ISBA

CHICAGO BOOTH
The University of Chicago Booth School of Business

ABC Tree

Jungeum Kim

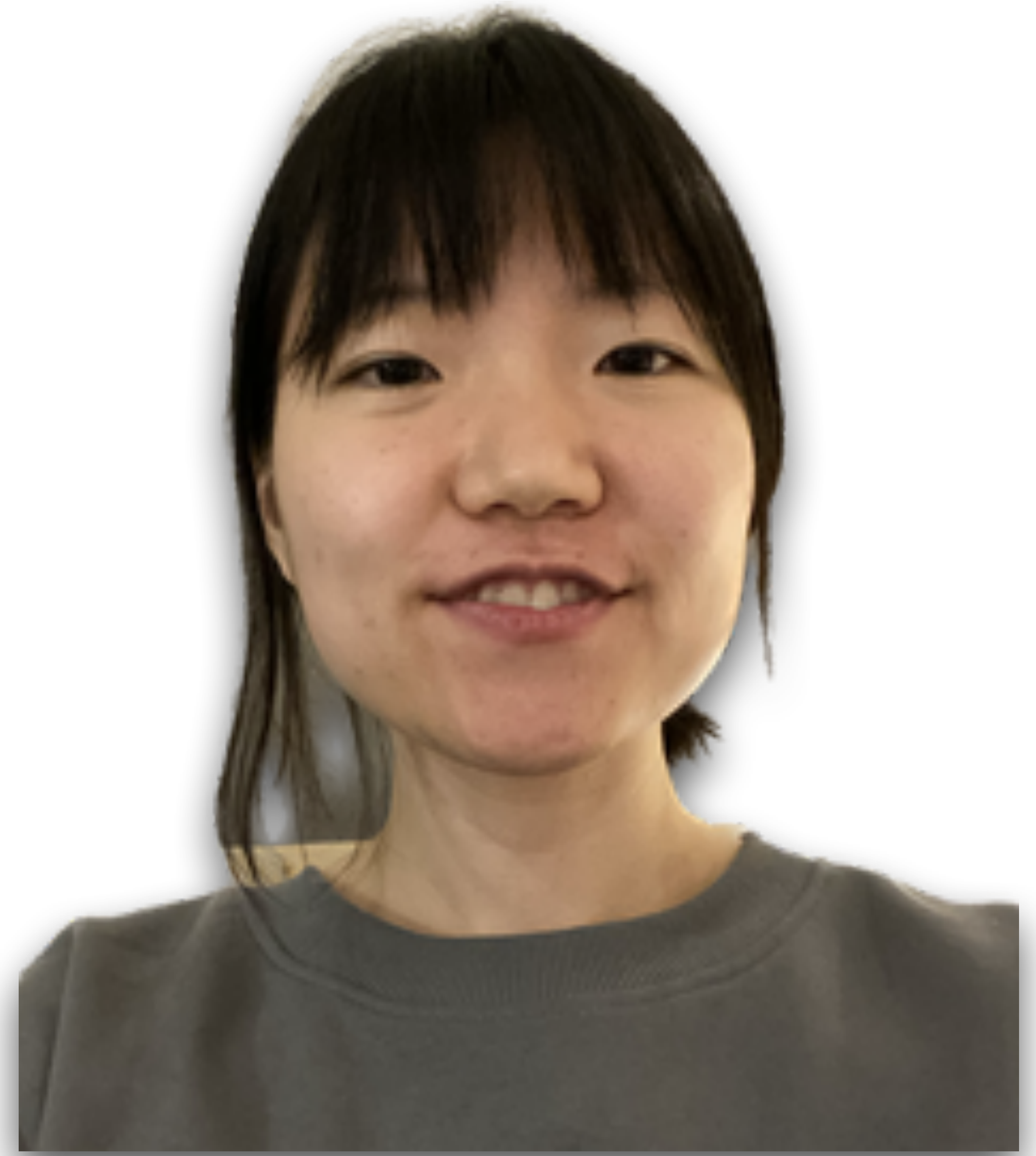
June 2024

ABC:Approximate Bayesian Computation

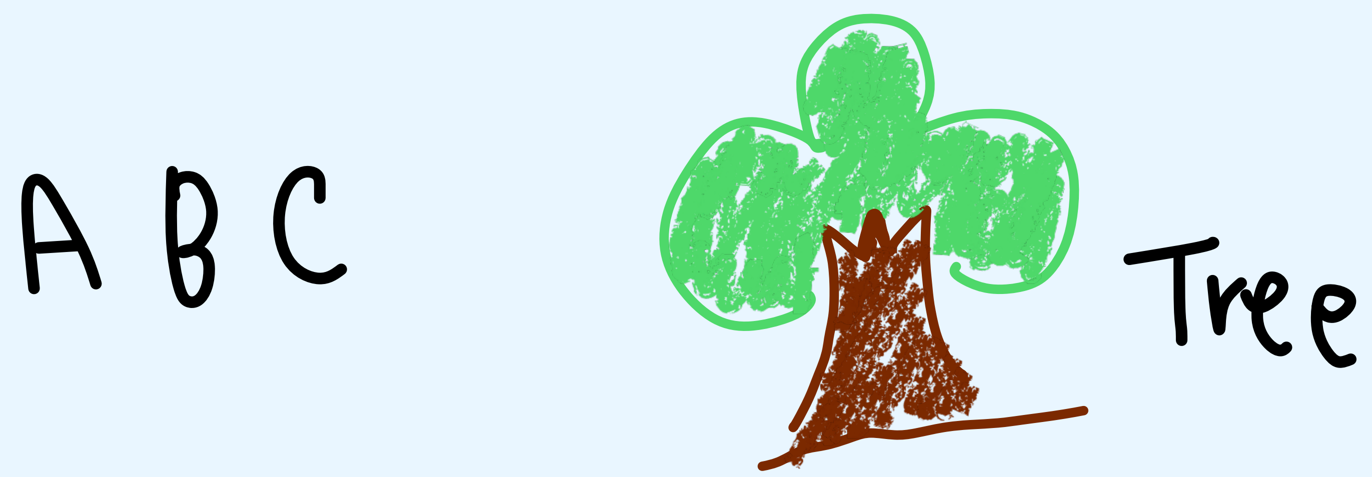
Authors

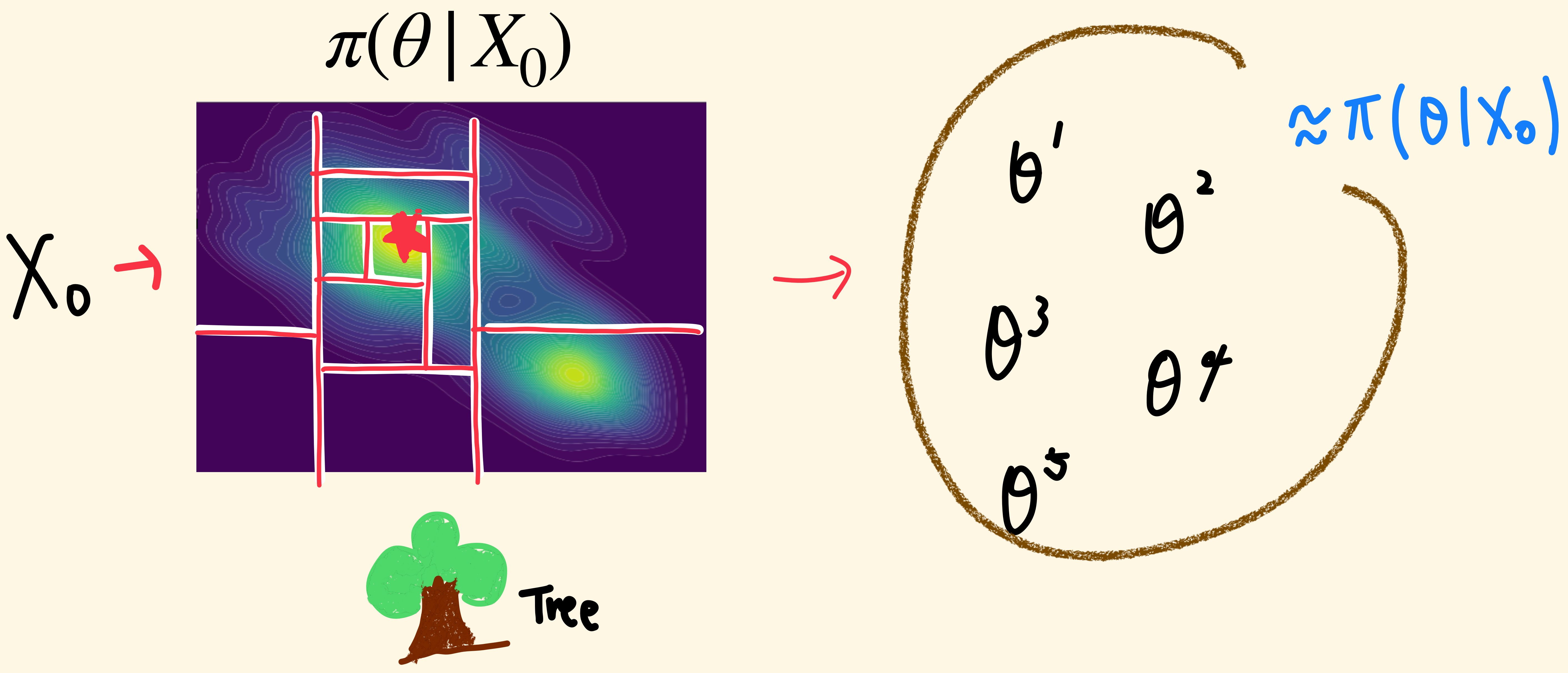
Sean O'Hagan

Jungeum Kim (Me)



Veronika Rockova





Our Contribution

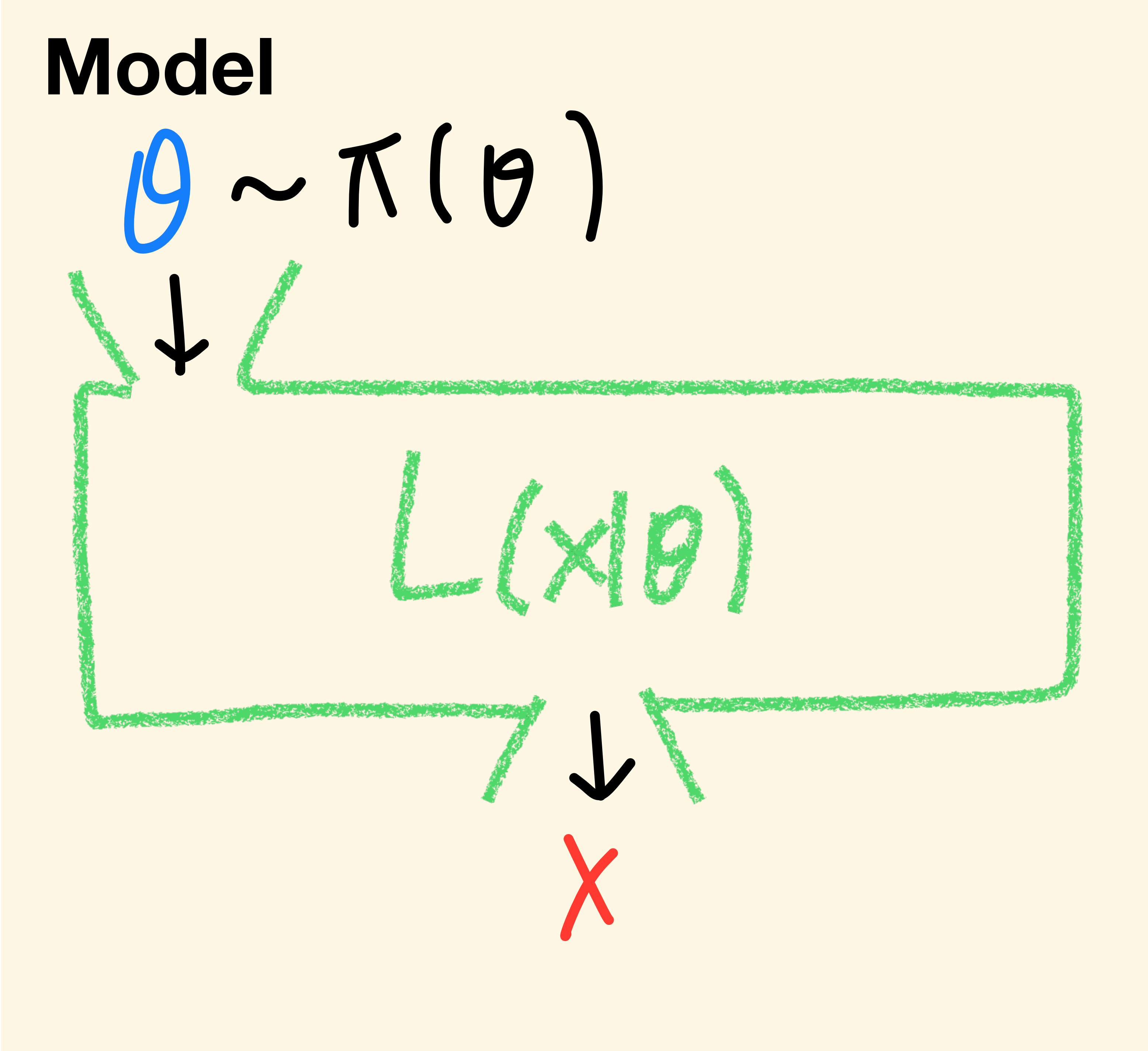
By Combining Tree with Active ABC learning

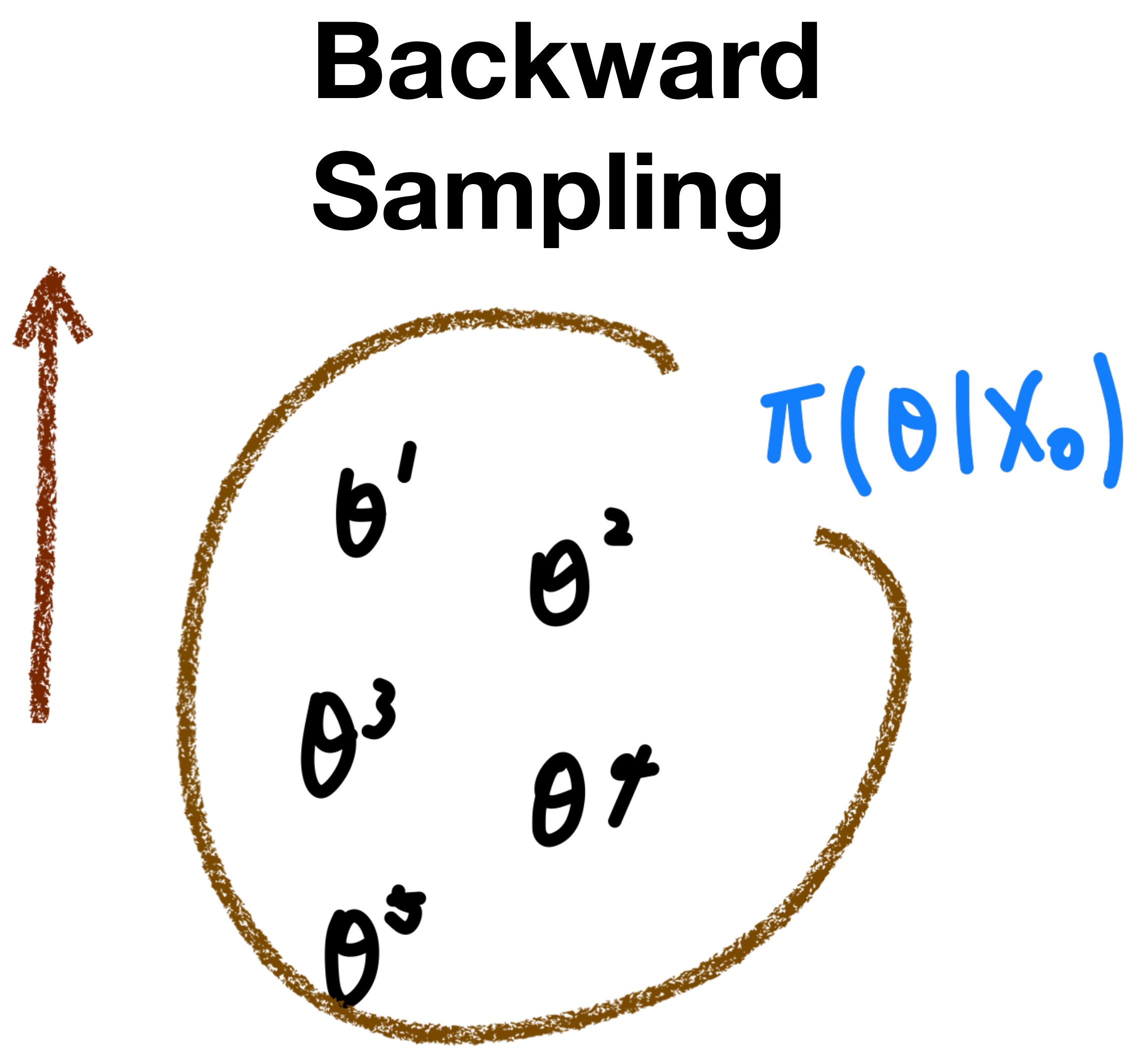
1. Query efficient ABC

2. Likelihood Free MAP (Maximum a Posteriori)

Background

Bayesian Posterior Sampling

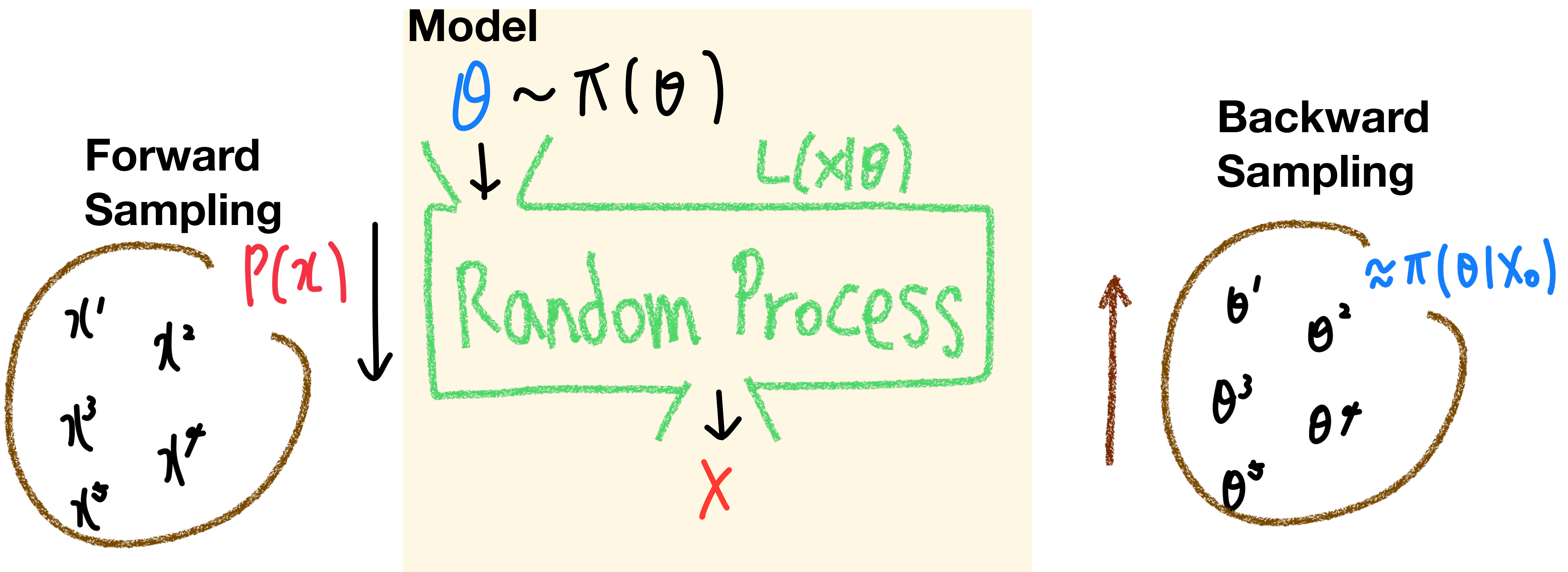




Conjugate Prior : Closed form $\pi(\theta \mid X)$

Likelihood given : MCMC —> $\pi(\theta \mid X) \propto L(X \mid \theta)\pi(\theta)$

Bayesian Posterior Sampling

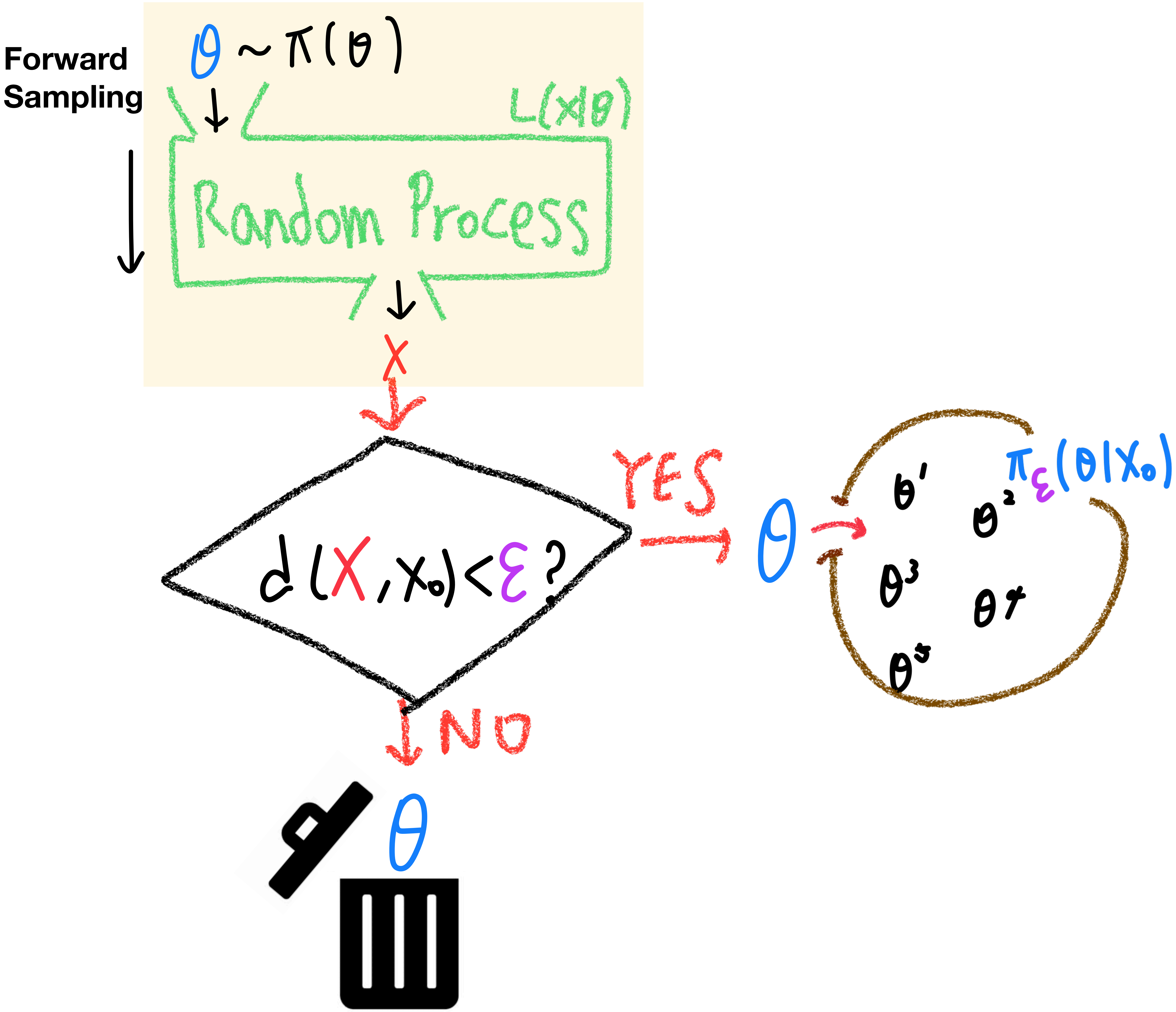


Conjugate Prior : closed form $\pi(\theta\,|\,X)$

Likelihood given : MCMC —> $\pi(\theta\,|\,X) \propto L(X\,|\,\theta)\pi(\theta)$

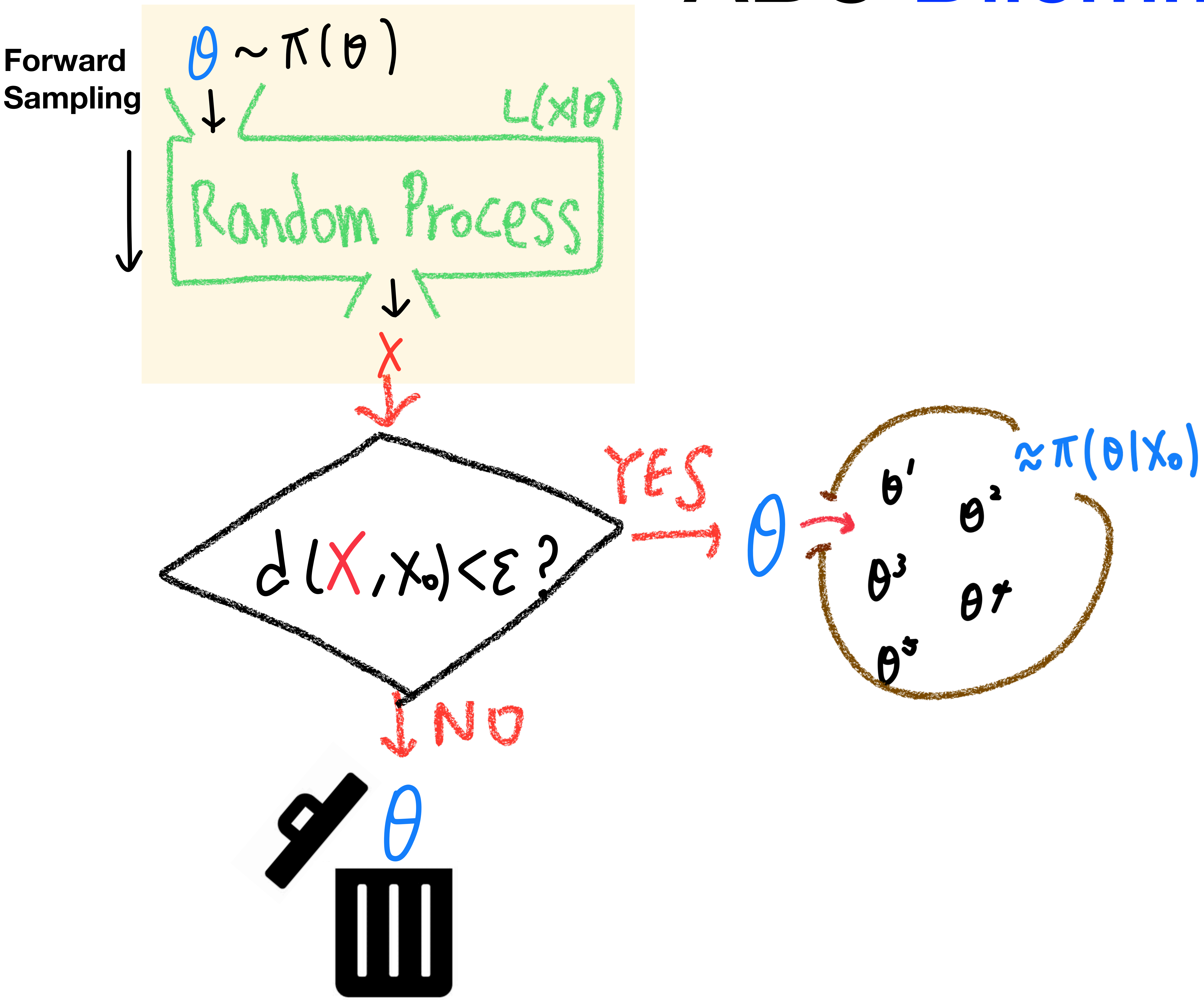
Likelihood unavailable: ABC - use forward sampling (simulation)

ABC (Approximate Bayesian Computation)



$$\pi_\varepsilon(\theta \mid X) \xrightarrow{\varepsilon \to 0} \pi(\theta \mid X)$$

ABC Dilemma



Small ε:

Sample quality ↑

Efficiency (acceptance) ↓

Large ε:

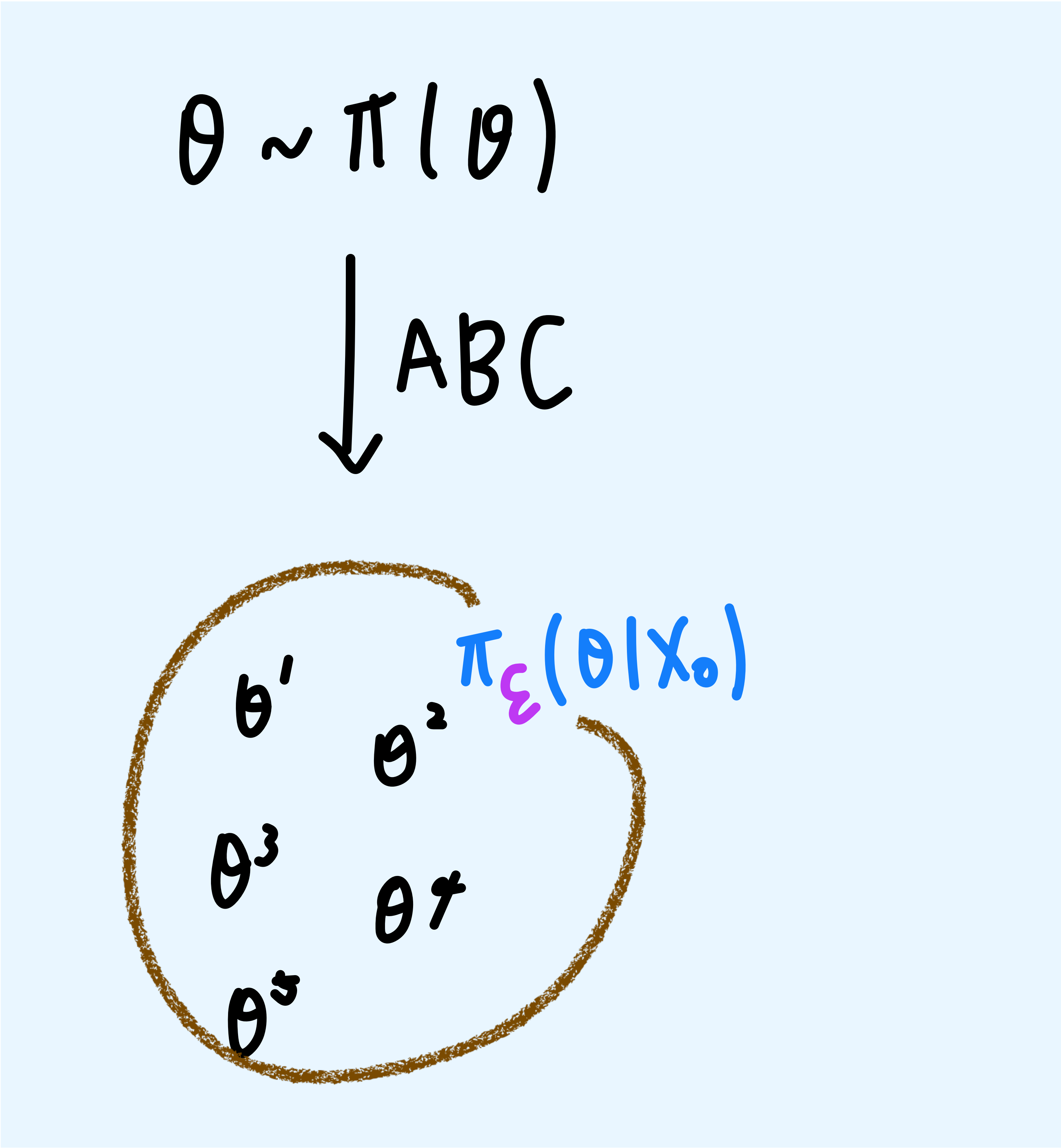
Sample quality ↓

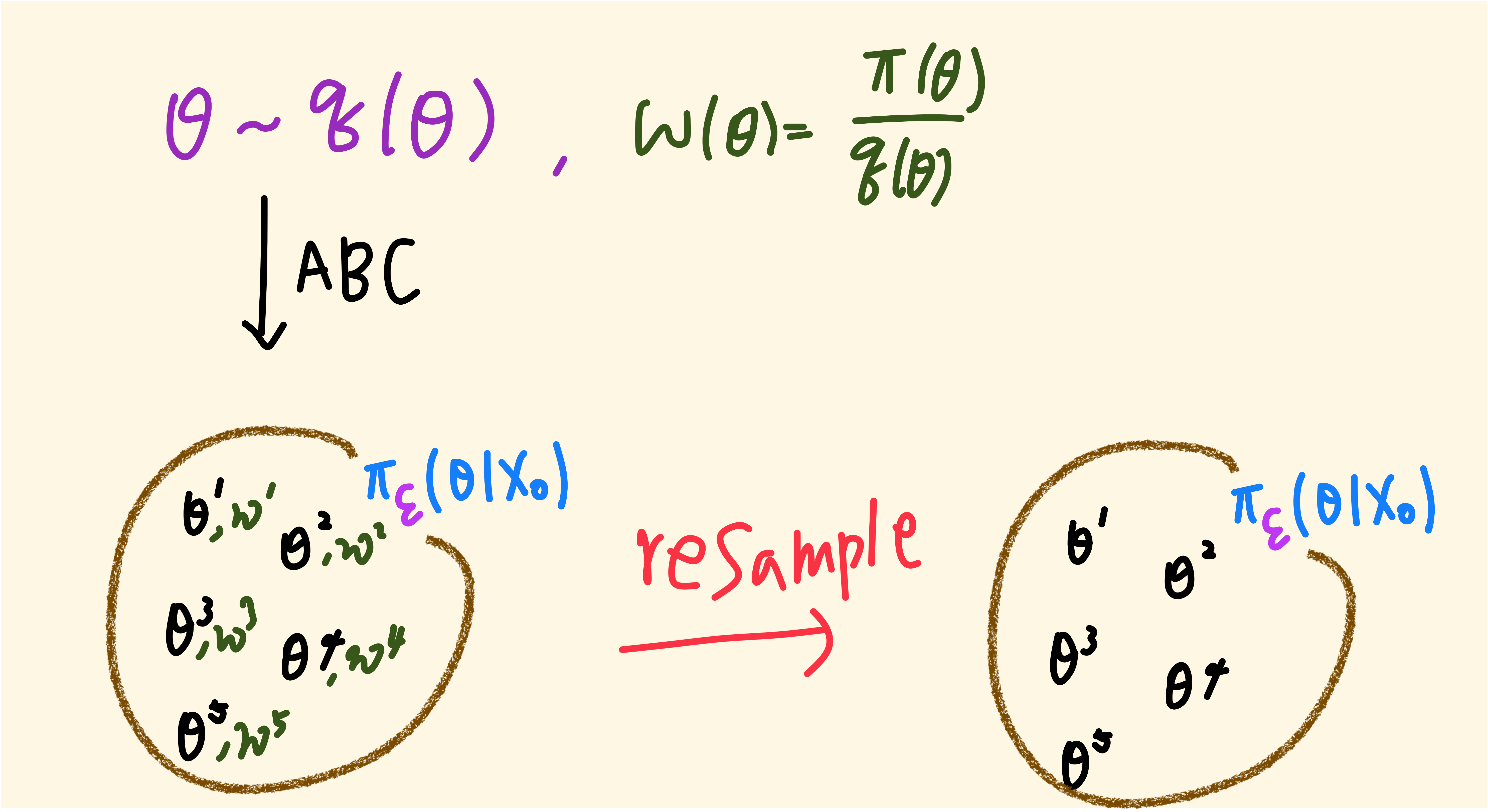
Efficiency (acceptance) ↑

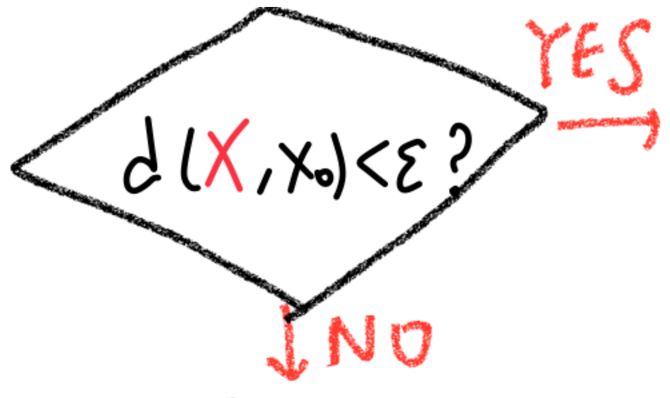
Goal: Query Efficient ABC

Small ε and acceptance ↑

Importance weighted ABC



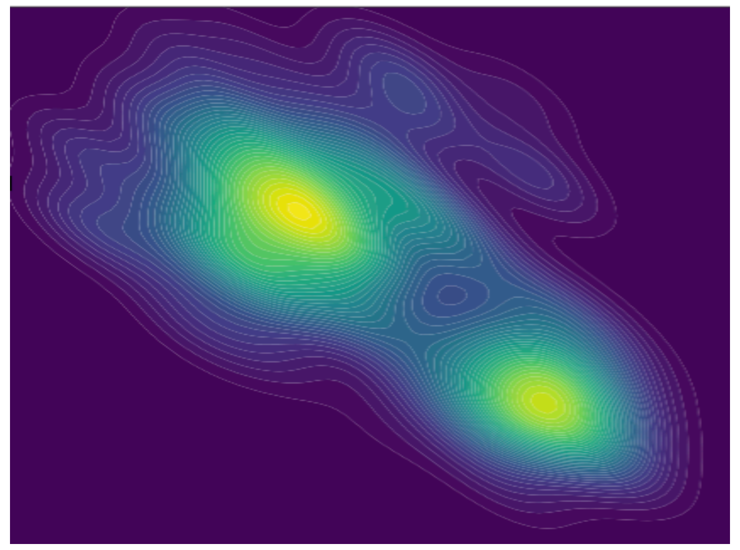


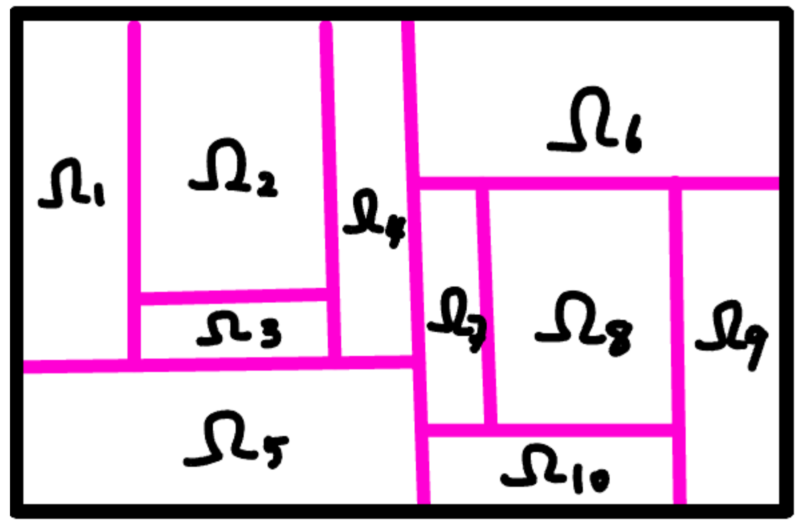
We will learn proposal $q(\theta)$ can be learned actively by the history ()

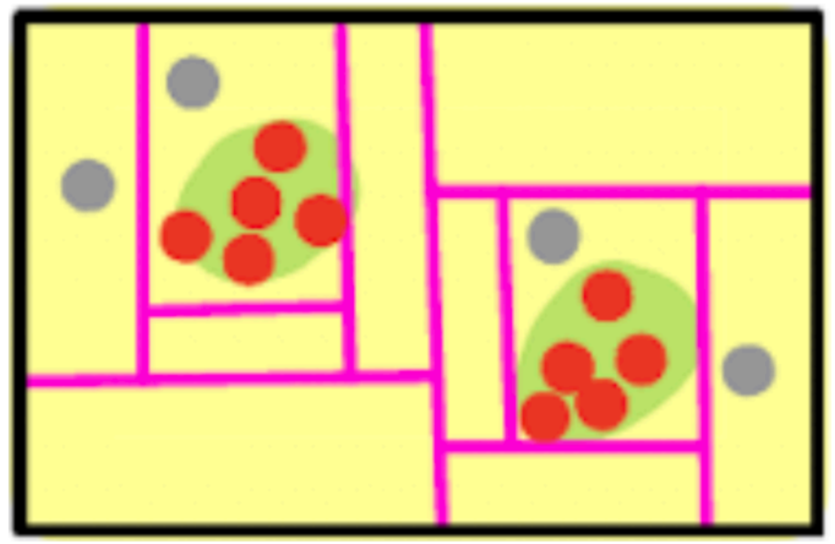
Our Approach

1. Partition Perspective

True Posterior







$$q(\theta) = \sum_{k=1}^{K} {\color{blue} q_k} \times \pi(\theta \mid \Omega_k)\mathbb{I}\{\theta \in \Omega_k\}, \text{ where } \Theta = \cup_{k=1}^{K} \Omega_k$$

2. Active Learning perspective

Question 1) How to get partition? (Outer Loop)

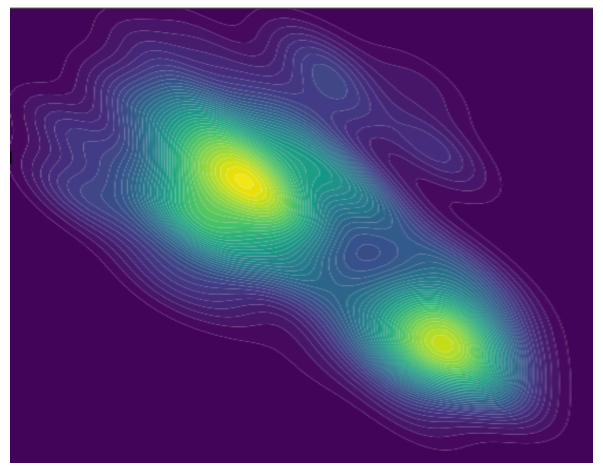
Question 2) How to get ${\color{blue} q_k}$? (Inner Loop)

Question 1) How to get partition?

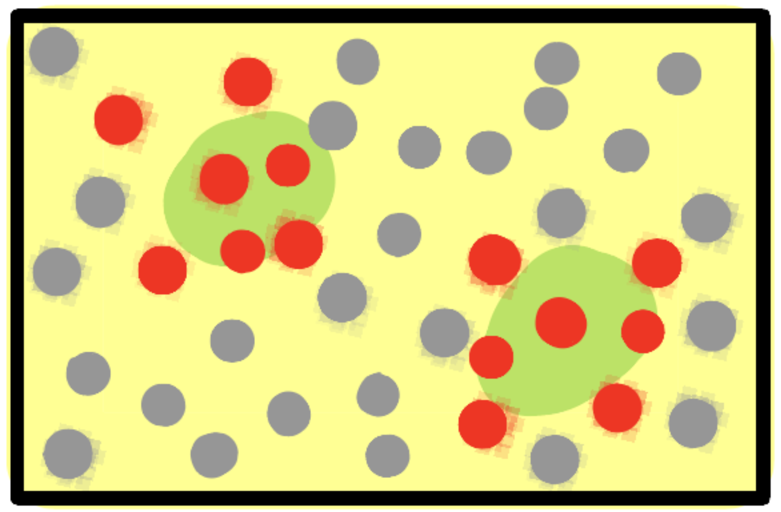
-> Tree fitting in the outer loop

Outer Loop: Partitioning by 

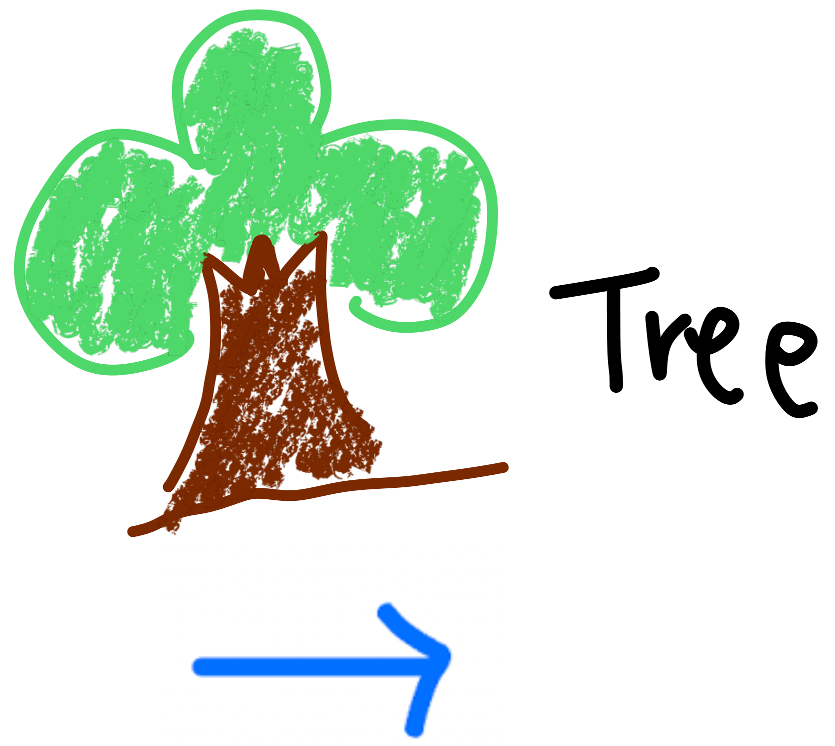
True Posterior

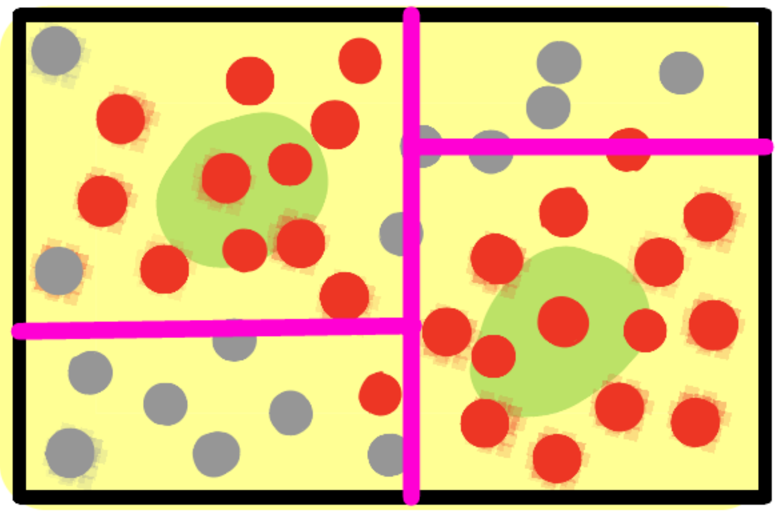


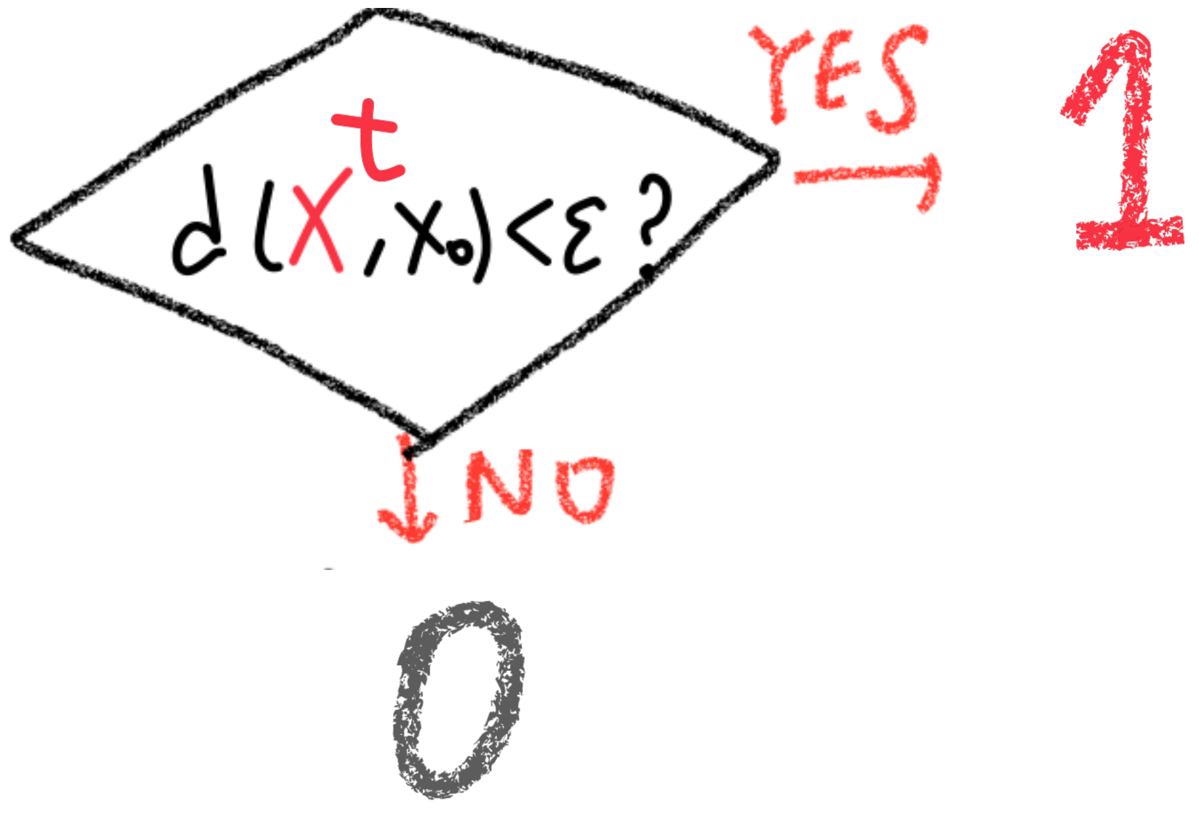
Θ



$Y^{(t)} = \mathrm{ABC}(X^t, X_0, \varepsilon)$

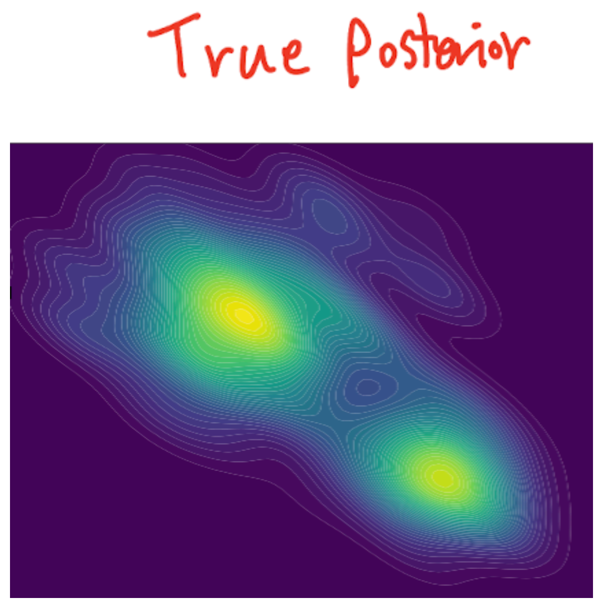


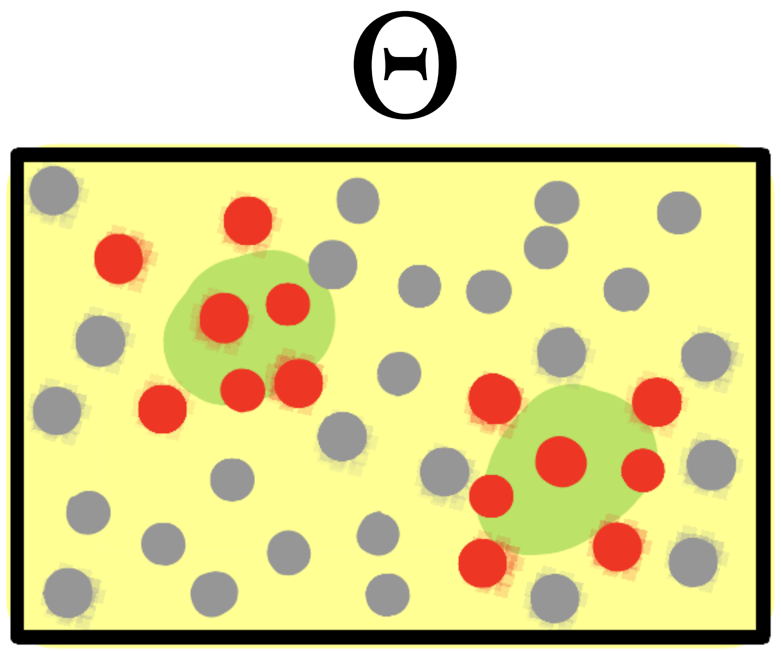


We will learn by $Y^{(t)} = \mathrm{ABC}(X^t, X_0) =$ 

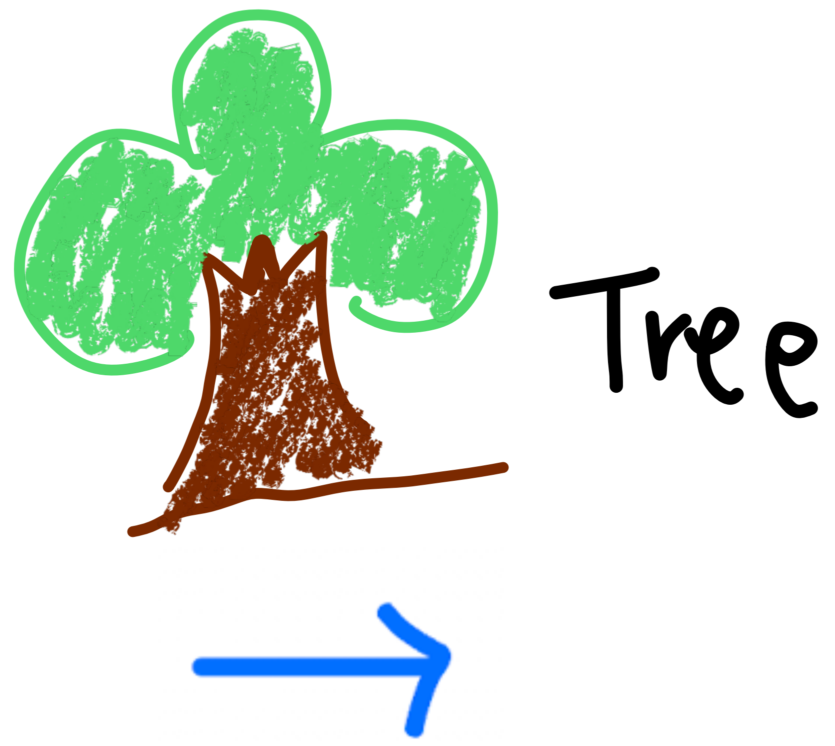
$$\pi_\varepsilon(\theta \mid X^t) \propto \pi(\theta)\mathbb{E}_{X^t|\theta}[\mathrm{ABC}(X^t, X_0, \varepsilon)]$$

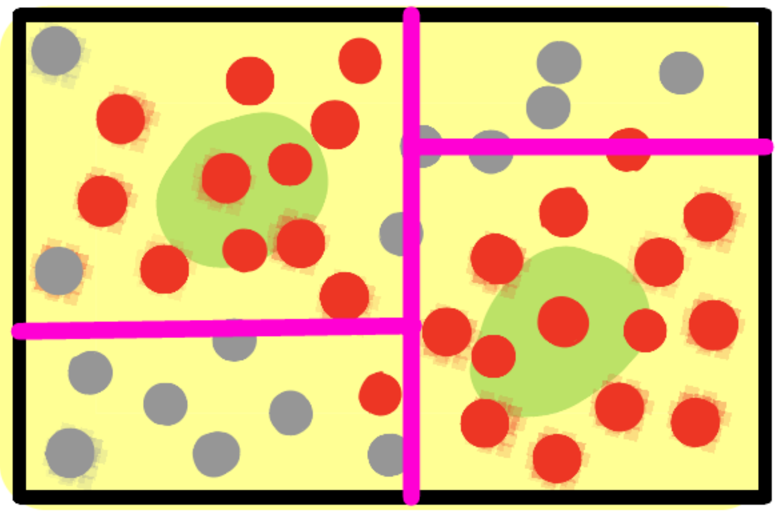
Outer Loop: Partitioning by 





$Y^{(t)} = \mathrm{ABC}(X^t, X_0, \varepsilon)$



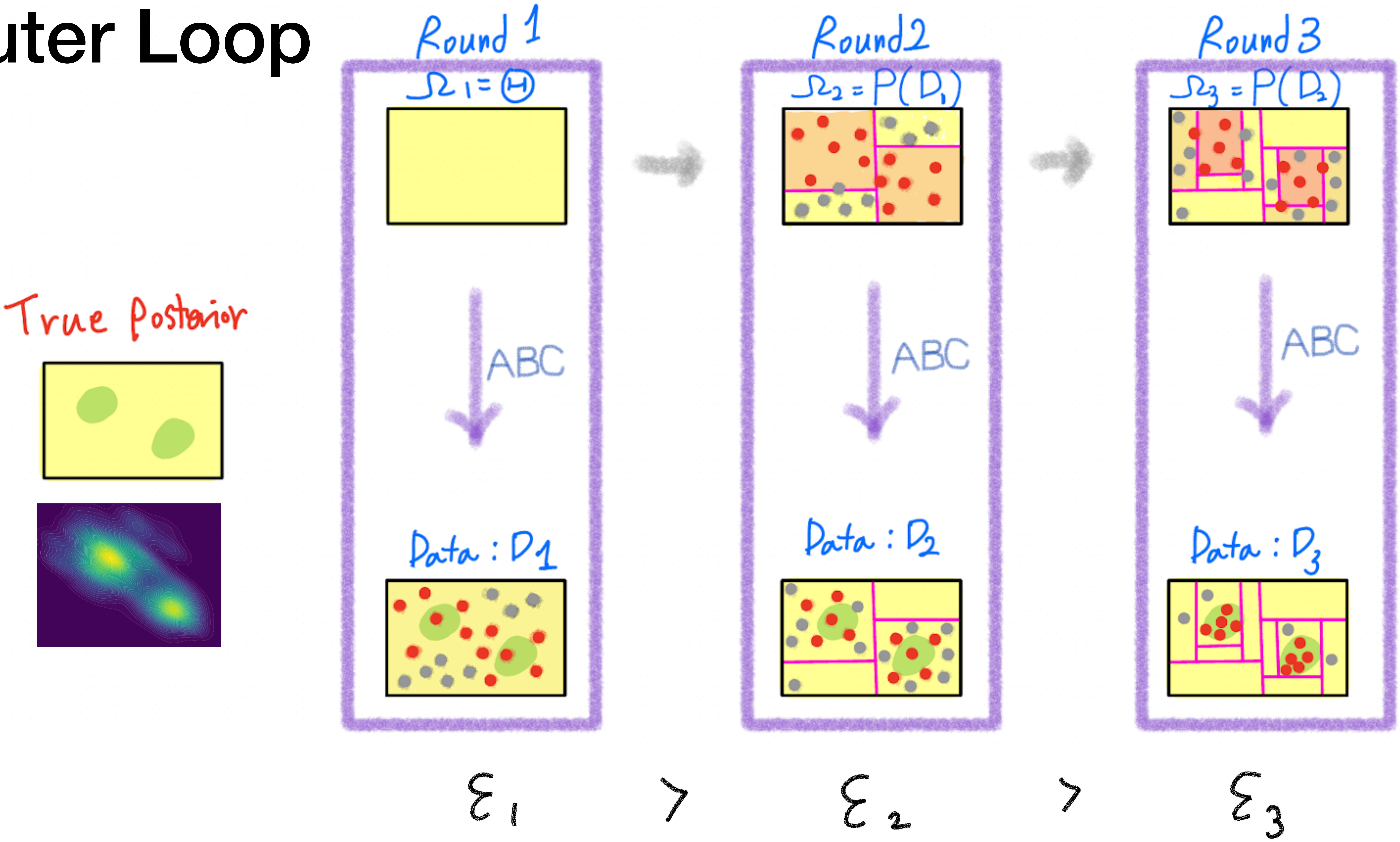


```
from sklearn.tree import DecisionTreeClassifier

tree_model = DecisionTreeClassifier()

tree_model.fit(thetas,Y)
```

Outer Loop


True Posterior
Round 1
$\Omega_1 = \Theta$
ABC
Data : D_1
Round2
$\Omega_2 = P(D_1)$
ABC
Data : D_2
Round 3
$\Omega_3 = P(D_2)$
ABC
Data : D_3
$\varepsilon_1 > \varepsilon_2 > \varepsilon_3$

Question 2) How to get q_k?

-> Thompson Sampling in the inner loop

Multi-arm Thompson sampling

Bandit Machine

$$\mu = \mathbb{E}[Y] = ?$$

$$Y \in \{0,1\}$$

$$\mu \sim Beta(\alpha, \beta)$$

$$\downarrow$$

$$Y \sim Ber(\mu)$$

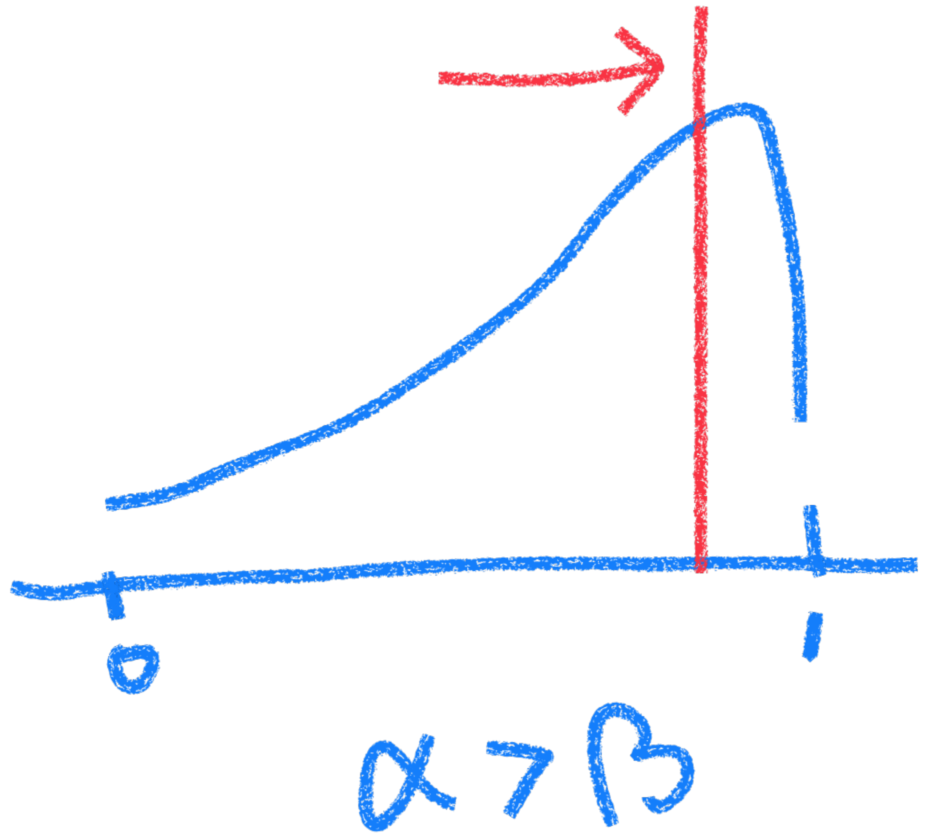
repeat

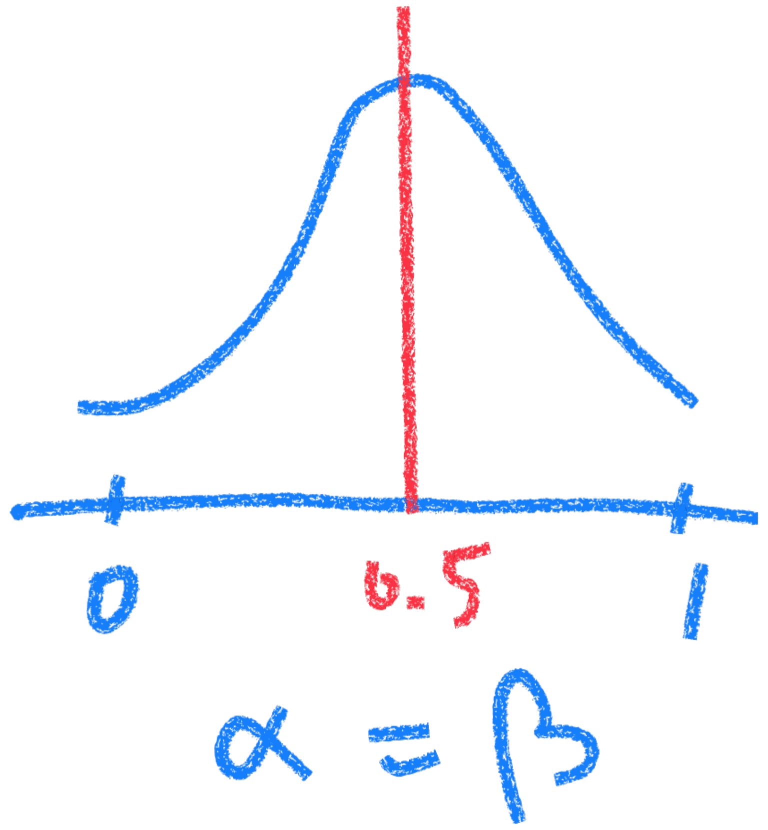
$$\downarrow$$

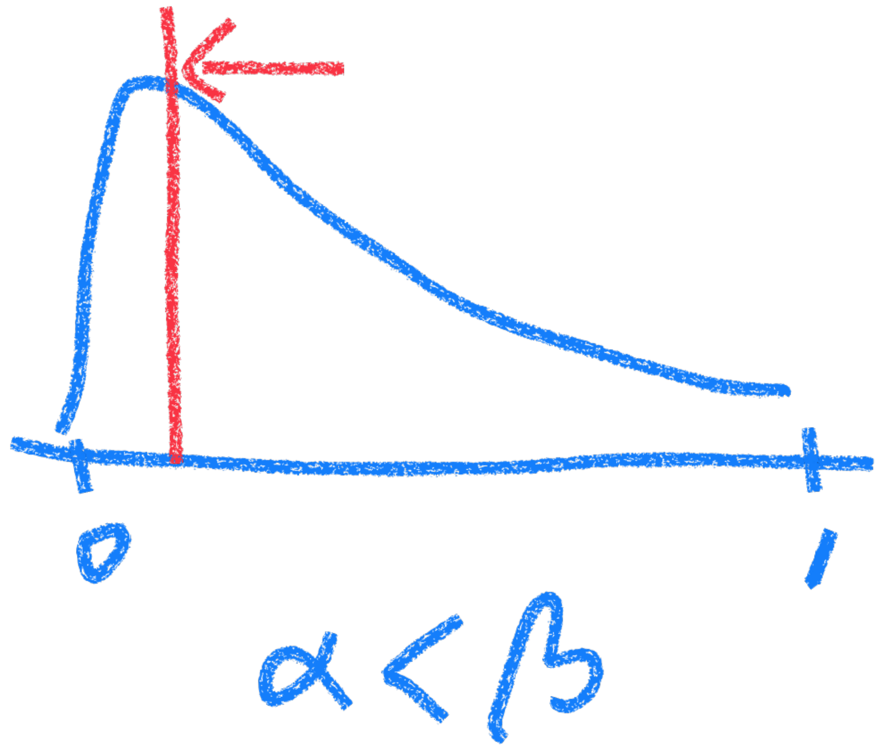
$$\mu \mid Y \sim Beta(\alpha', \beta')$$

$$\alpha' = \alpha + Y$$

$$\beta' = \beta + (1 - Y)$$







Multi-arm Thompson sampling

Bandit Machine

$\mu = \mathbb{E}[Y] = ?$

$Y \in \{0,1\}$

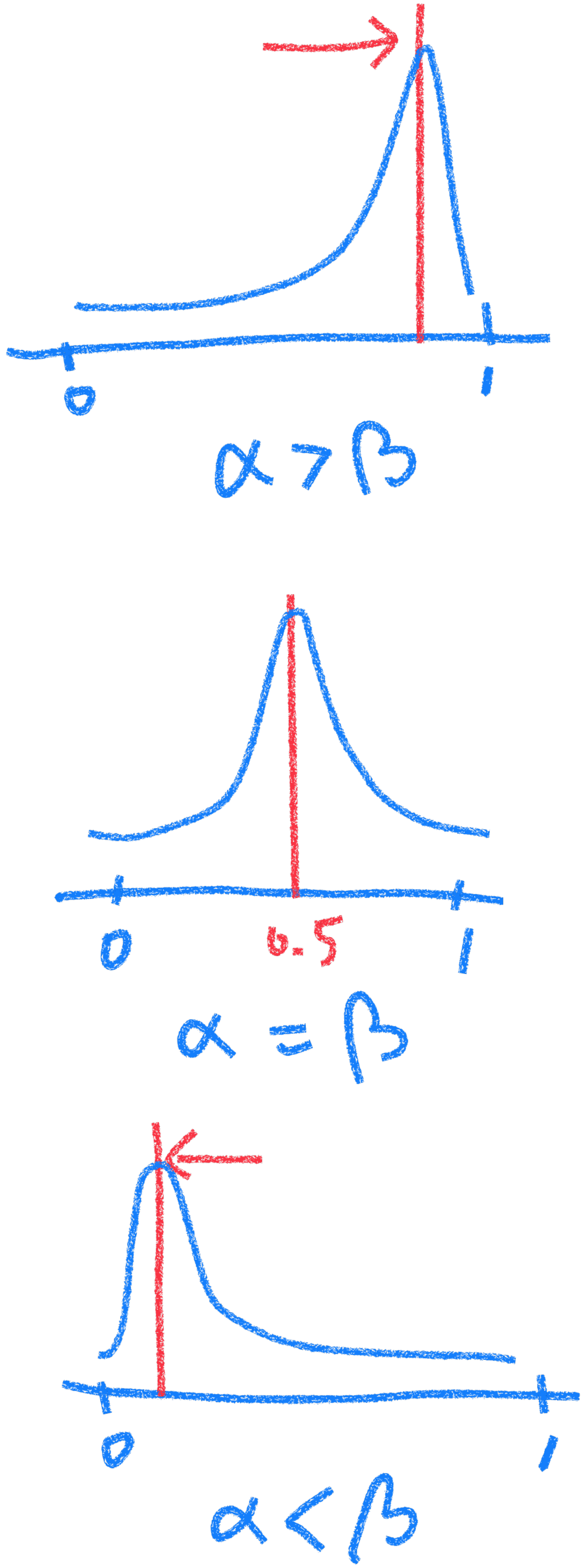
$\mu \sim Beta(\alpha, \beta)$

$Y \sim Ber(\mu)$ repeat

$\mu \,|\, Y \sim Beta(\alpha', \beta')$

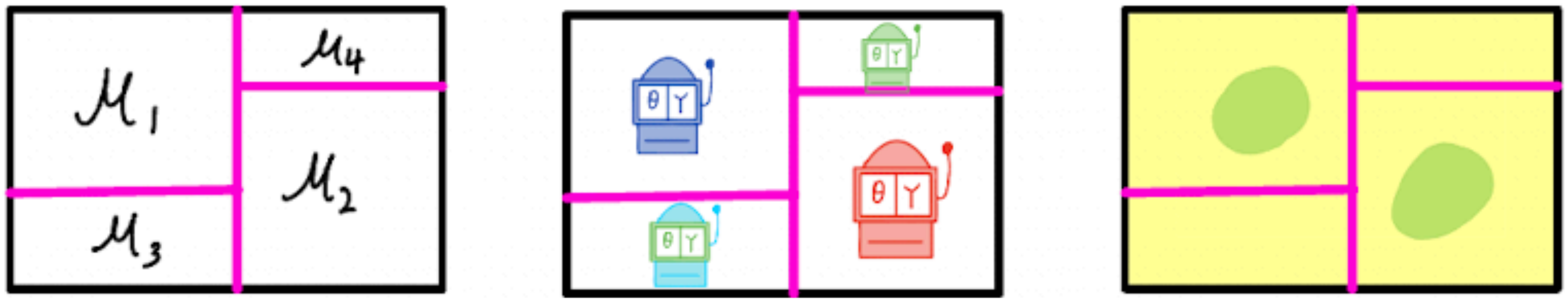
$\alpha' = \alpha + Y$

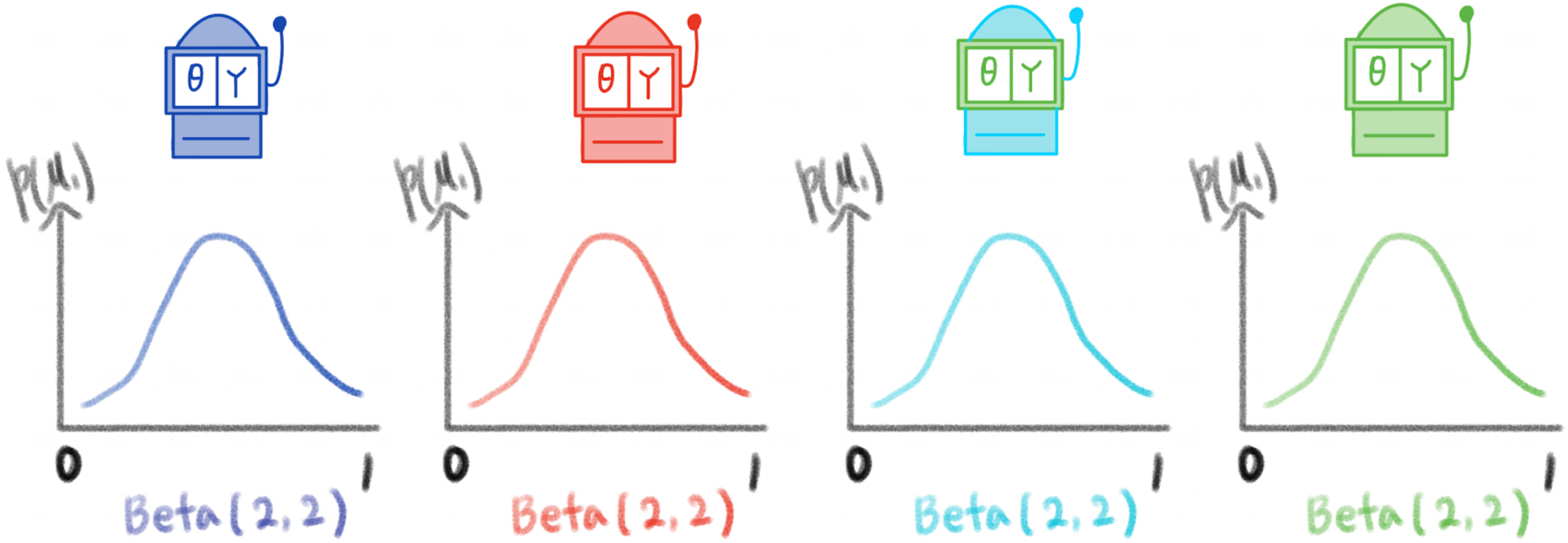
$\beta' = \beta + (1 - Y)$



Inner Loop: learning q_k

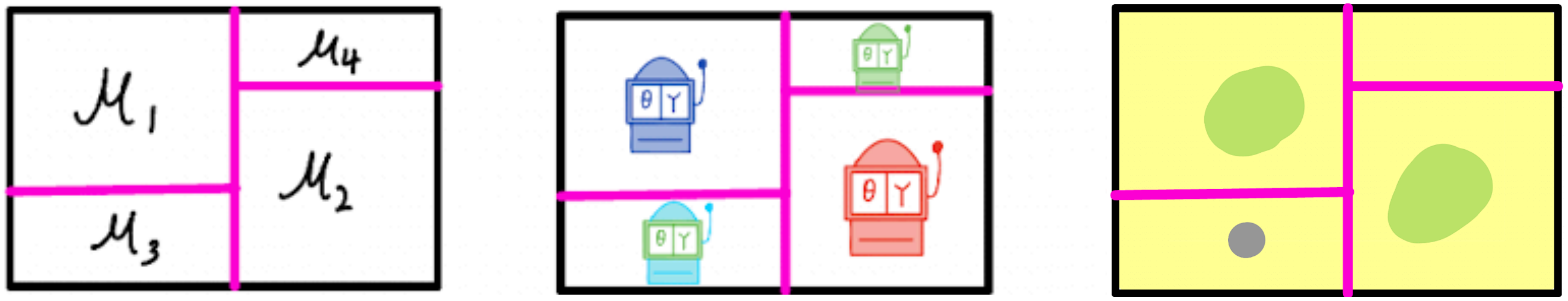
$\mu_k = \mathbb{E}[ABC \mid \Omega_k]$

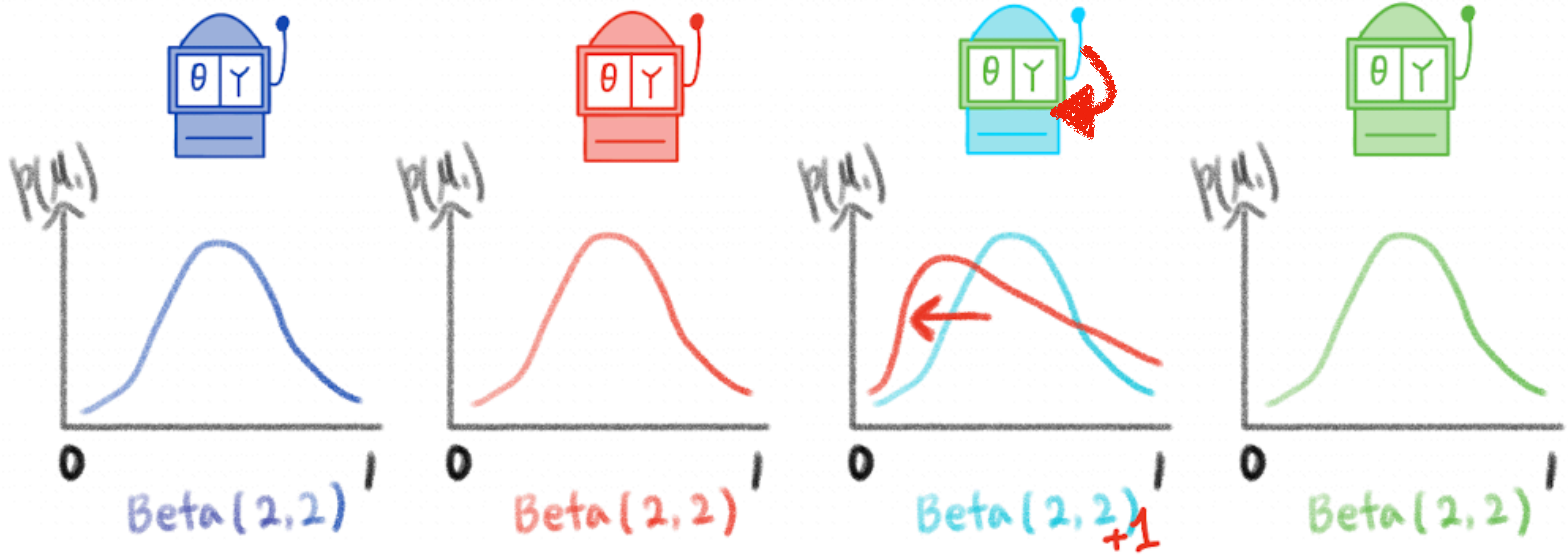




Inner Loop: learning q_k

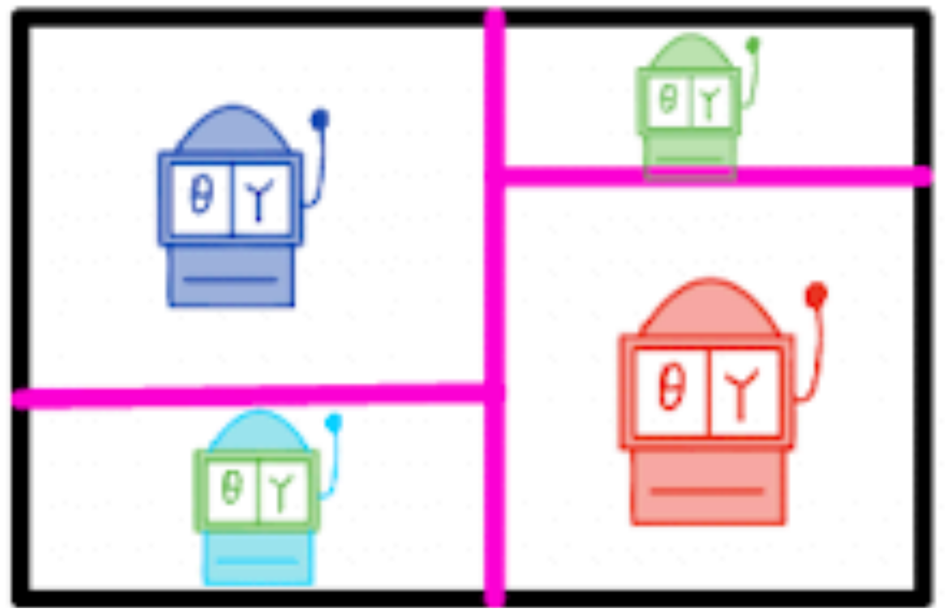
$$\mu_k = \mathbb{E}[ABC \mid \Omega_k]$$

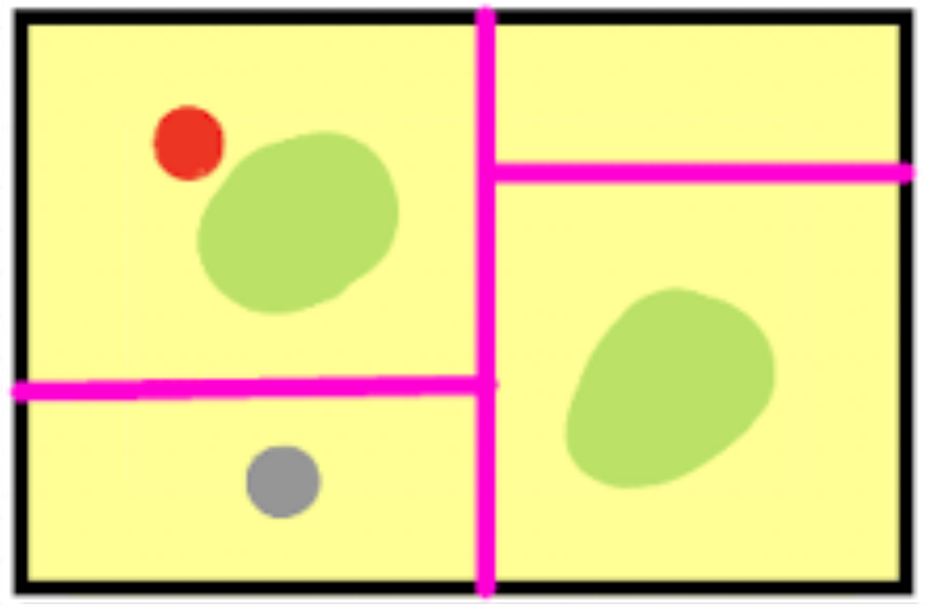


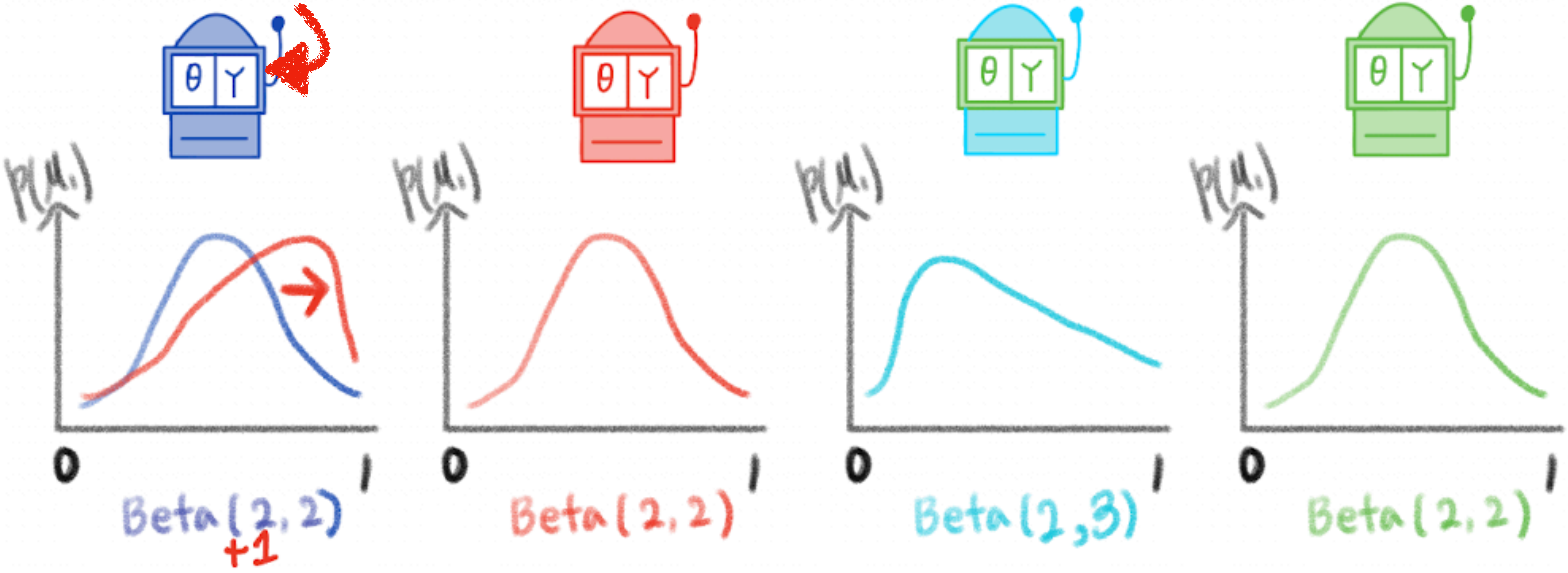


Inner Loop: learning q_k

$\mu_k = \mathbb{E}[ABC \mid \Omega_k]$

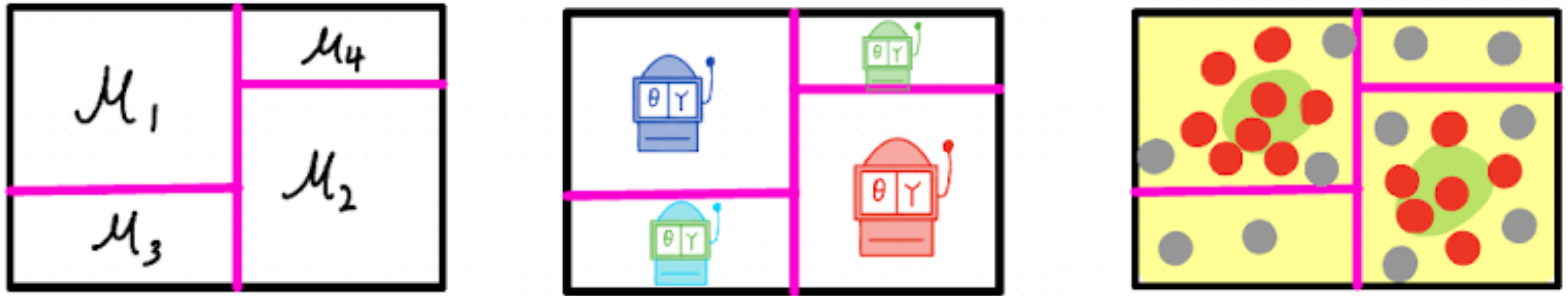


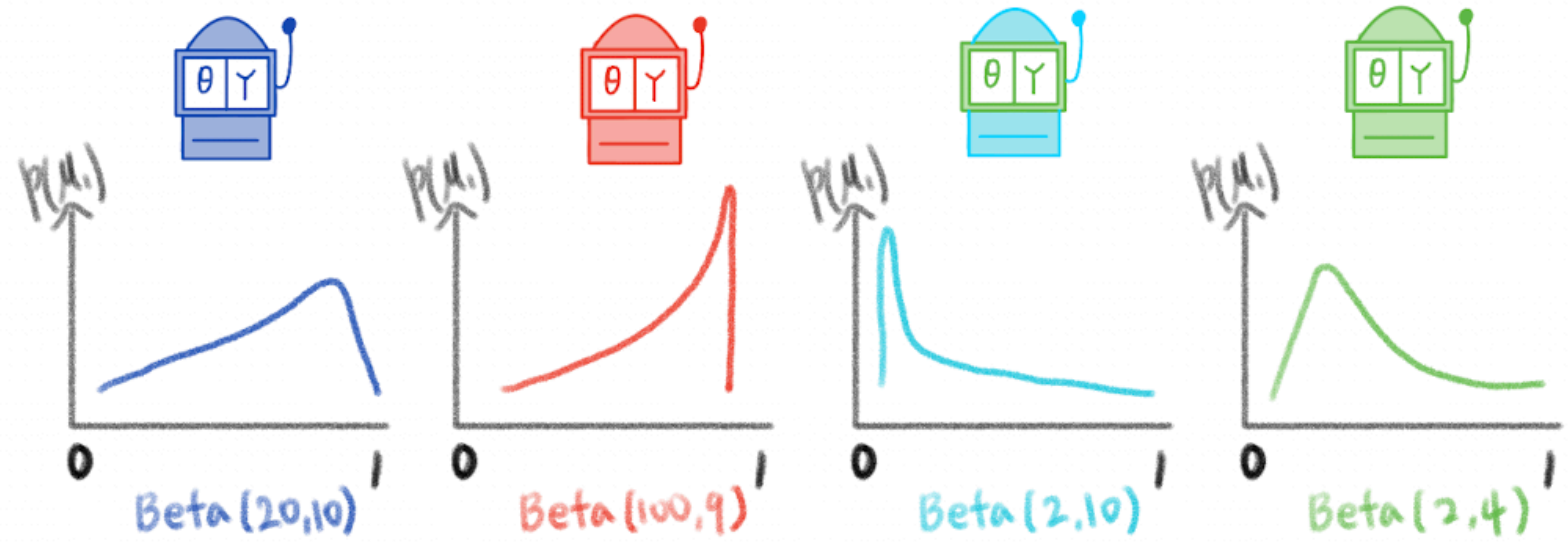




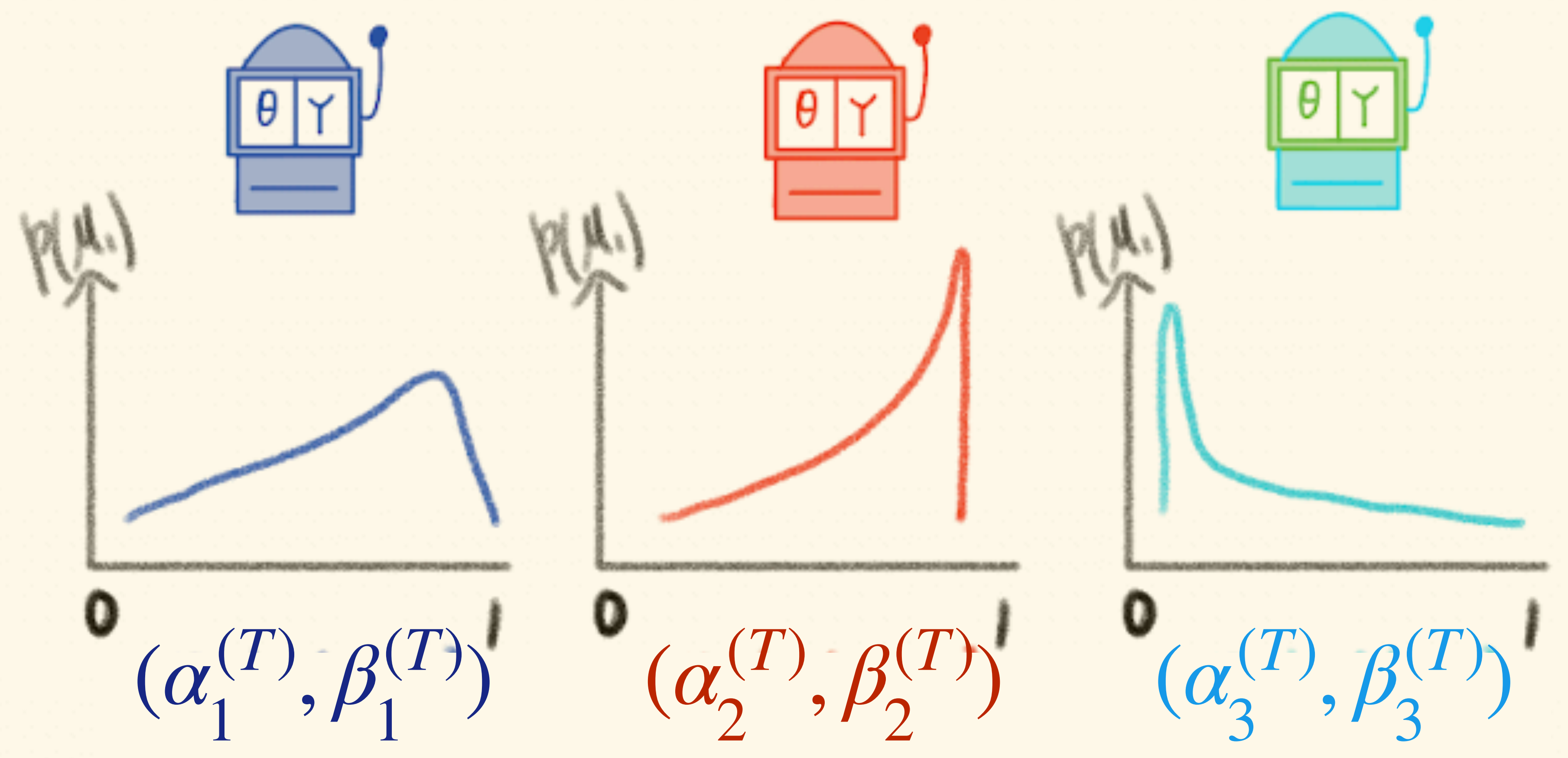
Inner Loop: learning q_k

$\mu_k = \mathbb{E}[ABC \mid \Omega_k]$





Arm Choice for Posterior Sampling



Exploitation and Exploration:

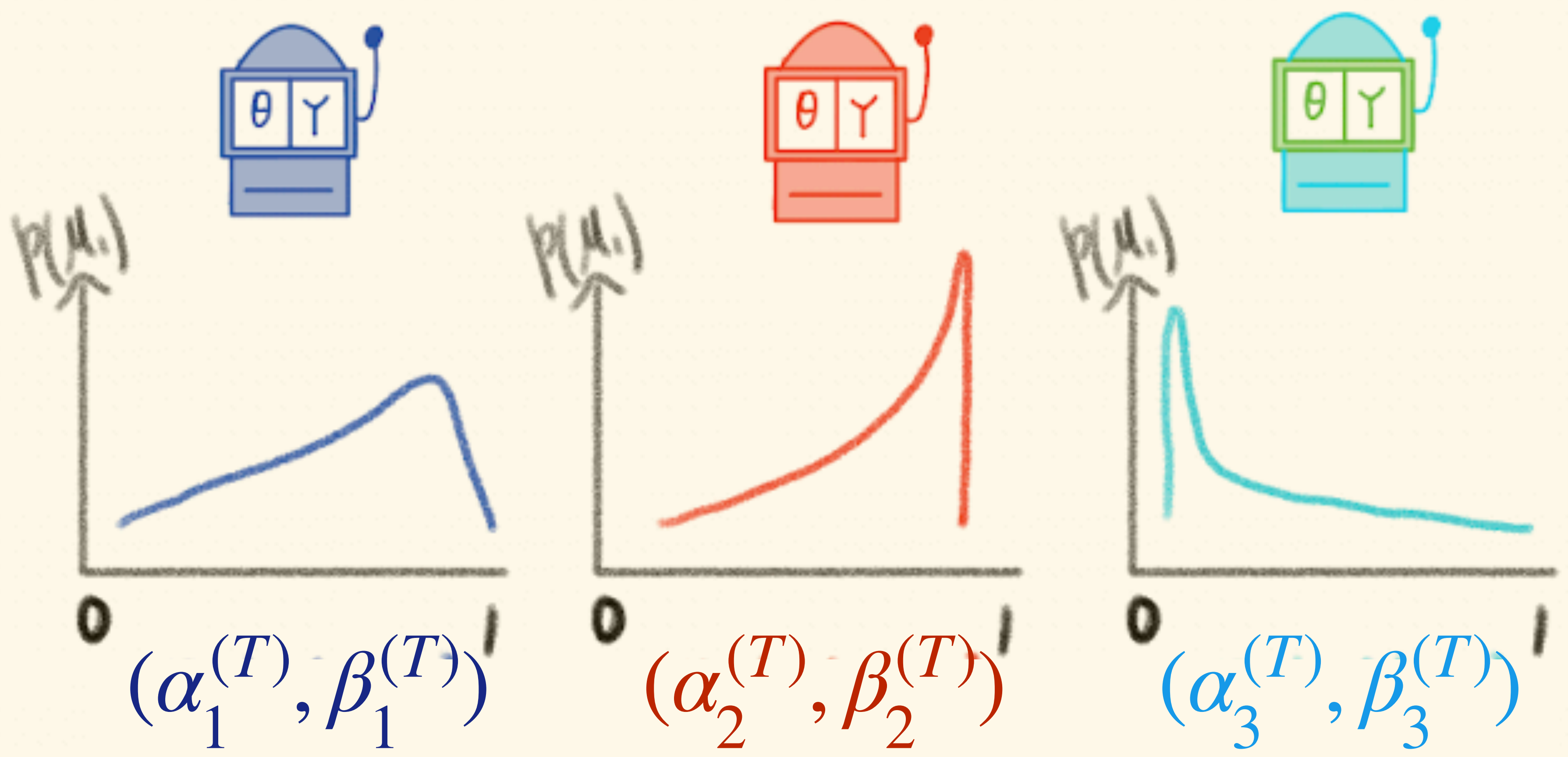
Estimate bin posterior by $\hat{p}^{(T)}_{\Omega_k} \propto \pi(\Omega_k) \dfrac{\alpha_k^{(T)}}{\alpha_k^{(T)} + \beta_k^{(T)}}$

Choose Arm by $Cat\{\hat{\mathbf{p}}^{(T)}_{\Omega}\}$, where $\hat{\mathbf{p}}^{(T)}_{\Omega} = (\hat{p}^{(T)}_{\Omega_1}, \hat{p}^{(T)}_{\Omega_2}, \hat{p}^{(T)}_{\Omega_3}, \hat{p}^{(T)}_{\Omega_4})$

$$\pi_\varepsilon(\theta \mid X^t) \propto \pi(\theta) \mathbb{E}_{X^t \mid \theta}[\mathrm{ABC}(X^t, X_0, \varepsilon)]$$

$$\rightarrow \qquad \pi_\varepsilon(\Omega_k \mid X^t) \propto \pi(\Omega_k) \mathbb{E}_{X^t \mid \theta, \theta \sim \pi(\theta \mid \Omega_k)}[\mathrm{ABC}(X^t, X_0, \varepsilon)]$$

Arm Choice for Posterior Sampling



Exploitation and Exploration:

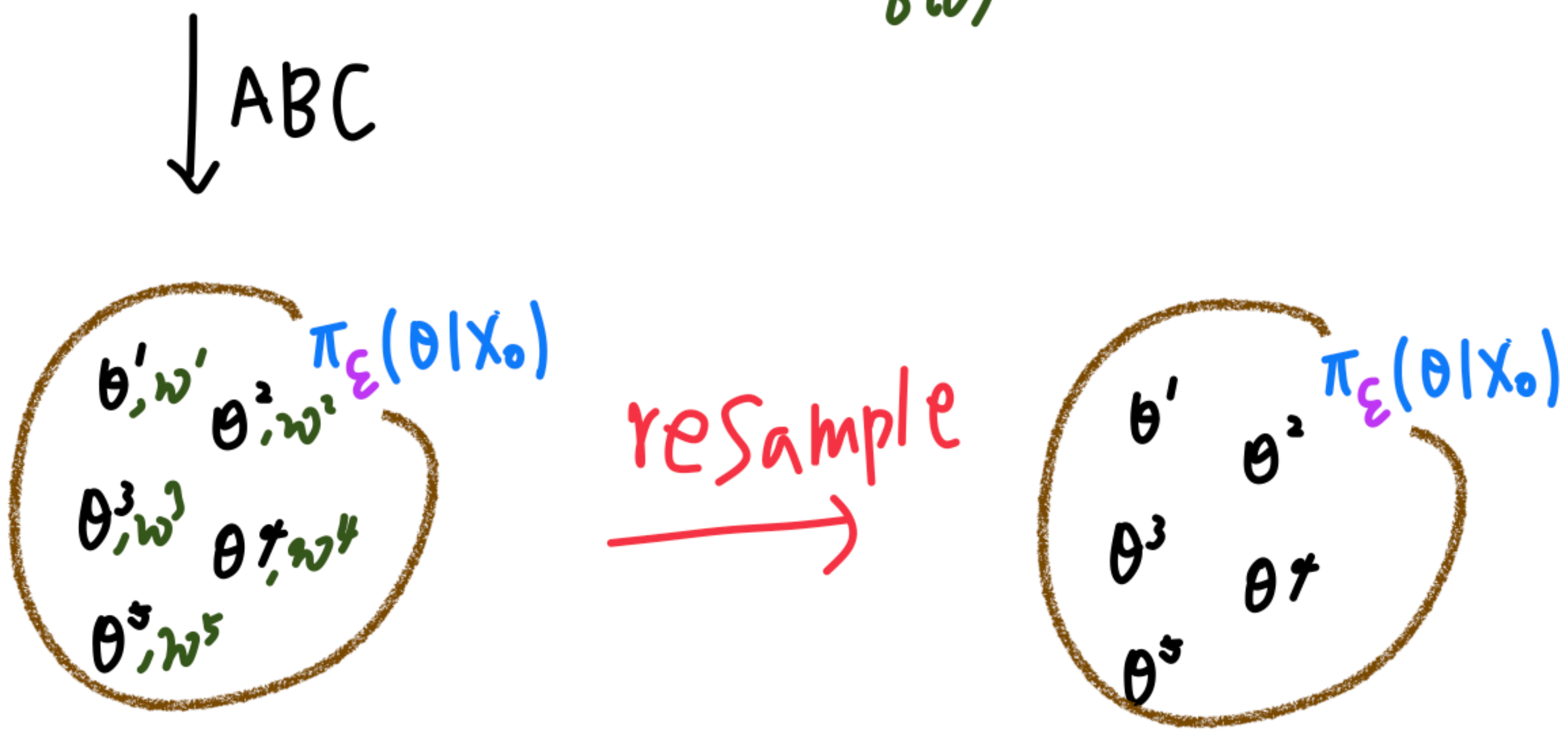
Estimate bin posterior by $\hat{p}^{(T)}_{\Omega_k} \propto \pi(\Omega_k)\dfrac{\alpha_k^{(T)}}{\alpha_k^{(T)} + \beta_k^{(T)}}$

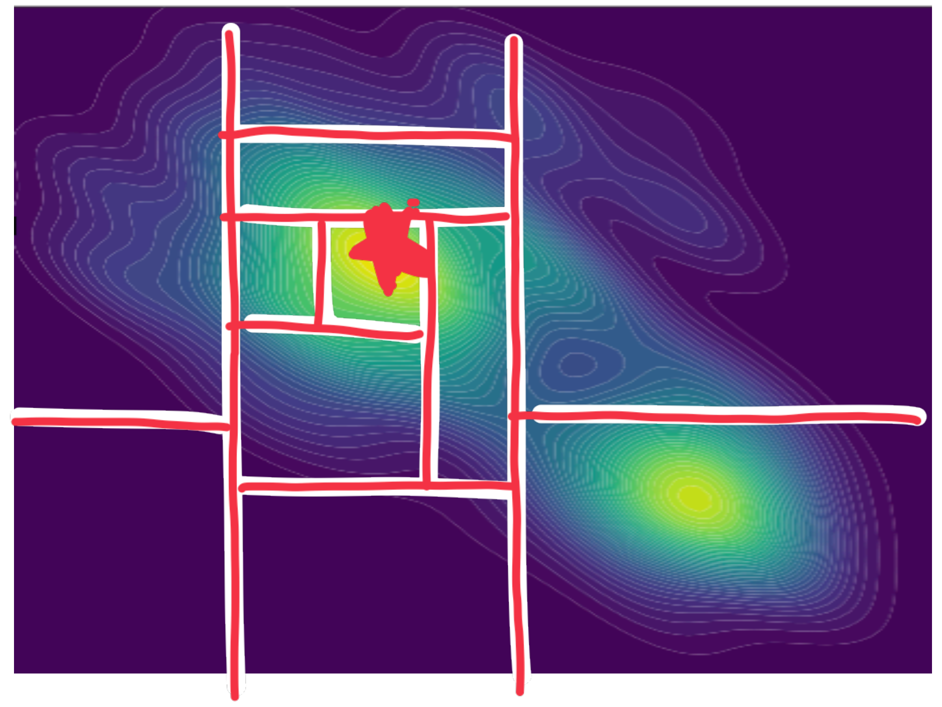
Choose Arm by $Cat\{\hat{\mathbf{p}}^{(T)}_{\Omega}\}$, where $\hat{\mathbf{p}}^{(T)}_{\Omega} = (\hat{p}^{(T)}_{\Omega_1}, \hat{p}^{(T)}_{\Omega_2}, \hat{p}^{(T)}_{\Omega_3}, \hat{p}^{(T)}_{\Omega_4})$

Pulling the best arm may not always be good!

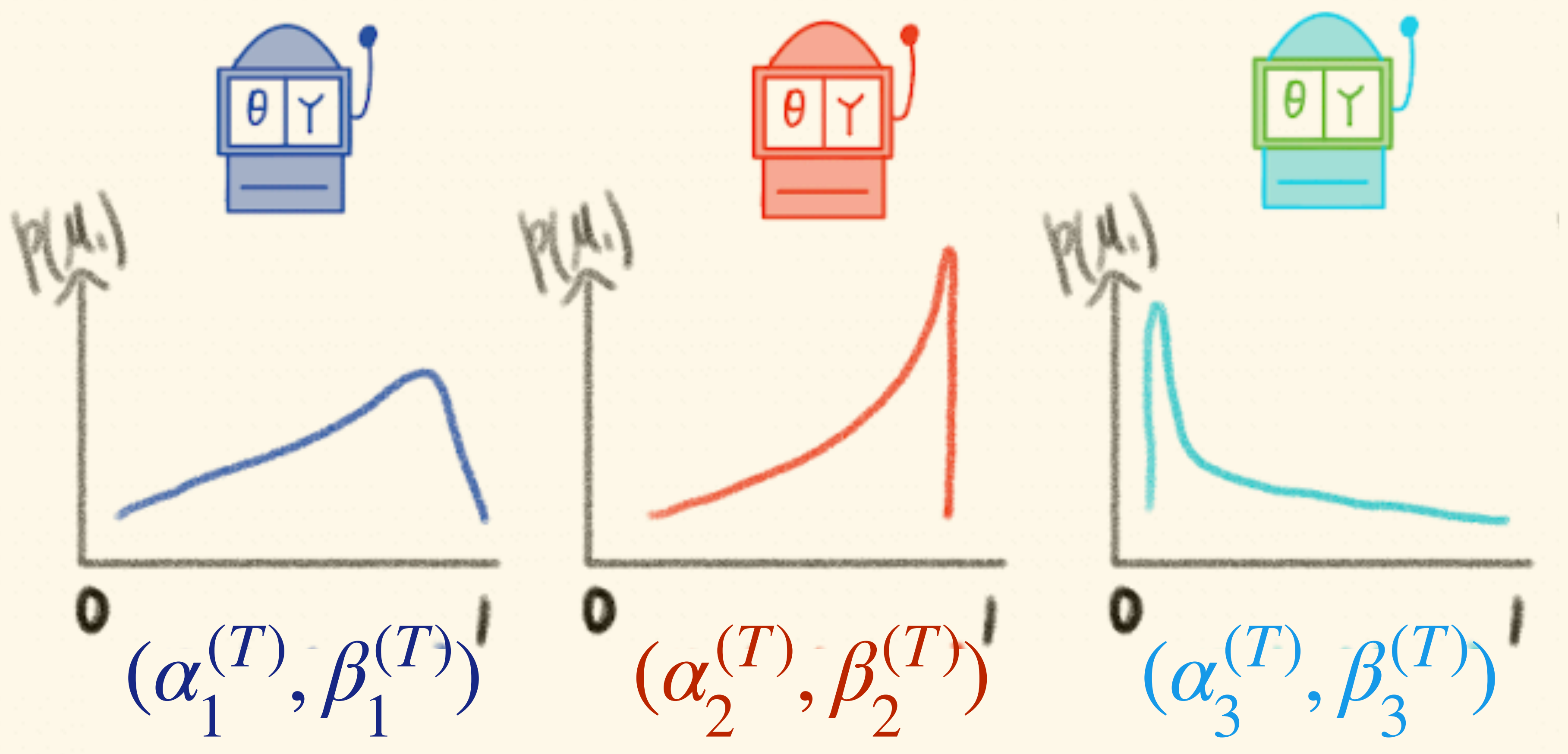
$\theta \sim q(\theta)$, $w(\theta) = \dfrac{\pi(\theta)}{q(\theta)}$

-> Importance weight variance





Arm Choice for Posterior Sampling



Regularized Exploitation and Exploration:

Estimate bin posterior by $\hat{p}_{\Omega_k}^{(T)} \propto \pi(\Omega_k) \dfrac{\alpha_k^{(T)}}{\alpha_k^{(T)} + \beta_k^{(T)}}$

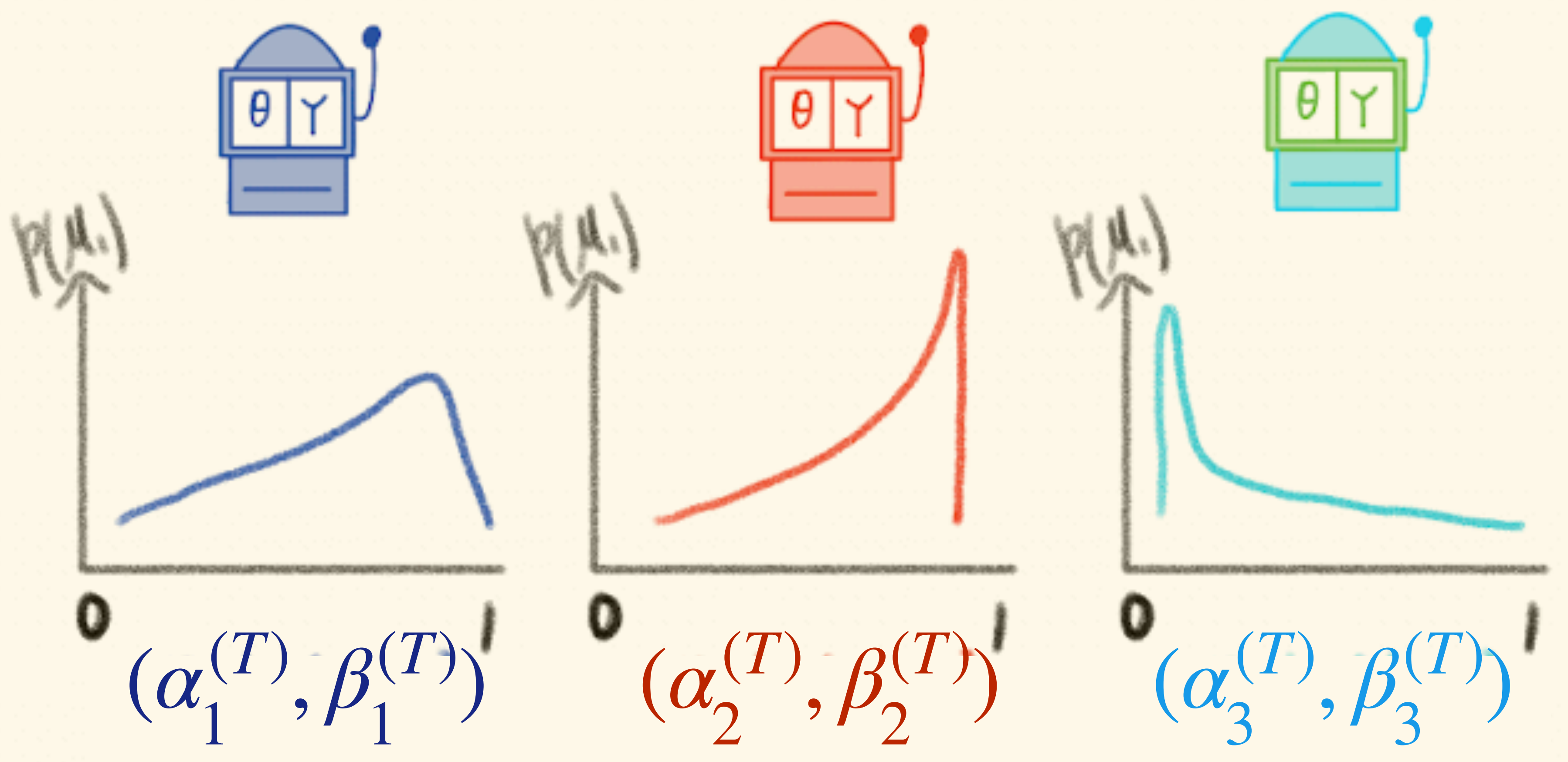
$(q_1, \ldots, q_4) = \mathbf{q}^*(\hat{\mathbf{p}}_\Omega^{(T)})$, where $\mathbf{q}^*(\mathbf{p}) = \arg \max_{\mathbf{q}:\|\mathbf{q}\|_1=1} \omega_{\mathbf{p}}(\mathbf{q})$.

Choose Arm by $Cat\{q_1, q_2, q_3, q_4\}$

For example: $\omega_{\mathbf{p}}^2(\mathbf{q}) = \dfrac{\sum_{k=1}^K q_k p_k / \pi_k}{\sum_{k=1}^K p_k \pi_k / q_k}$.

sampling efficiency (Alsing et al. 2018)

Arm Choice for Posterior Sampling



Regularized Exploitation:

Estimate bin posterior by $\hat{p}_{\Omega_k}^{(T)} \propto \pi(\Omega_k) \dfrac{\alpha_k^{(T)}}{\alpha_k^{(T)} + \beta_k^{(T)}}$

$(q_1, \ldots, q_4) = \mathbf{q}^*(\hat{\mathbf{p}}_\Omega^{(T)})$, where $\mathbf{q}^*(\mathbf{p}) = \arg\max_{\mathbf{q}:\|\mathbf{q}\|_1=1} \omega_{\mathbf{p}}(\mathbf{q})$.

Choose Arm by $Cat\{q_1, q_2, q_3, q_4\}$

Given regularity conditions, our algorithm is very good at optimizing the regret

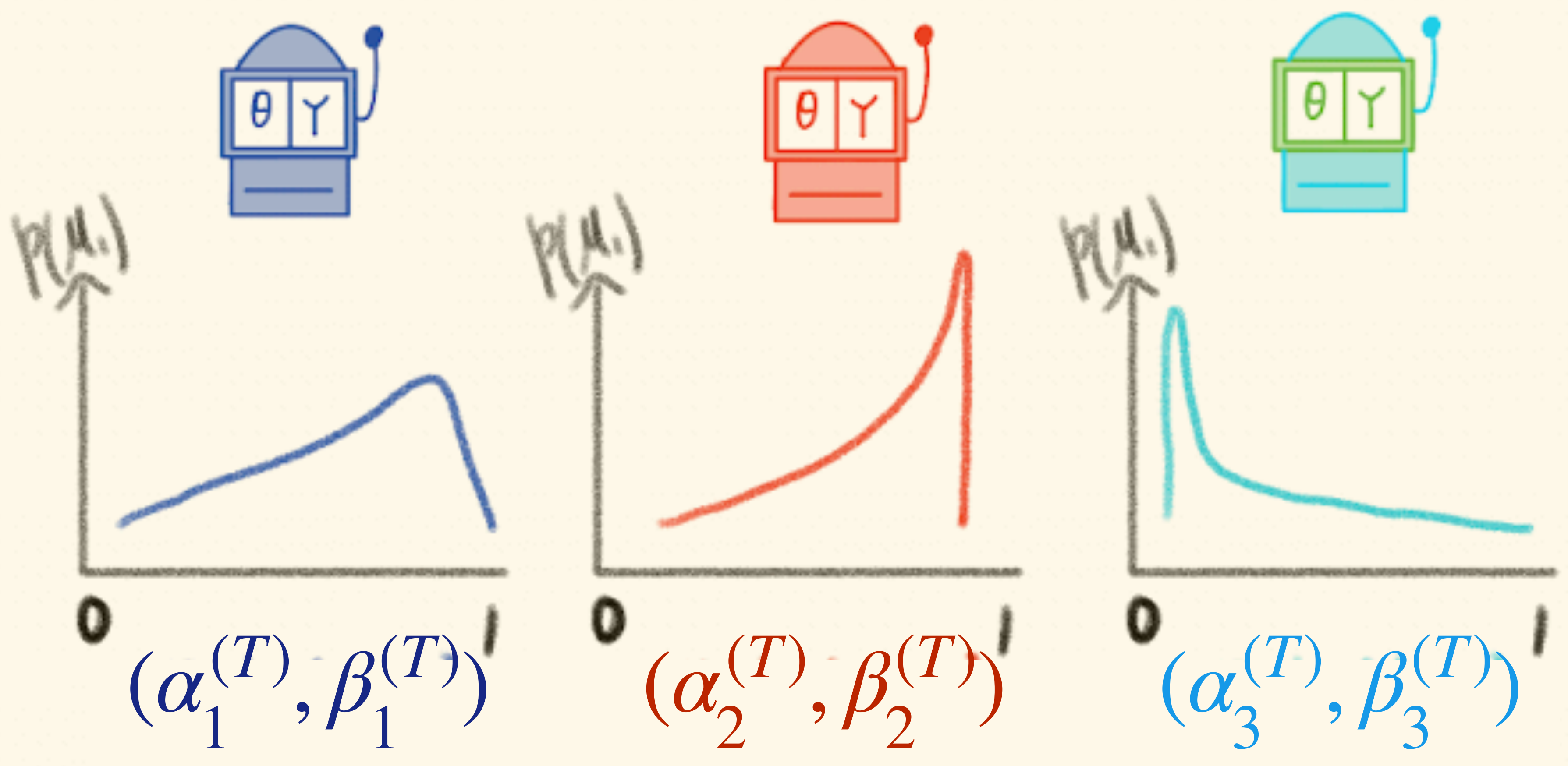
$$R_{TL}(\mathscr{A}, \mathbf{p}) := \omega_{\mathbf{p}}(\mathbf{q}^*) - \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \omega_{\mathbf{p}}\left(\hat{\mathbf{q}}^{(t)}\right)\right]$$

Theorem Given $\pi(\Omega_k) = 1/K, k = 1,\ldots,K$, and with $\mathscr{A}'$: our algorithm,

$$\sup_p R_{TL}(\mathscr{A}', p) \lesssim \frac{K^2 \log^2(T)}{T} \text{ and } \inf_{\mathscr{A}} \sup_p R_{TL}(\mathscr{A}, p) \gtrsim \frac{K^2}{T}$$

where the infimum is taken over all possible strategies $\mathscr{A}$. T : time, K : # of bins

Arm Choice for Maximum a Posteriori (MAP)

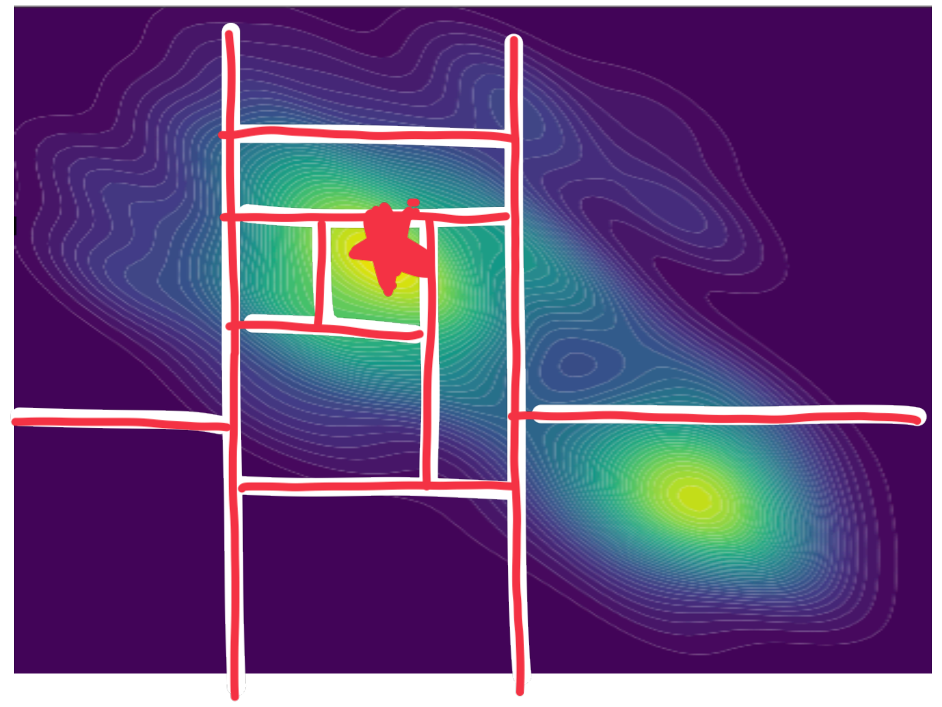


Estimated average posterior in each bin:

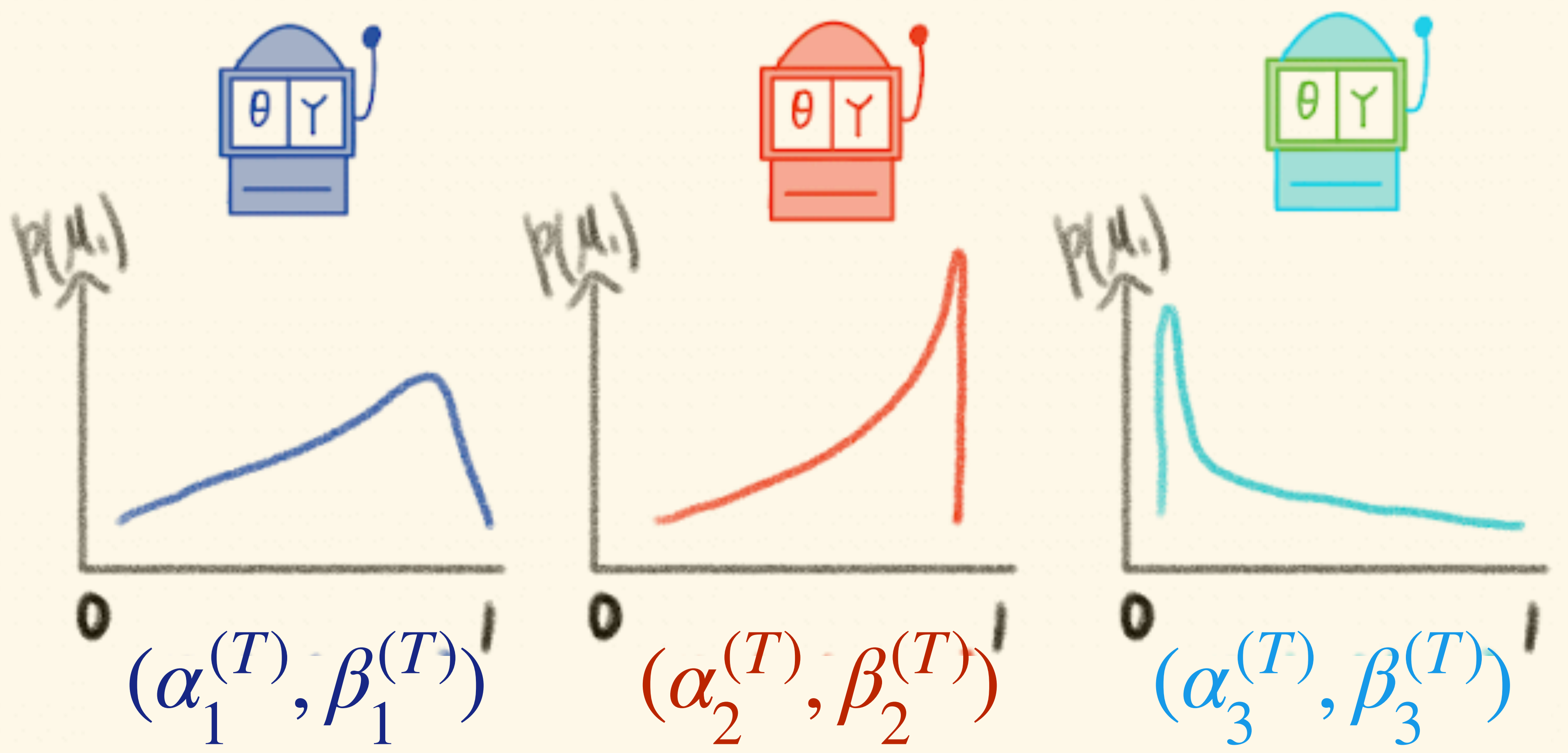
$$\hat{p}_k^{\Omega}(\theta \mid X) = \frac{\hat{p}_{\Omega_k}^{(T)}}{|\Omega_k|} \text{ for } \theta \in \Omega_k,\ \hat{p}_{\Omega_k}^{(T)} \propto \pi(\Omega_k) \times \frac{\alpha_k^{(T)}}{\alpha_k^{(T)} + \beta_k^{(T)}}$$

Exploitation: Pull $\arg\max\{\hat{p}_k^{\Omega}(\theta \mid X)\}_{k=1}^{K}$

Importance weight variance: No worries-!



Arm Choice for Maximum a Posteriori (MAP)



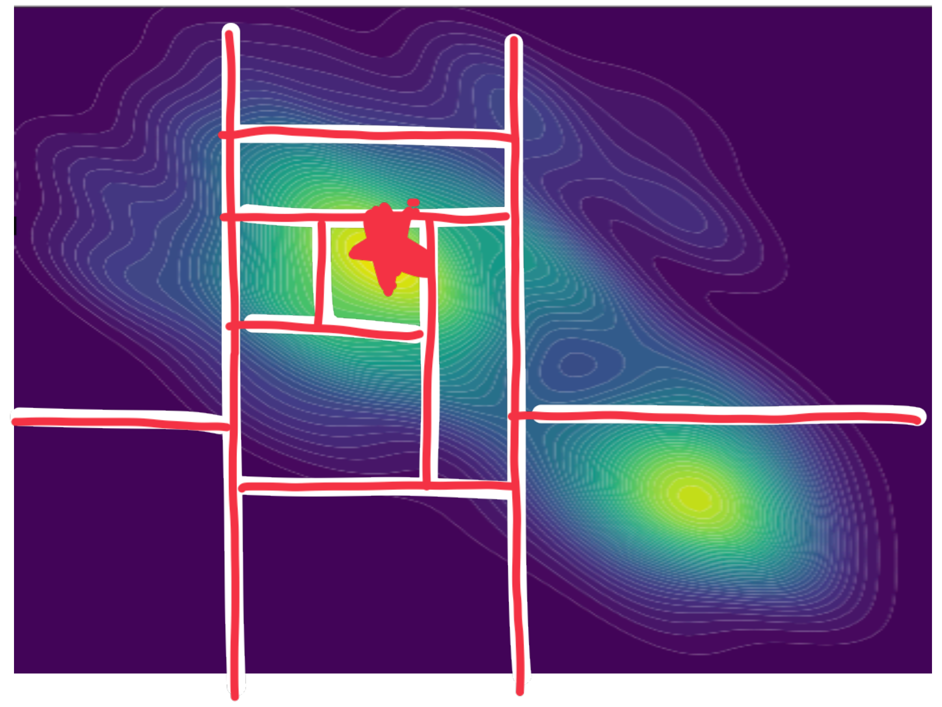
Exploration: $(\eta_1, \eta_2, \eta_3, \eta_4) \sim \bigotimes_{k=1}^{K} \mathrm{Beta}\left(\alpha_k^{(T)}, \beta_k^{(T)}\right)$

Sampled average posterior in each bin:

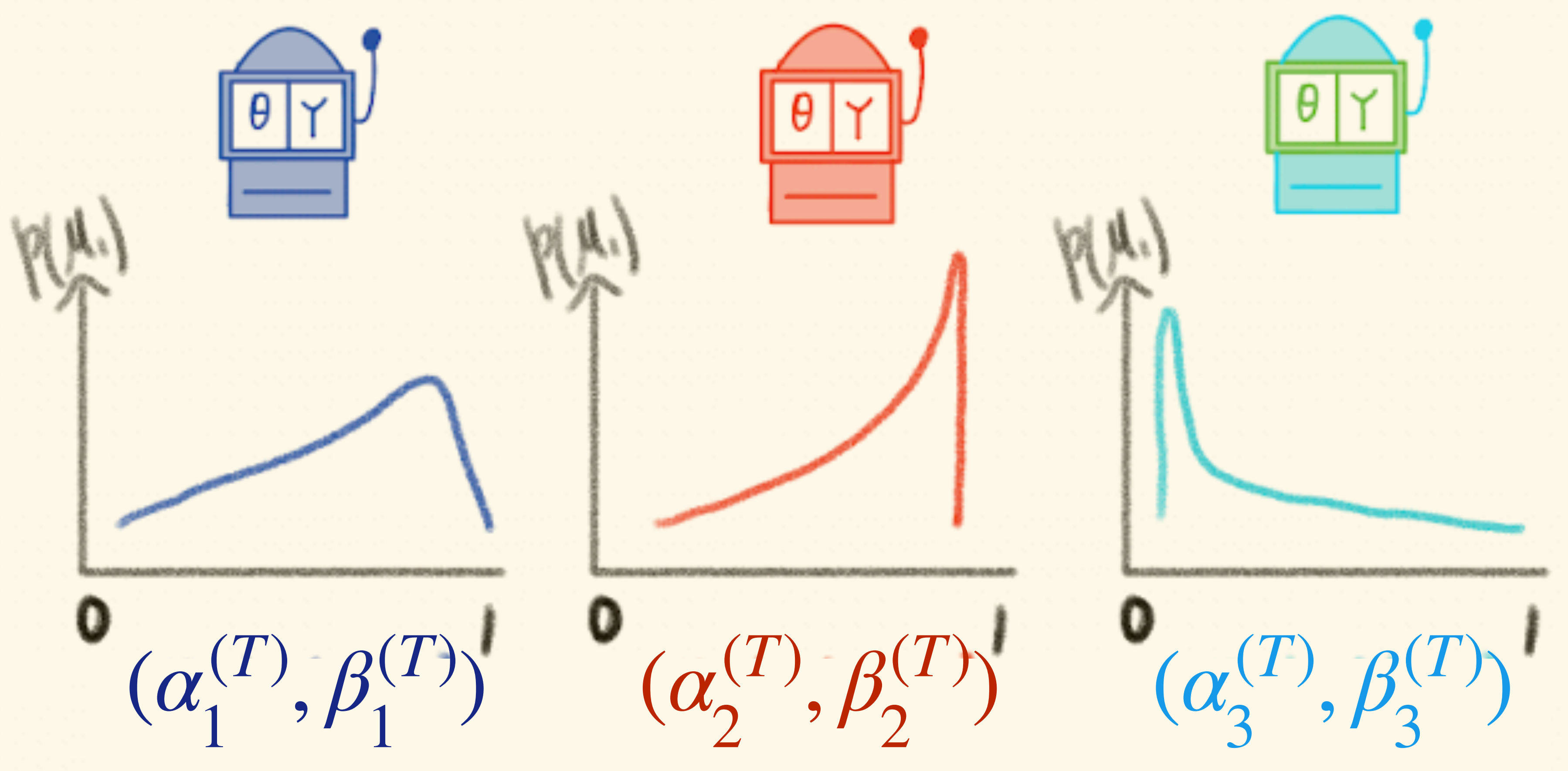
$$\hat{p}_\epsilon^\Omega(\theta \mid X) = \frac{\hat{p}_{\Omega_k}^{(T)}}{|\Omega_k|} \text{ for } \theta \in \Omega_k, \quad \hat{p}_{\Omega_k}^{(T)} \propto \pi(\Omega_k) \times \eta_k$$

Exploitation: Pull $\arg\max \{\hat{p}_k^\Omega(\theta \mid X)\}_{k=1}^K$

Importance weight variance: No worries-!



Arm Choice for Maximum a Posteriori (MAP)



Exploration: $(\eta_1, \eta_2, \eta_3, \eta_4) \sim \bigotimes_{k=1}^{K} \mathrm{Beta}\left(\alpha_k^{(T)}, \beta_k^{(T)}\right)$

Sampled average posterior in each bin:

$$\hat{p}_\epsilon^\Omega(\theta \,|\, X) = \frac{\hat{p}_{\Omega_k}^{(T)}}{|\Omega_k|} \text{ for } \theta \in \Omega_k, \quad \hat{p}_{\Omega_k}^{(T)} \propto \pi(\Omega_k) \times \eta_k$$

Exploitation: Pull $\arg\max\{\hat{p}_k^\Omega(\theta \,|\, X)\}_{k=1}^K$

Given regularity conditions, our algorithm is very good at optimizing the regret

$$r_T = \pi_\varepsilon(\theta^{\mathrm{MAP}} \,|\, X) - \pi_\varepsilon(\tilde{\theta}^{(T)} \,|\, X)$$, where $\tilde{\theta}^{(T)}$: the center of the bin chosen at time T

Theorem Given $\pi(\Omega_k) = 1/K, k = 1,\ldots,K$, with $\mathscr{A}'$: our algorithm,

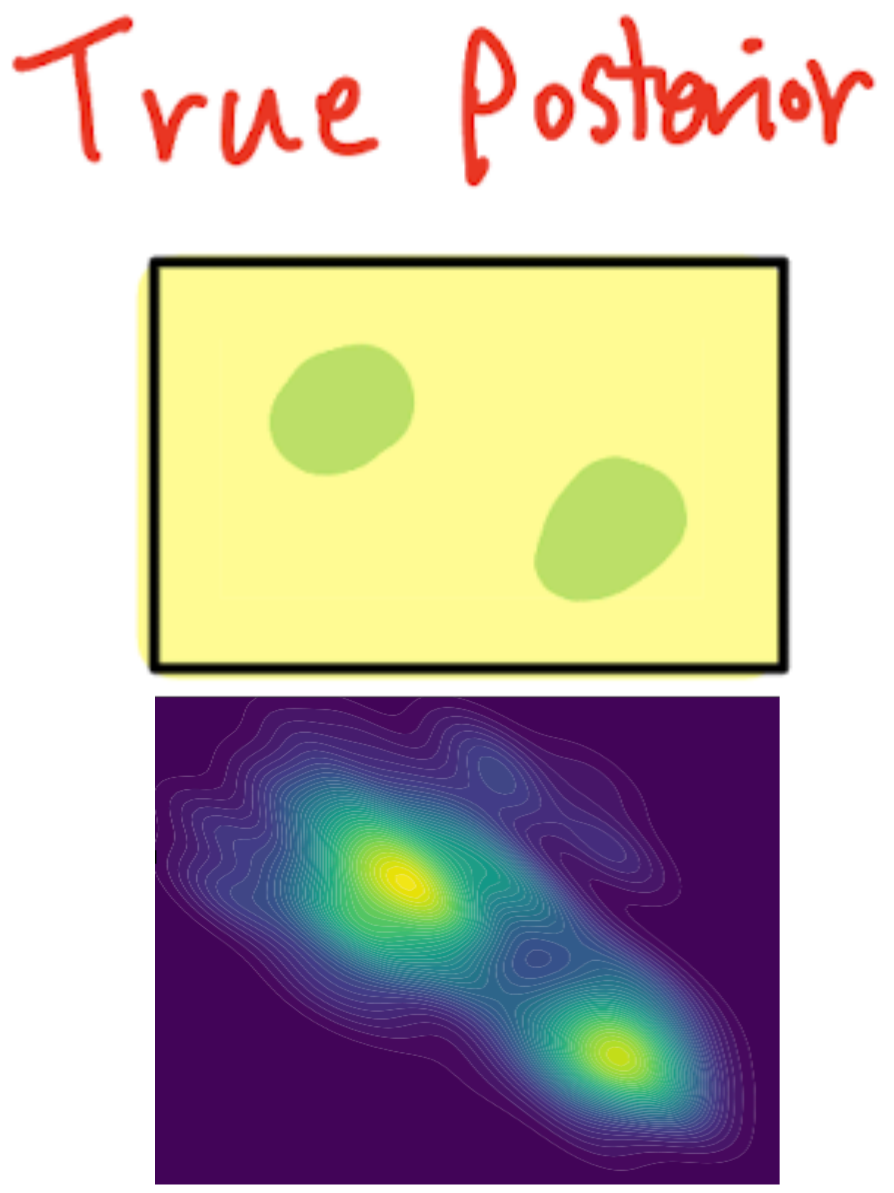
for any $\mu = (p_{\Omega_1}, \ldots, p_{\Omega_K})$, there exists a constant Γ_μ such

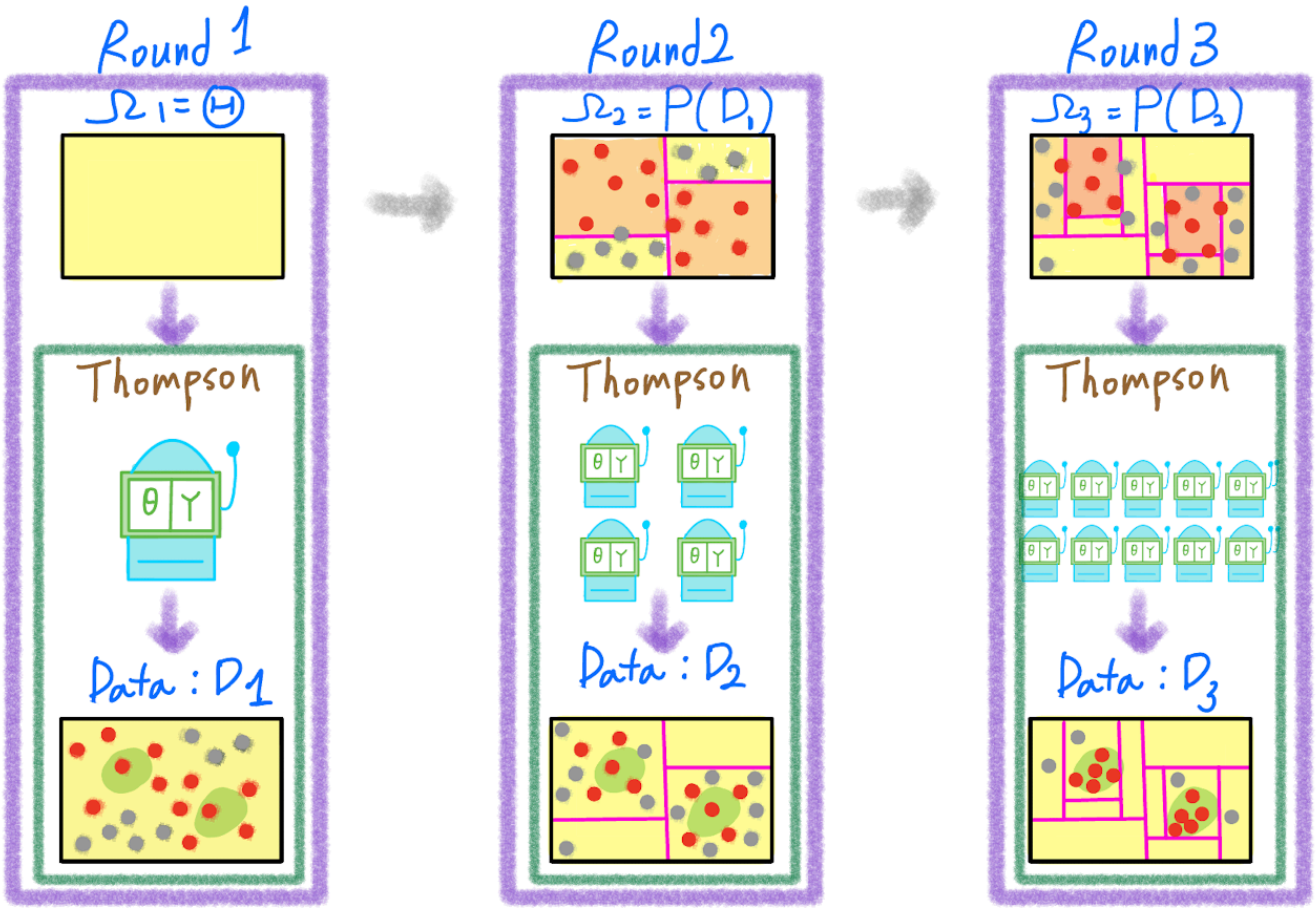
$$P(\hat{k}_{\mathscr{A}'}^{(T)} \neq k^*) = \exp(-T(\Gamma_\mu + o(1))) .$$

Moreover, any algorithm $\mathscr{A}$ must satisfy $P(\hat{k}_{\mathscr{A}}^{(T)} \neq k^*) \geq \exp(-T(2\Gamma_\mu + o(1)))$.

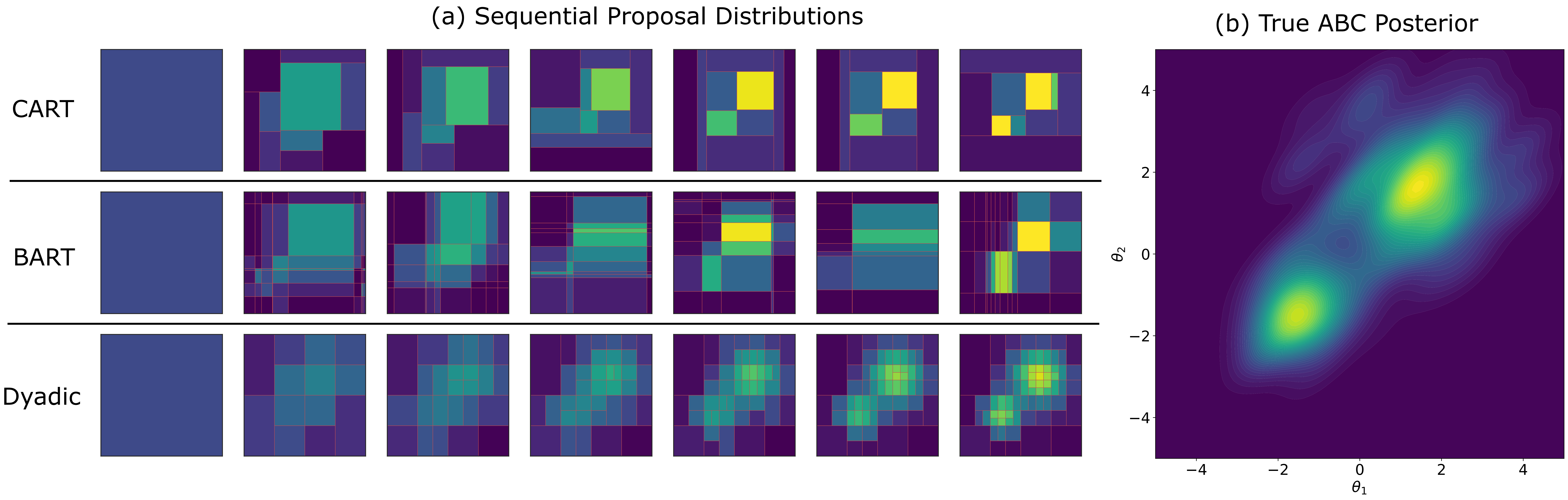
k^*: the bin index having the true MAP

Algorithm


True Posterior


Round 1
$\Omega_1 = \Theta$
Thompson
θ | Y
Data : D_1
Round2
$\Omega_2 = P(D_1)$
Thompson
Data : D_2
Round 3
$\Omega_3 = P(D_2)$
Thompson
Data : D_3

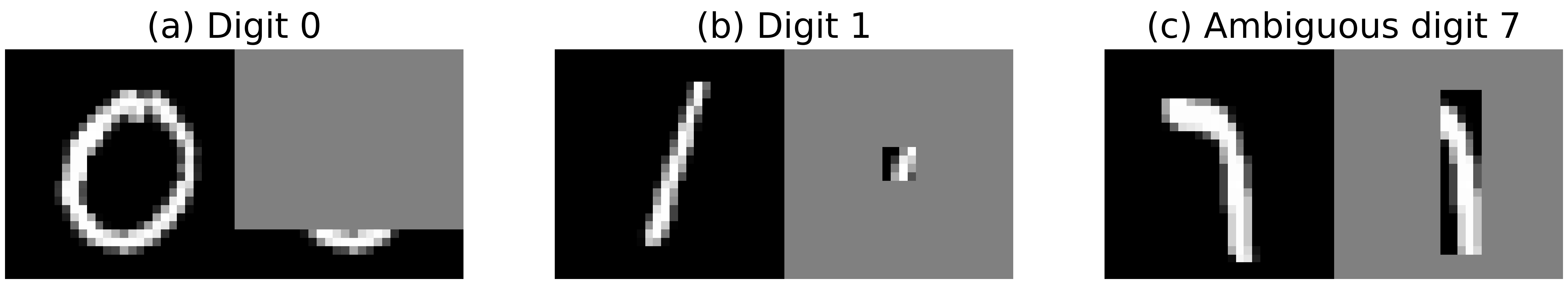
Example Posterior Sampling



Example MAP (Maximum a posteriori)

Experiment

Generative Unmasking Problem

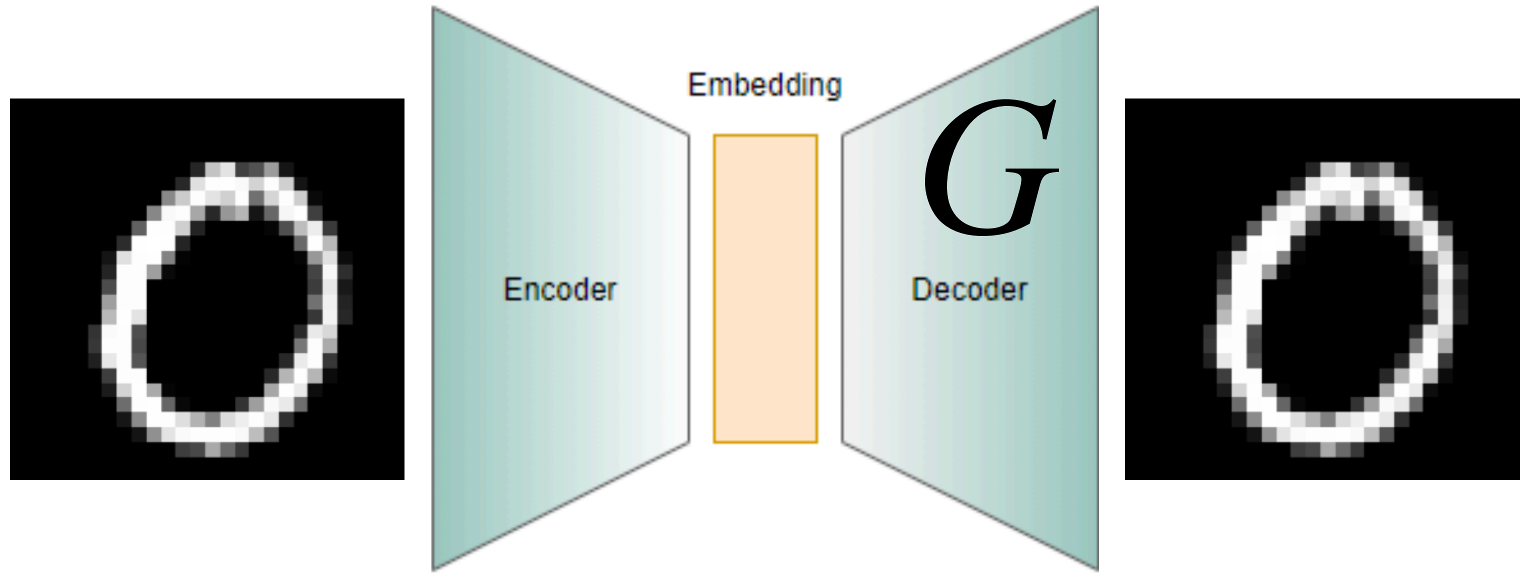


Model: $X = G(\theta) + \epsilon,\ \theta \sim N(0, I_{d_\theta}),\ \epsilon \sim N(0, \sigma_0^2 I_{d_x})$

Data: $X = (X_{\text{obs}}, X_{\text{mask}})$

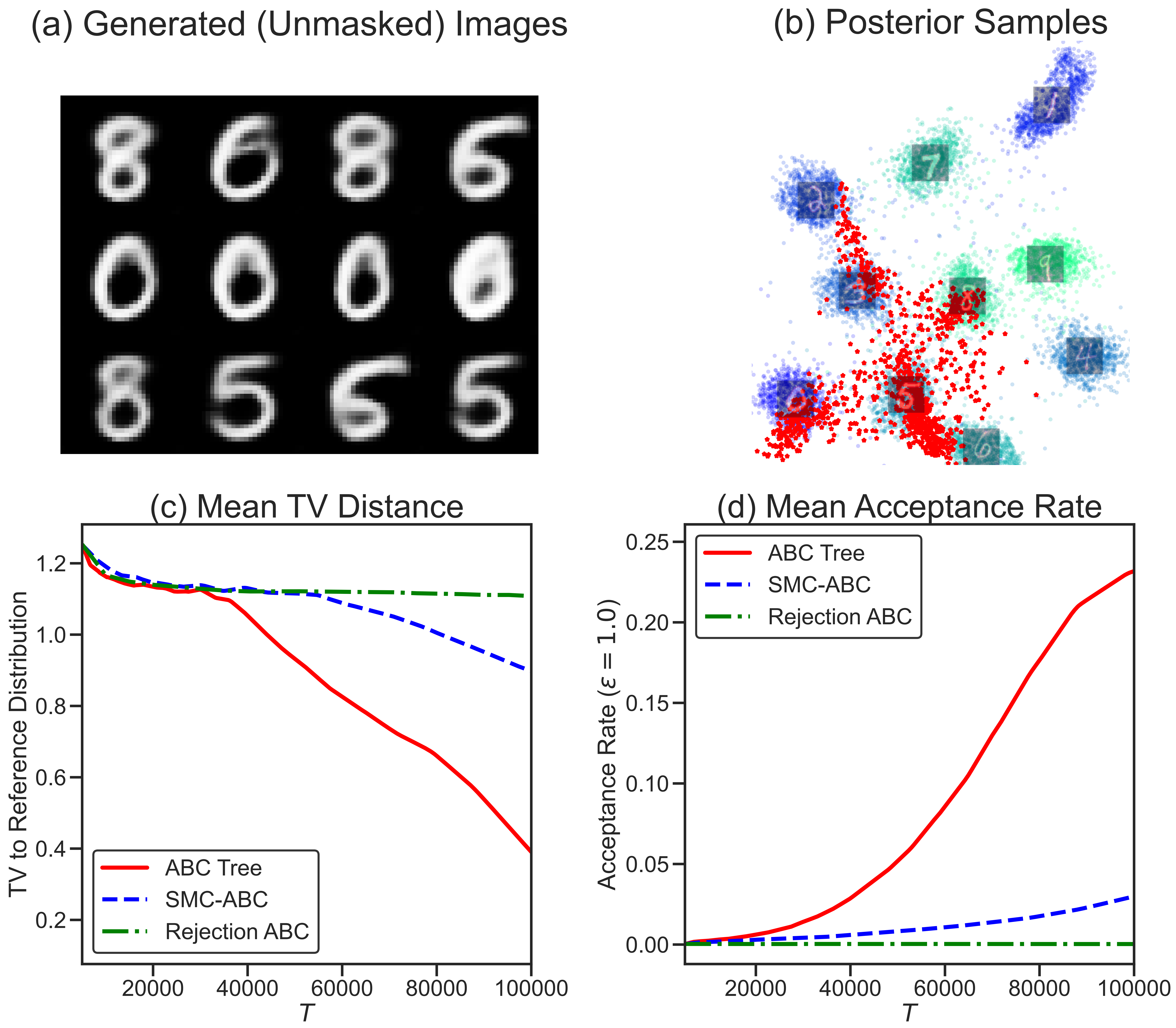
Sampling: $\tilde{\theta} \sim \pi(\theta \mid X_{\text{obs}})$

Unmasking: $\tilde{X}_{\tilde{\theta}} = G(\tilde{\theta}) + \epsilon,\ \epsilon \sim N(0, \sigma_0^2 I_{d_x})$



Posterior Sampling: Multi-modality captured-!

$$X = (X_{\text{obs}}, X_{\text{mask}})$$

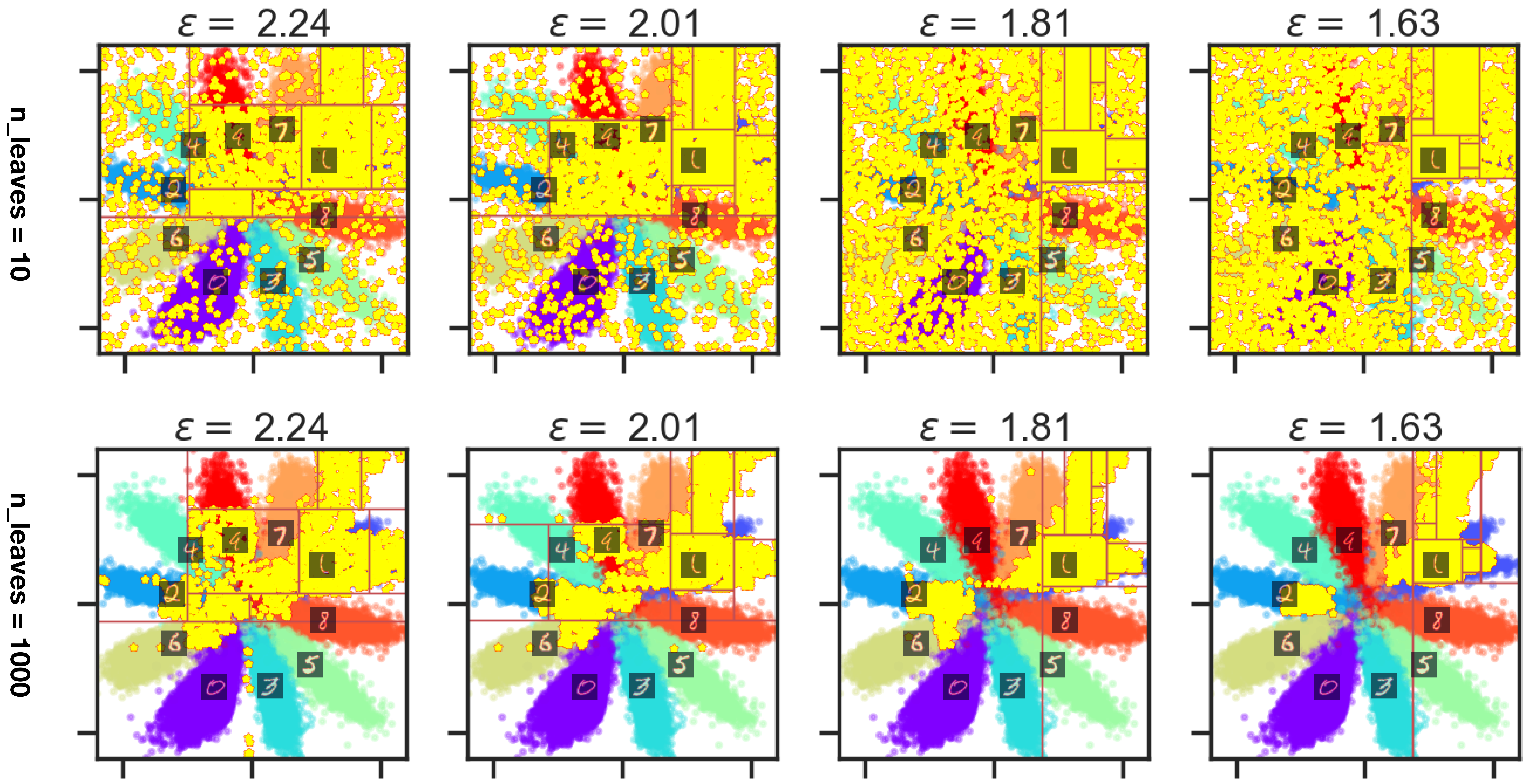


Likelihood-Free MAP:

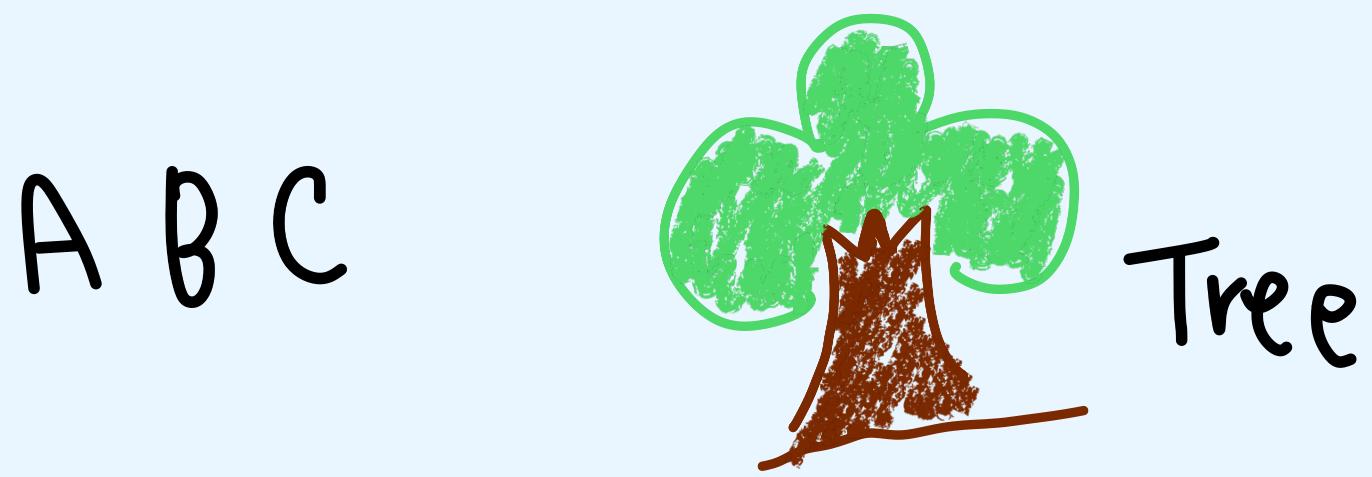
smaller MAP regions-!

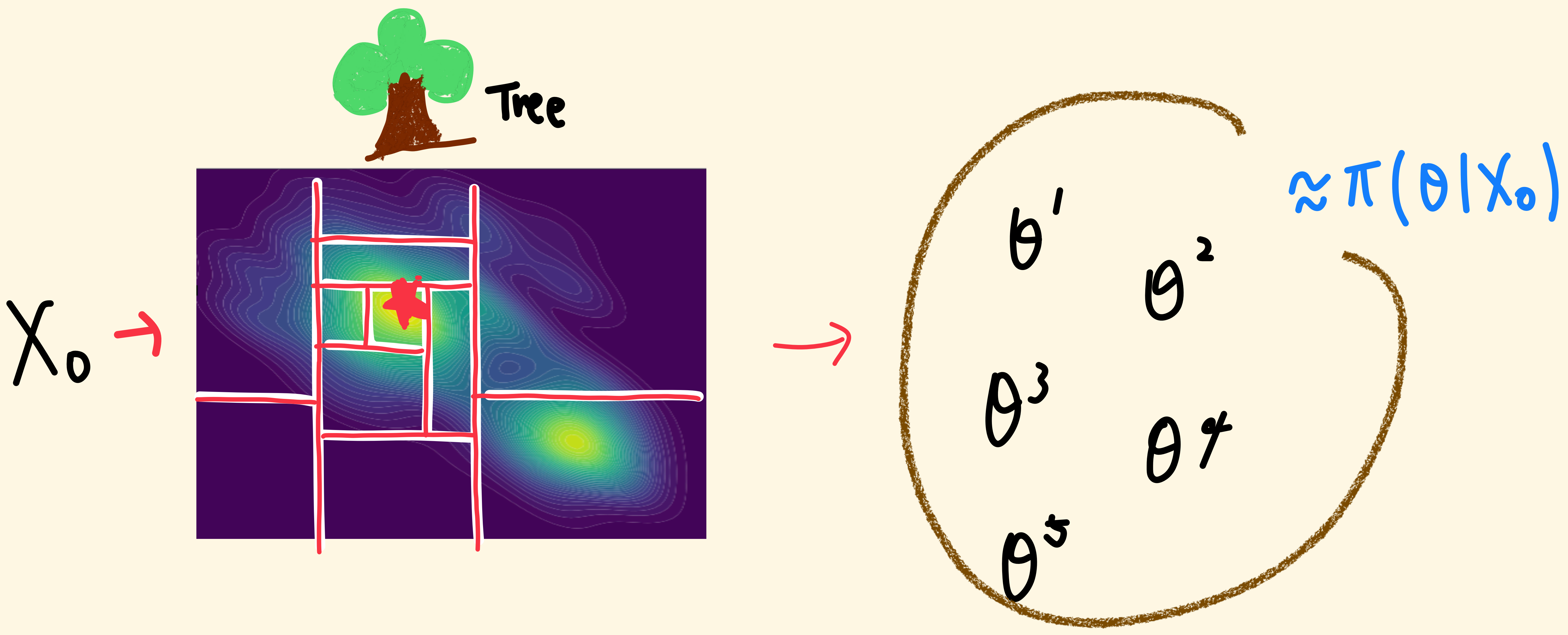
$$X = (X_{\text{obs}}, X_{\text{mask}})$$

Maximum size of trees



Concluding Remarks



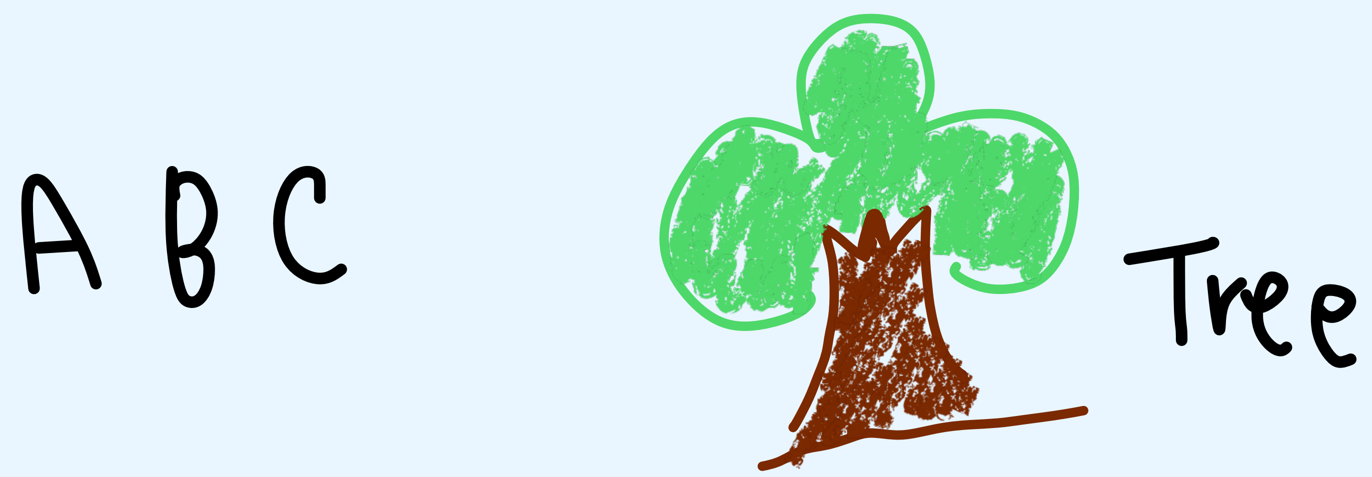


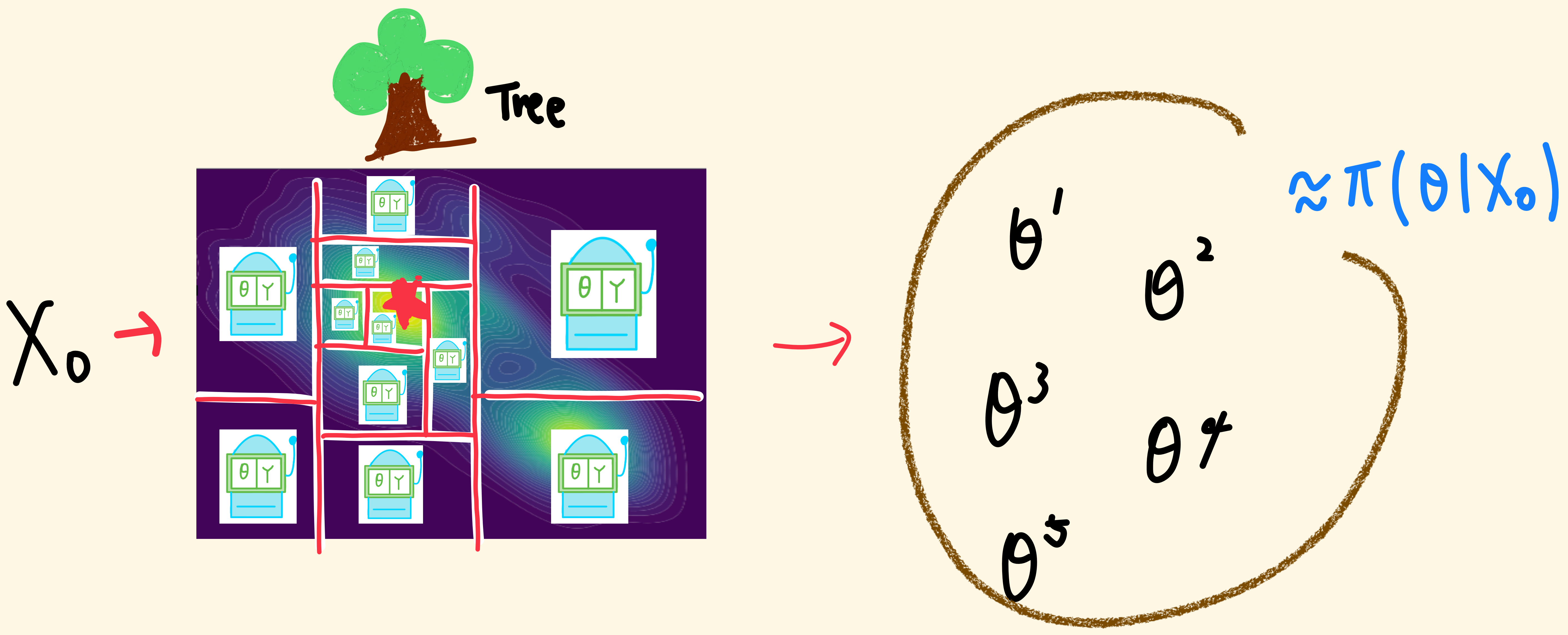
Our Contribution

By Combining Tree with Active ABC learning

1. Query efficient ABC

2. Likelihood Free MAP (Maximum a Posteriori)





Our Contribution

By Combining Tree with Active ABC learning

1. Query efficient ABC

2. Likelihood Free MAP (Maximum a Posteriori)

References

My Thompson algorithm explanation is inspired by and benefited from
a very helpful YouTube video by @aiinsights-riturajkaushik1618: https://www.youtube.com/watch?v=p701cYQeqew

Alsing, Justin, Benjamin D. Wandelt, and Stephen M. Feeney. "Optimal proposals for approximate Bayesian computation." *arXiv preprint arXiv:1808.06040* (2018).

Thank you!

And, thank you, Veronika and Sean-!

For more theoretical/numerical results:
Please check out https://arxiv.org/abs/2404.10436

Any comments on the slides and my presentation are appreciated-!!