Variational Monte Carlo - Unifying Approach: High Dimensional PDE's and Statistical Learning

Reinhold Schneider
TU Berlin
joint work with M. Eigel, P. Trunschke (WIAS Berlin) and S. Wolf (TU Berlin)

DFG Research Center MATHEON
Mathematics for key technologies

CRC 1114

My Motiviation

Dealing with Hierarchical (Tucker) Tensors:
low rank factorization

$$u(x,\xi) \approx \sum_{k=1}^{r} c_k(\xi)\varphi_k(x) = \Phi^T(x)C(\xi)$$

Problem: let $F : \mathbb{R}. \to \mathbb{R}$ a non-linear function, compute

$$F \circ u(x,\xi) \approx \sum_{k=1}^{R} d_k(\xi)\psi_k(x)\ ,\ \text{BUT } R >> r \text{ Too large!}$$

- $F(u) := \max\{0, u\}$ (ReLu) - optimal control with uncertainty - CVaR etc. (with M. Eigel et al.)
- $F(u) := e^{-u}$ - Inverse Bayesian - (with M. Eigel et al.)

Idea keep: $F \circ u$, $\Rightarrow$
sampling: given ξ, $F(u(\xi)$ and $\partial F \circ u(\xi)$ are computable.

Variational Formulation - Optimization Problems

Problem (Generic optimization problem (OP))

Given a target functional $\mathcal{J} : \mathcal{V} \to \mathbb{R}$

$$\textit{argmin } \{\mathcal{J}(W) : W \in \mathcal{V}\} \ .$$

Given further a family of (compact) model classes (specified later)

$$\mathcal{M}_h \subset \mathcal{V} \cap C_b^0(\mathbb{R}^d) \ , \ 0 < h < h_0 \ , \ \mathcal{M}_h \subset\subset \mathcal{V} \ \textit{(compac)t}$$

$$\textit{argmin } \{\mathcal{J}(W) : W \in \mathcal{V} \cap \mathcal{M}_h\} \ .$$

Example (Ritz Galerkin scheme)

Finite dimensional (feature) space
$\mathcal{M}_h = B_{R,0}|_{\mathcal{V}_h} \ , \ \mathcal{V}_h := \text{span}\{\varphi_k : k = 0, \ldots, n\} \subset \mathcal{V} \ , \ R, n \sim h^{-1}$

$$U := \text{argmin } \{\mathcal{J}(W) : W \in \mathcal{M}_h\} \qquad (1)$$

Examples

Approximation: for given $U \in \mathcal{H}$ minimize

$$\boxed{\mathcal{J}(W) = \|U - W\|^2 \ , \ W \in \mathcal{M}}$$

solving equations: where $A, g : \mathcal{V} \to \mathcal{H}$,

$$AU = B \quad or \quad g(U) = 0$$

here

$$\boxed{\mathcal{J}(W) := \|AW - B\|^2_{\mathcal{H}} \ \text{ resp. } \quad F(W) := \|g(W)\|^2_{\mathcal{H}} \ .}$$

or, if $A : \mathcal{V} \to \mathcal{V}'$ is symmetric and $B \in \mathcal{V}'$, $\mathcal{V} \subset \mathcal{H} \subset \mathcal{V}'$,

$$\boxed{\mathcal{J}(W) := \frac{1}{2}\langle AW, W\rangle - \langle B, W\rangle}$$

computing the lowest eigenvalue of a symmetric operator $A : \mathcal{V} \to \mathcal{V}'$,

$$\boxed{U = \text{argmin } \{\mathcal{J}(W) = \langle AW, W\rangle : \langle W, W\rangle = 1\} \ .}$$

constraint minimization - extended approach

Low rank model classes

- linear parametrization

$$\mathcal{M}_n := \mathrm{span}\{\varphi_1, \dots, \varphi_n\}$$

- Reduced basis ROM - model class

$$\mathcal{M}_{\leq r} := \{(x,\xi) \mapsto u(x,\xi) : \sum_{k=1}^{s} c_k(\xi)\varphi_k(x) = \Phi^T(x) \cdot C(\xi)\,,\ s \leq r\}$$

  is non-linear, bi -linear parametrization, manifold - *Absil et al.*
- Low rank tensors are given by multi -linear parametrizations
- Hierarchical (Tucker) tensors (HT) $\mathcal{M}_{\leq \mathbf{r}}$ are given by multi -linear parametrizations, which are compositions of bi-linear parametrizations
- HT are multi-layer layer generalization of RB-ROM
- HT are deep ROM: ROM - shallow ANN (artificial neuronal network) , HT - DNN (deep neuronal network)

HT - TT: *Hackbusch, Oseledets, Grasedyck, Nouy, Kressner, Bachmayr, Urban, ...*

Risk Minimization - Regression -

Given samples $(y^i, \xi^i) \in Y \times X$, find $\xi \mapsto f(\xi) \simeq y$.

Risk functional

$$\mathcal{J}(f) := \mathbb{E}[\ell(y, f(\xi))] = \int_{Y \times X} \ell(y, f(\xi)) d\mathbb{P}(y, \xi)$$

with loss function $\ell(y, f(\xi))$, $y \in Y$, $f(\xi) \in Y$.
Example: : regression $\ell : Y \times Y$, e.g. $Y = \mathbb{R}$,

$$\ell(f(\xi)) := \ell(y, f(\xi)) := \|y - f(\xi)\|^2 \ ,$$

Empirical risk - surrogate functional

$$\mathcal{J}_N(f) := \frac{1}{N} \sum_{i=1}^{N} \ell(y^i, f(\xi^i)) := \frac{1}{N} \sum_{i=1}^{N} \|y^i - f(\xi^i)\|^2$$

Model (hypothesis)) classes e.g. $\mathcal{M}_h = \mathcal{M} \cap B_R \subset\subset \mathcal{V}$, solve

$$\Psi_N = \ \text{argmin}_{\Phi \in \mathcal{M}_h} \mathcal{J}_N(\Phi)$$

compare work of Cohen & Leviatan and poster Haberstich, Nouy etc.

Least Squares Minimization

$$Au - f = 0. \ , \quad \text{e.g.} \quad A(\xi)u(\xi) - f(\xi) = 0$$

Objective functional

$$\mathcal{J}(u) := \int_X \ell(u(\xi))d\xi$$

with loss function $\ell(u(\xi))$, .

$$\ell(u(\xi)) := \|f(\xi) - Au(\xi)\|^2 = \langle Au(\xi) - f(\xi^i), Au(\xi) - f(\xi^i)\rangle \ ,$$

Empirical objective functional

$$\mathcal{J}_N(u) := \frac{1}{N}\sum_{i=1}^{N} \ell(u(x^i)) := \frac{1}{N}\sum_{i=1}^{N} \|f(\xi^i) - Au(\xi^i)\|^2$$

E.g. Restricting to low rank hierarchical tensors $\mathcal{M}_{\leq \mathbf{r}} \cap B_R(0)$
tensor collocation

Parametric Elliptic operators - Uncertainty quantification

$$A(\xi)u(\xi) - f = 0 \ , \ A(\xi) : V_h \to V_h' \ , \ \xi \in \mathbb{R}^d$$

Objective functional alternatively $u(\xi) - A^{-1}(\xi)f = 0$

$$\mathcal{J}(u) := \int_X \ell(u(\xi))d\xi = \int_X \langle u(\xi) - A^{-1}(\xi)f, u(\xi) - A^{-1}(\xi)f \rangle_{V_h} \rho(\xi)d\xi$$

with loss function $u(\xi) \mapsto \ell(u(\xi))$, .

$$\ell(u(\xi)) = \langle u(\xi) - A^{-1}(\xi^i)f, u(\xi) - A^{-1}(\xi^i)f \rangle \ ,$$

Empirical objective functional

$$\mathcal{J}_N(u) := \frac{1}{N}\sum_{i=1}^{N} \ell(u(\xi^i)) := \frac{1}{N}\sum_{i=1}^{N} \langle u(\xi) - A^{-1}(\xi^i)f, u(\xi) - A^{-1}(\xi^i)f \rangle$$

Restricting to model classes, e.g. low rank hierarchical tensors $\mathcal{M}_{\mathbf{r}} \cap B_R(0)$ *tensor regression* (S. Wolf et al.)

Semi -implicit time steps for parabolic operators - SKIP

e.g. Fokker Planck and/or Backward Kolmogorov equation

$$\partial_t u = -Au + F(u) \ , \ (I + h_k A)u_{k+1} = u_k + h_h F(u_k)$$

Objective functional

$$u_{k+1} = \text{argmin}_{u \in \mathcal{V}} \mathcal{J}(u) \ , \ \mathcal{J}(u) := \int_X \ell(u(\xi)) d\xi$$

with loss function $\ell(u(\xi))$, $u(\xi) \in \mathbb{R}$.
$\ell : X \to Y$, e.g. $Y = \mathbb{R}$,

$$\ell(u(\xi)) = \|u_k + h_k F(u_k(\xi)) - (I + h_k A)u(\xi)\|^2 \ ,$$

Empirical objective functional

$$\mathcal{J}_N(u) := \frac{1}{N} \sum_{i=1}^{N} \ell(u(x^i)) := \frac{1}{N} \sum_{i=1}^{N} \|u_k + h_k F(u_k(\xi^i)) - (I + h_k A)u(\xi^i)\|^2$$

Model (hypothesis)) classes e.g. $\mathcal{M}_h \cap B_R$ Restricting to to model classes, e.g. low rank hierarchical tensors $\mathcal{M}_{\mathbf{r}} \cap B_R(0)$

Multi-Class Classificatio - SKIP

$Y = \{1, \ldots, K\}$, $\xi \in \mathbb{R}^d$, Given samples (y^i, ξ^i), $i = 1, \ldots, N$. we estimate either the conditional probability density $h^i(\xi) = p_{y^i}(\xi) = d\mathbb{P}(y_i|\xi) \in \Delta Y$, or probability $h^i(\xi) = p_{y^i}(\xi) = d\mathbb{P}(y_i|\xi)\rho(\xi) \in \Delta Y$, $\xi \in \mathcal{X}$.

$$y = (y_k)_{k\in K}, p(\xi) \in \Delta Y \simeq \mathbb{R}^K_+ \quad , \quad \sum_{k=1}^{K} p_{y_k}(\xi) = 1 \quad , \quad .$$

Predictor $pred : X \to Y$

$$pred(\xi) := \text{argmax}_{y\in Y} p_y(\xi)$$

softmax function $\sigma : \mathbb{R}^K \to \Delta Y$, $\eta - k \in \mathbb{R}$,

$$\sigma_k(\eta) := \frac{e^{\eta_k}}{\sum_{j=1}^{K} e^{\eta_j}} \quad , \quad , \; k = 1, \ldots, K \; .$$

$$\begin{aligned} & \mathcal{M}_h \\ = \; & \{p_k = \sigma \circ h : h \in L_2(\mathbb{R}^d, \rho) | \|h\|_{L_2(\mathbb{R}^d,\rho)} \le R, k = 1, \ldots, K\} \end{aligned}$$

Multi-class Classification - SKIP

Kullback-Leibler-divergence as the distance between two probability densities, the true one p versus the estimated one q,

and (conditional) *cross entropy*. (Conditional) probability densities p, q

$$D(p,q) := \sum_{y \in Y} \int p_y(x) \log \frac{p_y(x)}{q_y(x)} d\mathbb{P}(x) \ ,$$

$$H(p,q) = -\sum_{y} \int p_y(x) \log q_y(x) d\mathbb{P}(x) = -\mathbb{E}(\log q_y(x))$$

$$\text{argmin}_{\ q \in L_{1,+}} D(p,q) = \text{argmin}_{\ q \in L_{1,+}} H(p,q)$$

(here $H(p,p) = 0$), $q =: \sigma \circ h$, empirical risk and loss function

$$\mathcal{R}_N(q) = \frac{1}{N} \sum_{i=1}^{N} \ell(y^i, q(x^i)) = \frac{-1}{N} \sum_{i=1}^{N} \log(\sigma_{y^i}(h(x^i)))$$

$$\ell(y, q_y(x)) := -\delta_{y,y^i} \log(\sigma_y(h(x))) \ , \ \ q \in \mathcal{M}_0$$

Remarks

- The model class is considered to be compact and ℓ continuous $\Rightarrow$ the empirical functional has a minimizer
- empirical functionals needs only point values $u(\xi)$.
- for optimization we need to compute (sub-) gradients at points ξ.
- no problems with variable coefficients and/or nonlinearities
- previous examples sketch the idea, they can be modified

Comparison: Numerics - Statistics

- Statistics: samples are given - data are noisy
- assumptions: samples are distributed according an underlying but not known density
- Numerics: samples can be produced - data have negligible variance
- density is known or to our disposal,

Empirical Objective Minimization - General Theory

Chervonenkis-Vapnik theory (1971), cf. Cucker & Smale (2003)

Given $\Phi_N \in \mathcal{M}_h$ there holds

$$|\mathcal{J}(\Phi_N) - \mathcal{J}_N(\Phi_N)| \leq \sup_{\Phi\in\mathcal{M}_h} |\mathcal{J}(\Phi) - \mathcal{J}_N(\Phi)| + \sigma_\rho$$

$$\begin{aligned} \Psi_N &:= \text{argmin}\,\{\mathcal{J}_N(\Phi) : \Phi \in \mathcal{M}_h\} (\text{ here } \sigma_\rho = 0) \\ \Psi_h &:= \text{argmin}\,\{\mathcal{J}(\Phi) : \Phi \in \mathcal{M}_h\}\ , \\ \Psi &:= \text{argmin}\,\{\mathcal{J}(\Phi) : \Phi \in \mathcal{V}\}\ , \end{aligned}$$

$$\begin{aligned} |\mathcal{J}_N(\Psi_N) - \mathcal{J}(\Psi_h)| &= |\mathcal{J}_N(\Psi_N) - \inf_{\Phi\in\mathcal{M}_h}\mathcal{J}(\Phi)| \\ &\leq 2 \sup_{\Phi\in\mathcal{M}_h} |\mathcal{J}(\Phi) - \mathcal{J}_N(\Phi)| \end{aligned}$$

$$\begin{aligned} |\mathcal{J}(\Psi_N) - \mathcal{J}(\Psi)| &\leq |\mathcal{J}(\Psi) - \mathcal{J}(\Psi_h)| \\ &\quad + 3 \sup_{\Phi\in\mathcal{M}_h} |\mathcal{J}(\Phi) - \mathcal{J}_N(\Phi)| \end{aligned}$$

Generalization error: $3 \sup_{\Phi\in\mathcal{M}_h} |\mathcal{J}(\Phi) - \mathcal{J}_N(\Phi)|$.

Numerical error: $|\mathcal{J}(\Psi) - \mathcal{J}(\Psi_h)| \leq C_2\|\Psi - \Psi_h\| (\lesssim \|\Psi - \Psi_h\|^2)$

Empirical Objective Minimization - Convergence in Probability

Instead of the deterministic estimate of the generalization error $|\mathcal{J}(\Phi) - \mathcal{J}_N(\Phi)| \leq \epsilon$ we estimate the probability with confidence $\delta \in (0,1)$,

$$\mathbb{P}\Big[\ \sup\ _{\Phi \in \mathcal{M}_h \cap \mathcal{A}} |\mathcal{J}(\Phi) - \mathcal{J}_N(\Phi)| \leq \epsilon\Big] \geq 1 - \delta \tag{2}$$

where $\delta = \delta(\epsilon, N) = \delta(h, \epsilon, N)$ depends only the model classes (hypothesis space(s)), desired accuracy ϵ, number of samples N.

$$\epsilon := |\mathcal{J}(\Psi) - \mathcal{J}_N(\Psi_N)|$$

$$\mathbb{P}\Big[|\mathcal{J}(\Psi) - \mathcal{J}(\Psi_N)| > \epsilon\Big] \leq \delta \tag{3}$$

$$\text{strong convexity} \ \Rightarrow\ \mathbb{P}\Big[\|\Psi - \Psi_N\|^2 \leq \epsilon\Big] \geq 1 - \delta \tag{4}$$

For fixed target accuracy ϵ, we need at least $N > N(\delta)$ samples.
For fixed N accuracy is limited to $\epsilon_{min} = \epsilon_{min}(N)$.

Generalization Error - General Theory

For $\Phi \in \mathcal{M}_h$, $0 < h \leq h_0$ compact model classes
$\mathcal{M}_h \subset\subset \mathcal{V} := V \otimes L_2(\mathbb{R}^d)$
Basic assumptions:

- Boundedness: For all $\Phi \in \mathcal{M}_h$, ,

$$\ell(\Phi(\xi)) \leq C_1 \; \forall \Phi \in \mathcal{M}_h \; , \quad \text{almost surely} \tag{5}$$

- Lipschitz continuity:

$$|\ell(\Phi_1(\xi)) - \ell(\Phi_2(\xi))| \leq C_2 \|\Phi_1 - \Phi_2\|_{\mathcal{V}} \; \forall \Phi_1, \Phi_2 \in \mathcal{M}_h \; , \quad \text{almost surely} \tag{6}$$

Covering number $\nu(\mathcal{M}_h, \delta)$, is the minimal number of balls of radius δ covering $\mathcal{M}_h$, w.r.t. to $\|. - .\|_{\mathcal{V}}$
Model classes $\mathcal{M}_h$ need to compact, but it do not to be linear subspaces nor convex, nor finite-dimensional. So far, this is not directly a numerical method.

Empirical Objective Minimization - Generalization Error

Theorem
For all $\epsilon > 0$,

$$\mathbb{P}\Big[\sup_{\Phi \in \mathcal{M}_h \cap \mathcal{A}} |\mathcal{J}(\Phi) - \mathcal{J}_N(\Phi)| > \epsilon \Big] \leq 2\nu(\mathcal{M}_h, \frac{\epsilon}{4C_2}) e^{-\frac{3\epsilon N}{4C_1^2}} \quad (7)$$

Proof is based on *Hoeffding inequality* $\Rightarrow$

$$\delta(\epsilon, N) := 2\nu(\mathcal{M}_h, \frac{\epsilon}{4C_2}) e^{-\frac{\epsilon^2 N}{2C_1^2}}$$

or better by *Bernstein inequality*, since $\sigma_\rho = 0$,

$$\delta(\epsilon, N) := 2\nu(\mathcal{M}_h, \frac{\epsilon}{4C_2}) e^{-\frac{3\epsilon N}{4C_1^2}}$$

These estimates may be improved ?!

Approximation Bounds - Strongly convex functionals

Assumption: (locally) Strong convexity $\Rightarrow$ for

$$\Psi := \text{ argmin } \{\mathcal{J}(\Phi) : \Phi \in \mathcal{V}\},$$

since $\mathcal{J}'(\psi)0$ there holds locally , i.e. $\forall \Phi : \|\Psi - \Phi\|_{\mathcal{V}} \leq \delta$

$$\|\Psi - \Phi\|^2 \quad \sim \quad |\mathcal{J}(\Psi) - \mathcal{J}(\Phi)|$$

Then

$$\begin{aligned}
\|\Psi - \Psi_h\|^2 &\lesssim |\mathcal{J}(\Psi) - \mathcal{J}(\Psi_h)| \\
&\lesssim \inf_{\Phi \in \mathcal{M}_h} \|\Psi - \Phi\|^2 \\
\|\Psi - \Psi_N\|^2 &\lesssim |\mathcal{J}(\Psi) - \mathcal{J}(\Psi_N)| \\
&\lesssim \inf_{\Phi \in \mathcal{M}_h} \|\Psi - \Phi\|^2 + 3 \sup{}_{\Phi \in \mathcal{M}_h} |\mathcal{J}(\Phi) - \mathcal{J}_N(\Phi)|
\end{aligned}$$

Open Problem: The exact computation of Ψ_N for non-convex models, e.g. HT (hierarchcial Tensors) or DNN (deep neural networks)

Model Classes – Linear Subspaces

$Au = f$, $\mathcal{V}_h \subset \mathcal{V}$ finite dimensional

$$\mathcal{M}_{h,\alpha} = \{u \in \mathcal{V}_h : \mathcal{J}(u) \leq \alpha\} \subset B_R(0)$$

is contained in a ball of radius R, we choose e.g.

$$\mathcal{M}_h := \mathcal{V}_h \cap B_R(0) \subset\subset \mathcal{V}$$

if $\mathcal{M}_{\alpha,N} = \{u \in \mathcal{V}_h : \mathcal{J}_N \leq \alpha\} \subset B_R(0)$ Galerkin matrices are approximated by Monte Carlo quadrature (variational Monte Carlo)
Covering Number:

$$\nu(\mathcal{M}_h,) \sim d^{\beta}(\frac{R}{\epsilon})^{\mathrm{dim}\mathcal{V}_h}$$

Kernel Methods - support vector machine - SKIP

Reproducing Kernel Hilbert spaces $\mathcal{K}$ - with kernels $k(x, y)$,

$$\|u\|_{\mathcal{K}}^2 = \langle u, Ku \rangle$$

$$B_R(0) := \{u \in \mathcal{K} : \|u\|_{\mathcal{K}} \leq R\} \subset\subset \mathcal{V}$$

Instead of

$$u_N := \text{argmin}\{\mathcal{J}_N(v) : \|u\|_{\mathcal{K}} \leq R\}$$

one solves the penalized problem

$$u_N := \text{argmin}\{\mathcal{J}_N(v) + \frac{\lambda}{2}\|v\|_{\mathcal{K}}^2\}$$

kernel ridge regression, for operators meshfree Galerkin
covering numbers: see e.g. Smale & Cucker , Cucker & Zhou

Sparsity models
ℓ_1 balls ....
Lasso, compressive sensing etc.

Deep Neuronal Networks - SKIP

$$u(\xi) := \sigma^j \circ (A^j \cdot - b^j) \circ \sigma^{j-1} \circ \cdots \circ \sigma^1(A^1\xi - b^1)$$

with nonlinear activation functions

$$\sigma : \mathbb{R} \to \mathbb{R} \ , \ \text{e.g. ReLu} : \sigma(x) := \max\{0, x\}$$

kernel networks $\sigma : \mathbb{R}^d \to \mathbb{R}$, e.g. $\sigma(x) := e^{-|x|^2}$
Hierarchical tensors - HT (Hackbusch 2009) (tree tensor networks)
includ. tensor trains - TT (Oseledets 2009) (matrix product states)
tensor product - multi-linear parametrization

$$\sigma : \mathbb{R}^2 \to \mathbb{R} \ : \ \sigma^\ell(\xi, \eta) := \xi \star \eta \ , \ \ell > 1 \ ,$$

$$\sigma : \mathbb{R}^d \to \mathbb{R} \ : \ \sigma^\ell(\xi_1, \cdots, \xi_d) := \xi_1 \star \cdots \star \xi_d \ , \ \ell > 1 \ .$$

Convolutionary networks (LeCun et al. 1996) pooling
$\xi \star \eta := \max\{\xi, \eta, 0\}$ (tensor product w.r.t. a tropical algebra?
Cohen & Shashua et al. (2017))

Covering Numbers - Examples

- Hierarchical Tensors (TT tensor trains)
  Covering number for HT been estimated by (Rauhhut & S. & Stojanac (2017))

$$\nu(\epsilon) \sim vol(B_1(0)) \left(\frac{R}{\epsilon}\right)^{n \log d} , \; n = dim \mathcal{M}_{\mathbf{r}}$$

Effective dimension ($\sim$ VC dim ?) $\log \nu = n |\log \epsilon| \log d$

Deep neural networks: e.g. Caffe, Tensorflow, etc.:

- $\mathcal{J}(\Phi(x^i))$ computable,

- optimization of parameters by back-propagation applicable.

- so far no theory

- Hierarchical Tensors (HT) are forming special deep networks (Cohen & Shashua et al. 2016-17),

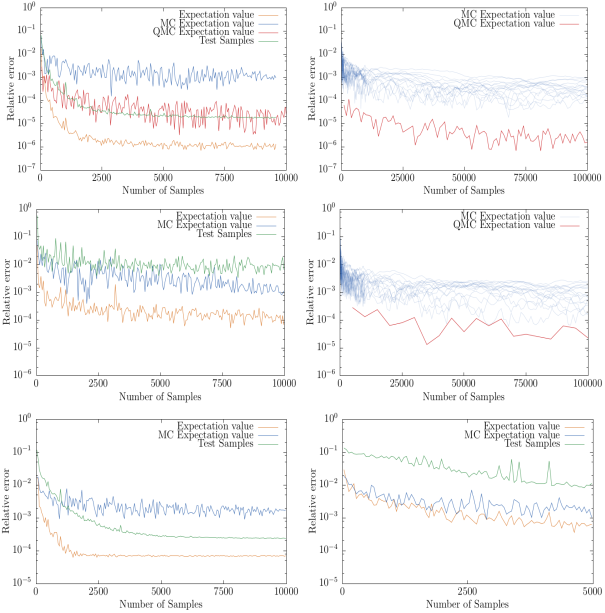
-but with theoretical results (algebraic varieties),...

e.g A. Nouy, A. Uschmajew, M. Bachmayr etc..

However, the present setting does not explain the small generalization operators of deep networks in connection with stochastic gradient.

work in progress

Expectation values: Affine, log-normal diffusion fields and cookie geometri



Transfer operator (Koopman operator) for MD simulation

Langevin dynamic: Joint work with F. Nüske & F. Vitalini & F. Noe (FUB), (extended dynamical mode decomposition)

First $m + 1 = 4(3)$ eigen functions corresponding to the largest m eigenvectors (the largest $\lambda_0 = 1$ is known)

A) $d = 18$ rotational vibrations, $m = 3$, $r_{max} = 3$, Deca Alanine

B) $d = 4$ rotational vibrations, $m = 3$, $r_{\max} = 4$, VGLAPG peptide

C) 1.05 ms simulation of 58-residue protein BPTI produced on the Anton supercomputer & provided by D. E. Shaw research
$d = 256$ distances, $m = 3$, $r_{max} = 6$. dim $\mathcal{V} = 2^{256} \approx 10^{85}$! dim $\mathcal{V}_{\text{als}} \leq nr^2 \leq 72$, dimension $\mathcal{M}_h \leq 1800$