# A note on gradient-based parameter estimation for energy-based models

Luca Martino[1], Salvatore Ingrassia[1], Sofia Mangano[2], and
Luca Scaffidi Domianello[1]

[1] Department of Economics and Business, University of Catania
Corso Italia, 55 - 95129 Catania (Italy)
[2] Department of Mathematics and Computer Science, University of Catania
Viale Andrea Doria, 6 - 95125 Catania (Italy)
[luca.martino, salvatore.ingrassia, luca.scaffidi]@unict.it
sofia.mangano@phd.unict.it

**Abstract.** Energy-based models (EBMs) are an important family of models where a piece of the likelihood is intractable, and hence unknown. For this reason, the parameter estimation in EBMs is a challenge for the standard estimation methods. In this paper, we present a critical discussion of gradient-based approaches for inference in energy-based models. We provide many details of different derivations, clarify connections and differences. We give practical suggestions for the application of the different schemes. Specifically, we focus on a suitable choice of the proposal/reference density that is crucial for the performance of the gradient-based procedures.

**Keywords:** Energy-based models, gradient descent, maximum likelihood

## 1 Introduction

An energy-based model (EBM) is a statistical model only specified up to the so-called partition function. The partition function normalizes the model so that it integrates to one for any choice of the parameters. However, it is often impossible to obtain it in closed form. Gibbs distributions, Markov and multi-layer networks are examples of models where analytical normalization is often impossible. See e.g. [8] for a review.
More specifically, considering empirical data coming from a sample space $\mathcal{X} \subseteq \mathbb{R}^d$, an EBM is defined as $p(\mathbf{x}|\boldsymbol{\theta}) = \frac{\phi(\mathbf{x}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$, depending on some parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^q$. The numerator $\phi(\mathbf{x}|\boldsymbol{\theta})$ is known and can be evaluated, whereas the denominator $Z(\boldsymbol{\theta}) = \int_{\mathcal{X}} \phi(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x}$ is analytically intractable and hence unknown. The intractable denominator $Z(\boldsymbol{\theta})$ is also referred as the partition function. Moreover, the function $E(\mathbf{x}|\boldsymbol{\theta}) = -\log \phi(\mathbf{x}|\boldsymbol{\theta})$ is often called *energy* (this term comes from the statistical mechanics). The energy function associates small values to good estimates of $\boldsymbol{\theta}$ and large values to bad estimates of $\boldsymbol{\theta}$. We remark

that traditional statistical models concerning classification, regression and density estimation can be reformulated in terms of energy-based models.
Even if nowadays the most common approaches avoid the normalization constant [4], like the *noise contrastive estimation* [3, 5] and the *score matching* first proposed in [7], in this paper we revisit some earlier approaches, that conversely approximate the normalization constant, based on the *noisy* gradient descent of negative log-likelihood function [2, 1] because we noticed that some important related statistical issues have been somehow overlooked (and/or missed) in the literature. We provide many details of different (complete) derivations, discussing connections and relationships and several practical suggestions for the choice of a proposal/reference density that is required in these gradient descent approaches. We remark also that, to handle more flexible distributions, or data characterized by heterogeneous sub-groups [12] quite recently proposed finite mixtures of energy-based models.

## 2 Energy-based models (EBMs)

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

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{\phi(\mathbf{x}|\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}, \tag{1}$$

For given $\boldsymbol{\theta}$, we can evaluate assume in general that the integral[3]

$$Z(\boldsymbol{\theta}) = \int_{\mathcal{X}} \phi(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}, \tag{2}$$

is unknown because it cannot be solved analytically in closed form, i.e., the integral is intractable.[4] Hence, the normalizing constant $Z(\boldsymbol{\theta})$, often called *partition function*, cannot be evaluated point-wise. This represents a challenge for a maximum likelihood estimation (MLE), as we discuss below.

## 3 MLE approaches to parameter estimation in EBMs

Let us assume that we have an observed dataset $\underline{\mathbf{x}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathcal{X}^N$, that contains iid realizations distributed as the the EBM in Eq. (1) for a specific unknown vector of parameters $\boldsymbol{\theta}^*$, i.e., $\mathbf{x}_n \sim p(\mathbf{x}|\boldsymbol{\theta}^*)$ for all $n = 1, ..., N$. In order

[3] All the integrals in this work are definite integrals. However, in the rest of the paper, for simplicity we avoid to write the integration domain.

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

to estimate the parameter of the distribution, the likelihood function of $\boldsymbol{\theta}$ given $\underline{\mathbf{x}}$ is given by

$$L(\boldsymbol{\theta}|\underline{\mathbf{x}}) = p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})^N} \prod_{n=1}^{N} \phi(\mathbf{x}_n|\boldsymbol{\theta}),$$

and then corresponding the log-likelihood $\mathcal{L}(\boldsymbol{\theta}|\underline{\mathbf{x}})$ is

$$\mathcal{L}(\boldsymbol{\theta}|\underline{\mathbf{x}}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \phi(\mathbf{x}_n|\boldsymbol{\theta}) - N \log Z(\boldsymbol{\theta}). \tag{3}$$

Since $E(\mathbf{x}_n|\boldsymbol{\theta}) = -\log \phi(\mathbf{x}_n|\boldsymbol{\theta})$, the log-likelihood (3) becomes

$$\mathcal{L}(\boldsymbol{\theta}|\underline{\mathbf{x}}) = -\sum_{n=1}^{N} E(\mathbf{x}_n|\boldsymbol{\theta}) - N \log Z(\boldsymbol{\theta}). \tag{4}$$

The maximum likelihood estimation (MLE) of $\boldsymbol{\theta}$ is often reformulated as the minimization of a loss function $J(\boldsymbol{\theta})$ defined as the negative log-likelihood function (NLL), i.e.,

$$J(\boldsymbol{\theta}) = \mathrm{NLL}(\boldsymbol{\theta}|\underline{\mathbf{x}}) = -\mathcal{L}(\boldsymbol{\theta}|\underline{\mathbf{x}}) = \sum_{n=1}^{N} E(\mathbf{x}_n|\boldsymbol{\theta}) + N \log Z(\boldsymbol{\theta}), \tag{5}$$

so that $\widehat{\boldsymbol{\theta}} = \arg_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max \mathcal{L}(\boldsymbol{\theta}|\underline{\mathbf{x}}) = \arg_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \min J(\boldsymbol{\theta})$. Generally, a standard widely-used optimization approach is based on the so called *gradient-descent*, i.e., computing a finite sequence of estimates $\{\boldsymbol{\theta}^{(t)}\}_{t=0}^{T}$, starting from some initial guess $\boldsymbol{\theta}^{(0)}$ (and suitable choices of the step value $\alpha_t$) according to

$$\widehat{\boldsymbol{\theta}}_t = \widehat{\boldsymbol{\theta}}_{t-1} - \alpha_t \nabla_{\boldsymbol{\theta}} J(\widehat{\boldsymbol{\theta}}_{t-1}). \tag{6}$$

For a suitable large number $T$ of iterations, we get the final estimate $\widehat{\boldsymbol{\theta}}_T \approx \widehat{\boldsymbol{\theta}} = \arg_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \min J(\boldsymbol{\theta})$, where $J(\boldsymbol{\theta})$ is given in Eq. (5). However, as we remarked above, $Z(\boldsymbol{\theta})$ is unknown for any $\boldsymbol{\theta}$, and cannot be computed in closed form. As a consequence, we cannot compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ neither. Therefore, suitable strategies must be adopted for approximating $Z(\boldsymbol{\theta})$ (or its gradient).

### 3.1 The Geyer's approach [2]

Theoretically speaking, the simplest approach relies on applications of some numerical method for computing $Z(\boldsymbol{\theta})$, and this approximation can be obtained by importance sampling (IS) [9]. It was proposed firstly in [2], and we also refer to it as *baseline* approach. Let $q(\cdot)$ be a known pdf with support $\mathcal{X}$ (which implies $\int q(\mathbf{y})d\mathbf{y} = 1$) and let $\underline{\mathbf{y}} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$, with $\mathbf{y} \in \mathcal{X} \subseteq \mathbb{R}^d$ $(m = 1, \ldots, M)$ be a sample of size $M$ generated from $q(\mathbf{y})$, i.e. $\mathbf{y}_1, \ldots, \mathbf{y}_M \sim q(\cdot)$ that is chosen by the user; in this literature $q(\cdot)$ is called *proposal or reference density* [2]. The IS estimator $\widehat{Z}(\boldsymbol{\theta})$ of $Z(\boldsymbol{\theta})$ is given by

$$\widehat{Z}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M}\frac{\phi(\mathbf{y}_m|\boldsymbol{\theta})}{q(\mathbf{y}_m)} = \frac{1}{M}\sum_{m=1}^{M} w_m^{(\boldsymbol{\theta})}, \qquad \mathbf{y}_m \sim q(\mathbf{y}), \tag{7}$$

where $w_m^{(\boldsymbol{\theta})} = \frac{\phi(\mathbf{y}_m|\boldsymbol{\theta})}{q(\mathbf{y}_m)}$ are the (unnormalized) importance weights. It is easy to show that the estimator above is unbiased [9]. Hence, we can consider the following approximation of Eq. (5):

$$\begin{aligned} J_{\texttt{BL}}(\boldsymbol{\theta}) &= \sum_{n=1}^{N} E(\mathbf{x}_n|\boldsymbol{\theta}) + N\log \widehat{Z}(\boldsymbol{\theta}), \\ &= \sum_{n=1}^{N} E(\mathbf{x}_n|\boldsymbol{\theta}) + N\log\left[\frac{1}{M}\sum_{m=1}^{M}\frac{\phi(\mathbf{y}_m|\boldsymbol{\theta})}{q(\mathbf{y}_m)}\right], \quad \mathbf{y}_m \sim q(\mathbf{y}). \end{aligned} \tag{8}$$

Note that $J_{\texttt{BL}}(\boldsymbol{\theta})$ is a random variable depending on $\mathbf{y}_m$'s, but for fixed $\underline{\mathbf{y}}$ $J_{\texttt{BL}}(\boldsymbol{\theta})$ becomes a deterministic function of $\boldsymbol{\theta}$. In Figure 1(a), we can see three different realizations of $J_{\texttt{BL}}(\boldsymbol{\theta})$. Then, the idea is to minimize this function [2],

$$\widehat{\boldsymbol{\theta}}_{\texttt{BL}} = \arg\min J_{\texttt{BL}}(\boldsymbol{\theta}). \tag{9}$$

We remark that the method requires *only one generation* of the artificial dataset, i.e., only $M$ artificial data points $\mathbf{y}_1, \ldots, \mathbf{y}_M \sim q(\mathbf{y})$. After this generation and given $\{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$, $J_{\texttt{BL}}(\boldsymbol{\theta})$ becomes a fixed, analytically known, and evaluable cost function. We point out also that there are two ways for reducing the variance of $J_{\texttt{BL}}(\boldsymbol{\theta})$:

- *increasing the sample size*: as $M \to \infty$, then $\widehat{Z}(\boldsymbol{\theta}) \to Z(\boldsymbol{\theta})$ and $J_{\texttt{BL}}(\boldsymbol{\theta}) \to J(\boldsymbol{\theta})$ as well.
- *select a good proposal density* $q(\mathbf{y})$ for fixed value of $M$; see the next section for more details.

### 3.2 On the choice of the proposal density $q(\cdot)$

The proposal density must be satisfy the following two assumption:

1. The analytic form of the proposal density $q(\cdot)$ must be available, and we need to be able to evaluate it point-wise.
2. We need to be able to draw samples from $q(\cdot)$.

In order to reduce the variance of the estimator in (7), a better choice would be to choose a proposal density that depends on $\boldsymbol{\theta}$, i.e., $q(\mathbf{y}|\boldsymbol{\theta})$. Indeed, the optimal reference density in this scenario is

$$q_{\texttt{opt}}(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta}) \propto \phi(\mathbf{y}|\boldsymbol{\theta}). \tag{10}$$

It can be shown that this choice minimizes the variance of $\widehat{Z}(\boldsymbol{\theta})$ [10], [9]. We point out that, even though the choice $q_{\texttt{opt}}(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})$ minimizes the the variance of $\widehat{Z}(\boldsymbol{\theta})$, this choice presents one important drawback and an additional computational cost:

a First of all, generally, we are not able to draw samples from the EBM model, i.e., from $q_{\texttt{opt}}(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})$. See Section 6 for further details.

b Secondly, for any proposal density $q(\mathbf{y}|\boldsymbol{\theta})$ depending on $\boldsymbol{\theta}$, we should generate a set of artificial data for each $\boldsymbol{\theta}$, i.e., $\mathbf{y}_1^{(\boldsymbol{\theta})}, \dots, \mathbf{y}_M^{(\boldsymbol{\theta})} \sim q(\mathbf{y}|\boldsymbol{\theta})$. Namely, for each evaluation of $J_{\texttt{BL}}(\boldsymbol{\theta})$ at some $\boldsymbol{\theta}$ we would require the generation of another set of artificial data. For instance, if we desire to evaluate $J_{\texttt{BL}}(\boldsymbol{\theta})$ in $L$ different points $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_L$, we would need to draw $L$ artificial samples of size $M$, since for each $\boldsymbol{\theta}_l$, for $l = 1, ..., L$, we would have a different set $\{\mathbf{y}_1^{(\theta_l)}, \dots, \mathbf{y}_M^{(\boldsymbol{\theta}_l)}\}$. This fact could also increase variability in the evaluation of the function $J_{\texttt{BL}}(\boldsymbol{\theta})$.

In particular, this last consideration is shared by any proposal density $q(\mathbf{y}|\boldsymbol{\theta})$ depending on $\boldsymbol{\theta}$, and not just the optimal one.

**Independent proposal.** Choosing a proposal $q(\mathbf{y})$ independent from $\boldsymbol{\theta}$ allows us to generate only one set of artificial data $\{\mathbf{y}_1, \dots, \mathbf{y}_M\}$, for all the processes. On the other hand, for some values of $\boldsymbol{\theta}$, the proposal $q(\mathbf{y})$ can be a non suitable choice, and the variance $\widehat{Z}(\boldsymbol{\theta})$ can vary drastically with $\boldsymbol{\theta}$. The experience with IS estimators suggests the use of a proposal $q(\mathbf{y})$ with great variance, bigger than the variance of $p(\mathbf{x}|\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$, if possible [10].

**Non-optimal but good choices.** From a theoretical point of view, it is interesting to observe that a non-optimal but good choice of the proposal density is any density close to the true model generating the observed data, i.e.,

$$q(\mathbf{y}) \approx p(\mathbf{y}|\boldsymbol{\theta}^*) = \frac{\phi(\mathbf{y}|\boldsymbol{\theta}^*)}{Z(\boldsymbol{\theta}^*)}, \tag{11}$$

This proposal ensures to have small variance in the estimation of $Z(\boldsymbol{\theta})$ around of the true value $\boldsymbol{\theta}^*$. An even better (and more robust) choice of of $q(\mathbf{y})$ is a density that mimics the shape of $p(\mathbf{y}|\boldsymbol{\theta}^*)$ but with more variance (the area under $q(\mathbf{y})$ is more diffuse). In this scenario, we can avoid catastrophic behaviors of the IS estimator, as shown in the numerical experiments of [10]. See also the results depicted in Figures 1-2 of Section 7. Clearly, we do not know $\boldsymbol{\theta}^*$ and we are not able to draw from $p(\mathbf{y}|\boldsymbol{\theta}^*)$. However, this consideration can drive the construction of a good proposal density.

## 4 Approximating the gradient: a first derivation

Noting that $\nabla_{\boldsymbol{\theta}} \log \widehat{Z}(\boldsymbol{\theta}) = \frac{1}{\widehat{Z}(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} \widehat{Z}(\boldsymbol{\theta})$, the gradient of the negative log-likelihood function in Eq. (5), scaled by a factor $1/N$, becomes:

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

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

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

Replacing above, we obtain

$$\begin{aligned}
\frac{1}{N} \nabla_{\boldsymbol{\theta}} J_{\mathtt{BL}}(\boldsymbol{\theta}) &= \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_n|\boldsymbol{\theta}) - \frac{1}{\sum_{j=1}^{M} \frac{\phi(\mathbf{y}_j|\boldsymbol{\theta})}{q(\mathbf{y}_j)}} \sum_{m=1}^{M} \frac{\nabla_{\boldsymbol{\theta}} E(\mathbf{y}_m|\boldsymbol{\theta}) \phi(\mathbf{y}_m|\boldsymbol{\theta})}{q(\mathbf{y}_m)} \\
&= \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_n|\boldsymbol{\theta}) - \sum_{m=1}^{M} \bar{w}_m^{(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} E(\mathbf{y}_m|\boldsymbol{\theta}), \qquad (12)
\end{aligned}$$

where we have defined the normalized importance weights as:

$$\bar{w}_m^{(\boldsymbol{\theta})} = \frac{\frac{\phi(\mathbf{y}_m|\boldsymbol{\theta})}{q(\mathbf{y}_m)}}{\sum_{j=1}^{M} \frac{\phi(\mathbf{y}_j|\boldsymbol{\theta})}{q(\mathbf{y}_j)}} = \frac{w_m^{(\boldsymbol{\theta})}}{\sum_{j=1}^{M} w_j^{(\boldsymbol{\theta})}}, \qquad (13)$$

and $w_m^{(\boldsymbol{\theta})} = \frac{\phi(\mathbf{y}_m|\boldsymbol{\theta})}{q(\mathbf{y}_m)}$ are the unnormalized importance weights.

**Remark 1** If we are able to draw samples from the model, i.e., $\mathbf{y}_m^{(\boldsymbol{\theta})} \sim q_{\mathtt{opt}}(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})$, using the optimal proposal density, the normalized IS weights become $\bar{w}_m^{(\boldsymbol{\theta})} = \frac{1}{M}$ for each $m$, and Eq. (12) becomes

$$\frac{1}{N} \nabla_{\boldsymbol{\theta}} J_{\mathtt{BL}}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_n|\boldsymbol{\theta}) - \frac{1}{M} \sum_{m=1}^{M} \nabla_{\boldsymbol{\theta}} E(\mathbf{y}_m^{(\boldsymbol{\theta})}|\boldsymbol{\theta}). \qquad (14)$$

However, recall that we have a new generation of artificial data $\mathbf{y}_m^{(\boldsymbol{\theta})}$ for each $\boldsymbol{\theta}$.

**Remark 2** Assume again that we are able to draw samples from the model, i.e., $\mathbf{y}_m^{(\boldsymbol{\theta})} \sim q_{\mathtt{opt}}(\mathbf{y}|\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})$. An additional issue is that we cannot evaluate completely $p(\mathbf{y}|\boldsymbol{\theta})$ since $Z(\boldsymbol{\theta})$ is unknown. Thus, we cannot evaluate the *unnormalized* IS weights

$$w_m^{(\boldsymbol{\theta})} = \frac{\phi(\mathbf{y}_m^{(\boldsymbol{\theta})}|\boldsymbol{\theta})}{q_{\mathtt{opt}}(\mathbf{y}_m^{(\boldsymbol{\theta})}|\boldsymbol{\theta})} = \frac{\phi(\mathbf{y}_m^{(\boldsymbol{\theta})}|\boldsymbol{\theta})}{p(\mathbf{y}_m^{(\boldsymbol{\theta})}|\boldsymbol{\theta})} = Z(\boldsymbol{\theta}) \frac{\cancel{\phi(\mathbf{y}_m^{(\boldsymbol{\theta})}|\boldsymbol{\theta})}}{\cancel{\phi(\mathbf{y}_m^{(\boldsymbol{\theta})}|\boldsymbol{\theta})}} = Z(\boldsymbol{\theta}), \quad \forall\, m, \qquad (15)$$

and, as a consequence, we cannot actually evaluate $J_{\mathtt{BL}}(\boldsymbol{\theta})$ in Eq. (8), but only its gradient $\nabla_{\boldsymbol{\theta}} J_{\mathtt{BL}}(\boldsymbol{\theta})$ in Eq. (14) (that only depends on the normalized weights, i.e., $\bar{w}_m^{(\boldsymbol{\theta})} = \frac{1}{M}$). However, if we are only interested in minimizing $J_{\mathtt{BL}}(\boldsymbol{\theta})$, evaluating its gradient $\nabla_{\boldsymbol{\theta}} J_{\mathtt{BL}}(\boldsymbol{\theta})$ is enough.

## 5 Approximating the gradient: a second classical derivation

Here, we follow a more classical derivation employed in different works about inference in EBMs [1, 6]. Instead of approximating $Z(\boldsymbol{\theta})$, let us consider directly the computation of the gradient of the negative log-likelihood function (5) scaled by the factor $1/N$,

$$\frac{1}{N}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_n|\boldsymbol{\theta}) + \frac{1}{Z(\boldsymbol{\theta})}\nabla_{\boldsymbol{\theta}} Z(\boldsymbol{\theta}).$$

Since $Z(\boldsymbol{\theta}) = \int \phi(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} = \int \exp(-E(\mathbf{x}|\boldsymbol{\theta}))d\mathbf{x}$, we have

$$\nabla_{\boldsymbol{\theta}} Z(\boldsymbol{\theta}) = -\int \nabla_{\boldsymbol{\theta}} E(\mathbf{x}|\boldsymbol{\theta}) \exp(-E(\mathbf{x}|\boldsymbol{\theta}))d\mathbf{x} = -\int \nabla_{\boldsymbol{\theta}} E(\mathbf{x}|\boldsymbol{\theta})\phi(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x},$$

and replacing above we obtain

$$\begin{aligned}
\frac{1}{N}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_i|\boldsymbol{\theta}) - \frac{1}{Z(\boldsymbol{\theta})}\int \nabla_{\boldsymbol{\theta}} E(\mathbf{y}|\boldsymbol{\theta})\phi(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x}, \\
&= \frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_i|\boldsymbol{\theta}) - \int \nabla_{\boldsymbol{\theta}} E(\mathbf{x}|\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x}, \qquad (16) \\
&\approx \frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} E(\mathbf{x}_i|\boldsymbol{\theta}) - \frac{1}{M}\sum_{m=1}^{M} \nabla_{\boldsymbol{\theta}} E(\mathbf{y}_m^{(\boldsymbol{\theta})}|\boldsymbol{\theta}), \quad \mathbf{y}_m^{(\boldsymbol{\theta})} \sim p(\cdot|\boldsymbol{\theta}). \qquad (17)
\end{aligned}$$

**Remark.** Note that Eq. (17) coincides with Eq. (14) and we remark that the following points (the first two have been previously discussed):

(a) Generally, we are not able to draw samples from the EBM model, i.e., $p(\mathbf{y}|\boldsymbol{\theta})$. See Section 6 for more details.

(b) We need to generate a different set of artificial data $\mathbf{y}_1^{(\boldsymbol{\theta})}, \ldots, \mathbf{y}_M^{(\boldsymbol{\theta})}$ for each $\boldsymbol{\theta}$. This point is valid for any proposal density $q(\mathbf{y}|\boldsymbol{\theta})$ depending on $\boldsymbol{\theta}$.

(c) Even if we are able to draw from $p(\mathbf{y}|\boldsymbol{\theta})$, we can approximate only the gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ but not $J(\boldsymbol{\theta})$ by Monte Carlo arguments. See last remark in Section 4 and Eq. (15). However, in order to minimize $J(\boldsymbol{\theta})$, the information of its gradient is enough. Moreover, other numerical integration methods could be employed for recovering $J(\boldsymbol{\theta})$ from its gradient, if required.

**Another possibility.** Here we describe an alternative procedure to avoid the generation of samples from the model $p(\mathbf{x}|\boldsymbol{\theta})$. We can use an IS approach for approximating the integral in Eq. (16), i.e., we can consider the following equality,

$$\begin{aligned}\int \nabla_{\boldsymbol{\theta}} E(\mathbf{x}|\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} &= \int \nabla_{\boldsymbol{\theta}} E(\mathbf{x}|\boldsymbol{\theta}) \frac{p(\mathbf{x}|\boldsymbol{\theta})}{q(\mathbf{x})} q(\mathbf{x}) d\mathbf{x}, \\ &= \frac{1}{Z(\boldsymbol{\theta})} \underbrace{\int \nabla_{\boldsymbol{\theta}} E(\mathbf{x}|\boldsymbol{\theta}) \frac{\phi(\mathbf{x}|\boldsymbol{\theta})}{q(\mathbf{x})} q(\mathbf{x}) d\mathbf{x},}_{(*)}\end{aligned}$$

so that $Z(\boldsymbol{\theta})$ can be approximated by (7) and the term $(*)$ can be approximated by similar IS arguments [9] and we get

$$\begin{aligned}\int \nabla_{\boldsymbol{\theta}} E(\mathbf{x}|\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} &\approx \frac{1}{\sum_{j=1}^{M} \frac{\phi(\mathbf{y}_j|\boldsymbol{\theta})}{q(\mathbf{y}_j)}} \sum_{m=1}^{M} \frac{\nabla_{\boldsymbol{\theta}} E(\mathbf{y}_m|\boldsymbol{\theta}) \phi(\mathbf{y}_m|\boldsymbol{\theta})}{q(\mathbf{y}_m)}, \\ &= \frac{1}{\sum_{j=1}^{M} w_j^{(\boldsymbol{\theta})}} \sum_{m=1}^{M} w_m^{(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} E(\mathbf{y}_m|\boldsymbol{\theta}), \\ &= \sum_{m=1}^{M} \bar{w}_m^{(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} E(\mathbf{y}_m|\boldsymbol{\theta}), \qquad \mathbf{y}_m \sim q(\mathbf{x}), \end{aligned} \tag{18}$$

where $\bar{w}_m^{(\boldsymbol{\theta})}$ are the normalized weights in Eq. (13). We solve the previous issues: (a) we are able to draw from $q(\cdot)$ (since we choose the proposal density $q(\cdot)$), and (b) we draw only once $M$ artificial data. However, $q(\cdot)$ can be not suitable for some values of $\boldsymbol{\theta}$.

## 6 Generating artificial data from $p(\mathbf{x}|\boldsymbol{\theta})$

Generally, we are not able to draw artificial data from the EBM $p(\mathbf{x}|\boldsymbol{\theta})$ for a given $\boldsymbol{\theta}$. In this case, there are mainly two alternatives: apply an MCMC algorithm or an "IS - plus - resampling" scheme. Below, we describe the details of both. As an example of MCMC, we consider the Metropolis-Hastings (MH) algorithm.

**MH algorithm.** As an example of possible MCMC method, we provide a Metropolis-Hastings schemes with an independent proposal density $q(\cdot)$ (independent from the previous state of the chain) [11]. We consider the use of an independent proposal density to facilitate the comparison with the other schemes in the rest of the work:

1. Starting with an arbitrary initial vector $\mathbf{y}_0^{(\boldsymbol{\theta})}$.
2. For $m = 1, \ldots, M$:
   (a) Draw $\mathbf{y}'$ from $q(\mathbf{y})$.

(b) Set $\mathbf{y}_m^{(\boldsymbol{\theta})} = \mathbf{y}'$ with probability,

$$\alpha = \min\left[1, \frac{p(\mathbf{y}'|\boldsymbol{\theta})}{p(\mathbf{y}_{m-1}^{(\boldsymbol{\theta})}|\boldsymbol{\theta})} \frac{q(\mathbf{y}_{m-1}^{(\boldsymbol{\theta})})}{q(\mathbf{y}')}\right] = \min\left[1, \frac{\phi(\mathbf{y}'|\boldsymbol{\theta})}{\phi(\mathbf{y}_{m-1}^{(\boldsymbol{\theta})}|\boldsymbol{\theta})} \frac{Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \frac{q(\mathbf{y}_{m-1}^{(\boldsymbol{\theta})})}{q(\mathbf{y}')}\right],$$

$$= \min\left[1, \frac{\phi(\mathbf{y}'|\boldsymbol{\theta})}{q(\mathbf{y}')} \frac{q(\mathbf{y}_{m-1}^{(\boldsymbol{\theta})})}{\phi(\mathbf{y}_{m-1}^{(\boldsymbol{\theta})}|\boldsymbol{\theta})}\right].$$

Otherwise, set $\mathbf{y}_m^{(\boldsymbol{\theta})} = \mathbf{y}_{m-1}^{(\boldsymbol{\theta})}$ with probability $1-\alpha$.

3. The output is $\{\mathbf{y}_1^{(\boldsymbol{\theta})}, \ldots, \mathbf{y}_M^{(\boldsymbol{\theta})}\}$.

**IS plus resampling.** Below, we present the details of the "IS plus resampling" scheme. Again, we need the use of a proposal density $q(\cdot)$ but the idea is get first a sample $\underline{\mathbf{z}} = \{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$ with $\mathbf{z}_m \sim q(\cdot)$ $(m = 1, \ldots, M)$ and then consider resamples points $\{\mathbf{y}_1^{(\boldsymbol{\theta})}, \ldots, \mathbf{y}_M^{(\boldsymbol{\theta})}\}$ from $\underline{\mathbf{z}}$:

1. Draw $\mathbf{z}_1, \ldots, \mathbf{z}_M \sim q(\cdot)$.
2. Assign the weights

$$w_m^{(\boldsymbol{\theta})} = \frac{\phi(\mathbf{z}_m|\boldsymbol{\theta})}{q(\mathbf{z}_m)}, \qquad m = 1, \ldots, M. \tag{19}$$

3. Compute the normalized weights,

$$\bar{w}_m^{(\boldsymbol{\theta})} = \frac{w_m^{(\boldsymbol{\theta})}}{\sum_{i=1}^M w_i^{(\boldsymbol{\theta})}}, \qquad m = 1, \ldots, M. \tag{20}$$

4. Resample (bootstrap) $M$ times with replacement, within $\{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$ according to normalized weights $\bar{w}_m^{(\boldsymbol{\theta})}$, $m = 1, \ldots, M$. The $M$ resampled samples will be denoted as $\{\mathbf{y}_1^{(\boldsymbol{\theta})}, \ldots, \mathbf{y}_M^{(\boldsymbol{\theta})}\}$. Note that $\mathbf{y}_m^{(\boldsymbol{\theta})} \in \{\mathbf{z}_1, \ldots, \mathbf{z}_M\}$ for any value of $m$. The outputs are the $M$ resampled particles $\{\mathbf{y}_1^{(\boldsymbol{\theta})}, \ldots, \mathbf{y}_M^{(\boldsymbol{\theta})}\}$.

Note that, with respect to the derivation in Section 3.1, we have the additional step of the resampling. We finally highlight that in both cases, MCMC or IS, an *internal* proposal density is required. Then, generally, even if we would like to employ the model $p(\mathbf{x}|\boldsymbol{\theta})$ as a reference density, we need the choice and use of another proposal density $q(\mathbf{y})$, within the Monte Carlo sampling schemes.
Furthermore, in both cases, we obtain samples distributed (approximately) as the target density $p(\mathbf{y}|\boldsymbol{\theta})$. Since the target density $p(\mathbf{y}|\boldsymbol{\theta})$ changes with $\boldsymbol{\theta}$, then the obtained samples $\{\mathbf{y}_1^{(\boldsymbol{\theta})}, \ldots, \mathbf{y}_M^{(\boldsymbol{\theta})}\}$ depend also on the specific fixed $\boldsymbol{\theta}$. Therefore, we need to generate a different set of artificial data $\mathbf{y}_1^{(\boldsymbol{\theta})}, \ldots, \mathbf{y}_M^{(\boldsymbol{\theta})}$ for each $\boldsymbol{\theta}$, i.e., the use of the MCMC or "IS plus resampling" algorithm must be repeated for different $\boldsymbol{\theta}$.

## 7 Some numerical considerations

For the sake of simplicity, and to know the ground-truth, in order to evaluate the different performance, we consider a one-dimensional Gaussian density with zero mean, $\mu = 0$, as observation model, i.e.,

$$p(x|\theta) = \mathcal{N}(x|0, \theta) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left(-\frac{x^2}{2\theta^2}\right), \tag{21}$$

where

$$\phi(x|\theta) = \exp\left(-\frac{x^2}{2\theta^2}\right), \quad E(x|\theta) = \frac{x^2}{2\theta^2}, \quad Z(\theta) = \sqrt{2\pi\theta^2}. \tag{22}$$

We assume that $Z(\theta)$ is unknown and apply the gradient approaches described in this work. We set $\theta^* = 2$ and $N = 100$, so that the observed data are drawn as

$$x_n \sim p(x|\theta^*) = \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{x^2}{8}\right), \tag{23}$$

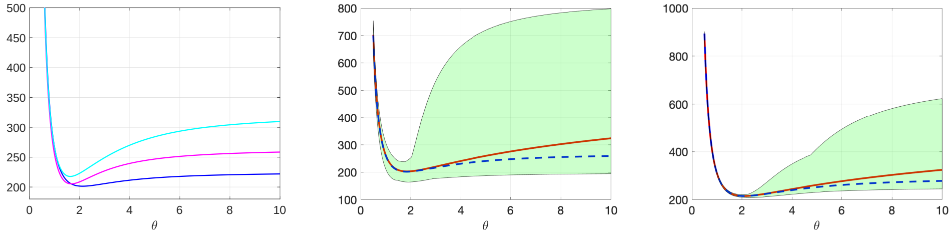
for $n = 1, \ldots, N$. We consider two possible Gaussian reference/proposal densities: the first one is an independent Gaussian proposal with $\mu_p = 2$ and $\sigma_p = 2$, whereas the second one is a Gaussian proposal (which depends on $\theta$) with $\mu_p = 2$ and $\sigma_p = 2 + \theta$,

$$q_1(y) = \mathcal{N}(y|2, 2) = \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{(y-2)^2}{8}\right), \tag{24}$$

$$q_2(y|\theta) = \mathcal{N}(y|2, 2+\theta) = \frac{1}{\sqrt{2\pi(2+\theta)^2}} \exp\left(-\frac{(y-2)^2}{2(2+\theta)^2}\right). \tag{25}$$

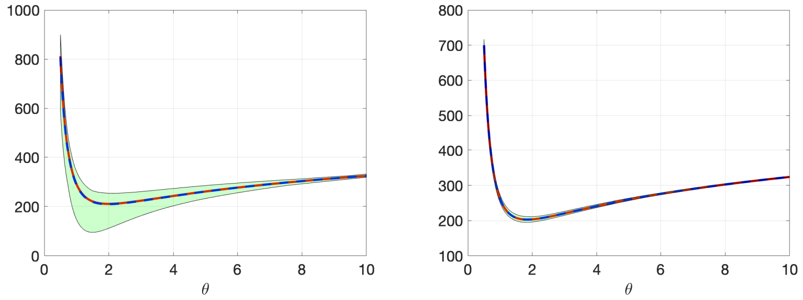
Namely, the artificial data are generated from $y_m \sim q_1(y)$ or $y_m^{(\theta)} \sim q_2(y|\theta)$, for $m = 1, \ldots, M$. We consider two values of $M \in \{100, 5000\}$. We test the results in 1000 independent runs to average the results. Figure 1 provides the results using $q_1(y)$. Figure 1(a) depicts three curves $J_{\texttt{BL}}(\theta)$ in Eq. (8) after generating three different realizations of artificial data $\{y_1, \ldots, y_M\} \sim q_1(y)$, with $M = 100$. In Figures 1(b)-1(c) (corresponding to $M = 100$ and $M = 5000$, respectively) we show the 100% of variability of 1000 curve $J_{\texttt{BL}}(\theta)$ with a green area, the empirical mean of the curves $J_{\texttt{BL}}(\theta)$ with a dashed blue line, and the true negative log-likelihood $J(\theta)$ with a solid red line. Figure 2 depicts the same curves and results but considering the second proposal $q_2(y|\theta)$.

In Figure 1, as $M \to \infty$ we can observe the converge of $J_{\texttt{BL}}(\theta)$ to $J(\theta)$ specially for values of $\theta < 2$, but the convergence struggles for values of $\theta > 2$, even with $M = 5000$. This is due to the choice of the proposal, which is not suitable for value of $\theta > 2$. Whereas, in Figure 2, the green area (i.e., the variability of $J_{\texttt{BL}}(\theta)$) is much more smaller than in in Figure 1, even with $M = 100$. For $M = 5000$, we have almost a perfect convergence. The reason is the $q_2(y|\theta)$ has a standard deviation, $2 + \theta$, always greater than the standard deviation of the model, that is 2 [10]. Therefore, $q_2(y|\theta)$ provides much better estimator $\widehat{\theta}$ in terms of mean square error, since it provides more efficient estimations of $\widehat{Z}(\theta)$.



(a) $N = M = 100$ (b) $N = M = 100$ (c) $N = 100, M = 5000$

**Fig. 1.** (a) Approximation of the negative log-likelihood $J(\theta)$ (shown with a solid red line) using the first proposal with $\mu_p = 2$ and $\sigma_p = 2$. (b) and (c) The green area shows the 100% of variability of $J_{\mathtt{BL}}(\theta)$. The empirical mean curve is depicted with a blue dashed line.



(a) $N = M = 100$ (b) $N = 100$, $M = 5000$

**Fig. 2.** Approximation of the negative log-likelihood $J(\theta)$ (shown with a solid red line) using the second proposal with $\mu_p = 2$ and $\sigma_p = 2 + \theta$ for $N = M = 100$ (a) and $N = M = 5000$ (b) . The green area shows the 100% of variability of $J_{\mathtt{BL}}(\theta)$. The empirical mean curve is depicted with a blue dashed line.

## 8 Conclusions

We have provided a detailed description of the use of gradient-based approaches for parameter estimation in EBMs. Different complete derivations are described discussing connections and relationships. We have given several practical suggestions for the choice of a proposal/reference density in other to ensure a suitable approximation of the negative log-likelihood and its gradient. We have also remarked several possible relevant issues that have been in some way overlooked (or missed) in the literature. Nevertheless, as said in Section 1, other approaches are available in the literature to estimate EBMs, such as the noise contrastive estimation and the score matching. As future work, we plan to study the connections with these possible other schemes. We finally remark that similar issues hold also modeling data characterized by heterogeneous sub-groups [12] propose a finite mixture of non-normalized densities, which is specified by

$$p(\cdot|\boldsymbol{\psi}) = \sum_{g=1}^{G} \frac{\pi_g p_g(\cdot|\boldsymbol{\theta}_k)}{Z(\boldsymbol{\theta}_g)} = \sum_{g=1}^{G} \xi_g p_g(\cdot|\boldsymbol{\theta}_g), \tag{26}$$

where $G$ is the number of components of the mixture, $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_G\}$ are the mixing weights which sum to one, $\sum_{g=1}^{G} \pi_g = 1$.

**Acknowledgements.** S. Ingrassia and L. Scaffidi Domianello acknowledge financial support from PNRR MUR project PE0000013-FAIR. L. Martino acknowledges financial support by the PIACERI Starting Grant BA-GRAPH (UPB 28722052144) of the University of Catania. The authors thank also Michael U. Gutmann for fruitful discussion during the preparation of the paper.

## References

1. Carreira-Perpiñán, M. A. and Hinton, G. E. (2005). On contrastive divergence learning. In *Tenth International Workshop on Artificial Intelligence and Statistics, Barbados.*
2. Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **56**(1), 261–274.
3. Gutmann, M. U. and Hyvärinen, A. (2012). Noise contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, **13**(2).
4. Gutmann, M. U. and Hyvärinen, A. (2013). Estimation of unnormalized statistical models without numerical integration. In *Proc. Workshop on Information Theoretic Methods in Science and Engineering ((WITMSE2013))*.
5. Gutmann, M. U., Kleinegesse, S., and Rhodes, B. (2022). Statistical applications of contrastive learning. *Behaviormetrika*, **49**(2), 277–301.
6. Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, **14**(8), 1771–1800.
7. Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**.
8. Le Cun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huanh, G. J. (2007). Energy-based models. In *Predicting Structured Data*, pages 191–246. MIT Press, Cambridge, Massachusetts.
9. Llorente, F. and Martino, L. (2025). Optimality in importance sampling: a gentle survey. *arXiv:2502.07396*.
10. Llorente, F., Martino, L., Delgado, D., and López-Santiago, J. (2023). Marginal likelihood computation for model selection and hypothesis testing: An extensive review. *SIAM Review*, **65**(1), 3–58.
11. Martino, L. and Elvira, V. (2017). *Metropolis Sampling*, pages 1–18. John Wiley & Sons, Ltd.
12. Matsuda, T. and Hyvärinen, A. (2019). Estimation of non-normalized mixture models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2555–2563. PMLR.