Supplementary Material
Multitask Learning without Label Correspondences

A Proof of lemma 2

Proof Denote by $\mathcal{B}$ a Banach space and let $\mathcal{B}^*$ be its dual. Denote space of conditional distributions $\mathcal{P} = \{p_{y|x} \mid p(y|x) \geq 0, \sum_{y\in\mathcal{Y}} p(y|x) = 1, \forall x \in \mathcal{X}, y \in \mathcal{Y}\}$. Let A be the conditional expectation operator of the feature map $\phi(x, y)$ with respect to conditional distribution $p(y|x)$, that is $Ap_{y|x} = \mathbf{E}_{y\sim p(y|x)}[\phi(x, y)]$. Fenchel's Duality [21, Theorem 4.4.3] states

$$\inf_{p_{y|x}\in\mathcal{P}} \{f(p_{y|x}) + g(Ap_{y|x})\} = \sup_{\theta\in\mathcal{B}^*} \{-f^*(A^*\theta) - g^*(-\theta)\}. \tag{12}$$

First, note that the adjoint of the linear operator A is $\left\langle Ap_{y|x}, \theta\right\rangle = \left\langle A^*\theta, p_{y|x}\right\rangle$, then we have $\left\langle \sum_{y\in\mathcal{Y}} p_{y|x}\phi(x, y), \theta\right\rangle = \sum_{y\in\mathcal{Y}} p_{y|x} \left\langle \phi(x, y), \theta\right\rangle$, thus $A^*\theta = \langle \phi(x, y), \theta\rangle$. Define $f(p_{y|x}) = p_{y|x} \log p_{y|x} + c \cdot p_{y|x} + \Lambda_{p_{y|x}}(\sum_{y\in\mathcal{Y}} p_{y|x} - 1)$ where c is the constant part w.r.t. $p_{y|x}$ (i.e. the gradient of the joint entropy), we the have $f^*(p^*_{y|x}) = \Lambda_{p^*_{y|x}} + \exp(p^*_{y|x} - 1 - c - \Lambda_{p^*_{y|x}})$ as its dual. Hence the dual of $\sum_{x\in\mathcal{X}} \left[-H(p_{y|x}) + \lambda \sum_{y\in\mathcal{Y}} g_y(x)p_{y|x}\right]$ is

$$\sum_{i=1}^{m} \left[\sum_y \exp(\langle \theta, \phi(x_i, y)\rangle - 1 - \lambda g_y(x_i) - \Lambda_{p^*_{y|x}}) + \Lambda_{p^*_{y|x}}\right] \tag{13}$$

Solving for optimality in $\Lambda_{p^*_{y|x}}$ gives $\sum_{i=1}^m \log \sum_y \exp(\langle \theta, \phi(x_i, y)\rangle - \lambda g_y(x_i))$. Similarly for $x' \in \mathcal{X}'$. The dual of the approximate moment matching constraint follows directly from [5, Lemma 6]. ■