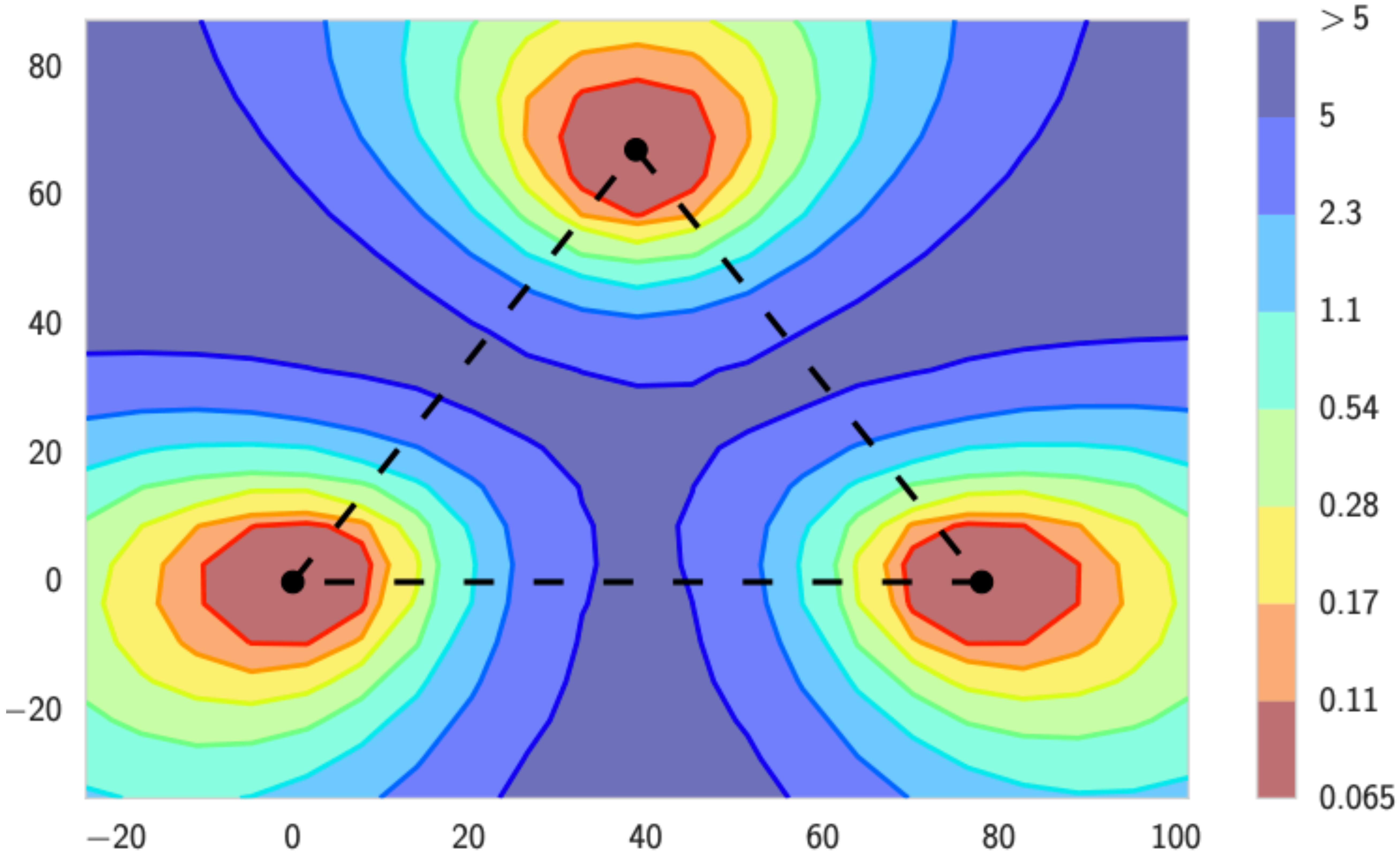
# Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

Prepared by the student of group 192:
Pozdeev Dmitrii Mikhailovich

# Introduction

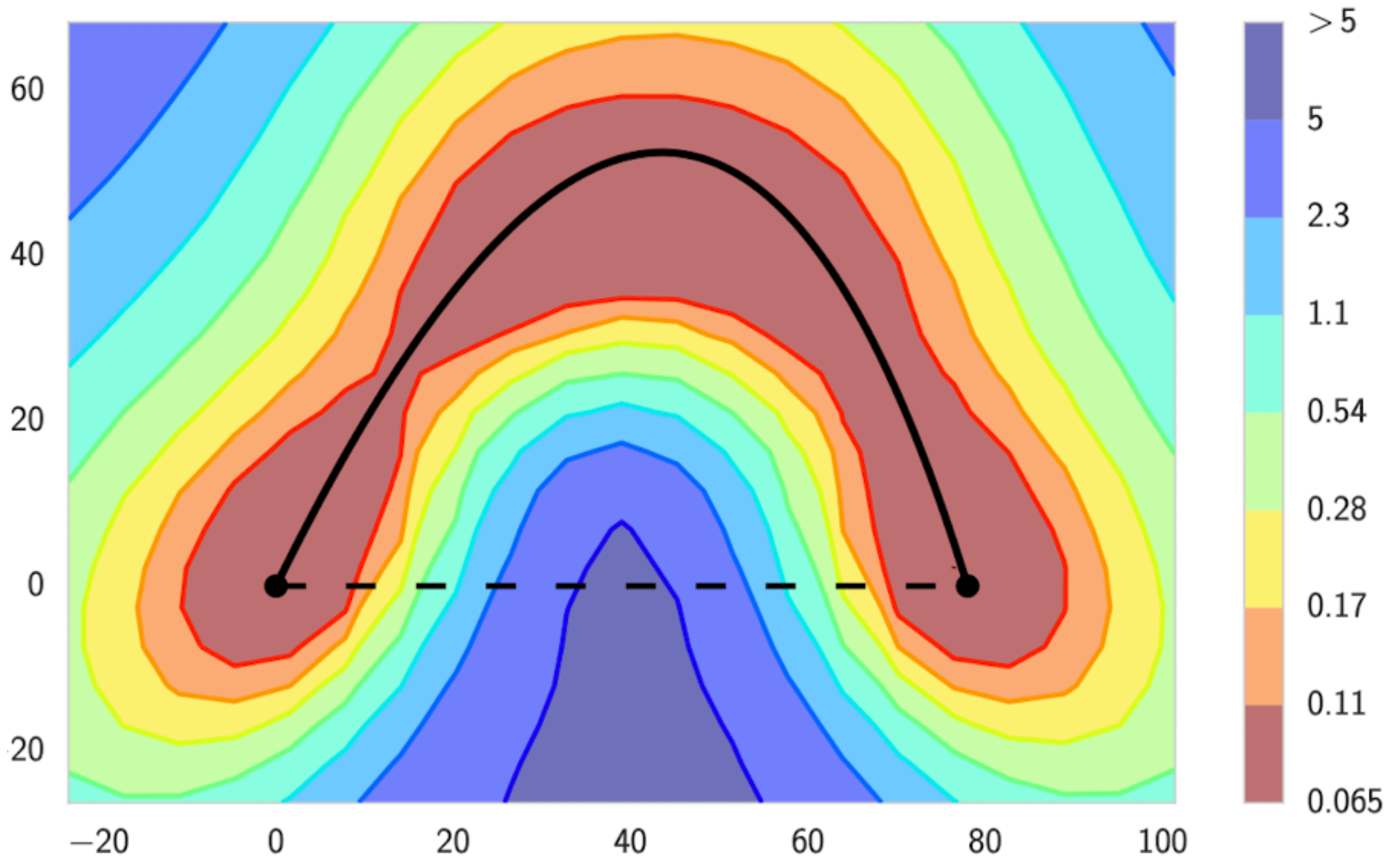
- Loss Surfaces are complicated
- Loss is high along a segment connecting two models




[https://arxiv.org/pdf/1802.10026.pdf]

# Plan

- Mode Connectivity
- Fast Geometric Ensembling
- Recent Work




[https://arxiv.org/pdf/1802.10026.pdf]

# Mode connectivity

Let $\hat{w}_1, \hat{w}_2 \in \mathbb{R}^{|net|}$ - two independently trained networks.

We want to find a path $\phi_\theta(t)$: $\phi_\theta(0) = \hat{w}_1$, $\phi_\theta(1) = \hat{w}_2$

$$\phi_\theta(t) = \begin{cases} 2\left(t\theta + (0.5 - t)\hat{w}_1\right), & 0 \le t \le 0.5 \\ 2\left((t - 0.5)\hat{w}_2 + (1 - t)\theta\right), & 0.5 \le t \le 1. \end{cases}$$

Polygonal chain

$$\phi_\theta(t) = (1 - t)^2 \hat{w}_1 + 2t(1 - t)\theta + t^2 \hat{w}_2, \;\; 0 \le t \le 1.$$

Bezier Curve

# Connection procedure

$$\hat{\ell}(\theta) = \frac{\int \mathcal{L}(\phi_\theta) d\phi_\theta}{\int d\phi_\theta} = \frac{\int\limits_0^1 \mathcal{L}(\phi_\theta(t)) \|\phi'_\theta(t)\| dt}{\int\limits_0^1 \|\phi'_\theta(t)\| dt} = \int\limits_0^1 \mathcal{L}(\phi_\theta(t)) q_\theta(t) dt = \mathbb{E}_{t \sim q_\theta(t)} \Big[ \mathcal{L}(\phi_\theta(t)) \Big], \quad (1)$$

Fair Loss Formula

where the distribution $q_\theta(t)$ on $t \in [0,1]$ is defined as: $q_\theta(t) = \|\phi'_\theta(t)\| \cdot \left( \int\limits_0^1 \|\phi'_\theta(t)\| dt \right)^{-1}$

On practice Fair Loss loss is intractable.

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

Our Loss

# Connection procedure

**Require:**

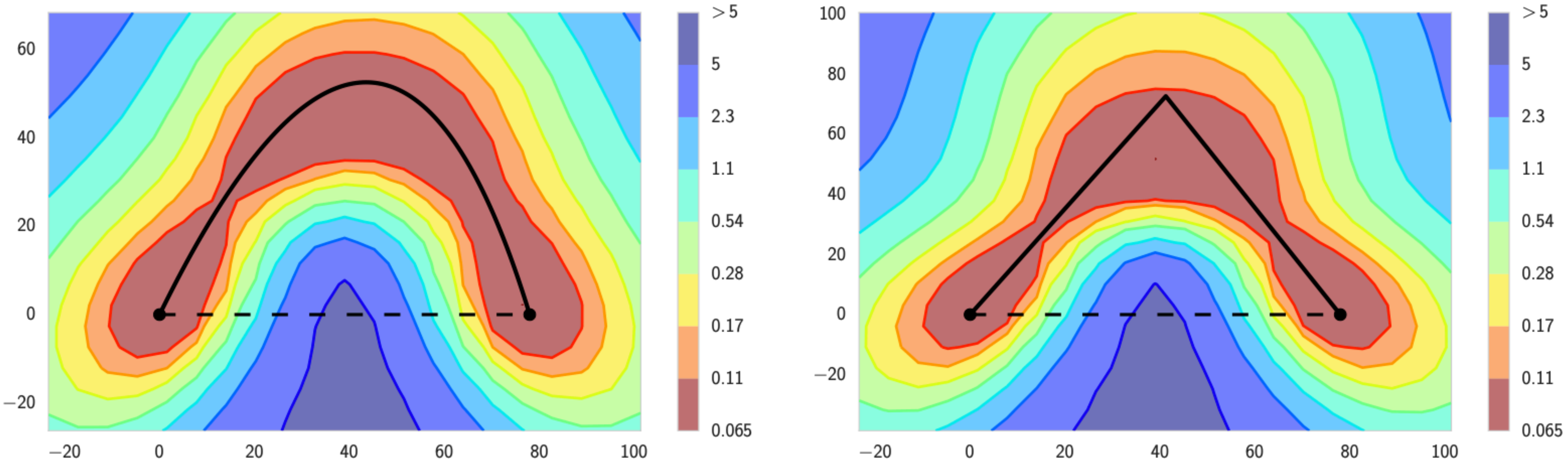
weights $\hat{w}_1, \hat{w}_2 \in \mathbb{R}^{|net|}$

**While** not converge:

Sample $\hat{t} \sim U(0,1)$

Make gradient step for $\theta$ with respect to the $L(\phi_\theta(\hat{t}))$

# Results




**Left**: Bezier Curve, **Right:** Polygonal chain
[https://arxiv.org/pdf/1802.10026.pdf]

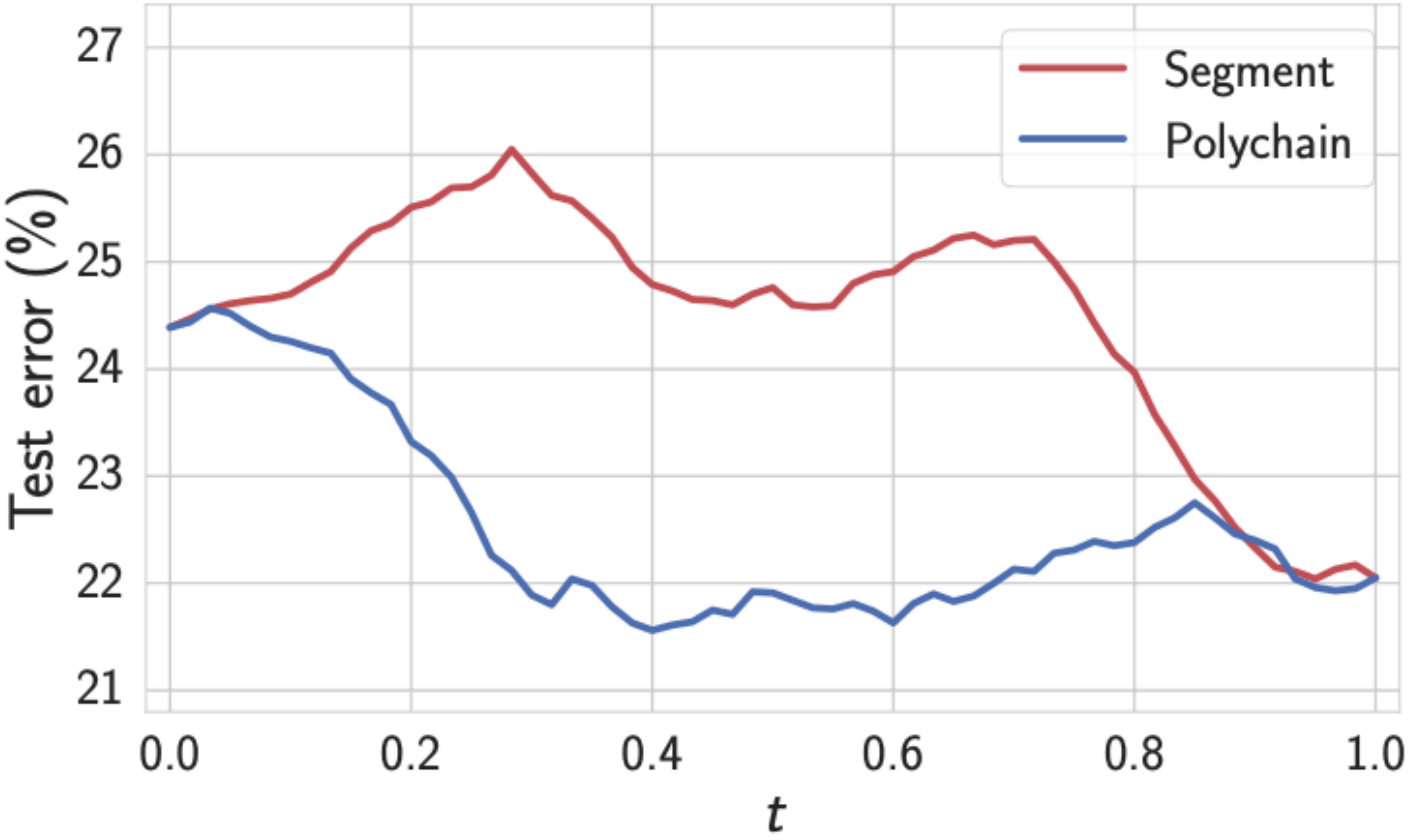
# Ensembles

**Ensemble learning** - combines several individual models to obtain better performance.

**Intuition:** diverse models form an efficient ensemble.

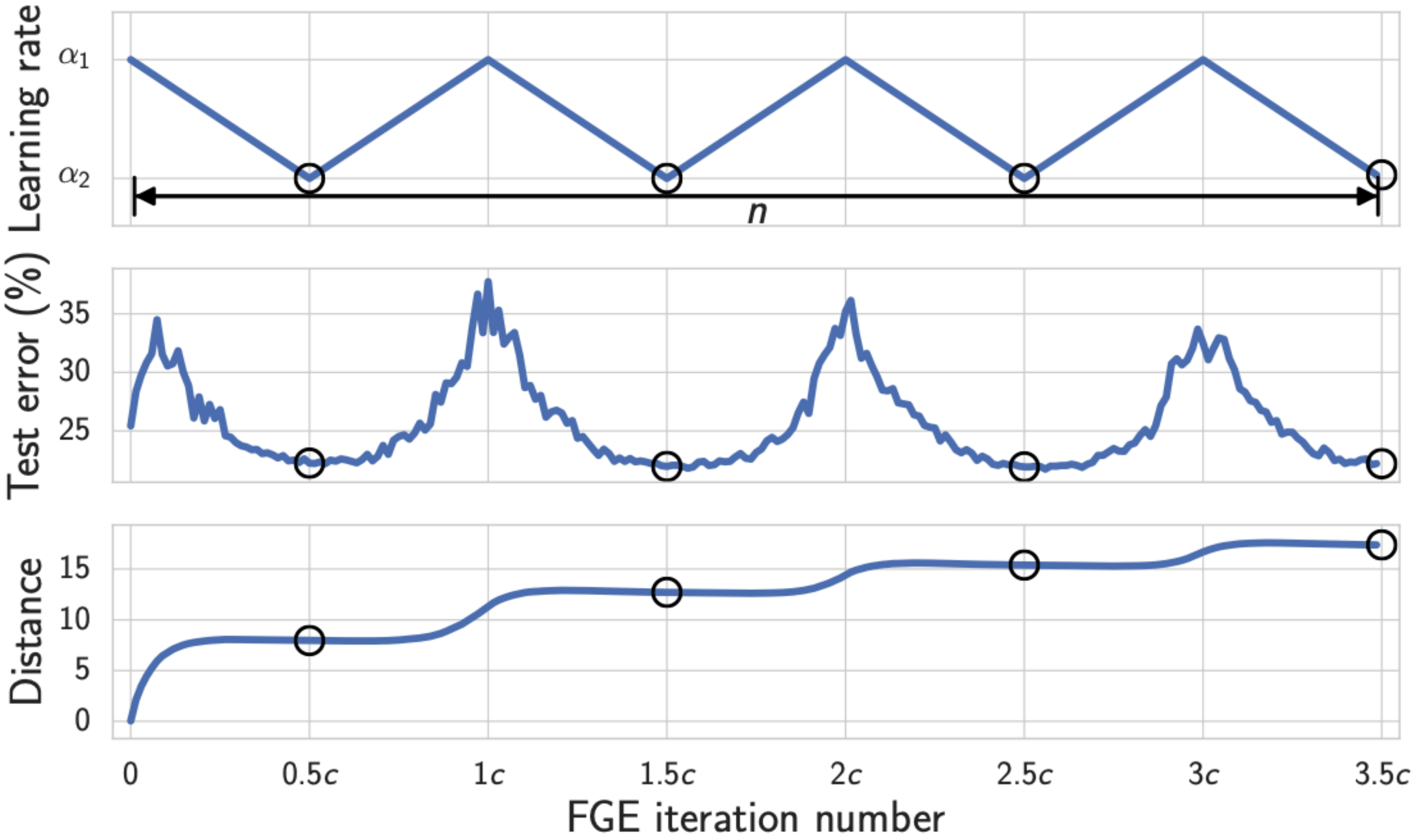
**Independent ensembles:** combine independently trained networks (from different random initialisations)

# Intuition behind FGE




[https://arxiv.org/pdf/1802.10026.pdf]

# FGE Learning rate




[https://arxiv.org/pdf/1802.10026.pdf]

# FGE Algorithm

---

**Algorithm 1** Fast Geometric Ensembling

---

**Require:**
  weights $\hat{w}$, LR bounds $\alpha_1, \alpha_2$,
  cycle length $c$ (even), number of iterations $n$
**Ensure:** ensemble
  $w \leftarrow \hat{w}$ {Initialize weight with $\hat{w}$}
  ensemble $\leftarrow$ [ ]
  **for** $i \leftarrow 1, 2, \ldots, n$ **do**
    $\alpha \leftarrow \alpha(i)$ {Calculate LR for the iteration}
    $w \leftarrow w - \alpha \nabla \mathcal{L}_i(w)$ {Stochastic gradient update}
    **if** $\mathrm{mod}(i, c) = c/2$ **then**
      ensemble $\leftarrow$ ensemble $+ [w]$ {Collect weights}
    **end if**
  **end for**

---

[https://arxiv.org/pdf/1802.10026.pdf]

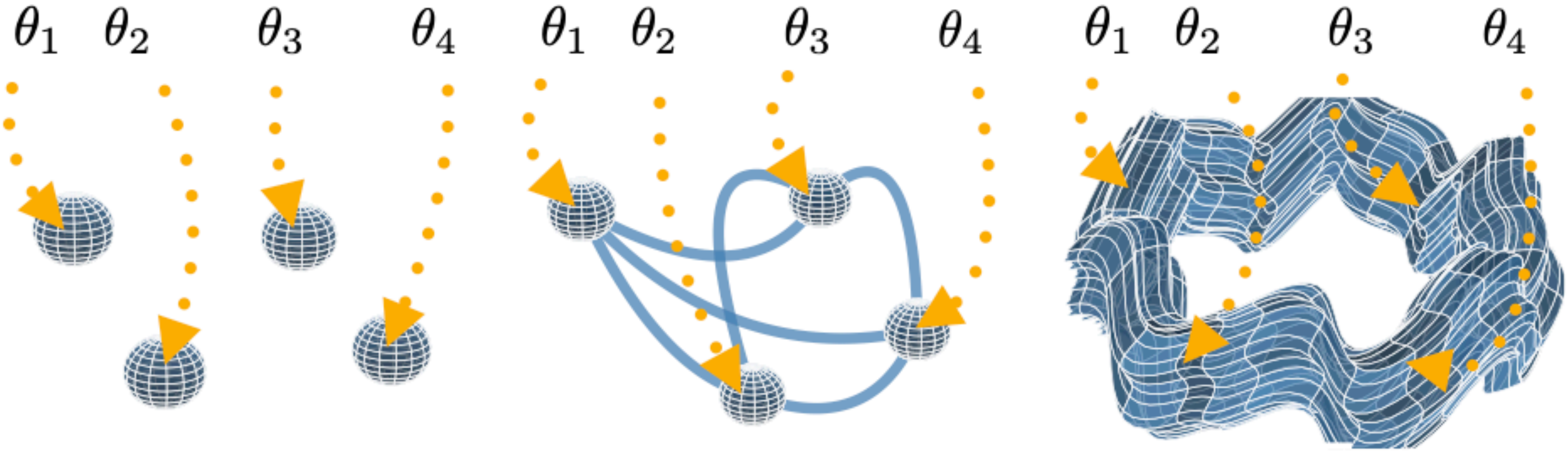
# FGE Results

Table 1: Error rates (%) on CIFAR-100 and CIFAR-10 datasets for different ensembling techniques and training budgets. The best results for each dataset, architecture, and budget are **bolded**.

| 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 | $27.4 \pm 0.1$ | 25.28 | 24.45 | $6.75 \pm 0.16$ | 5.89 | 5.9 |
| | SSE | $26.4 \pm 0.1$ | 25.16 | 24.69 | $6.57 \pm 0.12$ | 6.19 | 5.95 |
| | FGE | $\mathbf{25.7 \pm 0.1}$ | **24.11** | **23.54** | $\mathbf{6.48 \pm 0.09}$ | **5.82** | **5.66** |
| ResNet-164 (150) | Ind | $21.5 \pm 0.4$ | 19.04 | 18.59 | $4.72 \pm 0.1$ | **4.1** | **3.77** |
| | SSE | $20.9 \pm 0.2$ | 19.28 | 18.91 | $4.66 \pm 0.02$ | 4.37 | 4.3 |
| | FGE | $\mathbf{20.2 \pm 0.1}$ | **18.67** | **18.21** | $\mathbf{4.54 \pm 0.05}$ | 4.21 | 3.98 |
| WRN-28-10 (200) | Ind | $19.2 \pm 0.2$ | 17.48 | 17.01 | $3.82 \pm 0.1$ | 3.4 | **3.31** |
| | SSE | $17.9 \pm 0.2$ | 17.3 | 16.97 | $3.73 \pm 0.04$ | 3.54 | 3.55 |
| | FGE | $\mathbf{17.7 \pm 0.2}$ | **16.95** | **16.88** | $\mathbf{3.65 \pm 0.1}$ | **3.38** | 3.52 |

[https://arxiv.org/pdf/1802.10026.pdf]

# Recent Works




[https://arxiv.org/pdf/2102.13042.pdf]

# Recent Work

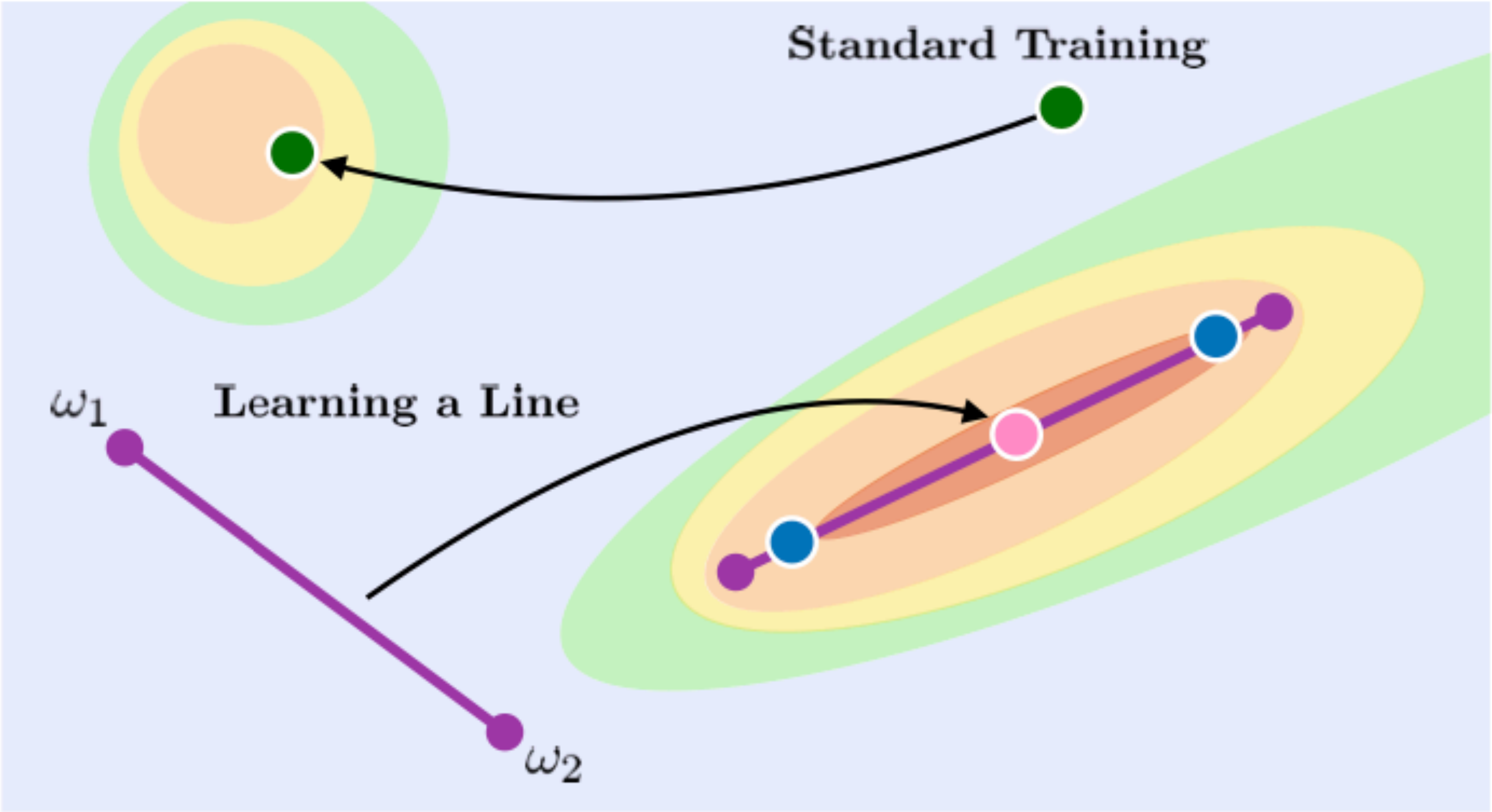



*Figure 1.* Schematic for learning a line of neural networks compared with standard training. The midpoint outperforms standard training in terms of accuracy, calibration, and robustness. Models near the endpoints enable high-accuracy ensembles in a single training run.

[https://arxiv.org/pdf/2102.10472.pdf]

# Conclusions

- Independent networks are connected by very simple curves
- There are methods that find such paths
- Using this insight we can build Fast Geometric Ensemble, which outperforms ensemble of independent models (if computational budget is fixed)

# References


- https://arxiv.org/pdf/1802.10026.pdf (main)
- https://arxiv.org/pdf/2102.13042.pdf
- https://arxiv.org/pdf/2102.10472.pdf