Marginal MCMC with Noisy Local Approximations

[1]**Andrew D. Davis**

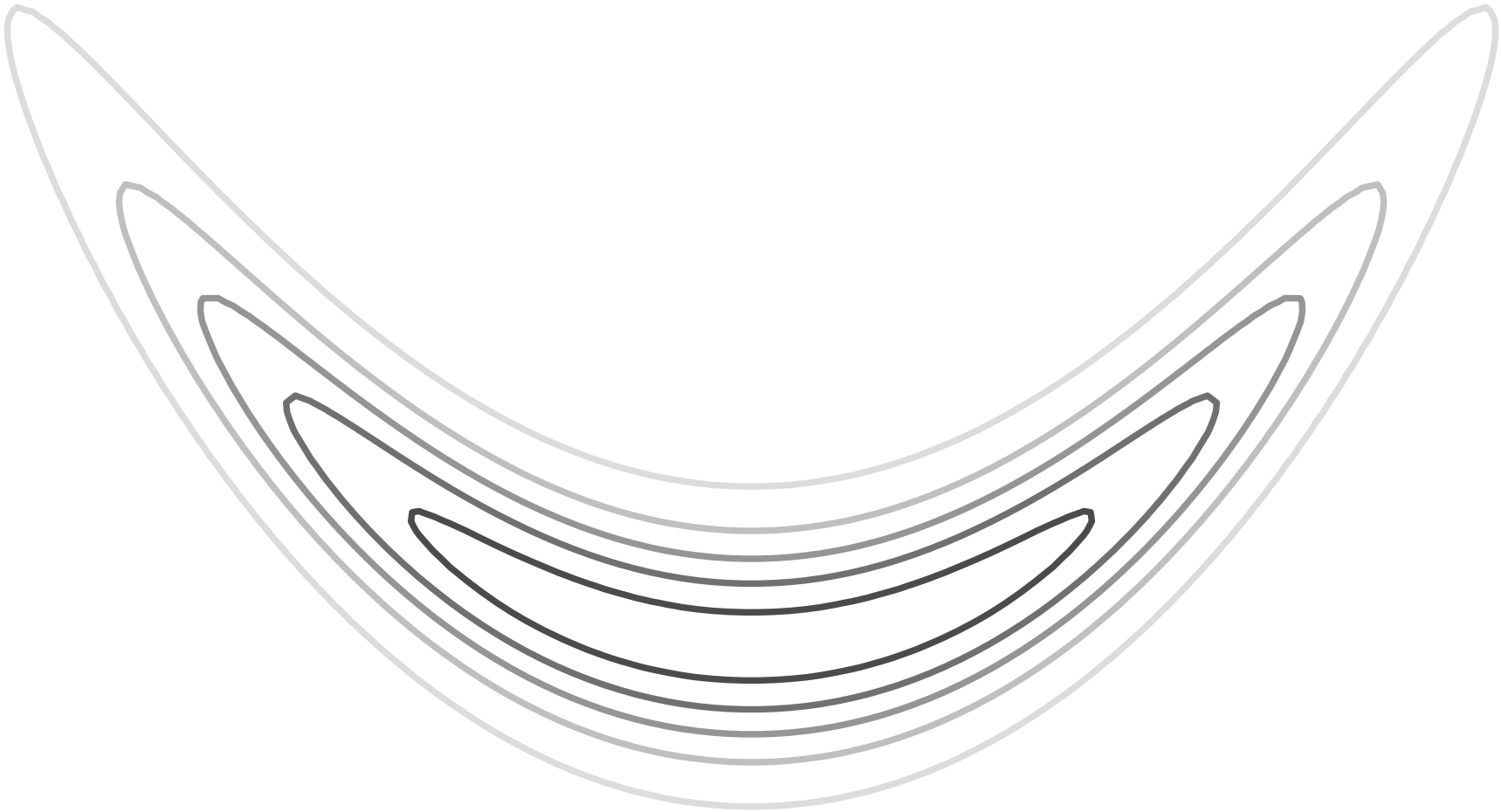
[2]Youssef Marzouk, [3]Aaron Smith, [4]Natesh Pillai

[1]Cold Region Research and Engineering Labs (CRREL)
[2]Massachusetts Institute of Technology
[3]University of Ottawa
[4]Harvard University

February 27, 2019

Problem statement

We want to **characterize the distribution π** using MCMC

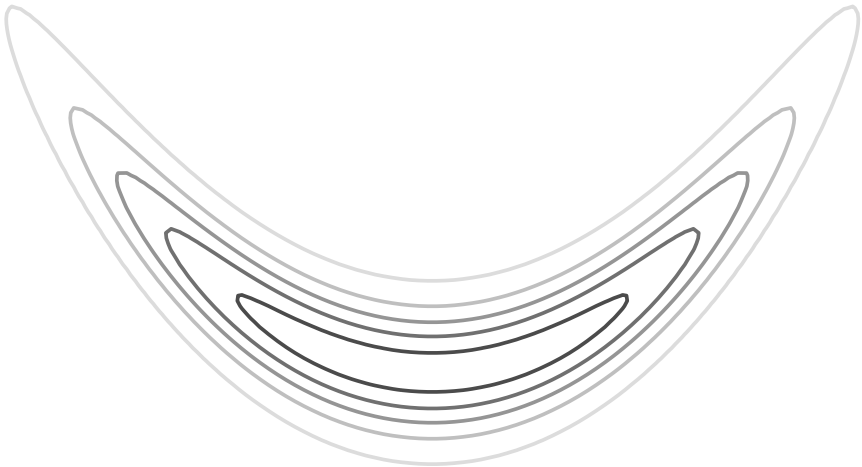


Problem statement

We want to **characterize the distribution π** using MCMC

Two problems:

1. The probability density function $\pi(x)$ is **computationally expensive**

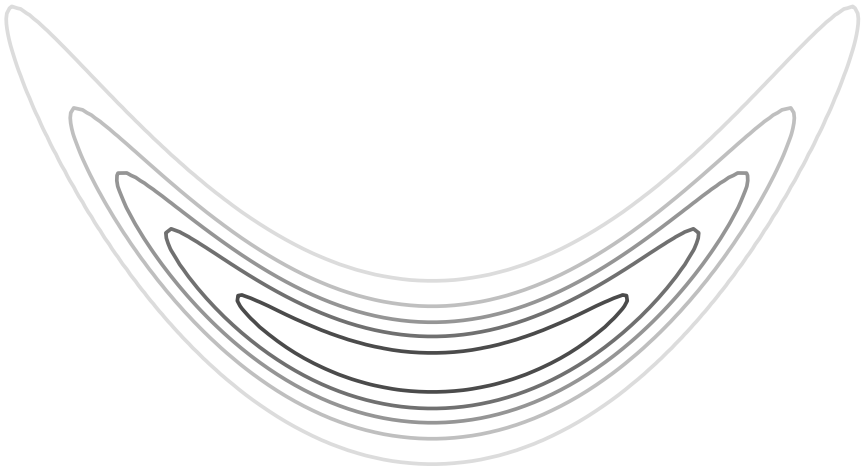


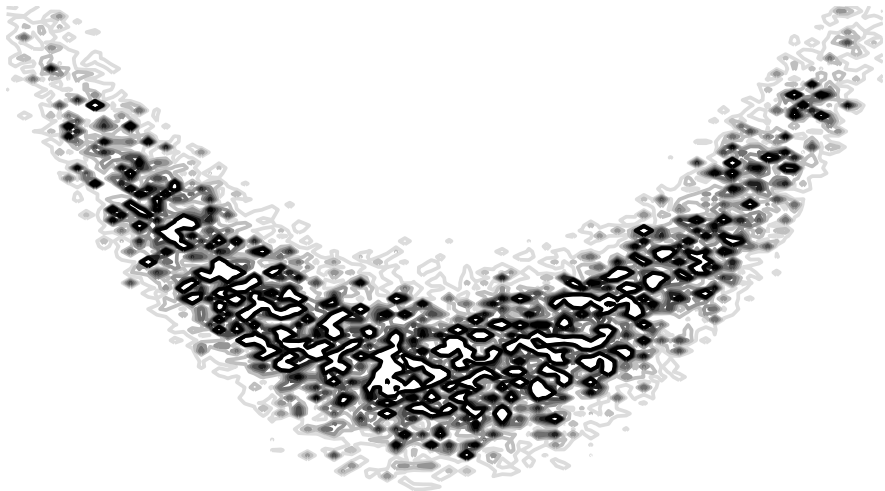
Problem statement

We want to **characterize the distribution π** using MCMC

Two problems:

1. The probability density function $\pi(x)$ is **computationally expensive**
2. Only **noisy** density evaluations are available $\hat{\pi}(x) \approx \pi(x)$





Bayesian inference

- Quantity of interest X (random variable)
- Data Y (random variable)

Bayes' theorem updates our degree of certainty given:

1. Data/observations
2. Physical models

Bayesian inference

- Quantity of interest X (random variable)
- Data Y (random variable)

Bayes' theorem updates our degree of certainty given:

1. Data/observations
2. Physical models

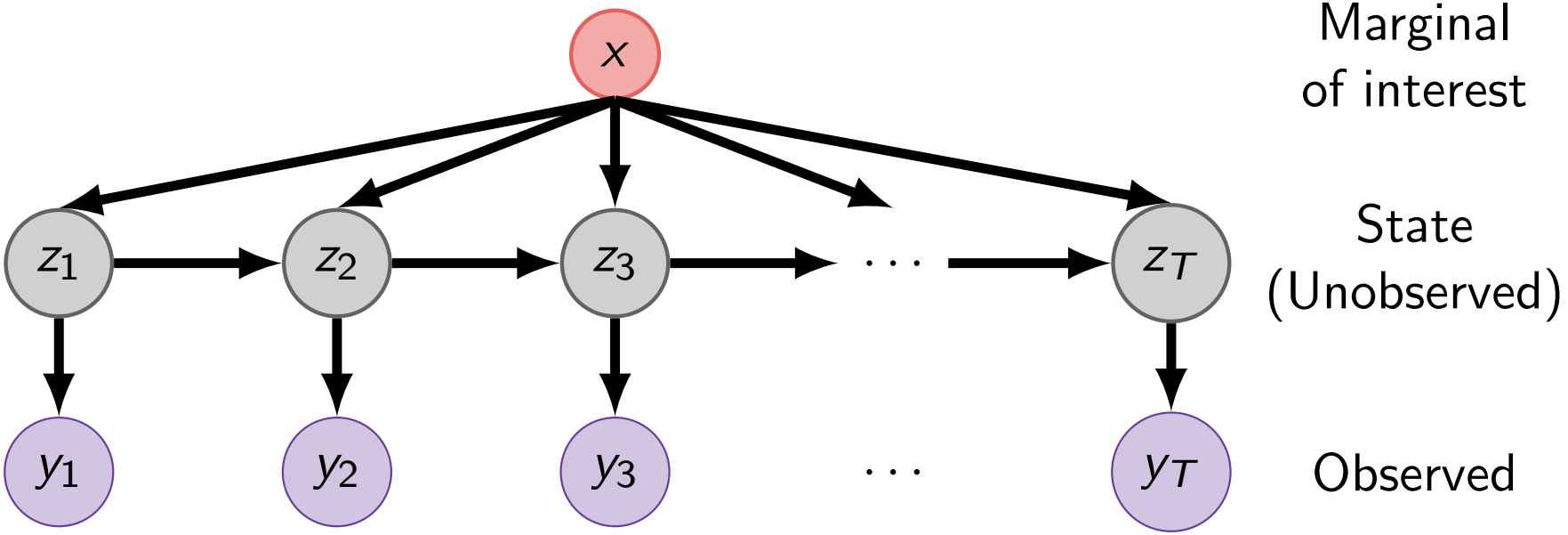
- **Prior** distribution: π_X models our degree of certainty about X
- **Likelihood** distribution: $\pi_{Y|X}$ quantifies model-data mismatch
- **Posterior** distribution: $\pi_{X|Y}$ updates the prior given Y
- **Bayes' rule** relates the densities using the law of total probability

$$\pi(x|y) \propto \pi(x)\pi(y|x)$$

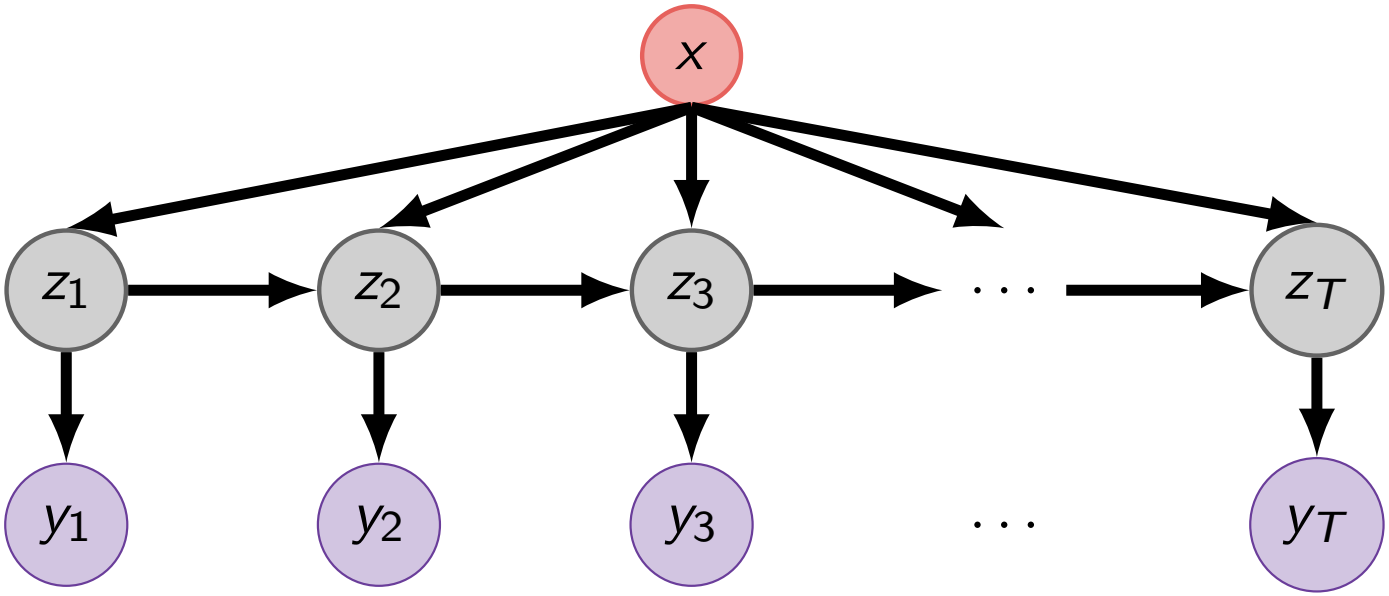
Motivating problem: state space modeling

Static parameter estimation—

- Given: data $y_{1:T}$ with an unobservable state $z_{1:T}$
- We want to characterize the distribution over static parameters $x|y_{1:T}$
  - x is low-dimensional



Marginalize state variables



Marginal of interest

State (Unobserved)

Observed

Marginalizing avoids characterizing the joint density over $[x, z_{1:T}]$

$$\underbrace{\pi(x|y_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(x)}_{\text{Prior}} \underbrace{\int \pi(z_{1:T}, y_{1:T}|x)\, dz}_{\text{Likelihood}}$$

Partitioned space

Partition parameter space:

1. Coordinates characterized by MCMC (x)
2. Coordinates to be "marginalized away" ($z_{1:T}$)

- Model evaluations are (even more) computationally expensive
- We only have noisy estimates of the likelihood

$$\hat{\pi}(y_{1:T}|x) \approx \int \pi(z_{1:T}, y_{1:T}|x)\, dz$$

Previous work

- Pseudo-marginal MCMC characterizes marginal distributions
  Beaumont, 2010, Andrieu and Roberts, 2009

- Exploit the target density's regularity to build a surrogate model
  - Polynomial approximations Marzouk et al., 2007, Marzouk and Xiu, 2009
  - Gaussian processes Rasmussen, 2006, Bernardo et al., 2008, Sacks et al., 1989, Santner et al., 2003

- Continual surrogate refinement asymptotically guarantees we characterize the true target distribution Conrad et al., 2016

Our contribution

- Trigger surrogate refinement using a specialized error indicator for local polynomial approximations
- Derive an ideal refinement rate using a surrogate-bias and MCMC-variance trade-off
- Build the surrogate model from noisy target density evaluations

Local polynomial surrogates

- *Given:* n (potentially noisy) density evaluations at points $\{x^{(i)}\}_{i=1}^n$
- **Minimize the least squares error** between $\pi(x)$ and a degree p polynomial $\breve{\pi}(x)$

  Conn, 2009, Stone, 1977, Kohler, 2002

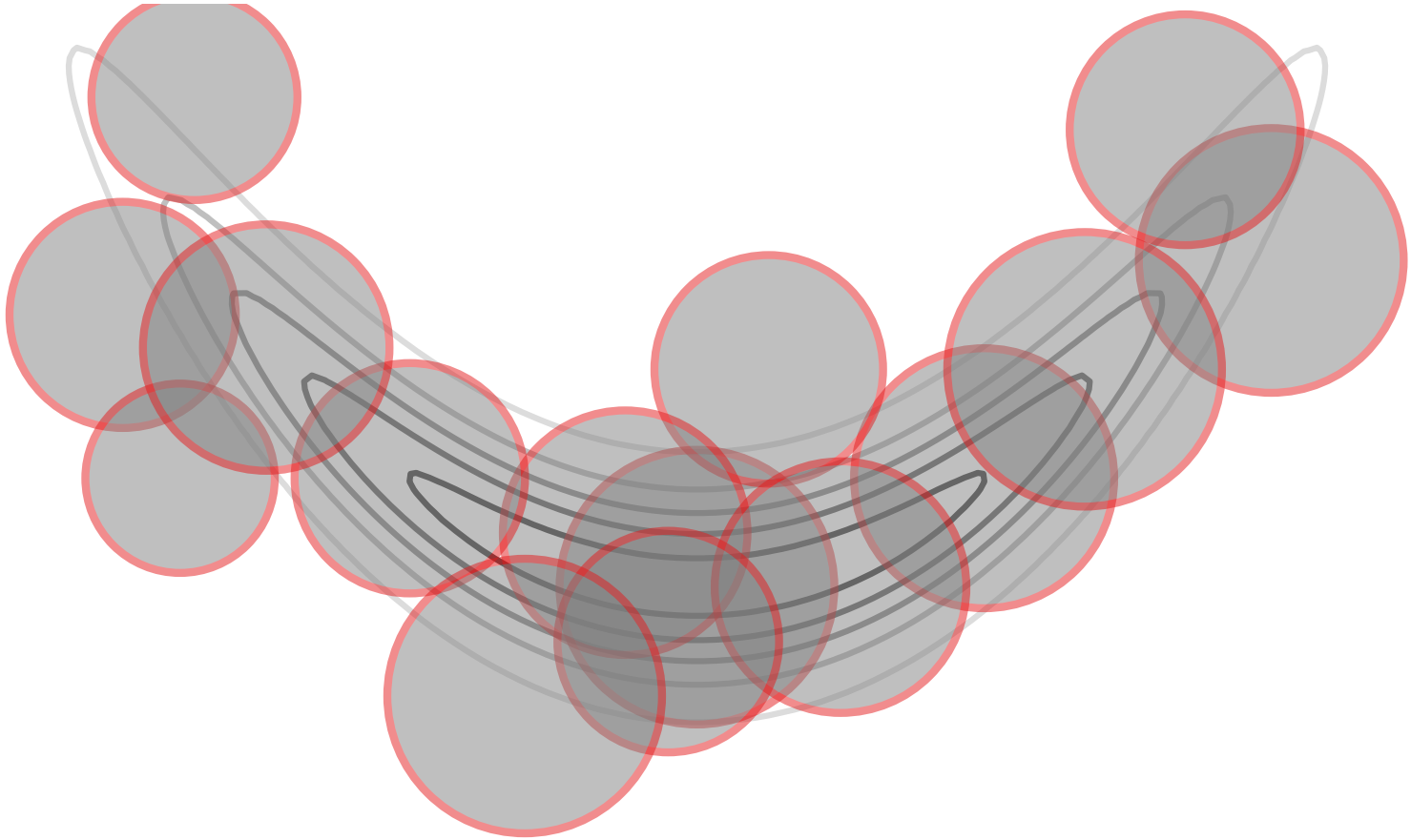
$$\breve{\pi}(x) = \arg\min_{\rho(x)} \sum_{i=1}^{n} \left(\rho(x^{(i)}) - \pi(x^{(i)})\right)^2 K(x^{(i)}, x)$$

- *Locally* supported kernel $K(x', x)$ finds k nearest neighbors
  - Note: this is NOT a Markov transition kernel
- $\mathcal{B}_k(x)$: smallest ball centered at x with k density evaluations

$$K(x', x) = \begin{cases} 1 & x' \in \mathcal{B}_k(x) \\ 0 & \text{otherwise} \end{cases}$$

Local polynomial surrogates

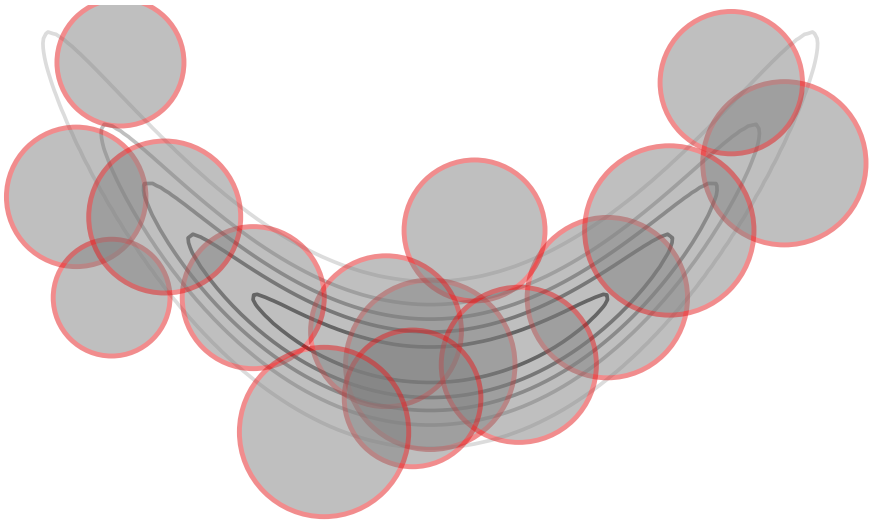
Build a local approximation in a ball around each point . . .



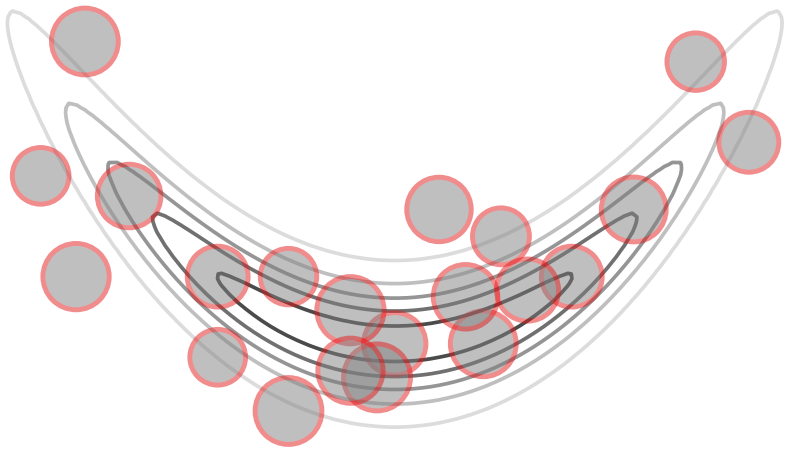
Surrogate convergence

In the **exact evaluation** case, the surrogate $\breve{\pi}(x)$ approaches $\pi(x)$ as:

1. The ball size $\Delta \to 0$ (number of evaluations $n \to \infty$)



Large balls

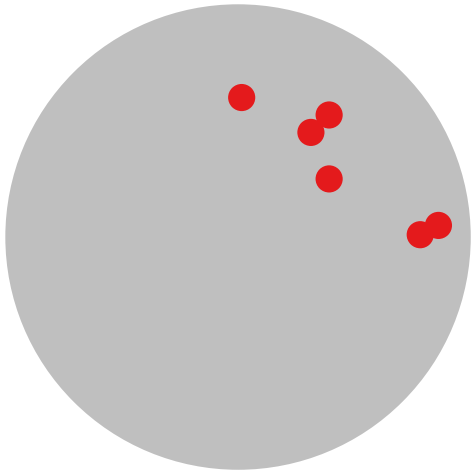


Small balls

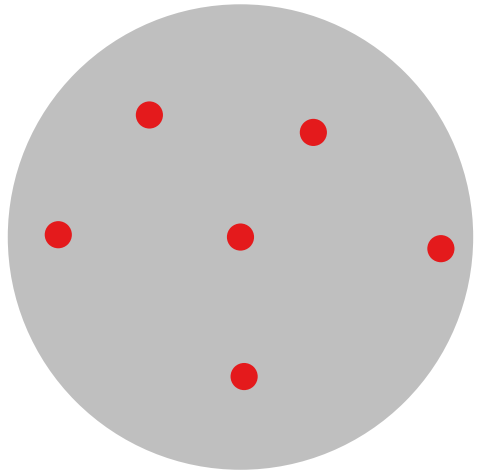
Surrogate convergence

In the **exact evaluation** case, the surrogate $\breve{\pi}(x)$ approaches $\pi(x)$ as:

1. The ball size $\Delta \to 0$ (number of evaluations $n \to \infty$)
2. Λ-poisedness is maintained inside each ball



Poorly poised



Well poised

Recall: Markov chain Monte Carlo (MCMC)

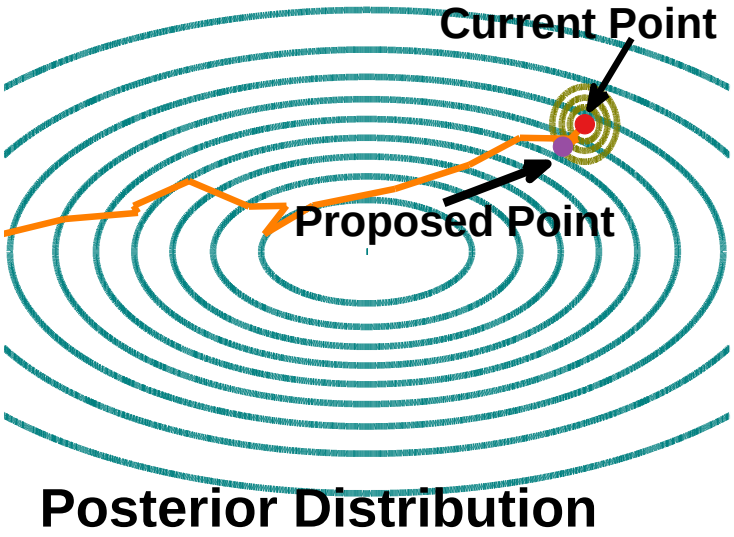
Three step algorithm:

1. Propose $x' \sim q_X(\cdot|x^{(t)})$
2. Acceptance probability

$$\alpha = \min\left(1, \frac{\pi(x')q(x^{(t)}|x')}{\pi(x^{(t)})q(x'|x^{(t)})}\right)$$

3. Accept/reject

$$x^{(t+1)} = \begin{cases} x' & \text{with probability } \alpha \\ x^{(t)} & \text{else} \end{cases}$$



Metropolis et al., 1953

Hastings, 1970

and variations . . .

Haario et al., 2006

Parno and Marzouk, 2014

Brooks et al., 2011

Recall: Markov chain Monte Carlo (MCMC)

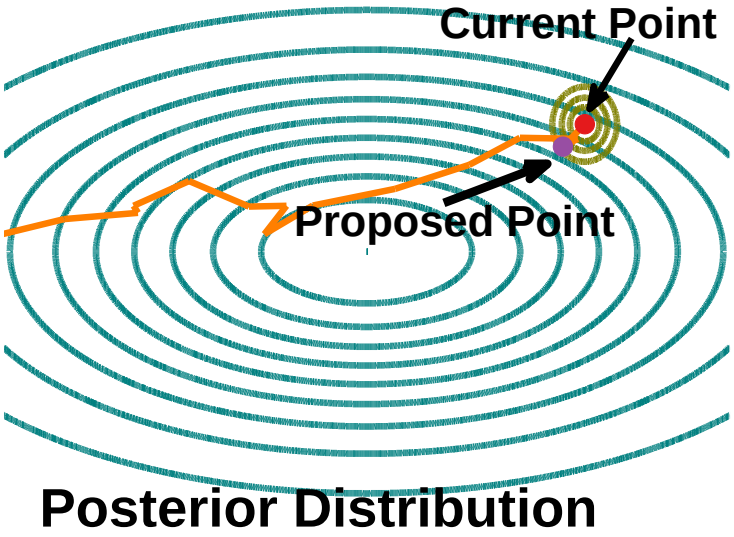
Three step algorithm:

1. Propose $x' \sim q_X(\cdot|x^{(t)})$
2. Acceptance probability

$$\alpha = \min\left(1, \frac{\pi(x')q(x^{(t)}|x')}{\pi(x^{(t)})q(x'|x^{(t)})}\right)$$

3. Accept/reject

$$x^{(t+1)} = \begin{cases} x' & \text{with probability } \alpha \\ x^{(t)} & \text{else} \end{cases}$$



- MCMC requires an expensive density evaluation every step
- Computationally prohibative

Metropolis et al., 1953
Hastings, 1970
and variations . . .
Haario et al., 2006
Parno and Marzouk, 2014
Brooks et al., 2011

Local approximation MCMC

$\breve{\pi}(x) \approx \pi(x)$ is a local polynomial approximation of the target density

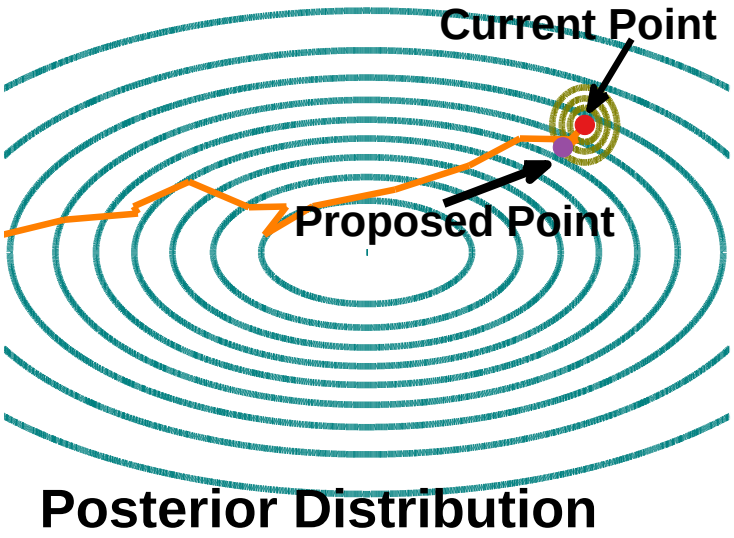
Four step algorithm:

1. **Possibly** refine $\breve{\pi}(\cdot)$ near $x^{(t)}$
2. Propose $x' \sim q_X(\cdot|x^{(t)})$
3. Acceptance probability

$$\alpha = \min\left(1, \frac{\breve{\pi}(x')q(x^{(t)}|x')}{\breve{\pi}(x^{(t)})q(x'|x^{(t)})}\right)$$

4. Accept/reject

$$x^{(t+1)} = \begin{cases} x' & \text{with probability } \alpha \\ x^{(t)} & \text{else} \end{cases}$$



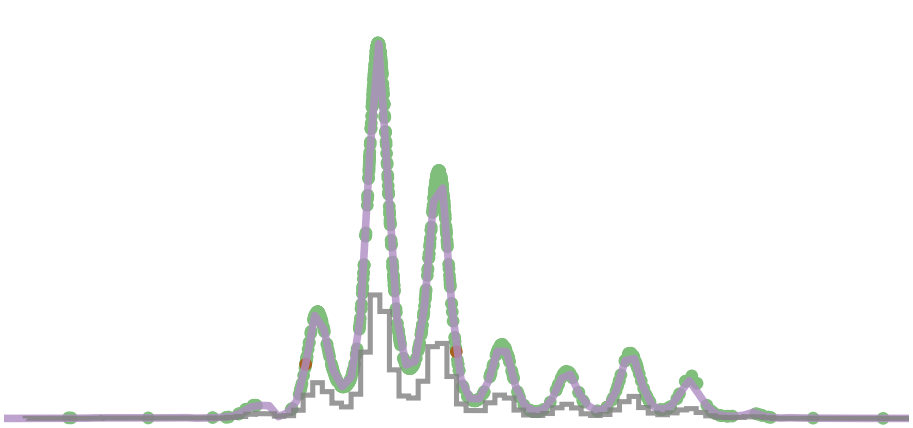
Refinement strategy

We devise a refinement strategy such that:

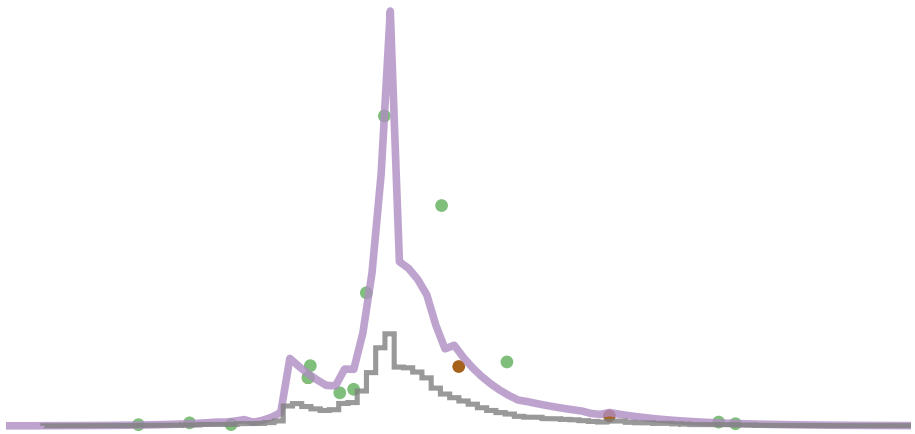
- Refinements cannot happen every MCMC step
- The expected number of refinements is infinite as $t \to \infty$
- The refinement frequency decays as $t \to \infty$
- We choose new points that maintain Λ-poisedness

Over- versus under-refinement

- Two error sources after a *finite* number of MCMC steps:
  1. MCMC variance (CLT: decays like $1/t$)
  2. Surrogate bias
- Frequent refinement $\Rightarrow$ MCMC variance dominates error
- Infrequent refinement $\Rightarrow$ surrogate bias dominates error



Over-refined



Under-refined

Bias-variance trade-off

- Assume the MCMC variance decays with the number of steps

$$[\text{MCMC variance}] \leq C_{\text{MCMC}} t^{-1}$$

- The surrogate bias is bounded

$$[\text{Surrogate bias}] = |\hat{\pi}(x) - \pi(x)| \leq C_{\text{Surrogate}} \bar{\Lambda} \Delta^{p+1}$$

- **Balance MCMC variance with surrogate bias squared**

$$C_{\text{MCMC}} t^{-1} \sim C_{\text{Surrogate}}^2 \bar{\Lambda}^2 \Delta^{2(p+1)}$$

Balance MCMC variance with surrogate bias squared

- Divide the chain into m levels
- Prescribe an error threshold $\gamma(m) = \gamma_0 m^{-\gamma_1}$ on each level
  - γ_0: initial error threshold
  - γ_1: error threshold decay rate
- Switch to level $m+1$ when

$$C_{\text{MCMC}} t_m^{-1} = {\gamma_0}^2 m^{-2\gamma_1}$$

- Since C_{MCMC} is unknown

$$t_m = \varphi_1 M^{2\gamma_1}$$

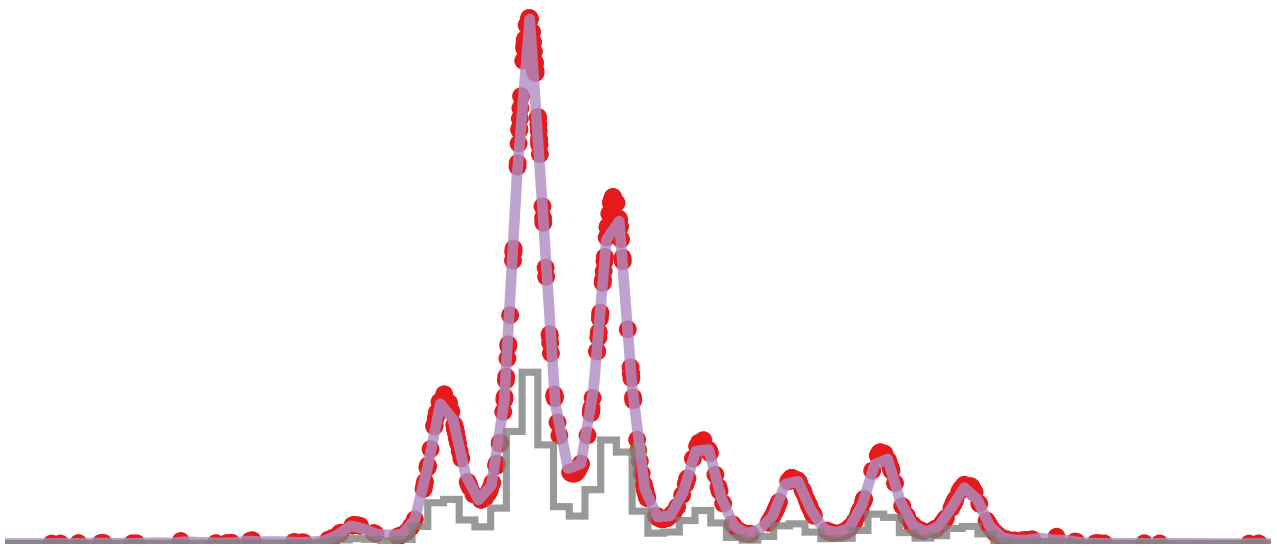
- Require $\gamma_1 > 0.5$ so the length of each level $t_m - t_{m-1}$ grows

Structural refinement

At $x^{(t)}$, refine the surrogate if:

1. The local error indicator is greater than the level's threshold

$$\Delta^{p+1}(x^{(t)}) > \gamma_0 M^{-\gamma_1}$$



Red dots:

Exact density evaluations

Purple line:

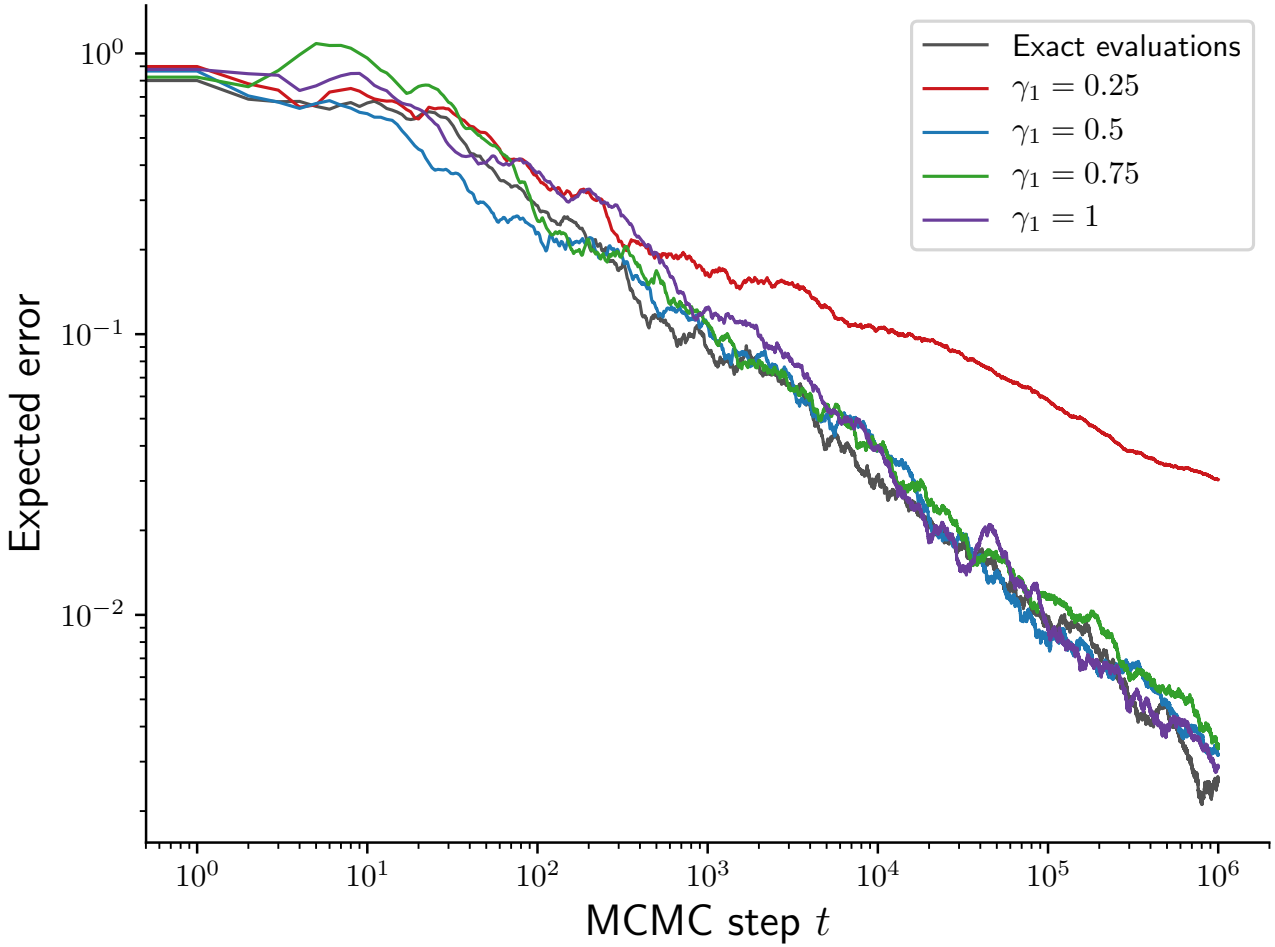
Local polynomial approximation

Grey line:

Binned MCMC samples

Expected error (structural refinement)

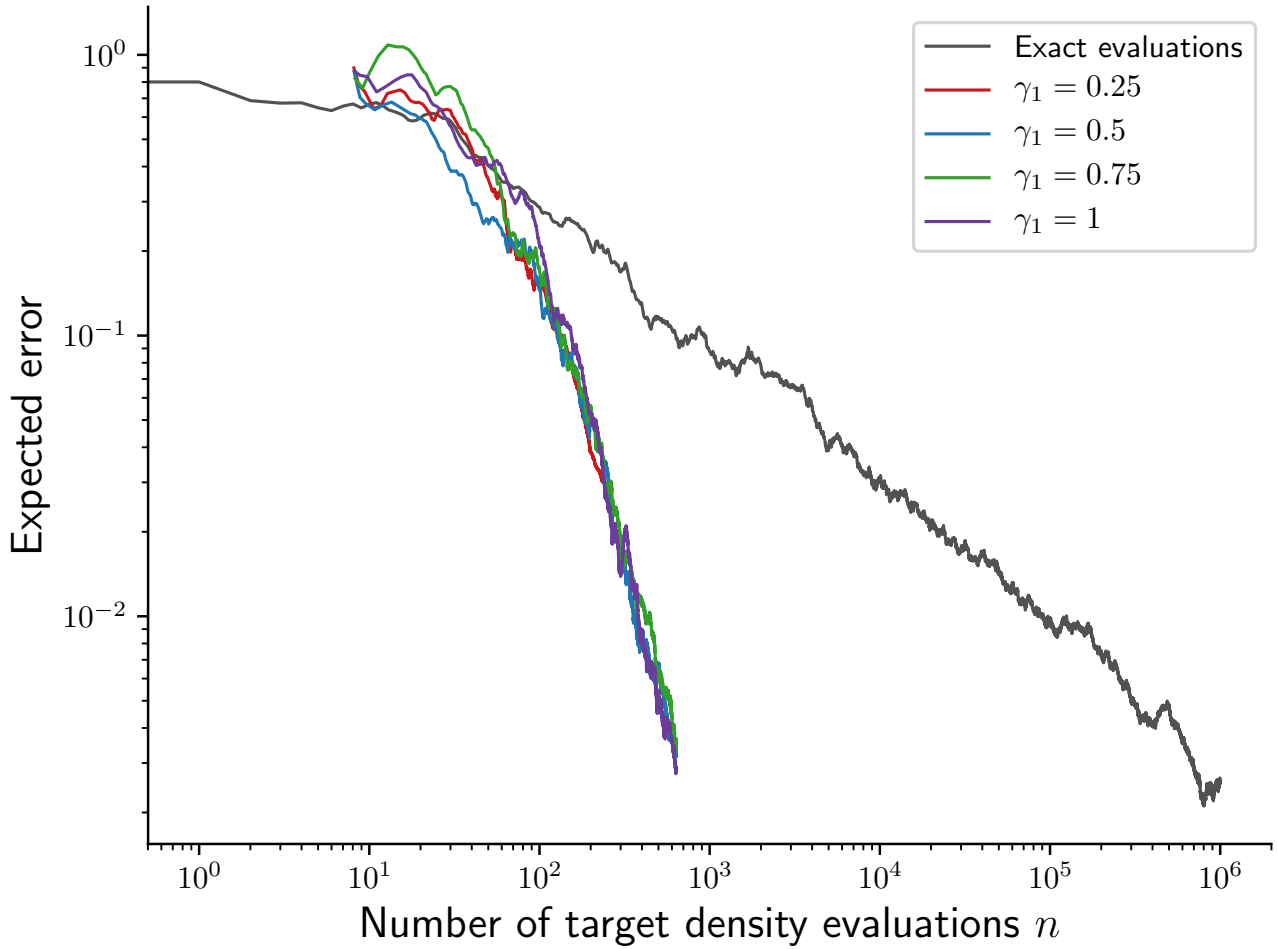
- MCMC with exact evaluations is a lower bound for the expected error
- The expected error decays at the same rate as MCMC with exact evaluations ($\gamma_1 > 0.5$)



- Refinement locations are chosen by the local error indicator

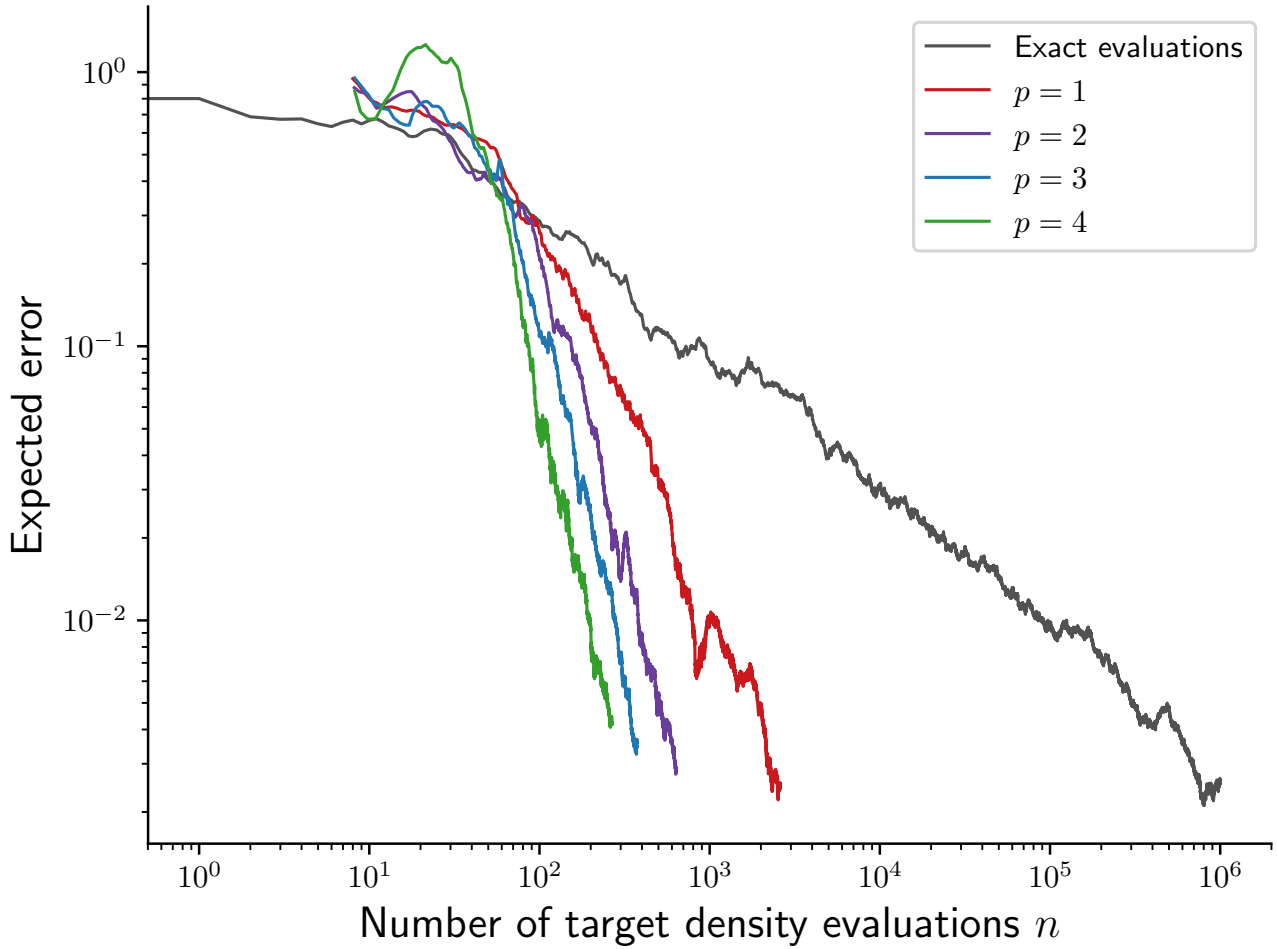
Expected error (structural refinement)

- The expected number of refinements is the same (when $\gamma_1 > 0.5$)
- When $\gamma_1 \leq 0.5$, the surrogate is underrefined
  - The error is dominated by the structural bias



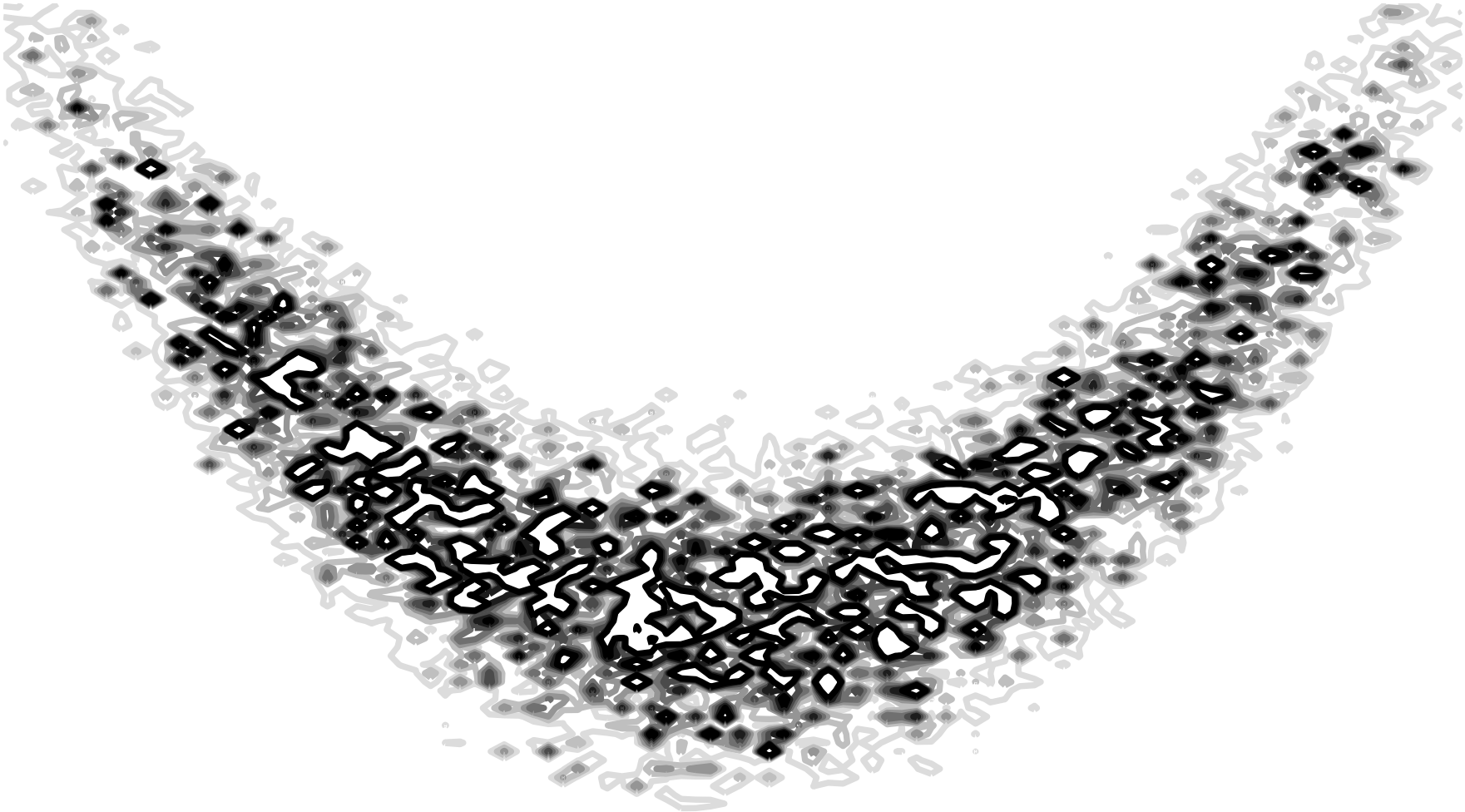
Expected error (structural refinement)

- Efficiency improves with polynomial order



Local polynomial surrogates

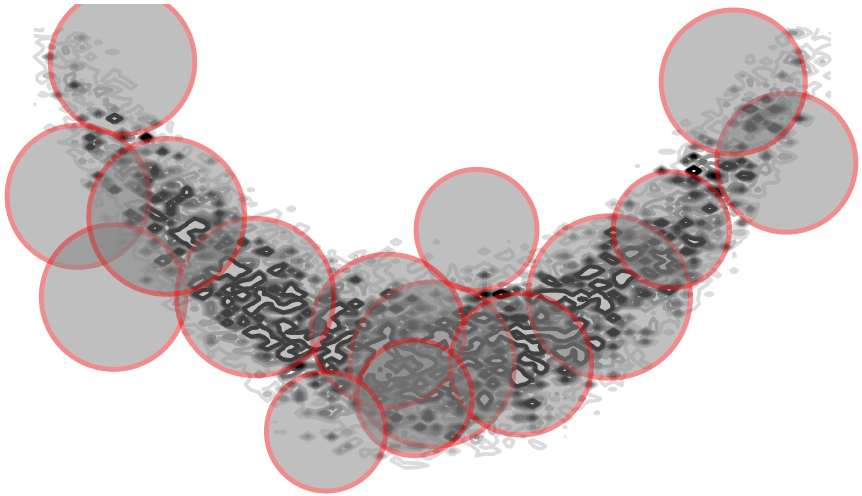
Consider situations with **noisy evaluations** of the target density



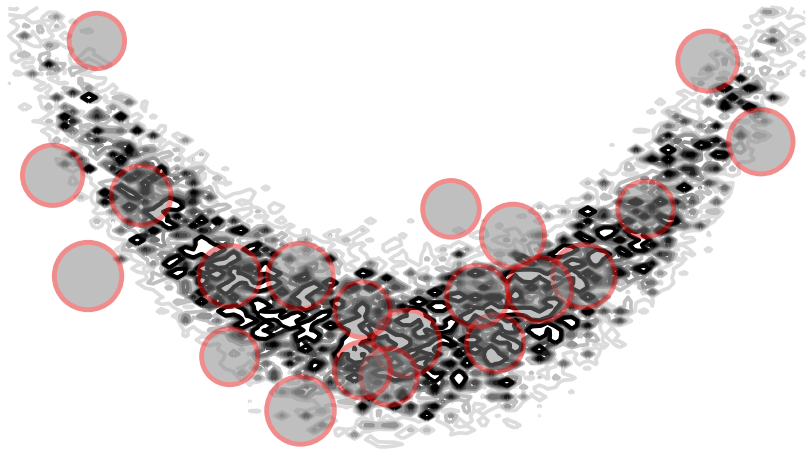
Surrogate convergence (noisy density evaluations)

The expected surrogate $\mathbb{E}[\breve{\pi}(x)]$ approaches $\pi(x)$ as:

1. The ball size $\Delta \to 0$ (number of evaluations $n \to \infty$)
2. Λ-poisedness is maintained inside each ball



Large balls

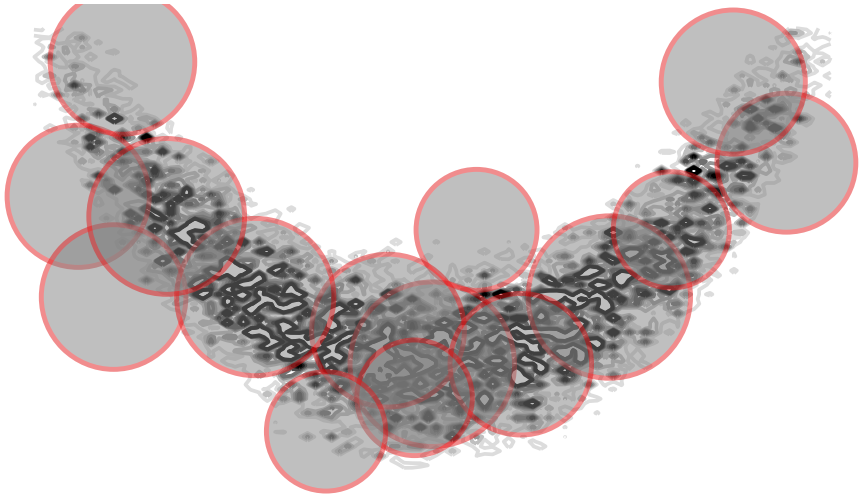


Small balls

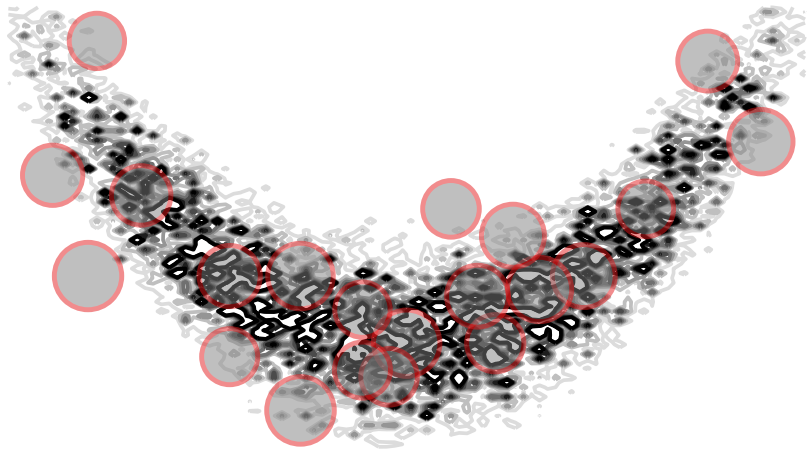
Surrogate convergence (noisy density evaluations)

The expected surrogate $\mathbb{E}[\breve{\pi}(x)]$ approaches $\pi(x)$ as:

1. The ball size $\Delta \to 0$ (number of evaluations $n \to \infty$)
2. Λ-poisedness is maintained inside each ball



Large balls



Small balls

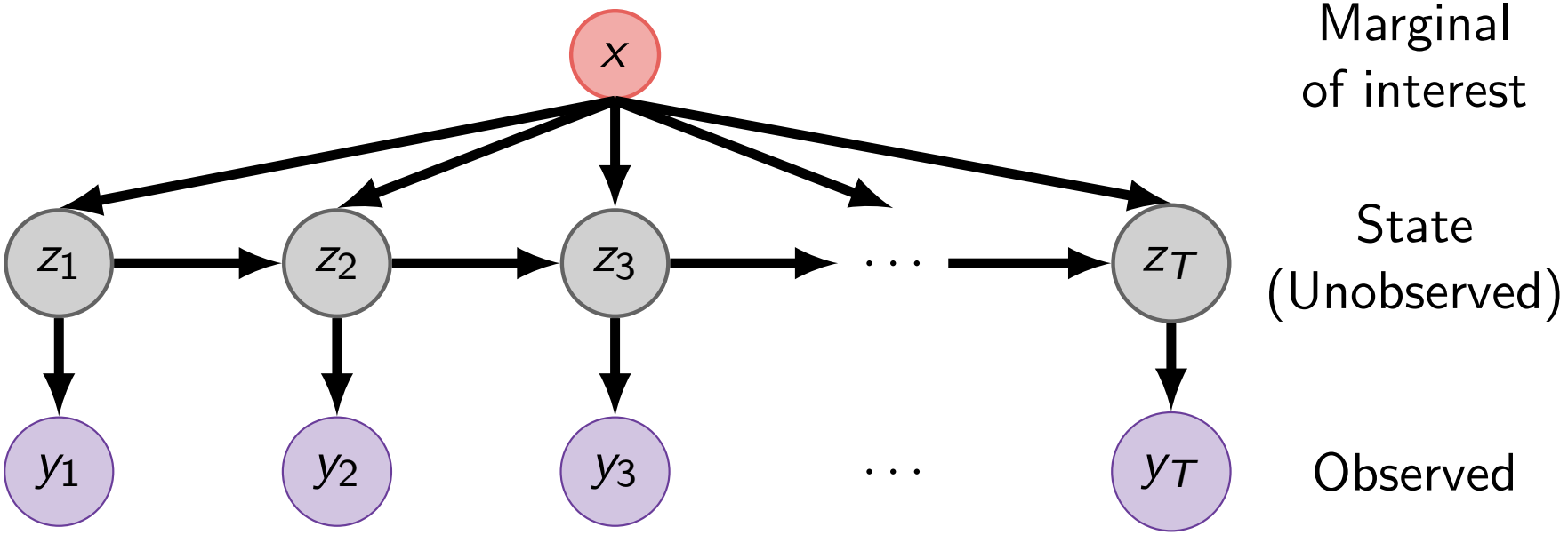
Replacing $\pi(x)$ with $\breve{\pi}(x)$ within MCMC asymptotically characterizes the target distribution Davis et al., in progress

State space example: stochastic volatility

- Hyperparameters: $[\nu, \phi, \sigma^2] = g(x_1, x_2, x_3)$
- g is nonlinear transformation such that $x \sim N(0, I)$ imples $[\nu, \phi, \sigma^2]$ are sampled from Gamma distributions

$$Z_1 \sim N(\nu, (1-\phi^2)^{-1}) \quad \text{and} \quad Z_t | Z_{t-1} \sim N(\nu + \phi(Z_{t-1} - \nu), \sigma^2)$$

$$Y_t \sim N(0, \exp(Z_t))$$



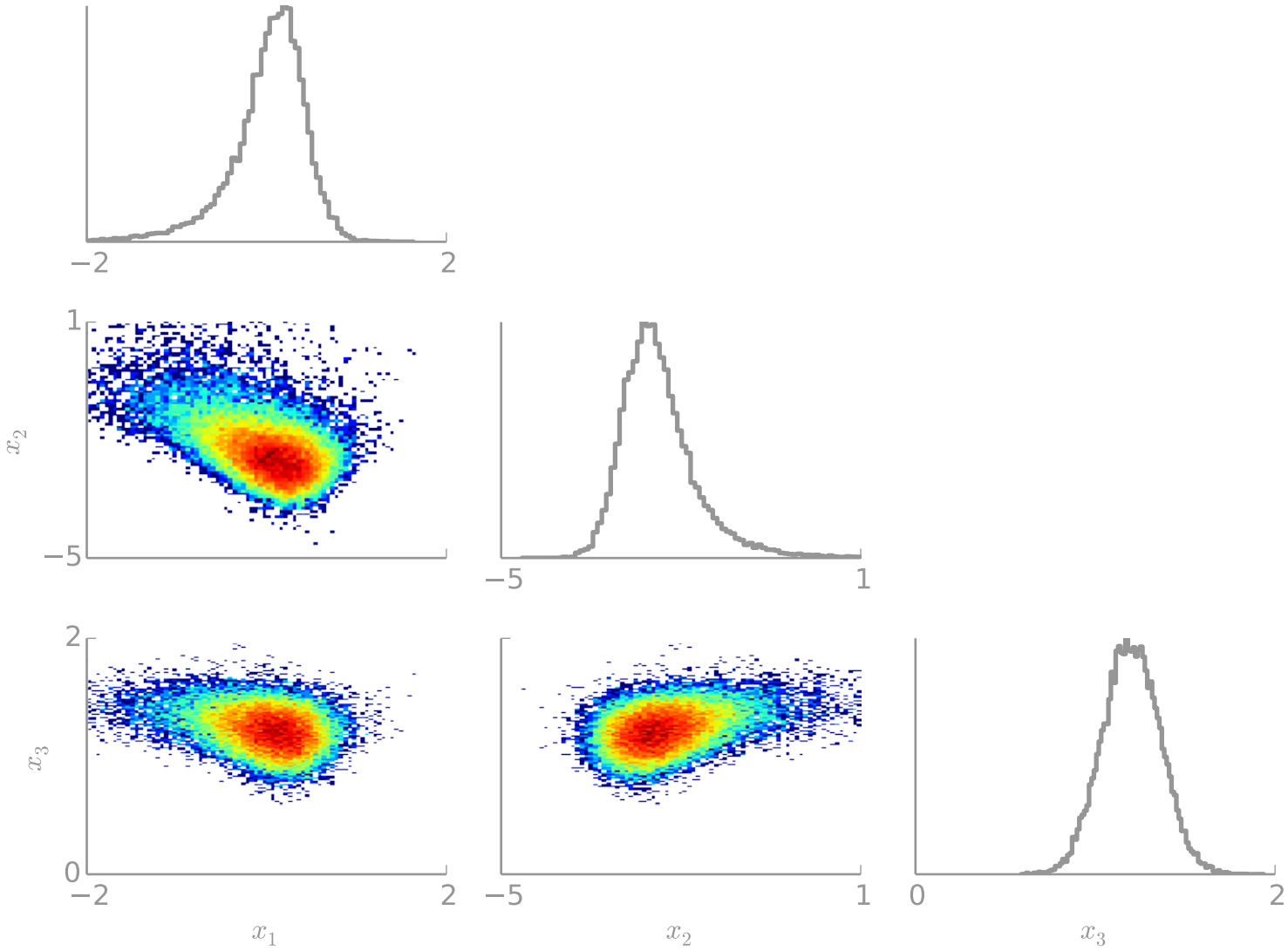
Problem statement: stochastic volatility

- Given: data $y_{1:T}$
- Define the target posterior marginal distribution

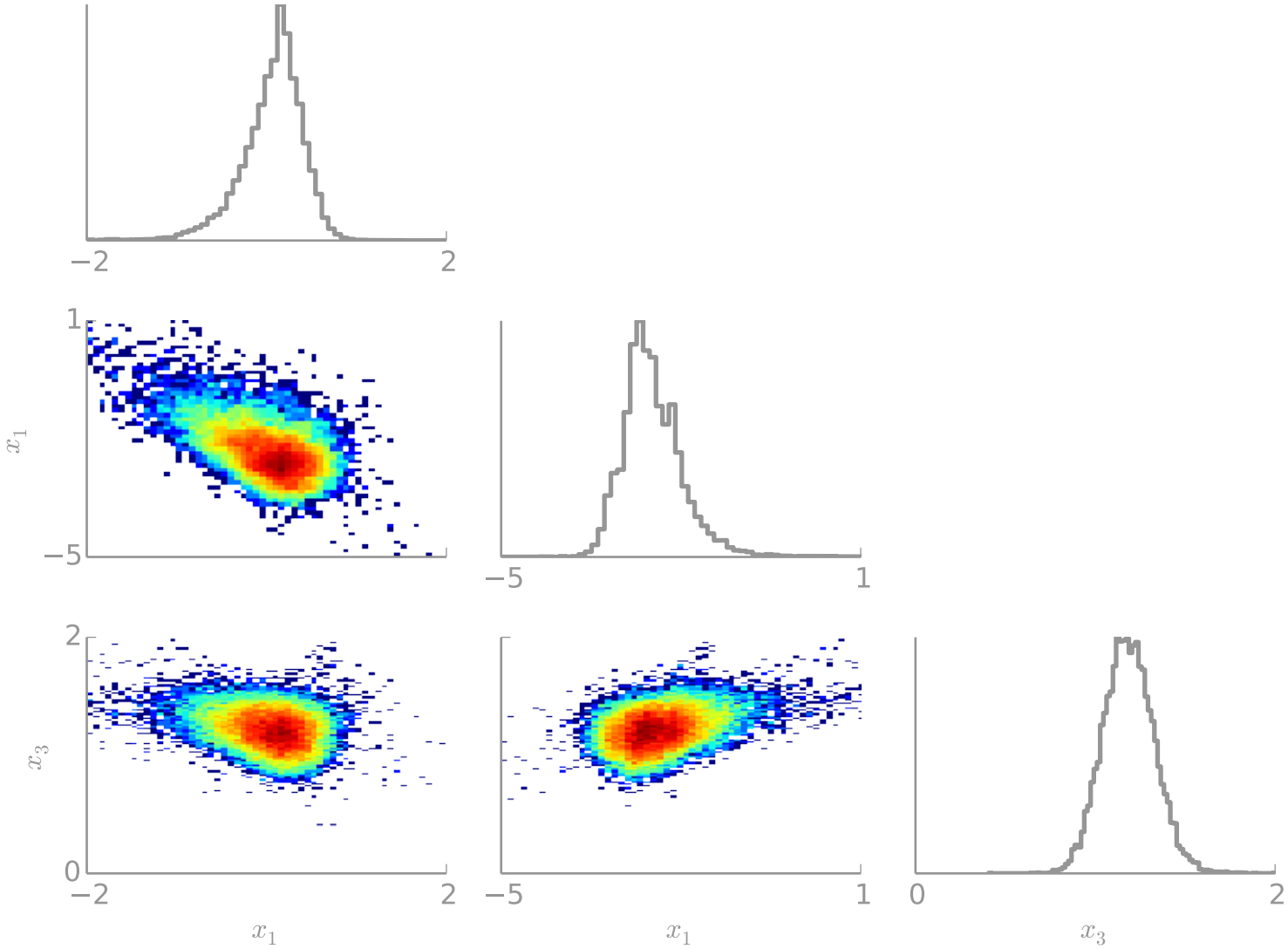
$$\begin{aligned}\pi(x|y_{1:T}) &= \int \pi(x, z_{1:T}|y_{1:T})\, dz_{1:T} \\ &\propto \underbrace{\pi(x)}_{\text{Prior}} \underbrace{\int \pi(y_{1:T}|z_{1:T}, x)\pi(z_{1:T}|x)\, dz_{1:T}}_{\text{Marginal likelihood}}\end{aligned}$$

- Estimate the marginal likelihood with importance sampling
  - Choosing the biasing distribution is nontrivial!

Stochastic volatility results: PM-MCMC



Stochastic volatility results: LA-MCMC



- Number of importance sampling estimates: $\approx 2.8 \times 10^3 \ll 2 \times 10^5$

Conclusions

- Building and refining a local polynomial approximations significantly reduces computational expense
- An ideal refinement rate comes from a bias-variance trade-off between surrogate model error and MCMC variance defines ideal
- We characterize distributions given *noisy* target density evaluations by building an asymptotically exact surrogate
  - e.g., posterior marginal distributions
- Code: MIT Uncertainty Quantification (MUQ) libraries

  `muq.mit.edu`