10-418 / 10-618 Machine Learning for Structured Data

Machine Learning Department
School of Computer Science
Carnegie Mellon University

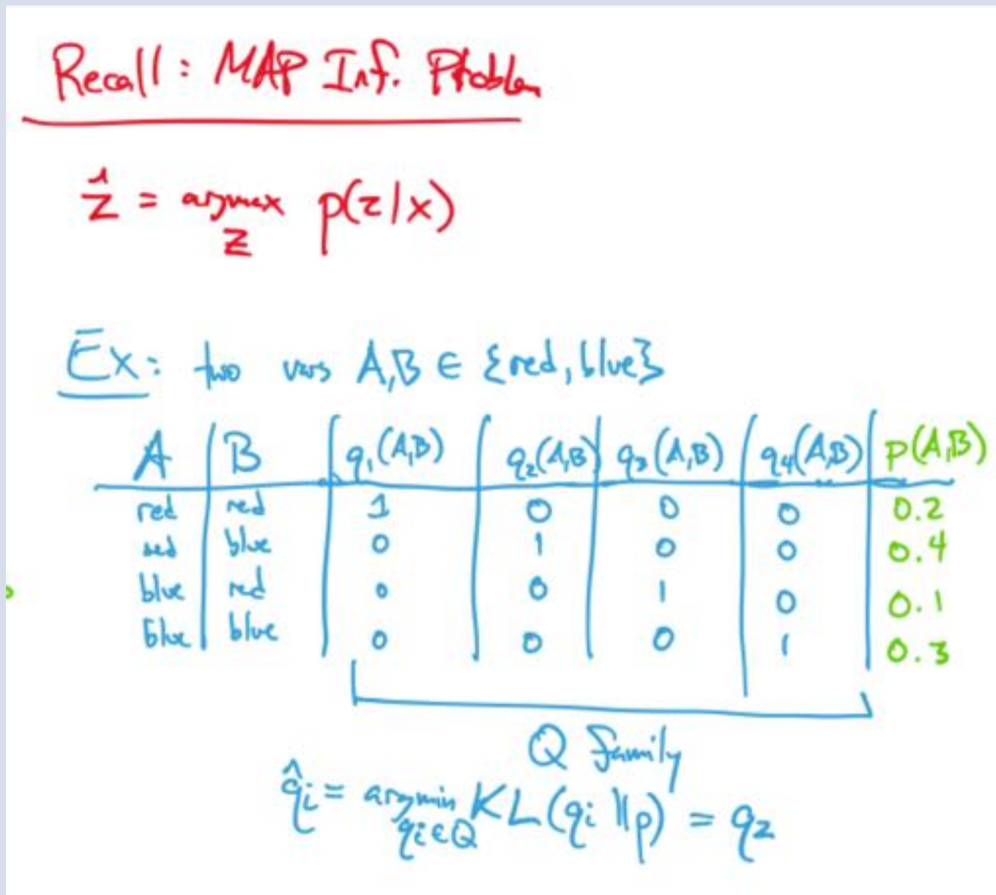
Variational Inference

Matt Gormley
Lecture 24
Nov. 18, 2019

Q&A

Q: How does the reduction of MAP Inference to Variational Inference work again…?

A: Let's look at an example…



Reminders

- Homework 4: Topic Modeling
  - Out: Wed, Nov. 6
  - Due: Mon, Nov. 18 at 11:59pm
- Homework 5: Variational Inference
  - Out: Mon, Nov. 18
  - Due: Mon, Nov. 25 at 11:59pm
- 618 Midway Poster:
  - Submission: Thu, Nov. 21 at 11:59pm
  - Presentation: Fri, Nov. 22 or Mon, Nov. 25

MEAN FIELD VARIATIONAL INFERENCE

Variational Inference

Whiteboard

– Background: KL Divergence
– Mean Field Variational Inference (overview)
– Evidence Lower Bound (ELBO)
– ELBO's relation to log p(x)

Variational Inference

Whiteboard

- Mean Field Variational Inference (derivation)
- Algorithm Summary (CAVI)
- Example: Factor Graph with Discrete Variables

Variational Inference

Whiteboard

- Example: two variable factor graph
  - Iterated Conditional Models
  - Gibbs Sampling
  - Mean Field Variational Inference

An example of why we need approximate inference

EXACT INFERENCE ON GRID CRF

Application: Pose Estimation

$\phi_i(y_i, x) \in \mathbb{R}^{\approx 1000}$: local image representation, e.g. HoG

$\rightarrow \quad \langle w_i, \phi_i(y_i, x) \rangle$: local confidence map

$\phi_{i,j}(y_i, y_j) = \mathit{good\_fit}(y_i, y_j) \in \mathbb{R}^1$: test for geometric fit

$\rightarrow \quad \langle w_{ij}, \phi_{ij}(y_i, y_j) \rangle$: penalizer for unrealistic poses

together: $\operatorname{argmax}_y p(y|x)$ is sanitized version of local cues

original | local classification | local + geometry

Feature Functions for CRF in Vision

$\phi_i(y_i, x)$: local representation, high-dimensional

$\rightarrow \quad \langle w_i, \phi_i(y_i, x)\rangle$: local classifier

$\phi_{i,j}(y_i, y_j)$: prior knowledge, low-dimensional

$\rightarrow \quad \langle w_{ij}, \phi_{ij}(y_i, y_j)\rangle$: penalize outliers

learning adjusts parameters:

- unary w_i: learn local classifiers and their importance
- binary w_{ij}: learn importance of smoothing/penalization

$\mathrm{argmax}_y\, p(y|x)$ is cleaned up version of local prediction

Case Study: Image Segmentation

- Image segmentation (FG/BG) by modeling of interactions btw RVs
  – Images are noisy.
  – Objects occupy continuous regions in an image.

[Nowozin,Lampert 2012]

Input image

Pixel-wise separate optimal labeling

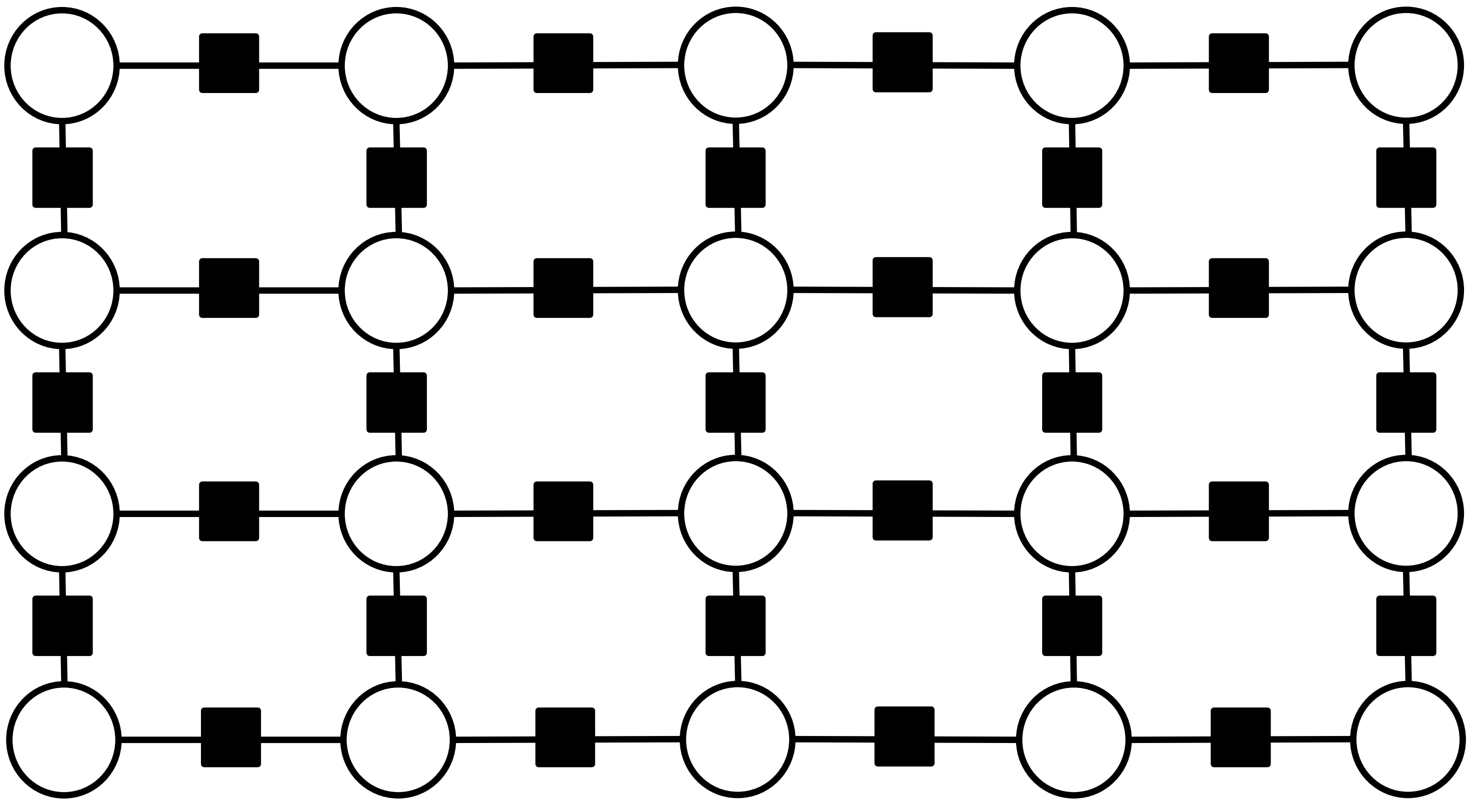
Locally-consistent joint optimal labeling

$$Y^* = \arg\max_{y\in\{0,1\}^n} \Bigg[\underbrace{\sum_{i\in S} V_i(y_i, X)}_{\text{Unary Term}} + \underbrace{\sum_{i\in S}\sum_{j\in N_i} V_{i,j}(y_i, y_j)}_{\text{Pairwise Term}} \Bigg].$$

Y: labels
X: data (features)
S: pixels
N_i: neighbors of pixel i

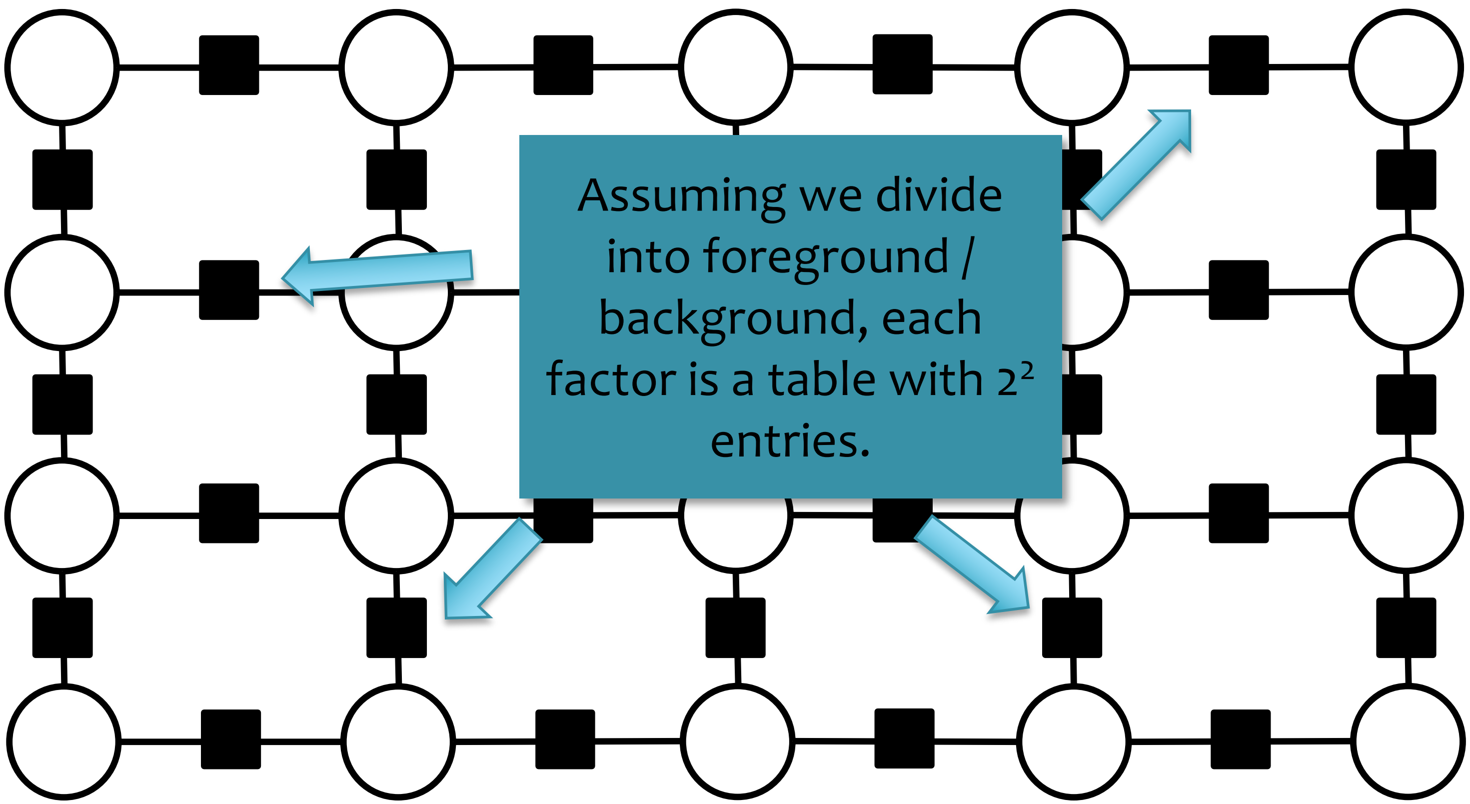
Grid CRF

- Suppose we want to image segmentation using a grid model



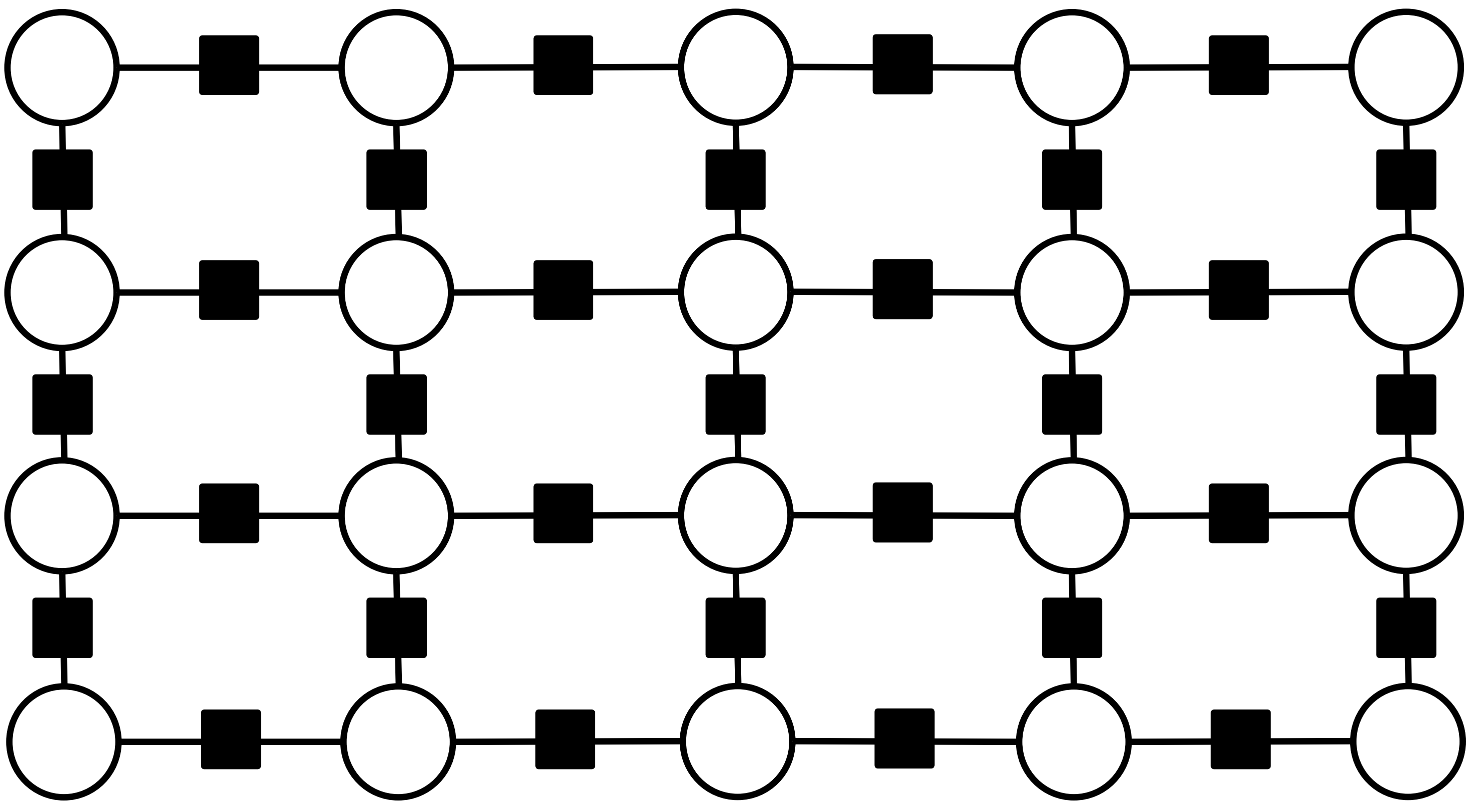
Grid CRF

- Suppose we want to image segmentation using a grid model



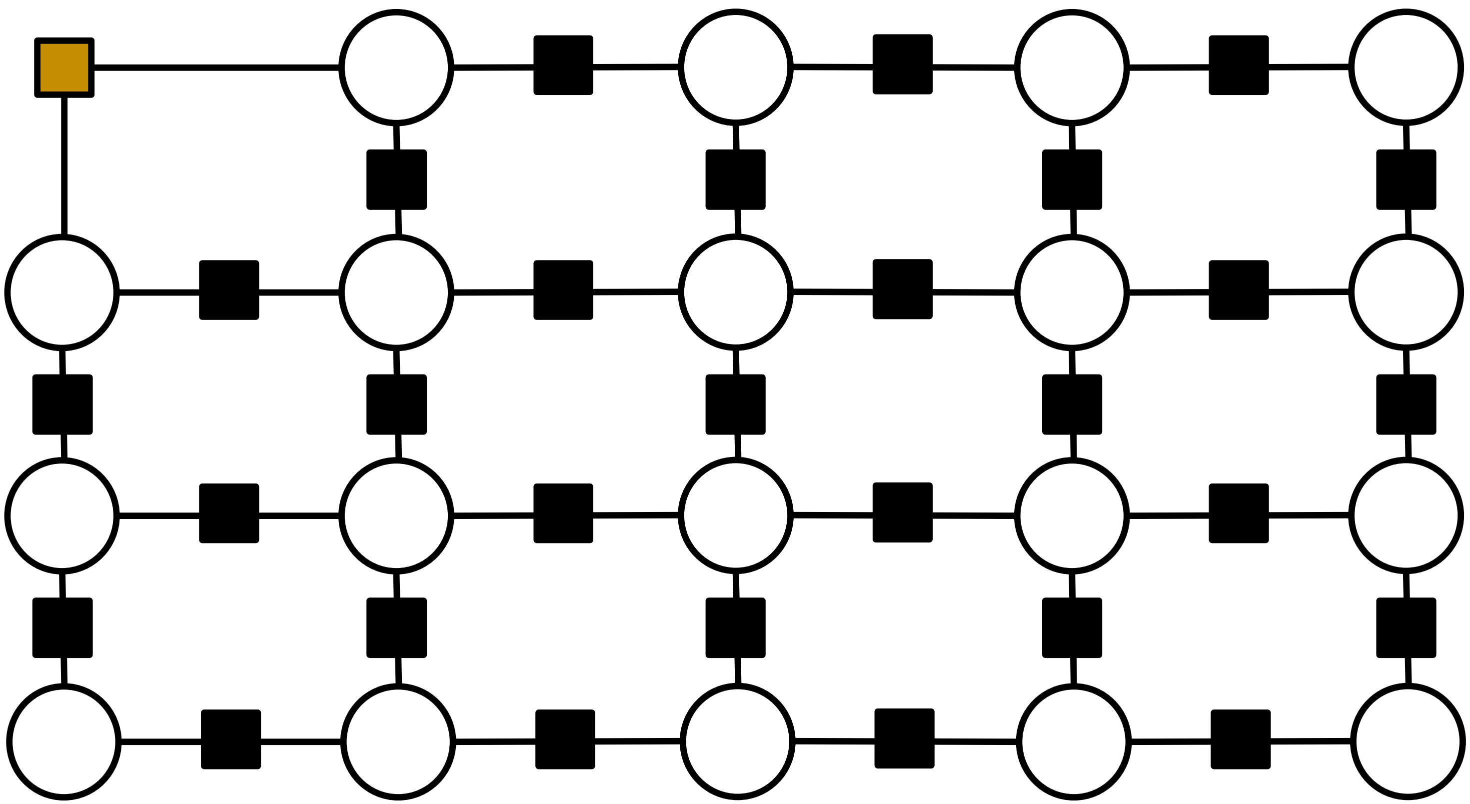
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



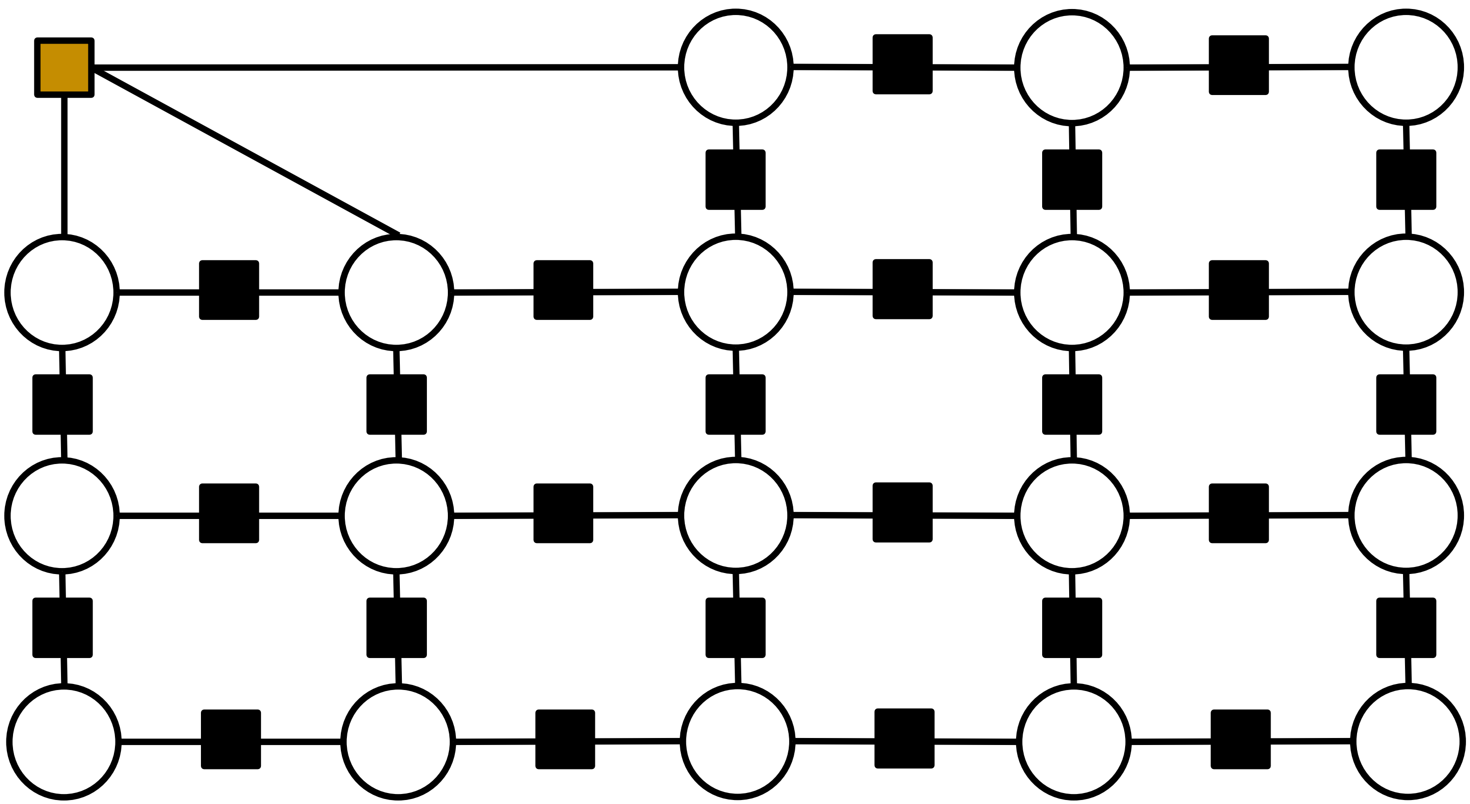
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



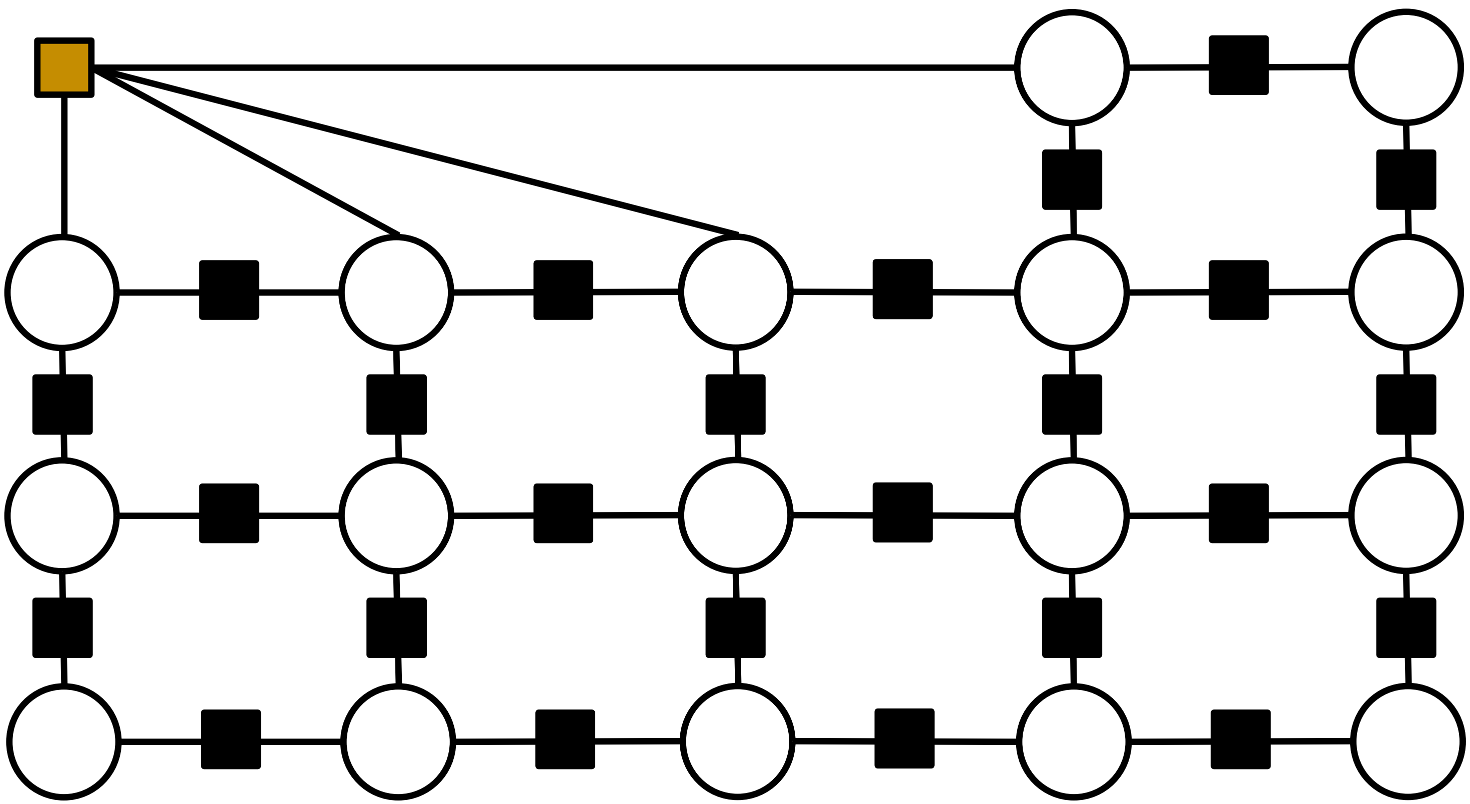
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



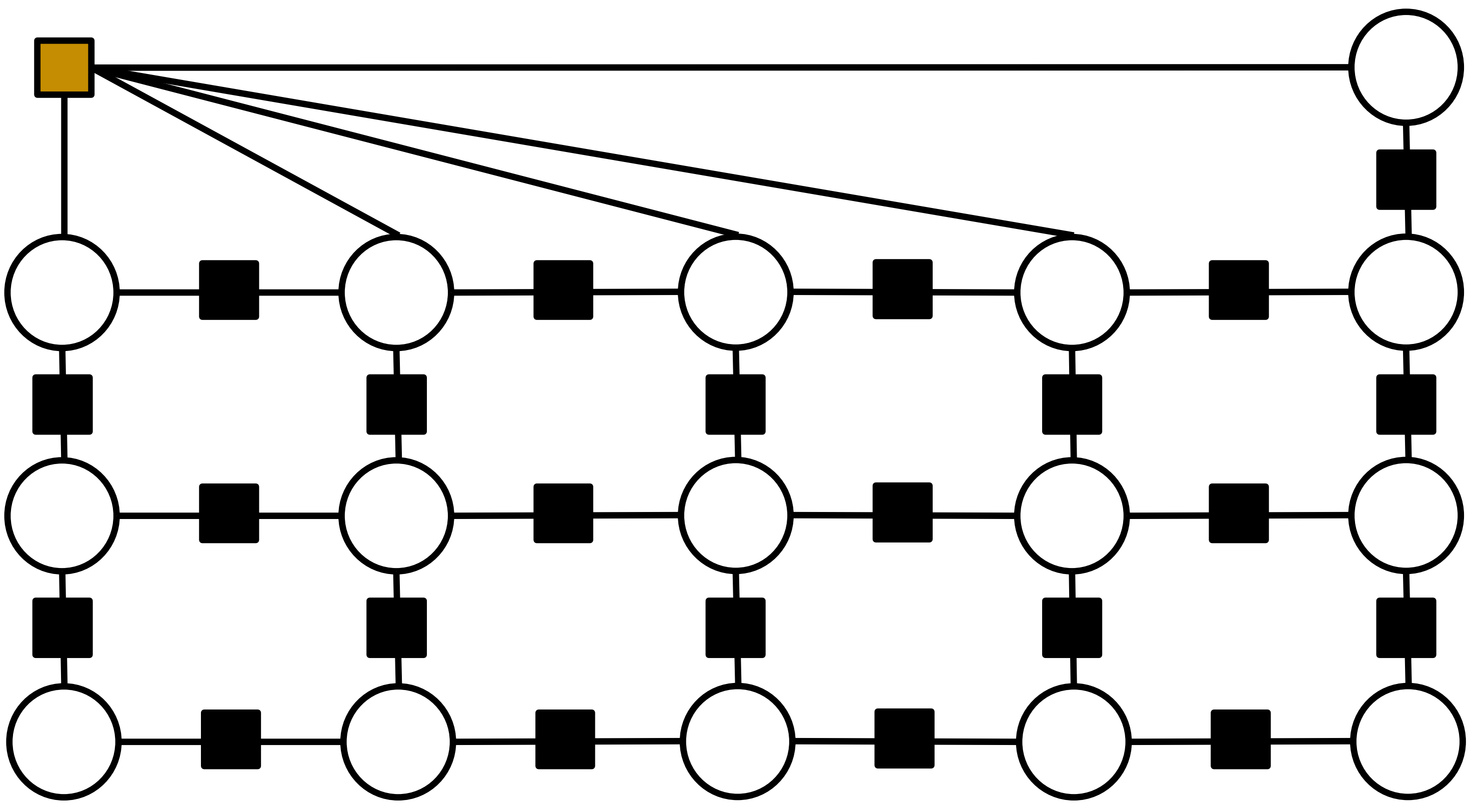
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



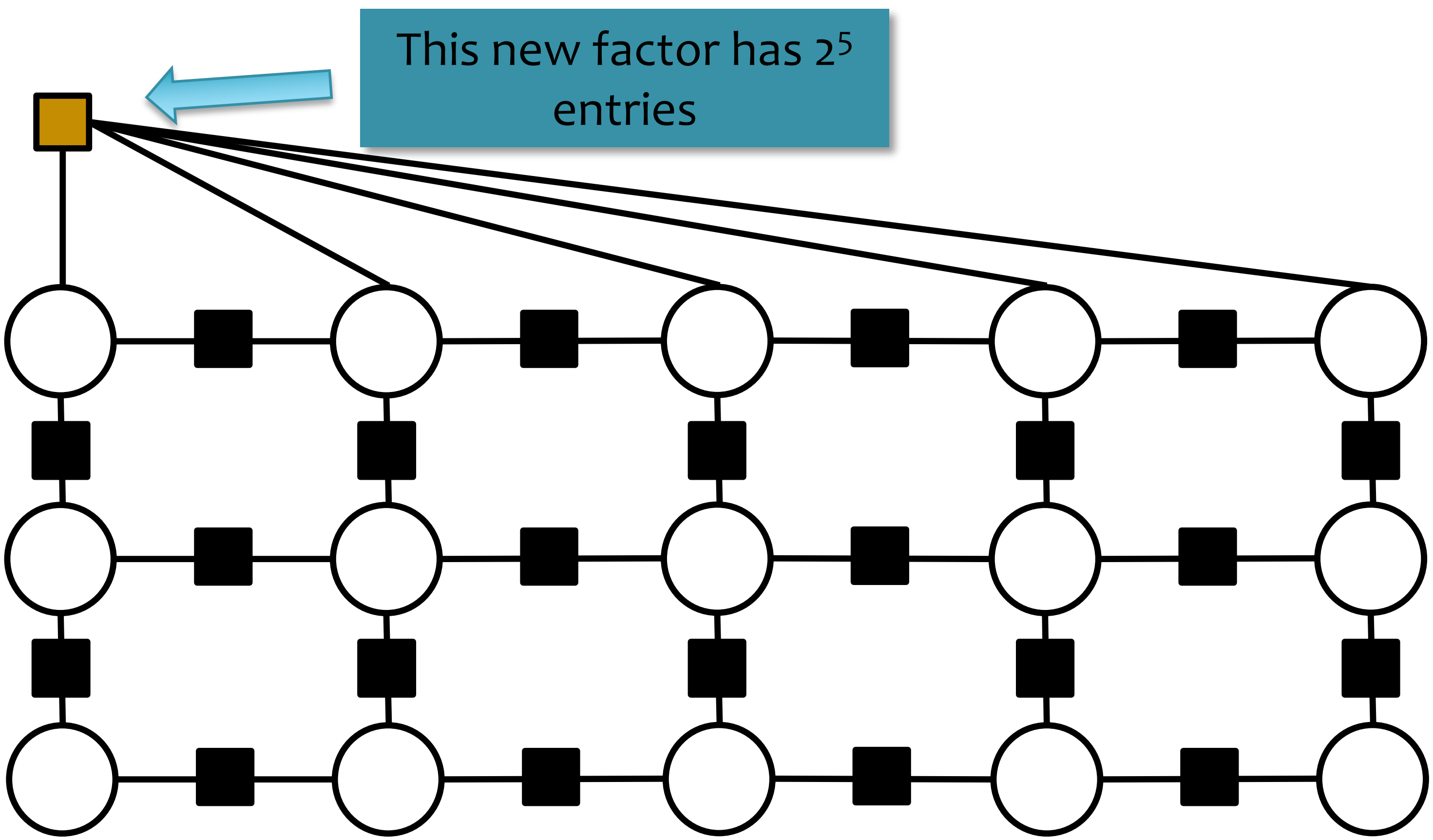
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



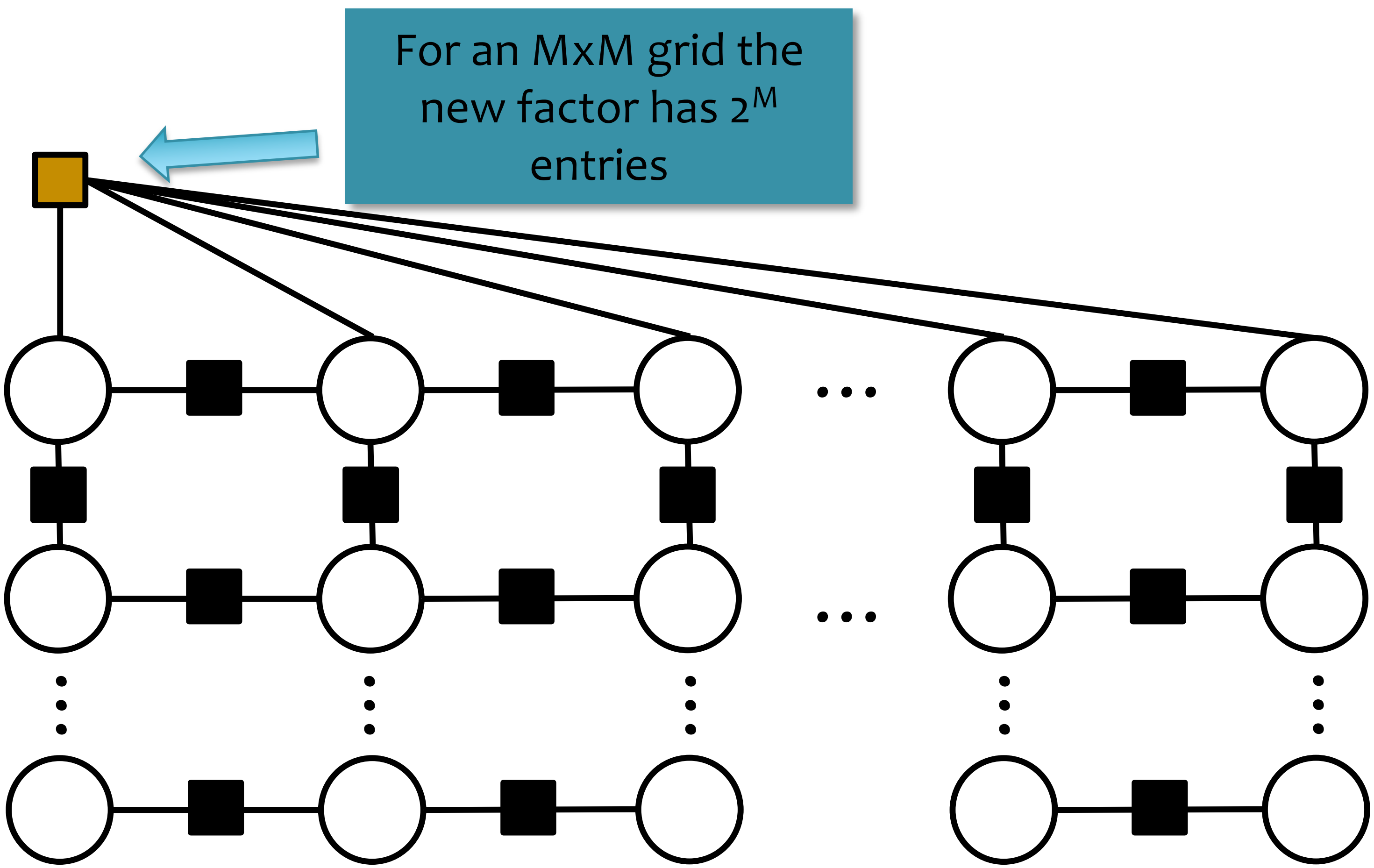
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



Grid CRF

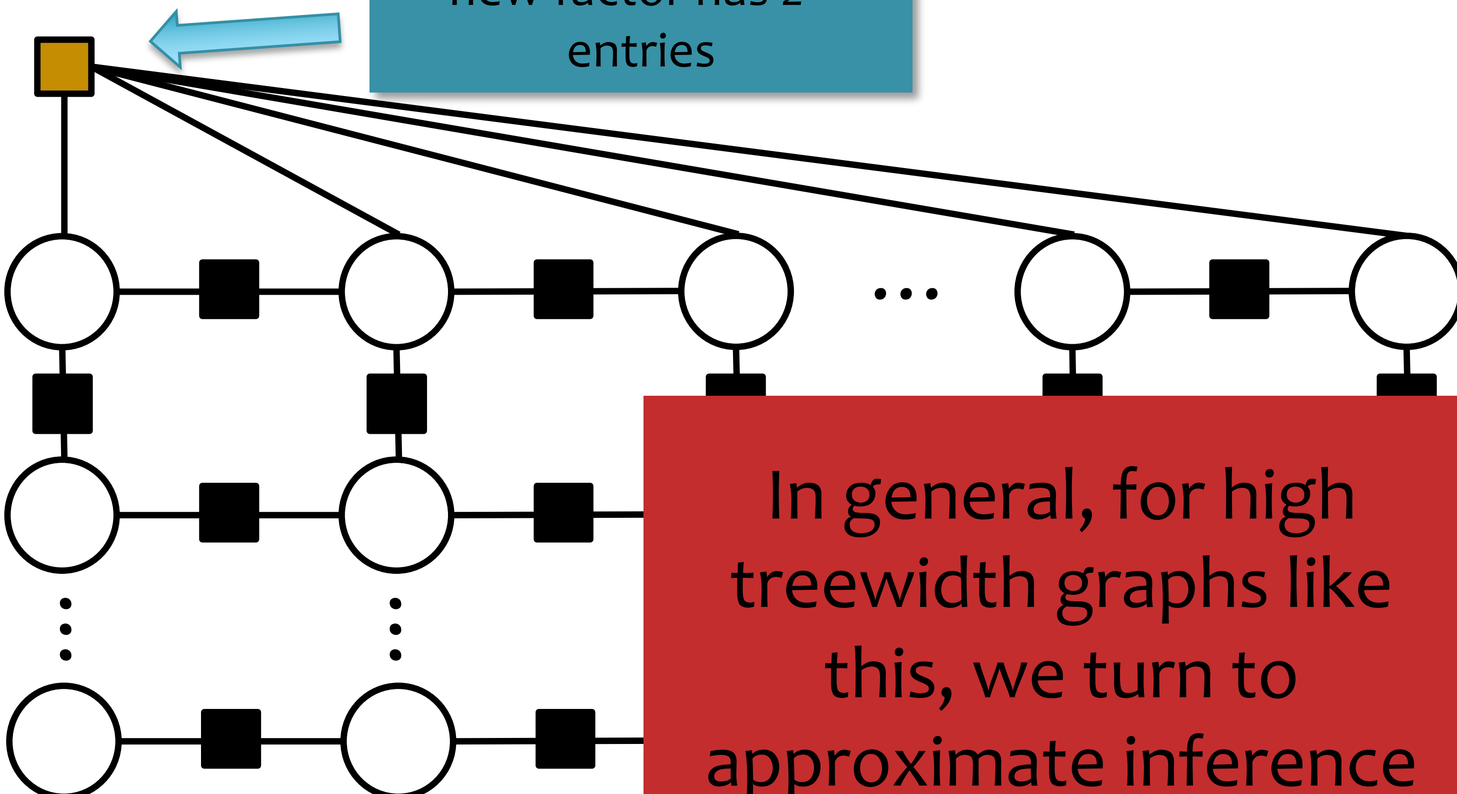
- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?



Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?

For an MxM grid the new factor has 2^M entries

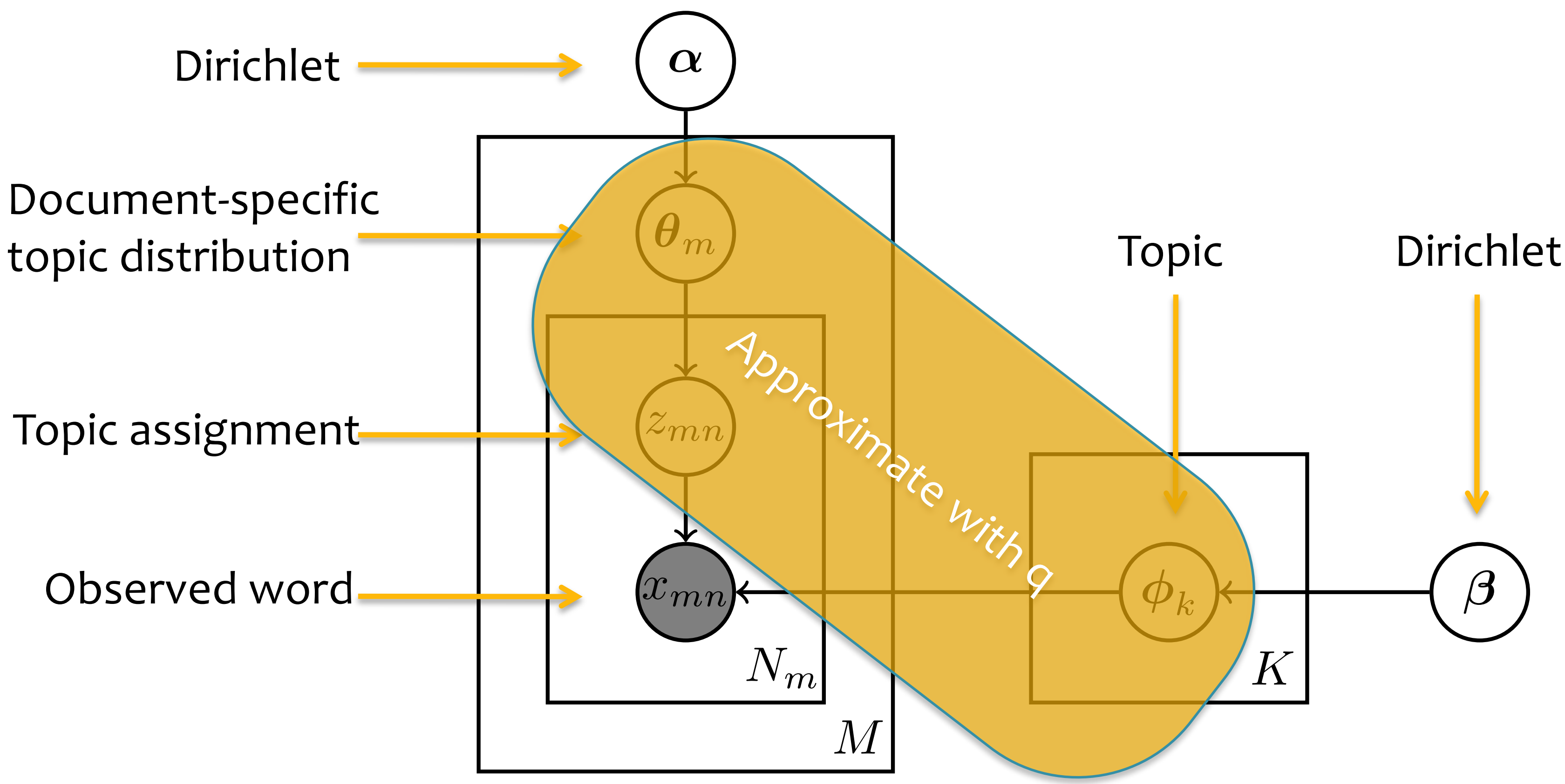


In general, for high treewidth graphs like this, we turn to approximate inference (which we'll cover soon!)

VARIATIONAL INFERENCE RESULTS

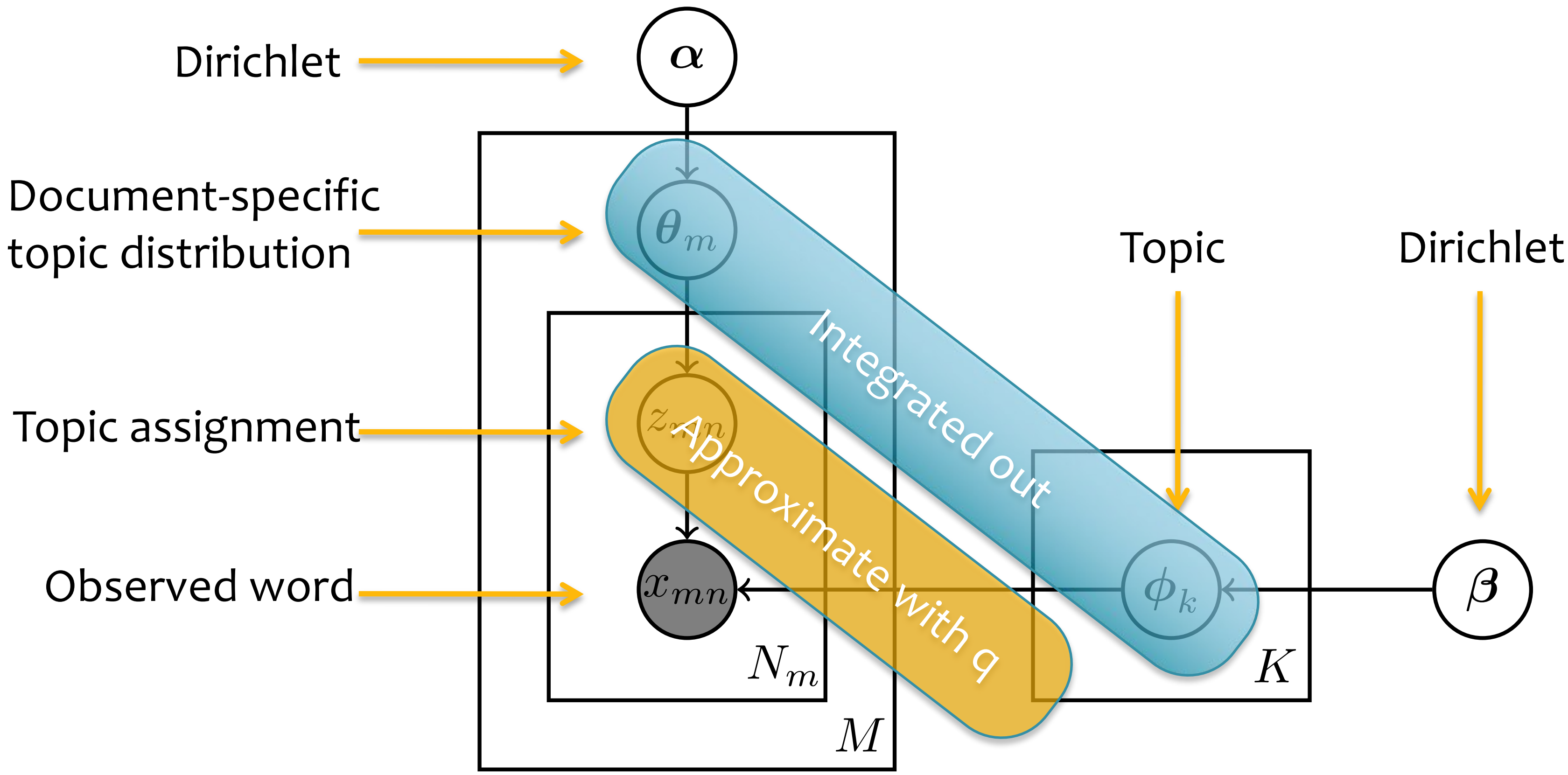
Collapsed Variational Bayesian LDA

- Explicit Variational Inference



Collapsed Variational Bayesian LDA

- Collapsed Variational Inference

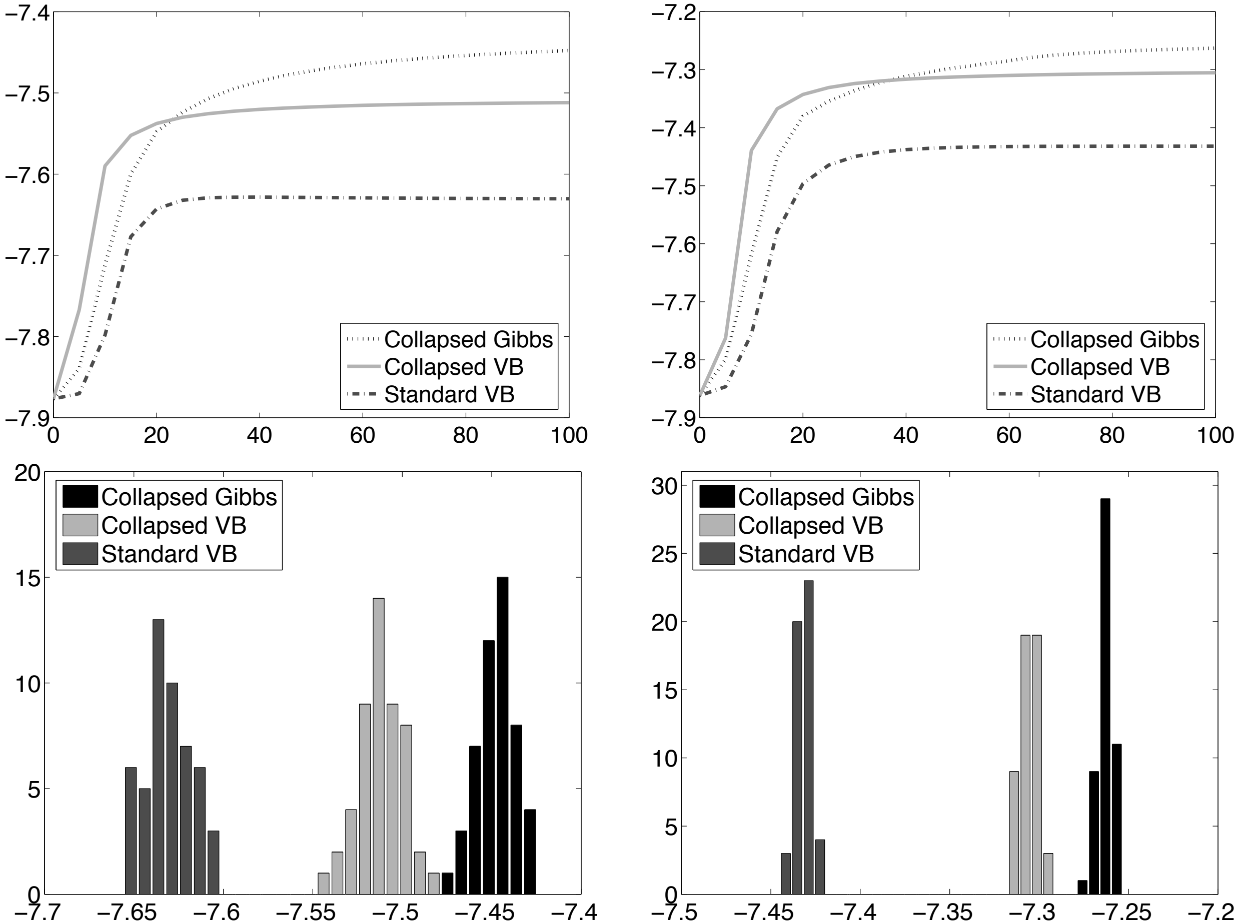


Collapsed Variational Bayesian LDA

- **First row:** test set per word log probabilities as functions of numbers of iterations for VB, CVB and Gibbs.
- **Second row:** histograms of final test set per word log probabilities across 50 random initializations.

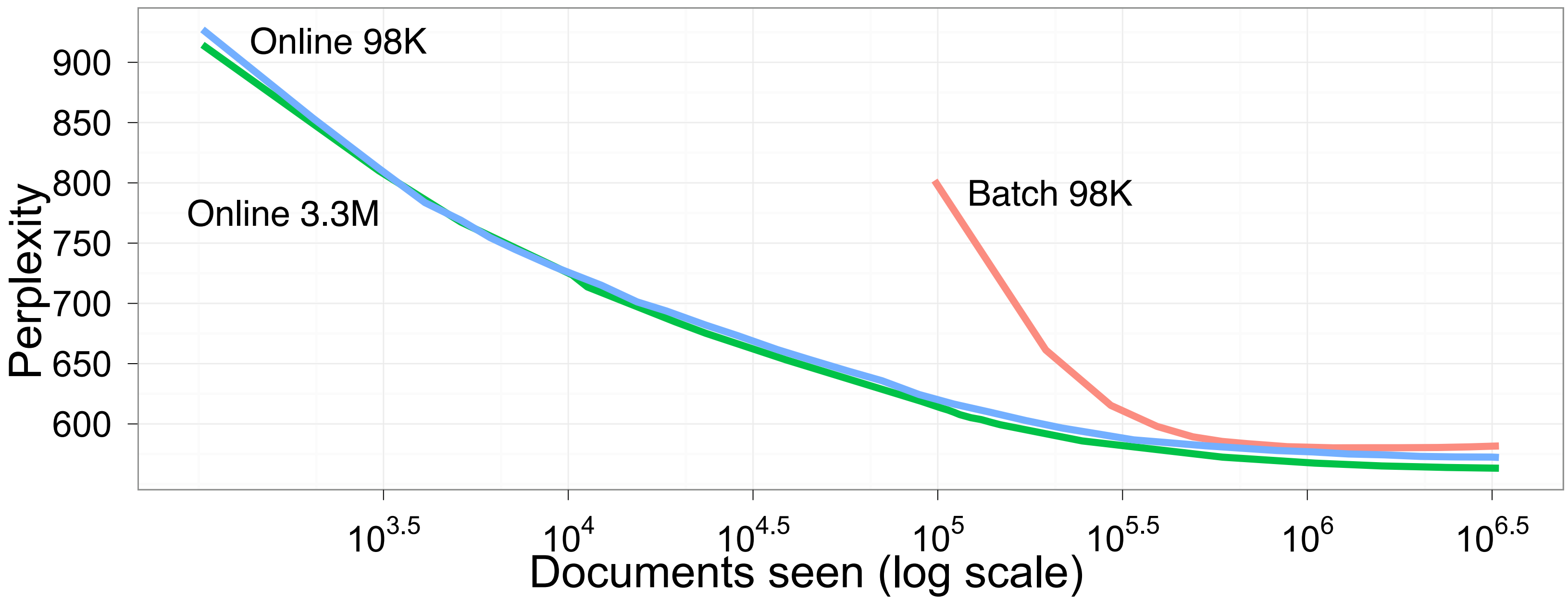
Data from dailykos.com

Data from NeurIPS proceedings



Slide from Teh et al (2007)

Online Variational Bayes for LDA



Documents analyzed	**2048**	**4096**	**8192**	**12288**	**16384**	**32768**	**49152**	**65536**
Top eight words	systems	systems	service	service	service	business	business	business
	road	health	systems	systems	companies	service	service	industry
	made	communication	health	companies	systems	companies	companies	service
	service	service	companies	business	business	industry	industry	companies
	announced	billion	market	company	company	company	services	services
	national	language	communication	billion	industry	management	company	company
	west	care	company	health	market	systems	management	management
	language	road	billion	industry	billion	services	public	public

Online Variational Bayes for LDA

Algorithm 1 Batch variational Bayes for LDA

Initialize $\boldsymbol{\lambda}$ randomly.
while relative improvement in $\mathcal{L}(\boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) > 0.00001$ **do**
  E step:
  for $d = 1$ to D **do**
    Initialize $\gamma_{dk} = 1$. (The constant 1 is arbitrary.)
    repeat
      Set $\phi_{dwk} \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log \beta_{kw}]\}$
      Set $\gamma_{dk} = \alpha + \sum_w \phi_{dwk} n_{dw}$
    until $\frac{1}{K} \sum_k |\text{change in} \gamma_{dk}| < 0.00001$
  end for
  M step:
  Set $\lambda_{kw} = \eta + \sum_d n_{dw} \phi_{dwk}$
end while

Algorithm 2 Online variational Bayes for LDA

Define $\rho_t \triangleq (\tau_0 + t)^{-\kappa}$
Initialize $\boldsymbol{\lambda}$ randomly.
for $t = 0$ to ∞ **do**
  E step:
  Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{twk} \propto \exp\{\mathbb{E}_q[\log \theta_{tk}] + \mathbb{E}_q[\log \beta_{kw}]\}$
    Set $\gamma_{tk} = \alpha + \sum_w \phi_{twk} n_{tw}$
  until $\frac{1}{K} \sum_k |\text{change in} \gamma_{tk}| < 0.00001$
  M step:
  Compute $\tilde{\lambda}_{kw} = \eta + D n_{tw} \phi_{twk}$
  Set $\boldsymbol{\lambda} = (1 - \rho_t)\boldsymbol{\lambda} + \rho_t \tilde{\boldsymbol{\lambda}}$.
end for

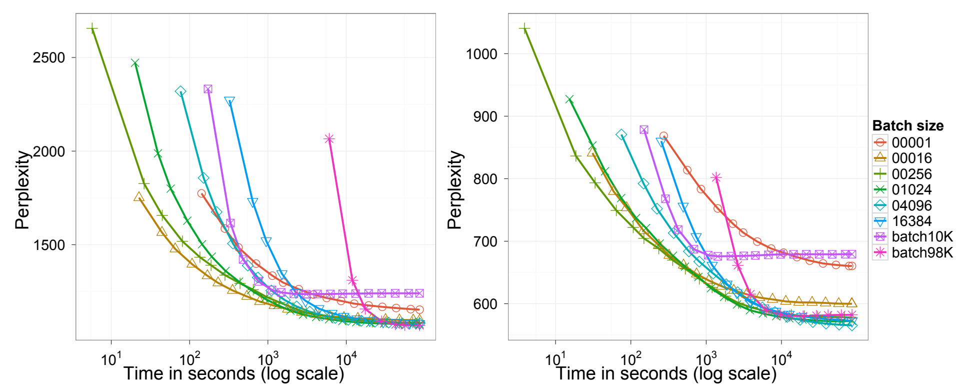


Figure 2: Held-out perplexity obtained on the *Nature* (left) and Wikipedia (right) corpora as a function of CPU time. For moderately large mini-batch sizes, online LDA finds solutions as good as those that the batch LDA finds, but with much less computation. When fit to a 10,000-document subset of the training corpus batch LDA's speed improves, but its performance suffers.

Figures from Hoffman et al. (2010)

Fully–Connected CRF

Model

$$p(\mathbf{x}|\mathbf{i}) = \frac{1}{Z(\mathbf{i})} \exp(-E(\mathbf{x}))$$

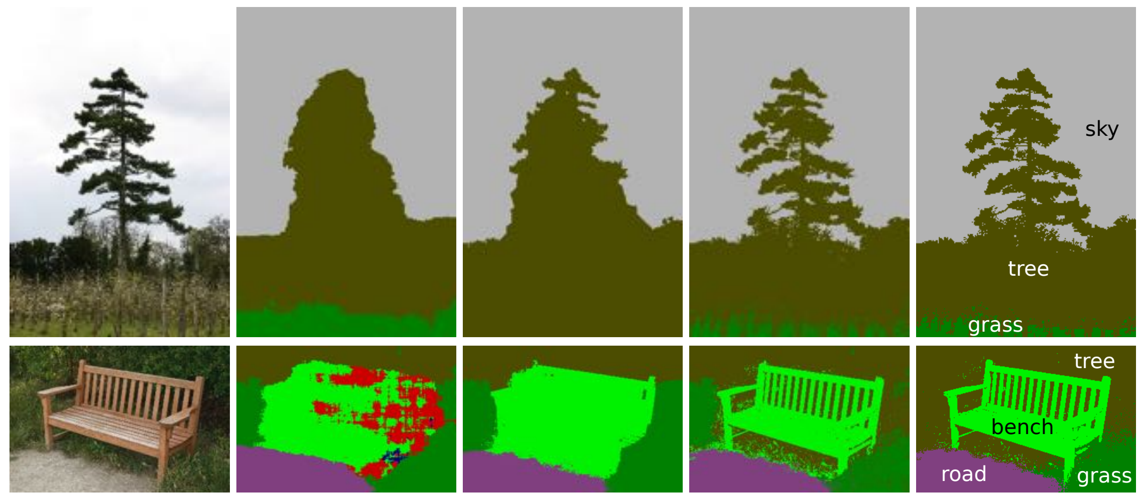
$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j),$$

This is a fully connected graph!

Inference

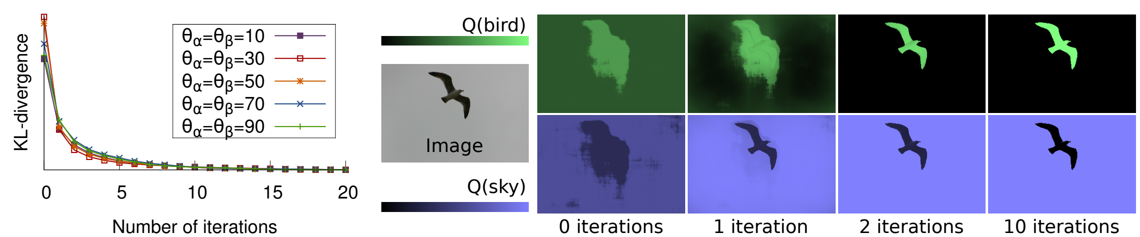
- Can do MCMC, but slow
- Instead use Variational Inference
- Then filter some variables for speed up

Results



(a) Image (b) Unary classifiers (c) Robust P^n CRF (d) Fully connected CRF, MCMC inference, 36 hrs (e) Fully connected CRF, our approach, 0.2 seconds

Figure 1: Pixel-level classification with a fully connected CRF. (a) Input image from the MSRC-21 dataset. (b) The response of unary classifiers used by our models. (c) Classification produced by the Robust P^n CRF [9]. (d) Classification produced by MCMC inference [17] in a fully connected pixel-level CRF model; the algorithm was run for 36 hours and only partially converged for the bottom image. (e) Classification produced by our inference algorithm in the fully connected model in 0.2 seconds.



(a) KL-divergence (b) Distributions $Q(X_i = \text{"bird"})$ (top) and $Q(X_i = \text{"sky"})$ (bottom)

Figure 2: Convergence analysis. (a) KL-divergence of the mean field approximation during successive iterations of the inference algorithm, averaged across 94 images from the MSRC-21 dataset. (b) Visualization of convergence on distributions for two class labels over an image from the dataset.

Figures from Krähenbühl & Koltun (2011)

Fully–Connected CRF

Model

$$p(\mathbf{x}|\mathbf{i}) = \frac{1}{Z(\mathbf{i})} \exp(-E(\mathbf{x}))$$

$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j),$$

This is a fully connected graph!

Inference

- Can do MCMC, but slow
- Instead use Variational Inference
- Then filter some variables for speed up

Follow-up Work (combine with CNN)

Published as a conference paper at ICLR 2015

SEMANTIC IMAGE SEGMENTATION WITH DEEP CONVOLUTIONAL NETS AND FULLY CONNECTED CRFS

Liang-Chieh Chen
Univ. of California, Los Angeles
lcchen@cs.ucla.edu

George Papandreou *
Google Inc.
gpapan@google.com

Iasonas Kokkinos
CentraleSupélec and INRIA
iasonas.kokkinos@ecp.fr

Kevin Murphy
Google Inc.
kpmurphy@google.com

Alan L. Yuille
Univ. of California, Los Angeles
yuille@stat.ucla.edu

ABSTRACT

Deep Convolutional Neural Networks (DCNNs) have recently shown state of the art performance in high level vision tasks, such as image classification and object detection. This work brings together methods from DCNNs and probabilistic graphical models for addressing the task of pixel-level classification (also called "semantic image segmentation"). We show that responses at the final layer of DCNNs are not sufficiently localized for accurate object segmentation. This is due to the very invariance properties that make DCNNs good for high level tasks. We overcome this poor localization property of deep networks by combining the responses at the final DCNN layer with a fully connected Conditional Random Field (CRF). Qualitatively, our "DeepLab" system is able to localize segment boundaries at a level of accuracy which is beyond previous methods. Quantitatively, our method sets the new state-of-art at the PASCAL VOC-2012 semantic image segmentation task, reaching 71.6% IOU accuracy in the test set. We show how these results can be obtained efficiently: Careful network re-purposing and a novel application of the 'hole' algorithm from the wavelet community allow dense computation of neural net responses at 8 frames per second on a modern GPU.

Figures from Krähenbühl & Koltun (2011)

Joint Parsing and Alignment with Weakly Synchronized Grammars

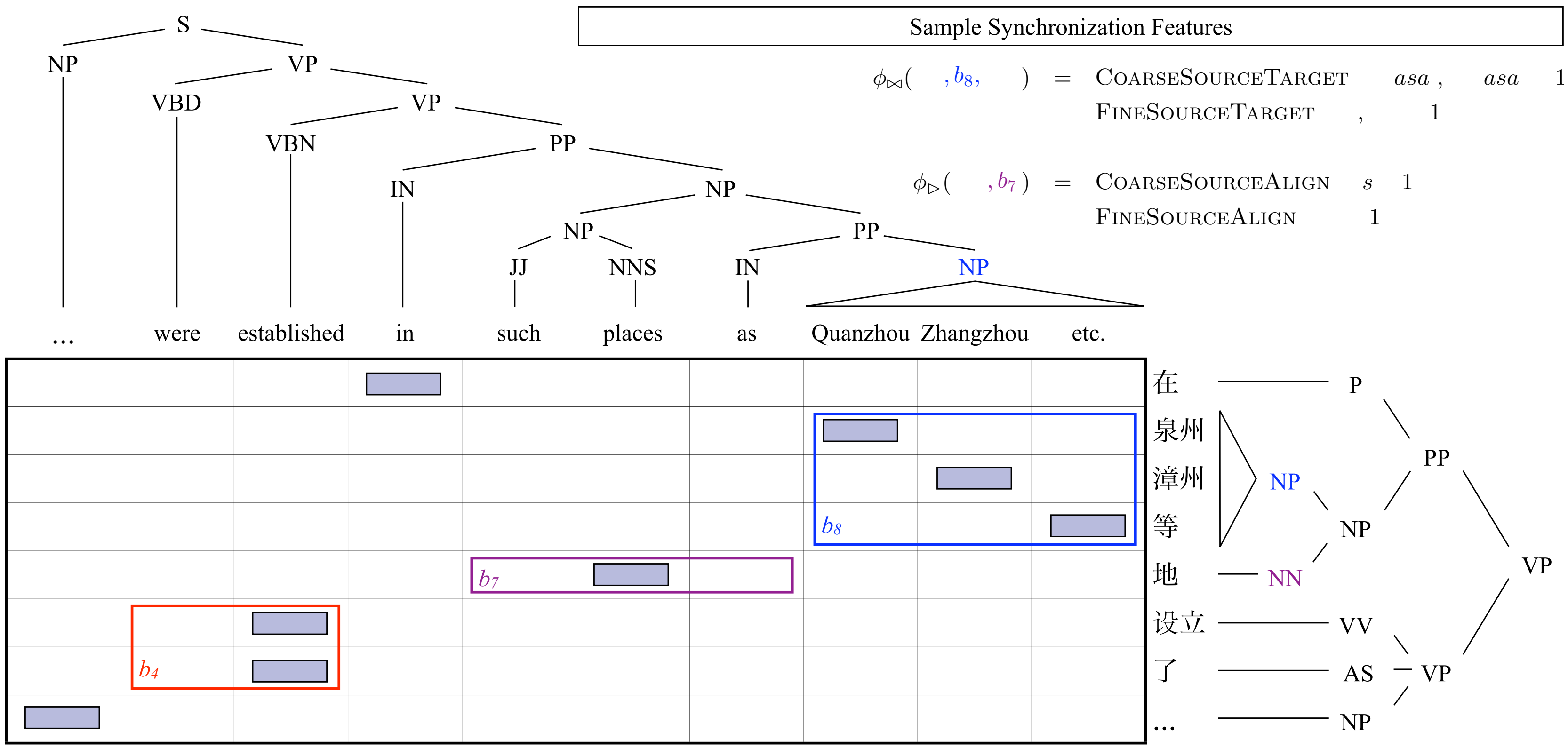
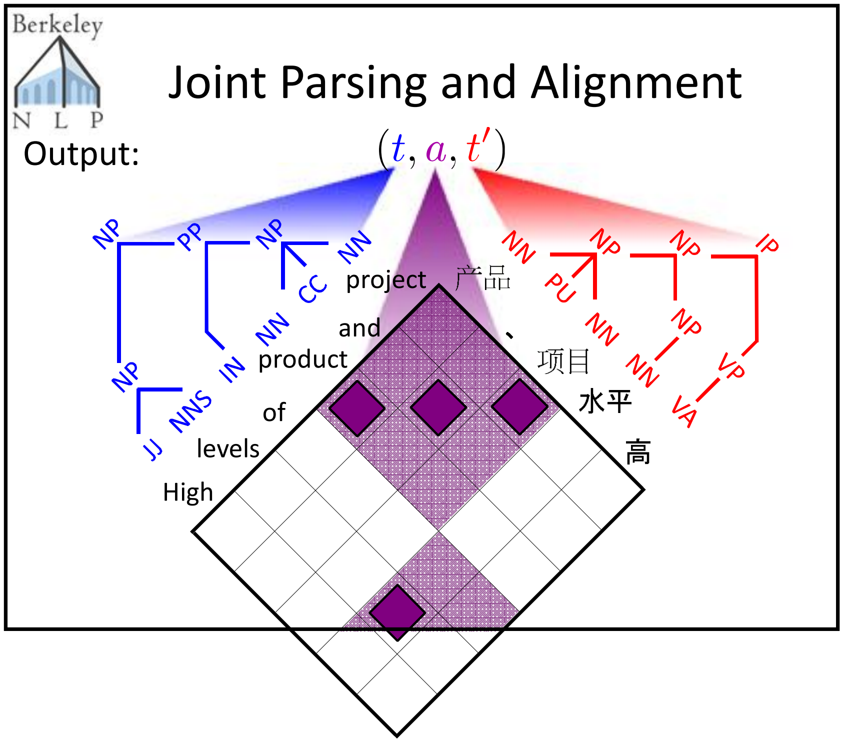


Figure 2: An example of a Chinese-English sentence pair with parses, word alignments, and a subset of the full optimal ITG derivation, including one totally unsynchronized bispan (b_4), one partially synchronized bispan (b_7), and and fully synchronized bispan (b_8). The inset provides some examples of active synchronization features (see Section 4.3) on these bispans. On this example, the monolingual English parser erroneously attached the lower PP to the VP headed by *established*, and the non-syntactic ITG word aligner misaligned 等 to *such* instead of to *etc.* Our joint model corrected both of these mistakes because it was rewarded for the synchronization of the two NPs joined by b_8.

Figures from Burkett et al. (2010)

Joint Parsing and Alignment with Weakly Synchronized Grammars



Figures from Burkett & Klein (ACL 2013 tutorial)

	Test Results		
	Ch F_1	Eng F_1	Tot F_1
Monolingual	83.6	81.2	82.5
Reranker	**86.0**	83.8	84.9
Joint	85.7	**84.5**	**85.1**

Table 1: Parsing results. Our joint model has the highest reported F_1 for English-Chinese bilingual parsing.

	Test Results			
	Precision	Recall	AER	F_1
HMM	86.0	58.4	30.0	69.5
ITG	**86.8**	73.4	20.2	79.5
Joint	85.5	**84.6**	**14.9**	**85.0**

Table 2: Word alignment results. Our joint model has the highest reported F_1 for English-Chinese word alignment.

Figures from Burkett et al. (2010)