Subspace inference for bayesian deep learning

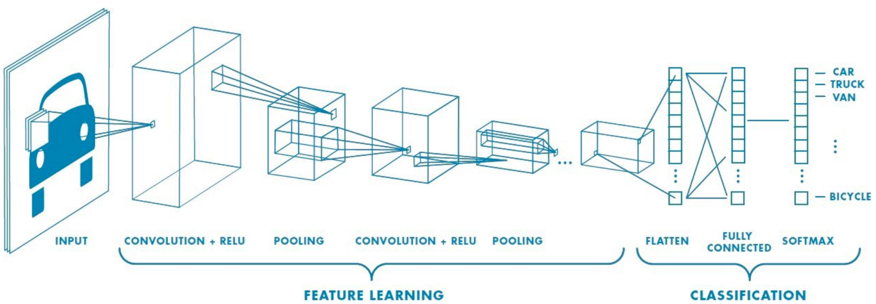
Pavel Izmailov, Wesley J. Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, Andrew Gorden Wilson

Uncertainty in Artificial Intelligence (UAI), 2019

Suspace inference: the idea

Parameter space. *All* the valid parameter settings available for a model; $\mathbb{R}^{|\theta|}$

Neural network has lots of parameters



Linear subspace

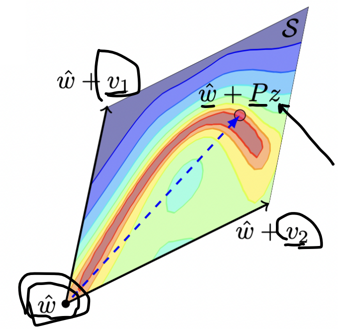


Figure 2: Illustration of subspace $\mathcal{S}$ with shift vector $\hat{w}$ and basis vectors v_1, v_2, with a contour plot of the posterior log-density over parameters z.

Assumption. Linear subspace of such a large parameter space might still contain a diverse set of high performing models! (given the subspace is chosen carefully)

Once we decided on basis vectors, the new "parameters" are the coefficients of the basis vectors. The basis vectors will all have dimensionality $|\boldsymbol{\theta}|$, but the number of coefficients might be small.

(Why is this ok to do? Maybe we just want to do Bayesian inference on this subspace to get **well-calibrated uncertainty estimations** / "Even though the parameter space is very high dimensional, a lot of functional variabilities can be captured in a low dimension subspace.")

Advantage. Apply standard bayesian inference techniques such as elliptical slide sampling (ESS) (an MCMC method) and variational inference with more flexible approximate posteriors.

Subspace inference: the algorithm

1. Construct subspace (i.e., choose its basis vectors)
   - Random subspace (shift vector: SWA solution; basis: 5 random vectors)
   - PCA subspace (shift vector: SWA solution; basis: 5 principle dims of SWA residual)*
   - Curved subspace (2d subspace*)
2. Posterior inference within the subspace (with simple prior choices $N(0, I)$)
   - Powerful, exact full-batch MCMC methods (HMC, ESS)
   - Deterministic approximation: variational approach with flexible variational family
3. Form a Bayesian model average with Monte Carlo samples of the following integral

$$p(y|\, x, D) = \int p(y|\, x, w = \hat{w} + Pz) p(z|\, D)\, dz$$

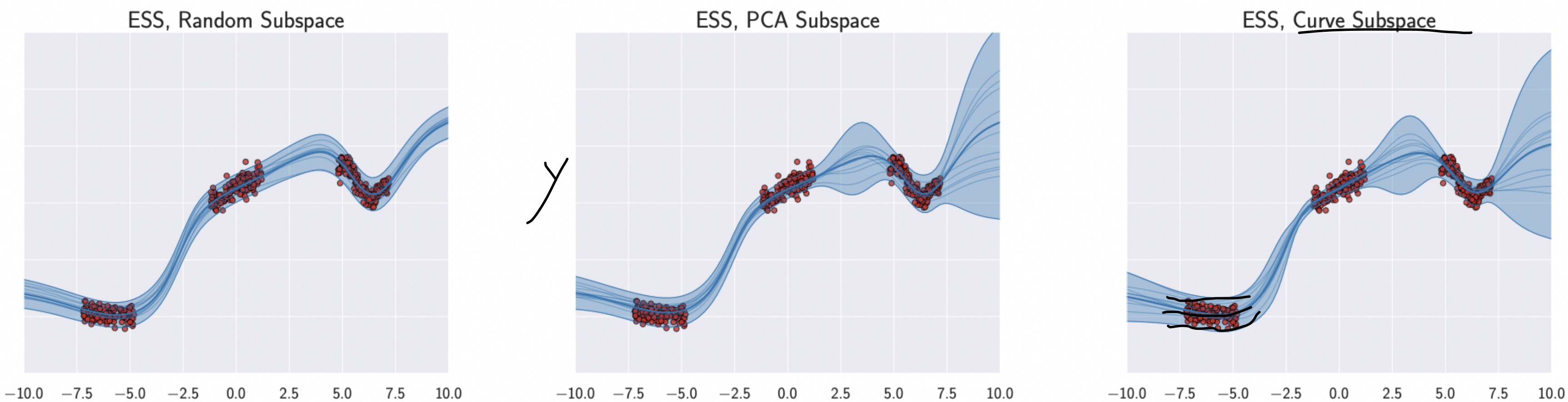
(So this is not just a reparametrization of the original model! It's a different model!)

*Talk about similarity with SWAG; it's quite clear why doing subspace inference would improve performance

*Show diagram in https://arxiv.org/pdf/1802.10026.pdf or https://arxiv.org/abs/1910.03867 (quite unexpected but useful results!)

Subspace inference: prelim results

On simple regression dataset:



NLL and Accuracy for PreResNet-164 for 10-d random, 10-d PCA, and 2-d curve subspaces. We report mean and stdev over 3 independent runs:

	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

Models. PreResNet-164, WideResNet28x10

Datasets. CIFAR10 (10 classes, 6000 per class), CIFAR100 (100 classes, 600 per class)

Prior. tempered priors

Technique. Remaining experiments we use the PCA subspace, generally provides good performance at a much lower computational cost than the curved subspace; ESS / simple VI in subspace

competitive with SWAG, which is the state of the art?

which kind of make sense?