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 INFEERENCE

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!
Choose shift $\hat{w}$ and basis vectors $\{d_1, \ldots, d_K\}$

Define subspace $S = \{w \mid w = \hat{w} + t_1 d_1 + \ldots + t_k d_K\}$

Likelihood $p(D \mid t) = p_M(D \mid w = \hat{w} + Pt)$. 

![Diagram of subspace](image)
**INFECTION**

- Approximate inference over parameters $t$
  - MCMC, Variational Inference, Normalizing Flows, ...

- Bayesian model averaging at test time:

\[
p(D^* | D) = \frac{1}{J} \sum_{i=1}^{J} p_M(D^* | \tilde{w} = \hat{w} + P\tilde{t}_i), \quad \tilde{t}_i \sim q(t | D)
\]
TEMPERING POSTERIOR

- In the subspace model \# parameters \ll \# data points
  - \sim 5-10 parameters, \sim 50K data points

- Posterior over \( t \) is extremely concentrated

- To address this issue, we utilize the tempered posterior:
  \[
  p_T(t | D) \propto p(D | t)^{1/T} \underbrace{p(t)}_{likelihood \; \; prior}
  \]

- \( T \) can be learned by cross-validation

- Heuristic: \[ T = \frac{\# \; data \; points}{\# \; parameters} \]
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 \( \hat{w} - w_i \)
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

Predictive Distribution
ESS, Random Subspace

Predictive Distribution
ESS, PCA Subspace

Predictive Distribution
ESS, Curve Subspace

Posterior log-density
ESS, Random Subspace

Posterior log-density
ESS, PCA Subspace

Posterior log-density
ESS, Curve Subspace
### Subspace Comparison on PrerResNet-164, CIFAR-100

<table>
<thead>
<tr>
<th></th>
<th>SGD</th>
<th>Random</th>
<th>PCA</th>
<th>Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLL</td>
<td>0.946 ± 0.001</td>
<td>0.686 ± 0.005</td>
<td>0.665 ± 0.004</td>
<td>0.646</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>78.50 ± 0.32</td>
<td>80.17 ±0.03</td>
<td>80.54 ± 0.13</td>
<td>81.28</td>
</tr>
</tbody>
</table>
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