Log-Likelihood Loss for Semantic Compression

Anuj Kumar Yadav*†, Dan Song†, Yanina Y. Shkel*, Ayfer Özgür†

*School of Computer & Communication Sciences, EPFL, Switzerland
†Department of Electrical Engineering, Stanford University, USA
Emails: {anuj.yadav, yanina.shkel}@epfl.ch,
{songdan, aozgur}@stanford.edu

***Abstract*—We study lossy source coding under a distortion measure defined by the negative log-likelihood induced by a prescribed conditional distribution $P_{X|U}$. This *log-likelihood distortion* models compression settings in which the reconstruction is a semantic representation from which the source can be probabilistically generated, rather than a pointwise approximation. We formulate the corresponding rate–distortion problem and characterize fundamental properties of the resulting rate–distortion function, including its connections to lossy compression under log-loss, classical rate–distortion problems with arbitrary distortion measures, and rate–distortion with perfect perception.**

***Index Terms*—log-likelihood, semantic compression, rate distortion, perception**

I. Introduction

Given a source $X \sim P_X$ taking values in an alphabet $\mathcal{X}$, and a conditional distribution $P_{X|U}$ such that $P_{X|U}(\cdot \mid u)$ is a valid probability distribution on $\mathcal{X}$ for every $u \in \mathcal{U}$, we study the rate–distortion trade-off for lossy compression under the distortion measure $d_{\ell\ell} : \mathcal{X} \times \mathcal{U} \to [0, \infty]$ defined as

$$d_{\ell\ell}(x, y) \triangleq \log \frac{1}{P_{X|U}(x \mid y)}. \tag{1}$$

We refer to $d_{\ell\ell}(x, y)$ as the *log-likelihood loss* (see Fig. 1).*

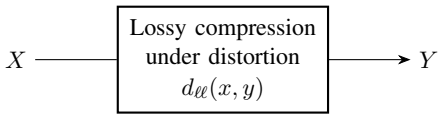
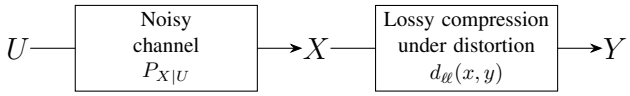
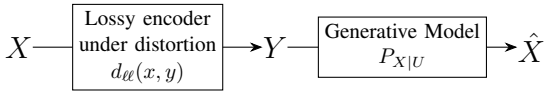


Fig. 1: Log-likelihood loss based lossy compression.

We propose this distortion measure to model modern compression settings in which lossy compression serves a dual purpose: producing a compact representation while preserving task-relevant features or semantic information about the source. For instance, X may represent an image and U a latent semantic representation describing its content, with $P_{X|U}$ modeling the probabilistic relationship between the semantics and the image, for example as induced by a trained generative or reconstruction model. Similarly, X may correspond to a text document and U to its semantic summary. In such settings, compression under the log-likelihood distortion in (1) seeks a representation Y of limited rate for which the original source realization X is maximally likely given Y through the probabilistic mapping $P_{X|U}$. Equivalently, the distortion quantifies the negative log-likelihood of reconstructing the source from its compressed representation, aligning compression with probabilistic reconstruction fidelity rather than traditional signal-level metrics, which quantify fidelity through a pointwise discrepancy between X and its reconstruction Y. As such, the log-likelihood loss naturally models scenarios in which the reconstruction Y does not represent a pointwise approximation of X, but rather an abstract description from which X can be probabilistically generated.



(a) Compression based denoising



(b) Compression with generative reconstruction

Fig. 2

Beyond semantic compression, the proposed distortion measure can be used to model several related but distinct application settings. In Fig. 2-(a), U represents an underlying source signal and X its noisy observation obtained through a channel $P_{X|U}$. The encoder observes X and compresses it under the distortion measure in (1). In this setting, the log-likelihood loss promotes denoising through compression. This effect was observed in [3], where it was shown that when compression is performed at the average distortion level $\mathbb{E}[d_{\ell\ell}(X, Y)] = H(X|U)$, the resulting representation serves as a universal denoising of X. A closely related idea appeared earlier in [4], where the distortion measure (1) was used as a cost function for entropic optimal transport in the context of generative modeling from privatized data.

Finally, Fig. 2-(b) illustrates a complementary application in which the decoder is fixed in advance to a given probabilistic

This work is partially funded by the Swiss NSF grant number 211337. For detailed proofs, please refer to the extended version [1].

*Our primary focus in this paper is the rate–distortion trade-off under the proposed log-likelihood loss. While the rate–distortion function characterizes the asymptotic trade-off for lossy compression of i.i.d. sources, it is also relevant in the one-shot setting, as shown in [2]. We adopt scalar notation in our figures for simplicity and because it is often more appropriate for semantic compression settings.

reconstruction model $P_{X|U}$. In this setting, the source is compressed with the knowledge that reconstruction will be performed probabilistically according to $P_{X|U}$, for example via a generative or AI-based model. The distortion measure (1) ensures that the compression strategy is matched to the decoder in the sense of maximizing the likelihood of reproducing the original source.

While these scenarios represent distinct applications of the proposed log-likelihood distortion measure—ranging from semantic compression to denoising and fixed-decoder reconstruction—they collectively highlight its relevance as a natural distortion measure for modern compression problems involving semantic representations.

Related works: Semantic compression has gained traction in recent years, and several models have been proposed for semantics-preserving lossy compression. In particular, foundation-model based semantic compression techniques have been studied in [5]–[7]. An approach to semantic compression based on information lattice learning was proposed in [8]. In [9] the authors studied an arithmetic coding based method for semantic lossless compression. [10] studied a rate-distortion framework for semantic compression via a distortion measure inspired by the information bottleneck constraint, along with an observation distortion measure. Other frameworks for semantic compression based on rate-distortion framework have been proposed under different metrics in [11]–[13]. Our approach significantly deviates from these earlier approaches as we capture semantics via a simple and intuitive novel distortion measure inspired by log-loss, where the goal is to compress X into Y such that X remains highly likely under $P_{X|U}(\cdot|Y)$.

Our contributions and organization: In this paper, we focus on studying the rate-distortion function under the proposed log-likelihood loss. In section III, we present several properties of the rate-distortion function and explore its relations with log-loss distortion. Though, the log-likelihood loss in (1) looks deceptively restricted, in section IV we show that it generalizes several commonly studied classical rate-distortion frameworks indicating its applicability across various rate-distortion scenarios, followed by an illustrative example. In section V, we show that our framework provides an achievable scheme to attain rate-distortion with perfect perception for a special class of rate-distortion problems.

II. Notations and Background

The PMF of a discrete random variable (PDF for continuous random variables) X is denoted using a upper case letter, say P_X, while the probability of an event is denoted using the bold-face letter $\mathbb{P}$. Given a random variable X, its support (and sets in general) is denoted by $\mathcal{X}$, while a realization is denoted by lower case letter, for example, $x \in \mathcal{X}$. We use $\Delta(\mathcal{X})$ to denote the probability simplex on $\mathcal{X}$. The expectation of the random variable X is $\mathbb{E}[X]$. We use $H(X)$ denote the Shannon entropy (differential entropy for continuous random variables) of a random variable X. All logarithms are to the base e, unless stated otherwise.

Definition 1 (Rate-Distortion Function (RDF))**.** *Given an information source $X \sim P_X$ taking values in $\mathcal{X}$. Let $Y \in \mathcal{Y}$ be the lossy reconstruction of X under the distortion measure $d : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$. Then, the rate-distortion function (RDF) for (X, d) is given by*

$$R(D) = \min_{W_{Y|X}:\ \mathbb{E}_{XY}[d(X,Y)] \le D} I(X;Y) \tag{2}$$

for any $D \ge 0$.

Definition 2 (Rate-Distortion Function under Log-Likelihood Loss)**.** *Consider the lossy compression of X into a reconstruction $Y \in \mathcal{U}$ under the log-likelihood distortion measure $d_{\ell\ell} : \mathcal{X} \times \mathcal{U} \to [0, \infty]$ such that $d_{\ell\ell}(x, y) := -\log P_{X|U}(x|y)$, where $P_{X|U}(x|u)$ for $x \in \mathcal{X}$ and $u \in \mathcal{U}$ is a given conditional distribution. The rate-distortion function under log-likelihood loss for $(X, P_{X|U})$ is defined as follows*

$$R_{\ell\ell}(D) = \min_{W_{Y|X}:\mathbb{E}_{XY}[d_{\ell\ell}(X,Y)] \le D} I(X;Y). \tag{3}$$

Remark 1. *$R_{\ell\ell}(D)$ can also be expressed as*

$$R_{\ell\ell}(D) = \min_{\substack{W_{Y|X}:\\ H(X|Y)+\mathbb{E}_Y[D_{\mathrm{KL}}(W_{X|Y}(\cdot|Y)||P_{X|U}(\cdot|Y))] \le D}} I(X;Y) \tag{4}$$

$$= \min_{W_{Y|X}:\mathbb{E}_Y[\mathrm{CE}(W_{X|Y}(\cdot|Y)||P_{X|U}(\cdot|Y))] \le D} I(X;Y). \tag{5}$$

where $\mathrm{CE}(\cdot\|\cdot)$ denotes the cross-entropy function.

III. Properties of $R_{\ell\ell}(D)$

In the following, we state some properties of the RDF under log-likelihood loss and its connections with other rate-distortion problems studied in the literature.

Theorem 1. *For given $(X, P_{X|U})$, $R_{\ell\ell}(D)$ is defined only for the distortion values $D \in [D_{\min}, D_{\max}]$ where*

$$D_{\min} := \mathbb{E}_X\left[\min_u \log \frac{1}{P_{X|U}(X|u)}\right], \tag{6}$$

$$D_{\max} := \min_u \mathbb{E}_X\left[\log \frac{1}{P_{X|U}(X|u)}\right]. \tag{7}$$

(i) *At distortion $D = D_{\min}$, we have*

$$R_{\ell\ell}(D_{\min}) = \min_{Q \in \Delta(\mathcal{U})} \mathbb{E}_X\left[-\log \sum_{u \in T(X)} Q_U(u)\right] \tag{8}$$

where $T(x) := \{u \in \mathcal{U} : u = \arg\max_{u'} P_{X|U}(x|u')\}$. The optimal reconstruction is a randomized maximum-likelihood (ML) decoder: for each x, the decoder outputs a random $y \in T(x)$, where the optimal tie-randomization strategy over $T(x)$ is given by the minimizer of (8)*.*

(ii) *At distortion $D = D_{\max}$, we have $R_{\ell\ell}(D_{\max}) = 0$.*

(iii) *Assume there exists $(U, X) \sim P_{U,X}$ which is consistent with $(X, P_{X|U})$. At distortion $D^* = H(X|U)$, $R_{\ell\ell}(D)$ admits a closed-form expression i.e.,*

$$R_{\ell\ell}(D^*) = I(U;X) = H(X) - D^*.$$

We refer to the distortion level D^* as a special operating point for $(X, P_{X|U})$. At distortion $D = D^*$, the reconstruction Y of the optimal compressor is a sample from the posterior $P_{U|X}$, which further implies $Y \sim P_U$.

A. Connection to Log-loss Distortion

Rate distortion under log-loss has been studied widely in the literature [14]–[16]. For a given source $X \sim P_X$, the decoder outputs a predictive distribution $W \in \Delta(\mathcal{X})$ under the log-loss distortion $d(x, W) := -\log W(x)$. At a given distortion level D, the log-loss RDF is given by

$$R_\ell(D) = \min_{W_{Y|X}: H(X|Y) \le D} I(X;Y) \tag{9}$$

$$= H(X) - D \tag{10}$$

where $D \in [0, H(X)]$.

The log-likelihood loss can also be interpreted as allowing the decoder to output a predictive distribution, however in this case the predictive distribution is restricted to a strict subset of the simplex, namely the family $\{P_{X|U}(\cdot|u)\}_{u \in \mathcal{U}} \subset \Delta(\mathcal{X})$. Concretely, if the decoder outputs a symbol $y \in \mathcal{U}$, this induces the predictive distribution $W_y := P_{X|U}(\cdot|y)$, and the log-likelihood loss becomes

$$d_{\ell\ell}(x, y) = -\log P_{X|U}(x|y) = d(x, W_y). \tag{11}$$

Hence, the log-likelihood loss can be viewed as a restricted version of the log-loss framework where the decoder is allowed to output a certain class of predictive distributions. As such, the trade-off under log-loss, serves as a lower bound for the trade-off under log-likelihood loss as stated in Theorem 2. From a practical perspective, this can model settings where the decoder is indeed restricted to use a given predictive model as illustrated in Fig. 2-(b).

The relationship between the log-likelihood loss and the log-loss can be further understood by comparing the corresponding distortion constraints in (4) and (9). Under the log-loss distortion in (9), the distortion constraint bounds only the conditional entropy $H(X|Y)$, thereby enforcing that the representation Y is predictive of X. In contrast, with the log-likelihood distortion, we bound $H(X|Y)$ together with an additional averaged KL divergence term that measures the discrepancy between the predictive distribution induced by Y and the prescribed probabilistic model $P_{X|U}$.

In addition to modeling a constrained decoder, this additional term can be interpreted as guiding the compression toward a desired notion of semantics. With log-loss alone, there is no restriction on the form of the predictive distribution. Therefore, the distortion does not distinguish, for example in the case of images, between a coarse quantization of the pixels and a semantic representation, provided both are equally predictive of the source image. In contrast, when using the log-likelihood distortion, a prescribed semantic structure encoded through $P_{X|U}$ allows the compression to be steered toward representations that align with that notion of semantics.

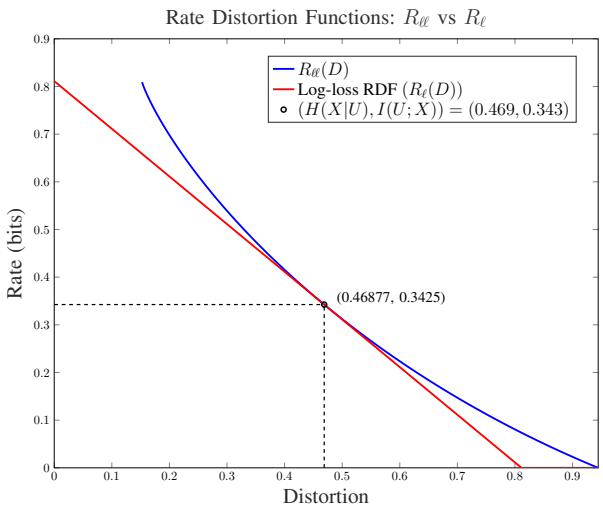


Fig. 3: $(X, P_{X|U})$: $X \sim \mathrm{Ber}(0.25)$ and $P_{X|U} \sim \mathrm{BSC}(0.1)$, the figure plots $R_\ell(D)$ for $D \in [0, H(X) = 0.8113]$ and $R_{\ell\ell}(D)$ for $D \in [D_{\min} = 0.152, D_{\max} = 0.945]$. At $D^* = 0.469$, both RDFs coincide. D^* is the special operating point for $(X, P_{X|U})$ with $U \sim \mathrm{Ber}(0.1875)$.

Theorem 2. *Let $R_\ell(D)$ be the RDF under log-loss for X, and let $R_{\ell\ell}(D)$ be the RDF under log-likelihood loss for $(X, P_{X|U})$. Then, for all feasible $D \in [D_{\min}, D_{\max}]$,*

$$R_{\ell\ell}(D) \ge R_\ell(D) \tag{12}$$

where the equality holds for all special operating points as defined in Theorem 1-(iii), i.e. whenever there exists $(U, X) \sim P_{U,X}$ which is consistent with $(X, P_{X|U})$ and $D = H(X|U)$.

Corollary 1. *Assume there exists $(U_1, X) \sim P_{U_1,X}$ and $(U_2, X) \sim P_{U_2,X}$ s.t. both $P_{U_1,X}$ and $P_{U_2,X}$ are consistent with $(X, P_{X|U})$. Let $D_1 = H(X|U_1)$ and $D_2 = H(X|U_2)$, then for all $D_1 \le D \le D_2$,*

$$R_{\ell\ell}(D) = H(X) - D.$$

The corollary follows as a consequence of the fact that $(X, P_{X|U})$ has multiple special operating points and $R_{\ell\ell}(D)$ touches the linear log-loss trade-off at all these points. In Fig. 4 we provide an example where this is the case.

IV. Connection to General Rate-Distortion Problems

In this section, we establish correspondence between lossy compression of a source $X \sim P_X$ under a general distortion measure d and log-likelihood loss. Given a source $X \sim P_X$ and a distortion measure $d : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$, suppose there exists a $\lambda > 0$ and a nonnegative function $\mu(x, \lambda)$, independent of y, such that for every $y \in \mathcal{Y}$ the following defines a valid conditional distribution

$$P^{\lambda}_{X|Y}(x|y) = \mu(x, \lambda) e^{-\lambda d(x,y)}. \tag{13}$$

Then, the rate-distortion problem for (X, d) at distortion level D can be reformulated as a log-likelihood loss problem $(X, P_{X|U})$ at distortion $\widetilde{D}$ related to D via an affine mapping.

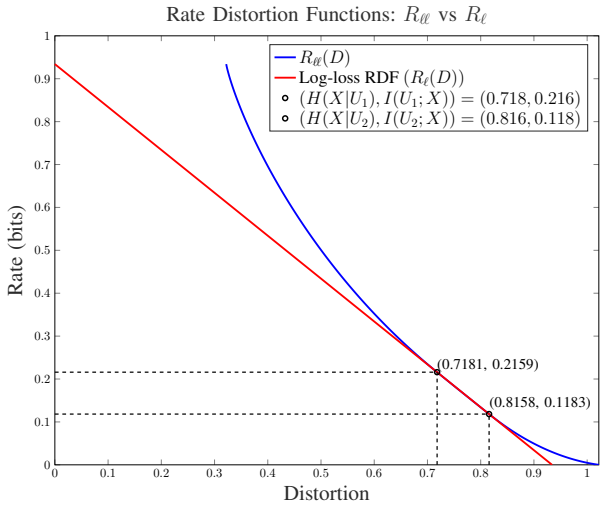


Fig. 4: $(X, P_{X|U})$: $X \sim \text{Ber}(0.35)$ and $P_{X|U} = [0.8, 0.4, 0.2; 0.2, 0.6, 0.8]$, the figure plots $R_\ell(D)$ for $D \in [0, 0.934]$ and $R_{\ell\ell}(D)$ for $D \in [0.322, 1.022]$. Both RDFs coincide for $D^* \in [0.718, 0.816]$. Every D^* in this interval is a special operating point with a corresponding (U_i, X) consistent with $(X, P_{X|U})$.

Now, given such a (X, d), consider compressing the same source $X \sim P_X$ under log likelihood loss $P_{X|U}$ chosen to coincide with $P^\lambda_{X|Y}$ above. Therefore, we have

$$\begin{aligned} d_{\ell\ell}(x, y) &= -\log P_{X|U}(x|y) && (14)\\ &= -\log\left(\mu(x, \lambda)e^{-\lambda d(x,y)}\right) && (15)\\ &= \lambda d(x, y) - \log(\mu(x, \lambda)). && (16)\end{aligned}$$

By taking expectation on both sides w.r.t (X, Y), we have

$$\mathbb{E}_{XY}[d_{\ell\ell}(X, Y)] = \lambda\mathbb{E}_{XY}[d(X, Y)] - \mathbb{E}_X[\log(\mu(X, \lambda))] \quad (17)$$

Then we have

$$\begin{aligned} R_{\ell\ell}(\widetilde{D}) &= \min_{W_{Y|X}:\mathbb{E}_{XY}[d_{\ell\ell}(X,Y)]\le\widetilde{D}} I(X;Y) && (18)\\ &= \min_{W_{Y|X}:\mathbb{E}_{XY}[d(X,Y)]\le\frac{\widetilde{D}+\mathbb{E}_X[\log(\mu(X,\lambda))]}{\lambda}} I(X;Y) && (19)\\ &= R(D) && (20)\end{aligned}$$

where we chose

$$\widetilde{D} = \lambda D - \mathbb{E}_X[\log(\mu(X, \lambda))]. \quad (21)$$

In other words, if $P^\lambda_{X|Y}$ in (13) exists then (X, d) and $(X, P_{X|U})$ with $P_{X|U} = P^\lambda_{X|Y}$ are equivalent in the sense that the corresponding RDFs are related through an affine transformation of the distortion value. This shows that the log-likelihood distortion problem $(X, P_{X|U})$ is as general as (X, d). Therefore, we cannot expect to characterize the RDF for $(X, P_{X|U})$ in full generality unless we can characterize (X, d) for any X and d.

In the sequel, we focus on a class of rate-distortion problems (X, d) for which the rate-distortion function can be computed by solving a single parameter optimization problem.

Theorem 3. *Given an information source $X \sim P_X$ with entropy $H(X)$ and a distortion measure $d(\cdot,\cdot)$, let $\mathcal{I}$ denote the set of all $\lambda > 0$ such that there exists*

(i) a real-valued function $\mu(\cdot, \lambda) : \mathcal{X} \to [0, \infty)$,
(ii) a coupling $P^\lambda_{Y,X}$ on $(\mathcal{Y} \times \mathcal{X})$, such that $\forall\ (y, x)$ we have $P^\lambda_{X|Y}(x|y) = \mu(x, \lambda)e^{-\lambda d(x,y)}$.

If $\mathcal{I}$ is non-empty, i.e., $\mathcal{I} \neq \varnothing$, the RDF $R(D)$ for (X, d) can be expressed as a single-parameter optimization problem i.e.,

$$R(D) = \max_{\lambda\in\mathcal{I}} \Big(H(X) + \mathbb{E}_X[\log(\mu(X, \lambda))] - \lambda D\Big). \quad (22)$$

The proof is based on solving the dual form for $R(D)$ through the Lagrangian form and applying KKT conditions while invoking the feasibility conditions to restrict the optimization to $\lambda \in \mathcal{I}$. The dual form of $R(D)$ also appears in [17], [18].

Remark 2. *Theorem 3 characterizes a class of rate–distortion problems for which a 'generalized Shannon-type lower bound' is tight, yielding the single-parameter representation in* (22)*. In the special case of continuous sources with difference distortions, this tight bound reduces to the classical Shannon lower bound [19].*

Note that the conditions of Theorem 3 are stronger than the condition in (13) as they also require the existence of $Y \sim Q_Y$ (equivalently, a coupling $P^\lambda_{X,Y}$ with $P^\lambda_X = P_X$) such that a single function $\mu(x, \lambda)$ simultaneously normalizes $P^\lambda_{X|Y}(\cdot|y)$ for all y. The set $\mathcal{I}$ is a collection of those values of λ for which such a coupling exists; accordingly, the maximization in (22) is restricted to $\lambda \in \mathcal{I}$.

Since the conditions of Theorem 3 imply (13), the RDF for the corresponding $(X, P_{X|U})$ with $P_{X|U} = P^\lambda_{X|Y}$ can also be expressed in the form (22) through the translation in (20). This gives us a family of log-likelihood loss problems for which we can characterize the corresponding RDF as we illustrate in the following example.

Example 1 (Binary Source with Hamming Distortion)**.** *Let $X \sim \text{Ber}(p)$, where $p \in [0, 1/2]$, and the distortion measure is the Hamming distortion i.e.,*

$$d_H(x, y) = \begin{cases} 0\ ;\ \text{if } x = y \\ 1\ ;\ \text{if } x \neq y \end{cases} \quad (23)$$

Existence of couplings: Fix a $\lambda > 0$. Then, due to condition (ii) in Theorem 3, the conditional distribution $P^\lambda_{X|Y}$ is given by

$$P^\lambda_{X|Y}(x|y) = \begin{bmatrix} \frac{1}{1+e^{-\lambda}} & \frac{e^{-\lambda}}{1+e^{-\lambda}} \\ \frac{e^{-\lambda}}{1+e^{-\lambda}} & \frac{1}{1+e^{-\lambda}} \end{bmatrix} \quad (24)$$

where $\mu(0, \lambda) = \mu(1, \lambda) = \frac{1}{1+e^{-\lambda}}$. Let $Y \sim \text{Ber}(q)$ for some q. We will show that there exists a $\lambda > 0$ which guarantees

the existence of a coupling. Using Bayes rule, we have that $P^{\lambda}_{Y|X}$ is

$$P^{\lambda}_{Y|X}(y|x) = \begin{bmatrix} \frac{\left(\frac{1}{1+e^{-\lambda}}\right)(1-q)}{1-p} & \frac{\left(\frac{e^{-\lambda}}{1+e^{-\lambda}}\right)q}{1-p} \\ \frac{\left(\frac{e^{-\lambda}}{1+e^{-\lambda}}\right)(1-q)}{p} & \frac{\left(\frac{1}{1+e^{-\lambda}}\right)q}{p} \end{bmatrix}. \tag{25}$$

We see that the rows of $P^{\lambda}_{Y|X}$ sum to 1 iff $q = \frac{p(1+e^{\lambda})-1}{e^{\lambda}-1}$. For $q \in [0,1]$, we have that $\lambda \geq \log(\frac{1-p}{p})$. Thus, given $p \in [0, 1/2]$ we observe that a valid coupling P^{λ}_{XY} exists for any $\lambda \in \mathcal{I} := [\log(\frac{1-p}{p}), \infty)$. Therefore, from (22)

$$R(D) = \max_{\lambda \in \mathcal{I}} \left(H(p) + \log \frac{1}{1+e^{-\lambda}} - \lambda D \right) \tag{26}$$

$$= H(p) + \log(1-D) - D \log\left(\frac{1-D}{D}\right) \tag{27}$$

$$= H(p) - H(D), \tag{28}$$

where $D \leq P$ and the maximizer is $\lambda^* = \log(\frac{1-D}{D})$.

Translation into log-likelihood loss $(X, P_{X|U})$: Fix a $\lambda_0 \in \mathcal{I}$. Now, choose $P_{X|U}$ in $(X, P_{X|U})$ to be $P^{\lambda_0}_{X|Y}$. Thus,

$$P_{X|U}(x|u) = \begin{bmatrix} \frac{1}{1+e^{-\lambda_0}} & \frac{e^{-\lambda_0}}{1+e^{-\lambda_0}} \\ \frac{e^{-\lambda_0}}{1+e^{-\lambda_0}} & \frac{1}{1+e^{-\lambda_0}} \end{bmatrix}. \tag{29}$$

From Eq. 21 we have

$$\widetilde{D} = \lambda_0 D - \mathbb{E}_X \left[\log\left(\frac{1}{1+e^{-\lambda_0}} \right) \right] \tag{30}$$

$$= \lambda_0 D + \log(1+e^{-\lambda_0}). \tag{31}$$

Therefore, we obtain that

$$R_{\ell\ell}(\widetilde{D}) = H(p) - H\left(\frac{\widetilde{D} - \log(1+e^{-\lambda_0})}{\lambda_0} \right) \tag{32}$$

where $\widetilde{D} \in [\log(1+e^{-\lambda_0}), \lambda_0 p + \log(1+e^{-\lambda_0})]$.

Examples with this translation include binary source with asymmetric distortion, Gaussian source with squared-error distortion, among others.

V. Connection to Rate-Distortion-Perception

Next, we show that our framework provides an achievable scheme for attaining rate distortion with prefect perception.

Definition 3 (Rate Distortion-Perception [20], [21])**.** *Given an information source $X \sim P_X$ taking values in $\mathcal{X}$. Let $Y \in \mathcal{X}$ be the lossy reconstruction of X under the distortion measure $d : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ and perception measure $\gamma : \Delta(\mathcal{X}) \times \Delta(\mathcal{X}) \to [0, \infty]$. We assume $\gamma(P, Q) = 0$ iff $P = Q$. Then, the rate distortion perception function for (X, d, γ) is given by*

$$R(D, \Gamma) = \min_{\substack{W_{Y|X}: \\ \mathbb{E}_{XY}[d(X,Y)] \leq D,\ \gamma(P_X; Q_Y) \leq \Gamma}} I(X;Y) \tag{33}$$

for any $D, \Gamma \geq 0$.

From Theorem 1-(iii), given $(X, P_{X|U})$ if there exists (U, X) consistent with $(X, P_{X|U})$ and $D = H(X|U)$, then the optimal reconstruction Y is such that $P_Y = P_U$. This observation can be translated to an achievable strategy for the rate-distortion problem with perfect perception as follows. Given X assume we choose any $P_{Z|X}(z|x)$ and process X into Z according to $P_{Z|X}$. Then, we compress Z under the log-likelihood distortion $d_{\ell\ell}(z, x) = -\log P_{Z|X}(z|x)$ at distortion level $H(Z|X)$ which ensures that the reconstruction Y has the distribution P_X. End to end, this yields an achievable scheme for the rate-distortion-perception problem with perfect perception. In other words, perfect perception can be achieved by first taking X and passing it through an arbitrary noisy channel $P_{Z|X}(z|x)$ and then lossily compressing the noisy observation under the log-likelihood distortion induced by the noisy channel at the fixed distortion level $H(Z|X)$.

In [20], a similar construction is used in the case of MSE distortion to upper bound $R(D, 0)$. In the next theorem, we will characterize for what class of rate-distortion problems (X, d) is such a scheme based on log-likelihood distortion $(Z, P_{X|Z})$ optimal in terms of achieving $R(D, 0)$.

Theorem 4. *Let (X, d, γ) describe a rate-distortion-perception problem such that $|\mathcal{X}| < \infty$ and $d(\cdot, \cdot)$ is such that the element-wise exponential matrix V, defined by $V(x, y) := e^{-\lambda d(x,y)}$ is completely positive (CP), for every $\lambda > 0$.* [†] *Then, there exists a Z such that the lossy compression of $(Z, P_{Z|X})$ under log-likelihood loss at $H(Z|X)$ induces a coupling $[W_{XZY}]_{XY}$ which achieves rate-distortion with zero perception error i.e.,*

$$R(D, 0) = \min_{P_{Z|X}:} I(X;Y) \tag{34}$$

$$\text{subject to } \mathbb{E}_{XY}[d(X,Y)] \leq D, \tag{35}$$

$$W_{X,Z,Y} = P_X P_{Z|X} P_{X|Z} \tag{36}$$

Lemma 1. *Let $d(\cdot, \cdot)$ be a squared euclidean distance on a finite subset of $\mathbb{R}^n$ or the Hamming distortion measure. Then, the condition that V is a CP matrix holds.*

VI. Conclusion

We introduced the log-likelihood distortion measure and motivated it for semantic compression. We characterized the basic properties of the resulting rate-distortion function and clarified its connection to classical log-loss as a constrained predictive compression framework. We also showed that, under suitable conditions, this framework recovers several standard rate-distortion settings, including Hamming and squared-error distortion, demonstrating its broader generality. Finally, we connected the framework to rate-distortion-perception theory, suggesting promising directions for finite-blocklength analysis, extensions to more general conditional models, and practical semantic compression schemes.

[†] A matrix A is completely positive if there exists a matrix B with nonnegative entries such that $A = BB^T$ [22].

REFERENCES

[1] A. K. Yadav, D. Song, Y. Shkel, and A. Özgür, "Log-likelihood loss for semantic compression," 2026. [Online]. Available: https://arxiv.org/abs/2601.16461

[2] C. T. Li and A. E. Gamal, "Strong Functional Representation Lemma and Applications to Coding Theorems," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6967–6978, 2018.

[3] D. Song, A. Özgür, and T. Weissman, "A markov property of empirical distributions and the performance of compression-based denoisers," in *2025 IEEE International Symposium on Information Theory (ISIT)*, 2025, pp. 1–6.

[4] D. Reshetova, W.-N. Chen, and A. Özgür, "Training generative models from privatized data via entropic optimal transport," *IEEE Journal on Selected Areas in Information Theory*, vol. 5, pp. 221–235, 2024.

[5] H. Gilbert, M. Sandborn, D. C. Schmidt, J. Spencer-Smith, and J. White, "Semantic compression with large language models," in *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2023, pp. 1–8.

[6] M. Li, R. Jin, L. Xiang, K. Shen, and S. Cui, "Crossword: A semantic approach to text compression via masking," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 9171–9175.

[7] R. Shen, H. Wu, W. Zhang, J. Hu, and D. Gunduz, "Compression beyond pixels: Semantic compression with multimodal foundation models," 2025. [Online]. Available: https://arxiv.org/abs/2509.05925

[8] H. Yu and L. R. Varshney, "Semantic compression with information lattice learning," in *2024 IEEE International Symposium on Information Theory Workshops (ISIT-W)*, 2024, pp. 1–6.

[9] Z. Liang, K. Niu, J. Xu, and P. Zhang, "Semantic arithmetic coding using synonymous mappings," *Entropy*, vol. 27, no. 4, 2025. [Online]. Available: https://www.mdpi.com/1099-4300/27/4/429

[10] Y.-Q. Zhao, Z.-M. Ma, G. Y. Li, S. Yuan, T. Ye, and C. Zhou, "Semantic rate-distortion theory with applications," 2025. [Online]. Available: https://arxiv.org/abs/2509.10061

[11] J. Chai, H. Zhu, Y. Xiao, G. Shi, and P. Zhang, "On the rate-distortion theory for task-specific semantic communication," *Entropy*, vol. 27, no. 8, 2025. [Online]. Available: https://www.mdpi.com/1099-4300/27/8/775

[12] T. Guo, Y. Wang, J. Han, H. Wu, B. Bai, and W. Han, "Semantic compression with side information: A rate-distortion perspective," *arXiv preprint arXiv:2208.06094*, Aug 2022.

[13] J. Chai, Y. Xiao, G. Shi, and W. Saad, "Rate-distortion-perception theory for semantic communication," in *2023 IEEE 31st International Conference on Network Protocols (ICNP)*, 2023, pp. 1–6.

[14] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[15] Y. Shkel, M. Raginsky, and S. Verdú, "Sequential prediction with coded side information under logarithmic loss," in *Proceedings of Algorithmic Learning Theory*, ser. Proceedings of Machine Learning Research, F. Janoos, M. Mohri, and K. Sridharan, Eds., vol. 83. PMLR, 07–09 Apr 2018, pp. 753–769. [Online]. Available: https://proceedings.mlr.press/v83/shkel18a.html

[16] Y. Y. Shkel and S. Verdú, "A single-shot approach to lossy source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 129–147, 2018.

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

[18] R. Gallager, *Information Theory and Reliable Communication*, ser. Courses and lectures. Wiley, 1968. [Online]. Available: https://books.google.com/books?id=Uc3uAAAAMAAJ

[19] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *Institute of Radio Engineers, International Convention Record*, vol. 7, 1959, pp. 325–350.

[20] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 675–685. [Online]. Available: https://proceedings.mlr.press/v97/blau19a.html

[21] J. Chen, L. Yu, J. Wang, W. Shi, Y. Ge, and W. Tong, "On the rate-distortion-perception function," *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 4, pp. 664–673, 2022.

[22] A. Berman and N. Shaked-Monderer, *Completely Positive Matrices*. WORLD SCIENTIFIC, 2003, _eprint: https://www.worldscientific.com/doi/pdf/10.1142/5273. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/5273