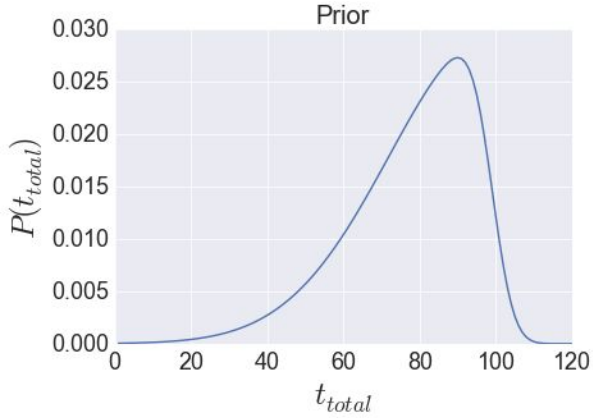
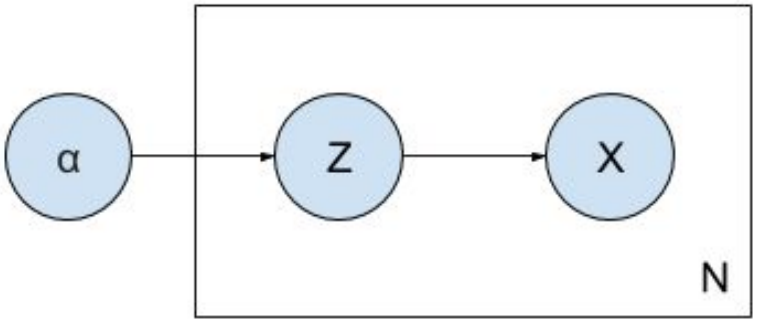
Learning Priors

So far..

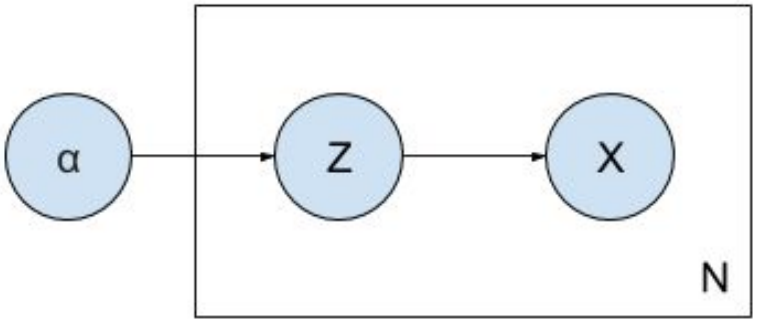


- $p(t_{\text{total}}|t) \propto p(t|t_{\text{total}})p(t_{\text{total}})$
- $p(t|t_{\text{total}}) = 1/t_{\text{total}}$ for $t_{\text{total}} \geq t$ and 0 otherwise

- Z => t_{total}
- X => t
- Z = $\{z_1, z_2, \ldots. z_N\}$
- X = $\{x_1, x_2, \ldots. x_N\}$



- If $z_1 = 80$, then $x_1 <= 80$
- X is superset of total lifespans
- You might encounter or hear about an alive or dead person
- Implication: even people who are alive inform our prior of total lifespan



Expectation Maximization

$$
\begin{aligned}
Q(\alpha|\alpha^{(t)}) &= E_{Z|X,\alpha^{(t)}}\left[\log L(\alpha; X, Z)\right] \\
&= \sum_{i=1}^{N} E_{z_i|x_i,\alpha^{(t)}}\left[\log L(\alpha; x_i, z_i)\right] \\
&= \sum_{i=1}^{N}\sum_{z_i} p(z_i|x_i, \alpha^{(t)}) \log p(x_i, z_i|\alpha) \\
&= \sum_{i=1}^{N}\sum_{z_i} T(x_i, z_i) \log \left(p(x_i|z_i)p(z_i|\alpha)\right) \\
&= \sum_{i=1}^{N}\sum_{z_i} T(x_i, z_i) \log \left(p(z_i|\alpha)/z_i\right)
\end{aligned}
$$

where:

- $T(x_i, z_i) := p(z_i|x_i, \alpha^{(t)})$ is a fixed function with respect to $\alpha^{(t)}$
- $p(x_i, z_i|\alpha) = p(x_i|z_i)p(z_i|\alpha)$ is the same formula from Tenenbaum's paper, with $\alpha^{(t)}$ included, since $p(x_i|z_i, \alpha) = p(x_i|z_i)$ by assumption
- Substituting $p(x_i|z_i) = 1/z_i$ for all z_i is okay since $T(x_i, z_i) = 0$ for $z_i \leq x_i$ by definition.

$$
T(x_i, z_i) = p(z_i|x_i, \alpha^{(t)}) = \frac{p(z_i, x_i|\alpha^{(t)})}{\sum_{z_i} p(z_i, x_i|\alpha^{(t)})}
$$

- Normally - maximize expectation wrt α
- Substitute α and repeat until convergence
- Analogous to batch gradient descent

$$\sum_{i=1}^{N} \sum_{z_i} T(x_i, z_i) \log (p(z_i|\alpha)/z_i)$$

Stepwise Expectation Maximization

- Update α for every incoming sample x_i
- Analogous to stochastic gradient descent
- Since single sample is a bad approximation we interpolate between s_i' and μ

$$\sum_{z_i} T(x_i, z_i) \log (p(z_i|\alpha)/z_i)$$

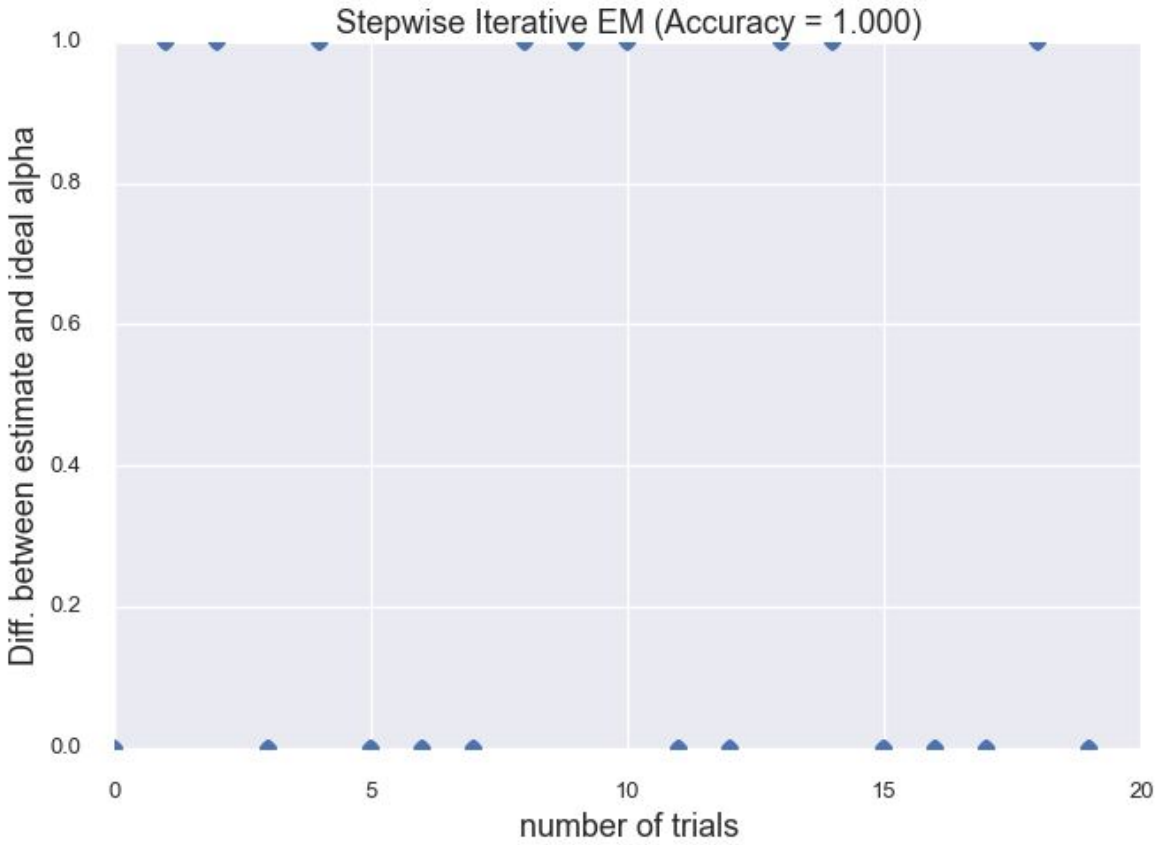
$\mu \leftarrow$ initialization; $k = 0$
for each iteration $t = 1, \ldots, T$:
  for each example $i = 1, \ldots, n$ in random order:
    $s_i' \leftarrow \sum_{\mathbf{z}} T(x_i, z_i) \log (p(z_i|\alpha)/z_i)$ [inference]
    $\mu \leftarrow (1-\eta_k)\mu + \eta_k s_i'; k \leftarrow k+1$ [towards new]

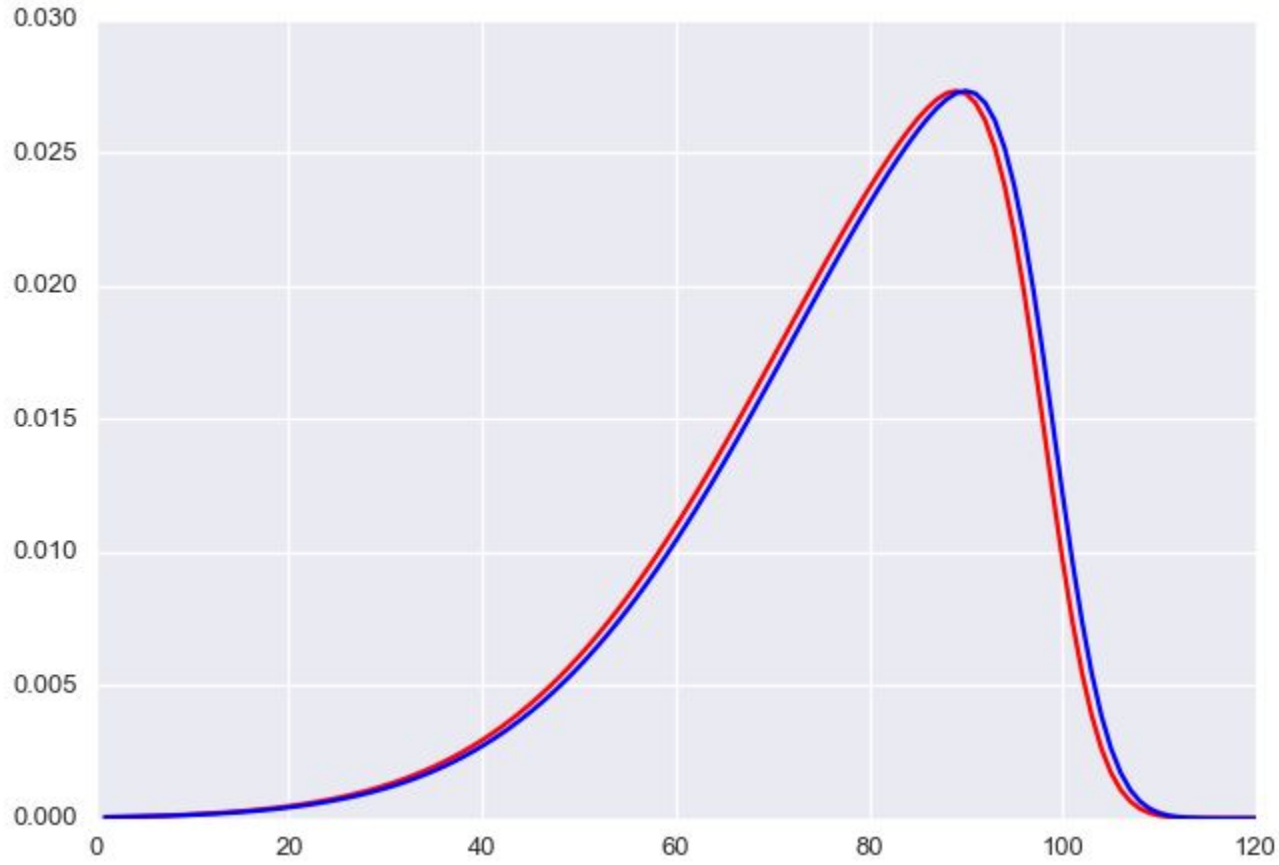
Properties of stepwise EM

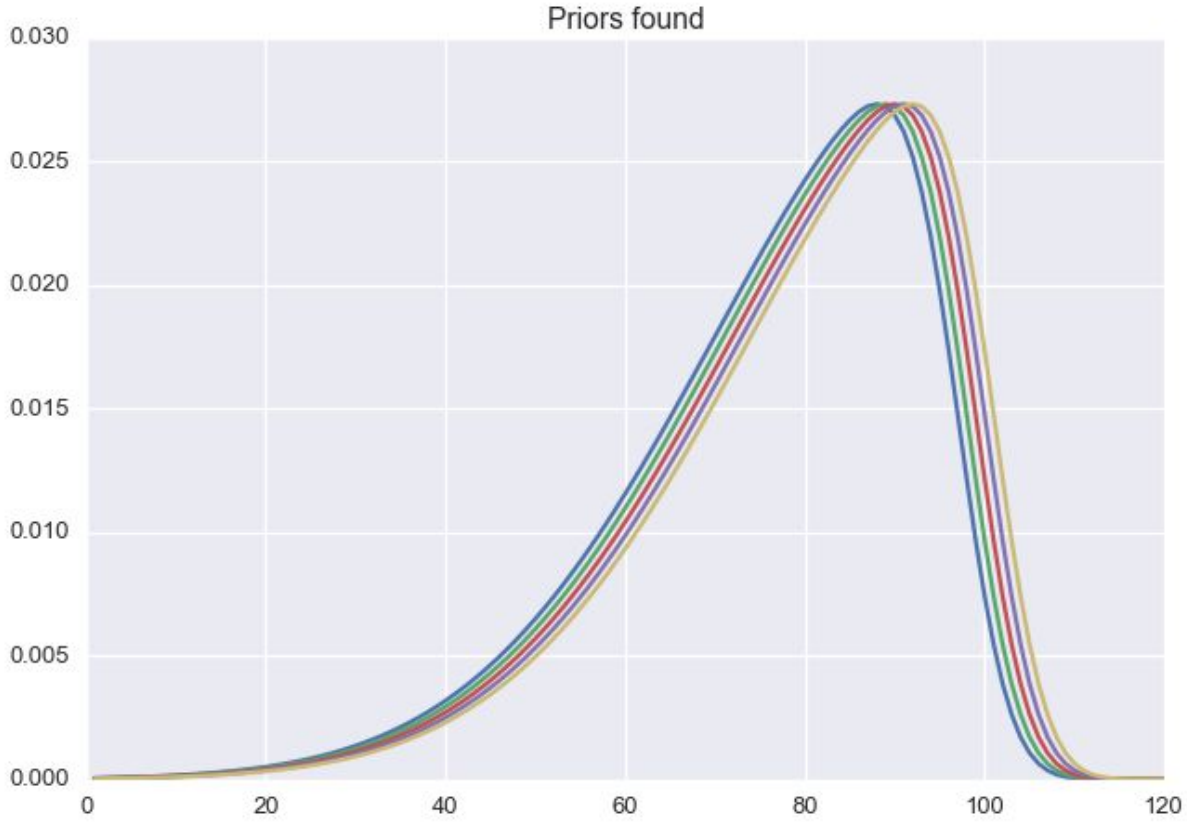
- Standard results from stochastic approximation literature
- If we take $\eta_k = (k+2)^{-\alpha}$, then any $0.5 < \alpha \leq 1$ is valid
- The smaller the alpha, the larger the updates and more quickly we forget (decay) the old statistics.
- Swift progress but generates instability in the trajectory
- One of the principal motivations: Speed
- Liang, P., & Klein, D. (2009, May)

Preliminary Results

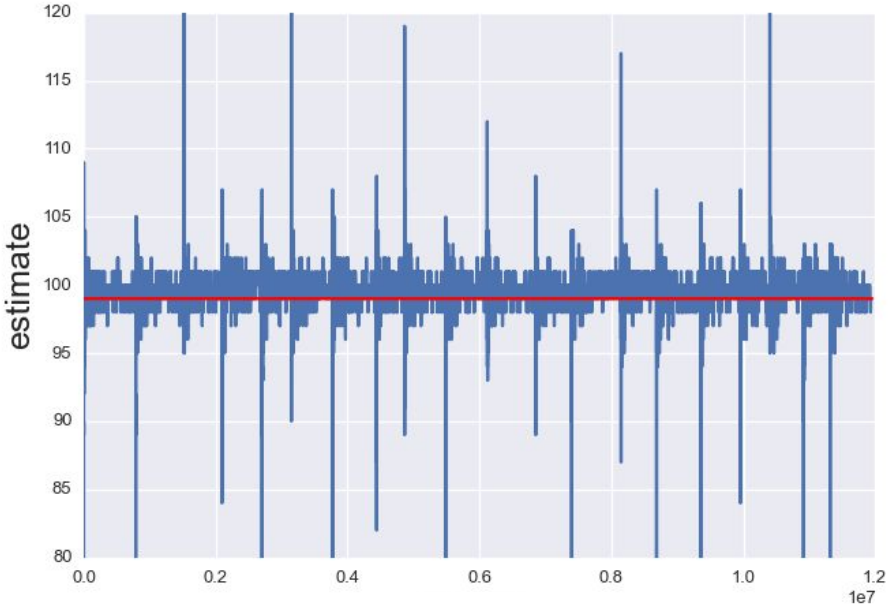
- α = 0.5
- Iter per sample = 5
- Batch iter = 5
- Trials = 20
- Samples/trial = 50,000

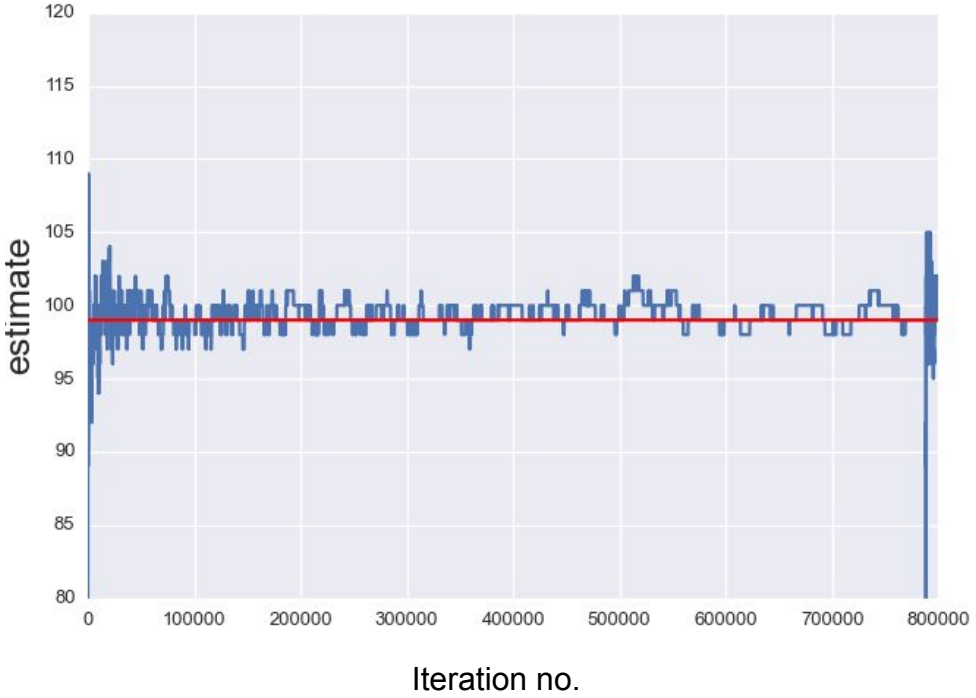


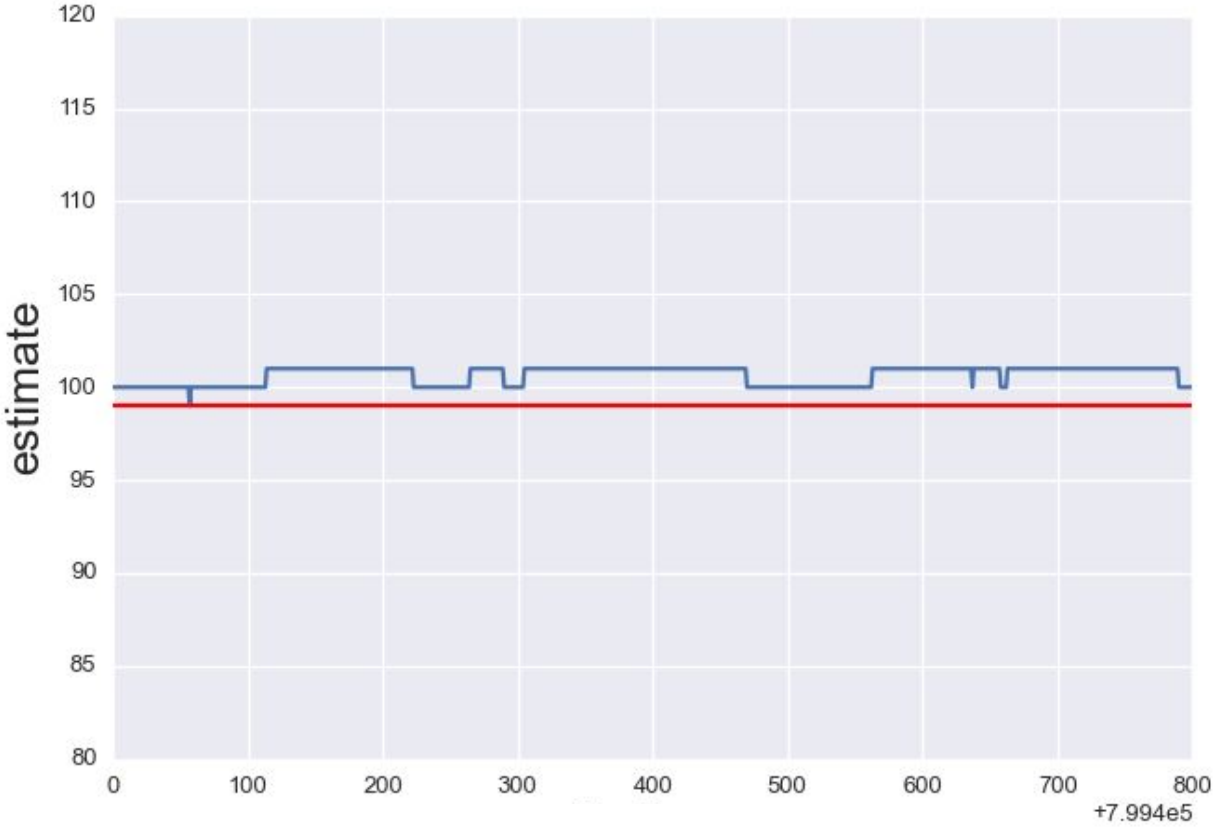

0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120



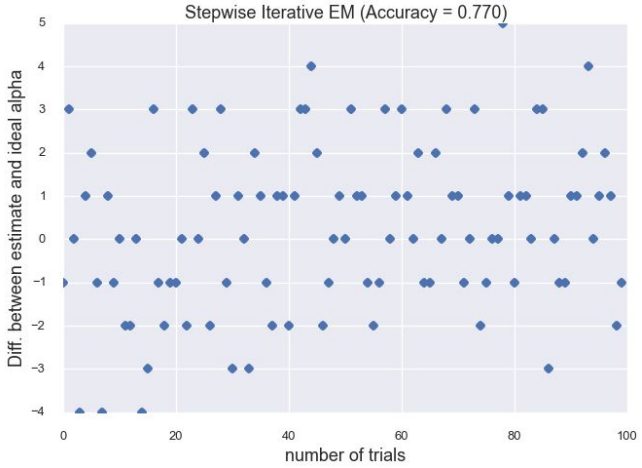
Accuracy measure = Should be within +/-2 of the optimal estimate

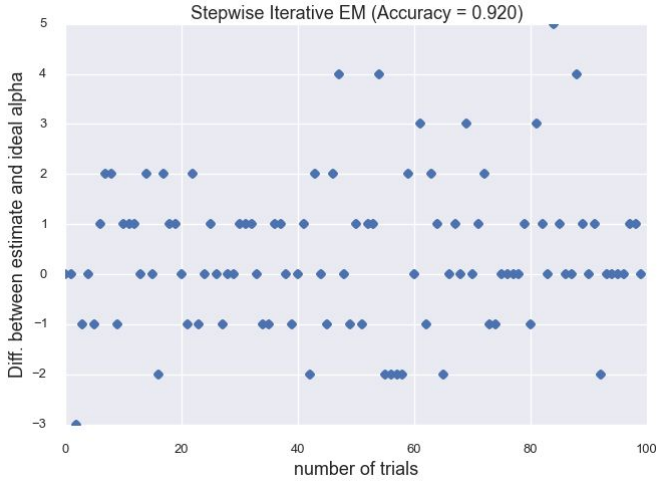

estimate
120
115
110
105
100
95
90
85
80
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e7
Iteration no.

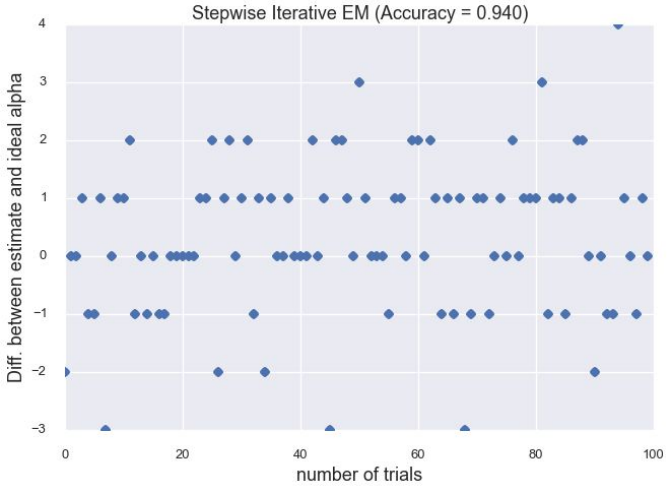

estimate
120
115
110
105
100
95
90
85
80
0
100000
200000
300000
400000
500000
600000
700000
800000
Iteration no.

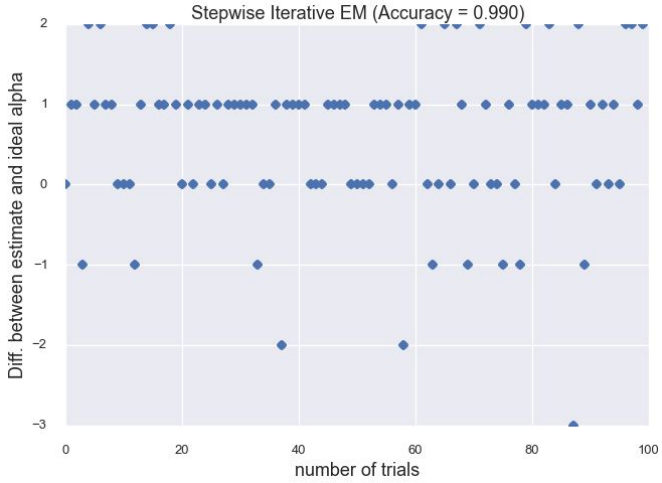

120
115
110
105
100
95
90
85
80
estimate
0
100
200
300
400
500
600
700
800
+7.994e5
Iteration no.

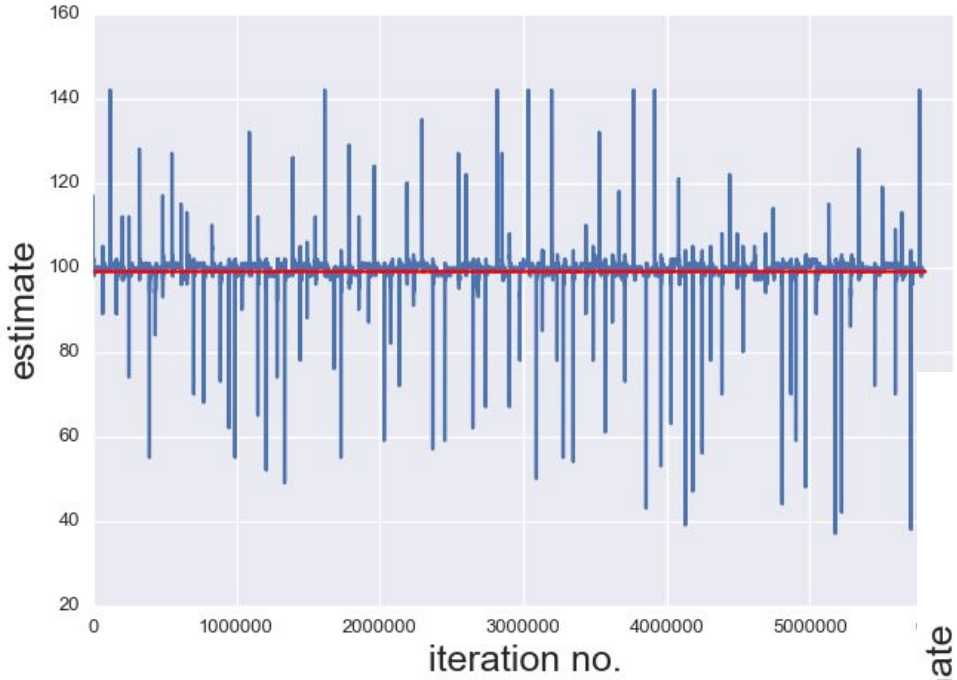
Samples/trials :
500, 1000,
3000, and 5000





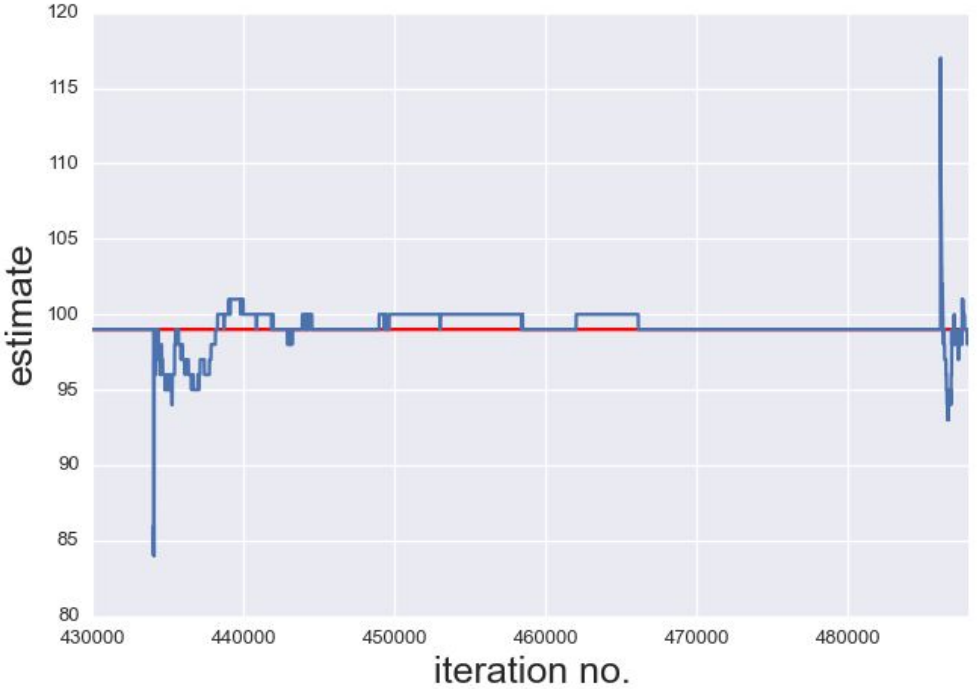




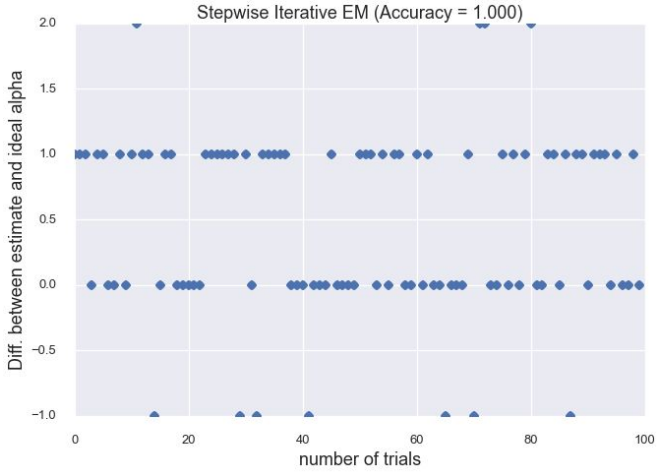


$\alpha = 0.65$

Increase in stability



Samples/trials :
500, 1000,
3000, and 5000



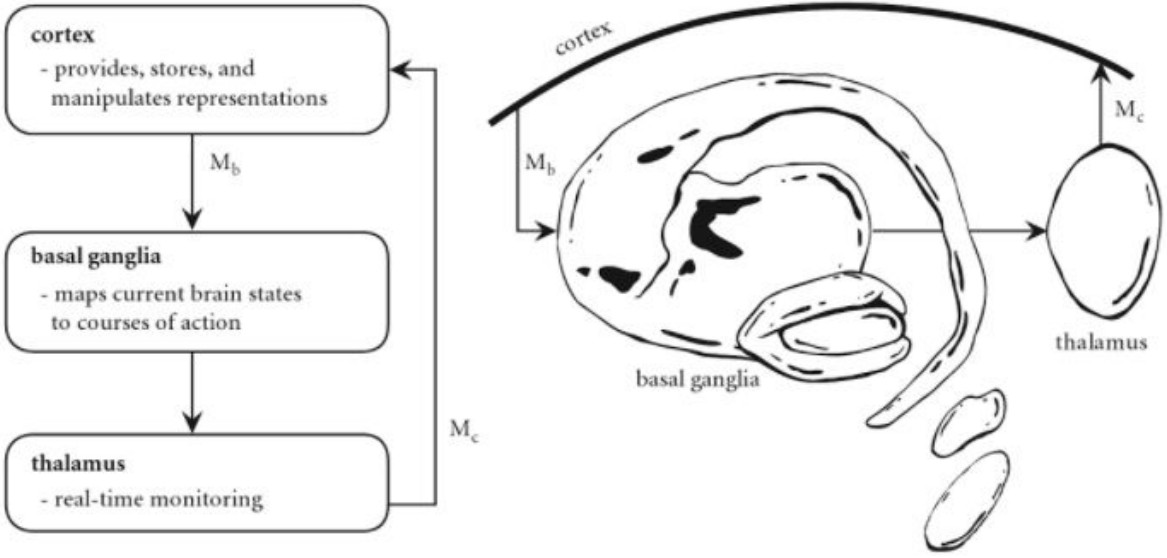


FIGURE 5.5 The cortex-basal ganglia-thalamus loop. Arrows indicate connections between the three areas. At a functional level, brain states from the cortex are mapped through the $\mathbf{M}_b$ matrix to the basal ganglia. Each row in such a matrix specifies a known context for which the basal ganglia will choose an appropriate action. The product of the current cortical state and $\mathbf{M}_b$ provides a measure of how similar the current state is to each of the known contexts. The output of the basal ganglia disinhibits the appropriate areas of thalamus. Thalamus, in turn, is mapped through the matrix $\mathbf{M}_c$ back to the cortex. Each column of this matrix specifies an appropriate cortical state that is the consequence of the selected action. The relevant anatomical structures are pictured on the right based on a simplified version of Figure 5.1.

Nengo Implementation

- Step1: Sample x_i comes in (cortical input)
- Step2: Compute the likelihood function p(x/z) - 120 dim vector (cortical state)

$$p(t|t_{\text{total}}) = 1/t_{\text{total}} \text{ for } t_{\text{total}} \geq t \text{ and } 0 \text{ otherwise}$$

- Step3: p(x, z) = p(x/z) p(z/α_0) where α_0 is our initial guess - 120 dim vector (cortical state)
- Step4: s_i' = (p(x,z) / np.sum(p(x,z))).dot(log p(Z/A).T) (matrix M_b)
  - (120) dim . (120*M) dim => M dim where M is the size of search space for α
- Step5:
  - $\eta_k = (k+2)^{-\alpha}$, *Increment k*
  - $\mu \leftarrow (1-\eta_k)\mu + \eta_k s_i'; k \leftarrow k+1$ (Bg input)
- Step6: ᾱ = argmax (μ) (automatically happens in Bg)
- Step7: Find p(z/ᾱ) (Bg output)

(update cortical state in step 3, omit step 1 and 2 until a new sample comes in.)