

Transformers can optimally learn regression mixture models

Reese Pathak[◊] Rajat Sen[†] Weihao Kong[†] Abhimanyu Das[†]

UC Berkeley, EECS[◊]
Google Research[†]

Abstract

Mixture models arise in many regression problems, but most methods have seen limited adoption partly due to these algorithms’ highly-tailored and model-specific nature. On the other hand, transformers are flexible, neural sequence models that present the intriguing possibility of providing general-purpose prediction methods, even in this mixture setting. In this work, we investigate the hypothesis that transformers can learn an optimal predictor for mixtures of regressions. We construct a generative process for a mixture of linear regressions for which the decision-theoretic optimal procedure is given by data-driven exponential weights on a finite set of parameters. We observe that transformers achieve low mean-squared error on data generated via this process. By probing the transformer’s output at inference time, we also show that transformers typically make predictions that are close to the optimal predictor. Our experiments also demonstrate that transformers can learn mixtures of regressions in a sample-efficient fashion and are somewhat robust to distribution shifts. We complement our experimental observations by proving constructively that the decision-theoretic optimal procedure is indeed implementable by a transformer.

1 Introduction

In several machine learning applications—federated learning [Wang et al., 2021], crowd-sourcing [Steinhardt et al., 2016] and recommendations systems [Wang et al., 2006]—data is collected from multiple sources. Each source generally provides a small batch of data: for instance in recommendation systems, a user can provide a source of rating data on a subset of items that she has encountered. Such batches, on their own, are often too small to learn an effective model for the desired application. On the other hand, by pooling many batches together, improvements can typically be made in the quality of the predictors that can be learned.

An issue with this “pooling” approach is that if it is done carelessly, then the models which are learned may lack personalization [Ting et al., 1999]. For instance, in a recommendation system, such an approach could yield a model that selects similar actions for dissimilar users. A better approach, however, is to model the problem as a *mixture* of distributions: for instance, we can model the sources as arising from m subpopulations, assuming that sources arising from the subpopulation have similar underlying distributions [Kleinberg and Sandler, 2004]. The sources from a single subpopulation can then be aggregated for the purposes of learning. For instance, in the recommendation systems example, users in the same subpopulation might be identified as having similar preferences and tastes for item genres.

A supervised learning formulation of the above setup is that the sources arise from a subpopulation indexed by an integer $i \in [m] := \{1, 2, \dots, m\}$. Additionally, assume that within each subpopulation the input-output pair (x, y) follows a relation of the form $y = f_i^*(x) + \eta$, where η is a zero-mean noise, and $x \in \mathbf{R}^d$. A batch of i.i.d data from such a source can be represented as $\{(x_j, y_j)\}_{j=1}^k$ where k is the batch size. Given many such batches, each having

examples only from one source, the task is to learn the functions $\{f_i^*\}_{i=1}^m$ well enough to make good predictions on another input, sometimes also referred to as a *query*, x_{k+1} . For instance, given the past ratings of an user, we should be able to determine their subpopulation well enough to infer their preferences on an unseen item.

The simplest version of the formulation above additionally imposes the assumption that the functions f_i^* are linear: $f_i^*(x) = \langle w_i^*, x_i \rangle$. This setting has been studied theoretically in [Kong et al., 2020, Jain et al., 2023]. Kong et al. [2020] introduced the problem and designed an algorithm in the setting where there are as many as $O(d)$ batches with size $k = O(1)$, and fewer medium-sized batched of size $k = O(\sqrt{m})$. However, that work imposed strong assumptions on the covariate distribution, which lead to the paper Jain et al. [2023], where these assumptions were relaxed. This latter work proposes a different algorithm that even allows covariate distributions to vary among subpopulations. Nonetheless, their algorithm needs to know problem parameters, such as a $L^2 - L^4$ hypercontractivity, a spectral norm bound on the covariance matrix, the noise level, and the number of subpopulations m . It is not clear how these algorithms will fare with model misspecification, or if they could be extended to applications like federated learning where it is unlikely that the correct model is linear, and distributed algorithms are required.

In this work, we ask the question: *Is there a deep learning architecture that can be trained using standard gradient decent, yet learns mixture models from batched data and can leverage small batches from a source to make predictions for its appropriate subpopulation?* If so, this would allow us to solve these type of mixture problem without needing highly specialized algorithms that could potentially be brittle with respect to knowing the correct form of the mixture model. Moreover, standard (stochastic) gradient descent would naturally extends to distributed training, using standard techniques from federated learning [Wang et al., 2006].

A natural candidate to address this question is the widely used transformer architecture [Vaswani et al., 2017]. Motivated by their immense success in NLP [Radford et al., 2019], computer vision [Han et al., 2022] and in context learning abilities demonstrated by large models [Chowdhery et al., 2022], several recent works have been aimed to analyze whether transformers can learn algorithms [Akyürek et al., 2023, Garg et al., 2022, Von Oswald et al., 2023]. These papers train decoder only transformers using prompts of the form $(x_1, f(x_1), \dots, x_j, f(x_j), \dots, x_k, f(x_k), x_{k+1})$ where the task is to predict $f(x_{k+1})$ after seeing the portion of the prompt preceding it. These papers show empirically that when f is sampled from a class of linear functions, then transformers learn to perform linear regression in-context. Akyürek et al. [2023] also show that transformers can represent gradient descent in the case of linear regression provided through a construction.

Our contributions: Motivated by the above papers, we investigate whether transformers can solve the problem of learning mixture models from batched data. Our contributions are as follows:

- We demonstrate that transformers can learn mixtures of linear regressions by training on such mixture data and exhibiting near-Bayes-optimal error at inference time.
- We strengthen this observation by proving constructively that transformers can implement the optimal method for the mixture distribution on which the transformer was trained.

- Our experiments show that transformers are sample-efficient: the transformers’ performance is similar (or better) than model-specific methods, when fixing the training set size.
- We evaluate certain inference-time metrics that capture the nearness of predictions made by the transformer versus another predictor. We show these metrics are smallest when taking the comparator to be the decision-theoretic optimal method, thereby further corroborating the hypothesis that transformers optimally learn mixtures of regressions.
- We suggest that transformers tolerate “small” distribution shifts by investigating transformers’ performance on both covariate and label shifts to the mixture model.

These contributions, taken together, are evidence that transformers can optimally, efficiently, and robustly learn mixtures of linear regressions.

Related work: The related work can be broadly divided into a thread that studies the theoretical properties of algorithms for estimation or prediction in a regression mixture model as well as another thread that studies the empirical and theoretical properties of transformers on learning regression models. Due to space considerations, we present a more detailed overview in Appendix A.

1.1 Generative model for data

Underlying the mixture of linear regressions, we consider the discrete mixture

$$\pi := \frac{1}{m} \sum_{i=1}^m \delta_{w_i^*}, \quad (1)$$

where $\{w_i^*\}_{i=1}^m \in \mathbf{R}^d$ are normalized such that $\|w_i^*\|_2 = \sqrt{d}$ for each $i \in [m]$. We consider *prompts* or *batches*, denoted $P = (x_1, y_1, \dots, x_k, y_k, x_{k+1})$. Here, for noise level $\sigma \geq 0$, we have

$$w \sim \pi, \quad x_i \stackrel{\text{i.i.d.}}{\sim} \mathbf{N}(0, I_d), \quad \text{and} \quad y_i \mid x_i \sim \mathbf{N}(\langle w, x_i \rangle, \sigma^2).^1 \quad (2)$$

The goal is then to predict y_{k+1} , the label for the query x_{k+1} .

1.2 Transformers

Transformers are deep neural networks that map sequences to sequences [Vaswani et al., 2017]. In this work, we specifically focus on decoder-only, *autoregressive transformers*. These models are comprised of multiple layers that map an input matrix $H \in \mathbf{R}^{p \times q}$ to an output matrix $H' \in \mathbf{R}^{p \times q}$. Here p denotes the hidden dimension, and q corresponds to the number of input tokens. The output is then fed successively to more such layers. Since the computation in each layer is the same (apart from parameters), we describe the computation occurring in a single layer. Write $h_j = (H_{ij})_{1 \leq i \leq p}$ for the j th column of H , and h'_j for the j th column of H' . Additionally, the prefix matrix $H_{:,i}$ is the $p \times (i-1)$ submatrix of H obtained by concatenating the first $i-1$ columns of H .²

¹When $\sigma = 0$, by $\mathbf{N}(v, 0)$ we mean the point mass δ_v .

²In the case $i = 1$, the submatrix can be interpreted as 0.

A layer is parameterized by a sequence of weights. Let n_{heads} denote the number of attention heads and d_{att} denote the hidden dimension for the attention layer and d_{ff} denote the hidden dimension for the feedforward (*i.e.*, dense) layer. Then, a layer has the following weights:

$$\begin{aligned} & \{W_i^{\text{Q}}\}_{i=1}^{n_{\text{heads}}}, \{W_i^{\text{V}}\}_{i=1}^{n_{\text{heads}}}, \{W_i^{\text{K}}\}_{i=1}^{n_{\text{heads}}}, \subset \mathbf{R}^{d_{\text{att}} \times p}, \\ & \{W_i^{\text{C}}\}_{i=1}^{n_{\text{heads}}} \subset \mathbf{R}^{p \times d_{\text{att}}} \quad W^{\text{in}} \in \mathbf{R}^{d_{\text{ff}} \times p}, \quad \text{and} \quad W^{\text{out}} \in \mathbf{R}^{p \times d_{\text{ff}}} \end{aligned} \quad (3)$$

for each column $i \in [p]$, the computation proceeds in the following fashion.

Self-attention: The layer begins with computing the attention vector, $a_i \in \mathbf{R}^p$, by

$$\begin{aligned} s_{ij} &:= \text{softmax} \left((W_j^{\text{K}} H_{:i})^\top W_j^{\text{Q}} h_i \right), \quad \text{for } j \in [n_{\text{heads}}], \quad \text{and,} \\ a_i &:= \sum_{j=1}^{n_{\text{heads}}} W_j^{\text{C}} W_j^{\text{V}} H_{:i} s_{ij} \end{aligned}$$

Above, with a slight abuse of notation, we define for any integer $\ell > 0$, $\text{softmax}: \mathbf{R}^\ell \rightarrow \mathbf{R}^\ell$ by the formula $\text{softmax}(v) = (e^{v_t} / \sum_{t'=1}^\ell e^{v_{t'}})_{t=1}^\ell$. Note that, above, $s_{ij} \in \mathbf{R}_+^{i-1}$.³

Feedforward network: The layer then continues by passing the attention vector (along with the original input column h_i) through a nonlinear dense layer. This is defined by

$$h'_i := a_i + h_i + W^{\text{out}} \sigma_\bullet(W^{\text{in}} \lambda(a_i + h_i))$$

Above the notation σ_\bullet indicates that the map $\sigma: \mathbf{R} \rightarrow \mathbf{R}$ is applied *componentwise* to its argument. In this work we take the nonlinearity to be the Gaussian error linear unit (GeLU) [Hendrycks and Gimpel, 2016] which is defined by

$$\sigma(u) = \frac{u}{2} \left(1 + \text{erf} \left(\frac{u}{\sqrt{2}} \right) \right), \quad \text{for any } u \in \mathbf{R}.$$

Above, erf denotes the Gauss error function. The function $\lambda: \mathbf{R}^p \rightarrow \mathbf{R}^p$ denotes layer normalization [Ba et al., 2016], and is given by

$$\lambda(v) = \sqrt{p} \frac{v - \bar{v} \mathbf{1}_p}{\|v - \bar{v} \mathbf{1}_p\|_2}, \quad \text{where} \quad \bar{v} = \frac{1}{p} \sum_{i=1}^p v_i.$$

This is a form of standardization where \bar{v} is interpreted as the mean (averaging the components) and $\|v - \bar{v} \mathbf{1}_p\|_2^2/p$ is interpreted as the variance (averaging the squared deviation to the mean).

2 Representation

In this section, we prove that transformers can actually represent the minimum mean squared error procedure. Indeed, let $f: P \mapsto \hat{y} \in \mathbf{R}$, by any procedure which takes a prompt P and outputs an estimate \hat{y} on the query, and define the mean squared error (MSE) by

$$\text{MSE}(f) := \mathbf{E}_P \left[(f(P) - y_{k+1})^2 \right].$$

³When $i = 1$, $s_{ij} = 0$.

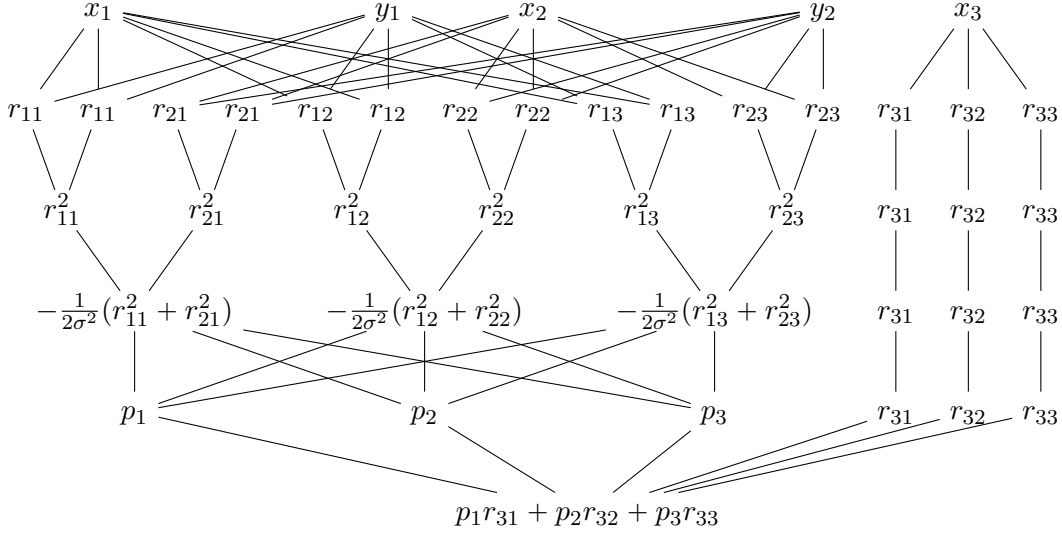


Figure 1. Illustration of an arithmetic circuit, implementable by a transformer, that computes the posterior mean as defined in display (4). Here, r_{ij} are residuals as defined in display (5) and p_j are probabilities obtained via a softmax operation, as defined in display (6). See main text for a description of the computation occurring at each level.

Then by standard Bayesian decision theory, under the observational model described in Section 1.1, it follows that the mean squared error is minimized at the posterior mean f_π^\star , which is given by

$$f_\pi^\star(P) = \langle \hat{w}(P), x_{k+1} \rangle \quad \text{where} \quad \hat{w}(P) := \frac{\sum_{j=1}^m w_j^\star \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^k (\langle w_j^\star, x_i \rangle - y_i)^2 \right)}{\sum_{\ell=1}^m \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^k (\langle w_\ell^\star, x_i \rangle - y_i)^2 \right)}. \quad (4)$$

Formally, $\text{MSE}(f) \geq \text{MSE}(f_\pi^\star)$, for all (measurable) f . Note above that \hat{w} does not depend on x_{k+1} .

Then our main result is that the function f_π^\star can be computed by a transformer.

Theorem 1. *There is an autoregressive transformer which implements the function f_π^\star as defined in equation (4).*

See Section B for a proof of this claim.

For an illustration of the underlying idea behind Theorem 1, see Figure 1 for an arithmetic circuit that computes the function f_π^\star , in the case $m = 3, k = 2$. The objects r_{ij} are residuals, defined as

$$r_{ij} = \langle w_j^\star, x_i \rangle - y_i \mathbf{1}\{i \neq k+1\}, \quad \text{for } i \in [k], j \in [m]. \quad (5)$$

The first layer computes the values $\{r_{ij}\}$, the second layer computes the squares of these values, the third layer computes the (scaled) sum of these values over the index i , which runs over the samples in the prompt, excluding the query. The fourth layer, computes the softmax of these sums,

$$p_j := \frac{\exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^k (\langle w_j^\star, x_i \rangle - y_i)^2 \right)}{\sum_{\ell=1}^m \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^k (\langle w_\ell^\star, x_i \rangle - y_i)^2 \right)}, \quad \text{for } j \in [m] \quad (6)$$

And the final layer computes

$$\sum_{j=1}^m p_j r_{(k+1),j} = \left\langle \sum_{j=1}^m p_j w_j^\star, x_{k+1} \right\rangle = \langle \hat{w}(P), x_{k+1} \rangle = f_\pi^\star(P)$$

where the last equation follows from the definitions in display (4). Therefore, the circuit depicted in Figure 1 is able to compute the posterior mean f_π^\star , at least for the choices $k = 2, m = 3$. Generalizing the circuit to general (k, m) is straightforward; therefore, our proof amounts to exploiting the circuit and demonstrating that each operation: linear transforms in the first and final layers, squaring in the second layer, summation in the third layer, softmax in the fourth layers are all implementable by a transformer.

3 Experimental results

In this section, we present results of training transformers on batches as described in Section 1.1. Our methodology closely follows the training procedure described in [Garg et al., 2022]. In the notation of Section 1.2, our transformer models set the hidden dimension as $p = 256$, feedforward network dimension as $d_{\text{ff}} = 4p = 1024$, and the number of attention heads as $n_{\text{heads}} = 8$. Our models have 12 layers. Additional details on the training methodology can be found in Appendix C. We also release our training and simulation code along with this paper.

3.1 Transformers can learn mixtures of linear regressions

To begin with, we investigate the performance of transformers on mixture models with various numbers of components and varying noise levels. We plot the performance of the transformer when prompted with a prompt P of length k , for $1 \leq k \leq 60$. The normalized MSE is the mean-squared error between the true labels and the estimated labels, divided by the dimension $d = 20$. Above, the algorithms that we compare against are:⁴

- *Ordinary least squares (OLS)*. For a prompt of length k , computes an element $\hat{w} \in \arg \min_{w \in \mathbf{R}^d} \sum_{j \leq k} (w^\top x_j - y_j)^2$. Outputs $\hat{y}_{k+1} = \hat{w}^\top x_{k+1}$.
- *Posterior mean*. This is an oracle algorithm. Given a prompt P of length k , computes the posterior mean $\hat{y}_{k+1} = f_\pi^\star(P)$, as defined in display (4).
- *Argmin*. This is an oracle algorithm. Given a prompt P of length k , computes

$$\hat{w} = \arg \min_{w \in \{w_j^\star\}_{j=1}^m} \sum_{j \leq k} \sum_{j \leq k} (w^\top x_j - y_j)^2. \quad (7)$$

The prediction is then $\hat{y}_{k+1} = \hat{w}^\top x_{k+1}$

Strikingly, we see that the transformer predictions are as good as—or nearly as good as—the oracle procedures which have knowledge of the true mixture components $\{w_j^\star\}$. It is important to note that OLS is suboptimal in general for mixtures of linear regressions. Nonetheless, the transformer is performing much better than OLS, indicating the trained transformer implements a better predictor to adapted to the mixtures of linear regressions setting.

⁴For interpretability of the figures, we omit the oracle algorithms above in the noiseless case ($\sigma = 0$) as the error is multiple orders of magnitude smaller than the data-driven procedures.

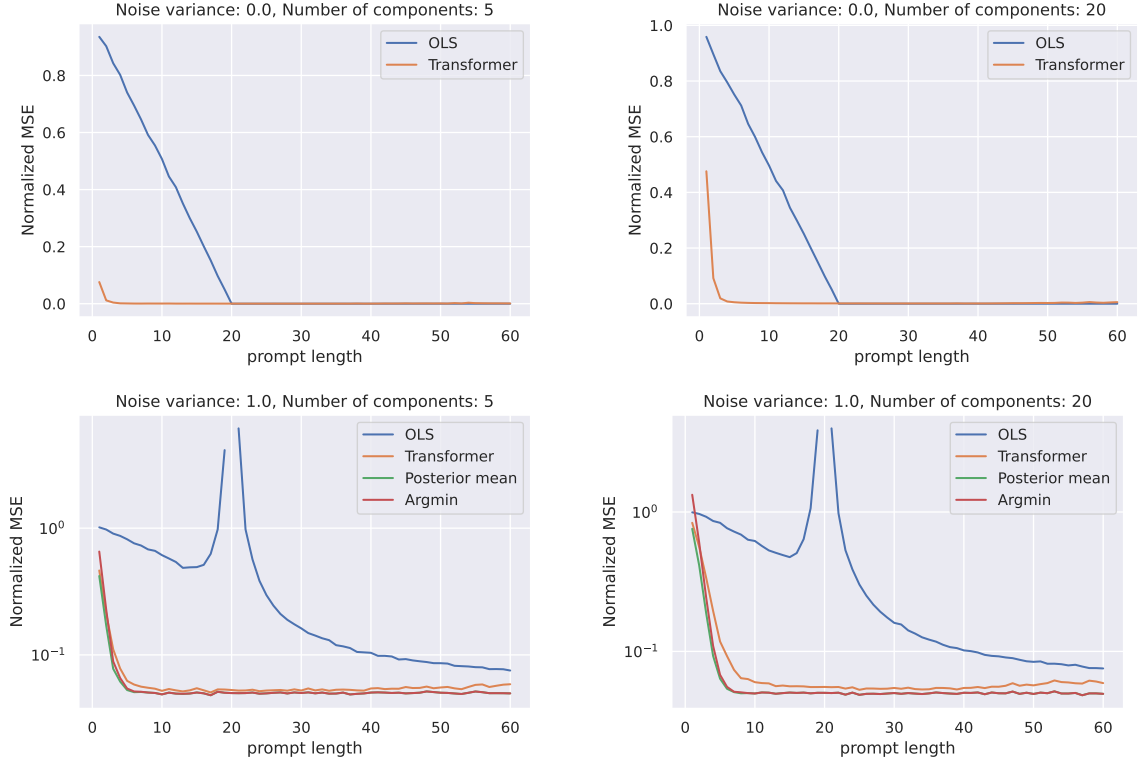


Figure 2. Transformer model trained on mixture of linear regressions data with 5 and 20 components. Top row is mixture data with no noise added to y ($\sigma = 0$) and bottom row is mixture data with noise added ($\sigma = 1$).

3.2 Comparison of performance for fixed training set size

Next, we investigate whether or not transformers learn mixtures of linear regressions in a sample efficient way. To do this, we depart slightly from the training methodology in [Garg et al., 2022]. We first sample a fixed training set of size $n \in \{15000, 30000, 45000, 60000\}$. Then—with some hyperparameter tuning to avoid overfitting, as well as a modification to the curriculum training, described in Appendix C.1—we train the transformer as in that paper. We then compare the inference time performance by computing the mean-squared error on prompts of length $k \in [1, 60]$.

The results of our simulation are shown below in Figure 3. We compared against two other procedures, which have the form of “plug-in” procedures:

- *Posterior mean, EM weights.* Here, we first estimate the component means w_j^* using batch expectation-maximization (see Appendix C.2 and Algorithm 1 therein for details). Then, we form $\hat{\pi}$, the uniform distribution over the estimated weights, and then predict y_{k+1} by $f_{\hat{\pi}}^*(P)$.
- *Posterior mean, SA weights.* We follow the approach above, but estimate the weights by using the subspace algorithm (SA), which is Alg. 1 in [Jain et al., 2023].

Note that the ‘oracle prediction error’ quantities appearing in Figure 3 are essentially the best possible error achievable using the weights estimated by the set of weights $\widehat{\mathcal{W}}$ estimated

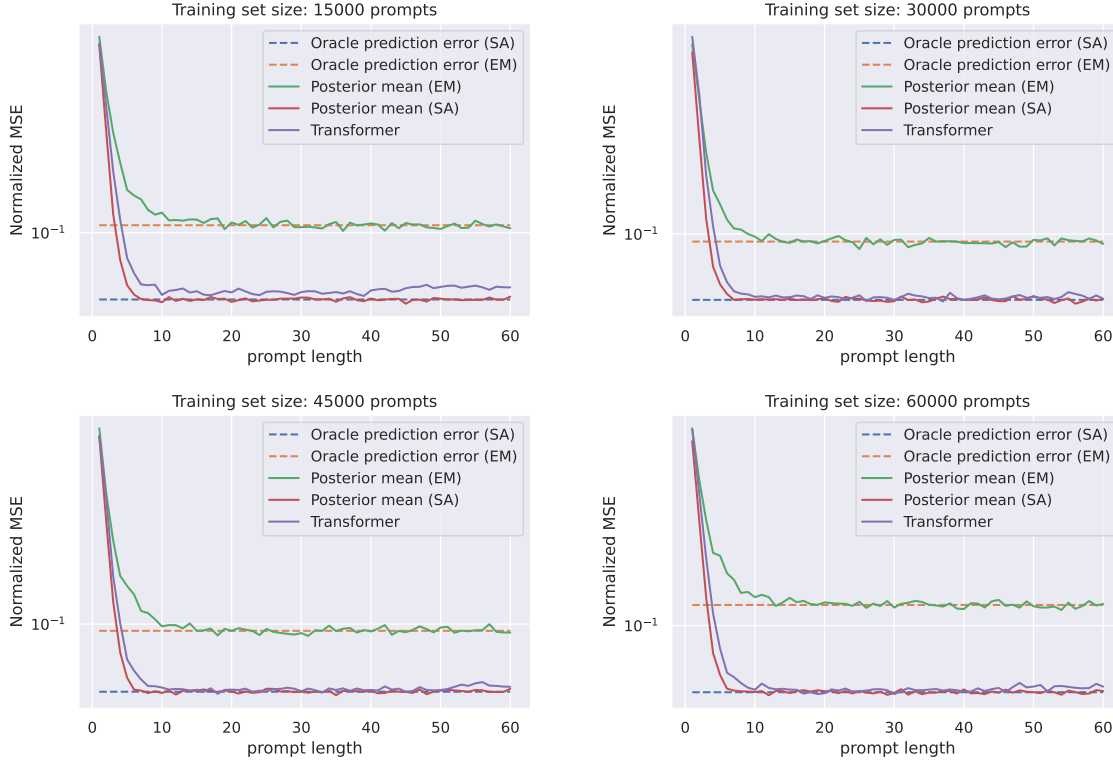


Figure 3. Comparison of EM, to subspace algorithm (SA) in [Jain et al., 2023], and Transformers.

by an algorithm. Before normalization by the dimension, it is the noise level plus

$$\frac{1}{m} \sum_{j=1}^m \min_{w \in \hat{\mathcal{W}}} \|w_j^* - w\|_2^2,$$

which is easily verified to be the prediction error with oracle knowledge of the nearest element in $\hat{\mathcal{W}}$ to the true component mean w_j^* , under our observational model (2). The main take-away from this simulation is that the transformer is able to get very close to the performance of the state-of-the-art model-specific algorithms, even when keeping the sample size the same.

3.3 What is the transformer actually learning?

In this section, we try to understand somewhat better, what algorithm the transformer is implementing at inference time. To do this, we define the squared distance, for two algorithms f, g that map a prompt P of length k to a prediction \hat{y}_{k+1} of x_{k+1} :

$$d_k^{\text{sq}}(f, g) := \mathbf{E}_P \left[(f(P) - g(P))^2 \right], \quad \text{where } k \geq 1.$$

Figure 4 depicts k versus $d_k^{\text{sq}}(f, g)$, taking f to be the transformer, and g to be a candidate algorithm listed below, as k varies between 1 and 60. The compared algorithms are:

- *Posterior mean, oracle weights.* Outputs $f_\pi^*(P)$ on prompt P .

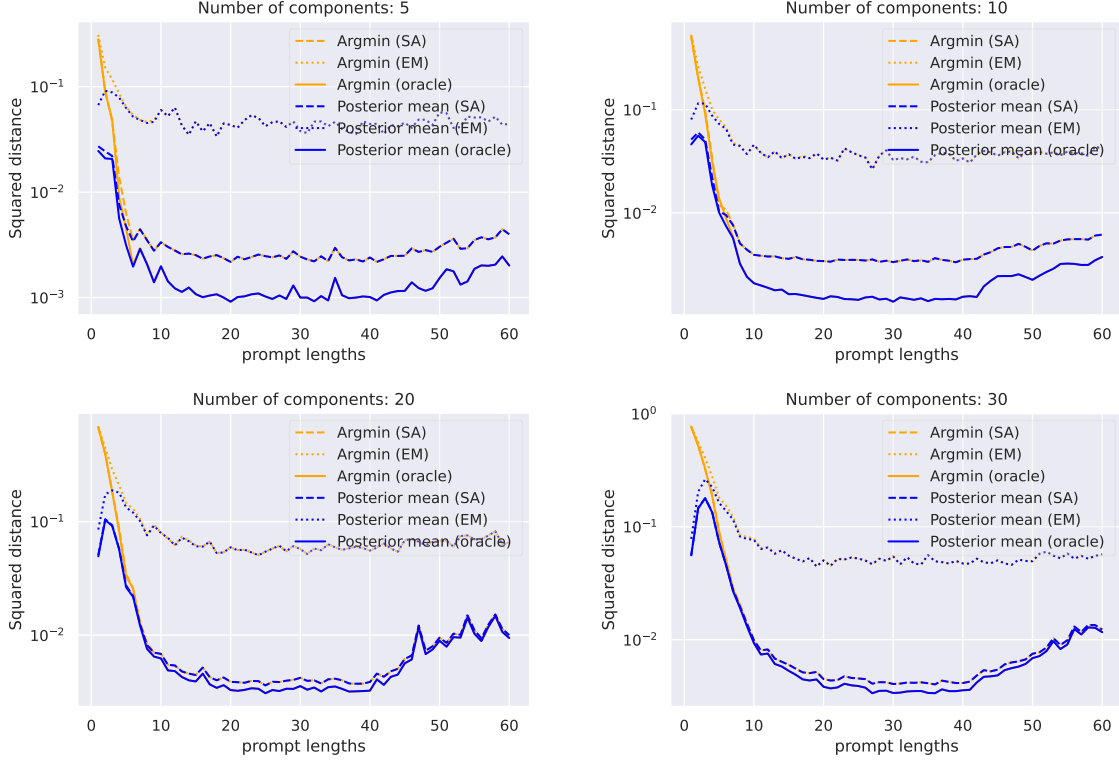


Figure 4. Comparing inference-time predictions from transformer versus posterior mean and argmin procedures with oracle or estimated weights. In these simulations, the noise level is set as $\sigma = 1.0$.

- *Posterior mean, SA weights.* Outputs $f_{\hat{\pi}}^*(P)$, with $\hat{\pi}$ from the subspace algorithm (SA).
- *Posterior mean, EM weights.* Same as above, but $\hat{\pi}$ from expectation-maximization (EM).
- *Argmin, oracle weights.* Outputs $\hat{w}(P)^\top x_{k+1}$ where $\hat{w}(P)$ follows display (7).
- *Argmin, SA weights.* Outputs $\hat{w}(P)^\top x_{k+1}$ where $\hat{w}(P)$ follows display (7), with w_j^* replaced by SA-estimated weights.
- *Argmin, EM weights.* Same as above, but with EM-estimated weights.

As seen from Figure 4, in all of the simulated settings, the algorithm closest to the transformer at inference time is the posterior mean procedure, with the oracle weights. Impressively, this observation holds regardless of our choice of the number of mixture components.

3.4 Evaluation on covariate shift and label shift

In this section, we evaluate transformers on distribution shift settings. The experimental results are presented in Figures 5 and 6. The distribution shift settings are described below, where we studied one setting of covariate shift and two settings of label shift.

Covariate scaling: We evaluate the transformer on prompts of length k where the covariates (including the query) are sampled as $x_i \sim \mathcal{N}(0, \kappa^2 I_d)$ for $i \in [k + 1]$. This is a shift from the training distribution when $\kappa \neq 1$. Figure 5 shows the results when taking $\kappa \in \{0.33, 0.5, 1, 2, 3\}$.

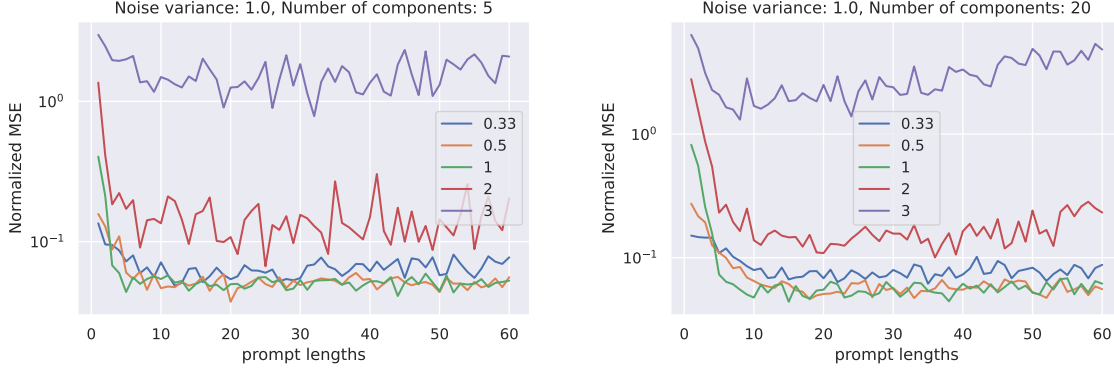


Figure 5. Evaluating transformer performance on covariate shifts.

As we see from the figure, the transformer is able to handle, to some extent, small shifts, such as $\kappa \in \{0.33, 0.5, 2\}$, but not shifts much larger than this.

Weight scaling: We evaluate the transformer on prompts sampled from the mixture distribution

$$\pi_{\alpha}^{\text{scale}} := \frac{1}{m} \sum_{i=1}^m \delta_{\alpha w_j^*}, \quad \text{where } \alpha > 0.$$

So, the weights w_j^* are scaled up or down by the factor α . Note that $\pi_1^{\text{scale}} = \pi$, meaning that $\alpha = 1$ is no shift. The left panels of Figure 6 depict results for $\alpha \in \{0.33, 0.5, 1.0, 2, 3\}$.

Weight shift: We evaluate the transformer on weights sampled from the mixture distribution

$$\pi_{\varepsilon}^{\text{add}} := \frac{1}{m} \sum_{i=1}^m \delta_{w_j(\varepsilon)}, \quad \text{where } w_j(\varepsilon) := w_j^* + \frac{\varepsilon}{\sqrt{d}} \mathbf{1}_d.$$

Thus, $\pi_{\varepsilon}^{\text{add}}$ shifts each component by an additive perturbation of norm ε . The right panels of Figure 6 depict the results for $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1.0\}$. Note that $\varepsilon = 0$ is no shift: $\pi_0^{\text{add}} = \pi$.

As seen from above, the transformer is fairly sensitive to weight scaling, as seen from the left panels in Figure 6. On the other hand, the transformer can handle small additive weight shifts, such as $\varepsilon = 0.25$, as depicted in the right panels in Figure 6.

Comparison to posterior mean procedure: In Appendix D, we replicate the figures above, with the change that in place of the transformer, we evaluate the performance of the posterior mean procedure, f_{π}^* , defined in display (4). At a high-level, the posterior mean algorithm is less sensitive to covariate scaling, but exhibits similar behavior to the transformer on the two label shift settings.

4 Discussion

In this work, we studied the behavior of transformers on mixtures of linear regressions, and showed that they can learn these mixture models near-optimally, sample-efficiently, and somewhat robustly. The fact that transformers—which, importantly, are general purpose prediction

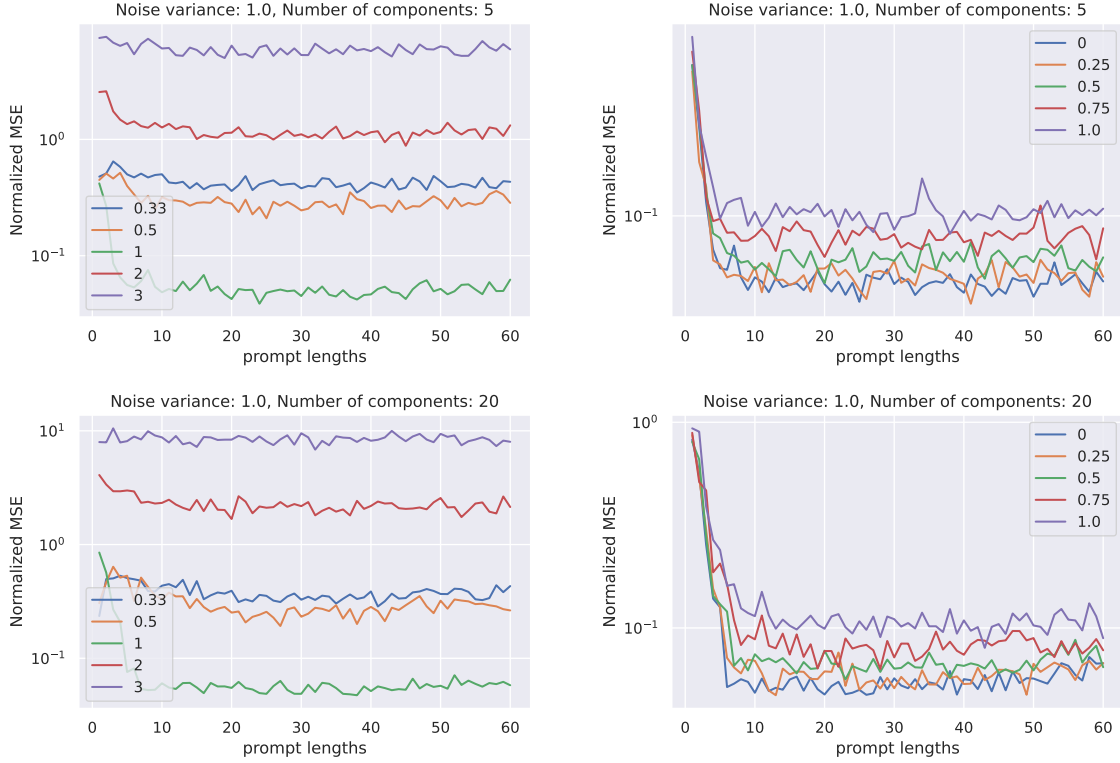


Figure 6. Evaluating transformers on weight scaling (left panels) and weight shift (right panels).

methods—can perform well in this statistically-complex mixture setting could be quite useful for practical problems, where it may be undesirable to use methods requiring a well-specified mixture model.

Additionally, our empirical and theoretical observations also suggest some lines of future investigation. For instance, in this work we have studied mixtures of linear regressions. However, in practice, the regression function within each component could potentially be *nonlinear*. To what extent do transformers perform well in these settings? Additionally, it would be interesting to study the in-context problem as was done in Garg et al. [2022], but in the mixture setting. Here, the mixture distribution would be sampled from a distribution *over* mixture models for each prompt. In general, the decision-theoretic optimal method could be more complicated to compute, as implementing the posterior mean would require computing a high-dimensional integral. Nonetheless, is it possible to approximate the optimal method with a trained transformer? We view this as an intriguing direction for future work.

References

- K. Ahuja, M. Panwar, and N. Goyal. In-context learning through the bayesian prism. *arXiv preprint arXiv:2306.04891*, 2023.
- E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.
- L. J. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. URL <http://arxiv.org/abs/1607.06450>.
- C. M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006. ISBN 978-0387-31073-2; 0-387-31073-8.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- S. Chen, J. Li, and Z. Song. Learning mixtures of linear regressions in subexponential time via Fourier moments. In *STOC*. <https://arxiv.org/pdf/1912.07629.pdf>, 2020.
- A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- S. Garg, D. Tsipras, P. S. Liang, and G. Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.
- K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.
- D. Hendrycks and K. Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL <http://arxiv.org/abs/1606.08415>.
- A. Jain, R. Sen, W. Kong, A. Das, and A. Orlitsky. Linear regression using heterogeneous data batches. *arXiv preprint arXiv:2309.01973*, 2023.
- J. Kleinberg and M. Sandler. Using mixture models for collaborative filtering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 569–578, 2004.
- W. Kong, R. Somani, Z. Song, S. Kakade, and S. Oh. Meta-learning for mixed linear regression. In *International Conference on Machine Learning*, pages 5394–5404. PMLR, 2020.
- Y. Li and Y. Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *COLT*. arXiv preprint arXiv:1802.07895, 2018.
- Y. Li, M. E. Ildiz, D. Papailiopoulos, and S. Oymak. Transformers as algorithms: Generalization and stability in in-context learning. 2023.

- S. Müller, N. Hollmann, S. P. Arango, J. Grabocka, and F. Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- J. Steinhardt, G. Valiant, and M. Charikar. Avoiding imposters and delinquents: Adversarial crowdsourcing and peer prediction. *Advances in Neural Information Processing Systems*, 29, 2016.
- K. M. Ting, B. T. Low, and I. H. Witten. Learning from batched data: Model combination versus data combination. *Knowledge and Information Systems*, 1:83–106, 1999.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.
- J. Von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.
- J. Wang, A. P. De Vries, and M. J. Reinders. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 501–508, 2006.
- J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.
- X. Yi, C. Caramanis, and S. Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621. PMLR, 2014.
- X. Yi, C. Caramanis, and S. Sanghavi. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016.
- R. Zhang, S. Frei, and P. L. Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.
- K. Zhong, P. Jain, and I. S. Dhillon. Mixed linear regression with multiple components. In *Advances in neural information processing systems (NIPS)*, pages 2190–2198, 2016.

A Related Work

The related work can be broadly divided into two categories: (i) theoretical works on learning mixture models and (ii) analyzing theoretically and empirically the learning abilities of transformers.

In the context of (i), there are numerous works that study the well known mixed linear regression problem with a batch size of 1 [Vempala and Wang, 2004, Yi et al., 2014, 2016, Chen et al., 2020, Li and Liang, 2018, Zhong et al., 2016]. In general the problem is NP-Hard as shown in [Yi et al., 2016]. Therefore most of the above works with the exception of [Li and Liang, 2018] makes the assumption that the covariates of all the mixture components are isotropic Gaussians. However, even with this strong assumption the time-complexity of all these algorithms are at least super-polynomial in m rendering them impractical.

Kong et al. [2020] pioneered the study of the problem in batch setting where they were motivated by meta-learning multiple tasks. They showed that they can recover well separated mixture of linear models from batched data with polynomial dependence on d, m and the inverse of the fraction of the smallest mixture component. However, this work still had the isotropic covariate assumption. Recent work [Jain et al., 2023] removed this assumption and further improved the sample complexity and the length of the medium size batches that is required for learning. We compare the training sample complexity of learning using transformers with that of the latter, as well as the popular EM method [Zhong et al., 2016], modified to work with batched data.

In the context of (ii), following the emergence of several hundred billion parameter large language models (LLM)’s like [Radford et al., 2019, Chowdhery et al., 2022], it has been observed that such models can learn from few examples supplied in a prompt during inference [Brown et al., 2020]. This ability to learn in-context has been studied in simpler settings in many recent works [Garg et al., 2022, Von Oswald et al., 2023, Akyürek et al., 2023, Zhang et al., 2023]. [Garg et al., 2022] showed empirically that transformers can learn to perform linear regression in context. [Akyürek et al., 2023] then showed that transformers can represent gradient decent for linear regression in context. A similar result was shown in [Von Oswald et al., 2023] but using linear self attention. Zhang et al. [2023] go one-step further by showing that gradient flow in linear self-attention based transformers can learn to do population gradient decent for linear regression. More general algorithm learning behavior has been demonstrated in [Li et al., 2023] and they also provide stability bounds for in-context learning.

Note that none of these prior works imply ability of transformers to learn mixture models from batch or non batch setting. Müller et al. [2021], Ahuja et al. [2023] look at in context learning from a Bayesian perspective. Müller et al. [2021] show that transformers fitted on the respective prior can emulate Gaussian processes. [Ahuja et al., 2023] has a section on learning multiple function classes in-context where they empirically study gaussian mixture models with two mixture components. However, they do not study the representation learning problem and training sample complexity is not investigated in depth.

B Proof of Theorem 1

In this section, we present the proof of Theorem 1. We begin, in Section B.1 by stating some preliminaries, such as the necessary operators we need to show that the transformer can implement. We then present the proof, assuming that these operators are transformer-

representable in Section B.2. Finally, the proof of the representation capacity of these operators by transformers is provided in Section B.3.

B.1 Operators that a transformer can implement

We now list some operators, for a matrix $H \in \mathbf{R}^{p \times q}$ that output a matrix $H' \in \mathbf{R}^{p \times q}$. The following list includes all the operators we need.

- **copy_down**($H; k, k', \ell, \mathcal{I}$): For columns with index $i \in \mathcal{I}$, outputs H' where $H'_{k':\ell',i} = H_{k:\ell,i}$, and the remaining entries are unchanged. Here, $\ell' = k' + (\ell - k)$ and $k' \geq k$, so that entries are copied “down” within columns $i \in \mathcal{I}$. Note, we assume $\ell \geq k$ and that $k' \leq q$ so that the operator is well-defined.
- **copy_over**($H; k, k', \ell, \mathcal{I}$): For columns with index $i \in \mathcal{I}$, outputs H' with $H'_{k':\ell',i} = H_{k:\ell,i-1}$. The remaining entries stay the same. Here entries from column $i - 1$ are copied “over” to column i .
- **mul**($H; k, k', k'', \ell, \mathcal{I}$): For columns with index $i \in \mathcal{I}$, outputs H' where

$$H'_{k''+t,i} = H_{k+t,i} H_{k'+t,i}, \quad \text{for } t \in \{0, \dots, \ell - k\}.$$

for $t \in [k, \ell]$. The remaining entries stay the same.

- **aff**($H; k, k', k'', \ell, \ell', \ell'', W, W', b, \mathcal{I}$): For columns with index $i \in \mathcal{I}$, outputs H' where

$$H'_{k'':\ell'',i} = W H_{k:\ell,i} + W' H_{k':\ell',i} + b.$$

Note that $\ell'' = k'' + \delta''$ where $W \in \mathbf{R}^{\delta'' \times \delta}$, $W' \in \mathbf{R}^{\delta'' \times \delta'}$ and $\ell = k + \delta$, $\ell' = k' + \delta'$. We assume $\delta, \delta', \delta'' \geq 0$. The remaining entries of H are copied over to H' , unchanged.

- **scaled_agg**($H; \alpha, k, \ell, k', i, \mathcal{I}$): Outputs a matrix H' with entries

$$H'_{k'+t,i} = \alpha \sum_{j \in \mathcal{I}} H_{k+t,j} \quad \text{for } t \in \{0, 1, \dots, \ell - k\}.$$

The set \mathcal{I} is causal, so that $\mathcal{I} \subset [i - 1]$. The remaining entries of H are copied over to H' , unchanged.

- **soft**($H; k, \ell, k'$): For the final column q , outputs a matrix H' with entries

$$H'_{k'+t,q} = \frac{e^{H_{k+t,q}}}{\sum_{t'=0}^{\ell-k} e^{H_{k+t',q}}}, \quad \text{for } t \in \{0, 1, \dots, \ell - k\}.$$

The remaining entries of H are copied over to H' , unchanged.

The important property of the above list of operators is that can all be implemented in a single layer of a autoregressive transformer.

Proposition 1. *Each of the operators **copy_down**, **copy_over**, **mul**, **aff**, **scaled_agg**, and **soft**, can be implemented by a single layer of an autoregressive transformer.*

See Section B.3 for a proof of this claim.

B.2 Proof of Theorem 1

In this section, we present the proof of Theorem 1, assuming Proposition 1. We need to introduce a bit of notation:

$$\mathcal{I}_{\text{even}}(k) := \{2j : j \in [k]\} \quad \text{and} \quad \mathcal{I}_{\text{odd}}(k) := \{2j - 1 : j \in [k + 1]\}.$$

Additionally, we define $\mathcal{W}^* \in \mathbf{R}^{m \times d}$ to have rows $w_i^* \in \mathbf{R}^d$, which as we recall from (1), are the true mixture weights.

We begin by assuming that the input prompt P is provided as $H^{(0)} \in \mathbf{R}^{(2d+4m+2) \times (2k+1)}$. This matrix is such that the only nonzero entries are $H_{1:d,2j-1}^{(0)} = x_j \in \mathbf{R}^d$ for each $j \in [k + 1]$. Additionally, $H_{1,2j}^{(0)} = y_j$ for each $j \in [k]$. Then, by leveraging the operators described above, we can see that $f^*(P) = H_{2d+4m+2,2k+1}^{(9)}$, where the matrix $H^{(8)}$ is constructed by the following process:

- $H^{(1)} = \text{copy_down}(H^{(0)}; 1, d + 1, d, \mathcal{I}_{\text{odd}}(k))$
- $H^{(2)} = \text{copy_over}(H^{(1)}; d + 1, d + 1, 2d, \mathcal{I}_{\text{even}}(k))$
- $H^{(3)} = \text{copy_down}(H^{(2)}; 1, 2d + 1, 1, \mathcal{I}_{\text{odd}}(k))$
- $H^{(4)} = \text{aff}(H^{(3)}; d + 1, 2d + 1, 2d + 2, 2d, 2d + 1, 2d + m + 1, \mathcal{W}^*, \mathbf{1}_{d \times 1}, 0, \mathcal{I}_{\text{even}}(k) \cup \{2k + 1\})$
- $H^{(5)} = \text{mul}(H^{(4)}; 2d + 2, 2d + 2, 2d + m + 2, 2d + m + 1, \mathcal{I}_{\text{even}}(k))$
- $H^{(6)} = \text{scaled_agg}(H^{(5)}; -\frac{1}{2\sigma^2}, 2d + m + 2, 2d + 2m + 1, 2d + m + 2, 2k + 1, \mathcal{I}_{\text{even}}(k))$
- $H^{(7)} = \text{soft}(H^{(6)}; 2d + m + 2, 2d + 2m + 1, 2d + 2m + 2)$
- $H^{(8)} = \text{mul}(H^{(7)}; 2d + 2, 2d + 2m + 2, 2d + 3m + 2, 2d + m + 1, \{2k + 1\})$
- $H^{(9)} = \text{aff}(H^{(8)}; 2d + 3m + 2, d + 1, 2d + 4m + 1, 2d + 2m, d + 1, 2d + 4m + 2, \mathbf{1}_m, 0, 0, \{2k + 1\})$

The process above is illustrated in Section B.2.1. By Proposition 1, each operation above is implementable by a layer of an autoregressive transformer. Therefore, this completes the proof.

B.2.1 Illustration of proof of Theorem 1

We illustrate the steps taken by the transformer to implement the softmax operation. To begin with, the matrix input to the transformer is modelled as below, in the case where $k = 2$. Below, $\tilde{y}_i = (y_1, 0, \dots, 0) \in \mathbf{R}^d$. Throughout we only show the nonzero entries (*i.e.*, missing rows and columns are always assumed 0). Then, our input is

$$H^{(0)} = \begin{bmatrix} x_1 & \tilde{y}_1 & x_2 & \tilde{y}_2 & x_3 \end{bmatrix}.$$

After the `copy_down` operation, we have

$$H^{(1)} = \begin{bmatrix} x_1 & \tilde{y}_1 & x_2 & \tilde{y}_2 & x_3 \\ x_1 & 0 & x_2 & 0 & x_3 \end{bmatrix}.$$

After the `copy_over` operation, we have

$$H^{(2)} = \begin{bmatrix} x_1 & \tilde{y}_1 & x_2 & \tilde{y}_2 & x_3 \\ x_1 & x_1 & x_2 & x_2 & x_3 \end{bmatrix}.$$

After another `copy_down` operation, we have

$$H^{(3)} = \begin{bmatrix} x_1 & \tilde{y}_1 & x_2 & \tilde{y}_2 & x_3 \\ x_1 & x_1 & x_2 & x_2 & x_3 \\ 0 & y_1 & 0 & y_2 & 0 \end{bmatrix}.$$

After the `aff` operation, we have,

$$H^{(4)} = \begin{bmatrix} x_1 & \tilde{y}_1 & x_2 & \tilde{y}_2 & x_3 \\ x_1 & x_1 & x_2 & x_2 & x_3 \\ 0 & y_1 & 0 & y_2 & 0 \\ 0 & r_1 & 0 & r_2 & r_3 \end{bmatrix}.$$

Note that $r_i = \mathcal{W}^* x_i - y_i \mathbf{1}$ for $i \neq k+1$ and otherwise $r_{k+1} = \mathcal{W}^* x_{k+1}$. After the `mul` operation, we obtain

$$H^{(5)} = \begin{bmatrix} x_1 & \tilde{y}_1 & x_2 & \tilde{y}_2 & x_3 \\ x_1 & x_1 & x_2 & x_2 & x_3 \\ 0 & y_1 & 0 & y_2 & 0 \\ 0 & r_1 & 0 & r_2 & r_3 \\ 0 & r_1^2 & 0 & r_2^2 & 0 \end{bmatrix}$$

Above the square should be interpreted element wise on the vectors r_i . Then, after the `scaled_agg` operation, we obtain

$$H^{(6)} = \begin{bmatrix} x_1 & \tilde{y}_1 & x_2 & \tilde{y}_2 & x_3 \\ x_1 & x_1 & x_2 & x_2 & x_3 \\ 0 & y_1 & 0 & y_2 & 0 \\ 0 & r_1 & 0 & r_2 & r_3 \\ 0 & r_1^2 & 0 & r_2^2 & -\frac{1}{2\sigma^2}(r_1^2 + r_2^2) \end{bmatrix}.$$

Then, after the `softmax` operation, we obtain

$$H^{(7)} = \begin{bmatrix} x_1 & \tilde{y}_1 & x_2 & \tilde{y}_2 & x_3 \\ x_1 & x_1 & x_2 & x_2 & x_3 \\ 0 & y_1 & 0 & y_2 & 0 \\ 0 & r_1 & 0 & r_2 & r_3 \\ 0 & r_1^2 & 0 & r_2^2 & -\frac{1}{2\sigma^2}(r_1^2 + r_2^2) \\ 0 & 0 & 0 & 0 & p \end{bmatrix}.$$

Here, $p = \text{softmax}(-\frac{1}{2\sigma^2}(r_1^2 + r_2^2))$. Finally, after yet another `mul`, we obtain

$$H^{(8)} = \begin{bmatrix} x_1 & \tilde{y}_1 & x_2 & \tilde{y}_2 & x_3 \\ x_1 & x_1 & x_2 & x_2 & x_3 \\ 0 & y_1 & 0 & y_2 & 0 \\ 0 & r_1 & 0 & r_2 & r_3 \\ 0 & r_1^2 & 0 & r_2^2 & -\frac{1}{2\sigma^2}(r_1^2 + r_2^2) \\ 0 & 0 & 0 & 0 & p \\ 0 & 0 & 0 & 0 & p \circ r_3 \end{bmatrix}.$$

Above, \circ denotes elementwise multiplication. Finally, after an **aff** operation, we obtain

$$H^{(9)} = \begin{bmatrix} x_1 & \tilde{y}_1 & x_2 & \tilde{y}_2 & x_3 \\ x_1 & x_1 & x_2 & x_2 & x_3 \\ 0 & y_1 & 0 & y_2 & 0 \\ 0 & r_1 & 0 & r_2 & r_3 \\ 0 & r_1^2 & 0 & r_2^2 & -\frac{1}{2\sigma^2}(r_1^2 + r_2^2) \\ 0 & 0 & 0 & 0 & p \\ 0 & 0 & 0 & 0 & p \circ r_3 \\ 0 & 0 & 0 & 0 & f_\pi^\star(P) \end{bmatrix}.$$

Note that the bottom-right entry of $H^{(9)}$ contains the output $f_\pi^\star(P)$. In other words, the desired result is $H_{2d+4m+2,2k+1}^{(9)}$. Note that $H \in \mathbf{R}^{p \times q}$ where $p = 2d + 4m + 2$, and $q = 2k + 1$.

B.3 Proof of Proposition 1

To begin with, we recall a few definitions, introduced in recent work [Akyürek et al., 2023].

Definition 1 (RAW operator). The Read-Arithmetic-Write (RAW) operators are maps on matrices, $\mathbf{R}^{p \times q} \rightarrow \mathbf{R}^{p \times q}$,

$$\text{RAW}_\bullet(H; \mathcal{I}, \mathcal{J}, \mathcal{K}, \Theta_{\mathcal{I}}, \Theta_{\mathcal{J}}, \Theta_{\mathcal{K}}, \pi) = H', \quad \text{where } \bullet \in \{\otimes, \oplus\}.$$

Here π is a causal set-valued map, with $\pi(i) \subset [i - 1]$. The operators \otimes, \oplus denote elementwise multiplication and addition, respectively. The entries of H' are given by

$$H'_{\mathcal{K},i} := \Theta_{\mathcal{K}} \left(\Theta_{\mathcal{J}} H_{\mathcal{J},i} \bullet \left(\frac{\Theta_{\mathcal{I}}}{\max\{|\pi(i)|, 1\}} \sum_{i' \in \pi(i)} H_{\mathcal{I},i'} \right) \right), \quad \text{and} \quad (8a)$$

$$H'_{\mathcal{K}^c,i} = H_{\mathcal{K}^c,i}, \quad (8b)$$

for each $i \in [q]$. Note that above $\mathcal{K}^c = [p] \setminus \mathcal{K}$, and for some positive integer r , $\Theta_{\mathcal{I}} \in \mathbf{R}^{r \times |\mathcal{I}|}$, $\Theta_{\mathcal{J}} \in \mathbf{R}^{r \times |\mathcal{J}|}$, and $\Theta_{\mathcal{K}} \in \mathbf{R}^{|\mathcal{K}| \times r}$.

In Akyürek et al. [2023], they show that the RAW operator can be implemented in one autoregressive transformer layer. They also argue that (with a slight change in parameterization) that the **mul** and **aff** operators are transformer-implementable. Therefore, we simply need to argue that the operators **soft**, **copy_down**, **copy_over**, and **scaled_agg** are all implementable by a transformer.

To begin with, note that, by inspection, we have, with $\delta = \ell - k$,

$$\text{copy_over}(H; k, k', \ell, \mathcal{I}) = \text{RAW}_{\oplus}(H; [k, \ell], \emptyset, [k', k' + \delta], I_{\delta+1}, 0, I_{\delta+1}, \pi_{\mathcal{I}}) \quad (9a)$$

$$\text{copy_down}(H; k, k', \ell, \mathcal{I}) = \text{RAW}_{\oplus}(H; [k, \ell], \emptyset, [k', k' + \delta], I_{\delta+1}, 0, I_{\delta+1}, \pi'_{\mathcal{I}}) \quad (9b)$$

$$\text{scaled_agg}(H; \alpha, k, \ell, k', i, \mathcal{I}) = \text{RAW}_{\oplus}(H; [k, \ell], \emptyset, [k', k' + \delta], I_{\delta+1}, 0, \alpha I_{\delta+1}, \pi''_{\mathcal{I},i}) \quad (9c)$$

Above, note that the intervals $[a, b]$ are just the integers between a and b (inclusive) and that we have defined

$$\pi_{\mathcal{I}}(i) = \begin{cases} \{i - 1\} & i \geq 2 \\ \emptyset & \text{otherwise} \end{cases}, \quad \pi'_{\mathcal{I}}(i) = \begin{cases} \{i\} & i \in \mathcal{I} \\ \emptyset & \text{otherwise} \end{cases}, \quad \text{and} \quad \pi''_{\mathcal{I},i}(j) = \begin{cases} \mathcal{I} & j = i \\ \emptyset & \text{otherwise} \end{cases}.$$

Therefore, the displays (9) establish the following result.

Lemma 1. *The operators `copy_over`, `copy_down`, and `scaled_agg` are all implementable via the RAW operator in a single autoregressive transformer layer.*

Finally, in Section B.3.1 we demonstrate the following result.

Lemma 2. *The softmax operation is implementable by an autoregressive transformer.*

This completes the proof of Proposition 1.

B.3.1 Proof of Lemma 2

In order to implement the softmax operation, we need to introduce a few other operations:

- $\text{div}(H; j, k, k', \ell, \mathcal{I})$: For columns with index $i \in \mathcal{I}$, outputs H' where $H'_{k'+t,i} = H_{k+t,i}/H_{j,i}$ for all $t \in \{0, \dots, \ell - k\}$. The remaining entries of H are copied as is into H' .
- $\text{mov}(H; k, k', \ell, \mathcal{I})$: For columns with index $i \in \mathcal{I}$, outputs H' where $H'_{k'+t,i} = H_{k+t,i}$ for all $t \in \{0, \dots, \ell - k\}$. The remaining entries of H are copied as is into H' .
- $\text{sigmoid}(H; k, k')$: In the final column q , outputs H' with $H'_{k',q} = \frac{1}{1+e^{-H_{k,q}}}$. The remaining entries of H are copied as is into H' .

The operations `div`, `mov` are special cases of the same operations as introduced in the paper Akyürek et al. [2023]. Thus, we only need to demonstrate that `sigmoid` is transformer-implementable. Assuming this for the moment, note that the softmax operation `softmax` is then implementable by the following sequence of operations. Let H denote the input to the softmax layer, and let $s = H_{k:\ell,q}$. Using the affine operation (`aff`) together with the softmax operation (`soft`) we can compute the values $1/(1+e^{s_i})$. Using the affine operation (`aff`) together with the `div` operation, we can invert these values to compute e^{s_i} . Finally, we can compute the sum of these values $S = \sum_i e^{s_i}$ with an affine operation (`aff`) and we can divide by this sum using another `div` operation. The result values are $e^{s_i}/\sum_j e^{s_j}$, which is the softmax of the vector s . A move operation (`mov`) then can move these values into the correct locations, $H_{k':(k'+\ell-k),q}$, as required.

Thus, to complete the proof, we need to show how to implement the sigmoid operation. For this, we can begin by using the affine operation to insert a value of 1 in the final column, and another affine operation to insert a 2×2 identity matrix in the first 2 columns of H . Then by selecting W^K and W^Q to select the identity matrix and to select $(H_{k,i}, 1)$, respectively, we can ensure that $W^K H_{:,i}$ is a $2 \times i - 1$ matrix with alternating columns $(1, 0), (0, 1)$. We also have $W^Q h_i = (H_{k,i}, 1)$. The corresponding softmax values in the self-attention layer are $s_i = (e^{H_{k+t,i}}/(1 + H_{k,i}), 1/(e^{H_{k,i}} + 1))$. By selecting W^V to select the identity matrix in $H_{:,i}$, and W^C to select the first value of s_i and place it in position $k' + t$, we can ensure that $a = 1/(e^{-H_{k,i}} + 1)e_{k'}$, where e_j denotes the j th standard basis vector. This value is precisely the sigmoid, as needed. To place this value in the correct location, we simply set the feedforward network matrices $W^{\text{in}}, W^{\text{out}} = 0$. Then, to preserve the output, we need to delete the value 1, and identity matrices placed into H at the beginning; this can clearly be done by two affine operations.

C Additional details on training methodology

Our training approach closely follows that of Garg et al. [2022] and Akyürek et al. [2023]. After some hyperparameter optimization, we settled on the choice of hidden dimension of 256, 8 attention heads, and 12 layers. We trained our transformers using Adam, with a constant step size of 0.1. We used curriculum training, as in Garg et al. [2022], with the exception of Figure 3, where the sample size was fixed. Our curriculum phases were 2000 steps each, with a batch size of 64. The final stage of training had 250000 steps with 64 batches. All of our figures presented mean squared errors computed over batch sizes of 256. The dimension of the original covariates was $d = 20$ throughout this paper.

C.1 Details on fixed sample-size training

In this setting, we used hyperparameter tuning over the dropout parameter, $\rho \in \{0, 0.05, 0.1\}$, and found the following choices to be best, for Figure 3:

- for $n = 15000$, we took $\rho = 0.0$.
- for $n = 30000$, we took $\rho = 0.1$.
- for $n = 45000$, we took $\rho = 0.0$.
- for $n = 60000$, we took $\rho = 0.1$.

We also used curriculum training in this setup, but obtained the samples by subsampling the fixed dataset. This was done by first randomly sampling a batch from the full dataset, and then randomly dropping and shuffling the prefix of each prompt so as to obtain a prompt of the shorter, desired length. Otherwise, the entire procedure was the same as the other figures, as described above.

C.2 Batch expectation maximization (EM) algorithm

Batch expectation-maximization is a variant of the standard expectation-maximization method (see, for instance, Section 14.5.1 in Bishop [2006]). For completeness, we describe the algorithm formally here. Note that ϕ denotes the standard univariate Gaussian pdf below. For notation, we also denote the prompts as

$$P^{(i)} := (x_1^{(i)}, y_1^{(i)}, \dots, x_k^{(i)}, y_k^{(i)}, x_{k+1}^{(i)}), \quad \text{for } i \in [n].$$

The algorithm is then stated below as Algorithm 1

Algorithm 1 Batch expectation-maximization for a discrete mixture of linear regressions with Gaussian noise

Require: Length k prompts $\{P^{(i)}\}_{i=1}^n$ noise variance $\sigma > 0$, number of components $m > 0$.

Initialize $\pi^{(0)} \in [0, 1]^m$, drawn uniformly on the probability simplex.

Initialize $w_j^{(0)} \in \mathbf{R}^d$, drawn uniformly on the sphere of radius \sqrt{d} for $j \in [m]$.

Initialize $\gamma_{ij}^{(0)} = 0$ for all $i \in [n], j \in [m]$.

while have not converged **do**

 update prompt-component assignment probabilities,

$$\gamma_{ij}^{(t+1)} = \frac{\pi_j^{(t)} \prod_{l=1}^k \phi\left(\frac{y_l^{(i)} - (x_l^{(i)})^\top w_j^{(t)}}{\sigma}\right)}{\sum_{j'=1}^m \pi_{j'}^{(t)} \prod_{l=1}^k \phi\left(\frac{y_l^{(i)} - (x_l^{(i)})^\top w_{j'}^{(t)}}{\sigma}\right)}, \quad \text{for all } i \in [n], j \in [m].$$

 update the marginal component probabilities by the formula

$$\pi_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \gamma_{ij}^{(t+1)}, \quad \text{for all } j \in [m]$$

 update the parameter estimates by solving,

$$w_j^{(t+1)} = \arg \min_{w \in \mathbf{R}^d} \left\{ \sum_{i=1}^n \sum_{l=1}^k \gamma_{ij}^{(t+1)} (y_l^{(i)} - w^\top x_l^{(i)})^2 \right\}, \quad \text{for all } j \in [m].$$

 update the iteration counter, $t \leftarrow t + 1$.

end while

return final set of component centers, $\{w_j^{(t)}\}_{j=1}^m$

In our implementation we stop (or declare the algorithm converged) if $t > t_{\max}$, or if

$$\max_j \min_{j'} \|w_j^{(t)} - w_{j'}^{(t-1)}\|_2 \leq \varepsilon.$$

In our experiments we took $t_{\max} = 20000$ and $\varepsilon = 0.001$.

D Comparison to distribution shift with the posterior mean estimator

In this section, we replicate the figures presented in Section 3.4, except we evaluate the distribution shift settings on the posterior mean procedure, f_π^* as defined in display (4).

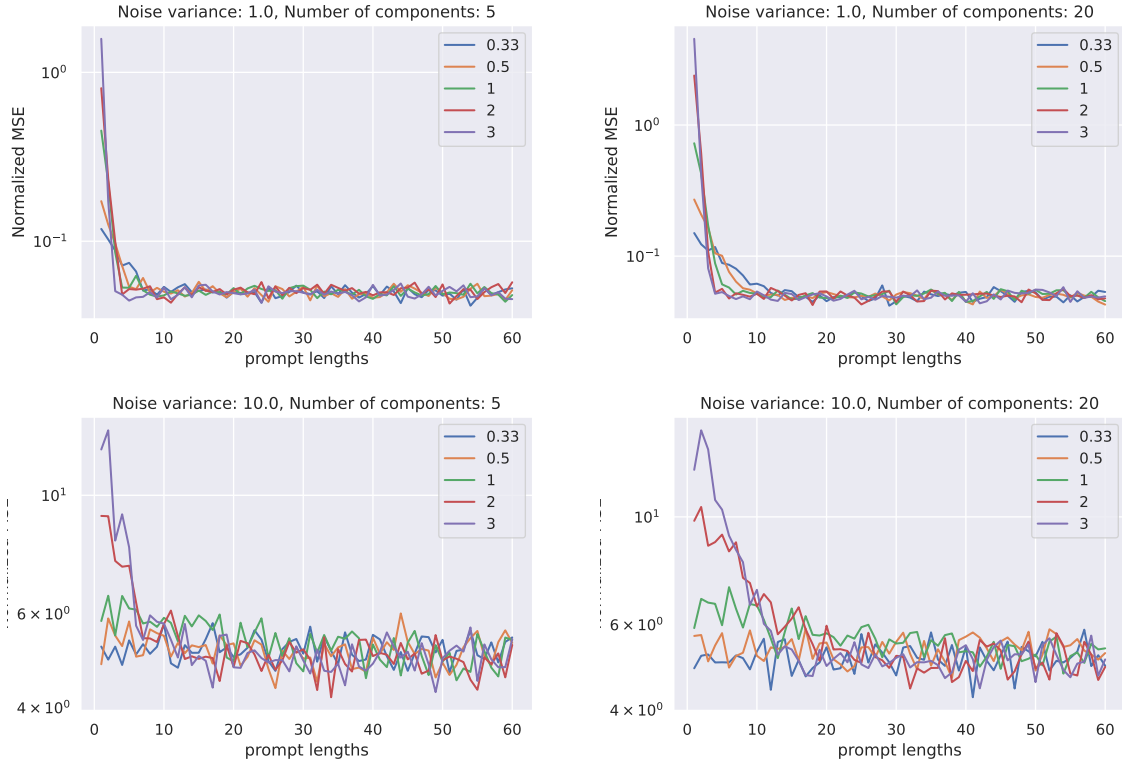


Figure 7. Posterior mean algorithm on covariate scaling distribution shift setting.

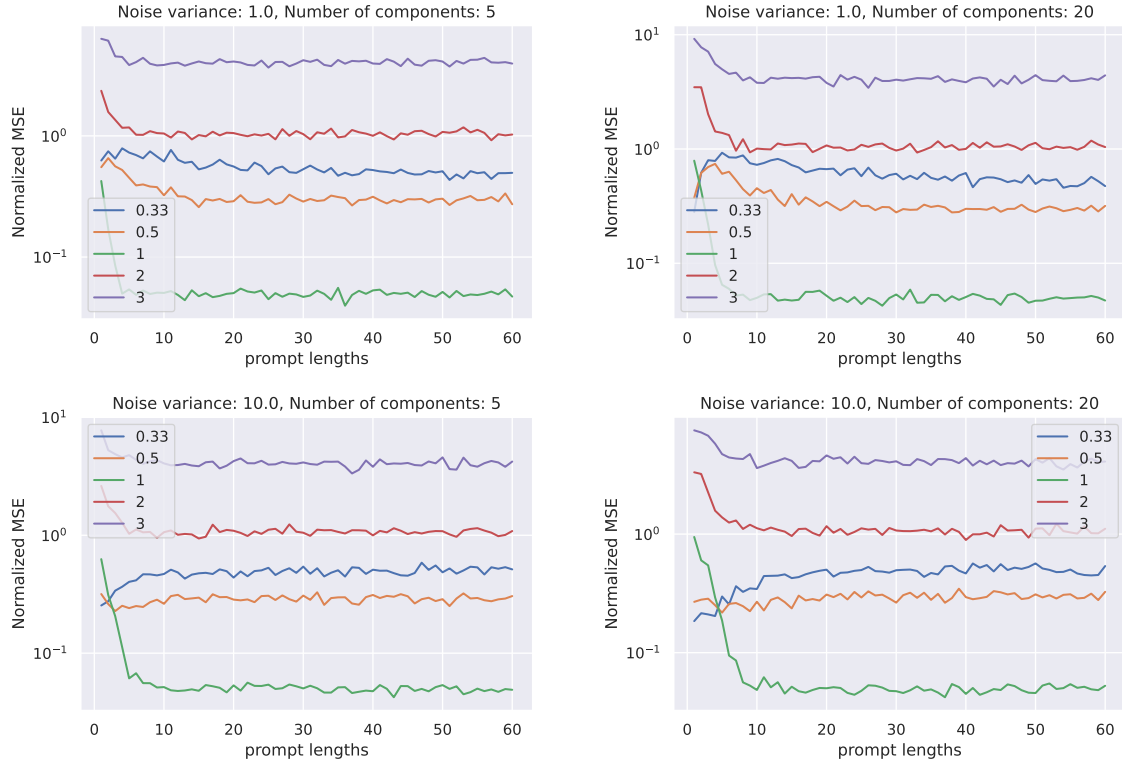


Figure 8. Posterior mean algorithm on weight scaling distribution shift setting.

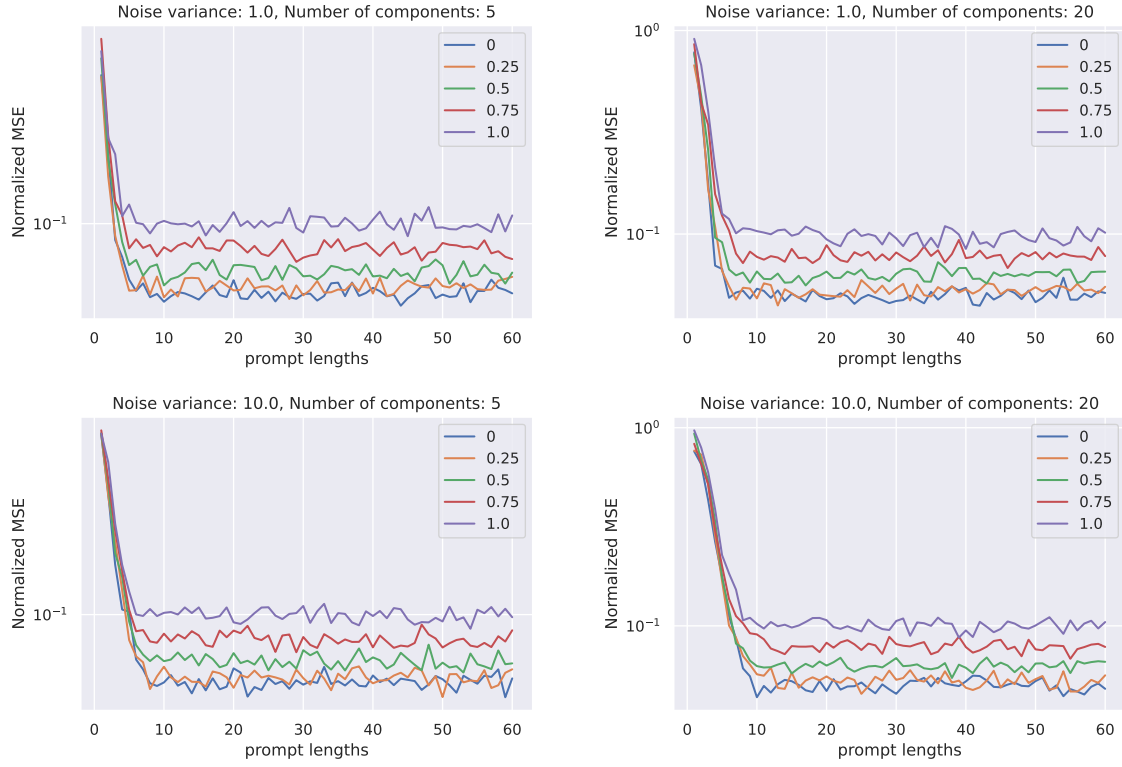


Figure 9. Posterior mean algorithm on weight additive shift setting.