Max-margin Learning for Lower Linear Envelope Potentials in Binary Markov Random Fields

Stephen Gould
stephen.gould@anu.edu.au

Australian National University

ICML, 29 June 2011
Motivation: Image Labeling

**Image labeling**: Label every pixel in an image with a class label from some pre-defined set, i.e., $y_p \in \mathcal{L}$. 
Motivation: Image Labeling

**Image labeling:** Label every pixel in an image with a class label from some pre-defined set, i.e., $y_p \in \mathcal{L}$.

- **Interactive segmentation** (Boykov and Jolly, 2001; Boykov and Funka-Lea, 2006)
- **Surface context** (Hoiem et al., 2005)
- **Semantic labeling** (He et al., 2004; Shotton et al., 2006; Gould et al., 2009)
- **Stereo matching** (Scharstein and Szeliski, 2002)
- **Photo montage** (Agarwala et al., 2004)
- **Denoising**
Motivation: Image Labeling

These problems are typically solved using a pairwise conditional Markov random field.

However, pairwise terms are often not expressive enough.
Consistency Potentials

Suppose an oracle told us which pixels belong together, e.g., for the figure-ground segmentation problem we might have
Consistency Potentials

Suppose an oracle told us which pixels belong together, e.g., for the figure-ground segmentation problem we might have

Then we would only need to label the so-called superpixels rather than individual pixels.
Consistency Potentials

Unfortunately we don’t have a perfect oracle. So what can we do?

[penalty]

[number of disagreements]

[Kohli et al., 2007]
Consistency Potential

Unfortunately we don’t have a perfect oracle. So what can we do?

[Kohli et al., 2007]  [Kohli et al., 2008]
Consistency Potentials

Unfortunately we don’t have a perfect oracle. So what can we do?

[Kohli et al., 2007]  [Kohli et al., 2008]  [Kohli and Kumar, 2010]
Higher-order Markov Random Fields

The **energy function** for a higher-order MRF over discrete random variables \( y = \{y_1, \ldots, y_n\} \) can be written as:

\[
E(y; x, \theta) = \sum_{c} \psi_c(y_c)
\]
Higher-order Markov Random Fields

The energy function for a higher-order MRF over discrete random variables \( y = \{y_1, \ldots, y_n\} \) can be written as:

\[
E(y; x, \theta) = \sum_c \psi_c(y_c)
\]

\[
= \sum_i \psi^U_i(y_i) + \sum_{ij} \psi^P_{ij}(y_i, y_j) + \sum_c \psi^H_c(y_c)
\]

where the potential functions \( \psi^U_i \), \( \psi^P_{ij} \) and \( \psi^H_c \) encode preferences for unary, pairwise and \( k \)-ary variable assignments, respectively.

The goal of inference is to find \( y^* = \arg\min_y E(y; x, \theta) \).
Binary Lower Linear Envelope MRFs

\[ \psi^H_c(y_c) \triangleq \min_k \left\{ a_k \sum_{i \in C} y_i + b_k \right\} \]
Binary Lower Linear Envelope MRFs

\[ \psi_c^H(y_c) \triangleq \min_k \left\{ a_k \sum_{i \in C} y_i + b_k \right\} \]
Energy Minimization (Kohli and Kumar, CVPR 2010)

\[ \psi_c^H(y_c) \triangleq \min_k \left\{ a_k \sum_{i \in C} y_i + b_k \right\} = \min_k \left\{ f_k(y_c) \right\} \]
Energy Minimization ([Kohli and Kumar, CVPR 2010])

$$\psi^H_c(y_c) \triangleq \min_k \left\{ a_k \sum_{i \in C} y_i + b_k \right\} = \min_k \{ f_k(y_c) \}$$

Introduce multi-valued auxiliary random variable $z \in \{1, \ldots, K\}$ and write

$$\tilde{\psi}^H_c(y_c, z) = \sum_k \left[ z = k \right] f_k(y_c).$$

Now minimize jointly over $y$ and $z$. 

![Diagram](attachment:diagram.png)
Energy Minimization (Attempt 2)

\[
\psi^H_c(y_c) \triangleq \min_k \left\{ a_k \sum_{i \in C} y_i + b_k \right\} = \min_k \left\{ f_k(y_c) \right\}
\]

Introduce auxiliary binary random variables \( z = (z_1, \ldots, z_K) \) with mutual exclusion constraint and write

\[
\tilde{\psi}^H_c(y_c, z) = \sum_k z_k f_k(y_c) \quad \text{s.t.} \quad \sum_k z_k = 1.
\]

Now minimize jointly over \( y \) and \( z \).
Energy Minimization (Attempt 3)

\[ \psi^H_c(y_c) \triangleq \min_k \left\{ a_k \sum_{i \in C} y_i + b_k \right\} = \min_k \{ f_k(y_c) \} \]

Assume sorted on \( a_k \). Introduce auxiliary binary random variables \( z = (z_1, \ldots, z_{K-1}) \) with *inclusion* constraints and write

\[ \tilde{\psi}^H_c(y_c, z) = \text{unary} f_1(y_c) + \sum_k \text{unary and pairwise } z_k (f_{k+1}(y_c) - f_k(y_c)) \text{ s.t. } z_k \geq z_{k+1}. \]

Now minimize jointly over \( y \) and \( z \).
Each transformation results in a different latent variable Markov random field:

Attempt 1
(multi-valued; pairwise)

Attempt 2
(binary; non-pairwise)

Attempt 3
(binary; pairwise)
Claim 1: The binary pairwise MRF induced by “Energy Minimization Attempt 3” is submodular (see paper for proof)
Exact Inference

- **Claim 1**: The binary pairwise MRF induced by “Energy Minimization Attempt 3” is submodular (see paper for proof)
- **Claim 2**: Submodular binary MRFs can be minimized in time polynomial in the number of variables ([Hammer, 1965])
Exact Inference

- **Claim 1:** The binary pairwise MRF induced by “Energy Minimization Attempt 3” is submodular (see paper for proof)
- **Claim 2:** Submodular binary MRFs can be minimized in time polynomial in the number of variables ([Hammer, 1965])
  - Empirically, very fast algorithms exist for quadratic submodular problems ([Boykov and Kolmogorov, 2004])
Exact Inference

- **Claim 1:** The binary pairwise MRF induced by “Energy Minimization Attempt 3” is submodular (see paper for proof)
- **Claim 2:** Submodular binary MRFs can be minimized in time polynomial in the number of variables ([Hammer, 1965])
  - Empirically, very fast algorithms exist for quadratic submodular problems ([Boykov and Kolmogorov, 2004])
- **We can perform exact inference in lower linear envelope binary Markov random fields**
Max-margin Learning for Structured Prediction
Max-margin Learning for Structured Prediction

- **Energy function.** Parameterized by $\theta \in \mathbb{R}^d$,

$$E(y; x, \theta) = \sum_c \psi_c(y_c; x, \theta_c) = \theta^T \phi(y, x)$$

- Easy inference
- Easy learning
Max-margin Learning for Structured Prediction

- **Energy function.** Parameterized by $\theta \in \mathbb{R}^d$,

  \[
  E(y; x, \theta) = \sum_c \psi_c(y_c; x, \theta_c) = \theta^T \phi(y, x)
  \]

  \[
  \text{easy learning}
  \]

  \[
  \text{easy inference}
  \]

- **Structured loss function.** e.g., $\Delta(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^{n} [\hat{y}_i \neq y_i]$

  \[
  \text{stephen gould}
  \]
Max-margin Learning for Structured Prediction

- **Energy function.** Parameterized by $\theta \in \mathbb{R}^d$,

$$E(y; x, \theta) = \sum_c \psi_c(y_c; x, \theta_c) = \theta^T \phi(y, x)$$

- **Structured loss function.** e.g., $\Delta(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^{n} [\hat{y}_i \neq y_i]$  

- **Learning algorithm.** Given a training set $\{(x_t, y_t)\}_{t=1}^{T}$, solve the *margin-rescaling* optimization problem ([Taskar et al., 2005; Tsochantaridis et al., 2004]).
Max-margin Learning for Structured Prediction

**QP for max-margin learning**

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| \theta \|^2_2 + \frac{C}{T} \sum_t \xi_t \\
\text{subject to} & \quad \theta^T \phi_t(y) - \theta^T \phi_t(y_t) \geq \Delta(y, y_t) - \xi_t, \quad \forall t, \ y \in \mathcal{Y}_t \\
& \quad \xi_t \geq 0, \quad \forall t
\end{align*}
\]

- energy difference
- rescaled margin
- \(\Delta(y, y_t)\) very large for very large \(\xi_t\)
### Max-margin Learning for Structured Prediction

**QP for max-margin learning**

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2}\|\theta\|_2^2 + \frac{C}{T} \sum_t \xi_t \\
\text{subject to} & \quad \theta^T \phi_t(y) - \theta^T \phi_t(y_t) \geq \Delta (y, y_t) - \xi, \quad \forall t, \ y \in \mathcal{Y}_t \\
& \quad \xi_t \geq 0, \quad \forall t
\end{align*}
\]

**Re-writing constraints**

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2}\|\theta\|_2^2 + \frac{C}{T} \sum_t \xi_t \\
\text{subject to} & \quad \xi_t \geq \max_{y \in \mathcal{Y}_t} \left\{ \Delta (y, y_t) - \theta^T \phi_t(y) \right\} + \theta^T \phi_t(y_t), \quad \forall t \\
& \quad \xi_t \geq 0, \quad \forall t
\end{align*}
\]
Lower Linear Envelope Representation

It remains to represent the lower linear envelope in a form that is amenable to learning.

\[
\psi^H_c(y_c) \triangleq \min_k \left\{ a_k \sum_{i \in C} y_i + b_k \right\} = \theta^T \phi(y_c)
\]
Lower Linear Envelope Representation for Learning

- Sample-based representation with concavity constraints:
  \[2\theta_k - \theta_{k-1} - \theta_{k+1} \geq 0\]
Lower Linear Envelope Representation for Learning

- Sample-based representation with concavity constraints:
  \[ 2\theta_k - \theta_{k-1} - \theta_{k+1} \geq 0 \]

- Features \( \phi(y_c) \) are 1-of-\( n \) indicator vectors

- Can extend to be invariant of clique size
Max-margin Learning for Lower Linear Envelope MRFs

QP for lower linear envelope MRF learning

\[
\begin{align*}
\text{minimize} \quad & \frac{1}{2}\|\theta\|_2^2 + \frac{C}{T} \sum_t \xi_t \\
\text{subject to} \quad & \theta^T \phi_t(y) - \theta^T \phi_t(y_t) \geq \Delta(y, y_t) - \xi_t, \quad \forall t, y \in Y_t \\
& \xi_t \geq 0, \quad \forall t \\
& D^2 \theta \geq 0
\end{align*}
\]
Max-margin Learning for Lower Linear Envelope MRFs

QP for lower linear envelope MRF learning

\[
\begin{align*}
\text{minimize}_{\theta, \xi} & \quad \frac{1}{2} \| \theta \|_2^2 + \frac{C}{T} \sum_t \xi_t \\
\text{subject to} & \quad \theta^T \phi_t(y) - \theta^T \phi_t(y_t) \geq \Delta(y, y_t) - \xi_t, \quad \forall t, y \in \mathcal{Y}_t \\
& \quad \xi_t \geq 0, \quad \forall t \\
& \quad \mathbf{D}^2 \theta \geq 0
\end{align*}
\]

- Learning algorithm repeatedly
  - solves above QP using sampled representation \( \theta \)
  - finds violated constraints using lower linear envelope representation \( \{(a_k, b_k)\} \)

- Variants of the feature representation and corresponding learning objective can also be used.
Synthetic Experiments
In these experiments the ground-truth location of the squares are given.
Synthetic Results
Synthetic Results

groundtruth | data | pairwise crf | 3rd iteration | final iteration
Synthetic Results

groundtruth | data | pairwise crf | 3rd iteration | final iteration

---
groundtruth | data | pairwise crf | 3rd iteration | final iteration
Learned Parameters for Synthetic Experiments

\[ \eta = (0.1, 0.1) \]

\[ \eta = (0.5, 0.1) \]

\( \eta \) is the signal-to-noise ratio.
Interactive Image Segmentation

“GrabCut” [Rother et al., SIGGRAPH 2004]
Interactive Image Segmentation

- “GrabCut” [Rother et al., SIGGRAPH 2004]

- Our experimental setup
  - leave-one-out cross-validation on 50 images
  - baseline: 8-neighborhood pairwise CRF
  - higher-order: lower linear envelope potential on non-overlapping superpixels
“GrabCut” Experiments
“GrabCut” Experiments

- Superpixels determined via a bottom-up unsupervised approach.
“GrabCut” Results

<table>
<thead>
<tr>
<th>image</th>
<th>truth</th>
<th>baseline</th>
<th>higher-order</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image.png" alt="Flower Image" /></td>
<td><img src="truth.png" alt="Flower Truth" /></td>
<td><img src="baseline.png" alt="Flower Baseline" /></td>
<td><img src="higher-order.png" alt="Flower Higher-Order" /></td>
</tr>
</tbody>
</table>
“GrabCut” Results

<table>
<thead>
<tr>
<th>image</th>
<th>truth</th>
<th>baseline</th>
<th>higher-order</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="image" /></td>
<td><img src="truth1.jpg" alt="truth" /></td>
<td><img src="baseline1.jpg" alt="baseline" /></td>
<td><img src="higher-order1.jpg" alt="higher-order" /></td>
</tr>
<tr>
<td><img src="image2.jpg" alt="image" /></td>
<td><img src="truth2.jpg" alt="truth" /></td>
<td><img src="baseline2.jpg" alt="baseline" /></td>
<td><img src="higher-order2.jpg" alt="higher-order" /></td>
</tr>
</tbody>
</table>
“GrabCut” Results

<table>
<thead>
<tr>
<th>Image</th>
<th>Truth</th>
<th>Baseline</th>
<th>Higher-Order</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Image" /></td>
<td><img src="truth1.jpg" alt="Truth" /></td>
<td><img src="baseline1.jpg" alt="Baseline" /></td>
<td><img src="higher-order1.jpg" alt="Higher-Order" /></td>
</tr>
<tr>
<td><img src="image2.jpg" alt="Image" /></td>
<td><img src="truth2.jpg" alt="Truth" /></td>
<td><img src="baseline2.jpg" alt="Baseline" /></td>
<td><img src="higher-order2.jpg" alt="Higher-Order" /></td>
</tr>
<tr>
<td><img src="image3.jpg" alt="Image" /></td>
<td><img src="truth3.jpg" alt="Truth" /></td>
<td><img src="baseline3.jpg" alt="Baseline" /></td>
<td><img src="higher-order3.jpg" alt="Higher-Order" /></td>
</tr>
</tbody>
</table>
“GrabCut” Results

<table>
<thead>
<tr>
<th>image</th>
<th>truth</th>
<th>baseline</th>
<th>higher-order</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image 1]</td>
<td>![Truth 1]</td>
<td>![Baseline 1]</td>
<td>![Higher-order 1]</td>
</tr>
<tr>
<td>![Image 2]</td>
<td>![Truth 2]</td>
<td>![Baseline 2]</td>
<td>![Higher-order 2]</td>
</tr>
<tr>
<td>![Image 3]</td>
<td>![Truth 3]</td>
<td>![Baseline 3]</td>
<td>![Higher-order 3]</td>
</tr>
</tbody>
</table>

- Quantitatively we see a 15% reduction in error rate.
- Simply enforcing hard consistency within superpixels results in a 1% increase in error rate.
Summary

- motivation
  - higher-order models are important for image understanding
Summary

- **motivation**
  - higher-order models are important for image understanding

- **this work—binary lower linear envelope potentials**
  - telescoping-sum construction for exact MAP inference in time polynomial in the number of variables and number of linear envelope functions
  - representation for learning parameters of lower linear envelope potentials using max-margin framework
  - demonstrated in the context of figure-ground segmentation
Summary

- **motivation**
  - higher-order models are important for image understanding

- **this work—binary lower linear envelope potentials**
  - telescoping-sum construction for exact MAP inference in time polynomial in the number of variables and number of linear envelope functions
  - representation for learning parameters of lower linear envelope potentials using max-margin framework
  - demonstrated in the context of figure-ground segmentation

- **future work**
  - apply to multi-class setting
  - explore relationship with latent-variable SVMs
Summary

- **motivation**
  - higher-order models are important for image understanding

- **this work—binary lower linear envelope potentials**
  - telescoping-sum construction for exact MAP inference in time polynomial in the number of variables and number of linear envelope functions
  - representation for learning parameters of lower linear envelope potentials using max-margin framework
  - demonstrated in the context of figure-ground segmentation

- **future work**
  - apply to multi-class setting
  - explore relationship with latent-variable SVMs

- **questions?**