# Credal Rate–Distortion Theory

Nicolò Bonacorsi*, Michele Caprio†, Francesca Meneghello‡, Francesco Restuccia‡
* Columbia University, † University of Manchester, ‡ Northeastern University
E-mails: nb3328@columbia.edu, michele.caprio@manchester.ac.uk, {fr.meneghello, f.restuccia}@northeastern.edu

***Abstract*—THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD. Traditional rate–distortion theory compresses inputs described by a single probability distribution. In this paper, we propose a new compression framework for robust inference where the uncertainty attached to an input is captured by considering a set of plausible likelihoods for the downstream task, i.e., a *credal set*. Existing distances between credal sets compare uncertainty models, but they do not specify the distortion induced when several inputs are merged and the decoder receives only one aggregate credal set for the whole cell. To fill this gap, we introduce a directed free-energy replacement distortion that measures the worst-case downstream cost of compression. We show that zero directed distortion occurs exactly when the replacement does not enlarge the set of plausible likelihoods. We show that the one-shot problem is exactly graph coloring, i.e., two inputs can be merged precisely when their credal likelihoods are compatible under the distortion constraint, and the minimum number of codewords is the chromatic number of the associated incompatibility graph. For product credal families and coordinate tests, block admissibility is coordinatewise and the block incompatibility graph is the OR power of the one-shot graph. The corresponding asymptotic local prefix rate is Körner graph entropy. Thus, once the credal tolerance is fixed, the compression rule is certified by the incompatibility graph, without reducing each credal set to a point prediction. On predictive credal sets built from the SVHN image-classification dataset, exact credal coloring compresses 60 inputs into 14 codewords (with a 4.286 compression ratio) while a barycenter baseline with the same number of codewords merges inputs that are credally incompatible 18.6% of the time.**

## I. Introduction

Traditional information theory starts from precise stochastic models of source law, channel law and distortion criterion [12], [14], [15]. This is acceptable when the uncertainty in the model is only *aleatoric*, i.e., when randomness is described by a single probability distribution. In many robust inference problems, however, the uncertainty is also *epistemic*; for example, a sensor may be poorly calibrated, a classifier may be trained under label ambiguity, or a data-reduction system may be deployed under model misspecification. In these cases, the object attached to an input is not a single downstream likelihood, but a *family* of plausible likelihoods. Mathematically, this family can be represented by a *credal set*, which is a nonempty closed convex set of probability distributions [3], [4]. Hence, under epistemic uncertainty, there are multiple distributions to compress. *This breaks the traditional information theory framework and requires redefining the underlying theory leveraging imprecise probability*.

Motivated by this aggregation problem, we study the rate-distortion problem in the credal setting. Our credal compressor replaces the individual credal sets of several inputs by one shared compressed credal set. The inputs sharing this description form a compression *cell*, whose aggregate credal set is determined by a replacement rule. As such, the replacement should be designed so that the plausible likelihood families associated with inputs in the cell are safely represented by the common aggregate, i.e., the common aggregate guarantees low distortion. Hence, while in traditional rate–distortion theory the distortion measures the cost of reconstructing one object by another [13], [16], in our credal rate–distortion theory the loss metric is associated with the propagation of the epistemic uncertainty over the compression and should measure the resulting increase in downstream uncertainty. *The key challenge in formulating such credal distortion metric is that credal replacement is not symmetric.* The aggregate must contain the plausible laws from all inputs in the cell, while minimizing the increase in the worst-case downstream score. As such, we introduce a directed replacement distortion that measures this increase and allows identifying cells for safe compression.

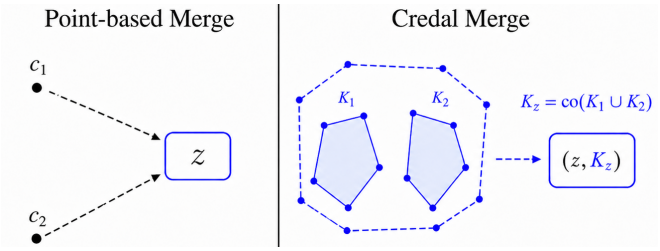


Fig. 1. Point-based VS credal merging. Point-based merging uses only representative points of the likelihood $(c_1, c_2)$, whereas credal merging acts on the full sets and assigns an aggregate set $K_z = \mathrm{co}(K_1 \cup K_2)$.

Existing work provides important tools for designing a safe replacement rule. Divergences and distances between credal sets compare uncertainty models [5], [6], while robust free-energy and entropic-risk representations give variational tools for model ambiguity [7]. Zero-error information theory shows that exact compatibility constraints can lead to graph coloring and graph entropy [8], [9]. *However, existing work does not determine the distortion induced when several credal likelihoods in a cell are replaced by the single aggregate set available to the decoder*. To fill this gap, our new theory provides guarantees on the epistemic uncertainty associated with compressed information that are not available in classical rate–distortion theory. Our replacement theory is based on the free-energy principle (detailed in Section III) and provides a one-shot graph coloring formulation for compression (Section IV). Our results in Section VII, obtained on predictive credal sets from SVHN, show that exact credal coloring can compress $N = 60$ credal likelihoods to $14$ codewords with a compression ratio of $4.286$, while a barycenter baseline with the same number of codewords has a credal violation rate of $18.6\%$, i.e., $18.6\%$ of its within-cluster pairs are incompatible under the credal replacement criterion.

## II. Problem Description

We consider a finite input alphabet $\mathcal{X}$ and a finite downstream task alphabet $\mathcal{Y}$. Each input $x \in \mathcal{X}$ denotes the observed object to be compressed, for example a sensor reading, an image, or a feature vector, while $y \in \mathcal{Y}$ denotes the quantity relevant for a downstream task, for example a class label or a state of the environment. In a sharply specified model, each input $x$ would determine a single conditional distribution on $\mathcal{Y}$. Instead, in this paper, the downstream law attached to $x$ is a credal set $K_x \subseteq \Delta(\mathcal{Y})$, where $\Delta(\mathcal{Y})$ is the probability simplex on $\mathcal{Y}$. Thus $K_x$ is the set of plausible conditional laws for $Y$ given $x$.

A compressor maps the original input to a lower-complexity description in a finite alphabet $\mathcal{Z}$. We write this deterministic map as $\mathsf{e} : \mathcal{X} \to \mathcal{Z}$, where $z = \mathsf{e}(x)$ is the extracted feature, compressed symbol, or codeword. We define a cell $\mathsf{e}^{-1}(z)$ as the set of all inputs mapped to the same $z$, i.e., $\mathsf{e}^{-1}(z) = \{x \in \mathcal{X} : \mathsf{e}(x) = z\}$. Inputs in the same cell are indistinguishable after compression.

The decoder must then assign one uncertainty description to the whole cell. Since the cell may contain several different credal likelihoods, the natural aggregate is the convex hull of all credal sets in that cell:

$$K_z = \mathrm{co}\left(\bigcup_{x \in \mathsf{e}^{-1}(z)} K_x\right), \tag{1}$$

where $\mathrm{co}(A)$ denotes the convex hull of $A$. Thus $K_z$ is the smallest credal set containing all credal sets in the cell, and represents the uncertainty left after the compressor has forgotten which $x$ was observed.

The resulting admissibility question is whether this aggregate can stand in for the individual credal sets in the cell. A pointwise reconstruction criterion would compare a representative distribution, such as a barycenter, with the original conditional law. That is not sufficient here, because the whole credal family should be preserved. Therefore, we test whether replacing each $K_x$ by the aggregate $K_z$ increases a downstream score by more than a tolerance. In the next section, we formalize the geometry of this replacement relation.

## III. Free-energy Replacement Geometry

Let $K \subseteq \Delta(\mathcal{Y})$ be a compact convex credal set and $\mathcal{K}(\mathcal{Y})$ denote the family of nonempty compact convex subsets of $\Delta(\mathcal{Y})$.

For a positive test function $g : \mathcal{Y} \to (0, \infty)$ and a credal set $K \in \mathcal{K}(\mathcal{Y})$, we define the support envelope, or upper expectation of $g$ over $K$, by

$$S_K(g) := \sup_{p \in K} \sum_{y \in \mathcal{Y}} p(y) g(y). \tag{2}$$

Let $f : \mathcal{Y} \to \mathbb{R}$ be a bounded downstream score. For example, $f(y)$ may encode the utility, loss, or log-weight assigned to outcome $y$. We define the free-energy envelope as

$$\Psi_K(f) := \log S_K(e^f) = \sup_{p \in K} \log \sum_{y \in \mathcal{Y}} p(y) e^{f(y)}. \tag{3}$$

Therefore, $\Psi_K$ is the logarithm of an upper expectation over the credal set $K$ and the supremum defines the upper robust free-energy envelope over the plausible laws in $K$. When $f$ is interpreted as a loss or cost, this is the worst-case exponential free-energy value. In this work, we restrict our analysis to the normalized test class

$$\mathcal{F} := \{f : \mathcal{Y} \to \mathbb{R} : \min f = 0,\ \max f \le 1\}, \tag{4}$$

where the constraints $\min f = 0$ and $\max f \le 1$ remove irrelevant additive constants and fix the scale of comparison.

Importantly, the logarithm in Equation (3) converts the multiplicative support identity for independent products into additivity of free energy. To see this, define the independent product of two credal sets as

$$K_1 \otimes K_2 := \mathrm{co}\left(\{p_1 \otimes p_2 : p_1 \in K_1,\ p_2 \in K_2\}\right). \tag{5}$$

By defining $(g_1 \otimes g_2)(y_1, y_2) := g_1(y_1) g_2(y_2)$ and $(f_1 \oplus f_2)(y_1, y_2) := f_1(y_1) + f_2(y_2)$, we have

$$S_{K_1 \otimes K_2}(g_1 \otimes g_2) = S_{K_1}(g_1) S_{K_2}(g_2). \tag{6}$$

Consequently, for additive scores,

$$\Psi_{K_1 \otimes K_2}(f_1 \oplus f_2) = \Psi_{K_1}(f_1) + \Psi_{K_2}(f_2). \tag{7}$$

Now, for two credal sets $K, L \subseteq \Delta(\mathcal{Y})$, we define the directed credal replacement distortion as

$$d_+(K, L) := \sup_{f \in \mathcal{F}} \{\Psi_L(f) - \Psi_K(f)\}, \tag{8}$$

which measures the largest normalized increase in robust free energy caused by using $L$ instead of $K$. Its symmetrization is

$$d(K, L) := \max\{d_+(K, L), d_+(L, K)\}. \tag{9}$$

So, a cell $C = \mathsf{e}^{-1}(z)$ is called $\varepsilon$-admissible if

$$d_+(K_x, K_z) \le \varepsilon, \quad \text{for every } x \in C, \tag{10}$$

and an encoder is called $\varepsilon$-faithful if all of its cells are $\varepsilon$-admissible. The aggregate description assigned to a codeword can safely replace the individual credal sets in that cell only up to tolerance $\varepsilon$. The compression problem reduces to a replacement question: when can the aggregate credal set $K_z$ stand in for each original credal set $K_x$ without increasing robust downstream free energy by more than $\varepsilon$? As a first structural step, we identify the zero-distortion case. The following theorem shows that directed zero distortion is exactly reverse inclusion.

*Theorem 1 (Inclusion zero set and metricity):* For all $K, L \in \mathcal{K}(\mathcal{Y})$,

$$d_+(K, L) = 0 \quad \Longleftrightarrow \quad L \subseteq K. \tag{11}$$

Consequently, $d$ is a metric on $\mathcal{K}(\mathcal{Y})$.

*Proof sketch:* If $L \subseteq K$, then $\Psi_L(f) \le \Psi_K(f)$ for every $f \in \mathcal{F}$, so $d_+(K, L) = 0$. Conversely, let $q_0 \in L \setminus K$. Since $K$ is compact and convex in a finite-dimensional simplex, strong separation gives a function $h : \mathcal{Y} \to \mathbb{R}$ and $\eta > 0$ such that $\langle q_0, h\rangle \ge \sup_{p \in K} \langle p, h\rangle + \eta$.

After shifting $h$ by its minimum and rescaling, choose $g = 1 + \alpha(h - \min h)$ so that $1 \le g \le e$ and the separation is preserved. Setting $f = \log g$ gives $f \in \mathcal{F}$ and $S_L(g) > S_K(g)$, hence $\Psi_L(f) > \Psi_K(f)$. Therefore $d_+(K, L) > 0$. The triangle inequality follows by decomposing $\Psi_M - \Psi_K = (\Psi_M - \Psi_L) + (\Psi_L - \Psi_K)$ and taking suprema over $f \in \mathcal{F}$; definiteness follows from the two directed inclusions. □

## IV. Exact One-shot Aggregation and Coloring

We now show that the admissibility criterion in Equation (10) is exactly a graph-coloring constraint. The key is that the convex-hull aggregate of a cell has a free-energy envelope equal to the maximum of the envelopes of its members.

Let $K_1, \dots, K_m \in \mathcal{K}(\mathcal{Y})$ and set $K_* := \mathrm{co}\,(\bigcup_{i=1}^m K_i)$. Since a linear functional attains the same supremum over a union and over its convex hull, we have

$$\Psi_{K_*}(f) = \max_{1 \le i \le m} \Psi_{K_i}(f). \tag{12}$$

For a nonempty cluster $C \subseteq \mathcal{X}$, define its aggregate credal set by $K_C := \mathrm{co}(\cup_{x \in C} K_x)$ and define the worst replacement cost inside the cluster as $\Delta(C) := \sup_{x \in C} d_+(K_x, K_C)$. Using Equation (12), this cost has the exact pairwise form

$$\Delta(C) = \max_{x, x' \in C} d_+(K_x, K_{x'}), \tag{13}$$

where the maximum is over ordered pairs. A cluster is $\varepsilon$-admissible if and only if every pair of inputs in the cluster has symmetric credal distance at most $\varepsilon$:

$$\Delta(C) \le \varepsilon \iff d(K_x, K_{x'}) \le \varepsilon \quad \text{for all } x, x' \in C.$$

This pairwise characterization defines a graph. Two inputs are *$\varepsilon$-compatible* if they can safely share an $\varepsilon$-admissible compression cell, equivalently if $d(K_x, K_{x'}) \le \varepsilon$. Let $G_\varepsilon = (\mathcal{X}, E_\varepsilon)$ be the incompatibility graph with vertex set $\mathcal{X}$. Its edge set $E_\varepsilon$ consists of the unordered pairs of inputs whose credal sets cannot be placed in the same $\varepsilon$-admissible cell:

$$\{x, x'\} \in E_\varepsilon \iff d(K_x, K_{x'}) > \varepsilon, \qquad x \ne x'. \tag{14}$$

Thus, *an edge means incompatibility* and a subset of vertices is independent if it contains no edge. Therefore, the $\varepsilon$-admissible compression cells are exactly the independent sets of $G_\varepsilon$.

Let $M^*(\varepsilon)$ be the minimum number of codewords used by a deterministic encoder $\mathsf{e} : \mathcal{X} \to \mathcal{Z}$ whose fibers are $\varepsilon$-admissible. Equivalently, $M^*(\varepsilon)$ is the smallest number of admissible cells needed to partition $\mathcal{X}$. Let $\chi(G_\varepsilon)$ denote the chromatic number of $G_\varepsilon$, namely the smallest number of colors needed to color the vertices so that adjacent vertices have different colors. A coloring is a map $\varphi : \mathcal{X} \to \mathcal{C}$, where $\mathcal{C}$ is a finite set of colors (codewords). It is proper for $G_\varepsilon$ if $\varphi(x) \ne \varphi(x')$ whenever $\{x, x'\} \in E_\varepsilon$ [8], [9].

A randomized encoder is a stochastic kernel $Q(\cdot|x)$ on a finite set of codewords. Its support fiber at codeword $z$ is $\{x \in \mathcal{X} : Q(z|x) > 0\}$. A randomized encoder is $\varepsilon$-faithful in the support sense if every support fiber is $\varepsilon$-admissible.

*Theorem 2 (One-shot coloring law):* For every $\varepsilon \in [0, 1]$,

$$M^*(\varepsilon) = \chi(G_\varepsilon). \tag{15}$$

If $X \sim \pi$, then the minimum output entropy over deterministic $\varepsilon$-faithful encoders is the chromatic entropy of the probabilistic graph $(G_\varepsilon, \pi)$, i.e.,

$$H_\chi(G_\varepsilon, \pi) = \min_{\varphi \text{ proper for } G_\varepsilon} H(\varphi(X)), \tag{16}$$

where $H$ is Shannon entropy in bits. Support randomization does not reduce either the fixed-length or the entropy objective.

*Proof sketch:* By Equation (13), every $\varepsilon$-faithful encoder has $\varepsilon$-admissible fibers, hence each fiber is an independent set in $G_\varepsilon$. The encoder therefore defines a proper coloring of $G_\varepsilon$. Conversely, every proper coloring has independent color classes, and these color classes are $\varepsilon$-admissible by the definition of $G_\varepsilon$. This proves Equation (15). Equation (16) follows from the same equivalence, with the output entropy equal to the entropy of the color assigned to $X$. This is the minimum-entropy coloring problem [11]. For support randomization, every emitted codeword has an independent support fiber. Choosing one supported codeword for each input gives a deterministic encoder using no more support symbols, so randomization cannot reduce the fixed-length objective. For the entropy objective, fixing the support pattern gives a convex polytope of stochastic encoders, and the output entropy is concave in the encoder. Hence its minimum is attained at an extreme point, which is deterministic row by row. □

### A. Computed Example and Robust Decision

For binary intervals $I[a, b] = \{p \in \Delta(\{0, 1\}) : p(1) \in [a, b]\}$, every normalized test has one of the two forms $(0, t)$ or $(t, 0)$ with $0 \le t \le 1$, and each positive contribution to the supremum is attained at $t = 1$. Therefore

$$d_+(I[a, b], I[c, d]) = \max \left\{ 0, \log \frac{1 + d(e-1)}{1 + b(e-1)}, \log \frac{1 + (1-c)(e-1)}{1 + (1-a)(e-1)} \right\}. \tag{17}$$

For $K_1 = I[0.90, 1]$, $K_2 = I[0.65, 0.85]$, $K_3 = I[0.40, 0.60]$, $K_4 = I[0.15, 0.35]$, $K_5 = I[0, 0.10]$, the incompatibility graphs at $\varepsilon = 0.40$ and $\varepsilon = 0.55$ are shown in Figure 2.

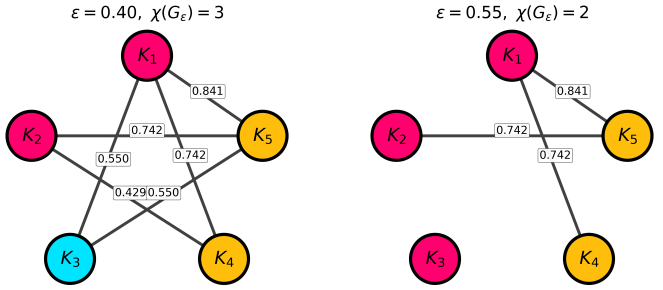


Fig. 2. Incompatibility graphs for the five-set interval family at $\varepsilon = 0.40$ (left) and $\varepsilon = 0.55$ (right). Edges indicate pairs with $d(K_i, K_j) > \varepsilon$; the displayed colorings are optimal.

At $\varepsilon = 0.40$, the optimal coloring is $\{K_1, K_2\}$, $\{K_3\}$, $\{K_4, K_5\}$, so $\chi(G_\varepsilon) = 3$. At $\varepsilon = 0.55$, the admissible clusters enlarge to $\{K_1, K_2, K_3\}$ and $\{K_4, K_5\}$, so $\chi(G_\varepsilon) = 2$. The transition occurs exactly when $\varepsilon$ crosses one of the pairwise symmetric replacement distances.

The same calculation gives a robust-classification example. Interpret $Y = 1$ as the 'cat' label and $Y = 0$ as the 'dog' label. Let $a$ be a clear cat state, $c$ a clear dog state, and $b$ an ambiguous cat-like state, modeled by $K_a = I[1, 1]$, $K_b = I[0.3, 0.8]$, $K_c = I[0, 0]$. Then $d(K_a, K_b) \simeq 0.7897$, $d(K_b, K_c) \simeq 0.8648$, and $d(K_a, K_c) = 1$. For $\varepsilon \in [0.7897, 0.8648)$, the exact coloring law forces the two-cell partition $\{a, b\} \mid \{c\}$. Under a uniform prior, assign one decision to each cell and evaluate the worst-case 0–1 loss over the original credal set attached to each input. For the partition $\{a, b\} \mid \{c\}$, the mixed cell $\{a, b\}$ has aggregate interval $I[0.3, 1]$, so the cell decision is cat. The resulting worst-case errors are 0 on $a$, 0.7 on $b$, and 0 on $c$, giving risk $(0 + 0.7 + 0)/3 = 0.2333$. By contrast, the competing partition $\{a\} \mid \{b, c\}$ has mixed-cell aggregate interval $I[0, .8]$, so the cell decision is dog. Its worst-case errors are 0 on $a$, .8 on $b$, and 0 on $c$, giving risk $(0 + .8 + 0)/3 = 0.2667$. Thus the admissible merge selected by the replacement graph also has lower worst-case decision risk than the competing merge.

## V. Intrinsic local block lift

Let $x^n = (x_1, \dots, x_n) \in \mathcal{X}^n$ and $y^n = (y_1, \dots, y_n) \in \mathcal{Y}^n$. For a block $x^n$, define $K^{\otimes}_{x^n} := K_{x_1} \otimes \cdots \otimes K_{x_n} \subseteq \Delta(\mathcal{Y}^n)$. For $1 \le i \le n$ and $f \in \mathcal{F}$, define the lifted block score as $(\Gamma_i f)(y^n) := f(y_i)$. We call $\Gamma_i f$ a local cylindrical observable: it is a function of the block alphabet $\mathcal{Y}^n$, but it depends only on the $i$-th coordinate and is constant along other coordinates. The corresponding local block test class is $\mathcal{F}^{\mathrm{loc}}_n := \{\Gamma_i f : \ 1 \le i \le n, \ f \in \mathcal{F}\}$. For arbitrary nonempty compact convex credal sets $K^n, L^n \subseteq \Delta(\mathcal{Y}^n)$, we define the local directed block distortion by

$$d^{\mathrm{loc}}_{+,n}(K^n, L^n) := \sup_{g \in \mathcal{F}^{\mathrm{loc}}_n} \{\Psi_{L^n}(g) - \Psi_{K^n}(g)\}. \tag{18}$$

Let $C \subseteq \mathcal{X}^n$ be a nonempty block cluster, and $K^{(n)}_C := \mathrm{co}\left(\bigcup_{x^n \in C} K^{\otimes}_{x^n}\right)$ be its aggregate credal set. We define $C$ as locally $\varepsilon$-admissible if

$$d^{\mathrm{loc}}_{+,n}(K^{\otimes}_{x^n}, K^{(n)}_C) \le \varepsilon \quad \text{for every } x^n \in C.$$

A deterministic $n$-block encoder is locally $\varepsilon$-faithful if all of its fibers are locally $\varepsilon$-admissible.

Let $G^{\vee n}_\varepsilon$ be the OR power of the incompatibility graph $G_\varepsilon$ with vertex set $\mathcal{X}^n$. Two distinct blocks $x^n, u^n$ are adjacent if there exists a coordinate $i$ such that $\{x_i, u_i\}$ is an edge of $G_\varepsilon$. We prove the following theorem.

*Theorem 3 (Intrinsic local block lift):* For $K_i, L_i \in \mathcal{K}(\mathcal{Y})$, $i = 1, \dots, n$,

$$d^{\mathrm{loc}}_{+,n}(K_1 \otimes \cdots \otimes K_n, L_1 \otimes \cdots \otimes L_n) = \max_{1 \le i \le n} d_+(K_i, L_i). \tag{19}$$

Consequently, a nonempty block cluster $C \subseteq \mathcal{X}^n$ is locally $\varepsilon$-admissible if and only if $C$ is an independent set of the OR power $G^{\vee n}_\varepsilon$.

*Proof sketch:* For a local cylindrical observable $g = \Gamma_i f$, we can write $\Gamma_i f = 0 \oplus \cdots \oplus 0 \oplus f \oplus 0 \oplus \cdots \oplus 0$, where $f$ appears in the $i$-th coordinate. Product additivity therefore gives $\Psi_{K_1 \otimes \cdots \otimes K_n}(\Gamma_i f) = \Psi_{K_i}(f)$, and similarly for $L_1 \otimes \cdots \otimes L_n$. Taking the supremum over $i$ and $f$ gives Equation (19). For block cluster $C$, the convex-hull max law in Equation (12) applied on $\mathcal{Y}^n$, gives $\Psi_{K^{(n)}_C}(\Gamma_i f) = \max_{u^n \in C} \Psi_{K^{\otimes}_{u^n}}(\Gamma_i f) = \max_{u^n \in C} \Psi_{K_{u_i}}(f)$. For every $x^n \in C$, $d^{\mathrm{loc}}_{+,n}(K^{\otimes}_{x^n}, K^{(n)}_C) = \max_{u^n \in C} \max_{1 \le i \le n} d_+(K_{x_i}, K_{u_i})$. Thus, $C$ is locally $\varepsilon$-admissible if and only if $\max_{1 \le i \le n} d(K_{x_i}, K_{u_i}) \le \varepsilon$ for all $x^n, u^n \in C$. By the OR power definition, this is exactly the condition that $C$ is an independent set of $G^{\vee n}_\varepsilon$. □

For a prior $\pi \in \Delta(\mathcal{X})$, let $X^n \sim \pi^{\otimes n}$ and $\mathsf{H}^{*,\mathrm{loc}}_{\pi,n}(\varepsilon)$ be the minimum output entropy of a deterministic locally $\varepsilon$-faithful $n$-block encoder. The theorem gives the exact finite-$n$ identity

$$\mathsf{H}^{*,\mathrm{loc}}_{\pi,n}(\varepsilon) = H_\chi(G^{\vee n}_\varepsilon, \pi^{\otimes n}). \tag{20}$$

By the classical OR-power graph-entropy theorem [8]–[10],

$$\lim_{n \to \infty} \frac{1}{n} \mathsf{H}^{*,\mathrm{loc}}_{\pi,n}(\varepsilon) = H_{G_\varepsilon}(\pi) := \min_{a \in \mathrm{VP}(G_\varepsilon)} \sum_{x \in \mathcal{X}} \pi(x) \log_2 \frac{1}{a_x}, \tag{21}$$

where $\mathrm{VP}(G)$ is the vertex-packing polytope, i.e. the convex hull of incidence vectors of independent sets of $G$. The minimum expected binary prefix rate per source symbol $R^{*,\mathrm{loc,pfx}}_{\pi,n}(\varepsilon)$ satisfies

$$\frac{1}{n} \mathsf{H}^{*,\mathrm{loc}}_{\pi,n}(\varepsilon) \le R^{*,\mathrm{loc,pfx}}_{\pi,n}(\varepsilon) < \frac{1}{n} \mathsf{H}^{*,\mathrm{loc}}_{\pi,n}(\varepsilon) + \frac{1}{n}. \tag{22}$$

Therefore, the local prefix rate converges to $H_{G_\varepsilon}(\pi)$. In particular, under local tests, the one-shot incompatibility graph already contains all the information needed for block coding: finite blocks correspond to OR powers, and the limiting rate is graph entropy.

## VI. Finite Blocks and Nonlocal Contrast

Computed colorings at finite blocklength can be amortized into explicit certificates for longer codes. Let $a_n(\varepsilon, \pi) := \mathsf{H}^{*,\mathrm{loc}}_{\pi,n}(\varepsilon) - nH_{G_\varepsilon}(\pi)$, and $\delta_n(\varepsilon, \pi) := a_n(\varepsilon, \pi)/n$. Products of independent sets in $G^{\vee m}_\varepsilon$ and $G^{\vee n}_\varepsilon$ are independent in $G^{\vee(m+n)}_\varepsilon$. Therefore $\mathsf{H}^{*,\mathrm{loc}}_{\pi,m+n}(\varepsilon) \le \mathsf{H}^{*,\mathrm{loc}}_{\pi,m}(\varepsilon) + \mathsf{H}^{*,\mathrm{loc}}_{\pi,n}(\varepsilon)$. It follows that

$$H_{G_\varepsilon}(\pi) = \inf_n \frac{1}{n} \mathsf{H}^{*,\mathrm{loc}}_{\pi,n}(\varepsilon), \tag{23}$$

and hence $a_n \ge 0$, $a_{m+n} \le a_m + a_n$, and $\delta_{km} \le \delta_m$ for all integers $k, m \ge 1$. Thus any certified coloring of a pilot block gives an explicit amortized certificate for all multiples of that block length. The entropy–prefix sandwich in Equation (22) then converts these finite-block entropy certificates into prefix-rate certificates. In particular, a certified $m$-block coloring can be repeated over longer blocks, with only the usual prefix and leftover-coordinate overhead.

If arbitrary block tests are allowed, coordinatewise separations can be pooled across the block, so the OR-power description no longer captures the full replacement geometry. To make this contrast explicit, let $\mathcal{F}^{\mathrm{glob}}_n$ be the class of all functions $F : \mathcal{Y}^n \to \mathbb{R}$ with $\min F = 0$ and $\max F \le 1$. Define $d^{\mathrm{glob}}_{+,n}$ by replacing the local test class with this unrestricted class.

For a credal set $K$, write the $n$-fold independent product as $K^{\otimes n} := K \otimes \cdots \otimes K$ , and define $L^{\otimes n}$ analogously. Then strict one-shot noninclusion is amplified by products: if $L \not\subseteq K$, then $d_{+,n}^{\mathrm{glob}}(K^{\otimes n}, L^{\otimes n}) \longrightarrow 1$. Indeed, choose $q \in L \setminus K$ and a function $h : \mathcal{Y} \to \mathbb{R}$ such that $\langle q, h\rangle > \sup_{p\in K}\langle p, h\rangle$. Choose $t$ strictly between these two quantities and consider the event $A_n := \left\{n^{-1}\sum_{i=1}^{n} h(Y_i) \geq t\right\}$. Let $F_n := \mathbf{1}_{A_n}$. Then $F_n \in \mathcal{F}_n^{\mathrm{glob}}$. Since $q^{\otimes n} \in L^{\otimes n}$ and $q^{\otimes n}(A_n) \to 1$, we have $\Psi_{L^{\otimes n}}(F_n) \to \log(e) = 1$. On the other hand, every product measure $p_1 \otimes \cdots \otimes p_n$ with $p_i \in K$ satisfies $n^{-1}\sum_{i=1}^{n} \mathbb{E}_{p_i}[h] \leq \sup_{p\in K}\langle p, h\rangle < t$. Since $h$ is bounded on the finite alphabet $\mathcal{Y}$, Hoeffding's inequality for independent bounded variables [1] gives $(p_1 \otimes \cdots \otimes p_n)(A_n) \to 0$ exponentially, uniformly over all $p_i \in K$. Therefore $\Psi_{K^{\otimes n}}(F_n) \to \log(1) = 0$.

Thus, unrestricted block tests can distinguish any strict one-shot noninclusion in the product limit. The OR-power reduction proved in Theorem 3 is therefore an exact law for local replacement.

## VII. SVHN Credal Compression Experiment

We instantiate the one-shot compression construction on predictive credal sets obtained from an SVHN multi-view credal-fusion pipeline [2]. The experiment starts from learned probabilistic outputs, forms one finitely generated credal likelihood for each test image, computes the induced free-energy replacement distances, and constructs the corresponding incompatibility graphs. Exact graph-coloring certificates then give the admissible codebooks predicted by Theorem 2.

For each image $x_i$, the pipeline produces three views $v \in \{1, 2, 3\}$. Each view is evaluated by a five-member ensemble $m \in \{1, \ldots, 5\}$, giving predictive distributions $p_{i,v,m} \in \Delta(\{0, \ldots, 9\})$. We attach to $x_i$ the credal likelihood $K_i = \mathrm{co}\left(\{p_{i,v,m} : v \in \{1, 2, 3\},\ m \in \{1, \ldots, 5\}\}\right)$. All directed distances are computed by finite support-envelope optimizations. For each tolerance $\varepsilon$, $G_\varepsilon$ joins two inputs whenever $d(K_i, K_j) > \varepsilon$. Recomputing all distances with stricter tolerances changed symmetrized distances by at most $1.45 \cdot 10^{-10}$ and produced no edge flips in the reported graphs.

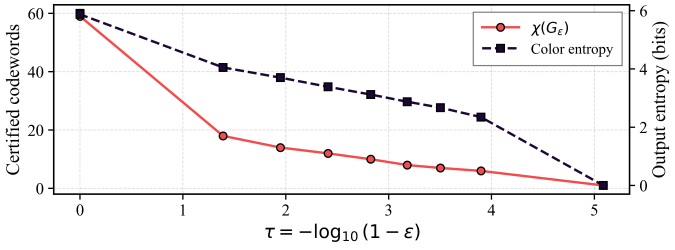


Fig. 3. Exact one-shot credal coloring on $N = 60$ SVHN credal sets.

Figure 3 shows the exact one-shot credal-coloring curve on the $N = 60$ SVHN credal likelihoods, while Table I reports selected certified operating points. For example, we select $\varepsilon = 0.988745$ and $\tau = -\log_{10}(1 - \varepsilon) = 1.949$, which lead to $\chi(G_\varepsilon) = 14$. Thus the 60 credal likelihoods are compressed into 14 admissible codewords, giving compression ratio 4.286 and cluster majority-label purity 0.783, where purity denotes the cluster-size-weighted majority-label fraction. Increasing $\varepsilon$ to $\varepsilon = 0.998508$, we obtain $\chi(G_\varepsilon) = 10$ and compression ratio 6.000, with purity 0.700. This increase in $\varepsilon$ removes edges from $G_\varepsilon$, so fewer codewords are required (see Theorem 2). However, in this experiment, this leads to coarser and less label-pure clusters.

TABLE I
Examples of certified operating points for credal compression.

| $\varepsilon$ | $\chi(G_\varepsilon)$ | $N/\chi$ | entropy | purity |
|---|---|---|---|---|
| 0.959367 | 18 | 3.333 | 4.052 | 0.767 |
| 0.988745 | 14 | 4.286 | 3.702 | 0.783 |
| 0.996122 | 12 | 5.000 | 3.387 | 0.733 |
| 0.998508 | 10 | 6.000 | 3.121 | 0.700 |
| 0.999339 | 8 | 7.500 | 2.876 | 0.650 |

We compare our credal-based metric with a minimax barycenter surrogate. Let $\bar{p}_i = (1/15)\sum_{v,m} p_{i,v,m}$. At each operating point, we take $k = \chi(G_\varepsilon)$ and compute an exact minimax $k$-center compression of the barycenters, minimizing worst within-cluster total-variation radius. This gives the point-summary baseline the same number of clusters as the credal coloring. The reported point-TV radius is this minimax within-cluster total-variation (TV) radius computed on the barycenter predictions. At $\varepsilon = 0.988745$ and $k = 14$, the barycenter baseline reduces the point-TV radius from 0.9629 to 0.2990, but its credal violation rate is 0.186. Thus about 18.6% of the within-cluster pairs produced by the barycenter clustering are credally incompatible. Figure 4 reports this comparison. The barycenter surrogate improves reconstruction of averaged predictions, but exact credal coloring gives the relevant guarantee for the credal replacement objective because it keeps the violation rate at zero by construction.

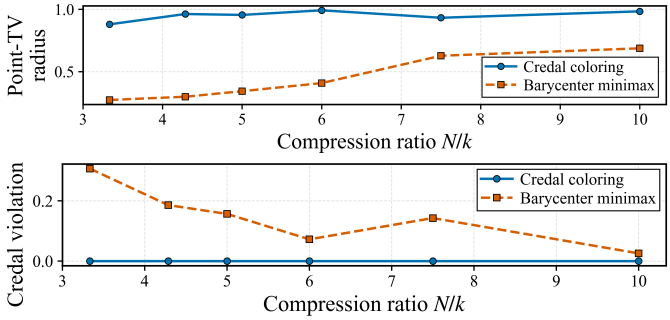


Fig. 4. Exact credal coloring VS minimax barycenter compression with $k = \chi(G_\varepsilon)$ clusters. Top: point-TV radius. Bottom: credal violation rate.

## VIII. Concluding Remarks

In this paper, we introduced credal rate–distortion theory for compression under epistemic uncertainty. We formulated a free-energy replacement distortion that turns local uncertainty-preserving compression of finite credal likelihoods into an exact graph problem. Its exact zero-set validates it as a correct criterion for credal replacement. Under local coordinate tests this yields an exact zero-error coding law — one-shot chromatic number and asymptotic Körner graph entropy.

## REFERENCES


[1] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[2] A. Q. M. S. Sayyed, M. Caprio, N. D. Bastian, and F. Restuccia, "Decision-driven credal information fusion," in *Proc. of the 29th International Conference on Information Fusion (FUSION)*, 2026.

[3] P. Walley, *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[4] M. C. M. Troffaes and G. de Cooman, *Lower Previsions*. Wiley, 2014.

[5] J. Abellán and M. Gómez, "Measures of divergence on credal sets," *Fuzzy Sets and Systems*, vol. 157, no. 11, pp. 1514–1531, 2006.

[6] S. L. Chau, M. Caprio, and K. Muandet, "Integral imprecise probability metrics," arXiv preprint arXiv:2505.16156, 2025.

[7] H. Föllmer and T. Knispel, "Entropic risk measures: coherence vs. convexity, model ambiguity, and robust large deviations," *Stochastics and Dynamics*, vol. 11, no. 2–3, pp. 333–351, 2011.

[8] J. Körner, "Coding of an information source having ambiguous alphabet and the entropy of graphs," in *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, Academia, Prague, 1973, pp. 411–425.

[9] H. S. Witsenhausen, "The zero-error side information problem and chromatic numbers," *IEEE Transactions on Information Theory*, vol. 22, no. 5, pp. 592–593, 1976.

[10] N. Alon and A. Orlitsky, "Source coding and graph entropies," *IEEE Transactions on Information Theory*, vol. 42, no. 5, pp. 1329–1339, 1996.

[11] J. Cardinal, S. Fiorini, and G. Van Assche, "On minimum entropy graph colorings," in *Proceedings of the IEEE International Symposium on Information Theory*, 2004, p. 43.

[12] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948; vol. 27, no. 4, pp. 623–656, 1948.

[13] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record*, vol. 7, part 4, pp. 142–163, 1959.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, Hoboken, NJ, 2006.

[15] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, Cambridge, 2011.

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

[17] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time, I," *Communications on Pure and Applied Mathematics*, vol. 28, no. 1, pp. 1–47, 1975.

[18] R. T. Rockafellar, *Convex Analysis*. Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, NJ, 1970.

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

[20] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, 2012.

[21] T. Augustin, F. P. A. Coolen, G. de Cooman, and M. C. M. Troffaes, eds., *Introduction to Imprecise Probabilities*. Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester, 2014.

[22] G. de Cooman, E. Miranda, and M. Zaffalon, "Independent natural extension," *Artificial Intelligence*, vol. 175, no. 12–13, pp. 1911–1950, 2011.

[23] M. Caprio, S. Dutta, K. J. Jang, V. Lin, R. Ivanov, O. Sokolsky, and I. Lee, "Credal Bayesian deep learning," *Transactions on Machine Learning Research*, 2024.

[24] M. Caprio, M. Sultana, E. G. Elia, and F. Cuzzolin, "Credal learning theory," *Advances in Neural Information Processing Systems*, vol. 37, 2024.