Collapsed inference for Bayesian deep learning

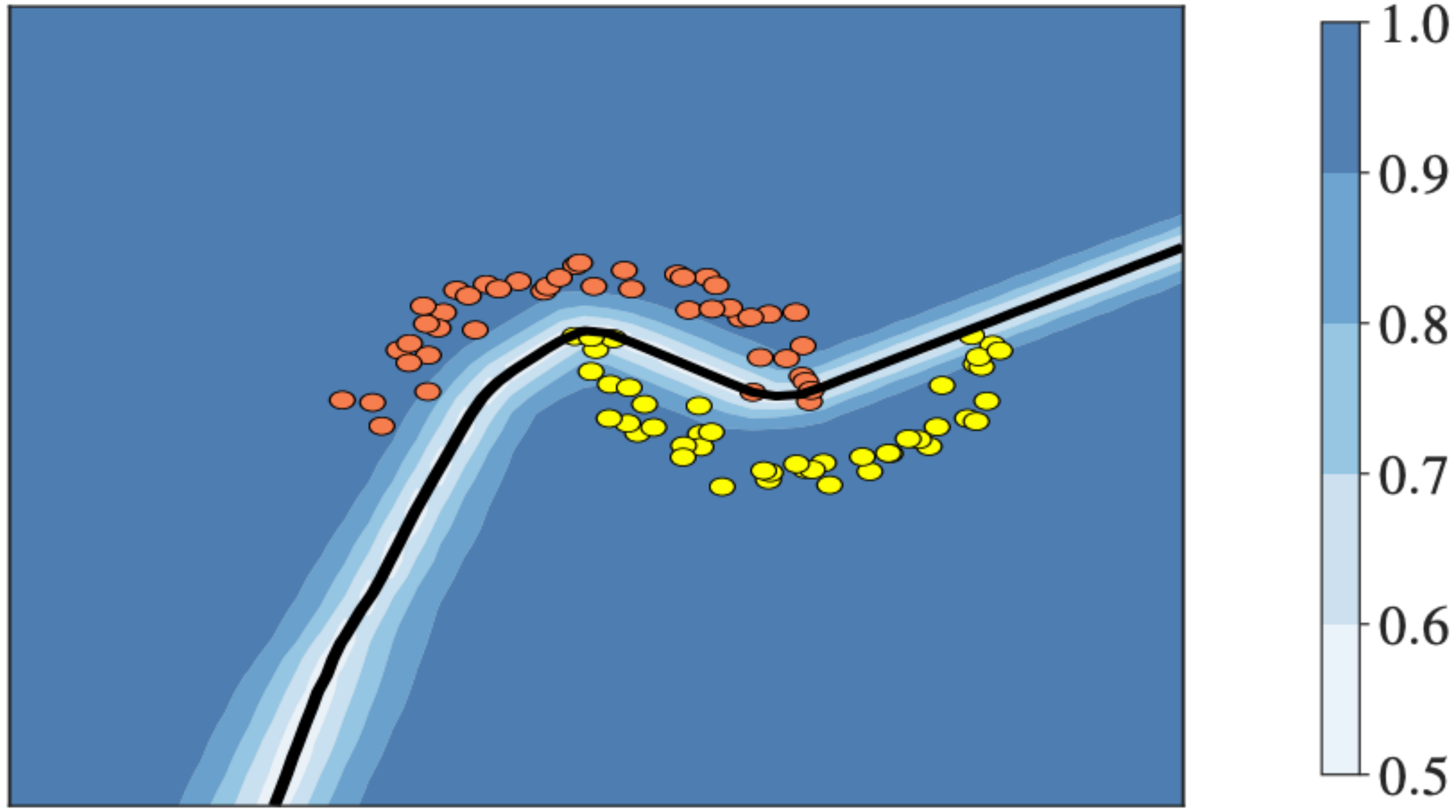
Zhe Zeng, Guy Van den Broeck

University of California, Los Angeles

July 28th - ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling

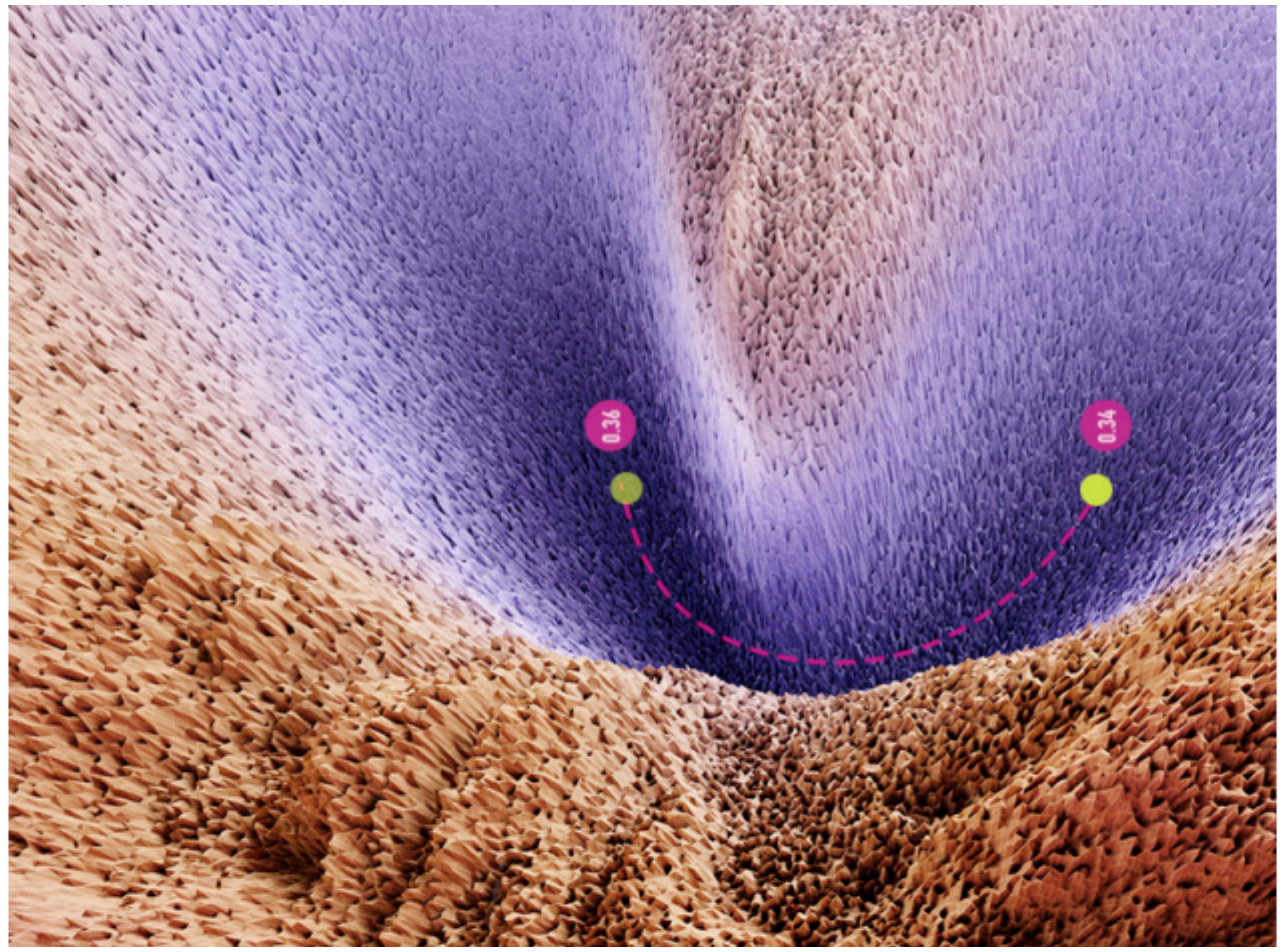
Motivation

Bad Uncertainty Estimation



Confidence by a ReLU neural network [2]

Risky Point Estimation

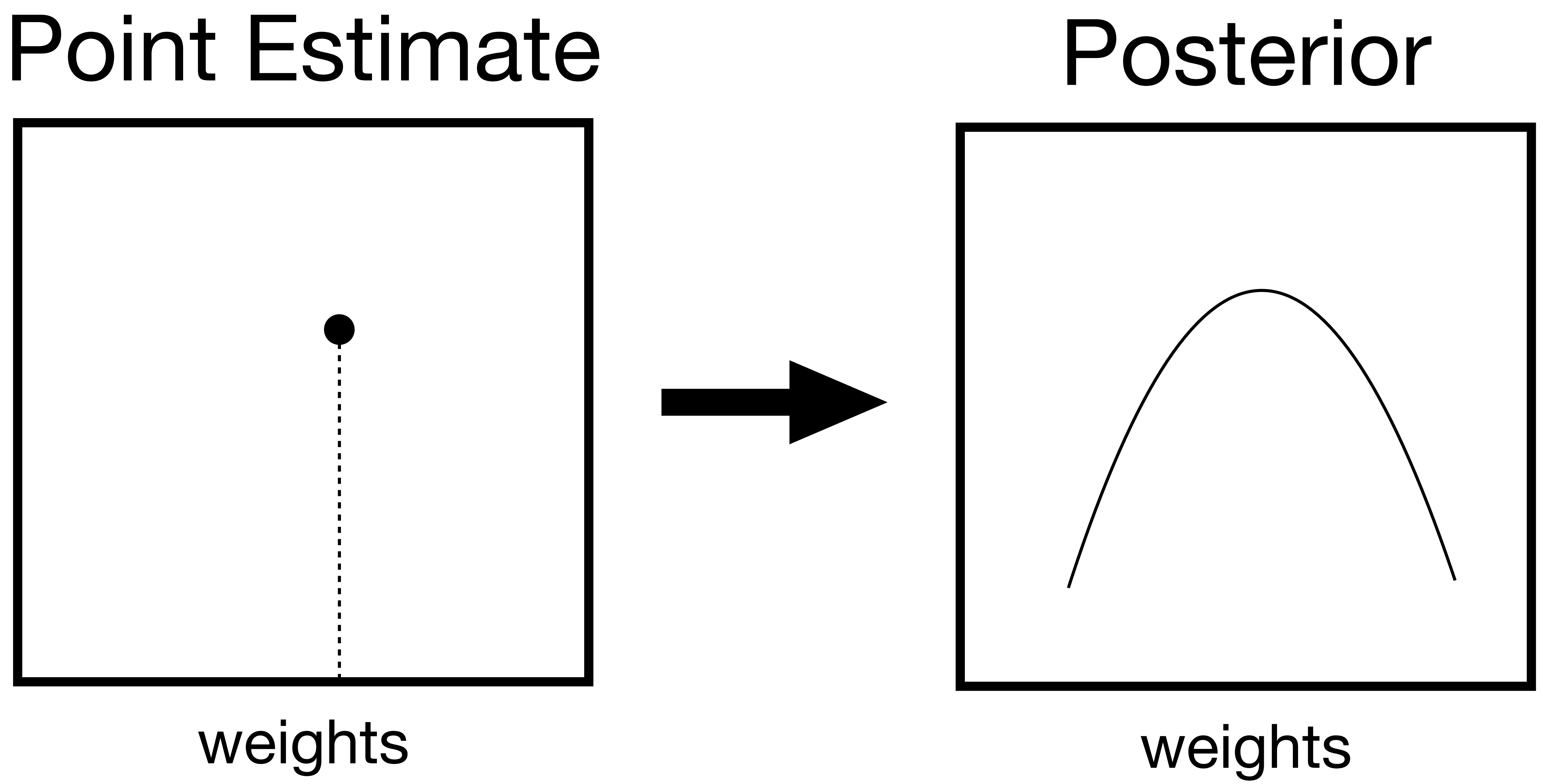


Loss surface [3]

➡**Bayesian Deep Learning** for *robust* and *reliable* predictions

Bayesian Model Average (BMA)

Key idea



$p(y \mid x, w)$

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

Motivation

- **Goal:** Bayesian model average

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

  Expected prediction $\mathbb{E}[y] = \int y \; p(y \mid x) \, dy$

- **Challenge:** DNNs are too big!

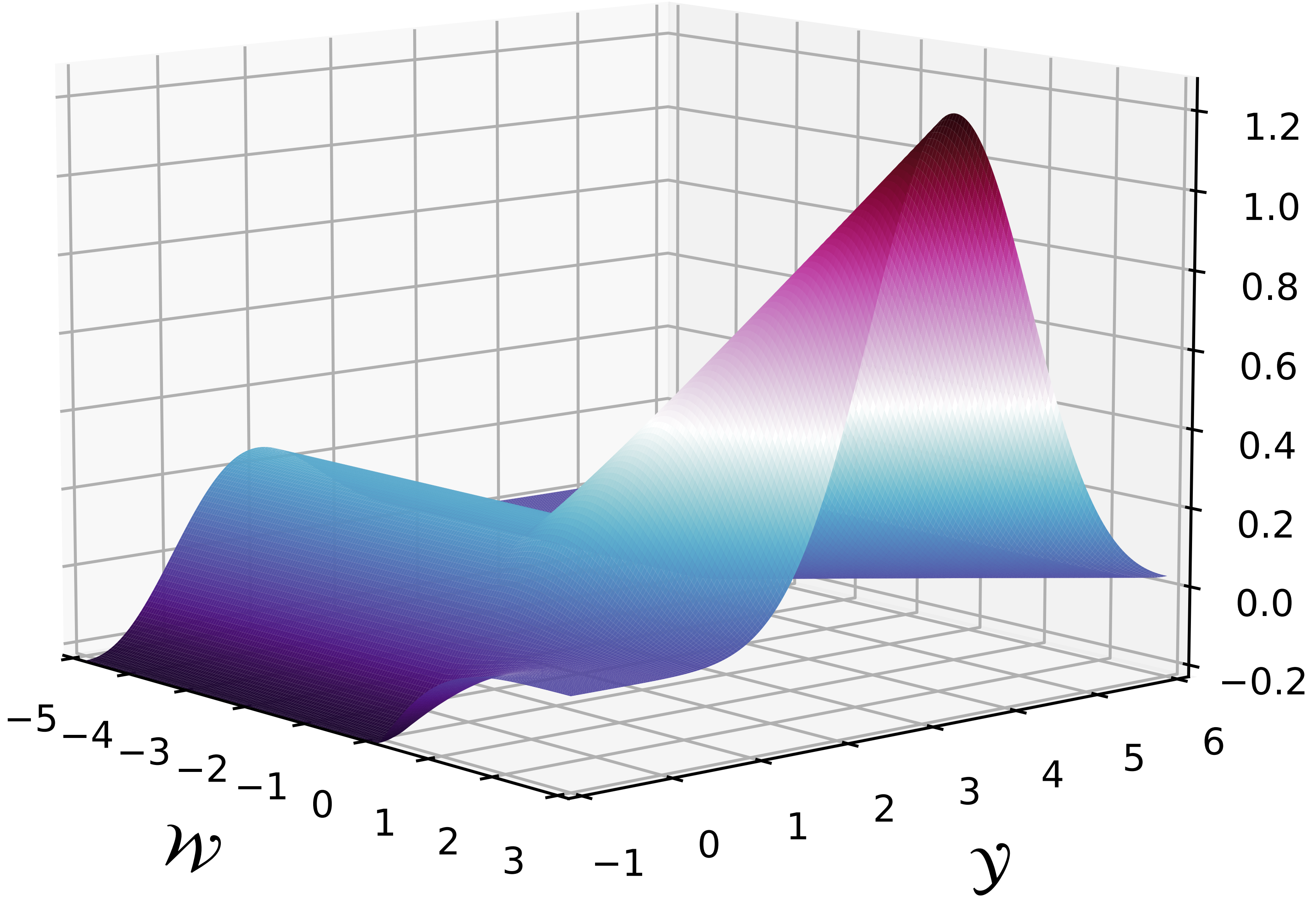
  ➡ *Costly* to maintain too many samples

  ➡ *Low sample efficiency* given the complex integrand

How complex? 🤔

How *complex* is the integrand? Simplest case!

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



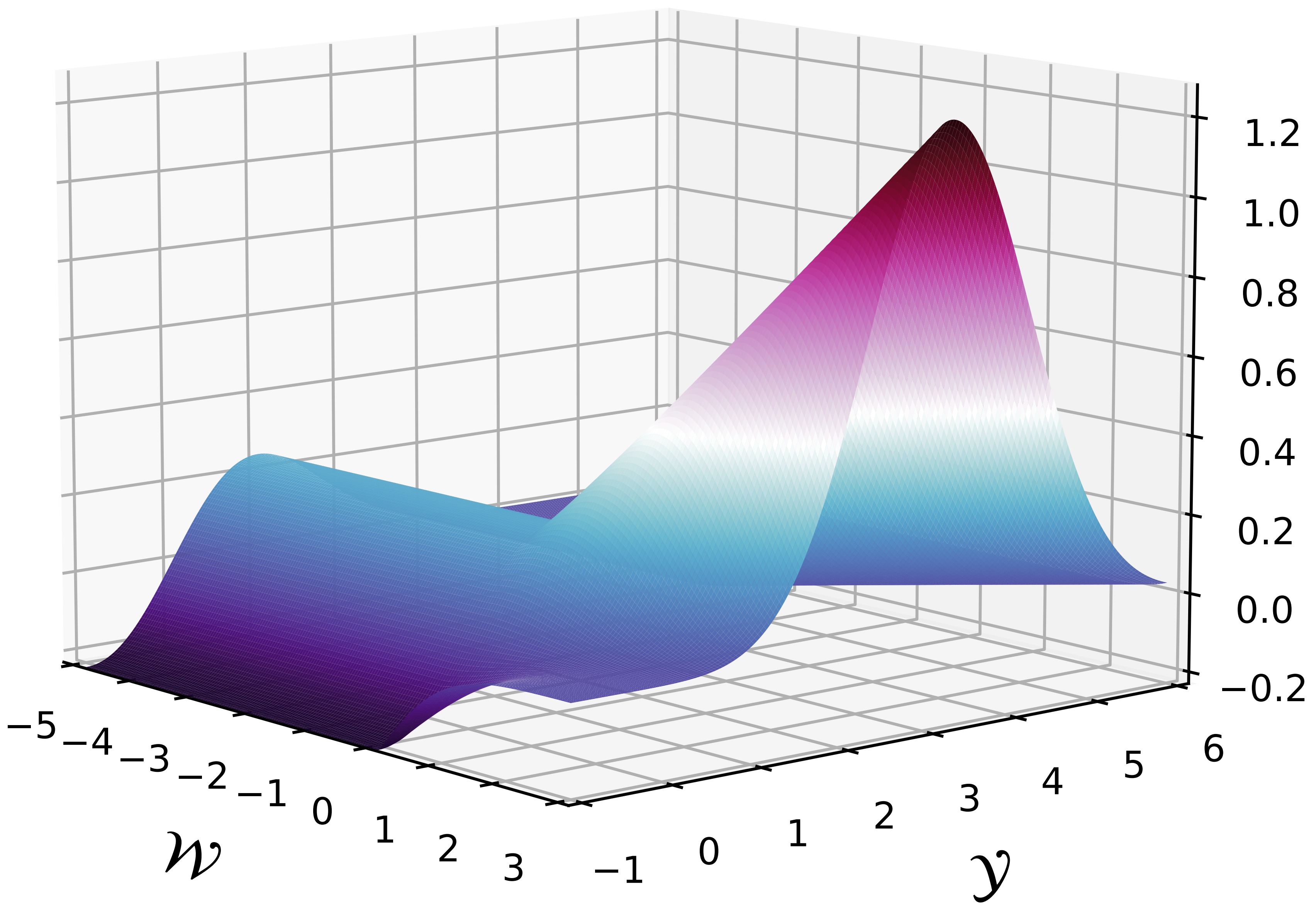
Non-convex, multi-modal, no closed form 🤯

Motivation

- **Goal:** Bayesian model average

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

Expected prediction $\mathbb{E}[y] = \int y \; p(y \mid x) \, dy$



Is there a *better* way to estimate the integral than *sampling*?

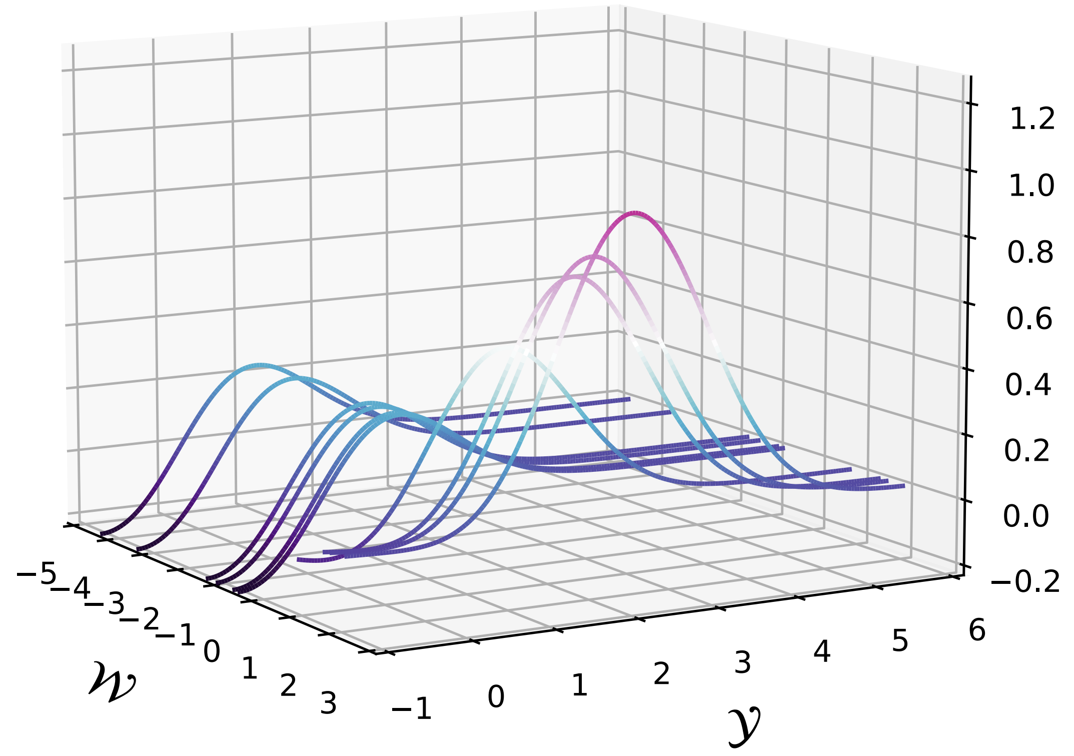
Yes! 🤩

Idea

A reduction from BMA to Exact Integration

BMA

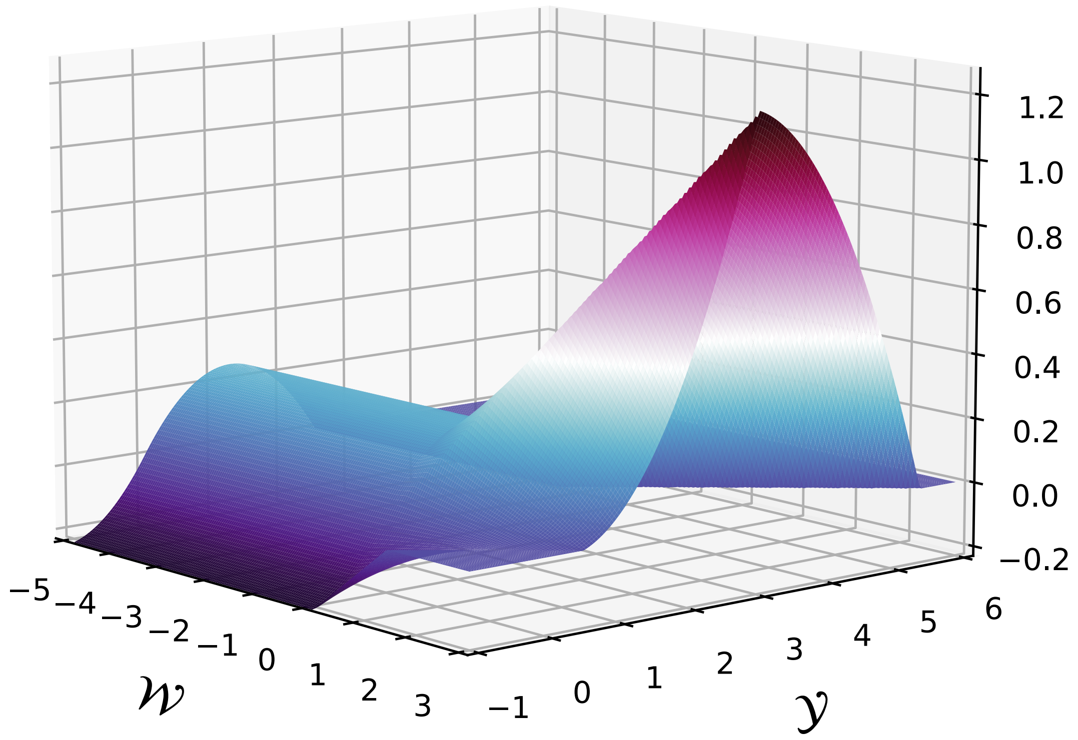
Sampling

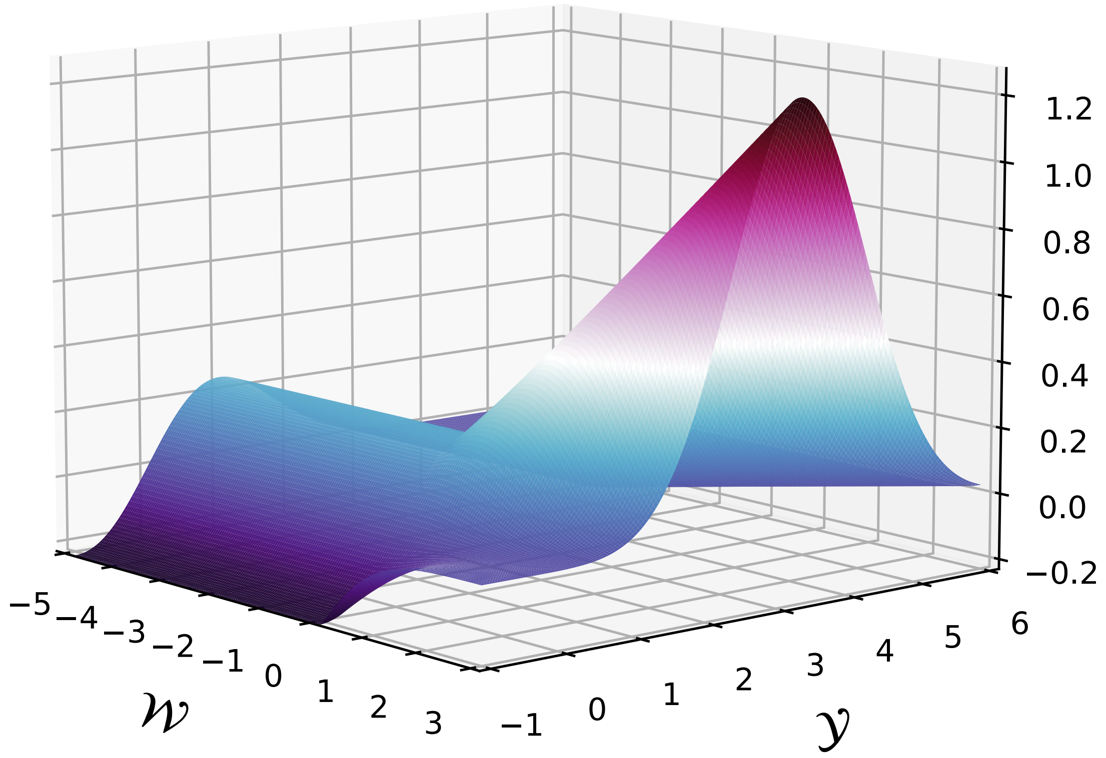


Approximate Representation

≈BMA

Exact Integration





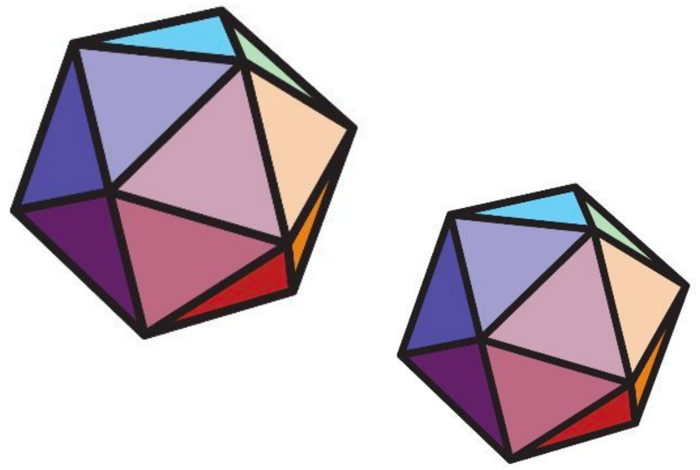
Ground truth BMA

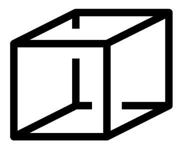
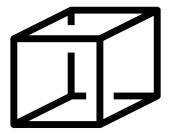
Weighted Model Integration

Idea

A reduction from BMA to ~~Exact Integration~~ *WMI*

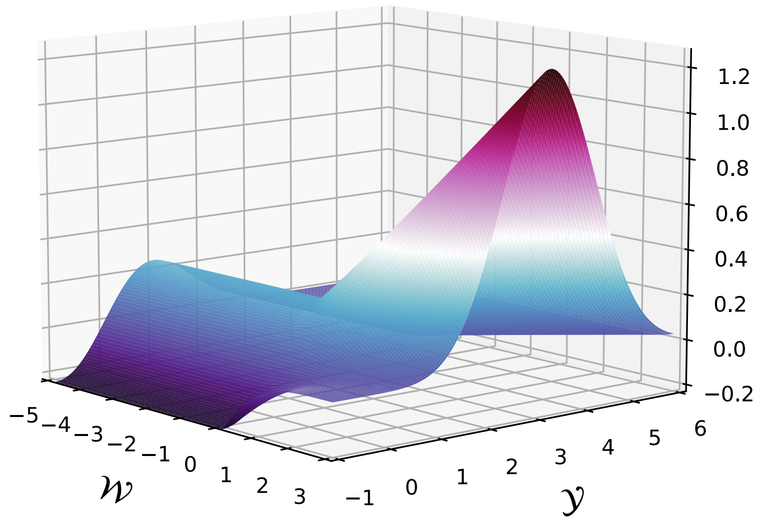
Weighted Model Integration (WMI) [4] is a class of *weighted volume computation* problems



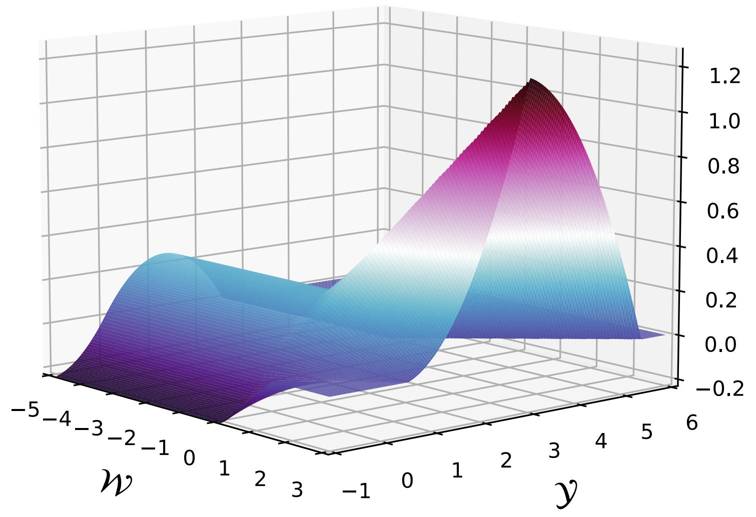
- *Region*:  a logical combination of arithmetic constraints
- *Polynomial Weight function* $\phi :$  $\rightarrow \mathbb{R}$

*Why **WMI**?*

- Existing WMI solvers are able to give ***exact marginalization*** results



$\approx$



Accurate approximation!

… but scalability?

Limitations

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

* Limited to fully connected layers

** Integration over polytopes in arbitrarily high dimensions is #P-hard

How to combine good from both worlds? 🤔

Limitations

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

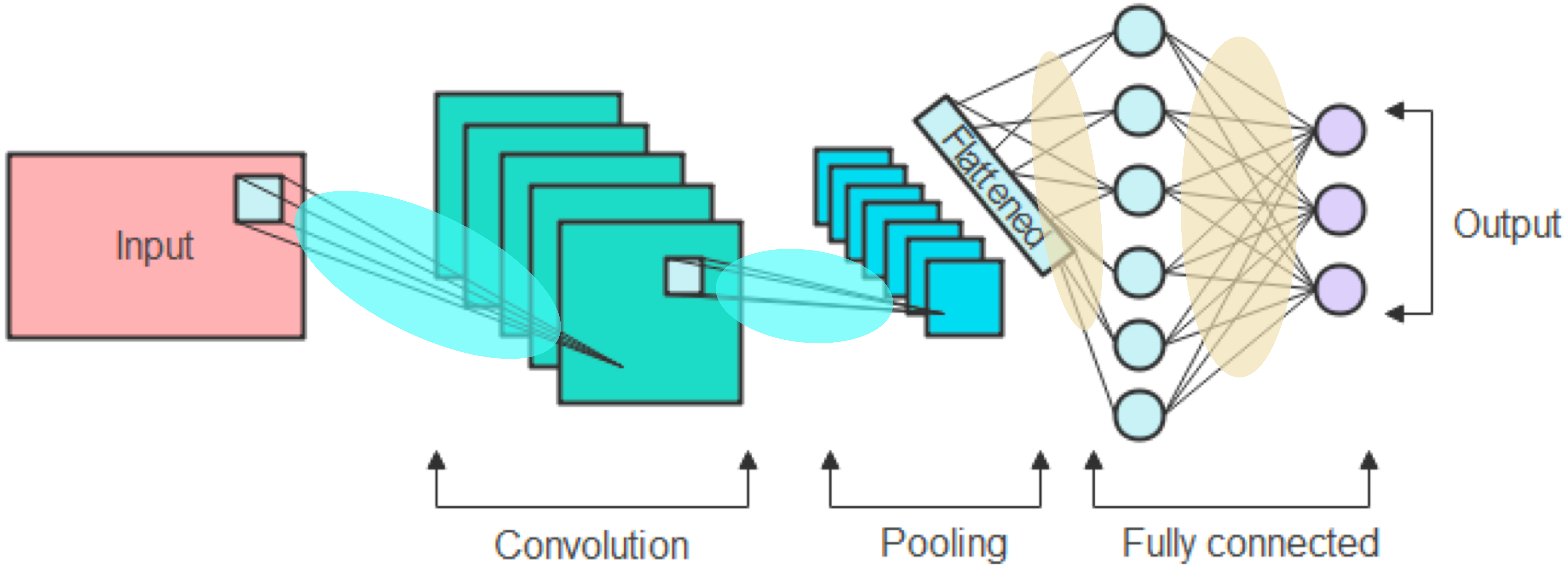
* Limited to fully connected layers

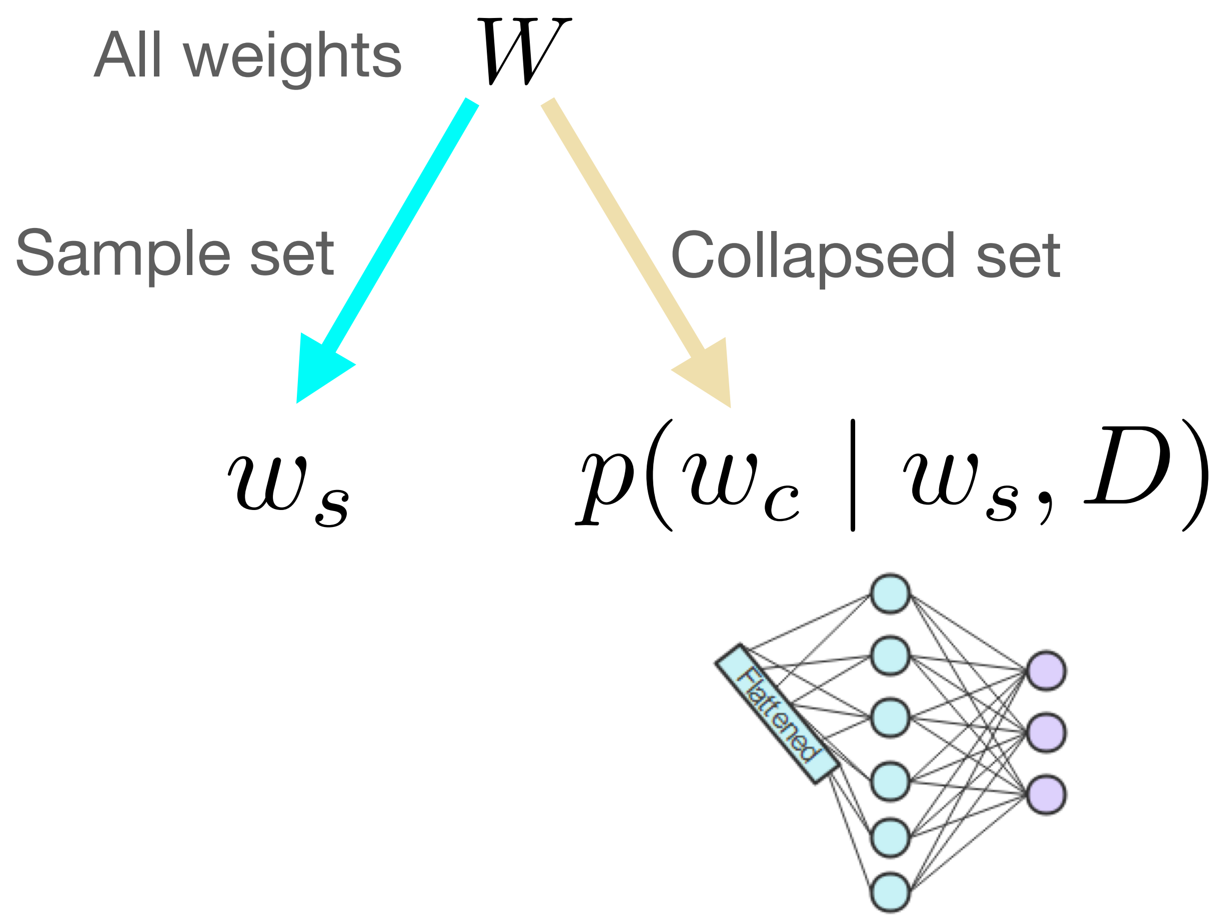
** Integration over polytopes in arbitrarily high dimensions is #P-hard

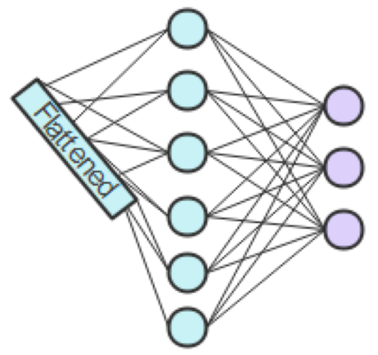
How to combine good from both worlds? 🤔

➡ *Collapsed inference scheme!* 💪

Collapsed Inference[5]





Expected prediction in BMA $\mathbb{E}[y] = \frac{1}{n} \sum_{w_s} \texttt{WMI}($  $)$

Accuracy + Flexibility, Scalability! 🥳

Experiment: UCI Regression

	Boston	Concrete	Yacht	Naval	Energy
CIBER (second)	**-2.471 ± 0.140**	-2.975 ± 0.102	-0.678 ± 0.301	7.276 ± 0.532	**<u>-0.716 ± 0.211</u>**
CIBER (last)	**-2.471 ± 0.140**	**<u>-2.959 ± 0.109</u>**	-0.687 ± 0.301	**<u>7.482 ± 0.188</u>**	**<u>-0.716 ± 0.211</u>**
SWAG	-2.761 ± 0.132	-3.013 ± 0.086	-0.404 ± 0.418	6.708 ± 0.105	-1.679 ± 1.488
PCA+ESS (SI)	-2.719 ± 0.132	-3.007 ± 0.086	**<u>-0.225 ± 0.400</u>**	6.541 ± 0.095	-1.563 ± 1.243
PCA+VI (SI)	-2.716 ± 0.133	-2.994 ± 0.095	-0.396 ± 0.419	6.708 ± 0.105	-1.715 ± 1.588
SGD	-2.752 ± 0.132	-3.178 ± 0.198	-0.418 ± 0.426	6.567 ± 0.185	-1.736 ± 1.613
DVI	-2.410 ± 0.020	-3.060 ± 0.010	-0.470 ± 0.030	6.290 ± 0.040	-1.010 ± 0.060
DGP	<u>-2.330 ± 0.060</u>	-3.130 ± 0.030	-1.390 ± 0.140	3.600 ± 0.330	-1.320 ± 0.030
VI	-2.430 ± 0.030	-3.040 ± 0.020	-1.680 ± 0.040	5.870 ± 0.290	-2.380 ± 0.020
MCD	-2.400 ± 0.040	-2.970 ± 0.020	-1.380 ± 0.010	4.760 ± 0.010	-1.720 ± 0.010
VSD	-2.350 ± 0.050	-2.970 ± 0.020	-1.140 ± 0.020	4.830 ± 0.010	-1.060 ± 0.010

CIBER Wins on 7/11!

	Elevators	KeggD	KeggU	Protein	Skillcraft	Pol
CIBER (second)	-0.378 ± 0.026	**<u>1.245 ± 0.090</u>**	**<u>1.125 ± 0.269</u>**	-0.720 ± 0.036	-1.003 ± 0.035	**<u>2.555 ± 0.115</u>**
CIBER (last)	-0.371 ± 0.023	1.178 ± 0.088	0.964 ± 0.231	-0.720 ± 0.036	**<u>-1.001 ± 0.032</u>**	2.506 ± 0.150
SWAG	-0.374 ± 0.021	1.080 ± 0.035	0.749 ± 0.029	**<u>-0.700 ± 0.051</u>**	-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)	**<u>-0.325 ± 0.019</u>**	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

Experiment: Image Classification

METRIC	NLL		ACC		ECE	
DATASET	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100
CIBER	$\mathbf{0.1927 \pm 0.0029}$	$\mathbf{0.9193 \pm 0.0027}$	$\mathbf{93.64 \pm 0.09}$	$\mathbf{74.71 \pm 0.18}$	0.0130 ± 0.0011	$\mathbf{0.0168 \pm 0.0025}$
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	$\mathbf{0.0082 \pm 0.0012}$	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	$\mathbf{0.9869 \pm 0.0102}$	$\mathbf{72.56 \pm 0.23}$	$\mathbf{0.0925 \pm 0.0028}$
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

- applicable to large NNs
- achieves accurate estimations of uncertainty
- boosts predictive performance

Takeaway

Volume Computation Solvers + Statistical ML

=

Reliable & Scalable Inference!

Thanks! Welcome to Poster #102 at 4-5 pm

[1] Zhe Zeng and Guy Van den Broeck. Collapsed inference for Bayesian deep learning. In ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling, 2023.

[2] Kristiadi, Agustinus, Matthias Hein, and Philipp Hennig. "Being bayesian, even just a bit, fixes overconfidence in relu networks." *International conference on machine learning*. PMLR, 2020.

[3] Garipov, Timur, et al. "Loss surfaces, mode connectivity, and fast ensembling of dnns." *Advances in neural information processing systems* 31 (2018).

[4] V. Belle, A. Passerini, and G. Van den Broeck. Probabilistic inference in hybrid domains by weighted model integration. In Proceedings of 24th International Joint Conference on Artificial Intelligence (IJCAI), pages 2770–2776, 2015.