1 Abstract

At initialization, artificial neural networks (ANNs) are equivalent to Gaussian processes in the infinite-width limit (16; 4; 7; 13; 6), thus connecting them to kernel methods. We prove that the evolution of an ANN during training can also be described by a kernel: during gradient descent on the parameters of an ANN, the network function fθ (which maps input vectors to output vectors) follows the kernel gradient of the functional cost (which is convex, in contrast to the parameter cost) w.r.t. a new kernel: the Neural Tangent Kernel (NTK). This kernel is central to describe the generalization features of ANNs. While the NTK is random at
initialization and varies during training, in the infinite-width limit it converges to an explicit limiting kernel and it stays constant during training. This makes it possible to study the training of ANNs in function space instead of parameter space. Convergence of the training can then be related to the positive-definiteness of the
limiting NTK. We prove the positive-definiteness of the limiting NTK when the data is supported on the sphere and the non-linearity is non-polynomial. We then focus on the setting of least-squares regression and show that in the infinitewidth limit, the network function fθ follows a linear differential equation during training. The convergence is fastest along the largest kernel principal components of the input data with respect to the NTK, hence suggesting a theoretical motivation for early stopping. Finally we study the NTK numerically, observe its behavior for wide networks, and compare it to the infinite-width limit.

초기화 시점에 인공 신경망(ANN)은 무한 너비 극한(infinite-width limit)에서 가우시안 프로세스(Gaussian processes)와 동일하며, 이를 통해 신경망은 커널 방법론과 연결됩니다. 본 논문은 학습 중 ANN의 진화 역시 커널로 설명될 수 있음을 증명합니다. ANN의 매개변수에 대한 경사 하강법이 진행되는 동안, 입력 벡터를 출력 벡터로 매핑하는 네트워크 함수 f_{θ}는 새로운 커널인 신경 접전 커널(Neural Tangent Kernel, NTK)에 대한 함수 비용(매개변수 비용과 달리 볼록함)의 커널 경사를 따릅니다. 이 커널은 ANN의 일반화 특성을 설명하는 데 핵심적인 역할을 합니다. NTK는 초기화 시점에는 무작위적이고 학습 과정에서 변화하지만, 무한 너비 극한에서는 명시적인 극한 커널(limiting kernel)로 수렴하며 학습 내내 상수로 유지됩니다. 이를 통해 매개변수 공간이 아닌 함수 공간(function space)에서 ANN의 학습을 연구하는 것이 가능해집니다. 학습의 수렴은 극한 NTK의 양의 정치성(positive-definiteness)과 연관될 수 있습니다. 본 논문은 데이터가 구(sphere) 위에 존재하고 비선형성이 비다항식인 경우, 극한 NTK의 양의 정치성을 증명합니다. 이어 최소 제곱 회귀(least-squares regression) 설정을 중점적으로 다루며, 무한 너비 극한에서 네트워크 함수 f_θ가 학습 중에 선형 미분 방정식을 따른다는 것을 보여줍니다. 수렴은 NTK에 대한 입력 데이터의 가장 큰 커널 주성분을 따라 가장 빠르게 일어나며, 이는 조기 종료(early stopping)에 대한 이론적 동기를 제시합니다. 마지막으로 NTK를 수치적으로 연구하여 넓은 네트워크에서의 동작을 관찰하고, 이를 무한 너비 극한과 비교합니다.

2 정리

NTK는 다음과 같이 두 데이터에서 gradient의 내적으로 정의된다. (일종의 gram matrix)

$$\Theta(x, x') = \nabla_\theta f(\theta, x) \cdot \nabla_\theta f(\theta, x')$$

충분히 넓고 얕은 모델은 gradient가 거의 변하지 않기 때문에 NTK도 거의 변하지 않는다.

모델은 일종의 kernel method처럼 일정한 형태를 유지한 채로 선형적으로 학습하게 된다.

NTK를 일정하게 근사할 수 있는 조건

$$y(w) \approx y(w_0) + \nabla_w y(w_0)^T (w - w_0)$$

w_0로 부터의 학습은 이와 같이 1차 테일러 근사로 표현할 수 있다.

1차 근사와 실제 학습 결과의 차이가 충분히 작다면, NTK가 일정하고 선형적으로 학습한다고 볼 수 있다.

즉, 테일러 근사가 성립하기 위해서는 가중치의 **Jacobian의 변화가 적어야 한다**.

$$\text{net change in } y(w) \lesssim ||y(w_0) - \bar{y}||$$

학습으로 인한 출력 값의 변화는 초기 출력과 ground trouth와의 차이보다는 작을 것이다.

이를 통해 가중치의 변화량 d를 다음과 같이 근사할 수 있다.

$$d = \frac{y(w)\text{의 변화량}}{w\text{에 대한 } y\text{의 변화량}} \lesssim \frac{||y(w_0) - \bar{y}||}{||\nabla_w y(w_0)||}$$

Jacobian의 변화량은 $d \cdot ||\nabla_w^2 y(w_0)||$로 나타낼 수 있고, Jacobian의 변화율 $\kappa(w_0)$를 나타내보면

$$\kappa(w_0) \triangleq \frac{d \cdot \left\|\nabla_w^2 y(w_0)\right\|}{\left\|\nabla_w y(w_0)\right\|} = \|y(w_0) - \bar{y}\| \frac{\left\|\nabla_w^2 y(w_0)\right\|}{\left\|\nabla_w y(w_0)\right\|^2}$$

따라서 변화율 κ가 충분히 작다면 모델의 학습을 선형적으로 나타낼 수 있다.

$$\kappa(w_0) = \|y(w_0) - \bar{y}\| \frac{\left\|\nabla_w^2 y(w_0)\right\|}{\left\|\nabla_w y(w_0)\right\|^2} \ll 1$$

Guide Proceedings On lazy training in differentiable programming | Proceedings…에서는 $\kappa(w_0)$를 줄이는 방법으로 모델의 output에 α를 곱하는 방법을 제시하였다.

$$\kappa_\alpha(w_0) = \|\alpha y(w_0) - \bar{y}\| \frac{\left\|\alpha\nabla_w^2 y(w_0)\right\|}{\left\|\alpha\nabla_w y(w_0)\right\|^2} \sim \frac{1}{\alpha} \frac{\left\|\nabla_w^2 y(w_0)\right\|}{\left\|\nabla_w y(w_0)\right\|^2}$$

3 Conclusion

4 참고자료

EigenTales Understanding the Neural Tangent Kernel → 대부분 여기서 가져옴