

APPENDIX

A.1 MMH: Max-margin Harmonium

For the special max-margin Harmonium (MMH), the learning problem is the same as defined in Section 4.1.1, 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 3.1, 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 $p(\mathbf{x}, \mathbf{z})$

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

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 4.1.2 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 4.1.2, where the step of estimating \mathbf{V} with Θ fixed is to learn a multi-class SVM

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

Note that in this case, the latent representation (i.e., expectation of \mathbf{H}) is simply written as $\mathbb{E}_{p(\mathbf{h}|\mathbf{x}, \mathbf{z})}[\mathbf{h}] = \Upsilon$, where $\Upsilon_k = \mathbf{x}^\top \mathbf{W}_{\cdot k} + \mathbf{z}^\top \mathbf{U}_{\cdot k}$, $\forall 1 \leq k \leq K$, 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 4.4.

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

$$\begin{aligned} \partial \alpha_i &= -\mathbb{E}_{q_0}[x_i] + \mathbb{E}_{q_1}[x_i], \\ \partial \beta_j &= -\mathbb{E}_{q_0}[z_j] + \mathbb{E}_{q_1}[z_j], \\ \partial(\sigma_j^{-1}) &= -\mathbb{E}_{q_0}[z_j^2 \sigma_j^{-1}] + \mathbb{E}_{q_1}[z_j^2 \sigma_j^{-1}], \\ \partial \mathbf{W}_{ik} &= -\mathbb{E}_{q_0}[x_i h'_k] + \mathbb{E}_{q_1}[x_i h'_k] - C_2 \sum_d (\mathbf{V}_{y_d k} - \mathbf{V}_{\bar{y}_d k}) \mathbb{E}_{q_0}[x_i] \\ \partial \mathbf{U}_{jk} &= -\mathbb{E}_{q_0}[z_j h'_k] + \mathbb{E}_{q_1}[z_j h'_k] - C_2 \sum_d (\mathbf{V}_{y_d k} - \mathbf{V}_{\bar{y}_d k}) \mathbb{E}_{q_0}[z_j], \end{aligned}$$

where $h'_k = \mathbf{x}^\top \mathbf{W}_{\cdot k} + \mathbf{z}^\top \mathbf{U}_{\cdot k}$ and $\bar{y}_d = \arg \max_y [\Delta \ell_d(y) + \mathbf{V}^\top \mathbb{E}_{q_0}[\mathbf{f}(y, \mathbf{h}_d)]]$ 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.

A.2 Multi-view Latent Subspace MN for Modeling Paragraph Ordering Information

In this section, we formally define the structured multi-view latent subspace Markov network used in Section 5.4. Let \mathbf{x} be an $P \times N$ observation matrix, where P is the number of paragraphs in a document and N is the vocabulary size. Each row \mathbf{x}_p is a vector, of which the element $x_{pi} = 1$ if word i appears in paragraph p ; otherwise $x_{pi} = 0$. Each column $\mathbf{x}_{\cdot i}$ represents the appearance pattern of word i in all paragraphs. To consider the paragraph ordering information, we define a first-order Markov chain on each $\mathbf{x}_{\cdot i}$ while assuming that different $\mathbf{x}_{\cdot i}$'s are conditional independent. More formally, we define the factorial conditional distribution $p(\mathbf{x}|\mathbf{h}) = \prod_{i=1}^N p(\mathbf{x}_{\cdot i}|\mathbf{h})$, where each $p(\mathbf{x}_{\cdot i}|\mathbf{h})$ is a linear chain CRF [27]. This model is actually an N -view latent subspace MN, where the i th view has P variables $\{X_{pi}\}_{p=1}^P$ that are connected via a linear chain.

By the constructive definition as discussed in Section 3 (please see Eq. (1) and Eq. (2)), we need to specify the edge set for each view and define the feature functions. Specifically, let E denote the set of generalized edges (including both singleton vertices and pairwise edges) of a linear chain on each view¹³, then $E = \{(1), \dots, (P), (1, 2), \dots, (P-1, P)\}$, where (p) is a degenerate edge (i.e., the node p). Accordingly, the feature functions ϕ consists of a *singleton* feature function g that is defined on a single variable and 4 *pairwise* feature functions ϕ_0, ϕ_1, ϕ_2 and ϕ_3 that defined on a pair of variables. Mathematically, we define the singleton feature function as

$$g(x_{pi}) = x_{pi},$$

and we define the 4 pairwise feature functions as

$$\forall j = 0, \dots, 3 : \phi_j(x_{pi}, x_{p+1, i}) = \begin{cases} 1, & \text{if } 2x_{pi} + x_{p+1, i} = j \\ 0, & \text{otherwise} \end{cases}$$

We denote the corresponding weights by α and β_j , $j = 0, \dots, 3$. To make the model reasonably rich, we assume

13. By definition, all views have the same set of edges

different views have different feature weights, while within each view, the weights of these feature functions are shared by all edges. For notation simplicity, we define $g(\mathbf{x}_{.i}) \triangleq \sum_{p=1}^P g(x_{pi})$, which is the accumulated function value of g evaluated on the sequence $\mathbf{x}_{.i}$, and $\phi_j(\mathbf{x}_{.i}) \triangleq \sum_{p=1}^{P-1} \phi_j(x_{pi}, x_{p+1,i})$, which is again an accumulated function value of ϕ_j . Now, we define the interaction terms between \mathbf{X} and latent variables \mathbf{H} as

$$\sum_{i=1}^N (g(\mathbf{x}_{.i}) \mathbf{U}^i \mathbf{h} + \sum_j \phi_j(\mathbf{x}_{.i}) \mathbf{W}_j^i \mathbf{h}),$$

where \mathbf{W}_j^i and \mathbf{U}^i are K -dimensional real vectors. Finally, we include a quadratic energy term of the real variables \mathbf{H} in the exponent of the joint distribution.

Putting the above definitions together, we define the joint distribution

$$p(\mathbf{x}, \mathbf{h}) \propto \exp \left\{ \sum_i g(\mathbf{x}_{.i}) (\alpha^i + \mathbf{U}^i \mathbf{h}) + \sum_{ij} \phi_j(\mathbf{x}_{.i}) (\beta_j^i + \mathbf{W}_j^i \mathbf{h}) - \frac{1}{2} \mathbf{h}^\top \mathbf{h} \right\},$$

and we have the conditional distributions as

$$p(\mathbf{x}_{.i} | \mathbf{h}) \propto \exp \left\{ g(\mathbf{x}_{.i}) (\alpha^i + \mathbf{U}^i \mathbf{h}) + \sum_j \phi_j(\mathbf{x}_{.i}) (\beta_j^i + \mathbf{W}_j^i \mathbf{h}) \right\}$$

$$p(h_k | \mathbf{X}) = \mathcal{N}(h_k | \sum_{i=1}^N (g(\mathbf{x}_{.i}) \mathbf{U}_k^i + \sum_j \phi_j(\mathbf{x}_{.i}) \mathbf{W}_{jk}^i), 1).$$

Now, we can follow the procedure in Section 4 to perform parameter estimation and inference. Since each view is a linear-chain Markov network, we can perform inference with a forward-backward message passing scheme, which can be done in the same way as in [27]. We omit the details for brevity.

The message passing for each document is of a complexity $O(N \times P \times S^2)$, where S is the number of possible values for each variable X_{pi} . Since X_{pi} is binary, we have $S^2 = 4$ (a very small constant). Moreover, the number of paragraphs P is on average very small, e.g., the average P is 9 in the hotel review dataset. Therefore, the time complexity is linear in terms of the feature dimension N (i.e., the number of terms in the given vocabulary).

A.3 Additional Experimental Results

In this section, we present additional experimental results.

Fig. 11 shows the 5-bottom ranked (i.e., with small expectation values of H_k) images for each of the 5 topics, as presented in Fig. 3. For convenience, we also include the average probability of each category distributed on the particular topic, which is the same as that in Fig. 3.

Fig. 12 shows the average value of $\mathbb{E}[\mathbf{H}]$ discovered by a 60-topic MMH, together with the variance across the images from the 13 classes of animals. The variance of each $\mathbb{E}[H_k]$ indicates the discriminative power of topic k over all the 13-class images.

Table 2 shows the complete table which contains the average distributions over topics for all the three methods, i.e., MMH, TWH, and DWH.

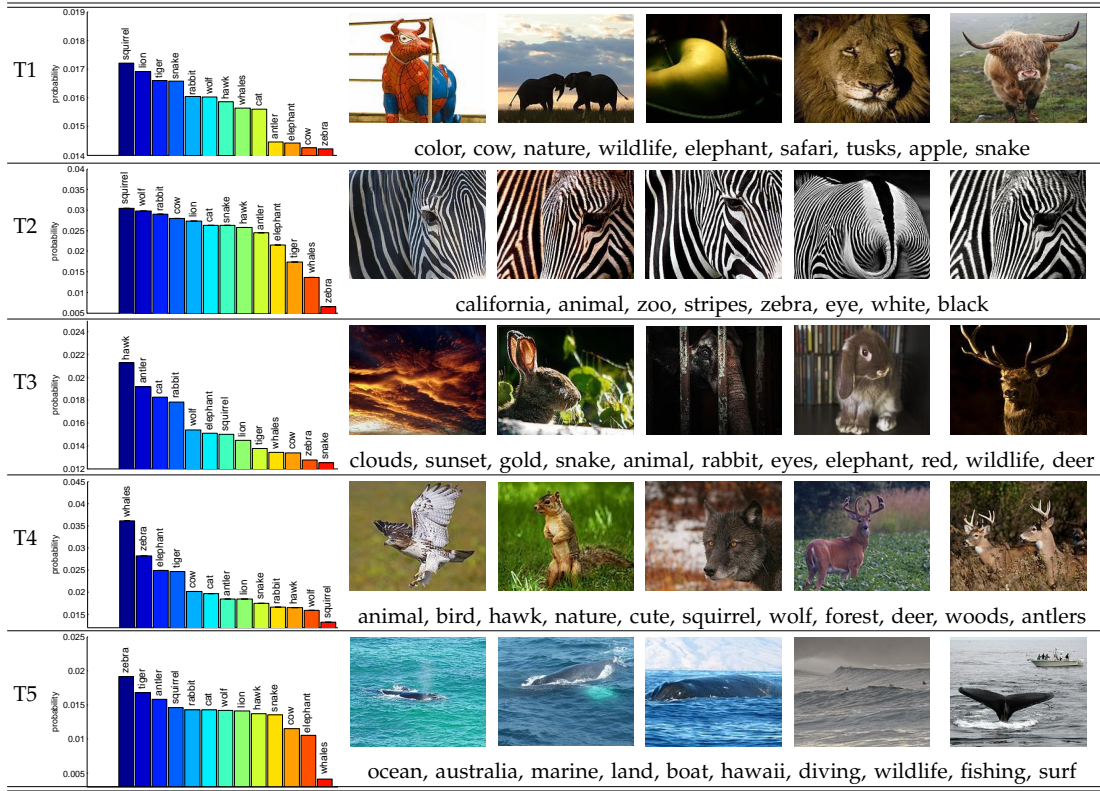


Fig. 11. Example topics discovered by the 60-topic MMH as in Fig. 3 on the Flickr animal dataset. For each topic, we show 5 bottom-ranked images as well as the average probabilities of that topic on representing images from the 13 categories.

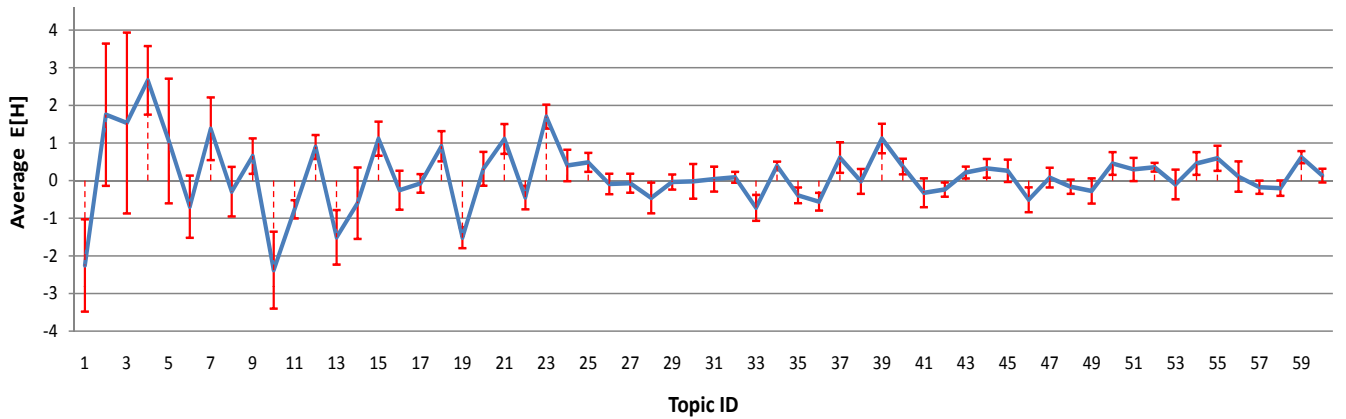
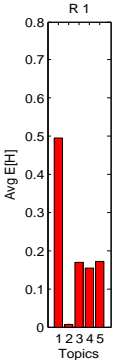
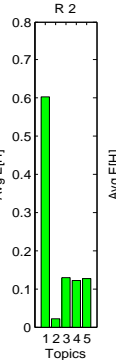
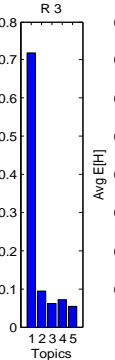
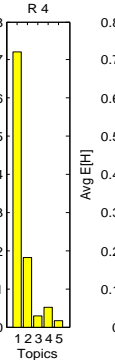

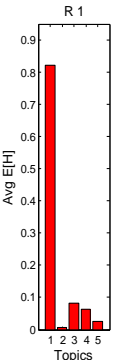
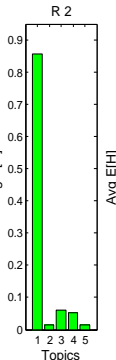
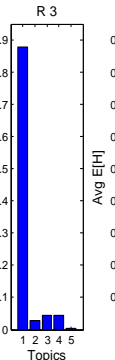
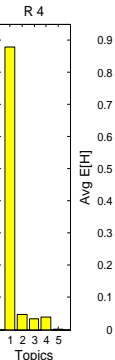
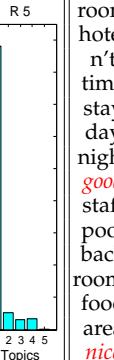
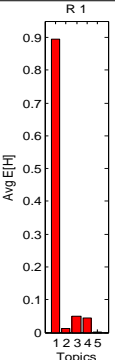
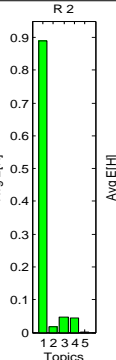
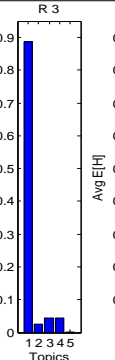
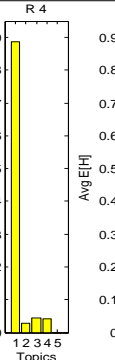
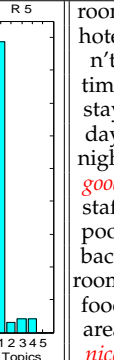


Fig. 12. Average value and variance of $E[H]$ discovered by a 60-topic MMH across 13class-animal Flickr images.

TABLE 2

Average distributions over the topics for documents with different rating scores by a 5-topic MMH, 5-topic TWH and 5-topic DWH.

Max Margin Harmonium (Avg-KL: 3.568)					
Average $\mathbb{E}_{p(h x,z)}[h]$ in 5-level Rating Score examples					
					
T1	T2	T3	T4	T5	
room	<i>great</i>	<i>small</i>	<i>worst</i>	<i>worst</i>	
hotel	<i>loved</i>	<i>worst</i>	<i>small</i>	<i>dirty</i>	
n't	arrived	<i>dirty</i>	<i>dirty</i>	<i>small</i>	
time	<i>enjoyed</i>	shower	shower	shower	
stay	<i>fantastic</i>	<i>broken</i>	<i>broken</i>	<i>broken</i>	
day	bit	smell	smell	smell	
night	<i>wonderful</i>	paying	paying	paying	
<i>good</i>	<i>lovely</i>	bathroom	bathroom	<i>poor</i>	
staff	pool	<i>poor</i>	<i>poor</i>	toilet	
pool	trip	toilet	toilet	refund	
back	beach	staying	refund	manager	
rooms	<i>fun</i>	refund	staying	bathroom	
food	<i>happy</i>	breakfast	walls	walls	
area	pools	hotel	hotel	carpet	
<i>nice</i>	<i>perfect</i>	walls	carpet	paid	
Tri-Wing Harmonium (Avg-KL: 0.045)					
Average $\mathbb{E}_{p(h x,z)}[h]$ in 5-level Rating Score examples					
					
T1	T2	T3	T4	T5	
room	beach	bathroom	resort	experience	
hotel	food	parking	beach	breakfast	
n't	pool	tv	ocean	stay	
time	<i>great</i>	area	trip	service	
stay	bar	coffee	vacation	<i>beautiful</i>	
day	resort	kitchen	desk	dinner	
night	restaurants	bed	check	guests	
<i>good</i>	drinks	street	time	made	
staff	restaurant	floor	front	<i>wonderful</i>	
pool	view	<i>large</i>	<i>great</i>	feel	
back	lunch	<i>small</i>	called	trip	
rooms	<i>good</i>	tub	call	house	
food	sea	<i>comfortable</i>	property	visit	
area	<i>beautiful</i>	location	<i>beautiful</i>	special	
<i>nice</i>	walk	internet	people	<i>comfortable</i>	
Dual-Wing Harmonium (Avg-KL: 0.038)					
Average $\mathbb{E}_{p(h x,z)}[h]$ in 5-level Rating Score examples					
					
T1	T2	T3	T4	T5	
room	beach	food	breakfast	belize	
hotel	food	told	reception	brett	
n't	pool	asked	bathroom	cam	
time	resort	holiday	bed	canapes	
stay	<i>great</i>	reception	shower	canoeing	
day	restaurants	day	holiday	caracol	
night	bar	bar	coffee	hosts	
<i>good</i>	drinks	staff	evening	nadege	
staff	restaurant	back	<i>small</i>	underway	
pool	lunch	manager	<i>clean</i>	wineries	
back	sea	people	bar	adopted	
rooms	<i>beautiful</i>	evening	hotel	amanda	
food	entertainment	entertainment	<i>good</i>	aurora	
area	pools	arrived	main	begun	
<i>nice</i>	view	hotel	tea	boasted	