MEDIATEK research

Exact, Tractable Gauss-Newton Optimization in Deep Reversible Architectures Reveal Poor Generalization

Davide Buffelli*, Jamie McGowan*, Wangkun Xu, Alexandru Cioba, Da-shan Shiu, Guillaume Hennequin, Alberto Bernacchia

*Equal Contribution

NEURAL INFORMATION PROCESSING SYSTEMS

Preliminaries

- Theoretical studies of generalization in GN/NGD have been limited to simplified models:
  - linear models [Amari et al., 2021]
  - nonlinear models taken to their NTK limit [Zhang et al., 2019]
- Studies in “real-world” data so far have required approximations

Our Contributions:

- We derive an exact, computationally tractable expression for Gauss-Newton updates in deep reversible networks
- We study the generalization properties of GN in models up to 147 million parameters

Challenges with GN (and GGN)

- The update involves a pseudoinversion

$$\theta(t+1) = \theta(t) - \alpha J^{+} \nabla_f \tilde{L} \qquad \text{[recall } J = \frac{df}{d\theta}\text{]}$$

- J has dimensions $nd_{out} \times |\theta|$
  - Computing J requires $\min(nd_{out}, |\theta|)$ forward/backward passes (using VJPs or JVPs – could be batched if it can fit in memory)
  - Pseudoinverting requires $O(nd_{out}|\theta| \cdot \min(nd_{out}, |\theta|))$ compute and $O(nd_{out}|\theta|)$ memory

- We need to find an efficient way of computing $\boldsymbol{J}^{+}$

MEDIATEK research

Making the GN update tractable…

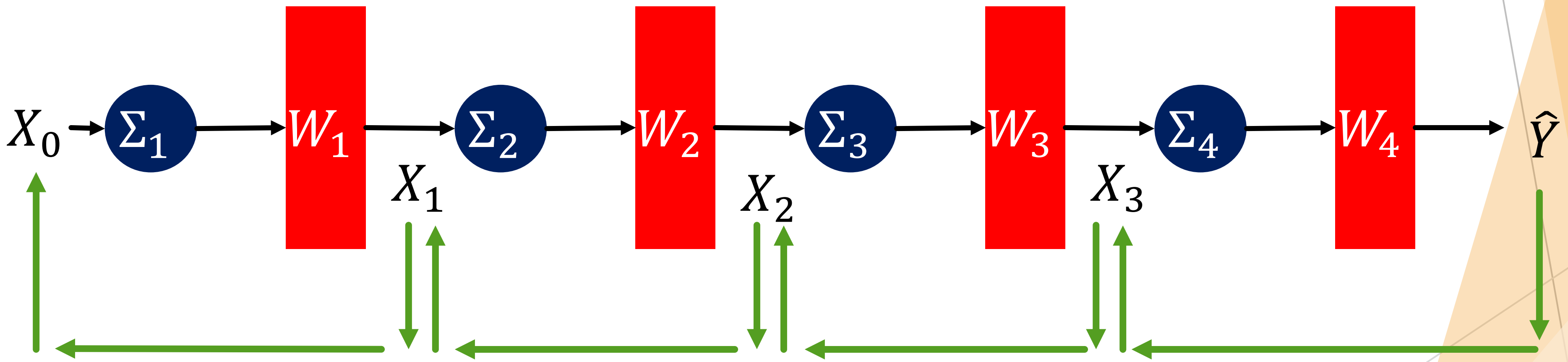
… for reversible neural networks

MEDIATEK research

Model Requirements

Models that are amenable to our method have two key properties:

- Every layer has output which is linear in the parameters
- Every layer is reversible (you can obtain the input starting from the output)



$$X_\ell = W_\ell \Sigma_\ell (X_{\ell-1})$$

MEDIATEK research

Practical GN Update

$$W_\ell(t+1) = W_\ell(t) - \alpha\, J_\ell^{+} \epsilon$$
$$= W_\ell(t) - \alpha\left[\frac{\partial \hat{Y}}{\partial X_\ell}\, \Sigma_\ell(X_{\ell-1})\right]^{+} \epsilon$$
$$= W_\ell(t) - \alpha\, \Sigma_\ell(X_{\ell-1})^{+} \frac{\partial X_\ell}{\partial \hat{Y}} \epsilon$$

Pseudoinversion of matrix with size $n \times d_{layer}$ which costs $O(nd_{layer} \cdot \min(n, d_{layer}))$

JVP – can be computed through autodiff at the cost of 1 forward pass

Our update

- computational cost of $O(Lnd^2 + Ln^2d)$
- memory cost of $O(nd + |\theta|)$

SGD

- computational cost of $O(Lnd^2)$
- memory cost of $O(nd + |\theta|)$

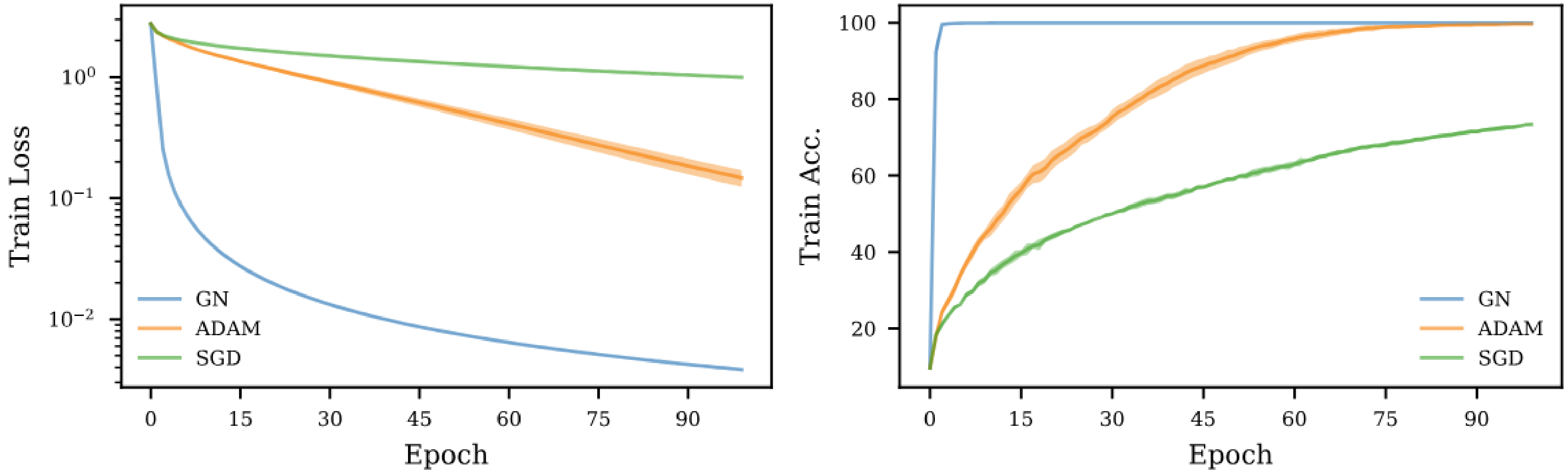
Theoretical Result

Our update with $J^{\dashv}$ has the **same convergence properties** as Gauss-Newton with J^{+} assuming J has linearly independent rows for all $\boldsymbol{\theta}$

Experiments

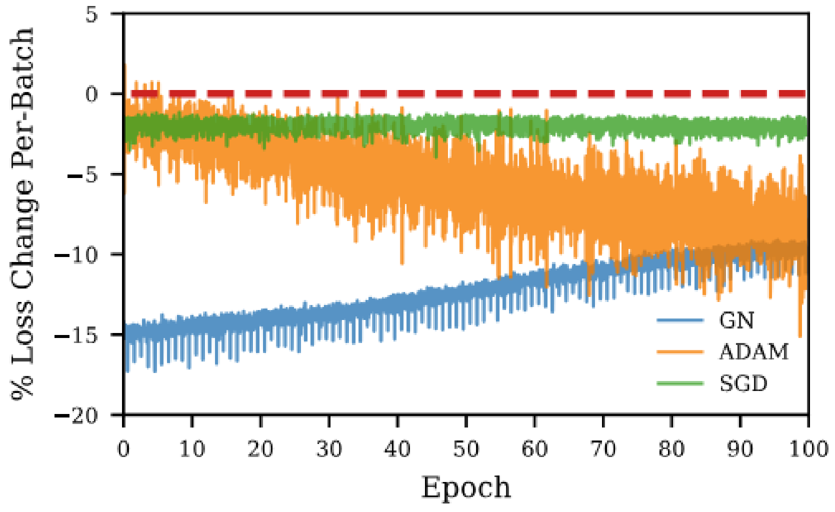
Full Batch

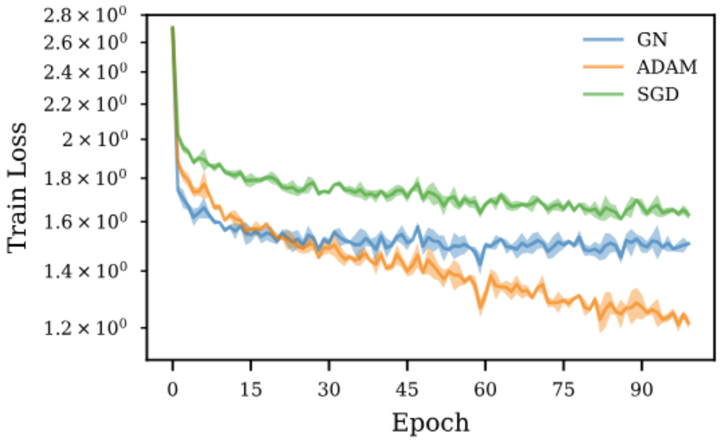
- Full Batch = random **subset** of data of size 1024
- All the theory is in full batch, these experiments verify that the theory is **correct**.
- Gauss-Newton trains extremely **quickly** and has a **little variation** across seeds compared to Adam/SGD.

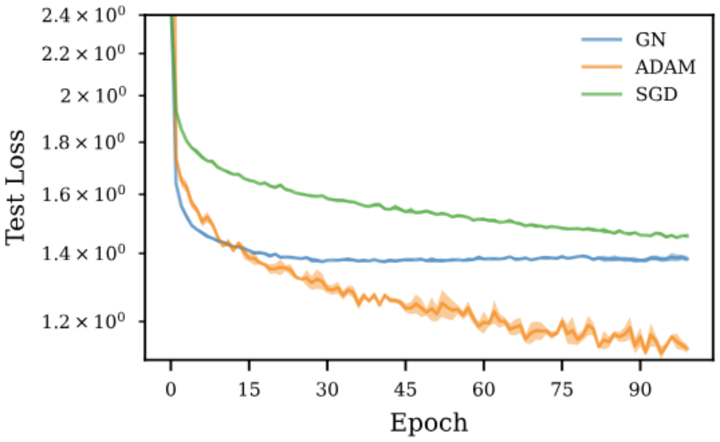


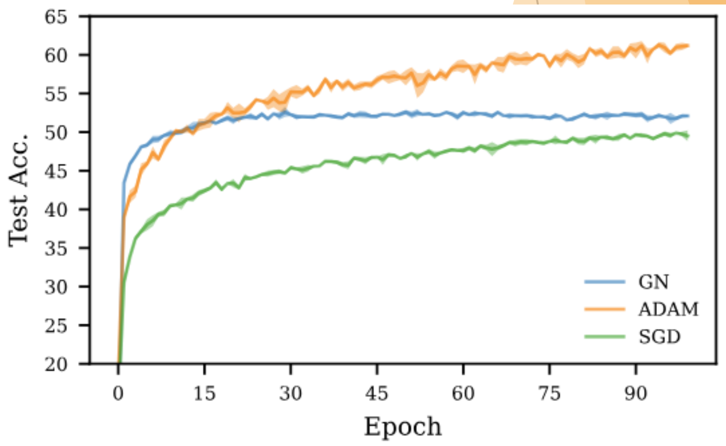
Mini-Batch

- In a mini-batch setting, the method breaks down
- Our hypothesis is that GN "*overfits*" to each mini-batch
- We test this by measuring the loss on the same mini-batch **before and after the update**.
- GN leads to a much stronger immediate **decrease** in the loss



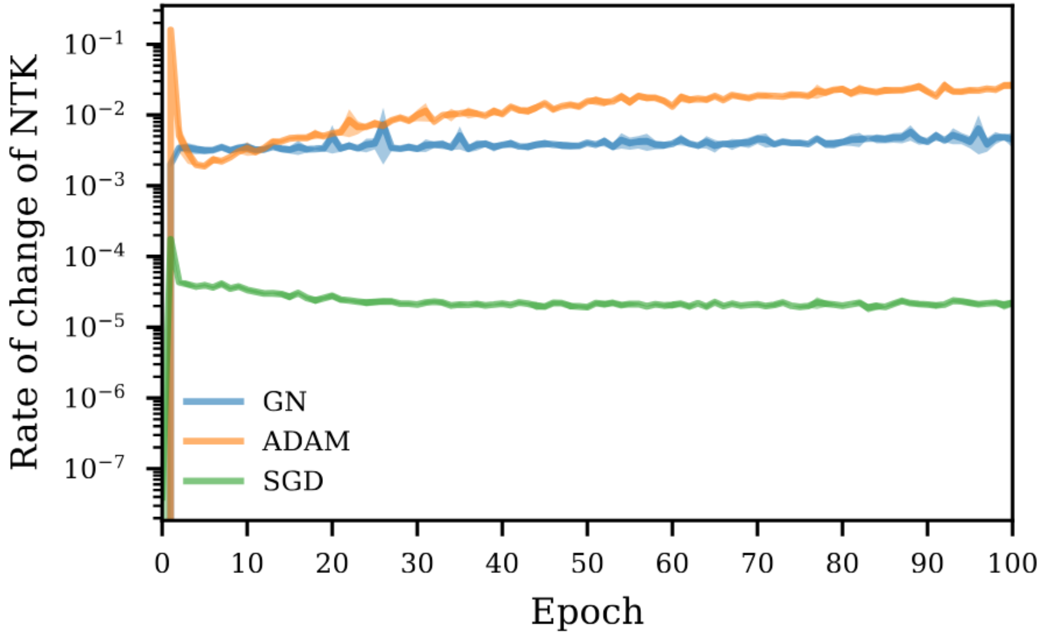


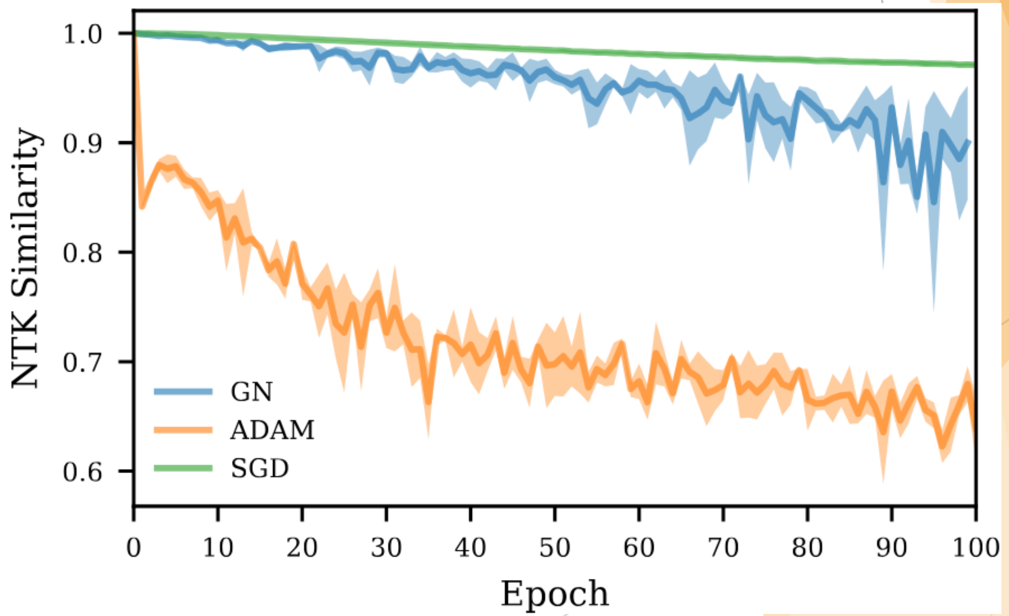




Evolution of the Neural Tangent Kernel

- The community considers "*feature learning*" as a change in the NTK
- We observe **almost no change** in the NTK
- This signals that Gauss-Newton is not learning features.

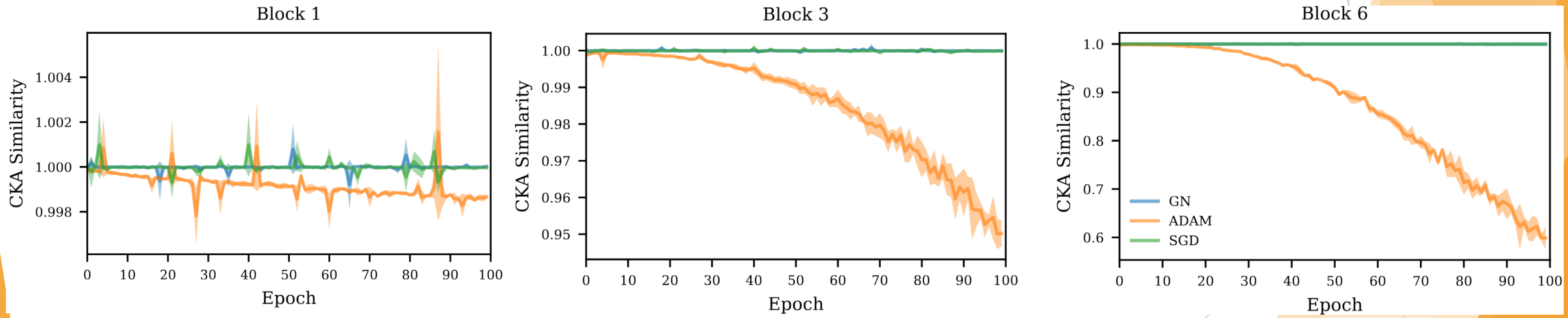




MEDIATEK research

Feature Learning with Gauss-Newton

- As another diagnostic test, we check the **CKA similarity across training** for each layer – with respect to initialization.
- This shows that the neural representations do not change during training
  - In addition to the *constant NTK* this is indicative of a “lazy” learning regime.
- Gauss-Newton seems to not be able to promote feature learning



Many more experiments available in the paper!

Thanks for Listening

Come and chat with us at our booth!

MEDIATEK research

Is Our Practical Update "the Same" as the Theoretical One?

Yes! *(under some assumptions)*

Assumption 4.1. Assume $J(\boldsymbol{\theta})$ has linearly independent rows (is surjective) for all $\boldsymbol{\theta}$ in the domain where GN dynamics takes place.

Theorem 4.3. *Under Assumption 4.1 so that there is a right inverse $J^{\dashv}$ satisfying $JJ^{\dashv} = I$, consider the update in parameter space with respect to the flow induced by an arbitrary right inverse $J^{\dashv}$:*

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha J^{\dashv} \nabla_{\mathbf{f}} \tilde{\mathcal{L}}. \tag{8}$$

Then the loss along these trajectories is the same up to $\mathcal{O}(\alpha)$, i.e. for any two choices $J_1^{\dashv}$ and $J_2^{\dashv}$, the corresponding iterates $\boldsymbol{\theta}_t^{(1)}$ and $\boldsymbol{\theta}_t^{(2)}$ satisfy:

$$\|\nabla_{\mathbf{f}} \tilde{\mathcal{L}}(\mathbf{f}(\boldsymbol{\theta}_t^{(1)})) - \nabla_{\mathbf{f}} \tilde{\mathcal{L}}(\mathbf{f}(\boldsymbol{\theta}_t^{(2)}))\| \leq \mathcal{O}(\alpha). \tag{9}$$

Moreover, as the Moore-Penrose pseudo-inverse is a right inverse under the assumptions, the result applies to J^+, and consequently to the dynamics of (5).

MEDIATEK research