A. Duality

The following is a version of the Fenchel duality theorem (see (Rockafellar, 1997)).

Theorem 5. *Let X and Y be Banach spaces, and $f: X \to \mathbb{R} \cup \{+\infty\}$ and $g: Y \to \mathbb{R} \cup \{+\infty\}$ convex functions. Let $A: X \to Y$ be a bounded linear map. If g is continuous at some point $y \in A\,\mathrm{dom}(f)$, then the following holds:*

$$\inf_{x \in X}(f(x) + g(Ax)) = \sup_{y^* \in Y^*}(-f^*(A^*y^*) - g^*(-y^*)), \tag{16}$$

where f^* and g^* are conjugate functions of f and g respectively, and A^* the adjoint of A. Furthermore, the supremum in (16) is attained if it is finite.

The following lemma gives the expression of the conjugate function of the (extended) relative entropy, which is a standard result (Boyd & Vandenberghe, 2004).

Lemma 6 (Conjugate function of the relative entropy)**.** *Let $f: \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ be defined by $f(\mathsf{p}) = D(\mathsf{p} \parallel \mathsf{p}_0)$ if $\mathsf{p} \in \Delta$ and $f(\mathsf{p}) = +\infty$ elsewhere. Then, the conjugate function of f is the function $f^*: \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ defined for all $\mathsf{q} \in \mathbb{R}^{\mathcal{X}}$ by*

$$f^*(\mathsf{q}) = \log\Big(\sum_{x \in \mathcal{X}} \mathsf{p}_0[x] e^{\mathsf{q}[x]}\Big) = \log\Big(\mathop{\mathrm{E}}_{x \sim \mathsf{p}_0}[e^{\mathsf{q}[x]}]\Big).$$

Proof. By definition of f, for any $\mathsf{q} \in \mathbb{R}^{\mathcal{X}}$, we can write

$$\sup_{\mathsf{p} \in \mathbb{R}^{\mathcal{X}}}\big(\langle \mathsf{p}, \mathsf{q}\rangle - D(\mathsf{p} \parallel \mathsf{p}_0)\big) = \sup_{\mathsf{p} \in \Delta}\big(\langle \mathsf{p}, \mathsf{q}\rangle - D(\mathsf{p} \parallel \mathsf{p}_0)\big).$$

Fix $\mathsf{q} \in \mathbb{R}^{\mathcal{X}}$ and let $\bar{\mathsf{q}} \in \Delta$ be defined for all $x \in \mathcal{X}$ by

$$\bar{\mathsf{q}}[x] = \frac{\mathsf{p}_0[x]e^{\mathsf{q}[x]}}{\sum_{x \in \mathcal{X}} \mathsf{p}_0[x]e^{\mathsf{q}[x]}} = \frac{\mathsf{p}_0[x]e^{\mathsf{q}[x]}}{\mathrm{E}_{p_0}[e^{\mathsf{q}}]}. \tag{17}$$

Then, the following holds for all $\mathsf{p} \in \Delta$:

$$\begin{aligned}\langle \mathsf{p}, \mathsf{q}\rangle - D(\mathsf{p} \parallel \mathsf{p}_0) &= \mathop{\mathrm{E}}_{\mathsf{p}}[\log(e^{\mathsf{q}})] - \mathop{\mathrm{E}}_{\mathsf{p}}\Big[\log \frac{\mathsf{p}}{\mathsf{p}_0}\Big] \\ &= \mathop{\mathrm{E}}_{\mathsf{p}}\Big[\log \frac{\mathsf{p}_0 e^{\mathsf{q}}}{p}\Big] \\ &= -D(\mathsf{p} \parallel \bar{\mathsf{q}}) + \log \mathop{\mathrm{E}}_{\mathsf{p}_0}[e^{\mathsf{q}}].\end{aligned}$$

Since $D(\mathsf{p} \parallel \bar{\mathsf{q}}) \geq 0$ and $D(\mathsf{p} \parallel \bar{\mathsf{q}}) = 0$ for $\mathsf{p} = \bar{\mathsf{q}}$, this shows that $\sup_{\mathsf{p} \in \Delta}\big(\mathsf{p} \cdot \mathsf{q} - D(\mathsf{p} \parallel \mathsf{p}_0)\big) = \log\big(\mathrm{E}_{\mathsf{p}_0}[e^{\mathsf{q}}]\big)$ and concludes the proof. ☐

Theorem 1. *Problem (2) and (3) are equivalent to the dual optimization problem $\sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w})$:*

$$\sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w}) = \min_{\mathsf{p}} F(\mathsf{p}). \tag{18}$$

Furthermore, let $\mathsf{p}^* = \mathrm{argmin}_{\mathsf{p}} F(\mathsf{p})$, then, for any $\epsilon > 0$ and any $\mathbf{w}$ such that $|G(\mathbf{w}) - \sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w})| < \epsilon$, the following inequality holds: $D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) \leq \epsilon$.

Proof. The proof follows by application of the Fenchel duality theorem (Theorem 5, Appendix A) to the optimization problem (3) with the functions f and g defined for all p and $\mathbf{u}$ by $f(\mathsf{p}) = D(\mathsf{p} \parallel \mathsf{p}_0) + I_\Delta(\mathsf{p})$ and $g(\mathbf{u}) = I_C(\mathbf{u})$ and with A the linear map defined by $A\mathsf{p} = \mathrm{E}_{\mathsf{p}}[\mathbf{\Phi}]$.

A is a bounded linear map since $\|A\| \leq \|\mathbf{\Phi}\|_\infty \leq \Lambda$ and $A^*\mathbf{w} = \mathbf{w} \cdot \mathbf{\Phi}$. Furthermore, define $\mathbf{u} \in \mathbb{F}$ by $\mathbf{u}_k = \mathrm{E}_{\widehat{\mathsf{p}}}[\mathbf{\Phi}_k]$. Then, $\mathbf{u}$ is in $A\,\mathrm{dom} f$ and is in C. Since $\beta_k > 0$ for all k, $\mathbf{u}$ is contained in $\mathrm{int}(C)$. $g = I_C$ equals zero over $\mathrm{int}(C)$ and is therefore continuous over $\mathrm{int}(C)$, thus g is continuous at $\mathbf{u} \in A\,\mathrm{dom} f$.

By Lemma 6, the conjugate of f is the function $f^*: \mathbb{R}^{\mathcal{X}} \to \mathbb{R}$ defined by $f^*(\mathsf{q}) = \log\big(\sum_{x \in \mathcal{X}} \mathsf{p}_0[x]e^{\mathsf{q}[x]}\big)$ for all $\mathsf{q} \in \mathbb{R}^{\mathcal{X}}$. The conjugate function of $g = I_C$ is the function g^* defined for all $\mathbf{w} \in \mathbb{R}^N$ by

$$\begin{aligned}g^*(\mathbf{w}) &= \sup_{\mathbf{u} \in C}\big(\mathbf{w} \cdot \mathbf{u} - I_C(\mathbf{u})\big) \\ &= \sup_{\mathbf{u} \in C}(\mathbf{w} \cdot \mathbf{u}) \\ &= \sup_{\mathbf{u} \in C}\Big(\sum_{k=1}^p \mathbf{w}_k \cdot \mathbf{u}_k\Big) \\ &= \sum_{k=1}^p \sup_{\|\mathbf{u}_k - \mathrm{E}_S[\mathbf{\Phi}_k]\|_\infty \leq \beta_k}(\mathbf{w}_k \cdot \mathbf{u}_k) \\ &= \sum_{k=1}^p \mathbf{w}_k \cdot \mathop{\mathrm{E}}_S[\mathbf{\Phi}_k] + \sum_{k=1}^p \sup_{\|\mathbf{u}_k\|_\infty \leq \beta_k}(\mathbf{w}_k \cdot \mathbf{u}_k) \\ &= \mathop{\mathrm{E}}_S[\mathbf{w} \cdot \mathbf{\Phi}] + \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1,\end{aligned}$$

where the penultimate equality holds by definition of the dual norm. In view of these identities, we can write

$$\begin{aligned}&-f^*(A^*\mathbf{w}) - g^*(-\mathbf{w}) \\ &= -\log\Big(\sum_{x \in \mathcal{X}} \mathsf{p}_0[x]e^{\mathbf{w} \cdot \mathbf{\Phi}(x)}\Big) + \mathop{\mathrm{E}}_S[\mathbf{w} \cdot \mathbf{\Phi}] - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 \\ &= -\log Z_{\mathbf{w}} + \frac{1}{m}\sum_{i=1}^m \mathbf{w} \cdot \mathbf{\Phi}(x_i) - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 \\ &= \frac{1}{m}\sum_{i=1}^m \log \frac{e^{\mathbf{w} \cdot \mathbf{\Phi}(x_i)}}{Z_{\mathbf{w}}} - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 \\ &= \frac{1}{m}\sum_{i=1}^m \log\Big[\frac{\mathsf{p}_{\mathbf{w}}[x_i]}{\mathsf{p}_0[x_i]}\Big] - \sum_{k=1}^p \beta_k \|\mathbf{w}_k\|_1 = G(\mathbf{w}),\end{aligned}$$

which proves that $\sup_{\mathbf{w} \in \mathbb{R}^N} G(\mathbf{w}) = \min_{\mathsf{p}} F(\mathsf{p})$. For any

$\mathbf{w} \in \mathbb{R}^N$, we can write

$$
\begin{aligned}
&G(\mathbf{w}) - D(\mathsf{p}^* \parallel \mathsf{p}_0)\\
&= \mathop{\mathrm{E}}_{x\sim\widehat{p}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] - \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|_1 - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}\left[\log \frac{\mathsf{p}^*[x]}{\mathsf{p}_0[x]}\right]\\
&= \mathop{\mathrm{E}}_{x\sim\widehat{p}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] - \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|_1 -\\
&\qquad \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}\left[\log \frac{\mathsf{p}^*[x]}{\mathsf{p}_{\mathbf{w}}[x]}\frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right]\\
&= -D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) - \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|_1\\
&\qquad + \mathop{\mathrm{E}}_{x\sim\widehat{p}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}\left[\log \frac{\mathsf{p}_{\mathbf{w}}(x)}{\mathsf{p}_0(x)}\right].
\end{aligned}
$$

The difference of the last two terms can be expressed as follows

$$
\begin{aligned}
&\mathop{\mathrm{E}}_{x\sim\widehat{p}}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}\left[\log \frac{\mathsf{p}_{\mathbf{w}}[x]}{\mathsf{p}_0[x]}\right]\\
&= \mathop{\mathrm{E}}_{x\sim\widehat{p}}[\mathbf{w}\cdot\mathbf{\Phi}(x) - \log Z_{\mathbf{w}}] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}[\mathbf{w}\cdot\mathbf{\Phi}(x) - \log Z_{\mathbf{w}}]\\
&= \mathop{\mathrm{E}}_{x\sim\widehat{p}}[\mathbf{w}\cdot\mathbf{\Phi}(x)] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}[\mathbf{w}\cdot\mathbf{\Phi}(x)].
\end{aligned}
$$

Plugging back this equality and rearranging yields

$$
\begin{aligned}
&D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) = D(\mathsf{p}^* \parallel \mathsf{p}_0) - G(\mathbf{w})\\
&\qquad - \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|_1 + \mathbf{w}\cdot\left(\mathop{\mathrm{E}}_{x\sim\widehat{p}}[\mathbf{\Phi}(x)] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}[\mathbf{\Phi}(x)]\right).
\end{aligned}
$$

The solution of the primal optimization, p^*, verifies the constraint $I_C(\mathrm{E}_{\mathsf{p}^*}[\mathbf{\Phi}]) = 0$, that is $\| \mathrm{E}_{x\sim\widehat{p}}[\mathbf{\Phi}_k(x)] - \mathrm{E}_{x\sim\mathsf{p}^*}[\mathbf{\Phi}_k(x)]\|_\infty \leq \beta_k$ for all $k \in [1, p]$. By Hölder's inequality, this implies that

$$
\begin{aligned}
&- \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|_1 + \mathbf{w}\cdot\left(\mathop{\mathrm{E}}_{x\sim\widehat{p}}[\mathbf{\Phi}(x)] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}[\mathbf{\Phi}(x)]\right)\\
&= - \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|_1 + \sum_{k=1}^{p} \mathbf{w}_k\cdot\left(\mathop{\mathrm{E}}_{x\sim\widehat{p}}[\mathbf{\Phi}_k(x)] - \mathop{\mathrm{E}}_{x\sim\mathsf{p}^*}[\mathbf{\Phi}_k(x)]\right)\\
&\leq - \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|_1 + \sum_{k=1}^{p} \beta_k \|\mathbf{w}_k\|_1 = 0.
\end{aligned}
$$

Thus, we can write, for any $\mathbf{w} \in \mathbb{R}^N$,

$$
D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) \leq D(\mathsf{p}^* \parallel \mathsf{p}_0) - G(\mathbf{w}).
$$

Now, assume that $\mathbf{w}$ verifies $|G(\mathbf{w}) - \sup_{\mathbf{w}\in\mathbb{R}^N} G(\mathbf{w})| \leq \epsilon$ for some $\epsilon > 0$. Then, $D(\mathsf{p}^* \parallel \mathsf{p}_0) - G(\mathbf{w}) = \sup_{\mathbf{w}} G(\mathbf{w}) - G(\mathbf{w}) \leq \epsilon$ implies $D(\mathsf{p}^* \parallel \mathsf{p}_{\mathbf{w}}) \leq \epsilon$. This concludes the proof of the theorem. □

Theorem 4. *Problem (10) is equivalent to dual optimization problem* $\sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w})$*:*

$$
\sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w}) = \min_{\mathsf{p}} \widetilde{F}(\mathsf{p}). \tag{19}
$$

Furthermore, let $\mathsf{p}^* = \operatorname{argmin}_{\mathsf{p}} \widetilde{F}(\mathsf{p})$*. Then, for any* $\epsilon > 0$ *and any* $\mathbf{w}$ *such that* $|\widetilde{G}(\mathbf{w}) - \sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w})| < \epsilon$*, we have* $\mathrm{E}_{x\sim\widehat{p}}\left[D\big(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big)\right] \leq \epsilon$*.*

Proof. The proof follows by application of the Fenchel duality theorem (Theorem 5, Appendix A) to the optimization problem (11) with the functions $\widetilde{f}$ and $\widetilde{g}$ defined for all p and $\mathbf{u}$ by $\widetilde{f}(\mathsf{p}) = \mathrm{E}_{x\sim\widehat{p}}\left[D\big(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x]\big) + I_\Delta\big(\mathsf{p}[\cdot|x]\big)\right]$ and $\widetilde{g}(\mathbf{u}) = I_C(\mathbf{u})$ and with A the linear map defined by $A\mathsf{p} = \mathop{\mathrm{E}}_{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\mathsf{p}[\cdot|x]}}[\mathbf{\Phi}(x,y)]$.

A is a bounded linear map since $\|A\| \leq \|\mathbf{\Phi}\|_\infty \leq \Lambda$. Note that

$$
\begin{aligned}
A\mathsf{p} = \mathop{\mathrm{E}}_{\substack{x\sim\widehat{\mathsf{p}}\\ y\sim\mathsf{p}[\cdot|x]}}[\mathbf{\Phi}(x,y)] &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \mathbf{\Phi}(x,y)\widehat{\mathsf{p}}[x]\mathsf{p}[y|x]\\
&= \sum_{x\in\operatorname{supp}(\widehat{p})} (\widehat{\mathsf{p}}[x]\mathbf{\Phi}(x,\cdot))\cdot(\mathsf{p}[\cdot|x]).
\end{aligned}
$$

Thus, the conjugate of A is defined for all $\mathbf{w} \in \mathbb{R}^N$ by $A^*\mathbf{w} = \mathbf{w}\cdot\big(\widehat{p}(x)\mathbf{\Phi}(x,y)\big)$. Furthermore, define $\mathbf{u} \in \mathbb{F}$ by $\mathbf{u}_k = \mathrm{E}_{(x,y)\sim S}[\mathbf{\Phi}_k(x,y)]$. Then, $\mathbf{u}$ is in $A\,\mathrm{dom}f$ and is in C. Since $\beta_k > 0$ for all k, $\mathbf{u}$ is contained in $\mathrm{int}(C)$. $g = I_C$ equals zero over $\mathrm{int}(C)$ and is therefore continuous over $\mathrm{int}(C)$, thus g is continuous at $\mathbf{u} \in A\,\mathrm{dom}f$.

The conjugate function of $\widetilde{f}$ is defined for all $\mathsf{q} = (\mathsf{q}[\cdot|x_i])_{i\in[1,m]}$ by

$$
\begin{aligned}
\widetilde{f}^*(\mathsf{q}) &= \sup_{\mathsf{p}[\cdot|x]\in\Delta} \Big\{\langle \mathsf{p},\mathsf{q}\rangle - \sum_{x\in\mathcal{X}} \widehat{\mathsf{p}}[x]\, D(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])\Big\}\\
&= \sup_{\mathsf{p}[\cdot|x]\in\Delta} \Big\{ \sum_{x\in\operatorname{supp}(\widehat{\mathsf{p}})} \widehat{\mathsf{p}}[x] \sum_{y\in\mathcal{Y}} \mathsf{p}[y|x]\mathsf{q}[y|x](\widehat{\mathsf{p}}[x])^{-1}\\
&\qquad\qquad - \sum_{x\in\operatorname{supp}(\widehat{\mathsf{p}})} \widehat{\mathsf{p}}[x] D(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])\Big\}\\
&= \sum_{x\in\operatorname{supp}(\widehat{\mathsf{p}})} \widehat{\mathsf{p}}[x] \sup_{\mathsf{p}[\cdot|x]} \Big\{ \sum_{y\in\mathcal{Y}} \mathsf{p}[y|x] \left(\frac{\mathsf{q}[y|x]}{\widehat{\mathsf{p}}[x]}\right)\\
&\qquad\qquad - D(\mathsf{p}[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])\Big\}\\
&= \sum_{x\in\operatorname{supp}(\widehat{\mathsf{p}})} \widehat{\mathsf{p}}[x] f_x^*\left(\frac{\mathsf{q}[y|x]}{\widehat{\mathsf{p}}[x]}\right)
\end{aligned}
$$

where f_x is defined for all $x \in \mathcal{X}$ and $\mathsf{p}' \in \mathbb{R}^{\mathcal{Y}}$ by $f(\mathsf{p}') = D(\mathsf{p}' \parallel \mathsf{p}_0[\cdot|x])$ if $\mathsf{p}' \in \Delta$, $f(\mathsf{p}') = +\infty$ otherwise. By

Lemma 6, $f_x^*\left(\frac{\mathsf{q}[y|x]}{\widehat{\mathsf{p}}[x]}\right) = \log\left(\sum_{y\in\mathcal{Y}} \mathsf{p}_0[y|x] e^{\frac{\mathsf{q}[y|x]}{\widehat{\mathsf{p}}[x]}}\right)$, thus, $\widetilde{f}^*$ is given by

$$\widetilde{f}^*(\mathsf{q}) = \mathop{\mathrm{E}}_{x\sim\widehat{\mathsf{p}}}\left[\log\left(\sum_{y\in\mathcal{Y}} \mathsf{p}_0[y|x] e^{\frac{\mathsf{q}[y|x]}{\widehat{\mathsf{p}}[x]}}\right)\right].$$

In view of these identities, we can write

$$\begin{aligned}
&-\widetilde{f}^*(A^*\mathbf{w}) - \widetilde{g}^*(-\mathbf{w}) \\
&= -\mathop{\mathrm{E}}_{x\sim\widehat{\mathsf{p}}}\left[\log\left(\sum_{y\in\mathcal{Y}} \mathsf{p}_0[y|x] e^{\mathbf{w}\cdot\mathbf{\Phi}(x,y)}\right)\right] \\
&\qquad + \mathop{\mathrm{E}}_{S}[\mathbf{w}\cdot\mathbf{\Phi}] - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 \\
&= -\mathop{\mathrm{E}}_{x\sim\widehat{\mathsf{p}}}[\log Z_{\mathbf{w}}(x)] + \frac{1}{m}\sum_{i=1}^m \mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i) - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 \\
&= \frac{1}{m}\sum_{i=1}^m \log\frac{e^{\mathbf{w}\cdot\mathbf{\Phi}(x_i,y_i)}}{Z_{\mathbf{w}}(x_i)} - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 \\
&= \frac{1}{m}\sum_{i=1}^m \log\left[\frac{\mathsf{p}_{\mathbf{w}}[y_i|x_i]}{\mathsf{p}_0[y_i|x_i]}\right] - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 = \widetilde{G}(\mathbf{w}),
\end{aligned}$$

which proves that $\sup_{\mathbf{w}\in\mathbb{R}^N} \widetilde{G}(\mathbf{w}) = \min_{\mathsf{p}} \widetilde{F}(\mathsf{p})$. The second part of the proof is similar to that of Theorem 1. For any $\mathbf{w}\in\mathbb{R}^N$, we can write

$$\begin{aligned}
&\widetilde{G}(\mathbf{w}) - \mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])] \\
&= \mathop{\mathrm{E}}_{(x,y)\sim S}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 \\
&\qquad - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log\frac{\mathsf{p}^*[y|x]}{\mathsf{p}_0[y|x]}\right] \\
&= \mathop{\mathrm{E}}_{(x,y)\sim S}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 - \\
&\qquad \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log\frac{\mathsf{p}^*[y|x]}{\mathsf{p}_{\mathbf{w}}[y|x]}\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] \\
&= -\mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])] - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 \\
&\qquad + \mathop{\mathrm{E}}_{(x,y)\sim S}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right].
\end{aligned}$$

The difference of the last two terms can be expressed as follows

$$\begin{aligned}
&\mathop{\mathrm{E}}_{(x,y)\sim S}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}\left[\log\frac{\mathsf{p}_{\mathbf{w}}[y|x]}{\mathsf{p}_0[y|x]}\right] \\
&= \mathop{\mathrm{E}}_{(x,y)\sim S}[\mathbf{w}\cdot\mathbf{\Phi}(x,y) - \log Z_{\mathbf{w}}(x)] \\
&\qquad - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{w}\cdot\mathbf{\Phi}(x,y) - \log Z_{\mathbf{w}}(x)] \\
&= \mathop{\mathrm{E}}_{(x,y)\sim S}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)].
\end{aligned}$$

Plugging back this equality and rearranging yields

$$\begin{aligned}
&\mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])] \\
&= \mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])] - \widetilde{G}(\mathbf{w}) - \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 \\
&\qquad + \mathbf{w}\cdot\left[\mathop{\mathrm{E}}_{(x,y)\sim S}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)]\right].
\end{aligned}$$

The solution of the primal optimization, p^*, verifies the constraint $I_C(\mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{\Phi}(x,y)]) = 0$, that is $\|\mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{\Phi}_k(x,y)] - \mathrm{E}_{(x,y)\sim S}[\mathbf{\Phi}_k(x,y)]\|_\infty \leq \beta_k$ for all $k\in[1,p]$. By Hölder's inequality, this implies that

$$\begin{aligned}
&-\sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 \\
&\qquad + \mathbf{w}\cdot\left[\mathop{\mathrm{E}}_{(x,y)\sim S}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)] - \mathop{\mathrm{E}}_{\substack{x\sim\widehat{p}\\ y\sim\mathsf{p}^*[\cdot|x]}}[\mathbf{w}\cdot\mathbf{\Phi}(x,y)]\right] \\
&\leq -\sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 + \sum_{k=1}^p \beta_k\|\mathbf{w}_k\|_1 = 0.
\end{aligned}$$

Thus, we can write, for any $\mathbf{w}\in\mathbb{R}^N$,

$$\begin{aligned}
&\mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])] \\
&\qquad\qquad \leq \mathop{\mathrm{E}}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])] - \widetilde{G}(\mathbf{w}).
\end{aligned}$$

Now, assume that $\mathbf{w}$ verifies $|\widetilde{G}(\mathbf{w}) - \sup_{\mathbf{w}\in\mathbb{R}^N}\widetilde{G}(\mathbf{w})| \leq \epsilon$ for some $\epsilon > 0$. Then, $\mathrm{E}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_0[\cdot|x])] - \widetilde{G}(\mathbf{w}) = \sup_{\mathbf{w}}\widetilde{G}(\mathbf{w}) - \widetilde{G}(\mathbf{w}) \leq \epsilon$ implies $\mathrm{E}_{x\sim\widehat{p}}[D(\mathsf{p}^*[\cdot|x] \parallel \mathsf{p}_{\mathbf{w}}[\cdot|x])] \leq \epsilon$. This concludes the proof of the theorem. □

B. Pseudocode of StructMaxent1

Figure 2 shows the pseudocode of StructMaxent1.

StructMaxent1$(S = (x_1, \ldots, x_m))$

for $t \leftarrow 1$ **to** T **do**
$\quad$ **for** $k \leftarrow 1$ **to** p **and** $j \leftarrow 1$ **to** N_k **do**
$\quad\quad$ **if** $(w_{t-1,k,j} \neq 0)$ **then**
$\quad\quad\quad$ $d_{k,j} \leftarrow \beta_k \operatorname{sgn}(w_{t-1,k,j}) + \epsilon_{t-1,k,j}$
$\quad\quad$ **elseif** $|\epsilon_{t-1,k,j}| \leq \beta_k$ **then**
$\quad\quad\quad$ $d_{k,j} \leftarrow 0$
$\quad\quad$ **else** $d_{k,j} \leftarrow -\beta_k \operatorname{sgn}(\epsilon_{t-1,k,j}) + \epsilon_{t-1,k,j}$
$\quad$ $(k, j) \leftarrow \operatorname*{argmax}_{(k,j)\in[1,p]\times[1,N_k]} |d_{k,j}|$
$\quad$ $\beta \leftarrow \frac{\overline{\Phi}^+_{t-1,k,j}\overline{\Phi}^-_{k,j} e^{-2w_{k,j}\Lambda} - \overline{\Phi}^+_{k,j}\overline{\Phi}^-_{t-1,k,j}}{\overline{\Phi}^+_{t-1,k,j} e^{-2w_{k,j}\Lambda} - \overline{\Phi}^-_{t-1,k,j}}$
$\quad$ **if** $(|\beta| \leq \beta_k)$ **then**
$\quad\quad$ $\eta \leftarrow -w_{t-1,k,j}$
$\quad$ **elseif** $(\beta > \beta_k)$ **then**
$\quad\quad$ $\eta \leftarrow \frac{1}{2\Lambda} \log\left[\frac{\overline{\Phi}^-_{t-1,k,j}(\beta_k - \overline{\Phi}^+_{k,j})}{\overline{\Phi}^+_{t-1,k,j}(\beta_k - \overline{\Phi}^-_{k,j})}\right]$
$\quad$ **else** $\eta \leftarrow \frac{1}{2\Lambda} \log\left[\frac{\overline{\Phi}^-_{t-1,k,j}(\beta_k + \overline{\Phi}^+_{k,j})}{\overline{\Phi}^+_{t-1,k,j}(\beta_k + \overline{\Phi}^-_{k,j})}\right]$
$\quad$ $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} + \eta \mathbf{e}_{k,j}$
$\quad$ $\mathsf{p}_{\mathbf{w}_t} \leftarrow \frac{\mathsf{p}_0[x] e^{\mathbf{w}_t \cdot \mathbf{\Phi}(x)}}{\sum_{x\in\mathcal{X}} \mathsf{p}_0[x] e^{\mathbf{w}_t \cdot \mathbf{\Phi}(x)}}$
return $\mathsf{p}_{\mathbf{w}_t}$

Figure 2. Pseudocode of the StructMaxent1 algorithm. For all $(k, j) \in [1, p] \times [1, N_k]$, $\beta_k = 2\mathfrak{R}_m(H_k) + \beta$, $\epsilon_{t-1,k,j} = \mathrm{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] - \mathrm{E}_S[\Phi_{k,j}]$ and, for any $s \in \{-1, +1\}$, $\overline{\Phi}^s_{t-1,k,j} = \mathrm{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] + s\Lambda$ and $\overline{\Phi}^s_{k,j} = \mathrm{E}_S[\Phi_{k,j}] + s\Lambda$. The closed-form solutions for the step size given here assume that the conditions (8) hold.

C. Algorithm

In this section we derive the step size for the StructMaxent1 and StructMaxent2 algorithms presented in Section 2.4 and Appendix B.

Observe that

$$F(\mathbf{w}_{t-1} + \eta\, \mathbf{e}_{k,j}) - F(\mathbf{w}_{t-1}) \tag{20}$$
$$= \beta_k(|w_{k,j} + \eta| - |w_{k,j}|) - \eta \mathop{\mathrm{E}}_{S}[\Phi_{k,j}] + \log\left[\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta\Phi_{k,j}}]\right].$$

Since $\Phi_{k,j} \in [-\Lambda, +\Lambda]$, by the convexity of $x \mapsto e^{\eta x}$, we can write

$$e^{\eta\Phi_{k,j}} \leq \frac{\Lambda - \Phi_{k,j}}{2\Lambda} e^{-\eta\Lambda} + \frac{\Phi_{k,j} + \Lambda}{2\Lambda} e^{\eta\Lambda}.$$

Taking the expectation and the log yields

$$\log \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta\Phi_{k,j}}] \leq \log\left[\frac{\overline{\Phi}^+_{t-1,k,j} e^{\eta\Lambda} - \overline{\Phi}^-_{t-1,k,j} e^{-\eta\Lambda}}{2\Lambda}\right]$$
$$= -\eta\Lambda + \log\left[\frac{\overline{\Phi}^+_{t-1,k,j} e^{2\eta\Lambda} - \overline{\Phi}^-_{t-1,k,j}}{2\Lambda}\right],$$

where we used the following notation:

$$\overline{\Phi}^s_{t-1,k,j} = \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] + s\Lambda \qquad \overline{\Phi}^s_{k,j} = \mathop{\mathrm{E}}_{S}[\Phi_{k,j}] + s\Lambda,$$

for all $(k, j) \in [1, p] \times [1, N_k]$ and $s \in \{-1, +1\}$.

Plugging back this inequality in (20) and ignoring constant terms, minimizing the resulting upper bound on $F(\mathbf{w}_{t-1} + \eta\, \mathbf{e}_{k,j}) - F(\mathbf{w}_{t-1})$ becomes equivalent to minimizing $\psi(\eta)$ defined for all $\eta \in \mathbb{R}$ by

$$\psi(\eta) = \beta_k|w_{k,j} + \eta| - \eta\overline{\Phi}^+_{k,j} + \log\left[\overline{\Phi}^+_{t-1,k,j} e^{2\eta\Lambda} - \overline{\Phi}^-_{t-1,k,j}\right].$$

Let η^* denote the minimizer of $\psi(\eta)$. If $w_{t-1,k,j} + \eta^* = 0$, then the subdifferential of $|w_{t-1,k,j} + \eta|$ at η^* is the set $\{\nu \colon \nu \in [-1, +1]\}$. Thus, in that case, the subdifferential $\partial\psi(\eta^*)$, contains 0 iff there exists $\nu \in [-1, +1]$ such that

$$\beta_k\nu - \overline{\Phi}^+_{k,j} + \frac{2\Lambda\overline{\Phi}^+_{t-1,k,j} e^{2\eta^*\Lambda}}{\overline{\Phi}^+_{t-1,k,j} e^{2\eta^*\Lambda} - \Phi^-_{t-1,k,j}} = 0$$
$$\Leftrightarrow \overline{\Phi}^+_{k,j} - \frac{2\Lambda\overline{\Phi}^+_{t-1,k,j} e^{-2w_{t-1,k,j}\Lambda}}{\overline{\Phi}^+_{t-1,k,j} e^{-2w_{t-1,k,j}\Lambda} - \Phi^-_{t-1,k,j}} = \beta_k\nu.$$

Thus, the condition is equivalent to

$$\left|\overline{\Phi}^+_{k,j} - \frac{2\Lambda\overline{\Phi}^+_{t-1,k,j} e^{-2w_{t-1,k,j}\Lambda}}{\overline{\Phi}^+_{t-1,k,j} e^{-2w_{t-1,k,j}\Lambda} - \Phi^-_{t-1,k,j}}\right| \leq \beta_k,$$

which can be rewritten as

$$\left|\frac{\overline{\Phi}^+_{t-1,k,j}\overline{\Phi}^-_{k,j} e^{-2w_{k,j}\Lambda} - \overline{\Phi}^+_{k,j}\overline{\Phi}^-_{t-1,k,j}}{\overline{\Phi}^+_{t-1,k,j} e^{-2w_{k,j}\Lambda} - \overline{\Phi}^-_{t-1,k,j}}\right| \leq \beta_k.$$

If $w_{t-1,k,j} + \eta^* > 0$, then ψ is differentiable at η^* and $\psi'(\eta^*) = 0$, that is

$$\beta - \overline{\Phi}^+_{k,j} + \frac{2\Lambda\overline{\Phi}^+_{t-1,k,j} e^{2\eta^*\Lambda}}{\overline{\Phi}^+_{t-1,k,j} e^{2\eta^*\Lambda} - \overline{\Phi}^-_{t-1,k,j}} = 0$$
$$\Leftrightarrow e^{2\eta^*\Lambda} = \frac{\overline{\Phi}^-_{t-1,k,j}(\beta_k - \overline{\Phi}^+_{k,j})}{\overline{\Phi}^+_{t-1,k,j}(\beta_k - \overline{\Phi}^-_{k,j})}$$
$$\Leftrightarrow \eta^* = \frac{1}{2\Lambda}\log\left[\frac{\overline{\Phi}^-_{t-1,k,j}(\beta_k - \overline{\Phi}^+_{k,j})}{\overline{\Phi}^+_{t-1,k,j}(\beta_k - \overline{\Phi}^-_{k,j})}\right].$$

For the step size η^* to be in $\mathbb{R}$, the following conditions must be met:

$$(\overline{\Phi}^-_{t-1,k,j} \neq 0) \wedge (\overline{\Phi}^+_{t-1,k,j} \neq 0) \wedge$$
$$((\beta_k - \overline{\Phi}^+_{k,j}) < 0) \wedge ((\beta_k - \overline{\Phi}^-_{k,j}) \neq 0),$$

that is

$$(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] \notin \{-\Lambda, +\Lambda\}) \wedge (\mathop{\mathrm{E}}_{S}[\Phi_{k,j}] > -\Lambda + \beta_k). \tag{21}$$

The condition $w_{t-1,k,j} + \eta^* > 0$ is equivalent to $e^{2\eta^*\Lambda} > e^{-2w_{t-1,k,j}\Lambda}$, which, in view of the expression of $e^{2\eta^*\Lambda}$ given above can be written as

$$\frac{\overline{\Phi}^+_{t-1,k,j}\overline{\Phi}^-_{k,j}e^{-2w_{k,j}\Lambda} - \overline{\Phi}^+_{k,j}\overline{\Phi}^-_{t-1,k,j}}{\overline{\Phi}^+_{t-1,k,j}e^{-2w_{k,j}\Lambda} - \overline{\Phi}^-_{t-1,k,j}} > \beta_k.$$

Similarly, if $w_{t-1,k,j} + \eta^* < 0$, ψ is differentiable at η^* and $\psi'(\eta^*) = 0$, which gives

$$\eta^* = \frac{1}{2\Lambda}\log\left[\frac{\overline{\Phi}^-_{t-1,k,j}(\beta_k + \overline{\Phi}^+_{k,j})}{\overline{\Phi}^+_{t-1,k,j}(\beta_k + \overline{\Phi}^-_{k,j})}\right].$$

Again for the step size η^* to be in $\mathbb{R}$, the following conditions must be met:

$$(\overline{\Phi}^-_{t-1,k,j} \neq 0) \wedge (\overline{\Phi}^+_{t-1,k,j} \neq 0)\wedge \\ ((\beta_k + \overline{\Phi}^+_{k,j}) \neq 0) \wedge ((\beta_k + \overline{\Phi}^-_{k,j}) < 0),$$

that is

$$(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] \notin \{-\Lambda, +\Lambda\}) \wedge (\mathop{\mathrm{E}}_{S}[\Phi_{k,j}] < \Lambda - \beta_k).$$

Combining with condition 21, the following condition on Φ, Λ and β_k must be satisfied:

$$(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] \notin \{-\Lambda, +\Lambda\})\wedge \\ (-\Lambda + \beta_k < \mathop{\mathrm{E}}_{S}[\Phi_{k,j}] < \Lambda - \beta_k).$$

Figure 2 shows the pseudocode of our algorithm using the closed-form solution for the step size just presented.

An alternative method consists of using a somewhat looser upper bound for $\log \mathrm{E}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta\Phi_{k,j}}]$ using Hoeffding's lemma and $\Phi_{k,j} \in [-\Lambda, +\Lambda]$:

$$\log \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[e^{\eta\Phi_{k,j}}] \leq \eta \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_{t-1}}}[\Phi_{k,j}] + \frac{\eta^2\Lambda^2}{2}.$$

Combining this inequality with (20) and disregarding constant terms, minimizing the resulting upper bound on $F(\mathbf{w}_{t-1} + \eta\,\mathbf{e}_{k,j}) - F(\mathbf{w}_{t-1})$ becomes equivalent to minimizing $\varphi(\eta)$ defined for all $\eta \in \mathbb{R}$ by

$$\varphi(\eta) = \beta_k|w_{k,j} + \eta| + \eta\epsilon_{t-1,k,j} + \frac{\eta^2\Lambda^2}{2}.$$

Let η^* denote the minimizer of $\varphi(\eta)$. If $w_{t-1,k,j} + \eta^* = 0$, then the subdifferential of $|w_{t-1,k,j} + \eta|$ at η^* is the set $\{\nu\colon \nu \in [-1, +1]\}$. Thus, in that case, the subdifferential $\partial\varphi(\eta^*)$ contains 0 iff there exists $\nu \in [-1, +1]$ such that

$$\beta_k\nu + \epsilon_{t-1,k,j} + \eta^*\Lambda^2 = 0 \Leftrightarrow w_{t-1,k,j}\Lambda^2 - \epsilon_{t-1,k,j} = \beta_k\nu.$$

The condition is therefore equivalent to

$$|w_{t-1,k,j}\Lambda^2 - \epsilon_{t-1,k,j}| \leq \beta_k.$$

If $w_{t-1,k,j} + \eta^* > 0$, then φ is differentiable at η^* and $\varphi'(\eta^*) = 0$, that is

$$\beta_k + \epsilon_{t-1,k,j} + \eta^*\Lambda^2 = 0 \Leftrightarrow \eta^* = \frac{1}{\Lambda^2}[-\beta_k - \epsilon_{t-1,k,j}].$$

In view of that expression, the condition $w_{t-1,k,j} + \eta^* > 0$ is equivalent to

$$w_{t-1,k,j}\Lambda^2 - \epsilon_{t-1,k,j} > \beta_k.$$

Similarly, if $w_{t-1,k,j} + \eta^* < 0$, φ is differentiable at η^* and $\varphi'(\eta^*) = 0$, which gives

$$\eta^* = \frac{1}{\Lambda^2}[\beta_k - \epsilon_{t-1,k,j}].$$

Figure 1 shows the pseudocode of our algorithm using the closed-form solution for the step size just presented.

D. Convergence analysis

In this section, we give convergence guarantees for both versions of the StructMaxent algorithm.

Theorem 3. *Let $(\mathbf{w}_t)_t$ be the sequence of parameter vectors generated by StructMaxent1 or StructMaxent2. Then, $(\mathbf{w}_t)_t$ converges to the optimal solution $\mathbf{w}^*$ of* (6).

Proof. We begin with the proof for StructMaxent2. Our proof is based on Lemma 19 of (Dudík et al., 2007), which implies that it suffices to show that $F(\mathbf{w}_t)$ admits a finite limit and that there exists a sequence $\mathbf{u}_t$ such that $R(\mathbf{u}_t, \mathbf{w}_t) \to 0$ as $t \to \infty$, where R is some auxiliary function. A function R is said to be *auxiliary* if

$$R(\mathbf{u}, \mathbf{w}) = I_C(\mathbf{u}) + \sum_{k=1}^{p}\beta_k\|\mathbf{w}_k\|_1 + \mathbf{w}\cdot\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}] + \mathbf{w}\cdot\mathbf{u} \\ + B(\mathbf{u} \parallel \mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}]),$$

where B is a Bregman divergences. We will use the Bregman divergence based on the squared difference:

$$B(\mathbf{u} \parallel \mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}]) = \frac{\|\mathbf{u} - \mathrm{E}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}]\|_2^2}{2\Lambda^2}.$$

Let $g_0(\mathbf{u}) = I_C(\mathbf{u}) + \mathbf{w}\cdot\mathbf{u}$ and observe that using the same arguments as in the proof of Theorem 1, we can write

$$g_0^*(\mathbf{r}) = \sup_{\mathbf{u}\in C}\big((\mathbf{r} - \mathbf{w})\cdot\mathbf{u} - I_C(\mathbf{u})\big) \\ = (\mathbf{r} - \mathbf{w})\cdot\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}] + \sum_{k=1}^{p}\beta_k\|\mathbf{r}_k - \mathbf{w}_k\|_1.$$

Similarly, if $f_0(\mathbf{u}) = B(\mathbf{u} \parallel \mathrm{E}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}])$, then

$$\begin{aligned} f_0^*(\mathbf{r}) &= \sup_{\mathbf{u}}(\mathbf{r}\cdot\mathbf{u} - B(\mathbf{u} \parallel \mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}])) \\ &= \frac{\Lambda^2\|\mathbf{r}\|_2}{2} + \mathbf{r}\cdot \mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}]. \end{aligned}$$

Therefore, applying Theorem 5 with $A = I$, we obtain

$$\begin{aligned} \inf_{\mathbf{u}} R(\mathbf{u}, \mathbf{w}_t) = \sup_{\mathbf{r}} \Big(&- \frac{\Lambda^2\|\mathbf{r}\|_2}{2} - \mathbf{r}\cdot \mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_t}}[\mathbf{\Phi}] - \mathbf{r}\cdot\mathop{\mathrm{E}}_{S}[\mathbf{\Phi}] \\ &+ \sum_{k=1}^{p} \beta_k(\|\mathbf{w}_{t,k}\|_1 - \|\mathbf{r}_k + \mathbf{w}_{t,k}\|_1)\Big), \end{aligned}$$

and we define $\mathbf{u}_t$ to be the solution of this optimization problem, which, in view of Theorem 5, does exist. We will now argue that $R(\mathbf{u}_t, \mathbf{w}_t) \to 0$ as $t \to \infty$. Note that

$$\begin{aligned} R(\mathbf{u}_t, \mathbf{w}_t) = -\sum_{k=1}^{p}\sum_{j=1}^{N_k} \inf_r \Big(&\frac{\lambda^2 r}{2} + r(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_t}}[\Phi_{k,j}] - \mathop{\mathrm{E}}_{S}[\Phi_{k,j}]) \\ &+ \beta_k|w_{t,k,j}| - \beta_k|r + w_{t,k,j}|\Big). \end{aligned}$$

Recall that, by definition of StructMaxent2, the following holds for all $(k, j) \in [1, p] \times [1, N_k]$:

$$\begin{aligned} &F(\mathbf{w}_t) - F(\mathbf{w}_{t+1}) \\ &\geq -\inf_r \Big(\frac{\Lambda^2 r}{2} + r(\mathop{\mathrm{E}}_{\mathsf{p}_{\mathbf{w}_t}}[\Phi_{k,j}] - \mathop{\mathrm{E}}_{S}[\Phi_{k,j}]) \\ &\qquad + \beta_k|w_{t,k,j}| - \beta_k|r + w_{t,k,j}|\Big) \\ &\geq 0, \end{aligned} \tag{22}$$

where the last inequality follows by taking $r = 0$. Therefore, to complete the proof, it suffices to show that $\lim_{t\to\infty} F(\mathbf{w}_t)$ is finite, since then $F(\mathbf{w}_t) - F(\mathbf{w}_{t+1}) \to 0$ and $R(\mathbf{u}_t, \mathbf{w}_t) \to 0$. By (22), $F(\mathbf{w}_t)$ is decreasing and it suffices to show that $F(\mathbf{w}_t)$ is bounded below. This is an immediate consequence of the feasibility of the optimization problem $\inf_\mathbf{w} F(\mathbf{w})$ which was established in Section 2.2 and the proof for StructMaxent2 is now complete.

The proof for StructMaxent1 requires the use a different Bregman divergence B defined as follows:

$$B(\mathbf{u} \parallel \mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}]) = \sum_{k=1}^{p}\sum_{j=1}^{N_k} D_0(\varphi_{kj}(\mathbf{u}) \parallel \varphi_{kj}(\mathop{\mathrm{E}}_{\mathsf{p}_\mathbf{w}}[\mathbf{\Phi}])),$$

where D_0 is unnormalized relative entropy, $\varphi_{kj}(\mathbf{u}) = ((\Lambda - u_{k,j}), (\Lambda + u_{k,j}))$ and $\|\mathbf{u}\|_\infty \leq \Lambda$. The rest of the argument remains the same. $\square$

E. Bounds on Rademacher complexities

In this section, we give the proof of the upper bounds on Rademacher complexities given in (15):

$$\begin{aligned} \mathfrak{R}_m(H_k^{\text{mono}}) &\leq \sqrt{\frac{2k\log d}{m}} \\ \mathfrak{R}_m(H_k^{\text{trees}}) &\leq \sqrt{\frac{(4k+2)\log_2(d+2)\log(m+1)}{m}}. \end{aligned}$$

The first inequality is an immediate consequence of Massart's lemma, which states that

$$\frac{1}{m}\mathop{\mathrm{E}}_{\sigma}\Big[\sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\Big] \leq \frac{r\sqrt{2\log|A|}}{m},$$

where $A \subset \mathbb{R}^n$ is a finite set, $r = \max_{\mathbf{x}\in A}\|\mathbf{x}\|_2$ and σ_is are Rademacher random variables. If we take A to be the image of the sample under H_k^{mono} then $|A| \leq |H_k^{\text{mono}}| \leq d^k$. Moreover, if the features in H_1^{mono} are normalized to belong to $[-1, 1]$ then $\Lambda = 1$ and $r = \sqrt{m}$. Combining these results with Massart's lemma leads to the desired bound.

Now we derive the second bound of (15). Since each binary decision tree in H_k^{trees}, can be viewed as a binary classifier, Massart's lemma yields that

$$\mathfrak{R}_m(H_k^{\text{trees}}) \leq \sqrt{\frac{2\log\Pi_{H_k^{\text{trees}}}(m)}{m}},$$

where $\Pi_{H_k^{\text{trees}}}(m)$ is the growth function of H_k^{trees}. We use Sauer's lemma to bound the growth function: $\Pi_{H_k^{\text{trees}}}(m) \leq (em)^{\text{VC-dim}(H_k^{\text{trees}})}$. For the family of binary decision trees in dimension d it is known that $\text{VC-dim}(H_k^{\text{trees}}) \leq (2k+1)\log_2(d+2)$ (Mansour, 1997) and the desired bound follows.