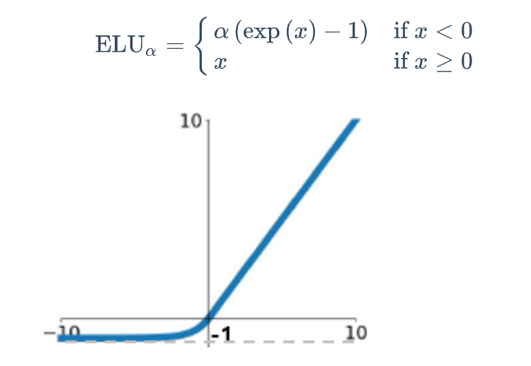
*From https://excelsior-cjh.tistory.com/177?category=940400*

1. **Activation function**

* Sigmoid
* tanh
* ReLU
* Leaky ReLU
* ELU:



(x < 0) 평균 -> 0이므로 bias shift감소 => gradient 손실 줄여줌

(x > 0) 에서 기울기 != 0

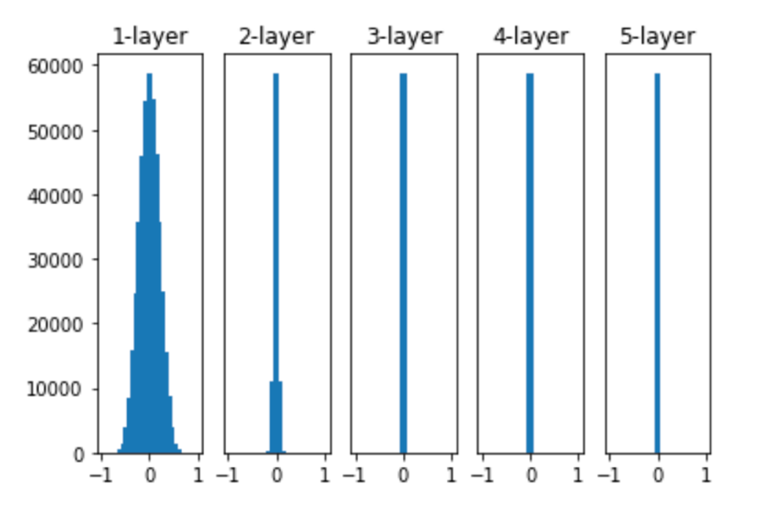
(Alpha = 1) x=0에서 급격히 변하지 않아 gradient descent에서 속도 빠름

1. **Weight initialization**

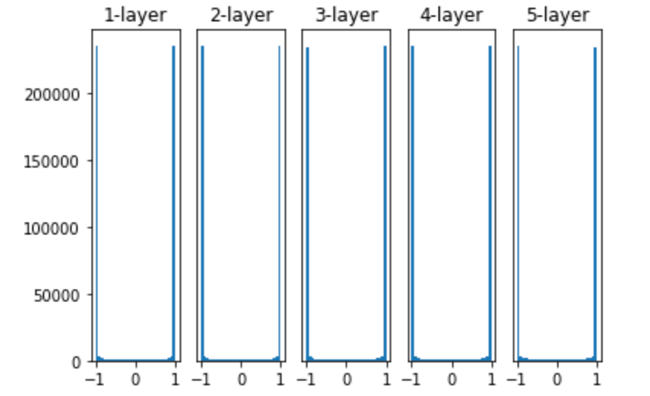
* Weight = 0 or w1=w2=….
* Small random numbers:

*N(0, 0.01^2)*으로 초기화 -> shallow할 땐 괜찮을지 몰라도 deep할 땐 문제발생

Tanh사용하면 첫번째 hidden layer를 제외한 나머지 layer들이 모두 0을 출력하게 됨

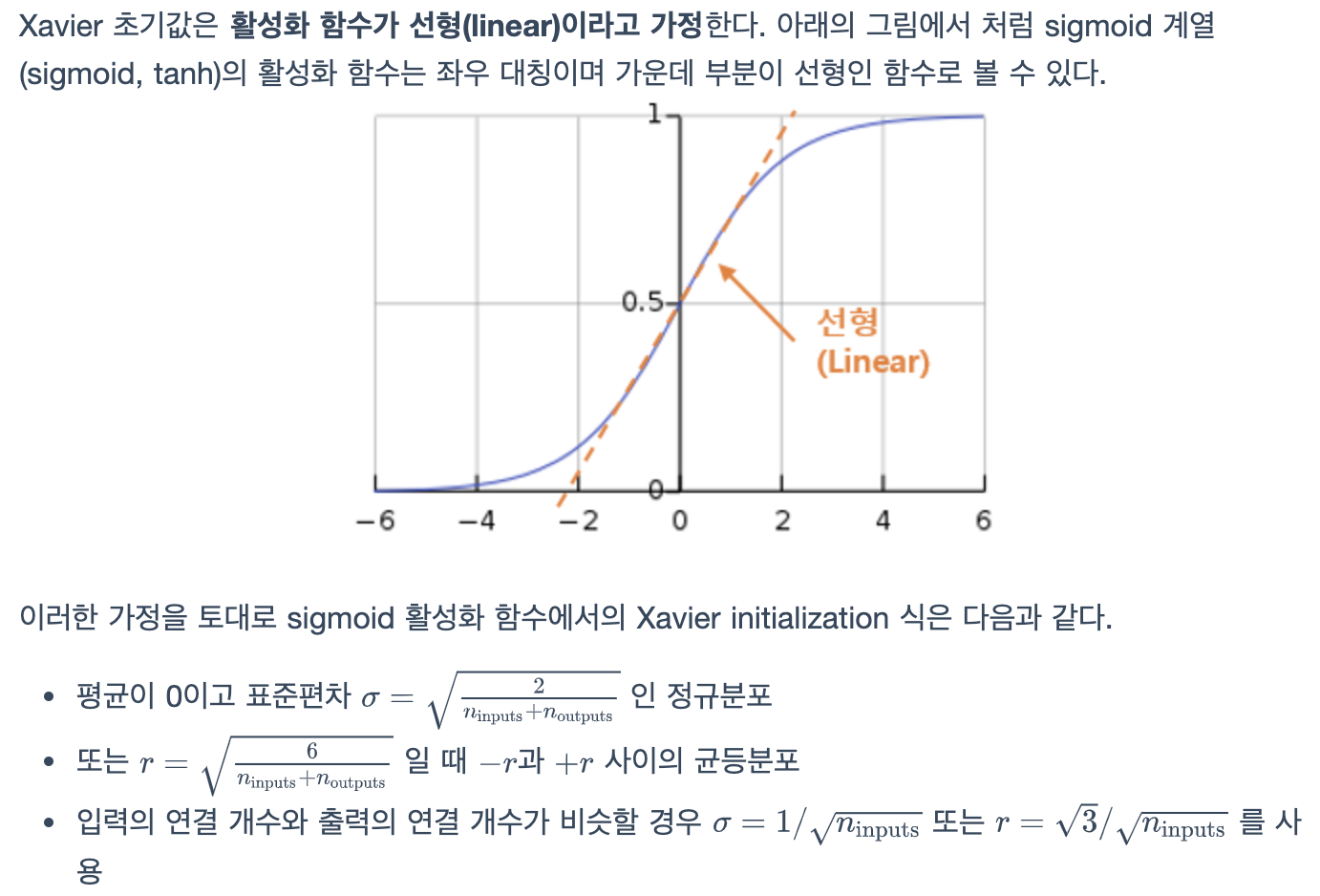
-> 미분 값 같아서 학습 X

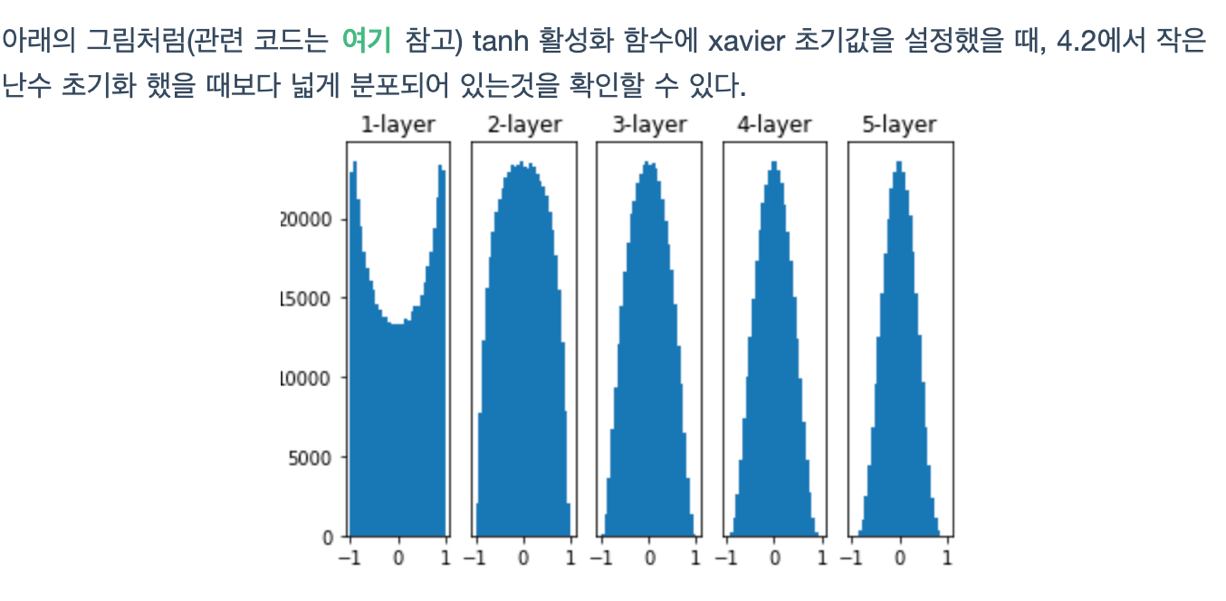
*N(0, 1)*으로 초기화 -> tanh 사용하면 vanishing gradient



* Xavier initialization

적절한 데이터가 흐르기 위한 조건: **각 층의 출력에 대한 분산 = 입력에 대한 분산**





* He initialization

