SUBSPACE INFERENCE FOR BAYESIAN DEEP LEARNING

PAVEL IZMAILOV, WESLEY MADDOX, **POLINA KIRICHENKO**, TIMUR GARIPOV, DMITRY VETROV, ANDREW GORDON WILSON

WHY BAYESIAN INFERENCE?

- Combining models for better predictions 📊
- Uncertainty representation (crucial for decision making) 🤷‍♀️
- Interpretably incorporate prior knowledge and domain expertise 👩‍⚕️

WHY BAYESIAN INFERENCE?

- Combining models for better predictions 📊
- Uncertainty representation (crucial for decision making) 🤷‍♀️
- Interpretably incorporate prior knowledge and domain expertise 👩‍⚕️

WHY NOT?

- Challenging for Deep NNs due to high dimensional weight spaces 😩

SUBSPACE INFERENCE

A modular approach:

- Design subspace
- Approximate posterior over parameters in the subspace
- Sample from approximate posterior for Bayesian model averaging

SUBSPACE INFERENCE

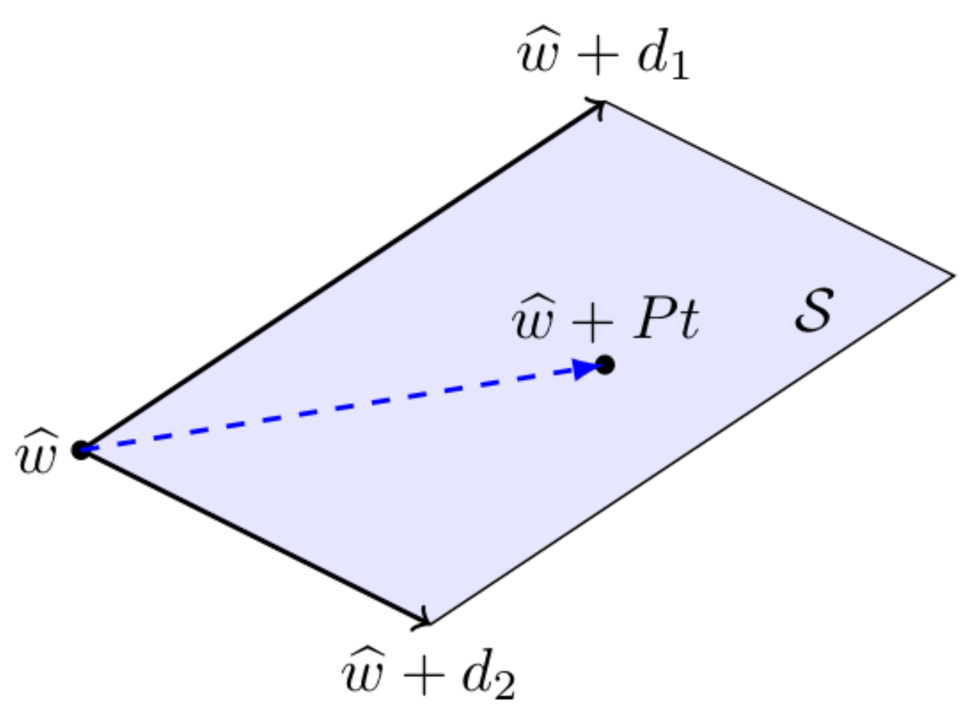
A modular approach:

- Design subspace
- Approximate posterior over parameters in the 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!

SUBSPACE

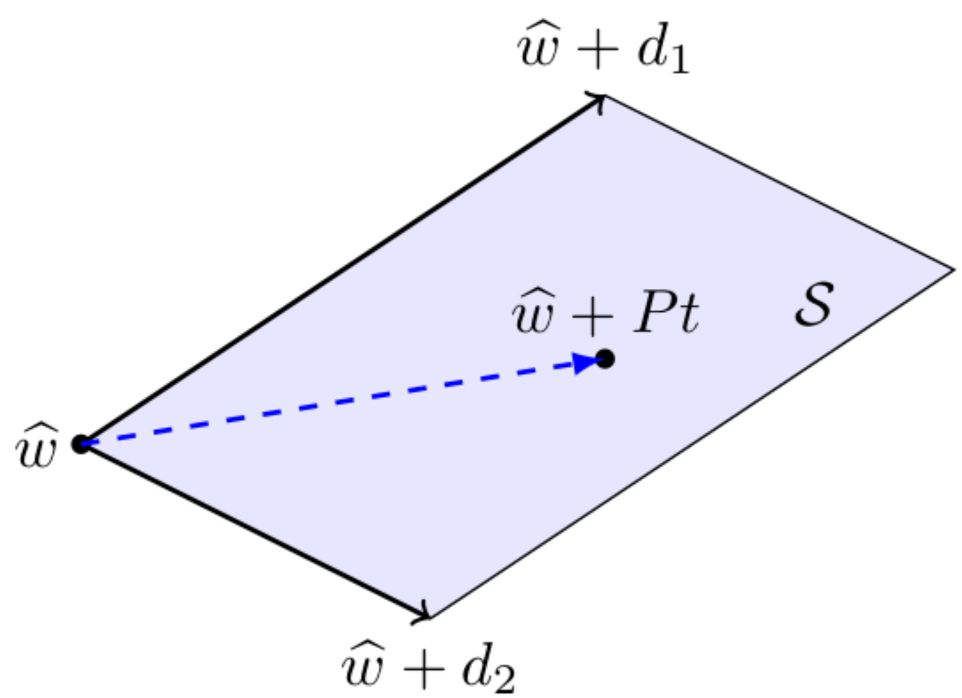
- Choose shift $\hat{w}$ and basis vectors $\{d_1, \ldots, d_K\}$
- Define subspace $S = \{w \,|\, w = \hat{w} + \underbrace{t_1 d_1 + \ldots + t_k d_K}_{Pt}\}$
- Likelihood $p(D \,|\, t) = p_M(D \,|\, w = \hat{w} + Pt)\,.$



INFERENCE

- Approximate inference over parameters t
    - MCMC, Variational Inference, Normalizing Flows, …
- Bayesian model averaging at test time:

$$p(D^* \mid D) = \frac{1}{J} \sum_{i=1}^{J} p_M(D^* \mid \tilde{w} = \hat{w} + P\tilde{t}_i), \quad \tilde{t}_i \sim q(t \mid D)$$



TEMPERING POSTERIOR

- In the subspace model *# parameters << # data points*
  - *~5-10 parameters, ~50K data points*
- Posterior over t is extremely concentrated
- To address this issue, we utilize the tempered posterior:

$$p_T(t \mid D) \propto \underbrace{p(D \mid t)^{1/T}}_{\text{likelihood}} \underbrace{p(t)}_{\text{prior}}$$

- *T* can be learned by cross-validation
- Heuristic: $T = \dfrac{\text{\# data points}}{\text{\# parameters}}$

SUBSPACE CHOICE

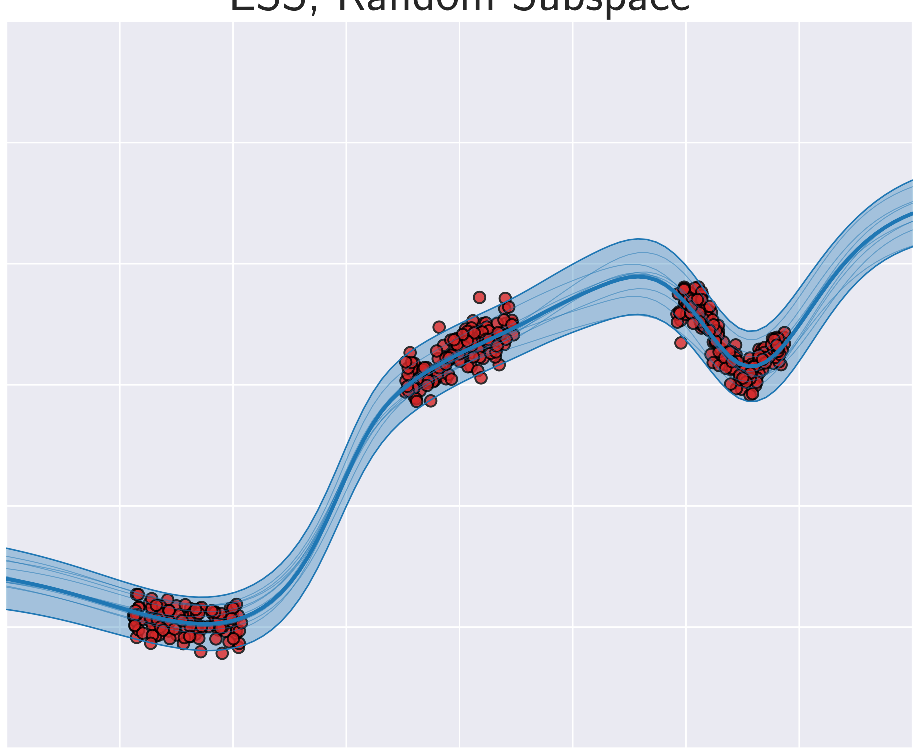
We want a subspace that

- Contains **diverse** models
- **Cheap** to construct

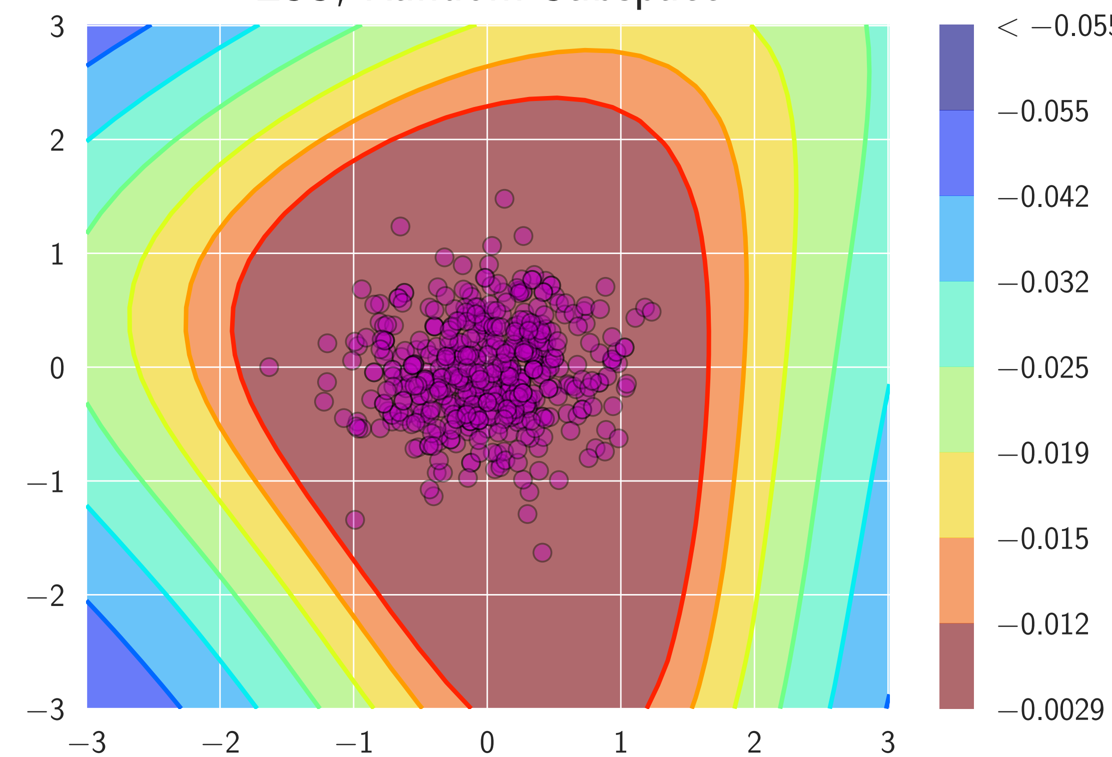
RANDOM SUBSPACE

- Directions $d_1, \ldots, d_K \sim N(0, I_p)$
- Use pre-trained solution as shift $\hat{w}$
- Subspace $S = \{w \,|\, w = \hat{w} + Pt\}$

Predictive Distribution
ESS, Random Subspace



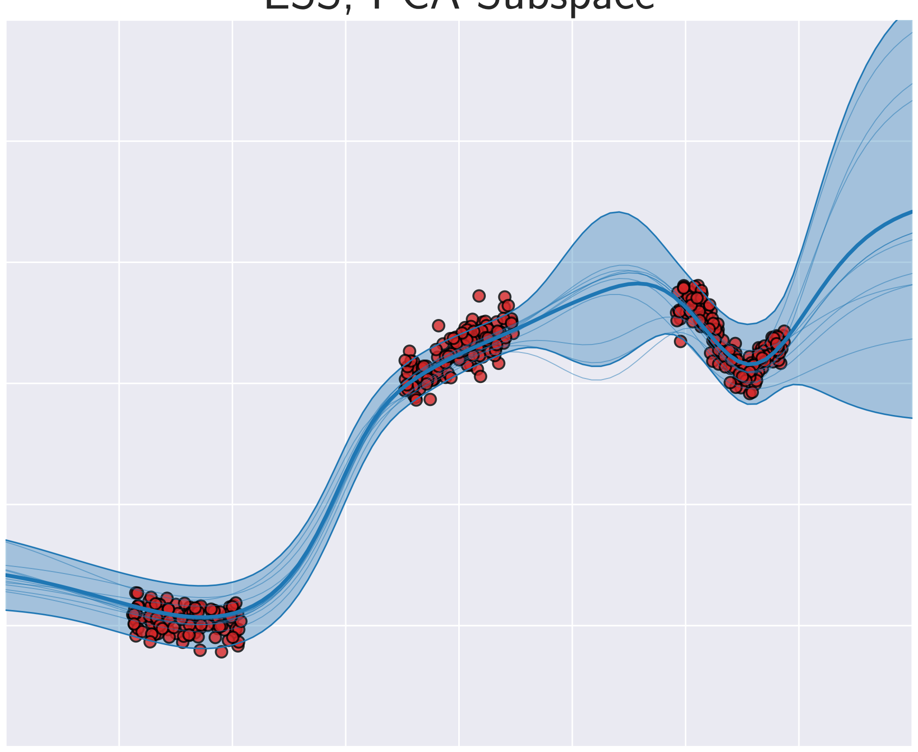
Posterior log-density
ESS, Random Subspace



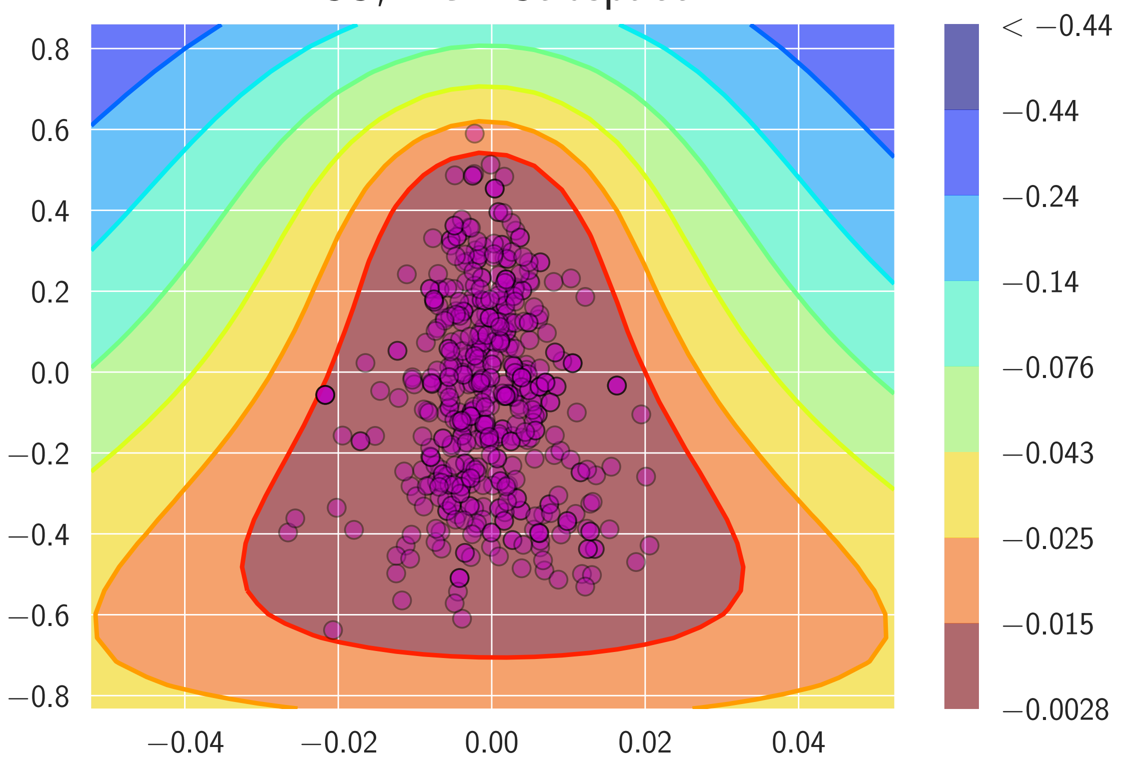
PCA OF THE SGD TRAJECTORY

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

Predictive Distribution
ESS, PCA Subspace

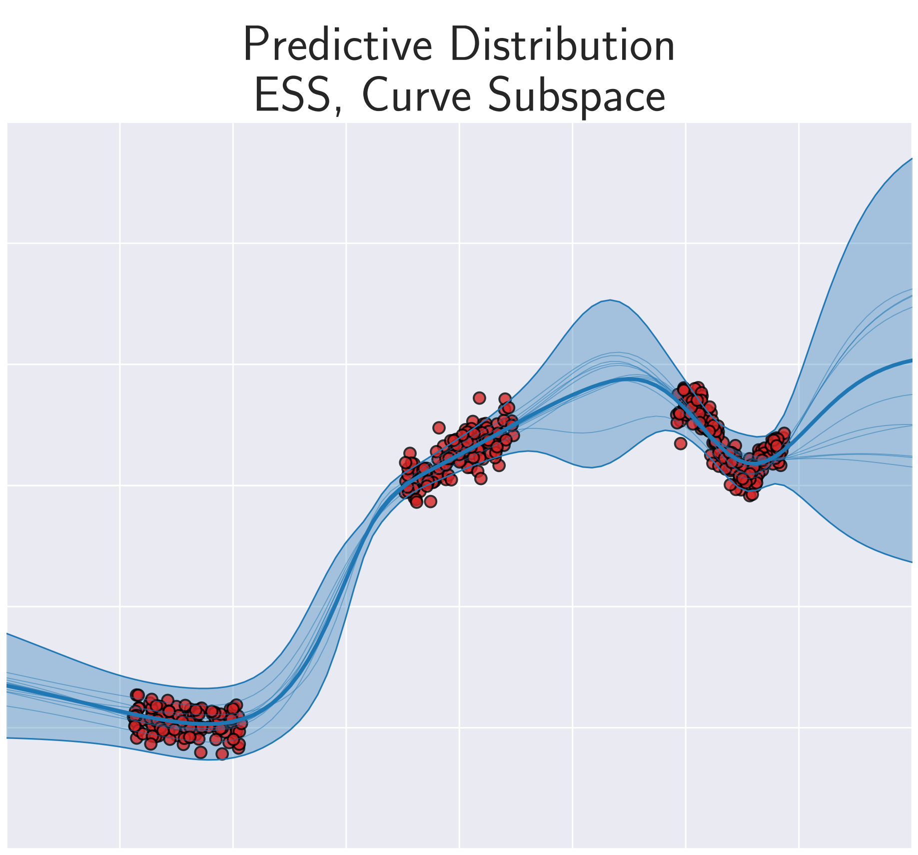


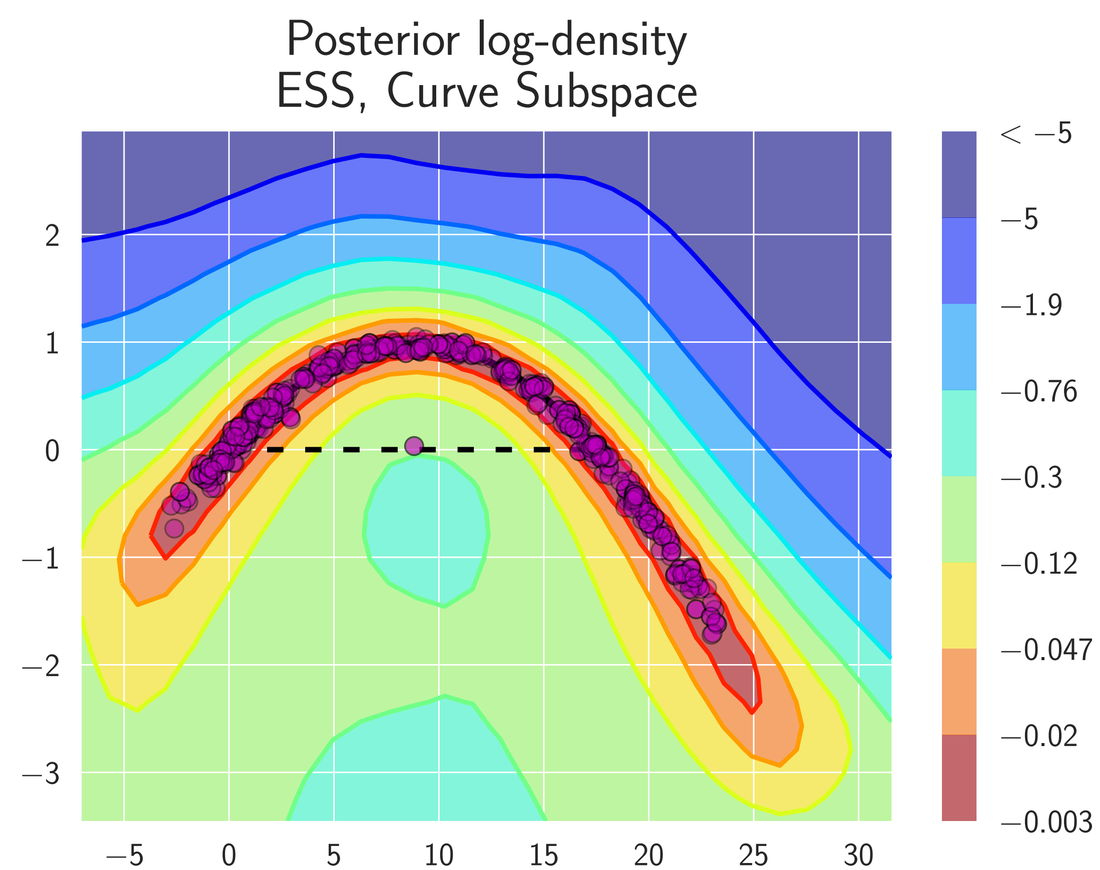
Posterior log-density
ESS, PCA Subspace



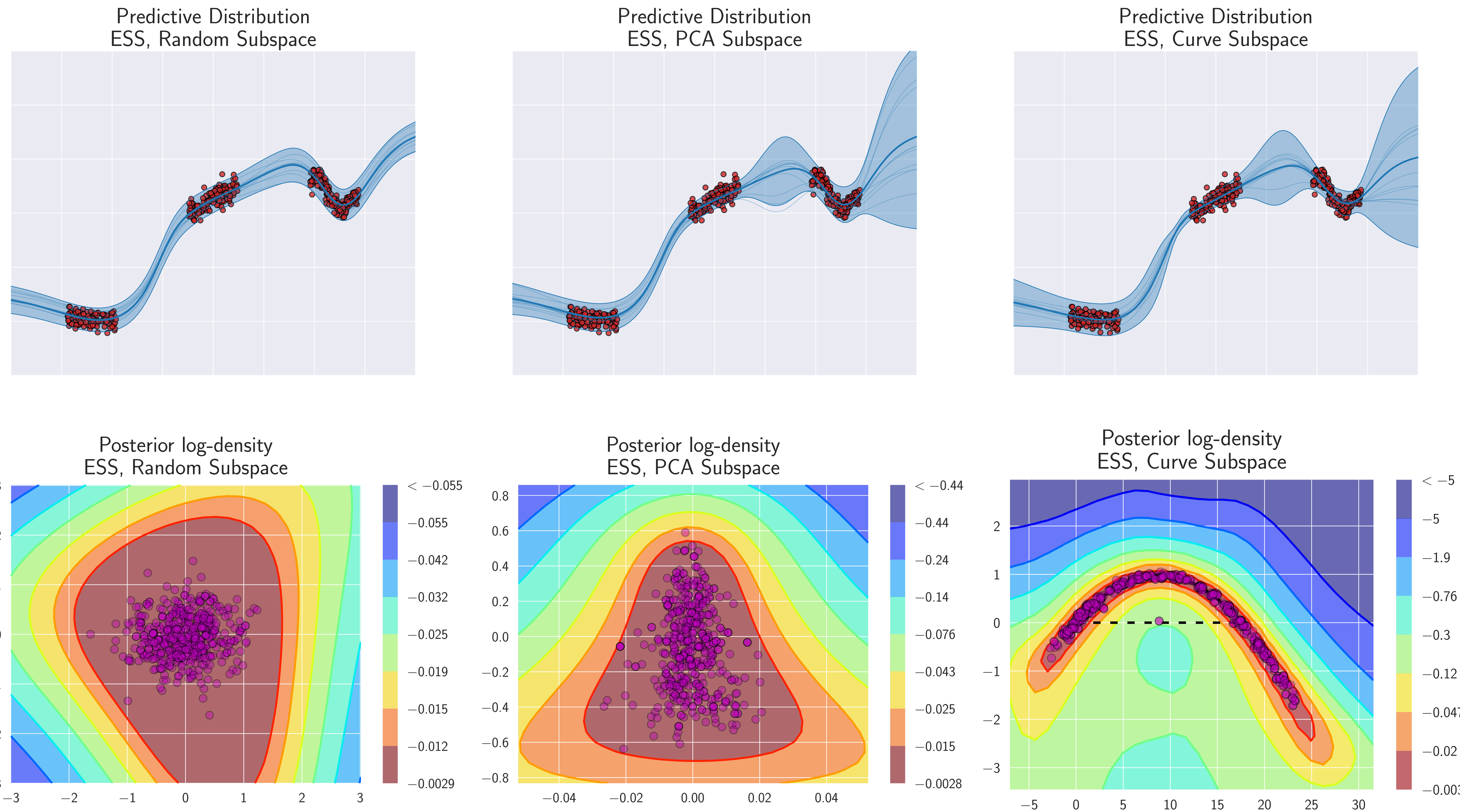
CURVE SUBSPACE

- Garipov et al. 2018 proposed a method to find 2D subspaces containing a path of low loss between weights of two independently trained neural networks

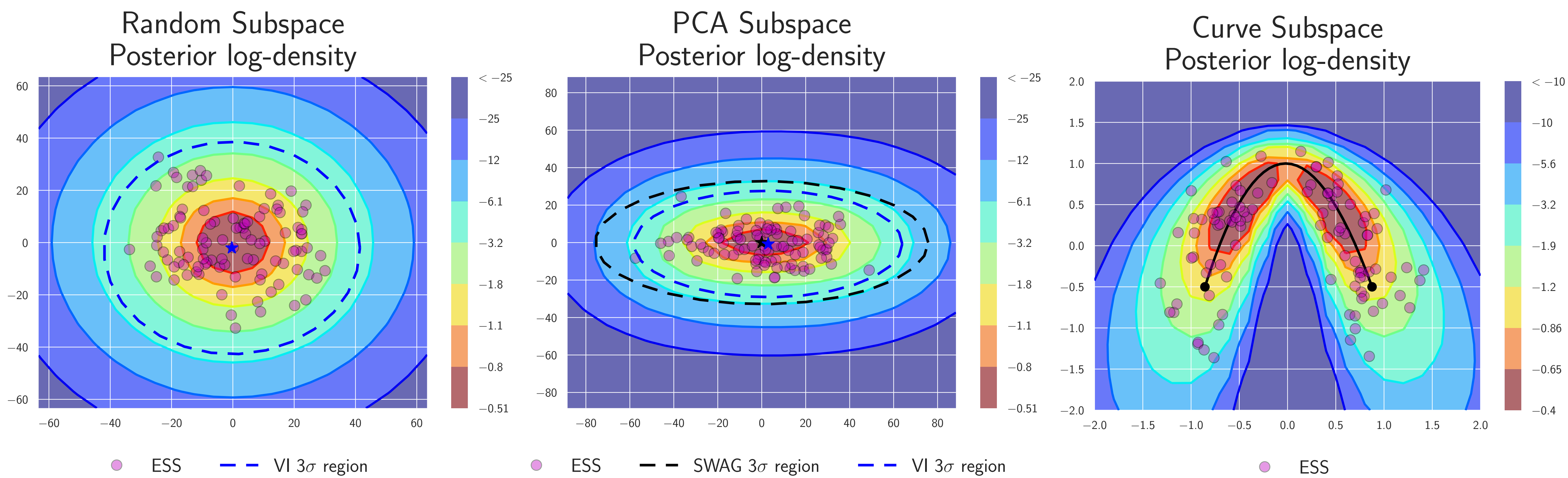




SUBSPACE COMPARISON



SUBSPACE COMPARISON ON PRERESNET-164, CIFAR-100



	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

TAKEAWAYS

- We can apply standard approximate inference methods in subspaces of parameter space
- More diverse subspaces => better performance:
  Curve Subspace > **PCA Subspace** > Random Subspace
- Subspace Inference in the PCA subspace is competitive with SWAG (Maddox et al., 2019), MC-Dropout (Gal & Ghahramani, 2016) and Temperature Scaling (Guo et al., 2017) on image classification and UCI regression