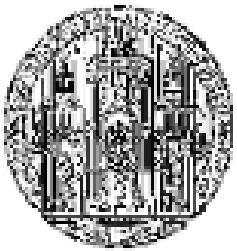
# Stein Variational Newton & other Sampling-Based Inference Methods

**Robert Scheichl**

Interdisciplinary Center for Scientific Computing
& Institute of Applied Mathematics
Universität Heidelberg



Collaborators:
**G. Detommaso** (Bath); T. Cui (Monash); A. Spantini & Y. Marzouk (MIT);
K. Anaya-Izquierdo & **S. Dolgov** (Bath); C. Fox (Otago)

**RICAM Special Semester on Optimization**

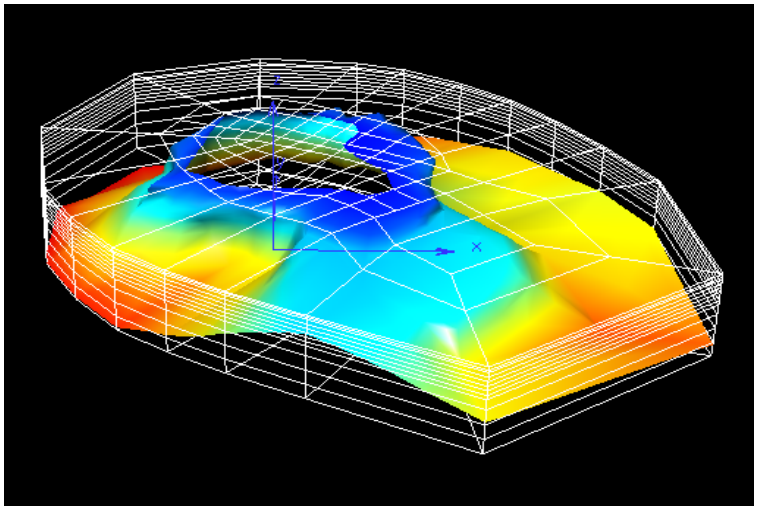
Workshop 3 – Optimization and Inversion under Uncertainty

Linz, November 11, 2019

# Inverse Problems

Data

Parameter



$$y = F(x) + e$$

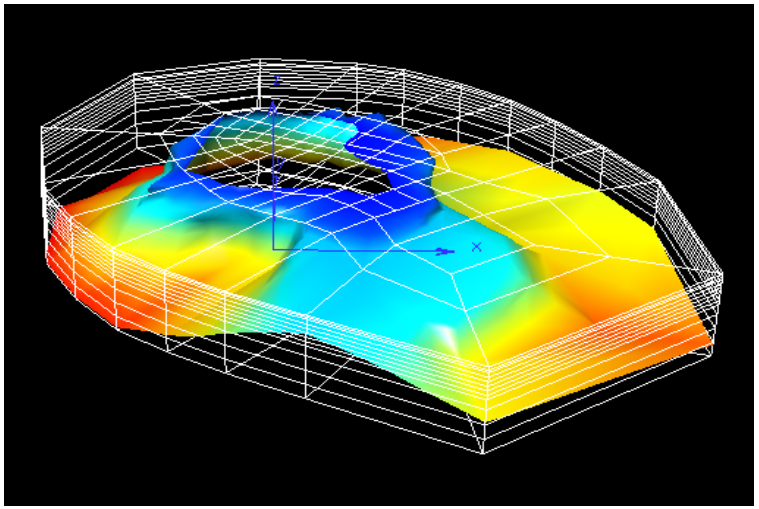
forward model (PDE)

observation/model errors

# Inverse Problems

Data

Parameter



$$y = F(x) + e$$

forward model (PDE)

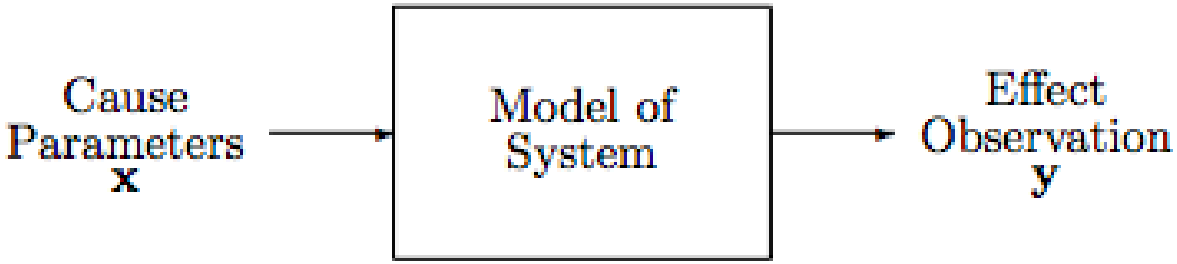
observation/model errors

$y \in \mathbb{R}^{N_y}$ Data $y$ are limited in number, noisy, and indirect.

$x \in X$ Parameter $x$ often a function (discretisation needed).

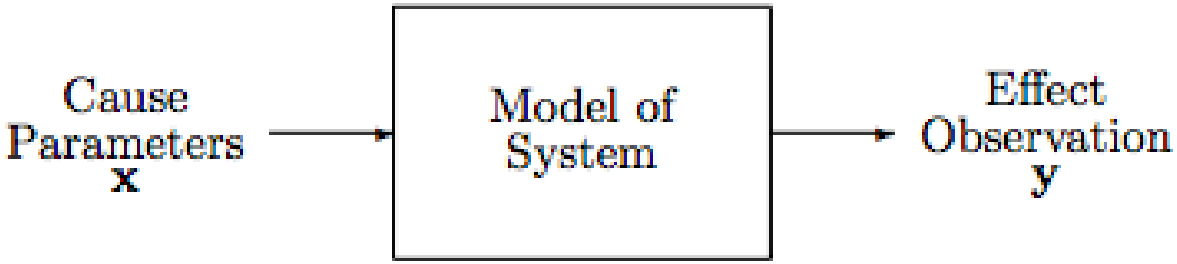
$F : X \to \mathbb{R}^{N_y}$ Continuous, bounded, and sufficiently smooth.

# Bayesian interpretation




The (physical) model gives $\pi(y|x)$, the *conditional probability of observing $y$ given $x$*. However, to predict, control, optimise or quantify uncertainty, the interest is often really in $\pi(x|y)$, the *conditional probability of possible causes $x$ given the observed data $y$* – the **inverse problem**:

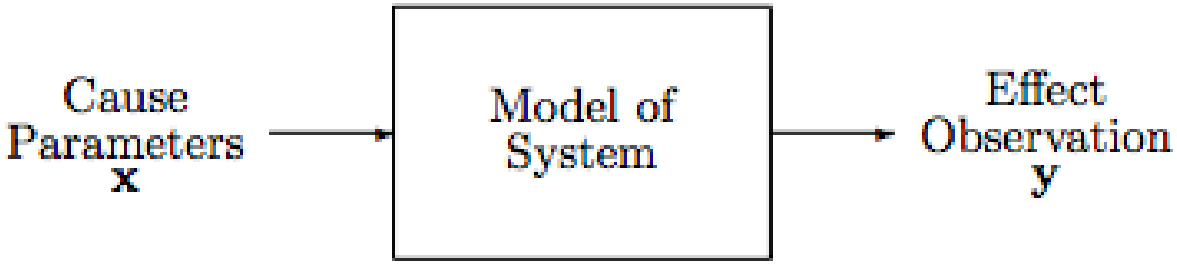
# Bayesian interpretation




The (physical) model gives $\pi(y|x)$, the *conditional probability of observing $y$ given $x$*. However, to predict, control, optimise or quantify uncertainty, the interest is often really in $\pi(x|y)$, the *conditional probability of possible causes $x$ given the observed data $y$* – the **inverse problem**:

$$\pi_{\text{pos}}(x) := \underbrace{\pi(x|y) \propto \pi(y|x)\,\pi_{\text{pr}}(x)}_{\textbf{Bayes' rule}}$$

# Bayesian interpretation




The (physical) model gives $\pi(y|x)$, the *conditional probability of observing $y$ given $x$*. However, to predict, control, optimise or quantify uncertainty, the interest is often really in $\pi(x|y)$, the *conditional probability of possible causes $x$ given the observed data $y$* – the **inverse problem**:

$$\pi_{\text{pos}}(x) := \underbrace{\pi(x|y) \propto \pi(y|x)\,\pi_{\text{pr}}(x)}_{\textbf{Bayes' rule}}$$

Extract information from $\pi_{\text{pos}}$ *(means, covariances, event probabilities, predictions)* by evaluating **posterior expectations:**

$$\mathbb{E}_{\pi_{\text{pos}}}[h(x)] = \int h(x)\pi_{\text{pos}}(x)dx$$

# Bayes' Rule and Classical Inversion

Classically [Hadamard, 1923]: Inverse map "$F^{-1}$" ($y \to x$) is typically ill-posed, i.e. lack of (a) **existence**, (b) **uniqueness** or (c) **boundedness**

# Bayes' Rule and Classical Inversion

Classically [Hadamard, 1923]: Inverse map "$F^{-1}$" ($y \to x$) is typically ill-posed, i.e. lack of (a) **existence**, (b) **uniqueness** or (c) **boundedness**

- least squares solution $\hat{x}$ is *maximum likelihood estimate*
- prior distribution $\pi_{\text{pr}}$ "acts" as regulariser – **well-posedness !**
- solution of regularised least squares problem is *maximum a posteriori (MAP) estimator*

# Bayes' Rule and Classical Inversion

Classically [Hadamard, 1923]: Inverse map "$F^{-1}$" ($y \to x$) is typically ill-posed, i.e. lack of (a) **existence**, (b) **uniqueness** or (c) **boundedness**

- least squares solution $\hat{x}$ is *maximum likelihood estimate*
- prior distribution $\pi_{\text{pr}}$ "acts" as regulariser – **well-posedness !**
- solution of regularised least squares problem is *maximum a posteriori (MAP) estimator*

However, in the Bayesian setting, the **full posterior** $\pi_{\text{pos}}$ **contains more information** than the MAP estimator alone, e.g. the posterior covariance matrix reveals components of $x$ that are (relatively) more or less certain.

# Bayes' Rule and Classical Inversion

Classically [Hadamard, 1923]: Inverse map "$F^{-1}$" ($y \to x$) is typically ill-posed, i.e. lack of (a) **existence**, (b) **uniqueness** or (c) **boundedness**

- least squares solution $\hat{x}$ is *maximum likelihood estimate*
- prior distribution $\pi_{\text{pr}}$ "acts" as regulariser – **well-posedness !**
- solution of regularised least squares problem is *maximum a posteriori (MAP) estimator*

However, in the Bayesian setting, the **full posterior $\pi_{\text{pos}}$ contains more information** than the MAP estimator alone, e.g. the posterior covariance matrix reveals components of $x$ that are (relatively) more or less certain.

- Possible to sample/explore via Metropolis-Hastings MCMC (in theory)

# Variational Bayes (as opposed to Metropolis-Hastings MCMC)

**Aim** to characterise the posterior distribution (density $\pi_{\text{pos}}$) **analytically** (at least approximately) for more efficient inference.

# Variational Bayes (as opposed to Metropolis-Hastings MCMC)

**Aim** to characterise the posterior distribution (density $\pi_{\text{pos}}$) **analytically** (at least approximately) for more efficient inference.

This is a **challenging task** since:

- $x \in \mathbb{R}^d$ is typically **high-dimensional** (e.g., discretised function)
- $\pi_{\text{pos}}$ is in general **non-Gaussian**
  (even if $\pi_{\text{pr}}$ and observation noise are Gaussian)
- evaluations of likelihood may be **expensive** (e.g., solution of a PDE)

# Variational Bayes (as opposed to Metropolis-Hastings MCMC)

**Aim** to characterise the posterior distribution (density $\pi_{\text{pos}}$) **analytically** (at least approximately) for more efficient inference.

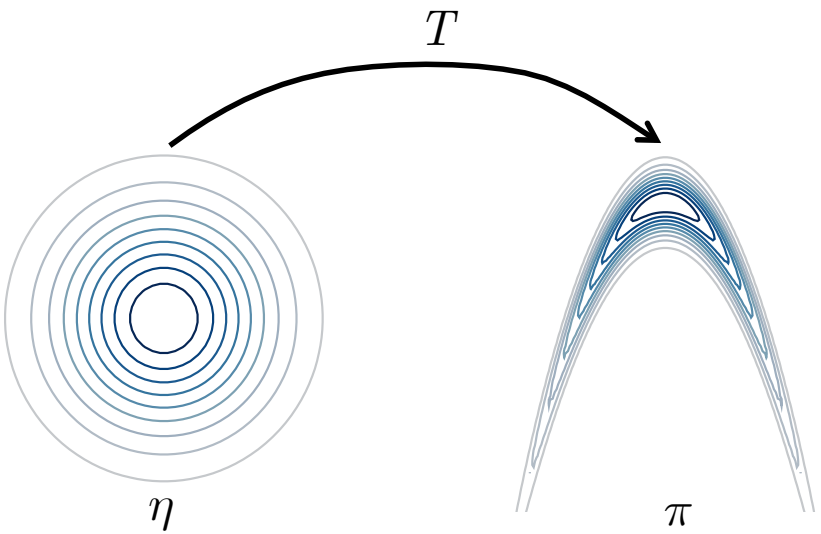
This is a **challenging task** since:

- $x \in \mathbb{R}^d$ is typically **high-dimensional** (e.g., discretised function)
- $\pi_{\text{pos}}$ is in general **non-Gaussian**
  (even if $\pi_{\text{pr}}$ and observation noise are Gaussian)
- evaluations of likelihood may be **expensive** (e.g., solution of a PDE)

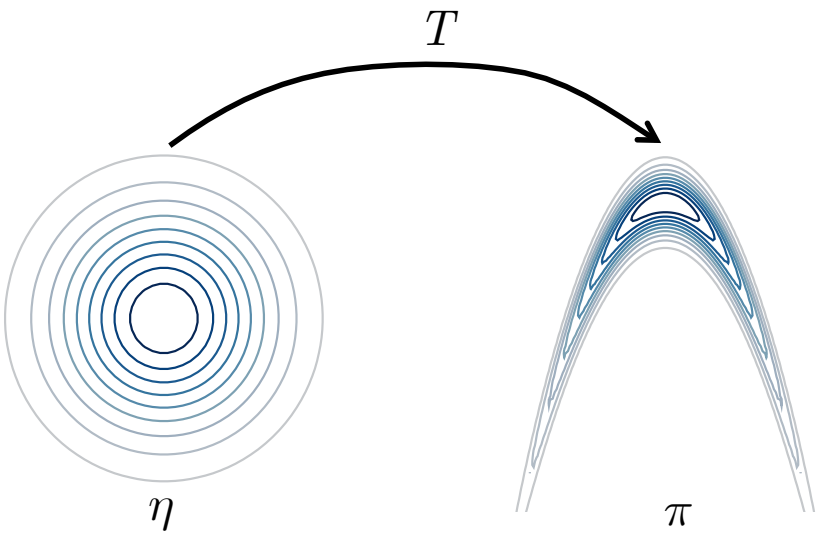
## Key Tools

Transport Maps, **Optimisation**, Principle Component Analysis, Model Order Reduction, Hierarchies, Sparsity, Low Rank Approximation

# Deterministic Couplings of Probability Measures



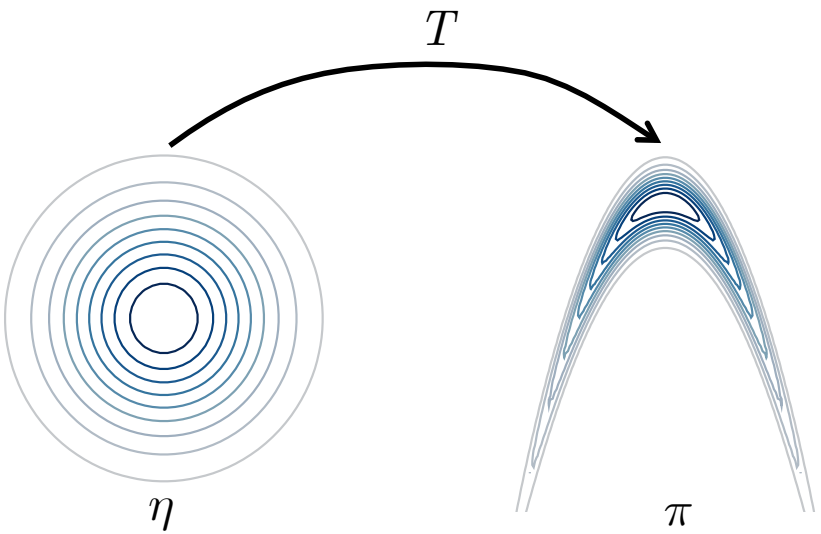
# Deterministic Couplings of Probability Measures




## Core idea [Moselhy, Marzouk, 2012]

- Choose a *reference distribution* $\eta$ (e.g., standard Gaussian)
- Seek transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_\sharp \eta = \pi$
  (or equivalently its inverse $S = T^{-1}$)

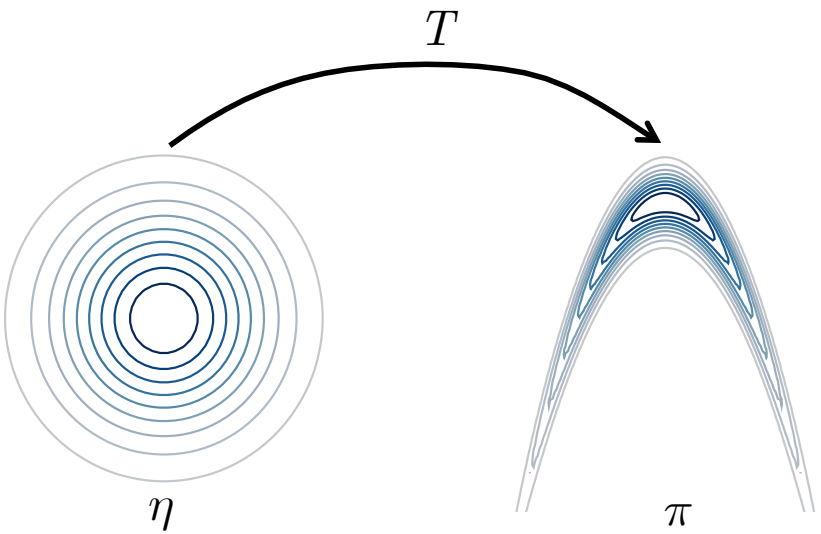
# Deterministic Couplings of Probability Measures




## Core idea [Moselhy, Marzouk, 2012]

- Choose a *reference distribution* $\eta$ (e.g., standard Gaussian)
- Seek transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_\sharp \eta = \pi$
  (or equivalently its inverse $S = T^{-1}$)
- In principle, enables *exact* (independent, unweighted) sampling!

# Deterministic Couplings of Probability Measures




## Core idea [Moselhy, Marzouk, 2012]

- Choose a *reference distribution* $\eta$ (e.g., standard Gaussian)
- Seek transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_\sharp \eta = \pi$
  (or equivalently its inverse $S = T^{-1}$)
- In principle, enables *exact* (independent, unweighted) sampling!
- Satisfying these conditions only **approximately** can still be useful!

# Variational Inference

- **Goal:** Sampling from target density $\pi(x)$

# Variational Inference

- **Goal:** Sampling from target density $\pi(x)$
- Given a reference density $p$, find an invertible map $\hat{T}$ such that

$$\hat{T} := \operatorname*{argmin}_T \; \mathscr{D}_{\mathrm{KL}}(T_\sharp\, p \,\|\, \pi) = \operatorname*{argmin}_T \; \mathscr{D}_{\mathrm{KL}}(p \,\|\, T_\sharp^{-1}\pi)$$

where

$$T_\sharp(x) := p\left(T^{-1}(x)\right) \left|\det\left(\nabla_x T^{-1}(x)\right)\right| \quad \ldots \quad \textbf{push-forward of } p$$

$$\mathscr{D}_{\mathrm{KL}}(p \,\|\, q) := \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\,\mathrm{d}x \quad \ldots \quad \textbf{Kullback-Leibler divergence}$$

# Variational Inference

- **Goal:** Sampling from target density $\pi(x)$
- Given a reference density $p$, find an invertible map $\hat{T}$ such that

$$\hat{T} := \operatorname*{argmin}_T \; \mathscr{D}_{\mathrm{KL}}(T_\sharp\, p \,\|\, \pi) = \operatorname*{argmin}_T \; \mathscr{D}_{\mathrm{KL}}(p \,\|\, T_\sharp^{-1}\pi)$$

where

$$T_\sharp(x) := p\left(T^{-1}(x)\right) \left|\det\left(\nabla_x T^{-1}(x)\right)\right| \quad \ldots \quad \textbf{push-forward of } p$$

$$\mathscr{D}_{\mathrm{KL}}(p \,\|\, q) := \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\,\mathrm{d}x \quad \ldots \quad \textbf{Kullback-Leibler divergence}$$

- Advantage of using $\mathscr{D}_{\mathrm{KL}}$: do **not** need normalising constant for $\pi$

# Variational Inference

- **Goal:** Sampling from target density $\pi(x)$
- Given a reference density $p$, find an invertible map $\hat{T}$ such that

$$\hat{T} := \operatorname*{argmin}_T \; \mathscr{D}_{\mathrm{KL}}(T_\sharp\, p \,\|\, \pi) = \operatorname*{argmin}_T \; \mathscr{D}_{\mathrm{KL}}(p \,\|\, T_\sharp^{-1}\pi)$$

where

$$T_\sharp(x) := p\left(T^{-1}(x)\right) \left|\det\left(\nabla_x T^{-1}(x)\right)\right| \quad \ldots \quad \textbf{push-forward of } p$$

$$\mathscr{D}_{\mathrm{KL}}(p \,\|\, q) := \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\,\mathrm{d}x \quad \ldots \quad \textbf{Kullback-Leibler divergence}$$

- Advantage of using $\mathscr{D}_{\mathrm{KL}}$: do **not** need normalising constant for $\pi$
- Minimise over some suitable class $\mathscr{T}$ of maps $T$
  (where ideally Jacobian determinant $\left|\det\left(\nabla_x T^{-1}(x)\right)\right|$ is easy to evaluate)

# Variational Inference

- **Goal:** Sampling from target density $\pi(x)$
- Given a reference density $p$, find an invertible map $\hat{T}$ such that

$$\hat{T} := \operatorname*{argmin}_T \mathscr{D}_{\mathrm{KL}}(T_\sharp\, p \,\|\, \pi) = \operatorname*{argmin}_T \mathscr{D}_{\mathrm{KL}}(p \,\|\, T_\sharp^{-1}\pi)$$

where

$$T_\sharp(x) := p\left(T^{-1}(x)\right) \left|\det\left(\nabla_x T^{-1}(x)\right)\right| \quad \ldots \quad \textbf{push-forward of } p$$

$$\mathscr{D}_{\mathrm{KL}}(p \,\|\, q) := \int \log\left(\frac{p(x)}{q(x)}\right) p(x)\,\mathrm{d}x \quad \ldots \quad \textbf{Kullback-Leibler divergence}$$

- Advantage of using $\mathscr{D}_{\mathrm{KL}}$: do **not** need normalising constant for $\pi$
- Minimise over some suitable class $\mathscr{T}$ of maps $T$
  (where ideally Jacobian determinant $\left|\det\left(\nabla_x T^{-1}(x)\right)\right|$ is easy to evaluate)
- **To improve:** enrich class $\mathscr{T}$ or use samples of $T_\sharp^{-1}\pi$ as **proposals for MCMC** or in **importance sampling** (see below)

# Many Choices ("Architectures") for $\mathscr{T}$ possible

**Examples:** (list not comprehensive!!)

1. Optimal Transport & Knothe-Rosenblatt Rearrangement
   [Moselhy, Marzouk, 2012], [Marzouk, Moselhy, Parno, Spantini, 2016]
2. Normalizing Flows [Rezende, Mohamed, 2015]
   (and related methods in the ML literature)

# Many Choices ("Architectures") for $\mathscr{T}$ possible

**Examples:** (list not comprehensive!!)

1. Optimal Transport & Knothe-Rosenblatt Rearrangement
   [Moselhy, Marzouk, 2012], [Marzouk, Moselhy, Parno, Spantini, 2016]
2. Normalizing Flows [Rezende, Mohamed, 2015]
   (and related methods in the ML literature)
3. Kernel-based variational inference: **Stein Variational Methods**
   [Liu, Wang, 2016], **[Detommaso, Cui, Spantini, Marzouk, RS, 2018]**,
   [Chen, Wu, Chen, O'Leary-Roseberry, Ghattas, arXiv 2019]

# Many Choices ("Architectures") for $\mathscr{T}$ possible

**Examples:** (list not comprehensive!!)

1. Optimal Transport & Knothe-Rosenblatt Rearrangement
   [Moselhy, Marzouk, 2012], [Marzouk, Moselhy, Parno, Spantini, 2016]
2. Normalizing Flows [Rezende, Mohamed, 2015]
   (and related methods in the ML literature)
3. Kernel-based variational inference: **Stein Variational Methods**
   [Liu, Wang, 2016], **[Detommaso, Cui, Spantini, Marzouk, RS, 2018]**,
   [Chen, Wu, Chen, O'Leary-Roseberry, Ghattas, arXiv 2019]
4. Layers of low-rank maps [Bigoni, Zahm, Spantini, Marzouk, arXiv 2019]
5. Layers of hierarchical invertible neural networks (HINT) not today!
   [Detommaso, Kruse, Ardizzone, Rother, Köthe, RS, arXiv 2019]

# Many Choices ("Architectures") for $\mathscr{T}$ possible

**Examples:** (list not comprehensive!!)

1. Optimal Transport & Knothe-Rosenblatt Rearrangement
   [Moselhy, Marzouk, 2012], [Marzouk, Moselhy, Parno, Spantini, 2016]
2. Normalizing Flows [Rezende, Mohamed, 2015]
   (and related methods in the ML literature)
3. Kernel-based variational inference: **Stein Variational Methods**
   [Liu, Wang, 2016], **[Detommaso, Cui, Spantini, Marzouk, RS, 2018]**,
   [Chen, Wu, Chen, O'Leary-Roseberry, Ghattas, arXiv 2019]
4. Layers of low-rank maps [Bigoni, Zahm, Spantini, Marzouk, arXiv 2019]
5. Layers of hierarchical invertible neural networks (HINT) not today!
   [Detommaso, Kruse, Ardizzone, Rother, Köthe, RS, arXiv 2019]
6. **Low-rank tensor approx.** & Knothe-Rosenblatt rearrangement
   **[Dolgov, Anaya-Izquierdo, Fox, RS, 2019]**

# A Stein Variational Newton (SVN) Method

[Detommaso, Cui, Spantini, Marzouk, RS, 2018]

# Stein variational gradient descent [Liu, Wang, 2016]

- Construct $\hat{T}$ as a composition of simple maps $\hat{T}_\ell$:

$$\hat{T} := \hat{T}_1 \circ \cdots \circ \hat{T}_\ell \circ \cdots, \quad \text{where } \hat{T}_\ell := I + \hat{Q}_\ell$$

- Stein Variational Gradient Descent (SVGD) picks **steepest descent direction** in a Reproducing Kernel Hilbert Space (RKHS) $\mathscr{H}^d$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

# Stein variational gradient descent [Liu, Wang, 2016]

- Construct $\hat{T}$ as a composition of simple maps $\hat{T}_\ell$:

$$\hat{T} := \hat{T}_1 \circ \cdots \circ \hat{T}_\ell \circ \cdots, \quad \text{where } \hat{T}_\ell := I + \hat{Q}_\ell$$

- Stein Variational Gradient Descent (SVGD) picks **steepest descent direction** in a Reproducing Kernel Hilbert Space (RKHS) $\mathscr{H}^d$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

- Given a reference measure $p_\ell$ in the $\ell$th step, define

$$J_{p_\ell} : \mathscr{H}^d \to \mathbb{R} \quad s.t. \quad J_{p_\ell}[Q] := \mathscr{D}_{\mathrm{KL}}\Big( \underbrace{(I+Q)_\sharp}_{T_\sharp} p_\ell \,\|\, \pi \Big)$$

- Then $\hat{Q}_\ell$ is chosen to satisfy $J_{p_\ell}[\hat{Q}_\ell] < J_{p_\ell}[\mathbf{0}]$

# Stein variational gradient descent [Liu, Wang, 2016]

- Construct $\hat{T}$ as a composition of simple maps $\hat{T}_\ell$:

$$\hat{T} := \hat{T}_1 \circ \cdots \circ \hat{T}_\ell \circ \cdots, \quad \text{where } \hat{T}_\ell := I + \hat{Q}_\ell$$

- Stein Variational Gradient Descent (SVGD) picks **steepest descent direction** in a Reproducing Kernel Hilbert Space (RKHS) $\mathscr{H}^d$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$
- Given a reference measure $p_\ell$ in the $\ell$th step, define

$$J_{p_\ell} : \mathscr{H}^d \to \mathbb{R} \quad s.t. \quad J_{p_\ell}[Q] := \mathscr{D}_{\mathrm{KL}}\Big( \underbrace{(I+Q)_\sharp}_{T_\sharp} p_\ell \,\|\, \pi \Big)$$

- Then $\hat{Q}_\ell$ is chosen to satisfy $J_{p_\ell}[\hat{Q}_\ell] < J_{p_\ell}[\mathbf{0}]$
- SVGD uses (functional) gradient descent in $\mathscr{H}^d$ and picks

$$\hat{Q}_\ell(z) := -\nabla J_{p_\ell}[\mathbf{0}] = \mathbb{E}_{x\sim p_\ell}[\nabla_x \log \pi(x) k(x,z) + \nabla_x k(x,z)]$$

# Stein variational gradient descent [Liu, Wang, 2016]

- Finally one defines $p_{\ell+1} := (\hat{T}_\ell)_\sharp p_\ell = (I + \hat{Q}_\ell)_\sharp p_\ell$
- In practice, $p_\ell$ taken as the empirical density of $N$ particles $\left(x_j^{(\ell)}\right)_{j=1}^N$ (as in filtering or sequential Monte Carlo methods) such that

$$\hat{Q}_\ell(z) := \frac{1}{N} \sum_{j=1}^{N} \left[ \nabla_x \log \pi(x_i^{(\ell)}) k(x_i^{(\ell)}, z) + \nabla_x k(x_i^{(\ell)}, z) \right]$$

# Stein variational gradient descent [Liu, Wang, 2016]

- Finally one defines $p_{\ell+1} := (\hat{T}_\ell)_\sharp p_\ell = (I + \hat{Q}_\ell)_\sharp p_\ell$
- In practice, $p_\ell$ taken as the empirical density of $N$ particles $\left(x_j^{(\ell)}\right)_{j=1}^N$ (as in filtering or sequential Monte Carlo methods) such that

$$\hat{Q}_\ell(z) := \frac{1}{N} \sum_{j=1}^{N} \left[ \nabla_x \log \pi(x_i^{(\ell)}) k(x_i^{(\ell)}, z) + \nabla_x k(x_i^{(\ell)}, z) \right]$$

---

**Algorithm 2:** Stein variational gradient descent (SVGD)

---

**Input** : Particles $(x_j^{(\ell)})_{j=1}^N$, step size $\varepsilon$

**Output:** Particles $(x_j^{(\ell+1)})_{j=1}^N$

**for** $j = 1, 2, \ldots, N$ **do**

$$x_j^{(\ell+1)} \leftarrow T_\ell(x_j^{(\ell)}) := x_j^{(\ell)} + \varepsilon \hat{Q}_\ell(x_j^{(\ell)})$$

**end for**

---

# 1st Improvement: Using second-order information

- Particles are evolved sequentially from initial distribution $p_0 = p$ to final distribution $p_L \approx \pi$.

# 1st Improvement: Using second-order information

- Particles are evolved sequentially from initial distribution $p_0 = p$ to final distribution $p_L \approx \pi$.
- SVGD is a deterministic **first-order** optimisation algorithm.
  We can accelerate it by introducing **second-order** information!

# 1st Improvement: Using second-order information

- Particles are evolved sequentially from initial distribution $p_0 = p$ to final distribution $p_L \approx \pi$.
- SVGD is a deterministic **first-order** optimisation algorithm.
  We can accelerate it by introducing **second-order** information!
- Representing $\hat{Q}_\ell(x) = \sum_{j=1}^N c_j k_j(x)$, where $k_j(x) := k(x, x_j^{(\ell)})$, the (exact) **Newton step** can be computed by solving the linear system

$$Hc = g$$

where

$$H_{mn} := \mathbb{E}_{p_\ell}[-\nabla^2 \log \pi \, k_m \, k_n + \nabla k_m \nabla k_n^\top], \qquad m, n = 1, \ldots, N,$$

$$g_m := \mathbb{E}_{p_\ell}[\nabla \log \pi \, k_m + \nabla k_m], \qquad m = 1, \ldots, N.$$

# 1st Improvement: Using second-order information

- Particles are evolved sequentially from initial distribution $p_0 = p$ to final distribution $p_L \approx \pi$.
- SVGD is a deterministic **first-order** optimisation algorithm.
  We can accelerate it by introducing **second-order** information!
- Representing $\hat{Q}_\ell(x) = \sum_{j=1}^N c_j k_j(x)$, where $k_j(x) := k(x, x_j^{(\ell)})$, the (exact) **Newton step** can be computed by solving the linear system

$$Hc = g$$

  where

$$H_{mn} := \mathbb{E}_{p_\ell}[-\nabla^2 \log \pi \, k_m \, k_n + \nabla k_m \nabla k_n^\top], \qquad m, n = 1, \ldots, N,$$

$$g_m := \mathbb{E}_{p_\ell}[\nabla \log \pi \, k_m + \nabla k_m], \qquad m = 1, \ldots, N.$$

- In practice, use block-diagonal approximation (inexact Newton)

$$\mathbb{H}_{mm} c_m = g_m \,, \quad \text{for } m = 1, \ldots, N, \quad \text{and set} \quad \hat{Q}_\ell(x_m) = c_m \,.$$

# A Stein variational Newton method

**Algorithm 3:** Stein variational (inexact) Newton

**Input** : Particles $(x_j^{(\ell)})_{j=1}^N$, step size $\varepsilon$

**Output:** Particles $(x_j^{(\ell+1)})_{j=1}^N$

1: **for** $m = 1, 2, \ldots, N$ **do**

2: Evaluate gradient $g_m$ and Hessian $\mathbb{H}_{mm}$, replacing $\nabla^2 \log \pi$ with Gauss-Newton approximation (only needs gradient info and is SPD)

3: Solve linear system

$$\mathbb{H}_{mm} c_m = g_m \quad \text{and set} \quad \hat{Q}_\ell(x_m^{(\ell)}) := c_m$$

4: Update particle $m$:

$$x_m^{(\ell+1)} \leftarrow x_m^{(\ell)} + \varepsilon \hat{Q}_\ell(x_m^{(\ell)})$$

5: **end for**

# 2nd Improvement: Kernel based on Hessian information

- [Liu, Wang, 2016] chose simple isotropic Gaussian kernel

$$k(x, z) = \exp(-\gamma \|x - z\|_2^2)$$

- However, kernel should mimic the shape of the target distribution

# 2nd Improvement: Kernel based on Hessian information

- [Liu, Wang, 2016] chose simple isotropic Gaussian kernel

$$k(x, z) = \exp(-\gamma \|x - z\|_2^2)$$

- However, kernel should mimic the shape of the target distribution
- We use a **scaled & averaged Hessian** (available at **no extra cost!**):

$$M \approx \frac{1}{d} \mathbb{E}_{p_\ell}[-\nabla^2 \log \pi]$$

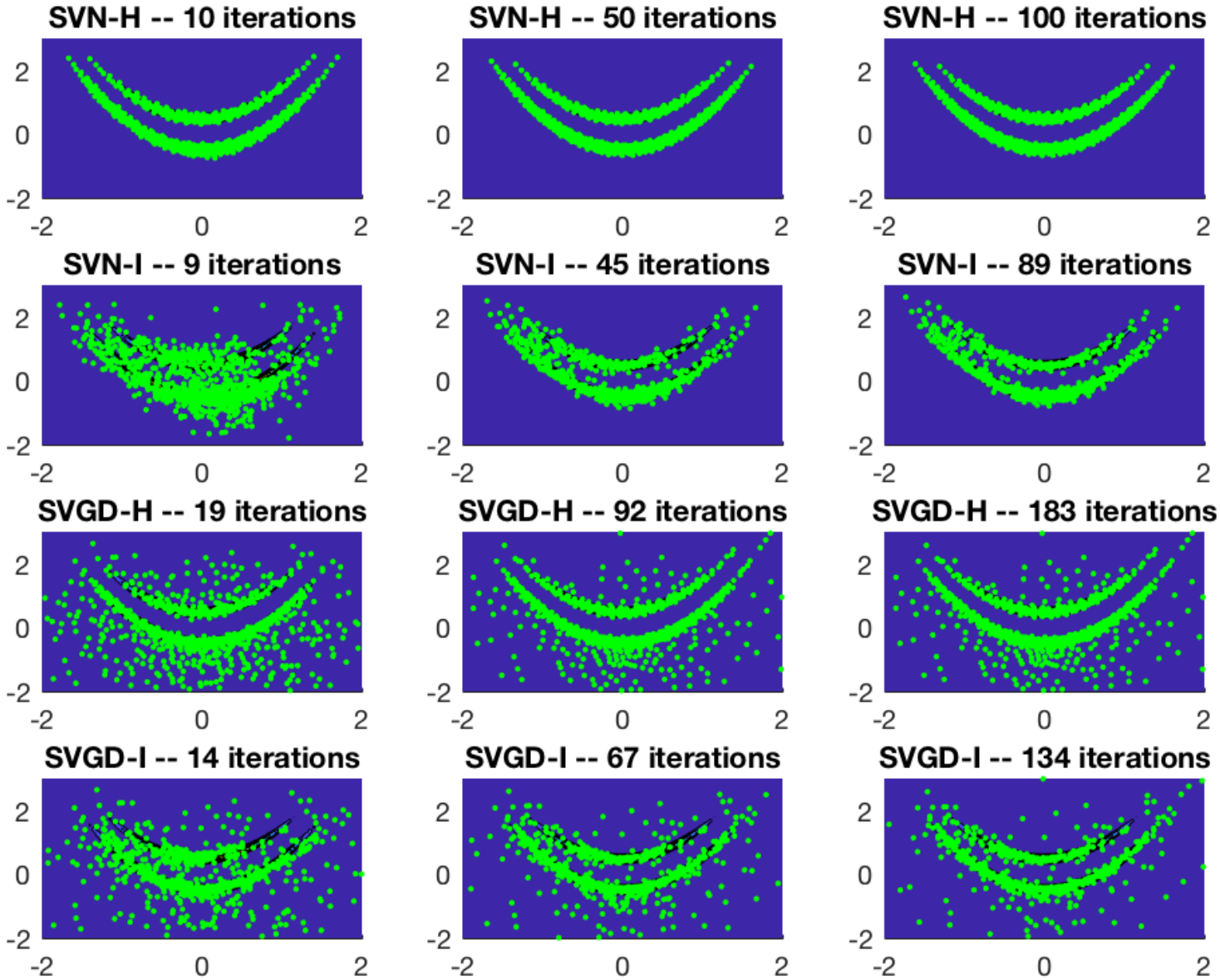
and then construct the (data-informed) kernel

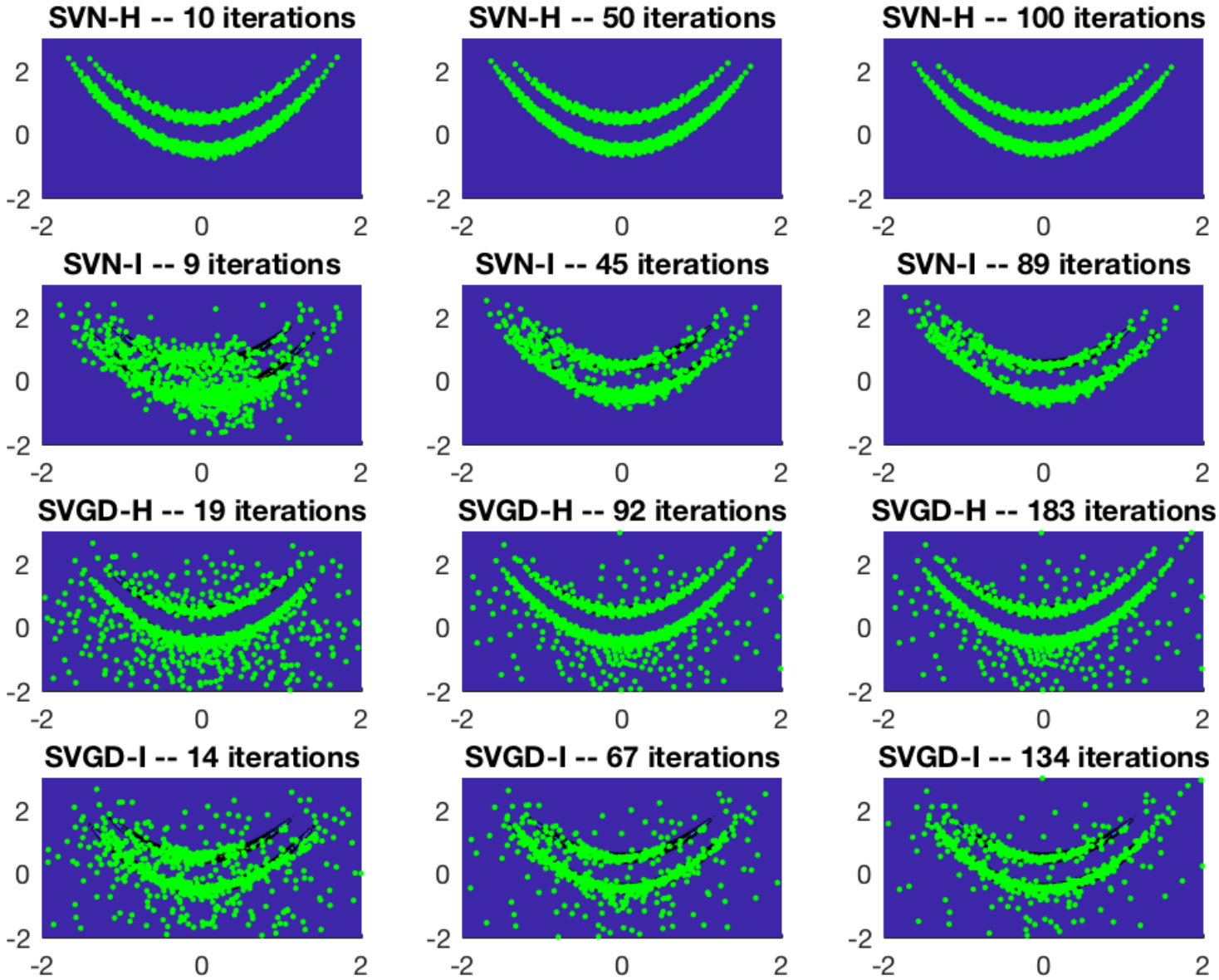
$$k(x, z) = \exp\left(-\frac{1}{2}\|x - z\|_M^2\right)$$

(In practice, use Gauss-Newton Hessian approximation $\mathbb{H}$ and MC average.)

# Test Case 1: two-dimensional "double-banana"

- Reference distribution (prior): $p = N(0, I)$
- Forward model: $\mathscr{F}(x) = \log\left((1 - x_1)^2 + 100(x_2 - x_1^2)^2\right)$
  (Rosenbrock function)
- Observation: $y = \mathscr{F}(x_{\text{true}}) + \xi$, with $x_{\text{true}} \sim N(0, I), \xi \sim N(0, 0.09I)$
- Number of particles: $N = 1000$
- Compare **SVN-H, SVN-I, SVGD-H** and **SVGD-I**
  ("H" stands for scaled Hessian kernel and "I" stands for isotropic kernel)


SVN-H -- 10 iterations
SVN-H -- 50 iterations
SVN-H -- 100 iterations
SVN-I -- 9 iterations
SVN-I -- 45 iterations
SVN-I -- 89 iterations
SVGD-H -- 19 iterations
SVGD-H -- 92 iterations
SVGD-H -- 183 iterations
SVGD-I -- 14 iterations
SVGD-I -- 67 iterations
SVGD-I -- 134 iterations
2
0
-2
-2
0
2


SVN-H -- 10 iterations
SVN-H -- 50 iterations
SVN-H -- 100 iterations
SVN-I -- 9 iterations
SVN-I -- 45 iterations
SVN-I -- 89 iterations
SVGD-H -- 19 iterations
SVGD-H -- 92 iterations
SVGD-H -- 183 iterations
SVGD-I -- 14 iterations
SVGD-I -- 67 iterations
SVGD-I -- 134 iterations
2
0
-2
-2
0
2

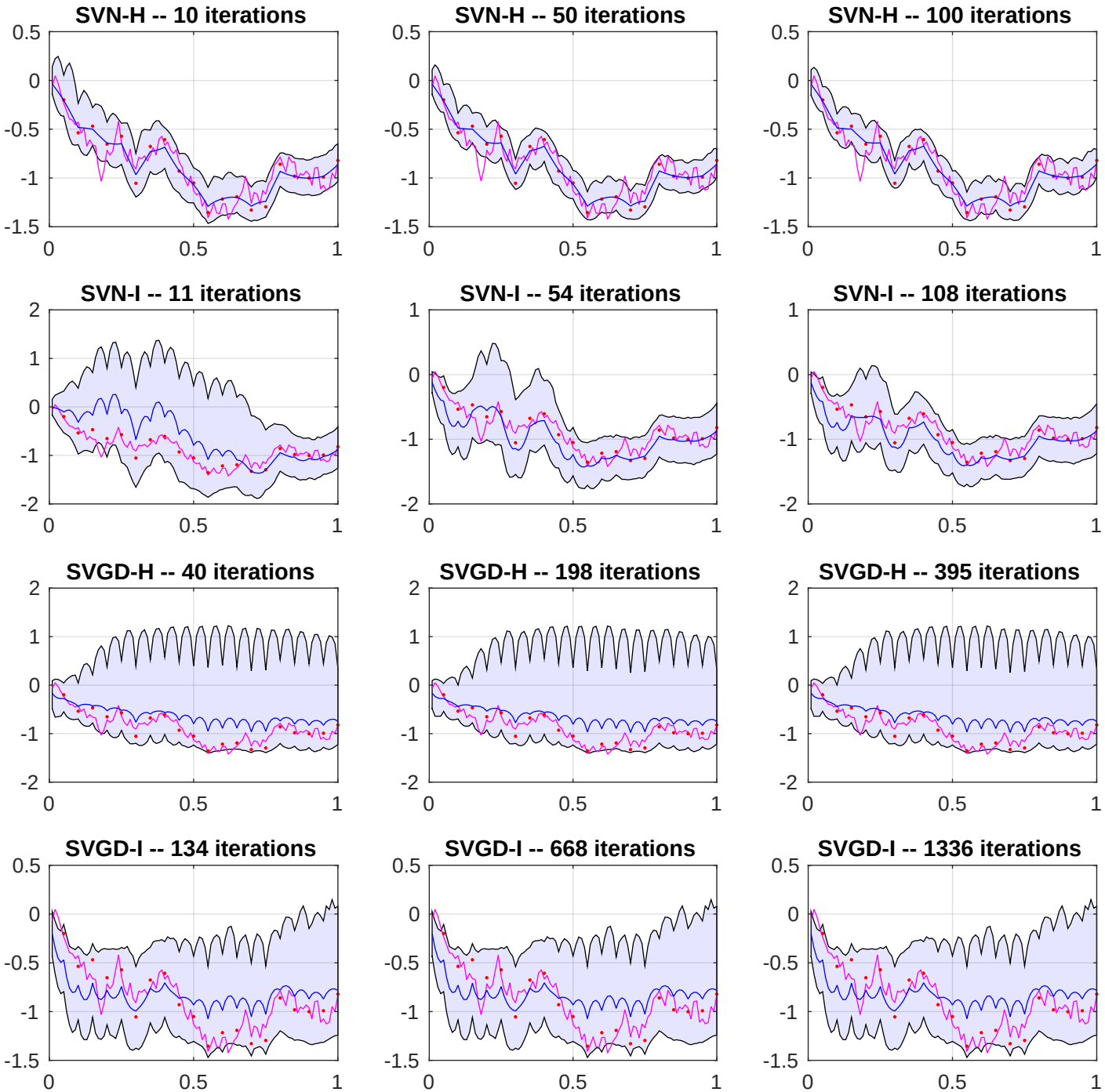
**Video**

# Test Case 2: 100-dimensional conditional diffusion

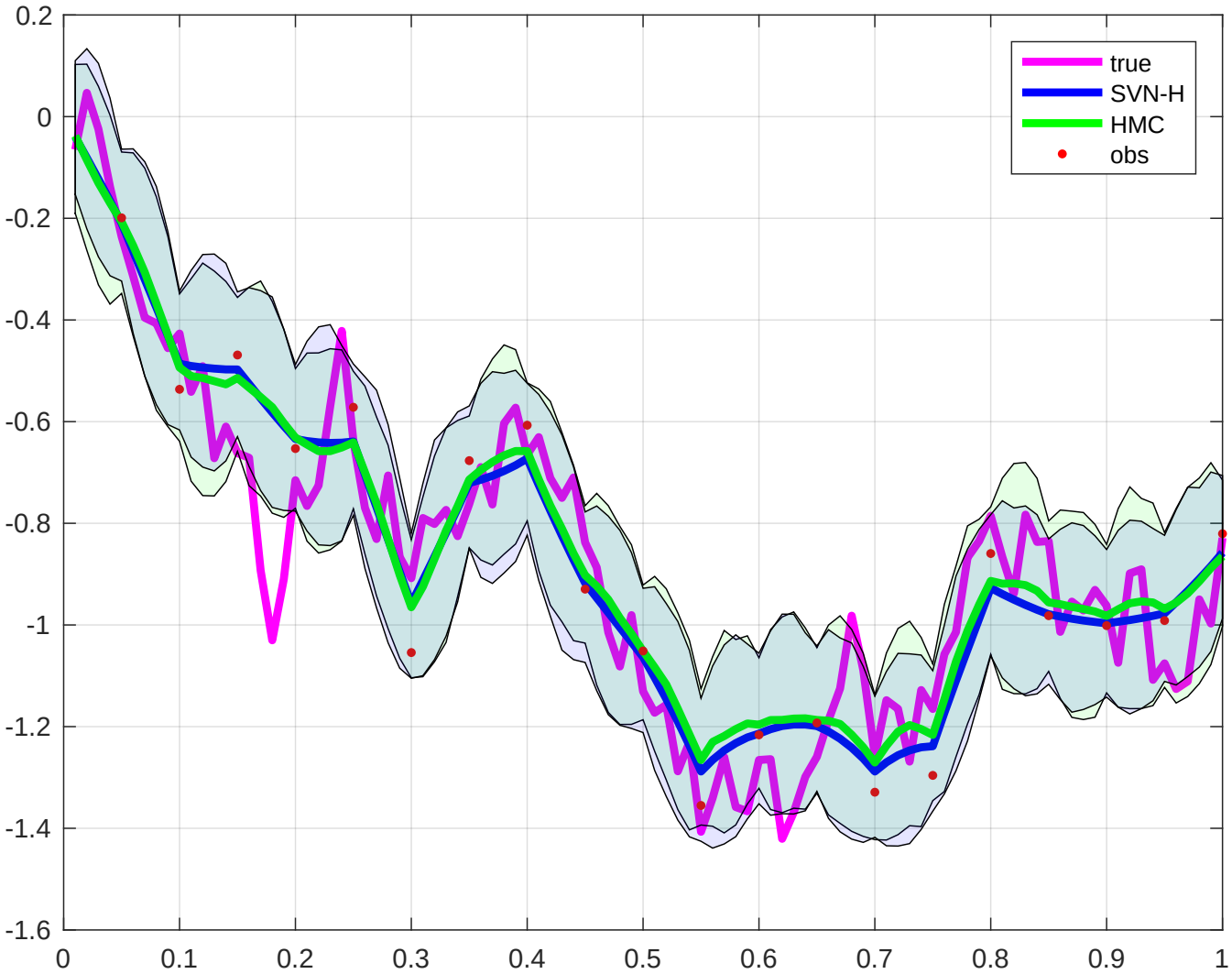
- Reference distribution: $p = N(0, C)$ with $C(t, t') = \min(t, t')$
- Forward model: $\mathscr{F}(u) = [\hat{u}_{t_5}, \hat{u}_{t_{10}}, \ldots, \hat{u}_{t_{100}}]^\top \in \mathbb{R}^{20}$, where $(\hat{u}_{t_i})_{i=1}^{100}$ is the Euler-Maruyama discretisation of

$$du_t = \frac{\beta u (1 - u^2)}{(1 + u^2)} dt + dx_t, \quad u_0 = 0$$

  for $t \in [0, 1]$ with step size $\Delta t = 1/100$
- Observation: $y = \mathscr{F}(x_{\text{true}}) + \xi$ with $x_{\text{true}} \sim N(0, I), \xi \sim N(0, 0.01I)$
- Number of particles: $N = 1000$
- Compare **SVN-H, SVN-I, SVGD-H** and **SVGD-I**
  ("H" stands for scaled Hessian kernel and "I" stands for isotropic kernel)


SVN-H -- 10 iterations
SVN-H -- 50 iterations
SVN-H -- 100 iterations
SVN-I -- 11 iterations
SVN-I -- 54 iterations
SVN-I -- 108 iterations
SVGD-H -- 40 iterations
SVGD-H -- 198 iterations
SVGD-H -- 395 iterations
SVGD-I -- 134 iterations
SVGD-I -- 668 iterations
SVGD-I -- 1336 iterations

# Compare SVN-H with Hamiltonian MCMC (HMC)



# Approximation and Sampling of Multivariate Probability Distributions in the Tensor Train Decomposition

[Dolgov, Anaya-Izquierdo, Fox, RS, 2019]

# Recall: General Variational Inference

- In general, in **Variational Inference** aim to find

$$\operatorname*{argmin}_{T} \mathscr{D}_{\mathsf{KL}}(T_\sharp \eta \,||\, \pi)$$

# Recall: General Variational Inference

- In general, in **Variational Inference** aim to find

$$\operatorname*{argmin}_{T} \mathscr{D}_{\mathsf{KL}}(T_\sharp \eta \,||\, \pi)$$

- Note

$$\mathscr{D}_{\mathsf{KL}}(T_\sharp \eta \,||\, \pi) = -\mathbb{E}_{\boldsymbol{u}\sim\eta}\Big[\log \pi(\boldsymbol{T}(\boldsymbol{u})) + \log|\det \nabla \boldsymbol{T}(\boldsymbol{u})|\Big] + \text{const}$$

- Particularly useful family are Knothe-Rosenblatt rearrangements (see [Marzouk, Moshely, Parno, Spantini, 2016]):

$$T(x) = \begin{bmatrix} T_1(x_1) \\ T_2(x_1, x_2) \\ \vdots \\ T_d(x_1, x_2, \ldots, x_d) \end{bmatrix}$$

Then: $\log|\det \nabla \boldsymbol{T}(\boldsymbol{u})| = \sum_k \log \partial_{x_k} T^k$

# Knothe-Rosenblatt via Conditional Distribution Sampling

In fact, $\exists!$ triangular map satisfying $T_\sharp \eta = \pi$ (for abs. cont. $\eta, \pi$ on $\mathbb{R}^d$)

Can be computed **explicitly** via **Conditional Distribution Sampling**[1]:

[1] Rosenblatt '52; Devroye '86; Hormann, Leydold, Derflinger '04

# Knothe-Rosenblatt via Conditional Distribution Sampling

In fact, $\exists!$ triangular map satisfying $T_\sharp \eta = \pi$ (for abs. cont. $\eta, \pi$ on $\mathbb{R}^d$)

Can be computed **explicitly** via **Conditional Distribution Sampling**[1]:

- Any density factorises into product of conditional densities:

$$\pi(x_1, \ldots, x_d) = \pi_1(x_1)\pi_2(x_2|x_1)\cdots\pi_d(x_d|x_1, \ldots, x_{d-1})$$

- Can sample (up to normalisation with known scaling factor)

$$x_k \sim \pi_k(x_k|x_1, \ldots, x_{k-1}) \sim \int \pi(x_1, \ldots, x_d) dx_{k+1} \cdots dx_d$$

[1] Rosenblatt '52; Devroye '86; Hormann, Leydold, Derflinger '04

# Knothe-Rosenblatt via Conditional Distribution Sampling

In fact, $\exists!$ triangular map satisfying $T_\sharp \eta = \pi$ (for abs. cont. $\eta, \pi$ on $\mathbb{R}^d$)

Can be computed **explicitly** via **Conditional Distribution Sampling**[1]:

- Any density factorises into product of conditional densities:

$$\pi(x_1, \ldots, x_d) = \pi_1(x_1)\pi_2(x_2|x_1)\cdots\pi_d(x_d|x_1, \ldots, x_{d-1})$$

- 1st step: Produce sample $x_1^i$ via *1D CDF-inversion* from

$$\pi_1(x_1) \sim \int \pi(x_1, x_2, \ldots, x_d) dx_2 \cdots dx_d$$

[1] Rosenblatt '52; Devroye '86; Hormann, Leydold, Derflinger '04

# Knothe-Rosenblatt via Conditional Distribution Sampling

In fact, $\exists!$ triangular map satisfying $T_\sharp \eta = \pi$ (for abs. cont. $\eta, \pi$ on $\mathbb{R}^d$)

Can be computed **explicitly** via **Conditional Distribution Sampling**[1]:

- Any density factorises into product of conditional densities:

$$\pi(x_1, \ldots, x_d) = \pi_1(x_1)\pi_2(x_2|x_1)\cdots\pi_d(x_d|x_1, \ldots, x_{d-1})$$

- 1st step: Produce sample $x_1^i$ via *1D CDF-inversion* from

$$\pi_1(x_1) \sim \int \pi(x_1, x_2, \ldots, x_d)\, dx_2 \cdots dx_d$$

- $k$-th step: Given $x_1^i, \ldots, x_{k-1}^i$, sample $x_k^i$ via *1D CDF-inversion* from

$$\pi_k(x_k|x_1^i, \ldots, x_{k-1}^i) \sim \int \pi(x_1^i, \ldots, x_{k-1}^i, x_k, x_{k+1}, \ldots, x_d)\, dx_{k+1} \cdots dx_d$$

[1] Rosenblatt '52; Devroye '86; Hormann, Leydold, Derflinger '04

# Knothe-Rosenblatt via Conditional Distribution Sampling

In fact, $\exists!$ triangular map satisfying $T_\sharp \eta = \pi$ (for abs. cont. $\eta, \pi$ on $\mathbb{R}^d$)

Can be computed **explicitly** via **Conditional Distribution Sampling**[1]:

- Any density factorises into product of conditional densities:

$$\pi(x_1, \ldots, x_d) = \pi_1(x_1)\pi_2(x_2|x_1)\cdots\pi_d(x_d|x_1, \ldots, x_{d-1})$$

- 1st step: Produce sample $x_1^i$ via *1D CDF-inversion* from

$$\pi_1(x_1) \sim \int \pi(x_1, x_2, \ldots, x_d)\, dx_2 \cdots dx_d$$

- $k$-th step: Given $x_1^i, \ldots, x_{k-1}^i$, sample $x_k^i$ via *1D CDF-inversion* from

$$\pi_k(x_k|x_1^i, \ldots, x_{k-1}^i) \sim \int \pi(x_1^i, \ldots, x_{k-1}^i, x_k, x_{k+1}, \ldots, x_d)\, dx_{k+1} \cdots dx_d$$

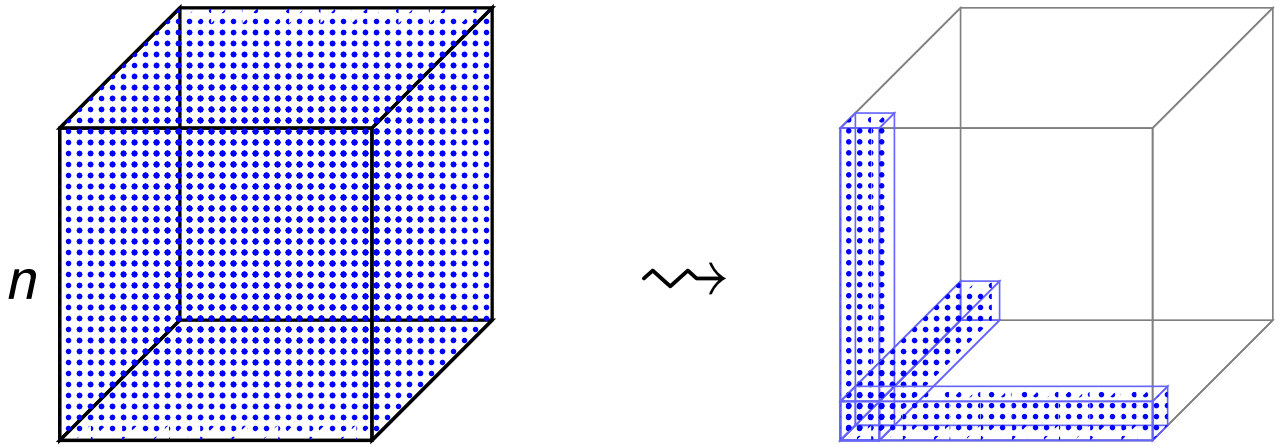
**Problem:** $(d-k)$**-dimensional integration** at $k$-th step!

[1] Rosenblatt '52; Devroye '86; Hormann, Leydold, Derflinger '04

# Low-rank Tensor Approximation of Distributions

## Presented already several times

**Low-rank tensor decomposition** $\Leftrightarrow$ separation of variables:




- Tensor grid with $n$ points per direction (or $n$ polynomial basis fcts.)
- Approximate: $\underbrace{\pi(x_1, \ldots, x_d)}_{\text{tensor}} \approx \underbrace{\sum_{|\alpha| \leq r} \pi_\alpha^1(x_1)\pi_\alpha^2(x_2)\cdots\pi_\alpha^d(x_d)}_{\text{tensor product decomposition}}$
- Construction, integrals, samples **all available at $\mathcal{O}(dn)$ cost !**

# Tensor Train (TT) surrogates for high-dim. distributions

[Dolgov, Anaya-Izquierdo, Fox, RS, 2019]

- Generic – not problem specific ("black box")
- Cross approximation: *"sequential"* design along *1D lines*
- Separable product form: $\tilde{\pi}(x_1, \ldots, x_d) = \sum_{|\alpha| \leq r} \pi_\alpha^1(x_1) \ldots \pi_\alpha^d(x_d)$

  Cheap construction/storage & low # model evals — linear in $d$

  Cheap integration w.r.t. $x$ — linear in $d$

  Cheap samples via *conditional distribution method* — linear in $d$

  **(see below)**

# Tensor Train (TT) surrogates for high-dim. distributions

[Dolgov, Anaya-Izquierdo, Fox, RS, 2019]

- Generic – not problem specific ("black box")
- Cross approximation: *"sequential"* design along *1D lines*
- Separable product form: $\tilde{\pi}(x_1, \ldots, x_d) = \sum_{|\alpha| \leq r} \pi_\alpha^1(x_1) \ldots \pi_\alpha^d(x_d)$

  Cheap construction/storage & low # model evals — linear in $d$

  Cheap integration w.r.t. $x$ — linear in $d$

  Cheap samples via *conditional distribution method* — linear in $d$

  **(see below)**

- Tuneable approximation error $\varepsilon$ (by adapting ranks $r$):

  $\Longrightarrow$ cost & storage *(poly)logarithmic in* $\varepsilon$

# Tensor Train (TT) surrogates for high-dim. distributions

[Dolgov, Anaya-Izquierdo, Fox, RS, 2019]

- Generic – not problem specific ("black box")
- Cross approximation: *"sequential"* design along *1D lines*
- Separable product form: $\tilde{\pi}(x_1, \ldots, x_d) = \sum_{|\alpha| \leq r} \pi_\alpha^1(x_1) \ldots \pi_\alpha^d(x_d)$

  Cheap construction/storage & low # model evals — linear in $d$

  Cheap integration w.r.t. $x$ — linear in $d$

  Cheap samples via *conditional distribution method* — linear in $d$

  **(see below)**

- Tuneable approximation error $\varepsilon$ (by adapting ranks $r$):

  $\Longrightarrow$ cost & storage *(poly)logarithmic in* $\varepsilon$

- Many known ways to use this surrogate for fast inference!

  (as proposals for MCMC, as control variates, importance weighting, ...)

# A Theoretical Result

[Rohrbach, Dolgov, Grasedyck, RS, in preparation]

For Gaussian distributions $\pi(x)$ we have the following result: Let

$$\pi : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto \exp\left(-\tfrac{1}{2} x^T \Sigma x\right)$$

and define

$$\Sigma := \begin{bmatrix} \Sigma_{11}^{(k)} & \Gamma_k^T \\ \Gamma_k & \Sigma_{22}^{(k)} \end{bmatrix} \quad \text{where} \quad \Gamma_k \in \mathbb{R}^{(d-k)\times k}.$$

# A Theoretical Result

[Rohrbach, Dolgov, Grasedyck, RS, in preparation]

For Gaussian distributions $\pi(x)$ we have the following result: Let

$$\pi : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto \exp\left(-\tfrac{1}{2} x^T \Sigma x\right)$$

and define

$$\Sigma := \begin{bmatrix} \Sigma_{11}^{(k)} & \Gamma_k^T \\ \Gamma_k & \Sigma_{22}^{(k)} \end{bmatrix} \quad \text{where} \quad \Gamma_k \in \mathbb{R}^{(d-k)\times k}.$$

**Theorem.** Let $\Sigma$ be SPD with $\lambda_{\min} > 0$, $\rho := \max_k \operatorname{rank}(\Gamma_k)$ and $\sigma := \max_{k,i} \sigma_i^{(k)}$, where $\sigma_i^{(k)}$ are the singular values of $\Gamma_k$. Then, for all $\varepsilon > 0$, there exists TT-approximation $\tilde{\pi}_\varepsilon$ s.t.

$$\|\pi - \tilde{\pi}_\varepsilon\|_{L^2(\mathbb{R}^d)} \leq \varepsilon \|\pi\|_{L^2(\mathbb{R}^d)}$$

and the TT-ranks of $\tilde{\pi}_\varepsilon$ are bounded by

$$r \leq \left(\left(1 + 7\frac{\sigma}{\lambda_{\min}}\right) \log\left(7\,\rho\,\frac{d}{\varepsilon}\right)\right)^\rho .$$

# Conditional Distribution Sampler for TT (TT-CD sampler)

For the TT approximation

$$\tilde{\pi}(x) = \sum_{\substack{\alpha_k=1 \\ 0<k<d}}^{r_k} \pi^1_{\alpha_1}(x_1) \cdot \pi^2_{\alpha_1,\alpha_2}(x_2) \cdot \pi^3_{\alpha_2,\alpha_3}(x_3) \cdots \pi^d_{\alpha_{d-1}}(x_d)$$

the $k$-th step of the CD sampler, given $x_1^i, \ldots, x_{k-1}^i$, simplifies to

$$\begin{aligned}\tilde{\pi}_k(x_k | x_1^i, \ldots, x_{k-1}^i) \sim \sum_{\alpha_1,\ldots,\alpha_{d-1}} & \pi^1_{\alpha_1}(x_1^i) \cdots \pi^{k-1}_{\alpha_{k-2},\alpha_{k-1}}(x_{k-1}^i) \cdots \\ & \cdots \pi^k_{\alpha_{k-1},\alpha_k}(x_k) \cdots \\ & \cdots \int \pi^{k+1}_{\alpha_k,\alpha_{k+1}}(x_{k+1}) dx_{k+1} \cdots \int \pi^d_{\alpha_{d-1}}(x_d) dx_d\end{aligned}$$

# Conditional Distribution Sampler for TT (TT-CD sampler)

For the TT approximation

$$\tilde{\pi}(x) = \sum_{\substack{\alpha_k=1 \\ 0<k<d}}^{r_k} \pi^1_{\alpha_1}(x_1) \cdot \pi^2_{\alpha_1,\alpha_2}(x_2) \cdot \pi^3_{\alpha_2,\alpha_3}(x_3) \cdots \pi^d_{\alpha_{d-1}}(x_d)$$

the $k$-th step of the CD sampler, given $x_1^i, \ldots, x_{k-1}^i$, simplifies to

$$\tilde{\pi}_k(x_k | x_1^i, \ldots, x_{k-1}^i) \sim \sum_{\alpha_1,\ldots,\alpha_{d-1}} \pi^1_{\alpha_1}(x_1^i) \cdots \pi^{k-1}_{\alpha_{k-2},\alpha_{k-1}}(x_{k-1}^i) \cdots$$
$$\cdots \pi^k_{\alpha_{k-1},\alpha_k}(x_k) \cdots$$
$$\cdots \int \pi^{k+1}_{\alpha_k,\alpha_{k+1}}(x_{k+1}) dx_{k+1} \cdots \int \pi^d_{\alpha_{d-1}}(x_d) dx_d$$

*To sample:* Simple 1D CDF-inversion

**linear** in $d$

# How to use TT-CD sampler to estimate $\mathbb{E}_\pi Q$ ?

*Problem:* We are sampling from approximate $\tilde{\pi} = \pi + \mathcal{O}(\varepsilon)$.

# How to use TT-CD sampler to estimate $\mathbb{E}_\pi Q$ ?

*Problem:* We are sampling from approximate $\tilde{\pi} = \pi + \mathcal{O}(\varepsilon)$.

*Option 0:*

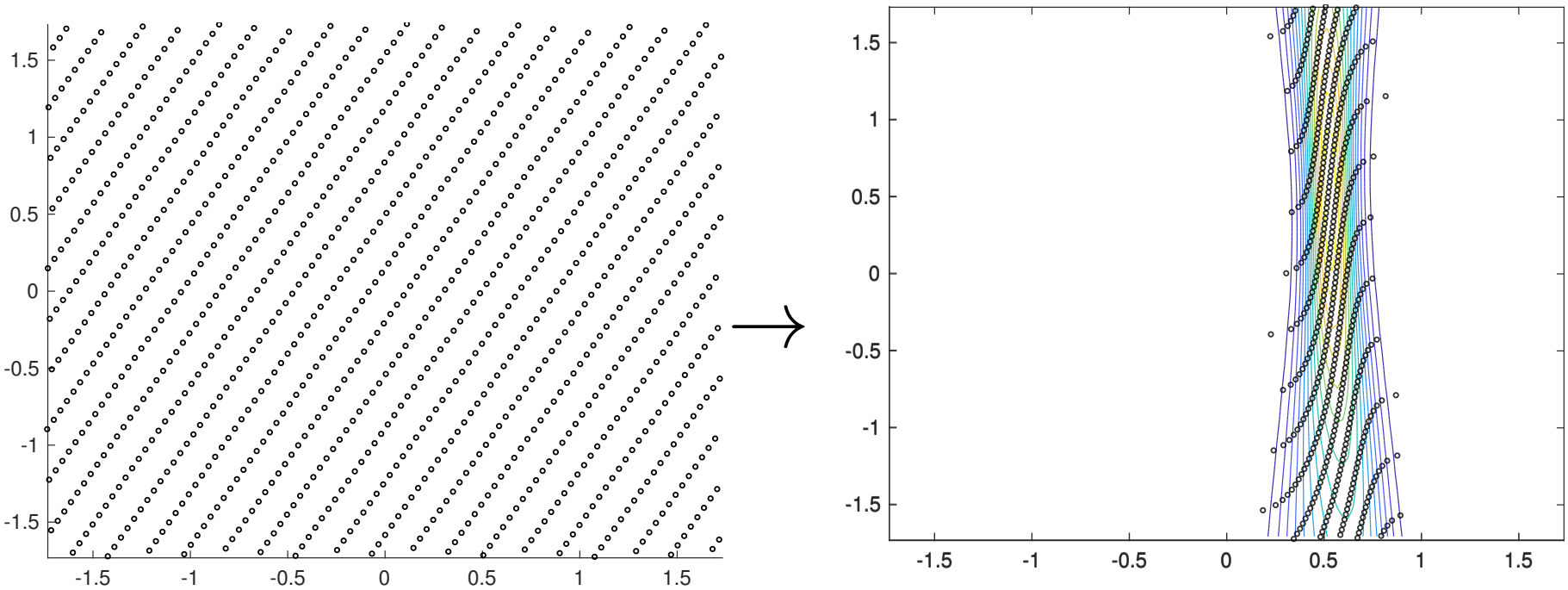
- Biased estimator $\mathbb{E}_\pi Q \approx \mathbb{E}_{\tilde{\pi}} Q$ via i.i.d. MC quadrature

# How to use TT-CD sampler to estimate $\mathbb{E}_\pi Q$ ?

*Problem:* We are sampling from approximate $\tilde{\pi} = \pi + \mathcal{O}(\varepsilon)$.

*Option 0:*

- Biased estimator $\mathbb{E}_\pi Q \approx \mathbb{E}_{\tilde{\pi}} Q$ via i.i.d. MC quadrature
- Can use *QMC "seeds"* instead of random ones



# Sampling from exact $\pi$: Unbiased estimates of $\mathbb{E}_\pi Q$

*Option 1:* Use $\{x^i_{\tilde{\pi}}\}$ as (i.i.d.) *proposals* in Metropolis-Hastings:

- Accept proposal $x^i_{\tilde{\pi}}$ with probability $\alpha = \min\left(1, \frac{\pi(x^i_{\tilde{\pi}})\tilde{\pi}(x^{i-1}_{\pi})}{\pi(x^{i-1}_{\pi})\tilde{\pi}(x^i_{\tilde{\pi}})}\right)$
- Can prove that **rejection rate** $\sim \varepsilon$ and **IACT** $\tau \sim 1 + \varepsilon$

# Sampling from exact $\pi$: Unbiased estimates of $\mathbb{E}_\pi Q$

*Option 1:* Use $\{x^i_{\tilde\pi}\}$ as (i.i.d.) *proposals* in Metropolis-Hastings:

- Accept proposal $x^i_{\tilde\pi}$ with probability $\alpha = \min\left(1, \frac{\pi(x^i_{\tilde\pi})\tilde\pi(x^{i-1}_\pi)}{\pi(x^{i-1}_\pi)\tilde\pi(x^i_{\tilde\pi})}\right)$
- Can prove that **rejection rate** $\sim \varepsilon$ and **IACT** $\tau \sim 1 + \varepsilon$

*Option 2:* Use $\tilde\pi$ for *importance weighting* + QMC quadrature:

$$\mathbb{E}_\pi Q \approx \frac{1}{Z}\frac{1}{N}\sum_{i=1}^N Q(x^i_{\tilde\pi})\frac{\pi(x^i_{\tilde\pi})}{\tilde\pi(x^i_{\tilde\pi})} \quad \text{with} \quad Z = \frac{1}{N}\sum_{i=1}^N \frac{\pi(x^i_{\tilde\pi})}{\tilde\pi(x^i_{\tilde\pi})}$$

- We can use an unbiased (randomised) QMC rule for both integrals.

# Sampling from exact $\pi$: Unbiased estimates of $\mathbb{E}_\pi Q$

*Option 1:* Use $\{x^i_{\tilde\pi}\}$ as (i.i.d.) *proposals* in Metropolis-Hastings:

- Accept proposal $x^i_{\tilde\pi}$ with probability $\alpha = \min\left(1, \frac{\pi(x^i_{\tilde\pi})\tilde\pi(x^{i-1}_\pi)}{\pi(x^{i-1}_\pi)\tilde\pi(x^i_{\tilde\pi})}\right)$
- Can prove that **rejection rate** $\sim \varepsilon$ and **IACT** $\tau \sim 1 + \varepsilon$

*Option 2:* Use $\tilde\pi$ for *importance weighting* + QMC quadrature:

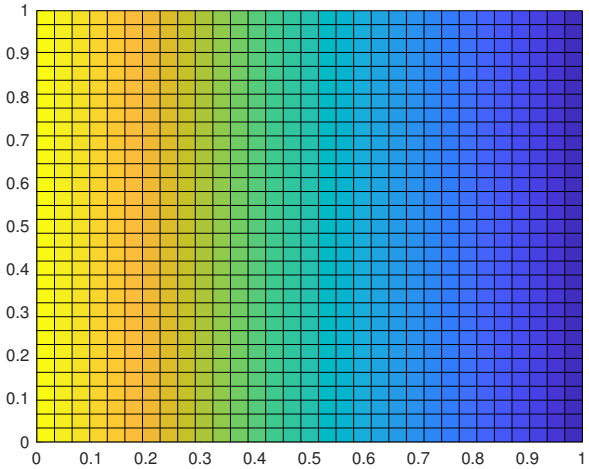
$$\mathbb{E}_\pi Q \approx \frac{1}{Z}\frac{1}{N}\sum_{i=1}^N Q(x^i_{\tilde\pi})\frac{\pi(x^i_{\tilde\pi})}{\tilde\pi(x^i_{\tilde\pi})} \quad \text{with} \quad Z = \frac{1}{N}\sum_{i=1}^N \frac{\pi(x^i_{\tilde\pi})}{\tilde\pi(x^i_{\tilde\pi})}$$

- We can use an unbiased (randomised) QMC rule for both integrals.

*Option 3:* Use biased QMC estimator as a *control variate* (**MLMCMC**)

# Numerical experiments: (Artificial) Inverse Diffusion Problem

$$
\begin{aligned}
-\nabla\kappa(s,x)\nabla u &= 0 \qquad s \in (0,1)^2 \\
u|_{s_1=0} &= 1, \qquad u|_{s_1=1} = 0, \\
\frac{\partial u}{\partial n}\Big|_{s_2=0} &= \frac{\partial u}{\partial n}\Big|_{s_2=1} = 0.
\end{aligned}
$$

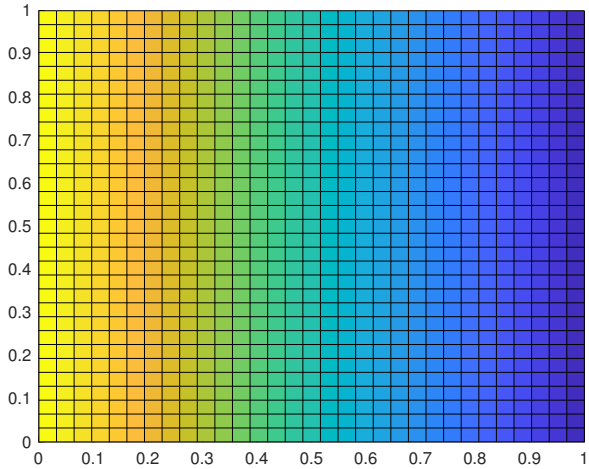



- Karhunen-Loève expansion[2] of $\log\kappa(s,x) = \sum_{k=1}^{d} \phi_k(s)x_k$ with prior $x_k \sim U[-1,1]$, $\|\phi_k\|_\infty = \mathcal{O}(k^{-\frac{3}{2}})$ & $d = 11$.

[2] Eigel, Pfeffer, Schneider, 2016.

# Numerical experiments: (Artificial) Inverse Diffusion Problem

$$
\begin{aligned}
-\nabla\kappa(s,x)\nabla u &= 0 \qquad s \in (0,1)^2 \\
u|_{s_1=0} &= 1, \qquad u|_{s_1=1} = 0, \\
\frac{\partial u}{\partial n}\Big|_{s_2=0} &= \frac{\partial u}{\partial n}\Big|_{s_2=1} = 0.
\end{aligned}
$$

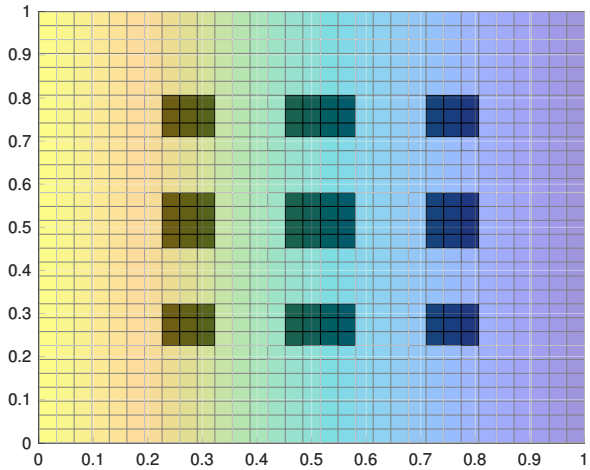



- Karhunen-Loève expansion[2] of $\log\kappa(s,x) = \sum_{k=1}^{d} \phi_k(s)x_k$ with prior $x_k \sim U[-1,1]$, $\|\phi_k\|_\infty = \mathcal{O}(k^{-\frac{3}{2}})$ & $d = 11$.
- Discretisation with bilinear FEs on uniform mesh with $h = 1/64$.

[2] Eigel, Pfeffer, Schneider, 2016.

# Numerical experiments: (Artificial) Inverse Diffusion Problem

$$
\begin{aligned}
-\nabla\kappa(s,x)\nabla u &= 0 \qquad s \in (0,1)^2 \\
u|_{s_1=0} &= 1, \qquad u|_{s_1=1} = 0, \\
\frac{\partial u}{\partial n}\Big|_{s_2=0} &= \frac{\partial u}{\partial n}\Big|_{s_2=1} = 0.
\end{aligned}
$$

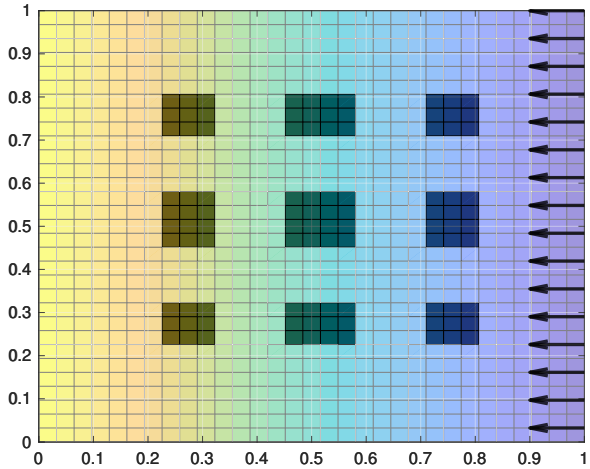



- Karhunen-Loève expansion[2] of $\log\kappa(s,x) = \sum_{k=1}^{d} \phi_k(s)x_k$ with prior $x_k \sim U[-1,1]$, $\|\phi_k\|_\infty = \mathcal{O}(k^{-\frac{3}{2}})$ & $d = 11$.
- Discretisation with bilinear FEs on uniform mesh with $h = 1/64$.
- **Data:** average pressure in 9 locations (synthetic, i.e. for some $s^*$)

[2] Eigel, Pfeffer, Schneider, 2016.

# Numerical experiments: (Artificial) Inverse Diffusion Problem

$$
\begin{aligned}
-\nabla\kappa(s,x)\nabla u &= 0 \qquad s \in (0,1)^2 \\
u|_{s_1=0} &= 1, \qquad u|_{s_1=1} = 0, \\
\frac{\partial u}{\partial n}\Big|_{s_2=0} &= \frac{\partial u}{\partial n}\Big|_{s_2=1} = 0.
\end{aligned}
$$

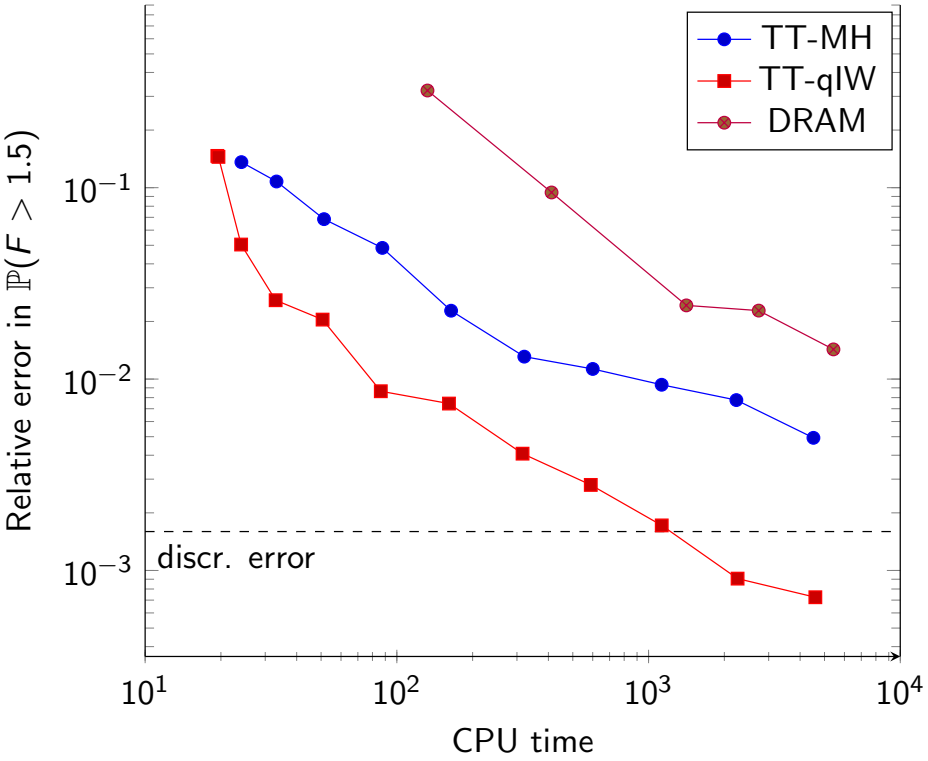



- Karhunen-Loève expansion[2] of $\log\kappa(s,x) = \sum_{k=1}^{d} \phi_k(s)x_k$ with prior $x_k \sim U[-1,1]$, $\|\phi_k\|_\infty = \mathcal{O}(k^{-\frac{3}{2}})$ & $d = 11$.
- Discretisation with bilinear FEs on uniform mesh with $h = 1/64$.
- **Data:** average pressure in 9 locations (synthetic, i.e. for some $s^*$)
- **QoI:** probability that flux exceeds 1.5

[2] Eigel, Pfeffer, Schneider, 2016.

# Comparison against DRAM (for inverse diffusion problem)

noise level $\sigma_e^2 = 0.01$



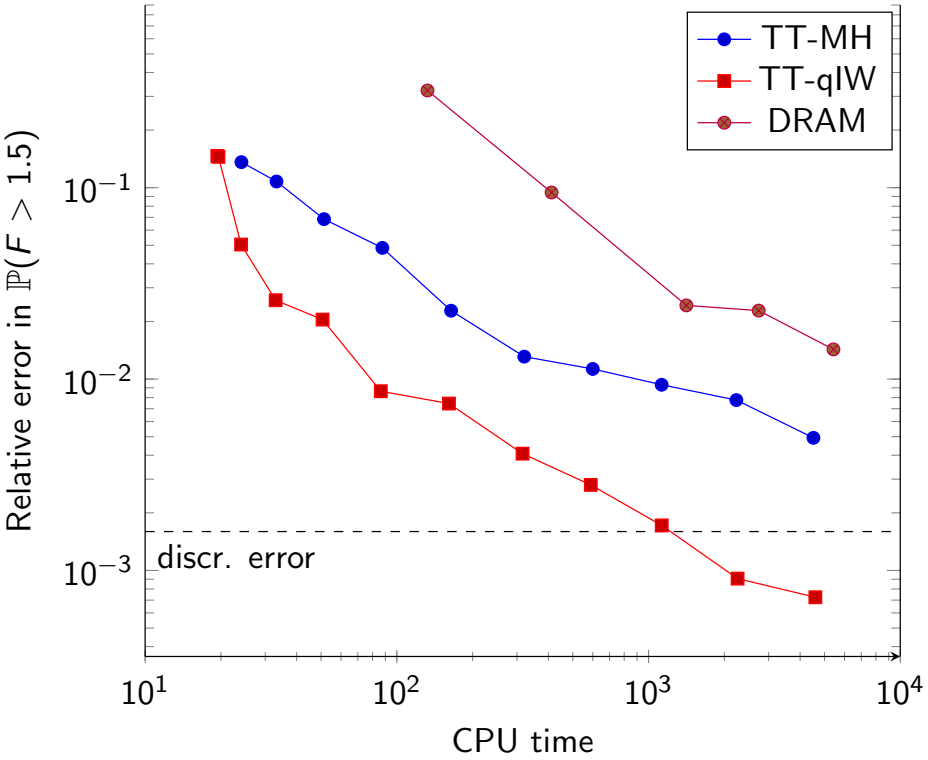

TT-MH TT conditional distribution samples (iid) as proposals for MCMC

TT-qIW TT surrogate for importance sampling with QMC

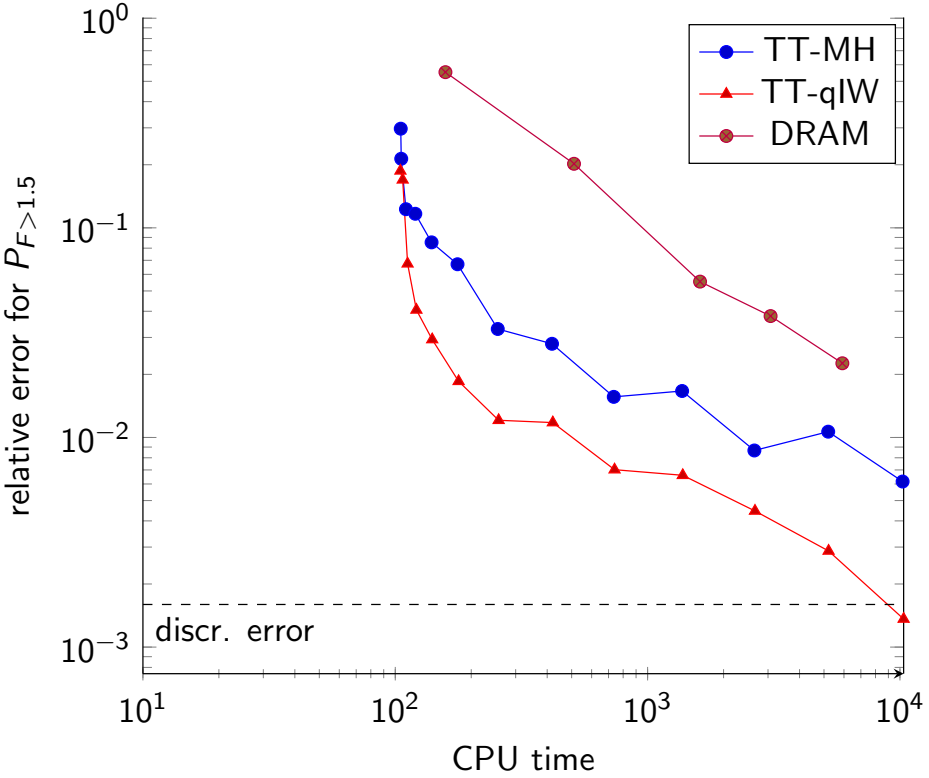
DRAM Delayed Rejection Adaptive Metropolis [Haario et al, 2006]

# Comparison against DRAM (for inverse diffusion problem)

noise level $\sigma_e^2 = 0.01$




noise level $\sigma_e^2 = 0.001$



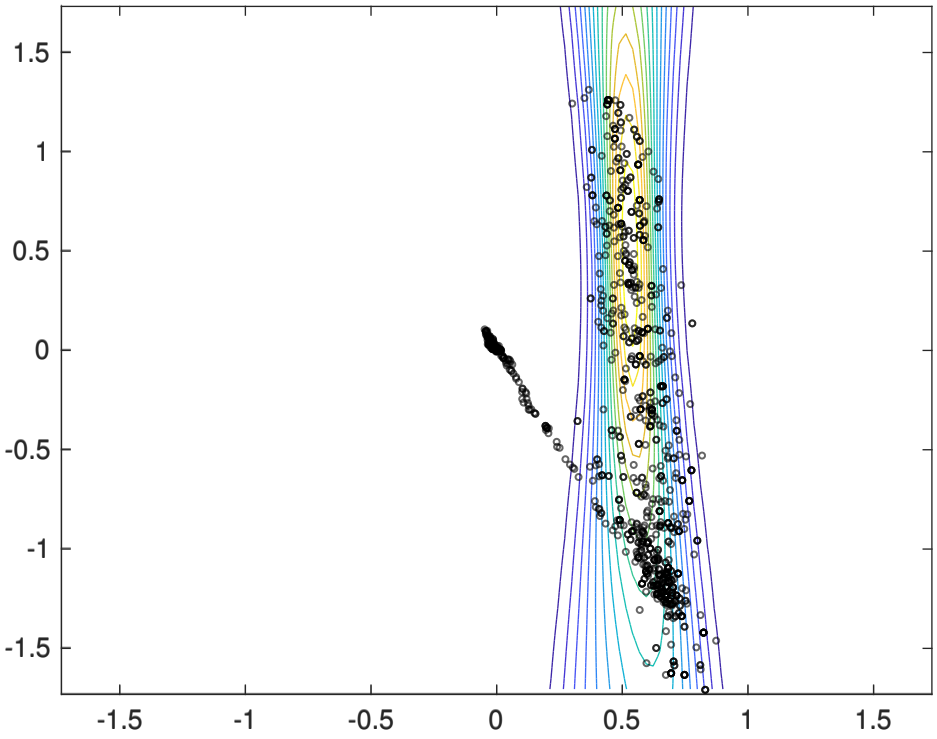

TT-MH TT conditional distribution samples (iid) as proposals for MCMC

TT-qIW TT surrogate for importance sampling with QMC

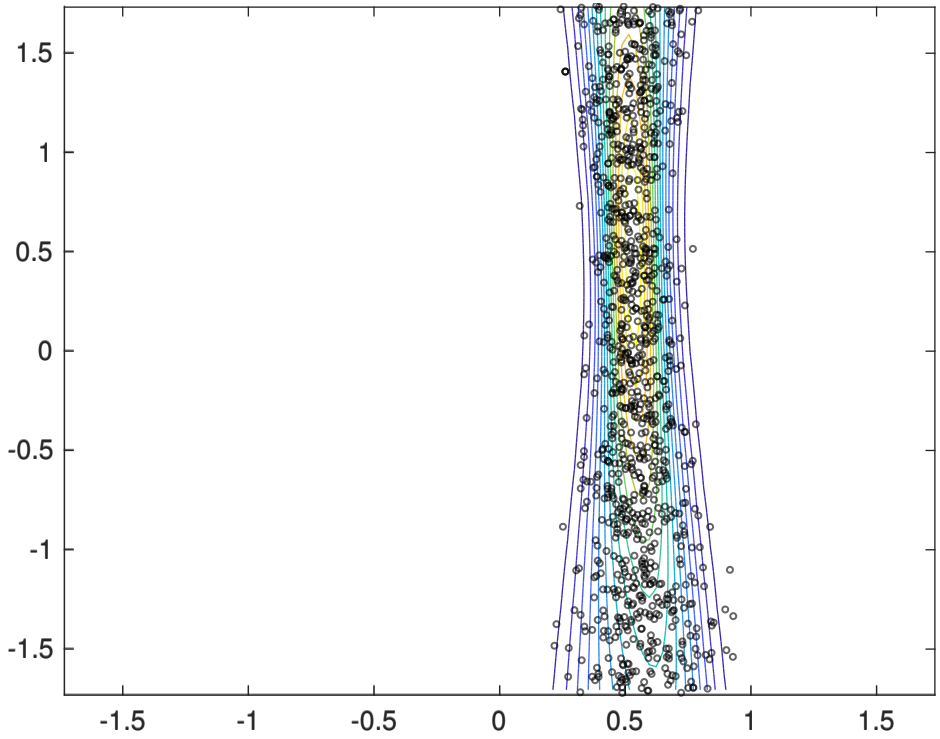
DRAM Delayed Rejection Adaptive Metropolis [Haario et al, 2006]

# Samples – Comparison TT-CD vs. DRAM

DRAM




TT-MH (i.i.d. seeds)



# Conclusions

- Inverse Problems under Uncertainty – Variational Inference
- **Central idea:** characterise complex/intractable distributions by constructing deterministic *couplings*
- **Central tool:** Optimisation of Kullback-Leibler divergence

# Conclusions

- Inverse Problems under Uncertainty – Variational Inference
- **Central idea:** characterise complex/intractable distributions by constructing deterministic *couplings*
- **Central tool:** Optimisation of Kullback-Leibler divergence
- Many types of approximation classes (non-exhaustive):
  - Sparse maps, decomposable maps, neural nets
  - Kernel-based approaches
  - Low rank structure

# Conclusions

- Inverse Problems under Uncertainty – Variational Inference
- **Central idea:** characterise complex/intractable distributions by constructing deterministic *couplings*
- **Central tool:** Optimisation of Kullback-Leibler divergence
- Many types of approximation classes (non-exhaustive):
  - Sparse maps, decomposable maps, neural nets
  - Kernel-based approaches
  - Low rank structure
- **Main Topic 1:** Newton-acceleration and data-informed kernels for Stein Variational Methods
- **Main Topic 2:** TT surrogates for efficient samplers in high dimensions
- Use approximate maps to accelerate MCMC or in importance sampler

# References


1. Moselhy, Marzouk, *Bayesian inference with optimal maps*, J Comput Phys 231, 2012 [arXiv:1109.1516]
2. Rezende, Mohamed, *Variational inference with normalizing flows*, ICML'15 Proc. 32nd Inter. Conf. Machine Learning, Vol. 37, 2015 [arXiv:1505.05770]
3. Marzouk, Moselhy, Parno, Spantini, *Sampling via measure transport: An introduction*, Handbook of Uncertainty Quantification (Ghanem, Higdon, Owhadi, Eds.), 2016 [arXiv:1602.05023]
4. Liu, Wang, *Stein variational gradient descent: A general purpose Bayesian inference algorithm*, NIPS 2016, Vol. 29, 2016 [arXiv:1608.04471]
5. Detommaso, Cui, Spantini, Marzouk, RS, *A Stein variational Newton method*, NIPS 2018, Vol. 31, 2018 [arXiv:1806.03085]
6. Dolgov, Anaya-Izquierdo, Fox, RS, *Approximation and sampling of multivariate probability distributions in the tensor train decomposition*, Statistics & Comput. (online first), 2019 [arXiv:1810.01212]
7. Detommaso, Kruse, Ardizzone, Rother, Köthe, RS, *HINT: Hierarchical invertible neural transport for general & sequential Bayesian inference*, 2019 [arXiv:1905.10687]