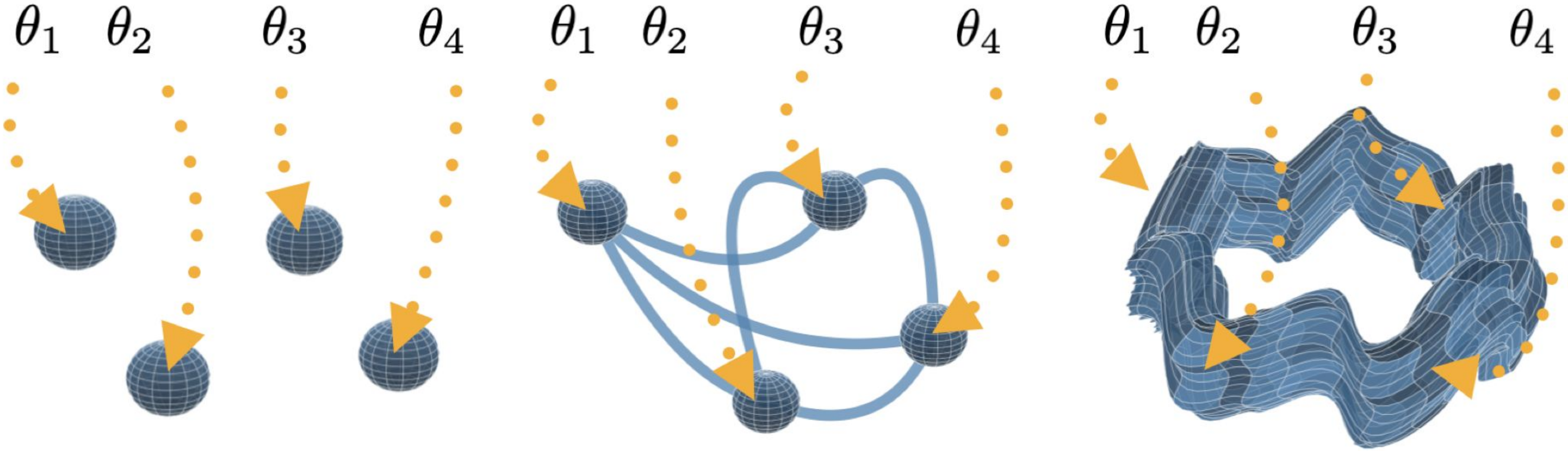
Loss Surface Simplexes for Mode Connecting Volumes and Fast Ensembling

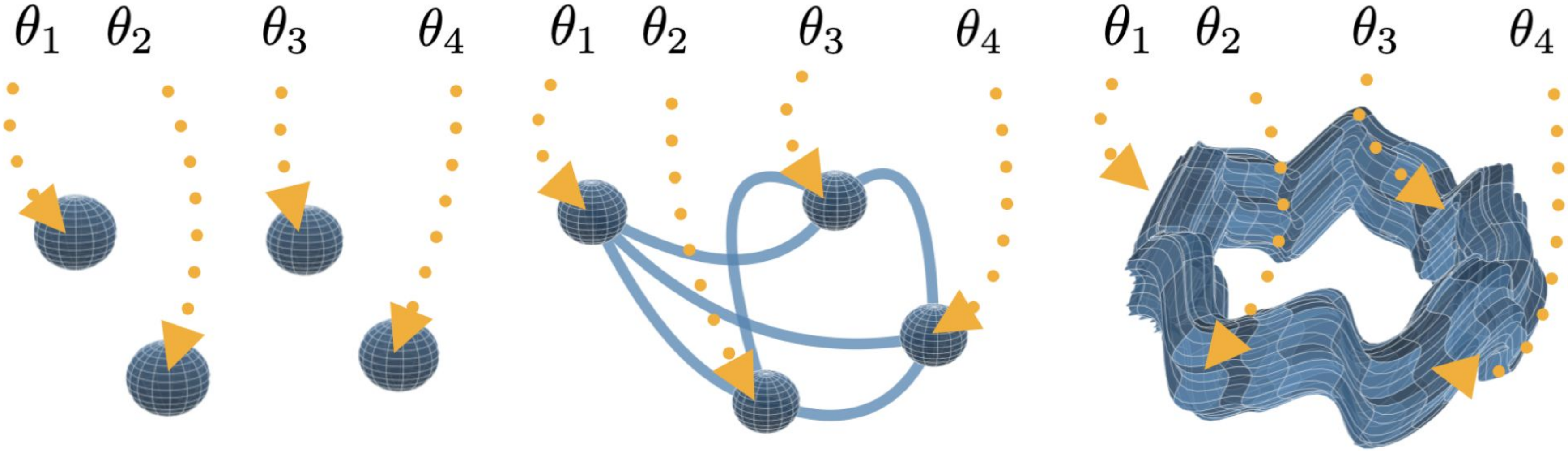
Greg Benton, Wesley Maddox, Sanae Lotfi, Andrew Gordon Wilson

The Structure of Loss Surfaces



Independent, distinct *modes* in parameter space (Choromanska et al, '15)

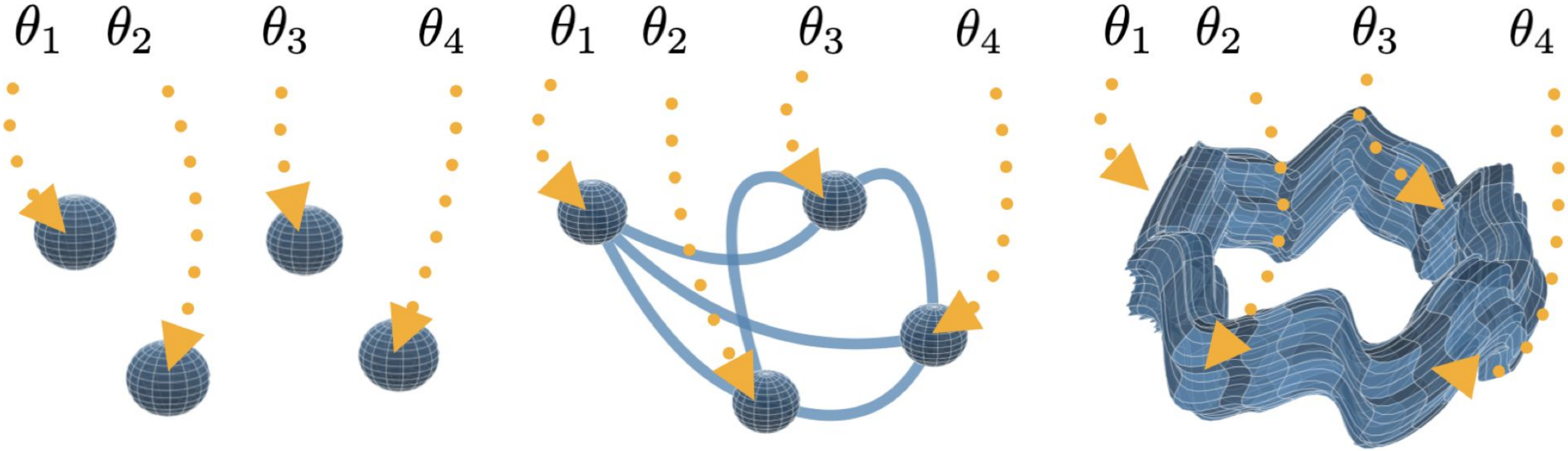
The Structure of Loss Surfaces



Independent, distinct *modes* in parameter space (Choromanska et al, ‘15)

Modes are connected along tunnels of low loss (Garipov et al, ‘18)

The Structure of Loss Surfaces

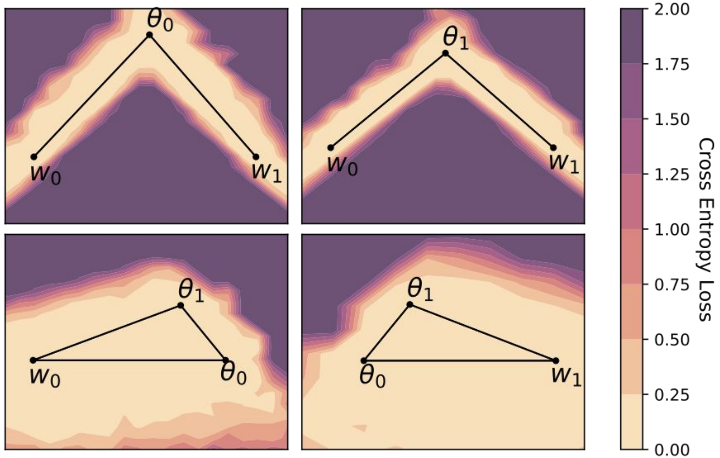


Independent, distinct *modes* in parameter space (Choromanska et al, ‘15)

Modes are connected along tunnels of low loss (Garipov et al, ‘18)

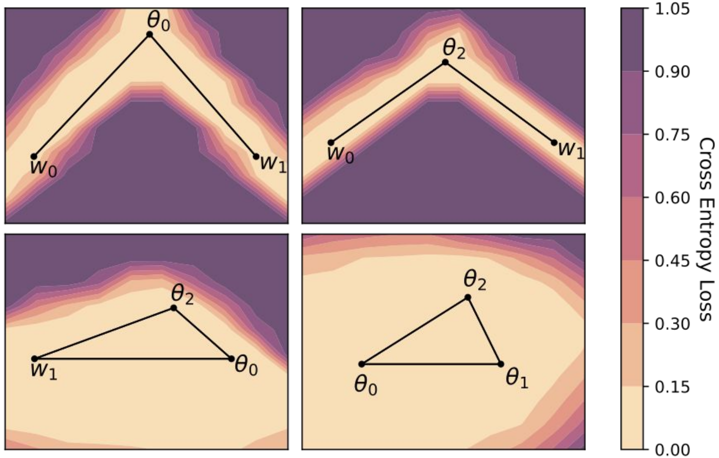
Modes are connected along volumes of low loss (this work; Wortsmann et al, ICML, ‘21))

Expanding Mode Connecting Paths



- Rather than mode connecting paths, learn mode connecting spaces
  - Form *simplicial complexes* in parameter space
  - Add multiple connecting points such that within each simplex of points we have low loss
  - Here we have two modes (w_i's) connected through **two** shared connecting points (θ_i's)

Expanding Mode Connecting Paths



- Rather than mode connecting paths, learn mode connecting spaces:
  - Form *simplicial complexes* in parameter space
  - Add multiple connecting points such that within each simplex of points we have low loss
  - Here we have two modes (w_i's) connected through **three** shared connecting points (θ_i's)

Expanding Mode Connecting Paths

- Finding connecting points is easy! Use a regularized loss function:

$$\mathscr{L}_{reg}(\mathscr{K}) = \boxed{\frac{1}{M}\sum_{\phi_m \sim \mathscr{K}} \mathscr{L}(\mathscr{D}, \phi_m)} - \lambda_j \log(\mathrm{V}(\mathscr{K}))$$

  - First term: Loss over the simplicial complex, where ϕ_m are sampled uniformly from simplicial complex $\mathcal{K}$

$$\boxed{\frac{1}{M}\sum_{m=1}^{M} \mathscr{L}(\mathscr{D}, \phi_m)} \approx \mathbb{E}_{\phi \sim \mathscr{K}} \mathscr{L}(\mathscr{D}, \phi)$$

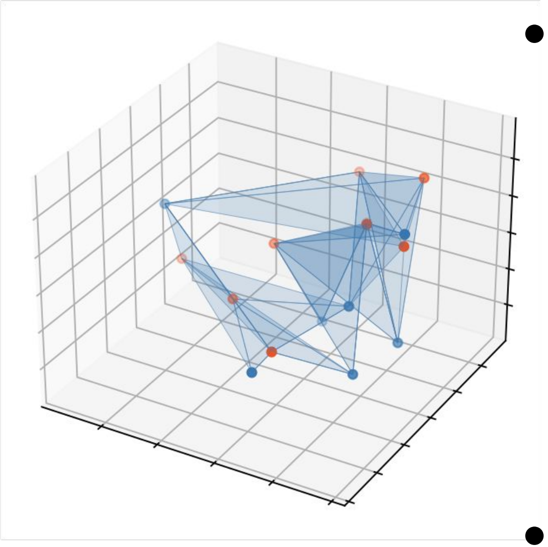
Expanding Mode Connecting Paths

- Finding connecting points is easy! Use a regularized loss function:

$$\mathcal{L}_{reg}(\mathcal{K}) = \frac{1}{M} \sum_{\phi_m \sim \mathcal{K}} \mathcal{L}(\mathcal{D}, \phi_m) - \boxed{\lambda_j \log(\mathrm{V}(\mathcal{K}))}$$

  - Second term: Regularizer attempts to maximize the volume of the simplicial complex in parameter space
    - $\boxed{\lambda_j}$ is the regularization penalty in finding the j^{th} vertex

Expanding Mode Connecting Paths



- Finding connecting points is easy! Use a regularized loss function:

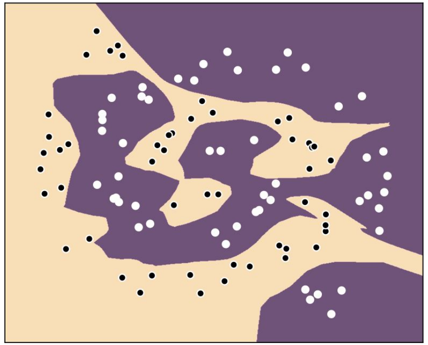
$$\mathcal{L}_{reg}(\mathcal{K}) = \frac{1}{M} \sum_{\phi_m \sim \mathcal{K}} \mathcal{L}(\mathcal{D}, \phi_m) - \lambda_j \log(\mathrm{V}(\mathcal{K}))$$

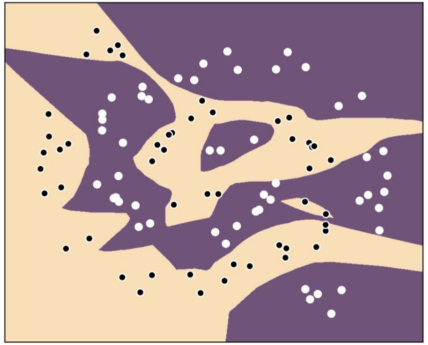
  - First term: Loss over the simplicial complex, where $\boldsymbol{\phi}_m$ are sampled uniformly from simplicial complex $\mathcal{K}$
  - Second term: Regularizer attempts to maximize the volume of the simplicial complex in parameter space
- Method is: *Simplicial Pointwise Random Optimization* (SPRO)

Ensembling In Low Loss Simplexes

- Using the same loss function we can turn SPRO into *Ensembled SPRO* or ESPRO
  1. Train a standard model (i.e. a 0-simplex)
  2. Add a new simplex vertex and train with the regularized loss function; here $\mathcal{K}$ is a single simplex (not a complex)

     $$\mathcal{L}_{reg}(\mathcal{K}) = \frac{1}{M} \sum_{\phi_m \sim \mathcal{K}} \mathcal{L}(\mathcal{D}, \phi_m) - \lambda_j \log(\mathsf{V}(\mathcal{K}))$$
  3. Fix the trained simplex vertex
  4. Repeat





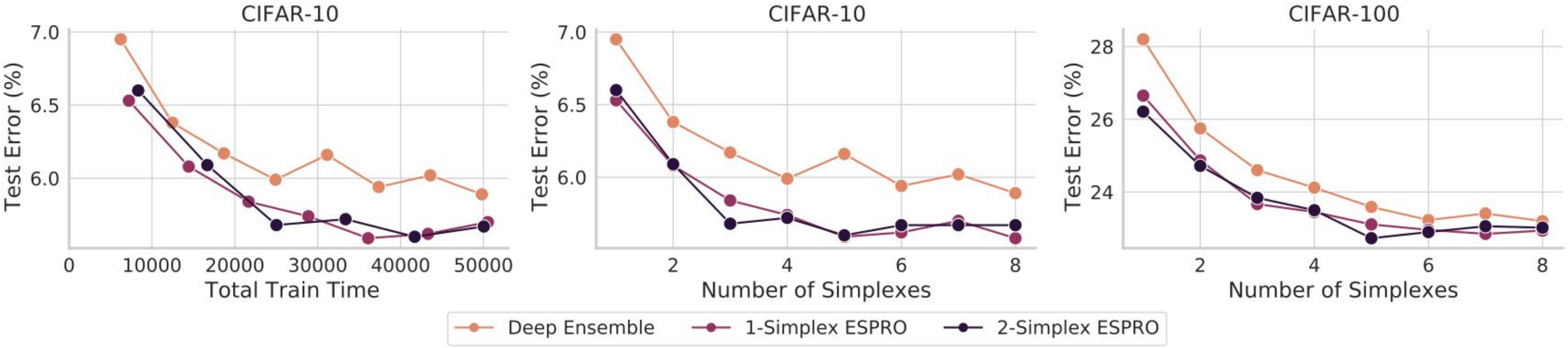
Ensembling In Low Loss Simplexes

- The regularizer encourages *parameter diversity as a proxy for functional diversity*
- Models from within a single ESPRO simplex find distinct representations of the data
- Ensemble over samples from simplex

$$\hat{y} = \frac{1}{J} \sum_{\phi_j \sim \mathcal{K}} f(x, \phi_j) \approx \int_{\mathcal{K}} f(x, \phi) d\phi$$

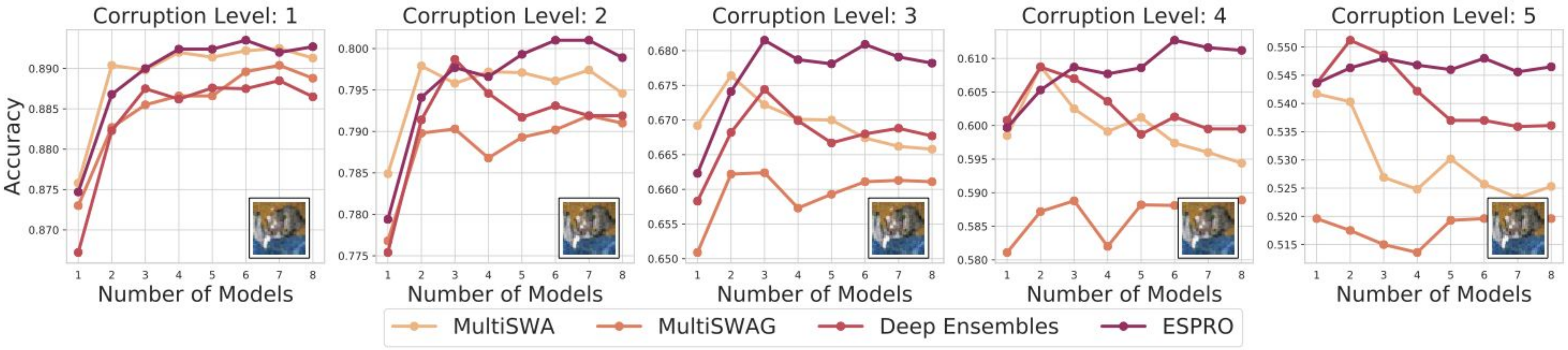
Ensembling In Low Loss Simplexes

- ESPRO is more accurate than traditional deep ensembles
  - Given small number of additional epochs, faster to train than equally accurate deep ensembles



Ensembling In Low Loss Simplexes

- ESPRO is more accurate than traditional deep ensembles
  - Given small number of additional epochs, faster to train than equally accurate deep ensembles
- Is more robust to dataset shift



Thank You

- *github.com/g-benton/loss-surface-simplexes*
- *Paper number 4477*