Importance Sampling Write-up

Andrew Vaughn

1 General Importance Sampling

We consider trying to maximize the value

$$\begin{aligned} L(s) &= P(D|s) && (1)\\ &= \int P(D,G|s)dG && (2)\\ &= E_{G|D,s_0}\left[\frac{P(D,G|s)}{P(G|D,s_0)}\right] && (3)\end{aligned}$$

We divide this expression through by

$$L(s_0) = P(D|s_0)$$

to get

$$\begin{aligned} \frac{L(s)}{L(s_0)} &= E_{G|D,s_0}\left[\frac{P(D,G|s)}{P(D,G|s_0)}\right] && (4)\\ &= E_{G|D,s_0}\left[\frac{P(D|G,s)P(G|s)}{P(D|G,s_0)P(G|s_0)}\right] && (5)\end{aligned}$$

We then assume the data are independent of the ARG given the selection coefficient (should check theory on whether this is always true or not). Therefore, this becomes

$$E_{G|D,s_0}\left[\frac{P(G|s)}{P(G|s_0)}\right]$$

We then replace this with the local tree T to obtain the importance sampling estimate of the log likelihood ratio

$$\frac{1}{M}\sum_{m=1}^{M}\frac{P(T_m|s)}{P(T_m|s_0)}$$

where G_m is sampled from $P(G|D,s_0)$.

We now consider the coalescence process $P(C|X,N) = \prod_{i=0}^{K-1} P(C_{i+1}|C_iX_i,N_i)$ where X_i is the derived allele frequency at time i.

Because G contains the allele labeling of the leaf nodes, it is sufficient for D, so this becomes (letting $\mathcal{T}$ and L be the tree without labels and letting L be the labels of the leaf nodes)

$$E_{G|D,s_0,\Theta}\left[\frac{P(G|s,\Theta)}{P(G|s_0,\Theta)}\right] = E_{G|D,s_0,\Theta}\left[\frac{P(\mathcal{T},L|s,\Theta)}{P(\mathcal{T},L|s_0,\Theta)}\right]$$

If the times are all sampled at the present then both $P(L|s,\Theta)$ and $P(L|s_0,\Theta)$ depend only on X_0 and not on s or s_0. Therefore, these terms cancel. However, if the leaves are sampled at ancient time points, then we need to consider the probability of seeing the given allelic state, which should be a simple binomial.

2 Coalescent model for a site under selection

$X(t)$ is the derived allele frequency. $N(t)$ is the effective population size, indexed by time before the present.

$C = (C^{der}, C^{anc}, C^{mix})$ where C^{der} is coalescence between derived allele classes, C^{anc} is coalescence between ancestral allele classes, and C^{mix} is coalescence before the mutation (includes all uncoalesced ancestral plus exactly one derived lineage). We consider Tavare's descent formula and take discrete timepoints. Let K be the absolute maximum time interval.

$$P(C|X,N) = \prod_{i=0}^{\tau-1} P(C_{i+1}^{der}|C_i^{der}, X_i, N_i)P(C_{i+1}^{anc}|C_i^{anc}, X_i, N_i) \times \prod_{i=\tau}^{K-1} P(C_{i+1}^{mix}|C_i^{mix}, N_i)$$

We have that

$$P(\mathcal{T}, L|s,\Theta) = \sum_{X\in\mathcal{X}} P(\mathcal{T}, L|s, X, \Theta)P(X|s,\Theta) = \sum_{X\in\mathcal{X}} P(\mathcal{T}, L|X,\Theta)P(X|s,\Theta) = \sum_{X\in\mathcal{X}} P(L|X,\Theta)P(\mathcal{T}|L,X,\Theta)P(X|s,\Theta)$$

so we have to affect both the coalescence process by adding the relevant lineages back in. We also have to do the emissions of the leaf samples at the ancient time points. See where these things happen and edit them.