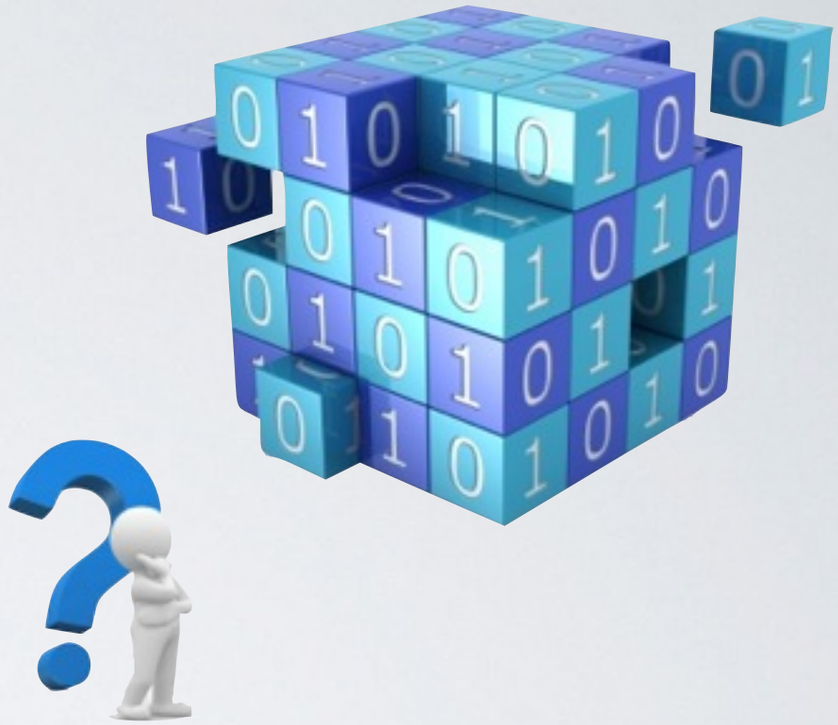
Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget

Anoop Korattikara Yutian Chen Max Welling

UCIrvine University of Cambridge Universiteit van Amsterdam

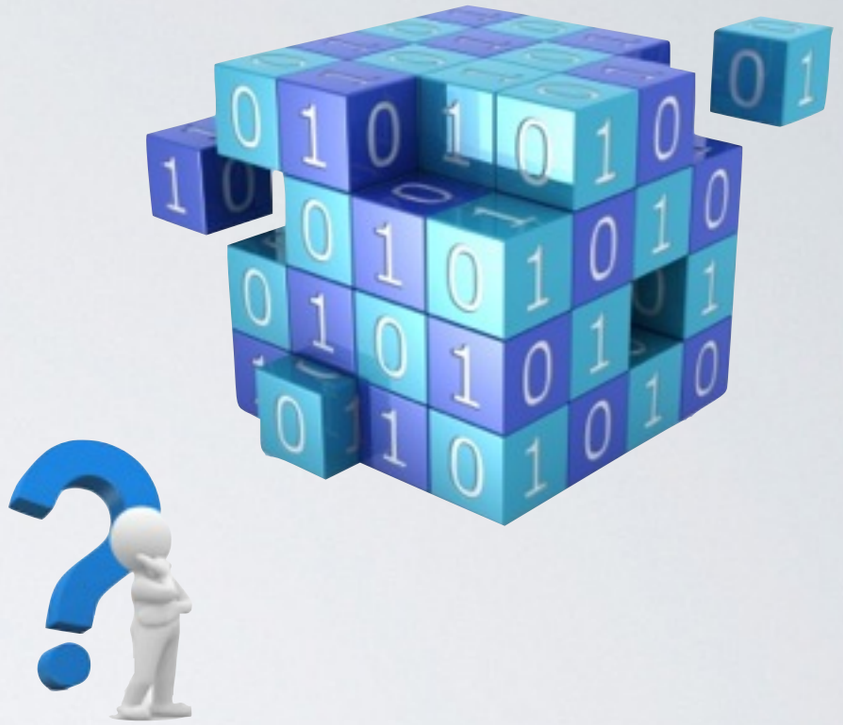
University of Sheffield, 6 March 2014

Big Data Challenge for Machine Learning



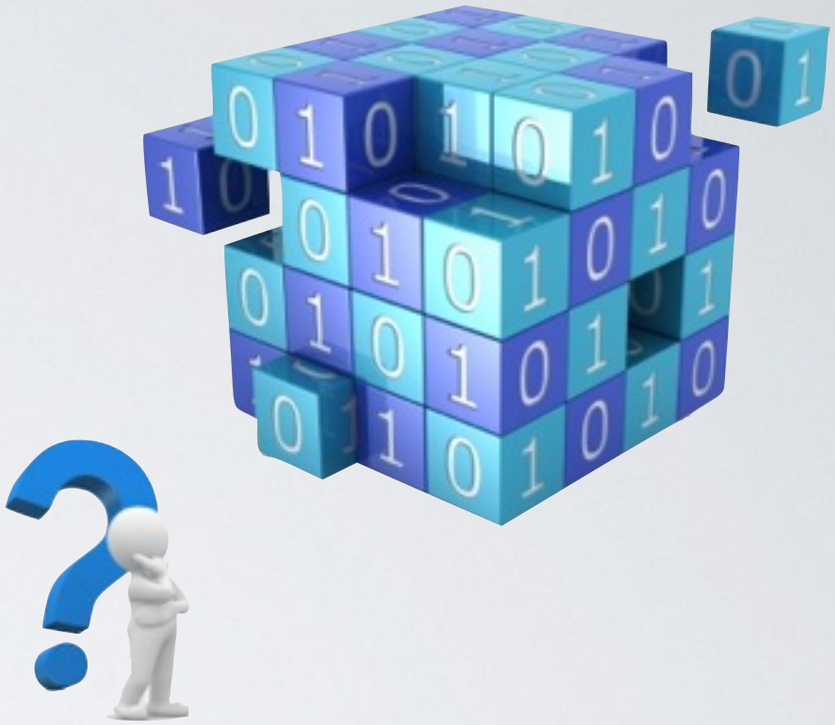
Big Data Challenge for Machine Learning

- Standard settings, moderate size of data
  - Generalization error, $\mathrm{E}(N)$
  - Computational efficiency, $\mathrm{T}(N)$

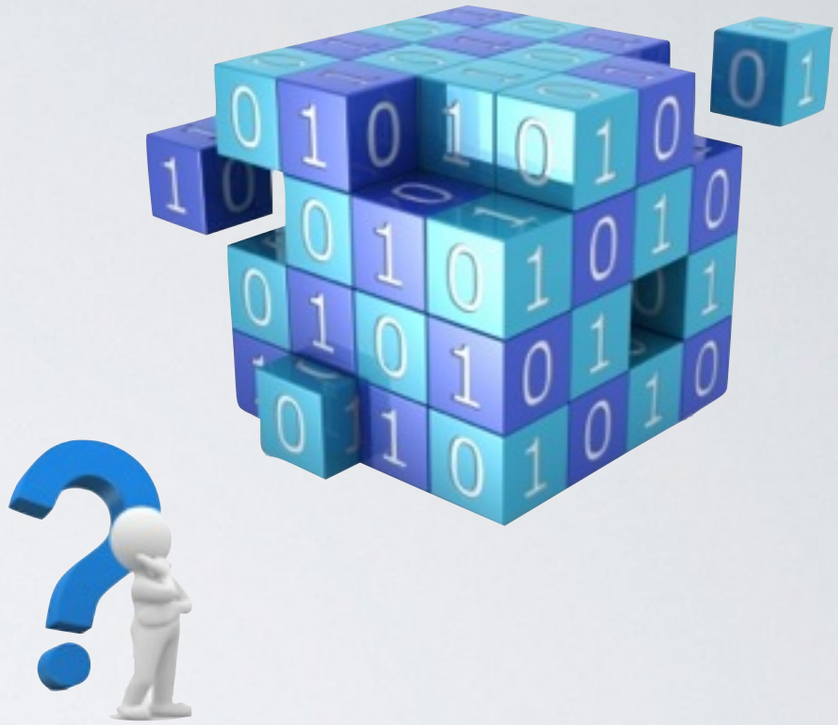


Big Data Challenge for Machine Learning

- Standard settings, moderate size of data
  - Generalization error, $\mathrm{E}(N)$
  - Computational efficiency, $\mathrm{T}(N)$
- Big data settings
  - Limited budget on time
  - $\mathrm{E}(N, T)$ or $\mathrm{E}(T)$ (Bottou & Bousquet, 2007)



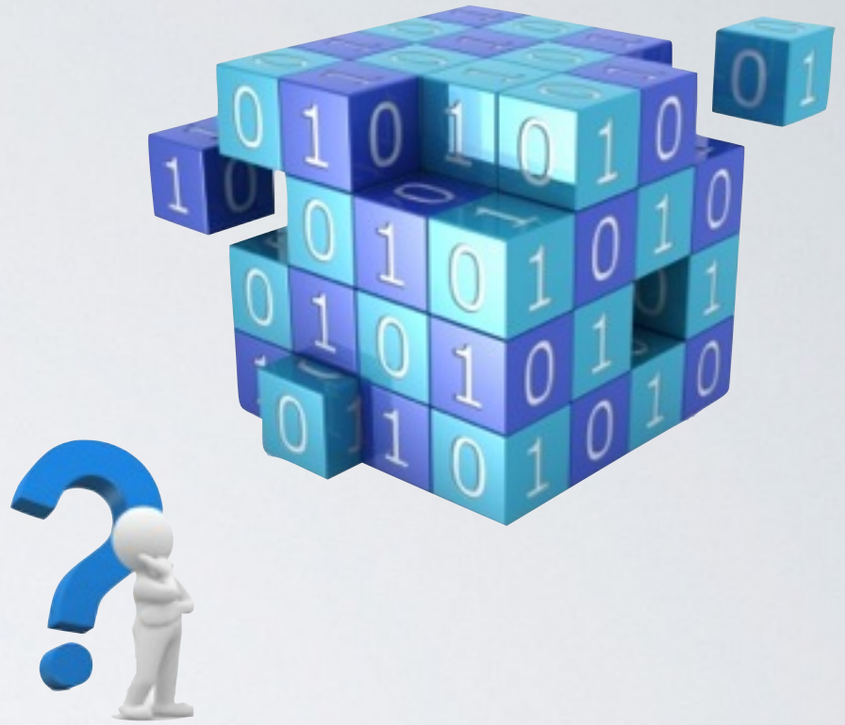
Big Data Challenge for Machine Learning



Big Data Challenge for Machine Learning

- Learning with optimization

$$\theta^* = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} l(x_i; \theta)$$

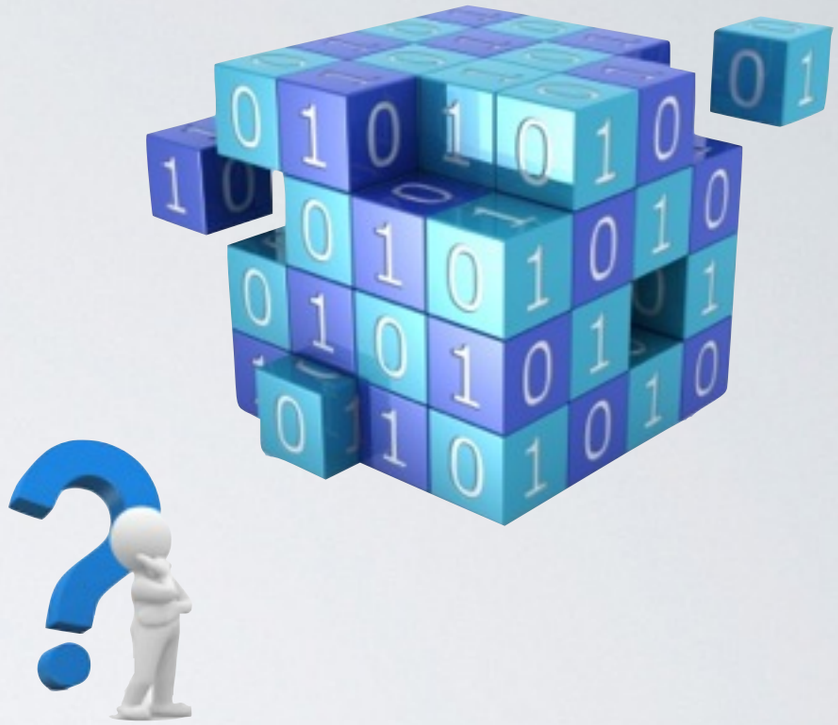


Big Data Challenge for Machine Learning

- Learning with optimization

$$\theta^* = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} l(x_i;\theta)$$

O(N)



Big Data Challenge for Machine Learning

- Learning with optimization

$$\theta^* = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} l(x_i;\theta)$$ O(N)

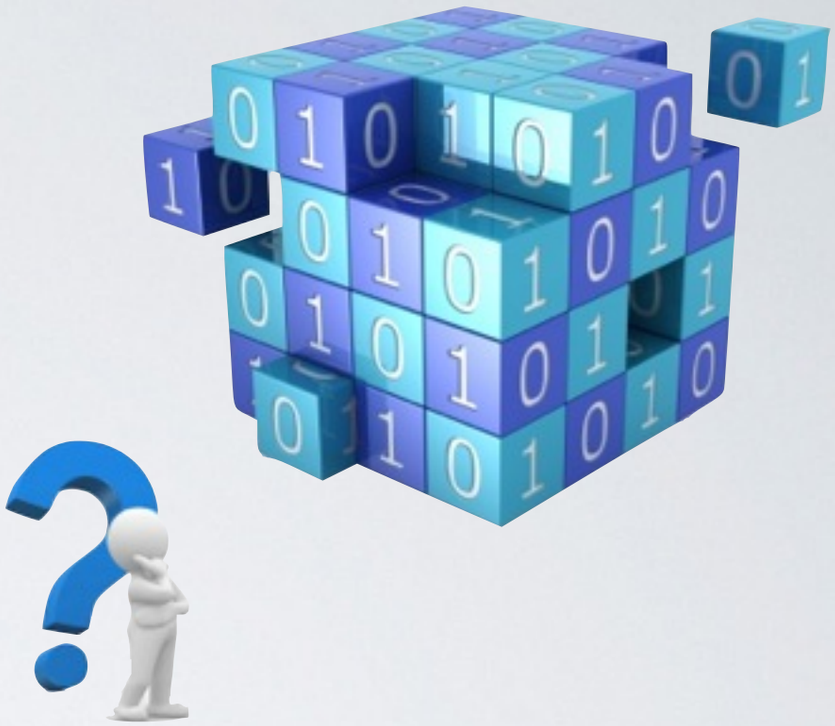
  - Stochastic gradient descent O(n)

Big Data Challenge for Machine Learning

- Learning with optimization

$$\theta^* = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} l(x_i;\theta)$$ O(N)

  - Stochastic gradient descent O(n)



- Bayesian inference

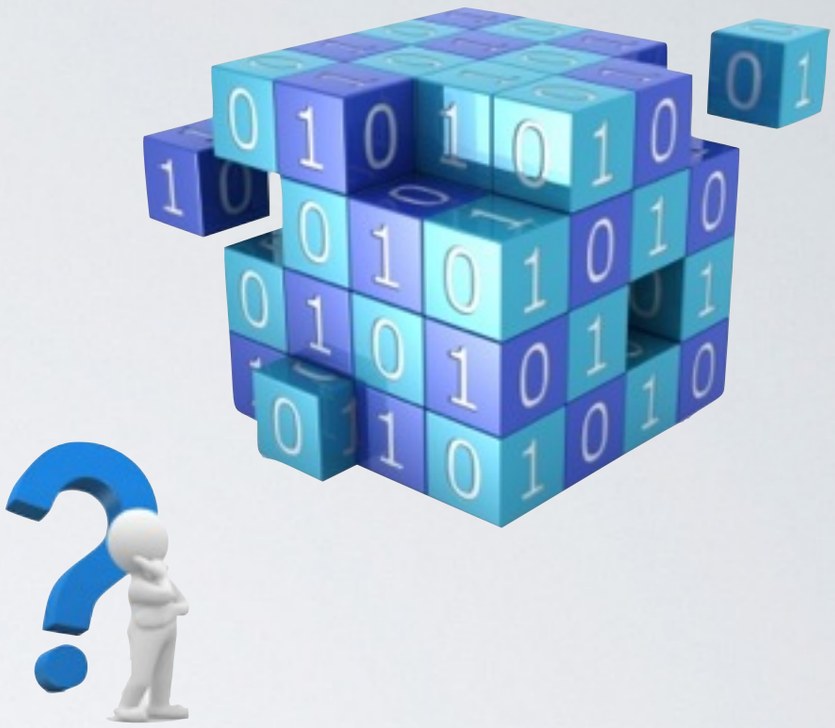
$$P(\theta \mid \mathcal{D}) \propto P(\theta)\prod_{i=1}^{N} P(x_i \mid \theta)$$

Big Data Challenge for Machine Learning

- Learning with optimization

$$\theta^* = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} l(x_i;\theta)$$ O(N)

  - Stochastic gradient descent O(n)



- Bayesian inference

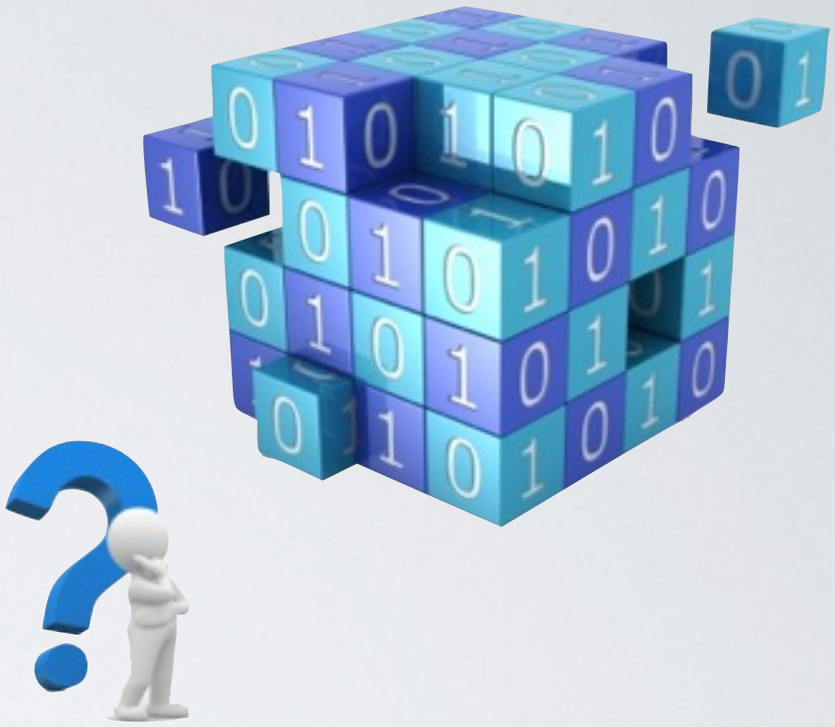
$$P(\theta \mid \mathcal{D}) \propto P(\theta)\prod_{i=1}^{N} P(x_i \mid \theta)$$ O(N)

Big Data Challenge for Machine Learning

- Learning with optimization

$$\theta^* = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} l(x_i;\theta)$$ O(N)

  - Stochastic gradient descent O(n)



- Bayesian inference

$$P(\theta \mid \mathcal{D}) \propto P(\theta)\prod_{i=1}^{N} P(x_i \mid \theta)$$ O(N)

  - Stochastic Variational Inference (Hoffman, et al, 2013) O(n)

Big Data Challenge for Machine Learning

- Learning with optimization

$$\theta^* = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} l(x_i;\theta)$$ O(N)

  - Stochastic gradient descent O(n)

- Bayesian inference

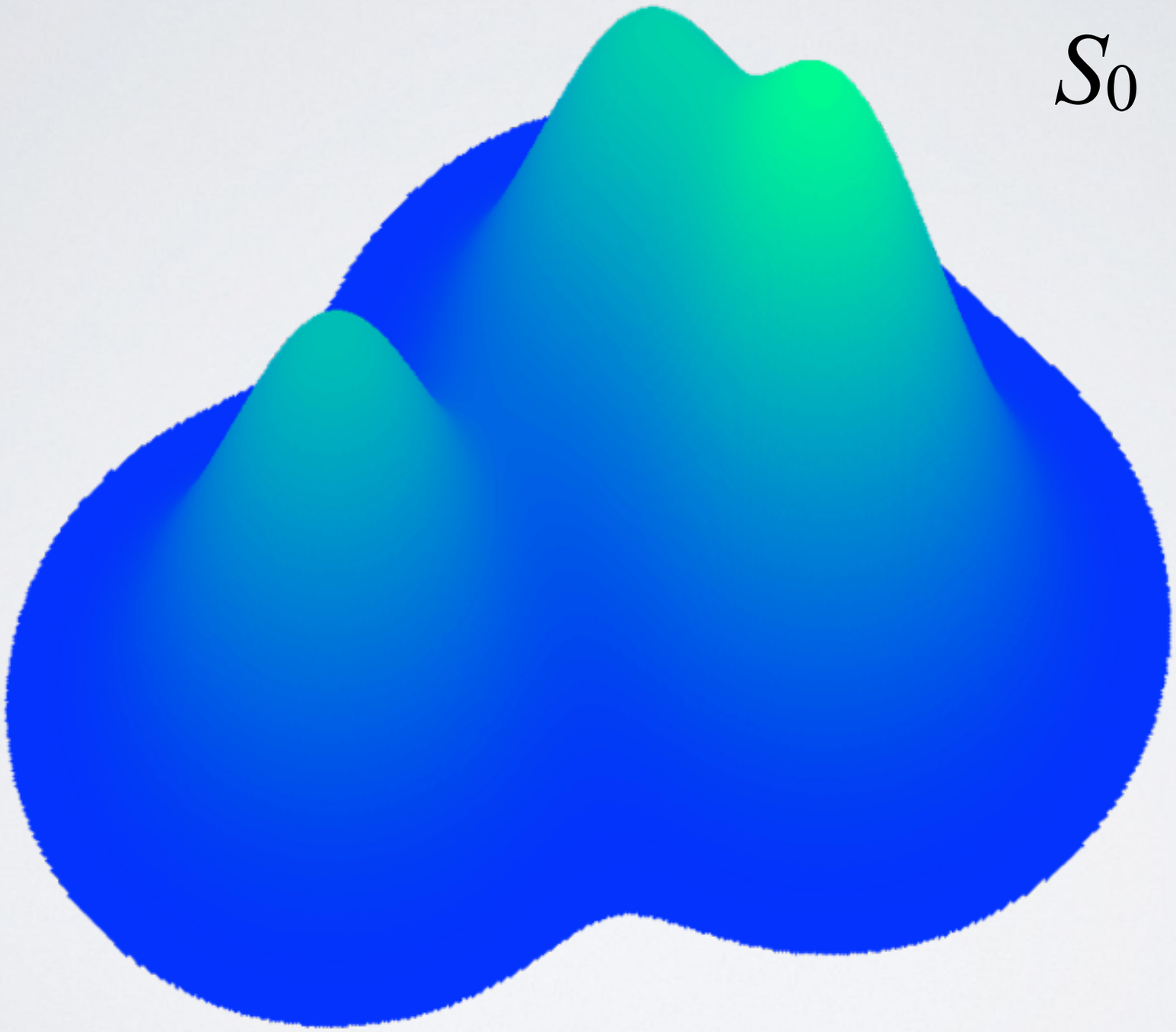
$$P(\theta \mid \mathcal{D}) \propto P(\theta)\prod_{i=1}^{N} P(x_i \mid \theta)$$ O(N)

  - Stochastic Variational Inference (Hoffman, et al, 2013) O(n)

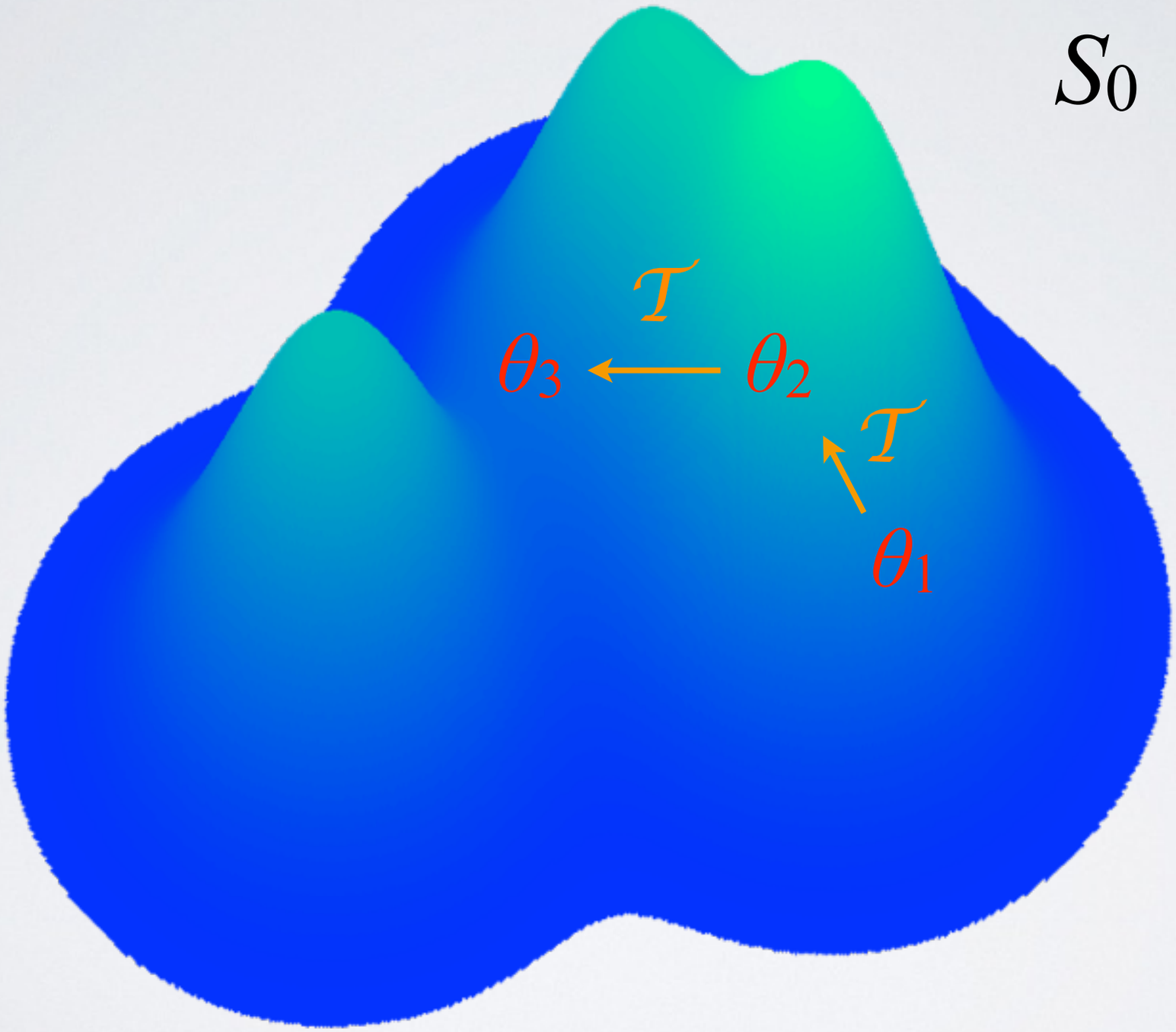
  - MCMC

MCMC Review

UNIVERSITY OF CAMBRIDGE

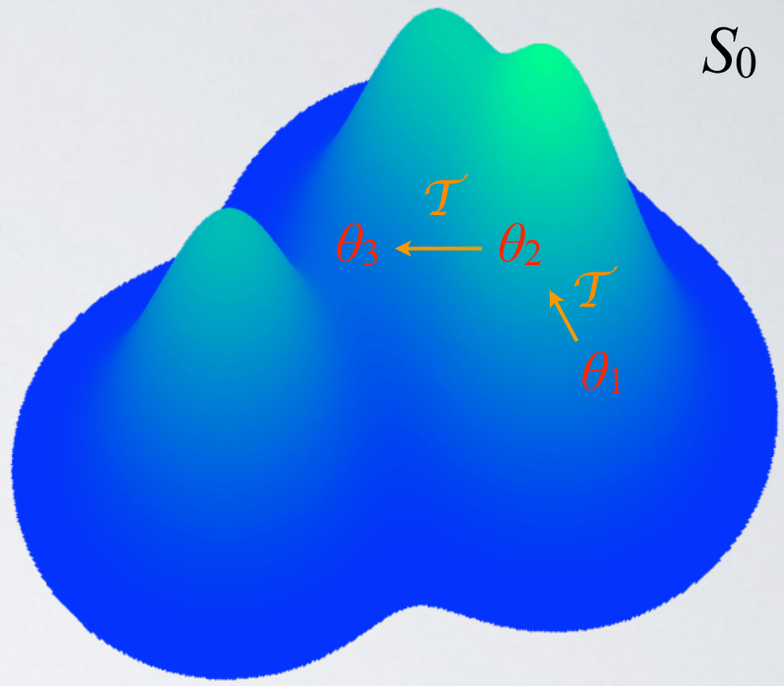


MCMC Review



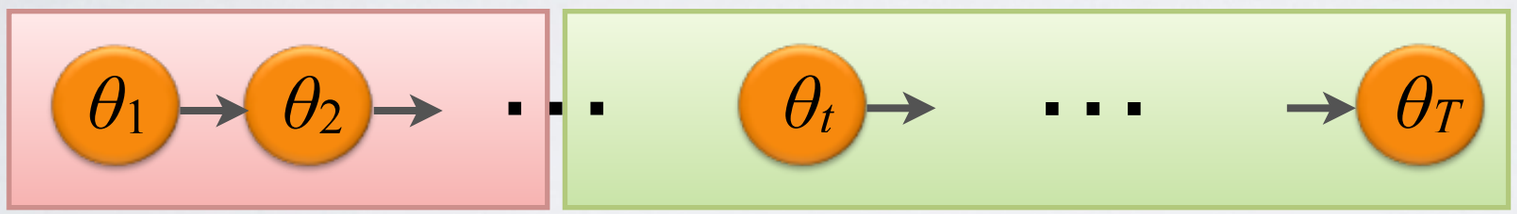
MCMC Review

Choose $\mathcal{T}$ s.t. the marginals $P(\theta_t) \rightarrow S_0$ as $t \rightarrow \infty$

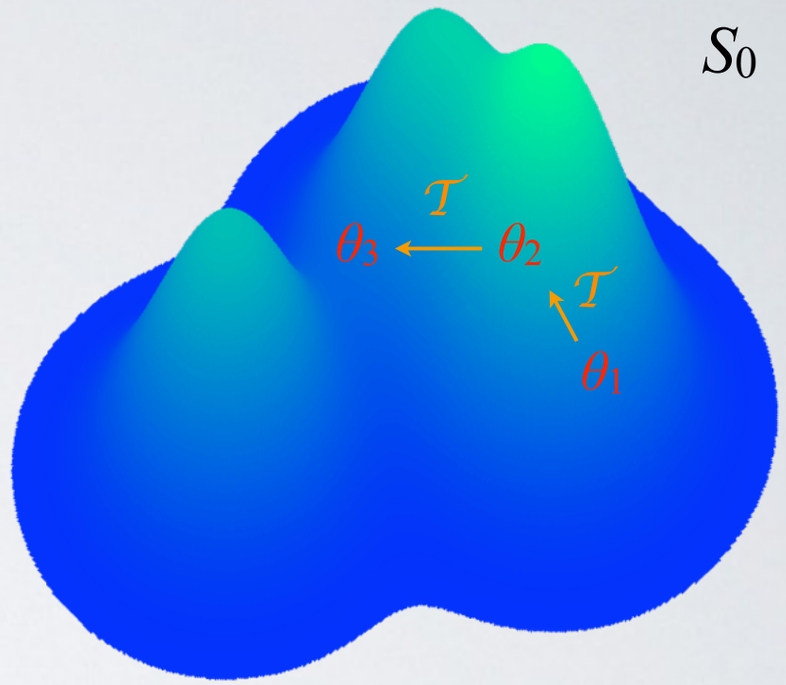


MCMC Review

Choose $\mathcal{T}$ s.t. the marginals $P(\theta_t) \rightarrow S_0$ as $t \rightarrow \infty$

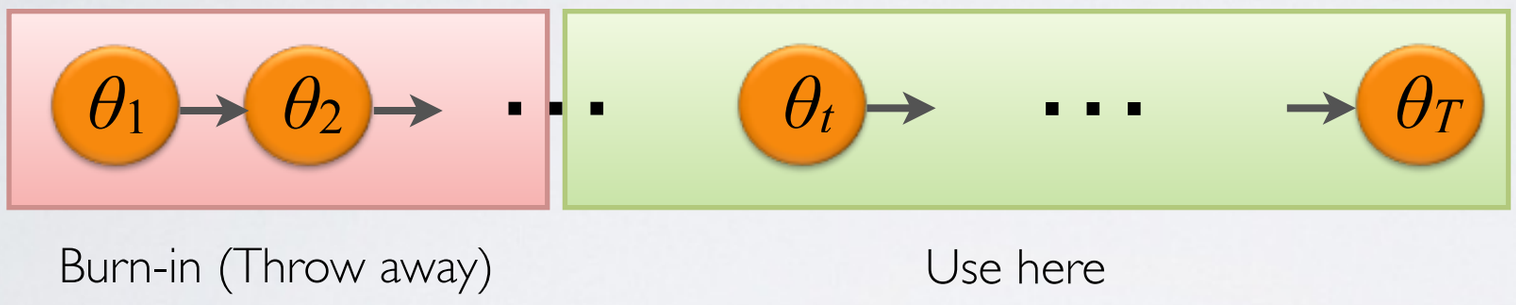


Burn-in (Throw away)



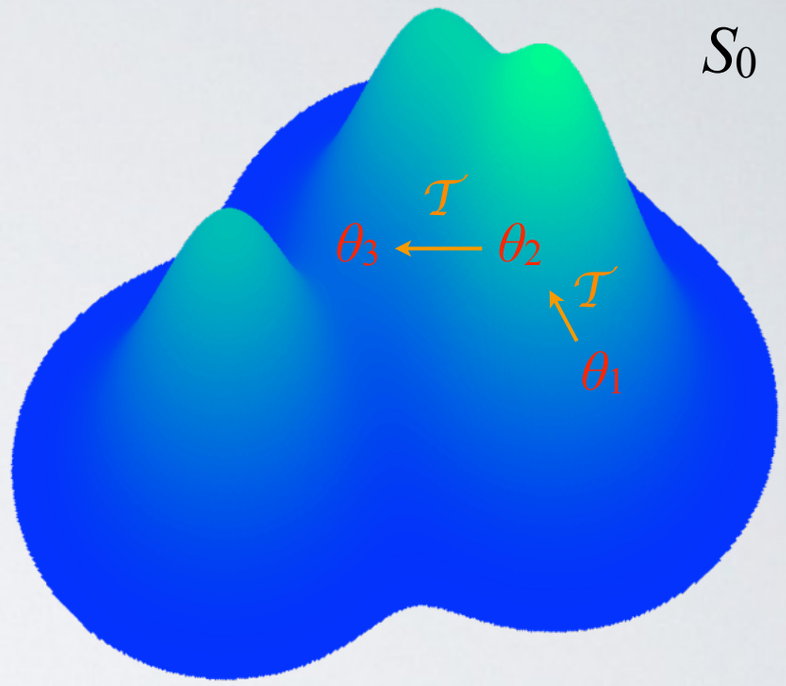
MCMC Review

Choose $\mathcal{T}$ s.t. the marginals $P(\theta_t) \rightarrow S_0$ as $t \rightarrow \infty$



Burn-in (Throw away)

Use here



$$I = \langle f \rangle_{S_0} \approx \hat{I} = \overline{f(\theta_t)}, \quad \text{where } \theta_t \sim S_0$$

Metropolis-Hastings Sampling

UNIVERSITY OF CAMBRIDGE

θ_t →

Propose

$$\theta' \sim q(\theta' \mid \theta_t)$$

Accept/Reject

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$

$$\theta_{t+1} \leftarrow \begin{cases} \theta' & \text{with probability } P_a \\ \theta_t & \text{with probability } 1\text{-}P_a \end{cases}$$

→ θ_{t+1}

Metropolis-Hastings Sampling

UNIVERSITY OF CAMBRIDGE

θ_t →

Propose

$$\theta' \sim q(\theta' \mid \theta_t)$$

Accept/Reject

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$

$$\theta_{t+1} \leftarrow \begin{cases} \theta' & \text{with probability } P_a \\ \theta_t & \text{with probability } 1\text{-}P_a \end{cases}$$

→ θ_{t+1}

Detailed Balance

Metropolis-Hastings Sampling

UNIVERSITY OF CAMBRIDGE

θ_t →

Propose

$$\theta' \sim q(\theta' \mid \theta_t)$$

Accept/Reject

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$

$$\theta_{t+1} \leftarrow \begin{cases} \theta' & \text{with probability } P_a \\ \theta_t & \text{with probability } 1\text{-}P_a \end{cases}$$

→ θ_{t+1}

Detailed Balance ⇒ $P(\theta_t) \rightarrow S_0$

Metropolis-Hastings Sampling

UNIVERSITY OF CAMBRIDGE

θ_t →

Propose

$$\theta' \sim q(\theta' \mid \theta_t)$$

Accept/Reject

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$

$$\theta_{t+1} \leftarrow \begin{cases} \theta' & \text{with probability } P_a \\ \theta_t & \text{with probability } 1\text{-}P_a \end{cases}$$

→ θ_{t+1}

Detailed Balance ⇒ $P(\theta_t) \rightarrow S_0$ ⇒ $\mathbb{E}\left[\hat{I}\right] = I$

Metropolis-Hastings Sampling

UNIVERSITY OF CAMBRIDGE

θ_t →

Propose

$$\theta' \sim q(\theta' \mid \theta_t)$$

Accept/Reject

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$

$$\theta_{t+1} \leftarrow \begin{cases} \theta' & \text{with probability } P_a \\ \theta_t & \text{with probability } 1\text{-}P_a \end{cases}$$

→ θ_{t+1}

Detailed Balance → $P(\theta_t) \rightarrow S_0$ → $\mathbb{E}\left[\hat{I}\right] = I$

$$S_0(\theta) = P(\theta \mid \mathcal{D}) \propto P(\theta)\prod_{i=1}^{N} P(x_i \mid \theta)$$

Metropolis-Hastings Sampling

UNIVERSITY OF CAMBRIDGE

θ_t →

Propose

$$\theta' \sim q(\theta' \mid \theta_t)$$

Accept/Reject

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$

$$\theta_{t+1} \leftarrow \begin{cases} \theta' & \text{with probability } P_a \\ \theta_t & \text{with probability } 1\text{-}P_a \end{cases}$$

→ θ_{t+1}

Detailed Balance → $P(\theta_t) \rightarrow S_0$ → $\mathbb{E}\left[\hat{I}\right] = I$

$$S_0(\theta) = P(\theta \mid \mathcal{D}) \propto P(\theta)\prod_{i=1}^{N} P(x_i \mid \theta)$$

O(N)

Risk Decomposition

Risk	=	Bias	^{2}	+	Variance
$\mathbb{E}\left[\left(I-\hat{I}\right)^2\right]$		$I-\langle f\rangle_S$			$\sigma^2\frac{\tau}{T}$

Risk Decomposition

Risk	=	Bias^2	+	Variance
$\mathbb{E}\left[\left(I-\hat{I}\right)^2\right]$		$I-\langle f\rangle_S$		$\sigma^2\frac{\tau}{T}$

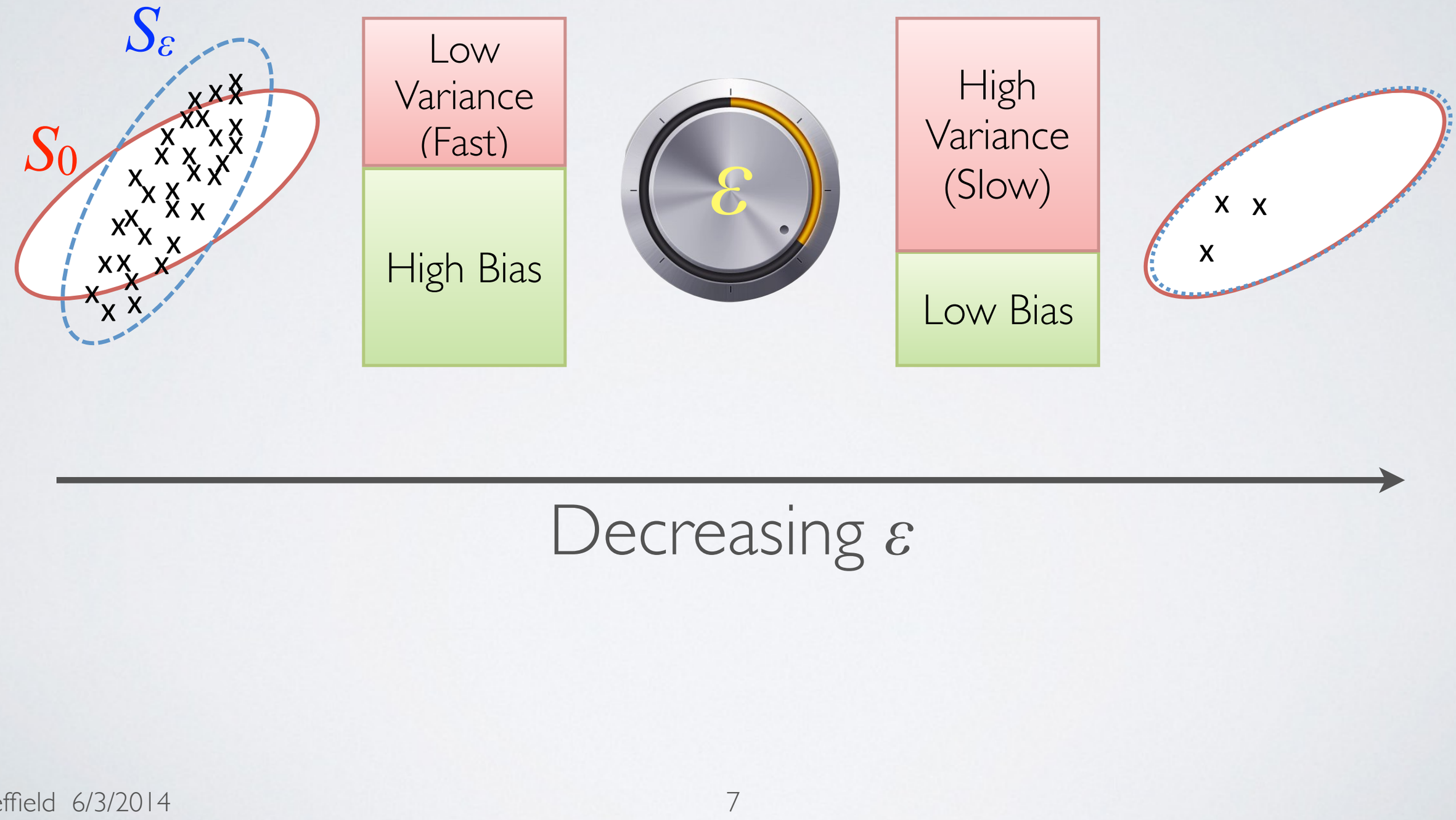
$S = S_0$ ⇒ $\langle f\rangle_S = I$

Risk Decomposition

Risk	=	Bias^2	+	Variance
$\mathbb{E}\left[\left(I-\hat{I}\right)^2\right]$		$I-\langle f\rangle_S$		$\sigma^2\frac{\tau}{T}$

$S = \int_\varepsilon$ ⇒ $\langle f\rangle_S \neq I$

Fixed Computational Budget



Efficient Approximate MH

UNIVERSITY OF CAMBRIDGE

θ_t →

Propose

$$\theta' \sim q(\theta' \mid \theta_t)$$

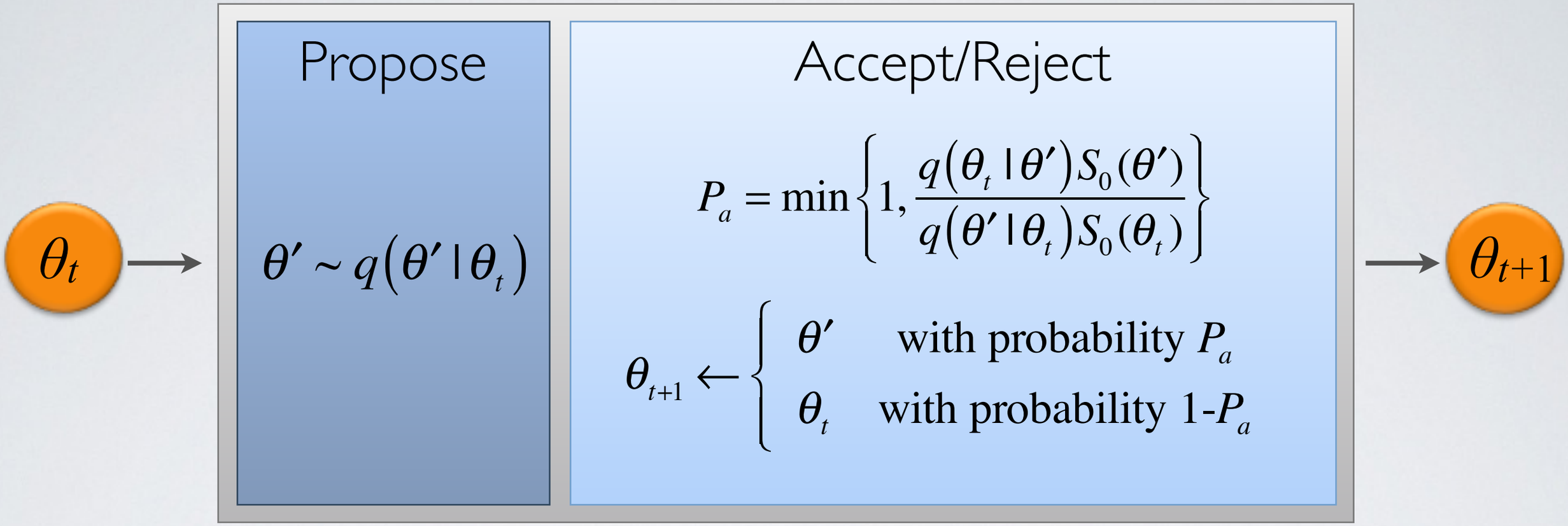
Accept/Reject

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$

$$\theta_{t+1} \leftarrow \begin{cases} \theta' & \text{with probability } P_a \\ \theta_t & \text{with probability } 1\text{-}P_a \end{cases}$$

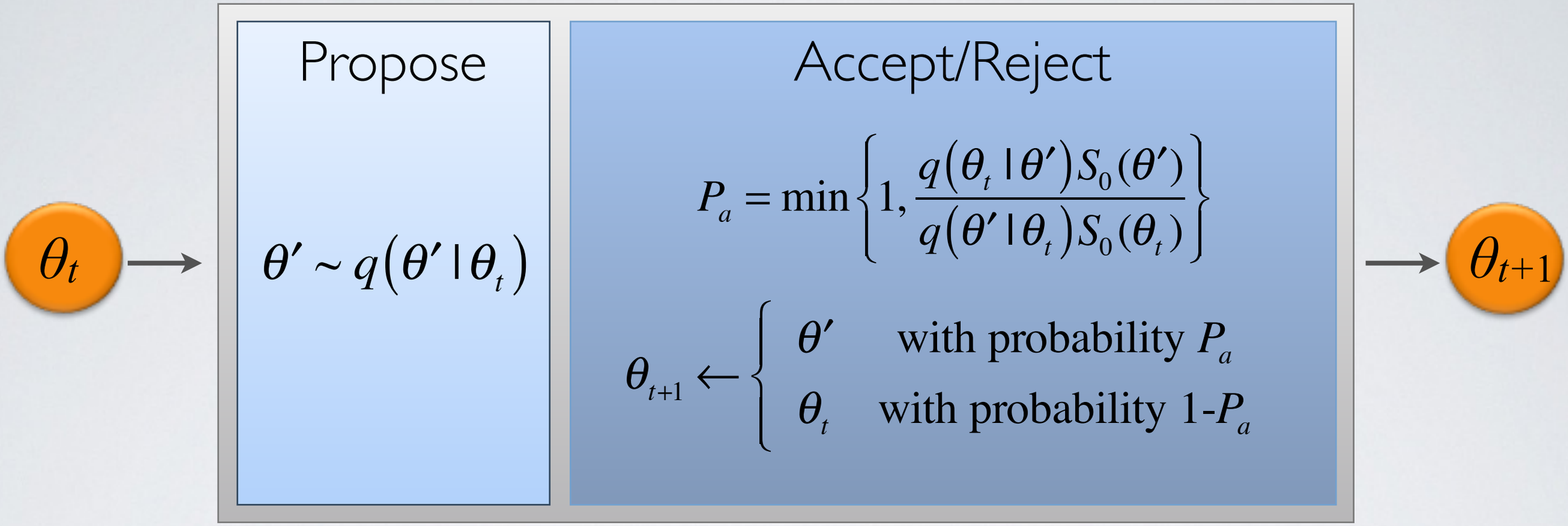
→ θ_{t+1}

Efficient Approximate MH



- Efficient Proposal
  - Stochastic Gradient Langevin Dynamics
    (Welling&Teh, 2011; Ahn et.al. 2012)

Efficient Approximate MH



- Efficient Accept/Reject?

Accept/Reject as a Hypothesis Testing

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$

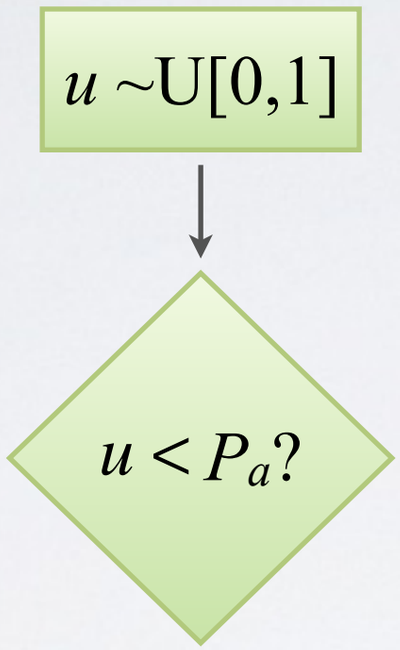
Accept/Reject as a Hypothesis Testing

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$

$u \sim \mathrm{U}[0,1]$

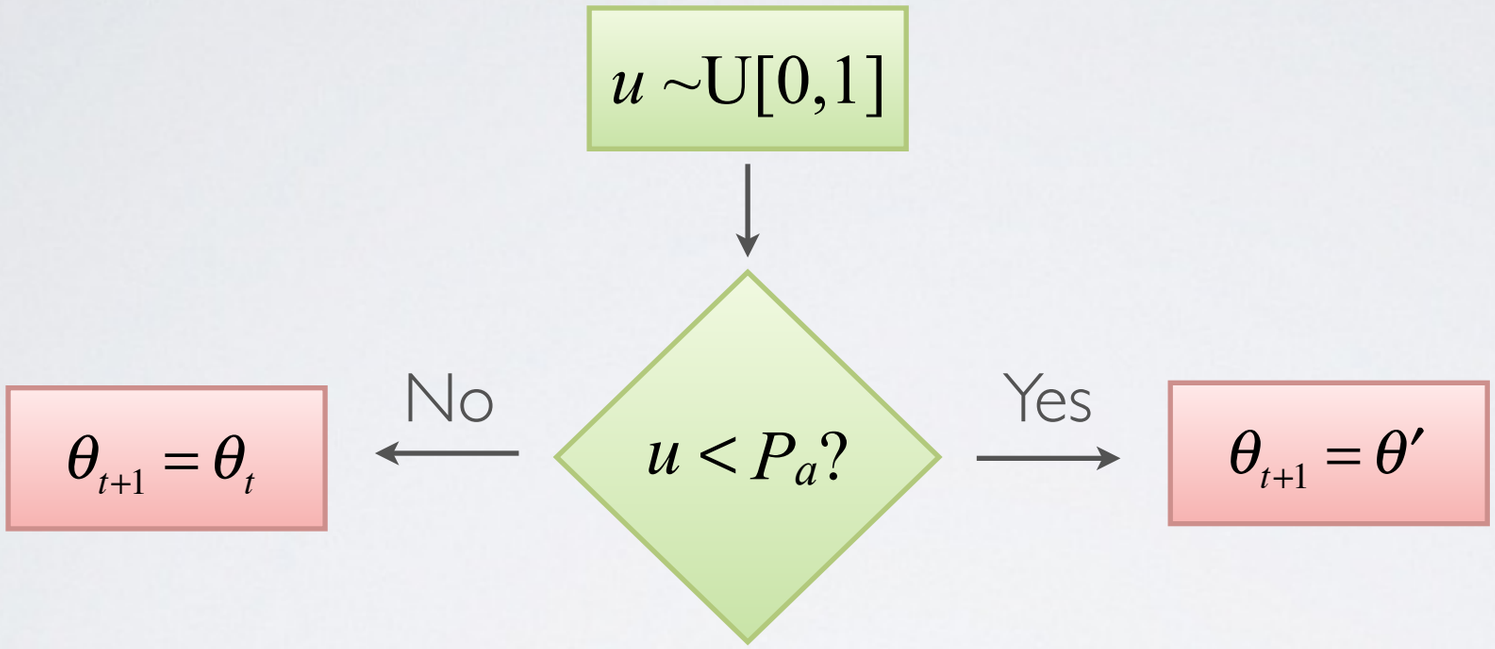
Accept/Reject as a Hypothesis Testing

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$



Accept/Reject as a Hypothesis Testing

$$P_a = \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$



Accept/Reject as a Hypothesis Testing

$$u < \min\left\{1, \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}\right\}$$

Accept/Reject as a Hypothesis Testing

$$u < \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}$$

Accept/Reject as a Hypothesis Testing

$$u < \frac{q(\theta_t \mid \theta') S_0(\theta')}{q(\theta' \mid \theta_t) S_0(\theta_t)}$$

$$u < \frac{q(\theta_t \mid \theta') P(\theta')}{q(\theta' \mid \theta_t) P(\theta_t)} \cdot \prod_{i=1}^{N} \frac{P(x_i \mid \theta')}{P(x_i \mid \theta_t)}$$

Accept/Reject as a Hypothesis Testing

$$u < \frac{q(\theta_t \mid \theta') S_0(\theta')}{q(\theta' \mid \theta_t) S_0(\theta_t)}$$

$$u < \frac{q(\theta_t \mid \theta') P(\theta')}{q(\theta' \mid \theta_t) P(\theta_t)} \cdot \prod_{i=1}^{N} \frac{P(x_i \mid \theta')}{P(x_i \mid \theta_t)}$$

$$\log(u) < \log\left(\frac{q(\theta_t \mid \theta') P(\theta')}{q(\theta' \mid \theta_t) P(\theta_t)}\right) + \sum_{i=1}^{N} \log\left(\frac{P(x_i \mid \theta')}{P(x_i \mid \theta_t)}\right)$$

Accept/Reject as a Hypothesis Testing

$$u < \frac{q(\theta_t \mid \theta')S_0(\theta')}{q(\theta' \mid \theta_t)S_0(\theta_t)}$$

$$u < \frac{q(\theta_t \mid \theta')P(\theta')}{q(\theta' \mid \theta_t)P(\theta_t)} \cdot \prod_{i=1}^{N} \frac{P(x_i \mid \theta')}{P(x_i \mid \theta_t)}$$

$$\log(u) < \log\left(\frac{q(\theta_t \mid \theta')P(\theta')}{q(\theta' \mid \theta_t)P(\theta_t)}\right) + \sum_{i=1}^{N} \log\left(\frac{P(x_i \mid \theta')}{P(x_i \mid \theta_t)}\right)$$

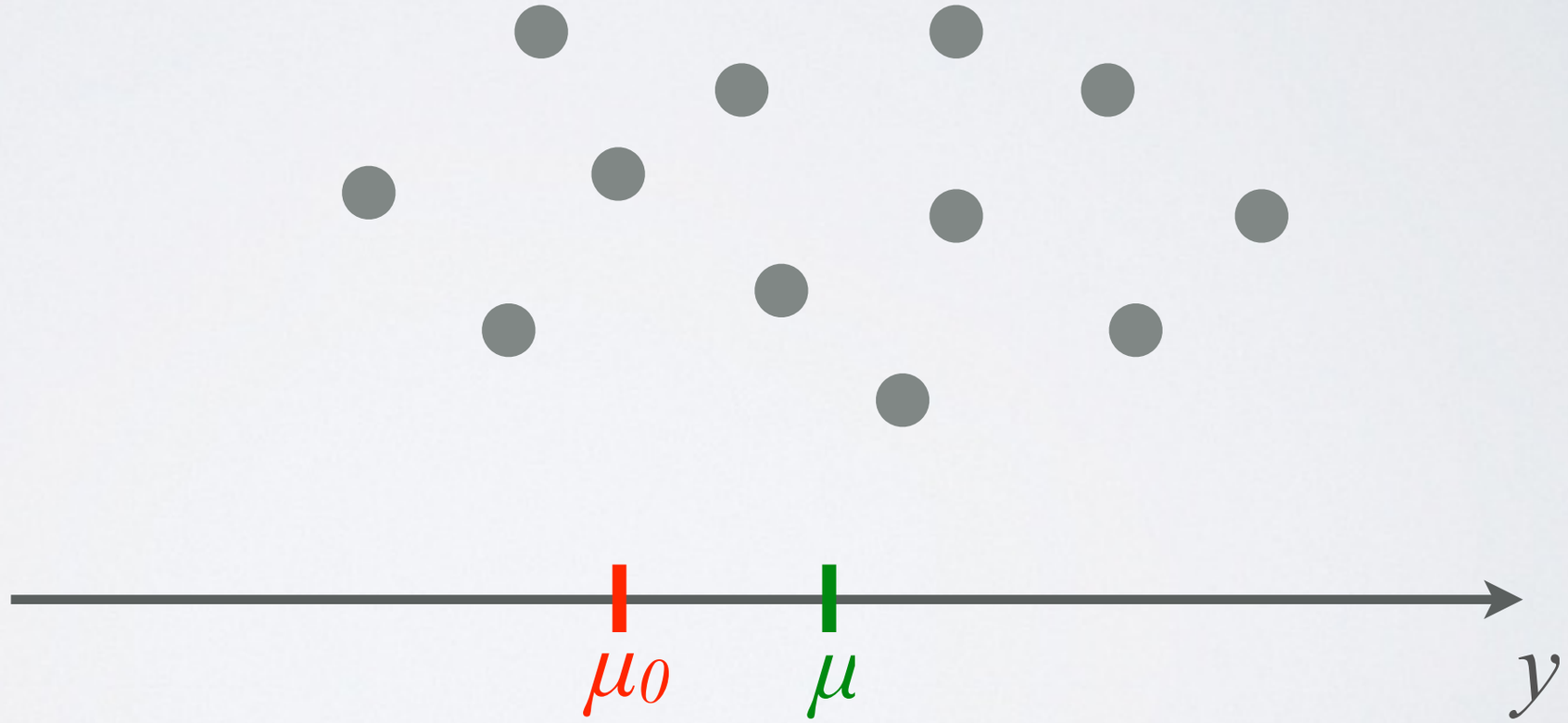
$$\frac{1}{N}\sum_{i=1}^{N} y_i > \mu_0 \,?$$

Accept/Reject as a Hypothesis Testing

- Given a set $\mathcal{Y}= \{y_i\}_{i=1}^{N}$, with a mean μ and a constant μ_0, test:
  - H_0: $\mu>\mu_0$
  - H_a: $\mu<\mu_0$

$$\frac{1}{N}\sum_{i=1}^{N} y_i > \mu_0 \, ?$$

$$\log\left(\frac{P(x_i|\theta')}{P(x_i|\theta_t)}\right) \qquad \frac{1}{N}\log\left(u\frac{q(\theta'|\theta_t)P(\theta_t)}{q(\theta_t|\theta')P(\theta')}\right)$$



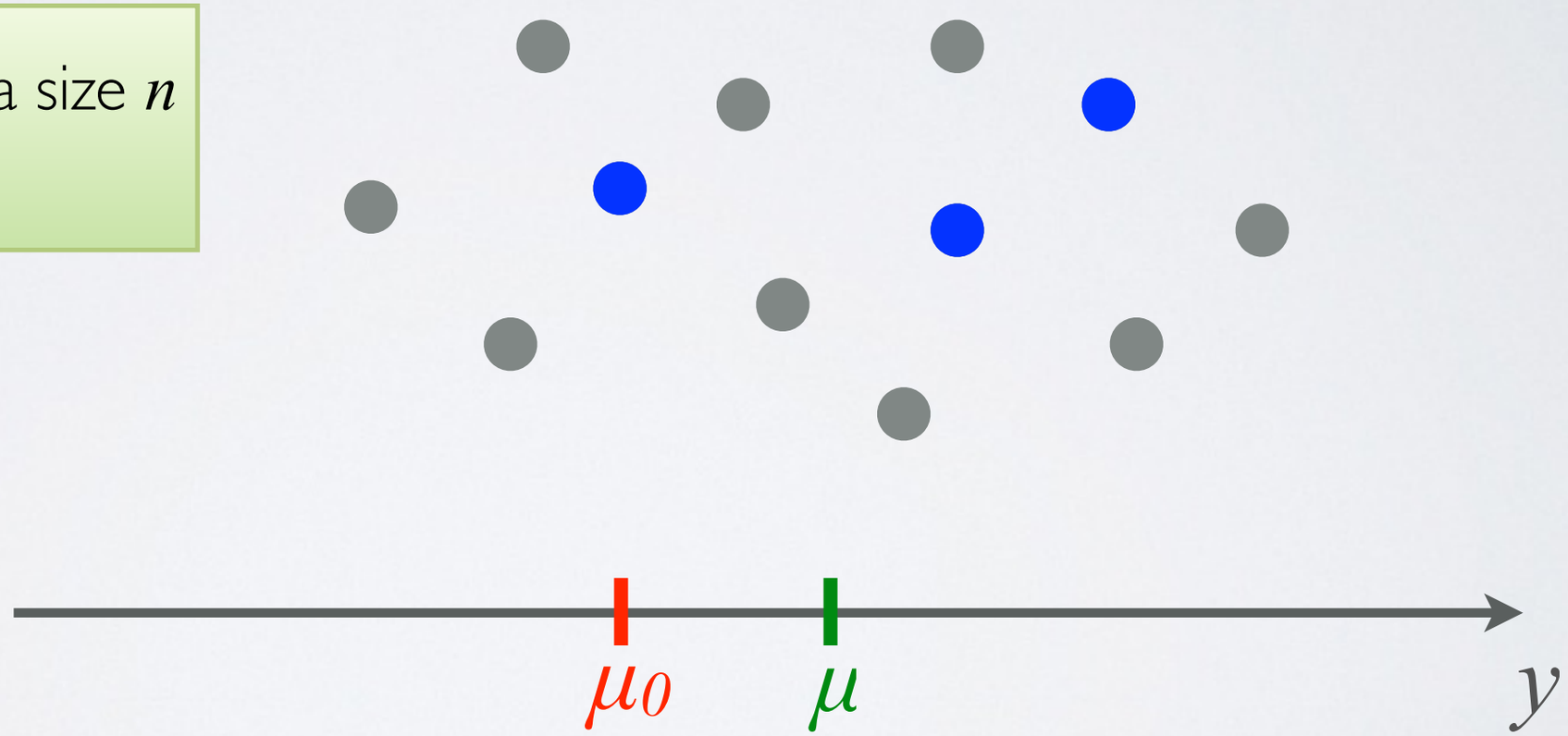
Accept/Reject as a Hypothesis Testing

- Given a set $\mathcal{Y} = \{y_i\}_{i=1}^{N}$, with a mean μ and a constant μ_0, test:
  - H_0: $\mu > \mu_0$
  - H_a: $\mu < \mu_0$

$$\frac{1}{N}\sum_{i=1}^{N} y_i > \mu_0 \, ?$$

$$\log\left(\frac{P(x_i|\theta')}{P(x_i|\theta_t)}\right) \qquad \frac{1}{N}\log\left(u\frac{q(\theta'|\theta_t)P(\theta_t)}{q(\theta_t|\theta')P(\theta')}\right)$$

1. Draw $Y \subseteq \mathcal{Y}$ of a size n



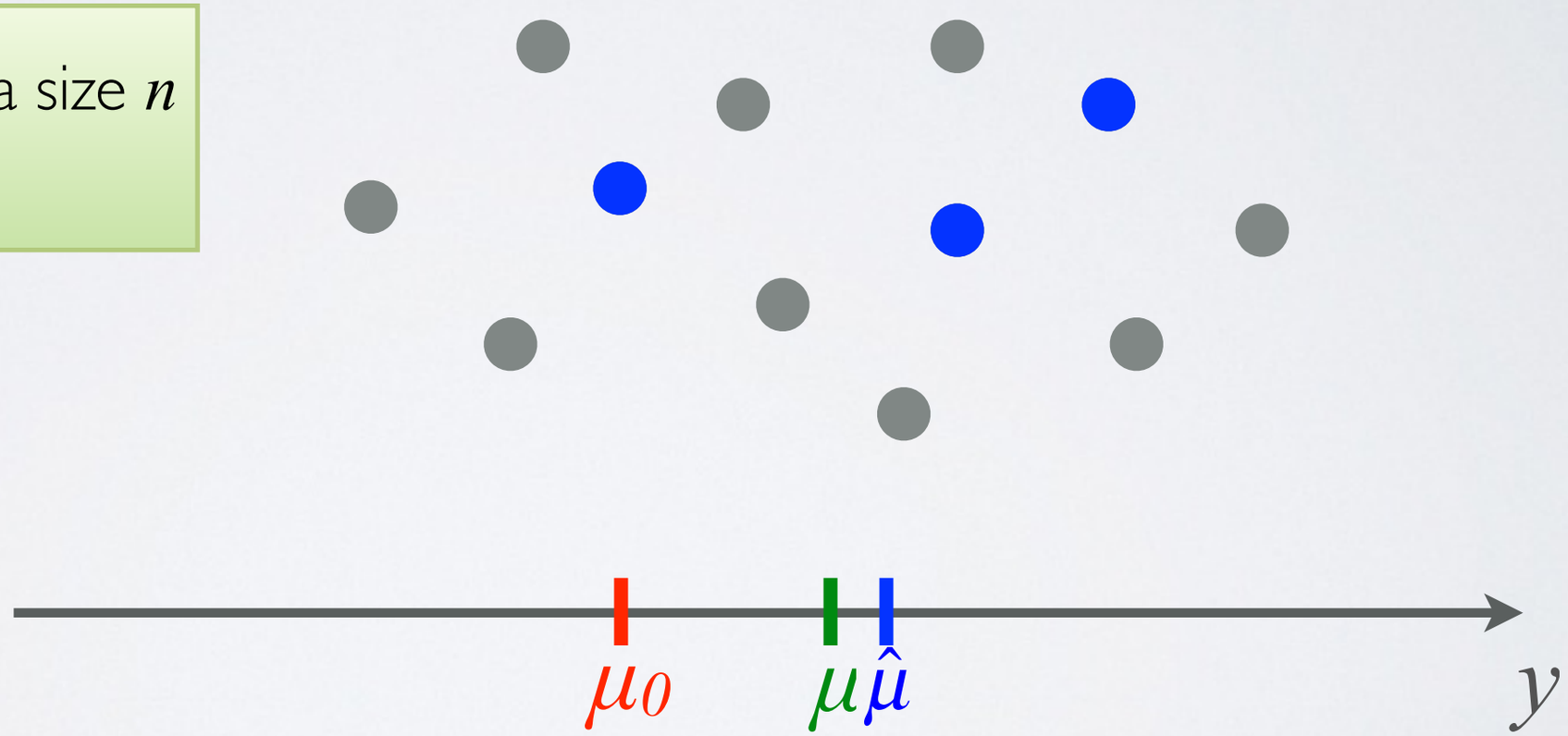
Accept/Reject as a Hypothesis Testing

- Given a set $\mathcal{Y} = \{y_i\}_{i=1}^{N}$, with a mean μ and a constant μ_0, test:
  - H_0: $\mu > \mu_0$
  - H_a: $\mu < \mu_0$

$$\frac{1}{N}\sum_{i=1}^{N} y_i > \mu_0 \, ?$$

$$\log\left(\frac{P(x_i|\theta')}{P(x_i|\theta_t)}\right) \qquad \frac{1}{N}\log\left(u\frac{q(\theta'|\theta_t)P(\theta_t)}{q(\theta_t|\theta')P(\theta')}\right)$$

1. Draw $Y \subseteq \mathcal{Y}$ of a size n
2. Check $\hat{\mu} > \mu_0$?



Accept/Reject as a Hypothesis Testing

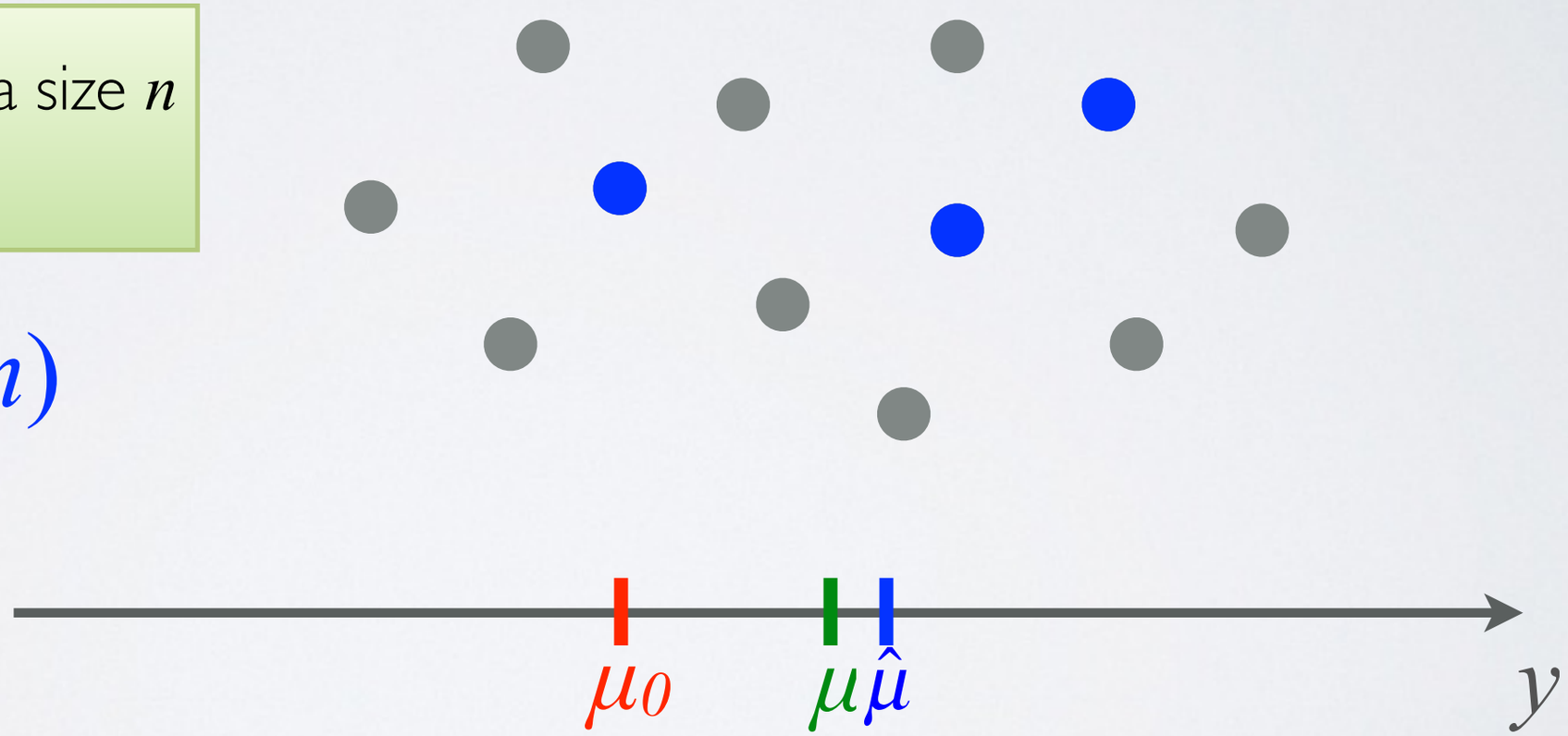
- Given a set $\mathcal{Y}=\{y_i\}_{i=1}^{N}$, with a mean μ and a constant μ_0, test:
  - H_0: $\mu>\mu_0$
  - H_a: $\mu<\mu_0$

$$\frac{1}{N}\sum_{i=1}^{N} y_i > \mu_0 \,?$$

$$\log\left(\frac{P(x_i|\theta')}{P(x_i|\theta_t)}\right) \qquad \frac{1}{N}\log\left(u\frac{q(\theta'|\theta_t)P(\theta_t)}{q(\theta_t|\theta')P(\theta')}\right)$$

1. Draw $Y \subseteq \mathcal{Y}$ of a size n
2. Check $\hat{\mu} > \mu_0$?

$$O(N) \rightarrow O(n)$$



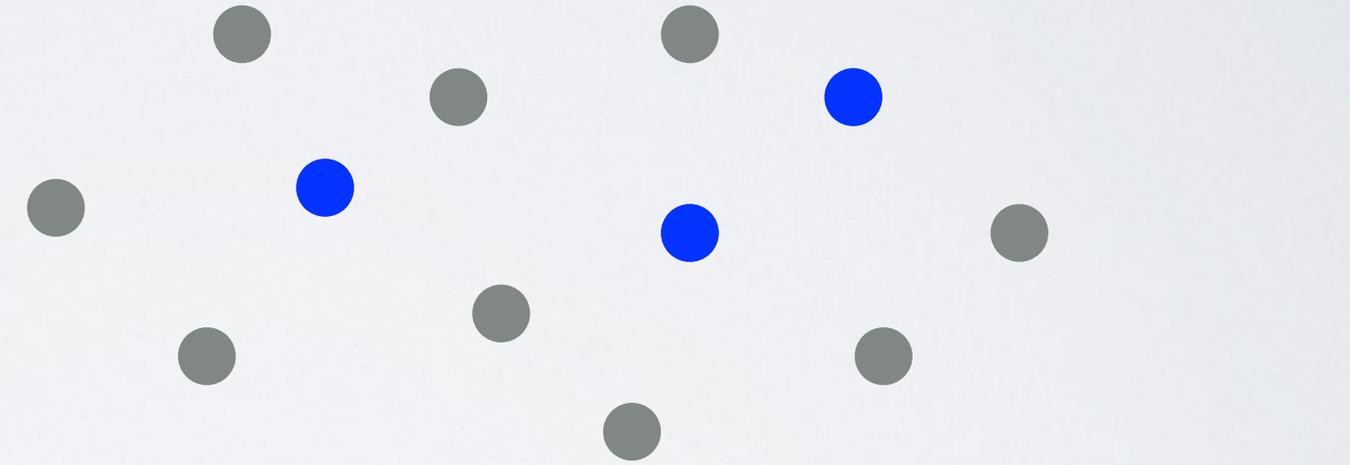
Accept/Reject as a Hypothesis Testing

- Given a set $\mathcal{Y} = \{y_i\}_{i=1}^{N}$, with a mean μ and a constant μ_0, test:
  - H_0: $\mu > \mu_0$
  - H_a: $\mu < \mu_0$

$$\frac{1}{N}\sum_{i=1}^{N} y_i > \mu_0 \, ?$$

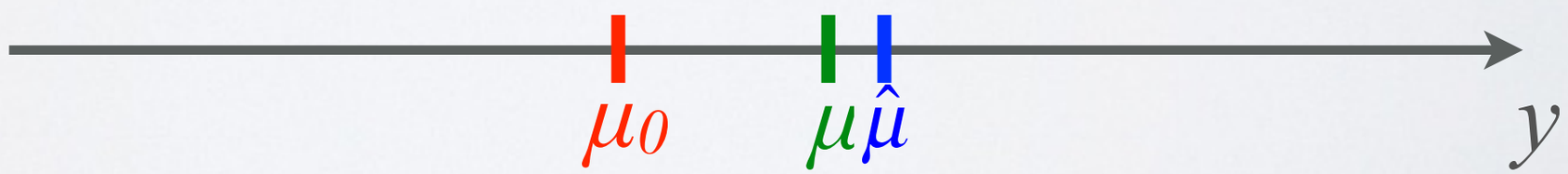
$$\log\left(\frac{P(x_i|\theta')}{P(x_i|\theta_t)}\right) \qquad \frac{1}{N}\log\left(u\frac{q(\theta'|\theta_t)P(\theta_t)}{q(\theta_t|\theta')P(\theta')}\right)$$

1. Draw $Y \subseteq \mathcal{Y}$ of a size n
2. Check $\hat{\mu} > \mu_0$?



$O(N) \rightarrow O(n)$

n?



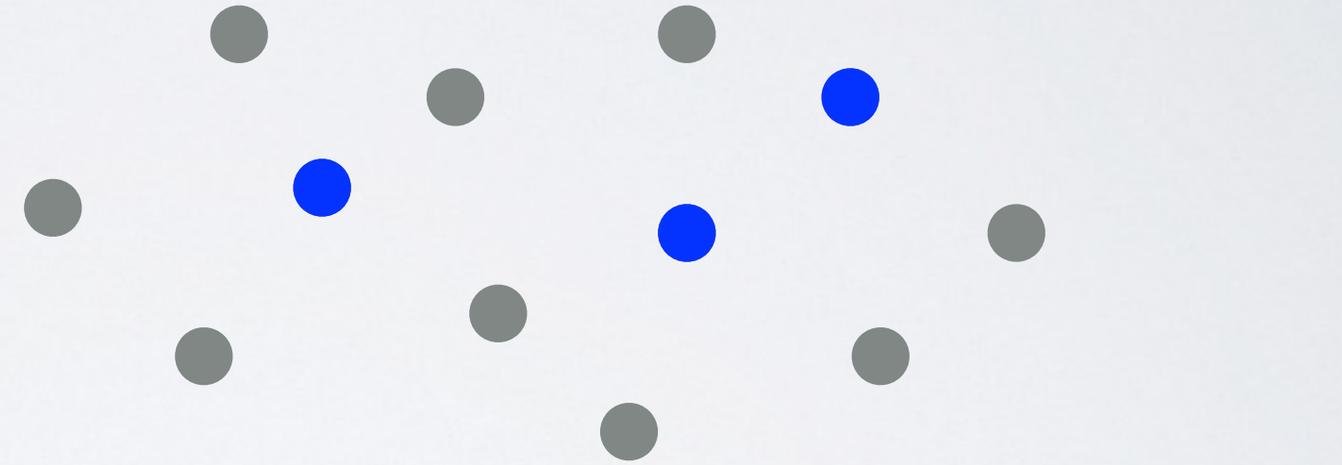
Accept/Reject as a Hypothesis Testing

- Given a set $\mathcal{Y}=\{y_i\}_{i=1}^{N}$, with a mean μ and a constant μ_0, test:
  - H_0: $\mu>\mu_0$
  - H_a: $\mu<\mu_0$

$$\frac{1}{N}\sum_{i=1}^{N} y_i > \mu_0 \,?$$

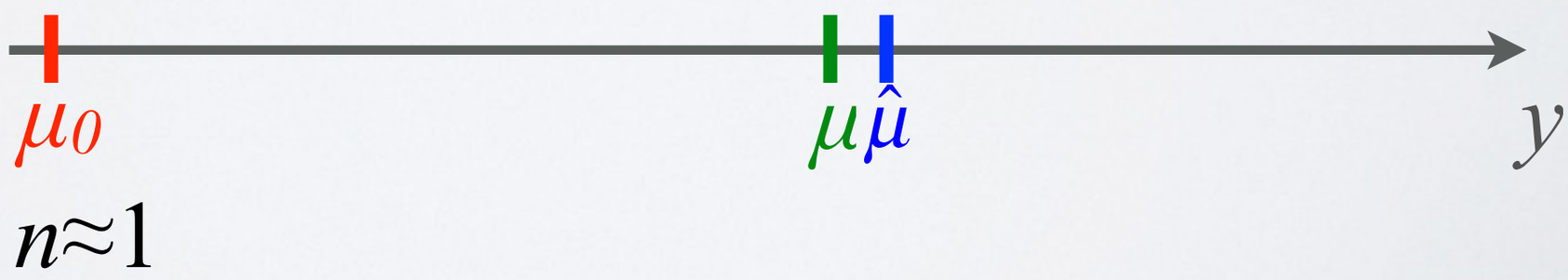
$$\log\left(\frac{P(x_i|\theta')}{P(x_i|\theta_t)}\right) \qquad \frac{1}{N}\log\left(u\frac{q(\theta'|\theta_t)P(\theta_t)}{q(\theta_t|\theta')P(\theta')}\right)$$

1. Draw $Y \subseteq \mathcal{Y}$ of a size n
2. Check $\hat{\mu} > \mu_0$?



$O(N) \rightarrow O(n)$

n?



$n \approx 1$

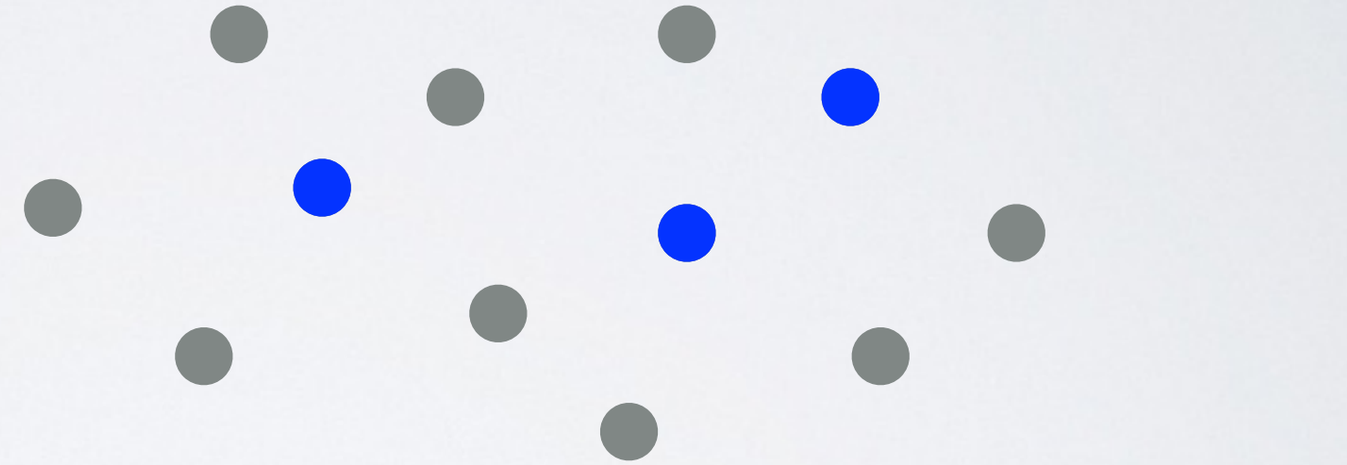
Accept/Reject as a Hypothesis Testing

- Given a set $\mathcal{Y}=\{y_i\}_{i=1}^{N}$, with a mean μ and a constant μ_0, test:
  - H_0: $\mu>\mu_0$
  - H_a: $\mu<\mu_0$

$$\frac{1}{N}\sum_{i=1}^{N} y_i > \mu_0 \,?$$

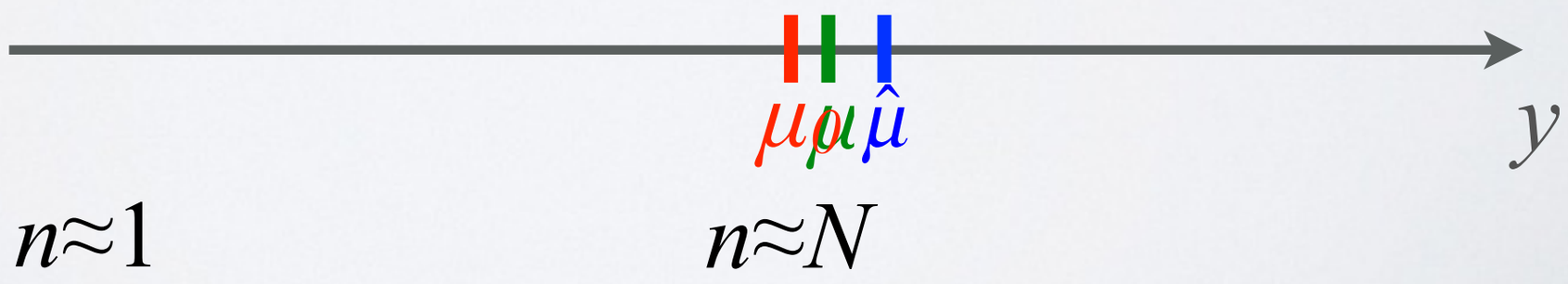
$$\log\left(\frac{P(x_i|\theta')}{P(x_i|\theta_t)}\right) \qquad \frac{1}{N}\log\left(u\frac{q(\theta'|\theta_t)P(\theta_t)}{q(\theta_t|\theta')P(\theta')}\right)$$

1. Draw $Y \subseteq \mathcal{Y}$ of a size n
2. Check $\hat{\mu} > \mu_0$?



$O(N) \rightarrow O(n)$

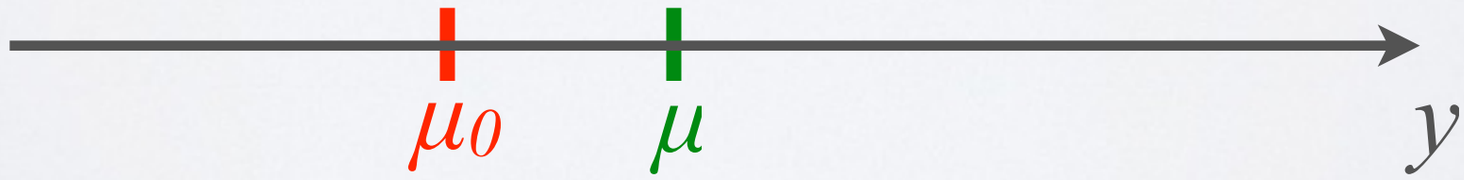
n?



$n \approx 1$ $\qquad$ $n \approx N$

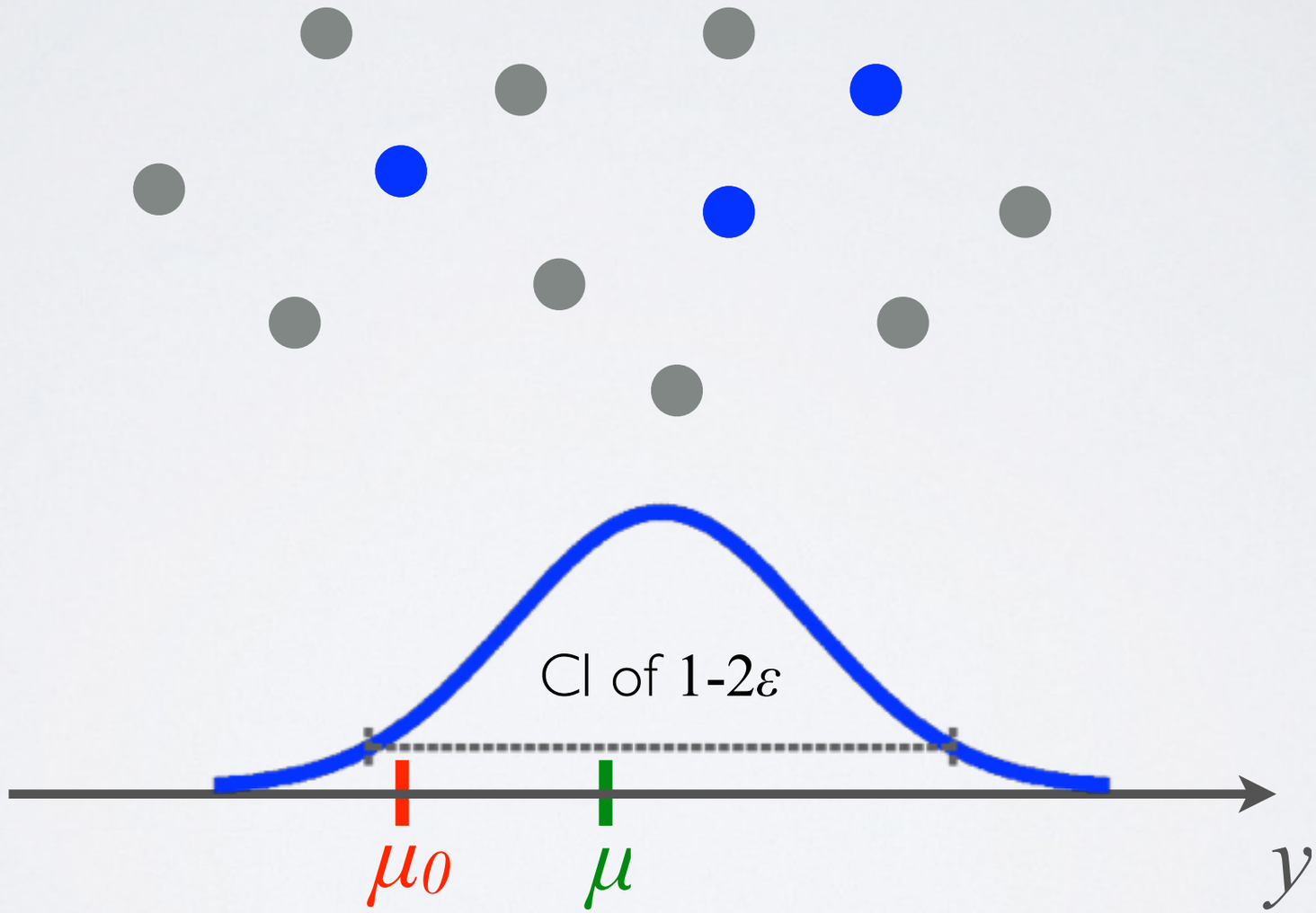
Sequential Hypothesis Testing

- Keep drawing data cases randomly ***w/o replacement*** until we are confident about our approximation.



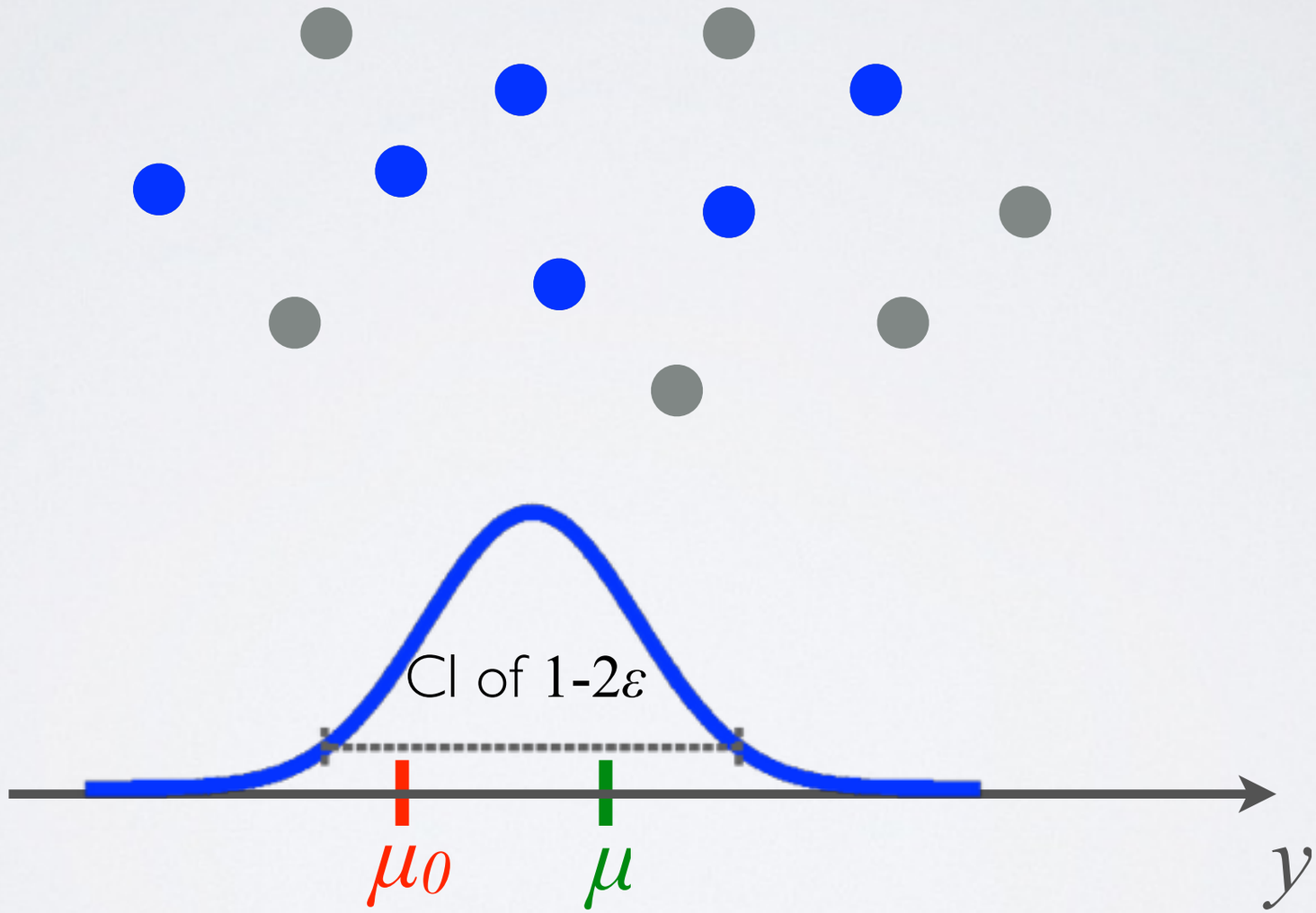
Sequential Hypothesis Testing

- Keep drawing data cases randomly ***w/o replacement*** until we are confident about our approximation.



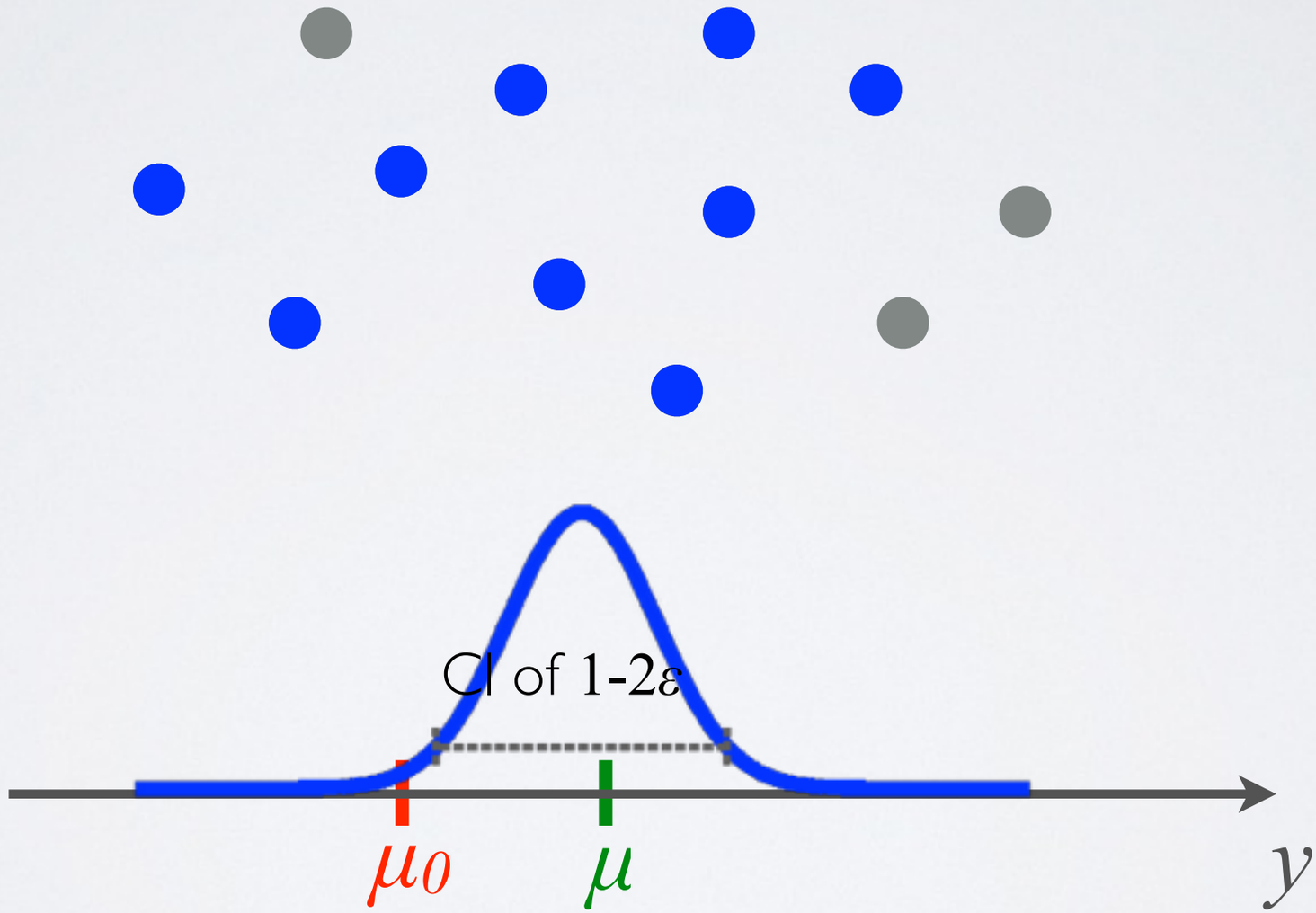
Sequential Hypothesis Testing

- Keep drawing data cases randomly ***w/o replacement*** until we are confident about our approximation.



Sequential Hypothesis Testing

- Keep drawing data cases randomly ***w/o replacement*** until we are confident about our approximation.



Sequential Hypothesis Testing

UNIVERSITY OF CAMBRIDGE

Sequential Hypothesis Testing

Draw $Y_j \subseteq \mathcal{Y}$ w/o replacement

$n \leftarrow n+m$

Sequential Hypothesis Testing

Draw $Y_j \subseteq \mathcal{Y}$ w/o replacement

$n \leftarrow n+m$

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}_j = \frac{\hat{\sigma}(y_i)}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

Finite population correction factor

Sequential Hypothesis Testing

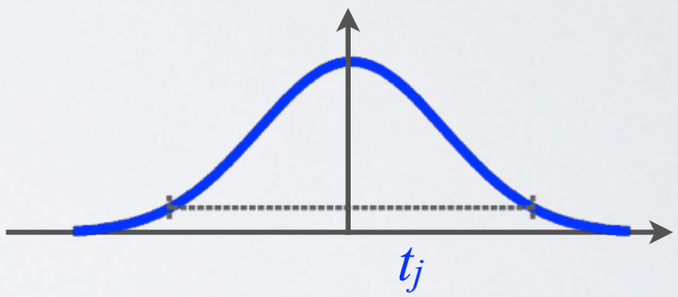
Draw $Y_j \subseteq \mathcal{Y}$ w/o replacement

$n \leftarrow n+m$

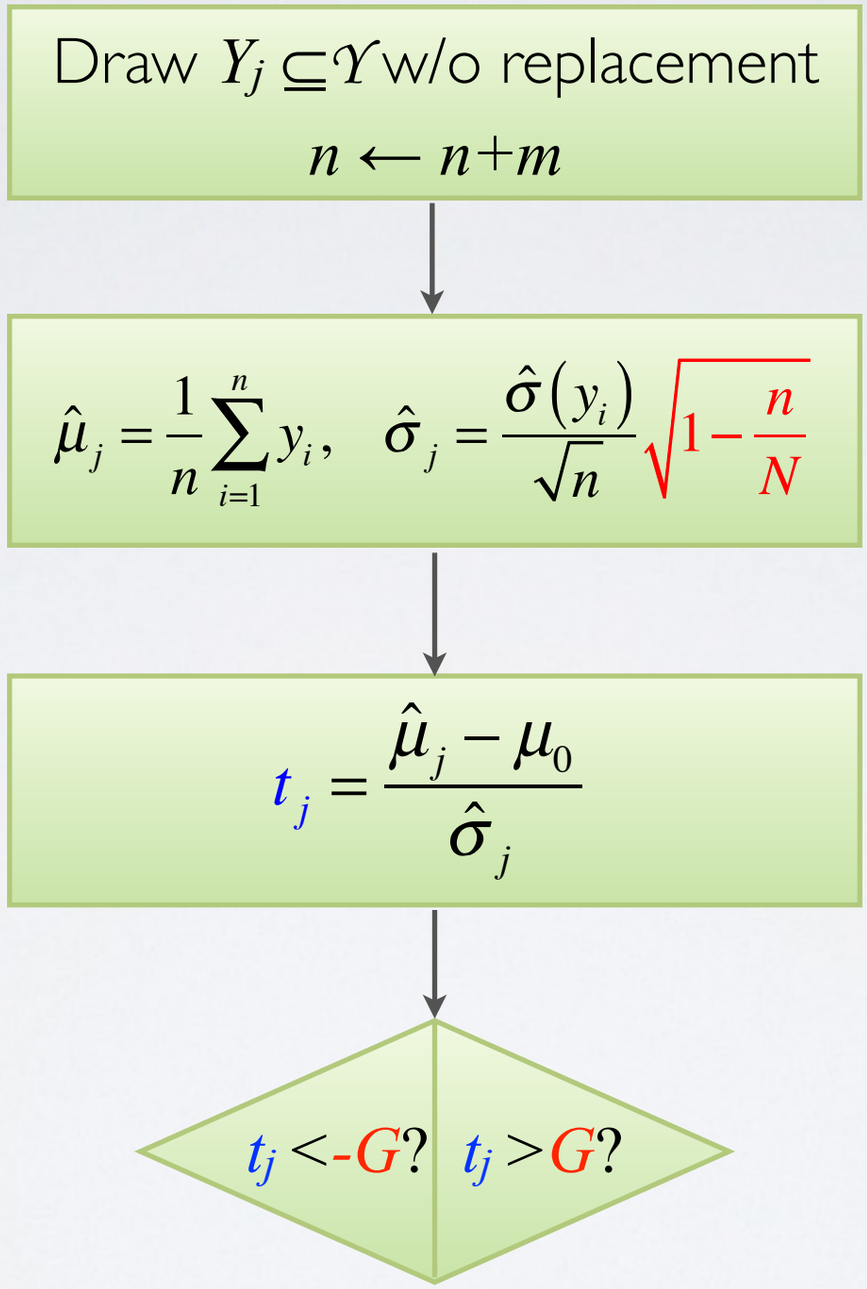
$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}_j = \frac{\hat{\sigma}(y_i)}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

Finite population correction factor

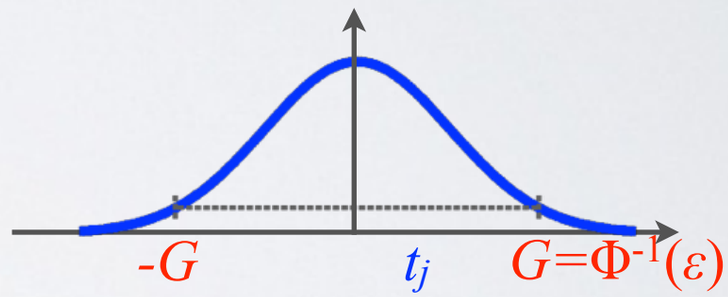
$$t_j = \frac{\hat{\mu}_j - \mu_0}{\hat{\sigma}_j}$$



Sequential Hypothesis Testing



Finite population correction factor



Sequential Hypothesis Testing

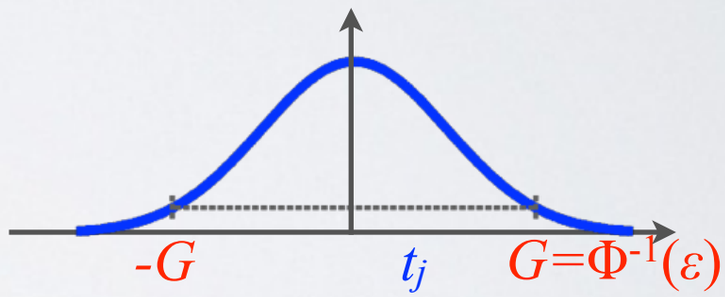
Draw $Y_j \subseteq \mathcal{Y}$ w/o replacement

$n \leftarrow n+m$

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}_j = \frac{\hat{\sigma}(y_i)}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

Finite population correction factor

$$t_j = \frac{\hat{\mu}_j - \mu_0}{\hat{\sigma}_j}$$



$t_j < -G$? $t_j > G$?

$\theta_{t+1} = \theta_t$

$\theta_{t+1} = \theta'$

Sequential Hypothesis Testing

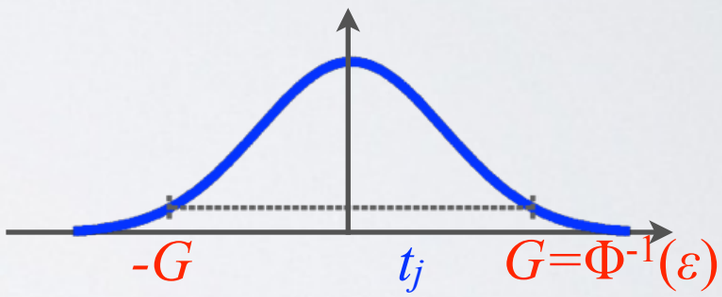
Draw $Y_j \subseteq \mathcal{Y}$ w/o replacement

$n \leftarrow n+m$

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \hat{\sigma}_j = \frac{\hat{\sigma}(y_i)}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

Finite population correction factor

$$t_j = \frac{\hat{\mu}_j - \mu_0}{\hat{\sigma}_j}$$



$t_j < -G$? $t_j > G$?

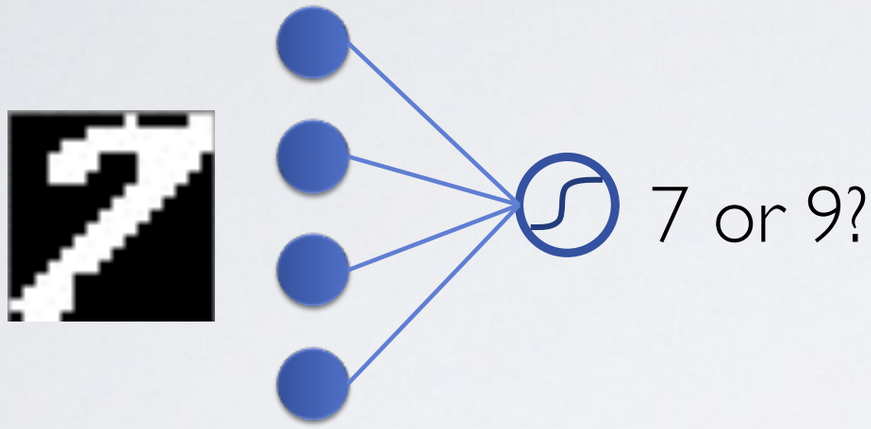
Need more data

$j \leftarrow j+1$

$\theta_{t+1} = \theta_t$

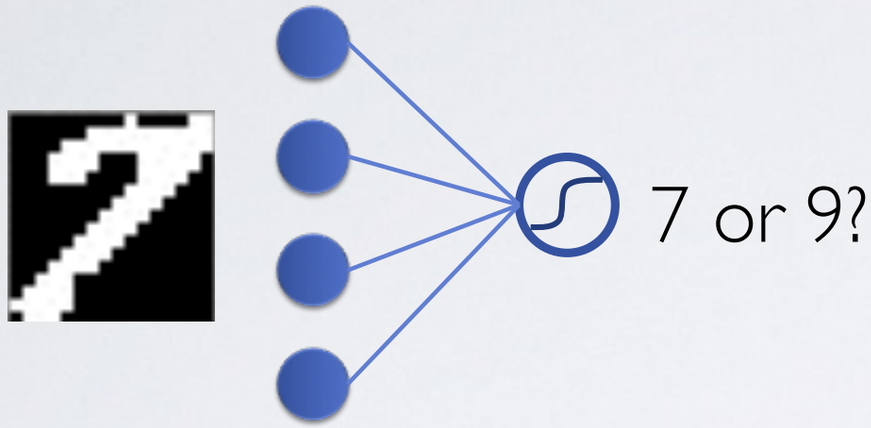
$\theta_{t+1} = \theta'$

Application I: Posterior Inference for Bayesian Logistic Regression



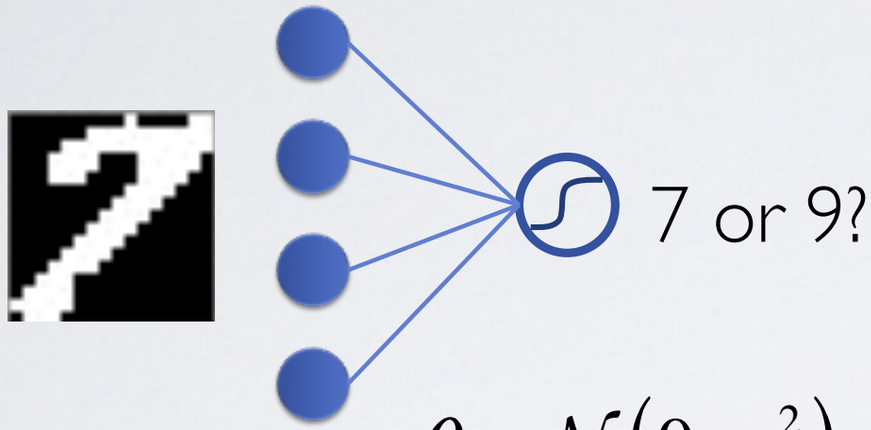
Application I: Posterior Inference for Bayesian Logistic Regression

- Data: MNIST digit 7 and 9, 12214 data points.



Application I: Posterior Inference for Bayesian Logistic Regression

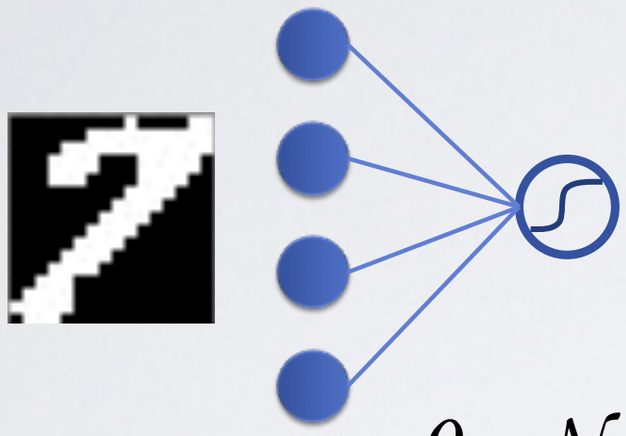
- Data: MNIST digit 7 and 9, 12214 data points.



$$\theta \sim \mathcal{N}\left(0,\sigma^2\right)$$

Application I: Posterior Inference for Bayesian Logistic Regression

- Data: MNIST digit 7 and 9, 12214 data points.

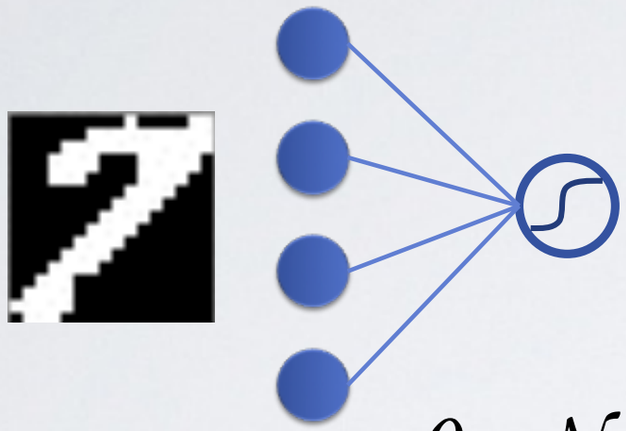


$$\theta \sim \mathcal{N}\left(0,\sigma^2\right)$$

$$P(\theta \mid \mathcal{D}) \propto \mathcal{N}(\theta) \prod_{i=1}^{N} \text{logit}\left(y_i \mid x_i, \theta\right)$$

Application I: Posterior Inference for Bayesian Logistic Regression

- Data: MNIST digit 7 and 9, 12214 data points.



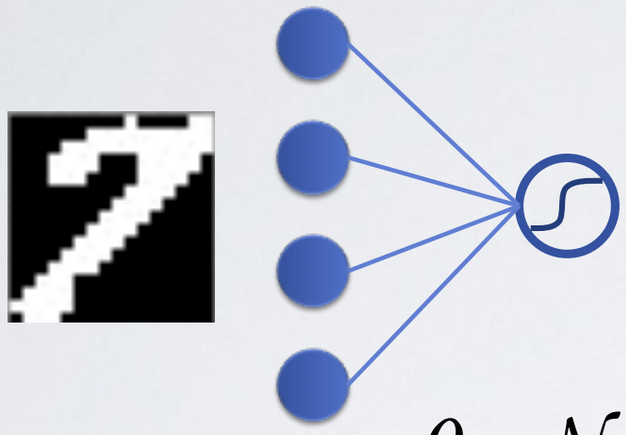
$$\theta \sim \mathcal{N}\left(0,\sigma^2\right)$$

$$P(\theta \mid \mathcal{D}) \propto \mathcal{N}(\theta) \prod_{i=1}^{N} \text{logit}\left(y_i \mid x_i, \theta\right)$$

- Sampler: M-H with Random walk proposal.

Application I: Posterior Inference for Bayesian Logistic Regression

- Data: MNIST digit 7 and 9, 12214 data points.



$$\theta \sim \mathcal{N}\left(0,\sigma^2\right)$$

$$P(\theta \mid \mathcal{D}) \propto \mathcal{N}(\theta) \prod_{i=1}^{N} \text{logit}\left(y_i \mid x_i, \theta\right)$$

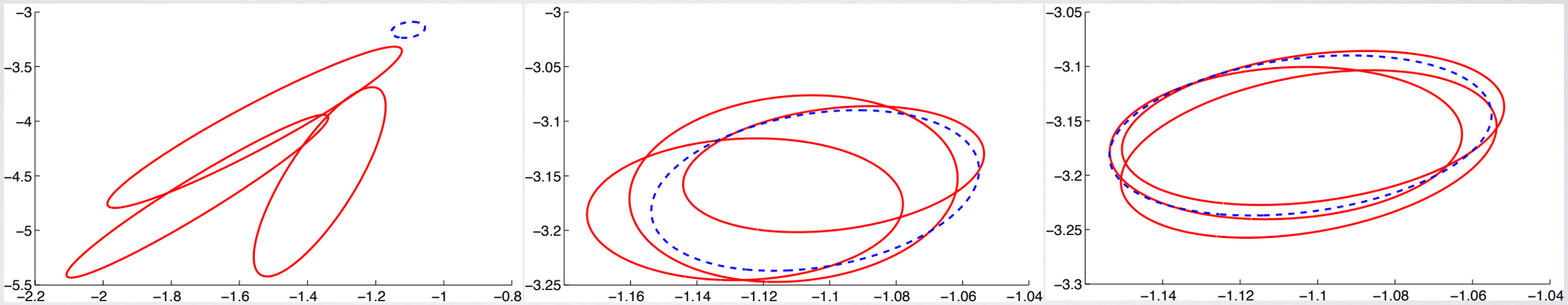
- Sampler: M-H with Random walk proposal.

- Quantity of interest: $P(x^* \mid \mathcal{D}) = \mathbb{E}_{P(\theta \mid \mathcal{D})}\left[\text{logit}(x^* \mid \theta)\right]$

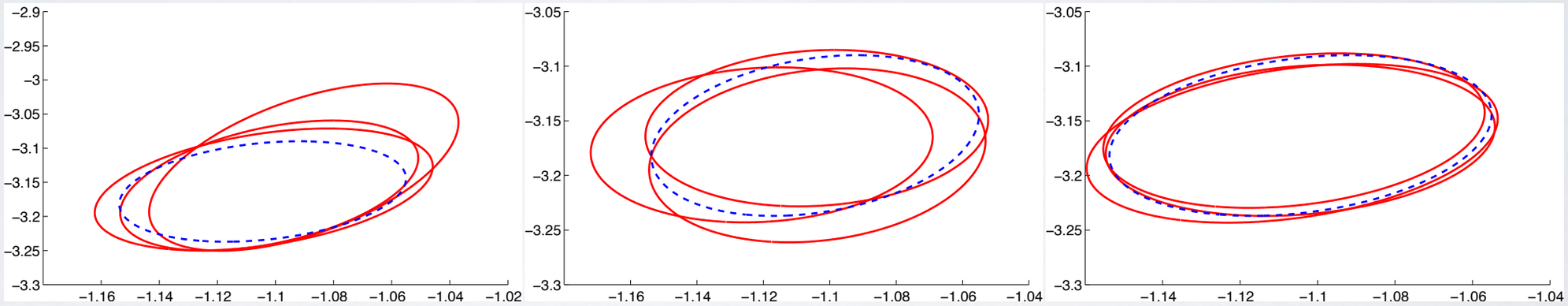
$$\downarrow$$

$$f(\theta)$$

Exact MH



Approximate
ε=0.01



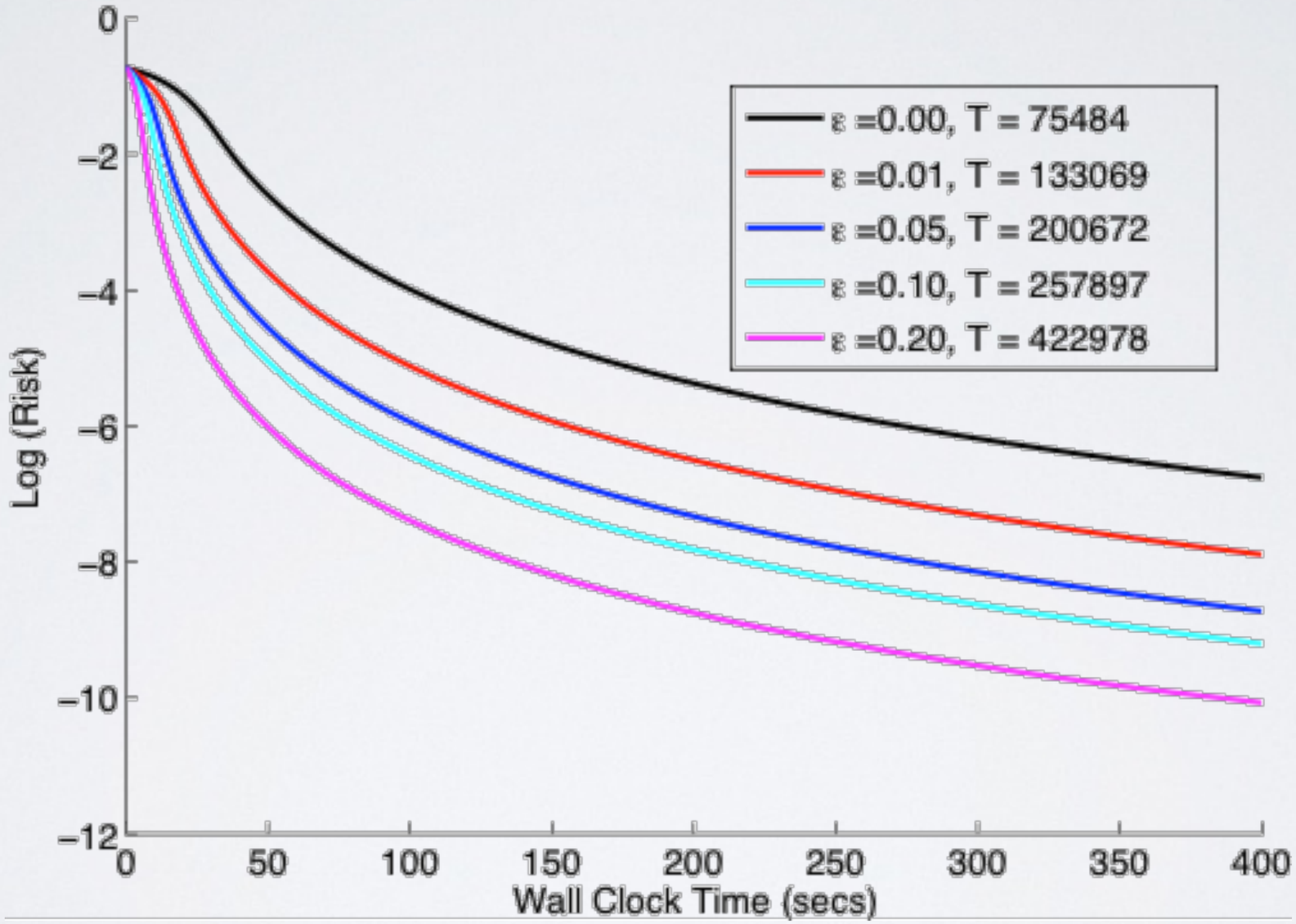
50 sec

100 sec

400 sec

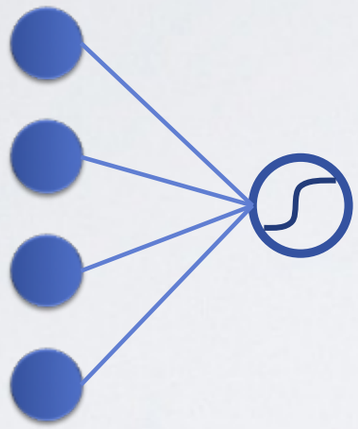
Wall Clock Time

$$Risk = \mathbb{E}\left[\left(\langle f \rangle_{S_0} - \overline{f(\theta_t)}\right)^2\right]$$



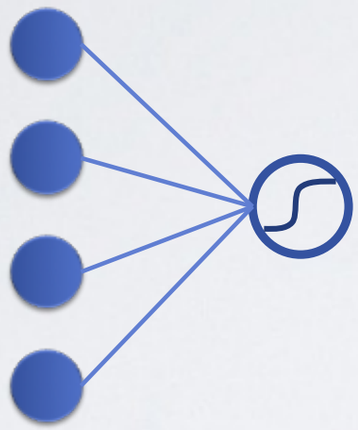
Application II: Variable selection in Logistic Regression

- Data: MiniBooNE 130,065 data points with 50 features
- Sampler: Reversible jump MCMC



Application II: Variable selection in Logistic Regression

- Data: MiniBooNE 130,065 data points with 50 features
- Sampler: Reversible jump MCMC



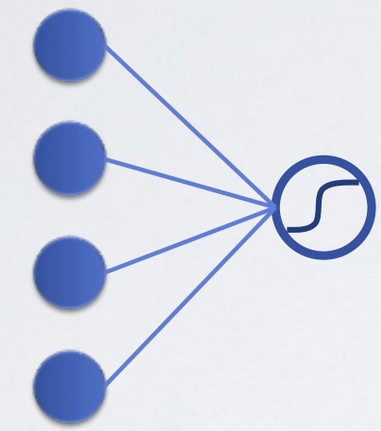
$$\theta_i = \beta_i \gamma_i$$

$$\beta_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

$$|\gamma| \sim \text{Poisson}$$

Application II: Variable selection in Logistic Regression

- Data: MiniBooNE 130,065 data points with 50 features
- Sampler: Reversible jump MCMC



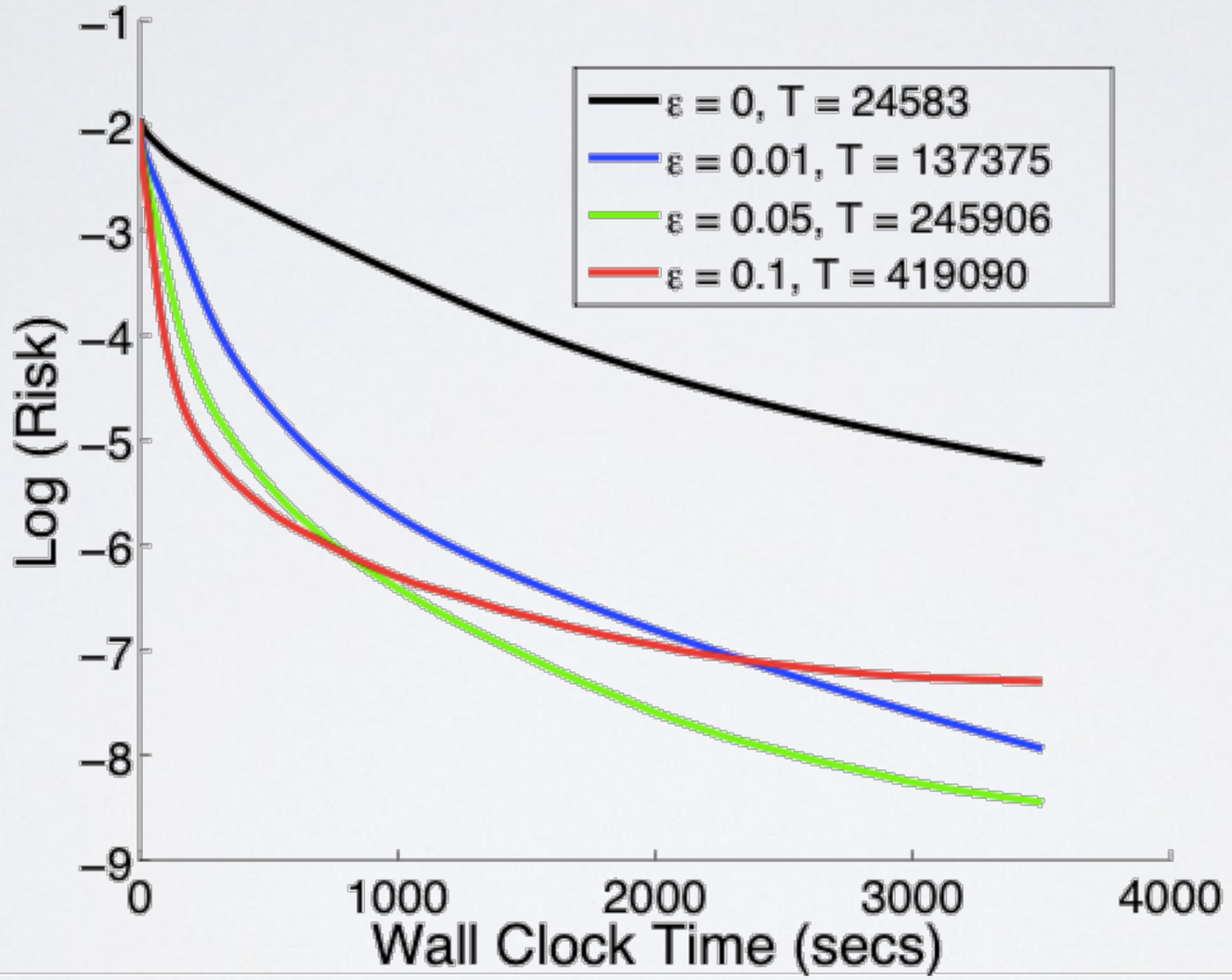
$$\theta_i = \beta_i \gamma_i$$

$$\beta_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

$$|\gamma| \sim \text{Poisson}$$

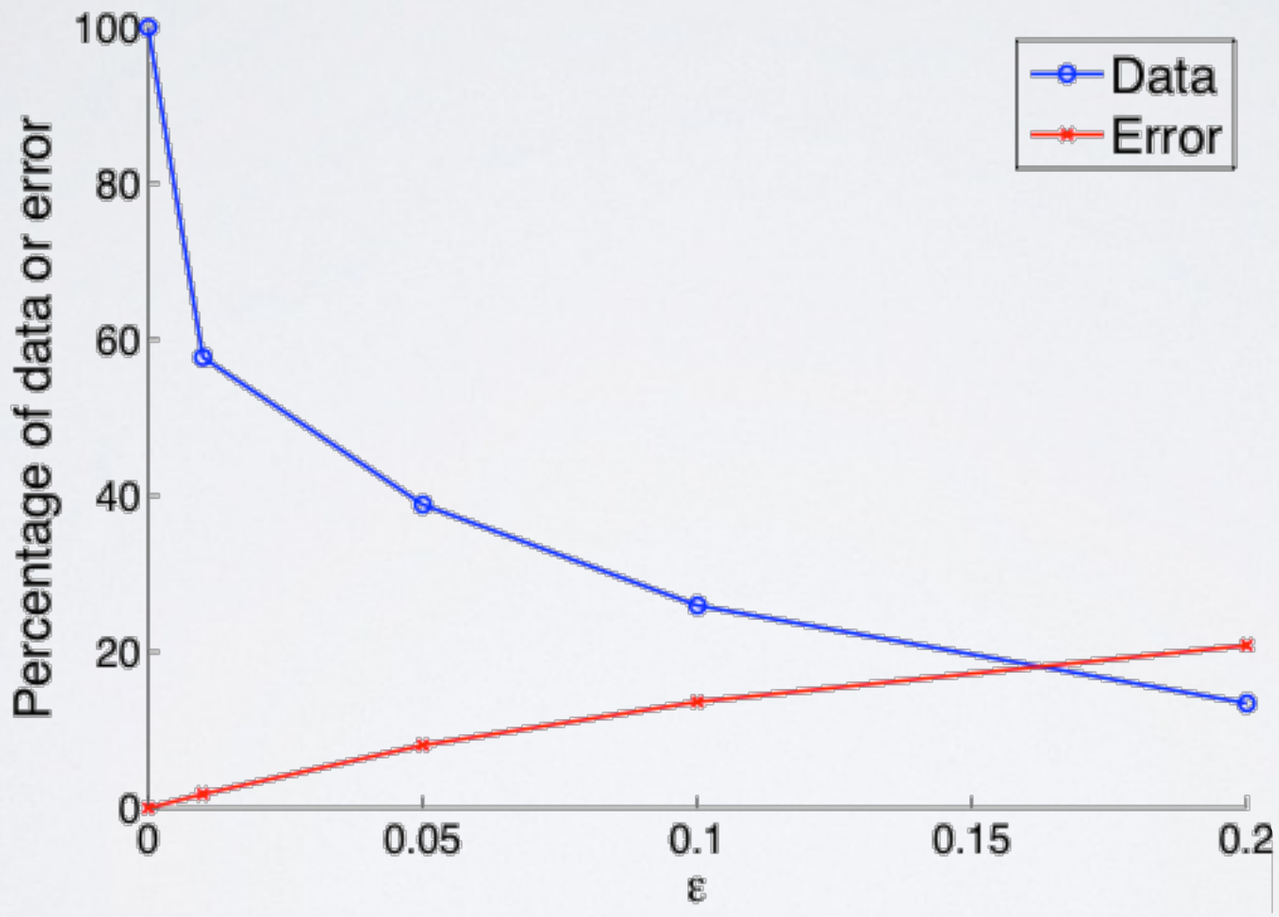
$$P\left(\beta, \gamma \mid \mathcal{D}\right) \propto \mathcal{N}\left(\beta\right) P(\gamma) \prod_{i=1}^{N} \text{logit}\left(y_i \mid x_i, \beta\gamma\right)$$

$$Risk = \mathbb{E}\left[\left(\langle f \rangle_{S_0} - \overline{f(\theta_t)}\right)^2\right]$$

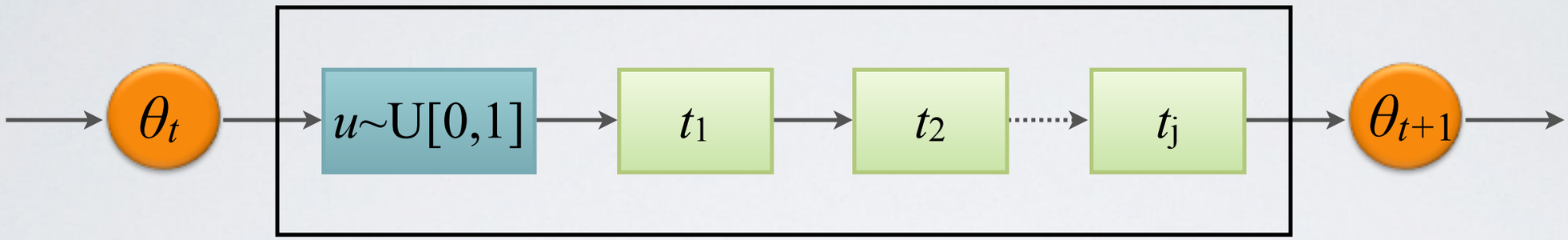


How to Design the Sequential Test?

$\mu > \mu_0$?

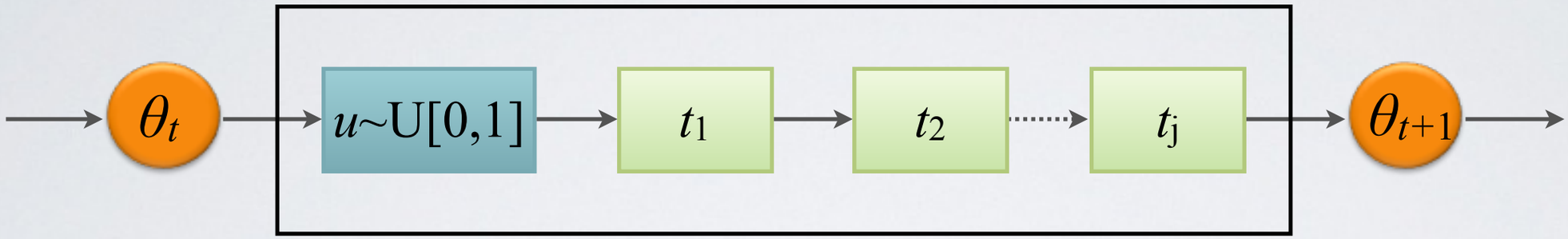


Sequential Test Design



Sequential Test Design

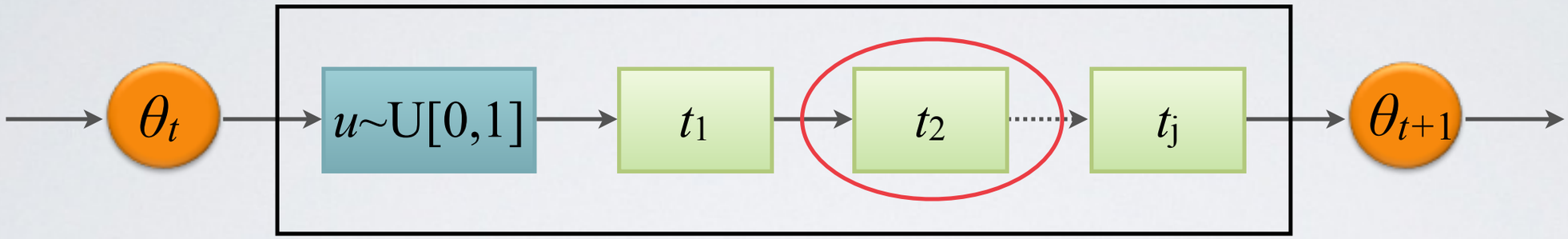
- How to design the sequential test to minimize the ***usage of data*** subject to a tolerance of the ***error***?



- The stationary distribution, S_ε

Sequential Test Design

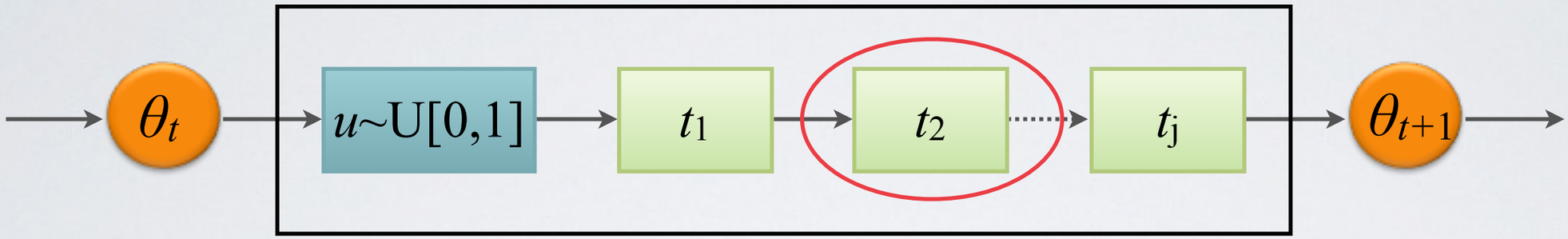
- How to design the sequential test to minimize the ***usage of data*** subject to a tolerance of the ***error***?



- One step of test, error: ε, data: m
- The stationary distribution, S_ε

Sequential Test Design

- How to design the sequential test to minimize the ***usage of data*** subject to a tolerance of the ***error***?

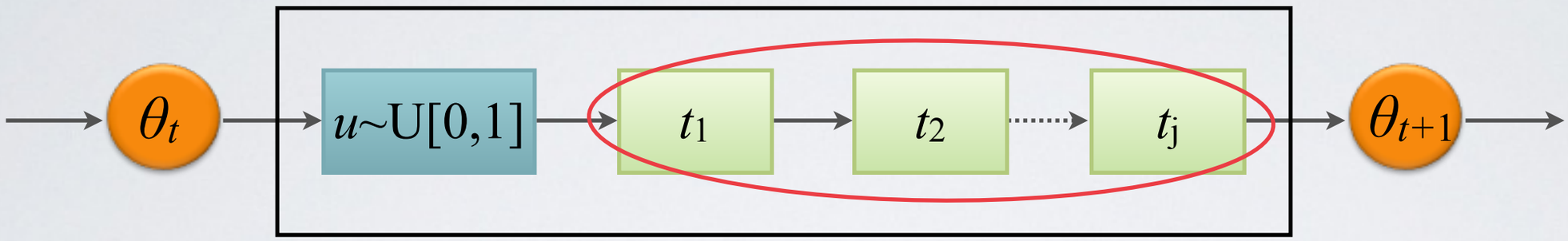


- One step of test, error: ε, data: m

- The stationary distribution, S_ε

Sequential Test Design

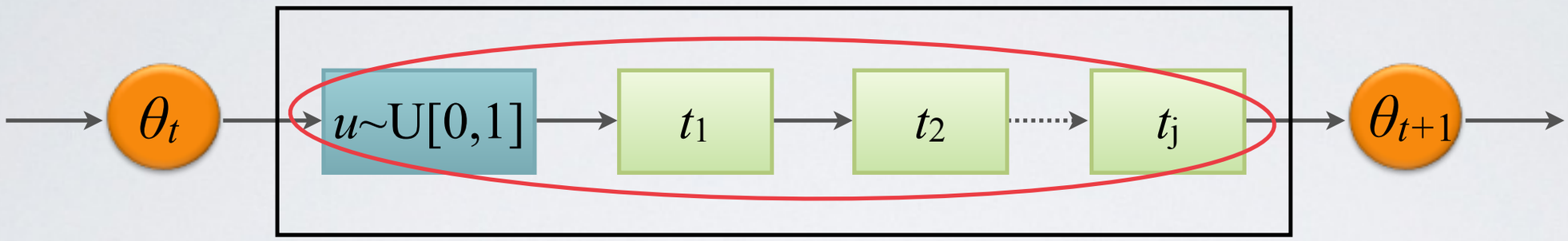
- How to design the sequential test to minimize the ***usage of data*** subject to a tolerance of the ***error***?



- One step of test, error: ε, data: m
- One sequential test, conditioned on u
- The stationary distribution, S_ε

Sequential Test Design

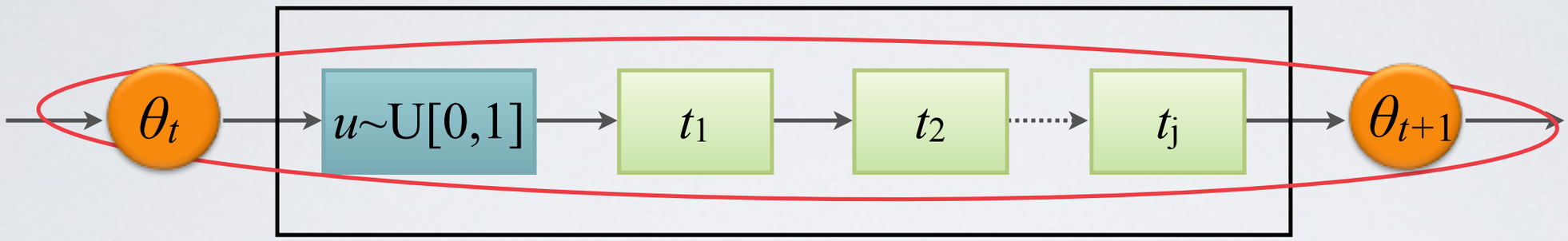
- How to design the sequential test to minimize the ***usage of data*** subject to a tolerance of the ***error***?



- One step of test, error: ε, data: m
- One sequential test, conditioned on u
- One accept/reject step (marginalized over u)
- The stationary distribution, S_ε

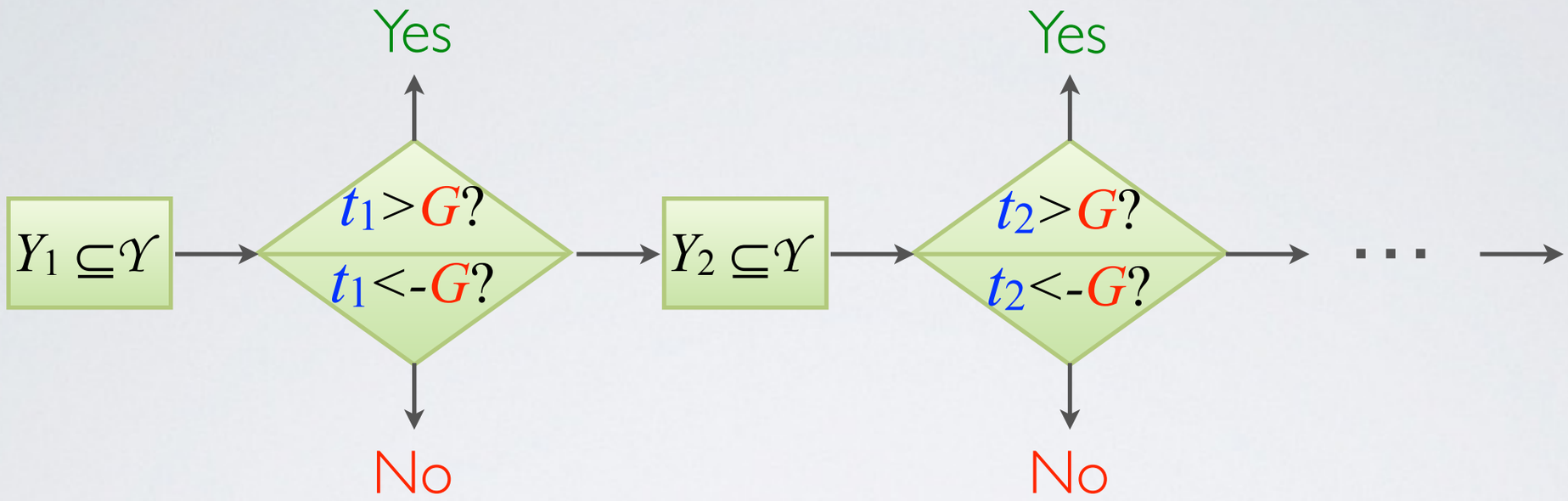
Sequential Test Design

- How to design the sequential test to minimize the ***usage of data*** subject to a tolerance of the ***error***?



- One step of test, error: ε, data: m
- One sequential test, conditioned on u
- One accept/reject step (marginalized over u)
- The stationary distribution, S_ε

One Sequential Test



$\mu > \mu_0$

$$\bar{\pi} = \mathbb{E}\left[j^* \right] / J$$

$$\mathcal{E} = P(t_{j^*} < -G)$$

Random Walk of t_j

- Two assumptions about Y_j, when m is large $Y_1 \subseteq \mathcal{Y} \longrightarrow Y_2 \subseteq \mathcal{Y} \longrightarrow \cdots$
  1. $\{\bar{y}_j\}$ is jointly Normal
  2. Accurate estimate of σ_y, $t_j = \dfrac{\overline{\{\bar{y}_j\}} - \mu_0}{\hat{\sigma}_j} \approx \dfrac{\overline{\{\bar{y}_j\}} - \mu_0}{\sigma_j} \sim \mathcal{N}(t_j)$

Random Walk of t_j

- Two assumptions about Y_j, when m is large $Y_1 \subseteq \mathcal{Y} \longrightarrow Y_2 \subseteq \mathcal{Y} \longrightarrow \cdots$
  1. $\{\bar{y}_j\}$ is jointly Normal
  2. Accurate estimate of σ_y, $t_j = \dfrac{\overline{\{\bar{y}_j\}} - \mu_0}{\hat{\sigma}_j} \approx \dfrac{\overline{\{\bar{y}_j\}} - \mu_0}{\sigma_j} \sim \mathcal{N}(t_j)$

$$P\left(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_j\right) = \mathcal{N}(\mu, \Sigma)$$

Random Walk of t_j

- Two assumptions about Y_j, when m is large $Y_1 \subseteq \mathcal{Y} \longrightarrow Y_2 \subseteq \mathcal{Y} \longrightarrow \cdots$
  1. $\{\bar{y}_j\}$ is jointly Normal
  2. Accurate estimate of σ_y, $t_j = \dfrac{\overline{\{\bar{y}_j\}} - \mu_0}{\hat{\sigma}_j} \approx \dfrac{\overline{\{\bar{y}_j\}} - \mu_0}{\sigma_j} \sim \mathcal{N}(t_j)$

$$P\left(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_j\right) = \mathcal{N}(\mu, \Sigma)$$

- Sample with Replacement

$$\Sigma(\bar{y}) = \frac{\sigma_y^2}{m}\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}$$

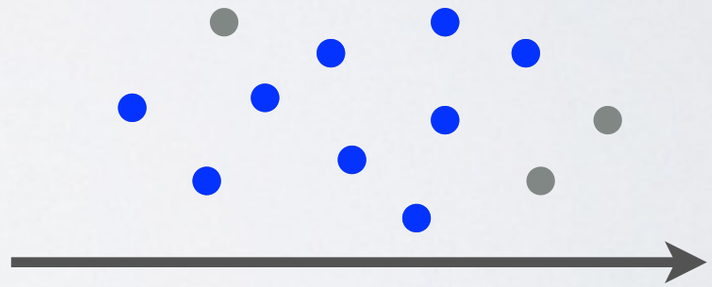
Random Walk of t_j

- Two assumptions about Y_j, when m is large $Y_1 \subseteq \mathcal{Y} \longrightarrow Y_2 \subseteq \mathcal{Y} \longrightarrow \cdots$
  1. $\{\bar{y}_j\}$ is jointly Normal
  2. Accurate estimate of σ_y, $t_j = \dfrac{\overline{\{\bar{y}_j\}} - \mu_0}{\hat{\sigma}_j} \approx \dfrac{\overline{\{\bar{y}_j\}} - \mu_0}{\sigma_j} \sim \mathcal{N}(t_j)$

$$P\left(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_j\right) = \mathcal{N}(\mu, \Sigma)$$

- Sample without Replacement

$$\Sigma(\bar{y}) = \frac{\sigma_y^2}{m}\begin{bmatrix} 1 & -\frac{1}{N} & -\frac{1}{N} \\ -\frac{1}{N} & 1 & -\frac{1}{N} \\ -\frac{1}{N} & -\frac{1}{N} & 1 \end{bmatrix}$$

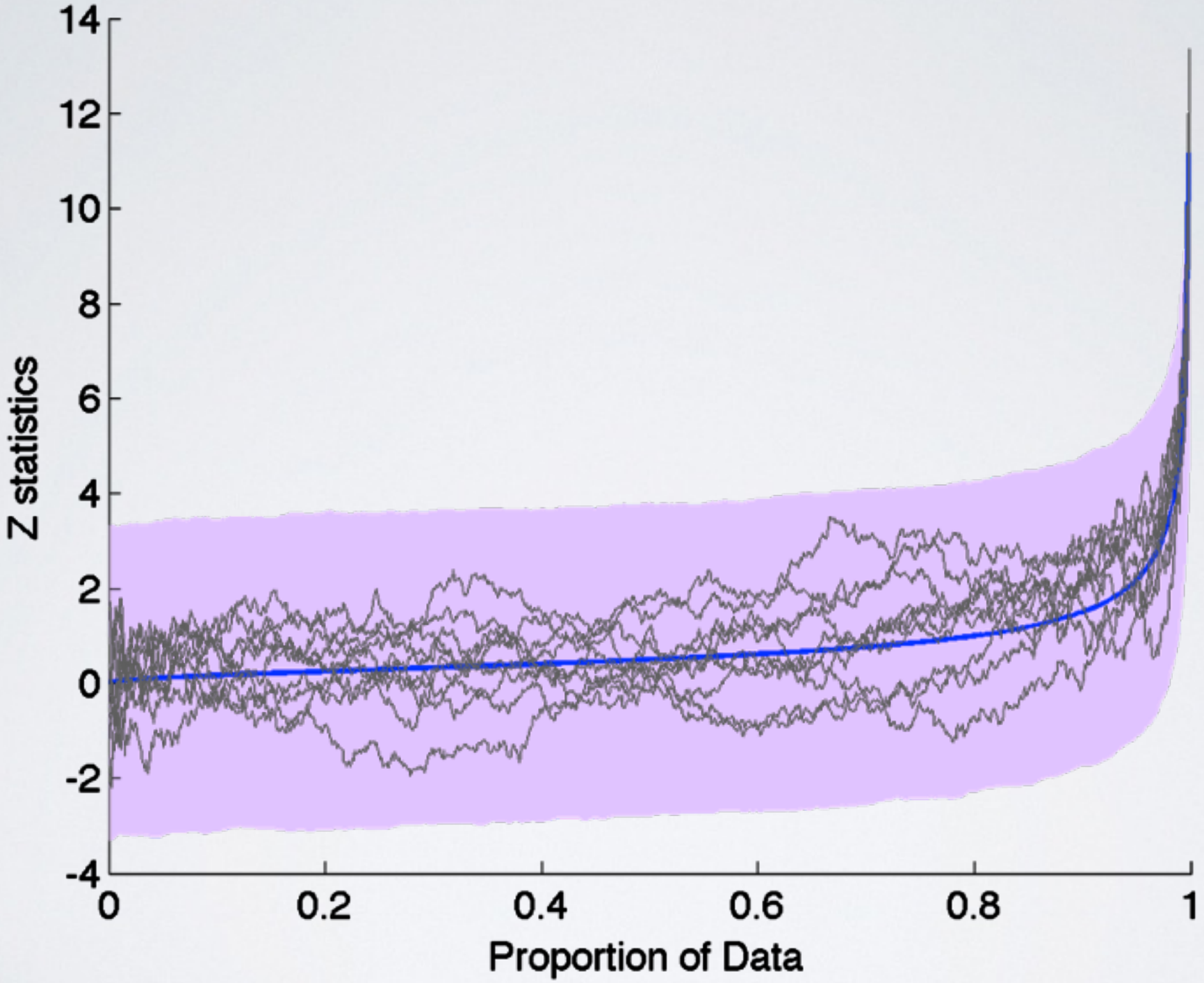


Random Walk of t_j

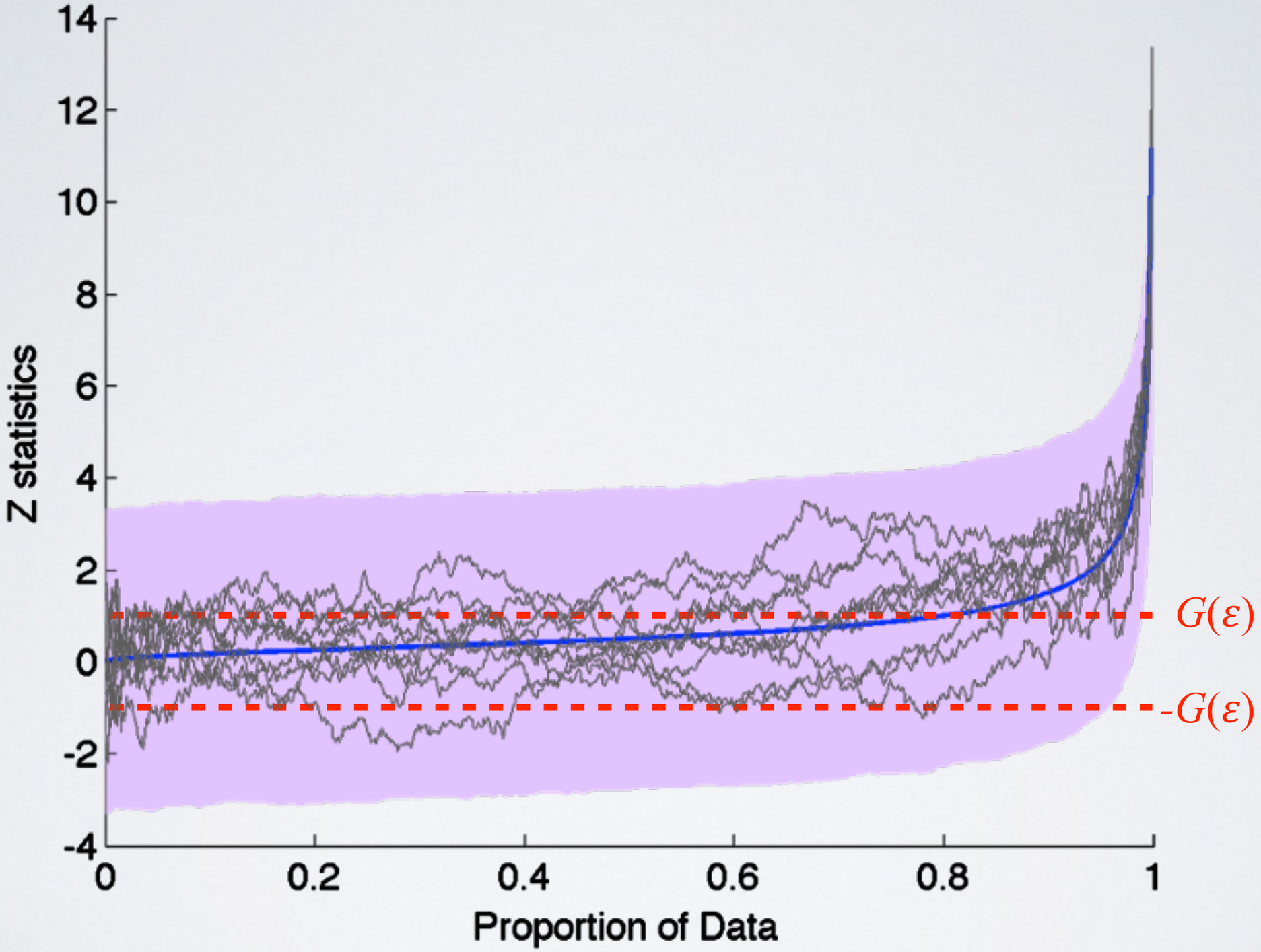
- Two assumptions about Y_j, when m is large $Y_1 \subseteq \mathcal{Y} \longrightarrow Y_2 \subseteq \mathcal{Y} \longrightarrow \cdots$
  1. $\{\bar{y}_j\}$ is jointly Normal
  2. Accurate estimate of σ_y, $t_j = \dfrac{\overline{\{\bar{y}_j\}} - \mu_0}{\hat{\sigma}_j} \approx \dfrac{\overline{\{\bar{y}_j\}} - \mu_0}{\sigma_j} \sim \mathcal{N}(t_j)$

$$P\left(t_j \mid t_{j-1}\right) = \mathcal{N}\left(\mu_t\left(t_{j-1}\right), \Sigma_t\right)$$

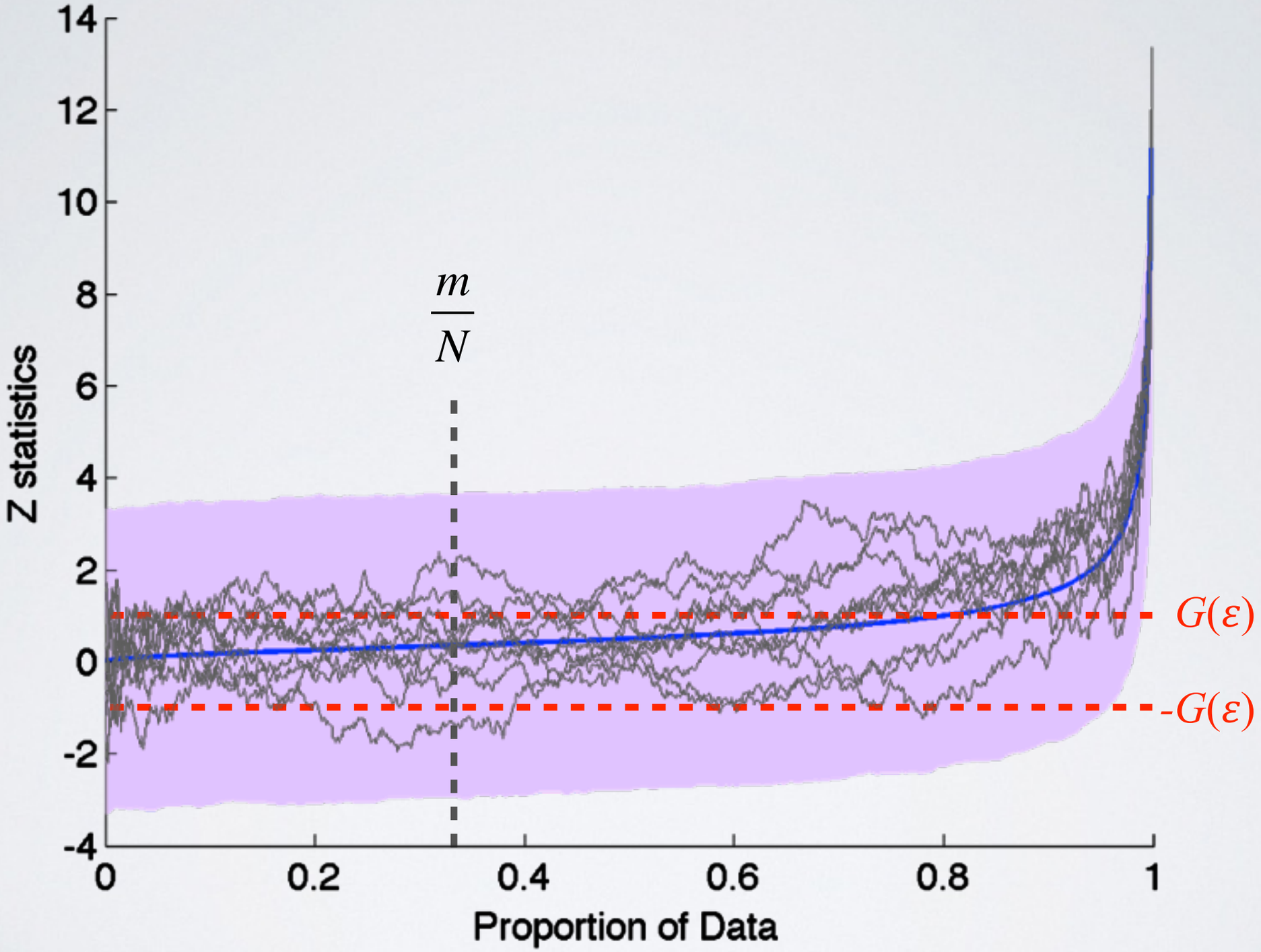
Example: $\mu > \mu_0$



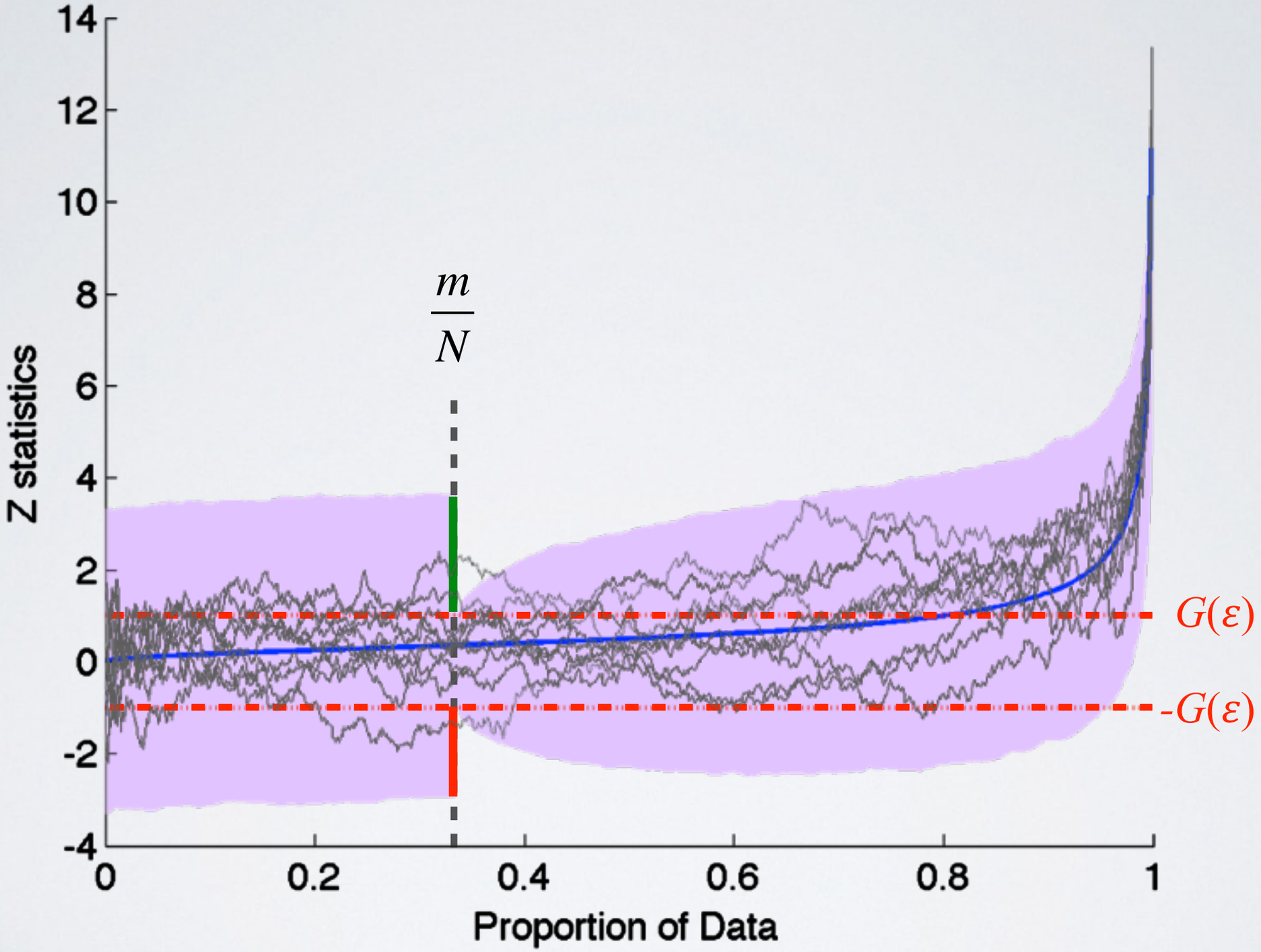
Example: $\mu > \mu_0$



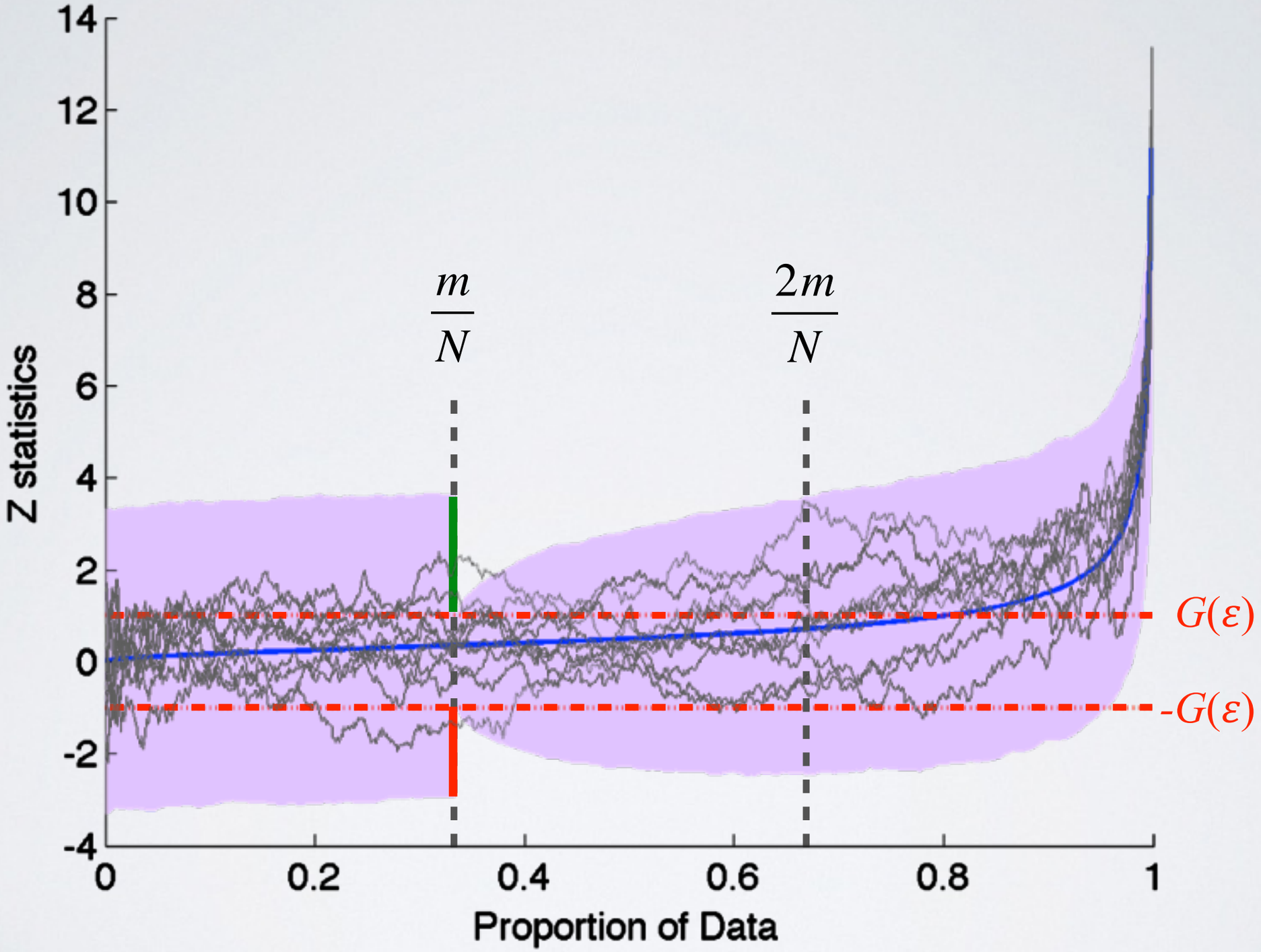
Example: $\mu > \mu_0$



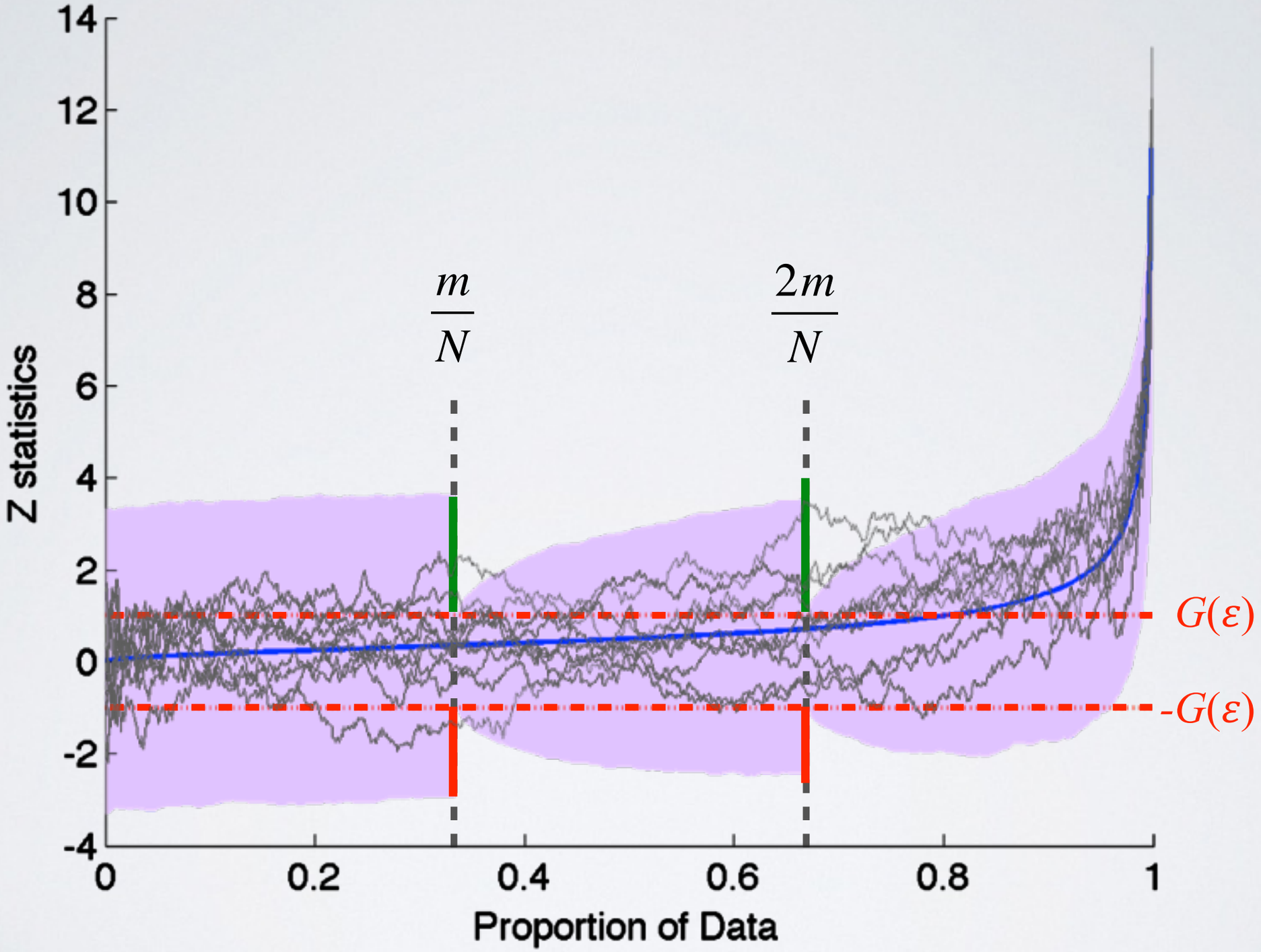
Example: $\mu > \mu_0$



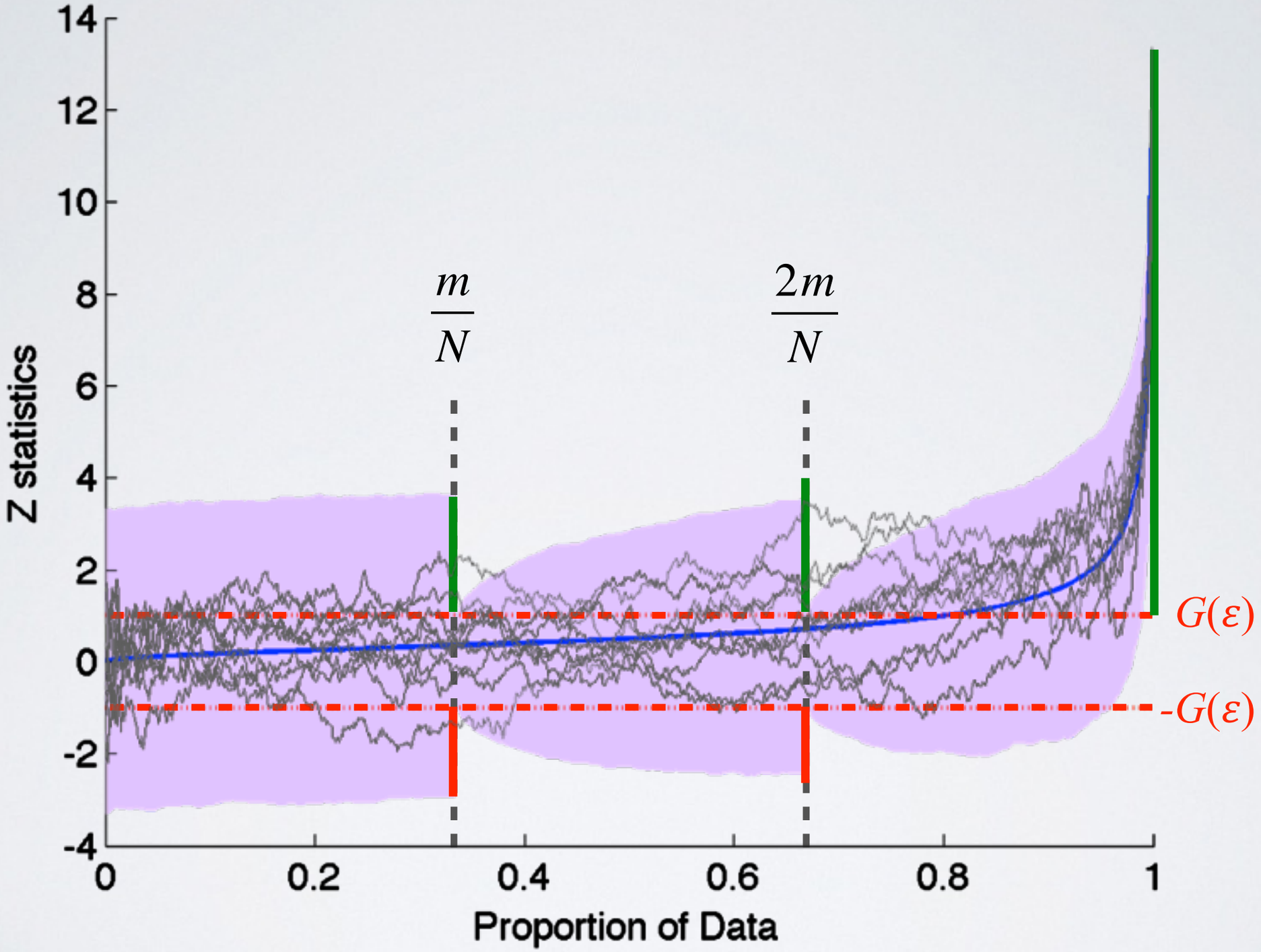
Example: $\mu > \mu_0$

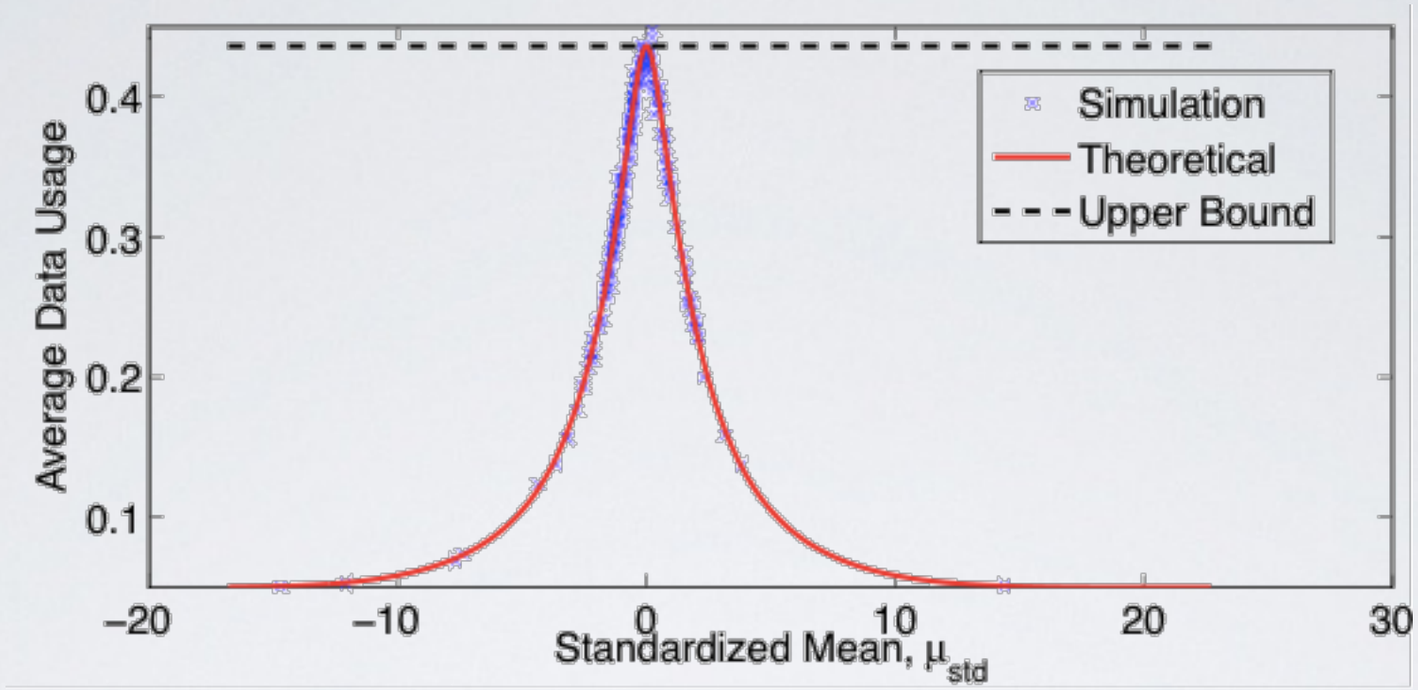


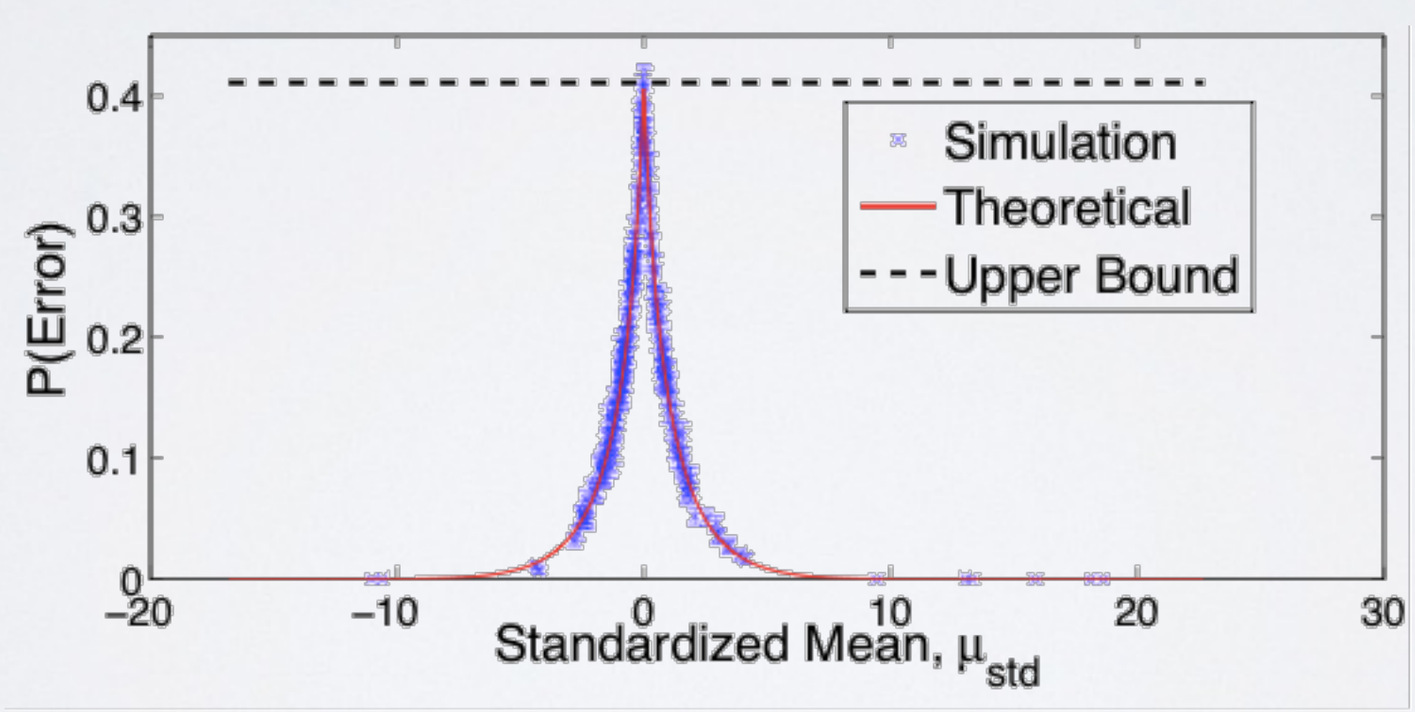
Example: $\mu > \mu_0$



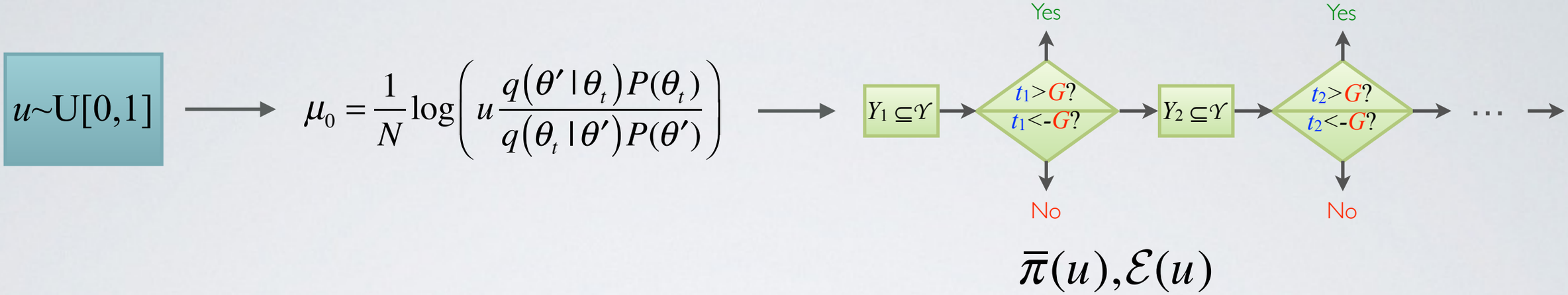
Example: $\mu > \mu_0$



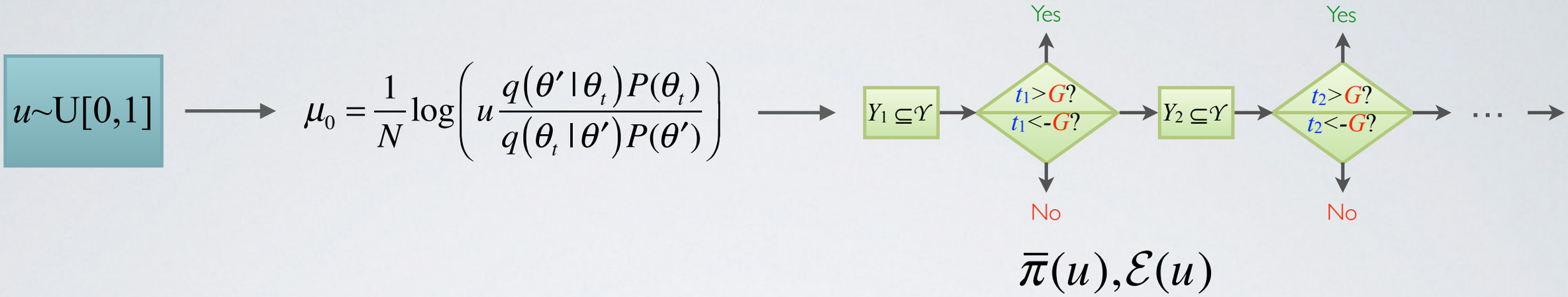

Average Data Usage
0.4
0.3
0.2
0.1
−20
−10
0
10
20
30
Standardized Mean, μstd
Simulation
Theoretical
Upper Bound


P(Error)
0.4
0.3
0.2
0.1
0
−20
−10
0
10
20
30
Standardized Mean, μstd
Simulation
Theoretical
Upper Bound

One Accept/Reject Step of M-H Iteration

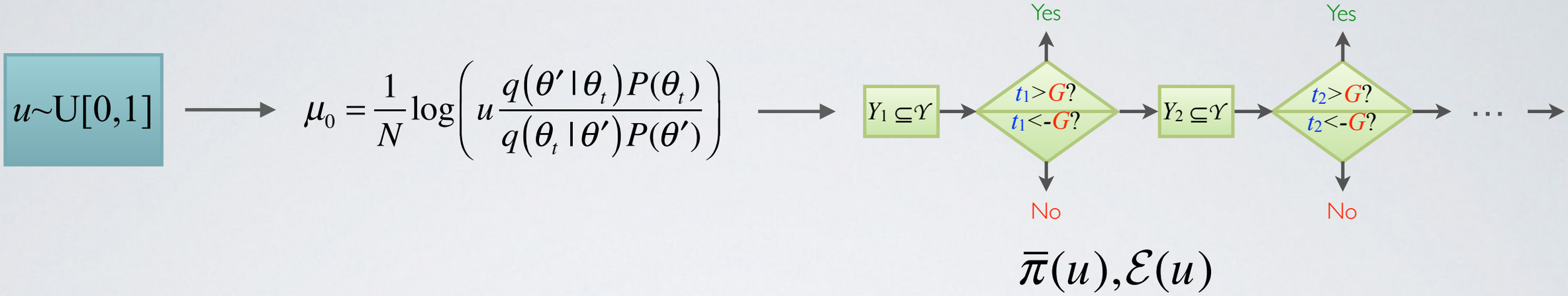


One Accept/Reject Step of M-H Iteration



- Average data usage $\mathbb{E}[\bar{\pi}(u)] = \int_{u=0}^{1} \bar{\pi}(u)du$

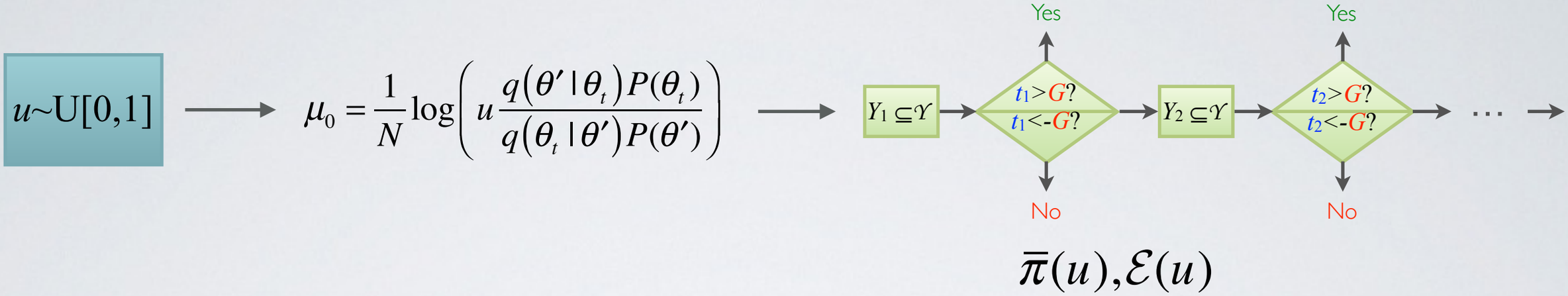
One Accept/Reject Step of M-H Iteration



- Average data usage $\mathbb{E}[\bar{\pi}(u)] = \int_{u=0}^{1} \bar{\pi}(u)\,du$

- Error in the approximate acceptance probability

One Accept/Reject Step of M-H Iteration

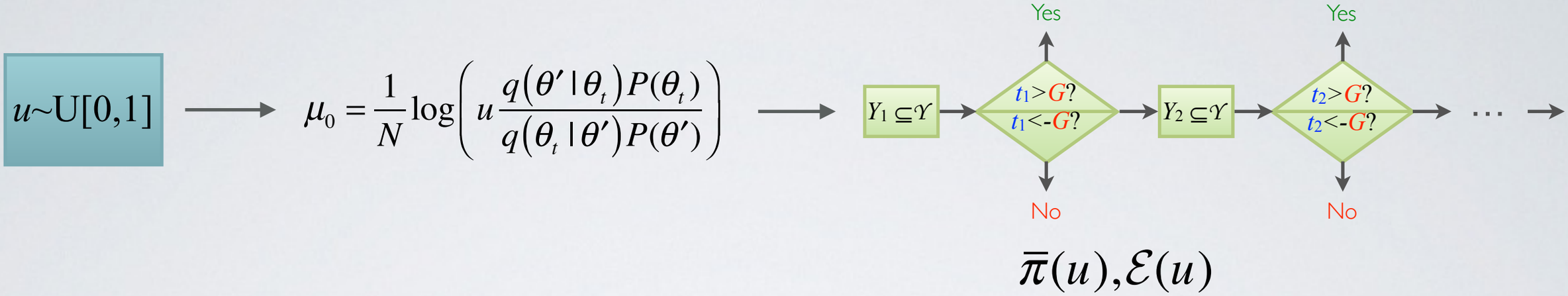


- Average data usage $\mathbb{E}[\bar{\pi}(u)] = \int_{u=0}^{1} \bar{\pi}(u)\,du$

- Error in the approximate acceptance probability

$$\Delta P_a = \left|P_{a,\epsilon} - P_a\right| = \left\|\left[\int_0^{P_a}(1-\mathcal{E}(u))\,du + \int_{P_a}^{1}\mathcal{E}(u)\,du\right] - \left[\int_0^{P_a} 1\,du + \int_{P_a}^{1} 0\,du\right]\right\|$$

One Accept/Reject Step of M-H Iteration



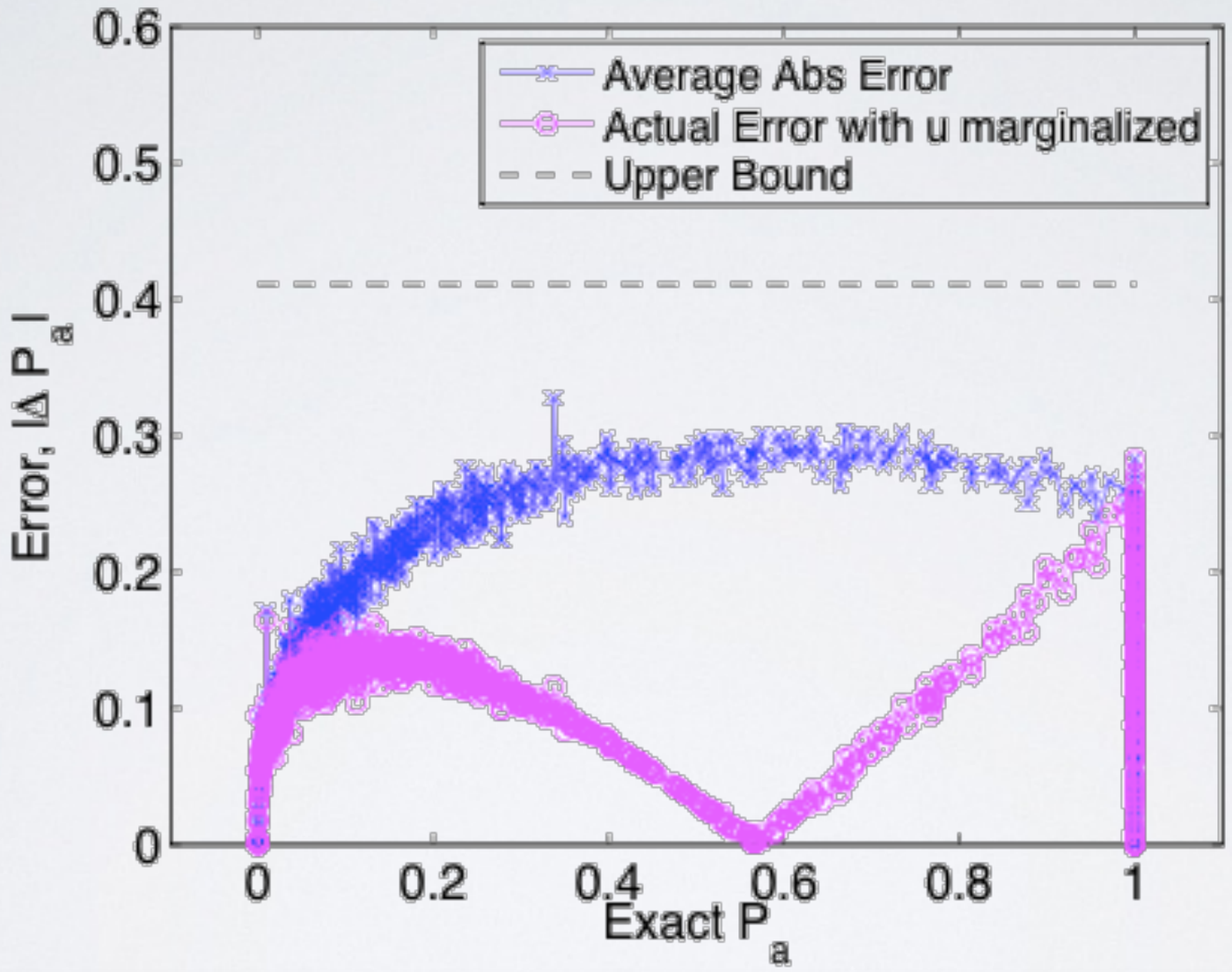
- Average data usage $\mathbb{E}[\bar{\pi}(u)] = \int_{u=0}^{1} \bar{\pi}(u)\,du$

- Error in the approximate acceptance probability

$$\Delta P_a = \left|P_{a,\epsilon} - P_a\right| = \left|\left[\int_0^{P_a}(1-\mathcal{E}(u))\,du + \int_{P_a}^{1}\mathcal{E}(u)\,du\right] - \left[\int_0^{P_a} 1\,du + \int_{P_a}^{1} 0\,du\right]\right|$$

$$= \left|-\int_0^{P_a}\mathcal{E}(u)\,du + \int_{P_a}^{1}\mathcal{E}(u)\,du\right|$$

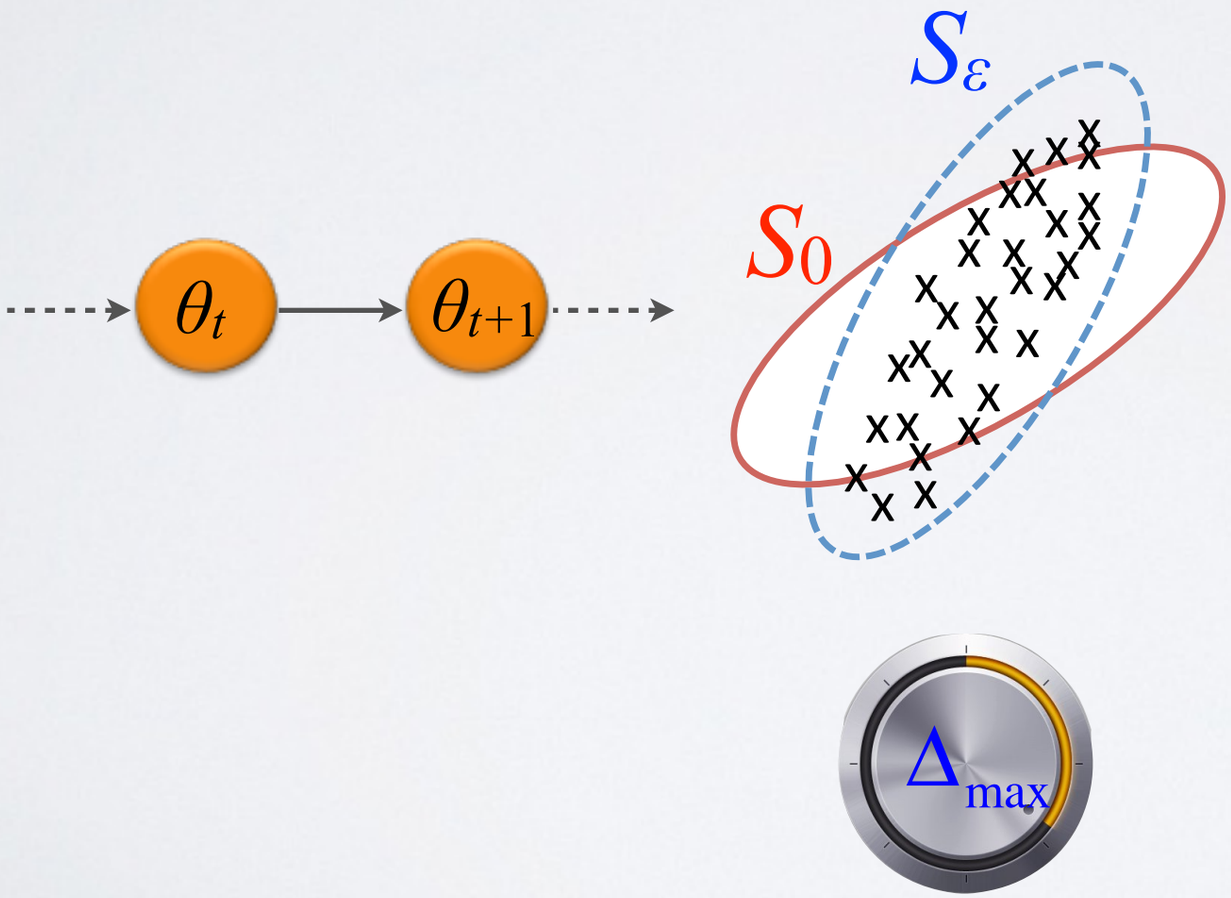
False negative False positive


Average Abs Error
Actual Error with u marginalized
Upper Bound
Error, |Δ P_a|
Exact P_a
0
0.1
0.2
0.3
0.4
0.5
0.6
0
0.2
0.4
0.6
0.8
1

Approximate stationary distribution, S_ε

Theorem 1. If the exact transition kernel $\mathcal{T}_0$ satisfies the contraction condition: $\exists \eta \in [0, 1)$ s.t. $d_v(P\mathcal{T}_0, S_0) \leq \eta d_v(P, S_0)\ \forall P$,
the distance between the posterior distribution S_0 and the stationary distribution of our approximate Markov chain S_ε is upper bounded as:

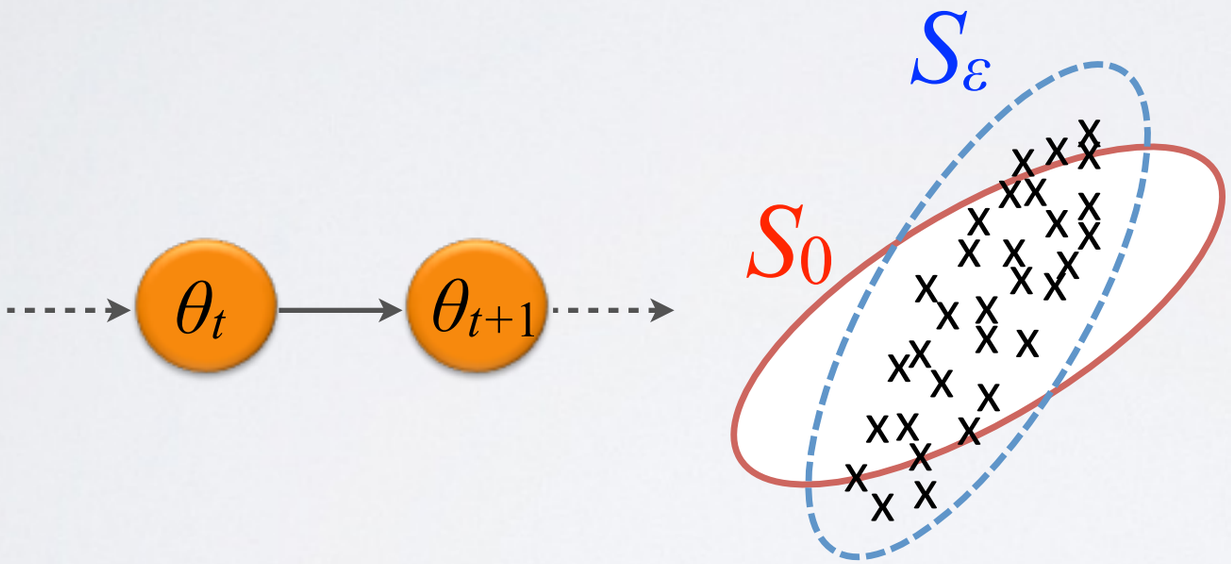
$$d_v\left(S_0, S_\epsilon\right) \leq \frac{\Delta_{\max}}{1-\eta}$$



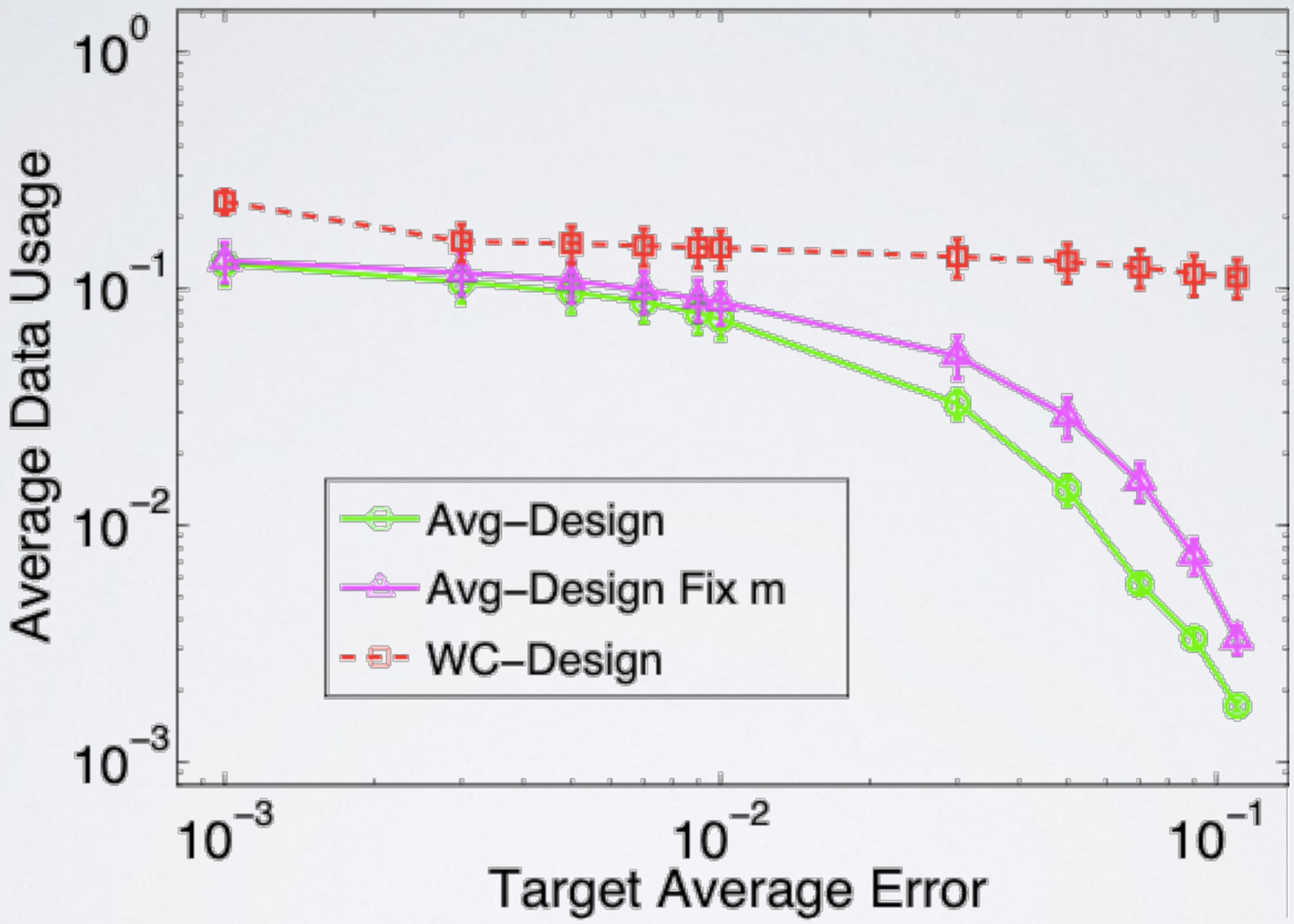
Approximate stationary distribution, S_ε

Theorem 1. If the exact transition kernel $\mathcal{T}_0$ satisfies the contraction condition: $\exists \eta \in [0, 1)$ s.t. $d_v(P\mathcal{T}_0, S_0) \leq \eta d_v(P, S_0)\ \forall P$,
the distance between the posterior distribution S_0 and the stationary distribution of our approximate Markov chain S_ε is upper bounded as:

$$d_v\left(S_0, S_\epsilon\right) \leq \frac{\Delta_{\max}}{1-\eta}$$



$$\approx \frac{1}{M}\sum_{m=1}^{M}\Delta(\theta_m)$$


Average Data Usage
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-3}
10^{-2}
10^{-1}
Target Average Error
Avg–Design
Avg–Design Fix m
WC–Design

Summary

- Proposed to design MCMC algorithms by balancing bias and variance under the big data setting.
- Proposed a scalable approximate M-H algorithm based on sequential hypothesis test on the accept/reject step.
- Analysis on the approximation error and speed gain.

UNIVERSITY OF CAMBRIDGE

Thanks!

Q&A

References

- Ahn, Sungjin, Anoop Korattikara, and Max Welling. "Bayesian posterior sampling via stochastic gradient Fisher scoring." In International Conference on Machine Learning, 2012.
- Bottou, Léon, and Olivier Bousquet. "The Tradeoffs of Large Scale Learning." NIPS. Vol. 4. 2007.
- Hoffman, Matthew D., et al. "Stochastic variational inference." The Journal of Machine Learning Research 14.1 (2013): 1303-1347.
- Singh, Sameer, Wick, Michael, and McCallum, Andrew. Monte Carlo MCMC: efficient inference by approximate sampling. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Process- ing and Computational Natural Language Learning, pp. 1104–1113. Association for Computational Linguistics, 2012.
- Welling, Max, and Yee W. Teh. "Bayesian learning via stochastic gradient Langevin dynamics." Proceedings of the 28th International Conference on Machine Learning. 2011.