STOCHSEQ: DERIVATIONS

KEVIN EMMETT

1. Notation

1.1. Model Inputs.

- p : forward bias of random walk
- e : base-call error rate
- L : length of input DNA sequence
- N : number of i.i.d. reads of input DNA sequence
- d : number of output symbols (DNA sequence: $d = 4$)

1.2. Data.

- $\mathbf{S} = \{s_1, \dots, s_L\}$: input DNA sequence.
  - $s_l \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$: base at position l of input DNA sequence.
- $\mathbf{Z} = \{\mathbf{Z}^1, \dots, \mathbf{Z}^N\}$: set of N latent state sequences.
  - T_n : length of read sequence n
  - $\mathbf{Z}^n = \mathbf{z}^n_{1:T_n} = \{z^n_1, \dots, z^n_{T_n}\}$: latent state sequence n
  - $z^n_t \in \{1, \dots, L\}$: latent state at position t of read sequence n.
- $\mathbf{X} = \{\mathbf{X}^1, \dots, \mathbf{X}^N\}$: set of N observed state sequences.
  - $\mathbf{X}^n = \mathbf{x}^n_{1:T_n} = \{x^n_1, \dots, x^n_{T_n}\}$: observed base sequence n.
  - $x^n_t \in \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$: observed base at position t of read sequence n.

1.3. Parameters.

- $\mathbf{\Theta} = (\mathbf{\Pi^i}, \mathbf{\Pi^f}, \mathbf{A}, \mathbf{\Sigma}, \mathbf{\Omega}, \mathbf{\Phi})$: set of model parameters.
- ${\mathbf{\Pi^i}}^n = p(\mathbf{z}^n_1)$: initial state vector for sequence n.
- ${\mathbf{\Pi^f}}^n = p(\mathbf{z}^n_{T_n})$: final state vector for sequence n.
- $\mathbf{A}^n_t = p(\mathbf{z}^n_t | \mathbf{z}^n_{t-1})$: state transition matrix. Can be time-independent or time-dependent.
- $\mathbf{\Sigma}$: sequence estimate matrix
  - $\Sigma_{ld} = p(S_l = d)$
- $\mathbf{\Omega}^n$: forward transition vector for sequence n.
  - $\Omega^n_t = p(z_t = z_{t-1} + 1)$
- $\mathbf{\Phi}^n$: read error vector for sequence n.
  - $\Phi^n_t = p(\phi^n_t = 1)$

2. Probability Model

The complete-data likelihood is written as

$$p(\mathbf{X}|\mathbf{\Theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\mathbf{\Theta}). \tag{2.1}$$

Date: February 8, 2013.

Assuming a set of N i.i.d. read sequences, the likelihood factorizes as

$$p(\mathbf{X}|\boldsymbol{\Theta}) = \prod_{n=1}^{N} p(\mathbf{X}^n|\boldsymbol{\Theta}^n) = \prod_{n=1}^{N} \sum_{\mathbf{Z}^n} p(\mathbf{X}^n, \mathbf{Z}^n|\boldsymbol{\Theta}^n). \tag{2.2}$$

Factorizing the total likelihood in this way allows us to perform EM updates on the parameters of the individual read sequences, with an additional step combining an update on $\boldsymbol{\Sigma}$ across all reads. We use the conditional independence properties of a first order Markov chain to factorize the likelihood for each read as

$$p(\mathbf{X}^n, \mathbf{Z}^n|\boldsymbol{\Theta}^n) = p(\mathbf{z}_1^n)p(\mathbf{x}_1^n|\mathbf{z}_1^n) \prod_{t=2}^{T_n} p(\mathbf{z}_t^n|\mathbf{z}_{t-1}^n)p(\mathbf{x}_t^n|\mathbf{z}_t^n). \tag{2.3}$$

Each of these terms can be represented using the set of model parameters $\boldsymbol{\Theta}^n = (\boldsymbol{\Pi^{\mathbf{i}}}^n, \boldsymbol{\Pi^{\mathbf{f}}}^n, \mathbf{A}^n, \boldsymbol{\Sigma})$.

$$\boldsymbol{\Pi^{\mathbf{i}}}^n = \{\pi_l^i\}^n : \pi_l^{i,n} = p(z_1^n = l) \tag{2.4}$$

$$\boldsymbol{\Pi^{\mathbf{f}}}^n = \{\pi_l^f\}^n : \pi_l^{f,n} = p(z_{T_n}^n = l) \tag{2.5}$$

$$\mathbf{A}^n = \{a_{t,ll'}\}^n : A_{t,ll'} = p(z_t^n = l'|z_{t-1}^n = l) \tag{2.6}$$

$$\boldsymbol{\Sigma} = \{\Sigma_{ld}\} : \Sigma_{ld} = p(x_t = d|z_t = l) \tag{2.7}$$

Writing the above expressions in vector form,

$$p(\mathbf{z}_1^n|\boldsymbol{\Pi^{\mathbf{i}}}^n) = \prod_{l=1}^{L} \pi_l^{i,n,z_{1l}} \tag{2.8}$$

$$p(\mathbf{z}_{T_n}^n|\boldsymbol{\Pi^{\mathbf{f}}}^n) = \prod_{l=1}^{L} \pi_l^{f,n,z_{T_n l}} \tag{2.9}$$

$$p(\mathbf{z}_t^n|\mathbf{z}_{t-1}^n, \mathbf{A}_t^n) = \prod_{l=1}^{L} \prod_{l'=1}^{L} {A_{t,ll'}^n}^{\mathbf{z}_{t,l}\mathbf{z}_{t-1,l'}} \tag{2.10}$$

$$p(\mathbf{x}_t^n|\mathbf{z}_t^n, \boldsymbol{\Sigma}) = \prod_{l=1}^{L} \prod_{d=1}^{D} {\Sigma_{ld}}^{\mathbf{z}_{t,l}\mathbf{x}_{t,d}} \tag{2.11}$$

3. Expectation Maximization (EM) for Hidden Markov Model

3.1. **E-Step.**

Calculate posterior distributions using the forward-backward algorithm. Two quantites of interest: (1) the marginal posterior at each time step, denoted by $\gamma(\mathbf{z}_t^n)$,

$$\gamma(\mathbf{z}_t^n) = p(\mathbf{z}_t^n|\mathbf{x}_{1:T_n}^n, \boldsymbol{\Theta}^n), \tag{3.1}$$

and (2) the joint posterior between successive states, denoted by $\xi(\mathbf{z}_{t-1}^n, \mathbf{z}_t^n)$,

$$\xi(\mathbf{z}_{t-1}^n, \mathbf{z}_t^n) = p(\mathbf{z}_{t-1}^n, \mathbf{z}_t^n|\mathbf{x}_{1:T_n}^n, \boldsymbol{\Theta}^n). \tag{3.2}$$

To do this we use the standard forward-backward algorithm. First, construct the quantities

$$\alpha(\mathbf{z}_t^n) = p(\mathbf{x}_{1:t}^n, \mathbf{z}_t^n) \tag{3.3}$$

$$\beta(\mathbf{z}_t^n) = p(\mathbf{x}_{t+1:T_n}^n|\mathbf{z}_t^n) \tag{3.4}$$

Recursion relations for $\alpha(\mathbf{z}_t^n)$ and $\beta(\mathbf{z}_t^n)$ can be derived:

$$\alpha(\mathbf{z}_t) = p(\mathbf{x}_t^n|\mathbf{z}_t^n) \sum_{\mathbf{z}_{t-1}^n} \alpha(\mathbf{z}_{t-1}^n)p(\mathbf{z}_t^n|\mathbf{z}_{t-1}^n) \tag{3.5}$$

$$\beta(\mathbf{z}_t^n) = \sum_{\mathbf{z}_{t+1}} \beta(\mathbf{z}_{t+1}^n) p(\mathbf{x}_{t+1}^n|\mathbf{z}_{t+1}^n) p(\mathbf{z}_{t+1}^n|\mathbf{z}_t^n) \tag{3.6}$$

Because we know where the random walk begins and ends, initial conditions for the recursion are fixed:

$$\alpha(z_{1,1}^n) = 1 \tag{3.7}$$

$$\beta(z_{T_n,L}^n) = 1 \tag{3.8}$$

The normalization condition on $\alpha(\mathbf{z}_t)$ is denoted by c_t^n,

$$c_t^n = \sum_{l=1}^{L} \alpha(z_{tl}^n) \tag{3.9}$$

Then we can write γ and ξ as

$$\gamma(\mathbf{z}_t^n) = \alpha(\mathbf{z}_t^n)\beta(\mathbf{z}_t^n) \tag{3.10}$$

$$\xi(\mathbf{z}_{t-1}^n, \mathbf{z}_t^n) = c_t^n \hat{\alpha}(\mathbf{z}_{t-1}^n) p(\mathbf{x}_t^n|\mathbf{z}_t^n) p(\mathbf{z}_t^n|\mathbf{z}_{t-1}^n) \hat{\beta}(\mathbf{z}_t^n) \tag{3.11}$$

3.2. **M-Step.** In which maximum likelihood estimates of model parameters are computed.

3.2.1. *Sequence Estimate.*

$$\Sigma_{ld}^n = \frac{\displaystyle\sum_{t=1}^{T_n} \gamma(z_{tl}^n) x_{td}^n}{\displaystyle\sum_{t=1}^{T_n} \gamma(z_{tl}^n)} \tag{3.12}$$

3.2.2. *Forward Transition Vector.*

$$\Omega_t^n = \sum_{l=1}^{L-1} \xi(z_{t-1}^n = l, z_t^n = l+1) \tag{3.13}$$

Sum the elements on the upper diagonal and normalize to get the probability of a forward transition at step t. The backward transition vector is simply $\mathbf{1} - \boldsymbol{\Omega}^{\mathbf{n}}$, or the sum along the lower diagonal.

3.3. **H-Step.** In which model parameters are updated by combining results of multiple reads.

3.3.1. *Sequence Inference.*

$$\Sigma_{ld} = \frac{1}{N} \sum_{n=1}^{N} \Sigma_{ld}^n \tag{3.14}$$

Use this as input $\boldsymbol{\Sigma}$ in the next iteration of EM.

3.3.2. *Error Inference.* To do. You can imagine deriving an estimate of where errors are likely to have occured in each read sequence. This could be a forward pass through the set of sequences where if a given location did not agree with the consensus at that point is liable to have been an error. Denote this vector by $\boldsymbol{\Phi}$.

4. Inference Evaluation

We define several ways of evaluating the sequence inference.

4.0.3. *Inference Likelihood.* $p(\mathbf{X})$ is the total data likelihood function, given by

$$p(\mathbf{X}) = \sum_{n=1}^{N} \prod_{t=1}^{T_n} c_t^n \tag{4.1}$$

4.0.4. *Sequence Inference Entropy.* H^l_{seq} measures the normalized entropy of the sequence inference at position l, where $\Sigma_{ld} = p(S_l = d)$,

$$H^l_{\text{seq}} = -\frac{1}{\log D} \sum_{d=1}^{D} \Sigma_{ld} \log \Sigma_{ld}. \tag{4.2}$$

The normalized total inference entropy is given by

$$H^{\text{tot}}_{\text{seq}} = \frac{1}{L} \sum_{l=1}^{L} H^l_{\text{seq}} \tag{4.3}$$

4.0.5. *Path Inference Entropy.* H^t_{path} measures the normalized entropy of the path inference at time t, where $\gamma_{tl} = p(z_t = l|\mathbf{X})$,

$$H^t_{\text{path}} = -\frac{1}{\log L} \sum_{l=1}^{L} \gamma_{tl} \log \gamma_{tl}. \tag{4.4}$$

This is a read sequence dependent measure.