Evaluating and Improving Subspace Inference in Bayesian Deep Learning

Yifei Xiong

Department of Statistics, Purdue University

Joint with Nianqiao P. Ju (Dept. of Statistics, Purdue)
and Ruqi Zhang (Dept. of Computer Science, Purdue)

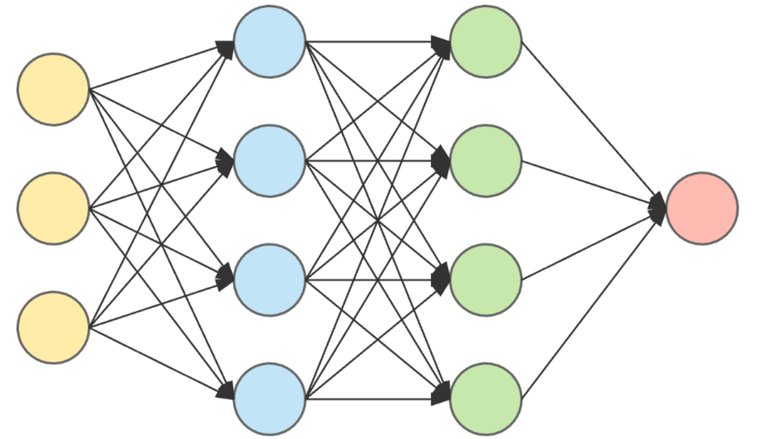
Nov 21, 2024

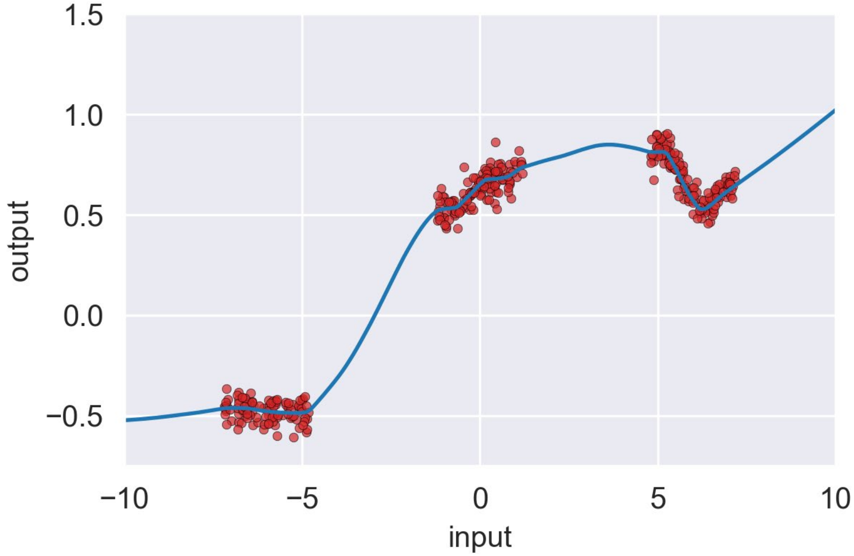
yifei-xiong.github.io

Neural Networks

- Neural Networks: $f_w(\cdot)$, $w \in \mathbb{R}^d$ are the network weights
  - Given training dataset $D = \{X, Y\}$, neural network methods typically find the optimal weights w^* (e.g., by stochastic gradient descent (SGD))

$$w^* = \underset{w}{\operatorname{argmin}} \operatorname{Loss}(f_w(X), Y)$$





Neural Networks

- After obtaining the weights w^*, the neural network can provide an output

$$\tilde{Y} = f_{w^*}(\tilde{X})$$

for testing data $\tilde{X}$.

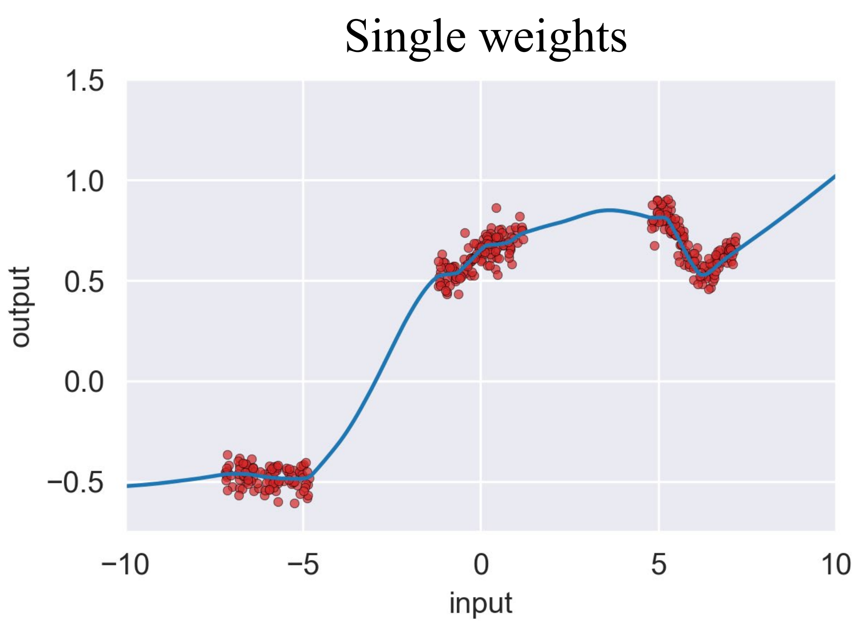
Neural Networks

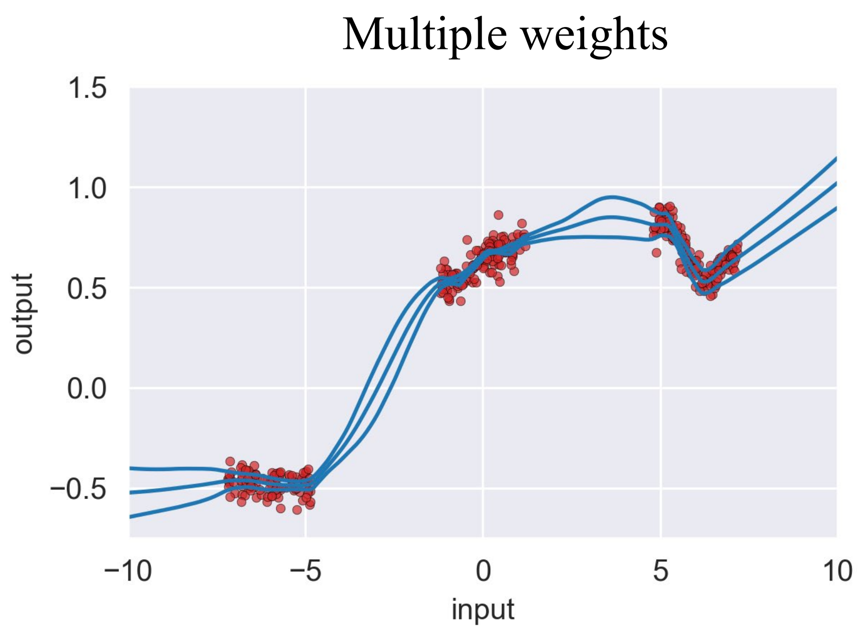
- After obtaining the weights w^*, the neural network can provide an output

$$\tilde{Y} = f_{w^*}(\tilde{X})$$

for testing data $\tilde{X}$.

without quantifying the uncertainty for w^* or $\tilde{Y}$





Bayesian Inference in Neural Networks

- Bayes' Theorem:

$$p(w|D) \propto p(w)\, p(D|w)$$

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

- We transform the loss function into a likelihood with

$$\ell(w; D) = \log p_w(Y|X) = -\mathrm{Loss}(f_w(X), Y)$$

and aim to study the posterior

$$p(w|D) \propto p(w)\, p_w(Y|X).$$

Bayesian Inference in Neural Networks

- Posterior of weights

$$p(w|D) \propto p(w)p(D|w)$$

- Posterior predictive distribution for another dataset D'

$$p(D'|D) = \int p(D'|w)p(w|D)\,\mathrm{d}w$$

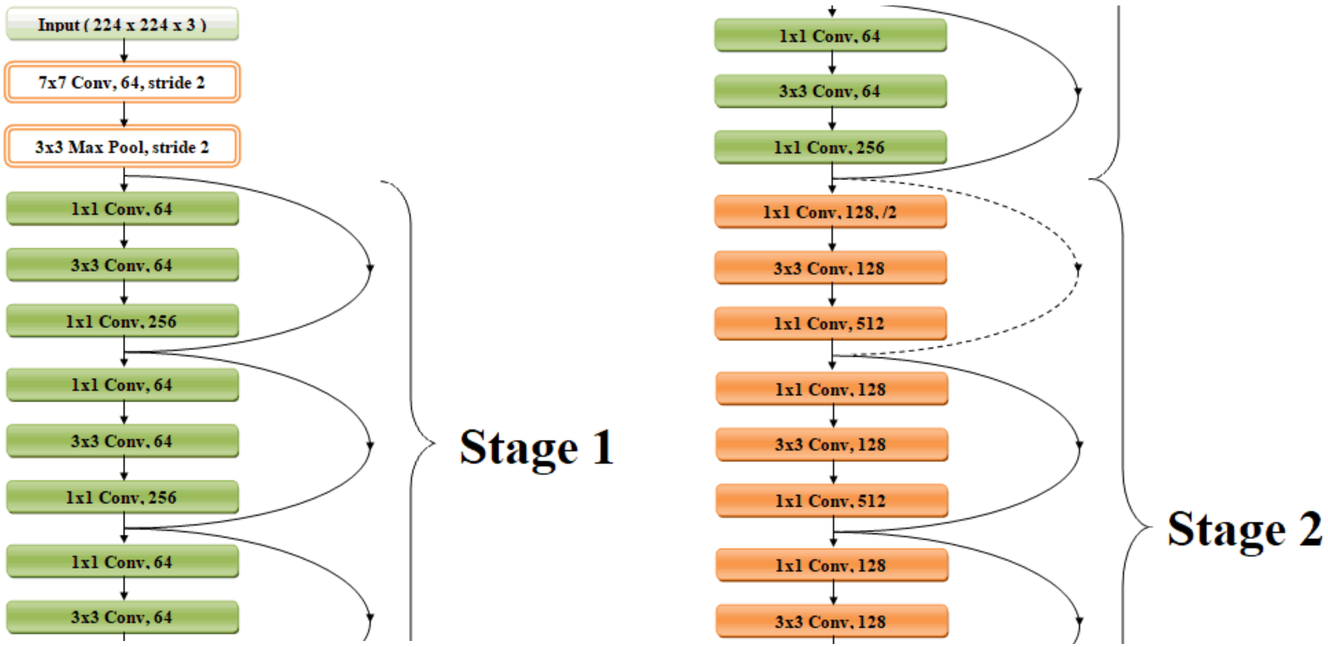
(Monte Carlo estimator) $$\approx \frac{1}{N}\sum_{i=1}^{N} p(D'|w_i)\,, w_i \sim p(w|D)$$

Bayesian Inference in Neural Networks

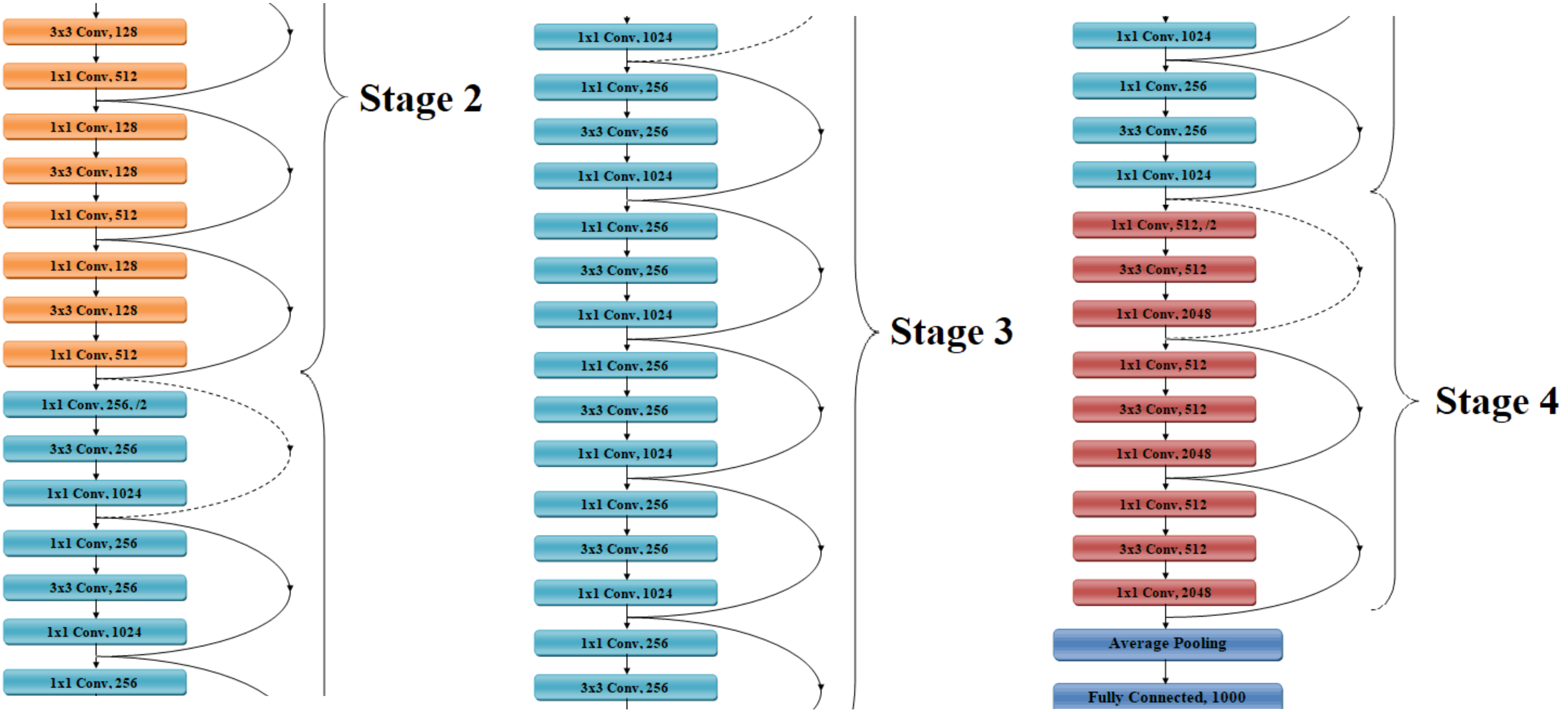
- We aim to study the posterior

$$p(w|D) \propto p(w)p(D|w).$$

- The dimension of deep neural network weights are very high:
  - Structure of ResNet-164



Bayesian Inference in Neural Networks



About 1.7 million parameters

Subspace Inference

- The dimension of the neural network weights are very high:
  - ResNet-164: ~ 1.7 million parameters
  - VGG-16: ~ 138 million parameters

- Sampling from the exact posterior $p(w|D)$ is difficult due to high dimension.

Subspace Inference

- Sampling from the exact posterior $p(w|D)$ is difficult due to high dimension.

- Subspace inference methods aim to construct a low dimensional posterior
  - Retains meaningful properties of $p(w|D)$ and $p(D'|D)$

- We can perform inference in a k-dimensional subspace
  - Linear subspaces $\mathcal{Z} = \{\hat{w} + Pz | z \in \mathbb{R}^k\} \subset \mathcal{W} = \mathbb{R}^d$
  - (Li et al., 2018; Izmailov et al., 2020)

PURDUE UNIVERSITY

Subspace Inference

- Perform inference in a k-dimensional subspace $\mathcal{Z} = \{\hat{w} + Pz | z \in \mathbb{R}^k\}$
  - Induced likelihood: $p_{\mathcal{Z}}(\mathcal{D}|z) = p(\mathcal{D}|w = \hat{w} + Pz)$
  - Induced posterior: $p_{\mathcal{Z}}(z|\mathcal{D}) \propto p_{\mathcal{Z}}(z) p_{\mathcal{Z}}(\mathcal{D}|z)$

Subspace Inference

- Perform inference in a k-dimensional subspace $\mathcal{Z} = \{\hat{w} + Pz | z \in \mathbb{R}^k\}$
  - Induced likelihood: $p_{\mathcal{Z}}(\mathcal{D}|z) = p(\mathcal{D}|w = \hat{w} + Pz)$
  - Induced posterior: $p_{\mathcal{Z}}(z|\mathcal{D}) \propto p_{\mathcal{Z}}(z) p_{\mathcal{Z}}(\mathcal{D}|z)$

- Posterior predictive:

$$p_{\mathcal{Z}}(\mathcal{D}'|\mathcal{D}) = \int_{\mathcal{Z}} p_{\mathcal{Z}}(\mathcal{D}'|z) p_{\mathcal{Z}}(z|\mathcal{D}) \, \mathrm{d}z \approx \frac{1}{N} \sum_{i=1}^{N} p_{\mathcal{Z}}(\mathcal{D}'|z_i)$$

with $z_i \sim p_{\mathcal{Z}}(z|\mathcal{D})$.

Subspace Construction

- Perform inference in a k-dimensional subspace $\mathcal{Z} = \left\{\hat{w} + Pz | z \in \mathbb{R}^k\right\}$
  - Key Question: How to choose $\hat{w}$ and P based on training data?
  - When minimizing the loss, we have the corresponding SGD trajectory $w_1, \cdots, w_n$

- Stochastic weight averaging (SWA, Izmailov et al., 2018)
  - Averaging weights $\hat{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$ along the SGD trajectory can find a generalizable solutions compared to SGD's final point w_n

Subspace Construction

- Perform inference in a k-dimensional subspace $\mathcal{Z} = \{\hat{w} + Pz | z \in \mathbb{R}^k\}$

➢ How to choose $\hat{w}$ and P based on training data?

- When minimizing the loss, we have the corresponding SGD trajectory $w_1, \cdots, w_n$

➢ Full trajectory (FT) subspace:

- set $\hat{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$, and obtain projection matrix P via PCA on all trajectories.

➢ Tail trajectory (TT) subspace (Izmailov et al., 2020):

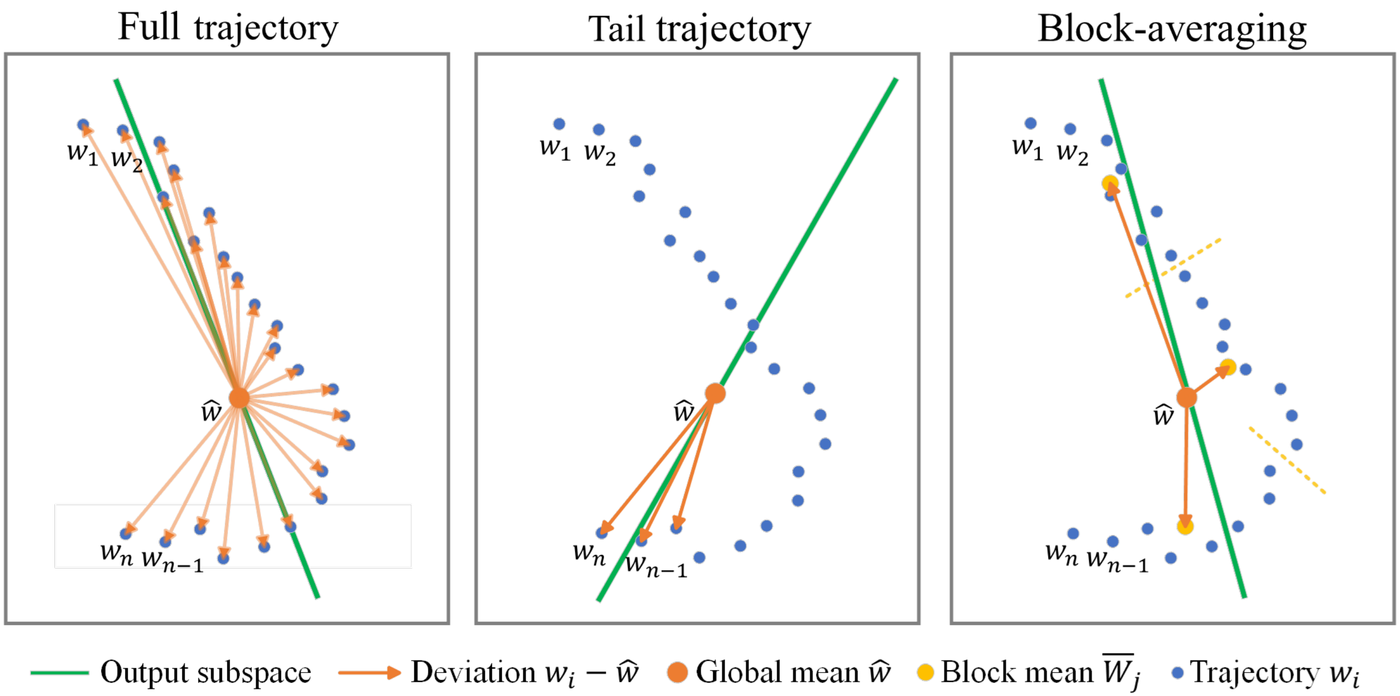
- set $\hat{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$, and obtain projection matrix P from the deviations of the last M points using SVD.

	FT subspace	**TT subspace**	**BA subspace**
Memory Cost	$O(nd)$	$O(Md)$	$O(Md)$
Computational Cost	$O(nd \log k)$	$O(Md \log k)$	$O(Md \log k)$

PURDUE UNIVERSITY

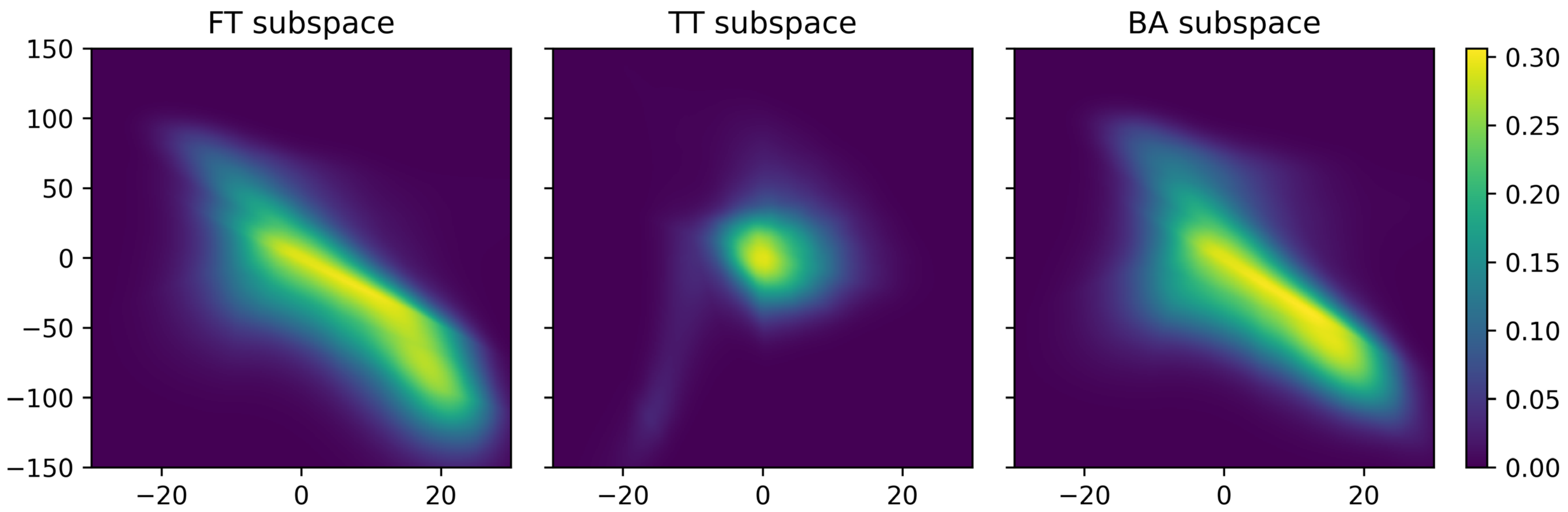
Our method: Block-Averaging Subspace

- We propose a block-averaging (BA) construction strategy
  - partitions the trajectory into M equidistant blocks and perform randomized SVD Halko et al., 2011) on block centers.
  - captures the global features of the entire trajectory



Our method: Block-Averaging Subspace

- Heat maps of induced likelihood across different subspaces



- BA has same algorithmic complexity and memory cost as the TT subspace
- BA subspace is similar respect to FT subspace & contains more high likelihood points

Subspace Quality Evaluation

- Current works skip subspace evaluation and directly start making predictions using the posterior predictive distribution.

➢ We need the quality evaluations for different subspaces.

Subspace Quality Evaluation

- **Def. 1** Subspace evidence (marginal likelihood)

$$p(\mathcal{D}|\mathcal{Z}) = \int_{w\in\mathcal{Z}} p_{\mathcal{W}}(\mathcal{D}|w)p_{\mathcal{W}}(w)\,\mathrm{d}w = \int_{z\in\mathbb{R}^k} p_{\mathcal{Z}}(\mathcal{D}|z)p_{\mathcal{Z}}(z)\,\mathrm{d}z.$$

- **Def. 2** Bayes factor for subspaces

$$\mathrm{BF}_{1,2} = \frac{p(\mathcal{D}|\mathcal{Z}_1)}{p(\mathcal{D}|\mathcal{Z}_2)}.$$

- If $\mathrm{BF}_{1,2}$ is large than 1, $\mathcal{Z}_1$ is better.

Subspace Quality Evaluation

$$p(\mathcal{D}|\mathcal{Z}) = \int_{w\in\mathcal{Z}} p_{\mathcal{W}}(\mathcal{D}|w)p_{\mathcal{W}}(w)\,\mathrm{d}w = \int_{z\in\mathbb{R}^k} p_{\mathcal{Z}}(\mathcal{D}|z)p_{\mathcal{Z}}(z)\,\mathrm{d}z.$$

$$\mathrm{BF}_{1,2} = \frac{p(\mathcal{D}|\mathcal{Z}_1)}{p(\mathcal{D}|\mathcal{Z}_2)}.$$

- Jeffery's scale of evidence (Kass and Raftery, 1995) gives an interpretation for Bayes factors:

- With $\mathrm{BF}_{1,2} > 10$ there is **strong evidence** favoring subspace $\mathcal{Z}_1$ and $\mathrm{BF}_{1,2} > \sqrt{10} \approx 3.2$ gives substantial evidence for $\mathcal{Z}_1$.
- Similarly, $\mathrm{BF}_{1,2} < 0.1$ or $\mathrm{BF}_{1,2} < 0.32$ gives **strong / substantial evidence** for choosing $\mathcal{Z}_2$.

PURDUE UNIVERSITY

Subspace Quality Evaluation

- Evidence ratios (y-axis) for different subspace construction methods using different M values (x-axis). Blue: TT against FT; Orange: BA against FT.
- Subspaces constructed from the BA trajectory outperform those from TT.

Efficient Posterior Predictive Checks

- Posterior predictive approximation with $z_i \sim p_Z(z|\mathcal{D})$:

$$p_Z(\mathcal{D}'|\mathcal{D}) = \int_Z p_Z(\mathcal{D}'|z) p_Z(z|\mathcal{D})\,\mathrm{d}z \approx \frac{1}{N}\sum_{i=1}^{N} p_Z(\mathcal{D}'|z_i)$$

➢ The evaluation of $p_Z(z|\mathcal{D})$ has a large computational overhead

  - (Perform a forward pass on the training dataset)

Efficient Posterior Predictive Checks

- Posterior predictive approximation with $z_i \sim p_{\mathcal{Z}}(z|\mathcal{D})$:

$$p_{\mathcal{Z}}(\mathcal{D}'|\mathcal{D}) = \int_{\mathcal{Z}} p_{\mathcal{Z}}(\mathcal{D}'|z) p_{\mathcal{Z}}(z|\mathcal{D})\, \mathrm{d}z \approx \frac{1}{N}\sum_{i=1}^{N} p_{\mathcal{Z}}(\mathcal{D}'|z_i)$$

- The evaluation of $p_{\mathcal{Z}}(z|\mathcal{D})$ has a large computational overhead
  - (Perform a forward pass on the training dataset)
- Weights z_i from the trajectory used to train $\mathcal{Z}$ have an empirical mean of 0 and an empirical covariance of I_k after projection
  - We can approximate this by importance sampling (IS)

Efficient Posterior Predictive Checks

- Self-normalized importance sampling (SNIS) estimator:

$$\mathbb{E}_p[f] = \int f(x)p(x)dx = \int f(x)q(x)\frac{p(x)}{q(x)}dx = \mathbb{E}_q\left[\frac{fp}{q}\right]$$

- Root mean squared error (RMSE) of SNIS estimator: $\mathcal{O}(N^{-0.5})$

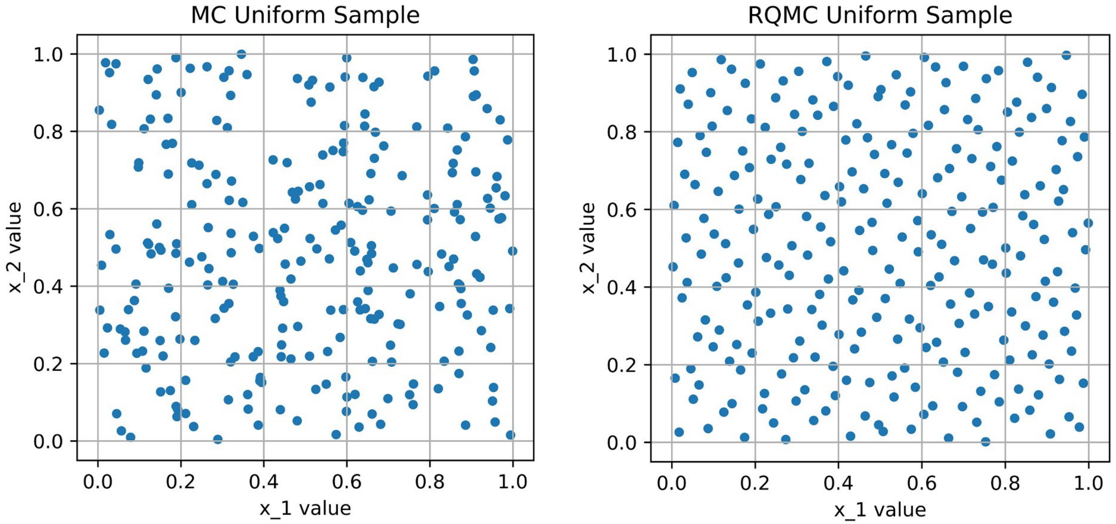
Efficient Posterior Predictive Checks

- Self-normalized importance sampling (SNIS) estimator:

$$\mathbb{E}_p[f] = \int f(x)p(x)dx = \int f(x)q(x)\frac{p(x)}{q(x)}dx = \mathbb{E}_q\left[\frac{fp}{q}\right]$$

- Root mean squared error (RMSE) of SNIS estimator: $\mathcal{O}(N^{-0.5})$

- When the dimensionality of $\mathcal{Z}$ is small (e.g., $k \leq 5$), we can use randomized quasi-Monte Carlo (RQMC) (Owen, 1997a; L'Ecuyer, 2018) to further reduce the RMSE.



Efficient Posterior Predictive Checks

- **Thm. 1** (Convergence rate). Under the Assumption 3 and 4, the RMSE for the RQMC-IS estimator satisfies

$$\sqrt{\mathbb{E}\left[\left(\hat{p}_{\mathrm{RQMC}}(N, q; \mathcal{D}, \mathcal{D}') - p_{\mathcal{Z}}(\mathcal{D}'|\mathcal{D})\right)^2\right]} = \mathcal{O}(N^{-1+\epsilon})$$

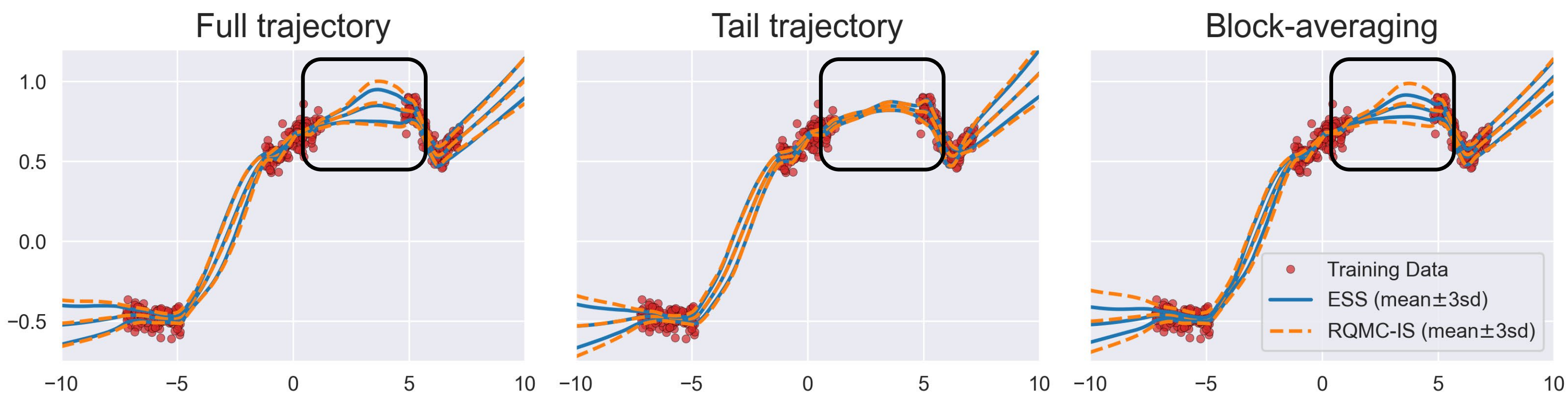
- for arbitrarily small $\epsilon > 0$.

Table 2: RMSE of posterior predictive estimations in different subspaces. The cost is measured by the number of forward passes through the model on the training set.

Method	Full trajectory RMSE	Full trajectory Cost	Tail trajectory RMSE	Tail trajectory Cost	Block-averaging RMSE	Block-averaging Cost
(MCMC) ESS	0.0091	6716±119.9	0.0110	5630±104.5	0.0091	6663±103.7
VI	0.0488	2000	0.0606	2000	0.0479	2000
SNIS ($N = 256$)	0.0137	256	0.0102	256	0.0141	256
SNIS ($N = 1024$)	0.0064	1024	0.0052	1024	0.0065	1024
RQMC-IS ($N = 256$)	0.0103	256	0.0031	256	0.0092	256
RQMC-IS ($N = 1024$)	**0.0026**	1024	**0.0006**	1024	**0.0028**	1024

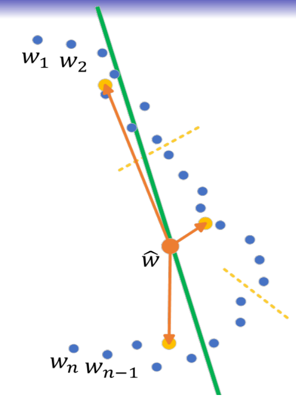
Case Study: Uncertainty quantification

➢ Visualizing uncertainty using posterior predictive:

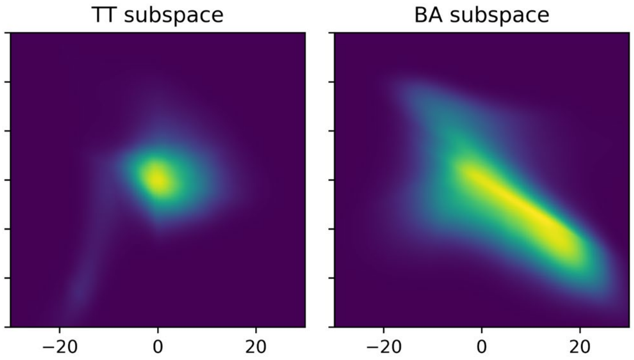


- The full trajectory (FT) and block-averaging (BA) subspace reflect higher uncertainty in data-sparse regions and higher confidence in data-rich regions, while the tail trajectory (TT) tends to be overconfident.

Summary



- Subspace Construction
  - Subspace should include representative and diverse points



- Quality Evaluation
  - Subspace Evidence; Bayes Factors
- Sampling from Posterior
  - MC (RQMC, Importance Sampling, ...)
  - MCMC (Hamiltonian Monte Carlo, Stochastic Gradient Langevin Dynamics, ...)
  - Other Machine Learning Methods (Variational Inference, Normalizing flows, ...)

Reference

Berger, J. O. (2003). Could fisher, jeffreys and neyman have agreed on testing? Statistical Science, 18(1):1–32.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American statistical Association, 112(518):859–877.

Daxberger, E., Nalisnick, E., Allingham, J. U., Antorán, J., and Hernández-Lobato, J. M. (2021b). Bayesian deep learning via subnetwork inference. In International Conference on Machine Learning, pages 2510–2521. PMLR.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of dnns. In Advances in Neural Information Processing Systems.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2023). A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 56(Suppl 1):1513–1589.

Ghosh, S., Yao, J., and Doshi-Velez, F. (2019). Model selection in bayesian neural networks via horseshoe priors. Journal of Machine Learning Research, 20(182):1–46.

Gressmann, F., Eaton-Rosen, Z., and Luschi, C. (2020). Improving neural network training in low dimensional random bases. Advances in Neural Information Processing Systems, 33:12140– 12150.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review, 53(2):217–288.

He, Z., Zheng, Z., and Wang, X. (2023). On the error rate of importance sampling with randomized quasi-monte carlo. SIAM Journal on Numerical Analysis, 61(2):515–538.

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. J. Mach. Learn. Res., 15(1):1593–1623.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. Conference on Uncertainty in Artificial Intelligence

Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., and Wilson, A. G. (2020). Subspace inference for bayesian deep learning. In Uncertainty in Artificial Intelligence, pages 1169–1179. PMLR.

Jiang, W., Kwok, J., and Zhang, Y. (2022). Subspace learning for effective meta-learning. In International Conference on Machine Learning, pages 10177–10194. PMLR.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. Journal of the american statistical association, 90(430):773–795.

Li, C., Farkhoor, H., Liu, R., and Yosinski, J. (2018). Measuring the intrinsic dimension of objective landscapes. arXiv preprint arXiv:1804.08838.

Li, J., Miao, Z., Qiu, Q., and Zhang, R. (2024). Training bayesian neural networks with sparse subspace variational inference. International Conference on Learning Representations.

L'Ecuyer, P. (2018). Randomized quasi-Monte Carlo: An introduction for practitioners. Springer. Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. Advances in neural information processing systems, 32.

Murray, I., Adams, R., and MacKay, D. (2010). Elliptical slice sampling. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pages 541–548. JMLR Workshop and Conference Proceedings

Neal, R. M. (2012). Mcmc using hamiltonian dynamics. arXiv preprint arXiv:1206.1901. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. NIPS Workshop on Deep Learning and Unsupervised Feature Learning.

Owen, A. B. (1997a). Monte carlo variance of scrambled net nds, quadrature. SIAM Journal on Numerical Analysis, 34(5):1884–1910.

Owen, A. B. (1997b). Scrambled net variance for integrals of smooth functions. The Annals of Statistics, 25(4):1541–1562.

Owen, A. B. (2013). Monte Carlo theory, methods and examples. https://artowen.su.domains/ mc/.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016). Deep kernel learning. In Artificial intelligence and statistics, pages 370–378. PMLR.

UCI Regression

Table 3: Bayes factors and testing data evidence ratios on UCI dataset (tail trajectory subspace against block-averaging subspace).

(a) Small UCI Regression Datasets

	boston	concrete	energy	naval	yacht
Bayes factor	0.123 ± 0.031	0.340 ± 0.244	0.214 ± 0.255	0.018 ± 0.020	0.335 ± 0.705
Evidence ratios	0.157 ± 0.052	0.545 ± 0.196	0.291 ± 0.212	0.140 ± 0.112	0.199 ± 0.091

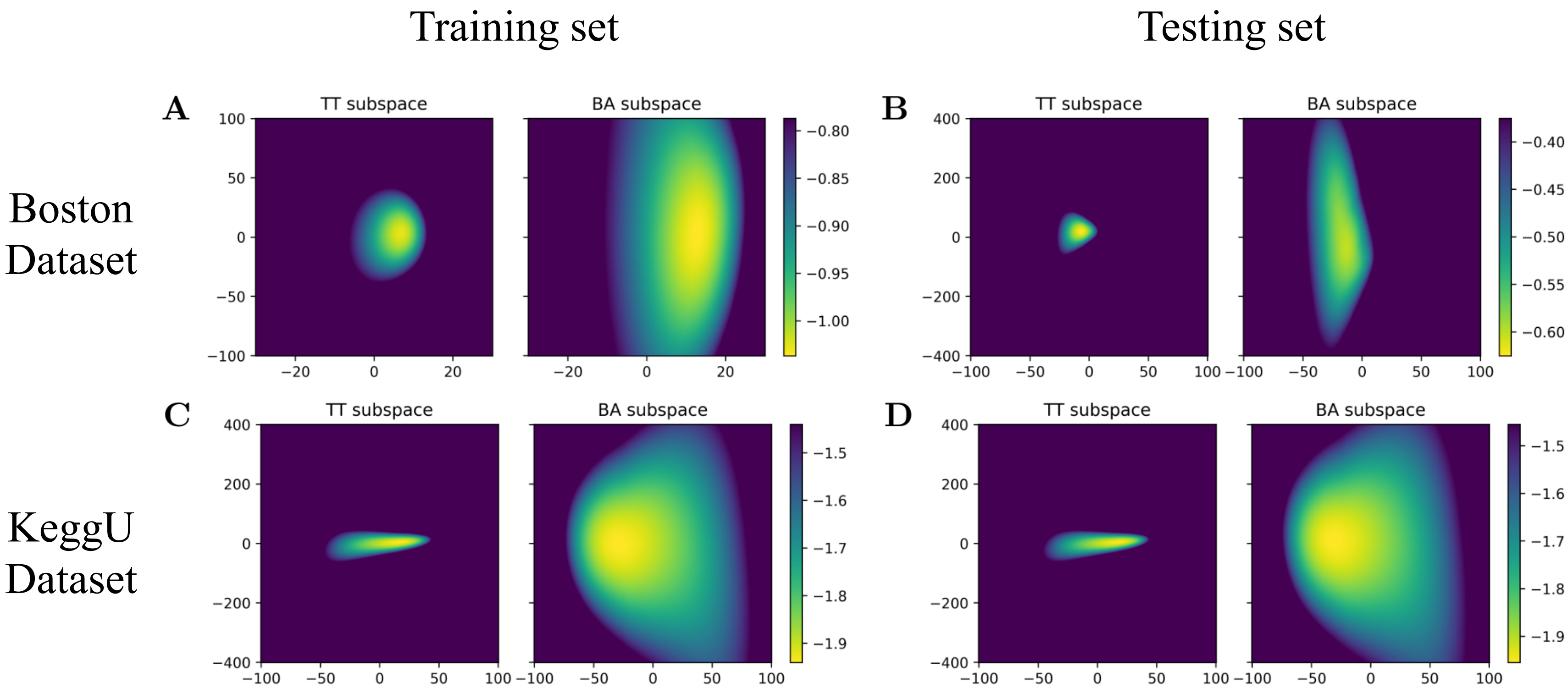
(b) Large UCI Regression Datasets

	elevators	protein	pol	keggD	keggU	skillcraft
Bayes factor	0.215 ± 0.121	0.091 ± 0.096	0.178 ± 0.077	0.269 ± 0.223	0.214 ± 0.252	0.474 ± 0.263
Evidence ratios	0.544 ± 0.184	0.537 ± 0.238	0.205 ± 0.089	0.359 ± 0.257	0.218 ± 0.288	0.608 ± 0.497

- Bayes factor and evidence ratios showing substantial evidence in favor of the BA subspace over the TT subspace.

PURDUE UNIVERSITY

UCI Regression



- The BA subspaces contain more ‘low-loss’ or ‘high-likelihood’ points, reflecting higher subspace quality.

Image Classification

Table 6: Bayes factors and testing data evidence ratios on CIFAR datasets (Tail trajectory subspace against Block-averaging subspace).

Dataset	VGG-16 on `CIFAR10`	PreResNet164 on `CIFAR10`	VGG-16 on `CIFAR100`	PreResNet164 on `CIFAR100`
Bayes factor	0.280 ± 0.031	0.270 ± 0.054	0.227 ± 0.004	0.381 ± 0.026
Evidence ratios	0.193 ± 0.031	0.285 ± 0.122	0.111 ± 0.016	0.353 ± 0.040

Table 7: Classification accuracy (ACC(%)) on CIFAR datasets.

Models	TT (ESS)	BA (ESS)	TT (VI)	BA (VI)	BA (RQMC)
VGG-16 on `CIFAR10`	91.98±0.43	91.92±0.40	91.80±0.42	**92.00±0.44**	91.94±0.51
PreResNet164 on `CIFAR10`	94.99±0.17	95.08±0.11	94.96±0.15	**95.13±0.11**	94.92±0.06
VGG-16 on `CIFAR100`	**68.32±0.47**	68.18±0.42	68.07±0.47	68.17±0.52	**68.33±0.49**
PreResNet164 on `CIFAR100`	76.99±0.03	77.06±0.15	76.94±0.14	77.14±0.27	**77.30±0.35**

- Table 6 show substantial evidence in favor of the BA subspace over the TT subspace, and The BA-based subspace combined with VI and RQMC methods, achieves higher accuracy.