Collapsed Inference for Bayesian Deep Learning

Zhe Zeng, Guy Van den Broeck
{zhezeng, guyvdb}@cs.ucla.edu

Paper Link

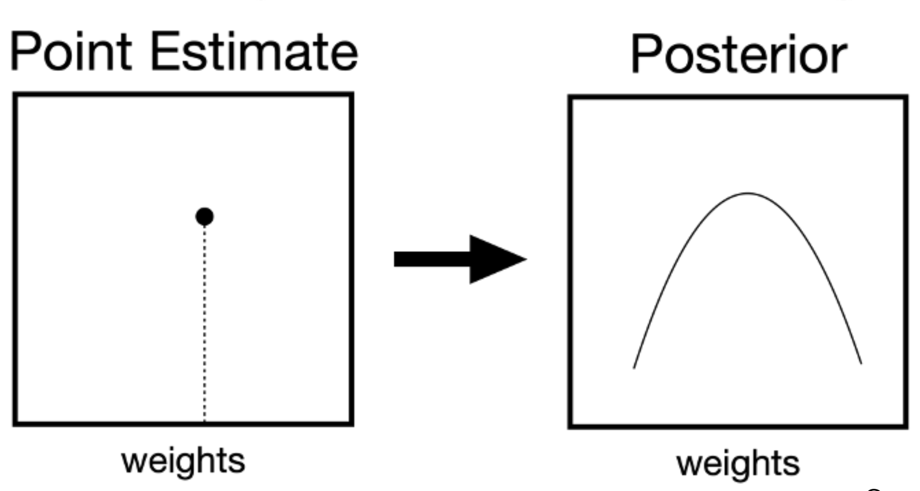
Bayesian Deep Learning

Motivations

— Neural networks are bad at uncertainty estimation.
— It can be risky to estimate using one set of weights

Thus, *Bayesian deep learning* is proposed as a framework for incorporating ***uncertainty*** into deep learning models. By treating neural network weights as random variables, it allows for more ***robust* and *reliable* predictions**.

Goal: Bayesian model average (BMA)



$p(y \mid x, w)$ $\qquad$ $p(y \mid x) = \int p(y \mid x, w) p(w)\, dw$

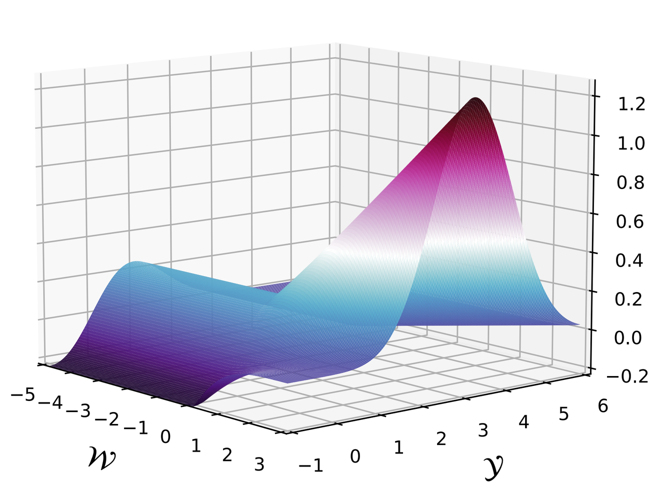
— Expected Prediction: $\mathbb{E}[y] = \int y\ p(y \mid x)\, dy$

Complex Integration

Challenges

DNNs are too big for sampling-based methods!
— It can be costly to maintain too many samples.
— Sample efficiency is low given the **complex** integrand.

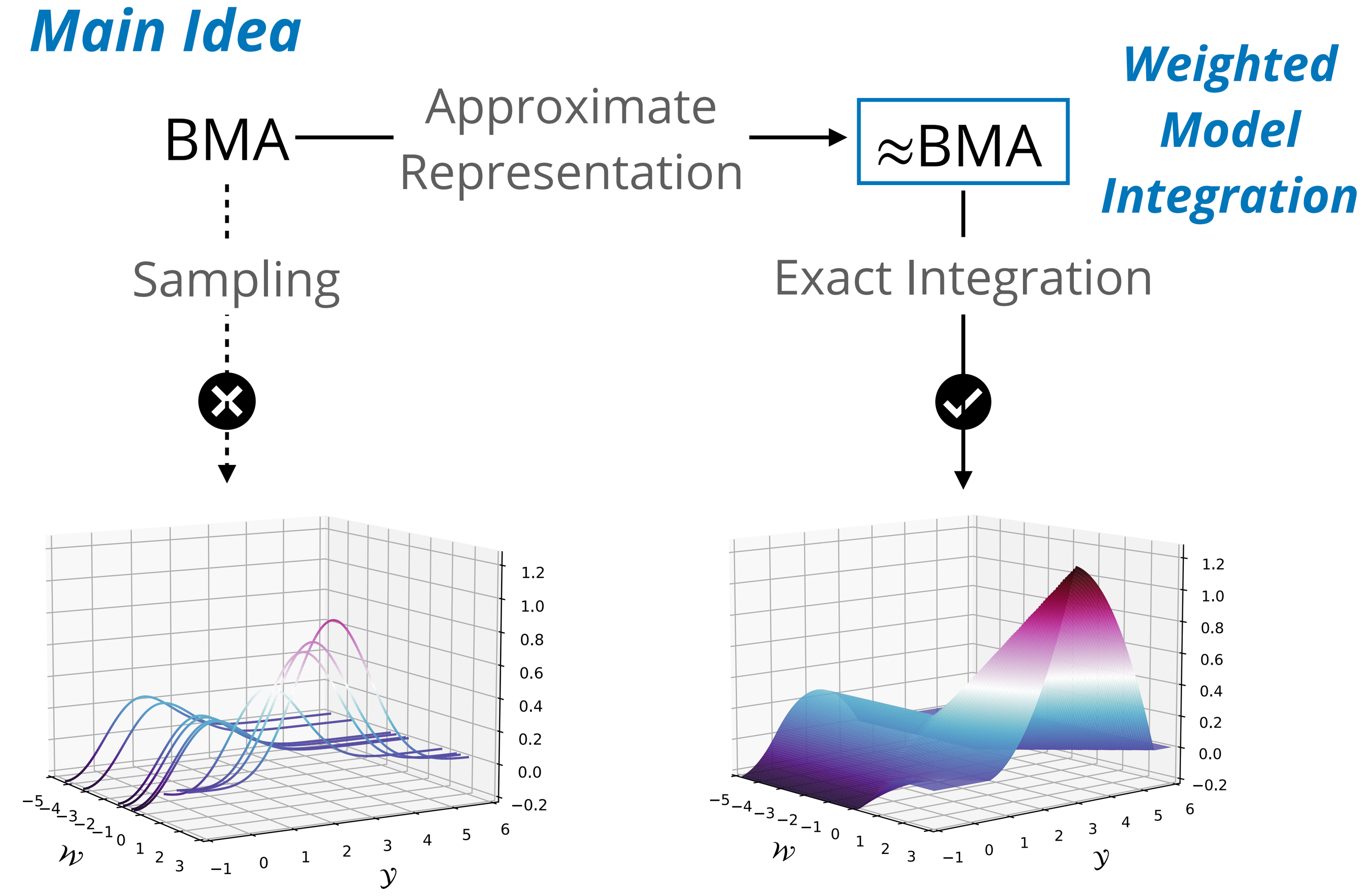
$$\mathbb{E}[y] = \int y\ \underbrace{p(w \mid D)}_{\text{Uniform}}\ \underbrace{p(y \mid \underbrace{f(x)}_{\text{Single ReLU}}, w)}_{\text{Gaussian}}\, dw\, dy$$



Non-convex, multi-modal, no closed form 🤯

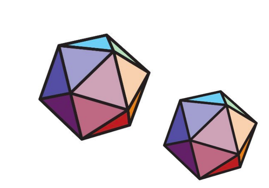
BMA as Volume Computation

Main Idea



where ***Weighted Model Integration*** (***WMI***) is a class of *weighted volume computation* problems:
— Regions: logical combination of arithmetic constraints
— Polynomial Weights: $\phi : ⬚ \to \mathbb{R}$

*Why **WMI**?*
— *Existing WMI solvers are able to give **exact marginalization** results*

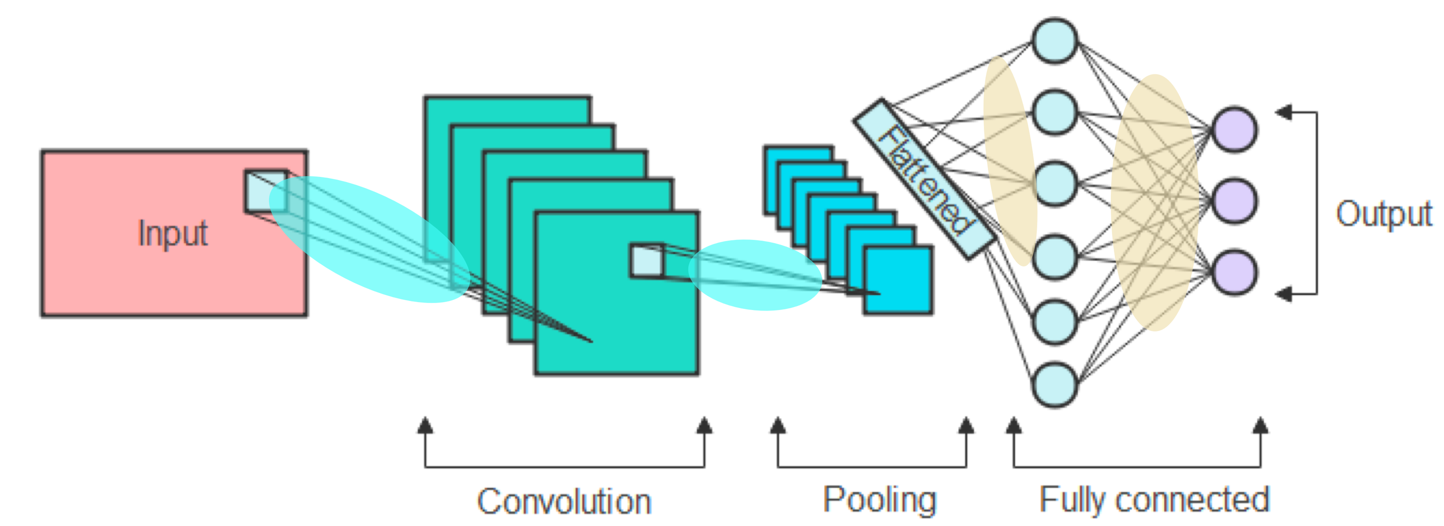


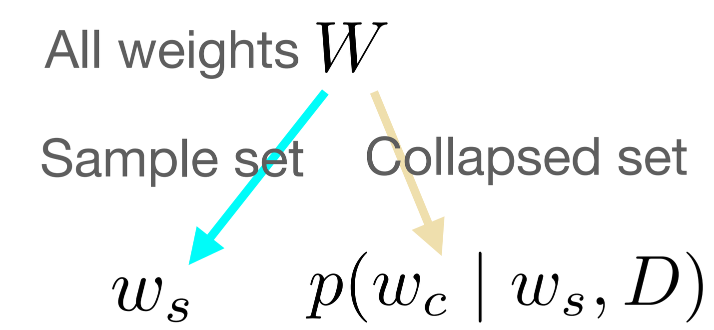
Collapsed Inference

A Scheme to Combine Good from Both Worlds!

	Sampling	BMA via WMI	Collapsed Inference
Accuracy	❌*	✅	✅
Flexibility	✅	❌**	✅
Scalability	✅	❌***	✅

*** Hard to estimate using a few samples**
**** Limited to fully connected layers**
***** Integration over polytopes in arbitrarily high dimensions is #P-hard**





Expected prediction in BMA

$$\mathbb{E}[y] = \frac{1}{n} \sum_{w_s} \mathtt{WMI}(\ \cdot\ )$$

Accuracy* + *Flexibility, Scalability! 🥰

Experiments

Regression

	Elevators	KeggD	KeggU	Protein	Skillcraft	Pol
CIBER (second)	-0.378 ± 0.026	**1.245 ± 0.090**	**1.125 ± 0.269**	-0.720 ± 0.036	-1.003 ± 0.035	**2.555 ± 0.115**
CIBER (last)	-0.371 ± 0.023	1.178 ± 0.088	0.964 ± 0.231	-0.720 ± 0.036	**-1.001 ± 0.032**	2.506 ± 0.150
SWAG	-0.374 ± 0.021	1.080 ± 0.035	0.749 ± 0.029	**-0.700 ± 0.051**	-1.180 ± 0.033	1.533 ± 1.084
PCA+ESS (SI)	-0.351 ± 0.030	1.074 ± 0.034	0.752 ± 0.025	-0.734 ± 0.063	-1.181 ± 0.033	-0.185 ± 2.779
PCA+VI (SI)	**-0.325 ± 0.019**	1.085 ± 0.031	0.757 ± 0.028	-0.712 ± 0.057	-1.179 ± 0.033	1.764 ± 0.271
SGD	-0.538 ± 0.108	1.012 ± 0.154	0.602 ± 0.224	-0.854 ± 0.085	-1.162 ± 0.032	1.073 ± 0.858
OrthVGP	-0.448	1.022	0.701	-0.914	—	0.159
NL	-0.698 ± 0.039	0.935 ± 0.265	0.670 ± 0.038	-0.884 ± 0.025	-1.002 ± 0.050	-2.840 ± 0.226

Image Classification

Metric	NLL		ACC		ECE	
Dataset	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100
CIBER	**0.1927 ± 0.0029**	**0.9193 ± 0.0027**	**93.64 ± 0.09**	**74.71 ± 0.18**	0.0130 ± 0.0011	**0.0168 ± 0.0025**
SWAG	0.2503 ± 0.0081	1.2785 ± 0.0031	93.59 ± 0.14	73.85 ± 0.25	0.0391 ± 0.0020	0.1535 ± 0.0015
SGD	0.3285 ± 0.0139	1.7308 ± 0.0137	93.17 ± 0.14	73.15 ± 0.11	0.0483 ± 0.0022	0.1870 ± 0.0014
SWA	0.2621 ± 0.0104	1.2780 ± 0.0051	93.61 ± 0.11	74.30 ± 0.22	0.0408 ± 0.0019	0.1514 ± 0.0032
SGLD	0.2001 ± 0.0059	0.9699 ± 0.0057	93.55 ± 0.15	74.02 ± 0.30	**0.0082 ± 0.0012**	0.0424 ± 0.0029
KFAC	0.2252 ± 0.0032	1.1915 ± 0.0199	92.65 ± 0.20	72.38 ± 0.23	0.0094 ± 0.0005	0.0778 ± 0.0054

Transfer Learning

Metric	NLL	ACC	ECE
CIBER	**0.9869 ± 0.0102**	**72.56 ± 0.23**	**0.0925 ± 0.0028**
SWAG	1.3425 ± 0.0015	72.30 ± 0.11	0.1988 ± 0.0028
SWA	1.3993 ± 0.0502	71.92 ± 0.01	0.2082 ± 0.0056
SGD	1.6528 ± 0.0390	72.42 ± 0.07	0.2149 ± 0.0027

Takeaway

Volume Computation Solvers + Statistical ML

=

Reliable & Scalable Inference!