NOISY ACTIVATION FUNCTIONS

CAGLAR GULCEHRE*, MARCIN MOCZULSKI†, MISHA DENIL†, YOSHUA BENGIO*

*University of Montreal, †University of Oxford

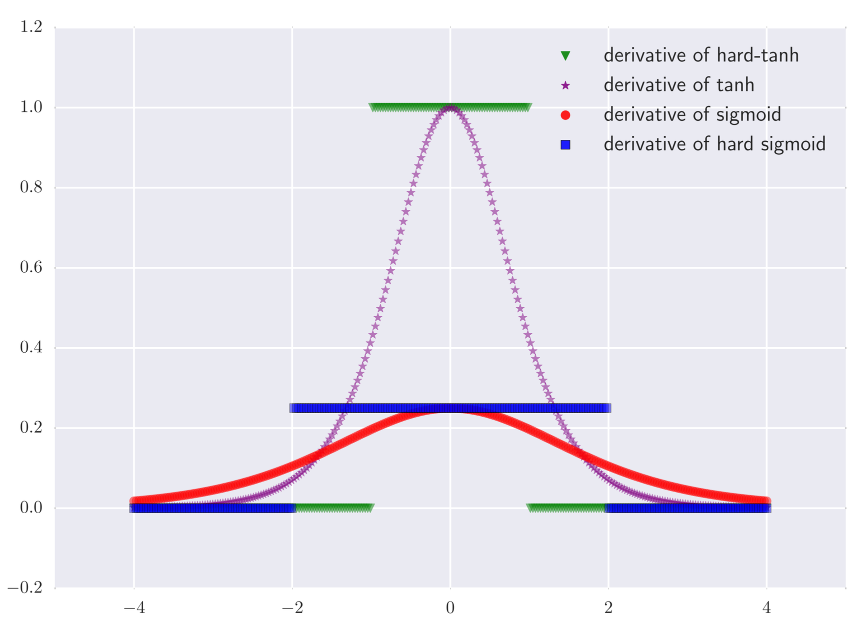
MOTIVATION

- Common nonlinear activation functions can have training difficulties.
- Logistic functions (sigmoid and tanh) can be difficult to train.
- Piecewise linear activation functions (i.e. ReLU) are easier to optimize.

OUR CONTRIBUTIONS

- Applying piecewise linear activation functions to gates of the recurrent models, i.e. LSTMs.
- Investigation of injecting noise to the activations.
- An efficient way to learn the std of noise for each unit.
- Annealing the activation noise can have a continuation effect.

SATURATING ACTIVATIONS



Definition 1. *Soft-saturating activation: An activation function softly saturates if it converges to a particular value as $x \to \infty$ and/or $x \to -\infty$.*

Definition 2. *Hard-saturating activation: An activation function hardly saturates if it becomes constant when its input gets larger than a threshold c.*

Linearize the activation function and clip it at the threshold:

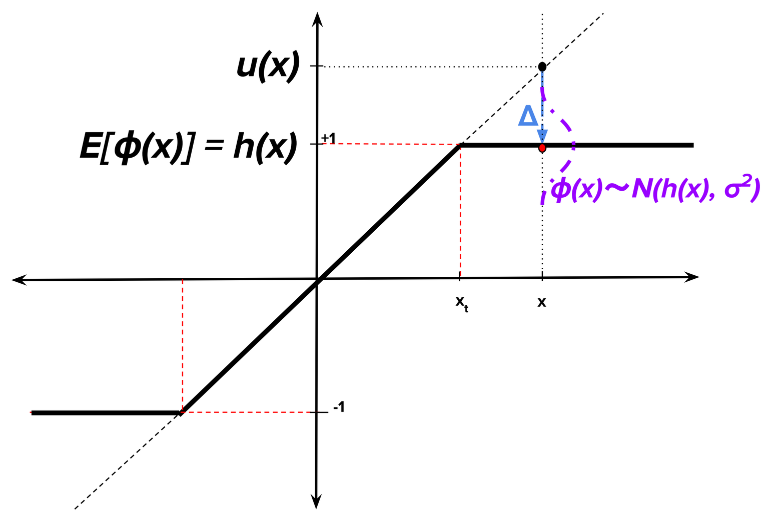
$$\text{hard-sigmoid}(x) = \max(\min(\mathrm{u}^s(x),\ 1),\ 0)$$
$$\text{hard-tanh}(x) = \max(\min(\mathrm{u}^t(x),\ 1),\ -1)$$

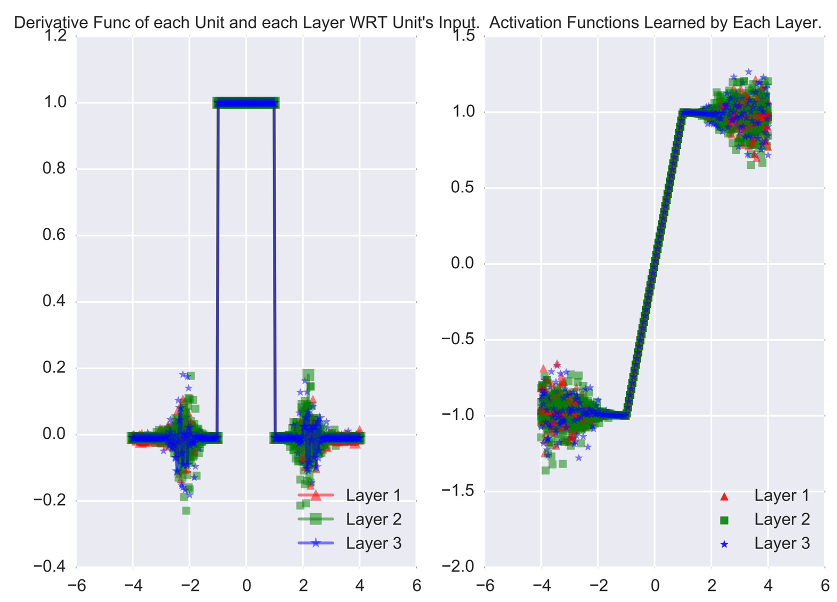
NOISY ACTIVATIONS(UNBIASED)

$\mathrm{h}(x)$: hard activation function.
$\mathrm{u}(x)$: soft activation function.

$$\phi(x, \xi) = \mathrm{u}(x) + s$$
$$s = \mu + \sigma\xi$$
$$\mathrm{E}_{\xi \sim \mathcal{N}(0,\ 1)} \approx \mathrm{h}(x)$$

NOISY ACTIVATIONS (BIASED)





Injecting Biased Noise:

$$\mathrm{d}(x) = -\mathrm{sgn}(x)\mathrm{sgn}(1-\alpha)$$
For $\epsilon = |\xi|$,
$$s = \mu(x) + \mathrm{d}(x)\sigma(x)\epsilon,$$
$$\phi(x,\ \xi) = \alpha\mathrm{h}(x) + (1-\alpha)\mathrm{u}(x) + \mathrm{d}(x)\sigma(x)\epsilon.$$

Use the expectation of the noise at the test time:

$$\mathrm{E}[\phi(x,\ \xi)] = \alpha\mathrm{h}(x) + (1-\alpha)\mathrm{u}(x) + \mathrm{d}(x)\sigma(x)\mathrm{E}[\epsilon].$$

Injecting Noise at the Input:
Noise injection to the input of the activation function can be written as:

$$\phi(x,\ \xi) = \mathrm{h}(x + s).$$

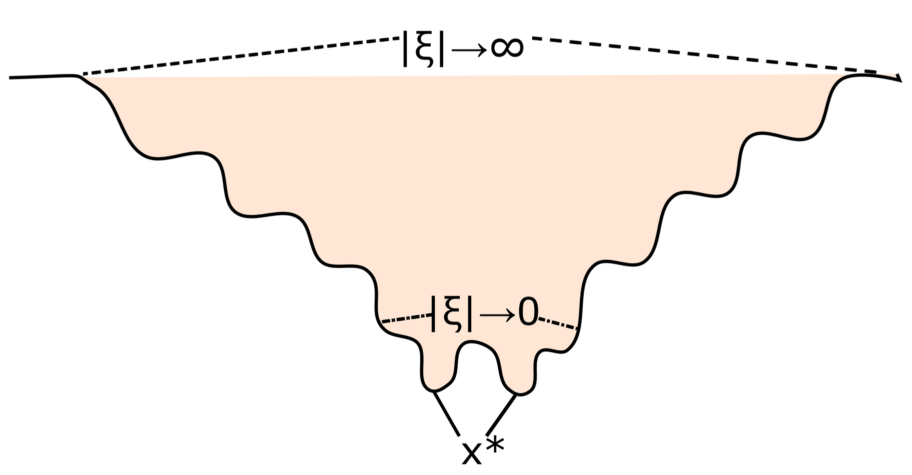
and s can have different formulations, e.g. $s = \sigma\xi$, $s = \sigma(x)\xi$ or $s = \mathbf{1}_{|x| \geq |x_t|}(\sigma\xi)$

ANNEALING THE NOISE

Start with large noise resulting in larger exploration and anneal the noise.

$$\lim_{|\xi| \to \infty} \left|\frac{\partial \phi(x, \xi)}{\partial x}\right| \to \infty.$$

A pathological one-dimensional case for SGD:



EXPERIMENTS

Used the same hyperparameters with baselines.

- **NAN** - NormAl Noise at the output.
- **NAH** - Half-NormAl (biased) Noise at the output.
- **NANI** - NormAl Noise at the Input.
- **NANIL** - NormAl Noise with Learned $\sigma(x)$ at the Input.
- **NANIS** - NormAl Noise at the Input when unit Saturates.

Neural Machine Translation

	Valid nll	BLEU
Sigmoid and Tanh NMT (Reference)	65.26	20.18
Hard-Tanh and Hard-Sigmoid NMT	64.27	21.59
Noisy (**NAH**) Tanh and Sigmoid NMT	**63.46**	**22.57**

Learning to Execute

Image Caption Generation

*Image caption generation is both without dropout.

	BLEU -1	BLEU-2	BLEU-3	BLEU-4	METEOR	Test NLL
Soft (Ref.)	**67**	44.8	29.9	19.5	18.9	40.33
Soft (**NAH**)	66	**45.8**	**30.69**	**20.9**	**20.5**	40.17
Soft (**NAH***)	64.9	44.2	**30.7**	**20.9**	20.3	**39.8**
Soft (**NANI**)	66	45.0	30.6	20.7	**20.5**	40.0
Soft (**NANIL**)	66	44.6	30.1	20.0	**20.5**	39.9
Hard	67	45.7	31.4	21.3	19.5	-

PennTreeBank Experiments

	Valid ppl	Test ppl
Noisy LSTM + **NAN**	111.7	**108.0**
Noisy LSTM + **NAH**	112.6	**108.7**
LSTM (Reference)	119.4	115.6

Annealing Experiments

	Test Error %
LSTM+MLP(Reference)	33.28
Noisy LSTM+MLP(**NAN**)	31.12
Curriculum LSTM+MLP	14.83
Noisy LSTM+MLP(**NAN**) Annealed Noise	**9.53**
Noisy LSTM+MLP(**NANIL**) Annealed Noise	20.94

NTM Experiments