Lower Bounds for the Algorithmic Complexity of Learned Indexes

Luis Alberto Croquevielle
Imperial College London, London, United Kingdom

Roman Sokolovskii
Imperial College London, London, United Kingdom

Thomas Heinis
Imperial College London, United Kingdom

Abstract

Learned index structures aim to accelerate queries by training machine learning models to approximate the rank function associated with a database attribute. While effective in practice, their theoretical limitations are not fully understood. We present a general framework for proving lower bounds on query time for learned indexes, expressed in terms of their space overhead and parameterized by the model class used for approximation. Our formulation captures a broad family of learned indexes, including most existing designs, as piecewise model-based predictors.

We solve the problem of lower bounding query time in two steps: first, we use probabilistic tools to control the effect of sampling when the database attribute is drawn from a probability distribution. Then, we analyze the approximation-theoretic problem of how to optimally represent a cumulative distribution function with approximators from a given model class. Within this framework, we derive lower bounds under a range of modeling and distributional assumptions, paying particular attention to the case of piecewise linear and piecewise constant model classes, which are common in practical implementations.

Our analysis shows how tools from approximation theory, such as quantization and Kolmogorov widths, can be leveraged to formalize the space-time tradeoffs inherent to learned index structures. The resulting bounds illuminate core limitations of these methods.

2012 ACM Subject Classification Theory of computation → Predecessor queries; Theory of computation → Data structures and algorithms for data management

Keywords and phrases Learned Indexes, Stochastic Processes, Approximation Theory

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

1 Introduction

Efficient query processing is a fundamental challenge in database systems. For relational databases, key areas of study include join optimization techniques [31, 17], cardinality estimation methods [2, 27, 19], and spatial indexing for multidimensional data [10, 14]. In general, improvements in query latency involve trade-offs with other resources, such as space overhead or the number of I/O operations.

For one-dimensional indexing, tree-based structures such as B+trees [12] are the predominant choice in practice. These methods aim to answer point and range queries over a single, ordered attribute. Typically, the attribute values (or *keys*) are stored in a sorted structure A, and queries are expressed as search operations over A. In particular, a point query for a value q requires locating the position of q in A, while a range query over an interval $[q, q']$ requires identifying the first key $\geq q$ and scanning forward until a key $> q'$ is reached.

In recent years, machine learning has emerged as a promising tool for query processing. A prominent example for one-dimensional indexing is the Recursive Model Index [18], which maintains a hierarchical structure similar to B+trees but embeds predictive models in internal nodes to estimate key positions within subsets of sorted data. This work has inspired a

© Author: Please fill in the \Copyright macro;
licensed under Creative Commons License CC-BY 4.0

42nd Conference on Very Important Topics (CVIT 2016).
Editors: John Q. Open and Joan R. Access; Article No. 23; pp. 23:1–23:34
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

broad line of research into "learned indexes" [11, 16, 15, 5], which extend this idea to support dynamic workloads [5, 29], streaming data [30, 21], and multi-dimensional queries [25, 6, 20].

While empirical studies have highlighted their potential advantages over traditional methods [23, 28], learned indexes generally lack formal guarantees. Their performance has also been observed to vary across different datasets and query workloads, in contrast to the consistent behavior of well-established methods such as B+trees. Proving algorithmic guarantees for learned indexes and characterizing the data-dependent behavior of these methods are all important theoretical questions that have been explored only partially.

In this work, we introduce a formal framework that captures a broad class of learned index structures. Within this framework, we establish lower bounds on query time, explicitly characterizing the trade-off between time and space complexity. These bounds hold under general assumptions, covering a wide range of data-generating processes. As an interesting application, these lower bounds allow us to identify a parameter regime where learned indexes offer no asymptotic advantage over traditional methods.

The paper is organized as follows: in Section 2, we present a more detailed account of related work and our own contributions. Section 3 introduces the mathematical setting and notation. Section 4 develops our general theoretical approach, while Sections 5 and 6 instantiate this framework by applying it to specific function approximation scenarios. Finally, Section 7 contains concluding remarks and a discussion of future work.

2 Related Work and Contributions

The next section introduces our full notation, but we provide some preliminary definitions here. We model an attribute A in a database relation as formed by a sequence $X_1, \ldots, X_n$ of numeric values. The central object of study in our analysis is the **rank** function, defined as $\mathbf{rank}(q) = \sum_{i=1}^{n} \mathbf{1}_{(-\infty, X_i]}(q)$, which counts how many keys are less than or equal to a given query value q. In the context of learned indexes, query time is typically understood as the time required to evaluate $\mathbf{rank}(q)$.

Classical methods such as B+trees are agnostic to the process by which data is generated. In contrast, learned indexes derive performance guarantees by making assumptions about it, e.g., that the $\{X_i\}$ are i.i.d. with respect to some distribution. This is to be expected in light of related areas of research, in particular statistical learning theory, where generalization guarantees usually rely on i.i.d. assumptions for both the training and testing data.

As a baseline, B+trees support search, insertion, and deletion in $O(\log n)$ time using $O(n)$ space, with worst-case guarantees that hold independently of the data-generating process. Initial theoretical analysis of learned indexes, beginning with the PGM-Index [9], showed performance comparable to B+trees in the static setting, but with $O(\log^2 n)$ search complexity for dynamic workloads. Under (restrictive) assumptions on the data-generating process, a constant improvement in space complexity relative to B+trees was shown [8].

More recent work has established that specific learned indexes can asymptotically outperform B+trees in expectation, for the case of static databases. Zeighami and Shahabi [32] proved that $O(\log(\log n))$ expected search time is achievable with linear space. Croquevielle et al. [3] further improved this to $O(1)$ expected time. For dynamic databases, Zeighami and Shahabi [33] considered a model where the data distribution can vary over time, with a parameter δ capturing the amount of distribution shift. They introduce a learned index that supports insertions and queries in $O(\log(\log n))$ expected time when $\delta = 0$ (i.e., no distribution shift), an asymptotic improvement over B+trees. For $\delta > 0$, an extra $O(\log(\delta n))$ cost is incurred, so that asymptotic complexity remains the same as for B+trees.

A complementary line of work was recently introduced by Zeighami and Shahabi [34], who study lower bounds for learned database operations from an information-theoretic perspective. In their framework, a learned index is modeled as a parameterized function f_θ where the parameters θ are encoded using a total budget of σ bits. The authors derive lower bounds on the model size σ, showing that if it is too small, there exists at least one input sequence $X_1, \ldots, X_n$ for which a desired level of approximation accuracy cannot be satisfied.

Our work takes a fundamentally different approach. Instead of analyzing a general function f_θ, we focus on a structured class $\mathcal{I}$ of indexing algorithms that captures common design elements found in most learned index constructions. While this narrows the scope of our work compared to the general setting of [34], it enables sharper lower bounds that are more directly relevant to learned indexes as commonly defined in the literature. Within this class, we establish lower bounds on query time as a function of the space overhead used by the indexing structure.

More specifically, we model learned indexes as directed acyclic graphs (DAGs), where internal nodes direct the search, and sink nodes encode predictive models and store the keys. We develop a general framework for proving query time lower bounds as a function of the model class complexity, the number of sink nodes K, the size of the dataset n, and the choice of local search algorithm (e.g., linear or binary search).

For a broad class of learned indexes, capturing most known designs, we prove that query time is lower bounded by $\Omega(\log(n/K))$ in the worst case. A key implication is that, in order to achieve asymptotic improvement over traditional structures like B+trees (which guarantee $O(\log n)$ query time with linear space), a learned index must use nearly linear space. Specifically, if the number of sink nodes satisfies $K = O(n^\alpha)$ for any $0 < \alpha < 1$, the worst-case query time remains $\Omega(\log n)$—matching that of B+trees. Thus, learned indexes can only hope for worst-case asymptotic gains when space usage is close to linear. When the model class is restricted to piecewise constant functions, this same conclusion further applies to *average-case query time*, where the expectation is taken over the query distribution.

3 Preliminaries

3.1 Data as Stochastic Process

We consider a probabilistic model for a database attribute. Formally, we model the attribute as a stochastic process $X = \{X_i\}_{i\in\mathbb{N}}$, where each X_i is a real-valued random variable. At time $n \in \mathbb{N}$, the attribute consists of the first n values in the sequence, sorted in increasing order. We denote the sorted values by $X_{(1)} \leq \ldots \leq X_{(n)}$, so that the attribute is represented as an array A_n with $A_n[i] = X_{(i)}$ for all $i \in \{1, \ldots, n\}$. This framework accommodates both worst-case and average-case analyses: we may consider a single sequence $X_1, \ldots, X_n$ to study worst-case behavior, or incorporate the probability measure $\mathbb{P}$ governing the stochastic process to analyze expected performance.

No complexity lower bounds will apply in full generality. For instance, if $X_i = i$, then the rank function **rank** can be computed in $O(1)$ time and space, and existing structures such as the PGM-Index easily attain that performance. In this work, we derive lower bounds under the assumption that the X_i are independent random variables drawn according to a common cumulative distribution function (CDF) F. We further assume that the query value q is generated independently from the X_i according to a probability measure μ which

has support $\mathcal{Q} \subseteq \mathbb{R}$.[1] We emphasize that independence assumptions are only required for average-case analyses; our worst-case bounds hold unconditionally, though independence is used as a technical tool in their derivation.

Throughout the paper, we use $\mathbb{E}_X[\cdot]$ to denote expectation with respect to the measure $\mathbb{P}$ governing the key sequence $\{X_i\}_{i=1}^n$. The notation $\mathbb{E}_q[\cdot]$ denotes expectation with respect to the query distribution μ, treating the key values $\{X_i\}_{i=1}^n$ as fixed (i.e., conditioning on a realization of the data). When expectations are taken jointly, we write $\mathbb{E}_{X,q}[\cdot]$, corresponding to integration with respect to the product measure $\mathbb{P} \otimes \mu$.

3.2 Computational Model and Learning Problem

As introduced in Section 2, we consider the **rank** function over the attribute A_n, given by

$$\mathbf{rank}_n(q) = \sum_{i=1}^{n} \mathbf{1}_{(-\infty, X_i]}(q)\,,$$

where $\mathbf{1}_U$ denotes the indicator function of a set U. Hence, $\mathbf{rank}_n(q)$ maps a query value $q \in \mathbb{R}$ to the number of keys less than or equal to it. Since it is usually clear from context, we omit the sub-index n and write **rank** instead. In learned indexes, the goal is to accurately approximate $\mathbf{rank}(q)$ while minimizing the time and space overhead needed to compute it.

We adopt the standard RAM model of computation under the uniform cost criterion [1], where each basic instruction requires one time unit and each register, storing an arbitrary integer, uses one space unit. *Time complexity* is defined as the number of basic instructions required to evaluate $\mathbf{rank}(q)$ using an index. *Space complexity* measures the redundant memory used by the index, excluding the space needed to store the attribute A_n itself.

Let I be an indexing procedure that builds a data structure $I(A_n)$ from the attribute A_n. For a query value q, let $T(I(A_n), q)$ denote the time required to compute $\mathbf{rank}(q)$ using the data structure $I(A_n)$, and let $S(I(A_n))$ denote the space overhead used by $I(A_n)$. Since it is always clear from context, we usually write $T(q)$ in place of $T(I(A_n), q)$.

3.3 Learned Index Structures

We model a learned index as a directed acyclic graph (DAG), as shown in Figure 1. Internal nodes implement routing functions, and the sink nodes contain predictive models and store the keys. Each query follows a path from the root to a sink node, where a prediction of its rank is refined via local search. This structure captures a wide range of learned indexes, most of which include hierarchical routing—such as RMI [18], ALEX [5], and PGM-Index [9].

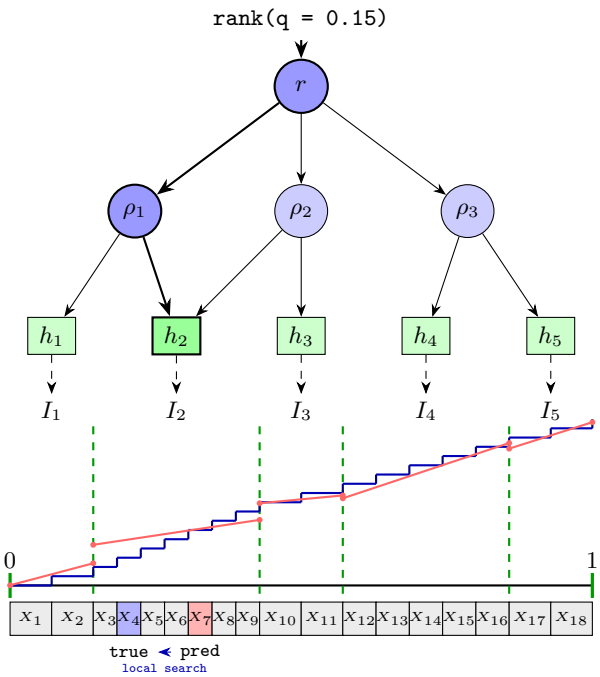
For the purpose of lower bounds, we focus on the *local correction step*. Regardless of how queries are routed, the exact rank of q is ultimately recovered by searching around the predicted value $h(q)$ in the sorted array. Since we aim for lower bounds that hold regardless of internal structure—which can vary substantially between different designs—the cost of this local search becomes the main driver of query complexity in our analysis.

We therefore abstract the index as a piecewise function defined by its sink-node models, and let K denote the number of sinks. This corresponds to a partitioning of the query domain into K regions $\{I_v\}_{v=1}^K$, each associated with a model h_v. Although we do not analyze the

[1] Formally, one must specify measurable spaces for the data and query distributions. For simplicity, we assume that all relevant spaces are equipped with the Borel σ-algebra and that all functions we integrate are measurable; these assumptions are standard and introduce no meaningful restrictions.

routing structure, we include it in the model to emphasize that our lower bounds apply regardless of graph topology or the complexity of the routing functions. Understanding how this structure impacts query time remains an open question and may lead to sharper bounds.

We denote by $\mathcal{I}$ the class of learned indexes defined in this way, and by $\mathcal{I}_{K,\mathcal{L}}$ those whose sink-node models belong to a function class $\mathcal{L}$ (e.g., $\mathcal{L}$ could be the set of polynomial functions) and use at most K sinks. For the rest of the paper, we use K as a lower bound (and proxy) for space usage, and we focus on how the approximation power of $\mathcal{L}$ interacts with K to characterize performance.



Root node: All queries enter here and are routed through the DAG. An example route for $q = 0.15$ is highlighted with thicker traversal lines.

Internal nodes: Each internal node v applies a routing function $\rho_v(q)$ to direct the query to one of its children.

Sink nodes: Each sink node v contains a model $h_v(q)$ that estimates $\mathbf{rank}(q)$ within the subinterval I_v.

In this example, the true rank function is shown as a blue step function, and the predictive models h_v are represented as red linear segments.

Local search: The estimate $h_v(q)$ is corrected through a local search over a subarray stored at the sink node.

Figure 1 Illustration of a learned index modeled as a directed acyclic graph (DAG).

4 General Framework for Lower Bounds

We now present a general framework for lower bounding the query time of learned indexes. We consider the query time $T(q)$ as the number of operations needed to compute $\mathbf{rank}(q)$ using the index built on the array A_n. The bounds are expressed in terms of the space overhead $S(R(A_n))$, via the number of sink nodes K. The analysis is based on three ideas:

(a) Relating the query time $T(q)$ to the prediction error $\varepsilon(q) = |h_v(q) - \mathbf{rank}(q)|$. Our analysis focuses on the final step of query execution: the local search after the model prediction. Since this step is always required to confirm or correct the predicted rank, its cost provides a valid lower bound on $T(q)$, which grows with the magnitude of $\varepsilon(q)$.

(b) Reducing the analysis of the prediction error $\varepsilon(q)$ to a function approximation problem for the CDF F, by using the approximation $\mathbf{rank}(q) \approx nF(q)$.

(c) A probabilistic analysis that formalizes the approximation $\mathbf{rank}(q) \approx nF(q)$ and allows us to derive lower bounds on $T(q)$ under various probabilistic assumptions.

We begin by describing components (b) and (c), and defer discussion of (a) to the end of this section, where we provide general lower bounds for $T(q)$. For clarity, most proofs are omitted during this exposition, and all omitted technical details are included in the appendix.

4.1 From Learned Indexes to Approximation Error

A learned index as defined in Section 3.3 can be seen as a structure that approximates $\mathbf{rank}(q)$ through a piecewise-defined prediction function

$$h(q) = \sum_{v=1}^{K} h_v(q)\mathbf{1}_{I_v}(q)\,, \tag{1}$$

where each h_v is a predictive model and $I_v \subseteq \mathcal{Q}$ its associated subdomain. For notational convenience, we identify an index I with its induced predictive model h; thus, we take $h \in \mathcal{I}_{K,\mathcal{L}}$ to mean that h is defined piecewise over K disjoint intervals using functions from the class $\mathcal{L}$.

Given a predictive function h, the prediction error is $\varepsilon(q) = |\mathbf{rank}(q) - h(q)|$. To analyze $\varepsilon(q)$ from the perspective of function approximation, we use the *empirical cumulative distribution function* (ECDF). Recall that our standing assumption is that the keys $X_1, \ldots, X_n$ are drawn i.i.d. from a distribution with CDF F. The ECDF is correspondingly defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty, X_i]}(x)\,.$$

The rank function can then be expressed as $\mathbf{rank}(q) = nF_n(q)$. Now, for both average and worst-case analysis, it is relevant to consider the expected error over all query values q, conditioned on $X_1, \ldots, X_n$. Since q is generated independently from $\{X_i\}$, we have

$$\mathbb{E}_q\big[\varepsilon(q)\big] = \int_{\mathcal{Q}} \varepsilon(q)d\mu = \int_{\mathcal{Q}} |\mathbf{rank}(q) - h(q)|d\mu = \int_{\mathcal{Q}} |nF_n(q) - h(q)|d\mu = \|nF_n - h\|_\mu\,, \tag{2}$$

where $\|\cdot\|_\mu$ denotes the absolute norm in $L^1(\mu)$, the space of μ-integrable functions over $\mathcal{Q}$. In other words, the best h function on average would be the function that minimizes the absolute distance to $\mathbf{rank} = nF_n$. This motivates the introduction of the *minimum approximation error*:

$$R_{K,\mathcal{L}}(G) = \inf_{h \in \mathcal{I}_{K,\mathcal{L}}} \|G - h\|_\mu\,,$$

for any $G \in L^1(\mu)$. When $G = \mathbf{rank} = nF_n$, this yields

$$R_{K,\mathcal{L}}(\mathbf{rank}) = R_{K,\mathcal{L}}(nF_n) = nR_{K,\mathcal{L}}(F_n)\,, \tag{3}$$

by the linear scaling property of the approximation error (see Lemma 7 in Appendix A). This allows us to reduce the analysis of $\mathbb{E}_q[\varepsilon(q)]$ to an approximation problem. The key idea is that $\mathbb{E}_q[\varepsilon(q)]$ can be lower bounded by the optimal $L^1(\mu)$-approximation of F, up to a small deviation term arising from sampling randomness. This deviation is captured by the function $\delta_n(x) = F(x) - F_n(x)$, where F_n is the ECDF. The following result makes this relationship precise. The proof is straightforward and can be found in Appendix A.

▶ **Proposition 1.** *Let $X_1, \ldots, X_n$ be i.i.d. samples from a distribution with CDF F, and let $q \sim \mu$ independently from the $\{X_i\}$. Then for any predictive model $h \in \mathcal{I}_{K,\mathcal{L}}$ it holds that*

$$\mathbb{E}_q\big[\varepsilon(q)\big] = \|nF_n - h\|_\mu \geq n\left[R_{K,\mathcal{L}}(F) - \|\delta_n\|_\mu\right]\,,$$

provided $\mathcal{L}$ is invariant under scalar multiplication.

This inequality shows that the expected prediction error is lower bounded by a deterministic approximation term $nR_{K,\mathcal{L}}(F)$, minus a data-dependent fluctuation term $n\|\delta_n\|_\mu$. Controlling this fluctuation is the main probabilistic step in our analysis and will be addressed in the next subsection. In addition to $\mathbb{E}_q\big[\varepsilon(q)\big]$, we also need bounds on $\mathbb{E}_q\big[\log_2 \varepsilon(q)\big]$, which arises in the analysis of exponential search. To this end, we define a proxy for query time: $\widetilde{T}(q) = \log_2 (\max\{2, \varepsilon(q)\})$, and establish a lower bound for its expectation. This requires a careful analysis and specific assumptions on F, μ, and $\mathcal{L}$. The following result illustrates the type of bound that can be derived (the proof can be found in Appendix B).

▶ **Proposition 2.** *Let $q \sim \mu$ independently from the keys $\{X_i\}_{i=1}^n$. Assume that the CDF F and μ have respective densities f and g, such that $0 < c_F \leq f(q), g(q) \leq C_F < \infty$ for all $q \in \mathcal{Q}$. Then, there exists a realization of $X_1, \ldots, X_n$ such that:*

$$\mathbb{E}_q\big[\widetilde{T}(q)\big] \geq C_1 \left[\log_2 (nR_{K,\mathcal{L}}(F)) - C_2\right] ,$$

where $C_1, C_2 > 0$ are independent of n and K, and $\mathcal{L} = P_0$—the class of constant functions.

Proposition 2 provides a logarithmic analogue to Proposition 1, which is useful for analyzing exponential search. The use of $\max\{2, \cdot\}$ in $\widetilde{T}(q)$ ensures that the proxy for query time is at least 1. The constants C_1 and C_2 depend on c_F and C_F, and might be disadvantageous. Still, they are independent of n and K, and do not change the asymptotic analysis. Moreover, the condition $0 < c_F \leq f, g \leq C_F < \infty$ need not be satisfied on all of $\mathcal{Q}$; it suffices that it holds over a subinterval $\mathcal{Q}' \subseteq \mathcal{Q}$. The main limitation of this variant, in contrast to Proposition 1, is the constraint $\mathcal{L} = P_0$, and the fact that this bound only holds for a specific realization $X_1, \ldots, X_n$: it does not necessarily hold for all $\{X_i\}_{i=1}^n \sim \mathbb{P}$.

4.2 Controlling the Randomness: Concentration of F_n

The lower bounds in Propositions 1 and 2 depend on $R_{K,\mathcal{L}}(F)$ and the behavior of δ_n. On the one hand, $R_{K,\mathcal{L}}(F)$ is a deterministic quantity that depends only on F, the model class $\mathcal{L}$, and the measure μ, and can be studied using approximation theory. On the other hand, we can use probabilistic tools to control the term $\|\delta_n\|_\mu$, ensuring it is small enough that the dominant source of error stems from approximating F rather than from sampling variability.

Generally, we wish to say that δ_n behaves like $1/r(n)$ for some function $r(n)$ that quantifies the rate of convergence of F_n to F. This allows us to isolate all randomness in a single term, $1/r(n)$, and express lower bounds on $\mathbb{E}_q\left[\varepsilon(q)\right]$ in terms of the deterministic approximation error for F. To make this idea precise, we rely on two types of results (all proofs are included in Appendix D): expected and worst-case concentration.

Expected concentration The main result of this type is given by Lemma 3, which we will use in conjunction with Proposition 1.

▶ **Lemma 3** (Expected bounds). *Assume $X_1, \ldots, X_n$ are i.i.d. and let μ be an arbitrary probability measure over $\mathcal{Q}$, independent from the $\{X_i\}$. Then, it holds that:*

(a) $\mathbb{E}_X\left[\|\delta_n\|_\infty\right] \leq \sqrt{\frac{\pi}{2n}}$.

(b) *As a consequence,* $\mathbb{E}_X\left[\|\delta_n\|_\mu\right] \leq \sqrt{\frac{\pi}{2n}}$.

These bounds allow us to reason about the concentration of δ_n in expectation over the random samples $\{X_i\}$, providing an average-case view of the approximation error—again, without making any assumptions on the query distribution μ.

Worst-case concentration The main result of this type is given by Lemma 4, which we use to analyze the worst-case algorithmic performance of learned indexes.

▶ **Lemma 4** (Existence of small-deviations). *Assume $X_1, \ldots, X_n$ are i.i.d. and let μ be an arbitrary probability measure over $\mathcal{Q}$, independent from the $\{X_i\}$. Then, it holds that:*

(a) *There exists a realization of $\{X_i\}_{i=1}^n$ such that $\|\delta_n\|_\mu \leq \sqrt{\frac{\pi}{2n}}$.*

(b) *If μ is equal to the data-generating distribution, i.e., $d\mu = dF$, then there exists a realization of $\{X_i\}_{i=1}^n$ such that $\|\delta_n\|_\mu \leq \frac{1}{\sqrt{6}\,n}$.*

This highlights how stronger assumptions on μ can yield sharper bounds—specifically, a $1/n$ rate instead of the general $1/\sqrt{n}$. In contrast, bounds that hold for arbitrary μ must remain more conservative to account for unfavorable scenarios: they may be loose in most cases but cannot be improved without additional assumptions. While a full characterization of how different choices of μ affect the convergence rate is outside the scope of this paper, we include the $d\mu = dF$ case as it reflects a common setting in the learned indexes literature.

4.3 From Prediction Error to Query Time

We now complete our framework by translating prediction error into actual query time. As discussed in Section 3.3, the final stage of query processing involves a local search to correct the predicted rank $h(q)$ into the true value $\mathbf{rank}(q)$. The cost of this step depends on the prediction error $\varepsilon(q) = |\mathbf{rank}(q) - h(q)|$, and varies depending on the algorithm used. We analyze the three most common search strategies: linear, exponential, and binary search.

▶ **Proposition 5** (Lower bounds for query time via prediction error). *Let $q \in \mathcal{Q}$ be a query with predicted rank $h(q)$ and actual rank $\mathbf{rank}(q)$, and let $\varepsilon(q) = |\mathbf{rank}(q) - h(q)|$ be the prediction error. Then, under the respective search strategies:*

(a) ***Linear search:*** $T(q) \geq \varepsilon(q)$.

(b) ***Exponential search:*** $T(q) \geq \log_2 (\max\{2, \varepsilon(q)\})$.

(c) ***Binary search:*** $\sup_q T(q) \geq \log_2(\sup_q \varepsilon(q))$.

Proof. We analyze each strategy in turn.

(a) **Linear search.** This method scans sequentially from the predicted position $h(q)$ until it reaches the true position $\mathbf{rank}(q)$, as illustrated in Figure 2. In practice, the search may advance by 2, 3, or any constant number of positions per step, but this only affects the total number by a constant factor and does not change asymptotic bounds. We therefore assume 1 step per comparison. The total number is thus at least $\varepsilon(q) = |\mathbf{rank}(q) - h(q)|$, so we obtain the lower bound $T(q) \geq \varepsilon(q)$.

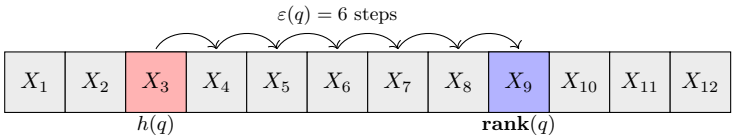


Figure 2 Linear search performs a step-by-step scan from the predicted position $h(q)$ to the true rank $\mathbf{rank}(q)$. The total cost is proportional to the prediction error $\varepsilon(q)$.

(b) **Exponential search.** Starting at $h(q)$, the algorithm expands geometrically (typically doubling the offset) until it brackets the true position $\mathbf{rank}(q)$ within a window, as visualized in Figure 3. It then performs binary search within that window. The number of expansion steps required to reach a position $\geq \mathbf{rank}(q)$ is $\lceil \log_2 \varepsilon(q) \rceil + 1 \geq \log_2 \varepsilon(q)$. The cost of the

final binary search is also logarithmic in the window size. Since query time cannot involve less than 1 operation, we write:

$$T(q) \geq \widetilde{T}(q) = \log_2 \left(\max\{2, \varepsilon(q)\}\right),$$

where $\widetilde{T}$ corresponds to the proxy function introduced earlier.

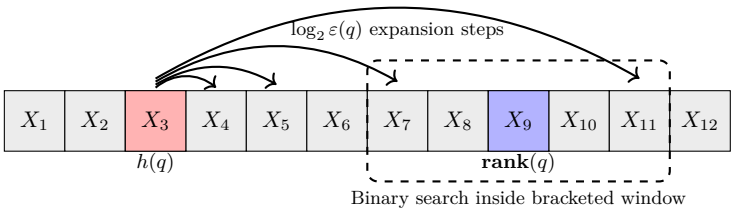


Figure 3 Exponential search expands a geometric window around the predicted position $h(q)$ until the true rank $\mathbf{rank}(q)$ is within range. The cost is logarithmic in the prediction error $\varepsilon(q)$.

(c) **Binary search.** In this case, the algorithm relies on a fixed window size that is guaranteed (by construction) to contain the true rank. The size of this window must be chosen to cover the *maximum possible prediction error* over all q, i.e., $\sup_q \varepsilon(q)$. Since binary search over a window of size M takes at least $\log_2 M$ steps in the worst case, we obtain:

$$\sup_q T(q) \geq \log_2(\sup_q \varepsilon(q)).$$

This is illustrated in Figure 4: although for this specific value of q the prediction error is smaller, the search range must still be wide enough to accommodate the worst-case error.

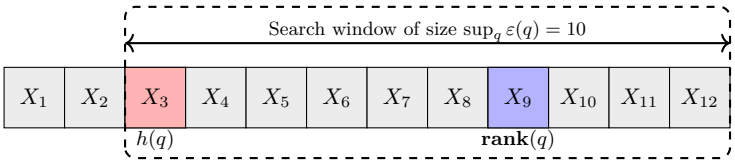


Figure 4 Binary search requires a search interval guaranteed to contain the true position $\mathbf{rank}(q)$. Therefore, the cost depends on the worst-case prediction error and is logarithmic in the interval size.

◀

4.4 Lower Bounds on Query Time

We are now ready to combine the results from the previous subsections to derive concrete lower bounds on the query time $T(q)$ under the three search strategies described in Section 4.3. We consider three natural types of lower bounds, which differ in the strength of their guarantees and the assumptions they require. In increasing order of generality, we consider:

(A) Full expectation bounds (average-case):

$$\mathbb{E}_{X,q}[T(q)] \geq C(n, K, \mathcal{L}, F, \mu).$$

This is the strongest type of result, holding in expectation over both the random dataset $\{X_i\}_{i=1}^n$ and the query value $q \sim \mu$. It typically requires strong assumptions (e.g., i.i.d. sampling) but yields average-case bounds across the entire system.

(B) Conditional expectation bounds (mixed-case):

$$\sup_{\{X_i\}_{i=1}^n} \mathbb{E}_q[T(q)] \geq C(n, K, \mathcal{L}, F, \mu).$$

This corresponds to worst-case bound over data and average-case bound over queries.

(C) Pointwise bounds (worst-case):

$$\sup_{\{X_i\}_{i=1}^n, q \in \mathcal{Q}} T(q) \geq C(n, K, \mathcal{L}, F, \mu)\,.$$

This is a purely worst-case guarantee, asserting that there exists some dataset and query for which the query time is large. It requires the fewest assumptions.

These regimes form a natural hierarchy: a type (A) bound implies one of type (B), which in turn implies one of type (C). This implication structure follows directly from the probabilistic method and reflects a trade-off between generality and the strength of assumptions. Importantly, the bounds hold uniformly over the model class $\mathcal{I}_{K,\mathcal{L}}$, i.e., they apply to every learned index structure in $\mathcal{I}_{K,\mathcal{L}}$.

Table 1 summarizes the resulting lower bounds on query time $T(q)$ under different search strategies, assumptions on the query distribution μ, and the three types of lower bounds (A)–(C) discussed above. These results show how the strongest bounds are attainable for linear search, followed by exponential search, and finally binary search, where we can only derive worst-case guarantees. Formal statements and proofs are deferred to Appendix E; most follow directly from the general framework developed earlier in this section.

Notice that the type (B) bound for exponential search is written asymptotically, without the constant C_1 from Proposition 2, in order to highlight its structural similarity to the rest.

Table 1 Lower bounds on query time $T(q)$ for various search strategies, bound types, and assumptions on μ. Type (A) bounds further assume that the $\{X_i\}$ are i.i.d. and that q is independent from the $\{X_i\}$. Type (B) bounds assume $\mathcal{L} = P_0$, that q is independent from the $\{X_i\}$, and the condition $0 < c_F \leq f, g \leq C_F < \infty$.

Bound	Search	Bound Type	Assumption	Lower Bound
L1	Linear	(A) (avg/avg)	Any μ	$n\left[R_{K,\mathcal{L}}(F) - \sqrt{\frac{\pi}{2n}}\right]$
L2	Linear	(B) (worst/avg)	$d\mu = dF$	$n\left[R_{K,\mathcal{L}}(F) - \frac{1}{\sqrt{6}\,n}\right]$
E1	Exponential	(B) (worst/avg)	Any μ	$\log_2\left(n\ R_{K,\mathcal{L}}(F)\right) - C_2$
E2	Exponential	(C) (worst/worst)	$d\mu = dF$	$\log_2\left(n\left[R_{K,\mathcal{L}}(F) - \frac{1}{\sqrt{6}\,n}\right]\right)$
B1	Binary	(C) (worst/worst)	Any μ	$\log_2\left(n\left[R_{K,\mathcal{L}}(F) - \sqrt{\frac{\pi}{2n}}\right]\right)$
B2	Binary	(C) (worst/worst)	$d\mu = dF$	$\log_2\left(n\left[R_{K,\mathcal{L}}(F) - \frac{1}{\sqrt{6}\,n}\right]\right)$

These results highlight the trade-offs between model complexity, dataset size, search strategy, and query performance in learned indexes. Note that Table 1 considers two representative regimes for the query distribution: (*i*) arbitrary μ, and (*ii*) the matched case $d\mu = dF$, where queries are drawn from the same distribution as the dataset. The latter is especially relevant in practice, as it is widely used in evaluations of learned indexes [9, 23, 21].

Altogether, this framework provides a general and flexible method for deriving query time lower bounds. In the following sections, we analyze the error term $R_{K,\mathcal{L}}(F)$ for concrete model classes commonly used in learned index structures. Section 5 focuses on $\mathcal{L} = P_0$, corresponding to piecewise constant models [32, 3], while Section 6 considers $\mathcal{L} = P_1$, corresponding to piecewise linear approximators, the most widely adopted choice in practice.

5 Lower Bounds for Piecewise Constant Models

The general lower bounds developed in Section 4 gain full significance once we derive explicit formulas for the approximation error $R_{K,\mathcal{L}}(F)$. This requires lower bounding this term for specific choices of the model class $\mathcal{L}$.

Virtually all learned index structures rely on polynomial models, as they have few parameters to store, and are fast to train and evaluate. Accordingly, we focus on model classes of the form $\mathcal{L} = P_m$, where P_m denotes the class of polynomials of degree at most m. In this section, we analyze the case $\mathcal{L} = P_0$, which corresponds to piecewise constant models. Next, in Section 6, we address the most widely used case in practice: $\mathcal{L} = P_1$, corresponding to piecewise linear models.

The key idea in our analysis is that, as we show below, for piecewise constant models the problem of approximating the CDF F reduces to finding the optimal quantization of a random variable distributed over $[0, 1]$. This allows us to borrow fundamental results from quantization theory to characterize the approximation error $R_{K,P_0}(F)$.

Our results in this section are derived for a broad class of distributions F and μ—specifically, the case where both are characterized by density functions with respect to the Lebesgue measure. We denote these densities f and g, respectively, and consider both the case $f = g$ (Section 5.1) and the general case (relegated to Appendix F). In both cases, we show that the approximation error decreases linearly with K; the difference is that for $f = g$ this can be shown for any K, whereas for $f \neq g$ the result holds asymptotically as $K \to \infty$.

More specifically, throughout Sections 5 and 6, we focus on the case where F is Lipschitz continuous. This regularity assumption is relatively mild, and it encompasses virtually all common continuous distributions (e.g., Gaussian, exponential, uniform). At the same time, it guarantees the existence of a density f, and it ensures that lower bounds reflect the limitations of the model class (as governed by K and $\mathcal{L}$), rather than artifacts of highly pathological target functions.

5.1 Same dataset and query distributions $f = g$

For the relatively simple case of piecewise constant approximating functions $h \in \mathcal{I}_{K,P_0}$ and $f = g$, the minimum approximation error $R_{K,P_0}(F)$ can be calculated exactly for any K. We achieve this by first showing that this problem setup reduces to optimal quantization of a uniformly distributed random variable, and then relying on a well-known result in quantization theory to calculate the error of the associated optimal quantizer.

Let $h \in \mathcal{I}_{K,P_0}$. Since F is a CDF with range $[0, 1]$, without loss of generality we can assume $0 \leq h(q) \leq 1$ for all q. Therefore, h can be written as

$$h(q) = \textstyle\sum_{k=1}^{K} c_k \mathbf{1}_{I_k}(q)\,,$$

where the K intervals $\{I_k\}$ partition $\mathcal{Q}$, and $c_k \in [0, 1]$ are the corresponding approximating constants. This means that when $f = g$, the absolute error $\|F - h\|_\mu$ can be written as

$$\|F - h\|_\mu = \int_{\mathcal{Q}} |F(q) - h(q)|\, f(q)dq = \sum_{k=1}^{K} \int_{I_k} |F(q) - c_k|\, f(q)dq = \sum_{k=1}^{K} \int_{J_k} |u - c_k|\, du\,, \quad (4)$$

where $J_k = F(I_k) = \{F(q) : q \in I_k\}$ (which is a continuous interval), and the last equality comes from substituting $u = F(q)$ and $du = f(q)dq$ inside the integrals. Since F is a CDF we know that $\{J_k\}$ is a partition of $[0, 1]$. If we now define $Q_K(u) = \sum_{k=1}^{K} c_k \mathbf{1}_{J_k}(u)$, then

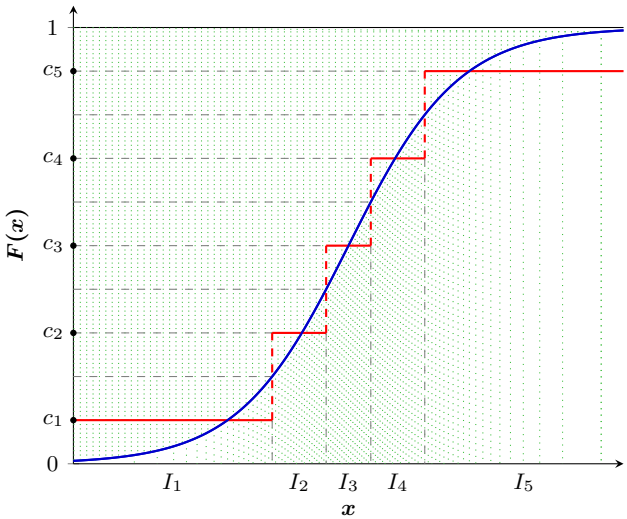


Figure 5 Optimal approximation of the CDF F (blue) via a piecewise constant function h (red) with $K = 5$ segments when $f = g$—i.e., when the dataset and queries are drawn from the same distribution. In this case, the optimal approximating values $\{c_k\}$ are evenly spaced (marked on the y-axis), and approximating the CDF is equivalent to quantizing Uniform($[0, 1]$) along the y-axis. The quantization boundaries (unmarked black horizontal dashed lines) are reflected onto the x-axis via F^{-1} to split $\mathcal{Q}$ into $K = 5$ equiprobable regions $\{I_k\}$ along the quintiles of F (black vertical dashed lines). The green dotted lines illustrate how the distribution of $X \sim F$ becomes uniform when transformed via $F(X)$.

(4) can be expressed as

$$\|F - h\|_\mu = \mathbb{E}\big[|U - Q_K(U)|\big] \,, \tag{5}$$

where Q_K is a piecewise constant function over $[0, 1]$ and the expectation is taken over a Uniform($[0, 1]$) random variable U. Note that (4)–(5) reflect the fact that $F(X)$ is uniformly distributed when $X \sim F$.

In other words, minimizing $\|F - h\|_\mu$ is equivalent to choosing a set of intervals $\{J_k\}$ and the associated points $\{c_k\}$ so as to minimize the expected distance between a realization of U and its corresponding value c_k. We recognize this as the problem of optimal quantization of U under the absolute error.

It is a well-established result in quantization theory that uniform quantization is optimal for Uniform($[0, 1]$). Specifically, the quantization points should be equally spaced as $\{c_k = \frac{2k-1}{2K} : k = 1, \ldots, K\}$, and the quantization intervals $\{J_k\}$ should split the domain $[0, 1]$ evenly. The associated minimum quantization error is equal to [13, Example 5.5]

$$T_{K,r}(U) = \inf_{Q_K} \{\mathbb{E}\left[|U - Q_K(U)|^r\right] : Q_K \text{ has } K \text{ quantization points}\} = \frac{1}{K^r(1+r)2^r} \,;$$

this means that, in our notation,

$$R_{K,P_0}(F) = T_{K,1}(U) = \frac{1}{4K} \,. \tag{6}$$

It is interesting to note that the minimum approximation error (6) and the optimal $\{c_k\}$ and $\{J_k\}$ are the same for any F; the only dependence on F is via the regions $\{I_k\}$ that should split $\mathcal{Q}$ at the K-quantiles of F to yield equal-sized $\{J_k\}$.

We illustrate this setting graphically in Figure 5, where we show an example CDF curve $F(x)$ (blue), optimally approximated by a piecewise constant function h (red) with $K = 5$

segments. The approximation points $\{c_k\}$ are evenly spaced along the y-axis, and the mid-points between them, when reflected onto the x-axis via F^{-1}, mark the quintiles of F, which split $\mathcal{Q}$ into $K = 5$ equiprobable regions $\{I_k\}$. The "uniformization" of a distribution by its own CDF that we used in (4)–(5) is illustrated by the green dotted lines—their density below the blue CDF curve is proportional to f, whereas above the CDF curve, when reflected off it onto the y-axis, it becomes uniform.

6 Lower Bounds for Piecewise Linear Models

As shown in Section 5, learned indexes based on piecewise constant approximations admit a clean lower bound analysis via quantization theory. There, the approximation error $R_{K,\mathcal{L}}(F)$ was analyzed for individual (Lipschitz continuous) CDFs F. In this section, we extend our analysis to more expressive model classes, such as piecewise linear functions. Our central object of study will be the worst-case approximation error over a function class $\mathcal{F}$:

$$\sup_{F\in\mathcal{F}} R_{K,\mathcal{L}}(F)\,.$$

By lower bounding this quantity, we can show that there exist functions within $\mathcal{F}$ which are not well approximated by any $h \in \mathcal{I}_{K,\mathcal{L}}$. In line with the rest of the paper, we focus on the class of Lipschitz continuous CDFs, and show that for $\mathcal{L} = P_1$, the lower bound $\Omega(1/K)$ still holds in this worst-case sense, matching classical upper bounds (see Appendix G.1).

We establish this result through two complementary approaches. In Section 6.1, we describe an explicit family of Lipschitz continuous CDFs $\{F_M\}$ for which $R_{K,P_1}(F) = \Omega(1/K)$ holds. This shows that the class $\mathcal{I}_{K,P_1}$ cannot guarantee $o(1/K)$ approximation error uniformly across all Lipschitz continuous functions, and the construction works for a broad class of measures μ.

In Section 6.2, we approach the problem through *Kolmogorov widths*, an important tool from approximation theory. This approach typically assumes μ as the Lebesgue measure, and only allows for existential (rather than constructive) proofs. Nonetheless, it is a very flexible framework: it allows us to analyze broader model classes beyond P_1 and to derive sharper bounds for more regular function spaces. Together, these two methods highlight complementary aspects of the approximation limits faced by learned index structures.

6.1 An Adversarial CDF Construction

To illustrate the limitations of piecewise linear approximators, we construct a family of CDFs $\{F_M\}_{M\in\mathbb{N}}$ with the following property: for any $M > (1+\alpha)K$ with $\alpha > 0$, the approximation error satisfies $R_{K,P_1}(F_M) = \Omega(1/K)$. The construction, detailed in Appendix G.2, is based on stitching together rescaled copies of the quadratic function x^2 over M disjoint intervals. While this is presented under the assumption that μ is the uniform distribution, the argument naturally extends to a broader class: specifically, any measure μ with a density g satisfying $g(q) \geq \lambda > 0$ over an interval. In such cases, the lower bound continues to hold up to constants depending on λ.

Thus, this construction provides a worst-case lower bound for $R_{K,P_1}(F)$ within the class of Lipschitz continuous CDFs. Importantly, it also highlights a fundamental aspect of learned index structures: the approximation error depends on both the location and the number of difficult-to-approximate regions. Hence, the choice of partition and local model must be adapted to the data. Fixing the model class and budget K in advance cannot guarantee uniformly good approximation.

6.2 Kolmogorov Widths and Approximation Lower Bounds

To further understand the limitations of learned index models beyond explicit constructions, we turn to a more general framework from approximation theory: Kolmogorov widths. This concept provides a way to quantify how well an entire function class $\mathcal{F}$ can be approximated by models of bounded complexity. Formally, the Kolmogorov width $d_\kappa(\mathcal{F})$ captures the best possible worst-case approximation error over $\mathcal{F}$:

$$d_\kappa(\mathcal{F}) = \inf_{H \in \mathcal{H}_\kappa} \sup_{F \in \mathcal{F}} \inf_{h \in H} \|F - h\|_\mu \,,$$

where the outermost infimum is taken over the set $\mathcal{H}_\kappa$ of all linear subspaces of dimension κ.

We relate $d_\kappa(\mathcal{F})$ to our context in the following way. Throughout, we have considered approximation by functions in $\mathcal{I}_{K,\mathcal{L}}$. This class is not necessarily a linear subspace of $L^1(\mu)$, since the sum of two K-piecewise functions is not necessarily K-piecewise. Nonetheless, if we fix a partition $\mathcal{P} = \{I_k\}_{k=1}^K$ of $\mathcal{Q}$, the subclass

$$\mathcal{I}_{\mathcal{P},\mathcal{L}} = \left\{ h : h(q) = \textstyle\sum_k h_k(q) \mathbf{1}_{I_k}(q),\ h_k \in \mathcal{L} \right\}$$

does form a linear subspace, provided that $\mathcal{L}$ is itself a linear subspace of $L^1(\mu)$. In such cases—e.g., when $\mathcal{L}$ consists of polynomials of bounded degree—any function class $\mathcal{I}_{\mathcal{P},\mathcal{L}}$ has dimension at most $\kappa = K \cdot \dim(\mathcal{L})$. This characterization leads to the following inequality.

▶ **Lemma 6.** *For any $\mathcal{F}, \mathcal{L} \subseteq L^1(\mu)$ where $\mathcal{L}$ is a linear subspace of finite dimension, it holds*

$$\sup_{F \in \mathcal{F}} R_{K,\mathcal{L}}(F) \geq d_\kappa(\mathcal{F}), \qquad \textit{where } \kappa = K \cdot \dim(\mathcal{L}) \,.$$

The proof can be found in Appendix G. We now focus on the case where $\mathcal{F}$ is the class of Lipschitz continuous functions, denoted by $W^{1,\infty}$. This notation comes from Sobolev space theory, which provides a convenient formalism for characterizing function smoothness, and within which most Kolmogorov width results are typically stated. When μ is the uniform distribution over a compact interval $[a, b]$, it holds that (see [22, Chapter 14, Theorem 3.8]):

$$d_\kappa(W^{1,\infty}) = \Omega(1/\kappa) \,. \tag{7}$$

Notice that for $\mathcal{L} = P_m$, the class of degree-m polynomials, we have $\kappa = (m+1)K$, since each piece has $m+1$ parameters. Piecewise linear models correspond to $m = 1$ and thus $\kappa = 2K$. Hence, for fixed m, Lemma 6 and equation (7) imply $\sup_{F \in W^{1,\infty}} R_{K,P_m}(F) = \Omega(1/K)$.

This has a direct implication for learned index structures: even the best possible model using K adaptive segments and bounded-degree polynomial predictors cannot achieve approximation error $o(1/K)$ for general Lipschitz continuous CDFs.[2]

These results match the $\Omega(1/K)$ result obtained in Section 6.1. While the explicit construction from Section 6.1 can accommodate a wide class of distributions μ, including non-uniform densities, the Kolmogorov width approach offers a wider perspective in regards to the approximating functions. In particular, it applies to any piecewise polynomial class $\mathcal{I}_{K,P_m}$ and, more generally, to any linear span of κ functions, such as truncated Fourier series.

[2] This conclusion requires some care. While the Kolmogorov width bound is stated for the worst-case over the class $W^{1,\infty}$, it is not immediate that the worst-case functions realizing the $\Omega(1/\kappa)$ bound include any CDFs. In principle, one might worry that these worst-case functions do not include any valid CDFs. In Appendix G.3, we provide a rigorous argument showing that any hard-to-approximate function $F \in W^{1,\infty}$ can be transformed into a CDF that remains equally hard to approximate.

The Kolmogorov width framework makes it explicit that the $\Omega(1/K)$ bound is inherently a worst-case result. For smoother subclasses $\mathcal{F} \subseteq W^{1,\infty}$, the theory yields sharper rates: for example, for $W^{2,\infty}$ (i.e., functions with Lipschitz continuous derivatives), one obtains $d_\kappa(W^{2,\infty}) = \Theta(1/\kappa^2)$. This highlights how stronger regularity assumptions can enable faster approximation. In contrast, the lower bound for piecewise constant models (Section 5) applies uniformly to all Lipschitz continuous CDFs and does not improve with additional regularity.

Concluding Remarks. A key takeaway from this section is that for the class $\mathcal{F} = W^{1,\infty}$ of Lipschitz continuous functions, the worst-case approximation error under learned index models in the class $\mathcal{I}_{K,P_m}$ satisfies $\Omega(1/K)$ for any fixed m. This resembles the $\Omega(1/K)$ rate derived in Section 5 for piecewise constant approximation, but here it holds only in a worst-case sense over $\mathcal{F}$, not uniformly for all $F \in \mathcal{F}$. In particular, the result does not extend to smoother subclasses such as $W^{2,\infty}$, where approximation can be better, but it still constitutes a meaningful lower bound for the general performance of learned indexes.

7 Discussion and Conclusions

This paper establishes an analytical framework for lower bounding query time using learned index structures (Sections 3–4). We consider two classes of approximating functions—piecewise constant and piecewise linear approximators—that are most commonly used in practice, and relate query time to the error of approximating the CDF of the data-generating distribution. For piecewise constant functions, we show that the approximation error decreases linearly with the number of segments K for a broad class of CDFs (Section 5). For piecewise linear approximators, the situation is more subtle: we show that the approximation error depends on the regularity of the CDF, and that in general using a richer approximating model does not guarantee an improved lower bound on the approximation error (Section 6). However, under additional assumptions on the smoothness of the data-generating CDF, the bound on the approximation error can be lowered further.

More specifically, Sections 5 and 6 establish a lower bound of $\Omega(1/K)$ on approximation error that, in general, cannot be improved for the class of Lipschitz continuous CDFs. For piecewise constant models, this holds uniformly (Section 5); for piecewise linear models, the bound applies in a worst-case sense over the choice of CDF (Section 6). In either case, the $\Omega(1/K)$ behavior governs the worst-case regime. Combining these results with the analysis summarized in Table 1 (e.g., see rows E2 and B2) we obtain an $\Omega(\log(n/K))$ lower bound on worst-case query time. As discussed in Section 2, this implies that the number of segments K (and therefore, space usage) must be nearly linear in n to yield an asymptotic advantage over traditional methods, a conclusion that holds for a broad class of learned indexes.

It is instructive to compare our bounds with the information-theoretic results in [34]. Their worst-case bounds require a finite query domain $\mathcal{Q}$, a restriction not present in our framework. Moreover, our analysis yields sharper and more applicable insights for typical learned index designs, particularly those based on piecewise constant or linear models. In the average-case setting, their results imply that achieving expected error η requires $\log(1/\eta)$ bits of model capacity. In contrast, our bounds—under independence assumptions—show a $1/\eta$ scaling, which is larger when $\eta \to 0$. Additionally, when the query and data distributions are the same—i.e., $d\mu = dF$—we show that the number of parameters can scale as $\Omega(n)$, compared to $\Omega(\sqrt{n})$ in their framework.

Aside from their theoretical importance, our results offer some practical insights for the design of learned indexes: first, they show that the number of segments K should not be

chosen without regard to the underlying data, as there exist "adversarial" distributions for which the approximation error can be poor for a given K (see Section 6.1). Second, the results of Section 6.2 show that approximation capacity depends not only on the expressiveness of the model class but also on the smoothness of the CDF. This suggests that selecting the model class in a data-dependent manner—adapted to the regularity of the CDF—could lead to more efficient index designs, an idea that warrants further exploration.

Several open directions remain: extending average-case analysis to binary search; relaxing independence assumptions on the data; considering a wider class of index structures; and analyzing learned index performance under updates or dynamic workloads. We hope this work offers some of the foundational tools needed to explore these ideas further.

References

1. Alfred V. Aho and John E. Hopcroft. *The Design and Analysis of Computer Algorithms*. Addison-Wesley Longman Publishing Co., Inc., USA, 1st edition, 1974.
2. Albert Atserias, Martin Grohe, and Dániel Marx. Size bounds and query plans for relational joins. *SIAM Journal on Computing*, 42(4):1737–1767, 2013. `arXiv:https://doi.org/10.1137/110859440`, `doi:10.1137/110859440`.
3. Luis Alberto Croquevielle, Guang Yang, Liang Liang, Ali Hadian, and Thomas Heinis. Beyond Logarithmic Bounds: Querying in Constant Expected Time with Learned Indexes. In Sudeepa Roy and Ahmet Kara, editors, *28th International Conference on Database Theory (ICDT 2025)*, volume 328 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 19:1–19:21, Dagstuhl, Germany, 2025. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. URL: `https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICDT.2025.19`, `doi:10.4230/LIPIcs.ICDT.2025.19`.
4. Sándor CSöRgő and Julian J. Faraway. The exact and asymptotic distributions of cramér-von mises statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):221–234, 12 2018. `arXiv:https://academic.oup.com/jrsssb/article-pdf/58/1/221/49098785/jrsssb\_58\_1\_221.pdf`, `doi:10.1111/j.2517-6161.1996.tb02077.x`.
5. Jialin Ding, Umar Farooq Minhas, Jia Yu, Chi Wang, Jaeyoung Do, Yinan Li, Hantian Zhang, Badrish Chandramouli, Johannes Gehrke, Donald Kossmann, David Lomet, and Tim Kraska. Alex: An updatable adaptive learned index. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 969–984, New York, NY, USA, 2020. Association for Computing Machinery. `doi:10.1145/3318464.3389711`.
6. Jialin Ding, Vikram Nathan, Mohammad Alizadeh, and Tim Kraska. Tsunami: a learned multi-dimensional index for correlated data and skewed workloads. *Proc. VLDB Endow.*, 14(2):74–86, October 2020. `doi:10.14778/3425879.3425880`.
7. Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, Providence, R.I., 2010.
8. Paolo Ferragina, Fabrizio Lillo, and Giorgio Vinciguerra. Why are learned indexes so effective? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3123–3132, Virtual, 13–18 Jul 2020. PMLR. URL: `http://proceedings.mlr.press/v119/ferragina20a.html`.
9. Paolo Ferragina and Giorgio Vinciguerra. The pgm-index: A fully-dynamic compressed learned index with provable worst-case bounds. *Proc. VLDB Endow.*, 13(8):1162–1175, apr 2020. `doi:10.14778/3389133.3389135`.
10. R. A. Finkel and J. L. Bentley. Quad trees A data structure for retrieval on composite keys. *Acta Inf.*, 4(1):1–9, March 1974. `doi:10.1007/BF00288933`.
11. Alex Galakatos, Michael Markovitch, Carsten Binnig, Rodrigo Fonseca, and Tim Kraska. Fiting-tree: A data-aware index structure. In *Proceedings of the 2019 International Conference*

on Management of Data, SIGMOD '19, page 1189–1206, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3299869.3319860.

12 Goetz Graefe et al. Modern B-tree techniques. *Foundations and Trends in Databases*, 3(4):203–402, 2011. doi:10.1561/1900000028.

13 Siegfried Graf and Harald Luschgy. *Foundations of quantization for probability distributions*. Springer Science & Business Media, 2000.

14 Antonin Guttman. R-trees: a dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data*, SIGMOD '84, page 47–57, New York, NY, USA, 1984. Association for Computing Machinery. doi:10.1145/602259.602266.

15 Ali Hadian and Thomas Heinis. Shift-table: A low-latency learned index for range queries using model correction. In Yannis Velegrakis, Demetris Zeinalipour-Yazti, Panos K. Chrysanthis, and Francesco Guerra, editors, *Proceedings of the 24th International Conference on Extending Database Technology, EDBT 2021, Nicosia, Cyprus, March 23 - 26, 2021*, pages 253–264. OpenProceedings.org, 2021. doi:10.5441/002/edbt.2021.23.

16 Andreas Kipf, Ryan Marcus, Alexander van Renen, Mihail Stoian, Alfons Kemper, Tim Kraska, and Thomas Neumann. Radixspline: A single-pass learned index. In *Proceedings of the Third International Workshop on Exploiting Artificial Intelligence Techniques for Data Management*, aiDM '20, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3401071.3401659.

17 Paraschos Koutris, Shaleen Deep, Austen Fan, and Hangdong Zhao. The Quest for Faster Join Algorithms. In Sudeepa Roy and Ahmet Kara, editors, *28th International Conference on Database Theory (ICDT 2025)*, volume 328 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 1:1–1:12, Dagstuhl, Germany, 2025. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICDT.2025.1, doi:10.4230/LIPIcs.ICDT.2025.1.

18 Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. The case for learned index structures. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, page 489–504, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3183713.3196909.

19 Viktor Leis, Andrey Gubichev, Atanas Mirchev, Peter Boncz, Alfons Kemper, and Thomas Neumann. How good are query optimizers, really? *Proc. VLDB Endow.*, 9(3):204–215, November 2015. doi:10.14778/2850583.2850594.

20 Pengfei Li, Hua Lu, Qian Zheng, Long Yang, and Gang Pan. Lisa: A learned index structure for spatial data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 2119–2133, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3318464.3389703.

21 Liang Liang, Guang Yang, Ali Hadian, Luis Alberto Croquevielle, and Thomas Heinis. Swix: A memory-efficient sliding window learned index. *Proc. ACM Manag. Data*, 2(1), March 2024. doi:10.1145/3639296.

22 G. Lorentz, George and Manfred V. Golitschek. *Constructive approximation : advanced problems*. Springer-Verlag, Berlin, 1996.

23 Ryan Marcus, Andreas Kipf, Alexander van Renen, Mihail Stoian, Sanchit Misra, Alfons Kemper, Thomas Neumann, and Tim Kraska. Benchmarking learned indexes. *Proc. VLDB Endow.*, 14(1):1–13, sep 2020. doi:10.14778/3421424.3421425.

24 P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269 – 1283, 1990. doi:10.1214/aop/1176990746.

25 Vikram Nathan, Jialin Ding, Mohammad Alizadeh, and Tim Kraska. Learning multi-dimensional indexes. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 985–1000, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3318464.3380579.

26 Theodore J Rivlin. *An introduction to the approximation of functions*. Dover, New York, 1981.

27 Dan Suciu. Applications of information inequalities to database theory problems. In *38th Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2023, Boston, MA, USA, June 26-29, 2023*, pages 1–30. IEEE, 2023. doi:10.1109/LICS56636.2023.10175769.

28 Zhaoyan Sun, Xuanhe Zhou, and Guoliang Li. Learned index: A comprehensive experimental evaluation. *Proc. VLDB Endow.*, 16(8):1992–2004, April 2023. doi:10.14778/3594512.3594528.

29 Jiacheng Wu, Yong Zhang, Shimin Chen, Jin Wang, Yu Chen, and Chunxiao Xing. Updatable learned index with precise positions. *Proc. VLDB Endow.*, 14(8):1276–1288, April 2021. doi:10.14778/3457390.3457393.

30 Guang Yang, Liang Liang, Ali Hadian, and Thomas Heinis. FLIRT: A fast learned index for rolling time frames. In Julia Stoyanovich, Jens Teubner, Nikos Mamoulis, Evaggelia Pitoura, Jan Mühlig, Katja Hose, Sourav S. Bhowmick, and Matteo Lissandrini, editors, *Proceedings 26th International Conference on Extending Database Technology, EDBT 2023, Ioannina, Greece, March 28-31, 2023*, pages 234–246, Ioannina, Greece, 2023. OpenProceedings.org. doi:10.48786/edbt.2023.19.

31 Mihalis Yannakakis. Algorithms for acyclic database schemes. In *Proceedings of the Seventh International Conference on Very Large Data Bases - Volume 7*, VLDB '81, page 82–94. VLDB Endowment, 1981.

32 Sepanta Zeighami and Cyrus Shahabi. On distribution dependent sub-logarithmic query time of learned indexing. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

33 Sepanta Zeighami and Cyrus Shahabi. Theoretical analysis of learned database operations under distribution shift through distribution learnability. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

34 Sepanta Zeighami and Cyrus Shahabi. Towards establishing guaranteed error for learned database operations. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL: https://openreview.net/forum?id=6tqgL8VluV.

A Proof of Scaling Lemma and Proposition 1

▶ **Lemma 7** (Scaling of Approximation Error). *Let $\lambda > 0$, and suppose the model class $\mathcal{L}$ is invariant under scalar multiplication. Then for any function $G \in L^1(\mu)$ and any $K \in \mathbb{N}$,*

$$R_{K,\mathcal{L}}(\lambda G) = \lambda \cdot R_{K,\mathcal{L}}(G)\,,$$

where $R_{K,\mathcal{L}}(G) = \inf_{h \in \mathcal{I}_{K,\mathcal{L}}} \|G - h\|_\mu$.

Proof. Let $\mathcal{H} = \mathcal{I}_{K,\mathcal{L}}$ denote the class of piecewise models defined over K intervals and using functions from $\mathcal{L}$. Since $\mathcal{L}$ is invariant under scalar multiplication, so is $\mathcal{H}$. That is, for any $h \in \mathcal{H}$ and any $\lambda > 0$, the function λh also belongs to $\mathcal{H}$. Now observe:

$$R_{K,\mathcal{L}}(\lambda G) = \inf_{h \in \mathcal{H}} \|\lambda G - h\|_\mu = \inf_{h \in \mathcal{H}} \|\lambda(G - \lambda^{-1} h)\|_\mu = \inf_{h \in \mathcal{H}} \lambda \cdot \|(G - \lambda^{-1} h)\|_\mu\,.$$

Using the positive homogeneity of the infimum:

$$R_{K,\mathcal{L}}(\lambda G) = \lambda \cdot \inf_{h \in \mathcal{H}} \|G - \lambda^{-1} h\|_\mu\,.$$

Since $\mathcal{H}$ is invariant under scalar multiplication, it follows that for any $h \in \mathcal{H}$, we have $\lambda^{-1} h \in \mathcal{H}$. Moreover, any $h \in \mathcal{H}$ can be written as $\lambda^{-1}\tilde{h}$ where $\tilde{h} \in \mathcal{H}$. This implies that the map $h \mapsto \lambda^{-1} h$ is a bijection on $\mathcal{H}$, and thus

$$R_{K,\mathcal{L}}(\lambda G) = \lambda \cdot \inf_{\tilde{h} \in \mathcal{H}} \|G - \tilde{h}\|_\mu = \lambda \cdot R_{K,\mathcal{L}}(G)\,.$$

◀

Proof of Proposition 1. Applying the triangle inequality in $L^1(\mu)$, we have for any $h \in \mathcal{I}_{K,\mathcal{L}}$:

$$\|F - h\|_\mu \leq \|F - F_n\|_\mu + \|F_n - h\|_\mu \quad \Longrightarrow \quad \|F_n - h\|_\mu \geq \|F - h\|_\mu - \|\delta_n\|_\mu \,.$$

Taking the infimum over $h \in \mathcal{I}_{K,\mathcal{L}}$ at the last inequality gives:

$$R_{K,\mathcal{L}}(F_n) \geq R_{K,\mathcal{L}}(F) - \|\delta_n\|_\mu \,.$$

Then, by the scaling identity $R_{K,\mathcal{L}}(nF_n) = nR_{K,\mathcal{L}}(F_n)$ (Lemma 7) and the definition of expected prediction error $\mathbb{E}_q\big[\varepsilon(q)\big] = \|nF_n - h\|_\mu \geq R_{K,\mathcal{L}}(nF_n)$, the result follows. ◀

B Proof of Proposition 2

The focus of this section is to prove Proposition 2. For clarity, some auxiliary results used in this section are proven in Appendix C. We start by recalling the definition of $\widetilde{T}(q)$:

$$\widetilde{T}(q) = \log_2\left(\max\{2, \varepsilon(q)\}\right) = \log_2\left(\max\{2, |nF_n(q) - h(q)|\}\right) \,.$$

Ultimately, we are interested in lower bounding $\mathbb{E}_q\left[\widetilde{T}(q)\right]$. To this end, recall first that (by hypothesis) F and μ have respective densities f and g which satisfy

$$0 < c_F \leq f(q), g(q) \leq C_F < \infty \tag{8}$$

for constants c_F, C_F, and for all $q \in \mathcal{Q}$. Proposition 2 (and this proof) can be generalized straightforwardly to the case where this property is only satisfied in a continuous interval $\mathcal{Q}' \subseteq \mathcal{Q}$ with positive measure. Nonetheless, for simplicity of exposition, we assume (8) holds for all $q \in \mathcal{Q}$.

Note that without loss of generality, we may assume $d\mu = dF = f\,dq$. This follows from:

$$\mathbb{E}_q\left[\widetilde{T}(q)\right] = \int_{\mathcal{Q}} \widetilde{T}(q) d\mu = \int_{\mathcal{Q}} \widetilde{T}(q)\, g(q) dq \geq \frac{c_F}{C_F} \int_{\mathcal{Q}} \widetilde{T}(q)\, f(q) dq \,,$$

where we have used $d\mu = g\,dq$ and the inequalities expressed in (8). Hence, if we can lower bound the last integral $\int_{\mathcal{Q}} \widetilde{T}(q)\, f(q) dq$ as stated in Proposition 2, the bound will also apply to $\mathbb{E}_q\left[\widetilde{T}(q)\right]$ with an extra (constant) factor of $\frac{c_F}{C_F}$. Hence, throughout this section we write dF in place of $d\mu$.

Finally, we assume $nR_{K,P_0} \geq 1$. Otherwise, Proposition 2 holds trivially since its right hand side is negative. With these assumptions in place, we can proceed to the proof.

Proof of Proposition 2. We divide the proof into four logical steps.

Step 1: Piecewise expansion. Any $h \in \mathcal{I}_{K,P_0}$ can be expanded as

$$h(q) = n \sum_{k=1}^{K} c_k \mathbf{1}_{I_k}(q),$$

where the $\{c_k\}$ are constants and the $\{I_k\}$ are intervals which partition $\mathcal{Q}$. Now, the quantity of interest can be lower bounded as:

$$\begin{aligned}
\mathbb{E}_q\big[\widetilde{T}(q)\big] &= \int_{\mathcal{Q}} \log_2\left(\max\{2, \varepsilon(q)\}\right) dF \\
&= \int_{\mathcal{Q}} \log_2\left(\max\left\{2, |nF_n(q) - h(q)|\right\}\right) dF \\
&\geq \int_{\mathcal{Q}} \log_2\left(\max\left\{2, n|F(q) - n^{-1}h(q)| - n|\delta_n(q)|\right\}\right) dF \\
&= \sum_{k=1}^{K} \underbrace{\int_{I_k} \log_2\left(\max\left\{2, n|F(q) - c_k| - n|\delta_n(q)|\right\}\right) dF}_{E_k}\,, \qquad (9)
\end{aligned}$$

where we have used the triangle inequality to lower bound $\varepsilon(q)$:

$$\begin{aligned}
\varepsilon(q) &= |nF_n(q) - h(q)| \\
&= |nF_n(q) - nF(q) + nF(q) - h(q)| \\
&\geq n|F(q) - n^{-1}h(q)| - n|F_n(q) - F(q)| \\
&= n|F(q) - n^{-1}h(q)| - n|\delta_n(q)|\,.
\end{aligned}$$

Now, without loss of generality, we can assume that h intersects with F in all intervals, that is, for each k there exists $q_k \in I_k$ such that $F(q_k) = c_k$. Otherwise, there would exist a different $h' \in \mathcal{I}_{K,P_0}$ such that $|F(q) - n^{-1}h'(q)| \leq |F(q) - n^{-1}h(q)|$ for all $q \in \mathcal{Q}$, meaning

$$\begin{aligned}
\mathbb{E}_q\big[\widetilde{T}(q)\big] &\geq \int_{\mathcal{Q}} \log_2\left(\max\left\{2, n|F(q) - n^{-1}h(q)| - n|\delta_n(q)|\right\}\right) dF \\
&\geq \int_{\mathcal{Q}} \log_2\left(\max\left\{2, n|F(q) - n^{-1}h'(q)| - n|\delta_n(q)|\right\}\right) dF\,,
\end{aligned}$$

so that any lower bound we derive for the case of h' would also apply to h. Hence, we assume the existence of $\{q_k\}_{k=1}^{K}$ with the above property.

From these derivations, it is clear that it is sufficient for our proof to lower bound the expression $\sum_{k=1}^{K} E_k$ as defined in (9), where each c_k is assumed to intersect F in at least one point $q_k \in I_k$. In the next part of the proof, we focus on lower bounding each individual term E_k, given by

$$E_k = \int_{I_k} \log_2\left(\max\left\{2, n|F(q) - c_k| - n|\delta_n(q)|\right\}\right) dF\,. \qquad (10)$$

To do this, we first define a setting in which the term $n|\delta_n(q)|$ is negligible, and then derive some concentration results for the remaining term $n|F(q) - c_k|$. Finally, in the last part of the proof, we turn to the aggregate error $\sum_{k=1}^{K} E_k$.

Step 2: A large-measure set where δ_n is small. Now, define the set

$$A = \left\{q \in \mathcal{Q} \text{ such that } |\delta_n(q)| \leq \frac{1}{n}\right\}.$$

Since $d\mu = dF$, by Lemma 4(b) there exists a realization of $X_1, \ldots, X_n$ such that $\|\delta_n\|_\mu \leq \frac{1}{\sqrt{6}\,n}$. We work under this realization for the rest of the proof. By Markov's inequality:

$$\mu\left(|\delta_n| > \frac{1}{n}\right) \leq \frac{\|\delta_n\|_\mu}{\frac{1}{n}} \leq \frac{n}{\sqrt{6}\,n} \leq \frac{1}{2}\,,$$

meaning that $\mu(A) \geq 1/2$. Now, for each $k = 1, \ldots, K$ define the set $J_k = I_k \cap A$, that is, the subset of interval I_k where $|\delta_n| \leq \frac{1}{n}$ holds. Starting from (10), we have

$$\begin{aligned} E_k &= \int_{I_k} \log_2 \left(\max\left\{2, n|F(q) - c_k| - n|\delta_n(q)|\right\}\right) dF \\ &\geq \int_{J_k} \log_2 \left(\max\left\{2, n|F(q) - c_k| - n|\delta_n(q)|\right\}\right) dF \\ &\geq \frac{1}{2} \int_{J_k} \log_2 \left(2 + n|F(q) - c_k| - n|\delta_n(q)|\right) dF \\ &\geq \frac{1}{2} \int_{J_k} \log_2 \left(1 + n|F(q) - c_k|\right) dF \,, \end{aligned} \tag{11}$$

where we have used the fact that $|\delta_n| \leq \frac{1}{n}$ over each set J_k.

Step 3: A concentration result for the approximation error. We now focus on lower bounding expression (11). For this, we define the first and second moments of $|F(q) - c_k|$ over the set J_k:

$$m_k = \int_{J_k} |F(q) - c_k| \, dF \,, \qquad s_k = \int_{J_k} |F(q) - c_k|^2 dF \,.$$

Next, denote by J'_k the set $J'_k = \left\{q \in J_k \text{ such that } |F(q) - c_k| \geq \frac{m_k}{2\mu(J_k)}\right\}$, that is, the subset of J_k where the quantity $|F(q) - c_k|$ is not much smaller than its normalized average over J_k. Since $\log_2 (1 + n|F(q) - c_k|)$ is never negative, it holds that

$$\begin{aligned} E_k &\geq \frac{1}{2} \int_{J_k} \log_2 \left(1 + n|F(q) - c_k|\right) dF \\ &\geq \frac{1}{2} \int_{J'_k} \log_2 \left(1 + n|F(q) - c_k|\right) dF \\ &\geq \frac{\mu(J'_k)}{2} \log_2 \left(1 + \frac{n m_k}{2\mu(J_k)}\right) \\ &\geq \frac{m_k^2}{8\, s_k} \cdot \log_2 \left(1 + \frac{n m_k}{2\mu(J_k)}\right) \,, \end{aligned} \tag{12}$$

where the last step comes from the Paley-Zygmund inequality (see Lemma 8). Therefore, the only remaining challenge is to lower bound the quantity m_k and upper bound s_k. To that end, we leverage the bi-Lipschitz nature of F: since f is bounded as $c_F \leq f \leq C_F$, this directly implies

$$c_F|x - y| \leq |F(x) - F(y)| \leq C_F|x - y| \tag{13}$$

for all $x, y \in \mathcal{Q}$. Now, recall our assumption that for each k, there exists $q_k \in I_k$ such that $F(q_k) = c_k$. Hence, we may turn to inequality (13) with $y = q_k$ to get, for all $q \in I_k$:

$$\begin{aligned} |F(q) - c_k| &= |F(q) - F(q_k)| \geq c_F|q - q_k| \,, \\ |F(q) - c_k| &= |F(q) - F(q_k)| \leq C_F|q - q_k| \,. \end{aligned}$$

Now, under these last conditions, we may invoke Lemma 9 to obtain bounds on m_k and s_k:

$$m_k \geq \frac{c_F^2}{4C_F^2} \, \mu(J_k)^2 \,, \qquad s_k \leq \frac{C_F^3}{3c_F^3} \, \mu(I_k)^3 \,.$$

We now substitute these values into the Paley-Zygmund bound from (12), yielding:

$$E_k \geq 3\left(\frac{c_F}{2C_F}\right)^7 \frac{\mu(J_k)^4}{\mu(I_k)^3} \log_2\left(1 + \frac{1}{2}\left(\frac{c_F}{2C_F}\right)^2 n\,\mu(J_k)\right). \tag{14}$$

This provides an explicit, fully non-asymptotic lower bound for each E_k in terms of c_F, C_F, n, and the measures $\mu(J_k)$ and $\mu(I_k)$. For the last part of the proof, we incorporate further information about these measures and aggregate over all $k = 1, \ldots, K$.

Step 4: Summing over k and final aggregation. We first partition the sets $\{J_k\}$ into two distinct groups, based on the extent to which their measures have decreased relative to the corresponding original I_k. Specifically, we define the index set $\mathcal{K}$ as

$$\mathcal{K} = \{k \in \{1, \ldots, K\} \text{ such that } \mu(J_k) \geq \mu(I_k)/3\}.$$

A key observation is that, when summing over all k outside of $\mathcal{K}$, we get

$$\sum_{k \notin \mathcal{K}} \mu(J_k) < \frac{1}{3}\sum_{k \notin \mathcal{K}} \mu(I_k) \leq \frac{1}{3}\mu(\mathcal{Q}) = \frac{1}{3}.$$

From this, we derive a key condition:

$$\sum_{k \in \mathcal{K}} \mu(J_k) = \sum_{k=1}^{K} \mu(J_k) - \sum_{k \notin \mathcal{K}} \mu(J_k) = \mu(A) - \sum_{k \notin \mathcal{K}} \mu(J_k) \geq \frac{1}{2} - \frac{1}{3} = \frac{1}{6}, \tag{15}$$

where we have used the fact that $\mu(A) \geq 1/2$, as derived earlier via Markov's inequality. Notice that this necessarily means $|\mathcal{K}|$ is not empty; hence, we can assume $|\mathcal{K}| \geq 1$. Now, by definition, for any $k \in \mathcal{K}$ it holds that $3\mu(J_k) \geq \mu(I_k)$. Combining with (14) and denoting $\gamma = \frac{c_F}{2C_F}$:

$$\begin{aligned}
\sum_{k=1}^{K} E_k &\geq \sum_{k \in \mathcal{K}} E_k \\
&\geq \frac{\gamma^7}{9} \sum_{k \in \mathcal{K}} \mu(J_k) \log_2\left(1 + \frac{\gamma^2 n}{2}\,\mu(J_k)\right) \\
&\geq \frac{\gamma^7}{9} \sum_{k \in \mathcal{K}} \mu(J_k) \log_2\left(\frac{\gamma^2 n}{2}\,\mu(J_k)\right) \\
&= \frac{\gamma^7}{9}\left[\log_2\left(\frac{\gamma^2 n}{2}\right)\sum_{k \in \mathcal{K}} \mu(J_k) + \sum_{k \in \mathcal{K}} \mu(J_k)\log_2\left(\mu(J_k)\right)\right].
\end{aligned}$$

Denote by S the sum $S = \sum_{k \in \mathcal{K}} \mu(J_k)$. Then, by Lemma 10, we get:

$$\begin{aligned}
\sum_{k=1}^{K} E_k &\geq \frac{\gamma^7}{9}\left[S\log_2\left(\frac{\gamma^2 n}{2}\right) + S\log_2\left(\frac{S}{|\mathcal{K}|}\right)\right] \\
&= \frac{\gamma^7}{9}\left[S\log_2\left(\frac{n}{4|\mathcal{K}|}\right) + S\log_2\left(2S\right) + S\log_2\left(\gamma^2\right)\right].
\end{aligned}$$

Now, the function $S \mapsto S\log_2(2S)$ is never smaller than $-0.5 = \log_2(1/\sqrt{2})$ for $S > 0$. Hence:

$$\geq \frac{\gamma^7}{9}\left[S\log_2\left(\frac{n}{4|\mathcal{K}|}\right) + \log_2\left(\frac{1}{\sqrt{2}}\right) + S\log_2\left(\gamma^2\right)\right].$$

On the other hand, we know $\gamma^2 < 1$, so that $\log_2\left(\gamma^2\right) < 0$. Since $S \leq \mu(\mathcal{Q}) = 1$:

$$\geq \frac{\gamma^7}{9}\left[S\log_2\left(\frac{n}{4|\mathcal{K}|}\right) + \log_2\left(\frac{\gamma^2}{\sqrt{2}}\right)\right].$$

Finally, consider the fact that $|\mathcal{K}| \leq K$. Moreover, from Section 5.1, we know that when $d\mu = dF$ it holds that $R_{K,P_0} = \frac{1}{4K}$ (see equation (6)). Replacing above:

$$\sum_{k=1}^{K} E_k \geq \frac{\gamma^7}{9}\left[S\log_2\left(nR_{K,P_0}\right) + \log_2\left(\frac{\gamma^2}{\sqrt{2}}\right)\right].$$

Now, by assumption, we have $nR_{K,P_0} \geq 1$. Hence $\log_2\left(nR_{K,P_0}\right) \geq 0$, and by (15) we conclude:

$$\begin{aligned}\sum_{k=1}^{K} E_k &\geq \frac{\gamma^7}{9}\left[\frac{1}{6}\log_2\left(nR_{K,P_0}\right) + \log_2\left(\frac{\gamma^2}{\sqrt{2}}\right)\right]\\ &\geq \frac{\gamma^7}{54}\left[\log_2\left(nR_{K,P_0}\right) - \log_2\left(\frac{8}{\gamma^{12}}\right)\right].\end{aligned}$$

The proof concludes by setting $C_1 = \frac{\gamma^7}{54}$ and $C_2 = \log_2\left(\frac{8}{\gamma^{12}}\right)$. ◀

C Auxiliary Results from Appendix B

▶ **Lemma 8** (Paley–Zygmund inequality). *Consider any measure μ over $\mathbb{R}$. Let $J \subseteq \mathbb{R}$ be a measurable set with $\mu(J) > 0$, and $G : J \to [0, \infty)$ be a measurable, non-negative function. Define its first and second moments over J as*

$$m_J = \int_J G(q)\, d\mu\,, \qquad s_J = \int_J G(q)^2 d\mu\,.$$

Then, for any $\theta \in (0, 1)$, the subset

$$J'_\theta = \left\{q \in J \;\text{ such that }\; G(q) \geq \theta\frac{m_J}{\mu(J)}\right\}$$

satisfies the lower bound

$$\mu(J'_\theta) \geq (1-\theta)^2\,\frac{m_J^2}{s_J}\,.$$

Proof. Notice that the first moment can be written as

$$m_J = \int_J G(q)\, d\mu = \int_J G(q)\,\mathbf{1}_{J'_\theta}(q)\, d\mu + \int_J G(q)\,\mathbf{1}_{J\setminus J'_\theta}(q)\, d\mu\,, \tag{16}$$

where $\mathbf{1}_U$ represents the indicator function of a set U. The first term on the right hand side can be bounded via Cauchy–Schwarz inequality:

$$\int_J G(q)\,\mathbf{1}_{J'_\theta}(q)\, d\mu \leq \left(\int_J \mathbf{1}_{J'_\theta}(q)\, d\mu\right)^{\frac{1}{2}}\left(\int_J G^2(q) d\mu\right)^{\frac{1}{2}} = \sqrt{\mu(J'_\theta)\cdot s_J}\,.$$

On the other hand, the last term in (16) can be upper bounded as

$$\int_J G(q)\,\mathbf{1}_{J\setminus J'_\theta}(q)\, d\mu = \int_{J\setminus J'_\theta} G(q)\, d\mu \leq \int_{J\setminus J'_\theta} \theta\frac{m_J}{\mu(J)}\, d\mu \leq \theta\frac{m_J}{\mu(J)}\mu(J) = \theta\cdot m_J\,.$$

Replacing these bounds in equation (16), we get

$$m_J \leq \sqrt{\mu(J'_\theta) \cdot s_J} + \theta \cdot m_J \,.$$

The result follows after subtracting $\theta \cdot m_J$ from both sides, squaring, and dividing by s_J. ◀

▶ **Lemma 9.** *Let $I \subseteq \mathbb{R}$ be a finite interval and $J \subseteq I$ a measurable subset of I with positive measure $\mu(J) > 0$, where μ is defined by $d\mu = f\,dq$ for a density f satisfying $0 < c_F \leq f(q) \leq C_F < \infty$, for all $q \in I$. Suppose $F : J \to \mathbb{R}$ is a measurable function such that there exists a point $q_0 \in I$ satisfying the bi-Lipschitz condition*

$$c_F|q - q_0| \leq |F(q) - F(q_0)| \leq C_F|q - q_0| \tag{17}$$

for all $q \in I$, where $c_F, C_F > 0$ are the same constants as above. Then, defining the first and second moments over J with respect to μ:

$$m_J = \int_J |F(q) - F(q_0)|\, d\mu \,, \qquad s_J = \int_J |F(q) - F(q_0)|^2\, d\mu \,,$$

we have the following bounds:

$$m_J \geq \frac{c_F^2}{4C_F^2}\mu(J)^2 \,, \qquad s_J \leq \frac{C_F^3\,\mu(I)^3}{3c_F^3} \,.$$

Proof. From the bi-Lipschitz condition (17), we obtain the following bound on the first moment:

$$m_J = \int_J |F(q) - F(q_0)|\, d\mu \geq c_F \int_J |q - q_0|\, d\mu \,.$$

Since $d\mu = f\,dq$ and $f \geq c_F$, we have:

$$m_J \geq c_F \int_J |q - q_0|\, f(q) dq \geq c_F^2 \int_J |q - q_0|\, dq \,.$$

The right-hand side is minimized when J is a symmetric interval of length $\lambda(J)$ centered at q_0, where λ denotes the Lebesgue measure. In this case, the integral becomes

$$m_J \geq c_F^2 \int_{-\lambda(J)/2}^{\lambda(J)/2} |q|\, dq = 2\, c_F^2 \int_0^{\lambda(J)/2} q\, dq = \frac{c_F^2}{4}\lambda(J)^2 \,.$$

Finally, notice that $\lambda(J) = \int_J 1 dq \geq \frac{1}{C_F}\int_J f(q) dq = \frac{\mu(J)}{C_F}$, which gives the desired bound:

$$m_J \geq \frac{c_F^2}{4C_F^2}\mu(J)^2 \,.$$

For the second moment s_J, we first use the upper bound in (17), which implies

$$s_J = \int_J |F(q) - F(q_0)|^2 d\mu \leq C_F^2 \int_J |q - q_0|^2\, d\mu \,.$$

Since $d\mu = f\,dq$ and $f \leq C_F$, we have:

$$s_J \leq C_F^2 \int_J |q - q_0|^2\, f(q) dq \leq C_F^3 \int_J |q - q_0|^2\, dq \,.$$

This last quantity is maximized when $J = I$ and q_0 lies at one of its endpoints. This corresponds to the following integral:

$$s_J \leq C_F^3 \int_0^{\lambda(I)} q^2 dq = \frac{C_F^3}{3} \lambda(I)^3 ,$$

where λ denotes the Lebesgue measure. Finally, notice that $\lambda(I) = \int_I 1 dq \leq \frac{1}{c_F} \int_I f(q) dq = \frac{\mu(I)}{c_F}$, which gives the desired bound:

$$s_J \leq \frac{C_F^3}{3c_F^3} \mu(I)^3 .$$

This completes the proof. ◀

▶ **Lemma 10** (Entropy bound under partial mass). *Let K be a positive integer and $p_1, \ldots, p_K$ be non-negative real numbers such that $\sum_{k=1}^K p_k = S > 0$. Then, it holds that*

$$\sum_{k=1}^K p_k \log_2(p_k) \geq S \log_2 \left(\frac{S}{K} \right) .$$

Proof. Let us define the normalized values $q_k = \frac{p_k}{S}$, for all $k = 1, \ldots, K$. Then, it holds that $q_k \geq 0$ and $\sum_{k=1}^K q_k = 1$, i.e., $\{q_k\}_{k=1}^K$ constitutes a discrete probability distribution. We compute:

$$\begin{aligned} \sum_{k=1}^K p_k \log_2(p_k) &= \sum_{k=1}^K S\, q_k \log_2(S\, q_k) \\ &= S \sum_{k=1}^K q_k \log_2(S) + S \sum_{k=1}^K q_k \log_2(q_k) \\ &= S \log_2(S) + S \sum_{k=1}^K q_k \log_2(q_k) . \end{aligned} \tag{18}$$

The term $\sum_{k=1}^K q_k \log_2(q_k)$ can be identified as the (negative) Shannon entropy, which is maximized for the uniform distribution (i.e., when $q_k = 1/K$). This means that:

$$-\sum_{k=1}^K q_k \log_2(q_k) \leq \log_2(K) \implies \sum_{k=1}^K q_k \log_2(q_k) \geq \log_2 \left(\frac{1}{K} \right) .$$

Applying this last inequality to equation (18), we get:

$$\begin{aligned} \sum_{k=1}^K p_k \log_2(p_k) &\geq S \log_2(S) + S \log_2 \left(\frac{1}{K} \right) \\ &= S \log_2 \left(\frac{S}{K} \right) , \end{aligned}$$

as claimed. ◀

D Proofs from Section 4.2

An important result in the study of empirical processes is the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [24], which we recap here. If $X_1, \ldots, X_n$ are i.i.d. and $\delta_n = F - F_n$ is the difference between the true CDF F and the empirical CDF F_n, then for any $t > 0$ it holds

$$\mathbb{P}\left(\|\delta_n\|_\infty > t \right) \leq 2e^{-2nt^2} .$$

We use this result in the proof of Lemma 3.

Proof of Lemma 3. Part (a) follows from the tail formula for expected value and the DKW inequality [24]:

$$\mathbb{E}_X\left[\|\delta_n\|_\infty\right] = \int_0^\infty \mathbb{P}\left(\|\delta_n\|_\infty > t\right) dt \leq \int_0^\infty 2e^{-2nt^2} dt\,.$$

Substituting $u = 2t\sqrt{n}$, $du = 2\sqrt{n}\,dt$ we get:

$$\mathbb{E}_X\left[\|\delta_n\|_\infty\right] \leq \frac{1}{\sqrt{n}} \int_0^\infty e^{-\frac{u^2}{2}} du = \sqrt{\frac{2\pi}{n}} \underbrace{\int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du}_{=\frac{1}{2}} = \sqrt{\frac{\pi}{2n}}\,,$$

where the last integral equals 1/2 because the integrand is a standard Gaussian density. For Part (b), we use the fact that μ is a probability measure (i.e., $\mu(\mathcal{Q}) = 1$):

$$\|\delta_n\|_\mu = \int_{\mathcal{Q}} |\delta_n(q)| d\mu \leq \sup_{q\in\mathcal{Q}} |\delta_n(q)| \int_{\mathcal{Q}} 1\, d\mu \leq \|\delta_n\|_\infty \cdot \mu(\mathcal{Q}) = \|\delta_n\|_\infty\,,$$

so that $\|\delta_n\|_\mu \leq \|\delta_n\|_\infty$ for any realization of $X_1, \ldots, X_n$. Applying monotonicity of expectation:

$$\mathbb{E}_X\left[\|\delta_n\|_\mu\right] \leq \mathbb{E}_X\left[\|\delta_n\|_\infty\right] \leq \sqrt{\frac{\pi}{2n}}\,.$$

◀

Proof of Lemma 4. For Part (a) we use Lemma 3, which shows that $\mathbb{E}_X\left[\|\delta_n\|_\mu\right] \leq \sqrt{\frac{\pi}{2n}}$. The expectation of a random variable cannot be strictly smaller than every individual value. Therefore, there must exist at least one realization of $\{X_i\}_{i=1}^n$ such that $\|\delta_n\|_\mu \leq \sqrt{\frac{\pi}{2n}}$. For Part (b), we make use of a result by Csörgő and Faraway [4]. Define the Cramér–von Mises statistic as

$$\omega_n^2 = n \int (F(q) - F_n(q))^2 dF\,.$$

Then, for any $t \in \left[\frac{1}{12n}, \frac{n+3}{12n^2}\right]$ it holds ([4, Equation 2.4]):

$$\mathbb{P}\left(\omega_n^2 \leq t\right) = \frac{n!\pi^{n/2}}{\Gamma(n/2+1)} \left(t - \frac{1}{12n}\right)^{n/2} > 0\,. \tag{19}$$

Now, choose $t = \frac{1}{12n} + \frac{n}{12n^2} = \frac{1}{6n}$. Notice that:

$$\omega_n^2 = n \int (F(q) - F_n(q))^2 dF = n\|\delta_n^2\|_\mu \geq n\|\delta_n\|_\mu^2\,,$$

where the last step is due to Jensen's inequality. Hence, we can write:

$$\mathbb{P}\left(\|\delta_n\|_\mu \leq \frac{1}{\sqrt{6}\,n}\right) = \mathbb{P}\left(n\|\delta_n\|_\mu^2 \leq \frac{1}{6n}\right) \geq \mathbb{P}\left(\omega_n^2 \leq \frac{1}{6n}\right) = \frac{n!\pi^{n/2}}{\Gamma(n/2+1)} \left(\frac{1}{12n}\right)^{n/2} > 0\,,$$

where the last two steps are an application of equation (19). Since $\left\{\|\delta_n\|_\mu \leq \frac{1}{\sqrt{6}\,n}\right\}$ has probability greater than 0, there must be a realization of $\{X_i\}_{i=1}^n$ where it occurs. This concludes the proof. ◀

E Lower Bounds on Query Time: Formal Results and Proofs

We now derive concrete lower bounds on the query time $T(q)$ under the three search strategies described in Section 4.3. We do this by combining the approximation-theoretic results from Propositions 1 and 2 with the deviation bounds from Lemmas 3 and 4. We begin with the case of linear search, for which $T(q) \geq \varepsilon(q)$. From this inequality, we obtain:

▶ **Corollary 11** (Lower bound for linear search). *Let $X_1, \ldots, X_n$ be i.i.d. samples from a distribution with CDF F, and let $q \sim \mu$ independently from the $\{X_i\}$. Let $I(A_n)$ be a learned index from the class $\mathcal{I}_{K,\mathcal{L}}$, then the expected query time under linear search satisfies:*

$$\mathbb{E}_{X,q}\left[T(q)\right] \geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\right].$$

Moreover, if $d\mu = dF$, then there exists a realization $X_1, \ldots, X_n$ such that

$$\mathbb{E}_{q}\left[T(q)\right] \geq n\left[R_{K,\mathcal{L}}(F) - \tfrac{1}{\sqrt{6}\,n}\right].$$

Proof. From Proposition 5 we know $T(q) \geq \varepsilon(q)$. Since q is independent of the $\{X_i\}$:

$$\mathbb{E}_{X,q}\left[T(q)\right] \geq \mathbb{E}_{X,q}\left[\varepsilon(q)\right] = \mathbb{E}_{X}\left[\mathbb{E}_{q}\left[\varepsilon(q)\right]\right].$$

Now, we lower bound $\mathbb{E}_q\left[\varepsilon(q)\right]$ via Proposition 1:

$$\begin{aligned}\mathbb{E}_{X,q}\left[T(q)\right] &\geq \mathbb{E}_{X}\left[n\left[R_{K,\mathcal{L}}(F) - \|\delta_n\|_\mu\right]\right] \\ &= n\left[R_{K,\mathcal{L}}(F) - \mathbb{E}_{X}\left[\|\delta_n\|_\mu\right]\right] \\ &\geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\right],\end{aligned}$$

where the last step comes from Lemma 3(b). This proves the first part of the corollary. The second part follows directly from Proposition 1 and Lemma 4(b). ◀

Notice that the first part of the corollary provides a stronger type of guarantee, as it holds in full expectation over both the data and query distributions ($\mathbb{E}_{X,q}[\cdot]$), and applies to any query distribution. In contrast, the second part assumes $d\mu = dF$ and yields only a worst-case guarantee over the data sample. However, this stronger assumption enables a sharper convergence rate: the deviation term scales as $1/n$ rather than $1/\sqrt{n}$. Similar trade-offs hold for other search algorithms.

In the case of exponential search, Proposition 5 gives $T(q) \geq \widetilde{T}(q) = \log_2\left(\max\{2, \varepsilon(q)\}\right)$. For notational simplicity, we will write $T(q) \geq \log_2 \varepsilon(q)$, where $\log_2(x)$ is assumed to output 1 for $x \leq 2$. In this regime, we can derive the following lower bounds.

▶ **Corollary 12** (Lower bound for exponential search). *Let $I(A_n)$ be a learned index from the class $\mathcal{I}_{K,\mathcal{L}}$. Then, there exists a realization of $\{X_i\}_{i=1}^n$ and $q \in \mathcal{Q}$ such that*

$$T(q) \geq \begin{cases} \log_2\left(n\left[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\right]\right) & \text{for arbitrary } \mu, \\ \log_2\left(n\left[R_{K,\mathcal{L}}(F) - \tfrac{1}{\sqrt{6}\,n}\right]\right) & \text{if } d\mu = dF. \end{cases}$$

Moreover, if the assumptions from Proposition 2 are satisfied, then there exists a realization of $\{X_i\}_{i=1}^n$ such that the expected query time under exponential search satisfies:

$$\mathbb{E}_q\left[T(q)\right] \geq C_1\left[\log_2\left(nR_{K,\mathcal{L}}(F)\right) - C_2\right],$$

where $C_1, C_2 > 0$ are constants independent of n and K.

Proof. We know $T(q) \geq \widetilde{T}(q)$ (Proposition 5). Hence, the second part of the corollary is a direct application of Proposition 2. For the first part, consider first an arbitrary μ. Then by Lemma 4(a), there is a realization of $\{X_i\}_{i=1}^n$ such that $\|\delta_n\|_\mu \leq \sqrt{\frac{\pi}{2n}}$. At the same time, by Proposition 1:

$$\mathbb{E}_q\big[\varepsilon(q)\big] \geq n\Big[R_{K,\mathcal{L}}(F) - \|\delta_n\|_\mu\Big] \geq n\Big[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\Big]\,,$$

so there exists at least one q for which $\varepsilon(q) \geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\frac{\pi}{2n}}\right]$. Combining this with $T(q) \geq \log_2 \varepsilon(q)$ proves the result. The case $d\mu = dF$ is proven in the same way, appealing to case (b) of Lemma 4 to argue that there exists a realization of $\{X_i\}_{i=1}^n$ such that $\|\delta_n\|_\mu \leq \frac{1}{\sqrt{6}\,n}$. ◀

Finally, for binary search, we consider the worst-case query time over both the data and the query value $q \in \mathcal{Q}$. The cost is determined by the worst-case prediction error, as stated by Proposition 5.

▶ **Corollary 13** (Lower bound for binary search). *Let $I(A_n)$ be a learned index from the class $\mathcal{I}_{K,\mathcal{L}}$. Then, there exists a realization of $\{X_i\}_{i=1}^n$ and $q \in \mathcal{Q}$ such that*

$$\sup_q T(q) \geq \begin{cases} \log_2\left(n\Big[R_{K,\mathcal{L}}(F) - \sqrt{\frac{\pi}{2n}}\Big]\right) & \textit{for arbitrary } \mu\,, \\[2ex] \log_2\left(n\Big[R_{K,\mathcal{L}}(F) - \frac{1}{\sqrt{6}\,n}\Big]\right) & \textit{if } d\mu = dF\,. \end{cases}$$

Proof. Consider first the case of arbitrary μ. By Lemma 4(a), there exists a realization of $\{X_i\}_{i=1}^n$ such that $\|\delta_n\|_\mu \leq \sqrt{\frac{\pi}{2n}}$. At the same time, by Proposition 1, we know

$$\mathbb{E}_q\big[\varepsilon(q)\big] \geq n\Big[R_{K,\mathcal{L}}(F) - \|\delta_n\|_\mu\Big] \geq n\Big[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\Big]\,.$$

Hence, there exists at least one q for which $\varepsilon(q) \geq n\left[R_{K,\mathcal{L}}(F) - \sqrt{\frac{\pi}{2n}}\right]$, which implies

$$\sup_q \varepsilon(q) \geq n\Big[R_{K,\mathcal{L}}(F) - \sqrt{\tfrac{\pi}{2n}}\Big]\,.$$

Combining this with $\sup_q T(q) \geq \log_2(\sup_q \varepsilon(q))$ (Proposition 5) proves the result. The case $d\mu = dF$ is proven in the same way, appealing to case (b) of Lemma 4 to argue that there exists a realization of $\{X_i\}_{i=1}^n$ such that $\|\delta_n\|_\mu \leq \frac{1}{\sqrt{6}\,n}$. ◀

These results flesh out the general framework developed in Section 4, and express concrete lower bounds on query time in terms of n, the number of keys, and $R_{K,\mathcal{L}}(F)$, the best approximation error achievable with a given model class.

F Piecewise Constant Approximation: Mismatched Dataset and Query Distributions

In contrast to matching dataset and query distributions $f = g$ considered in Section 5.1, the case $f \neq g$ is relatively more involved; here too, however, the problem of approximating F can be reduced to optimal quantization. The only difference is that the quantized random variable is no longer uniformly distributed, and our results hold asymptotically for $K \to \infty$ as opposed to any given K as in Section 5.1.

The results in this section assume that F is invertible. This is a relatively mild condition, which holds for most standard examples of continuous distributions. Moreover, this assumption can be relaxed: if F is only invertible on a subinterval $[a, b] \subseteq \mathcal{Q}$, the analysis still

applies by conditioning the query distribution to $[a, b]$. This affects only the scaling constants, through changes in the conditional query distribution, but preserves the asymptotic behavior in K. For clarity of exposition, we focus on the case where F is invertible over all of $\mathcal{Q}$.

▶ **Proposition 14** (Reduction to Quantization). *Let $h \in \mathcal{I}_{K,P_0}$ and $Y = F(Q)$ with $Q \sim \mu$ and $d\mu = g\,dq$. Then*

$$\|F - h\|_{\mu,r} \stackrel{\text{def}}{=\!=} \int_{\mathcal{Q}} |F(q) - h(q)|^r g(q) dq = \mathbb{E}[|Y - Q_K(Y)|^r]\,.$$

Proof. As in (4), we expand $\|F - h\|_{\mu,r}$ as

$$\|F - h\|_{\mu,r} = \sum_{k=1}^{K} \int_{I_k} |F(q) - c_k|^r g(q) dq\,.$$

Introducing $y = F(q),\ dy = f(q)dq,\ q = F^{-1}(y)$, and $dq = 1/f(q)\,dy$, we rewrite this as

$$\sum_{k=1}^{K} \int_{J_k} |y - c_k|^r \cdot \frac{g\left(F^{-1}(y)\right)}{f\left(F^{-1}(y)\right)} dy = \mathbb{E}[|Y - Q_K(Y)|^r]\,,$$

where the last equality follows because we recognize that the second term in the integral is the probability density function (PDF) of $Y = F(Q)$ when Q distributes according to the PDF g:

$$f_Y(y) = g\left(F^{-1}(y)\right) \cdot \left|\frac{d}{dy} F^{-1}(y)\right| = g\left(F^{-1}(y)\right) \cdot \left|\frac{1}{F'\left(F^{-1}(y)\right)}\right| = \frac{g\left(F^{-1}(y)\right)}{f\left(F^{-1}(y)\right)}\,; \tag{20}$$

the first equality expands the density of a function of a random variable, and the second the derivative of an inverse function. We remark that this requires the CDF F to be invertible. ◀

Proposition 14 implies that approximating F reduces to quantizing a random variable $Y = F(Q)$ distributed according to (20) over $[0, 1]$. When $f \neq g$, this distribution is no longer uniform, and uniform quantization is no longer optimal. To carry on with our visual analogy in Figure 5, one can imagine the density of the green dashed lines below F no longer proportional to f, and therefore the ones above F no longer uniform, in which case the approximating values $\{c_k\}$ would have to be shifted accordingly.

We are now in a position to leverage an important result in quantization theory. The asymptotic quantization error is known to satisfy [13, Theorem 6.2]

$$\lim_{K\to\infty} K^r\ T_{K,r}(Y) = \frac{1}{(1+r)2^r} \left(\int f_Y(y)^{1/(1+r)} dy\right)^{1+r}\,.$$

The right-hand side of the equality is known as the r-th quantization coefficient and is a fundamental property of a probability distribution, similar to its moments; quantization coefficients play an important role in quantization theory. Therefore, for the absolute error

$$\lim_{K\to\infty} K\ R_{K,P_0}(F) = \lim_{K\to\infty} K\ T_{K,1}(Y) = \frac{1}{4} \left(\int \sqrt{f_Y(y)} dy\right)^2\,.$$

In other words, asymptotically, the absolute error decreases linearly with K, just like in the case of the matching distribution considered in Section 5.1, yielding $R_{K,\mathcal{L}}(F) = \Omega\left(1/K\right)$ when $\mathcal{L} = P_0$.

G Proofs from Section 6

Proof of Lemma 6. $R_{K,\mathcal{L}}(F)$ can be written as

$$R_{K,\mathcal{L}}(F) = \inf_{h \in \mathcal{I}_{K,\mathcal{L}}} \|F - h\|_\mu = \inf_{\mathcal{P} \in \mathbf{P}_K(\mathcal{Q})} \inf_{h \in \mathcal{I}_{\mathcal{P},\mathcal{L}}} \|F - h\|_\mu \,,$$

where $\inf_{\mathcal{P} \in \mathbf{P}_K(\mathcal{Q})}$ is taken over the set $\mathbf{P}_K(\mathcal{Q})$ of all partitions of $\mathcal{Q}$ into K intervals. By definition of infimum, for any $t > 0$ there exists $\mathcal{P}_t \in \mathbf{P}_K(\mathcal{Q})$ such that

$$R_{K,\mathcal{L}}(F) \geq \inf_{h \in \mathcal{I}_{\mathcal{P}_t,\mathcal{L}}} \|F - h\|_\mu - t \,.$$

Taking supremum with respect to F:

$$\sup_{F \in \mathcal{F}} R_{K,\mathcal{L}}(F) \geq \sup_{F \in \mathcal{F}} \inf_{h \in \mathcal{I}_{\mathcal{P}_t,\mathcal{L}}} \|F - h\|_\mu - t \,.$$

Now, $\mathcal{I}_{\mathcal{P}_t,\mathcal{L}}$ is a linear space of dimension at most $\kappa = K \cdot \dim(\mathcal{L})$. Hence, the last expression can be lower bounded by taking infimum over all such spaces:

$$\sup_{F \in \mathcal{F}} R_{K,\mathcal{L}}(F) \geq \inf_{H \in \mathcal{H}_\kappa} \sup_{F \in \mathcal{F}} \inf_{h \in H} \|F - h\|_\mu - t = d_\kappa(\mathcal{F}) - t \,.$$

Notice that the inf-sup-inf sequence now has no dependence on t. Therefore, since the inequality holds for all $t > 0$, it holds for $t = 0$ too, by standard properties of the real numbers. ◀

G.1 Upper Bounds for Piecewise Linear Approximation

We briefly sketch standard upper bounds for the approximation error $R_{K,\mathcal{L}}(F)$ when $\mathcal{L} = P_1$, the class of degree-one polynomials, under two common regularity assumptions on the target function F. Throughout, we assume the domain is $[0, 1]$, the measure μ is the Lebesgue measure, and the interval is divided into K uniform subintervals $[a_i, a_{i+1}]$, where $a_i = \frac{i-1}{K}$ and $a_{i+1} = \frac{i}{K}$. We denote the length of each subinterval by $\ell = 1/K$.

Case 1: F is twice continuously differentiable. Let us approximate F over each interval $[a_i, a_{i+1}]$ using a linear function based on a first-order Taylor expansion:

$$L_i(x) = F(a_i) + (x - a_i)F'(a_i) \,.$$

Since F is $C^2([0, 1])$, we can bound the pointwise error using Taylor's theorem:

$$|F(x) - L_i(x)| \leq \frac{1}{2} M \cdot (x - a_i)^2 \,,$$

where $M = \max_{x \in [0,1]} |F''(x)|$. We know that M exists and is bounded because F'' is continuous and $[0, 1]$ is a compact interval. Integrating this over each interval yields:

$$\int_{a_i}^{a_{i+1}} |F(x) - L_i(x)| dx \leq \frac{1}{2} M \int_{a_i}^{a_{i+1}} (x - a_i)^2 dx = \frac{1}{6} M \ell^3 \,.$$

Finally, summing over all K intervals gives:

$$\|F - h\|_\mu \leq K \cdot \frac{1}{6} M \ell^3 = O(1/K^2) \,.$$

Case 2: F is M-Lipschitz. Let us now approximate F using linear interpolation between endpoints:

$$L_i(x) = F(a_i) + \frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i} \cdot (x - a_i)\,.$$

By the Lipschitz property, for any $x \in [a_i, a_{i+1}]$ we have:

$$\begin{aligned}
|F(x) - L_i(x)| &= \left|F(x) - F(a_i) - \frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i}(x - a_i)\right| \\
&= \left|\frac{F(x) - F(a_i)}{x - a_i} - \frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i}\right| \cdot |x - a_i| \\
&\leq \left(\left|\frac{F(x) - F(a_i)}{x - a_i}\right| + \left|\frac{F(a_{i+1}) - F(a_i)}{a_{i+1} - a_i}\right|\right) \cdot |x - a_i| \\
&\leq (M + M)|x - a_i|\,.
\end{aligned}$$

Therefore, the absolute error over each interval is at most $2M|x - a_i|$. Integrating over each interval:

$$\int_{a_i}^{a_{i+1}} |F(x) - L_i(x)|dx \leq 2M \int_{a_i}^{a_{i+1}} |x - a_i|dx = M\ell^2\,.$$

Finally, summing over all K intervals gives:

$$\|F - h\|_\mu \leq K \cdot M\ell^2 = O(1/K)\,.$$

These asymptotic rates for $R_{K,P_1}(F)$—i.e., $O(1/K)$ under Lipschitz continuity and $O(1/K^2)$ under C^2 assumption—hold more generally for functions defined on any compact interval, and under any measure μ that has a bounded density. The specific constants may depend on the domain and the density of μ, but the convergence rates remain unaffected. Thus, piecewise linear functions can provide different approximation guarantees under specific regularity assumptions.

G.2 Construction of Adversarial CDF with Respect to K

We will illustrate the case $\mathcal{Q} = [0, 1]$, with μ being the uniform distribution. This will allows us to perform some exact calculations, but the asymptotic properties we derive in this section, and the argument we use to prove them, apply for very general choices of $\mathcal{Q}$ and μ. Consider the function $F(x) = x^2$ and the optimal approximation error $R_{1,P_1}(F)$ by a single linear function. Notice that x^2 over $[0, 1]$ is a valid CDF. From [26, Corollary 3.4.1], it is easy to deduce that

$$R_{1,P_1}(F) = 1/16\,, \tag{21}$$

which is attained by $h(x) = x - 3/16$. Now, take a positive integer M. The intuition is that we can construct a staircase-type function F_M, with M total steps and each step F_M^i being equal to $F(x) = x^2$ after appropriate translation and scaling, such that the full F_M constitutes a valid CDF and is hard to approximate. For any $i = 0, \ldots, M - 1$ define F_M^i as:

$$F_M^i(x) = \frac{1}{M}\left[M\left(x - \frac{i}{M}\right)\right]^2 + \frac{i}{M} \quad \text{for} \quad x \in I_i = \left[\frac{i}{M}, \frac{i+1}{M}\right].$$

Now, the full F_M is defined as

$$F_M(x) = \sum_{i=0}^{M-1} F_M^i(x)\mathbf{1}_{I_i}(x) \quad \text{for} \quad x \in [0,1]\,.$$

It is easy to check that F_M is continuous, increasing, and satisfies $F(0) = 0$, $F(1) = 1$, thus being a valid CDF on $[0,1]$. Consider now any linear approximation $L_i \in P_1$ for the function F_M^i over the interval I_i. The approximation error is then given by:

$$\begin{aligned} E_i &= \int_{\frac{i}{M}}^{\frac{i+1}{M}} |F_M^i(x) - L_i(x)|dx \\ &= \int_{\frac{i}{M}}^{\frac{i+1}{M}} \left| \frac{1}{M}\left[M\left(x - \frac{i}{M}\right)\right]^2 + \frac{i}{M} - L_i(x) \right| dx\,. \end{aligned}$$

Substituting $u = M(x - i/M),\ du = Mdx$:

$$\begin{aligned} &= \frac{1}{M}\int_0^1 \left| \frac{u^2}{M} + \frac{i}{M} - L_i\left(\frac{u+i}{M}\right) \right| du \\ &= \frac{1}{M^2}\int_0^1 \left| u^2 + i - ML_i\left(\frac{u+i}{M}\right) \right| du\,. \end{aligned}$$

We focus now on the integral term. Since $i - ML_i((u+1)/M) \in P_1$, it cannot induce lower error when approximating u^2 than the optimal approximator $h(u) = u - 3/16$, hence by equation (21):

$$E_i \geq \frac{1}{16M^2}\,. \tag{22}$$

We now use this to derive a lower bound on $R_{K,P_1}(F_M)$. Since F_M consists of M disjoint quadratic steps, each over an interval of length $1/M$, a K-piecewise linear function can assign at most K linear segments across all M sub-parabolas. Therefore, at least $M - K + 1$ sub-parabolas must be covered using only one linear piece each, and on each corresponding interval, bound (22) applies. Summing over these $M - K + 1$ intervals, we obtain

$$R_{K,P_1}(F_M) \geq \frac{M - K + 1}{16M^2}\,.$$

This expression is maximized when $M = 2(K-1)$, yielding

$$R_{K,P_1}(F_{2(K-1)}) \geq \frac{1}{64(K-1)} = \Omega\left(\frac{1}{K}\right).$$

This construction shows that even when F is a valid CDF, it is possible to exhibit $\Omega(1/K)$ lower bounds for approximation using K-piecewise linear functions. In particular, the worst-case behavior of $R_{K,P_1}(F)$ is no better than that of piecewise constant models, despite the greater expressive power of linear segments. This highlights the critical role of regularity assumptions on F: without sufficient smoothness, more expressive models do not necessarily yield improved approximation rates.

G.3 Kolmogorov Widths for Lipschitz CDFs

Without loss of generality, we consider $[a,b] = [0,1]$ and restrict our attention to the class $\mathcal{F}$ of Lipschitz continuous functions with Lipschitz constant 1, that is:

$$\mathcal{F} = \left\{F \in L^1(\mu) : |F(x) - F(y)| \leq |x - y|\right\}.$$

Let $d_\kappa(\mathcal{F})$ denote the Kolmogorov width of $\mathcal{F}$ in $L^1(\mu)$, then:

$$d_\kappa(\mathcal{F}) = \inf_{H \in \mathcal{H}_\kappa} \sup_{F \in \mathcal{F}} \inf_{h \in H} \|F - h\|_\mu = \Omega(1/\kappa)\,, \tag{23}$$

where $d_\kappa(\mathcal{F}) = \Omega(1/\kappa)$ comes from classic results referenced in Section 6.2. However, not all functions in $\mathcal{F}$ are valid cumulative distribution functions (CDFs), as they need not be nonnegative or monotonically increasing. We now show that from any $F \in \mathcal{F}$, we can construct a valid CDF $\widehat{F}$ that remains equally hard to approximate up to constants.

For notational simplicity, denote the distance from F to the set H as $\|F - H\|_\mu$, that is, $\|F - H\|_\mu = \inf_{h \in H} \|F - h\|_\mu$. Notice that the outermost infimum in equation (23) is taken over all linear subspaces $H \subseteq L^1(\mu)$ of dimension at most κ. Hence, the bound applies to any such space. Consider then any H^* which includes at least all constant and linear functions (this can always be done as long as $\kappa \geq 2$, since $\operatorname{span}\{1, x\}$ is an example of such a space). Now, fix $t > 0$. By equation (23) and the definition of supremum, there exists F such that

$$\|F - H^*\|_\mu \geq \frac{C}{\kappa} - t\,,$$

where $C > 0$ is some positive constant independent of t.

Step 1: Constructing a strictly increasing function. It is well-known that $\mathcal{F}$ coincides with the Sobolev ball $W^{1,\infty}([0,1])$ of weakly-differentiable functions, such that their weak derivatives are essentially bounded by 1 [7, Chapter 5.8, Theorem 4]. This implies that $F' \geq -1$ up to a set of measure 0. Now, define

$$\widetilde{F}(x) = F(x) + 2x\,.$$

Then, $\widetilde{F}$ is monotonic on $[0,1]$, because its weak derivative $\widetilde{F}'$ is positive:

$$\widetilde{F}'(x) = F'(x) + 2 \geq -1 + 2 = 1\,.$$

Therefore, $\widetilde{F}$ is strictly increasing, and it is continuous since both F and $2x$ are Lipschitz continuous.

Step 2: Rescaling to a valid CDF. We now normalize $\widetilde{F}$ to obtain a function $\widehat{F}$ satisfying both $\widehat{F}(0) = 0$ and $\widehat{F}(1) = 1$:

$$\widehat{F}(x) = \frac{\widetilde{F}(x) - \widetilde{F}(0)}{\widetilde{F}(1) - \widetilde{F}(0)} = \frac{F(x) + 2x - F(0)}{F(1) + 2 - F(0)}\,.$$

$\widehat{F}$ is strictly increasing and continuous in $[0,1]$, because $\widetilde{F}$ is. Hence, it is a valid CDF. Moreover, we can upper and lower bound the scaling constant:

$$(\widetilde{F}(1) - \widetilde{F}(0)) = (F(1) + 2 - F(0)) \in [1, 3]\,, \tag{24}$$

since $(F(1) - F(0)) \in [-1, 1]$ by the 1-Lipschitz property.

Step 3: Preserving approximation difficulty. Recall that

$$\|F - H^*\|_\mu \geq \frac{C}{\kappa} - t\,.$$

Now, denote $M = F(1) + 2 - F(0)$ and consider

$$\|\widehat{F} - H^*\|_\mu = \left\| \frac{F + 2x - F(0)}{M} - H^* \right\|_\mu = \frac{1}{M} \|F + 2x - F(0) - M \cdot H^*\|_\mu \,.$$

Since H^* is a linear subspace, it is invariant under scalar multiplication, so that $M \cdot H^* = H^*$ provided $M \neq 0$. Furthermore, it also holds that $2x - F(0) - M \cdot H^* = H^*$, because we have assumed that H^* spans the set of constant and linear functions. By equation (24), we know that $M \leq 3$, so all in all we can conclude that:

$$\|\widehat{F} - H^*\|_\mu = \frac{1}{M} \|F - H^*\|_\mu \geq \frac{1}{3} \|F - H^*\|_\mu \geq \frac{C}{3\kappa} - \frac{t}{3} \,. \tag{25}$$

Let $\widehat{\mathcal{F}} = \{\widehat{F}$ constructed from $F \in \mathcal{F}$ as described in Steps 1 and 2$\}$. Then, taking supremum over that class in equation (25) we get:

$$\sup_{\widehat{F} \in \widehat{\mathcal{F}}} \|\widehat{F} - H^*\|_\mu \geq \frac{C}{3\kappa} - \frac{t}{3} \,,$$

where C does not depend on t. Since this holds for any $t > 0$, it holds for $t = 0$ as well. Now, taking infimum with respect to all linear spaces of dimension κ, we can conclude:

$$d_\kappa(\widehat{\mathcal{F}}) \geq \frac{C}{3\kappa} = \Omega\left(\frac{1}{\kappa}\right) .$$

This construction shows that any lower bound on approximation for Lipschitz continuous functions extends to a lower bound for valid CDFs, up to a constant multiplicative factor. Therefore, all conclusions from Section 6.2 remain valid even when restricted specifically to CDFs.