Incremental local approximations for computationally intensive MCMC on targeted marginals

[1]**Andrew D. Davis**, [1]Youssef Marzouk, [2]Patrick Heimbach

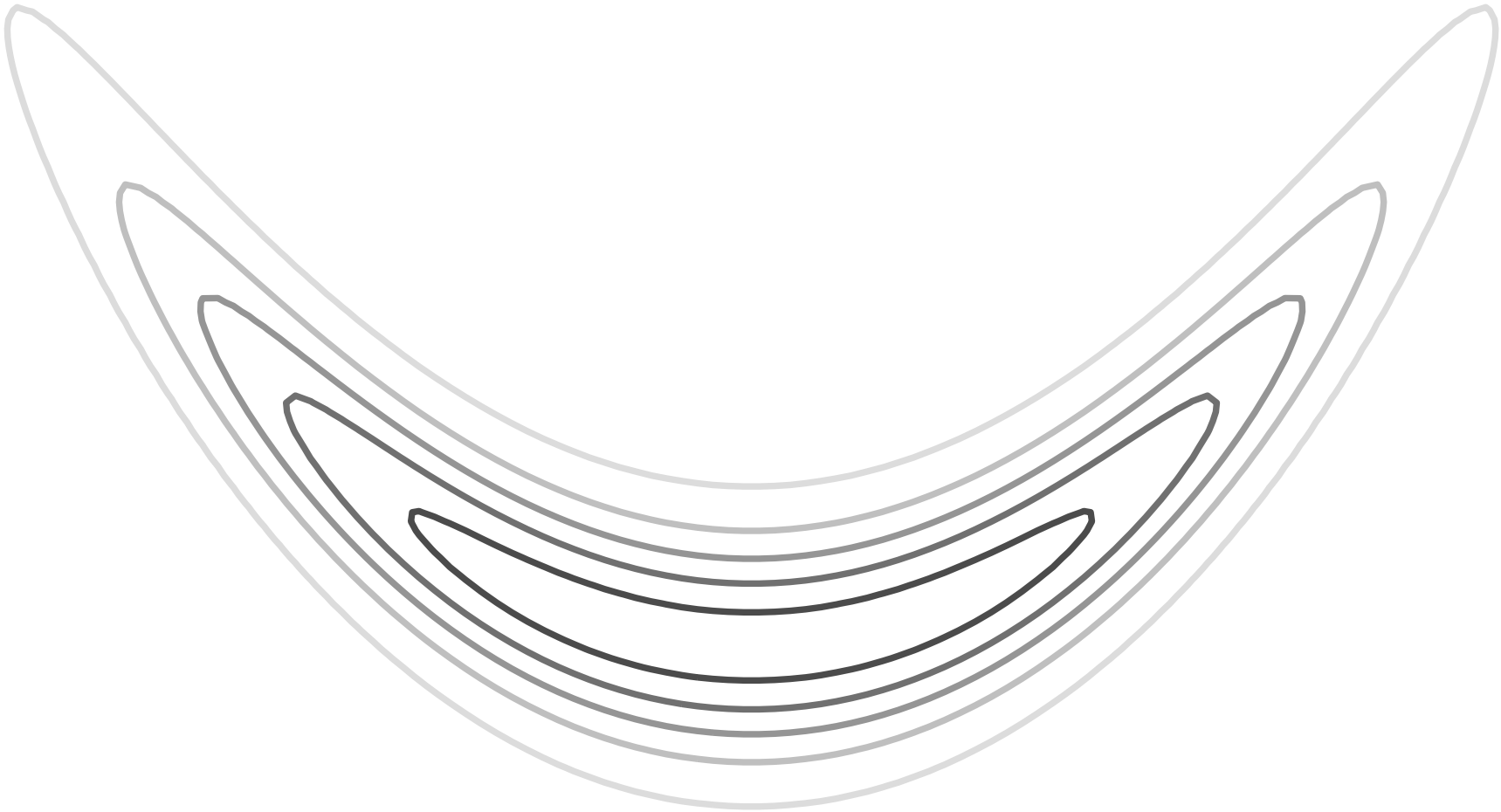
[1]Department of Aeronautics and Astronautics
Center for Computational Engineering
Massachusetts Institute of Technology
`http://uqgroup.mit.edu`

[2]Department of Earth, Atmosphere, and Planetary Sciences
Massachusetts Institute of Technology

July 2016

Problem statement

We want to **characterize the distribution π** using a sampling method

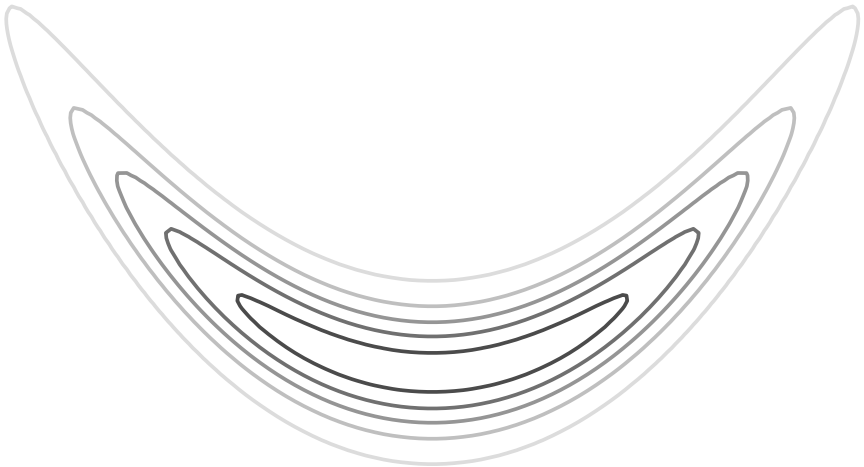


Problem statement

We want to **characterize the distribution π** using a sampling method

Two problems:

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

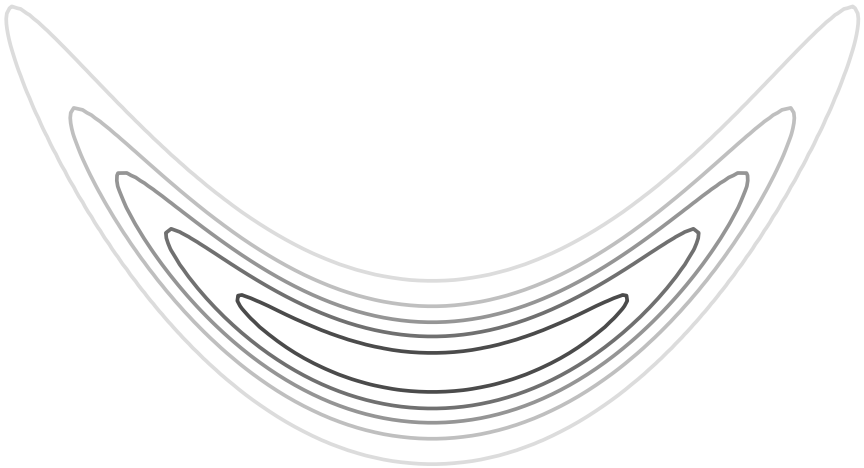


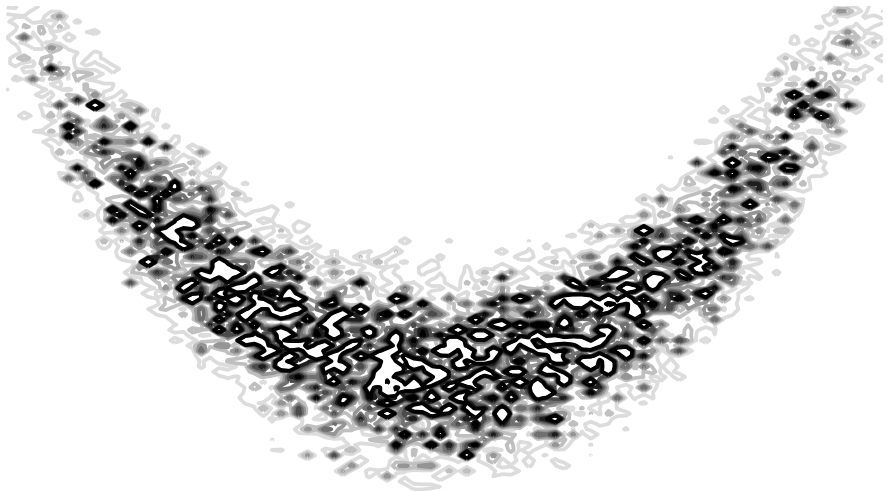
Problem statement

We want to **characterize the distribution π** using a sampling method

Two problems:

1. The probability density function $\pi(\theta)$ is **computationally expensive**
2. Only **noisy** density evaluations are available $\tilde{\pi}(\theta) = \pi(\theta) + \varepsilon$

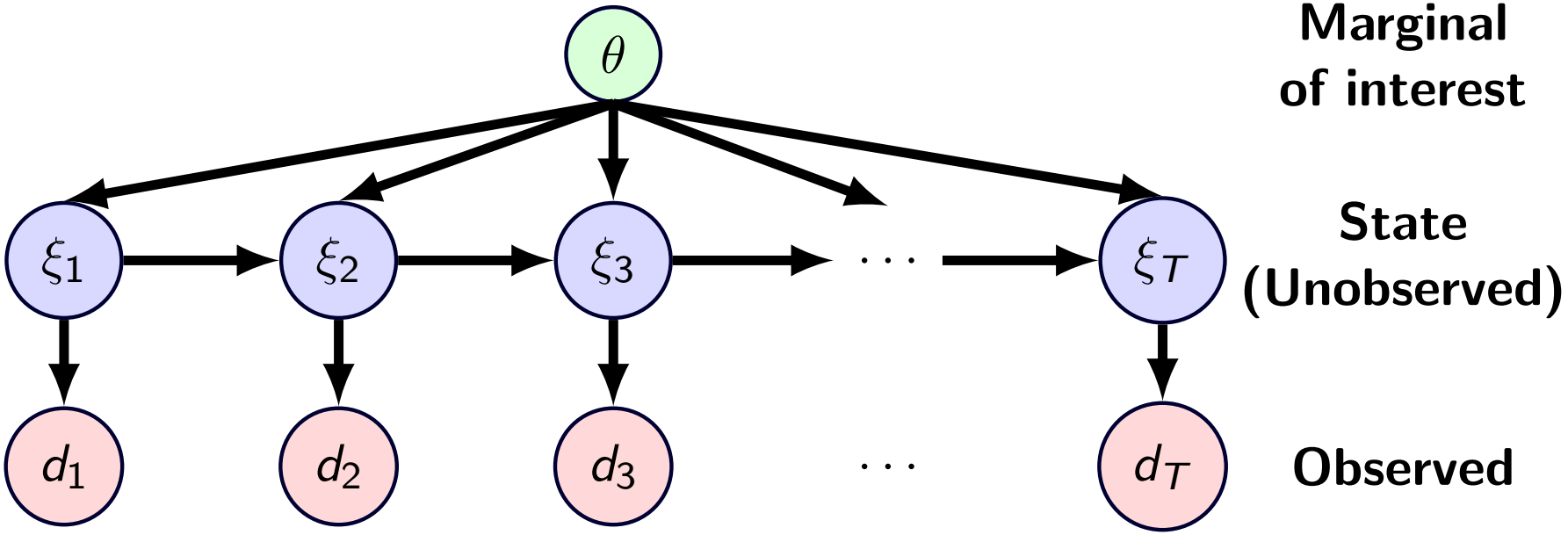




Example: state space modelling

State space model, parameter estimation —

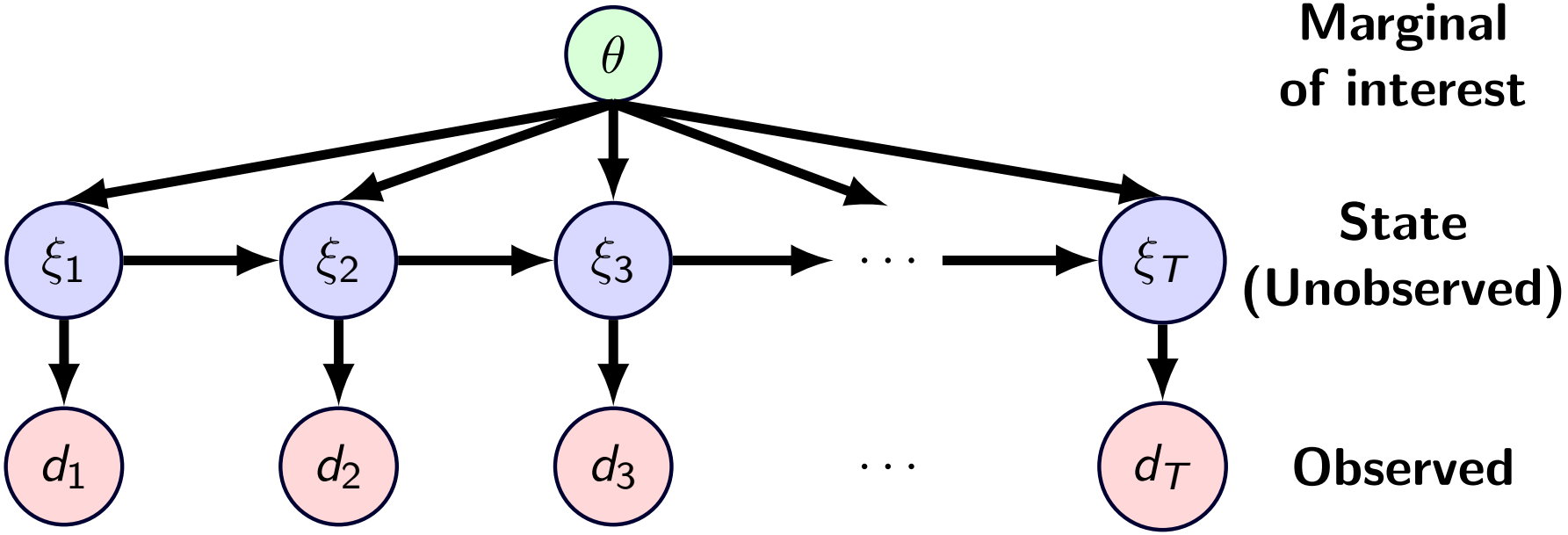
- Given observed data $d_{1:T}$ with an unobservable state $\xi_{1:T}$
- Estimate the static parameters θ



Example: state space modelling

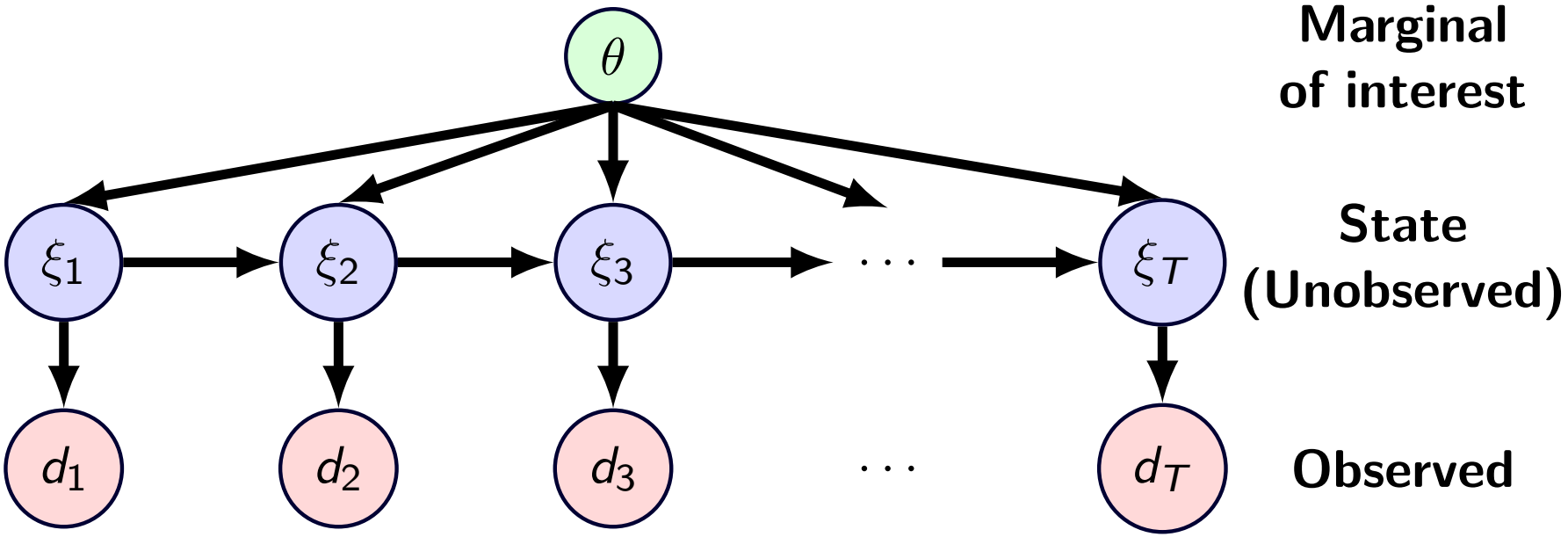
State space model, parameter estimation —

- Given observed data $d_{1:T}$ with an unobservable state $\xi_{1:T}$
- Estimate the static parameters θ



- We want to characterize the distribution of $\theta | d_{1:T}$

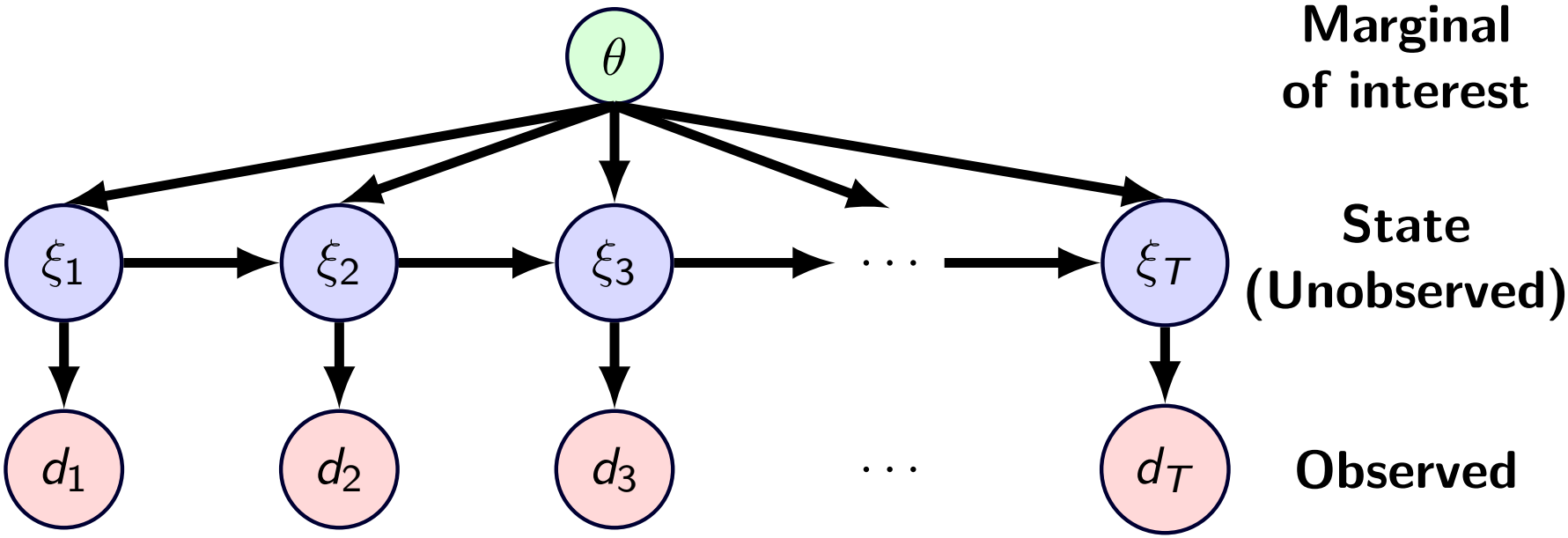
Example: state space modelling (joint inference)



Characterize the Bayesian posterior

$$\underbrace{\pi(\theta, \xi_{1:T} | d_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(\theta)\pi(\xi_{1:T}|\theta)}_{\text{Prior}} \underbrace{\pi(d_{1:T}|\theta, \xi_{1:T})}_{\text{Likelihood}}$$

Example: state space modelling (joint inference)



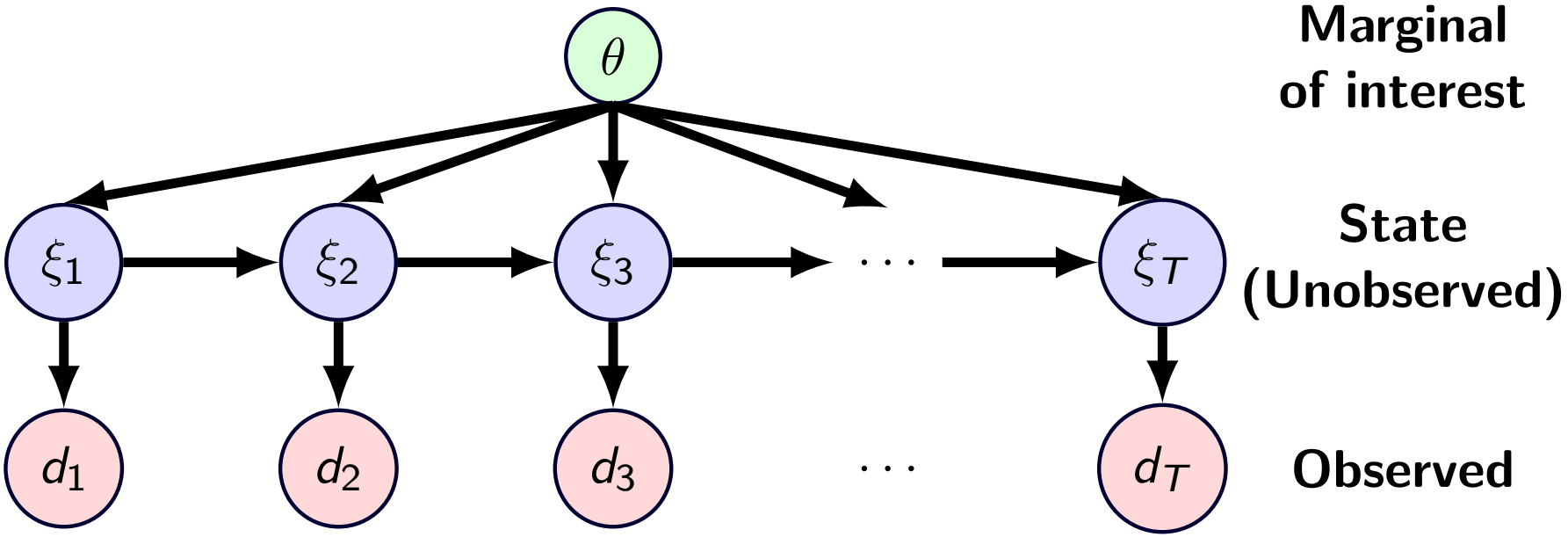
Characterize the Bayesian posterior

$$\underbrace{\pi(\theta, \xi_{1:T}|d_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(\theta)\pi(\xi_{1:T}|\theta)}_{\text{Prior}} \underbrace{\pi(d_{1:T}|\theta, \xi_{1:T})}_{\text{Likelihood}}$$

Challenges:

- The joint parameter $[\theta, \xi_{1:T}]$ may be high dimensional

Example: state space modelling (joint inference)



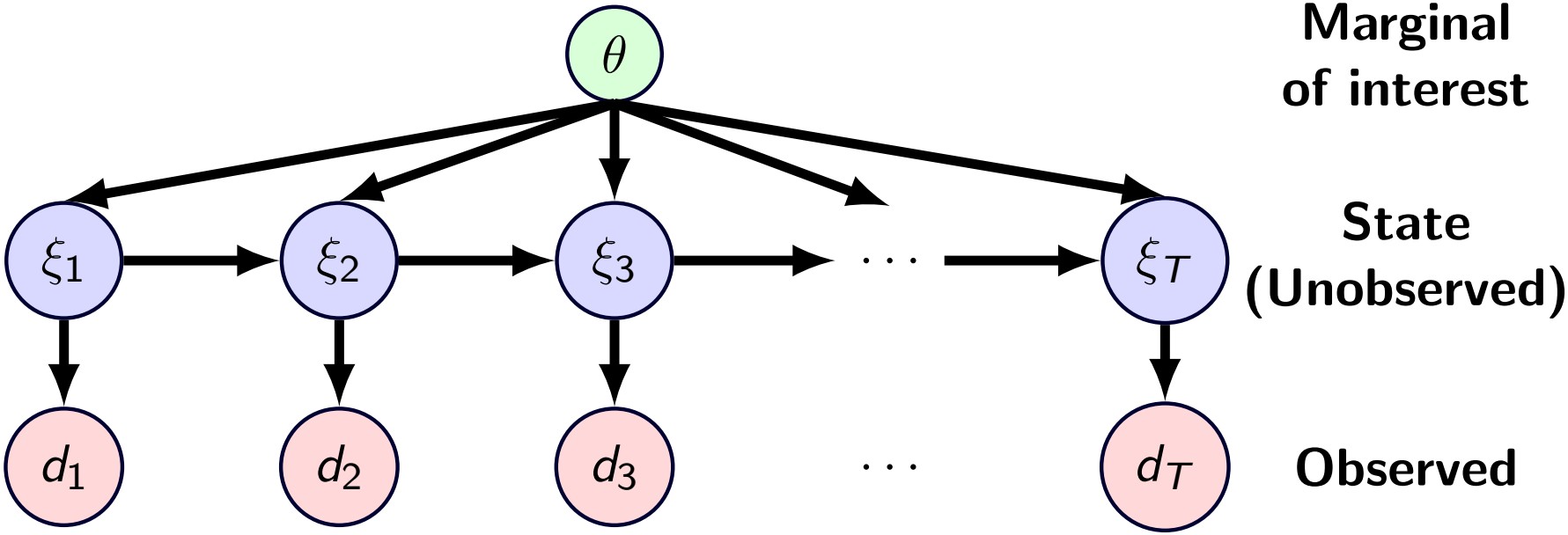
Characterize the Bayesian posterior

$$\underbrace{\pi(\theta, \xi_{1:T}|d_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(\theta)\pi(\xi_{1:T}|\theta)}_{\text{Prior}} \underbrace{\pi(d_{1:T}|\theta, \xi_{1:T})}_{\text{Likelihood}}$$

Challenges:

- The joint parameter $[\theta, \xi_{1:T}]$ may be high dimensional
- The posterior density $\pi(\theta, \xi_{1:T}|d)$ is computationally expensive

Example: state space modelling (marginal inference)



Characterize the posterior marginal

$$\underbrace{\pi(\theta|d_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(\theta)}_{\text{Prior}} \underbrace{\int_{\Xi} \pi(d_{1:T}, \xi_{1:T}|\theta)\, \mathrm{d}\xi_{1:T}}_{\text{Likelihood}}$$

Example: state space modelling (marginal inference)

Characterize the posterior marginal

$$\underbrace{\pi(\theta|d_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(\theta)}_{\text{Prior}} \underbrace{\int_{\Xi} \pi(d_{1:T}, \xi_{1:T}|\theta)\, \mathrm{d}\xi_{1:T}}_{\text{Likelihood}}$$

Example: state space modelling (marginal inference)

Characterize the posterior marginal

$$\underbrace{\pi(\theta|d_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(\theta)}_{\text{Prior}} \underbrace{\int_{\Xi} \pi(d_{1:T}, \xi_{1:T}|\theta)\, \mathrm{d}\xi_{1:T}}_{\text{Likelihood}}$$

Suppose we have a noisy estimate of the likelihood

$$\pi(d_{1:T}|\theta) \approx \sum_{i=1}^{N} w^{(i)} \pi\left(d_{1:T}, \xi_{1:T}^{(i)}|\theta\right)$$

Example: state space modelling (marginal inference)

Characterize the posterior marginal

$$\underbrace{\pi(\theta|d_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(\theta)}_{\text{Prior}} \underbrace{\int_{\Xi} \pi(d_{1:T}, \xi_{1:T}|\theta)\, \mathrm{d}\xi_{1:T}}_{\text{Likelihood}}$$

Suppose we have a noisy estimate of the likelihood

$$\pi(d_{1:T}|\theta) \approx \sum_{i=1}^{N} w^{(i)} \pi\left(d_{1:T}, \xi_{1:T}^{(i)}|\theta\right)$$

- **Pseudo-marginal MCMC** can characterize the posterior marginal distribution (Beaumont, 2010) and (Andrieu and Roberts, 2009)

Example: state space modelling (marginal inference)

Characterize the posterior marginal

$$\underbrace{\pi(\theta|d_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(\theta)}_{\text{Prior}} \underbrace{\int_{\Xi} \pi(d_{1:T}, \xi_{1:T}|\theta)\, \mathrm{d}\xi_{1:T}}_{\text{Likelihood}}$$

Suppose we have a noisy estimate of the likelihood

$$\pi(d_{1:T}|\theta) \approx \sum_{i=1}^{N} w^{(i)} \pi\left(d_{1:T}, \xi_{1:T}^{(i)}|\theta\right)$$

- **Pseudo-marginal MCMC** can characterize the posterior marginal distribution (Beaumont, 2010) and (Andrieu and Roberts, 2009)
- Parameter space is partitioned into coordinates characterized by MCMC (θ) and coordinates to be "marginalized away" ($\xi_{1:T}$)

Example: state space modelling (marginal inference)

Characterize the posterior marginal

$$\underbrace{\pi(\theta|d_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(\theta)}_{\text{Prior}} \underbrace{\int_{\Xi} \pi(d_{1:T}, \xi_{1:T}|\theta)\, \mathrm{d}\xi_{1:T}}_{\text{Likelihood}}$$

Suppose we have a noisy estimate of the likelihood

$$\pi(d_{1:T}|\theta) \approx \sum_{i=1}^{N} w^{(i)} \pi\left(d_{1:T}, \xi_{1:T}^{(i)}|\theta\right)$$

- **Pseudo-marginal MCMC** can characterize the posterior marginal distribution (Beaumont, 2010) and (Andrieu and Roberts, 2009)
- Parameter space is partitioned into coordinates characterized by MCMC (θ) and coordinates to be "marginalized away" ($\xi_{1:T}$)
- MCMC now targets a lower dimensional parameter θ

Example: state space modelling (marginal inference)

Characterize the posterior marginal

$$\underbrace{\pi(\theta|d_{1:T})}_{\text{Posterior}} \propto \underbrace{\pi(\theta)}_{\text{Prior}} \underbrace{\int_{\Xi} \pi(d_{1:T}, \xi_{1:T}|\theta)\, \mathrm{d}\xi_{1:T}}_{\text{Likelihood}}$$

Suppose we have a noisy estimate of the likelihood

$$\pi(d_{1:T}|\theta) \approx \sum_{i=1}^{N} w^{(i)} \pi\left(d_{1:T}, \xi_{1:T}^{(i)}|\theta\right)$$

- **Pseudo-marginal MCMC** can characterize the posterior marginal distribution (Beaumont, 2010) and (Andrieu and Roberts, 2009)
- Parameter space is partitioned into coordinates characterized by MCMC (θ) and coordinates to be "marginalized away" ($\xi_{1:T}$)
- MCMC now targets a lower dimensional parameter θ
- Model evaluations are (even more) computationally expensive and now we only have noisy target density evaluations

Approaches & outline

- Exploit regularity of the target density to build a surrogate model

Approaches & outline

- Exploit regularity of the target density to build a surrogate model
- Continual refinement of surrogate model asymptotically guarantees sampling from the true target distribution (Conrad et al., JASA 2015)

Approaches & outline

- Exploit regularity of the target density to build a surrogate model
- Continual refinement of surrogate model asymptotically guarantees sampling from the true target distribution (Conrad et al., JASA 2015)
- A tradeoff between error due to the surrogate model and Monte Carlo variance implies an ideal refinement rate

Approaches & outline

- Exploit regularity of the target density to build a surrogate model
- Continual refinement of surrogate model asymptotically guarantees sampling from the true target distribution (Conrad et al., JASA 2015)
- A tradeoff between error due to the surrogate model and Monte Carlo variance implies an ideal refinement rate
- Partitioning the parameter space isolates coordinate directions we care about

Approaches & outline

- Exploit regularity of the target density to build a surrogate model
- Continual refinement of surrogate model asymptotically guarantees sampling from the true target distribution (Conrad et al., JASA 2015)
- A tradeoff between error due to the surrogate model and Monte Carlo variance implies an ideal refinement rate
- Partitioning the parameter space isolates coordinate directions we care about
- A surrogate model built from noisy target density evaluations can still characterize the true target distribution

Surrogate modelling: computationally cheaper options

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

Surrogate modelling: computationally cheaper options

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

- We cannot afford density evaluations at every MCMC step!

Surrogate modelling: computationally cheaper options

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

- We cannot afford density evaluations at every MCMC step!
- We **leverage the density's underlying regularity** to build computationally cheaper surrogate models:
  - Polynomial approximations (Marzouk et al., 2007) and (Marzouk and Xiu, 2009)
  - Gaussian processes (Rasmussen, 2006), (Bernardo et al., 2008), (Sacks et al., 1989), and (Santner et al., 2003)
  - And many others ...

Surrogate modelling: computationally cheaper options

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

- We cannot afford density evaluations at every MCMC step!
- We **leverage the density's underlying regularity** to build computationally cheaper surrogate models:
  - Polynomial approximations (Marzouk et al., 2007) and (Marzouk and Xiu, 2009)
  - Gaussian processes (Rasmussen, 2006), (Bernardo et al., 2008), (Sacks et al., 1989), and (Santner et al., 2003)
  - And many others ...
- We focus on **local polynomial approximations** (Conn, 2009), (Stone, 1977), and (Kohler, 2002)
  - In contrast with previous methods, these approximations enable **asymptotically exact sampling**

Local polynomial surrogates

- Given **exact** density evaluations $y_i = \pi(\theta)$

$$\mathcal{Y}(\boldsymbol{\theta}_n) \equiv \{(\theta_1, y_1), \ldots (\theta_n, y_n)\}$$

Local polynomial surrogates

- Given **exact** density evaluations $y_i = \pi(\theta)$

$$\mathcal{Y}(\boldsymbol{\theta}_n) \equiv \{(\theta_1, y_1), \ldots (\theta_n, y_n)\}$$

- Find the degree p polynomial $h_n(\theta; \mathcal{Y}(\boldsymbol{\theta}_n))$ such that

$$h_n(\theta; \mathcal{Y}(\boldsymbol{\theta}_n)) \equiv \arg\min_{m \in \mathcal{P}_p} \sum_{i=1}^{n} \left(m(\theta^{(i)}) - y^{(i)} \right)^2 \mathcal{K}(\theta^{(i)}, \theta)$$

  - Locally supported kernel $\mathcal{K}(\cdot, \theta)$

Local polynomial surrogates

- Given **exact** density evaluations $y_i = \pi(\theta)$

$$\mathcal{Y}(\boldsymbol{\theta}_n) \equiv \{(\theta_1, y_1), \ldots (\theta_n, y_n)\}$$

- Find the degree p polynomial $h_n(\theta; \mathcal{Y}(\boldsymbol{\theta}_n))$ such that

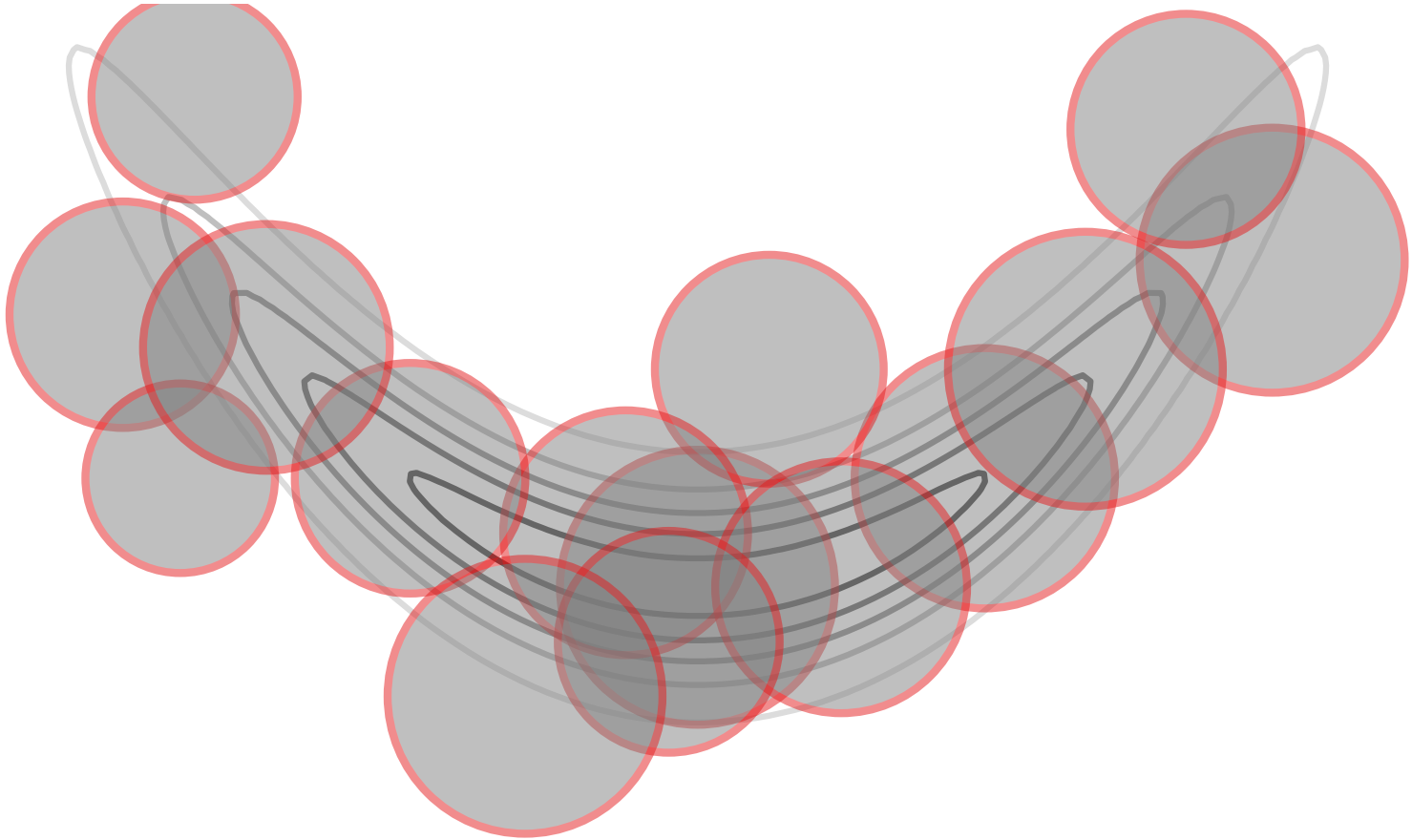
$$h_n(\theta; \mathcal{Y}(\boldsymbol{\theta}_n)) \equiv \arg\min_{m \in \mathcal{P}_p} \sum_{i=1}^{n} \left(m(\theta^{(i)}) - y^{(i)} \right)^2 \mathcal{K}(\theta^{(i)}, \theta)$$

  - Locally supported kernel $\mathcal{K}(\cdot, \theta)$

- **Intuition:** minimize the weighted least squares difference between the surrogate and the k_n nearest neighbors

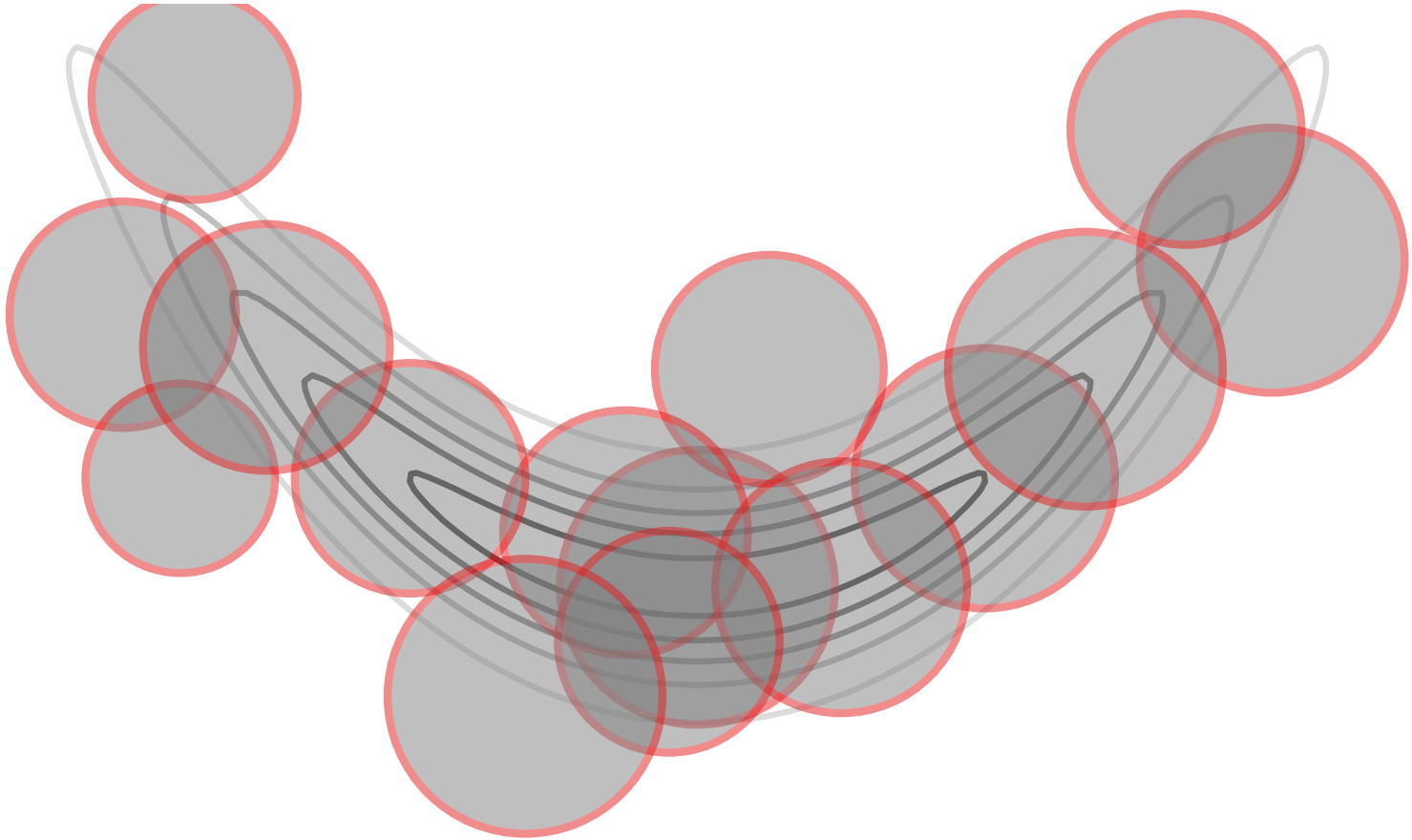
Local polynomial surrogates

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



Local polynomial surrogates

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

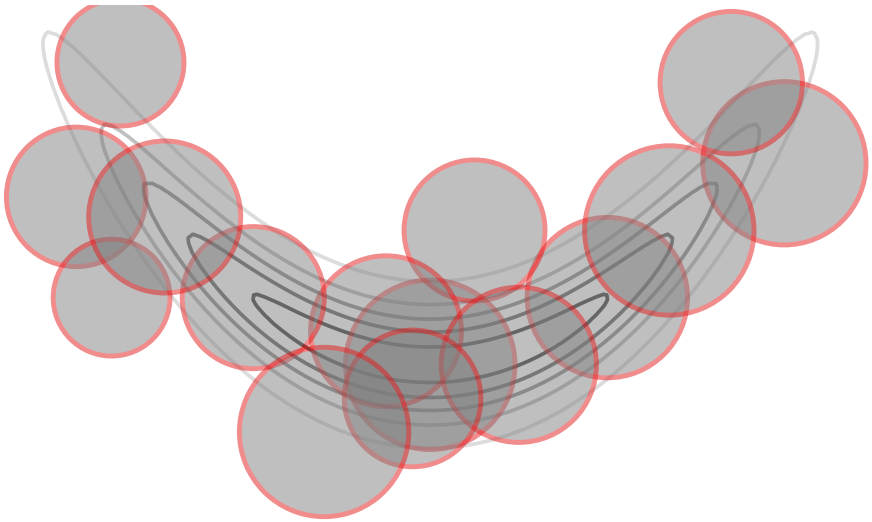


Ball size is determined by the prescribed **number of nearest neighbors**

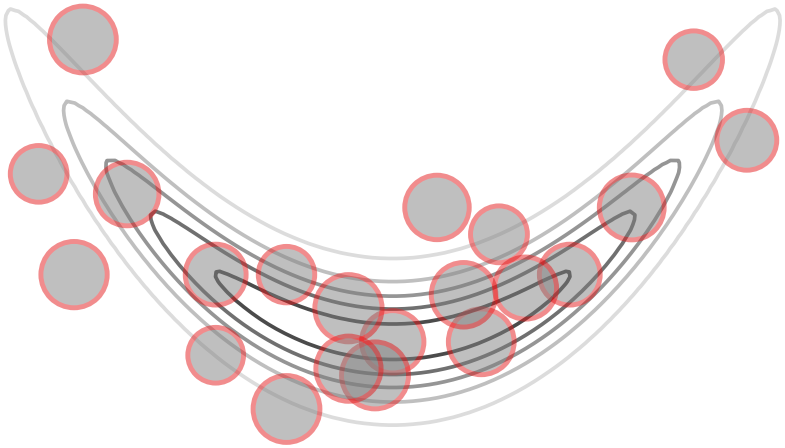
Surrogate convergence

The local polynomial approximation becomes exact as:

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



Large balls

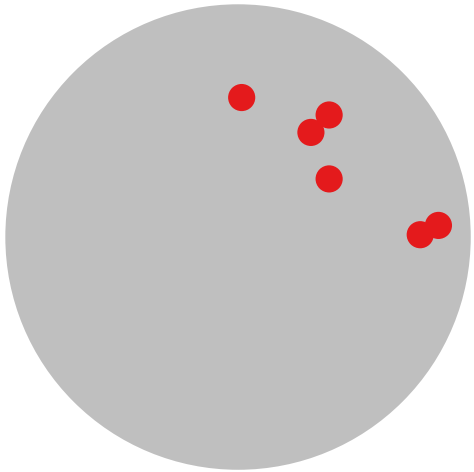


Small balls

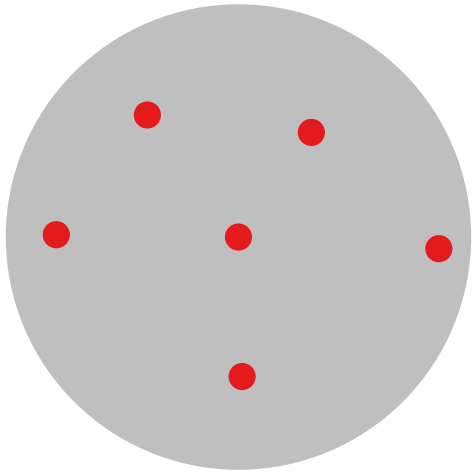
Surrogate convergence

The local polynomial approximation becomes exact 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

Markov chain Monte Carlo (MCMC) overview

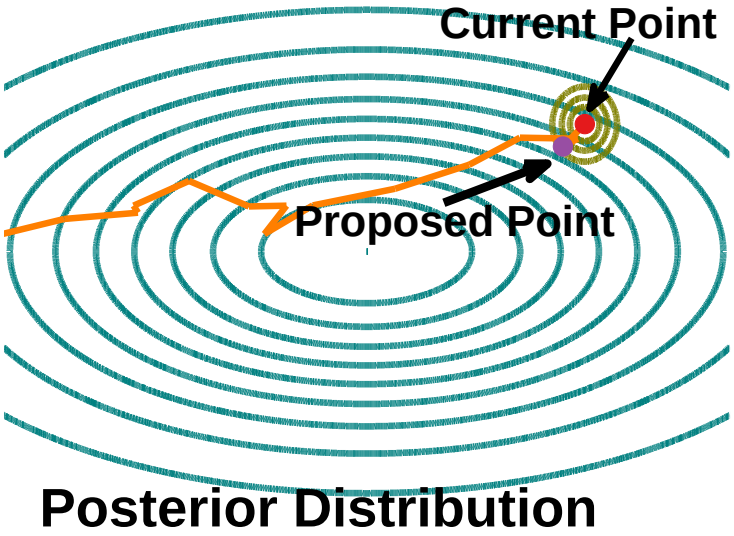
Three step algorithm:

1. *Propose* $\theta' \sim p(\cdot|\theta_t)$
2. *Acceptance probability*

$$\alpha \equiv \min\left(1, \frac{\pi(\theta')p(\theta_t|\theta')}{\pi(\theta_t)p(\theta'|\theta_t)}\right)$$

3. *Accept/reject*

$$\theta_{t+1} = \begin{cases} \theta' & \text{with probability } \alpha \\ \theta_t & \text{else} \end{cases}$$



Metropolis et al., 1953

Hastings, 1970

and variations . . .

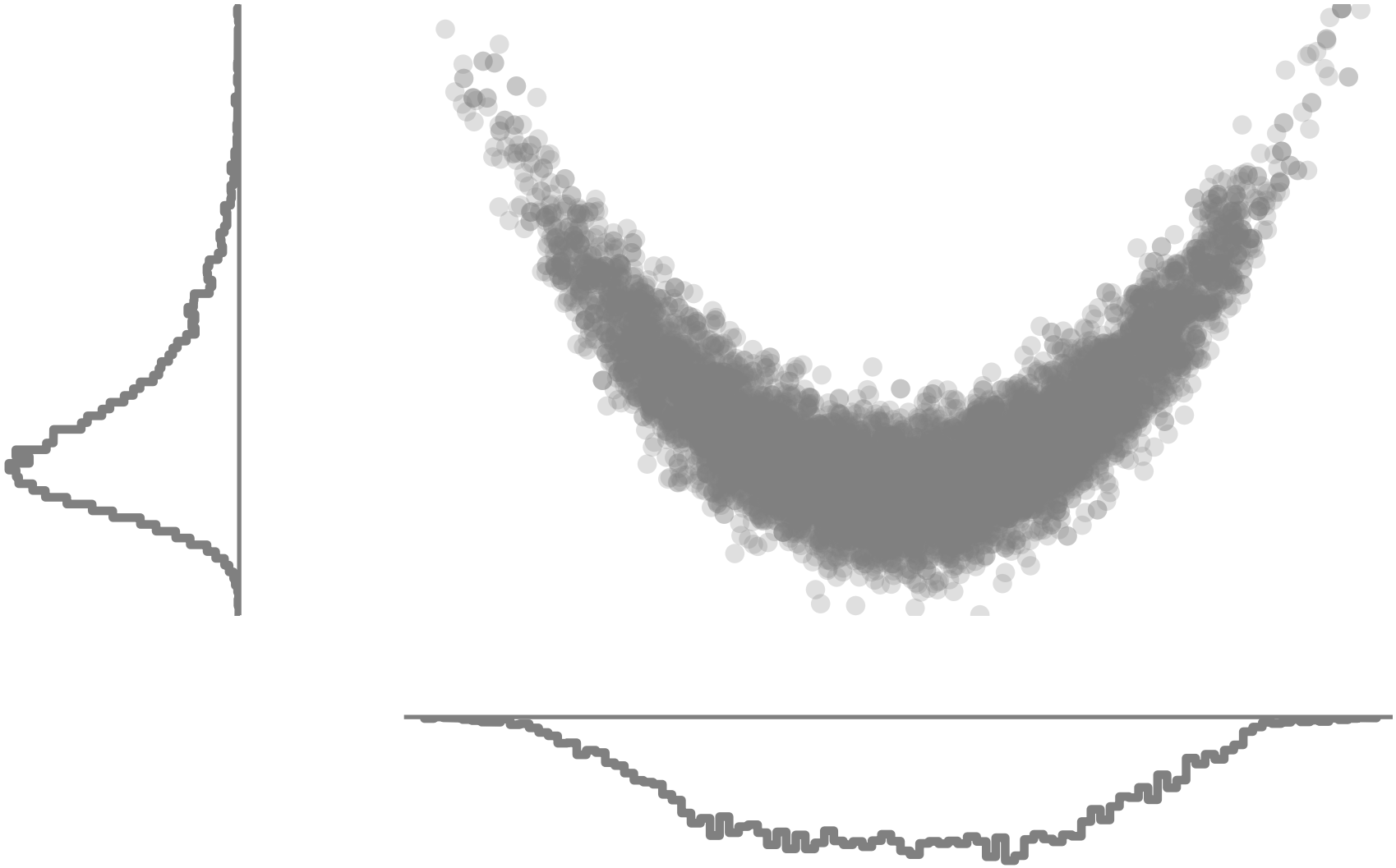
Haario et al., 2006

Parno and Marzouk, 2014

Brooks et al., 2011

MCMC with exact evaluations

We correctly characterize the distribution π and its marginals



MCMC with local approximations

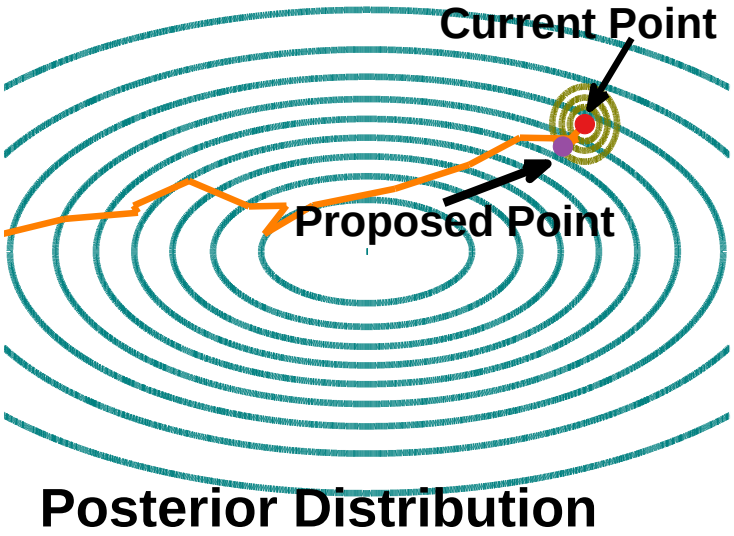
Four step algorithm:

1. *Propose* $\theta' \sim p(\cdot|\theta_t)$
2. *Refine* $\hat{\pi}(\cdot)$ near θ_t and θ'
3. *Acceptance probability*

$$\alpha \equiv \min\left(1, \frac{\hat{\pi}(\theta')p(\theta_t|\theta')}{\hat{\pi}(\theta_t)p(\theta'|\theta_t)}\right)$$

4. *Accept/reject*

$$\theta_{t+1} = \begin{cases} \theta' & \text{with probability } \alpha \\ \theta_t & \text{else} \end{cases}$$



MCMC with local approximations

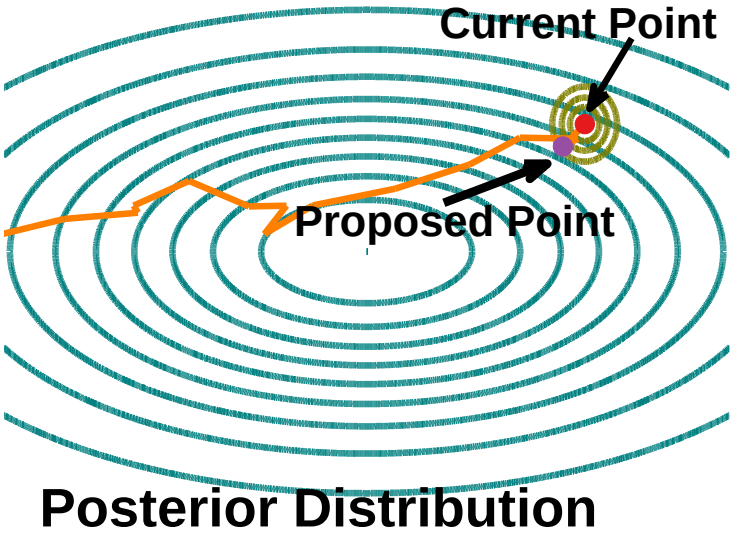
Four step algorithm:

1. *Propose* $\theta' \sim p(\cdot|\theta_t)$
2. *Refine* $\hat{\pi}(\cdot)$ near θ_t and θ'
3. *Acceptance probability*

$$\alpha \equiv \min\left(1, \frac{\hat{\pi}(\theta')p(\theta_t|\theta')}{\hat{\pi}(\theta_t)p(\theta'|\theta_t)}\right)$$

4. *Accept/reject*

$$\theta_{t+1} = \begin{cases} \theta' & \text{with probability } \alpha \\ \theta_t & \text{else} \end{cases}$$

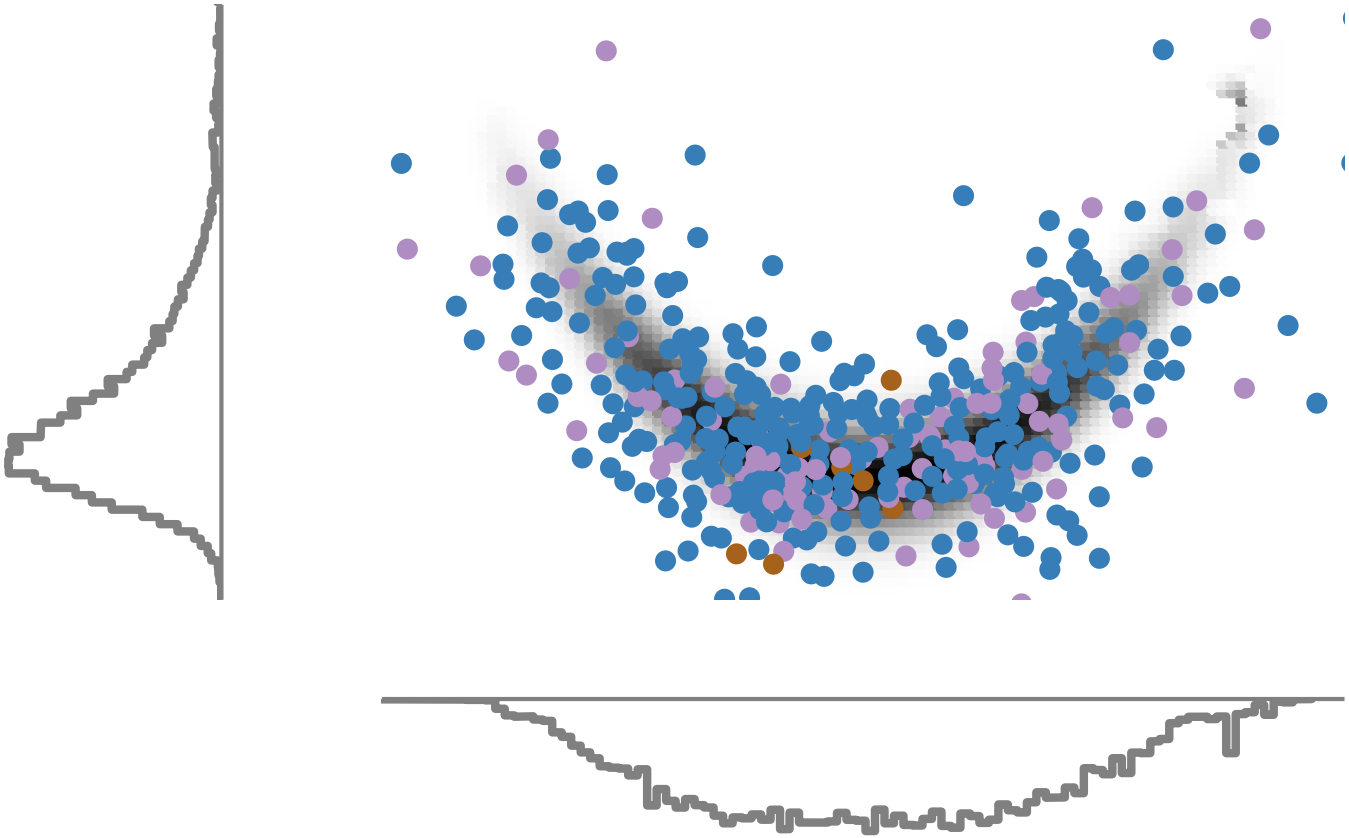


- **Refine** by evaluating the true density $\pi(\cdot)$ at a carefully chosen point near θ_t or θ'

Refinement strategy (random refinement)

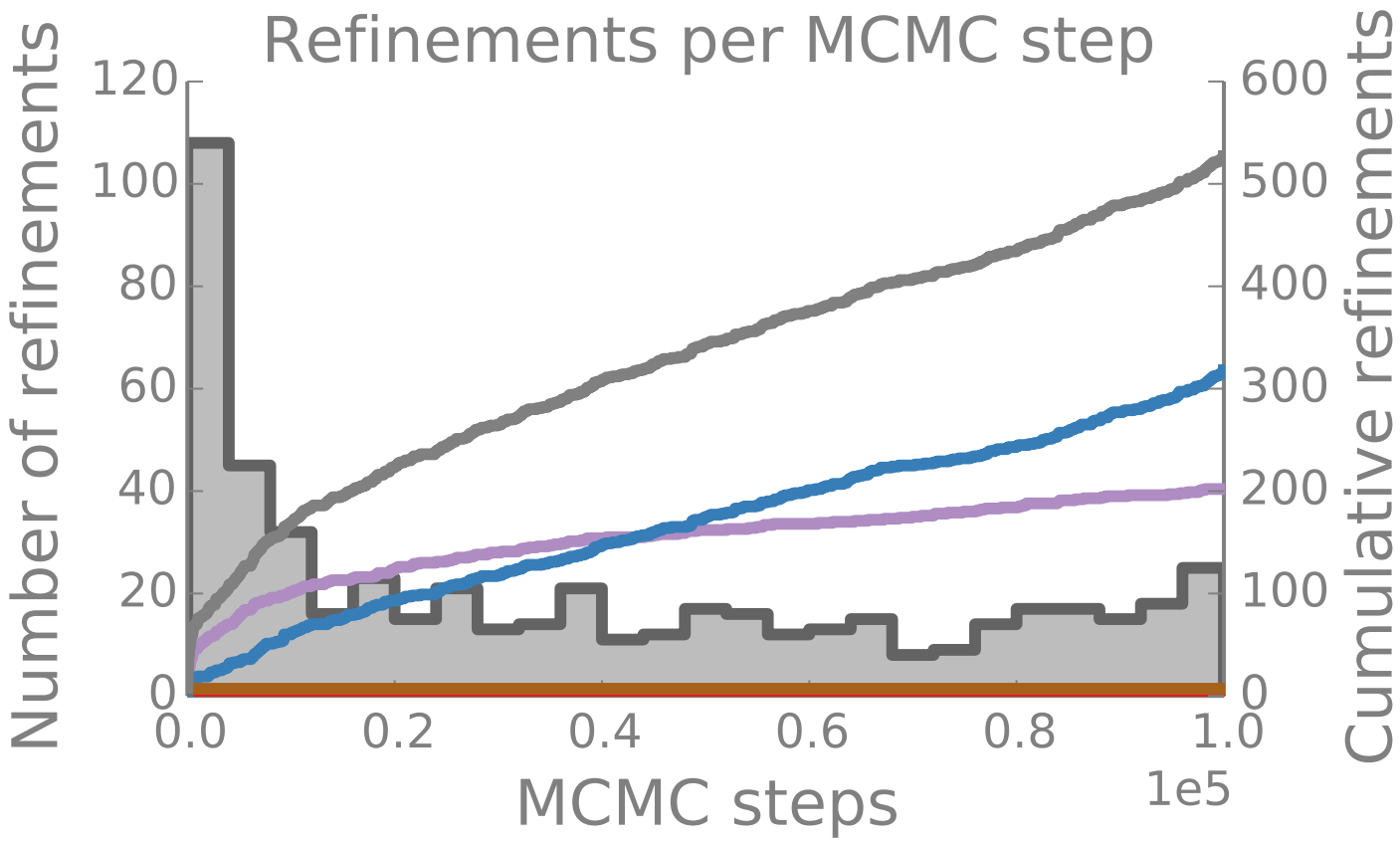
At θ_t and θ', refine the surrogate under two conditions:

1. With decaying probability $\beta_1 t^{-\beta_0}$
2. If the poisedness constant is large $\Lambda > \Lambda_T$



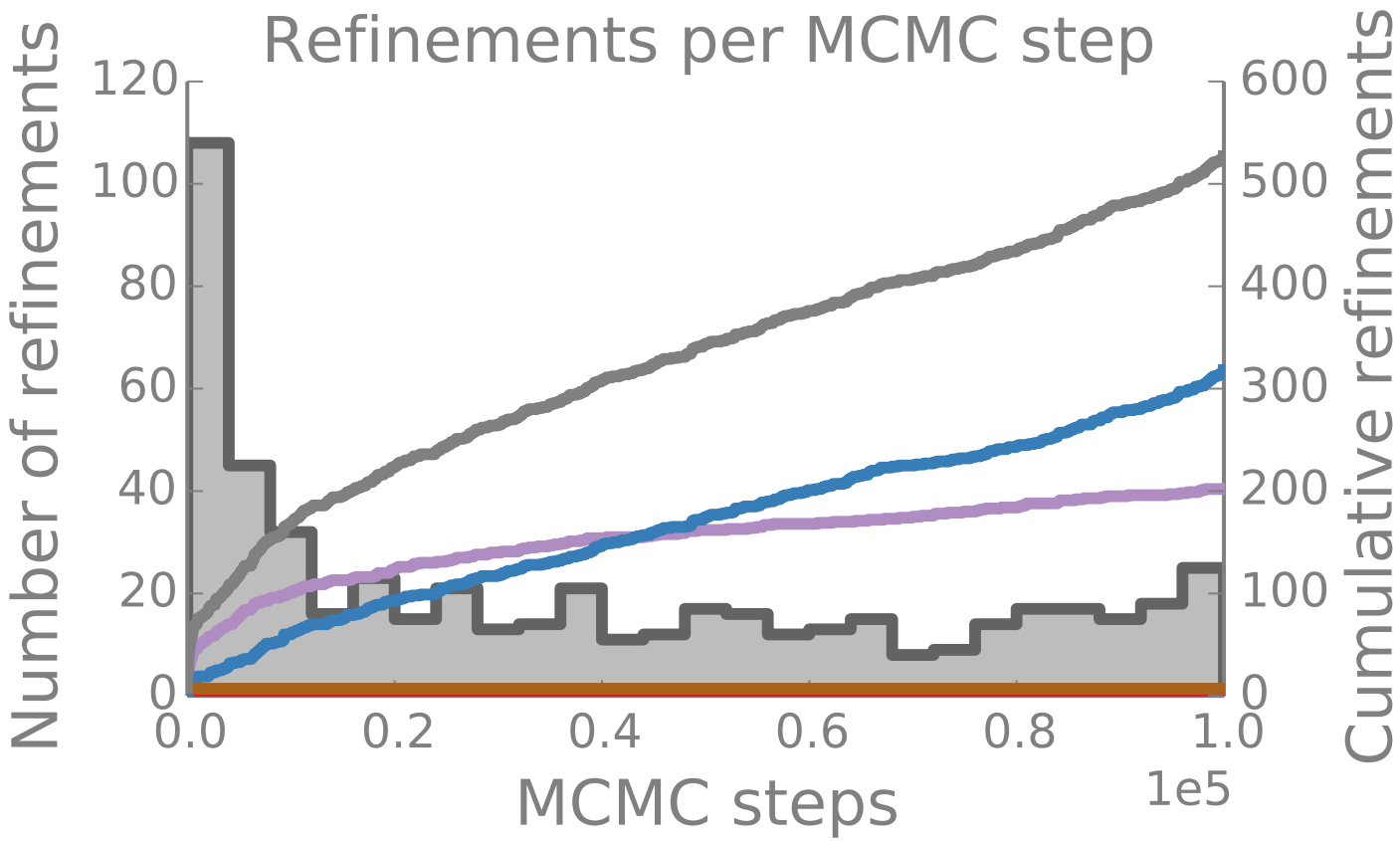
Refinement strategy (random refinement)

- Refinement frequency decays as MCMC progresses

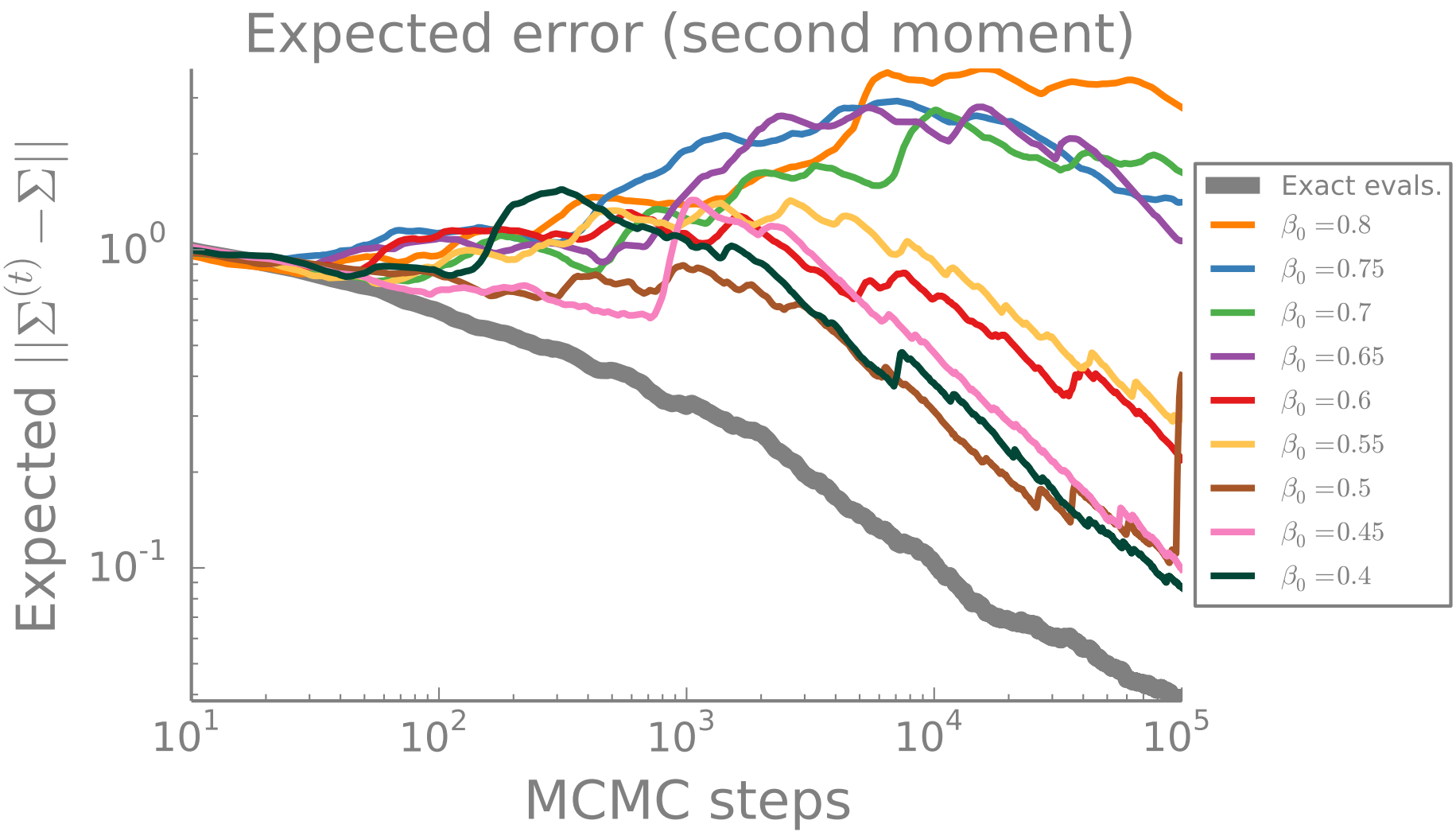


Refinement strategy (random refinement)

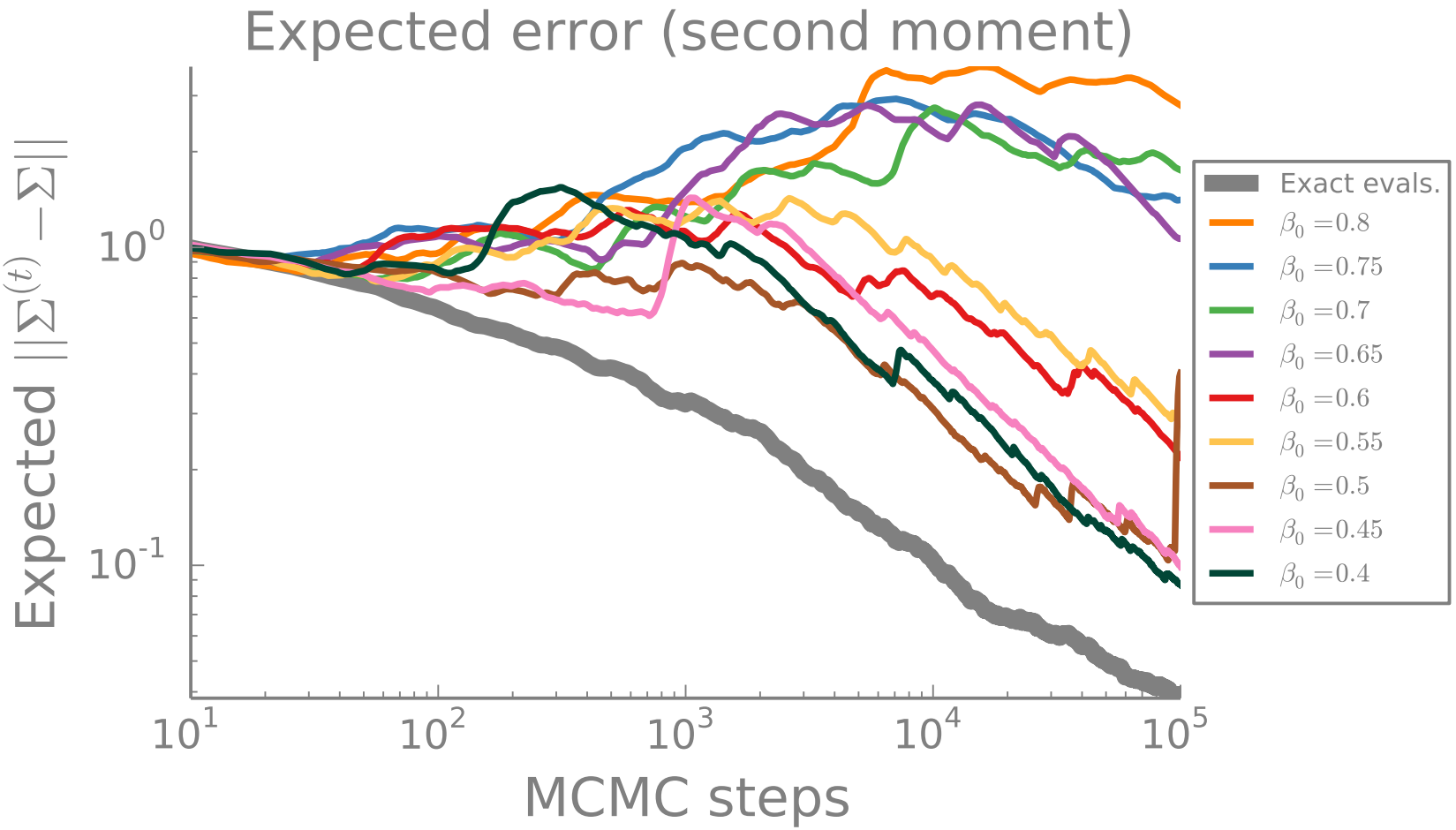
- Refinement frequency decays as MCMC progresses
- How quickly should the refinement frequency decay?



Expected error (random refinement)

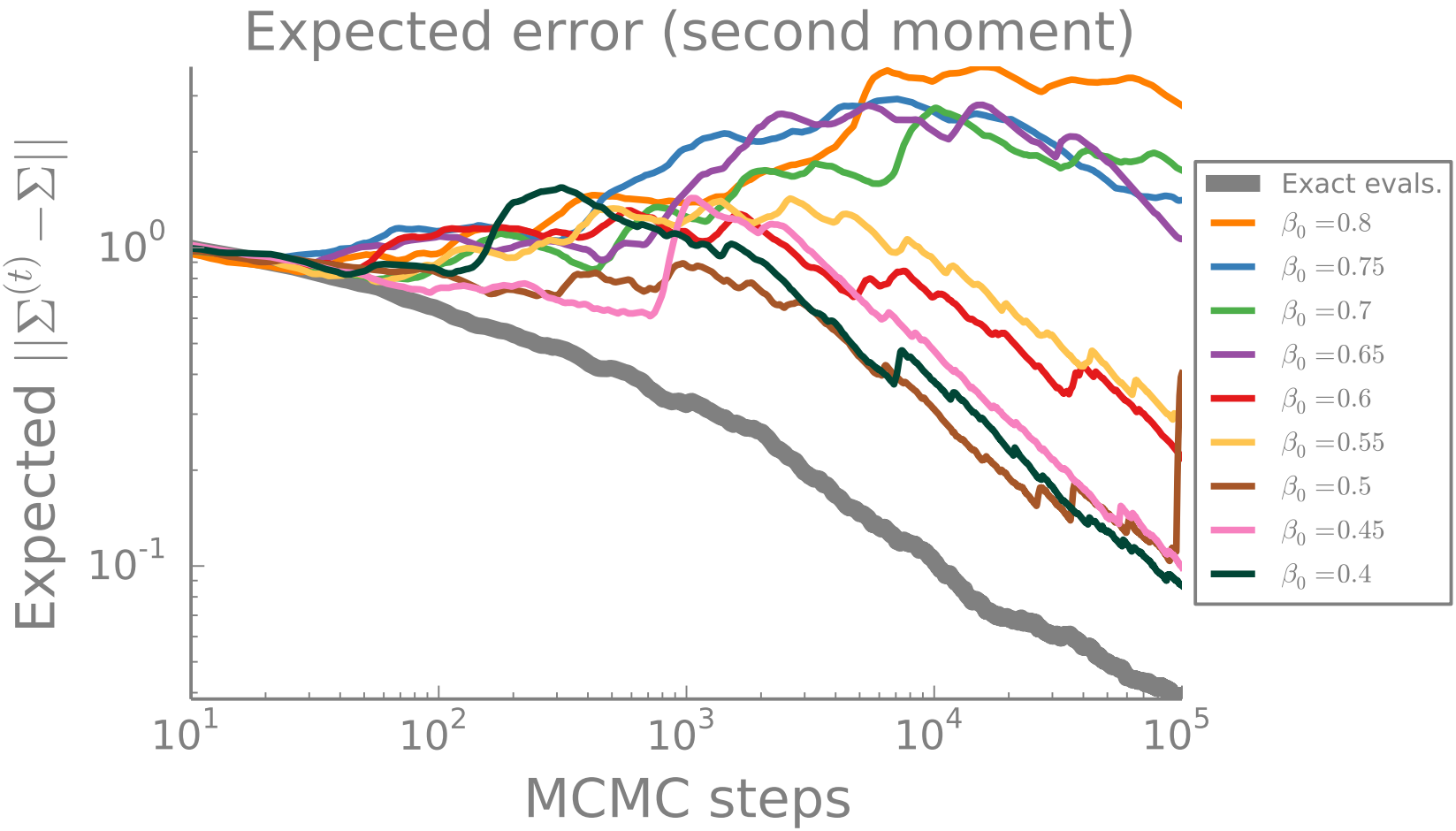


Expected error (random refinement)



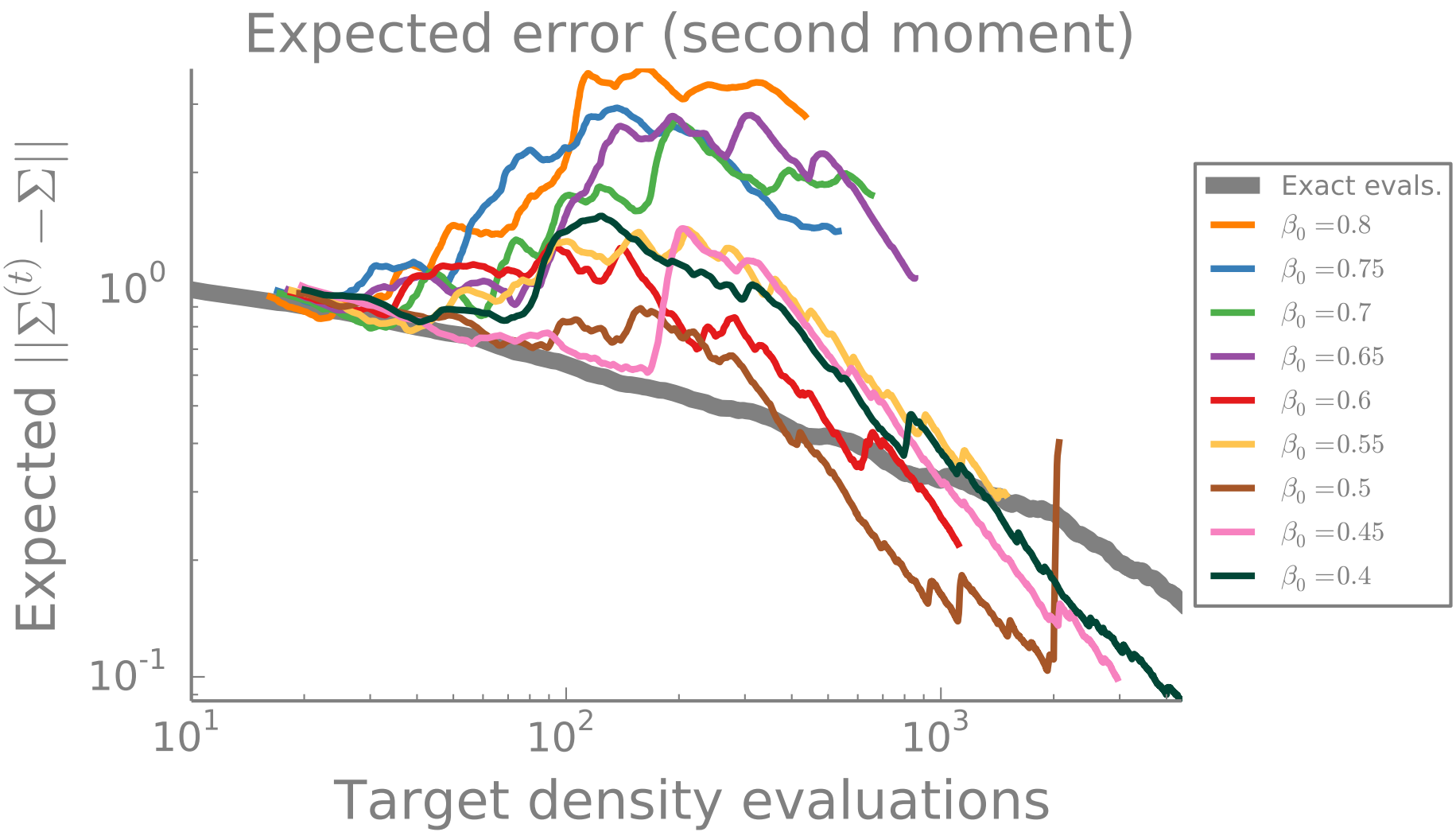
- MCMC with exact evaluations provides a lower bound for expected error decay

Expected error (random refinement)

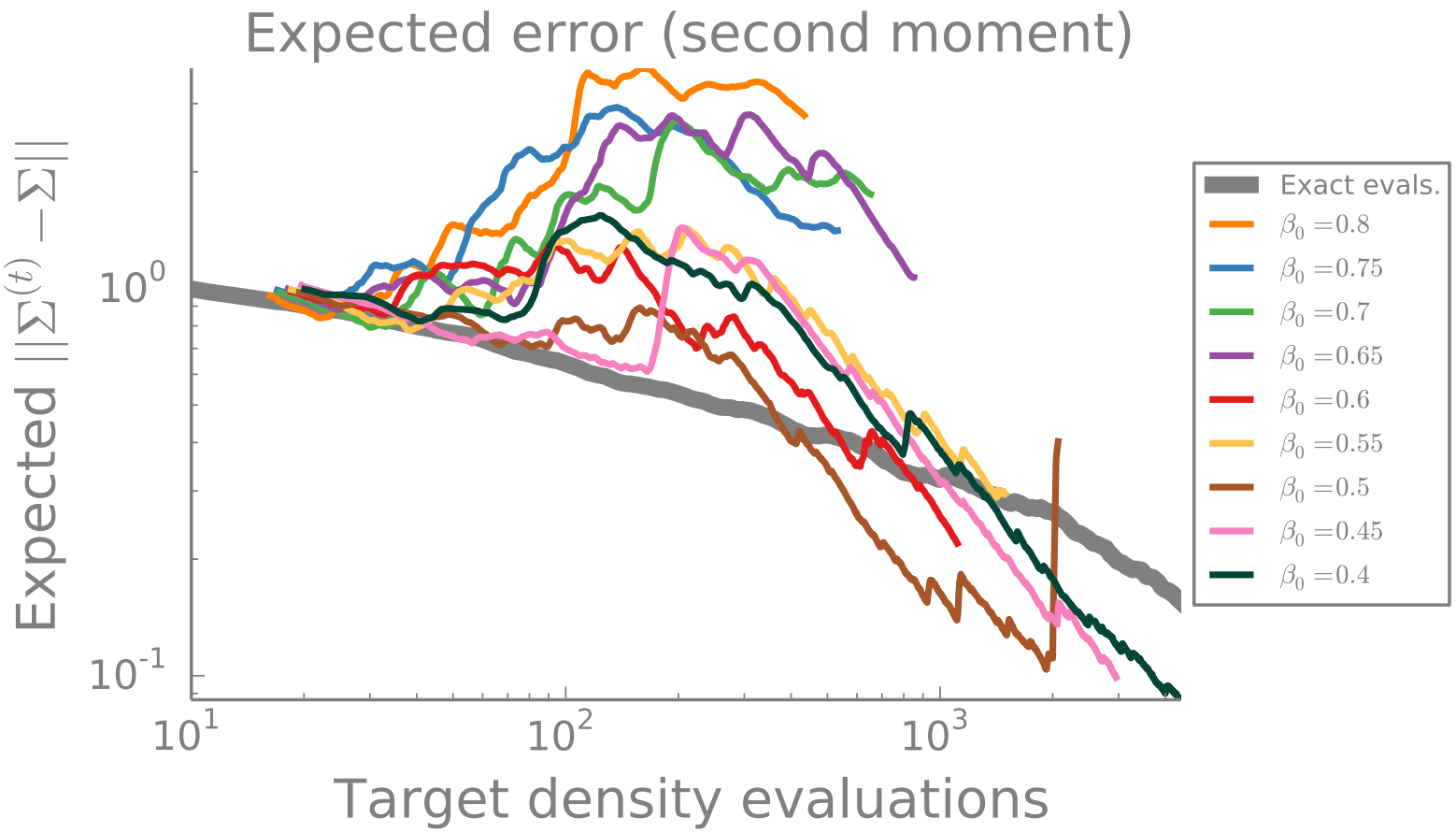


- MCMC with exact evaluations provides a lower bound for expected error decay
- Additional error is incurred because of the surrogate model

Expected error (random refinement)

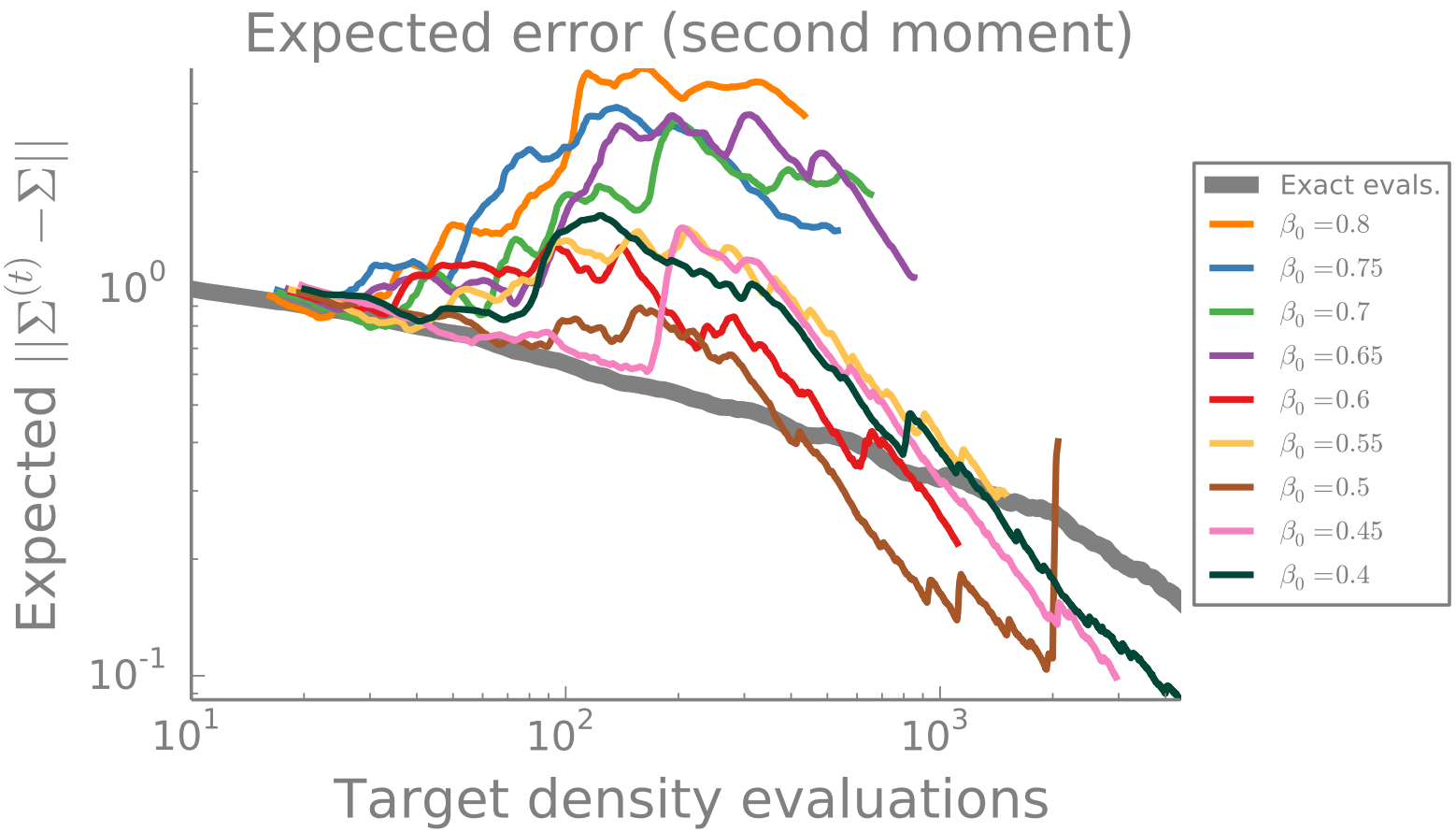


Expected error (random refinement)



- Converges more quickly, in terms of error per # of density evaluations

Expected error (random refinement)



- Converges more quickly, in terms of error per # of density evaluations
- Refinement of the surrogate model is beneficial — up to a point

Balancing Monte Carlo variance and surrogate error

- Monte Carlo variance estimate

$$[\text{Monte Carlo variance}] \leq C_{MC}\, t^{-1}$$

Balancing Monte Carlo variance and surrogate error

- Monte Carlo variance estimate

$$[\text{Monte Carlo variance}] \leq C_{MC}\, t^{-1}$$

- Suppose we can sample from the surrogate density

$$\begin{aligned}
[\text{Structural error}] &= \|\mathbb{E}_{\hat{\pi}}[f(\theta)] - \mathbb{E}_{\pi}[f(\theta)]\| \\
&= \frac{1}{T}\left\| \sum_{t=1}^{\infty} f(\theta_t)(1 - w(\theta_t)) \right\|
\end{aligned}$$

Balancing Monte Carlo variance and surrogate error

- Monte Carlo variance estimate

$$[\text{Monte Carlo variance}] \leq C_{MC}\, t^{-1}$$

- Suppose we can sample from the surrogate density

$$\begin{aligned} [\text{Structural error}] &= \|\mathbb{E}_{\hat{\pi}}[f(\theta)] - \mathbb{E}_{\pi}[f(\theta)]\| \\ &= \frac{1}{T}\left\| \sum_{t=1}^{\infty} f(\theta_t)(1 - w(\theta_t)) \right\| \end{aligned}$$

- Ideally — $[\text{Structural error (bias)}]^2 \sim [\text{Monte Carlo variance}]$

Balancing Monte Carlo variance and surrogate error

- Monte Carlo variance estimate

$$[\text{Monte Carlo variance}] \leq C_{MC} t^{-1}$$

- Suppose we can sample from the surrogate density

$$\begin{aligned} [\text{Structural error}] &= \|\mathbb{E}_{\hat{\pi}}[f(\theta)] - \mathbb{E}_{\pi}[f(\theta)]\| \\ &= \frac{1}{T}\left\|\sum_{t=1}^{\infty} f(\theta_t)(1 - w(\theta_t))\right\| \end{aligned}$$

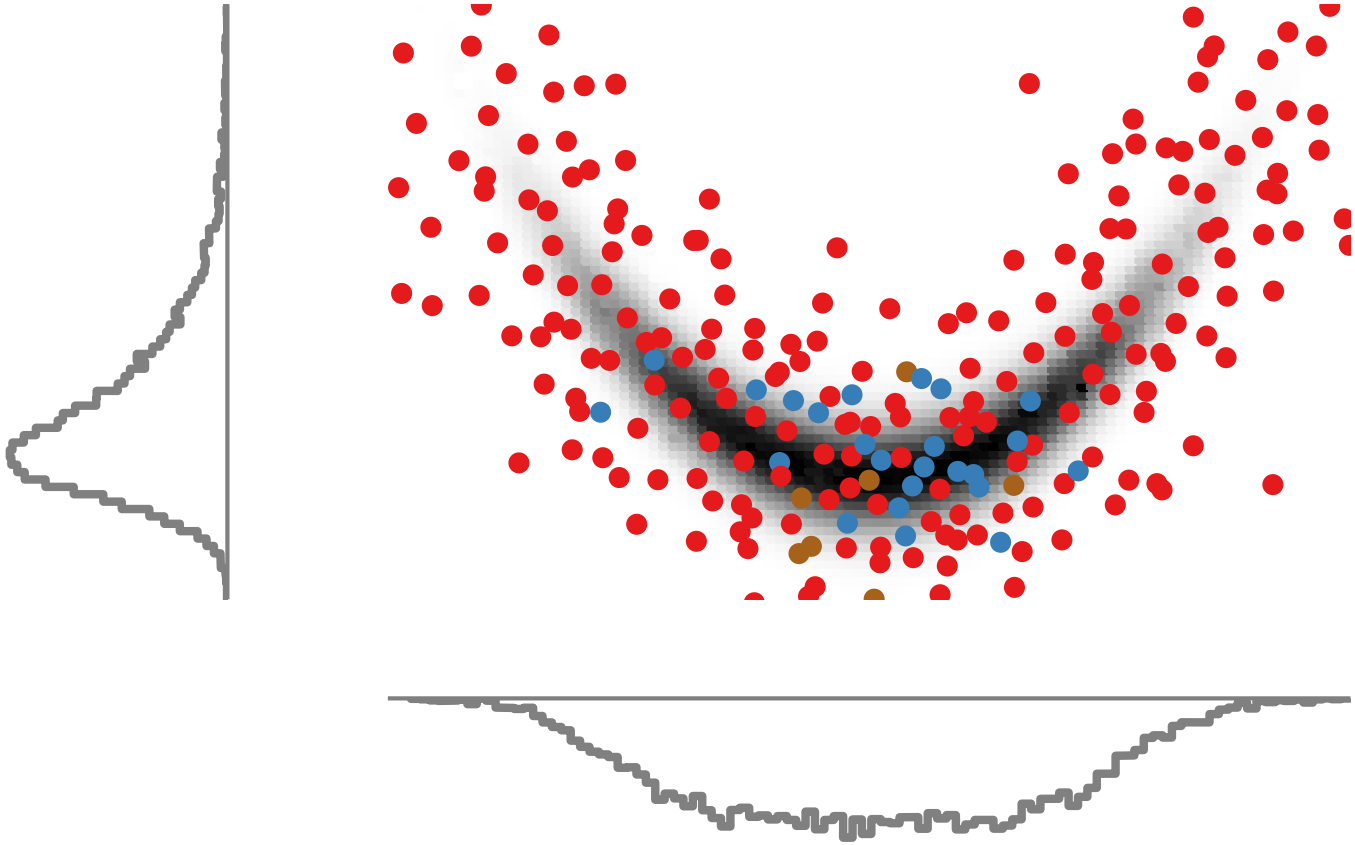
- Ideally — $[\text{Structural error (bias)}]^2 \sim [\text{Monte Carlo variance}]$
- Prescribe refinement strategy such that —

$$k_n \bar{\Lambda} \bar{\Delta}^{p+1} \sim \gamma_1 t^{-\gamma_0} \sim C_{MC} t^{-\frac{1}{4}}$$

Refinement strategy (structural refinement)

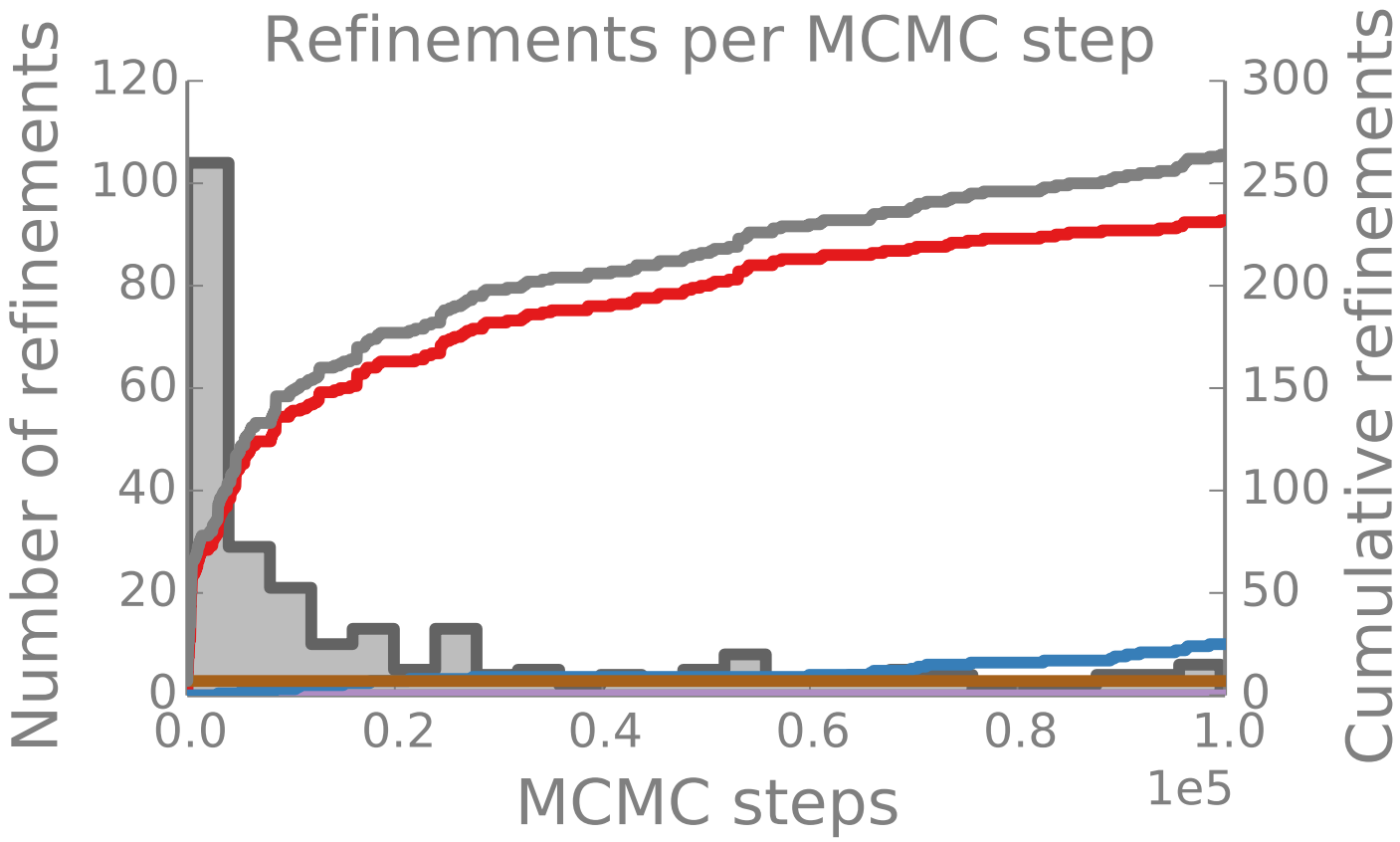
At θ_t, refine the surrogate model under two conditions:

1. With decaying error threshold $k_n \Lambda \Delta^{p+1} \sim q \Lambda_T \gamma_1 t^{-\gamma_0}$
2. If the poisedness constant is large $\Lambda > \frac{q}{k_n} \Lambda_T$



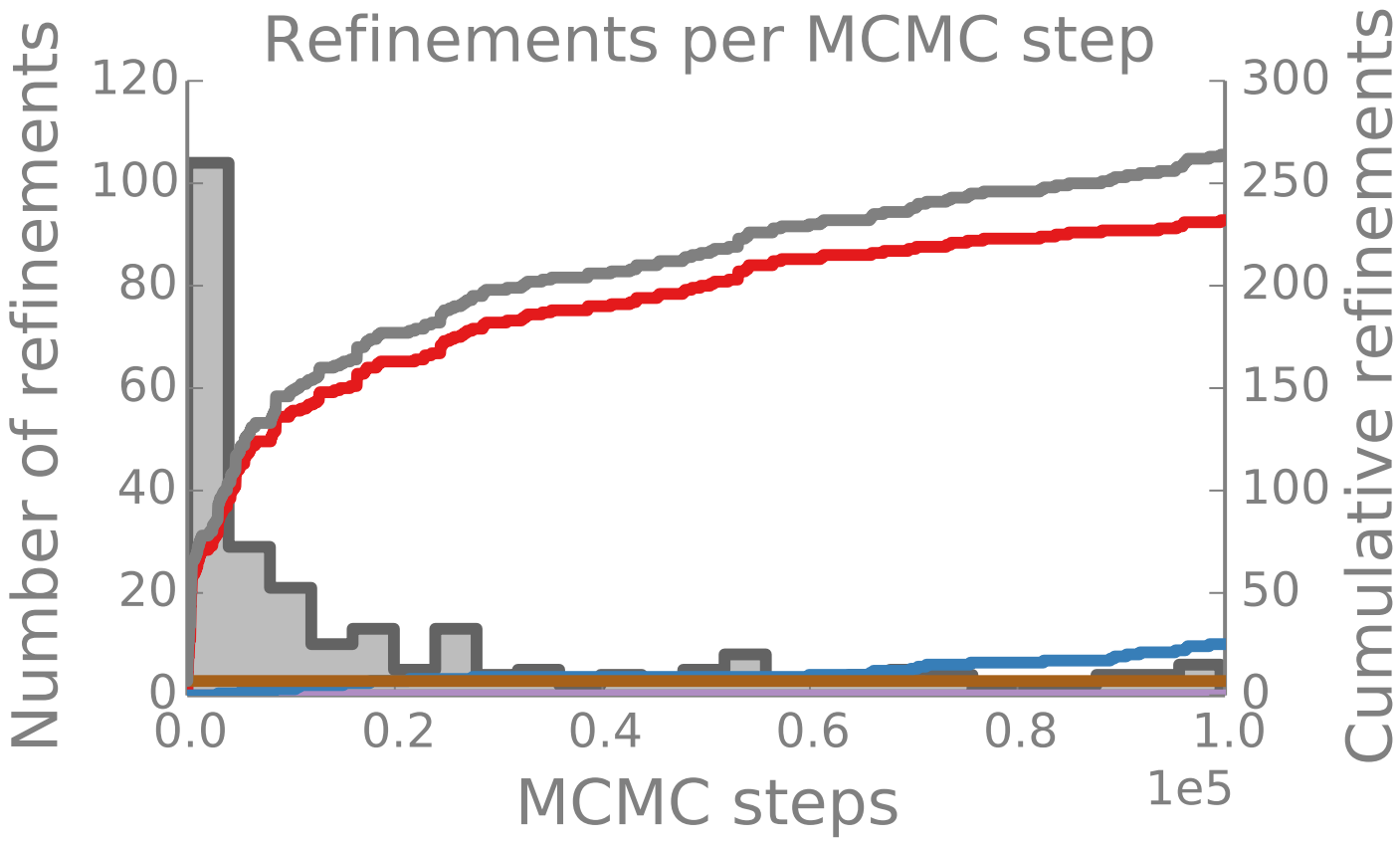
Refinement strategy (structural refinement)

- Refinement frequency decays as MCMC progresses

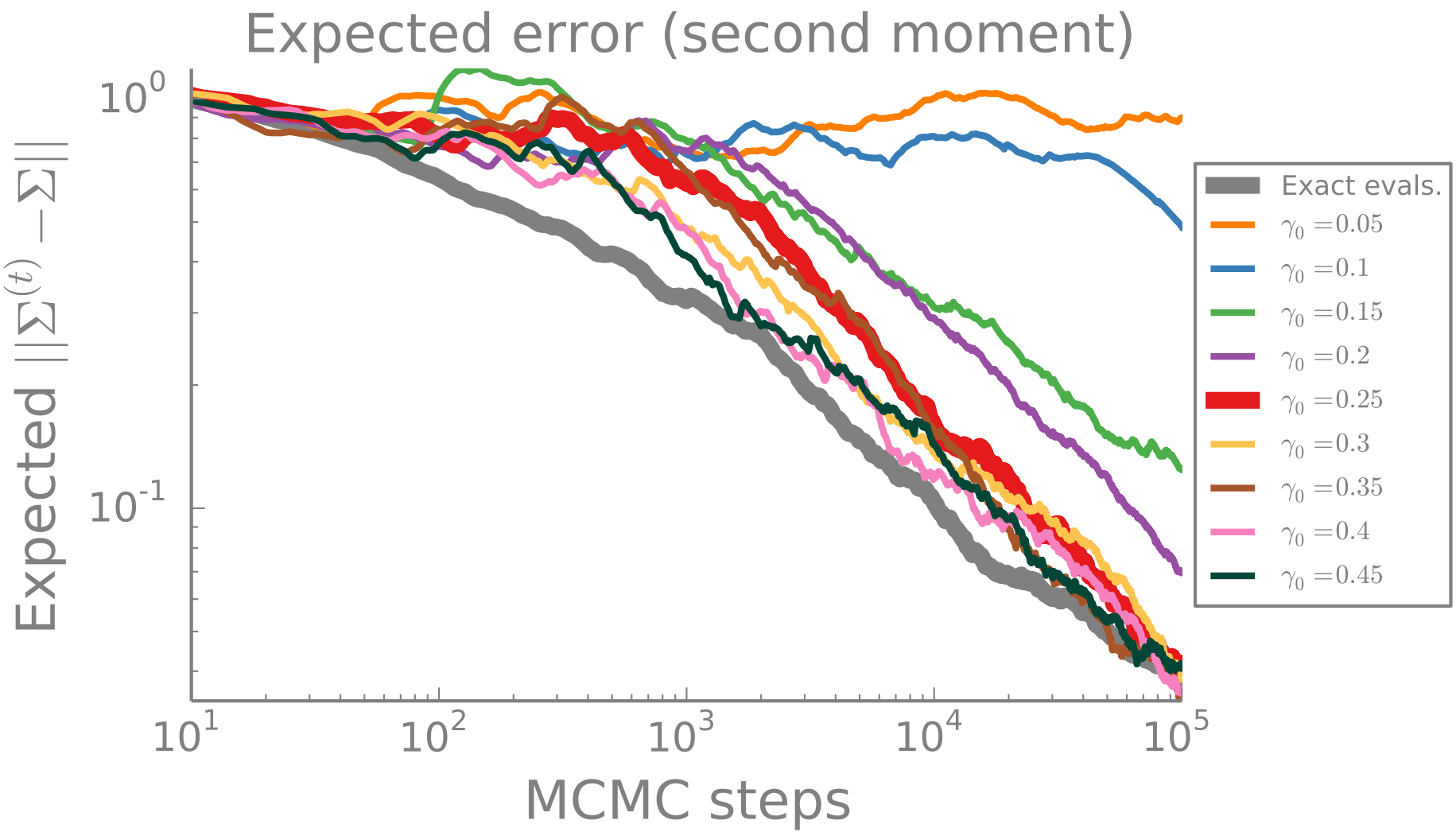


Refinement strategy (structural refinement)

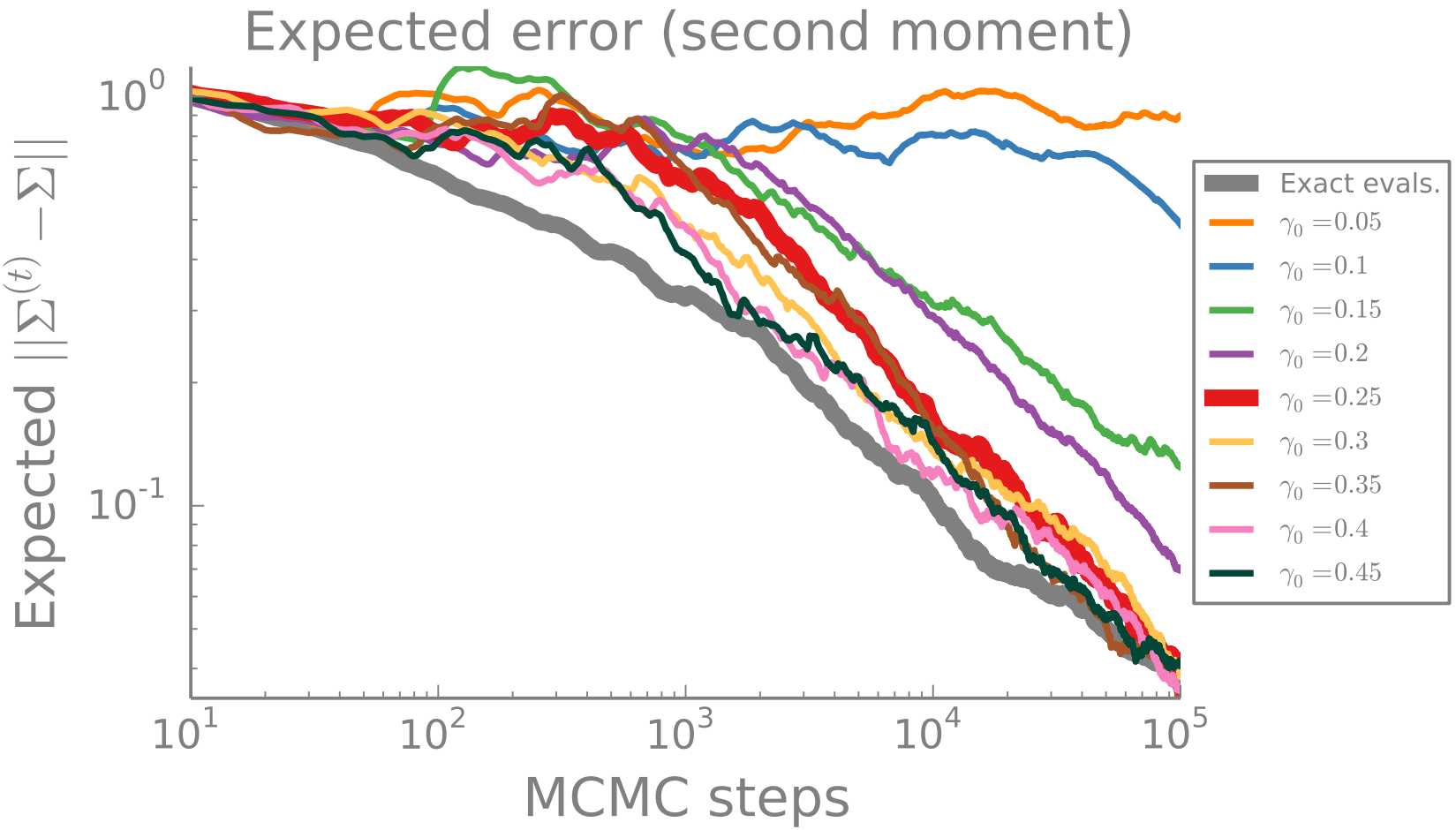
- Refinement frequency decays as MCMC progresses
- Ideally, the error threshold decays such that Monte Carlo variance and structural error balance



Expected error (structural refinement)

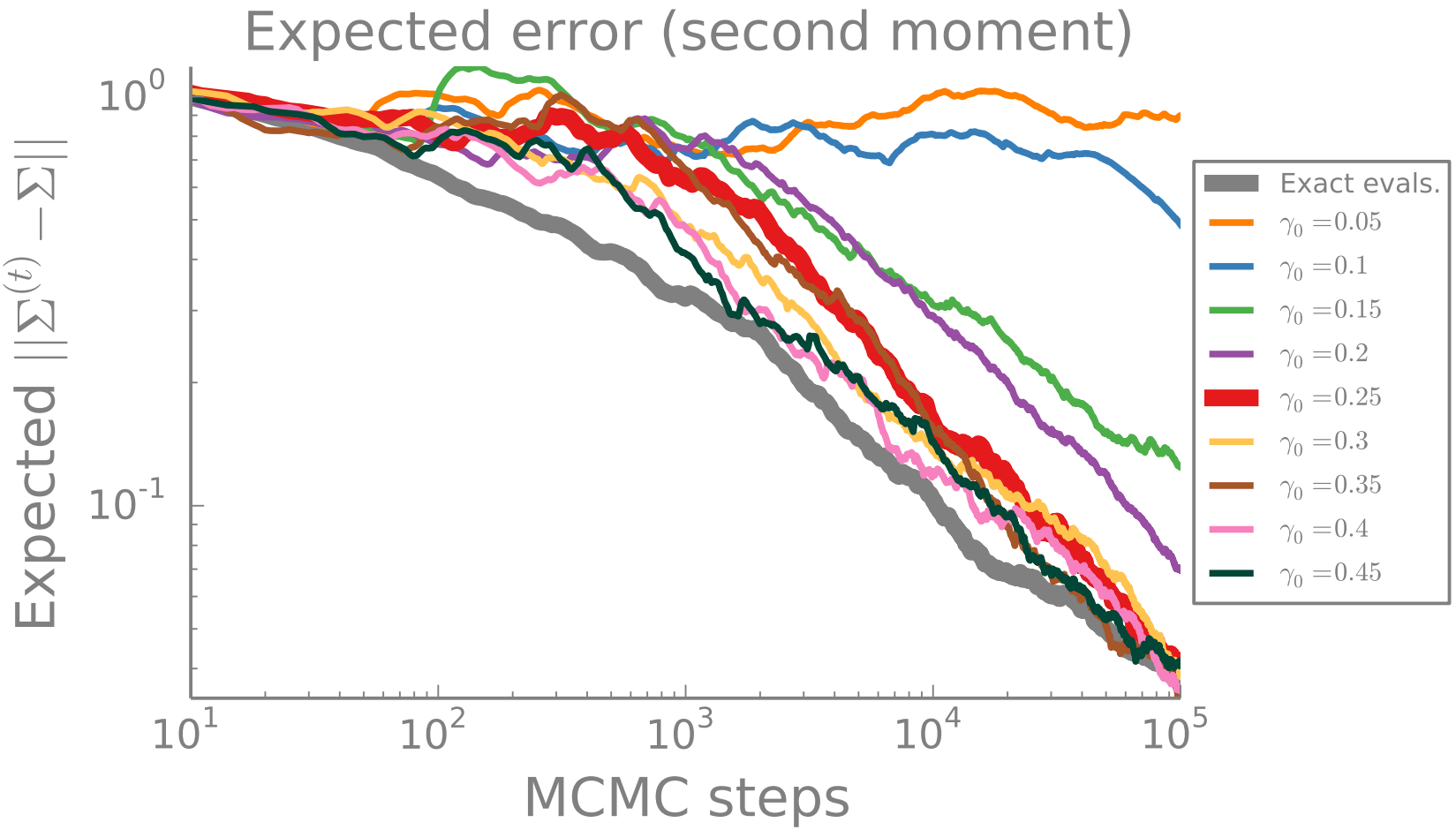


Expected error (structural refinement)



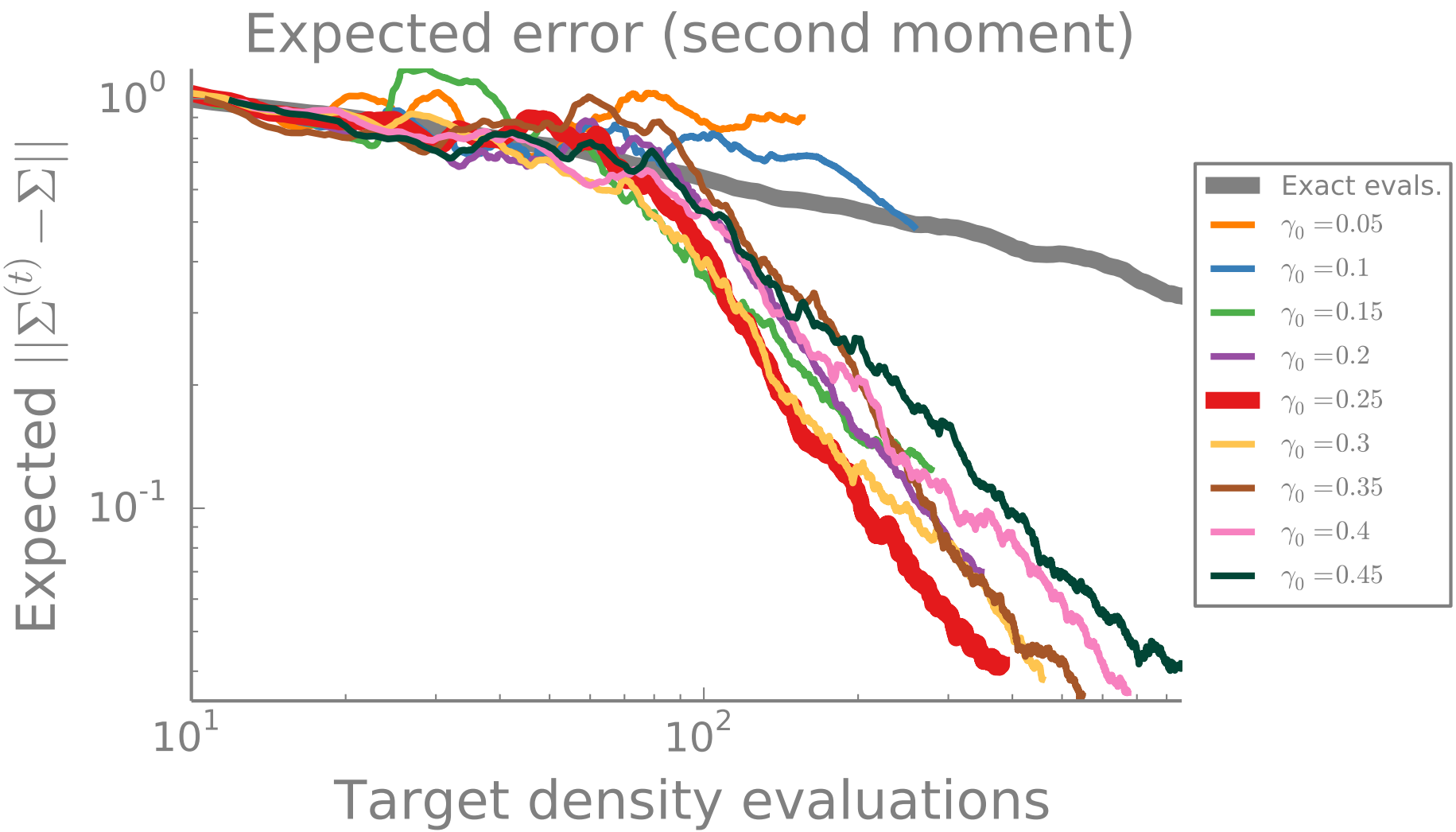
- **Infrequent refinement:** structural error dominates and error decays more slowly than with exact evaluations

Expected error (structural refinement)

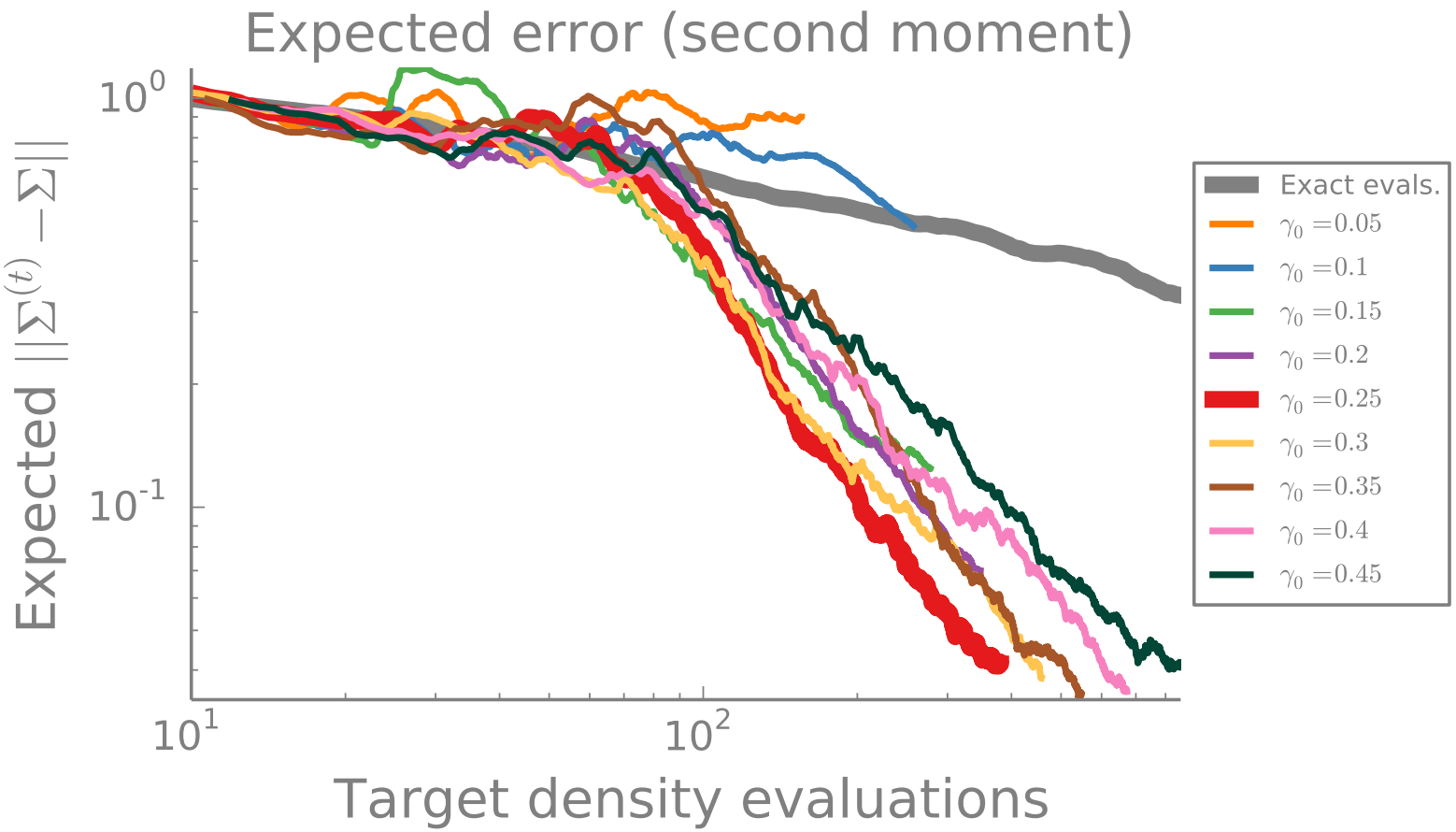


- **Infrequent refinement:** structural error dominates and error decays more slowly than with exact evaluations
- **Frequent refinement:** Monte Carlo variance dominates and error decays as if we had exact evaluations

Expected error (structural refinement)



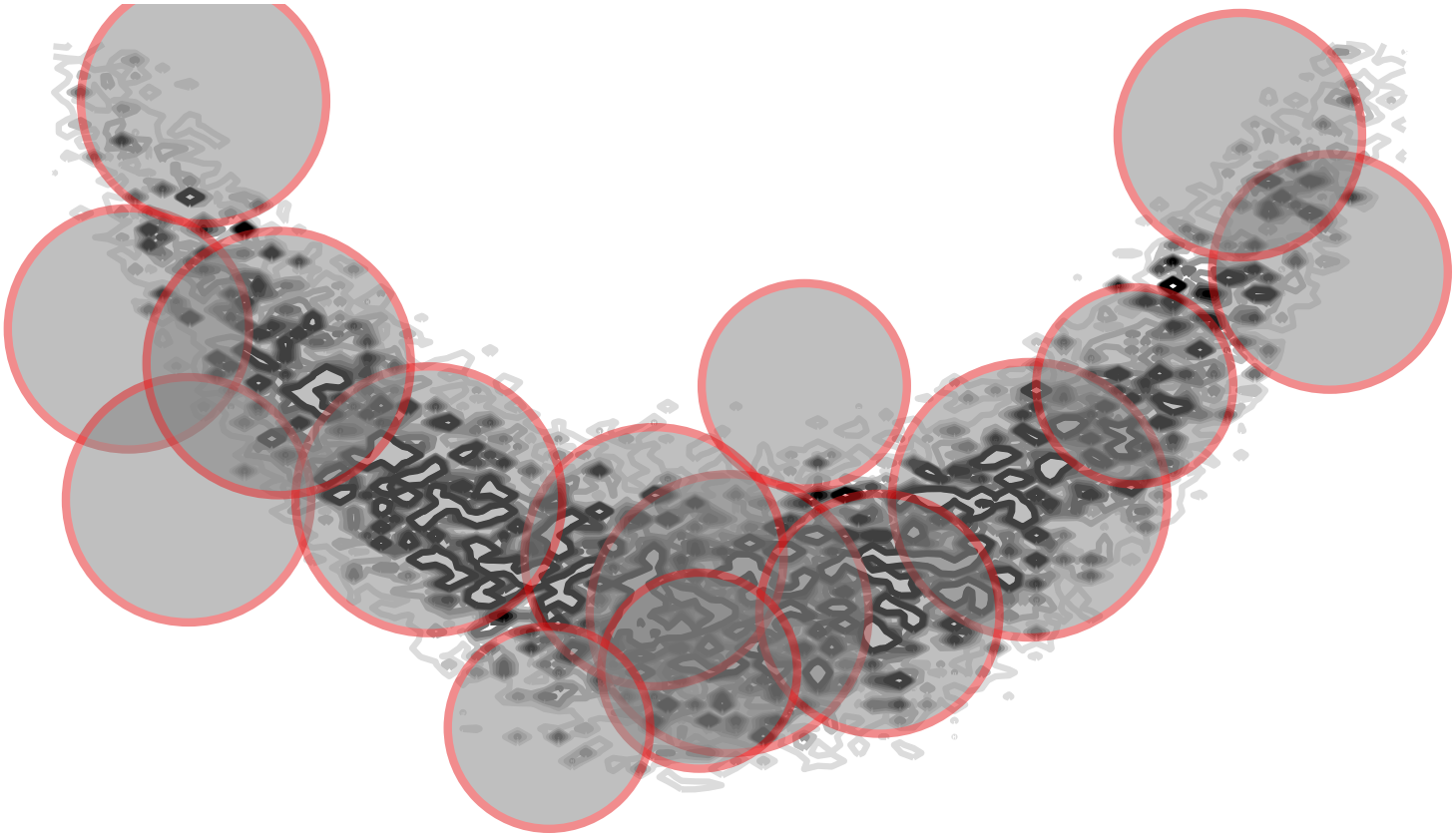
Expected error (structural refinement)



- **Optimal refinement strategy**: Balance Monte Carlo sampling variance with structural error; *fastest* decay of error with # density evaluations

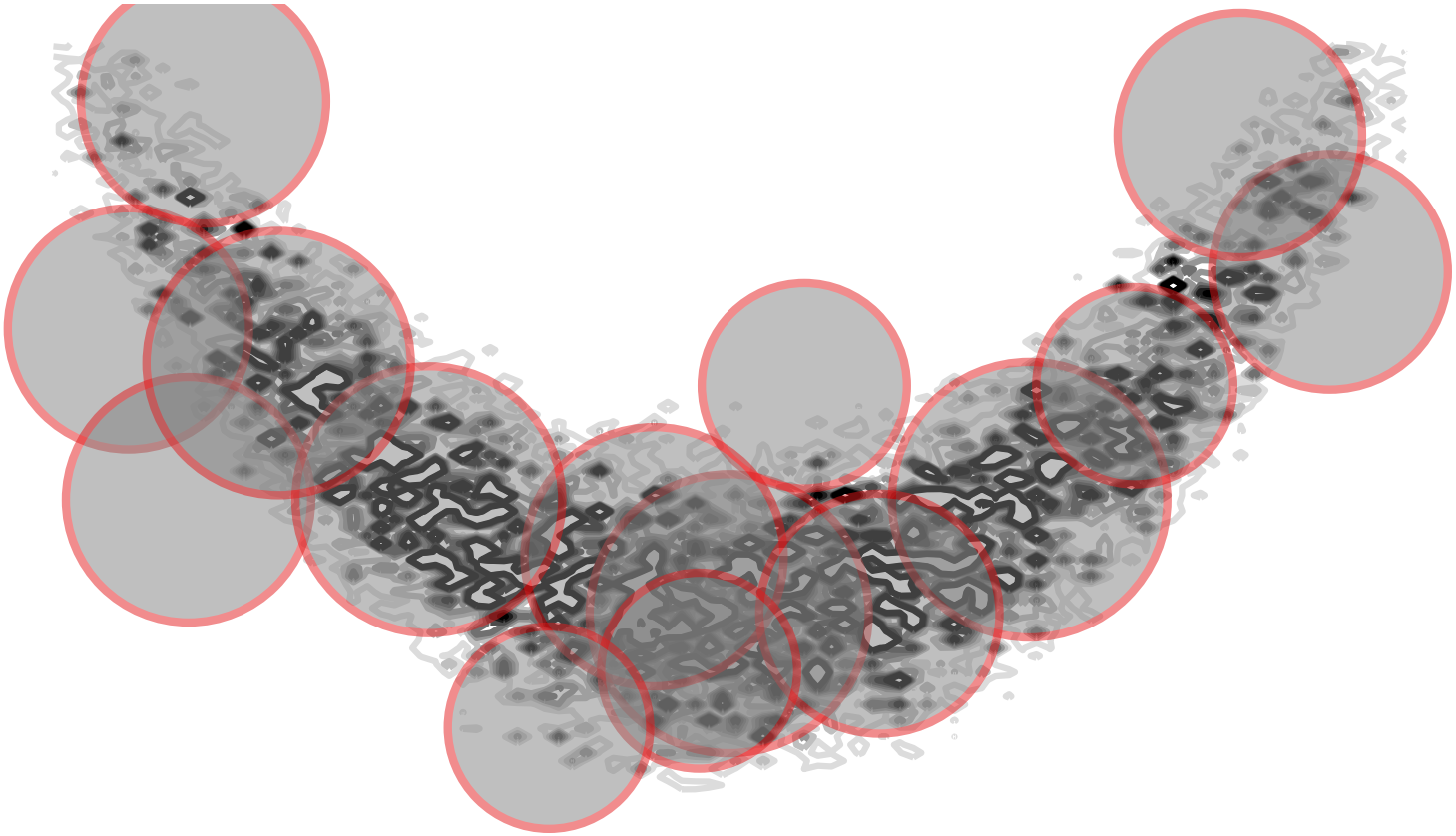
Local polynomial surrogates

- Now consider situations with **noisy evaluations** of the target density



Local polynomial surrogates

- Now consider situations with **noisy evaluations** of the target density

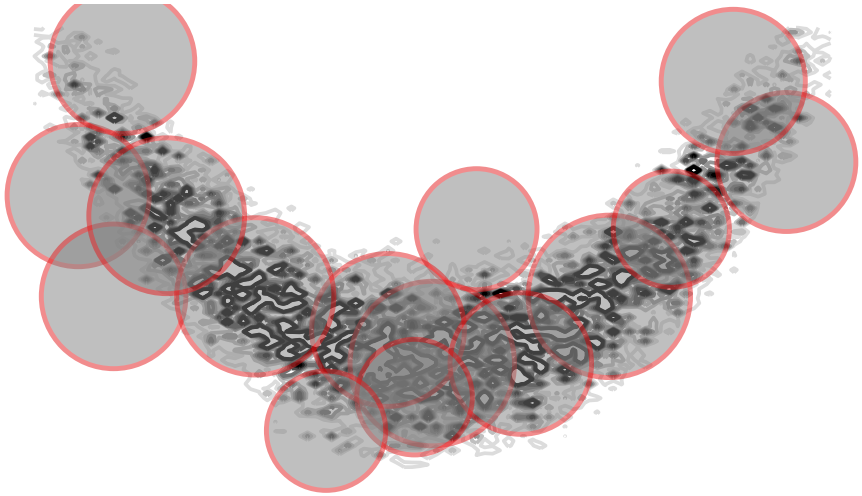


- Noise is due to marginalization along uninteresting coordinates, as in pseudo-marginal MCMC

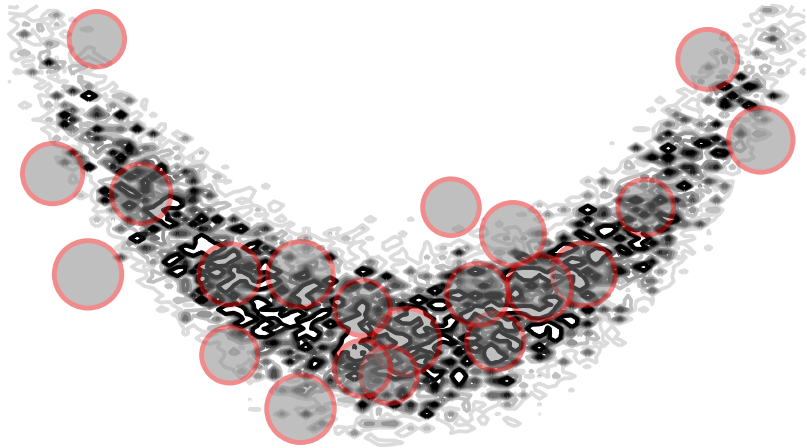
Surrogate convergence

The local polynomial approximation becomes exact as:

1. The ball size $\Delta \to 0$ (number of evaluations $n \to \infty$)
2. Λ-poisedness is maintained inside each ball
3. The number of nearest neighbors $k_n \to \infty$ (while $\frac{k_n}{n} \to 0$)



Large balls



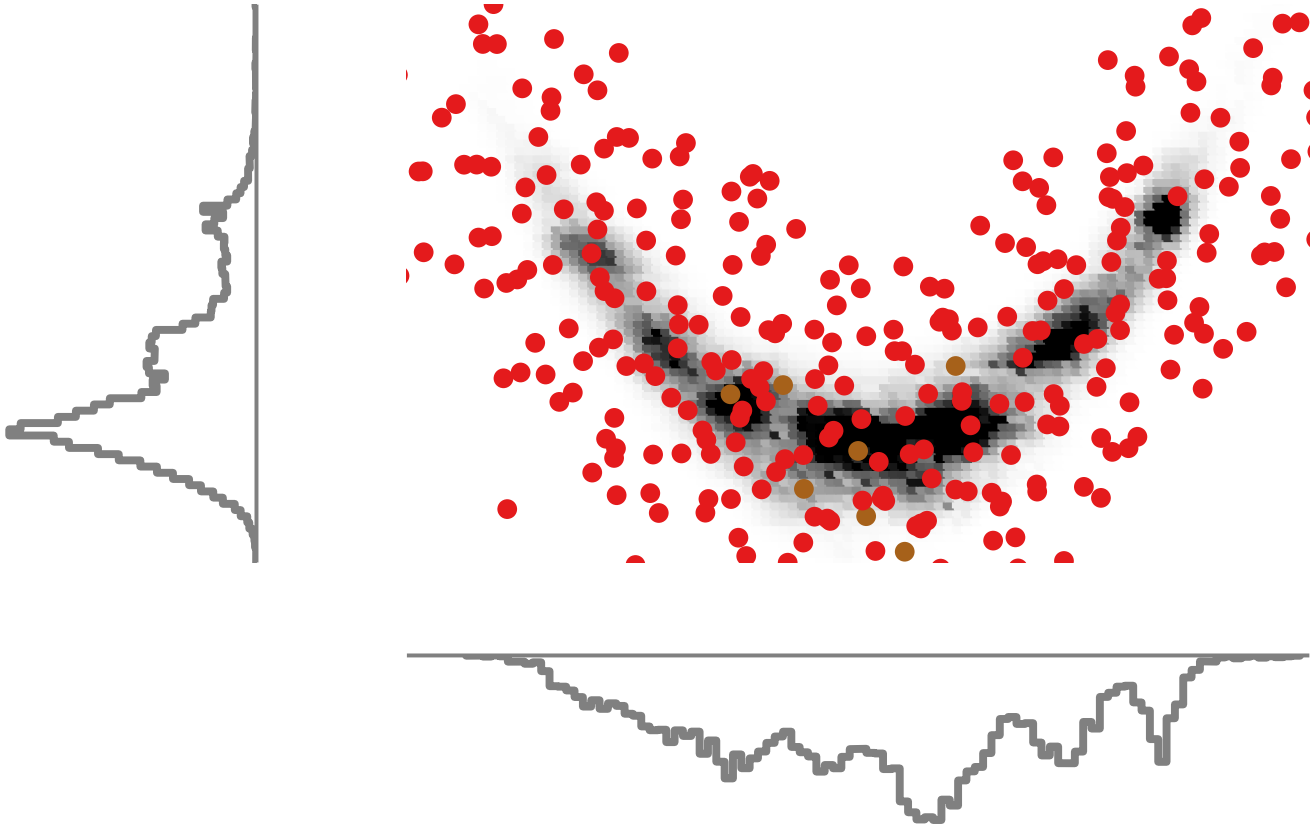
Small balls

Refinement strategy (structural refinement & noisy evaluations)

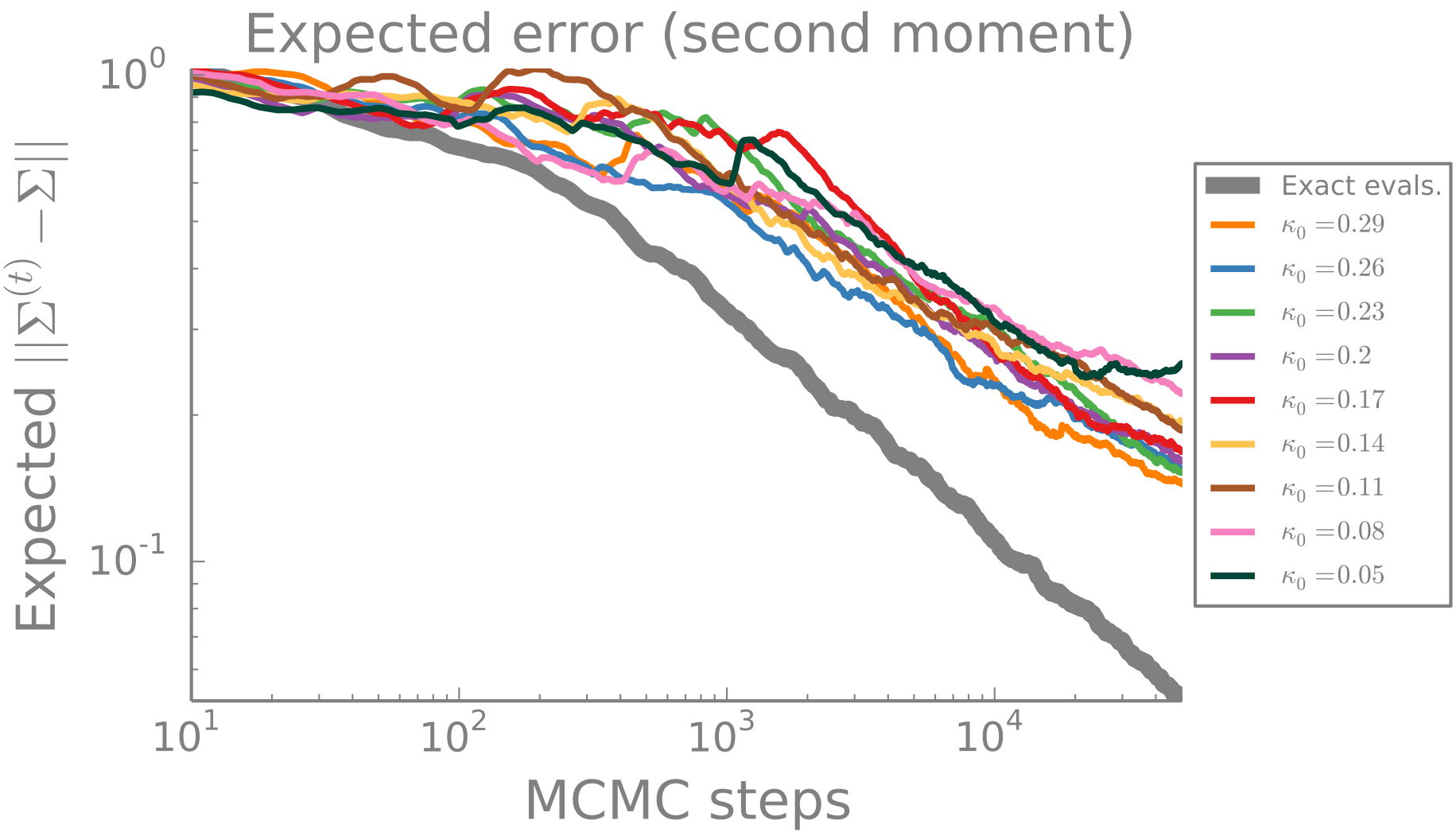
- Increment the number of nearest neighbors $k_n \sim k_n^{(0)} + \lfloor \kappa_1 t^{\kappa_0} \rfloor$

And refine at θ_t

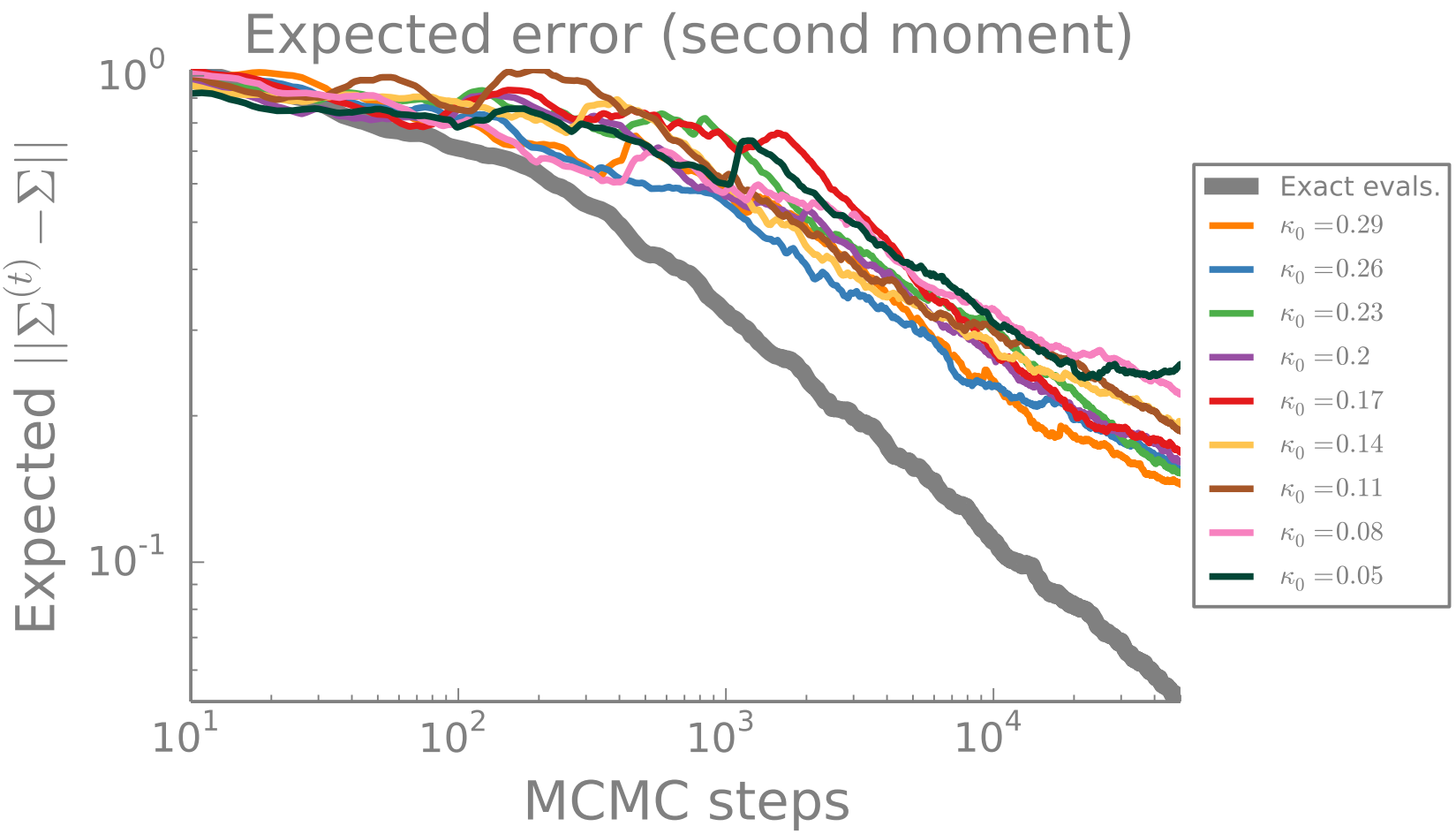
- With decaying error threshold $k_n \Lambda \Delta^{p+1} \sim q \Lambda_T \gamma_1 t^{-\gamma_0}$
- If the poisedness constant is large $\Lambda > \frac{q}{k_n} \Lambda_T$



Expected error (structural refinement & noisy evaluations)

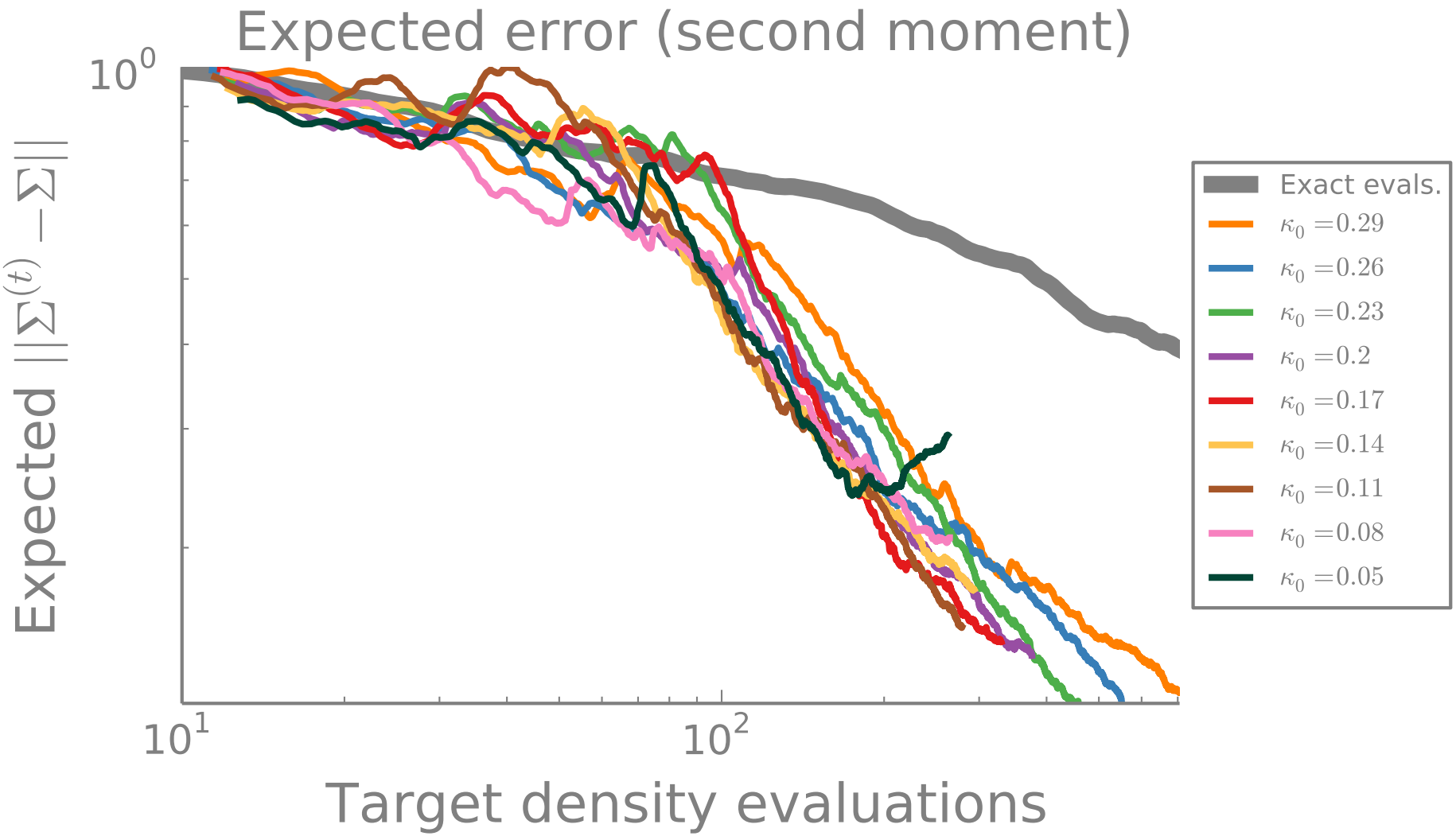


Expected error (structural refinement & noisy evaluations)

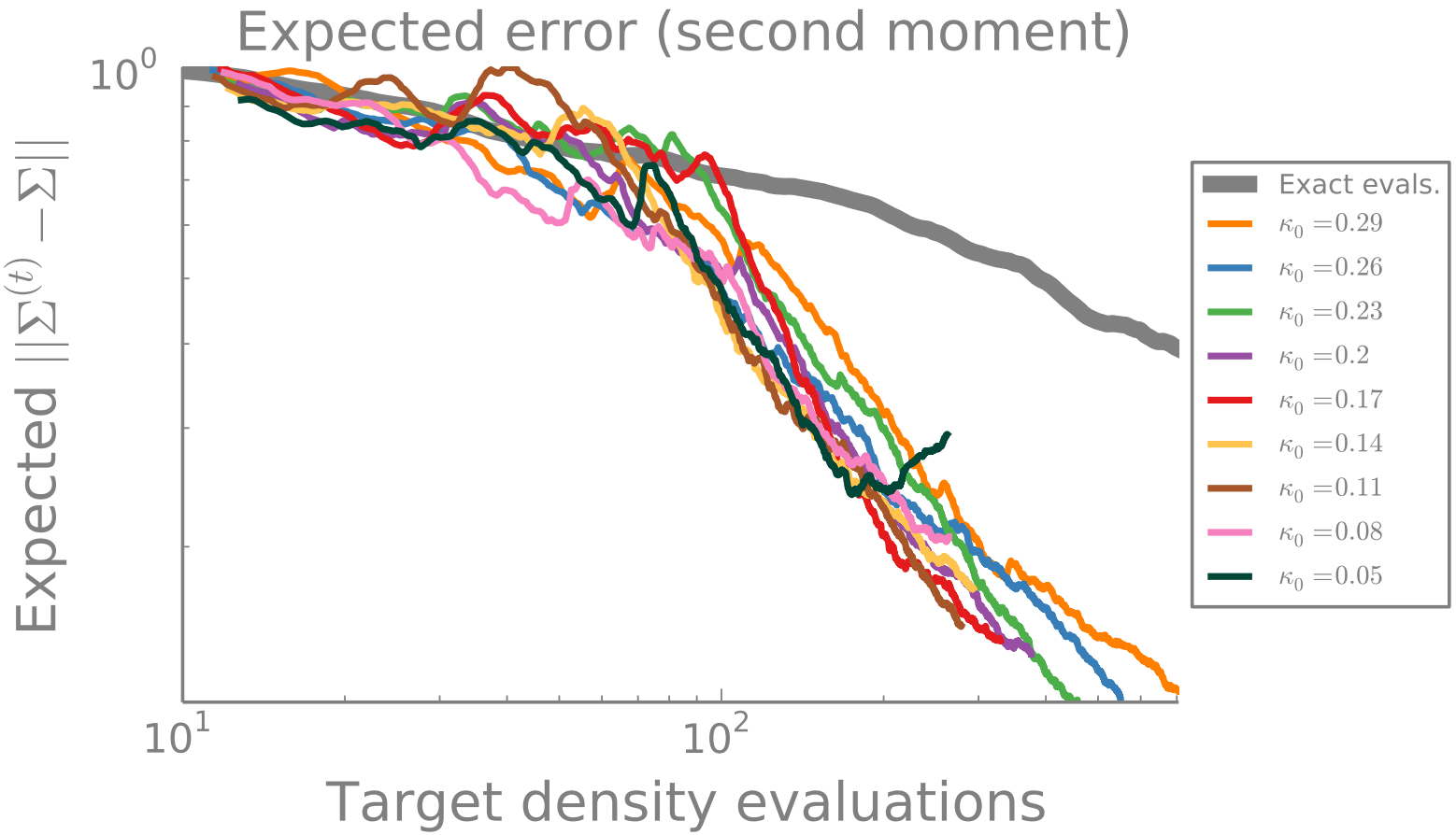


- MCMC with local approximations asymptotically characterizes the true target distribution even with noisy density evaluations

Expected error (structural refinement & noisy evaluations)



Expected error (structural refinement & noisy evaluations)



- Replacing the target density with a surrogate model built from noisy density evaluations is computationally advantageous

Conclusions

- Building and refining a local polynomial approximations significantly reduces computational expense

Conclusions

- Building and refining a local polynomial approximations significantly reduces computational expense

- Balancing error incurred by using the surrogate model with Monte Carlo error defines ideal rates for surrogate refinement

Conclusions

- Building and refining a local polynomial approximations significantly reduces computational expense
- Balancing error incurred by using the surrogate model with Monte Carlo error defines ideal rates for surrogate refinement
- Incrementing the number of nearest neighbors enables characterization of a distribution given only *noisy* evaluations of the target density
  - For instance, when the target is a particular marginal of a posterior distribution