

---

# Shift is Good: Mismatched Data Mixing Improves Test Performance

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 We consider training and testing on mixture distributions with different training  
2 and test proportions. We show that in many settings, and in some sense generi-  
3 cally, distribution shift can be beneficial, and test performance can improve due  
4 to mismatched training proportions. In a variety of scenarios, we identify the  
5 optimal training proportions and the extent to which such distribution shift can be  
6 beneficial.

## 7 1 Introduction

8 Imagine that you are taking a high-stakes exam next week. The exam will be 90% on European  
9 history and 10% on Chinese history. Both topics are equally familiar to you and equally difficult, and  
10 additional study will help you with each topic similarly. You have unlimited access to study material  
11 and practice questions for both. How should you spend your limited studying budget? Should your  
12 training match your test distribution, studying 90% European and 10% Chinese? Or would you  
13 benefit from a distribution shift? Studying more Chinese history? Less? Only European history? *We*  
14 *encourage the reader to pause and make an intuitive guess.*

15 The answer depends on the specific learning curve for improvement in test performance within a  
16 topic as a function of the number of training examples from that topic. But at least for a generic  $1/n$   
17 scaling (as obtained from e.g., both learning VC classes and in parametric regression), the answer,  
18 as we will see in Section 3, is that you would benefit from a distribution shift, and should study  
19 75% European History and 25% Chinese history—this would reduce your test error by 20% over the  
20 90/10 non-shifted training.

21 We just saw an example of what we term **Positive Distribution Shift**: Even if we have unlimited data  
22 from the target test distribution  $D_{\text{test}}$ , training on a shifted distribution  $D_{\text{train}} \neq D_{\text{test}}$  can actually  
23 *improve* test performance. This contrasts the typical study of *distribution shift*, i.e. training on one  
24 distribution but then applying the predictor, or testing, on another. Typically, it is implicitly assumed  
25 that the ideal case would be to train on the test distribution, that training on a different distribution  
26 is a compromise, either because we don’t know or have access to the true  $D_{\text{test}}$ , or it’s expensive  
27 to sample from it, or we have only a limited number of samples and want to supplement them with  
28 additional data from related distributions. Distribution shift is usually studied as “how much worse  
29 do things get if we train on  $D_{\text{train}} \neq D_{\text{test}}$ ”, with answers of the form “if  $D_{\text{train}}$  is close or related  
30 enough to  $D_{\text{test}}$ , then it’s not much worse”. In this paper, we investigate one of several ways in which  
31 distribution shift can be *positive*.

32 Specifically, we systematically study the benefit of such distribution shift when training with mis-  
33 matched mixing proportions relative to the test distribution. We model the test distribution as a  
34 mixture of  $K$  components, with known mixing proportions  $\{p_k\}_{k=1}^K$ , and consider training distribu-  
35 tions which are mixtures over the same components but with different mixing proportions  $\{q_k\}_{k=1}^K$ .

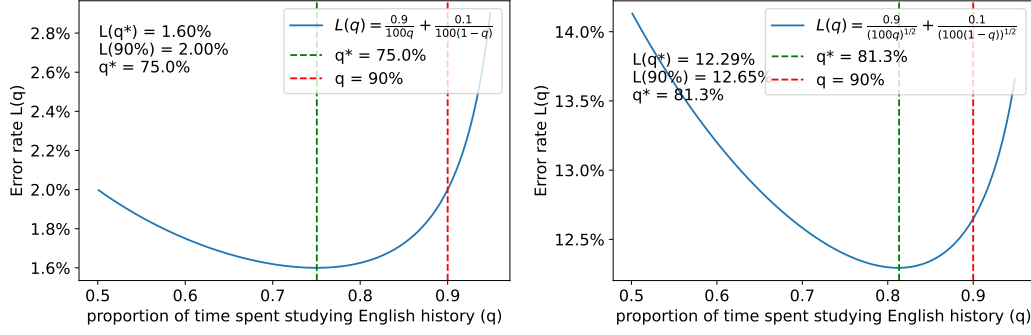


Figure 1: We plot the error rate for a hypothetical scenario modelling the high stakes exam described in Section 1. We model the error rate on each of the test portions as being proportional to  $\propto \frac{1}{n_i^\alpha}$ , where  $n_i$  represents the studying budget spent on that portion of the exam, so  $i = 1$  corresponds to European History and  $i = 2$  to the Chinese History and set  $n_1 + n_2 = N$  to be the total studying budget, with  $N = 100$  hours. The exponent  $\alpha$  is  $\alpha = 1$  on the left plot and  $\alpha = 2$  on the right plot. In both cases, we consider  $n_1 = qN$  and  $n_2 = (1 - q)N$ , where  $q$  is the proportion of time spent studying for the European History portion of the exam. This way, the error rate on the exam can be written as a function of  $q$  as  $L(q) = 0.9 \frac{1}{(100q)^\alpha} + 0.1 \frac{1}{(100(1-q))^\alpha}$ . We can see on both plots that shifting away from the testing proportion (red line, i.e.  $q = 90\%$ ) can lead to a better error rate with the optimal test proportion (green line, i.e.  $q^*$  whose values are displayed accordingly). See also Corollary 3.3.

We can either think of this as providing guidance when we can actively control mixing between different known components, or as helping us understand how and why a mismatched training distribution can actually be beneficial. In Section 5 we discuss how the analysis is also applicable to a setting where we are not testing on a mixture, but rather on compositional tasks, requiring composing multiple skills, and the skills appear with differing frequencies—this compositional setting served as a major motivation for our study.

We consider different per-component learning curves, capturing different error decays, differing hardness among the components, and the possibility of transfer between components. In Section 3 we consider power law error decay, both the  $1/n$  decay mentioned earlier and more general power laws, including with differing component hardnesses or error decays. In Section 4 we consider learning curves corresponding to “fact memorization” scenarios (discussed in Section 4), including those applicable to the skill composition setting, and which correspond to coupon-collector type learning curves. In Section 6 we consider the possibility of transfer between components. In all of these, we show that a mismatched training distribution can be beneficial, characterize the optimal training mixture, and the extent to which mismatch can improve test performance and reduce the training complexity.

Beyond all the specific scenarios, we then argue, in Section 7, that benefiting from mismatch is not the exception but rather the rule. We show that only in rare situations (either measure zero or satisfying a conservation property that does not generally hold) is the optimal training distribution equal to the test distribution, while in “most” cases shift is good.

## 2 Setup

**Learning Setup and Loss** For concreteness, let  $\ell(h, z)$  be the loss function that describes how well a model  $h$  performs on and instance  $z \in \mathcal{Z}$ . For example, in supervised learning,  $z$  can be an input-output pair  $(x, y)$ , and  $\ell(h, z)$  can be the prediction error of  $h(x)$  vs  $y$ . Or, in next-word prediction,  $z$  can be a document and  $\ell(h, z)$  can be the average cross-entropy loss when using  $h$  to predict each of the next tokens in the document. In any case, for a test distribution  $D_{\text{test}}$  over  $z$ , we evaluate the model through the *test loss*  $\mathcal{L}_{D_{\text{test}}}(h) := \mathbb{E}_{z \sim D_{\text{test}}}[\ell(h, z)]$ .

**Test Distribution.** We consider test distributions consisting of a mixture of  $K$  components  $\mathcal{D}_1, \dots, \mathcal{D}_K$ . A mixture  $\mathcal{D}_p = \sum_k p_k \mathcal{D}_k$  is then specified by mixing proportions  $p =$

( $p_1, \dots, p_K$ )  $\in \Delta_K$  on the probability simplex  $\Delta_K$ . We let  $\mathbf{p}$  be the mixing proportions in the test distribution, i.e.  $D_{\text{test}} = \mathcal{D}_{\mathbf{p}}$ , and so the test loss is  $\mathcal{L}_{\mathcal{D}_{\mathbf{p}}}(h) = \mathcal{L}_{\mathbf{p}}(h)$ , where here and elsewhere we use the subscript  $\mathbf{p}$  to denote the mixture  $\mathcal{D}_{\mathbf{p}}$ .

**Learning Algorithm.** We consider abstract “learning algorithm”  $\mathcal{A}$ , which, given training data (or sequence of training examples)  $S \in \mathcal{Z}^N$  of size  $N$ , outputs a model  $\mathcal{A}(S)$  with test loss  $\mathcal{D}_{\mathbf{p}}(\mathcal{A}(S))$ .

**Training Distribution.** We consider training on i.i.d. samples  $S \sim \mathcal{D}_{\mathbf{q}}^N$  from mixtures  $\mathcal{D}_{\mathbf{q}}$  of the same  $K$  components, but with potentially different mixing proportions  $\mathbf{q} \in \Delta_K$ . For training mixing proportions  $\mathbf{q}$ , we denote  $L_N(\mathbf{p}, \mathbf{q}) = \mathbb{E}_{S \sim \mathcal{D}_{\mathbf{q}}^N}[\mathcal{L}_{\mathbf{p}}(\mathcal{A}(S))]$  the expected test error on  $D_{\text{test}} = \mathcal{D}_{\mathbf{p}}$  when training with  $D_{\text{train}} = \mathcal{D}_{\mathbf{q}}$  (we frequently drop the subscript  $N$  if its clear from context). The “non-shifted” expected test loss is then denoted  $L_N^{\text{same}}(\mathbf{p}) = L_N(\mathbf{p}, \mathbf{p})$ . In contrast, we denote  $L_N^*(\mathbf{p}) = \min_{\mathbf{q} \in \Delta_K} L_N(\mathbf{p}, \mathbf{q})$  the test error with the best mixing ratios, and  $\mathbf{q}^*$  the minimizing ratios. When  $L^* < L^{\text{same}}$  and so  $\mathbf{q}^* \neq \mathbf{p}$ , this means we can benefit from mismatched training. **Our main analysis objective is to characterize  $\mathbf{q}^*$ ,  $L^*$  and the improvement over  $L^{\text{same}}$ .**

We can measure the mismatch benefit through the improvement in test error for a fixed training budget  $L_N^{\text{ratio}} = L_N^*/L_N^{\text{same}}$ . Or, we can consider the training complexity  $N_{\epsilon}(\mathbf{p}, \mathbf{q}) = \min N$  s.t.  $L_N(\mathbf{p}, \mathbf{q}) \leq \epsilon$  and the improvement  $N_{\epsilon}^{\text{ratio}} := \frac{N_{\epsilon}^*(\mathbf{p})}{N_{\epsilon}^{\text{same}}(\mathbf{p})}$ .

**Specifying the Learning Model** The expected test loss  $L_N(\mathbf{p}, \mathbf{q})$ , and so  $\mathbf{q}^*$  and the benefit of mismatch, depend on the data distributions and learning behaviour of the algorithm. We capture these by modeling the *subpopulation error function*  $e_k(\mathbf{n})$ , i.e. the error on each component  $\mathcal{D}_k$  when training with  $n_i$  examples from each component  $\mathcal{D}_i$ . That is, for a vector of sample sizes  $\mathbf{n} = (n_1, \dots, n_K) \in \mathbb{Z}_{\geq 0}^K$ , denote  $\mathcal{D}^{\mathbf{n}} = (\mathcal{D}_1)^{n_1} \times \dots \times (\mathcal{D}_K)^{n_K}$  the distributions over samples with  $n_i$  examples from each component  $\mathcal{D}_i$ . Then  $e_k(\mathbf{n}) = \mathbb{E}_{S \sim \mathcal{D}^{\mathbf{n}}}[\mathcal{L}_{\mathcal{D}_k}(\mathcal{A}(S))]$ . When  $e_k(\mathbf{n}) = g_k(n_k)$  depends only on the amount of within-component data, we say the components are *orthogonal*, meaning there is no transfer between them (as in our Chinese and European history example). The scalar function  $g_k(n_k)$  then captures the *learning curve* for each component. But more generally, there might also be transfer, with data from one component helping learning on another.

In any case, the learnability function  $e : \mathbb{Z}_{\geq 0}^K \rightarrow \mathbb{R}^K$ , captures our “learning model”. In each Section, we consider different forms of learning models and characterize  $\mathbf{q}^*$  and  $L^*$  for these models.

**Data Sets and Training Sequences** In our analysis, we refer to the training budget  $N$  and our learning model specifying learning based on  $n_k$  examples per component  $k$ . We can think of  $N$  and  $\mathbf{n}$  as specifying the number of training examples, in which case the training complexity is a sample complexity. Or, we can think of  $N$  as indicating the number of training steps, and  $n_k$  as indicating the number of steps in which an example from component  $k$  is used. In this case, training complexity is a measure of training time. Either interpretation is valid. But we should emphasize that we only study a dependence on *how many* examples are used from each component, *not* on the *order* (as in curriculum learning).

**Learnabilities and Mixing Ratios.** We model learning as a function of the *number* of examples from each component, but for our analysis, it will useful to introduce the function  $\bar{e}_{N,k}(\mathbf{q}) = \mathbb{E}_{S \sim (\mathcal{D}_{\mathbf{q}})^N}[\mathcal{L}_k(\mathcal{A}(S))]$ , which captures the expected error on component  $k$  with mixing proportions  $\mathbf{q}$ . We will refer to  $\bar{e}_k(\mathbf{q})$  as the subpopulation error function in terms of the mixture  $\mathbf{q}$ . Since the per-component counts  $\mathbf{n}$  are multinomial, we have  $\bar{e}_N(\mathbf{q}) = \mathbb{E}_{\mathbf{n} \sim \text{Mult}(\mathbf{q}, N)}[e(\mathbf{n})] \in \mathbb{R}^K$  and  $L_N(\mathbf{p}, \mathbf{q}) = \langle \mathbf{p}, \bar{e}_N(\mathbf{q}) \rangle$ . Frequently for large sample size  $N$ ,  $\bar{e}_N(\mathbf{q})$  will concentrate around  $e(\mathbf{q}N)$ , and we will sometimes exploit this in the analysis, or analyze for  $\bar{e}(\mathbf{q}) \approx e(\mathbf{q}N)$ .

### 3 Orthogonal Power Law

Many machine learning tasks can be captured with power law error functions. Some classic examples include linear regression or learning VC classes, both of which have error rate  $\propto \frac{1}{n}$ , where  $n$  is the number of data samples. More recently, there have been many papers studying the loss curves for large language models for various tasks as a function of the compute budget in various scaling laws, such as the Chinchilla Scaling Law [Hoffmann et al., 2022].

To model these situations, we will first consider a setup where each of the  $K$  tasks is orthogonal and their subpopulation error functions in terms of the number of samples follow a simple power law.

116 **Model 3.1** (Orthogonal Power Law Error Tasks). There are  $K$  orthogonal tasks, each of which takes  
 117 data from one of the  $K$  subpopulations  $\mathcal{D}_i$  that appear in the test distribution with probability  $p_i$   
 118 and whose subpopulation error function  $e_k(\mathbf{n})$  follows a power law, i.e.  $e_k(\mathbf{n}) = \frac{A_k}{n_k^{\alpha_k} + B_k}$  for some  
 119  $A_k > 0, B_k \geq 0$ , and  $0 < \alpha_k \leq 1$ .<sup>1</sup>

120 In Proposition 3.2, we characterize the test error improvement from the positive distribution shift  
 121 from optimal data mixing ratios in Model 3.1 when the size of the training data  $n$  is large.

122 **Proposition 3.2** (Optimal Data Mixing Ratios For General Power Law). *In Model 3.1, if for the*  
 123 *exponents it holds that  $\alpha_1 = \alpha_2 = \dots = \alpha_S < \alpha_{S+1} \leq \alpha_{S+2} \leq \dots \leq \alpha_K$  for some  $S$*   
 124 *then there exist  $\varepsilon_1, \varepsilon_2 \geq 0$  that depend on  $\alpha_i$  such that for any test data mixing ratio  $\mathbf{p}$  and any*  
 125  *$n > n_0(A_i, B_i, \alpha_i, p_i)$  we have that the following holds*

$$q_i^* = \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \left( \frac{(\alpha_i p_i A_i)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{\alpha_i + 1}} \right)^{\frac{1}{\alpha_i + 1}} + o\left( \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \right) \quad (1)$$

126

$$L^{\text{same}}(\mathbf{p}) = \frac{1}{N^{\alpha_1}} \sum_{i=1}^S p_i^{1-\alpha_1} A_i + o\left( \frac{1}{N^{\alpha_1 + \varepsilon_1}} \right). \quad (2)$$

$$L^*(\mathbf{p}) = \frac{1}{N^{\alpha_1}} \left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{\alpha_1} \left( \sum_{i=1}^S \frac{(p_i A_i)^{\frac{1}{\alpha_i + 1}}}{\alpha_i^{\frac{\alpha_i}{\alpha_i + 1}}} \right) + o\left( \frac{1}{N^{\alpha_1 + \varepsilon_2}} \right). \quad (3)$$

127 The  $o(\cdot)$  notation hides dependence on  $A_i, B_i, p_i, K$  and  $\alpha_i$ .

128 Proposition 3.2 shows that in the power law Model 3.1, positive distribution shift from optimal data  
 129 mixing ratios improves the prefactor of the test error dependence on the number of data samples  $N$   
 130 but does not change the decay rate in terms of  $N$ . For the proof of Proposition 3.2 and a more precise  
 131 statement, see Appendix A.1.

132 To show that this can have significant implications for making training more data efficient, we show  
 133 the improvement from this positive distribution shift on the sample complexity in the case where we  
 134 have one majority population and  $K - 1$  minority populations that all have the same power exponent  
 135  $\alpha$ . This will also include the test-taking example from Section 1.

136 **Corollary 3.3** (Sample Complexity Improvement From Optimal Data Mixing For General Power  
 137 Law). *Consider Model 3.1 with  $S = K$ , i.e.  $\alpha_1 = \dots = \alpha_K = \alpha$  and  $A_1 = \dots = A_K = A$  with*  
 138  *$\mathbf{p} = (p, \frac{1-p}{K-1}, \dots, \frac{1-p}{K-1})$ . We have that for any  $\epsilon > 0$*

$$N_\epsilon^{\text{ratio}}(\mathbf{p}) \leq (1-p) + 2 \frac{\alpha+1}{\alpha} \left( \frac{p}{1-p} \right)^{\frac{1}{\alpha+1}} K^{-\frac{\alpha}{\alpha+1}}.$$

139 Furthermore, the optimal mixing ratios are given by  $q_1^* \propto p^{\frac{1}{\alpha+1}}$  and  $q_i^* \propto \left( \frac{1-p}{K-1} \right)^{\frac{1}{\alpha+1}}$  for  $i \geq 2$ .

140 Corollary 3.3 demonstrates an example case, that if we have one majority population and a number  
 141 of minority populations, the positive distribution shift from optimal data mixing ratio significantly  
 142 improves sample complexity. For fixed  $p$ , if  $K$  is large enough,  $N^{\text{ratio}}(\mathbf{p})$  will be close to  $N^{\text{ratio}}(\mathbf{p}) \approx$   
 143  $1-p < 1$ , i.e. we get sample complexity improvement of up to  $p$ . For example, for  $p = 0.7$ ,  
 144  $\alpha = 0.28$ , and  $K = 100$ , for any  $\epsilon > 0$ ,  $N_\epsilon^{\text{ratio}}(\mathbf{p}) \approx 0.75$ , i.e. we achieve the same error with  $\approx 25\%$   
 145 less samples. We illustrate this in Figure 2. For the proof of Corollary 3.3, see Appendix A.1.

146 Furthermore, the test taking example considered in the introduction Section 1 follows from Corol-  
 147 lary 3.3, by taking  $K = 2$ ,  $\alpha = 1$ , and  $\mathbf{p} = (0.9, 0.1)$ . In particular, this shows that the optimal  
 148 studying budget allocation is  $\mathbf{q}^* = (0.75, 0.25)$  and the improvement is  $N^{\text{ratio}}(\mathbf{p}) = 0.8$ . This means  
 149 that if you study for the exam with the right mixing ratio  $\mathbf{q}^*$ , you would need to study 20% less time  
 150 to achieve the same score as compared to using the test mixing ratio  $\mathbf{p}$ . Further, taking  $\alpha = \frac{1}{2}$  we  
 151 get the second example on Figure 2. This shows that we indeed get  $\mathbf{q}^* = (0.812 \dots, 0.188 \dots)$  and  
 152  $N^{\text{ratio}}(\mathbf{p}) = 0.944$ .

<sup>1</sup>We will also use the convention that if  $B_k = 0$  then  $e_k(\mathbf{n}) = \min\{C_k, \frac{A_k}{n_k^{\alpha_k}}\}$  for some  $C_k > 0$ . This will prevent  $L(\mathbf{p}, \mathbf{q})$  from blowing up to infinity.

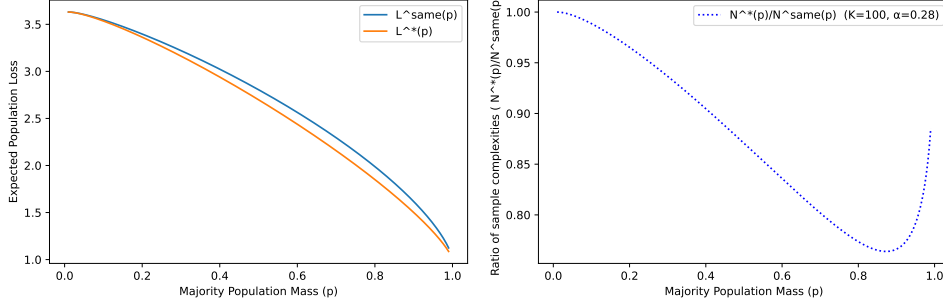


Figure 2: We consider the setup of Corollary 3.3 with  $A = 1$ ,  $\alpha = 0.28$ ,  $K = 100$ , and some fixed  $N$ . On the left plot, we show the "non-shifted" expected population loss  $L^{\text{same}}(\mathbf{p})$  and the optimally mixed expected population loss  $L^*(\mathbf{p})$  as a function of majority population mass  $p$ . On the right plot, we show the ratio of sample complexities for any fixed  $\epsilon > 0$ ,  $N_\epsilon^{\text{ratio}}(\mathbf{p})$  as a function of the mass of the majority population,  $p$ . We can see significant improvement in the sample complexity from the positive distribution shift from using optimal mixing ratio, even up to  $\approx 25\%$ .

## 4 Orthogonal Memorization Tasks

We consider a task of memorizing a number of unique elements from a dataset of fixed size, where the test distribution is a mixture of the tasks we are trying to memorize.

**Model 4.1** (Orthogonal Memorization Tasks). Suppose there are  $K$  tasks, each of which is a memorization of a unique element. The test distribution is a mixture of these  $K$  tasks, where the  $k$ -th task appears with probability  $p_k$ . In this case the subpopulation error functions in terms of  $\mathbf{n}$  is given by  $e_k(\mathbf{n}) = \mathbf{1}_{\{n_k=0\}}$ .

The following theorem characterizes the test error improvement from the positive distribution shift from optimal data mixing ratios in the Orthogonal Memorization Task Model 4.1.

**Theorem 4.2** (Optimal Data Mixing Test Error Improvement For Orthogonal Memorization Task). In Model 4.1, for all  $\mathbf{p} \in \Delta^{K-1}$  with  $p_1 \geq p_2 \geq \dots \geq p_K$ , the expected loss when training on  $n$  samples is given by

$$L^{\text{same}}(\mathbf{p}) = \sum_{k=1}^K p_k (1 - p_k)^N \quad (4)$$

$$L^*(\mathbf{p}) = (K_N(\mathbf{p}) - 1) \delta_N(\mathbf{p}) + \sum_{k=K_N(\mathbf{p})+1}^K p_k, \quad (5)$$

where  $\delta_N(\mathbf{p}) \in [p_{K_N(\mathbf{p})+1}, p_{K_N(\mathbf{p})})$  and  $K_N(\mathbf{p})$  is defined as follows:

$$K_N(\mathbf{p}) := \max \left\{ s \leq K : \sum_{k=1}^{s-1} (1 - (p_s/p_k)^{1/(K-1)}) < 1 \right\}. \quad (6)$$

To understand the magnitude of the test error improvement in Theorem 4.2, we will assume that the test proportions  $\mathbf{p}$  follow a power law  $p_k = \Theta(k^{-\alpha})$  for some  $\alpha > 1$  and that the number of tasks to memorize  $K$  is larger than the size of the training set  $N$ . In this case, we show that the improvement from positive distribution shift Theorem 4.2 improves even the test error scaling in terms of  $N$ . For the proof of Theorem 4.2, see Appendix A.2.

**Corollary 4.3** (Test Error Improvement For Orthogonal Memorization Tasks with Power Law Test Mixing Ratios). If  $p_k = \Theta(k^{-\alpha})$  for some  $\alpha > 1$  and  $K = \Omega(N)$ , then

$$L^{\text{same}}(\mathbf{p}) = \Theta(N^{-1+\frac{1}{\alpha}}), \quad L^*(\mathbf{p}) = \Theta(N^{-\alpha+1}).$$

For example, when  $\alpha = 1.5$ , we have  $L^{\text{same}}(\mathbf{p}) = \Theta(N^{-1/3})$  and  $L^*(\mathbf{p}) = \Theta(N^{-1/2})$ . For the proof of Corollary 4.3, see Appendix A.2.

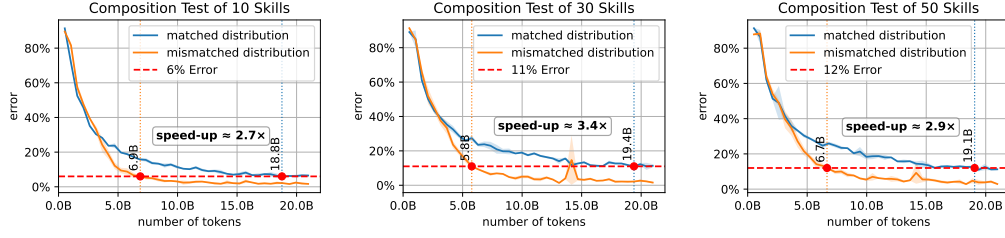


Figure 3: Mismatched distribution improves the test accuracy of a language model in solving a synthetic CoT reasoning task on skill composition (Section 5). During test, the model is asked to compose several functions following a power law. Instead of training directly on this task (blue curve), mixing with another task that uniformly samples the functions improves the final accuracy (orange curve).

## 5 Connection to Skill Composition

All the above analyses focus on the case where tasks are orthogonal. However, if we already know that the test distribution can be decomposed into  $M$  tasks, then maybe we should deal with these  $M$  tasks independently. So why do we have test mixing ratios in the first place?

We note here that in some cases, we may need to compose these  $M$  tasks later at inference time, and the test mixing ratios can come from the proportions in the composition. Imagine that we are training a language model to do mathematical reasoning. Each problem may involve several math skills, and a language model can acquire a math skill only if it sees the skill enough times during training. This can be conceptually modeled as the orthogonal memorization task discussed above, but at inference time, the language model has to sequentially apply the math skills in its chain of thought (CoT). The natural distribution of math skills then determines the test mixing ratios we care about.

We demonstrate this in a concrete synthetic task on skill composition. There are  $M$  skills, where the  $i$ -th skill is a function  $g_i$  that maps a number from  $\{0, \dots, 9\}$  to  $\{0, \dots, 9\}$ . Each skill has a unique English name. Assume that all these skills are randomly sampled: the names are uniformly random from a name set, and each  $g_i$  is uniformly random among all possible functions that map from  $\{0, \dots, 9\}$  to  $\{0, \dots, 9\}$ . At inference time, a set of  $k$  skills  $g_{i_1}, \dots, g_{i_k}$  are sampled IID following a power law with exponent  $\alpha = 1.5$ . The language model is prompted with the names of these skills and a number  $x \in \{0, \dots, 9\}$ : “[ $x$ ] -> [skill name 1] -> [skill name 2] ->  $\dots$  -> [skill name  $k$ ]”. The model is expected to output the result after function composition:  $y = g_{i_k}(g_{i_{k-1}}(\dots g_{i_1}(x) \dots))$ .

Let  $D_{\text{test}}$  be the distribution of the above prompt and a CoT calculating the correct answer, with  $M = 10^5$ ,  $k$  sampled uniformly from 10 to 50. Is the best strategy just training on the same distribution ( $D_{\text{train}} = D_{\text{test}}$ )? Inspired by our calculation for the orthogonal memorization task above, properly adjusting the occurrence probability for each skill may lead to better test accuracy. To demonstrate this, we construct another distribution  $D_{\text{uniform}}$  consisting of strings in the form of “[ $x$ ] [skill name] = [expected output]”, where the skill and input number are uniformly sampled. In Figure 3, we conduct experiments with a model with GPT-2 architecture and  $\sim 50M$  parameters. We show that training with  $D_{\text{train}} = 30\% \cdot D_{\text{uniform}} + 70\% \cdot D_{\text{test}}$  significantly outperform training with  $D_{\text{test}}$  directly. We defer the experiment details to Appendix C.

## 6 Non-orthogonal Tasks and Transfer Learning

Many transfer learning setups, such as multi-task learning of linear classifiers over linear representation with feature learning Baxter [2011], Maurer [2009], Pontil and Maurer [2013], Aliakbarpour et al. [2024] and multi-task learning with shared sparsity Wang et al. [2016, 2017], the subpopulation error functions  $e_k(\mathbf{n})$  can be written in the form  $e_k(\mathbf{n}) = \frac{A_{0,k}}{(n_1 + \dots + n_k)^{\alpha_k}} + \frac{A_{1,k}}{n_k^{\alpha_k}}$ . For example, in multi-task learning of shared sparsity Wang et al. [2017], the error bound takes this form with  $\alpha_1 = \dots = \alpha_K = 1$ .

To model all of these cases, we consider the following model of transfer learning.

**Model 6.1** (Standard Transfer Learning Model). There are  $K$  subpopulations, each of which appears in the test distribution with proportion  $p_k$ . The subpopulation error functions depend on the number of samples  $\mathbf{n}$  as  $e_k(\mathbf{n}) = \frac{A_{0,k}}{(n_1 + \dots + n_k)^{\alpha_k}} + \frac{A_{1,k}}{n_k^{\alpha_k}}$ , for some  $A_{0,k}, A_{1,k} > 0$  and  $0 < \alpha_k \leq 1$ .



Interestingly, the Standard Transfer Learning Model 6.1 is equivalent to the setup of Orthogonal Power Law Tasks Model 3.1 in the sense that we can understand optimal data mixing ratio  $\mathbf{q}^*$  and the error improvement of the Standard Transfer Learning model from a specific instance of the Orthogonal Power Law model. Namely, the transfer term in each of the subpopulation loss functions can be decomposed into a transfer error term and a specific task error term  $e_k(\mathbf{n}) = e_k^{\text{transfer}}(\mathbf{n}) + e_k^{\text{spec}}(\mathbf{n})$ , where  $e_k^{\text{transfer}}(\mathbf{n}) = \frac{A_{0,k}}{(n_1 + \dots + n_k)^{\alpha_k}}$  is independent of the distribution of samples across different tasks, and  $e_k^{\text{spec}}(\mathbf{n}) = \frac{A_{1,k}}{n_k^{\alpha_k}}$  only depends on  $n_k$ . Therefore, the transfer error term  $e_k^{\text{transfer}}(\mathbf{n})$  in each of the subpopulation error functions will only offset the final expected loss  $L(\mathbf{p}, \mathbf{q})$  by  $\sum_{i=1}^K p_i \frac{A_{0,k}}{N^{\alpha_k}}$ , which only depends on the total number of samples  $N$ . On the other hand, the specific task error terms  $e_k^{\text{spec}}(\mathbf{n})$  can be thought of as orthogonal tasks and will behave the same as in Model 3.1. So, for the Standard Transfer Learning Model 6.1, the optimal data mixing ratio  $\mathbf{q}^*$  and the expected test losses  $L^*(\mathbf{p})$  and  $L^{\text{same}}(\mathbf{p})$  are given by Equation (1) and Equation (2) respectively in Proposition 3.2 with  $A_k$  being replaced by  $A_{1,k}$ .

## 6.1 Data Mixing Transfer Learning.

Ye et al. [2025] consider the problem of estimating the outcome performance of a large language model trained on a mixture of domains. In particular, they find that an exponential function over the linear combinations of mixing proportions leads to good prediction. Namely, they fix the training budget  $N$  and only vary the mixing ratio  $\mathbf{q}$  and show that the validation loss on  $i$ -th domain can be predicted well by a function of the form  $c_i + b_i \exp\left(-\sum_{j=1}^K t_{ij} q_j\right)$ , where  $c_i, b_i, t_{ij}$  are parameters to fit. Following their work, we propose the following model for the Data Mixing Transfer Learning.

**Model 6.2** (Data Mixing Transfer Learning). There are  $K$  subpopulations, each of which appears with probability  $p_k$  in the test distribution. Each of the subpopulation error functions in terms of the mixing ratio  $\mathbf{q}$  are  $\bar{e}_k(\mathbf{q}) = c_k + b_k \exp\left(-\sum_{j=1}^K t_{ij} q_j\right)$  for some constants  $c_k$  and  $b_k > 0, t_{ij}$ .

We note that even though Model 6.2 is indeed not defined by the subpopulation error functions  $e_k(\mathbf{n})$ , it is precisely the setup that Ye et al. [2025] consider. This slightly deviates from our main setup, which focuses on specifying models by their error functions. However, when the number of samples  $N$  is large, it is reasonable to make the approximation that  $e_k(\mathbf{n}) \approx e_k(\mathbf{q}N)$ , and Model 6.2 can be interpreted as being defined by the subpopulation error functions of the form  $e_k(\mathbf{n}) = c_k(|\mathbf{n}|) + b_k(|\mathbf{n}|) \exp\left(-\sum_{j=1}^K t_{ij}(|\mathbf{n}|) n_j\right)$ , where  $c_k, b_k$ , and  $t_{ij}$  are functions that depend only on the total compute budget  $N = |\mathbf{n}|$ .

The following proposition characterizes the test error improvement from the positive distribution shift coming from the optimal data mixing ratio in the data mixing transfer model.

**Proposition 6.3** (Optimal Train Data Mixing Ratio for Data Mixing Transfer Learning Model). *In Model 6.2, if the coefficients  $t_{ij}$  are such that  $\mathbf{T}$  is invertible and  $(\mathbf{T}^T)^{-1} \mathbf{I} > 0$ , and  $p_i \neq 0$  for all  $i$ , the following hold*

$$\begin{aligned} \mathbf{q}^* &= (\mathbf{T})^{-1} \left( \frac{1 + \mathbf{I}^T \mathbf{T}^{-1} \boldsymbol{\tau}}{\mathbf{I} \mathbf{T}^{-1} \mathbf{I}} \mathbf{I} - \boldsymbol{\tau} \right) \\ L^{\text{same}}(\mathbf{p}) &= \sum_{i=1}^K c_i p_i + \sum_{i=1}^K p_i b_i \exp\left(-\sum_{j=1}^K t_{ij} p_j\right) \\ L^*(\mathbf{p}) &= \sum_{i=1}^K c_i p_i + \exp\left(\frac{-1 - \mathbf{I}^T \mathbf{T}^{-1} \boldsymbol{\tau}}{\mathbf{I}^T \mathbf{T}^{-1} \mathbf{I}}\right) \mathbf{I}^T (\mathbf{T}^T)^{-1} \mathbf{I}, \end{aligned}$$

where  $\boldsymbol{\tau}$  is a vector with entries  $\tau_l = \log\left(\frac{[(\mathbf{T}^T)^{-1} \mathbf{I}]_l}{p_l b_l}\right)$ .

Proposition 6.3 shows the positive distribution from the optimal data mixing for Model 6.2. Note that the additional conditions on  $\mathbf{T}, p_i$  are technical conditions used in order to simplify presentation. For the complete statement and the proof of Proposition 6.3, see Appendix A.3.

To demonstrate how large the gap can be, we consider the problem of data mixing transfer learning Model 6.2 with  $K = 2$  tasks and a one-directional transfer from the second to the first task.

**Corollary 6.4** (Optimal Data Mixing Ratio Can Have Significant Improvement in the Transfer Learning Model). *Let  $K = 2$ , let  $\mathbf{p} = (\frac{1}{2}, \frac{1}{2})$ , and let  $b_1 = b_2 = b > 0$ . If  $\mathbf{T} = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}$  then we have that*

$$L^{\text{same}} - L^* = 2be^{-\frac{1}{2}} \left( 1 - \frac{1}{4}\alpha + O(\alpha^2) \right).$$

Furthermore, if we let  $C = \frac{c_1 + c_2}{2}$  and  $B = be^{-\frac{1}{2}}$  then we have that

$$L^{\text{ratio}} = \frac{L_N}{L^*} = \frac{C - B}{C + B} + \frac{BC}{2(B + C)^2} \alpha + O(\alpha^2)$$

Corollary 6.4 shows that for two tasks with a small of transfer between the second to the first we can have error improvement from the positive distribution shift by mismatching training and test distribution, that is  $L^{\text{ratio}} \approx \frac{C-B}{C+B} < 1$  for small  $\alpha$ . For the proof of Corollary 6.4, see Appendix A.3.

## 7 It's Almost Always Better to Mismatch

So far, we have shown the existence of and quantified the positive distribution shift coming from mismatched test and train data mixing ratios for the cases of orthogonal power law tasks in Section 3, orthogonal memorization tasks in Section 4, and standard transfer learning and data mixing transfer learning in Section 6. that positive distribution shift from mismatching test and train mixing ratios exists. In this section, we will provide further mathematical justification that a positive distribution shift coming from the data mixing ratio almost always exists. That is, we show that it's almost always better to mismatch the training and test distributions:  $\mathbf{q}^* \neq \mathbf{p}$  and  $L^*(\mathbf{p}, \mathbf{q}^*) < L^{\text{same}}(\mathbf{p})$ .

More precisely, we will show that either the test data mixing ratio is on a measure zero set of the simplex or the subpopulation error functions  $e_k(\mathbf{n})$  have to be very specific functions, which are meaningless. For example, in the case of orthogonal tasks, either the test mixing ratio is on a measure zero subset or the subpopulation error functions  $e_k(\mathbf{n})$  are all constants, which we show in Corollary 7.4.

We define the probability simplex  $\Delta^{K-1} := \{\mathbf{p} \in \mathbb{R}^K : \mathbf{p} \geq 0, |\mathbf{p}| = 1\}$ , and its interior  $\Delta_+^{K-1} := \{\mathbf{p} \in \mathbb{R}^K : \mathbf{p} > 0, |\mathbf{p}| = 1\}$ , where  $|\mathbf{p}| := \sum_{k=1}^K p_k$ . We will define  $f_k(\mathbf{p})$  by extending the domain of each  $\bar{e}_k(\mathbf{p})$  to the set of non-zero, non-negative vectors  $\mathbb{R}_{\geq 0}^K \setminus \{\mathbf{0}\}$  by defining  $f_k(\mathbf{p}) := \bar{e}_k(\frac{\mathbf{p}}{|\mathbf{p}|})$ .

We further define  $L^{\text{same}}(\mathbf{p}) := \sum_{k=1}^K p_k f_k(\mathbf{p})$ , which extends the definition of  $L^{\text{same}}$  to the set of non-zero, non-negative vectors  $\mathbb{R}_{\geq 0}^K \setminus \{\mathbf{0}\}$ .

**Condition 7.1** (Conservation Condition).  $(f_1(\mathbf{p}), \dots, f_K(\mathbf{p})) = \nabla L^{\text{same}}(\mathbf{p})$  for all  $\mathbf{p} \in \mathbb{R}_{\geq 0}^K \setminus \{\mathbf{0}\}$ .

**Theorem 7.2** (Positive Distribution Shift Almost Always Exists For Data Mixing). *For any set of subpopulations  $\mathcal{D}_1, \dots, \mathcal{D}_K$  and any learning algorithm  $\mathcal{A}$ , either Condition 7.1 holds, or there exists a zero-measure set  $U$  on  $\Delta^{K-1}$  such that for all  $\mathbf{p} \in \Delta^{K-1} \setminus U$ ,  $L_N^*(\mathbf{p}) < L^{\text{same}}(\mathbf{p})$ .*

Theorem 7.2 shows that either  $\mathbf{p}$  is on a measure zero set  $U$  on  $\Delta^{K-1}$  or the Conservation Condition 7.1 must hold. We will show that Conservation Condition 7.1 happens only for very specific cases of subpopulation error functions.

**Conservation Condition Rarely Holds.** First, we will show that if the subtasks are orthogonal, the conservation condition Condition 7.1 is only satisfied if all of the subpopulation error functions are constants.

**Lemma 7.3** (Orthogonal Tasks). *If  $K \geq 3$ , and if for all  $k \in [K]$ ,  $f_k(\mathbf{p}) = g_k(\frac{p_k}{|\mathbf{p}|})$  for some function  $g_k$ , then Condition 7.1 holds if and only if  $g_k$ 's are all constant functions.*

Theorem 7.2 and Lemma 7.3 together show that in the case of orthogonal tasks, positive distribution shift always exists by changing the training data mixing ratio away from the test mixing ratio, unless all the subpopulation error functions are constant.



**Corollary 7.4** (Positive Distribution Shift Always Exists for Orthogonal Tasks). *For any set of  $K \geq 3$  subpopulations  $\mathcal{D}_1, \dots, \mathcal{D}_K$  and any learning algorithm  $\mathcal{A}$ , if there exists subpopulation  $k \in [K]$  such that its error function  $e_k$  is not a constant functions over  $[N]$  where  $N$  is the number of total samples then there exists a measure zero set  $U$  on  $\Delta^{K-1}$  such that for all  $\mathbf{p} \in \Delta^{K-1} \setminus U$  positive distribution shift from data mixing exists in the sense that there is  $\mathbf{q}^* \neq \mathbf{p}$  for which  $L_N(\mathbf{p}, \mathbf{q}) = L^*(\mathbf{p}) < L^{\text{same}}(\mathbf{p})$ .*

Further, we show that if the Conservation Condition 7.1 is satisfied, then one function  $f_i$  determines the rest up to a constant.

**Lemma 7.5.** *If both  $(f_1, \dots, f_K, L^{\text{same}})$  and  $(\hat{f}_1, \dots, \hat{f}_K, \hat{L}^{\text{same}})$  satisfy Condition 7.1, and if  $f_i = \hat{f}_i$  for some  $i \in [m]$ , then for all  $k \neq i$ ,  $f_k(\mathbf{p}) = \hat{f}_k(\mathbf{p}) + C_k$  for some constant  $C_k$ .*

The above Lemma 7.5 implies that for every  $k$  and corresponding error function  $e_k(\mathbf{n})$ , there exists at most one tuple of error functions  $\{e_j\}_{j=1, j \neq k}^K$  (up to a individual constant offset for each error function  $e_j$ ) that positive distribution shift does not happen for  $\mathbf{p}$  of positive measure. This further implies the following corollary.

**Corollary 7.6** (Positive Distribution Shift Almost Always Exists for General Tasks). *For any set of  $K \geq 3$  subpopulations  $\mathcal{D}_1, \dots, \mathcal{D}_K$  and any learning algorithm  $\mathcal{A}$ , for all  $\mathbf{p} \in \Delta_+^{K-1}$ , the configuration of  $[e_k(\mathbf{n})]_{k \in [K], \mathbf{n}}$  that positive distribution shift does not happen is zero-measure.*

Corollary 7.6 shows that either the test mixing ratio  $\mathbf{p}$  is on a set of measure zero on the simplex or the configuration of subpopulation error functions  $e_k(\mathbf{n})$  is on a set of measure zero. This implies that positive distribution shift exists *almost* always.

## 8 Related Works

**Distribution Shift That is Not Harmful.** The benefits of mismatching the training and test distribution has already been in studied in some settings. González and Abu-Mostafa [2015] demonstrate in many linear regression problems that mismatched training and test distributions can outperform matched ones. Unlike in our paper, they do not restrict to changing the train distribution only through data mixing, so their results do not fit our framework. On the other hand, we explicitly characterize the positive distribution shift, while González and Abu-Mostafa [2015] only show its existence for linear regression problems and are only able to characterize the distribution explicitly in very special cases. Canatar et al. [2021] show how in high-dimensional kernel regression problems to numerically optimize the training distribution for better test performance. However, they do not characterize the positive distribution shift, but rather only show how to numerically find it for kernel regression. Similarly, they do not restrict the test distribution to one coming from a data mixture, so their results do not fit our framework.

**Data Mixing.** There a number of recent empirically works that consider the same setting of data mixing as we do. Ye et al. [2025] introduce data mixing laws, quantitative empirical predictions of large language model performance based on the data mixture proportions. Furthermore, they show experimental results demonstrating that their approach significantly decreases the number of steps needed to reach certain performance. This paper informed our data mixing transfer model and fits in our framework. Goyal et al. [2024] show that data curation for VLMs cannot be compute agnostic. They introduce neural scaling laws that allow for estimating performance on multiple data pools without jointly training on them. Their work fits our framework. Similarly, we also find that optimal mixing ratios are not compute agnostic, specifically in the orthogonal power law tasks, orthogonal memorization task, and standard transfer learning task. Jiang et al. [2025] introduce an algorithm for online optimization of data distributions, that adjusts mixture based on the estimated per-domain learning potential, achieving comparable or better performance than previous methods while maintaing computational efficiency. While all of these works consider the same phenomena of changing the training mixing ratio to improve test performacne, the main difference between our work and theirs is that we consider positive distribution shift from data mixing ratio in a broader context and from the theoretical standpoint as well.

## References

- Maryam Aliakbarpour, Konstantina Bairaktari, Gavin Brown, Adam Smith, Nathan Srebro, and Jonathan Ullman. Metalearning with very few samples per task. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 46–93. PMLR, 30 Jun–03 Jul 2024. URL <https://proceedings.mlr.press/v247/aliakbarpour24a.html>.
- Jonathan Baxter. A model of inductive bias learning. *CoRR*, abs/1106.0245, 2011. URL <http://arxiv.org/abs/1106.0245>.
- Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-distribution generalization in kernel regression. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12600–12612. Curran Associates, Inc., 2021. URL [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/691dcb1d65f31967a874d18383b9da75-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/691dcb1d65f31967a874d18383b9da75-Paper.pdf).
- Carlos R. González and Yaser S. Abu-Mostafa. Mismatched training and test distributions can outperform matched ones. *Neural Computation*, 27(2):365–387, 2015. doi: 10.1162/NECO\_a\_00697.
- Sachin Goyal, Pratyush Maini, Zachary C. Lipton, Aditi Raghunathan, and J. Zico Kolter. Scaling laws for data filtering—data curation cannot be compute agnostic. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22702–22711, 2024. doi: 10.1109/CVPR52733.2024.02142.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=iBBcRU1OAPR>.
- Yiding Jiang, Allan Zhou, Zhili Feng, Sadhika Malladi, and J Zico Kolter. Adaptive data optimization: Dynamic sample selection with scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=aqok1UX7Z1>.
- Andreas Maurer. Transfer bounds for linear feature learning. *Machine Learning*, 75:327–350, 2009. URL <https://api.semanticscholar.org/CorpusID:14682470>.
- Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 55–76, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL <https://proceedings.mlr.press/v30/Pontil13.html>.
- Jialei Wang, Mladen Kolar, and Nathan Srebro. Distributed multi-task learning. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 751–760, Cadiz, Spain, 09–11 May 2016. PMLR. URL <https://proceedings.mlr.press/v51/wang16d.html>.
- Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3636–3645. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/wang17f.html>.
- Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=jjCB27TMK3>.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: Yes, the main claim accurately reflects the paper’s contribution and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Yes, we discuss the limitations of our work and clearly define the scope of each of our claims.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We provide full set of assumptions and complete and corrected proofs in the appendix. For some of the claims, we only state an informal or a limited scope version in the main body for the ease of presentation.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Yes, we disclose the information needed to reproduce the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we provide the access in to the code and data in the appendix.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

550 Answer: [\[Yes\]](#)

551 Justification: Yes, we specify all the details of the experiment necessary to understand and  
 552 reproduce the experiments.

553 Guidelines:

- 554 • The answer NA means that the paper does not include experiments.
- 555 • The experimental setting should be presented in the core of the paper to a level of detail that  
 556 is necessary to appreciate the results and make sense of them.
- 557 • The full details can be provided either with the code, in appendix, or as supplemental material.

558 **7. Experiment statistical significance**

559 Question: Does the paper report error bars suitably and correctly defined or other appropriate  
 560 information about the statistical significance of the experiments?

561 Answer: [\[Yes\]](#)

562 Justification: Yes, we provide information about statistical significance of results where appropri-  
 563 ate.

564 Guidelines:

- 565 • The answer NA means that the paper does not include experiments.
- 566 • The authors should answer "Yes" if the results are accompanied by error bars, confidence  
 567 intervals, or statistical significance tests, at least for the experiments that support the main  
 568 claims of the paper.
- 569 • The factors of variability that the error bars are capturing should be clearly stated (for example,  
 570 train/test split, initialization, random drawing of some parameter, or overall run with given  
 571 experimental conditions).
- 572 • The method for calculating the error bars should be explained (closed form formula, call to a  
 573 library function, bootstrap, etc.)
- 574 • The assumptions made should be given (e.g., Normally distributed errors).
- 575 • It should be clear whether the error bar is the standard deviation or the standard error of the  
 576 mean.
- 577 • It is OK to report 1-sigma error bars, but one should state it. The authors should preferably  
 578 report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality  
 579 of errors is not verified.
- 580 • For asymmetric distributions, the authors should be careful not to show in tables or figures  
 581 symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- 582 • If error bars are reported in tables or plots, The authors should explain in the text how they  
 583 were calculated and reference the corresponding figures or tables in the text.

584 **8. Experiments compute resources**

585 Question: For each experiment, does the paper provide sufficient information on the computer  
 586 resources (type of compute workers, memory, time of execution) needed to reproduce the  
 587 experiments?

588 Answer: [\[Yes\]](#)

589 Justification: Yes, we provide sufficient information on the computer resources needed to  
 590 reproduce the experiments in the appendix.

591 Guidelines:

- 592 • The answer NA means that the paper does not include experiments.
- 593 • The paper should indicate the type of compute workers CPU or GPU, internal cluster, or  
 594 cloud provider, including relevant memory and storage.
- 595 • The paper should provide the amount of compute required for each of the individual experi-  
 596 mental runs as well as estimate the total compute.
- 597 • The paper should disclose whether the full research project required more compute than the  
 598 experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it  
 599 into the paper).

600 **9. Code of ethics**



601 Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS  
602 Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

603 Answer: [Yes]

604 Justification: Yes, our research conforms in every aspect to the NeurIPS Code of Ethics.

605 Guidelines:

- 606 • The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- 607 • If the authors answer No, they should explain the special circumstances that require a deviation  
608 from the Code of Ethics.
- 609 • The authors should make sure to preserve anonymity (e.g., if there is a special consideration  
610 due to laws or regulations in their jurisdiction).

## 611 10. Broader impacts

612 Question: Does the paper discuss both potential positive societal impacts and negative societal  
613 impacts of the work performed?

614 Answer: [NA]

615 Justification: As this is mainly a theoretical paper, there is no immediate societal impact of the  
616 work.

617 Guidelines:

- 618 • The answer NA means that there is no societal impact of the work performed.
- 619 • If the authors answer NA or No, they should explain why their work has no societal impact or  
620 why the paper does not address societal impact.
- 621 • Examples of negative societal impacts include potential malicious or unintended uses (e.g.,  
622 disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deploy-  
623 ment of technologies that could make decisions that unfairly impact specific groups), privacy  
624 considerations, and security considerations.
- 625 • The conference expects that many papers will be foundational research and not tied to  
626 particular applications, let alone deployments. However, if there is a direct path to any  
627 negative applications, the authors should point it out. For example, it is legitimate to point out  
628 that an improvement in the quality of generative models could be used to generate deepfakes  
629 for disinformation. On the other hand, it is not needed to point out that a generic algorithm  
630 for optimizing neural networks could enable people to train models that generate Deepfakes  
631 faster.
- 632 • The authors should consider possible harms that could arise when the technology is being  
633 used as intended and functioning correctly, harms that could arise when the technology is  
634 being used as intended but gives incorrect results, and harms following from (intentional or  
635 unintentional) misuse of the technology.
- 636 • If there are negative societal impacts, the authors could also discuss possible mitigation  
637 strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms  
638 for monitoring misuse, mechanisms to monitor how a system learns from feedback over time,  
639 improving the efficiency and accessibility of ML).

## 640 11. Safeguards

641 Question: Does the paper describe safeguards that have been put in place for responsible release  
642 of data or models that have a high risk for misuse (e.g., pretrained language models, image  
643 generators, or scraped datasets)?

644 Answer: [NA]

645 Justification: The paper poses no such risks.

646 Guidelines:

- 647 • The answer NA means that the paper poses no such risks.
- 648 • Released models that have a high risk for misuse or dual-use should be released with necessary  
649 safeguards to allow for controlled use of the model, for example by requiring that users adhere  
650 to usage guidelines or restrictions to access the model or implementing safety filters.
- 651 • Datasets that have been scraped from the Internet could pose safety risks. The authors should  
652 describe how they avoided releasing unsafe images.

653 • We recognize that providing effective safeguards is challenging, and many papers do not  
654 require this, but we encourage authors to take this into account and make a best faith effort.

## 655 12. Licenses for existing assets

656 Question: Are the creators or original owners of assets (e.g., code, data, models), used in the  
657 paper, properly credited and are the license and terms of use explicitly mentioned and properly  
658 respected?

659 Answer: [Yes]

660 Justification: Yes, we properly credit all the original owners of assets where due.

661 Guidelines:

- 662 • The answer NA means that the paper does not use existing assets.
- 663 • The authors should cite the original paper that produced the code package or dataset.
- 664 • The authors should state which version of the asset is used and, if possible, include a URL.
- 665 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 666 • For scraped data from a particular source (e.g., website), the copyright and terms of service  
667 of that source should be provided.
- 668 • If assets are released, the license, copyright information, and terms of use in the package  
669 should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated  
670 licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- 671 • For existing datasets that are re-packaged, both the original license and the license of the  
672 derived asset (if it has changed) should be provided.
- 673 • If this information is not available online, the authors are encouraged to reach out to the  
674 asset's creators.

## 675 13. New assets

676 Question: Are new assets introduced in the paper well documented and is the documentation  
677 provided alongside the assets?

678 Answer: [NA]

679 Justification: We do not release new assets.

680 Guidelines:

- 681 • The answer NA means that the paper does not release new assets.
- 682 • Researchers should communicate the details of the dataset/code/model as part of their sub-  
683 missions via structured templates. This includes details about training, license, limitations,  
684 etc.
- 685 • The paper should discuss whether and how consent was obtained from people whose asset is  
686 used.
- 687 • At submission time, remember to anonymize your assets (if applicable). You can either create  
688 an anonymized URL or include an anonymized zip file.

## 689 14. Crowdsourcing and research with human subjects

690 Question: For crowdsourcing experiments and research with human subjects, does the paper  
691 include the full text of instructions given to participants and screenshots, if applicable, as well as  
692 details about compensation (if any)?

693 Answer: [NA]

694 Justification: The paper does not involve crowdsourcing nor research with human subjects.

695 Guidelines:

- 696 • The answer NA means that the paper does not involve crowdsourcing nor research with  
697 human subjects.
- 698 • Including this information in the supplemental material is fine, but if the main contribution of  
699 the paper involves human subjects, then as much detail as possible should be included in the  
700 main paper.
- 701 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or  
702 other labor should be paid at least the minimum wage in the country of the data collector.

703 **15. Institutional review board (IRB) approvals or equivalent for research with human subjects**

704 Question: Does the paper describe potential risks incurred by study participants, whether such  
705 risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or  
706 an equivalent approval/review based on the requirements of your country or institution) were  
707 obtained?

708 Answer: [NA]

709 Justification: See previous point.

710 Guidelines:

- 711 • The answer NA means that the paper does not involve crowdsourcing nor research with  
712 human subjects.
- 713 • Depending on the country in which research is conducted, IRB approval (or equivalent) may  
714 be required for any human subjects research. If you obtained IRB approval, you should  
715 clearly state this in the paper.
- 716 • We recognize that the procedures for this may vary significantly between institutions and  
717 locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines  
718 for their institution.
- 719 • For initial submissions, do not include any information that would break anonymity (if  
720 applicable), such as the institution conducting the review.

721 **16. Declaration of LLM usage**

722 Question: Does the paper describe the usage of LLMs if it is an important, original, or non-  
723 standard component of the core methods in this research? Note that if the LLM is used only for  
724 writing, editing, or formatting purposes and does not impact the core methodology, scientific  
725 rigorousness, or originality of the research, declaration is not required.

726 Answer: [NA]

727 Justification: The core methods developed in this research do not involve LLMs as any important,  
728 original, or non-standard components.

729 Guidelines:

- 730 • The answer NA means that the core method development in this research does not involve  
731 LLMs as any important, original, or non-standard components.
- 732 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what  
733 should or should not be described.

## 734 A Proofs of Error Rate Improvements

### 735 A.1 Case 1: General Power Law Tasks

736 **Definition A.1** (Approximate Subpopulation Error Function). For Power Law Model 3.1, let  $f_k(\mathbf{q})$   
 737 be *approximate subpopulation error function* defined as

$$f_k(\mathbf{q}) = \frac{A_i}{(q_i N)^{\alpha_i} + B_i}.$$

738 We will define the *approximate expected population loss* as

$$\tilde{L}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^K p_i f_k(\mathbf{q}) = \sum_{i=1}^K p_i \frac{A_i}{(q_i N)^{\alpha_i} + B_i}. \quad (7)$$

739 We will show that for Power Law Model 3.1 and large number of samples  $N$ , it is sufficient to  
 740 optimize the over the approximate expected population loss in to find  $\mathbf{q}^*$  up to error of the order  $\frac{1}{N}$ .

741 **Proposition A.2** (Sufficient to Consider Expectation). *For the approximate error function  $f_k(\mathbf{q})$  in*  
 742 *Definition A.1, we have that*

$$|f_k(\mathbf{q}) - \bar{e}_k(\mathbf{q})| \leq 320 \min\left\{\frac{A_k}{B_k}, 1\right\} \frac{1}{(Nq_k)^2} + \frac{\alpha_k A_k}{(Nq_k)^{\alpha_k + \frac{1}{4}}} + \frac{\alpha_k A_k}{(Nq_k)^{\alpha_k + \frac{1}{2}}}.$$

743 *Proof of Proposition A.2.* Let  $g(x) = \frac{A_k}{x^{\alpha_k} + B_k}$ . Note that for  $n_k \sim \text{Binom}(N, q_k)$  we have that  
 744  $\mu = \mathbb{E}[X] = Nq_k$ . So, we have that  $f_k(\mathbf{q}) = g(\mu)$  and  $\bar{e}_k(\mathbf{n}) = \mathbb{E}(g(n_k))$ . Note also that on  $(0, \infty)$ ,  
 745  $g(x)$  is twice differentiable with

$$g'(x) = -\frac{A_k \alpha_k x^{\alpha_k - 1}}{(x^{\alpha_k} + B_k)^2}$$

$$g''(x) = \frac{A_k \alpha_k (1 - \alpha_k) x^{\alpha_k - 2}}{(x^{\alpha_k} + B_k)^2} + \frac{A_k \alpha_k^2 x^{2\alpha_k - 2}}{(x^{\alpha_k} + B_k)^3}.$$

746 Note that because  $\mathbb{E}(n_k - \mu)$

$$\mathbb{E}[g(n_k) - g(\mu)] = \mathbb{E}\left[\frac{1}{2}g''(\xi)(n_k - \mu)^2\right].$$

747 We will bound the following

$$\bar{e}_k(\mathbf{q}) - f_k(\mathbf{q}) = \mathbb{E}[g(n_k) - g(\mu)] = \mathbb{E}[(g(n_k) - g(\mu))\mathbf{1}_{|n_k - \mu| > \mu^{\frac{3}{4}}}] + \mathbb{E}[(g(n_k) - g(\mu))\mathbf{1}_{|n_k - \mu| \leq \mu^{\frac{3}{4}}}].$$

748 First, note that by the Multiplicative Chernoff bound we have that if  $\mu > 1$

$$P\left(|n_k - \mu| \geq \mu^{\frac{3}{4}}\right) \leq 2 \exp\left(-\frac{\sqrt{\mu}}{2}\right).$$

749 If  $\mu \leq 1$  then  $|n_k - \mu| \leq \mu < \mu^{\frac{3}{4}}$ . Therefore, we have that<sup>2</sup>

$$\mathbb{E}[(g(n_k) - g(\mu))\mathbf{1}_{|n_k - \mu| \geq \mu^{3/4}}] \leq 4 \min\left\{\frac{A_k}{B_k}, 1\right\} \exp\left(-\frac{\sqrt{\mu}}{2}\right),$$

750 where we also used that  $|g(n_k) - g(\mu)| \leq 2 \max_x g(x)$ .

751 Furthermore, for the second term, we have that  $|n_k - \mu| < \mu^{3/4}$  implies that  $n_k > (\mu - \mu^{3/4})$ .

752 By Taylor's Theorem there exists  $\xi \in (n_k, \mu)$  (or  $(\mu, n_k)$ ) so that

$$g(n_k) = g(\mu) + g'(\mu)(n_k - \mu) + \frac{1}{2}g''(\xi)(n_k - \mu)^2.$$

---

<sup>2</sup>In the case that  $B_k = 0$  we take the minimum with 1.

753 Therefore, we have that

$$\begin{aligned} E[(g(n_k) - g(\mu))\mathbf{1}_{|n_k - \mu| < \mu^{3/4}}] &= E[(g(n_k) - g(\mu)) | |n_k - \mu| < \mu^{3/4}] \\ &= E[g'(\mu)(n_k - \mu) + \frac{1}{2}g''(\xi)(n_k - \mu)^2 | |n_k - \mu| < \mu^{3/4}] \end{aligned}$$

754 So we have that

$$\begin{aligned} |\bar{e}_k(\mathbf{q}) - f_k(\mathbf{q})| &\leq |E[g(n_k) - g(\mu)]| \leq 4 \min\left\{\frac{A_k}{B_k}, 1\right\} \exp\left(-\frac{\sqrt{\mu}}{2}\right) \\ &\quad + E[|g'(\mu)| | |n_k - \mu| < \mu^{3/4}] + E\left[\frac{1}{2}|g''(\xi)| |n_k - \mu|^2 \mid |n_k - \mu| < \mu^{3/4}\right] \\ &\leq 4 \min\left\{\frac{A_k}{B_k}, 1\right\} \exp\left(-\frac{\sqrt{\mu}}{2}\right) + \frac{\alpha_k A_k}{\mu^{\alpha_k + 1}} \mu^{\frac{3}{4}} + \frac{\alpha_k(1 - \alpha_k)A_k}{\mu^{\alpha_k + 2}} \mu^{\frac{3}{2}} + \frac{\alpha_k^2 A_k}{\mu^{\alpha_k + 2}} \mu^{\frac{3}{2}} \\ &\leq 4 \min\left\{\frac{A_k}{B_k}, 1\right\} \exp\left(-\frac{\sqrt{\mu}}{2}\right) + \frac{\alpha_k A_k}{\mu^{\alpha_k + \frac{1}{4}}} + \frac{\alpha_k A_k}{\mu^{\alpha_k + \frac{1}{2}}} \\ &\leq 320 \min\left\{\frac{A_k}{B_k}, 1\right\} \frac{1}{\mu^2} + \frac{\alpha_k A_k}{\mu^{\alpha_k + \frac{1}{4}}} + \frac{\alpha_k A_k}{\mu^{\alpha_k + \frac{1}{2}}}. \end{aligned}$$

755

□

756 **Proposition A.3** (Optimum of the Approximate Power Law). *Let  $\tilde{q}^*$  be the minimum of the ap-*  
757 *proximate population loss defined in Equation (7). For  $N > N_0(p_i, A_i, B_i, \alpha_i)$ , we have that*  
758 *then*

$$\begin{aligned} \tilde{q}_i^* &= \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \left( \frac{(\alpha_i p_i A_i)}{\left(\sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}}\right)^{\alpha_i + 1}} \right)^{\frac{1}{\alpha_i + 1}} + o\left(\frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}}\right) \\ \tilde{L}(\tilde{\mathbf{q}}^*) &= \frac{1}{N^{\alpha_1}} \left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{\alpha_1} \left( \sum_{i=1}^S \frac{(p_i A_i)^{\frac{1}{\alpha_i + 1}}}{\alpha_i^{\frac{\alpha_i}{\alpha_i + 1}}} \right) + O\left(\frac{1}{N^{\alpha_1 + \frac{3\alpha_1^2}{2\alpha_1 + 2}}}\right). \quad (8) \end{aligned}$$

759 *Proof of Proposition A.2.* We will take  $N$  large enough so that we force  $\tilde{q}_i^* \neq 0$ . First take

$$r_i = \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \left( \frac{(\alpha_i p_i A_i)}{\left(\sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}}\right)^{\alpha_i + 1}} \right)^{\frac{1}{\alpha_i + 1}}$$

760 Take

$$\begin{aligned} \bar{q}_1 &= \left( \frac{(\alpha_1 p_1 A_1)}{\left(\sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}}\right)^{\alpha_1 + 1}} \right)^{\frac{1}{\alpha_1 + 1}} - \sum_{i=S+1}^K \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \left( \frac{(\alpha_i p_i A_i)}{\left(\sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}}\right)^{\alpha_i + 1}} \right)^{\frac{1}{\alpha_i + 1}} \\ \bar{q}_i &= r_i \text{ for } i > 1. \end{aligned}$$

761 This way  $\sum_{i=1}^K \bar{q}_i = 1$ . Take  $N$  large enough so that  $\bar{q}_1 \in (0, 1)$ , i.e.

$$N > \left( 2 \left( \frac{(\alpha_1 p_1 A_1)}{\left(\sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}}\right)^{\alpha_1 + 1}} \right)^{\frac{-1}{\alpha_1 + 1}} \left( \sum_{i=S+1}^K \left( \frac{(\alpha_i p_i A_i)}{\left(\sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}}\right)^{\alpha_i + 1}} \right)^{\frac{1}{\alpha_i + 1}} \right) \right)^{\frac{\alpha_S + 1 + 1}{\alpha_S + 1 - \alpha_1}} \quad (9)$$

762 suffices because  $\frac{1}{N^{\frac{\alpha_{S+1}-\alpha_1}{\alpha_{S+1}+1}}} \geq \frac{1}{N^{\frac{\alpha_i-\alpha_1}{\alpha_i+1}}}$  for all  $i \geq S+1$ . Note that for these  $\bar{q}_i$ , we have that for all  
 763  $i > 2$

$$f_i(\bar{\mathbf{q}}) \leq \frac{1}{N^{\alpha_i \frac{1+\alpha_1}{1+\alpha_i}}} A_i \left( \frac{(\alpha_i p_i A_i)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i+1}} \right)^{\alpha_i+1}} \right)^{-\frac{\alpha_i}{\alpha_i+1}}.$$

764 For  $i = 1$ , we have that  $\bar{q}_1 \geq \frac{1}{2} \left( \frac{(\alpha_1 p_1 A_1)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i+1}} \right)^{\alpha_1+1}} \right)^{\frac{1}{\alpha_1+1}}$ , so

$$f_1(\bar{\mathbf{q}}) \leq A_1 2^{\alpha_1} \left( \frac{(\alpha_1 p_1 A_1)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i+1}} \right)^{\alpha_1+1}} \right)^{\frac{-\alpha_1}{\alpha_1+1}}.$$

765 Therefore, we have that for the approximate expected population loss

$$\begin{aligned} \tilde{L}(\bar{\mathbf{q}}) &\leq \frac{1}{N^{\alpha_1}} p_1 A_1 2^{\alpha_1} \left( \frac{(\alpha_1 p_1 A_1)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i+1}} \right)^{\alpha_1+1}} \right)^{\frac{-\alpha_1}{\alpha_1+1}} + \sum_{i=2}^K p_i \frac{1}{N^{\alpha_i \frac{1+\alpha_1}{1+\alpha_i}}} A_i \left( \frac{(\alpha_i p_i A_i)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i+1}} \right)^{\alpha_i+1}} \right)^{-\frac{\alpha_i}{\alpha_i+1}} \\ &\leq \frac{2^{\alpha_1}}{N^{\alpha_1}} \left( \sum_{i=1}^K p_i A_i \left( \frac{(\alpha_i p_i A_i)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i+1}} \right)^{\alpha_i+1}} \right)^{-\frac{\alpha_i}{\alpha_i+1}} \right). \end{aligned}$$

766 Therefore, taking  $N$  large enough so that  $\tilde{L}(\bar{\mathbf{q}}) < \min_i \{1, \frac{A_i}{B_i}\}$  shows that  $\tilde{L}$  at  $\bar{\mathbf{q}}$  is smaller than  $\tilde{L}$   
 767 for any  $\mathbf{q}$  with one of  $q_i = 0$ . For this it suffices to take

$$N > 2 \left( \frac{1}{\min_i \{1, \frac{A_i}{B_i}\}} \left( \sum_{i=1}^K p_i A_i \left( \frac{(\alpha_i p_i A_i)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i+1}} \right)^{\alpha_i+1}} \right)^{-\frac{\alpha_i}{\alpha_i+1}} \right)^{-1} \right)^{\frac{1}{\alpha_1}}. \quad (10)$$

768 Therefore, we have shown that for  $N$  larger than the expressions in Equation (9) and Equation (10),  
 769  $\tilde{\mathbf{q}}^*$  has no zero coordinates.

770 Now we can find  $\tilde{\mathbf{q}}^*$  inside  $(0, 1)^K$  using Lagrange multipliers. Note that each  $f_i$  is continuously  
 771 differentiable on  $(0, 1)^K$ , which is an open set containing the feasible set. Note that the constraint  
 772 now is  $\sum_{i=1}^K q_i - 1 = 0$ . We have that there exists  $\lambda > 0$  such that

$$\begin{aligned} \frac{p_i A_i}{((N q_i)^{\alpha_i} + B_i)^2} \alpha_i N^{\alpha_i} q_i^{\alpha_i-1} &= \lambda \\ \sum_{i=1}^K q_i &= 1. \end{aligned}$$

773 Note that this equation has a unique solution in  $(0, \infty)$  since  $\frac{A_i}{((N q_i)^{\alpha_i} + B_i)^2} \alpha_i N^{\alpha_i}$  is a decreasing  
 774 function and  $\lambda q_i^{1-\alpha_i}$  is an increasing function, and for  $q_i = 0$  we have that  $\frac{A_i}{B_i^2} \alpha_i N^{\alpha_i} > 0$ . Let  $\lambda^*$   
 775 and  $\tilde{q}_i = \tilde{q}_i(\lambda^*)$  be the unique solution. Note that  $\tilde{L}(\tilde{\mathbf{q}}^*) \leq \tilde{L}(\bar{\mathbf{q}})$  so in particular we have that

$$p_i \frac{A_i}{(N q_i)^{\alpha_i} + B_i} \leq C \frac{1}{N^{\alpha_1}},$$



776 where  $C = 2^{\alpha_1} \left( \sum_{i=1}^K p_i A_i \left( \frac{(\alpha_i p_i A_i)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i+1}} \right)^{\alpha_1+1}} \right)^{-\frac{\alpha_i}{\alpha_i+1}} \right)$ . So we have that

$$\begin{aligned} \frac{p_i A_i}{C} N^{\alpha_1} &\leq (N q_i)^{\alpha_i} + B_i \\ N q_i &\geq \left( \frac{p_i A_i}{C} N^{\alpha_1} - B_i \right)^{\frac{1}{\alpha_i}}. \end{aligned}$$

777 Taking

$$N \geq \frac{2CB_i}{p_i A_i}$$

778 for all  $i$ . Therefore, we have that

$$\begin{aligned} N q_i &\geq \left( \frac{1}{2} \frac{p_i A_i}{C} N^{\alpha_1} \right)^{\frac{1}{\alpha_i}} \\ q_i &\geq N^{\frac{\alpha_1 - \alpha_i}{\alpha_i}} \end{aligned}$$

779 Therefore as long as

$$N > \left( \max_i \left\{ \frac{2B_i C}{p_i A_i} \right\} \right)^{\frac{1}{\alpha_1}} \quad (11)$$

780 we have that

$$\begin{aligned} \frac{A_i}{((N q_i)^{\alpha_i} + B_i)^2} &\geq \frac{p_i A_i}{(N q_i)^{\alpha_i}} \left( 1 - \frac{B_i}{(N q_i)^{\alpha_i}} \right)^2 \geq \frac{p_i A_i}{(N q_i)^{\alpha_i}} \left( 1 - \frac{2B_i C}{p_i A_i N^{\alpha_1}} \right)^2 \\ &\geq \frac{p_i A_i}{(N q_i)^{\alpha_i}} \left( 1 - \frac{4B_i C}{p_i A_i N^{\alpha_1}} \right) \end{aligned}$$

781 for

$$N > \max_i \left\{ B_i^{\frac{1}{\alpha_i}} \right\}. \quad (12)$$

782 Therefore, the equation

$$\frac{p_i A_i}{((N q_i)^{\alpha_i} + B_i)^2} \alpha_i N^{\alpha_i} q_i^{\alpha_i-1} = \lambda$$

783 implies that

$$\frac{p_i A_i}{(N q_i)^{2\alpha_i}} \alpha_i N^{\alpha_i} q_i^{\alpha_i-1} \left( 1 - \frac{4B_i C}{p_i A_i N^{\alpha_1}} \right) \leq \lambda \leq \frac{p_i A_i}{(N q_i)^{2\alpha_i}} \alpha_i N^{\alpha_i} q_i^{\alpha_i-1}.$$

784 Therefore, for all  $q$  we have that

$$\left( \frac{p_i A_i \alpha_i}{N^{\alpha_i} \lambda} \right)^{\frac{1}{\alpha_i+1}} \left( 1 - \frac{4B_i C}{p_i A_i N^{\alpha_1}} \right)^{\frac{1}{\alpha_i+1}} \leq q_i \leq \left( \frac{p_i A_i \alpha_i}{N^{\alpha_i} \lambda} \right)^{\frac{1}{\alpha_i+1}}.$$

785 Plugging this back into  $\sum_{i=1}^K q_i = 1$  we have that for  $\lambda$  it holds that

$$\sum_{i=1}^K \left( \frac{p_i A_i \alpha_i}{N^{\alpha_i} \lambda} \right)^{\frac{1}{\alpha_i+1}} \left( 1 - \frac{4B_i C}{p_i A_i N^{\alpha_1}} \right)^{\frac{1}{\alpha_i+1}} \leq 1 \leq \sum_{i=1}^K \left( \frac{p_i A_i \alpha_i}{N^{\alpha_i} \lambda} \right)^{\frac{1}{\alpha_i+1}}$$

786 Therefore, we have that

$$\lambda^* = \frac{1}{N^{\alpha_1}} \left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i+1}} \right)^{\alpha_1+1} + O\left(\frac{1}{N^{2\alpha_1}}\right).$$

787 From this we can compute that

$$\tilde{q}_i^* = \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \left( \frac{(\alpha_i p_i A_i)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{\alpha_i + 1}} \right)^{\frac{1}{\alpha_i + 1}} + O \left( \frac{1}{N^{\frac{\alpha_i - \alpha_1 + 2\alpha_1}{\alpha_i + 1}}} \right). \quad (13)$$

788 This finishes the proof. The lower bound on  $N$  i.e.  $N_0(p_i, A_i, B_i, \alpha_i)$  is given by the minimum of  
789 Equations (9) to (11) and Equation (12). This shows that

$$\tilde{L}(\tilde{\mathbf{q}}^*) = \frac{1}{N^{\alpha_1}} \left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{\alpha_1} \left( \sum_{i=1}^S \frac{(p_i A_i)^{\frac{1}{\alpha_i + 1}}}{\alpha_i^{\frac{\alpha_i}{\alpha_i + 1}}} \right) + O \left( \frac{1}{N^{\alpha_1 + \frac{2\alpha_1^2}{\alpha_1 + 1}}} \right)$$

790

□

791 **Proposition A.4** (Approximate Optimal is Close to Optimal). *Let  $\tilde{\mathbf{q}}^*$  be the minimum of the approx-*  
792 *imate population error  $\tilde{L}(\mathbf{q})$  in Equation (7) and let  $\mathbf{q}^*$  be the minimum of the loss in Power Law*  
793 *Model 3.1. Then if  $N \geq N_1(p_i, A_i, B_i, \alpha_i)$*

$$\begin{aligned} L(\mathbf{q}^*) &= \frac{1}{N^{\alpha_1}} \left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{\alpha_1} \left( \sum_{i=1}^S \frac{(p_i A_i)^{\frac{1}{\alpha_i + 1}}}{\alpha_i^{\frac{\alpha_i}{\alpha_i + 1}}} \right) + O \left( \frac{1}{N^{\alpha_1 + \frac{2\alpha_1^2}{\alpha_1 + 1}}} \right) \\ |\tilde{q}_i^* - q_i^*| &\leq o \left( \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \right) \\ q_i^* &= \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \left( \frac{(\alpha_i p_i A_i)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{\alpha_i + 1}} \right)^{\frac{1}{\alpha_i + 1}} + o \left( \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \right). \end{aligned}$$

794 *Proof of Proposition A.4.* Note that by Proposition A.2, we have that for  $\tilde{\mathbf{q}}^*$  defined in Equation (13)

$$\begin{aligned} L(\tilde{\mathbf{q}}^*) &\leq \tilde{L}(\tilde{\mathbf{q}}^*) + \sum_{k=1}^K p_k \left( 320 \min \left\{ \frac{A_k}{B_k}, 1 \right\} \frac{1}{(N\tilde{q}_k)^2} + \frac{\alpha_k A_k}{(N\tilde{q}_k)^{\alpha_k + \frac{1}{4}}} + \frac{\alpha_k A_k}{(N\tilde{q}_k)^{\alpha_k + \frac{1}{2}}} \right) \\ &\leq \tilde{L}(\tilde{\mathbf{q}}^*) + \frac{C_L}{N^{\frac{\alpha_1 + 1}{\alpha_1 + 1}(\alpha_1 + \frac{1}{4})}} \leq \tilde{L}(\tilde{\mathbf{q}}^*) + \frac{C_L}{N^{\alpha_1 + \frac{1}{4}}}, \end{aligned}$$

795 where  $C_L = 320 \min \left\{ \frac{A_k}{B_k}, 1 \right\} + 2\alpha_k A_k$ . Note additionally that by analogous logic from Proposi-  
796 tion A.2 the inequality also holds the other way. By Proposition A.2, we have that

$$L(\mathbf{q}^*) \geq \tilde{L}(\mathbf{q}^*) - \sum_{k=1}^K p_k \left( 320 \min \left\{ \frac{A_k}{B_k}, 1 \right\} \frac{1}{(N\tilde{q}_k)^2} + \frac{\alpha_k A_k}{(N\tilde{q}_k)^{\alpha_k + \frac{1}{4}}} + \frac{\alpha_k A_k}{(N\tilde{q}_k)^{\alpha_k + \frac{1}{2}}} \right).$$

797 Note that since  $\tilde{L}(\tilde{\mathbf{q}}^*)$  is the minimum, we have that

$$\tilde{L}(\mathbf{q}^*) \geq \tilde{L}(\tilde{\mathbf{q}}^*).$$

798 Therefore, we conclude

$$L(\mathbf{q}^*) \geq \tilde{L}(\tilde{\mathbf{q}}^*) - \sum_{k=1}^K p_k \left( 320 \min \left\{ \frac{A_k}{B_k}, 1 \right\} \frac{1}{(N\tilde{q}_k)^2} + \frac{\alpha_k A_k}{(N\tilde{q}_k)^{\alpha_k + \frac{1}{4}}} + \frac{\alpha_k A_k}{(N\tilde{q}_k)^{\alpha_k + \frac{1}{2}}} \right).$$

799 This finishes the proof of the first claim that

$$L(\mathbf{q}^*) = \frac{1}{N^{\alpha_1}} \left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{\alpha_1} \left( \sum_{i=1}^S \frac{(p_i A_i)^{\frac{1}{\alpha_i + 1}}}{\alpha_i^{\frac{\alpha_i}{\alpha_i + 1}}} \right) + O \left( \frac{1}{N^{\alpha_1 + \frac{2\alpha_1^2}{\alpha_1 + 1}}} \right).$$

800 Note that the above equations imply that

$$|\tilde{L}(\tilde{\mathbf{q}}^*) - \tilde{L}(\mathbf{q}^*)| \leq 2 \sum_{k=1}^K p_k \left( 320 \min\left\{\frac{A_k}{B_k}, 1\right\} \frac{1}{(N\tilde{q}_k)^2} + \frac{\alpha_k A_k}{(N\tilde{q}_k)^{\alpha_k + \frac{1}{4}}} + \frac{\alpha_k A_k}{(N\tilde{q}_k)^{\alpha_k + \frac{1}{2}}} \right) \leq 2 \frac{C_L}{N^{\alpha_1 + \frac{1}{4}}}.$$

801 Note now that for all  $k$  we have that for all  $q, q+h \in (0, 1)$  that there is  $\xi \in (q+h, q)$  with

$$f_k(q+h) - f_k(q) = f'_k(\xi)h.$$

802 Therefore, for  $k = 1, \dots, S$  we have that

$$f_i(\tilde{q}_i^* + h) - f_i(\tilde{q}_i^*) = f'_i(\xi_i)h$$

803 for some  $\xi_i \in (\tilde{q}_i^*, \tilde{q}_i^* + h)$ . Therefore,

$$|f_i(q_i^*) - f_i(\tilde{q}_i^*)| = |f'_i(\xi_i)| |q_i^* - \tilde{q}_i^*|.$$

804 If for  $i = 1, 2, \dots, S$  we have that  $q_i^* > 2\tilde{q}_i^*$ , say  $i = 1$ , there there exists index  $j$  such that

805  $q_j^* < \tilde{q}_i^* - \frac{\tilde{q}_1^*}{K}$ . Note that all  $|f'_i(x)|$  are decreasing, so then  $|f'_j(\xi_j)| \geq |f'_j(\tilde{q}_j^*)| = |\lambda| \geq \frac{1}{N^{\alpha_1 + \frac{1}{8}}}$  for

806  $N$  large enough, i.e. it suffices to have

$$N > 16 \left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{-8(\alpha_1 + 1)}. \quad (14)$$

807 Then we have that

$$\frac{p_i C_0}{N^{\alpha_1 + \frac{1}{8}}} \leq \frac{p_i}{N^{\alpha_1 + \frac{1}{8}}} |\tilde{q}_i^*| \leq p_i |f_i(q_i^*) - f_i(\tilde{q}_i^*)| \leq |\tilde{L}(\tilde{\mathbf{q}}^*) - \tilde{L}(\mathbf{q}^*)| \leq \frac{2C_L}{N^{\alpha_1 + \frac{1}{4}}},$$

808 which is impossible for

$$N > \left( \max_i \left\{ \frac{2C_L}{p_i} \right\} \right)^8. \quad (15)$$

809 Therefore, for all  $i = 1, \dots, S$  we have that  $q_i^* \leq 2\tilde{q}_i^*$ . Therefore, we have that for all  $i = 1, \dots, S$ ,

810  $|f'_i(\xi)| \geq \frac{1}{2^{2\alpha_1}} |f'_i(\tilde{q}_i^*)| = \frac{1}{2^{2\alpha_1}} |\lambda| \geq \frac{1}{2^{2\alpha_1}} \frac{1}{N^{\alpha_1}} C_\lambda$ , where  $C_\lambda = \left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{(\alpha_1 + 1)}$ .

811 Therefore, for all  $i = 1, \dots, S$  we have that

$$\frac{1}{2^{2\alpha_1}} \frac{1}{N^{\alpha_1}} C_\lambda |q_i^* - \tilde{q}_i^*| \leq p_i |f_i(q_i^*) - f_i(\tilde{q}_i^*)| |\tilde{L}(\tilde{\mathbf{q}}^*) - \tilde{L}(\mathbf{q}^*)| \leq 2 \frac{C_L}{N^{\alpha_1 + \frac{1}{4}}}.$$

812 Therefore, for all  $i = 1, \dots, S$  we have that

$$|q_i^* - \tilde{q}_i^*| < \frac{2^{2\alpha_1 + 1} C_L}{C_\lambda N^{\frac{1}{4}}}.$$

813 This shows that for  $i = 1, \dots, S$

$$q_i^* = \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \left( \frac{(\alpha_i p_i A_i)}{\left( \sum_{i=1}^S (\alpha_i p_i A_i)^{\frac{1}{\alpha_i + 1}} \right)^{\alpha_1 + 1}} \right)^{\frac{1}{\alpha_i + 1}} + o \left( \frac{1}{N^{\frac{\alpha_i - \alpha_1}{\alpha_i + 1}}} \right).$$

814 □

815 *Proof of Proposition 3.2.* Follows directly from Proposition A.4 and Proposition A.3.

816 **Proposition A.5** (Minimizer is in the interior). *Consider the approximate population error given by*  
 817 *Equation (7). Let  $\mathbf{q}^*$  be the minimum on  $\Delta^{K-1}$ . Then it holds that  $\mathbf{q}_i^* \neq 0$  for all  $i$  for which  $\alpha_i < 1$ .*

818 *Proof of Proposition A.5.* Assume that  $q^*$  is such that  $q_i^* = 0$  with  $\alpha_i < 1$  for  $i \in I$ , where  $I$  is a set  
 819 of indices. There exists  $j$  with  $q_j^* \neq 0$  since  $\sum_{i=1}^K q_i^* = 1$ . Consider the following function

$$g(x) = \sum_{i \in I} \frac{p_i A_i}{(xN)^{\alpha_i} + B_i} + \frac{p_j A_j}{((q_j^* - |I|x)N)^{\alpha_j} + B_j}.$$

820 Note that

$$g'(x) = - \sum_{i \in I} \frac{p_i A_i}{((xN)^{\alpha_i} + B_i)^2} \alpha_i N^{\alpha_i} x^{\alpha_i-1} + \frac{p_j A_j}{(((q_j^* - |I|x)N)^{\alpha_j} + B_j)^2} \alpha_j N^{\alpha_j} (q_j^* - |I|x)^{\alpha_j-1}.$$

821 for  $x \in (0, \frac{q_j^*}{|I|})$ . Note also that on  $x \in (0, \frac{q_j^*}{|I|})$ , the function  $g(x)$  is continous. We have that  
 822  $\lim_{x \rightarrow 0+} g'(x) = -\infty$ . There exists  $0 < \delta < \frac{q_j^*}{|I|}$  such that  $g'(x) < 0$  for all  $x \in (0, \delta)$ . To see  
 823 this, not that if this were not the case, there would have to exist a sequence of points  $x_1, x_2, \dots$   
 824 such that  $x_i \rightarrow 0$ . To see why, note that  $\lim_{x \rightarrow 0} g'(x) = -\infty$  implies that if  $g'(x_0) > 0$  then there  
 825 is  $\tilde{x}_0 \in (0, x_0)$  with  $g'(\tilde{x}_0) < 0$  and so by IVT we have that there has to exist  $x_1 \in (\tilde{x}_0, x_0)$  with  
 826  $g'(x_1) = 0$ . Repeated this procedure gives the sequence  $x_1, x_2, \dots$ . This is a contradiction since  
 827  $\lim_{n \rightarrow \infty} g'(x_n) = 0$ . Therefore, we have that  $g(x)$  is decreasing on  $(0, \delta)$ . Assume that  $g(0) \leq g(x)$   
 828 for all  $x \in (0, \delta)$ . Therefore, we have that  $\frac{g(x)-g(0)}{x} \geq 0$  for all  $x \in (0, \delta)$ . By MVT, for each  
 829  $x \in (0, \delta)$  there exists  $\xi_x \in (0, x)$  with  $g'(\xi_x) = \frac{g(x)-g(0)}{x} \geq 0$ . Again, this is a contradiction, since  
 830 for  $x \rightarrow 0+$  we have that  $\xi_x \rightarrow 0+$  so in particular  $0 \leq \lim_{x \rightarrow 0+} g'(\xi_x) = \lim_{x \rightarrow 0+} g'(x) = -\infty$ .  
 831 Therefore, there exists  $y \in (0, \delta)$  with  $g(0) > g(y)$ . This contradicts the assumption that  $q_i^* = 0$  for  
 832 all  $i \in I$  because if  $\tilde{q}_i = \begin{cases} q_i^* & i \neq j, i \notin I \\ y & i \in I \\ q_j^* - |I|y & i = j \end{cases}$ , then  $\tilde{L}(\tilde{q}, p) < \tilde{L}(q^*, p)$ . Therefore,  $q^*$  has  
 833 nonzero coordinates for all  $q_i^*$  for which  $\alpha_i \neq 1$ . □

834 □

835 *Proof of Corollary 3.3.* From Proposition 3.2, by directly plugging in we have that since here  $S = K$

$$\begin{aligned} q_i^* &= \frac{p_i^{\frac{1}{\alpha+1}}}{\sum_{i=1}^m p_i^{\frac{1}{\alpha+1}}} \\ q_1^* &= \frac{p^{\frac{1}{\alpha+1}}}{p^{\frac{1}{\alpha+1}} + (K-1) \left( \frac{1-p}{K-1} \right)^{\frac{1}{\alpha+1}}} \\ q_{i \geq 2} &= \frac{\left( \frac{1-p}{K-1} \right)^{\frac{1}{\alpha+1}}}{p^{\frac{1}{\alpha+1}} + (K-1) \left( \frac{1-p}{K-1} \right)^{\frac{1}{\alpha+1}}} \end{aligned}$$

836 Therefore, this immediately shows the claim about  $q_i^*$ . Therefore, we have that

$$N^{\text{ratio}} = \left( \frac{(p^{\frac{1}{\alpha+1}} + (K-1) \left( \frac{1-p}{K-1} \right)^{\frac{1}{\alpha+1}})^{\alpha+1}}{p^{1-\alpha} + (K-1)^{\alpha} (1-p)^{1-\alpha}} \right)^{\frac{1}{\alpha}}.$$

837 The only thing left to prove is the inequality. Let  $\delta = \left( \frac{p}{1-p} \right)^{\frac{1}{\alpha+1}} (K-1)^{-\frac{\alpha}{\alpha+1}}$ . Note that we can  
 838 write

$$p^{\frac{1}{\alpha+1}} + (K-1) \left( \frac{1-p}{K-1} \right)^{\frac{1}{\alpha+1}} = (K-1)^{\frac{\alpha}{\alpha+1}} (1-p)^{\frac{1}{\alpha+1}} (1+\delta).$$

839 Note that  $p^{1-\alpha} + (K-1)^{\alpha} (1-p)^{1-\alpha} \geq (K-1)^{\alpha} (1-p)^{1-\alpha}$ . Therefore, we can write that

$$N^{\text{ratio}} \leq (1-p)(1+\delta)^{\frac{\alpha+1}{\alpha}}.$$

840 Note also that  $(1 + \delta)^t \leq 1 + 2^t \delta$  for  $\delta < 1$ , since  $f(x) = (1 + x)^t$  has  $f''(x) = t(t-1)(1+x)^{t-2}$   
 841 so for  $t > 1$  it is convex. Therefore,  $f(\delta) \leq f(0) + (f(1) - f(0))\delta = 1 + (2^t - 1)\delta \leq 1 + 2^t \delta$ .  
 842 Using this for  $t = \alpha$ , we have that

$$N^{\text{ratio}}(\mathbf{p}) \leq (1 - p) + 2^{\frac{\alpha+1}{\alpha}} \left( \frac{p}{1-p} \right)^{\frac{1}{\alpha+1}} m^{-\frac{\alpha}{\alpha+1}}.$$

843 This finishes the proof. □

## 845 A.2 Case 2: Orthogonal Memorization Tasks

846 *Proof of Theorem 4.2.* Follows from Lemma A.6 and Lemma A.8. □

847 *Proof Corollary 4.3.* First, note that in this case  $\left( \frac{p_K}{p_k} \right)^{\frac{1}{N-1}} = \Theta\left( \left( \frac{k}{K} \right)^{\alpha/(N-1)} \right)$ . Therefore, only for  
 848  $l = \Theta(K)$  do we have  $f_N(l) = \Theta(1)$ , so indeed then  $K_N = \Theta(K)$  in this case. We directly compute  
 849 that  $L^{\text{same}}(\mathbf{p}) = \Theta(N^{-1+\frac{1}{\alpha}})$ .  $L^*(\mathbf{p})$  follows directly from Lemma A.8 by using  $K_N = \Theta(K)$ , and  
 850 we get  $L^*(\mathbf{p}) = \Theta(N^{\alpha-1})$  □

852 **Proofs for the Memorization Case** For every task  $k$ , we only need to memorize the unique  
 853 hypothesis that appears together with the task.

$$\bar{e}_k(\mathbf{q}) = (1 - q_k)^N, \quad L_N(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^K p_k (1 - q_k)^N.$$

854 Let  $\{q_k^*(N)\}_{k=1}^M = \arg \min_{\{q_k\}_{k=1}^M} \{L_N(\mathbf{p}, \mathbf{q})\}$ .

855 **Lemma A.6.** For all  $N \geq 1$ , there exists  $\beta_N > 0$  such that the following holds for  $q_k^*(N)$ :

$$q_k^*(N) = \max \left\{ 0, 1 - \beta_N \cdot p_k^{-1/(N-1)} \right\}.$$

856 *Proof.* By the method of Lagrange multipliers, there exists  $\lambda \in \mathbb{R}$  such that

$$-N p_k (1 - q_k^*(N))^{N-1} + \lambda = 0, \quad \forall k \in [M] \quad \text{s.t.} \quad q_k^*(N) > 0.$$

857 Then we have

$$q_k^*(N) = 1 - \left( \frac{\lambda}{N p_k} \right)^{1/(N-1)}.$$

858 Setting  $Z_N := \left( \frac{N}{\lambda} \right)^{1/(N-1)}$  and  $\beta_N = \frac{1}{Z_N}$  finishes the proof. □

859 Let  $K_N := \max\{k \in [K] : q_k^*(N) \neq 0\}$ .  $K_N$  and  $\beta_N$  satisfy the following relationship.

860 **Lemma A.7.** For all  $N \geq 1$ ,

$$\beta_N = \frac{K_N - 1}{\sum_{k=1}^{K_N} p_k^{-1/(N-1)}} \in \left[ p_{K_N+1}^{1/(N-1)}, p_{K_N}^{1/(N-1)} \right],$$

$$K_N = \max\{l \mid f_N(l) < 1\} \quad \text{where} \quad f_N(l) := \sum_{k=1}^{l-1} \left( 1 - \left( \frac{p_K}{p_k} \right)^{1/(N-1)} \right).$$

861 *Proof.* Since  $\sum_{k=1}^K q_k^*(N) = 1$  and  $q_k^*(N) = 0$  for all  $k > K_N$ , by Lemma A.6, we have

$$\sum_{k=1}^{K_N} \left( 1 - \beta_N \cdot p_k^{-1/(N-1)} \right) = 1.$$

862 Rearranging the terms, we obtain

$$\beta_N \sum_{k=1}^{K_N} p_k^{-1/(N-1)} = K_N - 1,$$

863 which implies  $\beta_N = \frac{K_N - 1}{\sum_{k=1}^{K_N} p_k^{-1/(N-1)}}$ .

864 By definition of  $K_N$ ,  $1 - \beta_N \cdot p_{K_N}^{-1/(N-1)} > 0$  and  $1 - \beta_N \cdot p_{K_N+1}^{-1/(N-1)} \leq 0$ . This implies  
 865  $\beta_N \in [p_{K_N+1}^{1/(N-1)}, p_{K_N}^{1/(N-1)})$ . Then we have

$$\begin{aligned} 1 &> \sum_{k=1}^{K_N} \left(1 - p_{K_N}^{1/(N-1)} \cdot p_k^{-1/(N-1)}\right) = \sum_{k=1}^{K_N-1} \left(1 - p_{K_N}^{1/(N-1)} \cdot p_k^{-1/(N-1)}\right). \\ 1 &\leq \sum_{k=1}^{K_N} \left(1 - p_{K_N+1}^{1/(N-1)} \cdot p_k^{-1/(N-1)}\right). \end{aligned}$$

866 Let  $f_N(K) := \sum_{k=1}^{K-1} \left(1 - p_K^{1/(N-1)} \cdot p_k^{-1/(N-1)}\right)$ . Then  $K_N = \max\{K : f_N(K) < 1\}$ .  $\square$

867 **Lemma A.8.** Test errors for sampling with  $\mathbf{q} = \mathbf{p}$  and  $\mathbf{q} = \mathbf{q}^*$  are

$$\begin{aligned} L^{\text{same}}(\mathbf{p}) &= \sum_{k=1}^K p_k (1 - p_k)^N, \\ L^*(\mathbf{p}) &= \sum_{k=K_N+1}^K p_k + (K_N - 1) \beta_N^{N-1} \in \left[ \sum_{k=K_N+1}^K p_k + (K_N - 1) p_{K_N+1}, \sum_{k=K_N+1}^K p_k + (K_N - 1) p_{K_N} \right). \end{aligned}$$

868 *Proof.* The first equation is straightforward. For the second equation,

$$\begin{aligned} L^*(\mathbf{p}) &= \sum_{k=K_N+1}^K p_k + \sum_{k=1}^N p_k (1 - q_k^*)^N = \sum_{k=1}^{K_N} p_k (\beta_N \cdot p_k^{-1/(N-1)})^N \\ &= \sum_{k=K_N+1}^K p_k + \beta_N^N \sum_{k=1}^{K_N} p_k^{-1/(N-1)} \\ &= \sum_{k=K_N+1}^K p_k + \beta_N^N \cdot \left( \frac{K_N - 1}{\beta_N} \right) \\ &= \sum_{k=K_N+1}^K p_k + (K_N - 1) \beta_N^{N-1}. \end{aligned}$$

869 Further noting that  $\beta_N \in [p_{K_N+1}^{1/(N-1)}, p_{K_N}^{1/(N-1)})$  completes the proof.  $\square$

### 870 A.3 Case 3: Transfer Learning

871 **Conditions on  $\mathbf{T}$ .** In Proposition 6.3, the conditions  $\mathbf{T}$  invertible and  $(\mathbf{T}^\top)^{-1} \mathbf{1} > 0$  are presented  
 872 as technical conditions. They will be used to simplify analysis. We can motivate these conditions by  
 873 the following. Consider  $\mathbf{T}$  to be a stochastic matrix with each column summing up to 1. This scenario  
 874 would model, for example,  $K$  things to be memorized, where the  $i$ -th element has distribution given  
 875 by  $\mathbf{T} e_i$ . Similarly to the case of orthogonal memorization, i.e. Model 4.1, the loss is expected to  
 876 have exponential behavior, i.e. for large number of samples to behave like  $\exp(-q_i)$  if we have  $q_i$   
 877 proportion of samples from the  $i$ -th subtask. This would further motivate the formula in Transfer  
 878 Learning Model 6.2. This way, the condition  $(\mathbf{T}^\top)^{-1} \mathbf{1} = \mathbf{1} > 0$  is naturally satisfied.



879 *Proof of Proposition 6.3.* Note first that  $L(\mathbf{q}) = L(\mathbf{p}, \mathbf{q})$  is a strictly convex and  $C^\infty$  function.

880 To see this, note that  $\nabla_{\mathbf{q}}^2 L(\mathbf{q}) = (\mathbf{T}^\top) \text{diag} \left( \begin{pmatrix} \vdots \\ p_l b_l \exp(\sum_{j=1}^M t_{lj} q_j) \\ \vdots \end{pmatrix} \right) \mathbf{T}$ . Let's denote  $\mathbf{w} =$   
881  $\begin{pmatrix} \vdots \\ p_l b_l \exp(\sum_{j=1}^M t_{lj} q_j) \\ \vdots \end{pmatrix}$ . Note that  $w_i > 0$  for all  $i$ . Therefore, note that

$$\mathbf{x}^T (\mathbf{T}^\top) \text{diag}(\mathbf{w}) \mathbf{T} \mathbf{x} = (\mathbf{T} \mathbf{x})^T \text{diag}(\mathbf{w}) (\mathbf{T} \mathbf{x}) = \sum_i w_i (x_{t,i})^2 > 0$$

882 for all  $\mathbf{x} \neq 0$ . Note that since  $L(\mathbf{q})$  is strictly convex and smooth on  $[0, 1]^n$  we can apply Lagrange  
883 Multipliers, KKT conditions, and Slater's condition to find the minimum of  $L(\mathbf{q})$  over  $\mathbf{q} \in \Delta^{K-1}$ .

884 Note that the optimization problem<sup>3</sup> here is

$$\begin{aligned} \min_{\mathbf{q} \in [0,1]^n} \quad & L(\mathbf{q}) \\ & q_i \geq 0 \\ & \sum_i q_i = 1. \end{aligned}$$

885 Note that since  $L(\mathbf{q})$  is convex, and all the constraints are affine so this is a convex optimization  
886 problem for which Slater's condition implies strong duality. This means that any point described  
887  $\mathbf{q}^*, \lambda^*, \mu_i^*$ , where  $\lambda$  and  $\mu$  are the dual variables, then satisfying KKT conditions is sufficient for the  
888 optimality of the primal and the dual. Note that  $\mathbf{q}_S = (\frac{1}{K}, \dots, \frac{1}{K})$  satisfies the Slater's condition.  
889 Therefore, it suffices to demonstrate a point  $\mathbf{q}^*, \lambda^*, \mu_i^*$  satisfying the KKT conditions. We will show  
890 that we have  $\mu_i^* = 0$  here.

891 The first of the KKT conditions will be.

$$\begin{aligned} \frac{\partial L}{\partial q_i} + \lambda + \mu_i &= 0 \\ - \sum_{l=1}^M p_l b_l \exp(- \sum_{j=1}^M t_{lj} q_j) t_{li} + \lambda + \mu_i &= 0 \end{aligned}$$

892 Therefore, if we set  $\mu_i^* = 0$ , we have that for  $\mathbf{w} = \begin{pmatrix} \vdots \\ p_l b_l \exp(\sum_{j=1}^M t_{lj} q_j) \\ \vdots \end{pmatrix}$  it holds

$$\begin{aligned} -\mathbf{T}^T \begin{pmatrix} \vdots \\ p_l b_l \exp(- \sum_{j=1}^M t_{lj} q_j) \\ \vdots \end{pmatrix} &= -\lambda \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ \begin{pmatrix} \vdots \\ p_l b_l \exp(- \sum_{j=1}^M t_{lj} q_j) \\ \vdots \end{pmatrix} &= \lambda (\mathbf{T}^T)^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \lambda \begin{pmatrix} \vdots \\ \sum_{j=1}^M \tilde{t}_{ij} \\ \vdots \end{pmatrix} \end{aligned}$$

---

<sup>3</sup>For a reference, see <https://www.stat.cmu.edu/ryantibs/convexopt-F16/scribes/kkt-scribed.pdf>

893 Note that because of the condition  $(\mathbf{T}^\top)^{-1} > 0$  we can solve for each of the  $p_l b_l \exp(-\sum_{j=1}^M t_{lj} q_j)$ .  
 894 So we have that (since  $\mathbf{T}$  is also invertible)

$$\mathbf{q} = -(\mathbf{T})^{-1} \begin{pmatrix} \vdots \\ \log \left( \lambda \frac{\sum_{j=1}^M \tilde{t}_{ij}}{p_i b_i} \right) \\ \vdots \end{pmatrix}.$$

895 Here  $\tilde{t}_{ij}$  are entries of  $(\mathbf{T}^\top)^{-1}$ . Note now that

$$\mathbf{T}\mathbf{q} = - \begin{pmatrix} \vdots \\ \log \left( \lambda \frac{\sum_{j=1}^M \tilde{t}_{ij}}{p_i b_i} \right) \\ \vdots \end{pmatrix}.$$

896 Therefore, we have that

$$\begin{aligned} f_k(\mathbf{p}) &= c_k + b_k \exp \left( \log \left( \lambda \frac{\sum_{j=1}^M \tilde{t}_{kj}}{p_k b_k} \right) \right) \\ &= c_k + \frac{\lambda}{p_k} \left( \sum_{j=1}^M \tilde{t}_{kj} \right) \end{aligned}$$

897 We also have that  $\lambda$  satisfies the following

$$\begin{aligned} \mathbf{1}^T \mathbf{q} &= 1 \\ \mathbf{1}^T \mathbf{T}^{-1} \begin{pmatrix} \vdots \\ \log \left( \lambda \frac{\sum_{j=1}^M \tilde{t}_{ij}}{p_i b_i} \right) \\ \vdots \end{pmatrix} &= -1 \\ \log(\lambda) \left( \sum_{l=1}^M \sum_{i=1}^M \tilde{t}_{il} \right) + \sum_{i=1}^M \sum_{l=1}^M \tilde{t}_{il} \log \left( \frac{\sum_{j=1}^M \tilde{t}_{lj}}{p_l b_l} \right) &= -1 \\ \lambda^* &= \exp \left( \frac{-1 - \sum_{i=1}^M \sum_{l=1}^M \tilde{t}_{il} \log \left( \frac{\sum_{j=1}^M \tilde{t}_{lj}}{p_l b_l} \right)}{\sum_{l=1}^M \sum_{i=1}^M \tilde{t}_{il}} \right) = \exp \left( \frac{-1 - \sum_{i=1}^M \sum_{l=1}^M \tilde{t}_{il} \log \left( \frac{\sum_{j=1}^M \tilde{t}_{lj}}{p_l b_l} \right)}{\mathbf{1}^T \mathbf{T}^{-1} \mathbf{1}} \right). \end{aligned}$$

898 Here  $\tilde{t}_{ij}$  are the entries of  $\mathbf{T}^{-1}$ .

899 Now to check that the KKT conditions are satisfied, note that by design the first will be satisfied  
 900  $\frac{\partial L}{\partial q_i} + \lambda = 0$ . Note also that  $\mu_i^* = 0$  so the complementary slackness is satisfied. Further, note that  
 901  $\lambda^* > 0$  holds and also that  $\mathbf{1}^T \mathbf{q} = 1$  holds. Finally, we check that this value of  $\lambda^*$  given  $q_i \geq 0$ ,  
 902 which we can see by simplifying

$$\lambda^* = \exp \left( \frac{-1 - \mathbf{1}^T \mathbf{T}^{-1} \begin{pmatrix} \vdots \\ \log \left( \frac{[(\mathbf{T}^\top)^{-1} \mathbf{1}]_l}{p_l b_l} \right) \\ \vdots \end{pmatrix}}{\mathbf{1}^T \mathbf{T}^{-1} \mathbf{1}} \right).$$

903 Combining all of these we have that

$$\begin{aligned}
L^{\text{same}}(\mathbf{p}) &= \sum_{i=1}^M c_i p_i + \sum_{i=1}^M p_i b_i \exp \left( - \sum_{j=1}^M t_{ij} p_j \right) \\
L_N(\mathbf{p}) &= \sum_{i=1}^M c_i p_i + \sum_{i=1}^M \lambda \left( \sum_{j=1}^M \tilde{t}_{ij} \right) = \sum_{i=1}^M c_i p_i + \exp \left( \frac{-1 - \sum_{i=1}^M \sum_{l=1}^M \tilde{t}_{il} \log \left( \frac{\sum_{j=1}^M \tilde{t}_{lj}}{p_l b_l} \right)}{\mathbf{1}^T \mathbf{T}^{-1} \mathbf{1}} \right) \mathbf{1}^T (\mathbf{T}^T)^{-1} \mathbf{1} \\
L_N(\mathbf{p}) &= \sum_{i=1}^M c_i p_i + \exp \left( \frac{-1 - \mathbf{1}^T \mathbf{T}^{-1} \left( \begin{array}{c} \vdots \\ \log \left( \frac{[(\mathbf{T}^T)^{-1} \mathbf{1}]_i}{p_i b_i} \right) \\ \vdots \end{array} \right)}{\mathbf{1}^T \mathbf{T}^{-1} \mathbf{1}} \right) \mathbf{1}^T (\mathbf{T}^T)^{-1} \mathbf{1}.
\end{aligned}$$

904 This finishes the proof.  $\square$

905 *Proof of Corollary 6.4.* We plug the into the formula of Proposition 6.3.

$$\begin{aligned}
L^{\text{same}}(\mathbf{p}) &= \frac{c_1 + c_2}{2} + \frac{b}{2} \exp \left( -\frac{1 + \alpha}{2} \right) + \frac{b}{2} \exp \left( -\frac{1}{2} \right) \\
L^*(\mathbf{p}) &= \frac{c_1 + c_2}{2} + \frac{b(2 - \alpha)}{2} \exp \left( -\frac{1}{2 - \alpha} \right) (1 - a)^{-\frac{1 - \alpha}{2 - \alpha}}.
\end{aligned}$$

906 Note that we have that

$$L^{\text{same}}(\mathbf{p}) = \frac{c_1 + c_2}{2} + b e^{-1/2} \left( 1 - \frac{\alpha}{4} + \frac{\alpha^2}{16} + O(\alpha^3) \right), \quad (16)$$

$$L^*(\mathbf{p}) = \frac{c_1 + c_2}{2} + b e^{-1/2} \left( 1 - \frac{\alpha}{4} - \frac{7\alpha^2}{32} + O(\alpha^3) \right). \quad (17)$$

907 Therefore, the first statement follows directly. For the second statement, we have that

$$L^{\text{ratio}} = \frac{C + B \left( 1 - \frac{\alpha}{4} + \frac{\alpha^2}{16} + O(\alpha^3) \right)}{C + B \left( 1 - \frac{\alpha}{4} - \frac{7\alpha^2}{32} + O(\alpha^3) \right)} = 1 + \frac{9B}{32(C + B)} \alpha^2 + O(\alpha^3).$$

908  $\square$

## 909 B Proof of the Existence of PDS in the General Case

### 910 B.1 Proof of Main Theorem

911 We provide a functional-analytic characterization of when positive distribution shift is guaranteed to  
912 exist. The key idea is to study the loss  $L_N(\mathbf{p}, \mathbf{r})$  as a function of both the target mixing ratios  $\mathbf{p}$  and  
913 the training mixing ratios  $\mathbf{r}$ , and show that  $\mathbf{r} = \mathbf{p}$  almost never minimizes  $L_N(\mathbf{p}, \mathbf{r})$  except for the  
914 degenerate cases described in Theorem 7.2.

915 The following key property of  $f_k$  is useful for our analysis:

916 **Lemma B.1.** *For all  $k \in [m]$ , the function  $f_k$  is a 0-homogeneous rational function.*

917 *Proof.* It is easy to see that  $f_k$  is 0-homogeneous, since by definition, it holds for all  $c > 0$  that  
 918  $f_k(c\mathbf{r}) = f_k(\frac{c\mathbf{r}}{|c\mathbf{r}|}) = f_k(\mathbf{r})$ .

919 To show that  $f_k$  is a rational function, recall that for  $\mathbf{r} \in \Delta^{m-1}$ ,  $f_k(\mathbf{r})$  is defined as the expected loss  
 920 of the model trained on a dataset  $S$  sampled with mixing ratio  $\mathbf{r}$  and evaluated on subpopulation  $\mathcal{D}_k$ .

921 Sampling from  $\mathbf{r}$  corresponds to first sampling subpopulation indices  $i_1, \dots, i_n \in [m]$  according to  
 922  $\mathbf{r}$ , and then drawing the  $j$ -th sample in the dataset  $S$  from the subpopulation  $\mathcal{D}_{i_j}$ . This allows us to  
 923 rewrite the expectation as:

$$F_k(\mathbf{r}) := \sum_{1 \leq i_1, \dots, i_n \leq m} (r_{i_1} \cdots r_{i_n} \cdot \mathbb{E}_{S \sim \mathcal{D}_{i_1} \times \mathcal{D}_{i_2} \times \cdots \times \mathcal{D}_{i_n}} \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_k} [\ell(\mathcal{A}(S), \mathbf{z})])$$

924 Each term in this sum  $F_k(\mathbf{r})$  is the product of a monomial of degree  $n$  in  $\mathbf{r}$  and a constant that does  
 925 not depend on  $\mathbf{r}$ . Therefore,  $F_k(\mathbf{r})$  is a degree- $n$  polynomial in  $\mathbf{r} \in \Delta^{m-1}$ . Since  $f_k(\mathbf{r}) := F_k(\frac{\mathbf{r}}{|\mathbf{r}|})$   
 926 by definition, it follows that  $f_k$  is a rational function on  $\mathbb{R}_{\geq 0}^m \setminus \{\mathbf{0}\}$ .  $\square$

927 Define the total population loss when testing under  $\mathbf{p}$  but training under  $\mathbf{r}$  as  $L_N(\mathbf{p}, \mathbf{r}) :=$   
 928  $\sum_{k=1}^m p_k f_k(\mathbf{r})$ . We now characterize when  $\mathbf{r} = \mathbf{p}$  is a minimizer of  $L_N(\mathbf{p}, \mathbf{r})$  over  $\mathbf{r} \in \Delta^{m-1}$ .

929 **Lemma B.2.** For any  $\mathbf{p} \in \Delta_+^{m-1}$ , if  $L^{\text{same}}(\mathbf{p}) = L_N^*(\mathbf{p})$ , then

$$\sum_{k=1}^m p_k \frac{\partial f_k(\mathbf{p})}{\partial p_i} = 0 \quad \text{for all } i \in [m]. \quad (18)$$

*Proof.* We minimize  $L_N(\mathbf{p}, \mathbf{r})$  over  $\mathbf{r} \in \Delta^{m-1}$  using the method of Lagrange multipliers. Define the Lagrangian:

$$\mathcal{J}(\mathbf{r}, \lambda) = \sum_{k=1}^m p_k f_k(\mathbf{r}) - \lambda \left( \sum_{k=1}^m r_k - 1 \right).$$

At a minimizer  $\mathbf{r} = \mathbf{p}$ , the stationarity condition requires  $\frac{\partial}{\partial r_i} \mathcal{J}(\mathbf{r}, \lambda) = 0$  for all  $i \in [m]$ . This yields

$$\frac{\partial}{\partial r_i} \left( \sum_{k=1}^m p_k f_k(\mathbf{r}) \right) \Big|_{\mathbf{r}=\mathbf{p}} = \lambda \quad \text{for all } i \in [m].$$

That is,

$$\sum_{k=1}^m p_k \frac{\partial f_k(\mathbf{p})}{\partial p_i} = \lambda.$$

930 Multiplying both sides by  $p_i$  and summing over  $i \in [m]$  gives:

$$\sum_{i=1}^m p_i \lambda = \sum_{i=1}^m p_i \sum_{k=1}^m p_k \frac{\partial f_k(\mathbf{p})}{\partial p_i} = \sum_{k=1}^m \left( p_k \cdot \langle \mathbf{p}, \nabla f_k(\mathbf{p}) \rangle \right) = \sum_{k=1}^m (p_k \cdot 0) = 0,$$

931 where the third equality holds because  $f_k$  is 0-homogeneous and thus  $\langle \mathbf{p}, \nabla f_k(\mathbf{p}) \rangle = 0$  by Euler's  
 932 theorem. Thus,  $\lambda = 0$ , and we have  $\sum_{k=1}^m p_k \frac{\partial f_k(\mathbf{p})}{\partial p_i} = 0$ , as claimed.  $\square$

933 We now connect this condition to a gradient field characterization.

934 **Theorem B.3.** For any learning algorithm  $\mathcal{A}$ , one of the following two scenarios must hold:

- 935 1.  $L^{\text{same}}(\mathbf{p}) = L_N^*(\mathbf{p})$  holds only for a zero-measure subset of  $\mathbf{p} \in \Delta^{m-1}$ ;
- 936 2.  $\nabla L^{\text{same}}(\mathbf{p}) = (f_1(\mathbf{p}), \dots, f_m(\mathbf{p}))$ .

937 *Proof.* Let  $\Omega_i$  denote the set of  $\mathbf{p} \in \Delta_+^{m-1}$  for which the gradient condition (18) holds for index  
 938  $i \in [m]$ . By Lemma B.1, the function  $f_k$  is a rational function of  $\mathbf{p}$ . It follows that both  $\frac{\partial f_k(\mathbf{p})}{\partial p_i}$  and  
 939  $\sum_{k=1}^m p_k \frac{\partial f_k(\mathbf{p})}{\partial p_i}$  are also rational functions of  $\mathbf{p}$ . Therefore,  $\Omega_i$  is the zero set of a rational function,  
 940 and must be either a measure-zero subset of  $\Delta_+^{m-1}$  or the entire domain.

941 Let  $\Omega := \bigcap_{i \in [m]} \Omega_i$  be the intersection of all  $\Omega_i$ . Then  $\Omega$  is either a zero-measure subset of  $\Delta_+^{m-1}$   
 942 or the entire domain. If  $\Omega$  is a zero-measure subset, then by Lemma B.2, we are in the first case  
 943 of the theorem. If  $\Omega$  is the entire domain, then the gradient condition (18) holds for all  $i \in [m]$ ,  
 944  $\mathbf{p} \in \mathbb{R}_{\geq 0}^m \setminus \{\mathbf{0}\}$ .

Recall that  $L^{\text{same}}(\mathbf{p}) := \sum_{k=1}^m p_k f_k(\mathbf{p})$ . Then we compute:

$$\frac{\partial L^{\text{same}}(\mathbf{p})}{\partial p_i} = f_i(\mathbf{p}) + \sum_{k=1}^m p_k \frac{\partial f_k(\mathbf{p})}{\partial p_i}.$$

945 By the gradient condition (18),  $\sum_{k=1}^m p_k \frac{\partial f_k(\mathbf{p})}{\partial p_i} = 0$ . Thus,  $\frac{\partial L^{\text{same}}(\mathbf{p})}{\partial p_i} = f_i(\mathbf{p})$ , which implies  
 946  $\nabla L^{\text{same}}(\mathbf{p}) = (f_1(\mathbf{p}), \dots, f_m(\mathbf{p}))$ , which is the second case of the theorem.  $\square$

947 Finally, Theorem B.3 implies Theorem 7.2.

## 948 B.2 Characterization of Conservation Conditions

949 *Proof of Lemma 7.3.* If Condition 7.1 holds, then for all  $i, j \in [m]$  ( $i \neq j$ ),

$$\frac{\partial}{\partial p_j} f_i(\mathbf{p}) = \frac{\partial^2}{\partial p_i \partial p_j} L^{\text{same}}(\mathbf{p}) = \frac{\partial}{\partial p_i} f_j(\mathbf{p}).$$

950 By the chain rule, we have  $\frac{\partial}{\partial p_j} f_i(\mathbf{p}) = -\frac{p_i}{|\mathbf{p}|^2} g'_i(\frac{p_i}{|\mathbf{p}|})$  and  $\frac{\partial}{\partial p_i} f_j(\mathbf{p}) = -\frac{p_j}{|\mathbf{p}|^2} g'_j(\frac{p_j}{|\mathbf{p}|})$ . Thus, for all  
 951  $\mathbf{p} \in \mathbb{R}_{\geq 0}^m \setminus \{\mathbf{0}\}$ ,

$$-\frac{p_i}{|\mathbf{p}|^2} g'_i(\frac{p_i}{|\mathbf{p}|}) = -\frac{p_j}{|\mathbf{p}|^2} g'_j(\frac{p_j}{|\mathbf{p}|}).$$

952 For any  $x, y > 0$  with  $x + y < 1$ , we can choose  $\mathbf{p}$  such that  $\frac{p_i}{|\mathbf{p}|} = x$  and  $\frac{p_j}{|\mathbf{p}|} = y$ . Then we  
 953 have  $x g'_i(x) = y g'_j(y)$  for all such  $x, y$ . This is only possible if there exists a constant  $C$  such that  
 954  $x g'_i(x) = C$  for all  $x \in (0, 1)$ . Solving this gives  $g'_i(x) = \frac{C}{x}$ , which implies that  $g_i(x) = C \ln x + A$   
 955 for some constant  $A$ .

956 Since  $g_i(x)$  has no singularity at  $x = 0$ , we must have  $C = 0$ . Thus,  $g_i(x)$  is a constant function.  $\square$

957 *Proof of Lemma 7.5.* Let  $\Delta(\mathbf{p}) = L^{\text{same}}(\mathbf{p}) - \hat{L}^{\text{same}}(\mathbf{p})$  be the difference between the two losses  
 958 when training on the same distribution. By Condition 7.1, we have

$$\frac{\partial}{\partial p_i} \Delta(\mathbf{p}) = \frac{\partial}{\partial p_i} L^{\text{same}}(\mathbf{p}) - \frac{\partial}{\partial p_i} \hat{L}^{\text{same}}(\mathbf{p}) = f_i(\mathbf{p}) - \hat{f}_i(\mathbf{p}) = 0. \quad (19)$$

959 Therefore,  $\Delta(\mathbf{p})$  is independent of  $p_i$ , and there exists a function  $C : \mathbb{R}^{m-1} \rightarrow \mathbb{R}$ ,  $\mathbf{p}_{-i} \mapsto C(\mathbf{p}_{-i})$   
 960 such that  $\Delta(\mathbf{p}) = C(\mathbf{p}_{-i})$  for all  $\mathbf{p} \in \mathbb{R}_{\geq 0}^m \setminus \{\mathbf{0}\}$ , where  $\mathbf{p}_{-i} = (p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_m)$ . This  
 961 is because we can set  $C(\mathbf{p}_{-i}) = \Delta(\mathbf{p})|_{p_i=0}$  and then take the integral of  $\frac{\partial}{\partial p_i} \Delta(\mathbf{p})$  over  $p_i$  to get  
 962  $\Delta(\mathbf{p}) = C(\mathbf{p}_{-i})$ .

963 Next, note that both  $L^{\text{same}}(\mathbf{p})$  and  $\hat{L}^{\text{same}}(\mathbf{p})$  can be written as rational functions of the form

$$L^{\text{same}}(\mathbf{p}) = \frac{S(\mathbf{p})}{|\mathbf{p}|^n}, \quad \hat{L}^{\text{same}}(\mathbf{p}) = \frac{\hat{S}(\mathbf{p})}{|\mathbf{p}|^n},$$

964 where  $n$  is the dataset size. This is because  $L^{\text{same}}(\mathbf{p}) = \sum_{k=1}^m p_k f_k(\mathbf{p})$ .

965 Now we show that  $\Delta(\mathbf{p})$  must have the form  $\Delta(\mathbf{p}) = \sum_{k \neq i} C_k p_k$  for some constants  $C_k$ . Let  
 966  $D(\mathbf{p}) := S(\mathbf{p}) - \hat{S}(\mathbf{p})$ . Since  $D(\mathbf{p})$  is a polynomial,  $C(\mathbf{p}_{-i}) = \frac{D(\mathbf{p})}{|\mathbf{p}|^n}$  must be a rational function.

967 Let  $C(\mathbf{p}_{-i}) = \frac{A(\mathbf{p}_{-i})}{B(\mathbf{p}_{-i})}$  for some polynomials  $A(\mathbf{p}_{-i})$  and  $B(\mathbf{p}_{-i})$ . Then

$$D(\mathbf{p}) B(\mathbf{p}_{-i}) = A(\mathbf{p}_{-i}) |\mathbf{p}|^n.$$

968 If  $A = 0$ , then  $\Delta(\mathbf{p}) = 0$ . Otherwise, both  $A(\mathbf{p}_{-i})$  and  $B(\mathbf{p}_{-i})$  are non-zero polynomials. Since  
 969  $B(\mathbf{p}_{-i})$  cannot be divisible by  $|\mathbf{p}|^n$ ,  $D$  must be divisible by  $|\mathbf{p}|^n$ . Note that  $D$  is a  $(n+1)$ -  
 970 homogeneous polynomial and  $|\mathbf{p}|^n$  is  $n$ -homogeneous, so  $C(\mathbf{p}_{-i}) = \frac{D}{|\mathbf{p}|^n}$  must be a 1-homogeneous

polynomial. The only 1-homogeneous polynomials in variables  $\mathbf{p}_{-i}$  are linear functions of the form  $C(\mathbf{p}_{-i}) = \sum_{k \neq i} C_k p_k$  for some constants  $C_k$ . Thus, no matter  $A = 0$  or not, we have  $\Delta(\mathbf{p}) = \sum_{k \neq i} C_k p_k$ .

Finally, by Condition 7.1, we can compute for all  $k \neq i$  that

$$f_k(\mathbf{p}) - \hat{f}_k(\mathbf{p}) = \frac{\partial}{\partial p_k} \Delta(\mathbf{p}) = C_k,$$

which implies that  $f_k(\mathbf{p}) = \hat{f}_k(\mathbf{p}) + C_k$ , as desired.  $\square$

*Proof of Corollary 7.6.* This follows from Lemma 7.5. Note that Lemma 7.5 implies that for every  $k$  and corresponding error function  $e_k(\mathbf{n})$ , there exists at most one tuple of error functions  $\{e_j\}_{j=1, j \neq k}^K$  (up to a individual constant offset for each error function  $e_j$ ) that positive distribution shift does not happen for  $\mathbf{p}$  of positive measure. This implies the corollary.  $\square$

## C Experiment Details

**Model Architecture and Tokenizer.** We use a model architecture similar to GPT-2, except that we use RoPE instead of absolute position embedding. Our model has 6 layers, 8 attention heads, and 512 embedding dimensions. We use the same tokenizer as GPT-2, which is a byte-pair encoding (BPE) tokenizer.

**Generation of Skills.** We randomly generate  $M = 10^5$  skills. For each skill, we randomly sample 3 English tokens and concatenate them to form the skill name. The first token is sampled from a set of 1000 tokens that start with a blank space and then a capital letter. The second and third tokens are sampled from a set of 1000 tokens that start with a capital letter without a blank space. The starting blank space is to ensure that the skill name is tokenized into exactly 3 tokens when placed in a prompt with space-separated skill names. For example, “CourtClientCheck” can be a skill name (with blank space removed). Then, for each skill  $i$ , we uniformly randomly sample a function  $g_i$  that maps a number from  $\{0, \dots, 9\}$  to  $\{0, \dots, 9\}$ .

**Distribution: Skill Composition.** For each data point, a number  $k$  is sampled uniformly from  $\{10, \dots, 50\}$ , then a set of  $k$  skills  $g_{i_1}, \dots, g_{i_k}$  are sampled IID following a power law  $p(i) \propto (i + 50)^{-\alpha}$  with exponent  $\alpha = 1.5$ . The text consists of two parts. The input part is as follows:

```
<|begin_of_text|> Input:
[x] -> [skill name 1] -> [skill name 2] -> ... -> [skill name k]
```

The output part is as follows:

```
Output:
[x] -> [skill name 1] = [x1]
[x1] -> [skill name 2] = [x2]
[x2] -> [skill name 3] = [x3]
...
[xk-1] -> [skill name k] = [xk]
[xk]
```

The input and output parts are concatenated together with a blank line in between.

**Distribution: Uniform Skills.** For each data point, we randomly sample a skill uniformly from the skill set. Then the text is as follows:

```
<|begin_of_text|> [x] [skill name] = [expected output]
```

**Evaluation.** We evaluate the test accuracy of the model on skill composition task with CoT reasoning. We sample 400 data points from the skill composition task, but fix  $k$  to be 10, 30, 50. For each data point, only the input part is given to the model, and the model’s autoregressive output is considered as correct if the last line of the output is the same as the expected output.



1014 **Training with Matched Distribution.** In the training, we use batch size 512 and maximum  
1015 sequence length 2048. We perform sequence packing for each sequence in the batch: we sample data  
1016 points from the skill composition task until the maximum sequence length is reached. We train the  
1017 model on 4 A4000 GPUs for at most 40K steps.

1018 **Training with Mismatched Distribution.** Similar as above, but for each sequence, we first choose  
1019 the task to be skill composition or uniform skills with probability 70% and 30% respectively. Then  
1020 we sample data points from the chosen task until the maximum sequence length is reached.

1021 **Results.** We show the results in Figure 3. We see that training with mismatched distribution  
1022 significantly outperforms training with matched distribution.