Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

Timur Garipov*[1,2], Pavel Izmailov*[3], Dmitrii Podoprikhin*[4], Dmitry Vetrov[5], and Andrew Gordon Wilson[3]

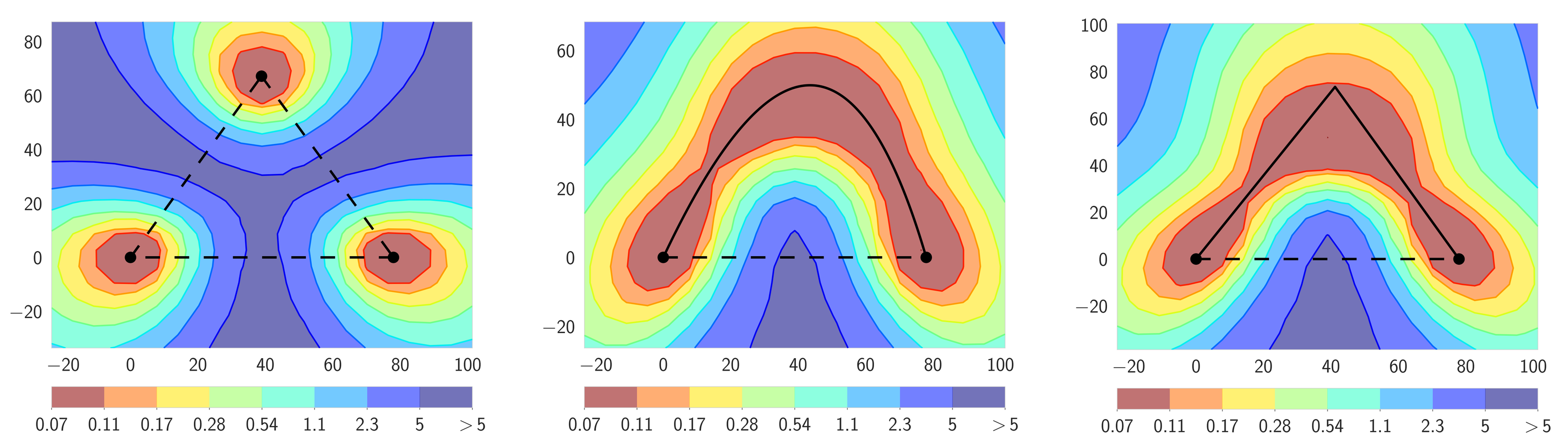
[1]Samsung AI Center in Moscow, [2]Skolkovo Institute of Science and Technology, [3]Cornell University,
[4]Samsung-HSE Laboratory, [5]National Research University Higher School of Economics

Outline

- Local optima for modern deep neural networks are connected by very simple curves of near-constant loss and accuracy.
- We propose a simple method to find such curves.
- Mode connectivity holds for a wide range of architectures and hyper-parameter settings, such as batch size, optimizer, and learning rate schedule.
- Inspired by these observations, we propose Fast Geometric Ensembling (FGE). FGE explores the region of low loss and ensembles multiple models from this region.

Loss Surfaces

2D slices of loss surfaces, CIFAR-100, ResNet-164



- **Left:** Three optima for independently trained networks.
- **Middle** and **Right**: A quadratic Bezier curve, and a polygonal chain with one bend, connecting the lower two optima on the left panel along a path of near-constant loss.

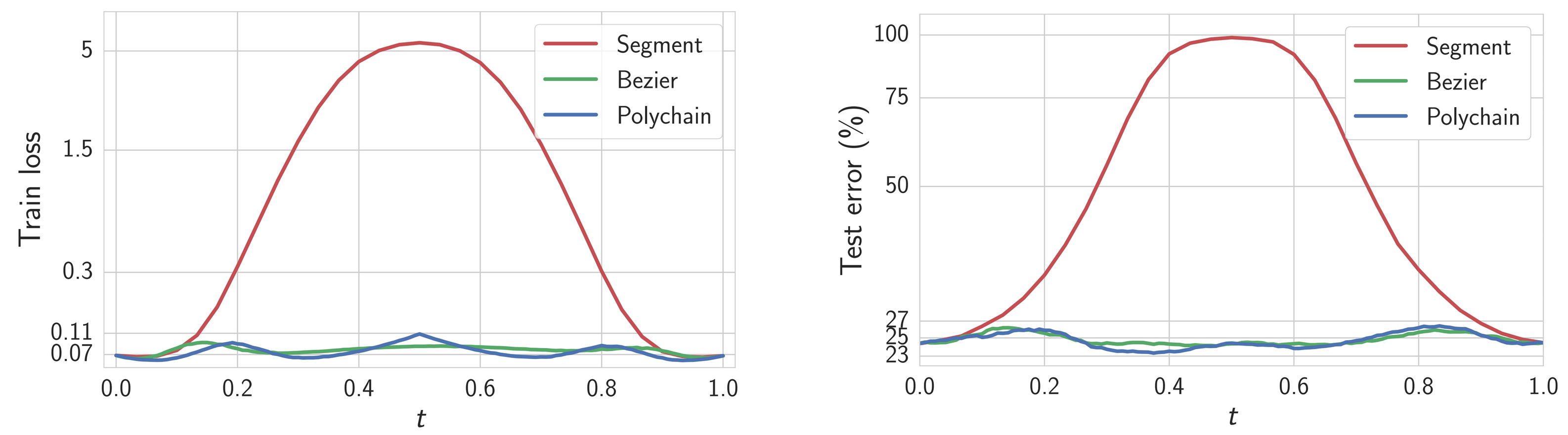
Finding Paths between Modes

- Endpoints: $\widehat{w}_1, \widehat{w}_2 \in \mathbb{R}^{|net|}$, sets of weights of DNNs
- Loss function: $\mathcal{L}(w)$
- Curve parametrization: $\phi_\theta : [0,1] \to \mathbb{R}^{|net|}$ with parameters θ

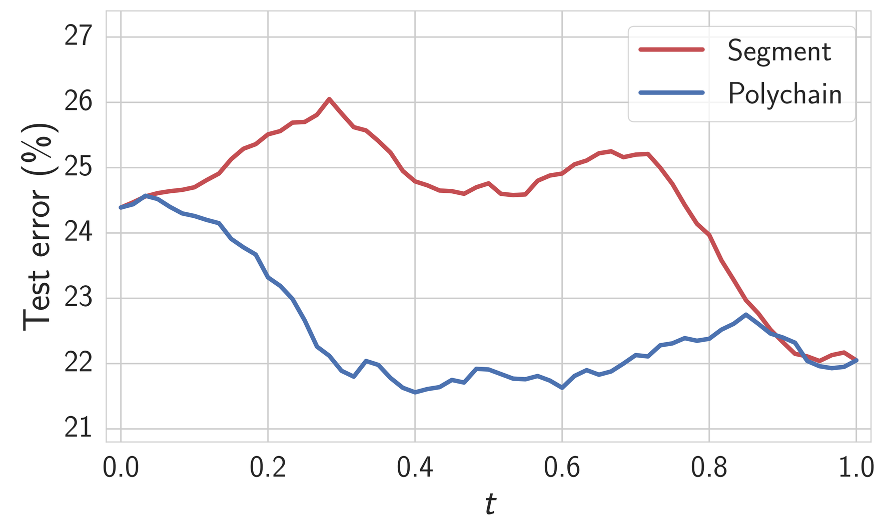
$$\phi_\theta(0) = \widehat{w}_1, \quad \phi_\theta(1) = \widehat{w}_2$$

$$\underset{\theta}{\text{minimize}} \;\; \ell(\theta) = \int_0^1 \mathcal{L}(\phi_\theta(t))dt = \mathbb{E}_{t \sim U(0,1)} \mathcal{L}(\phi_\theta(t)).$$

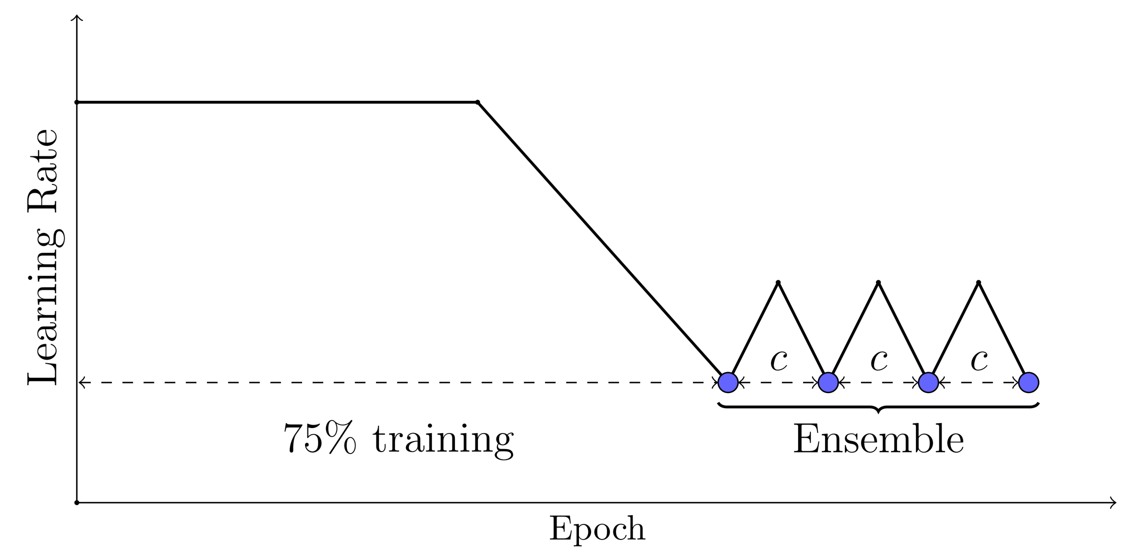
Curve Evaluation

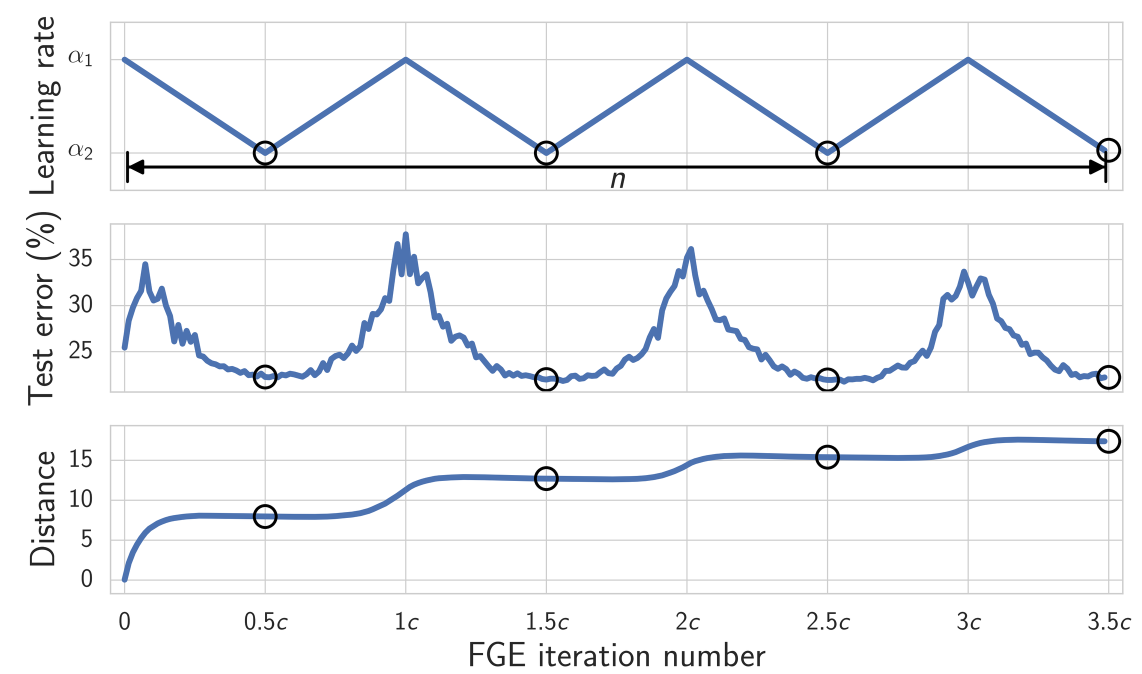


Ensembling along the curve

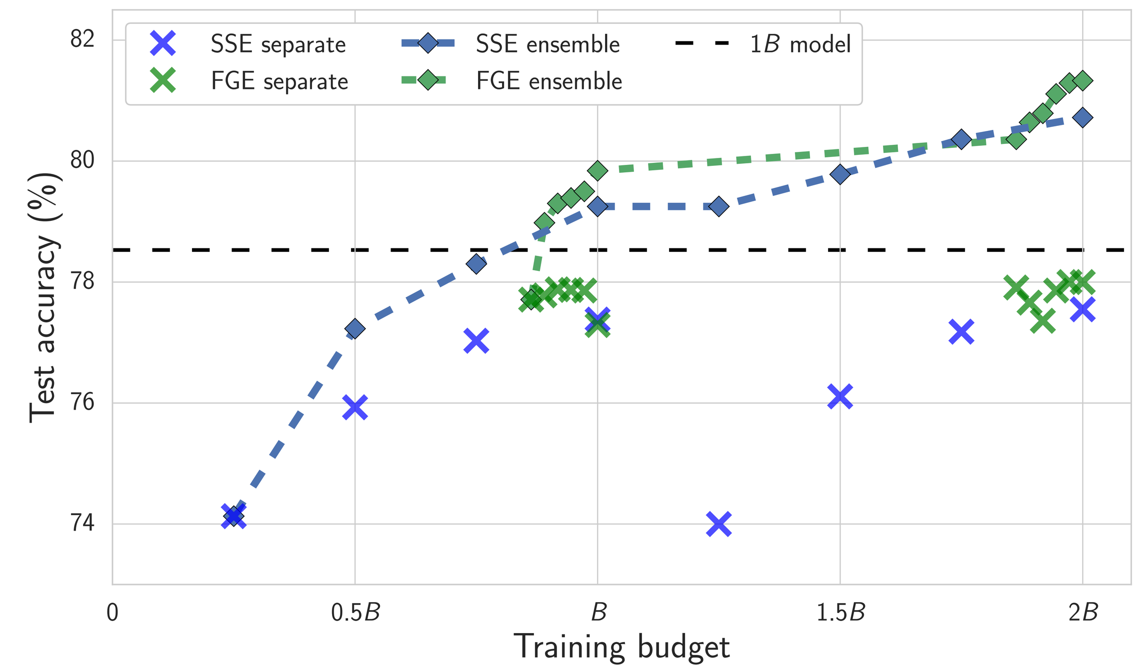


Fast Geometric Ensembling





Ensembling Results



Error rates (%) on CIFAR-100 and CIFAR-10 datasets for different ensembling methods and training budgets.

DNN (Budget)	method	CIFAR-100 $1B$	CIFAR-100 $2B$	CIFAR-100 $3B$	CIFAR-10 $1B$	CIFAR-10 $2B$	CIFAR-10 $3B$
VGG-16 (200)	Ind	27.4	25.28	24.45	6.81	5.89	5.9
	SSE	26.4	25.16	24.69	6.5	6.19	5.95
	FGE	**25.74**	**24.11**	**23.54**	**6.48**	**5.82**	**5.66**
ResNet-110 (150)	Ind	21.37	19.04	18.59	4.7	**4.1**	**3.77**
	SSE	20.75	19.28	18.91	4.66	4.37	4.3
	FGE	**20.16**	**18.67**	**18.21**	**4.55**	4.21	3.98
WRN-28-10 (200)	Ind	19.1	17.48	17.01	3.74	3.4	**3.31**
	SSE	17.78	17.3	16.97	3.74	3.54	3.55
	FGE	**17.73**	**16.95**	**16.88**	**3.64**	**3.38**	3.52

ImageNet ResNet-50:

- We run FGE starting from a pretrained model.
- Form ensemble of 4 models in only 5 epochs.
- Achieve 0.56% improvement of top-1 error-rate.

Discussion

- New posterior approximation families for Bayesian deep learning.
- Geometric insights in this paper could be used to accelerate the convergence, stability and accuracy of optimization procedures.

Code

github.com/timgaripov/dnn-mode-connectivity