ECE 6504: Advanced Topics in Machine Learning

Probabilistic Graphical Models and Large-Scale Learning

Topics

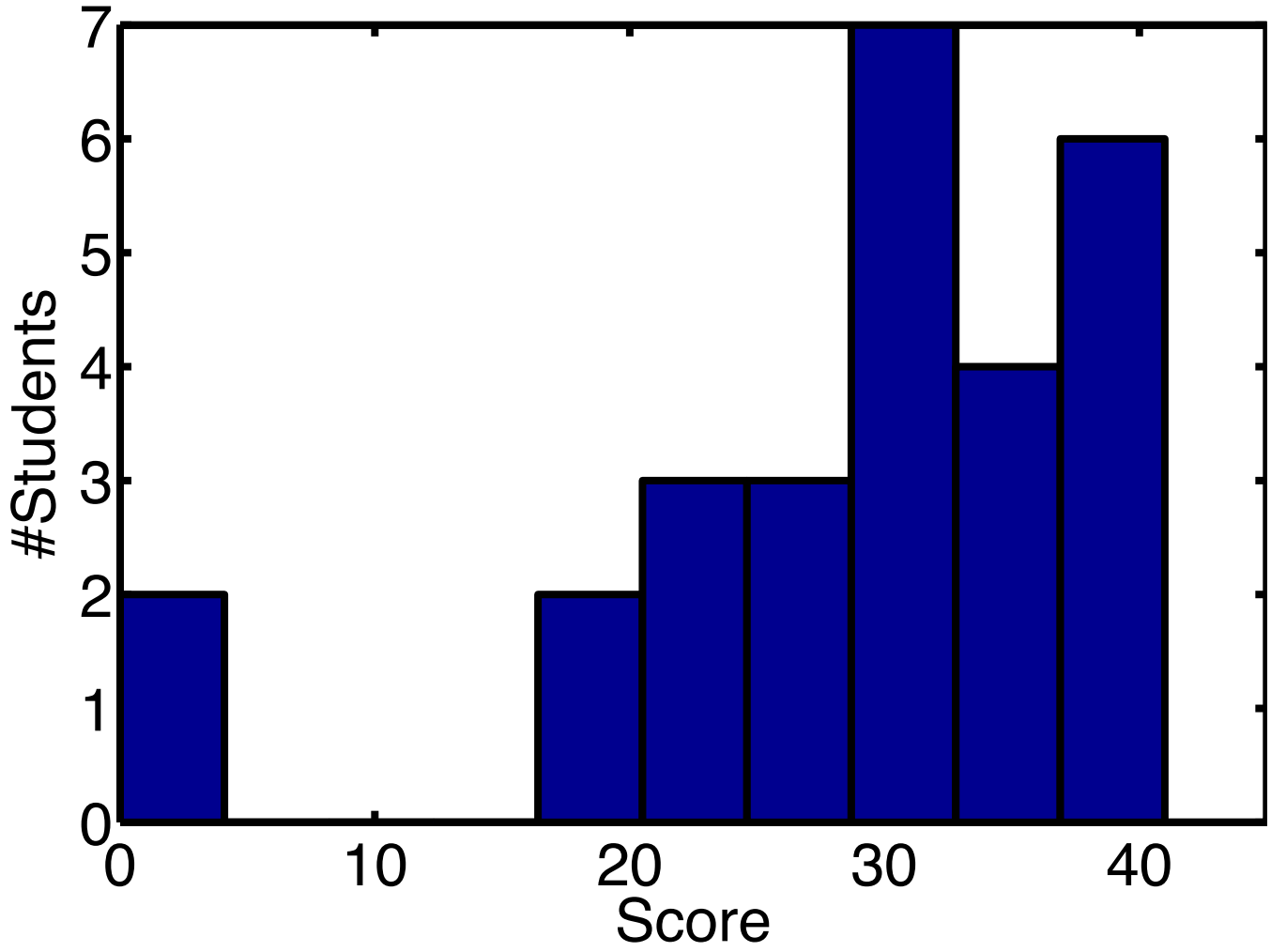
- Summary of Class
- Advanced Topics

Dhruv Batra

Virginia Tech

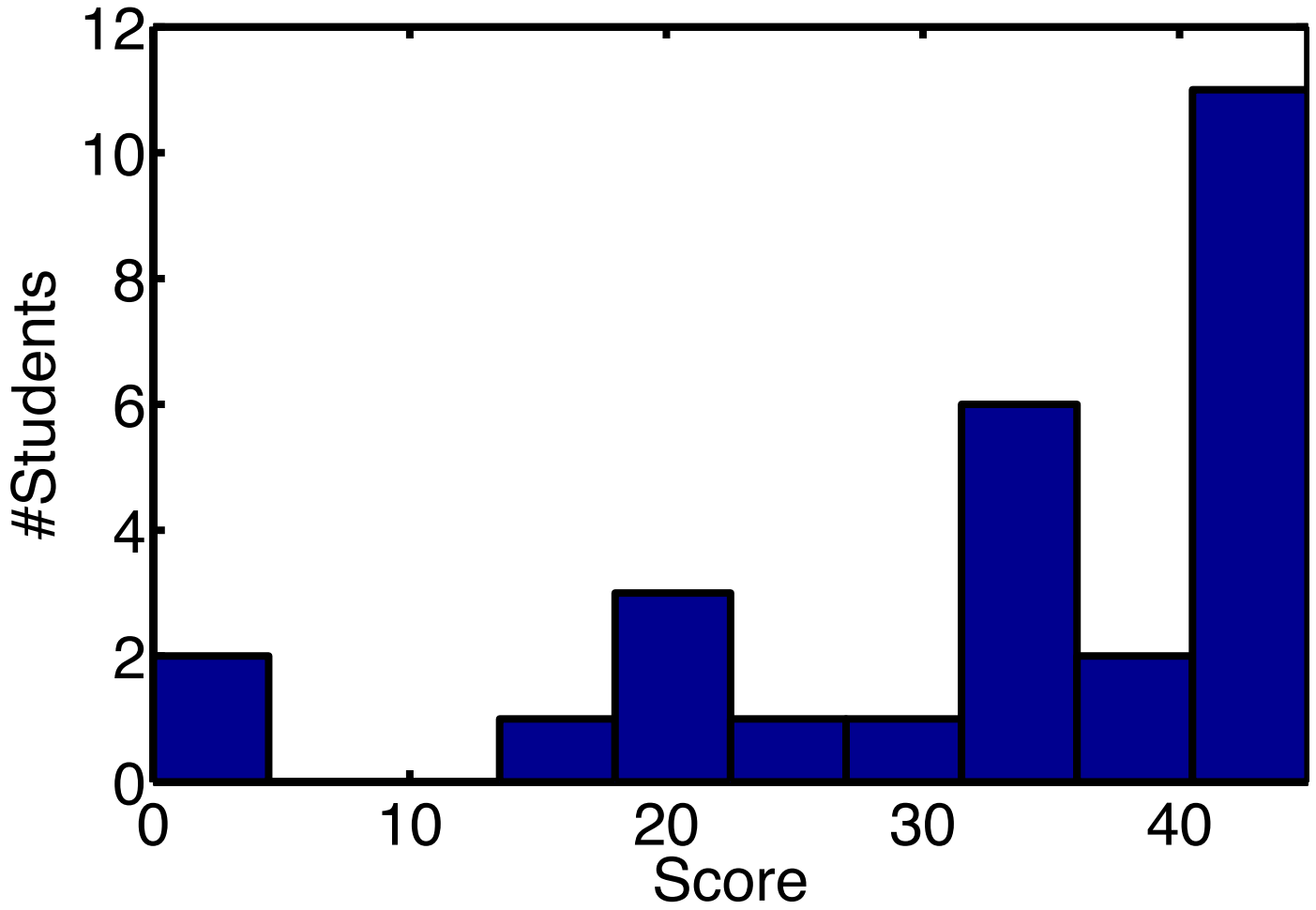
HW1 Grades

- Mean: 28.5/38 ~= 74.9%



HW2 Grades

- Mean: 33.2/45 ~= 73.8%



Administrativia

- (Mini-)HW4
  - Out now
  - Due: May 7, 11:55pm
  - Implementation:
    - Parameter Learning with Structured SVMs and Cutting-Plane

- Final Project Webpage
  - Due: ~~May 7~~, May 13 11:55pm
  - ~~Can use late days~~ Can't use late days any more
  - 1-3 paragraphs
    - Goal
    - Illustrative figure
    - Approach
    - Results (with figures or tables)

- Take Home Final
  - Out: May 8
  - Due: May 13, 11:55pm
  - No late days
  - Open book, open notes, open internet. Cite your sources.
  - No discussions!

A look back: PGMs

- One of the most exciting advancements in statistical AI in the last 10-20 years
- Marriage
  - Graph Theory + Probability
- Compact representation for exponentially-large probability distributions
  - Exploit conditional independencies
- Generalize
  - naïve Bayes
  - logistic regression
  - Many more …

A look back: what you learnt

- Directed Graphical Models (Bayes Nets)
  - Representation: Directed Acyclic Graphs (DAGs), Conditional Probability Tables (CPTs), d-Separation, v-structures, Markov Blanket, I-Maps
  - Parameter Learning: MLE, MAP, EM
  - Structure Learning: ***Chow-Liu***, Decomposable scores, hill climbing
  - Inference: Marginals, MAP/MPE, ***Variable Elimination***

- Undirected Graphical Models (MRFs/CRFs)
  - Representation: Junction trees, Factor graphs, treewidth, Local Makov Assumptions, Moralization, Triangulation
  - Inference: ***Belief Propagation***, Message Passing, Linear Programming Relaxations, Dual-Decomposition, Variational Inference, Mean Field
  - Parameter Learning: MLE, gradient descent
  - Structured Prediction: ***Structured SVMs, Cutting-Plane training***

- Large-Scale Learning
  - Online learning: perceptrons, stochastic (sub-)gradients
  - Distributed Learning: Dual Decomposition, Alternating Direction Method of Multipliers (ADMM)

Main Issues in PGMs

- Representation
  - How do we store $P(X_1, X_2, \ldots, X_n)$
  - What does my model mean/imply/assume? (Semantics)
- Inference
  - How do I answer questions/queries with my model? such as
  - Marginal Estimation: $P(X_5 \mid X_1, X_4)$
  - Most Probable Explanation: argmax $P(X_1, X_2, \ldots, X_n)$
- Learning
  - How do we learn parameters and structure of $P(X_1, X_2, \ldots, X_n)$ from data?
  - What model is the right for my data?

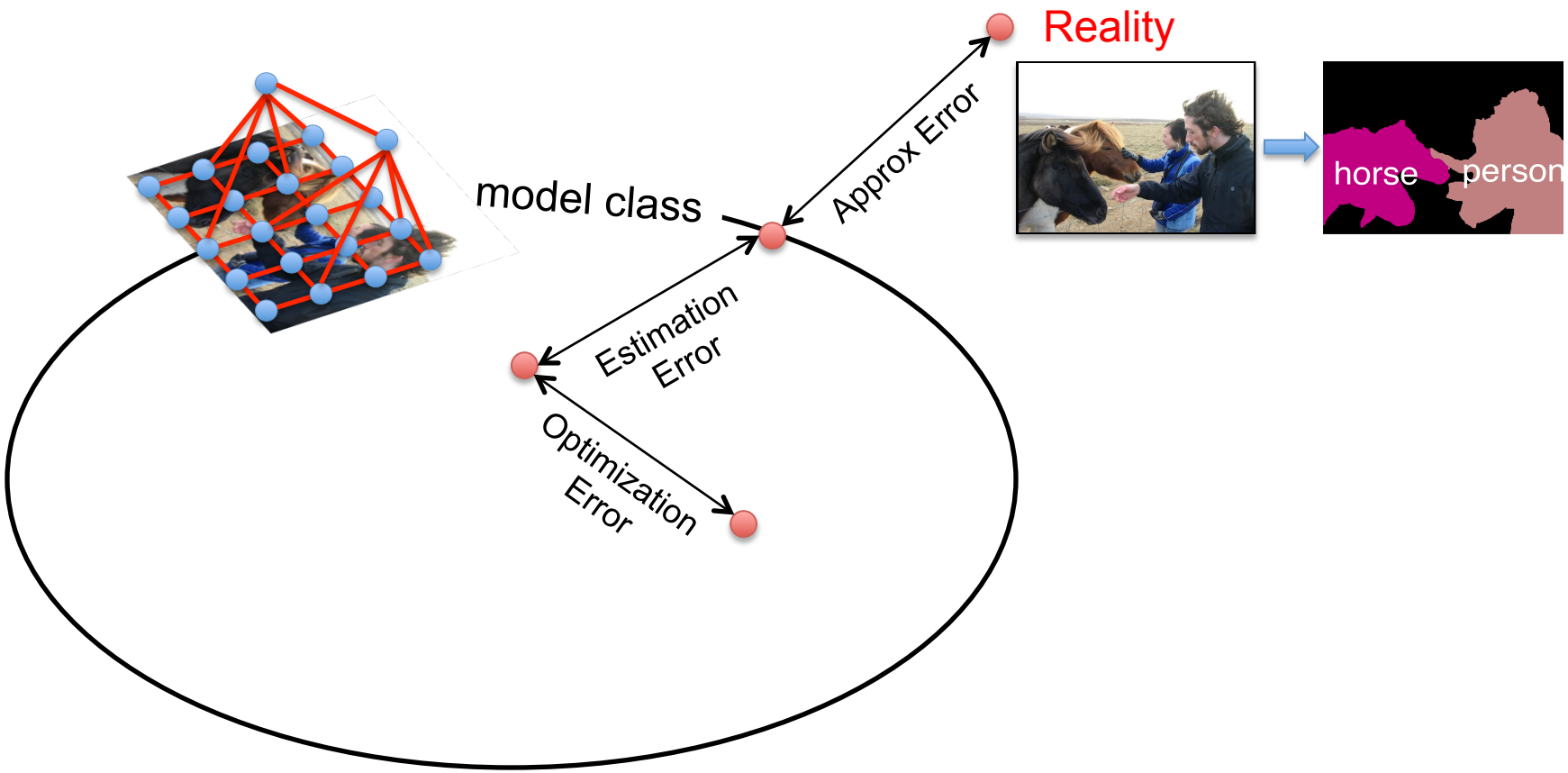
What is this class about?

- Making global predictions from local observations

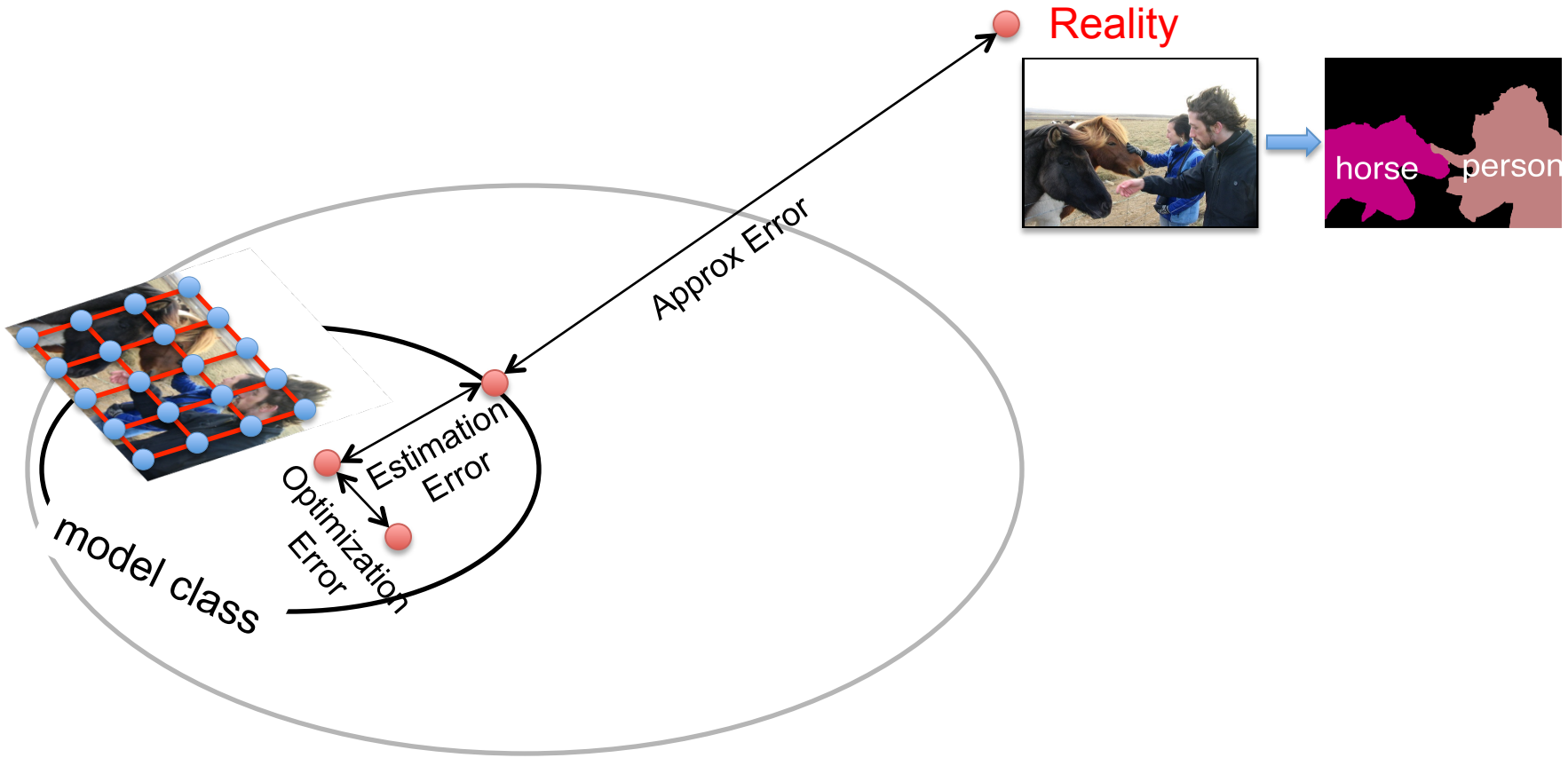
A look forward

- Stuff we couldn't teach you
  - A.K.A: Stuff that's not on the exam!
- What do people in this area work on?
  - What is being published in PGMs / Structured Prediction?

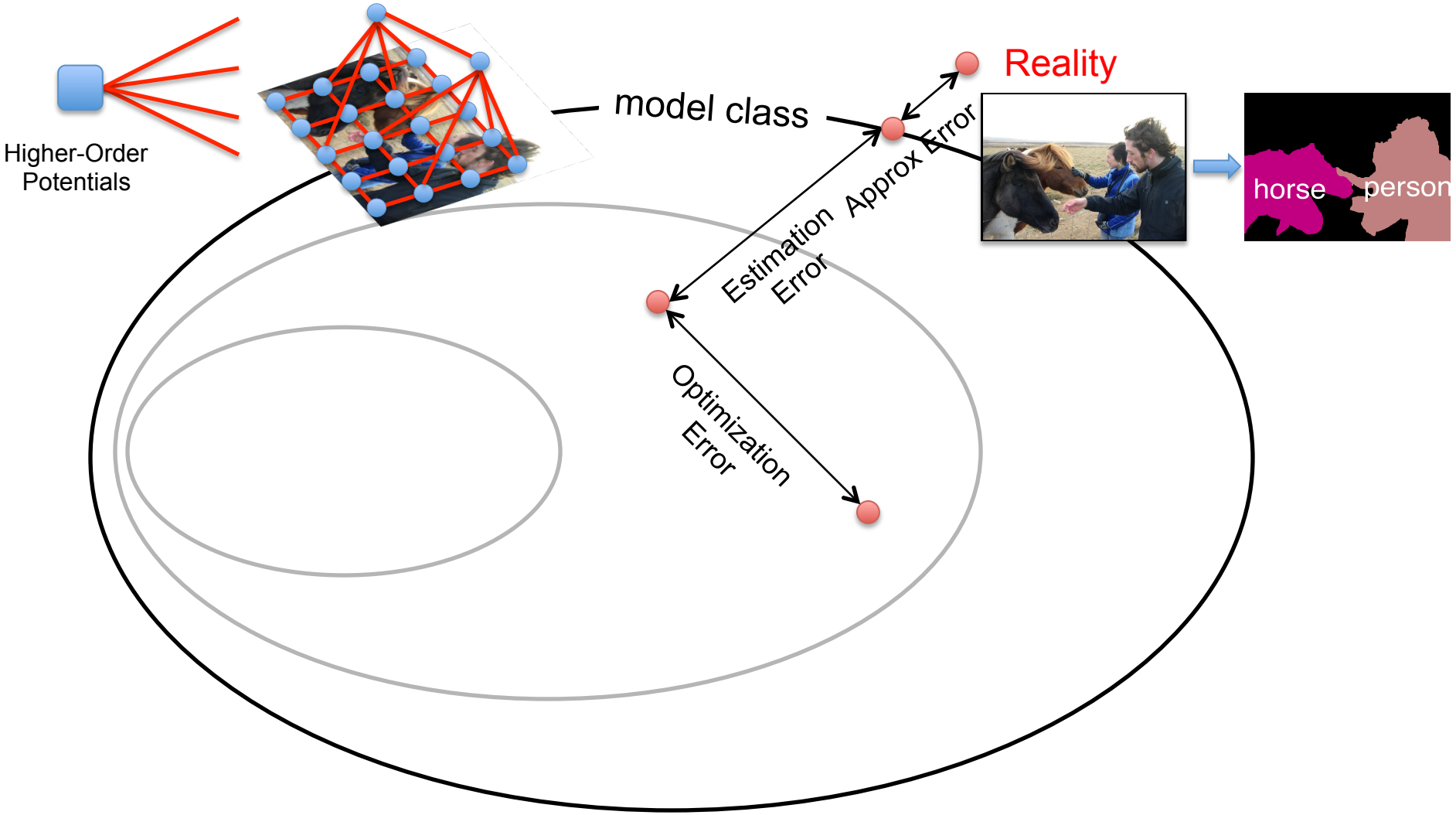
Error Decomposition



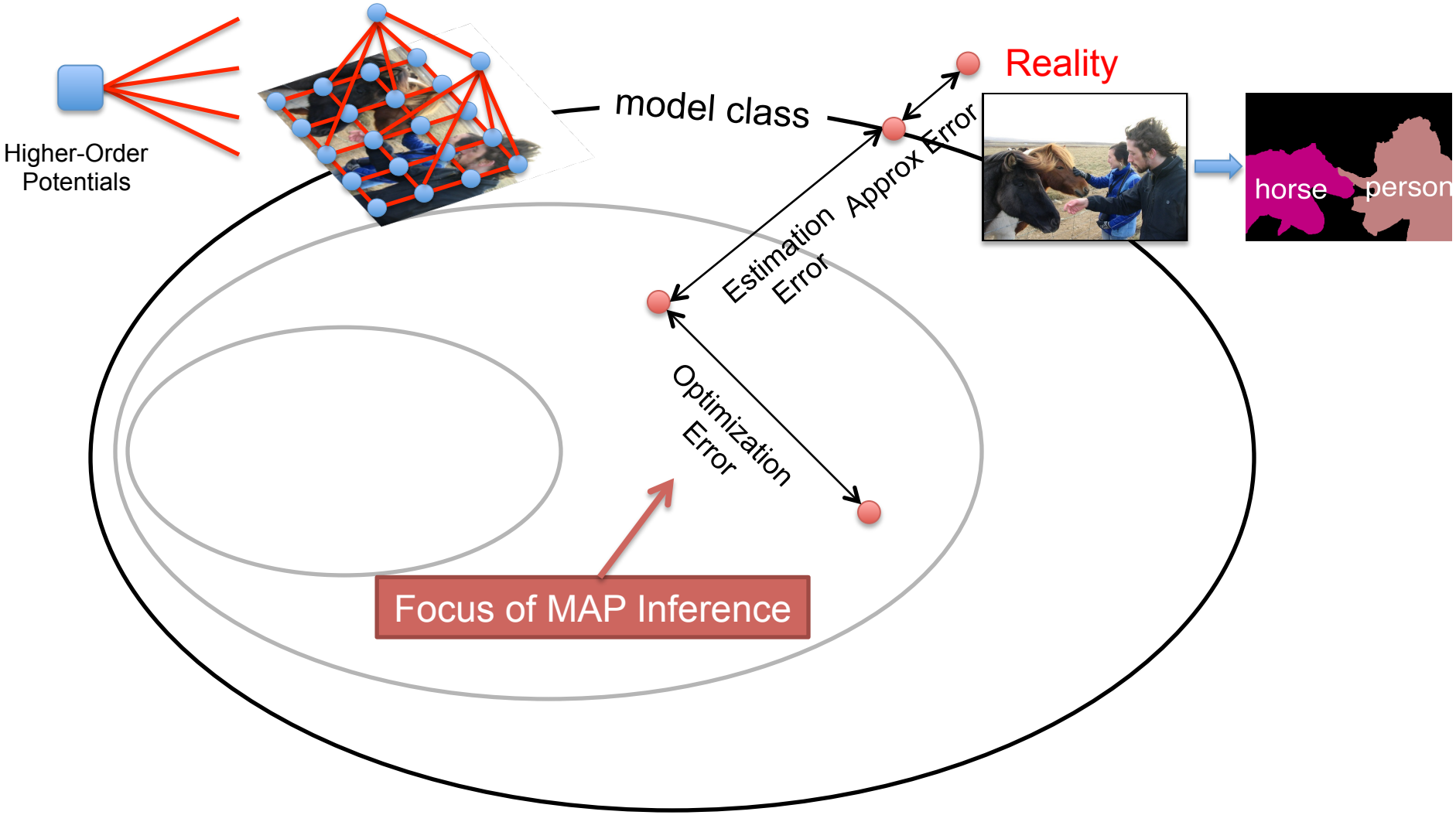
Error Decomposition



Error Decomposition



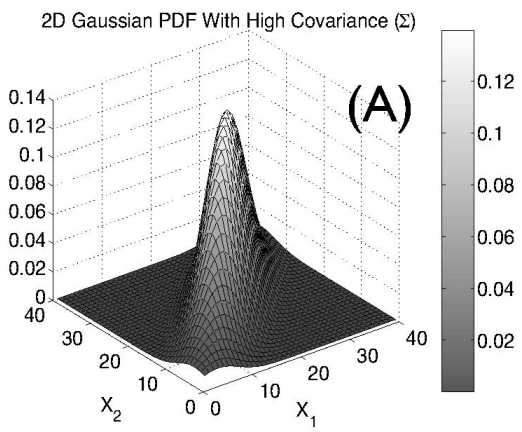
Error Decomposition

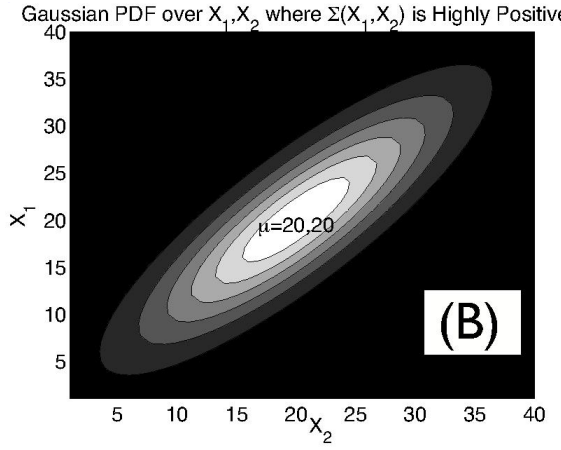


Continuous-Variable PGMs

Multivariate Gaussian

$$p(X_1, \ldots, X_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$





Canonical form

$$
\begin{aligned}
p(X_1, \ldots, X_n) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\} \\
&= K \exp\left\{\eta^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\Lambda\mathbf{x}\right\}
\end{aligned}
$$

- Standard form and canonical forms are related:

$$
\begin{aligned}
\mu &= \Lambda^{-1}\eta \\
\Sigma &= \Lambda^{-1}
\end{aligned}
$$

- Conditioning is easy in canonical form
- Marginalization easy in standard form

Sampling

What you've learned so far

- VE & Junction Trees
  - Exact inference
  - Exponential in tree-width

- Belief Propagation, Mean Field
  - Approximate inference for marginals/conditionals
  - Fast, but can get inaccurate estimates

- Sample-based Inference
  - Approximate inference for marginals/conditionals
  - With "enough" samples, will converge to the right answer (or a high accuracy estimate)

(If you want to be cynical, replace "enough" with "ridiculously many")

Goal

- Often we want expectations given samples x[1] … x[M] from a distribution P.

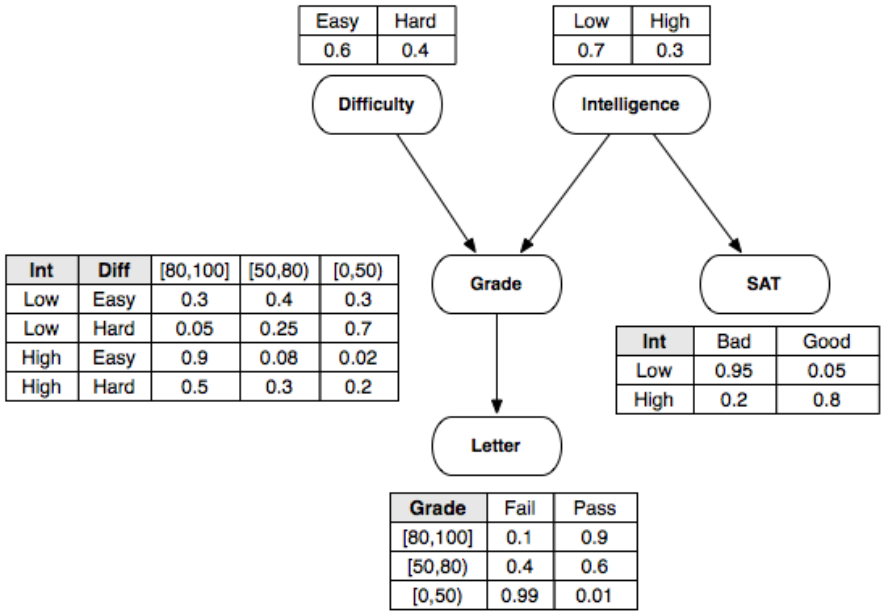
$$E_P[f] \approx \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{x}[m]) \qquad \mathbf{x}[i] \sim P(\mathbf{X})$$

$$P(\mathbf{X} = \mathbf{x}) \approx \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}(\mathbf{x}[m] = \mathbf{x})$$

Discrete Random Variables: $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$

Number of samples from P(X): M

Forward Sampling



Int	Diff	[80,100]	[50,80)	[0,50)
Low	Easy	0.3	0.4	0.3
Low	Hard	0.05	0.25	0.7
High	Easy	0.9	0.08	0.02
High	Hard	0.5	0.3	0.2

Int	Bad	Good
Low	0.95	0.05
High	0.2	0.8

Grade	Fail	Pass
[80,100]	0.1	0.9
[50,80)	0.4	0.6
[0,50)	0.99	0.01

- Sample nodes in topological order
- Assignment to parents selects P(X|Pa(X))
- End result is one sample from P(X)
- Repeat to get more samples

D $\mathbf{x}[m, D] \sim (Easy : 0.6, Hard : 0.4)$ **D = Easy**

I $\mathbf{x}[m, I] \sim (Low : 0.7, High : 0.3)$ **I = High**

G $\mathbf{x}[m, G|D = d, I = i] \sim ([80, 100] : 0.9, [50, 80) : 0.08, [0, 50) : 0.02)$ **G = [80,100]**

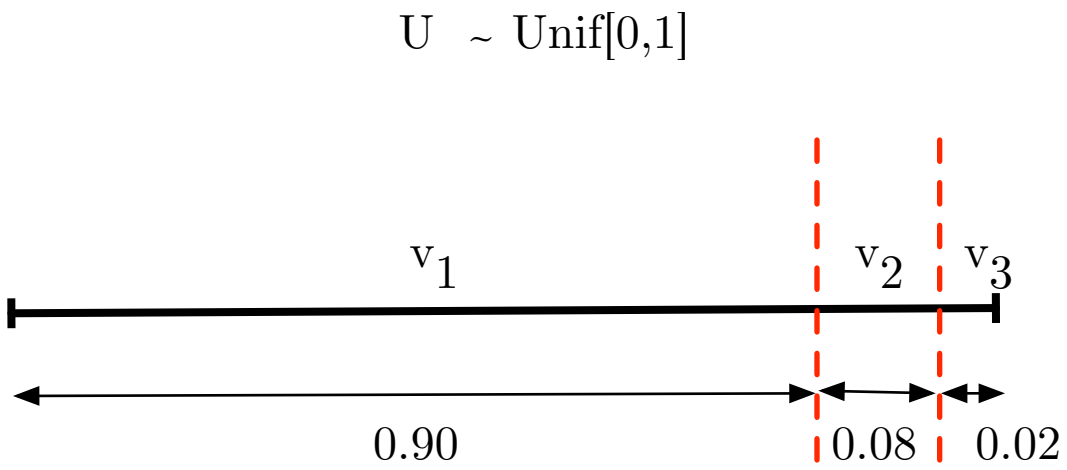
S $\mathbf{x}[m, S|I = i] \sim (Bad : 0.2, Good : 0.8)$ **S = Bad**

L $\mathbf{x}[m, L|G = g] \sim (Fail : 0.1, Pass : 0.9)$ **L = Pass**

Multinomial Sampling

- Given an assignment to its parents, X_i is a multinomial random variable.

$$\mathbf{x}[m, G|D = d, I = i] \sim (v_1 : 0.9, v_2 : 0.08, v_3 : 0.02)$$



Sample-based probability estimates

- Have a set of *M* samples from P(X)
- Can estimate any probability by counting records:

Marginals:

$$\hat{P}(D = \text{Easy}, S = \text{Bad}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}(x[m, D] = \text{Easy}, x[m, S] = \text{Bad})$$

Conditionals:

$$\hat{P}(D = \text{Easy} | S = \text{Bad}) = \frac{\sum_{m=1}^{M} \mathbf{1}(\mathbf{x}[m, D] = \text{Easy}, \mathbf{x}[m, S] = \text{Bad})}{\sum_{m=1}^{M} \mathbf{1}(\mathbf{x}[m, S] = \text{Bad})}$$

Rejection sampling: once the sample and evidence disagree, throw away the sample.

Rare events: If the evidence is unlikely, i.e., P(E = e) small, then the sample size for P(X|E=e) is low

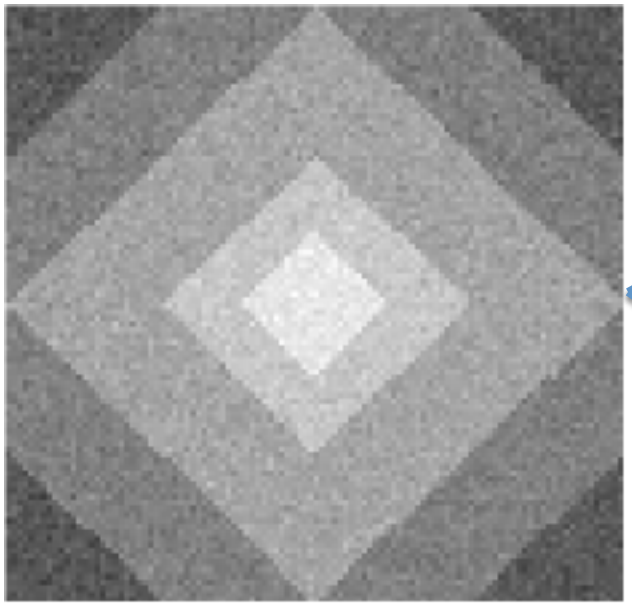
IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 23, NO. 11, NOVEMBER 2001

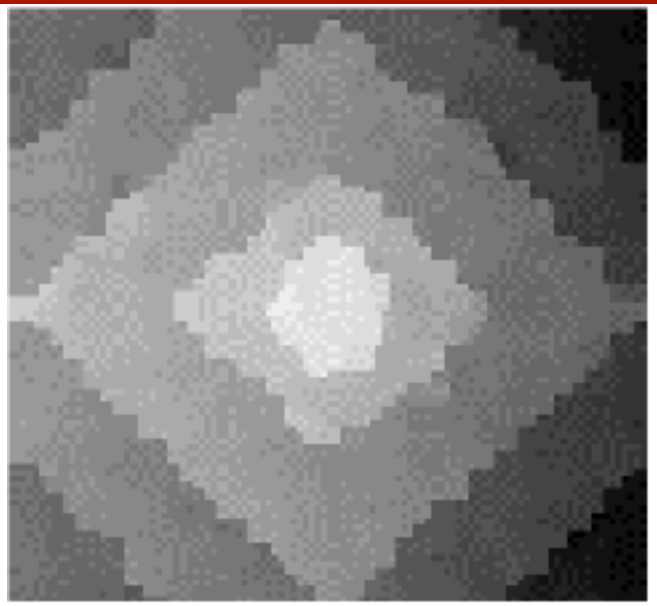
Fast Approximate Energy Minimization via Graph Cuts

Yuri Boykov, *Member*, *IEEE*, Olga Veksler, *Member*, *IEEE*, and Ramin Zabih, *Member*, *IEEE*

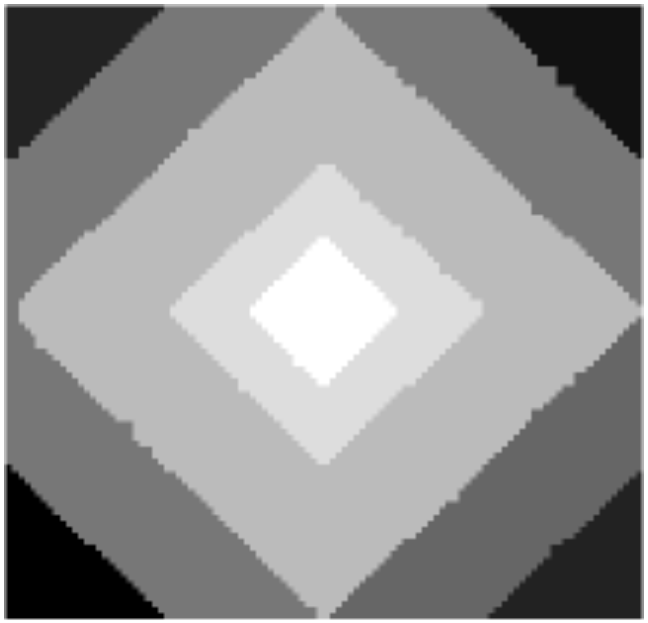
Abstract—Many tasks in computer vision involve assigning a label (such as disparity) to every pixel. A common constraint is that the labels should vary smoothly almost everywhere while preserving sharp discontinuities that may exist, e.g., at object boundaries. These tasks are naturally stated in terms of energy minimization. In this paper, we consider a wide class of energies with various smoothness constraints. Global minimization of these energy functions is NP-hard even in the simplest discontinuity-preserving case. Therefore, our focus is on efficient approximation algorithms. We present two algorithms based on graph cuts that efficiently find a local minimum with respect to two types of large moves, namely *expansion* moves and *swap* moves. These moves can simultaneously change the labels of arbitrarily large sets of pixels. In contrast, many standard algorithms (including simulated annealing) use small moves where only one pixel changes its label at a time. Our expansion algorithm finds a labeling within a known factor of the global minimum, while our swap algorithm handles more general energy functions. Both of these algorithms allow important cases of discontinuity preserving energies. We experimentally demonstrate the effectiveness of our approach for image restoration, stereo and motion. On real data with ground truth, we achieve 98 percent accuracy.

Index Terms—Energy minimization, early vision, graph algorithms, minimum cut, maximum flow, stereo, motion, image restoration, Markov Random Fields, Potts model, multiway cut.



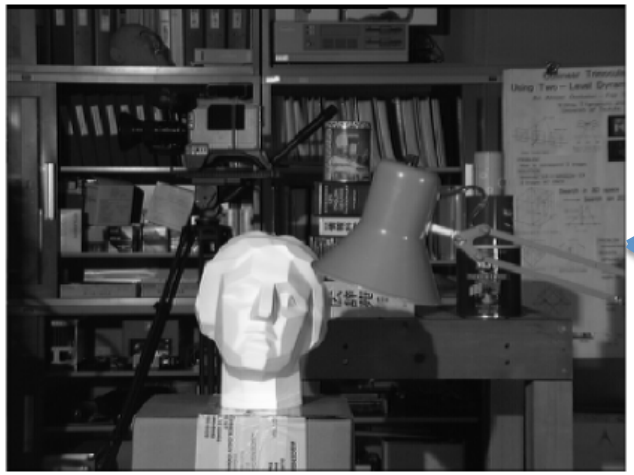


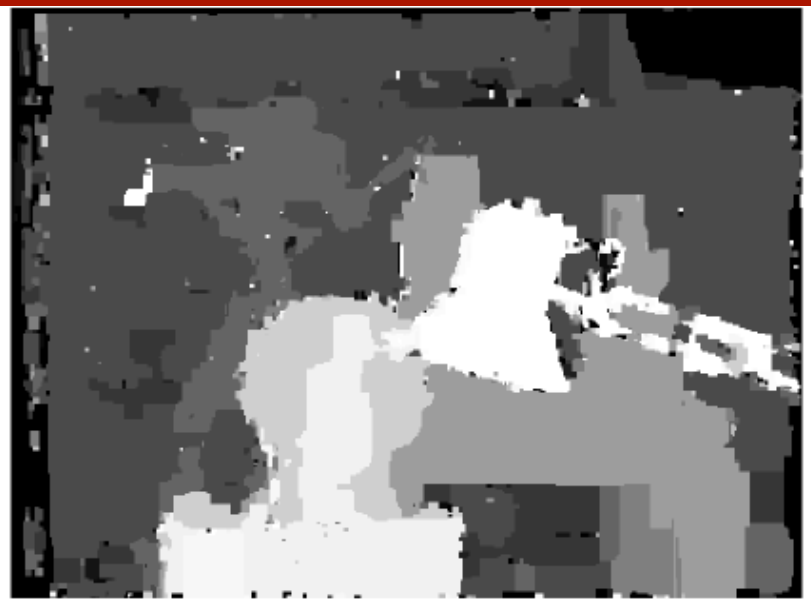
Simulated Annealing



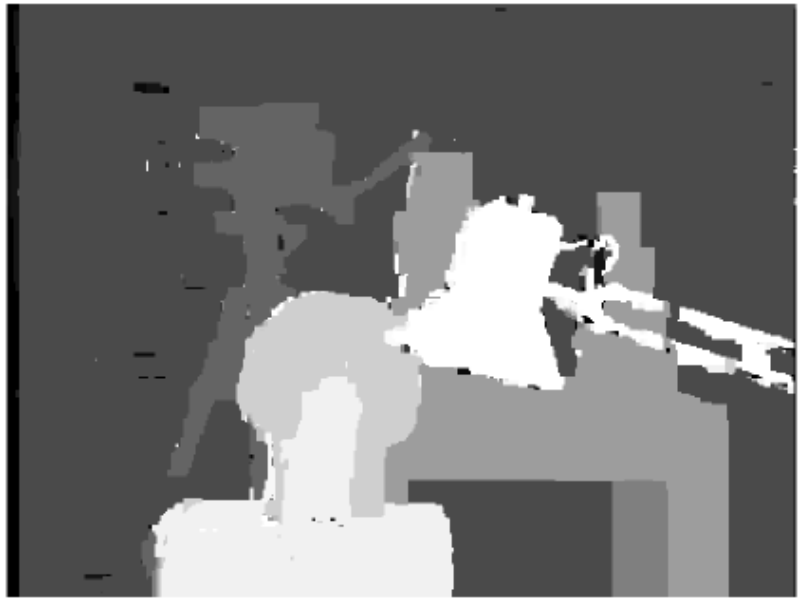
Alpha-Expansion

Image Credit: [BVZ, PAMI01]





Simulated Annealing



Alpha-Expansion

Image Credit: [BVZ, PAMI01]

Effect of Exact MAP

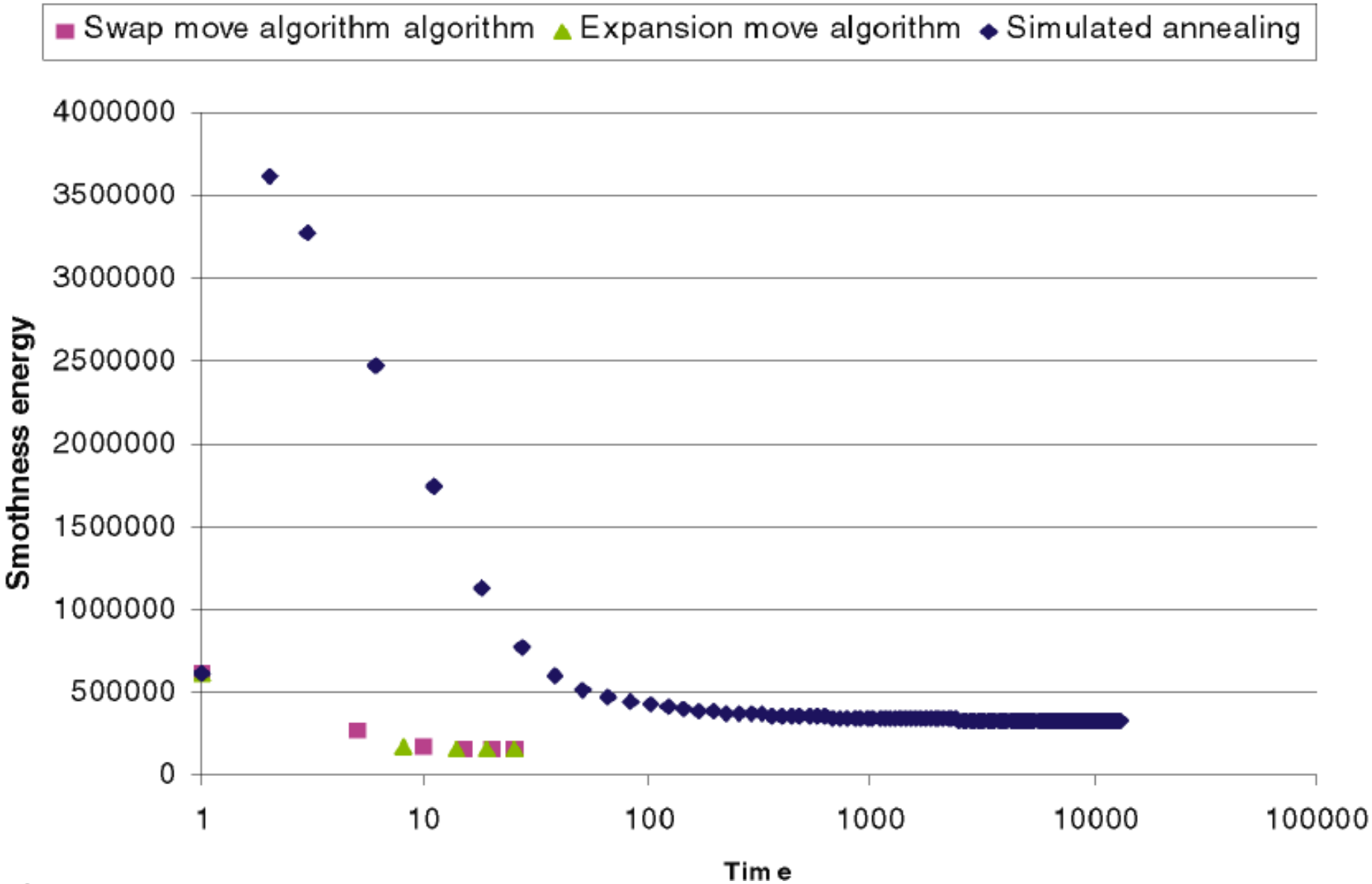


Image Credit: [BVZ, PAMI01]

Sontag NIPS10

More data means less inference: A pseudo-max approach to structured learning

David Sontag
Microsoft Research

Ofer Meshi
Hebrew University

Tommi Jaakkola
CSAIL, MIT

Amir Globerson
Hebrew University

Abstract

The problem of learning to predict structured labels is of key importance in many applications. However, for general graph structure both learning and inference are intractable. Here we show that it is possible to circumvent this difficulty when the distribution of training examples is rich enough, via a method similar in spirit to pseudo-likelihood. We show that our new method achieves consistency, and illustrate empirically that it indeed approaches the performance of exact methods when sufficiently large training sets are used.

Soft-Margin Structured SVM

- Minimize $\frac{1}{2}w^2 + \frac{C}{N}\sum_j \xi_j$

  subject to $w^T\phi(\mathbf{x}^j, \mathbf{y}^j) \geq w^T\phi(\mathbf{x}^j, \mathbf{y}) + \Delta(\mathbf{y}^j, \mathbf{y}) - \xi_j$

  Too many constraints!

Cutting-Plane Method

$$\frac{1}{2}w^2 + \frac{C}{N}\sum_j \xi_j$$

$$w^T \phi(\mathbf{x}^j, \mathbf{y}^j) \geq w^T \phi(\mathbf{x}^j, \mathbf{y}) + \Delta(\mathbf{y}^j, \mathbf{y}) - \xi_j$$

- Key insight of NIPS10 paper
  - What if we replace exponentially many constraints with a smaller set (without using Cutting-Plane)?
- Key contribution
  - There exist a rich set of distributions where this approximation results in optimal parameter in the infinite data setting

Meshi ICML10

Learning Efficiently with Approximate Inference via Dual Losses

Ofer Meshi MESHI@CS.HUJI.AC.IL
David Sontag DSONTAG@CSAIL.MIT.EDU
Tommi Jaakkola TOMMI@CSAIL.MIT.EDU
Amir Globerson GAMIR@CS.HUJI.AC.IL

Abstract

Many structured prediction tasks involve complex models where inference is computationally intractable, but where it can be well approximated using a linear programming relaxation. Previous approaches for learning for structured prediction (e.g., cutting-plane, subgradient methods, perceptron) repeatedly make predictions for some of the data points. These approaches are computationally demanding because each prediction involves solving a linear program to optimality. We present a scalable algorithm for learning for structured prediction. The main idea is to instead solve the dual of the structured prediction loss. We formulate the learning task as a convex minimization over both the weights and the dual variables corresponding to each data point. As a result, we can begin to optimize the weights even before completely solving any of the individual prediction problems. We show how the dual variables can be efficiently optimized using coordinate descent. Our algorithm is competitive with state-of-the-art methods such as stochastic subgradient and cutting-plane.

would be to explicitly model the interactions between the labels, which then results in the labels being jointly predicted. Structured prediction models do this by using classifiers of the form $\boldsymbol{y} = \arg\max_{\hat{\boldsymbol{y}}} \boldsymbol{w} \cdot f(\boldsymbol{x}, \hat{\boldsymbol{y}})$, where $f(\boldsymbol{x}, \boldsymbol{y})$ is a given function and $\boldsymbol{w}$ are weights to be learned from data.

Much of the early work on structured prediction (Lafferty et al., 2001; Taskar et al., 2004) focused on the case where prediction (i.e., maximization over $\boldsymbol{y}$) could be done using efficient combinatorial algorithms such as dynamic programming or maximum-weight matching. However, this restricted the types of interactions that these models were capable of capturing to tractable structures such as tree graphs. Recent work on graphical models has shown that even when the maximization over $\boldsymbol{y}$ is not known a priori to be tractable, linear programming (LP) relaxations often succeed at finding the true maximum, even giving certificates of optimality (Sontag et al., 2008). This strongly motivates learning structured prediction models which use LP relaxations for prediction, and indeed several recent works show that this yields empirically effective results (Finley and Joachims, 2008; Martins et al., 2009).

Learning with large scale data necessitates efficient algorithms for finding the optimal weight vector $\boldsymbol{w}$. Al-

Tarlow UAI10

Graph Cuts is a Max-Product Algorithm

Daniel Tarlow, Inmar E. Givoni, Richard S. Zemel, Brendan J. Frey
University of Toronto
Toronto, ON M5S 3G4
{dtarlow@cs, inmar@psi, zemel@cs, frey@psi}.toronto.edu

Abstract

The maximum a posteriori (MAP) configuration of binary variable models with submodular graph-structured energy functions can be found efficiently and exactly by graph cuts. Max-product belief propagation (MP) has been shown to be suboptimal on this class of energy functions by a canonical counterexample where MP converges to a suboptimal fixed point (Kulesza & Pereira, 2008).

In this work, we show that under a particular scheduling and damping scheme, MP is equivalent to graph cuts, and thus optimal. We explain the apparent contradiction by showing that with proper scheduling and damping, MP always converges to an optimal fixed point. Thus, the canonical counterexample only shows the suboptimality of MP with a particular suboptimal choice of schedule and damping. With proper choices, MP is optimal.

eral, but also occasionally erratic, algorithm is max-product belief propagation (MP).

Our aim in this work is to establish the precise relationship between MP and graph cuts, namely that graph cuts is a special case of MP. To do so, we map analogous aspects of the algorithms to each other: message scheduling in MP to selecting augmenting paths in graph cuts; passing messages on a chain to pushing flow through an augmenting path; message damping to limiting flow to be the bottleneck capacity of an augmenting path; and letting messages reinforce themselves on a loopy graph to connected components decoding scheme of graph cuts.

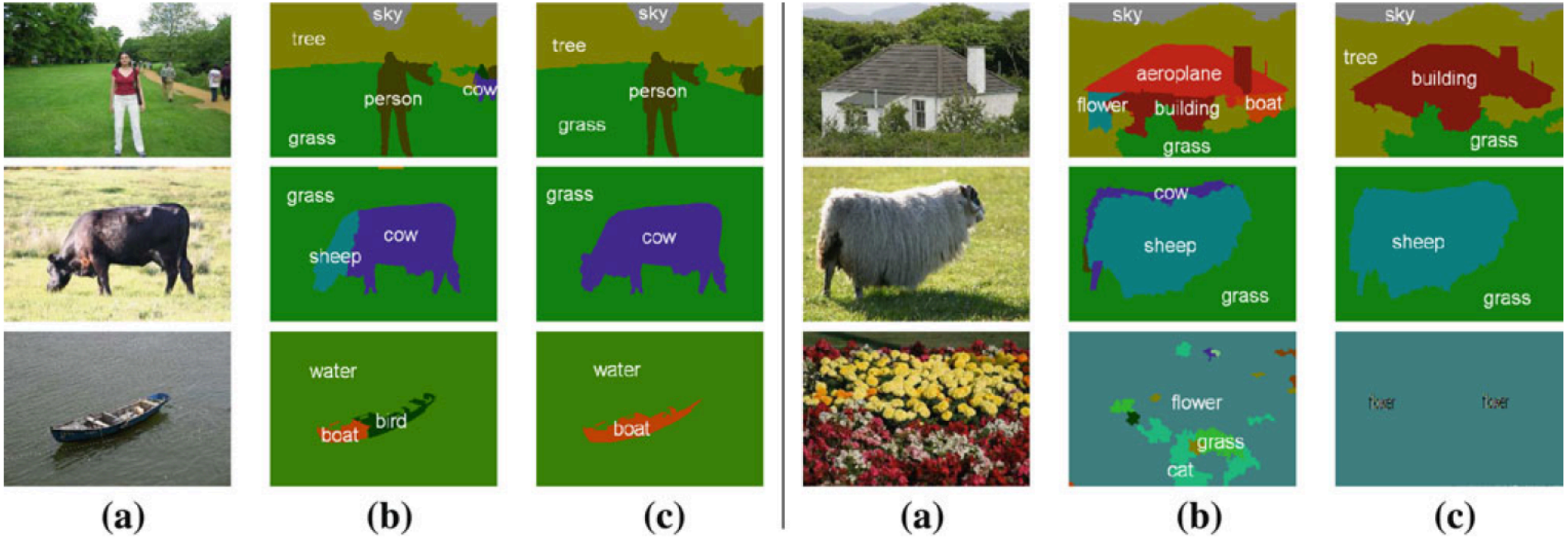
This equivalence implies strong statements regarding the optimality of MP on binary submodular energies defined on graphs with arbitrary topology, which may appear to contradict much of what is known about MP—all empirical results showing MP to be suboptimal on binary submodular problems, and the theoretical results of Kulesza and Pereira (2008); Wainwright and Jordan (2008) which show analytically that MP converges to the wrong solution. We analyze this is-

Ladicky IJCV12

Int J Comput Vis (2013) 103:213–225
DOI 10.1007/s11263-012-0583-y

Inference Methods for CRFs with Co-occurrence Statistics

L'ubor Ladický · Chris Russell · Pushmeet Kohli · Philip H. S. Torr



Lempitsky ICCV09

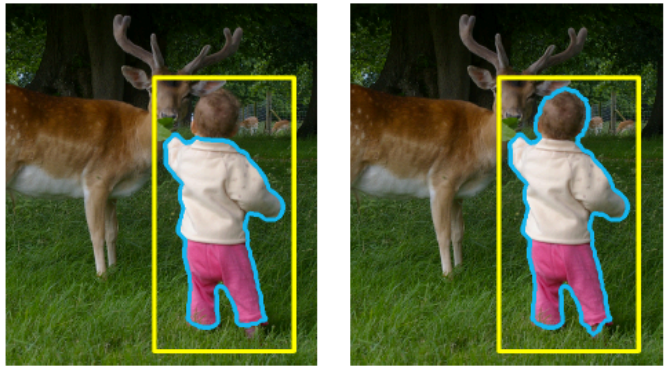
Image Segmentation with A Bounding Box Prior

Victor Lempitsky, Pushmeet Kohli, Carsten Rother, Toby Sharp
Microsoft Research Cambridge

Abstract

User-provided object bounding box is a simple and popular interaction paradigm considered by many existing interactive image segmentation frameworks. However, these frameworks tend to exploit the provided bounding box merely to exclude its exterior from consideration and sometimes to initialize the energy minimization. In this paper, we discuss how the bounding box can be further used to impose a powerful topological prior, which prevents the solution from excessive shrinking and ensures that the user-provided box bounds the segmentation in a sufficiently tight way.

The prior is expressed using hard constraints incorporated into the global energy minimization framework leading to an NP-hard integer program. We then investigate the possible optimization strategies including linear relaxation as well as a new graph cut algorithm called pinpointing. The latter can be used either as a rounding method for the fractional LP solution, which is provably better than thresholding-based rounding, or as a fast standalone heuristic. We evaluate the proposed algorithms on a publicly available dataset, and demonstrate the practical benefits of the new prior both qualitatively and quantitatively.



without the prior with the prior

Figure 1. **Our tightness prior.** The segmentation on the left computed with graph cut is consistent with the low level image cues, yet inconsistent with the user input (in yellow) being too loose for this bounding box. By minimizing the same graph cut energy under a set of constraints, our method computes the segmentation that fits the bounding box in a sufficiently tight way, obtaining a better result (right).

to the interior of the bounding box. This property is easy to incorporate into any algorithm, as one can simply assign all exterior pixels to the 'background' class. The second property is much harder to incorporate or even to formalize.

Kohli & Rother

Chapter 1

Higher-order models in Computer Vision

PUSHMEET KOHLI

Machine Learning and Perception
Microsoft Research
Cambridge, UK
Email: pkohli@microsoft.com

CARSTEN ROTHER

Machine Learning and Perception
Microsoft Research
Cambridge, UK
Email: carrot@microsoft.com

Perturb and MAP

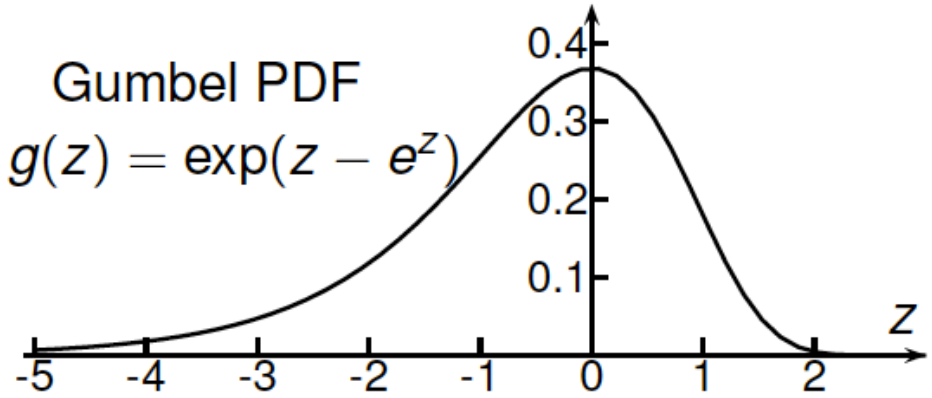
$$S(\mathbf{y}) = \sum_{i \in \mathcal{V}} \theta_i(y_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(y_i, y_j)$$

- Approach
  - Perturb: $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + \epsilon, \quad \epsilon \sim p(\epsilon)$
  - MAP: $\underset{\mathbf{y}}{\operatorname{argmax}} \; S_{\tilde{\theta}}(\mathbf{y})$

Perturb and MAP

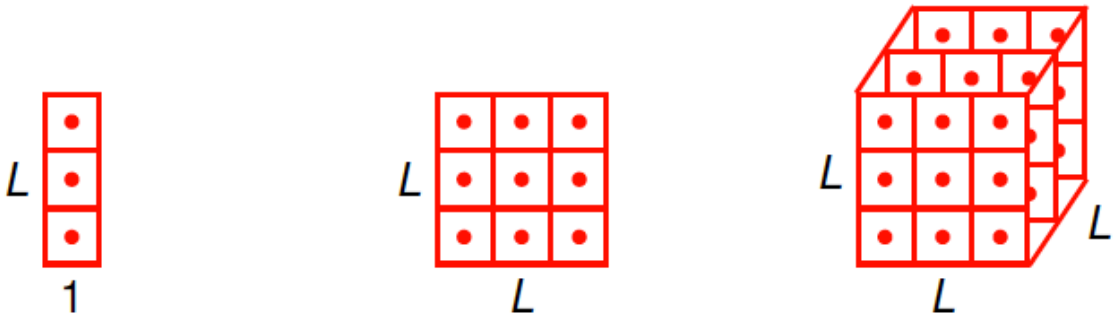
- Perturb: $\tilde{\theta} = \theta + \epsilon, \quad \epsilon \sim p(\epsilon)$

- Theorem: If IID Gumbel, then EXACT samples.
  - [Papandreou & Yuille, ICCV11]
  - [Hazan & Jaakkola, ICML12]



Perturb and MAP

- Full Order Gumbel is hard!



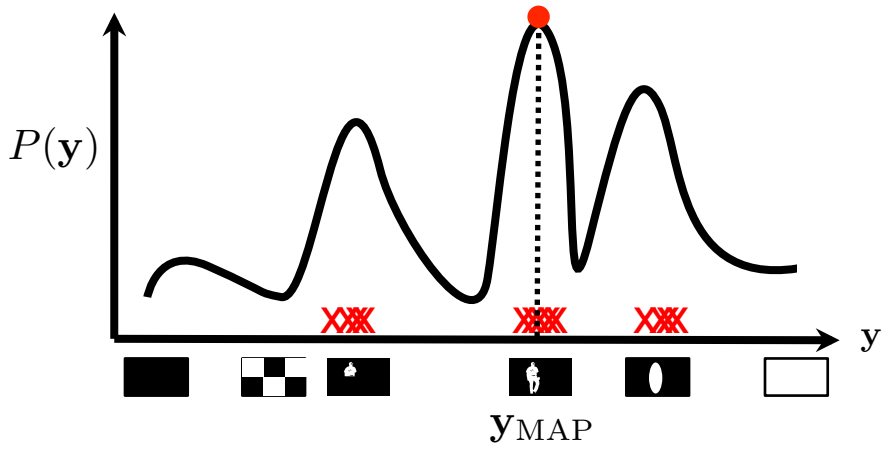
Perturb and MAP

- Reduced Order Gumbel

$$S(\mathbf{y}) = \sum_{i \in \mathcal{V}} \theta_i(y_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(y_i, y_j)$$

- Approach
  - Perturb: $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + \epsilon, \quad \epsilon \sim p(\epsilon)$ $\qquad \tilde{\theta}_i = \theta_i + \epsilon, \quad \epsilon \sim p(\epsilon)$
  - MAP: $\underset{\mathbf{y}}{\operatorname{argmax}} \; S_{\tilde{\theta}}(\mathbf{y})$

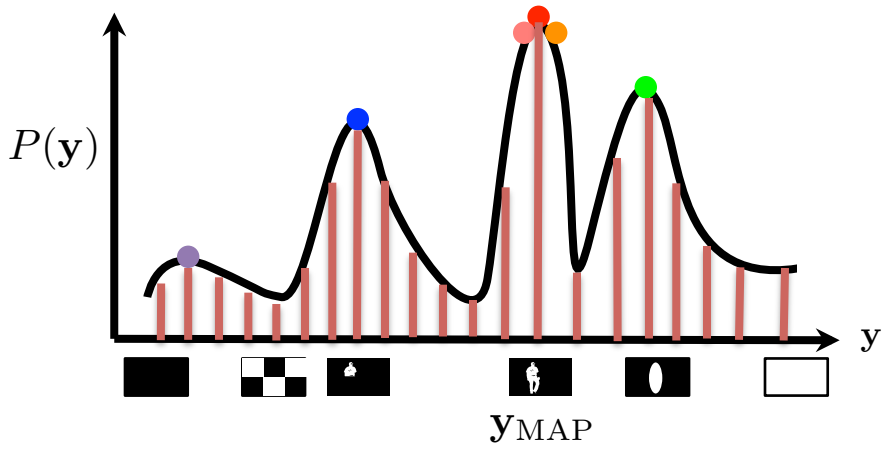
My Research: Multiple Predictions



Sampling

Porway & Zhu, 2011
TU & Zhu, 2002
Rich History

My Research: Multiple Predictions



Sampling

Porway & Zhu, 2011
TU & Zhu, 2002
Rich History

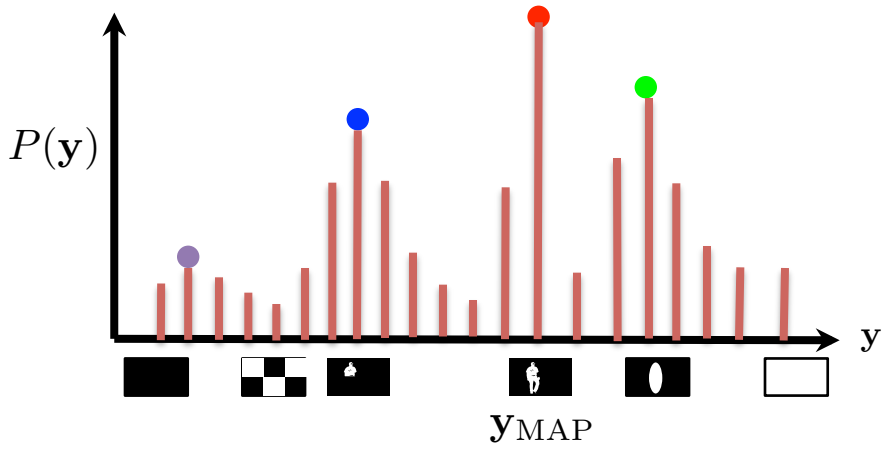
M-Best MAP

Flerova et al., 2011
Fromer et al., 2009
Yanover et al., 2003

Ideally:

M-Best Modes

My Research: Multiple Predictions



Our work: Diverse M-Best in MRFs [ECCV '12]

- Don't *hope* for diversity. Explicitly encode it.
- Not guaranteed to be modes.

What next?

- Seminars:
  - CV-ML Reading Group
    - https://filebox.ece.vt.edu/~cvmlreadinggroup/
- Conferences:
  - Neural Information Processing Systems (NIPS)
  - International Conference in Machine Learning (ICML)
  - Uncertainty in Artificial Intelligence (UAI)
  - Artificial Intelligence & Statistics (AISTATS)
- Classes (at some point in the near future)
  - ECE6504: Fundamental Ideas in Machine Learning
    - Paper reading class showing lineage of ideas
    - Story from ’89 first backprop papers to CNNs today
  - ECE6504: Machine Learning for Big Data
    - Large-Scale Distributed Machine Learning
    - Use frameworks such as Graphlab
    - Implement things in CloudCV

Feedback

- Student Perception of Teaching (SPOT)
  - https://eval.scholar.vt.edu/portal
  - Tell us how we're doing
  - What would you like to see more
  - What would you like to see less
  - ENDS MAY 8