Structural Maxent Models

Vitaly Kuznetsov$^1$

Joint work with Corinna Cortes$^2$, Mehryar Mohri$^{1,2}$ and Umar Syed$^2$

$^1$Courant Institute of Mathematical Sciences, New York University

$^2$Google Research, New York
Species Habitat Modeling
Input: a corpus $C$, a sequence of words over vocabulary $V$.

Goal: estimate the probability of a sequence of words $P(w_1, \ldots, w_n)$.

Critical component of speech recognition and other natural language processing models.
Density Estimation

Unsupervised Learning Scenario:
- $S = (x_1, \ldots, x_m)$ an i.i.d. sample from unknown distribution $\mathcal{D}$.
- A feature mapping $\Phi : \mathcal{X} \rightarrow \mathcal{F}$.
- Find a distribution $p$ that estimates $\mathcal{D}$.
Maxent Principle: Select the distribution that is the closest to the uniform, so that the average value of each feature matches its empirical value.

Key benefits:
- Diverse features can be used.
- Good theoretical guarantees.

Challenge: Is it possible to use a richer set of features and yet not overfit?
Outline

- Learning Scenario.
- Structural Maxent Principal.
- Duality.
- Learning Guarantees.
- Algorithm.
- Experiments.
Prior Work

- Maxent principle. (Jaynes, 1957, 1983)
- Maxent modeling in NLP. (Berger et al. 1996; Pietra et al. 1997; Malouf 2002; Manning & Klein, 2003; Ratnaparkhi, 2010)
- Maxent modeling in ecology. (Philips et. al., 2004, 2006; Elith et. al. 2011)
- Regularized Maxent. (Kazama & Tsujii, 2003; Chen & Rosenfeld, 2000; Lebanon & Lafferty, 2001; Dudik et. al., 2007)
- Bayesian interpretation. (Williams, 1994; Goodman, 2004)
Learning Setup: Families of Features

- $H_1, \ldots, H_p$ families of feature maps.
- Feature mapping $\Phi : \mathcal{X} \rightarrow \mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_p$.
- $\forall x \in \mathcal{X}, \Phi(x) = (\Phi_1(x), \ldots, \Phi_p(x))$.
- $\forall k \in \{1, \ldots, p\}, \Phi_{j,k} \in H_k$ and $\|\Phi_k\|_{\infty} \leq \Lambda$. 

![Diagram showing families of feature maps](attachment:diagram.png)
Uniform Convergence Bounds

(Koltchinskii & Panchenko, 2002)

For any $\delta > 0$, with probability at least $1 - \delta$:

$$\left\| \mathbb{E}_{x \sim D} [\Phi_k(x)] - \mathbb{E}_{x \sim S} [\Phi_k(x)] \right\|_\infty \leq \sqrt{2 \mathcal{R}_m(H_k) + 2 \sqrt{\log \frac{2p}{\delta}} + \sqrt{\frac{\log 2m}{2m}}}$$

for all $k \in [1, p]$, where $\mathcal{R}_m(H_k)$ is the Rademacher complexity of $H_k$. 
Structural Maxent Principal

Find distribution p closest to some distribution \( p_0 \) subject to uniform convergence constraints:

\[
\min_{p \in \Delta} D(p \parallel p_0), \quad \text{s.t. } \forall k \in [1, p] : \\
\left\| \mathbb{E}_{x \sim p} [\Phi_k(x)] - \mathbb{E}_{x \sim S} [\Phi_k(x)] \right\|_\infty \leq 2 \mathfrak{K}_m(H_k) + \beta,
\]

where

\[
D(p \parallel q) = \sum_{x \in \mathcal{X}} p[x] \log \frac{p[x]}{q[x]}
\]

is the relative entropy.
Dual Objective

\[ G(w) = \frac{1}{m} \sum_{i=1}^{m} \log \left[ \frac{p_w(x_i)}{p_0(x_i)} \right] - \sum_{k=1}^{p} \beta_k \|w_k\|_1, \]

- **Weights:** \( \beta_k = 2\mathcal{H}_m(H_k) + \beta. \)
- **Gibbs distribution:** \( p_w = \frac{p_0[x] e^{w \cdot \Phi(x)}}{Z_w}. \)
- **Partition function:** \( Z_w = \sum_{x \in \mathcal{X}} p_0[x] e^{w \cdot \Phi(x)}. \)
Theorem

*Primal Struct Maxent problem is equivalent to the dual Struct Maxent problem* $\sup_{w \in \mathbb{R}^N} G(w)$:

$$\sup_{w \in \mathbb{R}^N} G(w) = \min_{p} F(p).$$

Furthermore, let $p^* = \arg\min_{p} F(p)$, then, for any $\epsilon > 0$ and any $w$ such that $\left| G(w) - \sup_{w \in \mathbb{R}^N} G(w) \right| < \epsilon$, the following holds:

$$D(p^* \parallel p_w) \leq \epsilon.$$
Generalization Bound

Theorem

Fix $\delta > 0$. Let $\hat{\mathbf{w}}$ be a solution to the dual Struct Maxent problem with $\beta = \Lambda \sqrt{\frac{\log \frac{2p}{\delta}}{2m}}$. Then with probability at least $1 - \delta$,

$$
\mathcal{L}_D(\hat{\mathbf{w}}) \leq \inf_{\mathbf{w}} \mathcal{L}_D(\mathbf{w}) + 2 \sum_{k=1}^{p} \|\mathbf{w}_k\|_1 \left[ 2\mathcal{K}_m(H_k) + \Lambda \sqrt{\frac{\log \frac{2p}{\delta}}{2m}} \right],
$$

where $\mathcal{L}_D(\mathbf{w}) = \mathbb{E}_{x \sim D}[-\log p_\mathbf{w}[x]]$. 
Generalization Bound: Consequences

- Favorable learning guarantees for Structural Maxent models.
- Learning with complex features is possible as long as their weight is small.
Unconstrained optimization: \( \inf_w F(w) \) with

\[
F(w) = \sum_{k=1}^{p} \beta_k \| w_k \|_1 - w \cdot \mathbb{E}[\Phi] + \log \left[ \sum_{x \in \mathcal{X}} p_0[x] e^{w \cdot \Phi(x)} \right],
\]

and \( \beta_k = \beta + 2 \Re m(H_k) \) for \( k \in [1, p] \).
Coordinate Descent

Repeat until convergence:

• Find descent direction

\[(k, j) = \arg\max_{(k,j)} |\delta_{(k,j)} F(w_{t-1})|,\]

where \(\delta_{(k,j)} F(w_{t-1})\) is the subgradient of \(F(w_{t-1})\) in the direction \((k, j)\).

• Find step size \(\eta\) (line search; closed-form size).

\[w_t = w_{t-1} + \eta e_{(k,j)}.\]

Convergence analysis:

• Convergence for closed-form step sizes.
• Linear convergence for line search.
## Experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Maxent</th>
<th>$L_1$-Maxent</th>
<th>Structural Maxent</th>
</tr>
</thead>
<tbody>
<tr>
<td>b. variegatus</td>
<td>19.95</td>
<td>15.67</td>
<td>13.36</td>
</tr>
<tr>
<td></td>
<td>(0.54)</td>
<td>(0.33)</td>
<td>(0.28)</td>
</tr>
<tr>
<td>m. minutus</td>
<td>17.41</td>
<td>12.20</td>
<td>10.25</td>
</tr>
<tr>
<td></td>
<td>(0.87)</td>
<td>(0.78)</td>
<td>(0.30)</td>
</tr>
<tr>
<td>arson</td>
<td>5.93</td>
<td>5.75</td>
<td>5.68</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>murders</td>
<td>5.38</td>
<td>5.23</td>
<td>5.17</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>rapes</td>
<td>6.42</td>
<td>6.22</td>
<td>6.16</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>robberies</td>
<td>5.14</td>
<td>5.00</td>
<td>4.94</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>burglary</td>
<td>6.04</td>
<td>5.85</td>
<td>5.78</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>assault</td>
<td>6.65</td>
<td>6.35</td>
<td>6.30</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>larceny</td>
<td>5.83</td>
<td>5.65</td>
<td>5.58</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>auto theft</td>
<td>5.93</td>
<td>5.75</td>
<td>5.68</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.01)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>traffic</td>
<td>14.72</td>
<td>13.85</td>
<td>13.00</td>
</tr>
<tr>
<td></td>
<td>(1.11)</td>
<td>(0.24)</td>
<td>(1.01)</td>
</tr>
</tbody>
</table>
Species Habitat Model
Conclusion

- New family of density estimation models.
- Strong data-dependent learning guarantees.
- Good performance in practice.
- Generalization using arbitrary Bregman divergences.