Mixture-of-Products Model Structure for BirdFlow

Mixture of Products: Motivation

Goal: train models over a loss function with a term that depends on the site fidelity marginal (to get sampled tracks that start and end at the same location)

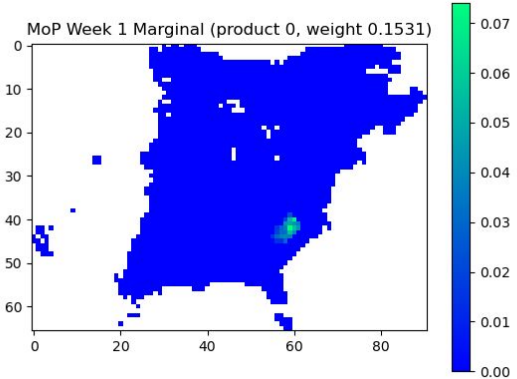
Challenge: We can't necessarily achieve this goal using the markov chain model structure, due to storage concerns

Potential Solution: Use a Mixture-Of-Products model structure instead, which automatically eliminates these storage concerns

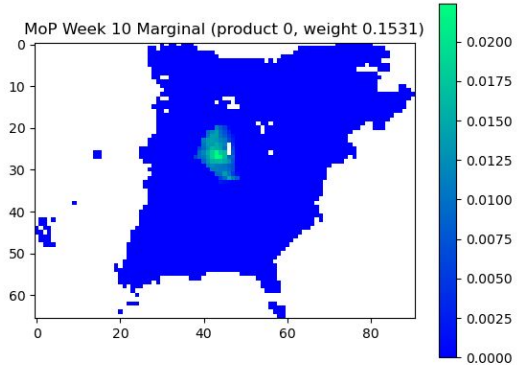
- Overview of the Mixture-Of-Products model
  - What is a Mixture-Of-Products?
  - Calculating single-timestep, pairwise, **site-fidelity marginals**
  - Sampling, conditional sampling
  - Forecasting
- Overview of experimentation done so far!

What is a Product Distribution?

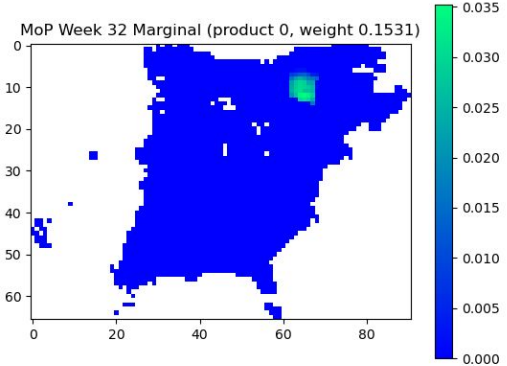
- Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T$ be random variables for a bird's location in weeks 1 through T
- $p^{(k)}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T) = p^{(k)}(\mathbf{X}_1)\, p^{(k)}(\mathbf{X}_2) \ldots p^{(k)}(\mathbf{X}_T)$ is a **product distribution** over tracks



$p^{(k)}(\mathbf{X}_1)$



$p^{(k)}(\mathbf{X}_{10})$



$p^{(k)}(\mathbf{X}_{32})$

What is a Mixture of Product Distributions?

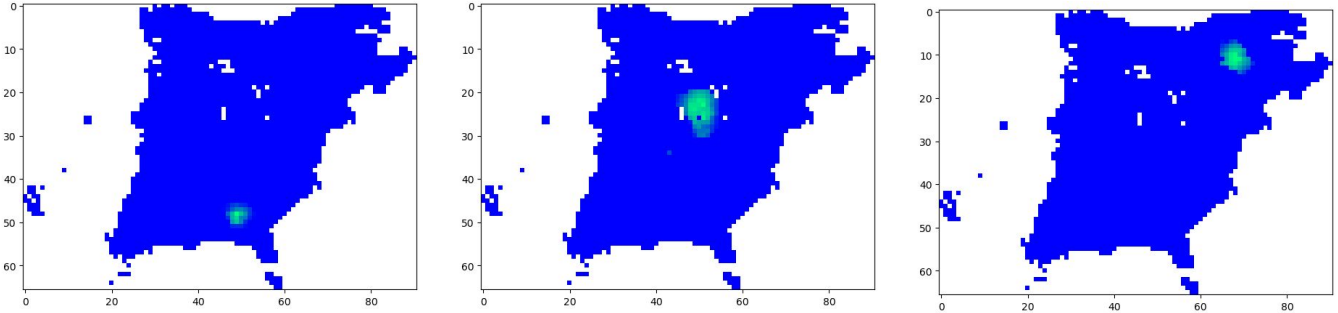
- A weighted average of **n** product distributions, called **components**.

$$p_\theta(X_1, ..., X_T) = \sum_{k=1}^{n} \alpha_k p_\theta^{(k)}(X_1, \ldots X_T)$$

- $\boldsymbol{\alpha}$ is a vector of **n** positive weights that sum to one

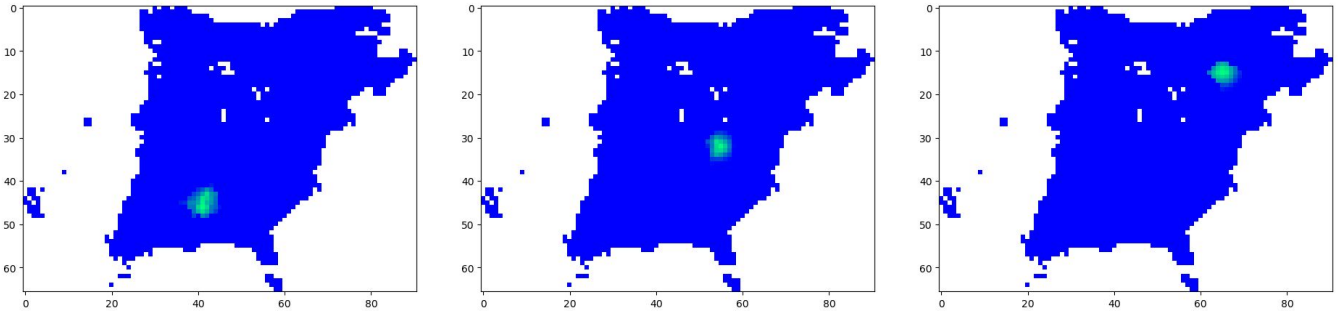
Visualizing a Mixture-Of-Products: toy example

$p_\theta^{(1)}(X_1)$ $p_\theta^{(1)}(X_2)$ $p_\theta^{(1)}(X_3)$



$\alpha_1 = 0.5$

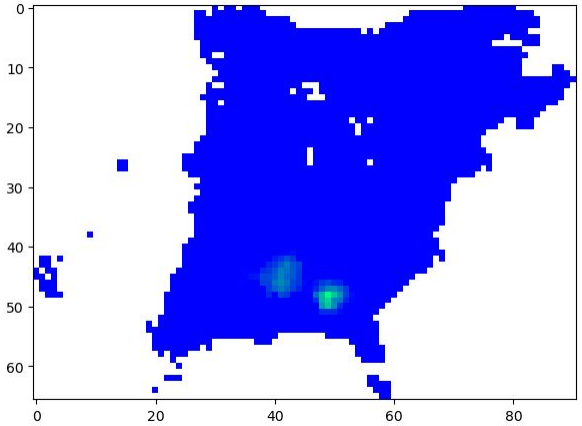
$p_\theta^{(2)}(X_1)$ $p_\theta^{(2)}(X_2)$ $p_\theta^{(2)}(X_3)$



$\alpha_2 = 0.5$

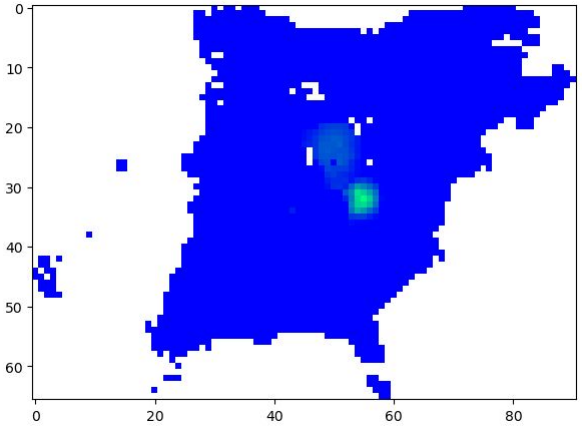
Visualizing a Mixture-Of-Products: toy example

$p_\theta(X_1)$



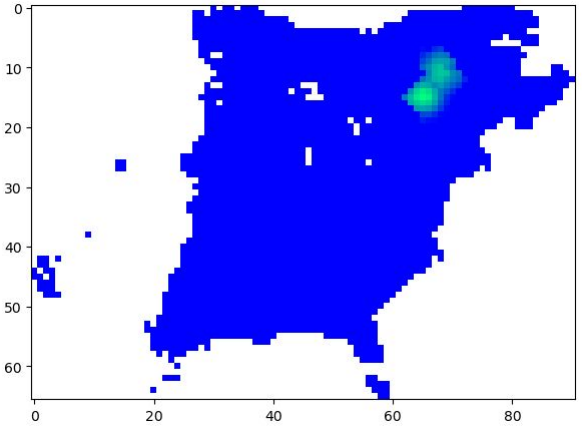
$p_\theta(X_1) = \sum_{k=1}^{2} \alpha_k p_\theta^{(k)}(X_1)$

$p_\theta(X_2)$



$p_\theta(X_2) = \sum_{k=1}^{2} \alpha_k p_\theta^{(k)}(X_2)$

$p_\theta(X_3)$



$p_\theta(X_3) = \sum_{k=1}^{2} \alpha_k p_\theta^{(k)}(X_3)$

Computing Single-Timestep and Pairwise Marginals of a Mixture-of-Products

- Single-timestep marginals:

$$p_\theta(X_t) = \sum_{k=1}^{n} \alpha_k p_\theta^{(k)}(X_t)$$

- Pairwise marginals:

$$p_\theta(X_t, X_{t+1}) = \sum_{k=1}^{n} \alpha_k p_\theta^{(k)}(X_t) p_\theta^{(k)}(X_{t+1})$$

Computing the Site Fidelity Marginal: a motivation for the Mixture-Of-Products model structure

$$p_\theta(X_1, X_T) = \sum_{k=1}^{n} \alpha_k p_\theta^{(k)}(X_1) p_\theta^{(k)}(X_T)$$

Key Point: Computing the site fidelity marginal when modeling the track distribution with a **Markov Chain** requires $O(n^3)$ space. Not feasible if we have a lot of grid cells.

When we model the track distribution with a **Mixture-Of-Products**, computing the site fidelity marginal only requires $O(n^2)$ space.

Sampling a Track from a Mixture of Products: A Two-Step Process

1) Pick a product distribution with probability given by the product distribution's weight (product distribution **k** is picked with probability $\boldsymbol{\alpha}_{\mathbf{k}}$)

2) Sample $\mathbf{x}_1$ from $p^{(k)}(\mathbf{X}_1)$, sample $\mathbf{x}_2$ from $p^{(k)}(\mathbf{X}_2)$, …, sample $\mathbf{x}_T$ from $p^{(k)}(\mathbf{X}_T)$, yielding the sampled track ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_T$).

Conditional Sampling: Problem Setup

Suppose we observe a bird to be at cell x_t in week t, and wish to sample the bird's location in weeks $j_1, \ldots, j_q$ given this observation.

We sample from the conditional $p_\theta(X_{j_1}, \ldots X_{j_q} | X_t = x_t)$

Conditional Sampling: the two-step process

1. Select a product distribution k with probability proportional to $\alpha_k \cdot p_\theta^{(k)}(X_t = x_t)$
2. Sample x_{j_1} from $p_\theta^{(k)}(X_{j_1})$, sample x_{j_2} from $p_\theta^{(k)}(X_{j_2})$..., sample x_{j_T} from $p_\theta^{(k)}(X_{j_T})$

Forecasting with a Mixture-Of-Products

Suppose we observe a bird's location at cell $\mathbf{x}_t$ in week **t**, and wish to compute the distribution of the bird's location in week **j**.

$$p_\theta(X_j|X_t = x_t) = \sum_{k=1}^{n} \pi_k p_\theta^{(k)}(X_j)$$

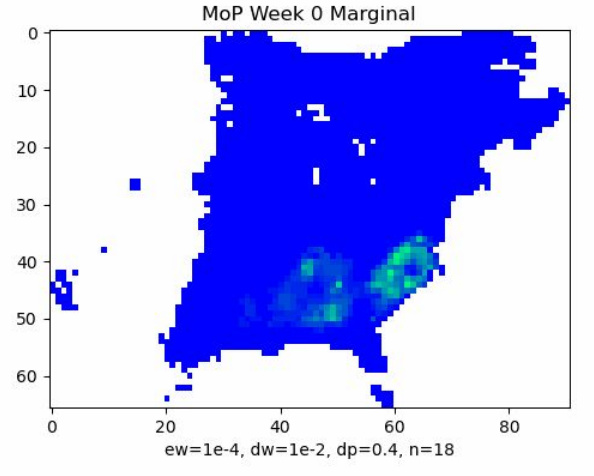
$\pi_k = n_c \alpha_k p_\theta^{(k)}(X_t = x_t)$, n_c is a constant used to scale the weights of π to sum to one

Overview of Experiments

- Initial Experiments
  - Trained models with up to 1000 components
- Evaluating the loss of Mixture-Of-Products made from tracks sampled from a the markov chain
- Fine tuning from sampled tracks

Initial Experiments: Methodology

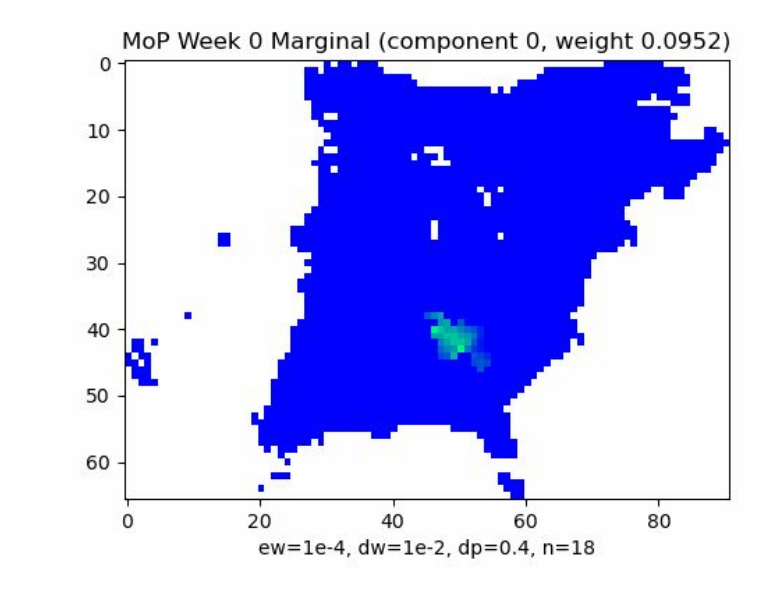
- Trained Mixture of Products models with 6-1000 components using gradient descent over the standard BirdFlow loss function (with observation, distance, and entropy terms)
- Learned the weights and weekly distributions of each component
- Initial weights / weekly distributions sampled from a normal distribution

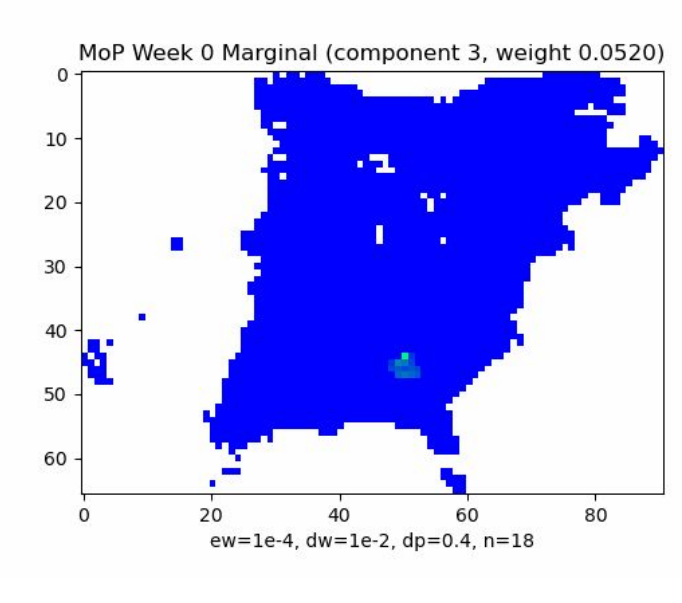


This gif shows
single-timestep marginals of
a model with 18 components.

ew=1e-4, dw=1e-2, dp=0.4,
n=18 (# of components)

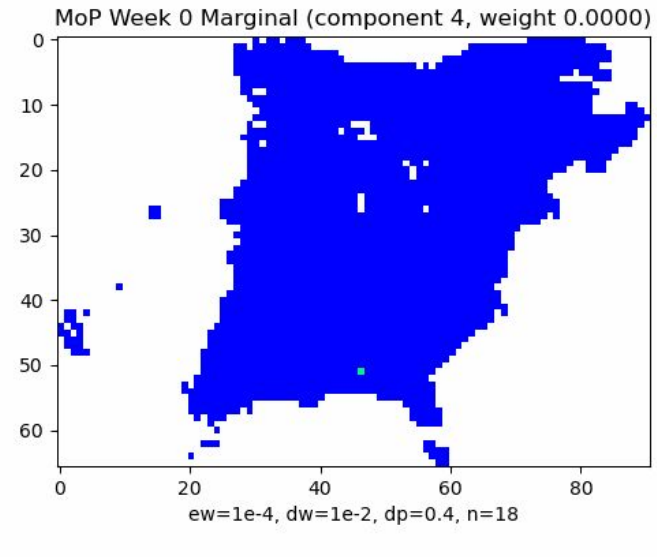
Observation 1: Learned components resemble routes, but tend to wander around

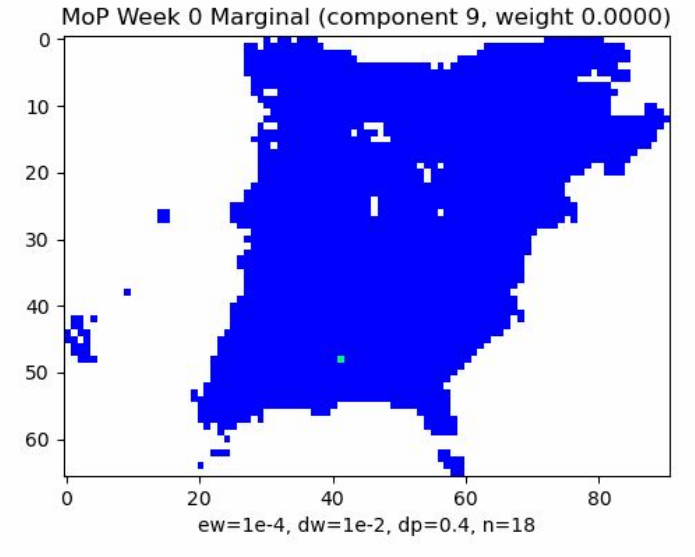




ew=1e-4, dw=1e-2, dp=0.4, n=18 (components 0, 3)

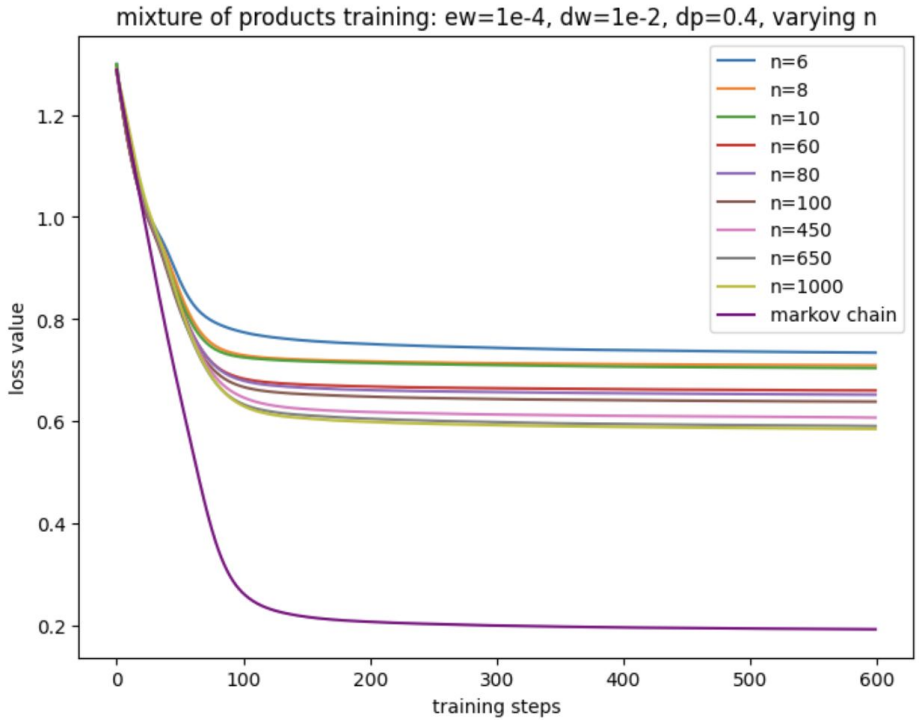
Observation 2: The presence of Collapsed Components





ew=1e-4, dw=1e-2, dp=0.4, n=18 (components 4, 9)

Initial Experiments: Mixture-Of-Products Optima are far from that of the Markov Chain



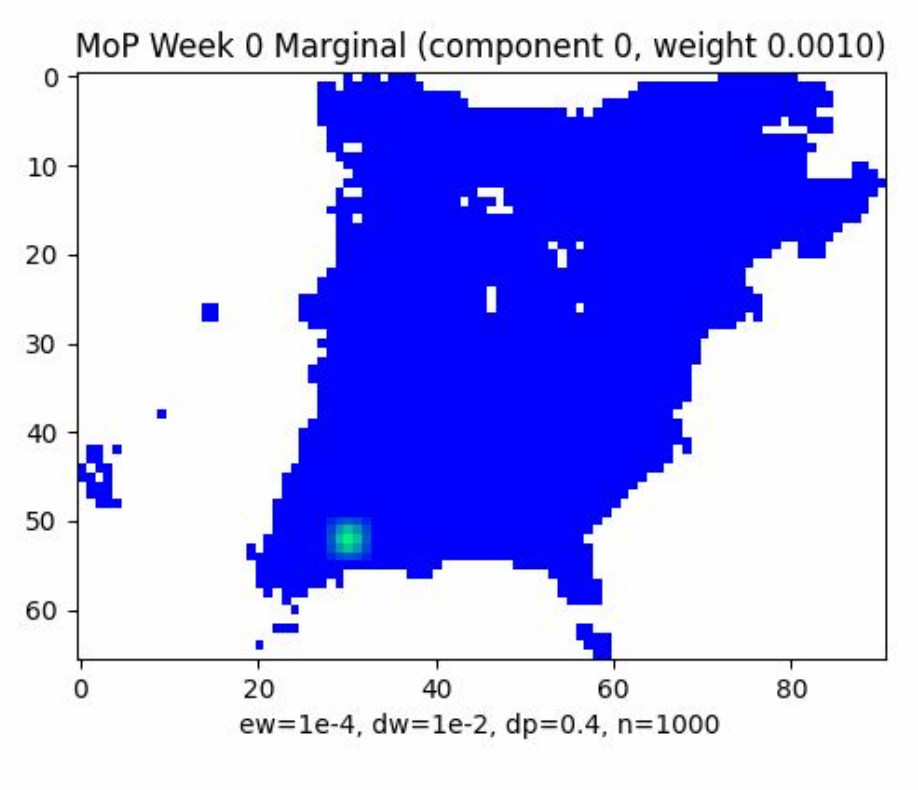
Methodology:

- Train a Markov Chain using the same hyperparameters as for the Mixture-Of-Products
- Train mixture-of-products models with up to 1000 components
- Plot loss over number of training steps

Creating Mixture-Of-Products Components by sampling routes from the Markov Chain: Methodology

- Sample routes from the Markov Chain
- Create components from these routes, weight them equally
  - weekly marginals are a “box” centered at the bird’s location during that week of the route
  - compute probability values with a spherical gaussian
- Parameters varied:
  - box size
  - variance of the spherical gaussian (lower variance ~ probability is more concentrated)
- **Question: how do the loss values of these Markov Chain - sampled MoPs compare to the optima found when training MoPs?**

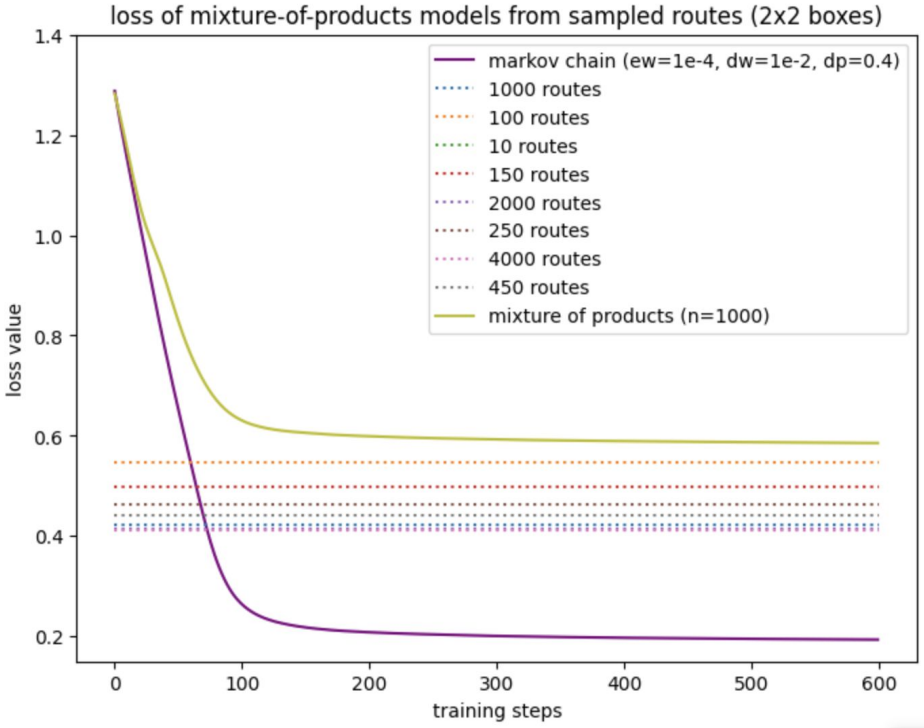
Mixture-Of-Products Component from a Sampled Route

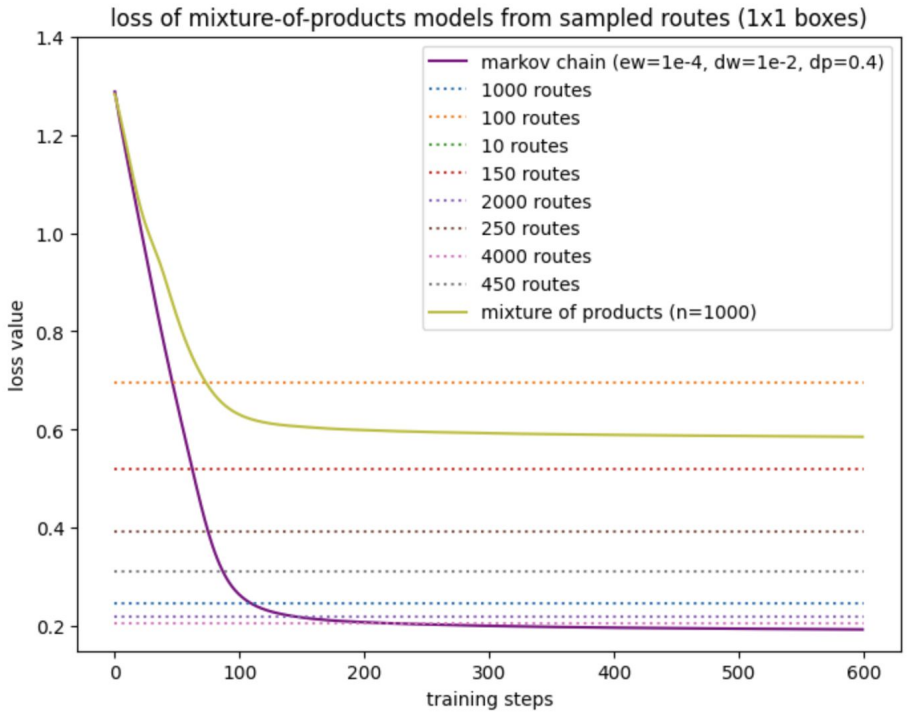


5x5 box size

scale=2 (spread of the gaussian used to generate probabilities)

Markov-Chain-Sampled Mixture-Of-Products Models have low loss values! (which demonstrates our training got stuck in a local optimum initially)

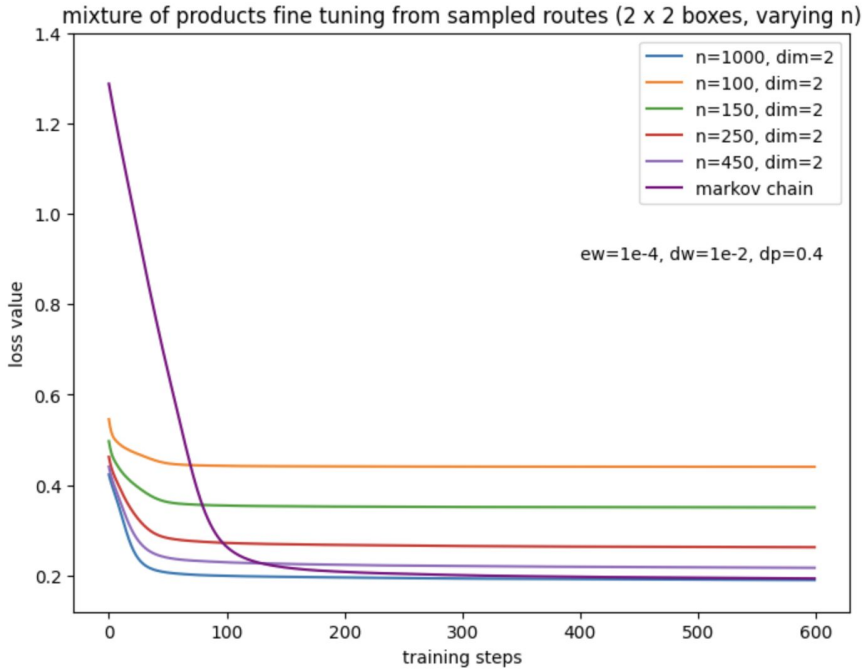


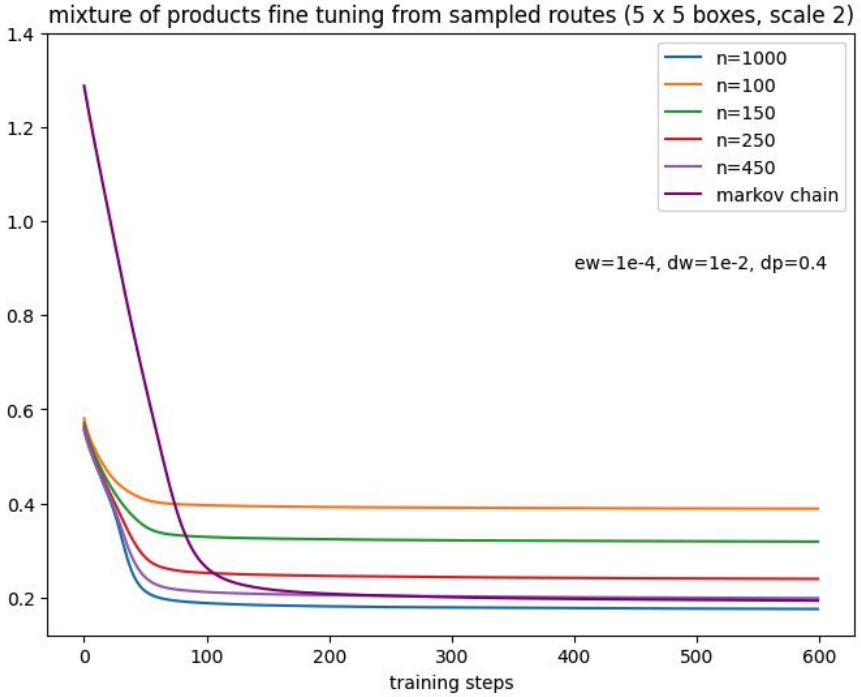


Fine tuning Mixture-Of-Products from Markov-Chain sampled components: Methodology

- Initialize the Mixture of Products components to be routes sampled from the markov chain, initialize the weights to be equal
  - experiment with different box sizes and scales
- Train using gradient descent over the standard BirdFlow loss function

Fine-tuned mixture of products have comparable loss to the markov chain!





Conclusions, and Future Directions

Fine tuning mixture of products models from routes sampled from the Markov Chain produces MoP models with comparable loss to the Markov Chain!

Future Directions:

- Analyze the effect of box size and scale on the quality of the fine-tuned models - grid search!
- Experiment with new optimization and initialization techniques
- Grid search over entropy weight, distance weight / power
- Training MoPs (and Markov Chains?) on a loss function with a site fidelity term

Discussion!