Output-Constrained Bayesian Neural Networks

Wanqian Yang*, Lars Lorch*, Moritz A. Graule*, Srivatsan Srinivasan, Anirudh Suresh, Jiayu Yao, Melanie F. Pradier, Finale Doshi-Velez

Harvard University

*equal contribution

Presented at ICML 2019 Workshop on Uncertainty & Robustness in Deep Learning.

Motivation: A general method to impose human-interpretable (function space) constraints on BNNs

Main idea:

$$p_{\mathcal{C}}(\mathcal{W}) := p(\mathcal{W})\, g(\mathcal{W} \mid \mathcal{C})$$

"standard" Gaussian prior

Our contribution:

$g(\mathcal{W} \mid \mathcal{C})$

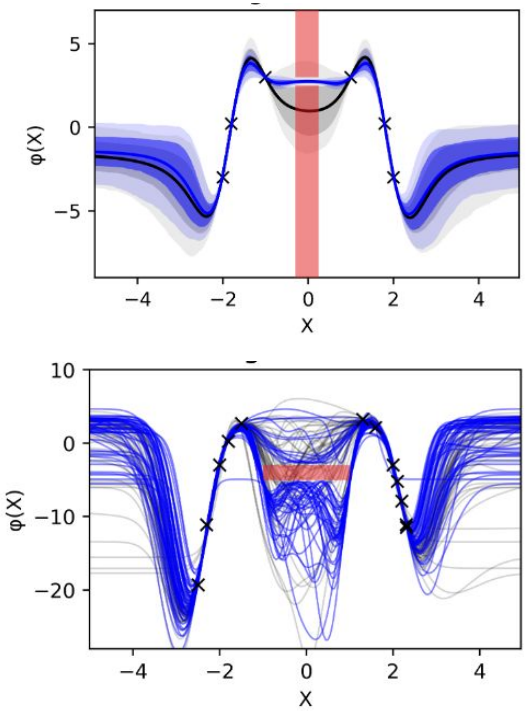
samples points in constrained region

"positive" formulation: rewards output mass within desired region (in function space)

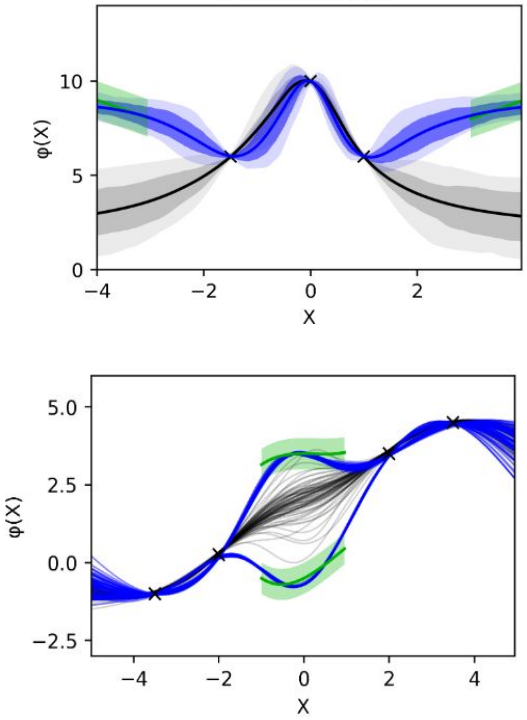
"negative" formulation: penalizes output mass within region to avoid (in function space)

Examples

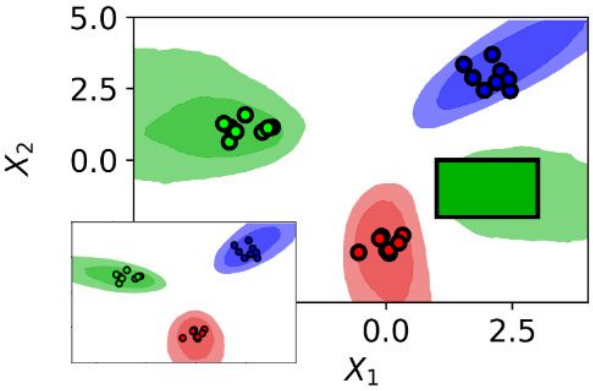
Negative constraint priors
"avoid red region"



Positive constraint priors
"pass through green region"



Classification
with positive constraint prior



gray/black/inset: standard prior
blue: constraint prior

Come visit us at our poster!