Feature Learning as a Virtual Covariance Learning

Taehun Cha and Donghun Lee
Department of Mathematics, Korea University

NEURAL INFORMATION PROCESSING SYSTEMS

AIML@K

Research Gap

- ML/DL theory has mainly focused on model structures, e.g.:
  - Model Complexity (Generalization Analysis)
  - Static Infinite width NN (NTK, NNGP, …)
- **Feature learning happens depending on data structure, e.g.:**
  - Transformers trained on language data (GPT) vs. image data (ViT) learn different features
- We need a tool **inspecting NN from data & hidden space**
- Recently, **Neural Feature Ansatz (NFA)** is proposed: $W^\top W \propto \nabla_x f \cdot (\nabla_x f)^\top$
  - which suggests the relationship between learnt model and data covariance structure
  - Limited to explain ***"trained"* neural network, not neural network *"training"***

Virtual Covariance

For a neural network with hidden states $h_l = \sigma(W_{l-1}h_{l-1})$, where non-linearity σ, learning rate γ, input data $h_0 = x$, and loss function $\mathcal{L}$, define:

- **Virtual update**: $h_l^+ = h_l - \gamma\nabla_{h_l}\mathcal{L}$
- **Actual Update**: $W_l^+ = W_l - \gamma\nabla_{W_l}\mathcal{L}$
- **Virtual Covariance**: $\widetilde{cov}(h) = h \cdot h^\top$
- **Virtual Covariance Shift**: $\widetilde{cov}(h_l^+) - \widetilde{cov}(h_l)$

Rethinking SGD

- **Theorem 2**: The following holds up to a residual term of order γ^2

$$(W_l^+)^\top W_l^+ - (W_l)^\top W_l \approx \widetilde{cov}(h_l^+) - \widetilde{cov}(h_l)$$

(SGD-updated weight learns Virtual Covariance structure)

- **Theorem 3**: If σ is increasing and L-Lipschitz and $||h_{l-1}|| = 1$, then, element-wisely,

$$|\sigma(W_{l-1}^+ h_{l-1}) - \sigma(W_{l-1}h_{l-1})| < L^2|h_l^+ - h_l| \text{ and}$$

$$sgn\big(\sigma(W_{l-1}^+ h_{l-1}) - \sigma(W_{l-1}h_{l-1})\big) = sgn(h_l^+ - h_l)$$

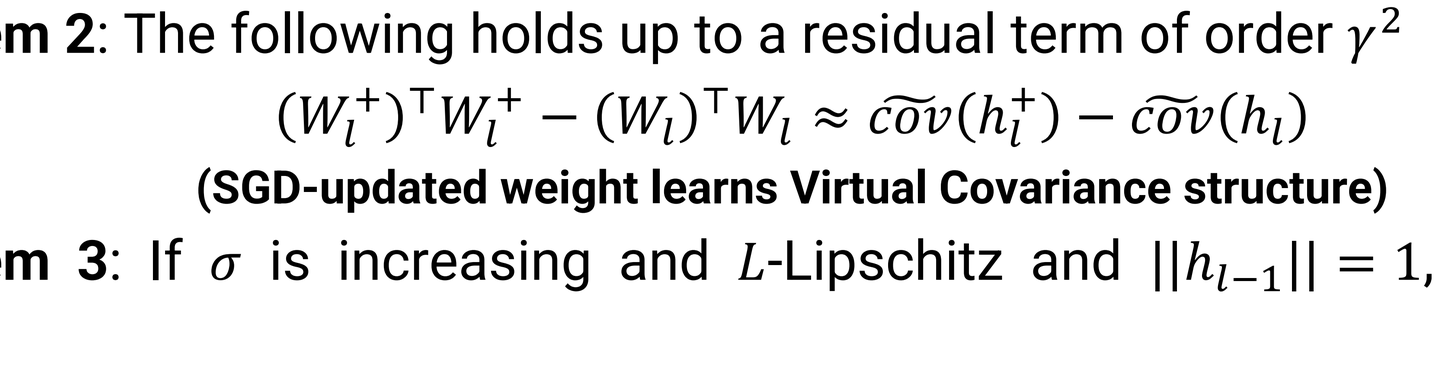
(The learnt Virtual Covariance structure doesn't deviate far from actually updated input)

Virtual Covariance Learning

- **Directly update the VC structure** with well-known orthogonal Procrustes problem

$$arg_{W_l^+}min||W_l^+ - W_l|| \; subject\ to\ (W_l^+)^\top W_l^+ - (W_l)^\top W_l = \widetilde{cov}(h_l^+) - \widetilde{cov}(h_l)$$

- **Results:**



Efficient
Robust
Non-overfitting

Discussion & Future Works

- VCL provides a tool to **analyze DNN "training" from the data (or hidden) space**
  - Feature learning *emerges* as DNN is implicitly trained to take the virtually updated input
- **Also works with CNN and deep NN** → Self-attention is left!
- **Also works with linear logistic regressor** → Easy to theoretically investigate!
- Need one or two SVD (or EVD) → Need further optimization!

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-24873052 and RS-2025-25433902)

I'm currently looking for a postdoctoral position in the mathematical foundations of DNNs and LLMs!

Personal