

# A Rate-Distortion Theory for Membership Testing: From Filters to LLM Hallucination

November 4, 2025

## Abstract

We introduce generalized membership testing, a unifying abstraction that captures classical approximate set membership (including Bloom filters) and large-language-model (LLM) decision losses under a single, task-dependent error framework. In the sparse-key regime, we develop an information-theoretic characterization of the optimal space–error tradeoff, revealing a rate–distortion–style frontier whose leading term is governed by relative entropy.

Applying our results on LLMs with cross-entropy loss, we show that an optimally trained/compressed LLM must hallucinate. Specifically, it has to assign fact-level confidence score on a fraction of unseen “non-facts” to minimize loss. Our analysis of the hallucination phenomenon is based solely on memory and optimal compression, independent of the computational and data/sample aspect. For two-sided filters allowing false negatives and false positives, we refine a known space lower bound into its tight form, which we show to be achievable via a simple hash function-based filter.

# 1 Introduction

Approximate membership data structures—epitomized by Bloom filters [10]—address the task of storing a *key* set  $S \subseteq U$  in a vast universe using space proportional to  $|S|$  while tolerating a prescribed error. A typical modern filter permits a false positive rate (FPR)  $\varepsilon$ , guarantees a false negative rate (FNR) of zero, and approaches the information-theoretic optimal space usage of  $n \log_2(1/\varepsilon)$  bits, where  $n = |S|$ . In parallel, recent work on large language models (LLMs) [31] frames *hallucination control* as a decision problem over a universe of *factoids*: given a query  $e \in U$ , a model outputs a calibrated score  $\hat{x} \in [0, 1]$  indicating its belief that  $e$  is true, and is penalized by a loss that depends on whether  $e$  is a fact or not. At a technical level, the two seemingly disparate problems are both sparse membership tests, distinguished primarily by their error metrics.

**A unifying abstraction.** We develop a common framework, called *generalized membership testing*, that captures both problems and more. For a large universe  $U$  with key density  $p = n/|U| \ll 1$ , a *generalized membership tester* is a data structure that, after initialization on *any* key set  $S \subset U$  with  $|S| = n$ , can store the information efficiently using space proportional to  $n$ , and can output a confidence score  $\hat{x} \in [0, 1]$  for any query  $e \in U$ . Errors of its output are measured separately on keys and non-keys via two *error metrics*  $d^K(\hat{x})$  and  $d^N(\hat{x})$ : we use  $d^K(\hat{x}) = 1 - \hat{x}$  and  $d^N(\hat{x}) = \hat{x}$  for two-sided filters (controlling both FNR/FPR), and  $d^K(\hat{x}) = -\ln \hat{x}$ ,  $d^N(\hat{x}) = -\ln(1 - \hat{x})$  for LLMs (cross-entropy). This framework is general and can potentially extend to other settings by specifying a task-dependent error metric. In this paper, we develop theories for custom error metrics under this general framework; in particular, we answer the following question:

*In the generalized membership testing problem with custom error metrics, what is the fundamental trade-off between the error rates and space usage in the sparse regime?*

## 1.1 Our contributions

In this paper, we provide a complete answer to this question by developing an analog for rate-distortion theorem for the generalized membership testing problem. Following that, we apply the theory on both the settings of two-sided filters and LLM decision problem. We summarize the main results as follows.

### 1.1.1 A unifying theory for generalized membership testing

Let  $\text{KL}(\cdot \parallel \cdot)$  denote the Kullback-Leibler divergence with base-2 logarithm, and let  $\chi^2(\cdot \parallel \cdot)$  denote the chi-squared divergence. Our main theorem characterizes the fundamental error-space trade-off in the generalized membership testing problem.

**Theorem 1.1** (informal, see Theorem 3.1). *Consider a generalized membership tester of  $n$  keys in a universe of size  $u$ , where the key density  $p = \frac{n}{u}$  is fixed. Given custom error metrics satisfying certain assumptions, the asymptotic space usage per key, as  $n \rightarrow \infty$ , is at least*

$$\text{KL}(\mu_1^* \parallel \mu_0^*) - \frac{1}{2 \ln 2} \chi^2(\mu_1^* \parallel \mu_0^*) \cdot p + o(p)$$

bits, where  $\mu_1^*$  and  $\mu_0^*$  are distributions on  $[0, 1]$  that minimize the first KL divergence term, with respect to the constraint that a tester drawing answers from  $\mu_1^*$  for keys and  $\mu_0^*$  for non-keys can achieve the desired error rates.

Moreover, this space bound is tight and can be achieved by non-constructive membership testers.

The leading term,  $\text{KL}(\mu_1^* || \mu_0^*)$ , represents the fundamental information cost: it is the minimum information gain needed per key to distinguish the key distribution ( $\mu_1^*$ ) from the background non-key distribution ( $\mu_0^*$ ) in the sparse limit ( $p \rightarrow 0$ ).

While the generality of Theorem 1.1 extends beyond its application to filters and LLM decision problems, in this paper, we focus on the two above-mentioned settings. The results are obtained by solving the respective convex optimization problems over a Banach space. It is an interesting open question to identify other applicable settings with different error metrics.

### 1.1.2 Hallucination in LLMs

Following from the work of Kalai et al. [31, 30], we model hallucination control as a decision task over a universe of factoids. The output of LLM is a confidence score  $\hat{x} \in [0, 1]$ , and errors are measured separately on facts and non-facts via the log-losses  $-\ln \hat{x}$  and  $-\ln(1 - \hat{x})$ , respectively. This places LLMs within our membership-testing framework and isolates *memory/space* as the sole limiting resource, abstracting away architectural and optimization details.

An application of our main theorem shows the following:

**Theorem 1.2** (informal, see Theorems 4.1 and 4.2). *Suppose an LLM, when optimally trained on any  $n$  facts in factoid universe of  $U$  of size  $u = n/p$ , can achieve expected loss  $\varepsilon_K$  on facts (keys) and  $\varepsilon_N$  on non-facts, where  $e^{-\varepsilon_K} + e^{-\varepsilon_N} > 1$ . Then, its asymptotic space usage as  $n \rightarrow \infty$  is at least*

$$\log_2 (-\ln(1 - e^{-\varepsilon_K})) - \log_2(\varepsilon_N) - \Theta_{\varepsilon_K, \varepsilon_N}(p) \quad (*)$$

bits per fact. In particular, the per-fact space usage scales with  $\log \log(1/\varepsilon_K)$  as  $\varepsilon_K \rightarrow 0$ , and scales with  $\log(1/\varepsilon_N)$  as  $\varepsilon_N \rightarrow 0$ .

Moreover, the unique optimal output distributions  $\mu_1^*, \mu_0^*$  on facts and non-facts are the following discrete distributions, respectively:

$$\begin{aligned} \mu_1^* &= \delta_{e^{-\varepsilon_K}}, \\ \mu_0^* &= \left(1 - \frac{\varepsilon_N}{-\ln(1 - e^{-\varepsilon_K})}\right)\delta_0 + \frac{\varepsilon_N}{-\ln(1 - e^{-\varepsilon_K})}\delta_{e^{-\varepsilon_K}}. \end{aligned}$$

Here,  $\delta_x$  is the Dirac delta at  $x$ .

The optimal distributions are graphically illustrated in Figure 1. Apart from demonstrating a fundamental space-error trade-off for parametric memory in LLM, we also stress the following interpretations of our result:

**High-confidence hallucination.** The optimal  $\mu_0^*$  has exactly two atoms: a mass at 0 (the model firmly rejects most non-facts) and a mass at  $e^{-\varepsilon_K}$ , the same confidence it assigns to true, seen facts. This is an explicit, quantitative *hallucination channel*: an optimally compressed model under log-loss must sometimes assert falsehoods with fact-level confidence. In this case, no thresholding strategy can separate the facts from hallucinations. One might hope the opposite would happen, where all losses on non-facts are cumulated over low output values and causes no hallucination; yet this is impossible if the model achieves optimal loss.

**Hallucination is inevitable.** Since the solution is unique, any sufficiently well-trained model that minimizes the two losses must (at least in theory) converge to these distributions. Thus, high-confidence hallucinations are not a bug; they are a feature of the compression task defined by the log-loss objectives. While it is believed that lossless compression leads to intelligence [9], we show that *successful lossy compression provably leads to hallucination*.

**Simplicity and generality of the setting.** The theorem is proved in an extremely stripped-down, *fundamental* setting: a decision model outputs calibrated confidences on factoids. We allow arbitrarily powerful training and inference rules for parametric memory—yet *space alone* still enforces the bound (\*). Our minimalist setting allows the analysis to carry over for any choice of model inference and training, complementing the model- and inference-specific explanations for hallucination in the existing literature (see Section 1.2).

**Benevolent assumptions on samples.** We assume the LLM sees all facts during training, tasking it to report what it is sure of. This contrasts with Kalai et al. [31, 30], where hallucinations primarily stem from unseen facts that require guessing. We also allow the LLM to have small non-fact loss, meaning it is trained to reject all unknowns, despite this being highly impractical due to the sheer size of non-facts. Our stricter setting is complementary to all previous results, as it isolates compression as the cause for hallucination even with complete information.

A numeric illustration of the implication of our theorem on LLM hallucination can be found in Section A in appendix. Detailed results on LLM are presented in Section 4.

### 1.1.3 Optimal space–error trade-off for two-sided filters

We study two-sided filters where a false negative rate (FNR)  $\varepsilon_K$  as well as FPR are allowed. This is a well-motivated [27, 55, 20] generalization of the ordinary membership filter, though its theoretical foundations not yet well-developed: the only known result is a bound of

$$(1 - \varepsilon_K) \log_2 \frac{1}{\varepsilon_N} - \frac{1 - \varepsilon_K}{\varepsilon_N \ln 2} \cdot p - \Theta(1)$$

bits per key [50]. The prior bound’s trailing  $-\Theta(1)$  blurs the true space–error frontier for two-sided filters. This is not innocuous: an additive  $\Theta(1)$  bits per key is exactly the premium separating dynamic from static filters [40, 7], so hiding it masks the inherent trade-off. In this paper, we make this constant explicit and pin down the dependence on  $\varepsilon_K$  and  $\varepsilon_N$ .

1. Applying Theorem 1.1, the optimal distributions are  $\mu_1^* = \text{Bern}(1 - \varepsilon_K)$  and  $\mu_0^* = \text{Bern}(\varepsilon_N)$ , which yields an asymptotic space lower bound of

$$(1 - \varepsilon_K) \log_2 \frac{1 - \varepsilon_K}{\varepsilon_N} + \varepsilon_K \log_2 \frac{\varepsilon_K}{1 - \varepsilon_N} - \frac{(1 - \varepsilon_K - \varepsilon_N)^2}{2\varepsilon_N(1 - \varepsilon_N) \ln 2} \cdot p + o(p)$$

per key, for two-sided filters with false negative rate (FNR)  $\varepsilon_K$  and FPR  $\varepsilon_N$ . This improves upon the previous result with a tighter dependence on  $p$  (by a factor of  $\geq 1/2$ ), and removes the  $\Theta(1)$  per-key space overhead left unspecified.

2. In practice, most filters operate on the hash values of elements  $e_i \in U$  (we call them *universe-independent* in Section 5.2). For these filters, this bound is naturally just:

$$\text{KL}(\text{Bern}(1 - \varepsilon_K) \| \text{Bern}(\varepsilon_N)).$$

This is coherent with the classic  $\log_2 \frac{1}{\varepsilon_N}$  result [13] for one-sided filters when  $\varepsilon_K = 0$ .

3. We design a simple (albeit computationally inefficient) universe-independent filter and prove that it matches the above lower bound up to  $o(1)$  space overhead per key, asserting its tightness.

Results on filters are presented in Section 5.

## 1.2 Related works

**LLMs and hallucination.** Hallucination [28, 1, 26] is often defined as the generation of content that is fluent and plausible but factually inaccurate, nonsensical, or unfaithful to source material. A large body of prior works have identified causal factors throughout the entire LLM development pipeline, attributing them to *data-centric causes* (noisy web corpora) [19, 4, 52], *model-centric causes* (objective/architectural limits of next-token prediction) [29, 3, 25, 59], and *inference-centric causes* (inference-time choices such as stochastic decoding) [24, 43, 37, 39].

A complementary line argues hallucination is not a bug, but inherent: Kalai et al. reduce generation to a decision task over “factoids”, tying errors to calibration and epistemic uncertainty [31, 30]; Xu et al. give impossibility results for the LLM to learn certain computable functions [61]. Our contribution fits this latter thread but isolates *compression/space* as the driver: under log-loss, optimal compression forces a structured channel for high-confidence false positives. This complements mitigation-oriented work [32], and suggests a fundamental limit of purely parametric memory, motivating non-parametric augmentation, e.g. RAG [33, 38].

Also related is the well-known notion that lossless compression is deeply related to LLMs [17], and even the general idea of intelligence itself [14]. In our study, we looked at parametric memory of LLMs through the lens of lossy compression, where the parameters of a model are the compressed code itself. We found that, while optimal lossless compression may lead to intelligence, lossy compression under cross entropy loss will provably lead to hallucination.

**Filters.** Approximate membership filters (classically one-sided) achieved tremendous success in many applications [12, 56, 44, 42]. Bloom filter [10], the design that started this field, has an multiplicative overhead of  $\log_2 e$  over space lower bound. Most works on filter design target practical speed/space improvements [23], approaching the classic  $\log_2(1/\varepsilon_N)$  bound up to  $O(1)$  bits per key at extreme load [6, 5, 51, 22, 11, 58, 49, 8, 21, 7], with some even up to  $o(1)$  per-key overhead [18, 53]. Despite the fundamental distinction, the filters with  $O(1)$  overhead also sometimes claim to be “space optimal”, as they focus on the practical regime where  $\varepsilon_N \rightarrow 0$ . Furthermore, a fundamental  $\Theta(1)$  gap separates dynamic from static filters—filters that allow insertion need  $\Omega(1)$  extra bits per key, with the state of the art attaining  $\log_2(1/\varepsilon_N) + \log_2 e$  [40, 7].

Two-sided filters (allowing FNR to further lower FPR) are motivated by applications such as reconciliation and routing [27, 55], and were also explored heuristically in other standard settings [20, 36, 47, 34]; however, space lower bounds remain sparse, with an early bound that leaves a  $\Theta(1)$  term unspecified [50]. A specific motivation for allowing FNR in filters can be found in appendix B. Finally, there are also many works in learning-augmented data structures [35], including for filters [46, 57]. These designs can further reduce space usage by using ML model to classify the elements in  $U$  before storing and querying them.

## 2 Preliminaries and Technical Overview

All logarithms in the rest of the paper are base-2 unless otherwise specified as ln. We always use  $u$  to denote the universe size  $|U|$ , and use  $n$  to denote the key size  $|S|$ .

### 2.1 Generalized membership testers and error metrics

First, we will define the appropriate error metrics we use in this paper:

**Definition 2.1** (Error metrics). *Let  $d^K, d^N : [0, 1] \rightarrow \mathbb{R}_{\geq 0}$  be our **error metric functions** for keys  $S$  and non-keys  $U \setminus S$ . In particular, on output  $\hat{x} \in [0, 1]$ , the error on a single key is  $d^K(\hat{x})$ , and the error on a single non-key is  $d^N(\hat{x})$ .*

In this paper, we assume the two functions are nonnegative lower semi-continuous, and in case of perfect output we must have  $d^K(1) = d^N(0) = 0$ . Additionally, we assume there exists some  $c \in [0, 1]$  achieving finite error under both metrics:

$$d^K(c) < \infty, \quad \text{and} \quad d^N(c) < \infty.$$

We now formally define the generalized membership tester.

**Definition 2.2** (Generalized membership tester). *Given universe  $U = [u]$ , key set size  $n$ , and two error metric functions, a **generalized membership tester**  $\mathcal{M}$  is a tuple of the following two (random) algorithms:*

- **INITIALIZE**, where the tester takes input  $(S, \varepsilon_K, \varepsilon_N)$  and outputs the bit representation of a data structure,  $D$ .
- **QUERY**, where the tester takes input  $e \in U$  and the data structure  $D$ , and returns a confidence score  $\hat{x} \in [0, 1]$ , indicating its belief of the probability that  $e \in S$ .

We say that  $\mathcal{M}$  achieves average error  $\varepsilon_K$  on keys and  $\varepsilon_N$  on non-keys, if for any  $S \subseteq U$  of size  $n$ , over the randomness of initialize and query, we have:

$$\mathbb{E}_{D \leftarrow \text{INITIALIZE}(S)} \left[ \mathbb{E}_{e \sim \text{Unif}(S)} [d^K(\text{QUERY}(e, D))] \right] \leq \varepsilon_K,$$

and

$$\mathbb{E}_{D \leftarrow \text{INITIALIZE}(S)} \left[ \mathbb{E}_{e \sim \text{Unif}(U \setminus S)} [d^N(\text{QUERY}(e, D))] \right] \leq \varepsilon_N,$$

To give examples for the error metrics, for two-sided filters, we want  $d^K$  and  $d^N$  to characterize the FNR and FPR of the filter, respectively. We hence define the distortion functions as  $d^K(\hat{x}) = 1 - \hat{x}$ ,  $d^N(\hat{x}) = \hat{x}$ .

For LLMs, we want  $d^K$  and  $d^N$  to characterize the *loss* it incurs on a key (fact) and a non-key (non-fact), respectively. In practice, given any prior distribution over the keys and non-keys, this is the cross entropy loss, which translates to the error metric of  $d^K(\hat{x}) = -\ln \hat{x}$ , and  $d^N(\hat{x}) = -\ln(1 - \hat{x})$ .

The core topic we study is the minimum space usage of the data structure  $D$  that the tester initializes. Formally, we define the following:

**Definition 2.3** (Space function). *For a pair of error metrics  $d^K, d^N$ , let*

$$L_{d^K, d^N}(n, u, \varepsilon_K, \varepsilon_N)$$

*denote the minimum space usage  $L$  in bits, such that there exists a generalized membership tester  $\mathcal{M}$  on  $n$  keys in universe  $[u]$  achieving average error  $\varepsilon_K, \varepsilon_N$  on keys and non-keys respectively, which always initializes its data structure  $D$  to at most  $L$  bits.*

We will omit the subscript of space function  $L$  when the error metrics are clear from context.

## 2.2 Overview for convex optimization on Banach spaces

To apply Theorem 1.1, we need to find optimal distributions with minimum KL divergence subject to error constraints. This optimization is over  $\mathcal{P}([0, 1])$ , the space of all Borel probability distributions over  $[0, 1]$ . This is a convex subset of the Banach space of all signed Borel measures over  $[0, 1]$ . Note that we cannot make continuity assumption for  $\mu_1^*, \mu_0^*$  (indeed, our optimal solutions are discrete), so we cannot directly solve for their densities.

In Section 4 and Section 5, we will solve the following program.

$$\min_{\mu_1, \mu_0 \in \mathcal{P}([0,1])} \text{KL}(\mu_1 \| \mu_0), \quad \text{subject to} \quad \begin{cases} \mathbb{E}_{\hat{X} \sim \mu_1} [d^K(\hat{X})] \leq \varepsilon_K, \\ \mathbb{E}_{\hat{X} \sim \mu_0} [d^N(\hat{X})] \leq \varepsilon_N. \end{cases} \quad (1)$$

Since KL divergence is jointly convex on  $(\mu_1, \mu_0)$ , and that the feasible region is a convex set, this program is convex. The Lagrangian of this program is:

$$\mathcal{L}(\mu_1, \mu_0, \lambda_1, \lambda_0) = \text{KL}(\mu_1 \| \mu_0) + \lambda_1 (\mathbb{E}_{\mu_1} [d^K(\hat{X})] - \varepsilon_K) + \lambda_0 (\mathbb{E}_{\mu_0} [d^N(\hat{X})] - \varepsilon_N).$$

Now we describe the general method we follow for solving this optimization problem in both the filter and LLM settings.

**Step 0: KKT point gives tight error metrics.** We check for Slater's condition: the feasible region clearly contains a relative interior, as we can choose  $\mu_1 = \delta_1$  and  $\mu_0 = \delta_0$  to achieve zero error. Therefore, the KKT conditions are necessary and sufficient for optimal solution. Next, notice the  $d^K$  constraint only depends on  $\mu_1$ , and the  $d^N$  constraint only depends on  $\mu_0$ . Unless we are in the trivial regime where  $\text{KL}(\mu_1^* \| \mu_0^*) = 0$ , neither  $\mu_0$  nor  $\mu_1$  are free to optimize over the entire  $\mathcal{P}([0, 1])$ ; otherwise, the two distributions will always tend to become equal. We therefore conclude that both error metrics are *tight* at the KKT point, and their corresponding multipliers must have  $\lambda_1 > 0, \lambda_0 > 0$ .

**Step 1: Stationary condition for  $\mu_1$ .** Using Donsker-Varadhan variational formula (Theorem C.8), we can write out the closed-form expression for the optimal distribution  $\mu_1^*$  that minimizes the Lagrangian, in terms of  $\mu_0^*, \lambda_1, \lambda_0$ .

**Step 2: Stationary condition for  $\mu_0$ .** Once we plug in the above closed-form, the Lagrangian  $\mathcal{L}$  is now a functional of  $\mu_0$  for each choice of  $\lambda_1, \lambda_0$ . We call this new functional  $J(\mu_0)$ . To determine the stationary point, we take the Gâteaux derivative (definition C.5) of  $J$  with respect to changes in  $\mu_0$ . The global optimal solution for a convex functional is characterized by the property that the derivative is nonnegative in *all directions*, as described in Theorem C.6.

In both cases, the Gâteaux derivative, which is a functional over the measures, has the *representation* of integrating a continuous function  $g$  over the signed measure  $R - \mu_0$ :

$$\delta J(\mu_0; R - \mu_0) = \int g(x) dR(x) - \int g(x) d\mu_0(x), \quad \forall R \in \mathcal{P}([0, 1]).$$

It then follows that, for this derivative to be nonnegative in all directions,  $\mu_0^*$  must be *supported on the global minima* of  $g(x)$ .

**Step 3: Solving for support and probability mass.** The derivative representation  $g(x)$  is a function of  $x$  on  $[0, 1]$  parametrized by  $\lambda_1, \lambda_0$ . In this step we deduce the likely range of parameters  $\lambda_1$  and  $\lambda_0$  in order to understand the shape of  $g(x)$ , and in both cases we conclude that  $g(x)$  has exactly two global minima for our choice of parameters.

Assuming  $\mu_0^*$  is a discrete distribution on these two minima, we can plug in the condition that the error constraints are tight, and obtain an explicit expression of  $\mu_1^*, \mu_0^*, \lambda_1, \lambda_0$ , in terms of the error rates  $\varepsilon_K, \varepsilon_N$ .

**Step 4: Verifying KKT conditions.** Finally, we verify the KKT conditions. Stationary conditions are automatically satisfied, and we mainly need to verify that  $\lambda_1, \lambda_0$  (and therefore  $g(x)$ ) are actually in the form that what we assumed earlier. This certifies that our solutions are optimal.

### 3 A Rate-Distortion Theory for Generalized Membership Testers

In this section, we state and prove our main theorem, which characterizes the space-error trade-off for generalized membership testers with custom error metrics and error rates, when the keys are sparse in the universe.

**Theorem 3.1.** *Fix any  $\varepsilon_K, \varepsilon_N \in (0, 1)$  and  $p = \frac{n}{u} \in (0, 1)$ . Given per-key error metric  $d^K$  and per-non-key error metric  $d^N$ , suppose  $\mu_1^*, \mu_0^*$  are the unique minimizers of  $\text{KL}(\mu_1 \| \mu_0)$  over the set of distributions satisfying the constraints:*

$$\mathbb{E}_{\hat{X} \sim \mu_1} [d^K(\hat{X})] \leq \varepsilon_K, \text{ and } \mathbb{E}_{\hat{X} \sim \mu_0} [d^N(\hat{X})] \leq \varepsilon_N.$$

*Then, the minimum asymptotic space usage per key, as  $n \rightarrow \infty$ , is characterized by:*

$$\lim_{n \rightarrow \infty} \frac{1}{n} L_{d^K, d^N}(n, n/p, \varepsilon_K, \varepsilon_N) = \text{KL}(\mu_1^* \| \mu_0^*) - \frac{1}{2 \ln 2} \chi^2(\mu_1^* \| \mu_0^*) \cdot p + o(p).$$

The emergence of KL-divergence is clean and natural: from a hypothesis testing perspective, this KL term intuitively represents the “information gain” needed per key to distinguish the key’s output distribution ( $\mu_1^*$ ) from the background non-key distribution ( $\mu_0^*$ ). Our theorem can also be viewed as the natural rate-distortion counterpart for the problem of sparse set membership, where the fundamental quantity characterizing the necessary “information” is KL divergence instead of mutual information.

The rest of the section is organized as follows. In Section 3.1, we define key concepts in the classical rate-distortion theory and extend it to work with two distortion measures. Here, we assume *i.i.d. membership*, where each  $e \in U$  independently has probability  $p$  of being in  $S$ . Then, in Section 3.2, we show that lower bounds from i.i.d. membership setting can naturally extend to the setting for our membership testers, where  $S$  is a set of fixed size. Finally, in Section 3.3, we prove the main theorem by studying the rate-distortion function as  $p \rightarrow 0$ .

#### 3.1 Rate-Distortion theory for i.i.d. membership information

First, we formalize the *lossy compression* problem (see, e.g., Chapter 10 of [16]) that is closely related to the problem of generalized membership testing. As discussed, we consider the *i.i.d. membership* setting: suppose each  $e \in U$  has a fixed  $p = \frac{n}{u}$  probability of being a key in  $S$ , independent from all other elements:

1. Alice observes membership information  $X_i = \mathbb{1}\{i \in S\}$  for  $i \in [u] = U$ . For each  $i$ , this is an independent  $\text{Bern}(p)$  random variable.
2. Alice and Bob agree on a membership tester. Alice will initialize the tester using  $S$  and send its data structure  $D$  to Bob, who will query every element in  $U$  to decode the membership information  $\hat{X}_i$  for  $i \in [u]$ .
3. The **distortion** of this protocol is measured by the average of a **distortion measure**  $d$  over all  $(X_i, \hat{X}_i)$  pairs:

$$\mathbb{E} \left[ \frac{1}{u} \sum_{i=1}^u d(X_i, \hat{X}_i) \right].$$

In the rate-distortion theory view, the bits transmitted by Alice is the output of a randomized encoding function  $f_u : \{0, 1\}^u \rightarrow \{0, 1\}^{uR}$ , where  $R$  is the **rate** of the code, i.e., the number of bits transmitted per element. Bob’s randomized decoding function  $g_u : \{0, 1\}^{uR} \rightarrow [0, 1]^u$  is then used to recover the approximations of the membership information  $\{\hat{X}_i\}_{i=1}^u$ .

We make the following adjustments to suit our purpose. First, we use *two* distortion measures to constraint on both error metrics  $d^K$  and  $d^N$  simultaneously. Specifically, the two distortion measures are just  $d^K$  and  $d^N$  with the proper scaling with respect to  $p$ . They should capture the fact that we average the errors over keys/non-keys, instead of over the entire  $U$ . Since the expected density of keys and non-keys are  $p$  and  $1 - p$  respectively in the universe, we define:

$$d_p^K(x, \hat{x}) = \frac{1}{p} d^K(\hat{x}) \mathbb{1}\{x = 1\} \quad \text{and} \quad d_p^N(x, \hat{x}) = \frac{1}{1-p} d^N(\hat{x}) \mathbb{1}\{x = 0\}.$$

Now, given any distortion measures  $d_p^K$  and  $d_p^N$  which satisfy the scaling requirements, we can state and extend the classical results on rate-distortion trade-off, assuming the membership information on  $U$  consists of  $u$  copies of  $\text{Bern}(p)$  random variables.

There are two ways of characterizing the rate-distortion trade-off under a certain distortion measure, and the central result in this field is that, as  $n \rightarrow \infty$  they are equal.

**Definition 3.2.** For any  $u \in \mathbb{N}$  and  $p \in (0, 1)$ , the **rate-distortion function**  $R_{p,u}(\varepsilon_K, \varepsilon_N)$  is the minimum per-element rate for codes with expected distortion at most  $(\varepsilon_K, \varepsilon_N)$ :

$$R_{p,u}(\varepsilon_K, \varepsilon_N) := \min \left\{ R : \exists f_u, g_u \text{ of rate } R \text{ such that } \mathbb{E} \left[ \frac{1}{u} \sum_{i=1}^u d_p^K(X_i, \hat{X}_i) \right] \leq \varepsilon_K, \right. \\ \left. \mathbb{E} \left[ \frac{1}{u} \sum_{i=1}^u d_p^N(X_i, \hat{X}_i) \right] \leq \varepsilon_N \right\}.$$

Meanwhile, for each  $p$ , the **information rate-distortion function**  $R_p^{(I)}(\varepsilon_K, \varepsilon_N)$  is:

$$R_p^{(I)}(\varepsilon_K, \varepsilon_N) := \min \{ I(X; \hat{X}) : X \sim \text{Bern}(p), \mathbb{E}[d_p^K(X, \hat{X})] \leq \varepsilon_K, \mathbb{E}[d_p^N(X, \hat{X})] \leq \varepsilon_N \},$$

where the minimum is taken over all joint distributions of  $(X, \hat{X})$ .

For our purpose, it suffices to prove that  $R_{p,u}(\varepsilon_K, \varepsilon_N)$  is at least as large as  $R_p^{(I)}(\varepsilon_K, \varepsilon_N)$ . The proof can be found in Section C.1.

**Lemma 3.3.** For all  $p \in (0, 1)$  and  $u \in \mathbb{N}$ , we have:

$$R_{p,u}(\varepsilon_K, \varepsilon_N) \geq R_p^{(I)}(\varepsilon_K, \varepsilon_N).$$

### 3.2 From i.i.d. membership to fixed key size

The number of keys  $|S|$  follows a  $\text{Binom}(u, p)$  distribution under the i.i.d. setting, which is close to a  $\text{Poisson}(n)$  distribution instead of a fixed size  $n$  when  $n$  is fixed and  $u$  is large. However, we can show that this difference is negligible if  $n \rightarrow \infty$  and  $p$  is fixed, by applying the law of large numbers: when  $p$  is fixed and  $n, u \rightarrow \infty$ , with high probability we would have  $|S| = (1 + o(1))n$ .

**Lemma 3.4.** Fix any pair of error metrics  $d^K$  and  $d^N$ . For all  $\varepsilon^K, \varepsilon^N \geq 0$ , and for all fixed rational  $p = \frac{n}{u} \in (0, 1)$ , the space function for generalized membership testers must satisfy:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} L_{d^K, d^N}(n, u, \varepsilon_K, \varepsilon_N) \geq \frac{1}{p} R_p^{(I)}(\varepsilon_K, \varepsilon_N).$$

*Proof.* Let  $n_1, u_1$  be coprime integers such that  $\frac{n_1}{u_1} = p$ , and define sequences  $\{n_j\}, \{u_j\}$  to be  $n_j = jn_1$  and  $u_j = ju_1$ , for all  $j \in \mathbb{N}$ . Let  $S_j$  be the set of keys generated by the i.i.d. membership process. Clearly,  $|S_j| \sim \text{Binom}(u_j, p)$ , with  $\text{var}(|S_j|/n_j) = O(1/n_j)$ . By Chebyshev's inequality, there is a sequence of real numbers  $\{\delta_j\}_{j \in \mathbb{N}}$  such that  $\delta_j = \Theta(n_j^{-1/3})$  and for each  $j$  we have

$$\mathbb{P}[|S_j - n_j| \geq \delta_j n_j] \leq \frac{\text{var}(|S_j|/n_j)}{\delta_j^2} = O(n_j^{-1/3}) = O(\delta_j).$$

Then, using a filter of appropriate size, we can construct a protocol for transmitting the membership information of the universe  $[u_j]$ :

1. If  $|S_j - n_j| \leq \delta_j n_j$ , then we build a membership tester of  $n_j$  elements with expected error rates  $\varepsilon_K$  and  $\varepsilon_N$ , using at most  $L(n_j, u_j, \varepsilon_K, \varepsilon_N)$  bits. When  $|S_j| \neq n_j$ , we will either drop elements, or add dummy elements to the filter, which we can remove during decoding. We use  $O(\delta_j n_j) \cdot \log u_j$  bits to explicitly specify the exact elements we dropped or the dummy elements added.
2. If  $|S_j| > (1 + \delta_j)n_j$  or  $|S_j| < (1 - \delta_j)n_j$ , then we send a dummy message indicating  $\hat{X}_i = c$  for all  $i \in [u_j]$ . Here,  $c \in [0, 1]$  is the value where  $d^K(c) < C$  and  $d^N(c) < C$ , as required by our definition of error metric.

Note that in case 2, both expected distortions measured by  $d_p^K$  and  $d_p^N$  are at most  $C$ . For case 1, the expected  $d_p^K$  distortion is at most  $\varepsilon_K$  and the expected  $d_p^N$  distortion is at most  $\varepsilon_N$ , since the explicit encoding can only decrease the error rates, as perfect answers result in both errors being zero. Thus, we can bound the expected distortions of this protocol by

$$\begin{cases} \mathbb{E} \left[ \frac{1}{u_j} \sum_{i=1}^{u_j} d_p^K(X_i, \hat{X}_i) \right] \leq (1 - \delta_j)\varepsilon_K + \delta_j C &=: a_j, \\ \mathbb{E} \left[ \frac{1}{u_j} \sum_{i=1}^{u_j} d_p^N(X_i, \hat{X}_i) \right] \leq (1 - \delta_j)\varepsilon_N + \delta_j C &=: b_j, \end{cases}$$

where  $a_j \rightarrow \varepsilon_K$  and  $b_j \rightarrow \varepsilon_N$  as  $j \rightarrow \infty$ .

Note that, the total number of bits used by this protocol is at most  $L(n_j, u_j, \varepsilon_K, \varepsilon_N) + O(\delta_j n_j \log u_j)$ . By Theorem 3.3, we have:

$$\frac{1}{u_j} L(n_j, u_j, \varepsilon_K, \varepsilon_N) \geq R_p^{(I)}(a_j, b_j) - \frac{1}{u_j} O(\delta_j n_j \log u_j), \text{ for all } j \in \mathbb{N}.$$

As  $j \rightarrow \infty$ , we have  $\delta_j n_j \log u_j = o(u_j)$ , so the claim follows.  $\square$

We provide a matching upper bound to Theorem 3.4, establishing that the rate-distortion function exactly characterizes the fundamental limit for generalized membership testers with fixed key size. We show the full proof in Section C.2. This proof also follows a standard strategy in rate-distortion theory.

**Theorem 3.5.** *Fix any pair of error metrics  $d^K$  and  $d^N$ . For all  $\varepsilon_K, \varepsilon_N \geq 0$ , and for all fixed rational  $p = \frac{n}{u} \in (0, 1)$ , the space function for generalized membership testers satisfies:*

$$\limsup_{n \rightarrow \infty} \frac{1}{n} L_{d^K, d^N}(n, u, \varepsilon_K, \varepsilon_N) \leq \frac{1}{p} R_p^{(I)}(\varepsilon_K, \varepsilon_N).$$

Together, we now have:

$$\lim_{n \rightarrow \infty} \frac{1}{n} L_{d^K, d^N}(n, u, \varepsilon_K, \varepsilon_N) = \frac{1}{p} R_p^{(I)}(\varepsilon_K, \varepsilon_N).$$

### 3.3 Combining the results

Now we are ready to prove the space lower bound for approximate membership structures. First, to understand the asymptotic behavior of  $\frac{1}{p} R_p^{(I)}(\varepsilon_K, \varepsilon_N)$  at the limit of  $p \rightarrow 0$ , we prove the following lemma about mutual information.

**Lemma 3.6.** *Fix any  $\varepsilon_K, \varepsilon_N \geq 0$ . For any pair of conditional distributions  $(\mu_1, \mu_0)$  that are independent of  $p$ , let  $F_p(\mu_1, \mu_0) = \frac{1}{p} I(X; \hat{X})$ , where  $X \sim \text{Bern}(p)$ ,  $\hat{X}|(X = 1) \sim \mu_1$ , and  $\hat{X}|(X = 0) \sim \mu_0$ . Then:*

1.  $\lim_{p \rightarrow 0} F_p(\mu_1, \mu_0) = \text{KL}(\mu_1 \| \mu_0)$ , and
2.  $\lim_{p \rightarrow 0} \frac{\partial}{\partial p} F_p(\mu_1, \mu_0) = -\frac{1}{2 \ln 2} \chi^2(\mu_1 \| \mu_0)$ .

*Proof.* See Section C.3. □

To prove the main theorem, we need to consider the behavior of  $R_p^{(I)}(\varepsilon_K, \varepsilon_N)$  as  $p \rightarrow 0$ . Since this is the objective function of an optimization problem over  $F_p(\mu_1, \mu_0)$ , we need the following envelope theorem:

**Lemma 3.7** (Corollary 4 of [45]). *Let  $f(p, \mu)$  be a function of two variables, defined for  $p \in [0, 1]$  and  $\mu \in \mathcal{X}$ , where  $\mathcal{X}$  is a nonempty compact topological space. Let  $V(p) := \inf_{\mu \in \mathcal{X}} f(p, \mu)$ . Let  $\mathcal{C}^*(p) \subseteq \mathcal{X}$  be the set of  $\mu$  that achieves this infimum for each  $p$ .*

*If  $f(p, \mu)$  is lower semi-continuous in  $\mu$  for each  $p$ , and  $\frac{\partial}{\partial p} f(p, \mu)$  exists and is jointly continuous in  $(\mu, p)$ , then  $\mathcal{C}^*(p)$  is always non-empty, and:*

$$V'(p+) := \lim_{h \rightarrow 0+} \frac{V(p+h) - V(p)}{h} = \min_{\mu \in \mathcal{C}^*(p)} \frac{\partial}{\partial p} f(p, \mu) \text{ for all } p \in [0, 1).$$

Now we are ready to state and prove the main theorem.

*Proof of theorem 3.1.* We omit the subscript in  $L$  for convenience. We will start with:

$$\lim_{n \rightarrow \infty} \frac{1}{n} L(n, n/p, \varepsilon_K, \varepsilon_N) = \frac{1}{p} R_p^{(I)}(\varepsilon_K, \varepsilon_N).$$

Note that  $\frac{1}{p} R_p^{(I)}(\varepsilon_K, \varepsilon_N) = \min_{(\mu_1, \mu_0) \in \mathcal{C}} F_p(\mu_1, \mu_0)$ , where  $\mathcal{C}$  is the feasible set from the theorem statement.  $\mathcal{C}$  is independent from the choice of  $p$ , given the scaling requirement on the distortion measures with respect to  $p$ . We are interested in the asymptotic behavior of this minimum, as  $p \rightarrow 0$ .

Let  $F_0(\mu_1, \mu_0) = \text{KL}(\mu_1 \| \mu_0)$ . We want to apply the envelope theorem with  $F_p(\mu_1, \mu_0)$  in place of  $f(p, \mu)$ . By extending the definition to include  $p = 0$ ,  $F$  satisfies the continuity and differentiability conditions of Theorem 3.7 by the continuity of mutual information and KL divergence. We also note that the space of distributions is compact under the weak topology, and the feasible region  $\mathcal{C}$  is also compact by the lower semi-continuity of  $d^K$  and  $d^N$ . Under the assumption that optimal solutions  $\mu_1^*, \mu_0^*$  are unique, there is only one minimizer  $(\mu_1^*, \mu_0^*)$  in the  $\min_{\mu \in \mathcal{C}^*(p)}$  operator in the envelope theorem, so we can conclude that:

$$\frac{\partial}{\partial p} \left[ \liminf_{n \rightarrow \infty} \frac{1}{n} L(n, n/p, \varepsilon_K, \varepsilon_N) \right] \Big|_{p=0} = \frac{\partial}{\partial p} F_p(\mu_1^*, \mu_0^*) \Big|_{p=0} = -\frac{1}{2 \ln 2} \chi^2(\mu_1^* \| \mu_0^*),$$

from which the theorem follows. □

## 4 Hallucination in LLMs

As discussed in introduction, we apply the same reduction as Kalai et al.[31, 30], by reducing the generative hallucination problem to the hallucination detection problem, or LLM decision. In this section, we will first justify the treatment of LLMs as membership testers, and then state and prove our main result.

We view LLM decision as an instance of membership testing, where we take  $S$  to be the set of *facts*,  $U$  the set of potential facts or *factoids*, and  $U \setminus S$  are the *non-facts*. An LLM is then trained on the facts and non-facts, with the goal of achieving expected loss  $\varepsilon_K$  on facts and  $\varepsilon_N$  on non-facts. We will assume  $|S| \ll |U|$ , and that  $S$  is a random subset of  $U$  of a fixed size. Before stating the main theorem, we first discuss the validity of this view point.

**Number of factoids.** There are often much more potential facts than the facts seen by the LLM, an phenomenon observed in prior work [31]. Plausible-looking made-up citations, for example, vastly outnumber the already enormous set of existing research papers.

An extreme setting is to take  $U$  as the set of all potential inputs. In this case,  $|U|$  is exponential in context length, and most factoids are lengthy gibberish, which clearly results in  $|U| \gg |S|$ . But even this view is not completely vacuous: using LLM to distinguish facts from gibberish, or the more general OOD detection, is itself a core challenge in many research areas [15, 60].

**Arbitrariness of  $S$ .** To apply the membership tester result, we require that for all factoids  $U$ , LLM should be able to train on any subset  $S$  of  $n$  facts, and achieve the same target loss on facts and non-facts. This assumption is natural for plausible-sounding factoids  $U$ , as there is no reason to prefer one factoid over another, when they sound equally likely on the surface.

In the view where  $U$  is all possible inputs, the models used in practice admittedly have strong prior of certain input structures over others. This is why the numerical space lower bound in the extreme example (appendix A) does not apply to real-world models per se. In this scenario, our bound highlights the importance of model/optimizer designs that induce an “implicit bias” [2, 54] towards learning real-world data over random data.

**Main theorem.** Now we state the main theorem and prove it by solving the convex optimization. Recall that the error metrics or *loss* for LLM are  $d^K(\hat{x}) = -\ln \hat{x}$  and  $d^N(\hat{x}) = -\ln(1 - \hat{x})$ . It is easy to check that they satisfy our requirements on error metrics.

We are interested in the regime where achieving the desired error rates requires a non-trivial structure, i.e., where the minimum KL divergence is strictly positive. This occurs when  $e^{-\varepsilon_K} + e^{-\varepsilon_N} > 1$ . If this condition is not met, one can find a single distribution  $\mu_1^* = \mu_0^*$  (e.g., a Dirac mass  $\delta_x$  where  $e^{-\varepsilon_K} \leq x \leq 1 - e^{-\varepsilon_N}$ ) that satisfies both constraints simultaneously, resulting in a zero space lower bound.

Our main result about LLMs is the following.

**Theorem 4.1.** *In the LLM decision problem, consider an LLM capable of achieving average loss  $\varepsilon_K \geq 0$  on facts and average loss  $\varepsilon_N \geq 0$  on non-facts for all  $S \subseteq U$ , and suppose  $e^{-\varepsilon_K} + e^{-\varepsilon_N} > 1$ . Suppose there are  $u = \frac{n}{p}$  factoids for some  $p \ll 1$ , and the LLM is trained on  $n$  facts. Then the asymptotic space usage of the LLM is at least*

$$\text{KL}(\mu_1^* \parallel \mu_0^*) - \Theta_{\varepsilon_K, \varepsilon_N}(p)$$

bits per fact, where the uniquely optimal distributions  $\mu_1^*, \mu_0^*$  are:

$$\mu_1^* = \delta_{x^*}, \quad \mu_0^* = (1 - q^*)\delta_0 + q^*\delta_{x^*}.$$

Here,  $\delta_x$  denotes the Dirac measure at  $x$ , and the parameters are  $x^* = e^{-\varepsilon_K}$  and  $q^* = \frac{\varepsilon_N}{-\ln(1-x^*)}$ .

We call  $q^*$  the **hallucination probability**, and the leading KL divergence of the space lower bound is characterized by  $q^*$ :

$$\text{KL}(\mu_1^* \parallel \mu_0^*) = -\log q^* = \log(-\ln(1 - e^{-\varepsilon_K})) - \log(\varepsilon_N).$$

We give a graphical illustration of the optimal distributions for LLMs in Figure 1.

The behavior of this bound reveals an interesting asymmetry in the difficulty of learning keys versus non-keys under log-loss.

**Corollary 4.2.** *The leading term of the space lower bound in Theorem 4.1 exhibits the following asymptotic behavior:*

1. *If  $\varepsilon_K$  is fixed and  $\varepsilon_N \rightarrow 0$ , the per-key space bound grows as  $\Theta(\log(1/\varepsilon_N))$ .*

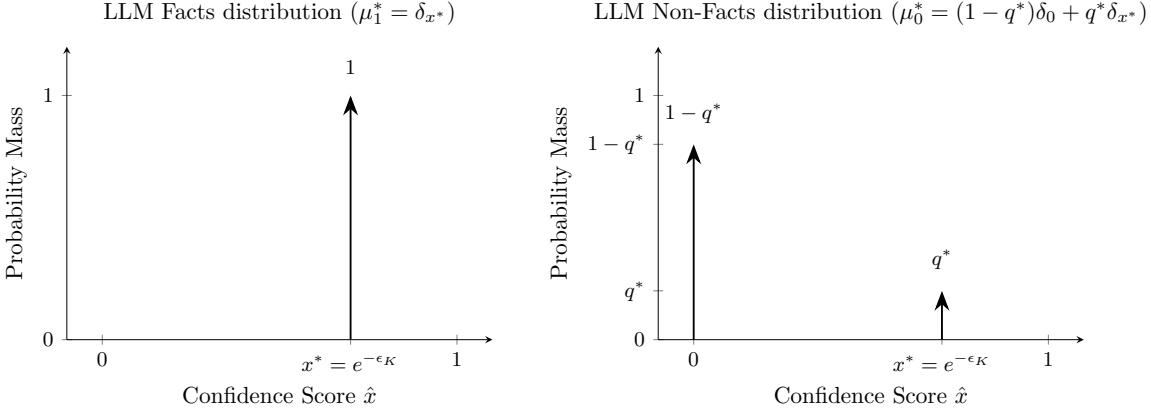


Figure 1: Space-optimal distributions for LLM output

2. If  $\varepsilon_N$  is fixed and  $\varepsilon_K \rightarrow 0$ , the per-key space bound grows as  $\Theta(\log \log(1/\varepsilon_K))$ .

*Proof.*  $L(\varepsilon_K, \varepsilon_N) = \log(-\ln(1 - e^{-\varepsilon_K})) - \log(\varepsilon_N)$ . For item 1, the first term is constant, so the growth is dominated by  $-\log(\varepsilon_N)$ . For item 2, as  $\varepsilon_K \rightarrow 0$ , we use the Taylor expansion  $e^{-\varepsilon_K} \approx 1 - \varepsilon_K$ . Then  $-\ln(1 - e^{-\varepsilon_K}) \approx -\ln(\varepsilon_K) = \log(1/\varepsilon_K)$ . Thus, the first term grows as  $\log(\log(1/\varepsilon_K))$ .  $\square$

Despite the seemingly benevolent  $\log(1/\varepsilon_N)$  growth, in A we demonstrate an extreme case where this can still be catastrophic.

#### 4.1 Solving the Convex Optimization

Our proof is an application of Theorem 3.1, for which we need to solve the following convex optimization.

$$\min_{\mu_1, \mu_0 \in \mathcal{P}([0,1])} \text{KL}(\mu_1 \| \mu_0), \quad \text{subject to} \quad \begin{cases} \mathbb{E}_{X \sim \mu_1}[-\ln X] \leq \varepsilon_K, \\ \mathbb{E}_{X \sim \mu_0}[-\ln(1 - X)] \leq \varepsilon_N. \end{cases} \quad (2)$$

We now solve the optimization problem (2) via the procedure described in Section 2.2. The Lagrangian is:

$$\mathcal{L}(\mu_1, \mu_0, \lambda_1, \lambda_0) = \text{KL}(\mu_1 \| \mu_0) + \lambda_1 \left( \mathbb{E}_{\mu_1}[-\ln X] - \varepsilon_K \right) + \lambda_0 \left( \mathbb{E}_{\mu_0}[-\ln(1 - X)] - \varepsilon_N \right).$$

**Step 1: stationary condition for  $\mu_1$ .** Applying the variational formula Theorem C.8 with  $h(x) = -\lambda_1 \ln X$ , the first two terms are minimized when the following inf is attained:

$$\inf_{\mu_1} \left\{ \text{KL}(\mu_1 \| \mu_0) + \mathbb{E}_{\mu_1}[\lambda_1(-\ln X)] \right\}.$$

By Theorem C.8, this is when

$$\frac{d\mu_1}{d\mu_0}(x) \propto e^{-h(x)} = x^{\lambda_1}. \quad (3)$$

Let  $C(\mu_0) = \mathbb{E}_{\mu_0}[X^{\lambda_1}]$  be the normalization constant. The value of the infimum is  $-\ln C(\mu_0)$ .

**Step 2: stationary condition for  $\mu_0$ .** Plugging the optimized  $\mu_1$  back into the Lagrangian, we obtain the dual function which we seek to minimize over  $\mu_0$ :

$$J(\mu_0) = -\ln \mathbb{E}_{\mu_0}[X^{\lambda_1}] + \lambda_0 \mathbb{E}_{\mu_0}[-\ln(1-X)] - \lambda_1 \varepsilon_K - \lambda_0 \varepsilon_N.$$

$J(\mu_0)$  is a convex functional of  $\mu_0$ . We use Gâteaux derivatives  $\delta J(\mu_0^*; R - \mu_0)$  to find the optimality condition.

The derivative of the first term  $J_1(\mu_0) = -\ln \mathbb{E}_{\mu_0}[X^{\lambda_1}]$  is:

$$\delta J_1(\mu_0; R - \mu_0) = -\frac{\mathbb{E}_R[X^{\lambda_1}] - \mathbb{E}_{\mu_0}[X^{\lambda_1}]}{\mathbb{E}_{\mu_0}[X^{\lambda_1}]} = -\frac{\mathbb{E}_R[X^{\lambda_1}] - \mathbb{E}_{\mu_0}[X^{\lambda_1}]}{C(\mu_0)}.$$

The derivative of the second (linear) term  $J_2(\mu_0)$  is:

$$\delta J_2(\mu_0; R - \mu_0) = \lambda_0 \left( \mathbb{E}_R[-\ln(1-X)] - \mathbb{E}_{\mu_0}[-\ln(1-X)] \right).$$

Combining the terms,  $\delta J = \delta J_1 + \delta J_2$  can be represented in the integral form, as in Theorem C.7, by the following function  $g_{\mu_0}(x)$ :

$$g_{\mu_0}(x) = -\frac{x^{\lambda_1}}{C(\mu_0)} - \lambda_0 \ln(1-x). \quad (4)$$

By the KKT Support Condition (Theorem C.7), the optimal distribution  $\mu_0^*$  must be supported on the set of global minima of  $g_{\mu_0^*}(x)$ .

**Step 3: solving for support and probability mass.** Let  $g := g_{\mu_0^*}$  and  $C^* := C(\mu_0^*)$ . To analyze the minima of  $g$ , we now establish a crucial property of  $\lambda_1$ .

**Lemma 4.3.** *In the non-trivial regime ( $e^{-\varepsilon_K} + e^{-\varepsilon_N} > 1$ ), the optimal Lagrange multiplier satisfies  $\lambda_1 > 1$ .*

*Proof.* Consider the derivatives of  $g(x)$ :

$$g'(x) = -\frac{\lambda_1 x^{\lambda_1-1}}{C^*} + \frac{\lambda_0}{1-x}, \quad g''(x) = -\frac{\lambda_1(\lambda_1-1)x^{\lambda_1-2}}{C^*} + \frac{\lambda_0}{(1-x)^2}.$$

Note that  $C^* > 0$ , otherwise we must have  $\mu_0^* = \delta_0$ . But if  $\mu_0^*$  collapses to a point mass, then we either have  $\mu_1^* = \mu_0^*$  (infeasible solution) or  $\text{KL}(\mu_1^* \parallel \mu_0^*) = \infty$ . We will try to avoid this and solve for a  $\mu_0^*$  with a non-singleton support.

Suppose  $\lambda_1 \leq 1$ . If  $\lambda_1 = 1$ ,  $g''(x) > 0$ . If  $0 < \lambda_1 < 1$ , then  $\lambda_1 - 1 < 0$ , so the first term is positive, and  $g''(x) > 0$ . In both cases,  $g(x)$  is strictly convex and has a unique global minimum  $x^*$ . Thus,  $\mu_0^* = \delta_{x^*}$ , and hence  $\mu_1^* = \delta_{x^*}$  by the previous result on  $\frac{d\mu_1^*}{d\mu_0^*}(x)$ . This implies the KL divergence objective is zero, contradicting the assumption of the non-trivial regime.  $\square$

Knowing  $\lambda_1 > 1$ , we analyze the shape of  $g(x)$ . Since  $\lambda_1 - 1 > 0$ ,  $x^{\lambda_1-1} \rightarrow 0$  as  $x \rightarrow 0$ , so  $g'(0) = \lambda_0 > 0$ . It follows that  $g$  is increasing both as  $x \rightarrow 0$  and  $x \rightarrow 1$ . We will show that there is at most one local minimum at some  $x^* \in (0, 1)$ , so 0 must and  $x^*$  must both be global minima, with  $g(x^*) = g(0) = 0$ .

Consider stationary condition  $g'(x) = -\frac{\lambda_1 x^{\lambda_1-1}}{C^*} + \frac{\lambda_0}{1-x} = 0$ . Note that the first term is negative, the second term is positive, and  $g'(x) > 0$  iff  $\frac{\lambda_0}{1-x} > \frac{\lambda_1 x^{\lambda_1-1}}{C^*}$ , which is true iff  $C^* \frac{\lambda_0}{\lambda_1} > x^{\lambda_1-1}(1-x) =: h(x)$ . Clearly  $h(0) = h(1) = 0$ , corresponding to  $g'(0)$  being positive at these endpoints. Differentiating  $h$ , we have:

$$h'(x) = x^{\lambda_1-2}(\lambda_1 - 1 - \lambda_1 x).$$

Since  $x^{\lambda_1-2}$  is always positive on  $(0, 1)$ , this derivative only changes sign once from positive to negative, and therefore the equation  $C^* \frac{\lambda_0}{\lambda_1} = h(x)$  has at most two solutions in  $(0, 1)$ , in which case  $h(x)$  will start from zero, increase to a local maximum, and then drop back to 0 at  $x = 1$ . Hence, there are at most two points where  $g'(x) = 0$ , the first one being a local maximum and the second a local minimum, which we call  $x^*$ .

In fact, if the optimal KL divergence were to be finite, the global minimum of  $g = 0$  must be attained at both  $x = 0$  and  $x^*$ :

1. If the minimum is uniquely at  $x = 0$ . Then  $\mu_0^* = \delta_0$ . This implies  $C^* = 0^{\lambda_1} = 0$  (since  $\lambda_1 > 1$ ). This leads to  $J(\mu_0^*) = \infty$ , which is not optimal.
2. If the minimum is uniquely at  $x^*$ . Then  $\mu_0^* = \delta_{x^*}$ . This implies  $\mu_1^* = \mu_0^*$  and  $\text{KL} = 0$ , a contradiction.

We conclude that  $\mu_0^*$  is a two-point distribution:

$$\mu_0^* = (1 - q^*)\delta_0 + q^*\delta_{x^*}.$$

We can now determine  $\mu_1^*$  using the relative density in (3): because  $\frac{d\mu_1^*}{d\mu_0^*}(0) = 0^{\lambda_1} = 0$ , it follows that it is a point mass  $\mu_1^* = \delta_{x^*}$ .

Now we use the tight constraints to determine  $x^*$  and  $q^*$  for  $\mu_0^*$ .

1. Constraint on  $\mu_1^*$ :  $\mathbb{E}_{\mu_1^*}[-\ln X] = \varepsilon_K \implies -\ln(x^*) = \varepsilon_K \implies x^* = e^{-\varepsilon_K}$ .
2. Constraint on  $\mu_0^*$ :  $\mathbb{E}_{\mu_0^*}[-\ln(1 - X)] = \varepsilon_N$ .

$$(1 - q^*)(-\ln 1) + q^*(-\ln(1 - x^*)) = \varepsilon_N \implies q^* = \frac{\varepsilon_N}{-\ln(1 - x^*)}.$$

The condition  $e^{-\varepsilon_K} + e^{-\varepsilon_N} > 1$  ensures that  $\varepsilon_N < -\ln(1 - e^{-\varepsilon_K})$ , so  $0 < q^* < 1$ .

**Step 4: verifying KKT conditions.** Note that calculate we  $C^* = \mathbb{E}_{\mu_0^*}[X^{\lambda_1}] = q^*(x^*)^{\lambda_1}$  (since  $\lambda_1 > 1$ ).

Condition 1:  $g(x^*) = 0$ .

$$g(x^*) = -\frac{(x^*)^{\lambda_1}}{C^*} - \lambda_0 \ln(1 - x^*) = -\frac{1}{q^*} - \lambda_0 \ln(1 - x^*) = 0.$$

This yields  $\lambda_0 = \frac{-1}{q^* \ln(1 - x^*)}$ . Since  $q^* > 0$  and  $\ln(1 - x^*) < 0$ , we have  $\lambda_0 > 0$ .

Condition 2:  $g'(x^*) = 0$ .

$$g'(x^*) = -\frac{\lambda_1(x^*)^{\lambda_1-1}}{C^*} + \frac{\lambda_0}{1 - x^*} = 0.$$

Rearranging and substituting  $C^*$ :

$$\lambda_1 = \frac{\lambda_0 C^*}{(1 - x^*)(x^*)^{\lambda_1-1}} = \frac{\lambda_0 q^*(x^*)^{\lambda_1}}{(1 - x^*)(x^*)^{\lambda_1-1}} = \frac{\lambda_0 q^* x^*}{1 - x^*}.$$

Substituting the expression for  $\lambda_0$ :

$$\lambda_1 = \left( \frac{-1}{q^* \ln(1 - x^*)} \right) \frac{q^* x^*}{1 - x^*} = \frac{-x^*}{(1 - x^*) \ln(1 - x^*)}.$$

We verify that  $\lambda_1 > 1$ . Let  $y = 1 - x^* \in (0, 1)$ . We need to show  $\frac{-(1-y)}{y \ln y} > 1$ . Since  $y \ln y < 0$ , this is equivalent to  $-(1 - y) < y \ln y$ , or  $1 - 1/y < \ln y$ . Consider the function  $h(y) = \ln y - (1 - 1/y)$ . Its derivative is  $h'(y) = 1/y - 1/y^2 = (y - 1)/y^2$ . On  $(0, 1)$ ,  $h'(y) < 0$ . Since  $\lim_{y \rightarrow 1^-} h(y) = 0$ , we have  $h(y) > 0$  for  $y \in (0, 1)$ . Thus,  $\lambda_1 > 1$ .

Since  $\lambda_1 > 1$ , the shape analysis in Step 3 holds:  $g$  has exactly one local maximum and one local minimum in  $(0, 1)$ , confirming that  $\{0, x^*\}$  are indeed the global minima of  $g(x)$ . The KKT conditions are fully satisfied, and therefore  $\mu_1^*$  and  $\mu_0^*$  are the global optimum.

## 5 Two-Sided Filters

In this section, we apply our main theorem (Theorem 3.1) to analyze two-sided membership filters. We aim to determine the optimal distributions  $\mu_1^*$  and  $\mu_0^*$  that achieve the space lower bound under constraints on the False Negative Rate (FNR) and False Positive Rate (FPR). We will first state and prove our main result for filters, and then focus on a special class of *universe-independent filters*, for which we show matching space upper and lower bounds.

Let  $\hat{x} \in [0, 1]$  be the confidence score output by the tester, which in the case of filters can be understood as the *probability that the filter accepts a given element  $e \in U$* . Recall the error metrics  $d^K(\hat{x}) = 1 - \hat{x}$  and  $d^N(\hat{x}) = \hat{x}$ .

We are interested in the regime where the required error rates  $\varepsilon_K$  (FNR) and  $\varepsilon_N$  (FPR) necessitate a non-trivial structure. This occurs when  $\varepsilon_K + \varepsilon_N < 1$ . If  $\varepsilon_K + \varepsilon_N \geq 1$ , then we can choose a constant  $x_0 \in [1 - \varepsilon_K, \varepsilon_N]$ . The trivial tester characterized by  $\mu_1^* = \mu_0^* = \delta_{x_0}$  satisfies both constraints simultaneously, which results in  $\text{KL}(\mu_1^* \parallel \mu_0^*) = 0$ .

We prove that in the non-trivial regime, the optimal distributions are Bernoulli distributions.

**Theorem 5.1.** *For any two-sided filter with FNR  $\varepsilon_K \in [0, 1)$  and FPR  $\varepsilon_N \in [0, 1)$ , suppose  $\varepsilon_K + \varepsilon_N < 1$ . Let the key density be  $p = n/u$ . Then its asymptotic space usage as  $n \rightarrow \infty$  is at least*

$$\text{KL}(\mu_1^* \parallel \mu_0^*) - \frac{1}{2 \ln 2} \chi^2(\mu_1^* \parallel \mu_0^*) \cdot p + o(p)$$

bits per key, and the uniquely optimal distributions  $\mu_1^*$  and  $\mu_0^*$  are:

$$\mu_1^* = \text{Bern}(1 - \varepsilon_K), \quad \mu_0^* = \text{Bern}(\varepsilon_N).$$

The leading term is the binary KL divergence  $D(1 - \varepsilon_K \parallel \varepsilon_N)$ .

We give a graphical illustration of the optimal distributions for filters in Figure 2.

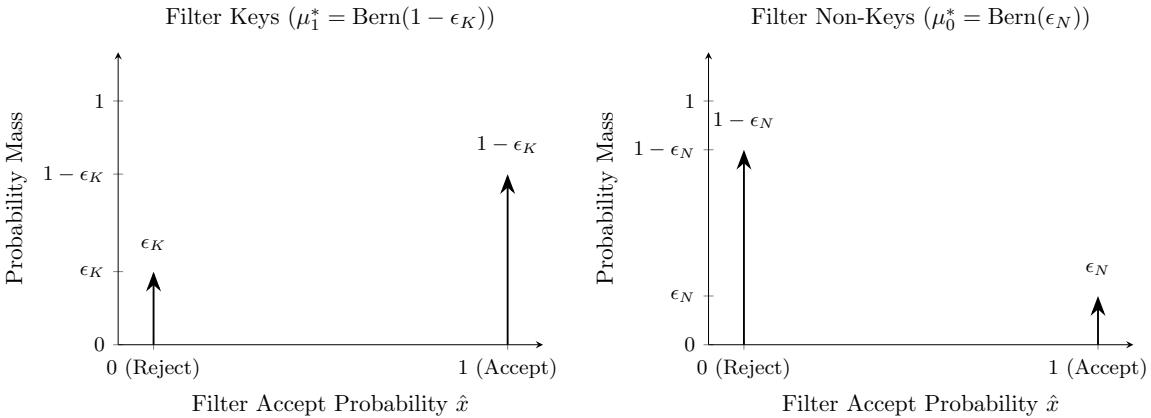


Figure 2: Space-optimal distributions for filter output

Our proof is again an application of Theorem 3.1, by solving the following program:

$$\min_{\mu_1, \mu_0 \in \mathcal{P}([0,1])} \text{KL}(\mu_1 \parallel \mu_0), \quad \text{subject to} \quad \begin{cases} \mathbb{E}_{X \sim \mu_1}[1 - X] \leq \varepsilon_K, \\ \mathbb{E}_{X \sim \mu_0}[X] \leq \varepsilon_N. \end{cases} \quad (5)$$

### 5.1 Solving the Convex Optimization

Writing out the Lagrangian:

$$\mathcal{L}(\mu_1, \mu_0, \lambda_1, \lambda_0) = \text{KL}(\mu_1 \parallel \mu_0) + \lambda_1 \left( \mathbb{E}_{\mu_1}[1 - X] - \varepsilon_K \right) + \lambda_0 \left( \mathbb{E}_{\mu_0}[X] - \varepsilon_N \right).$$

**Step 1: stationary condition for  $\mu_1$ .** Following the same variational analysis as before, we minimize the Lagrangian over  $\mu_1$  via the variational formula Theorem C.8 by taking:

$$\frac{d\mu_1}{d\mu_0}(x) \propto e^{-\lambda_1(1-x)} = e^{\lambda_1(x-1)}.$$

Plugging this back, the dual problem requires minimizing

$$J(\mu_0) = -\ln \mathbb{E}_{\mu_0}[e^{\lambda_1(X-1)}] + \lambda_0 \mathbb{E}_{\mu_0}[X]$$

over  $\mu_0$ . By the KKT Support Condition (Theorem C.7), the optimal  $\mu_0^*$  must be supported on the global minima of the function  $g(x)$ :

$$g(x) = -\frac{e^{\lambda_1(x-1)}}{C^*} + \lambda_0 x,$$

where  $C^* = \mathbb{E}_{\mu_0^*}[e^{\lambda_1(X-1)}]$ .

**Step 2: stationary condition for  $\mu_0$ .** We analyze the shape of  $g(x)$  on  $[0, 1]$ . We compute its derivatives:

$$g'(x) = -\frac{\lambda_1 e^{\lambda_1(x-1)}}{C^*} + \lambda_0, \quad g''(x) = -\frac{\lambda_1^2 e^{\lambda_1(x-1)}}{C^*}.$$

Since  $\lambda_1 > 0$  (required for the non-trivial regime) and  $C^* > 0$  (as it is an expectation of a strictly positive function), we have  $g''(x) < 0$ . Hence,  $g(x)$  is strictly concave on  $[0, 1]$ , and its global minimum must occur at  $x = 0$  or  $x = 1$  (or both).

By the KKT Support Condition, the support of  $\mu_0^*$  must be a subset of  $\{0, 1\}$ , so  $\mu_0^*$  must be a Bernoulli distribution.

**Step 3 and 4.** For the KL divergence  $\text{KL}(\mu_1 \| \mu_0)$  to be finite,  $\mu_1^*$  must also be a Bernoulli distribution. From the tight primal constraints, we have a natural choice of  $\mu_1^* = \text{Bern}(1 - \varepsilon_K)$  and  $\mu_0^* = \text{Bern}(\varepsilon_N)$ , whose optimality we can now verify using the KKT conditions.

Now we try to find  $\lambda_1 > 0, \lambda_0 > 0$  that satisfy the KKT conditions. Since  $0 < \varepsilon_N < 1$ ,  $\mu_0^*$  has support  $\{0, 1\}$ . By step 2, this implies that the global minimum of  $g(x)$  must be at both endpoints, so  $g(0) = g(1)$ .

$$\begin{aligned} C^* &= \mathbb{E}_{\mu_0^*}[e^{\lambda_1(X-1)}] = (1 - q^*)e^{-\lambda_1} + q^*e^0 = (1 - \varepsilon_N)e^{-\lambda_1} + \varepsilon_N. \\ g(0) &= -\frac{e^{-\lambda_1}}{C^*} \\ g(1) &= -\frac{1}{C^*} + \lambda_0. \end{aligned}$$

Equating  $g(0) = g(1)$  yields  $\lambda_0 = \frac{1 - e^{-\lambda_1}}{C^*}$ . Since  $C^* > 0$ , we have  $\lambda_0 > 0$  as long as  $\lambda_1 > 0$ .

The stationary condition on  $\mu_1$  must also hold. Plugging in, we have  $\mu_1^*(\{1\}) = \frac{1}{C^*} \mu_0^*(\{1\})$ . In our choice of parameters, we need  $1 - \varepsilon_K = \frac{\varepsilon_N}{C^*}$ , which gives  $C^* = \frac{\varepsilon_N}{1 - \varepsilon_K}$ . Finally, we equate the two expressions for  $C^*$  to solve for  $\lambda_1$ :

$$\begin{aligned} (1 - \varepsilon_N)e^{-\lambda_1} + \varepsilon_N &= \frac{\varepsilon_N}{1 - \varepsilon_K} \\ (1 - \varepsilon_N)e^{-\lambda_1} &= \frac{\varepsilon_N - \varepsilon_N(1 - \varepsilon_K)}{1 - \varepsilon_K} = \frac{\varepsilon_N \varepsilon_K}{1 - \varepsilon_K} \\ e^{-\lambda_1} &= \frac{\varepsilon_N \varepsilon_K}{(1 - \varepsilon_K)(1 - \varepsilon_N)}. \end{aligned}$$

This yields the explicit value for  $\lambda_1 = \ln \left( \frac{(1 - \varepsilon_K)(1 - \varepsilon_N)}{\varepsilon_N \varepsilon_K} \right)$ . This value is positive exactly when we are in the non-trivial regime  $\varepsilon_K + \varepsilon_N < 1$ . Thus  $\lambda_1 > 0$ , which also implies  $\lambda_0 > 0$ . This finishes the proof that the two Bernoulli distributions are the global optimum solution.

## 5.2 Optimal space-error trade-off for universe-independent filters

In this subsection we go further by limiting our scope to the universe-independent filters. A key property of all filters used in practice is that they are agnostic of the universe, in the sense that they operate only on the hash values of the elements in  $U$ , not  $U$  itself. This encapsulates most mainstream filters, including Bloom filter [10], variants of the Cuckoo filter [22], and variants of quotient filter [6].

For convenience of analysis, we will assume access to random oracle hash functions over  $U$  whenever we use a universe-independent filter over  $U$ . Since many filter practical designs require a different number of hash functions as  $n$  and the FPR  $\varepsilon_N$  changes, we will model the random oracles by associating random bits to every element in  $U$ .

**Definition 5.2.** Suppose each element  $e_i \in U$  is identified with an independent sequence of random bits  $b_{i,0}, b_{i,1}, \dots$ . A **universe-independent filter** is a membership tester working under the filter loss, whose initialization and query implementations are independent of the keys and query input element, and instead work with the corresponding random bits.

For universe-independent filters, we apply our main theorems to show a space lower bound of  $\text{KL}(1 - \varepsilon_K \| \varepsilon_N) + o(1)$  per element, and prove its tightness.

**Theorem 5.3.** For any universe-independent filter initialized on  $n$  keys and achieves FNR of  $\varepsilon_K$  and FPR of  $\varepsilon_N$ , its asymptotic space usage as  $n \rightarrow \infty$  is at least

$$\text{KL}\left(\text{Bern}(1 - \varepsilon_K) \| \text{Bern}(\varepsilon_N)\right)$$

bits per key.

*Proof.* Since the filter achieves the target error rates regardless of the universe size, we can use the filter as a membership tester of arbitrarily large  $U$ . For universe-independent filter initialized on  $n$  elements, its space usage is at least  $\sup_{u \geq n} L(n, u, \varepsilon_K, \varepsilon_N)$ .

Let  $L_{\text{indep}}(n, \varepsilon_K, \varepsilon_N)$  be the minimum per-key space usage for such a filter. It follows that:

$$\begin{aligned} \liminf_{n \rightarrow \infty} L_{\text{indep}}(n, \varepsilon_K, \varepsilon_N) &\geq \lim_{n \rightarrow \infty} \sup_{p \in (0,1)} L(n, n/p, \varepsilon_K, \varepsilon_N) \\ &\geq \sup_{p \in (0,1)} \lim_{n \rightarrow \infty} L(n, n/p, \varepsilon_K, \varepsilon_N) \\ &= \sup_{p \in (0,1)} \left\{ \text{KL}\left(\text{Bern}(1 - \varepsilon_K) \| \text{Bern}(\varepsilon_N)\right) - \Theta_{\varepsilon_K, \varepsilon_N}(p) \right\} \\ &= \text{KL}\left(\text{Bern}(1 - \varepsilon_K) \| \text{Bern}(\varepsilon_N)\right). \end{aligned}$$

□

**Theorem 5.4.** When  $\varepsilon_N = \frac{1}{q}$  for some prime power  $q$ , there is a two-sided filter achieving error rates  $(\varepsilon_K, \varepsilon_N)$  with probability at least 0.9, and it uses  $\text{KL}(\text{Bern}(1 - \varepsilon_K) \| \text{Bern}(\varepsilon_N)) + o(1)$  bits per key as  $n \rightarrow \infty$ .

*Proof.* Let  $n = |S|$  and let  $\mathbb{F}_q$  be the finite field with  $q = 1/\varepsilon_N$  elements. Assume we have a vector-valued random oracle function  $h : U \rightarrow \mathbb{F}_q^m$ . For each  $e \in U$ , and for each entry  $j \in [m]$ ,  $h(e)_j$  is i.i.d. uniform on  $\mathbb{F}_q$ . WLOG, let  $S = \{e_1, \dots, e_n\}$ . We consider the following linear system of equations, where  $y$  is not the all-zero vector:

$$\begin{cases} \langle h(e_1), y \rangle = 0, \\ \vdots \\ \langle h(e_n), y \rangle = 0. \end{cases}$$

where  $\langle u, v \rangle$  denotes  $\sum_{j=1}^m u_j v_j$  in  $\mathbb{F}_q$ , and  $y$  is an unknown vector in  $\mathbb{F}_q^m$ . Define  $m$  to be the smallest number such that the non-zero solution  $y$  satisfying the most equations will, with probability at least 0.9, satisfy at least a  $1 - \varepsilon_K$  fraction of the equations.

We use this system to construct a strongly optimal filter  $\mathcal{F}$ . We first describe the initialization and query process, and then prove its optimality.

**Initialization.** On input  $\varepsilon_K$ ,  $\varepsilon_N$ , and key set  $S = \{e_1, \dots, e_n\}$ , the filter stores the binary encoding of an optimal solution  $y$  to the above system of equations. This takes  $m \log q + O(1)$  bits. The construction succeeds if the stored  $y$  satisfies at least a  $1 - \varepsilon_K$  fraction of the equations.

**Query.** On query  $e \in U$ , the tester query the oracle to get  $h(e)$  and outputs 1 iff  $\langle h(e), y \rangle = 0$ . The probability of a false positive is exactly  $\frac{1}{q}$  on non-keys by the property of the random oracle.

**Analysis.** The filter satisfies the target FNR of  $\varepsilon_K$  by design, since that is the fraction of the linear equations during initialization that do not hold. For FPR, we have  $\mathbb{P}_{x \sim \text{Unif}(U \setminus S)}[\text{QUERY}(x) = 1] = \frac{1}{q}$  by the following lemma:

**Lemma 5.5.** *For any fixed  $y_1, \dots, y_m \in \mathbb{F}_q$  not all zero, and for i.i.d.  $h_1, \dots, h_m$  uniform on  $\mathbb{F}_q$ , we have:*

$$\mathbb{P}\left[\sum_{j=1}^m y_j h_j = 0\right] = \frac{1}{q}.$$

*Proof.* WLOG suppose  $y_1 \neq 0$ . Then, for any  $c \in \mathbb{F}_q$ , we have:

$$\mathbb{P}\left[\sum_{j=1}^m y_j h_j = c\right] = \mathbb{P}\left[h_1 = y_1^{-1} \left(c - \sum_{j=2}^m y_j h_j\right)\right] = \frac{1}{q}.$$

□

We now bound its space usage by showing that, for any fixed  $\delta$ , for sufficiently large  $n$ , it suffices to take  $m$  to be:

$$m = \left\lceil n \left( \frac{1}{\log q} \cdot \text{KL}(1 - \varepsilon_K \| \varepsilon_N) + \delta \right) \right\rceil$$

to ensure that the filter is strongly optimal with FNR at most  $\varepsilon_K$  with sufficient probability.

We use the second moment method to show that,  $\mathbb{P}[Z \geq 1] = 1 - o(1)$  as  $n \rightarrow \infty$ . Let  $Y$  be the set of all possible non-zero values of  $y$  with  $|Y| = q^m - 1$ , and let  $Z$  be the number of solutions  $y$  satisfying the FNR requirement. Let  $X_y$  be the indicator that a fixed non-zero  $y$  satisfies  $(1 - \varepsilon_K)n$  equations. Then, we have:

$$\begin{aligned} \mathbb{E}[Z] &= \sum_{y \in Y} \mathbb{E}[X_y] \\ &= (q^m - 1) \cdot 2^{-n \cdot \text{KL}(1 - \varepsilon_K \| \varepsilon_N) - o(n)}, \end{aligned}$$

where the last step is by Chernoff-Hoeffding theorem. We denote this probability  $\mathbb{P}[X_y = 1]$  by  $p_n$ .

To bound  $\mathbb{E}[Z^2]$ , write

$$\mathbb{E}[Z^2] = \sum_{y \in Y} \mathbb{E}[X_y] + \sum_{y \neq y'} \mathbb{E}[X_y X_{y'}].$$

Fix distinct  $y, y'$ . For each  $i \in [n]$ , the pair of events  $A_i := \{\langle h(e_i), y \rangle = 0\}$  and  $B_i := \{\langle h(e_i), y' \rangle = 0\}$  are independent whenever  $y$  and  $y'$  are linearly independent over  $\mathbb{F}_q$ ; indeed, the mapping from  $h(e_i)$  to  $(\langle h(e_i), y \rangle, \langle h(e_i), y' \rangle)$  is a linear surjection from  $\mathbb{F}_q^m$  to  $\mathbb{F}_q^2$ . The probability of each pair of values is just  $\frac{1}{q^2}$ .

Moreover, all linear functions  $\langle h(e_i), \cdot \rangle$  are i.i.d. distributed, so the set of events  $\{A_i\}$ 's (and  $\{B_i\}$ 's) are independent among themselves. It follows that the random variables  $\{\mathbb{1}\{A_i\}\} \cup \{\mathbb{1}\{B_i\}\}$  are all independent. Therefore,  $X_y$  and  $X_{y'}$  are independent.

If instead  $y'$  is a non-zero scalar multiple of  $y$  (there are exactly  $q - 2$  such  $y'$  for each fixed  $y$  when excluding  $y$  itself), then  $\langle h(e_i), y' \rangle = 0$  iff  $\langle h(e_i), y \rangle = 0$  for all  $i$ , so  $X_y = X_{y'}$  and  $\mathbb{E}[X_y X_{y'}] = \mathbb{E}[X_y] = p_n$ . Hence the second moment equals:

$$\begin{aligned} \mathbb{E}[Z^2] &= \sum_{y \in Y} \mathbb{E}[X_y] + \sum_{\substack{y \neq y' \\ y' \in \langle y \rangle \setminus \{y\}}} \mathbb{E}[X_y X_{y'}] + \sum_{\substack{y \neq y' \\ y, y' \text{ lin. indep.}}} \mathbb{E}[X_y X_{y'}] \\ &\leq (q^m - 1)p_n + (q^m - 1)(q - 2)p_n + (q^m - 1)(q^m - q)p_n^2 \\ &= (q^m - 1)((q - 1)p_n + (q^m - q)p_n^2). \end{aligned}$$

By Paley–Zygmund inequality,

$$\begin{aligned} \mathbb{P}[Z \geq 1] &\geq \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]} \\ &\geq \frac{(q^m - 1)^2 p_n^2}{(q^m - 1)((q - 1)p_n + (q^m - q)p_n^2)} \\ &= \frac{(q^m - 1)p_n}{(q - 1) + (q^m - q)p_n}. \end{aligned}$$

With our choice of  $m$ , we have  $q^m p_n = 2^{\delta n - o(n)}$  and all other terms are negligible. Hence this probability tends to 1 as  $n \rightarrow \infty$ .  $\square$

## References

- [1] Aisha Alansari and Hamzah Luqman. Large language models hallucination: A comprehensive survey, 2025.
- [2] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7411–7422, 2019.
- [3] Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction, 2025.
- [4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Madeleine Clare Elish, William Isaac, and Richard S. Zemel, editors, *FAccT ’21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM, 2021.
- [5] Michael A. Bender, Martin Farach-Colton, Mayank Goswami, Rob Johnson, Samuel McCauley, and Shikha Singh. Bloom filters, adaptivity, and the dictionary problem. In Mikkel Thorup, editor, *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 182–193. IEEE Computer Society, 2018.

- [6] Michael A. Bender, Martin Farach-Colton, Rob Johnson, Russell Kraner, Bradley C. Kuszmaul, Dzejla Medjedovic, Pablo Montes, Pradeep Shetty, Richard P. Spillane, and Erez Zadok. Don’t thrash: How to cache your hash on flash. *Proc. VLDB Endow.*, 5(11):1627–1637, 2012.
- [7] Michael A. Bender, Martin Farach-Colton, John Kuszmaul, William Kuszmaul, and Mingmou Liu. On the optimal time/space tradeoff for hash tables. In Stefano Leonardi and Anupam Gupta, editors, *STOC ’22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 1284–1297. ACM, 2022.
- [8] Ioana O. Bercea and Guy Even. A dynamic space-efficient filter with constant time operations. In Susanne Albers, editor, *17th Scandinavian Symposium and Workshops on Algorithm Theory, SWAT 2020, June 22-24, 2020, Tórshavn, Faroe Islands*, volume 162 of *LIPICS*, pages 11:1–11:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.
- [9] Léonard Blier and Yann Ollivier. The description length of deep learning models. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2220–2230, 2018.
- [10] Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, 13(7):422–426, 1970.
- [11] Alexander Dodd Breslow and Nuwan Jayasena. Morton filters: Faster, space-efficient cuckoo filters via biasing, compression, and decoupled logical sparsity. *Proc. VLDB Endow.*, 11(9):1041–1055, 2018.
- [12] Andrei Z. Broder and Michael Mitzenmacher. Survey: Network applications of bloom filters: A survey. *Internet Math.*, 1(4):485–509, 2003.
- [13] Larry Carter, Robert W. Floyd, John Gill, George Markowsky, and Mark N. Wegman. Exact and approximate membership testers. In Richard J. Lipton, Walter A. Burkhard, Walter J. Savitch, Emily P. Friedman, and Alfred V. Aho, editors, *Proceedings of the 10th Annual ACM Symposium on Theory of Computing, May 1-3, 1978, San Diego, California, USA*, pages 59–65. ACM, 1978.
- [14] Gregory J Chaitin. On the intelligibility of the universe and the notions of simplicity, complexity and irreducibility. *arXiv preprint math/0210035*, 2002.
- [15] Valeriia Cherepanova and James Zou. Talking nonsense: Probing large language models’ understanding of adversarial gibberish inputs. *CoRR*, abs/2404.17120, 2024.
- [16] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2001.
- [17] Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- [18] Martin Dietzfelbinger and Rasmus Pagh. Succinct data structures for retrieval and approximate membership (extended abstract). In Luca Aceto, Ivan Damgård, Leslie Ann Goldberg, Magnús M. Halldórsson, Anna Ingólfssdóttir, and Igor Walukiewicz, editors, *Automata, Languages and Programming, 35th International Colloquium, ICALP 2008, Reykjavík, Iceland, July 7-11, 2008, Proceedings, Part I: Track A: Algorithms, Automata, Complexity, and Games*, volume 5125 of *Lecture Notes in Computer Science*, pages 385–396. Springer, 2008.

- [19] Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneweld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1286–1305. Association for Computational Linguistics, 2021.
- [20] Benoit Donnet, Bruno Baynat, and Timur Friedman. Retouched bloom filters: allowing networked applications to trade off selected false positives against false negatives. In Christophe Diot and Mostafa H. Ammar, editors, *Proceedings of the 2006 ACM Conference on Emerging Network Experiment and Technology, CoNEXT 2006, Lisboa, Portugal, December 4-7, 2006*, page 13. ACM, 2006.
- [21] Tomer Even, Guy Even, and Adam Morrison. Prefix filter: Practically and theoretically better than bloom. *Proc. VLDB Endow.*, 15(7):1311–1323, 2022.
- [22] Bin Fan, David G. Andersen, Michael Kaminsky, and Michael Mitzenmacher. Cuckoo filter: Practically better than bloom. In Aruna Seneviratne, Christophe Diot, Jim Kurose, Augustin Chaintreau, and Luigi Rizzo, editors, *Proceedings of the 10th ACM International Conference on emerging Networking Experiments and Technologies, CoNEXT 2014, Sydney, Australia, December 2-5, 2014*, pages 75–88. ACM, 2014.
- [23] Thomas Mueller Graf and Daniel Lemire. Xor filters. *ACM J. Exp. Algorithms*, 25:1–16, 2020.
- [24] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [25] Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, and Bing Qin. Alleviating hallucinations from knowledge misalignment in large language models via selective abstention learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24564–24579. Association for Computational Linguistics, 2025.
- [26] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55, 2025.
- [27] Paul Hurley and Marcel Waldvogel. Bloom filters: One size fits all? In *32nd Annual IEEE Conference on Local Computer Networks (LCN 2007), 15-18 October 2007, Clontarf Castle, Dublin, Ireland, Proceedings*, pages 183–190. IEEE Computer Society, 2007.
- [28] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023.
- [29] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai,

- Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022.
- [30] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *CoRR*, abs/2509.04664, 2025.
  - [31] Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In Bojan Mohar, Igor Shinkar, and Ryan O’Donnell, editors, *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, pages 160–171. ACM, 2024.
  - [32] Ioannis Kazlaris, Efstathios Antoniou, Konstantinos Diamantaras, and Charalampos Bratsas. From illusion to insight: A taxonomic survey of hallucination mitigation techniques in llms. *AI*, 2025.
  - [33] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
  - [34] Denis Kleyko, Abbas Rahimi, Ross W. Gayler, and Evgeny Osipov. Autoscaling bloom filter: controlling trade-off between true and false positives. *Neural Comput. Appl.*, 32(8):3675–3684, 2020.
  - [35] Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. The case for learned index structures. In Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein, editors, *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 489–504. ACM, 2018.
  - [36] Rafael P. Laufer, Pedro B. Velloso, and Otto Carlos Muniz Bandeira Duarte. A generalized bloom filter to secure distributed network applications. *Comput. Networks*, 55(8):1804–1819, 2011.
  - [37] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation, 2023.
  - [38] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
  - [39] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In Anna Rogers, Jordan L. Boyd-Graber, and Naoki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12286–12312. Association for Computational Linguistics, 2023.

- [40] Shachar Lovett and Ely Porat. A lower bound for dynamic approximate membership data structures. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 797–804. IEEE Computer Society, 2010.
- [41] David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.
- [42] Lailong Luo, Deke Guo, Richard T. B. Ma, Ori Rottenstreich, and Xueshan Luo. Optimizing bloom filter: Challenges, solutions, and comparisons. *IEEE Commun. Surv. Tutorials*, 21(2):1912–1949, 2019.
- [43] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics, 2023.
- [44] Páll Melsted and Jonathan K. Pritchard. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinform.*, 12:333, 2011.
- [45] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.
- [46] Michael Mitzenmacher. A model for learned bloom filters and optimizing by sandwiching. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 462–471, 2018.
- [47] Negar Mosharraf, Anura P. Jayasumana, and Indrakshi Ray. Compacted bloom filter. In *2nd IEEE International Conference on Collaboration and Internet Computing, CIC 2016, Pittsburgh, PA, USA, November 1-3, 2016*, pages 304–311. IEEE Computer Society, 2016.
- [48] OpenAI. GPT-4o System Card, 2024.
- [49] Anna Pagh, Rasmus Pagh, and S. Srinivasa Rao. An optimal bloom filter replacement. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2005, Vancouver, British Columbia, Canada, January 23-25, 2005*, pages 823–829. SIAM, 2005.
- [50] Rasmus Pagh and Flemming Friche Rodler. Lossy dictionaries. In Friedhelm Meyer auf der Heide, editor, *Algorithms - ESA 2001, 9th Annual European Symposium, Aarhus, Denmark, August 28-31, 2001, Proceedings*, volume 2161 of *Lecture Notes in Computer Science*, pages 300–311. Springer, 2001.
- [51] Prashant Pandey, Alex Conway, Joe Durie, Michael A. Bender, Martin Farach-Colton, and Rob Johnson. Vector quotient filters: Overcoming the time/space trade-off in filter design. In Guoliang Li, Zhanhuai Li, Stratos Idreos, and Divesh Srivastava, editors, *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, pages 1386–1399. ACM, 2021.
- [52] Michal Perelkiewicz and Rafal Poswiata. A review of the challenges with massive web-mined corpora used in large language models pre-training. In Leszek Rutkowski, Rafal Scherer, Marcin Korytkowski, Witold Pedrycz, Ryszard Tadeusiewicz, and Jacek M. Zurada, editors, *Artificial Intelligence and Soft Computing - 23rd International Conference, ICAISC 2024*,

- [53] Ely Porat. An optimal bloom filter replacement based on matrix solving. In Anna E. Frid, Andrey Morozov, Andrey Rybalchenko, and Klaus W. Wagner, editors, *Computer Science - Theory and Applications, Fourth International Computer Science Symposium in Russia, CSR 2009, Novosibirsk, Russia, August 18-23, 2009. Proceedings*, volume 5675 of *Lecture Notes in Computer Science*, pages 263–273. Springer, 2009.
- [54] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron C. Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 2019.
- [55] Ori Rottenstreich and Isaac Keslassy. The bloom paradox: When not to use a bloom filter. *IEEE/ACM Trans. Netw.*, 23(3):703–716, 2015.
- [56] Sasu Tarkoma, Christian Esteve Rothenberg, and Eemil Lagerspetz. Theory and practice of bloom filters for distributed systems. *IEEE Commun. Surv. Tutorials*, 14(1):131–155, 2012.
- [57] Kapil Vaidya, Eric Knorr, Michael Mitzenmacher, and Tim Kraska. Partitioned learned bloom filters. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [58] Minmei Wang, Mingxun Zhou, Shouqian Shi, and Chen Qian. Vacuum filters: More space-efficient and faster replacement for bloom and cuckoo filters. *Proc. VLDB Endow.*, 13(2):197–210, 2019.
- [59] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [60] Ruiyao Xu and Kaize Ding. Large language models for anomaly and out-of-distribution detection: A survey. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5992–6012. Association for Computational Linguistics, 2025.
- [61] Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *CoRR*, abs/2401.11817, 2024.

## A Numeric illustration: space lower bound for LLM

We present an illuminating example below where even  $\log(1/\varepsilon_N)$  space growth, as in the case of LLMs, can be catastrophic in the extreme case. In particular, consider the *arbitrary parametric memory* regime, where we let  $U$  to be the set of *all possible inputs*, and we require the model to be able to memorize any  $S \subset U$  of a given size.

Let  $\mathcal{M}$  be a perfectly trained decider LLM, achieving expected loss  $\varepsilon_K$  on facts and  $\varepsilon_N$  on non-facts. Suppose  $\mathcal{M}$  has the same vocabulary and context window length as GPT-4o [48] (100,000 and 128,000, respectively), then we have  $|U| = 10^{640,000}$ . Suppose there are  $10^6$  facts, then in order to have the same number of facts and hallucinations in the universe, we need

the hallucination probability to be  $10^{-639,994}$ . This results in a space usage of  $639,994 \log 10 \approx 2.126 \times 10^6$  bits per fact, hence a total of 2.1 trillion bits.

Suppose there is a generative LLM that faithfully represents  $\mathcal{M}$ 's beliefs on what is fact. Then even in the above setting with size in the trillions, the generative model would, when asked to generate a random fact, hallucinate and output some lengthy gibberish with probability  $\frac{1}{2}$ .

**Note on practical implications.** Because the models and optimizers used in practice have implicit biases, and their training data have intrinsic structures, LLMs in practice do not face the same space lower bound as the case where  $S$  is truly an arbitrary set of inputs. However, adversarial gibberish attack against LLMs is indeed an active area of study [15].

## B Motivation for two-sided filters

Despite their success, one-sided filters sometimes have limited performance and versatility due to the stringent no-false-negative requirement. As pointed out by Hurley and Waldvogel [27], certain filter applications in networking [12] would tolerate or even *prefer* false negatives. An example is *set difference reconciliation*, where Bob sends a filter of his local storage  $S_B$  to Alice, such that Alice can compute  $S_A \setminus S_B$  and update Bob on what he does not have. Here, a false negative in Bob's filter only causes a minor overhead, yet a false positive will lead to synchronization errors. For more examples, see Table 1.

Application area	Necessity of one-sided error
Distributed caching	No
Object location in P2P systems	No
Approximate set reconciliation	False negative preferred
Resource routing	No
Loop detection	False negative preferred
Flow detection	Yes
Multicast	Yes
Hyphenation Exceptions	Yes
Set intersection	Yes
Differential files	Yes

Table 1: A summary of network applications of filters with the role of false negatives highlighted. This table is taken from Hurley and Waldvogel [27].

Moreover, even in traditionally FPR-heavy tasks, the idea of introducing FNR to further reduce FPR is often employed, as exemplified by designs such as the retouched Bloom filter [20], generalized Bloom filter [36], and autoscaling Bloom filter [34]. In *the Bloom paradox* [55], the authors also discussed *selective insertion* and *selective query* as a means to reduce FPR and introduce FNR, in order to minimize the overall cost. All these approaches can be viewed as heuristics for the FPR-FNR trade-off for a given space usage, and it is therefore of great interest to study the fundamental limit of this approach, as well as methods that can achieve this limit.

## C Missing Proofs from Section 3

### C.1 Proof for lemma 3.3

**Lemma C.1** (Same as lemma 3.3). *For all  $p \in (0, 1)$  and  $u \in \mathbb{N}$ , we have:*

$$R_{p,u}(\varepsilon^K, \varepsilon^N) \geq R_p^{(I)}(\varepsilon^K, \varepsilon^N).$$

Before the proof itself, we first show the useful fact that  $R_p^{(I)}(\varepsilon^K, \varepsilon^N)$  is jointly convex in both inputs.

**Lemma C.2.** For all  $u \in \mathbb{N}$ ,  $p \in (0, 1)$ , inputs  $\varepsilon^K, \varepsilon'^K, \varepsilon^N, \varepsilon'^N \in [0, 1]$ , and  $\lambda \in (0, 1)$ , we have:

$$R_{p,u}(\lambda\varepsilon^K + (1-\lambda)\varepsilon'^K, \lambda\varepsilon^N + (1-\lambda)\varepsilon'^N) \leq \lambda R_{p,u}(\varepsilon^K, \varepsilon^N) + (1-\lambda)R_{p,u}(\varepsilon'^K, \varepsilon'^N)$$

*Proof.* Fixing  $X$ , let  $\hat{X}$  and  $\hat{X}'$  be random variables achieving the minimum mutual information:

$$\begin{cases} I(X; \hat{X}) &= R_{p,u}(\varepsilon^K, \varepsilon^N), \\ I(X; \hat{X}') &= R_{p,u}(\varepsilon'^K, \varepsilon'^N). \end{cases}$$

Now consider random variable  $\hat{X}_\lambda$  defined by:

$$\hat{X}_\lambda = \begin{cases} \hat{X} & \text{with probability } \lambda, \\ \hat{X}' & \text{with probability } 1 - \lambda, \end{cases}$$

independently from  $X, \hat{X}, \hat{X}'$ . From linearity of expectation and the fact that  $\hat{X}, \hat{X}'$  both satisfy the distortion constraints, we have:

$$\begin{cases} \mathbb{E}[d_p^K(X, \hat{X}_\lambda)] &\leq \lambda\varepsilon^K + (1-\lambda)\varepsilon'^K, \\ \mathbb{E}[d_p^N(X, \hat{X}_\lambda)] &\leq \lambda\varepsilon^N + (1-\lambda)\varepsilon'^N. \end{cases}$$

Since  $I(X; \hat{X})$  is convex in the conditional distribution  $\hat{X} | X$ , the desired statement follows from:

$$\begin{aligned} R_{p,u}(\lambda\varepsilon^K + (1-\lambda)\varepsilon'^K, \lambda\varepsilon^N + (1-\lambda)\varepsilon'^N) &\leq I(X; \hat{X}_\lambda) \\ &\leq \lambda I(X; \hat{X}) + (1-\lambda)I(X; \hat{X}') \\ &= \lambda R_{p,u}(\varepsilon^K, \varepsilon^N) + (1-\lambda)R_{p,u}(\varepsilon'^K, \varepsilon'^N). \end{aligned}$$

□

Now we are ready for the proof. We use the following shorthand for the tuples of random variables:  $X^u = (X_1, \dots, X_u)$  and  $\hat{X}^u = (\hat{X}_1, \dots, \hat{X}_u)$ .

*Proof for Lemma 3.3.* For all  $R$  such that there exists pair of (random) functions  $f : \{0, 1\}^u \rightarrow \{0, 1\}^{uR}$  and  $g : [0, 1]^{uR} \rightarrow \{0, 1\}^u$  satisfying distortion requirements  $\varepsilon^K, \varepsilon^N$ , we must have:

$$\begin{aligned} uR &\geq H(f(X^u)) \\ &\geq I(X^u; f(X^u)) \\ &\geq I(X^u; \hat{X}^u) \\ &= H(X^u) - H(X^u | \hat{X}^u) \\ &= \sum_{i=1}^u H(X_i) - \sum_{i=1}^u H(X_i | \hat{X}^u, X_1, \dots, X_{i-1}) \\ &\geq \sum_{i=1}^u H(X_i) - \sum_{i=1}^u H(X_i | \hat{X}_i) \\ &= \sum_{i=1}^u I(X_i; \hat{X}_i) \\ &\geq \sum_{i=1}^u R(\mathbb{E}[d_p^K(X_i, \hat{X}_i)], \mathbb{E}[d_p^N(X_i, \hat{X}_i)]) \\ &\geq uR_p^{(I)} \left( \mathbb{E} \left[ \frac{1}{u} \sum_i^u d_p^K(X_i, \hat{X}_i) \right], \mathbb{E} \left[ \frac{1}{u} \sum_i^u d_p^N(X_i, \hat{X}_i) \right] \right) \text{ by convexity,} \\ &\geq uR_p^{(I)}(\varepsilon^K, \varepsilon^N). \end{aligned}$$

□

## C.2 Proof for theorem 3.5

**Theorem C.3** (Same as theorem 3.5). *Fix any pair of error metrics  $d^K$  and  $d^N$ . For all  $\varepsilon_K, \varepsilon_N \geq 0$ , and for all fixed rational  $p = \frac{n}{u} \in (0, 1)$ , the space function for membership testers satisfies:*

$$\limsup_{n \rightarrow \infty} \frac{1}{n} L_{d^K, d^N}(n, u, \varepsilon_K, \varepsilon_N) \leq \frac{1}{p} R_p^{(I)}(\varepsilon_K, \varepsilon_N).$$

To prove this theorem, we employ the method of *types* and *strong typicality*, as defined in Section 10.6 of [16].

Let  $T_n^u$  denote the *type class* of all binary sequences of length  $u$  with exactly  $n$  ones, corresponding to all possible sets  $S$  of size  $n$ . When  $p = n/u$ , sequences in  $T_n^u$  have an empirical distribution exactly equal to  $\text{Bern}(p)$ . Thus, they are strongly typical with respect to this distribution. We aim to show that a randomly generated codebook can cover this entire type class  $T_n^u$  with high probability.

*Proof of Theorem 3.5.* Let  $p = n/u$ . We aim to show that for any  $R > R_p^{(I)}(\varepsilon_K, \varepsilon_N)$ , there exists a membership tester using  $uR$  bits that satisfies the distortion constraints for every set  $S$  corresponding to  $T_n^u$ . The encoding step implements INITIALIZE, and a query  $e_i \in U$  is answered by decoding the stored index and outputting the corresponding  $\hat{X}_i$ .

Let  $p^*(x, \hat{x})$  be the joint distribution achieving the minimum in the definition of  $R_p^{(I)}(\varepsilon_K, \varepsilon_N)$ , where the marginal  $p^*(x)$  is  $\text{Bern}(p)$ . Let  $I^* = R_p^{(I)}(\varepsilon_K, \varepsilon_N)$ . Let  $p^*(\hat{x})$  be the corresponding marginal distribution for  $\hat{X}$ .

**Codebook Generation.** Fix  $\epsilon > 0$ . Fix  $R > I^*$ . Generate a random codebook  $\mathcal{C}$  consisting of  $M = 2^{uR}$  codewords  $\{\hat{X}^u(w)\}_{w=1}^M$ , drawn independently according to the product distribution  $\prod_{i=1}^u p^*(\hat{x}_i)$ .

**Encoding (Initialization).** For a given set  $S$  (represented by  $X^u \in T_n^u$ ), the encoder searches for a codeword  $\hat{X}^u(w) \in \mathcal{C}$  such that  $(X^u, \hat{X}^u(w))$  is strongly jointly  $\epsilon$ -typical with respect to  $p^*(x, \hat{x})$ . We denote the set of strongly jointly typical sequences as  $A_\epsilon^{*(u)}$ . If such a codeword is found, the index  $w$  is stored as the data structure  $D$ . If not found, an encoding failure is declared.

**Error Analysis.** If  $(X^u, \hat{X}^u) \in A_\epsilon^{*(u)}$ , the empirical distribution of the pairs  $(X_i, \hat{X}_i)$  is close to  $p^*(x, \hat{x})$ . By the definition of strong typicality and the fact that the distortions have finite expectation over  $(X_i, \hat{X}_i)$ , we can apply the law of large numbers, and conclude that the realized distortions must be close to their expectations under  $p^*$ . Specifically, there exists  $\epsilon'$  (where  $\epsilon' \rightarrow 0$  as  $\epsilon \rightarrow 0$ ) such that:

$$\begin{aligned} \frac{1}{u} \sum_{i=1}^u d_p^K(X_i, \hat{X}_i) &\leq \mathbb{E}_{p^*}[d_p^K(X, \hat{X})] + \epsilon' \leq \varepsilon_K + \epsilon', \\ \frac{1}{u} \sum_{i=1}^u d_p^N(X_i, \hat{X}_i) &\leq \mathbb{E}_{p^*}[d_p^N(X, \hat{X})] + \epsilon' \leq \varepsilon_N + \epsilon'. \end{aligned}$$

We verify that this implies the constraints for the membership tester. For a fixed set  $S$  of size  $n = up$  (vector  $X^u$  contains exactly  $n$  1's), the average error on keys is:

$$\begin{aligned} \frac{1}{n} \sum_{i \in S} d^K(\hat{X}_i) &= \frac{1}{n} \sum_{i=1}^u d^K(\hat{X}_i) \mathbb{1}\{X_i = 1\} \\ &= \frac{1}{up} \sum_{i=1}^u p \cdot d_p^K(X_i, \hat{X}_i) \\ &= \frac{1}{u} \sum_{i=1}^u d_p^K(X_i, \hat{X}_i) \leq \varepsilon_K + \epsilon'. \end{aligned}$$

Similarly, the average error on non-keys is  $\frac{1}{u-n} \sum_{i \notin S} d^N(\hat{X}_i) = \frac{1}{u} \sum_{i=1}^u d_p^N(X_i^u, \hat{X}_i) \leq \varepsilon_N + \epsilon'$ .

**Failure Probability.** We analyze the probability of encoding failure over the random choice of the codebook  $\mathcal{C}$ . We must ensure that the codebook works for all  $X^u \in T_n^u$  simultaneously.

Let  $E(X^u)$  be the event that the sequence  $X^u$  fails encoding. We fix  $X^u \in T_n^u$ . Since  $X^u$  is strongly typical with respect to  $p^*(x)$ , we can bound the probability that a single randomly drawn codeword  $\hat{X}^u$  (drawn i.i.d.  $\sim p^*(\hat{x})$ ) is jointly typical with  $X^u$ . We rely Lemma 10.6.2 of [16], which states that this probability is lower bounded as:

$$\Pr((X^u, \hat{X}^u) \in A_\epsilon^{*(u)}) \geq 2^{-u(I(X; \hat{X}) + \delta(\epsilon))},$$

where  $\delta(\epsilon) \rightarrow 0$  as  $\epsilon \rightarrow 0$  and  $u \rightarrow \infty$ . Since  $I(X; \hat{X}) = I^*$ , the probability is at least  $2^{-u(I^* + \delta(\epsilon))}$ . Let  $\delta = \delta(\epsilon)$ .

The probability of failure for this specific  $X^u$  (i.e., none of the  $M$  independent codewords are jointly typical) is:

$$\begin{aligned} \mathbb{P}(E(X^u)) &\leq \left(1 - 2^{-u(I^* + \delta)}\right)^M \\ &\leq \exp(-M \cdot 2^{-u(I^* + \delta)}) \\ &= \exp(-2^{uR} \cdot 2^{-u(I^* + \delta)}) = \exp(-2^{u(R - I^* - \delta)}). \end{aligned}$$

Now, consider the event  $E$  that there exists at least one sequence in  $T_n^u$  that fails encoding:  $E = \bigcup_{X^u \in T_n^u} E(X^u)$ . We apply the union bound. The size of the type class is bounded by  $|T_n^u| \leq \text{poly}(u) \cdot 2^{uH(p)}$  (e.g., using bounds from Chapter 11 of the same book).

$$\begin{aligned} \mathbb{P}[E] &\leq \sum_{X^u \in T_n^u} \mathbb{P}[E(X^u)] \\ &\leq |T_n^u| \cdot \exp(-2^{u(R - I^* - \delta)}) \\ &\leq \text{poly}(u) \cdot 2^{uH(p)} \cdot \exp(-2^{u(R - I^* - \delta)}). \end{aligned}$$

Since  $R > I^*$ , we can choose  $\epsilon$  small enough such that  $R - I^* - \delta > 0$ . As  $u \rightarrow \infty$ , the term  $2^{u(R - I^* - \delta)}$  grows exponentially. The double-exponential decay term  $\exp(\dots)$  dominates the  $\text{poly}(u) \cdot 2^{uH(p)}$  term. Thus,  $\mathbb{P}[E] \rightarrow 0$ .

This implies that for sufficiently large  $u$  (and thus  $n$ ), there exists at least one deterministic codebook  $\mathcal{C}^*$  such that every sequence  $X^u \in T_n^u$  can be successfully encoded. This codebook defines the membership tester.

For this tester, the distortion constraints are satisfied (up to  $\epsilon'$ ) for all sets  $S$  of size  $n$ . The space used is  $\approx uR$  bits. The space per key is  $\frac{uR}{n} = \frac{R}{p}$ . Since this holds for any  $R > R_p^{(I)}(\varepsilon_K, \varepsilon_N)$  and  $\epsilon'$  can be made arbitrarily small by choosing  $\epsilon$  small, by continuity of  $R_p^{(I)}$  we conclude:

$$\limsup_{n \rightarrow \infty} \frac{1}{n} L(n, u, \dots) \leq \frac{1}{p} R_p^{(I)}(\varepsilon_K, \varepsilon_N).$$

□

### C.3 Proof for lemma 3.6

**Lemma C.4** (Same as lemma 3.6). *Fix any  $\varepsilon_K, \varepsilon_N \in (0, 1)$ . For any pair of conditional distributions  $(\mu_1, \mu_0)$  that are independent of  $p$ , let  $F_p(\mu_1, \mu_0) = \frac{1}{p} I(X; \hat{X})$ , where  $X \sim \text{Bern}(p)$ ,  $\hat{X}|(X = 1) \sim \mu_1$ , and  $\hat{X}|(X = 0) \sim \mu_0$ . Then:*

1.  $\lim_{p \rightarrow 0} F_p(\mu_1, \mu_0) = \text{KL}(\mu_1 \| \mu_0)$ , and

$$2. \lim_{p \rightarrow 0} \frac{\partial}{\partial p} F_p(\mu_1, \mu_0) = -\frac{1}{2 \ln 2} \chi^2(\mu_1 \| \mu_0).$$

*Proof.* First note that the mutual information vanishes at  $p = 0$ , so we can apply L'Hôpital's rule to compute the first item. Let  $\mu_p = p\mu_1 + (1-p)\mu_0$ . Then, we have:

$$\begin{aligned} \lim_{p \rightarrow 0} F_p(\mu_1, \mu_0) &= \lim_{p \rightarrow 0} \frac{\partial}{\partial p} I(X; \hat{X}) \\ &= \lim_{p \rightarrow 0} \frac{\partial}{\partial p} (H(\hat{X}) - H(\hat{X}|X)) \\ &= \lim_{p \rightarrow 0} \frac{\partial}{\partial p} (H(\mu_p) - pH(\mu_1) - (1-p)H(\mu_0)) \\ &= \lim_{p \rightarrow 0} \frac{\partial}{\partial p} H(\mu_p) - H(\mu_1) + H(\mu_0). \end{aligned}$$

To compute  $\frac{d}{dp} H(\mu_p)$ , we write the entropy in the most general form. Let  $\nu$  be a measure such that  $\mu_0, \mu_1, \mu_p$  are all absolutely continuous with respect to  $\nu$ . Let  $f_0, f_1, f_p$  be the Radon-Nikodym derivatives of  $\mu_0, \mu_1, \mu_p$  with respect to  $\nu$ . Then, we have:

$$\begin{aligned} \frac{\partial}{\partial p} H(\mu_p) &= -\frac{1}{\ln 2} \frac{\partial}{\partial p} \int f_p \ln f_p d\nu \\ &= -\frac{1}{\ln 2} \int \frac{\partial}{\partial p} (f_p \ln f_p) d\nu \\ &= -\frac{1}{\ln 2} \int \left( \frac{\partial}{\partial p} f_p \right) \ln f_p + f_p \left( \frac{\partial}{\partial p} \ln f_p \right) d\nu \\ &= -\frac{1}{\ln 2} \int \left( \frac{\partial}{\partial p} f_p \right) (\ln f_p + 1) d\nu \\ &= -\frac{1}{\ln 2} \int (f_1 - f_0)(\ln f_p + 1) d\nu \\ &= -\int (f_1 - f_0) \log f_p d\nu. \end{aligned}$$

Thus, as  $p \rightarrow 0$ , we have:

$$\begin{aligned} \lim_{p \rightarrow 0} F_p(\mu_1, \mu_0) &= -H(\mu_1) + H(\mu_0) - \int (f_1 - f_0) \log f_0 d\nu. \\ &= -H(\mu_1) - \int f_1 \log f_0 d\nu. \\ &= \text{KL}(\mu_1 \| \mu_0). \end{aligned}$$

We can similarly compute the second derivative of mutual information:

$$\begin{aligned} \frac{\partial^2}{\partial p^2} I(X; \hat{X}) &= \frac{\partial}{\partial p} \left( -\frac{1}{\ln 2} \int (f_1 - f_0) \ln f_p d\nu \right) \\ &= -\frac{1}{\ln 2} \int (f_1 - f_0) \left( \frac{\partial}{\partial p} \ln f_p \right) d\nu \\ &= -\frac{1}{\ln 2} \int \frac{(f_1 - f_0)^2}{pf_1 + (1-p)f_0} d\nu. \end{aligned}$$

Hence, the second item follows from L'Hôpital's rule again:

$$\begin{aligned}
\lim_{p \rightarrow 0} \frac{\partial}{\partial p} F_p(\mu_1, \mu_0) &= \lim_{p \rightarrow 0} \frac{\partial}{\partial p} \frac{I(X; \hat{X})}{p} \\
&= \lim_{p \rightarrow 0} \frac{p \cdot \frac{\partial}{\partial p} I(X; \hat{X}) - I(X; \hat{X})}{p^2} \\
&= \lim_{p \rightarrow 0} \frac{p \cdot \frac{\partial^2}{\partial p^2} I(X; \hat{X})}{2p} \\
&= -\frac{1}{2 \ln 2} \chi^2(\mu_1 \| \mu_0).
\end{aligned}$$

□

#### C.4 Technical lemmas for optimization over measure spaces

To solve the optimization problem (2) in full generality without assuming the existence of densities, we utilize tools from variational analysis on measure spaces.

To analyze the optimality conditions for functionals over the space of measures, we use the concept of Gâteaux derivatives.

**Definition C.5** (Gâteaux Derivative). *Let  $J : \mathcal{P}([0, 1]) \rightarrow \mathbb{R}$  be a functional. The Gâteaux derivative of  $J$  at  $Q \in \mathcal{P}([0, 1])$  in the direction of  $R - Q$ , where  $R \in \mathcal{P}([0, 1])$ , is defined as:*

$$\delta J(Q; R - Q) = \lim_{\eta \rightarrow 0^+} \frac{J((1 - \eta)Q + \eta R) - J(Q)}{\eta},$$

provided the limit exists.

If  $J$  is a convex functional, the Gâteaux derivative allows us to state the necessary and sufficient conditions for global optimality.

**Theorem C.6** (First-Order Optimality Condition; see e.g. [41]). *If  $J : \mathcal{P}([0, 1]) \rightarrow \mathbb{R}$  is convex and Gâteaux differentiable, then  $Q^* \in \mathcal{P}([0, 1])$  minimizes  $J$  if and only if  $\delta J(Q^*; R - Q^*) \geq 0$  for all  $R \in \mathcal{P}([0, 1])$ .*

When the Gâteaux derivative can be represented in an integral form (and in particular linear in  $R - Q$ ), this optimality condition translates into a structural property regarding the support of the optimal measure.

**Lemma C.7** (KKT Support Condition). *Suppose the Gâteaux derivative of a convex functional  $J$  at  $Q$  can be represented as*

$$\delta J(Q; R - Q) = \int_0^1 g_Q(x) dR(x) - \int_0^1 g_Q(x) dQ(x) = \mathbb{E}_R[g_Q(X)] - \mathbb{E}_Q[g_Q(X)],$$

for some measurable function  $g_Q(x)$  (which may depend on  $Q$ ). Then  $Q^*$  minimizes  $J$  if and only if  $Q^*$  is supported on the set of global minima of the function  $g_{Q^*}(x)$ .

That is,  $x \in \text{supp}(Q^*) \implies g_{Q^*}(x) = \inf_{y \in [0, 1]} g_{Q^*}(y)$ .

*Proof.* By Theorem C.6,  $Q^*$  is optimal iff  $\mathbb{E}_R[g_{Q^*}(X)] \geq \mathbb{E}_{Q^*}[g_{Q^*}(X)]$  for all  $R$ . Let  $K^* = \mathbb{E}_{Q^*}[g_{Q^*}(X)]$ . If we take  $R = \delta_y$  for any  $y \in [0, 1]$ , we get  $g_{Q^*}(y) \geq K^*$ . Thus  $K^*$  is the global minimum value of  $g_{Q^*}(x)$ . Since  $Q^*$  is a probability measure, the equality  $\mathbb{E}_{Q^*}[g_{Q^*}(X)] = K^*$  can only hold if  $Q^*$  is supported entirely on the set where  $g_{Q^*}(x) = K^*$ . □

We also recall the following standard result regarding the minimization of KL divergence subject to linear constraints.

**Lemma C.8** (Donsker-Varadhan variational formula). *Suppose  $Q \in \mathcal{P}([0, 1])$ , and let  $h$  be a  $Q$ -measurable real function such that  $\mathbb{E}_Q[e^{-h(X)}] < \infty$ . Then,*

$$-\ln \mathbb{E}_{X \sim Q}[e^{-h(X)}] = \inf_{P \in \mathcal{P}([0, 1])} \left\{ \text{KL}(P \| Q) + \mathbb{E}_P[h(X)] \right\}.$$

The infimum is attained when  $P$  has the Radon-Nikodym derivative  $\frac{dP}{dQ}(x) \propto e^{-h(x)}$ .