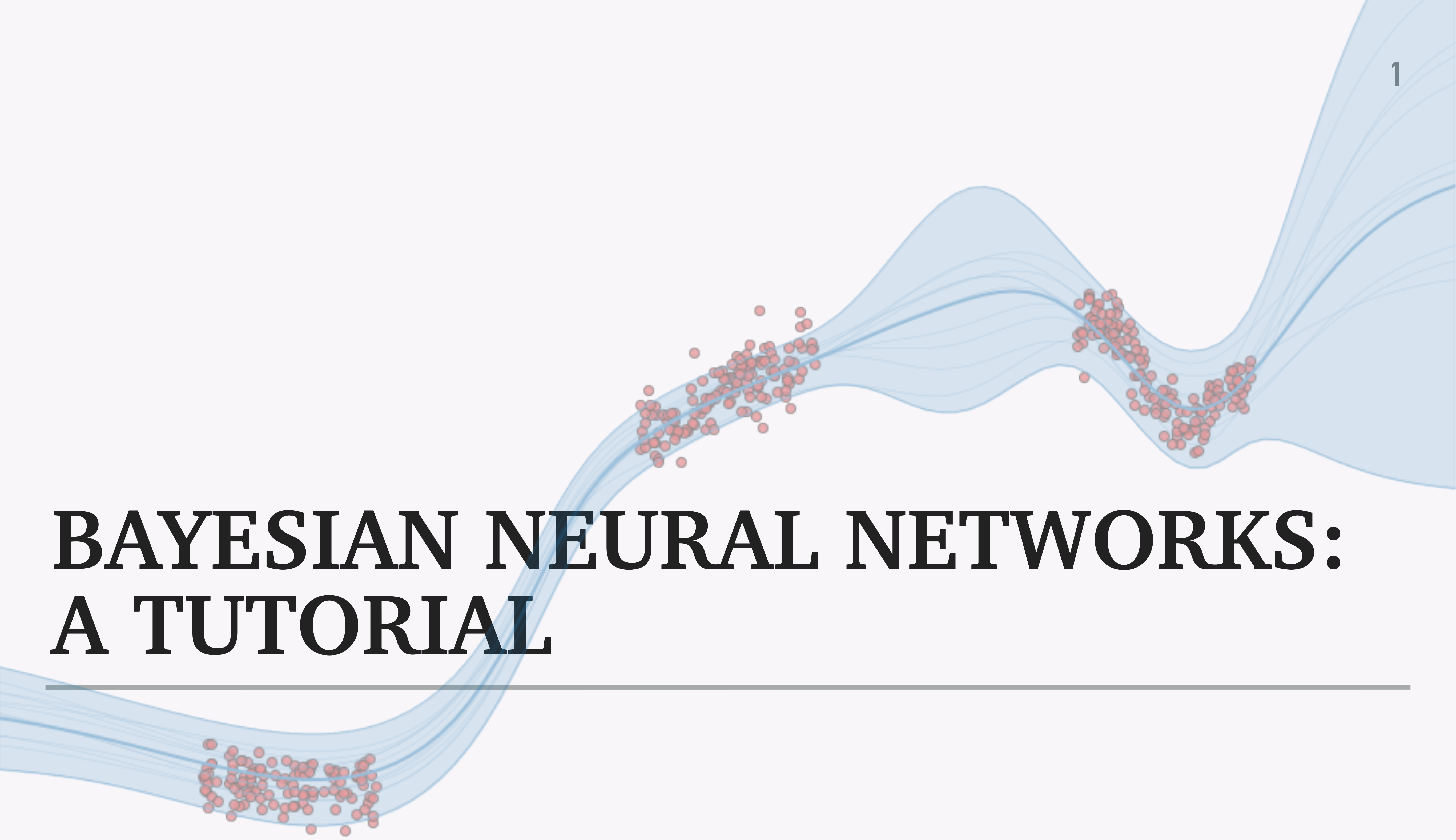


BAYESIAN NEURAL NETWORKS: A TUTORIAL

WESLEY MADDOX

NYU

THANKS TO + MY COLLABORATORS

- Andrew Wilson
- Pavel Izmailov & Polina Kirichenko
  - For the slides :)
- Greg Benton
- Slides available at: https://wjmaddox.github.io/assets/BNN_tutorial_CILVR.pdf

STRUCTURE

- Motivation
- Intro to Bayesian Inference
- Approximate Inference
  - Variational Inference
  - Laplace Approximations
  - MCMC
- Loss-Geometry Inspired Methods (our work)

MOTIVATION

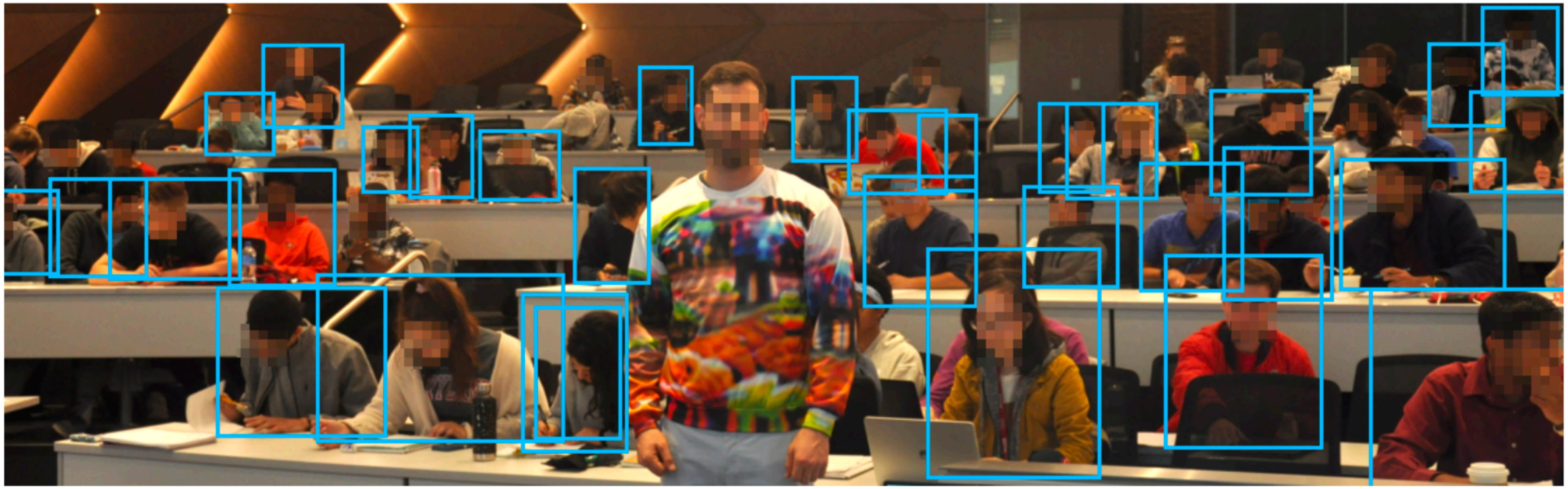
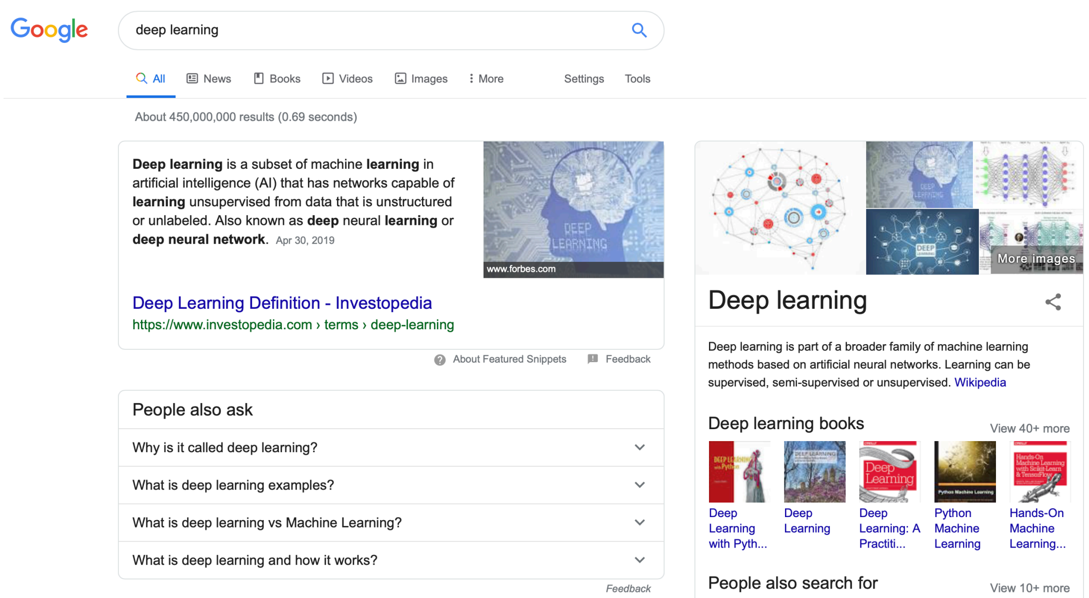
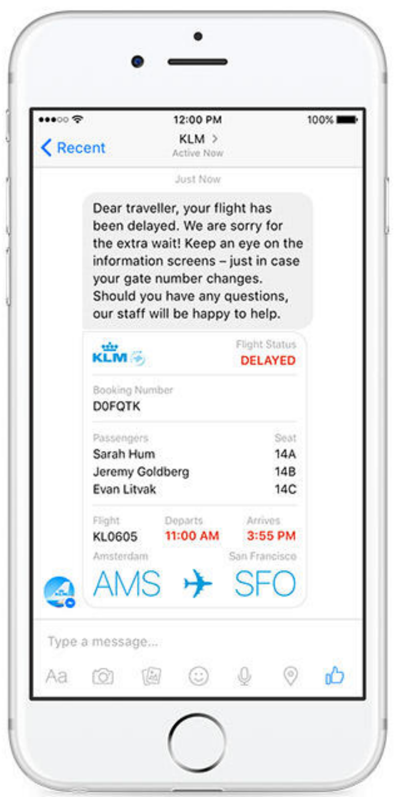


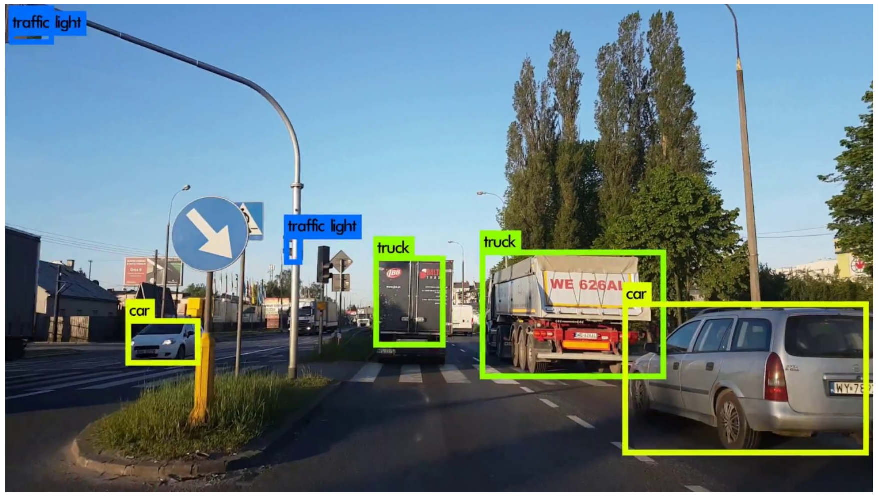
Figure 1: This stylish pullover is a great way to stay warm this winter, whether in the office or on-the-go. It features a stay-dry microfleece lining, a modern fit, and adversarial patterns the evade most common object detectors. In this demonstration, the YOLOv2 detector is evaded using a pattern trained on the COCO dataset with a carefully constructed objective.

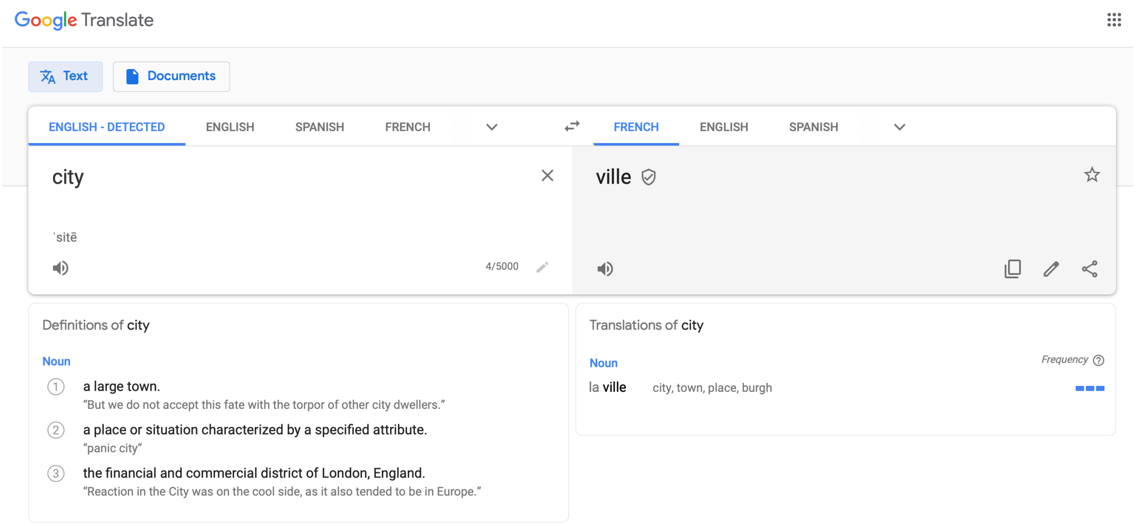
From: "Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors," Wu, Lim, Davis, Goldstein, https://arxiv.org/pdf/1910.14667.pdf

DEEP LEARNING SUCCESS



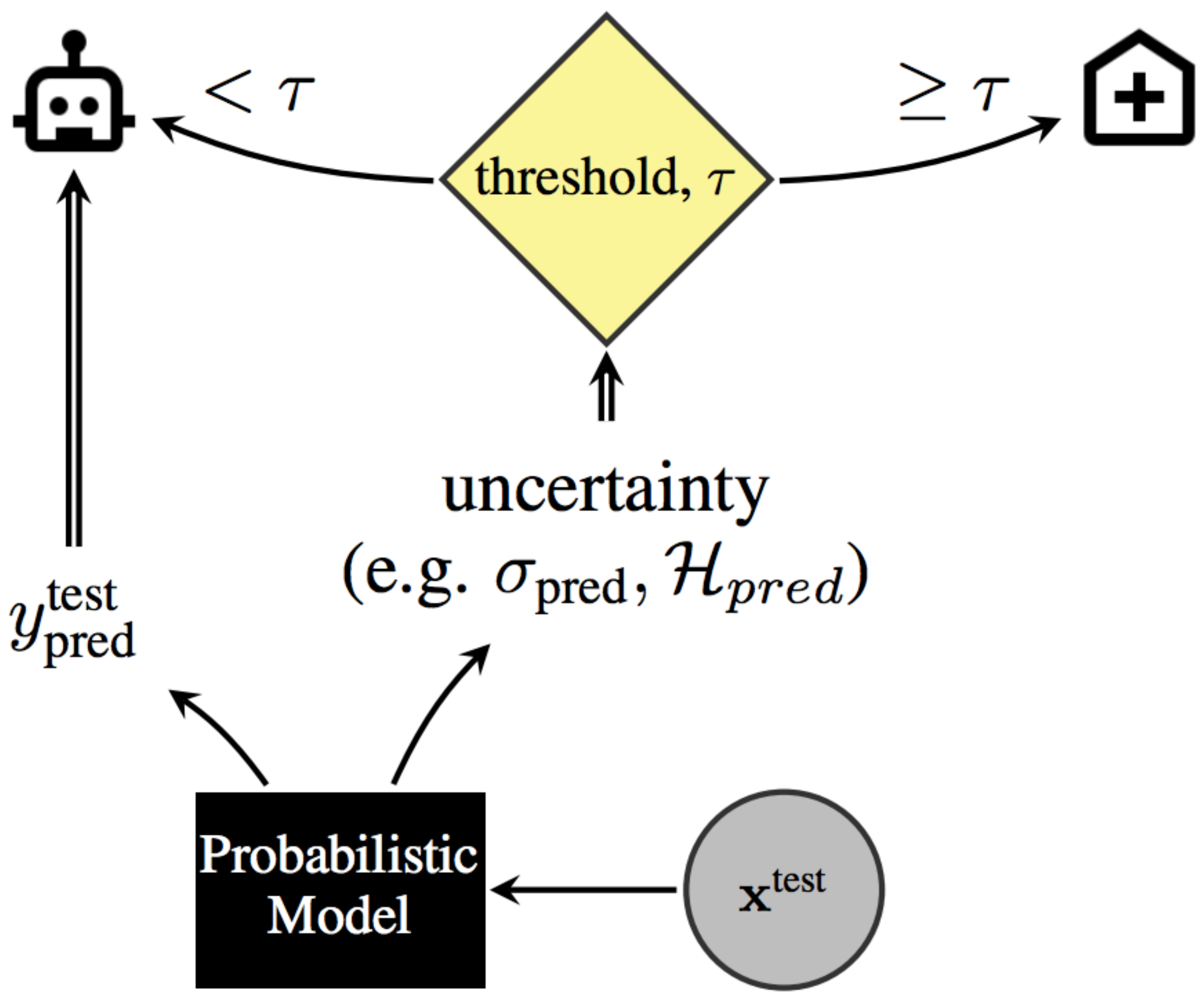






UNCERTAINTY IN DEEP LEARNING

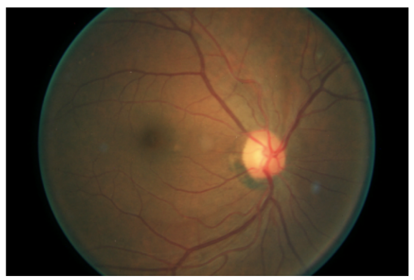
Automated diagnosis: human-in-the-loop

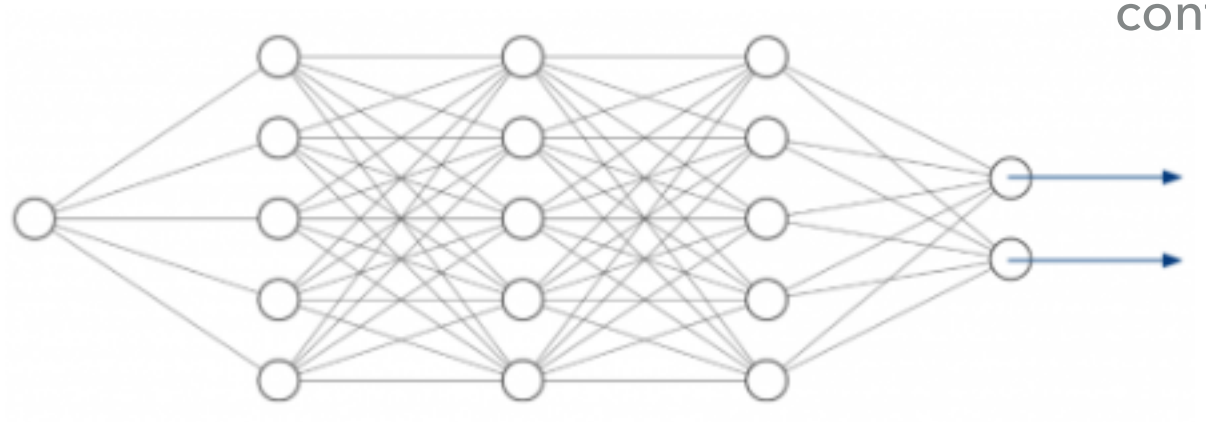


"Benchmarking Bayesian Deep Learning with Diabetic Retinopathy Diagnosis" by Angelos Filos et al.

CALIBRATION

confidence prediction

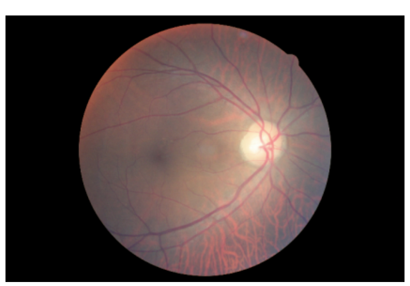


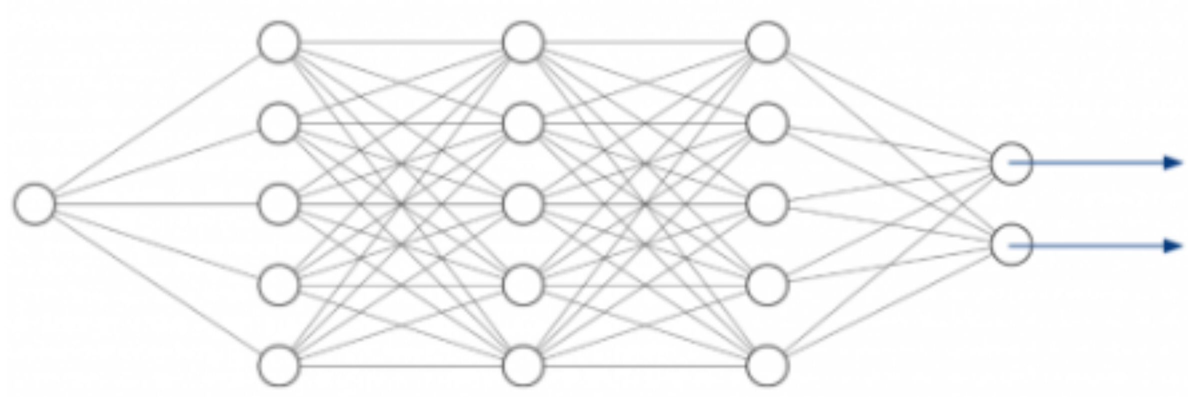


80%

20%

healthy





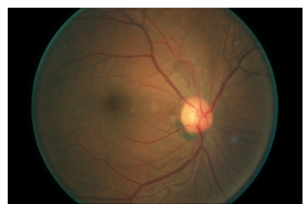
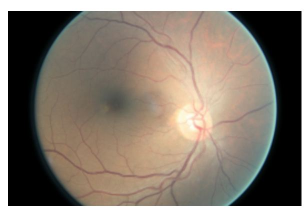
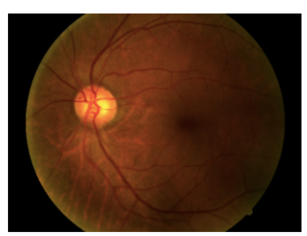
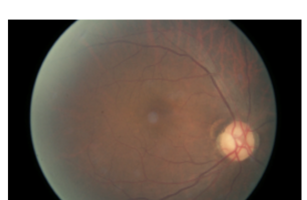
10%

90%

unhealthy

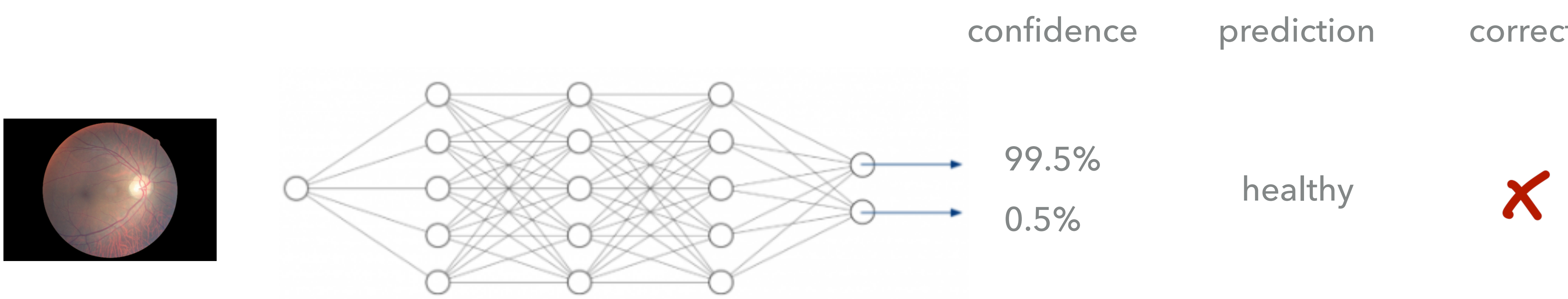
"Benchmarking Bayesian Deep Learning with Diabetic Retinopathy Diagnosis" by Angelos Filos et al.

CALIBRATION

	confidence	prediction	correct
	80% 20%	healthy	✓
	80% 20%	healthy	✓
	80% 20%	healthy	✓
	80% 20%	healthy	✗
	80% 20%	healthy	✓

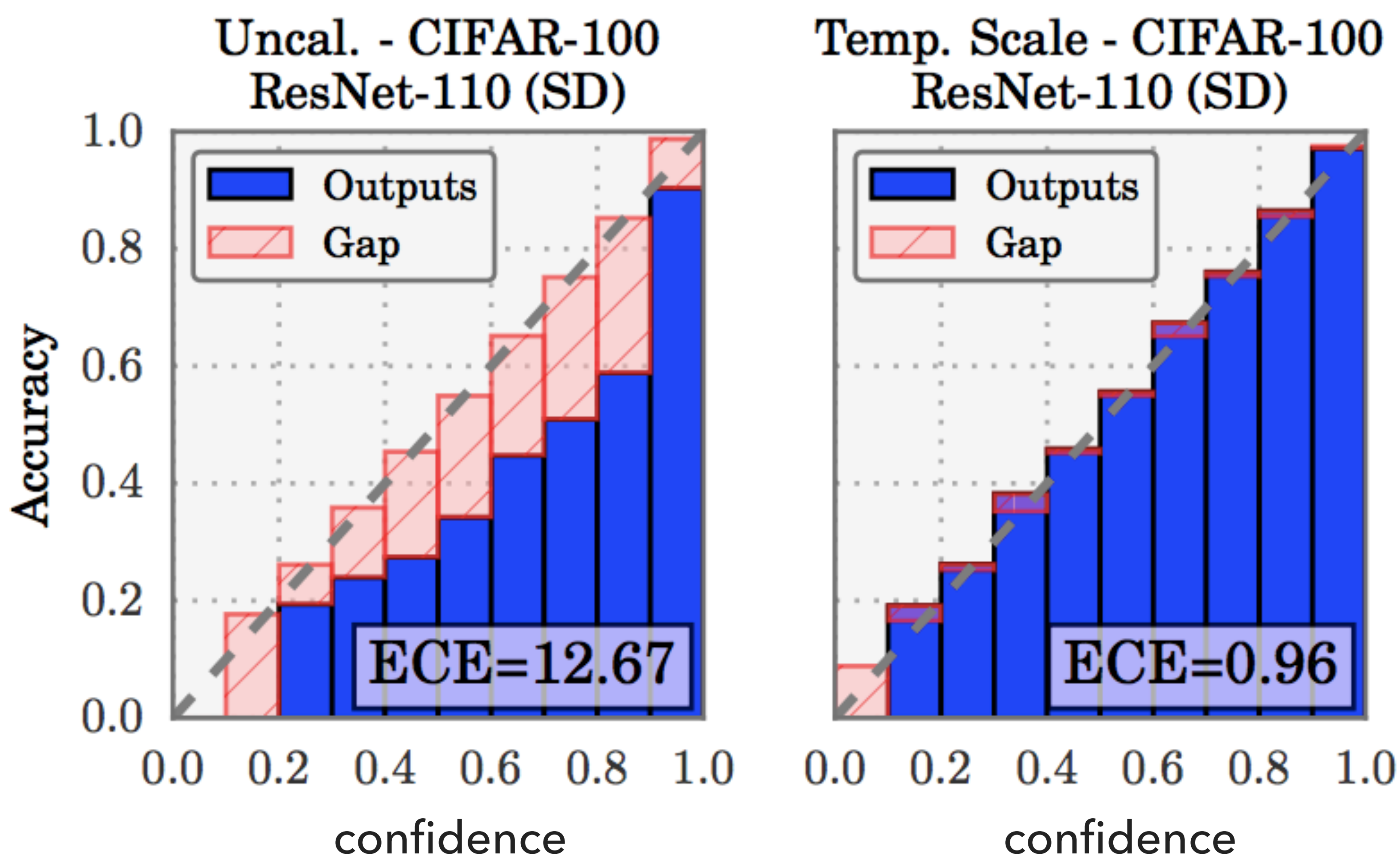
UNCERTAINTY: OVERCONFIDENCE IN NEURAL NETWORKS

- $p(y|x)$ should represent probabilities of belonging to a class
- Neural networks are often over-confident in their predictions



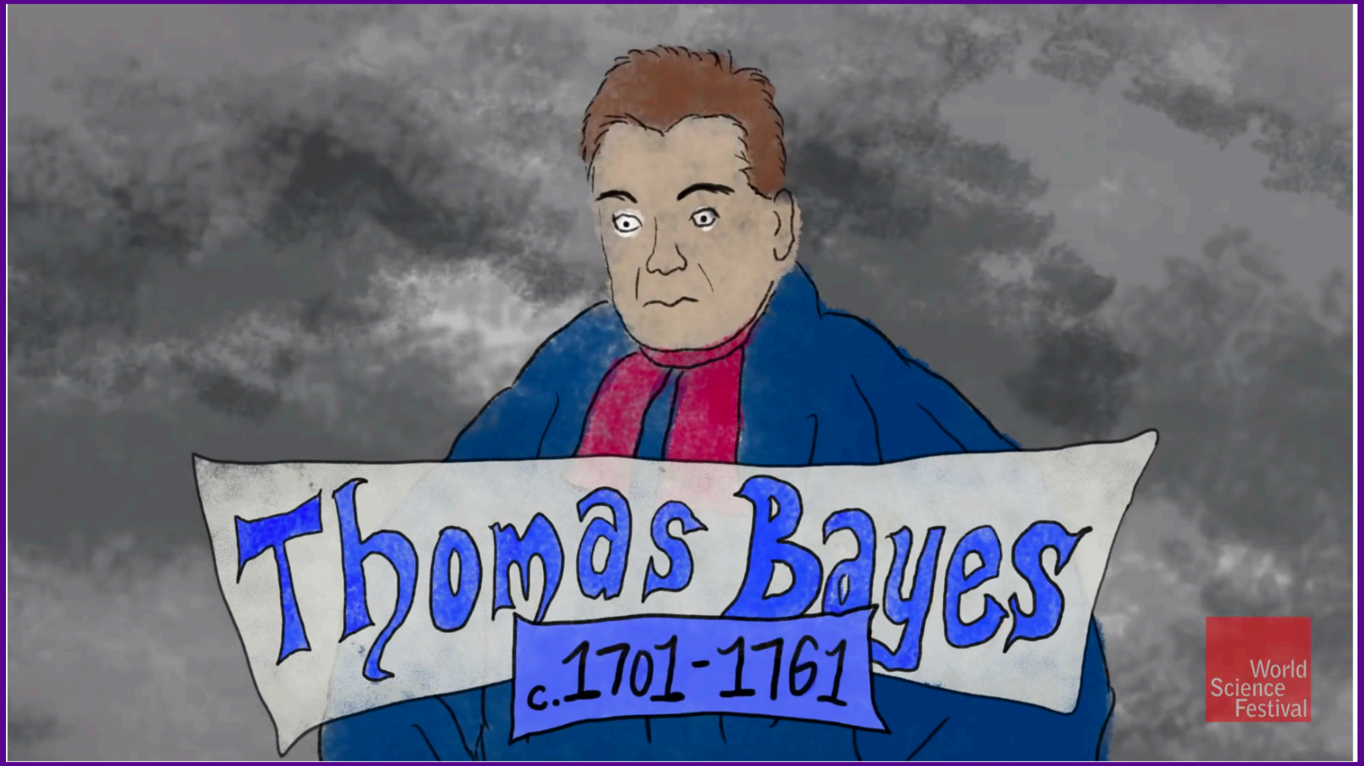
EXPECTED CALIBRATION ERROR (ECE)

ECE is the expected difference between model's confidence and its accuracy



"On Calibration of Modern Neural Networks" by Chuan Guo, Geoff Pleiss, Yu Sun and Kilian Q. Weinberger

BAYESIAN INFERENCE: A QUICK REVIEW



https://www.britannica.com/biography/Thomas-Bayes

BAYESIAN INFERENCE

- Likelihood $p(\mathcal{D}|\theta) = p(y|f(x;\theta))$
- Prior $p(\theta)$
  - Possibly implicit to the training method
- Posterior $p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \approx q(\theta|\mathcal{D})$
- Inference (Bayesian model averaging)

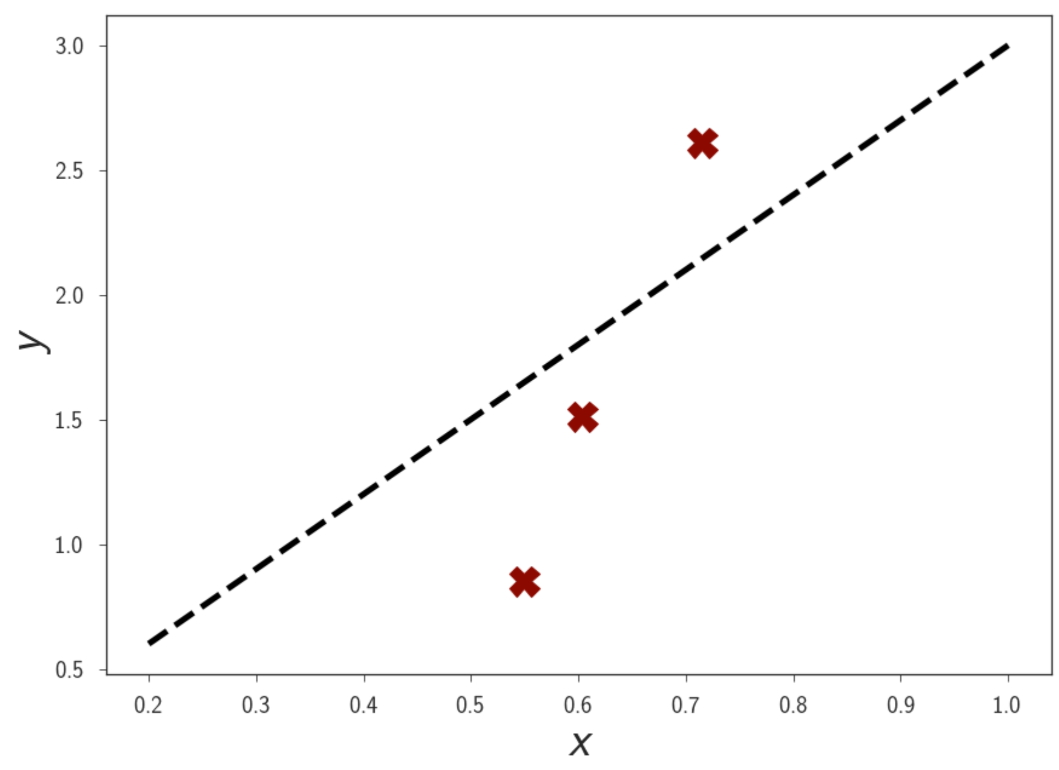
$$p(y^*|\mathcal{D}) = \mathbb{E}_{p(\theta|\mathcal{D})} p(y^*|\theta) \approx \frac{1}{K}\sum_{k=1}^{K} p(y^*|\theta_k)$$

$$\theta_k \sim q(\theta|\mathcal{D})$$

BAYESIAN MACHINE LEARNING

Consider a simple linear regression problem:

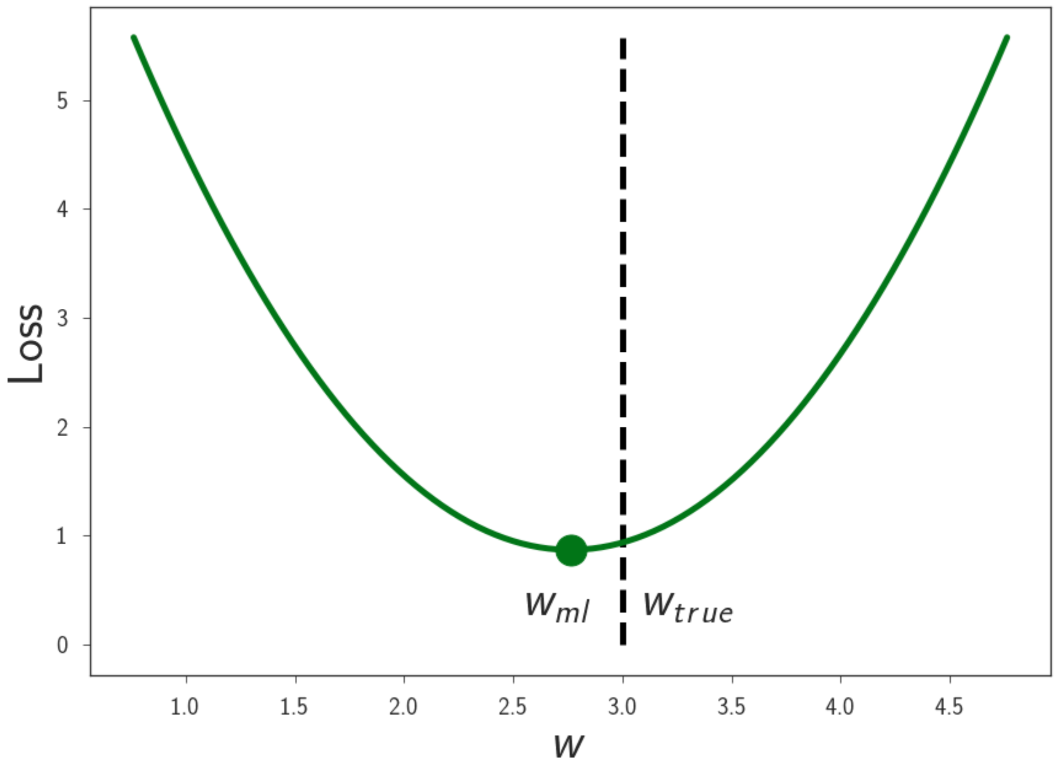
$$y = wx + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

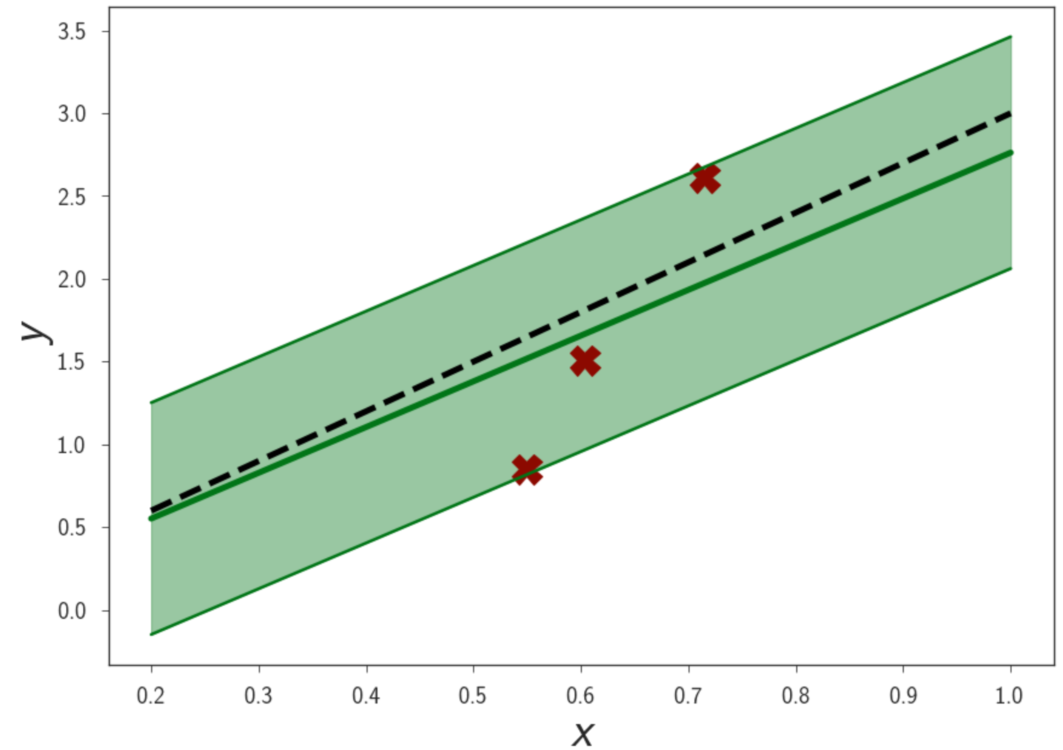


BAYESIAN MACHINE LEARNING

Standard linear regression:

$$\max_w \sum_{i=1}^{N} \log \mathcal{N}(y_i | w x_i, \sigma^2) \quad \Longleftrightarrow \quad \min_w \frac{1}{N} \sum_{i=1}^{N} (y_i - w x_i)^2$$



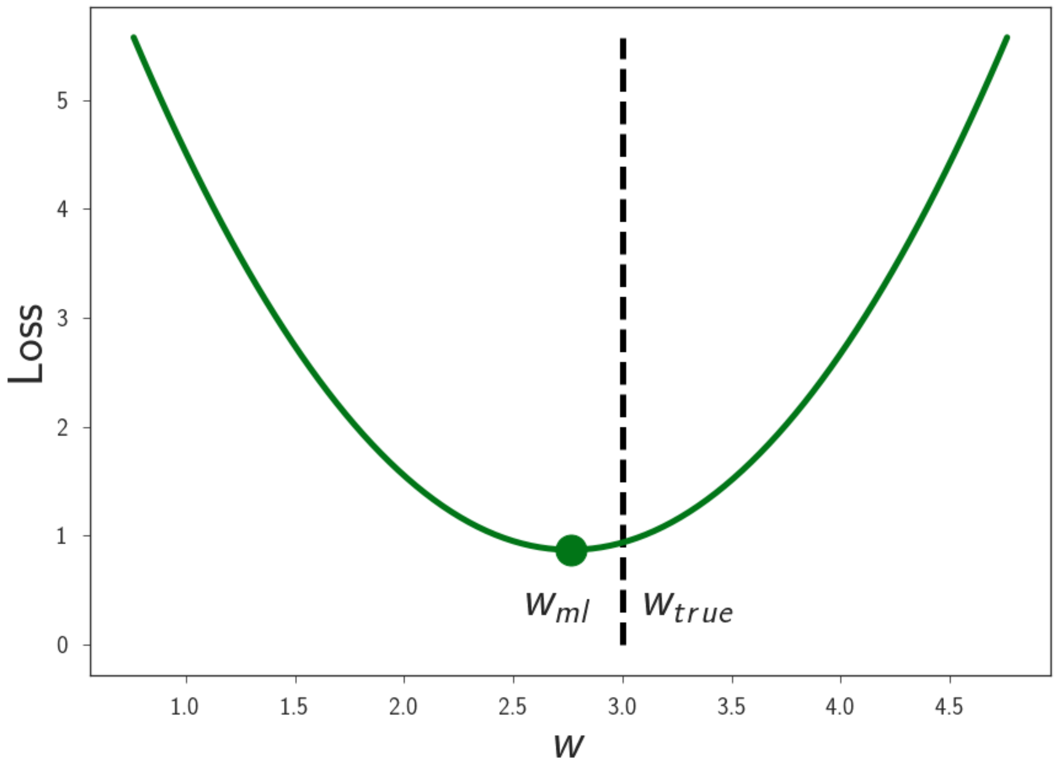


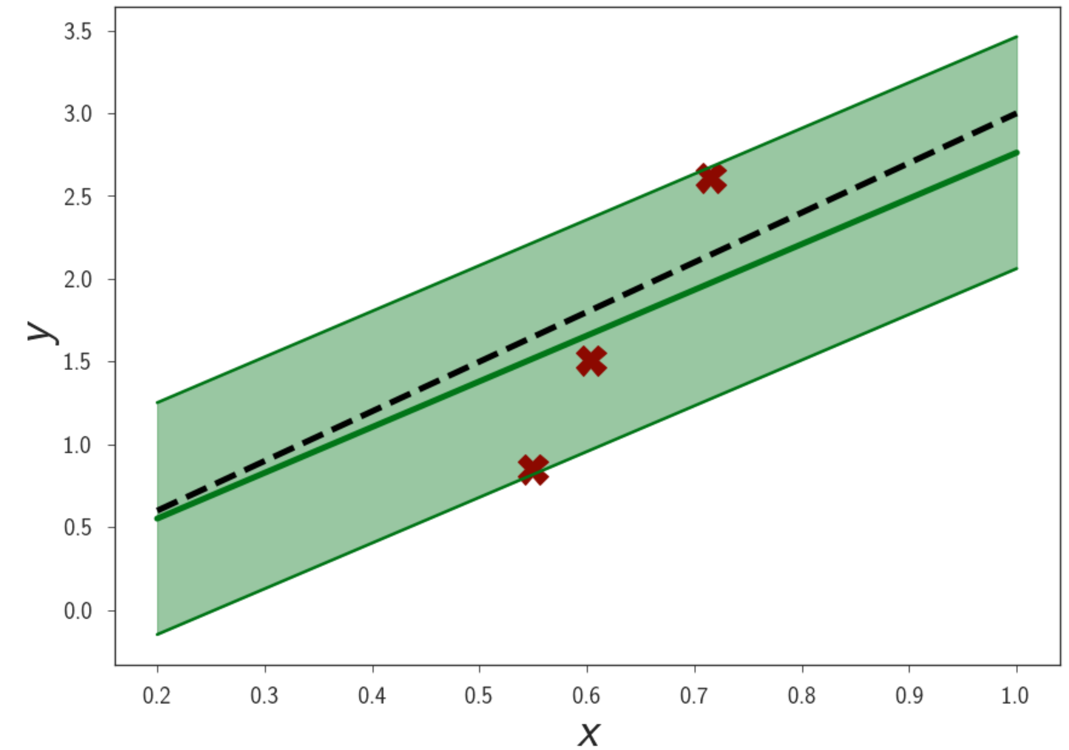
BAYESIAN MACHINE LEARNING

Standard linear regression:

$$\max_w \sum_{i=1}^{N} \log \mathcal{N}(y_i | w x_i, \sigma^2) \quad \Longleftrightarrow \quad \min_w \frac{1}{N} \sum_{i=1}^{N} (y_i - w x_i)^2$$

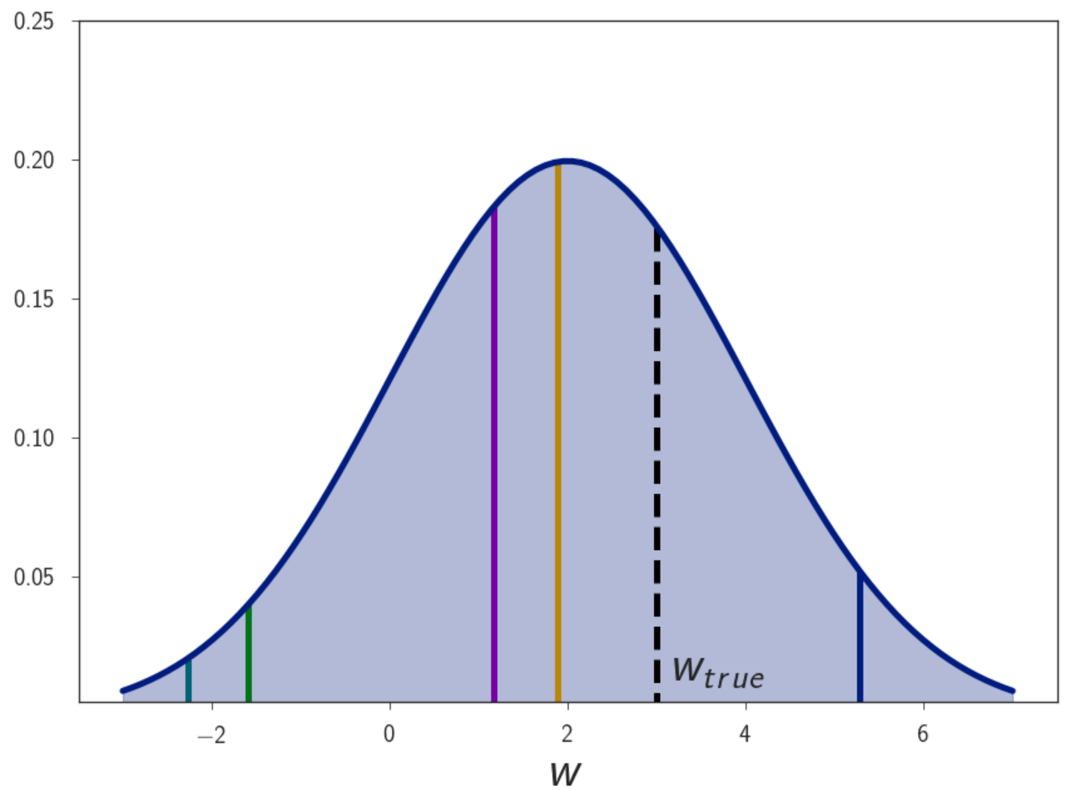
We want to model uncertainty over parameters of the model

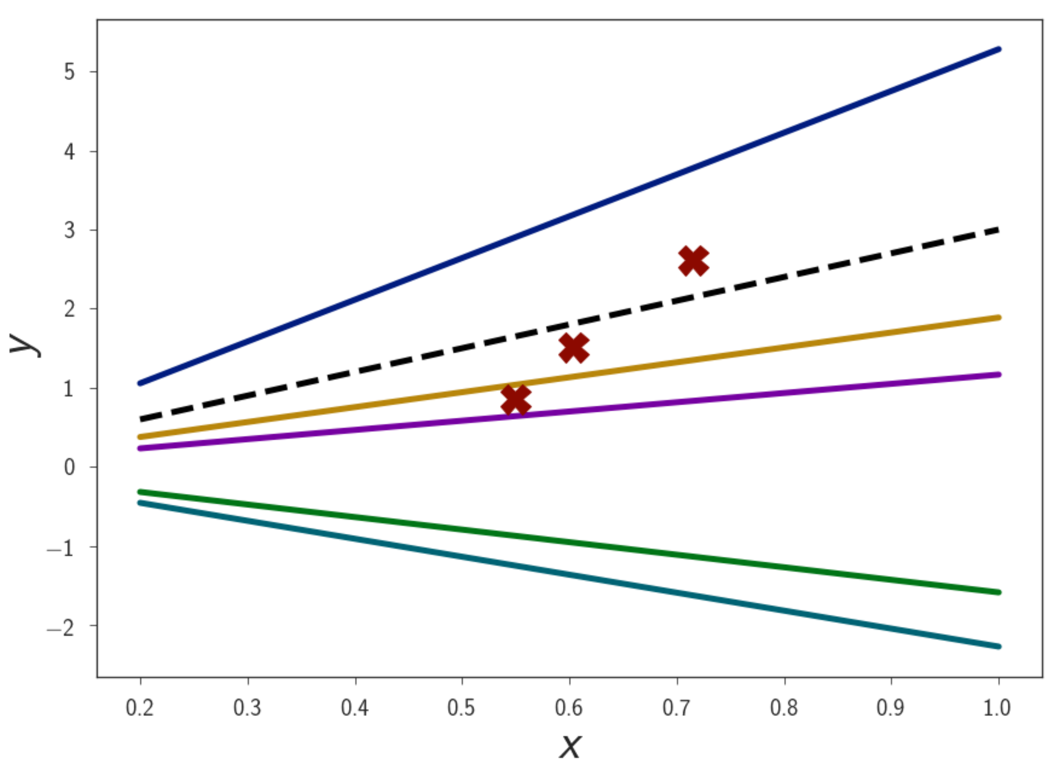




BAYESIAN LEARNING

Step 1: introduce a prior distribution $p(w)$ over parameters

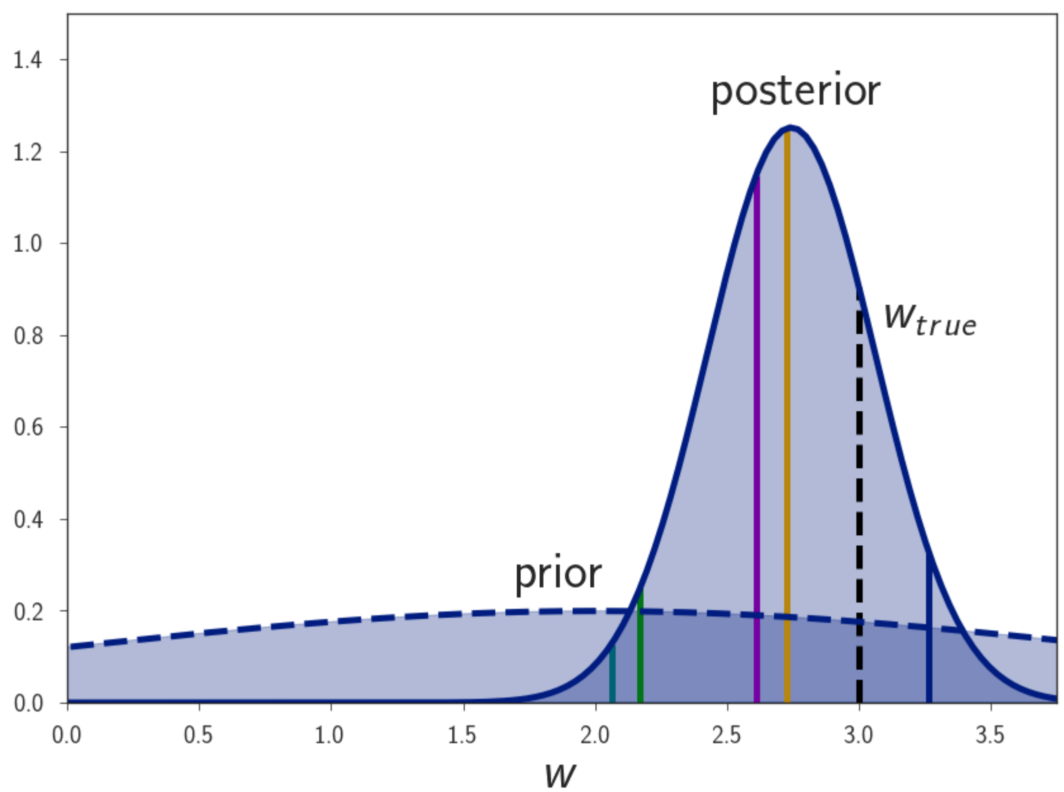


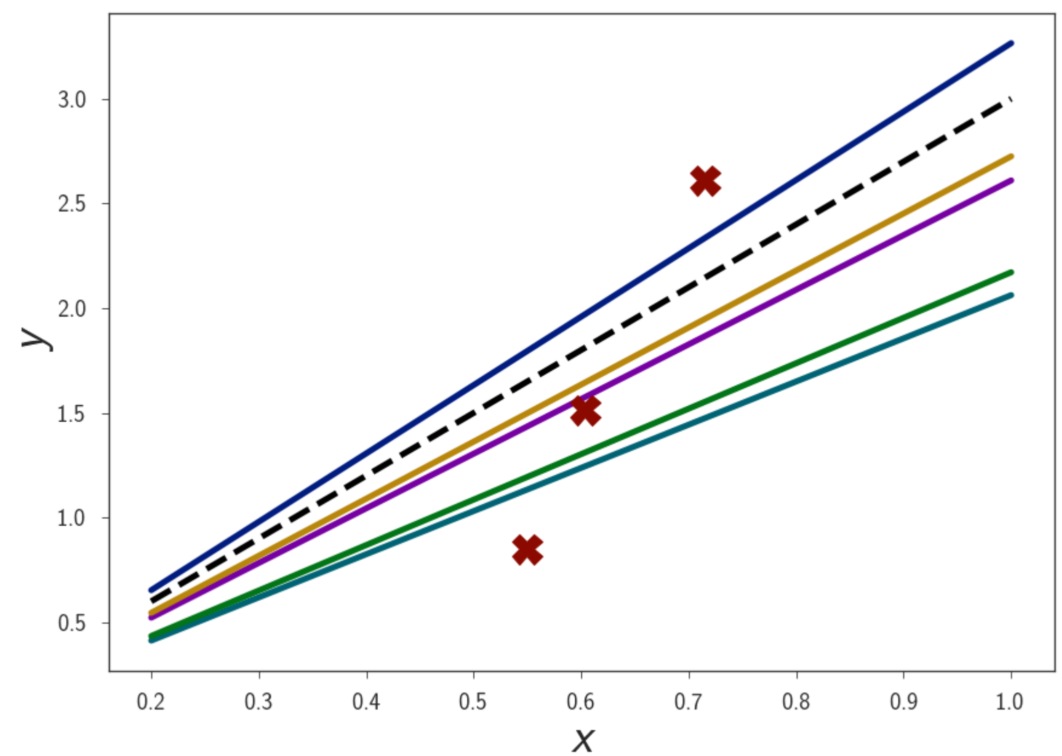


BAYESIAN LEARNING

Step 2: Compute posterior $p(w|D)$ using Bayes rule

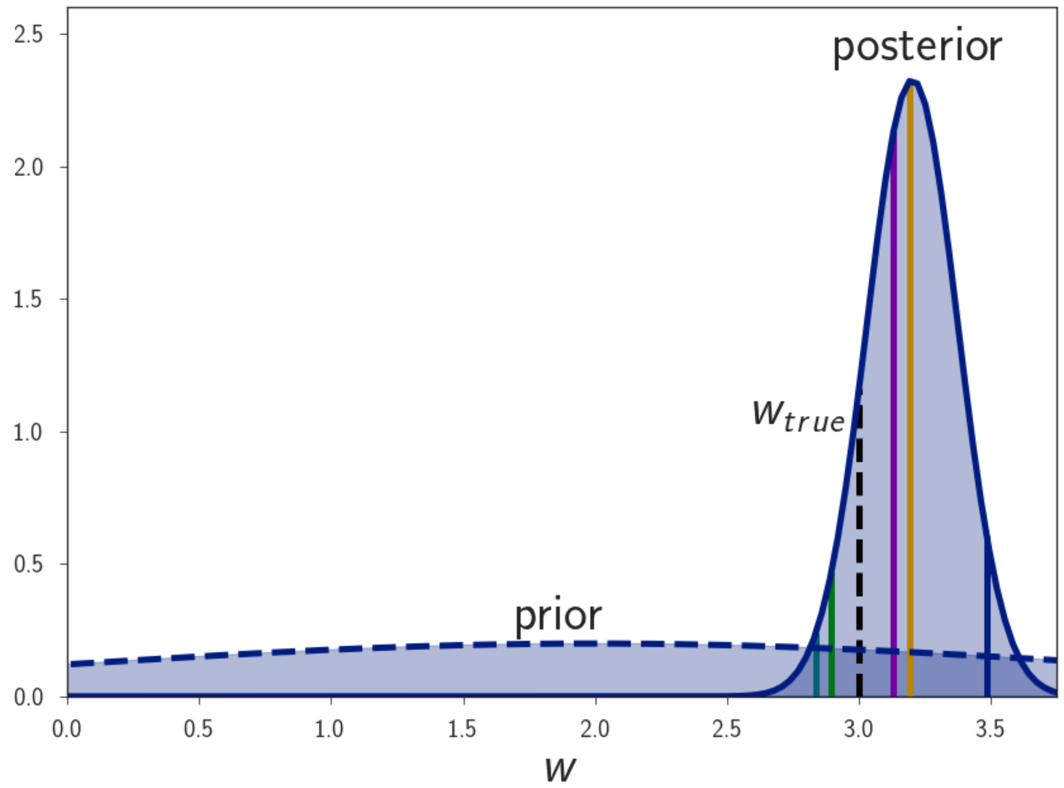
$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

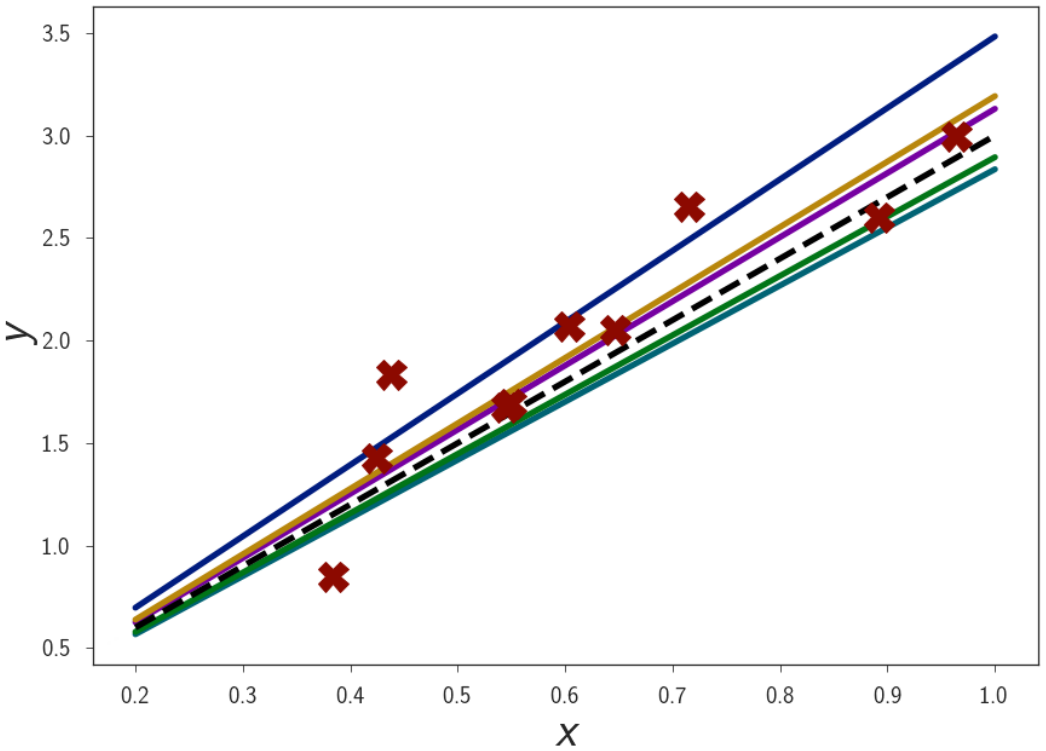




BAYESIAN LEARNING: POSTERIOR CONTRACTION (1)

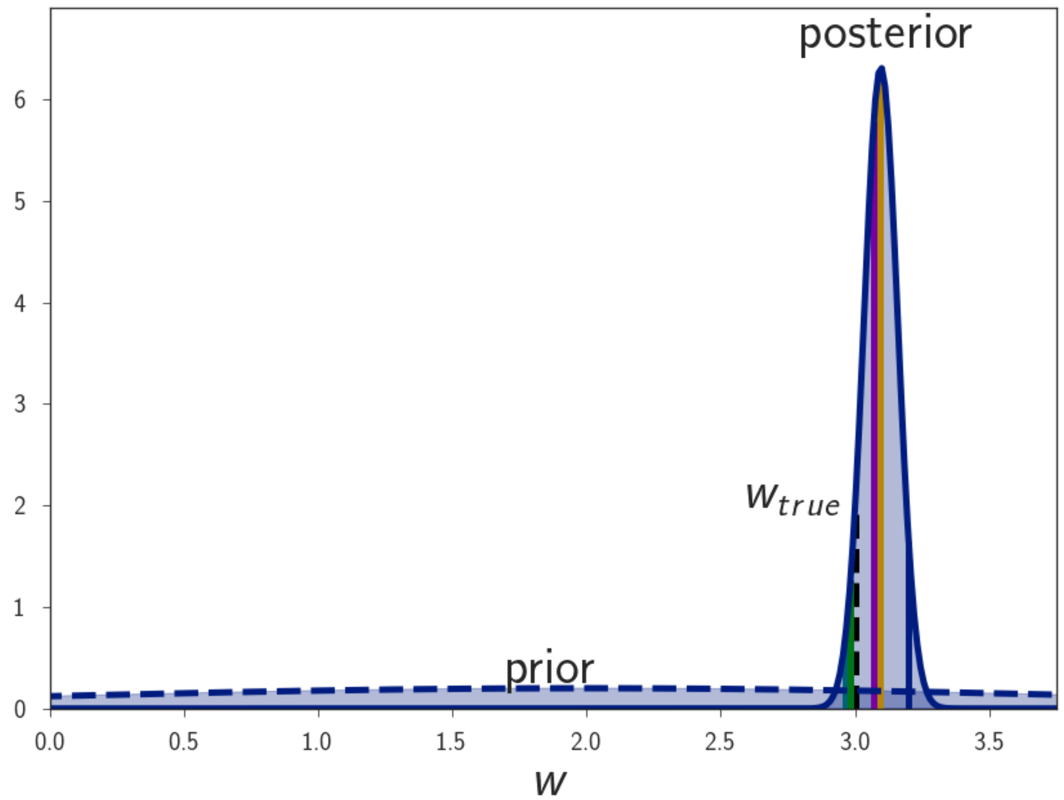
$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

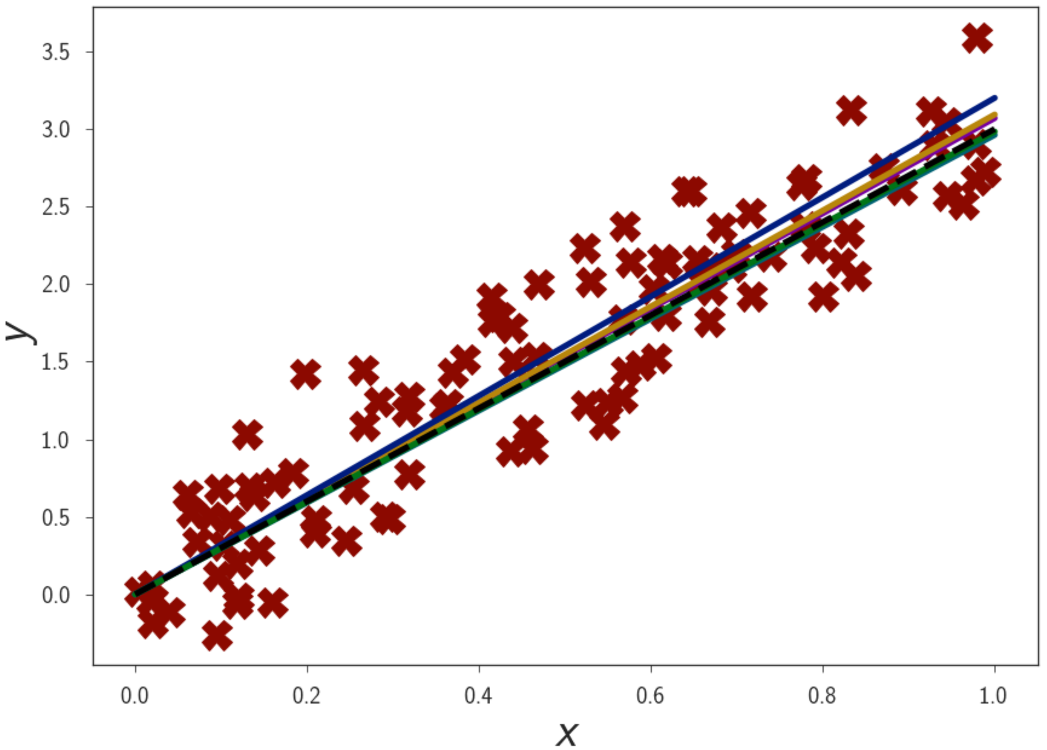




BAYESIAN LEARNING: POSTERIOR CONTRACTION (2)

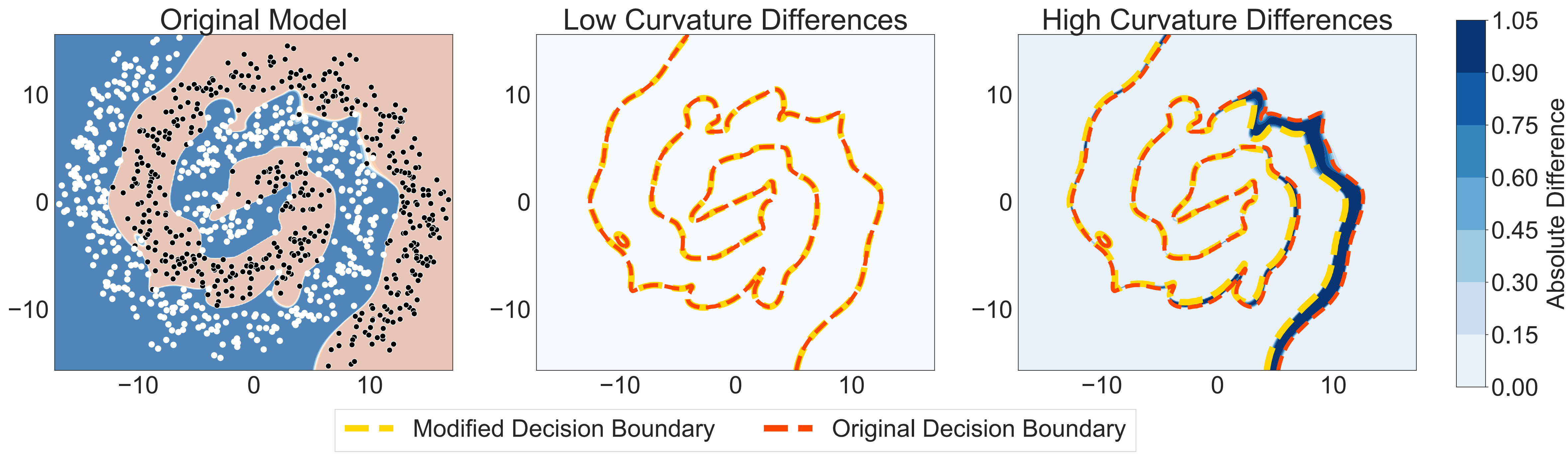
$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$





BAYESIAN LEARNING: POSTERIOR CONTRACTION (3)

What happens if we have more parameters than data points???



Theorem (Function-Space Homogeneity in Linear Models). *Let* $\Phi = \Phi(x) \in \mathbb{R}^{n\times k}$ *be a feature map of* n *data observations,* x*, with* $n < k$ *and assign isotropic prior* $\beta \sim \mathcal{N}(0_k, S_0 = \alpha^2 I_k)$ *for parameters* $\beta \in \mathbb{R}^k$*. The minimal eigenvectors of the Hessian define a* $k - n$ *dimensional subspace in which parameters can be perturbed without changing the training predictions in function-space.*

Will revisit these results later….

Maddox, Benton, & Wilson, in preparation

BAYESIAN MODEL AVERAGING

- We combine aleatoric and epistemic uncertainties via BMA:

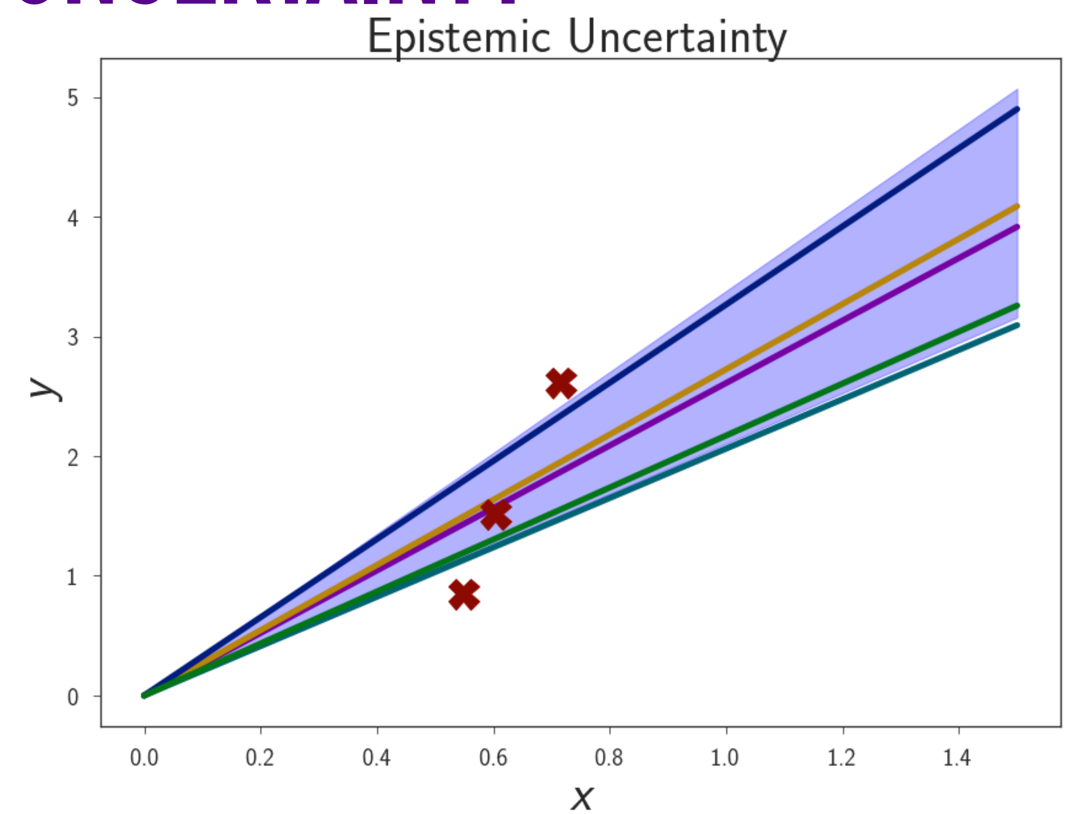
$$p(y^*|x^*, D) = \int_w p(y^*|x^*, w)p(w|D)dw$$

- Ignoring the uncertainty in the posterior over w leads to overconfident predictions

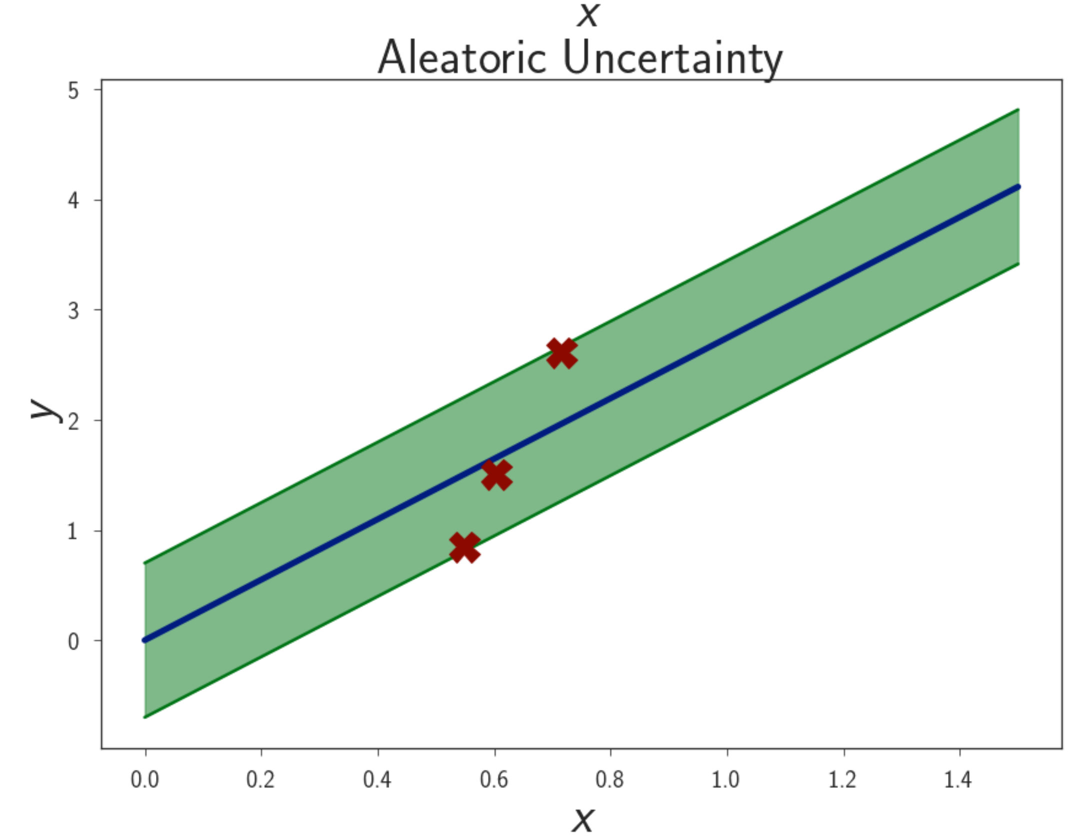
BAYESIAN LEARNING: TWO TYPES OF UNCERTAINTY

Epistemic uncertainty is our uncertainty over the model

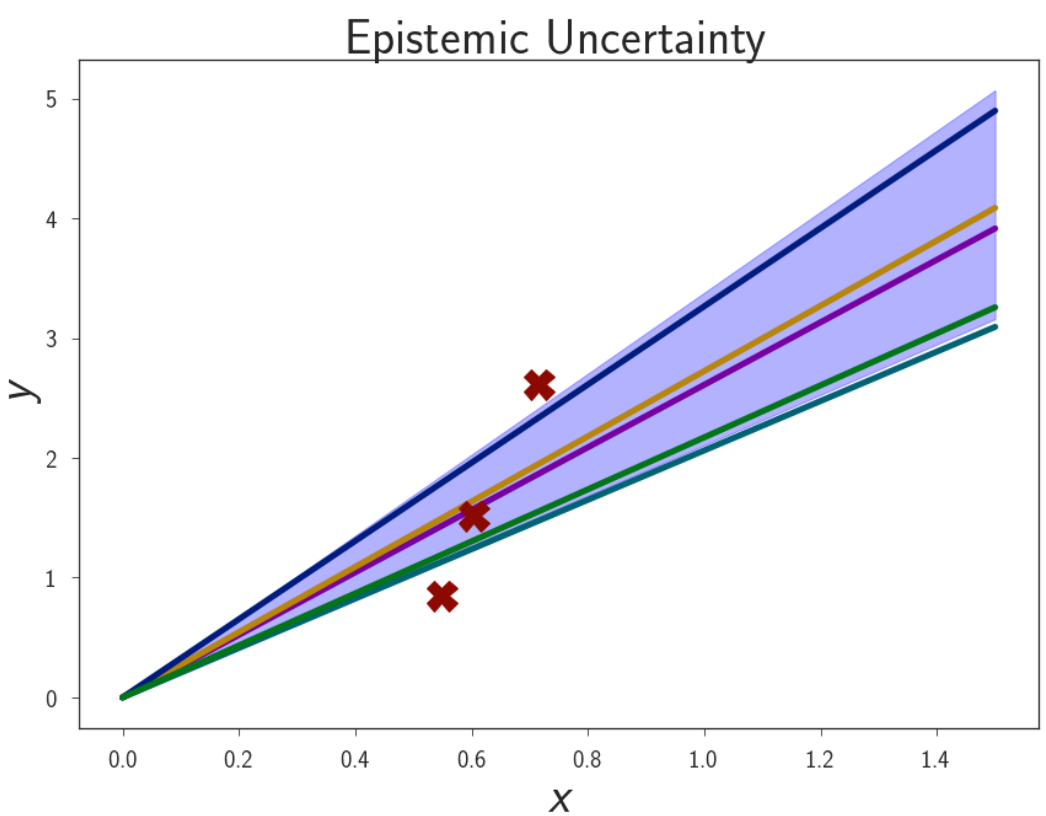
- Grows with x because uncertainty in w is multiplied by x



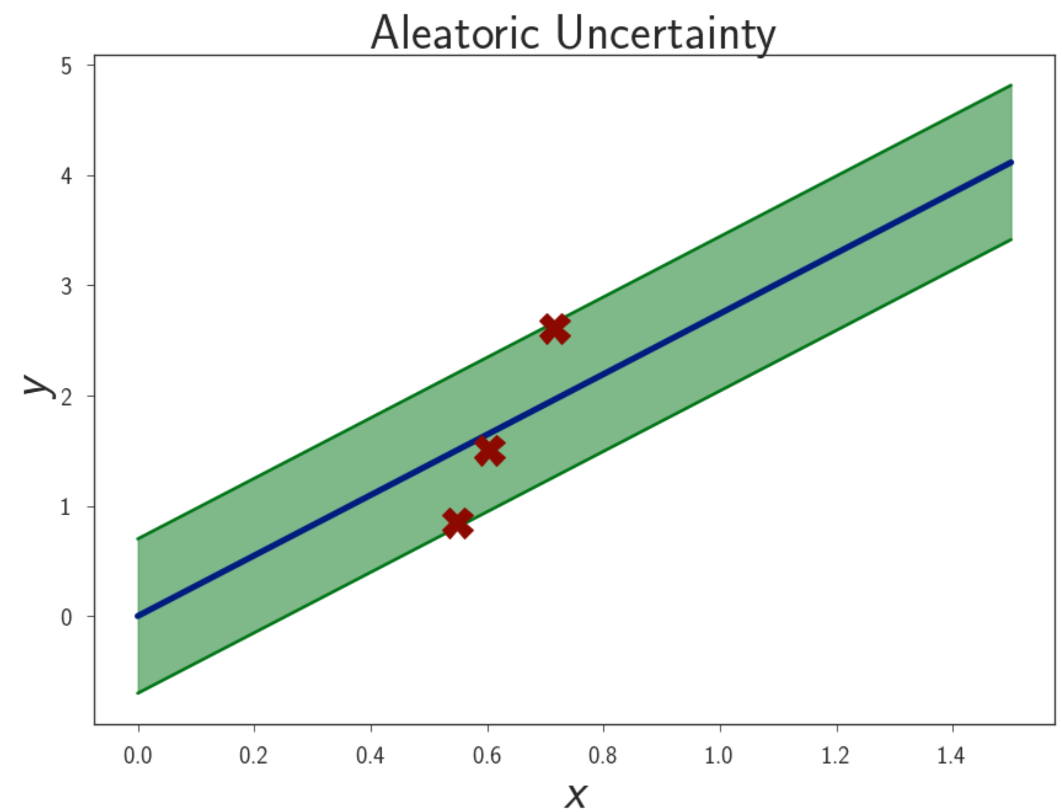
Aleatoric uncertainty is our uncertainty over the data for a fixed model, e.g. noise.

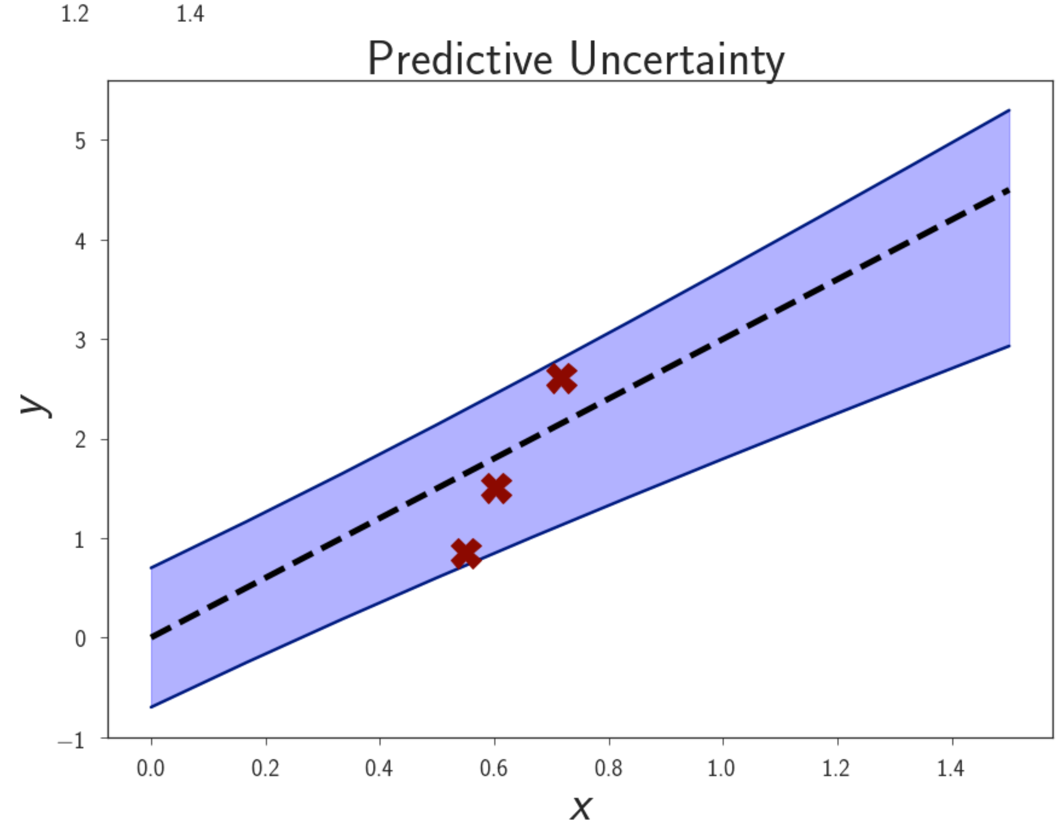


BAYESIAN LEARNING: BAYESIAN MODEL AVERAGING



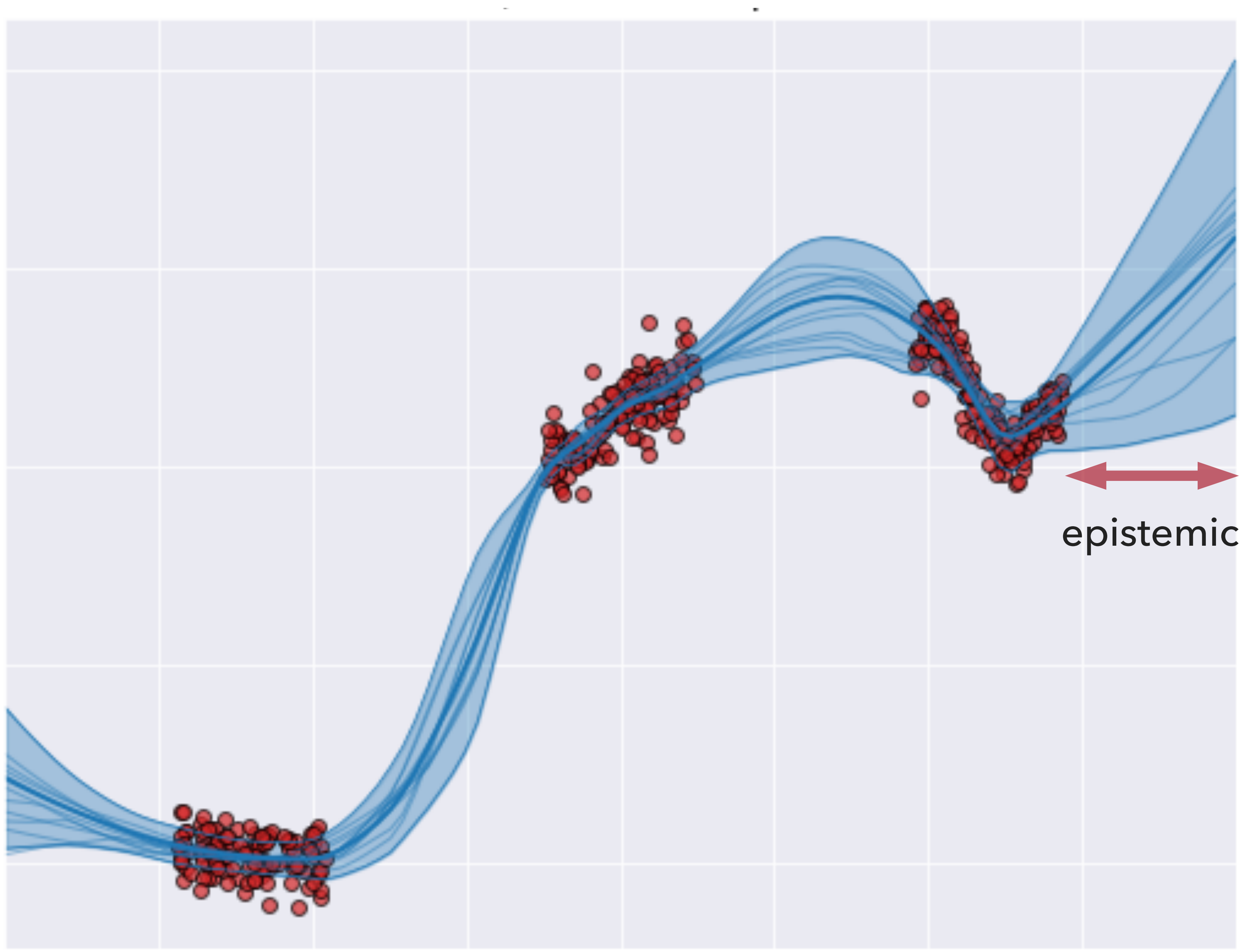
+





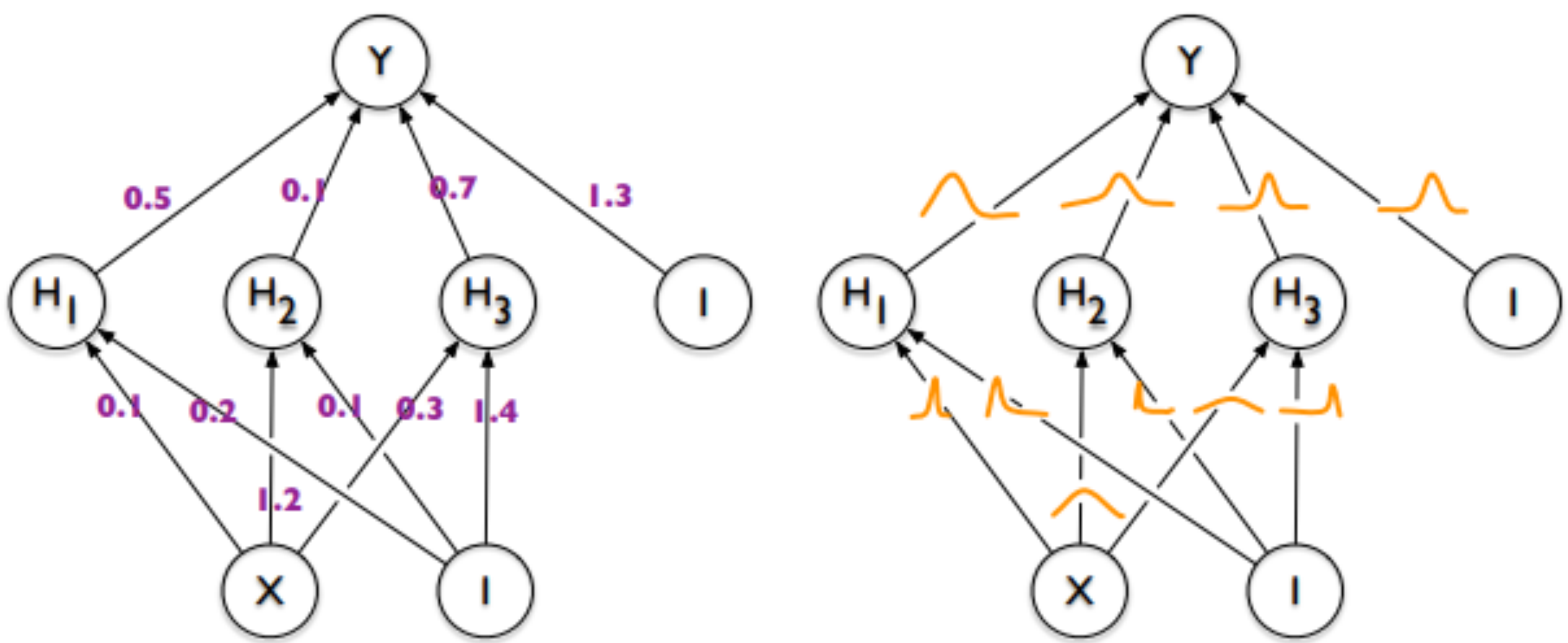
BAYESIAN LEARNING: TWO TYPES OF UNCERTAINTY

Epistemic uncertainty: non-linear model



BAYESIAN DEEP LEARNING

- In Bayesian deep learning we model posterior distribution over the weights of neural networks
- In theory, leads to better predictions and well-calibrated uncertainty



Standard DNN Bayesian DNN

"Weight Uncertainty in Neural Networks" by Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, Daan Wierstra

BAYESIAN DEEP LEARNING: CHALLENGES

Bayesian inference for deep neural networks is extremely challenging

- Posterior is intractable
- Millions of parameters
- Large datasets
- Unclear which priors to use

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} = \frac{p(D|w)p(w)}{\int_{w'} p(D|w')p(w')dw'}$$

BAYESIAN DEEP LEARNING: CHALLENGES

Bayesian inference for deep neural networks is extremely challenging

- Posterior is intractable — Is the likelihood correct?
- Millions of parameters — What do these parameters mean?
- Large datasets — Can we run MCMC for 1 million steps on ImageNet??
- Unclear which priors to use — Is the prior correct?

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} = \frac{p(D|w)p(w)}{\int_{w'} p(D|w')p(w')dw'}$$

BAYESIAN DEEP LEARNING: CHALLENGES

Bayesian inference for deep neural networks is extremely challenging

- Posterior is intractable
- Millions of parameters
- Large datasets
- Unclear which priors to use

Is the likelihood correct? Probably

What do these parameters mean?
Care about functions instead

Can we run MCMC for 1 million steps on ImageNet??
We don't need to

Is the prior correct?
Probably

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} = \frac{p(D|w)p(w)}{\int_{w'} p(D|w')p(w')dw'}$$

BAYESIAN DEEP LEARNING: CHALLENGES

Bayesian inference for deep neural networks is extremely challenging

- Posterior is intractable
- Millions of parameters
- Large datasets
- Unclear which priors to use

Is the likelihood correct? Probably

What do these parameters mean?
Care about functions instead

Can we run MCMC for 1 million steps on ImageNet??

We don't need to

Is the prior correct?

Probably

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} = \frac{p(D|w)p(w)}{\int_{w'} p(D|w')p(w')dw'}$$

APPROXIMATE INFERENCE

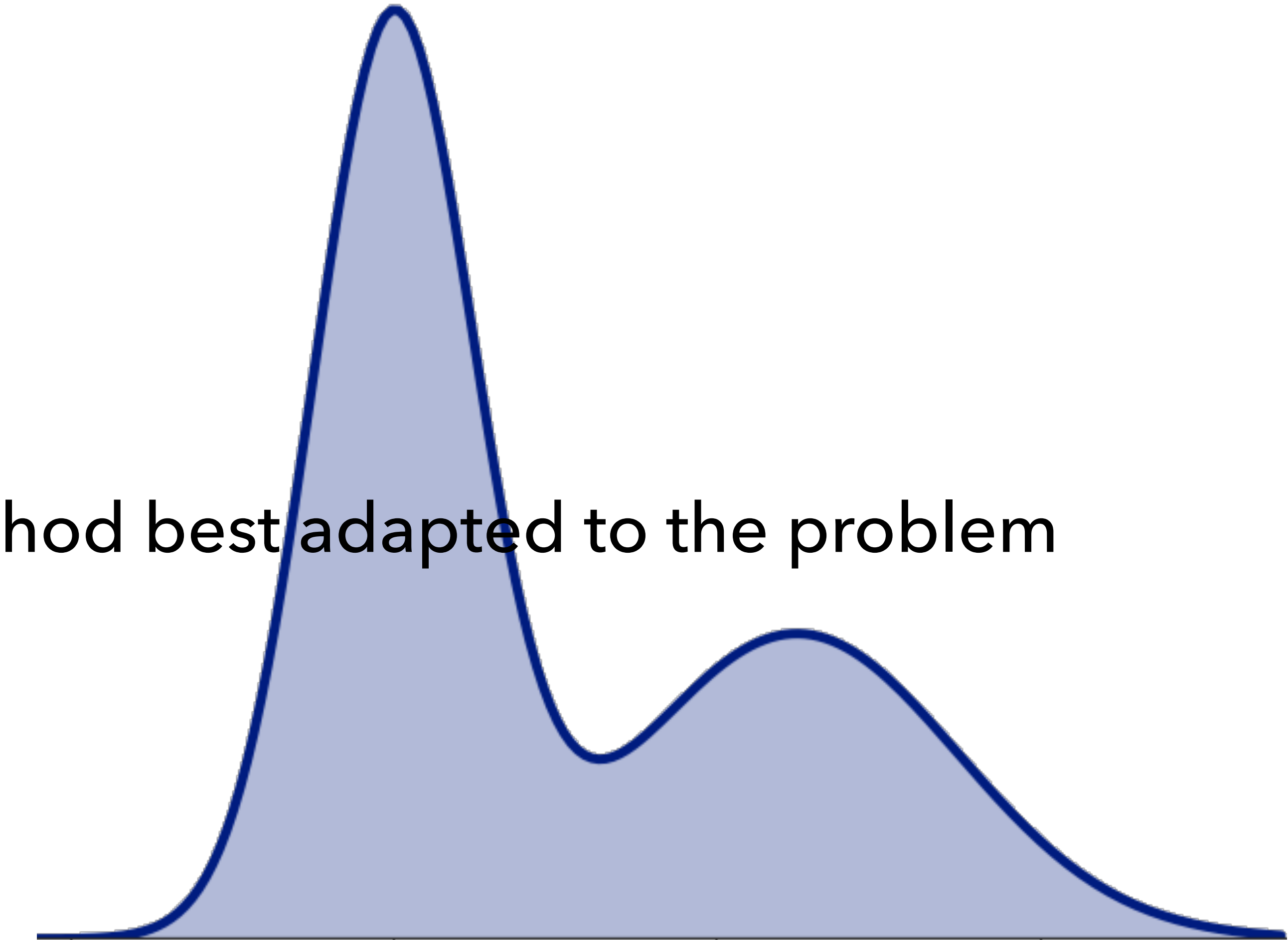
http://www.understandingrelationships.com/wp-content/uploads/Fotolia_70137791_Subscription_Monthly_M.jpg

HOW CAN WE DO APPROXIMATE BAYESIAN INFERENCE?

Posterior Approximation:

- Laplace Approximation
- Variational Inference
- Markov Chain Monte Carlo
- Geometrically Inspired Methods

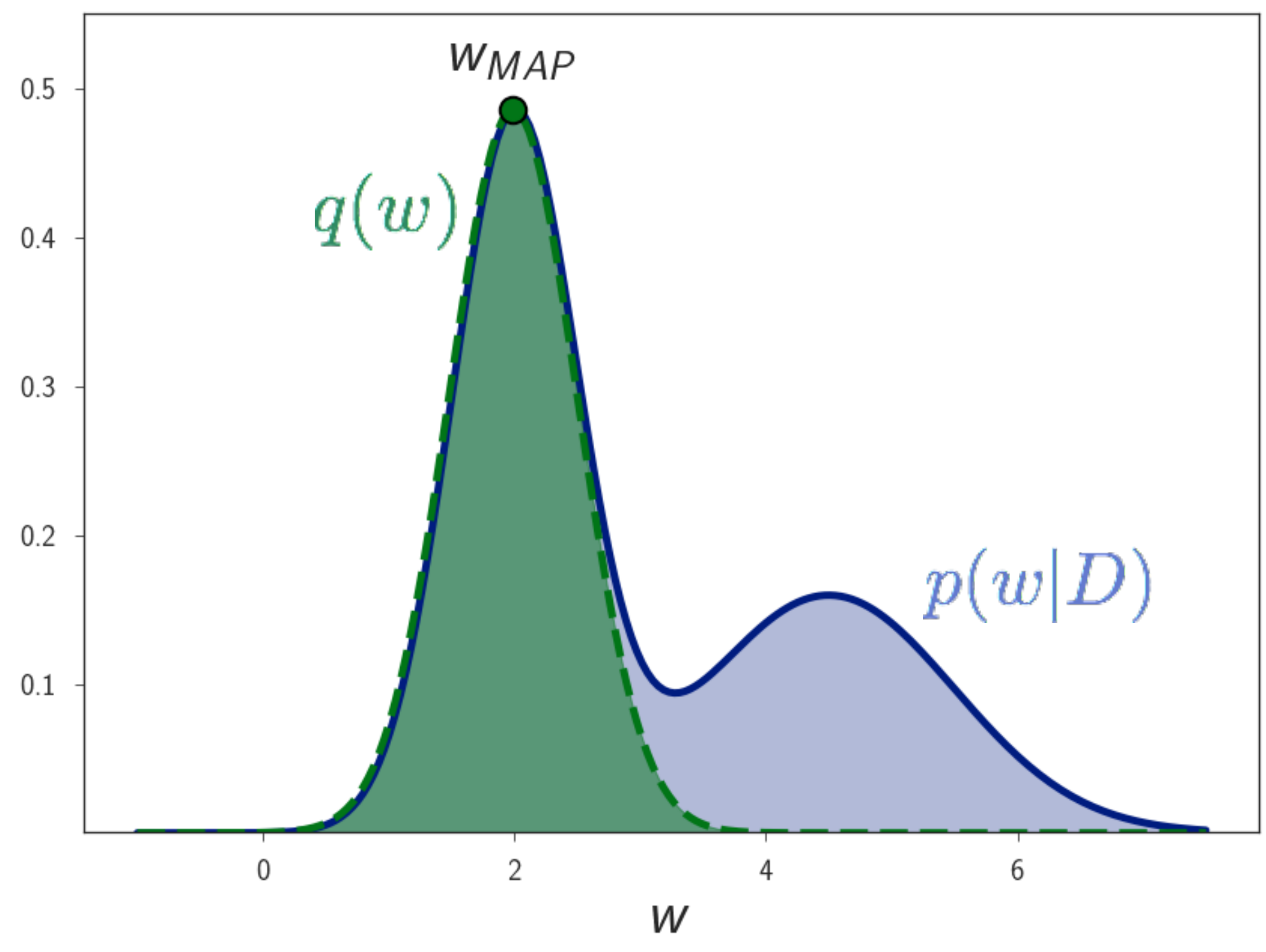
There's no one best method - use the method best adapted to the problem



LAPLACE APPROXIMATION

Approximate posterior with a Gaussian $\mathcal{N}(w|\mu, A^{-1})$

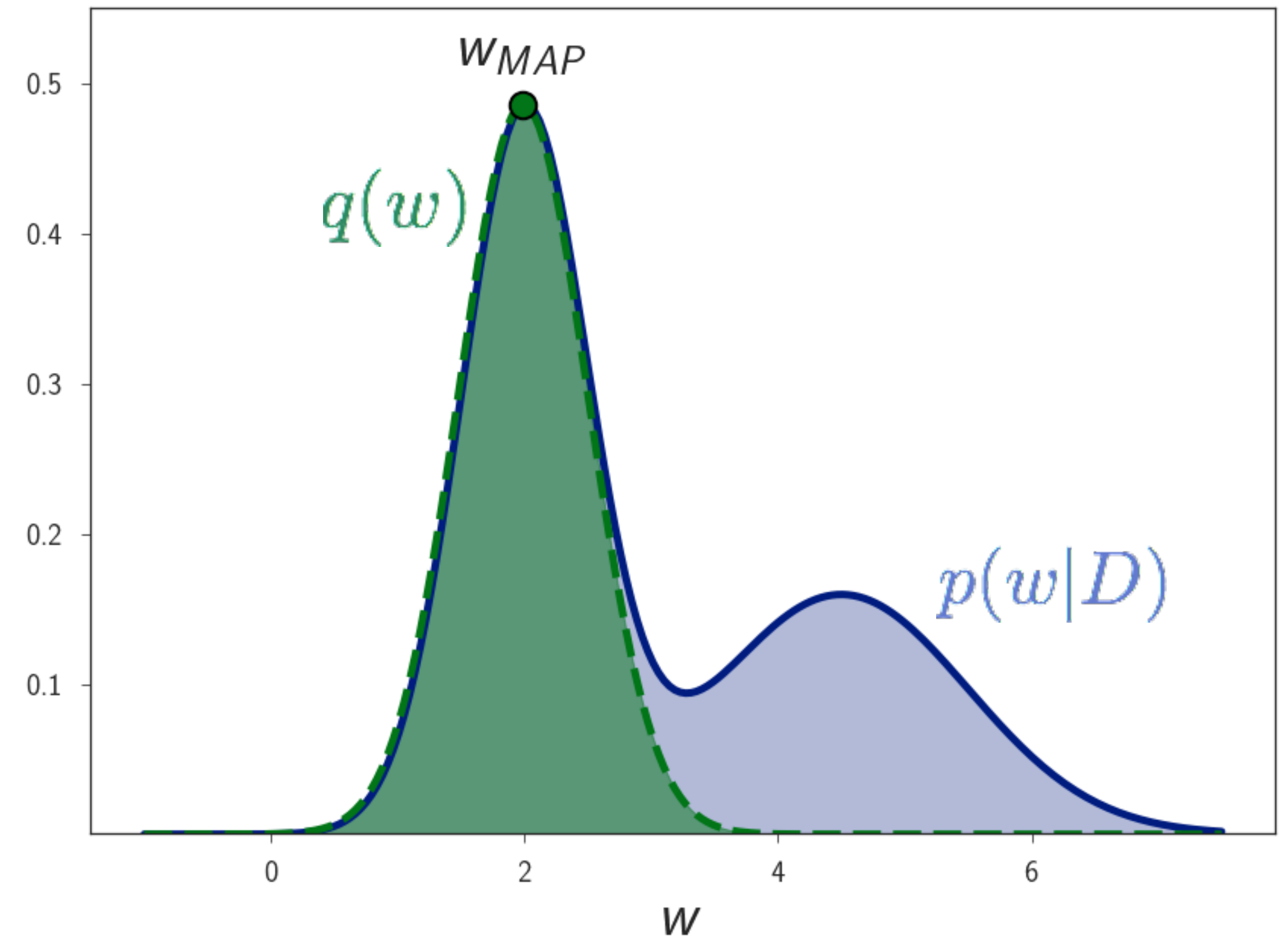
- $w = w_{MAP}$ mode (local maximum) of $p(w|D)$
- $A = -\nabla\nabla \log[p(D|w)p(w)]$
- Only captures a single mode



LAPLACE APPROXIMATION

Approximate posterior with a Gaussian $\mathcal{N}(w|\mu, A^{-1})$

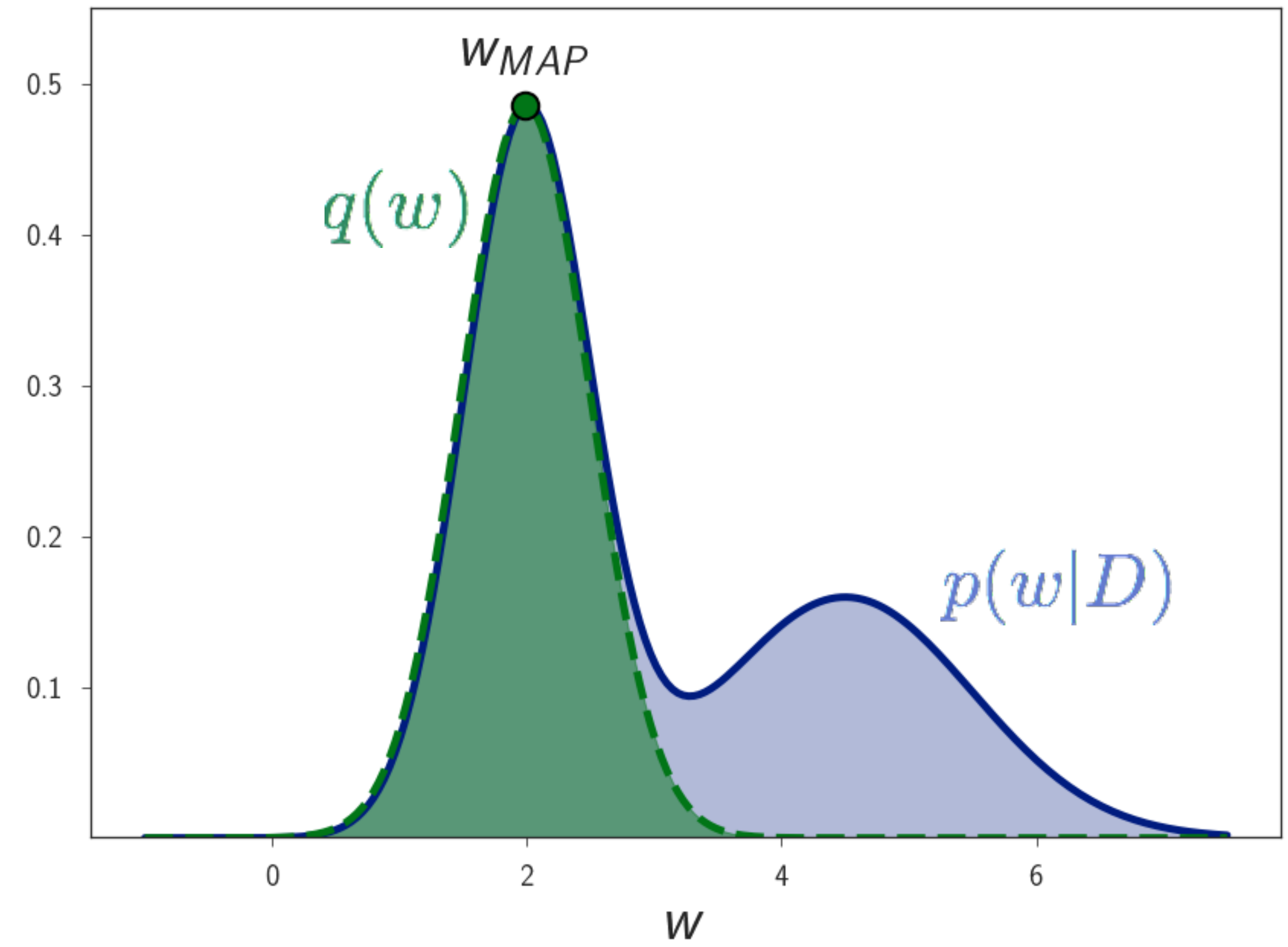
- $w = w_{MAP}$ mode (local maximum) of $p(w|D)$
- $A = -\nabla\nabla \log[p(D|w)p(w)]$
- Only captures a single mode
- **Is a single mode a bad thing?**



LAPLACE APPROXIMATION

Approximate posterior with a Gaussian $\mathcal{N}(w|\mu, A^{-1})$

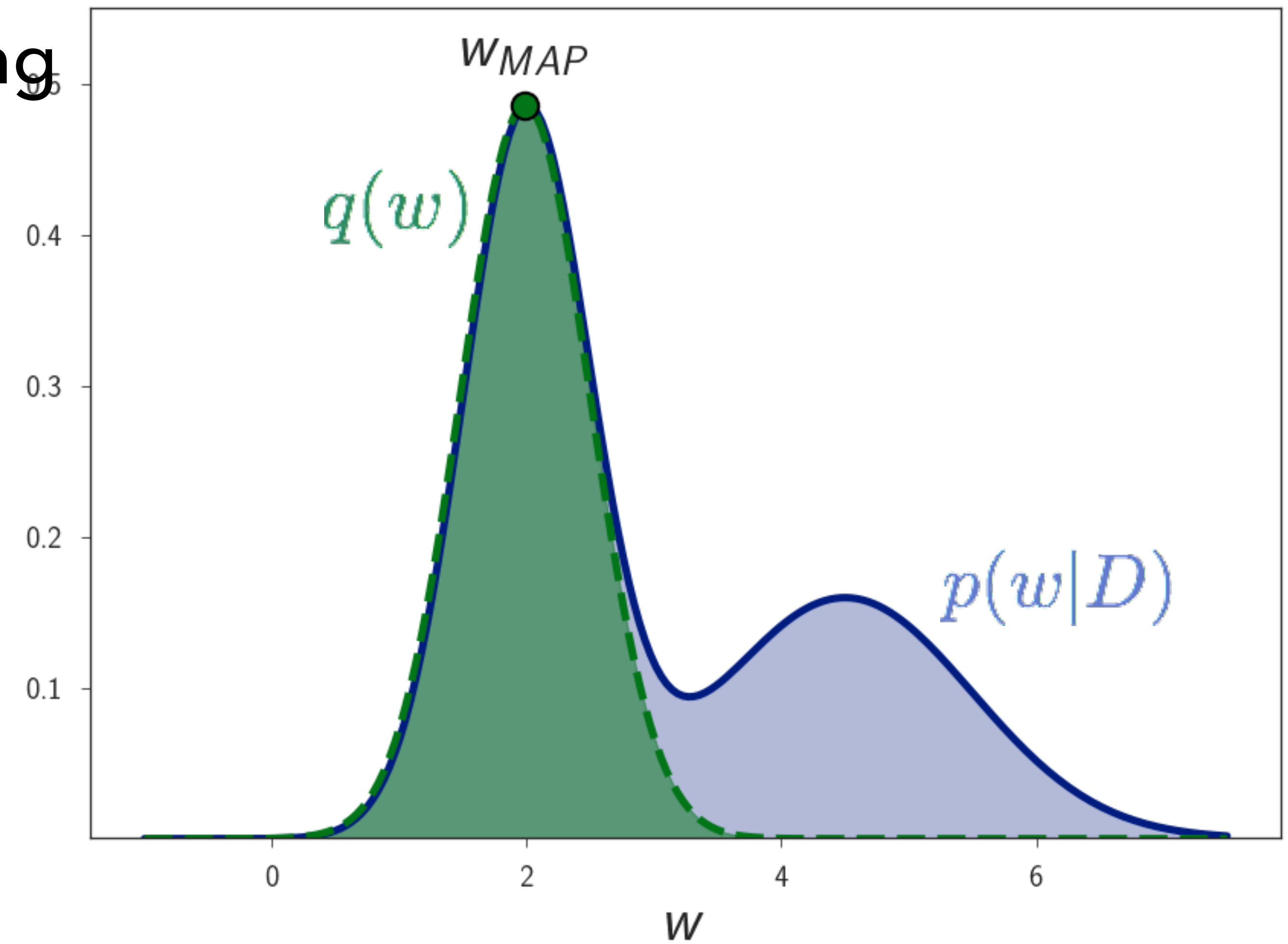
- $w = w_{MAP}$ mode (local maximum) of $p(w|D)$
- $A = -\nabla\nabla \log[p(D|w)p(w)]$
- Only captures a single mode
- **Is a single mode a bad thing?**



LAPLACE APPROXIMATION: DEEP LEARNING

Approximate posterior with a Gaussian $\mathcal{N}(w|\mu, A^{-1})$

- $w = w_{MAP}$ mode (local maximum) of $p(w|D)$
- Approximate A with a KFAC (tri-diagonal) – Ritter et al., 2018a
- Application: Catastrophic forgetting (Ritter et al, 2018b)
- Originally from Mackay, '92

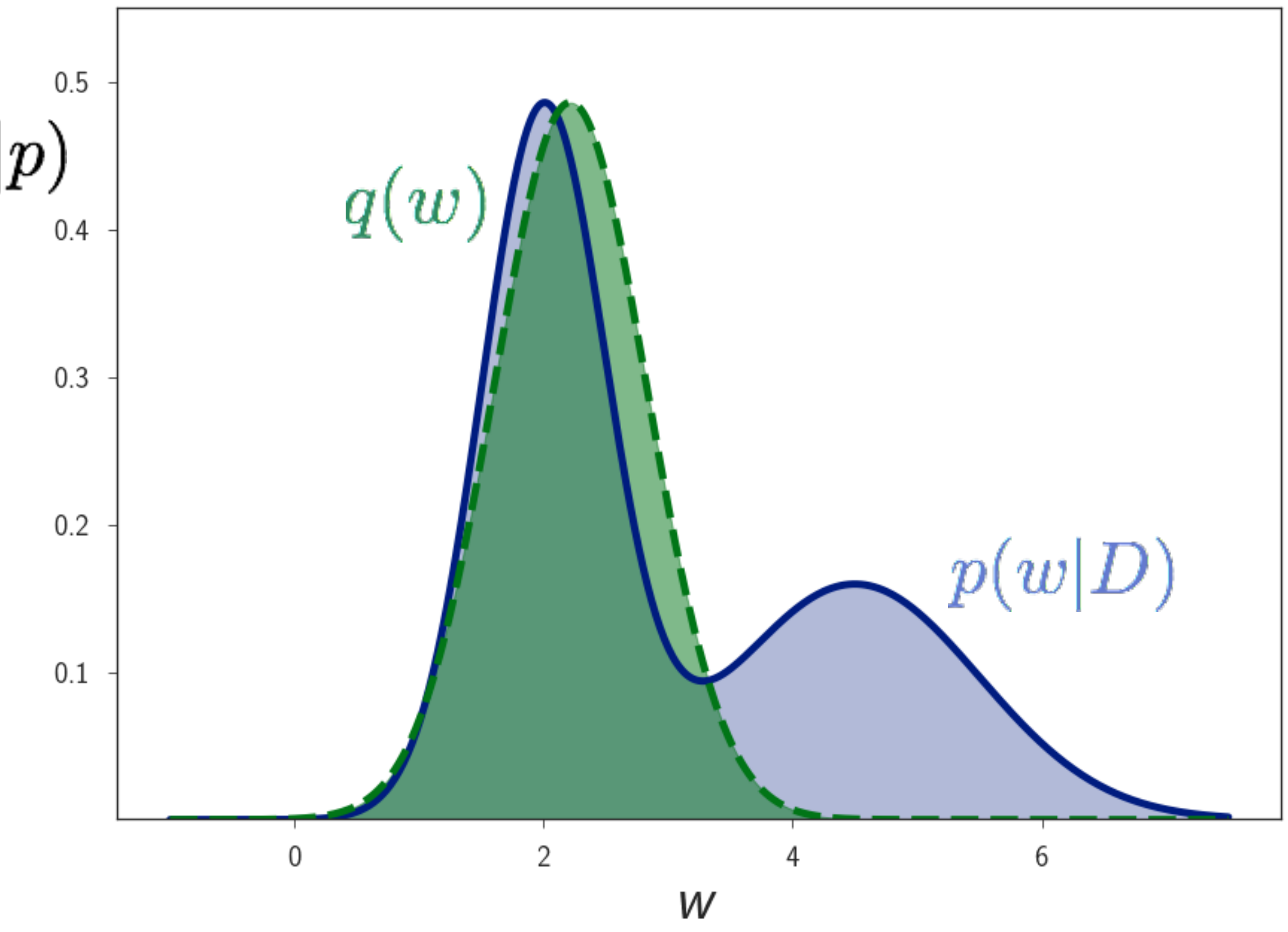


VARIATIONAL INFERENCE

We can find the best approximating distribution within a given family with respect to KL-divergence

- $KL(q||p) = \int_w q(w) \log \frac{q(w)}{p(w|D)} dw$

- If $q = \mathcal{N}(\mu, \Sigma)$, then $\min_{\mu,\Sigma} KL(q||p)$



VARIATIONAL INFERENCE

We can find the best approximating distribution within a given family with respect to KL-divergence

- $KL(q||p) = \int_w q(w) \log \frac{q(w)}{p(w|D)} dw$
- Stochastic variational inference (Hoffman et al, '13, Kucelkibir, et al, '17, Graves, 2011)

$$ELBO(w) = E_{q(w)}(\log p(\mathcal{D}|w)) - KL(q(w)||p(w))$$

Traditionally… $q(w) = \mathcal{N}(\mu_i, \sigma_i^2)$

- Minimizing the KL divergence is "consistent" statistically (Wang & Blei, '19) & optimal in other settings (Knoblauch, et al, '19)

- Can somewhat evaluate if it works (Yao, et al, '18)

VARIATIONAL INFERENCE: DEEP LEARNING (2)

We can find the best approximating distribution within a given family with respect to KL-divergence

- $KL(q||p) = \int_w q(w) \log \frac{q(w)}{p(w|D)} dw$
- Other bounds exist…
  - Chi-Square (Dieng, et al, '17), F-divergences (Wang, et al, '17), Perturbative divergences (Bamler et al, '17), VPNG (Tang & Ranganath, '19)
- Better approximation distributions….
  - Matrix-variate Gaussians (Louizos, et al, '16), Normalizing flows (Louizos, et al, '17), Bayes by Backprop (Blundell, et al, 16)

VARIATIONAL INFERENCE: DEEP LEARNING (3)

We can find the best approximating distribution within a given family with respect to KL-divergence

- $KL(q||p) = \int_w q(w) \log \frac{q(w)}{p(w|D)} dw$
- Other bounds exist…
  - Renyi-divergences (Li & Turner, '16), robust divergences (Futami, et al, '17), operator divergences (Ranganath, et al, 16), etc…
- Better approximation distributions….
  - Implicit distributions (Tran, Ranganath, Blei, '17), GANs (Huszar, '17), boosting (Miller et al, '17), smoothed dropout (Gal, et al, '17), etc….

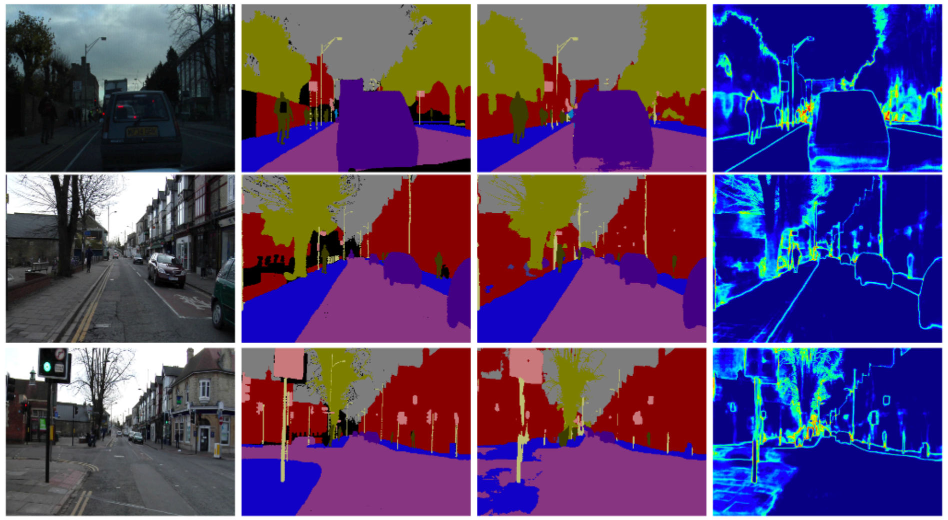
VARIATIONAL INFERENCE: DEEP LEARNING

We can find the best approximating distribution within a given family with respect to KL-divergence

- $KL(q||p) = \int_w q(w) \log \frac{q(w)}{p(w|D)} dw$
- Dropout at test time (Gal & Ghahramani, '15, Gal & Ghahramani, '16, Gal, '16, Gal & Li, '17)
    - $q(w) = Bernoulli(p)\mathcal{N}(\mu_i, \sigma)$
    - KL un-defined so it's actually minimizing a quasi-KL divergence… (Hron et al, '18)

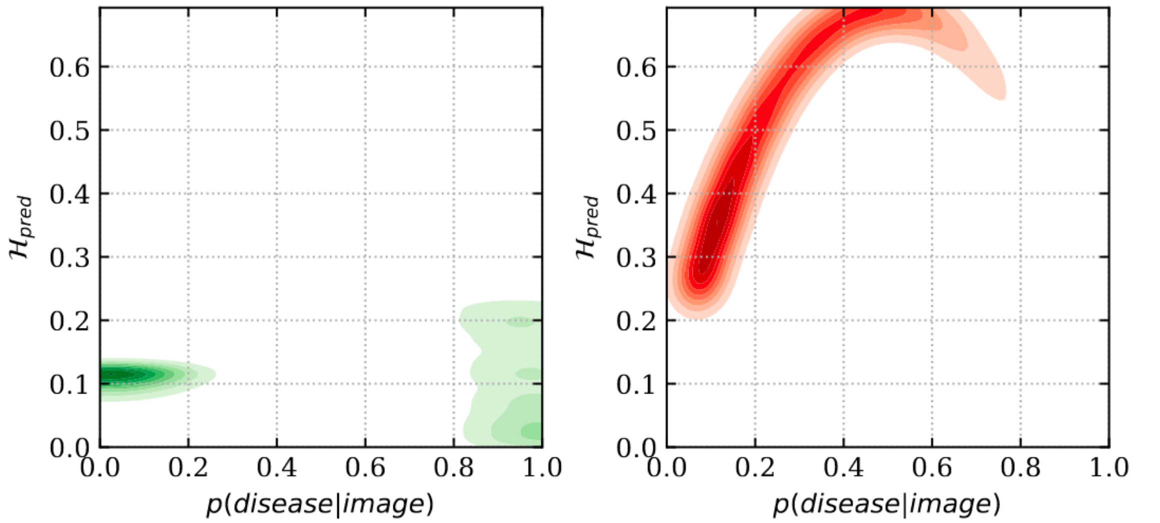
VARIATIONAL INFERENCE: DROPOUT

- Applications of dropout
  - Segmentation for autonomous driving(Kendall & Gal '17)



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty

From Kendall & Gal



(a) Correctly classified test images (b) Erroneously classified test images

Figure 4: Relation between predictive uncertainty (i.e. entropy), $\mathcal{H}_{\text{pred}}$, of MC Dropout model, and maximum-likelihood, i.e. sigmoid probabilities $p(\text{disease}|\text{image})$. The model has higher uncertainty for the miss-classified images, hence it can be used as an indicator to drive referral.

From Filios et al, 2019

  - Segementation for clinical applications

VARIATIONAL INFERENCE: DEEP LEARNING (6)

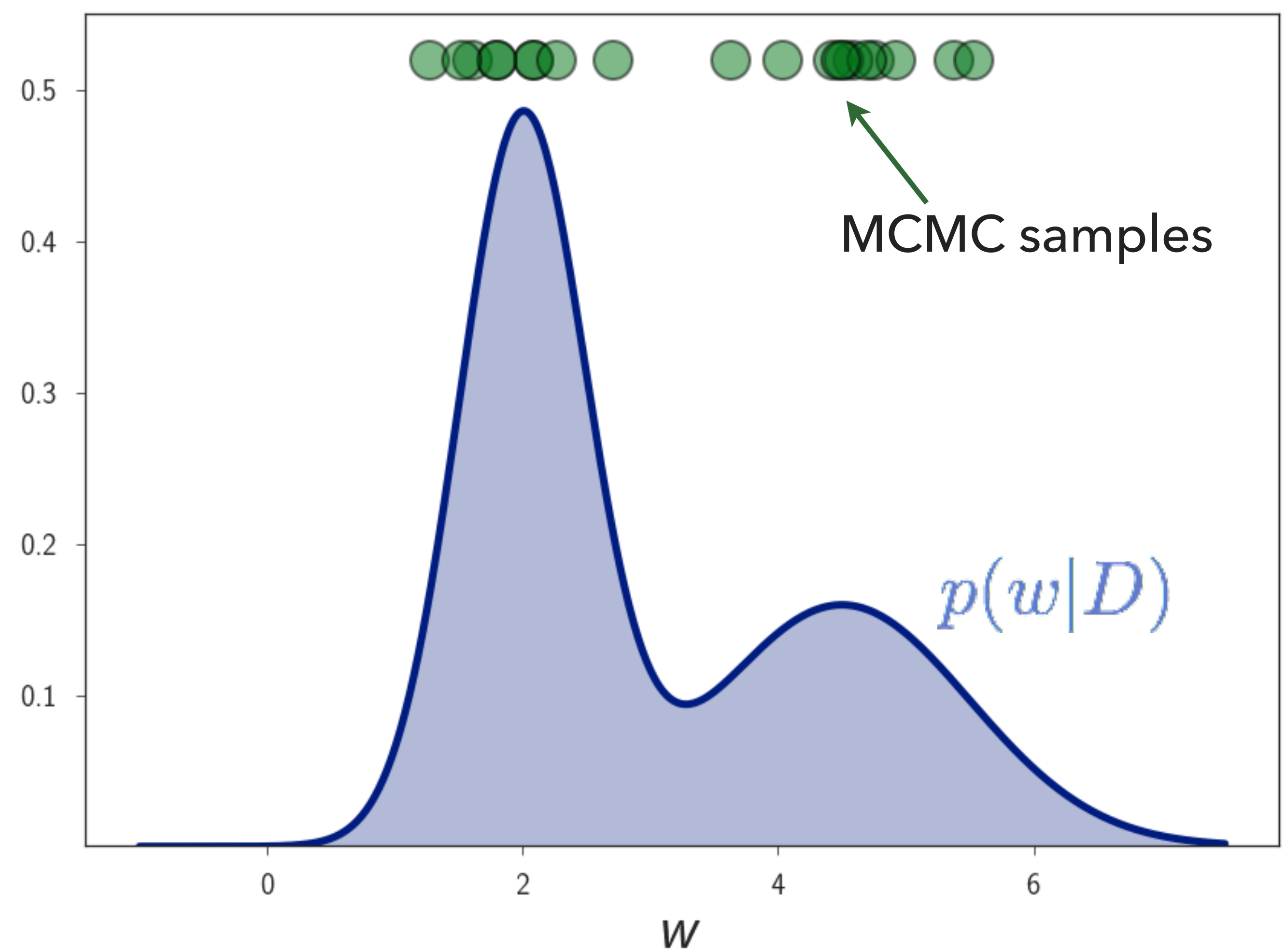
We can find the best approximating distribution within a given family with respect to KL-divergence

- $$KL(q||p) = \int_w q(w) \log \frac{q(w)}{p(w|D)} dw$$
- Variational interpretations of stochastic optimization….
    - Early stopping (Duvenaud, et al, '16)
    - Constant SGD (Mandt, Hoffman, Blei, '17) **[more later]**
    - Adam (Khan et al, '18, Osawa et al, '19)
    - Natural Gradient descent (Zhang et al, '18, Bae et al, 19)
    - MCMC (Hoffman & Ma, '19)

MARKOV CHAIN MONTE CARLO

We can produce samples from the exact posterior by defining specific Markov Chains

- Software packages:
  - Stan, PyMC4, Pyro
- Langevin dynamics (SGLD) (Neal '93, Welling & Teh, '11)
- Hamiltonian dynamics
  - Neal, '95, '96
  - Stochastic version - Chen et al, '14



CYCLIC SGMCMC (ZHANG ET AL, ICLR 2020)

https://github.com/ruqizhang/csgmcmc

- Run stochastic Hamiltonian Monte Carlo with a cyclic learning rate

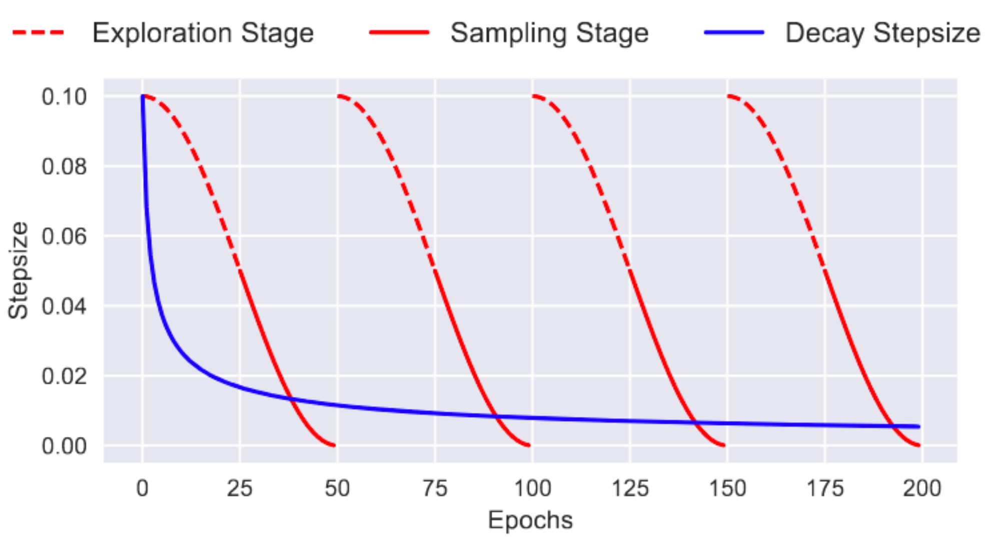
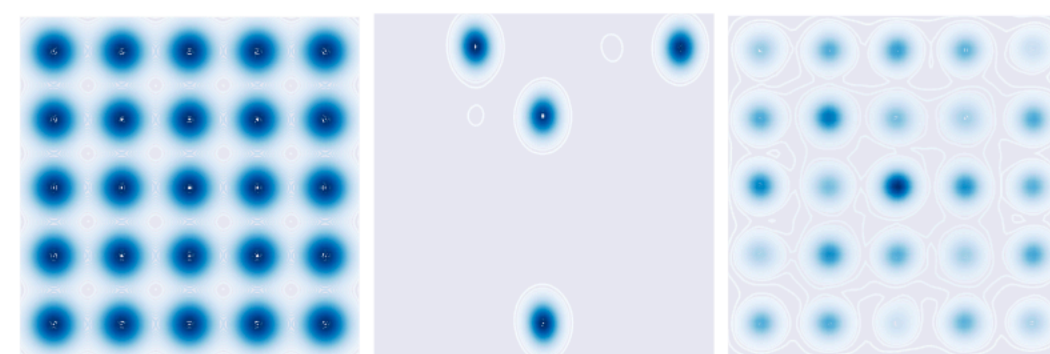


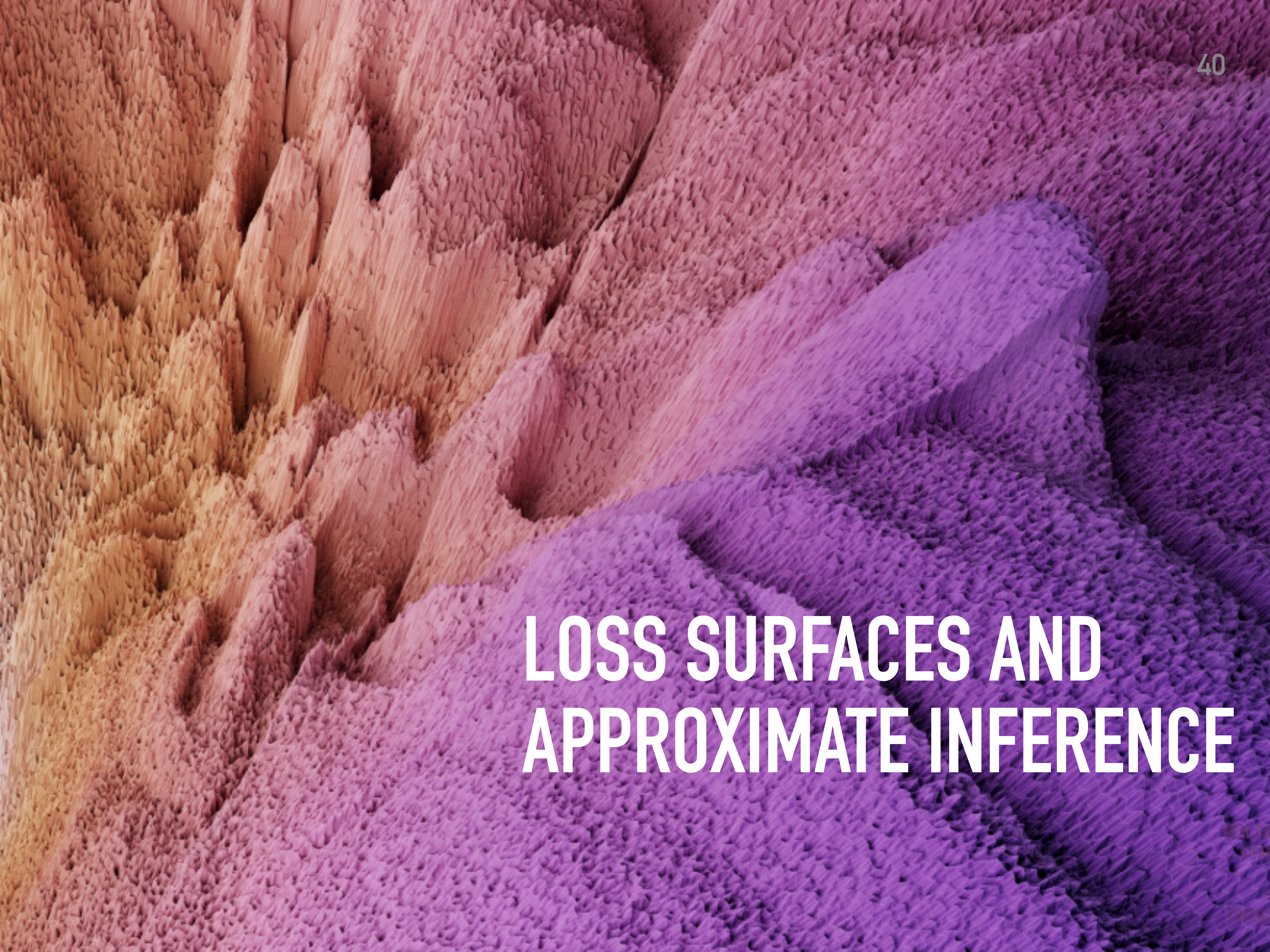
Figure 1. Illustration of the proposed cyclical stepsize schedule (red) and the traditional decreasing stepsize schedule (blue) for SG-MCMC algorithms.



(a) Target (b) SGLD (c) cSGLD

Figure 2. Sampling from a mixture of 25 Gaussians shown in (a) for the parallel setting. With a budget of 50k$\times$4=200k samples, traditional SGLD in (b) has only discovered 4 of the 25 modes, while our cSGLD in (c) has fully explored the distribution.

Converges faster to the posterior than standard SGHMC in terms of Wasserstein distance



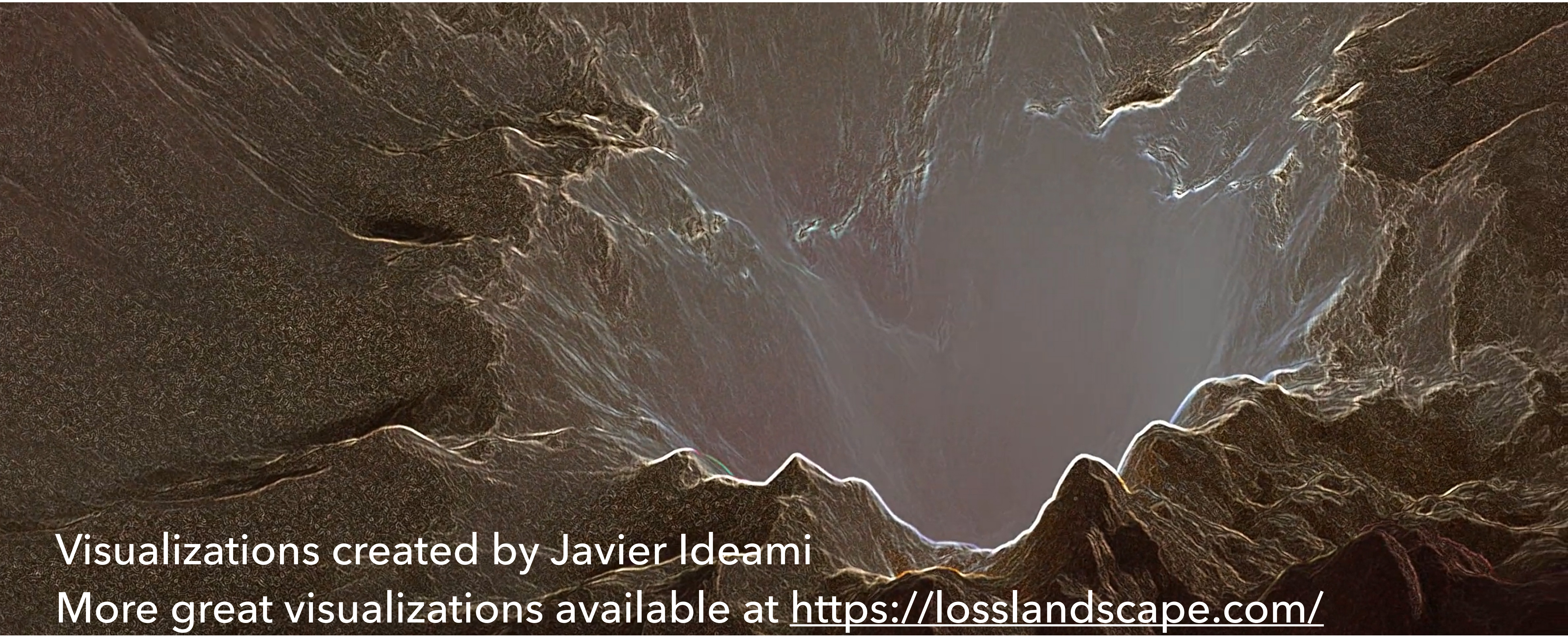
LOSS SURFACES AND APPROXIMATE INFERENCE

BASED OFF OF

- "A Simple Baseline for Bayesian Uncertainty in Deep Learning," **Maddox**, Garipov, *Izmailov*, Vetrov, *Wilson*, https://arxiv.org/abs/1902.02476, NeurIPS, 2019
  - Code: https://github.com/wjmaddox/swa_gaussian
- "Subspace Inference for Bayesian Deep Learning," *Izmailov*, **Maddox**, *Kirichenko*, Garipov, Vetrov, *Wilson*, https://arxiv.org/abs/1907.07504, UAI, 2019.
  - Code: https://github.com/wjmaddox/drbayes

LOSS SURFACES: WHY DO WE CARE?

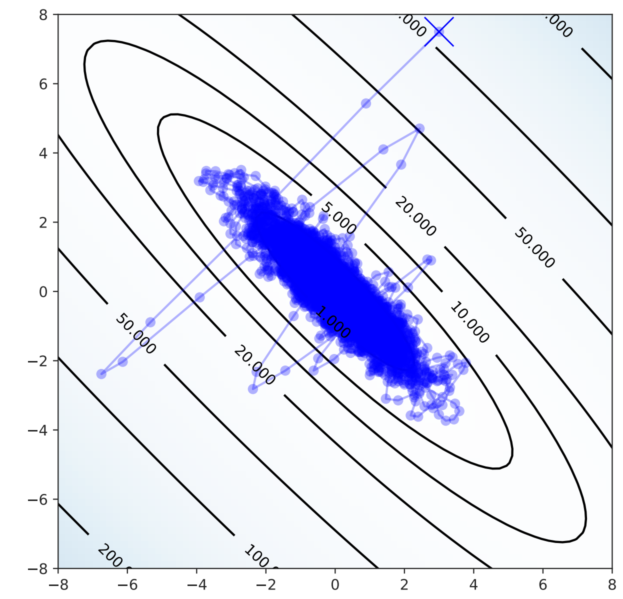
- Better approximate Bayesian Inference
  - $\text{loss} = -\log p(w|D)$, so understanding loss surfaces is crucial for approximate Bayesian inference



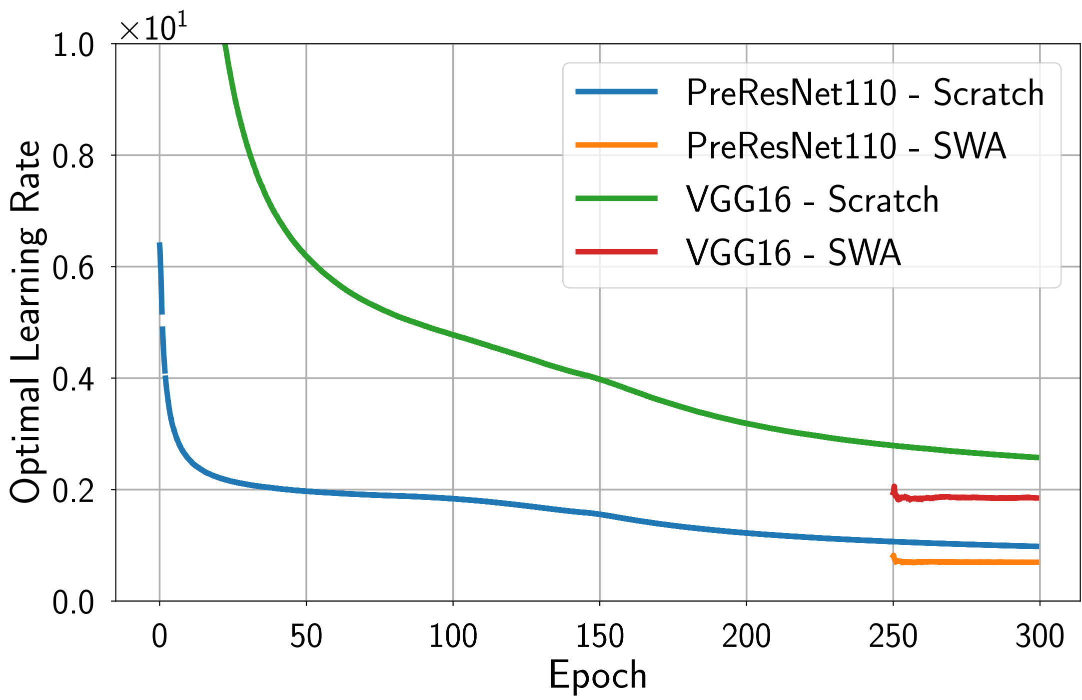
Visualizations created by Javier Ideami
More great visualizations available at https://losslandscape.com/

SGD AS APPROXIMATE BAYESIAN INFERENCE – MANDT, ET AL, JMLR, '17

- SGD with isotropic noise follows the shape of the posterior



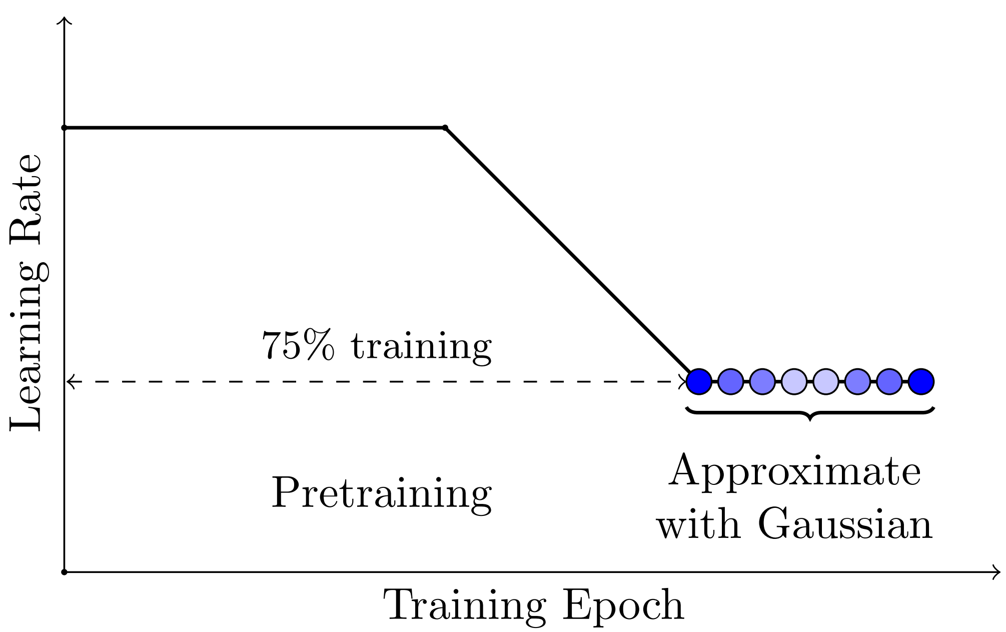
- Assumptions of analysis don't quite hold for DNNs
- But… we can use the same idea to approximate the posterior for DNNs

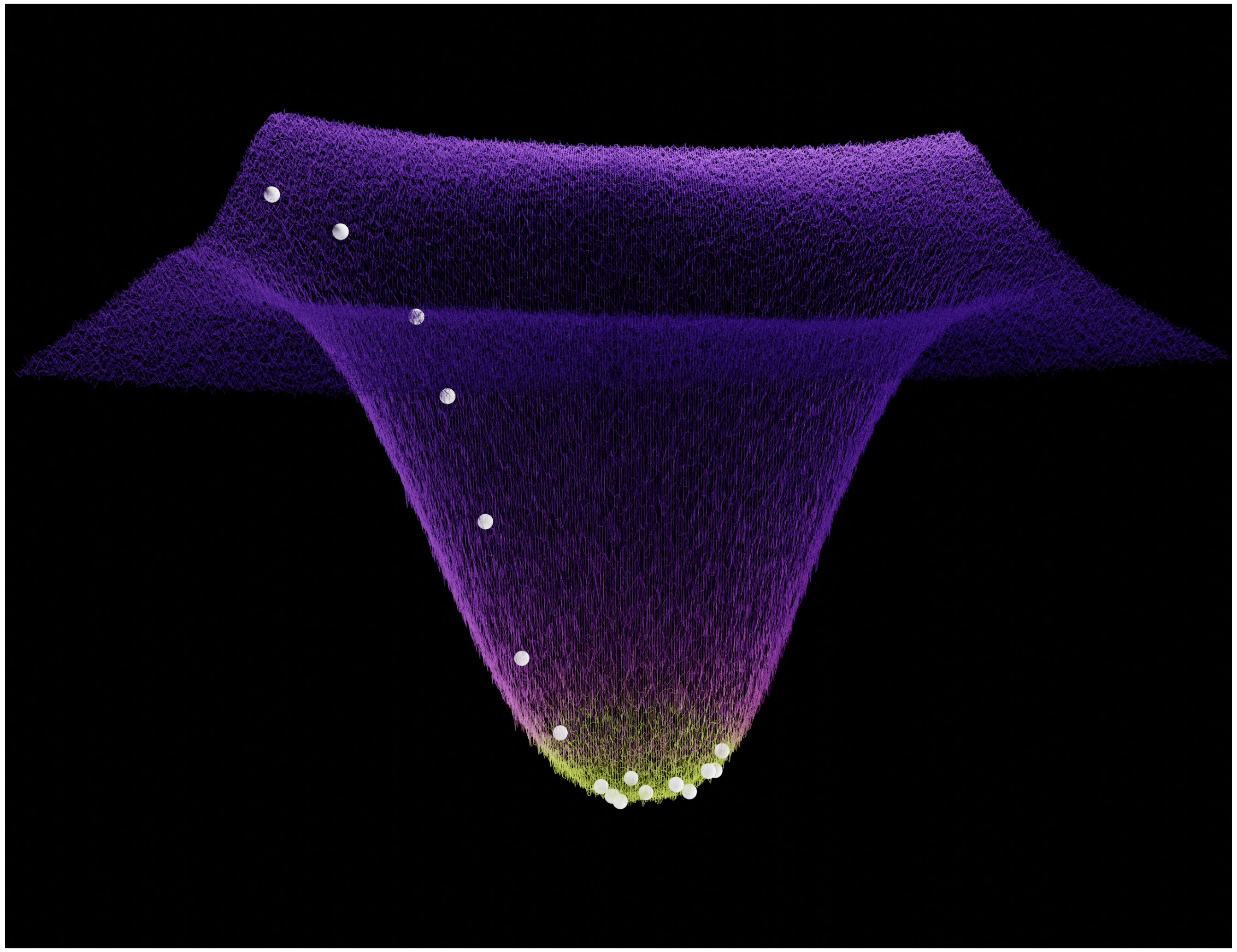


STOCHASTIC WEIGHT AVERAGING GAUSSIAN (SWAG)- MADDOX ET AL, NEURIPS, '19

https://github.com/wjmaddox/swa_gaussian

During training

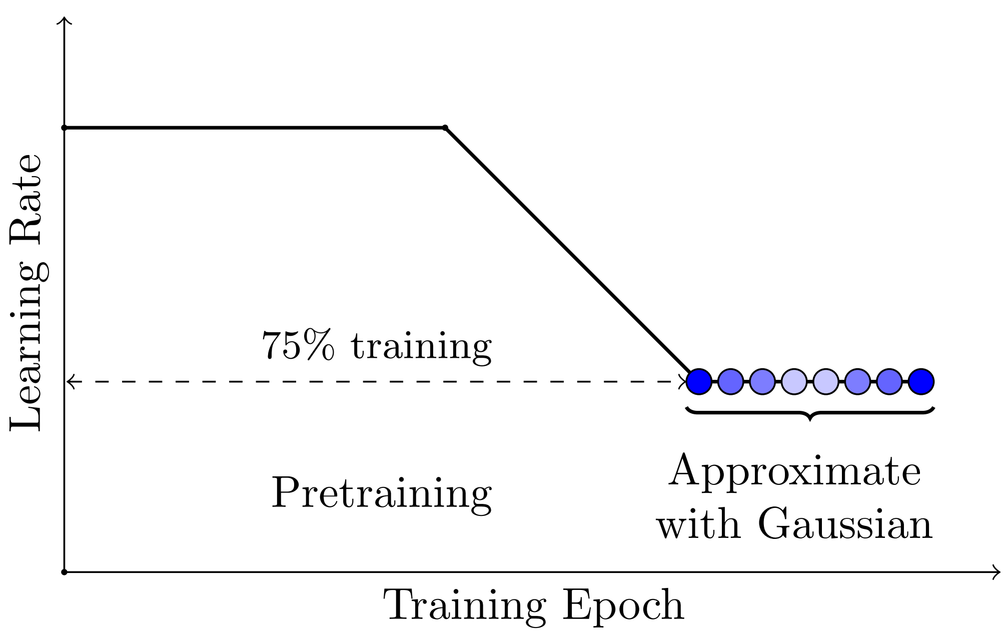


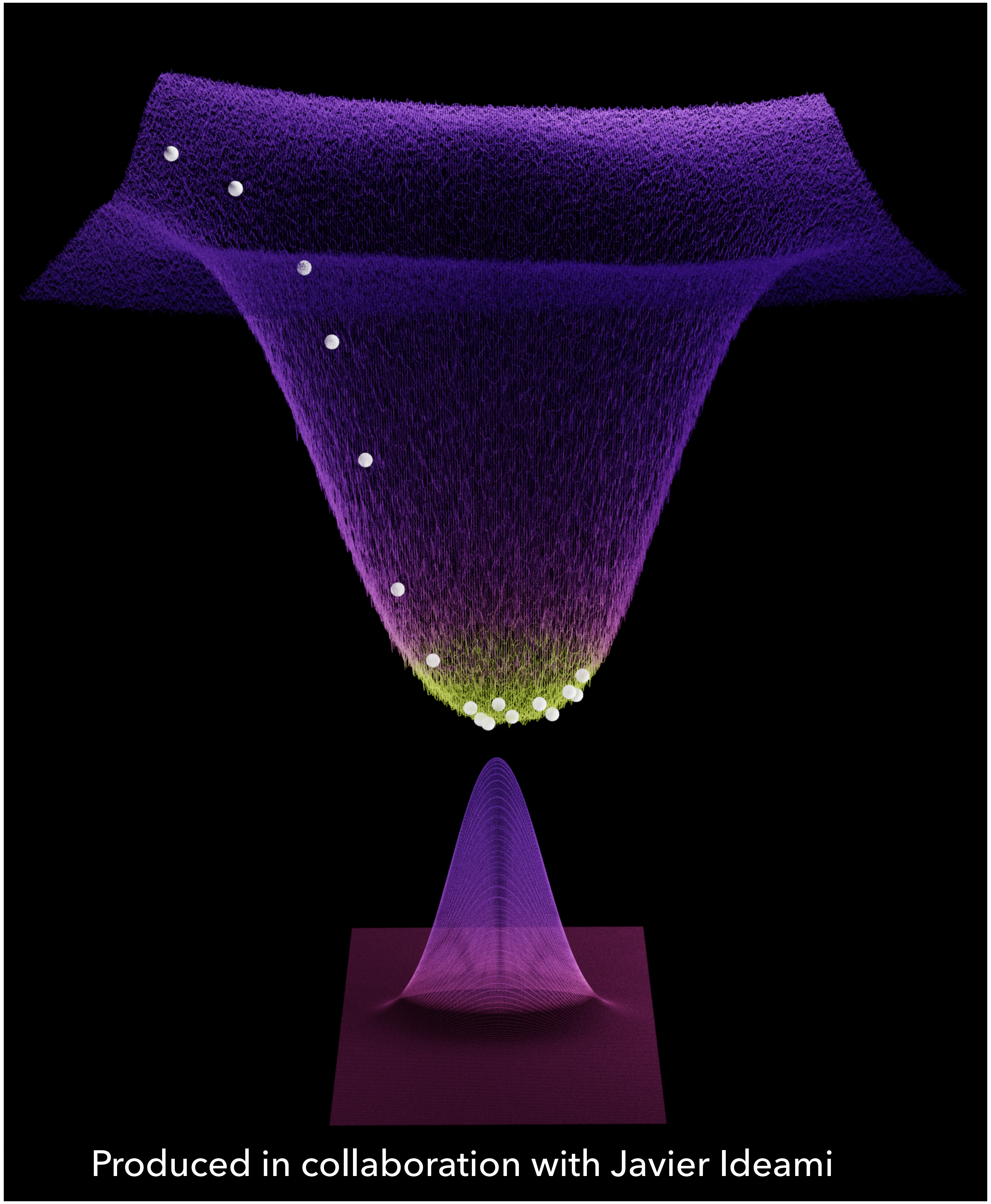


Produced in collaboration with Javier Ideami

SWAG

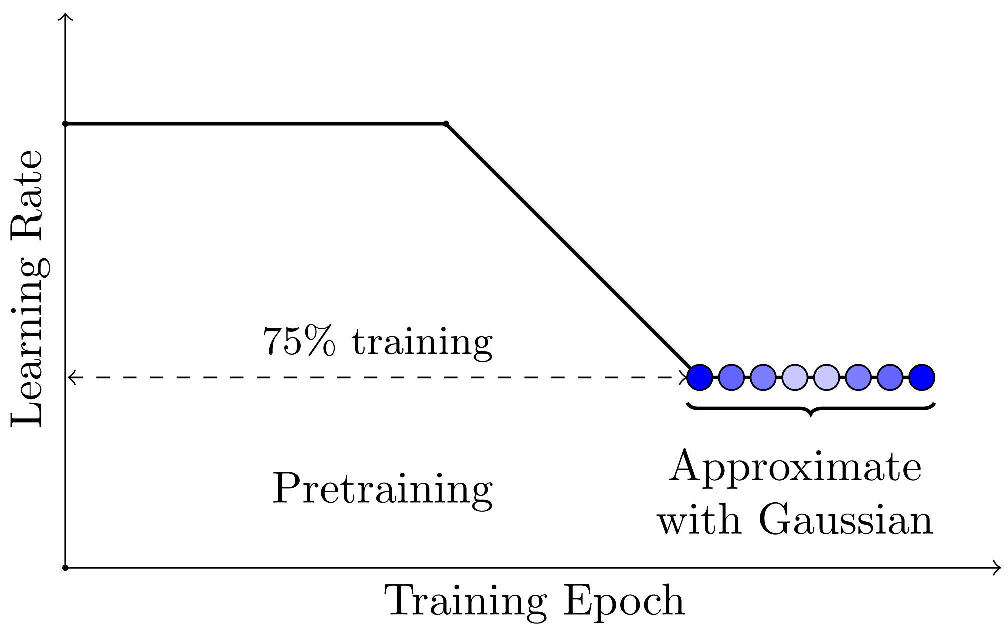
During training



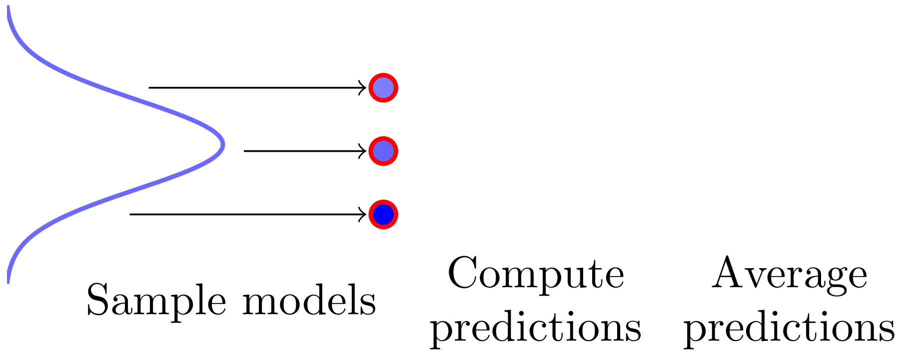


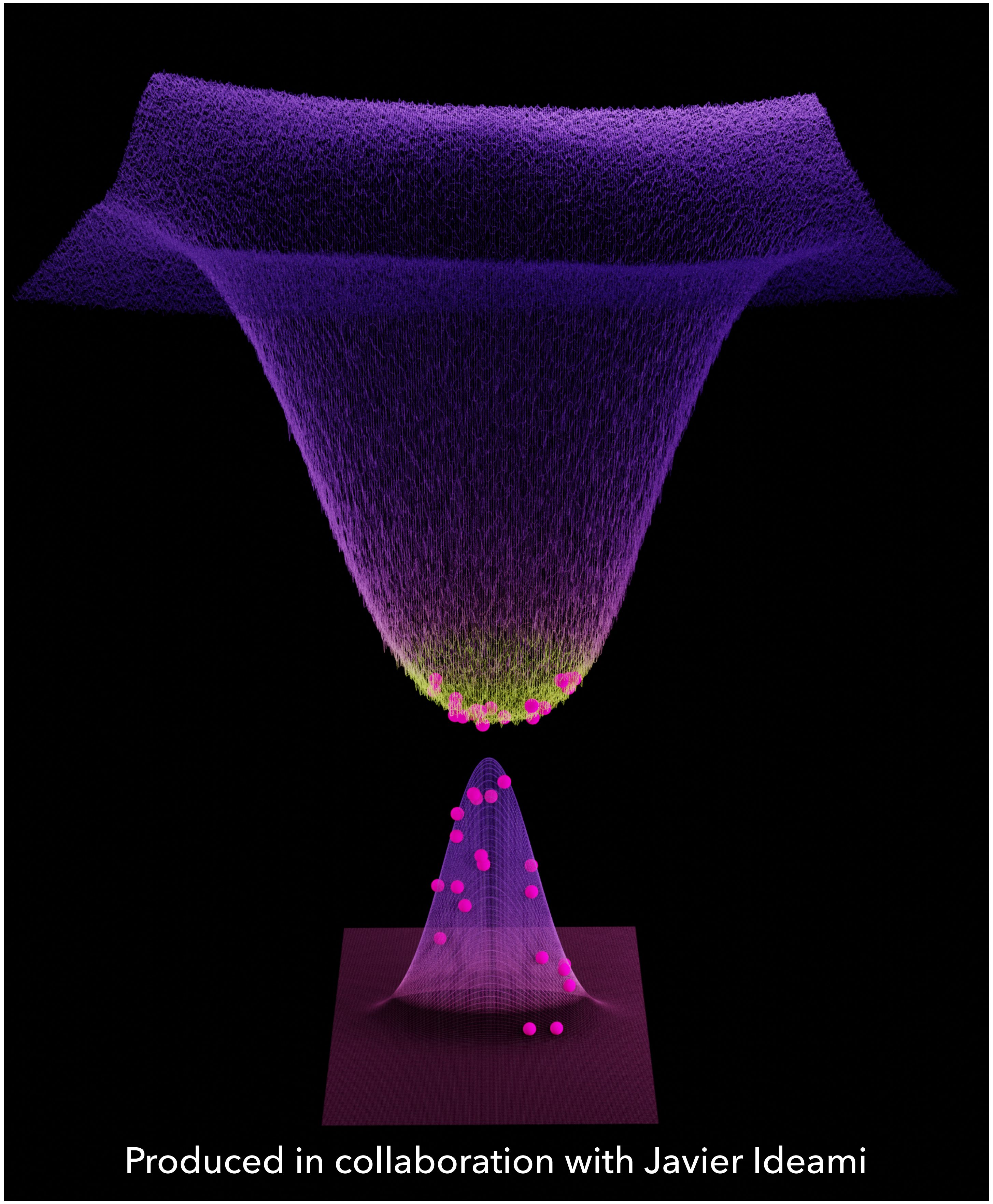
SWAG

During training

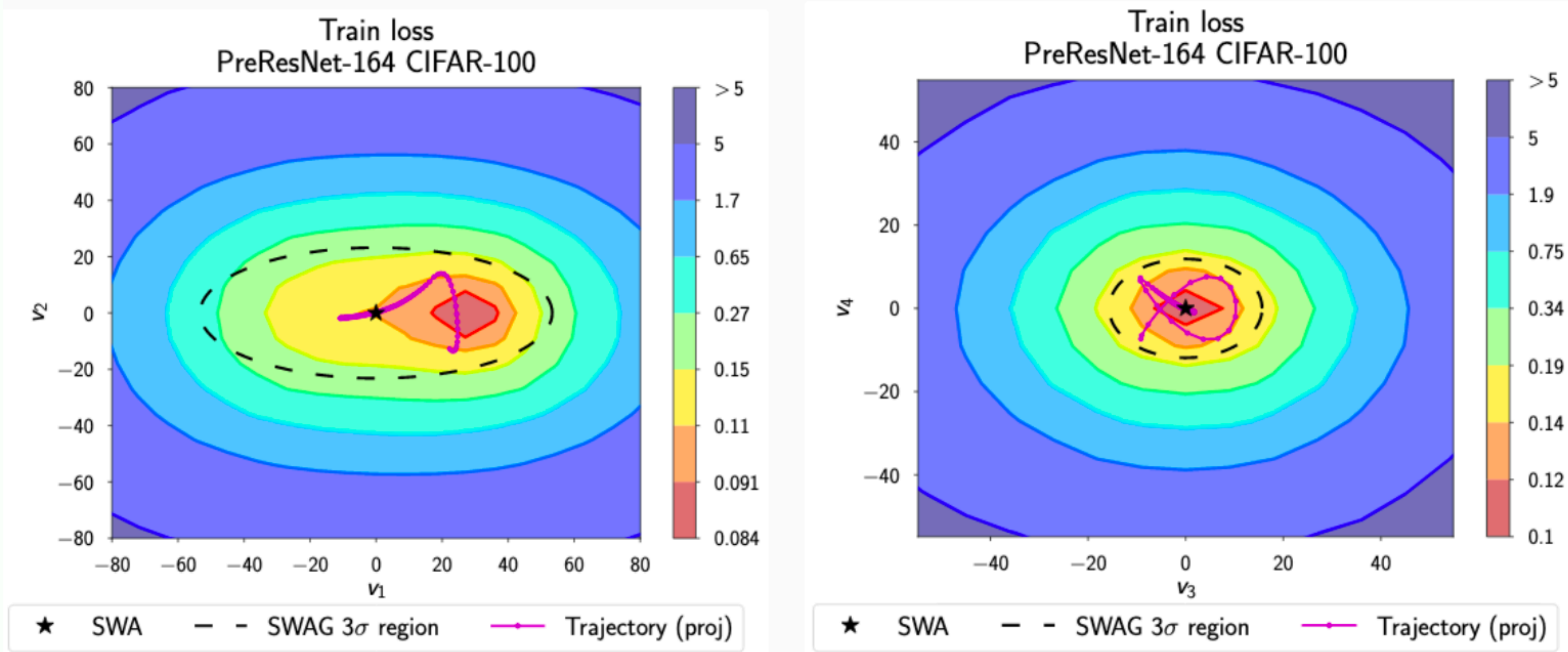


At test time



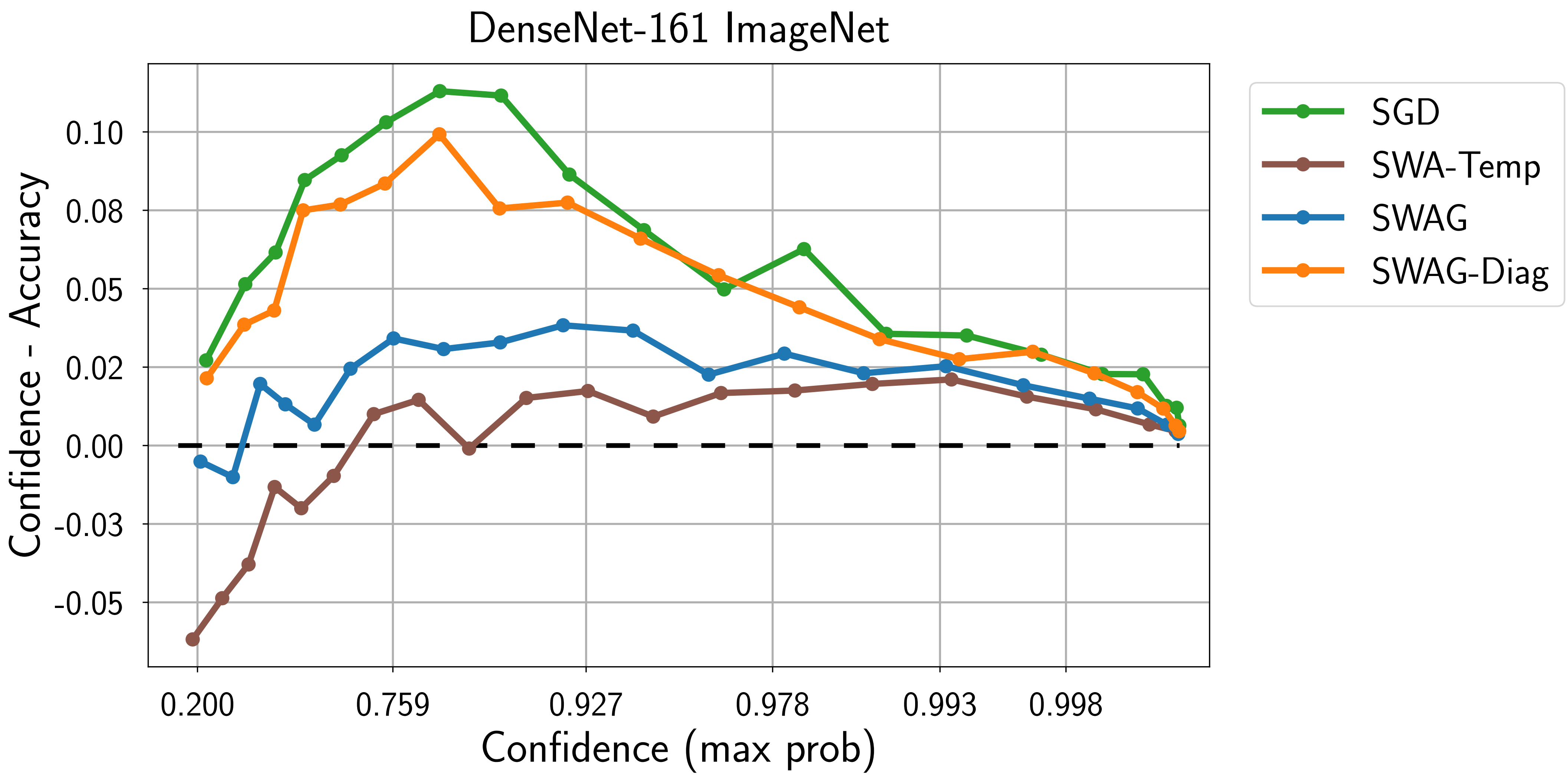


SWAG – EMPIRICAL MOTIVATION



$$\psi(t_1, t_2) = \mathcal{L}(\theta_{\text{SWA}} + t_1 \cdot \frac{v_i}{\|v_i\|} + t_2 \cdot \frac{v_j}{\|v_j\|})$$

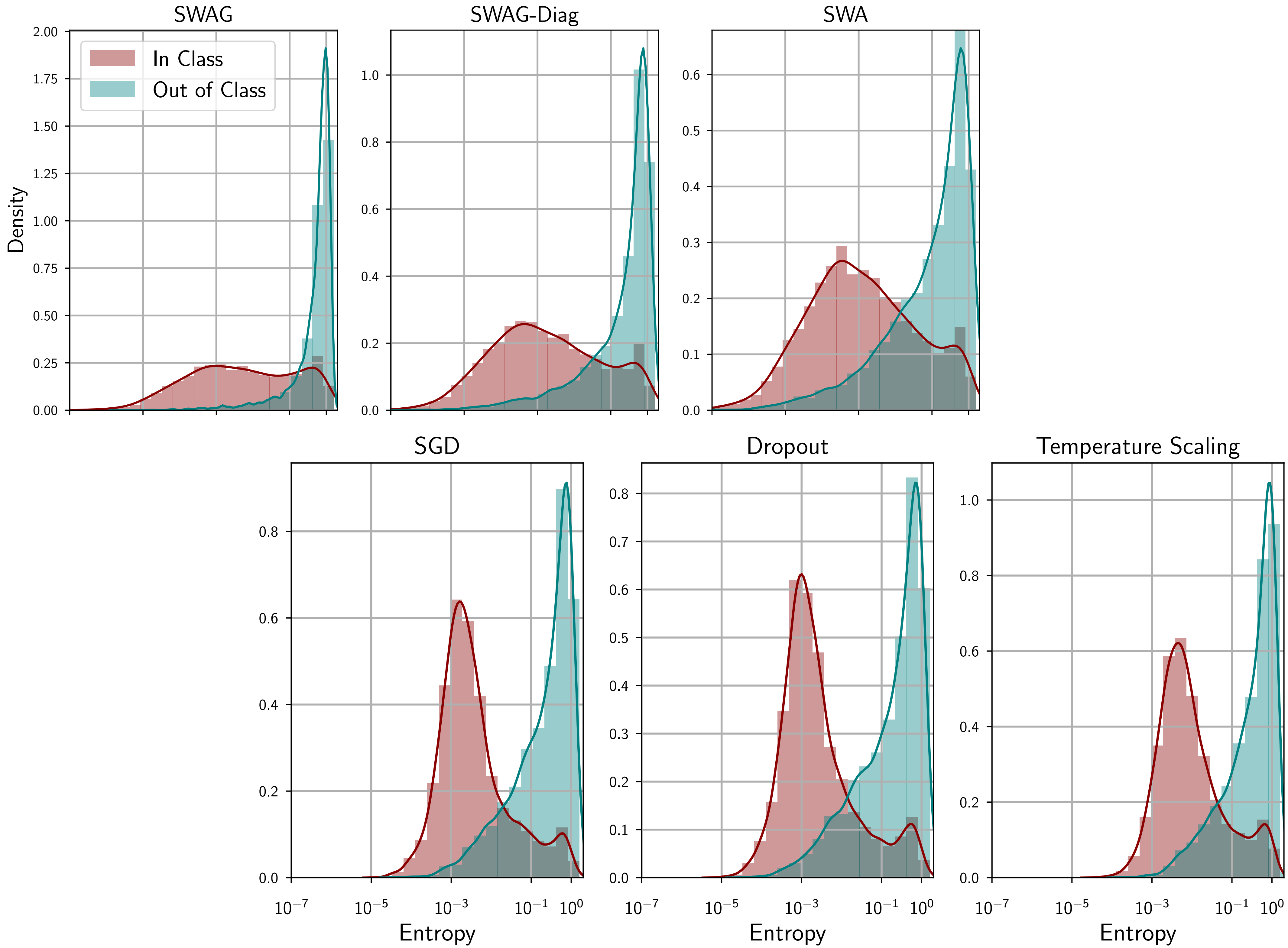
SWAG – CALIBRATION



SWAG – BAYESIAN MODEL AVERAGING

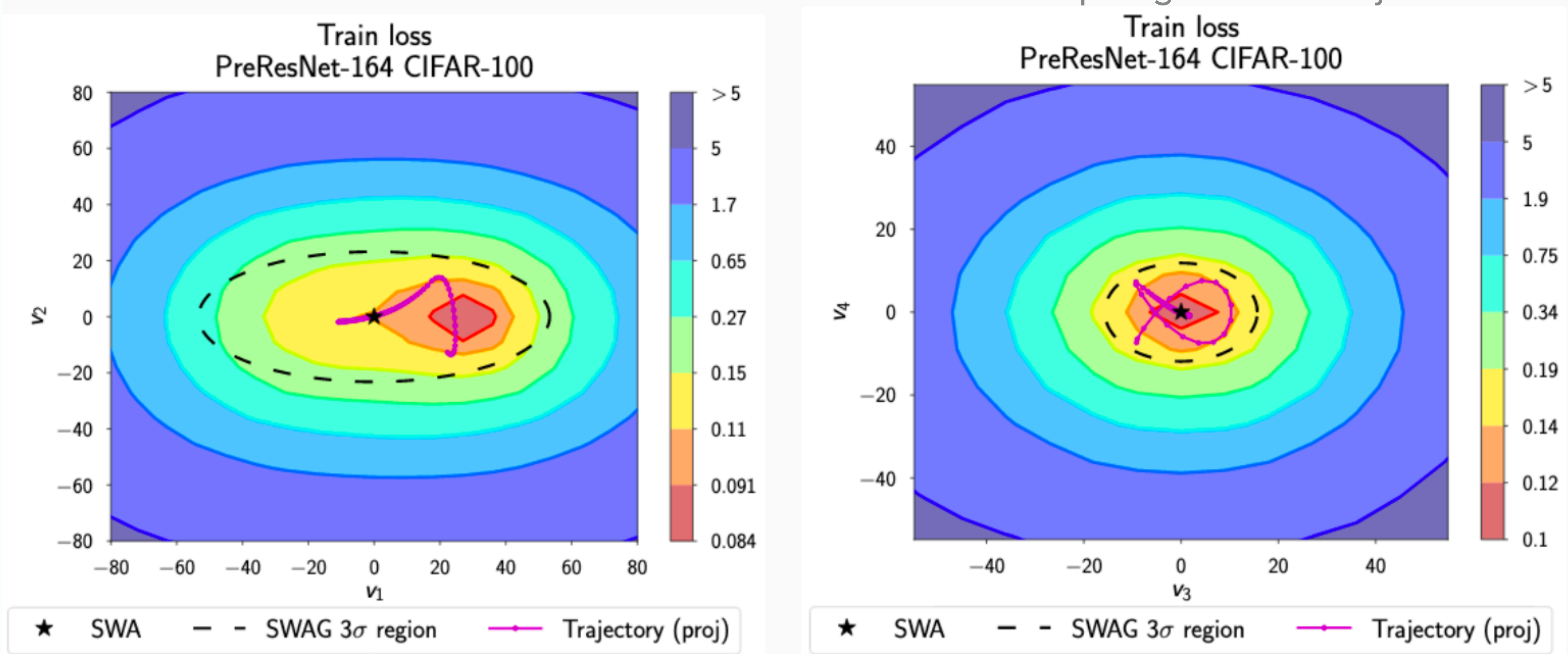
Dataset	Model	SGD	SWA	SWAG-Diag	SWAG	KFAC-Laplace	SWA-Dropout	SWA-Temp
CIFAR-10	VGG-16	93.17	93.61	**93.66**	93.60	92.65	93.23	93.61
CIFAR-10	PreResNet-164	95.49	96.09	96.03	96.03	95.49	**96.18**	96.09
CIFAR-10	WideResNet28x10	96.41	**96.46**	96.41	96.32	96.17	96.39	96.46
CIFAR-100	VGG-16	73.15	74.30	74.68	**74.77**	72.38	72.50	74.30
CIFAR-100	PreResNet-164	78.50	**80.19**	80.18	79.90	78.51		80.19
CIFAR-100	WideResNet28x10	80.76	82.40	**82.40**	82.23	80.94	82.30	82.40
ImageNet	DenseNet-161	77.79	**78.60**	78.59	78.59			78.60
ImageNet	ResNet-152	78.39	78.92	78.96	**79.08**			78.92
CIFAR10 $\rightarrow$ STL10	VGG-16	**72.42**	71.92	72.09	72.19		71.45	71.92
CIFAR10 $\rightarrow$ STL10	PreResNet-164	75.56	**76.02**	75.95	75.88			76.02
CIFAR10 $\rightarrow$ STL10	WideResNet28x10	76.75	**77.50**	77.26	77.09		76.91	77.50

SWAG – OUT OF SAMPLE DETECTION



SUBSPACE INFERENCE FOR BAYESIAN DEEP LEARNING – IZMAILOV, ET AL, UAI, '19

https://github.com/wjmaddox/drbayes

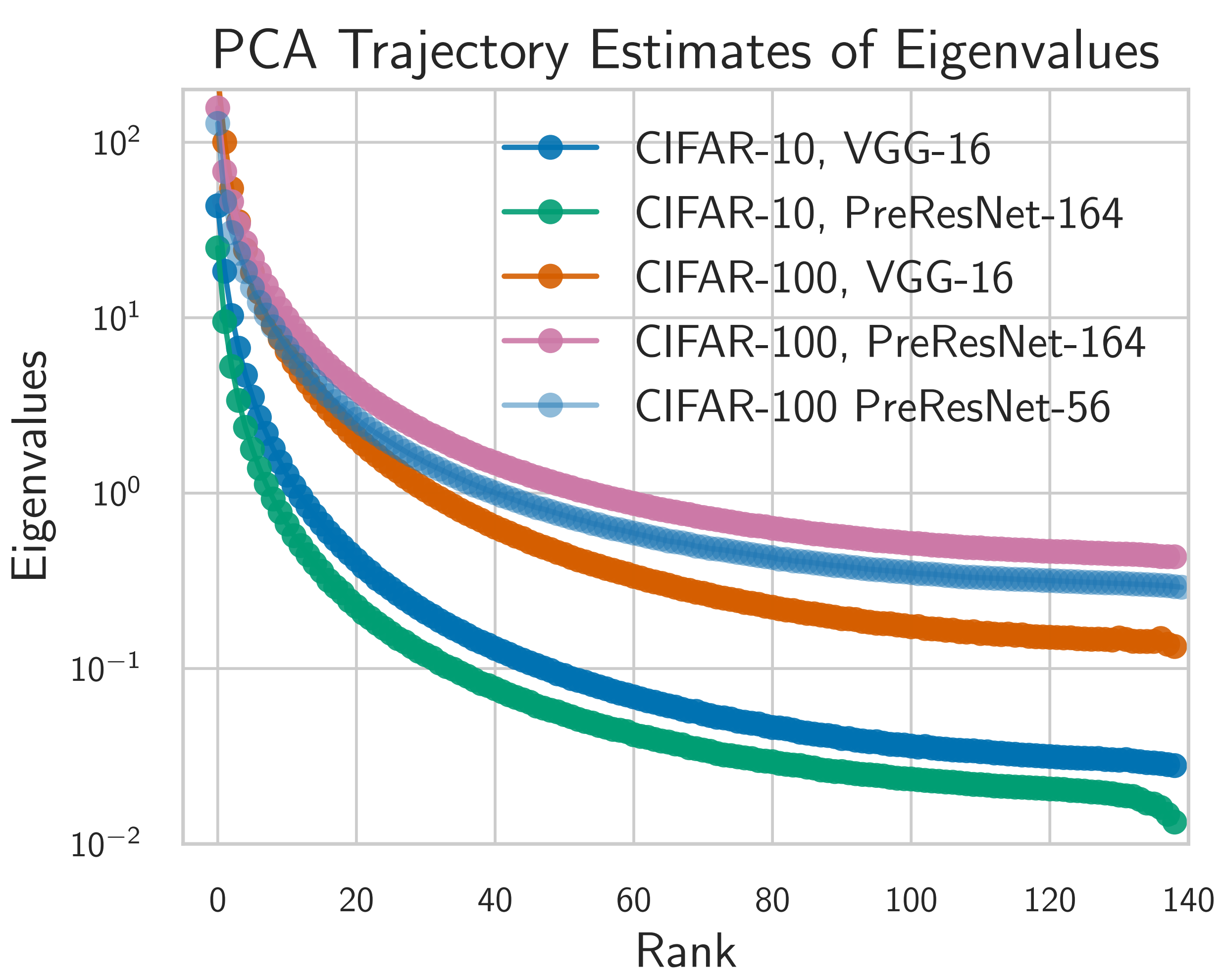


$$\psi(t_1, t_2) = \mathcal{L}(\theta_{\mathsf{SWA}} + t_1 \cdot \frac{v_i}{\|v_i\|} + t_2 \cdot \frac{v_j}{\|v_j\|})$$

- Remember this plot?

SUBSPACE INFERENCE

- SGD trajectory happens in a very small subspace
  - Summarize the information from the trajectory in very low dimensions
  - Also seen in Gur-Ari, et al, '19



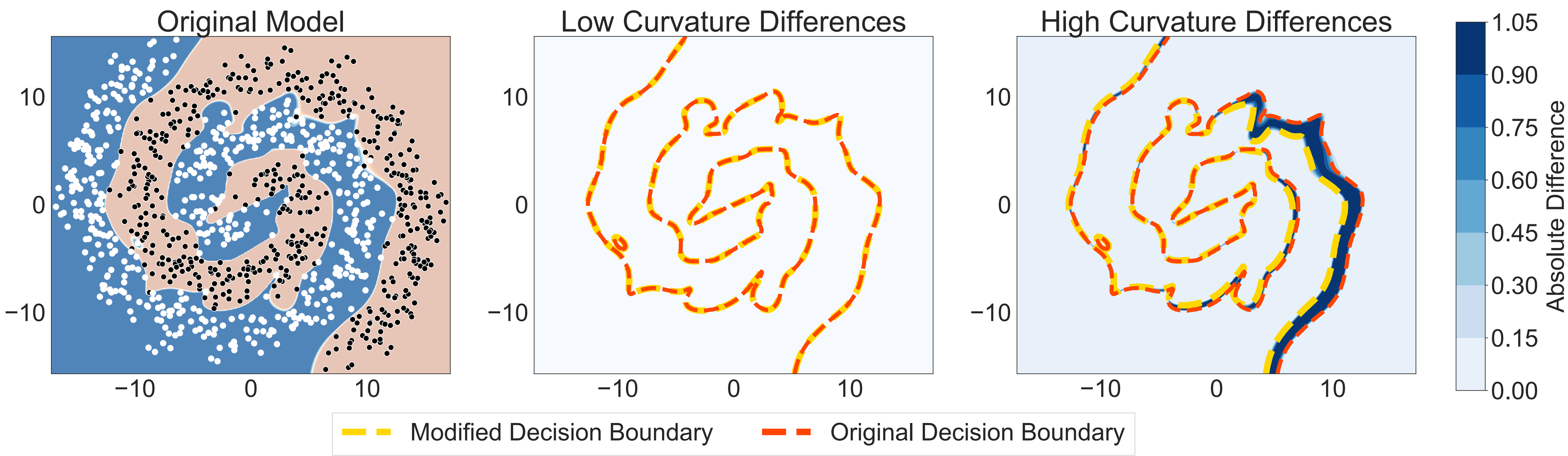
SUBSPACE INFERENCE

- A modular approach
  - Design subspace
  - Approximate posterior over parameters in that subspace
  - Sample from approximate posterior for bayesian model averaging

We can approximate posterior of 36 million dimensional WideResNet in 5D subspace and get state-of-the-art results!

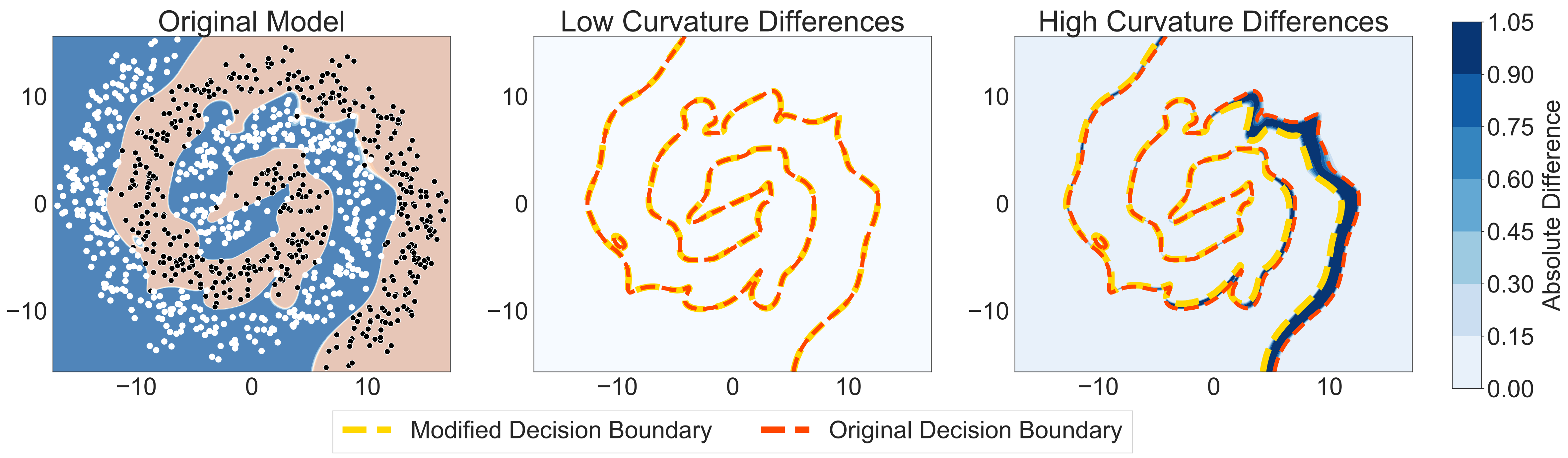
POSTERIOR CONTRACTION (REVISITED)

- If N >> p, can we even learn interesting distributions in p dimensions?



POSTERIOR CONTRACTION (REVISITED)

- If N >> p, can we even learn interesting distributions in p dimensions?

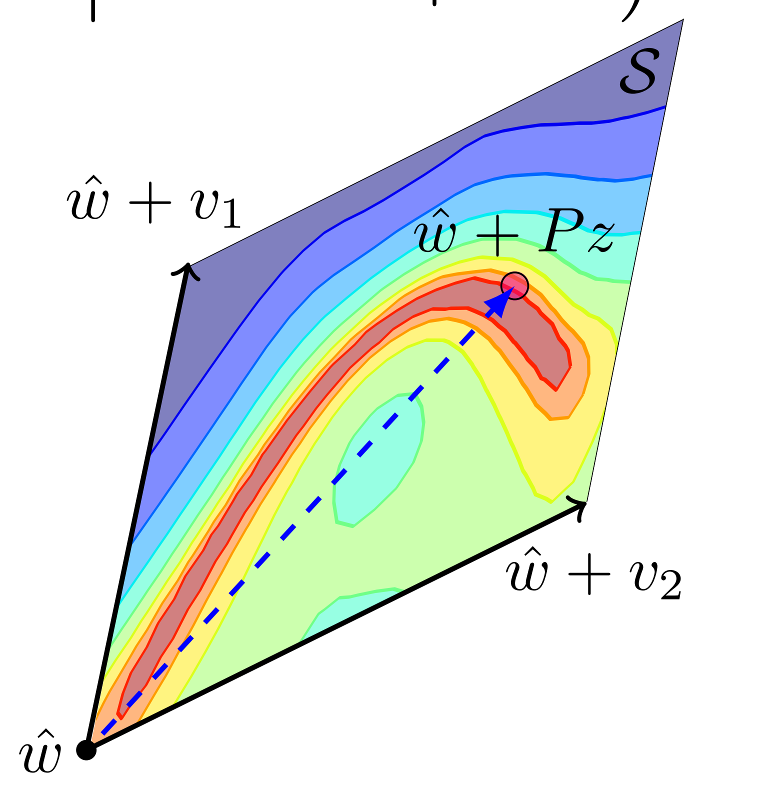


Probably not

CREATING THE SUBSPACE

- Choose shift $\hat{w}$ and basis vectors $\{d_1, \cdots, d_K\}$
- Define subspace $\mathcal{S} = \{w | w = \hat{w} + \underbrace{d_1 z_1 + \cdots d_K z_K}_{Pz}\}$
- Likelihood

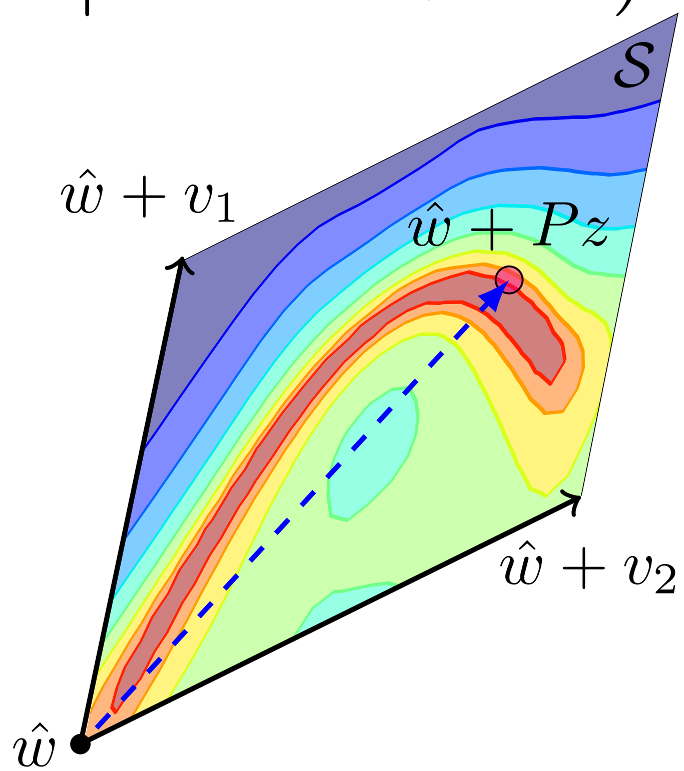
$$p(\mathcal{D}|z) = p_{\mathcal{M}}(\mathcal{D}|w = \hat{w} + Pz)^{1/T}$$



CREATING THE SUBSPACE

- Choose shift $\hat{w}$ and basis vectors $\{d_1, \cdots, d_K\}$
- Define subspace $\mathcal{S} = \{w | w = \hat{w} + \underbrace{d_1 z_1 + \cdots d_K z_K}_{Pz}\}$
- Likelihood

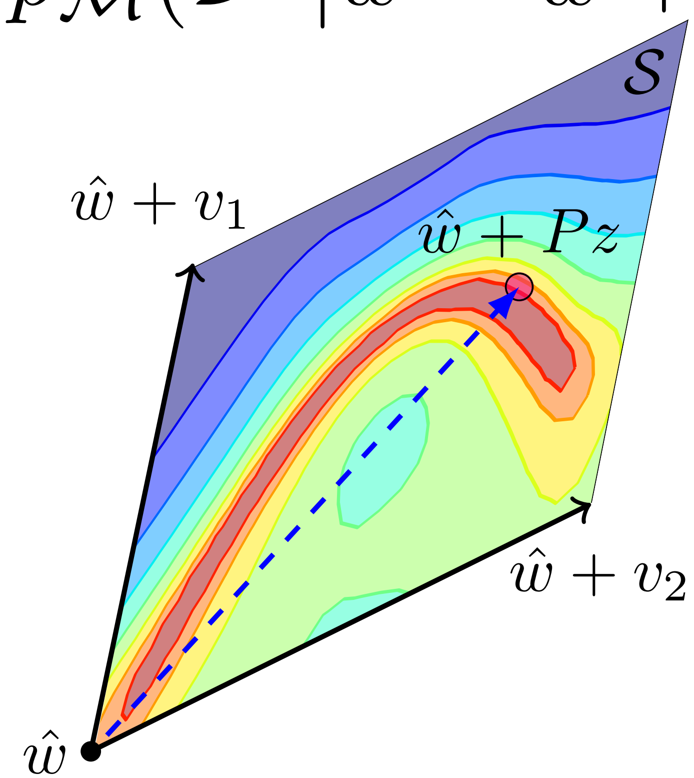
$$p(\mathcal{D}|z) = p_{\mathcal{M}}(\mathcal{D}|w = \hat{w} + Pz)^{1/T}$$



T>>1: to increase prior dependency & reduce effect of likelihood

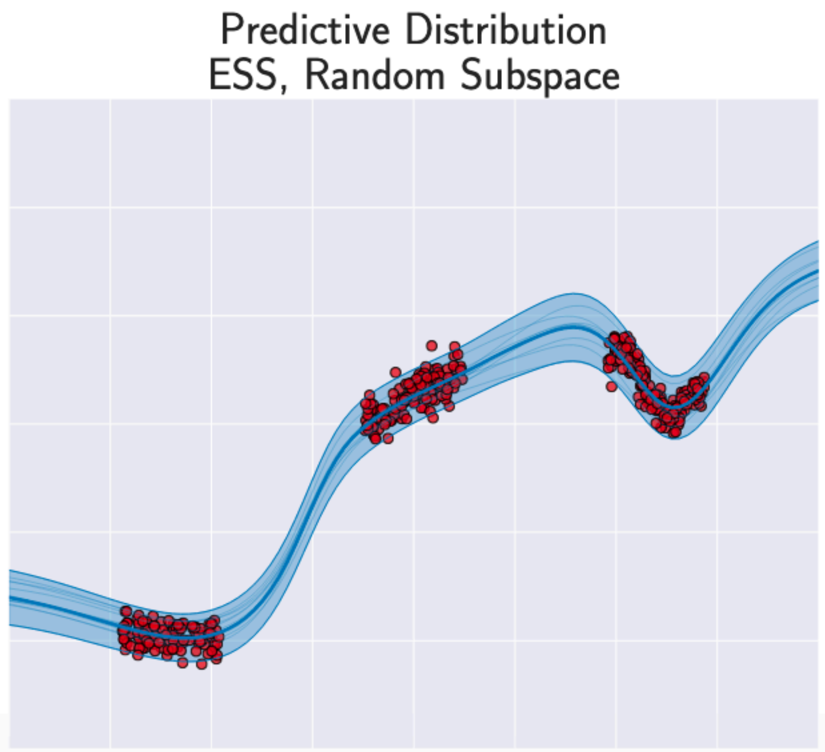
INFERENCE IN THE SUBSPACE

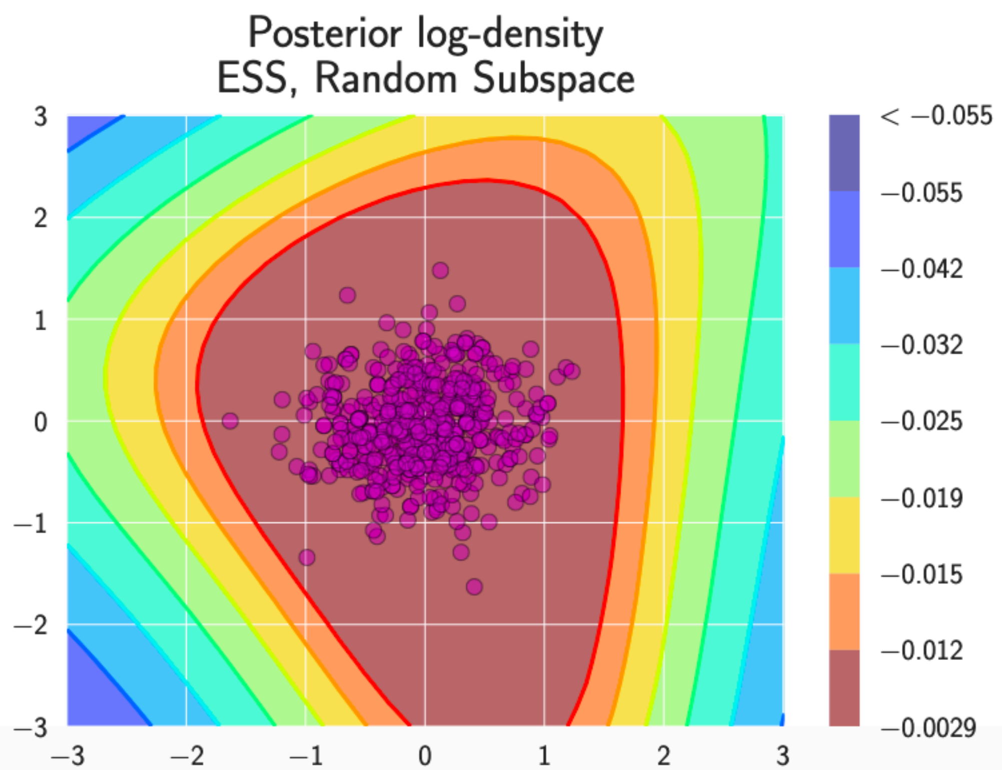
- Approximate inference over parameters
  - MCMC, VI, Normalizing Flows, …
- Bayesian model averaging at test time
- $$p(\mathcal{D}^*|\mathcal{D}) = \frac{1}{J}\sum_{j=1}^{J} p_{\mathcal{M}}(\mathcal{D}^*|w = \hat{w} + P\tilde{z}_j),\ \tilde{z}_j \sim q(\tilde{z}|\mathcal{D})$$



WHICH SUBSPACE? (RANDOM SUBSPACE)

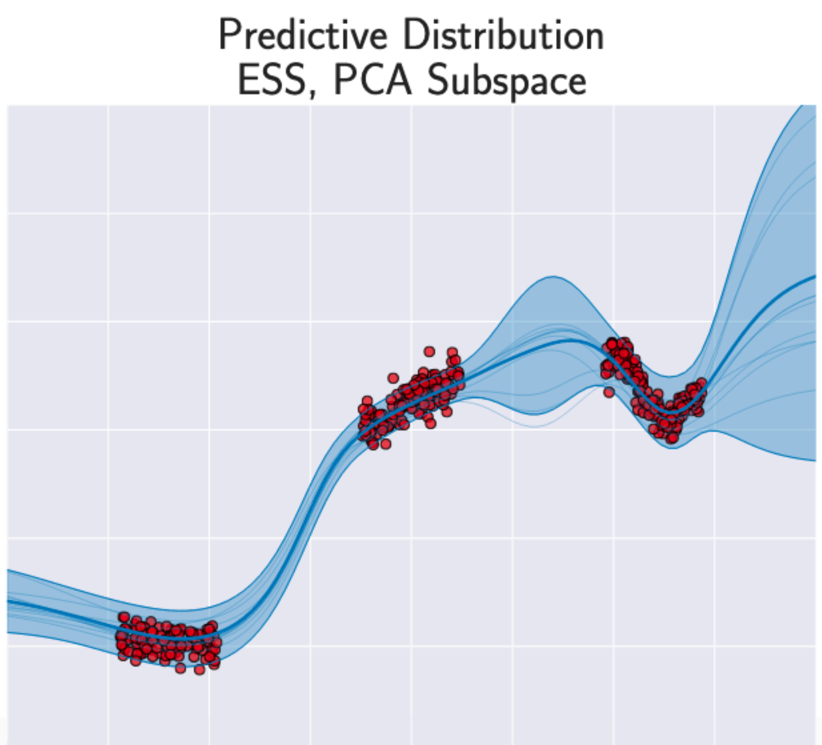
- Random directions: $d_1, \cdots, d_K \sim \mathcal{N}(0, I_K)$
- Use pre-trained solution as shift $\hat{w}$
- Subspace $S = \{w | w = \hat{w} + Pz\}$

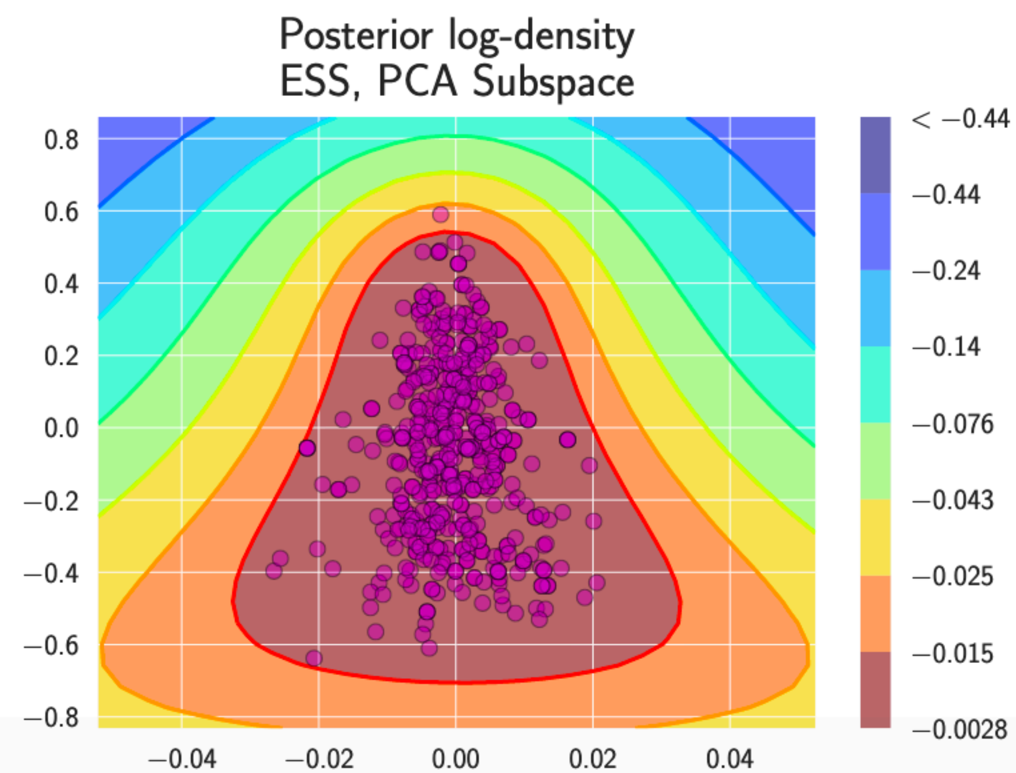




WHICH SUBSPACE? (PCA OF THE SGD TRAJECTORY)

- Run SGD with high constant learning rate from a pre-trained solution
- Collect snapshots of weights w_i
- Use SWA solution as shift $$\hat{w} = \frac{1}{T}\sum_{i=1}^{T} w_i$$
- $\{d_1, \cdots, d_K\}$ – first K PCA components of vectors $\hat{w} - w_i$



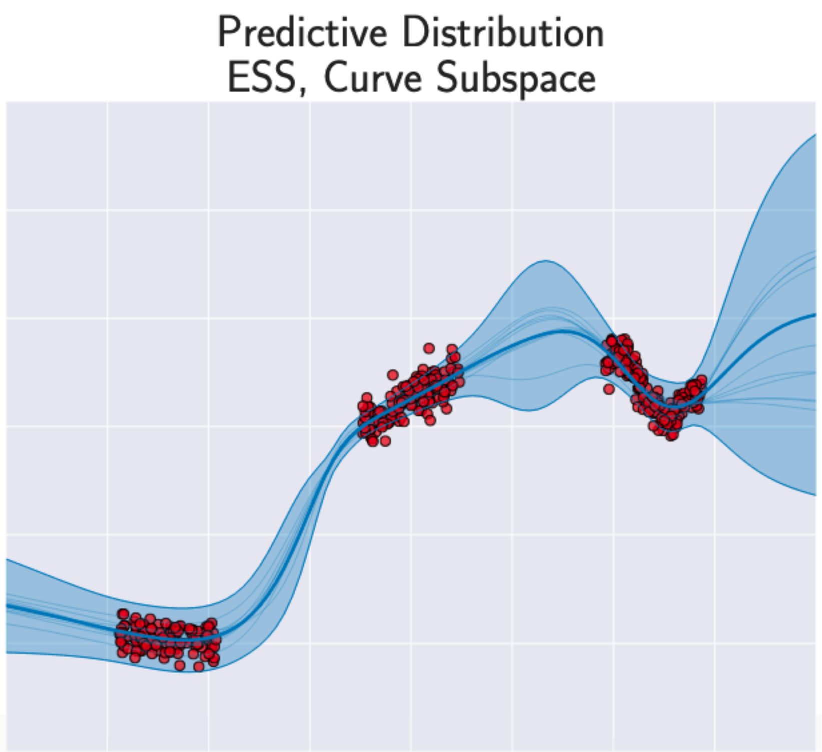


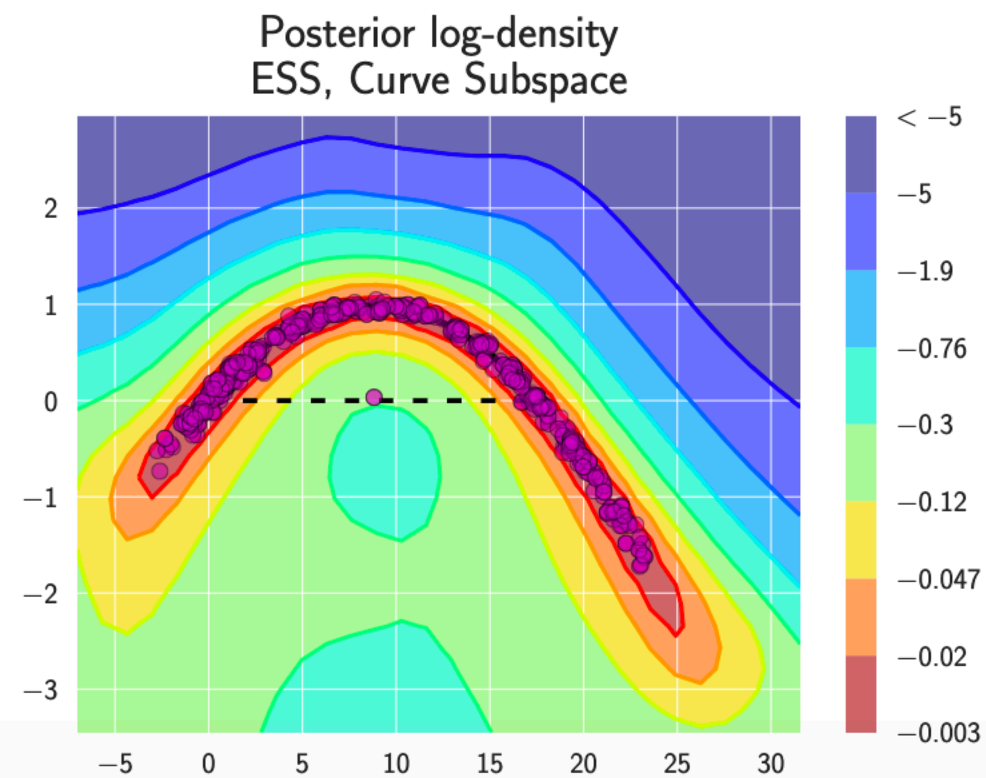
WHICH SUBSPACE? (CURVES — GARIPOV ET AL, '18)

- Garipov et al, '18 proposed a method to find 2D subspaces containing a path of low loss between weights of two independently trained neural networks

$$\arg\min_{\theta} \mathbb{E}_{t\sim U(0,1)}(\mathcal{L}(\phi_\theta(t)))$$

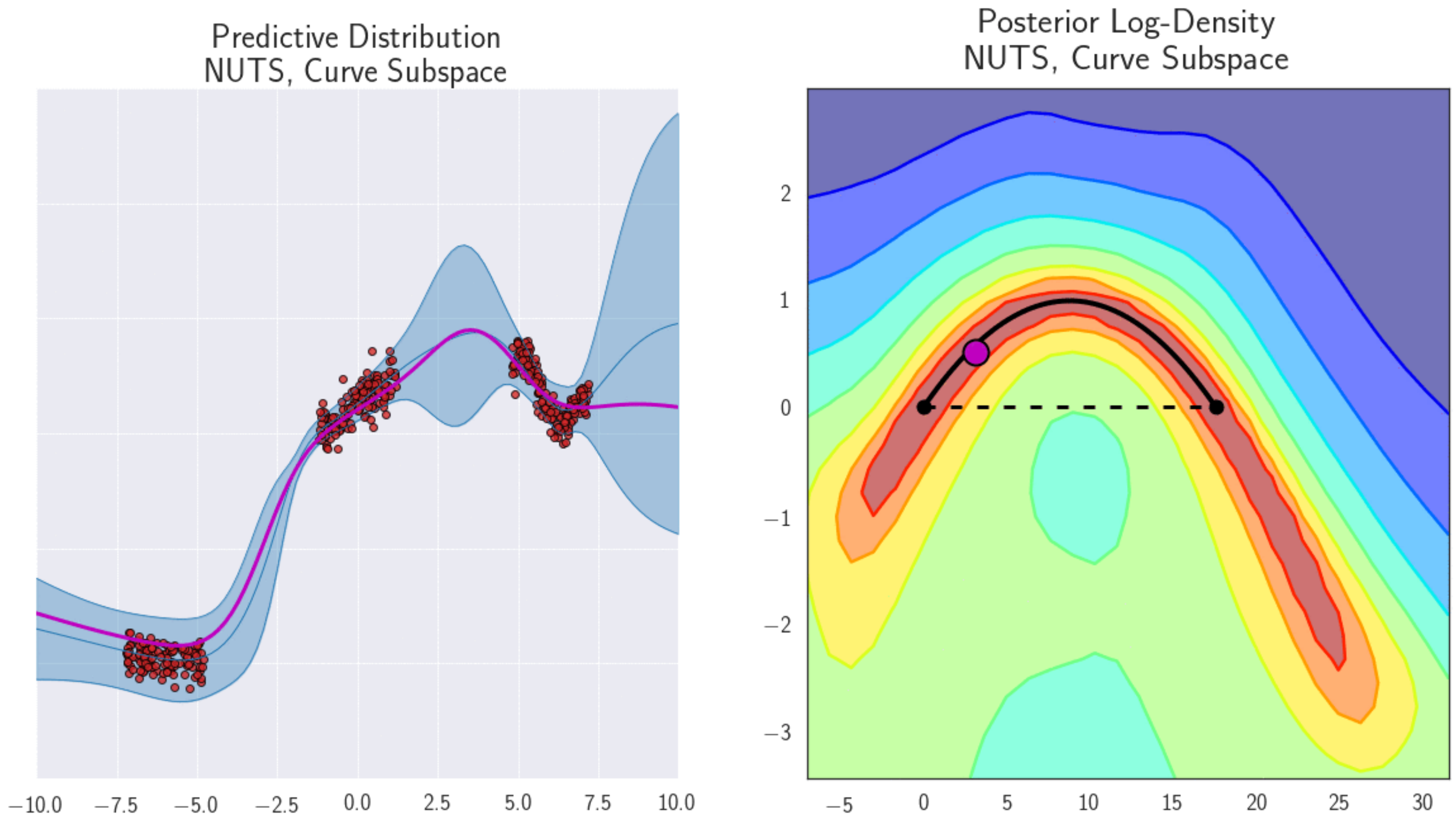
$$\phi_\theta(t) = (1-t)^2\hat{w}_1 + 2t(1-t)\theta + t^2\hat{w}_2^2$$





WHICH SUBSPACE? (CURVES — GARIPOV ET AL, '18)

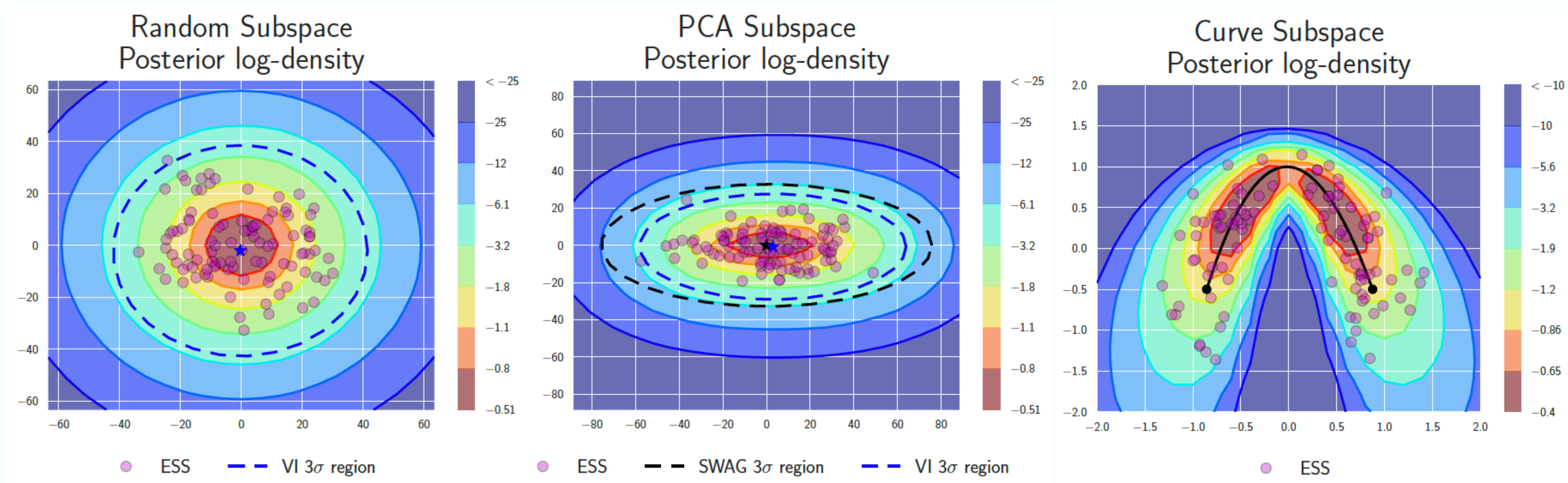
- Garipov et al, '18 proposed a method to find 2D subspaces containing a path of low loss between weights of two independently trained neural networks



RESULTS (PRERESNET164, CIFAR100)

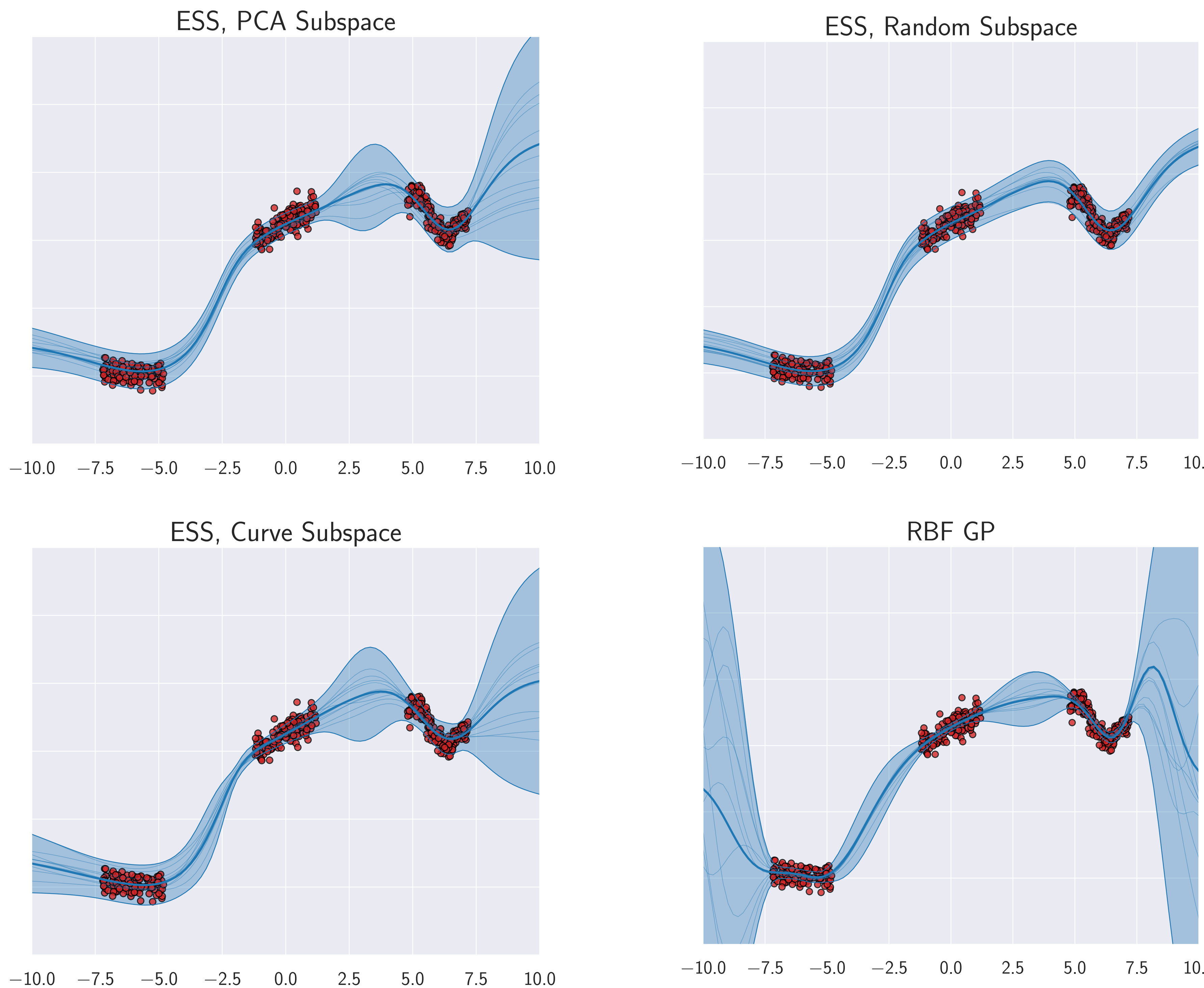
PCA of the SGD trajectory

Curves: chaining ind. models (Garipov, et al, '18)

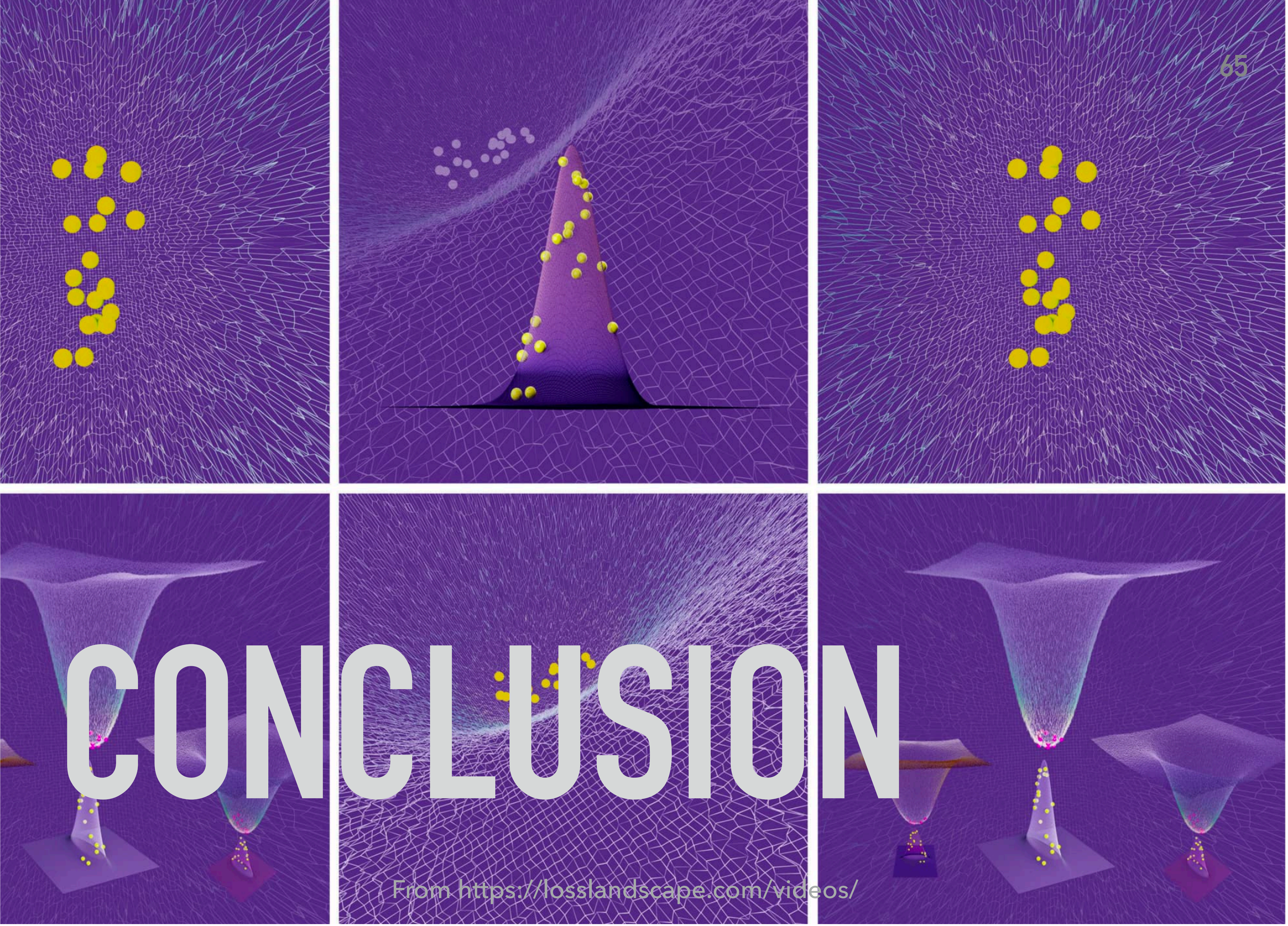


	SGD	Random	PCA	Curve
NLL	0.946 ± 0.001	0.686 ± 0.005	0.665 ± 0.004	0.646
Accuracy (%)	78.50 ± 0.32	80.17 ±0.03	80.54 ± 0.13	81.28

RESULTS – REGRESSION



Gold standard on regression tasks



CONCLUSION

From https://losslandscape.com/videos/

BAYESIAN DEEP LEARNING: CHALLENGES

Bayesian inference for deep neural networks is extremely challenging

- Posterior is intractable
- Millions of parameters
- Large datasets
- Unclear which priors to use

Is the likelihood correct? Probably

What do these parameters mean?
Care about functions instead

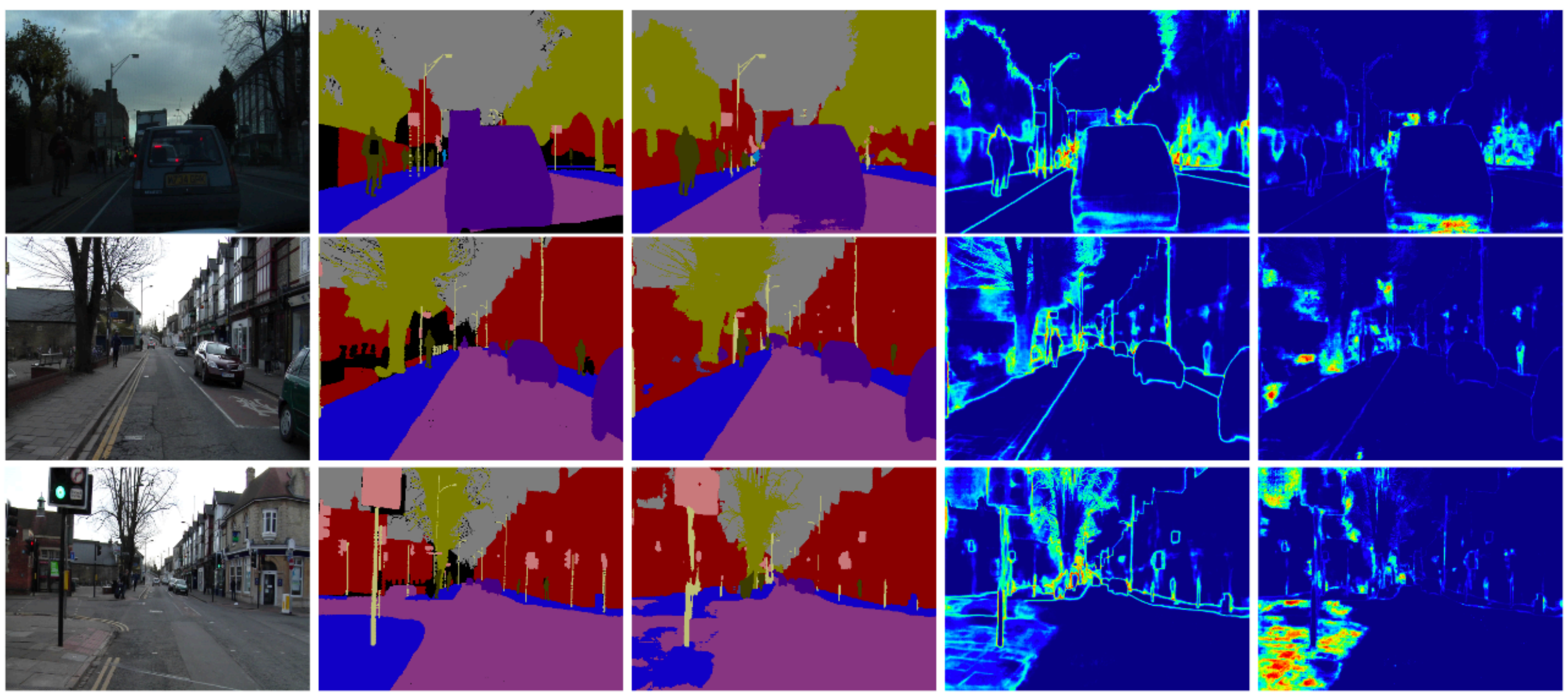
Can we run MCMC for 1 million steps on ImageNet??
We don't need to

Is the prior correct?
Probably

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} = \frac{p(D|w)p(w)}{\int_{w'} p(D|w')p(w')dw'}$$

BAYESIAN DEEP LEARNING: SUCCESSES

- But it doesn't mean we shouldn't try…



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Again from Kendall & Gal, "What Uncertainties do we need for bayesian deep learning for computer vision?"

BAYESIAN DEEP LEARNING: PRIOR CHOICE

- Typically use a iid Gaussian prior $\mathcal{N}(0, \alpha^2 I)$
- Choices may not be adversarial…
- But also not fantastic…

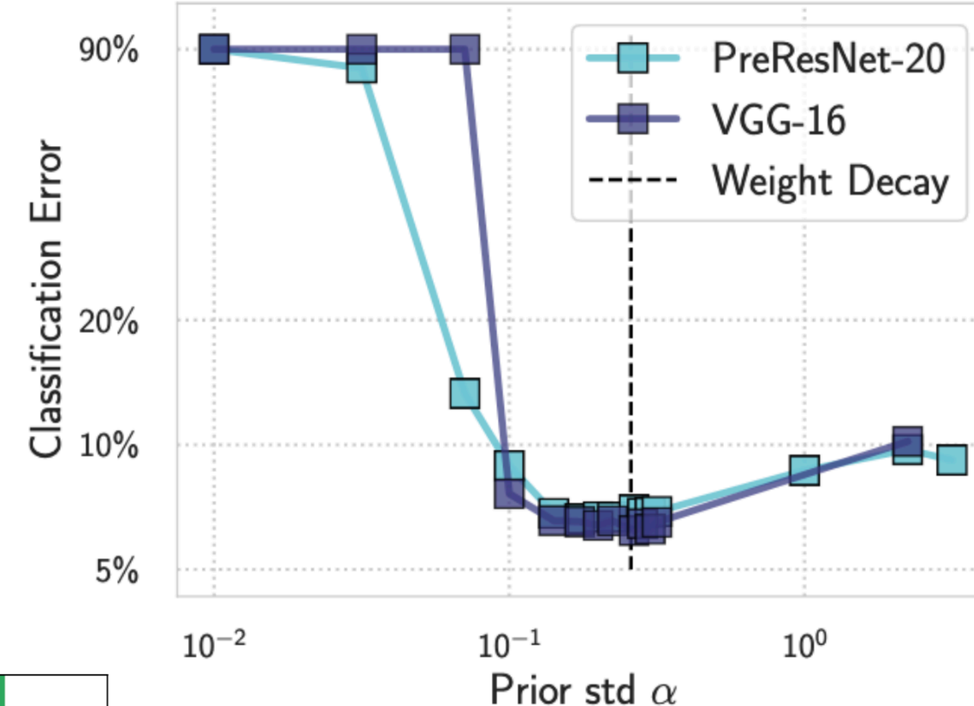


Figure 10g of Wilson & Izmailov '20
https://arxiv.org/pdf/2002.08791.pdf

$\alpha = 1$

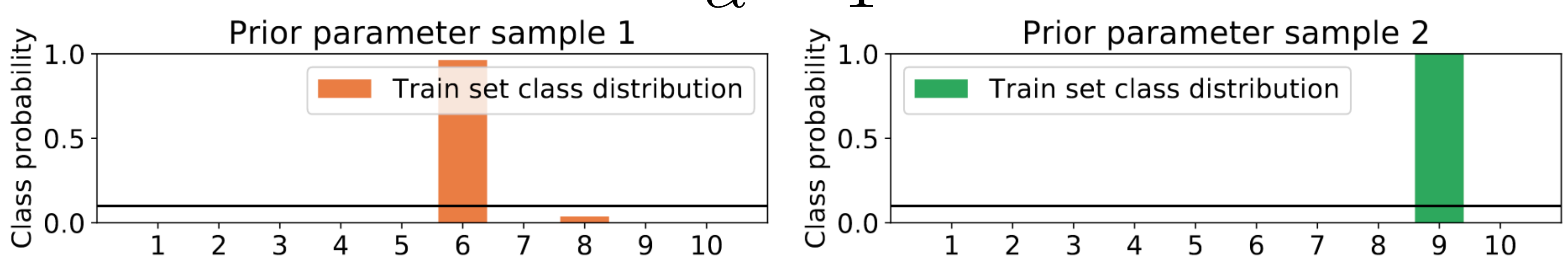


Figure 7. ResNet-20/CIFAR-10 typical prior predictive distributions for 10 classes under a $\mathcal{N}(0, I)$ prior averaged over the entire training set, $\mathbb{E}_{x \sim p(x)}[p(y|x, \boldsymbol{\theta}^{(i)})]$. Each plot is for one sample $\boldsymbol{\theta}^{(i)} \sim \mathcal{N}(0, I)$ from the prior. Given a sample $\boldsymbol{\theta}^{(i)}$ the average training data class distribution is highly concentrated around the same classes for all x.

Figure 7 of Wenzel et al, '20 https://arxiv.org/pdf/2002.02405.pdf

BAYESIAN DEEP LEARNING: COMPARISONS

Method	Accuracy	Calibration	Train time	Test time	Code
Ensembles (Lakshminarayanan et al, '17)	**high**	Often more overconfident	K times standard training	K times slower	**Train K models**
Swag (Maddox et al, '19)	Slightly better than MAP	**Less overconfident**	**Standard training**	K times slower	**Store models at train time**
Dropout (Gal & Gharamani, '16)	About the same as MAP	Slightly less overconfident	**Standard training**	K times slower	**Apply dropout at test time**
VOGN (Osawa et al, '19)	Slightly worse than MAP?*	Less overconfident	2x standard training	K times slower	Modify Adam

BAYESIAN DEEP LEARNING: COMPARISONS

Method	Accuracy	Calibration	Train time	Test time	Code
Ensembles (Lakshminarayanan et al, '17)	**high**	Often more overconfident	K times standard training	K times slower	**Train K models**
Swag (Maddox et al, '19)	Slightly better than MAP	**Less overconfident**	**Standard training**	K times slower	**Store models at train time**
Dropout (Gal & Gharamani, '16)	About the same as MAP	Slightly less overconfident	**Standard training**	K times slower	**Apply dropout at test time**
VOGN (Osawa et al, '19)	Slightly worse than MAP?*	Less overconfident	2x standard training	K times slower	Modify Adam

*: see figure 4, table 1 of Osawa et al, '19 (https://arxiv.org/pdf/1906.02506.pdf)

QUESTIONS?

Slides at https://wjmaddox.github.io/assets/BNN_tutorial_CILVR.pdf

REFERENCES (LAPLACE APPROXIMATIONS)

MacKay, David JC. "Bayesian interpolation." *Neural computation* 4.3 (1992): 415-447.
MacKay, David JC. "Bayesian model comparison and backprop nets." *Advances in neural information processing systems*. 1992.
MacKay, David JC. "A practical Bayesian framework for backpropagation networks." *Neural computation* 4.3 (1992): 448-472.
Foresee, F. Dan and Hagan, M. T. "Gauss-Newton Approximation to Bayesian Learning." ICNN (1997). 10.1109/ICNN.1997.614194

Ritter, Hippolyt, Aleksandar Botev, and David Barber. "A scalable laplace approximation for neural networks." *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*. Vol. 6. International Conference on Representation Learning, 2018.
Ritter, Hippolyt, Aleksandar Botev, and David Barber. "Online structured laplace approximations for overcoming catastrophic forgetting." *Advances in Neural Information Processing Systems*. 2018.

REFERENCES (VARIATIONAL INFERENCE)

Stochastic VI

Hoffman, Matthew D., et al. "Stochastic variational inference." *The Journal of Machine Learning Research* 14.1 (2013): 1303-1347.

Kucukelbir, Alp, et al. "Automatic differentiation variational inference." *The Journal of Machine Learning Research* 18.1 (2017): 430-474.

Ranganath, Rajesh, Sean Gerrish, and David Blei. "Black box variational inference." *Artificial Intelligence and Statistics*. 2014.

Graves, Alex. "Practical variational inference for neural networks." *Advances in neural information processing systems*. 2011.

Theory

Wang, Yixin, and David M. Blei. "Frequentist consistency of variational Bayes." *Journal of the American Statistical Association* 114.527 (2019): 1147-1161.

Yao, Yuling, et al. "Yes, but did it work?: Evaluating variational inference." *ICML, arXiv:1802.02538* (2018).

Knoblauch, Jeremias, Jack Jewson, and Theodoros Damoulas. "Generalized variational inference." *arXiv preprint arXiv:1904.02063* (2019).

REFERENCES (VARIATIONAL BOUNDS)

Bounds

Hernández-Lobato, José Miguel, and Ryan Adams. "Probabilistic backpropagation for scalable learning of bayesian neural networks." *International Conference on Machine Learning*. 2015.

Ranganath, Rajesh, et al. "Operator variational inference." *Advances in Neural Information Processing Systems*. 2016.

Dieng, Adji Bousso, et al. "Variational Inference via $\chi $ Upper Bound Minimization." *Advances in Neural Information Processing Systems*. 2017.

Bamler, Robert, et al. "Perturbative black box variational inference." *Advances in Neural Information Processing Systems*. 2017.

Ambrogioni, Luca, et al. "Wasserstein variational inference." *Advances in Neural Information Processing Systems*. 2018.

Hernández-Lobato, J. M., et al. "Black-Box α-divergence minimization." *Proceedings of the 33rd International Conference on Machine Learning*. Vol. 48. International Machine Learning Society, 2016.

Li, Yingzhen, and Richard E. Turner. "Rényi divergence variational inference." *Advances in Neural Information Processing Systems*. 2016.

Wu, Anqi, et al. "Deterministic variational inference for robust bayesian neural networks." *ICLR, arXiv:1810.03958* (2019).

Futami, Futoshi, Issei Sato, and Masashi Sugiyama. "Variational inference based on robust divergences." *arXiv preprint arXiv:1710.06595* (2017).

Tang, Da, and Rajesh Ranganath. "The Variational Predictive Natural Gradient." *ICML, arXiv:1903.02984* (2019).

REFERENCES (VARIATIONAL POSTERIORS)

Techniques

Blundell, Charles, et al. "Weight uncertainty in neural networks." *ICML, arXiv:1505.05424* (2015).

Fortunato, Meire, Charles Blundell, and Oriol Vinyals. "Bayesian recurrent neural networks." *WiML Workshop, arXiv:1704.02798* (2017).

Huszár, Ferenc. "Variational inference using implicit distributions." *arXiv preprint arXiv:1702.08235* (2017).

Louizos, Christos, and Max Welling. "Structured and efficient variational deep learning with matrix gaussian posteriors." *International Conference on Machine Learning*. 2016.

Louizos, Christos, and Max Welling. "Multiplicative normalizing flows for variational bayesian neural networks." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Miller, Andrew C., Nicholas J. Foti, and Ryan P. Adams. "Variational boosting: Iteratively refining posterior approximations." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Sun, Shengyang, Changyou Chen, and Lawrence Carin. "Learning structured weight uncertainty in bayesian neural networks." *Artificial Intelligence and Statistics*. 2017.

Tran, Dustin, Rajesh Ranganath, and David Blei. "Hierarchical implicit models and likelihood-free variational inference." *Advances in Neural Information Processing Systems*. 2017.

REFERENCE (VARIATIONAL INFERENCE)

Dropout

Li, Yingzhen, and Yarin Gal. "Dropout inference in Bayesian neural networks with alpha-divergences." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Kingma, Durk P., Tim Salimans, and Max Welling. "Variational dropout and the local reparameterization trick." *Advances in neural information processing systems*. 2015.

Li, Yingzhen, and Yarin Gal. "Dropout inference in Bayesian neural networks with alpha-divergences." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

Gal, Yarin, Jiri Hron, and Alex Kendall. "Concrete dropout." *Advances in neural information processing systems*. 2017.

Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. 2016.

Gal, Yarin, and Zoubin Ghahramani. "Bayesian convolutional neural networks with Bernoulli approximate variational inference." *ICLR Workshop Track, preprint arXiv:1506.02158* (2015).

Gal, Yarin. "Uncertainty in deep learning." *University of Cambridge* 1 (2016): 3.

Hron, Jiri, Alexander G. de G. Matthews, and Zoubin Ghahramani. "Variational gaussian dropout is not bayesian." NeurIPS BDL Workshop, *arXiv:1711.02989* (2017).

Hron, Jiri, Alex Matthews, and Zoubin Ghahramani. "Variational Bayesian dropout: pitfalls and fixes." *International Conference on Machine Learning*. 2018.

REFERENCES (OPTIMIZATION AS VI)

Duvenaud, David, Dougal Maclaurin, and Ryan Adams. "Early stopping as nonparametric variational inference." *Artificial Intelligence and Statistics*. 2016.
Khan, Mohammad Emtiyaz, et al. "Fast and scalable bayesian deep learning by weight-perturbation in adam." ICML, *arXiv:1806.04854* (2018).
Bae, Juhan, Guodong Zhang, and Roger Grosse. "Eigenvalue corrected noisy natural gradient." *NeurIPS BDL Workshop, arXiv:1811.12565* (2018).
Osawa, Kazuki, et al. "Practical deep learning with bayesian principles." *Advances in Neural Information Processing Systems*. 2019.
Hoffman, Matthew D., and Yian Ma. "Langevin Dynamics as Nonparametric Variational Inference." AABI (2019).

REFERENCES (MCMC)

Welling, Max, and Yee W. Teh. "Bayesian learning via stochastic gradient Langevin dynamics." *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011.
Neal, Radford M. "Bayesian learning via stochastic dynamics." *Advances in neural information processing systems*. 1993.

Neal, Radford M. *Bayesian learning for neural networks*. Vol. 118. Springer Science & Business Media, 2012.
Chen, Tianqi, Emily Fox, and Carlos Guestrin. "Stochastic gradient hamiltonian monte carlo." *International conference on machine learning*. 2014.
Zhang, Ruqi, et al. "Cyclical stochastic gradient mcmc for bayesian deep learning." *ICLR, arXiv:1902.03932* (2020).