Lecture 10: Structured Output SVM

Vojtěch Franc

May 3, 2015

8.B: Structured output SVM

- Convex surrogates of the empirical risk
- Risk surrogates for the max-sum classifier

XEP33SML – Structured Model Learning, Summer 2015

8.B: Structured Output Support Vector Machines

- Given training examples $\mathcal{T} = \{(\boldsymbol{x}^i, \boldsymbol{y}^i) \in \mathcal{X} \times \mathcal{Y} \mid i \in \mathcal{I}\}$, the goal is to learn parameters $\boldsymbol{w} \in \mathbb{R}^n$ of a general linear classifier

$$h(\boldsymbol{x}; \boldsymbol{w}) = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \left\langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y}) \right\rangle$$

  where $\boldsymbol{\Psi} \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^n$ is fixed feature map.
- Regularized empirical risk minimization based learning leads to solving

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n} \left(\lambda \underbrace{\Omega(\boldsymbol{w})}_{\text{regularizer}} + \underbrace{\hat{R}_{\mathcal{T}}(\boldsymbol{w})}_{\substack{\text{surrogate of} \\ \text{empirical risk}}} \right)$$

  where $\Omega \colon \mathbb{R}^n \to \mathbb{R}$ is a (convex) regularizer and $\hat{R}_{\mathcal{T}} \colon \mathbb{R}^n \to \mathbb{R}$ is a surrogate of the empirical risk

$$R_{\mathcal{T}}(h(\cdot; \boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \ell(\boldsymbol{y}^i, h(\boldsymbol{x}^i; \boldsymbol{w}))$$

  and $\ell \colon \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ such that $\ell(\boldsymbol{y}, \boldsymbol{y}') = 0$ iff $\boldsymbol{y} = \boldsymbol{y}'$ is the loss.

Questions:

- How to construct the surrogate $\hat{R}_{\mathcal{T}}$ for a generic linear classifier and loss ?
- How to solve the risk minimization problem ?

8.B: Slack-rescaling loss

A loss of the linear classifier $h(\boldsymbol{x}; \boldsymbol{w})$ on $(\boldsymbol{x}^i, \boldsymbol{y}^i)$ can be upper-bounded by a convex function:

$$\begin{aligned}
\ell(\boldsymbol{y}^i, h(\boldsymbol{x}^i, \boldsymbol{w})) &= \max_{\boldsymbol{y} \in \mathcal{Y}} \ell(\boldsymbol{y}^i, \boldsymbol{y}) [\![\max_{\boldsymbol{y}' \in \mathcal{Y}} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}') \rangle - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \leq 0]\!] \\
&\leq \max_{\boldsymbol{y} \in \mathcal{Y}} \ell(\boldsymbol{y}^i, \boldsymbol{y}) [\![\langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \leq 0]\!] \\
&\leq \max_{\boldsymbol{y} \in \mathcal{Y}} \ell(\boldsymbol{y}^i, \boldsymbol{y}) \max\{0, 1 - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle\} \\
&= \max_{\boldsymbol{y} \in \mathcal{Y}} \ell(\boldsymbol{y}^i, \boldsymbol{y}) \Big[1 - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \Big] \\
&= \hat{\ell}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) \, .
\end{aligned}$$

Remark: The slack-rescaling loss is non-zero if the margin of incorrect label $\gamma = \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle$ is less than 1.

Learning: RRM leads to a convex problem $\boldsymbol{w}^* = \operatorname{argmin}_{\boldsymbol{w} \in \mathbb{R}^n} \Big[\lambda \Omega(\boldsymbol{w}) + \hat{R}_{\mathcal{T}}(\boldsymbol{w}) \Big]$ where

$$\hat{R}_{\mathcal{T}}(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \hat{\ell}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \max_{\boldsymbol{y} \in \mathcal{Y}} \ell(\boldsymbol{y}^i, \boldsymbol{y}) \Big[1 - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \Big]$$

8.B: Slack-rescaling loss

RRM with the slack-rescaling loss is scaling invariant. Let $\ell^{sc} = K\ell$, $K > 0$ be a scaled version of the loss ℓ. Then

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}} \left[\lambda\Omega(\boldsymbol{w}) + \sum_{i=1}^{m} \hat{\ell}(\boldsymbol{x}^i, \boldsymbol{y}^i; \boldsymbol{w})\right] = \operatorname*{argmin}_{\boldsymbol{w}} \left[\frac{\lambda}{K}\Omega(\boldsymbol{w}) + \sum_{i=1}^{m} \hat{\ell}^{sc}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})\right]$$

First order oracle: compute the value $\hat{\ell}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})$ and its sub-gradient $\hat{\ell}'(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})$ at $\boldsymbol{w}$

1. Solve $\hat{\boldsymbol{y}}^i = \operatorname{argmax}_{\boldsymbol{y}\in\mathcal{Y}} \ell(\boldsymbol{y}^i, \boldsymbol{y})\Big[1 - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i)\rangle + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y})\rangle\Big]$

2. Compute sub-gradient $\hat{\ell}'(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) = \ell(\boldsymbol{y}^i, \hat{\boldsymbol{y}}^i)\Big[\boldsymbol{\Psi}(\boldsymbol{x}^i, \hat{\boldsymbol{y}}^i) - \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i)\Big]$.

Remarks:

- An efficient oracle can be constructed in special cases, e.g. when $|\mathcal{Y}|$ is small or the loss is simple like the 0/1-loss $\ell(\boldsymbol{y}, \boldsymbol{y}') = [\![\boldsymbol{y} \neq \boldsymbol{y}']\!]$.
- It is remains an open problem how to do it for a general loss.

8.B: Margin-rescaling loss

A loss of the linear classifier $h(\boldsymbol{x}; \boldsymbol{w})$ on $(\boldsymbol{x}^i, \boldsymbol{y}^i)$ can be upper-bounded by a convex function:

$$\begin{aligned}
\ell(\boldsymbol{y}^i, h(\boldsymbol{x}^i, \boldsymbol{w})) &= \max_{\boldsymbol{y}\in\mathcal{Y}} \ell(\boldsymbol{y}^i, \boldsymbol{y}) [\![\max_{\boldsymbol{y}'\in\mathcal{Y}} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}') \rangle - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \leq 0]\!] \\
&\leq \max_{\boldsymbol{y}\in\mathcal{Y}} \ell(\boldsymbol{y}^i, \boldsymbol{y}) [\![\langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \leq 0]\!] \\
&\leq \max_{\boldsymbol{y}\in\mathcal{Y}} \max \left\{ 0, \ell(\boldsymbol{y}^i, \boldsymbol{y}) - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle + \langle \boldsymbol{w}, \Psi(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \right\} \\
&= \max_{\boldsymbol{y}\in\mathcal{Y}} \left[\ell(\boldsymbol{y}^i, \boldsymbol{y}) - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \right] \\
&= \hat{\ell}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})
\end{aligned}$$

Remark: The margin-rescaling loss is non-zero if the margin of incorrect label $\gamma = \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle$ is less than $\ell(\boldsymbol{y}^i, \boldsymbol{y})$.

Learning: RRM leads to a convex problem $\boldsymbol{w}^* = \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left[\lambda\Omega(\boldsymbol{w}) + \hat{R}_{\mathcal{T}}(\boldsymbol{w}) \right]$ where

$$\hat{R}_{\mathcal{T}}(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \hat{\ell}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \max_{\boldsymbol{y}\in\mathcal{Y}} \left[\ell(\boldsymbol{y}^i, \boldsymbol{y}) - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \right]$$

8.B: Margin-rescaling loss

The margin-rescaling loss

$$\hat{\ell}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) = \max_{\boldsymbol{y} \in \mathcal{Y}} \Big[\ell(\boldsymbol{y}^i, \boldsymbol{y}) - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) \rangle + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \Big]$$

Note that the slack-rescaling and the margin-rescaling surrogates coincide for the 0/1-loss $\ell(\boldsymbol{y}, \boldsymbol{y}') = [\![\boldsymbol{y} \neq \boldsymbol{y}']\!]$

First order oracle: compute the value $\hat{\ell}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})$ and its sub-gradient $\hat{\ell}'(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})$ at $\boldsymbol{w}$

1. Solve augmented classification problem (ACP)

$$\hat{\boldsymbol{y}}^i \in \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} \Big[\ell(\boldsymbol{y}^i, \boldsymbol{y}) + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}) \rangle \Big]$$

2. Compute sub-gradient $\hat{\ell}'(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) = \boldsymbol{\Psi}(\boldsymbol{x}^i, \hat{\boldsymbol{y}}^i) - \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i)$.

Notice the analogy with Perceptron algorithm: in order to learn you have to be able to classify.

8.B: Margin-rescaling loss for max-sum classifier

Consider the max-sum classifier

$$\hat{\boldsymbol{y}} \in \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} \big\langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\big\rangle = \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} \left\langle \boldsymbol{w}, \sum_{v\in\mathcal{V}} \boldsymbol{\Psi}_v(\boldsymbol{x}, y_v) + \sum_{\{v,v'\}\in\mathcal{E}} \boldsymbol{\Psi}_{vv'}(\boldsymbol{x}, y_v, y_{v'})\right\rangle$$

and an additive loss $\ell(\boldsymbol{y},\boldsymbol{y}') = \sum_{v\in\mathcal{V}} \ell_v(y_v, y'_{v'})$ where $\ell_v \colon \mathcal{Y} \to [0,\infty)$, $v \in \mathcal{V}$
(e.g. the Hamming loss)

Then the augmented classification problem boils down to solving

$$\begin{aligned}
\hat{\boldsymbol{y}}^i &\in \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} \Big[\ell(\boldsymbol{y}^i,\boldsymbol{y}) + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i,\boldsymbol{y})\rangle\Big] \\
&= \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} \Big[\sum_{v\in\mathcal{V}} \Big(\ell_v(y^i_v, y_v) + \langle \boldsymbol{w}, \boldsymbol{\Psi}_v(\boldsymbol{x}^i, y_v)\rangle\Big) + \sum_{\{v,v'\}\in\mathcal{E}} \langle \boldsymbol{w}, \boldsymbol{\Psi}_{vv'}(\boldsymbol{x}^i, y_v, y_{v'})\rangle\Big]
\end{aligned}$$

The ACP is tractable if the graph $(\mathcal{V},\mathcal{E})$ is acyclic.

8.B: Margin-rescaling loss for sub-modular max-sum classifier

Consider a class of the max-sum classifier

$$\hat{\boldsymbol{y}} \in \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y})\rangle = \operatorname*{argmax}_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} \left\langle \boldsymbol{w}, \sum_{v\in\mathcal{V}} \boldsymbol{\Psi}_v(\boldsymbol{x}, y_v) + \sum_{\{v,v'\}\in\mathcal{E}} \boldsymbol{\Psi}_{vv'}(y_v, y_{v'}) \right\rangle$$

where $\boldsymbol{w} \in \mathcal{W} \subset \mathbb{R}^n$ such that $-g_{vv'}(y, y') = \langle \boldsymbol{w}, \boldsymbol{\Psi}_{vv'}(y, y')\rangle$ is sub-modular w.r.t. an ordering of $\mathcal{Y} = \{1, \ldots, K\}$.

Learning: RRM with additive loss $\ell(\boldsymbol{y}, \boldsymbol{y}') = \sum_{v\in\mathcal{V}} \ell_v(y_v, y'_{v'})$ leads to a constrained convex problem

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left[\lambda\Omega(\boldsymbol{w}) + \sum_{i=1}^{m} \hat{\ell}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) \right]$$

subject to

$$\left\langle \boldsymbol{w}, \boldsymbol{\Psi}_{vv'}(y, y') + \boldsymbol{\Psi}_{vv'}(y+1, y'+1) - \boldsymbol{\Psi}_{vv'}(y, y'+1) - \boldsymbol{\Psi}(y+1, y') \right\rangle \geq 0,$$

$$\{v, v'\} \in \mathcal{E}, y, y' \in \{1, \ldots, K-1\}$$

Provided the solver maintains the intermediate solution $\boldsymbol{w}$ feasible the max-sum problems associated with the ACP are sub-modular and thus tractable.

8.B: LP relaxation of margin-rescaling loss

The ACP associated to a general max-sum classifier and additive loss

$$\max_{\boldsymbol{y}\in\mathcal{Y}} \Big[\ell(\boldsymbol{y}^i,\boldsymbol{y}) + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i,\boldsymbol{y})\rangle\Big]$$
$$= \max_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} \underbrace{\Bigg[\sum_{v\in\mathcal{V}} \Big(\ell_v(y^i_v, y_v) + \langle \boldsymbol{w}, \boldsymbol{\Psi}_v(\boldsymbol{x}^i, y_v)\rangle\Big) + \sum_{\{v,v'\}\in\mathcal{E}} \langle \boldsymbol{w}, \boldsymbol{\Psi}_{vv'}(\boldsymbol{x}^i, y_v, y_{v'})\rangle\Bigg]}_{f(\boldsymbol{y};\boldsymbol{x}^i,\boldsymbol{y}^i,\boldsymbol{w})}$$

can be upper bounded via the LP relaxation:

$$\max_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} f(\boldsymbol{y};\boldsymbol{x}^i,\boldsymbol{y}^i,\boldsymbol{w}) \le \min_{\boldsymbol{\varphi}} U(\boldsymbol{\varphi};\boldsymbol{x}^i,\boldsymbol{y}^i,\boldsymbol{w})$$

where

$$U(\boldsymbol{\varphi};\boldsymbol{x}^i,\boldsymbol{y}^i,\boldsymbol{w}) = \Bigg[\sum_{v\in\mathcal{V}} \max_{y\in\mathcal{Y}} q^{\boldsymbol{\varphi}}_v(y;\boldsymbol{x}^i, y^i_v, \boldsymbol{w}) + \sum_{\{v,v'\}\in\mathcal{E}} \max_{(y,y')\in\mathcal{Y}^2} g^{\boldsymbol{\varphi}}_{vv'}(y,y';\boldsymbol{x}^i,\boldsymbol{w})\Bigg]$$

and $\boldsymbol{\varphi}\in\mathbb{R}^{2|\mathcal{E}||\mathcal{Y}|}$ is a vector of dual variables $\varphi_{vv'}\colon \mathcal{Y}\to\mathbb{R}$, $\varphi_{v'v}\colon \mathcal{Y}\to\mathbb{R}$, $\{v,v'\}\in\mathcal{E}$ and

$$q^{\boldsymbol{\varphi}}_v(y;\boldsymbol{x}^i,y^i_v,\boldsymbol{w}) = \ell_v(y^i_v,y_v) + \langle \boldsymbol{w}, \boldsymbol{\Psi}_v(\boldsymbol{x}^i,y)\rangle - \sum_{v'\in\mathcal{N}(v)} \varphi_{vv'}(y), \quad v\in\mathcal{V}, y\in\mathcal{Y}$$
$$g^{\boldsymbol{\varphi}}_{vv'}(y,y';\boldsymbol{x}^i,\boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{\Psi}_{vv'}(\boldsymbol{x}^i,y,y')\rangle + \varphi_{vv'}(y) + \varphi_{v'v}(y'), \quad \{v,v'\}\in\mathcal{E}, y,y'\in\mathcal{Y}$$

8.B: LP relaxation of margin-rescaling loss

The LP-relaxed margin-rescaling surrogate is obtained by substituting the LP dual:

$$\begin{aligned}\hat{\ell}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) &= \max_{\boldsymbol{y} \in \mathcal{Y}^{\mathcal{V}}} \Big[\ell(\boldsymbol{y}^i, \boldsymbol{y}) + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y})\rangle\Big] - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i)\rangle \\ &\leq \min_{\boldsymbol{\varphi}} U(\boldsymbol{\varphi}; \boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i)\rangle \\ &= \hat{\ell}_{\mathrm{LP}}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})\end{aligned}$$

Learning: RRM leads to a convex problem $\boldsymbol{w}^* = \operatorname{argmin}_{\boldsymbol{w} \in \mathbb{R}^n} \Big[\lambda \Omega(\boldsymbol{w}) + \hat{R}_{\mathcal{T}}(\boldsymbol{w})\Big]$ where

$$\hat{R}_{\mathcal{T}}(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \hat{\ell}_{\mathrm{LP}}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \min_{\boldsymbol{\varphi}} \Big[U(\boldsymbol{\varphi}; \boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i)\rangle\Big]$$

First order oracle: compute the value $\hat{\ell}_{\mathrm{LP}}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})$ and its sub-gradient $\hat{\ell}'_{\mathrm{LP}}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})$

1. Solve the LP dual
$$\boldsymbol{\varphi}^i = \operatorname*{argmin}_{\boldsymbol{\varphi}} U(\boldsymbol{\varphi}; \boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})$$

2. Compute the sub-gradient
$$\hat{\ell}'_{\mathrm{LP}}(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) \in \partial_{\boldsymbol{w}} \big[U(\boldsymbol{\varphi}^i; \boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i)\rangle\big]$$

8.B: PosLearn: surrogate of additive loss

A linear classifier $h(\boldsymbol{x};\boldsymbol{w}) = (h_v(\boldsymbol{x};\boldsymbol{w}) \in \mathcal{Y} \mid v \in \mathcal{V}) \in \operatorname{argmax}_{\boldsymbol{y}\in\mathcal{Y}^{\mathcal{V}}} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\rangle$ can be evaluated component-wise

$$h_v(\boldsymbol{x};\boldsymbol{w}) \in \operatorname*{argmax}_{y_v\in\mathcal{Y}} \max_{\boldsymbol{y}\in\mathcal{Y}(y_v)} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{y})\rangle \qquad \text{where} \qquad \mathcal{Y}(y_v) = \{\boldsymbol{y}' \in \mathcal{Y}^{\mathcal{V}} \mid y'_v = y_v\}$$

An additive loss $\ell(\boldsymbol{y}^i, h(\boldsymbol{x}^i;\boldsymbol{w})) = \sum_{v\in\mathcal{V}} \ell_v(y^i_v, h_v(\boldsymbol{x};\boldsymbol{w}))$ can be bound by $\hat{\ell}(\boldsymbol{x}^i,\boldsymbol{y}^i,\boldsymbol{w}) = \sum_{v\in\mathcal{V}} \hat{\ell}_v(\boldsymbol{x}^i, y^i_v, \boldsymbol{w})$ where

$$\begin{aligned}
\ell_v(y^i_v, h_v(\boldsymbol{x}^i,\boldsymbol{w})) &= \max_{y_v\in\mathcal{Y}} \ell_v(y^i_v, y_v)[\![\max_{\boldsymbol{y}'\in\mathcal{Y}} \langle \boldsymbol{w}, \Psi(\boldsymbol{x}^i,\boldsymbol{y}')\rangle - \max_{\boldsymbol{y}\in\mathcal{Y}(y_v)} \langle \boldsymbol{w}, \Psi(\boldsymbol{x}^i,\boldsymbol{y})\rangle \le 0]\!] \\
&\le \max_{y_v\in\mathcal{Y}} \ell_v(y^i_v, y_v)[\![\langle \boldsymbol{w}, \Psi(\boldsymbol{x}^i,\boldsymbol{y}^i)\rangle - \max_{\boldsymbol{y}\in\mathcal{Y}(y_v)} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i,\boldsymbol{y})\rangle \le 0]\!] \\
&\le \max_{y_v\in\mathcal{Y}} \ell_v(y^i_v, y_v) \max\left\{0, 1 - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i,\boldsymbol{y}^i)\rangle + \max_{\boldsymbol{y}\in\mathcal{Y}(y_v)} \langle \boldsymbol{w}, \Psi(\boldsymbol{x}^i,\boldsymbol{y})\rangle\right\} \\
&= \max_{y_v\in\mathcal{Y}} \ell_v(y^i_v, y_v)\left[1 - \langle \boldsymbol{w}, \Psi(\boldsymbol{x}^i,\boldsymbol{y}^i)\rangle + \max_{\boldsymbol{y}\in\mathcal{Y}(y_v)} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i,\boldsymbol{y})\rangle\right] \\
&= \hat{\ell}_v(\boldsymbol{x}^i,\boldsymbol{y}^i,\boldsymbol{w})
\end{aligned}$$

8.B: PosLearn: surrogate of additive loss

Learning: RRM leads to a convex problem $\boldsymbol{w}^* = \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \Big[\lambda\Omega(\boldsymbol{w}) + \hat{R}_{\mathcal{T}}(\boldsymbol{w})\Big]$ where

$$\begin{aligned}\hat{R}_{\mathcal{T}}(\boldsymbol{w}) &= \frac{1}{m}\sum_{i=1}^{m}\sum_{v\in\mathcal{V}} \hat{\ell}_v(\boldsymbol{x}^i, y_v^i, \boldsymbol{w}) \\ &= \frac{1}{m}\sum_{i=1}^{m}\sum_{v\in\mathcal{V}} \max_{y_v\in\mathcal{Y}} \ell_v(y_v^i, y_v)\Big[1 - \langle \boldsymbol{w}, \Psi(\boldsymbol{x}^i, \boldsymbol{y}^i)\rangle + \max_{\boldsymbol{y}\in\mathcal{Y}(y_v)} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y})\rangle\Big]\end{aligned}$$

For example, in the case of the Hamming loss $\ell_v(y, y') = [\![y \neq y']\!]$ the surrogate simplifies to

$$\hat{\ell}_v(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) = \max\left\{0, 1 - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i)\rangle + \max_{y_v\in\mathcal{Y}\setminus\{y_v^i\}} \max_{\boldsymbol{y}\in\mathcal{Y}(y_v)} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y})\rangle\right\}$$

8.B: PosLearn: surrogate of additive loss

Consider surrogate of the Hamming loss $\ell_v(y, y') = [\![y \neq y']\!]$ which reads

$$\hat{\ell}_v(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) = \max\left\{0, 1 - \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i)\rangle + \max_{y_v \in \mathcal{Y}\setminus\{y_v^i\}} \max_{\boldsymbol{y} \in \mathcal{Y}(y_v)} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y})\rangle\right\}$$

First order oracle: compute the value $\hat{\ell}_v(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})$ and its sub-gradient $\hat{\ell}'_v(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w})$ at $\boldsymbol{w}$

1. Compute
$$\hat{\boldsymbol{y}}_v^i \in \operatorname*{argmax}_{y_v \in \mathcal{Y}\setminus\{y_v^i\}} \max_{\boldsymbol{y} \in \mathcal{Y}(y_v)} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y})\rangle$$
which is the best labeling not containing y_v^i at position v.

2. Compute sub-gradient
$$\hat{\ell}'_v(\boldsymbol{x}^i, \boldsymbol{y}^i, \boldsymbol{w}) = \begin{cases} \mathbf{0} & \text{if} \quad \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i)\rangle \geq 1 + \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}^i, \hat{\boldsymbol{y}}_v^i)\rangle \\ \boldsymbol{\Psi}(\boldsymbol{x}^i, \hat{\boldsymbol{y}}_v^i) - \boldsymbol{\Psi}(\boldsymbol{x}^i, \boldsymbol{y}^i) & \text{otherwise} \end{cases}$$

For example, the max-marginals $\max_{\boldsymbol{y} \in \mathcal{Y}(y_v)} \langle \boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y})\rangle$ can be computed efficiently for max-sum classifier with acyclic graph $(\mathcal{V}, \mathcal{E})$.