1 Energy-based Model

Definition 1. *[Energy-based Model]*

Let $\mathcal{M}$ a measure space, and $E: \mathbb{R}^m \to (\mathcal{M} \to \mathbb{R})$. Then define probabilitic model based on E as

$$p(x;\theta) = \frac{\exp(-E(x;\theta))}{\int_{\mathcal{M}} \mathrm{d}x' \exp(-E(x';\theta))},$$

where $\theta \in \mathbb{R}^m$ and $x \in \mathcal{M}$.

We call this an energy-based model, where $E(\cdot;\theta)$ is called a energy function parameterized by θ.

Theorem 2. *[Universality]*

For any probability density $q: \mathcal{M} \to \mathbb{R}$ and for $\forall C \in \mathbb{R}$, define, for $\forall x \in \operatorname{supp}(q)$,

$$E_q(x) := -\ln q(x) + C,$$

then, for $\forall x \in \operatorname{supp}(q)$,

$$q(x) = \frac{\exp(-E_q(x))}{\int_{\operatorname{supp}(q)} \mathrm{d}x' \exp(-E_q(x'))}.$$

That is, for any probability density, there exists an energy function (up to constant) that can describe the probability density.

Theorem 3. *[Activity Rule]*

The local maximum of $p(\cdot;\theta)$ is the local minimum of $E(\cdot;\theta)$, and vice versa.

Theorem 4. *[Learning Rule]*

For any probability density $p_D: \mathcal{M} \to \mathbb{R}$, define Lagrangian $L(\theta; p_D) := -\int_{\mathcal{M}} \mathrm{d}x \, p_D(x) \ln p(x;\theta)$. Then, the gradient of Lagrangian w.r.t. component θ^α is

$$\frac{\partial L}{\partial \theta^\alpha}(\theta; p_D) = \int_{\mathcal{M}} \mathrm{d}x \, p_D(x) \frac{\partial E}{\partial \theta^\alpha}(x;\theta) - \int_{\mathcal{M}} \mathrm{d}x \, p(x;\theta) \frac{\partial E}{\partial \theta^\alpha}(x;\theta),$$

or in more compact format,

$$\frac{\partial L}{\partial \theta^\alpha}(\theta; p_D) = \mathbb{E}_{x \sim p_D}\left[\frac{\partial E}{\partial \theta^\alpha}(x;\theta)\right] - \mathbb{E}_{x \sim p(x;\theta)}\left[\frac{\partial E}{\partial \theta^\alpha}(x;\theta)\right].$$

2 Effective Theory

Definition 5. *[Effective Energy]*

Suppose exists $(\mathcal{V}, \mathcal{H})$, s.t. $\mathcal{M} = \mathcal{V} \oplus \mathcal{H}$. Re-denote $E(x;\theta) \to E(v,h;\theta)$ and $p(x;\theta) \to p(v,h;\theta)$. Then, define effective energy $E_{\text{eff}}: \mathcal{V} \to \mathbb{R}$ as

$$E_{\text{eff}}(v;\theta) := -\ln \int_{\mathcal{H}} \mathrm{d}h \exp(-E(v,h;\theta)).$$

Theorem 6. *[Effective Theory]*

Recall that $p(v;\theta) := \int_{\mathcal{H}} \mathrm{d}h \, p(v,h;\theta)$. Then,

$$p(v;\theta) = \frac{\exp(-E_{\text{eff}}(v;\theta))}{\int_{\mathcal{V}} \mathrm{d}v' \exp(-E_{\text{eff}}(v';\theta))}.$$

Lemma 7. *[Gradient of Effective Energy]*

$$\frac{\partial E_{\text{eff}}}{\partial \theta^\alpha}(v,\theta) = \int_{\mathcal{H}} \mathrm{d}h \, p(h|v;\theta) \frac{\partial E}{\partial \theta^\alpha}(v,h;\theta).$$

Theorem 8. *[Learning Rule of Effective Theory]*

For any probability density $p_D\colon \mathcal{V} \to \mathbb{R}$, define Lagrangian $L(\theta; p_D) := -\int_{\mathcal{V}} \mathrm{d}v\, p_D(v) \ln p(v; \theta)$. Then, the gradient of Lagrangian w.r.t. component θ^α is

$$\frac{\partial L}{\partial \theta^\alpha}(\theta; p_D) = \int_{\mathcal{V}} \mathrm{d}v \int_{\mathcal{H}} \mathrm{d}h\, p_D(v)\, p(h|v;\theta)\, \frac{\partial E}{\partial \theta^\alpha}(v, h; \theta) - \int_{\mathcal{V}} \mathrm{d}v \int_{\mathcal{H}} \mathrm{d}h\, p(v, h; \theta)\, \frac{\partial E}{\partial \theta^\alpha}(v, h; \theta),$$

or in more compact format,

$$\frac{\partial L}{\partial \theta^\alpha}(\theta; p_D) = \mathbb{E}_{v \sim p_D, h \sim p(h|v;\theta)}\left[\frac{\partial E}{\partial \theta^\alpha}(v, h; \theta)\right] - \mathbb{E}_{v, h \sim p(v,h;\theta)}\left[\frac{\partial E}{\partial \theta^\alpha}(v, h; \theta)\right].$$

3 Examples

Example 9. [Boltzmann Machine]

- Let $\mathcal{M} = \{0,1\}^n$, $W \in \mathbb{R}^{(n \times n)}$, $b \in \mathbb{R}^n$, $\theta := (W, b)$. Then a Boltzmann machine is defined by energy function
$$E(x; W, b) := -(1/2) \sum_{\alpha, \beta \neq \alpha} W_{\alpha\beta}\, x^\alpha\, x^\beta - \sum_\alpha b_\alpha\, x^\alpha.$$
- Direct calculation gives
$$p(x_\alpha = 1 | x_{\setminus \alpha}) = \sigma\left(\sum_{\beta \neq \alpha} W_{\alpha\beta}\, x^\beta + b_\alpha\right).$$

Example 10. [Restricted Boltzmann Machine]

- Let $\mathcal{V} = \{0,1\}^n$ and $\mathcal{H} = \{0,1\}^m$. Let $W \in \mathbb{R}^{(n \times m)}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}^m$, $\theta := (W, b, c)$. Then a restricted Boltzmann machine is defined by energy function
$$E(v, h; W, b, c) := -(1/2) \sum_{\alpha, \beta \neq \alpha} W_{\alpha\beta}\, v^\alpha\, h^\beta - \sum_\alpha b_\alpha\, v^\alpha - \sum_\alpha c_\alpha\, h^\alpha.$$
- Direct calculation gives
$$E_{\mathrm{eff}}(v; W, b, c) = -\sum_\alpha b_\alpha\, v^\alpha - \sum_\beta s_+\left(\sum_\alpha W_{\alpha\beta}\, v^\alpha + c_\beta\right),$$
where s_+ represents soft-plus function.