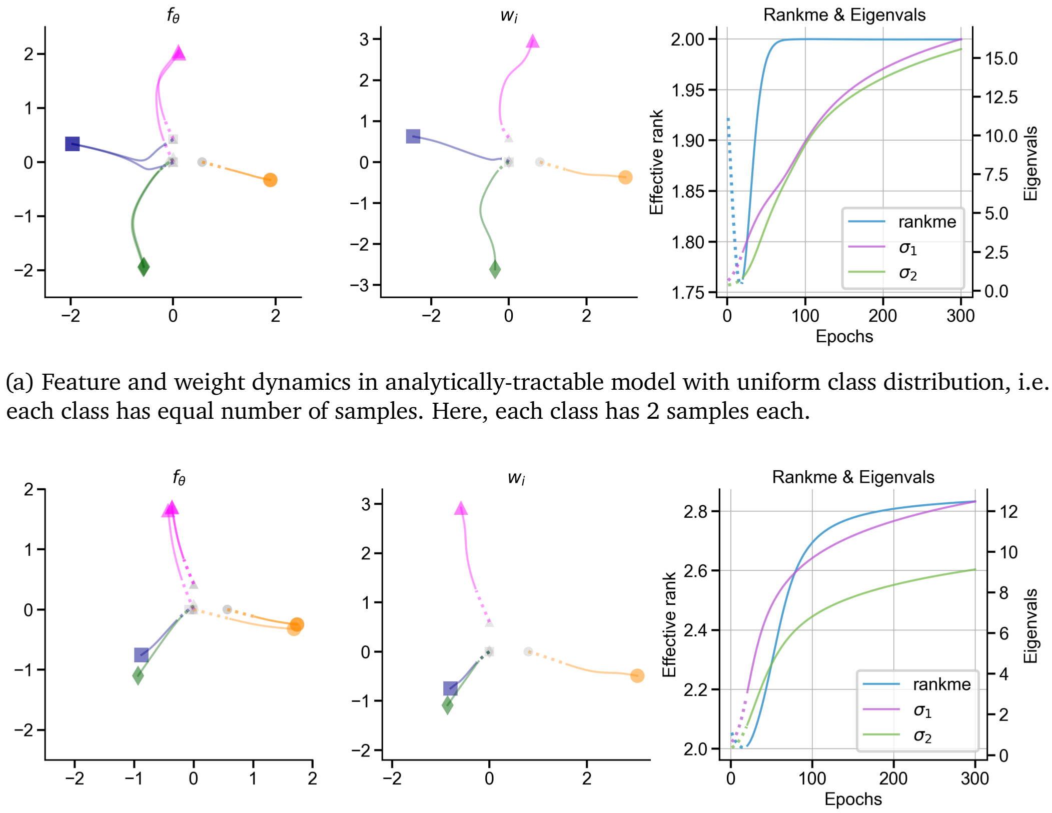
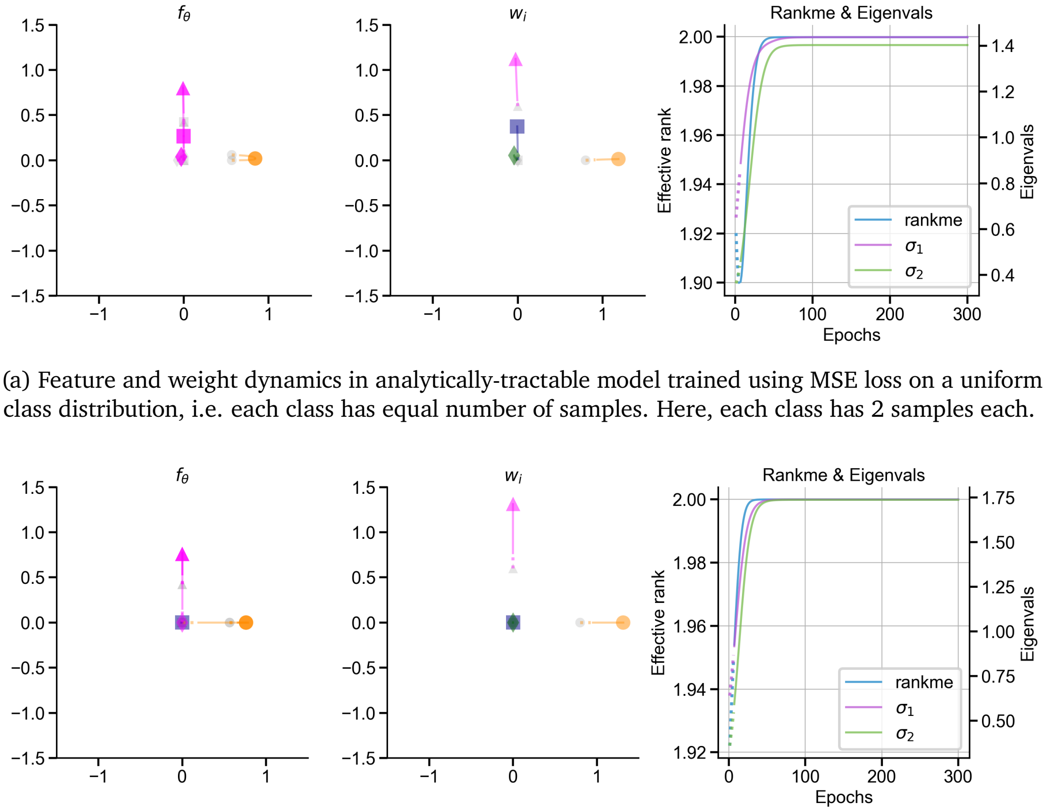
C.3. Control experiments verifying the necessity condition for multiphase learning dynamics



(a) Feature and weight dynamics in analytically-tractable model with uniform class distribution, i.e. each class has equal number of samples. Here, each class has 2 samples each.

(b) Feature and weight dynamics in analytically-tractable model with no information bottleneck, i.e. feature dimensionality, d, is comparable to number of classes, $|\mathcal{V}|$. Here, $d = 3$ and $|\mathcal{V}| = 4$. Note that we only plot the first two dimensions for ease and consistency of visualization.

Figure 12: Negative control experiments analogous to Figure 5. Removing either the skewed class distribution or the information bottleneck gets rid of the three distinct phases of learning. In each case, the resulting dynamics is an initial “warmup” , followed by an “entropy-seeking” phase wherein effective rank continues to grow monotonically.



(a) Feature and weight dynamics in analytically-tractable model trained using MSE loss on a uniform class distribution, i.e. each class has equal number of samples. Here, each class has 2 samples each.

(b) Feature and weight dynamics in analytically-tractable model trained using MSE loss on a skewed class distribution. Note that only information about the most frequently occurring classes are learned.

Figure 13: Negative control experiments analogous to Figure 5, with mean squared error instead of cross-entropy as the training loss. In both uniform and skewed label distribution settings, the resulting dynamics is an initial "warmup" , followed by an "entropy-seeking" phase wherein effective rank grows monotonically and quickly saturates.

C.4. Supervised finetuning

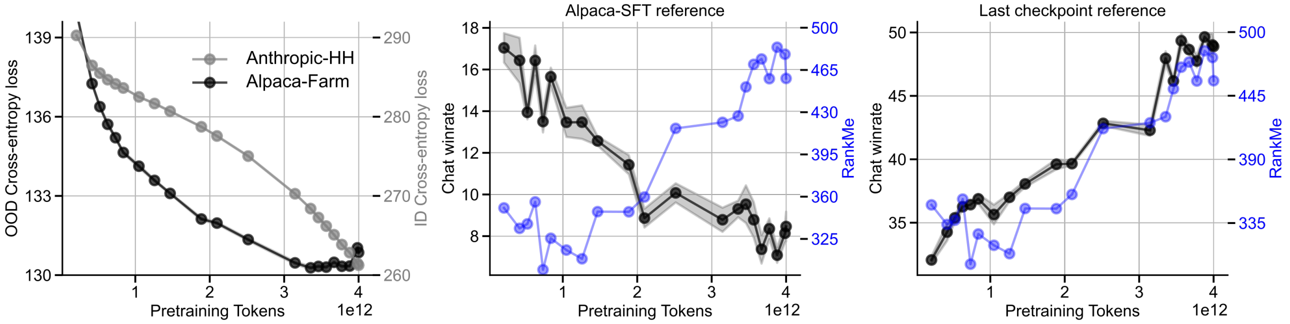


Figure 14: Loss and chat win rates after SFT on Anthropic-HH dataset. **(Left)** Cross-entropy loss on in-distribution (Anthropic-HH) test set and out-of-distribution (Alpaca-Farm-human-ANN chat) dataset. While in-distribution loss after SFT decreases monotonically, ood loss after SFT saturates or gets slightly worse with longer pretraining. **(Center)** Length-controlled chat win rates for Anthropic-SFT vs Alpaca-SFT version of a base model on AlpacaEval. Longer pretraining increases the sensitivity of the model's behavior to the SFT dataset. **(Right)** Length-controlled win rates for Anthropic-SFT version of intermediate base models compared to the Anthropic-SFT version of the final base model checkpoint. Models obtained from later in pretraining are equivalent chat models, demonstrating nearly 50% win rate compared to the final checkpoint. Choosing the ideal checkpoint to use for SFT requires navigating the tradeoff between an improvement in base model's capability (note an increased `RankMe`) and reduction in robustness with longer pretraining. Shaded bars indicate standard deviation computed over 5 seeds.