Smoothing/Regularization Techniques for Probabilistic and Structured Classification

Mathieu Blondel

NTT, CS Laboratories, Kyoto, Japan

2018/6/08

Joint work with Vlad Niculae, Andre Martins, Claire Cardie, Arthur Mensch

Outline

- Background: structured prediction
- Regularized prediction functions
- A new family of loss functions
- Generalized entropies, sparsity and separation margins
- Applications and experimental results

Outline

- **Background: structured prediction**
- Regularized prediction functions
- A new family of loss functions
- Generalized entropies, sparsity and separation margins
- Applications and experimental results

Structured prediction

$\boxed{\textbf{Goal}\text{: predict } \boldsymbol{y} \in \mathcal{Y} \text{ from } \boldsymbol{x} \in \mathcal{X}}$

- Both $\mathcal{X}$ and $\mathcal{Y}$ may be complex structured spaces (sequences, permutations, etc)
- Assumption 1: a function $\boldsymbol{f}_W \colon \mathcal{X} \to \mathbb{R}^d$ is available. Converts $\boldsymbol{x}$ into $\boldsymbol{\theta} = \boldsymbol{f}_W(\boldsymbol{x})$ ("potentials" or "features")
- Assumption 2: $\boldsymbol{y} \in \mathcal{Y}$ can be represented as a d-dimensional binary vector, i.e., $\boldsymbol{y} \in \{0, 1\}^d$

Maximum a-posteriori (MAP) inference

- The inner product $\langle \boldsymbol{y}, \boldsymbol{\theta} \rangle$ can be thought as the affinity score between $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$
- Find the highest-scoring $\boldsymbol{y}$:

$$\boxed{\hat{\boldsymbol{y}} \in \mathrm{MAP}(\boldsymbol{\theta}) := \underset{\boldsymbol{y} \in \mathcal{Y}}{\operatorname{argmax}} \; \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle}$$

Corresponds to finding the mode of posterior distribution $p(\boldsymbol{y}|\boldsymbol{\theta}) \propto \exp\langle \boldsymbol{y}, \boldsymbol{\theta} \rangle$ (Gibbs distribution)

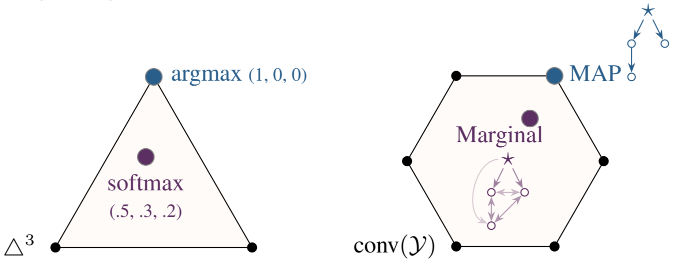
Combinatorial problem: $|\mathcal{Y}|$ potentially exponential in input size

Marginal polytope and marginal inference

- $\operatorname{conv}(\mathcal{Y}) := \{\mathbb{E}_{\boldsymbol{p}}[Y] : \boldsymbol{p} \in \triangle^{|\mathcal{Y}|}\}$ forms a convex polytope, called the marginal polytope [Wainwright & Jordan '08]
- Marginal inference consists in computing

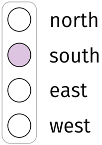
$$\boxed{\operatorname{marginals}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{p}}[Y] \in \operatorname{conv}(\mathcal{Y})}$$

where $p(\boldsymbol{y}; \boldsymbol{\theta}) \propto \exp\langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$ is the Gibbs distribution



Examples of structured inference

One-of-k classification

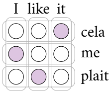


$$
\begin{array}{r} \\ \text{north} \\ \text{south} \\ \text{east} \\ \text{west} \end{array}
\begin{array}{c}
\begin{array}{cccc} y_{\mathrm{n}} & y_{\mathrm{s}} & y_{\mathrm{e}} & y_{\mathrm{w}} \end{array} \\
\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}
\end{array}
$$

MAP: $\operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle = \operatorname*{argmax}_{i \in [k]} \theta_i$

marginals: $\exp \boldsymbol{\theta} \Big/ \sum_{i=1}^{k} \theta_i$ (softmax)

Linear assignment

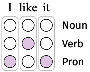


$$
\begin{array}{r|cccccc}
 & y_{123} & y_{132} & y_{213} & y_{231} & y_{312} & y_{321} \\
\hline
\text{I–cela} & 1 & 1 & 0 & 0 & 0 & 0 \\
\text{I–me} & 0 & 0 & 1 & 1 & 0 & 0 \\
\text{I–plait} & 0 & 0 & 0 & 0 & 1 & 1 \\
\hline
\text{like–cela} & 0 & 0 & 1 & 0 & 1 & 0 \\
\text{like–me} & 1 & 0 & 0 & 0 & 0 & 1 \\
\text{like–plait} & 0 & 1 & 0 & 1 & 0 & 0 \\
\hline
\text{it–cela} & 0 & 0 & 0 & 1 & 0 & 1 \\
\text{it–me} & 0 & 1 & 0 & 0 & 1 & 0 \\
\text{it–plait} & 1 & 0 & 1 & 0 & 0 & 0
\end{array}
$$

MAP: Hungarian algorithm

marginals: intractable [Valiant '79; Taskar '04]

Sequence prediction



$$
\begin{array}{r|ccccc}
 & y_{\mathrm{NNN}} & y_{\mathrm{NNV}} & & y_{\mathrm{PVP}} & \\
\hline
\text{I=N} & 1 & 1 & & 0 & \\
\text{I=V} & 0 & 0 & & 0 & \\
\text{I=P} & 0 & 0 & & 1 & \\
\hline
\text{like=N} & 1 & 1 & & 0 & \\
\text{like=V} & 0 & 0 & \cdots & 1 & \cdots \\
\text{like=P} & 0 & 0 & & 0 & \\
\hline
\text{it=N} & 1 & 0 & & 0 & \\
\text{it=V} & 0 & 1 & & 0 & \\
\text{it=P} & 0 & 0 & & 1 &
\end{array}
\qquad
\begin{array}{r|ccccc}
 & y_{\mathrm{NNN}} & y_{\mathrm{NNV}} & & y_{\mathrm{PVP}} & \\
\hline
\text{I,like=NN} & 1 & 1 & & 0 & \\
\text{I,like=NV} & 0 & 0 & & 0 & \\
\text{I,like=NP} & 0 & 0 & & 0 & \\
\text{I,like=VN} & 0 & 0 & & 0 & \\
\text{I,like=VV} & 0 & 0 & \cdots & 0 & \cdots \\
\text{I,like=VP} & 0 & 0 & & 0 & \\
\text{I,like=PN} & 0 & 0 & & 0 & \\
\text{I,like=PV} & 0 & 0 & & 1 & \\
\text{I,like=PP} & 0 & 0 & & 0 & \\
\hline
\text{like,it=NN} & 1 & 0 & & 0 & \\
 & & \cdots & & & \\
\text{like,it=VP} & 0 & 0 & & 1 & \\
 & & \cdots & & &
\end{array}
$$

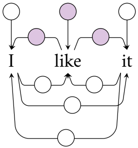
MAP: Viterbi algorithm

marginals: forward-backward algorithm

Image credit: Vlad Niculae (PhD thesis, to appear)

Examples of structured inference

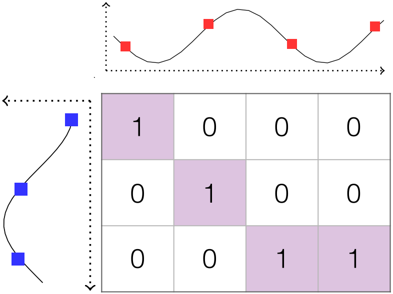
Dependency parsing



$$\begin{array}{r} \star\to\text{I} \\ \text{like}\to\text{I} \\ \text{it}\to\text{I} \\ \star\to\text{like} \\ \text{I}\to\text{like} \\ \text{it}\to\text{like} \\ \star\to\text{it} \\ \text{I}\to\text{it} \\ \text{like}\to\text{it} \end{array} \begin{bmatrix} 1 & & 0 & 0 & \\ 0 & & 1 & 1 & \\ 0 & & 0 & 0 & \\ \hline 0 & & 1 & 1 & \\ 1 & \dots & 0 & 0 & \dots \\ 0 & & 0 & 0 & \\ \hline 0 & & 0 & 0 & \\ 0 & & 1 & 0 & \\ 1 & & 0 & 1 & \end{bmatrix}$$

MAP: maximal arborescence algorithms
marginals: Koo et al '07, Smith & Smith '07

Time-series alignment



MAP: dynamic time warping (DTW)
marginals: soft-DTW [CB'17]

Relation between loss and inference

$$\boxed{\min_{W} \sum_{i=1}^{n} L(\boldsymbol{\theta}_i; \boldsymbol{y}_i) \quad \boldsymbol{\theta}_i \equiv \boldsymbol{f}_W(\boldsymbol{x}_i)}$$

- Structured SVM loss:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \max_{\boldsymbol{y}' \in \mathcal{Y}} \langle \boldsymbol{\theta}, \boldsymbol{y}' \rangle - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$$

Subgradient requires a call to MAP inference

- Conditional random field (CRF) loss:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \log \sum_{\boldsymbol{y}' \in \mathcal{Y}} \exp \langle \boldsymbol{\theta}, \boldsymbol{y}' \rangle - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$$

Gradient requires a call to marginal inference

Issues with MAP inference

- Can't deal with ambiguous ouputs

  MAP inference returns only one output: the highest-scoring one. For difficult cases, we may want to know other likely outputs.

- Lack of differentiability

$$\boldsymbol{x} \in \mathcal{X} \to \boxed{\boldsymbol{f_W}} \to \boldsymbol{\theta} \in \mathbb{R}^d \to \boxed{\text{MAP}} \to \widehat{\boldsymbol{y}} \in \mathcal{Y} \to \cdots$$

  Can't use MAP as layer in a neural net pipeline

Issues with marginal inference

- Every $\boldsymbol{y}$ gets non-zero probability since $p(\boldsymbol{y}; \boldsymbol{\theta}) \propto \exp \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$

  How to assign exactly zero probability to irrelevant $\boldsymbol{y}$?

- Intractable for some output spaces $\mathcal{Y}$

  Can we make inference differentiable and at the same time tractable for more output spaces?

We provide an answer based on convex duality
and smoothing / regularization!

Outline

- Background: structured prediction
- **Regularized prediction functions**
- A new family of loss functions
- Generalized entropies, sparsity and separation margins
- Applications and experimental results

Prediction function as a linear program

View a combinatorial problem as continuous optimization

$$\widehat{\boldsymbol{y}}(\boldsymbol{\theta}) \in \underset{\boldsymbol{y} \in \mathcal{Y}}{\operatorname{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle = \underset{\boldsymbol{y} \in \operatorname{conv}(\mathcal{Y})}{\operatorname{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$$

i.e., max of a linear form over a convex polytope

Note that when $\mathcal{Y} = \{\boldsymbol{e}_i\}_{i=1}^{d}$, $\operatorname{conv}(\mathcal{Y}) = \triangle^d$

Regularized prediction functions [NB'17,MB'18]

$$\boxed{\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \underset{\boldsymbol{\mu} \in \mathrm{conv}(\mathcal{Y})}{\mathrm{argmax}} \; \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \Omega(\boldsymbol{\mu})}$$

where Ω is a convex regularization function

$$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \boldsymbol{\mu}^\star = \mathbb{E}_{\boldsymbol{p}}[Y] \in \mathrm{conv}(\mathcal{Y})$$

for some, not necessarily unique, $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$

Relation with the convex conjugate

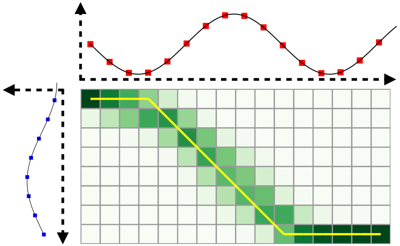
$$\boxed{\widehat{y}_\Omega(\boldsymbol{\theta}) \in \underset{\boldsymbol{\mu} \in \mathrm{dom}(\Omega)}{\mathrm{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \Omega(\boldsymbol{\mu})}$$

- $\Omega^*(\boldsymbol{\theta}) := \max_{\boldsymbol{\mu} \in \mathrm{dom}(\Omega)} \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \Omega(\boldsymbol{\mu}) = \langle \boldsymbol{\theta}, \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \rangle - \Omega(\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}))$
- $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) \in \partial \Omega^*(\boldsymbol{\theta})$ (from Danskin's theorem)
  - $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \nabla \Omega^*(\boldsymbol{\theta})$ if Ω is strictly convex

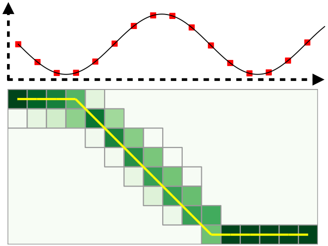
Benefit of regularization 1

Dealing with ambiguous predictions

Regularization moves $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ away from the vertices of the marginal polytope: $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$ = convex combination of $\boldsymbol{y} \in \mathcal{Y}$



entropic regularization (marginals)



quadratic regularization

Benefit of regularization 2

Smoothing effect

If Ω is strongly convex then

- Ω^* is smooth (differentiable with Lipschitz continuous gradient)
- $\widehat{\boldsymbol{y}}_\Omega = \nabla\Omega^*$ is differentiable almost everywhere

$$\boldsymbol{x} \in \mathcal{X} \to \boxed{\boldsymbol{f_W}} \to \boldsymbol{\theta} \in \mathbb{R}^d \to \boxed{\widehat{\boldsymbol{y}}_\Omega} \to \ldots$$

Differentiable pipeline, can be trained end-to-end using backpropagation!

Outline

- Background: structured prediction
- Regularized prediction functions
- **A new family of loss functions**
- Generalized entropies, sparsity and separation margins
- Applications and experimental results

Fenchel-Young losses

- Fenchel-Young loss generated by Ω [NMBC'17, BMN '18]

$$\boxed{L_{\Omega}(\boldsymbol{\theta}; \boldsymbol{y}) := \Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle}$$

where $\boldsymbol{\theta} \in \mathrm{dom}(\Omega^*) = \mathbb{R}^d$ and $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathrm{dom}(\Omega)$ is the ground-truth

- Grounded in the Fenchel-Young inequality

$$\Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{\mu}) \geq \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle \quad \forall \boldsymbol{\theta} \in \mathrm{dom}(\Omega^*), \boldsymbol{\mu} \in \mathrm{dom}(\Omega).$$

Properties of Fenchel-Young losses

$$\boxed{\mathrm{L}_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) := \Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{y}) - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle}$$

1. **Non-negativity:** $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) \geq 0$
2. **Zero loss:** $L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = 0 \Leftrightarrow \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \boldsymbol{y}$
3. **Convex** and **differentiable** in $\boldsymbol{\theta}$

Properties stated for strictly convex Ω for notational simplicity.

Learning with Fenchel-Young losses

$$\textbf{Primal:} \quad \min_{W} \sum_{i=1}^{n} L_\Omega(\boldsymbol{\theta}_i; \boldsymbol{y}_i) + G(W) \text{ s.t. } \boldsymbol{\theta}_i \equiv \boldsymbol{f}_W(\boldsymbol{x}_i)$$

Gradients: $\nabla_{\boldsymbol{\theta}} L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}$ ("residual vector")

If $\boldsymbol{f}_W(\boldsymbol{x}) = W\boldsymbol{x}$ then

$$\textbf{Dual:} \quad \max_{\beta} -D(\beta) \text{ s.t. } \beta_i \in \mathrm{dom}(\Omega) \; \forall i \in [n]$$

$$D(\beta) \coloneqq \sum_i \Omega(\beta_i) - \Omega(\boldsymbol{y}_i) + G^* \left(\sum_{i=1}^{n} (\boldsymbol{y}_i - \beta_i) \boldsymbol{x}_i^\top \right)$$

Learning with Fenchel-Young losses

$$\boxed{\textbf{Primal:} \quad \min_{W} \sum_{i=1}^{n} L_\Omega(\boldsymbol{\theta}_i; \boldsymbol{y}_i) + G(W) \text{ s.t. } \boldsymbol{\theta}_i \equiv \boldsymbol{f}_W(\boldsymbol{x}_i)}$$

Gradients: $\nabla_{\boldsymbol{\theta}} L_\Omega(\boldsymbol{\theta}; \boldsymbol{y}) = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) - \boldsymbol{y}$ ("residual vector")

If $\boldsymbol{f}_W(\boldsymbol{x}) = W\boldsymbol{x}$ then

$$\boxed{\textbf{Dual:} \quad \max_{\beta} -D(\beta) \text{ s.t. } \boldsymbol{\beta}_i \in \operatorname{dom}(\Omega)\ \forall i \in [n]}$$

$$D(\beta) := \sum_i \Omega(\boldsymbol{\beta}_i) - \Omega(\boldsymbol{y}_i) + G^* \left(\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\beta}_i) \boldsymbol{x}_i^\top \right)$$

Relation with Bregman divergences

- Bregman divergence generated by strictly convex Ω

$$B_\Omega(\boldsymbol{y}||\boldsymbol{\mu}) := \Omega(\boldsymbol{y}) - \Omega(\boldsymbol{\mu}) - \langle \nabla\Omega(\boldsymbol{\mu}), \boldsymbol{y} - \boldsymbol{\mu}\rangle$$

- Using $\boldsymbol{\theta} = \nabla\Omega(\boldsymbol{\mu})$ we get

$$\boxed{\mathrm{B}_\Omega(\boldsymbol{y}||\boldsymbol{\mu}) = L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})}$$

Proof uses that if Ω is a l.s.c. proper convex function, then

$$\Omega^*(\boldsymbol{\theta}) + \Omega(\boldsymbol{\mu}) = \langle \boldsymbol{\theta}, \boldsymbol{\mu}\rangle \Leftrightarrow \boldsymbol{\mu} = \nabla\Omega^*(\boldsymbol{\theta}) \Leftrightarrow \boldsymbol{\theta} = \nabla\Omega(\boldsymbol{\mu})$$

Relation with Bregman divergences

- Bregman divergences are defined in primal space

$$\boxed{\mathrm{B}_\Omega \colon \ \mathrm{dom}(\Omega) \times \mathrm{dom}(\Omega) \to \mathbb{R}_+}$$

- Fenchel-Young losses are defined in "mixed space"

$$\boxed{\mathrm{L}_\Omega \colon \ \mathrm{dom}(\Omega^*) \times \mathcal{Y} \subseteq \mathrm{dom}(\Omega) \to \mathbb{R}_+}$$

$B_\Omega(\boldsymbol{y}||\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})) = B_\Omega(\boldsymbol{y}||\nabla\Omega^*(\boldsymbol{\theta}))$ not necessarily convex!

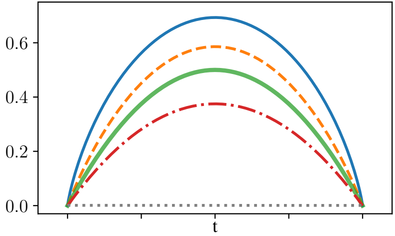
Tsallis α-entropies [Tsallis '88]

Choose $\mathrm{dom}(\Omega) = \triangle^{|\mathcal{Y}|}$ and $\Omega = -\mathrm{H}^{\mathrm{T}}_{\alpha}$

$$\boxed{\mathrm{H}^{\mathrm{T}}_{\alpha}(\boldsymbol{p}) \coloneqq \sum_{j=1}^{|\mathcal{Y}|} h_{\alpha}(p_j) \quad \text{with} \quad h_{\alpha}(t) \coloneqq \frac{t - t^{\alpha}}{\alpha(\alpha - 1)}}$$

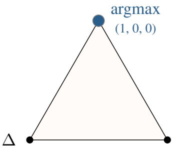
A parametric family of separable entropies

$\mathrm{H}^{\mathrm{T}}_{\alpha}([t, 1-t])$



Delta distribution, perceptron loss

$$\boxed{\Omega(\boldsymbol{p}) = -\mathsf{H}^{\mathrm{T}}_{\infty}(\boldsymbol{p}) = 0}$$

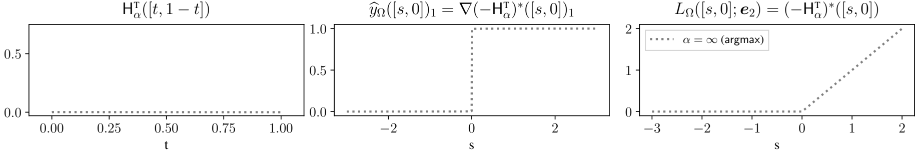


“delta” distribution

$$\boxed{\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta}) \in \operatorname*{argmax}_{\boldsymbol{y} \in \{\boldsymbol{e}_i\}} \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle}$$

perceptron loss

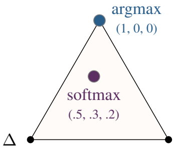
$$\boxed{\mathsf{L}_{\Omega}(\boldsymbol{\theta}; \boldsymbol{e}_j) = \max_{i \in [k]} \theta_i - \theta_j}$$



Softmax distribution, logistic loss

negative Shannon entropy

$$\boxed{\Omega(\boldsymbol{p}) = -\mathsf{H}^{\mathrm{T}}_1(\boldsymbol{p}) = \sum_i p_i \log p_i}$$

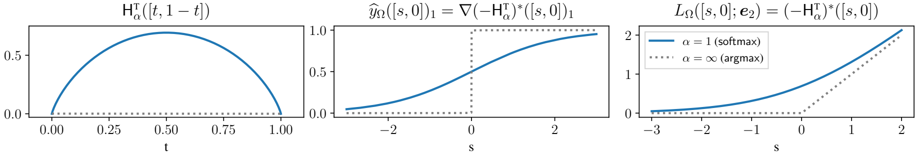


softmax

$$\boxed{\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \frac{\exp \boldsymbol{\theta}}{\sum_{i=1}^k \exp \theta_i}}$$

logistic loss

$$\boxed{\mathsf{L}_\Omega(\boldsymbol{\theta}; \boldsymbol{e}_j) = \log \sum_{i \in [k]} \exp \theta_i - \theta_j}$$



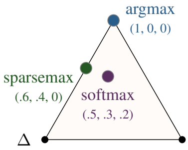
sparsemax distribution, loss [Martins & Astudillo '16]

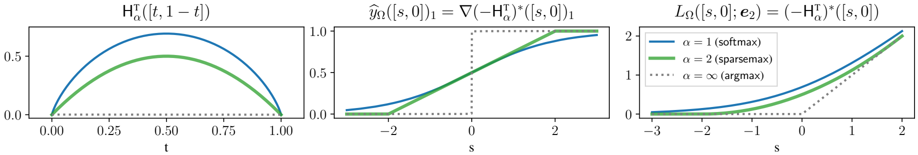
negative Gini index [Gini 1912]

$$\boxed{\Omega(\boldsymbol{p}) = -\mathsf{H}^{\mathrm{T}}_{2}(\boldsymbol{p}) = \frac{1}{2}\sum_{i} p_i(p_i - 1) = \frac{1}{2}\|\boldsymbol{p}\|^2 - \frac{1}{2}}$$

projection onto the simplex / sparsemax

$$\boxed{\widehat{\boldsymbol{y}}_{\Omega}(\boldsymbol{\theta}) = \underset{\boldsymbol{p}\in\triangle^k}{\operatorname{argmin}} \|\boldsymbol{p} - \boldsymbol{\theta}\|^2}$$





CRFs and structured sparsemax

Choose $\mathrm{dom}(\Omega) = \mathrm{conv}(\mathcal{Y})$

- Conditional Random Fields: maximum entropy principle

$$-\Omega(\boldsymbol{\mu}) = \max_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \mathrm{H}^{\mathrm{S}}(\boldsymbol{p}) \text{ s.t. } \mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}$$

Then $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta}) = \mathrm{marginals}(\boldsymbol{\theta})$; tractable for some $\mathcal{Y}$

- Structured sparsemax: minimum norm

$$\Omega(\boldsymbol{\mu}) = \min_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \|\boldsymbol{p}\|^2 \text{ s.t. } \mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}$$

Computing $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) =: \text{sparsemax-mean}(\boldsymbol{\theta})$ likely intractable for structured $\mathcal{Y}$

CRFs and structured sparsemax

Choose $\mathrm{dom}(\Omega) = \mathrm{conv}(\mathcal{Y})$

- Conditional Random Fields: maximum entropy principle

$$-\Omega(\boldsymbol{\mu}) = \max_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \mathsf{H}^{\mathrm{S}}(\boldsymbol{p}) \text{ s.t. } \mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}$$

Then $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \nabla\Omega^*(\boldsymbol{\theta}) = \mathrm{marginals}(\boldsymbol{\theta})$; tractable for some $\mathcal{Y}$

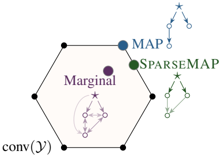
- Structured sparsemax: minimum norm

$$\Omega(\boldsymbol{\mu}) = \min_{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}} \|\boldsymbol{p}\|^2 \text{ s.t. } \mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}$$

Computing $\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) =: \mathrm{sparsemax\text{-}mean}(\boldsymbol{\theta})$ likely intractable for structured $\mathcal{Y}$

sparseMAP: mean space regularization [NMBC '18]

$$\boxed{\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta}) = \underset{\boldsymbol{\mu} \in \operatorname{conv}(\mathcal{Y}) \subseteq \mathbb{R}^d}{\operatorname{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \|\boldsymbol{\mu}\|^2}$$



- $\widehat{\boldsymbol{y}}_\Omega$ can be computed using the conditional gradient algorithm (a.k.a. Frank-Wolfe)
- Main ingredient is the linear (min|max)imization oracle

$$\underset{\boldsymbol{y} \in \mathcal{Y}}{\operatorname{argmax}} \ \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle = \operatorname{MAP}(\boldsymbol{\theta})$$

- FW returns both $\boldsymbol{\mu}^\star$ and one possible $\boldsymbol{p}$ s.t. $\mathbb{E}_{\boldsymbol{p}}[Y] = \boldsymbol{\mu}^\star$

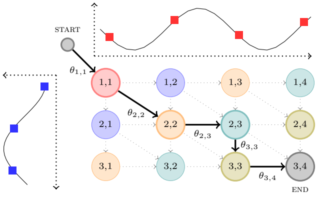
Smoothed dynamic programming [CB' 17, MB '18]

- When $\mathcal{Y}$ can be represented as a DAG, MAP inference can be computed by dynamic programming

- Key idea: Smooth the max/min operator within Bellman's recursion

- Entropic regul: marginals$(\boldsymbol{\theta}) = \nabla \mathrm{DP}_\Omega(\boldsymbol{\theta}) \in \mathrm{conv}(\mathcal{Y})$

- Quadratic regul: sparsemax-mean$(\boldsymbol{\theta}) \approx \nabla \mathrm{DP}_\Omega(\boldsymbol{\theta}) \in \mathrm{conv}(\mathcal{Y})$



$$\langle \boldsymbol{Y}, \boldsymbol{\theta} \rangle = \theta_{1,1} + \theta_{2,2} + \theta_{2,3} + \theta_{3,3} + \theta_{3,4}$$

- initialize v at edge cases
- for all (i,j) in topological order:
  $v_{i,j} = \theta_{i,j} + \mathrm{softmin}_\Omega\{v_{i-1,j}, v_{i,j-1}, v_{i-1,j-1}\}$
- Output: $\mathrm{DP}_\Omega(\boldsymbol{\theta}) := v_{m,n}(\boldsymbol{\theta})$ (convex in $\boldsymbol{\theta}$!)

Backpropagating through $\widehat{\boldsymbol{y}}_\Omega$

$$\boldsymbol{x} \in \mathcal{X} \to \boxed{\boldsymbol{f_W}} \to \boldsymbol{\theta} \in \mathbb{R}^d \to \boxed{\widehat{\boldsymbol{y}}_\Omega} \to \dots$$

- Since $\widehat{\boldsymbol{y}}_\Omega = \nabla\Omega^*$, backpropagating through $\widehat{\boldsymbol{y}}_\Omega$ requires multipications with the Hessian: $\nabla^2\Omega^*(\boldsymbol{\theta})\boldsymbol{z}$ for some $\boldsymbol{z}$
- Can be computed from the CG/FW solution by solving a linear system derived from the KKT conditions [NMBC '18]
- Another way is to backpropagate through the directional derivative at $\boldsymbol{\theta}$ along $\boldsymbol{z}$ [Pearlmutter '94, MB '18]

$$\nabla^2 \mathrm{DP}_\Omega(\boldsymbol{\theta})\boldsymbol{z} = \nabla\langle \nabla \mathrm{DP}_\Omega(\boldsymbol{\theta}), \boldsymbol{z}\rangle$$

Summary of losses recovered

	$\mathsf{dom}(\Omega)$	$\Omega(\boldsymbol{\mu})$	$\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})$	$L_\Omega(\boldsymbol{\theta}; \boldsymbol{y})$
Squared loss	$\mathbb{R}^{\|\mathcal{Y}\|}$	$\frac{1}{2}\|\boldsymbol{\mu}\|^2$	$\boldsymbol{\theta}$	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|^2$
Perceptron loss	$\triangle^{\|\mathcal{Y}\|}$	0	$\mathsf{argmax}(\boldsymbol{\theta})$	$\mathsf{max}_i\, \theta_i - \theta_k$
Logistic loss	$\triangle^{\|\mathcal{Y}\|}$	$-\mathsf{H}^{\mathrm{S}}(\boldsymbol{\mu})$	$\mathsf{softmax}(\boldsymbol{\theta})$	$\log \sum_i \exp \theta_i - \theta_k$
Sparsemax loss	$\triangle^{\|\mathcal{Y}\|}$	$\frac{1}{2}\|\boldsymbol{\mu}\|^2$	$\mathsf{sparsemax}(\boldsymbol{\theta})$	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|^2 - \frac{1}{2}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})-\boldsymbol{\theta}\|^2$
Struct. perceptron	$\mathsf{conv}(\mathcal{Y})$	0	$\mathsf{MAP}(\boldsymbol{\theta})$	$\mathsf{max}_{\boldsymbol{y}'}\, \langle \boldsymbol{\theta}, \boldsymbol{y}' \rangle - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$
CRF	$\mathsf{conv}(\mathcal{Y})$	$\min\limits_{\mathbb{E}_{\boldsymbol{p}}[Y]=\boldsymbol{\mu}} -\mathsf{H}^{\mathrm{S}}(\boldsymbol{p})$	$\mathsf{marginals}(\boldsymbol{\theta})$	$\log \sum_{\boldsymbol{y}'} \exp \langle \boldsymbol{\theta}, \boldsymbol{y}' \rangle - \langle \boldsymbol{\theta}, \boldsymbol{y} \rangle$
Struct. sparsemax	$\mathsf{conv}(\mathcal{Y})$	$\min\limits_{\mathbb{E}_{\boldsymbol{p}}[Y]=\boldsymbol{\mu}} \|\boldsymbol{p}\|^2$	intractable*	intractable*
SparseMAP	$\mathsf{conv}(\mathcal{Y})$	$\frac{1}{2}\|\boldsymbol{\mu}\|^2$	$\mathsf{sparseMAP}(\boldsymbol{\theta})$	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{\theta}\|^2 - \frac{1}{2}\|\widehat{\boldsymbol{y}}_\Omega(\boldsymbol{\theta})-\boldsymbol{\theta}\|^2$

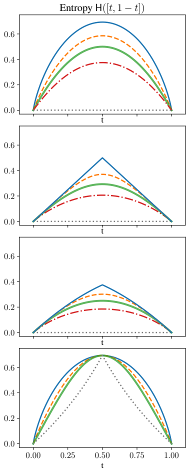
* Can be approximated by smoothed dynamic programming [MB '18]

Outline

- Background: structured prediction
- Regularized prediction functions
- A new family of loss functions
- **Generalized entropies, sparsity and separation margins**
- Applications and experimental results

Generalized entropies [DeGroot '62, Grunwald & Dawid '04]

Use a concave function $\mathrm{H}(\boldsymbol{p})$ to measure the "uncertainty" in $\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}$



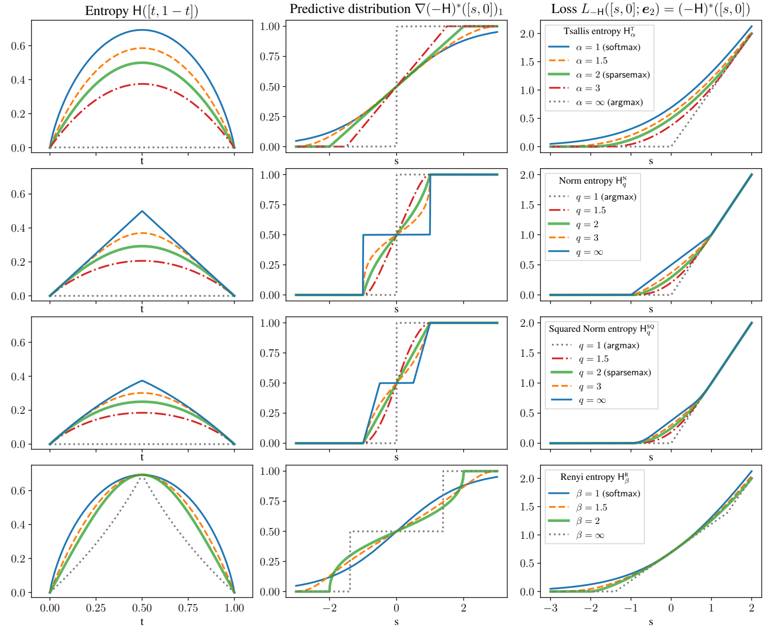
Tsallis: $\mathrm{H}^{\mathrm{T}}_{\alpha}(\boldsymbol{p}) := \frac{1}{\alpha(\alpha-1)} \sum_{j=1}^{|\mathcal{Y}|} p_j - p_j^{\alpha}$

***q*-Norm:** $\mathrm{H}^{\mathrm{N}}_{q}(\boldsymbol{p}) := 1 - \|\boldsymbol{p}\|_q$

Squared *q*-Norm: $\mathrm{H}^{\mathrm{SQ}}_{q}(\boldsymbol{p}) := \frac{1}{2}(1 - \|\boldsymbol{p}\|_q^2)$

Rényi: $\mathrm{H}^{\mathrm{R}}_{\beta}(\boldsymbol{p}) := \frac{1}{1-\beta} \log \sum_{j=1}^{|\mathcal{Y}|} p_j^{\beta}.$

A wealth of new loss and prediction functions [BMN '18]



Properties of generalized entropies

- **Assumption 1:** $\mathrm{H}(\boldsymbol{p}) = 0$ if $\boldsymbol{p} \in \{\boldsymbol{e}_i\}$
- **Assumption 2:** H is strictly concave over $\mathrm{dom}(\Omega) = \triangle^{|\mathcal{Y}|}$
- **Assumption 3:** $\mathrm{H}(P\boldsymbol{p})$ for any permutation matrix P

$$\Downarrow$$

- **Non-negativity:** $H(\boldsymbol{p}) \geq 0$
- **Maximum:** $\underset{\boldsymbol{p} \in \triangle^{|\mathcal{Y}|}}{\mathrm{argmax}}\, H(\boldsymbol{p}) = \dfrac{\mathbf{1}}{|\mathcal{Y}|}$
- **Order-preservingness:** If $\boldsymbol{p} = \widehat{\boldsymbol{y}}_\Omega(\boldsymbol{s}) = \nabla(-\mathrm{H})^*(\boldsymbol{s})$ then

$$s_i > s_j \Rightarrow p_i \geq p_j$$

Condition for sparse prediction function

When is $\hat{\boldsymbol{y}}_\Omega = \nabla(-\mathrm{H})^*$ sparse?

Under assumptions 1 to 3:

$$\boxed{\forall \boldsymbol{p} \in \triangle^{|\mathcal{Y}|} : \partial(-\mathrm{H})(\boldsymbol{p}) \neq \varnothing \Leftrightarrow \nabla(-\mathrm{H})^*(\mathbb{R}^{|\mathcal{Y}|}) = \triangle^{|\mathcal{Y}|}}$$

i.e., $\nabla(-\mathrm{H})^*$ covers the full simplex

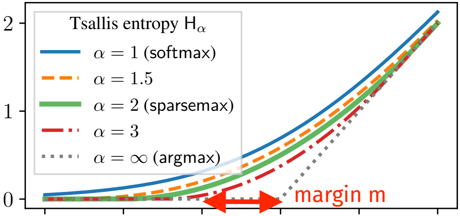
Functions whose gradient "explode" at the boundary (e.g., Shannon entropy) are called "essentially smooth". For those functions, $\nabla(-\mathrm{H})^*$ maps only to the relative interior of $\triangle^{|\mathcal{Y}|}$.

Separation margin of a loss

A loss $L(\boldsymbol{s}; \boldsymbol{y})$ over $\mathbb{R}^{|\mathcal{Y}|} \times \{\boldsymbol{e}_i\}_{i=1}^{|\mathcal{Y}|}$, where $\boldsymbol{y} = \boldsymbol{e}_k$ is the ground truth, has a separation margin $m > 0$ if

$$\boxed{s_k \geq m + \max_{j \neq k} s_j \quad \Rightarrow \quad L(\boldsymbol{s}; \boldsymbol{y}) = 0}$$

We denote the smallest such m by $\mathrm{margin}(L)$.



Condition for separation margin and value

$$L_{-\mathsf{H}}(\boldsymbol{s}; \boldsymbol{e}_k) \text{ has a separation margin } m$$
$$\Updownarrow$$
$$m\boldsymbol{e}_k \in \partial(-\mathsf{H})(\boldsymbol{e}_k)$$

Tight link between margins and sparse prediction functions!

For twice differentiale H:

$$\mathrm{margin}(L_{-\mathsf{H}}) = \nabla_j \mathsf{H}(\boldsymbol{e}_k) - \nabla_k \mathsf{H}(\boldsymbol{e}_k).$$

For separable entropies $\mathsf{H} = \sum_j h(p_j)$:

$$\mathrm{margin}(L_{-\mathsf{H}}) = h'(0) - h'(1)$$

Outline

- Background: structured prediction
- Regularized prediction functions
- A new family of loss functions
- Generalized entropies, sparsity and separation margins
- **Applications and experimental results**

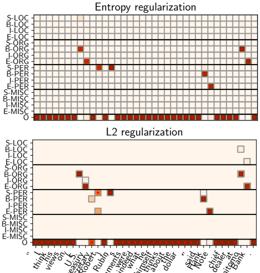
Named Entity Recognition [MB '18]

- Identify blocks of words corresponding to names, locations, etc
- Pipeline

$$\text{sentence } \boldsymbol{x} \in \mathcal{X} \rightarrow \boxed{\text{bi-LSTM}} \rightarrow \boldsymbol{\theta} \in \mathbb{R}^d \rightarrow \boxed{\mathrm{L}_\Omega} \rightarrow \mathbb{R}_+$$

$$\text{sentence } \boldsymbol{x} \in \mathcal{X} \rightarrow \boxed{\text{bi-LSTM}} \rightarrow \boldsymbol{\theta} \in \mathbb{R}^d \rightarrow \boxed{\widehat{\boldsymbol{y}}_\Omega} \rightarrow \underset{\substack{\uparrow \\ \boldsymbol{y}}}{\boxed{\Delta(\cdot,\cdot)}} \rightarrow \mathbb{R}+$$

- Results on CoNLL 2013 shared task:



Ω	Loss	English	Spanish	German	Dutch
Negentropy	Surrogate	90.80	**86.68**	77.35	**87.56**
	Relaxed	90.47	86.20	**77.56**	87.37
ℓ_2^2	Surrogate	**90.86**	85.51	76.01	86.58
	Relaxed	89.49	84.07	76.91	85.90
(Lample et al., 2016)		*90.96*	*85.75*	*78.76*	*81.74*

Machine Translation with Attention [MB '18]

- Translate source language into target language

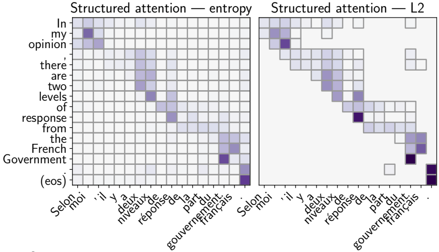
- RNN pipeline: decoding step for outputting the next word

$$\text{encoding } \boldsymbol{x} \rightarrow \boxed{\text{scoring}} \rightarrow \boldsymbol{\theta} \rightarrow \boxed{\widehat{\boldsymbol{y}}_\Omega} \rightarrow \text{ attention weights}$$

$$\uparrow$$

$$\text{RNN decoder state } \boldsymbol{z}$$

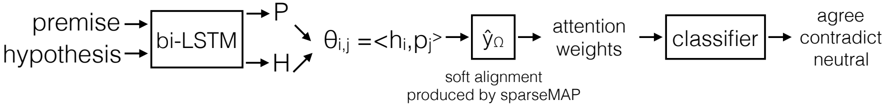
- ℓ_2^2 reg achieves similar accuracy with more interpretable maps



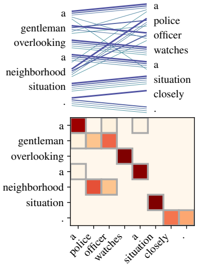
Attention model	WMT14 1M fr→en	WMT14 en→fr
Softmax	**27.96**	**28.08**
Entropy regularization	**27.96**	27.98
ℓ_2^2 reg.	27.21	27.28

Natural Language Inference [NMBC '18]

- Infer whether two sentence agree, contradict, are neutral
- Pipeline:



- Results on the SNLI and multi-SNLI dataset



Accuracy scores and percentage of non-aligned pairs

ESIM variant	MultiNLI	SNLI
softmax	76.05 (100%)	86.52 (100%)
sequential	75.54 (13%)	**86.62** (19%)
matching	**76.13** (8%)	86.05 (15%)

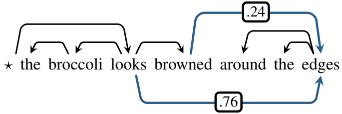
Dependency parsing [NMBC '18]

- Predict the directed tree of grammatical dependencies between words in a sentence

- Pipeline:

$$\text{sentence } \boldsymbol{x} \in \mathcal{X} \to \boxed{\text{bi-LSTM}} \to \boldsymbol{\theta} \in \mathbb{R}^d \to \overset{\overset{\boldsymbol{y}}{\downarrow}}{\boxed{\mathsf{L}_\Omega}} \to \mathbb{R}_+$$

- Results on Universal Dependency data (CoNLL 2017 shared task)



Loss	en	zh	vi	ro	ja
Structured SVM	87.02	81.94	69.42	87.58	**96.24**
CRF	86.74	83.18	69.10	87.13	96.09
SparseMAP	86.90	**84.03**	69.71	87.35	96.04
m-SparseMAP	**87.34**	82.63	**70.87**	**87.63**	96.03
UDPipe baseline	87.68	82.14	69.63	87.36	95.94

Conclusion

- Regularization / smoothing allows to deal with ambiguous outputs and brings differentiability
- FY losses allow to learn such regularized prediction functions and unify a wealth of existing losses
- Link between sparsity of $\widehat{\boldsymbol{y}}_\Omega = \nabla\Omega^*$, sparsity of dual variables and margin of L_Ω
- FY losses support arbitrary $\mathrm{dom}(\Omega)$, allowing a wide variety of (unexplored) applications

References

- Blondel et al. Learning Classifiers with Fenchel-Young Losses: Generalized Entropies, Margins, and Algorithms. arXiv preprint, 2018.
- Cuturi & Blondel. Soft-DTW: a differentiable loss function for time-series. ICML, 2017.
- DeGroot. Uncertainty, information, and sequential experiments. The Annals of Mathematical Statistics, 1962.
- Grunwald & Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. Annals of Statistics, 2004.
- Koo et al. Structured prediction models via the matrix-tree theorem. EMNLP, 2007.
- Martins & Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. ICML, 2016.
- Mensch & Blondel. Differentiable Dynamic Programming for Structured Prediction and Attention. ICML, 2018.
- Niculae & Blondel. A regularized framework for sparse and structured neural attention. NIPS, 2017.

References

- Niculae et al. SparseMAP: Differentiable Sparse Structured Inference. ICML 2018.
- Pearlmutter. Fast exact multiplication by the Hessian. Neural computation, 1994.
- Smith & Smith. Probabilistic models of nonprojective dependency trees. EMNLP, 2007
- Taskar et al. Max-Margin Markov Networks, NIPS 2003.
- Tsallis. Possible generalization of Boltzmann-Gibbs statistics. Journal of Statistical Physics, 1988.
- Valiant. The complexity of computing the permanent. Theor. Comput. Sci., 1979.
- Wainwright & Jordan. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 2008.