Rare Events and their Optimization

Vishwas Rao
vhebbur@anl.gov

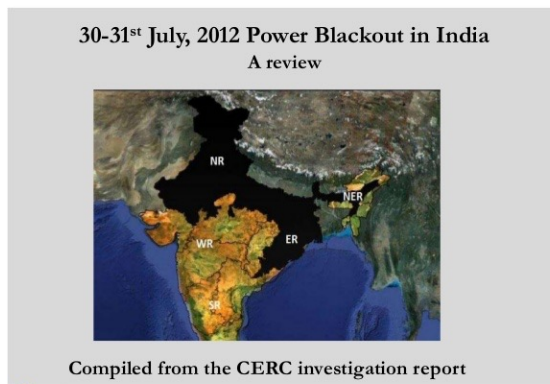
Collaborators: Shanyin Tong (NYU), Anirudh Subramanyam (ANL), Mihai Anitescu (ANL & UChicago), Emil Constantinescu (ANL), and Romit Maulik (ANL)

HEP (05/18/2023)

Motivation

- Extreme weather events
  - Flooding
  - Earthquakes
  - Tornadoes
- Cascading power failures
  - 2003 in USA
  - 2012 in India
  - 2012 in NoVA region
  - 2021 in Texas
- Huge costs incurred (US only)
  - 2017 – 174 Billion
  - 2018 – 155 Billion

Hurricane Harvey caused extreme flooding in parts of Houston, TX - 2017



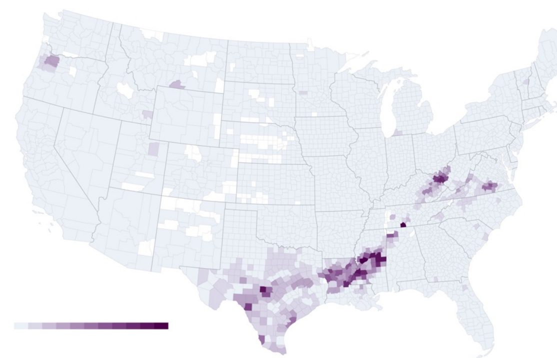
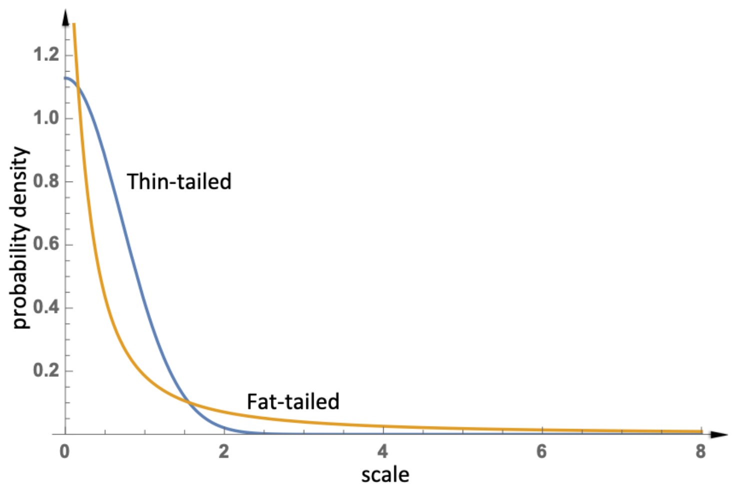


Image credit: AICHE, CERC investigation report

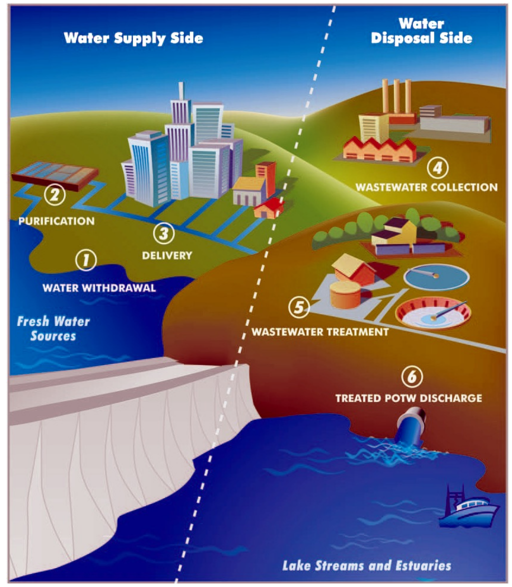
A thought experiment

- Components of a system are independent from one another (above a certain scale)
  - At much larger scales the magnitude of fluctuations of the system follows a normal distribution (loosely follows from the central limit theorem)
  - Probabilities of the events many std deviations from the mean **are astronomically improbable.**
  - For example: Consider 100 independent ladders each with 1/10 probability of falling.
- Components of a system are interdependent
  - Interdependencies can lead to a distribution of fluctuations in which the probability of an extreme event, while still small is **not astronomically small.**
  - If we tie all the ladders together, while the probability of an individual ladder falling is smaller – but we would have significantly increased the probability of all the ladders collapsing.

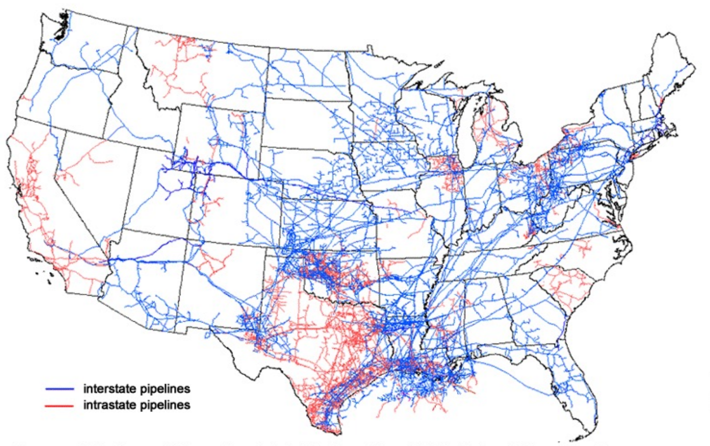


Concept and Figure borrowed from **An Introduction to Complex Systems Science and its Applications**

Examples of Complex Systems



Map of U.S. interstate and intrastate natural gas pipelines



Source: U.S. Energy Information Administration, *About U.S. Natural Gas Pipelines*

Texas Blackout: Before

Texas Blackout: After. About 4.4 Million people were affected

Blackout in India: 2012. About 620 Million people were affected.

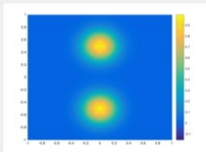
Causes for disruption

- Weather events
  - Texas 2021, California 2020
- Poor planning
  - Texas 2021,
  - NE US 2003 (A software bug in the alarm system)
  - California 2020.
- Terrorist attacks
  - Ukraine Cyberattack (2015)

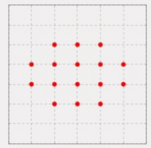
Common Themes

- **Rare**
  - O(100) Power outages per year in US
  - O(1000) large floods in the last 35 years
- **High-Impact**
  - Costs O($1B)
  - Huge societal costs
- **Questions**
  - Risk Quantification
  - Risk Mitigation

Design and Control Space



Continuous
and Discrete

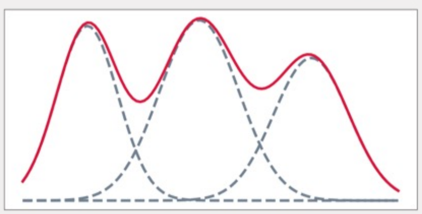


Risk Quantification
for a fixed design

Risk Mitigation
to find robust designs

Probabilistic constraints

Uncertainty in Data and Models

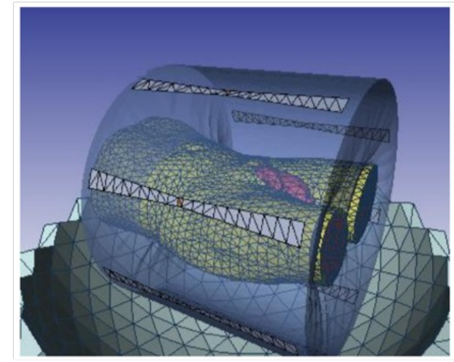


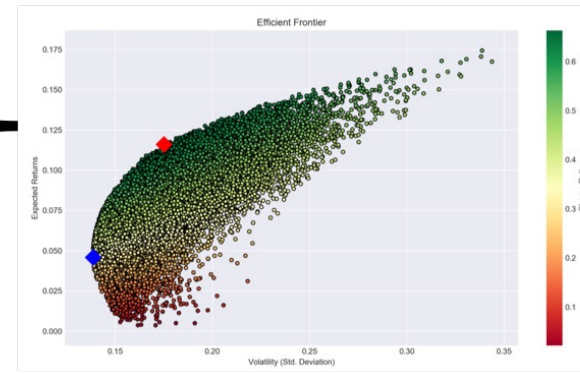
Model:

$$\vec{\nabla} \cdot (\rho \vec{u}) = 0$$

$$\frac{\partial(\rho \vec{u})}{\partial t} + \vec{\nabla} \cdot \rho \vec{u} \otimes \vec{u} = -\vec{\nabla} p + \vec{\nabla} \cdot \bar{\bar{\tau}} + \rho \vec{f}$$

$$\frac{\partial(\rho e)}{\partial t} + \vec{\nabla} \cdot (\rho e + p) \vec{u} = \vec{\nabla} \cdot (\bar{\bar{\tau}} \cdot \vec{u}) + \rho \vec{f} \vec{u} + \vec{\nabla} \cdot \vec{q} + r$$





Rare event problem formulation

- Estimating the probability of rare events: $\mathbb{P}(F(\mathbf{x},\theta) \geq z)$
- Optimization under rare events:

Risk Mitigation

Risk Quantification

$$\begin{aligned} \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \quad & c(\mathbf{x}) \\ \text{subject to} \quad & \mathbb{P}(F(\mathbf{x},\theta) \geq z) \leq \alpha \quad \text{for some fixed } \alpha \ll 1. \end{aligned}$$

$F : \mathcal{X} \times \Theta \to \mathbb{R}$: Limit state function

$\theta \in \Theta$: random parameter

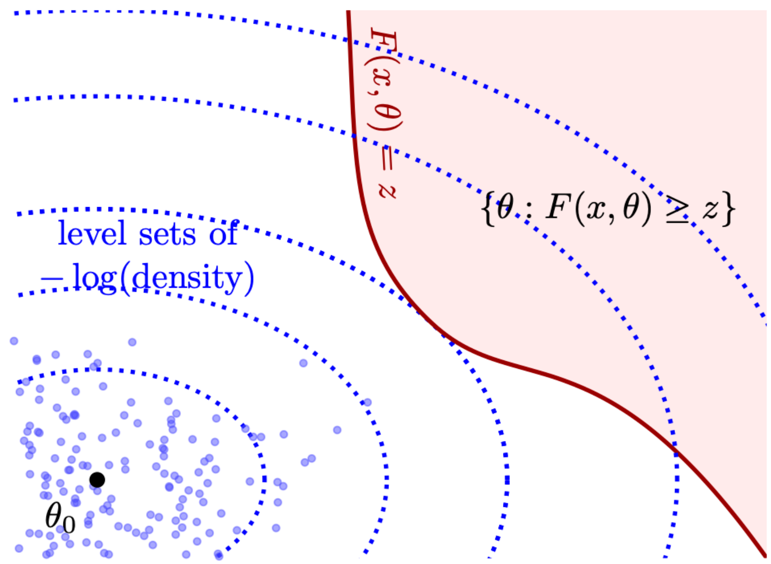
$\mathbf{x} \in \mathcal{X}$: control, can be infinite-dimensional

$z \in \mathbb{R}$: threshold at which the system fails

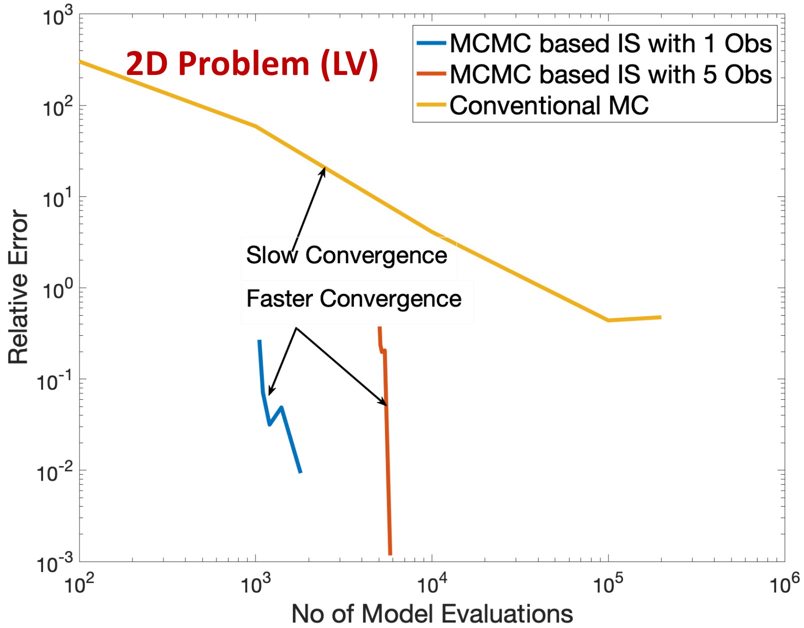
$\alpha \in (0,1)$: risk of failure

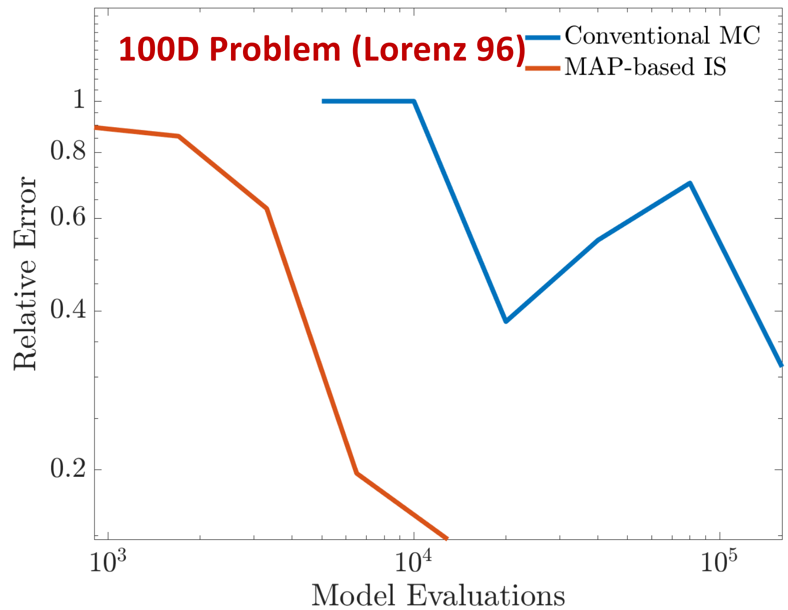
Challenges

- Not enough data that corresponds to the rare events
- Computationally expensive
  - For example: estimating the odds of an event whose probability is ~ 1e-3, for an underlying simulation that requires 10 minutes per simulation, requires two years of serial computation for a std. dev of 10%.
- Mitigation is even more harder
  - Sampling-based methods require $O(\alpha^{-1})$ samples
  - Optimization problem size grows linearly with samples



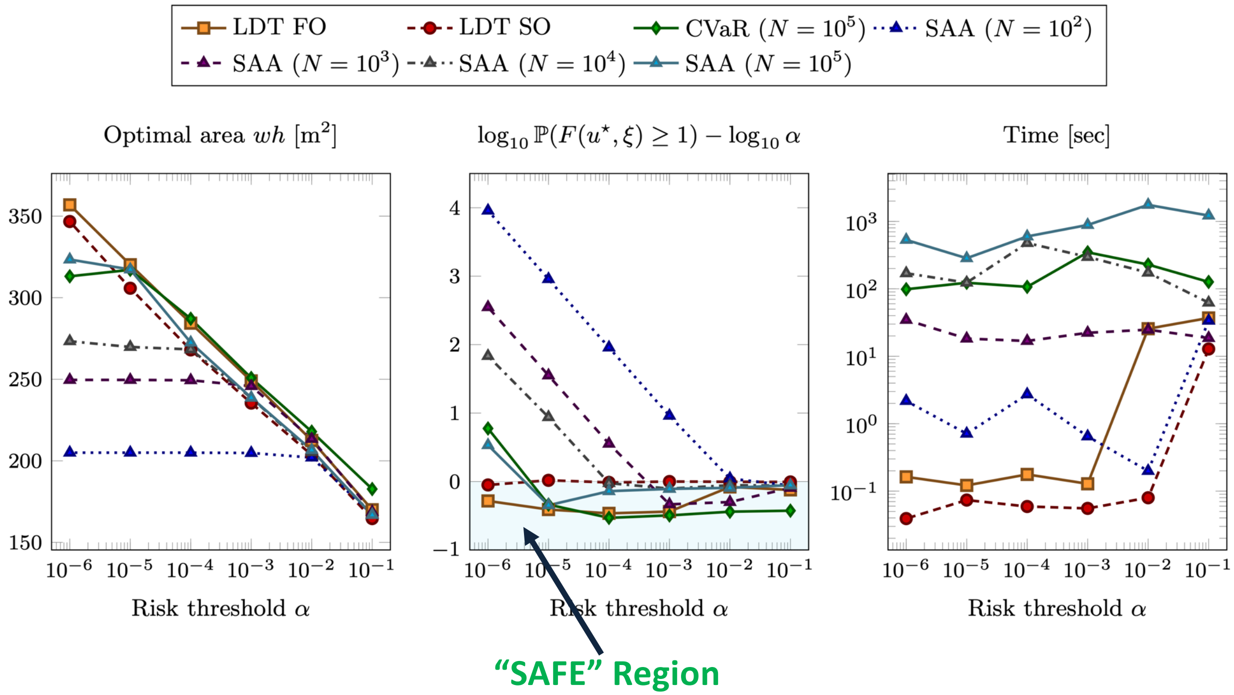
Overview of results: Risk Quantification





- Accurate results with O(1000) samples for small systems
- Uses **Rice's formula** + Bayesian Inference
- Extensible to larger systems

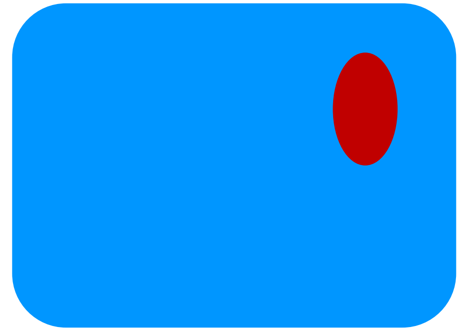
Overview of results: Risk Mitigation



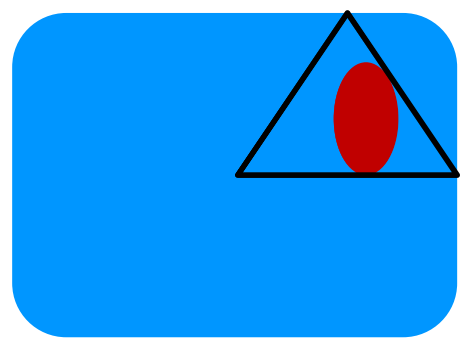
- Sampling free methods (**LDT + Bilevel optimization**)
- Scalable for ``Extremely rare events"
- Works for Gaussian and Gaussian mixtures

Existing approaches

- **Monte Carlo simulation**
  - Draw random samples from p
  - Simulate the dynamics with each random sample
  - Compute the fraction of the samples that exceed the threshold
  - $P_u^{MC}(\mathbf{x}_0^1, \ldots, \mathbf{x}_0^M) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(\mathbf{x}_0^i)$



- **Importance sampling**:
  - Construct a ``suitable” biasing distribution
  - Draw random samples from the biasing distribution
  - Sum the “importance weights” for the samples that exceed threshold
  - $P_u^{IS}(\bar{\mathbf{x}}_0^1, \ldots, \bar{\mathbf{x}}_0^M) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(\bar{\mathbf{x}}_0^i)\frac{p(\bar{\mathbf{x}}_0^i)}{q(\bar{\mathbf{x}}_0^i)}$



Problem formulation

- Consider the following dynamical system

$$\mathbf{x}' = f(t, \mathbf{x})\,, \quad t = [0, T]$$
$$\mathbf{x}(0) = \mathbf{x}_0\,, \quad \mathbf{x}_0 \sim p\,, \quad \mathbf{x} \in \Omega\,,$$

- We are interested in estimating $P_T(z) := \mathbb{P}(F(\mathbf{x}, \theta) > z)$

$$F(\mathbf{x}, \theta) := \max \mathbf{c}^\top(\mathbf{x}(t, \mathbf{x}_0)) \geq z$$

- The initial state of the system has the PDF \`p’. When f is linear and p is Gaussian, the evolution of x is analytically tractable. However, when f is nonlinear, the evolution of x is not analytically tractable.

Sketch of our Approach

- Let $\Omega(z) \subset \Omega$ denote the set of all initial conditions that cause an excursion. That is:

$$\Omega(z) := \left\{ \mathbf{x}_0 : \sup_{[0,T]} \mathbf{c}^\top \mathbf{x}(t, \mathbf{x}_0) \geq z \right\}$$

- Then $P_T(z)$ is the $\mathtt{measure}(\Omega(z))$

- We use Rice's formula to gain insights about $\Omega(z)$

Our Approach: Rice's formula + Bayesian Inference

- Rice's formula: $$\mathbb{E}\left\{N_z^+(0,T)\right\} = \int_0^T \int_0^\infty y\varphi_t(z,y)\,\mathrm{d}y\,\mathrm{d}t\,.$$

Expected # of excursions **Derivative of the SP** **Joint PDF**

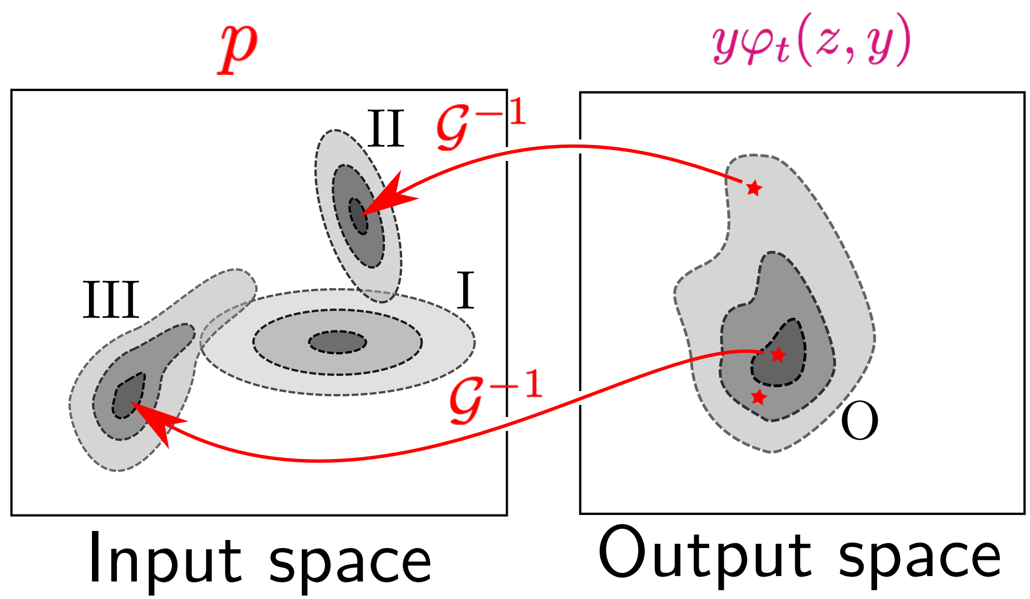
- For Gaussian processes: $$\left|\mathbb{P}\left\{\sup_{t\in[0,T]} \mathbf{c}^\top \mathbf{x}(t) \geq z\right\} - \mathbb{E}\left\{N_z^+(0,T)\right\}\right| \leq \mathcal{O}\left(e^{-\beta z^2}\right),$$

Faster than exponential convergence

- Unfortunately, $\varphi_t(z,y)$ is analytically computable only for Gaussian processes
- The key idea is that values 't' and `y' at which $y\varphi_t(z,y)$ is large contribute the most to integral in Rice's formula. We use this idea to construct a biasing distribution.

Ideas for constructing Biasing distribution

- Consider the forward map $\mathcal{G} : \mathbb{R}^{d\times 1} \to \mathbb{R}^2$ that evaluates $\begin{bmatrix} \mathbf{c}^\intercal \mathbf{x}(t) \\ \mathbf{c}^\intercal \mathbf{x}'(t) \end{bmatrix}$ using the dynamics for a given initial conditions and at a specified time t.
- $\Omega(z)$ can be approximated by $\mathcal{G}^{-1}\left(\begin{bmatrix} z \\ y_i \end{bmatrix}\right)$
- This is ill-posed; there are multiple x_0's that map to a given $\begin{bmatrix} z \\ y_i \end{bmatrix}$.



Biasing distribution is II U III

Overview of our approach

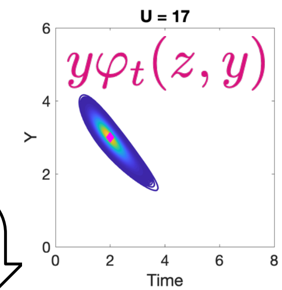
Prior

Gaussian Uncertainties
$\mathbf{x}_0 \sim \mathcal{N}(\overline{\mathbf{x}}_0, \Sigma)$

Nonlinear Dynamical systems
$\mathbf{x}' = f(t, \mathbf{x}), \quad t \in [0, T]$

Non-Gaussian Processes

Linearization + Rice's formula
$\mathbf{x}' \approx \mathbf{F} \cdot (\mathbf{x} - \overline{\mathbf{x}}_0) + f(\overline{\mathbf{x}}_0)$

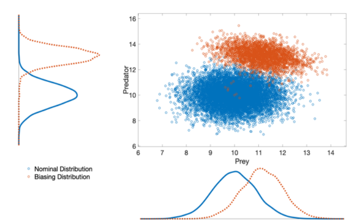


Likelihood
$\exp\left(-\frac{1}{2}\left\|\mathbf{y}_i - \mathcal{G}(\mathbf{x}_i, t_i)\right\|^2_{\Gamma_i^{-1}}\right)$

MAP/MCMC or ML-based methods

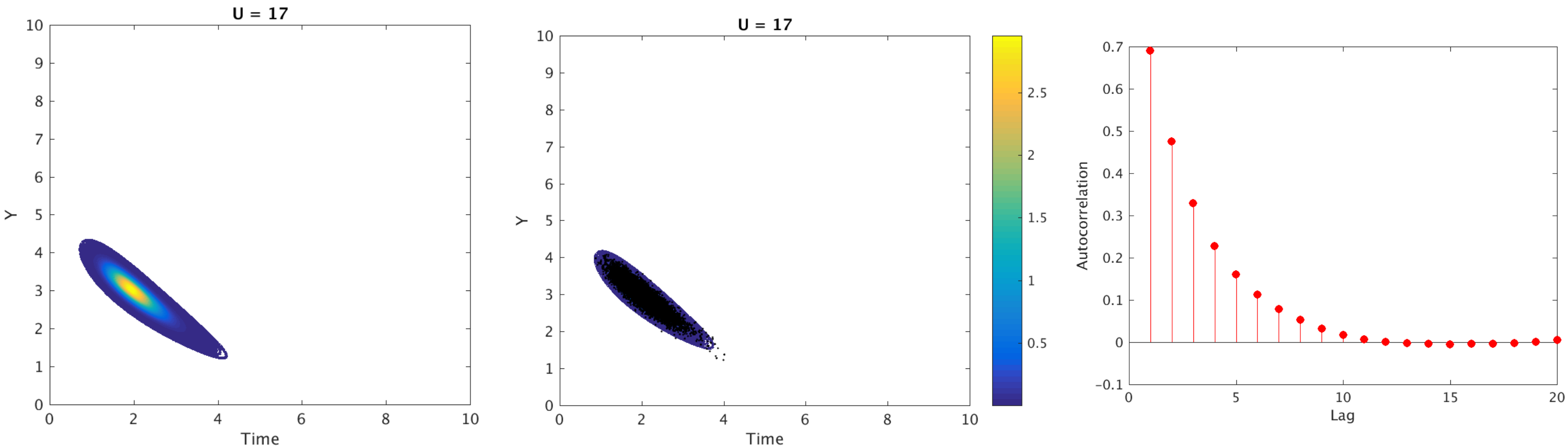
Posterior/ IBD

Importance Sampling



Excursion Probability Estimates

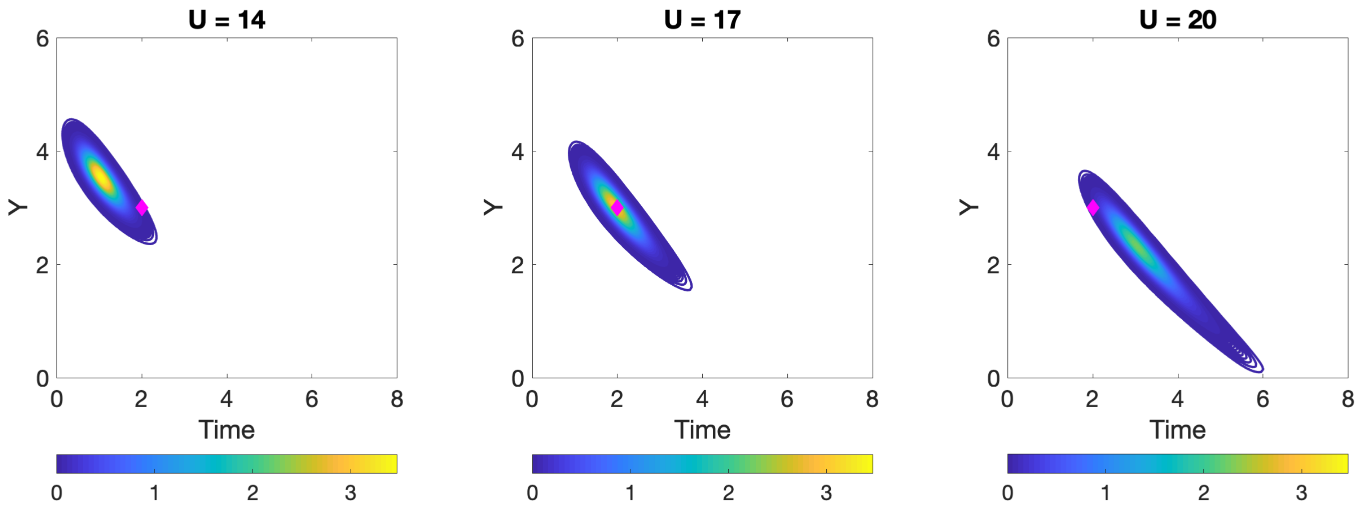
Sampling from $y\varphi_t(u, y)$



Left: The product of the derivative and the joint PDF of the state and its derivative for u=17. **Center**: Samples drawn from $y\varphi_t(u, y)$ using DRAM MCMC. These samples will be used to construct the likelihood. **Right**: Autocorrelation between the samples. Picking every tenth sample will “ensure” independence

Choice of mean and covariance of likelihood

- Sampling from $y\varphi_t(u, y)$ gives us the likely time at which there is an excursion. So to determine the mean and covariance of the likelihood, we look at $y_i, u_i \mid t_i$



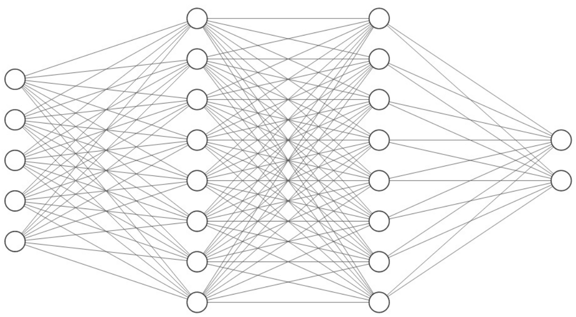
- Use Laplace approximation to estimate mean and covariance

Approaches to solve Bayesian inverse problems

- Laplace approximation at MAP (MAP-based IS) → **[R, Anitescu, In Press, SIAM JUQ]**
  - Solve the inverse problem using the negative log likelihood as cost function
  - Use the Hessian inverse at the MAP point to approximate the covariance of the posterior
  - Use LBFGS to solve the optimization problem (Poblano toolbox)
  - Adjoints to obtain the gradient
- MCMC-based IS → **[R, Anitescu, In Press, SIAM JUQ]**
  - Sample directly from the posterior
  - DRAM algorithm
- Machine Learning based approach to find the inverse maps →**[MLDADS (2020)]**

Inverse map using neural networks

- **The inverse map may be expensive to obtain**
  - Solving Bayesian inverse problems is expensive
  - Can we build data-driven surrogates for approximating pre-images?
  - We may utilize a fully connected neural network to approximate this map.

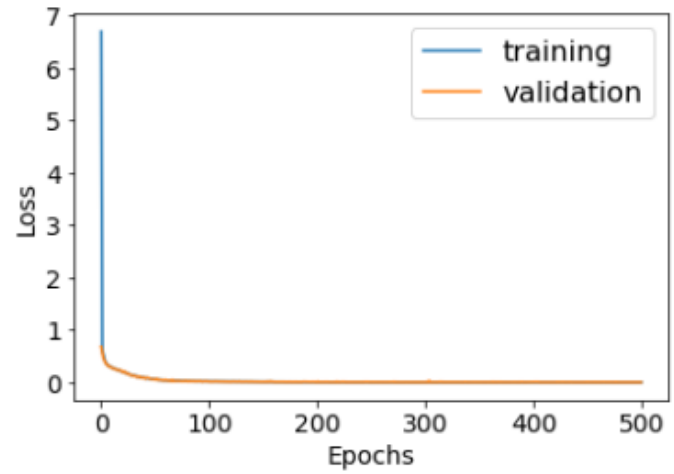


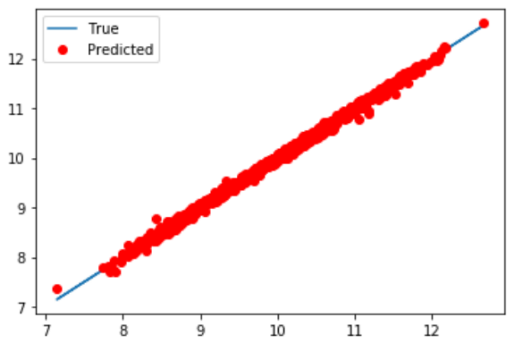
A fully connected artificial neural network

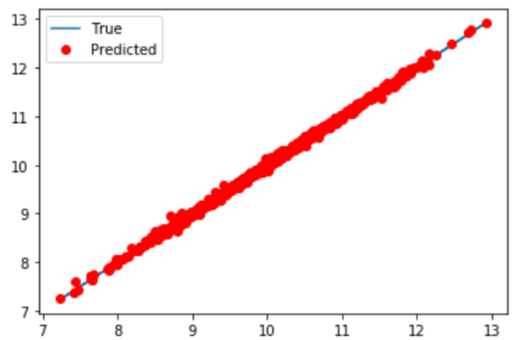
- **Learning**
  - We train our map given multiple examples of forward simulations.
  - The input to the map is the outcome of the simulation and the initial condition is the output.
  - Our neural network training is a non-convex optimization – we use the ADAM Stochastic Gradient optimizer with a learning rate of 0.001. Consistently reducing accuracy on a held out data-set is used to terminate optimization (i.e., the prevention of overfitting)

Inverse map using neural networks

- **Quantifying map fidelity**
  - We track the reducing objective function until improvement has stalled
  - The held-out (validation) data is used to select the best model
  - Scatter plots on the 'test' data (i.e., data completely unused till this point) show quality of predictions







Two plots for each dimension of outcome
Map evaluation in <1e-3 seconds

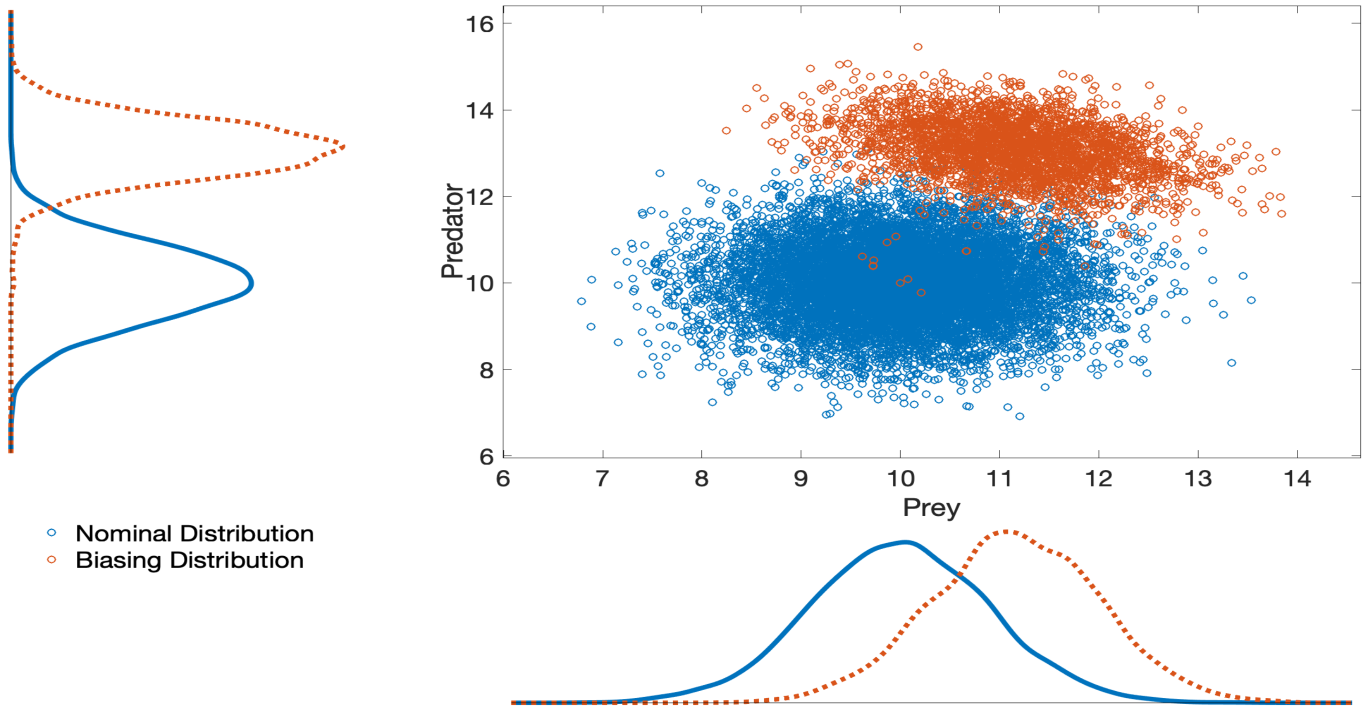
Nonlinear Example

- We consider the Lotka Volterra example:

$$\frac{dx_1}{dt} = \alpha x_1 - \beta x_1 x_2 \,,$$
$$\frac{dx_2}{dt} = \delta x_1 x_2 - \gamma x_2 \,, \quad \mathbf{x}(0) \sim \mathcal{N}\left(\begin{bmatrix}10\\10\end{bmatrix}, 0.8 \times I_2\right)$$

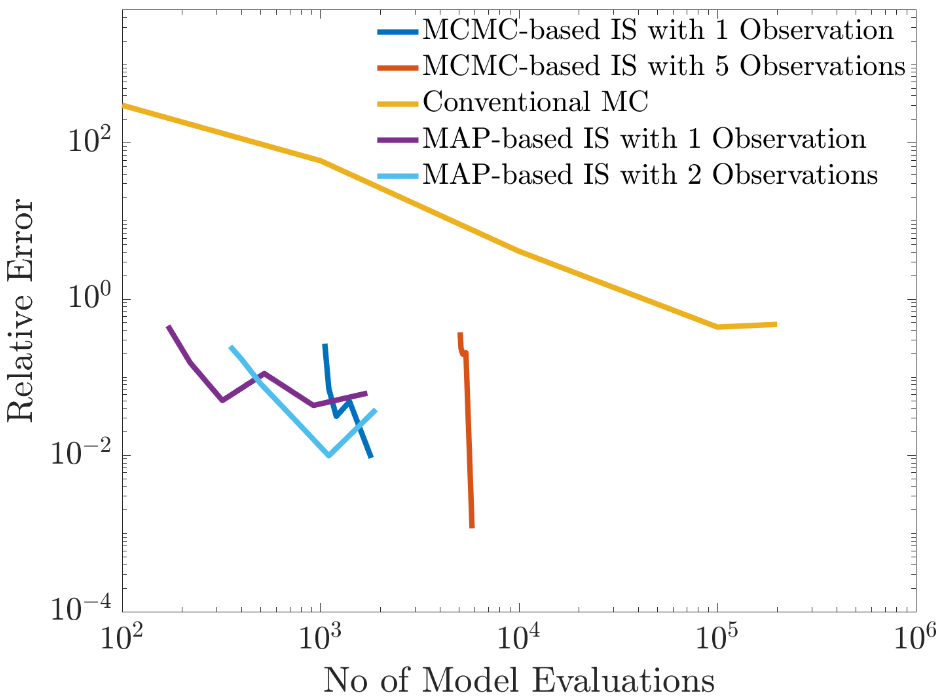
- We are interested in estimating x2 exceeding 17
- We also look at Lorenz96 system (100D)

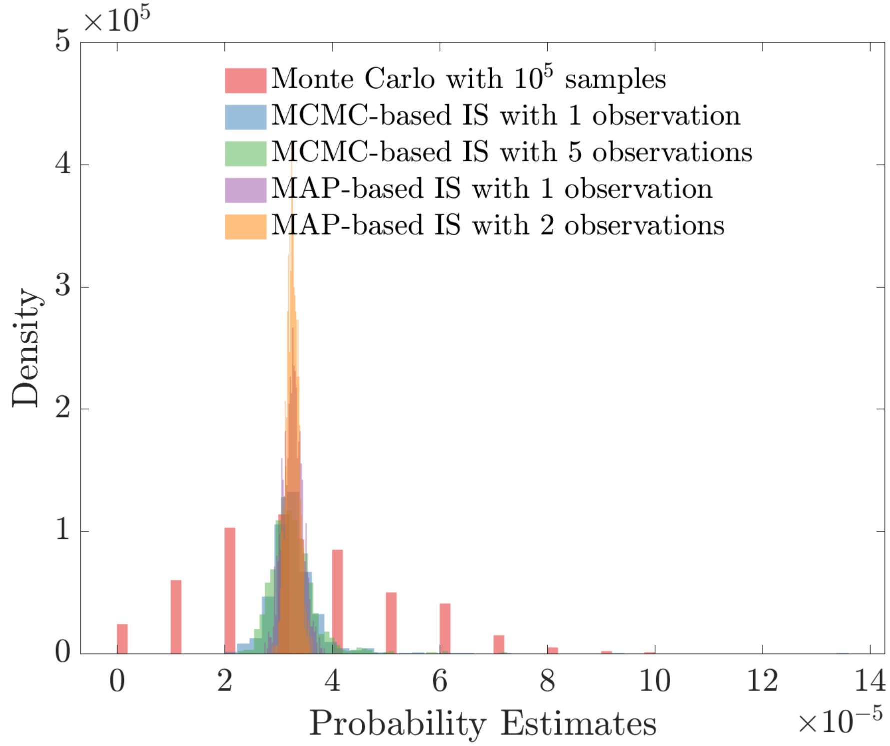
Nominal and biasing distributions



Samples from Nominal and Biasing distributions. The Biasing distributions mainly picks samples from the tails. This is obtained using the MCMC based IS approach.

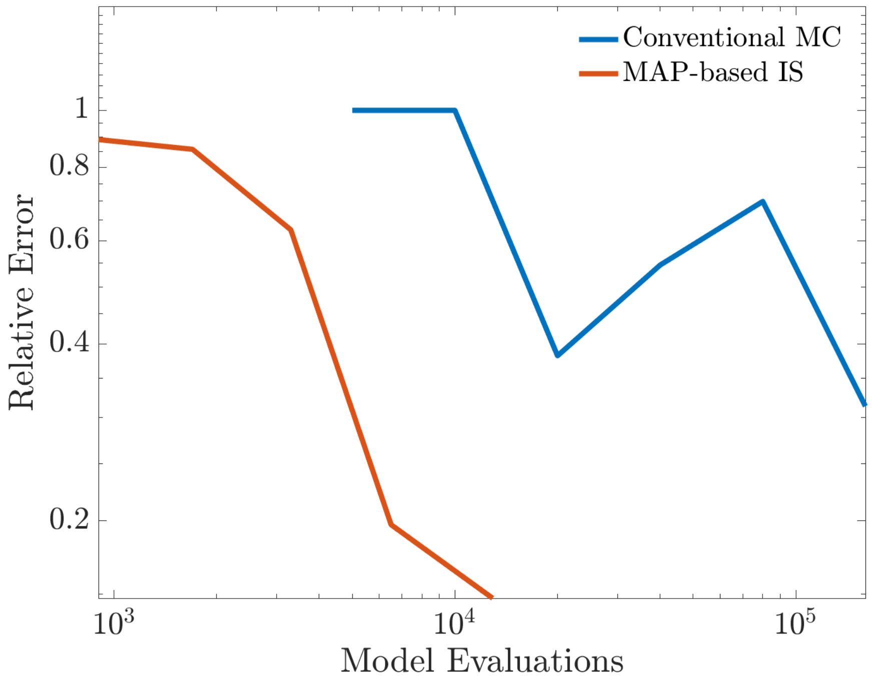
Results (Gaussian Uncertainties)

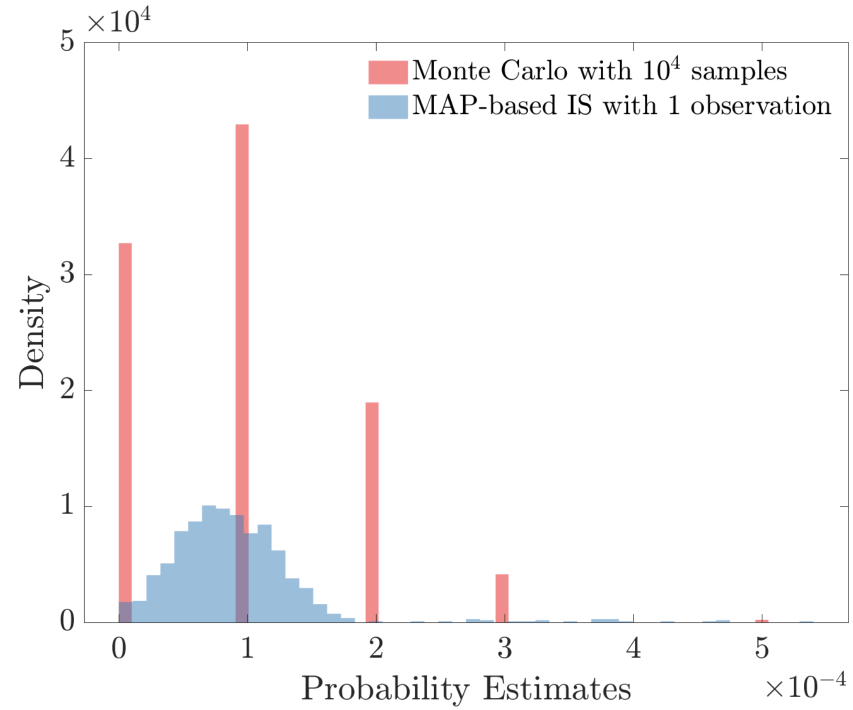




- **``True" probability is 3.28e-5 (Obtained with 10 Million samples of MC)**
- **Both MCMC and MAP based IS yield comparable results for similar amount of "work"**

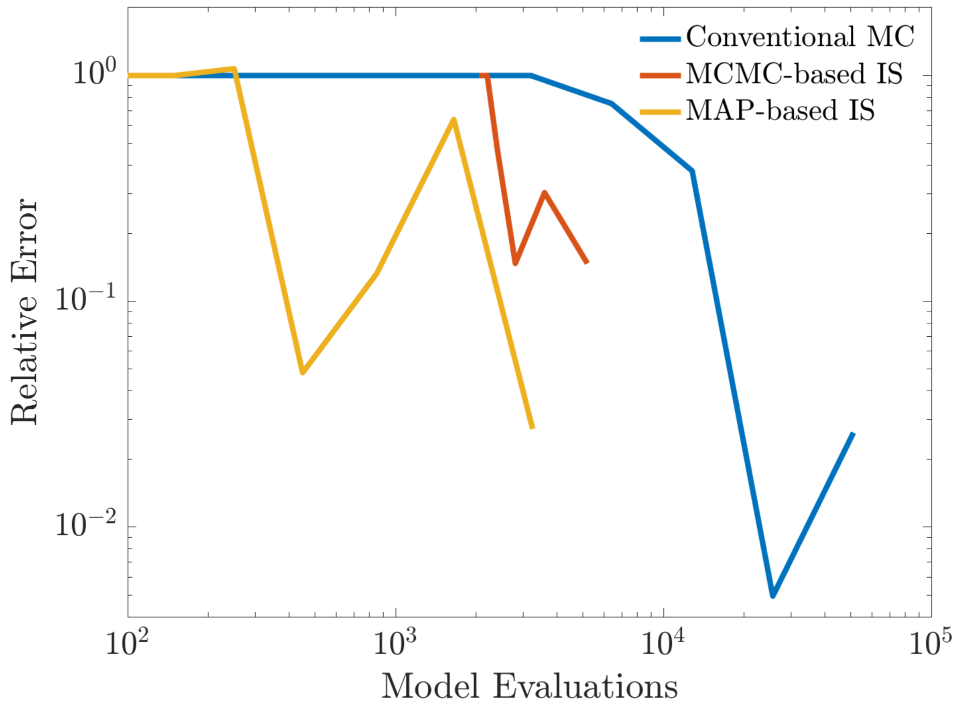
Results (Gaussian Uncertainties)





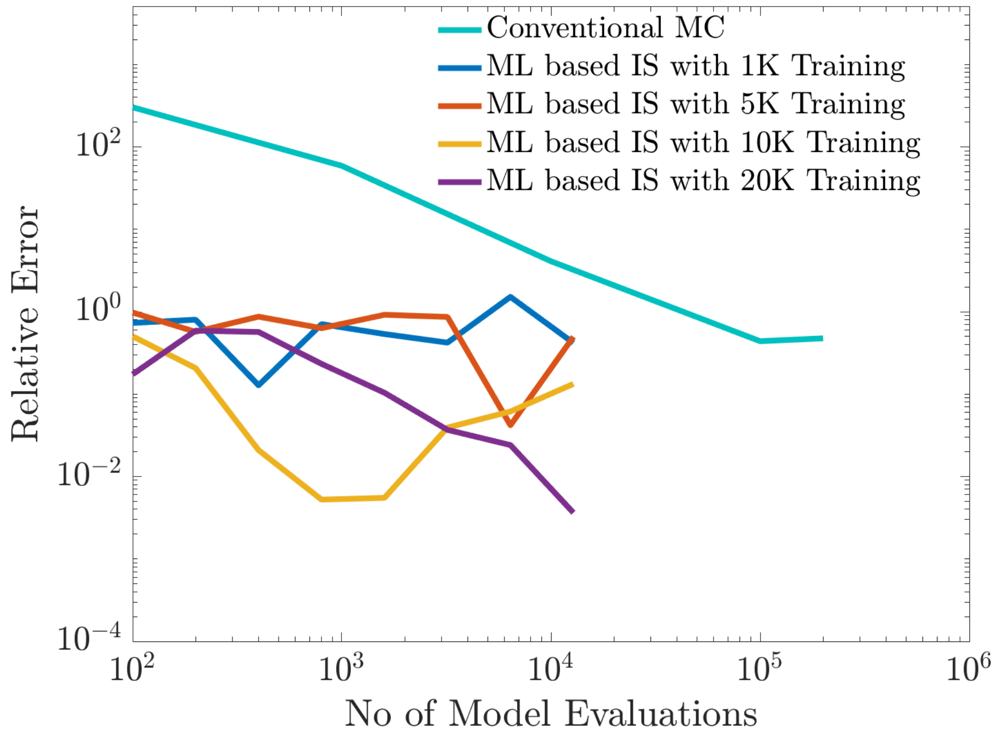
- ``True" probability is 8.09e-5 (Obtained with 10 Million samples of MC)

Results (non-Gaussian Uncertainties)



- **The true probability here is 6.281 x 10^{-4}.**
- **Convergence of different approaches with an uniform initial distribution of the state.**
- **The convergence is not as smooth as it is for a Gaussian initial distribution, and we attribute the cause to the edge effects of a uniform distribution.**

Results



Comparison between Conventional MCS and ML-based IS. We observe even with small number training data, we obtain fairly accurate estimates and as we increase the training data, the accuracy improves dramatically.

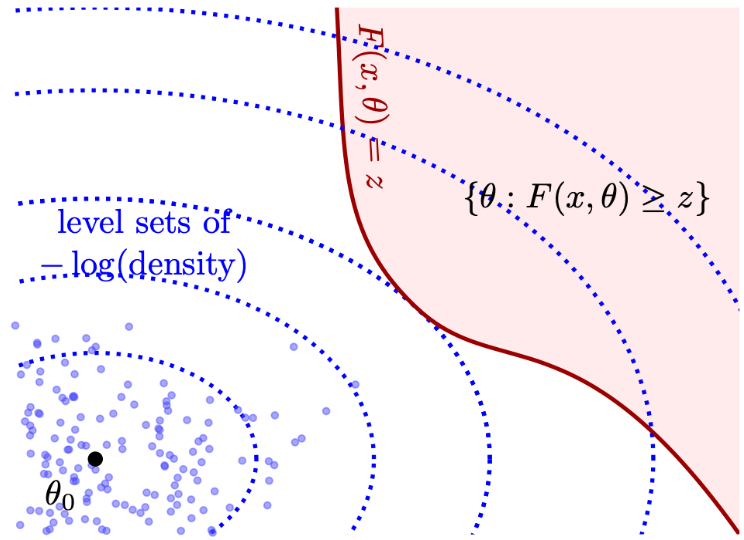
Computational Cost

- Generating 20K training data points cost approximately equivalent to 400 Model evaluations
- Training the dataset required 180 seconds on an 8th generation Intel Core I7 with python 3.6.8 (this is equivalent to about 50 model evaluations).
- Inference costs were negligible.

Approaches to approximate probabilistic constraints

- **Sample average approximation**

$$\min_{\mathbf{x}\in\mathcal{X}} \quad c(\mathbf{x})$$
$$\text{s.t.} \quad \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}_{F(\mathbf{x},\theta_k)\geq z} \leq \alpha$$



- **CVaR:** Convex approximation of constraint

$$\min_{\mathbf{x}\in\mathcal{X}} \quad c(\mathbf{x})$$
$$\text{s.t.} \quad \mathrm{CVaR}_{1-\alpha}[F(\mathbf{x},\theta)-z] \leq 0$$

$$\mathrm{CVaR}_{1-\alpha}(Z) := \inf_{t\in\mathbb{R}} \left[t + \tfrac{1}{\alpha}\mathbb{E}[Z-t]_+\right]$$

Large Deviation Theory - A quick introduction

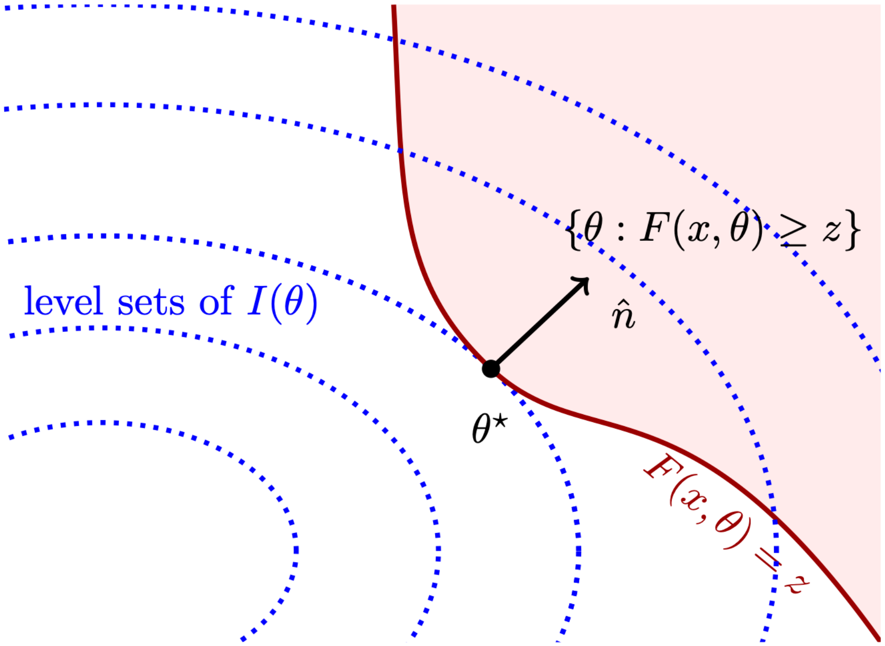
- LDT is concerned with the **asymptotic behavior of tails** of probability distributions – specifically the rates of exponential decay of probabilistic measures of extreme events.

- It uses the **rate functions** to characterize the asymptotic behavior of rare probabilities.

- Recently **Grafke, Vanden Eijnden, and Dematteis (2018, 2019)** and later **Tong, Stadler, and Vanden Eijnden (2020)** adapted **the classical LDT and sharp asymptotics** to study the behavior of rare events.

- The key idea is to find **a dominating point** in the rare event set by solving an optimization problem and estimate probabilities solely based on this.

Large Deviation Theory



- $I(\theta)$ is the Legendre-Fenchel transform of cumulant generating functions. Example:

$$\theta \sim \mathcal{N}(\theta_0, \Sigma), \quad I(\theta) = \tfrac{1}{2}\|\theta - \theta_0\|^2_{\Sigma^{-1}}.$$

- Probability computation requires optimization:

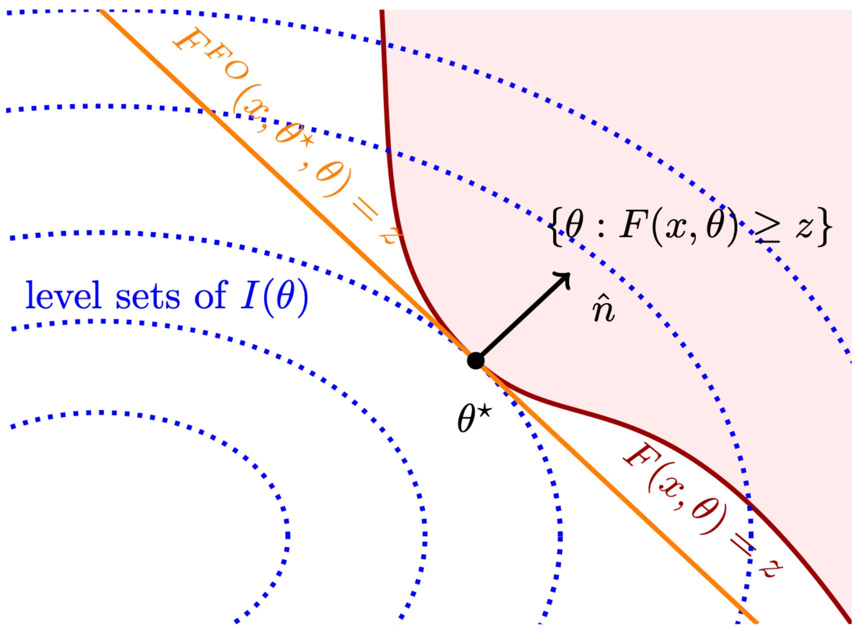
$$\mathbb{P}(F(\mathbf{x},\theta) \geq z) \asymp \exp(-I(\theta^\star)) \text{ as } z \to \infty,$$

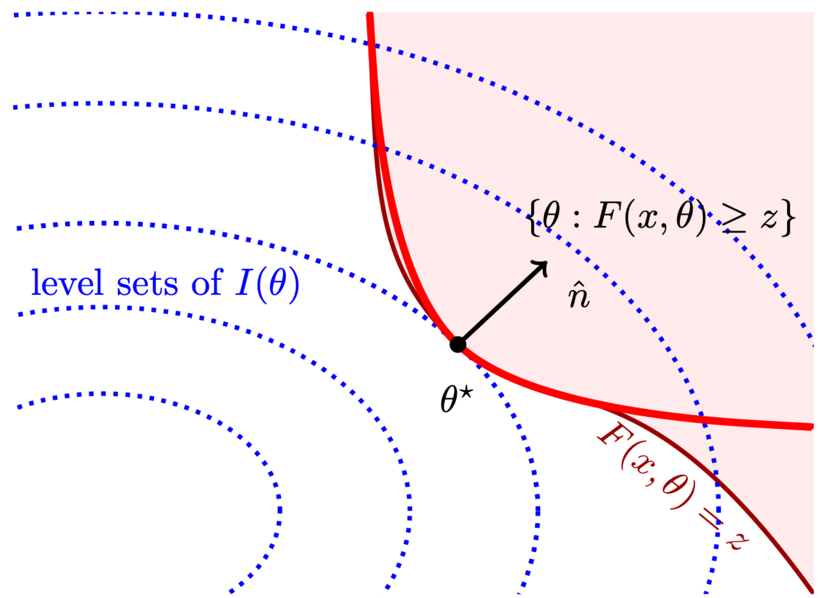
$$\theta^\star = \operatorname*{argmin}_{\theta : F(\mathbf{x},\theta) = z} I(\theta)$$

Regularity Conditions:

- F is concave w.r.t θ
- $\nabla_\theta F$ is Lipschitz continuous

Explicit formulae for chance constraints

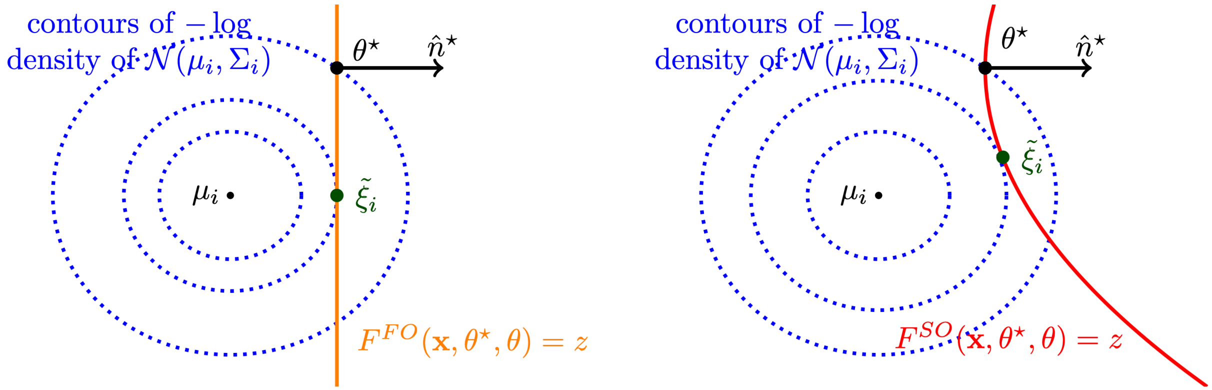




$$P_1(\mathbf{x}, \theta^\star) = \Phi(-\sqrt{2I(\theta^\star)})$$

$$P_2 \approx P_1 \left[(\hat{n}^\top H^{-1} \hat{n}) \det H\right]^{-\frac{1}{2}}$$

Probability estimation with Gaussian Mixtures



$$\tilde{\xi}_i = \operatorname*{argmin}_{F^{FO}(\mathbf{x},\theta^\star,\theta)=z} \frac{1}{2}\|\theta - \mu_i\|^2_{\Sigma_i^{-1}} = \operatorname*{argmin}_{\langle \nabla_\xi F(\mathbf{x},\theta^\star), \theta-\theta^\star\rangle = 0} \frac{1}{2}\|\theta - \mu_i\|^2_{\Sigma_i^{-1}},$$

- **First order approximation:**

$$P^{FO}(\mathbf{x}, \theta^\star) = \sum_{i=1}^{m} w_i \Phi(-\|\xi_i^\star\|),$$

$$\|\xi_i^\star\| = \frac{\langle \nabla_\theta F(\mathbf{x}, \theta^\star), \theta^\star - \mu_i\rangle}{\|\Sigma_i^{\frac{1}{2}} \nabla_\theta F(\mathbf{x}, \theta^\star)\|}.$$

- **Second order approximation:**

$$P^{SO}(\mathbf{x}, \theta^\star) = \sum_{i=1}^{m} w_i \Phi\left(-\sqrt{2I_i(\tilde{\theta}_i)}\right) \det\nolimits_\perp \left(I_d - \lambda_i \Sigma_i^{\frac{1}{2}} \nabla^2_\theta F(\mathbf{x}, \theta^\star) \Sigma_i^{\frac{1}{2}}\right)^{-\frac{1}{2}},$$

$$\lambda_i = \|\Sigma_i^{\frac{1}{2}} \nabla_{\tilde{\theta}} I_i(\tilde{\theta}_i)\| / \|\Sigma_i^{\frac{1}{2}} \nabla_{\tilde{\theta}} F^{SO}(\mathbf{x}, \theta^\star, \tilde{\theta}_i)\|,$$

$$\begin{aligned} \tilde{\theta}_i = \operatorname{argmin} \quad & \frac{1}{2}(\theta - \mu_i)^\top \Sigma_i^{-1}(\theta - \mu_i) =: I_i(\theta) \\ \text{subject to} \quad & F^{SO}(\mathbf{x}, \theta^\star, \theta) = z, \ \ i = 1, \ldots, m. \end{aligned}$$

Chance Constraints to Bilevel optimization

Optimization under Rare Chance constraints

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \quad c(\mathbf{x})$$
$$\text{subject to} \quad \mathbb{P}(F(\mathbf{x},\theta) \geq z) \leq \alpha$$

LDT approximation

Bilevel optimization

$$\min_{\mathbf{x}\in\mathcal{X}} \quad c(\mathbf{x})$$
$$\text{s.t.} \quad P_k(\mathbf{x},\theta^\star) \leq \alpha$$
$$\theta^\star = \underset{\theta: F(\mathbf{x},\theta)=z}{\text{argmin}} \; I(\theta)$$

- **Sampling free**
- **Works for Gaussian mixtures**

KKT Reformulation

- Accurate if: Regularity conditions hold or $z \to \infty$
- ``Good” if F is close to concave and z is “large”

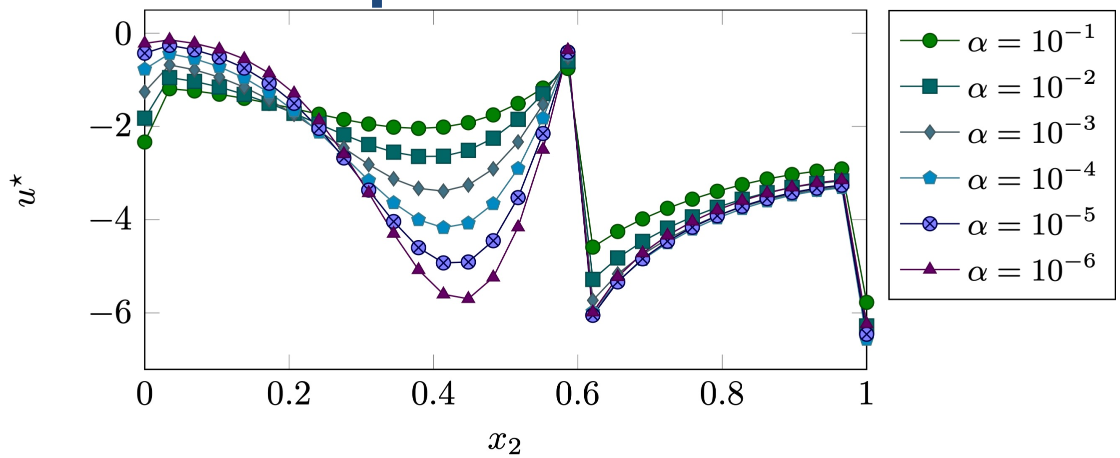
Optimal boundary control for steady state advection diffusion problem

$$
\left.
\begin{aligned}
-\nabla \cdot (\kappa(x,\xi)\nabla y(x)) + w(x)\cdot\nabla y(x) &= f(x,\xi), && x \in \Omega,\\
(\kappa(x,\xi)\nabla y(x))\cdot n(x) &= \frac{1}{\epsilon_0}\left(u(x) - y(x)\right), && \text{on } \Gamma_c,\\
(\kappa(x,\xi)\nabla y(x))\cdot n(x) &= 0, && \text{on } \Gamma_n,
\end{aligned}
\right\} \text{ PDE}
$$

$$
F(u,\xi) := \frac{1}{|\Omega_0|}\int_{\Omega_0} y(x;u,\xi)\,dx,
$$

$$
\begin{aligned}
\underset{u}{\text{minimize}} \quad & \frac{1}{2}\int_{\Gamma_c} u^2(x)\,dx,\\
\text{subject to} \quad & \mathbb{P}(F(u,\xi) \geq z) \leq \alpha, \longleftarrow \text{Chance Constraints}
\end{aligned}
$$

Optimal boundary control for steady state advection diffusion problem



The optimal boundary conditions for different α using SORM

α	$\frac{1}{2}\int_{\Omega_0}(u^\star)^2 dx$	$\mathbb{P}(F(u^\star,\xi) \geq z)$
10^{-1}	3.52	8.90e-02
10^{-2}	4.43	9.50e-03
10^{-3}	5.22	9.15e-04
10^{-4}	5.88	9.54e-05
10^{-5}	6.41	9.60e-06

The objective values and feasibilities of the boundary control problem for different α using SORM

Concluding Remarks and References

- Rice’s formula + Bayesian Inference for risk quantification
  - Can be extended to Gaussian mixtures and by controlling the variance of mixture components, this can be used in optimization under rare chance constraints
- LDT + Bilevel optimization holds a lot of promise
  - Works well for Concave or nearly concave limit state functions

arXiv:2001.11904 [pdf, other] math.NA math.DS math.PR
Efficient computation of extreme excursion probabilities for dynamical systems
Authors: Vishwas Rao, Mihai Anitescu

arXiv:2006.03466 [pdf, other] physics.comp-ph
A Machine-Learning-Based Importance Sampling Method to Compute Rare Event Probabilities
Authors: Vishwas Rao, Romit Maulik, Emil Constantinescu, Mihai Anitescu

arXiv:2011.06052 [pdf, other] math.OC math.PR stat.CO
Optimization under rare chance constraints
Authors: Shanyin Tong, Anirudh Subramanyam, Vishwas Rao