).

Note that, with the second term, $\sum_{i=1}^{N+M} \log\left[\mathtt{MIX}(\mathbf{u}_i|\boldsymbol{\theta}, Z)\right]$, we are applying the deterministic mixture idea [28, 29, 30].

Remark 8. Let us focus on the inference of $\boldsymbol{\theta}$. The first term, $J_{\mathtt{NLL}}(\boldsymbol{\theta}|Z, \underline{\mathbf{y}})$, corresponds to the negative log-likelihood in Eq. (7), which must be minimized to estimate $\boldsymbol{\theta}$, providing or approximating $Z_{\mathtt{tr}}(\boldsymbol{\theta})$. The second term, instead, is the (*positive*) log-likelihood associated with fitting the mixture $\mathtt{MIX}(\mathbf{u}|\boldsymbol{\theta}, Z) = \alpha_1 \psi(\mathbf{u}|\boldsymbol{\theta}, Z) + \alpha_2 q(\mathbf{u})$ to the data $\mathbf{u}_1, \ldots, \mathbf{u}_{M+N}$; in this case, *the maximization* provides an estimator of $\boldsymbol{\theta}_{\mathtt{tr}}$. Thus, assuming that $Z_{\mathtt{tr}}(\boldsymbol{\theta})$ is known, both terms separately yield consistent estimators of $\boldsymbol{\theta}_{\mathtt{tr}}$, the first through minimization, the second through maximization. Surprisingly, however, their sum (despite combining a minimization and a maximization objective) still defines a suitable cost function to minimize, also leading to a consistent estimator of $\boldsymbol{\theta}_{\mathtt{tr}}$.

7 Possible alternative techniques

Given the previous considerations, we may formulate and evaluate various alternative cost functions. The performance of the novel functions can be compared with $J_{\texttt{GA}}$ and $J_{\texttt{CL}}$. Table 2 presents six functional forms under consideration, four of which, up to our knowledge, are novel in the context of EBMs. Note that standard CL is also denoted as CL-1 in Table 2. We denote by CL-2 the method based on a cost function identical to the standard CL, except for the sign reversal of the second term (associated with the mixture), so that both terms contribute to a single cost function to be minimized for estimating $\boldsymbol{\theta}_{\texttt{tr}}$. The cost function referred to as CL-3 almost coincides again with that of the standard CL, but considers only the true observed data, $\{\mathbf{y}_n\}_{n=1}^N$, in the second (mixture) term. Finally, the two techniques denoted as Mixture-1 and Mixture-2 rely solely on the mixture component. Mixture-1 uses all available data $\{\mathbf{u}_n\}_{n=1}^{N+M}$, whereas Mixture-2 considers only the observed samples $\{\mathbf{y}_n\}_{n=1}^N$.

Furthermore, from the same functional form, different procedures can be derived depending on whether Z is treated: (a) as a variable to optimize or (b) replaced by an estimation. For instance, we can include IS estimators of $Z_{\texttt{tr}}(\boldsymbol{\theta})$ and use the cost functions only to find an estimator of $\boldsymbol{\theta}_{\texttt{tr}}$, exactly as in the GA. In this case, we have two possibilities:

1. Use the same samples $\mathbf{x}_m$'s as artificial and reference data, hence a unique $q(\cdot)$ playing both roles: proposal and reference densities. Thus, the estimator of $Z_{\texttt{tr}}(\boldsymbol{\theta})$ is

$$\widehat{Z}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \frac{\phi(\mathbf{x}_m|\boldsymbol{\theta})}{q(\mathbf{x}_m)}, \qquad \mathbf{x}_m \sim q(\mathbf{x}). \tag{26}$$

2. Alternatively, we may employ a different density $q_2(\cdot)$ as proposal, and generate M_2 different samples $\mathbf{s}_m$'s, which are used exclusively to obtain an estimator of $Z_{\texttt{tr}}(\boldsymbol{\theta})$, i.e.,

$$\widehat{Z}_S(\boldsymbol{\theta}) = \frac{1}{M_2}\sum_{k=1}^{M_2} \frac{\phi(\mathbf{s}_k|\boldsymbol{\theta})}{q_2(\mathbf{s}_k)}, \qquad \mathbf{s}_k \sim q_2(\mathbf{x}), \tag{27}$$

   with $k = 1, ..., M_2$.

The computational cost in terms of evaluation of the model numerator, $\phi(\cdot|\boldsymbol{\theta})$, is clearly different. Indeed, in the first case, only M auxiliary/reference samples are drawn, whereas in the second case $M+M_2$ samples are generated: M serving as reference points (i.e., $\mathbf{x}_m$), and M_2 as auxiliary data (i.e., $\mathbf{s}_j$) for obtaining $\widehat{Z}_S$. Note that the first case can be recovered as a special case of the second one, setting $M_2 = M$ and $\mathbf{s}_m = \mathbf{x}_m$. Thus, we can assert that, in the first scenario, we are *recycling* the samples $\mathbf{x}_m$ to obtain also an estimator of $Z_{\mathtt{tr}}(\boldsymbol{\theta})$.
Combining the functional forms in Table 2 and the two alternative estimators, we obtain the Geyer's approach, the standard CL, and several possible sub-methods, that are summarized in Table 3. Note that all the cost functions (representing the different techniques) in Table 3 can be easily and fairly compared since they must be minimized only with respect to $\boldsymbol{\theta}$. The only technique can work in the extended space $(\boldsymbol{\theta}, Z)$ is standard CL (denoted also as CL-1).

Table 2: Functional forms of different possible cost functions. Several procedures can be derived from the same functional form, depending on whether the variable Z is included in the optimization or replaced by an estimate. We have removed constants, useless in term of optimization.

Method	**Variable**	**Cost function**	**Z included in the opt.**	**New**
GA approach Max. Likelihood	$\boldsymbol{\theta}$	$J_{\mathtt{NLL}}(\boldsymbol{\theta}\|Z,\underline{\mathbf{y}}) = -\sum_{n=1}^{N}\log\left[\psi(\mathbf{y}_n\|\boldsymbol{\theta},Z)\right]$	—	✗
Standard CL CL-1	$(\boldsymbol{\theta}, Z)$	$J_{\mathtt{CL\text{-}1}}(\boldsymbol{\theta},Z) = J_{\mathtt{NLL}}(\boldsymbol{\theta}\|Z,\underline{\mathbf{y}}) + \sum_{i=1}^{N+M}\log\left[\mathtt{MIX}(\mathbf{u}_i\|\boldsymbol{\theta},Z)\right]$	✓	✗
CL-2	$(\boldsymbol{\theta}, Z)$	$J_{\mathtt{CL\text{-}2}}(\boldsymbol{\theta},Z) = J_{\mathtt{NLL}}(\boldsymbol{\theta}\|Z,\underline{\mathbf{y}}) - \sum_{i=1}^{N+M}\log\left[\mathtt{MIX}(\mathbf{u}_i\|\boldsymbol{\theta},Z)\right]$	✗	✓
CL-3	$(\boldsymbol{\theta}, Z)$	$J_{\mathtt{CL\text{-}3}}(\boldsymbol{\theta},Z) = J_{\mathtt{NLL}}(\boldsymbol{\theta}\|Z,\underline{\mathbf{y}}) + \sum_{n=1}^{N}\log\left[\mathtt{MIX}(\mathbf{y}_n\|\boldsymbol{\theta},Z)\right]$	✗	✓
Mixture-1	$(\boldsymbol{\theta}, Z)$	$J_{\mathtt{MIX\text{-}1}}(\boldsymbol{\theta},Z) = -\sum_{i=1}^{N+M}\log\left[\mathtt{MIX}(\mathbf{u}_i\|\boldsymbol{\theta},Z)\right]$	✗	✓
Mixture-2	$(\boldsymbol{\theta}, Z)$	$J_{\mathtt{MIX\text{-}2}}(\boldsymbol{\theta},Z) = -\sum_{n=1}^{N}\log\left[\mathtt{MIX}(\mathbf{y}_n\|\boldsymbol{\theta},Z)\right]$	✗	✓

Remark 9. Ideally, if we set $Z(\boldsymbol{\theta}) = Z_{\mathtt{tr}}(\boldsymbol{\theta})$, the cost functions of CL-3, Mixture-2, and the maximum likelihood approach do not depend on the reference samples

Table 3: Techniques using IS estimators of $Z_{\texttt{tr}}(\boldsymbol{\theta})$: recycling the same $\mathbf{x}_m \sim q(\mathbf{x})$ and creating the IS estimators $\widehat{Z}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \frac{\phi(\mathbf{x}_m,\boldsymbol{\theta})}{q(\mathbf{x}_m)}$, or generating other auxiliary samples $\mathbf{s}_m \sim q_2(\mathbf{x})$ and building an independent estimator $\widehat{Z}_S(\boldsymbol{\theta}) = \frac{1}{M_2}\sum_{m=1}^{M_2} \frac{\phi(\mathbf{s}_m,\boldsymbol{\theta})}{q_2(\mathbf{s}_m)}$. With the exception of Standard CL, all the methods provide a an estimation of the full function $Z_{\texttt{tr}}(\boldsymbol{\theta})$. Standard CL gives only the approximation in one point, $Z_{\texttt{tr}}(\boldsymbol{\theta}_{\texttt{tr}})$.

Method		**Variable**	**Cost function**	**Notes**	**Estim. of $Z_{\texttt{tr}}$**
Geyer's approach (GA)		$\boldsymbol{\theta}$	$J_{\texttt{GA}}(\boldsymbol{\theta}) = J_{\texttt{NLL}}(\boldsymbol{\theta}\|\widehat{Z}, \underline{\mathbf{y}})$		all funct. $Z_{\texttt{tr}}(\boldsymbol{\theta})$
Standard CL (CL-1)		$(\boldsymbol{\theta}, Z)$	$J_{\texttt{CL-1}}(\boldsymbol{\theta}, Z)$		only $Z_{\texttt{tr}}(\boldsymbol{\theta}_{\texttt{tr}})$
IS-within-CL-j $j = 1, 2, 3$	IS-Rec-CL-j	$\boldsymbol{\theta}$	$J_{\texttt{CL-j}}(\boldsymbol{\theta}, \widehat{Z})$	recycling $\mathbf{x}_m \sim q(\mathbf{x})$	all funct. $Z_{\texttt{tr}}(\boldsymbol{\theta})$
	IS-CL-j		$J_{\texttt{CL-j}}(\boldsymbol{\theta}, \widehat{Z}_S)$	using $\mathbf{s}_k \sim q_2(\mathbf{x})$	
IS-within-Mixture-j $j = 1, 2$	IS-Rec-Mix-j	$\boldsymbol{\theta}$	$J_{\texttt{MIX-j}}(\boldsymbol{\theta}, \widehat{Z})$	recycling $\mathbf{x}_m \sim q(\mathbf{x})$	all funct. $Z_{\texttt{tr}}(\boldsymbol{\theta})$
	IS-Mix-j		$J_{\texttt{MIX-j}}(\boldsymbol{\theta}, \widehat{Z}_S)$	using $\mathbf{s}_k \sim q_2(\mathbf{x})$	

$\mathbf{x}_1, ..., \mathbf{x}_M$. However, the cost functions of CL-3 and Mixture-2, even in the ideal case $Z(\boldsymbol{\theta}) = Z_{\texttt{tr}}(\boldsymbol{\theta})$, still depend on the proposal/reference density $q(\mathbf{x})$, but evaluated at the observed data $\mathbf{y}_1, ..., \mathbf{y}_N$.

In Table 3, IS-within-CL-1 represents the standard CL but using inside an IS estimator of $Z(\boldsymbol{\theta})$ and estimating only $\boldsymbol{\theta}$. Hence, we have 3 methods using the functional $J_{\texttt{CL-1}}(\boldsymbol{\theta}, Z)$. More generally, considering also CL-2 and CL-3, we have:

- Standard CL (denoted also as CL-1), where we obtain an estimation of $\boldsymbol{\theta}$ and of the value $Z_{\texttt{tr}}(\boldsymbol{\theta}_{\texttt{tr}})$.

- IS-Rec-CL-1, IS-Rec-CL-2 and IS-Rec-CL-3, using the estimator $\widehat{Z}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \frac{\phi(\mathbf{x}_m|\boldsymbol{\theta})}{q(\mathbf{x}_m)}$ built recycling the reference points $\mathbf{x}_m \sim q(\mathbf{x})$.

- IS-CL-1, IS-CL-2 and IS-CL-3, but using the estimator $\widehat{Z}_S(\boldsymbol{\theta}) = \frac{1}{M_2}\sum_{k=1}^{M_2} \frac{\phi(\mathbf{s}_k|\boldsymbol{\theta})}{q_2(\mathbf{s}_k)}$, where $\mathbf{s}_k \sim q_2(\mathbf{x})$, $k = 1, ..., M_2$.

8 Optimal independent proposal density

8.1 Related theory

For a given fixed value of $\boldsymbol{\theta}$, we know that $q_{\texttt{opt}}(\mathbf{x}|\boldsymbol{\theta}) = \psi(\mathbf{x}|\boldsymbol{\theta})$ [20]. However, in our search for $\widehat{\boldsymbol{\theta}}_{\text{ML}}$, e.g., using a gradient descent, we consider several densities $\psi(\mathbf{x}|\boldsymbol{\theta})$ associated to different $\boldsymbol{\theta}$-values. Thus, our goal is to select an independent proposal density $q_{\texttt{opt}}(\mathbf{x})$ that performs adequately for different parameter values (e.g., $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$), and, if possible, for the entire parameter space $\boldsymbol{\Theta}$. For this reason, a possible idea is to consider the following quantity,

$$\bar{Z} = \int_{\boldsymbol{\Theta}} Z_{\texttt{tr}}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \left(\int_{\mathcal{D}} \phi(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \right) d\boldsymbol{\theta}, \quad \text{(assuming } \bar{Z} < \infty\text{)}, \tag{28}$$

and try to minimize the variance of its estimation. Note that we are assuming that $\bar{Z}$ is a finite value. Let consider a finite set $\boldsymbol{\Theta}_{\texttt{sub}} \subset \boldsymbol{\Theta}$ with $|\boldsymbol{\Theta}_{\texttt{sub}}| < \infty$, such that $\widehat{\boldsymbol{\theta}}_{\text{ML}} \in \boldsymbol{\Theta}_{\texttt{sub}}$. Hence, we can draw uniformly in $\boldsymbol{\Theta}_{\texttt{sub}}$, i.e.,

$$\boldsymbol{\theta}_i \sim \mathcal{U}(\boldsymbol{\Theta}_{\texttt{sub}}),$$

where $i = 1, ..., Q$. Then, we can write

$$\bar{Z} \approx \frac{1}{Q} \sum_{i=1}^{Q} \left[\int_{\mathcal{D}} \phi(\mathbf{x}|\boldsymbol{\theta}_i) d\mathbf{x} \right] = \frac{1}{Q} \sum_{i=1}^{Q} Z_{\texttt{tr}}(\boldsymbol{\theta}_i), \tag{29}$$

where $Z_{\texttt{tr}}(\boldsymbol{\theta}_i) = \int_{\mathcal{D}} \phi(\mathbf{x}|\boldsymbol{\theta}_i) d\mathbf{x}$ is a scalar value. It is possible to show that for minimizing the variance of the finite sum $\sum_{i=1}^{Q} Z_{\texttt{tr}}(\boldsymbol{\theta}_i)$ (estimating each $Z_{\texttt{tr}}(\boldsymbol{\theta}_i)$ by IS), we have to use

$$q_{\texttt{opt}}(\mathbf{x}) \propto \sqrt{[\phi(\mathbf{x}|\boldsymbol{\theta}_1)]^2 + [\phi(\mathbf{x}|\boldsymbol{\theta}_2)]^2 + ... + [\phi(\mathbf{x}|\boldsymbol{\theta}_Q)]^2}, \tag{30}$$

as proposal density. For a proof see [20]. The optimal density above is relevant from a theoretical point of view, since generally we cannot evaluate it point-wise and we cannot draw from it. However, these considerations above can drive the construction of an adaptive proposal density (e.g., using variational inference). Namely, this result is useful for designing adaptive schemes.

8.2 Practical use

Generally, a widely-used optimization approach is based on a gradient-descent method. We can consider a set of possible initial nodes $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_Q \sim \mathcal{U}(\boldsymbol{\Theta}_{\texttt{sub}})$. This set can be also used for selecting a good starting points for the gradient descent trying to ensure the convergence to the global minimum. Then, to obtain a suitable set of artificial data, we can apply a IS plus resampling scheme:

1. Draw $M_0 >> M$ samples from a initial proposal density,
$$\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_{M_0} \sim q_0(\mathbf{x}). \tag{31}$$
2. Assign the weights
$$\rho_i = \frac{q_{\texttt{opt}}(\bar{\mathbf{x}}_i)}{q_0(\bar{\mathbf{x}}_i)}, \tag{32}$$
$$= \frac{\sqrt{[\phi(\bar{\mathbf{x}}_i|\boldsymbol{\theta}_1)]^2 + [\phi(\bar{\mathbf{x}}_i|\boldsymbol{\theta}_2)]^2 + ... + [\phi(\bar{\mathbf{x}}_i|\boldsymbol{\theta}_Q)]^2}}{q_0(\bar{\mathbf{x}}_i)}, \tag{33}$$
where $q_{\texttt{opt}}$ is given in Eq. (30).
3. Resample M times within $\{\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_{M_0}\}$ according to the probability mass $\bar{\rho}_i = \frac{\rho_i}{\sum_{k=1}^{M_0} \rho_k}$, $i = 1, ..., M_0$, obtaining a new set $\{\mathbf{x}_1, ..., \mathbf{x}_M\}$, where $\mathbf{x}_k \in \{\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_{M_0}\}$ for all $k = 1, ..., M$.
4. Compute
$$C_{\texttt{opt}} = \frac{1}{M_0} \sum_{k=1}^{M_0} \rho_k.$$
5. Return $\{\mathbf{x}_1, ..., \mathbf{x}_M\}$ or directly
$$\widehat{Z}(\boldsymbol{\theta}) = \frac{C_{\texttt{opt}}}{M} \sum_{m=1}^{M} \frac{\phi(\mathbf{x}_m|\boldsymbol{\theta})}{\sqrt{[\phi(\mathbf{x}_m|\boldsymbol{\theta}_1)]^2 + [\phi(\mathbf{x}_m|\boldsymbol{\theta}_2)]^2 + ... + [\phi(\mathbf{x}_m|\boldsymbol{\theta}_Q)]^2}}. \tag{34}$$

The algorithm above returns $\widehat{Z}(\boldsymbol{\theta})$ that will be used within $J_{\texttt{GA}}(\boldsymbol{\theta})$, for instance, or any other IS-within-CL scheme $J_{\texttt{CL-j}}(\boldsymbol{\theta})$. We can also update the estimation $\widehat{Z}(\boldsymbol{\theta})$ incorporating $\boldsymbol{\theta}$-points selected by the gradient-descent. For instance, let us consider a finite sequence of estimates $\{\boldsymbol{\theta}_{Q+1}, ..., \boldsymbol{\theta}_{Q+T}\}$, obtained starting from the last point $\boldsymbol{\theta}_Q$ drawn from q_0 above, according to the gradient-descent rule

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_t &= \widehat{\boldsymbol{\theta}}_{t-1} - \alpha_{t-1} \nabla_{\boldsymbol{\theta}} J_{\texttt{GA}}(\widehat{\boldsymbol{\theta}}_{t-1}), \qquad (35) \\ &= \widehat{\boldsymbol{\theta}}_{t-1} - \alpha_{t-1} \nabla_{\boldsymbol{\theta}} J_{\texttt{NLL}}(\widehat{\boldsymbol{\theta}}_{t-1} | \widehat{Z}, \underline{\mathbf{y}}), \qquad (36) \end{aligned}$$

with $t = Q+1, ..., Q+T$ and with suitable choices of the step value α_t. Hence, we have the first Q starting points, i.e., $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_Q$ and the new T points, i.e., $\boldsymbol{\theta}_{Q+1}, ..., \boldsymbol{\theta}_{Q+T}$. Then, we can recompute the weights ρ_i and improve $\widehat{Z}(\boldsymbol{\theta})$ before continuing with the gradient descent:

1. Given the samples drawn as in Eq. (31), i.e., $\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_{M_0} \sim q_0(\mathbf{x})$, assign the weights

$$\rho_i = \frac{\sqrt{\sum_{k=1}^{Q+T} \phi(\bar{\mathbf{x}}_i | \boldsymbol{\theta}_k)^2}}{q_0(\bar{\mathbf{x}}_i)}. \qquad (37)$$

2. Resample M times within $\{\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_{M_0}\}$ according to the probability mass $\bar{\rho}_i = \frac{\rho_i}{\sum_{k=1}^{M_0} \rho_k}$, $i = 1, ..., M_0$, obtaining a new set $\{\mathbf{x}_1, ..., \mathbf{x}_M\}$, where $\mathbf{x}_k \in \{\bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_{M_0}\}$ for all $k = 1, ..., M$.

3. Compute $C_{\texttt{opt}} = \frac{1}{M_0} \sum_{k=1}^{M_0} \rho_k$.

4. Return $\{\mathbf{x}_1, ..., \mathbf{x}_M\}$ or directly

$$\widehat{Z}(\boldsymbol{\theta}) = \frac{C_{\texttt{opt}}}{M} \sum_{m=1}^{M} \frac{\phi(\mathbf{x}_m | \boldsymbol{\theta})}{\sqrt{\sum_{k=1}^{Q+T} \phi(\mathbf{x}_m | \boldsymbol{\theta}_k)^2}}. \qquad (38)$$

The gradient descent and the procedure above can be iterated several times until convergence, i.e., $||\widehat{\boldsymbol{\theta}}_t - \widehat{\boldsymbol{\theta}}_{t-1}|| < \epsilon$ with $\epsilon > 0$ (clearly, we expect $\widehat{\boldsymbol{\theta}}_t \approx \widehat{\boldsymbol{\theta}}_{\texttt{ML}}$ if $\epsilon \approx 0$ and the convergence is on a global mode). All the considerations above are valid for any methodology using an IS estimator of $Z_{\texttt{tr}}(\boldsymbol{\theta})$ with an independent proposal density $q(\mathbf{x})$.

9 Numerical experiments

The numerical study in this section focuses on comparing the different schemes, previously described. We consider a simple one-dimensional Gaussian setting with an instructive purpose. First of all, the Gaussian configuration ensures analytical tractability: both the partition function $Z_{\texttt{tr}}$ is known in closed form, allowing theory and simulation to be matched precisely. Moreover, the presentation is clearer and more pedagogically effective. Hence, clearly the goal of the example is not to demonstrate performance on a complex model, but rather to illustrate the performance under controlled conditions and to help the reproducibility.[4]

9.1 Settings

More specifically, we consider a target distribution of type

$$\psi(y|\theta, Z_{\texttt{tr}}) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left(-\frac{y^2}{2\theta^2}\right), \tag{39}$$

$$\phi(y|\theta) = \exp\left(-\frac{y^2}{2\theta^2}\right), \qquad Z_{\texttt{tr}}(\theta) = \sqrt{2\pi\theta^2}, \tag{40}$$

$$E(y|\theta) = \frac{y^2}{2\theta^2}. \tag{41}$$

We also assume that the observed data $y_1, ..., y_N$ are generated from the model above with $\theta_{\texttt{tr}} = 1.5$, i.e.,

$$y_n \sim \psi(y|\theta_{\texttt{tr}}, Z_{\texttt{tr}}) = \frac{1}{Z_{\texttt{tr}}(\theta_{\texttt{tr}})} \exp\left(-\frac{y^2}{2\theta_{\texttt{tr}}^2}\right), \qquad Z_{\texttt{tr}}(\theta_{\texttt{tr}}) = \sqrt{2\pi\theta_{\texttt{tr}}^2}, \tag{42}$$

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

$$x_m \sim q(x) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right), \tag{43}$$

[4] Related Matlab and R code is given at `http://www.lucamartino.altervista.org/PUBLIC_CODE_ISCL.zip`

drawing clearly M artificial data $\{x_1, ..., x_M\}$ from $q(x)$. For simplicity, we keep fixed $\mu_p = 0$. We test the different techniques considering different cases varying N, M, and σ_p. The goal is to estimate the ground-truth value $\theta_{\texttt{tr}} = 1.5$, hence we focus on the parameter estimation problem (i.e., estimating θ). Moreover, we design different scenarios where:

- **Scenario 1:** since the ground-truth is known in the described experiment, we replace $Z(\theta)$ in the different functionals with the true function $Z_{\texttt{tr}}(\theta)$.
- **Scenario 2:** We replace $Z(\theta)$ with $\widehat{Z}(\theta) = \frac{1}{M}\sum_{m=1}^{M} \frac{\phi(x_m,\theta)}{q(x_m)}$ employing the same M points $\{x_1, ..., x_M\}$ used also as reference/noise data.
- **Scenario 3:** We replace $Z(\theta)$ with $\widehat{Z}_S(\theta) = \frac{1}{M_2}\sum_{m=1}^{M_2} \frac{\phi(s_m|\theta)}{q_2(s_m)}$ employing other M_2 points $\{s_1, ..., s_M\}$ generated from another proposal $s_m \sim q_2(x)$, with $m = 1, ..., M_2$. We set $q_2(x) = \mathcal{N}(x|0, 25)$. Note that we could also choose $q_2(x) = q(x)$.

Note that **Scenario 1** is a extreme special case of **Scenario 3** with $M_2 \to \infty$. There would be a "scenario 4" that consists on the joint estimating of $[\widehat{\theta}, \widehat{Z}(\widehat{\theta})]$ where $\widehat{Z}(\widehat{\theta})$ approximates a single value $Z_{\texttt{tr}}(\theta_{\texttt{tr}})$. However, the only procedure able to do that is the standard CL (also denoted as CL-1). For comparing all the techniques, we compute the mean square error (MSE) in the estimation of $\theta_{\texttt{tr}} = 1.5$. Hence, we do not consider the error in estimation of $Z(\theta)$.

We compute the MSE in estimation of $\theta_{\texttt{tr}}$ and, in all scenarios, we keep fixed $\mu_p = 0$ and vary σ_p.

9.2 Results for Scenario 1 – Scenario 3 with $M_2 \to \infty$

In all the techniques, inside the cost functions, we are using the true function $Z_{\texttt{tr}}(\theta)$. This setting can be considered as an extreme case of using $\widehat{Z}_S(\theta)$ with

$$M_2 \to \infty,$$

(or $M \to \infty$ in GA). For this reason, GA coincides perfectly with the maximum likelihood approach and the solution does not depends on σ_p. Since we are using the ground-truth $Z_{\texttt{tr}}(\theta)$ directly in the cost function, this "ideal" GA does not depend on the proposal q. We set $N = 10$ and $M = 10$ for the reference points

$x_1, ..., x_M$ (used for building the corresponding cost functions). The results are shown in Figure 1. We also test and compare standard CL working in the extended bidimensional space (θ, Z). We split the figure into two plots due to the large number of curves. Looking at Figure 1, we can observe the following relevant points:

- The maximum likelihood approach (i.e., the "ideal" GA) is the best method for all the σ_p values, except at $\sigma_p = 1.5$, where IS-CL-3 provides the best results. Very close results have been obtained by IS-CL-2, that is the second best technique close to the "ideal" GA.

- IS-CL-1 and standard CL-1 (considering only the error in θ) provide results far to the best ones.

- IS-CL-3 always outperforms IS-CL-1. IS-CL-2 also outperforms IS-CL-1 and standard CL1.

- IS-CL-3, IS-CL-1 and standard CL-1 perform particularly bad for small values of σ_p. Moreover, IS-CL-3 and IS-CL-1 seems to suffer high values of σ_p.

- The two methods based on fitting the mixture, IS-Mix-1 and IS-Mix-2, do not provide particularly good results. They seems to work better for small values of σ_p but also for great values of σ_p. They suffer in the range of intermediate values of σ_p.

- IS-CL-3 and IS-CL-1 present a best value of σ_p where they seem to reach a minimum MSE.

- Both IS-Mix-j schemes perform well for values of σ_p below $\theta_{\texttt{tr}} = 1.5$ (the standard deviation of the true model). However, their performance deteriorates for intermediate values of σ_p

9.3 Results for Scenario 2

In this framework, we employ the M points $\{x_1, ..., x_M\}$ drawn from $q(x_m)$ for both (a) as 'reference noise data', and (b) building the estimator $\widehat{Z}(\theta) = \frac{1}{M}\sum_{m=1}^{M} \frac{\phi(x_m,\theta)}{q(x_m)}$. Namely, the reference points $\{x_1, ..., x_M\}$ are recycled to build

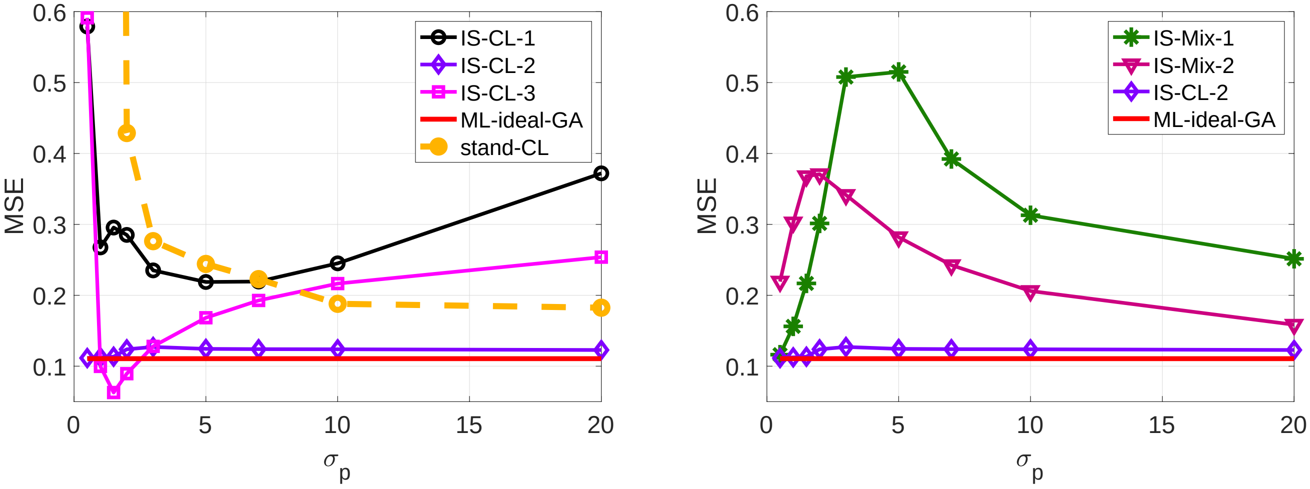

Figure 1: MSE versus the standard deviation σ_p of the proposal density in Scenario 1, with $N = 10$ and $M = 10$. We split the figure into two plots due to the large number of curves. The considered setting is particularly challenging for mixture-based cost functions since the two pdfs (target and proposal) have the same locations, $\mu_p = 0$.

$\widehat{Z}(\theta)$. The results are given in Figure 2. We depict only the curves corresponding to IS-Rec-CL-1, IS-Rec-CL-2, GA, and the standard CL. The curve for IS-Rec-CL-3 is not shown because it is completely overlapped by those of IS-Rec-CL-1, IS-Rec-CL-2 and GA, while the mixture schemes yield substantially higher MSE values. Hence, IS-Rec-CL-1, IS-Rec-CL-2, IS-Rec-CL-3 and GA provide very similar (good) results, generally outperforming the standard CL except for the value $\sigma_p = 3$. Therefore, this experiment suggests that there is no advantage in optimizing over the higher-dimensional space (θ, Z), that is, in using the standard CL. However, these results have been obtained with high value of $M = 5000$ that allows a good estimation of $Z(\theta)$. It is possible to show that, with small values of M, the estimation of $Z(\theta)$ provided by $\widehat{Z}(\theta)$ is poor and the standard CL provides better results.

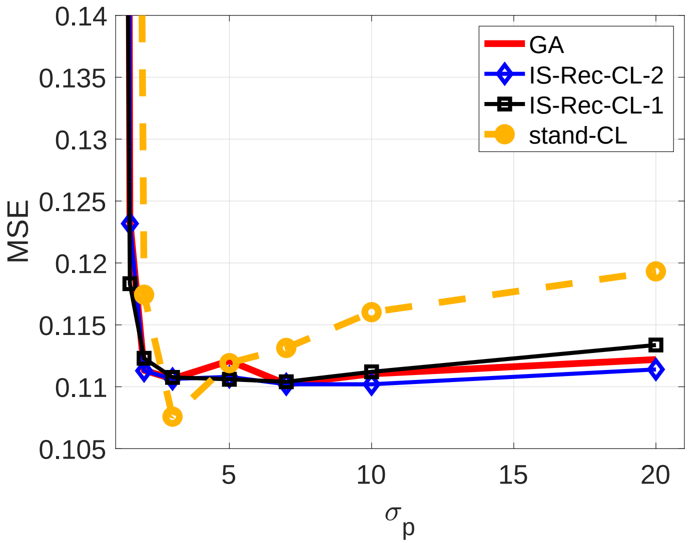

Figure 2: MSE versus the standard deviation σ_p, obtained with some recycling schemes and the standard CL with $N = 10$ and $M = 5000$. We report only the curves corresponding to IS-Rec-CL-1, IS-Rec-CL-2, GA, and the standard CL. The curve for IS-Rec-CL-3 is not shown because it is completely overlapped by those of IS-Rec-CL-1 and IS-Rec-CL-2, while the mixture schemes yield substantially higher MSE values.

9.4 Summary of the results

Focusing mainly on the inference of $\boldsymbol{\theta}$, we can summarize the main conclusions as follows:

- Without computational restrictions on the evaluation of the numerator $\phi(\mathbf{x}, \boldsymbol{\theta})$, the estimation $Z(\boldsymbol{\theta})$ with an IS estimator seems to be a better strategy (using $M_2 \rightarrow \infty$ samples), and then apply the GA or IS-within-CL-2 scheme.

- Considering the cost of evaluating $\phi(\mathbf{x}, \boldsymbol{\theta})$, we should compare with the *IS-recycling schemes* (IS-Rec-CL-j). from the results:

  – the standard CL is preferable for small values of M,

  – while the IS-Rec-CL-j methods, with $j = 1, 2, 3$, are preferable for greater values of M.

Moreover, focusing only on $\boldsymbol{\theta}$, the IS-within-CL-2 scheme seems to work better than standard CL. For some specific values of the proposal parameters, the IS-within-CL-3 schemes could outperform even the "ideal" GA (i.e., the classical maximum likelihood approach). More generally, the IS-RC-CL-j techniques seem to be more robust than Stand-CL when the scale parameter of the proposal is smaller of target density (i.e., $\sigma_p < \sigma_{\texttt{true}}$).

Regarding $Z(\boldsymbol{\theta})$, we can remark that:

- With GA and the IS-RC-CL-j schemes, we obtain the complete approximation $Z(\boldsymbol{\theta})$, i.e., $\widehat{Z}(\boldsymbol{\theta})$. With the standard CL, we only obtain an approximation of $Z_{\texttt{tr}}(\boldsymbol{\theta}_{\texttt{tr}})$.

Once the entire partition function $Z(\boldsymbol{\theta})$ has been approximated by $\widehat{Z}(\boldsymbol{\theta})$, Bayesian inference can be carried out straightforwardly. Consequently, the GA and IS-within-CL-j schemes can be employed for a *pre-Bayesian analysis*. Moreover, the pointwise estimator $\widehat{\boldsymbol{\theta}}$ can also be useful for designing an efficient Bayesian analysis. Finally recall that the use of the optimal independent proposal density $q_{\texttt{opt}}(\mathbf{x})$ can improve the performance of the IS-RC-CL-j schemes.

10 Conclusions

In this work, we have provided a comprehensive analysis of Geyer's approach and contrastive learning (CL) for inference in non-normalized models, clarifying their theoretical connections and practical differences. By adopting a complementary perspective to existing asymptotic comparisons [1], we have shown that meaningful comparisons between GA and CL can be carried out without increasing the dimensionality of Geyer's method. Instead, by simplifying the CL framework through the use of importance sampling estimators of the partition function, we have reduced the optimization problem to the original parameter space, leading to the proposed IS-within-CL schemes.

The theoretical comparison between the resulting cost functions not only sheds light on the relationship between GA and CL, but also naturally motivates the design of novel variants that interpolate between these methodologies. These schemes retain the interpretability and computational advantages of lower-dimensional optimization while benefiting from the flexibility of contrastive formulations. Our experimental study demonstrates that the proposed approaches are competitive and, in several scenarios (for instance, with an high value of the artificial data M and small number of true data N), advantageous compared to the standard CL methods. Additionally, an important outcome of IS-within-CL-j or GA schemes is that these methodologies yield an explicit approximation of the entire partition function $Z_{\texttt{tr}}(\boldsymbol{\theta})$. This can be used as a pre-Bayesian analysis that offers $\widehat{Z}(\boldsymbol{\theta})$ and the information about high-probability regions of the parameter space, facilitating the construction of efficient Bayesian inference procedures. Last but not least, we have introduced the optimal independent proposal density for both GA and IS-within-CL schemes and provided an algorithm for its practical implementation.

Acknowledgements

This work has been partially supported by the PIACERI Starting Grant BA-GRAPH (UPB 28722052144) and PIACERI project LikeFree-BA-GRAPH (UPB 28722052159) of the University of Catania.

References

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

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[9] H. O. Georgii, *Gibbs Measures and Phase Transitions.* De Gruyter, 2011.

[10] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications.* American Mathematical Society, 1980.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

[17] ——, “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.

[18] 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.

[19] 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.

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

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

[22] ——, “Some extensions of score matching,” *Computational Statistics & Data Analysis*, vol. 51, no. 5, pp. 2499–2512, 2007.

[23] 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.

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

[25] A. Hyvärinen, “Consistency of pseudolikelihood estimation of fully visible boltzmann machines,” *Neural Computation*, vol. 18, no. 10, pp. 2283–2292, 2006.

[26] 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.

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

[28] 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.

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

[30] L. Martino, V. Elvira, D. Luengo, and J. Corander, “Layered adaptive importance sampling,” *Statistics and Computing*, vol. 27, no. 3, pp. 599–623, 2017.

A Direct approximation of the gradient

Let us consider the gradient of the cost function $J_{\texttt{GA}}(\boldsymbol{\theta})$ scaled by a factor $1/N$, i.e.,

$$\frac{1}{N}\nabla_{\boldsymbol{\theta}} J_{\texttt{GA}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}} E(\mathbf{y}_n|\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\widehat{Z}(\boldsymbol{\theta}).$$

Recalling that $\nabla_{\boldsymbol{\theta}}\log\widehat{Z}(\boldsymbol{\theta}) = \frac{1}{\widehat{Z}(\boldsymbol{\theta})}\nabla_{\boldsymbol{\theta}}\widehat{Z}(\boldsymbol{\theta})$, the equation above becomes:

$$\begin{aligned}\frac{1}{N}\nabla_{\boldsymbol{\theta}} J_{\texttt{GA}}(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}} E(\mathbf{y}_n|\boldsymbol{\theta}) + \frac{1}{\widehat{Z}(\boldsymbol{\theta})}\nabla_{\boldsymbol{\theta}}\widehat{Z}(\boldsymbol{\theta}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}} E(\mathbf{y}_n|\boldsymbol{\theta}) + \frac{1}{\frac{1}{M}\sum_{j=1}^{M}\frac{\phi(\mathbf{x}_j|\boldsymbol{\theta})}{q(\mathbf{x}_j)}}\frac{1}{M}\sum_{m=1}^{M}\frac{\nabla_{\boldsymbol{\theta}}\phi(\mathbf{x}_m|\boldsymbol{\theta})}{q(\mathbf{x}_m)}.\end{aligned}$$

Moreover, since $\phi(\mathbf{y}|\boldsymbol{\theta}) = e^{-E(\mathbf{y}|\boldsymbol{\theta})}$, we have

$$\nabla_{\boldsymbol{\theta}}\phi(\mathbf{y}|\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} E(\mathbf{y}|\boldsymbol{\theta})e^{-E(\mathbf{y}|\boldsymbol{\theta})} = -\nabla_{\boldsymbol{\theta}} E(\mathbf{y}|\boldsymbol{\theta})\phi(\mathbf{y}|\boldsymbol{\theta}).$$

Replacing above, we obtain

$$\begin{aligned}\frac{1}{N}\nabla_{\boldsymbol{\theta}} J_{\texttt{GA}}(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}} E(\mathbf{y}_n|\boldsymbol{\theta}) - \frac{1}{\sum_{j=1}^{M}\frac{\phi(\mathbf{x}_j|\boldsymbol{\theta})}{q(\mathbf{x}_j)}}\sum_{m=1}^{M}\frac{\nabla_{\boldsymbol{\theta}} E(\mathbf{x}_m|\boldsymbol{\theta})\phi(\mathbf{x}_m|\boldsymbol{\theta})}{q(\mathbf{x}_m)} \\ &= \frac{1}{N}\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}} E(\mathbf{y}_n|\boldsymbol{\theta}) - \sum_{m=1}^{M}\bar{w}_m^{(\boldsymbol{\theta})}\nabla_{\boldsymbol{\theta}} E(\mathbf{x}_m|\boldsymbol{\theta}), \qquad (44)\end{aligned}$$

where we have defined the normalized IS weights as:

$$\bar{w}_m^{(\boldsymbol{\theta})} = \frac{\frac{\phi(\mathbf{x}_m|\boldsymbol{\theta})}{q(\mathbf{x}_m)}}{\sum_{j=1}^{M} \frac{\phi(\mathbf{x}_j|\boldsymbol{\theta})}{q(\mathbf{x}_j)}} = \frac{w_m^{(\boldsymbol{\theta})}}{\sum_{j=1}^{M} w_j^{(\boldsymbol{\theta})}}, \tag{45}$$

and $w_m^{(\boldsymbol{\theta})} = \frac{\phi(\mathbf{x}_m|\boldsymbol{\theta})}{q(\mathbf{x}_m)}$ are the unnormalized IS weights. If we are able to draw samples from the model, i.e., $\mathbf{x}_m^{(\boldsymbol{\theta})} \sim \psi(\mathbf{y}|\boldsymbol{\theta})$, the normalized IS weights become $\bar{w}_m^{(\boldsymbol{\theta})} = \frac{1}{M}$ for each m, and Eq. (44) can be rewritten as

$$\frac{1}{N}\nabla_{\boldsymbol{\theta}} J_{\mathtt{GA}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} E(\mathbf{y}_n|\boldsymbol{\theta}) - \frac{1}{M}\sum_{m=1}^{M} \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_m^{(\boldsymbol{\theta})}|\boldsymbol{\theta}), \qquad \mathbf{x}_m^{(\boldsymbol{\theta})} \sim p(\cdot|\boldsymbol{\theta}). \tag{46}$$

However, recall that we need to generate of artificial data $\mathbf{x}_m^{(\boldsymbol{\theta})}$ for every $\boldsymbol{\theta}$-value, hence the total number of evaluation of the energy function grows.