Optimisation of Overparametrized Sum-Product Networks

Martin Trapp[1,2], Robert Peharz[3] and Franz Pernkopf[1]

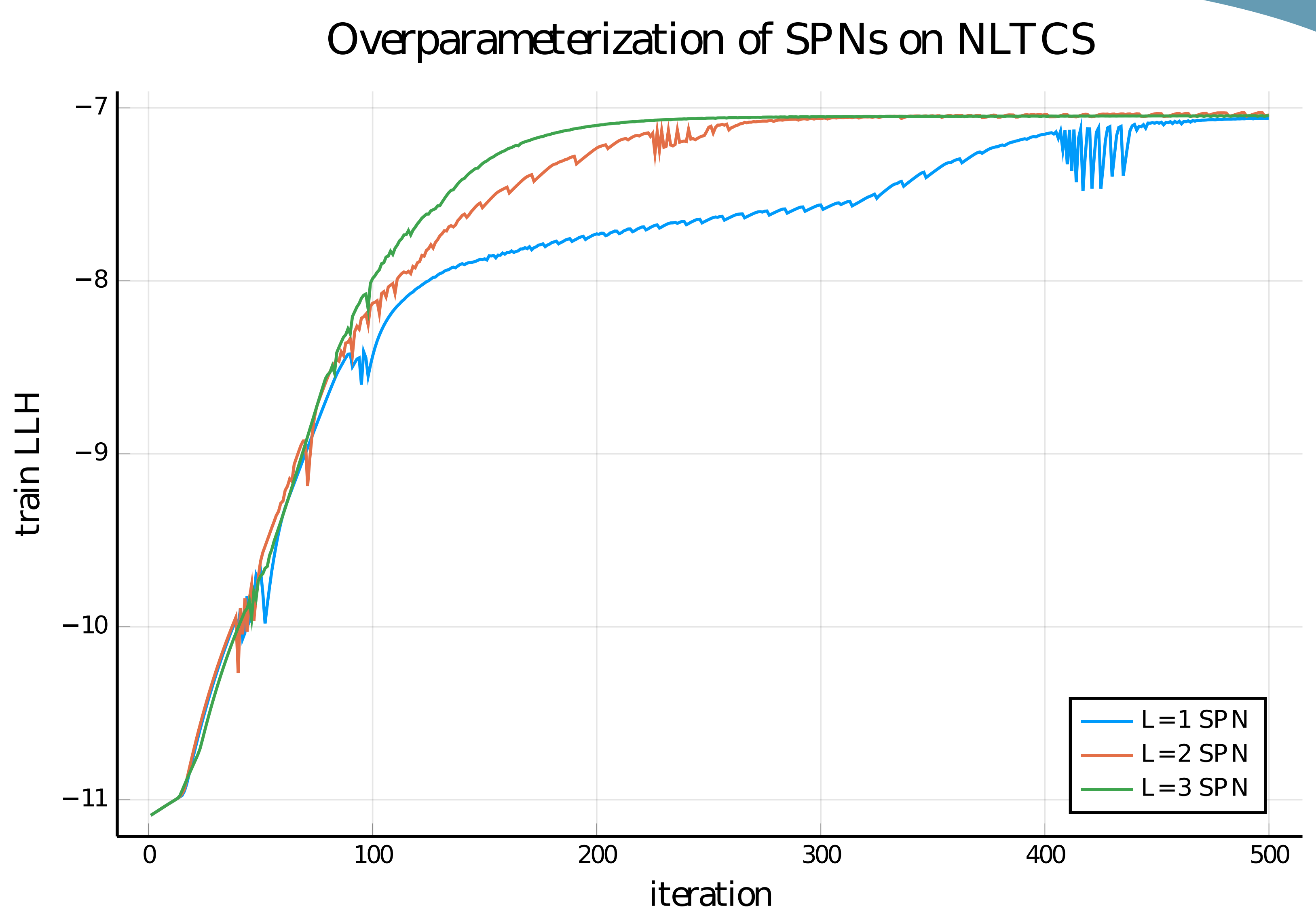
[1] SPSC Lab, Graz University of Technology, Graz, Austria
[2] Austrian Research Institute for Artificial Intelligence, Vienna, Austria
[3] Computational and Biological Learning Lab, University of Cambridge, Cambridge, UK

Sum-Product Networks (SPNs) are an expressive class of probabilistic models allowing exact and efficient inference.

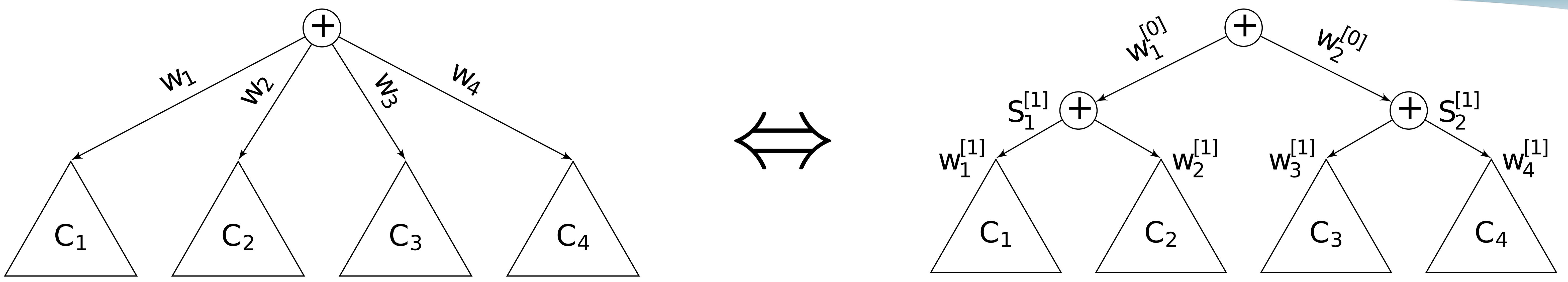
There seems to be a pearl of conventional wisdom that parameter learning is surprisingly fast in deep SPNs.

It turns out that gradient-based optimisation in deep tree-structured SPNs is equal to gradient optimisation with adaptive learning rate and momentum term.

Those acceleration effects are directly influence by the depth of the network even if the SPN model is non-linear e.g., the leaves are Gaussian.



Acceleration Effects in Sum-Product Networks



Generative learning of overparametrized SPNs

$$\mathcal{L}(\theta \,|\, \mathcal{X}) = \sum_{n=1}^{N} \log \mathrm{g}(\mathbf{x}_n \,|\, \theta) \underbrace{- \log \mathrm{f}(* \,|\, \theta)}_{\text{Partition function}}$$

$$\mathrm{g}(\mathbf{x}_n \,|\, \theta) = \sum_{k=1}^{K} \prod_{l=0}^{L} w^{[l]}_{\phi(k,l)} C_k(\mathbf{x}_n)$$

PDF of child k

Weight of component k decomposes into L weights

Deep tree-structured SPNs are equivalent to overparametrized SPNs

$$\mathcal{S} = \sum_{k=1}^{\kappa} \prod_{w_{S,C} \in \mathcal{T}_k} w_{S,C} \prod_{L \in \mathcal{T}_k} p(\mathbf{x}_n \,|\, \theta_L)$$

Induced tree

$$\prod_{l=0}^{L_{\mathcal{T}_k}} w^{[l]}_{\phi(k,l)} = \prod_{w_{S,C} \in \mathcal{T}_k} w_{S,C}$$

Depth of Tk

Indexing function: $\phi : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbf{S} \times \mathbf{N}$

Gradient of decomposed w_k at time t+1

$$w_k^{(t+1)} = w_\gamma^{[0](t+1)} \cdot w_k^{[1](t+1)}$$

Learning rate, close to zero

$$= \left[w_\gamma^{[0](t)} + \eta \nabla_{w_\gamma^{[0](t)}} \right] \left[w_k^{[1](t)} + \eta \nabla_{w_k^{[1](t)}} \right]$$

$$\approx w_k^{(t)} + \rho^{(t)} \nabla_{w_k^{(t)}} + \lambda^{(t)} w_k^{(t)}$$

w_k is sum of past gradients, due to $w_k^{(0)}$ close to 0

Gradient updates are updates with adaptive learning rate and momentum term

$$w_k^{(t)} \approx w_k^{(t)} + \rho^{(t)} \nabla_{w_k^{(t)}} + \sum_{\tau=1}^{t-1} \mu^{(t,\tau)} \nabla_{w_k^{(\tau)}}$$

Momentum term

$$\rho^{(t)} := \eta (w^{[0]}_{\phi(k,0)})^2$$

$$\lambda^{(t)} := \sum_{l=0}^{L-1} \eta \nabla_{w^{[l]}_{\phi(k,l)}} (w^{[l]}_{\phi(k,l)})^{-1}$$

ICML 2019 Workshop on Tractable Probabilistic Models
June 14th, 2019
Long Beach, USA

UNIVERSITY OF CAMBRIDGE

TU Graz

OFAI