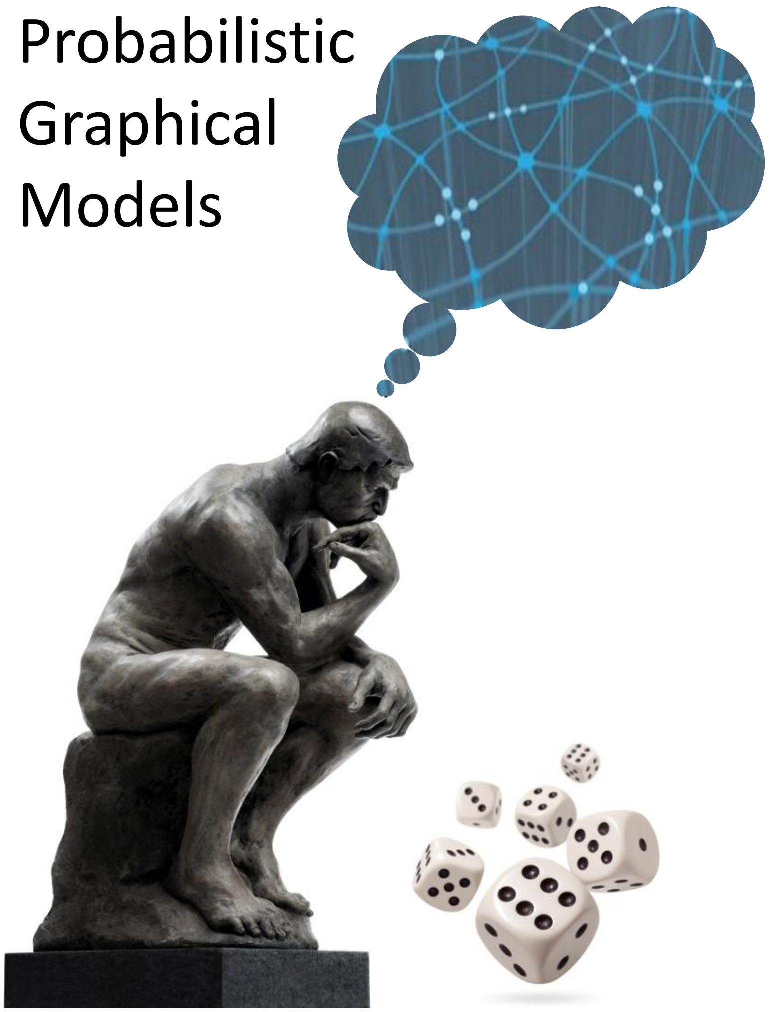
Probabilistic
Graphical
Models



Learning

Parameter Estimation

Max Likelihood for CRFs

Estimation for CRFs

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y} \mid \boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{x}}(\boldsymbol{\theta})} \tilde{P}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y}) \qquad Z_{\boldsymbol{x}}(\boldsymbol{\theta}) = \sum_{\boldsymbol{Y}} \tilde{P}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{Y})$$

log conditional likelihood

$$\mathcal{D} = \{(\boldsymbol{x}[m], \boldsymbol{y}[m])\}_{m=1}^{M} \qquad \ell_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{\theta} : \mathcal{D}) = \sum_{m=1}^{M} \ln P_{\boldsymbol{\theta}}(\boldsymbol{y}[m] \mid \boldsymbol{x}[m], \boldsymbol{\theta})$$

$$\ell_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{\theta} : (\boldsymbol{x}[m], \boldsymbol{y}[m])) = \left(\sum_i \theta_i f_i(\boldsymbol{x}[m], \boldsymbol{y}[m])\right) - \ln Z_{\boldsymbol{x}[m]}(\boldsymbol{\theta})$$

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} \ell_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{\theta} : \mathcal{D}) = \frac{1}{M} \sum_{m=1}^{M} \left(\boxed{f_i(\boldsymbol{x}[m], \boldsymbol{y}[m])} - \boxed{\boldsymbol{E}_{\boldsymbol{\theta}}[f_i(\boldsymbol{x}[m], \boldsymbol{Y})]} \right)$$

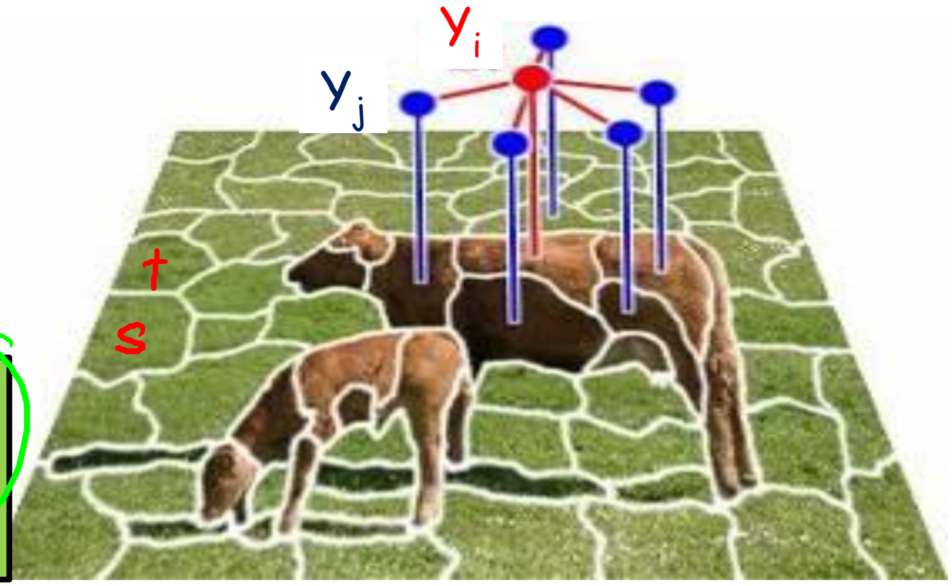
Daphne Koller

Example

$f_1(Y_s, X_s) = \mathbf{1}(Y_s = g) \times G_s$

average intensity of green channel for pixels in superpixel s

$f_2(Y_s, Y_t) = \mathbf{1}(Y_s = Y_t)$



$$\frac{\partial}{\partial \theta_i} \ell_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{\theta} : (\boldsymbol{x}[m], \boldsymbol{y}[m])) = (f_i(\boldsymbol{x}[m], \boldsymbol{y}[m]) - \boldsymbol{E}_{\boldsymbol{\theta}}[f_i(\boldsymbol{x}[m], \boldsymbol{Y})])$$

m

$$\frac{\partial}{\partial \theta_1} = \sum_s \mathbf{1}\{y_s[m] = g\} G_s[m] - \sum_s P_{\boldsymbol{\theta}}(Y_s = g \mid \boldsymbol{x}[m]) G_s[m]$$

$$\frac{\partial}{\partial \theta_2} = \sum_{(s,t) \in \mathcal{N}} \mathbf{1}\{y_s[m] = y_t[m]\} - \sum_{(s,t) \in \mathcal{N}} P_{\boldsymbol{\theta}}(Y_s = Y_t \mid \boldsymbol{x}[m])$$

Computation

MRF

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} \ell(\boldsymbol{\theta} : \mathcal{D}) = \boldsymbol{E}_{\mathcal{D}}[f_i(\boldsymbol{X})] - \boldsymbol{E}_{\boldsymbol{\theta}}[f_i]$$

- Requires inference at each gradient step

1

CRF

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} \ell_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{\theta} : \mathcal{D}) = \frac{1}{M} \sum_{m=1}^{M} \left(f_i(\boldsymbol{x}[m], \boldsymbol{y}[m]) - \boxed{\boldsymbol{E}_{\boldsymbol{\theta}}[f_i(\boxed{\boldsymbol{x}[m]}, \boldsymbol{Y})]} \right)$$

- Requires inference for each x[m] at each gradient step

M = # training instances

Daphne Koller

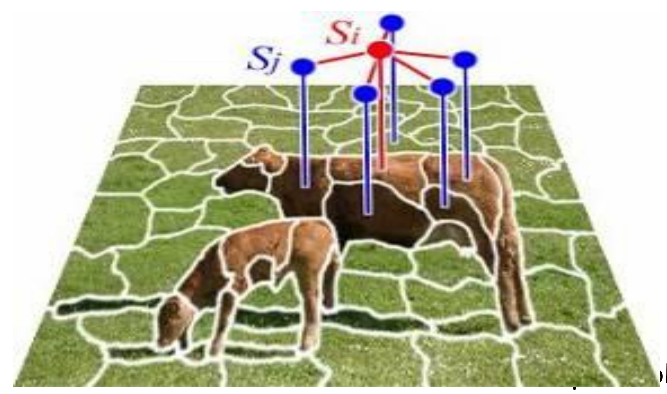
However…

- For inference of $P(\mathbf{Y} \mid \mathbf{x})$, we need to compute distribution only over $\mathbf{Y}$
- If we learn an MRF, need to compute $P(\mathbf{Y},\mathbf{X})$, which may be much more complex

$$f_1(Y_s, X_s) = \mathbf{1}(Y_s = g) \times G_s$$

$$f_2(Y_s, Y_t) = \mathbf{1}(Y_s = Y_t)$$

average intensity of green channel for pixels in superpixel i



Summary

- CRF learning very similar to MRF learning
  - Likelihood function is concave
  - Optimized using gradient ascent (usually L-BFGS)
- Gradient computation requires inference: one per gradient step, data instance
  - c.f., once per gradient step for MRFs
- But conditional model is often much simpler, so inference cost for CRF, MRF is not the same