Log-Likelihood Loss for Semantic Compression

Anuj Yadav
EPFL

Joint work with Dan Song (Stanford University), Yanina Shkel (EPFL), and Ayfer Özgür (Stanford University)

August 11, 2026

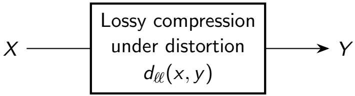
Outline

Outline

Lossy Compression Setup $(X, P_{X|U})$

- 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' $d_{\ell\ell}(x,y) : \mathcal{X} \times \mathcal{U} \to [0,\infty]$ is:

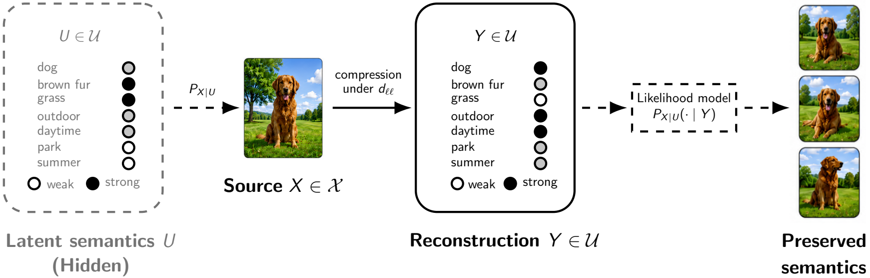
$$d_{\ell\ell}(x,y) \triangleq \log \frac{1}{P_{X|U}(x|y)}$$



- Reconstruction $Y \in \mathcal{U}$ is a semantic representation of X.

Motivation - Semantic Compression

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



Rate-Distortion framework

- 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

$$R_{\ell\ell}(D) = \min_{W_{Y|X}:\mathbb{E}_{XY}[-\log P_{X|U}(x|y)]\leq D} I(X;Y)$$

Rate-Distortion framework

- 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

$$R_{\ell\ell}(D) = \min_{W_{Y|X}:\mathbb{E}_{XY}[-\log P_{X|U}(x|y)]\leq D} I(X;Y)$$

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

Properties of $R_{\ell\ell}(D)$: $D_{\min}$

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

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

Properties of $R_{\ell\ell}(D)$: $D_{\min}$

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

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

- $R_{\ell\ell}(D_{\min}) = \min_{Q \in \Delta(\mathcal{U})} \mathbb{E}_X \left[-\log \sum_{u \in T(X)} Q_U(u)\right]$
- Optimal reconstruction is a randomized maximum-likelihood (ML) decoder.

Properties of $R_{\ell\ell}(D)$: $D_{\min}$

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

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

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

Properties of $R_{\ell\ell}(D)$: $D_{\max}$

- Given $(X, P_{X|U})$, $R_{\ell\ell}(D)$ is defined only for $D \in [D_{\min}, D_{\max}]$.
- $D_{\max} \triangleq \min_u \mathbb{E}_X \left[- \log P_{X|U}(X|u)\right]$.
- Best global explanation without observing x.
- $R_{\ell\ell}(D_{\max}) = 0$.

Properties of $R_{\ell\ell}(D)$: Special operating point D^*

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

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

Properties of $R_{\ell\ell}(D)$: Special operating point D^*

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

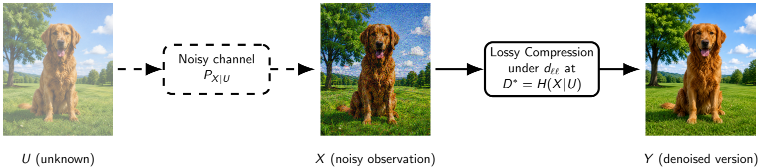
$$R_{\ell\ell}(D^*) = I(U; X) = H(X) - D^*$$

- D^* is the 'special operating point' for $(X, P_{X|U})$.
- Y is a sample from the posterior $P_{U|X}$ i.e., $Y \sim P_U$.

$$\begin{tikzcd} \boxed{U} \arrow[r, "P_{X|U}"] & \boxed{X} \arrow[r, "P_{U|X}"] & \boxed{Y} \end{tikzcd}$$

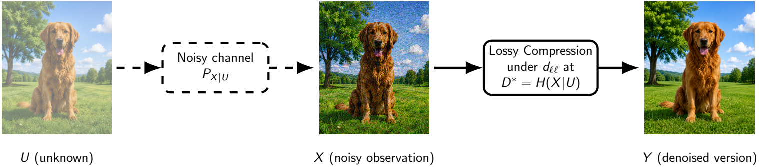
Motivation - Compression based denoising

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



Motivation - 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 $d_{\ell\ell}$ at the special operating point $D^* = H(X|U)$.



- Log-likelihood compression based denoising for stationary ergodic sources [Song, Özgür, Weissman '25].

Outline

Log-loss and Log-likelihood loss

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

$$R_\ell(D) = [H(X) - D]^+.$$

Log-loss and Log-likelihood loss

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

$$R_\ell(D) = [H(X) - D]^+.$$

- Log-likelihood loss: The semantic reconstruction y induces a predictive distribution $W_y \in \{P_{X|U}(\cdot|u)\}_{u \in \mathcal{U}} \subset \Delta(\mathcal{X})$.
- Can be viewed as a restricted version of the log-loss framework.

Log-loss and Log-likelihood loss

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

$$R_\ell(D) = [H(X) - D]^+.$$

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

$$R_{\ell\ell}(D) \geq R_\ell(D).$$

Log-loss and Log-likelihood loss

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

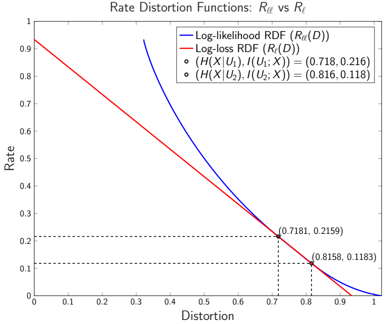
$$R_{\ell\ell}(D^*) = R_{\ell}(D^*) = H(X) - D^*$$

Log-loss and Log-likelihood loss

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

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

- Example: $(X, P_{X|U})$ s.t. $X \sim \mathrm{Bern}(0.35)$ & $P_{X|U} = \begin{bmatrix} 0.8, 0.4, 0.2 \\ 0.2, 0.6, 0.8 \end{bmatrix}$.
- $D_{\min} = 0.322$ and $D_{\max} = 1.022$.



Outline

Translation into Log-likelihood framework

- Consider a rate-distortion problem (X, d, D).

Translation into Log-likelihood framework

- 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).

Translation into Log-likelihood framework

- 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, $\forall\ \lambda \in \mathcal{J}$,

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

  where $\tilde{D} = \lambda D - \mathbb{E}_X[\log(\mu(X, \lambda))]$.
- RDFs are related through an affine transformation of the distortion D.

Translation into Log-likelihood framework

- 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, $\forall\ \lambda \in \mathcal{J}$,

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

where $\tilde{D} = \lambda D - \mathbb{E}_X[\log(\mu(X, \lambda))]$.
- RDFs are related through an affine transformation of the distortion D.
- Our framework generalizes various rate-distortion problems.
- Characterizing the rate-distortion function (RDF) is equally difficult.

RDF as 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).

RDF as 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

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

RDF as 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

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

- Also appears in a different form in [Kostina and Verdú '12].

RDF as 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

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

- Also appears in a different form in [Kostina and Verdú '12].
- $\mathcal{I} \subset \mathcal{J}$ i.e., (Cond. B). is stronger than the (Cond. A).
- Eg., Gaussian source with absolute error - violates (Cond. B).

Outline

The Perception Problem

- Source: $X \sim P_X$
- Lossy reconstruction of X: $Y \in \mathcal{X}$
- Distortion measure : $d : \mathcal{X} \times \mathcal{X} \to [0, \infty]$
- Perception measure: $\gamma : \Delta(\mathcal{X}) \times \Delta(\mathcal{X}) \to [0, \infty]$
- Rate-distortion-perception function for (X, d, γ) is

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

The Perception Problem

- Source: $X \sim P_X$
- Lossy reconstruction of X: $Y \in \mathcal{X}$
- Distortion measure : $d : \mathcal{X} \times \mathcal{X} \to [0, \infty]$
- Perception measure: $\gamma : \Delta(\mathcal{X}) \times \Delta(\mathcal{X}) \to [0, \infty]$
- Rate-distortion-perception function for (X, d, γ) is

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

- For $\Gamma = 0$, perfect perception is achieved.

The Perception Problem

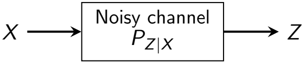
- Source: $X \sim P_X$
- Lossy reconstruction of X: $Y \in \mathcal{X}$
- Distortion measure : $d : \mathcal{X} \times \mathcal{X} \to [0, \infty]$
- Perception measure: $\gamma : \Delta(\mathcal{X}) \times \Delta(\mathcal{X}) \to [0, \infty]$
- Rate-distortion-perception function for (X, d, γ) is

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

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

Achieving perfect perception

- Source: $X \sim P_X$ (U in the log-likelihood framework).
- Choose $P_{Z|X}$ such that



Achieving perfect perception

- Source: $X \sim P_X$ (U in the log-likelihood framework).
- Choose $P_{Z|X}$ such that

$$X \longrightarrow \boxed{\begin{matrix}\text{Noisy channel}\\ P_{Z|X}\end{matrix}} \longrightarrow Z$$

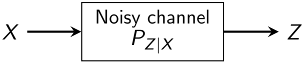
- Compress Z into Y under the log-likelihood loss i.e.,

$$d_{\ell\ell}(z, y) = -\log P_{Z|X}(z|y)$$

  at distortion level $D^* = H(Z|X)$.

Achieving perfect perception

- Source: $X \sim P_X$ (U in the log-likelihood framework).
- Choose $P_{Z|X}$ such that



- Compress Z into Y under the log-likelihood loss i.e.,

$$d_{\ell\ell}(z, y) = -\log P_{Z|X}(z|y)$$

at distortion level $D^* = H(Z|X)$.

- Our framework induces a RDP coupling achieving perfect perception.

Achieving perfect perception

- Source: $X \sim P_X$ (U in the log-likelihood framework).
- Choose $P_{Z|X}$ such that

$$X \longrightarrow \boxed{\begin{array}{c}\text{Noisy channel}\\ P_{Z|X}\end{array}} \longrightarrow Z$$

- Compress Z into Y under the log-likelihood loss i.e.,

$$d_{\ell\ell}(z, y) = -\log P_{Z|X}(z|y)$$

  at distortion level $D^* = H(Z|X)$.
- Our framework induces a RDP coupling achieving perfect perception.
- For which class of RDP problems (X, d, γ) does a log-likelihood based compression framework $(Z, P_{Z|X})$ exist, that achieves $R(D, 0)$?

Achieving perfect perception

- Consider a RDP problem (X, d, γ) such that $|\mathcal{X}| < \infty$.

Achieving perfect perception

- Consider a RDP problem (X, d, γ) such that $|\mathcal{X}| < \infty$.
- Let $\mathbf{V}$ be the element-wise exponential matrix defined as

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

Achieving perfect perception

- Consider a RDP problem (X, d, γ) such that $|\mathcal{X}| < \infty$.
- Let **V** be the element-wise exponential matrix defined as

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

- If **V** is Completely Positive (CP) for every $\lambda > 0$: compression under the log-likelihood framework induces a RDP optimal coupling $[W_{XYZ}]_{XY}$.
- **V** is completely positive if $\mathbf{V} = \mathbf{B}\mathbf{B}^T$ where $\mathbf{B} \geq 0$.

Achieving perfect perception

- Consider a RDP problem (X, d, γ) such that $|\mathcal{X}| < \infty$.
- Let $\mathbf{V}$ be the element-wise exponential matrix defined as

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

- If $\mathbf{V}$ is Completely Positive (CP) for every $\lambda > 0$: compression under the log-likelihood framework induces a RDP optimal coupling $[W_{XYZ}]_{XY}$.
- $\mathbf{V}$ is completely positive if $\mathbf{V} = \mathbf{B}\mathbf{B}^T$ where $\mathbf{B} \geq 0$.
- Examples: Squared Euclidean distance on a finite subset of $\mathbb{R}^n$, q-ary Hamming distortion measure.

Summary

- Log-Likelihood loss based framework for Semantic lossy compression.
- The rate-distortion problem and properties of $R_{\ell\ell}(D)$.
  - $D \in [D_{\min}, D_{\max}]$.
  - At $D^* = H(X|U)$, $R_{\ell\ell}(D) = I(U;X)$.
- Comparison with the log-loss framework : $R_{\ell\ell}(D) \geq R_{\ell}(D)$.
- Translation of existing RD problems into Log-likelihood framework.
- Achieving perfect perception in certain RDP problems via log-likelihood framework.

Summary

- Log-Likelihood loss based framework for Semantic lossy compression.
- The rate-distortion problem and properties of $R_{\ell\ell}(D)$.
  - $D \in [D_{\min}, D_{\max}]$.
  - At $D^* = H(X|U)$, $R_{\ell\ell}(D) = I(U; X)$.
- Comparison with the log-loss framework : $R_{\ell\ell}(D) \geq R_{\ell}(D)$.
- Translation of existing RD problems into Log-likelihood framework.
- Achieving perfect perception in certain RDP problems via log-likelihood framework.

Thank You!