# Optimization of Neural Networks with an Explicit Regularization: Generalized Gauss-Newton Method

Adeyemi D. Adeoye [1], Philipp Christian Petersen [2], Alberto Bemporad [1]

[1] *IMT School for Advanced Studies Lucca, Italy*, [2] *Faculty of Mathematics, University of Vienna, Austria*

IMT SCHOOL FOR ADVANCED STUDIES LUCCA

## Neural Network Training

Let $n$ = number of hidden neurons, $n_0$ = input dimension. The one-hidden layer NN is:

$$\mathbb{R}^{n_0} \ni x \mapsto \Phi(x;\theta) \triangleq \kappa(n)\sum_{i=1}^{n} v_i \varrho(u_i x)$$

The training task: Find $\theta$ minimizing

$$\min_{\theta\in\mathbb{R}^p} \mathcal{L}(\theta) \triangleq \underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell(\Phi(x_i;\theta), y_i)}_{\hat{R}_s(\Phi)} \; + \; g(\theta)$$

$g$ is a convex regularizer

## Setup & key assumptions

$Q_t = \nabla^2_{\Phi_t}\hat{R}_s(\Phi_t)$, $e_t = \nabla_{\Phi_t}\hat{R}_s(\Phi_t)$, $H_t = \nabla^2 g(\theta_t)$, $J_t = (\nabla_\theta \Phi(x_1,\theta_t), \ldots, \nabla_\theta \Phi(x_m,\theta_t))^\top$

**Regularized GGN iterations**: augment $Q_t$, $e_t$ and $J_t$, resp. by 0, 1 and $\nabla g(\theta_t)$ in appropriate dimensions; denote by $\hat{Q}_t$, $\hat{e}_t$ and $\hat{J}_t$, resp.

$$\theta_{t+1} = \theta_t - \alpha_t(\hat{J}_t^\top \hat{Q}_t \hat{J}_t + H_t)^{-1}\hat{J}_t^\top \hat{e}_t$$

where $\alpha_t$ are step sizes (or *learning rates*)

Convenient form for overparameterized models:

$$\theta_{t+1} = \theta_t - \alpha_t H_t^{-1}\hat{J}_t^\top (I + \hat{Q}_t\hat{J}_t H_t^{-1}\hat{J}_t^\top)^{-1}\hat{e}_t$$

- $\varrho$ is twice differentiable, Lipschitz, and smooth
- $g$ is thrice differentiable and $(M_g,\nu)$-GSC (generalized self-concordant): $\forall u, v \in \mathbb{R}^p$, $|\langle \nabla^3 g(x)[v]u, u\rangle| \le M_g \|u\|_x^2 \|v\|_x^{\nu-2}\|v\|^{3-\nu}$
- $\hat{R}_s$ is $\gamma_R$-strongly convex, and has upper-bounded gradients and Hessian; $g$, $\hat{Q}_t$ and $\hat{e}_t$ are locally bounded

## GGN-NTK (zero damping limit)

Let $g(\theta) \equiv \tau\bar{g}(\theta)$; $\tau$ controls the regularization strength

Zero damping limit ($\tau \to 0$) + Infinite overparameterization $\iff$ **dynamics are stable**, and the (unregularized) GGN iterations:

$$\theta_{t+1} = \theta_t - \alpha_t J_t^\top G_t^{-1} e_t$$

**NTK matrix**: $G_{t_{i,j}} = \langle \nabla_\theta \Phi(x_i,\theta_t), \nabla_\theta \Phi(x_j,\theta_t)\rangle$

Moore-Penrose inverse ≡ Overparameterized NNs

**Note:** In the infinite-width limit, the GD reduces to the kernel gradient descent:

$$\Phi_{t+1} = \Phi_t - \alpha_t G_t \nabla_{\Phi_t}\hat{R}_s(\Phi_t)$$

Correspondingly, the NTK regression is:

$$\theta_{t+1} = \operatorname*{argmin}_\theta \frac{1}{2}\|\langle J_t, \theta - \theta_t\rangle + \nabla_{\Phi_t}\hat{R}_s(\Phi_t)\|^2$$

$\iff$ **linearization of $\Phi$ around $\theta_t$**

(the **GGN-NTK relation**)

- When $\theta_t$ is close $\theta_0$, the linearization provides a good approximation to $\Phi$
- Key properties: **stability**, **generalization**

## Regularization with stability

If $\tau > 0$ ($g \neq 0$), the relation between gradient descent and NTK will probably break.

However, the GGN dynamics still enjoy stability and generalization with infinite overparameterization:

- $g$ is $(M_g,\nu)$-GSC $\iff$ locally stable $H_t$
- Bounded terms in the GGN iterates

In this case,

$$\Phi_{t+1} = \Phi_t - \alpha_t \hat{G}_t \hat{e}_t$$

where $\hat{G}_t \triangleq J_t H_t^{-1}\hat{J}_t^\top (I + \hat{Q}_t\hat{J}_t H_t^{-1}\hat{J}_t^\top)^{-1}$

- Empirically, this can be simulated with small step sizes for GGN (equivalently, the *hidden learning* phenomenon)

## Theory: setup

For the regularized GGN updates, consider the adaptive learning rate selection rule

$$\alpha_t = \frac{\bar{\alpha}_t}{1 + M_g \eta_t}$$

where $0 < \bar{\alpha}_t \le 1$ and $\eta_t = \|\nabla g(\theta_t)\|^*_{\theta_t}$

- For convenience, let $\tilde{\Phi}_t \in \mathbb{R}^{m+1}$ denote the vector obtained by augmenting $\Phi_t$ by 1
This $\tilde{\Phi}_t$ corresponds to a different augmented version of $J_t$ denoted by $\tilde{J}_t$. We have
$$\tilde{\Phi}_{t+1} = \tilde{\Phi}_t - \alpha_t \tilde{G}_t \hat{e}_t$$
where $\tilde{G}_t \triangleq \tilde{J}_t H_t^{-1}\hat{J}_t^\top (I + \hat{Q}_t\hat{J}_t H_t^{-1}\hat{J}_t^\top)^{-1}$
Also, let $\tilde{\Phi}^* \in \mathbb{R}^{m+1}$ denote the vector obtained by augmenting $\Phi^*_t$ by 0
- Let $B_R$, $B_\Phi$, $B_g$, $d_g$, $d_q$, $\beta$, $D_g$ and $D_R$ be fixed constant terms defined by the regularity assumption. Also, introduce $\hat{\beta}_m \triangleq \sigma_{\min}(J)$, the smallest singular value of $J$
- Suppose the GGN iterates remain inside the ball $\mathcal{B}_{r_0}(\theta_0) \subset \mathcal{E}_r(\theta_0)$, where $\mathcal{E}_r(\theta_0)$ is an ellipsoid

## Theory: convergence

**Theorem.** Fix $0 < \bar{\alpha}_t \equiv \bar{\alpha} < 1$, and choose $T \triangleq \frac{1}{\bar{\alpha}}\log(\|\tilde{\Phi}_0 - \tilde{\Phi}^*\|^2/\epsilon)$ for any $\epsilon \in (0,1)$. It holds that $\|\Phi_T - \Phi^*\|^2 \le \epsilon$ (or equivalently, $\|\tilde{\Phi}_T - \tilde{\Phi}^*\|^2 \le \epsilon + 1$) after $T$ iterations, if $1 + M_g\eta_t \le \|\tilde{G}_t\|_F$ and $|\tilde{G}_{22}| \ge |\langle \tilde{G}_{21}^\top + \tilde{G}_{12}, \tilde{v}\rangle|$ for some $\tilde{v}$ depending on $t \le T$ where, given a $2\times 2$ block partitioning of $\tilde{G}_t$, $\tilde{G}_{22} \in \mathbb{R}^{1\times 1}$, $\tilde{G}_{21} \in \mathbb{R}^{1\times(m+1)}$, and $\tilde{G}_{12} \in \mathbb{R}^{(m+1)\times 1}$ respectively denote the lower right, lower left and upper right blocks of $\tilde{G}_t$

- It is reasonable to choose $\tau$ satisfying $1 + \tau M_g \eta_0 \le \sqrt{(m+1)\lambda_1(\tilde{G}_0^\top \tilde{G}_0)}$

## Theory: loss decay

**Theorem.** The loss decays according to $\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t) - \left[\vartheta L^2_{D_t}(1 + D_g\varpi_t) - \xi L_{D_t}\right]$, for $t \ge 0$, where $L_{D_t} \triangleq \frac{\alpha_t\beta\hat{\beta}_1 D_g}{d_g(D_g + d_q\hat{\beta}_m^2)}$, $\xi \triangleq B_R B_\Phi + B_g$, $\vartheta \triangleq B^2_\Phi(\gamma_R - D_R)$, $\varpi_t \triangleq \omega_\nu(d_\nu(\theta_t,\theta_{t+1})) - \omega_\nu(-d_\nu(\theta_t,\theta_{t+1}))$, $\omega_\nu$ is an increasing univariate function, $d_\nu$ is a scaled metric term associated with the self-concordance of $g$, and we assume $d_\nu(\theta_t,\theta_{t+1}) < 1$

## Simulation: teacher-student

Let $\theta^* \equiv (u^*, v^*)$, $\varrho(x) \triangleq x/(1+\exp(-x))$ and

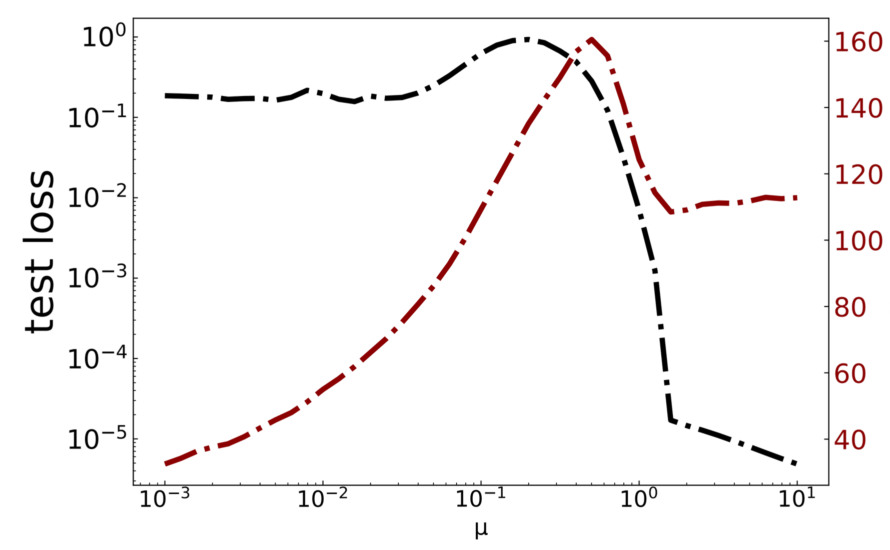
$$\mathbb{R}^{n_0} \ni x \mapsto \Phi^*(x;\theta^*) \triangleq \sum_{i=1}^{n^*} v_i^* \varrho(u_i^* x)$$

$$\bar{g}(\theta) = \sum_{i=1}^{p} \frac{\mu^2 - \mu\sqrt{\mu^2 + \theta_i^2} + \theta_i^2}{\sqrt{\mu^2 + \theta_i^2}}$$

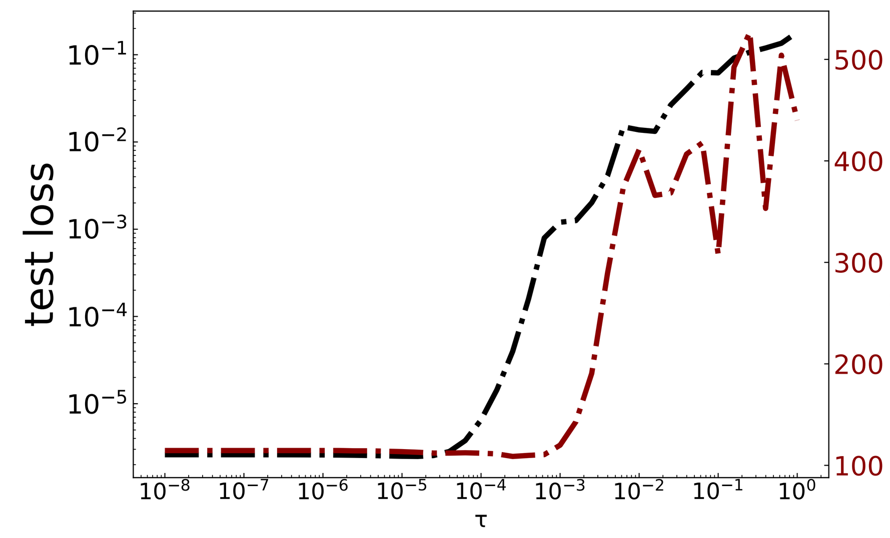
where $M_g = 2\mu^{-0.7}p^{0.2}$, $\nu = 2.6$, $\mu = 1/\kappa(n)$

Train: 500, test: 1000, $n = 500$, $n^* = 5$

$\tau = 10^{-4}$

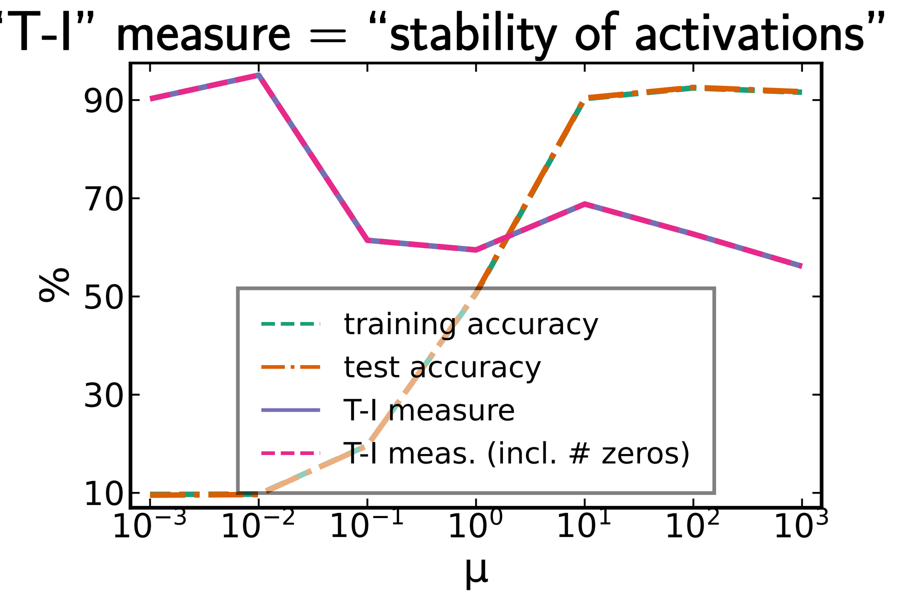



$\mu = 1/\sqrt{n}$




MNIST

"T-I" measure = "stability of activations"




## References

[1] Wei, C., Lee, J. D., Liu, Q., & Ma, T. (2019). Regularization matters: Generalization and optimization of neural nets vs their induced kernel, *Advances in Neural Information Processing Systems, 32*

[2] Cai, T., Gao, R., Hou, J., Chen, S., Wang, D., He, D., ... & Wang, L. (2019). Gram-gauss-newton method: Learning overparameterized neural networks for regression problems, *arXiv preprint arXiv:1905.11675*

[3] Arbel, M., Menegaux, R., & Wolinski, P. (2024). Rethinking Gauss-Newton for learning over-parameterized models, *Advances in Neural Information Processing Systems, 36*

[4] Adeoye, A. D., Petersen, P. C., & Bemporad, A. (2024). Regularized Gauss-Newton for Optimizing Overparameterized Neural Networks, *arXiv preprint arXiv:2404.14875*