Variational Inference

Christoph Teichmann Antoine Venant

November 29, 2017

Past Lectures and Today

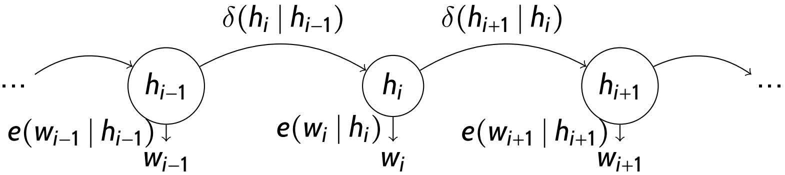
1. **General Principles of Bayesian Inference:** define a random quantity of interest $\rightarrow$ define a *joint* density of probability $\rightarrow$ condition on observed data to obtain a predictive *posterior* density.
2. **The Dirichlet-Multinomial model:** how to define prior densities over discrete (finite or countably infinite) probability distributions.
3. **MCMC Methods:** how to *Sample* from (and compute expected values under) the posterior distribution when direct computation of the posterior density is not directly feasible.

Past Lectures and Today

1. **General Principles of Bayesian Inference:** define a random quantity of interest $\rightarrow$ define a *joint* density of probability $\rightarrow$ condition on observed data to obtain a predictive *posterior* density.
2. **The Dirichlet-Multinomial model:** how to define prior densities over discrete (finite or countably infinite) probability distributions.
3. **MCMC Methods:** how to *Sample* from (and compute expected values under) the posterior distribution when direct computation of the posterior density is not directly feasible.
4. **Variational Inference:** *Approximate* the posterior distribution.

Problem Reminder

- Recall the HMM language model from last session:



- (observed) words $w_i \in L$. hidden tags / latent variables $h_i \in H$.
- Transition probabilities $\delta(h_{i+1} \mid h_i)$ from hidden states to hidden states.
- Emission probabilities $e(w_i \mid h_i)$ in every hidden states.
- **prior densities $p_0(\delta(\cdot \mid h))$ (over probability vectors over H) for every h.**
- **prior densities $p_0(e(\cdot \mid h))$ (over probability vectors over L) for every h.**

Problem Reminder (cont'd)

Inference

After observing the sequence of words $\mathbf{w} = w_0 \dots w_n$, what are the posterior densities over transitions and emission probabilities?

- Assume for simplicity $w_0 = \texttt{start}$, $t_0 = \texttt{<S>}$ with prob. 1.

$$p(\langle \delta(\cdot \mid h) \rangle_{h \in H}, \langle e(\cdot \mid h)_{h \in H} \rangle \mid \mathbf{w}) = \frac{p(\langle \delta(\cdot \mid h) \rangle_{h \in H}, \langle e(\cdot \mid h)_{h \in H} \rangle, \mathbf{w})}{p(\mathbf{w})}$$

$$= \frac{\prod_{h \in H} p_0(\delta(\cdot \mid h)) \times p_0(e(\cdot \mid h)) \times \overbrace{\sum_{\mathbf{h_0 \dots h_n}} \prod_{i=1}^{n} \delta(h_i \mid h_{i-1}) \times e(w_i \mid h_i)}^{\text{computable in } O(n|H|^2) \text{ (forward-backward algo.)}}}{\underbrace{p(\mathbf{w})}_{\text{Expensive computation: marginalize twice over } |H|-1 \text{ simplex.}}}.$$

More generally

- $\mathbf{Z}$ random variable describing latent variables.
- $\mathbf{X}$ random variable describing observed events.
- Joint density $p(\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \overbrace{p(\mathbf{Z} = \mathbf{z})}^{\text{prior}} \times p(\mathbf{X} = \mathbf{X} \mid \mathbf{Z} = \mathbf{z})$.
- We're interested in posterior density $p(\mathbf{Z} = \mathbf{z} \mid \mathbf{X} = \mathbf{x}) = \frac{p(\mathbf{Z}=z, \mathbf{X}=\mathbf{x})}{p(\mathbf{X}=\mathbf{x})}$. But too expensive to compute (in particular $p(\mathbf{X} = \mathbf{x})$).
- Last time: find way to sample without explicit computation.
- Today, *variational inference*: find $q^*(\mathbf{Z} = \mathbf{z})$ the best approximation of $p(\mathbf{Z} = \mathbf{z} \mid \mathbf{X} = \mathbf{x})$ over a family of probability densities $\mathcal{Q}$.

Why another inference technique?

- Metropolis-Hastings guarantees convergence in probability, but convergence time might be very slow (random walk effect).
- Variational inference generally faster but yield approximate distribution.
- Hence variational inference can be useful to quickly evaluate a wide range of model over large data.
- Sometimes Gibbs Sampling not possible, MCMC methods not straightforwardly usable.

Variational Inference

1. Define a set of probability densities over latent variables $\mathcal{Q}$ (in pratice $\mathcal{Q} = \{q_\theta(\mathbf{Z}) \mid \theta \in \Theta\}$, θ vector of so called *variational parameters*).
2. Search for $q^* \in \mathcal{Q}$ s.t. q^* mimizes the Kullback-Leibler divergence to $p(\mathbf{Z} \mid \mathbf{x})$.

$$q^* = argmin_{q \in \mathcal{Q}} KL(q(\mathbf{Z}) \,||\, p(\mathbf{Z} \mid \mathbf{x}))$$

KL divergence

$$\begin{aligned} KL(p_1(\mathbf{Z}) \,||\, p_2(\mathbf{Z})) &\triangleq \int p_1(\mathbf{z})(log(p_1(\mathbf{z})) - log(p_2(\mathbf{z})))d\mathbf{z} \\ &= \mathbb{E}_{p_1}(log(p_1(\mathbf{Z}))) - \mathbb{E}_{p_1}(log(p_2(\mathbf{Z}))). \end{aligned}$$

- Information theoretic quantity.
- is 0 only when densities are equal.
- is always positive.

Evidence Lower Bound

- $KL(q(\mathbf{Z}) \,||\, p(\mathbf{Z} \mid \mathbf{x}) = \mathbb{E}_q(log(q(\mathbf{Z}))) - \mathbb{E}_q(log(p(\mathbf{Z} \mid \mathbf{x})))$ depends on $p(\mathbf{Z} \mid \mathbf{x})$ which we don't know how to compute.
- $KL(q(\mathbf{Z}) \,||\, p(\mathbf{Z} \mid \mathbf{x}) = \mathbb{E}_q(log(q(\mathbf{Z}))) - \mathbb{E}_q(log(p(\mathbf{Z}, \mathbf{X} = \mathbf{x}))) + log(p(\mathbf{x}))$ (Exercise).
- We can mimize instead $\mathbb{E}_q(log(q(\mathbf{Z}))) - \mathbb{E}_q(log(p(\mathbf{Z}, \mathbf{X} = \mathbf{x}))$, or equivalently maximize

$$\begin{aligned} elb(q) &= \mathbb{E}_q(log(p(\mathbf{Z}, \mathbf{X} = \mathbf{x}))) - \mathbb{E}_q(log(q(\mathbf{Z}))) \\ &= \mathbb{E}_q(log(p(\mathbf{X} = \mathbf{x} \mid \mathbf{Z}))) - KL(q(\mathbf{Z}) \,||\, p(\mathbf{Z})) \end{aligned}$$

(Exercise: proove this).

Evidence Lower Bound (cont'd)

$$elb(q) = \mathbb{E}_q(log(p(\mathbf{X} = \mathbf{x} \mid \mathbf{Z}))) - KL(q(\mathbf{Z}) \,||\, p(\mathbf{Z}))$$

- Does not depend on the normalization factor $p(\mathbf{x})$ anymore!
- $elb(q) \leq log(p(\mathbf{x}))$ (Exercise).
- But what should $\mathcal{Q}$ look like? How do we find optimal q^*?

Mean-field Variational Inference

- Assume $\mathbf{Z} = \langle Z_1, \ldots, Z_n \rangle$.
- Simplifying assumption: let $\mathcal{Q}$ be such that latent variables Z_i and Z_j are independant under every $q \in \mathcal{Q}$.
- $\mathcal{Q} = \prod_{i=1}^{n} \mathcal{Q}_i$, for every $q = \langle q_1, \ldots, q_n \rangle \in \mathcal{Q}$

$$q(Z_1 = z_1, \ldots, Z_n = z_n) = \prod_{i=1}^{n} q_i(Z_i = z_i).$$

- This is known has the *mean-field variational family*.
- Idea: can approximate marginals $p(Z_i \mid x)$ closely, but won't account for dependence of the latent variables on one another under the *true* joint posterior $p(\mathbf{Z} \mid x)$.

Optimization

Recall Gibbs Sampling from last session:

Optimization

Recall Gibbs Sampling from last session:

- From current state $\langle z_1^{t+1}, \ldots, z_{i-1}^{t+1}, z_i^t, \ldots, z_n^t \rangle$.

Optimization

Recall Gibbs Sampling from last session:

- From current state $\langle z_1^{t+1}, \ldots, z_{i-1}^{t+1}, z_i^t, \ldots, z_n^t \rangle$.
- Fix $\mathbf{z}_{\neg i} = \langle z_1^{t+1}, \ldots, z_{i-1}^{t+1}, z_{i+1}^t, \ldots, z_n^t \rangle$.

Optimization

Recall Gibbs Sampling from last session:

- From current state $\langle z_1^{t+1}, \ldots, z_{i-1}^{t+1}, z_i^t, \ldots, z_n^t \rangle$.
- Fix $\mathbf{z}_{\neg i} = \langle z_1^{t+1}, \ldots, z_{i-1}^{t+1}, z_{i+1}^t, \ldots, z_n^t \rangle$.
- Sample z_i^{t+1} from conditional distribution $p(z^{t+1} \mid \mathbf{z}_{\neg i}, x)$.

Optimization

Recall Gibbs Sampling from last session:

- From current state $\langle z_1^{t+1}, \ldots, z_{i-1}^{t+1}, z_i^t, \ldots, z_n^t \rangle$.
- Fix $\mathbf{z}_{\neg i} = \langle z_1^{t+1}, \ldots, z_{i-1}^{t+1}, z_{i+1}^t, \ldots, z_n^t \rangle$.
- Sample z_i^{t+1} from conditional distribution $p(z^{t+1} \mid \mathbf{z}_{\neg i}, x)$.

Successive (manageable) coordinate updates yield new samples!

Optimization (cont'd)

Coordinate Ascent Mean-field V.I.

To find a (local) optimum $q^* = \langle q_1^*, \ldots, q_n^* \rangle \in \mathcal{Q}$:

Optimization (cont'd)

Coordinate Ascent Mean-field V.I.

To find a (local) optimum $q^* = \langle q_1^*, \ldots, q_n^* \rangle \in \mathcal{Q}$:

- Assume approximation after step t: $q_0^t = \langle q_1^t, \ldots q_n^t \rangle$.

Optimization (cont'd)

Coordinate Ascent Mean-field V.I.

To find a (local) optimum $q^* = \langle q_1^*, \ldots, q_n^* \rangle \in \mathcal{Q}$:

- Assume approximation after step t: $q_0^t = \langle q_1^t, \ldots q_n^t \rangle$.
- Update coordinate $1, \ldots, n$ successively.

Optimization (cont'd)

Coordinate Ascent Mean-field V.I.

To find a (local) optimum $q^* = \langle q_1^*, \ldots, q_n^* \rangle \in \mathcal{Q}$:

- Assume approximation after step t: $q_0^t = \langle q_1^t, \ldots q_n^t \rangle$.
- Update coordinate $1, \ldots, n$ successively.
- If coordinate $1, \ldots, i-1$ have been updated:

$$q_{i-1}^t = \langle q_1^{t+1}, \ldots, q_{i-1}^{t+1}, q_i^t, q_{i+1}^t, \ldots, q_n^t \rangle$$

.

Optimization (cont'd)

Coordinate Ascent Mean-field V.I.

To find a (local) optimum $q^* = \langle q_1^*, \ldots, q_n^* \rangle \in \mathcal{Q}$:

- Assume approximation after step t: $q_0^t = \langle q_1^t, \ldots q_n^t \rangle$.
- Update coordinate $1, \ldots, n$ successively.
- If coordinate $1, \ldots, i-1$ have been updated:

$$q_{i-1}^t = \langle q_1^{t+1}, \ldots, q_{i-1}^{t+1}, q_i^t, q_{i+1}^t, \ldots, q_n^t \rangle$$

.

- Then update coordinate i following

$$q_i^{t+1} = argmax_{q_i' \in \mathcal{Q}_i} elb(\langle q_1^{t+1}, \ldots, q_{i-1}^{t+1}, q_i', q_{i+1}^t, \ldots, q_n^t \rangle)$$

.

Optimization (cont'd)

Coordinate Ascent Mean-field V.I.

To find a (local) optimum $q^* = \langle q_1^*, \ldots, q_n^* \rangle \in \mathcal{Q}$:

- Assume approximation after step t: $q_0^t = \langle q_1^t, \ldots q_n^t \rangle$.
- Update coordinate $1, \ldots, n$ successively.
- If coordinate $1, \ldots, i-1$ have been updated:

$$q_{i-1}^t = \langle q_1^{t+1}, \ldots, q_{i-1}^{t+1}, q_i^t, q_{i+1}^t, \ldots, q_n^t \rangle$$

.

- Then update coordinate i following

$$q_i^{t+1} = argmax_{q_i' \in \mathcal{Q}_i} elb(\langle q_1^{t+1}, \ldots, q_{i-1}^{t+1}, q_i', q_{i+1}^t, \ldots, q_n^t \rangle)$$

.

Successive (manageable) coordinate updates yield refined approximations!

Update Rule

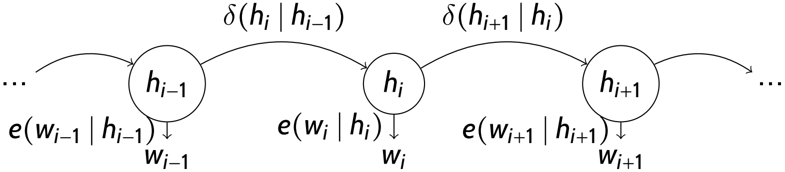
Update rule

- How find $q_i^{t+1} = argmax_{q_i' \in \mathcal{Q}_i} elb(\langle q_1^{t+1}, \dots, q_{i-1}^{t+1}, q_i', \dots, q_n^t \rangle)$?
- (depending on time) we admit the following result:

$$log(q_i^{t+1}(Z_i = z_i)) = \frac{\overbrace{\mathbb{E}_{q_{\neg i}}(log(p(Z_i = z_i, \mathbf{Z}_{\neg \mathbf{i}} = \mathbf{z}_{\neg \mathbf{i}}, \mathbf{x})))}^{\text{Will generally decompose over the } q_i}}{\underbrace{\int_z \mathbb{E}_{q_{\neg i}}(log(p(Z_i = z \mid \mathbf{z}_{\neg \mathbf{i}}, \mathbf{x})))dz}_{\text{Summation over one coordinate only}}}$$

.

Back to the HMM example



- We let priors follow Dirichlet distributions:

$$p_O(\delta(\cdot \mid h)) = \frac{\prod_{h' \in H} \delta(h' \mid h)^{\alpha_h^{h'}}}{B(\alpha_{\mathbf{h}}^{\delta})} \qquad p_O(e(\cdot \mid h)) = \frac{\prod_{w \in L} \delta(w \mid h)^{\alpha_h^{w}}}{B(\alpha_{\mathbf{h}}^{\mathbf{e}})}$$

with $\alpha_h^{\delta} = \langle \alpha_h^{h'} \rangle_{h' \in H}$ and $\alpha_h^{e} = \langle \alpha_h^{w} \rangle_{w \in L}$.