Collapsed inference for Bayesian deep learning

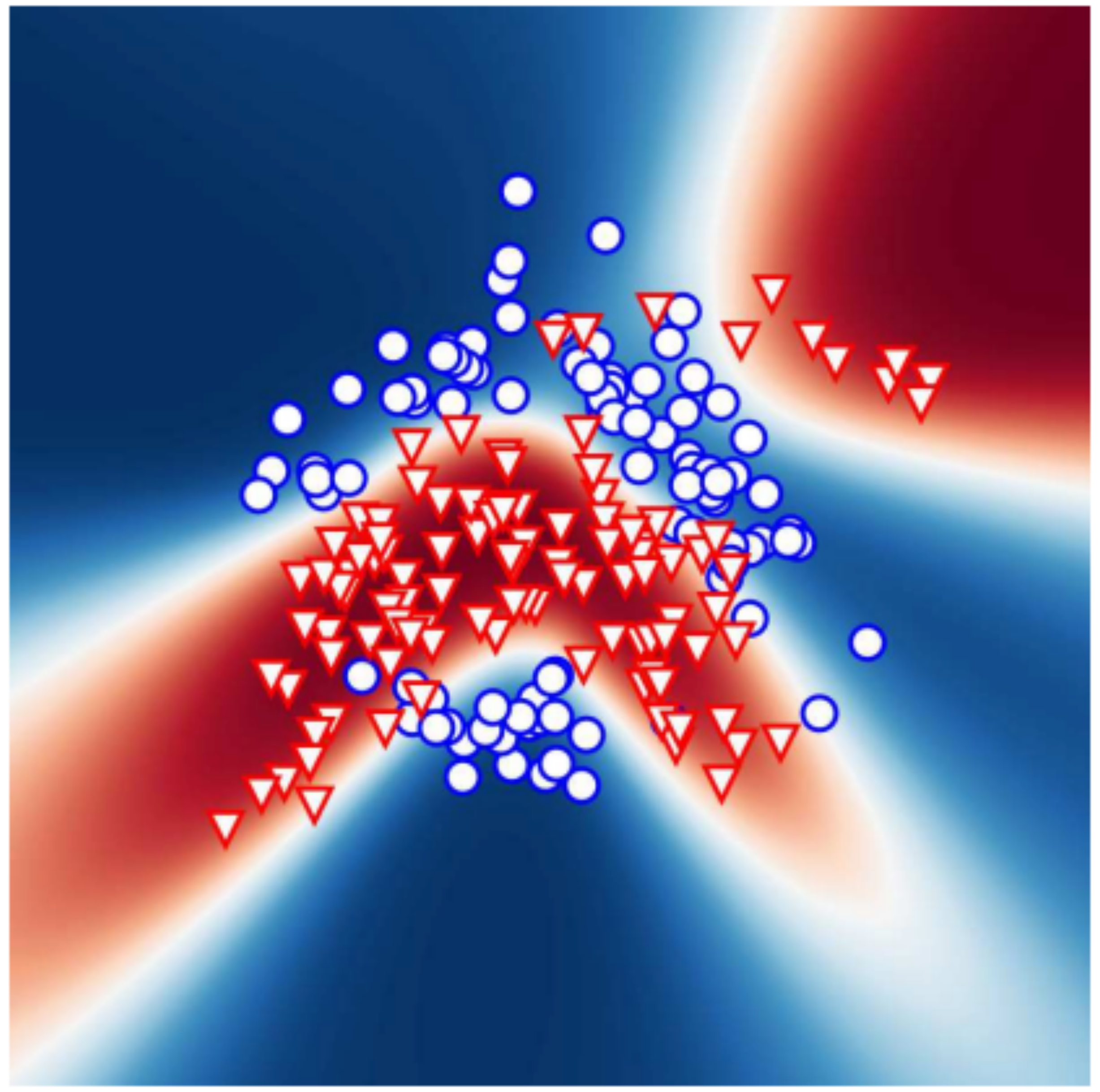
Zhe Zeng, Guy Van den Broeck

University of California, Los Angeles

NeurIPS 2023

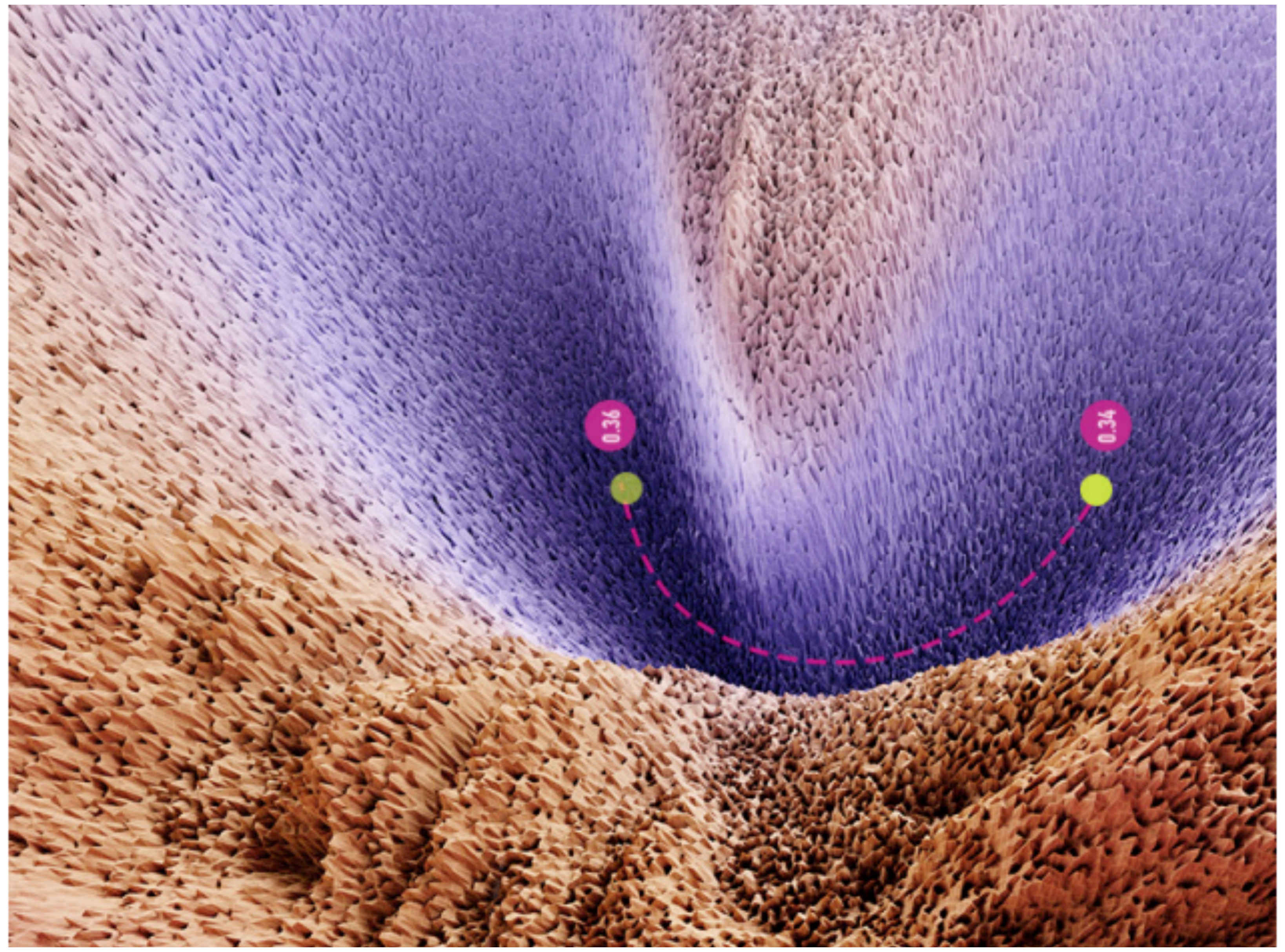
Motivation

Bad Uncertainty Estimation



Confidence by a ReLU neural network [1]

Risky Point Estimation

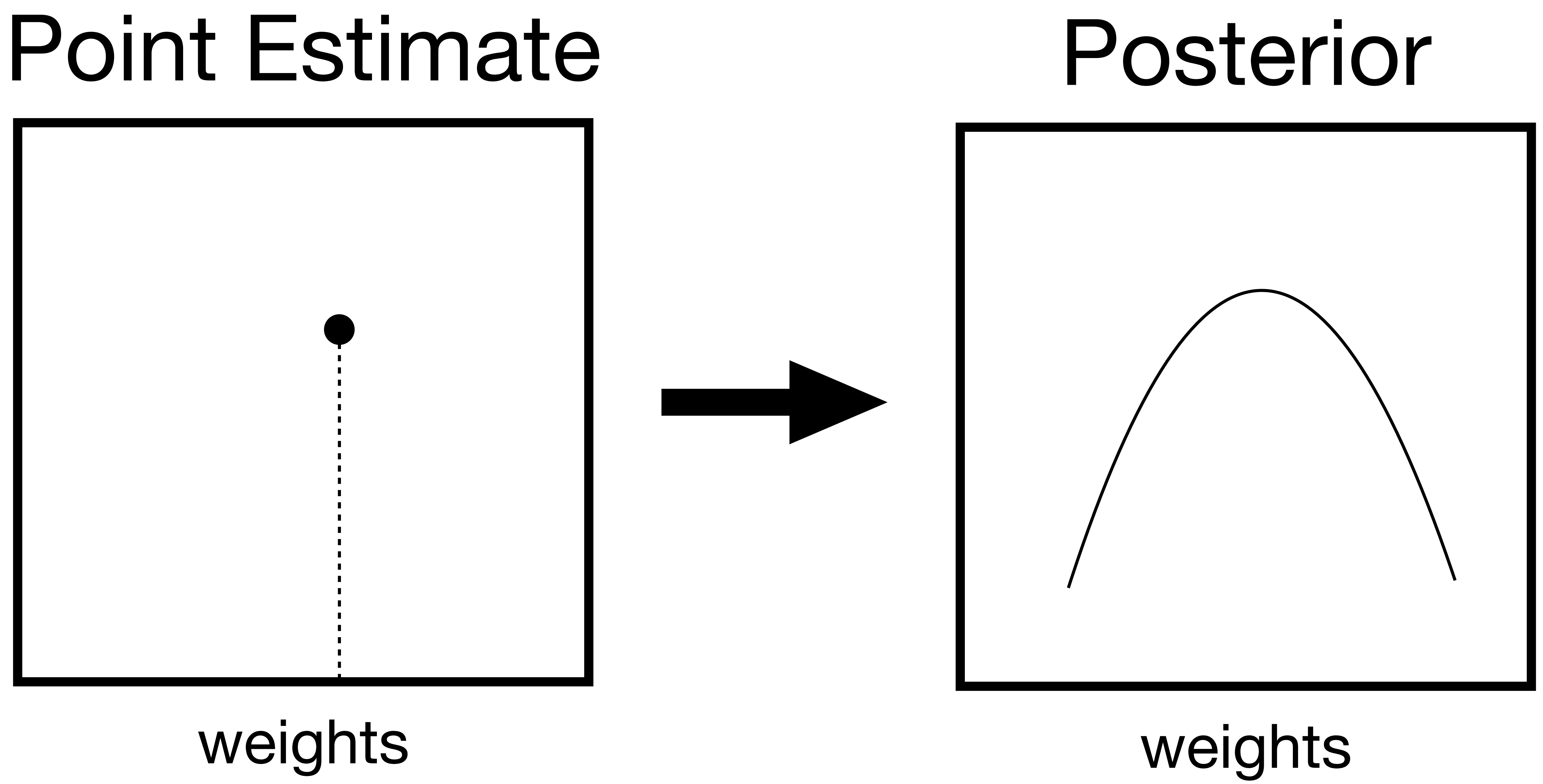


Loss surface [2]

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

Bayesian Model Average (BMA)

Key idea



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

Bayesian Model Average (BMA)

Key idea

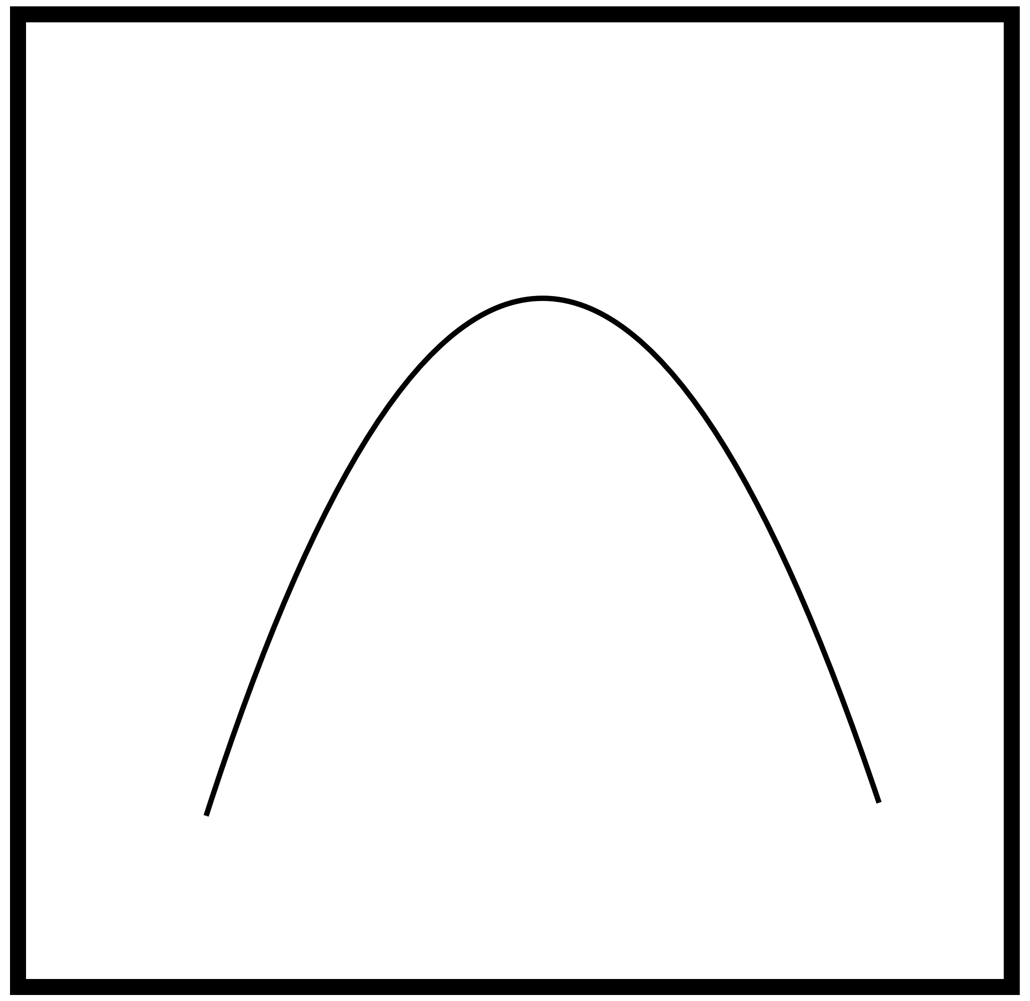
- **Challenge:** DNNs are too big!

➡ *Costly* to maintain many samples

➡ *Low sample efficiency* given the $\boxed{\text{complex}}$ integrand

How complex? 🤔

Posterior

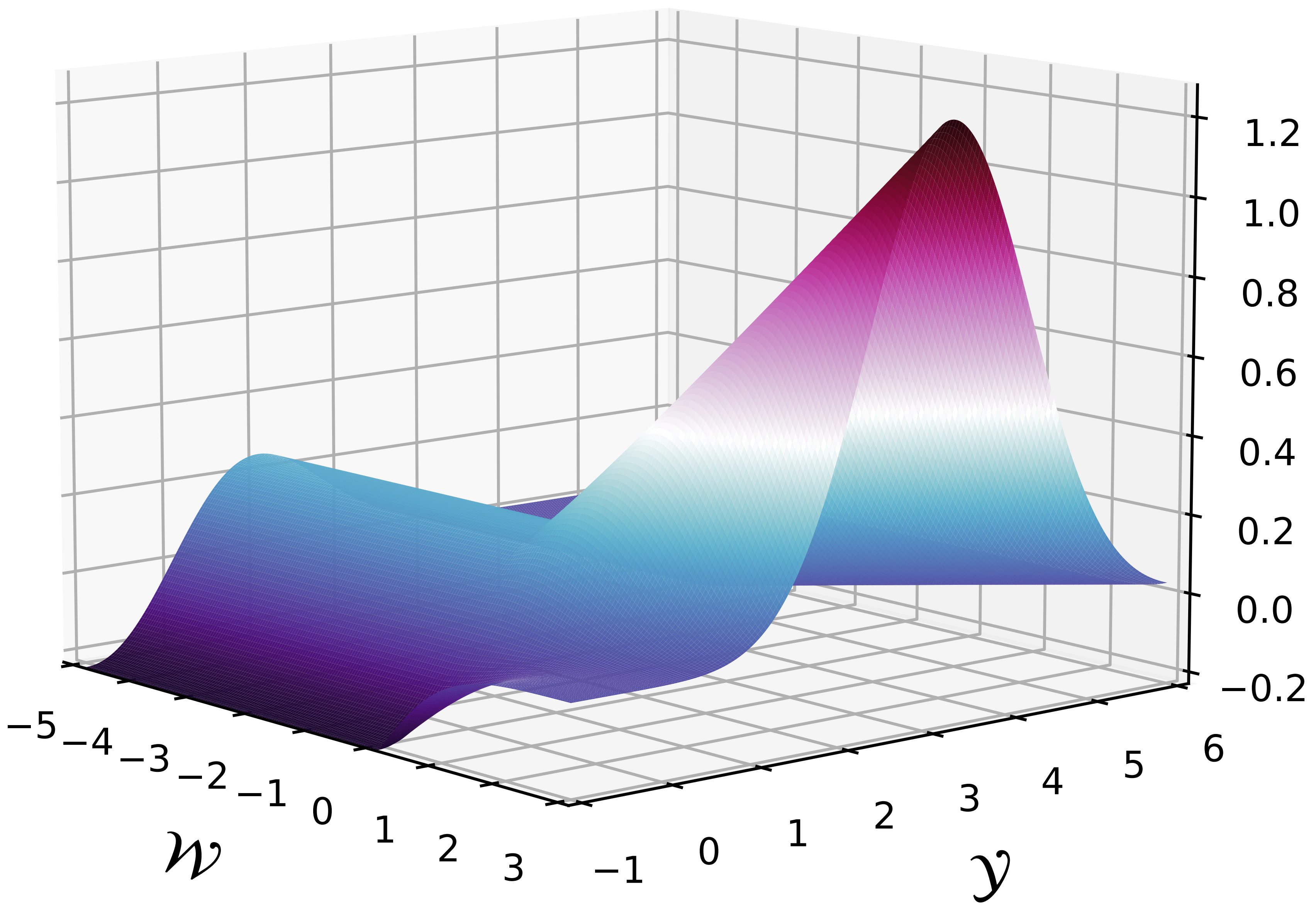


weights

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

How *complex* is the integrand? Simplest case!

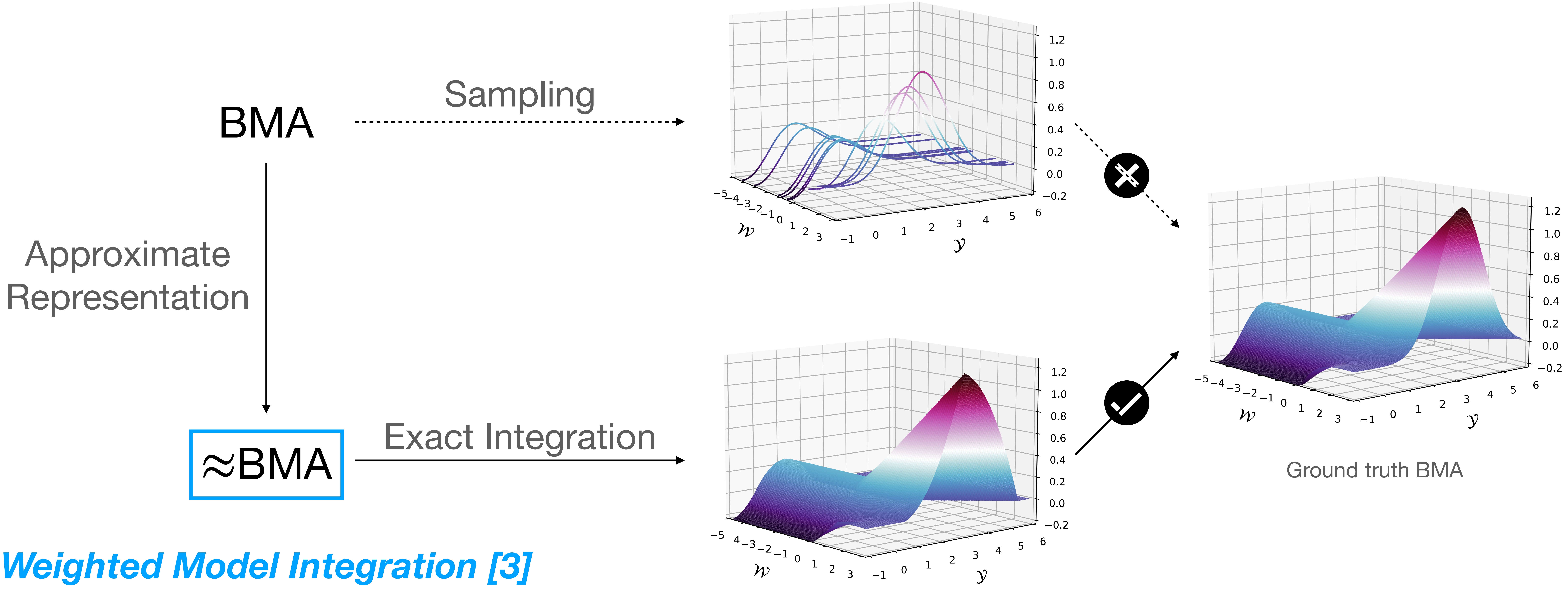
Expected prediction $\mathbb{E}[y] = \int y \, \underbrace{p(w \mid D)}_{\text{Uniform}} \, \underbrace{p(y \mid \underbrace{f(x)}_{\text{Single ReLU}}, w)}_{\text{Gaussian}} \, dw \, dy$



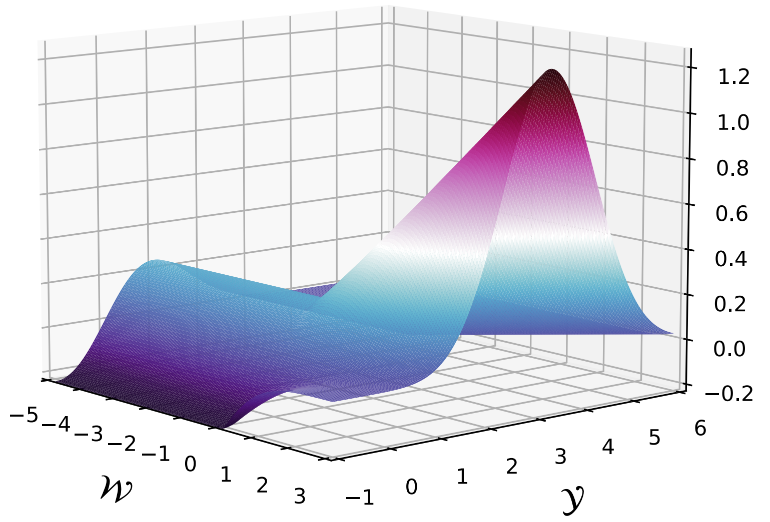
Non-convex, multi-modal, no closed form 🤯

Idea

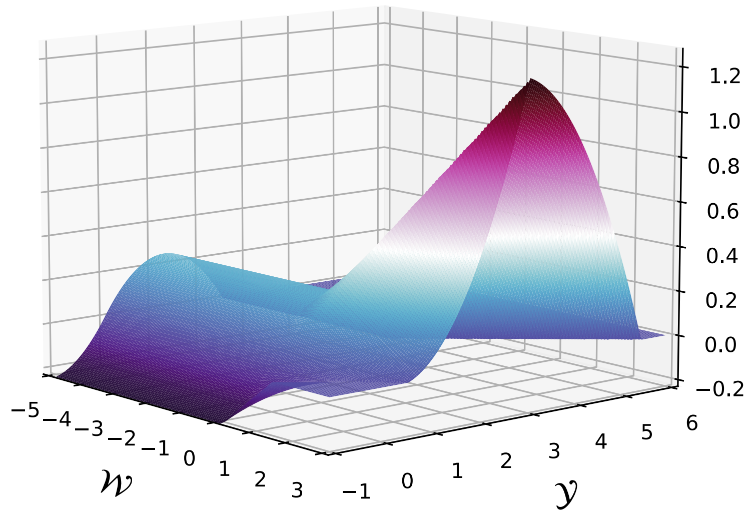
A reduction from BMA to Exact Integration



Weighted Model Integration [3]



$\approx$



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	✅	❌**	✅

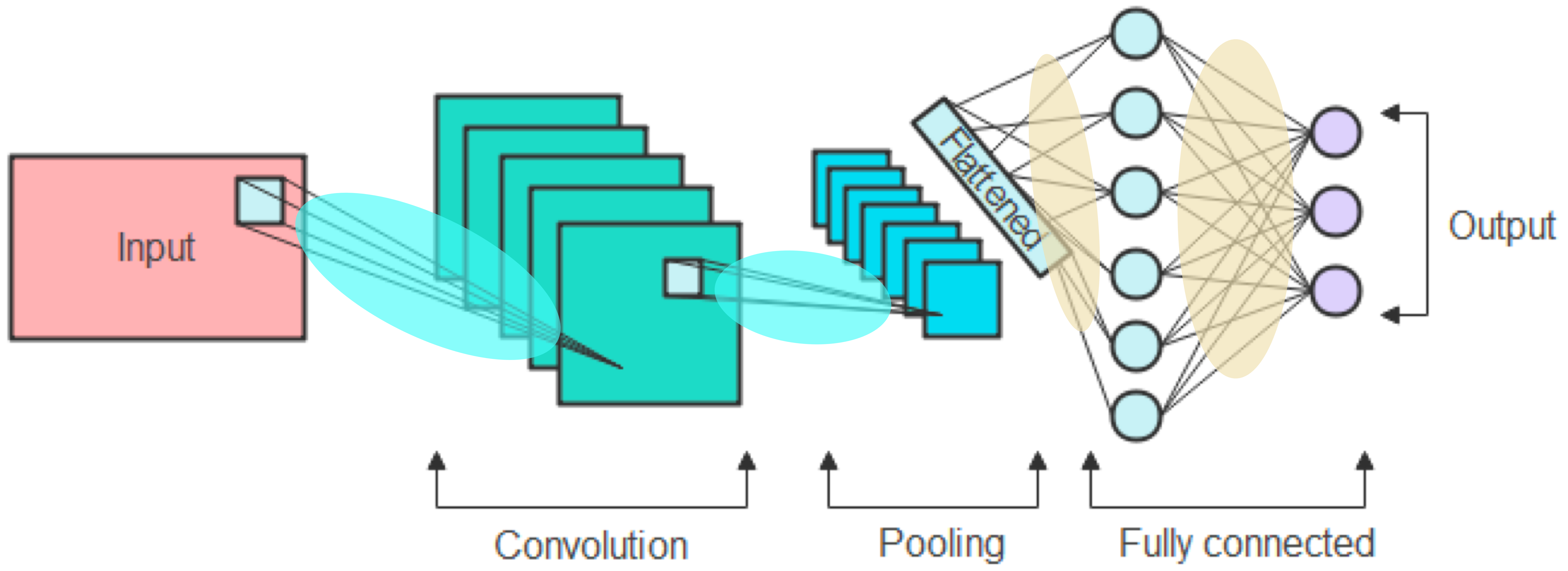
* Limited to fully connected layers

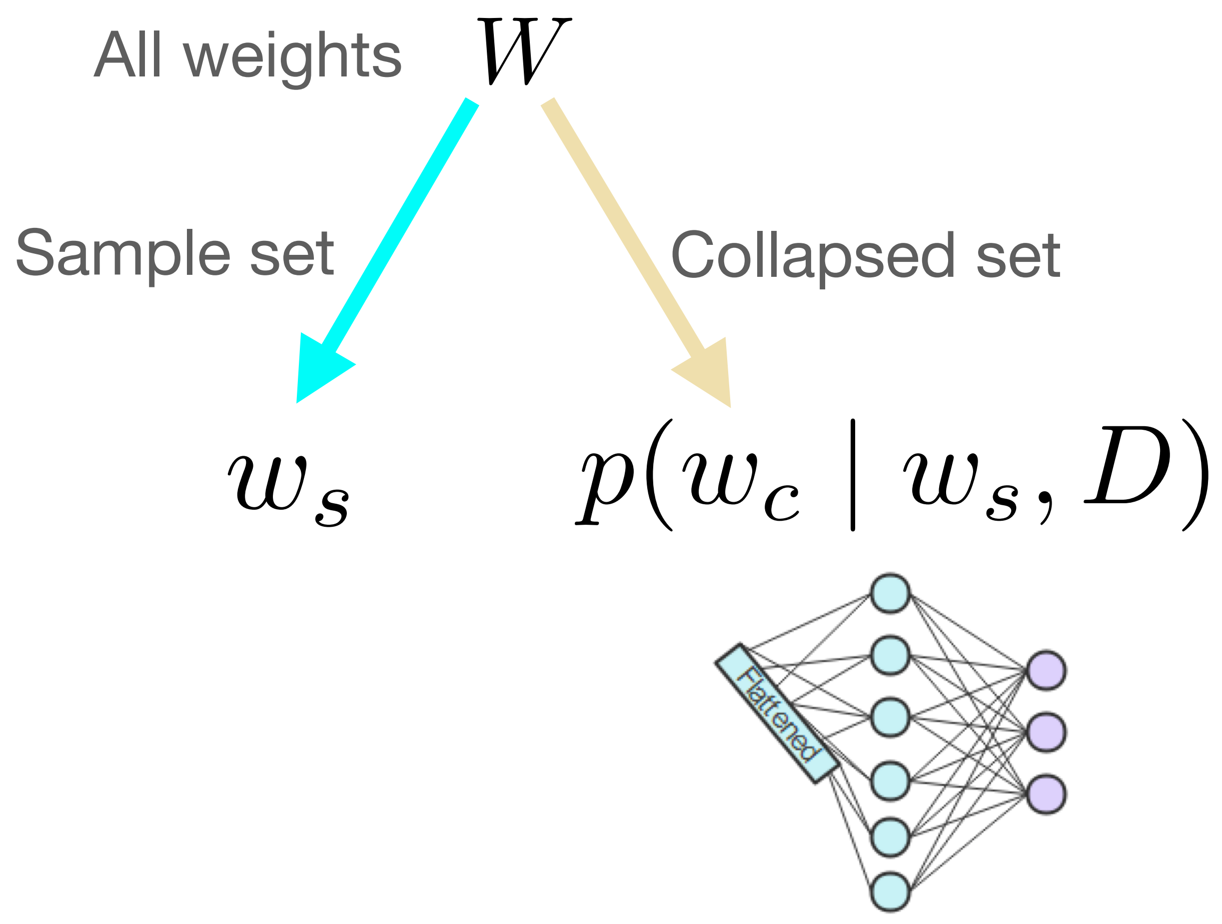
** Integration over polytopes in arbitrarily high dimensions is #P-hard

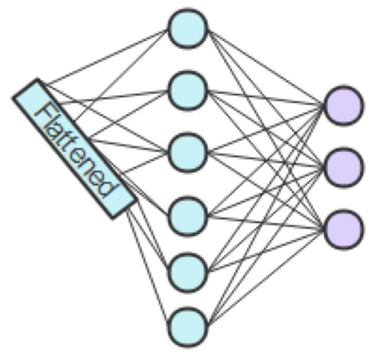
How to combine good from both worlds? 🤔

➡ *Collapsed inference scheme!* 💪

Collapsed Inference





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

Accuracy + Flexibility, Scalability! 🥳

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

- Our approach is applicable to large NNs
- It achieves accurate estimations of uncertainty
- It boosts predictive performance

Takeaway

Volume Computation Solvers + Statistical ML

=

Reliable & Scalable Inference!

Thanks!

[1] Meronen, Lassi, Martin Trapp, and Arno Solin. "Periodic activation functions induce stationarity." *Advances in Neural Information Processing Systems* 34 (2021): 1673-1685.

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

[3] 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.