Natural Neural Tangent Kernels in DL and RL

Youssef Mousaaid

July 11, 2022

Outline:

1. Neural Tangent Kernels in DNN
2. Natural Neural Tangent Kernels in DNN
3. Towards Neural Tangent Kernels in RL

References:

Neural Tangent Kernel: Convergence and Generalization in Neural Networks, arXiv/1806.07572 (Jacot et. al.)

Motivation

- Loss

$$L(\theta) = \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i), \qquad x_i \in X$$

- Gradient flow

$$\partial_t \theta = -\nabla_\theta L(\theta) = -\sum_{j=1}^{n} \left(\frac{\partial f_\theta(x_j)}{\partial \theta} \right)^T \frac{\partial \ell(z, y_j)}{\partial z}\Big|_{z=f_\theta(x_j)}$$

- Dynamics of f_θ during training

$$\begin{aligned} \partial_t f_\theta(x_i) &= \frac{\partial f_\theta(x_i)}{\partial \theta} \partial_t \theta \\ &= -\frac{\partial f_\theta(x_i)}{\partial \theta} \sum_{j=1}^{n} \left(\frac{\partial f_\theta(x_j)}{\partial \theta} \right)^T \frac{\partial \ell(z, y_j)}{\partial z}\Big|_{z=f_\theta(x_j)} \end{aligned}$$

Motivation

Dynamics of f_θ during training:

$$\partial_t f_\theta(x_i) = -\partial_\theta f_\theta(x_i) \sum_{j=1}^{n} \left(\partial_\theta f_\theta(x_j)\right)^T \partial_z \ell(z, y_j)|_{z=f_\theta(x_j)}$$

$$= -\partial_\theta f_\theta(x_i) \begin{pmatrix} \partial_\theta f_\theta(x_1) \\ \vdots \\ \partial_\theta f_\theta(x_n) \end{pmatrix}^T \begin{pmatrix} \partial_z \ell(z, y_1)|_{z=f_\theta(x_1)} \\ \vdots \\ \partial_z \ell(z, y_n)|_{z=f_\theta(x_n)} \end{pmatrix}$$

$$\Rightarrow \partial_t \underbrace{\begin{pmatrix} f_\theta(x_1) \\ \vdots \\ f_\theta(x_n) \end{pmatrix}}_{f_\theta(X)} = - \underbrace{\begin{pmatrix} \partial_\theta f_\theta(x_1) \\ \vdots \\ \partial_\theta f_\theta(x_n) \end{pmatrix} \begin{pmatrix} \partial_\theta f_\theta(x_1) \\ \vdots \\ \partial_\theta f_\theta(x_n) \end{pmatrix}^T}_{\Theta(X,X)} \begin{pmatrix} \partial_z \ell(z, y_1)|_{z=f_\theta(x_1)} \\ \vdots \\ \partial_z \ell(z, y_n)|_{z=f_\theta(x_n)} \end{pmatrix}$$

Neural tangent kernels

If ℓ is the mean squared error loss, then the dynamics simplify to

$$\partial_t f_\theta(X) = -\Theta(X, X)\left(f_\theta(X) - y\right),$$

where $y = (y_1, \ldots, y_n)^T$.

- This means

$$f_{\theta+\epsilon} = f_\theta - \epsilon\Theta(X, X)\left(f_\theta(X) - y\right) + o(\epsilon).$$

  Hence the name neural tangent kernel.

- Furthermore, if Θ stays/becomes constant, the linear ODE has as a solution

$$f_{\theta(t)}(X) = f_{\theta^*}(X) + \exp\left(-t\Theta(X, X)\right)\left(f_{\theta(0)}(X) - f_{\theta^*}(X)\right)$$

The search for constant neural tangent kernels

The setup:

- a fully connected neural net with layers numbers from 0 (input layer) to L (output layer), each containing $n_0, \dots, n_L$ neurons,
- a Lipschitz, twice differentiable nonlinearity function $\sigma \colon \mathbb{R} \to \mathbb{R}$, with bounded second derivative,
- the input space $\mathbb{R}^{n_0}$ follows a fixed distribution p.
- Denote the function of this network by $f_\theta \colon \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$, where $\theta \in \mathbb{R}^P$.

The setup

The network function is given by

$$f_\theta(x) := \tilde{\alpha}^{(L)}(x;\theta),$$

where the functions $\tilde{\alpha}^{(\ell)}(\cdot;\theta) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_\ell}$ (called preactivations) and $\alpha^{(\ell)}(\cdot;\theta) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_\ell}$ (called activations) are defined from the 0-th to the L-th layer by:

$$\begin{aligned}
\alpha^{(0)}(x;\theta) &= x \\
\tilde{\alpha}^{(\ell+1)}(x;\theta) &= \frac{1}{\sqrt{n_\ell}} W^{(\ell)} \alpha^{(\ell)}(x;\theta) + \beta b^{(\ell)} \\
\alpha^{(\ell)}(x;\theta) &= \sigma(\tilde{\alpha}^{(\ell)}(x;\theta)).
\end{aligned}$$

Here $\beta > 0$ is some fixed parameter.

At initialization

$$W^{\ell}_{i,j}, b^{\ell}_{i} \sim \mathcal{N}(0,1)$$

$$\Theta(x, x') = \sum_{i=1}^{P} \partial_p f_\theta(x) \partial_p f_\theta(x')$$

“Infinite width neural nets” are Gaussian processes

Proposition (Jacot, Gabriel, Hongler)

For a network as above, at initialization, in the limit as $n_1, ..., n_{L-1} \to \infty$, the output functions $f_{\theta,k}$, for $k = 1, ..., n_L$, tend (in law) to iid centered Gaussian processes of covariance $\Sigma^{(L)}$, where $\Sigma^{(L)}$ is defined recursively by:

$$\Sigma^{(1)}(x, x') = \frac{1}{n_0} x^T x' + \beta^2$$

$$\Sigma^{(L+1)}(x, x') = \mathbb{E}_{f \sim \mathcal{N}\left(0, \Sigma^{(L)}\right)}[\sigma(f(x))\sigma(f(x'))] + \beta^2,$$

taking the expectation with respect to a centered Gaussian process f of covariance $\Sigma^{(L)}$.

Proof.

Follows by induction on layers, using the law of large numbers. $\square$

Theorem (Jacot, Gabriel, Hongler)

For a network as above, at initialization, in the limit as the layers width $n_1, ..., n_{L-1} \to \infty$, the NTK $\Theta^{(L)}$ converges in probability to a deterministic limiting kernel:

$$\Theta^{(L)} \to \Theta_\infty^{(L)}.$$

The scalar kernel $\Theta_\infty^{(L)} : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}$ is defined recursively by

$$\begin{aligned} \Theta_\infty^{(1)}(x, x') &= \Sigma^{(1)}(x, x') \\ \Theta_\infty^{(L+1)}(x, x') &= \Theta_\infty^{(L)}(x, x')\dot{\Sigma}^{(L+1)}(x, x') + \Sigma^{(L+1)}(x, x'), \end{aligned}$$

where

$$\dot{\Sigma}^{(L+1)}\left(x, x'\right) = \mathbb{E}_{f \sim \mathcal{N}\left(0, \Sigma^{(L)}\right)}\left[\dot{\sigma}\left(f\left(x\right)\right)\dot{\sigma}\left(f\left(x'\right)\right)\right].$$

Proof.

Follows by induction on layers, using the law of large numbers. □

Asymptotics during training

Theorem

With the same assumptions as before and with σ being Lipschitz. During training, as $n_1, \ldots, n_L \to \infty$, we have uniformly in $t \in [0, T]$

$$\Theta^{(L)} \to \Theta_{\infty}^{(L)}.$$

Natural Gradient Descent

One of a family of algorithms with update rule

$$\theta_{t+1} = \theta_t - \eta_t G(\theta_t)^{-1} \nabla_{\theta_t} L(\theta).$$

This is the:

- standard gradient descent if $G = I$,
- Gauss-Newton method if $G(\theta_t) = \nabla_\theta^2 L(\theta) = H(\theta)$, the Hessian matrix,
- saddle-free Newton method (SFN), if $G(\theta) = |H(\theta)|$, where

$$H(\theta) = O^T \mathrm{diag}(\lambda_1, \ldots, \lambda_r) O, \quad |H(\theta)| = O^T \mathrm{diag}(|\lambda_1|, \ldots, |\lambda_r|) O.$$

Natural Gradient Descent - Fisher information matrix

Suppose we are given a model $p(x|\theta)$ parameterized by θ. The fisher information matrix is

$$F = \mathop{\mathbb{E}}_{p(x|\theta)} \left[\nabla \log p(x|\theta)^{\mathsf{T}} \, \nabla \log p(x|\theta) \right] .$$

In practice, we use an empirical distribution given by training data points $\{x_1, \ldots, x_N\}$. The empirical FIM is

$$F = \frac{1}{N} \sum_{i=1}^{N} \nabla \log p(x_i|\theta)^{\mathsf{T}} \, \nabla \log p(x_i|\theta) .$$

Natural Gradient Descent

The natural gradient descent is given by

$$\theta_{t+1} = \theta_t - \eta_t F(\theta_t)^{-1} \nabla_{\theta_t} L(\theta),$$

where $F(\theta)$ is the Fisher information matrix.

Motivation

One can regard the distribution space $\{p_\theta | \theta\}$ as a Riemannian manifold in which the metric tensor is given by the Fisher information matrix. Then it can be proven that $F(\theta_t)^{-1} \nabla_{\theta_t} L(\theta)$ is the steepest descent in the space $\{p_\theta | \theta\}$.

Advantages

Fisher efficient (Amari (1998)) + saddle points free (Rattray, Saad and Amari (1998), and Rattray and Saad (2000)).

NTK for Natural Gradient Descent

The dynamics of the natural gradient descent

$$\partial_t f_{\theta(t)}(X) = -\eta \nabla_\theta f_{\theta(t)}(X) F(\theta)^{-1} \nabla_\theta f_{\theta(t)}(X)^T \nabla_f L(X, Y; f_\theta).$$

Define

$$\Theta(X, X) = \nabla_\theta f_{\theta(t)}(X) F(\theta)^{-1} \nabla_\theta f_{\theta(t)}(X)^T,$$

so that

$$\partial_t f_{\theta(t)}(X) = -\eta \Theta(X, X) \nabla_f L(X, Y; f_\theta).$$

The Fisher information matrix of a DNN with Gaussian noise

The FIM of a DNN is

$$F(\theta|x) = \mathbb{E}_{p(y|x)}\left[\left(\nabla_\theta \log p(y, f_\theta(x))\right)^T \nabla_\theta \log p(y, f_\theta(x))\right].$$

In practice, we use the empirical FIM obtained by assuming the empirical distribution of the data

$$\widetilde{F}(\theta) = \frac{1}{n}\left(\nabla_\theta f_\theta(X)\right)^T E_{p(y|X)}\left[\left(\nabla_f \log p(y, f_\theta(X))\right)^T \nabla_f \log p(y, f_\theta(X))\right] \nabla_\theta f_\theta(X).$$

If p is Gaussian with variance σ^2, then

$$\widetilde{F} = \frac{1}{n\sigma^2}\left(\nabla_\theta f_\theta(X)\right)^T \nabla_\theta f_\theta(X)$$

Natural NTK with Gaussian noise

Assumption: Assume that the DNN is overparameterized, that is, $nk \leq p$.

$\Rightarrow$ there is a high probability the rows of $\nabla_\theta f_\theta(X)$ are linearly independent

$\Rightarrow$ $\widetilde{F}(\theta)$ has a pseudoinverse (that satisfies $\widetilde{F}(\theta)\widetilde{F}(\theta)^+ = I$)

$$\widetilde{F}(\theta)^+ = n\sigma^2 \nabla_\theta f_\theta(X)^+ \left(\nabla_\theta f_\theta(X)^+\right)^T.$$

The empirical NTK is

$$\Theta(X, X) = \nabla_\theta f_\theta(X) \widetilde{F}(\theta)^+ \nabla_\theta f_\theta(X)^T = n\sigma^2 I_{nk}.$$

Exact solution for natural NTK

Thus

$$\partial_t f_{\theta(t)}(X) = -\eta n \sigma^2 I_{nk} \nabla_f L(X, Y; f_\theta).$$

If $L(X, Y; f_\theta)$ is MSE, then

$$f_{\theta(t)}(X) = f_{\theta^*}(X) + \exp\left(-t\eta n \sigma^2\right)\left(f_{\theta(0)}(X) - f_{\theta^*}(X)\right).$$

Towards NTK in RL

The dissimilarity function

$$\mathrm{KL}(q\|p) = \mathbb{E}_q\left[\log q - \log p\right].$$

For $p \sim \mathcal{N}(\mu_1, \Sigma_1), q \sim \mathcal{N}(\mu_2, \Sigma_2)$, we have

$$\begin{aligned}\mathrm{KL}(p\|q) &= \int p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx \\ &= \frac{1}{2}\left(\log\frac{\det\Sigma_2}{\det\Sigma_1} - n + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)\right).\end{aligned}$$

Σ is constant

The Fisher information matrix, with constant Σ, is

$$F(\theta) = \left(\frac{\partial \mu_2}{\partial \theta}\right)^T \Sigma^{-1} \left(\frac{\partial \mu_2}{\partial \theta}\right).$$

Assume $\frac{\partial \mu}{\partial \theta}$ has linearly independent rows, then $F(\theta)$ has a pseudoinverse $F(\theta)^+ = \left(\frac{\partial \mu_2}{\partial \theta}\right)^+ \Sigma \left(\frac{\partial \mu_2}{\partial \theta}\right)^{T+}$. Then the (pseudo) natural gradient is

$$\begin{aligned}\frac{d\theta}{dt} &= -\eta F(\theta)^+ \left(\frac{\partial \mu_2}{\partial \theta}\right)^T \Sigma^{-1}(\mu_2 - \mu_1) \\ &= -\eta \left(\frac{\partial \mu_2}{\partial \theta}\right)^+ \Sigma \left(\frac{\partial \mu_2}{\partial \theta}\right)^{T+} \left(\frac{\partial \mu}{\partial \theta}\right)^T \Sigma^{-1}(\mu_2 - \mu_1) \\ &= -\eta \left(\frac{\partial \mu}{\partial \theta}\right)^+ (\mu_2 - \mu_1).\end{aligned}$$

The dynamics

$$\frac{\partial \mu_2}{\partial t} = -\eta(\mu_2 - \mu_1) \Rightarrow \mu(\theta_t) = \exp\left(-\eta t\right)\left(\mu_2(\theta_0) - \mu_1(\theta_0)\right).$$

Remark

The NTK is still a scalar multiple of the identity matrix in this case.