Neural Networks as Model Selection with MDL Normalization

BAIHAN LIN

Center for Theoretical Neuroscience, Columbia University, New York, NY 10027, USA
Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA
Department of Applied Mathematics, University of Washington, Seattle, WA 98105, USA

Abstract

If we consider the neural network optimization process as a model selection problem, the implicit space can be constrained by the normalizing factor, the minimum description length of the optimal universal code. Inspired by the adaptation phenomenon of biological neuronal firing, we propose a class of reparameterization of the activation in the neural network that take into account the statistical regularity in the implicit space under the Minimum Description Length (MDL) principle. We introduce an incremental version of computing this universal code as normalized maximum likelihood and demonstrated its flexibility to include data prior such as top-down attention and other oracle information and its compatibility to be incorporated into batch normalization and layer normalization. The empirical results showed that the proposed method outperforms existing normalization methods in tackling the limited and imbalanced data from a non-stationary distribution benchmarked on computer vision and reinforcement learning tasks. As an unsupervised attention mechanism given input data, this biologically plausible normalization has the potential to deal with other complicated real-world scenarios as well as reinforcement learning setting where the rewards are sparse and non-uniform. Further research is proposed to discover these scenarios and explore the behaviors among variants.

MDL, Optimal Codes, NML

Among the optimal universal codes, NML probability minimizes the worst-case regret:

$$P_{NML}(x) = \frac{P(x|\hat{\theta}(x))}{\sum_{x'} P(x'|\hat{\theta}(x'))} \tag{1}$$

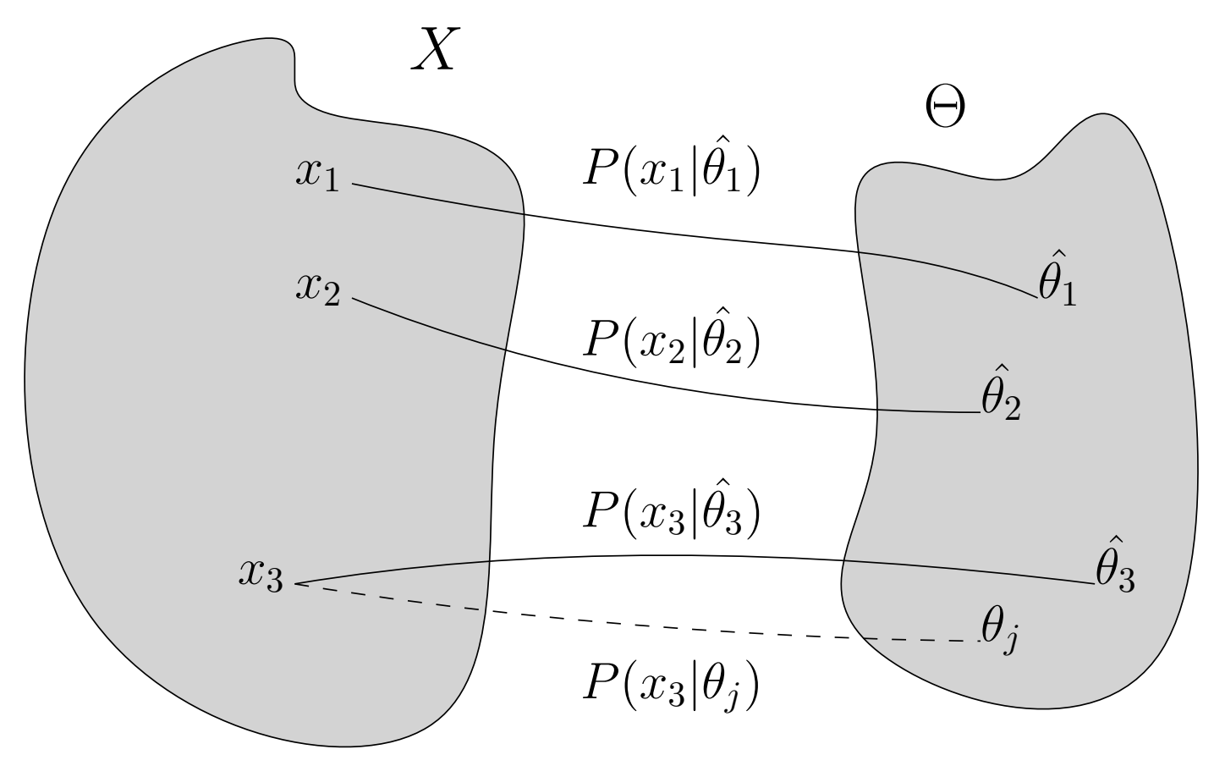

Figure 1: Normalized maximal likelihood. Data sample x_i are drawn from the entire data distribution X and model $\hat{\theta}_i$ is the optimal model that describes data x_i with the shortest code length.

Regularity Normalization

```
Algorithm 1 Regularity Normalization (RN)
Input: Values of $x$ over a mini-batch: $\mathcal{B} = \{x_{1,\cdots,m}\}$;
Parameter: $COMP_t$, $\hat{\theta}_t$
Output: $y_i = RN(x_i)$

$COMP_{t+1}$ = increment($COMP_t$, $P(x_i|\hat{\theta}_t(x_i))$)
$L_{x_i} = COMP_{t+1} - \log P(x_i|\hat{\theta}_t(x_i))$
$y_i = L_{x_i} * x_i$
```

RN can also include a data prior function, $s(x)$ as variant **Saliency Normalization**:

$$P_{NML}(x) = \frac{s(x)P(x|\hat{\theta}(x))}{\sum_{x'} s(x')P(x'|\hat{\theta}(x'))} \tag{2}$$

RL Problem with DQN

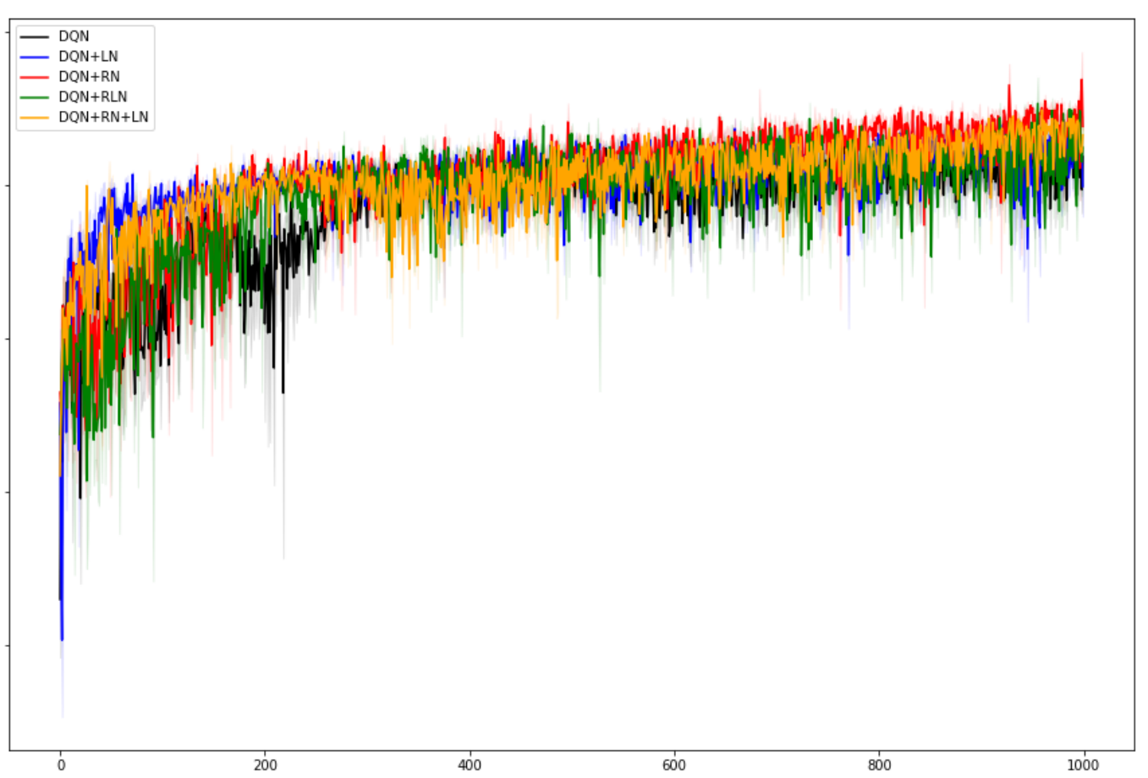

Figure 2: Learning in LunarLander-V2. By the averaged final scores over 1000 episodes, DQN+RN (76.95 ± 4.44) performs the best, followed by DQN+RN+LN (65.82 ± 10.91) and DQN+RLN (49.27 ± 40.35). All three proposed agents beat DQN (37.17 ± 8.82) and DQN-LN (-1.54 ± 39.14) by a large marginal.

Neural networks as model selection

We generalize the optimal universal code with NML formulation as:

$$P_{NML}(x_i) = \frac{P(x_i|\hat{\theta}_i(x_i))}{\sum_{j=0}^{i} P(x_j|\hat{\theta}_j(x_j))} \tag{3}$$

where $\hat{\theta}_i(x_i)$ refers to the model parameter already optimized for $i-1$ steps and have seen sequential data sample x_0 through x_{i-1}. This distribution is updated every time a new data sample is given, and can therefore be computed incrementally, as in batch-based training.

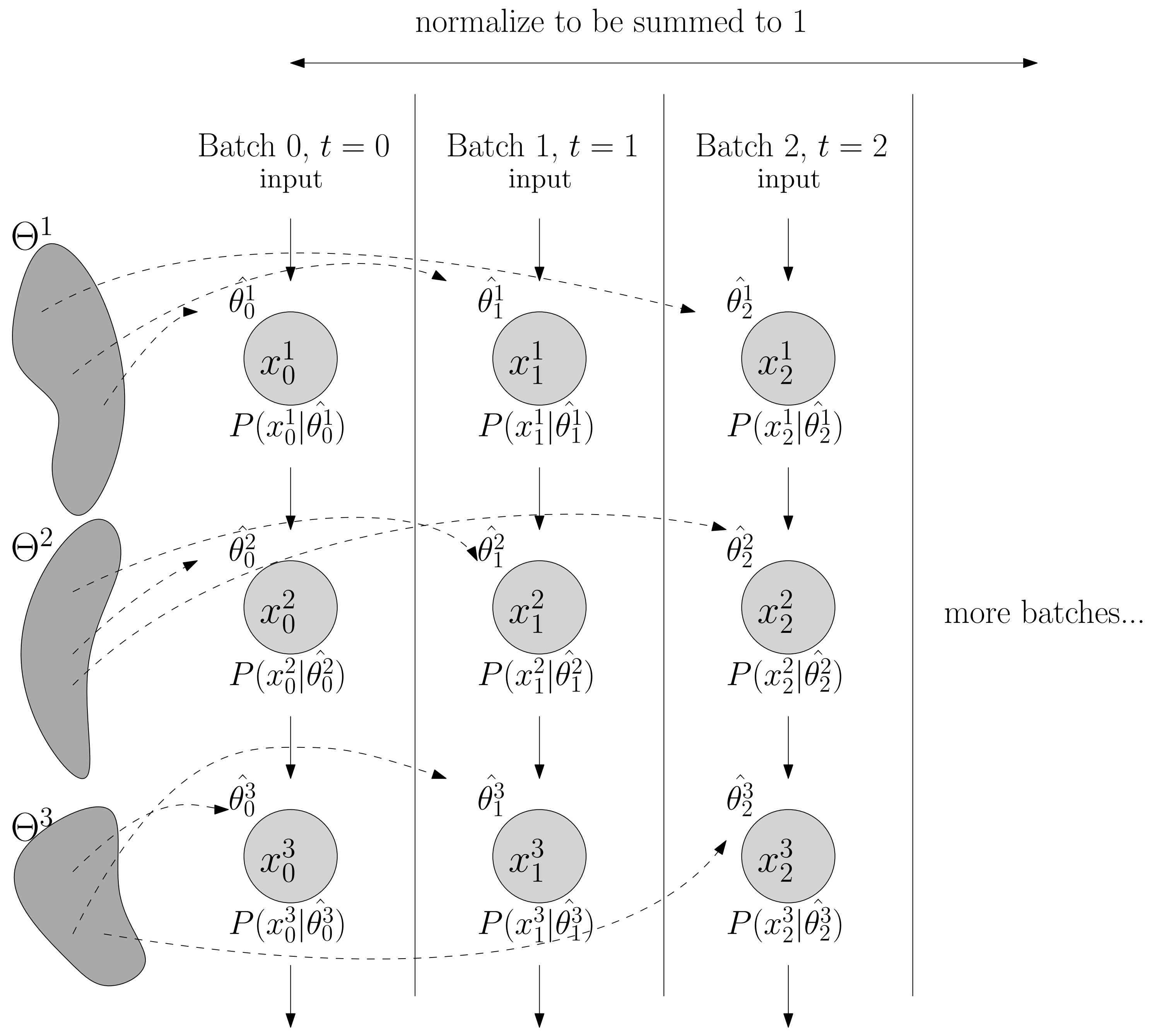

Figure 3: Consider each optimization step as the process of choose the optimal model from model class Θ^i for ith layer of the neural networks, the optimized parameter $\hat{\theta}_j^i$ with subscript j as time step $t=j$ and superscript i as layer i can be assumed to be the optimal model among all models in the model class Θ^i. The normalized maximum likelihood can be computed by choosing $P(x_j^i|\hat{\theta}_j^i)$, the "optimal" model with shortest code length given data x_j^i, as the summing component in the normalization.

Imbalanced MNIST Problem with FFNN

	"Balanced"	"Rare minority"			"Highly imbalanced"				"Dominant oligarchy"	
	$n=0$	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$
baseline	4.80 ± 0.15	14.48 ± 0.28	23.74 ± 0.28	32.80 ± 0.22	42.01 ± 0.45	51.99 ± 0.32	60.86 ± 0.19	70.81 ± 0.40	80.67 ± 0.36	90.12 ± 0.25
BN	**2.77 ± 0.05**	12.54 ± 0.30	21.77 ± 0.25	30.75 ± 0.30	40.67 ± 0.45	49.96 ± 0.46	59.08 ± 0.70	67.25 ± 0.54	76.55 ± 1.41	80.54 ± 2.38
LN	3.09 ± 0.11	8.78 ± 0.84	14.22 ± 0.65	20.62 ± 1.46	26.87 ± 0.97	34.23 ± 2.08	36.87 ± 0.64	41.73 ± 2.74	**41.20 ± 1.13**	**41.26 ± 1.30**
WN	4.96 ± 0.11	14.51 ± 0.44	23.72 ± 0.39	32.99 ± 0.28	41.95 ± 0.46	52.10 ± 0.30	60.97 ± 0.18	70.87 ± 0.39	80.76 ± 0.36	90.12 ± 0.25
RN	4.91 ± 0.39	8.61 ± 0.86	14.61 ± 0.58	19.49 ± 0.45	**23.35 ± 1.22**	33.84 ± 1.69	41.47 ± 1.91	60.46 ± 2.88	81.96 ± 0.59	90.11 ± 0.24
RLN	5.01 ± 0.29	9.47 ± 1.21	**12.32 ± 0.56**	22.17 ± 0.94	23.76 ± 1.56	32.23 ± 1.66	43.06 ± 3.56	57.30 ± 6.33	88.36 ± 1.77	89.55 ± 0.32
LN+RN	4.59 ± 0.29	**8.41 ± 1.16**	12.46 ± 0.87	**17.25 ± 1.47**	25.65 ± 1.91	**28.71 ± 1.97**	**33.14 ± 2.49**	**36.08 ± 2.09**	44.54 ± 1.74	82.29 ± 4.44
SN	7.00 ± 0.18	12.27 ± 1.30	16.12 ± 1.39	24.91 ± 1.61	31.07 ± 1.41	41.87 ± 1.78	52.88 ± 2.09	68.44 ± 1.42	83.34 ± 1.85	82.41 ± 2.30

Figure 4: Test errors of the imbalanced permutation-invariant MNIST 784-1000-1000-10 task

Supervised + unsupervised, top-down + bottom-up attention

In concept, the regularity-based normalization can also be considered as an unsupervised attention mechanism imposed on the input data, with the flexibility to directly install top-down attention from either oracle supervision or other meta information.