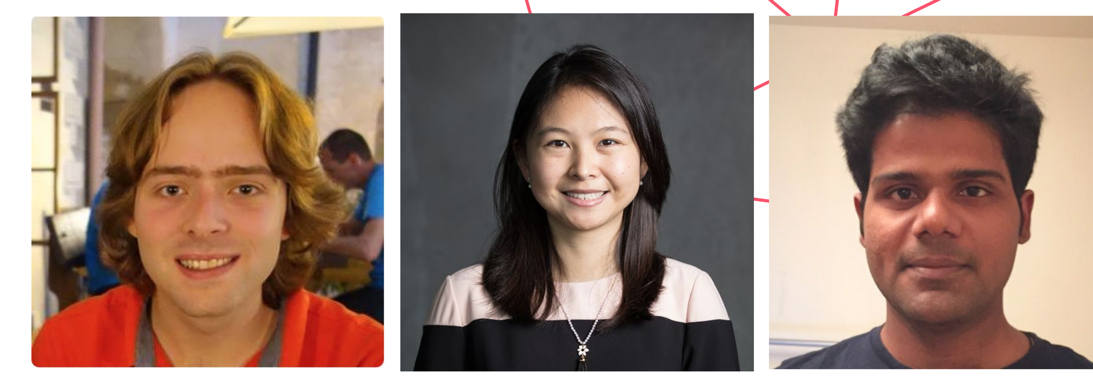
DeepMind

Deep Ensembles: A Loss Landscape Perspective

Stanislav Fort*[2], Huiyi Hu*[1], Balaji Lakshminarayanan[1]

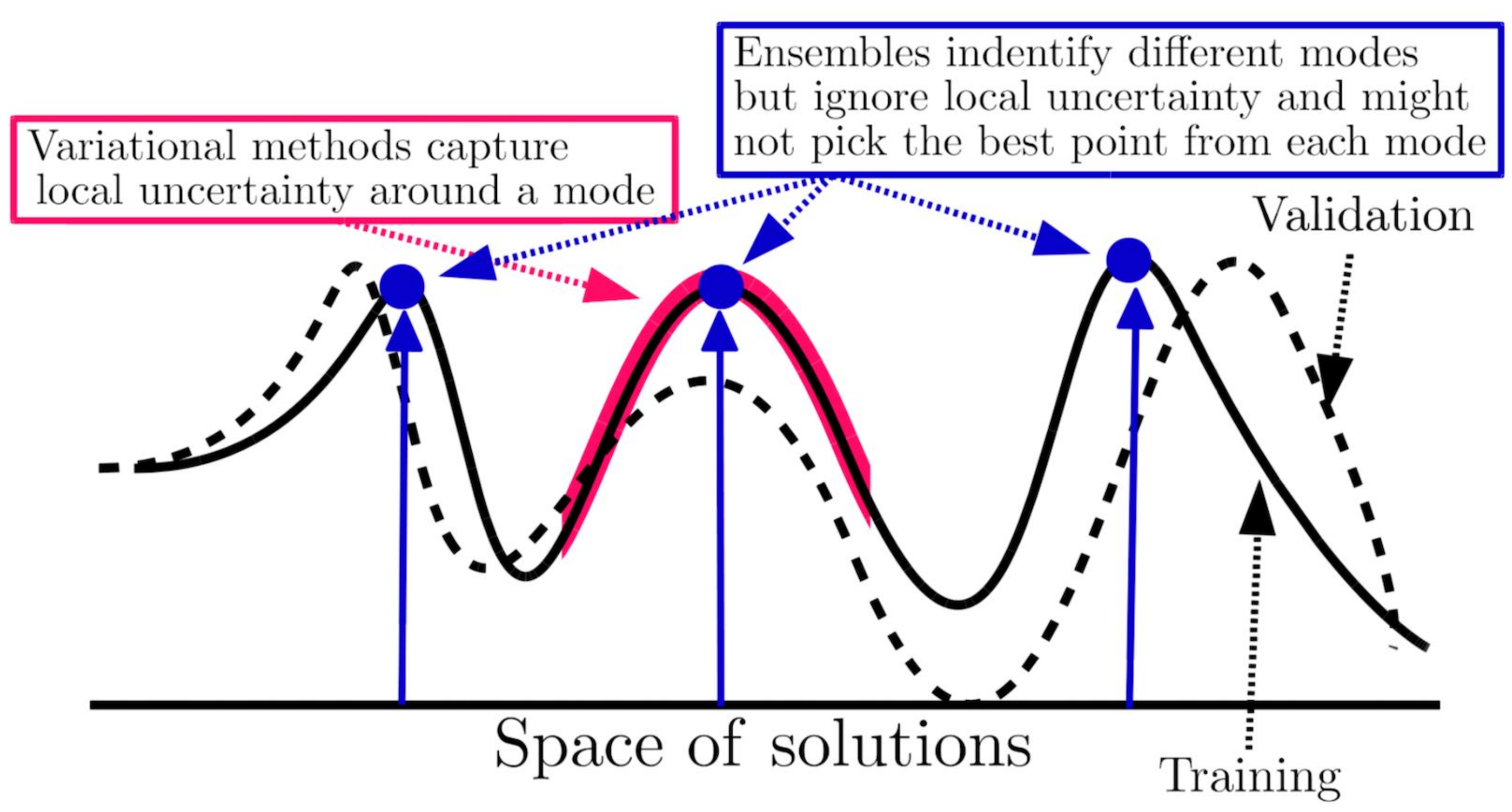
[1]DeepMind, Mountain View [2]Google Research (now at Stanford)



Deep Ensembles vs Bayesian Neural Networks

Question: Why do deep ensembles trained with just random initializations outperform Bayesian neural nets in practice?

Hypothesis: Scalable Bayesian neural nets (based on variational inference) are effective at averaging uncertainty within a single mode, but fail to explore the diversity of multiple modes due to the structure of neural network loss landscape.

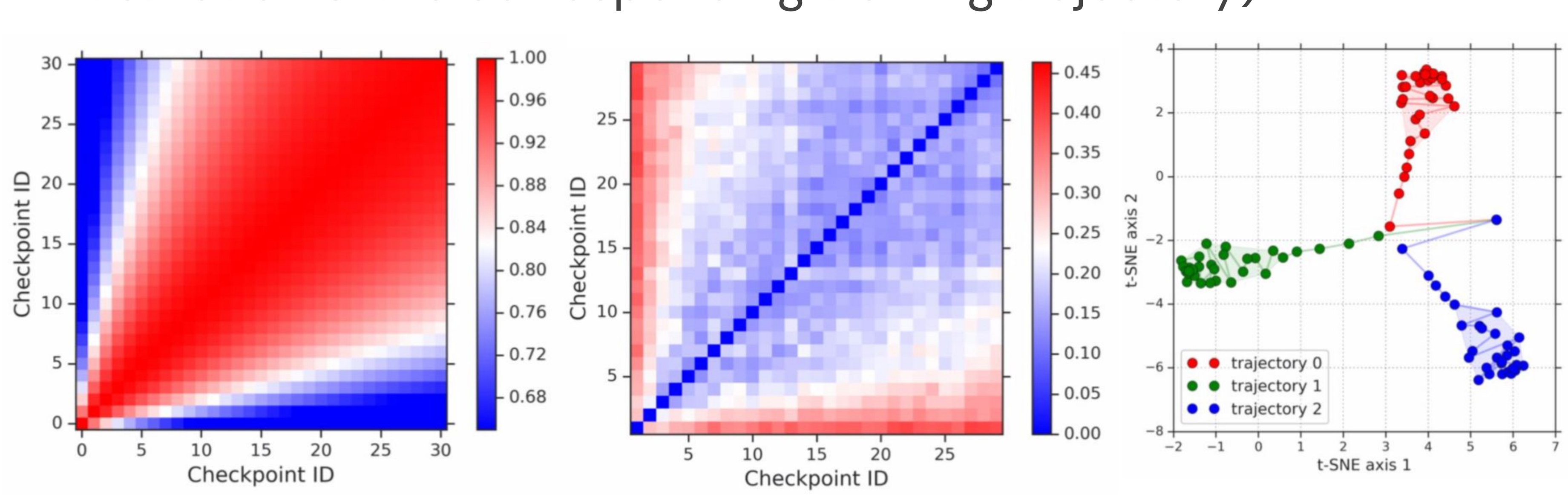


Similarity of functions on CIFAR-10

Setup: Train a simple CNN on CIFAR-10 several times. Evaluate similarity of solutions along trajectory:

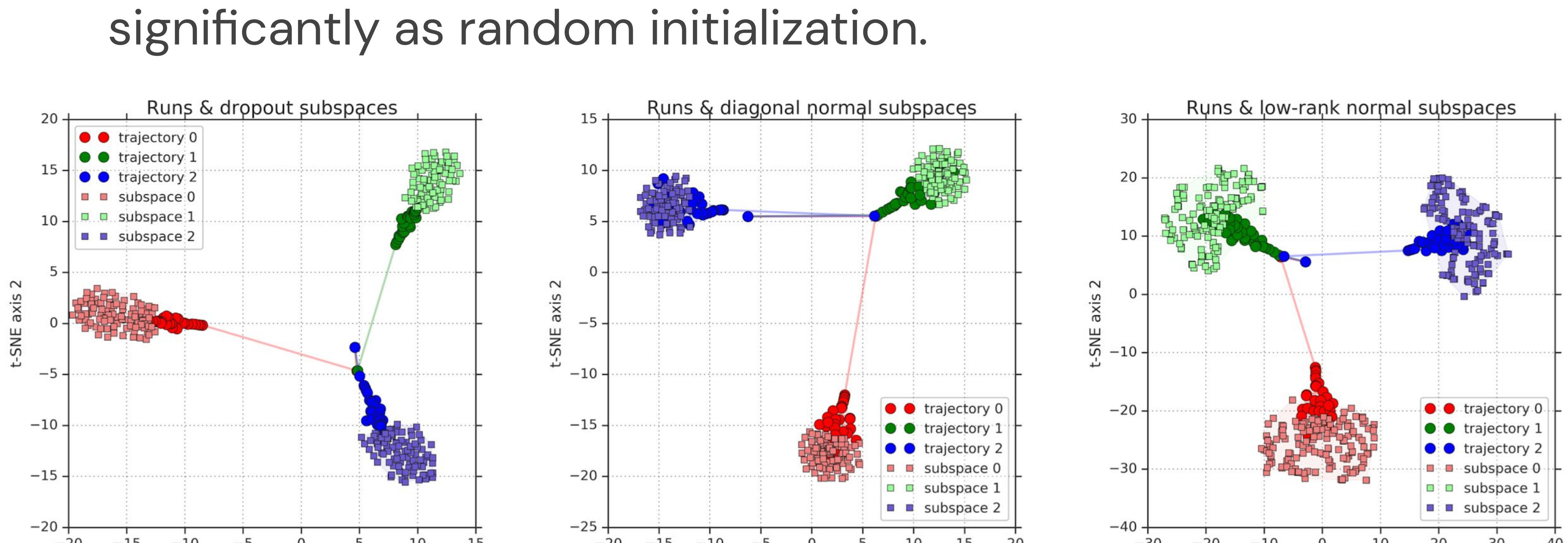
- **Weight space**: cosine similarity
- **Function space**: fraction of images on which class label predictions disagree $\frac{1}{N}\sum_{n=1}^{N}\left(f_{w_1}(x_n) \neq f_{w_2}(x_n)\right)$

t-SNE plots: predictions along training trajectory, as well as multiple random initializations (each color represents an initialization & the corresponding training trajectory)

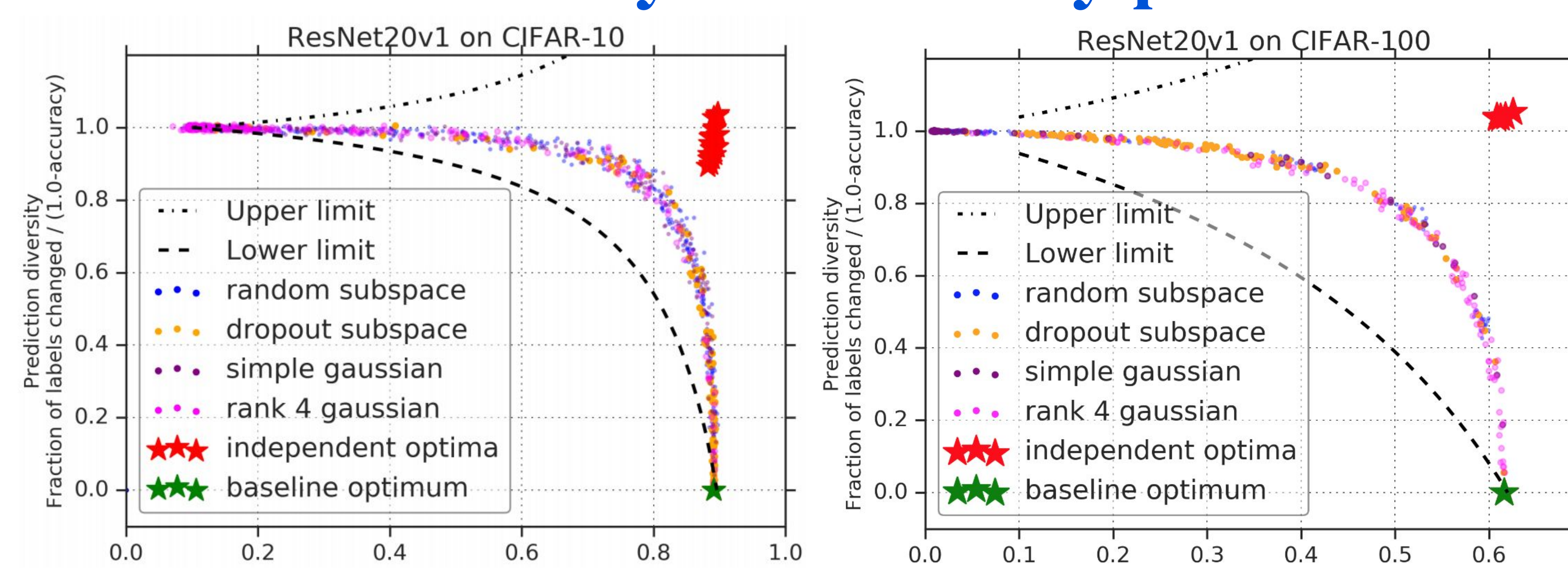


Cosine similarity (weight space) **Disagreement (function space)** **t-SNE of predictions (along trajectory)**

- Checkpoints along the trajectory are very similar functions in terms of both weight space and function space similarity
- **t-SNE plots show that functions along the same trajectory are more similar than functions from different trajectories**
- Subspace sampling methods: Dropout, Gaussian (diagonal, low-rank) increase diversity of functions but not as significantly as random initialization.

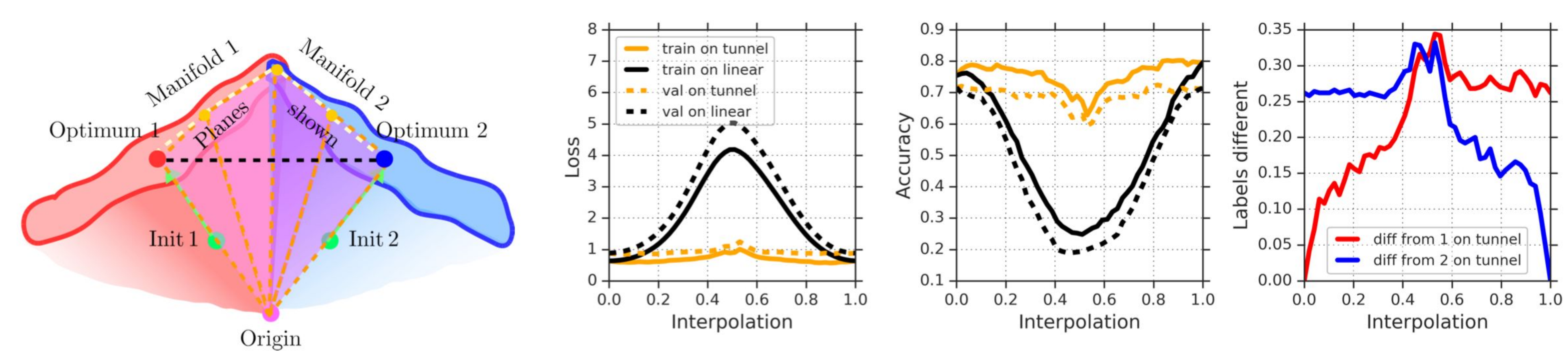


Diversity vs Accuracy plots

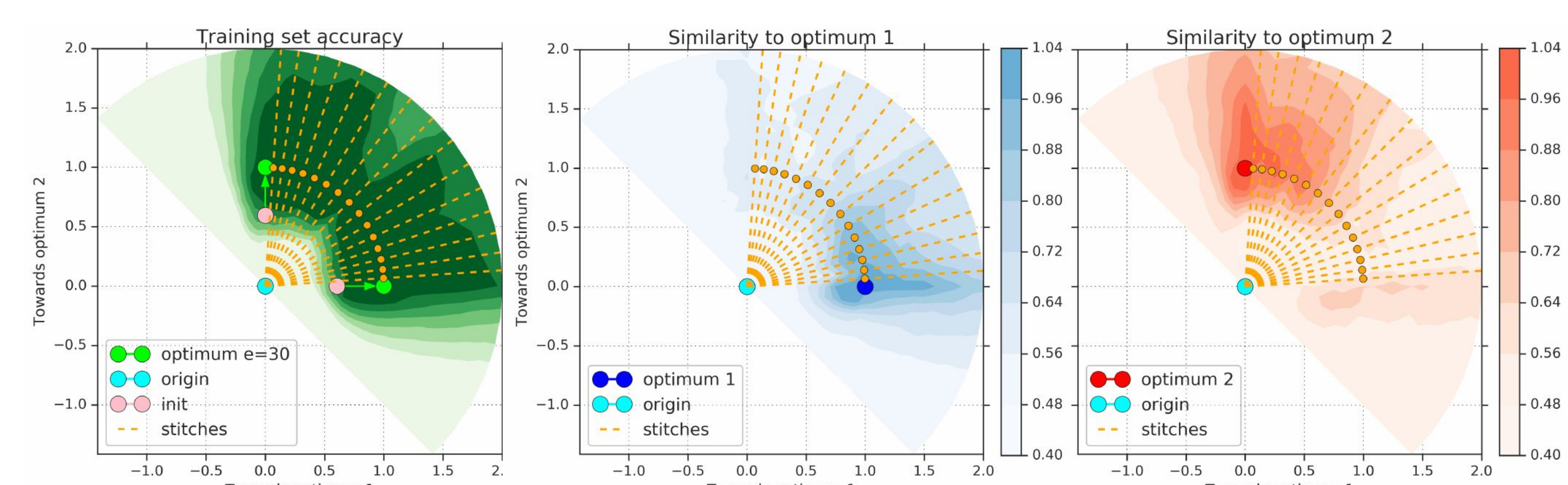


- **Bias-variance perspective:** we care about both accurate solutions (low bias) and diverse solutions (decorrelation reduces variance)
- **Diversity** = function space (i.e. predictions) disagreement normalized by (1-accuracy)
- **Random initializations are much more effective at sampling diverse and accurate solutions.**
- Subspace sampling methods exchange accuracy for diversity suboptimally compared to independent solutions
- Similar trends on other architectures and datasets.

Identical Loss ≠ Identical Functions

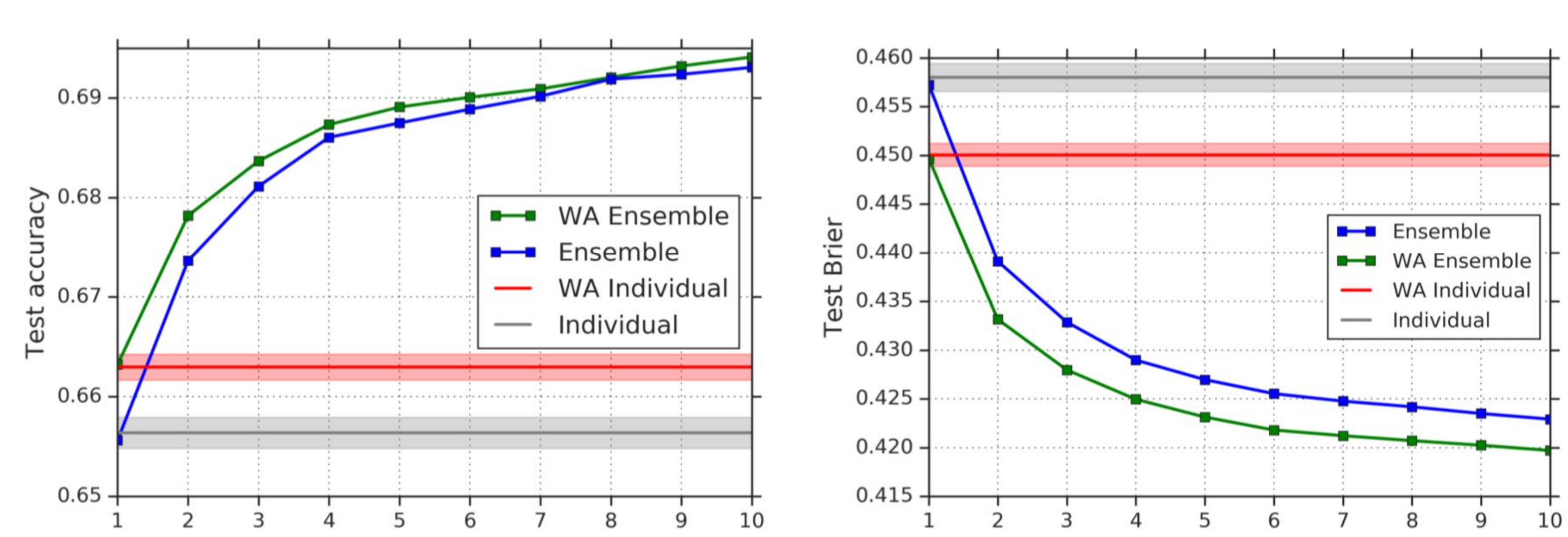


(a) Cartoon illustration (b) Low-loss tunnel



Combining Ensemble and Subspace sampling

- Random initialization leads to multiple modes while subspace methods are effective at averaging uncertainty within a mode
- **Best of both worlds**:
  - Use weight averaging (WA) within a mode to pick the best point (can use variance too, but increases #parameters per mode)
  - Ensemble over multiple random inits
- Results on ImageNet showing the relative benefits of ensemble & WA. Ensembling helps more, particularly under dataset shift.



Paper: https://arxiv.org/abs/1912.02757