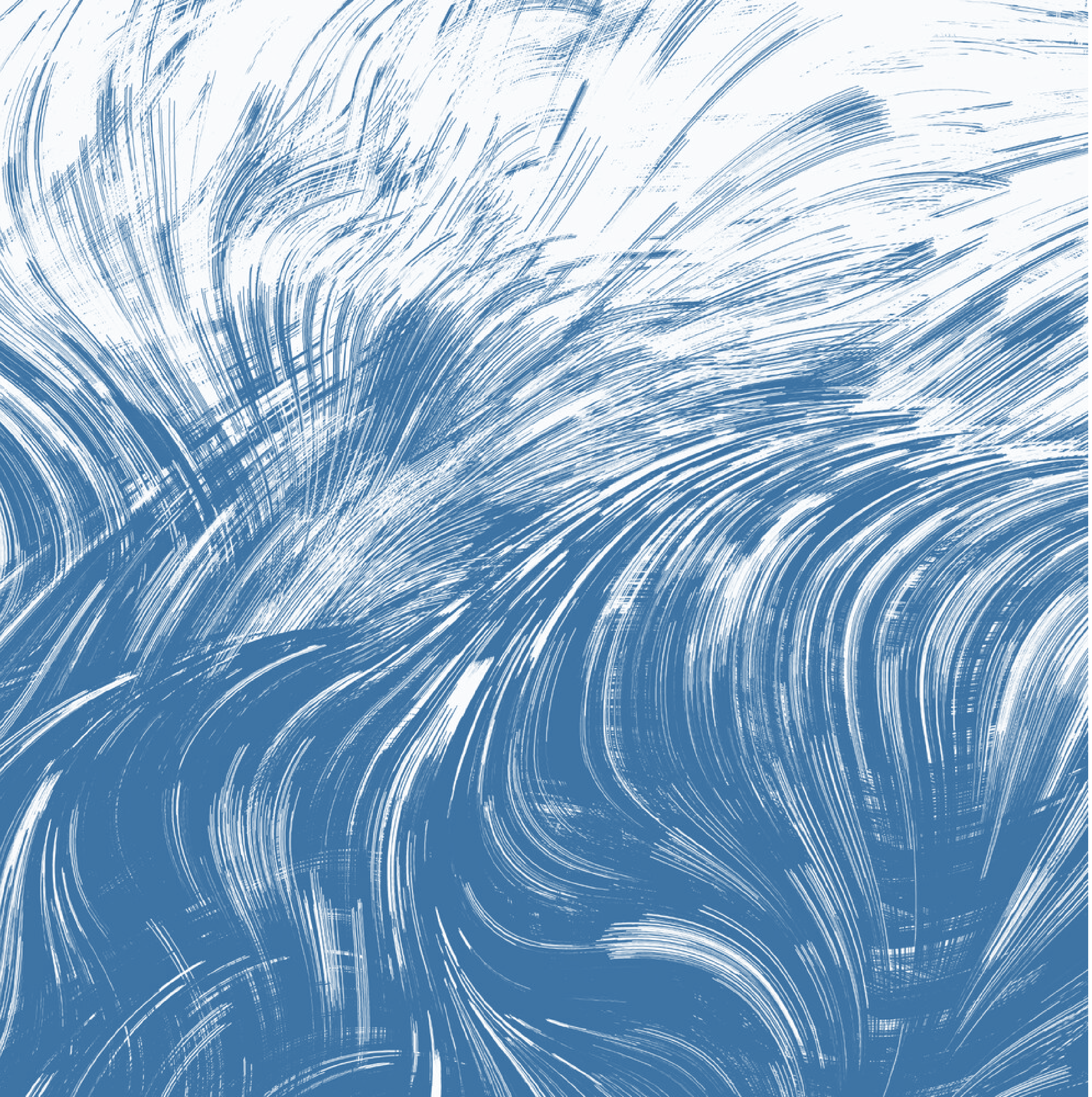


An energy-based model approach to the estimation of rare event probabilities

Lea Friedli

Joint work with:
Arnaud Doucet, University of Oxford
David Ginsbourger, University of Bern
Niklas Linde, University of Lausanne

Schweizerischer Nationalfonds

UNIL | Université de Lausanne

No risk, no fun
Rare event probabilities

No risk, no fun
Rare event probabilities

- Bayesian inversion inferring property field $\boldsymbol{\theta}$ given measurements $\boldsymbol{y}$
- Interested in quantity depending on field through $\boldsymbol{\theta} \mapsto \mathcal{R}(\boldsymbol{\theta})$
  → $\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq T \mid \boldsymbol{y})$, risk of failure of a system

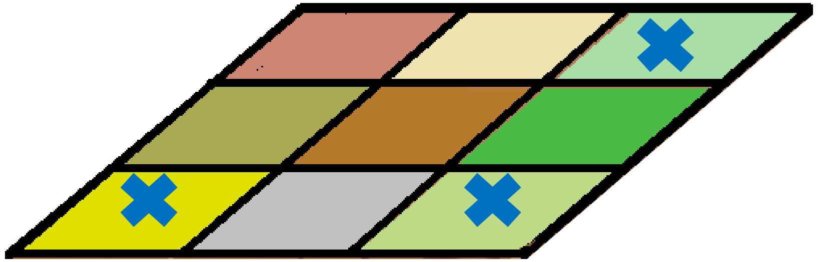
No risk, no fun
Rare event probabilities

- Bayesian inversion inferring property field $\boldsymbol{\theta}$ given measurements $\boldsymbol{y}$
- Interested in quantity depending on field through $\boldsymbol{\theta} \mapsto \mathcal{R}(\boldsymbol{\theta})$
  → $\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq T \mid \boldsymbol{y})$, risk of failure of a system
- High-dimensional target space, low-dimensional risk space
- Applications in finance, engineering, environmental science,…

No risk, no fun
Rare event probabilities

- Bayesian inversion inferring property field $\boldsymbol{\theta}$ given measurements $\boldsymbol{y}$
- Interested in quantity depending on field through $\boldsymbol{\theta} \mapsto \mathcal{R}(\boldsymbol{\theta})$
  → $\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq T \mid \boldsymbol{y})$, risk of failure of a system
- High-dimensional target space, low-dimensional risk space
- Applications in finance, engineering, environmental science,…
- Traditional Monte Carlo approach: excessive number of samples needed
- Variational approach (Valsson & Parrinello 2014) → Energy based models

Analytical toy example

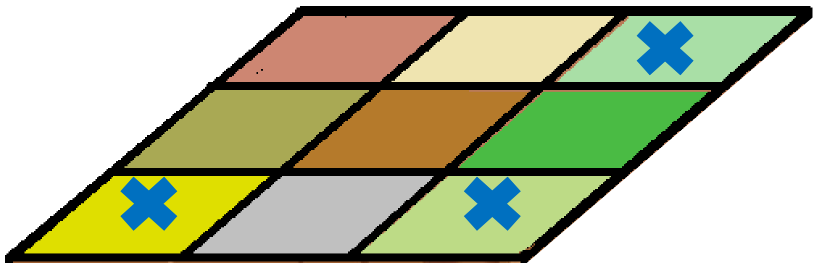


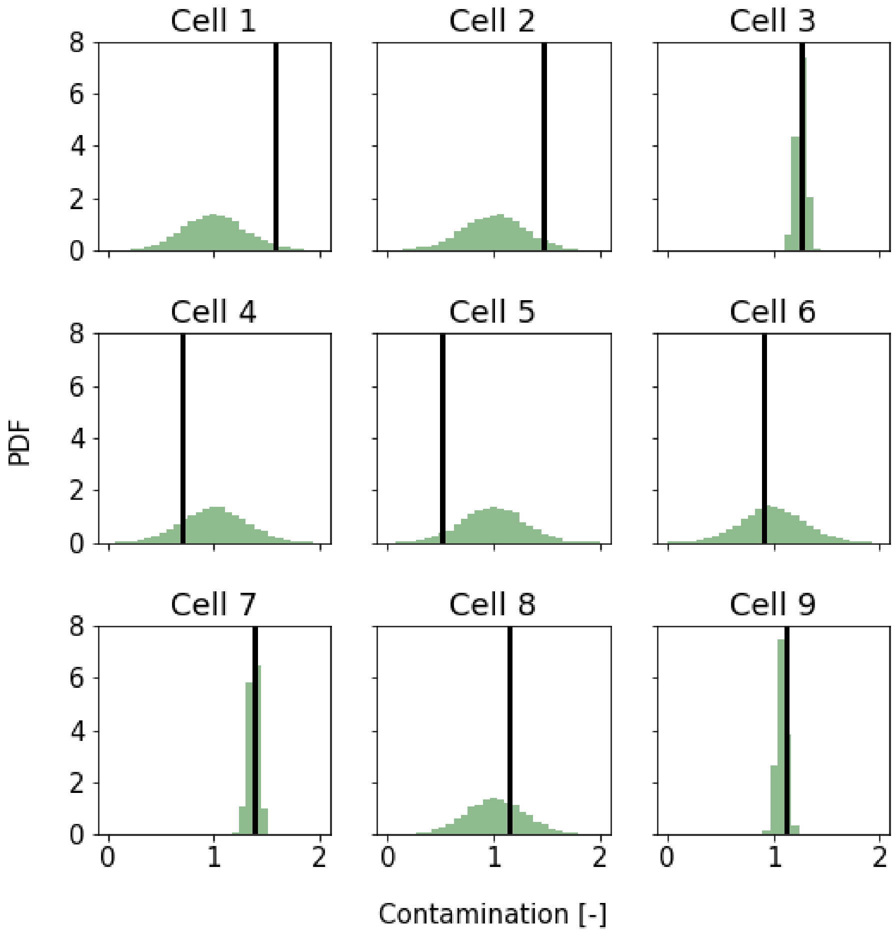
$\boldsymbol{\theta}$: Property field (contamination values)

$\boldsymbol{y} = \mathcal{G}(\boldsymbol{\theta}) + \epsilon$: (Local) measurements

→ Posterior PDF $p(\boldsymbol{\theta}|\boldsymbol{y})$

Analytical toy example



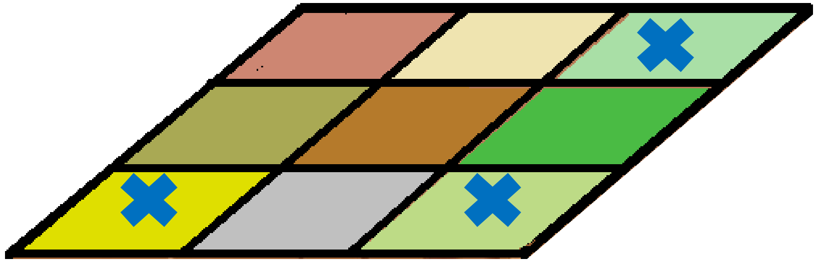


$\boldsymbol{\theta}$: Property field (contamination values)

$\boldsymbol{y} = \mathcal{G}(\boldsymbol{\theta}) + \epsilon$: (Local) measurements

→ Posterior PDF $p(\boldsymbol{\theta}|\boldsymbol{y})$

Analytical toy example



$\mathcal{R}(\boldsymbol{\theta})$: Quantity of property field

→ $\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq T \mid \boldsymbol{y}) = ?$

$\boldsymbol{\theta}$: Property field (contamination values)

$\boldsymbol{y} = \mathcal{G}(\boldsymbol{\theta}) + \epsilon$: (Local) measurements

→ Posterior PDF $p(\boldsymbol{\theta}|\boldsymbol{y})$

Analytical toy example

$\mathcal{R}(\boldsymbol{\theta})$: Quantity of property field

→ $\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq T \mid \boldsymbol{y}) = ?$

$\boldsymbol{\theta}$: Property field (contamination values)

$\boldsymbol{y} = \mathcal{G}(\boldsymbol{\theta}) + \epsilon$: (Local) measurements

→ Posterior PDF $p(\boldsymbol{\theta}|\boldsymbol{y})$

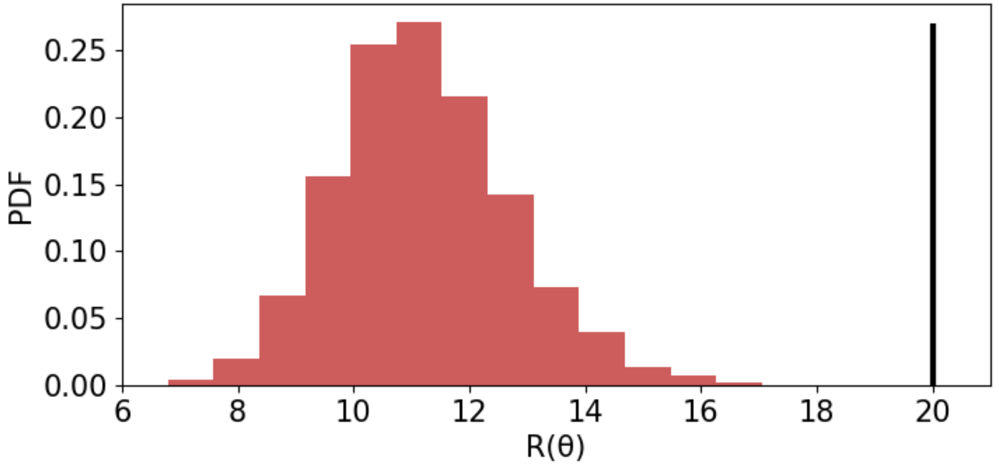
Linear quadratic dose response

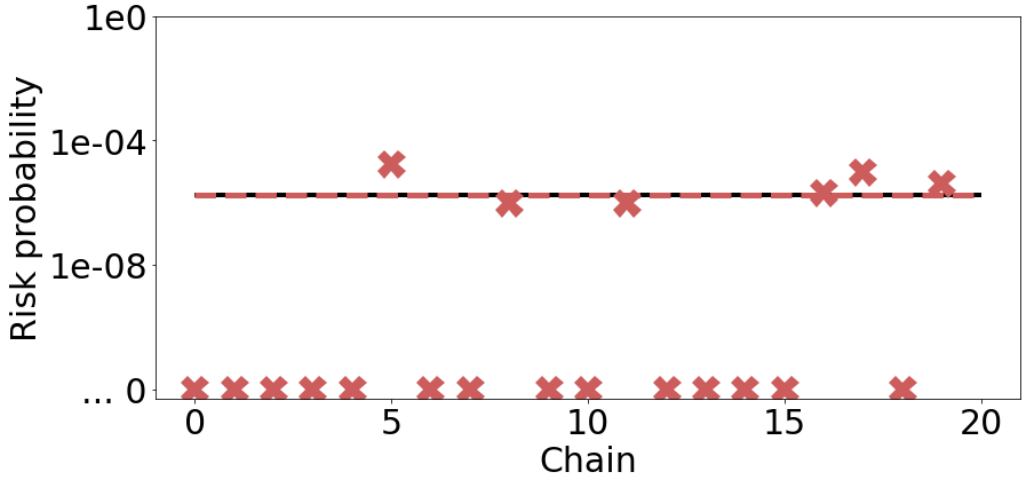
$$\mathcal{R}(\boldsymbol{\theta}) = \sum_{m=1}^{9} \theta_m^2$$

→ $\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq 20.0 \mid \boldsymbol{y}) = 1.76 \times 10^{-6}$

Metropolis Hastings

- 20 chains, 1 Million iterations each
- Gaussian proposals, AR 30%
- Convergence after 2'000 iterations (Gelman-Rubin)





Energy based model approach

- $\mathbb{P}(\mathcal{R}(\theta) \geq T \mid y) = \int_T^\infty p(r|y)\, dr$
- $R = \mathcal{R}(\theta),\ p(r|\, y) = \frac{\exp(-F(r))}{\int \exp(-F(s)) ds}$ → Find free energy $F(r)$

Energy based model approach

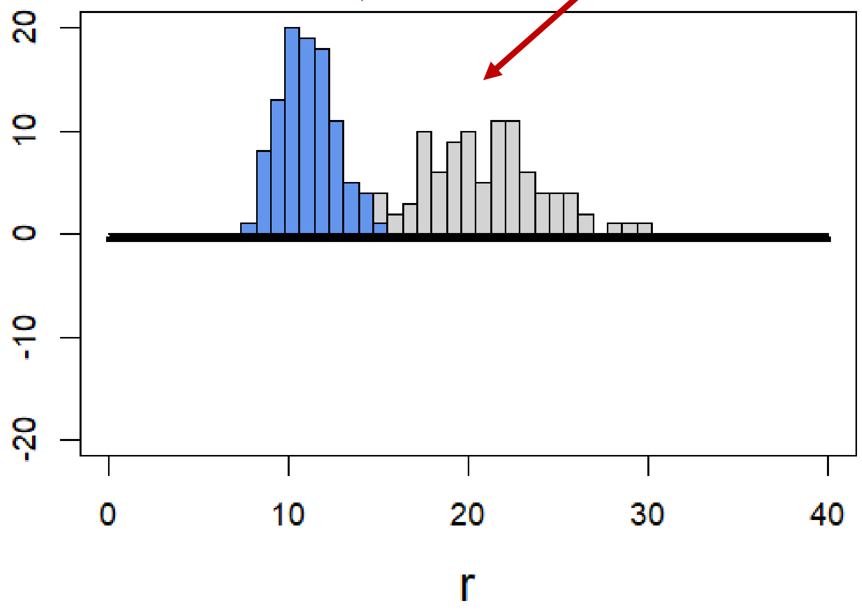
- $\mathbb{P}(\mathcal{R}(\theta) \geq T \mid y) = \int_T^\infty p(r|y)\, dr$
- $R = \mathcal{R}(\theta),\ p(r|\, y) = \frac{\exp(-F(r))}{\int \exp(-F(s))ds}$ → Find free energy $F(r)$
- Introduce bias potential V: $\mathbb{R} \to \mathbb{R}, r \mapsto V(r)$

$$p_V(r) = \frac{\exp\left(-(F(r)+V(r))\right)}{\int \exp\left(-(F(s)+V(s))\right)ds}$$

- Select PDF $p(r)$ with mass on $[T, \infty)$

$\boldsymbol{p(r|y) = p_0(r)}$

$\boldsymbol{p(r)}$



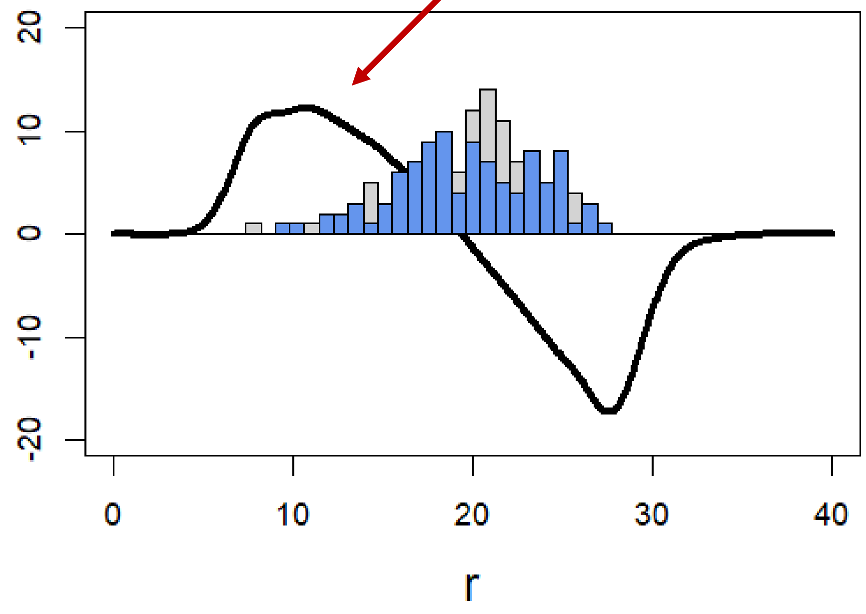
Energy based model approach

- $\mathbb{P}(\mathcal{R}(\theta) \geq T \mid y) = \int_T^\infty p(r|y)\, dr$
- $R = \mathcal{R}(\theta),\ p(r|\, y) = \frac{\exp(-F(r))}{\int \exp(-F(s)) ds}$ → Find free energy $F(r)$
- Introduce bias potential V: $\mathbb{R} \to \mathbb{R}, r \mapsto V(r)$

$$p_V(r) = \frac{\exp\left(-\left(F(r)+V(r)\right)\right)}{\int \exp\left(-\left(F(s)+V(s)\right)\right) ds}$$

- Select PDF $p(r)$ with mass on $[T, \infty)$
- Find $V_{opt}(r)$ minimizing V $\mapsto KL(p||p_V)$

$V_{opt}(r)$



Energy based model approach

- $\mathbb{P}(\mathcal{R}(\theta) \geq T \mid y) = \int_T^\infty p(r|y)\, dr$
- $R = \mathcal{R}(\theta),\ p(r|\, y) = \frac{\exp(-F(r))}{\int \exp(-F(s)) ds}$ → Find free energy $F(r)$
- Introduce bias potential V: $\mathbb{R} \to \mathbb{R}, r \mapsto V(r)$

$$p_V(r) = \frac{\exp\left(-\left(F(r)+V(r)\right)\right)}{\int \exp\left(-\left(F(s)+V(s)\right)\right) ds}$$

- Select PDF $p(r)$ with mass on $[T, \infty)$
- Find $V_{opt}(r)$ minimizing V $\mapsto KL(p||p_V)$
- $F(r) = -\log\left(p(r)\right) - V_{opt}(r)$

Variational approach
Valsson and Parrinello (2014)

- Loss function
- Rare event probabilities

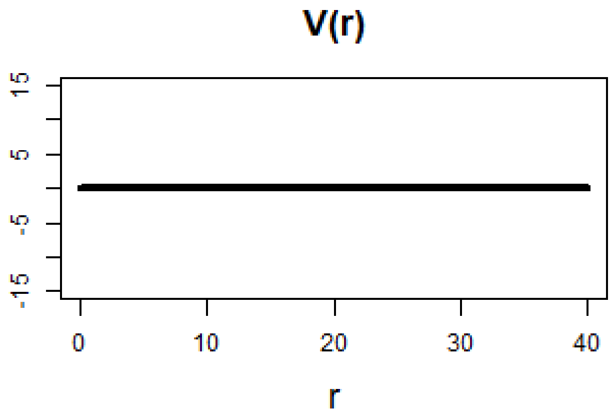
Implementation

Initialize V_ψ

Sample $r_i \sim p_{V_\psi}(r)$

Update V_ψ

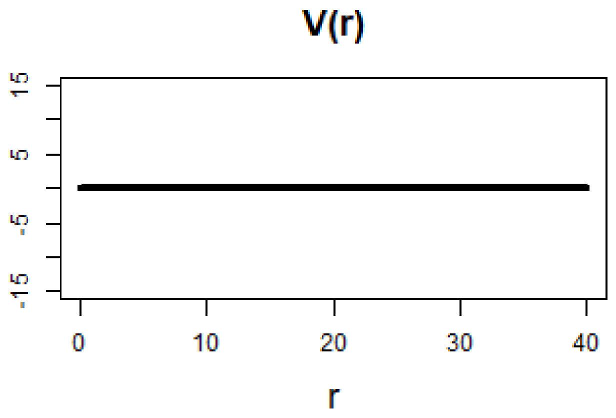
Parameterize bias potential $r \mapsto V_\psi(r)$ with neural network, radial basis functions, splines,…

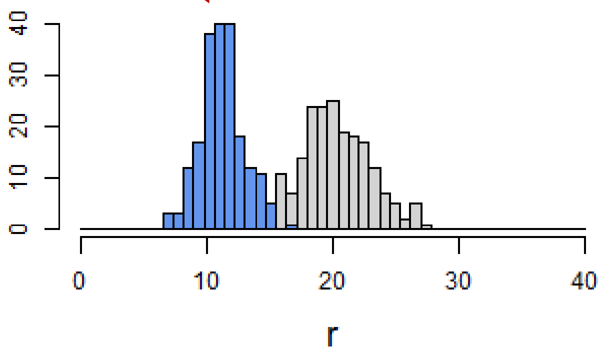


Implementation

Initialize V_ψ → Sample $r_i \sim p_{V_\psi}(r)$ → Update V_ψ

Sample $r_i \sim p_{V_\psi}(r)$ using MCMC

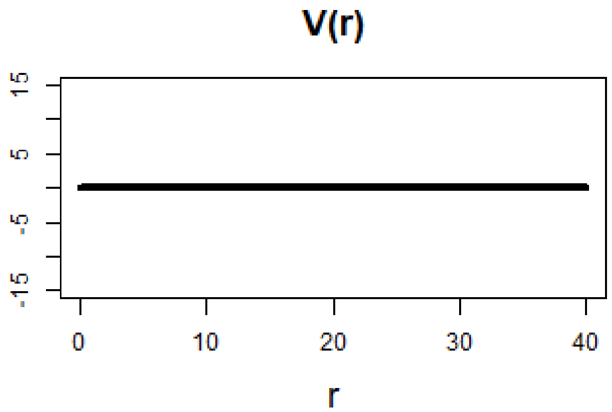


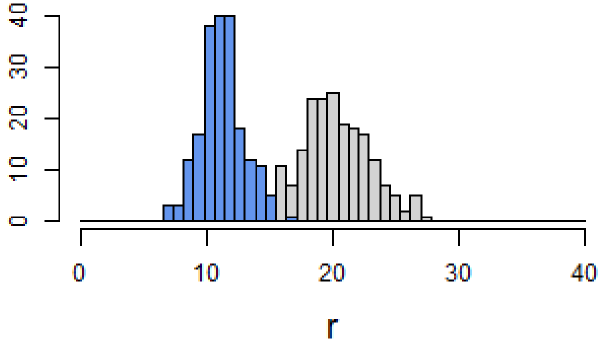


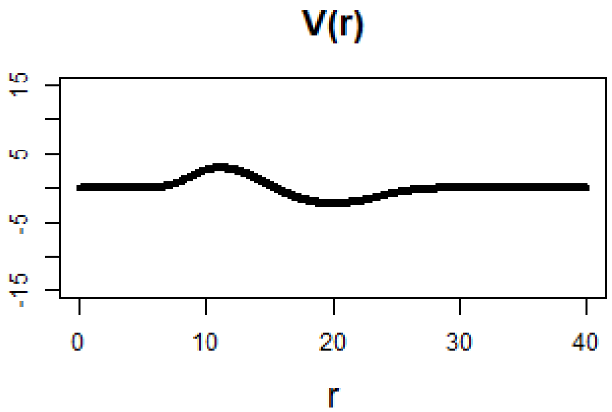
Implementation

Initialize V_ψ → Sample $r_i \sim p_{V_\psi}(r)$ → Update V_ψ

Update ψ using stochastic gradient descent to minimize $KL(p||p_V)$

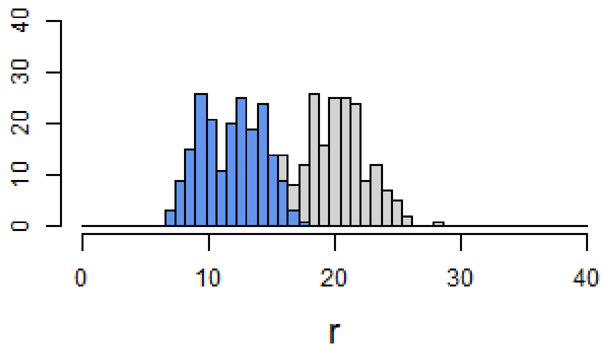


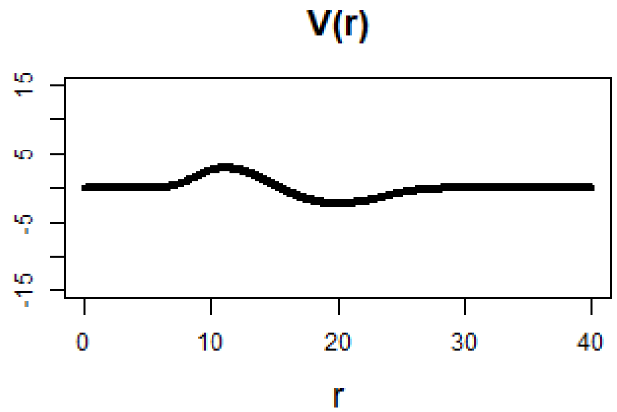




Implementation

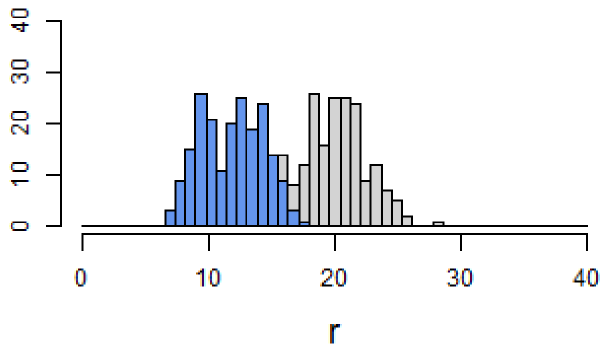
Initialize V_ψ → Sample $r_i \sim p_{V_\psi}(r)$ → Update V_ψ

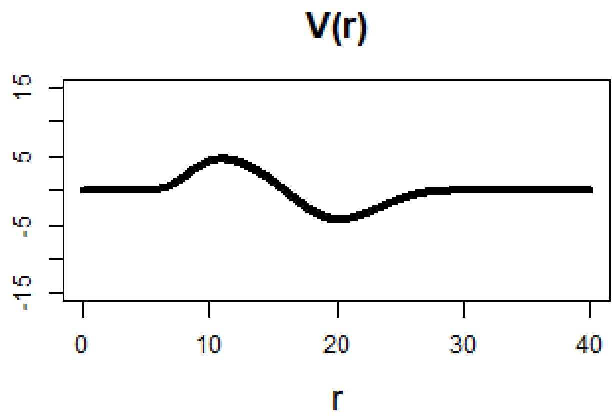




Implementation

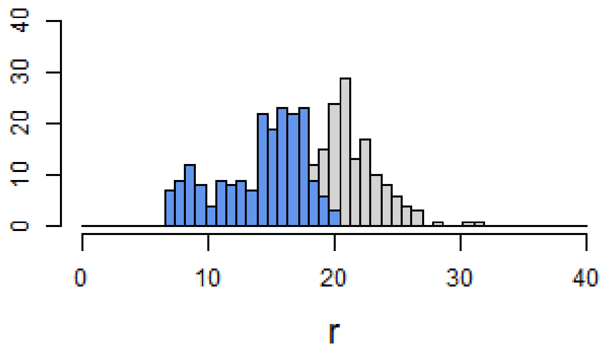
Initialize V_ψ → Sample $r_i \sim p_{V_\psi}(r)$ → Update V_ψ

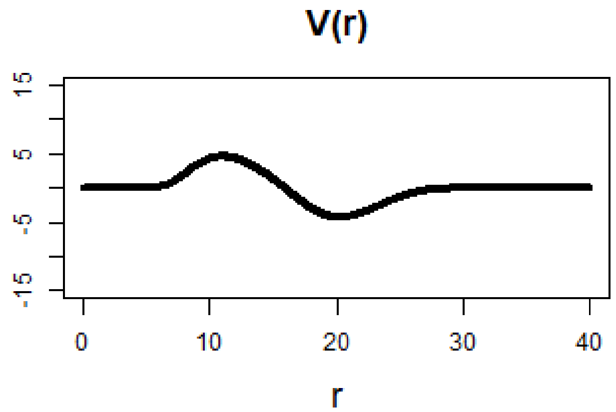




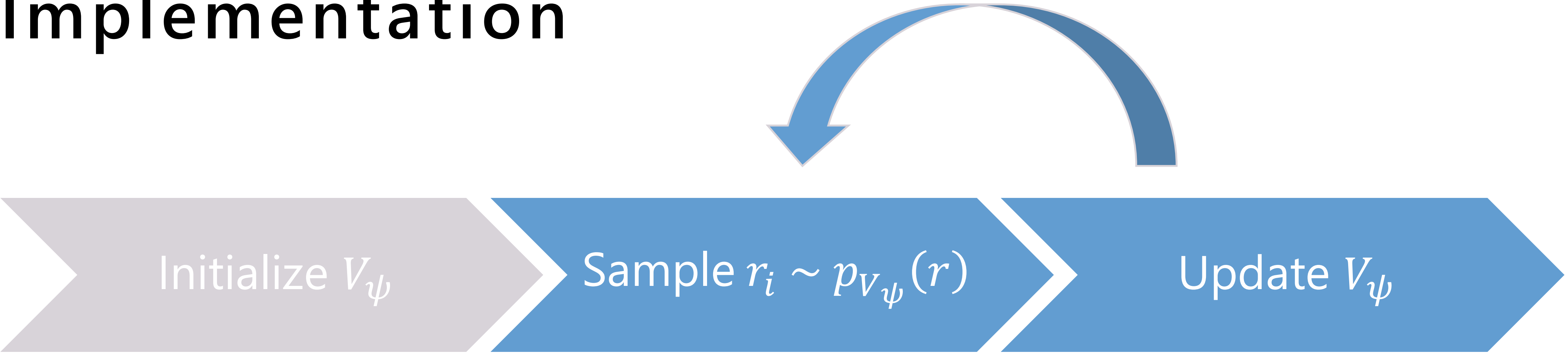
Implementation

Initialize V_ψ → Sample $r_i \sim p_{V_\psi}(r)$ → Update V_ψ



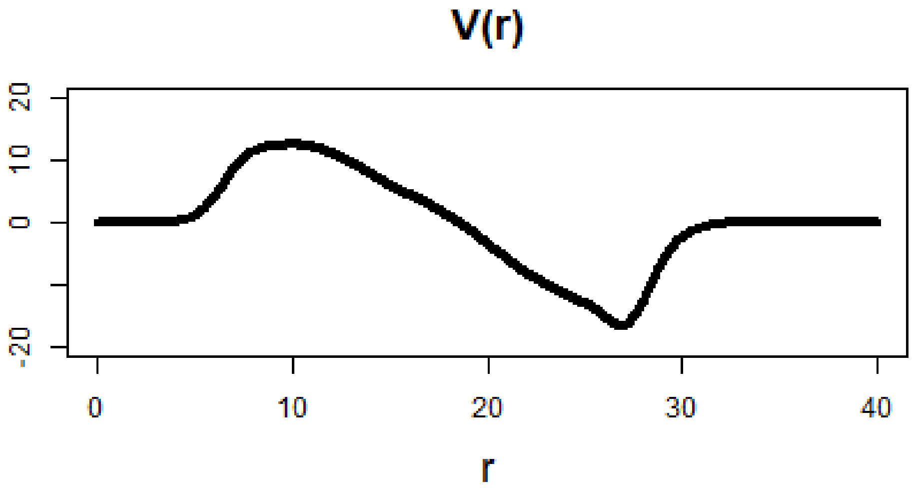


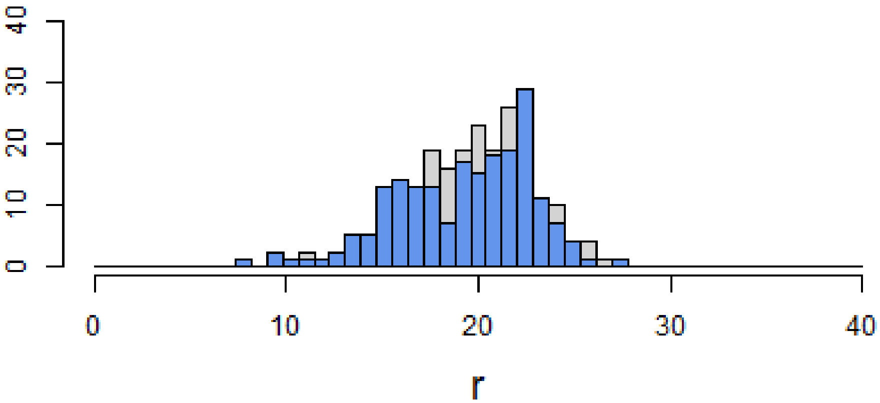
Implementation



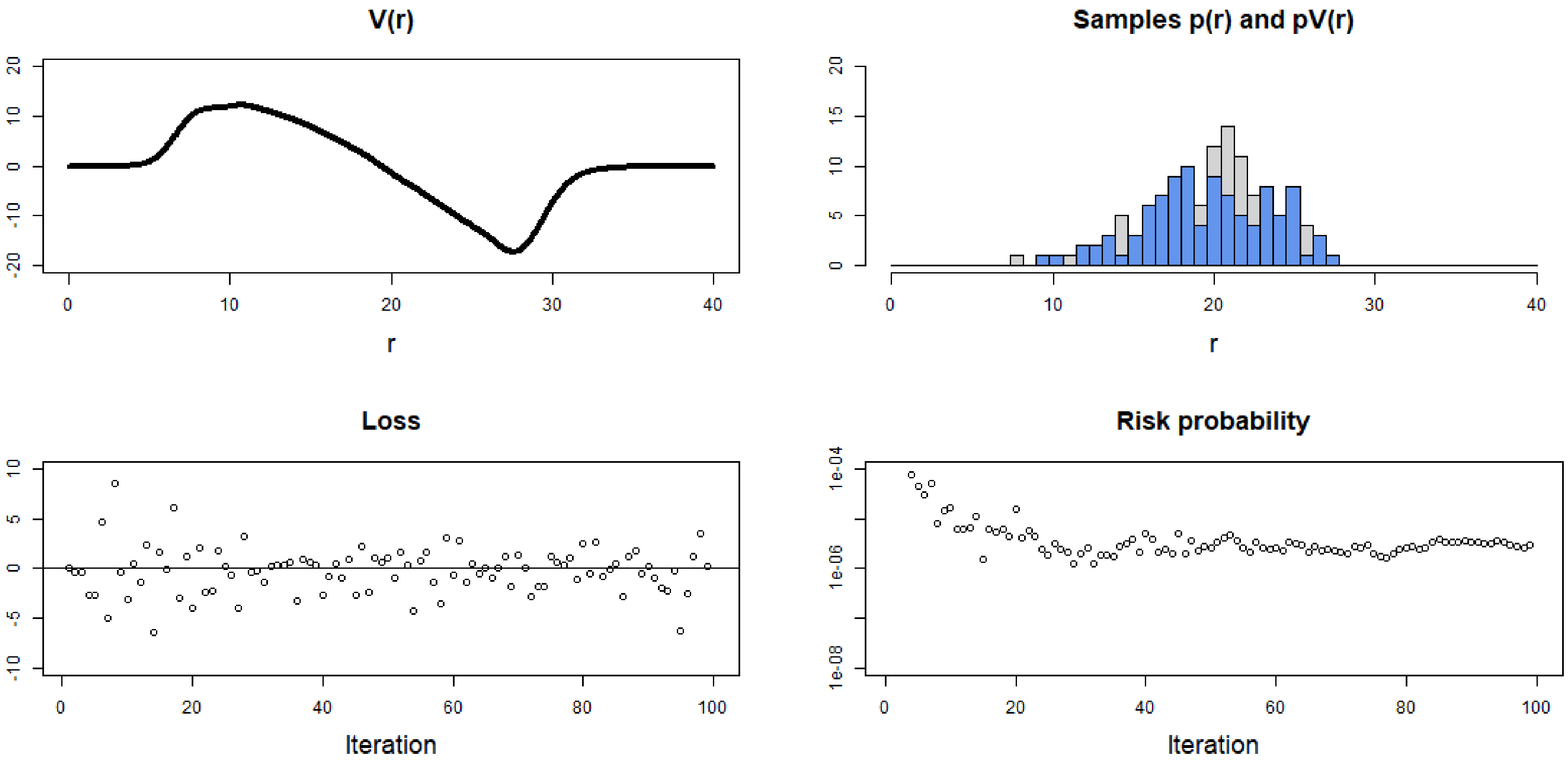
Implementation

Initialize V_ψ → Sample $r_i \sim p_{V_\psi}(r)$ → Update V_ψ





When do we stop?



Kernel Stein discrepancy

Kernel Stein discrepancy between distributions (Riabiz et al. 2022)

$$KSD\left(p|\, p_{V_\psi}\right) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} k_p\left(x_i, x_j\right)},\ \ x_i \sim p_{V_\psi}(\cdot),$$

$$k_p(x,y) = \ \nabla_x \nabla_y k(x,y) + \ \left\langle \nabla_x k(x,y), \nabla_y \log p(y) \right\rangle + \left\langle \nabla_y k(x,y), \nabla_x \log p(x) \right\rangle$$
$$+\, k(x,y) \left\langle \nabla_x \log p(x), \nabla_y \log p(y) \right\rangle$$

Kernel Stein discrepancy

Kernel Stein discrepancy between distributions (Riabiz et al. 2022)

$$KSD\left(p|\, p_{V_\psi}\right) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} k_p\left(x_i, x_j\right)}\,,\;\; x_i \sim p_{V_\psi}(\cdot),$$

$$\begin{aligned} k_p(x,y) = {} & \nabla_x \nabla_y k(x,y) + \left\langle \nabla_x k(x,y), \nabla_y \log p(y) \right\rangle + \left\langle \nabla_y k(x,y), \nabla_x \log p(x) \right\rangle \\ & + k(x,y) \left\langle \nabla_x \log p(x), \nabla_y \log p(y) \right\rangle \end{aligned}$$

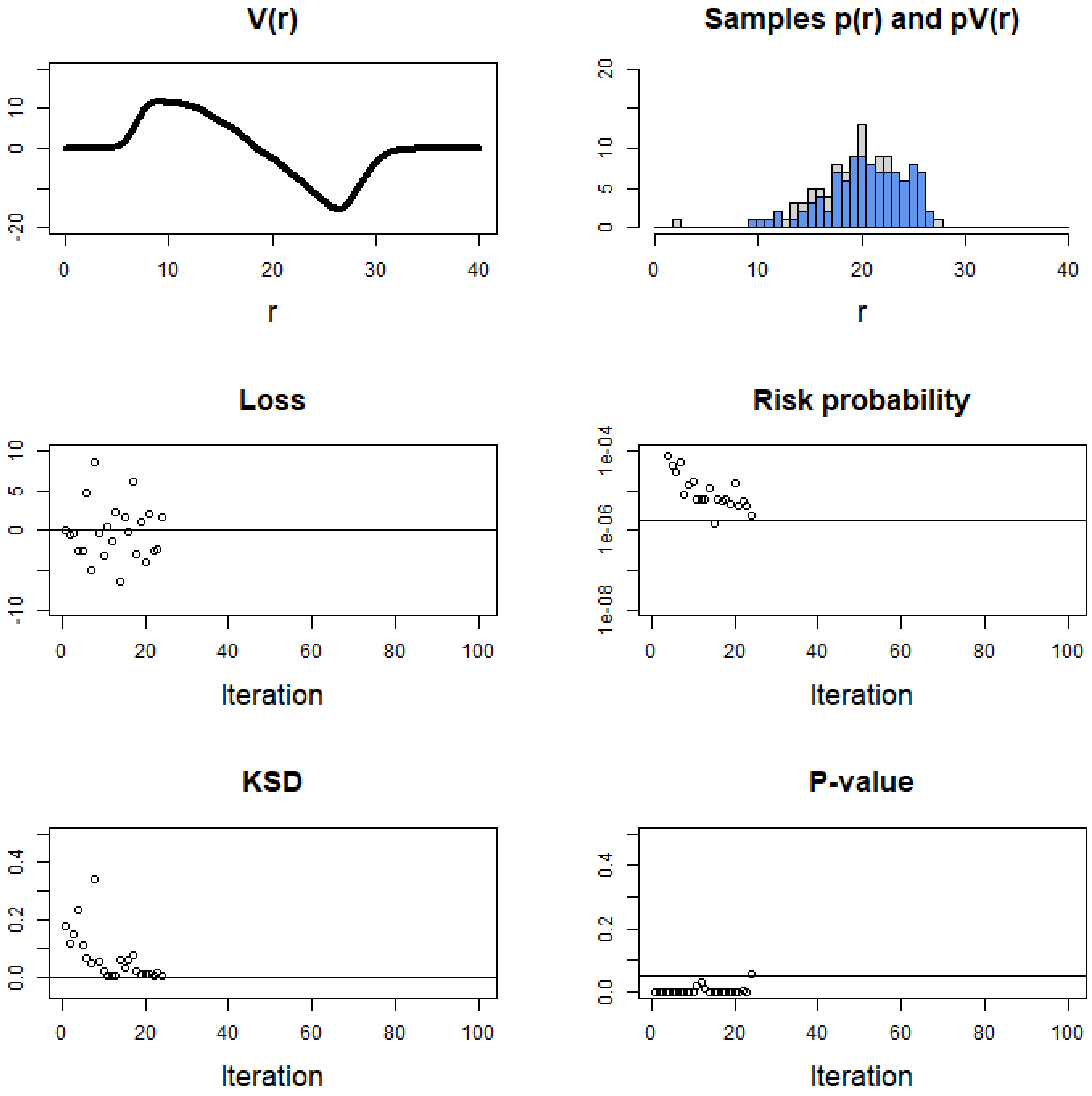
Use kernelized Stein discrepancy for goodness-of-fit tests with $H_0 : \; p = p_{V_\psi}$

→ Employing bootstrap procedure

→ Stop when H_0 cannot be rejected anymore (significance level α)

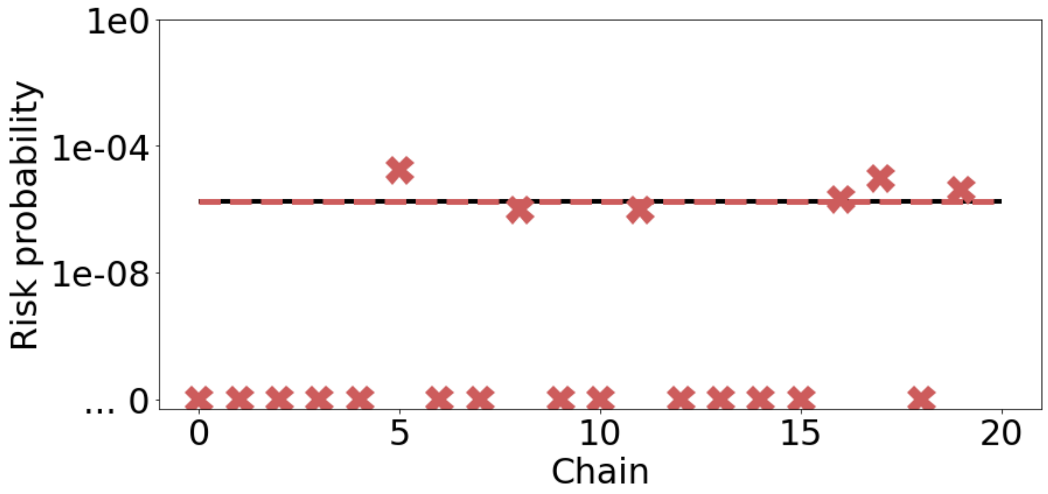
→ Conservative for correlated samples (Chwialkowski et al. 2016)

Stopping criteria



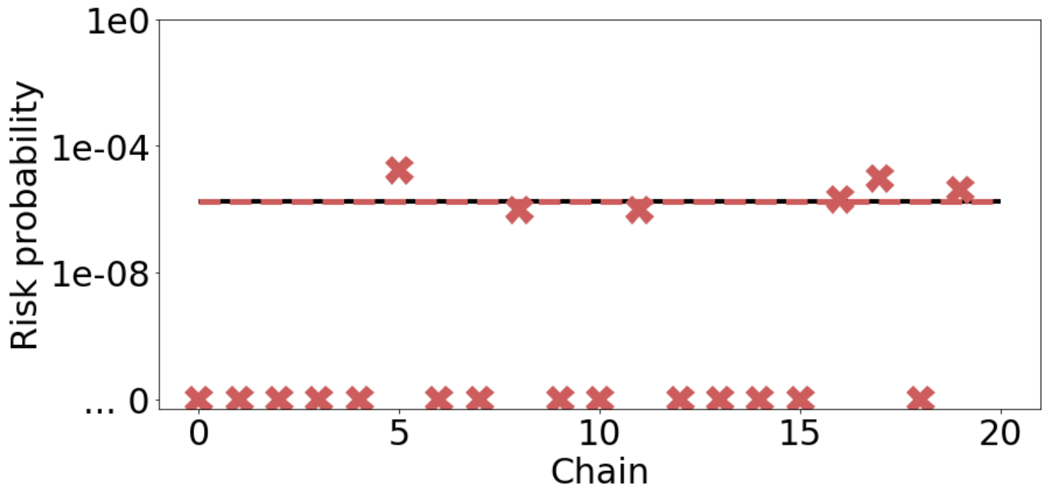
Energy based models vs. MCMC

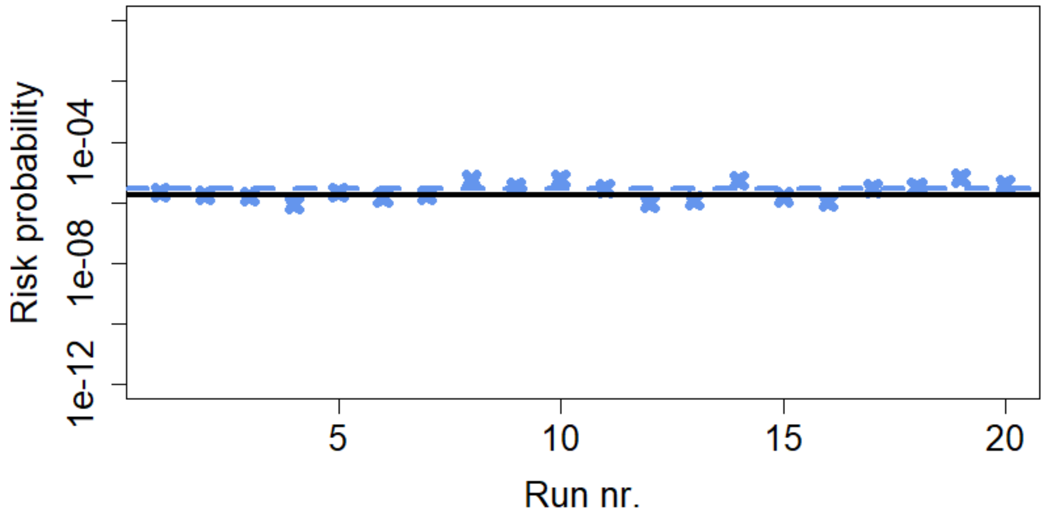
- $\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq 20.0 \mid \boldsymbol{y}) = 1.76 \times 10^{-6}$
- MCMC (left): 1’000’000 iterations → SD = 4.2×10^{-6}
-



Energy based models vs. MCMC

- $\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq 20.0 \mid \boldsymbol{y}) = 1.76 \times 10^{-6}$
- MCMC (left): 1’000’000 iterations → SD = 4.2×10^{-6}
- EBM (right): 14’000-50’000 iterations → SD = 1.5×10^{-6}





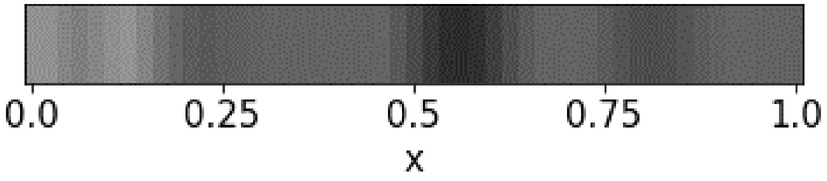
1D flow example (Straub et al. 2016)

- Hydraulic diffusivity field $\boldsymbol{a}(\boldsymbol{x}), x \in [0\mathrm{m}, 1\mathrm{m}]$

  → Log-diffusivity Gaussian, Karhunen-Loève expansion
- Diffusivity = speed at which pressure pulse propagates through aquifer

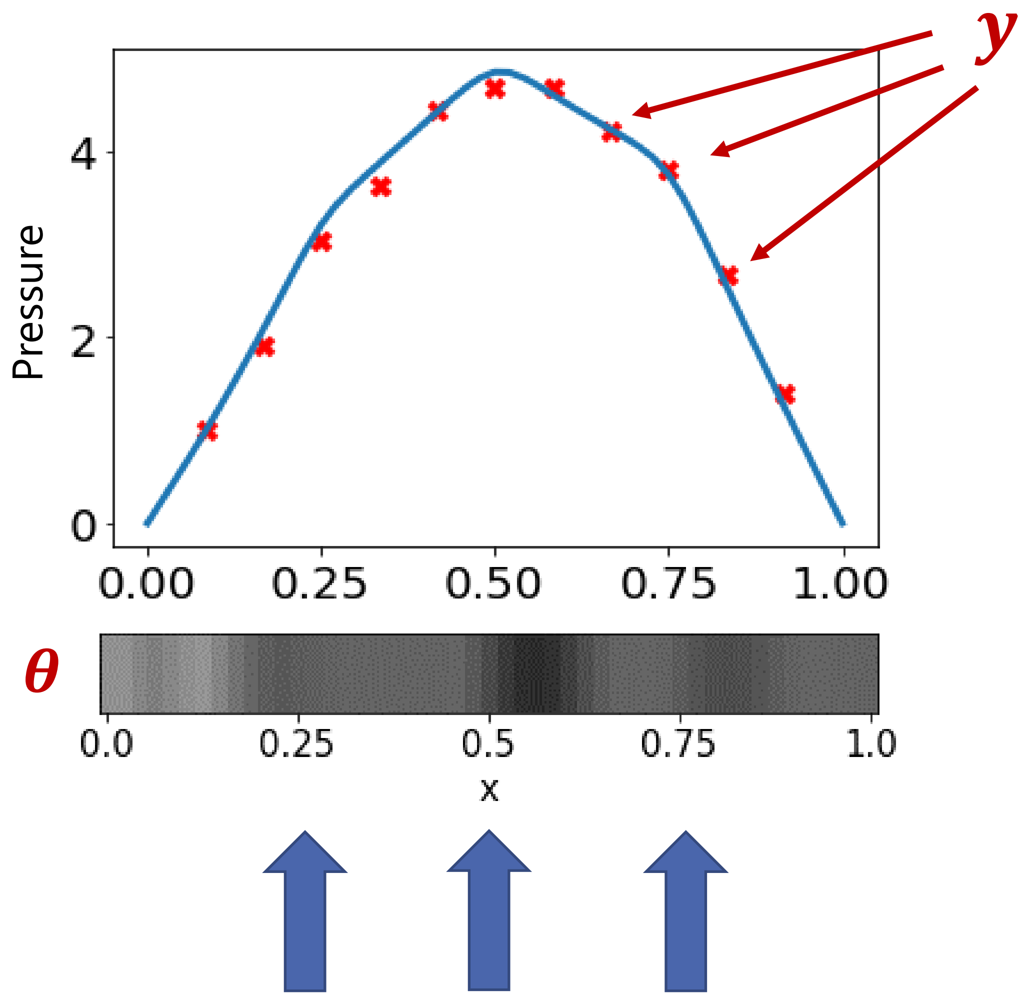
$$\ln \boldsymbol{a}(\boldsymbol{x}) = \boldsymbol{\mu}_{\ln\boldsymbol{\theta}} + \boldsymbol{\sigma}_{\ln\boldsymbol{\theta}} \sum_{\boldsymbol{i}=1}^{10} \sqrt{\boldsymbol{w}_{\boldsymbol{i}}}\,\boldsymbol{v}_{\boldsymbol{i}}(\boldsymbol{x})\boldsymbol{Z}_{\boldsymbol{i}}$$

$\boldsymbol{\theta}$



1D flow example

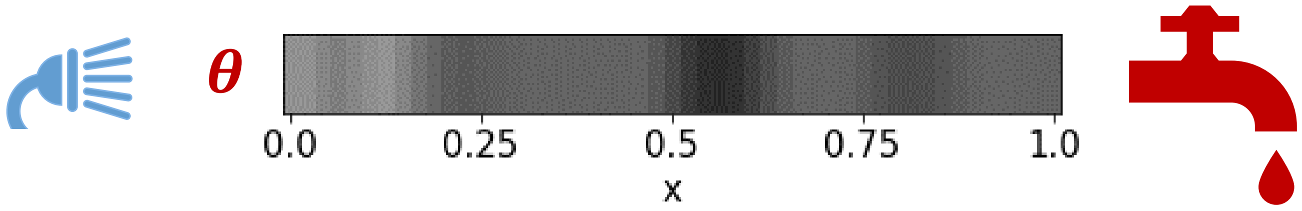
Data



1D flow example

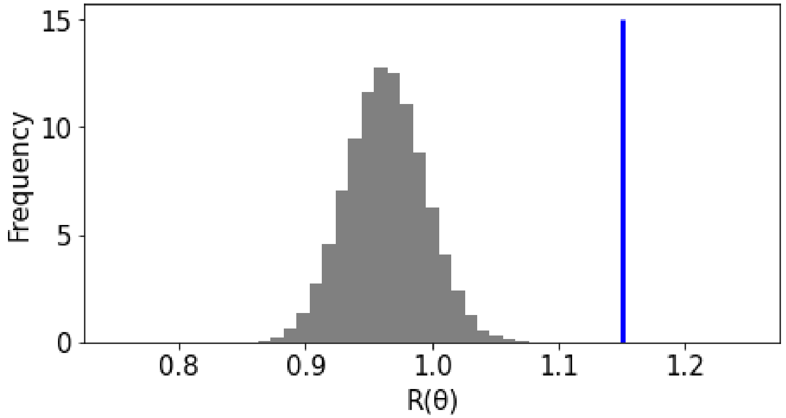
Risk

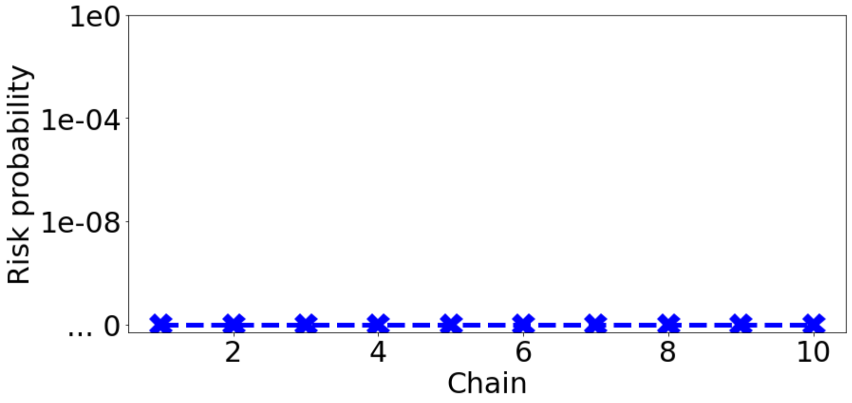
$R(\theta)$ = **outflow rate**



Energy based models vs. MCMC

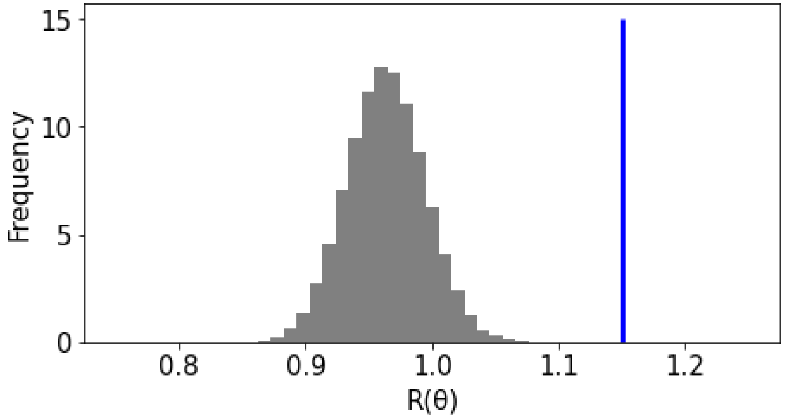
$$\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq 1.15 \mid \boldsymbol{y})$$

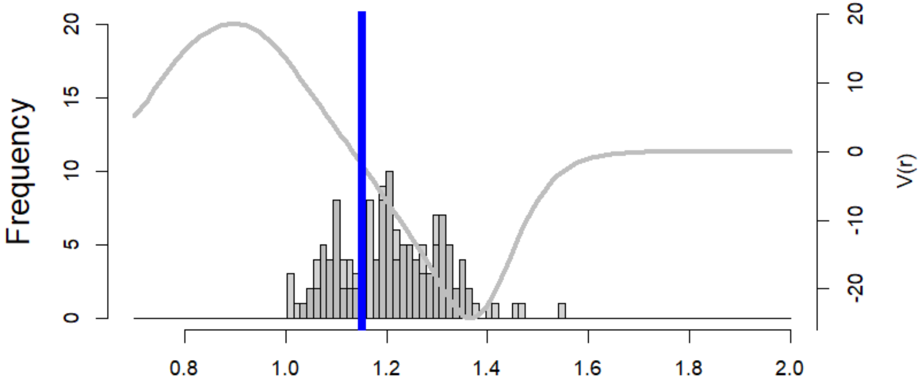


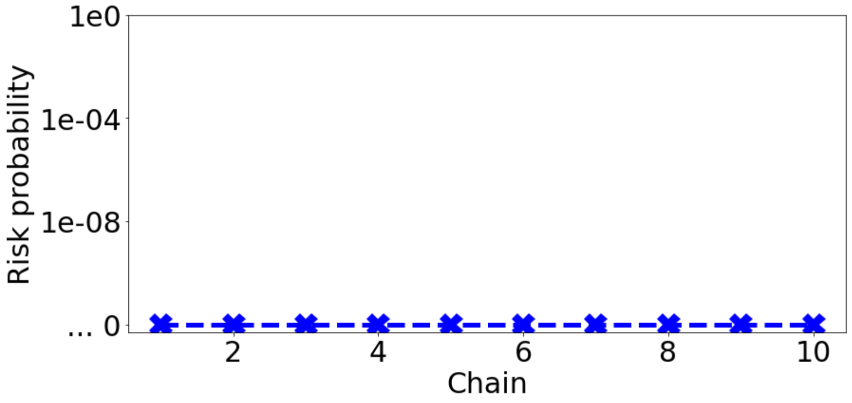


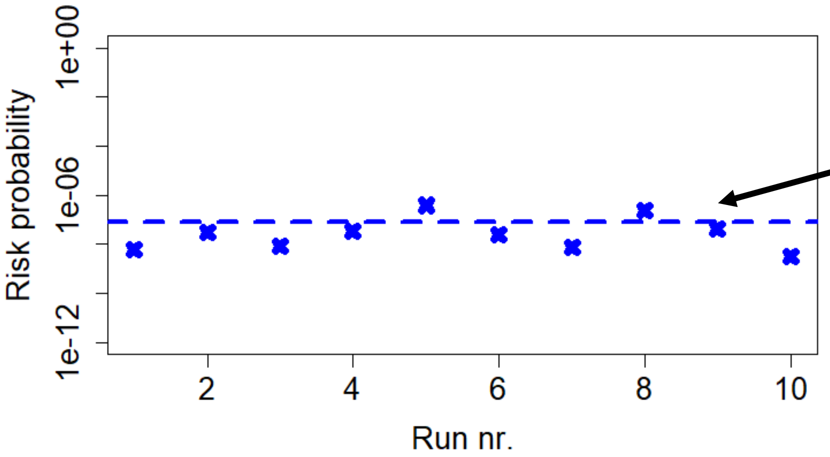
Energy based models vs. MCMC

$$\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq 1.15 \mid \boldsymbol{y})$$





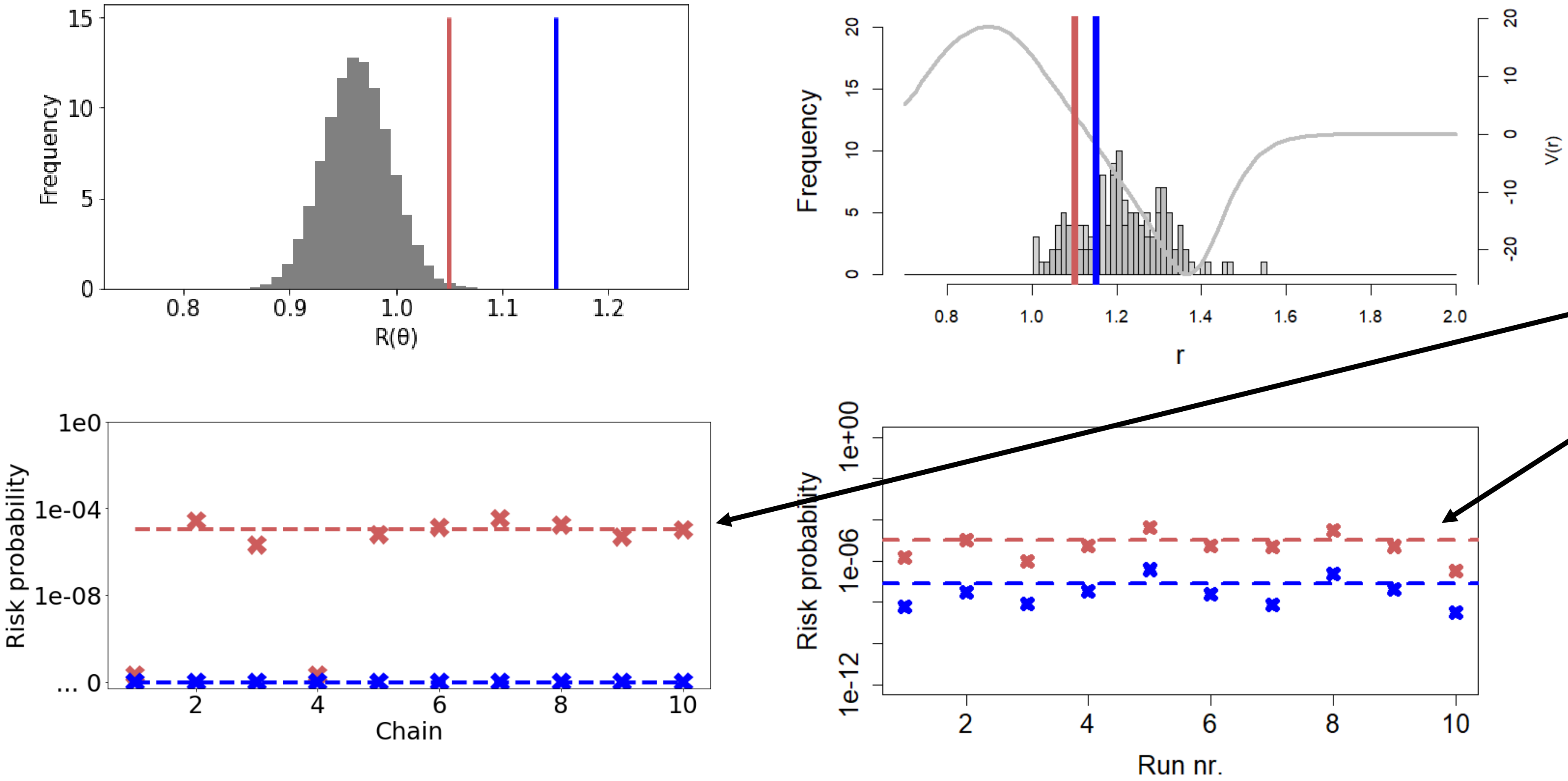




7.8×10^{-8}

Energy based models vs. MCMC

$+\,\mathbb{P}(\mathcal{R}(\boldsymbol{\theta}) \geq 1.10 \mid \boldsymbol{y})$



1.2×10^{-5}

1.1×10^{-5}

Conclusions and ongoing work

- **EBM approach reduces high-dimensional problem to optimization of one-dimensional function, compared to MCMC fraction of model evaluations needed**

Conclusions and ongoing work

- **EBM approach reduces high-dimensional problem to optimization of one-dimensional function, compared to MCMC fraction of model evaluations needed**

- **Configuration crucial**

- Parameterization $V(r)$	- Sampling $p_V(r)$
- Choice of $p(r)$	- Numerical integration
- Learning rate	- Stopping criteria

Conclusions and ongoing work

- **EBM approach reduces high-dimensional problem to optimization of one-dimensional function, compared to MCMC fraction of model evaluations needed**

- **Configuration crucial**

  - Parameterization $V(r)$
  - Choice of $p(r)$
  - Learning rate
  - Sampling $p_V(r)$
  - Numerical integration
  - Stopping criteria

- **Assessment and comparison**

  - MCMC to validate probabilities down to 10^{-6}
  - Reliability literature (Straub et al. 2016)
  - Sequential Monte Carlo

Thank you!

lea.friedli@unil.ch

Schweizerischer Nationalfonds

UNIL | Université de Lausanne

References

Kacper Chwialkowski, Heiko Strathmann and Arthur Gretton. A kernel test of goodness of fit. In International conference on machine learning (pp. 2606-2615). PMLR, 2016.

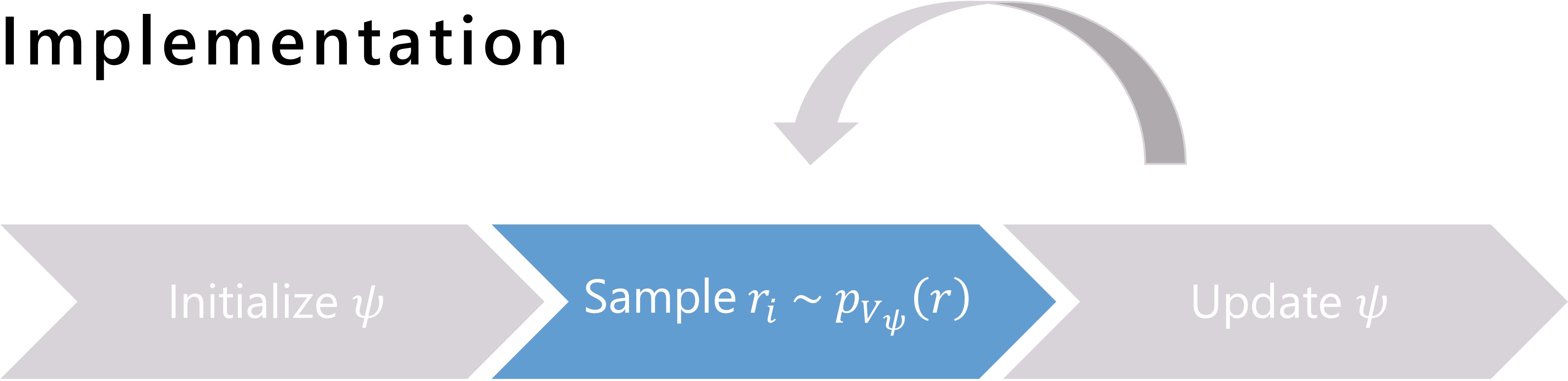
Michele Invernizzi, Omar Valsson, and Michele Parrinello. Coarse graining from variationally enhanced sampling applied to the ginzburg–landau model. Proceedings of the National Academy of Sciences, 114(13):3370–3374, 2017.

Marina Riabiz, Wilson Ye Chen, Jon Cockayne, Pawel Swietach, Steven A. Niederer, Lester Mackey, and Chris. J. Oates. Optimal Thinning of MCMC Output. arXiv:2005.03952v5, 2022.

Daniel Straub, Iason Papaioannou and Wolfgang Betz. Bayesian analysis of rare events. Journal of Computational Physics, 314, 538-556, 2016.

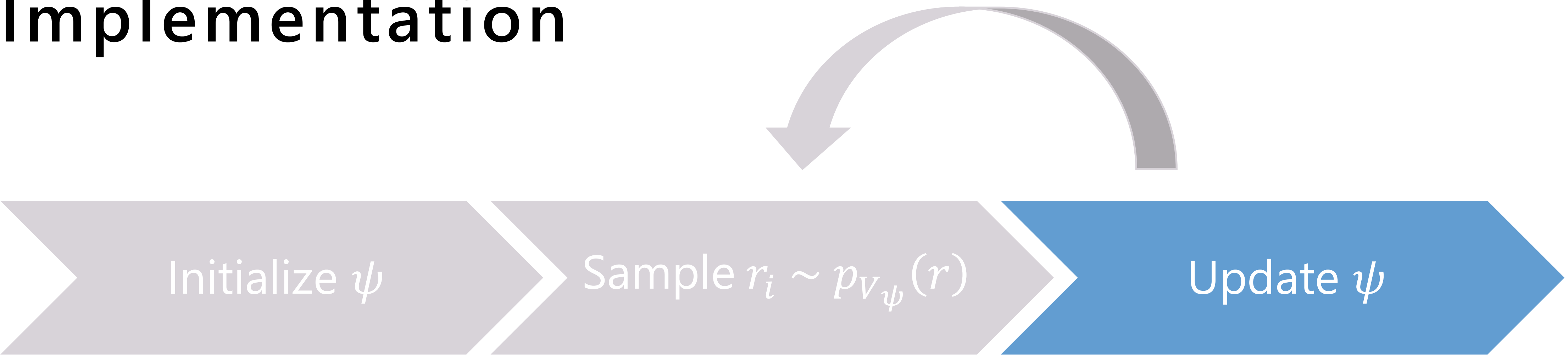
Omar Valsson and Michele Parrinello. Variational approach to enhanced sampling and free energy calculations. Physical Review Letters, 113(9):090601, 2014.

Implementation



- $p_{V_\psi}(r) = \frac{\exp\left(-\left(F(r)+V_\psi(r)\right)\right)}{\int \exp\left(-\left(F(s)+V_\psi(s)\right)\right)ds}$
- Posterior PDF: $p(\theta|y) = \frac{\exp(-U(\theta))}{\int \exp(-U(\xi))d\xi}$ with $U(\theta) = -\log p(y|\theta) - \log p(\theta)$
- MCMC to draw proportional to $p_{V_\psi}(\theta) = \frac{\exp\left(-\left(U(\theta)+V_\psi(\mathcal{R}(\boldsymbol{\theta}))\right)\right)}{\int \exp\left(-\left(U(\xi)+V_\psi(\mathcal{R}(\xi))\right)\right)d\xi}$
- Transform samples with $\theta \longmapsto \mathcal{R}(\theta)$

Implementation



- Stochastic gradient descent $\psi_{n+1} = \psi_n - \gamma \frac{\partial J(\psi)}{\partial \psi}$
- Valsson and Parrinello (2014) employ loss related to KL divergence
- $\frac{\partial J(\psi)}{\partial \psi} \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \psi} V_\psi(r_i) - \frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \psi} V_\psi(s_i),\ s_i \sim p(\cdot),\ r_i \sim p_{V_\psi}(\cdot)$

Configuration analytical toy example

- **Parameterization $\boldsymbol{V(r)}$**

  1000 Gaussian radial basis functions (RBF, eps=1, from 0 to 40), learn weights

- **Learning rate**

  0.5, geometric decrease with factor 1/1.025

- **MCMC:**

  Metropolis-Hastings, Gaussian proposals, step width for AR 30%, 1100 steps, burn-in after 100 steps, thinning with factor 10

- **Choice of $\boldsymbol{p(r)}$**

  $\mathcal{N}(20,4)$

- **Stopping criteria**

  $\alpha = 0.05$, 1000 bootstrap samples

Configuration 1D flow example

- **Parameterization $\boldsymbol{V}(\boldsymbol{r})$**

  1000 Gaussian radial basis functions (RBF, eps=20 from 0 to 2), learn weights
- **Learning rate**

  0.25, geometric decrease with factor 1/1.025
- **MCMC:**

  Metropolis-Hastings, Gaussian proposals, step width for AR 30%, 1200 steps, burn-in after 200 steps, thinning with factor 10
- **Choice of $\boldsymbol{p}(\boldsymbol{r})$**

  $\mathcal{N}(1.2, 0.125)$
- **Stopping criteria**

  $\alpha = 0.01$, 1000 bootstrap samples