Approximation and learning with compositional tensor trains

Martin Eigel[§,1], Charles Miranda[†,2], Anthony Nouy[§,2], David Sommer[1]

[†] PhD student (presenting author). [§] PhD supervisor

PhD expected duration: Nov. 2023 – Oct. 2026

[1] Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
{eigel,sommer}@wias-berlin.de

[2] Centrale Nantes, Nantes Université, Laboratoire de Mathématiques Jean Leray UMR CNRS 6629, France
{charles.miranda,anthony.nouy}@ec-nantes.fr

Abstract

We introduce compositional tensor trains (CTTs) for the approximation of multivariate functions, a class of models obtained by composing low-rank functions in the tensor-train format. This format can encode standard approximation tools, such as (sparse) polynomials, deep neural networks (DNNs) with fixed width, or tensor networks with arbitrary permutation of the inputs, or more general affine coordinate transformations, with similar complexities.

Formally, a CTT u is defined by linear operators $\mathfrak{L} : \mathbb{R}^d \to \mathbb{R}^p$ and $\mathfrak{R} : \mathbb{R}^p \to \mathbb{R}^{d_o}$ called respectively *lift* and *retraction*, and a univariate basis $\Phi = \{\phi_j : \mathbb{R} \to \mathbb{R}\}_{j=1}^n$, such that

$$u(x) = \mathfrak{R} \circ (\mathrm{Id} + \psi_L) \circ \cdots \circ (\mathrm{Id} + \psi_1) \circ \mathfrak{L}(x),$$

where ψ_k are tensors in the *Tensor-Train format* [6].

This format can be viewed as a DNN with width exponential in the input dimension and structured weights matrices. Compared to DNNs, this format enables controlled compression at the layer level using efficient tensor algebra.

On the optimization side, we derive a layerwise algorithm inspired by natural gradient descent [1], allowing to exploit efficient low-rank tensor algebra. The natural gradient descent tries to mimic the update in the functional space by an update in the parameter space. In the case of L^2 functions, this update simplifies to

$$\theta_{k+1} = \theta_k - \alpha_k G(\theta_k)^\dagger \nabla_\theta L(\theta_k),$$

where $G(\theta)_{ij} := \langle \partial_{\theta_i} u_\theta, \partial_{\theta_j} u_\theta \rangle$ is the *Gram matrix* and $L : \Theta \to \mathbb{R}$ is a *loss function* e.g. $L(\theta) = \frac{1}{2}\|u_\theta - v\|_{L^2}^2$.

In the case of CTT, the Gram matrix G can be stored efficiently due to the *Tensor-Train format* and its inherit low-rank format. Computing the update direction can be done using algorithms such as *alternating linear scheme (ALS)* [3]. However, the Gram matrix associated to each layer ℓ may have a bad condition number, so that ALS without preconditioning may show a slow convergence and yield a highly suboptimal low-rank approximation of the update direction.

A well-established approach to mitigate this issue is to approximate G_ℓ with a low-rank surrogate. In particular, the randomized Nyström method [5, 7, 2, 4] achieves this by projecting

G_ℓ onto a randomly generated, low-dimensional subspace. In the context of this work, the Gram matrix G_ℓ is a linear operator acting on tensor spaces, and so we can compute random projections using a tensor-structured sketch efficiently. The key advantage of this Gaussian sketching approach is that, with high probability, the span of the sketch captures the dominant eigenspace of G_ℓ Viewing the format as a discrete dynamical system, we also derive an optimization algorithm inspired by numerical methods in optimal control.

Numerical experiments on regression tasks demonstrate the expressivity of the new format and the relevance of the proposed optimization algorithms.

The Figure 1 shows the performance of the optimizer for a recovery problem where the TT ranks have provably high. Moreover, by computing layerwise updates, the optimizer is faster than the state of the art solvers.

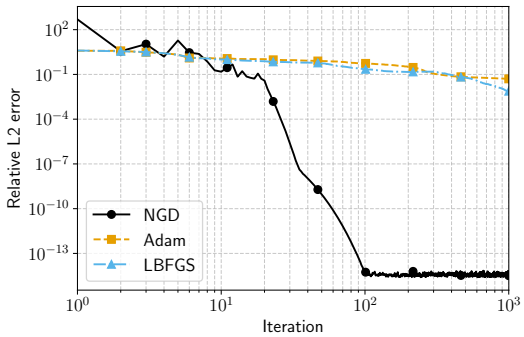

Figure 1: Convergence plot for the optimizers Adam, NGD and L-BFGS for a recovery problem, in log-log scale, for dimensions $d = 4$.

We also studied the *effective condition number* $\kappa_\ell(\theta) := \|G_\ell(\theta)\|_{2\to 2}\|G_\ell(\theta)^\dagger\|_{2\to 2}$ during the optimization. Experiments show that the Gram matrices become highly ill-conditioned during optimization, with condition numbers κ_ℓ ranging from 10^6 up to 10^{14}. In fact, the condition number increases rapidly, reaching values around 10^{13} after approximately 20 iterations. Initially, directions associated with small eigenvalues play a useful role by guiding the optimizer toward a good configuration. However, as the solution approaches optimality, these directions contribute progressively less to the reduction of the loss.

Finally, we applied the randomized method to the recovery problem and studying the convergence behavior for various sketching sizes. We have observed that retaining a rank-30 approximation of the Gram matrix is sufficient to achieve convergence to an optimal solution, which is less than half of the total eigendirections.

Overall, CTTs combine the expressivity of compositional models with the algorithmic efficiency of tensor algebra, offering a scalable alternative to standard deep neural networks.

Short biography (PhD student)

I am a third year PhD student working on *Approximation and learning with compositional functions networks*, supervised by Anthony NOUY (École Centrale de Nantes, Nantes, France), and Martin EIGEL (Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany). My research focuses on the approximation capabilities of compositional functions networks, and the design of optimization algorithms for these specific model classes. My research is funded by DFG-ANR Cofnet, and DR Centrale Nantes.

References

[1] Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, February 1998.

[2] Alex Gittens and Michael Mahoney. Revisiting the Nyström method for improved large-scale machine learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 567–575, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[3] Sebastian Holtz, Thorsten Rohwedder, and Reinhold Schneider. The Alternating Linear Scheme for Tensor Optimization in the Tensor Train Format. *SIAM Journal on Scientific Computing*, 34(2):A683–A713, January 2012.

[4] Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, May 2020.

[5] E. J. Nyström. Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*, 54(0):185–204, 1930.

[6] I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, January 2011.

[7] Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.