Paired-Dual Learning for Fast Training of Latent Variable Hinge-Loss MRFs

Stephen H. Bach* Maryland

Bert Huang* Virginia Tech

Jordan Boyd-Graber Colorado

Lise Getoor UC Santa Cruz

* Equal Contributors

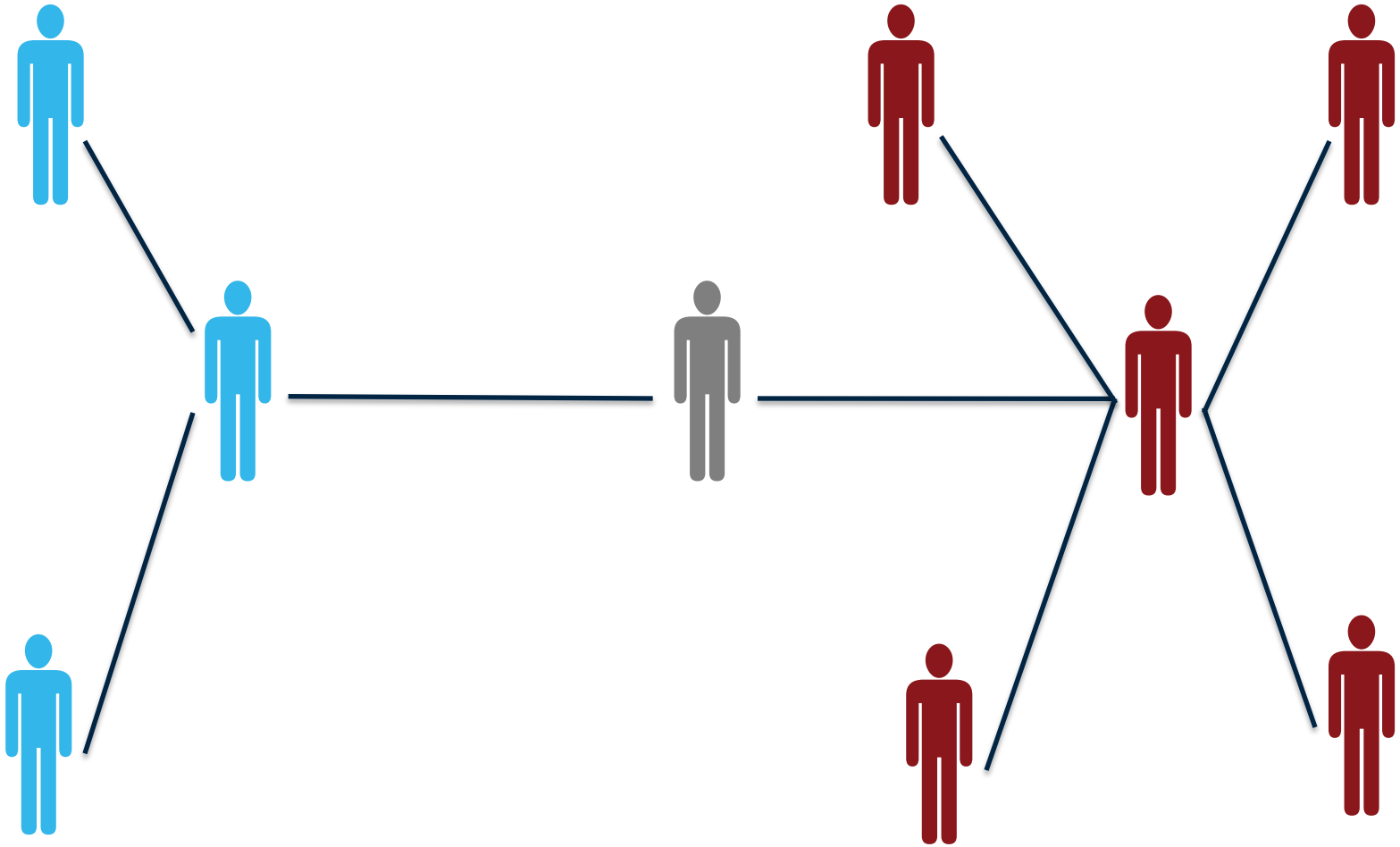
ICML 2015

This Talk

- In rich, structured domains, **latent variables** can capture fundamental aspects and increase accuracy
- Learning with latent variables needs **repeated inferences**
- Recent work has **overcome the inference bottleneck** in discrete models, but using **continuous variables** introduces **new challenges**
- We introduce **paired-dual learning** (PDL)
- PDL is **so fast** that is often finishes before traditional methods make a single parameter update

Latent Variable Models

Community Detection



Latent User Attributes

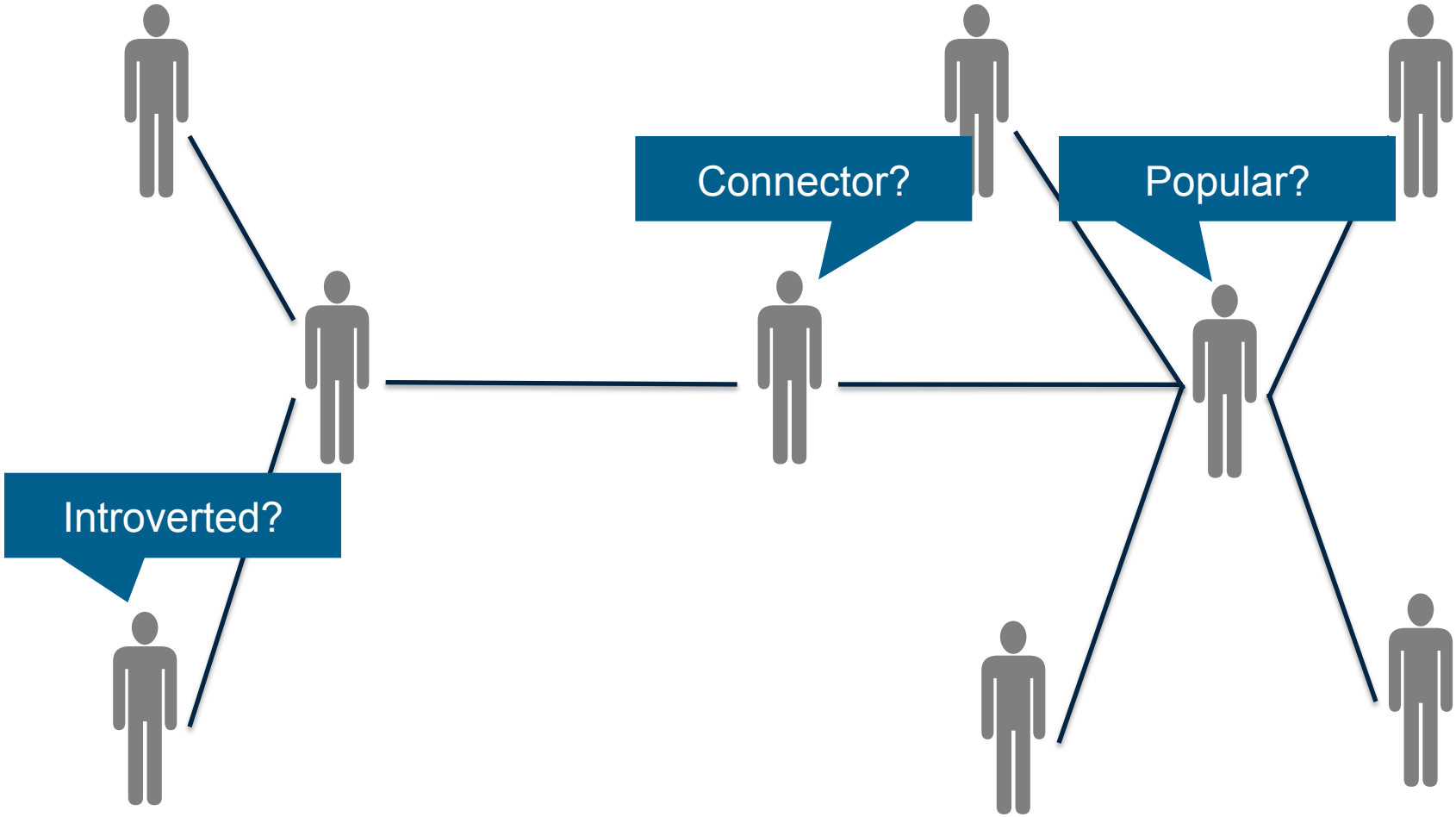


Image Reconstruction

- Latent variables can represent archetypical components

Originals With LVs Without

- Learned components for face reconstruction:

Learning with Latent Variables

Model

- Observations $\boldsymbol{x}$
- Targets $\boldsymbol{y}$ with ground-truth labels $\hat{\boldsymbol{y}}$
- Latent (unlabeled) $\boldsymbol{z}$
- Parameters $\boldsymbol{w}$

$$P(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{x}; \boldsymbol{w}) = \frac{1}{Z(\boldsymbol{x}; \boldsymbol{w})} \exp\left(-\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\right)$$

$$Z(\boldsymbol{x}; \boldsymbol{w}) = \sum_{\boldsymbol{y}, \boldsymbol{z}} \exp\left(-\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\right)$$

Learning Objective

$$\begin{aligned}\log P(\hat{\boldsymbol{y}}|\boldsymbol{x};\boldsymbol{w}) &= \log Z(\boldsymbol{x},\hat{\boldsymbol{y}};\boldsymbol{w}) - \log Z(\boldsymbol{x};\boldsymbol{w}) \\ &= \min_{\rho\in\Delta(\boldsymbol{y},\boldsymbol{z})} \max_{q\in\Delta(\boldsymbol{z})} \mathbb{E}_\rho\left[\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\right] - H(\rho) \\ &\quad - \mathbb{E}_q\left[\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x},\hat{\boldsymbol{y}},\boldsymbol{z})\right] + H(q)\end{aligned}$$

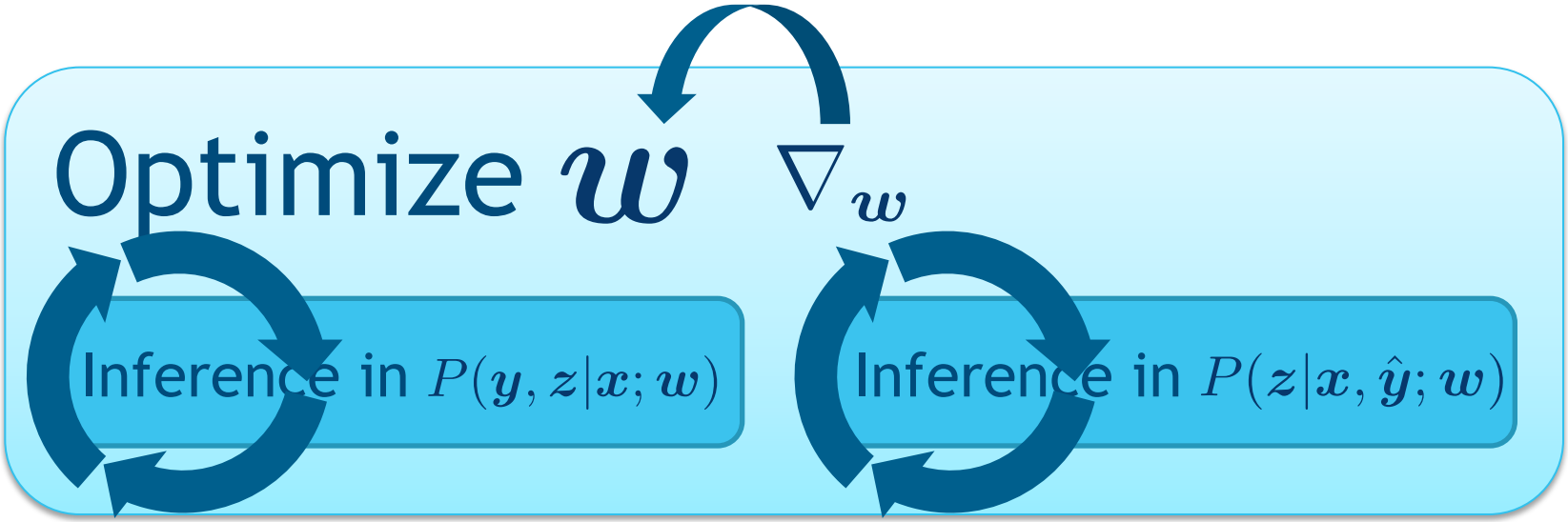
Optimize $\boldsymbol{w}$

Inference in $P(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x};\boldsymbol{w})$

Inference in $P(\boldsymbol{z}|\boldsymbol{x},\hat{\boldsymbol{y}};\boldsymbol{w})$

Traditional Method

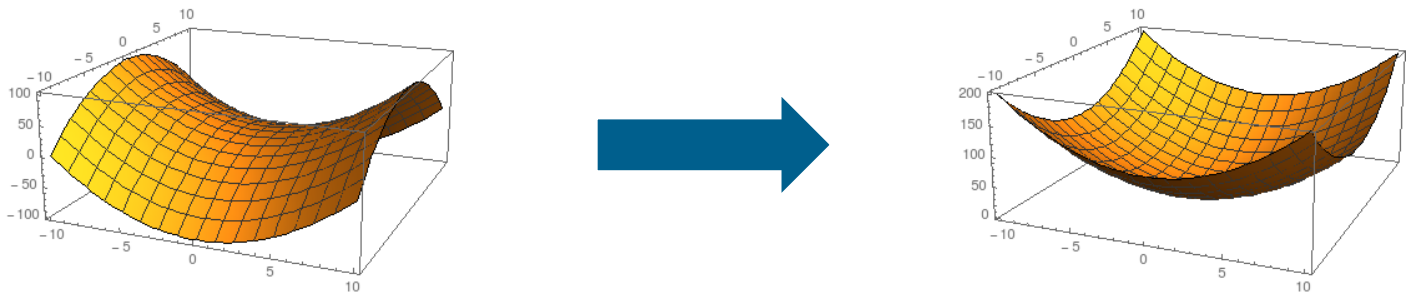
- Perform full inference in each distribution
- Compute the gradient with respect to $\boldsymbol{w}$
- Update $\boldsymbol{w}$ using the gradient



How can we solve the inference bottleneck?

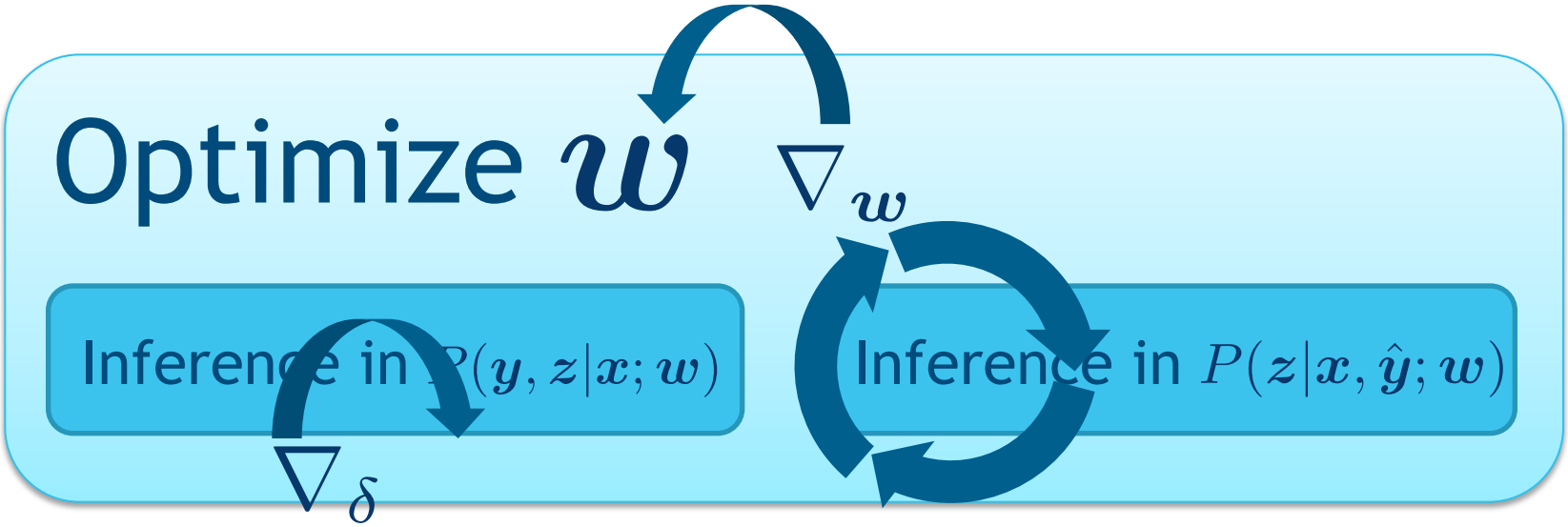
Smart Supervised Learning

- Supervised learning objective contains an inner inference
- Interleave inference and learning
  - e.g., Taskar et al. [ICML 2005], Meshi et al. [ICML 2010], Hazan and Urtasun [NIPS 2010]
- Idea: turn saddle-point optimization into joint minimization by **dualizing inner inference problem:**



Smart Latent Variable Learning

- For **discrete models**, Schwing et al. [ICML 2012] proposed dualizing one of the inferences and interleaving with parameter updates



How can we solve the inference bottleneck for continuous models?

Continuous Structured Prediction

- The learning objective contains expectations and entropy functions that are intractable for continuous distributions

- Recently, there's been a lot of work on developing
  - **continuous probabilistic graphical models**
  - **continuous probabilistic programming languages**

Hinge-Loss Markov Random Fields

- **Natural language processing**
  - Beltagy et al. [ACL 2014], Foulds et al. [ICML 2015]
- **Social network analysis**
  - Huang et al. [SBP 2013], West et al. [TACL 2014], Li et al. [2014]
- **Massive open online course (MOOC) analysis**
  - Ramesh et al. [AAAI 2014, ACL 2015]
- **Bioinformatics**
  - Fakhraei et al. [TCBB 2014]

Hinge-Loss Markov Random Fields

- MRFs over **continuous variables** in [0,1] and hinge-loss potential functions

$$P(\boldsymbol{y}) \propto \exp\left(-\sum_{j=1}^{m} w_j \left(\max\left\{\ell_j(\boldsymbol{y}), 0\right\}\right)^{p_j}\right)$$

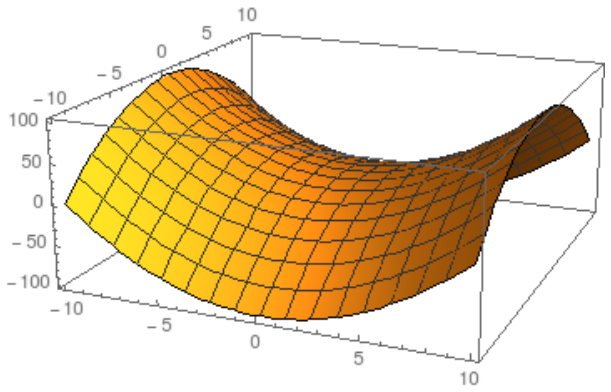
where ℓ_j is a linear function and $p_j \in \{1, 2\}$

MAP Inference in HL-MRFs

- Exact MAP inference in HL-MRFs is very fast, thanks to the alternating direction method of multipliers (ADMM)

- ADMM decomposes inference by
  - Forming augmented Lagrangian
  - Iteratively updating blocks of variables

$$L_w(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\alpha}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{z}})$$



Paired-Dual Learning

Continuous Latent Variables

- The objective is the same, but the expectations and entropies are intractable

$$\arg\min_{\boldsymbol{w}} \max_{\rho \in \Delta(\boldsymbol{y},\boldsymbol{z})} \min_{q \in \Delta(\boldsymbol{z})} \frac{\lambda}{2}\|\boldsymbol{w}\|^2 - \mathbb{E}_\rho\left[\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})\right] + H(\rho) + \mathbb{E}_q\left[\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x},\hat{\boldsymbol{y}},\boldsymbol{z})\right] - H(q)$$

Variational Approximations

- We can restrict the distribution families to single points
  - In other words, we can approximate expectations with MAP
  - Great for models with fast, convex inference, like HL-MRFs

- But, the entropy of a point distribution is always zero

$$\arg\min_{\boldsymbol{w}} \max_{\boldsymbol{y},\boldsymbol{z}} \min_{\boldsymbol{z}'} \frac{\lambda}{2}\|\boldsymbol{w}\|^2 - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) + \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}, \hat{\boldsymbol{y}}, \boldsymbol{z}')$$

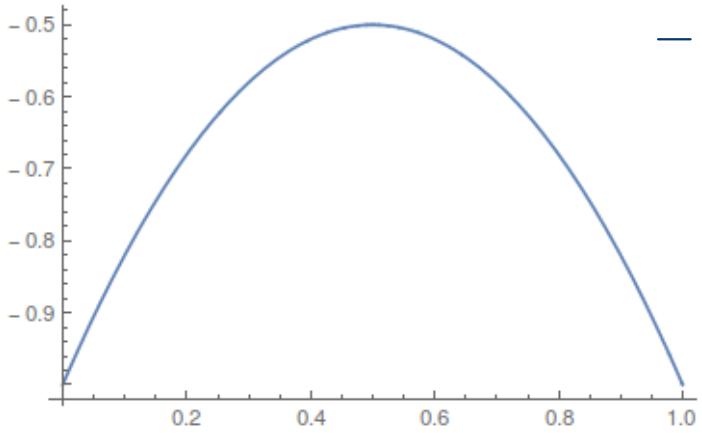
- Therefore, $\boldsymbol{w} = \boldsymbol{0}$ is always a global optimum

Entropy Surrogates

- We design surrogates to fill the role of entropy terms
  - They need to be tractable
  - Choice should be tailored to problem and model
  - Options include curvature and one-sided vs. two-sided

- Goal: require non-zero parameters to predict ground truth

- Example:



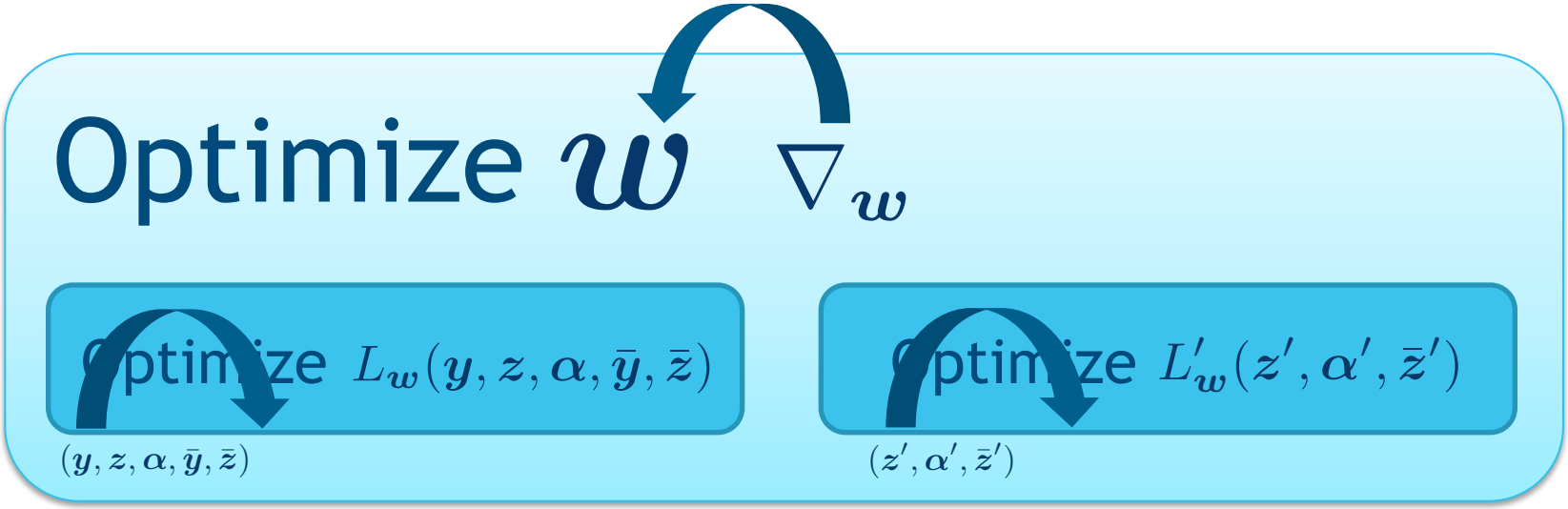
$$-\max\{y,0\}^2 - \max\{1-y,0\}^2$$

Paired-Dual Learning

$$\arg\min_{\boldsymbol{w}} \max_{\boldsymbol{y},\boldsymbol{z}} \min_{\boldsymbol{z}'} \frac{\lambda}{2}\|\boldsymbol{w}\|^2 - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) + h(\boldsymbol{y},\boldsymbol{z}) + \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x},\hat{\boldsymbol{y}},\boldsymbol{z}') - h(\hat{\boldsymbol{y}},\boldsymbol{z}')$$

- **Repeatedly solving** the inner inference problems with ADMM still becomes **expensive**
- But we can replace the inference problems with their augmented Lagrangians

Paired-Dual Learning



- If the inner maxes and mins were solved to convergence this objective would be equivalent

- Instead, paired-dual learning **iteratively updates** the parameters and blocks of Lagrangian variables

Evaluation

Evaluation

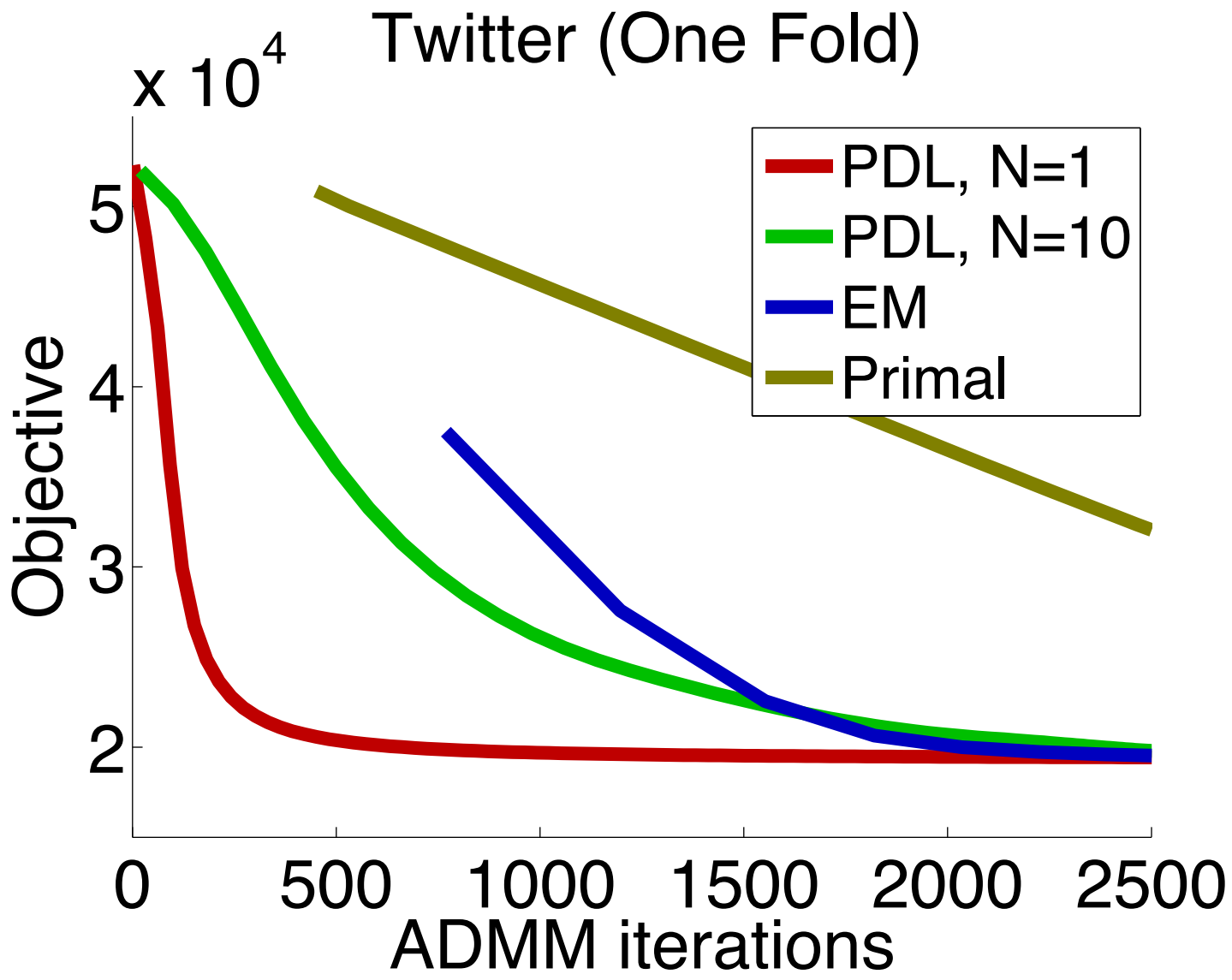
- Three real-world problems:
  - Community detection
  - Latent user attributes
  - Image reconstruction
- Learning methods:
  - Paired-dual learning (PDL) (N=1, N=10)
  - Expectation maximization (EM)
  - Primal gradient descent (Primal)
- Evaluated:
  - Learning objective
  - Predictive performance
  - Vs. ADMM (inference) iterations

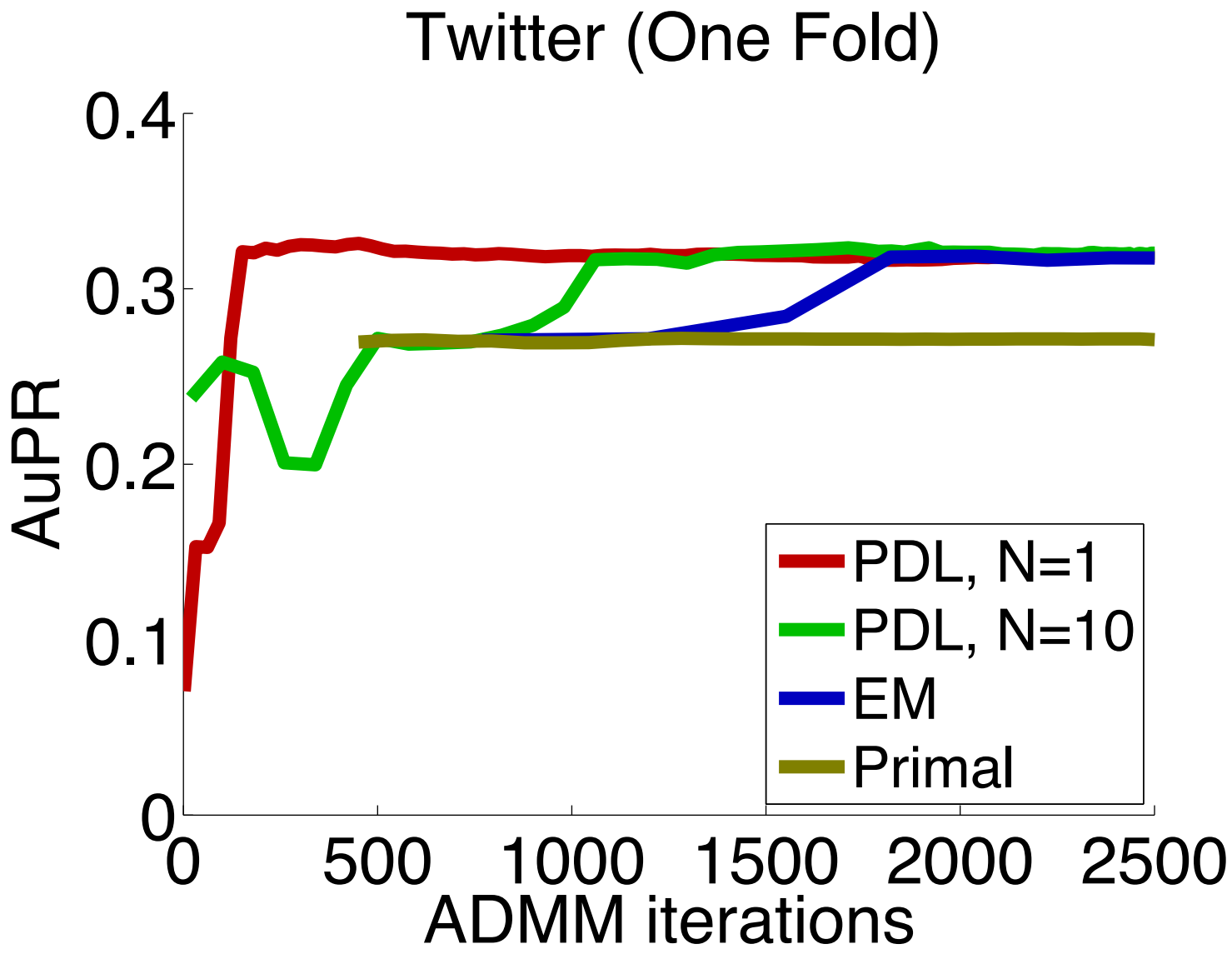
Community Detection

- Case Study: 2012 Venezuelan Presidential Election
  - Incumbent: Hugo Chávez
  - Challenger: Henrique Capriles

Chávez

Capriles

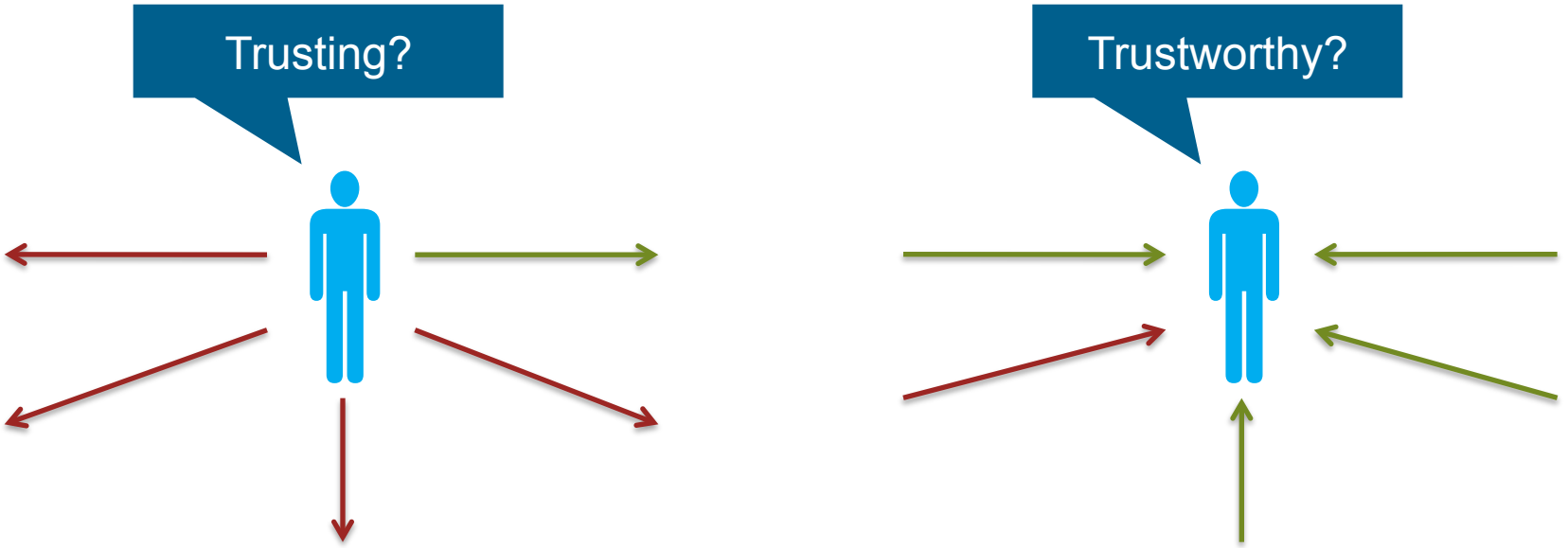

Twitter (One Fold)
x 10^4
Objective
5
4
3
2
0
500
1000
1500
2000
2500
ADMM iterations
PDL, N=1
PDL, N=10
EM
Primal

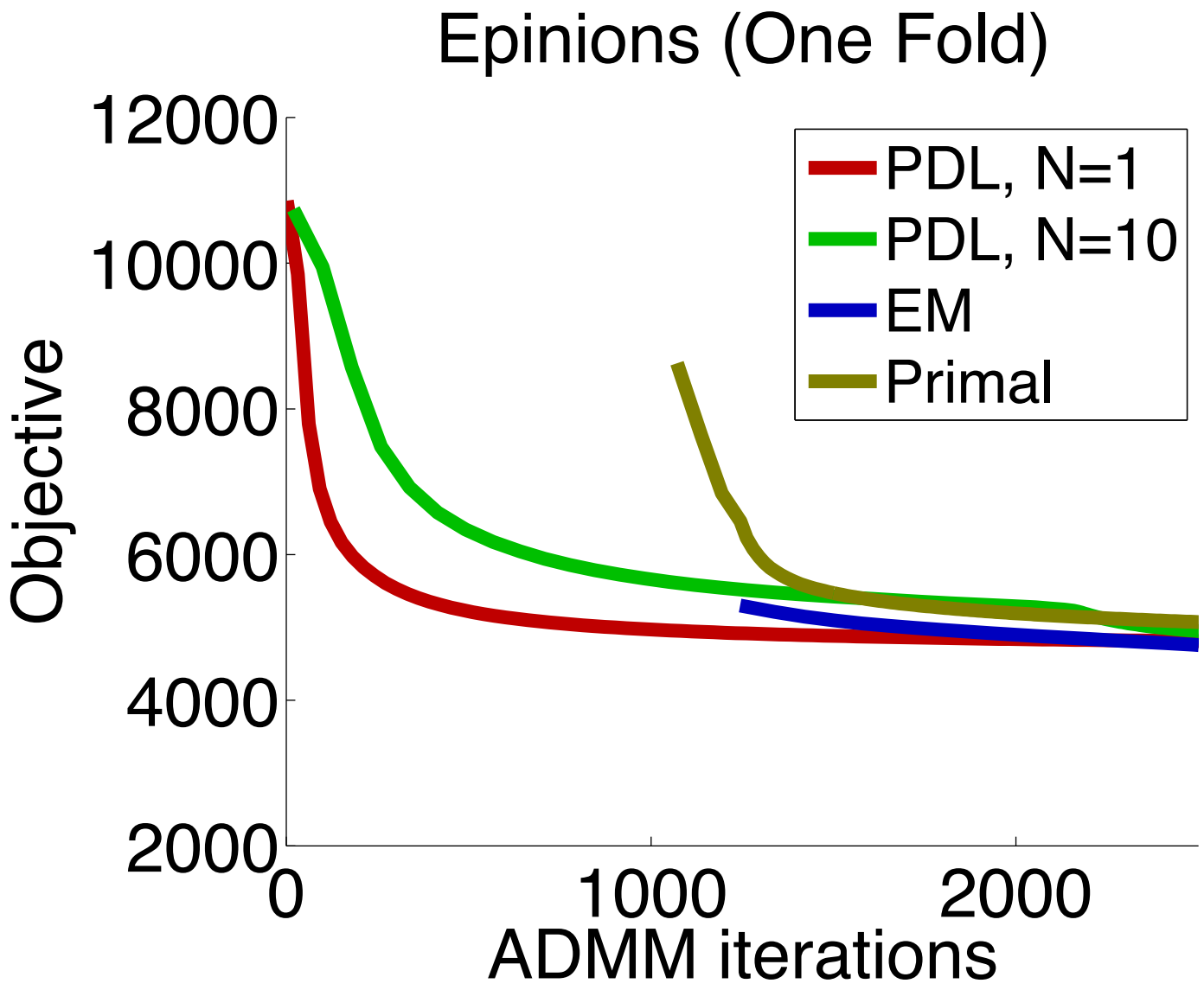

Twitter (One Fold)
AuPR
0.4
0.3
0.2
0.1
0
0
500
1000
1500
2000
2500
ADMM iterations
PDL, N=1
PDL, N=10
EM
Primal

Latent User Attributes

- Task: trust prediction in Epinions social network [Richardson et al., ISWC 2003]

- Latent variables represent whether users are:




Epinions (One Fold)
12000
10000
8000
6000
4000
2000
Objective
0
1000
2000
ADMM iterations
PDL, N=1
PDL, N=10
EM
Primal

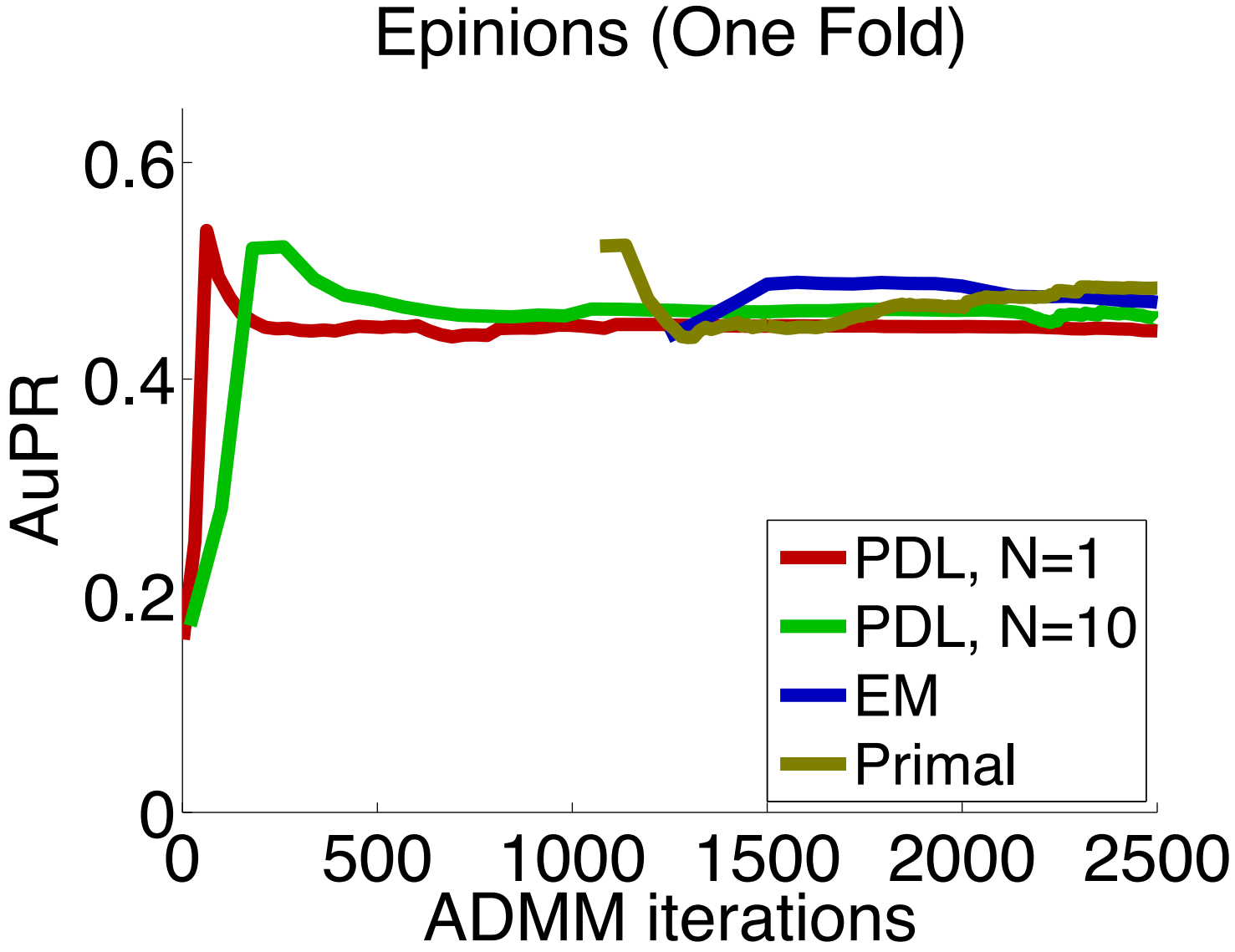

Epinions (One Fold)
AuPR
0.6
0.4
0.2
0
0
500
1000
1500
2000
2500
ADMM iterations
PDL, N=1
PDL, N=10
EM
Primal

Image Reconstruction

- Tested on Olivetti faces [Famaria and Harter, 1994], using experimental protocol of Poon and Domingos [UAI 2012]
- Latent variables capture facial structure

Originals With LVs Without

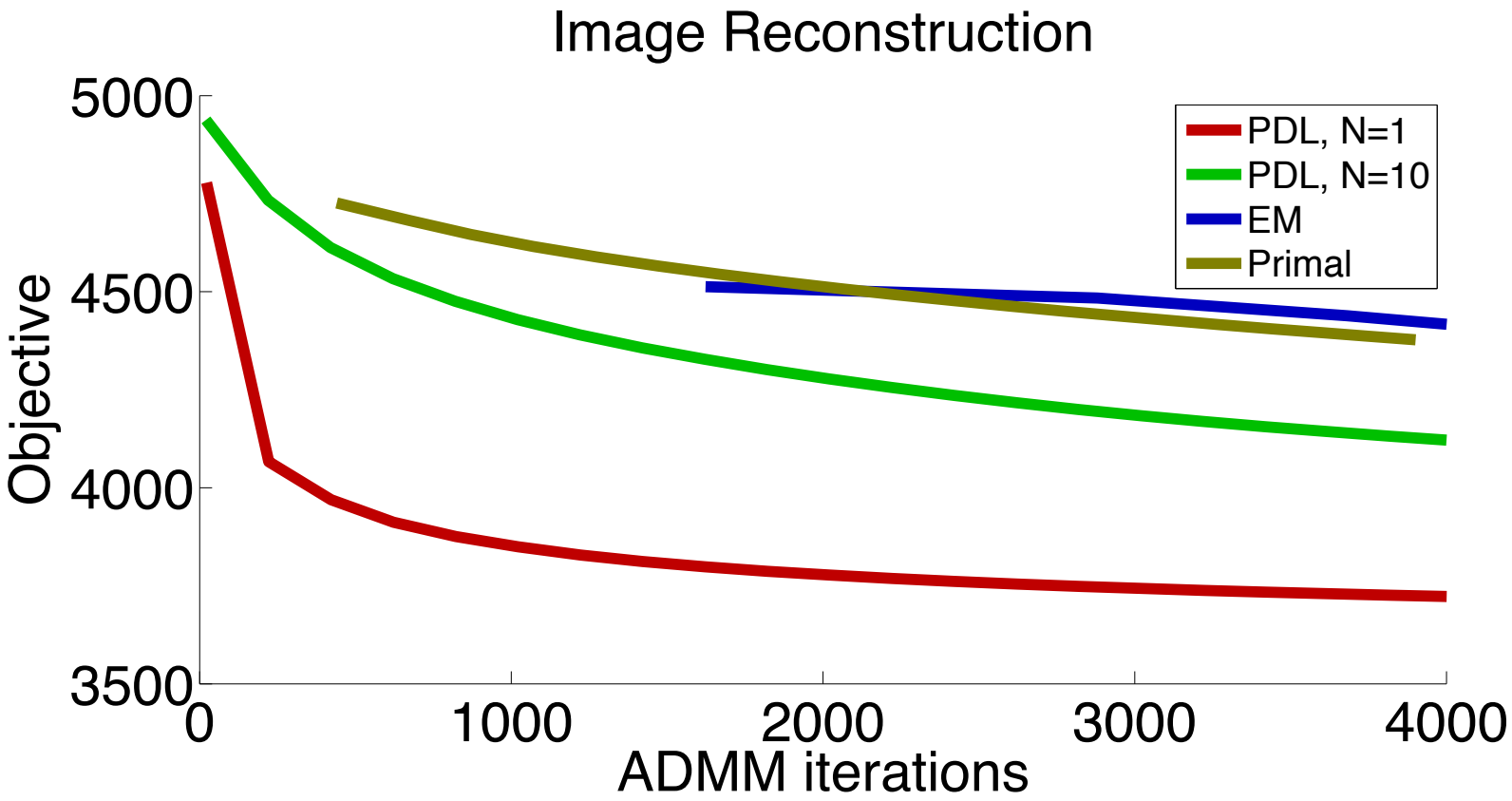


Image Reconstruction
5000
4500
4000
3500
Objective
0
1000
2000
3000
4000
ADMM iterations
PDL, N=1
PDL, N=10
EM
Primal

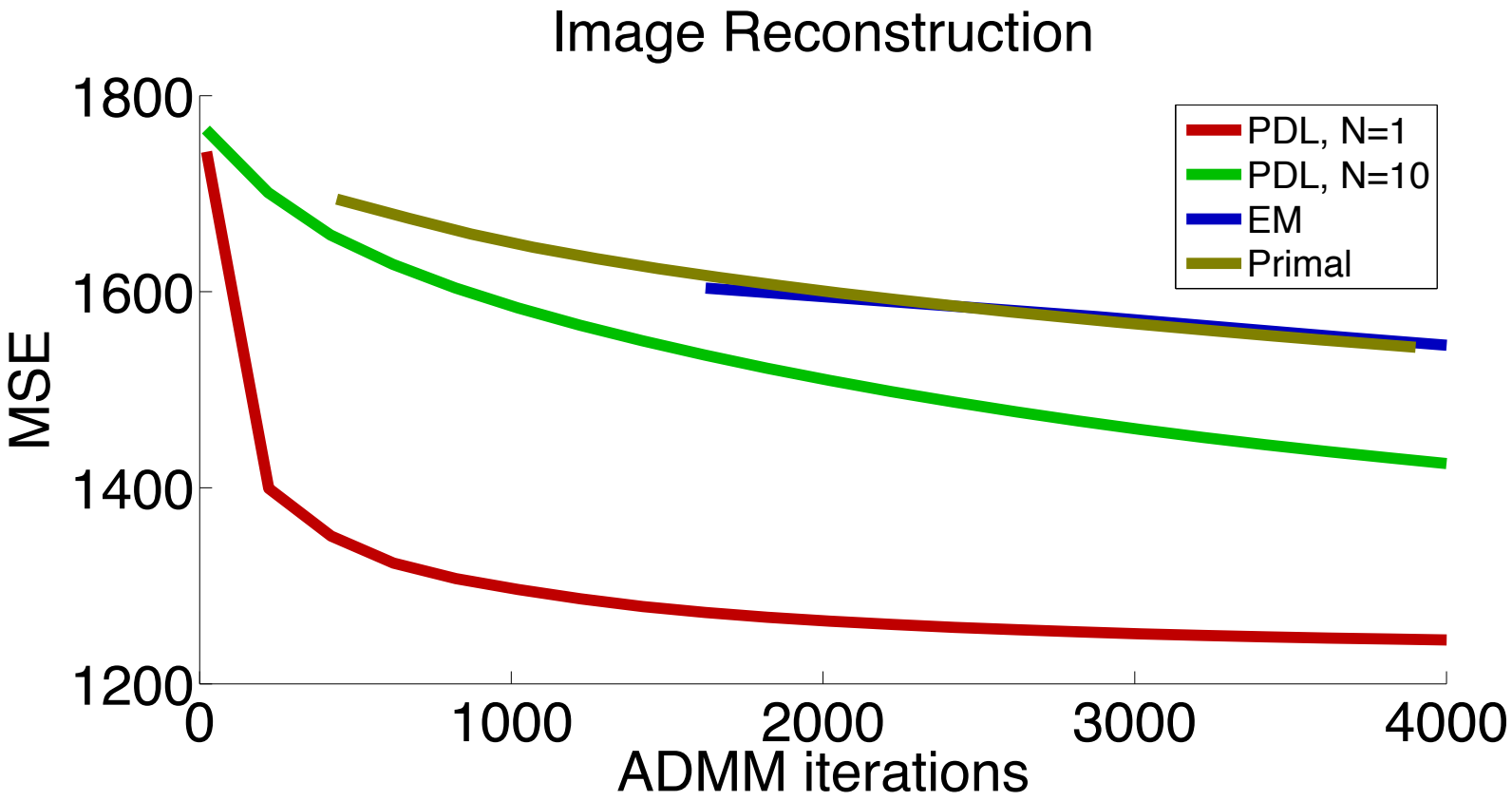


Image Reconstruction
MSE
1800
1600
1400
1200
0
1000
2000
3000
4000
ADMM iterations
PDL, N=1
PDL, N=10
EM
Primal

Conclusion

Conclusion

- Continuous latent variables
  - Capture rich, nuanced information in structured domains
  -
- Pa
  -
  - Makes large-scale, latent variable hinge-loss MRFs practical

- Open questions
  - Convergence proof for paired-dual learning
  - Should we also use it for discrete models?

Thank You!

bach@cs.umd.edu @stevebach