Mutual Information Constrained Chernoff Bottleneck

Dier Tang[*] and Guangyue Han[†]

Department of Mathematics, The University of Hong Kong

Abstract

The classical information bottleneck (IB) measures the relevance of a representation U of X to a target Y by $I(U;Y)$, which does not directly characterize the error of downstream decisions. For a binary hypothesis Y inferred from many separately encoded observations, the optimal error exponent is the Chernoff information between the two conditional distributions of U given Y. We study the mutual information constrained Chernoff bottleneck, which seeks an encoder that maximizes this Chernoff information subject to a rate constraint $I(U;X) \le R$. We show that its optimal value $C(R)$ increases strictly up to $R = H(V)$, where V merges the symbols of X with equal likelihood ratio, remains at the uncompressed exponent beyond, and, unlike the IB curve, need not be concave. We further show that $k+1$ outputs suffice to attain $C(R)$, where k is the cardinality of V. We propose an alternating algorithm that updates the encoder via a generalized Blahut–Arimoto algorithm and the Chernoff parameter s via a nonlinear equation, and prove that its iterates remain feasible, with nondecreasing and convergent Chernoff information. Comparing with the classical IB for n samples, we show that, whenever the two objectives disagree, the IB encoder has the smaller equivocation $H(Y|U^n)$ at $n = 1$, whereas the Chernoff bottleneck has the smaller error probability and equivocation for all large n. Numerical experiments confirm the theory, and on real topic-detection data from the 20 Newsgroups corpus, compressing each word to only 17% of its entropy retains 90% of the error exponent and nearly the accuracy of the uncompressed classifier.

Keywords: Information bottleneck; Chernoff information; rate–distortion theory; cardinality bound; hypothesis testing; generalized Blahut–Arimoto algorithm

1 Introduction

Representation learning, a fundamental paradigm in machine learning [1], aims to automatically learn meaningful and compact representations of raw data that make downstream tasks—such as inference and prediction—more effective [2, 3, 4, 5, 6, 7]. The information bottleneck (IB) [8] operationalizes this goal by seeking a representation U that preserves the information relevant to a target Y while discarding redundant details of the observation X, expressed as the trade-off between maximizing $I(U;Y)$ and minimizing $I(U;X)$. Since then, the IB principle has been used both as a tool for analyzing deep neural networks and as an objective for learning robust, generalizable representations [9, 10, 11, 12, 13, 14, 15].

However, while the mutual information $I(U;Y)$ quantifies average statistical dependence, it does not directly characterize the fundamental limit on the error probability of downstream decision-making. Specifically, when $Y \in \{0,1\}$ is a binary hypothesis to be inferred from n representation samples $U_1, \dots, U_n$, obtained by applying the same encoder $P_{U|X}$ to observations $X_1, \dots, X_n$ that

[*]Email: tangde_math@connect.hku.hk.

[†]Email: ghan@hku.hk.

are i.i.d. given Y, the minimum Bayes error probability decays exponentially in n, and the optimal error exponent is the Chernoff information [16]

$$C(P_0, P_1) = \max_{0 \le s \le 1} \Big(-\log \sum_u P_0(u)^s P_1(u)^{1-s} \Big)$$

between the conditional distributions $P_{U|0}$ and $P_{U|1}$ of U given $Y = 0$ and $Y = 1$, for every prior with $P_Y(0), P_Y(1) > 0$ [17, 18]. In contrast to $I(U;Y)$, which depends on the prior, this exponent depends only on $P_{U|0}$ and $P_{U|1}$, and an encoder that is optimal for the classical IB need not be optimal for testing. This operational interpretation motivates replacing $I(U;Y)$ by the Chernoff information, thereby leading to the *mutual information (MI) constrained Chernoff bottleneck*

$$C(R) = \max_{P_{U|X}} \; C(P_{U|0}, P_{U|1}) \quad \text{s.t.} \quad I(U;X) \le R,$$

which seeks the representation under which the error of the downstream test decays fastest, subject to a budget R on its complexity. The encoder is applied to each sample separately, as when the samples are observed by different sensors or edge devices, or processed by a per-sample feature extractor before the decision is made [19, 20, 21], in contrast to testing under communication constraints, where a whole block of observations is encoded jointly [22, 23].

Although the MI constrained Chernoff bottleneck has the same form as the classical IB, the change of objective means that the theory and algorithms developed for the IB do not carry over directly. They rely on two properties of $I(U;Y)$. First, the IB loss $I(X;Y) - I(U;Y) = \mathbb{E}[\mathrm{D}(P_{Y|X} \| P_{Y|U})]$ is an expectation over (X,U) of a per-letter divergence, which leads to self-consistent equations and Blahut–Arimoto-type iterations [8, 24, 25]. Second, $I(U;Y) = H(Y) - \sum_u P_U(u) H(Y|U=u)$ is a sum over the outputs of the encoder, weighted by their probabilities, which underlies the time-sharing and cardinality arguments [26]. The Chernoff information has neither property: it is not an expectation of a per-letter quantity, and it involves an optimization over the parameter s that couples all outputs. Consequently, neither rate–distortion theory [27, 28] nor the algorithms for the classical IB apply directly, and the shape of the optimal curve, cardinality bounds on the representation, and the computation of optimal encoders must be studied anew.

In this paper, we address these problems. Our contributions are as follows.

(i) *Shape of the curve.* We prove that $C(R)$ is strictly increasing on $[0, H(V)]$ and saturates at $C(P_{X|0}, P_{X|1})$ for all $R \ge H(V)$, where V merges the symbols of X with equal likelihood ratio (Theorem 3.5), and that for $R \le H(V)$ every optimal encoder uses the full rate budget (Corollary 3.6). In contrast to the IB curve, $C(R)$ need not be concave (Example 3.7).

(ii) *Cardinality bounds.* We prove that an optimal encoder exists with at most $|\mathcal{X}| + 1$ outputs (Theorem 4.2), and sharpen the bound to $k + 1$, where k is the number of distinct likelihood ratios, by a reduction to V (Theorem 4.4).

(iii) *Computation.* We develop an algorithm (Algorithm 2) that alternately updates the encoder, by a generalized Blahut–Arimoto algorithm (Algorithm 1) with a bisection over the trade-off, and the parameter s, by solving a nonlinear equation. We prove that Algorithm 1 converges to a global minimizer at rate $O(1/t)$ with a stopping certificate (Theorem 5.8, Corollary 5.9), that the bisection has an explicit bracket and certificate (Lemma 5.10), and that the iterates of Algorithm 2 remain feasible, with s uniformly bounded away from 0 and 1 and the Chernoff information monotonically convergent (Lemma 5.12, Theorem 5.13, Remark 5.14).

(iv) *Comparison with the IB.* We show that the equivocation $H(Y|U^n)$ decays at the same exponential rate as the error probability (Lemma 6.1), so that, when the two objectives disagree, the

IB encoder, which is the best for a single sample under log-loss, is overtaken by the Chernoff bottleneck as n grows (Theorem 6.2). For the error probability, the IB need not be better even for small n (Remark 6.3).

(v) *Experiments.* On synthetic sources, the computed $C(R)$ curves confirm the theory of Section 3, and the optimal encoders are randomized likelihood-ratio quantizers (Sections 7.1 and 7.2). On real data, in topic detection on the 20 Newsgroups corpus [29], the Chernoff bottleneck compresses each word to 1 nat, only 17% of its entropy, while keeping 90% of the error exponent and nearly the accuracy of the uncompressed classifier (Section 7.3). In a comparison with the IB, the IB has the smaller error only while the error is about 0.2 or larger, and below that the Chernoff bottleneck needs 18–25% fewer samples (Section 7.4).

The rest of the paper is organized as follows. Section 2 formulates the problem. Section 3 characterizes the shape of the $C(R)$ curve, and Section 4 establishes the cardinality bounds and reformulates the problem over a finite output alphabet. Section 5 develops and analyzes the algorithm, Section 6 compares the Chernoff bottleneck with the IB for a finite number of samples, Section 7 reports the numerical experiments, and Section 8 concludes the paper.

2 Problem Setup

Let X, Y be discrete random variables with finite alphabet $\mathcal{X}$ and $\mathcal{Y} = \{0, 1\}$, and joint distribution $P_{X,Y}$ with marginals $P_X(x) > 0, P_Y(y) > 0$ for any $x \in \mathcal{X}$ and $y \in \{0, 1\}$. We derive $P_{X|Y=0}$ and $P_{X|Y=1}$ via Bayes formula, denote as $P_{X|0}$ and $P_{X|1}$ respectively. Assume there are no symbols that are "instantly recognizable", i.e.,

$$P_{X|0}(x) > 0, \ P_{X|1}(x) > 0, \ \forall x \in \mathcal{X}, \tag{1}$$

together with the nondegeneracy assumption $P_{X|0} \not\equiv P_{X|1}$.

Suppose samples $X^n = (X_1, \cdots, X_n)$ are i.i.d. drawn from either $P_{X|0}$ or $P_{X|1}$, the error exponent of using X^n to infer Y is given by the Chernoff information

$$C(P_{X|0}, P_{X|1}) = \max_{0 \le s \le 1} \Big(-\log \sum_x P_{X|0}(x)^s P_{X|1}(x)^{1-s} \Big) = -\min_{0 \le s \le 1} \log \sum_x P_{X|0}(x)^s P_{X|1}(x)^{1-s}.$$

In this paper, we set $0^0 = \lim_{s \to 0} 0^s = 0$. All logarithms are natural.

For an encoder $K = K_{U|X}$ that compresses $X \to U$ with finite output alphabet, denote by $P^K_{U|0}$, $P^K_{U|1}$ the induced conditional distributions, P^K_U the marginal, and $I_K(U; X)$ the mutual information. Suppose samples $U^n = (U_1, \cdots, U_n)$ are i.i.d. drawn from either $P^K_{U|0}$ or $P^K_{U|1}$, the error exponent of using U^n to infer Y is given by the Chernoff information $C(P^K_{U|0}, P^K_{U|1})$. Define the distortion of the encoder $K_{U|X}$ as

$$d(K) = C(P_{X|0}, P_{X|1}) - C(P^K_{U|0}, P^K_{U|1}). \tag{2}$$

The distortion $d(K)$ measures how much information of inferring Y is lost in the encoding process, and is well-defined according to the following lemma.

Lemma 2.1. *For any encoder $K_{U|X}$, $d(K) \ge 0$.*

Proof. Since $U \leftrightarrow X \leftrightarrow Y$ forms a Markov chain, the proof is straightforward by the data processing inequality of Chernoff information [30]. A direct proof via Hölder's inequality is given in Lemma 3.3. □

Given rate $R \geq 0$ and alphabet $\mathcal{U}_M = \{1, \cdots, M\}$ with finite M, define the *feasible encoder set*

$$\mathcal{K}_R^{(M)} = \Big\{ K_{U|X} : K(u|x) \geq 0, \sum_u K(u|x) = 1, \forall (u,x) \in \mathcal{U}_M \times \mathcal{X};\ I_K(U;X) \leq R \Big\}, \tag{3}$$

with its strictly positive subset

$$\mathcal{K}_{R,+}^{(M)} = \Big\{ K_{U|X} \in \mathcal{K}_R^{(M)} : K(u|x) > 0,\ \forall (u,x) \in \mathcal{U}_M \times \mathcal{X} \Big\}, \tag{4}$$

and define the *entire feasible set*

$$\mathscr{K}_R = \bigcup_{M \geq 1} \mathcal{K}_R^{(M)}. \tag{5}$$

Then, our problem can be formulated as

Problem 1. *Given $R \geq 0$, seek*

$$K^* \in \arg \inf_{K \in \mathscr{K}_R} d(K),$$

equivalently,

$$K^* \in \arg \sup_{K \in \mathscr{K}_R} C(P_{U|0}^K, P_{U|1}^K). \tag{6}$$

We refer to (6) as the *mutual information (MI) constrained Chernoff bottleneck*, for which the attainability of the optimal solutions is established in Section 4.

For rate $R \geq 0$, define the supremum in (6) as

$$C(R) \triangleq \sup_{K \in \mathscr{K}_R} C(P_{U|0}^K, P_{U|1}^K), \tag{7}$$

which is our object of study. The prior P_Y enters Problem 1 only through $P_X = P_Y(0)P_{X|0} + P_Y(1)P_{X|1}$, the distribution of the observations seen by the encoder, which defines the rate $I_K(U;X)$; the objective $C(P_{U|0}^K, P_{U|1}^K)$ depends only on $P_{X|0}$ and $P_{X|1}$. Let $F(R) \triangleq e^{-C(R)}$.

Given alphabet $\mathcal{U}_M = \{1, \cdots, M\}$ with finite M. For $K \in \mathcal{K}_R^{(M)}$ and $s \in [0,1]$, define

$$B_M(K,s) \triangleq \sum_{u \in \mathcal{U}_M} P_{U|0}^K(u)^s P_{U|1}^K(u)^{1-s}, \tag{8}$$

and for $K \in \mathcal{K}_R^{(M)}$, define the corresponding Chernoff information

$$\mathcal{C}(K) \triangleq -\log \min_{s \in [0,1]} B_M(K,s). \tag{9}$$

We abbreviate $C_X = C(P_{X|0}, P_{X|1}) > 0$, which is strictly positive since $P_{X|0} \not\equiv P_{X|1}$. Denote by

$$\ell \triangleq \min_{x \in \mathcal{X}} \frac{P_{X|1}(x)}{P_{X|0}(x)}, \quad L \triangleq \max_{x \in \mathcal{X}} \frac{P_{X|1}(x)}{P_{X|0}(x)}, \tag{10}$$

which are both well defined according to (1), and let

$$\Gamma \triangleq \max\{L, \ell^{-1}\} = \exp\Big(\max_x \Big| \log \frac{P_{X|1}(x)}{P_{X|0}(x)} \Big| \Big) > 1. \tag{11}$$

For all $x \in \mathcal{X}$, let $\rho(x) \triangleq \dfrac{P_{X|1}(x)}{P_{X|0}(x)} \in [\ell, L]$, and $\{\rho_1, \cdots, \rho_k\}$ be the distinct values taken by ρ, then $2 \le k \le |\mathcal{X}|$ (the lower bound is because $P_{X|0} \not\equiv P_{X|1}$). Define $\nu : \mathcal{X} \to \{1, \cdots, k\}$ by $\nu(x) = j$ if and only if $\rho(x) = \rho_j$, and set variable

$$V = \nu(X). \tag{12}$$

Thus V merges exactly those symbols that have the same likelihood ratio, equivalently the same posterior $P_{Y|X=x}$; it is the minimal sufficient statistic of X for Y [17], so that $I(V;Y) = I(X;Y)$. Moreover, $P_V(j) = \sum_{x:\nu(x)=j} P_X(x)$, and since V is a function of X, we have $I(V;X) = H(V) \le H(X)$, with equality if and only if all $|\mathcal{X}|$ ratios are distinct. We write ν^{-1} as the inverse mapping of ν, and define

$$\mathcal{X}_j \triangleq \nu^{-1}(j) = \{x \in \mathcal{X} : \rho(x) = \rho_j\}, \quad j = 1, \cdots, k, \tag{13}$$

for the *likelihood-ratio classes*, so that $\mathcal{X} = \bigsqcup_{j=1}^{k} \mathcal{X}_j$. Denote by K^X the identity encoder that outputs X itself (with no compression), and

$$K^V(u|x) = \mathbf{1}\{u = \nu(x)\}, \quad (u, x) \in \mathcal{U}_k \times \mathcal{X}, \tag{14}$$

the deterministic encoder that outputs V.

3 Global Shape of the $C(R)$ Curve

With $C(R)$ defined in (7), the following lemma shows some basic shape of the $C(R)$ curve.

Lemma 3.1. *(i)* $C(0) = 0$*;*

(ii) $C(R)$ *is nondecreasing on* $[0, \infty)$*;*

(iii) $0 \le C(R) \le C_X$ *for all* $R \in [0, \infty)$*.*

Proof. For (i), Markovity and $I_K(U;X) = 0$ implies $P^K_{U|0} \equiv P^K_{U|1}$. Hence, $C(0) = 0$. For (ii), monotonicity holds because $R \mapsto \mathscr{K}_R$ is nondecreasing for inclusion. For (iii), the upper bound follows from Lemma 2.1, i.e., the data processing inequality for Chernoff information [30, 31]. □

Theorem 3.5 gives a stronger result for the shape of $C(R)$, and its proof relies on the following lemmas. Throughout, $B_M(K, s)$ and V are defined in (8) and (12), respectively.

Lemma 3.2 (Sufficiency of V)**.** *Let* $|\mathcal{V}| = k$*, and* K^V *be defined in* (14)*, we have*

$$I_{K^V}(U;X) = H(V), \quad B_k(K^V, s) = B_{|\mathcal{X}|}(K^X, s), \ \forall s \in [0,1].$$

In particular, $C(P^{K^V}_{U|0}, P^{K^V}_{U|1}) = C_X$*.*

Proof. Appendix B.1. □

Lemma 3.3 (Hölder)**.** *For any encoder* $K \in \mathcal{K}^{(M)}_R$ *and* $s \in [0,1]$*,*

$$B_M(K, s) \ge B_{|\mathcal{X}|}(K^X, s). \tag{15}$$

Moreover, if $I_K(U;X) < H(V)$*, then the inequality* (15) *is strict for every* $s \in (0,1)$*.*

Proof. Appendix B.2. □

Lemma 3.4 (Time-sharing)**.** *Let K_1, K_2 be encoders with output alphabets $\mathcal{U}_{M_1}, \mathcal{U}_{M_2}$, producing U_1 and U_2 from X, respectively. For $\theta \in [0,1]$, let $Q \sim \mathrm{Bern}(\theta)$ be independent of X, and let K be the encoder with output alphabet $\mathcal{U}_{M_1+M_2}$ producing $U = U_1$ if $Q = 1$ and $U = M_1 + U_2$ if $Q = 0$. Then, for any $s \in [0,1]$,*

$$I_K(U;X) = \theta I_{K_1}(U;X) + (1-\theta) I_{K_2}(U;X), \quad B_{M_1+M_2}(K,s) = \theta B_{M_1}(K_1,s) + (1-\theta) B_{M_2}(K_2,s).$$

Proof. Appendix B.3. □

Theorem 3.5 (Saturation and strict monotonicity)**.** *$C(R)$ is strictly increasing on $[0, H(V)]$, and $C(R) = C_X$ for every $R \geq H(V)$.*

Proof. Saturation. Let $R \geq H(V)$. By Lemma 3.2, $I_{K^V}(U;X) = H(V) \leq R$, so that $K^V \in \mathscr{K}_R$, and $C(P_{U|0}^{K^V}, P_{U|1}^{K^V}) = C_X$. Hence $C(R) \geq C_X$, while Lemma 3.1 gives $C(R) \leq C_X$. Therefore $C(R) = C_X$.

Strict monotonicity. Let $0 \leq R_1 < R_2 \leq H(V)$, and let $K^* \in \mathscr{K}_{R_1}$ with output alphabet $\mathcal{U}_{M^*}$ attain $C(R_1)$, i.e., $\mathcal{C}(K^*) = C(R_1)$ (the existence of K^* is guaranteed by Theorem 4.2). Write $I^* = I_{K^*}(U;X) \leq R_1 < H(V)$.

By Young's inequality (Fact A.1), for every $s \in [0,1]$,

$$B_{M^*}(K^*, s) \leq \sum_{u \in \mathcal{U}_{M^*}} \left(s P_{U|0}^{K^*}(u) + (1-s) P_{U|1}^{K^*}(u)\right) = 1,$$

while $B_{M^*}(K^*, 0) = B_{M^*}(K^*, 1) = 1$. If $B_{M^*}(K^*, s) \equiv 1$ on $[0,1]$, we select $\bar{s} = \frac{1}{2}$. Otherwise, $\min_{s \in [0,1]} B_{M^*}(K^*, s) < 1$, then every $\bar{s} \in \arg\min_{s \in [0,1]} B_{M^*}(K^*, s)$ lies in $(0,1)$. In both cases, we can select $\bar{s} \in (0,1)$ and

$$B_{M^*}(K^*, \bar{s}) = \min_{s \in [0,1]} B_{M^*}(K^*, s) = e^{-\mathcal{C}(K^*)} = F(R_1). \tag{16}$$

Moreover, since $I^* < H(V)$, Lemma 3.3 gives

$$B_{M^*}(K^*, \bar{s}) > B_{|\mathcal{X}|}(K^X, \bar{s}). \tag{17}$$

Now let K' be the encoder with output alphabet $\mathcal{U}_{k+M^*}$ given by Lemma 3.4, with $K_1 = K^V$, $K_2 = K^*$ and

$$\theta = \frac{R_2 - I^*}{H(V) - I^*} \in (0,1].$$

By Lemma 3.4 and Lemma 3.2,

$$I_{K'}(U;X) = \theta H(V) + (1-\theta) I^* = R_2,$$

so that $K' \in \mathscr{K}_{R_2}$. Using Lemma 3.4, then Lemma 3.2, then (17) with $\theta > 0$, and finally (16),

$$\begin{aligned} F(R_2) \leq e^{-\mathcal{C}(K')} \leq B_{k+M^*}(K', \bar{s}) &= \theta B_k(K^V, \bar{s}) + (1-\theta) B_{M^*}(K^*, \bar{s}) \\ &= \theta B_{|\mathcal{X}|}(K^X, \bar{s}) + (1-\theta) B_{M^*}(K^*, \bar{s}) < B_{M^*}(K^*, \bar{s}) = F(R_1). \end{aligned}$$

Therefore, $C(R_2) > C(R_1)$, which completes the proof. □

Corollary 3.6 (The rate constraint is active)**.** *For $R \leq H(V)$, every optimal encoder of Problem 1 satisfies $I_K(U;X) = R$.*

Proof. Assume K is optimal at rate R with $I_K(U;X) < R$. Since $K \in \mathscr{K}_{I_K(U;X)}$, we have $C(I_K(U;X)) \geq C(R)$, contradicting the strict monotonicity of C on $[0, H(V)]$ established in Theorem 3.5. □

The IB curve $R \mapsto \max\{I(U;Y) : I(U;X) \leq R\}$ also reaches its maximum $I(X;Y)$ exactly at $R = H(V)$: in both problems, the maximum is attained only when U is sufficient for Y, which requires $I(U;X) \geq H(V)$. However, time-sharing guarantees the concavity of the IB curve, whereas $C(R)$ need not be concave, as the following example shows.

Example 3.7 ($C(R)$ curve not necessarily concave)**.** Let $\mathcal{X} = \{1, 2\}$, $P_Y(0) = P_Y(1) = \frac{1}{2}$, $P_{X|0} = (1-a, a)$ and $P_{X|1} = (a, 1-a)$ with $0 < a \leq 10^{-3}$. Then, $C(R)$ is not concave on $[0, \log 2]$.

Proof. Appendix B.4. □

4 Cardinality Bounds and Attainability

Since Problem 1 is searching K in the entire feasible set, which is the union of an infinite number of sets, the attainability of the optimal solution has not been determined. Theorem 4.2 provides a guarantee in this regard, while Lemma 4.1 shows many good properties of the feasible encoder set $\mathcal{K}_R^{(M)}$, and is used in the proof of Theorem 4.2.

Lemma 4.1. *Given finite M, for any $R \geq 0$, the feasible encoder set $\mathcal{K}_R^{(M)}$ is non-empty, compact and convex.*

Proof. Appendix B.5. □

Theorem 4.2 (Cardinality bound)**.** *For any $R \geq 0$,*

$$\sup_{K \in \mathscr{K}_R} C(P_{U|0}^K, P_{U|1}^K) = \max_{K \in \mathcal{K}_R^{(|\mathcal{X}|+1)}} C(P_{U|0}^K, P_{U|1}^K), \tag{18}$$

which indicates the cardinality of U can be restricted to satisfy $|\mathcal{U}| \leq |\mathcal{X}|+1$ without loss of generality. Moreover, the maximum in (18) is achievable.

Proof. Assume $K \in \mathscr{K}_R$ is an arbitrary feasible encoder with output alphabet $\mathcal{U}_m$. Let $B_m(K, s)$ be defined in (8), and choose

$$s_K \in \arg\min_{0 \leq s \leq 1} B_m(K, s). \tag{19}$$

Then, the corresponding Chernoff information is

$$\mathcal{C}(K) = -\log B_m(K, s_K).$$

Assume $P_U^K(u) > 0$ for all $u \in \mathcal{U}_m$, otherwise, it could be directly removed and reduce m. For each $u \in \mathcal{U}_m$, define $\lambda_u \triangleq P_U^K(u)$, $w_u \triangleq P_{X|U=u}^K \in \Delta_{|\mathcal{X}|-1}$, and for $s \in [0,1]$,

$$g_s(w) \triangleq \Big(\frac{\sum_x w(x) P_{Y|X}(0|x)}{P_Y(0)}\Big)^s \Big(\frac{\sum_x w(x) P_{Y|X}(1|x)}{P_Y(1)}\Big)^{1-s}.$$

Then,

$$B_m(K, s) = \sum_{u=1}^m \lambda_u g_s(w_u).$$

Construct vectors

$$v_u = (1, w_u(1), \cdots, w_u(|\mathcal{X}|-1), \mathrm{D}(w_u\|P_X)) \in \mathbb{R}^{|\mathcal{X}|+1}, \quad \forall u \in \mathcal{U}_m.$$

If $m \geq |\mathcal{X}|+2$, then vectors $v_1, \cdots, v_m$ are linearly dependent, which means there exists $c_1, \cdots, c_m$ not all zeros such that $\sum_{u=1}^m c_u v_u = 0$. In this case, expand each coordinate gives

$$\begin{aligned} \sum_{u=1}^m c_u &= 0; \\ \sum_{u=1}^m c_u w_u(x) &= 0, \quad x = 1, \cdots, |\mathcal{X}|-1; \\ \sum_{u=1}^m c_u \mathrm{D}(w_u\|P_X) &= 0. \end{aligned} \tag{20}$$

Moreover, since

$$\sum_{u=1}^m c_u w_u(x_{|\mathcal{X}|}) = \sum_{u=1}^m c_u (1 - \sum_{x=1}^{|\mathcal{X}|-1} w_u(x)) = \sum_{u=1}^m c_u - \sum_{x=1}^{|\mathcal{X}|-1} \sum_{u=1}^m c_u w_u(x) = 0,$$

the equation (20) holds for all $x \in \mathcal{X}$.

For any $t \in \mathbb{R}$, define the disturbance weight

$$\lambda_u(t) = \lambda_u + t c_u.$$

Because $c_1, \cdots, c_m$ are not all zeros, and $\sum_{u=1}^m c_u = 0$, there must be both positive and negative numbers in the sequence $\{c_u\}_{u=1}^m$. Define

$$t_- = \max_{u:c_u>0} \Big(-\frac{\lambda_u}{c_u} \Big) < 0, \quad t_+ = \min_{u:c_u<0} \Big(-\frac{\lambda_u}{c_u} \Big) > 0.$$

Then, for any $t \in (t_-, t_+)$, all $\lambda_u(t)$ remains positive; when $t = t_-$ or t_+, at least one weight $\lambda_u(t)$ goes exactly to zero; for any $t \in [t_-, t_+]$, we have

$$\begin{aligned} \sum_{u=1}^m \lambda_u(t) &= \sum_{u=1}^m \lambda_u + t \sum_{u=1}^m c_u = 1; \\ \sum_{u=1}^m \lambda_u(t) w_u(x) &= P_X(x) + t \sum_{u=1}^m c_u w_u(x) = P_X(x); \\ I_t(U;X) = \sum_{u=1}^m \lambda_u(t) D(w_u\|P_X) &= I_K(U;X) + t \sum_{u=1}^m c_u D(w_u\|P_X) = I_K(U;X). \end{aligned}$$

Furthermore, for the s_K selected in (19), define

$$\hat{B}_{s_K}(t) \triangleq \sum_{u=1}^m \lambda_u(t) g_{s_K}(w_u) = \hat{B}_{s_K}(0) + t \sum_{u=1}^m c_u g_{s_K}(w_u).$$

If $\sum_{u=1}^m c_u g_{s_K}(w_u) \geq 0$, let $t = t_- < 0$; if $\sum_{u=1}^m c_u g_{s_K}(w_u) < 0$, let $t = t_+ > 0$. Thus, we can always select $t_0 \in \{t_-, t_+\}$ such that

$$\hat{B}_{s_K}(t_0) \leq \hat{B}_{s_K}(0) = B_m(K, s_K),$$

and at least one $\lambda_u(t_0) = 0$. Define the updated weights $\tilde{\lambda}_u = \lambda_u(t_0), \forall u \in \mathcal{U}_m$, and removing those symbols with zero probability gives the updated alphabet $\mathcal{U}_{\tilde{m}}$ with $\tilde{m} \leq m-1$. Define a new encoder $\tilde{K}$ on $\mathcal{U}_{\tilde{m}}$ as

$$\tilde{K}(u|x) = \frac{\tilde{\lambda}_u w_u(x)}{P_X(x)}, \quad \forall u \in \mathcal{U}_{\tilde{m}}.$$

It satisfies $I_{\tilde{K}}(U;X) = I_K(U;X) \leq R$, and $B_{\tilde{m}}(\tilde{K}, s_K) \leq B_m(K, s_K)$. Hence,

$$\mathcal{C}(\tilde{K}) = -\log \min_{0 \leq s \leq 1} B_{\tilde{m}}(\tilde{K}, s) \geq -\log B_{\tilde{m}}(\tilde{K}, s_K) \geq -\log B_m(K, s_K) = \mathcal{C}(K).$$

Therefore, for any encoder K on $\mathcal{U}_m$ with $m \geq |\mathcal{X}| + 2$, we can always construct a new encoder $\tilde{K}$, whose positive probability output symbol is reduced by at least one, and

$$I_{\tilde{K}}(U;X) = I_K(U;X), \quad \mathcal{C}(\tilde{K}) \geq \mathcal{C}(K).$$

Accordingly,

$$\max_{K \in \mathcal{K}_R^{(|\mathcal{X}|+1)}} \mathcal{C}(K) \geq \max_{K \in \mathcal{K}_R^{(|\mathcal{X}|+2)}} \mathcal{C}(K) \geq \cdots \geq \max_{K \in \mathcal{K}_R^{(|\mathcal{X}|+n)}} \mathcal{C}(K), \quad n \to \infty.$$

And since

$$\mathscr{K}_R = \bigcup_{M \geq 1} \mathcal{K}_R^{(M)},$$

we have

$$\sup_{K \in \mathscr{K}_R} \mathcal{C}(K) = \max_{K \in \mathcal{K}_R^{(|\mathcal{X}|+1)}} \mathcal{C}(K).$$

Furthermore, by Lemma 4.1, $\mathcal{K}_R^{(|\mathcal{X}|+1)}$ is non-empty and compact, and since $\mathcal{C}(K)$ is continuous with respect to K, the maximum in (18) is attainable. □

Theorem 4.2 demonstrates the supremum in the entire feasible set $\mathscr{K}_R$ can be achieved in a feasible encoder set $\mathcal{K}_R^{(M)}$ with cardinality $M = |\mathcal{X}| + 1$. In fact, the cardinality bound can be further sharpened, as Theorem 4.4 shows, while Lemma 4.3 is applied in the proof. Let V be defined in (12), taking k distinct values.

Lemma 4.3 (Reduction to the sufficient statistic)**.** *Consider an encoder $\bar{K} = \bar{K}_{U|V}$ of V with output alphabet $\mathcal{U}_M$, denote by $P_{U|y}^{\bar{K}}(u) = \sum_{j=1}^{k} P_{V|y}(j)\bar{K}(u|j)$, $y \in \{0,1\}$, its induced conditional distributions, and by $I_{\bar{K}}(U;V)$ its mutual information under P_V. Then,*

(i) For every encoder $K = K_{U|X}$ with output alphabet $\mathcal{U}_M$, there exists an encoder $\bar{K}$ of V with output alphabet $\mathcal{U}_M$ such that

$$P_{U|y}^{\bar{K}} = P_{U|y}^{K}, \ y \in \{0,1\}, \qquad I_{\bar{K}}(U;V) \leq I_K(U;X). \tag{21}$$

(ii) For every encoder $\bar{K}$ of V with output alphabet $\mathcal{U}_M$, the encoder $K(u|x) = \bar{K}(u|\nu(x))$ of X satisfies

$$P_{U|y}^{K} = P_{U|y}^{\bar{K}}, \ y \in \{0,1\}, \qquad I_K(U;X) = I_{\bar{K}}(U;V). \tag{22}$$

Proof. Appendix B.6. □

Theorem 4.4 (Sharpened cardinality bound)**.** *For any $R \geq 0$,*

$$\sup_{K \in \mathscr{K}_R} C(P^K_{U|0}, P^K_{U|1}) = \max_{K \in \mathcal{K}_R^{(k+1)}} C(P^K_{U|0}, P^K_{U|1}), \tag{23}$$

and the maximum is attained by an encoder of the form $K(u|x) = \bar{K}(u|\nu(x))$ for some encoder $\bar{K} = \bar{K}_{U|V}$ of V with output alphabet $\mathcal{U}_{k+1}$. Hence the cardinality of U can be restricted to $|\mathcal{U}| \leq k+1$ without loss of generality.

Proof. Consider the pair (V, Y) with joint distribution

$$P_{V,Y}(j, y) = \sum_{x \in \mathcal{X}_j} P_{X,Y}(x, y), \quad (j, y) \in \{1, \cdots, k\} \times \{0, 1\}.$$

The marginals are $P_Y(y) > 0$ and $P_V(j) = \sum_{x \in \mathcal{X}_j} P_X(x) > 0$, and by (1), $P_{V|y}(j) = P_{X|y}(\mathcal{X}_j) > 0$ for all j and y. Moreover, summing $P_{X|1}(x) = \rho_j P_{X|0}(x)$ over $\mathcal{X}_j$ derives $P_{V|1}(j) = \rho_j P_{V|0}(j)$. Thus, $P_{X|0} \not\equiv P_{X|1}$ indicates $P_{V|0} \not\equiv P_{V|1}$.

With the notation of Lemma 4.3, and similar to the definition of $\mathscr{K}_R$ in (5), denote by the entire feasible set from V to U as $\mathscr{K}_R$, and define the curve of the reduced problem as

$$C_V(R) \stackrel{\Delta}{=} \sup_{\bar{K} \in \mathscr{K}_R} C(P^{\bar{K}}_{U|0}, P^{\bar{K}}_{U|1}). \tag{24}$$

Since the proof of Theorem 4.2 uses only the standing assumptions of Section 2, it applies verbatim to the pair (V, Y), whose alphabet has cardinality k. Therefore, there exists an encoder $\bar{K}^*$ of V with output alphabet $\mathcal{U}_{k+1}$ such that

$$I_{\bar{K}^*}(U; V) \leq R, \quad C(P^{\bar{K}^*}_{U|0}, P^{\bar{K}^*}_{U|1}) = C_V(R). \tag{25}$$

We now show that $C(R) = C_V(R)$. Let $K \in \mathscr{K}_R$ be arbitrary, with output alphabet $\mathcal{U}_M$, and let $\bar{K}$ be defined by Lemma 4.3(i). By (21), $I_{\bar{K}}(U; V) \leq I_K(U; X) \leq R$ and $P^{\bar{K}}_{U|y} = P^K_{U|y}$ for $y \in \{0, 1\}$. Hence

$$\mathcal{C}(K) = C(P^K_{U|0}, P^K_{U|1}) = C(P^{\bar{K}}_{U|0}, P^{\bar{K}}_{U|1}) \leq C_V(R),$$

and taking the supremum over $K \in \mathscr{K}_R$ gives $C(R) \leq C_V(R)$. Conversely, let $K^*(u|x) = \bar{K}^*(u|\nu(x))$. By (22) and (25), K^* has output alphabet $\mathcal{U}_{k+1}$, $I_{K^*}(U; X) = I_{\bar{K}^*}(U; V) \leq R$, and $P^{K^*}_{U|y} = P^{\bar{K}^*}_{U|y}$ for $y \in \{0, 1\}$, so that

$$K^* \in \mathcal{K}_R^{(k+1)}, \quad \mathcal{C}(K^*) = C_V(R) \geq C(R).$$

Combining the two directions, and using $\mathcal{K}_R^{(k+1)} \subseteq \mathscr{K}_R$,

$$C(R) \leq \mathcal{C}(K^*) \leq \max_{K \in \mathcal{K}_R^{(k+1)}} \mathcal{C}(K) \leq \sup_{K \in \mathscr{K}_R} \mathcal{C}(K) = C(R).$$

Hence all inequalities hold with equality, and the supremum over $\mathcal{K}_R^{(k+1)}$ is attained by K^*, which proves (23) with a maximizer of the form $K^*(u|x) = \bar{K}^*(u|\nu(x))$. □

Remark 4.5. Theorem 4.4 strictly improves Theorem 4.2 when $k < |\mathcal{X}|$, i.e., at least two symbols share the same likelihood ratio. Lemma 4.3 also shows that the curve $C(R)$ depends on the source only through the joint distribution $P_{V,Y}$: compressing X is never better than compressing its minimal sufficient statistic, at any rate.

Therefore, we can reformulate the MI constrained Chernoff bottleneck as follows:

Problem 2. *Given $R \geq 0$ and $\mathcal{U}_M = \{1, \cdots, M\}$, seek*

$$K^* \in \arg \max_{K \in \mathcal{K}_R^{(M)}} C(P_{U|0}^K, P_{U|1}^K). \tag{26}$$

Theorem 4.4 implies that we can recover the optimal solution of Problem 1 by setting $M \geq k+1$ in Problem 2, for which we establish an alternating optimization algorithm in Section 5.

5 An Alternating Optimization Algorithm

In this section, we fix the alphabet $\mathcal{U}_M = \{1, \cdots, M\}$. With $B_M(K, s)$ defined in (8), the maximization in (26) can be rewritten as

$$\min_{K \in \mathcal{K}_R^{(M)}} \min_{0 \leq s \leq 1} B_M(K, s).$$

We establish an iterative algorithm for Problem 2 by alternately minimizing over K and s:

(a) Fix s, update K with

$$K \in \arg \min_{K \in \mathcal{K}_R^{(M)}} B_M(K, s); \tag{27}$$

(b) Fix K, update s with

$$s = \arg \min_{0 \leq s \leq 1} B_M(K, s). \tag{28}$$

5.1 Step (a): Generalized Blahut-Arimoto Algorithm

In this step, we fix s and update K. For fixed $s \in (0, 1)$, and any $w(u) > 0$, Young's inequality (Fact A.1) gives

$$P_{U|0}^K(u)^s P_{U|1}^K(u)^{1-s} \leq s w(u) P_{U|0}^K(u) + (1-s) w(u)^{-\frac{s}{1-s}} P_{U|1}^K(u),$$

and equality holds when $w(u) P_{U|0}^K(u) = w(u)^{-\frac{s}{1-s}} P_{U|1}^K(u)$. Namely,

$$P_{U|0}^K(u)^s P_{U|1}^K(u)^{1-s} = \min_{w(u)>0} \left(s w(u) P_{U|0}^K(u) + (1-s) w(u)^{-\frac{s}{1-s}} P_{U|1}^K(u) \right),$$

with the minimum achieved by

$$w(u) = \left(\frac{P_{U|1}^K(u)}{P_{U|0}^K(u)} \right)^{1-s}. \tag{29}$$

Lemma 5.1. *Let $\mathcal{K}_{R,+}^{(M)}$ be defined in (4), and ℓ and L be defined in (10). If $K_{U|X} \in \mathcal{K}_{R,+}^{(M)}$, then for any $u \in \mathcal{U}_M$,*

$$0 < \ell \leq \frac{P_{U|1}^K(u)}{P_{U|0}^K(u)} \leq L < +\infty.$$

Hence, $\ell^{1-s} \leq w(u) \leq L^{1-s}$.

Proof. Appendix B.7. □

Remark 5.2. Lemma 5.1 gives a uniform bound of $w(u)$. The condition $K_{U|X} \in \mathcal{K}_{R,+}^{(M)}$ holds for every iterate of Algorithm 2 by Theorem 5.13.

Thus, the minimization in (27) becomes

$$\min_{K \in \mathcal{K}_R^{(M)}} \min_{w(u)>0} \sum_{u \in \mathcal{U}_M} s w(u) P_{U|0}^K(u) + (1-s) w(u)^{-\frac{s}{1-s}} P_{U|1}^K(u). \tag{30}$$

For $y \in \{0,1\}$, Markovity gives

$$P_{U|y}^K(u) = \sum_{x \in \mathcal{X}} K(u|x) P_{X|y}(x).$$

Define the parameter matrix $C^s \in \mathbb{R}_+^{M \times |\mathcal{X}|}$ with entries

$$C_{u,x}^s = s w(u) P_{X|0}(x) + (1-s) w(u)^{-\frac{s}{1-s}} P_{X|1}(x), \tag{31}$$

where $w(u)$ is given by (29) at the current encoder. With $w(u)$ fixed, the objective in (30) is linear in K, and the update of K solves

$$\min_{K \in \mathcal{K}_R^{(M)}} \sum_{(u,x) \in \mathcal{U}_M \times \mathcal{X}} C_{u,x}^s K(u|x). \tag{32}$$

Remark 5.3 (Concavity and the linear majorizer). For fixed $s \in [0,1]$, $B_M(K,s)$ is concave in K, since $(a,b) \mapsto a^s b^{1-s}$ is concave on $\mathbb{R}_+^2$ and $P_{U|y}^K$ is linear in K. Consequently, maximizing $\mathcal{C}(K)$ amounts to minimizing the concave function $\min_{s \in [0,1]} B_M(K,s)$ over the convex set $\mathcal{K}_R^{(M)}$, for which global optimality cannot be expected in general. Moreover, let $w(u)$ and $C_{u,x}^s$ be computed from a current encoder $\tilde{K} \in \mathcal{K}_{R,+}^{(M)}$ via (29) and (31). For $s \in (0,1)$, direct differentiation gives

$$\frac{\partial B_M(K,s)}{\partial K(u|x)}\Big|_{K=\tilde{K}} = s w(u) P_{X|0}(x) + (1-s) w(u)^{-\frac{s}{1-s}} P_{X|1}(x) = C_{u,x}^s.$$

Since $B_M(\cdot, s)$ is positively homogeneous of degree one, Euler's identity gives

$$B_M(\tilde{K}, s) = \sum_{u,x} C_{u,x}^s \tilde{K}(u|x).$$

Hence the objective in (32) is exactly the tangent plane of $B_M(\cdot, s)$ at $\tilde{K}$: by concavity, it upper bounds $B_M(K,s)$ for every K, with equality at $K = \tilde{K}$. Step (a) is therefore a successive linearization of a concave function.

Write the Lagrangian function as

$$\mathcal{L}(K, \lambda, \mu) = \sum_{u,x} C_{u,x}^s K(u|x) + \lambda (I_K(U;X) - R) + \sum_x \mu(x) \Big(\sum_u K(u|x) - 1 \Big),$$

with the trade-off parameter $\lambda \geq 0$. Taking derivative over K gives

$$\frac{\partial \mathcal{L}}{\partial K} = C_{u,x}^s + \lambda P_X(x) \log \frac{K(u|x)}{P_U(u)} + \mu(x).$$

Let $\frac{\partial \mathcal{L}}{\partial K} = 0$ and normalize, we have

$$K(u|x) = \frac{P_U(u) \exp\Big(-\dfrac{C_{u,x}^s}{\lambda P_X(x)} \Big)}{\sum_{u'} P_U(u') \exp\Big(-\dfrac{C_{u',x}^s}{\lambda P_X(x)} \Big)}. \tag{33}$$

And for fixed $K(u|x)$, the output distribution $P_U(u)$ is given by $P_U(u) = \sum_x K(u|x)P_X(x)$. Thus, we have established the Generalized Blahut-Arimoto (GBA) algorithm for (32).

Algorithm 1 Generalized Blahut-Arimoto Algorithm for (32)

Input: P_X; parameter matrix $C^s_{u,x}$; current $K_{U|X}$; trade-off $\lambda > 0$

Output: converged encoder $\hat{K}_{U|X}$

1: $K^{(0)}_{U|X} \leftarrow K_{U|X}$; $P^{(0)}_U(u) \leftarrow \sum_x K^{(0)}_{U|X}(u|x)P_X(x)$; $t \leftarrow 0$

2: **repeat**

3: Given $P^{(t)}_U$, update $K^{(t+1)}_{U|X}$:

$$K^{(t+1)}(u|x) \leftarrow \frac{P^{(t)}_U(u)\exp\Big(-\dfrac{C^s_{u,x}}{\lambda P_X(x)}\Big)}{\sum_{u'} P^{(t)}_U(u')\exp\Big(-\dfrac{C^s_{u',x}}{\lambda P_X(x)}\Big)}$$

4: Given $K^{(t+1)}_{U|X}$, update $P^{(t+1)}_U$:

$$P^{(t+1)}_U(u) \leftarrow \sum_x K^{(t+1)}(u|x)P_X(x)$$

5: $t \leftarrow t+1$

6: **until** $\max_{u\in\mathcal{U}_M}\left(P^{(t)}_U(u) - P^{(t-1)}_U(u)\right)/P^{(t-1)}_U(u)$ very small ▷ as stated in Rem. 5.4

7: **return** $\hat{K}_{U|X} = K^{(t)}_{U|X}$

Remark 5.4. The stopping criterion is that of the Blahut–Arimoto algorithm [24, 25], and it comes with an optimality certificate. Let $\epsilon_t = \max_u \left(P^{(t)}_U(u) - P^{(t-1)}_U(u)\right)/P^{(t-1)}_U(u) \geq 0$ be the quantity checked at termination. By Corollary 5.9, the output $\hat{K}_{U|X} = K^{(t)}_{U|X}$ satisfies

$$\mathcal{L}_\lambda(\hat{K}) - \min_K \mathcal{L}_\lambda(K) \leq \lambda \log(1+\epsilon_t) \leq \lambda\epsilon_t,$$

where $\mathcal{L}_\lambda$ is defined in Section 5.4.

5.2 Step (b): Solving a Nonlinear Equation

In this step, we fix K and update s. We consider the encoder $K \in \mathcal{K}^{(M)}_{R,+}$, which holds for every iterate of Algorithm 2 by Theorem 5.13.

For fixed $\mathcal{U}_M$ and $K_{U|X} \in \mathcal{K}^{(M)}_{R,+}$, define

$$g_K(s) \triangleq \sum_{u\in\mathcal{U}_M} P^K_{U|0}(u)^s P^K_{U|1}(u)^{1-s}, \quad s \in [0,1] \tag{34}$$

Lemma 5.5. *For $K_{U|X} \in \mathcal{K}^{(M)}_{R,+}$, $g_K(s)$ is continuous and infinitely differentiable on $[0,1]$.*

Proof. $K_{U|X} \in \mathcal{K}^{(M)}_{R,+}$ and the Markovity guarantees for any $u \in \mathcal{U}_M$, $P^K_{U|0}(u) > 0, P^K_{U|1}(u) > 0$. Since $g_K(s)$ is a summation of finite terms, considering $h^u_K(s) \triangleq P^K_{U|0}(u)^s P^K_{U|1}(u)^{1-s}$, we have

$$h^u_K(s) = \exp\left(s\log P^K_{U|0}(u) + (1-s)\log P^K_{U|1}(u)\right),$$

which is certainly continuous and infinitely differentiable. □

Calculate the derivatives

$$g'_K(s) = \sum_{u\in\mathcal{U}_M} P^K_{U|0}(u)^s P^K_{U|1}(u)^{1-s} \log \frac{P^K_{U|0}(u)}{P^K_{U|1}(u)}, \tag{35}$$

and

$$g''_K(s) = \sum_{u\in\mathcal{U}_M} P^K_{U|0}(u)^s P^K_{U|1}(u)^{1-s} \Big(\log \frac{P^K_{U|0}(u)}{P^K_{U|1}(u)} \Big)^2. \tag{36}$$

Thus, $g''_K(s) \geq 0$ on $[0,1]$. Moreover, if $P^K_{U|0} \not\equiv P^K_{U|1}$ on $\mathcal{U}_M$, we have $g''_K(s) > 0$ on $[0,1]$, i.e., $g'_K(s)$ is monotonically increasing on $[0,1]$, and

$$g'_K(0) = \sum_{u\in\mathcal{U}_M} P^K_{U|1}(u) \log \frac{P^K_{U|0}(u)}{P^K_{U|1}(u)} = -\mathrm{D}(P^K_{U|1} \| P^K_{U|0}) < 0, \tag{37}$$

$$g'_K(1) = \sum_{u\in\mathcal{U}_M} P^K_{U|0}(u) \log \frac{P^K_{U|0}(u)}{P^K_{U|1}(u)} = \mathrm{D}(P^K_{U|0} \| P^K_{U|1}) > 0. \tag{38}$$

Hence, we have the following lemma.

Lemma 5.6. *For $K_{U|X} \in \mathcal{K}^{(M)}_{R,+}$, $P^K_{U|0} \not\equiv P^K_{U|1}$, there exists the unique $s^* \in (0,1)$ such that*

$$g_K(s^*) = \min_{s\in[0,1]} g_K(s).$$

Proof. The conditions give (37) and (38). And since $g''_K(s) > 0$ on $[0,1]$, there exists the unique $s^* \in (0,1)$ such that $g'_K(s^*) = 0$. □

Searching $s^* = \arg\min_{s\in[0,1]} g_K(s)$ is equivalent to searching s^* such that

$$g'_K(s^*) = 0, \tag{39}$$

where $g'_K(s)$ is given by (35). Non-linear equation (39) can be solved by numerical methods such as the bisection method, Newton's method, or Brent's method [32, 33, 34].

5.3 Alternating Optimization Algorithm for Problem 2

Combining step (a) and (b), we establish the algorithm for Problem 2 as below.

Algorithm 2 Algorithm for Problem 2

Input: P_X; $P_{X|0}$; $P_{X|1}$; cardinality M; rate $R > 0$
Output: Encoder K^*

Initialize: $K_{U|X}^{(0)} \in \mathcal{K}_{R,+}^{(M)}$; $s^{(0)} \in (0,1)$; $l \leftarrow 0$
Compute $P_U^{(0)}$, $P_{U|0}^{(0)}$, $P_{U|1}^{(0)}$ from $K_{U|X}^{(0)}$; $B^{(0)} \leftarrow \sum_{u \in \mathcal{U}_M} P_{U|0}^{(0)}(u)^{s^{(0)}} P_{U|1}^{(0)}(u)^{1-s^{(0)}}$
if $B^{(0)} = 1$ **then go to** line 1 ▷ re-initialize to avoid $P_{U|0}^{(0)} \equiv P_{U|1}^{(0)}$
Compute Γ by (11)
repeat
  For each u, x, compute $w(u)$ from (29) and $C_{u,x}^{s^{(l)}}$ from (31)
  $\lambda_{\text{small}} \leftarrow 0$; $\lambda_{\text{big}} \leftarrow \Gamma/R$; $K_{U|X}^{(l+1)} \leftarrow K_{U|X}^{(l)}$ ▷ initial bracket (Lem. 5.10)
  repeat
    $\lambda \leftarrow (\lambda_{\text{small}} + \lambda_{\text{big}})/2$ ▷ bisection (Lem. 5.10)
    Start from $K_{U|X}^{(l)}$, run Algorithm 1 with λ to derive $\hat{K}_{U|X}$; compute $\hat{I}_K(U;X)$
    if $\hat{I}_K(U;X) > R$ **then**
      $\lambda_{\text{small}} \leftarrow \lambda$
    else
      $\lambda_{\text{big}} \leftarrow \lambda$
      if $\sum_{u,x} C_{u,x}^{s^{(l)}} \hat{K}(u|x) \le \sum_{u,x} C_{u,x}^{s^{(l)}} K^{(l+1)}(u|x)$ **then** ▷ safeguard
        $K_{U|X}^{(l+1)} \leftarrow \hat{K}_{U|X}$
      end if
    end if
  until $\lambda_{\text{big}} - \lambda_{\text{small}}$ very small
  Compute $P_U^{(l+1)}$, $P_{U|0}^{(l+1)}$, $P_{U|1}^{(l+1)}$ from $K_{U|X}^{(l+1)}$
  Compute $g'_K(s)$ from (35); solve equation (39) to derive $s^{(l+1)}$
  $B^{(l+1)} \leftarrow \sum_{u \in \mathcal{U}_M} P_{U|0}^{(l+1)}(u)^{s^{(l+1)}} P_{U|1}^{(l+1)}(u)^{1-s^{(l+1)}}$ ▷ update objective B
  $l \leftarrow l + 1$
until $|B^{(l)} - B^{(l-1)}|$ very small
return $K^* = K_{U|X}^{(l)}$

5.4 Convergence Analysis

We analyze the convergence of Algorithms 1 and 2.

Lemma 5.7 (Uniform boundedness)**.** *For $K_{U|X} \in \mathcal{K}_{R,+}^{(M)}$ and $s \in (0,1)$, let $w(u)$ be given by (29) and $C_{u,x}^s$ by (31). Then, for all $(u,x) \in \mathcal{U}_M \times \mathcal{X}$,*

$$0 < P_{X|0}(x)^s P_{X|1}(x)^{1-s} \le C_{u,x}^s \le \Gamma\big(sP_{X|0}(x) + (1-s)P_{X|1}(x)\big).$$

Proof. Appendix B.8. □

For fixed $C_{u,x}^s$ and $\lambda \ge 0$, denote by $\mathcal{K}_\infty^{(M)}$ the set of all encoders with output alphabet $\mathcal{U}_M$, with no rate constraint, and define

$$G(K) \triangleq \sum_{u,x} C_{u,x}^s K(u|x), \quad \mathcal{L}_\lambda(K) \triangleq G(K) + \lambda I_K(U;X),$$

so that (32) reads $\min_{K\in\mathcal{K}_R^{(M)}} G(K)$. Let $\mathcal{L}_\lambda^* = \min_{K\in\mathcal{K}_\infty^{(M)}} \mathcal{L}_\lambda(K)$, which is attained since $\mathcal{K}_\infty^{(M)}$ is compact and $\mathcal{L}_\lambda$ is continuous. For $\lambda > 0$, Algorithm 1 alternately minimizes

$$\hat{\mathcal{L}}_\lambda(K_{U|X}, P_U) = \sum_{u,x} C_{u,x}^s K(u|x) + \lambda \sum_x P_X(x) \sum_u K(u|x) \log \frac{K(u|x)}{P_U(u)} \tag{40}$$

over $K_{U|X} \in \mathcal{K}_\infty^{(M)}$ and P_U. Since $\sum_x P_X(x) \sum_u K(u|x) \log \frac{K(u|x)}{P_U(u)} = I_K(U;X) + \mathrm{D}(P_U^K \| P_U)$,

$$\hat{\mathcal{L}}_\lambda(K, P_U) = \mathcal{L}_\lambda(K) + \lambda \mathrm{D}(P_U^K \| P_U), \tag{41}$$

so that $\min_{P_U} \hat{\mathcal{L}}_\lambda(K, P_U) = \mathcal{L}_\lambda(K)$, attained at $P_U = P_U^K$, and the global minimum of $\hat{\mathcal{L}}_\lambda$ equals $\mathcal{L}_\lambda^*$.

Theorem 5.8 (Global convergence of Algorithm 1)**.** *Given $C_{u,x}^s$ and $\lambda > 0$, assume $P_U^{(0)}(u) > 0$ for all $u \in \mathcal{U}_M$ in Algorithm 1, which holds if $K_{U|X}^{(0)} \in \mathcal{K}_{R,+}^{(M)}$. Then $K_{U|X}^{(t)}$ converges to a global minimizer $\bar{K}$ of $\mathcal{L}_\lambda$ over $\mathcal{K}_\infty^{(M)}$, and $(K_{U|X}^{(t)}, P_U^{(t)})$ converges to $(\bar{K}, P_U^{\bar{K}})$, a global minimizer of $\hat{\mathcal{L}}_\lambda$. Moreover, $\mathcal{L}_\lambda(K_{U|X}^{(t)})$ is nonincreasing in t, and for every global minimizer K^* of $\mathcal{L}_\lambda$ over $\mathcal{K}_\infty^{(M)}$ and every $t \geq 1$,*

$$0 \leq \mathcal{L}_\lambda(K_{U|X}^{(t)}) - \mathcal{L}_\lambda^* \leq \frac{\lambda \mathrm{D}(P_U^{K^*} \| P_U^{(0)})}{t}. \tag{42}$$

Proof. Appendix B.9. □

Corollary 5.9 (Stopping certificate)**.** *Under the assumptions of Theorem 5.8, for every $t \geq 1$,*

$$0 \leq \mathcal{L}_\lambda(K_{U|X}^{(t)}) - \mathcal{L}_\lambda^* \leq \lambda \log \max_{u\in\mathcal{U}_M} \frac{P_U^{(t)}(u)}{P_U^{(t-1)}(u)}.$$

Proof. Let K^* be a global minimizer of $\mathcal{L}_\lambda$ and $q^* = P_U^{K^*}$. Applying (57) at step $t-1$ with $K = K^*$, in the notation of Appendix B.9,

$$\mathcal{L}_\lambda(K_t) - \mathcal{L}_\lambda^* \leq \lambda(\mathrm{D}(q^*\|q_{t-1}) - \mathrm{D}(q^*\|q_t)) = \lambda \sum_u q^*(u) \log \frac{q_t(u)}{q_{t-1}(u)} \leq \lambda \log \max_u \frac{q_t(u)}{q_{t-1}(u)}.$$

□

In Algorithm 2, the trade-off parameter λ is searched by bisection. To justify this, let $K(\lambda)$ be any minimizer of $\mathcal{L}_\lambda(K)$ in $\mathcal{K}_\infty^{(M)}$, i.e.,

$$K(\lambda) \in \arg\min_{K\in\mathcal{K}_\infty^{(M)}} \mathcal{L}_\lambda(K), \tag{43}$$

which Algorithm 1 approximates by Theorem 5.8, and let $I(\lambda) \triangleq I_{K(\lambda)}(U;X)$. The minimizer in (43) need not be unique, and the following result holds for every choice of it.

Lemma 5.10 (Bisection search of λ)**.** *Let $C_{u,x}^s$ be given by (31) for some $K_{U|X} \in \mathcal{K}_{R,+}^{(M)}$ and $s \in (0,1)$, let $R > 0$, Γ be defined in (11), and $G^* = \min_{K\in\mathcal{K}_R^{(M)}} G(K)$ be the optimal value of (32). Then, for every choice of $K(\lambda)$ in (43):*

(i) $I(\lambda)$ is nonincreasing on $[0, +\infty)$, i.e., $I(\lambda_1) \geq I(\lambda_2)$ for all $0 \leq \lambda_1 < \lambda_2$.

(ii) If $I(\lambda) \le R$, then $K(\lambda) \in \mathcal{K}_R^{(M)}$ and

$$0 \le G(K(\lambda)) - G^* \le \lambda(R - I(\lambda)). \tag{44}$$

In particular, $K(\lambda)$ solves (32) if $I(\lambda) = R$ or $\lambda = 0$.

(iii) There exists $\lambda^* \in [0, \Gamma/R)$ such that $I(\lambda) > R$ for $\lambda < \lambda^*$, and $I(\lambda) \le R$ for $\lambda > \lambda^*$.

Consequently, the bisection in lines 7–19 of Algorithm 2, initialized with $\lambda_{small} = 0$ and $\lambda_{big} = \Gamma/R$, keeps $\lambda^* \in [\lambda_{small}, \lambda_{big}]$ and $I(\lambda_{big}) \le R$ at every step, and after n steps, $\lambda_{big} - \lambda_{small} = \Gamma/(R\,2^n)$.

Proof. Appendix B.10. □

Remark 5.11. Lemma 5.10 justifies lines 7–19 of Algorithm 2. The bracket $[0, \Gamma/R]$ is valid at every outer iteration, since the bound $\lambda^* < \Gamma/R$ relies only on Lemma 5.7. Locating λ^* to accuracy $\delta > 0$ takes $\lceil \log_2(\Gamma/(R\delta)) \rceil$ runs of Algorithm 1. Finally, $K_{U|X}^{(l+1)}$ is only replaced by encoders of rate at most R, so it stays feasible for (32), and its suboptimality is bounded by (44).

Many previous theorems and lemmas assume $s \in (0, 1)$; the following lemma ensures this.

Lemma 5.12 (Interiority of s^*)**.** *For $K_{U|X} \in \mathcal{K}_{R,+}^{(M)}$ with $P_{U|0}^K \not\equiv P_{U|1}^K$, let $\mathcal{C}(K)$ be defined in (9) as the corresponding Chernoff information, Γ be defined in (11), and s^* be the unique root of (39). Then,*

$$\frac{\mathcal{C}(K)}{\log \Gamma} \le s^* \le 1 - \frac{\mathcal{C}(K)}{\log \Gamma}. \tag{45}$$

Moreover, $\mathcal{C}(K) \le \frac{1}{2} \log \Gamma$, which indicates the interval (45) is non-empty.

Proof. By Lemma 5.1, for any $u \in \mathcal{U}_M$,

$$\left| \log \frac{P_{U|1}^K(u)}{P_{U|0}^K(u)} \right| \le \max\{\log L, \log \ell^{-1}\} = \log \Gamma.$$

Thus,

$$D(P_{U|1}^K \| P_{U|0}^K) \le \log \Gamma, \quad D(P_{U|0}^K \| P_{U|1}^K) \le \log \Gamma. \tag{46}$$

With $g_K(s)$ defined in (34), $K_{U|X} \in \mathcal{K}_{R,+}^{(M)}$ indicates $g_K(s) > 0$. Let $\psi_K(s) \triangleq \log g_K(s)$. While Lemma 5.5 guarantees the differentiability, calculate the derivatives gives

$$\psi_K'(s) = \frac{g_K'(s)}{g_K(s)}, \quad \psi_K''(s) = \frac{g_K''(s) g_K(s) - (g_K'(s))^2}{(g_K(s))^2}. \tag{47}$$

By (35), (36) and Cauchy-Schwarz inequality, we have $(g_K'(s))^2 \le g_K''(s) g_K(s)$, and equality holds if and only if $P_{U|0}^K \equiv P_{U|1}^K$. Hence, $\psi_K(s)$ is strictly convex on $[0, 1]$. Moreover, $g_K(0) = g_K(1) = 1$. Thus, $\psi_K(0) = \psi_K(1) = 0$, and by (37), (38) and (47),

$$\psi_K'(0) = g_K'(0) = -D(P_{U|1}^K \| P_{U|0}^K), \quad \psi_K'(1) = g_K'(1) = D(P_{U|0}^K \| P_{U|1}^K).$$

Since $\arg\min_{s \in [0,1]} \psi_K(s) = \arg\min_{s \in [0,1]} g_K(s) = s^*$, and $\psi_K(s^*) = -\mathcal{C}(K)$, convexity of $\psi_K(s)$ gives

$$-\mathcal{C}(K) = \psi_K(s^*) \ge \psi_K(0) + \psi_K'(0) s^* = -D(P_{U|1}^K \| P_{U|0}^K) s^*, \tag{48}$$

and

$$-\mathcal{C}(K) = \psi_K(s^*) \ge \psi_K(1) + \psi_K'(1)(s^* - 1) = D(P_{U|0}^K \| P_{U|1}^K)(s^* - 1). \tag{49}$$

Combining (46), (48), (49) derives (45) and $\mathcal{C}(K) \le \frac{1}{2} \log \Gamma$. □

Theorem 5.13 (Monotone convergence of Algorithm 2)**.** *In Algorithm 2, assume that, at every outer iteration $l \geq 0$, line 21 returns the exact minimizer $s^{(l+1)}$ of $B_M(K^{(l+1)}, \cdot)$ over $[0,1]$. Then, for every $l \geq 0$, $K_{U|X}^{(l)} \in \mathcal{K}_{R,+}^{(M)}$, $s^{(l)} \in (0,1)$, and*

$$0 < F(R) \leq B^{(l+1)} \leq B^{(l)} \leq \cdots \leq B^{(0)} < 1. \tag{50}$$

Consequently, $\{B^{(l)}\}_{l=0}^{\infty}$ converges to some $B^* \in [F(R), B^{(0)}] \subset (0,1)$, and the Chernoff information $\mathcal{C}(K^{(l)}) = -\log B^{(l)}$, $l \geq 1$, is nondecreasing and converges to $-\log B^* \leq C(R)$.

Proof. By lines 2 and 22, $B^{(l)} = B_M(K^{(l)}, s^{(l)})$ for every $l \geq 0$. For $l \geq 0$ with $K_{U|X}^{(l)} \in \mathcal{K}_{R,+}^{(M)}$ and $s^{(l)} \in (0,1)$, let $C_{u,x}^{s^{(l)}}$ be computed from $K_{U|X}^{(l)}$ in line 6, and write $G^{(l)}(K) = \sum_{u,x} C_{u,x}^{s^{(l)}} K(u|x)$. We prove by induction on l that

$$K_{U|X}^{(l)} \in \mathcal{K}_{R,+}^{(M)}, \quad s^{(l)} \in (0,1), \quad B^{(l)} \leq B^{(0)} < 1. \tag{51}$$

Base case. By line 1, $K_{U|X}^{(0)} \in \mathcal{K}_{R,+}^{(M)}$ and $s^{(0)} \in (0,1)$. By Young's inequality (Fact A.1) with $w(u) = 1$,

$$B^{(0)} \leq \sum_{u \in \mathcal{U}_M} (s^{(0)} P_{U|0}^{K^{(0)}}(u) + (1 - s^{(0)}) P_{U|1}^{K^{(0)}}(u)) = 1,$$

with equality if and only if $P_{U|0}^{K^{(0)}} \equiv P_{U|1}^{K^{(0)}}$, which is excluded by line 3. Hence $B^{(0)} < 1$.

Induction step. Assume (51) holds for l. In lines 7–19, $K_{U|X}^{(l+1)}$ is initialized as $K_{U|X}^{(l)}$ and replaced only by an output $\hat{K}_{U|X}$ of Algorithm 1 started from $K_{U|X}^{(l)}$ with $\hat{I}_K(U;X) \leq R$ and $G^{(l)}(\hat{K}) \leq G^{(l)}(K^{(l+1)})$. Such an output is strictly positive, by the proof of Theorem 5.8. Hence

$$K_{U|X}^{(l+1)} \in \mathcal{K}_{R,+}^{(M)}, \qquad G^{(l)}(K^{(l+1)}) \leq G^{(l)}(K^{(l)}).$$

By Remark 5.3, $G^{(l)}$ is the tangent plane of $B_M(\cdot, s^{(l)})$ at $K^{(l)}$, so that $B_M(K, s^{(l)}) \leq G^{(l)}(K)$ for every K, with equality at $K = K^{(l)}$. Since $s^{(l+1)}$ minimizes $B_M(K^{(l+1)}, \cdot)$,

$$\begin{aligned} B^{(l+1)} = B_M(K^{(l+1)}, s^{(l+1)}) &\leq B_M(K^{(l+1)}, s^{(l)}) \leq G^{(l)}(K^{(l+1)}) \\ &\leq G^{(l)}(K^{(l)}) = B_M(K^{(l)}, s^{(l)}) = B^{(l)} \leq B^{(0)} < 1. \end{aligned}$$

If $P_{U|0}^{K^{(l+1)}} \equiv P_{U|1}^{K^{(l+1)}}$, then $B_M(K^{(l+1)}, \cdot) \equiv 1$, contradicting $B^{(l+1)} = \min_{s \in [0,1]} B_M(K^{(l+1)}, s) < 1$. Hence Lemma 5.6 gives $s^{(l+1)} \in (0,1)$, which completes the induction and proves the monotonicity in (50).

Lower bound. Since $K_{U|X}^{(l)} \in \mathcal{K}_R^{(M)} \subseteq \mathscr{K}_R$,

$$B^{(l)} \geq \min_{s \in [0,1]} B_M(K^{(l)}, s) = e^{-\mathcal{C}(K^{(l)})} \geq e^{-C(R)} = F(R),$$

and $F(R) > 0$ since $C(R) \leq C_X < \infty$ by Lemma 3.1.

Convergence. $\{B^{(l)}\}$ is nonincreasing and bounded below by $F(R)$, hence converges to some $B^* \in [F(R), B^{(0)}]$. For $l \geq 1$, $s^{(l)}$ minimizes $B_M(K^{(l)}, \cdot)$, so that $\mathcal{C}(K^{(l)}) = -\log B^{(l)}$. This sequence is nondecreasing and converges to $-\log B^* \leq -\log F(R) = C(R)$. □

Remark 5.14. Lemma 5.12 ensures that $s^{(l)}$ lies in the interior of $(0,1)$ for each iteration $l \geq 1$ of Algorithm 2. Moreover, by Theorem 5.13, $\mathcal{C}(K^{(l)}) = -\log B^{(l)} \geq -\log B^{(0)} > 0$ for every $l \geq 1$, so that $s^{(l)}$ is uniformly bounded away from 0 and 1:

$$s^{(l)} \in \Big[\frac{-\log B^{(0)}}{\log \Gamma},\ 1 - \frac{-\log B^{(0)}}{\log \Gamma}\Big] \subset (0,1), \quad \forall\, l \geq 1.$$

6 When Does the Chernoff Bottleneck Outperform the IB?

The Chernoff bottleneck and the classical IB optimize different objectives over the same set of encoders. In this section, we compare their optimal encoders when Y is inferred from a finite number n of samples. Fix $R > 0$ and $M \geq k+1$, and let

$$K_{\mathrm{IB}} \in \arg\max_{K \in \mathcal{K}_R^{(M)}} I_K(U;Y), \qquad K_{\mathrm{CB}} \in \arg\max_{K \in \mathcal{K}_R^{(M)}} \mathcal{C}(K),$$

where both maxima exist since $\mathcal{K}_R^{(M)}$ is compact (Lemma 4.1). By Theorem 4.4, $\mathcal{C}(K_{\mathrm{CB}}) = C(R)$. Since $I_K(U;Y)$ also depends on K only through $P_{U|0}^K$ and $P_{U|1}^K$ (for the given P_Y), Lemma 4.3 and the cardinality bound of the IB [26], applied to V, show that K_{IB} attains the IB curve as well.

For an encoder K, let $U_1, \cdots, U_n$ be i.i.d. given Y with law $P_{U|Y}^K$, and write $P_{U|y}^K(u^n) = \prod_{i=1}^n P_{U|y}^K(u_i)$. Given $U^n = u^n$, the MAP test decides the y with the larger $P_Y(y)P_{U|y}^K(u^n)$, so each u^n contributes the smaller of the two to the error probability:

$$P_e(n;K) = \sum_{u^n \in \mathcal{U}_M^n} \min\left(P_Y(0)P_{U|0}^K(u^n),\ P_Y(1)P_{U|1}^K(u^n)\right). \tag{52}$$

By Chernoff's theorem [16], [17, Theorem 11.9.1],

$$\lim_{n\to\infty} -\frac{1}{n}\log P_e(n;K) = \mathcal{C}(K). \tag{53}$$

We also consider the equivocation $H_K(Y|U^n)$, i.e., the log-loss of the Bayes predictor of Y from U^n. For $n = 1$, $H_K(Y|U) = H(Y) - I_K(U;Y)$, so K_{IB} is the best encoder for a single sample under log-loss. For large n, the posterior $P_{Y|U^n}$ is close to 0 or 1 with high probability, and near the endpoints the binary entropy $h(t)$ is about $t\log\frac{1}{t}$; since $\log\frac{1}{P_e}$ grows only linearly in n, the equivocation and the error probability have the same exponent.

Lemma 6.1 (Equivocation and error probability)**.** *For every encoder K and $n \geq 1$, with $P_e = P_e(n;K)$,*

$$2\log 2 \cdot P_e \leq H_K(Y|U^n) \leq P_e(1 - \log P_e). \tag{54}$$

Consequently, $\lim_{n\to\infty} -\frac{1}{n}\log H_K(Y|U^n) = \mathcal{C}(K)$.

Proof. By (1), $P_{U|0}^K(u) > 0$ if and only if $P_{U|1}^K(u) > 0$, so that $P_e > 0$. Write $P(u^n) = \sum_y P_Y(y)P_{U|y}^K(u^n)$ and $q(u^n) = \min_y P_{Y|U^n=u^n}(y) \leq \frac{1}{2}$. Since $P_Y(y)P_{U|y}^K(u^n) = P(u^n)P_{Y|U^n=u^n}(y)$ by Bayes' rule, (52) and the definition of conditional entropy give

$$P_e = \sum_{u^n} P(u^n)\, q(u^n), \qquad H_K(Y|U^n) = \sum_{u^n} P(u^n)\, h(q(u^n)),$$

where $h(t) = -t\log t - (1-t)\log(1-t)$ and we used $h(t) = h(1-t)$. Since h is concave on $[0, \frac{1}{2}]$ with $h(0) = 0$ and $h(\frac{1}{2}) = \log 2$, it lies above its chord, $h(t) \geq 2\log 2 \cdot t$, which gives the lower bound. By the concavity of h and Jensen's inequality, $H_K(Y|U^n) \leq h(P_e)$, which is Fano's inequality for binary Y [17, Theorem 2.10.1]. Moreover, $h(t) \leq t(1 - \log t)$, since $-(1-t)\log(1-t) \leq t$ by $\log z \leq z - 1$ with $z = \frac{1}{1-t}$. This gives the upper bound.

By (53), $-\log P_e = O(n)$, so that $\log(1 - \log P_e) = O(\log n)$. Taking logarithms in (54), dividing by n and using (53) proves the limit. □

Theorem 6.2 (The IB versus the Chernoff bottleneck)**.** *Assume $\mathcal{C}(K_{\mathrm{IB}}) < C(R)$. Then*

(i) $P_e(n; K_{\mathrm{CB}}) < P_e(n; K_{\mathrm{IB}})$ *and* $H_{K_{\mathrm{CB}}}(Y|U^n) < H_{K_{\mathrm{IB}}}(Y|U^n)$ *for all sufficiently large* n*;*

(ii) if moreover $I_{K_{\mathrm{CB}}}(U;Y) < I_{K_{\mathrm{IB}}}(U;Y)$*, then* $H_{K_{\mathrm{IB}}}(Y|U) < H_{K_{\mathrm{CB}}}(Y|U)$*, so that* $n \mapsto H_{K_{\mathrm{IB}}}(Y|U^n) - H_{K_{\mathrm{CB}}}(Y|U^n)$ *changes sign at least once.*

Proof. (i) By (53) and Lemma 6.1, both

$$\log \frac{P_e(n; K_{\mathrm{IB}})}{P_e(n; K_{\mathrm{CB}})} \quad \text{and} \quad \log \frac{H_{K_{\mathrm{IB}}}(Y|U^n)}{H_{K_{\mathrm{CB}}}(Y|U^n)}$$

equal $n\big(C(R) - \mathcal{C}(K_{\mathrm{IB}})\big) + o(n)$, which tends to $+\infty$.

(ii) The inequality at $n = 1$ follows from $H_K(Y|U) = H(Y) - I_K(U;Y)$, and the sign change from (i). □

The two assumptions say that the two objectives disagree at rate R: K_{IB} is not Chernoff-optimal, and K_{CB} is not IB-optimal. They require $R < H(V)$: for $R \geq H(V)$, the IB encoder is sufficient for Y (Section 3), i.e., $X - U - Y$ is a Markov chain, and data processing gives $\mathcal{C}(K_{\mathrm{IB}}) \geq C_X = C(R)$. In terms of the learning curve $n \mapsto I_K(U^n;Y) = H(Y) - H_K(Y|U^n)$, the IB maximizes its first value $I_K(U;Y)$, while the Chernoff bottleneck maximizes the exponential rate $\mathcal{C}(K)$ at which it approaches $H(Y)$.

Remark 6.3 (Error probability for small n). Theorem 6.2(ii) has no counterpart for the error probability, since the IB maximizes $I_K(U;Y)$ rather than $1 - P_e(1;K)$. Indeed, using $\min(t, 1-t) = t - \max(2t-1, 0)$ with $t = P_{Y|U=u}(1)$,

$$P_e(1;K) = P_Y(1) - \sum_{u:\, P_{Y|U=u}(1) > \frac{1}{2}} P_U^K(u)(2P_{Y|U=u}(1) - 1).$$

Assume $P_Y(1) \leq \frac{1}{2}$, so that $P_Y(1)$ is the error without observation. Then an output reduces the error only if its posterior $P_{Y|U=u}(1)$ exceeds $\frac{1}{2}$, whereas $I_K(U;Y) = \sum_u P_U^K(u)\mathrm{D}(P_{Y|U=u}\|P_Y)$ rewards every output whose posterior differs from the prior. When $P_Y(1)$ is small, a posterior above $\frac{1}{2}$ requires a large likelihood ratio $P_{U|1}^K(u)/P_{U|0}^K(u)$, which can favor the Chernoff bottleneck already for small n; Section 7.4 gives an example.

7 Numerical Experiments

7.1 The $C(R)$ Curve

We compute $C(R)$ for the six sources in Table 1, all with $P_Y(0) = P_Y(1) = \frac{1}{2}$. In sources A, C, D and E, some symbols share a likelihood ratio, so that $k < |\mathcal{X}|$ and $H(V) < H(X)$, while in source B all likelihood ratios are distinct, so that $H(V) = H(X)$. Source A is symmetric, C is asymmetric, D has a weak signal, and E has a larger alphabet. Source F is the source of Example 3.7 with $a = 10^{-3}$.

Table 1: Sources used in Section 7.1, with $P_Y(0) = P_Y(1) = \frac{1}{2}$ and $a = 10^{-3}$ (natural logarithms).

	$P_{X\|0}$	$P_{X\|1}$	k	$H(V)$	$H(X)$	C_X
A	$(0.4, 0.2, 0.2, 0.2)$	$(0.2, 0.2, 0.2, 0.4)$	3	1.089	1.366	0.0349
B	$(0.5, 0.3, 0.2)$	$(0.2, 0.3, 0.5)$	3	1.096	1.096	0.0699
C	$(0.5, 0.2, 0.2, 0.1)$	$(0.1, 0.3, 0.3, 0.3)$	3	1.030	1.376	0.1208
D	$(0.3, 0.25, 0.25, 0.2)$	$(0.25, 0.25, 0.25, 0.25)$	3	1.037	1.384	0.0025
E	$(0.2, 0.15, 0.15, 0.1, 0.1, 0.1, 0.1, 0.1)$	$(0.05, 0.075, 0.075, 0.1, 0.1, 0.2, 0.2, 0.2)$	4	1.277	2.066	0.0658
F	$(1-a, a)$	$(a, 1-a)$	2	0.693	0.693	2.7612

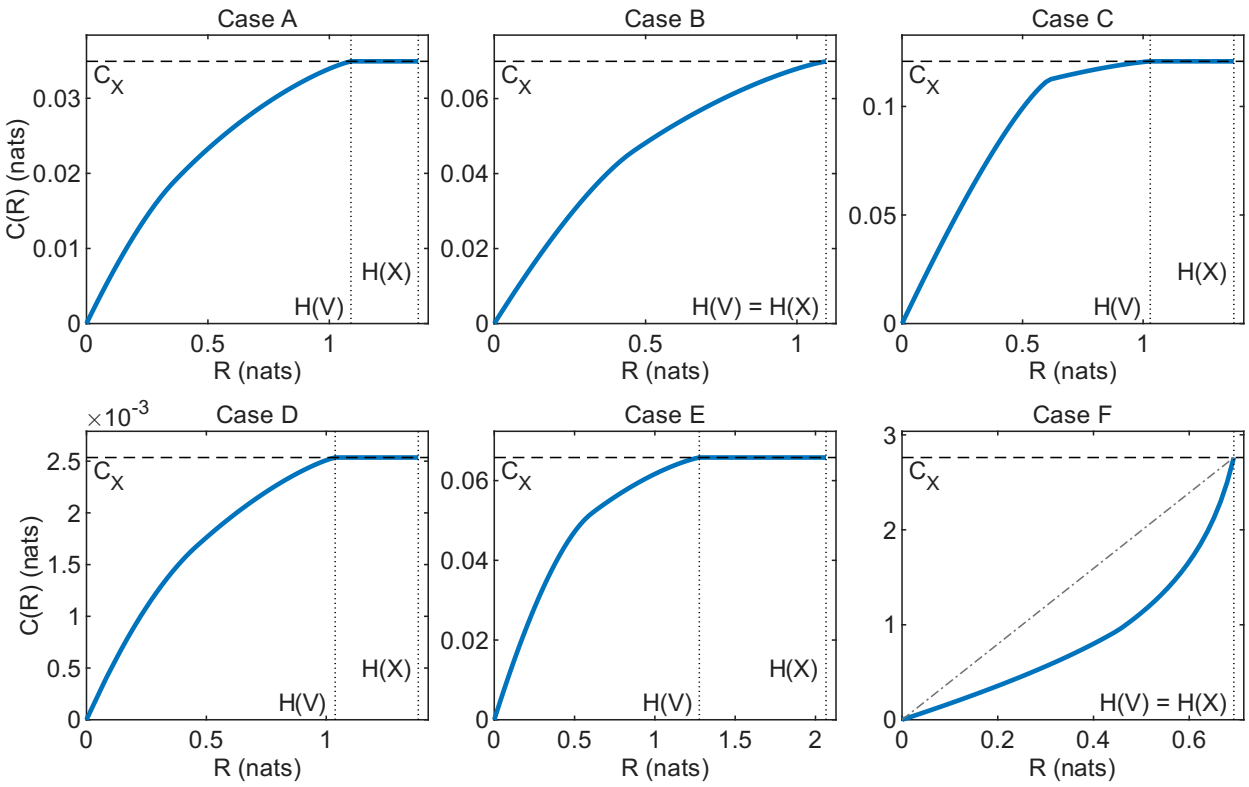


Figure 1: $C(R)$ for the sources of Table 1. Dashed: C_X. Dotted: $H(V)$ and $H(X)$. Dash-dotted (F): the chord from $(0,0)$ to $(H(V), C_X)$.

For each source we set $M = k+1$, which suffices by Theorem 4.4, and run Algorithm 2 at 50 evenly spaced rates in $(0, H(V)]$ and, when $H(V) < H(X)$, at 10 rates in $(H(V), H(X)]$. Since maximizing $\mathcal{C}(K)$ amounts to minimizing a concave function (Remark 5.3), the algorithm may stop at a local optimum, so at each rate we keep the best of $k+3$ initializations: two random encoders; the time-sharing encoder of Lemma 3.4 between K^V and a constant encoder; and, for each $j = 1, \cdots, k$, the encoder that outputs j with probability θ when $X \in \mathcal{X}_j$ and a constant symbol otherwise. In the last two, θ is the largest value with $I_K(U;X) \le R$, and the encoder is mixed with weight 10^{-9} with a random feasible encoder, so that it lies in $\mathcal{K}_{R,+}^{(M)}$. Every plotted value is attained by an explicit encoder and is therefore a lower bound on $C(R)$.

Figure 1 shows the results. In every case, $C(R)$ is strictly increasing on $[0, H(V)]$ and equals C_X on $[H(V), H(X)]$, in agreement with Theorem 3.5: rate beyond $H(V)$ only distinguishes symbols with the same likelihood ratio, which does not help the test. For source B, $H(V) = H(X)$ and the curve has no flat part. Source F illustrates Example 3.7: its whole curve lies below the chord from $(0,0)$ to $(H(V), C_X)$. Since B is concave while both B and F have $H(V) = H(X)$, the non-concavity is not a consequence of $H(V) = H(X)$.

7.2 Structure of the Optimal Encoder

We examine the optimal encoders of source C in Table 1, whose classes $\mathcal{X}_1 = \{1\}$, $\mathcal{X}_2 = \{2,3\}$, $\mathcal{X}_3 = \{4\}$ have likelihood ratios $\rho_1 = 0.2$, $\rho_2 = 1.5$, $\rho_3 = 3$, and whose curve in Figure 1 has a kink near $R = 0.61$. We take three rates: $R = 0.30$; $R = h(P_V(1)) = h(0.3) \approx 0.61$, where h is the binary entropy function, which is the rate of the deterministic encoder that separates $\mathcal{X}_1$ from the other

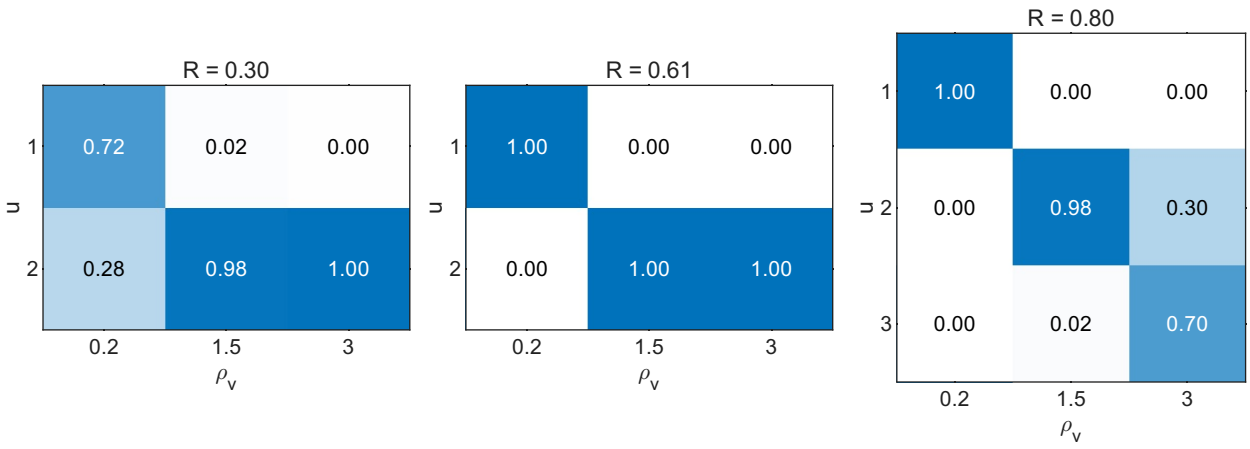


Figure 2: Optimal encoders $\bar{K}^*(u|v)$ of source C at three rates. Columns: classes v, labeled by ρ_v. Rows: used outputs u, sorted by $P_{U|1}(u)/P_{U|0}(u)$.

classes; and $R = 0.80$. At each rate we compute the best encoder K^* as in Section 7.1 and map it, as in Lemma 4.3(i), to the encoder of V

$$\bar{K}^*(u|j) = \sum_{x \in \mathcal{X}_j} \frac{P_X(x)}{P_V(j)} K^*(u|x),$$

which has the same $P_{U|0}$ and $P_{U|1}$ and no larger rate. Outputs with $P_U(u) \le 10^{-4}$ are dropped, outputs with equal likelihood ratio are merged, and the remaining outputs are sorted by $P_{U|1}(u)/P_{U|0}(u)$.

Figure 2 shows three features. First, the optimal encoders are randomized: at $R = 0.30$, class 1 is sent to output 1 with probability 0.72, and at $R = 0.80$, class 3 is split between outputs 2 and 3 with probabilities 0.30 and 0.70. This differs from quantizer design under an alphabet constraint, where deterministic likelihood-ratio quantizers are optimal [35]. Second, the encoders are monotone in the likelihood ratio: each output collects a contiguous range of classes in the order of ρ_v, so that $\bar{K}^*$ is a randomized likelihood-ratio quantizer. The same randomized, monotone structure appears for source E. Third, at $R = h(0.3)$ the optimal encoder is the deterministic quantizer $\{\mathcal{X}_1\}, \{\mathcal{X}_2 \cup \mathcal{X}_3\}$, and this rate is exactly where the curve of source C has its kink. Below it, the encoder randomizes the separation of $\mathcal{X}_1$; above it, $\mathcal{X}_1$ stays separated and the encoder randomizes the separation of $\mathcal{X}_2$ and $\mathcal{X}_3$ with a third output.

7.3 Real Dataset Practice

We apply the Chernoff bottleneck to topic detection, where a document is observed word by word and every word is compressed separately before the decision, as in decentralized detection [19, 20]. We use the 20 Newsgroups corpus [29] in its date-sorted train/test split, with the topics `rec.sport.hockey` ($Y = 0$) and `sci.med` ($Y = 1$), which have 598 and 594 training documents and 399 and 393 test documents. The alphabet $\mathcal{X}$ consists of the 1000 most frequent words in the training documents of the two topics, and all other words are discarded. From the training word counts $N_y(x)$ we estimate

$$P_{X|y}(x) = \frac{N_y(x) + 1}{\sum_{x'} N_y(x') + |\mathcal{X}|}, \quad y = 0, 1,$$

which satisfies (1), and set $P_Y(0) = P_Y(1) = \frac{1}{2}$. This gives $k = 826$, $H(X) = 5.80$ and $C_X = 0.158$ nats per word.

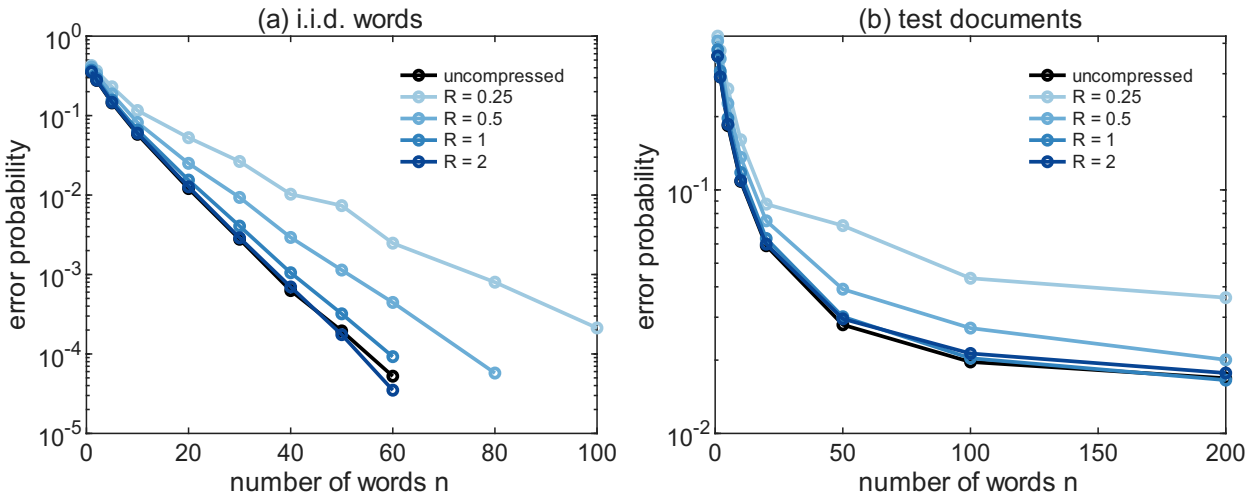


Figure 3: Test error of topic detection (hockey vs. medicine) from n words, each compressed to R nats by the Chernoff bottleneck, compared with the uncompressed naive Bayes classifier. (a) Words drawn i.i.d. from each topic. (b) Words drawn from each test document.

Table 2: Rate, Chernoff information, and the fitted decay rate of the error in Figure 3(a), i.e., the slope of $-\log P_e$ in n over $10^{-4} \le P_e \le 10^{-1}$ (nats).

R	$R/H(X)$	$C(R)$	$C(R)/C_X$	fitted slope
0.25	0.04	0.073	0.47	0.071
0.5	0.09	0.113	0.71	0.107
1	0.17	0.142	0.90	0.128
2	0.34	0.154	0.98	0.141
uncompressed	1	0.158	1	0.151

We compute encoders with $M = 16$ outputs at $R \in \{0.25, 0.5, 1, 2\}$ nats per word by Algorithm 2, keeping at each rate the best of four initializations: two random encoders, the encoder of the previous rate, and a likelihood-ratio quantizer that divides the words, in the order of $\rho(x)$, into $M-1$ groups of equal P_X-mass and is time-shared with a constant symbol so that $I_K(U;X) \le R$. Since $M < k+1$, the resulting values of $\mathcal{C}(K)$, reported as $C(R)$ in Table 2, are lower bounds on $C(R)$. Each word is encoded independently by the same K, and the decision on $U^n = (U_1, \cdots, U_n)$ is the MAP test

$$\hat{Y} = 1 \iff \sum_{i=1}^{n} \log \frac{P^K_{U|1}(U_i)}{P^K_{U|0}(U_i)} > 0,$$

where $P^K_{U|y}$ is computed from the training estimates. Without compression ($U_i = X_i$), this test is the naive Bayes classifier. We measure the error on the test data in two ways, in both cases averaged over the two topics: (a) the n words are drawn i.i.d. from the word frequencies of each topic in the test documents (2×10^5 repetitions per n and topic), so that the model of Section 2 holds up to the mismatch between training and test data; (b) the n words are drawn with replacement from a single test document (200 repetitions per document), which is the actual task.

The computed encoders use 3, 8, 12 and 12 of the 16 outputs, and all of them are randomized likelihood-ratio quantizers, as in Section 7.2: each output collects words of similar $\rho(x)$, and the

Table 3: IB versus Chernoff bottleneck. n_c: the IB has the smaller $H_K(Y|U^n)$ or $P_e(n;K)$ exactly for $n < n_c$. $P_e(n_c)$: error of K_{CB} at the crossover. n_ϵ: the smallest n with $P_e(n;K) \le \epsilon$, for K_{IB} / K_{CB}.

			n_c				
R	$C(R)$	$\mathcal{C}(K_{\mathrm{IB}})$	H	P_e	$P_e(n_c)$	$n_{10^{-2}}$	$n_{10^{-4}}$
0.0168	0.0157	0.0120	18	24	0.23	228 / 180	> 500 / 460
0.0337	0.0318	0.0240	8	14	0.20	112 / 89	288 / 226
0.0674	0.0638	0.0483	5	8	0.18	57 / 45	143 / 107
0.1347	0.1281	0.0986	3	4	0.17	28 / 23	71 / 53

assignment is monotone in $\rho(x)$. Table 2 shows how much of the exponent is kept: at $R = 1$ nat per word, which is 17% of $H(X)$, the encoder keeps 90% of C_X, and even at $R = 0.25$, which is 4% of $H(X)$, it keeps 47%.

Figure 3(a) confirms that $C(R)$ predicts the performance on real word statistics: the error decays exponentially in n, with fitted rates within about 10% of $C(R)$, the small shortfall being due to the polynomial factor in n and the mismatch between training and test data. Consequently, compression to R nats per word multiplies the number of words needed for a given error by about $C_X/C(R)$, i.e., by 1.11 at $R = 1$ and 1.40 at $R = 0.5$. Figure 3(b) shows the actual task. The curves are ordered as $C(R)$; at $R = 1$ the error is close to that of the uncompressed classifier for all n, and at $R = 2$ the two are indistinguishable. The error levels off near 2% rather than decaying as e^{-nC}, because the words of a document follow the document's own distribution rather than being i.i.d. given the topic; the uncompressed classifier has the same floor, so it is not caused by compression. Finally, the benefit of the mutual information constraint is largest at low rates: at $R = 0.25$, the best deterministic likelihood-ratio quantizer with at most five cells and entropy at most R has Chernoff information 0.059, against 0.073 for the Chernoff bottleneck, whereas at $R \ge 1$ both are within a few percent of C_X.

7.4 Comparison with the IB

We compare the encoders K_{IB} and K_{CB} of Section 6. When all likelihood ratios are moderate, the two nearly coincide, so we use the source

$$P_{X|0} = (0.26, 0.37, 0.365, 0.005), \qquad P_{X|1} = (0.25, 0.02, 0.005, 0.725),$$

with $P_Y(0) = P_Y(1) = \frac{1}{2}$, whose likelihood ratios $\rho(x) = 0.96, 0.054, 0.014, 145$ span a wide range, so that $k = 4$, $H(V) = H(X) = 1.347$ and $C_X = 0.820$. We take $M = k + 1 = 5$ and the low rates $R = 0.0168, 0.0337, 0.0674, 0.1347$.

The encoders K_{CB} are computed by Algorithm 2, and K_{IB} by the same scheme with $B_M(\cdot, s)$ replaced by $I_K(U;Y)$: since $I_K(U;Y)$ is convex in K, maximizing its tangent plane at the current encoder is the problem (32) with $C^s_{u,x}$ replaced by $P_X(x)\mathrm{D}(P_{Y|X=x}\|P_{Y|U=u})$, which Algorithm 1 solves with the same bisection and safeguard. Unlike the iterative IB algorithm [8], the bisection meets the rate constraint exactly. For both problems, we keep the best of several initializations, including the best encoder of the other problem. For $n = 1, \cdots, 500$, $P_e(n;K)$ and $H_K(Y|U^n)$ are computed without Monte Carlo, from the law of the sum of the log-likelihood ratios of $U_1, \cdots, U_n$, obtained by n-fold convolution via the FFT.

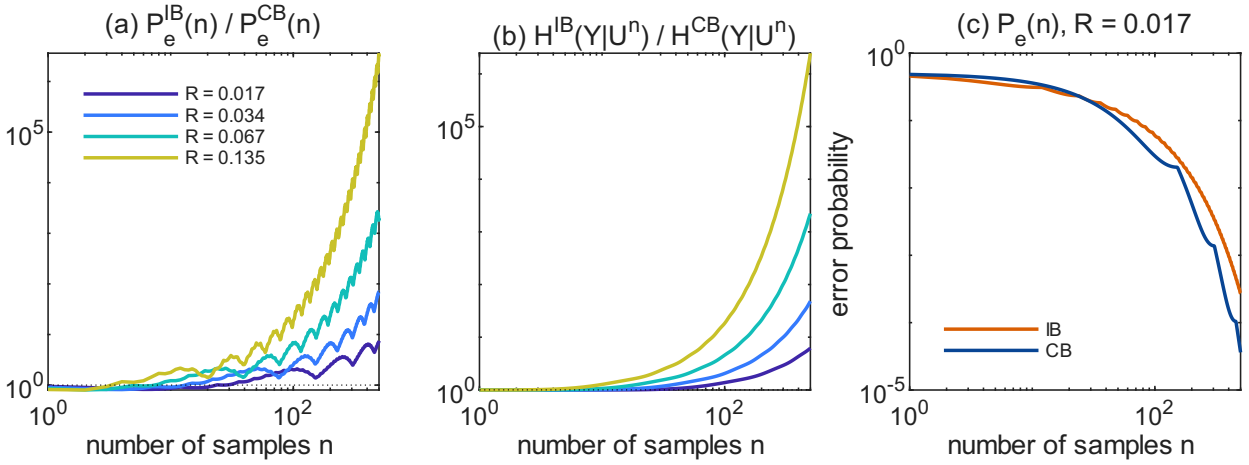


Figure 4: IB versus Chernoff bottleneck (CB) encoders. (a) Ratio of error probabilities $P_e(n; K_{\mathrm{IB}})/P_e(n; K_{\mathrm{CB}})$. (b) Ratio of equivocations $H_{K_{\mathrm{IB}}}(Y|U^n)/H_{K_{\mathrm{CB}}}(Y|U^n)$. Values above 1 (dotted) mean that the Chernoff bottleneck is better. (c) Error probabilities at $R = 0.017$.

At every rate, the IB has the larger $I_K(U;Y)$ and the Chernoff bottleneck an exponent larger by 23–25%, so Theorem 6.2 applies. As predicted, the ratio of equivocations in Figure 4(b) starts below 1 and crosses 1 exactly once. The error probabilities behave in the same way, with crossovers $n_c = 24, 14, 8, 4$. The crossover occurs while the error is still large, at $n_c C(R) \approx 0.4$–0.5 and $P_e \approx 0.2$, and n_c grows like $1/R$. For smaller errors, the Chernoff bottleneck needs 18–25% fewer samples (Table 3), close to the asymptotic saving $1 - \mathcal{C}(K_{\mathrm{IB}})/C(R) \approx 24\%$.

The encoders explain this. Both are essentially binary, with one output, the alarm, carrying the evidence for $Y = 1$. The Chernoff bottleneck sends the alarm only for $x = 4$, rarely ($P_U^K(u) = 0.016$ to 0.12) but with likelihood ratio $\rho(4) = 145$; the IB sends it also for the other symbols, more often (0.09 to 0.29), but with likelihood ratio only 2.8 to 4.6. A frequent but weak alarm is better for one sample, a rare but decisive one for many. Since $P_e(n; K)$ depends essentially only on the number of alarms, it decreases in steps, which causes the oscillations in Figure 4(a).

Finally, as Remark 6.3 indicates, the small-n advantage of the IB in error probability depends on the prior. For the same source with $P_Y(0) = 0.9$ and $R = H(V)/10 = 0.128$, the IB has the larger $I_K(U;Y)$, yet the Chernoff bottleneck has the smaller error for $n = 1, \cdots, 5$, e.g., 0.0229 against 0.0341 at $n = 4$. Both alarms have posterior $P_{Y|U=u}(1)$ above $\frac{1}{2}$, and the purer one of the Chernoff bottleneck (0.94 against 0.88) reduces $P_e(1; K)$ more.

8 Conclusion

We studied the MI constrained Chernoff bottleneck, which replaces the relevance $I(U;Y)$ of the classical IB by the Chernoff information of the downstream binary test. The curve $C(R)$ is governed by the variable V that merges the symbols of X with equal likelihood ratio: $C(R)$ increases strictly up to $R = H(V)$, where it reaches the uncompressed exponent C_X, and stays constant beyond, indicating the additional rates beyond $H(V)$ is wasted. (Theorem 3.5). For $R \le H(V)$, every optimal encoder uses the full rate budget (Corollary 3.6). Unlike the classical IB curve, which is concave by time-sharing, the $C(R)$ curve need not be concave (Example 3.7). An optimal encoder exists with at most $|\mathcal{X}| + 1$ outputs (Theorem 4.2) and, by the reduction to V, with at most $k + 1$ outputs, where k is the number of distinct likelihood ratios (Theorem 4.4).

Since the Chernoff information is not an expectation of a per-letter quantity, the classical Blahut–Arimoto algorithm does not apply directly. We therefore proposed Algorithm 2, which alternates between the encoder K and the parameter s. For fixed s, the encoder minimizes the tangent plane of the concave function $B_M(\cdot, s)$ at the current encoder under the MI constraint (Remark 5.3), which is a rate–distortion problem solved by a generalized Blahut–Arimoto algorithm (Algorithm 1) with a bisection over the trade-off parameter λ. Algorithm 1 converges to a global minimizer of the Lagrangian at rate $O(1/t)$ with a stopping certificate (Theorem 5.8, Corollary 5.9), and the bisection has an explicit bracket and certificate (Lemma 5.10). For fixed encoder K, the parameter s is the unique root of a nonlinear equation (Lemma 5.6). The iterates of Algorithm 2 remain feasible, s stays uniformly bounded away from 0 and 1, and the Chernoff information is nondecreasing and convergent (Lemma 5.12, Theorem 5.13, Remark 5.14). Since the problem amounts to minimizing a concave function, the limit need not be globally optimal, which we mitigate by structured initializations.

We also compared the Chernoff bottleneck with the classical IB when Y is inferred from n samples. The equivocation $H(Y|U^n)$ decays at the same exponential rate as the error probability, namely the Chernoff information (Lemma 6.1). Hence, when the two objectives disagree, the IB encoder, which is the best for a single sample under log-loss, is overtaken by the Chernoff bottleneck as n grows, in both equivocation and error probability (Theorem 6.2). For the error probability, the IB need not be better even for small n, since an output reduces the error only if it moves the posterior across $\frac{1}{2}$ (Remark 6.3).

The numerical experiments confirm the theory. Figure 1 shows the computed $C(R)$ curves of six synthetic sources: each curve increases strictly up to $H(V)$ and stays at C_X beyond; for source B, whose likelihood ratios are all distinct, $H(V) = H(X)$ and the curve has no flat part; and the curve of source F, which is the source of Example 3.7, lies entirely below its chord (Section 7.1). Figure 2 shows the computed encoders of source C at three rates. They are randomized likelihood-ratio quantizers, each output collecting a contiguous range of likelihood ratios, in contrast to quantizer design under an alphabet constraint, where deterministic likelihood-ratio quantizers are optimal. At the kink of $C(R)$, the encoder is exactly a deterministic quantizer (Section 7.2). In a real-data application, topic detection on the 20 Newsgroups corpus, compressing each word to 1 nat, only 17% of its entropy, keeps 90% of the error exponent (Table 2). When the words are drawn i.i.d. from the test word frequencies, the error decays exponentially at a rate within about 10% of $C(R)$ (Figure 3(a)). On the actual test documents, the errors are ordered as $C(R)$, and at 1 nat the classifier is nearly as accurate as the uncompressed one (Figure 3(b)). The advantage over deterministic likelihood-ratio quantization is largest at low rates (Section 7.3). Compared with the IB on a source with widely spread likelihood ratios, the IB has the smaller error only while the error is about 0.2 or larger, and below that the Chernoff bottleneck needs 18–25% fewer samples (Section 7.4).

More broadly, the Chernoff bottleneck judges a representation by how fast the error of the downstream decision decays, rather than by an average measure of relevance. We believe this makes it a principled criterion for learning compact representations for inference, and hope that the theory, algorithm, and experiments presented here serve as a foundation for its further study.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[4] Ravid Shwartz-Ziv and Yann LeCun, "To compress or not to compress—Self-supervised learning and information theory: A review," *Entropy*, vol. 26, no. 3, Art. no. 252, 2024.

[5] Oriane Siméoni, Huy Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski, "DINOv3," arXiv:2508.10104, 2025.

[6] Dier Tang, Jing Yee Tan, and Guangyue Han, "Sharp regret bounds and a task-covariance correction for spectral representation learning," arXiv:2609.15825v2, 2026.

[7] Achleshwar Luthra, Tianbao Yang, and Tomer Galanti, "Self-supervised contrastive learning is approximately supervised contrastive learning," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[8] Naftali Tishby, Fernando Pereira, and William Bialek, "The information bottleneck method," in *Proc. 37th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, 1999, pp. 368–377.

[9] Naftali Tishby and Noga Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Information Theory Workshop (ITW)*, Jerusalem, Israel, 2015, pp. 1–5.

[10] Ravid Shwartz-Ziv and Naftali Tishby, "Opening the black box of deep neural networks via information," arXiv:1703.00810, 2017.

[11] Alexander Alemi, Ian Fischer, Joshua Dillon, and Kevin Murphy, "Deep variational information bottleneck," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[12] Artemy Kolchinsky, Brendan Tracey, and David Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, no. 12, Art. no. 1181, 2019.

[13] Alessandro Achille and Stefano Soatto, "Emergence of invariance and disentanglement in deep representations," *Journal of Machine Learning Research*, vol. 19, no. 50, pp. 1–34, 2018.

[14] Ohad Shamir, Sivan Sabato, and Naftali Tishby, "Learning and generalization with the information bottleneck," *Theoretical Computer Science*, vol. 411, no. 29–30, pp. 2696–2711, 2010.

[15] Matías Vera, Pablo Piantanida, and Leonardo Rey Vega, "The role of the information bottleneck in representation learning," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, USA, 2018, pp. 1580–1584.

[16] Herman Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

[17] Thomas Cover and Joy Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[18] Amir Dembo and Ofer Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York, NY, USA: Springer, 1998.

[19] John Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing*, vol. 2, H. Vincent Poor and John Thomas, Eds. Greenwich, CT, USA: JAI Press, 1993, pp. 297–344.

[20] Jean-François Chamberland and Venugopal Veeravalli, "Decentralized detection in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 407–416, 2003.

[21] Jiawei Shao, Yuyi Mao, and Jun Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 197–211, 2022.

[22] Rudolf Ahlswede and Imre Csiszár, "Hypothesis testing with communication constraints," *IEEE Transactions on Information Theory*, vol. 32, no. 4, pp. 533–542, 1986.

[23] Te Sun Han, "Hypothesis testing with multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 33, no. 6, pp. 759–772, 1987.

[24] Richard Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[25] Suguru Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[26] Hans Witsenhausen and Aaron Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 5, pp. 493–501, 1975.

[27] Claude Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[28] Toby Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[29] Ken Lang, "NewsWeeder: Learning to filter netnews," in *Proc. 12th International Conference on Machine Learning (ICML)*, 1995, pp. 331–339.

[30] Tim van Erven and Peter Harremoës, "Rényi divergence and Kullback–Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[31] Imre Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.

[32] Richard Burden and J. Douglas Faires, *Numerical Analysis*, 9th ed. Boston, MA, USA: Brooks/Cole, 2011.

[33] Richard Brent, *Algorithms for Minimization without Derivatives*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1973.

[34] William Press, Saul Teukolsky, William Vetterling, and Brian Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge, U.K.: Cambridge University Press, 2007.

[35] John Tsitsiklis, "Extremal properties of likelihood-ratio quantizers," *IEEE Transactions on Communications*, vol. 41, no. 4, pp. 550–558, 1993.

[36] Walter Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York, NY, USA: McGraw-Hill, 1976.

A Several Facts Used in the Proof

Fact A.1 (Young's inequality). *For any $x, y > 0$, and $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, we have*

$$xy \le \frac{x^p}{p} + \frac{y^q}{q},$$

equality holds when $x^p = y^q$.

Fact A.2 (Hölder's inequality). *For nonnegative $a_1, \cdots, a_n$, $b_1, \cdots, b_n$, and $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\sum_{i=1}^{n} a_i b_i \le \Big(\sum_{i=1}^{n} a_i^p\Big)^{1/p} \Big(\sum_{i=1}^{n} b_i^q\Big)^{1/q},$$

with equality if and only if there exist $\alpha, \beta \ge 0$, not both zero, such that $\alpha a_i^p = \beta b_i^q$ for all i.

Fact A.3 (Weighted average inequality). *For finite k and $a_i, b_i > 0$, $i = 1, \cdots, k$, we have*

$$\min_{1 \le i \le k} \frac{a_i}{b_i} \le \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} b_i} \le \max_{1 \le i \le k} \frac{a_i}{b_i}.$$

Fact A.4 (Heine-Borel Theorem). *A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded. [36, Theorem 2.41]*

Fact A.5 (Joint convexity of relative entropy). *For probability distributions P_i, Q_i, $i = 1, \cdots, n$, on a finite set and weights $\pi_i \ge 0$ with $\sum_i \pi_i = 1$,*

$$\mathrm{D}\Big(\sum_{i=1}^{n} \pi_i P_i \,\Big\|\, \sum_{i=1}^{n} \pi_i Q_i\Big) \le \sum_{i=1}^{n} \pi_i \mathrm{D}(P_i \| Q_i).$$

[17, Theorem 2.7.2]

Fact A.6 (Pinsker's inequality). *For probability distributions P, Q on a finite set, with natural logarithms,*

$$\mathrm{D}(P\|Q) \ge \frac{1}{2}\|P - Q\|_1^2.$$

[17, Lemma 11.6.1]

B Proofs

B.1 Proof of Lemma 3.2

$V = \nu(X)$ gives $H(V|X) = 0$. Thus, $I_{K^V}(U; X) = H(V) - H(V|X) = H(V)$.
Fix $j \in \{1, \cdots, k\}$ and $s \in [0, 1]$. On the class $\mathcal{X}_j$ we have $P_{X|1}(x) = \rho_j P_{X|0}(x)$, hence

$$\sum_{x \in \mathcal{X}_j} P_{X|0}(x)^s P_{X|1}(x)^{1-s} = \sum_{x \in \mathcal{X}_j} P_{X|0}(x)^s (\rho_j P_{X|0}(x))^{1-s} = \rho_j^{1-s} P_{X|0}(\mathcal{X}_j).$$

And since $P_{U|y}^{K^V}(j) = P_{X|y}(\mathcal{X}_j)$ for $y \in \{0, 1\}$, and $P_{X|1}(\mathcal{X}_j) = \rho_j P_{X|0}(\mathcal{X}_j)$, we derive

$$P_{U|0}^{K^V}(j)^s P_{U|1}^{K^V}(j)^{1-s} = P_{X|0}(\mathcal{X}_j)^s (\rho_j P_{X|0}(\mathcal{X}_j))^{1-s} = \rho_j^{1-s} P_{X|0}(\mathcal{X}_j) = \sum_{x \in \mathcal{X}_j} P_{X|0}(x)^s P_{X|1}(x)^{1-s}.$$

Summing over j and using $\mathcal{X} = \bigsqcup_j \mathcal{X}_j$ yields $B_k(K^V, s) = B_{|\mathcal{X}|}(K^X, s)$. Taking the minimum over $s \in [0, 1]$ gives $C(P_{U|0}^{K^V}, P_{U|1}^{K^V}) = C_X$.

B.2 Proof of Lemma 3.3

For $s \in \{0,1\}$ both sides of (15) equal 1, so let $s \in (0,1)$. Applying Hölder's inequality (Fact A.2) with exponents $\frac{1}{s}$ and $\frac{1}{1-s}$ to $(K(u|x)P_{X|0}(x))^s$ and $(K(u|x)P_{X|1}(x))^{1-s}$ for all $x \in \mathcal{X}$ gives, for each $u \in \mathcal{U}_M$,

$$\begin{aligned}\sum_x K(u|x)P_{X|0}(x)^s P_{X|1}(x)^{1-s} &\le \Big(\sum_x K(u|x)P_{X|0}(x)\Big)^s \Big(\sum_x K(u|x)P_{X|1}(x)\Big)^{1-s} \\ &= P^K_{U|0}(u)^s P^K_{U|1}(u)^{1-s}.\end{aligned} \tag{55}$$

Summing (55) over $u \in \mathcal{U}_M$ and using $\sum_{u\in\mathcal{U}_M} K(u|x) = 1$ derives (15).

Suppose that $B_M(K,s) = B_{|\mathcal{X}|}(K^X, s)$ for some $s \in (0,1)$. Then (55) holds with equality for every $u \in \mathcal{U}_M$. By (1), for any $u \in \mathcal{U}_M$ with $P^K_U(u) > 0$, the equality condition of Fact A.2 indicates there exists $c_u > 0$ such that

$$K(u|x)P_{X|1}(x) = c_u K(u|x)P_{X|0}(x), \quad \forall x \in \mathcal{X}.$$

This means $\rho(x) = c_u$ for all x with $K(u|x) > 0$, which implies U determines $V = \nu(X)$, i.e., $V = g(U)$ almost surely for some function g. Hence,

$$I_K(U;X) \ge I(g(U);X) = I(V;X) = H(V) - H(V|X) = H(V).$$

Taking the contrapositive gives the conclusion.

B.3 Proof of Lemma 3.4

Since $Q = \mathbf{1}\{U \le M_1\}$ is a function of U, and Q is independent of X, we have

$$\begin{aligned}I_K(U;X) &= I(Q;X) + I(U;X|Q) \\ &= P(Q=1)I(U;X|Q=1) + P(Q=0)I(U;X|Q=0) \\ &= \theta I_{K_1}(U;X) + (1-\theta)I_{K_2}(U;X).\end{aligned}$$

For $y \in \{0,1\}$, $P^K_{U|y}(u) = \theta P^{K_1}_{U|y}(u)$ for $u \le M_1$ and $P^K_{U|y}(u) = (1-\theta)P^{K_2}_{U|y}(u-M_1)$ for $u > M_1$. Using $(c\alpha)^s(c\beta)^{1-s} = c\,\alpha^s\beta^{1-s}$ for $c, \alpha, \beta \ge 0$, we have

$$\begin{aligned}B_{M_1+M_2}(K,s) &= \sum_{u=1}^{M_1} \big(\theta P^{K_1}_{U|0}(u)\big)^s \big(\theta P^{K_1}_{U|1}(u)\big)^{1-s} + \sum_{u=M_1+1}^{M_1+M_2} \big((1-\theta)P^{K_2}_{U|0}(u-M_1)\big)^s \big((1-\theta)P^{K_2}_{U|1}(u-M_1)\big)^{1-s} \\ &= \theta \sum_{u=1}^{M_1} P^{K_1}_{U|0}(u)^s P^{K_1}_{U|1}(u)^{1-s} + (1-\theta)\sum_{u=1}^{M_2} P^{K_2}_{U|0}(u)^s P^{K_2}_{U|1}(u)^{1-s} \\ &= \theta B_{M_1}(K_1,s) + (1-\theta)B_{M_2}(K_2,s).\end{aligned}$$

B.4 Proof of Example 3.7

Here $V = X$, so $H(V) = \log 2$, and by Theorem 3.5 and symmetry, $C(\log 2) = C_X = -\log\big(2\sqrt{a(1-a)}\big) \ge 2.76$. Since $C(0) = 0$, concavity would give $C(R_0) \ge \frac{1}{10}C_X \ge 0.276$ at $R_0 = \frac{1}{10}\log 2$. We show $C(R_0) < 0.26$.

Let K be an encoder with $I_K(U;X) \le R_0$, write $K_x = K(\cdot|x)$ and $\beta(P,Q) = \sum_u \sqrt{P(u)Q(u)}$.

(i) $\psi(s) = \log B_M(K,s)$ is convex with $\psi(0) = \psi(1) = 0$. For $s \le \frac{1}{2}$, convexity on $[s,1]$ gives $\psi(\frac{1}{2}) \le \frac{\psi(s)}{2(1-s)}$, so $\psi(s) \ge 2\psi(\frac{1}{2})$, and symmetrically for $s \ge \frac{1}{2}$. Hence $\mathcal{C}(K) \le -2\log\beta(P^K_{U|0}, P^K_{U|1})$.

(ii) $P^K_{U|0} = (1-a)K_1 + aK_2$, $P^K_{U|1} = aK_1 + (1-a)K_2$, and β is jointly concave, so $\beta(P^K_{U|0}, P^K_{U|1}) \geq \beta(K_1, K_2)$.

(iii) With $m = \frac{K_1+K_2}{2}$, $t_u = \frac{K_1(u)}{2m(u)}$ and the binary entropy h, we have $I_K(U;X) = \sum_u m(u)\big(\log 2 - h(t_u)\big)$ and $\beta(K_1, K_2) = \sum_u 2m(u)\sqrt{t_u(1-t_u)}$. Since $\log y \leq \frac{y-1}{\sqrt{y}}$ for $y \geq 1$, taking $y = \frac{1}{t}$ and $y = \frac{1}{1-t}$ gives $h(t) \leq (1-t)\sqrt{t} + t\sqrt{1-t} \leq \sqrt{2t(1-t)}$, so $\beta(K_1, K_2) \geq \sqrt{2}\big(\log 2 - I_K(U;X)\big)$.

Combining (i)–(iii), $\mathcal{C}(K) \leq -\log\left(2(\log 2 - R_0)^2\right) = 0.2506$ for every such K. Hence $C(R_0) \leq 0.2506 < 0.276$.

B.5 Proof of Lemma 4.1

We prove non-emptiness, compactness, and convexity in turn.

For non-emptiness. Let K_0 be a constant encoder, i.e.,

$$K_0(1|x) = 1,\ K_0(u|x) = 0,\ \forall x \in \mathcal{X}, u = 2, \cdots, M.$$

Then $I_{K_0}(U;X) = 0 \leq R$. Thus, $\mathcal{K}_R^{(M)}$ is non-empty.

For compactness. Let $\{K_n\}_{n=1}^{\infty} \subseteq \mathcal{K}_R^{(M)}$ be a sequence such that $K_n \to K^*$. Since $\Delta_{U|X}$ is the direct product of a finite number of simplexes, which are closed, indicating that $\Delta_{U|X}$ is closed, thus we have $K^* \in \Delta_{U|X}$. Moreover, $I(U;X)$ is a continuous functional of $K_{U|X}$. Then,

$$I_{K^*}(U;X) = I_{\lim_{n\to\infty} K_n}(U;X) = \lim_{n\to\infty} I_{K_n}(U;X) \leq \lim_{n\to\infty} R = R.$$

Thus, $K^* \in \mathcal{K}_R^{(M)}$, i.e., $\mathcal{K}_R^{(M)}$ is closed. Since $\mathcal{K}_R^{(M)}$ is bounded, by Heine-Borel Theorem, $\mathcal{K}_R^{(M)}$ is compact.

For convexity, let $K_1, K_2 \in \mathcal{K}_R^{(M)}$ and $\theta \in [0,1]$. Define $K_\theta(u|x) = \theta K_1(u|x) + (1-\theta)K_2(u|x)$. For any $x \in \mathcal{X}$, we have $K_\theta(u|x) \geq 0$ and $\sum_{u\in\mathcal{U}_M} K_\theta(u|x) = \theta \sum_{u\in\mathcal{U}_M} K_1(u|x) + (1-\theta)\sum_{u\in\mathcal{U}_M} K_2(u|x) = 1$. Thus, $K_\theta \in \Delta_{U|X}$. For fixed P_X, the mutual information induced by encoder K is

$$I_K(U;X) = \mathrm{D}\left(P^K_{XU} \middle\| P_X P^K_U\right),$$

where

$$P^K_{XU}(x,u) = P_X(x)K(u|x), \qquad P^K_U(u) = \sum_{x\in\mathcal{X}} P_X(x)K(u|x).$$

Since both P^K_{XU} and $P_X P^K_U$ are affine functions of K, we have

$$P^{K_\theta}_{XU} = \theta P^{K_1}_{XU} + (1-\theta)P^{K_2}_{XU}, \quad P_X P^{K_\theta}_U = \theta P_X P^{K_1}_U + (1-\theta) P_X P^{K_2}_U.$$

By the joint convexity of relative entropy,

$$\begin{aligned} I_{K_\theta}(U;X) &= \mathrm{D}\left(P^{K_\theta}_{XU} \middle\| P_X P^{K_\theta}_U\right) \\ &\leq \theta \mathrm{D}\left(P^{K_1}_{XU} \middle\| P_X P^{K_1}_U\right) + (1-\theta)\mathrm{D}\left(P^{K_2}_{XU} \middle\| P_X P^{K_2}_U\right) \\ &= \theta I_{K_1}(U;X) + (1-\theta) I_{K_2}(U;X) \\ &\leq \theta R + (1-\theta)R = R, \end{aligned}$$

which gives $K_\theta \in \mathcal{K}_R^{(M)}$. Hence, $\mathcal{K}_R^{(M)}$ is convex.

B.6 Proof of Lemma 4.3

Proof of (i). Fix $j \in \{1, \cdots, k\}$. On $\mathcal{X}_j$ we have $P_{X|1}(x) = \rho_j P_{X|0}(x)$, and summing over $\mathcal{X}_j$ gives $P_{V|1}(j) = \rho_j P_{V|0}(j)$. Hence $P_{X|0}(x)/P_{V|0}(j) = P_{X|1}(x)/P_{V|1}(j)$ for $x \in \mathcal{X}_j$, and averaging over y with weights $P_Y(y)$ shows that this common value equals $P_X(x)/P_V(j)$, i.e.,

$$P_{X|y}(x) = P_{V|y}(j)\,\frac{P_X(x)}{P_V(j)}, \quad x \in \mathcal{X}_j,\ y \in \{0,1\}. \tag{56}$$

Define the encoder $\bar{K}$ of V with output alphabet $\mathcal{U}_M$ by

$$\bar{K}(u|j) = \sum_{x \in \mathcal{X}_j} \frac{P_X(x)}{P_V(j)} K(u|x),$$

which is well defined since $\sum_u \bar{K}(u|j) = \sum_{x\in\mathcal{X}_j} P_X(x)/P_V(j) = 1$. By (56),

$$P^{\bar{K}}_{U|y}(u) = \sum_{j=1}^{k} \sum_{x\in\mathcal{X}_j} P_{V|y}(j)\frac{P_X(x)}{P_V(j)} K(u|x) = \sum_{x\in\mathcal{X}} P_{X|y}(x) K(u|x) = P^{K}_{U|y}(u).$$

Moreover, the joint distribution of (V, U) under K is $\sum_{x\in\mathcal{X}_j} P_X(x)K(u|x) = P_V(j)\bar{K}(u|j)$, which is the one induced by $\bar{K}$ with input P_V. Since V is a function of X,

$$I_{\bar{K}}(U;V) = I_K(U;V) \le I_K(U;X,V) = I_K(U;X).$$

Proof of (ii). Grouping the symbols of $\mathcal{X}$ by class,

$$P^{K}_{U|y}(u) = \sum_{j=1}^{k} \sum_{x\in\mathcal{X}_j} P_{X|y}(x)\bar{K}(u|j) = \sum_{j=1}^{k} P_{V|y}(j)\bar{K}(u|j) = P^{\bar{K}}_{U|y}(u).$$

Under K, the output U depends on X only through $V = \nu(X)$, so $I_K(U;X|V) = 0$, and the joint distribution of (V, U) is $P_V(j)\bar{K}(u|j)$. Hence

$$I_K(U;X) = I_K(U;X,V) = I_K(U;V) + I_K(U;X|V) = I_{\bar{K}}(U;V).$$

B.7 Proof of Lemma 5.1

The Markovity gives

$$P^{K}_{U|y}(u) = \sum_{x\in\mathcal{X}} K(u|x) P_{X|y}(x), \quad y \in \{0,1\}.$$

Then,

$$\frac{P^{K}_{U|1}(u)}{P^{K}_{U|0}(u)} = \frac{\sum_{x\in\mathcal{X}} K(u|x)P_{X|1}(x)}{\sum_{x\in\mathcal{X}} K(u|x)P_{X|0}(x)}.$$

Since $K_{U|X} \in \mathcal{K}^{(M)}_{R,+}$, and by (1), applying the weighted average inequality (Fact A.3) derives

$$\ell = \min_{x\in\mathcal{X}} \frac{P_{X|1}(x)}{P_{X|0}(x)} \le \frac{P^{K}_{U|1}(u)}{P^{K}_{U|0}(u)} \le \max_{x\in\mathcal{X}} \frac{P_{X|1}(x)}{P_{X|0}(x)} = L.$$

B.8 Proof of Lemma 5.7

Applying Young's inequality (Fact A.1) gives

$$\begin{aligned} C_{u,x}^s &= sw(u)P_{X|0}(x) + (1-s)w(u)^{-\frac{s}{1-s}}P_{X|1}(x) \\ &\geq \left(w(u)P_{X|0}(x)\right)^s \left(w(u)^{-\frac{s}{1-s}}P_{X|1}(x)\right)^{1-s} \\ &= P_{X|0}(x)^s P_{X|1}(x)^{1-s}, \end{aligned}$$

with the lower bound strictly positive and independent of the encoder.

By Lemma 5.1, and $\ell \leq 1 \leq L$ (since $\sum_x \rho(x)P_{X|0}(x) = \sum_x P_{X|1}(x) = 1$), we have

$$w(u) \leq L^{1-s} \leq L \leq \Gamma, \quad w(u)^{-\frac{s}{1-s}} \leq \ell^{-s} \leq \ell^{-1} \leq \Gamma.$$

Substituting into (31) derives

$$C_{u,x}^s \leq \Gamma\big(sP_{X|0}(x) + (1-s)P_{X|1}(x)\big).$$

B.9 Proof of Theorem 5.8

Write $q_t = P_U^{(t)}$, $K_t = K_{U|X}^{(t)}$, and $Z_t(x) = \sum_{u'} q_t(u') \exp\big(-C_{u',x}^s/(\lambda P_X(x))\big)$. Since q_0 has full support and every exponential factor is positive, induction on t gives $q_t(u) > 0$ for all $t \geq 0$ and $K_t(u|x) > 0$ for all $t \geq 1$ and all (u,x). Hence every relative entropy below is finite. Note also that $q_t = P_U^{K_t}$ for every $t \geq 0$.

One-step inequality. Let $K \in \mathcal{K}_\infty^{(M)}$ and $q = P_U^K$. The update in Algorithm 1 gives $C_{u,x}^s = -\lambda P_X(x) \log\left(Z_t(x)K_{t+1}(u|x)/q_t(u)\right)$. Substituting it into (40),

$$\hat{\mathcal{L}}_\lambda(K, q_t) = \lambda \sum_x P_X(x)\Big(\mathrm{D}(K(\cdot|x) \,\|\, K_{t+1}(\cdot|x)) - \log Z_t(x)\Big).$$

Taking $K = K_{t+1}$ and subtracting, then applying Fact A.5 with $q = \sum_x P_X(x)K(\cdot|x)$ and $q_{t+1} = \sum_x P_X(x)K_{t+1}(\cdot|x)$,

$$\hat{\mathcal{L}}_\lambda(K, q_t) - \hat{\mathcal{L}}_\lambda(K_{t+1}, q_t) = \lambda \sum_x P_X(x)\mathrm{D}(K(\cdot|x) \,\|\, K_{t+1}(\cdot|x)) \geq \lambda \mathrm{D}(q\|q_{t+1}).$$

By (41), $\mathcal{L}_\lambda(K_{t+1}) \leq \hat{\mathcal{L}}_\lambda(K_{t+1}, q_t)$ and $\hat{\mathcal{L}}_\lambda(K, q_t) = \mathcal{L}_\lambda(K) + \lambda \mathrm{D}(q\|q_t)$. Hence

$$\mathcal{L}_\lambda(K_{t+1}) \leq \mathcal{L}_\lambda(K) + \lambda(\mathrm{D}(q\|q_t) - \mathrm{D}(q\|q_{t+1})), \quad \forall K \in \mathcal{K}_\infty^{(M)}. \tag{57}$$

Monotonicity and rate. Taking $K = K_t$, for which $q = q_t$, (57) gives $\mathcal{L}_\lambda(K_{t+1}) \leq \mathcal{L}_\lambda(K_t)$. Taking $K = K^*$ with $q^* = P_U^{K^*}$, summing (57) over $t = 0, \cdots, T-1$, and using the monotonicity,

$$T(\mathcal{L}_\lambda(K_T) - \mathcal{L}_\lambda^*) \leq \sum_{t=1}^T (\mathcal{L}_\lambda(K_t) - \mathcal{L}_\lambda^*) \leq \lambda(\mathrm{D}(q^*\|q_0) - \mathrm{D}(q^*\|q_T)) \leq \lambda \mathrm{D}(q^*\|q_0),$$

which proves (42).

Convergence of the iterates. Since $\mathcal{K}_\infty^{(M)}$ is compact, there is a subsequence $K_{t_j} \to \bar{K}$, and then $q_{t_j} \to \bar{q} = P_U^{\bar{K}}$. By (42) and the continuity of $\mathcal{L}_\lambda$, $\bar{K}$ is a global minimizer of $\mathcal{L}_\lambda$. Applying (57) with $K = \bar{K}$ and using $\mathcal{L}_\lambda(K_{t+1}) \geq \mathcal{L}_\lambda(\bar{K})$ shows that $\mathrm{D}(\bar{q}\|q_t)$ is nonincreasing in t. Moreover, $\mathrm{D}(\bar{q}\|q_{t_j}) = \sum_{u:\bar{q}(u)>0} \bar{q}(u) \log \frac{\bar{q}(u)}{q_{t_j}(u)} \to 0$, since $q_{t_j}(u) \to \bar{q}(u) > 0$ in every nonzero term. Hence the whole sequence $\mathrm{D}(\bar{q}\|q_t) \to 0$, and $q_t \to \bar{q}$ by Fact A.6. Since $Z_t(x) \geq \min_u \exp\big(-C_{u,x}^s/(\lambda P_X(x))\big) > 0$, K_{t+1} is a continuous function of q_t through the update in Algorithm 1. Therefore K_t converges, and its limit is $\bar{K}$. Finally, by (41), $\hat{\mathcal{L}}_\lambda(\bar{K}, \bar{q}) = \mathcal{L}_\lambda(\bar{K}) = \mathcal{L}_\lambda^*$, so $(K_t, q_t) \to (\bar{K}, P_U^{\bar{K}})$, a global minimizer of $\hat{\mathcal{L}}_\lambda$.

B.10 Proof of Lemma 5.10

Proof of (i). Let $0 \le \lambda_1 < \lambda_2$, and write $K_i = K(\lambda_i)$, $G_i = G(K_i)$ and $I_i = I(\lambda_i)$ for $i = 1, 2$. Since K_1 minimizes $\mathcal{L}_{\lambda_1}$ and K_2 minimizes $\mathcal{L}_{\lambda_2}$ over $\mathcal{K}_\infty^{(M)}$,

$$G_1 + \lambda_1 I_1 \le G_2 + \lambda_1 I_2, \qquad G_2 + \lambda_2 I_2 \le G_1 + \lambda_2 I_1.$$

Adding the two inequalities gives $(\lambda_2 - \lambda_1)(I_1 - I_2) \ge 0$, hence $I_1 \ge I_2$.

Proof of (ii). By Lemma 4.1, $\mathcal{K}_R^{(M)}$ is non-empty and compact, and G is continuous, so G^* is attained by some $K^\circ \in \mathcal{K}_R^{(M)}$. Since $K(\lambda)$ minimizes $\mathcal{L}_\lambda$ over $\mathcal{K}_\infty^{(M)} \supseteq \mathcal{K}_R^{(M)}$,

$$G(K(\lambda)) + \lambda I(\lambda) \le G(K^\circ) + \lambda I_{K^\circ}(U; X) \le G^* + \lambda R,$$

which gives the upper bound in (44). If $I(\lambda) \le R$, then $K(\lambda) \in \mathcal{K}_R^{(M)}$, so that $G(K(\lambda)) \ge G^*$, which gives the lower bound. If $I(\lambda) = R$ or $\lambda = 0$, the right-hand side of (44) vanishes, and $G(K(\lambda)) = G^*$.

Proof of (iii). Let $\beta = \sum_x P_{X|0}(x)^s P_{X|1}(x)^{1-s}$, which is strictly positive by (1). By the lower bound in Lemma 5.7, for any $K \in \mathcal{K}_\infty^{(M)}$,

$$G(K) \ge \sum_x \Big(\sum_u K(u|x) \Big) P_{X|0}(x)^s P_{X|1}(x)^{1-s} = \beta.$$

Let K_0 be the constant encoder in Appendix B.5, which satisfies $I_{K_0}(U; X) = 0$. By the upper bound in Lemma 5.7,

$$G(K_0) = \sum_x C_{1,x}^s \le \Gamma \sum_x \big(s P_{X|0}(x) + (1-s) P_{X|1}(x) \big) = \Gamma.$$

Hence, for any $\lambda > 0$, the minimality of $K(\lambda)$ gives

$$\beta + \lambda I(\lambda) \le \mathcal{L}_\lambda(K(\lambda)) \le \mathcal{L}_\lambda(K_0) = G(K_0) \le \Gamma,$$

that is, $I(\lambda) \le (\Gamma - \beta)/\lambda$, which is at most R whenever $\lambda \ge (\Gamma - \beta)/R$. Now define

$$\lambda^* \triangleq \inf \{ \lambda \ge 0 : I(\lambda) \le R \text{ for some choice of } K(\lambda) \},$$

so that $\lambda^* \le (\Gamma - \beta)/R < \Gamma/R$. If $\lambda > \lambda^*$, there exist $\lambda_1 < \lambda$ and a choice of $K(\lambda_1)$ with $I(\lambda_1) \le R$, and (i) gives $I(\lambda) \le I(\lambda_1) \le R$ for every choice of $K(\lambda)$. If $\lambda < \lambda^*$, then by the definition of λ^*, $I(\lambda) > R$ for every choice of $K(\lambda)$.

Consequences for the bisection. The bisection sets $\lambda_{\text{small}} \leftarrow \lambda$ only when $I(\lambda) > R$, which by (iii) forces $\lambda \le \lambda^*$, and sets $\lambda_{\text{big}} \leftarrow \lambda$ only when $I(\lambda) \le R$, which forces $\lambda \ge \lambda^*$. Initially, $\lambda_{\text{small}} = 0 \le \lambda^* < \Gamma/R = \lambda_{\text{big}}$, and $I(\Gamma/R) \le R$ by (iii). Hence $\lambda^* \in [\lambda_{\text{small}}, \lambda_{\text{big}}]$ and $I(\lambda_{\text{big}}) \le R$ at every step, and since each step halves the bracket, $\lambda_{\text{big}} - \lambda_{\text{small}} = \Gamma/(R\, 2^n)$ after n steps.