

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

L. Martino^{*†}, L. Scaffidi-Domianello[†], S. Mangano[†]

[†] Università degli Studi di Catania, Italy.

^{*} Corresponding author: luca.martino@unict.it

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 [34], 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 replicable experimental studies. The described IS-within-CL schemes yield an entire approximation of the partition function, so enabling a possible efficient classical 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, Python and R codes are also made freely available to help the reproducibility of the results.

1 Introduction

Energy-based models define probability distributions only up to an unknown normalizing function $Z_{\text{tr}}(\boldsymbol{\theta})$, commonly referred to as the partition function, where the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{d_{\boldsymbol{\theta}}}$ is the object of interest for inference. For this reason, they are also commonly referred to as non-normalized models. In this framework, the model is defined as $\psi(\mathbf{y}|\boldsymbol{\theta}, Z_{\text{tr}}) = \frac{\phi(\mathbf{y}|\boldsymbol{\theta})}{Z_{\text{tr}}(\boldsymbol{\theta})}$ where $\mathbf{y}$ are the observed data [7, 5, 22, 29, 36]. 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 [1, 9, 10, 20, 26]. In Bayesian inference, these models generate the so-called doubly intractable posteriors [2, 23, 31, 33]. 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 [11, 12, 13]. The main idea is to use importance sampling (IS) estimators to approximate $Z(\boldsymbol{\theta})$ [25, 24]. 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 $\phi(\mathbf{y}|\boldsymbol{\theta})$ [17, 19]. A third class includes the *contrastive learning* (CL) approach, more precisely noise-contrastive estimation (NCE)¹, which reformulates parameter estimation as a classification or discrimination problem between observed data and artificially generated samples [14, 15]. 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 [1, 18]. The CL approach has received considerable attention and has been extensively studied in the literature from different perspectives and in a wide range of applications [30, 34, 35], including in very recent works [27, 35].

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 [34] 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 Θ , whereas CL operates in the higher-dimensional space $\Theta \times \mathbb{R}$, 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 $\Theta \times \mathbb{R}$. 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 Θ (lower-dimensional with respect to $\Theta \times \mathbb{R}$). We refer to them as IS-within-CL schemes. These novel methodologies differ from GA in the functional form of their employed cost functions. The design of these novel cost functions is motivated by the theoretical comparison between Geyer’s and CL cost functions previously analyzed. Furthermore, we theoretically derive the optimal proposal density for GA, which also performs particularly well across all the IS-within-CL schemes, as proved by the numerical experiments. We also develop an algorithm for its practical implementation. Throughout the considerations and comparisons presented in this work, we mainly focus on the estimation of the true parameter vector $\boldsymbol{\theta}_{\text{tr}}$. For an exhaustive analysis of the performance of standard CL in estimating the normalizing constant $Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$ (i.e., the particular value of the partition function at $\boldsymbol{\theta}_{\text{tr}}$), see [27].

We discuss the potential benefits and compare all the proposed approaches through a replicable experimental study. The associated MATLAB, Python and R code is provided to facilitate the

¹Throughout the remainder of this work, the term contrastive learning (CL) refers specifically to noise-contrastive estimation (NCE), even when NCE is not explicitly mentioned.

reproducibility of the results.² Both GA and the IS-within-CL schemes yield an approximation $\hat{Z}(\boldsymbol{\theta})$ of the entire partition function $Z_{\text{tr}}(\boldsymbol{\theta})$. 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 $Z_{\text{tr}}(\boldsymbol{\theta})$. This possibility is specifically described in Section 5.3.

The rest of the work is structured as follows. After recalling the necessary background in Section 2, the main contributions of this work are given in Sections 3, 4, and 5. In Section 3, we first establish the relevant connections between GA and CL, which provide the basis for introducing the proposed schemes in Section 4. Section 5 is devoted to the derivation and description of the optimal proposal. Several numerical experiments and additional tests are presented in Section 6, while the final discussion and conclusions are provided in Section 7.

2 Problem statement and two benchmark solutions

2.1 Energy-based models (EBMs)

Let $\phi(\mathbf{y}|\boldsymbol{\theta}) = e^{-E(\mathbf{y}|\boldsymbol{\theta})} \geq 0$ be a function parametrized by a vector of parameters $\boldsymbol{\theta}$ taking values in $\Theta \subseteq \mathbb{R}^{d_\theta}$. The non-negative function $E(\mathbf{y}|\boldsymbol{\theta}) = -\log \phi(\mathbf{y}|\boldsymbol{\theta})$ 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 $\psi(\mathbf{y}|\boldsymbol{\theta}, Z_{\text{tr}})$, defined for each $\boldsymbol{\theta}$ as

$$\psi(\mathbf{y}|\boldsymbol{\theta}, Z_{\text{tr}}) = \frac{\phi(\mathbf{y}|\boldsymbol{\theta})}{Z_{\text{tr}}(\boldsymbol{\theta})}, \quad \mathbf{y} \in \mathcal{D} \subseteq \mathbb{R}^d. \quad (1)$$

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

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

Namely, $Z_{\text{tr}}(\boldsymbol{\theta})$ is unknown because the integral above cannot be solved analytically in closed form, i.e., is intractable.³ The subindex **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 $Z_{\text{tr}}(\boldsymbol{\theta})$, 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 $\underline{\mathbf{y}} = \mathbf{y}_{1:N} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\} \in \mathcal{D}^N$. Each vector $\mathbf{y}_n$ may contain different types of data, including discrete, continuous, and/or mixed-type components. The observed dataset $\underline{\mathbf{y}}$ contains i.i.d. realizations distributed as the EBM in Eq. (1) for a specific unknown vector of parameters $\boldsymbol{\theta}_{\text{tr}}$ (true vector of parameters), i.e.,

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

²Related code is given at http://www.lucamartino.altervista.org/PUBLIC_CODE_ISCL.zip

³We assume that $\mathbf{y}$ is a continuous vector, although several of the considerations also apply to the discrete case, as illustrated in the second numerical experiment. When the observed data are discrete, all integrals must be replaced by sums.

Note that $Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$ is a scalar normalizing constant, i.e., the true partition function evaluated at $\boldsymbol{\theta}_{\text{tr}}$. As a consequence, $\psi(\mathbf{y}|\boldsymbol{\theta}_{\text{tr}}, Z_{\text{tr}})$ 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.

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

Notation	Description
$\boldsymbol{\theta}_{\text{tr}}$	True vector of parameters
$Z_{\text{tr}}(\boldsymbol{\theta}) = \int_{\mathcal{D}} \phi(\mathbf{y} \boldsymbol{\theta}) d\mathbf{y}$	True partition function
$\psi(\mathbf{y} \boldsymbol{\theta}_{\text{tr}}, Z_{\text{tr}}) = \frac{\phi(\mathbf{y} \boldsymbol{\theta}_{\text{tr}})}{Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})}$	Generating EBM (true model that generates the observed data $\mathbf{y}_{1:N}$)
$\mathbf{y}_n \sim \psi(\mathbf{y} \boldsymbol{\theta}_{\text{tr}}, Z_{\text{tr}})$	n -th observed data vector
$Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$	Normalizing constant of the true model
$\psi(\mathbf{y} \boldsymbol{\theta}, Z_{\text{tr}}) = \frac{\phi(\mathbf{y} \boldsymbol{\theta})}{Z_{\text{tr}}(\boldsymbol{\theta})}$	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	
$q(\mathbf{x})$ $\mathbf{x}_m \sim q(\mathbf{x})$	Proposal density Auxiliary data
CL	
	Reference density Reference data
$\underline{\mathbf{y}} = \mathbf{y}_{1:N} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$	Set of true observed data
$\underline{\mathbf{x}} = \mathbf{x}_{1:M} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$	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

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

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

and the corresponding log-likelihood is

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

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

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

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

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

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

2.4 Geyer's scheme based on IS estimators

2.4.1 IS estimators for $Z_{\text{tr}}(\boldsymbol{\theta})$

Theoretically speaking, the simplest approaches rely on applications of some numerical method for computing $Z_{\text{tr}}(\boldsymbol{\theta})$. One way to obtain this approximation is based on importance sampling (IS) [24]. 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 does not depend on $\boldsymbol{\theta}$. Let $\underline{\mathbf{x}} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$, be a sample of size M generated from $q(\mathbf{x})$, i.e.,

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

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* [12]. In this case, the IS estimator is given by

$$\hat{Z}(\boldsymbol{\theta}) = \frac{1}{M} \sum_{m=1}^M \frac{\phi(\mathbf{x}_m|\boldsymbol{\theta})}{q(\mathbf{x}_m)}, \quad \mathbf{x}_m \sim q(\mathbf{x}), \quad (8)$$

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

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

where the proposal depends on $\boldsymbol{\theta}$, and a different set $\{\mathbf{x}_1^{(\boldsymbol{\theta})}, \dots, \mathbf{x}_M^{(\boldsymbol{\theta})}\}$ is generated for each value of $\boldsymbol{\theta}$. In this scenario, the optimal proposal is $q(\mathbf{x}|\boldsymbol{\theta}) \propto \phi(\mathbf{x}|\boldsymbol{\theta})$ [24]. Hence, for a fixed $\boldsymbol{\theta}$, the estimator $\widehat{Z}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ 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 [24]. Nevertheless, $\widehat{Z}(\boldsymbol{\theta})$ provides an approximation that is valid for all $\boldsymbol{\theta}$. Moreover, note that once the known part $\phi(\cdot)$, the proposal $q(\cdot)$, and the auxiliary samples $\mathbf{x}_m$ are fixed, $\widehat{Z}(\boldsymbol{\theta})$ becomes a deterministic, analytically known, and easily evaluable function of $\boldsymbol{\theta}$.

2.4.2 Geyer's approach (GA)

In [12], the first IS approximation (8) is substituted directly into Eq. (7):

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

By IS arguments, we have $J_{\text{GA}}(\boldsymbol{\theta}) \approx J_{\text{NLL}}(\boldsymbol{\theta}|Z_{\text{tr}}, \underline{\mathbf{y}})$. The idea is then to minimize this function [12], i.e., $\widehat{\boldsymbol{\theta}}_{\text{GA}} = \arg \min J_{\text{GA}}(\boldsymbol{\theta})$. Clearly, as $M \rightarrow \infty$,

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

Furthermore, an approximation of the gradient (scaled by $\frac{1}{N}$) can be also directly computed as:

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

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

Remark 3. In Geyer's approach, the same set of auxiliary samples $\underline{\mathbf{x}} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\} \sim q(\mathbf{x})$ 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 $\boldsymbol{\theta}$ [24]. There are two main strategies to reduce the variance of $J_{\text{GA}}(\boldsymbol{\theta})$: (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.

2.5 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

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

and

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

i.e., $p(\mathbf{u}|a = 1) = \psi(\mathbf{u}|\boldsymbol{\theta}_{\text{tr}}, Z_{\text{tr}})$ 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

$$\underbrace{\mathbf{u}_1 = \mathbf{y}_1, \dots, \mathbf{u}_N = \mathbf{y}_N}_{a=1}, \underbrace{\mathbf{u}_{N+1} = \mathbf{x}_1, \dots, \mathbf{u}_{N+M} = \mathbf{x}_M}_{a=0}. \quad (13)$$

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 $\mathbf{x}_1, \dots, \mathbf{x}_M$ 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 vector $\boldsymbol{\theta}_{\text{tr}}$ and the scalar value $Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$, 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 $\boldsymbol{\xi} = [\boldsymbol{\theta}^\top, Z]^\top$, the posterior probabilities are

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

$$= \frac{\phi(\mathbf{u}, \boldsymbol{\theta})}{\phi(\mathbf{u}, \boldsymbol{\theta}) + \nu Z(\boldsymbol{\theta})q(\mathbf{u})}, \quad (15)$$

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

Remark 4. Note that we are considering a generic vector $\boldsymbol{\theta}$ and a generic value of $Z(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta}$.

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

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

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 for obtaining a good approximation $\hat{\eta}(\mathbf{u})$ consists of minimizing an empirical risk:

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

where $V(\eta) : [0, 1] \rightarrow [0, 1]$ is a given *non-increasing function* in $[0, 1]$ (called *loss function*).⁴ Therefore, for each possible input $\mathbf{u}$, the CL procedure is given by the following formula:

$$\hat{\boldsymbol{\xi}} = \arg \min_{\boldsymbol{\xi}} J_{\text{CL}}(\boldsymbol{\xi}), \quad \text{and set} \quad \hat{\eta}(\mathbf{u}) = \eta(\mathbf{u}, \hat{\boldsymbol{\xi}}), \quad \forall \mathbf{u}. \quad (18)$$

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

$$V(\eta) = -\log(\eta). \quad (19)$$

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

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

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

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

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

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

$$[\hat{\boldsymbol{\theta}}_{\text{CL}}, \hat{Z}_{\text{CL}}] = \arg \min J_{\text{CL}}(\boldsymbol{\theta}, Z), \quad (23)$$

where $\hat{\boldsymbol{\theta}}_{\text{CL}} \rightarrow \boldsymbol{\theta}_{\text{tr}}$ and

$$\hat{Z}_{\text{CL}} \rightarrow Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}}), \quad (24)$$

is a scalar value, that is the approximation of Z_{tr} in one specific point, $\boldsymbol{\theta}_{\text{tr}}$. 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 of the reference density in CL, see [3, 4].

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

⁴The 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).

3 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_{GA} and J_{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 [24]; whereas, regarding the optimal reference density, see [3, 4].
4. The cost function J_{CL} takes values in a higher-dimensional space, i.e., $J_{\text{CL}} : \mathbb{R}^{d_\theta+1} \rightarrow \mathbb{R}$, whereas $J_{\text{GA}} : \mathbb{R}^{d_\theta} \rightarrow \mathbb{R}$. In other words, J_{CL} depends on one additional input, so its support domain has one higher dimension than that of J_{GA} . This difference arises because, in the GA scheme, $Z(\boldsymbol{\theta})$ is pre-approximated using importance sampling : this can be interpreted as a preliminary step that may affect the performance of the subsequent minimization. Furthermore, in standard CL, the scalar quantity $Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$ is treated as an additional parameter to be estimated, leading to the simultaneous estimation of $\boldsymbol{\theta}_{\text{tr}}$ and $Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$.
5. Using GA, we obtain an approximation of the full partition function $Z_{\text{tr}}(\boldsymbol{\theta})$, whereas with CL we only obtain a point-wise approximation of the specific value $Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$: this represents just a value, the function form remains unknown.

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

$$\begin{aligned}
 J_{\text{CL}}(\boldsymbol{\theta}, Z) &= - \sum_{n=1}^N \log \left[\frac{N\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z)}{N\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z) + Mq(\mathbf{y}_n)} \right] - \sum_{m=1}^M \log \left[\frac{Mq(\mathbf{x}_m)}{N\psi(\mathbf{x}_m|\boldsymbol{\theta}, Z) + Mq(\mathbf{x}_m)} \right], \\
 &= - \sum_{n=1}^N \log \left[\frac{\frac{N}{N+M}\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z)}{\frac{N}{N+M}\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z) + \frac{M}{N+M}q(\mathbf{y}_n)} \right] - \sum_{m=1}^M \log \left[\frac{\frac{M}{N+M}q(\mathbf{x}_m)}{\frac{N}{N+M}\psi(\mathbf{x}_m|\boldsymbol{\theta}, Z) + \frac{M}{N+M}q(\mathbf{x}_m)} \right], \\
 &= - \sum_{n=1}^N \log \left[\frac{\alpha_1\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z)}{\alpha_1\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z) + \alpha_2q(\mathbf{y}_n)} \right] - \sum_{m=1}^M \log \left[\frac{\alpha_2q(\mathbf{x}_m)}{\alpha_1\psi(\mathbf{x}_m|\boldsymbol{\theta}, Z) + \alpha_2q(\mathbf{x}_m)} \right],
 \end{aligned}$$

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

$$\begin{aligned}
 J_{\text{CL}}(\boldsymbol{\theta}, Z) &= - \sum_{n=1}^N \log \left[\frac{\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z)}{\alpha_1\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z) + \alpha_2q(\mathbf{y}_n)} \right] - \sum_{m=1}^M \log \left[\frac{q(\mathbf{x}_m)}{\alpha_1\psi(\mathbf{x}_m|\boldsymbol{\theta}, Z) + \alpha_2q(\mathbf{x}_m)} \right] - N \log \alpha_1 - M \log \alpha_2, \\
 &= - \sum_{n=1}^N \log \left[\frac{\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z)}{\alpha_1\psi(\mathbf{y}_n|\boldsymbol{\theta}, Z) + \alpha_2q(\mathbf{y}_n)} \right] - \sum_{m=1}^M \log \left[\frac{q(\mathbf{x}_m)}{\alpha_1\psi(\mathbf{x}_m|\boldsymbol{\theta}, Z) + \alpha_2q(\mathbf{x}_m)} \right] + C_0,
 \end{aligned}$$

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

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

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

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

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

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

Note that, with the second term, $\sum_{i=1}^{N+M} \log [\text{MIX}(\mathbf{u}_i|\boldsymbol{\theta}, Z)]$, we are applying the deterministic mixture idea [32, 8, 28].

Remark 8. Let us focus on the inference of $\boldsymbol{\theta}_{\text{tr}}$, keeping the other argument, Z , of the function $J_{\text{CL}}(\boldsymbol{\theta}, Z)$ fixed for the purposes of this discussion. The first term, $J_{\text{NLL}}(\boldsymbol{\theta}|Z, \underline{\mathbf{y}})$, corresponds to the negative log-likelihood in Eq. (7), which must be minimized to estimate $\boldsymbol{\theta}_{\text{tr}}$. The second term, instead, is the (*positive*) log-likelihood associated with fitting the mixture $\text{MIX}(\mathbf{u}|\boldsymbol{\theta}, Z) = \alpha_1 \psi(\mathbf{u}|\boldsymbol{\theta}, Z) + \alpha_2 q(\mathbf{u})$ to the data $\mathbf{u}_1, \dots, \mathbf{u}_{M+N}$; in this case, *the maximization* provides an estimator of $\boldsymbol{\theta}_{\text{tr}}$. Thus, assuming that $Z_{\text{tr}}(\boldsymbol{\theta})$ is known, both terms separately yield consistent estimators of $\boldsymbol{\theta}_{\text{tr}}$, 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 $\boldsymbol{\theta}_{\text{tr}}$, as shown in literature [14, 15].

4 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_{GA} and J_{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 θ_{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. Indeed, in a first scenario, (a) Z can be treated as a variable to optimize (which is the case of the standard CL approach) or (b) Z can be replaced by an approximation. Considering this second scenario, for instance, we can include IS estimators of $Z_{\text{tr}}(\theta)$ and use the cost functions only to find an estimator of θ_{tr} , exactly as in the GA. Moreover, in this case, we have two additional possible approaches:

1. **Schemes based on a single proposal.** Use the same samples $\mathbf{x}_m$ as artificial/auxiliary and reference data, hence a unique $q(\cdot)$ playing both roles: proposal and reference densities. Thus, the estimator of $Z_{\text{tr}}(\theta)$ is

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

We refer to these techniques - based on a single proposal - as *recycling* schemes, since the same samples $\mathbf{x}_m$ are employed for two different purposes: as reference data and for building the IS estimator of Z_{tr} .

2. **Schemes based on two proposals.** Alternatively, we can use a proposal density $q_1(\mathbf{x})$ for generating the reference samples, i.e., $\mathbf{x}_1, \dots, \mathbf{x}_M \sim q_1(\mathbf{x})$, and 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 $Z_{\text{tr}}(\theta)$, i.e.,

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

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

Computational cost. The computational cost, in terms of evaluations of the non-normalized model $\phi(\cdot|\theta)$, is clearly different between the two families of methods. 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 $\hat{Z}_S$. Note that the first case can be recovered as a special case of the second one, setting $M_2 = M$ and $\mathbf{s}_m = \mathbf{x}_m$. In this sense, we can assert that, in the first scenario, we are *recycling* the samples $\mathbf{x}_m$ to obtain also an estimator of $Z_{\text{tr}}(\theta)$. Then, since we must evaluate the function $\phi(\cdot|\theta)$ at these samples, the schemes based on two proposals are clearly more costly. With respect to the analytic form of the cost functions (and hence considering Z simply as a parameter/variable), CL-1, CL-2, CL-3, Mixture-1, and Mixture-2 in Table 2 differ in the presence or absence of specific

terms in their respective sums. In this sense, Mixture-2 has the smallest number of terms, namely N , since it depends only on the observed data $\underline{\mathbf{y}}$ and is independent of the samples $\mathbf{x}_m$ used for constructing the cost function. It is therefore the least costly scheme in terms of the number of terms involved in the cost function. CL-1 and CL-2 both contain $2N + M$ terms, whereas CL-3 contains $2N$ terms and is also independent of $\mathbf{x}_m$. Finally, Mixture-1 contains $N + M$ terms in its sum.

Combining the functional forms in Table 2 and the two alternative possibilities described above, 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 that 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 a preliminary estimate. We have removed constants, useless in term of optimization.

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

Remark 9. Ideally, if we set $Z(\boldsymbol{\theta}) = Z_{\text{tr}}(\boldsymbol{\theta})$, the cost functions of GA, CL-3, and Mixture-2 do not depend on the reference samples $\mathbf{x}_1, \dots, \mathbf{x}_M$ (e.g., GA becomes the maximum likelihood approach). However, the cost functions of CL-3 and Mixture-2, even in the ideal case $Z(\boldsymbol{\theta}) = Z_{\text{tr}}(\boldsymbol{\theta})$, still depend on the functional of the proposal/reference density $q(\mathbf{x})$, evaluated at the observed data $\mathbf{y}_1, \dots, \mathbf{y}_N$.

Remark 10. When an estimation $\hat{Z}(\boldsymbol{\theta})$, obtained using a set of reference samples $\mathbf{x}_1, \dots, \mathbf{x}_M$, is employed for replacing $Z_{\text{tr}}(\boldsymbol{\theta})$, GA, CL-3 and Mixture-2 depend on these reference samples, as

Table 3: Techniques using IS estimators of $Z_{\text{tr}}(\boldsymbol{\theta})$: single proposal-based schemes, recycling the same $\mathbf{x}_m \sim q(\mathbf{x})$ and creating the IS estimators $\hat{Z}(\boldsymbol{\theta}) = \frac{1}{M} \sum_{m=1}^M \frac{\phi(\mathbf{x}_m, \boldsymbol{\theta})}{q(\mathbf{x}_m)}$; or, two proposals-based schemes, where the samples $\mathbf{x}_m \sim q_1(\mathbf{x})$ are used for building the cost function and other auxiliary samples $\mathbf{s}_m \sim q_2(\mathbf{x})$ are drawn for building an independent estimator $\hat{Z}_S(\boldsymbol{\theta}) = \frac{1}{M_2} \sum_{m=1}^{M_2} \frac{\phi(\mathbf{s}_m, \boldsymbol{\theta})}{q_2(\mathbf{s}_m)}$. With the exception of Standard CL (CL-1), all the methods provide an estimation of the full function $Z_{\text{tr}}(\boldsymbol{\theta})$. Standard CL returns only the approximation in one point, $Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$.

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

well. Consequently, the resulting objective functions are not determined solely by the observed data $\underline{\mathbf{y}}$, but also by the particular reference samples used to construct $\hat{Z}(\boldsymbol{\theta})$.

In Table 3, the IS-within-CL schemes employ an IS estimator $\hat{Z}(\boldsymbol{\theta})$. They can be divided in two sub-sets of methods: IS-Rec-CL that use a single proposal q , and IS-CL that use two proposals q_1 and q_2 . More precisely, considering all the CL-j schemes, with $j = 1, 2, 3$, we have:

- Standard CL, simply denoted as CL-1, where we obtain an estimation of $\boldsymbol{\theta}$ and of the single value $Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$.
- IS-Rec-CL-1, IS-Rec-CL-2 and IS-Rec-CL-3, using the estimator $\hat{Z}(\boldsymbol{\theta}) = \frac{1}{M} \sum_{m=1}^M \frac{\phi(\mathbf{x}_m | \boldsymbol{\theta})}{q(\mathbf{x}_m)}$ 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 $\hat{Z}_S(\boldsymbol{\theta}) = \frac{1}{M_2} \sum_{k=1}^{M_2} \frac{\phi(\mathbf{s}_k | \boldsymbol{\theta})}{q_2(\mathbf{s}_k)}$, where $\mathbf{s}_k \sim q_2(\mathbf{x})$, $k = 1, \dots, M_2$. Here, the reference samples $\mathbf{x}_m \sim q_1(\mathbf{x})$ are used only for building the corresponding cost function.

As an example, note that CL-1, IS-Rec-CL-1 and IS-CL-1 are using the same functional cost $J_{\text{CL-1}}(\boldsymbol{\theta}, Z)$, as shown in Table 3. Note also that the IS-Rec-CL schemes can be viewed as lying between GA and CL, as the artificial samples drawn from the proposal q are used for both purposes: constructing the cost function and approximating $Z_{\text{tr}}(\boldsymbol{\theta})$.

5 Optimal independent proposal density

5.1 Related theory

For a given fixed value of $\boldsymbol{\theta}$, we know that $q_{\text{opt}}(\mathbf{x}|\boldsymbol{\theta}) = \psi(\mathbf{x}|\boldsymbol{\theta})$ for estimating $Z_{\text{tr}}(\boldsymbol{\theta})$ [24]. However, in our search for $\hat{\boldsymbol{\theta}}_{\text{ML}}$, e.g., using a gradient descent, we consider several densities $\psi(\mathbf{x}|\boldsymbol{\theta}, Z(\boldsymbol{\theta}))$ associated to different $\boldsymbol{\theta}$ -values. Thus, our goal is to select an independent proposal density $q_{\text{opt}}(\mathbf{x})$ that performs adequately for different parameter values (e.g., $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_Q$), and, if possible, for the entire parameter space Θ . For this reason, a possible idea is to consider the following quantity,

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

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

$$\boldsymbol{\theta}_i \sim \mathcal{U}(\Theta_{\text{sub}}),$$

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

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

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

$$\bar{q}_{\text{opt}}(\mathbf{x}) \propto q_{\text{opt}}(\mathbf{x}) = \sqrt{\phi(\mathbf{x}|\boldsymbol{\theta}_1)^2 + \phi(\mathbf{x}|\boldsymbol{\theta}_2)^2 + \dots + \phi(\mathbf{x}|\boldsymbol{\theta}_Q)^2}, \quad (30)$$

as proposal density. For a proof see [24]. 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.

Remark 11. The proposal density $\bar{q}_{\text{opt}}(\mathbf{x})$ in Eq. (30) is the optimal independent proposal pdf (or pmf) for estimating the partition function $Z_{\text{tr}}(\boldsymbol{\theta})$. Hence, this result is optimal for GA, and also optimal for the IS-CL schemes with respect to the second proposal $q_{2,\text{opt}}(\mathbf{x})$ that draws the auxiliary samples $\mathbf{s}_m$. More generally, we have:

- in the IS-CL schemes, we should additionally identify an optimal proposal $q_{1,\text{opt}}(\mathbf{x})$ for generating the reference samples $\mathbf{x}_m$ used to construct the cost function;
- in the IS-Rec-CL schemes, we should instead identify a single optimal proposal $q_{\text{opt}}(\mathbf{x})$ that accounts for both the estimation of the partition function and the fact that the generated samples $\mathbf{x}_m$ are also used to construct the corresponding cost function. However, it is worth noting that the proposal $q_{\text{opt}}(\mathbf{x})$ in Eq. (30) is also able to provide very good performance within the IS-Rec-CL schemes, as demonstrated by the numerical experiments; see, for instance, Figures 4, 5, and 6.

These considerations may provide interesting directions for future research, particularly in light of the previous works in [3, 4].

5.2 Practical use

Generally, a widely-used optimization approach is based on a gradient-descent method. We can consider a set of possible initial nodes $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_Q \sim \mathcal{U}(\boldsymbol{\Theta}_{\text{sub}})$. 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 $M_0 \gg M$ samples from an initial proposal density,

$$\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_{M_0} \sim q_0(\mathbf{x}). \quad (31)$$

2. Assign the weights

$$\rho_i = \frac{q_{\text{opt}}(\bar{\mathbf{x}}_i)}{q_0(\bar{\mathbf{x}}_i)}, \quad (32)$$

$$= \frac{\sqrt{[\phi(\bar{\mathbf{x}}_i|\boldsymbol{\theta}_1)]^2 + [\phi(\bar{\mathbf{x}}_i|\boldsymbol{\theta}_2)]^2 + \dots + [\phi(\bar{\mathbf{x}}_i|\boldsymbol{\theta}_Q)]^2}}{q_0(\bar{\mathbf{x}}_i)}, \quad (33)$$

where q_{opt} is given in Eq. (30).

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

$$C_{\text{opt}} = \frac{1}{M_0} \sum_{k=1}^{M_0} \rho_k.$$

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

$$\hat{Z}(\boldsymbol{\theta}) = \frac{C_{\text{opt}}}{M} \sum_{m=1}^M \frac{\phi(\mathbf{x}_m|\boldsymbol{\theta})}{\sqrt{[\phi(\mathbf{x}_m|\boldsymbol{\theta}_1)]^2 + [\phi(\mathbf{x}_m|\boldsymbol{\theta}_2)]^2 + \dots + [\phi(\mathbf{x}_m|\boldsymbol{\theta}_Q)]^2}}. \quad (34)$$

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

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t-1} - \alpha_{t-1} \nabla_{\boldsymbol{\theta}} J_{\text{GA}}(\hat{\boldsymbol{\theta}}_{t-1}), \quad (35)$$

$$= \hat{\boldsymbol{\theta}}_{t-1} - \alpha_{t-1} \nabla_{\boldsymbol{\theta}} J_{\text{NLL}}(\hat{\boldsymbol{\theta}}_{t-1}|\hat{Z}, \mathbf{y}), \quad (36)$$

with $t = Q+1, \dots, Q+T$ and with suitable choices of the step value α_t . Hence, we have the first Q starting points, i.e., $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_Q$ and the new T points, i.e., $\boldsymbol{\theta}_{Q+1}, \dots, \boldsymbol{\theta}_{Q+T}$. Then, we can recompute

the weights ρ_i and improve $\hat{Z}(\boldsymbol{\theta})$ before continuing with the gradient descent:

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

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

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

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

The gradient descent and the procedure above can be iterated several times until convergence, i.e., $\|\hat{\boldsymbol{\theta}}_t - \hat{\boldsymbol{\theta}}_{t-1}\| < \epsilon$ with $\epsilon > 0$ (clearly, we expect $\hat{\boldsymbol{\theta}}_t \approx \hat{\boldsymbol{\theta}}_{\text{ML}}$ 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 $Z_{\text{tr}}(\boldsymbol{\theta})$ with an independent proposal density $q(\mathbf{x})$.

5.3 Allowing an efficient classical Bayesian analysis

The importance of estimating the complete partition function $Z_{\text{tr}}(\boldsymbol{\theta})$ instead only one point, i.e., only $Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$, is clarified by this section. First of all, we need to highlight another relevant consideration regarding $Z_{\text{tr}}(\boldsymbol{\theta})$ given below.

Remark 12. The partition function $Z_{\text{tr}}(\boldsymbol{\theta})$ does *not* depend on the observed data $\underline{\mathbf{y}} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$. Hence, for a given unnormalized model $\phi(\cdot | \boldsymbol{\theta})$ defined over a parameter space of dimension d_θ , we can approximate $Z_{\text{tr}}(\boldsymbol{\theta})$ once and subsequently reuse this approximation for different inference problems involving different observed datasets, say $\underline{\mathbf{y}}$, $\underline{\mathbf{y}}'$, $\underline{\mathbf{y}}''$, etc. This property can provide a substantial computational advantage when multiple datasets are analyzed under the same underlying probabilistic model (i.e., fixing ϕ and d_θ), allowing also a standard Bayesian inference.

Hence, we can obtain an estimate $\hat{Z}_{\text{tr}}(\boldsymbol{\theta})$ once, and subsequently reuse this resulting approximation across several inference problems involving different datasets, such as $\underline{\mathbf{y}}$, $\underline{\mathbf{y}}'$, $\underline{\mathbf{y}}''$, etc (or we can refine the approximation if a higher level of accuracy is desired). This is particularly useful for allowing Bayesian inference (BI) where, assuming a prior distribution $p(\boldsymbol{\theta})$, a posterior density is analyzed:

$$p(\boldsymbol{\theta} | \underline{\mathbf{y}}) = \frac{1}{p(\underline{\mathbf{y}})} L(\underline{\mathbf{y}} | \boldsymbol{\theta}, Z_{\text{tr}}) p(\boldsymbol{\theta}) \propto L(\underline{\mathbf{y}} | \boldsymbol{\theta}, Z_{\text{tr}}) p(\boldsymbol{\theta}). \quad (39)$$

Recall that the complete likelihood is given by

$$L(\underline{\mathbf{y}}|\boldsymbol{\theta}, Z_{\text{tr}}) = p(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N|\boldsymbol{\theta}) = \frac{1}{Z_{\text{tr}}(\boldsymbol{\theta})^N} \prod_{n=1}^N \phi(\mathbf{y}_n|\boldsymbol{\theta}), \quad (40)$$

Hence, for allowing BI, we can replace the estimation of the partition function, i.e., $\log L(\underline{\mathbf{y}}|\boldsymbol{\theta}, \hat{Z}_{\text{tr}}) = \sum_{n=1}^N \log \phi(\mathbf{y}_n|\boldsymbol{\theta}) - N \log \hat{Z}_{\text{tr}}(\boldsymbol{\theta})$ and

$$\hat{p}(\boldsymbol{\theta}|\underline{\mathbf{y}}) \propto L(\underline{\mathbf{y}}|\boldsymbol{\theta}, \hat{Z}_{\text{tr}})p(\boldsymbol{\theta}). \quad (41)$$

Furthermore, we can use the estimator $\hat{\boldsymbol{\theta}}_{\text{tr}}$ (and the rest of information obtained with the methods above) to design a suitable proposal density $q(\boldsymbol{\theta})$, that is now defined in the $\boldsymbol{\theta}$ -space, for efficiently exploring the (approximated) posterior $\hat{p}(\boldsymbol{\theta}|\underline{\mathbf{y}})$ by Monte Carlo methods. This proposal density $q(\boldsymbol{\theta})$ can be further adapted as information about $\hat{p}(\boldsymbol{\theta}|\underline{\mathbf{y}})$ becomes available. Moreover, we can also use this proposal to draw the nodes $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_Q \sim q(\boldsymbol{\theta})$ used in Section 5.1 for building $q_{\text{opt}}(\mathbf{x})$. Hence, we can refine the choice of $q_{\text{opt}}(\mathbf{x})$ and, consequently, improve the accuracy of the approximation $\hat{Z}_{\text{tr}}(\boldsymbol{\theta})$. It is important to remark that another Bayesian approach for EBM including Z , as an additional parameter to estimate as CL, has been recently proposed in [35].

6 Numerical experiments

To illustrate and assess the proposed techniques, we consider three complementary models. **(a)** First, we consider a Gaussian model, for which the complete partition function is analytically available in closed form. **(b)** Second, we consider an Ising model, which is a genuine discrete EBM whose partition function is, in general, computationally very costly to evaluate because the number of possible configurations grows exponentially with the dimensionality of the model. **(c)** Finally, we consider a nonlinear, non-Gaussian, continuous EBM characterized by a quartic energy function, for which the normalizing constant generally does not admit a closed-form expression. Our primary goals are to facilitate the understanding of the proposed methodologies, provide reliable ground-truth results for comparison, and ensure that the numerical experiments are straightforward to reproduce. Additionally, to further support reproducibility, related code is also made available.⁵

6.1 First experiment

This first numerical study 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_{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.

⁵Related Matlab, Python and R code is given at http://www.lucamartino.altervista.org/PUBLIC_CODE_ISCL.zip

6.1.1 Settings

More specifically, we consider a target distribution of type

$$\psi(y|\theta, Z_{\text{tr}}) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left(-\frac{y^2}{2\theta^2}\right), \quad (42)$$

$$\phi(y|\theta) = \exp\left(-\frac{y^2}{2\theta^2}\right), \quad Z_{\text{tr}}(\theta) = \sqrt{2\pi\theta^2}, \quad (43)$$

$$E(y|\theta) = \frac{y^2}{2\theta^2}. \quad (44)$$

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

$$y_n \sim \psi(y|\theta_{\text{tr}}, Z_{\text{tr}}) = \frac{1}{Z_{\text{tr}}(\theta_{\text{tr}})} \exp\left(-\frac{y^2}{2\theta_{\text{tr}}^2}\right), \quad Z_{\text{tr}}(\theta_{\text{tr}}) = \sqrt{2\pi\theta_{\text{tr}}^2}, \quad (45)$$

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

$$x_m \sim q(x) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(x - \mu_p)^2}{2\sigma_p^2}\right), \quad (46)$$

drawing clearly M artificial data $\{x_1, \dots, x_M\}$ 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 $\theta_{\text{tr}} = 1.5$, 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 $Z_{\text{tr}}(\theta)$.
- **Scenario 2:** We replace $Z(\theta)$ with $\hat{Z}(\theta) = \frac{1}{M} \sum_{m=1}^M \frac{\phi(x_m, \theta)}{q(x_m)}$ employing the same M points $\{x_1, \dots, x_M\}$ used also as reference/noise data. Hence, this scenario is devoted to the methods based on a single proposal (i.e., the recycling schemes).
- **Scenario 3:** We replace $Z(\theta)$ with $\hat{Z}_S(\theta) = \frac{1}{M_2} \sum_{m=1}^{M_2} \frac{\phi(s_m, \theta)}{q_2(s_m)}$ employing other M_2 points $\{s_1, \dots, s_{M_2}\}$ generated from another proposal $s_m \sim q_2(x)$, with $m = 1, \dots, M_2$. 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 $M_2 \rightarrow \infty$. Then, both Scenario 1 and Scenario 3 are related to the methods based on two proposals. There would be a “scenario 4” that consists on the joint estimating of $[\hat{\theta}, \hat{Z}(\hat{\theta})]$ where $\hat{Z}(\hat{\theta})$ approximates a single value $Z_{\text{tr}}(\theta_{\text{tr}})$. 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 $\theta_{\text{tr}} = 1.5$. Hence, we do not consider the error in estimation of $Z(\theta)$. We compute the MSE in estimation of θ_{tr} and, in all scenarios, we keep fixed $\mu_p = 0$ and vary σ_p .

6.1.2 Results for Scenario 1 – Scenario 3 with $M_2 \rightarrow \infty$

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

$$M_2 \rightarrow \infty.$$

For this reason, GA coincides perfectly with the maximum likelihood approach and the solution does not depend on σ_p (since GA does not depend on the samples $x_1, \dots, x_M$). Namely, since we are using the ground-truth $Z_{\text{tr}}(\theta)$ 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 $x_1, \dots, x_M$ (used for building the corresponding cost functions). Since, for a finite and reasonably large value of M_2 , we obtain results very similar to those observed in Scenario 1, we do not provide specific figures for Scenario 3. However, Scenario 3 is included in the available code.

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 $\theta_{\text{tr}} = 1.5$ (the standard deviation of the true model). However, their performance deteriorates for intermediate values of σ_p .

6.1.3 Results for Scenario 2

In this framework, we test the methods based on a single proposal (i.e., the recycling schemes): we employ the M samples $\{x_1, \dots, x_M\}$ drawn from $q(x_m)$ for both (a) as “reference noise data”, and (b) yielding the estimator $\hat{Z}(\theta) = \frac{1}{M} \sum_{m=1}^M \frac{\phi(x_m, \theta)}{q(x_m)}$. Namely, the reference points $\{x_1, \dots, x_M\}$ are recycled to build $\hat{Z}(\theta)$. 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

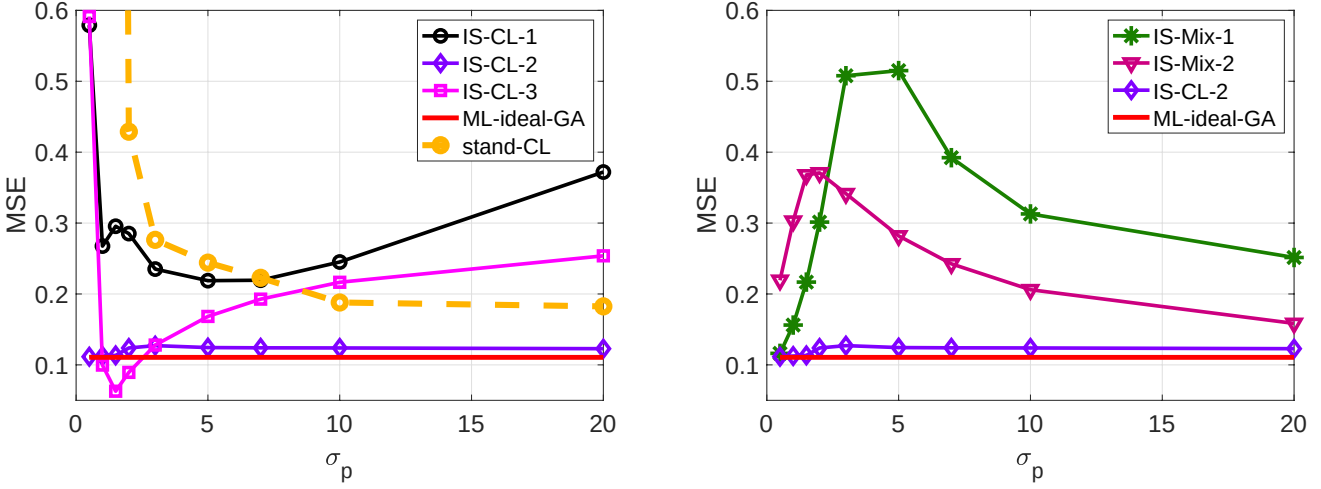


Figure 1: Results related to the first experiment in Section 6.1. MSE (estimating θ_{true}) versus the standard deviation σ_p of the proposal density q in Scenarios 1 and 3, with $N = 10$, $M = 10$ but $M_2 \rightarrow \infty$, hence using the true partition function $Z_{\text{tr}}(\theta)$. Hence, the samples x_m drawn from q are used only for building the cost functions. 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, $\mu_p = 0$.

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 (at least asymptotically as M grows; in Fig. 2, we use $M = 5000$) 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 $\hat{Z}(\theta)$ is poor and the standard CL provides better results.

6.1.4 Summary of the results

Focusing mainly on the inference of θ , we can summarize the main conclusions as follows:

- Without computational restrictions on the evaluation of the numerator $\phi(\mathbf{x}, \theta)$, the estimation $Z(\theta)$ 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 $\phi(\mathbf{x}, \theta)$, 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 .

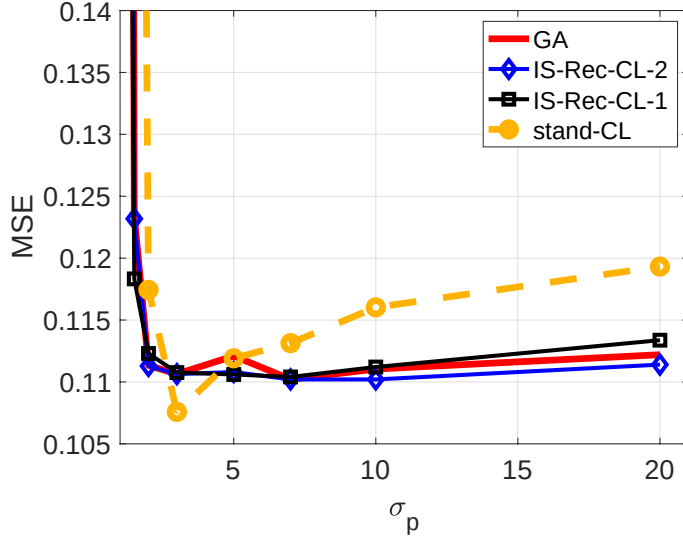


Figure 2: Results related to the first experiment in Section 6.1. MSE versus the standard deviation σ_p in Scenario 2, obtained with some recycling schemes and the standard CL with $N = 10$ and $M = 5000$. We report only the curves corresponding to IS-Rec-CL-1, IS-Rec-CL-2, GA, and the standard CL. The 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.

Moreover, focusing only on θ , with finite N , and with $M_2 \rightarrow \infty$ (or $M \rightarrow \infty$ in the recycling schemes), the IS-within-CL-1, IS-within-CL-2 and IS-within-CL-3 schemes seem to work better than standard CL (in Figure 2, the curve corresponding to IS-Rec-CL-3 is not shown just for visual clarity, as it substantially overlaps with the other curves). In Figure 1, for some specific values of the proposal parameters (and recalling $N = 10$, finite), 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., $\sigma_p < \sigma_{\text{true}}$).

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

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

Once the entire partition function $Z(\theta)$ has been approximated by $\hat{Z}(\theta)$, Bayesian inference can be carried out straightforwardly. Consequently, the GA and IS-within-CL schemes can be employed for a *pre-Bayesian analysis*. Moreover, the pointwise estimator $\hat{\theta}$ can also be useful for designing an efficient Bayesian analysis (as discussed in the Section 5.3). Finally recall that the use of the optimal independent proposal density $q_{\text{opt}}(\mathbf{x})$ can improve the performance of the IS-RC-CL schemes.

6.2 Second experiment: application to Ising model

6.2.1 General formulation of the Ising model: application and issues

Now we consider an *Ising model*, a.k.a. *Boltzmann machine* [16, 18, 21]. The Ising model is one of the fundamental models of statistical physics. It describes a system of interacting binary variables, or “spins”, typically taking values in $\{-1, +1\}$. The probability of a configuration is determined by an energy function involving external fields (represented by coefficients denoted as h_i) and pairwise interactions between spins (represented by coefficients denoted as $J_{i,j}$). The model has a simple formulation but is capable of describing complex and collective phenomena. For these reasons, it has subsequently found applications beyond physics, including image processing, spatial statistics, and network modeling. The connections with images and networks is clear since an Ising model can be naturally represented by an undirected graph, where each node corresponds to a binary spin $y_i \in \{-1, +1\}$, and each edge (i, j) represents a pairwise interaction $J_{i,j}$ between the corresponding variables. More specifically, let us consider that our data are vectors of d binary variables,

$$\mathbf{y} = [y_1, y_2, \dots, y_d]^\top \in \mathcal{D}, \quad y_i \in \{-1, +1\}. \quad (47)$$

where $\mathcal{D} = \{-1, +1\}^d$ represents the set containing of all possible 2^d configurations (i.e., the space of all possible data vectors). Since the data are discrete, the model is a probability mass function (pmf), $\text{Prob}(\mathbf{y}) = \text{Prob}(y_1, \dots, y_d)$, defined as

$$\text{Prob}(\mathbf{y}) = \psi(\mathbf{y}|\boldsymbol{\theta}, Z_{\text{tr}}) = \frac{1}{Z_{\text{tr}}(\boldsymbol{\theta})} \exp \left\{ \sum_{i=1}^d h_i y_i + \sum_{i < j} J_{i,j} y_i y_j \right\}, \quad (48)$$

where $h_1, h_2, \dots, h_d$, are the so-called external-field parameters, and $J_{i,j}$ (with $i < j$ for $i = 1, \dots, d-1$ and $j = 2, \dots, d$) are the pairwise interaction parameters. Thus, the complete parameter vector

$$\boldsymbol{\theta} = [h_1, \dots, h_d, J_{1,2}, \dots, J_{i,j}, \dots, J_{d-1,d}]^\top \in \mathbb{R}^{d_\theta}, \quad (49)$$

has in general a dimension of

$$d_\theta = d + \frac{d(d-1)}{2} = \frac{d(d+1)}{2}.$$

Note that all the components of $\boldsymbol{\theta}$ are real numbers and they can be positive or negative. If all $h_i = 0$, the model is also known as *zero-field Ising model* and the dimension of the problem is only $\frac{d(d-1)}{2}$. The partition function is the normalizing constant that sums the unnormalized probabilities over all possible configurations, and can be computed exactly as

$$Z_{\text{tr}}(\boldsymbol{\theta}) = \sum_{\forall \mathbf{y} \in \{-1, +1\}^d} \phi(\mathbf{y}|\boldsymbol{\theta}), \quad (50)$$

$$= \sum_{\forall \mathbf{y} \in \{-1, +1\}^d} \exp \left\{ \sum_{i=1}^d h_i y_i + \sum_{i < j} J_{i,j} y_i y_j \right\}. \quad (51)$$

The main computational issue is that evaluating the partition function generally requires summing over all 2^d possible configurations, making its exact computation exponentially expensive in the

number of variables. For instance, with $d = 16$, there are $2^{16} = 65536$ possible configurations, while with $d = 20$ this number increases to $2^{20} = 1048576$. Hence, the exact computation of $Z_{\text{tr}}(\boldsymbol{\theta})$ becomes rapidly impractical as the dimensionality d of the problem increases.

6.2.2 Details of the experiment

To clarify all the elements of the Ising model while keeping the exposition simple and facilitating reproducibility, we consider an example with $d = 3$, where

$$\begin{aligned}\boldsymbol{\theta}_{\text{tr}} &= [h_1 = 0.2, \quad h_2 = -0.3, \quad h_3 = 0.1, \\ &\quad J_{1,2} = 0.5, \quad J_{1,3} = 0.2, \quad J_{2,3} = -0.4]^\top,\end{aligned}\tag{52}$$

so that the model, which represents the probability of the vector $\mathbf{y} = [y_1, y_2, y_3]^\top$, can then be written explicitly as

$$\begin{aligned}\text{Prob}(\mathbf{y}) &= \psi(\mathbf{y}|\boldsymbol{\theta}_{\text{tr}}, Z_{\text{tr}}) = \frac{\phi(\mathbf{y}|\boldsymbol{\theta}_{\text{tr}})}{Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})}, \\ &= \frac{\exp\{0.2y_1 - 0.3y_2 + 0.1y_3 + 0.5y_1y_2 + 0.2y_1y_3 - 0.4y_2y_3\}}{Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})},\end{aligned}\tag{53}$$

where the energy $E(\mathbf{y}|\boldsymbol{\theta}_{\text{tr}}) = -\log \phi(\mathbf{y}|\boldsymbol{\theta}_{\text{tr}})$ is then

$$E(\mathbf{y}|\boldsymbol{\theta}_{\text{tr}}) = -0.2y_1 + 0.3y_2 - 0.1y_3 - 0.5y_1y_2 - 0.2y_1y_3 + 0.4y_2y_3.\tag{54}$$

Since $y_i \in \{-1, +1\}$ with $i = 1, 2, 3$, there are only $2^3 = 8$ possible configurations shown in Table 4. Hence, the partition function evaluated at $\boldsymbol{\theta}_{\text{tr}}$ can be computed exactly as:

$$\begin{aligned}Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}}) &= \sum_{k=1}^8 \phi(\mathbf{y}^{(k)}|\boldsymbol{\theta}_{\text{tr}}) = \sum_{k=1}^8 \exp\{0.2y_1^{(k)} - 0.3y_2^{(k)} + 0.1y_3^{(k)} \\ &\quad + 0.5y_1^{(k)}y_2^{(k)} + 0.2y_1^{(k)}y_3^{(k)} - 0.4y_2^{(k)}y_3^{(k)}\}, \\ &= 10.148.\end{aligned}\tag{55}$$

Thus, we can define a vector of 8 true probabilities associated to $\boldsymbol{\theta}_{\text{tr}}$ and the $2^3 = 8$ configurations,

$$\begin{aligned}\mathbf{p}_{\text{true}} &= \mathbf{p}_{\text{true}}(\boldsymbol{\theta}_{\text{tr}}) = [p_1^{(\text{true})}, p_2^{(\text{true})}, \dots, p_8^{(\text{true})}], \\ &= [0.13, 0.24, 0.06, 0.02, 0.05, 0.19, 0.16, 0.13].\end{aligned}\tag{56}$$

Note that the vector $\mathbf{p}_{\text{true}}$ above represents completely the model, i.e., it contains the 8 values taking by the Ising model, $p_k^{(\text{true})} = \psi(\mathbf{y}^{(k)}|\boldsymbol{\theta}_{\text{tr}}, Z_{\text{tr}})$. Clearly, it is the true pmf only at $\boldsymbol{\theta}_{\text{tr}}$. Indeed, as $\boldsymbol{\theta}$ changes, the true pmf is represented by

$$\mathbf{p}_{\text{true}}(\boldsymbol{\theta}) = [p_1^{(\text{true})}(\boldsymbol{\theta}), \dots, p_8^{(\text{true})}(\boldsymbol{\theta})],\tag{57}$$

where $p_k^{(\text{true})}(\boldsymbol{\theta}) = \psi(\mathbf{y}^{(k)}|\boldsymbol{\theta}, Z_{\text{tr}})$ for $k = 1, \dots, 8$.

Table 4: All the possible $2^3 = 8$ configurations $\mathbf{y}^{(k)} = [y_1^{(k)}, y_2^{(k)}, y_3^{(k)}]^\top$, with $k = 1, \dots, 8$, of the Ising model with $d = 3$, i.e., all the possible observed data. The table shows the corresponding value of the unnormalized model $\phi(\mathbf{y}^{(k)}|\boldsymbol{\theta}_{\text{tr}})$, the true value of $Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}}) = 10.148$ and the probabilities $\text{Prob}(\mathbf{y}^{(k)}) = \psi(\mathbf{y}^{(k)}|\boldsymbol{\theta}_{\text{tr}}, Z_{\text{tr}}) = \phi(\mathbf{y}^{(k)}|\boldsymbol{\theta}_{\text{tr}})/Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}})$ corresponding to each configuration.

k	$y_1^{(k)}$	$y_2^{(k)}$	$y_3^{(k)}$	$\phi(\mathbf{y}^{(k)} \boldsymbol{\theta}_{\text{tr}})$	$p_k^{(\text{true})} = \text{Prob}(\mathbf{y}^{(k)}) = \psi(\mathbf{y}^{(k)} \boldsymbol{\theta}_{\text{tr}}, Z_{\text{tr}})$
1	-1	-1	-1	1.3499	$p_1^{(\text{true})} = 0.1330$
2	-1	-1	1	2.4596	$p_2^{(\text{true})} = 0.2424$
3	-1	1	-1	0.6065	$p_3^{(\text{true})} = 0.0598$
4	-1	1	1	0.2231	$p_4^{(\text{true})} = 0.0220$
5	1	-1	-1	0.4966	$p_5^{(\text{true})} = 0.0489$
6	1	-1	1	2.0138	$p_6^{(\text{true})} = 0.1984$
7	1	1	-1	1.6487	$p_7^{(\text{true})} = 0.1625$
8	1	1	1	1.3499	$p_8^{(\text{true})} = 0.1330$
Summing:				$Z_{\text{tr}}(\boldsymbol{\theta}_{\text{tr}}) = 10.148$	1

First test. We keep fixed $h_1 = 0.2$, $h_2 = -0.3$, $h_3 = 0.1$, $J_{13} = 0.2$, $J_{23} = -0.4$ and consider as unknown variable only $\theta = J_{1,2}$ (recall that the true value $J_{1,2} = 0.5$). Hence, in this experiment, we have $\theta = J_{1,2}$ and we can write

$$\phi(\mathbf{y}|\theta) = \exp\{0.2y_1 - 0.3y_2 + 0.1y_3 + \theta y_1 y_2 + 0.2y_1 y_3 - 0.4y_2 y_3\}. \quad (58)$$

and $Z_{\text{tr}}(\theta) = \sum_{k=1}^8 \phi(\mathbf{y}^{(k)}|\theta)$, both function of the scalar value θ .

We generate N observations $\mathbf{y}_1, \dots, \mathbf{y}_N$ where each vector $\mathbf{y}_n = [y_{n,1}, y_{n,2}, y_{n,3}]^\top$ with $y_{n,i} \in \{-1, +1\}$ generated by the Ising model above with parameters defined in Eq. (52) (hence also considering $J_{12} = 0.5$). We draw M samples, $\mathbf{x}_1, \dots, \mathbf{x}_M$, where $\mathbf{x}_m = [x_{m,1}, x_{m,2}, x_{m,3}]^\top$ with $x_{m,i} \in \{-1, +1\}$, considering different independent proposal pmfs defined by the following vector $\mathbf{q} = [q_1, \dots, q_8]$, i.e.,

$$\begin{aligned} \text{Prob}(\mathbf{x}^{(1)}) &= \text{Prob}([-1, -1, -1]) = q_1, & \text{Prob}(\mathbf{x}^{(2)}) &= \text{Prob}([-1, -1, 1]) = q_2, \\ \text{Prob}(\mathbf{x}^{(3)}) &= \text{Prob}([-1, 1, -1]) = q_3, & \text{Prob}(\mathbf{x}^{(4)}) &= \text{Prob}([-1, 1, 1]) = q_4, \\ \text{Prob}(\mathbf{x}^{(5)}) &= \text{Prob}([1, -1, -1]) = q_5, & \text{Prob}(\mathbf{x}^{(6)}) &= \text{Prob}([1, -1, 1]) = q_6, \\ \text{Prob}(\mathbf{x}^{(7)}) &= \text{Prob}([1, 1, -1]) = q_7, & \text{Prob}(\mathbf{x}^{(8)}) &= \text{Prob}([1, 1, 1]) = q_8. \end{aligned}$$

Note that, for each generated artificial data, we have $\mathbf{x}_m \in \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(8)}\}$.

We compare several independent proposals within an IS scheme for obtaining $\hat{Z}(\theta)$, an estimation of the partition function $Z_{\text{tr}}(\theta)$ at different values of θ , denoted as $\theta_1, \dots, \theta_V$ and selected as a grid between -2 and 2 with step 0.1. Then, we compute the mean square error (MSE):

$$\text{MSE} = \frac{1}{V} \sum_{v=1}^V (Z_{\text{tr}}(\theta_v) - \hat{Z}(\theta_v))^2. \quad (59)$$

Moreover, by plugging the estimated normalizing constant $\widehat{Z}(\theta)$ into the likelihood function, we apply the GA and compute the MSE of the resulting maximum likelihood estimate $\widehat{\theta}_{\text{tr}}$ with respect to the true parameter value $\theta_{\text{tr}} = J_{1,2} = 0.5$. More specifically, we test the use of following proposal pmfs:

- the uniform pmf, denoted as $\mathbf{q}_{\text{uni}} = [1/8, \dots, 1/8]$;
- the optimal pmf, denoted as $\mathbf{q}_{\text{opt}} = [q_1^{(\text{opt})}, \dots, q_8^{(\text{opt})}]$, described in Eq. 30,

$$q_k^{(\text{opt})} = \bar{q}_{\text{opt}}(\mathbf{x}^{(k)}) \propto \sqrt{\phi(\mathbf{x}^{(k)}|\theta_1)^2 + \phi(\mathbf{x}^{(k)}|\theta_2)^2 + \dots + \phi(\mathbf{x}^{(k)}|\theta_V)^2}.$$

In this case, we have

$$\mathbf{q}_{\text{opt}} = [0.085, 0.155, 0.104, 0.038, 0.085, 0.344, 0.104, 0.085],$$

that is different from $\mathbf{p}_{\text{true}}$ given in Eq. (56). Note that, for instance, the 6-th configuration $\mathbf{y}^{(6)} = [1, -1, 1]^\top$ in Table 4 has the greatest probability, i.e., $q_6^{(\text{opt})} = 0.344$. Moreover, since here the optimal density is discrete (defined in a finite domain) we can easily draw from it;

- we also use $\mathbf{q} = \mathbf{p}_{\text{true}}$ (denoted also as $\mathbf{q}_{\text{true}}$) given in Eq. (56), i.e., the true pmf at $\theta_{\text{tr}} = J_{1,2} = 0.5$;
- and finally we test a random independent proposal $\mathbf{q}_{\text{rand}}$ generating at each run $u_k \sim \mathcal{U}([0, 1])$ for $k = 1, \dots, 8$, and then set $q_k = \frac{u_k}{\sum_{k=1}^8 u_k}$ as components of $\mathbf{q}_{\text{rand}}$.

One run, with $M = 100$ and $N = 10$, of the IS estimations of the partition function $Z_{\text{tr}}(\theta)$ is given in Figure 3, drawing $M = 100$ samples from the first three different proposals above. Recall that, as discussed in Section 2.4.1, the best choice (considering only the error in estimation) would be to have a different proposal for each θ_q (i.e., a proposal dependent on θ) equal to $\mathbf{p}_{\text{true}}(\theta)$ in Eq. (57), and generate different M samples for each θ_v , i.e., $\mathbf{x}_1^{(\theta_v)}, \dots, \mathbf{x}_M^{(\theta_v)}$ (hence, much more artificial data and model evaluations: MV instead of only M). However, we are comparing proposals, independent of θ , generating only M artificial data (and, hence, only M evaluations of the model). Note that the variance of the estimate depends on θ in all cases: the region with the smallest variance for $\mathbf{q} = \mathbf{p}_{\text{true}}$ is around θ_{tr} as expected by construction, and is more around $\theta = 0$ for $\mathbf{q}_{\text{opt}}$.

Discussion of the results. The results are provided in Figures 4, 5 and 6. Figure 4(a) reports the MSEs computed as in Eq. (59), for estimating the partition function $Z_{\text{tr}}(\theta)$ as function of M , with $N = 10$. We can see the the optimal proposal described in Section 5 provides always the smallest MSE, as expected by the theory. Figure 4(b) provides the MSE for estimating θ_{tr} after replacing the estimation of $Z_{\text{tr}}(\theta)$ in likelihood function, i.e., using GA. For estimating the partition function, $\mathbf{p}_{\text{true}}$ seems the second better possibility, as shown in Fig. 4(a), whereas it is the third possibility in terms of MSE for estimating the true parameter value $\theta_{\text{tr}} = J_{1,2} = 0.5$, outperformed also by the uniform proposal, as shown in Fig. 4(b). This apparent discrepancy can be explained by the fact that a higher variance of the estimated partition function at some specific values of θ (see the example in Fig. 3) may introduce larger errors in the likelihood surface, potentially affecting the location of its maximum (or minimum, if we consider J_{MLL}) and,

consequently, the accuracy of the resulting parameter estimate. However, this effect progressively vanishes as M increases.

Recall that $\mathbf{q}_{\text{opt}}$ is derived to be optimal only for estimating the partition function $Z_{\text{tr}}(\theta)$ (hence, optimal within GA). The optimality within the proposed schemes, where the generated samples also contribute to the construction of the cost functions, requires further investigation; see Remark 11. However, it is worth noting that the proposal $q_{\text{opt}}(\mathbf{x})$ in Eq. (30) also performs very well within the IS-Rec-CL schemes, as illustrated in Figures 6 and 5. Figure 6 depicts the MSE for the estimation of θ_{tr} and shows that $\mathbf{q}_{\text{opt}}$ also provides the best performance across all the IS-Rec-CL schemes for the tested values of M . Within IS-Rec-Mix-1, $\mathbf{q}_{\text{true}}$ also provides good results, whereas $\mathbf{q}_{\text{opt}}$ yields the smallest MSE only for very small values of M .

Finally, Figure 5 shows that IS-Rec-CL-3 generally provides the best performance, regardless of the choice of the proposal, except when $\mathbf{q}_{\text{true}}$ is used, in which case IS-Rec-Mix-1 outperforms all the other techniques. However, IS-Rec-CL-3 still achieves the second-best performance in this case. When using $\mathbf{q}_{\text{uni}}$, $\mathbf{q}_{\text{opt}}$, and $\mathbf{q}_{\text{rand}}$, IS-Rec-CL-2 generally outperforms GA. Similarly, IS-Rec-CL-1 outperforms GA when using $\mathbf{q}_{\text{opt}}$ and $\mathbf{q}_{\text{rand}}$, but performs worse than GA when using $\mathbf{q}_{\text{true}}$. Finally, with $\mathbf{q}_{\text{uni}}$, IS-Rec-CL-1, IS-Rec-CL-2, and GA provide similar performance.

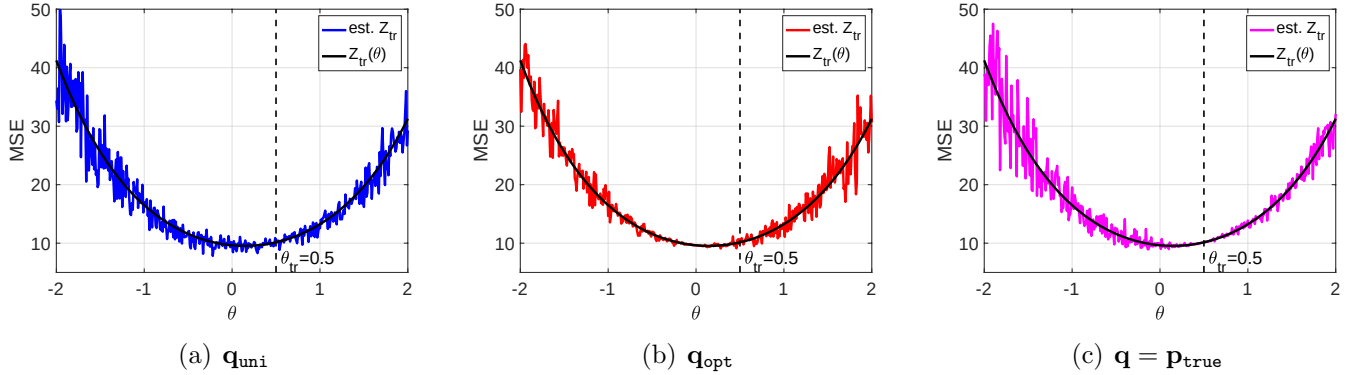
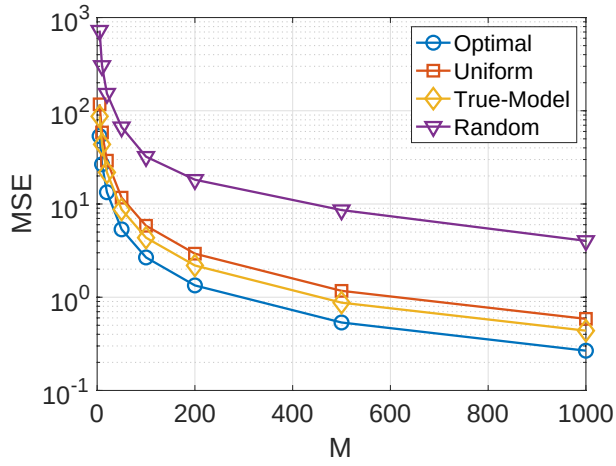
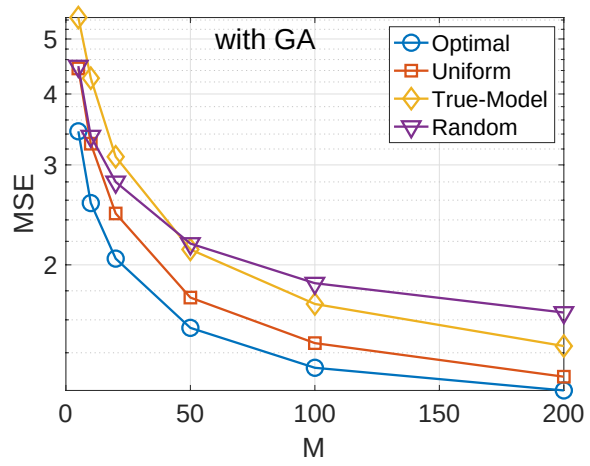


Figure 3: Figures related to the experiment in Section 6.2. One run ($M = 100$, $N = 10$) of estimation of the partition function $Z_{\text{tr}}(\theta)$ (shown with solid black line) using different proposal pmfs within an IS scheme: (a) $\mathbf{q}_{\text{uni}}$, (b) $\mathbf{q}_{\text{opt}}$, and (c) $\mathbf{q} = \mathbf{q}_{\text{true}} = \mathbf{p}_{\text{true}}$.

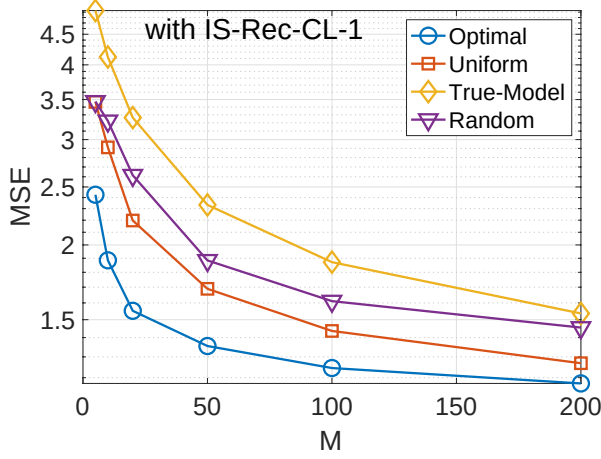


(a) Estimating $Z_{\text{tr}}(\theta)$.

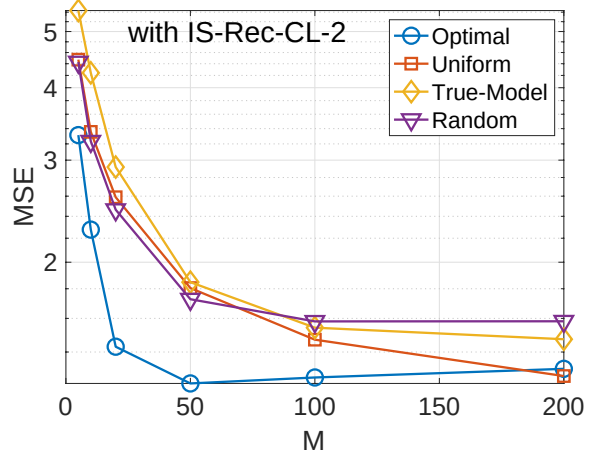


(b) Estimating θ_{tr} .

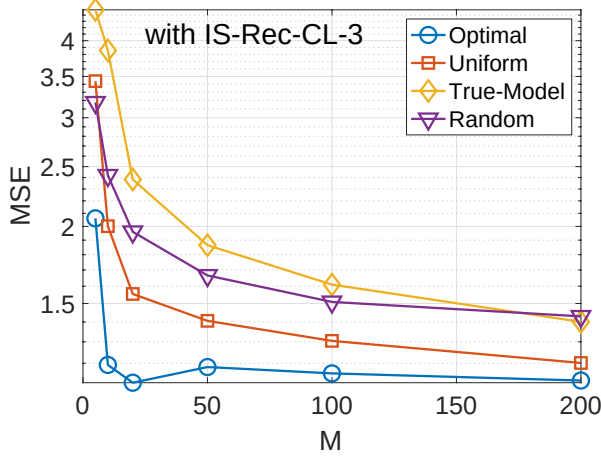
Figure 4: Results related to the experiment in Section 6.2. MSE versus M (averaged over 10^4 runs), with $N = 10$. **(a)** MSE, computed as in Eq. (59), for estimating the partition function $Z_{\text{tr}}(\theta)$. Note that the computation of this MSE involves several values of the function $Z_{\text{tr}}(\theta)$, as shown in Eq. (59). **(b)** MSE obtained by GA estimating the true parameter value $\theta_{\text{tr}} = J_{1,2} = 0.5$.



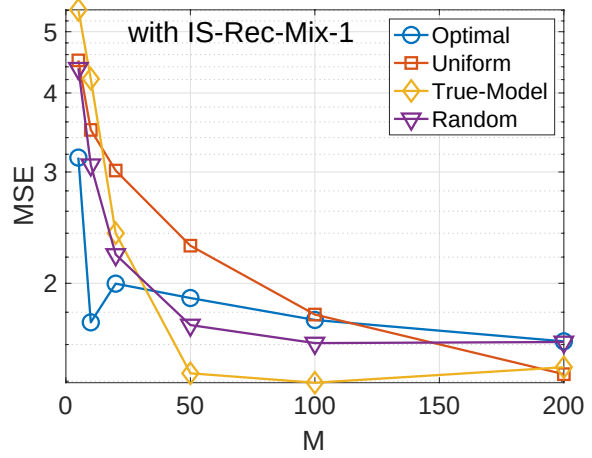
(a)



(b)



(c)



(d)

Figure 5: Results related to the experiment in Section 6.2. MSE versus M for estimating $\theta_{\text{tr}} = J_{1,2} = 0.5$, with $N = 10$ (averaged over 10^4 runs) testing different proposal pmfs and several recycling schemes (i.e., using a single proposal and the samples $\mathbf{x}_m$ are also used for obtaining $\hat{Z}$): more precisely: (a) IS-Rec-CL-1, (b) IS-Rec-CL-2, (c) IS-Rec-CL-3, and (d) IS-Rec-Mix-1.

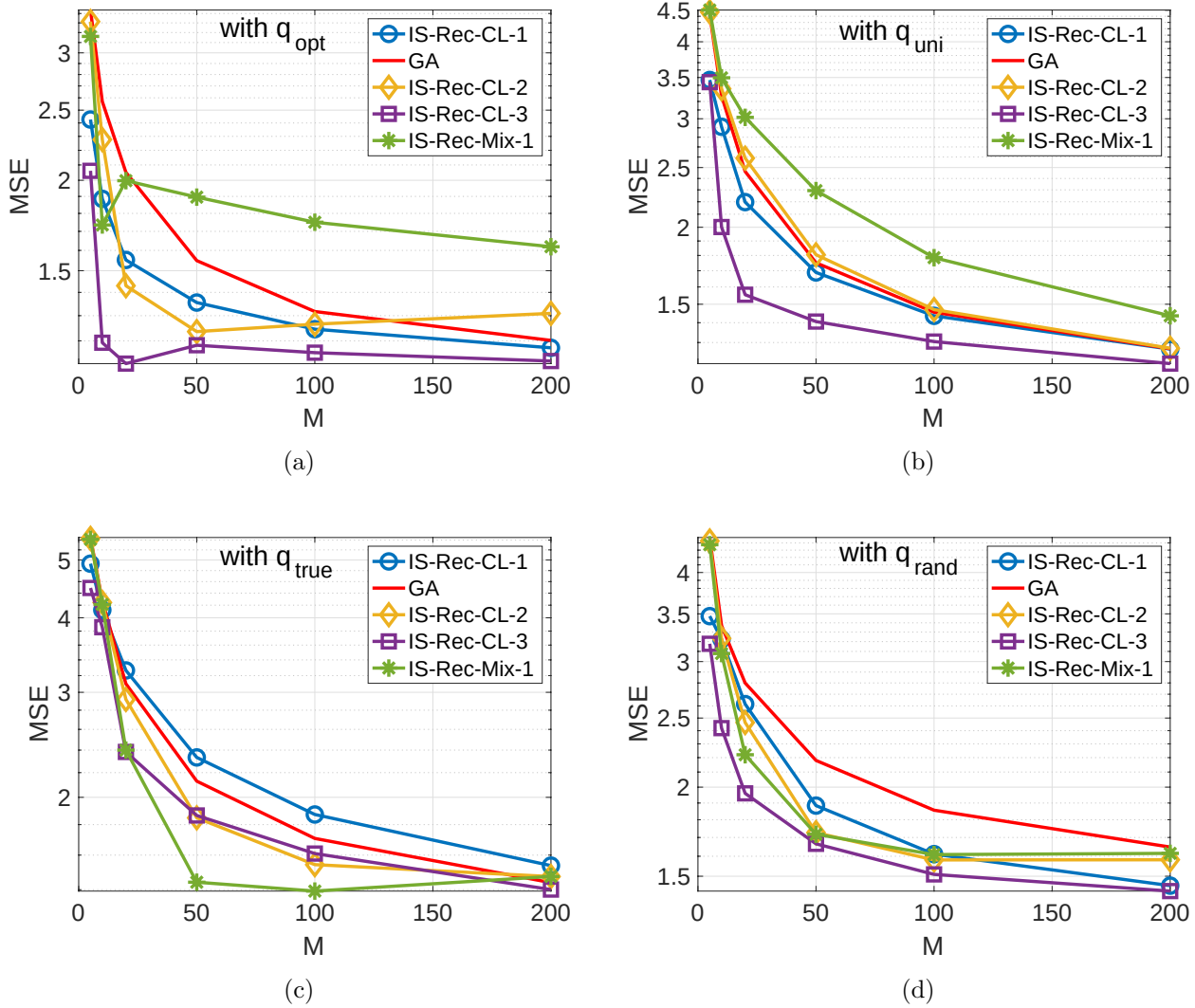


Figure 6: Results related to the experiment in Section 6.2. MSE versus M for estimating $\theta_{\text{tr}} = J_{1,2} = 0.5$, with $N = 10$ (averaged over 10^4 runs), with the different recycling schemes, i.e., using a single proposal and the samples $\mathbf{x}_m$ are also used for obtaining $\hat{Z}$; (a) $\mathbf{q}_{\text{opt}}$, (b) $\mathbf{q}_{\text{uni}}$, (c) $\mathbf{q}_{\text{true}} = \mathbf{p}_{\text{true}}$, and (d) $\mathbf{q}_{\text{rand}}$.

6.3 Third experiment

In this section, we consider a *nonlinear*, *non-Gaussian* and *continuous* EBM, in which the observed variable $y \in \mathbb{R}$ is described through an energy function parameterized by a scalar parameter $\theta \in \mathbb{R}$. We retain this simple one-dimensional setting to facilitate the availability of ground-truth results, while keeping the underlying concepts easy to understand and the numerical experiments straightforward to reproduce. Specifically, we define a nonlinear quartic energy function,

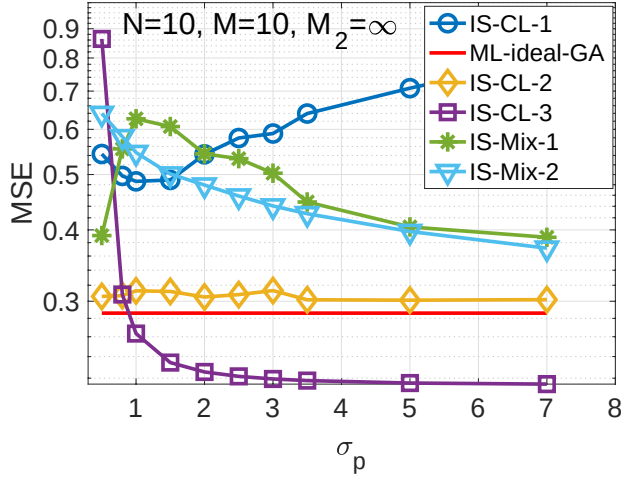
$$E(y|\theta) = y^4 - \theta y^2 + 16, \quad (60)$$

so that the corresponding unnormalized density is $\phi(y|\theta) = \exp[-E(y|\theta)]$. The partition function is given by the integral $Z_{\text{tr}}(\theta) = \int_{\mathbb{R}} \exp[-E(x|\theta)] dy$, that is in general analytically intractable. This nonlinear, non-Gaussian continuous EBM can appear simple but its quartic energy term prevents the density from belonging to the Gaussian family. Moreover, depending on the value of θ , the energy landscape can exhibit either a single minimum or two symmetric minima, resulting in a unimodal or bimodal probability density, respectively. The quartic term also guarantees that the integral defining $Z_{\text{tr}}(\theta)$ is finite for any finite value of θ , while, in general, $Z_{\text{tr}}(\theta)$ does not admit a simple closed-form expression.

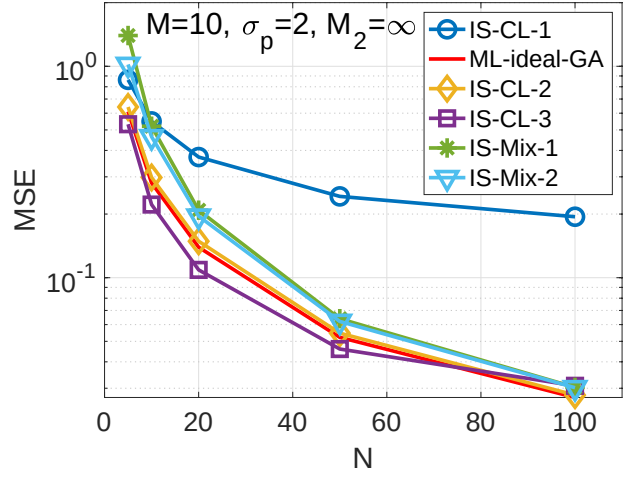
We generate N data points, $y_1, \dots, y_N$ according to the true model assuming $\theta_{\text{tr}} = 2$ (using Monte Carlo sampling methods for drawing these samples). Once again, we consider a Gaussian proposal/reference density, $x_m \sim q(x) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right)$, with $\mu_p = 0$. We test the different techniques considering several cases varying N , M , and σ_p . Note that here q belongs to a different family with respect to the analyzed model, unlike in the first experiment.

The results averaged over 10^4 runs are provided in Figure 7. The results in Figures 7(a) and 7(b) correspond to the Scenarios 1 and 3 of the first experiment, i.e., schemes using two proposals with $M_2 \rightarrow \infty$. Recall, when $M_2 \rightarrow \infty$, the estimator $\hat{Z}_S(\theta) \rightarrow \hat{Z}_{\text{tr}}(\theta)$. Figures 7(c) and 7(d) provide the results of the recycling schemes, i.e., when a single proposal is employed and the M reference data are also used for estimating the partition function $\hat{Z}_{\text{tr}}(\theta)$. The figures depict the MSE estimating $\theta_{\text{tr}} = 2$, versus σ_p of the proposal density q and versus the number of true data N .

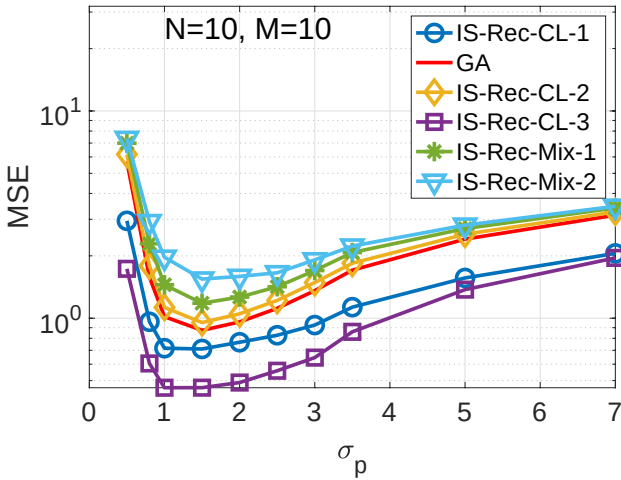
Discussion of the results. The IS-within-CL-3 schemes, both with two proposals (i.e., IS-CL-3) and with a single proposal (i.e., IS-Rec-CL-3), seem to provide the best performance in almost all cases. This result further confirms the good performance given by IS-within-CL-3 in the other numerical experiments, including the second example involving the Ising model. However, the performance of IS-within-CL-3 seems to be better in this example than in the first experiment in Section 6.1. Moreover, in this experiment, the IS-within-CL-2 schemes achieve results very close to those obtained by GA, although the IS-within-CL-2 schemes consistently perform slightly worse than GA. The CL-1 scheme with two proposals, i.e., IS-CL-1, and $M_2 \rightarrow \infty$ yields the worst performance in almost all cases. In contrast, its single-proposal counterpart, IS-Rec-CL-1, appears to be the second-best performing method (this is also in line with the results given the first example). Finally, the two IS-within-Mixture schemes do not perform particularly well in this numerical example, similarly to the results observed in the first experiment in Section 6.1. However, it can depend on the type of proposal density employed, as shown by the results obtained in the second experiment in Section 6.2.



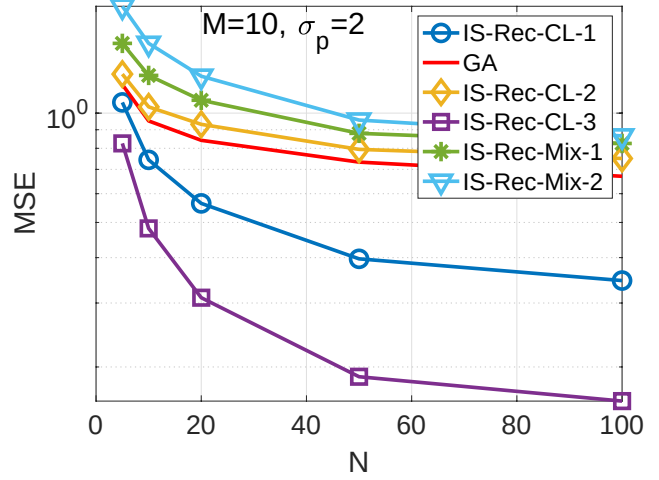
(a)



(b)



(c)



(d)

Figure 7: Results of the third experiment in Section 6.3. MSE for estimating $\theta_{\text{tr}} = 2$, with $M = 10$ (averaged over 10^4 runs): (a)-(c) versus σ_p ; (b)-(d) versus N . Figures (a)-(b) correspond to the two proposal schemes with $M_2 \rightarrow \infty$. Figures (c)-(d) correspond to the methods based on a single proposal (i.e., recycling schemes).

7 Conclusions and future research directions

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 [34], 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 main contributions of this work have been presented in Sections 3, 4, and 5. The theoretical comparison (given in Section 3) between the resulting cost functions not only sheds light on the relationship between GA and CL but also, in light of the considerations in Remark 8, naturally motivates the design of novel variants aimed at improving the performance of these two benchmark methodologies. The novel techniques have been introduced in Section 4. These schemes retain the interpretability and computational advantages of lower-dimensional optimization while benefiting from the flexibility of contrastive formulations. In Section 5, we have introduced an optimal independent proposal density for estimating the partition function in GA and IS-within-CL schemes, also providing an algorithm for its practical implementation. The experimental studies demonstrate 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. Furthermore, the IS-within-CL-2 schemes seem to achieve performance similar to GA. This behavior may be explained by the structure of their cost function, which is given by the sum of two non-negative log-likelihood terms acting “in the same direction”, unlike in CL-1, where one term is subtracted from the other: one corresponding to the EBM likelihood, exactly as in GA, and the other corresponding to the likelihood of a fitted mixture model. Consequently, the first term provides the same likelihood contribution as in GA, while the additional mixture-likelihood term may have a comparatively limited effect on the resulting parameter estimates, working always as a likelihood function. The IS-within-CL-3 schemes often provide the best performance, consistently outperforming IS-within-CL-1, although the reasons for this behavior require further investigation.

Last but not least, an important outcome of IS-within-CL- j or GA schemes is that these methodologies yield an explicit approximation of the entire partition function $Z_{\text{tr}}(\boldsymbol{\theta})$. This can be used as a pre-Bayesian analysis that offers $\hat{Z}(\boldsymbol{\theta})$ and the information about high-probability regions of the parameter space, facilitating the construction of efficient standard Bayesian inference procedures, as shown in Section 5.3. Since the partition function $Z_{\text{tr}}(\boldsymbol{\theta})$ does *not* depend on the observed data $\mathbf{y}$, but only on the unnormalized model $\phi(\cdot|\boldsymbol{\theta})$ and the dimensionality $d_{\boldsymbol{\theta}}$ of the space, it can be approximated once and subsequently reused across different Bayesian inference problems involving different observed datasets. Moreover, if desired, the approximation can be progressively refined.

Finally, several interesting directions for future research can be identified. A first possible direction concerns extending the analysis developed in this work to mixtures of EBMs. Regarding the estimation of normalizing constants, the relationships between standard CL, reverse logistic regression, and extended bridge sampling have already been investigated in the literature; see, in particular, [27]. More generally, the CL approach has recently been extended to mixtures of EBMs in [30, 35] and applied in [6]. Extending the analysis developed in the present work to this mixture setting would require several additional considerations, including the generalization of the IS methodology to the simultaneous estimation of multiple normalizing constants. Moreover, a second relevant direction for future research concerns extending the optimality analysis developed in this work. For instance, one could additionally identify an optimal proposal $q_{1,\text{opt}}(\mathbf{x})$ for generating the reference samples $\mathbf{x}_m$ used exclusively to construct the cost function. Alternatively, one could seek a single optimal proposal $q_{\text{opt}}(\mathbf{x})$ that jointly accounts for the estimation of the

partition function and the fact that the generated samples $\mathbf{x}_m$ are also used to construct the corresponding cost function. These research directions are particularly timely, as also highlighted by recent studies on the optimal design of proposal and noise distributions [3, 4, 24].

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. The work of Luca Scaffidi Domianello has been supported by European Union - Next Generation EU, Mission 4 Component 2 Investment 1.3 in the framework of PNRR MUR project PE0000013 - FAIR Future Artificial Intelligence Research (CUP E63C22001940006).

References

- [1] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36(2):192–236, 1974.
- [2] A. Caimo and A. Mira. Efficient computational strategies for doubly intractable problems with applications to bayesian social networks. *Statistics and Computing*, 25:113–125, 2015.
- [3] 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*, volume 180 of *Proceedings of Machine Learning Research*, pages 307–316, 2022.
- [4] Othmane Chehab, Alexandre Gramfort, and Aapo Hyvärinen. Optimizing the noise in self-supervised learning: From importance sampling to noise-contrastive estimation. *arXiv:2301.09696*, 2023.
- [5] 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*, 2024(10):104011, 2024.
- [6] C. Di Nuzzo, S. Ingrassia, and L. Scaffidi Domianello. Fitting mixtures of von Mises distributions via noise contrastive estimation. *Statistics & Probability Letters*, page 110608, 2025.
- [7] Y. Du and I. Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.
- [8] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.
- [9] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

- [10] H. O. Georgii. *Gibbs Measures and Phase Transitions*. De Gruyter, 2011.
- [11] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics*, 23:156–163, 1991.
- [12] C. J. Geyer. On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B*, 56(2):261–274, 1994.
- [13] C. J. Geyer and E. A. Thompson. Likelihood inference for spatial point processes. *Journal of the Royal Statistical Society, Series B*, 61(3):657–689, 1999.
- [14] M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Journal of Machine Learning Research*, 13:307–361, 2012.
- [15] M. U. Gutmann, S. Kleinegesse, and B. Rhodes. Statistical applications of contrastive learning. *Behaviormetrika*, 49:277–301, 2022.
- [16] G. E. Hinton. *A Practical Guide to Training Restricted Boltzmann Machines*, pages 599–619. Springer, 2012.
- [17] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.
- [18] A. Hyvärinen. Consistency of pseudolikelihood estimation of fully visible boltzmann machines. *Neural Computation*, 18(10):2283–2292, 2006.
- [19] A. Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007.
- [20] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. American Mathematical Society, 1980.
- [21] C. Kuelske. The Ising model: highlights and perspectives. *Mathematical Physics, Analysis and Geometry*, 28(3):20, Aug 2025.
- [22] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang. A tutorial on energy-based learning. *Predicting Structured Data*, pages 1–59, 2006.
- [23] F. Liang. A double metropolis-hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2010.
- [24] F. Llorente and L. Martino. Optimality in importance sampling: A gentle survey. *arXiv:2502.07396*, 2025.
- [25] F. Llorente, L. Martino, D. Delgado, and J. López-Santiago. Marginal likelihood computation for model selection and hypothesis testing: An extensive review. *SIAM Review*, 65(1):3–58, 2023.

- [26] 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*, 93(1):18–61, 2025.
- [27] L. Martino. A unifying view of contrastive learning, importance sampling, and bridge sampling for energy-based models. *arXiv:2604.08116*, 2026.
- [28] L. Martino, V. Elvira, D. Luengo, and J. Corander. Layered adaptive importance sampling. *Statistics and Computing*, 27(3):599–623, 2017.
- [29] 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>, pages 1–10, 2025.
- [30] T. Matsuda and A. Hyvärinen. Estimation of non-normalized mixture models. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2555–2563. PMLR, 2019.
- [31] I. Murray, Z. Ghahramani, and D. MacKay. Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*, 2012.
- [32] A. B. Owen and Y. Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.
- [33] J. Park and M. Haran. Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390, 2018.
- [34] L. Riou-Durand and N. Chopin. Noise contrastive estimation: Asymptotics and comparison with MC-MLE. *arXiv:1801.10381*, 2019.
- [35] N. Sonobe, S. Sugawara, D. Mochihashi, and T. Matsuda. Contrastive Bayesian inference for unnormalized models. *arXiv:2603.09306*, 2026.
- [36] M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers, 2008.

A Direct approximation of the gradient

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

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

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

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

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

$$\begin{aligned} \frac{1}{N} \nabla_{\boldsymbol{\theta}} J_{\text{GA}}(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} E(\mathbf{y}_n | \boldsymbol{\theta}) - \frac{1}{\sum_{j=1}^M \frac{\phi(\mathbf{x}_j | \boldsymbol{\theta})}{q(\mathbf{x}_j)}} \sum_{m=1}^M \frac{\nabla_{\boldsymbol{\theta}} E(\mathbf{x}_m | \boldsymbol{\theta}) \phi(\mathbf{x}_m | \boldsymbol{\theta})}{q(\mathbf{x}_m)} \\ &= \frac{1}{N} \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} E(\mathbf{y}_n | \boldsymbol{\theta}) - \sum_{m=1}^M \bar{w}_m^{(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_m | \boldsymbol{\theta}), \end{aligned} \quad (61)$$

where we have defined the normalized IS weights as:

$$\bar{w}_m^{(\boldsymbol{\theta})} = \frac{\frac{\phi(\mathbf{x}_m | \boldsymbol{\theta})}{q(\mathbf{x}_m)}}{\sum_{j=1}^M \frac{\phi(\mathbf{x}_j | \boldsymbol{\theta})}{q(\mathbf{x}_j)}} = \frac{w_m^{(\boldsymbol{\theta})}}{\sum_{j=1}^M w_j^{(\boldsymbol{\theta})}}, \quad (62)$$

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. (61) can be rewritten as

$$\frac{1}{N} \nabla_{\boldsymbol{\theta}} J_{\text{GA}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \nabla_{\boldsymbol{\theta}} E(\mathbf{y}_n | \boldsymbol{\theta}) - \frac{1}{M} \sum_{m=1}^M \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_m^{(\boldsymbol{\theta})} | \boldsymbol{\theta}), \quad \mathbf{x}_m^{(\boldsymbol{\theta})} \sim p(\cdot | \boldsymbol{\theta}). \quad (63)$$

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.