UPPSALA
UNIVERSITET

Learning Proposals for Practical Energy-Based Regression

Fredrik K. Gustafsson
Uppsala University

AISTATS 2022
March 14, 2022

Background: Energy-Based Models

UPPSALA UNIVERSITET

An **energy-based model (EBM)** specifies a probability distribution $p(x;\theta)$ over $x \in \mathcal{X}$ directly via a parameterized scalar function $f_\theta : \mathcal{X} \to \mathbb{R}$:

$$p(x;\theta) = \frac{e^{f_\theta(x)}}{Z(\theta)}, \quad Z(\theta) = \int e^{f_\theta(\tilde{x})} d\tilde{x}$$

By defining $f_\theta(x)$ using a deep neural network (DNN), the EBM $p(x;\theta)$ becomes expressive enough to learn practically any distribution from observed data.

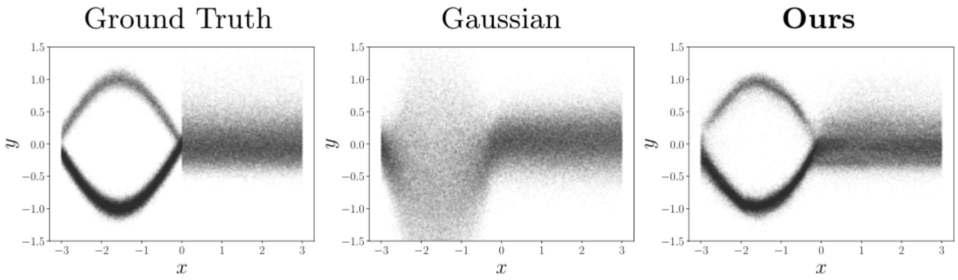
Drawback: the normalizing partition function $Z(\theta) = \int e^{f_\theta(\tilde{x})} d\tilde{x}$ is intractable, which complicates evaluating or sampling from the EBM $p(x;\theta)$.

Compare with normalizing flow models which are specifically designed to be easy to both evaluate and sample. EBMs instead prioritize maximum model expressivity.

Energy-Based Regression: train a DNN $f_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to predict a scalar value $f_\theta(x, y)$, then model $p(y|x)$ with the *conditional* EBM $p(y|x; \theta)$:

$$p(y|x;\theta) = \frac{e^{f_\theta(x,y)}}{Z(x,\theta)}, \quad Z(x,\theta) = \int e^{f_\theta(x,\tilde{y})} d\tilde{y}.$$

The EBM $p(y|x;\theta)$ can learn complex distributions $p(y|x)$ directly from data:



Background: Energy-Based Regression - Training

Energy-Based Regression: train a DNN $f_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to predict a scalar value $f_\theta(x, y)$, then model $p(y|x)$ with the *conditional* EBM $p(y|x; \theta)$:

$$p(y|x;\theta) = \frac{e^{f_\theta(x,y)}}{Z(x,\theta)}, \quad Z(x,\theta) = \int e^{f_\theta(x,\tilde{y})} d\tilde{y}.$$

The DNN $f_\theta(x, y)$ can be trained using various methods for fitting a distribution $p(y|x; \theta)$ to observed data $\{(x_i, y_i)\}_{i=1}^N$.

Energy-Based Regression: train a DNN $f_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to predict a scalar value $f_\theta(x, y)$, then model $p(y|x)$ with the *conditional* EBM $p(y|x; \theta)$:

$$p(y|x;\theta) = \frac{e^{f_\theta(x,y)}}{Z(x,\theta)}, \quad Z(x,\theta) = \int e^{f_\theta(x,\tilde{y})} d\tilde{y}.$$

The most straightforward training method is probably to approximate the negative log-likelihood $\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p(y_i|x_i;\theta)$ using importance sampling:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{1}{M}\sum_{m=1}^{M} \frac{e^{f_\theta(x_i, y_i^{(m)})}}{q(y_i^{(m)})}\right) - f_\theta(x_i, y_i),$$

$$\{y_i^{(m)}\}_{m=1}^{M} \sim q(y) \text{ (proposal distribution)}.$$

Background: Energy-Based Regression - Training

Energy-Based Regression: train a DNN $f_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ to predict a scalar value $f_\theta(x, y)$, then model $p(y|x)$ with the *conditional* EBM $p(y|x;\theta)$:

$$p(y|x;\theta) = \frac{e^{f_\theta(x,y)}}{Z(x,\theta)}, \quad Z(x,\theta) = \int e^{f_\theta(x,\tilde{y})} d\tilde{y}.$$

Previous work has also employed noise contrastive estimation (NCE):

$$J_{\text{NCE}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} J_{\text{NCE}}^{(i)}(\theta), \quad J_{\text{NCE}}^{(i)}(\theta) = \log \frac{\exp\{f_\theta(x_i, y_i^{(0)}) - \log q(y_i^{(0)})\}}{\sum_{m=0}^{M} \exp\{f_\theta(x_i, y_i^{(m)}) - \log q(y_i^{(m)})\}},$$

$$y_i^{(0)} \triangleq y_i, \quad \{y_i^{(m)}\}_{m=1}^{M} \sim q(y) \text{ (noise distribution)}.$$

Practical Limitations of Energy-Based Regression

In previous work, the proposal/noise distribution $q(y)$ was set to a mixture of K Gaussian components centered at the true regression target y_i,

$$q(y) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(y; y_i, \sigma_k^2 I).$$

$q(y)$ contains task-dependent hyperparameters K and $\{\sigma_k^2\}_{k=1}^{K}$.

$q(y)$ depends on the true target y_i and can thus only be utilized during training.

We address both these limitations by jointly learning a parameterized proposal/noise distribution $q(y|x;\phi)$ during EBM training.

We derive an efficient and convenient objective that can be employed to train $q(y|x;\phi)$ by directly minimizing its KL divergence to the EBM $p(y|x;\theta)$.

We want $q(y|x;\phi)$ to be a close approximation of the EBM $p(y|x;\theta)$. Specifically, we want to find ϕ that minimizes the KL divergence between $q(y|x;\phi)$ and $p(y|x;\theta)$.

Therefore, we seek to compute $\nabla_\phi D_{\mathrm{KL}}\big(p(y|x;\theta) \parallel q(y|x;\phi)\big)$. The gradient $\nabla_\phi D_{\mathrm{KL}}$ is generally intractable, but can be conveniently approximated by the following result:

Result 1: For a conditional EBM $p(y|x;\theta) = e^{f_\theta(x,y)} / \int e^{f_\theta(x,\tilde{y})} d\tilde{y}$ and distribution $q(y|x;\phi)$,

$$\nabla_\phi D_{\mathrm{KL}}\big(p \parallel q\big) \approx \nabla_\phi \log \left(\frac{1}{M} \sum_{m=1}^{M} \frac{e^{f_\theta(x,y^{(m)})}}{q(y^{(m)}|x;\phi)} \right),$$

where $\{y^{(m)}\}_{m=1}^{M}$ are M independent samples drawn from $q(y|x;\phi)$.

Result 1: For a conditional EBM $p(y|x;\theta) = e^{f_\theta(x,y)} / \int e^{f_\theta(x,\tilde{y})} d\tilde{y}$ and distribution $q(y|x;\phi)$,

$$\nabla_\phi D_{\mathrm{KL}}\big(p \parallel q\big) \approx \nabla_\phi \log\left(\frac{1}{M}\sum_{m=1}^{M} \frac{e^{f_\theta(x,y^{(m)})}}{q(y^{(m)}|x;\phi)}\right),$$

where $\{y^{(m)}\}_{m=1}^{M}$ are M independent samples drawn from $q(y|x;\phi)$.

Given data $\{x_i\}_{i=1}^{N}$, Result 1 implies that the proposal/noise distribution $q(y|x;\phi)$ can be trained to approximate the EBM $p(y|x;\theta)$ by minimizing the loss,

$$J_{\mathrm{KL}}(\phi) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{1}{M}\sum_{m=1}^{M} \frac{e^{f_\theta(x_i,y_i^{(m)})}}{q(y_i^{(m)}|x_i;\phi)}\right),$$

$$\{y_i^{(m)}\}_{m=1}^{M} \sim q(y|x_i;\phi).$$

Joint Training Method

Given data $\{x_i\}_{i=1}^N$, Result 1 implies that the proposal/noise distribution $q(y|x;\phi)$ can be trained to approximate the EBM $p(y|x;\theta)$ by minimizing the loss,

$$J_{\mathrm{KL}}(\phi) = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{M}\sum_{m=1}^{M}\frac{e^{f_\theta(x_i, y_i^{(m)})}}{q(y_i^{(m)}|x_i;\phi)}\right),$$

$$\{y_i^{(m)}\}_{m=1}^{M} \sim q(y|x_i;\phi).$$

Since $J_{\mathrm{KL}}(\phi)$ is identical to the first term of the EBM loss $J(\theta)$ from previous work,

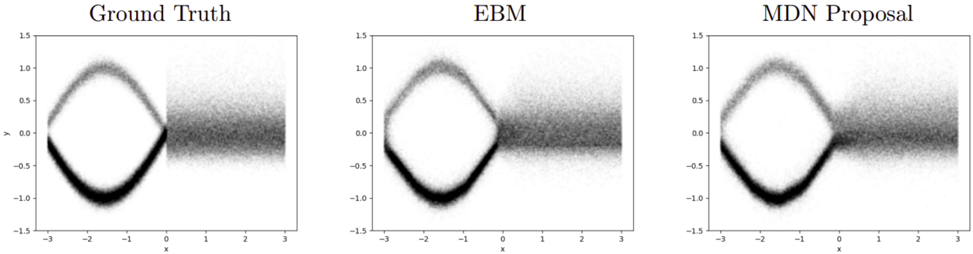
$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{M}\sum_{m=1}^{M}\frac{e^{f_\theta(x_i, y_i^{(m)})}}{q(y_i^{(m)})}\right) - f_\theta(x_i, y_i), \tag{1}$$

the EBM $p(y|x;\theta)$ and proposal/noise distribution $q(y|x;\phi)$ can be trained by jointly minimizing (1) w.r.t. both θ and ϕ.

$$J_{\mathrm{KL}}(\phi) = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{M}\sum_{m=1}^{M}\frac{e^{f_\theta(x_i, y_i^{(m)})}}{q(y_i^{(m)}|x_i;\phi)}\right),$$

$$\{y_i^{(m)}\}_{m=1}^{M} \sim q(y|x_i;\phi).$$

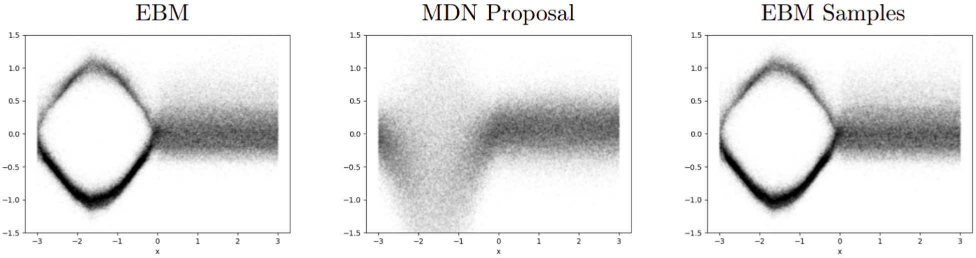
The EBM $p(y|x;\theta)$ and proposal/noise distribution $q(y|x;\phi)$ can also be jointly trained by updating ϕ via $J_{\mathrm{KL}}(\phi)$, and updating θ via $J_{\mathrm{NCE}}(\theta)$.



Utilizing the Proposal

As $q(y|x;\phi)$ has been trained to approximate the EBM $p(y|x;\theta)$, it can be utilized with self-normalized importance sampling to e.g. compute the EBM mean at test-time, thus producing a stand-alone prediction $y^\star$.

The proposal $q(y|x;\phi)$ can also be used to draw approximate samples from the EBM:



Contact

UPPSALA UNIVERSITET

Fredrik K. Gustafsson, Uppsala University

fredrik.gustafsson@it.uu.se

www.fregu856.com