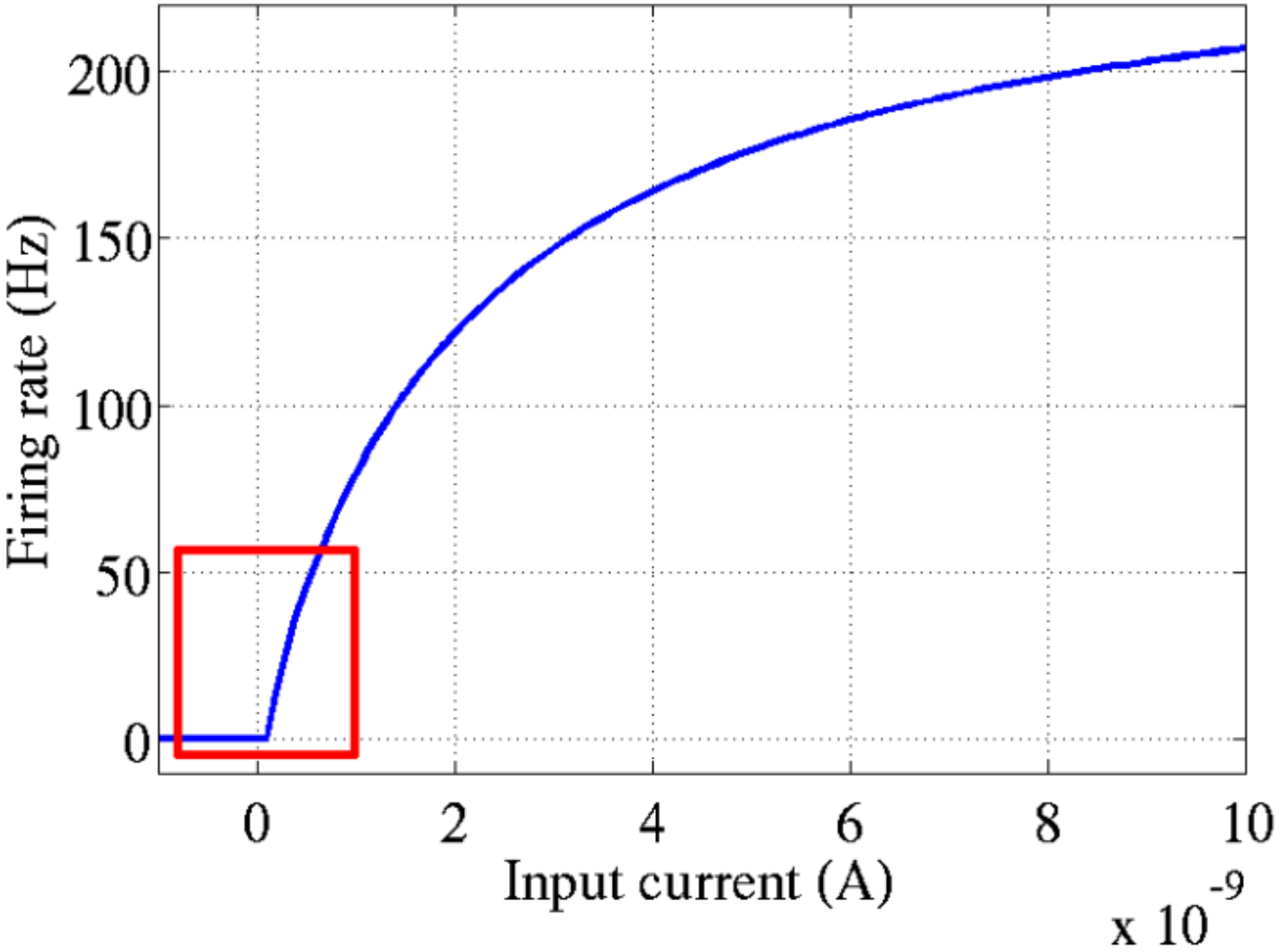
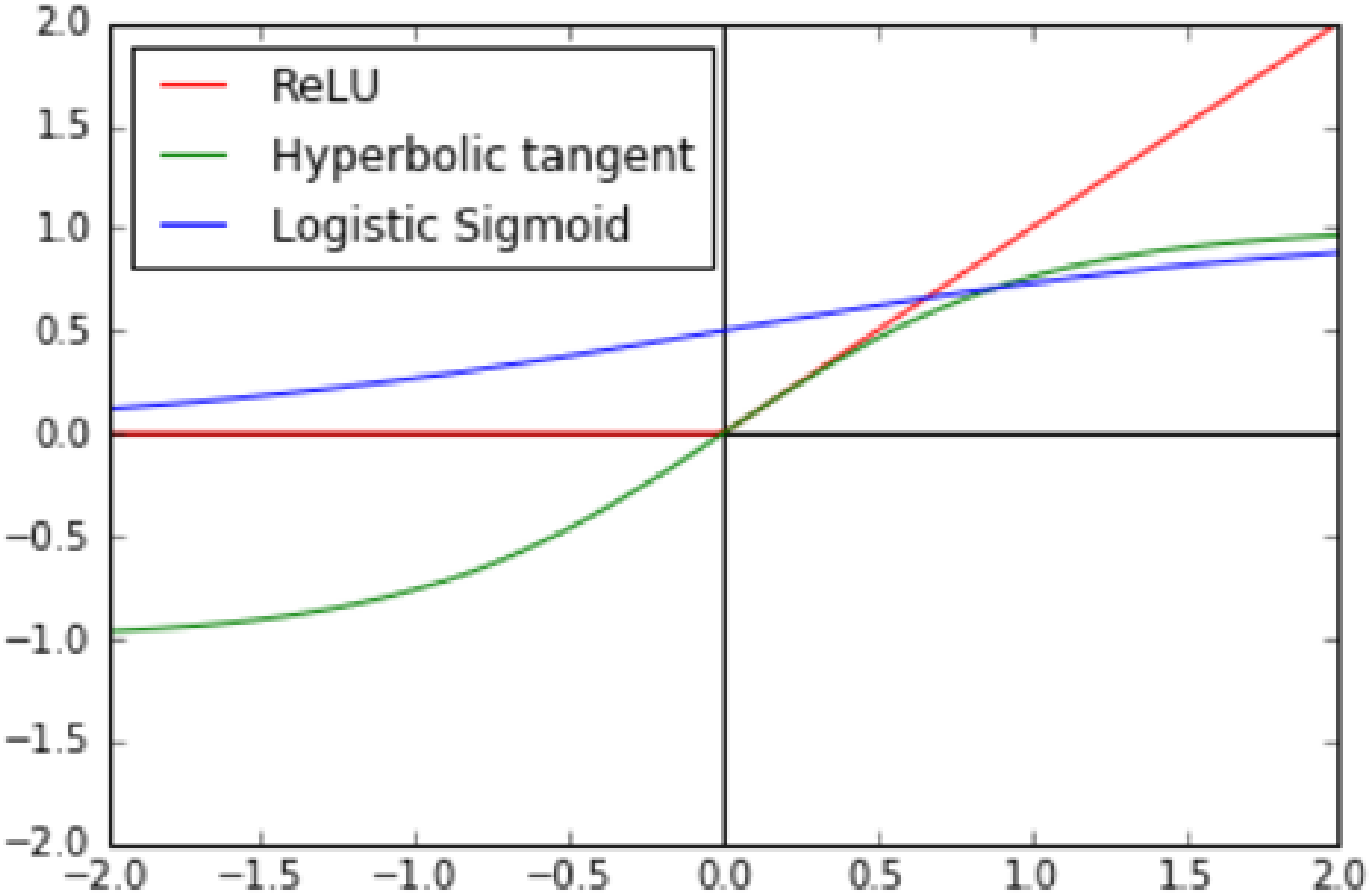
Why ReLU?

Radek Bartyzal
rbartyzal1@gmail.com
17. 8. 2016
Let’s talk ML in Prague

Biological inspiration



Activation functions



Logistic Sigmoid

- non-symmetric
- bounded
- more biologically accurate than Tanh
  - Positive values only
- when initialized with small weights output is around 0.5

  = not biologically accurate

  = hurts gradient-based optimization

Hyperbolic Tangent

- anti-symmetric
- bounded
- faster backprop convergence than Sigmoid
  - due to large gradient around 0
- works well even though the forced antisymmetry around 0 is absent in biological neurons

Rectified Linear Unit = ReLU

Pros:

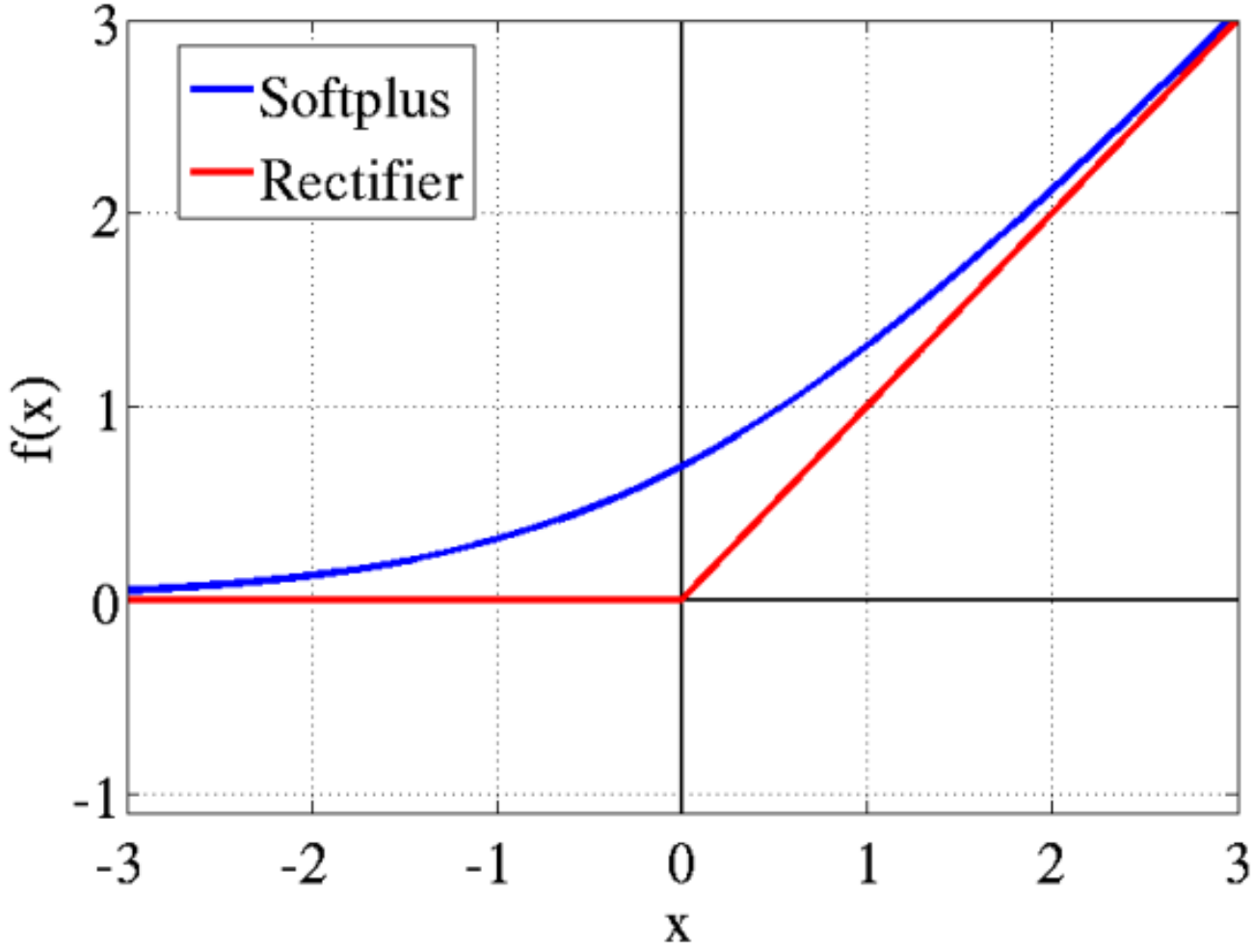
- most biologically accurate of the three mentioned ones
- allows true zeros
- leads to truly sparse networks
- computationally cheaper than exponential functions

Rectified Linear Unit = ReLU

Cons:

- too much sparsity may hurt predictive performance = reduced capacity of model
  - from 85% + of true zeros
- unbounded
  - use regularizer to prevent potential numerical problems
- not symmetric around 0
  - multiply half of the units output values by -1
- true zeros could hurt backprop
  - they don’t: equal or better results than Softplus

Rectified Linear Unit = ReLU



Results

- stacked denoising auto-encoders
- three hidden layers, 1000 units per layer

Neuron	MNIST	CIFAR10	NISTP	NORB
With **unsupervised pre-training**				
Rectifier	**1.20%**	**49.96%**	**32.86%**	**16.46%**
Tanh	**1.16%**	**50.79%**	35.89%	17.66%
Softplus	**1.17%**	**49.52%**	**33.27%**	19.19%
Without **unsupervised pre-training**				
Rectifier	**1.43%**	**50.86%**	**32.64%**	**16.40%**
Tanh	1.57%	52.62%	36.46%	19.29%
Softplus	1.77%	53.20%	35.48%	17.68%

Conlusion

- biologically credible
- almost no improvement when using unsupervised pre-training, contrary to tanh or softplus.
- rectifier networks are truly deep sparse networks
  - average sparsity of hidden layers = 50 – 80%
  - brain hypothetical sparsity = 95 – 99%
- great for image classification
- awesome for sentiment analysis

Sources

- Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "**Deep Sparse Rectifier Neural Networks**." *Aistats*. Vol. 15. No. 106. 2011.
- LeCun, Yann, Ido Kanter, and Sara A. Solla. **"Second order properties of error surfaces: Learning time and generalization."** Advances in neural information processing systems 3 (1991): 918-924.