SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Newton Design: Designing CNNs with the Family of Newton's Methods

Zhengyang Shen[1,2], Yibo Yang[3], Qi She[2], Changhu Wang[2], Jinwen Ma[1*] & Zhouchen Lin[4,5*]

[1]*School of Mathematical Sciences, Peking University, Beijing* 100871*, China;*
[2]*Bytedance AI Lab, Haidian District, Beijing* 100871*, China;*
[3]*JD Explore Academy, Beijing 100176, China;*
[4]*Key Lab. of Machine Perception, School of Intelligence Science and Technology, Peking University, Beijing* 100871*, China;*
[5]*Pazhou Lab, Guangzhou* 510320*, China*

Appendix A A Brief Introduction of The Family of Newton's Method

As for the optimization problem

$$\min_x \ F(x),$$

the iteration of Newton's method is

$$x_{k+1} = x_k - \left[\nabla^2 F(x_k)\right]^{-1} \nabla F(x_k),$$

where $\nabla F(x)$ and $\nabla^2 F(x)$ are the gradient and the Hessian matrix of $F(x)$, respectively. Since it is always time-consuming to calculate the inverse Hessian matrix, we can also use a matrix H_k to approximate $\left[\nabla^2 F(x_k)\right]^{-1}$, and obtain the iteration of the quasi-Newton method

$$x_{k+1} = x_k - H_k \nabla F(x_k).$$

H_k can be acquired using multiple methods, such as the rank one correction formula, DFP and BFGS algorithm, etc.

Appendix B The Convergence Rate of Quasi-Newton Method

Theorem 1. If $A \in \mathbb{S}^n$ and $\|A\|_2 < 1$,$x \in \mathbb{R}^n$, Φ is the ReLU function, then the iteration $x_{k+1} = \Phi(Ax_k)$ converges to $x^\star$ with a linear convergence rate.

Proof. Firstly, we prove that the map $y = \Phi(Ax)$ is a contractive map. Let $x^{(1)}, x^{(2)} \in \mathbb{R}^n, A = (a_1, a_2, \cdots, a_n)^T$. According to Lagrangian median theorem, there exists $s_i \in \mathbb{R}^n, i = 1, 2, \cdots, n$, such that

$$\begin{aligned} y_i^{(2)} - y_i^{(1)} &= \Phi(a_i^T x^{(2)}) - \Phi(a_i^T x^{(1)}) \\ &= \Phi'(a_i^T s_i) a_i^T (x^{(2)} - x^{(1)}). \end{aligned} \tag{B1}$$

thus

$$\begin{aligned} &\|y^{(2)} - y^{(1)}\|_2 = \|\Phi(Ax^{(2)}) - \Phi(Ax^{(1)})\|_2 \\ =&\|[\Phi'(a_1^T s_1)a_1, \cdots, \Phi'(a_n^T s_n)a_n]^T (x^{(2)} - x^{(1)})\|_2 \\ \leqslant&\|[\Phi'(a_1^T s_1)a_1, \cdots, \Phi'(a_n^T s_n)a_n]^T\|_2 \|x^{(2)} - x^{(1)}\|_2 \\ =&\|DA\|_2 \|x^{(2)} - x^{(1)}\|_2, . \end{aligned} \tag{B2}$$

where D is a diagnal matrix, and the diagnal elements are 0 or 1. In addition,

$$\begin{aligned} \|DA\|_2^2 &= \max_{\|y\|_2=1} \{y^T (DA)(DA)^T y\} \\ &= \max_{\|y\|_2=1} \{(D^T y)^T AA^T (D^T y)\} \\ &\leqslant \max_{\|y\|_2=1} \{y^T AA^T y\} = \|A\|_2^2, \end{aligned} \tag{B3}$$

* Corresponding author (email: jwma@math.pku.edu.cn, zlin@pku.edu.cn)

so

$$\|y^{(2)} - y^{(1)}\|_2 \leqslant \|A\|_2 \|x^{(2)} - x^{(1)}\|_2, (\|A\|_2 < 1), \tag{B4}$$

i.e., $y = \Phi(Ax)$ is a contractive map.

According to the contractive map principle, the iteration $x_{k+1} = \Phi(Ax_k)$ converges to $x^\star$, and similarly

$$\begin{aligned} \|x_{k+1} - x^\star\|_2 =& \|\Phi(Ax_k) - \Phi(Ax^\star)\|_2 \\ \leqslant& \|A\|_2 \|x_k - x^\star\|_2 \end{aligned} \tag{B5}$$

so it has a linear convergence rate. ■