Self-reinforced Knothe–Rosenblatt rearrangements for high-dimensional stochastic computation

In collaboration with Sergey Dolgov (Bath), Rob Scheichl (Heidelberg) and Olivier Zahm (INRIA)

Tiangang Cui
Monash University

tiangang.cui@monash.edu
https://www.fastfins.org/

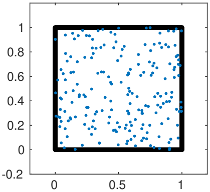
9th Workshop on High-Dimensional Approximation, Canberra, 2023.

Generating random variables

Uniform random variable

$$\pi(x) = \mathrm{ind}(x \in [0,1]^d)$$

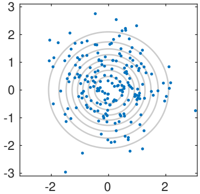
```
>> X = rand(2,200);
```



Gaussian random variable

$$\pi(x) = \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{1}{2}\|x\|^2\Big)$$

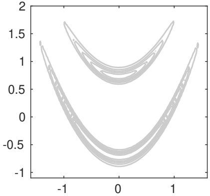
```
>> X = random('norm', 0, 1, 2, 200);
```



Random variable with potential function

$$-\log \pi(x) = -5(log((1-x_1^2+100(x_2-x_1^2)^2)-3)^2+\cdots$$

```
>> phi= @(x) ...;
>> base = ApproxBases(Legendre(30), ...
          BoundedDomain([-4,4]), d); % d = 2
>> kr = TTDIRT(phi,base,Tempering1());
```

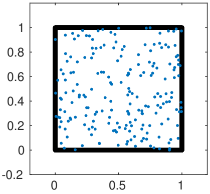


Generating random variables

Uniform random variable

$$\pi(x) = \mathrm{ind}(x \in [0,1]^d)$$

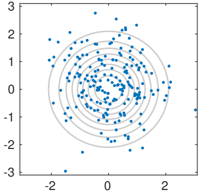
```
>> X = rand(2,200);
```



Gaussian random variable

$$\pi(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(- \frac{1}{2}\|x\|^2 \right)$$

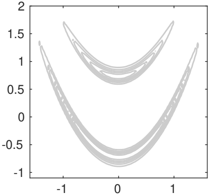
```
>> X = random('norm', 0, 1, 2, 200);
```



Random variable with potential function

$$-\log \pi(x) = -5(log((1 - x_1^2 + 100(x_2 - x_1^2)^2) - 3)^2 + \cdots$$

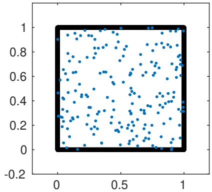
```
>> phi= @(x) ...;
>> base = ApproxBases(Legendre(30), ...
          BoundedDomain([-4,4]), d); % d = 2
>> kr = TTDIRT(phi,base,Tempering1());
```



Uniform random variable

$$\pi(x) = \mathrm{ind}(x \in [0,1]^d)$$

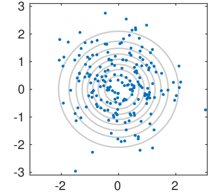
```
>> X = rand(2,200);
```



Gaussian random variable

$$\pi(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|x\|^2 \right)$$

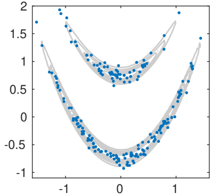
```
>> X = random('norm', 0, 1, 2, 200);
```



Random variable with potential function

$$-\log \pi(x) = -5(log((1-x_1^2+100(x_2-x_1^2)^2)-3)^2+\cdots$$

```
>> phi= @(x) ...;
>> base = ApproxBases(Legendre(30), ...
          BoundedDomain([-4,4]), d); % d = 2
>> kr = TTDIRT(phi,base,Tempering1());
>> X = random(kr, 200);
```

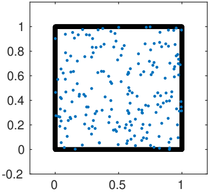


Generating random variables

Uniform random variable

$$\pi(x) = \mathrm{ind}(x \in [0,1]^d)$$

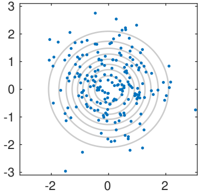
```
>> X = rand(2,200);
```



Gaussian random variable

$$\pi(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|x\|^2 \right)$$

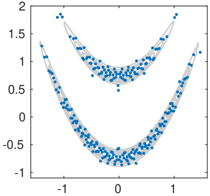
```
>> X = random('norm', 0, 1, 2, 200);
```



Random variable with potential function

$$-\log \pi(x) = -5(log((1-x_1^2+100(x_2-x_1^2)^2)-3)^2+\cdots$$

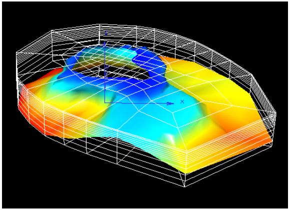
```
>> phi= @(x) ...;
>> base = ApproxBases(Legendre(30), ...
          BoundedDomain([-4,4]), d); % d = 2
>> kr = TTDIRT(phi,base,Tempering1());
>> X = sobol(kr, 200);
```



Motivation: inverse problems

Data

Parameter



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

forward model (PDE, ODE) ← G

e → observation/model errors

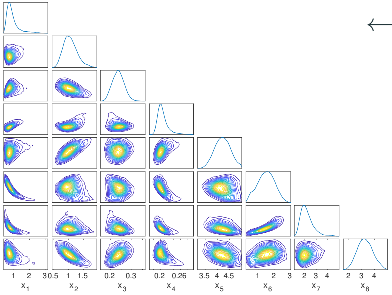
$$\textit{Bayes' rule}: \quad \underbrace{\pi(x|y)}_{\textit{Posterior}} \propto \underbrace{L(y|G(x))}_{\textit{Likelihood}} \underbrace{\pi_0(x)}_{\textit{Prior}}$$

Common operations:

* Sampling: $X|(Y{=}y) \sim \pi(x|y)$
* Expectation: $\int g(x)\,\pi(x|y)\,dx$
* Optimisation under uncertainty: $\underset{\theta}{\operatorname{argmin}} \int \phi(\theta, x)\,\pi(x|y)\,dx$

Motivation: inverse problems

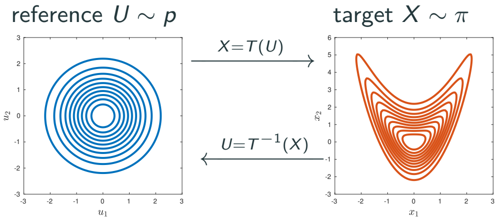
Why posterior random variables are special/difficult?



⟵ posterior of a circadian model

- ∗ high-dimensional
- ∗ analytically intractable
- ∗ nonlinear interactions among variables
- ∗ concentrated density
- ∗ classical Markov chain methods are hard to parallelise and optimise

Find an order-preserving transport map $T : \mathbb{R}^d \to \mathbb{R}^d$



Variational approach: minimises divergence of π from $T_\sharp\, p$ over a class of T

$$KL(\pi \| T_\sharp\, p) = \int \log \frac{\pi}{T_\sharp\, p} d\pi$$

$$\approx \int \log \pi d\pi - \frac{1}{N} \sum_{j=1}^{N} \left(\log p(T^{-1}(X^j)) + \log |\nabla T^{-1}(X^j)| \right), \quad X^j \sim \pi$$

- Rosenblatt transport [Bigoni, Marzouk, Parno, Spantini, Zahm, Zech ...]
- Normalizing flows [Kruse, Mohamed, Papamakarios, Rezende ...]
- Stein variational methods [Liu, Lu, Detommaso, Wang ...]
- Minmax formulation [Tabak, Trigila, Turner, Zhao ...]

Function approximation: we approximate π by a tensor train (TT) $\tilde{\pi}$

Knothe–Rosenblatt rearrangement

* Factorise a joint density into a product of marginals and conditionals:

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

* Marginal

$$\pi_{\leq k}(x_1, \ldots, x_k) = \int \pi(x_1, \ldots, x_k, x_{k+1}, x_d)dx_{k+1}\cdots dx_d$$

* Marginal CDF

$$F_{\leq 1}(x_1) = \int_{\infty}^{x_1} \pi_{\leq 1}(x_1)dx_1'$$

* Conditional

$$\pi_k(x_k|x_1, \ldots, x_{k-1}) = \frac{\pi_{\leq k}(x_1, \ldots, x_{k-1}, x_k)}{\pi_{\leq k-1}(x_1, \ldots, x_{k-1})}$$

* Conditional CDF

$$F_k(x_k|x_1, \ldots, x_{k-1}) = \int_{\infty}^{x_k} \pi_k(x_k|x_1, \ldots, x_{k-1})dx_k'$$

Knothe–Rosenblatt rearrangement

* Factorise a joint density into a product of marginals and conditionals:

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

* Marginal

$$\pi_{\leq k}(x_1, \dots, x_k) = \int \pi(x_1, \dots, x_k, x_{k+1}, x_d)dx_{k+1}\cdots dx_d$$

* Marginal CDF

$$F_{\leq 1}(x_1) = \int_{\infty}^{x_1} \pi_{\leq 1}(x_1)dx_1'$$

* Conditional

$$\pi_k(x_k|x_1, \dots, x_{k-1}) = \frac{\pi_{\leq k}(x_1, \dots, x_{k-1}, x_k)}{\pi_{\leq k-1}(x_1, \dots, x_{k-1})}$$

* Conditional CDF

$$F_k(x_k|x_1, \dots, x_{k-1}) = \int_{\infty}^{x_k} \pi_k(x_k|x_1, \dots, x_{k-1})dx_k'$$

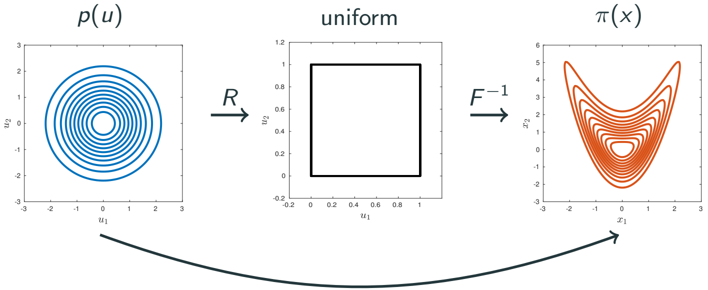
Knothe–Rosenblatt rearrangement

Rosenblatt transport: $F(X) = Z$, where $Z_k \sim \mathrm{unifom}[0,1]$

$$F(x) = \begin{cases} z_1 &= \Pr[X_1 \leq x_1] &= F_1(x_1) \\ &\vdots & \\ z_k &= \Pr[X_k \leq x_k | X_{<k} = x_{<k}] &= F_k(x_k|x_1, \ldots, x_{k-1}) \\ &\vdots & \\ z_d &= \Pr[X_d \leq x_d | X_{<d} = x_{<d}] &= F_d(x_d|x_1, \ldots, x_{d-1}) \end{cases},$$

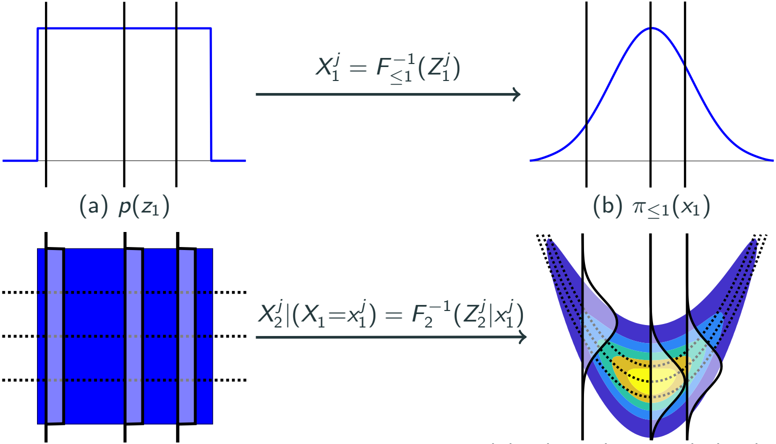
Inverse Rosenblatt transport: $X = F^{-1}(Z)$

Knothe–Rosenblatt rearrangement: general reference measures



$$\pi = (F^{-1} \circ R)_\sharp\, p$$

Uniform reference: $Z \sim \text{uniform}(0,1]^d$



(a) $p(z_1)$

(b) $\pi_{\leq 1}(x_1)$

(c) $p(z_1, z_2) = p_1(z_1)p_2(z_2)$

(d) $\pi(x_1, x_2) = \pi_{\leq 1}(x_1)\pi_2(x_2|x_1)$

The key is to compute all marginal densities:

$$\pi_{\leq k}(x_1, \ldots, x_k) = \int \pi(x_1, \ldots, x_k, x_{k+1}, x_d) dx_{k+1} \cdots dx_d$$

* Approximate multivariate $\pi(x_1, \ldots, x_d)$ (or a tensor after discretisation):

  Matrix Product States/Tensor-Train[1]:

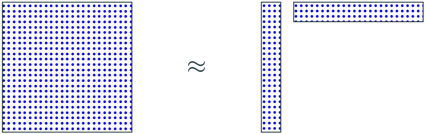
$$\pi(x_1, \ldots, x_d) \approx \underbrace{\sum_{\alpha_0, \alpha_1, \cdots, \alpha_d}^{r_0, r_1, \cdots, r_d} \mathcal{F}_1^{(\alpha_0, \alpha_1)}(x_1) \cdots \mathcal{F}_k^{(\alpha_{k-1}, \alpha_k)}(x_k) \cdots \mathcal{F}_d^{(\alpha_{d-1}, \alpha_d)}(x_d)}_{\equiv \tilde{\pi}(x_1, \ldots, x_d) \quad \text{tensor product decomposition}}$$

* What is this and what is the benefit?

[1] Wilson '75, White '93, Verstraete '04; Oseledets '09; Hackbusch, Kühn '09

Functional tensor train: what happened in 2D

* 2D: discretisation of $\pi(x_1, x_2)$ gives a matrix $A(i, j)$



* Function/matrix decomposition

$$\pi(x_1, x_2) = \sum_{\alpha_1=1}^{r} \mathcal{F}^{(\alpha_1)}(x_1)\mathcal{F}^{(\alpha_1)}(x_2) + \mathcal{O}(\varepsilon)$$

$$A(i, j) = \sum_{\alpha_1=1}^{r} V^{(\alpha_1)}(i)W^{(\alpha_1)}(j) + \mathcal{O}(\varepsilon)$$

* *Rank* $r \ll n$ to save storage and computing cost. Separates x_1 and x_2

$$\pi(x_2) = \int \pi(x_1, x_2)dx_1 \approx \sum_{\alpha_1=1}^{r} \left(\int \mathcal{F}^{(\alpha_1)}(x_1)dx_1\right) \mathcal{F}^{(\alpha_1)}(x_2)$$

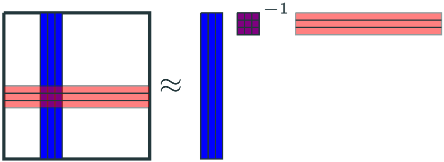
$$\pi(x_1) = \int \pi(x_1, x_2)dx_2 \approx \sum_{\alpha_1=1}^{r} \mathcal{F}^{(\alpha_1)}(x_1) \left(\int \mathcal{F}^{(\alpha_1)}(x_2)dx_2\right)$$

Tensor train: cross interpolation

* How to build this? SVD/Schmidt decomposition (unpractical for large d) needs high-dimensional integration
* A rank-r matrix can be recovered by its r indep. rows and cols:

$$A = \hat{A} \equiv A(:, \mathcal{J}) A^{-1}(\mathcal{I}, \mathcal{J}) A(\mathcal{I}, :),$$

for some *index sets* $\mathcal{I}, \mathcal{J} \subset \{1, \dots, n\}$ of cardinality r.



Tricks are in details:

* what are the *best* and *practically realizable* indices $\mathcal{I}, \mathcal{J}$?
* how to generalize to $d > 2$ dimensions?

Tensor train: MaxVol and alternating iteration

Maximum Volume (MaxVol): *best* indices in Chebyshev norm:

* If $\left|\det A(\mathcal{I},\mathcal{J})\right| = \max_{\hat{\mathcal{I}},\hat{\mathcal{J}}} \left|\det A(\hat{\mathcal{I}},\hat{\mathcal{J}})\right|$, then $\|A-\hat{A}\|_\infty \le (r+1)\sigma_{r+1}(A)$.
* NP-hard to look through all submatrices, so apply alternating iterations

Given an *initial set* $\mathcal{J} \subset \{1 \ldots n\}$, repeat the following

1. $V = A(:,\mathcal{J})$
2. $\mathcal{I} = \texttt{pivots}(V)$
3. $W = A(\mathcal{I},:)$.
4. $\mathcal{J} = \texttt{pivots}(W^\top)$.

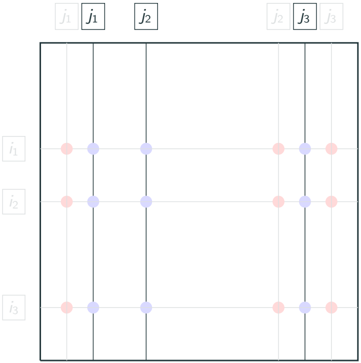
$\texttt{pivots}$ via LU or QR with $\mathcal{O}(nr^2)$ flops [Goreinov, Tyrtyshnikov 2001]

Easily generalisable to $d>2$ and functions $\pi(x_1,\ldots,x_{k-1},x_k,x_{k+1},\ldots,x_d)$

Maximum Volume (MaxVol): *best* indices in Chebyshev norm:

* If $\left|\det A(\mathcal{I},\mathcal{J})\right| = \max_{\hat{\mathcal{I}},\hat{\mathcal{J}}} \left|\det A(\hat{\mathcal{I}},\hat{\mathcal{J}})\right|$, then $\|A - \hat{A}\|_\infty \le (r+1)\sigma_{r+1}(A)$.
* NP-hard to look through all submatrices, so apply alternating iterations

Given an *initial set* $\mathcal{J} \subset \{1 \ldots n\}$, repeat the following



1. $V = A(:,\mathcal{J})$
2. $\mathcal{I} = \texttt{pivots}(V)$
3. $W = A(\mathcal{I},:)$.
4. $\mathcal{J} = \texttt{pivots}(W^\top)$.

pivots via LU or QR with $\mathcal{O}(nr^2)$ flops [Goreinov, Tyrtyshnikov 2001]

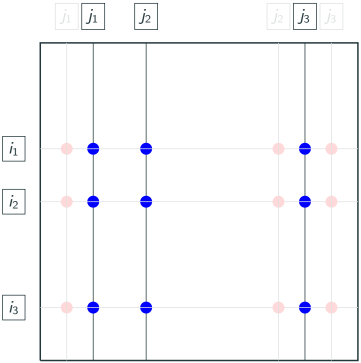
Easily generalisable to $d > 2$ and functions $\pi(x_1,\ldots,x_{k-1},x_k,x_{k+1},\ldots,x_d)$

Tensor train: MaxVol and alternating iteration

Maximum Volume (MaxVol): *best* indices in Chebyshev norm:

* If $\left|\det A(\mathcal{I},\mathcal{J})\right| = \max_{\hat{\mathcal{I}},\hat{\mathcal{J}}} \left|\det A(\hat{\mathcal{I}},\hat{\mathcal{J}})\right|$, then $\|A-\hat{A}\|_\infty \le (r+1)\sigma_{r+1}(A)$.
* NP-hard to look through all submatrices, so apply alternating iterations

Given an *initial set* $\mathcal{J} \subset \{1\dots n\}$, repeat the following



1. $V = A(:,\mathcal{J})$
2. $\mathcal{I} = \texttt{pivots}(V)$
3. $W = A(\mathcal{I},:)$.
4. $\mathcal{J} = \texttt{pivots}(W^\top)$.

pivots via LU or QR with $\mathcal{O}(nr^2)$ flops [Goreinov, Tyrtyshnikov 2001]

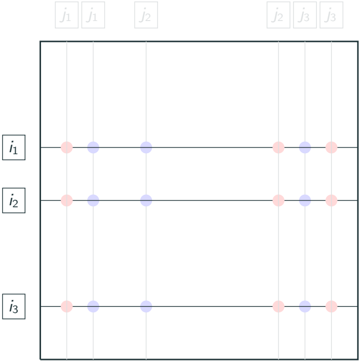
Easily generalisable to $d>2$ and functions $\pi(x_1,\dots,x_{k-1},x_k,x_{k+1},\dots,x_d)$

Tensor train: MaxVol and alternating iteration

Maximum Volume (MaxVol): *best* indices in Chebyshev norm:

* If $\left|\det A(\mathcal{I},\mathcal{J})\right| = \max_{\hat{\mathcal{I}},\hat{\mathcal{J}}} \left|\det A(\hat{\mathcal{I}},\hat{\mathcal{J}})\right|$, then $\|A-\hat{A}\|_\infty \le (r+1)\sigma_{r+1}(A)$.
* NP-hard to look through all submatrices, so apply alternating iterations

Given an *initial set* $\mathcal{J} \subset \{1\dots n\}$, repeat the following



1. $V = A(:,\mathcal{J})$
2. $\mathcal{I} = \texttt{pivots}(V)$
3. $W = A(\mathcal{I},:)$.
4. $\mathcal{J} = \texttt{pivots}(W^\top)$.

pivots via LU or QR with $\mathcal{O}(nr^2)$ flops [Goreinov, Tyrtyshnikov 2001]

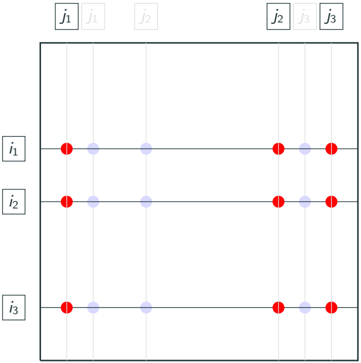
Easily generalisable to $d>2$ and functions $\pi(x_1,\dots,x_{k-1},x_k,x_{k+1},\dots,x_d)$

Tensor train: MaxVol and alternating iteration

Maximum Volume (MaxVol): *best* indices in Chebyshev norm:

* If $\left|\det A(\mathcal{I},\mathcal{J})\right| = \max_{\hat{\mathcal{I}},\hat{\mathcal{J}}} \left|\det A(\hat{\mathcal{I}},\hat{\mathcal{J}})\right|$, then $\|A - \hat{A}\|_\infty \le (r+1)\sigma_{r+1}(A)$.
* NP-hard to look through all submatrices, so apply alternating iterations

Given an *initial set* $\mathcal{J} \subset \{1 \dots n\}$, repeat the following



1. $V = A(:,\mathcal{J})$
2. $\mathcal{I} = \texttt{pivots}\,(V)$
3. $W = A(\mathcal{I},:)$.
4. $\mathcal{J} = \texttt{pivots}\,(W^\top)$.

pivots via LU or QR with $\mathcal{O}(nr^2)$ flops [Goreinov, Tyrtyshnikov 2001]

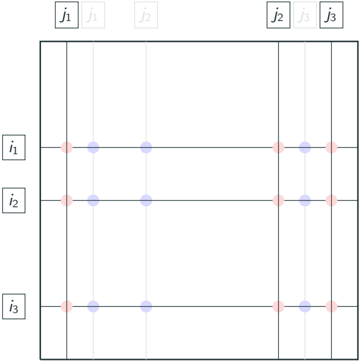
Easily generalisable to $d > 2$ and functions $\pi(x_1,\dots,x_{k-1},x_k,x_{k+1},\dots,x_d)$

Tensor train: MaxVol and alternating iteration

Maximum Volume (MaxVol): *best* indices in Chebyshev norm:

* If $\left|\det A(\mathcal{I},\mathcal{J})\right| = \max_{\hat{\mathcal{I}},\hat{\mathcal{J}}} \left|\det A(\hat{\mathcal{I}},\hat{\mathcal{J}})\right|$, then $\|A-\hat{A}\|_\infty \le (r+1)\sigma_{r+1}(A)$.
* NP-hard to look through all submatrices, so apply alternating iterations

Given an *initial set* $\mathcal{J} \subset \{1\dots n\}$, repeat the following



1. $V = A(:,\mathcal{J})$
2. $\mathcal{I} = \texttt{pivots}(V)$
3. $W = A(\mathcal{I},:)$.
4. $\mathcal{J} = \texttt{pivots}\left(W^\top\right)$.

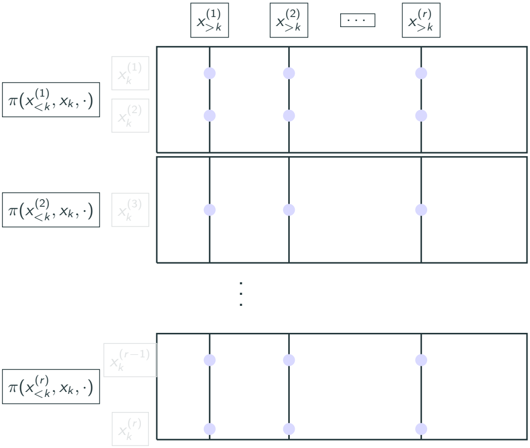
`pivots` via LU or QR with $\mathcal{O}(nr^2)$ flops [Goreinov, Tyrtyshnikov 2001]

Easily generalisable to $d>2$ and functions $\pi(x_1,\dots,x_{k-1},x_k,x_{k+1},\dots,x_d)$

Functional TT cross

Grouped coordinates: $x_{<k} = (x_1, x_2, \dots, x_{k-1})$, $x_{>k} = (x_{k+1}, \dots, x_{d-1}, x_d)$

Left and right interpolation sets: $\mathcal{I}_{<k} = \{x_{<k}^{(i)}\}_{i=1}^{r}$, $\mathcal{J}_{>k} = \{x_{>k}^{(j)}\}_{j=1}^{r}$



Iterate to next sets:

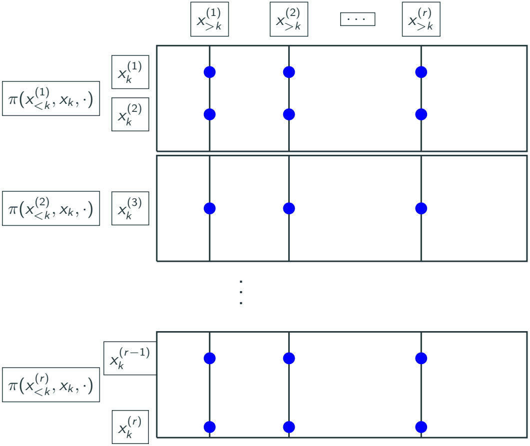
* Discretize x_k: $\pi(\mathcal{I}_{<k}, x_k, \mathcal{J}_{>k})$ for all $x_k \in \mathcal{I}_k$
* $\mathcal{I}_{<k+1} = \texttt{MaxVol} \subset \mathcal{I}_{<k} \otimes \mathcal{I}_k$
* $\mathcal{J}_{>k+1} = \{x_{>k+1}^{(j)}\}_{j=1}^{r}$
* $k \to k+1$, move to the x_{k+1}.
* $k = d$, switch direction and update $\mathcal{J}_{>k}$.
* Stop if converged, repeat otherwise.
* $O(dr^2)$ function evaluations.

TT cross: [Oseledets, Tyrtyshnikov, '10], [Gorodetsky, Karaman, Marzouk, '18]

Empirical interpolation: [Maday, Chaturantabut, Sorensen ...]

Grouped coordinates: $x_{<k} = (x_1, x_2, \dots, x_{k-1})$, $x_{>k} = (x_{k+1}, \dots, x_{d-1}, x_d)$

Left and right interpolation sets: $\mathcal{I}_{<k} = \{x_{<k}^{(i)}\}_{i=1}^{r}$, $\mathcal{J}_{>k} = \{x_{>k}^{(j)}\}_{j=1}^{r}$



Iterate to next sets:

* Discretize x_k: $\pi(\mathcal{I}_{<k}, \mathbf{x_k}, \mathcal{J}_{>k})$ for all $x_k \in \mathcal{I}_k$
* $\mathcal{I}_{<k+1} = \texttt{MaxVol} \subset \mathcal{I}_{<k} \otimes \mathcal{I}_k$
* $\mathcal{J}_{>k+1} = \{x_{>k+1}^{(j)}\}_{j=1}^{r}$
* $k \to k+1$, move to the x_{k+1}.
* $k = d$, switch direction and update $\mathcal{J}_{>k}$.
* Stop if converged, repeat otherwise.
* $O(dr^2)$ function evaluations.

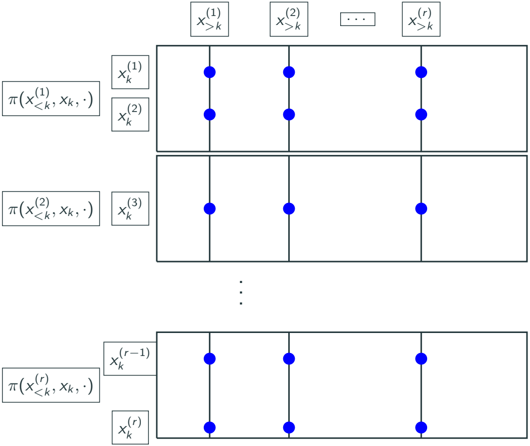
TT cross: [Oseledets, Tyrtyshnikov, '10], [Gorodetsky, Karaman, Marzouk, '18]

Empirical interpolation: [Maday, Chaturantabut, Sorensen ...]

Functional TT cross

Grouped coordinates: $x_{<k} = (x_1, x_2, \dots, x_{k-1})$, $x_{>k} = (x_{k+1}, \dots, x_{d-1}, x_d)$

Left and right interpolation sets: $\mathcal{I}_{<k} = \{x_{<k}^{(i)}\}_{i=1}^r$, $\mathcal{J}_{>k} = \{x_{>k}^{(j)}\}_{j=1}^r$



Iterate to next sets:

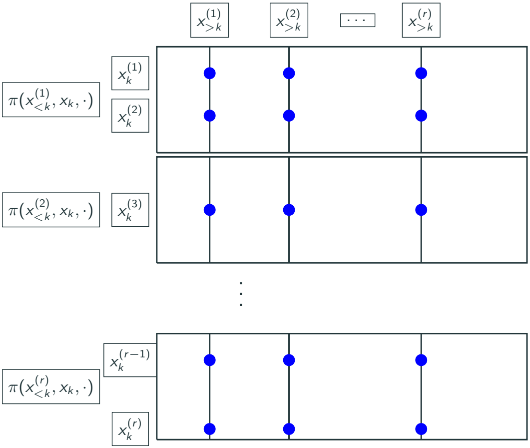
* Discretize x_k: $\pi(\mathcal{I}_{<k}, \mathbf{x_k}, \mathcal{J}_{>k})$ for all $x_k \in \mathcal{I}_k$
* $\mathcal{I}_{<k+1} = \texttt{MaxVol} \subset \mathcal{I}_{<k} \otimes \mathcal{I}_k$
* $\mathcal{J}_{>k+1} = \{x_{>k+1}^{(j)}\}_{j=1}^r$
* $k \to k+1$, move to the x_{k+1}.
* $k = d$, switch direction and update $\mathcal{J}_{>k}$.
* Stop if converged, repeat otherwise.
* $O(dr^2)$ function evaluations.

TT cross: [Oseledets, Tyrtyshnikov, '10], [Gorodetsky, Karaman, Marzouk, '18]

Empirical interpolation: [Maday, Chaturantabut, Sorensen ...]

Functional TT cross

Grouped coordinates: $x_{<k} = (x_1, x_2, \dots, x_{k-1})$, $x_{>k} = (x_{k+1}, \dots, x_{d-1}, x_d)$
Left and right interpolation sets: $\mathcal{I}_{<k} = \{x_{<k}^{(i)}\}_{i=1}^{r}$, $\mathcal{J}_{>k} = \{x_{>k}^{(j)}\}_{j=1}^{r}$



Iterate to next sets:

* Discretize x_k: $\pi(\mathcal{I}_{<k}, \mathbf{x_k}, \mathcal{J}_{>k})$ for all $x_k \in \mathcal{I}_k$
* $\mathcal{I}_{<k+1} = \texttt{MaxVol} \subset \mathcal{I}_{<k} \otimes \mathcal{I}_k$
* $\mathcal{J}_{>k+1} = \{x_{>k+1}^{(j)}\}_{j=1}^{r}$
* $k \to k+1$, move to the x_{k+1}.
* $k = d$, switch direction and update $\mathcal{J}_{>k}$.
* Stop if converged, repeat otherwise.
* $O(dr^2)$ function evaluations.

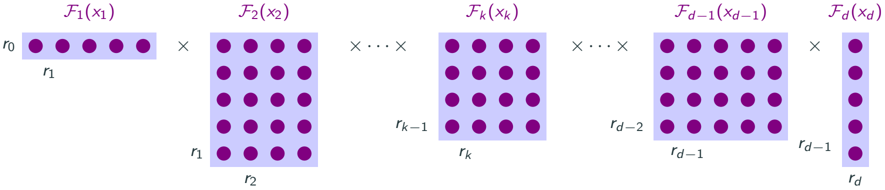
TT cross: [Oseledets, Tyrtyshnikov, '10], [Gorodetsky, Karaman, Marzouk, '18]
Empirical interpolation: [Maday, Chaturantabut, Sorensen ...]

Functional tensor train

TT factorisation $\implies$ separation of variable

$$\pi(x_1,\ldots,x_d) \approx \sum_{\alpha_0,\alpha_1,\cdots,\alpha_d}^{r_0,r_1,\cdots,r_d} \mathcal{F}_1^{(\alpha_0,\alpha_1)}(x_1)\cdots\mathcal{F}_k^{(\alpha_{k-1},\alpha_k)}(x_k)\cdots\mathcal{F}_d^{(\alpha_{d-1},\alpha_d)}(x_d)$$

Each $\mathcal{F}_k : \mathbb{R} \to \mathbb{R}^{r_{k-1}\times r_k}$ is a *matrix-valued univariate* function



Integration of a matrix-valued function $\mathcal{F}_k(x_k) \implies$ integration over x_k

From TT to KR rearrangement

* Given TT $\tilde{\pi} \approx \pi$, we approximate the marginal functions

$$\begin{aligned}\pi_{\leq k}(x_1, \ldots, x_k) &\approx \tilde{\pi}_{\leq k}(x_1, \ldots, x_k) \\ &= \int \tilde{\pi}_(x_1, \ldots, x_k, x_{k+1}, x_d) dx_{k+1} \cdots dx_d \\ &= \mathcal{F}_1(x_1) \cdots \mathcal{F}_k(x_k) \bar{\mathcal{F}}_{k+1} \cdots \bar{\mathcal{F}}_d\end{aligned}$$

where $\bar{\mathcal{F}}_k = \int \mathcal{F}_k(x_k) dx_k \in \mathbb{R}^{r_{k-1} \times r_k}$.

* Approximate Rosenblatt transport [Dolgov et. al., 2020]

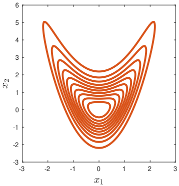
$$\tilde{F}(x) = \begin{bmatrix} \tilde{F}_1(x_1) \\ \vdots \\ \tilde{F}_k(x_k | x_1, \ldots, x_{k-1}) \\ \vdots \\ \tilde{F}_d(x_d | x_1, \ldots, x_{d-1}) \end{bmatrix}$$

* Approximate inverse Rosenblatt transport

$$x = \tilde{F}^{-1}(z) \equiv \left[\tilde{F}_1^{-1}(z_1), \ldots, \tilde{F}_k^{-1}(z_k | x_1, \ldots, x_{k-1}), \ldots, \tilde{F}_d^{-1}(z_d | x_1, \ldots, x_{d-1})\right]^\top$$

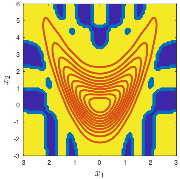
- Non-negative pdfs are not preserved by TT, so $\tilde{T}$ may not be order-preserving

$\pi(x_1, x_2)$



$\approx$

$\sum_{\alpha_1=1}^{r} \mathcal{F}^{(\alpha_1)}(x_1)\mathcal{F}^{(\alpha_1)}(x_2)$



$\not\geq 0$

- Build tensor train $\tilde{g} \approx \sqrt{\pi}$ [Cui, Dolgov, 2022]

$$\tilde{g}(x_1, x_2, \ldots, x_d) = \sum \mathcal{G}_1^{(\alpha_0,\alpha_1)}(x_1) \cdots \mathcal{G}_k^{(\alpha_{k-1},\alpha_k)}(x_k) \cdots \mathcal{G}_d^{(\alpha_{d-1},\alpha_d)}(x_d)$$

- $\pi \approx \tilde{g}^2 \implies$ order-preserving $\tilde{T}_\sharp\, p = \tilde{g}^2$

$$\pi_{\leq k}(x_1, \ldots, x_k) \approx \int \left(\sum \mathcal{G}_1^{(\alpha_0,\alpha_1)}(x_1) \cdots \mathcal{G}_d^{(\alpha_{d-1},\alpha_d)}(x_d)\right)^2 dx_{k+1} \cdots dx_d$$

- $\mathcal{O}(dr^3)$ for marginalisation (via QRF) and $\mathcal{O}(dr^2)$ for evaluation

Error of the KR rearrangement

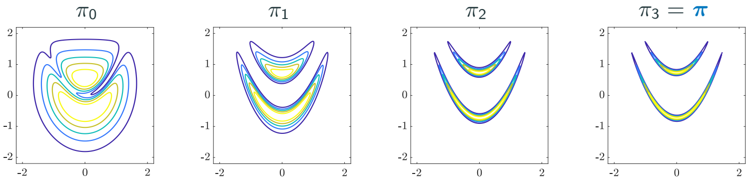
* Target density: $\pi(x) = \frac{1}{\zeta}\exp(-\Phi(x))$ for some potential function $\Phi(x)$ and ζ is unknown
* Build TT $\tilde{g}$ that $\approx \exp(-\frac{1}{2}\Phi(x))$
* Control the Hellinger error by the relative L^2 error [Cui & Dolgov, 2022],

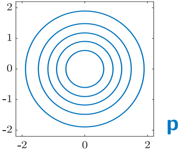
$$\frac{\| \exp(-\frac{1}{2}\Phi(x)) - \tilde{g} \|}{\| \exp(-\frac{1}{2}\Phi(x)) \|} \leq \tau \implies D_{\mathrm{Hell}}(\pi, \tilde{T}_\sharp p) \leq \sqrt{2}\tau$$

* The error rate of $\| \exp(-\frac{1}{2}\Phi(x)) - \tilde{g} \|$ may depend on the smoothness, see [Griebel & Harbrecht, 2023], [Griebel, Harbrecht & Schneider, 2022], ...

Concentrated density: deep IRT (DIRT)

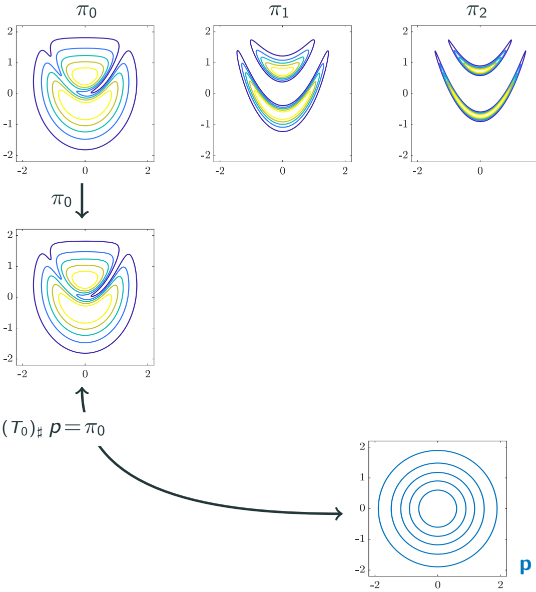
Guided by bridging densities $\pi_k(x) = \pi^{\beta_k}(x)$ with $\beta_0 < \beta_1 < \cdots < \beta_L = 1$.





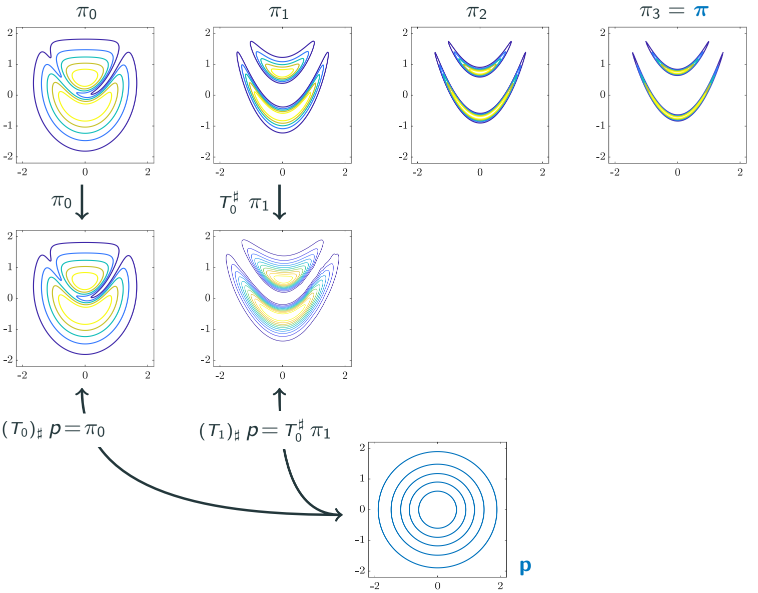
Concentrated density: deep IRT (DIRT)

Guided by bridging densities $\pi_k(x) = \pi^{\beta_k}(x)$ with $\beta_0 < \beta_1 < \cdots < \beta_L = 1$.



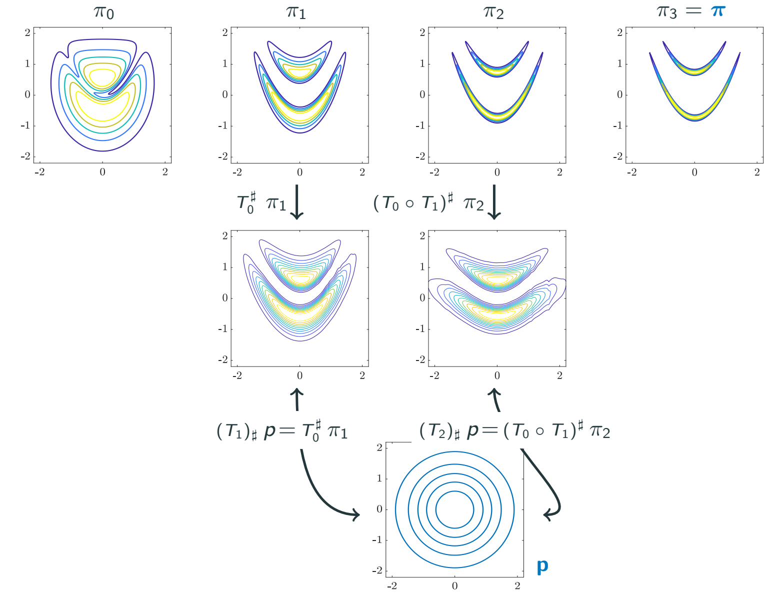
Concentrated density: deep IRT (DIRT)

Guided by bridging densities $\pi_k(x) = \pi^{\beta_k}(x)$ with $\beta_0 < \beta_1 < \cdots < \beta_L = 1$.



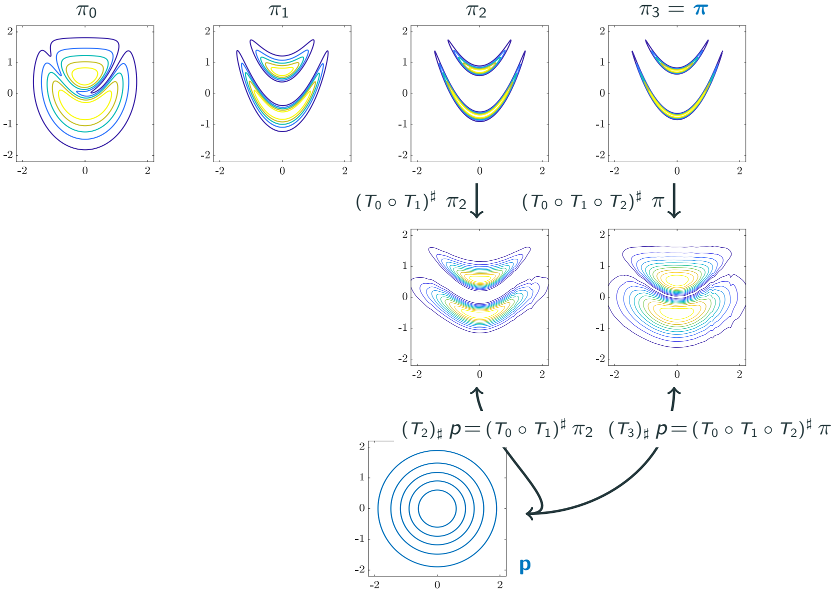
Concentrated density: deep IRT (DIRT)

Guided by bridging densities $\pi_k(x) = \pi^{\beta_k}(x)$ with $\beta_0 < \beta_1 < \cdots < \beta_L = 1$.



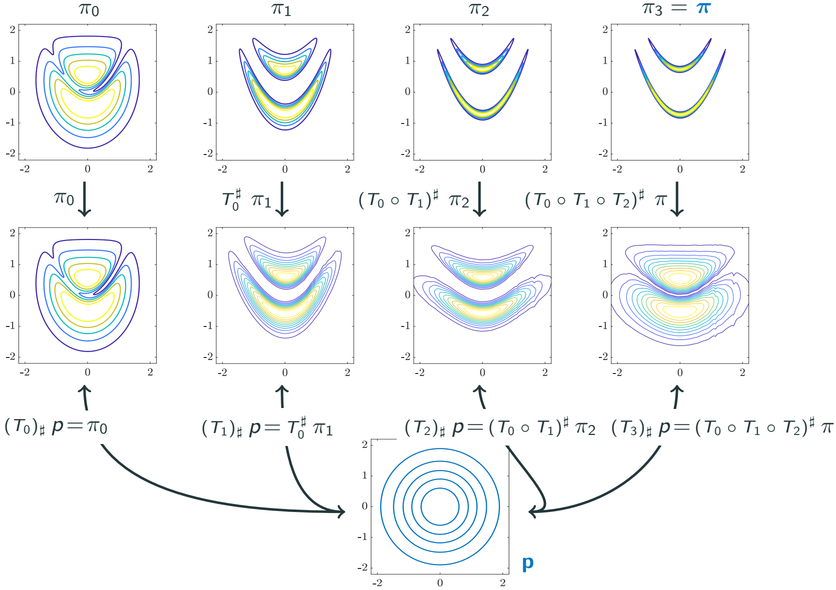
Concentrated density: deep IRT (DIRT)

Guided by bridging densities $\pi_k(x) = \pi^{\beta_k}(x)$ with $\beta_0 < \beta_1 < \cdots < \beta_L = 1$.



Concentrated density: deep IRT (DIRT)

Guided by bridging densities $\pi_k(x) = \pi^{\beta_k}(x)$ with $\beta_0 < \beta_1 < \cdots < \beta_L = 1$.



Stability of DIRT

* Target density: $\pi(x) \propto \exp(-\Phi(x) - \Phi_{\mathrm{ref}}(x))$
* Bridging densities $\pi_k(x) \propto \exp(-\beta_k \Phi(x) - \Phi_{\mathrm{ref}}(x))$
* Approximations
  * Iter 0: find $\tilde{g}_0$ such that $\|\tilde{g}_0 - \sqrt{\pi_0}\|_2 \leq \epsilon_0$
  * Iter j: find $\tilde{g}_j$ such that $\|\tilde{g}_j - \sqrt{r_{j,j-1}} \circ T_0 \cdots \circ T_{j-1}\|_2 \leq \epsilon_j$ where

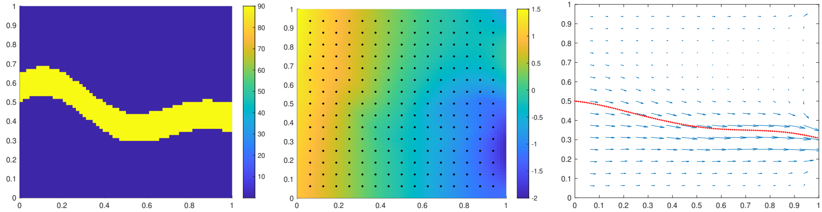
$$r_{k,j} \propto \exp(-(\beta_k - \beta_j)\Phi(x))$$

* Errors in L^2/Hellinger are linearly accumulated [Cui & Dolgov, 2022],

$$D_{\mathrm{Hell}}\big(\pi_k, (T_0 \cdots \circ T_k)_\sharp\, p\big) \leq \sqrt{2}\Big(\sqrt{c_{k,0}}\epsilon_0 + \textstyle\sum_{j=1}^{k} \sqrt{c_{k,j}\zeta_{j-1}}\epsilon_j\Big)$$

where $c_{k,j} = \sup_x r_{k,j}(x)$ and ζ_{j-1} is the normalising constant of π_{j-1}

Example 1: importance sampling for (data-driven) rare events



- $*$ Water table $u(s;x)$ given diffusivity $\ln \kappa(s;x) = \sum_{k=1}^{d} \psi_k(s) x_k$, $d = 20$

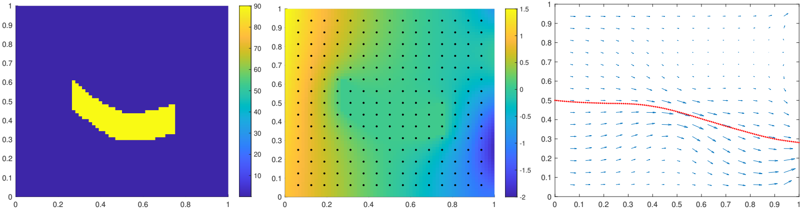
$$-\nabla \cdot (\kappa(s;x)\nabla u(s;x)) = 0, \quad s \in (0,1)^2$$

- $*$ Trajectory of a particle $s(t;x)$ given flow field $\nabla u(s;x)$:

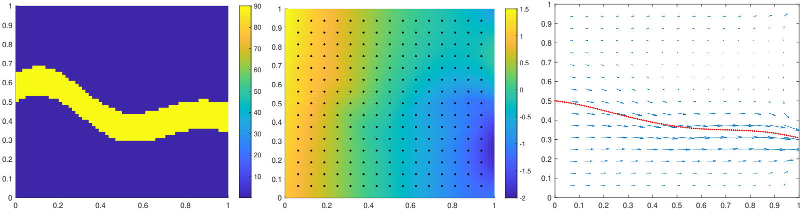
$$\frac{ds}{dt} = \kappa(s;x)\nabla u(s;x), \quad s(0) = [0, 0.4]^\top,$$

- $*$ Breakthrough time $\tau(x)$ s.t. $s_1(\tau;x) = 1$
- $*$ Diffusivity (parametrized by x) is unknown, what is the risk of $\tau(X) \leq \tau^*$
- $*$ Collect partial obaservations of u (black dots) to estimate κ (or $x \in \mathbb{R}^d$)

$x^{(1)}$: without a high-diffusivity channel, breakthrough time $\tau = 0.1886$

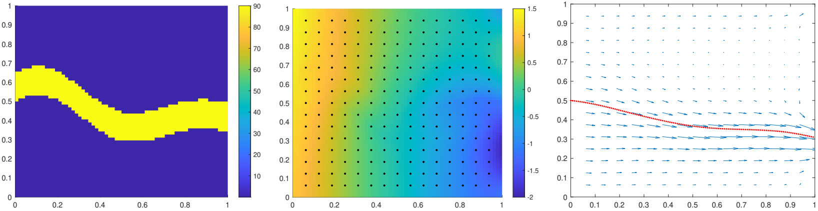


$x^{(2)}$: with a high-diffusivity channel, breakthrough time $\tau = 0.0929$



* left: true diffusivity $\log \kappa$
* middle: water tables u generated by true κ with observation locations (dots)
* right: flow fields (blue) and particle trajectories (red)

Example 1: importance sampling for (data-driven) rare events



Collect data y (at dots) and compute **posterior** failure probability

$$\mathbb{P}_{\pi(x|y)}\left[\tau(X) \le \tau^*\right] = \frac{Q}{Z} = \frac{\int \mathrm{ind}(\tau(x) \le \tau^*)\mathcal{L}(y|F(x))\pi_0(x)dx}{\int \mathcal{L}(y|F(x))\pi_0(x)dx} \quad \leftarrow \text{normalising constant}$$

Using DIRT to approximate optimal importance densities

- $\tilde{p}(x) \approx p^*(x) \propto \mathrm{ind}(\tau(x) \le \tau^*)\mathcal{L}(y|F(x))\pi_0(x)$

$$Q \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\mathrm{ind}(\tau(X^i) \le \tau^*)\mathcal{L}(y|F(X^i))\pi_0(X^i)}{\tilde{p}(X^i)}, \quad X^i \sim \tilde{p}$$

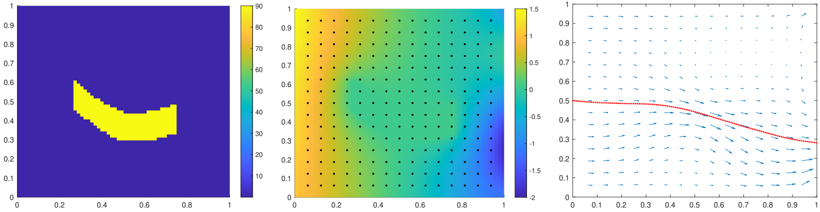
- $\tilde{q}(x) \approx q^*(x) \propto \mathcal{L}(y|F(x))\pi_0(x)$

$$Z \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\mathcal{L}(y|F(X^i))\pi_0(X^i)}{\tilde{q}(X^i)}, \quad X^i \sim \tilde{q}$$

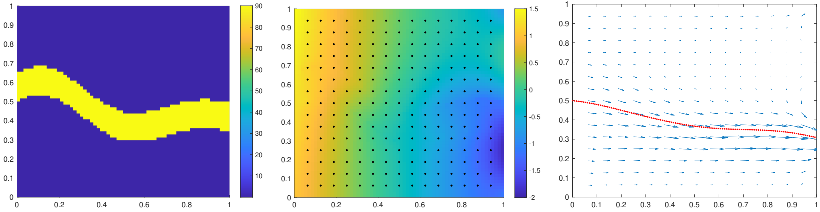
The integrand can be very concentrated for small τ^* (rare event)

Example 1: importance sampling for (data-driven) rare events, $\tau^* = 0.1$

without a high-diffusivity channel: $\mathbb{P}_{\pi(x|y)}\left[\tau(X) \leq \tau^*\right] = 9.4 \times 10^{-4}$

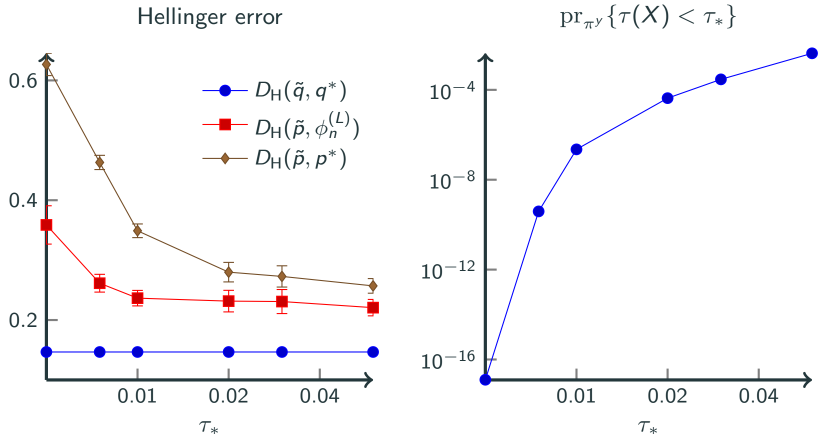


with a high-diffusivity channel: $\mathbb{P}_{\pi(x|y)}\left[\tau(X) \leq \tau^*\right] = 2.8 \times 10^{-2}$



Left: coefficient $\log \kappa$, middle: water table u and centra of observation subdomains D_i (black dots), right: flux (blue) with particle trajectory (red)

Example 1: posterior failure probability $\mathbb{P}_{\pi(x|y)}[\tau(X) \leq \tau^*]$



Changing τ^* with fixed TT rank 7 and $n = 17$ collocation points

References

TT sampling:

1. Cui, Dolgov & Zahm (2023). Conditional deep inverse Rosenblatt transports. arXiv:2106.04170
2. Cui, Dolgov & Scheichl (2022). Deep importance sampling using tensor-trains with application to a priori and a posteriori rare event estimation. arXiv:2209.01941
3. Cui & Dolgov (2022). Deep composition of tensor trains using squared inverse Rosenblatt transports. Found Comput Math
4. Dolgov, Anaya-Izquierdo, Fox & Scheichl (2020). Approximation and sampling of multivariate probability distributions in the tensor train decomposition. Stat Comput

TT complexity:

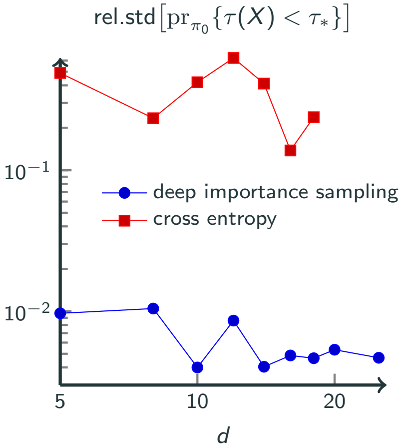
5. Griebel & Harbrecht (2023). Analysis of tensor approximation schemes for continuous functions. Found Comput Math
6. Griebel, Harbrecht & Schneider (2022). Low-rank approximation of continuous functions in Sobolev spaces with dominating mixed smoothness. arXiv:2203.04100
7. Rohrbach, Dolgov, Grasedyck & Scheichl (2022). Rank bounds for approximating gaussian densities in the tensor-train format. SIAM/ASA JUQ

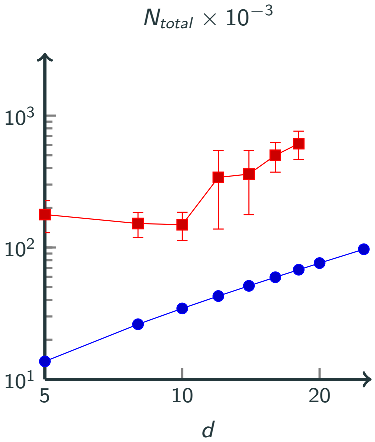
Software:

https://github.com/fastfins/ftt.m

https://github.com/dolgov/TT-IRT

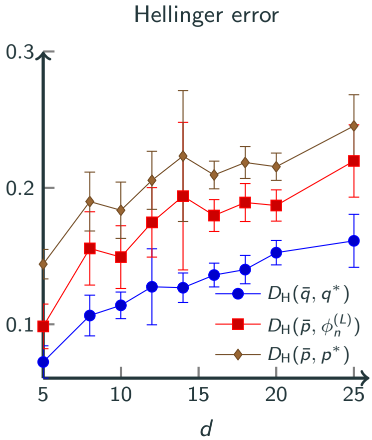
Example 1: prior rare events $\mathbb{P}_{\pi_0(x)}[\tau(X) \le \tau^*]$, $\tau^* = 0.03$

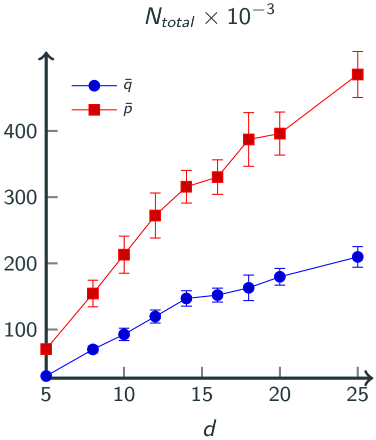




Approximate optimal importance density $p^*(x) \propto \mathrm{ind}(\tau(x) \le \tau^*)\pi_0(x)$ by DIRT

Example 1: posterior rare events $\mathbb{P}_{\pi(x|y)}[\tau(X) \leq \tau^*]$





$n = 17$ collocation points and adative TT rank

Example 2: conditional DIRT

Classical inference learns $X|Y{=}y$ for a realisation $Y = y$

- Need to evaluate $\pi_{X|Y=y}(x|y)$ many times, costly

Offline-online using conditional DIRT [Cui, Dolgov, Zahm, 2023]

* **Offline** phase: learn the law of $X|Y$ for all possible data Y
  - Approximate the joint pdf $\pi_{Y,X}(y,x) = \mathcal{L}(y|F(x))\pi_0(x)$ using DIRT
  - A triangular map: $\pi_{Y,X} \approx T_\sharp\, p_{Y,X}$ with $p_{Y,X} = p_X\, p_X$

$$\begin{bmatrix} y \\ x \end{bmatrix} = T(u_Y, u_X) = \begin{bmatrix} T_Y(u_Y) \\ T_X(u_Y, u_X) \end{bmatrix}$$

  - Still costly—$\mathcal{L}(y|F(x))$—but once in a lifetime
* **Online** phase: given data y, issue pre-computed conditional $X|Y = y$
  - Map the data to the reference space $u_Y = T_Y^{-1}(y)$
  - Define the conditional map $T_{X|Y=y}(\cdot) \equiv T_X(u_Y, \cdot)$

$$\pi_{X|Y=y}(x|y) \approx \left(T_{X|Y=y}\right)_\sharp p_X$$

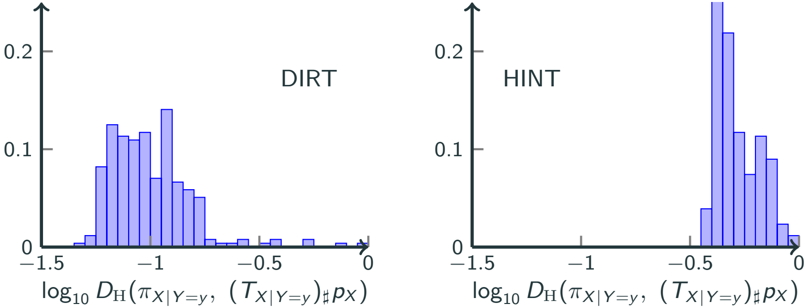
Example 2: susceptible-infected-recovered (SIR)

System of ODEs

$$\frac{dS(t)}{dt} = -\beta S I, \qquad \frac{dI(t)}{dt} = \beta S I - \gamma I, \qquad \frac{dR(t)}{dt} = \gamma I,$$

* started from $S(0) = 99$, $I(0) = 1$, $R(0) = 0$,
* unknown parameters $x = (\beta, \gamma)$
* data y: number of infected individuals at 4 times: $I(1.25), I(2.5), I(3.75), I(5)$

Online performance (posteriors based on multiple data sets):

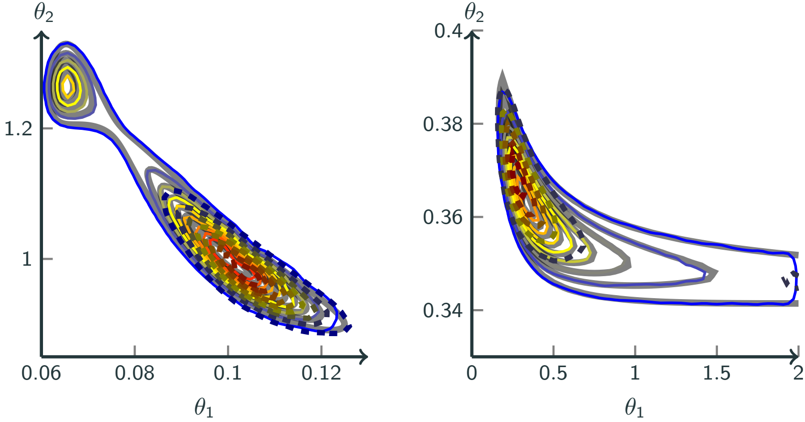


HINT: Hierarchical invertible neural networks.

DIRT (21 secs, MATLAB+Laptop). HINT (28 mins, Tensorflow+GPU).

Example 2: susceptible-infected-recovered (SIR)

Two realizations of data y:

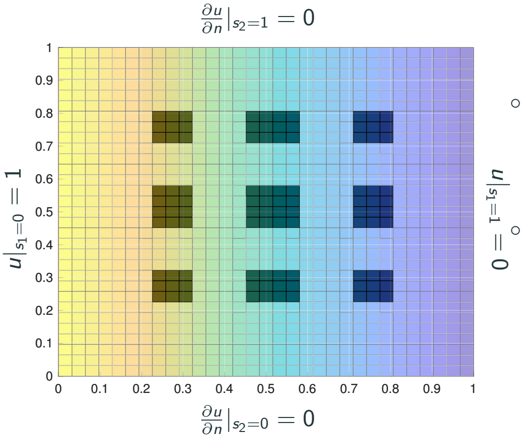


thick grey lines: exact posterior densities
blue solid lines: DIRT
dashed dark lines: HINT

Laplace equation: $-\nabla \cdot (\kappa(s; x) \nabla u(s)) = 0, \quad s \in (0,1)^2$

Unknown diffusivity: $\ln \kappa(s; x) = \sum_{k=1}^{d} \psi_k(s) x_k$, with $d = 11$

Data: average pressure in $m = 3^2$ places with noise $\mathcal{N}(0, \sigma^2)$



- Likelihood

  $$\mathcal{L}(y|x) \propto \exp(-\frac{1}{2\sigma^2}\|y - F(x)\|^2)$$

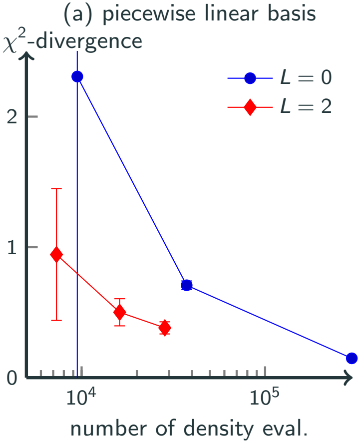
  with $\sigma = 0.1$
- Gaussian prior

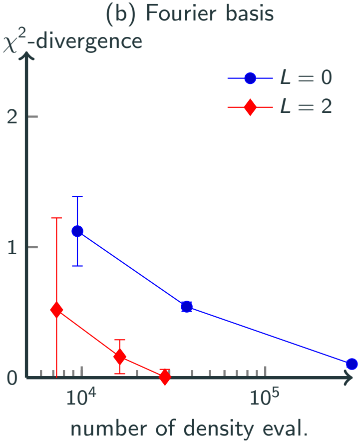
  $$p(x_k) = \mathcal{N}(0, \eta_k)$$

Example 3: accuracy vs. layers

$L = 0$ ($\beta = 1$ and $n = 32$) – single level

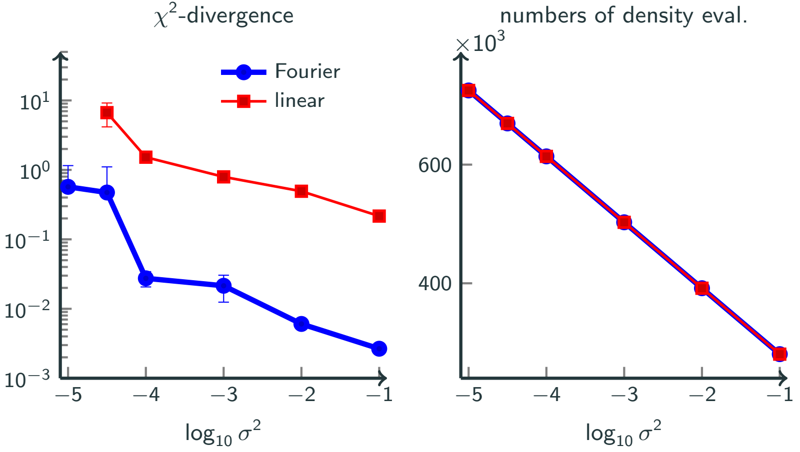
$L = 2$ ($\beta = \{0.1, \sqrt{0.1}, 1\}$ and $n = 16$)





Example 3: concentrated posterior

Varying noise variances σ^2 with $n = 16$ and TT rank 20.



Example 4: predator prey

A dynamical system: the populations of predator (Q) and prey (P) follows

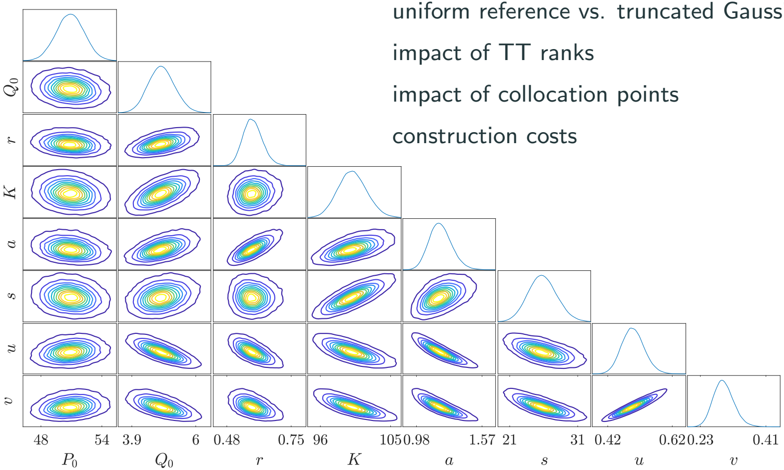
$$\begin{cases} \dfrac{dP}{dt} &= rP\Big(1 - \dfrac{P}{K}\Big) - s\Big(\dfrac{PQ}{a+P}\Big), \\ \dfrac{dQ}{dt} &= u\Big(\dfrac{PQ}{a+P}\Big) - vQ, \end{cases}$$

* Unknown initial conditions $P(t=0) = P_0$ and $Q(t=0) = Q_0$
* Without predator, prey evolves according to the logistic equation characterised by unknown r and K.
* Without prey, predator decreases exponentially with a unknown rate v.
* P and Q have nonlinear interaction modelled by unknowns a, s, and u.
* Estimate

$$x = [P_0, Q_0, r, K, a, s, u, v]^\top$$

using observed populations of predator and prey at time t_i for $i = 1, \ldots, n_T$.

Example 4: predator prey



uniform reference vs. truncated Gaussian

impact of TT ranks

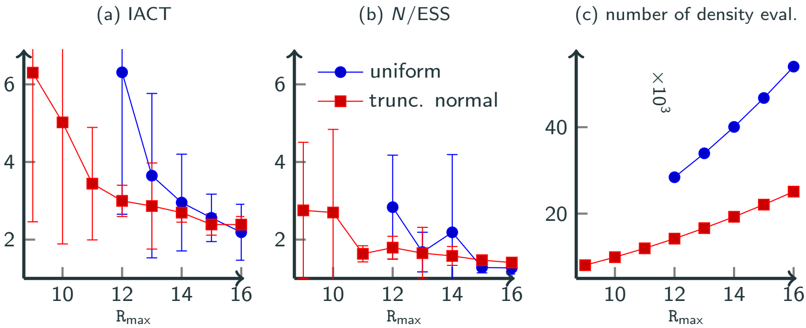
impact of collocation points

construction costs

Plots of marginal posterior densities of the predator-prey model. Estimated using 10^6 posterior samples.

Example 4: predator prey, $L = 8$ DIRT layers

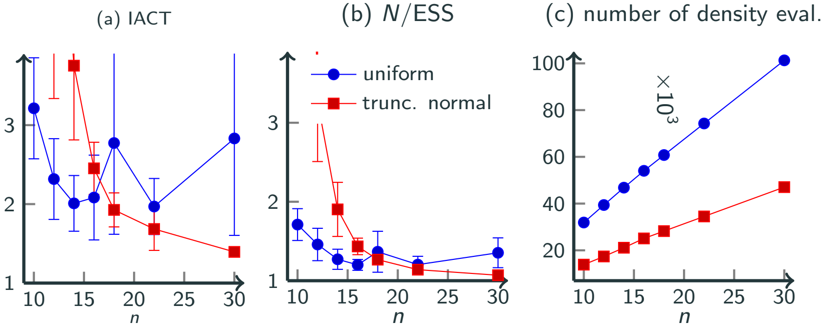
Varying maximum TT ranks and different reference measures.



Estimated from multiple simulations.

Example 4: predator prey

Varying numbers of collocation points n and different reference measures.



Estimated from multiple simulations.

Example 4: predator prey

Relative sampling errors in the estimated covariance from DIRT, DRAM MCMC and Stein Variational Newton (SVN) for different numbers of density evaluations (left) and CPU times (right).