Variational Information Planning Notes

Jason Pacheco

April 4, 2018

1 Information Theoretic Planning

Consider a model of latent variables x and conditionally independent observations $\mathcal{Y}_T = \{y_1, \ldots, y_T\}$. At each time t a discrete *action* $a_t \in \{1, \ldots, A\}$ parameterizes the likelihood, denoted $p_{a_t}(y_t \mid x)$. Given observations $\mathcal{Y}_T$ and actions $\mathcal{A}_T = \{a_1, \ldots, a_T\}$ the posterior is:

$$p(x \mid \mathcal{Y}_T; \mathcal{A}_T) \propto p(x) \prod_{t=1}^{T} p_{a_t}(y_t \mid x) \tag{1}$$

The choice of conditionally independent observations in the joint is for simplification. Our approach is easily extended to the case where nuisance variables must be integrated out, as shown in the labeled LDA example of Sec. 3. Given the posterior at stage $t-1$ information thoretic planning selects an action to maximize the posterior mutual information (MI):

$$\begin{aligned} a_t^* &= \arg\max_a \; I_a(X; Y_t \mid \mathcal{Y}_{t-1}) \\ &= \arg\max_a \; H(X \mid \mathcal{Y}_{t-1}) - H_a(X \mid Y_t, \mathcal{Y}_{t-1}) \end{aligned} \tag{2}$$

where $H_a(\cdot)$ denotes differential entropy under the hypothesized action a. Under this model the marginal entropy $H(X \mid \mathcal{Y})$ summarizes past observations and is thus invariant to the choice of action at time t. Having chosen action a_t^* the system draws a new observation from the corresponding likelihood: $y_t \sim p_{a_t^*}(\cdot \mid x)$. The posterior is then updated and the process is repeated. Note that this procedure can be interpreted as a general case of Bayesian sequential experiment design.

2 Variational Information Planning

Calculation of MI (2) is complicated by the posterior expectations required to calculate entropy. In this section we show a variational lower bound on MI that is maximized through moment-matching. We conclude with a brief description of a fully variational inference and planning algorithm.

2.1 Variational Information Bound

For any valid distribution $q \in \mathcal{Q}$ we have the following lower bound:

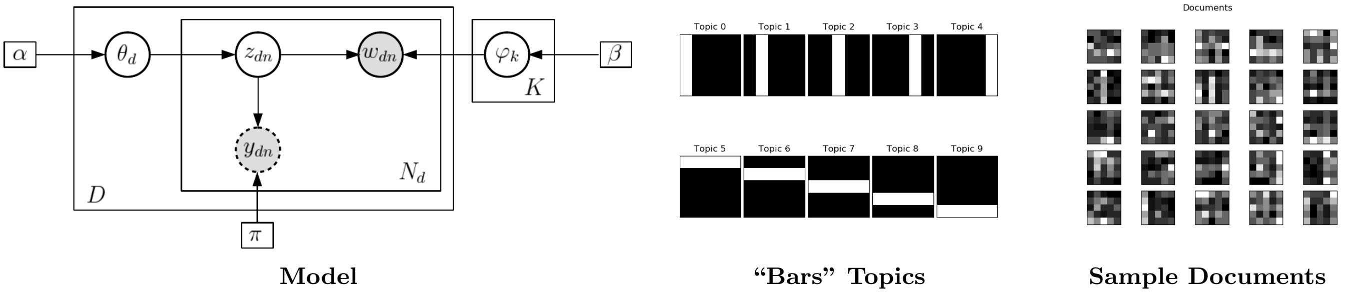
$$I(X; Y_t \mid \mathcal{Y}_{t-1}) \geq H(X \mid \mathcal{Y}_{t-1}) - H_p(q(X \mid Y_t)) \tag{3}$$

where the final term is the cross-entropy: $H_p(q(x \mid y)) = \langle -\log q(x \mid y) \rangle_{p(x,y)}$. The bound is a result of the nonnegativity of Kullback-Leibler divergence since $KL(p \| q) \geq 0$, then $H(p) \geq H_p(q)$ and is tight when $q(X \mid Y_t)$ equals the posterior $p(X \mid Y_t, \mathcal{Y}_{t-1})$. When $q \in \mathcal{Q}$ is an exponential family distribution the tightest bound is achieved via moment-matching. To see this we rewrite bound (3) as an equivalent constrained maximization over natural parameters:

$$\max_q H(x) + H_p(q(y)) - H_p(q(x, y)) \quad \text{such that} \quad q(y) = \int q(x, y) dx \tag{4}$$

For brevity we have dropped explicit conditioning on the observation set $\mathcal{Y}$ and time index subscripts t. The marginalization constraint ensures $q(x \mid y)$ is a valid conditional distribution. Rearranging the objective we have the equivalent minimization:

$$q^* = \arg\min_{q(x,y)} \mathrm{KL}(p(x, y) \| q(x, y)) - \arg\min_{q(y)} \mathrm{KL}(p(y) \| q(y)) \tag{5}$$



Model **"Bars" Topics** **Sample Documents**

Figure 1: **Labeled LDA** The graphical model of LLDA denotes semi-supervised annotation y_{dn} with a dashed line. Each word n of document d is assigned an annotation and at each planning stage we select among the remaining set of annotations. We use synthetic "bars" topics as introduced in Griffiths and Steyvers.

The above minimization is *equivalent* in the sense that q^* is the maximizer of (4) and the minimizer of (5), though their objective functions differ. However, the moment-matching property of exponential families ensures that (5) is a convex objective whose unique solution corresponds to matching the expected sufficient statistics of q to the corresponding moments of p. We choose $q(x, y)$ so that $q(y) = \int q(x, y)dx$ remains in the exponential family. This choice of $q(x, y)$ ensures that the solution to the unconstrained objective satisfies the marginalization constraints.

2.2 Algorithm Description

Suppose that at time $t-1$ we have a posterior approximation $q(x) \approx p(x \mid \mathcal{Y}_{t-1})$. For each hypothesized action $a_t \in \mathcal{A}$ we construct a local approximation of the joint distribution. The local approximation consists of the variational posterior, which summarizes previous observations, and the predicted future observation conditioned on the hypothesized action:

$$\hat{p}_{a_t}(x, y_t \mid \mathcal{Y}_{t-1}) \propto q(x)p_{a_t}(y_t \mid x) \tag{6}$$

Our variational planning objective maximizes the lower bound (3) of MI under the approximation $\hat{p}$,

$$\max_{a \in \mathcal{A}} I_{\hat{p}_a}(X; Y_t \mid \mathcal{Y}_{t-1}) \geq \max_{a \in \mathcal{A}} \max_{q_a \in \mathcal{Q}} H_{\hat{p}}(X \mid \mathcal{Y}_{t-1}) - H_{\hat{p}_a}(q_a(X \mid Y_t)). \tag{7}$$

As was shown in the previous section, the lower bound for each action can be found by matching moments of q_a and $\hat{p}$. This procedure of moment-matching to a local approximation is analogous to the steps in expectation propagation (EP). More specifically, the local approximation $\hat{p}$ is analogous to the *augmented distribution* in EP. The model requirement is that expected sufficient statistics under $\hat{p}$ can be calculated efficiently. Planning requires a total of $\mathcal{O}(|\mathcal{A}|)$ moment-matching operations, which can be done in parallel.

3 Labeled LDA Example

Labeled LDA (LLDA), introduced in Flaherty et al. (2005), extends LDA to include a semi-supervised label for each word in the corpus. The application considered in Flaherty et al. uses LLDA to infer latent groups of genes affected by various drugs. In that setting the annotation label assigns functional categories to individual genes, when they are known. In this way the LLDA model can discover interpretable topics. See Fig. 1 for the LLDA graphical model and notation we will reference below.

At stage $t-1$ of learning we observe a set of annotations $\mathcal{Y}_{t-1} = \{y_{d_1 n_1}, \ldots, y_{d_{t-1} n_{t-1}}\}$, which we treat as discrete in this example. We update our variational posterior approximation over K topics φ and document-topic proportions θ for D documents:

$$p(\theta, \varphi \mid \mathcal{Y}_{t-1}) \approx \prod_{d=1}^{D} q(\theta_d) \prod_{k=1}^{K} q(\varphi_k) \tag{8}$$

During stage t *planning* we select the index (d, n) which maximizes MI between topics φ and annotation Y_{dn} as in: $\max_{d,n} I(Y_{dn}; \varphi \mid \mathcal{Y}_{t-1})$. Unlike the simplef model (1), LLDA involves nuisance variables which must be marginalized during the planning stage. To generalize the algorithm we assume an EP-style variational approximation which are product distributions consisting of *messages* from each neighbor:

$$q(\varphi_k) = \text{Dirichlet}(\varphi_k \mid \beta + \sum_{d=1}^{D} \sum_{n=1}^{N} \chi_{kdn}), \qquad q(\theta_d) = \text{Dirichlet}(\theta_d \mid \alpha + \sum_{n=1}^{N} \xi_{dn}) \tag{9}$$

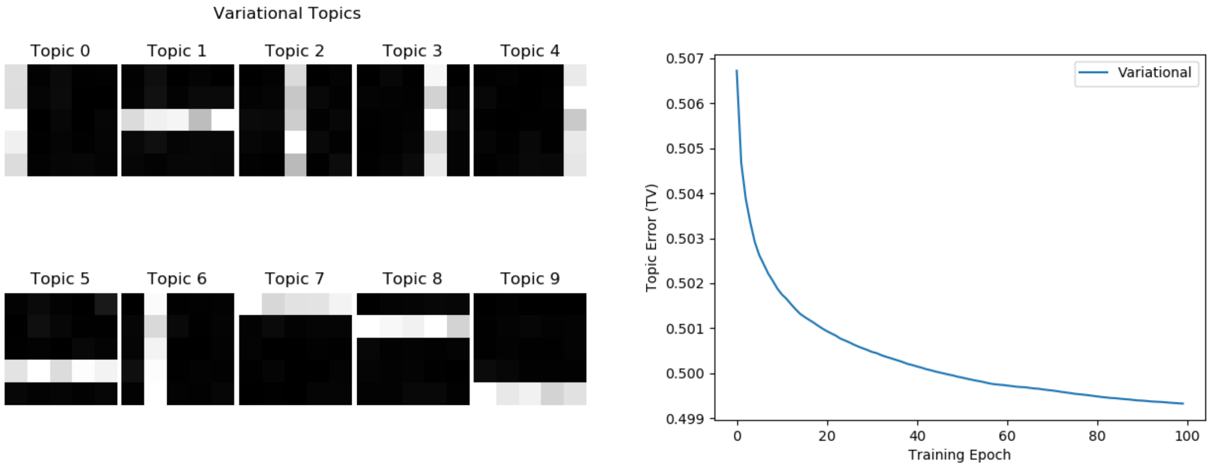


Figure 2: **Preliminary Results** *Left:* Posterior mean of inferred topics. *Right:* Total variation error of estimated topics. The posterior distribution concentrates on the correct topic ordering in the limit as all annotations are provided. As a result, the TV error thus penalizes incorrect orderings.

To form a local approximation we first remove the influence of the $(d,n)^{th}$ component by subtracting the corresponding natural statistic:

$$q_k^{\setminus dn}(\varphi_k) = \text{Dirichlet}(\varphi_k \mid \beta + \sum_{d=1}^{D} \sum_{n' \neq n} \chi_{kdn'}), \qquad q_d^{\setminus n}(\theta_d) = \text{Dirichlet}(\theta_d \mid \alpha + \sum_{n' \neq n} \xi_{dn'}) \tag{10}$$

The factor $q_k^{\setminus dn}(\varphi_k)$ represents the distribution on topic φ_k after removing influence of word n in document d. Similarly, $q_d^{\setminus n}(\theta_d)$ is the residual distribution on document-topic proportions. In EP these are sometimes referred to as the *cavity* distributions. The local approximation is thus a Dirichlet-Multinomial mixture:

$$\hat{p}(\varphi, y_{dn} \mid \mathcal{Y}_{t-1}) \propto \left[\prod_k q_k^{\setminus dn}(\varphi_k) \right] \sum_k \pi_k(y_{dn}) \varphi_k(w_{dn}) \mathbb{E}_{q_d^{\setminus n}} \left[\theta_d(k) \right] \tag{11}$$

$$\propto \sum_k C_k \pi_k(y_{dn}) \text{Dirichlet}(\varphi_k \mid \cdot) \tag{12}$$

Mixture weights C_k and Dirichlet parameters can be easily calculated, though we omit the details for brevity. For planning we approximate $\hat{p}$ with a single Dirichlet-Multinomial distribution via moment-matching. We can thus easily calculate the variational bound on MI used in planning.

As a preliminary experiment we fit LLDA to a simulated dataset of $D = 500$ documents, each with $N = 100$ words, and we use the *bars* dataset introduced in Griffiths and Steyvers. Specifically, when topics are arranged into a 5×5 square they place equal probability in a single column or row, and zero probability elsewhere. The annotation distribution assigns $y_{dn} = z_{dn}$ with probability proportional to $1 - \epsilon$ and ϵ otherwise. In this way annotations act as true topic assignments with high probability. We thus expect topic ordering to stabilize in the limit as all words are labeled, with the topic posterior concentrating on the correct ordering. Fig. 2 shows results after 100 training rounds.