Log-Likelihood Loss for Semantic Compression

Anuj K. Yadav — EPFL
Dan Song — Stanford University
Yanina Y. Shkel — EPFL
Ayfer Özgür — Stanford University

Summary and contributions

Log-likelihood loss. We introduce a log-likelihood loss function d_ℓ induced by a likelihood model $P_{X|U}$, and study semantic compression under a rate-distortion formulation.

Structure of $R_\ell(D)$. We characterize the feasible distortion range, identify a special operating point, and compare the resulting trade-off with the log-loss framework.

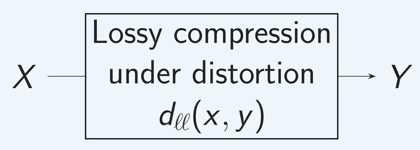
Generalized framework. We show that a broad class of classical rate-distortion problems can be embedded into our log-likelihood based rate-distortion framework.

Applications. Our framework also provides natural constructions for denoising, achieving rate-distortion with perfect perception, and generative modeling from private data.

The Log-likelihood loss

- Information source $X \sim P_X$.
- Conditional law $P_{X|U}$ (where $U \in \mathcal{U}$ is a latent semantic variable).
- The 'log-likelihood loss' $d_\ell(x,y) : \mathcal{X} \times \mathcal{U} \to [0,\infty]$ is:

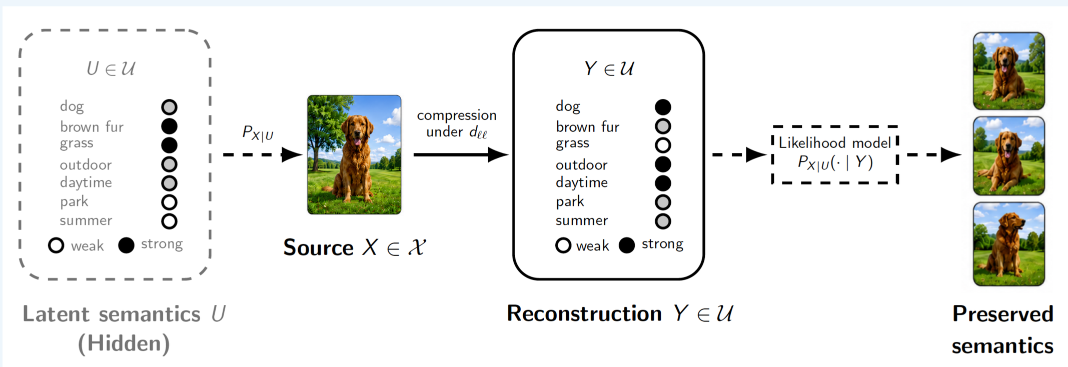
$$d_\ell(x,y) \triangleq \log \frac{1}{P_{X|U}(x|y)}$$



- Reconstruction $Y \in \mathcal{U}$ is a semantic representation of X, rather than a point-wise approximation.

Semantic Compression

- Serves a dual purpose:
  - lossy compression of source X,
  - while preserving semantic (task-relevant) information.



Rate-distortion formulation

- Source $X \in \mathcal{X}$; reconstruction $Y \in \mathcal{U}$
- Given $(X, P_{X|U})$, the rate-distortion function (RDF) under log-likelihood loss is

$$\begin{aligned} R_\ell(D) &= \min_{W_{Y|X}: \mathbb{E}_{XY}[-\log P_{X|U}(x|y)] \le D} I(X;Y) \\ &= \min_{\substack{W_{Y|X}: \\ H(X|Y) + \mathbb{E}_Y[D_{\mathrm{KL}}(W_{X|Y}(\cdot|Y) \| P_{X|U}(\cdot|Y))] \le D}} I(X;Y) \\ &= \min_{\substack{W_{Y|X}: \\ \mathbb{E}_Y[\mathrm{CE}(W_{X|Y}(\cdot|Y) \| P_{X|U}(\cdot|Y))] \le D}} I(X;Y) \end{aligned}$$

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

Properties of $R_\ell(D)$

Given $(X, P_{X|U})$, the RDF $R_\ell(D)$ is defined only for distortion level $D \in [D_{\min}, D_{\max}]$.

Minimum distortion level : $D_{\min}$.

- $D_{\min} = \mathbb{E}_X\left[\min_u \{-\log P_{X|U}(X|u)\}\right]$.
- Best semantic explanation: for every x, choose reconstruction y that maximizes the likelihood of x i.e.,

$$y \in T(x) = \{\arg\max_{u' \in \mathcal{U}} P_{X|U}(x|u')\}$$

- $R_\ell(D_{\min}) = \min_{Q \in \Delta(\mathcal{U})} \mathbb{E}_X\left[-\log \Sigma_{u \in T(X)} Q_U(u)\right]$.
- Optimal reconstruction is a randomized maximum-likelihood (ML) decoder.
- $D_{\min} = 0$ iff $\forall\, x \in \mathcal{X}$, there is a perfect semantic explanation $u \in \mathcal{U}$.
- $D_{\min} = 0 \implies R_\ell(D_{\min}) = H(X)$.

Maximum distortion level : $D_{\max}$.

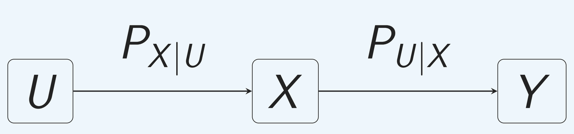
- $D_{\max} = \min_u \mathbb{E}_X\left[-\log P_{X|U}(X|u)\right]$.
- Best global explanation without observing x.
- $R_\ell(D_{\max}) = 0$.

Special operating point. Let $(U,X) \sim P_{U,X}$ be consistent with $(X, P_{X|U})$.

- Consider $D^* \in [D_{\min}, D_{\max}]$ s.t. $D^* = H(X|U)$.
- $R_\ell(D)$ admits the closed-form expression at D^* i.e.,

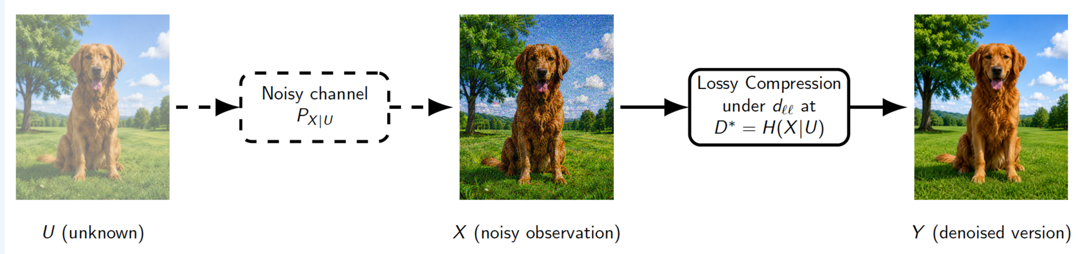
$$R_\ell(D^*) = I(U;X) = H(X) - D^*$$

- Y is a sample from the posterior $P_{U|X}$ i.e., $Y \sim P_U$.



Compression-based denoising

- Unknown clean source: U
- Known noisy channel: $P_{X|U}$
- Noisy observation: X
- Denoising for $(X, P_{X|U})$ via lossy compression.
- Compression under d_ℓ at the special operating point $D^* = H(X|U)$.



- Log-likelihood compression based denoising for stationary ergodic sources [Song et al. '25].

Relation with log-loss framework

- Log-loss: Given source $X \sim P_X$. The decoder outputs $W \in \Delta(\mathcal{X})$ s.t. $d_l(x,W) = -\log W(x)$ and

$$R_l(D) = [H(X) - D]^+.$$

- Log-likelihood loss: The semantic reconstruction y induces a predictive distribution W_y such that $W_y \in \{P_{X|U}(\cdot|u)\}_{u \in \mathcal{U}} \subset \Delta(\mathcal{X})$.
- Can be viewed as a restricted version of the log-loss framework.
- For every $D \in [D_{\min}, D_{\max}]$, we have

$$R_\ell(D) \ge R_l(D).$$

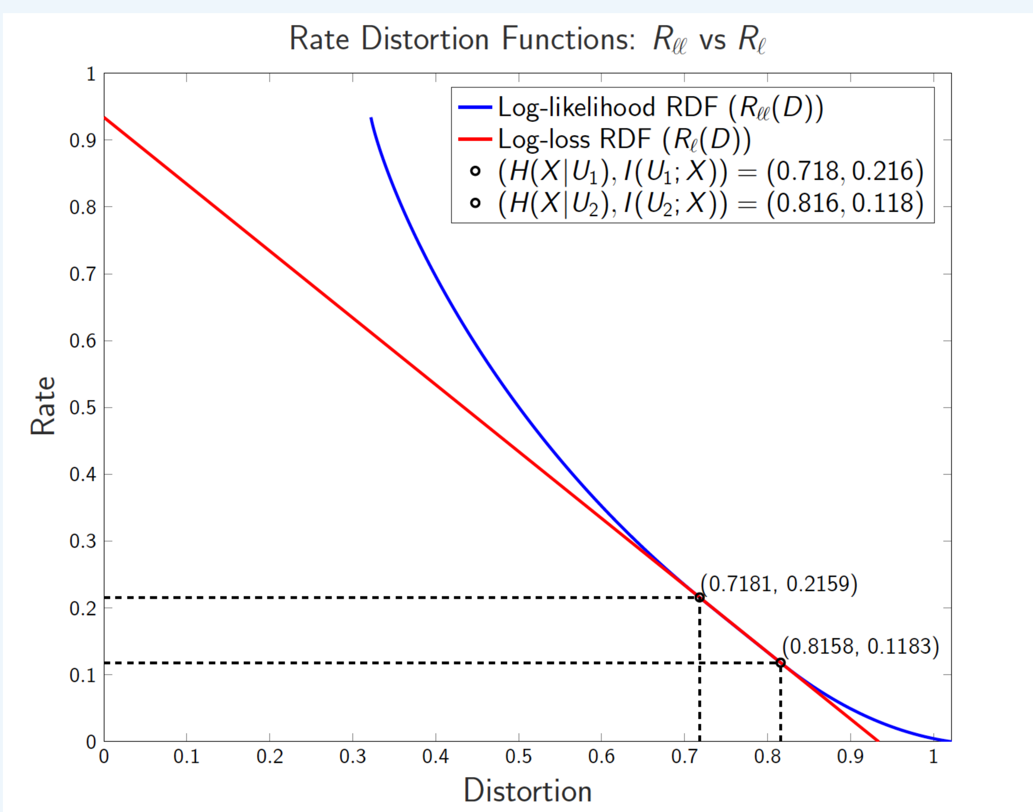
- At every special operating point $D^* \in [D_{\min}, D_{\max}]$, we have

$$R_\ell(D^*) = R_l(D^*) = H(X) - D^*$$

Example. Consider $(X, P_{X|U})$ such that $X \sim \mathrm{Bern}(0.35)$ &

$$P_{X|U} = \begin{bmatrix} 0.8, 0.4, 0.2 \\ 0.2, 0.6, 0.8 \end{bmatrix}.$$

- $D_{\min} = 0.322$ and $D_{\max} = 1.022$.



Generalization of classical RD problems

- Consider a rate-distortion problem (X, d, D).
- Let $\mathcal{J}$ be the set of all $\lambda > 0$ s.t. there exists
  1. a function $\mu(\cdot,\lambda) : \mathcal{X} \to (0,\infty)$,
  2. $P^{(\lambda)}_{X|Y}(x|y) = \mu(x,\lambda)e^{-\lambda d(x,y)}$ is a valid PMF $\forall y \in \mathcal{Y}$. (Cond. A).

If $\mathcal{J} \neq \emptyset$, then, $\forall\, \lambda \in \mathcal{J}$,

$$R_\ell(\tilde{D}) = R(D)$$

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

- Several classical RD problems can be embedded into the log-likelihood framework via an affine transformation of the distortion D.
- Characterizing the RDF in our framework is equally difficult as any classical RD problem.

RDF as a single-parameter optimization

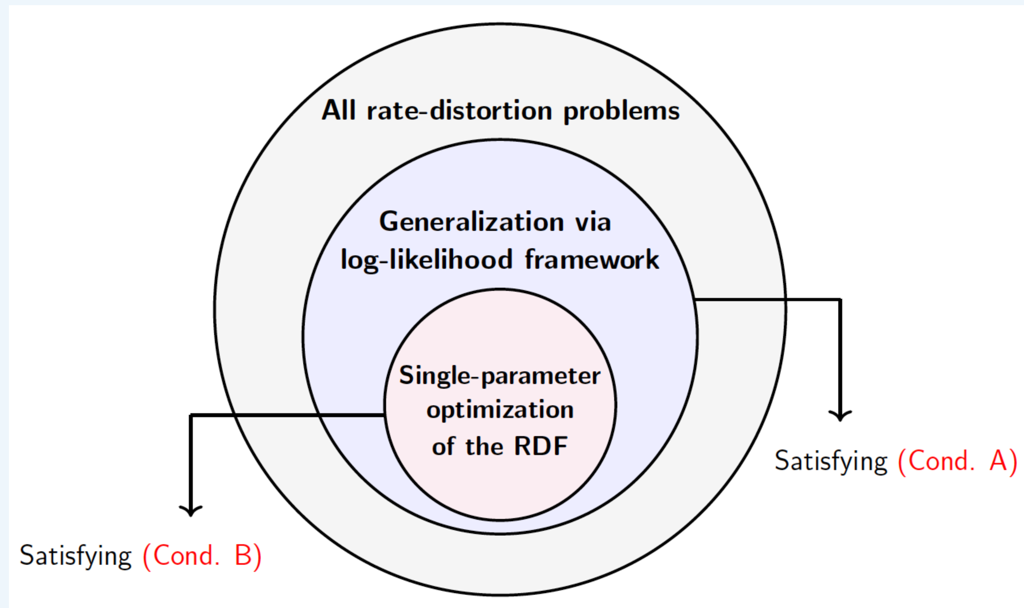
- Consider a rate-distortion problem (X, d, D).
- Let $\mathcal{I}$ be the set of all $\lambda > 0$ s.t. there exists
  1. a function $\mu(\cdot,\lambda) : \mathcal{X} \to (0,\infty)$,
  2. a coupling $P_{Y,X}$ s.t. $\forall\,(y,x)$ $P^{(\lambda)}_{X|Y}(x|y) = \mu(x,\lambda)e^{-\lambda d(x,y)}$. (Cond. B).

If $\mathcal{I} \neq \emptyset$, then

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

- $\mathcal{I} \subset \mathcal{J}$ i.e., (Cond. B) is stronger than the (Cond. A).
- Example: Gaussian source with absolute error - violates (Cond. B).

Classes of RD problems



Rate-distortion with perfect perception

For source $X \sim P_X$, reconstruction $Y \in \mathcal{X}$, distortion function $d : \mathcal{X} \times \mathcal{X} \to [0,\infty]$, and perception measure $\gamma : \Delta(\mathcal{X}) \times \Delta(\mathcal{X}) \to [0,\infty]$, the rate-distortion-perception function is

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

- For $\Gamma = 0$, perfect perception is achieved.
- Log-likelihood based framework can be leveraged to achieve perfect perception.

Log-likelihood construction. Transmit X through a noisy channel $P_{Z|X}$ to obtain Z. Then, compress Z into Y under the log-likelihood framework

$$d_\ell(z,y) = -\log P_{Z|X}(z|y), \quad \text{at} \quad D^* = H(Z|X).$$

- The induced coupling $[W_{XYZ}]_{XY}$ achieves perfect perception.
- For which class of RDP problems (X, d, γ) does this compression framework $(Z, P_{Z|X})$ achieve $R(D,0)$?

Sufficient condition. Given (X, d, γ) with finite $|\mathcal{X}|$, let the element-wise exponential matrix

$$\mathbf{V}(x,y) = e^{-\lambda d(x,y)}.$$

If $\mathbf{V}$ is completely positive for every $\lambda > 0$, i.e.

$$\mathbf{V} = \mathbf{B}\mathbf{B}^T \quad \text{for some } \mathbf{B} \ge 0,$$

then, a log-likelihood compression framework exists that induces an RDP-optimal perfect-perception coupling.

- Examples include squared Euclidean distance on finite subsets of $\mathbb{R}^n$ and q-ary Hamming distortion.

References

1. A. K. Yadav, D. Song, Y. Shkel, A. Özgür, "Log-Likelihood Loss for Semantic Compression," *IEEE ISIT*, 2026.
2. D. Song, A. Özgür, T. Weissman, "A markov property of empirical distributions and the performance of compression-based denoisers," *IEEE ISIT*, 2025.
3. V. Kostina and S. Verdú, "Fixed-Length Lossy Compression in the Finite Blocklength Regime," *IEEE Trans. IT*, 2012.
4. Y. Blau, T. Michaeli, "Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff," *ICML*, 2019.