Appendix

A.1 MMH: Max-margin Harmonium

For the special max-margin Harmonium (MMH), the learning problem is the same as defined in Section 3.1, and only several changes are needed to estimate parameters based on the general learning procedure. In this section, we present the necessary changes for learning MMH. For any other special cases of multi-view Markov networks, the learning can be similarly done.

With the definitions of local conditionals in Section 4, we can directly write the joint model distribution $p(\mathbf{x}, \mathbf{z}, \mathbf{h})$ based on the constructive definition and the marginal data likelihood $p(\mathbf{x}, \mathbf{z})$

$$p(\mathbf{x}, \mathbf{z}) \propto \exp \Big\{ \alpha^\top \mathbf{x} + \beta^\top \mathbf{z} - \frac{1}{2} \sum_j \frac{z_j^2}{\sigma_j^2} + \frac{1}{2} \sum_k (\mathbf{x}^\top \mathbf{W}_{\cdot k} + \mathbf{z}^\top \mathbf{U}_{\cdot k})^2 \Big\}.$$

Then, we use the contrastive divergence method and introduce two variational distribution q_0 and q_1. In this case, we can make a superficially simpler mean field assumption that $q(\mathbf{x}, \mathbf{z}, \mathbf{h}) = \prod_i q(x_i) \prod_j q(z_j) \prod_k q(h_k)$. Indeed, the general structured mean field assumption as made in Section 3.2 will lead to the same results, that is, a fully factorized form of $q(\mathbf{x})$, $q(\mathbf{z})$ and $q(\mathbf{h})$. Specifically, we have the following fully factorized update rules for posterior inference of q

$$\begin{aligned} q(\mathbf{x}) &= \prod_i q(x_i) = \prod_i p(x_i | \mathbb{E}_{q(\mathbf{H})}[\mathbf{H}]) \\ q(\mathbf{z}) &= \prod_j q(z_j) = \prod_j p(z_j | \mathbb{E}_{q(\mathbf{H})}[\mathbf{H}]) \\ q(\mathbf{h}) &= \prod_k q(h_k) = \prod_k p(h_k | \mathbb{E}_{q(\mathbf{X})}[\mathbf{X}], \mathbb{E}_{q(\mathbf{Z})}[\mathbf{Z}]). \end{aligned}$$

Similarly, (x_i, z_j) are clamped at their observed values for q_0, and only $q(h_k)$ is updated. The distribution q_1 is achieved by performing the above updates starting from q_0. Several iterations can yield a good q_1. After we have inferred q_0 and q_1, parameter estimation can be done by an alternating procedure as in Section 3.2. The first step of estimating $\mathbf{V}$ with Θ fixed is to learn a multi-class SVM, which is

$$\min_{\mathbf{V}} \;\; \frac{1}{2} C_1 \|\mathbf{V}\|_2^2 + C_2 \frac{1}{D} \sum_d \max_y [\Delta \ell_d(y) - \mathbf{V}^\top \mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\Delta \mathbf{f}_d(y)]].$$

Note that in this case, the latent representation (i.e., expectation of $\mathbf{H}$) is simply written as $\mathbb{E}_{p(\mathbf{h}|\mathbf{x},\mathbf{z})}[\mathbf{H}] = \mathbf{v}$, where $\mathbf{v}_k = \mathbf{x}^\top \mathbf{W}_{\cdot k} + \mathbf{z}^\top \mathbf{U}_{\cdot k}, \; \forall 1 \leq k \leq K$, when input data $\mathbf{x}$ and $\mathbf{z}$ are fully observed. If missing values exist in $\mathbf{x}$ or $\mathbf{z}$, the corresponding components are replaced with their expected values. Therefore, the prediction tasks (e.g., classification and retrieval) can be easily done in testing, as detailed in Section 3.2.

For the second step of estimating Θ, the sub-gradient is computed as

$$\nabla \alpha_i = -\mathbb{E}_{q_0}[x_i] + \mathbb{E}_{q_1}[x_i], \quad \nabla \beta_j = -\mathbb{E}_{q_0}[z_j] + \mathbb{E}_{q_1}[z_j], \quad \nabla(\sigma_k^{-1}) = -\mathbb{E}_{q_0}[z_k^2 \sigma^{-1}] + \mathbb{E}_{q_1}[z_k^2 \sigma^{-1}],$$

$$\nabla \mathbf{W}_{ik} = -\mathbb{E}_{q_0}[x_i h'_k] + \mathbb{E}_{q_1}[x_i h'_k] - C_2 \frac{1}{D} \sum_d (\mathbf{V}_{\bar{y}_d k} - \mathbf{V}_{y_d k}) \mathbb{E}_{q_0}[x_i]$$

$$\nabla \mathbf{U}_{jk} = -\mathbb{E}_{q_0}[z_j h'_k] + \mathbb{E}_{q_1}[z_j h'_k] - C_2 \frac{1}{D} \sum_d (\mathbf{V}_{\bar{y}_d k} - \mathbf{V}_{y_d k}) \mathbb{E}_{q_0}[z_j],$$

where $h'_k = \mathbf{x}^\top \mathbf{W}_{\cdot k} + \mathbf{z}^\top \mathbf{U}_{\cdot k}$ and $y_d = \arg\max_y [\Delta \ell_d(y) + \mathbf{V}^\top \mathbb{E}_{q_0}[\mathbf{f}(\mathbf{H}_d, y)]]$ is the *loss-augmented prediction*. Based on the definition of q_0, the expectations $\mathbb{E}_{q_0}[x_i]$ and $\mathbb{E}_{q_0}[z_j]$ are actually the count frequency of x_i and z_j, respectively.