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
No risk, no fun
Rare event probabilities
No risk, no fun
Rare event probabilities

• Bayesian inversion inferring property field $\theta$ given measurements $y$

• Interested in quantity depending on field through $\theta \mapsto R(\theta)$
  $\rightarrow P(R(\theta) \geq T \mid y)$, risk of failure of a system
Bayesian inversion inferring property field $\theta$ given measurements $y$

Interested in quantity depending on field through $\theta \mapsto \mathcal{R}(\theta)$

\[ \mathbb{P}(\mathcal{R}(\theta) \geq T | y), \text{ risk of failure of a system} \]

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 $\theta$ given measurements $y$
• Interested in quantity depending on field through $\theta \mapsto \mathcal{R}(\theta)$
  $\Rightarrow \mathbb{P}(\mathcal{R}(\theta) \geq T \mid y)$, 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) $\Rightarrow$ Energy based models
Analytical toy example

\( \theta \): Property field (contamination values)
\( y = g(\theta) + \epsilon \): (Local) measurements
\( \rightarrow \) Posterior PDF \( p(\theta | y) \)
Analytical toy example

\( \theta \): Property field (contamination values)
\( y = G(\theta) + \epsilon \): (Local) measurements
\( \rightarrow \) Posterior PDF \( p(\theta|y) \)
Analytical toy example

\[ \theta: \text{Property field (contamination values)} \]
\[ y = g(\theta) + \epsilon: \text{(Local) measurements} \]
\[ \rightarrow \text{Posterior PDF } p(\theta|y) \]

\[ R(\theta): \text{Quantity of property field} \]
\[ \rightarrow P(R(\theta) \geq T | y) = ? \]
Analytical toy example

$\theta$: Property field (contamination values)
y = $\mathcal{G}(\theta) + \epsilon$: (Local) measurements
→ Posterior PDF $p(\theta | y)$

$\mathcal{R}(\theta)$: Quantity of property field
→ $\mathbb{P}(\mathcal{R}(\theta) \geq T | y) = ?$

Linear quadratic dose response

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

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

- 20 chains, 1 Million iterations each
- Gaussian proposals, AR 30%
- Convergence after 2’000 iterations (Gelman–Rubin)
Energy based model approach

• \( P(R(\theta) \geq T \mid y) = \int_T^\infty p(r \mid y) \, dr \)

• \( R = R(\theta), \quad p(r \mid y) = \frac{\exp(-F(r))}{\int \exp(-F(s)) \, ds} \rightarrow \text{Find free energy } F(r) \)
Energy based model approach

1. \( P(R(\theta) \geq T \mid y) = \int_T^\infty p(r|y) \, dr \)

2. \( R = R(\theta), \quad p(r|y) = \frac{\exp(-F(r))}{\int \exp(-F(s)) \, ds} \Rightarrow \text{Find free energy } F(r) \)

3. Introduce bias potential \( V: \mathbb{R} \to \mathbb{R}, r \mapsto V(r) \)

   \[
p_V(r) = \frac{\exp(-(F(r) + V(r)))}{\int \exp(-(F(s) + V(s))) \, ds}
\]

4. Select PDF \( p(r) \) with mass on \([T, \infty)\)
Energy based model approach

- \( P(\mathcal{R}(\theta) \geq T | y) = \int_T^{\infty} p(r|y) \, dr \)

- \( R = \mathcal{R}(\theta), \ p(r|y) = \frac{\exp(-F(r))}{\int \exp(-F(s)) \, ds} \rightarrow \text{Find free energy } F(r) \)

- Introduce bias potential \( V: \mathbb{R} \rightarrow \mathbb{R}, r \mapsto V(r) \)

\[
p_V(r) = \frac{\exp(-(F(r)+V(r)))}{\int \exp(-(F(s)+V(s))) \, ds}
\]

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

- Find \( V_{opt}(r) \) minimizing \( V \mapsto KL(p||p_V) \)
Energy based model approach

- \( \mathbb{P}(R(\theta) \geq T \mid y) = \int_T^\infty p(r \mid y) \, dr \)
- \( R = R(\theta), \ p(r \mid y) = \frac{\exp(-F(r))}{\int \exp(-F(s))ds} \rightarrow \text{Find free energy } F(r) \)
- Introduce bias potential \( V: \mathbb{R} \rightarrow \mathbb{R}, r \mapsto V(r) \)
  \[ p_V(r) = \frac{\exp(-(F(r) + V(r)))}{\int \exp(-(F(s) + V(s)))ds} \]
- Select PDF \( p(r) \) with mass on \([T, \infty)\)
- Find \( V_{opt}(r) \) minimizing \( V \mapsto KL(p || p_V) \)
- \( F(r) = -\log(p(r)) - V_{opt}(r) \)

Variational approach
Valsson and Parrinello (2014)
- Loss function
- Rare event probabilities
Implementation

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

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

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

Initialize $V_\psi$

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

Update $V_\psi$

Update $\psi$ using stochastic gradient descent to minimize $KL(p||p_V)$
Implementation

1. Initialize $V_\psi$
2. Sample $r_i \sim p_{V_\psi}(r)$
3. Update $V_\psi$
Implementation

Initialize $V_{\psi}$

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

Update $V_{\psi}$
Implementation

Initialize $V_{\psi}$

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

Update $V_{\psi}$
Implementation

Initialize $V_\psi$

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

Update $V_\psi$
Implementation

Initialize $V_\psi$

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

Update $V_\psi$

$V(r)$

$\begin{align*}
V(r) & = \begin{cases} 
0 & \text{for } r < 0 \\
10 & \text{for } 0 \leq r < 10 \\
20 & \text{for } 10 \leq r < 20 \\
30 & \text{for } 20 \leq r < 30 \\
40 & \text{for } 30 \leq r < 40 \\
0 & \text{for } r \geq 40 
\end{cases}
\end{align*}$

$\begin{align*}
r & = \begin{cases} 
0 & \text{for } r < 0 \\
10 & \text{for } 0 \leq r < 10 \\
20 & \text{for } 10 \leq r < 20 \\
30 & \text{for } 20 \leq r < 30 \\
40 & \text{for } 30 \leq r < 40 \\
0 & \text{for } r \geq 40 
\end{cases}
\end{align*}$
When do we stop?
Kernel Stein discrepancy

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

\[
KSD \left( p \mid p_{V, \psi} \right) = \sqrt{\frac{1}{n^2} \sum_{i,j=1}^{n} k_p(x_i, x_j), \ x_i \sim p_{V, \psi}(\cdot),}
\]

\[
k_p(x, y) = \nabla_x \nabla_y k(x, y) + \langle \nabla_x k(x, y), \nabla_y \log p(y) \rangle + \langle \nabla_y k(x, y), \nabla_x \log p(x) \rangle
\]

\[
+ k(x, y) \langle \nabla_x \log p(x), \nabla_y \log p(y) \rangle
\]

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 \( \alpha \)) → Conservative for correlated samples (Chwialkowski et al. 2016)
Kernel Stein discrepancy

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

\[
KSD\left( p | p_{V,\psi}\right) = \sqrt{\frac{1}{n^2} \sum_{i,j=1}^{n} k_p(x_i, x_j), \ x_i \sim p_{V,\psi}()}.
\]

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

Use kernelized Stein discrepancy for goodness-of-fit tests with \( H_0 : p = p_{V,\psi} \)

\( \rightarrow \) Employing bootstrap procedure

\( \rightarrow \) Stop when \( H_0 \) cannot be rejected anymore (significance level \( \alpha \))

\( \rightarrow \) Conservative for correlated samples (Chwialkowski et al. 2016)
Stopping criteria

- V(r)
- Samples p(r) and pV(r)
- Loss
- Risk probability
- KSD
- P-value

Iteration vs. Various Metrics
Energy based models vs. MCMC

- $\mathbb{P}(R(\theta) \geq 20.0 \mid y) = 1.76 \times 10^{-6}$
- MCMC (left): 1,000,000 iterations $\rightarrow$ SD = $4.2 \times 10^{-6}$
Energy based models vs. MCMC

- $\mathbb{P}(R(\theta) \geq 20.0 \mid y) = 1.76 \times 10^{-6}$
- MCMC (left): 1'000'000 iterations $\Rightarrow$ SD = $4.2 \times 10^{-6}$
- EBM (right): 14’000-50’000 iterations $\Rightarrow$ SD = $1.5 \times 10^{-6}$
1D flow example (Straub et al. 2016)

- Hydraulic diffusivity field $a(x), x \in [0m, 1m]$
  \[ \ln a(x) = \mu_{\ln \theta} + \sigma_{\ln \theta} \sum_{i=1}^{10} \sqrt{w_i} v_i(x) Z_i \]
  \[ \theta \]

  $w_i$ Log-diffusivity Gaussian, Karhunen-Loève expansion
- Diffusivity = speed at which pressure pulse propagates through aquifer
1D flow example

Data

![Graph showing pressure data with x and y axes labeled. The graph includes a peak and two arrows pointing upwards.]
1D flow example

Risk

\[ R(\theta) = \text{outflow rate} \]
Energy based models vs. MCMC

\[ \mathbb{P}(R(\theta) \geq 1.15 \mid y) \]
Energy based models vs. MCMC

\[ \mathbb{P}(R(\theta) \geq 1.15 \mid y) \]

\[ 7.8 \times 10^{-8} \]
Energy based models vs. MCMC

\[ + \mathbb{P}(R(\theta) \geq 1.10 \mid y) \]
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 $V(r)$
  - Sampling $p_V(r)$
  - Choice of $p(r)$
  - Numerical integration
  - Learning rate
  - 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
References


Implementation

- Initialize $\psi$
- Sample $r_i \sim p_{V\psi}(r)$
- Update $\psi$

- $p_{V\psi}(r) = \frac{\exp(-F(r) + V_\psi(r))}{\int \exp(-F(s) + V_\psi(s)) ds}$

- 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 $p_{V\psi}(\theta) = \frac{\exp(-U(\theta) + V_\psi(\mathcal{R}(\theta)))}{\int \exp(-U(\xi) + V_\psi(\mathcal{R}(\xi))) d\xi}$

- Transform samples with $\theta \mapsto \mathcal{R}(\theta)$
Implementation

- Initialize $\psi$
- Sample $r_i \sim p_{V, \psi}(r)$
- Update $\psi$

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** $V(r)$
  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** $p(r)$
  $\mathcal{N}(20,4)$

- **Stopping criteria**
  $\alpha = 0.05$, 1000 bootstrap samples
Configuration 1D flow example

- **Parameterization** $V(r)$
  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** $p(r)$
  $\mathcal{N}(1.2,0.125)$

- **Stopping criteria**
  $\alpha = 0.01$, 1000 bootstrap samples