Homework 5

Assigned: 24/10/2011.
Due: 4/11/2011.

Exercise 1. Blahut-Arimoto Algorithm

In this exercise, you will analyze a special case of the EM algorith, which can be shown to converge to a unique optimum. This variant is older than EM itself, and is known as Blahut-Arimoto algorithm. You can find more information on the role of this algorithm for rate distortion theory in the textbook by Cover and Thomas.

Recall mixtures models from the lecture. For latent variable x, observed variable y (both discrete with finite range), the log marginal likelihood is

$$\log P(y = \tilde{y}) = \log \sum_{x=1}^{K} \pi_x P(y = \tilde{y}|x).$$

In the following, we write $P(\tilde{y})$ instead of $P(y = \tilde{y})$, but keep in mind the difference between y (a variable) and $\tilde{y}$ (a fixed observed value for it). For the purpose of this exercise, we will work with this single datapoint $\tilde{y}$, but all you'll show here holds for i.i.d. datasets just as well. In this exercise, assume that $P(\tilde{y}|x)$ is fixed and known, and the goal is maximize $\log P(\tilde{y})$ w.r.t. the distribution $P(x) = (\pi_x)$.

If (p_i), (q_i) are distributions over a finite set, the relative entropy is defined as

$$\mathrm{D}[\boldsymbol{q} \,\|\, \boldsymbol{p}] := \sum_i q_i \log \frac{q_i}{p_i}.$$

The function $(\boldsymbol{q}, \boldsymbol{p}) \mapsto \mathrm{D}[\boldsymbol{q} \,\|\, \boldsymbol{p}]$ is convex, although you don't need this fact here. Given two convex sets of distributions $\mathcal{P}$, $\mathcal{Q}$, Csiszár and Tusnády showed that the minimum

$$\mathrm{D}[\boldsymbol{q}_* \,\|\, \boldsymbol{p}_*] = \min_{\boldsymbol{p} \in \mathcal{P}} \min_{\boldsymbol{q} \in \mathcal{Q}} \mathrm{D}[\boldsymbol{q} \,\|\, \boldsymbol{p}]$$

is found by a simple *alternating minimization* algorithm: $\boldsymbol{q}_{j+1} = \mathrm{argmin}_{\boldsymbol{q}} \mathrm{D}[\boldsymbol{q} \,\|\, \boldsymbol{p}_j]$, $\boldsymbol{p}_{j+1} = \mathrm{argmin}_{\boldsymbol{p}} \mathrm{D}[\boldsymbol{q}_{j+1} \,\|\, \boldsymbol{p}]$, $j = 1, 2, \ldots$.

1. **[4 points]** Bring the EM problem $\min_{(\pi_x)}(-\log P(\tilde{y}))$ into the form of the alternating minimization algorithm. What is $\mathcal{P}$? What is $\mathcal{Q}$?
   Hint: Use the log partition function bound from the lecture. Use families $\mathcal{P}$, $\mathcal{Q}$ of *joint* distributions over (x, y), try the form $q_{(x,y)} = Q(x)\mathrm{I}_{\{y=\tilde{y}\}}$.

2. **[2 points]** Show that the sets $\mathcal{P}$, $\mathcal{Q}$ you determined above (for the EM problem) are both convex.

3. **[2 points]** Show that with your re-formulation of the EM problem, the E step update is equivalent to $\boldsymbol{q}_{j+1} = \mathrm{argmin}_{\boldsymbol{q}} \mathrm{D}[\boldsymbol{q} \,\|\, \boldsymbol{p}_j]$, and the M step update is equivalent to $\boldsymbol{p}_{j+1} = \mathrm{argmin}_{\boldsymbol{p}} \mathrm{D}[\boldsymbol{q}_{j+1} \,\|\, \boldsymbol{p}]$.
   This means that in this restricted case, EM is just alternating minimization between two convex sets, therefore converges to a unique minimum.

4. **[2 points]** Prove the information inequality: for any two distributions $\boldsymbol{p}$, $\boldsymbol{q}$, the relative entropy $\mathrm{D}[\boldsymbol{q} \,\|\, \boldsymbol{p}]$ is always nonnegative. Moreover, $\mathrm{D}[\boldsymbol{q} \,\|\, \boldsymbol{p}] = 0$ if and only if $q_i = p_i$ for all i.
   Hint: Use Jensen's inequality, in much the same way as in the lecture.

Exercise 2. Rauch-Tung-Striebel Smoother

In this exercise, you will work out the RTS smoother, based on an alternative view of HMM inference. Recall the definition of an HMM, based on observation probabilities $P(\boldsymbol{y}|\boldsymbol{x})$ and transition probabilities $P(\boldsymbol{x}_{\cdot}|\boldsymbol{x}_{\cdot-1})$. Given data $D = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$, recall the difference between filtering (computing $P(\boldsymbol{x}_t|\boldsymbol{y}_{\leq t})$, where $\boldsymbol{y}_{\leq t} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t\}$) and smoothing (computing the marginal posteriors $P(\boldsymbol{x}_t|D)$). Since filtering is important per se (for online prediction, given an incoming data stream), most smoothing algorithms make use of filtering code as far as possible. An example is the two-filter smoother discussed in the lecture. The RTS smoother is different, in that the backward pass is simpler, and the data is not required anymore. On the other hand, since the backward pass can start only once the forward (filtering) pass has been completed, the RTS smoother is less suitable for parallel hardware.

1. **[2 points]** Let us first consider a general HMM, not specifying the CPT types. You've learned about belief propagation in the lecture. For a chain, compute all messages, then combine them at each node. For LDS models, BP corresponds to the two-filter smoother. In this exercise, you will work out an alternative to BP (for Markov chains), that will lead to RTS smoothing.
   Suppose we have determined $P(\boldsymbol{x}_t|\boldsymbol{y}_{\leq t})$, $t = 1, \ldots, T$, in a forward pass. Instead of running an equivalent backward pass, we seek a recursion for the marginals $P(\boldsymbol{x}_t|D)$ directly. The start is easy: $P(\boldsymbol{x}_T|D) = P(\boldsymbol{x}_T|\boldsymbol{y}_{\leq T})$. Now assume that we have determined $P(\boldsymbol{x}_t|D)$, $t > 1$. Develop a recursion of the form

$$P(\boldsymbol{x}_{t-1}|D) = \int f_t(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t) P(\boldsymbol{x}_t|D)\, d\boldsymbol{x}_t$$

   (if $\boldsymbol{x}$ is discrete, then $\int$ becomes $\sum$), so that f_t can be computed easily (by local computations that do not need any further global propagation) from the filtering distributions $P(\boldsymbol{x}_{t'}|\boldsymbol{y}_{\leq t'})$ we have already determined. State what f_t is, and how to compute it.
   Hint: Draw the local part of the model at $\boldsymbol{x}_{t-1}$, $\boldsymbol{x}_t$. Look for conditional independencies that help you. It is helpful to think of $P(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{\leq (t-1)})$ as a "prior" in this context (state why).

2. **[4 points]** The RTS smoother is what you get if you apply this idea to the LDS model discussed in the lecture. Develop the RTS backward pass equations (for $P(\boldsymbol{x}_t|D)$, $t = 1, \ldots, T$), given that you have the Kalman filter outcomes available. Use the normal form (moment parameters), it is simpler this way. Do state the final recursion of means and covariance matrices.
   Hint: Yes, you need those Gaussian formulae. $f_t(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)$ above will be some simplification of $P(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, D)$. Work out its mean and covariance from the local filtering distributions' moments, and recall that the former is affine, the latter independent of $\boldsymbol{x}_t$. Then, average over $P(\boldsymbol{x}_t|D)$.