Your classifier is secretly an EBM and you should treat it like one

1. EBM Definition Recap

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z(\theta)}$$

ensures $p_\theta(x) \geq 0$

flexible in terms of model architecture

ensures $\int p_\theta(x)\,dx = 1$

more Facts?

$$Z(\theta) := \int \exp(-E_\theta(x))\,dx$$

$$E_\theta : \mathbb{R}^D \to \mathbb{R}$$

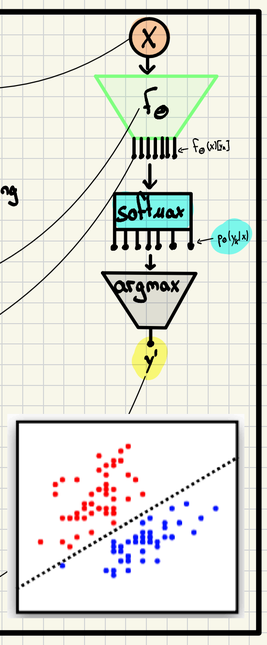
2. Classification Recap

- $x \in \mathbb{R}^D$, $y_i \in \{y_1, \dots, y_K\}$
  we have some data $\subseteq \mathbb{R}^D \times \{y_1, \dots, y_K\}$
- Goal:
  estimate $p(y_i|x)$ with $p_\theta(y_i|x)$ by maximizing the likelihood of the data.
- For that we use a mapping
  $f_\theta : \mathbb{R}^D \to \mathbb{R}^K$ and
- pass it through a softmax:

$$p_\theta(y_i|x) = \frac{\exp(f_\theta(x)[i])}{\sum_j \exp(f_\theta(x)[j])}$$

- Take the argmax as the class:

$$y' = \underset{\in\{y_1,\dots,y_K\}}{\operatorname{argmax}}\ p_\theta(y_i|x)$$



3. A different perspective:
- Redefine the meaning of embeddings

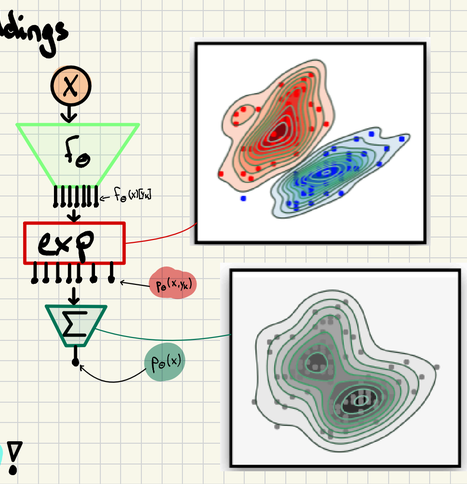
- EBM: $p_\theta(x, y_i) = \frac{\exp(f_\theta(x)[y_i])}{Z(\theta)}$
- with Energy $E_\theta(x, y_i) = -f_\theta(x)[y_i]$
- Aggregation over classes gives us the marginal: an EBM for $p(x)$:

$$p_\theta(x) = \sum_{y \in Y} p_\theta(x, y_i) = \frac{1}{Z(\theta)} \sum_{y \in Y} \exp(f_\theta(x)[y_i])$$

with energy $E_\theta(x) = -\log \sum_{y_i \in Y} \exp(f_\theta(x)[y_i])$

- Conditioning on x recovers the softmax!

$$p_\theta(y_i|x) = \frac{p_\theta(x, y_i)}{p_\theta(x)} = \frac{\frac{\exp(f_\theta(x)[y_i])}{Z(\theta)}}{\frac{1}{Z(\theta)} \sum_{y_i \in Y} \exp(f_\theta(x)[y_i])} = \frac{\exp(f_\theta(x)[y_i])}{\sum_{y_i \in Y} \exp(f_\theta(x)[y_i])}$$



4. Insight: We have 3 EBMs hidden in the classifier.

1. The obvious conditional $p_\theta(y_i|x)$
2. The joint $p_\theta(x, y_i)$
3. The marginal $p_\theta(x)$

(actually K conditionals and joint models)

5. How 2 train?

$$\frac{p_\theta(x, y_i)}{p_\theta(x)} = p_\theta(y_i|x)$$

$$\log p_\theta(x, y_i) - \log p_\theta(x) = \log p_\theta(y_i|x)$$

Vanilla Cross-Entropy Loss

$$\text{Loss}(x, y_i, \theta) = -\log p_\theta(x, y_i) = -\left(\log p_\theta(y_i|x) + \log p_\theta(x)\right)$$

Estimate using Stochastic Gradient Langevin Dynamics

6. Stochastic Gradient Langevin Dynamics (SGLD)

Recap:

$$p_\theta(x) = \sum_{y_i \in Y} p_\theta(x, y_i) = \frac{1}{Z(\theta)} \sum_{y_i \in Y} \exp(f_\theta(x)[y_i])$$

$$E_\theta(x) = -\log \sum_{y_i \in Y} \exp(f_\theta(x)[y_i])$$

Derivative of the log-likelihood for a single example x' with resp. to the parameters θ can be expressed as:

$$\frac{\partial \log p_\theta(x')}{\partial \theta} = E_{p_\theta(x)}\left[\frac{\partial E_\theta(x)}{\partial \theta}\right] - \frac{\partial E_\theta(x')}{\partial \theta}$$

intractable → MCMC-Sampling with SGLD + Persistent Contrastive (PCD) Divergence for $p_\theta(x)$ for faster sampling convergence

SGLD Sampling

(1) $x_0 \sim p_\theta(x)$ (which is typically a uniform dist over the input domain)

(2) $\alpha_0 \in \mathbb{R}^+$ (which should follow a polynomial decay schedule)

(3) for some iterations do:

(3.1) $\varepsilon_i \sim N(0, \alpha_i I)$

(3.2) $x_{i+1} = x_i - \frac{\alpha_i}{2} \frac{\partial E_\theta(x_i)}{\partial x_i} + \varepsilon_i$

(3.3) $\alpha_{i+1} = \text{PolynomialDecaySchedule}(\alpha_i)$

PCD
but with PCD consists of a fraction $B \in (0,1)$ of the samples from the last step and $(1-B)$ fraction of uniform noise.

Implementation Demo

EBM Definition Recap

$$p_\theta(x) = \frac{\exp(-E_\theta(x))}{Z(\theta)}$$

ensures $p_\theta(x) \geq 0$ · flexible in terms of model architecture · ensures $\int p_\theta(x)\,dx = 1$ · more Facts?

$$Z(\theta) := \int \exp(-E_\theta(x))\,dx$$

$$E_\theta : \mathbb{R}^D \to \mathbb{R}$$

Mixture of Gaussians as an EBM

$$p_{\theta_i}(x) = \frac{\exp\left(-\frac{(x-\mu_i)^2}{2}\right)}{\sqrt{2\pi}}$$

$$p_\theta(x) = \frac{1}{N} \sum_{i=1}^{N} p_{\theta_i} = \frac{1}{N\sqrt{2\pi}} \sum_{i=1}^{N} \exp\left(-\frac{(x-\mu_i)^2}{2}\right)$$

$$= \frac{1}{N\sqrt{2\pi}} \exp\left((-1)\left(-\log \sum_{i=1}^{N} \exp\left(-\frac{(x-\mu_i)^2}{2}\right)\right)\right)$$

$$Z(\theta) = N\sqrt{2\pi}$$

$$E_\theta(x) = -\log \sum_{i=1}^{N} \exp\left(-\frac{(x-\mu_i)^2}{2}\right)$$

7. Benefits

- Hybrid Model for Classification and Conditional Generation
- Robustness against adversarial attacks
- Better Calibration
  = Alignment of predictive confidence and classification rate