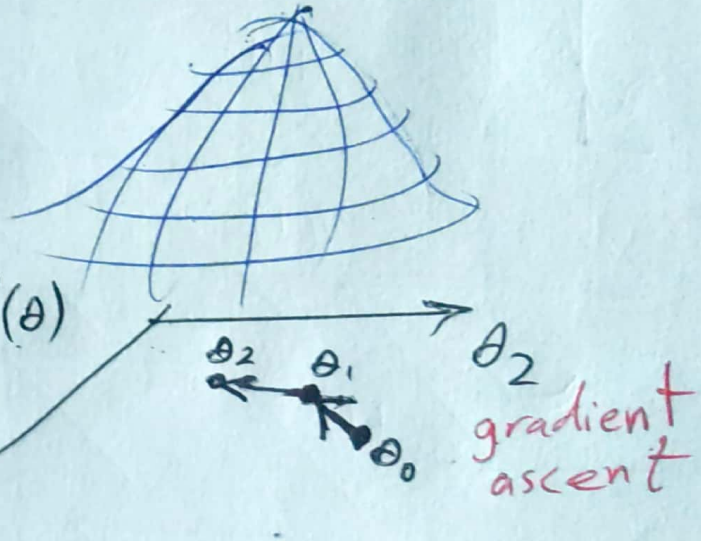
$$\frac{\partial}{\partial \theta_j} ll(\theta)$$

$$\begin{bmatrix} \frac{\partial}{\partial \theta_1} ll(\theta) \\ \frac{\partial}{\partial \theta_2} ll(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_p} ll(\theta) \end{bmatrix} \Rightarrow \nabla ll(\theta) = \frac{\partial}{\partial \theta} ll(\theta)$$

$$P(X) = \frac{1}{Z(\theta)} e^{F(X,\theta)}$$



گرادیان

$$\frac{\partial}{\partial \theta_j} ll(\theta) = m E_D\left\{\frac{\partial}{\partial \theta_j} F(X,\theta)\right\} - m E_{P_\theta(X)}\left\{\frac{\partial}{\partial \theta_j} F(X,\theta)\right\}$$

$t \leftarrow 0$

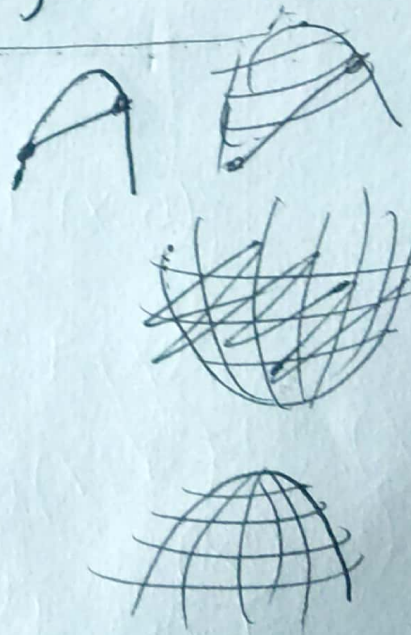
θ_0

while not converged

$$\frac{\partial}{\partial \theta} ll(\theta)\Big|_{\theta=\theta_t}$$

$$\theta_{t+1} = \theta_t + \lambda \frac{\partial}{\partial \theta} ll(\theta)\Big|_{\theta=\theta_t}$$

$t \leftarrow t+1$



problem: How to compute $E_{P_\theta(X)}\left\{\frac{\partial}{\partial\theta_j} F(X,\theta)\right\}$

take samples $X'^1, X'^2, \dots, X'^M$ from $P_\theta(X) = \frac{1}{Z(\theta)} e^{F(X,\theta)}$

$$E_{P_\theta(X)}\left\{\frac{\partial}{\partial\theta_j} F(X,\theta)\right\} \simeq \frac{1}{M} \sum_{i=1}^{m} \frac{\partial}{\partial\theta_j} F(X'^i,\theta)$$

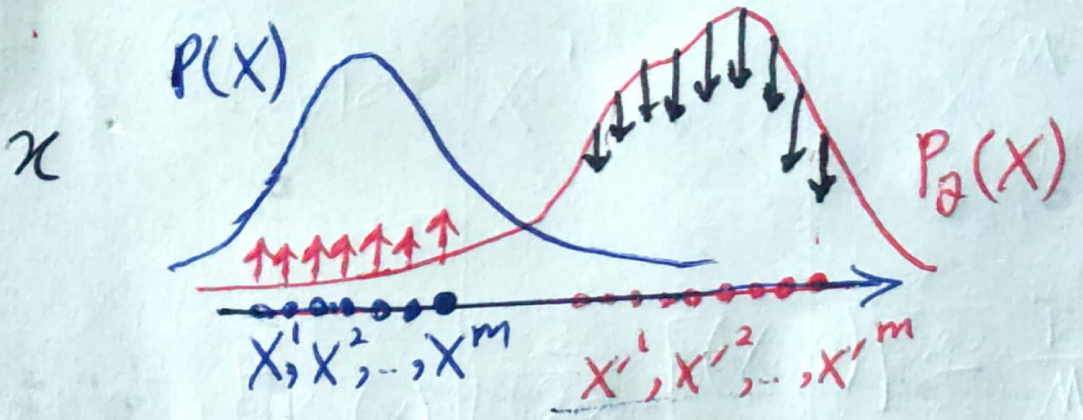
$$\frac{\partial}{\partial\theta_j} ll(\theta) = \sum_{i=1}^{m} \frac{\partial}{\partial\theta_j} F(X^i,\theta) - \frac{m}{M} \sum_{i=1}^{m} \frac{\partial}{\partial\theta_j} F(X'^i,\theta)$$

training data $X^1, X^2, \dots, X^m$

samples from $P_\theta(X)$ $X'^1, X'^2, \dots, X'^m$

usally $M = m$

$$\frac{\partial}{\partial\theta_j} ll(\theta) = \underbrace{\sum_{i=1}^{m} \frac{\partial F}{\partial\theta_j} F(X^i,\theta)}_{\text{positive force}} - \underbrace{\sum_{i=1}^{m} \frac{\partial F}{\partial\theta_j} F(X'^i,\theta)}_{\text{negative force}}$$



Contrastive divergence (CD)

CRF - $P_\theta(Y|X) = \frac{1}{Z(\theta, X)} e^{F_\theta(\theta X, Y)}$

Conditional likelihood

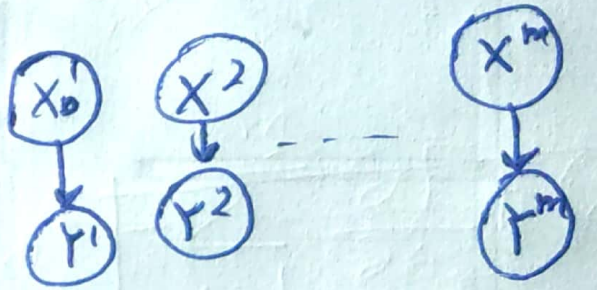
Data

$(X^1, Y^1), (X^2, Y^2), \ldots, (X^m, Y^m)$

$cll(\theta) = Pr(Y^1, Y^2, \ldots, Y^m | X^1, X^2, \ldots, X^m)$

$\prod_{i=1}^{m} Pr(Y^i | X^1, \ldots, X^m)$

$\prod_{i=1}^{m} Pr(Y^i | X^i) \Longrightarrow cll(\theta) = \prod_{i=1}^{m} P_\theta(Y^i | X^i)$



(IV)

$$cll(\theta) = \log cl(\theta) = \sum_{i=1}^{m} \log P_\theta(Y^i \mid X^i)$$

$$= \sum_{i=1}^{m} \log \frac{1}{Z_\theta(X^i)} e^{F_\theta(X^i, Y^i)}$$

$$= -\sum_{i=1}^{m} \log Z_\theta(X^i) + \sum_{i=1}^{m} F_\theta(X^i, Y^i)$$

$$\frac{\partial}{\partial \theta_j} cll(\theta) = -\sum_{i=1}^{m} \sum_{Y} P_\theta(Y \mid X^i) F_\theta(X^i, Y) + \sum_{i=1}^{m} F_\theta(X^i, Y^i)$$

$$-\sum_{i=1}^{m} \underbrace{E_{P_\theta(Y \mid X^i)} \{ F_\theta(X^i, Y) \}}_{\text{Needs inference per } X^i} + {}^{m}E_{D} \{ F_\theta(X, Y) \}$$

(IV)

$$cll(\theta) = \log cl(\theta) = \sum_{i=1}^{m} \log P_\theta(Y^i \mid X^i)$$

$$= \sum_{i=1}^{m} \log \frac{1}{Z_\theta(X^i)} e^{F_\theta(X^i, Y^i)}$$

$$= -\sum_{i=1}^{m} \log Z_\theta(X^i) + \sum_{i=1}^{m} F_\theta(X^i, Y^i)$$

$$\frac{\partial}{\partial \theta_j} cll(\theta) = -\sum_{i=1}^{m} \sum_{Y} P_\theta(Y \mid X^i) F_\theta(X^i, Y) + \sum_{i=1}^{m} F_\theta(X^i, Y^i)$$

$$-\sum_{i=1}^{m} \underbrace{E_{P_\theta(Y \mid X^i)} \{ F_\theta(X^i, Y) \}}_{\text{Needs inference per } X^i} + {}^{m}E_D \{ F_\theta(X, Y) \}$$

$$F_{\theta}(X) = \theta\left\{f_1(X_1, X_2) + f_2(X_2, X_3) + f_3(X_3, X_4)\right\}$$

$$E\left\{\frac{\partial F}{\partial \theta}\right\} = E\left\{f_1(X_1, X_2) + f_2(X_2, X_3) + f_3(X_3, X_4)\right\}$$

$$E\left\{f_1(X_1, X_2)\right\} + E\left\{f_2(X_2, X_3)\right\} + E\left\{f_3(X_3, X_4)\right\}$$