X represents data

θ_i represents i^{th} of N parameters ($i \in \{1, \ldots, N\}$).

$\theta_i{}^{(j)}$ represents j^{th} sampled value of i^{th} parameter

Gibbs Sampler:

Initialize with parameter values $\theta_1{}^{(0)}, \theta_2{}^{(0)}, \ldots, \theta_N{}^{(0)}$.

We require posterior probability of inital state to be positive

$$\Pr\left(\theta_1{}^{(0)}, \theta_2{}^{(0)}, \ldots, \theta_N{}^{(0)} \mid X\right) > 0$$

Set $k = 0$.

Step 1: Sample ${\theta_1}^{(k+1)}$ from $\Pr\left(\theta_1 \mid X, {\theta_2}^{(k)}, \ldots, {\theta_N}^{(k)}\right)$

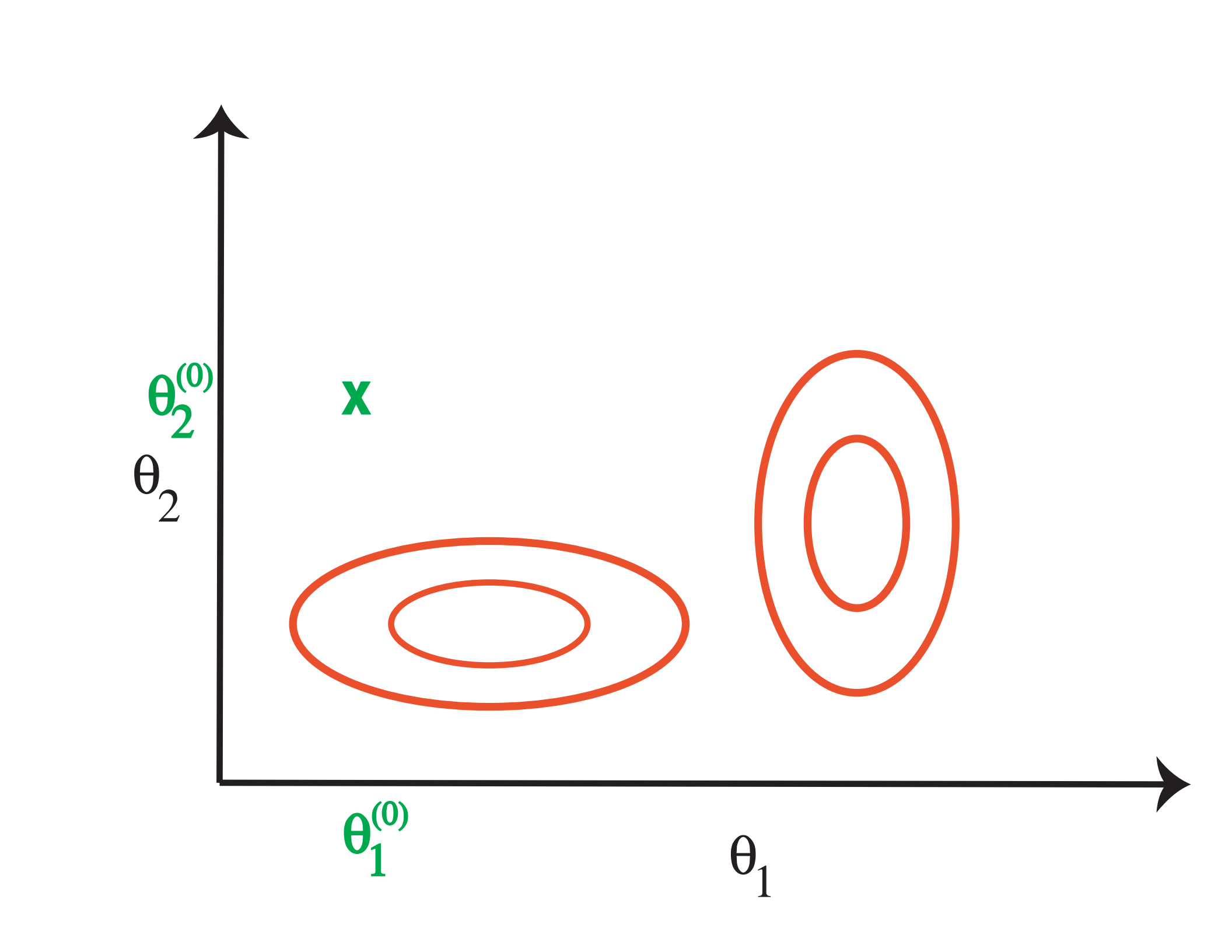
Step 2: Sample ${\theta_2}^{(k+1)}$ from $\Pr\left(\theta_2 \mid X, {\theta_1}^{(k+1)}, {\theta_3}^{(k)} \ldots, {\theta_N}^{(k)}\right)$

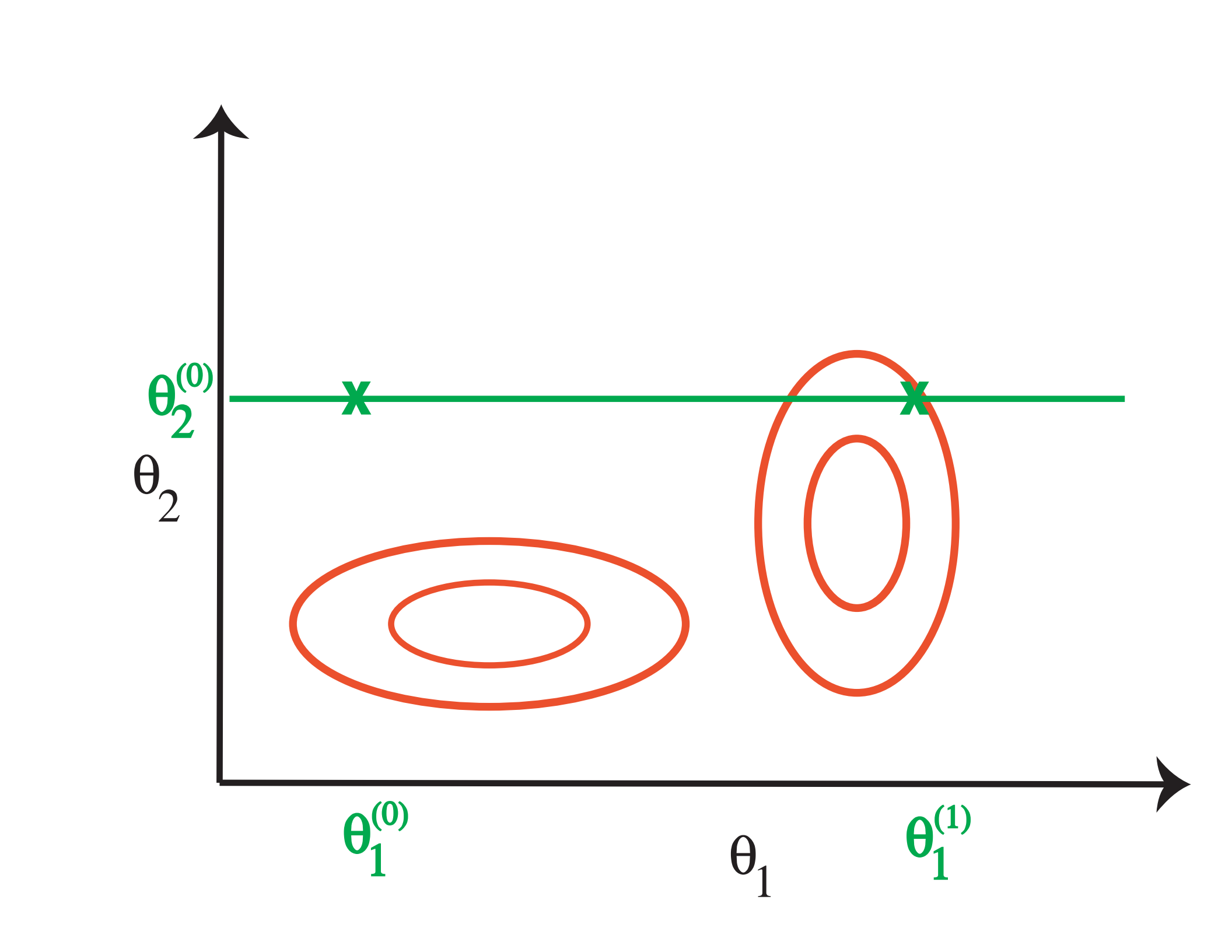
Step 3:
Sample ${\theta_3}^{(k+1)}$ from $\Pr\left(\theta_3 \mid X, {\theta_1}^{(k+1)}, {\theta_2}^{(k+1)}, {\theta_4}^{(k)} \ldots, {\theta_N}^{(k)}\right)$

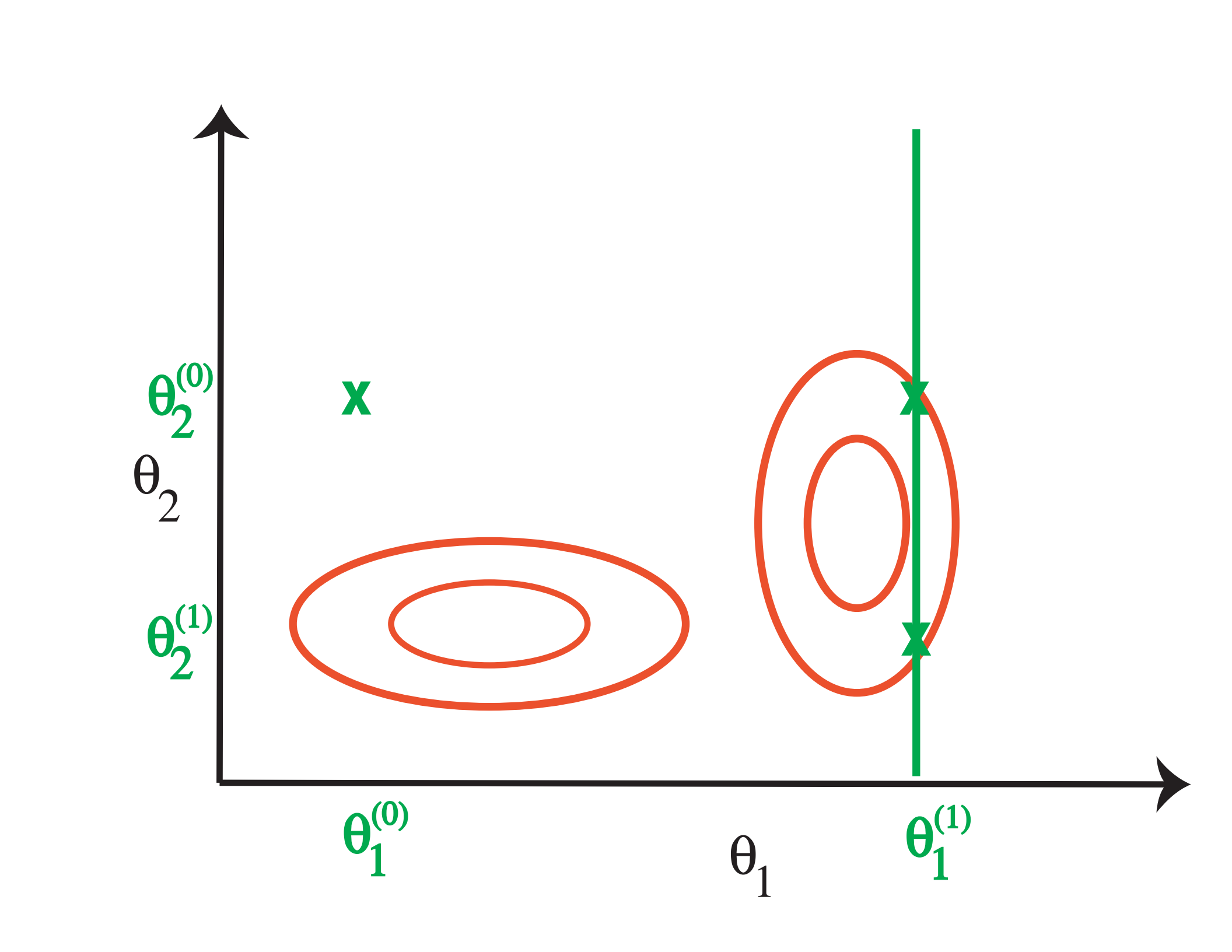
...

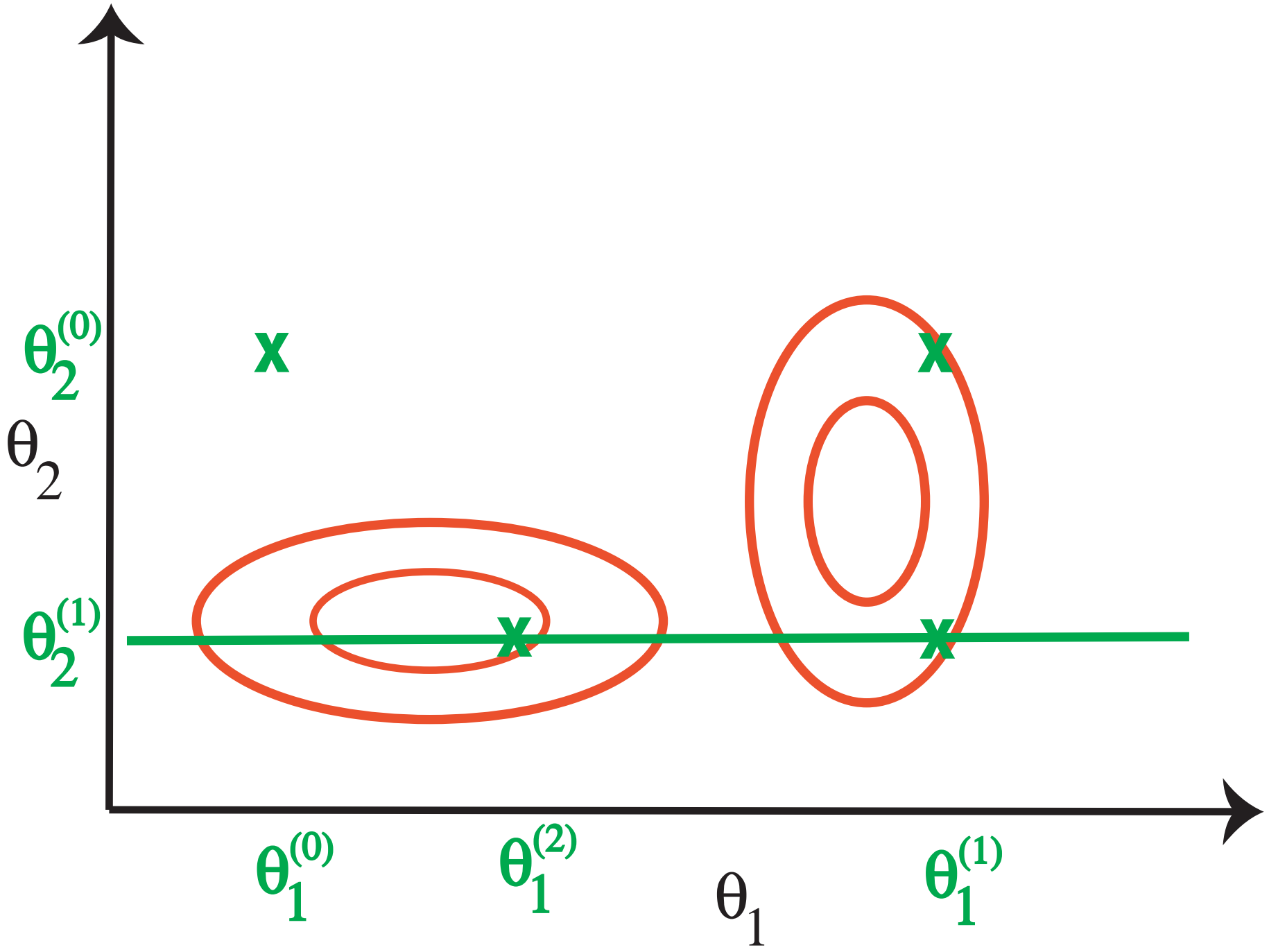
Step N: Sample ${\theta_N}^{(k+1)}$ from $\Pr\left(\theta_N \mid X, {\theta_1}^{(k+1)}, \ldots, {\theta_{N-1}}^{(k+1)}\right)$

Step N+1: Set $k = k + 1$. Goto Step 1.


$\theta_2^{(0)}$
x
θ_2
$\theta_1^{(0)}$
θ_1


$\theta_2^{(0)}$
θ_2
$\theta_1^{(0)}$
θ_1
$\theta_1^{(1)}$


$\theta_2^{(0)}$
θ_2
$\theta_2^{(1)}$
$\theta_1^{(0)}$
θ_1
$\theta_1^{(1)}$


$\theta_2^{(0)}$
θ_2
$\theta_2^{(1)}$
$\theta_1^{(0)}$
$\theta_1^{(2)}$
θ_1
$\theta_1^{(1)}$

Gibbs Sampling - another Markov chain Monte Carlo technique for sampling from posterior distributions (developed by Geman and Geman 1984)

Skeleton of an Example:

$\alpha^{(t)}$ represents t^{th} sampled pairwise alignment

$S^{(t)}$ represents t^{th} set of sampled parameters controlling nucleotide substitution (or amino acid replacement)

$\delta^{(t)}$ represents t^{th} set of sampled parameters controlling insertion and deletion

Data are two non-aligned sequences A and B

To sample from joint posterior density $P(\alpha, S, \delta|A, B)$:

1. Randomly pick initial values $\alpha^{(0)}, S^{(0)}, \delta^{(0)}$

2. Set $t = 1$

3. Sample $\alpha^{(t)}$ from $P(\alpha|S^{(t-1)}, \delta^{(t-1)}, A, B)$

4. Sample $S^{(t)}$ from $P(S|\alpha^{(t)}, \delta^{(t-1)}, A, B)$

5. Sample $\delta^{(t)}$ from $P(\delta|\alpha^{(t)}, S^{(t)}, A, B)$

6. Set $t = t + 1$, Go to Step 3

Chip Lawrence, Jun Liu, and collaborators have done much work on adapting Gibbs Sampling and related Markov chain Monte Carlo techniques to the detection and characterization of subtle sequence patterns.

Most basic case:

N non-aligned sequences

Let $S = \{S_1, S_2, \ldots, S_N\}$ represent these sequences

A priori knowledge is that each sequence will have exactly one ungapped motif of length W

a_i represents the position in sequence i where the ungapped motif of length W begins

Model:

1. All sequences are independent realizations of some process (Specifically, the process generates sequences according to some model but the length of the sequence is assumed given)

2. Amino acid residues at a site in a sequence are determined independently of residues at all other positions

3. Residue types of all sequence positions that are not in the motif are sampled from some background model probability distribution

4. Each of the W positions in the motif is associated with its own set of 20 amino acid frequencies

(Dirichlet priors or a mixture of Dirichlet priors can be used for both the background and motif-position amino acid frequencies.)

5. All possible starting positions for the motif within a sequence can be considered equally likely *a priori*

More Notation: Let π be the vector that contains information about residue frequencies at all background positions and at all motif positions

Goal:

1. Find the set of starting positions of the motif in Sequences 1, 2, ..., N.

2. Estimate the frequencies of all amino acids in position 1 of the motif, position 2 of the motif, ..., position W of the motif.

Outline of Gibbs Procedure (details follow):

Step 1. Make random initial guess as to values of all parameters in models and denote these guesses by $\pi^{(0)}, a_1^{(0)}, a_2^{(0)}, \ldots, a_N^{(0)}$.

Step 2. Set $t = 1$

Step 3. Sample $\pi^{(t)}$ from $P(\pi|a_1^{(t-1)}, a_2^{(t-1)}, \ldots, a_N^{(t-1)}, S)$

Step 4.
Sample $a_1^{(t)}$ from $P(a_1|\pi^{(t)}, a_2^{(t-1)}, \ldots, a_N^{(t-1)}, S) = P(a_1|\pi^{(t)}, S_1)$
$a_2^{(t)}$ from $P(a_2|\pi^{(t)}, a_1^{(t)}, a_3^{(t-1)}, \ldots, a_N^{(t-1)}, S) = P(a_2|\pi^{(t)}, S_2)$
...
$a_N^{(t)}$ from $P(a_N|\pi^{(t)}, a_1^{(t)}, a_2^{(t)}, \ldots, a_{N-1}^{(t)}, S) = P(a_N|\pi^{(t)}, S_N)$

Step 5. Set t=t+1

Step 6. Go to Step 3

Gibbs Motif Implementation (an "almost" Gibbs Sampler)

0. Start by randomly picking a position in each sequence where the motif begins. This defines the "motif alignment" and the background positions.

1. Randomly pick a sequence and temporarily remove it from remaining sequences (Note: this would not be done in a "proper" Gibbs Sampler). Estimate amino acid frequencies at site j in motif with $(c_{t,j,r} + b_{t,j,r})/(c_t + b_t)$ where: $c_{t,j,r}$ is # residues of type r at position j for motif t, $b_{t,j,r}$ is # residues that are pseudocounts for type r at position j for motif t, c_t is total # residues per position, and b_t is total number # pseudocounts per position (Note: we are effectively using estimates of posterior means rather than sampling from posterior of residue frequencies).

Estimate background frequencies of amino acid type r with $(c_r + q_r)/(c + q)$ where: c_r is # background counts of type r, q_r is # pseduocounts for background of type r, c is total # background counts, and q is total # background pseudocounts.

2. Add the removed sequence back to motif alignment. Do this by randomly picking starting position for motif according to its posterior probability.

A. Posterior probability for each starting position is joint probability of data and starting position divided by probability of data.

B. Probability of data is simply sum over all possible starting positions of the joint probability of data and starting position

C. Joint probability of data and starting position is product of probability of data given starting position and probability of starting position (probability of starting position could be treated uniform over all possible starting positions)

D. Prob. of data given starting position is product of appropriate frequencies (either motif or background) for all residues in sequence

(1) Example: $W = 2$ (motif length), $N = 3$ (# of sequences)

Π ~ contains parameters for nucleotide frequencies

$\Pi = \{\Pi_B, \Pi_1, \Pi_2\}$ where

$\Pi_B = \{\Pi_{BA}, \Pi_{BC}, \Pi_{BG}, \Pi_{BT}\}$ ~ background frequencies

$\Pi_1 = \{\Pi_{1A}, \Pi_{1C}, \Pi_{1G}, \Pi_{1T}\}$ ~ frequencies at motif position 1

$\Pi_2 = \{\Pi_{2A}, \Pi_{2C}, \Pi_{2G}, \Pi_{2T}\}$ ~ frequencies at motif position 2

Consider these 3 sequences: Sequence 1, Sequence 2, Sequence 3

$S_1 = GACC$, $S_2 = AAT$, $S_3 = ATC$

motif position 1, motif position 2

G A C C ← background
A A T
A T C ← background

What we get when $q_1 = 2, q_2 = 2, q_3 = 1$

background
background

$q = \{q_1, q_2, q_3\}$ ← motif starting points

Dirichlet Priors [Independent Across motif positions & background frequencies]
(α's below are called pseudocounts)

$$P(\Pi) = P(\Pi_B, \Pi_1, \Pi_2) = P(\Pi_B)\,P(\Pi_1)\,P(\Pi_2)$$

$$P(\Pi_B) = C_B\, \Pi_{BA}^{\alpha_{BA}-1}\, \Pi_{BC}^{\alpha_{BC}-1}\, \Pi_{BG}^{\alpha_{BG}-1}\, \Pi_{BT}^{\alpha_{BT}-1}$$

$$P(\Pi_1) = C_1\, \Pi_{1A}^{\alpha_{1A}-1}\, \Pi_{1C}^{\alpha_{1C}-1}\, \Pi_{1G}^{\alpha_{1G}-1}\, \Pi_{1T}^{\alpha_{1T}-1}$$

$$P(\Pi_2) = C_2\, \Pi_{2A}^{\alpha_{2A}-1}\, \Pi_{2C}^{\alpha_{2C}-1}\, \Pi_{2G}^{\alpha_{2G}-1}\, \Pi_{2T}^{\alpha_{2T}-1}$$

② Likelihoods $P(S|a,\pi) = P(s_1|a,\pi)\,P(s_2|a,\pi)\,P(s_3|a,\pi)$

$$P(S_1 = GACC \mid a,\pi) = P(S_1 = GACC \mid a_1 = 2, \pi)$$

$$= \pi_{BG}\,\pi_{1A}\,\pi_{2C}\,\pi_{BC}$$

$$P(S_2 = AAT \mid a_2 = 2, \pi) = \pi_{BA}\,\pi_{1A}\,\pi_{2A}$$

$$P(S_3 = ATC \mid a_3 = 1, \pi) = \pi_{1A}\,\pi_{2T}\,\pi_{BC}$$

$$P(\pi \mid a_1 = 2, a_2 = 2, a_3 = 1, S) = \frac{P(\pi|a_1,a_2,a_3)\,P(S|\pi,a_1,a_2,a_3)}{\sout{P(a_1,a_2,a_3)}\;P(S|a_1,a_2,a_3)}$$

$$= \frac{P(\pi)\,P(S|\pi,a_1,a_2,a_3)}{\sout{P(a_1,a_2,a_3)}\;P(S|a_1,a_2,a_3)}$$

$$= C\,P(\pi_B)\,P(\pi_1)\,P(\pi_2)\,P(S_1|a_1,\pi)\,P(S_2|a_2,\pi)\,P(S_3|a_3,\pi)$$

$$= C\left[\pi_{BA}^{\alpha_{BA}-1+1}\,\pi_{BC}^{\alpha_{BC}-1+2}\,\pi_{BG}^{\alpha_{BG}-1+1}\,\pi_{BT}^{\alpha_{BT}-1}\right]$$

$$\times\left[\pi_{1A}^{\alpha_{1A}-1+3}\,\pi_{1G}^{\alpha_{1G}-1}\,\pi_{1C}^{\alpha_{1C}-1}\,\pi_{1T}^{\alpha_{1T}-1}\right]$$

$$\times\left[\pi_{2A}^{\alpha_{2A}-1}\,\pi_{2C}^{\alpha_{2C}-1+1}\,\pi_{2G}^{\alpha_{2G}-1}\,\pi_{2T}^{\alpha_{2T}-1+2}\right]$$

All Dirichlet Distributions!

(3)

$$P(a_1=2 \mid \pi, s_1) = \frac{P(s_1, a_1=2 \mid \pi)}{P(s_1 \mid \pi)} = \frac{P(s_1 \mid a_1=2, \pi) P(a_1=2 \mid \pi)}{\sum_{j=1}^{3} P(s_1 \mid a_1=j, \pi) P(a_1=j \mid \pi)}$$

$$= \frac{P(s_1 \mid a_1=2, \pi) P(a_1=2)}{\sum_{j=1}^{3} P(s_1 \mid a_1=j, \pi) P(a_1=j)} = \frac{P(s_1 \mid a_1=2, \pi) \cdot \frac{1}{3}}{\sum_{j=1}^{3} P(s_1 \mid a_1=j, \pi) \cdot \frac{1}{3}}$$

Note: $P(s_1 \mid a_1=2, \pi)$ is explained at top of previous page