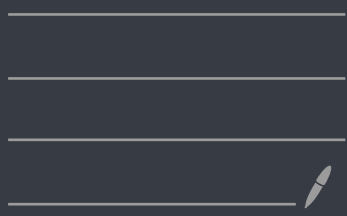


· Likelihood - Based Training

$\begin{cases} p(x) \geq 0 \\ \sum_x p(x) = 1 \end{cases}$ $\left(\int_x p(x)dx = 1\right)$

the "volume" is fixed. when the MLE training increase $P(X_{train})$,
it decrease $P(X_{test})$ normally ∴

coming up with a non-negative function $f_\theta(x)$ is easy (eg. $g_\theta(x) = f_\theta(x)^2$ or $\exp(f_\theta(x))$ where f_θ is a neural net)
but $g_\theta(x)$ may not sum/integrate to 1 $\left(\sum_x g_\theta(x) = Z(\theta) \neq 1\right)$

· Energy - Based Model

$P_\theta(x) = \frac{1}{Z(\theta)} \exp(f_\theta(x))$ $Z(\theta) = \int_\theta \exp(f_\theta(x))dx$

why exponential:

1. want to capture large variations in probability (otherwise we need highly non-smooth f_θ)
2. log-prob is the natural that we want to work with
3. exponential family: many common distributions can be written in this form
4. these distributions arise under fairly general assumptions in statistical physics $-f_\theta(x)$ is called the energy

pros: can use pretty much any $f_\theta(x)$

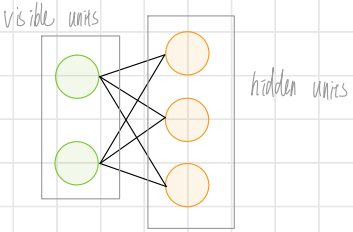
cons: sampling is hard
evaluating and optimizing likelihood $P_\theta(x)$ is hard (need to evaluate $\int_x \exp f_\theta(x)dx$)
no feature learning (but can add latent variables)

Given x and x', evaluating $P_\theta(x)$ and $P_\theta(x')$ is hard.

However, their ratio $P_\theta(x)/P_\theta(x') = \exp(f_\theta(x) - f_\theta(x'))$ does not involve $Z(\theta)$

can do anomaly detection, denoising

· Restricted Boltzmann Machine



RBM: Energy-based model with binary variable

$x \in \{0,1\}^n \quad z \in \{0,1\}^m$

the joint distribution is $P(x,z;w,b,c) = \frac{1}{Z}\exp(x^T W z + b^T x + c^T z)$

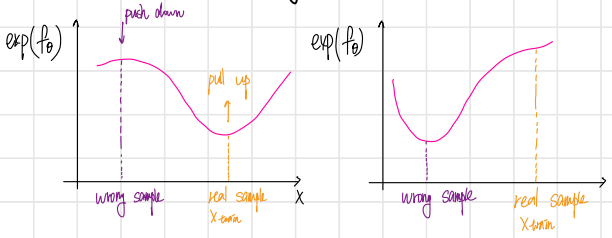
called restrict because theres no X-X connection or Z-Z connection

(can stack multiple RBM together)

∵ the normalizing constant is hard to compute : $Z(w,b,c) = \sum_{x\in\{0,1\}^n} \sum_{z\in\{0,1\}^m} \exp(x^T W z + b^T x + c^T z)$

optimizing the unnormalized prob exp() wrt w b c is easy, but optimizing $P(x,z;w,b,c)$, which depends on z, is hard

· Energy-based Model Training



Goal: maximize $\frac{1}{Z(\theta)}\exp(f_\theta(X_{train}))$ increase the numerator, decrease the denominator

Because the model is unnormalized, increasing $\exp(f_\theta(X_{train}))$ wrt θ doesn't guarantee that X_{train} become more likely compared to other X's

contrastive divergence:

instead of evaluating $Z(\theta)$ exactly, use a monte-carlo estimation

contrastive divergence: Sample $X_{sample} \sim P_\theta$

take step on $\nabla_\theta (f_\theta(X_{train}) - f_\theta(X_{sample}))$ ∵ sampling is hard

Sampling from energy-based models

no direct way to sample from P_θ since we cannot evaluate how likely each possible sample is

However, we can easily compare two samples

Markov chain Monte Carlo:

Initialize X^0 randomly, $t=0$

repeat {

$X' = X^t + noise$ (eg. flip some pixels, gaussian noise)

if $f_\theta(X') > f_\theta(X^t)$ let $X^{t+1} = X'$

else let $X^{t+1} = X'$ with prob $\exp(f_\theta(X') - f_\theta(X^t))$

}

run this procedure for sufficiently large number of iterations

this will give a sample from the distribution

Stochastic local search, may be stuck at local maximum and only goes one with very low prob