# Explaining Data Mixing Scaling Laws

**author names withheld**

**Under Review for the Workshop on High-dimensional Learning Dynamics, 2026**


## Abstract

Recent research has established empirical scaling laws to predict model performance on multi-domain data mixtures. However, a theoretical understanding of these model loss behaviors remains limited. In this work, we propose a unified framework to explain the underlying mechanics of data mixing. Our approach extends theoretical perspectives originally developed for standard neural scaling laws (e.g., Kaplan and Chinchilla) to the multi-domain setting. Based on the distributional assumption that domains overlap on fundamental skills while diverging on specialized skills, we identify two key factors that decide the domain loss of models trained on different data mixtures: *Capacity Competition*, where the allocation of finite model capacity couples domain losses globally, and *Noise Reduction*, where optimal weights shift toward harder-to-learn domains to minimize variance. Experiments demonstrate that our framework fits the loss landscape with lower Mean Relative Error than existing empirical baselines and accurately predicts optimal training mixtures. Most importantly, our model successfully extrapolates across scales, predicting the optimal mixture at a large, unseen scale using parameters trained on a smaller one. In addition, our model achieves these results using significantly fewer parameters compared to previous empirical laws.


## 1. Introduction

Large foundation models are typically trained on data from multiple domains, with the data mixture—the proportion of each domain used—playing a critical role in model performance. However, discovering the optimal data mixture is often a highly costly process that lacks principled methodologies. Practitioners often rely on expensive trial-and-error or static heuristics.

To address this, recent research has moved toward principled methods such as developing scaling laws for data mixing [4, 7, 17, 18]. These empirical frameworks attempt to predict a model's validation loss on specific domains as a function of the mixing weights used during training. Several functional forms have been proposed to model this relationship, as summarized in. These laws deviate from the standard power-form scaling laws and present non-trivial domain interaction: the loss on a domain depends not only on the weight of the domain itself, but also the weights of other domains, and this correlation does not exhibit a simple functional form.

Table 1: Comparison of empirical functional forms for predicting domain loss $L_i(h)$. $N$ and $D$ represent model parameters and training tokens, respectively.

| Reference | Functional Form | Reference | Functional Form |
|---|---|---|---|
| [17] | $L_i \approx E_i + \left(\sum_{j=1}^{K} C_{ij} h_j^{\gamma_{ij}}\right)^{-1}$ | [18] | $L_i \approx c_i + k_i \exp\left(\sum_{j=1}^{K} t_{ij} h_j\right)$ |

The paper has been accepted for ICML 2026.

Despite the utility of these empirical fits, a theoretical understanding of the mechanics driving domain interaction remains absent. Relying solely on empirically fitted curves presents a significant bottleneck. Not only is the fitting process resource-intensive, but the resulting laws act as "black boxes": it is unclear whether they generalize to larger scales or different datasets, nor is it obvious how to map the predicted domain loss to downstream task performance.

In this work, we propose a unified theoretical framework to explain the underlying mechanics of data mixing. Our framework extends previous theoretical frameworks for standard neural scaling laws—specifically the Quantization Model [12, 14] and the Projected Linear Regression Model [2, 10]—to the multi-domain setting. Based on a natural distributional assumption that different domains overlap on fundamental skills and diverge on specialized skill, we identify two key factors that decide the loss of models trained on different data mixtures:

**Model Capacity Competition:** The model has a finite capacity and can only learn a finite number of skills. The specialized skills from different domains compete for the model capacity. Adjusting domain weights will change the importance of the skills and thus change the model capacity allocated to each domain. The resulting model capacity allocation introduces a non-trivial domain interaction, and is a key factor that determines the trained model's loss on a domain.

**Noise Driven by Data Amount:** For each skill within the model's capacity, the loss incurred by the skill depends on the number of times that the model has seen the skill. As skills from different domains have different difficulty levels, the loss decreases at different speed as the domain weight increases; this dynamic shifts the optimal mixture weights toward domains that are harder to learn.

Leveraging this theoretical framework, we formulate loss prediction as a convex program that yields numerical estimates for arbitrary mixtures. Furthermore, we frame the search for the optimal training mixture as a bi-level optimization problem, which can be efficiently solved using Online Mirror Descent.

Experimentally, our results validate the correctness of our theoretical framework and demonstrate its effectiveness:

- **Superior Fitting Accuracy:** Our models fit the observed loss landscape with lower Mean Relative Error (MRE) than existing empirically fitted scaling laws.
- **Optimal Mixture Prediction:** Our framework effectively identifies optimal training mixtures that yield the lowest validation loss on the target average distribution.
- **Parameter Efficiency:** Crucially, we achieve these results while utilizing significantly fewer free parameters compared to leading empirical laws.
- **Model Size Extrapolation:** Most importantly, our model successfully extrapolates across scales, predicting the optimal mixture at a large, unseen scale using parameters trained on a smaller one.

## 2. Problem Description

In this section, we formally define the problem of data mixing and review prior empirical work on scaling laws in the multi-domain setting.

**Problem setup.** Consider a set of $K$ data domains $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$, where distinct domains represent different data sources such as GitHub, books, and Wikipedia. We construct a training data mixture by sampling from these domains according to a domain weight vector $h \in \Delta^K$, where $\Delta^K$ denotes the probability simplex $\Delta^K = \left\{h \in \mathbb{R}^K \mid \sum_{i=1}^K h_i = 1,\ h_i \geq 0\right\}$. Let $\theta(h)$ denote the parameters of a model trained on this mixture with fixed model size $N$ and training token count $D$.

The primary objective of data mixing is to determine the optimal domain weights $h^*$ that minimize the trained model's loss on a specific target distribution defined by importance weights $w \in \Delta^K$, under a fixed compute budget parameterized by model size $N$ and training token count $D$. Since training large models to evaluate every potential mixture is computationally prohibitive, this is typically achieved by fitting an *empirical data mixing law*. The goal of such laws is to predict the validation loss $L_i$ on the held-out set of each domain $\mathcal{D}_i$ as a function of the mixing weights $h$ and potentially the scale parameters $N$ and $D$:

$$L_i(h, N, D) \approx f_i(h, N, D).$$

Once the functional form $f_i$ is fitted, the optimal training mixture $h^*$ for a target distribution defined by importance weights $w \in \Delta^K$ can be identified by solving the following minimization problem:

$$h^* = \arg\min_{h \in \Delta^K} \sum_{i=1}^K w_i f_i(h, N, D).$$

**Empirical data mixing laws.** Recent research has proposed various functional forms to model the relationship between the mixing weights $h$ and the resulting domain losses. The standard methodology involves sampling a diverse set of mixture configurations from the simplex $\Delta^K$, training small-to-medium scale models on these mixtures, and fit the coefficients of $f_i$ against the observed validation losses. Table 3 summarizes the key functional forms established in the literature.

## 3. Theoretical Framework for Data Mixing

In this work, we propose a unified theoretical framework to explain the underlying mechanics of data mixing. Our approach extends the single-domain theoretical perspectives—specifically the Quantization Model [14] and the Linear Regression Model [2, 10]—to the multi-domain setting. We will first introduce the Extended Quantization Model, which views training as a process of model capacity allocation. This model serves as the foundation for our Extended Linear Regression Model, which implicitly solves this capacity allocation problem while incorporating an additional noise term. This allocation of finite model capacity across competing domains is viewed as a critical factor driving domain interaction. We begin by defining a structural assumption regarding domain overlap.

### 3.1. The "Shared Head, Disjoint Tail" Structure

Building on [15], we introduce a natural structural assumption regarding how information overlaps across different data domains (e.g., Code, Math, English). Intuitively, most domains share a common foundation of basic knowledge while diverging in specialized topics. We formalize this as the "Shared Head, Disjoint Tail" structure:

- **Power-Law Distribution:** Within each domain $i$, knowledge units (skills) follow a power-law distribution in terms of frequency.
- **Shared Head:** Different domains largely overlap in the head of the distribution—the region of high-probability, fundamental skills (e.g., basic grammar, logic, or arithmetic).
- **Disjoint Tail:** As we move to the tail of the distribution (rare, specialized knowledge), the domains become increasingly distinct and independent.

While real-world data is unlikely to be strictly disjoint in the tail, this idealization serves as a useful approximation. Based on our modeling in the following sections, when the data size is fixed and only the mixture weights vary, the change in loss induced by overlapping skills is likely to be small compared to the change induced by disjoint skills. In other words, the loss induced by overlapping skills will behave more like a constant compared to disjoint skills.

### 3.2. Main Theoretical Model

Under the "Shared Head, Disjoint Tail" assumption, we extend the single-domain theoretical models of [2, 10] to a multi-domain setting. In our framework, the expected test loss comprises two components.

**1. Model capacity allocation.** First, the model has a finite capacity $N$ and can learn at most $N$ skills. As the model is trained to minimize loss, the training process implicitly solves a capacity allocation problem: when training on a data mixture $h = (h_1, \ldots, h_K)$, the model implicitly determines the capacity $x_i$ allocated to each domain $i$ by minimizing the total expected loss:

$$\min_{x} \left\{ \sum_{i=1}^{K} h_i c_i x_i^{-b_i} \;\middle|\; \sum_{i=1}^{K} (x_i - H) \leq N - H, \quad x_i \geq H, \forall i \right\} \tag{1}$$

Letting $x^*(h) = (x_1^*(h), \ldots, x_K^*(h))$ denote the optimal allocation, the unlearned skills from domain $i$ incur a loss of $c_i (x_i^*(h))^{-b_i}$.

**2. Stochastic noise.** Second, for each skill within the model's capacity, the loss will depend on the number of times it has been observed; in other words, the stochastic nature of one-pass SGD introduces a noise term $A_i (D h_i)^{-a_i}$, which is determined by the number of training samples drawn from that domain.

**Theorem 1 (Informal)** *Consider a projected linear model $\theta \in \mathbb{R}^N$ trained via one-pass SGD on $D$ samples drawn from a mixture $h$, its expected test loss on domain $i$, denoted $L_i(h, N, D)$, satisfies*

$$L_i(h, N, D) \approx c_i x_i^*(h, N)^{-b_i} + A_i (D h_i)^{-a_i} + E_i \tag{2}$$

*where $a_i, A_i, E_i$ are constants that depend on the data distribution.*

## 4. Experiments

In this section, we empirically validate our proposed theoretical framework. We focus on two primary objectives:

1. **Predictive Accuracy:** We evaluate how well our theoretical model fits the observed loss landscape under various data mixtures compared to existing empirical baselines. Accuracy is assessed by the Mean Relative Error (MRE) and Mean Absolute Error (MAE).
2. **Optimal Mixture Extrapolation:** We leverage the fitted scaling laws to predict the optimal training mixture at a larger, unseen model scale, and then evaluate its test loss.

Table 2: Comparison of fitting accuracy on 64 1B-parameter models trained on $K = 17$ domains from the Pile dataset. Our theoretically grounded models achieve the lowest error rates (MRE and MAE) while using significantly fewer total parameters than heuristic baselines.

| **Method** | **MRE (%)** $\downarrow$ | **MAE** $\downarrow$ | **#Param** |
|---|---|---|---|
| *Empirical Baselines* | | | |
| Additive | 2.209 | 0.052 | $K(2K+1)$ |
| Exponential | 6.990 | 0.059 | $K(K+2)$ |
| BiMix [4] | 2.963 | 0.144 | $2K$ |
| RegMix [11] | 6.480 | 0.136 | $K^2$ |
| *Theoretical Models (Ours)* | | | |
| Ours (Eq. (3)) | 2.064 | 0.051 | $3K$ |
| **Ours (Eq. 4)** | **1.533** | **0.034** | $5K$ |

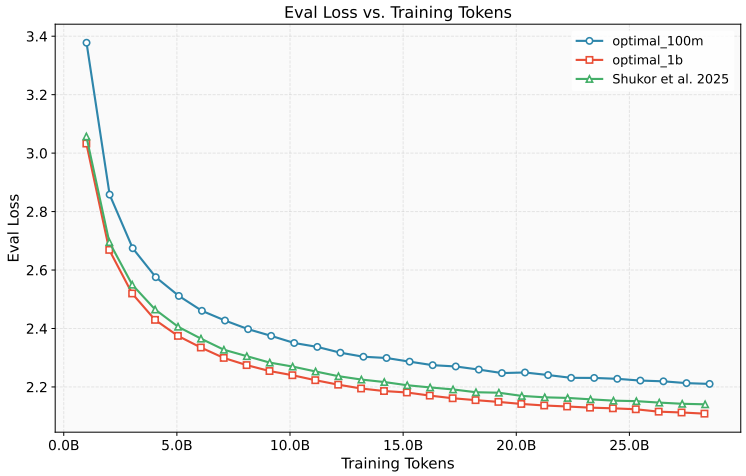


Figure 1: We fit both our proposed scaling law and the scaling law of [17] on 100M model/10B token mixtures. We then train 1B models with 30B tokens using three different mixtures: (1) the optimal 1B model/30B token mixture predicted by our law, extrapolated to the target scale by scaling $N$ and $D$ proportionally (orange curve). (2) our mixture optimized directly for the 100M model (blue curve), and (3) the optimal mixture predicted by the scaling law of [17] (green curve). The results demonstrate that our extrapolated mixture achieves the best overall training performance. Notably, at the 1B scale, it outperforms the mixture that was optimal for the smaller 100M model, confirming that our framework accurately predicts shifts in the ideal data mixture as model size increases.

## References


[1] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi: 10.1073/pnas.2311878121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2311878121.

[2] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, pages 4345–4382. PMLR, 2024.

[3] David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, Dec 1987. doi: 10.1364/JOSAA.4.002379. URL https://opg.optica.org/josaa/abstract.cfm?URI=josaa-4-12-2379.

[4] Ce Ge, Zhijian Ma, Daoyuan Chen, Yaliang Li, and Bolin Ding. Bimix: A bivariate data mixing law for language model pretraining. *arXiv preprint arXiv:2405.14908*, 2024.

[5] Xinran Gu, Kaifeng Lyu, Jiazheng Li, and Jingzhao Zhang. Data mixing can induce phase transitions in knowledge acquisition. *arXiv preprint arXiv:2505.18091*, 2025.

[6] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[7] Feiyang Kang, Yifan Sun, Bingbing Wen, Si Chen, Dawn Song, Rafid Mahmood, and Ruoxi Jia. Autoscale: Scale-aware data mixing for pre-training llms. *arXiv preprint arXiv:2407.20177*, 2024.

[8] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[9] Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. *arXiv preprint arXiv:2509.19189*, 2025.

[10] Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter L. Bartlett, and Jason D. Lee. Scaling laws in linear regression: compute, parameters, and data. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS 24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[11] Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*, 2024.

[12] Ziming Liu, Yizhou Liu, Eric J Michaud, Jeff Gore, and Max Tegmark. Physics of skill learning. *arXiv preprint arXiv:2501.12391*, 2025.

[13] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. URL https://arxiv.org/abs/1608.03983.

[14] Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36:28699–28722, 2023.

[15] Zhixuan Pan, Shaowen Wang, and Jian Li. Understanding llm behaviors via compression: Data generation, knowledge acquisition and scaling laws. *arXiv preprint arXiv:2504.09597*, 2025.

[16] Shikai Qiu, Lechao Xiao, Andrew Gordon Wilson, Jeffrey Pennington, and Atish Agarwala. Scaling collapse reveals universal dynamics in compute-optimally trained neural networks, 2025. URL https://arxiv.org/abs/2507.02119.

[17] Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. *arXiv preprint arXiv:2507.09404*, 2025.

[18] Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *arXiv preprint arXiv:2403.16952*, 2024.

Table 3: Comparison of empirical functional forms for predicting domain loss $L_i(h)$ based on mixture weights $h$. $N$ and $D$ represent model parameters and training tokens, respectively.

| **Law Name** | **Functional Form** ($f_i(h, N, D)$) |
|---|---|
| **Additive** [17] | $L_i \approx E_i + \left(\sum_{j=1}^{K} C_{ij} h_j^{\gamma_{ij}}\right)^{-1} + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$ |
| **Exponential** [18] | $L_i \approx c_i + k_i \exp\left(\sum_{j=1}^{K} t_{ij} h_j\right)$ |
| **BiMix** [4] | $L_i \approx (\frac{B}{D^{\beta}} + E)\frac{C}{h_i^{\gamma}}$ |
| **RegMix** [11] | $L_i \approx w_0 + \sum_{j=1}^{K} w_j h_j$ |

## Appendix A. Preliminaries and Problem Description

This section establishes the theoretical background and introduces the problem of data mixing. We begin by reviewing two primary theoretical frameworks that explain the standard neural scaling

laws in the single-domain setting: the Quantization Model and the Linear Regression Model. Subsequently, we formally define the data mixing problem and survey existing empirical data mixing laws used to predict multi-domain loss.

## A.1. Theoretical Foundations of Standard Neural Scaling Laws

### A.1.1. QUANTIZATION MODEL

Standard neural scaling laws describe the predictable power-law relationship between model performance and scale. Empirically, the test loss $L$ is typically modeled as a function of both model size $N$ and dataset size $D$ [6, 8]:

$$L(N, D) \approx \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + E.$$

While these laws are well-established empirically, theoretical understanding of their origin remains an active area of research. Below, we review two major frameworks that attribute this power-law scaling to the intrinsic power-law structure of the data distribution.

[14] propose the *Quantization Model*, which posits that the knowledge contained within large-scale training corpora decomposes into discrete knowledge units or skills termed "quanta," $\mathcal{Q} = \{q_1, q_2, \dots\}$. These quanta are assumed to follow a Zipfian distribution where the $k$-th most frequent quantum has probability:

$$p(q_k) \propto k^{-\alpha}, \quad (\alpha > 1).$$

Crucially, the framework asserts that when trained on this dataset, a model learns skills in a deterministic order based on frequency to minimize the expected loss. Specifically, a model of effective capacity $N$ is assumed to learn the top $N$ most frequent quanta (i.e., $\{q_1, \dots, q_N\}$). Assuming that each unlearned quantum contributes a constant error $c$, the total expected loss is determined by the cumulative probability mass of the unlearned quanta in the tail:

$$L(N) = c \sum_{k=N+1}^{\infty} p(q_k) \approx c \int_{N}^{\infty} k^{-\alpha}\, dk \propto N^{-(\alpha-1)}.$$

As a result, the loss scaling exponent regarding model size $N$ is a direct consequence of the power-law distribution of skills inherent in the data.

### A.1.2. THE LINEAR REGRESSION MODEL.

Despite its elegance, the Quantization Model is limited: it does not capture the stochasticity inherent in training dynamics. Building on similar intuitions, [2, 10] provide a rigorous derivation of scaling laws by analyzing the training dynamics of linear regression under one-pass Stochastic Gradient Descent (SGD). In this framework, neural scaling laws are governed by the spectral decay of the data. As training progresses, the model "resolves" eigenmodes in descending order of their eigenvalues—learning the dominant patterns first before fitting the fine-grained details. Below, we adopt the formal framework from [10] to provide a simplified theoretical explanation.

**Data Generation:** Consider a linear regression problem where the input covariates $\mathbf{x} \in \mathbb{R}^d$ (where $d$ can be infinite) are drawn from a distribution with zero mean and covariance matrix $\mathbf{H} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$. The target label $y$ is generated by a linear teacher with additive noise:

$$y = \langle \theta_*, \mathbf{x} \rangle + \epsilon,$$

where $\theta_*$ is the ground-truth parameter and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent Gaussian noise.

**Spectral Assumptions:** When training a linear regression model with one-pass SGD, the learning dynamics are determined by the spectrum of the covariance matrix $\mathbf{H} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$. Intuitively, the eigenvalues $\lambda_k$ of $\mathbf{H}$ represent the variance (or signal strength) of the data along the $k$-th principal component. Empirical studies on natural data (e.g., images and text) consistently observe that these eigenvalues follow a power distribution [1, 3]. It is therefore assumed that the eigenvalues, sorted in descending order, follow a power law:

$$\lambda_k \propto k^{-\alpha}, \quad \text{for } \alpha > 1.$$

**Finite Parameter Projection:** To model a neural network with a finite capacity of $N$ parameters, we project the high-dimensional input $\mathbf{x}$ into a lower-dimensional feature space using a "sketching matrix" $\mathbf{S} \in \mathbb{R}^{N \times d}$. The model learns a weight vector $\theta \in \mathbb{R}^N$ by minimizing the squared error on the projected features $\widetilde{\mathbf{x}} = \mathbf{S}\mathbf{x}$:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^N} \frac{1}{D} \sum_{i=1}^{D} (y_i - \theta^\top \mathbf{S}\mathbf{x}_i)^2.$$

**Result:** Under this setup, [10] derive the scaling law for the test loss $L(N, D)$ of the projected linear model $\widehat{w}$ (trained via one-pass SGD) as a function of the model size $N$ and training samples $D$:

$$L(N, D) \approx \underbrace{O\left(\frac{1}{N^{a_1}}\right)}_{\text{Model Scaling}} + \underbrace{O\left(\frac{1}{D^{a_2}}\right)}_{\text{Data Scaling}} + E.$$

### A.2. Empirical Data Mixing Laws

In this section, we formally define the problem of data mixing and review prior empirical work on scaling laws in the multi-domain setting.

**Problem setup.** Consider a set of $K$ data domains $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$, where distinct domains represent different data sources such as GitHub, books, and Wikipedia. We construct a training data mixture by sampling from these domains according to a domain weight vector $h \in \Delta^K$, where $\Delta^K$ denotes the probability simplex $\Delta^K = \left\{h \in \mathbb{R}^K \mid \sum_{i=1}^K h_i = 1,\ h_i \geq 0\right\}$. Let $\theta(h)$ denote the parameters of a model trained on this mixture with fixed model size $N$ and training token count $D$.

The primary objective of data mixing is to determine the optimal domain weights $h^*$ that minimize the trained model's loss on a specific target distribution defined by importance weights $w \in \Delta^K$, under a fixed compute budget parameterized by model size $N$ and training token count $D$. Since training large models to evaluate every potential mixture is computationally prohibitive, this is typically achieved by fitting an *empirical data mixing law*. The goal of such laws is to predict the validation loss $L_i$ on the held-out set of each domain $\mathcal{D}_i$ as a function of the mixing weights $h$ and potentially the scale parameters $N$ and $D$:

$$L_i(h, N, D) \approx f_i(h, N, D).$$

Once the functional form $f_i$ is fitted, the optimal training mixture $h^*$ for a target distribution defined by importance weights $w \in \Delta^K$ can be identified by solving the following minimization problem:

$$h^* = \arg\min_{h \in \Delta^K} \sum_{i=1}^{K} w_i f_i(h, N, D).$$

**Empirical data mixing laws.** Recent research has proposed various functional forms to model the relationship between the mixing weights $h$ and the resulting domain losses. The standard methodology involves sampling a diverse set of mixture configurations from the simplex $\Delta^K$, training small-to-medium scale models on these mixtures, and fit the coefficients of $f_i$ against the observed validation losses. Table 3 summarizes the key functional forms established in the literature.

## Appendix B. Theoretical Framework for Data Mixing

In this work, we propose a unified theoretical framework to explain the underlying mechanics of data mixing. Our approach extends the single-domain theoretical perspectives—specifically the Quantization Model [14] and the Linear Regression Model [10]—to the multi-domain setting. We will first introduce the Extended Quantization Model, which views training as a process of model capacity allocation. This model serves as the foundation for our Extended Linear Regression Model, which implicitly solves this capacity allocation problem while incorporating an additional noise term. This allocation of finite model capacity across competing domains is viewed as a critical factor driving domain interaction. We begin by defining a structural assumption regarding domain overlap.

### B.1. The "Shared Head, Disjoint Tail" Structure

Building on [15], we introduce a natural structural assumption regarding how information overlaps across different data domains (e.g., Code, Math, English). Intuitively, most domains share a common foundation of basic knowledge while diverging in specialized topics. We formalize this as the "Shared Head, Disjoint Tail" structure:

- **Power-Law Distribution:** Within each domain $i$, knowledge units (skills) follow a power-law distribution in terms of frequency.
- **Shared Head:** Different domains largely overlap in the head of the distribution—the region of high-probability, fundamental skills (e.g., basic grammar, logic, or arithmetic).
- **Disjoint Tail:** As we move to the tail of the distribution (rare, specialized knowledge), the domains become increasingly distinct and independent.

While real-world data is unlikely to be strictly disjoint in the tail, this idealization serves as a useful approximation. Based on our modeling in the following sections, when the data size is fixed and only the mixture weights vary, the change in loss induced by overlapping skills is likely to be small compared to the change induced by disjoint skills. In other words, the loss induced by overlapping skills will behave more like a constant compared to disjoint skills. We discuss the robustness of this approximation in detail in the following sections.

### B.2. Extended Quantization Model

We first extend the Quantization Model (Section A.1.1) to a multi-domain formulation. This extension serves as a foundation for the extended Linear Regression Model in the next section.

**Single-Domain Skill Distribution and Loss.** We associate each domain $\mathcal{D}_i$ with a continuous space of skills indexed by $k_i \in [1, \infty)$ with power-law density

$$p_i(k_i) = (\alpha_i - 1)k_i^{-\alpha_i}, \quad \text{for } \alpha_i > 1.$$

We consider a model with effective capacity $N$, representing the total volume of unique skills it can learn. Following the Quantization Model, we frame training as a **capacity allocation problem**. For each domain $i$, the model selects a coverage threshold $x_i \geq 1$, learning all high-frequency skills up to this cutoff ($k_i \leq x_i$) while discarding the tail. Assuming that each unlearned quantum in domain $i$ contributes a constant error $c_i$, the loss incurred on domain $i$ is determined by the probability mass of the unlearned tail ($k_i > x_i$):

$$L_i = c_i \int_{x_i}^{\infty} (\alpha_i - 1)k_i^{-\alpha_i}\, dk_i = c_i x_i^{-(\alpha_i - 1)} = c_i x_i^{-b_i},$$

where $b_i = \alpha_i - 1$.

**Multi-Domain Data Assumptions and Loss Minimization.** Under the "Shared Head, Disjoint Tail" assumption, we assume that the skill spaces are aligned such that the interval $k_i \in [1, H]$ corresponds to foundational knowledge shared across all domains (the "base" skills). We assume that the model capacity is sufficient to acquire all shared base skills. Consequently, the capacity allocation problem reduces to allocating the remaining model capacity to the disjoint tails. When trained on a data mixture $h = (h_1, \ldots, h_K)$, the model determines the coverage thresholds $x_i$ to minimize the weighted mixture loss:

$$\begin{aligned} \min_{x} \quad & L = \sum_{i=1}^{K} h_i c_i x_i^{-b_i} \\ \text{s.t.} \quad & \sum_{i=1}^{K} (x_i - H) \leq N - H, \\ & x_i \geq H, \quad \forall i. \end{aligned} \tag{3}$$

Let $x^*(h) = (x_1^*(h), \ldots, x_K^*(h))$ denote the optimal solution to the optimization problem in Eq. 3, which depends on the mixture weights $h$. Under this model, the resulting loss on domain $i$ is given by $L_i(h) = c_i(x_i^*(h))^{-b_i}$.

**Domain Interaction: Capacity Competition.** Under this formulation, domain interaction is driven by model capacity competition. The constraint $\sum(x_i - H) \leq N - H$ couples the domains together in a competition for model resources. In particular, when the scaling exponents are similar ($b_i \approx \bar{b}$) and the model capacity is large ($x_i \gg H$), we can derive an approximate closed-form solution using Lagrange multipliers:

$$x_i^*(h) \approx \frac{\left(b_i c_i h_i N^{\bar{b}+1}\right)^{\frac{1}{\bar{b}_i+1}}}{\left(\sum_{k=1}^{K} (b_k c_k h_k)^{\frac{1}{\bar{b}_k+1}}\right)^{\frac{\bar{b}+1}{\bar{b}_i+1}}}.$$

The resulting approximate loss on domain $i$ is given by $L_i(h) = c_i(x_i^*(h))^{-b_i}$. Crucially, this solution shows that the loss $L_i(h)$ on each domain depends on the "aggregate demand" of the mixture, represented by the denominator term $\sum(b_k c_k h_k)^{\frac{1}{b_k+1}}$. This term creates an explicit coupling: domain $i$'s loss is driven not just by its own properties, but by the weights and complexities of every competing domain (e.g., $h_j$ for $j \neq i$).

**The Effect of Tail Overlap.** Now suppose the tails of different domains overlap. To minimize expected loss, the model learns the $N$ skills with the highest expected loss, denoted by $c(\text{skill}) \cdot p(\text{skill})$, where $p(\text{skill})$ is the aggregate probability density across all domains and $c(\text{skill})$ is the loss incurred by not learning the skill. So the status of a skill is binary: learned vs. unlearned. Consequently, when the mixture weights vary, the loss fluctuation caused by overlapping skills is likely to be smaller than that of disjoint skills. This is because the aggregate probability density $p(\text{skill})$ of an overlapping skill will be relatively stable as the mixture shifts, making its status (learned vs. unlearned) less sensitive to weight changes. In other words, the loss induced by overlapping skills will behave more like a constant compared to disjoint skills.

**Limitation.** While this formulation naturally extends the Quantization Model, it has a critical limitation when predicting the optimal mixture. We formulate the search for the optimal mixture as a bi-level optimization problem. The outer objective minimizes the loss on a target distribution defined by importance weights $w = (w_1, \ldots, w_K)$, while the inner optimization determines the resource allocation $x^*(h)$ given the training mixture $h$:

$$h^* = \arg\min_h \sum_{i=1}^{K} w_i L_i(h) = \sum_{i=1}^{K} w_i c_i (x_i^*(h))^{-b_i}. \tag{4}$$

Here, $x_i^*(h)$ is the solution to the capacity allocation problem in Eq. (3). Crucially, because the inner optimization (Eq. 3) minimizes a weighted sum of losses with respect to $h$, and the outer objective (Eq. 4) minimizes a weighted sum of the same losses with respect to $w$, the bi-level problem is trivially solved by setting the training weights equal to the target weights $h^* \equiv w$.

This result contradicts empirical observations, where the optimal training mixture often deviates significantly from the target distribution. To resolve this discrepancy, we proceed to introduce the Extended Linear Regression Model.

### B.3. Extended Linear Regression Model

To address the limitations of the Extended Quantization Model and incorporate training dynamics, we extend the linear regression framework of [10] to the multi-domain setting. This Extended Linear Regression Model can be viewed as an extension of the previous Extended Quantization Model as well: the training process implicitly solves the capacity allocation problem defined in (3), while introducing an additional noise term.

**Problem Formulation.** Following Section A.1.2, we consider a linear regression problem over a union of $K$ domains. For each domain $i \in \{1, \ldots, K\}$, input covariates $\mathbf{x}_i \in \mathbb{R}^d$ (where $d$ can be infinite) are drawn from a distribution with zero mean and a covariance matrix defined as $\mathbf{H}_i = \mathbb{E}_{\mathcal{P}_i}[\mathbf{x}_i \mathbf{x}_i^\top]$. The label $y$ is generated by a global linear teacher $y_i = \langle \theta^*, \mathbf{x}_i \rangle + \epsilon_y$, where $\theta^* \sim \mathcal{N}(0, \mathbf{I})$ is the ground-truth parameter and noise $\epsilon_y \sim \mathcal{N}(0, \sigma^2)$. We consider a mixture distribution

$\mathcal{P}(h) = \sum_{i\in[K]} h_i\mathcal{P}_i$ defined by weights $h \in \Delta^{K-1}$ with covariance matrix $\mathbf{H}(h) = \sum_{i=1}^{K} h_i\mathbf{H}_i$. Following [9], we model a neural network with $N$ parameters by projecting the high-dimensional input $\mathbf{x}$ into a $N$-dimensional feature space using a "Top-N sketching matrix" $\mathbf{S} \in \mathbb{R}^{N\times d}$ (detailed in Section C.1). The model learns a weight vector $\theta \in \mathbb{R}^N$ by minimizing the squared error on the projected features $\widetilde{\mathbf{x}} = \mathbf{S}\mathbf{x}$.

**Spectral Assumption: Shared Head and Disjoint Tails.** We then formalize the "Shared Head, Disjoint Tail" assumption. In spectral analysis, an eigenvector represents a specific direction or pattern of variation in the data (e.g., a specific texture in images or topic in text), while its corresponding eigenvalue quantifies the variance (or strength) of that pattern. Intuitively, we assume real-world data consists of *universal patterns* shared across all domains (the head) and *specialized nuances* unique to each domain (the tails). To make the analysis tractable, we assume that all covariance matrices $\{\mathbf{H}_i\}_{i=1}^{K}$ share a common orthonormal basis of eigenvectors $\mathbf{U} = [u_1, \ldots, u_d]$ (i.e., they are simultaneously diagonalizable). In this basis, each matrix $\mathbf{H}_i$ is diagonal with eigenvalues denoted by $\lambda_k^{(i)}$. We classify these eigenvectors into two categories:

- **Shared Head ($k \le H$):** The first $H$ eigenvectors $\{u_1, \ldots, u_H\}$ represent universal components. We assume all domains possess non-zero variance along these directions ($\lambda_k^{(i)} > 0$ for all $i$), meaning these patterns are present in every domain.
- **Disjoint Tail ($k > H$):** The remaining eigenvectors represent domain-specific components. We assume each domain $i$ possesses a unique set of eigenvectors $\{u_k^{(i)}\}$ that are orthogonal to the specific components of other domains. Following Section A.1.2, we assume that the variance along these directions follows a domain-specific power law:

$$u_k^{(i)\top}\mathbf{H}_j u_k^{(i)} = \begin{cases} k^{-\alpha_i} & \text{if } j = i \\ 0 & \text{if } j \neq i \end{cases}.$$

  Consequently, for $k > H$, the spectral structures are completely decoupled: each domain $i$ has positive eigenvalues $\lambda_k^{(i)} = k^{-\alpha_i}$ along its own unique eigenvectors and zero along the eigenvectors of others.

Under this assumption, the mixture covariance matrix $\mathbf{H}(h) = \sum h_j\mathbf{H}_j$ exhibits a decoupled structure in the tail. Because the domain-specific eigenvectors do not overlap, the eigenvalue of a unique component in the mixture is simply its original variance scaled by the domain's proportion $h_j$. Specifically, for a tail eigenvector $u_k^{(i)}$ belonging to domain $i$, the corresponding eigenvalue in the mixture is:

$$\lambda(\mathbf{H}(h), u_k^{(i)}) = \sum_{j=1}^{K} h_j \cdot u_k^{(i)\top}\mathbf{H}_j u_k^{(i)} = h_i k^{-\alpha_i}.$$

**Connection with the Extended Quantization Model.** Here, each eigenvector $u_k^{(i)}$ can be viewed as a skill to be learned, and its corresponding eigenvalue $\lambda_k^{(i)} = h_i k^{-\alpha_i}$ can be viewed as the expected loss incurred if this skill is not learned (analogous to the $c(\text{skill}) \cdot p(\text{skill})$ term in our previous discussion). Therefore, our spectral assumption parallels the skill distribution assumption in the Extended Quantization Model. Furthermore, adjusting a domain's weight $h_i$ produces an

equivalent effect in both frameworks: reducing the weight $h_i$ linearly shrinks all eigenvalues of domain $i$ according to $\lambda_k^{(i)} = h_i k^{-\alpha_i}$; similarly, in the Extended Quantization Model, reducing $h_i$ linearly shrinks the expected loss of skills in domain $i$ by shrinking $p(\text{skill})$.

**Loss Analysis in the Multi-Domain Setting.** Under the spectral assumptions established above, we analyze the expected test loss $L(h, N, D)$ of a projected linear model $\theta \in \mathbb{R}^N$ trained via one-pass SGD on a dataset of size $D$, sampled from a mixture distribution with weights $h \in \Delta^{K-1}$. The expected loss is governed by two primary factors. First, the model has a finite size $N$ and is able to learn at most $N$ skills. As the model is trained to minimize loss, the training process implicitly solves the capacity allocation problem defined in the Extended Quantization Model (3), distributing the capacity $N$ across domains to the most valuable skills; consequently, the unlearned skills in each domain contribute a loss similar to $L_i(h) = c_i(x_i^*(h))^{-b_i}$ in the Extended Quantization Model. Second, for each skill within the model's capacity, the loss will depend on the number of times it has been observed; in other words, the stochastic nature of one-pass SGD introduces a noise term to the domain loss $L_i$, which is determined by the number of training samples drawn from that domain.

**Theorem 2** *Given a data size $D$ and a model size $N$, assume that the domains have mutually disjoint tails whose eigenvalues follow a power distribution with exponents $\alpha_i > 1$ and assume that the error contribution from the shared head is negligible. Given a mixture distribution with weights $h \in \Delta^{K-1}$, let $b_i = \alpha_i - 1$, $c_i = 1/(\alpha_i - 1)$, and let $x^*(h, N)$ be the optimal solution of the Extended Quantization Model (3) defined by model capacity $N$, mixture weights $h$, parameters $\{b_i\}$ and $\{c_i\}$. For a projected linear model $\theta \in \mathbb{R}^N$ trained via one-pass SGD on $D$ samples drawn from a mixture $h$, its expected test loss on domain $i$, denoted $L_i(h, N, D)$, satisfies*

$$L_i(h, N, D) \approx c_i x_i^*(h, N)^{-b_i} + A_i(Dh_i)^{-a_i} + E_i \tag{5}$$

*under some mild assumptions, where $a_i, A_i, E_i$ are constants that depend on $\alpha_i$.*

*Formally, for any $\epsilon > 0$, there exist $N_0$ and $D_0$ such that for all $N > N_0$, $D > D_0$, and $h$ satisfying Assumption 4, the expected test loss on domain $\tau$ is bounded by:*

$$L_\tau(h, N, D) \geq c_\tau(x^*(N, h)_\tau)^{-b_\tau}(1 - \epsilon) + \frac{A}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}}(1 - \epsilon) + \sigma^2,$$

$$L_\tau(h, N, D) \leq c_\tau(x^*(N, h)_\tau)^{-b_\tau}(1 + \epsilon) + \frac{A}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}}(1 + \epsilon) + \sigma^2,$$

*where* $A := \frac{\Gamma\left(1-\frac{1}{\alpha_\tau}\right)}{\alpha_\tau(2\eta_0)^{1-\frac{1}{\alpha_\tau}}} + \sigma^2 C_{\alpha_\tau} I_\gamma \eta_0^{1/\alpha_\tau}$.

**Domain Interaction and Optimal Data Mixture.** Analogous to the Extended Quantization Model, capacity competition is the key factor driving domain interaction. Specifically, the weights of other domains ($h_j$ for $j \neq i$) influence domain $i$'s loss $L_i$ solely through the capacity allocation term $x_i^*(h, N)^{-b_i}$, as the noise term $A_i(Dh_i)^{-a_i}$ depends solely on the domain's own weight $h_i$. However, when solving for the optimal training mixture $h^*$ given a target distribution with domain weights $w$, a key distinction emerges: the optimal mixture $h^*$ no longer equals $w$. To see this,

we formulate the optimal mixture calculation as a bi-level optimization problem:

$$\begin{aligned} h^* &= \arg\min_h \sum_{i=1}^K w_i L_i(h, N, D) \\ &= \sum_{i=1}^K \left[ w_i \left( c_i x_i^*(h, N)^{-b_i} + A_i (D h_i)^{-a_i} \right) + E_i \right] \end{aligned} \tag{6}$$

where $L_i(h, N, D)$ is determined by the optimal solution $x_i^*(h, N)$ of another optimization problem (3) with a different loss in the objective. The additional noise term $A_i(Dh_i)^{-a_i}$ in $L_i(h, N, D)$ introduces a misalignment between the objectives in (3) and (6), causing the optimal mixture $h^*$ to deviate from the target weights $w$.

This bi-level optimization problem can be solved with Online Mirror Descent (OMD) on the simplex. We first characterize the gradient of the outer objective $\mathcal{J}(h) := \sum_{i=1}^K w_i L_i(h, N, D)$.

**Proposition 3 (Gradient Characterization)** *Let $x^*(h)$ and $\lambda(h)$ be the optimal solution and the Lagrange multiplier of the inner capacity allocation problem. The gradient of the outer objective $\mathcal{J}(h)$ with respect to $h_k$ is: $\nabla_k \mathcal{J}(h) = -w_k a_k A_k D^{-a_k} h_k^{-a_k - 1} + \frac{\lambda x_k^*}{h_k(b_k+1)} \left( \bar{R} - \frac{w_k}{h_k} \right)$*

*where $\bar{R}$ is defined as: $\bar{R} = \dfrac{\sum_{j=1}^K \frac{x_j^*}{b_j+1} \left( \frac{w_j}{h_j} \right)}{\sum_{j=1}^K \frac{x_j^*}{b_j+1}}$.*

We then propose Algorithm 1 with Exponentiated Gradient updates to find the optimal mixture $h^*$.

**Effect of Overlap on Loss Components.** We now consider the scenario where domain tails overlap and analyze the impact of overlapping skills on $L_i$ as a function of $h$. We first examine the capacity allocation term, $c_i x_i^*(h, N)^{-b_i}$. Following the reasoning in Section B.2, overlapping skills contribute less to the fluctuation of this term compared to disjoint skills. Regarding the noise term $A_i(Dh_i)^{-a_i}$, the contribution from a skill depends on its aggregate observation count, $D \cdot p(\text{skill})$. Since the aggregate probability density $p(\text{skill})$ of an overlapping skill is more stable against mixture shifts, its contribution to the noise term will be more stable as well. Therefore, the overall change in loss induced by overlapping skills is likely to be small relative to that induced by disjoint skills.

## Appendix C. Sketched Linear Regression

Following [2, 9, 10], we consider a supervised learning setting where the goal is to learn a linear relationship from a stream of data. We analyze the dynamics of Stochastic Gradient Descent (SGD) under a *sketched observation* model, which characterizes the scenario where the model capacity (or feature access) is limited relative to the complexity of the data generating process. Our theoretical framework builds upon the work of [9], who demonstrated a scaling law with respect to the learning rate schedule (LRS). In this work, we extend their analysis to derive the scaling law with respect to the mixture ratio $h$. Crucially, we no longer assume that the entire second-order moment of $\mathbf{x}$ follows a single power law; instead, the moment is partitioned into several blocks, with each block exhibiting distinct power-law scaling. In this section, we will compute the validation loss on each domain given a data mixture $h \in \Delta^{K-1}$. We investigate the linear model proposed in Section B.3. Following the notation in Section B.3, we use $\lambda_k^{(i)}$, to represent the k-th eigenvalue in domain $k$.

**Algorithm 1:** Bi-Level Mixture Optimization via OMD

**Input:** Problem parameters, target $w$, constraints $N, D$, step size $\eta$
**Output:** Optimal mixture $h^*$
Initialize $h^{(0)} \leftarrow [1/K, \ldots, 1/K]$, $t \leftarrow 0$;
**while** *not converged* **do**
  `// 1. Inner Level: Capacity Response`
  Solve the inner optimization (Eq. 3) given $h^{(t)}$ to obtain the optimal allocation $x^*$ and multiplier $\lambda$;
  `// 2. Outer Level: Gradient Calculation`
  Compute the gradient vector $\nabla \mathcal{J}(h^{(t)})$ according to the closed-form expression in Proposition 3;
  `// 3. Optimization: Mirror Descent Step`
  Update weights:

$$\widetilde{h} \leftarrow h^{(t)} \odot \exp\Big(-\eta \nabla \mathcal{J}(h^{(t)})\Big)$$

  Normalize:

$$h^{(t+1)} \leftarrow \widetilde{h}/\|\widetilde{h}\|_1$$

  $t \leftarrow t+1$;
**end**
Return $h^* \leftarrow h^{(t)}$;

## C.1. Theoretical Framework

### C.1.1. DATA GENERATION

We first formally introduce the data generation process and assumptions therein.

Given a data mixture weight $h \in \Delta^{K-1}$, the input data $\mathbf{x} \in \mathbb{R}^d$ are drawn from a distribution $\mathcal{D} = \sum_{i=1}^{K} h_i \mathcal{P}_i$ with zero mean and covariance matrix $\mathbf{H}(h) := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] = \sum_{i=1}^{K} h_i \mathbf{A}_i$, where $\mathbf{A}_i := \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_i}[\mathbf{x}\mathbf{x}^\top]$. For technical reason, we introduce Assumption 4.

**Assumption 4** *There is a universal constant $\underline{h} > 0$ such that $h_i \geq \underline{h}$ holds for $i \in [K]$. We can set $\underline{h} = 10^{-2}$.*

This assumption ensure that every domain should contribute a non zero proportion to data mixture.

We assume $\mathbf{A}_i$ admits an eigen decomposition with eigenvalues $\lambda_1^{(i)} \geq \lambda_2^{(i)} \geq \cdots \geq \lambda_d^{(i)} \geq 0$. Without loss of generality, we work in the basis of the eigenvectors of $\mathbf{H}$, which means $\mathbf{A}_1, \mathbf{A}_2 \ldots \mathbf{A}_K, \mathbf{H}$ are diagonal. To facilitate the convergence analysis, we introduce Assumption 5.

**Assumption 5 (Hypercontractivity)** *For any domain $i \in [K]$ and any positive semi-definite (PSD) matrix $\mathbf{M}$, there exists a constant $C_0 > 0$ such that*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{P}_i}\Big[\mathbf{x}\mathbf{x}^\top \mathbf{M} \mathbf{x}\mathbf{x}^\top - \mathbf{A}_i \mathbf{M} \mathbf{A}_i\Big] \preceq C_0 tr(\mathbf{A}_i \mathbf{M}) \mathbf{A}_i.$$

Assumption 5 is a standard hypercontractivity condition which is also used in [9]. Intuitively, it requires the fourth moments of the data distribution to be reasonably bounded by its second moments, which ensures that the distribution $\mathcal{P}_i$ does not exhibit overly heavy tails. We note that this is

a very mild and standard assumption in theoretical analysis. A wide variety of standard distributions satisfy this condition for a small constant $C_0$, including but not limited to Gaussian distributions, sub-Gaussian distributions, and bounded distributions (such as uniform distributions on a sphere or a hypercube). Consequently, this assumption provides a robust foundation for concentration of measure without requiring strict distributional forms.

The target labels $y \in \mathbb{R}$ are generated by a true parameter $\theta^* \in \mathbb{R}^d$ via a linear model with additive noise:

$$y = \mathbf{x}^\top \theta^* + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{7}$$

where the noise $\epsilon$ is independent of $\mathbf{x}$. We also introduce the Assumption 6 on the prior of $\theta^*$.

**Assumption 6** *Assume that $\theta^*$ satisfies a prior such that $\mathbb{E}[\theta^*]^{\otimes 2} = \mathbf{I}$.*

**Assumption 7** *For all $i \in [K]$,$1 \leq \alpha_i \leq 3$.*

Assumption 7 ensure that the exponents do not vary too much, which is also empirically verified in [8].

### C.1.2. MODEL TRAINING

We then formalize the training process.

To model the constraint of finite model size or limited feature extraction, we adopt the setup from [9] and introduce a **sketching operator** $\mathbf{S} \in \mathbb{R}^{N \times d}$. In this work, we focus on the *spectral truncation* sketch, which projects the input data onto the subspace spanned by the top-$N$ eigenvectors of $\mathbf{H}$. Let $r_i$ denote the number of eigenvalues of $\mathbf{H}$ that are greater than $u_i^\top \mathbf{H} u_i$.

$$\mathbf{S}_{i,j} = \begin{cases} 1 & r_j = i \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

**Intuition of S** By eigende decomposition, $\mathbf{H} = \sum (u_i^\top \mathbf{H} u_i) u_i u_i^\top$, where $(u_i^\top \mathbf{H} u_i)$ can be seen as the *frequency* of skill $i$. As discussed in [5], model with capacity $N$ will learn the Top-$N$ frequent skills.

The learner does not observe the full input $\mathbf{x}$. Instead, at each step, the learner receives the sketched feature vector:

$$\widetilde{\mathbf{x}} := \mathbf{S}\mathbf{x}. \tag{9}$$

We train a parameter vector $\theta \in \mathbb{R}^N$ using SGD with cosine learning rate decay [13] on the sketched inputs. We initialize the model at $\theta_0 = \mathbf{0}$. At step $k$, the algorithm samples a data pair $(\mathbf{x}_k, y_k)$, constructs the observation $\widetilde{\mathbf{x}}_k = \mathbf{S}\mathbf{x}_k$, and performs a gradient descent update on the squared loss $\ell(\theta) = \frac{1}{2}(\widetilde{\mathbf{x}}_k^\top \theta - y_k)^2$:

$$\begin{aligned} \theta_{k+1} &= \theta_k - \eta_k \nabla_\theta \ell(\theta_k; \widetilde{\mathbf{x}}_k, y_k) \\ &= \theta_k - \eta_k \widetilde{\mathbf{x}}_k (\widetilde{\mathbf{x}}_k^\top \theta_k - y_k), \end{aligned} \tag{10}$$

where $\eta_k = \eta_0 \left(1 + \cos(\pi k / D)\right)$.

We assume that the training loss is always bounded.

We also impose an assumption on the parameter size $N$ and data size $D$:

**Assumption 8** *There is an absolute constant $\varepsilon > 0$ such that $N^{\alpha_i} \geq D^{1+\varepsilon}$ for all $i \in [K]$.*

Assumption 8 is a remarkably mild condition that is naturally satisfied in standard large-scale training regimes. Following empirical scaling laws [8], compute-optimal models typically scale the parameter size $N$ linearly with the dataset size $D$ (e.g., $D \approx 20N$). Under this linear scaling relationship ($N \propto D$), the left-hand side of our inequality scales as $\mathcal{O}(D^{\alpha_i})$. Since $\alpha_i > 1$ by definition, there strictly exists a sufficiently small constant $\varepsilon > 0$ such that $1 + \varepsilon \leq \alpha_i$. Consequently, the polynomial growth of $N^{\alpha_i}$ will trivially dominate $D^{1+\varepsilon}$ for large-scale $D$. Therefore, rather than imposing a restrictive capacity requirement, this assumption merely formalizes the standard overparameterized operational regime of modern language models, while simultaneously providing the necessary analytical bounds for our subsequent theoretical proofs.

### C.2. SDE Approximation

Following previous work [9, 16], we use continuous time approximation of one pass SGD, which simplify the analysis.

Since

$$\theta_{k+1} - \theta_k = -\eta_k \left( \mathbb{E}[\nabla_\theta \ell(\theta_k)] + \nabla_\theta \ell(\theta_k) - \mathbb{E}[\nabla_\theta \ell(\theta_k)] \right), \tag{11}$$

we have

$$\theta_D - \theta_0 = -\sum_{k=0}^{D-1} \eta_k \mathbb{E}[\nabla_\theta \ell(\theta_k)] - \sum_{k=0}^{D-1} \eta_k \left( \nabla_\theta \ell(\theta_k) - \mathbb{E}[\nabla_\theta \ell(\theta_k)] \right). \tag{12}$$

We generalize the discrete sequence $\{\theta_0, \ldots, \theta_D\}$ to a continuous function $\theta(\cdot)$, and similarly extend $\eta_k$ to $\eta(\cdot)$.

Now we compute $\mathbf{\Sigma}(\theta_k) := (\nabla_\theta \ell(\theta_k) - \mathbb{E}[\nabla_\theta \ell(\theta_k)])^{\otimes 2}$.

Since

$$\nabla_\theta \ell(\theta_k) = \mathbf{S}\mathbf{x}_k\mathbf{x}_k^\top \left( \mathbf{S}^\top \theta_k - \theta^* \right) - \epsilon_k \mathbf{S}\mathbf{x}_k, \tag{13}$$

we have

$$\mathbf{\Sigma}(\theta_k) = \mathbb{E}\left[ (\nabla_\theta \ell(\theta_k) - \mathbb{E}[\nabla_\theta \ell(\theta_k)]) \right]^{\otimes 2} = \left[ \mathbf{S} \left( \mathbf{x}_k\mathbf{x}_k^\top - \mathbf{H} \right) \left( \mathbf{S}^\top \theta - \theta^* \right) \right]^{\otimes 2} + \sigma^2 \mathbf{S}\mathbf{H}\mathbf{S}^\top. \tag{14}$$

Using Euler-Maclaurin equation, we have

$$\theta_D - \theta_0 \approx \theta(D) - \theta(0) = \int_0^D -\eta(k)\mathbf{S}\mathbf{H} \left( \mathbf{S}^\top \theta(k) - \theta^* \right) \mathrm{d}k + \int_0^D \eta(k) \sqrt{\mathbf{\Sigma}(\theta(k))} \mathrm{d}\mathbf{B}_k, \tag{15}$$

where $\mathbf{B}_k \in \mathbb{R}^N$ is a N-dimensional Brownian motion. Following [9], we define **intrinsic time**

$$\tau(t) := \int_0^t \eta(k) \mathrm{d}k.$$

We can rewrite equation 15 as

$$\theta(D) - \theta(0) = \int_0^D -\mathbf{S}\mathbf{H}(\mathbf{S}^\top \theta(k) - \theta^*) \mathrm{d}\tau(k) + \sqrt{\eta(k)\mathbf{\Sigma}(\theta(k))} \mathrm{d}\mathbf{B}_\tau. \tag{16}$$

Taking the derivative of Equation 16, we have the following lemma.

**Lemma 9**

*Define* $\mathbf{w}(t) := \mathbf{S}^\top \theta(\tau^{-1}(t)) - \theta^*, \gamma(t) := \eta(\tau^{-1}(t))$, *we have*

$$\mathrm{d}\mathbf{w} = -\mathbf{S}^\top \mathbf{S}\mathbf{H}\mathbf{w}\mathrm{d}t + \mathbf{S}^\top \sqrt{\gamma(t)\mathbf{\Sigma}(\theta(\tau^{-1}(t))}\mathbf{S}\mathrm{d}\mathbf{B}_t, \tag{17}$$

*where* $\mathbf{B}_t \in \mathbb{R}^d$ *is a d-dimensional Brownian motion.*

**Assumption 10**

*The train loss over the training process has a finite upper bound L. Formally, for all t,* $\mathbb{E}\left[(\langle \mathbf{S}\mathbf{x}, \theta(t)\rangle - y)^2\right] \leq L$.

### C.3. Test Loss

In the following discussion, we will compute the test loss on a certain domain $\tau \in [K]$ given mixture $h \in \Delta^{K-1}$.

Direct computation yield:

$$\begin{aligned} L_\tau &= \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_\tau}\left[(\langle \mathbf{S}\mathbf{x}, \theta\rangle - \langle \mathbf{x}, \theta^*\rangle)^2\right] + \sigma^2 && (18)\\ &= \mathbf{w}^\top(t)\mathbf{A}_\tau \mathbf{w}(t) + \sigma^2. && (19)\end{aligned}$$

**Theorem 11**

*Consider the projected linear regression model defined in Section C.1, let* $\mathbf{A}_\tau = diag(a_1, a_2 \dots a_d)$ *be domain* $\tau$*'s data covariance, then the expected test loss on domain* $\tau$*, denoted by* $L_\tau$*, satisfies*

$$Bias + Approx + Var_1 \leq L_\tau - \sigma^2 \leq Bias + Approx + Var_2, \tag{20}$$

*where*

$$Bias := \sum_{r_k \leq N} \exp(-2\lambda_k D\eta_0) a_k. \tag{21}$$

$$Approx := \sum_{r_i > N} a_k. \tag{22}$$

$$Var_1 := \sum_{r_k \leq N} \sigma^2 a_k \lambda_k \int_0^{D\eta_0} \exp(-2\lambda_k(D\eta_0 - t))\gamma(t)\mathrm{d}t. \tag{23}$$

$$Var_2 := \sum_{r_k \leq N} a_k \lambda_k \int_0^{D\eta_0} \exp(-2\lambda_k(D\eta_0 - t))\gamma(t)\left(\sigma^2 + \frac{C_0}{\underline{h}}\mathbf{w}(t)^\top \mathbf{H}\mathbf{w}(t)\right)\mathrm{d}t. \tag{24}$$

**Proof** By Itô's Lemma, for any diagonal matrix $\mathbf{A}$ (not necessarily $\mathbf{A}_i$), we have

$$\begin{aligned} \mathrm{d}\mathbf{w}^\top(t)\mathbf{A}\mathbf{w}(t) &= \mathbb{E}\left[2\mathbf{w}^\top(t)\mathbf{A}\mathrm{d}\mathbf{w}(t) + \frac{1}{2}\mathrm{tr}\left[2\mathbf{A}\cdot(\mathrm{d}\mathbf{w}(t))^{\otimes 2}\right]\right] && (25)\\ &= \mathbb{E}\left[-2\mathbf{w}^\top(t)\mathbf{A}\mathbf{S}^\top\mathbf{S}\mathbf{H}\mathbf{w}\mathrm{d}t + \gamma(t)\mathbf{w}^\top(\mathbf{x}\mathbf{x}^\top - \mathbf{H})\mathbf{S}^\top\mathbf{A}\mathbf{S}(\mathbf{x}\mathbf{x}^\top - \mathbf{H})\mathbf{w}\mathrm{d}t + \gamma(t)\sigma^2\mathrm{tr}[\mathbf{A}\mathbf{S}\mathbf{H}\mathbf{S}^\top]\mathrm{d}t\right]. && (26)\end{aligned}$$

We first take expectation over $\mathbf{x}$ with the help of Assumption 5 and Assumption 4, for term $\mathbf{w}^\top(\mathbf{x}\mathbf{x}^\top - \mathbf{H})\mathbf{S}^\top\mathbf{A}\mathbf{S}(\mathbf{x}\mathbf{x}^\top - \mathbf{H})\mathbf{w}$, we have

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}(h)}\left[\mathbf{w}^\top(\mathbf{x}\mathbf{x}^\top - \mathbf{H})\mathbf{S}^\top\mathbf{A}\mathbf{S}(\mathbf{x}\mathbf{x}^\top - \mathbf{H})\mathbf{w}\right] = \mathbf{w}^\top\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\mathbf{S}^\top\mathbf{A}\mathbf{S}\mathbf{x}\mathbf{x}^\top - \mathbf{H}\mathbf{S}^\top\mathbf{A}\mathbf{S}\mathbf{H}\right]\mathbf{w} \tag{27}$$

$$= \mathbf{w}^\top\sum h_i\mathbb{E}_{\mathbf{x}\sim\mathcal{P}_i}\left[\mathbf{x}\mathbf{x}^\top\mathbf{S}^\top\mathbf{A}\mathbf{S}\mathbf{x}\mathbf{x}^\top - \mathbf{A}_i\mathbf{S}^\top\mathbf{A}\mathbf{S}\mathbf{A}_i\right]\mathbf{w} \tag{28}$$

$$\leq \mathbf{w}^\top\sum h_i C_0\mathrm{tr}[\mathbf{S}^\top\mathbf{A}\mathbf{S}\mathbf{A}_i]\mathbf{A}_i\mathbf{w} \tag{29}$$

$$\leq \mathbf{w}^\top\frac{C_0}{\underline{h}}\sum\mathrm{tr}[\mathbf{S}^\top\mathbf{A}\mathbf{S}\mathbf{A}_i]h_i\mathbf{A}_i\mathbf{w} \tag{30}$$

$$= \frac{C_0}{\underline{h}}\mathrm{tr}[\mathbf{S}^\top\mathbf{A}\mathbf{S}\mathbf{H}]\mathbf{w}(t)^\top\mathbf{H}\mathbf{w}(t). \tag{31}$$

Set $\mathbf{A}$ as $\mathbf{E}_{k,k}$, let $y_k(t) = \mathbb{E}[\mathbf{w}_k^2(t)]$, we can get the two side bound with respect to $y_k$:

- For $r_k \leq N$,

$$-2\lambda_k y_k(t) + \gamma(t)\cdot\sigma^2\cdot\lambda_k \leq y_k'(t) \leq -2\lambda_k y_k(t) + \gamma(t)\lambda_k\left(\sigma^2 + \frac{C_0}{\underline{h}}\mathbf{w}(t)^\top\mathbf{H}\mathbf{w}(t)\right). \tag{32}$$

- For $r_k > N$, $y_k'(t) = 0$.

Solving the differential equation, for all $k$ such that $r_k \leq N$, we have

$$y_k(D\eta_0) \geq \exp(-2\lambda_k D\eta_0)y_k(0) + \sigma^2\lambda_k\int_0^{D\eta_0}\exp(-2\lambda_k(D\eta_0 - t))\gamma(t)\mathrm{d}t, \tag{33}$$

$$y_k(D\eta_0) \leq \exp(-2\lambda_k D\eta_0)y_k(0) + \lambda_k\int_0^{D\eta_0}\exp(-2\lambda_k(D\eta_0 - t))\gamma(t)\left(\sigma^2 + \frac{C_0}{\underline{h}}\mathbf{w}(t)^\top\mathbf{H}\mathbf{w}(t)\right)\mathrm{d}t. \tag{34}$$

By $\mathbb{E}[\mathbf{w}^\top(t)\mathbf{A}\mathbf{w}(t)] = \sum y_k(t)a_k$ and $\mathbb{E}[\mathbf{w}_i^2(0)] = 1$(by Assumption 6) we complete the proof. ■

In the following sections, we estimate each term in the bounds separately.

### C.4. The Approximation Error Term

Consider an arbitrary domain $\tau$, we first provide a bound for the approximation error term $\mathrm{Approx} := \sum_{r_k > N} a_k$. We show that this term is actually equivalent to the domain loss $L_i(h) = c_i x_i^*(h)^{-b_i}$ in our Extended Quantization Model, with a gap due to discretization.

**Theorem 12** *Consider the projected linear regression model defined in Section C.1. Let $x^*$ be the solution to the Problem 3 with parameter $c_i = \frac{1}{\alpha_i - 1}, b_i = \alpha_i - 1$. Then for any domain $\tau$, the term Approx in Theorem 11 satisfies*

$$\left|\mathit{Approx} - c_\tau(x_\tau^*)^{-b_\tau}\right| \leq \frac{\alpha_\tau}{N^{\alpha_\tau+1}}$$

*for sufficiently large $N$.*

**Remark:** By Lemma 13, as $x_\tau^* = \Omega\left(N^{\frac{\alpha_{\min}}{\alpha_\tau}}\right)$, $(x_\tau^*)^{1-\alpha_\tau} = \mathcal{O}\left(\frac{1}{N^{\frac{\alpha_\tau-1}{\alpha_\tau}\alpha_{\min}}}\right)$. Theorem 12 indicates that Approx $= \frac{(x_\tau^*)^{1-\alpha_\tau}}{\alpha_\tau-1}(1+o_N(1))$, which means the gap due to discretization can be ignored.

### C.4.1. PROOF OF THEOREM 12

**Lemma 13** *Consider the Extended Quantization Model in Section B.2 (Problem 3):*

$$\begin{aligned} \min_{\mathbf{x}} \quad & L(\mathbf{x}) = \sum_{i=1}^{K} h_i c_i x_i^{-b_i} \\ \text{s.t.} \quad & \sum_{i=1}^{K}(x_i - H) \le N - H, \\ & x_i \ge H, \quad \forall i \in \{1,\dots,K\}, \end{aligned} \tag{35}$$

*where $h_i, c_i, b_i > 0$ for all $i$. Let $b_{\min} = \min_{1\le i\le K} b_i$. As $N \to \infty$, the optimal solution $x^*$ satisfies the asymptotic lower bound:*

$$x_\tau^* = \Omega\left(N^{\frac{b_{\min}+1}{b_\tau+1}}\right).$$

**Proof** Let $A_i = h_i c_i > 0$ for all $i \in \{1,\dots,K\}$. Since $A_i > 0$ and $b_i > 0$, the objective function $L(\mathbf{x})$ is strictly monotonically decreasing with respect to each variable $x_i$. Consequently, to minimize the objective function, the variables $x_i$ must take the largest possible values permitted by the feasible region. This implies that the sum constraint must be active at the optimal solution, yielding the equality $\sum_{i=1}^{K} x_i = N + (K-1)H$. Furthermore, as the total available resource $N$ approaches infinity, the optimal values $x_i^*$ will also approach infinity. Thus, for sufficiently large $N$, the lower bound constraints $x_i \ge H$ become strictly inactive and can be omitted from the asymptotic analysis.

We proceed by applying the method of Lagrange multipliers. The Lagrangian associated with the equality constraint is given by

$$\mathcal{L}(\mathbf{x},\lambda) = \sum_{i=1}^{K} A_i x_i^{-b_i} + \lambda\left(\sum_{i=1}^{K} x_i - N - (K-1)H\right),$$

where $\lambda > 0$ is the Lagrange multiplier. Taking the partial derivative of $\mathcal{L}$ with respect to $x_i$ and equating it to zero yields the first-order necessary conditions for optimality:

$$\frac{\partial \mathcal{L}}{\partial x_i} = -b_i A_i x_i^{-(b_i+1)} + \lambda = 0, \quad \forall i \in \{1,\dots,K\}.$$

Rearranging this expression, we obtain a relationship between the optimal variable $x_i$ and the multiplier $\lambda$:

$$\lambda = b_i A_i x_i^{-(b_i+1)}.$$

Since $\lambda$ is a global constant across all dimensions, we can equate the expressions for an arbitrary index $i$ and the specific index $\tau$, yielding

$$b_i A_i x_i^{-(b_i+1)} = b_\tau A_\tau x_\tau^{-(b_\tau+1)}.$$

Solving this equation for $x_i$ in terms of $x_\tau$, we find

$$x_i = \left(\frac{b_i A_i}{b_\tau A_\tau}\right)^{\frac{1}{b_i+1}} x_1^{\frac{b_\tau+1}{b_i+1}}.$$

Substituting this relationship back into the active resource constraint gives

$$\sum_{i=1}^{K} \left(\frac{b_i A_i}{b_\tau A_\tau}\right)^{\frac{1}{b_i+1}} x_\tau^{\frac{b_\tau+1}{b_i+1}} = N + (K-1)H.$$

We now analyze the asymptotic behavior of this equation as $N \to \infty$. On the right-hand side, the constant term $(K-1)H$ becomes negligible, so the right-hand side is asymptotically equivalent to $N$. On the left-hand side, we have a sum of fractional powers of $x_1$. As $N \to \infty$ implies $x_1 \to \infty$, the behavior of the sum is completely dominated by the term with the highest exponent. The exponent of $x_1$ for the $i$-th term is $\frac{b_\tau+1}{b_i+1}$. This exponent is maximized when its denominator, $b_i + 1$, is minimized, which occurs exactly when $b_i = b_{\min} = \min_{1 \le j \le K} b_j$.

Let $\mathcal{I}_{\min} = \{i \mid b_i = b_{\min}\}$ be the index set of all terms achieving this minimum exponent. Extracting these dominant terms, we establish the asymptotic equivalence

$$\sum_{i \in \mathcal{I}_{\min}} \left(\frac{b_{\min} A_i}{b_\tau A_\tau}\right)^{\frac{1}{b_{\min}+1}} x_1^{\frac{b_\tau+1}{b_{\min}+1}} \sim N.$$

Letting $C = \sum_{i \in \mathcal{I}_{\min}} \left(\frac{b_{\min} A_i}{b_\tau A_\tau}\right)^{\frac{1}{b_{\min}+1}}$, which is a strictly positive constant, the relation simplifies to

$$C x_\tau^{\frac{b_\tau+1}{b_{\min}+1}} \sim N.$$

Solving this asymptotic equivalence for $x_\tau$ yields

$$x_\tau \sim \left(\frac{1}{C}\right)^{\frac{b_{\min}+1}{b_\tau+1}} N^{\frac{b_{\min}+1}{b_\tau+1}}.$$

This demonstrates that the growth rate of $x_\tau^*$ is proportional to $N^{\frac{b_{\min}+1}{b_\tau+1}}$. Therefore, we conclude that the optimal solution $x_\tau^*$ satisfies the strict asymptotic lower bound $x_\tau^* = \Omega\left(N^{\frac{b_{\min}+1}{b_\tau+1}}\right)$, completing the proof. ■

Recall that Problem 3 can be written as

$$\begin{aligned} \min_x \quad & L = \sum_{i=1}^{K} h_i \frac{1}{\alpha_i - 1} x_i^{1-\alpha_i} \\ \text{s.t.} \quad & \sum_{i=1}^{K} (x_i - H) \le N - H, \\ & x_i \ge H, \quad \forall i. \end{aligned}$$

By KKT condition, the optimal solution $x^*$ satisfies:

$$\sum_{i\in[K]} x_i^* = N + (K-1)H, \tag{36}$$

$$\frac{h_i}{x_i^{*\alpha_i}} + \mu_i = \lambda, \tag{37}$$

$$\mu_i(x_i^* - H) = 0. \tag{38}$$

Define

$$S(\lambda) := \sum_{i\in[K]} \left(\frac{h_i}{\lambda}\right)^{\frac{1}{\alpha_i}}. \tag{39}$$

For sufficiently large $N$ such that

$$N > \sum_{i\in[K]} \left(\frac{h_i}{\underline{h}}\right)^{\frac{1}{\alpha_i}} H^{\frac{\alpha_{\max}}{\alpha_i}} - (K-1)H, \tag{40}$$

we have

$$S\left(\frac{\underline{h}}{H^{\alpha_{\max}}}\right) < N + (K-1)H. \tag{41}$$

Therefore, $S\left(\frac{h_i}{H^{\alpha_i}}\right) < N + (K-1)H$. Let $\lambda_0$ be the solution to the equation

$$S(\lambda) = N + (K-1)H. \tag{42}$$

Therefore, $\lambda_0 < \frac{h_i}{H^{\alpha_i}}$ for any $i \in [K]$. We claim that there is an optimal solution satisfying $\mu_i^* = 0$. We choose $\lambda^* = \lambda_0$. Since $\lambda^* < \frac{h_i}{H^{\alpha_i}}$, we have $x_i^* > H$, which satisfies the KKT condition. Define $x_{i,0} := \lfloor x_i^* \rfloor, x_{i,1} := \lceil x_i^* \rceil$. Since $\sum x_i^* = N - (K-1)H$, we have

$$\sum x_{i,0} \le N - (K-1)H \le \sum x_{i,1}. \tag{43}$$

Therefore,

$$\sum_{k>x_{\tau,0}} k^{-\alpha_\tau} < \text{Approx} < \sum_{k>x_{\tau,1}} k^{-\alpha_\tau}, \tag{44}$$

which means

$$\left|\text{Approx} - \frac{(x_\tau^*)^{1-\alpha_\tau}}{\alpha_\tau - 1}\right| \le \frac{\alpha_\tau}{N^{\alpha_\tau+1}}.$$

### C.5. The Bias Term

In this section, we bound the bias term in the expected test loss for an arbitrary domain $\tau$.

**Theorem 14** *Consider the projected linear regression model defined in Section C.1. For any $N, D$, and any $h \in$, any domain $\tau \in [K]$, when the model is trained with mixture $h$, the bias term in the expected loss for domain $\tau$ in Theorem 11 has*

$$\textit{Bias} = \frac{\Gamma\left(1-\frac{1}{\alpha_\tau}\right)}{\alpha_\tau (2\eta_0)^{1-\frac{1}{\alpha_\tau}}} \frac{1}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}} + \mathcal{E},$$

*where the error $\mathcal{E}$ is strictly bounded by*

$$|\mathcal{E}| \leq \frac{C_{10,1}}{D} + \frac{C_{10,2}}{N^{\alpha_{\min}\left(1-\frac{1}{\alpha_\tau}\right)}},$$

*where $C_{10,1}$ and $C_{10,2}$ are constants that only depend on $\alpha$ and $\underline{h}$, but not $h, N, D$.*

**Proof** By definition, we have

$$\text{Bias} = \sum_{k=1}^{H} k^{-\alpha_1} \exp\left(-2D\eta_0 \sum_{j\in[K]} h_j \lambda_k^{(j)}\right) + \sum_{k=H+1}^{x_\tau^*} \exp(-2h_1\lambda_k^{(1)} D\eta_0)\lambda_k^{(1)} \tag{45}$$

$$\leq H^{1-\alpha_1} \exp(-2D\eta_0 H^{-\alpha_{\max}}) + \sum_{k=H+1}^{x_\tau^*} \exp(-2h_1\lambda_k^{(1)} D\eta_0)\lambda_k^{(1)}. \tag{46}$$

We simplify the notation by defining

$$A := H^{1-\alpha_\tau} \exp(-2D\eta_0 H^{-\alpha_{\max}}), \tag{47}$$

$$B := \sum_{k=H+1}^{x_\tau^*} \exp(-2h_\tau k^{-\alpha_\tau} D\eta_0) k^{-\alpha_\tau}, \tag{48}$$

and then by definition,

$$B \leq \text{Bias} \leq A + B. \tag{49}$$

We now estimate $B$ by converting the discrete sum into a continuous integral. Define a continuous version of $B$ as $f(x) := \exp(-2h_\tau x^{-\alpha_\tau} D\eta_0) x^{-\alpha_\tau}$. Since $f(x)$ first increases and then decreases, achieving its absolute maximum at $k_0 = (2h_\tau D\eta_0)^{\frac{1}{\alpha_\tau}}$, the approximation error between the discrete summation and the continuous integral is bounded by the total variation of $f(x)$ across the interval:

$$\left| B - \int_H^{x_\tau^*} f(x)\mathrm{d}x \right| \leq \int_H^{x_\tau^*} |f'(x)|\mathrm{d}x \leq 2f(k_0) = \frac{2}{e(2h_\tau D\eta_0)} = \frac{1}{eh_\tau D\eta_0}. \tag{50}$$

Next, we evaluate the continuous integral. Applying the change of variable $p = \frac{2h_\tau D\eta_0}{x^{\alpha_\tau}}$, we have $\mathrm{d}p = -\alpha_\tau \frac{2h_\tau D\eta_0}{x^{\alpha_\tau+1}}\mathrm{d}x$, which yields:

$$\int_H^{x_\tau^*} f(x)\mathrm{d}x = \frac{1}{\alpha_\tau (2h_\tau D\eta_0)^{1-\frac{1}{\alpha_\tau}}} \int_{p_{\min}}^{p_{\max}} p^{-\frac{1}{\alpha_\tau}} \exp(-p)\mathrm{d}p, \tag{51}$$

where the integration limits are $p_{\min} = \frac{2h_\tau D\eta_0}{(x_\tau^*)^{\alpha_\tau}}$ and $p_{\max} = \frac{2h_\tau D\eta_0}{H^{\alpha_\tau}}$. The integral can be decomposed into the complete Gamma function minus the two truncation tails:

$$\int_{p_{\min}}^{p_{\max}} p^{-\frac{1}{\alpha_\tau}} \exp(-p)\mathrm{d}p = \Gamma\left(1-\frac{1}{\alpha_\tau}\right) - \int_0^{p_{\min}} p^{-\frac{1}{\alpha_\tau}} \exp(-p)\mathrm{d}p - \int_{p_{\max}}^{\infty} p^{-\frac{1}{\alpha_\tau}} \exp(-p)\mathrm{d}p.$$

We bound the two truncation tails as follows. For the lower tail, since $\exp(-p) \le 1$, we have:

$$\int_0^{p_{\min}} p^{-\frac{1}{\alpha_\tau}} \exp(-p)\mathrm{d}p \le \int_0^{p_{\min}} p^{-\frac{1}{\alpha_\tau}}\mathrm{d}p = \frac{\alpha_\tau}{\alpha_\tau - 1} p_{\min}^{1-\frac{1}{\alpha_\tau}} = \frac{\alpha_\tau}{\alpha_\tau - 1}\left(\frac{2h_\tau D\eta_0}{(x_\tau^*)^{\alpha_\tau}}\right)^{1-\frac{1}{\alpha_\tau}}. \quad (52)$$

For the upper tail, since $p_{\max} > 1$, $p^{-\frac{1}{\alpha_\tau}} \le p_{\max}^{-\frac{1}{\alpha_\tau}}$ for all $p \ge p_{\max}$, leading to:

$$\int_{p_{\max}}^{\infty} p^{-\frac{1}{\alpha_\tau}} \exp(-p)\mathrm{d}p \le p_{\max}^{-\frac{1}{\alpha_\tau}} \int_{p_{\max}}^{\infty} \exp(-p)\mathrm{d}p = p_{\max}^{-\frac{1}{\alpha_\tau}} \exp(-p_{\max}) = \frac{H}{(2h_\tau D\eta_0)^{\frac{1}{\alpha_\tau}}} \exp\left(-\frac{2h_\tau D\eta_0}{H^{\alpha_\tau}}\right). \quad (53)$$

Multiplying these two tail bounds by the integral's prefactor $\frac{1}{\alpha_\tau (2h_\tau D\eta_0)^{1-\frac{1}{\alpha_\tau}}}$, the overall error from the continuous integral truncations are strictly bounded by $\frac{1}{\alpha_\tau - 1}\frac{1}{(x_\tau^*)^{\alpha_\tau - 1}}$ and $\frac{H}{2\alpha_\tau h_\tau D\eta_0}\exp\left(-\frac{2h_\tau D\eta_0}{H^{\alpha_\tau}}\right)$, respectively.

By setting the principal order as the main term:

$$M = \frac{\Gamma\left(1-\frac{1}{\alpha_\tau}\right)}{\alpha_\tau (2\eta_0)^{1-\frac{1}{\alpha_\tau}}} \frac{1}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}}, \quad (54)$$

and combining all sources of discrepancies $|\text{Bias} - M| \le |\text{Bias} - B| + |B - \int f| + (\text{Integral Truncation Errors})$, we obtain the final explicit bound for the total error $\mathcal{E} = \text{Bias} - M$:

$$\begin{aligned} |\mathcal{E}| &\le A + \frac{1}{eh_\tau D\eta_0} + \frac{1}{\alpha_\tau - 1}\frac{1}{(x_\tau^*)^{\alpha_\tau - 1}} + \frac{H}{2\alpha_\tau h_\tau D\eta_0}\exp\left(-\frac{2h_\tau D\eta_0}{H^{\alpha_\tau}}\right) && (55)\\ &\le \frac{H^{1+\alpha_{\max}-\alpha_\tau}}{2eD\eta_0} + \frac{C}{2(\alpha_\tau - 1)N^{\alpha_{\min}\left(1-\frac{1}{\alpha_\tau}\right)}} + \frac{H}{2\alpha_\tau h_\tau D\eta_0} && (56)\\ &\le \frac{C_{10,1}}{D} + \frac{C_{10,2}}{N^{\alpha_{\min}\left(1-\frac{1}{\alpha_\tau}\right)}}, && (57) \end{aligned}$$

where $C_{10,1}$ and $C_{10,2}$ only depends on $\alpha$. The second $\le$ is by $\exp(x) \ge ex$ for all $x \in \mathbb{R}$ and Lemma 13. ■

### C.6. The Variance Term

In this section, we analyze the variance terms $\text{Var}_1, \text{Var}_2$ for an arbitrary domain $\tau$. We provide a bound as follows.

**Theorem 15** *Consider the projected linear regression model defined in Section C.1. For any $N, D$, any mixture $h \in$, and for an arbitrary domain $\tau$, the terms $Var_1$, $Var_2$ in Theorem 11 satisfy*

$$Var_1 = \sigma^2 C_{\alpha_\tau} I_\gamma \eta_0^{1/\alpha_\tau} (Dh_\tau)^{1/\alpha_\tau - 1} + \mathcal{E}_1 \tag{58}$$

$$Var_1 \leq Var_2 \leq Var_1 + (C_{12} + C_{14,1})(Dh_\tau)^{-2/3} + C_{14,2} D N^{1-2\alpha_\tau} + C_{15}(Dh)^{1/\alpha_\tau - 1} D^{-1}. \tag{59}$$

*where*

$$|\mathcal{E}_1| \leq C_{12}(Dh_\tau)^{-2/3} + C_{13,2} D N^{1-2\alpha_\tau}, \tag{60}$$

*and $C_{\alpha_\tau}, I_\gamma, C_{12}, C_{13,2}, C_{14,1}, C_{14,2}, C_{15}$ are constants that depend only on $\eta_0, \alpha, H$ but not on mixture $h$.*

### C.6.1. PROOF OF THEOREM 15

Recall that

$$\text{Var}_1 := \sum_{r_k \leq N} \sigma^2 a_k \lambda_k \int_0^{D\eta_0} \exp(-2\lambda_k (D\eta_0 - t)) \gamma(t) \mathrm{d}t.$$

$$\text{Var}_2 := \sum_{r_k \leq N} a_k \lambda_k \int_0^{D\eta_0} \exp(-2\lambda_k (D\eta_0 - t)) \gamma(t) \left( \sigma^2 + \frac{C_0}{\underline{h}} \mathbf{w}(t)^\top \mathbf{H} \mathbf{w}(t) \right) \mathrm{d}t.$$

We start with analyzing $\gamma(t)$.

**Lemma 16** *Let $G : \mathbb{R} \to \mathbb{R}$ be the inverse of the map $y \mapsto y + \sin y$. We have*

$$\gamma(t) = \eta_0 \left( 1 + \cos G \left( \frac{t\pi}{D\eta_0} \right) \right).$$

**Proof** Since $\eta(x) = \eta_0 \left(1 + \cos\left(\frac{\pi k}{D}\right)\right)$, we have

$$\tau(x) = \int_0^x \eta(a) \mathrm{d}a \tag{61}$$

$$= \eta_0 \frac{D}{\pi} \left( \frac{\pi}{D} x + \sin\left( \frac{\pi}{D} x \right) \right). \tag{62}$$

When $\tau(x) = t$, we have

$$x = \frac{D}{\pi} G \left( \frac{t\pi}{D\eta_0} \right). \tag{63}$$

Therefore

$$\gamma(t) = \eta_0 \left( 1 + \cos G \left( \frac{t\pi}{D\eta_0} \right) \right). \tag{64}$$

■

Note that the difference between $\text{Var}_1$ and $\text{Var}_2$ is only $\mathbf{w}^\top(t)\mathbf{H}\mathbf{w}(t)$ term. Note that $\mathbf{w}^\top(t)\mathbf{H}\mathbf{w}(t) + \sigma^2$ is exactly the train loss. In the following discussion, we first show that train loss satisfies the following theorem.

**Lemma 17**

$$\mathbf{w}^\top(t)\mathbf{H}\mathbf{w}(t) \leq \frac{C'}{\max\left(t^{1-\frac{1}{\alpha_{\max}}}, LC'\right)} + \frac{C''}{N^{\alpha_{\min}\left(1-\frac{1}{\alpha_{\max}}\right)}},$$

*where*

$$L = \max_t \mathbf{w}^\top(t)\mathbf{H}\mathbf{w}(t) \text{ (Assumption 10 ensures that } L \text{ is a finite constant)} \tag{65}$$

**Proof** We can upper bound the train loss $\mathbf{w}^\top(t)\mathbf{H}\mathbf{w}(t)$ as:

$$\mathbf{w}^\top(t)\mathbf{H}\mathbf{w}(t) \leq \sum_{i=1}^{K}\sum_{k\geq 1} h_i k^{-\alpha_i}\exp(-2hk^{-\alpha_i}t) \tag{66}$$

$$+\sum_{i=1}^{K}\sum_{k\geq 1} h_i k^{-\alpha_i}\int_0^t \exp(-2k^{-\alpha_i}(D\eta_0 - s))\gamma(s)\sigma'^2 \mathrm{d}s \tag{67}$$

$$+\sum_{i=1}^{K}\frac{x_i^{*1-\alpha_i}}{\alpha_i - 1}, \tag{68}$$

where $\sigma'^2 = \sigma^2 + \frac{C_0}{h}L$. Following similar steps in Theorem 14, we have

$$\sum_{k\geq 1} h_i k^{-\alpha_i}\exp(-2hk^{-\alpha_i}t) \leq \frac{\Gamma\left(1-\frac{1}{\alpha_i}\right)}{\alpha_i 2^{1-\frac{1}{\alpha_i}}}\frac{1}{(h_i t)^{1-\frac{1}{\alpha_i}}} + \frac{C_{10,1}}{t} + \frac{C_{10,2}}{N^{\alpha_{\min}(1-1/\alpha_\tau)}}. \tag{69}$$

Let $\gamma_1(s) := 1 + \cos G\left(\frac{\pi s}{t}\right)$. Since

$$\sum_{k\geq 1} h_i k^{-\alpha_i}\int_0^t \exp(-2k^{-\alpha_i}(D\eta_0 - s))\gamma(s)\sigma'^2\mathrm{d}s \leq \sum_{k\geq 1} h_i k^{-\alpha_i}\int_0^t \exp(-2k^{-\alpha_i}(t-s))\gamma_1(s)\sigma'^2\mathrm{d}s, \tag{70}$$

using Lemma 18, substituting $D\eta_0$ as $t$, we have

$$\sum_{k\geq 1} h_i k^{-\alpha_i}\int_0^t \exp(-2k^{-\alpha_i}(D\eta_0 - s))\gamma(s)\sigma'^2\mathrm{d}s \leq \sigma'^2 C_{\alpha_i} I_{\gamma'}(h_i t)^{1/\alpha_i - 1} + \frac{C_{13,1}}{(h_i t)^{2/3}} + \frac{C_{13,2}t}{N^{\alpha_{\min}(2-1/\alpha_\tau)}}. \tag{71}$$

For the last part, by Lemma 13, we have

$$\sum_{i=1}^{K}\frac{x_i^{*1-\alpha_i}}{\alpha_i - 1} \leq \frac{K}{\alpha_{\min}}\frac{1}{N^{\alpha_{\min}\left(1-\frac{1}{\alpha_{\max}}\right)}}. \tag{72}$$

Adding these three terms up, we complete our proof. ■

Lemma 18 is used in the proof of Lemma 17 and closely related to $\text{Var}_1$, $\text{Var}_2$.

**Lemma 18** *Let $h, \alpha, \eta_0, \sigma > 0$ be constants such that $1 < \alpha < 2$. For an integer $N \in \mathbb{N}^+$ and a continuous variable $D > 0$, consider the sum:*

$$S = \sum_{k=1}^{N} hk^{-2\alpha} \int_0^{D\eta_0} \exp\left(-2hk^{-\alpha}(D\eta_0 - t)\right) \gamma(t)\sigma^2 dt,$$

*where $G : \mathbb{R} \to \mathbb{R}$ is the inverse function of $y \mapsto y + \sin y$, and $\gamma(t) = \eta_0 \left(1 + \cos G\left(\frac{t\pi}{D\eta_0}\right)\right)$.*

*We have*

$$S = \sigma^2 C_\alpha I_\gamma \eta_0^{1/\alpha} (Dh)^{1/\alpha - 1} + \mathcal{E},$$

*where $C_\alpha = \frac{2^{1/\alpha-2}}{\alpha}\Gamma(2 - 1/\alpha)$ and $I_\gamma = \int_0^1 \left[1 + \cos G(\pi(1-v))\right] v^{1/\alpha-2} dv$ are finite absolute constants. The approximation error $\mathcal{E}$ is strictly bounded by*

$$|\mathcal{E}| \le C_{13,1}(Dh)^{-2/3} + C_{13,2} D N^{1-2\alpha},$$

*where $C_{13,1}, C_{13,2}$ only depends on $\alpha, \sigma, \eta_0$.*

**Proof** Let $T = D\eta_0$ and let $s = T - t$. Define the function

$$f(x) := hx^{-2\alpha} \int_0^T \exp\left(-2hx^{-\alpha}s\right) \gamma(T - s)\sigma^2 ds.$$

The target summation can be written as $S = \sum_{k=1}^{N} f(k)$.

First, we need to bound the schedule function $\gamma(T - s)$ near $s = 0$. By definition, $\gamma(T - s) = \eta_0(1 - \cos x_s)$ where $x_s - \sin x_s = \frac{\pi s}{T}$. For $x_s \in [0, \pi]$, the function $(x_s - \sin x_s)/x_s^3$ achieves its minimum $1/\pi^2$ at $x_s = \pi$. Thus, $x_s - \sin x_s \ge x_s^3/\pi^2$, which implies $x_s \le \pi(s/T)^{1/3}$. Utilizing $1 - \cos x_s \le x_s^2/2$, we obtain a strict global bound:

$$\gamma(T - s) \le \eta_0 \frac{\pi^2}{2} \left(\frac{s}{T}\right)^{2/3}. \tag{73}$$

Now we bound the sum-to-integral gap $|S - \int_1^N f(x)dx| \le \int_1^N |f'(x)|dx$. Taking the derivative of $f(x)$ with respect to $x$:

$$|f'(x)| \le \int_0^T hx^{-2\alpha-1} \exp\left(-2hx^{-\alpha}s\right) \left(2\alpha + 2\alpha hsx^{-\alpha}\right) \gamma(T - s)\sigma^2 ds.$$

Integrating this bound over $x \in [1, \infty)$ and applying the change of variable $v = x^{-\alpha}$ gives

$$\int_1^\infty 2\alpha hx^{-2\alpha-1} e^{-2hx^{-\alpha}s} dx = 2h \int_0^1 ve^{-2hsv} dv \le \min\left(h, \frac{1}{2hs^2}\right),$$

and similarly the second term yields $\min\left(\frac{2}{3}h^2 s, \frac{1}{2hs^2}\right)$. Thus, the total variation is bounded by:

$$\int_1^N |f'(x)|dx \le \sigma^2 \int_0^T \gamma(T - s) \left[\min\left(h, \frac{1}{2hs^2}\right) + \min\left(\frac{2}{3}h^2 s, \frac{1}{2hs^2}\right)\right] ds.$$

Splitting the integral at $s = 1/h$ and substituting the bound (73), we get:

$$\begin{aligned}\int_1^N |f'(x)|dx &\le \sigma^2 \int_0^{1/h} \eta_0 \frac{\pi^2}{2} \left(\frac{s}{T}\right)^{2/3} \left(\frac{5}{3}h\right) ds + \sigma^2 \int_{1/h}^T \eta_0 \frac{\pi^2}{2} \left(\frac{s}{T}\right)^{2/3} \frac{1}{hs^2} ds \\ &= \frac{5\pi^2}{6}\sigma^2\eta_0 T^{-2/3} h \left[\frac{3}{5}\left(\frac{1}{h}\right)^{5/3}\right] + \frac{\pi^2}{2}\sigma^2\eta_0 T^{-2/3}\frac{1}{h}\left[3\left(\frac{1}{h}\right)^{-1/3}\right] \\ &= 2\pi^2\sigma^2\eta_0 (hT)^{-2/3} = 2\pi^2\sigma^2\eta_0^{1/3}(Dh)^{-2/3}.\end{aligned}$$

Next, we evaluate the continuous continuous integral over $x \in (0, \infty)$ and extract the main order $M$. Swapping the integration order yields:

$$\begin{aligned}\int_0^\infty f(x)dx &= \sigma^2 \int_0^T \gamma(T-s)\left(\int_0^\infty hx^{-2\alpha}\exp(-2hx^{-\alpha}s)dx\right) ds \\ &= \sigma^2 C_\alpha h^{1/\alpha - 1} \int_0^T \gamma(T-s)s^{1/\alpha - 2} ds.\end{aligned}$$

Using the dimensionless variable $v = s/T$, we factor out $T$ to match the stated integral $I_\gamma$:

$$\int_0^\infty f(x)dx = \sigma^2 C_\alpha I_\gamma \eta_0^{1/\alpha}(Dh)^{1/\alpha - 1} =: M. \tag{74}$$

To formalize the full sum $S = M + \mathcal{E}$, we decompose the error as $\mathcal{E} = \left(S - \int_1^N f(x)dx\right) - \int_0^1 f(x)dx - \int_N^\infty f(x)dx$. We bound the upper and lower truncation tails individually. For the upper integral tail $x \in [0, 1]$, we substitute (73) into $f(x)$ and evaluate the $s$-integral exactly:

$$\begin{aligned}f(x) &\le \sigma^2 \int_0^\infty hx^{-2\alpha}\exp(-2hx^{-\alpha}s)\left[\eta_0\frac{\pi^2}{2}\left(\frac{s}{T}\right)^{2/3}\right] ds \\ &= \frac{\pi^2}{2}\sigma^2\eta_0 h x^{-2\alpha} T^{-2/3}\Gamma(5/3)(2hx^{-\alpha})^{-5/3} = \frac{\pi^2\Gamma(5/3)}{2^{8/3}}\sigma^2\eta_0 T^{-2/3}h^{-2/3}x^{-\alpha/3}.\end{aligned}$$

Integrating this bounding function over $x \in [0, 1]$ directly gives:

$$\int_0^1 f(x)dx \le \frac{\pi^2\Gamma(5/3)}{2^{8/3}}\sigma^2\eta_0(hT)^{-2/3}\int_0^1 x^{-\alpha/3}dx = \frac{3\pi^2\Gamma(5/3)}{2^{8/3}(3-\alpha)}\sigma^2\eta_0^{1/3}(Dh)^{-2/3}.$$

For the lower integral tail $x \in [N, \infty)$, we use the trivial bound $\gamma(T-s) \le 2\eta_0$:

$$\begin{aligned}\int_N^\infty f(x)dx &\le \sigma^2 \int_N^\infty \left(\int_0^T hx^{-2\alpha}\exp(-2hx^{-\alpha}s)(2\eta_0)ds\right) dx \\ &\le \sigma^2 \int_N^\infty hx^{-2\alpha}(2\eta_0 T)dx = \frac{2\eta_0\sigma^2 hT}{2\alpha - 1}N^{1-2\alpha} = \frac{2\eta_0^2\sigma^2 Dh}{2\alpha-1}N^{1-2\alpha}.\end{aligned}$$

Combining all error sources yields

$$\mathcal{E} \le 2\pi^2\sigma^2\eta_0^{1/3}(Dh)^{-2/3} + \frac{3\pi^2\Gamma(5/3)}{2^{8/3}(3-\alpha)}\sigma^2\eta_0^{1/3}(Dh)^{-2/3} + \frac{2\eta_0^2\sigma^2 Dh}{2\alpha-1}N^{1-2\alpha} \tag{75}$$

$$\le C_{13,1}(Dh)^{-2/3} + C_{13,2}DN^{1-2\alpha}, \tag{76}$$

where $C_{13,1}, C_{13,2}$ only depends on $\alpha, \sigma, \eta_0$. ■

**Lemma 19** *Let $h, \alpha, \alpha_2, \eta_0, C > 0$ be positive constants. For an integer $N \in \mathbb{N}^+$, and a continuous scale parameter $D > 0$, consider the sum:*

$$S = \sum_{k=1}^{N} hk^{-2\alpha} \int_0^{D\eta_0} \exp\left(-2hk^{-\alpha}(D\eta_0 - t)\right) \gamma(t) \min(C, t^{-\alpha_2}) dt,$$

*where $G : \mathbb{R} \to \mathbb{R}$ is the inverse function of $y \mapsto y + \sin y$, and $\gamma(t) = \eta_0 \left(1 + \cos G\left(\frac{t\pi}{D\eta_0}\right)\right)$.*

*The summation $S$ is strictly bounded by:*

$$S \leq C_{14,1}(Dh)^{-2/3} + C_{14,2} DN^{1-2\alpha} + C_{15}(Dh)^{1/\alpha - 1} D^{-1},$$

*where $C_{15}$ only depends on $\alpha, C$ which are all constant.*

**Proof** Let $T = D\eta_0$. Define the function

$$f(x) := hx^{-2\alpha} \int_0^{T} \exp\left(-2hx^{-\alpha}(T - t)\right) \gamma(t) \min(C, t^{-\alpha_2}) dt.$$

The target summation can be exactly written as $S = \sum_{k=1}^{N} f(k)$.

Applying the change of variable $s = T - t$, we rewrite $f(x)$ as:

$$f(x) = hx^{-2\alpha} \int_0^{T} \exp\left(-2hx^{-\alpha} s\right) \gamma(T - s) \min(C, (T - s)^{-\alpha_2}) ds.$$

Since $\min(C, (T - s)^{-\alpha_2}) \leq C$ for all $s \in [0, T]$, the function $f(x)$ is uniformly bounded by the kernel analyzed in Lemma 18 by setting $\sigma^2 = C$. Namely:

$$f(x) \leq hx^{-2\alpha} \int_0^{T} \exp\left(-2hx^{-\alpha} s\right) \gamma(T - s) C ds.$$

We define the continuous principal integral as $\mathcal{I} = \int_0^{\infty} f(x) dx$. The total sum is $S = \mathcal{I} + \mathcal{E}$, where the approximation error decomposes as $\mathcal{E} = \left(\sum_{k=1}^{N} f(k) - \int_1^N f(x) dx\right) - \int_0^1 f(x) dx - \int_N^{\infty} f(x) dx$. By directly substituting $\sigma^2 = C$ into the established derivative bounds and integral tails of Lemma 18, we strictly bound the error magnitudes:

$$\begin{aligned}
|\mathcal{E}| &\leq \int_1^N |f'(x)| dx + \int_0^1 f(x) dx + \int_N^{\infty} f(x) dx \\
&\leq 2\pi^2 C \eta_0^{1/3} (Dh)^{-2/3} + \frac{3\pi^2 \Gamma(5/3)}{2^{8/3}(3 - \alpha)} C \eta_0^{1/3} (Dh)^{-2/3} + \frac{2\eta_0^2 Ch}{2\alpha - 1} DN^{1-2\alpha} \\
&= C_{14,1}(Dh)^{-2/3} + C_{14,2} DN^{1-2\alpha}.
\end{aligned}$$

Next, we evaluate the continuous principal integral $\mathcal{I}$. Swapping the order of integration yields:

$$\mathcal{I} = \int_0^{T} \gamma(T - s) \min(C, (T - s)^{-\alpha_2}) \left( \int_0^{\infty} hx^{-2\alpha} \exp(-2hx^{-\alpha} s) dx \right) ds.$$

Applying the substitution $v = x^{-\alpha}$ with $dx = -\frac{1}{\alpha}v^{-1/\alpha-1}dv$, the inner integral evaluates to $C_\alpha h^{1/\alpha-1}s^{1/\alpha-2}$. Reversing the initial temporal substitution ($t = T - s$), the principal integral becomes exactly:

$$\mathcal{I} = C_\alpha h^{1/\alpha-1} \int_0^T (T-t)^{1/\alpha-2}\gamma(t)\min(C, t^{-\alpha_2})dt. \tag{77}$$

To rigorously bound this, we split the integration domain at $t = T/2$ such that $\mathcal{I} = \mathcal{I}_1 + \mathcal{I}_2$. For the first half-domain $t \in [0, T/2]$, the term $(T-t) \geq T/2$. Since $1/\alpha - 2 < 0$, we have $(T-t)^{1/\alpha-2} \leq (T/2)^{1/\alpha-2}$. Furthermore, $\gamma(t) \leq 2\eta_0$. Bounding the integral of the cutoff function over this local domain by its global integral over $(0, \infty)$ yields:

$$\int_0^{T/2} \min(C, t^{-\alpha_2})dt \leq \int_0^{C^{-1/\alpha_2}} Cdt + \int_{C^{-1/\alpha_2}}^{\infty} t^{-\alpha_2}dt = \frac{\alpha_2}{\alpha_2 - 1}C^{1-1/\alpha_2}.$$

Multiplying these individual maximum bounds, we obtain:

$$\begin{aligned}\mathcal{I}_1 &\leq C_\alpha h^{1/\alpha-1}\left(\frac{T}{2}\right)^{1/\alpha-2}(2\eta_0)\left(\frac{\alpha_2}{\alpha_2-1}C^{1-1/\alpha_2}\right) \\ &= \frac{2^{3-1/\alpha}\alpha_2 C_\alpha \eta_0^{1/\alpha-1}}{\alpha_2 - 1}C^{1-1/\alpha_2}\frac{(Dh)^{1/\alpha-1}}{D}.\end{aligned}$$

For the second half-domain $t \in [T/2, T]$, the variable is bounded away from zero. Consequently, the minimum function is strictly bounded by its algebraic tail: $\min(C, t^{-\alpha_2}) \leq t^{-\alpha_2} \leq (T/2)^{-\alpha_2} = 2^{\alpha_2}T^{-\alpha_2}$. Factoring out this upper bound allows us to conservatively extend the remaining integral to the full domain $[0, T]$:

$$\begin{aligned}\mathcal{I}_2 &\leq 2^{\alpha_2}T^{-\alpha_2}C_\alpha h^{1/\alpha-1}\int_{T/2}^T (T-t)^{1/\alpha-2}\gamma(t)dt \\ &\leq 2^{\alpha_2}T^{-\alpha_2}C_\alpha h^{1/\alpha-1}\int_0^T (T-t)^{1/\alpha-2}\gamma(t)dt.\end{aligned}$$

Applying the dimensionless change of variable $v = 1 - t/T$ sets $dt = Tdv$ and $T - t = Tv$, fully recovering the constant $I_\gamma$:

$$\int_0^T (T-t)^{1/\alpha-2}\gamma(t)dt = T^{1/\alpha-1}\eta_0\int_0^1 v^{1/\alpha-2}\left[1 + \cos G(\pi(1-v))\right]dv = T^{1/\alpha-1}\eta_0 I_\gamma.$$

Substituting this back and recalling $T = D\eta_0$, we obtain:

$$\mathcal{I}_2 \leq 2^{\alpha_2}(D\eta_0)^{-\alpha_2}C_\alpha h^{1/\alpha-1}\left((D\eta_0)^{1/\alpha-1}\eta_0 I_\gamma\right) = 2^{\alpha_2}C_\alpha I_\gamma \eta_0^{1/\alpha-\alpha_2}\frac{(Dh)^{1/\alpha-1}}{D^{\alpha_2}}.$$

Summing the evaluated components gives

$$S \leq C_{14,1}(Dh)^{-2/3} + C_{14,2}DN^{1-2\alpha} + \frac{2^{3-1/\alpha}\alpha_2 C_\alpha \eta_0^{1/\alpha-1}}{\alpha_2-1}C^{1-1/\alpha_2}\frac{(Dh)^{1/\alpha-1}}{D} + 2^{\alpha_2}C_\alpha I_\gamma \eta_0^{1/\alpha-\alpha_2}\frac{(Dh)^{1/\alpha-1}}{D^{\alpha_2}} \tag{78}$$

$$\leq C_{14,1}(Dh)^{-2/3} + C_{14,2}DN^{1-2\alpha} + C_{15}(Dh)^{1/\alpha-1}D^{-1}, \tag{79}$$

where $C_{15}$ only depends on $\alpha, C$ which are all constant. ■

Now, we are ready to proof Theorem 15.

By Lemma 18, $\text{Val}_1$ satisfies

$$\text{Val}_1 = \sigma^2 C_{\alpha_\tau} I_\gamma \eta_0^{1/\alpha_\tau} (Dh_\tau)^{1/\alpha_\tau - 1} + \mathcal{E}_1, \tag{80}$$

where $C_{\alpha_\tau}$ and $I_\gamma$ are the absolute constants defined in Lemma 18 and $\mathcal{E}_1$ satisfies

$$|\mathcal{E}_1| \le C_{13,1}(Dh_\tau)^{-2/3} + C_{13,2} D N^{1-2\alpha_\tau} + \int_0^{D\eta_0} \exp(-2H^{-\alpha_{\max}}(D\eta_0 - t))\gamma(t)\sigma'^2 \mathrm{d}t \tag{81}$$

$$\le C_{13,1}(Dh_\tau)^{-2/3} + C_{13,2} D N^{1-2\alpha_\tau} + \frac{\pi^2}{2}(D\eta_0)^{-2/3}\Gamma(5/3)(2h_\tau H^{-\alpha_{\max}})^{-5/3} \tag{82}$$

$$\le C_{12}(Dh_\tau)^{-2/3} + C_{13,2} D N^{1-2\alpha_\tau}. \tag{83}$$

$\int_0^{D\eta_0} \exp(-2H^{-\alpha_{\max}}(D\eta_0 - t))\gamma(t)\sigma'^2 \mathrm{d}t$ is induced by the error of the shared head and $C_{12} = C_{13,1} + \frac{\pi^2}{2}\eta_0^{-2/3}\Gamma(5/3)H^{5\alpha/3}$ which does not depend on $h$.

By Lemma 17, we have

$$\text{Val}_1 \le \text{Val}_2 \le \text{Val}_1 + \sum a_k \lambda_k \int_0^{D\eta_0} \exp(-2\lambda_k(D\eta_0 - t))\gamma(t)\frac{C_0}{\underline{h}} \left( \frac{C'}{\max\left(t^{1-\frac{1}{\alpha_{\max}}}, LC'\right)} + \frac{C''}{N^{\alpha_{\min}\left(1-\frac{1}{\alpha_{\max}}\right)}} \right) \mathrm{d}t \tag{84}$$

By Lemma 19, we have

$$\sum a_k \lambda_k \int_0^{D\eta_0} \exp(-2\lambda_k(D\eta_0 - t))\gamma(t)\frac{C_0}{\underline{h}} \left( \frac{C'}{\max\left(t^{1-\frac{1}{\alpha_{\max}}}, LC'\right)} + \frac{C''}{N^{\alpha_{\min}\left(1-\frac{1}{\alpha_{\max}}\right)}} \right) \mathrm{d}t \le (C_{12}C_{14,1})(Dh_\tau)^{-2/}$$
$$+ C_{14,2} D N^{1-2\alpha_\tau}$$
$$+ C_{15}(Dh)^{1/\alpha_\tau - 1} D^{-1}$$

Therefore,

$$\text{Val}_2 \le \text{Val}_1 + (C_{12}C_{14,1})(Dh_\tau)^{-2/3} + C_{14,2} D N^{1-2\alpha_\tau} + C_{15}(Dh)^{1/\alpha_\tau - 1} D^{-1}. \tag{85}$$

### C.7. Proof of the Main Theorem

Combining Theorem 12, Theorem 14 and Theorem 15, we have our final main theorem:

**Theorem 20 (Formal Statement of Theorem 2)** *Consider the projected linear regression model defined in Section C.1 and consider an arbitrary domain $\tau \in [K]$. Let $x^*$ be the solution to Problem 3 with parameters $c_i = \frac{1}{\alpha_i - 1}, b_i = \alpha_i - 1$. Then for any $\epsilon > 0$, there exist $N_0$ and $D_0$ such that for all $N > N_0$, $D > D_0$, and $h$ satisfying Assumption 4, the expected test loss on domain $\tau$ is bounded by:*

$$L_\tau(h, N, D) \ge c_\tau (x^*(N,h)_\tau)^{-b_\tau}(1-\epsilon) + \frac{A}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}}(1-\epsilon) + \sigma^2,$$

$$L_\tau(h, N, D) \le c_\tau (x^*(N,h)_\tau)^{-b_\tau}(1+\epsilon) + \frac{A}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}}(1+\epsilon) + \sigma^2,$$

*where* $A := \frac{\Gamma\left(1-\frac{1}{\alpha_\tau}\right)}{\alpha_\tau(2\eta_0)^{1-\frac{1}{\alpha_\tau}}} + \sigma^2 C_{\alpha_\tau} I_\gamma \eta_0^{1/\alpha_\tau}$.

**Proof** By Theorem 12, we have

$$\frac{\left|\text{Approx} - \frac{(x_\tau^*)^{1-\alpha_\tau}}{\alpha_\tau - 1}\right|}{\frac{(x_\tau^*)^{1-\alpha_\tau}}{\alpha_\tau - 1}} \leq \frac{\frac{\alpha_\tau}{N^{\alpha_\tau+1}}}{\frac{(x_\tau^*)^{1-\alpha_\tau}}{\alpha_\tau-1}} \tag{86}$$

$$\leq \frac{\alpha_\tau(\alpha_\tau-1)}{N^{\alpha_\tau+1-\frac{\alpha_{\min}}{\alpha_\tau}(\alpha_\tau-1)}} \tag{87}$$

$$\leq \frac{\alpha_\tau(\alpha_\tau-1)}{N^{\alpha_\tau-\alpha_{\min}+1+\frac{\alpha_{\min}}{\alpha_\tau}}}. \tag{88}$$

By Theorem 14,

$$\frac{\left|\text{Bias} - \frac{\Gamma\left(1-\frac{1}{\alpha_\tau}\right)}{\alpha_\tau(2\eta_0)^{1-\frac{1}{\alpha_\tau}}} \frac{1}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}}\right|}{\frac{\Gamma\left(1-\frac{1}{\alpha_\tau}\right)}{\alpha_\tau(2\eta_0)^{1-\frac{1}{\alpha_\tau}}} \frac{1}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}}} \leq \frac{\frac{C_{10,1}}{D} + \frac{C_{10,2}}{N^{\alpha_{\min}\left(1-\frac{1}{\alpha_\tau}\right)}}}{\frac{\Gamma\left(1-\frac{1}{\alpha_\tau}\right)}{\alpha_\tau(2\eta_0)^{1-\frac{1}{\alpha_\tau}}} \frac{1}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}}} \tag{89}$$

$$\leq C'_{10,1} D^{-1/\alpha_\tau} + C'_{10,2} D^{-\varepsilon(1-1/\alpha_\tau)}, \tag{90}$$

where $C'_{10,1} := C_{10,1} \cdot \left(\frac{\Gamma\left(1-\frac{1}{\alpha_\tau}\right)}{\alpha_\tau(2\eta_0)^{1-\frac{1}{\alpha_\tau}}}\right)^{-1}$.

By Theorem 15,

$$\frac{\left|\text{Var} - \sigma^2 C_{\alpha_\tau} I_\gamma \eta_0^{1/\alpha_\tau} (Dh_\tau)^{1-1/\alpha_\tau}\right|}{\sigma^2 C_{\alpha_\tau} I_\gamma \eta_0^{1/\alpha_\tau} (Dh_\tau)^{1-1/\alpha_\tau}} \leq \frac{(C_{14,1} + 2C_{12} + C_{15})(Dh_\tau)^{-2/3} + (C_{14,2} + C_{13,2}) D N^{1-2\alpha_\tau}}{\sigma^2 C_{\alpha_\tau} I_\gamma \eta_0^{1/\alpha_\tau} (Dh_\tau)^{1-1/\alpha_\tau}} \tag{91}$$

$$\leq C_{17,1} D^{-1/\alpha_\tau+1/3} + C_{17,2} D^{-2(1+\varepsilon)(1-1/\alpha_\tau)}, \tag{92}$$

where $C_{17,1}, C_{17,2}$ do not depend on $h, \epsilon$.

For all $h$ satisfying Assumption 4, since $\alpha_\tau - \alpha_{\min} + 1 + \frac{\alpha_{\min}}{\alpha_\tau} > 0, -1/\alpha_\tau + 1/3 > 0$ , we can select $D_0, N_0$ such that

$$\frac{2\alpha_\tau(\alpha_\tau-1)}{N_0^{\alpha_\tau-\alpha_{\min}+1+\frac{\alpha_{\min}}{\alpha_\tau}}} \leq \epsilon \tag{93}$$

$$C'_{10,1} D_0^{-1/\alpha_\tau} + C'_{10,2} D_0^{-\varepsilon} \leq \epsilon \tag{94}$$

$$C_{17,1} D_0^{-1/\alpha_\tau+1/3} + C_{17,2} D_0^{-\varepsilon} \leq \epsilon. \tag{95}$$

For all $N > N_0, D > D_0$ satisfying Assumption 8, we have

$$\text{Approx} \in \left[ \frac{(x_\tau^*)^{1-\alpha_\tau}}{\alpha_\tau - 1}(1-\epsilon), \frac{(x_\tau^*)^{1-\alpha_\tau}}{\alpha_\tau - 1}(1+\epsilon) \right], \tag{96}$$

$$\text{Bias} \in \left[ \frac{\Gamma\left(1-\frac{1}{\alpha_\tau}\right)}{\alpha_\tau (2\eta_0)^{1-\frac{1}{\alpha_\tau}}} \frac{1}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}}(1-\epsilon), \frac{\Gamma\left(1-\frac{1}{\alpha_\tau}\right)}{\alpha_\tau (2\eta_0)^{1-\frac{1}{\alpha_\tau}}} \frac{1}{(Dh_\tau)^{1-\frac{1}{\alpha_\tau}}}(1+\epsilon) \right], \tag{97}$$

$$\text{Var} \in \left[ \sigma^2 C_{\alpha_\tau} I_\gamma \eta_0^{1/\alpha_\tau} (Dh_\tau)^{1-1/\alpha_\tau}(1-\epsilon), \sigma^2 C_{\alpha_\tau} I_\gamma \eta_0^{1/\alpha_\tau} (Dh_\tau)^{1-1/\alpha_\tau}(1+\epsilon) \right]. \tag{98}$$

By Theorem 11, adding up these three terms, we complete the proof. ■