Convergence bounds for nonlinear least squares approximation

Workshop on Optimal Sampling for Approximation

Philipp Trunschke

March 10, 2022

Berlin International Graduate School in Model and Simulation based Research

Overview

Setting

The best approximation in a nonlinear model class is given by

$$u_{\mathcal{M}} \in \operatorname*{arg\,min}_{v \in \mathcal{M}} \|u - v\|_{L^2(Y,\rho)},$$

- where $\mathcal{V} = L^\infty(Y, \rho)$ for a probability measure ρ,
- $u \in \mathcal{V}$ is the function to be approximated,
- and $\mathcal{M} \subseteq \mathcal{V}$ is the (nonlinear) *model class*.

The best approximation in a nonlinear model class is given by

$$u_{\mathcal{M}} \in \arg\min_{v\in\mathcal{M}} \|u - v\|_{L^2(Y,\rho)},$$

- where $\mathcal{V} = L^\infty(Y, \rho)$ for a probability measure ρ,
- $u \in \mathcal{V}$ is the function to be approximated,
- and $\mathcal{M} \subseteq \mathcal{V}$ is the (nonlinear) *model class*.

In general, this problem can only be solved empirically

- Given i.i.d. samples $y_i \sim \rho$ for $i = 1, \dots, n \in \mathbb{N}$, we can estimate $\|u - v\|_{L^2(Y,\rho)}$ by

$$\|u - v\|_n := \left(\frac{1}{n}\sum_{i=1}^{n} |u(y_i) - v(y_i)|^2\right)^{1/2}.$$

- The *empirical best approximation* of u in $\mathcal{M}$ is given by

$$u_{\mathcal{M},n} \in \arg\min_{v\in\mathcal{M}} \|u - v\|_n.$$

$u_{\mathcal{M},n}$ approximates u almost as well as $u_{\mathcal{M}}$

Definition

For any set $A \subseteq \mathcal{V}$ and any $\delta \in (0,1)$ define the *restricted isometry property*

$$\mathrm{RIP}_A(\delta) \quad :\Leftrightarrow \quad \forall u \in A : (1-\delta)\|u\|^2_{L^2(Y,\rho)} \leq \|u\|^2_n \leq (1+\delta)\|u\|^2_{L^2(Y,\rho)}.$$

$u_{\mathcal{M},n}$ approximates u almost as well as $u_{\mathcal{M}}$

Definition

For any set $A \subseteq \mathcal{V}$ and any $\delta \in (0,1)$ define the *restricted isometry property*

$$\mathrm{RIP}_A(\delta) \quad :\Leftrightarrow \quad \forall u \in A : (1-\delta)\|u\|_{L^2(Y,\rho)}^2 \le \|u\|_n^2 \le (1+\delta)\|u\|_{L^2(Y,\rho)}^2 .$$

Theorem (Eigel, Schneider, T – 2021)

If $\mathrm{RIP}_{\{u_{\mathcal{M}}\}-\mathcal{M}\cup\{u\}}(\delta)$ *holds, then*

$$\|u-u_{\mathcal{M}}\|_{L^2(Y,\rho)} \le \|u-u_{\mathcal{M},n}\|_{L^2(Y,\rho)} \le \left(1+2\sqrt{\tfrac{1+\delta}{1-\delta}}\right)\|u-u_{\mathcal{M}}\|_{L^2(Y,\rho)} .$$

$u_{\mathcal{M},n}$ approximates u almost as well as $u_{\mathcal{M}}$

Definition

For any set $A \subseteq \mathcal{V}$ and any $\delta \in (0,1)$ define the *restricted isometry property*

$$\mathrm{RIP}_A(\delta) \quad :\Leftrightarrow \quad \forall u \in A : (1-\delta)\|u\|_{L^2(Y,\rho)}^2 \leq \|u\|_n^2 \leq (1+\delta)\|u\|_{L^2(Y,\rho)}^2.$$

Theorem (Eigel, Schneider, T – 2021)

If $\mathrm{RIP}_{\{u_{\mathcal{M}}\}-\mathcal{M}\cup\{u\}}(\delta)$ *holds, then*

$$\|u-u_{\mathcal{M}}\|_{L^2(Y,\rho)} \leq \|u-u_{\mathcal{M},n}\|_{L^2(Y,\rho)} \leq \left(1+2\sqrt{\tfrac{1+\delta}{1-\delta}}\right)\|u-u_{\mathcal{M}}\|_{L^2(Y,\rho)}.$$

Since $\|\bullet\|_n$ is a random variable, $\mathrm{RIP}_{\{u_{\mathcal{M}}\}-\mathcal{M}\cup\{u\}}(\delta)$ is a random event.

The probability of $\mathrm{RIP}_A(\delta)$ can be bounded by standard concentration inequalities

Definition

For any set $A \subseteq \mathcal{V}$, define the *variation function* $\mathfrak{K}_A(y) := \sup_{a\in A} \frac{|a(y)|^2}{\|a\|_{L^2(Y,\rho)}^2}$.

The probability of $\mathrm{RIP}_A(\delta)$ can be bounded by standard concentration inequalities

Definition

For any set $A \subseteq \mathcal{V}$, define the *variation function* $\mathfrak{K}_A(y) := \sup_{a\in A} \frac{|a(y)|^2}{\|a\|_{L^2(Y,\rho)}^2}$.

Theorem (Eigel, Schneider, T – 2021)

For any set $A \subseteq \mathcal{V}$ with $\dim(\langle A\rangle) < \infty$ and any $\delta \in (0,1)$ there exists C such that

$$\mathbb{P}[\neg\,\mathrm{RIP}_A(\delta)] \le C \exp\left(-\frac{n}{2}\left(\frac{\delta}{\|\mathfrak{K}_A\|_{L^\infty(Y,\rho)}}\right)^2\right).$$

The constant C is independent of n and depends polynomially on δ and $\|\mathfrak{K}_A\|_{L^\infty(Y,\rho)}^{-1}$.

The probability of $\mathrm{RIP}_A(\delta)$ can be bounded by standard concentration inequalities

Definition

For any set $A \subseteq \mathcal{V}$, define the *variation function* $\mathfrak{K}_A(y) := \sup_{a\in A} \frac{|a(y)|^2}{\|a\|^2_{L^2(Y,\rho)}}$.

Theorem (Eigel, Schneider, T – 2021)

For any set $A \subseteq \mathcal{V}$ with $\dim(\langle A\rangle) < \infty$ and any $\delta \in (0,1)$ there exists C such that

$$\mathbb{P}[\neg\,\mathrm{RIP}_A(\delta)] \le C \exp\left(-\frac{n}{2}\left(\frac{\delta}{\|\mathfrak{K}_A\|_{L^\infty(Y,\rho)}}\right)^2\right).$$

The constant C is independent of n and depends polynomially on δ and $\|\mathfrak{K}_A\|^{-1}_{L^\infty(Y,\rho)}$.

Empirical best approximation requires a "small" $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{M}\cup\{u\}}$.

The sample complexity of tensor networks

Approximation by tensor networks

- Tensor networks are multilinear approximations that can break the curse of dimensionality.
- They can be interpreted as a subclass of neural networks.
- But they form manifolds and varieties.
- **They are a popular tool in the numerics of parametric PDEs.**

Approximation by tensor networks may not be feasible

- Tensor networks are multilinear approximations that can break the curse of dimensionality.
- They can be interpreted as a subclass of neural networks.
- But they form manifolds and varieties.
- **They are a popular tool in the numerics of parametric PDEs.**

Theorem (T – 2021)

- *Let* $\mathcal{V} := \mathcal{V}_1 \otimes \cdots \otimes \mathcal{V}_M$ *with* $\dim(\mathcal{V}_m) = d_m$ *for* $m = 1, \ldots, M$.
- *Consider a model class* $\mathcal{M} \subseteq \mathcal{V}$ *of tensor networks with* $\langle \mathcal{M} \rangle = \mathcal{V}$.
- *Then, for all* $u \in \mathcal{V}$,

$$\left\| \mathfrak{K}_{\{u_{\mathcal{M}}\} - \mathcal{M} \cup \{u\}} \right\|_{L^\infty(Y,\rho)} \geq \prod_{m=1}^{M} d_m .$$

Approximation by tensor networks may not be feasible

- Tensor networks are multilinear approximations that can break the curse of dimensionality.
- They can be interpreted as a subclass of neural networks.
- But they form manifolds and varieties.
- **They are a popular tool in the numerics of parametric PDEs.**

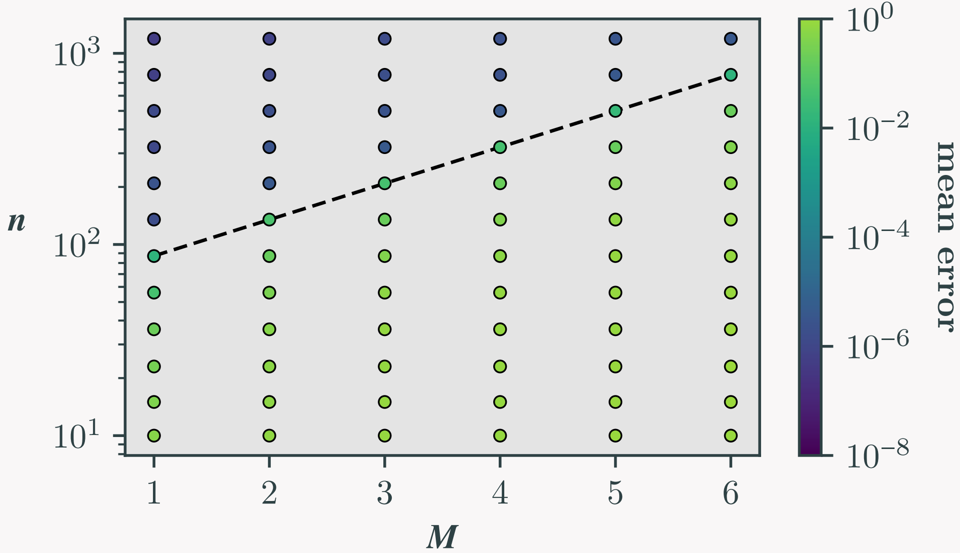
Theorem (T – 2021)

- *Let* $\mathcal{V} := \mathcal{V}_1 \otimes \cdots \otimes \mathcal{V}_M$ *with* $\dim(\mathcal{V}_m) = d_m$ *for* $m = 1, \dots, M$.
- *Consider a model class* $\mathcal{M} \subseteq \mathcal{V}$ *of tensor networks with* $\langle \mathcal{M} \rangle = \mathcal{V}$.
- *Then, for all* $u \in \mathcal{V}$,

$$\left\| \mathfrak{K}_{\{u_{\mathcal{M}}\} - \mathcal{M} \cup \{u\}} \right\|_{L^\infty(Y,\rho)} \geq \prod_{m=1}^{M} d_m .$$

The curse persists with respect to the number of samples.

A phase diagram for rank 1 approximation of $\exp(y_1 + \cdots + y_M)$



But approximation by tensor networks is feasible in practice!

Stationary diffusion

- Consider the random stationary diffusion equation

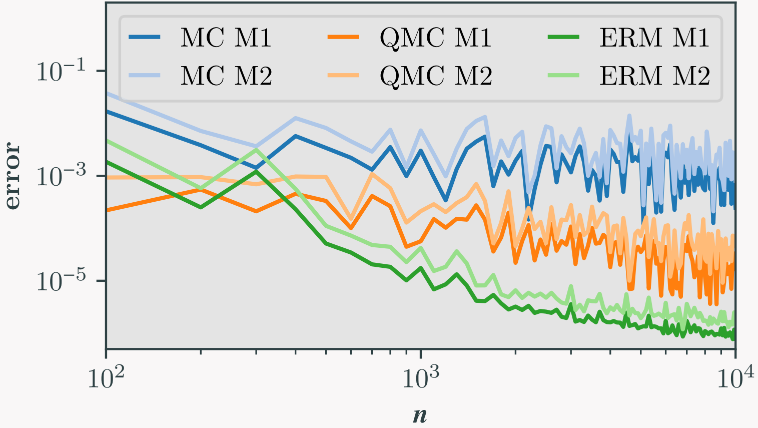
$$\begin{aligned} -\nabla_x \cdot (a(x,y)\nabla_x u(x,y)) &= f(x) \qquad \text{in } D \\ u(x,y) &= 0 \qquad \text{on } \partial D \end{aligned}$$

- $x \in D$ for a bounded Lipschitz domain $D \subseteq \mathbb{R}^d$
- $y \sim \rho$ for a measure ρ on the probability space (Ω, Σ, ρ)

Goal: Approximate u from samples $u(\bullet, y_i)$ with $y_i \sim \rho$.

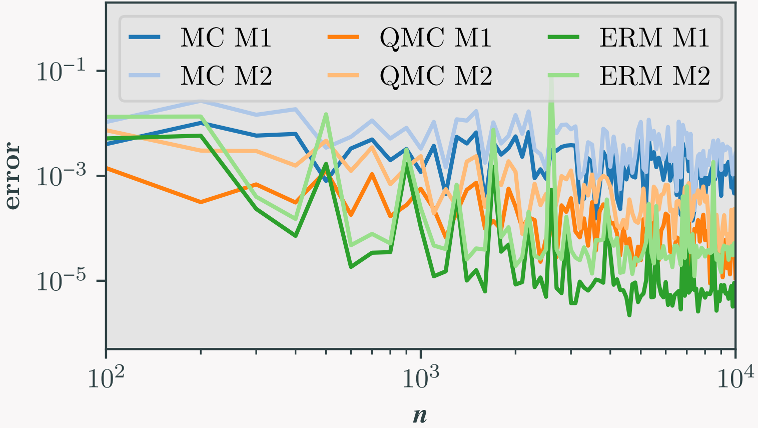
Stationary diffusion: Uniform diffusion coefficient

$$a(x,y) := 1 + \frac{6}{\pi^2}\sum_{m=1}^{20} m^{-2}\sin(\pi\lfloor\tfrac{m}{2}\rfloor x_1)\sin(\pi\lceil\tfrac{m}{2}\rceil x_2)y_m \qquad \text{and} \qquad y \sim \mathcal{U}([-1,1])^{\otimes 20}$$



Stationary diffusion: Log-normal diffusion coefficient

$$a(x, y) := \exp\left(\frac{6}{\pi^2}\sum_{m=1}^{20} m^{-2}\sin(\pi\lfloor\tfrac{m}{2}\rfloor x_1)\sin(\pi\lceil\tfrac{m}{2}\rceil x_2)y_m\right) \quad \text{and} \quad y \sim \mathcal{N}(0,1)^{\otimes 20}$$



The local sample complexity

The variation function may be small in the neighborhood $\mathcal{N}$ of $u_{\mathcal{M}}$ in $\mathcal{M}$

Consider $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}$ instead of $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}\cup\{u\}} = \max\{\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}, \mathfrak{K}_{\{u_{\mathcal{M}}-u\}}\}$.

The variation function may be small in the neighborhood $\mathcal{N}$ of $u_{\mathcal{M}}$ in $\mathcal{M}$

Consider $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}$ instead of $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}\cup\{u\}} = \max\{\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}, \mathfrak{K}_{\{u_{\mathcal{M}}-u\}}\}$.

Proposition (T – 2021)

- $\mathfrak{K}_\bullet$ *is continuous.*
- $\mathfrak{K}_\bullet$ *is monotonic, i.e.* $A \subseteq B$ *implies* $\mathfrak{K}_A \leq \mathfrak{K}_B$.

The variation function may be small in the neighborhood $\mathcal{N}$ of $u_{\mathcal{M}}$ in $\mathcal{M}$

Consider $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}$ instead of $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}\cup\{u\}} = \max\{\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}, \mathfrak{K}_{\{u_{\mathcal{M}}-u\}}\}$.

Proposition (T – 2021)

- *$\mathfrak{K}_\bullet$ is continuous.*
- *$\mathfrak{K}_\bullet$ is monotonic, i.e. $A \subseteq B$ implies $\mathfrak{K}_A \leq \mathfrak{K}_B$.*

Definition

Define the *local variation function* $\mathfrak{K}^{\mathrm{loc}}_{\mathcal{M},u_{\mathcal{M}}} := \lim_{\mathrm{diam}(\mathcal{N})\to 0} \mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}$.

The variation function may be small in the neighborhood $\mathcal{N}$ of $u_{\mathcal{M}}$ in $\mathcal{M}$

Consider $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}$ instead of $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}\cup\{u\}} = \max\{\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}, \mathfrak{K}_{\{u_{\mathcal{M}}-u\}}\}$.

Proposition (T – 2021)

- $\mathfrak{K}_{\bullet}$ *is continuous.*
- $\mathfrak{K}_{\bullet}$ *is monotonic, i.e. $A \subseteq B$ implies $\mathfrak{K}_A \leq \mathfrak{K}_B$.*

Definition

Define the *local variation function* $\mathfrak{K}^{\mathrm{loc}}_{\mathcal{M},u_{\mathcal{M}}} := \lim_{\mathrm{diam}(\mathcal{N})\to 0} \mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}$.

- Monotonicity implies $\mathfrak{K}^{\mathrm{loc}}_{\mathcal{M},u_{\mathcal{M}}} \leq \mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}} \leq \mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{M}}$.
- Continuity implies $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}} \approx \mathfrak{K}^{\mathrm{loc}}_{\mathcal{M},u_{\mathcal{M}}}$ if $\mathrm{diam}(\mathcal{N})$ is small.

The variation function may be small in the neighborhood $\mathcal{N}$ of $u_{\mathcal{M}}$ in $\mathcal{M}$

Consider $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}$ instead of $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}\cup\{u\}} = \max\{\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}, \mathfrak{K}_{\{u_{\mathcal{M}}-u\}}\}$.

Proposition (T – 2021)

- $\mathfrak{K}_{\bullet}$ *is continuous.*
- $\mathfrak{K}_{\bullet}$ *is monotonic, i.e. $A \subseteq B$ implies $\mathfrak{K}_A \leq \mathfrak{K}_B$.*

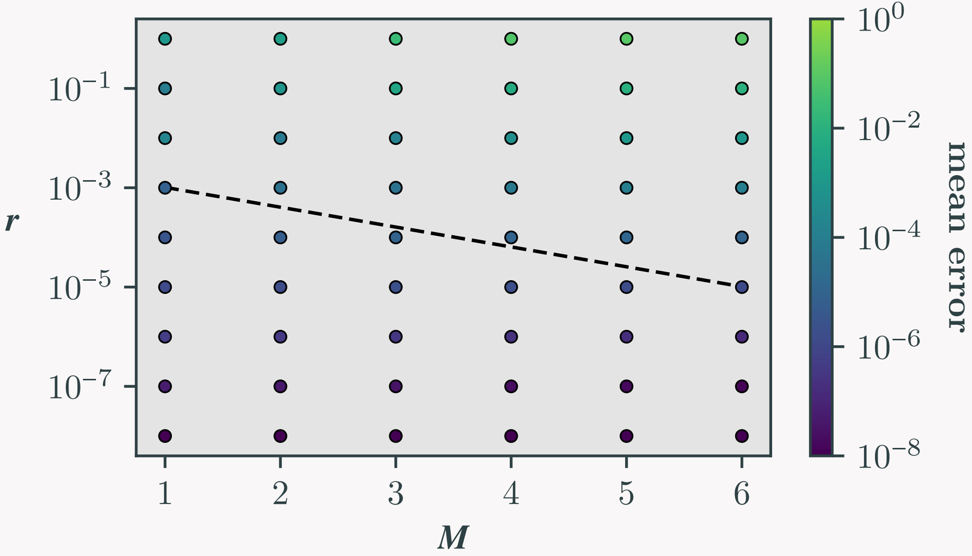
Definition

Define the *local variation function* $\mathfrak{K}^{\mathrm{loc}}_{\mathcal{M},u_{\mathcal{M}}} := \lim_{\mathrm{diam}(\mathcal{N})\to 0} \mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}$.

- Monotonicity implies $\mathfrak{K}^{\mathrm{loc}}_{\mathcal{M},u_{\mathcal{M}}} \leq \mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}} \leq \mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{M}}$.
- Continuity implies $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}} \approx \mathfrak{K}^{\mathrm{loc}}_{\mathcal{M},u_{\mathcal{M}}}$ if $\mathrm{diam}(\mathcal{N})$ is small.

$\mathfrak{K}^{\mathrm{loc}}_{\mathcal{M},u_{\mathcal{M}}}$ provides a tight lower bound for $\mathfrak{K}_{\{u_{\mathcal{M}}\}-\mathcal{N}}$.

Another phase diagram for rank 1 approximation of $\exp(y_1 + \cdots + y_M)$



The local variation function can be computed analytically

Definition

$\mathcal{M}$ is *locally linearizable* in $u_{\mathcal{M}} \in \mathcal{M}$ if there exists a neighborhood $\mathcal{N}$ of $u_{\mathcal{M}}$ in $\mathcal{M}$ such that $\mathcal{N}$ is an embedded, connected C^2-manifold with positive reach.
Then $\mathbb{T}_{u_{\mathcal{M}}}\mathcal{M}$ denotes the *tangent space* of $\mathcal{M}$ in $u_{\mathcal{M}}$.

The local variation function can be computed analytically

Definition

$\mathcal{M}$ is *locally linearizable* in $u_{\mathcal{M}} \in \mathcal{M}$ if there exists a neighborhood $\mathcal{N}$ of $u_{\mathcal{M}}$ in $\mathcal{M}$ such that $\mathcal{N}$ is an embedded, connected C^2-manifold with positive reach.
Then $\mathbb{T}_{u_{\mathcal{M}}}\mathcal{M}$ denotes the *tangent space* of $\mathcal{M}$ in $u_{\mathcal{M}}$.

Theorem (T – 2021)

If $\mathcal{M}$ is locally linearizable in $u_{\mathcal{M}} \in \mathcal{M}$, then $\mathfrak{K}^{\mathrm{loc}}_{\mathcal{M},u_{\mathcal{M}}} = \mathfrak{K}_{\mathbb{T}_{u_{\mathcal{M}}}\mathcal{M}}$.

The local variation function can be computed analytically

Definition

$\mathcal{M}$ is *locally linearizable* in $u_{\mathcal{M}} \in \mathcal{M}$ if there exists a neighborhood $\mathcal{N}$ of $u_{\mathcal{M}}$ in $\mathcal{M}$ such that $\mathcal{N}$ is an embedded, connected C^2-manifold with positive reach.
Then $\mathbb{T}_{u_{\mathcal{M}}}\mathcal{M}$ denotes the *tangent space* of $\mathcal{M}$ in $u_{\mathcal{M}}$.

Theorem (T – 2021)

If $\mathcal{M}$ is locally linearizable in $u_{\mathcal{M}} \in \mathcal{M}$, then $\mathfrak{K}^{\mathrm{loc}}_{\mathcal{M},u_{\mathcal{M}}} = \mathfrak{K}_{\mathbb{T}_{u_{\mathcal{M}}}\mathcal{M}}$.

$\mathfrak{K}_{\mathbb{T}_{u_{\mathcal{M}}}\mathcal{M}}$ grows exponentially for example ??.
But it is small, for example, if $u_{\mathcal{M}}$ is a low degree polynomial.

Empirical approximation requires a small variation function

- This may be satisfied in a neighborhood $\mathcal{N}$ of $u_{\mathcal{M}}$.
- And this provides a heuristic argument for the success of state-of-the-art algorithms.
- But we have also seen counterexamples.
- A low variation function can not be guaranteed in all practical applications.

→ **Algorithms should enforce a small variation function.**

- For approximation by tensor train networks this is realized in the *restricted alternating least squares* (RALS) algorithm.

Numerical experiments

Stationary diffusion

- Consider the random stationary diffusion equation

$$
\begin{aligned}
-\nabla_x \cdot (a(x,y)\nabla_x u(x,y)) &= f(x) \qquad \text{in } D \\
u(x,y) &= 0 \qquad \text{on } \partial D
\end{aligned}
$$

- $x \in D$ for a bounded Lipschitz domain $D \subseteq \mathbb{R}^d$
- $y \sim \rho$ for a measure ρ on the probability space (Ω, Σ, ρ)

Goal: Approximate $M(y) = \int_D u(x,y)\,\mathrm{d}x$ from samples $M(y_i)$ with $y_i \sim \rho$.

Stationary diffusion: Uniform diffusion coefficient

$$a(x, y) := 1 + \frac{6}{\pi^2} \sum_{m=1}^{20} m^{-2} \sin(\pi \lfloor \tfrac{m}{2} \rfloor x_1) \sin(\pi \lceil \tfrac{m}{2} \rceil x_2) y_m \qquad \text{and} \qquad y \sim \mathcal{U}([-1, 1])^{\otimes 20}$$

	$n = 9000$	$n = 1000$	$n = 500$	$n = 100$	$n = 45$
RALS	$1.13 \cdot 10^{-5}$	$5.88 \cdot 10^{-5}$	$2.52 \cdot 10^{-4}$	$9.73 \cdot 10^{-4}$	$1.35 \cdot 10^{-3}$
hard thresholding	$4.23 \cdot 10^{-5}$	$1.97 \cdot 10^{-4}$	$6.17 \cdot 10^{-4}$	$9.73 \cdot 10^{-3}$	$2.92 \cdot 10^{-2}$
SALSA	$8.24 \cdot 10^{-5}$	$4.49 \cdot 10^{-4}$	$1.46 \cdot 10^{-2}$	$4.89 \cdot 10^{-1}$	$4.91 \cdot 10^{-1}$
ALS + ℓ^2-regularization	$4.74 \cdot 10^{-4}$	$7.15 \cdot 10^{-4}$	$8.25 \cdot 10^{-3}$	$9.86 \cdot 10^{-1}$	$7.06 \cdot 10^{-1}$

Stationary diffusion: Log-normal diffusion coefficient

$$a(x, y) := \exp\left(\frac{6}{\pi^2}\sum_{m=1}^{20} m^{-2}\sin(\pi\lfloor\tfrac{m}{2}\rfloor x_1)\sin(\pi\lceil\tfrac{m}{2}\rceil x_2)y_m\right) \qquad \text{and} \qquad y \sim \mathcal{N}(0,1)^{\otimes 20}$$

	$n = 9000$	$n = 1000$	$n = 500$	$n = 100$	$n = 45$
RALS	$1.13 \cdot 10^{-5}$	$5.88 \cdot 10^{-5}$	$2.52 \cdot 10^{-4}$	$9.73 \cdot 10^{-4}$	$1.35 \cdot 10^{-3}$
hard thresholding	$4.23 \cdot 10^{-5}$	$1.97 \cdot 10^{-4}$	$6.17 \cdot 10^{-4}$	$9.73 \cdot 10^{-3}$	$2.92 \cdot 10^{-2}$
SALSA	$8.24 \cdot 10^{-5}$	$4.49 \cdot 10^{-4}$	$1.46 \cdot 10^{-2}$	$4.89 \cdot 10^{-1}$	$4.91 \cdot 10^{-1}$
ALS + ℓ^2-regularization	$4.74 \cdot 10^{-4}$	$7.15 \cdot 10^{-4}$	$8.25 \cdot 10^{-3}$	$9.86 \cdot 10^{-1}$	$7.06 \cdot 10^{-1}$

Algorithms should enforce a small variation function!