

What We Do (not?) Know About Quantization

Sergey Salishev

How Engineers See Quantization

$$f = q + r,$$

then

$$q = f - r$$

- Here f is analog value, q is quantized value, r is quantization noise (residue, error)
- In some systems while error is small everything works OK
- If error is critically large the system breaks down
- How large?
 - It is empirically checked

The Big Question

- Consider discrete feedforward networks
 - Digital Artificial NN
 - Digital Circuits, e.g. multiplier
- Artificial NN usually tolerate some noise, while Digital Circuits do not
- What is the difference?

Information Theory View on NN

- Latent space \rightarrow External Encoder \rightarrow NN Decoder \rightarrow Latent Space \rightarrow NN Encoder \rightarrow NN Decoder \rightarrow Latent Space
 - E.g. Meaning \rightarrow Human \rightarrow English text \rightarrow NN Decoder \rightarrow Meaning \rightarrow NN Encoder \rightarrow French text
- NN Decoder and Encoder are Shannon Channels
 - Assume External Encoder is lossless
 - As long as $R < C$ there is the code (Weights?) with $\sim 0\%$ probability of data decoding error
 - Here R is transfer rate, C is channel capacity
 - Moreover, almost any random code is good enough (2nd Shannon Theorem)
- $R = H(L)$, L is latent space, H is Entropy (1st Shannon Theorem)
- Why no one uses random codes?
 - Latent space is unknown
 - Overfitting to training data
 - Exponential worst case training complexity
- Still lots of randomness
 - Random weight initialization
 - Random training permutation

Why Does Gradient Descent Even Work on NN?

$$y = f_q(\Theta, x)$$
$$\Theta = g(x_0, y_0)$$

From information theory f_q is a purely discrete object, BUT we also expect robustness of training and stability
We ignore generalization for now

$$y + \Delta y = f(\Theta + \Delta\Theta, x + \Delta x)$$
$$\Theta + \Delta\Theta = g(x_0 + \Delta x_0, y_0 + \Delta y_0)$$

If $\Delta x_0, \Delta y_0, \Delta x \ll 1$ then $\Delta y \ll 1$

With limit conditions this means Limited Variations

Can be replaced with Differentiability almost everywhere (stronger)

$$f_a(x) = f_q(x) + r(x)$$

Here $r(x)$ is continuous differentiable regularization term

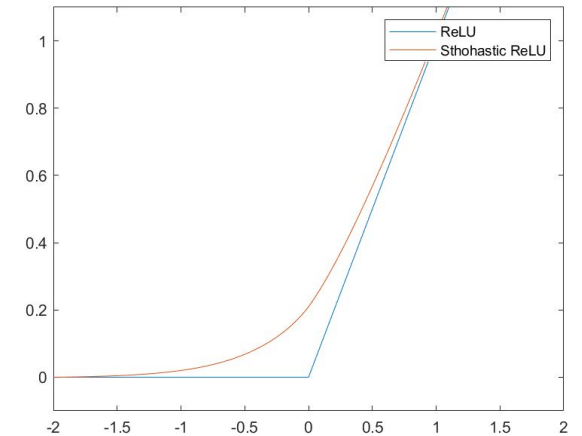
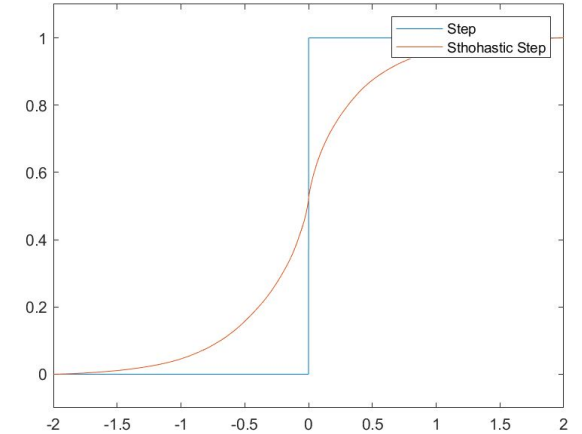
Stochastic Regularization

- What $r(x)$ term might be?
- An example

$$f_1(x) = c \int_{-\infty}^x E \left(\frac{f_0(t + n) - f_0(t)}{n} \right) dt, \quad n \sim N(0,1)$$

Here f_0 is discrete nonlinearity (step, ReLU, etc)

- Replace $f_0 \rightarrow f_1$
- Noisy training + Computation errors + Mini Batches + SGD
= Some Stochastic Regularization



Analog Meets Digital

- Analog NN have infinite capacity (good)
 - Infinitely complex digital implementation (bad)
- True quantization problem
 - Select discrete NN weights to make $\mathcal{C} = R + \varepsilon$, ε is small
 - Means lossless quantization as Latent space is not changed
- How to change analog channel capacity?
 - Just add some noise (3rd Shannon Theorem)

Back to Engineering

$$q = f - \alpha r, \quad E|r|^2 = 1$$

- Just optimize α with SGD along with other NN parameters

$$I_\alpha = \log_2 \frac{E|f|^2}{\alpha|r|^2} = \log_2 E|f|^2 - \log_2 \alpha$$

I_α is Shannon SNR in bits

- If $I_\alpha > 0$ (some signal left) it is quantization
- If $I_\alpha = 0$ (no signal) it is pruning

Open Questions

- Can we deduce differentiability almost everywhere by formalizing generalization?
- Is there a closed form for stochastic regularization depending on the noise distribution?
- What are sufficient conditions of convergence of differentiable stochastic quantization?
- What are sufficient conditions of convergence of stochastic approximation of stochastic quantization with quantization noise only?