Mutual Information Constrained Chernoff Bottleneck

Dier Tang
Department of Mathematics
The University of Hong Kong
Hong Kong, China
tangde_math@connect.hku.hk

Guangyue Han
Department of Mathematics
The University of Hong Kong
Hong Kong, China
ghan@hku.hk

***Abstract*—The classical information bottleneck (IB) measures the relevance of a representation to the target by their mutual information, which does not directly characterize the error of downstream decisions. For a binary hypothesis inferred from many separately encoded observations, the optimal error exponent is the Chernoff information between the two conditional distributions of the representation. We study the Chernoff bottleneck with a mutual information constraint, which maximizes this Chernoff information over encoders with bounded mutual information between the representation and the observation. We show that the optimal exponent increases strictly with the rate up to the entropy of the statistic that merges the observation symbols with equal likelihood ratio, remains at the uncompressed exponent beyond, and, unlike the IB curve, need not be concave. We prove a cardinality bound of $k+1$ on the representation, where k is the number of distinct likelihood ratios. We propose an alternating algorithm that updates the encoder via a generalized Blahut–Arimoto algorithm and the Chernoff parameter via a nonlinear equation, and prove that its iterates remain feasible, with nondecreasing and convergent Chernoff information. Numerical experiments confirm the theory, and in topic detection on 20 Newsgroups, compressing each word to 17% of its entropy retains 90% of the error exponent and nearly the uncompressed accuracy. The full version is available at [1].**

***Index Terms*—Information bottleneck, Chernoff information, hypothesis testing, cardinality bound, alternating optimization**

I. Introduction

Representation learning [2] aims to learn compact representations of raw data that make downstream tasks, such as inference and prediction, more effective [3]–[8]. The information bottleneck (IB) [9] seeks a representation U that preserves the information relevant to a target Y while discarding redundancy of the observation X, trading off maximizing $I(U;Y)$ against minimizing $I(U;X)$. It has been used both to analyze deep neural networks and as an objective for learning robust, generalizable representations [10]–[16].

However, $I(U;Y)$ does not directly characterize the error of downstream decisions. When $Y \in \{0,1\}$ is inferred from the outputs $U_1,\dots,U_n$ of the same encoder $P_{U|X}$ applied to observations $X_1,\dots,X_n$ i.i.d. given Y, the minimum Bayes error decays exponentially in n, with optimal exponent equal to the Chernoff information $C(P_{U|0},P_{U|1})$ for every prior with full support [17]–[19]. Unlike $I(U;Y)$, it does not depend on the prior, and an IB-optimal encoder need not be optimal for testing. This motivates the *mutual information (MI) constrained Chernoff bottleneck*

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

where samples are encoded separately, as by different sensors or a per-sample feature extractor [20]–[22], not jointly as in testing under communication constraints [23], [24].

The theory and algorithms of classical IB do not carry over directly, since they rely on the IB loss $I(X;Y)-I(U;Y)$ being an expectation of a per-letter divergence, which yields Blahut–Arimoto-type iterations [9], [25], [26], and on $I(U;Y)$ being a probability-weighted sum over the outputs, which underlies time-sharing and cardinality arguments [27]. The Chernoff information has neither property: its optimization over a parameter s couples all outputs. Our contributions are as follows.

- *Shape of the curve.* $C(R)$ is strictly increasing up to $R=H(V)$, where V merges the symbols of X with equal likelihood ratio, and equals the uncompressed exponent beyond (Theorem 1), so every optimal encoder at $R\le H(V)$ uses the full rate (Corollary 1); unlike the IB curve, $C(R)$ need not be concave (Example 1).
- *Cardinality bound.* An optimal encoder exists with at most $k+1$ outputs, where k is the number of distinct likelihood ratios (Theorem 2).
- *Algorithm.* An alternating algorithm updates the encoder by a generalized Blahut–Arimoto algorithm, which converges globally at rate $O(1/t)$, and the parameter s by solving a nonlinear equation (Algorithm 1). Its iterates remain feasible, with nondecreasing and convergent Chernoff information (Theorems 3 and 4).
- *Experiments.* Computed curves confirm the theory (Fig. 1), and in topic detection on 20 Newsgroups [28], compressing each word to 17% of its entropy keeps 90% of the error exponent and nearly the uncompressed accuracy (Fig. 2).

Section II formulates the problem, Sections III and IV study the shape of $C(R)$ and the cardinality bound, and Sections V and VI present the algorithm and the experiments. Omitted proofs and further details are in the full version [1].

II. Problem Setup

Let X and Y be random variables on finite alphabets $\mathcal{X}$ and $\mathcal{Y} = \{0, 1\}$, respectively, with $P_X(x) > 0, P_Y(y) > 0$ for any x, y, and write $P_{X|y} = P_{X|Y=y}$. Assume $P_{X|0} \not\equiv P_{X|1}$ and

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

For distributions P_0, P_1 on a finite set $\mathcal{U}$, the Chernoff information [17], [18] is

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

and denote by $C_X = C(P_{X|0}, P_{X|1}) > 0$. All logarithms are natural, and $0^0 = \lim_{s \to 0} 0^s = 0$.

Given output alphabet $\mathcal{U}_M = \{1, \dots, M\}$, an encoder $K = K_{U|X}$ induces $P^K_{U|y}(u) = \sum_x K(u|x) P_{X|y}(x)$, the marginal P^K_U, and the mutual information $I_K(U; X)$. Define the *feasible encoder set* $\mathcal{K}^{(M)}_R$ as the set of such encoders with $I_K(U; X) \le R$, its strictly positive subset $\mathcal{K}^{(M)}_{R,+}$ with $K(u|x) > 0$ for all (u, x), and $\mathscr{K}_R = \bigcup_{M \ge 1} \mathcal{K}^{(M)}_R$. Define

$$B_M(K, s) \triangleq \sum_{u \in \mathcal{U}_M} P^K_{U|0}(u)^s P^K_{U|1}(u)^{1-s}, \tag{2}$$

for any encoder K with output alphabet $\mathcal{U}_M$ and $s \in [0, 1]$. The Chernoff information of K and the MI constrained Chernoff bottleneck are

$$\mathcal{C}(K) = -\log \min_{0 \le s \le 1} B_M(K, s), \quad C(R) = \sup_{K \in \mathscr{K}_R} \mathcal{C}(K). \tag{3}$$

Clearly, $C(0) = 0$, $C(R)$ is nondecreasing, and $C(R) \le C_X$ by the data processing inequality of Chernoff information [29].

Let $\rho(x) = P_{X|1}(x)/P_{X|0}(x)$ take the distinct values $\rho_1, \dots, \rho_k$, so that $2 \le k \le |\mathcal{X}|$, and let $\nu(x) = j$ if and only if $\rho(x) = \rho_j$. Define $V = \nu(X)$, which merges the symbols with equal likelihood ratio, and is the minimal sufficient statistic of X for Y [18]. Moreover, $H(V) \le H(X)$, with equality if and only if $k = |\mathcal{X}|$. Let $K^V(u|x) = \mathbf{1}\{u = \nu(x)\}$ be the encoder that outputs V, K^X be the identity encoder that outputs X with no compression, and define $\Gamma \triangleq \max_x \max\{\rho(x), \rho(x)^{-1}\} > 1$.

III. Global Shape of the $C(R)$ Curve

The shape of $C(R)$ is governed by V, as Theorem 1 shows, whose proof relies on the following lemmas. Detailed proofs of all lemmas are included in [1].

Lemma 1 (Sufficiency of V): $I_{K^V}(U; X) = H(V)$ and $B_k(K^V, s) = B_{|\mathcal{X}|}(K^X, s)$ for all $s \in [0, 1]$. In particular, $\mathcal{C}(K^V) = C_X$.

Lemma 2 (Contraction): For any encoder K with output alphabet $\mathcal{U}_M$ and $s \in [0, 1]$,

$$B_M(K, s) \ge B_{|\mathcal{X}|}(K^X, s). \tag{4}$$

Moreover, if $I_K(U; X) < H(V)$, then the inequality (4) is strict for every $s \in (0, 1)$.

Lemma 3 (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 $s \in [0, 1]$,

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

Theorem 1 (Saturation and strict monotonicity): $C(R)$ is strictly increasing on $[0, H(V)]$, and $C(R) = C_X$ for every $R \ge H(V)$.

Proof: Let $R \ge H(V)$. By Lemma 1, $I_{K^V}(U; X) = H(V) \le R$, so that $K^V \in \mathscr{K}_R$, and $\mathcal{C}(K^V) = C_X$. Hence $C(R) \ge C_X$, while the data processing inequality gives $C(R) \le C_X$. Therefore $C(R) = C_X$.

Let $0 \le R_1 < R_2 \le H(V)$, 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 2). Write $I^* = I_{K^*}(U; X) \le R_1 < H(V)$. By Young's inequality,

$$B_{M^*}(K^*, s) \le \sum_{u \in \mathcal{U}_{M^*}} \big(s P^{K^*}_{U|0}(u) + (1 - s) P^{K^*}_{U|1}(u) \big) = 1$$

for every $s \in [0, 1]$, while $B_{M^*}(K^*, 0) = B_{M^*}(K^*, 1) = 1$. If $B_{M^*}(K^*, s) \equiv 1$ on $[0, 1]$, we directly 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

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

Moreover, since $I^* < H(V)$, Lemma 2 gives $B_{M^*}(K^*, \bar{s}) > B_{|\mathcal{X}|}(K^X, \bar{s})$. Let K' be the encoder with output alphabet $\mathcal{U}_{k+M^*}$ given by Lemma 3, with $K_1 = K^V$, $K_2 = K^*$ and

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

By Lemma 1 and Lemma 3,

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

so that $K' \in \mathscr{K}_{R_2}$. Then,

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

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

Corollary 1 (Active rate constraint): For $R \le H(V)$, every optimal encoder satisfies $I_K(U; X) = R$.

Proof: Suppose K is optimal at rate R with $I_K(U; X) < R$. Since $K \in \mathscr{K}_{I_K(U;X)}$, we have $C(I_K(U; X)) \ge C(R)$, contradicting the strict monotonicity of C on $[0, H(V)]$. ∎

By time-sharing, the IB curve $R \mapsto \max\{I(U; Y) : I(U; X) \le R\}$ is concave, whereas $C(R)$ need not be:

Example 1 (Non-concave $C(R)$): Let P_Y be uniform, X be binary with $P_{X|0} = (1 - a, a)$, $P_{X|1} = (a, 1 - a)$, $0 < a \le 10^{-3}$. Then $C(R)$ is not concave on $[0, \log 2]$. Proof in [1].

IV. Cardinality Bound and Attainability

The supremum in (3) is over encoders with arbitrarily many outputs. Theorem 2 proves the cardinality bound, and thereby guarantees that an optimal encoder exists. Detailed proofs of Lemmas 4 and 5 are included in [1].

Lemma 4 (Compactness and convexity): For any finite M and $R \geq 0$, the feasible encoder set $\mathcal{K}_R^{(M)}$ is non-empty, compact and convex.

Lemma 5 (Reduction to V): For an encoder $\bar{K} = \bar{K}_{U|V}$ of V, let $P_{U|y}^{\bar{K}}(u) = \sum_j P_{V|y}(j)\bar{K}(u|j)$ and $I_{\bar{K}}(U;V)$ be its mutual information under P_V. Then:

1) For every encoder K of X with output alphabet $\mathcal{U}_M$, some encoder $\bar{K}$ of V with output alphabet $\mathcal{U}_M$ satisfies $P_{U|y}^{\bar{K}} = P_{U|y}^{K}$ for $y \in \{0,1\}$ and $I_{\bar{K}}(U;V) \leq I_K(U;X)$.
2) For every encoder $\bar{K}$ of V, the encoder $K(u|x) = \bar{K}(u|\nu(x))$ of X satisfies $P_{U|y}^{K} = P_{U|y}^{\bar{K}}$ and $I_K(U;X) = I_{\bar{K}}(U;V)$.

Theorem 2 (Cardinality bound): For every $R \geq 0$,

$$C(R) = \max_{K \in \mathcal{K}_R^{(k+1)}} \mathcal{C}(K),$$

and the maximum is attained by an encoder of the form $K(u|x) = \bar{K}(u|\nu(x))$, where $\bar{K}$ is an encoder of V with output alphabet $\mathcal{U}_{k+1}$.

Proof: For encoders $\bar{K}$ of V, define $B_M(\bar{K}, s)$ and $\mathcal{C}(\bar{K})$ as in (2) and (3) with $P_{U|y}^{\bar{K}}$, let $\bar{\mathcal{K}}_R^{(M)}$ be the set of such encoders with output alphabet $\mathcal{U}_M$ and $I_{\bar{K}}(U;V) \leq R$, and let $\bar{\mathscr{K}}_R = \bigcup_{M \geq 1} \bar{\mathcal{K}}_R^{(M)}$.

By Lemma 5, every $K \in \mathscr{K}_R$ has some $\bar{K} \in \bar{\mathscr{K}}_R$ with $\mathcal{C}(\bar{K}) = \mathcal{C}(K)$, and every $\bar{K} \in \bar{\mathscr{K}}_R$ lifts to $K(u|x) = \bar{K}(u|\nu(x))$ in $\mathscr{K}_R$ with the same output alphabet. Hence

$$C(R) = \sup_{\bar{K} \in \bar{\mathscr{K}}_R} \mathcal{C}(\bar{K}). \tag{5}$$

Let $\bar{K} \in \bar{\mathcal{K}}_R^{(m)}$ with $m \geq k+2$, where, without loss of generality, $\lambda_u \triangleq P_U^{\bar{K}}(u) > 0$ for all u, and let $s_{\bar{K}}$ minimize $B_m(\bar{K}, \cdot)$. Write $w_u \triangleq P_{V|U=u}^{\bar{K}}$ and

$$\phi_s(w) \triangleq \Big(\sum_j \frac{w(j)P_{V|0}(j)}{P_V(j)}\Big)^s \Big(\sum_j \frac{w(j)P_{V|1}(j)}{P_V(j)}\Big)^{1-s}.$$

Then $P_{U|y}^{\bar{K}}(u) = \lambda_u \sum_j w_u(j)P_{V|y}(j)/P_V(j)$, so that

$$\textstyle\sum_{u=1}^m \lambda_u w_u = P_V, \quad I_{\bar{K}}(U;V) = \sum_{u=1}^m \lambda_u \mathrm{D}(w_u\|P_V),$$
$$B_m(\bar{K}, s) = \textstyle\sum_{u=1}^m \lambda_u \phi_s(w_u).$$

The $m \geq k+2$ vectors $v_u = \big(1, w_u(1), \ldots, w_u(k-1), \mathrm{D}(w_u\|P_V)\big) \in \mathbb{R}^{k+1}$ are linearly dependent, so there exists non-zero $c \in \mathbb{R}^m$ such that $\sum_{u=1}^m c_u v_u = 0$. Thus,

$$\sum_{u=1}^m c_u = 0, \quad \sum_{u=1}^m c_u \mathrm{D}(w_u\|P_V) = 0,$$

and $\sum_{u=1}^m c_u w_u(j) = 0$ for all j, including $j = k$ since $w_u(k) = 1 - \sum_{j<k} w_u(j)$. Hence $\lambda_u(t) \triangleq \lambda_u + tc_u$ satisfy $\sum_{u=1}^m \lambda_u(t) w_u = P_V$ and $\sum_{u=1}^m \lambda_u(t)\mathrm{D}(w_u\|P_V) = I_{\bar{K}}(U;V)$ for all t, while $\sum_{u=1}^m \lambda_u(t)\phi_{s_{\bar{K}}}(w_u)$ is affine in t. Since $c \neq 0$ sums to zero, it has entries of both signs. Let

$$t_- \triangleq \max_{u:c_u>0} \big(-\frac{\lambda_u}{c_u}\big) < 0, \qquad t_+ \triangleq \min_{u:c_u<0} \big(-\frac{\lambda_u}{c_u}\big) > 0.$$

Then, $\lambda_u(t) \geq 0$ for $t \in [t_-, t_+]$, and at $t \in \{t_-, t_+\}$, some $\lambda_u(t)$ vanishes. Choose $t_0 \in \{t_-, t_+\}$ such that

$$\sum_{u=1}^m \lambda_u(t_0)\phi_{s_{\bar{K}}}(w_u) \leq \sum_{u=1}^m \lambda_u(0)\phi_{s_{\bar{K}}}(w_u) = B_m(\bar{K}, s_{\bar{K}}),$$

and remove the outputs with $\lambda_u(t_0) = 0$. The encoder $\tilde{K}(u|j) = \lambda_u(t_0)w_u(j)/P_V(j)$ then has output alphabet $\mathcal{U}_{\tilde{m}}$ with $\tilde{m} \leq m-1$, $I_{\tilde{K}}(U;V) = I_{\bar{K}}(U;V) \leq R$, and

$$\mathcal{C}(\tilde{K}) \geq -\log B_{\tilde{m}}(\tilde{K}, s_{\bar{K}}) \geq -\log B_m(\bar{K}, s_{\bar{K}}) = \mathcal{C}(\bar{K}).$$

Repeating this, every $\bar{K} \in \bar{\mathscr{K}}_R$ is dominated by an encoder in $\bar{\mathcal{K}}_R^{(k+1)}$, padded with unused outputs if needed. Hence the supremum in (5) may be taken over $\bar{\mathcal{K}}_R^{(k+1)}$, and it is attained by some $\bar{K}^*$, since this set is compact by Lemma 4 applied to V and $\mathcal{C}$ is continuous.

By Lemma 5 and (5), $K^*(u|x) = \bar{K}^*(u|\nu(x))$ lies in $\mathcal{K}_R^{(k+1)}$ and $\mathcal{C}(K^*) = \mathcal{C}(\bar{K}^*) = C(R)$. Since $\mathcal{K}_R^{(k+1)} \subseteq \mathscr{K}_R$, $C(R) = \mathcal{C}(K^*) = \max_{K \in \mathcal{K}_R^{(k+1)}} \mathcal{C}(K)$. ∎

Remark 1: By Lemma 5, $C(R)$ depends only on $P_{V,Y}$, so compressing X is never better than compressing V.

Remark 2: Theorem 2 implies that to achieve $C(R)$, it suffices to set the cardinality $M = k+1$ and seek

$$K^* \in \arg\max_{K \in \mathcal{K}_R^{(M)}} \mathcal{C}(K). \tag{6}$$

V. An Alternating Optimization Algorithm

For $R > 0$ (since $C(0) = 0$) and any M ($M \geq k+1$ suffices), we seek (6) by minimizing $B_M(K, s)$ alternately over the encoder $K \in \mathcal{K}_R^{(M)}$ and the parameter $s \in [0,1]$.

A. Encoder Update

For $s \in (0,1)$ and any $w(u) > 0$, Young's inequality gives

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

with equality if $w(u) = \big(P_{U|1}^K(u)/P_{U|0}^K(u)\big)^{1-s}$. Fixing $w(u)$ by this equality condition at the current encoder $\tilde{K} \in \mathcal{K}_{R,+}^{(M)}$, we obtain $B_M(K,s) \leq G(K) \triangleq \sum_{u,x} C_{u,x}^s K(u|x)$ for all K, with equality at $\tilde{K}$, where

$$C_{u,x}^s = sw(u)P_{X|0}(x) + (1-s)w(u)^{-\frac{s}{1-s}} P_{X|1}(x). \tag{7}$$

Since $B_M(\cdot, s)$ is concave and positively homogeneous in K, G is its tangent plane at $\tilde{K}$. Thus, the encoder is updated by

$$\hat{K} \in \arg\min_{K \in \mathcal{K}_R^{(M)}} G(K), \tag{8}$$

which is a rate–distortion problem with distortion $C_{u,x}^s/P_X(x)$ [30], [31].

Since G is linear and $\mathcal{K}_R^{(M)}$ is convex (Lemma 4), (8) is a convex problem. We solve it through the Lagrangian $\mathcal{L}_\lambda(K) \triangleq G(K) + \lambda I_K(U;X)$, $\lambda > 0$, minimized over all encoders with output alphabet $\mathcal{U}_M$. Its stationarity condition $K(u|x) \propto P_U^K(u)\exp\big(-C_{u,x}^s/(\lambda P_X(x))\big)$ is a fixed-point equation, since P_U^K depends on K. As in the Blahut–Arimoto algorithm [25], [26], alternately updating K and P_U gives the generalized Blahut–Arimoto (GBA) iteration

$$\begin{aligned} K^{(t+1)}(u|x) &\propto P_U^{(t)}(u)\exp\Big(-\frac{C_{u,x}^s}{\lambda P_X(x)}\Big), \\ P_U^{(t+1)}(u) &= \textstyle\sum_x K^{(t+1)}(u|x)P_X(x). \end{aligned} \tag{9}$$

Since $C_{u,x}^s/P_X(x)$ depends on x only through $\rho(x)$, every iterate of (9) has the form $K(u|x) = \tilde{K}(u|\nu(x))$, so the algorithm effectively operates on V, consistent with Theorem 2.

Theorem 3 (Global convergence of GBA): If $P_U^{(0)}(u) > 0$ for all u, then $K^{(t)}$ converges to a global minimizer of $\mathcal{L}_\lambda$, and, with the minimum $\mathcal{L}_\lambda^*$ attained at K^*, for every $t \geq 1$,

$$0 \leq \mathcal{L}_\lambda(K^{(t)}) - \mathcal{L}_\lambda^* \leq \frac{\lambda}{t}\mathrm{D}(P_U^{K^*}\|P_U^{(0)}),$$

$$\mathcal{L}_\lambda(K^{(t)}) - \mathcal{L}_\lambda^* \leq \lambda \log \max_u \frac{P_U^{(t)}(u)}{P_U^{(t-1)}(u)}.$$

Proof sketch: Iteration (9) alternately minimizes $\mathcal{L}_\lambda(K) + \lambda\mathrm{D}(P_U^K\|P_U)$ over K and P_U. Its exponential form and joint convexity of D give $\mathcal{L}_\lambda(K^{(t+1)}) \leq \mathcal{L}_\lambda(K) + \lambda\big(\mathrm{D}(P_U^K\|P_U^{(t)}) - \mathrm{D}(P_U^K\|P_U^{(t+1)})\big)$ for all K, so $K = K^{(t)}$ gives monotonicity, and $K = K^*$ the bounds by telescoping and one step. ∎

The second bound is a computable stopping criterion. The trade-off λ is found by bisection: let $K(\lambda)$ be any minimizer of $\mathcal{L}_\lambda$, $I(\lambda) \triangleq I_{K(\lambda)}(U;X)$, and G^* the optimal value of (8).

Lemma 6 (Bisection): $I(\lambda)$ is nonincreasing. If $I(\lambda) \leq R$, then $K(\lambda)$ is feasible and $0 \leq G(K(\lambda)) - G^* \leq \lambda(R - I(\lambda))$. Moreover, there exists $\lambda^* < \Gamma/R$ with $I(\lambda) > R$ for $\lambda < \lambda^*$ and $I(\lambda) \leq R$ for $\lambda > \lambda^*$.

Proof sketch: Adding the optimality conditions at λ_1, λ_2 gives monotonicity, and comparing $K(\lambda)$ with a minimizer of (8) gives the suboptimality bound. By Young's inequality, $G \geq \beta \triangleq \sum_x P_{X|0}(x)^s P_{X|1}(x)^{1-s} > 0$, while $w(u), w(u)^{-\frac{s}{1-s}} \leq \Gamma$ give $G \leq \Gamma$ for a constant encoder. Comparing with it, as it has zero rate, gives $\lambda I(\lambda) \leq \Gamma - \beta$, hence $\lambda^* < \Gamma/R$. ∎

Hence the bisection on $[0, \Gamma/R]$ keeps λ^* in its bracket, and every accepted encoder is feasible with certified suboptimality.

B. Parameter Update

For fixed $K \in \mathcal{K}_{R,+}^{(M)}$, define $g_K(s) \triangleq B_M(K,s)$. Since $P_{U|y}^K(u) > 0$, a direct calculation gives

$$g_K'(s) = \sum_{u\in\mathcal{U}_M} P_{U|0}^K(u)^s P_{U|1}^K(u)^{1-s} \log \frac{P_{U|0}^K(u)}{P_{U|1}^K(u)},$$

$$g_K''(s) = \sum_{u\in\mathcal{U}_M} P_{U|0}^K(u)^s P_{U|1}^K(u)^{1-s} \Big(\log \frac{P_{U|0}^K(u)}{P_{U|1}^K(u)}\Big)^2.$$

If $P_{U|0}^K \not\equiv P_{U|1}^K$, then $g_K'' > 0$ on $[0,1]$, which implies $g_K'(s)$ is strictly increasing. Moreover,

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

Hence the minimizer of $g_K(s)$ on $[0,1]$, denoted by s^*, is the unique root in $(0,1)$ of the nonlinear equation

$$g_K'(s) = 0, \tag{10}$$

which can be found by bisection, Newton's or Brent's method [32]–[34]. Furthermore, s^* stays away from the endpoints.

Lemma 7 (Interiority): For $K \in \mathcal{K}_{R,+}^{(M)}$ with $P_{U|0}^K \not\equiv P_{U|1}^K$,

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

Proof sketch: Let $\psi_K \triangleq \log g_K$, which is convex by the Cauchy–Schwarz inequality, with $\psi_K(0) = \psi_K(1) = 0$ and $|\psi_K'(0)|, |\psi_K'(1)| \leq \log\Gamma$, since the likelihood ratios of U lie in $[\Gamma^{-1}, \Gamma]$. Its tangent lines at $s = 0$ and $s = 1$ give $\mathcal{C}(K) \leq s^* \log\Gamma$ and $\mathcal{C}(K) \leq (1 - s^*)\log\Gamma$. ∎

C. Monotone Convergence

Combining the two updates gives Algorithm 1.

Algorithm 1 Alternating algorithm for the Chernoff bottleneck

Input: $P_{X|0}$, $P_{X|1}$, P_X; M; $R > 0$
Output: encoder K

1: Choose $K^{(0)} \in \mathcal{K}_{R,+}^{(M)}$ with $P_{U|0}^{K^{(0)}} \not\equiv P_{U|1}^{K^{(0)}}$ and $s^{(0)} \in (0,1)$; set $B^{(0)} \leftarrow B_M(K^{(0)}, s^{(0)})$ and $l \leftarrow 0$
2: **repeat**
3: Compute $C_{u,x}^{s^{(l)}}$ from $K^{(l)}$ by (7)
4: $K^{(l+1)} \leftarrow K^{(l)}$; $\underline{\lambda} \leftarrow 0$; $\overline{\lambda} \leftarrow \Gamma/R$
5: **repeat**
6: $\lambda \leftarrow (\underline{\lambda} + \overline{\lambda})/2$; run (9) from $K^{(l)}$ to get $\hat{K}$
7: **if** $I_{\hat{K}}(U;X) > R$ **then** $\underline{\lambda} \leftarrow \lambda$
8: **else** $\overline{\lambda} \leftarrow \lambda$
9: **if** $G(\hat{K}) \leq G(K^{(l+1)})$ **then** $K^{(l+1)} \leftarrow \hat{K}$
10: **until** $\overline{\lambda} - \underline{\lambda}$ is small
11: $s^{(l+1)} \leftarrow$ root of (10) for $K^{(l+1)}$
12: $B^{(l+1)} \leftarrow B_M(K^{(l+1)}, s^{(l+1)})$; $l \leftarrow l + 1$
13: **until** $|B^{(l)} - B^{(l-1)}|$ is small
14: **return** $K^{(l)}$

Theorem 4 (Monotone convergence): In Algorithm 1, if (10) is solved exactly, then, for all $l \geq 0$, $K^{(l)} \in \mathcal{K}_{R,+}^{(M)}$ and

$$e^{-C(R)} \leq B^{(l+1)} \leq B^{(l)} \leq B^{(0)} < 1.$$

Hence, for $l \geq 1$, $\mathcal{C}(K^{(l)}) = -\log B^{(l)}$ is nondecreasing and converges to a limit at most $C(R)$, and $s^{(l)} \in [c/\log\Gamma, 1 - c/\log\Gamma]$ with $c = -\log B^{(0)}$.

Proof sketch: Line 9 and the positivity of (9) give $K^{(l+1)} \in \mathcal{K}_{R,+}^{(M)}$, and $B^{(l+1)} \leq B_M(K^{(l+1)}, s^{(l)}) \leq G(K^{(l+1)}) \leq G(K^{(l)}) = B^{(l)}$ by the optimality of $s^{(l+1)}$, majorization, line 9 and tangency at $K^{(l)}$. Young's inequality gives $B^{(0)} < 1$, feasibility gives $B^{(l)} \geq e^{-C(R)}$, and Lemma 7 bounds $s^{(l)}$. ∎

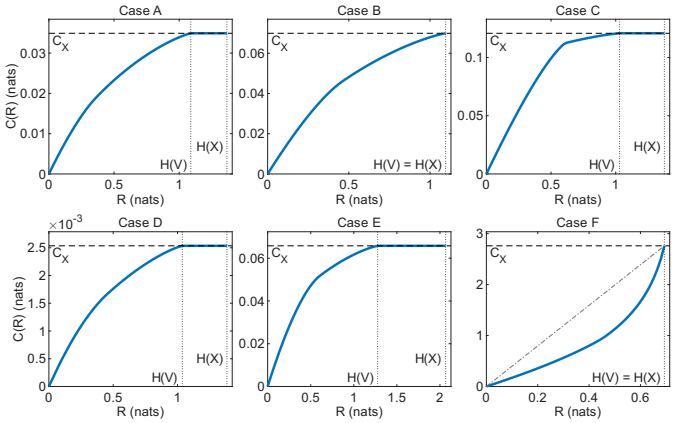


Fig. 1. $C(R)$ for sources A–F. Dashed: C_X. Dotted: $H(V)$ and $H(X)$. Dash-dotted (F): the chord from $(0,0)$ to $(H(V), C_X)$.

TABLE I
SOURCES IN FIG. 1, WITH $P_Y(0) = P_Y(1) = \frac{1}{2}$.

	$P_{X\|0}$	$P_{X\|1}$	k
A	$(0.4, 0.2, 0.2, 0.2)$	$(0.2, 0.2, 0.2, 0.4)$	3
B	$(0.5, 0.3, 0.2)$	$(0.2, 0.3, 0.5)$	3
C	$(0.5, 0.2, 0.2, 0.1)$	$(0.1, 0.3, 0.3, 0.3)$	3
D	$(0.3, 0.25, 0.25, 0.2)$	$(0.25, 0.25, 0.25, 0.25)$	3
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
F	$(1-10^{-3}, 10^{-3})$	$(10^{-3}, 1-10^{-3})$	2

VI. Numerical Experiments

A. The $C(R)$ Curve

We compute $C(R)$ for the six sources in Table I with $M = k+1$ (Theorem 2), where F is the source of Example 1. At each rate, we keep the best of several initializations of Algorithm 1, including the time-sharing of K^V and a constant encoder. Fig. 1 confirms Theorem 1: $C(R)$ increases strictly up to $H(V)$ and equals C_X beyond. For A, C, D and E, where $k < |\mathcal{X}|$, C_X is already reached at $H(V) < H(X)$. The curve of F lies below its chord, consistent with Example 1.

B. Application to Real Data

We detect the topic, `rec.sport.hockey` ($Y = 0$) or `sci.med` ($Y = 1$), of a 20 Newsgroups document [28] from its separately compressed words. With $\mathcal{X}$ the 1000 most frequent training words, add-one smoothed $P_{X|y}$ and uniform P_Y, we have $k = 826$, $H(X) = 5.80$ and $C_X = 0.158$ nats. We use Algorithm 1 with $M = 16$, far below $k+1 = 827$, and the MAP test on $U_1, \ldots, U_n$. At $R = 1$ nat (17% of $H(X)$) and 0.25 nat, the encoder keeps 90% and 47% of C_X. To test whether $\mathcal{C}(K)$ from the training data predicts the test error, we draw words i.i.d. from the test frequencies, where the empirical error exponent matches $\mathcal{C}(K)$ to within about 10% (Fig. 2(a)). On real test documents, whose words are not i.i.d., the error decays more slowly, but at $R = 1$ it nearly matches the uncompressed error (Fig. 2(b)).

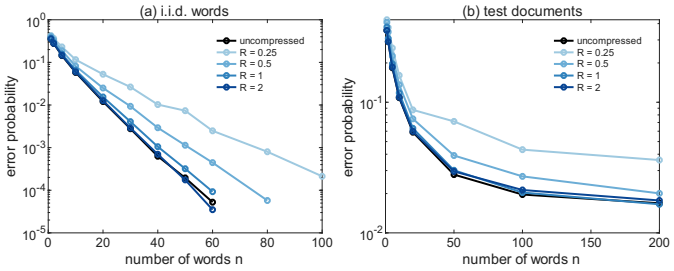


Fig. 2. Test error of topic detection from n words, each compressed to R nats, versus the uncompressed naive Bayes classifier. (a) Words i.i.d. from each topic's test frequencies. (b) Words from each test document.

VII. Conclusion and Future Directions

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. Its curve $C(R)$ increases strictly up to $H(V)$ and equals the uncompressed exponent C_X beyond (Theorem 1), so every optimal encoder at $R \le H(V)$ uses the full rate (Corollary 1). Unlike the IB curve, $C(R)$ need not be concave (Example 1). An optimal encoder exists with at most $k+1$ outputs, where k is the number of distinct likelihood ratios (Theorem 2). Because the Chernoff information is not a per-letter expectation, Algorithm 1 replaces $B_M(\cdot, s)$ by a tangent-plane majorizer, minimized by a generalized Blahut–Arimoto algorithm, and updates s by a nonlinear equation, with feasible iterates and nondecreasing and convergent Chernoff information (Theorems 3 and 4). The computed curves confirm the shape of $C(R)$ (Fig. 1), and, on real data, a small fraction of the entropy keeps most of the error exponent (Fig. 2).

Four directions remain open.

1) *Global optimality.* As $\min_s B_M(\cdot, s)$ is concave in K, our alternating algorithm (Algorithm 1) may stop at a local optimum. Searching over encoders that randomly quantize the likelihood ratio, or tracking the curve down from K^V at $R = H(V)$, may yield global optima.
2) *Comparison with the IB.* As the IB targets a single sample and the Chernoff bottleneck the error exponent, we expect the IB to be better for small n and the Chernoff bottleneck for large n, with a crossover that $P_e(n) \approx c_K n^{-1/2} e^{-n\mathcal{C}(K)}$ [35] could locate.
3) *Multiple hypotheses.* For m hypotheses, the Bayes exponent is the minimum pairwise Chernoff information [36], leading to a max–min Chernoff bottleneck.
4) *Neural encoders.* The tangent-plane majorizer is an expectation over (X, Y), so a neural encoder can be trained by stochastic gradients with a variational bound on $I(U;X)$ [12], extending the Chernoff bottleneck to continuous data.

By judging a representation by how fast the downstream error decays, rather than by an average measure of relevance, the Chernoff bottleneck offers a principled criterion for learning compact representations for inference, and we hope this work lays a foundation for its further study.

Declaration on the Use of AI

The authors used Claude Opus 5.5 to polish the writing (grammar, phrasing, readability and LaTeX formatting) of each section, and help write MATLAB code for the numerical experiments. All AI-assisted text and code were carefully reviewed, revised, and verified by the authors, who take full responsibility for the entire submission.

References

[1] D. Tang and G. Han, "Mutual information constrained Chernoff bottleneck," arXiv:2609.37994, 2026.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.

[4] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

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

[6] O. Siméoni *et al.*, "DINOv3," arXiv:2508.10104, 2025.

[7] D. Tang, J. Y. Tan, and G. Han, "Sharp regret bounds and a task-covariance correction for spectral representation learning," arXiv:2609.15825v2, 2026.

[8] A. Luthra, T. Yang, and T. Galanti, "Self-supervised contrastive learning is approximately supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2025.

[9] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, 1999, pp. 368–377.

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

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

[12] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[13] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, no. 12, Art. no. 1181, 2019.

[14] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *J. Mach. Learn. Res.*, vol. 19, no. 50, pp. 1–34, 2018.

[15] O. Shamir, S. Sabato, and N. Tishby, "Learning and generalization with the information bottleneck," *Theor. Comput. Sci.*, vol. 411, no. 29–30, pp. 2696–2711, 2010.

[16] M. Vera, P. Piantanida, and L. Rey Vega, "The role of the information bottleneck in representation learning," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, 2018, pp. 1580–1584.

[17] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Statist.*, vol. 23, no. 4, pp. 493–507, 1952.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[19] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York, NY, USA: Springer, 1998.

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

[21] J.-F. Chamberland and V. V. Veeravalli, "Decentralized detection in sensor networks," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 407–416, 2003.

[22] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, 2022.

[23] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 533–542, 1986.

[24] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, 1987.

[25] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[26] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[27] H. S. Witsenhausen and A. D. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, 1975.

[28] K. Lang, "NewsWeeder: Learning to filter netnews," in *Proc. 12th Int. Conf. Mach. Learn. (ICML)*, 1995, pp. 331–339.

[29] T. van Erven and P. Harremoës, "Rényi divergence and Kullback–Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[30] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

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

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

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

[34] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[35] R. R. Bahadur and R. Ranga Rao, "On deviations of the sample mean," *Ann. Math. Statist.*, vol. 31, no. 4, pp. 1015–1027, 1960.

[36] C. C. Leang and D. H. Johnson, "On the asymptotics of M-hypothesis Bayesian detection," *IEEE Trans. Inf. Theory*, vol. 43, no. 1, pp. 280–282, 1997.