

Supporting Material for ‘Statistical Inference for Nanopore Sequencing with a Biased Random Walk Model’

Kevin J. Emmett,^{†*} Jacob K. Rosenstein,[¶] Jan-Willem van de Meent,[‡] Ken L. Shepard,[§]
and Chris H. Wiggins[‡]

[†]Department of Physics and [‡]Department of Applied Physics and Applied Math and

[§]Department of Electrical Engineering, Columbia University, New York, New York; and

[¶]School of Engineering, Brown University, Providence, Rhode Island

S1 Model Derivation

In this section we provide a derivation of the model used to generate the results in the paper.

S1.1 Notation

S1.1.1 Model Inputs

- β : forward bias of random walk
- ϵ : 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: $d = 4$)

S1.1.2 Data

- $\mathbf{S} = \{s_1, \dots, s_L\}$: input DNA sequence.
 - $s_l \in \{\text{A, G, C, T}\}$: 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
 - $\mathbf{Z}^n = \mathbf{z}_{1:T_n}^n = \{z_1^n, \dots, z_{T_n}^n\}$: latent state sequence n
 - $z_t^n \in \{1, \dots, L\}$: latent state at position t of read sequence n .
- $\mathbf{X} = \{\mathbf{X}^1, \dots, \mathbf{X}^N\}$: set of N observed state sequences.
 - $\mathbf{X}^n = \mathbf{x}_{1:T_n}^n = \{x_1^n, \dots, x_{T_n}^n\}$: observed base sequence n .
 - $x_t^n \in \{\text{A, G, C, T}\}$: observed base at position t of read sequence n .

S1.1.3 Parameters

- $\Theta = (\Pi^i, \Pi^f, \mathbf{A}, \Sigma)$: set of model parameters.
- $\Pi^i^n = p(\mathbf{z}_1^n)$: initial state vector for sequence n .
- $\Pi^f^n = p(\mathbf{z}_{T_n}^n)$: final state vector for sequence n .
- $\mathbf{A}_t^n = p(\mathbf{z}_t^n | \mathbf{z}_{t-1}^n)$: state transition matrix. Can be time-independent or time-dependent.
- Σ : sequence estimate matrix
 $\cdot \Sigma_{ld} = p(S_l = d)$

S1.2 Probability Model

The complete-data likelihood is written as

$$p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta). \quad (\text{S1})$$

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

$$p(\mathbf{X}|\Theta) = \prod_{n=1}^N p(\mathbf{X}^n|\Theta^n) = \prod_{n=1}^N \sum_{\mathbf{Z}^n} p(\mathbf{X}^n, \mathbf{Z}^n|\Theta^n). \quad (\text{S2})$$

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 Σ across all reads. We use the conditional independence properties of a first order Markov chain to factorize the likelihood for each read as

$$p(\mathbf{X}^n, \mathbf{Z}^n|\Theta^n) = p(\mathbf{z}_1^n) p(\mathbf{x}_1^n | \mathbf{z}_1^n) \prod_{t=2}^{T_n} p(\mathbf{z}_t^n | \mathbf{z}_{t-1}^n) p(\mathbf{x}_t^n | \mathbf{z}_t^n). \quad (\text{S3})$$

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

$$\Pi^i^n = \{\pi_l^i\}^n : \pi_l^{i,n} = p(z_1^n = l) \quad (\text{S4})$$

$$\Pi^f^n = \{\pi_l^f\}^n : \pi_l^{f,n} = p(z_{T_n}^n = l) \quad (\text{S5})$$

$$\mathbf{A}^n = \{a_{t,ll'}\}^n : A_{t,ll'} = p(z_t^n = l' | z_{t-1}^n = l) \quad (\text{S6})$$

$$\Sigma = \{\Sigma_{ld}\} : \Sigma_{ld} = p(x_t = d | z_t = l) \quad (\text{S7})$$

Writing the above expressions in vector form,

$$p(\mathbf{z}_1^n | \Pi^i^n) = \prod_{l=1}^L \pi_l^{i,n,z_{1l}} \quad (\text{S8})$$

$$p(\mathbf{z}_{T_n}^n | \Pi^f^n) = \prod_{l=1}^L \pi_l^{f,n,z_{T_n l}} \quad (\text{S9})$$

$$p(\mathbf{z}_t^n | \mathbf{z}_{t-1}^n, \mathbf{A}^n) = \prod_{l=1}^L \prod_{l'=1}^L A_{t,ll'}^{n,z_t,l,z_{t-1},l'} \quad (\text{S10})$$

$$p(\mathbf{x}_t^n | \mathbf{z}_t^n, \Sigma) = \prod_{l=1}^L \prod_{d=1}^D \Sigma_{ld}^{z_t,l,x_t,d} \quad (\text{S11})$$

S1.3 Expectation Maximization (EM) for Hidden Markov Model

S1.3.1 E-Step

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

$$\gamma(\mathbf{z}_t^n) = p(\mathbf{z}_t^n | \mathbf{x}_{1:T_n}^n, \Theta^n), \quad (\text{S12})$$

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

$$\xi(\mathbf{z}_{t-1}^n, \mathbf{z}_t^n) = p(\mathbf{z}_{t-1}^n, \mathbf{z}_t^n | \mathbf{x}_{1:T_n}^n, \Theta^n). \quad (\text{S13})$$

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

$$\alpha(\mathbf{z}_t^n) = p(\mathbf{x}_{1:t}^n | \mathbf{z}_t^n) \quad (\text{S14})$$

$$\beta(\mathbf{z}_t^n) = p(\mathbf{x}_{t+1:T_n}^n | \mathbf{z}_t^n) \quad (\text{S15})$$

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

$$\alpha(\mathbf{z}_t) = p(\mathbf{x}_t^n | \mathbf{z}_t^n) \sum_{\mathbf{z}_{t-1}^n} \alpha(\mathbf{z}_{t-1}^n) p(\mathbf{z}_t^n | \mathbf{z}_{t-1}^n) \quad (\text{S16})$$

$$\beta(\mathbf{z}_t^n) = \sum_{\mathbf{z}_{t+1}^n} \beta(\mathbf{z}_{t+1}^n) p(\mathbf{x}_{t+1}^n | \mathbf{z}_{t+1}^n) p(\mathbf{z}_{t+1}^n | \mathbf{z}_t^n) \quad (\text{S17})$$

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

$$\alpha(\mathbf{z}_{1,1}^n) = 1 \quad (\text{S18})$$

$$\beta(\mathbf{z}_{T_n,L}^n) = 1 \quad (\text{S19})$$

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

$$c_t^n = \sum_{l=1}^L \alpha(\mathbf{z}_{t,l}^n) \quad (\text{S20})$$

Then we can write γ and ξ as

$$\gamma(\mathbf{z}_t^n) = \alpha(\mathbf{z}_t^n) \beta(\mathbf{z}_t^n) \quad (\text{S21})$$

$$\xi(\mathbf{z}_{t-1}^n, \mathbf{z}_t^n) = c_t^n \hat{\alpha}(\mathbf{z}_{t-1}^n) p(\mathbf{x}_t^n | \mathbf{z}_t^n) p(\mathbf{z}_t^n | \mathbf{z}_{t-1}^n) \hat{\beta}(\mathbf{z}_t^n) \quad (\text{S22})$$

S1.3.2 M-Step

In the M-step, maximum likelihood estimates of the model parameters are computed.

$$\Sigma_{ld}^n = \frac{\sum_{t=1}^{T_n} \gamma(\mathbf{z}_{t,l}^n) x_{t,d}^n}{\sum_{t=1}^{T_n} \gamma(\mathbf{z}_{t,l}^n)} \quad (\text{S23})$$

S1.3.3 H-Step

In the H-step, model parameters are updated by combining results of multiple reads.

$$\Sigma_{ld} = \frac{1}{N} \sum_{n=1}^N \Sigma_{ld}^n \quad (\text{S24})$$

Use this as input Σ in the next iteration of EM.

S1.4 Inference Evaluation

We can define several ways of evaluating the sequence inference.

S1.4.1 Inference Likelihood

$p(\mathbf{X})$ is the total data likelihood function, given by

$$p(\mathbf{X}) = \sum_{n=1}^N \prod_{t=1}^{T_n} c_t^n \quad (\text{S25})$$

S1.4.2 Sequence Inference Entropy

H_{seq}^l measures the normalized entropy of the sequence inference at position l , where $\Sigma_{ld} = p(S_l = d)$,

$$H_{\text{seq}}^l = -\frac{1}{\log D} \sum_{d=1}^D \Sigma_{ld} \log \Sigma_{ld}. \quad (\text{S26})$$

The normalized total inference entropy is given by

$$H_{\text{seq}}^{\text{tot}} = \frac{1}{L} \sum_{l=1}^L H_{\text{seq}}^l \quad (\text{S27})$$