

Introduction to Data Assimilation

Lecture 7

Quanling Deng*

Abstract

This lecture introduces information theory, focusing on Shannon's entropy, relative entropy, and mutual information, with their comparisons to RMSE and pattern correlations.

Keywords Entropy, relative entropy, mutual information, RMSE, Corr

1 Shannon's Entropy and Uncertainty Quantification

1.1 Definition and Intuition

Definition 1.1 (Shannon's Entropy, Discrete Case). Let X be a discrete random variable with probability function $p_i = \mathbb{P}(X = a_i)$ defined over sample space $A = \{a_1, \dots, a_n\}$. The *Shannon entropy* of X is defined as

$$S(p) = - \sum_{i=1}^n p_i \ln p_i, \quad (1.1)$$

where $\sum_j p_j = 1$.

1.1.1 Examples: Discrete Shannon Entropy and Uncertainty

To better illustrate how Shannon's entropy quantifies uncertainty, we consider several simple examples for discrete random variables with finite outcomes.

Example 1.2 (Two-Outcome System (Coin Toss)). Let X be a binary random variable representing a coin toss outcome:

$$X = \begin{cases} H, & \text{with probability } p, \\ T, & \text{with probability } 1-p. \end{cases}$$

Then, from definition (1.1),

$$S(p) = -p \ln p - (1-p) \ln(1-p). \quad (1.2)$$

*Yau Mathematical Science Center, Tsinghua University, Beijing, China 100048. E-mail address: qldeng@tsinghua.edu.cn

Case 1: Deterministic outcome. If $p = 1$ or $p = 0$, then

$$S(1) = S(0) = 0,$$

indicating no uncertainty — the outcome is fully predictable.

Case 2: Fair coin. If $p = \frac{1}{2}$, then

$$S\left(\frac{1}{2}\right) = -2 \times \frac{1}{2} \ln \frac{1}{2} = \ln 2 \approx 0.693.$$

This is the maximum entropy possible for two outcomes, corresponding to maximal uncertainty (completely random outcome).

Remark 1.3. Entropy thus increases as p approaches 0.5 and decreases as p approaches 0 or 1. In this sense, entropy quantifies how unpredictable the random variable is. A fair coin (maximal randomness) carries the greatest uncertainty and hence the largest entropy.

Example 1.4 (Three-Outcome System (Die Roll with Biased Probabilities)). Consider a die-like system with three possible outcomes $\{a_1, a_2, a_3\}$ and probabilities

$$p_1 = 0.5, \quad p_2 = 0.25, \quad p_3 = 0.25.$$

Then the entropy is

$$\begin{aligned} S(p) &= - \sum_{i=1}^3 p_i \ln p_i \\ &= -(0.5 \ln 0.5 + 0.25 \ln 0.25 + 0.25 \ln 0.25) \\ &= -(0.5 \ln 0.5 + 0.5 \ln 0.25) \\ &= -0.5 \ln 0.5 - 0.5 \ln 0.25 \\ &\approx 1.0397. \end{aligned}$$

For comparison, the entropy of a uniform three-outcome system ($p_i = 1/3$) is

$$S_{\max} = -3 \times \frac{1}{3} \ln \frac{1}{3} = \ln 3 \approx 1.099.$$

Hence, biasing toward one outcome ($p_1 = 0.5$) actually reduces the entropy relative to the uniform case.

Remark 1.5. Uniform distributions maximize entropy because they represent complete uncertainty — all outcomes are equally likely. Any deviation from uniformity (bias) introduces predictability and thus lowers entropy. Summary:

- Entropy measures the average uncertainty of a random variable. A smaller entropy implies higher predictability, while larger entropy corresponds to greater disorder or lack of information. Entropy increases with uncertainty and reaches its maximum when all outcomes are equally probable.

- Deterministic systems have zero entropy because there is no randomness.
- Entropy reduction (information gain) quantifies how much new information is learned from observations.

These examples emphasize Shannon's view that information is the *resolution of uncertainty* — the reduction of entropy when new knowledge is obtained.

Definition 1.6 (Shannon's Entropy — Continuous Case). Let X be a continuous random variable with probability density function $\rho(x) \geq 0$ satisfying $\int \rho(x) dx = 1$. The *continuous form of Shannon's entropy* of X is given by

$$S(\rho) = - \int \rho(x) \ln \rho(x) dx. \quad (1.3)$$

Remark 1.7 (Shannon's Story and Intuition). Claude Shannon introduced entropy in 1948 in his groundbreaking paper *A Mathematical Theory of Communication*. His inspiration came from thermodynamics and the concept of entropy in statistical physics. He sought a mathematical measure of uncertainty in communication, and noted that if all N outcomes are equally likely, the uncertainty should increase with N . This led to the logarithmic form $S = \ln N$. He later remarked that John von Neumann suggested the term “entropy” because “no one really knows what entropy is, so in a debate you will always have the advantage.”

1.2 Entropy in the Gaussian Framework

Lemma 1.8 (Entropy of a Multivariate Gaussian Distribution). *Let $\rho(x)$ be an n -dimensional Gaussian distribution $\mathcal{N}(\mu, R)$ with mean vector μ and covariance matrix R . The Shannon entropy is*

$$S(\rho) = \frac{n}{2} \left(1 + \ln 2\pi \right) + \frac{1}{2} \ln \det R. \quad (1.4)$$

Derivation (1D case). For $\rho(x) = \frac{1}{\sqrt{2\pi R}} e^{-(x-\mu)^2/(2R)}$,

$$\begin{aligned} S(\rho) &= - \int \rho(x) \ln \rho(x) dx \\ &= - \int \rho(x) \left[-\frac{1}{2} \ln(2\pi R) - \frac{(x-\mu)^2}{2R} \right] dx \\ &= \frac{1}{2} \ln(2\pi R) + \frac{1}{2R} \int (x-\mu)^2 \rho(x) dx \\ &= \frac{1}{2} \ln(2\pi e R). \end{aligned}$$

The general n -dimensional expression follows directly.

Remark 1.9. Equation (1.4) shows that entropy depends only on the covariance matrix R , not on the mean μ . Thus, entropy measures the uncertainty (or spread) in the system rather than its position in state space. A Gaussian with larger variance (or covariance determinant) carries higher uncertainty.

1.3 RMSE and Its Relation to Entropy

Definition 1.10 (Root Mean Square Error (RMSE)). Given two time series $\{x_i\}$ and $\{y_i\}$ of equal length N , their root mean square error is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - x_i)^2}. \quad (1.5)$$

If $U = y - x$ denotes the residual and $\text{Var}(U) = R = \text{RMSE}^2$, then for Gaussian residuals $U \sim \mathcal{N}(0, R)$, the associated entropy is

$$S(U) = \frac{1}{2} \ln(2\pi e R) = \frac{1}{2} \ln(2\pi e \text{RMSE}^2). \quad (1.6)$$

Thus, entropy and RMSE are closely related — both increase with uncertainty or variability in prediction. However, RMSE measures only second-order moments, while entropy incorporates higher-order information, making it a more general uncertainty measure.

Remark 1.11. In data assimilation and forecasting, RMSE serves as an intuitive “pathwise” accuracy metric, while entropy offers a statistical measure of spread or unpredictability. Together, they form a bridge between deterministic model accuracy and probabilistic uncertainty quantification.

2 Relative Entropy and Its Role in Data Assimilation

2.1 Definition and Properties

Definition 2.1 (Relative Entropy (Kullback–Leibler Divergence)). Let $p(x)$ denote the true probability density (or reference distribution) and $q(x)$ a model or approximating density. The *relative entropy* (or Kullback–Leibler divergence) from q to p is defined as

$$D_{\text{KL}}(p\|q) = \int p(x) \ln \frac{p(x)}{q(x)} dx. \quad (2.1)$$

The Kullback–Leibler divergence satisfies:

- $D_{\text{KL}}(p\|q) \geq 0$ (Gibbs’ inequality), with equality if and only if $p = q$.
- It is asymmetric: $D_{\text{KL}}(p\|q) \neq D_{\text{KL}}(q\|p)$.
- It measures the additional “surprise” or information loss when assuming q instead of the true p .

Remark 2.2 (Information-Theoretic Meaning). Relative entropy quantifies how much information is lost when the model q is used to approximate reality p . It measures the inefficiency of using q -based expectations to describe a system that truly follows p . This quantity connects naturally to Bayesian inference and data assimilation, where we repeatedly update a prior estimate to reduce D_{KL} with respect to the posterior truth.

2.2 Gaussian Framework

Definition 2.3 (KL Divergence between Gaussian Distributions). If $p \sim \mathcal{N}(m_p, R_p)$ and $q \sim \mathcal{N}(m_q, R_q)$ are n -dimensional Gaussian distributions, then

$$D_{\text{KL}}(p\|q) = \frac{1}{2}(m_p - m_q)^T R_q^{-1} (m_p - m_q) + \frac{1}{2}[\text{tr}(R_q^{-1} R_p) - n - \ln \det(R_p R_q^{-1})]. \quad (2.2)$$

The first term represents the mean (bias) mismatch, while the second measures the spread (covariance) mismatch.

Remark 2.4 (Physical Interpretation). Relative entropy in the Gaussian case can be interpreted as the energy difference between two distributions, separating the bias (mean difference) and dispersion (variance difference) components. In climate science, this form has been used to quantify model predictability and information gain. For Gaussian distributions, D_{KL} has a clear geometric meaning in information space — it represents the Mahalanobis distance between the two distributions, augmented by the difference in their “volumes” (determinants of covariance). This makes it a powerful diagnostic for comparing model uncertainty with observational uncertainty in data assimilation.

2.3 Connection with Prior and Posterior in Data Assimilation

In Bayesian data assimilation, we sequentially update a model (prior) state using new observational information to obtain a posterior estimate. Let $p_{\text{prior}}(x)$ denote the prior probability density, $p_{\text{post}}(x)$ the posterior, and $p_{\text{obs}}(y|x)$ the likelihood function associated with observations.

Definition 2.5 (Bayesian Update). The posterior distribution is given by Bayes’ theorem:

$$p_{\text{post}}(x) = \frac{p_{\text{obs}}(y|x) p_{\text{prior}}(x)}{p_{\text{obs}}(y)}, \quad p_{\text{obs}}(y) = \int p_{\text{obs}}(y|x) p_{\text{prior}}(x) dx. \quad (2.3)$$

The *information gain* due to the assimilation of new observations can be expressed in terms of relative entropy as

$$\Delta S = D_{\text{KL}}(p_{\text{post}}\|p_{\text{prior}}) = \int p_{\text{post}}(x) \ln \frac{p_{\text{post}}(x)}{p_{\text{prior}}(x)} dx. \quad (2.4)$$

This measures how much the posterior distribution differs from the prior — i.e., how much information (in bits or nats) the new data have contributed.

Remark 2.6 (Interpretation in Data Assimilation). Equation (2.4) provides a quantitative way to evaluate the efficiency of the assimilation process:

- A large $D_{\text{KL}}(p_{\text{post}}\|p_{\text{prior}})$ indicates that the observation has strongly updated the prior belief — high information gain.
- A small D_{KL} suggests the new data carry little information or are consistent with the prior.

In ensemble-based Kalman filters and variational assimilation, this measure can be used to diagnose how much the ensemble (prior) distribution is tightened by the assimilation step.

Remark 2.7 (Relation to Forecast Skill). In a forecasting context, the relative entropy between the model forecast p_f and a reference (truth) p_t can be interpreted as the amount of “ignorance” in the forecast. Minimizing $D_{\text{KL}}(p_t \| p_f)$ therefore corresponds to optimizing model predictive skill — a concept sometimes referred to as *relative entropy optimization*.

Remark 2.8 (Historical Connection). The use of relative entropy in data assimilation was inspired by its role in statistical mechanics and information theory. Jaynes (1957) proposed that the best estimate of a system’s probability distribution is the one that maximizes entropy subject to known constraints — a principle later generalized in Bayesian inference and DA to *minimizing relative entropy* between posterior and prior. This unifies physical intuition (energy minimization) with statistical learning (information update).

2.4 Summary

- Relative entropy measures the distance between two probability distributions in information space.
- In data assimilation, it quantifies the information gain from assimilating observations:

$$\text{Information Gain} = D_{\text{KL}}(p_{\text{post}} \| p_{\text{prior}}).$$

- It provides a rigorous criterion for evaluating the effectiveness of observation systems, model updates, and ensemble spread reduction.
- It unifies ideas from Bayesian inference, statistical physics, and communication theory in the mathematical foundation of data assimilation.

3 Mutual Information and Its Relation to Correlation

3.1 Definition

Definition 3.1 (Mutual Information). For two random variables X and Y with joint distribution $p(x, y)$ and marginals $p(x)$, $p(y)$, the mutual information is defined as

$$I(X; Y) = \int \int p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} dx dy. \quad (3.1)$$

Equivalently,

$$I(X; Y) = S(X) + S(Y) - S(X, Y), \quad (3.2)$$

where $S(\cdot)$ denotes Shannon’s entropy. It measures the reduction in uncertainty of X given Y , and vice versa.

Remark 3.2. Mutual information quantifies the total (linear and nonlinear) dependence between two variables. It vanishes only if the two are statistically independent.

3.2 Gaussian Framework and Relation to Correlation

Definition 3.3 (Mutual Information for Gaussian Variables). For jointly Gaussian variables with covariance matrices R_x , R_y , and cross-covariance R_{xy} ,

$$I(X;Y) = \frac{1}{2} \ln \frac{\det(R_x) \det(R_y)}{\det(R)}, \quad R = \begin{pmatrix} R_x & R_{xy} \\ R_{yx} & R_y \end{pmatrix}. \quad (3.3)$$

In 1D, using the correlation coefficient $r = R_{xy}/\sqrt{R_x R_y}$,

$$I(X;Y) = -\frac{1}{2} \ln(1 - r^2). \quad (3.4)$$

Definition 3.4 (Pattern Correlation Coefficient). For two datasets $\{x_i\}$ and $\{y_i\}$,

$$\text{Corr} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}. \quad (3.5)$$

In the Gaussian framework, mutual information generalizes correlation — while correlation captures linear dependence, mutual information captures all forms of dependency.

Remark 3.5 (Historical Note). Mutual information was first formalized by Shannon as a measure of the reduction in uncertainty between the input and output of a communication channel. It later became a cornerstone concept across machine learning, neuroscience, and data assimilation, quantifying how much one variable “tells” us about another.

3.3 Why Relative Entropy is Important

Relative entropy underpins both Shannon’s entropy and mutual information:

$$I(X;Y) = D_{\text{KL}}(p(x,y) \| p(x)p(y)). \quad (3.6)$$

Thus, mutual information is the relative entropy between the joint distribution and the product of its marginals — measuring how far the two variables are from being independent.

Remark 3.6. This connection unifies all three measures:

- Entropy: Uncertainty—RMSE
- Relative Entropy: Discrepancy
- Mutual Information: Shared Information—Corr

In practical applications — such as data assimilation, statistical learning, and climate prediction — relative entropy provides a rigorous way to quantify information gain or model error in a probabilistic sense. DA performance shall consider all three skill scores.