

## 1 Energy-based Model

### Definition 1. [Energy-based Model]

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

$$p_E(x; \theta) = \frac{\exp(-E(x; \theta))}{\int_{\mathcal{M}} dx' \exp(-E(x'; \theta))}, \quad (1)$$

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  $\theta$ .

### Theorem 2. [Universality]

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

$$E_q(x) := -\ln q(x) + C, \quad (2)$$

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

$$q(x) = \frac{\exp(-E_q(x))}{\int_{\text{supp}(q)} dx' \exp(-E_q(x'))}. \quad (3)$$

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

### Theorem 3. [Maximum Entropy Principle]

For any probability density  $p_D: \mathcal{M} \rightarrow \mathbb{R}$ , we have

$$p_E(x) = \text{argmax}_p H[X], \quad (4)$$

s.t. constrains

$$\mathbb{E}_{x \sim p_D} \left[ \frac{\partial E}{\partial \theta^\alpha}(x; \theta) \right] = \mathbb{E}_{x \sim p} \left[ \frac{\partial E}{\partial \theta^\alpha}(x; \theta) \right] \quad (5)$$

are satisfied.

### Theorem 4. [Activity Rule]

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

### Theorem 5. [Learning Rule]

For any probability density  $p_D: \mathcal{M} \rightarrow \mathbb{R}$ , define Lagrangian  $L(\theta; p_D) := -\int_{\mathcal{M}} dx p_D(x) \ln p_E(x; \theta)$ . Then, the gradient of Lagrangian w.r.t. component  $\theta^\alpha$  is

$$\frac{\partial L}{\partial \theta^\alpha}(\theta; p_D) = \int_{\mathcal{M}} dx p_D(x) \frac{\partial E}{\partial \theta^\alpha}(x; \theta) - \int_{\mathcal{M}} dx p_E(x; \theta) \frac{\partial E}{\partial \theta^\alpha}(x; \theta), \quad (6)$$

or in more compact format,

$$\frac{\partial L}{\partial \theta^\alpha}(\theta; p_D) = \mathbb{E}_{x \sim p_D} \left[ \frac{\partial E}{\partial \theta^\alpha}(x; \theta) \right] - \mathbb{E}_{x \sim p(x; \theta)} \left[ \frac{\partial E}{\partial \theta^\alpha}(x; \theta) \right]. \quad (7)$$

## 2 Effective Theory

### Definition 6. [Effective Energy]

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

$$E_{\text{eff}}(v; \theta) := -\ln \int_{\mathcal{H}} dh \exp(-E(v, h; \theta)). \quad (8)$$

**Theorem 7.** [Effective Theory]

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

$$p(v; \theta) = \frac{\exp(-E_{\text{eff}}(v; \theta))}{\int_{\mathcal{V}} dv' \exp(-E_{\text{eff}}(v'; \theta))}. \quad (9)$$

**Lemma 8.** [Gradient of Effective Energy]

$$\frac{\partial E_{\text{eff}}}{\partial \theta^\alpha}(v, \theta) = \int_{\mathcal{H}} dh p(h|v; \theta) \frac{\partial E}{\partial \theta^\alpha}(v, h; \theta). \quad (10)$$

**Theorem 9.** [Learning Rule of Effective Theory]

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

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

or in more compact format,

$$\frac{\partial L}{\partial \theta^\alpha}(\theta; p_D) = \mathbb{E}_{v \sim p_D, h \sim p(h|v; \theta)} \left[ \frac{\partial E}{\partial \theta^\alpha}(v, h; \theta) \right] - \mathbb{E}_{v, h \sim p(v, h; \theta)} \left[ \frac{\partial E}{\partial \theta^\alpha}(v, h; \theta) \right]. \quad (11)$$

### 3 Examples

**Example 10.** [Boltzmann Machine]

- Let  $\mathcal{M} = \{0, 1\}^n$ ,  $W \in \mathbb{R}^{(n \times n)}$ ,  $b \in \mathbb{R}^n$ ,  $\theta := (W, b)$ . Given dataset  $D := \{x_i | x_i \in \mathcal{M}, i = 1, \dots, N\}$ , denote expectation as  $\hat{x}^\alpha$ . Then a Boltzmann machine is defined by energy function

$$E(x; W, b) := -\frac{1}{2} \sum_{\alpha, \beta} W_{\alpha\beta} (x^\alpha - \hat{x}^\alpha) (x^\beta - \hat{x}^\beta) - \sum_{\alpha} b_{\alpha} x^{\alpha}. \quad (12)$$

- Direct calculation gives, for  $\forall \alpha$ ,

$$p(x_{\alpha} = 1 | x_{\setminus \alpha}) = \sigma \left( \sum_{\alpha \neq \beta} W_{(\alpha\beta)} (x^{\beta} - \hat{x}^{\beta}) + c_{\alpha} \right), \quad (13)$$

where  $W_{(\alpha\beta)} := (W_{\alpha\beta} + W_{\beta\alpha})/2$  and  $c_{\alpha} := b_{\alpha} + (1/2 - \hat{x}^{\alpha})W_{\alpha\alpha}$ . This relation is held even for arbitrary vector  $\hat{x}$ .

- Relating to MaxEnt principle, the observable that the model simulates is

$$\forall (\alpha, \beta), \mathbb{E}_{x \sim P_D} [(x^{\alpha} - \hat{x}^{\alpha})(x^{\beta} - \hat{x}^{\beta})], \quad (14)$$

for which it shall also simulate

$$\forall \alpha, \mathbb{E}_{x \sim P_D} [\hat{x}^{\alpha}]. \quad (15)$$

**Proof.** Here we proof the activity rule.

Directly, for  $\forall \gamma$ ,

$$\begin{aligned}
& \ln p(x_\gamma = 1 | x_{\setminus \gamma}) - \ln p(x_\gamma = 0 | x_{\setminus \gamma}) \\
& \{\alpha = \beta = \gamma\} = \frac{1}{2} W_{\gamma\gamma} (1 - \hat{x}^\gamma)^2 - \frac{1}{2} W_{\gamma\gamma} (-\hat{x}^\gamma)^2 \\
& \{\alpha \neq \gamma, \beta = \gamma\} + \frac{1}{2} \sum_{\alpha \neq \gamma} W_{\alpha\gamma} (x^\alpha - \hat{x}^\alpha) (1 - \hat{x}^\gamma) - \frac{1}{2} \sum_{\alpha \neq \gamma} W_{\alpha\gamma} (x^\alpha - \hat{x}^\alpha) (-\hat{x}^\gamma) \\
& \{\alpha = \gamma, \beta \neq \gamma\} + \frac{1}{2} \sum_{\beta \neq \gamma} W_{\gamma\beta} (1 - \hat{x}^\gamma) (x^\beta - \hat{x}^\beta) - \frac{1}{2} \sum_{\beta \neq \gamma} W_{\gamma\beta} (-\hat{x}^\gamma) (x^\beta - \hat{x}^\beta) \\
& \{\alpha, \beta \neq \gamma\} + \frac{1}{2} \sum_{\alpha, \beta \neq \gamma} W_{\alpha\beta} (x^\alpha - \hat{x}^\alpha) (x^\beta - \hat{x}^\beta) - \frac{1}{2} \sum_{\alpha, \beta \neq \gamma} W_{\alpha\beta} (x^\alpha - \hat{x}^\alpha) (x^\beta - \hat{x}^\beta) \\
& \{\alpha = \gamma\} + b^\gamma - 0 \\
& \{\alpha \neq \gamma\} + \sum_{\alpha \neq \gamma} b_\gamma x^\gamma - \sum_{\alpha \neq \gamma} b_\gamma x^\gamma \\
& = \frac{1}{2} W_{\gamma\gamma} - W_{\gamma\gamma} \hat{x}^\gamma \\
& \quad + \frac{1}{2} \sum_{\alpha \neq \gamma} W_{\alpha\gamma} (x^\alpha - \hat{x}^\alpha) \\
& \quad + \frac{1}{2} \sum_{\beta \neq \gamma} W_{\gamma\beta} (x^\beta - \hat{x}^\beta) \\
& \quad + 0 \\
& \quad + b_\gamma \\
& \quad + 0 \\
& = \left( \frac{1}{2} - \hat{x}^\gamma \right) W_{\gamma\gamma} + \sum_{\alpha \neq \gamma} W_{(\gamma\alpha)} (x^\alpha - \hat{x}^\alpha) + b_\gamma
\end{aligned}$$

Thus

$$p(x_\gamma = 1 | x_{\setminus \gamma}) = \sigma \left[ \sum_{\alpha \neq \gamma} \frac{1}{2} (W_{\alpha\gamma} + W_{\gamma\alpha}) (x^\alpha - \hat{x}^\alpha) + \left( b_\gamma + \left( \frac{1}{2} - \hat{x}^\gamma \right) W_{\gamma\gamma} \right) \right]. \quad \square$$

**Example 11.** [Restricted Boltzmann Machine]

- Let  $\mathcal{V} = \{0, 1\}^n$  and  $\mathcal{H} = \{0, 1\}^m$ ,  $\mathcal{M} = \mathcal{V} \times \mathcal{H}$ . Let  $U \in \mathbb{R}^{(n \times m)}$ ,  $b \in \mathbb{R}^n$ ,  $c \in \mathbb{R}^m$ . Then a restricted Boltzmann machine is defined by energy function

$$E(v, h; U, b, c) := - \sum_{\alpha, \beta} U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha) (h^\beta - \hat{h}^\beta) - \sum_{\alpha} b_\alpha v^\alpha - \sum_{\alpha} c_\alpha h^\alpha. \quad (16)$$

- Direct calculation gives

$$E_{\text{eff}}(v; U, b, c) = \sum_{\alpha} \left( \sum_{\beta} U_{\alpha\beta} v^\alpha \hat{h}^\beta - b_\alpha \right) - s_+ \left( \sum_{\alpha} U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha) + c_\beta \right), \quad (17)$$

where soft-plus  $s$  is defined as

$$s(x) := \ln(1 + e^x). \quad (18)$$

- Relating to Boltzmann machine,  $x \rightarrow (v, h)$ ,  $b \rightarrow (b, c)$ , and

$$W \rightarrow \begin{pmatrix} 0 & U \\ U^T & 0 \end{pmatrix}. \quad (19)$$

**Proof.** Here we proof the expression of  $E_{\text{eff}}(v; U, b, c)$ .

Directly,

$$\begin{aligned}
E_{\text{eff}}(v) \\
\{\text{Definition}\} &= -\ln \sum_h \exp(-E(v, h)) \\
\{\text{Definition}\} &= -\ln \sum_h \exp\left(\sum_{\alpha, \beta} U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha)(h^\beta - \hat{h}^\beta) + \sum_\alpha b_\alpha v^\alpha + \sum_\alpha c_\alpha h^\alpha\right) \\
\{\text{Combine}\} &= -\sum_\alpha b_\alpha v^\alpha - \ln \sum_h \exp\left[\sum_\beta \left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha)\right)(h^\beta - \hat{h}^\beta) + \sum_\beta c_\beta h^\beta\right] \\
\{\exp \sum = \prod \exp\} &= -\sum_\alpha b_\alpha v^\alpha - \ln \prod_\beta \sum_{h^\beta=0,1} \exp\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha)(h^\beta - \hat{h}^\beta) + c_\beta h^\beta\right) \\
\{\ln \prod = \sum \ln\} &= -\sum_\alpha b_\alpha v^\alpha - \sum_\beta \ln \sum_{h^\beta=0,1} \exp\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha)(h^\beta - \hat{h}^\beta) + c_\beta h^\beta\right).
\end{aligned}$$

Since

$$\begin{aligned}
&\sum_{h^\beta=0,1} \exp\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha)(h^\beta - \hat{h}^\beta) + c_\beta h^\beta\right) \\
&= \exp\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha)(1 - \hat{h}^\beta) + c_\beta\right) + \exp\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha)(-\hat{h}^\beta)\right) \\
\{\text{Extract}\} &= \exp\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha)(-\hat{h}^\beta)\right) \left[ \exp\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha) + c_\beta\right) + 1 \right],
\end{aligned}$$

we have

$$\begin{aligned}
E_{\text{eff}}(v) \\
\{\text{Previous}\} &= -\sum_\alpha b_\alpha v^\alpha - \sum_\beta \ln \sum_{h^\beta=0,1} \exp\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha)(h^\beta - \hat{h}^\beta) + c_\beta h^\beta\right) \\
\{\text{Plugin}\} &= -\sum_\alpha b_\alpha v^\alpha + \sum_\beta \sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha) \hat{h}^\beta \\
&\quad - \sum_\beta \ln \left[ \exp\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha) + c_\beta\right) + 1 \right] \\
\{s(x) := \dots\} &= -\sum_\alpha b_\alpha v^\alpha + \sum_{\alpha, \beta} U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha) \hat{h}^\beta - s\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha) + c_\beta\right) \\
\{\text{Extract Const}\} &= -\sum_\alpha b_\alpha v^\alpha + \sum_{\alpha, \beta} U_{\alpha\beta} v^\alpha \hat{h}^\beta - s\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha) + c_\beta\right) + \text{Const} \\
\{\text{Combine}\} &= \sum_\alpha \left( \sum_\beta U_{\alpha\beta} v^\alpha \hat{h}^\beta - b_\alpha \right) - s\left(\sum_\alpha U_{\alpha\beta} (v^\alpha - \hat{v}^\alpha) + c_\beta\right) + \text{Const}.
\end{aligned}$$

□

## 4 Perturbation Theory

### 4.1 Perturbation of Boltzmann Machine

Define  $p_i(x)$  by Taylor expansion  $p_E(x) = p_0(x) + p_1(x) + \dots + p_n(x) + \mathcal{O}(W^{n+1})$ . Denote  $\sigma_\alpha := \sigma(b_\alpha)$ .

#### 4.1.1 0th-order

**Lemma 12.** *[0th-order of Boltzmann Machine]*

We have

$$p_0(x) = \prod_{\alpha} p_{\alpha}(x^{\alpha}), \quad (20)$$

where

$$p_{\alpha}(x) := \frac{\exp(b_{\alpha} x)}{1 + \exp(b_{\alpha})}. \quad (21)$$

**Proof.** Since  $E_0(x; W, b) := -\sum_{\alpha} b_{\alpha} x^{\alpha}$ ,

$$\begin{aligned} p_0(x) &= \frac{\exp(\sum_{\alpha} b_{\alpha} x^{\alpha})}{\sum_{x'^1 \in \{0,1\}} \cdots \sum_{x'^n \in \{0,1\}} \exp(\sum_{\alpha} b_{\alpha} x'^{\alpha})} \\ &= \prod_{\alpha} \frac{\exp(b_{\alpha} x^{\alpha})}{\sum_{x'^{\alpha} \in \{0,1\}} \exp(b_{\alpha} x'^{\alpha})} \\ &= \prod_{\alpha} \frac{\exp(b_{\alpha} x^{\alpha})}{1 + \exp(b_{\alpha})} \\ &= \prod_{\alpha} p_{\alpha}(x). \end{aligned}$$

□

**Lemma 13.** We have

$$\frac{\partial p_{\alpha}}{\partial b_{\alpha}}(x) = p_{\alpha}(x)(x - \sigma_{\alpha}). \quad (22)$$

**Proof.** Directly,

$$\begin{aligned} \frac{\partial}{\partial b_{\alpha}} p_{\alpha}(x) &= \frac{\partial}{\partial b_{\alpha}} \frac{\exp(b_{\alpha} x)}{1 + \exp(b_{\alpha})} \\ &= \frac{\exp(b_{\alpha} x) x}{1 + \exp(b_{\alpha})} - \frac{\exp(b_{\alpha} x) [\exp(b_{\alpha})]}{[1 + \exp(b_{\alpha})]^2} \\ &= \frac{\exp(b_{\alpha} x)}{1 + \exp(b_{\alpha})} \left[ x - \frac{\exp(b_{\alpha})}{1 + \exp(b_{\alpha})} \right] \\ &= p_{\alpha}(x)(x - \sigma(b_{\alpha})). \end{aligned}$$

□

**Lemma 14.** For  $\forall \alpha$ , the mean of  $p_{\alpha}$   $V^{\alpha} := \sum_x p_0(x) x^{\alpha}$  is

$$V^{\alpha} = \sigma^{\alpha}. \quad (23)$$

**Proof.** Since  $(\partial p_{\alpha} / \partial b_{\alpha})(x) = p_{\alpha}(x)(x - \sigma(b_{\alpha}))$ ,

$$\begin{aligned} \sum_x \frac{\partial}{\partial b_{\alpha}} p_{\alpha}(x) &= \sum_x p_{\alpha}(x) x - \sum_x p_{\alpha}(x) \sigma(b_{\alpha}) \\ \frac{\partial}{\partial b_{\alpha}} \sum_x p_{\alpha}(x) &= \sum_x p_{\alpha}(x) x - \left( \sum_x p_{\alpha}(x) \right) \sigma(b_{\alpha}) \\ 0 &= \sum_x p_{\alpha}(x) x - \sigma(b_{\alpha}). \end{aligned}$$

□

**Lemma 15.** Variance  $V^{\alpha_1\alpha_2} := \sum_x p_0(x) \prod_{i=1}^2 (x - \sigma^{\alpha_i})$  is

$$V^{\alpha_1\alpha_2} = V_c^{\alpha_1\alpha_2}. \quad (24)$$

where

$$V_c^{\alpha_1\alpha_2} := \delta^{\alpha_1\alpha_2} \sigma^{\alpha_1} (1 - \sigma^{\alpha_1}). \quad (25)$$

**Proof.** Since  $(\partial p_\alpha / \partial b_\alpha)(x) = p_\alpha(x)(x - \sigma(b_\alpha))$ ,

$$\begin{aligned} \frac{\partial^2 p_0}{\partial b_\beta \partial b_\alpha}(x) &= \frac{\partial}{\partial b_\beta} [p_0(x)(x - \sigma^\alpha)] \\ &= p_0(x)(x - \sigma^\alpha)(x - \sigma^\beta) - \delta^{\alpha\beta} p_0(x) \sigma^\alpha (1 - \sigma^\alpha). \end{aligned}$$

Thus,

$$\begin{aligned} \sum_x \frac{\partial^2 p_0}{\partial b_\beta \partial b_\alpha}(x) &= \sum_x p_0(x)(x - \sigma^\alpha)(x - \sigma^\beta) - \sum_x \delta^{\alpha\beta} p_0(x) \sigma^\alpha (1 - \sigma^\alpha). \\ 0 &= \sum_x p_0(x)(x - \sigma^\alpha)(x - \sigma^\beta) - \delta^{\alpha\beta} \sigma^\alpha (1 - \sigma^\alpha). \\ \sum_x p_0(x)(x - \sigma^\alpha)(x - \sigma^\beta) &= \delta^{\alpha\beta} \sigma^\alpha (1 - \sigma^\alpha). \end{aligned} \quad \square$$

**Lemma 16.** 3-momentum  $V^{\alpha_1\alpha_2\alpha_3} := \sum_x p_0(x) \prod_{i=1}^3 (x - \sigma^{\alpha_i})$  is

$$V^{\alpha_1\alpha_2\alpha_3} = V_c^{\alpha_1\alpha_2\alpha_3}, \quad (26)$$

where

$$V_c^{\alpha_1\alpha_2\alpha_3} := \delta^{\alpha_1\alpha_2\alpha_3} \sigma^{\alpha_1} (1 - \sigma^{\alpha_1}) (1 - 2\sigma^{\alpha_1}). \quad (27)$$

**Lemma 17.** 4-momentum  $V^{\alpha_1\cdots\alpha_4} := \sum_x p_0(x) \prod_{i=1}^4 (x - \sigma^{\alpha_i})$  is

$$V^{\alpha_1\cdots\alpha_4} = V_c^{\alpha_1\cdots\alpha_4} + \sum_{\text{all pairs}} V^{\alpha_{m_1}\alpha_{m_2}} V^{\alpha_{n_1}\alpha_{n_2}}, \quad (28)$$

where “connected” part

$$V_c^{\alpha_1\cdots\alpha_4} := \delta^{\alpha_1\cdots\alpha_4} \sigma^{\alpha_1} (1 - \sigma^{\alpha_1}) [1 - 6\sigma^{\alpha_1} + 6(\sigma^{\alpha_1})^2], \quad (29)$$

and  $(m_1, m_2), (n_1, n_2)$  runs over all (three) pairs.

#### 4.1.2 1st-order

**Lemma 18.** For  $\forall \alpha$ ,

$$\hat{x}^\alpha = \sigma^\alpha + \mathcal{O}(W). \quad (30)$$

**Proof.** The gradient of loss gives

$$\begin{aligned} \hat{x}^\alpha &= \sum_x p_E(x) x^\alpha \\ \{\text{Taylor expand}\} &= \sum_x p_0(x) x^\alpha + \mathcal{O}(W) \\ \left\{ \sum_x p_0(x) x^\alpha = \sigma^\alpha \right\} &= \sigma^\alpha + \mathcal{O}(W). \end{aligned}$$

□

**Theorem 19.**

$$\frac{p_1(x)}{p_0(x)} = \frac{1}{2}W_{\alpha\beta}(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta) - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta}. \quad (31)$$

**Proof.** Directly,

$$\begin{aligned} p_E(x) &= \frac{\exp(b_\alpha x^\alpha + \frac{1}{2}W_{\alpha\beta}(x^\alpha - \hat{x}^\alpha)(x^\beta - \hat{x}^\beta))}{Z} \\ \{\text{Extract } b_\alpha x^\alpha\} &= \frac{\exp(b_\alpha x^\alpha) \exp(\frac{1}{2}W_{\alpha\beta}(x^\alpha - \hat{x}^\alpha)(x^\beta - \hat{x}^\beta))}{Z} \\ \{\text{Expand to } \mathcal{O}(W)\} &= \frac{\exp(b_\alpha x^\alpha) \{1 + \frac{1}{2}W_{\alpha\beta}(x^\alpha - \hat{x}^\alpha)(x^\beta - \hat{x}^\beta) + \dots\}}{Z_0(1 + Z_1 + \dots)} \\ \{p_0(x) = \dots\} &= p_0(x) \frac{\{1 + \frac{1}{2}W_{\alpha\beta}(x^\alpha - \hat{x}^\alpha)(x^\beta - \hat{x}^\beta) + \dots\}}{1 + Z_1 + \dots} \\ \left\{ \frac{1}{1+\epsilon} \sim 1 - \epsilon \right\} &= p_0(x) \left\{ 1 + \frac{1}{2}W_{\alpha\beta}(x^\alpha - \hat{x}^\alpha)(x^\beta - \hat{x}^\beta) + \dots \right\} \{1 - Z_1 + \dots\} \\ \{\text{Expand}\} &= p_0(x) \left\{ 1 + \frac{1}{2}W_{\alpha\beta}(x^\alpha - \hat{x}^\alpha)(x^\beta - \hat{x}^\beta) - Z_1 + \dots \right\} \\ &=: p_0(x) + p_1(x) + \dots \end{aligned}$$

Thus

$$\begin{aligned} \frac{p_1(x)}{p_0(x)} &= \frac{1}{2}W_{\alpha\beta}(x^\alpha - \hat{x}^\alpha)(x^\beta - \hat{x}^\beta) - Z_1 \\ \{\hat{x}^\alpha = \sigma^\alpha + \mathcal{O}(W)\} &= \frac{1}{2}W_{\alpha\beta}(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta) - Z_1. \end{aligned}$$

Now we compute  $Z_1$ . Since

$$\begin{aligned} 1 &= \sum_x p_E(x) = \sum_x p_0(x) \left\{ 1 + \frac{1}{2}W_{\alpha\beta}(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta) - Z_1 \right\} \\ \left\{ \sum_x p_0(x) = 1 \right\} &= 1 + \frac{1}{2}W_{\alpha\beta} \left[ \sum_x p_0(x)(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta) \right] - Z_1 \\ \{V^{\alpha\beta} := \dots\} &= 1 + \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta} - Z_1 \end{aligned}$$

we have

$$Z_1 = \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta}.$$

Then,

$$\begin{aligned} \frac{p_1(x)}{p_0(x)} &= \frac{1}{2}W_{\alpha\beta}(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta) - Z_1 \\ \{Z_1 = \dots\} &= \frac{1}{2}W_{\alpha\beta}(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta) - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta}. \end{aligned}$$

□

**Theorem 20.**  $Up$  to  $\mathcal{O}(W)$ , for  $\forall \gamma$ ,

$$\sum_x p_E(x)x^\gamma = V^\gamma + \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta\gamma}.. \quad (32)$$

**Proof.** Directly,

$$\begin{aligned}
\sum_x p_E(x)x^\gamma &= \sum_x p_0(x)x^\gamma + \sum_x p_1(x)x^\gamma \\
\{p_1(x) = \dots\} &= \sum_x p_0(x)x^\gamma + \sum_x p_0(x) \left[ \frac{1}{2}W_{\alpha\beta}(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta) - \frac{1}{2}W_{\alpha\alpha}\sigma^\alpha(1 - \sigma^\alpha) \right] x^\gamma \\
\{\text{Expand}\} &= \sum_x p_0(x)x^\gamma \\
&\quad + \frac{1}{2}W_{\alpha\beta} \sum_x p_0(x)(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta)x^\gamma \\
&\quad - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta} \sum_x p_0(x)x^\gamma \\
&= \sum_x p_0(x)x^\gamma \\
\{\text{Combine}\} &+ \frac{1}{2}W_{\alpha\beta} \sum_x p_0(x)(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta)(x^\gamma - \sigma^\gamma) + \frac{1}{2}W_{\alpha\beta} \sum_x p_0(x)(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta)\sigma^\gamma \\
&\quad - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta} \sum_x p_0(x)x^\gamma \\
&= V^\gamma \\
\{V^{\alpha\beta} = \dots\} &+ \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta\gamma} + \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta}\sigma^\gamma \\
&\quad - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta}\sigma^\gamma \\
&= V^\gamma + \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta\gamma}.
\end{aligned}$$

□

**Theorem 21.** Up to  $\mathcal{O}(W)$ , for  $\forall(\mu, \nu)$ ,

$$\sum_x p_E(x)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu) = V^{\mu\nu} + W_{(\alpha\beta)}V^{\alpha\mu}V^{\beta\nu} + \frac{1}{2}W_{\alpha\beta}V_c^{\alpha\beta\mu\nu}. \quad (33)$$

**Proof.** Directly,

$$\begin{aligned}
&\sum_x p_E(x)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu) \\
\{p_E = p_0 + p_1\} &= \sum_x p_0(x)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu) + \sum_x p_1(x)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu) \\
&= \sum_x p_0(x)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu) \\
\{p_1(x) = \dots\} &+ \sum_x p_0(x) \left[ \frac{1}{2}W_{\alpha\beta}(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta) - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta} \right] (x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu) \\
&= \sum_x p_0(x)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu) \\
\{\text{Expand}\} &+ \frac{1}{2}W_{\alpha\beta} \sum_x p_0(x)(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu) \\
&\quad - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta} \sum_x p_0(x)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu) \\
\{\hat{x} = \dots\} &= \sum_x p_0(x) \left( x^\mu - \sigma^\mu - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta\mu} \right) \left( x^\nu - \sigma^\nu - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta\nu} \right)
\end{aligned}$$



$$\begin{aligned}
& \{\hat{x}^\alpha = \sigma^\alpha + \mathcal{O}(W)\} + \frac{1}{2}W_{\alpha\beta} \sum_x p_0(x)(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta)(x^\mu - \sigma^\mu)(x^\nu - \sigma^\nu) \\
& \{\hat{x}^\alpha = \sigma^\alpha + \mathcal{O}(W)\} - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta} \sum_x p_0(x)(x^\mu - \sigma^\mu)(x^\nu - \sigma^\nu) \\
& \{\text{Expand}\} = \sum_x p_0(x)(x^\mu - \sigma^\mu)(x^\nu - \sigma^\nu) \\
& \quad - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta\mu} \sum_x p_0(x)(x^\mu - \sigma^\mu) \\
& \quad - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta\nu} \sum_x p_0(x)(x^\nu - \sigma^\nu) \\
& \quad + \frac{1}{2}W_{\alpha\beta} \sum_x p_0(x)(x^\alpha - \sigma^\alpha)(x^\beta - \sigma^\beta)(x^\mu - \sigma^\mu)(x^\nu - \sigma^\nu) \\
& \quad - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta} \sum_x p_0(x)(x^\mu - \sigma^\mu)(x^\nu - \sigma^\nu) \\
& \{V^{\mu\nu} = \dots\} = V^{\mu\nu} \\
& \{\sigma^\mu = V^\mu = \dots\} - 0 \\
& \{\sigma^\nu = V^\nu = \dots\} - 0 \\
& \{V^{\alpha\beta\mu\nu} = \dots\} + \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta\mu\nu} \\
& \{V^{\mu\nu} = \dots\} - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta}V^{\mu\nu} \\
& \quad = V^{\mu\nu} \\
& \{V^{\alpha\beta\mu\nu} = V_c^{\alpha\beta\mu\nu} + \dots\} + \frac{1}{2}W_{\alpha\beta}(V_c^{\alpha\beta\mu\nu} + V^{\alpha\beta}V^{\mu\nu} + V^{\alpha\mu}V^{\beta\nu} + V^{\alpha\nu}V^{\beta\mu}) \\
& \quad - \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta}V^{\mu\nu} \\
& \quad = V^{\mu\nu} + \frac{1}{2}W_{\alpha\beta}(V_c^{\alpha\beta\mu\nu} + V^{\alpha\mu}V^{\beta\nu} + V^{\alpha\nu}V^{\beta\mu}) \\
& \{\text{Combine}\} = V^{\mu\nu} + W_{(\alpha\beta)}V^{\alpha\mu}V^{\beta\nu} + \frac{1}{2}W_{\alpha\beta}V_c^{\alpha\beta\mu\nu}.
\end{aligned}$$

□

**Corollary 22.** Define  $\hat{C}^{\mu\nu} := \sum_x p_D(x)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu)$ . Let  $W$  symmetric. By loss gradient, we have

$$\hat{x}^\alpha = \sum_x p_E(x)x^\alpha; \quad (34)$$

$$\hat{C}^{\mu\nu} = \sum_x p_E(x)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu). \quad (35)$$

From these, we get, up to  $\mathcal{O}(W)$ , for  $\forall \mu$ ,

$$\hat{C}^{\mu\mu} = \hat{x}^\mu(1 - \hat{x}^\mu) + \mathcal{O}(W^2), \quad (36)$$

$$\sigma^\mu = \hat{x}^\mu - W_{\mu\mu}\hat{x}^\mu(1 - \hat{x}^\mu)\left(\frac{1}{2} - \hat{x}^\mu\right); \quad (37)$$

and for  $\forall \mu, \nu$  with  $\mu \neq \nu$ ,

$$W_{\mu\nu} = \frac{\hat{C}^{\mu\nu}}{\hat{x}^\mu(1 - \hat{x}^\mu)\hat{x}^\nu(1 - \hat{x}^\nu)}. \quad (38)$$

**Proof.** When  $\mu \neq \nu$ , we have

$$\begin{aligned}\hat{C}^{\mu\nu} &= \sum_x p_E(x)(x^\mu - \hat{x}^\mu)(x^\nu - \hat{x}^\nu) \\ \{V^{\mu\nu} \propto \delta^{\mu\nu}\} &= W_{(\alpha\beta)} V^{\alpha\mu} V^{\beta\nu} \\ \{W \text{ symmetric}\} &= W_{\alpha\beta} V^{\alpha\mu} V^{\beta\nu} \\ \{V^{\alpha_1\alpha_2} = \delta^{\alpha_1\alpha_2} \sigma^{\alpha_1}(1 - \sigma^{\alpha_1})\} &= W_{\mu\nu} \sigma^\mu(1 - \sigma^\mu) \sigma^\nu(1 - \sigma^\nu) \\ \{\hat{x}^\alpha = \sigma^\alpha + \mathcal{O}(W)\} &= W_{\mu\nu} \hat{x}^\mu(1 - \hat{\sigma}^\mu) \hat{x}^\nu(1 - \hat{x}^\nu)\end{aligned}$$

thus, for  $\forall \mu \neq \nu$ ,

$$W_{\mu\nu} = \frac{\hat{C}^{\mu\nu}}{\hat{x}^\mu(1 - \hat{x}^\mu) \hat{x}^\nu(1 - \hat{x}^\nu)}.$$

And for  $\mu = \nu$ ,

$$\begin{aligned}\hat{C}^{\mu\mu} &= \sum_x p_E(x)(x^\mu - \hat{x}^\mu)(x^\mu - \hat{x}^\mu) \\ \{W_{\mu\nu} \text{ symmetric}\} &= V^{\mu\mu} + W_{\alpha\beta} V^{\alpha\mu} V^{\beta\mu} + \frac{1}{2} W_{\alpha\beta} V^{\alpha\beta\mu\mu} \\ &= \sigma^\mu(1 - \sigma^\mu) \\ &\quad + W_{\alpha\beta} \delta^{\alpha\mu} \delta^{\beta\mu} [\sigma^\mu(1 - \sigma^\mu)]^2 \\ &\quad + \frac{1}{2} W_{\alpha\beta} \delta^{\alpha\beta\mu\mu} \sigma^\mu(1 - \sigma^\mu) [1 - 6\sigma^\mu + 6(\sigma^\mu)^2] \\ &= \sigma^\mu(1 - \sigma^\mu) \\ &\quad + W_{\mu\mu} [\sigma^\mu(1 - \sigma^\mu)]^2 \\ &\quad + \frac{1}{2} W_{\mu\mu} \sigma^\mu(1 - \sigma^\mu) [1 - 6\sigma^\mu + 6(\sigma^\mu)^2] \\ \{\hat{x} = \sigma + \dots\} &= \left( \hat{x}^\mu - \frac{1}{2} W_{\alpha\beta} V^{\alpha\beta\mu} \right) \left( 1 - \hat{x}^\mu + \frac{1}{2} W_{\alpha\beta} V^{\alpha\beta\mu} \right) \\ &\quad + W_{\mu\mu} [\sigma^\mu(1 - \sigma^\mu)]^2 \\ &\quad + \frac{1}{2} W_{\mu\mu} \sigma^\mu(1 - \sigma^\mu) [1 - 6\sigma^\mu + 6(\sigma^\mu)^2] \\ \{\text{Expand}\} &= \hat{x}^\mu(1 - \hat{x}^\mu) + W_{\alpha\beta} V^{\alpha\beta\mu} \left( \hat{x}^\mu - \frac{1}{2} \right) \\ &\quad + W_{\mu\mu} [\sigma^\mu(1 - \sigma^\mu)]^2 \\ &\quad + \frac{1}{2} W_{\mu\mu} \sigma^\mu(1 - \sigma^\mu) [1 - 6\sigma^\mu + 6(\sigma^\mu)^2] \\ \{V^{\alpha\beta\mu} = \dots\} &= \hat{x}^\mu(1 - \hat{x}^\mu) + W_{\mu\mu} \sigma^\mu(1 - \sigma^\mu) (1 - 2\sigma^\mu) \left( \hat{x}^\mu - \frac{1}{2} \right) \\ &\quad + W_{\mu\mu} [\sigma^\mu(1 - \sigma^\mu)]^2 \\ &\quad + \frac{1}{2} W_{\mu\mu} \sigma^\mu(1 - \sigma^\mu) [1 - 6\sigma^\mu + 6(\sigma^\mu)^2] \\ \{\hat{x}^\alpha = \sigma^\alpha + \mathcal{O}(W)\} &= \hat{x}^\mu(1 - \hat{x}^\mu) + W_{\mu\mu} \hat{x}^\mu(1 - \hat{x}^\mu) (1 - 2\hat{x}^\mu) \left( \hat{x}^\mu - \frac{1}{2} \right) \\ &\quad + W_{\mu\mu} [\hat{x}^\mu(1 - \hat{x}^\mu)]^2 \\ &\quad + \frac{1}{2} W_{\mu\mu} \hat{x}^\mu(1 - \hat{x}^\mu) [1 - 6\hat{x}^\mu + 6(\hat{x}^\mu)^2] \\ \{\text{Combine}\} &= \hat{x}^\mu(1 - \hat{x}^\mu) \\ &\quad + W_{\mu\mu} \hat{x}^\mu(1 - \hat{x}^\mu) \times \\ &\quad \times \left\{ (1 - 2\hat{x}^\mu) \left( \hat{x}^\mu - \frac{1}{2} \right) + \hat{x}^\mu(1 - \hat{x}^\mu) + \frac{1}{2} [1 - 6\hat{x}^\mu + 6(\hat{x}^\mu)^2] \right\} \\ \{\text{Simplify}\} &= \hat{x}^\mu(1 - \hat{x}^\mu),\end{aligned}$$

Thus,

$$\hat{C}^{\mu\nu} = \hat{x}^\mu(1 - \hat{x}^\mu) + \mathcal{O}(W^2).$$

Finally, we have

$$\begin{aligned}
 \hat{x}^\mu &= V^\mu + \frac{1}{2}W_{\alpha\beta}V^{\alpha\beta\mu} \\
 &= \sigma^\mu + \frac{1}{2}W_{\alpha\beta}\delta^{\alpha\beta\mu}\sigma^\alpha(1-\sigma^\alpha)(1-2\sigma^\alpha) \\
 &= \sigma^\mu + W_{\mu\mu}\sigma^\mu(1-\sigma^\mu)\left(\frac{1}{2}-\sigma^\mu\right). \\
 \{\hat{x}^\alpha = \sigma^\alpha + \mathcal{O}(W)\} &= \sigma^\mu + W_{\mu\mu}\hat{x}^\mu(1-\hat{x}^\mu)\left(\frac{1}{2}-\hat{x}^\mu\right)
 \end{aligned}$$

Thus

$$\sigma^\mu = \hat{x}^\mu - W_{\mu\mu}\hat{x}^\mu(1-\hat{x}^\mu)\left(\frac{1}{2}-\hat{x}^\mu\right).$$

□