Measuring the accuracy of likelihood-free inference

Aden Forrow & Ruth Baker

24 May, 2022

Outline

$$p(\theta|x^*) = \frac{p(x^*|\theta)p(\theta)}{p(x^*)}$$

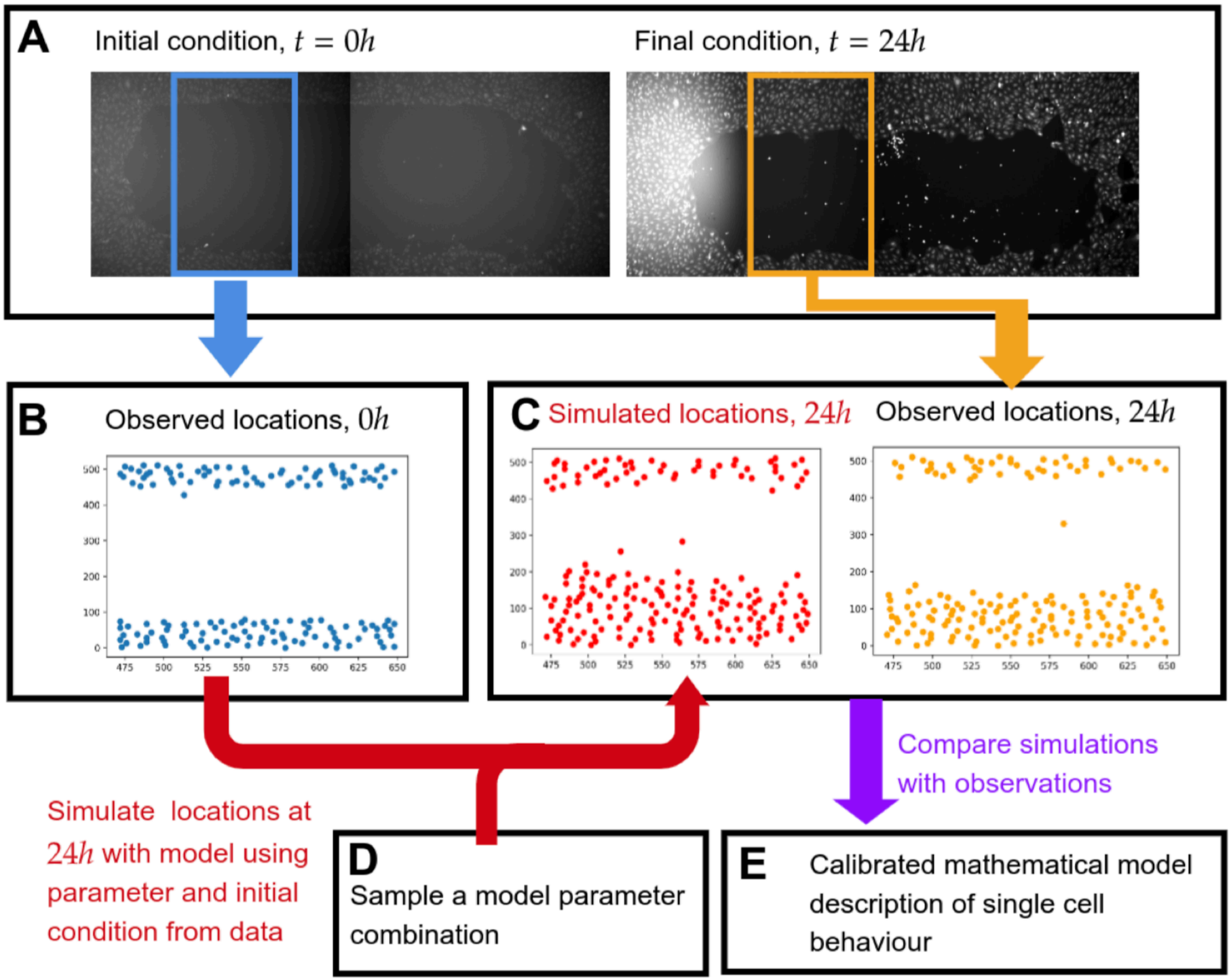
I Likelihood-free inference

- Bayesian inference without likelihoods
- ABC rejection sampling

II How well do algorithms work?

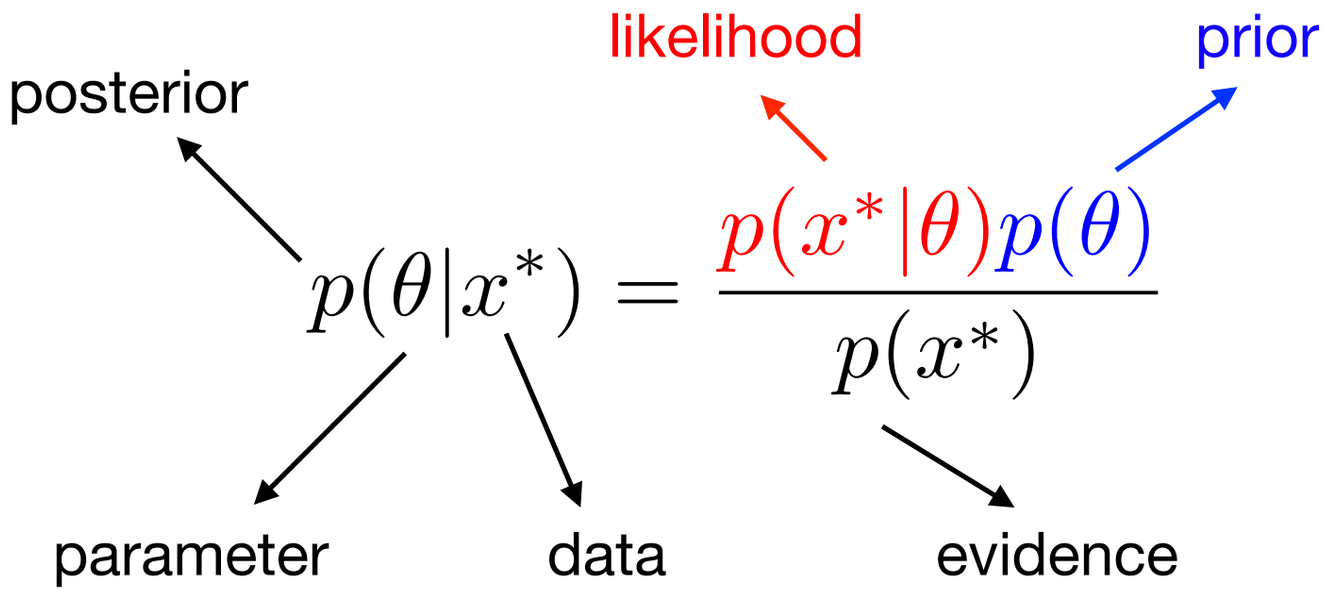
- Performance scores
  - Acceptance rate
  - Effective sample size
  - Error in expectations

III Comparing different scores



Martina-Perez et al., Efficient Bayesian inference for mechanistic modelling with high-throughput data, 2022

Bayesian inference



Bayesian inference

$$p(\theta|x^*) = \frac{p(x^*|\theta)p(\theta)}{p(x^*)}$$

Challenge 1: evidence can rarely be computed

Bayesian inference

$$p(\theta|x^*) = \frac{p(x^*|\theta)p(\theta)}{p(x^*)}$$

Challenge 1: evidence can rarely be computed

Challenge 2: likelihood may be intractable

Bayesian inference

$$p(\theta|x^*) = \frac{p(x^*|\theta)p(\theta)}{p(x^*)}$$

Challenge 1: evidence can rarely be computed

Challenge 2: likelihood may be intractable → this talk

Likelihood-free inference

$$p(\theta|x^*) = \frac{p(x^*|\theta)p(\theta)}{p(x^*)}$$

For many models, we can't compute $p(x|\theta)$,

but we can simulate to draw samples $x \sim p(x|\theta)$.

ABC rejection sampling

“approximate Bayesian computation”

Step 1: sample parameters from prior

$$\theta_i \sim p(\theta)$$

ABC rejection sampling

“approximate Bayesian computation”

Step 1: sample parameters from prior

$$\theta_i \sim p(\theta)$$

Step 2: simulate data from each parameter

$$x_i \sim p(x|\theta_i)$$

ABC rejection sampling

"approximate Bayesian computation"

Step 1: sample parameters from prior

$$\theta_i \sim p(\theta)$$

Step 2: simulate data from each parameter

$$x_i \sim p(x|\theta_i)$$

Step 3: keep parameters where simulation output matched observations

$$\{\theta_i | \Delta(x_i, x^*) < \epsilon\}$$

ABC rejection sampling

Step 1: sample parameters from prior

$$\theta_i \sim p(\theta)$$

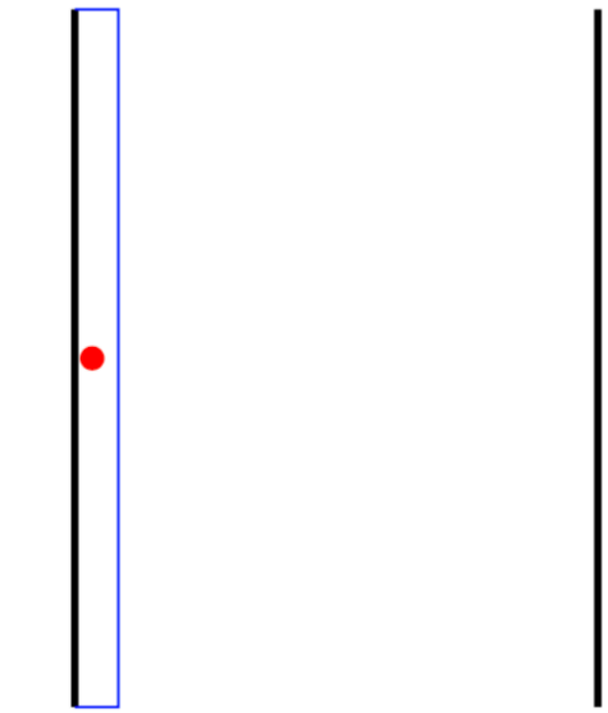
Step 2: simulate data from each parameter

$$x_i \sim p(x|\theta_i)$$

Step 3: keep parameters where simulation output matched observations

$$\{\theta_i | \Delta(x_i, x^*) < \epsilon\}$$

ABC rejection sampling

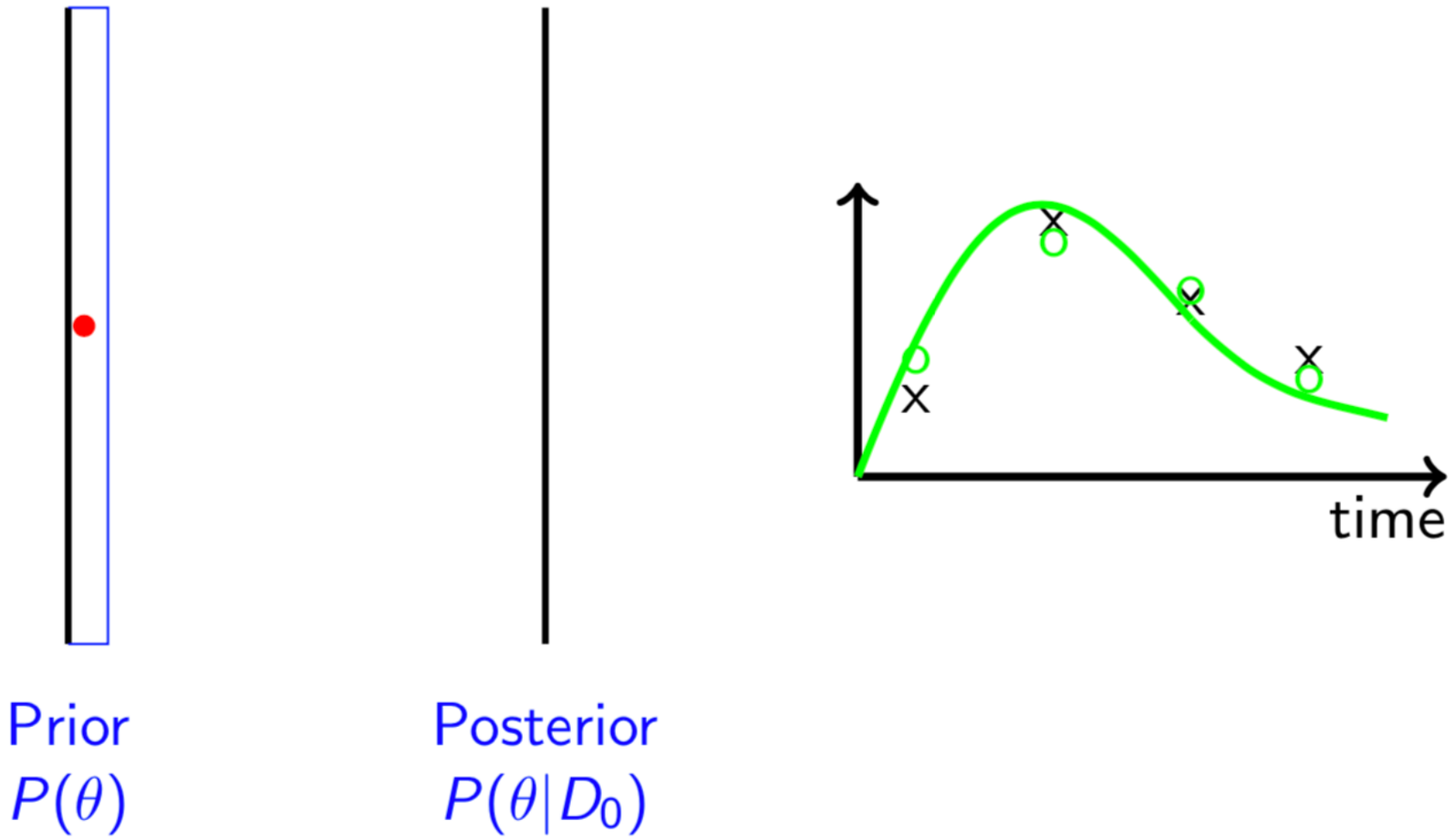


Prior
$P(\theta)$

Posterior
$P(\theta|D_0)$

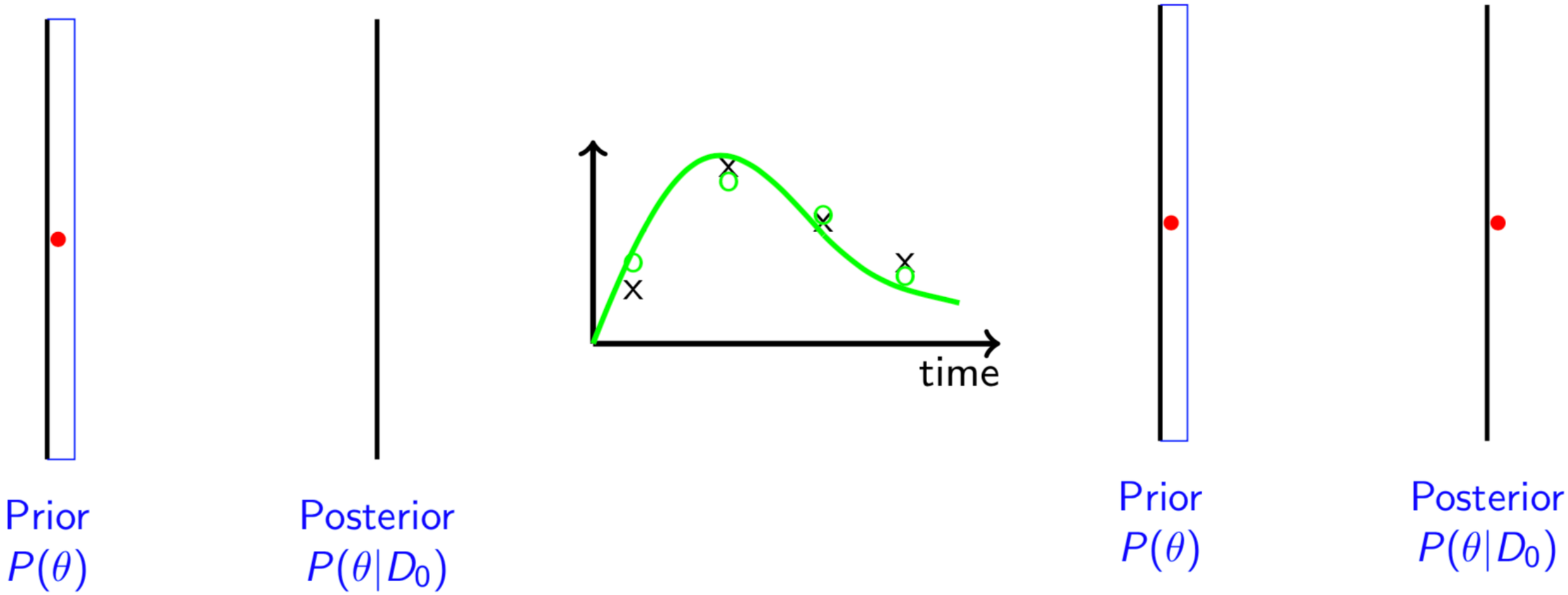
Toni & Stumpf, Tutorial on ABC rejection and ABC SMC for parameter estimation and model selection, 2009

ABC rejection sampling



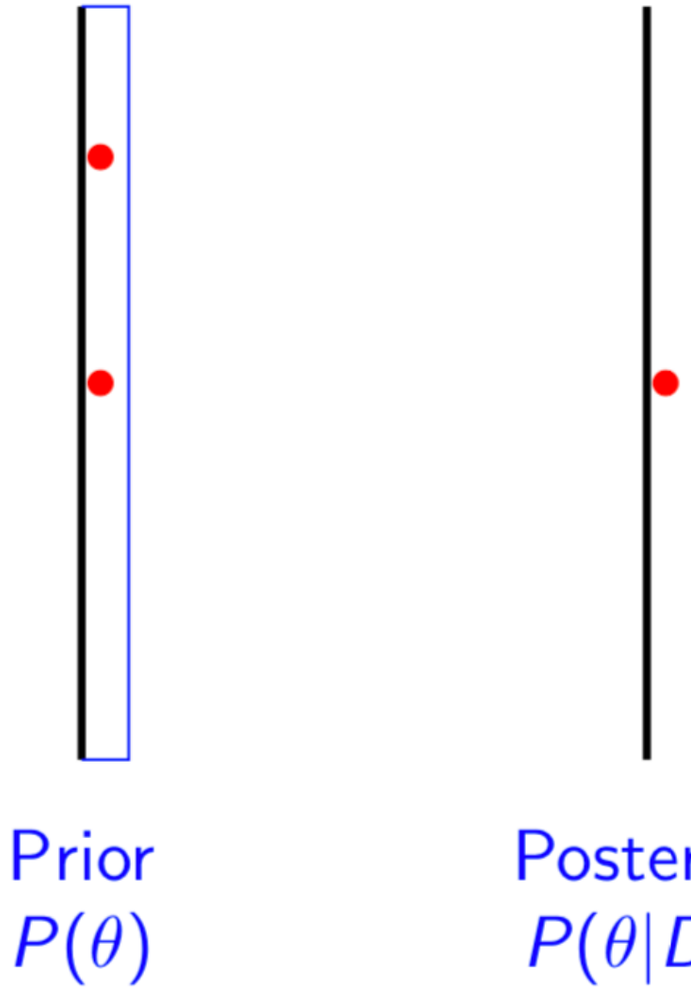
Toni & Stumpf, Tutorial on ABC rejection and ABC SMC for parameter estimation and model selection, 2009

ABC rejection sampling



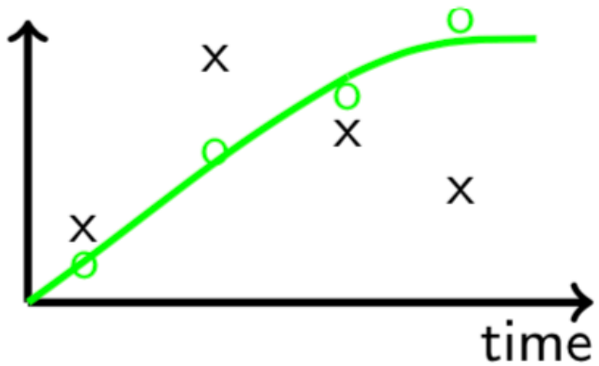
Toni & Stumpf, Tutorial on ABC rejection and ABC SMC for parameter estimation and model selection, 2009

ABC rejection sampling



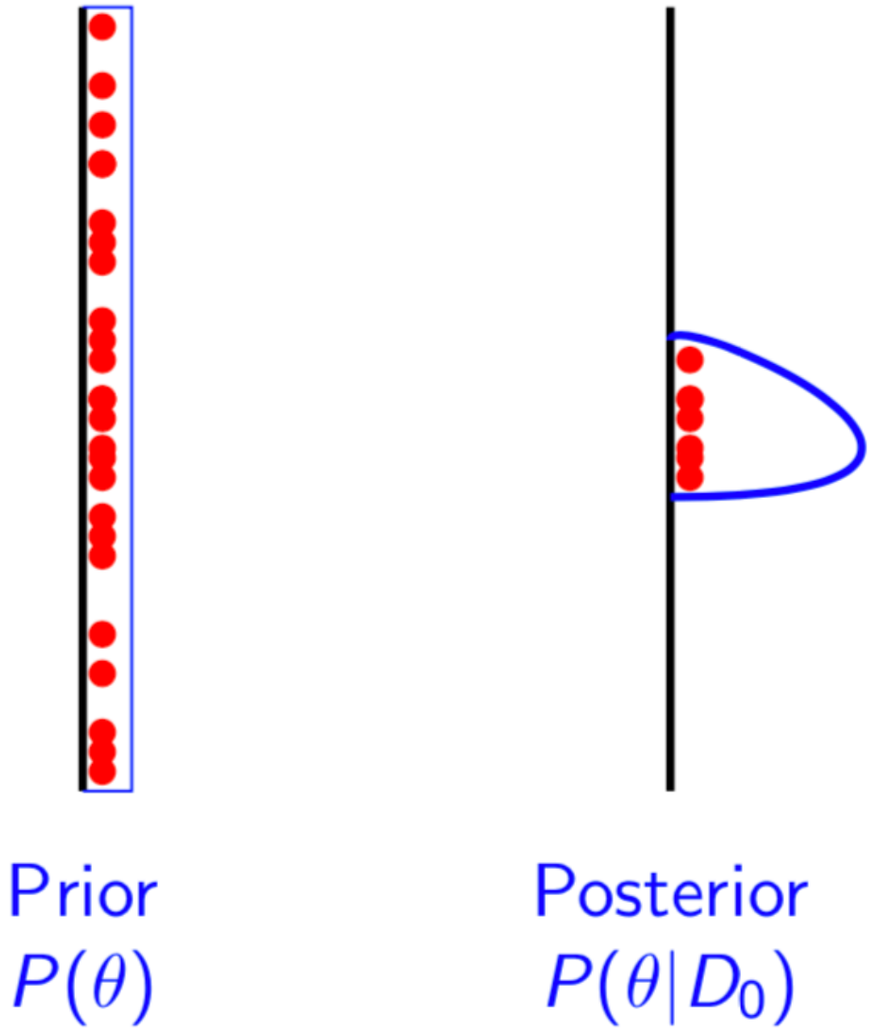
Prior
$P(\theta)$

Posterior
$P(\theta|D_0)$



Toni & Stumpf, Tutorial on ABC rejection and ABC SMC for parameter estimation and model selection, 2009

ABC rejection sampling

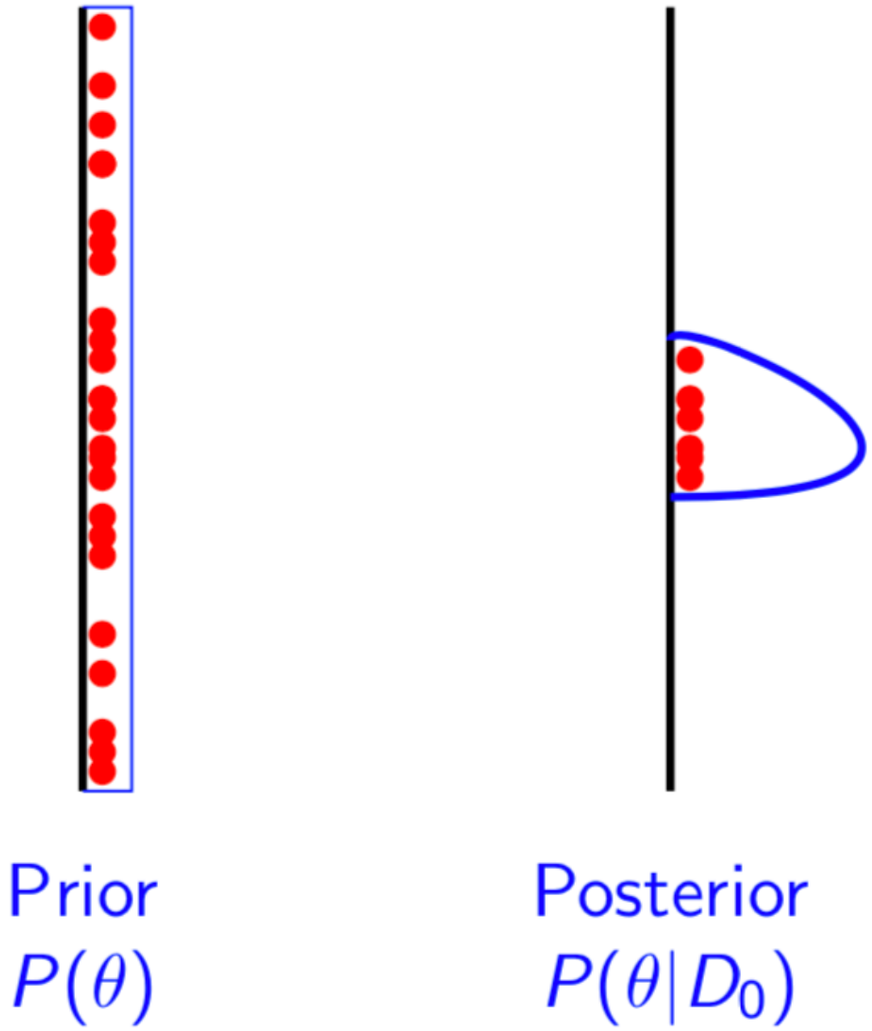


Prior
$P(\theta)$

Posterior
$P(\theta|D_0)$

Toni & Stumpf, Tutorial on ABC rejection and ABC SMC for parameter estimation and model selection, 2009

ABC rejection sampling



Many particles were rejected in the procedure, for which we have spent a lot of computational effort for simulation. ABC rejection is therefore computationally inefficient.

Toni & Stumpf, Tutorial on ABC rejection and ABC SMC for parameter estimation and model selection, 2009

How can we improve efficiency?

How can we improve efficiency?

With more accepted samples, our approximation should be better.

$\Longrightarrow$ *We want an algorithm with higher acceptance rate.*

How can we improve efficiency?

With more accepted samples, our approximation should be better.

$\Longrightarrow$ *We want an algorithm with higher acceptance rate.*

Samples are more likely to be accepted where the likelihood is higher.

$\Longrightarrow$ *Sampling from the posterior should be better than sampling from the prior.*

Candidate scores for algorithms

1. Acceptance rate

$$\frac{N_{acc}}{N}$$

Candidate scores for algorithms

1. Acceptance rate

$$\frac{N_{acc}}{N}$$

2. Effective sample size

$$\widehat{ESS} = \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}$$

What do we want to learn from ABC?

What do we want to learn from ABC?

- The posterior mean? $\mathbb{E}_{p(\theta|x^*)}\left[\theta\right]$

What do we want to learn from ABC?

- The posterior mean? $\mathbb{E}_{p(\theta|x^*)}\left[\theta\right]$
- The posterior variance? $\mathbb{E}_{p(\theta|x^*)}\left[(\theta-\bar{\theta})^2\right]$

What do we want to learn from ABC?

- The posterior mean? $\mathbb{E}_{p(\theta|x^*)}\left[\theta\right]$
- The posterior variance? $\mathbb{E}_{p(\theta|x^*)}\left[(\theta-\bar{\theta})^2\right]$
- The probability of event $\mathcal{A}$? $\mathbb{E}_{p(\theta|x^*)}\left[\mathbf{1}_{\theta\in\mathcal{A}}\right]$

What do we want to learn from ABC?

- The posterior mean? $\mathbb{E}_{p(\theta|x^*)}\left[\theta\right]$
- The posterior variance? $\mathbb{E}_{p(\theta|x^*)}\left[(\theta-\bar{\theta})^2\right]$
- The probability of event $\mathcal{A}$? $\mathbb{E}_{p(\theta|x^*)}\left[\mathbf{1}_{\theta\in\mathcal{A}}\right]$

All of these are function expectations: $\mathbb{E}_{p(\theta|x^*)}[f(\theta)]$

Candidate scores for algorithms

1. Acceptance rate $\frac{N_{acc}}{N}$

2. Effective sample size $\widehat{ESS} = \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}$

3. Error in posterior expectations $\left|\left|\mathbb{E}_{\hat{p}(\theta|x^*)}[f(\theta)] - \mathbb{E}_{p(\theta|x^*)}[f(\theta)]\right|\right|^2$

Candidate scores for algorithms

1. Acceptance rate $\frac{N_{acc}}{N}$

2. Effective sample size $\widehat{ESS} = \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}$

3. Error in posterior expectations $\left|\left|\mathbb{E}_{\hat{p}(\theta|x^*)}[f(\theta)] - \mathbb{E}_{p(\theta|x^*)}[f(\theta)]\right|\right|^2$

4. KL divergence from posterior $KL\left(\hat{p}(\theta|x^*)\middle|\middle|p(\theta|x^*)\right)$

Candidate scores for algorithms

1. Acceptance rate $\frac{N_{acc}}{N}$

2. Effective sample size $\widehat{ESS} = \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}$

3. Error in posterior expectations $\left|\left|\mathbb{E}_{\hat{p}(\theta|x^*)}[f(\theta)] - \mathbb{E}_{p(\theta|x^*)}[f(\theta)]\right|\right|^2$

4. KL divergence from posterior $KL\left(\hat{p}(\theta|x^*)\middle|\middle|p(\theta|x^*)\right)$

Model selection with discrete hypotheses

Suppose we have k hypotheses θ_i. We simulate n_i times from each θ_i.

$$n_i^* \text{ yield } x = x^*$$

Model selection with discrete hypotheses

Suppose we have k hypotheses θ_i. We simulate n_i times from each θ_i.

$$n_i^* \text{ yield } x = x^*$$

$$\hat{p}(x^*|\theta_i) = \frac{n_i^*}{n_i}$$

Model selection with discrete hypotheses

Suppose we have k hypotheses θ_i. We simulate n_i times from each θ_i.

$$n_i^* \text{ yield } x = x^*$$

$$\hat{p}(x^*|\theta_i) = \frac{n_i^*}{n_i}$$

$$\sum_i n_i = N$$

Model selection with discrete hypotheses

Suppose we have k hypotheses θ_i. We simulate n_i times from each θ_i.

$$n_i^* \text{ yield } x = x^*$$

$$\hat{p}(x^*|\theta_i) = \frac{n_i^*}{n_i} \qquad \sum_i n_i = N$$

How do our scores depend on n_i?

Model selection with discrete hypotheses

Suppose we have k hypotheses θ_i. We simulate n_i times from each θ_i.

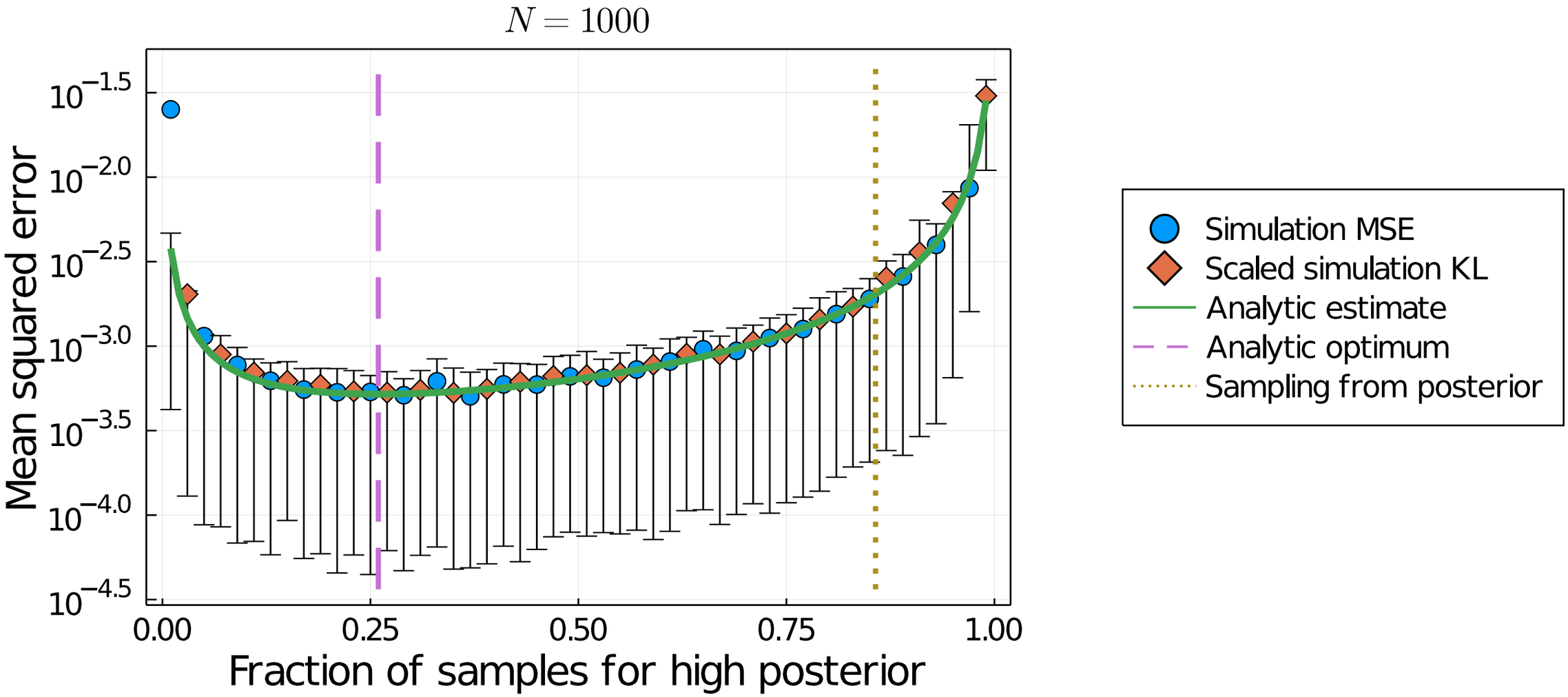
$$n_i^* \text{ yield } x = x^*$$

$$\hat{p}(x^*|\theta_i) = \frac{n_i^*}{n_i} \qquad \sum_i n_i = N$$

How do our scores depend on n_i?

Acceptance rate + ESS: always sample more where likelihood is higher.

Model selection with two hypotheses



Where did intuition go wrong?

Discrete parameter theory

i Variance calculation

ii Comparison of performance scores

Complications in practice

i Adaptive sampling

ii Continuous parameters

Expectation variance for discrete parameters

$$\mathbb{E}_{\hat{p}(\theta|x^*)}[f(\theta)] = \frac{\sum_i f(\theta_i)\hat{p}(x^*|\theta_i)p(\theta_i)}{\sum_i \hat{p}(x^*|\theta_i)p(\theta_i)} \equiv \frac{R}{S}$$

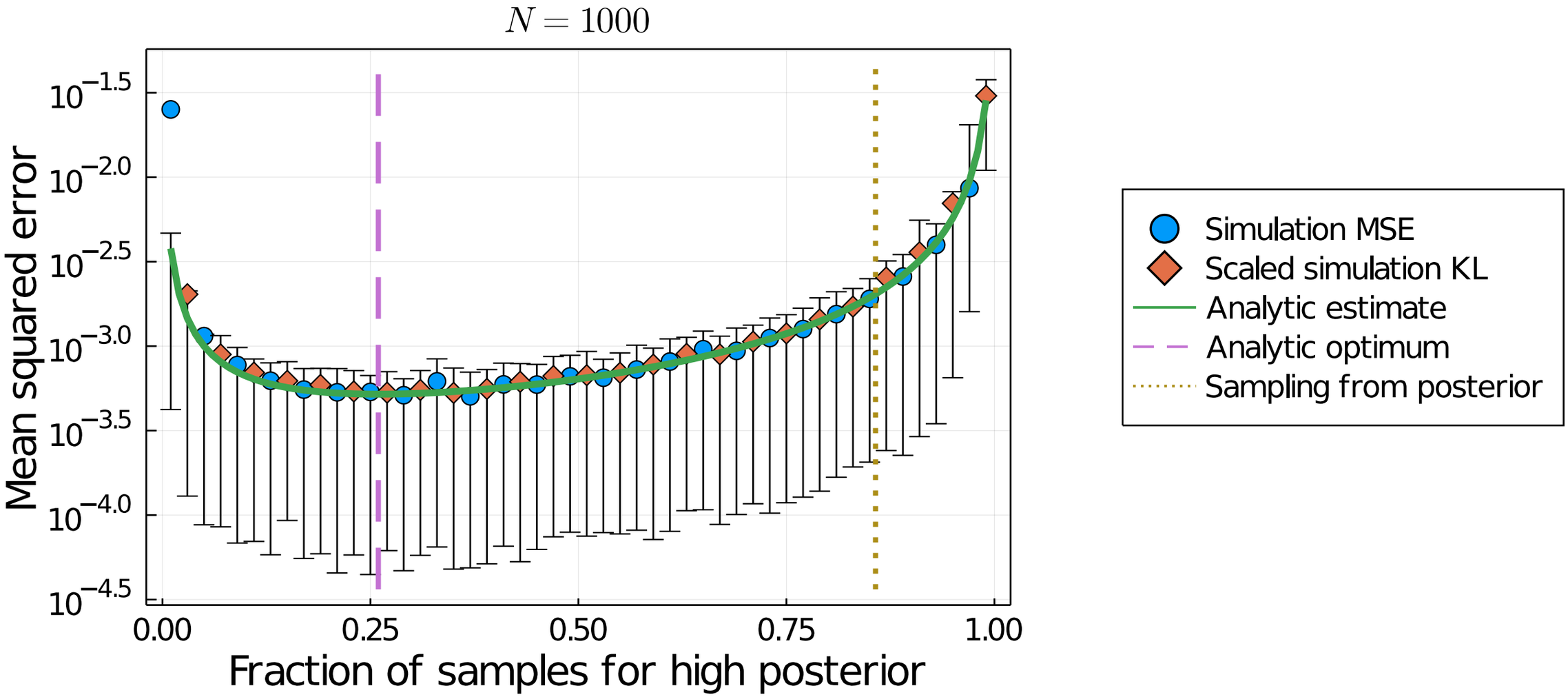
Delta method:

$$\mathrm{var}\left(\frac{R}{S}\right) \approx \frac{1}{\mu_S^2}\mathrm{var}(R) - 2\frac{\mu_R}{\mu_S^3}\mathrm{cov}(R,S) + \frac{\mu_R^2}{\mu_S^4}\mathrm{var}(S)$$

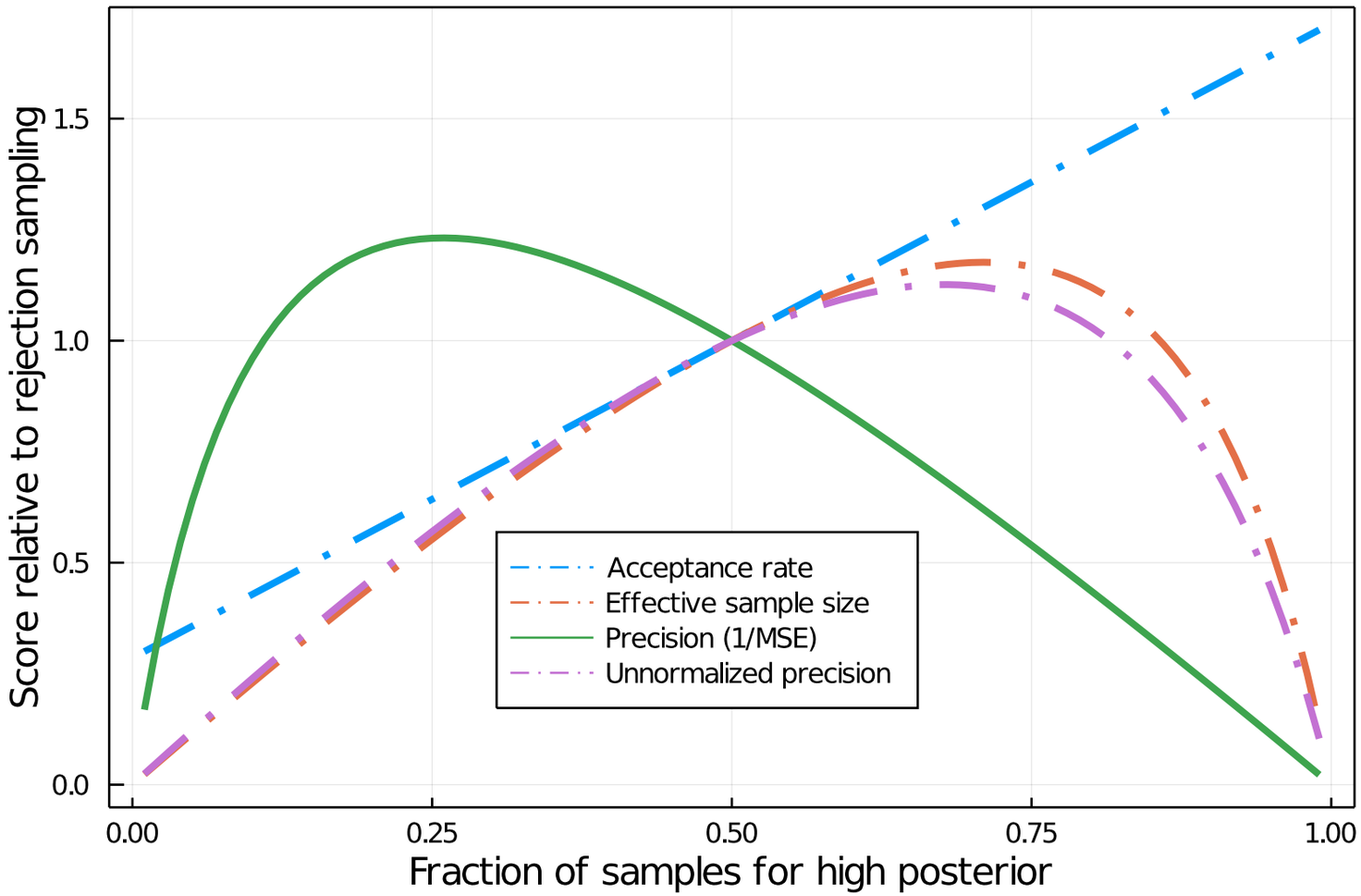
Expectation variance for discrete parameters

$$\mathrm{var}\left(\mathbb{E}_{\hat{p}(\theta|x^*)}\left[f(\theta)\right]\right) \approx p(x^*)^{-2} \sum_i p(\theta_i)^2 \frac{p(x^*|\theta_i)(1-p(x^*|\theta_i))}{n_i} \left(f(\theta_i) - \bar{f}\right)^2$$

Model selection with two hypotheses



Model selection with two hypotheses



Expectation variance for discrete parameters

$$\mathrm{var}\left(\mathbb{E}_{\hat{p}(\theta|x^*)}\left[f(\theta)\right]\right) \approx p(x^*)^{-2} \sum_i p(\theta_i)^2 \frac{p(x^*|\theta_i)(1-p(x^*|\theta_i))}{n_i} \left(f(\theta_i) - \bar{f}\right)^2$$

Expectation variance for discrete parameters

$$\mathrm{var}\left(\mathbb{E}_{\hat{p}(\theta|x^*)}\left[f(\theta)\right]\right) \approx \underline{p(x^*)^{-2}} \sum_i p(\theta_i)^2 \frac{p(x^*|\theta_i)(1-p(x^*|\theta_i))}{n_i} \left(f(\theta_i) - \bar{f}\right)^2$$

Acceptance rate with rejection sampling

Expectation variance for discrete parameters

$$\mathrm{var}\left(\mathbb{E}_{\hat{p}(\theta|x^*)}\left[f(\theta)\right]\right) \approx p(x^*)^{-2} \sum_i p(\theta_i)^2 \frac{p(x^*|\theta_i)(1-p(x^*|\theta_i))}{n_i} \left(f(\theta_i) - \bar{f}\right)^2$$

$$n_i \propto p(\theta_i)(p(x^*|\theta_i)(1-p(x^*|\theta_i))^{1/2} \left| f(\theta_i) - \bar{f} \right|$$

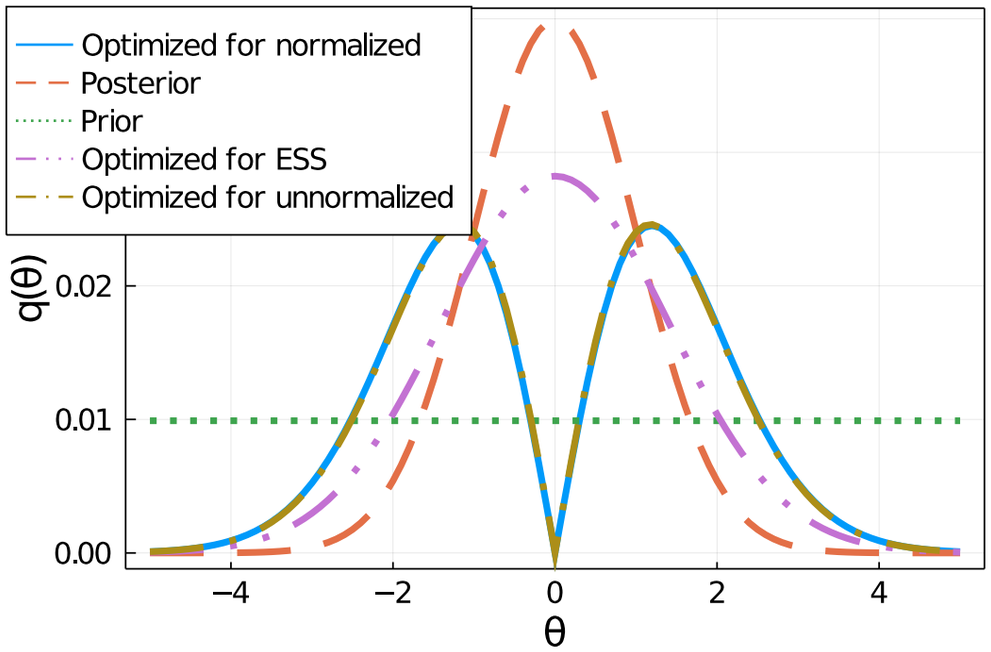
Expectation variance for discrete parameters

$$\mathrm{var}\left(\mathbb{E}_{\hat{p}(\theta|x^*)}\left[f(\theta)\right]\right) \approx p(x^*)^{-2} \sum_i p(\theta_i)^2 \frac{p(x^*|\theta_i)(1-p(x^*|\theta_i))}{n_i} \left(f(\theta_i) - \bar{f}\right)^2$$

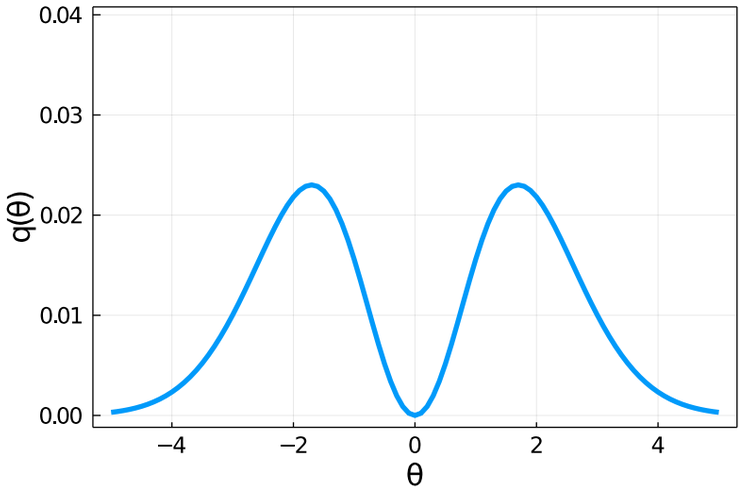
$$n_i \propto p(\theta_i)\left(p(x^*|\theta_i)(1-p(x^*|\theta_i)\right)^{1/2} \left|f(\theta_i) - \bar{f}\right|$$

$$n_1 \propto \left(p(x^*|\theta_2)(1-p(x^*|\theta_1))\right)^{1/2}$$

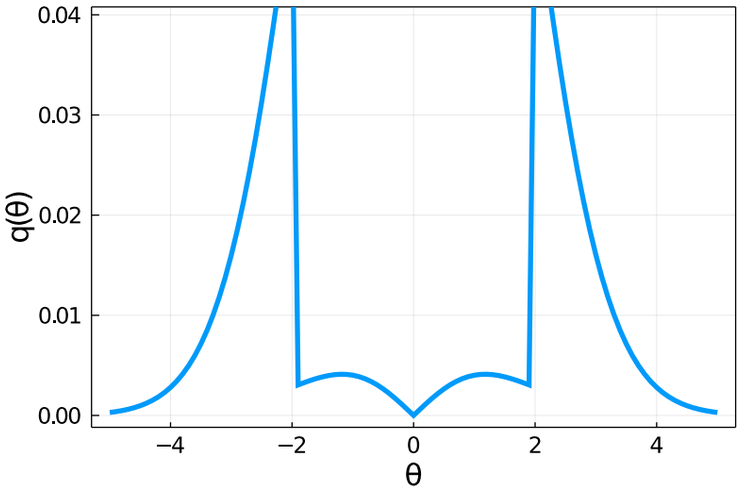
The optimal choice of simulations depends on *f*

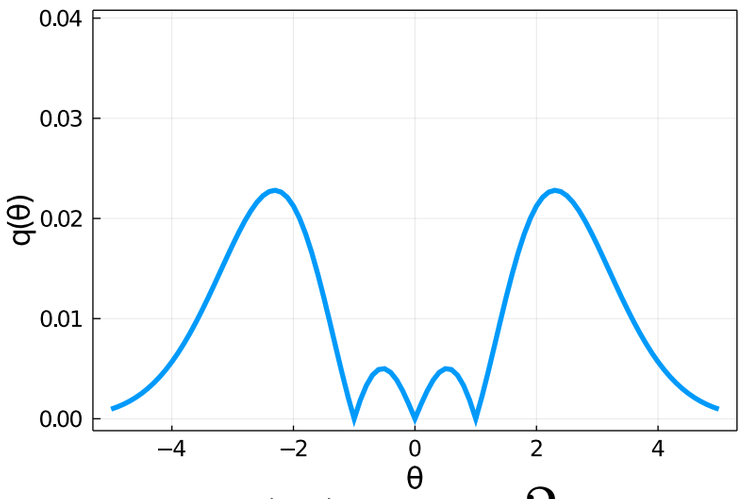


$f(\theta) = \theta$

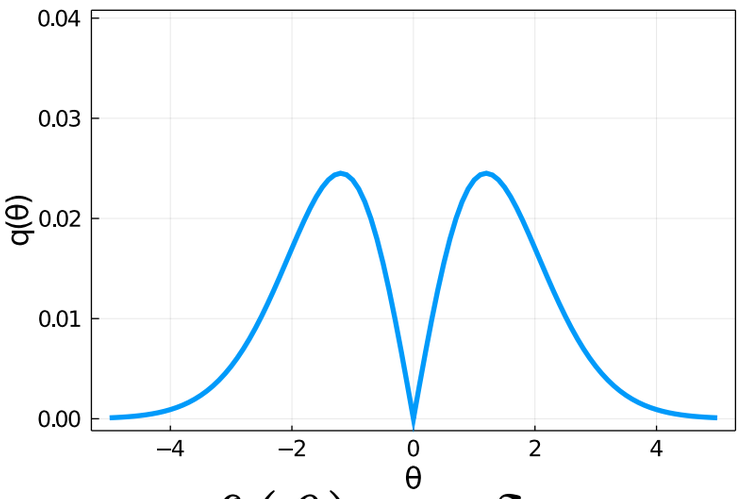


$f(\theta) = \mathbf{1}[|\theta| < 2]$



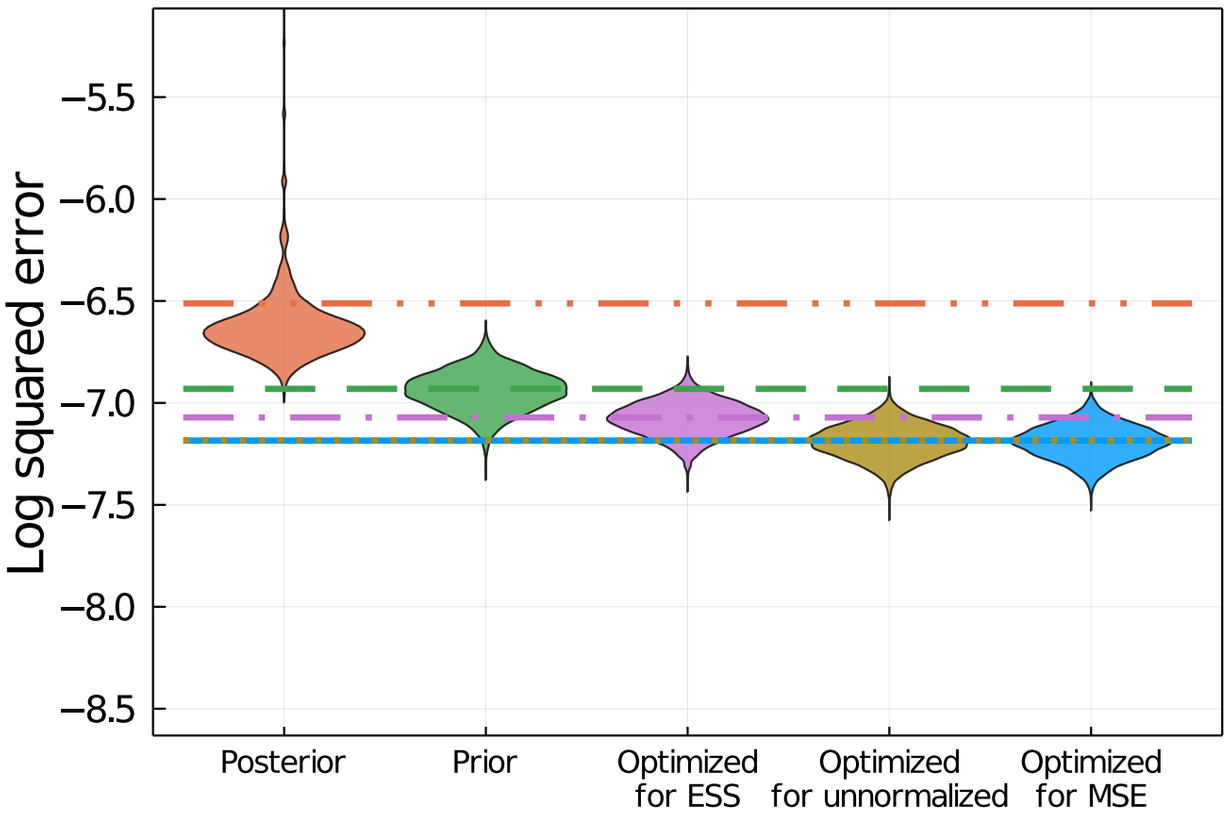


$f(\theta) = \theta^2$

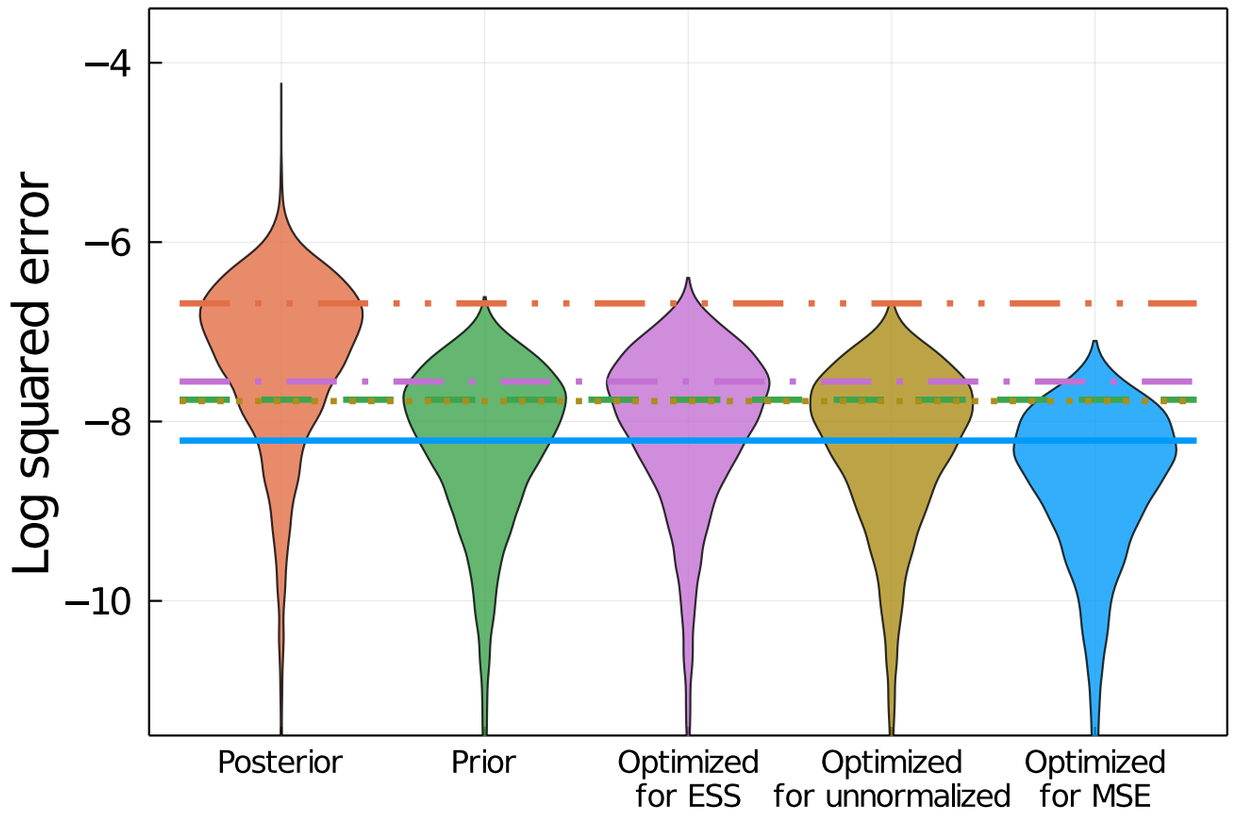


$f(\theta) = \delta_{\theta,\theta'}$

Simulating from the posterior performs poorly



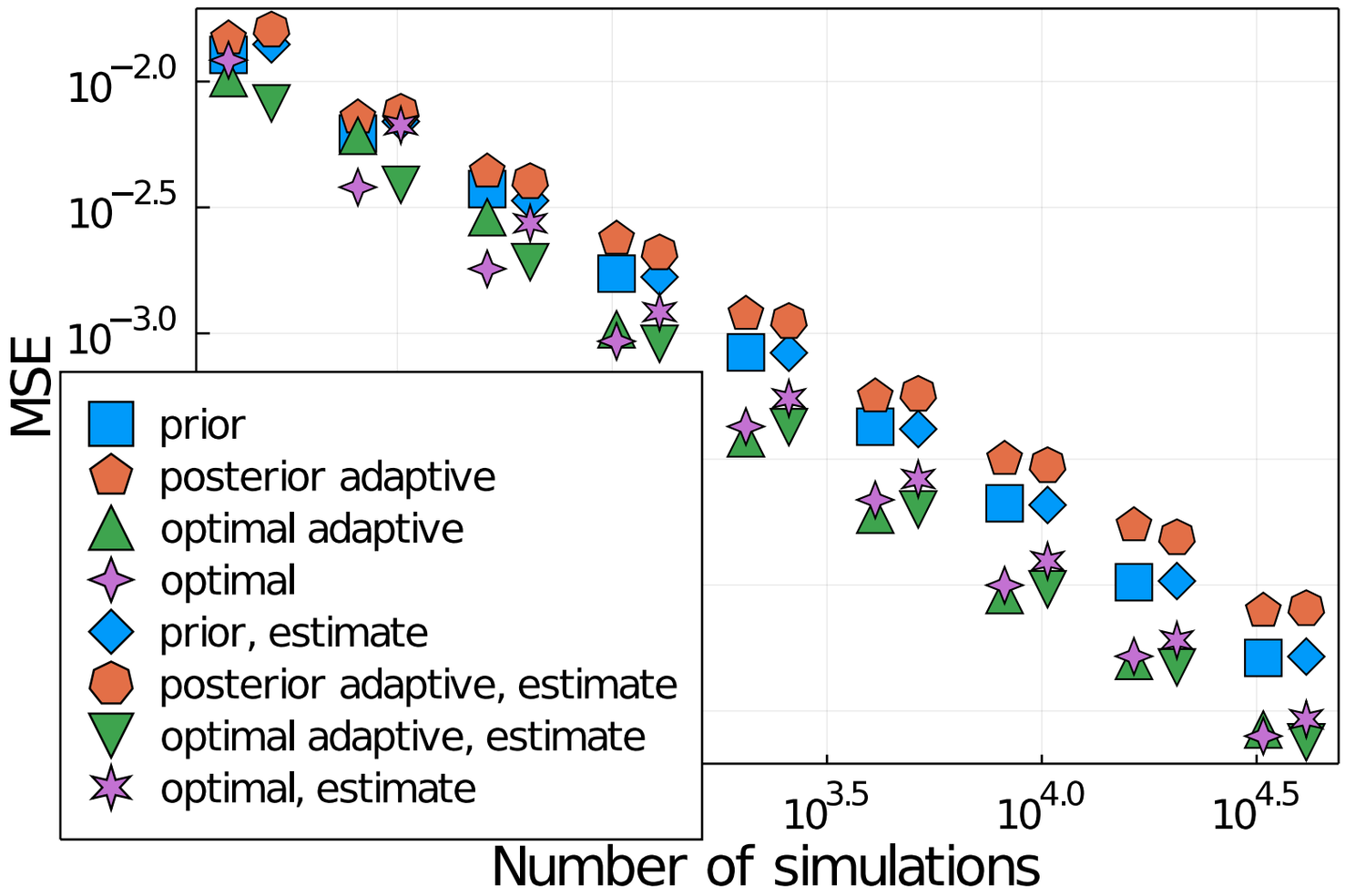
$$f(\theta) = \delta_{\theta,\theta'}$$



$$f(\theta) = \mathbf{1}[|\theta| < 2]$$

Adaptive sampling is straightforward

$$n_i \propto p(\theta_i)(p(x^*|\theta_i)(1-p(x^*|\theta_i))^{1/2}\left|f(\theta_i)-\bar{f}\right|$$



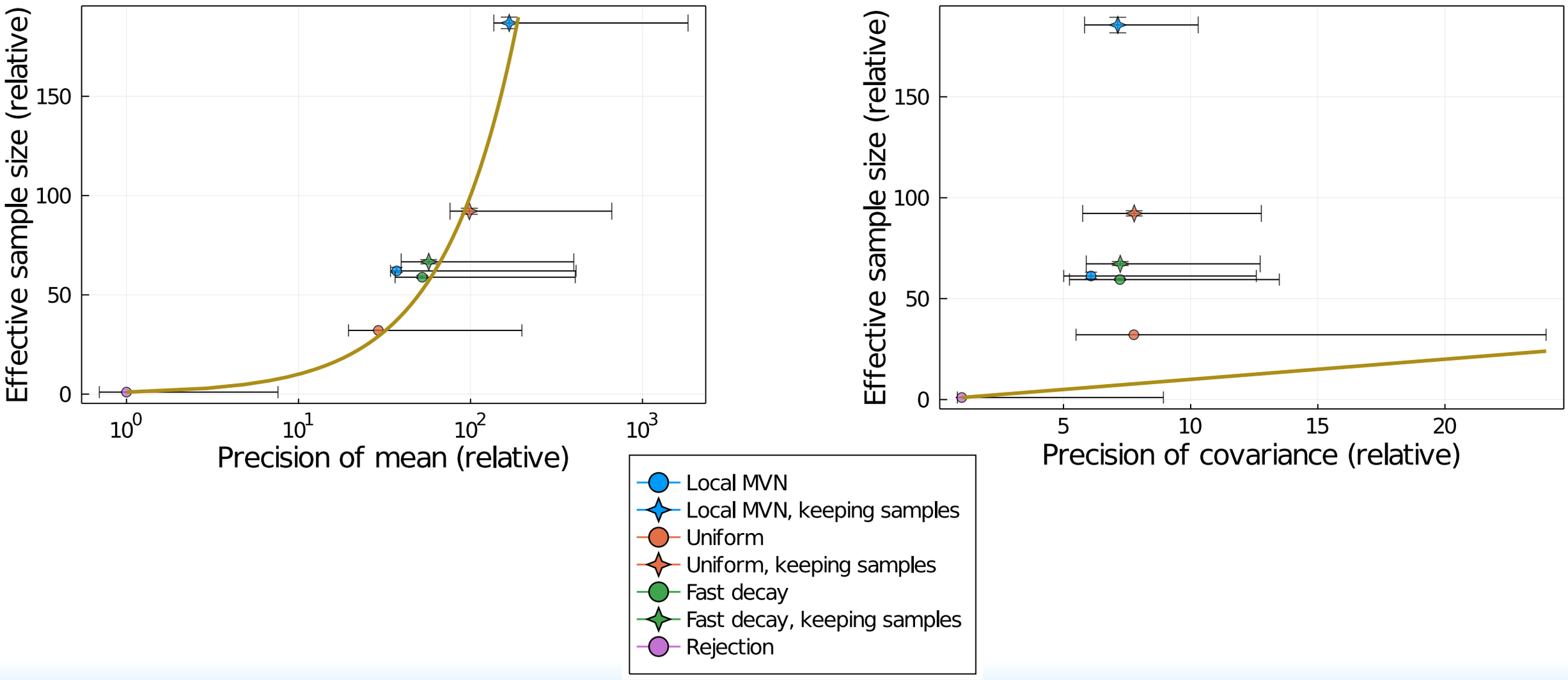
Results qualitatively translate to continuous parameters

Likelihood-free inference algorithms have two choices:

1. Which parameters to simulate from
2. How to approximate posterior from simulations

(2) is trivial for small discrete spaces, but not for continuous cases

Accuracy depends on target



Outline

I Likelihood-free inference

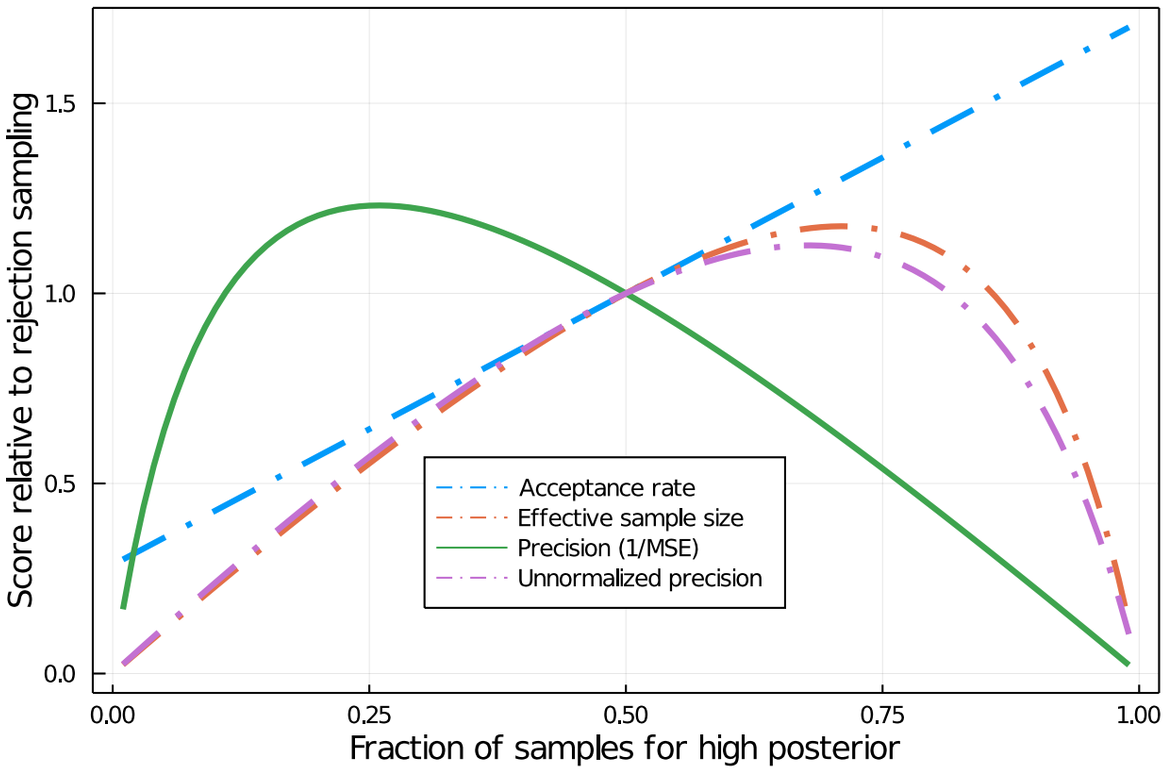
- ABC, ABC rejection

II How well do algorithms work?

- Performance scores

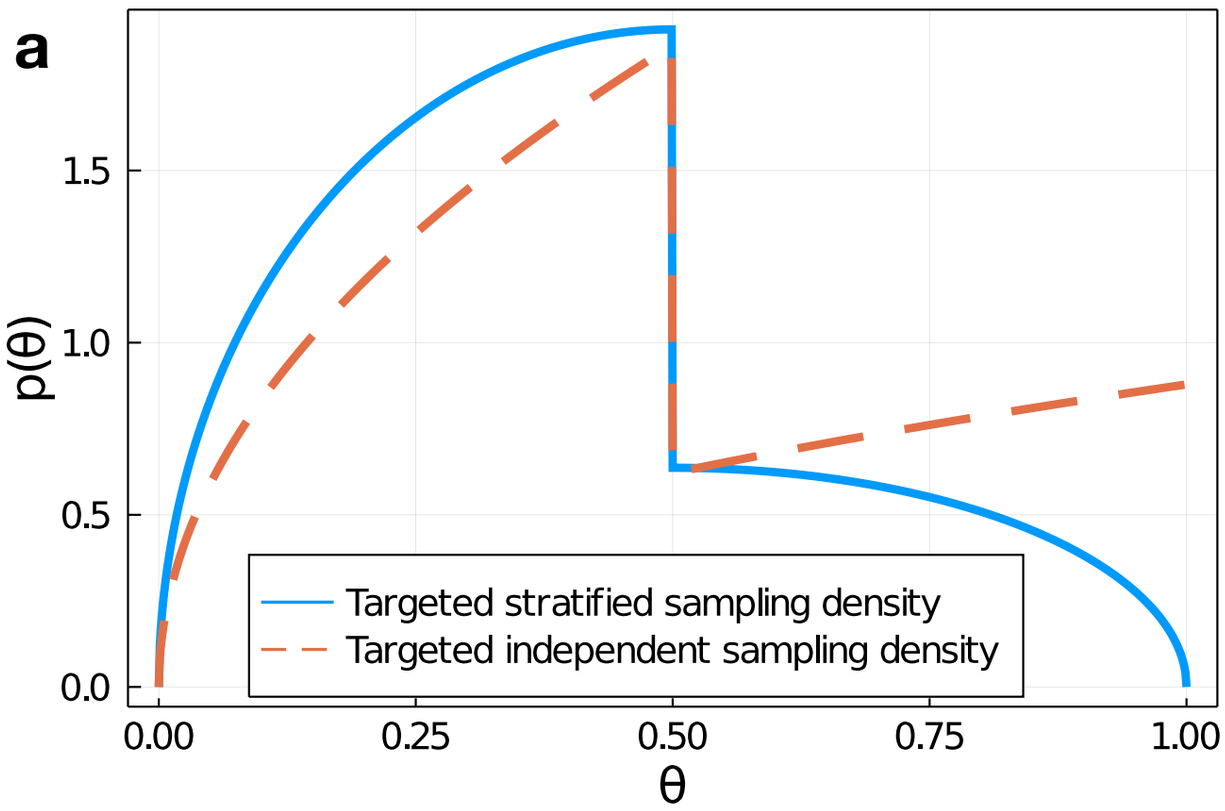
III Comparing different scores

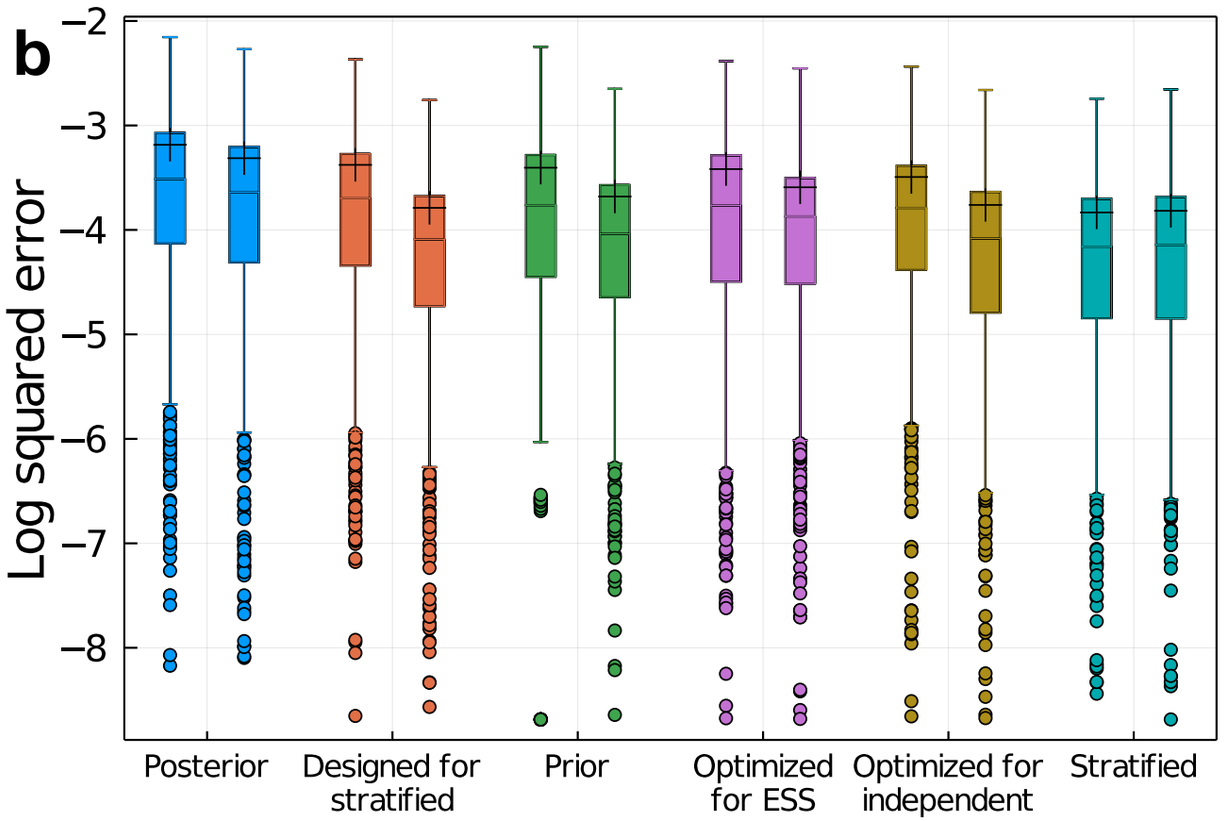
- How you measure performance matters
- ESS, and especially acceptance rate, are flawed heuristics
- If you know what you want to learn from your posterior, you can optimize your algorithm accordingly



Both sampling and estimation matter

$$\theta \in [0, 1] \qquad p(x^*|\theta) = \theta \qquad f(\theta) = \mathbf{1}[|\theta| > 0.5]$$





left: empirical average right: kernel regression

ABC-SMC

“sequential Monte Carlo”: rather than sampling parameters from the prior, sample parameters from an adaptive approximation to the posterior

ABC-SMC

“sequential Monte Carlo”: rather than sampling parameters from the prior, sample parameters from an adaptive approximation to the posterior

Step 1: sample parameters from previous generation

$$\tilde{\theta}_i \sim \{(\theta_j, w_j)\}_{k-1}$$

ABC-SMC

"sequential Monte Carlo": rather than sampling parameters from the prior, sample parameters from an adaptive approximation to the posterior

Step 1: sample parameters from previous generation

$$\tilde{\theta}_i \sim \{(\theta_j, w_j)\}_{k-1}$$

Step 2: perturb parameters, calculate weights w_i

$$\theta_i = \tilde{\theta}_i + \delta_i$$

ABC-SMC

"sequential Monte Carlo": rather than sampling parameters from the prior, sample parameters from an adaptive approximation to the posterior

Step 1: sample parameters from previous generation

$$\tilde{\theta}_i \sim \{(\theta_j, w_j)\}_{k-1}$$

Step 2: perturb parameters, calculate weights w_i

$$\theta_i = \tilde{\theta}_i + \delta_i$$

Step 3: simulate data from each parameter

$$x_i \sim p(x|\theta_i)$$

ABC-SMC

“sequential Monte Carlo”: rather than sampling parameters from the prior, sample parameters from an adaptive approximation to the posterior

Step 1: sample parameters from previous generation

$$\tilde{\theta}_i \sim \{(\theta_j, w_j)\}_{k-1}$$

Step 2: perturb parameters, calculate weights w_i

$$\theta_i = \tilde{\theta}_i + \delta_i$$

Step 3: simulate data from each parameter

$$x_i \sim p(x|\theta_i)$$

Step 4: keep parameters where simulation output matched observations with lower threshold

$$\{(\theta_i, w_i)\}_k = \{(\theta_i, w_i)|\Delta(x_i, x^*) < \epsilon_k\}$$

ABC-SMC

"sequential Monte Carlo": rather than sampling parameters from the prior, sample parameters from an adaptive approximation to the posterior

Step 1: sample parameters from previous generation

$$\tilde{\theta}_i \sim \{(\theta_j, w_j)\}_{k-1}$$

Step 2: perturb parameters, calculate weights w_i

$$\theta_i = \tilde{\theta}_i + \delta_i$$

Step 3: simulate data from each parameter

$$x_i \sim p(x|\theta_i)$$

Step 4: keep parameters where simulation output matched observations with lower threshold

$$\{(\theta_i, w_i)\}_k = \{(\theta_i, w_i)|\Delta(x_i, x^*) < \epsilon_k\}$$

ABC-SMC

"sequential Monte Carlo": rather than sampling parameters from the prior, sample parameters from an adaptive approximation to the posterior

Step 1: sample parameters from previous generation

$$\tilde{\theta}_{k,i} \sim \{(\theta_{k-1,j}, w_{k-1,j})\}_{k-1}$$

Step 2: perturb parameters, calculate weights $w_{k.i}$

$$\theta_{k,i} = \tilde{\theta}_{k,i} + \delta_{k,i}$$

Step 3: simulate data from each parameter

$$x_{k,i} \sim p(x|\theta_{k,i})$$

Step 4: keep parameters where simulation output matched observations with lower threshold

$$\{(\theta_{k,i}, w_{k,i})\}_k = \{(\theta_{k,i}, w_{k,i})|\Delta(x_{k,i}, x^*) < \epsilon_k\}$$

ABC-SMC

The standard final output is the weighted parameters from the last round:

$$\{(\theta_{K,i}, w_{K,i})\}_K = \{(\theta_{K,i}, w_{K,i}) | \Delta(x_{K,i}, x^*) < \epsilon_K\}$$

ABC-SMC

The standard final output is the weighted parameters from the last round:

$$\{(\theta_{K,i}, w_{K,i})\}_K = \{(\theta_{K,i}, w_{K,i}) | \Delta(x_{K,i}, x^*) < \epsilon_K\}$$

We could instead take parameters from any round:

$$\{(\theta_i, w_i)\}_{all} = \{(\theta_{k,i}, \alpha_k w_{k,i}) | \Delta(x_{k,i}, x^*) < \epsilon_K,\ 1 \leq k \leq K\}$$

ESS is moderately informative about accuracy