Neural Information Processing Systems
Lake Tahoe, 2012

# Dynamic Pruning of Factor Graphs for Maximum Marginal Prediction

Christoph H. Lampert
Institute of Science and Technology Austria (IST Austria)

IST AUSTRIA
Institute of Science and Technology

## Summary

**Situation: probabilistic graphical model with binary variables**
**Task: maximum marginal prediction (MMP):** decide state of each variable based on which value has higher *marginal* probability

Examples: *Bayes-optimal decision rule under Hamming loss, "maximum posterior marginals" prediction, multi-label classification, binary image restoration, ...*

**Contributions:**

- **early stopping rule for sampling-based inference**
- **method to adaptively shrink factor graph while sampling**

**Results: much faster decisions at negligible loss of accuracy**

## Common notation

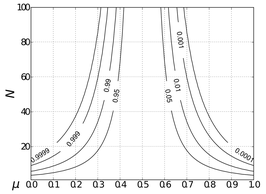
- $X_1, \dots, X_n$: $n$ random variable with joint state space $\mathcal{X} = \{0,1\}^n$,
- $\mathcal{G} = (V, \mathcal{F})$ factor graph with nodes $V = \{1, \dots, n\}$ and factors $\mathcal{F} = \{F_1, \dots, F_{|\mathcal{F}|}\}$; $F_k \subset V$,
- $\psi_F$: log-potential belonging to factor $F$,
- $p$: joint probability distribution over $\mathcal{X}$, $\quad p(X = x) \propto \exp\Big(\sum_{F \in \mathcal{F}} \psi_F(x_F)\Big)$,
- $\square_F$ subset of "$\square$" objects belonging to factor $F$, e.g. $x_F = (x_i)_{i \in F}$,
- $\mu_i = p(x_i = 1)$: marginal of variable $x_i$,
- $z_i = [\![\mu_i \geq \frac{1}{2}]\!]$: maximum marginal decision variable for variable $x_i$. $\quad \leftarrow$ main quantities of interest

## Early stopping rule: Single variable, i.i.d. samples

$x$ binary variable, $z$ decision variable, $\mathcal{S} = \{x^{(1)}, \dots, x^{(N)}\}$ i.i.d. sample from $p(x)$, $\; m = \sum_j x^{(j)}$.

$$p(z = 0|\mathcal{S}) = \int_0^{\frac{1}{2}} p(q|\mathcal{S})\,dq = \frac{1}{B_{m+1,N-m+1}} \int_0^{\frac{1}{2}} q^m (1-q)^{N-m}\,dq = I_{\frac{1}{2}}(m+1, N-m+1)$$

"Incomplete Beta-function"




## Early stopping rule: Single variable, non-i.i.d. samples

$x$ single binary variable, $\quad z$ its decision variable, $\quad \mathcal{S} = \{x^{(1)}, \dots, x^{(N)}\}$ non-i.i.d. sample from $p(x)$

**Idea:** Use **effective sample size (ESS)**, $N'$, estimated from first-order autocorrelation coefficient $r$:

$$p(z = 0|\mathcal{S}) = I_{\frac{1}{2}}(m' + 1, N' - m' + 1)$$

$$\hat{\mu} = \frac{1}{N}\sum_{j=1}^{N} x^{(j)}, \qquad r = \frac{1}{N-1}\sum_{j=1}^{N-1} \frac{(x^{(j)} - \hat{\mu})(x^{(j+1)} - \hat{\mu})}{\hat{\sigma}^2}, \qquad N' = \frac{1-r}{1+r}N, \qquad m' = \hat{\mu}N'$$

## Early stopping rule: Graphical model

$X$ several binary variables, joint distribution $p$ given by factor graph $\mathcal{G} = (V, \mathcal{F})$

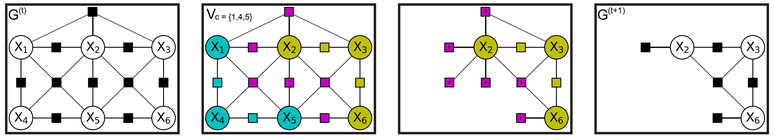
**Situation:** Sampling from joint $p(x)$ creates dependent samples $x^{(1)}, \dots, x^{(N)}$.
Each $x_i^{(j)}$ is a sample from marginal $p(x_i)$, forming $\mathcal{S}_i := \{x_i^{(1)}, \dots, x_i^{(N)}\}$.
From $\mathcal{S}_i$ we can estimate confidence for $z_i$, independently for each $i = 1, \dots, n$

**Algorithm:**

- input: factor graph $\mathcal{G} = (V, \mathcal{F})$, batch size $k$, confidence threshold $\epsilon > 0$
- initialize: $\mathcal{G}^{(0)} = (V^{(0)}, \mathcal{F}^{(0)}) \leftarrow \mathcal{G}$
- for $t = 0, \dots, T_{max}$, and while $V^{(t)} \neq \emptyset$:
  - $x^{(1)}, \dots, x^{(k)} \leftarrow$ sample $k$ times from $p^{(t)}(x)$ for $x \in \{0,1\}^{|V^{(t)}|}$
  - for all $j \in V^{(t)}$:
    - update statistics $(\hat{\mu}_i, \hat{\sigma}_i^2, r_i, N_i')$ by $x_i^{(1)}, \dots, x_i^{(k)}$
  - $V_c \leftarrow \{j \in V^{(t)} : (p(z_i = 0|\hat{\mu}_i, N_i') < \epsilon) \vee (p(z_i = 0|\hat{\mu}_i, N_i') > 1 - \epsilon)\}$ (confident set)
  - if $V_c \neq \emptyset$:
    - $\mathcal{G}^{(t+1)} \leftarrow$ **prune**($V_c$ **from** $\mathcal{G}^{(t)}$)
  - else:
    - $\mathcal{G}^{(t+1)} \leftarrow \mathcal{G}^{(t)}$

Remaining question: how to **prune** some variables from a factor graph?

## Main contribution: Dynamic pruning of factor graphs




**Algorithm:**

- input: factor graph $\mathcal{G} = (V, \mathcal{F})$, set of variables to prune $V_c$, set of variable to keep $V_u = V \setminus V_c$.
- initialize: $\tilde{\mathcal{F}} \leftarrow \emptyset$, $\tilde{\Psi} \leftarrow \emptyset$
- for $F \in \mathcal{F}$:
  - $F_u \leftarrow F \cap V_u, \quad F_c \leftarrow F \cap V_c$
  - if $F_u = \emptyset$: do nothing
  - else if $F_u = F$: $\quad \tilde{\mathcal{F}} \leftarrow \tilde{\mathcal{F}} \cup \{F\}; \quad \tilde{\Psi} \leftarrow \tilde{\Psi} \cup \{\Psi_F\}$,
  - else $\emptyset \subsetneq F_u \subsetneq F$:
    - $\Psi_{F_u}(x_u) = \sum_{\bar{x}_c \in \{0,1\}^{|F_c|}} q(\bar{x}_c)\Psi_F(\bar{x}_c, x_u)$ with $q(\bar{x}_c) = \prod_{i \in F_c} \hat{\mu}_i^{\bar{x}_i}(1 - \hat{\mu}_i)^{1 - \bar{x}_i}$,
    - $\tilde{\mathcal{F}} \leftarrow \tilde{\mathcal{F}} \cup \{F_u\}; \quad \tilde{\Psi} \leftarrow \tilde{\Psi} \cup \{\Psi_{F_u}\}$,
- merge factors of duplicate domain within $\tilde{F}$ by adding their log-potentials
- output: new factor graph: $\tilde{\mathcal{G}} = (V_u, \tilde{\mathcal{F}})$

## Derivation: Partial mean field approximation

**Idea:** Find best distributions $q_u(x_u)$, $q_c(x_c)$, such that $\quad p(x_u, x_c) \approx q_u(x_u)q_c(x_c)$.

- Assumption: $q_c(x_c) = \prod_{i \in V_c} q_i(x_i)$, with $\mathbb{E}_{x_i \sim q_i}\{x_i\} = \hat{\mu}_i \quad \rightarrow \quad q_c(x_c) = \prod_{i \in F_c} \hat{\mu}_i^{x_i}(1 - \hat{\mu}_i)^{1 - x_i}$,
- Minimizing $\mathrm{KL}(p||q_u q_c) \quad \rightarrow \quad q_u(x_c) \propto \exp(\,\Psi_u(x_u)\,)$ with $\Psi_u(x_u) = \mathbb{E}_{\bar{x}_c \sim q_c}\{\Psi(x_u, \bar{x}_c)\}$

Above algorithm computes this expectation, using the decomposition $\Psi(x) = \sum_{F \in \mathcal{F}} \Psi_F(x_F)$.

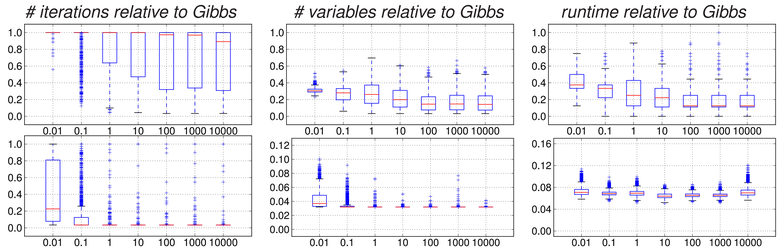
## Experiment I: Multi-label classification

Fully connected pairwise CRF: $V = \{1, \dots, n\}$, $\mathcal{F} = \{\text{all pairs}\}$, $p(x)$ learned from training data

**Sanity check:** Compare Hamming loss for
1) exact inference, where possible,
2) Gibbs sampling (500 iterations),
3) proposed adaptive sampling with tolerances $\epsilon \in \{10^{-2}, 10^{-5}, 10^{-8}\}$

No loss of accuracy from early stopping.

| Dataset | #Labels | #Train | #Test | [1] | [2] | Exact | Gibbs | Proposed $\epsilon = 10^{-2}$ | Proposed $\epsilon = 10^{-5}$ | Proposed $\epsilon = 10^{-8}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Synth1 | 6 | 471 | 5045 | 6.9 | – | 5.2 | 5.3 | 5.2 | 5.2 | 5.2 |
| Synth2 | 10 | 1000 | 10000 | 7.0 | – | 10.0 | 10.0 | 10.0 | 10.0 | 10.0 |
| Scene | 6 | 1211 | 1196 | 10.1 | 9.5 | 10.4 | 10.3 | 10.2 | 10.2 | 10.2 |
| Rcv1-10 | 10 | 2916 | 2914 | 5.6 | – | 4.2 | 4.2 | 4.6 | 4.4 | 4.2 |
| Mediamill-10 | 10 | 29415 | 12168 | 18.8 | – | 18.4 | 18.6 | 19.0 | 18.6 | 18.4 |
| Yeast | 14 | 1500 | 917 | 20.2 | 20.2 | 20.0 | 20.2 | 23.4 | 21.4 | 20.5 |
| Tmc2007 | 22 | 21519 | 7077 | – | 3.3 | 5.3 | 5.3 | 5.3 | 5.3 | 5.3 |
| AwA | 85 | 24295 | 6180 | – | – | – | 32.2 | 32.7 | 32.7 | 32.7 |
| Mediamill | 101 | 29415 | 12168 | – | 3.6 | – | 3.7 | 3.6 | 3.5 | 3.6 |
| Rcv1 | 103 | 3000 | 3000 | – | – | – | 1.5 | 1.7 | 1.6 | 1.5 |




Top row: Yeast dataset, bottom row: Rcv1 dataset. x-axis: inverse regularization strength

- often reduced number of iterations, but hard cases are respected
- always reduced number of samples and runtime, sometimes dramatically so

## Experiment II: Binary image inpainting

Grid MRF: $V = \{\text{all pixels}\}(\approx 80 \times 100)$, $\mathcal{F} = \{\textit{short and long range pairs}\}$, $p(x)$ from [3]

- run inference for 30 seconds: Gibbs vs. proposed adaptive sampling

*input image with (masked region in gray)* — *Gibbs reconstruction (marginals / # samples per pixel)* — *Adaptive reconstruction (marginals / # samples)*

Adaptive sampling stops early for 'easy' regions and concentrates samples in hard regions.

## Summary and outlook

- Adaptive sampling allows on-the-fly early stopping for data-dependent subsets of variables
- In our experiments: significant increase in speed at minor, if any, loss in accuracy
- Future work: better understand the approximation error, go beyond partial mean-field assumption

[1] T. Finley, T. Joachims: *Training structural SVMs when exact inference is intractable*, ICML 2008.
[2] G. Madjarov, D. Kocev, D. Gjorgjevikj, S. Dzeroski: *An extensive experimental comparison of methods for multi-label learning*, Pattern Recognition, 2012.
[3] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, P. Kohli: *Decision tree fields*, ICCV 2011.

Created with LaTeX beamerposter http://www-i6.informatik.rwth-aachen.de/~dreuw/latexbeamerposter.php