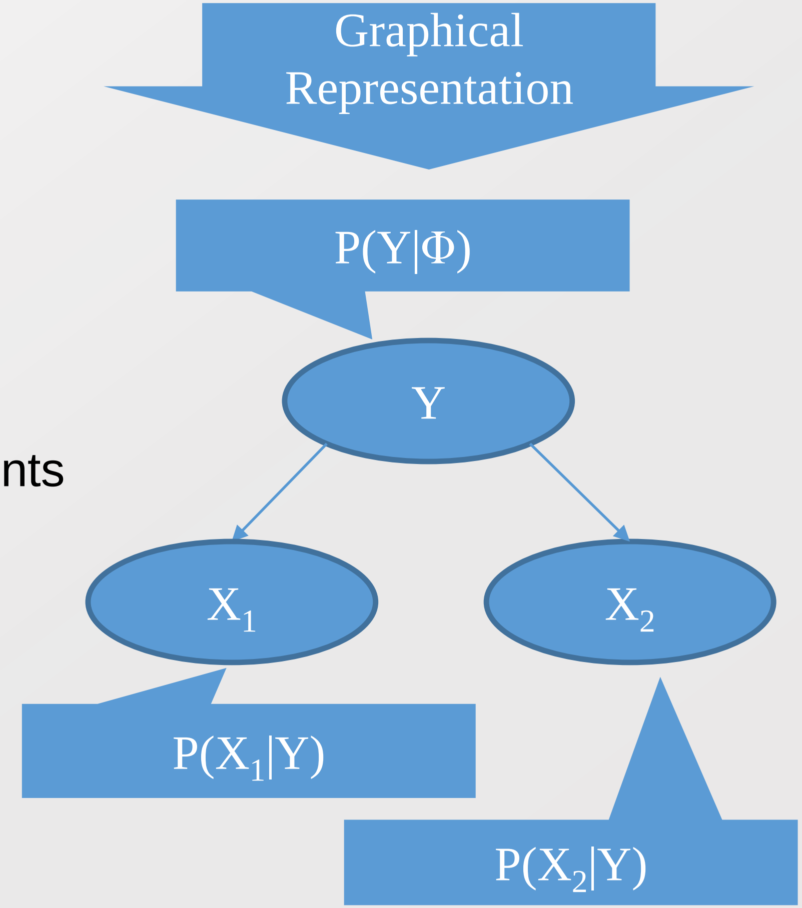
Conditional Random Field

Il-Chul Moon
Department of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

Detour: Bayesian Network

- A graphical notation of
  - Random variables
  - Conditional independence
  - To obtain a compact representation of the full joint distributions
- Syntax
  - A acyclic and directed graph
  - A set of nodes
    - A random variable
    - A conditional distribution given its parents
    - P (X_i|Parents(X_i))
  - A set of links
    - Direct influence from the parent to the child

$$P(Y=y)\prod_{1\le i\le d}P(X_i=x_i|Y=y)$$

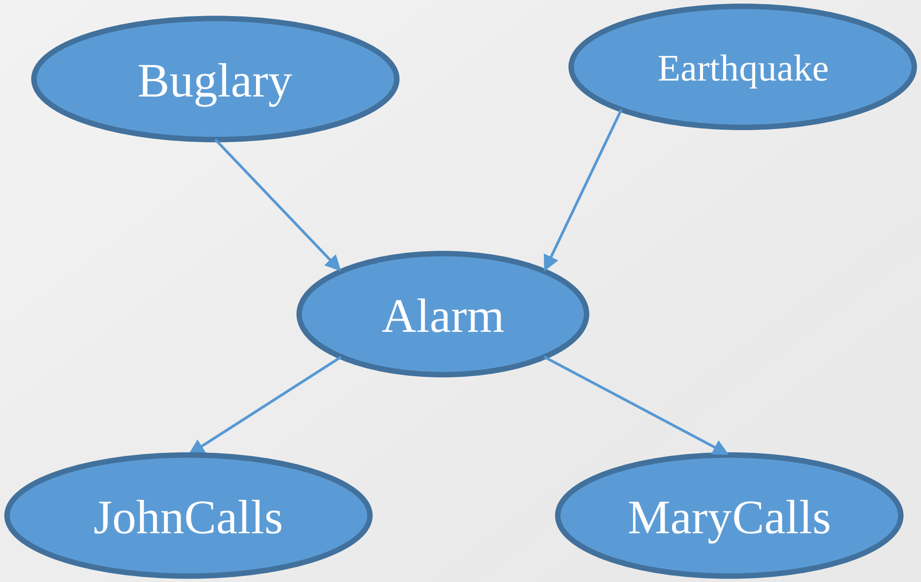


Detour: Components of Bayesian Network

- Qualitative components
  - Prior knowledge of causal relations
  - Learning from data
  - Frequently used structures
  - Structural aspects
- Quantitative components
  - Conditional probability tables
  - Probability distribution assigned to nodes
- Probability computing is related to both
  - Quantitative and Qualitative

P(B)=0.001

P(E)=0.002



B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

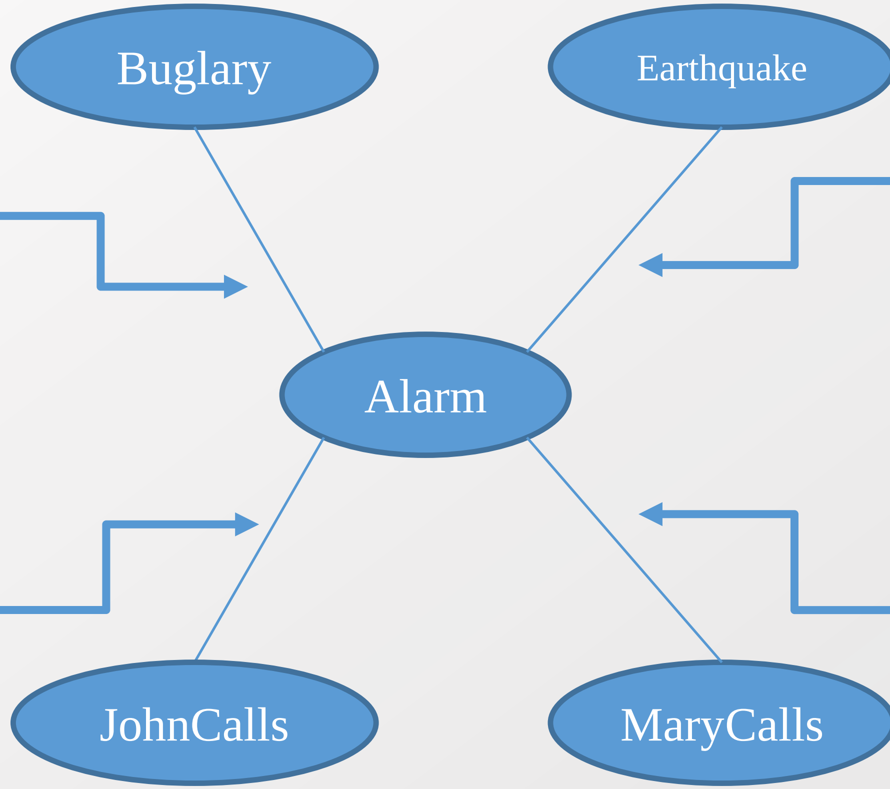
A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

Undirected Graphical Model

B	A	$\phi(B,A)$
1	1	0
1	0	5
0	1	3
0	0	1

E	A	$\phi(A,E)$
1	1	6
1	0	4
0	1	2
0	0	1



J	A	$\phi(A,J)$
1	1	3
1	0	2
0	1	2
0	0	1

M	A	$\phi(A,M)$
1	1	4
1	0	3
0	1	3
0	0	2

- Full joint distribution
  - $P(B,E,A,J,M) = \frac{\phi(A,B)\phi(A,E)\phi(A,J)\phi(A,M)}{\sum_{A,B,E,J,M}\phi(A,B)\phi(A,E)\phi(A,J)\phi(A,M)} = \frac{\phi(A,B)\phi(A,E)\phi(A,J)\phi(A,M)}{Z}$

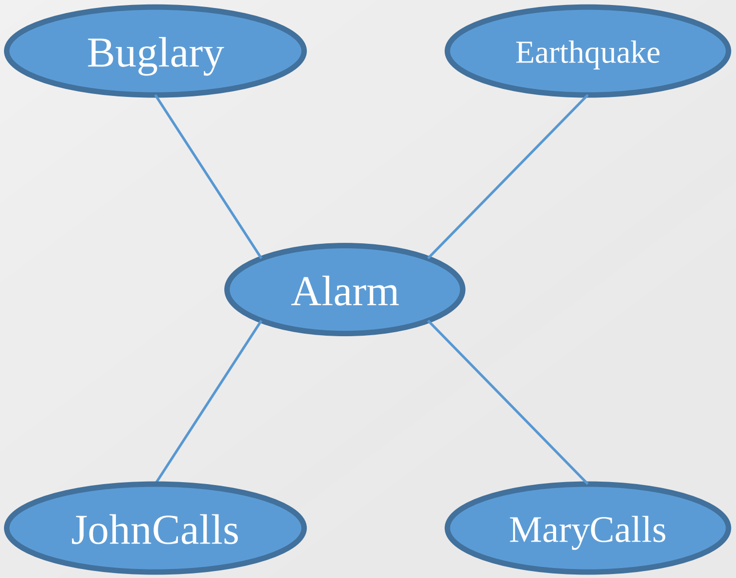
Markov Random Field

- Markov random field
  - A probability distribution p over variables $x_1 \dots x_n$ defined by an undirected graph G in which nodes correspond to variable x_i.

$$p(x_1 \dots x_n) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c)$$

  - Here, C is the set of cliques of G.
    - Edge is a minimum form of clique with two nodes
  - ϕ_c is a nonnegative function over the variables in a clique.
  - Z is the partition function to normalize.

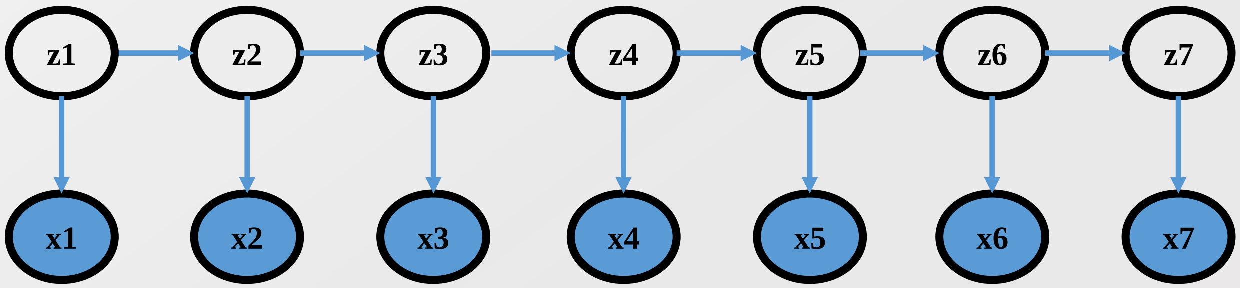
$$Z = \sum_{x_1 \dots x_n} \prod_{c \in C} \phi_c(x_c)$$



$$P(B,E,A,J,M) = \frac{\phi(A,B)\phi(A,E)\phi(A,J)\phi(A,M)}{\sum_{A,B,E,J,M} \phi(A,B)\phi(A,E)\phi(A,J)\phi(A,M)} = \frac{\phi(A,B)\phi(A,E)\phi(A,J)\phi(A,M)}{Z}$$

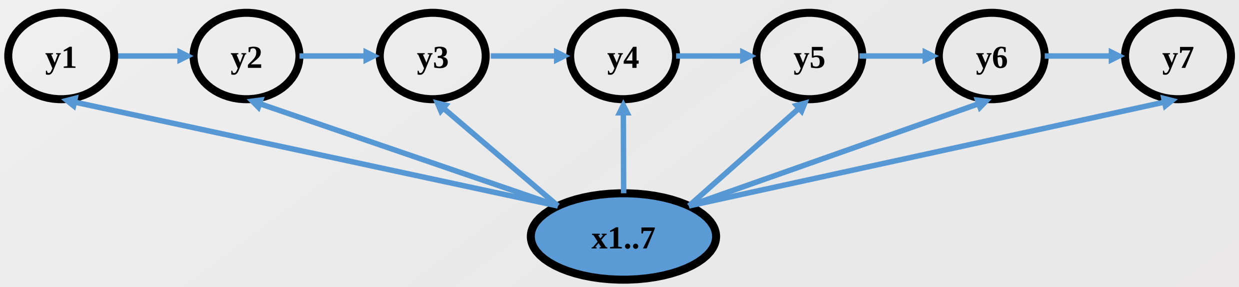
Limitations of Hidden Markov Model

- Hidden Markov model captures limited dependencies
  - State at the current time to Observation at the current time
  - State at the previous time to State at the current time
  - Only forward dependencies in the state
- Hidden Markov model is a generative model that could be used for a classification task
  - HMM maximizes the likelihood of $P(X, Z)$
  - Classification task optimizes $P(Z|X)$ instead of $P(X, Z)$



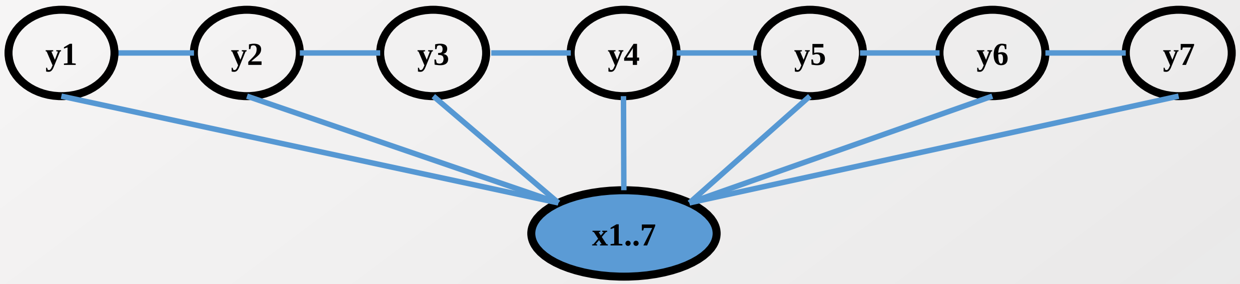
- MEMM captures dependencies
  - Observation from all time to State at the current time
  - Only forward dependencies in the state
- MEMM is a discriminative model for a classification task
  - MEMM optimizes $P(Z|X)$
- However, the formulation includes a new function of f$(y_i, y_{i-1}, X_{1:n})$
  - Potential function!

$$P(Y_{1:n}|X_{1:n}) = \prod_{i=1}^{n} P(y_i|y_{i-1}, X_{1:n}) = \prod_{i=1}^{n} \frac{\exp(w^T f(y_i, y_{i-1}, X_{1:n}))}{\sum_{y_j} \exp(w^T f(y_j, y_{i-1}, X_{1:n}))} = \prod_{i=1}^{n} \frac{\exp(w^T f(y_i, y_{i-1}, X_{1:n}))}{Z(y_{i-1}, X_{1:n})}$$



Conditional Random Field

- Conditional random field defines
  - The potential function between state transitions
  - The potential function between the state and the observations



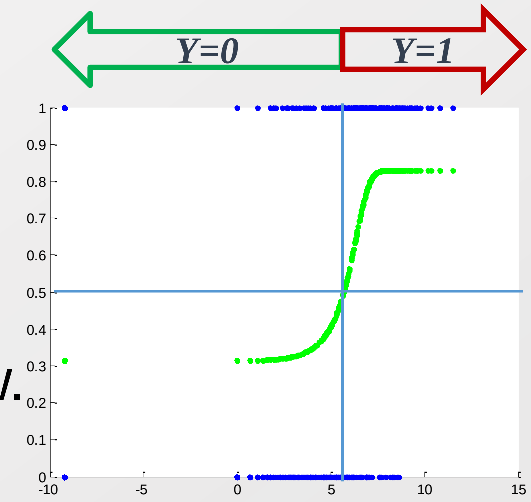
$$P(Y_{1:n}|X_{1:n}) = \frac{1}{Z(\lambda,\mu,X_{1:n})}\prod_{i=1}^{n}\phi(y_i,y_{i-1},X_{1:n})$$

$$= \frac{1}{Z(\lambda,\mu,X_{1:n})}\exp(\sum_{i=1}^{n}(\sum_{k}\lambda_k f_k(y_i,y_{i-1},X_{1:n}) + \sum_{l}\mu_l g_l(y_i,X_{1:n})))$$

$$Z(\lambda,\mu,X_{1:n}) = \sum_{Y_{1:n}}\exp(\sum_{i=1}^{n}(\sum_{k}\lambda_k f_k(y_i,y_{i-1},X_{1:n}) + \sum_{l}\mu_l g_l(y_i,X_{1:n})))$$

Detour: Logistic Regression

- Logistic regression is a probabilistic classifier to predict the binomial or the multinomial outcome
  - by fitting the conditional probability to the logistic function.
- You can see the problem from the different view.
  - This way is actually closer to the formal definition.



- Given the Bernoulli experiment
  - $P(y|x) = \mu(x)^y (1 - \mu(x))^{1-y}$
  - $\mu(x) = \frac{1}{1+e^{-\dot{\theta}^T x}} = P(y = 1|x)$
  - Here, $\mu(x)$ is the logistic function
- From the previous slide,
  - $X\theta = \log\left(\frac{P(Y|X)}{1-P(Y|X)}\right) \rightarrow P(Y|X) = \frac{e^{X\theta}}{1+e^{X\theta}}$

Logistic Function

$$f(x) = \frac{1}{1 + e^{-x}}$$

The goal, finally, becomes finding out $\boldsymbol{\theta}$, again

Comparison to Logistic Regression

- $P(Y_{1:n}|X_{1:n}) = \frac{1}{Z(\lambda,\mu,X_{1:n})}\prod_{i=1}^{n}\phi(y_i, y_{i-1}, X_{1:n})$

$$= \frac{1}{Z(\lambda,\mu,X_{1:n})}\exp(\sum_{i=1}^{n}(\sum_{k}\lambda_k f_k(y_i, y_{i-1}, X_{1:n}) + \sum_{l}\mu_l g_l(y_i, X_{1:n})))$$

- Assume that
  - y is a single dimension
  - $X_{1:n}$ has a binary value for each X_i
  - f_k is an indicator feature function as $f_k = \mathbf{1}_{X_i=1,y=1}$
- Then, the conditional random field becomes
  - $P(Y=1|X_{1:n}) = \frac{1}{Z(\lambda,\mu,X_{1:n})}\exp\left(\sum_{i=1}^{n}\lambda_k \mathbf{1}_{X_i=1,y=1}\right)$

$$= \frac{\exp\left(\sum_{i=1}^{n}\lambda_k \mathbf{1}_{X_i=1,y=1}\right)}{\exp\left(\sum_{i=1}^{n}\lambda_k \mathbf{1}_{X_i=1,y=0}\right) + \exp\left(\sum_{i=1}^{n}\lambda_k \mathbf{1}_{X_i=1,y=1}\right)}$$

$$= \frac{\exp(\sum_{i=1}^{n}\lambda_k X_i)}{\exp(0) + \exp(\sum_{i=1}^{n}\lambda_k X_i)} = \frac{\exp(\sum_{i=1}^{n}\lambda_k X_i)}{1 + \exp(\sum_{i=1}^{n}\lambda_k X_i)} = \frac{1}{1 + \exp(\sum_{i=1}^{n}-\lambda_k X_i)}$$

Detour: Exponential Family

- Exponential Family
  - $P(x|\theta) = h(x)\exp(\eta(\theta) \cdot T(x) - A(\theta))$
    - Sufficient statistics : $T(x)$, Natural parameter : $\eta(\theta)$
    - Underlying measure : $h(x)$, Log normalizer : $A(\theta)$
  - Normal Distribution : $P(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$
    - Sufficient statistics : $(x, x^2)^T$, Natural parameter :$(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})^T$
    - Underlying measure :$\frac{1}{\sqrt{2\pi}}$, Log normalizer :$\frac{\mu^2}{2\sigma^2} + \log|\sigma|$
  - Dirichlet Distribution : $P(x_1, \dots, x_K|\alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} {x_i}^{\alpha_i - 1}$
    - Sufficient statistics : $(\log x_1, \dots, \log x_K)^T$, Natural parameter :$(\alpha_1 - 1, \dots, \alpha_K - 1)^T$
    - Underlying measure :1, Log normalizer : $-\log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) + \log\prod_{i=1}^{K}\Gamma(\alpha_i)$

Conditional Random Field : $\boldsymbol{P(Y_{1:n}|X_{1:n}) = \frac{1}{Z(\lambda,\mu,X_{1:n})} exp(\sum_{i=1}^{n}(\sum_k \lambda_k f_k(y_i, y_{i-1}, X_{1:n}) + \sum_l \mu_l g_l(y_i, X_{1:n})))}$

- Derivative of log normalizer → Moments of sufficient statistics
  - $\frac{d}{d\eta}a(\eta) = \frac{d}{d\eta}\log\int h(x)\exp\{\eta^T T(x)\}dx = \frac{\int T(x)h(x)\exp\{\eta^T T(x)\}dx}{\int h(x)\exp\{\eta^T T(x)\}dx}$

$$= \frac{\int T(x)h(x)\exp\{\eta^T T(x)\}dx}{\exp(a(\eta))} = \int T(x)h(x)\exp\{\eta^T T(x) - a(\eta)\}\,dx$$

KAIST

Learning the CRF Parameter, λ and μ

$$Z(\lambda,\mu,X_{1:n}) = \sum_{Y_{1:n}} \exp(\sum_{i=1}^{n}(\sum_k \lambda_k f_k(y_i,y_{i-1},X_{1:n}) + \sum_l \mu_l g_l(y_i,X_{1:n}))$$

$P(x|\theta)$
$= h(x)\exp(\eta(\theta)\cdot T(x)$
$- A(\theta))$

- In supervised learning, the pair of $X_{1:n}$ and $Y_{1:n}$ are provided
  - Need to maximize $P(Y_{1:n}|X_{1:n})$ by adjusting CRF parameters
- $\lambda^*,\mu^* = argmax_{\lambda,\mu} L(\lambda,\mu) = argmax_{\lambda,\mu} \prod_{d\in D} P\left(Y_{d,1:n}\middle|X_{d,1:n};\lambda,\mu\right)$

$$= argmax_{\lambda,\mu} \prod_{d\in D} \frac{1}{Z(\lambda,\mu,X_{d,1:n})} \exp(\sum_{i=1}^{n}(\sum_k \lambda_k f_k(y_{d,i},y_{d,i-1},X_{d,1:n}) + \sum_l \mu_l g_l(y_{d,i},X_{d,1:n}))$$

$$= argmax_{\lambda,\mu} \sum_{d\in D}\left[\sum_{i=1}^{n}(\sum_k \lambda_k f_k(y_{d,i},y_{d,i-1},X_{d,1:n}) + \sum_l \mu_l g_l(y_{d,i},X_{d,1:n})) - logZ(\lambda,\mu,X_{d,1:n})\right]$$

- Simple gradient method can be applied to the objective function
- $\nabla_{\lambda_k} L(\lambda,\mu) = \sum_{d\in D}\left[\sum_{i=1}^{n} \lambda_k f_k(y_{d,i},y_{d,i-1},X_{d,1:n}) - \frac{d}{d\lambda_k} logZ(\lambda,\mu,X_{d,1:n})\right]$
  - $\frac{d}{d\lambda_k} logZ(\lambda,\mu,X_{d,1:n}) = E_{P\left(Y_{d,1:n}\middle|X_{d,1:n};\lambda,\mu\right)}[\sum_{i=1}^{n}\sum_k f_k(y_i,y_{i-1},X_{1:n})]$
    - $\because \frac{d}{d\eta} a(\eta) = E_P[T(x)]$
- $\nabla_{\lambda_k} L(\lambda,\mu) = \sum_{d\in D}\left[\sum_{i=1}^{n} \lambda_k f_k(y_{d,i},y_{d,i-1},X_{d,1:n}) - \sum_{Y_{d,1:n}} P\left(Y_{d,1:n}\middle|X_{d,1:n};\lambda,\mu\right)\sum_{i=1}^{n}\sum_k f_k(y_i,y_{i-1},X_{1:n})\right]$
- $=\sum_{d\in D}\left[\sum_{i=1}^{n} \lambda_k f_k(y_{d,i},y_{d,i-1},X_{d,1:n}) - \sum_{Y_{d,1:n}} P\left(Y_{d,1:n}\middle|X_{d,1:n};\lambda,\mu\right)\sum_{i=1}^{n}\sum_k f_k(y_i,y_{i-1},X_{1:n})\right]$
- $=\sum_{d\in D}[\sum_{i=1}^{n} \lambda_k f_k(y_{d,i},y_{d,i-1},X_{d,1:n}) - \sum_{i=1}^{n}\sum_{y_{d,i},y_{d,i-1}}\sum_k P\left(Y_{d,1:n}\middle|X_{d,1:n};\lambda,\mu\right) f_k(y_i,y_{i-1},X_{1:n})]$

Neural Networks and CRF

- Similarity on model structure
  - Neuron with logistic activation function == Logistic regression
  - CRF with assumptions == Logistic regression
- Similarity on model inference
  - Neuron with gradient descent
  - CRF with gradient descent
- Two models are easily interoperable and inferenced together