The principle of maximum entropy

Adithya C. Ganesh

April 1, 2019

Abstract

This expository paper covers the principle of maximum entropy, a key idea from information theory. We will start by defining the Shannon entropy, which provides a measure of "disorder" or "expected surprise" of a random variable. We will discuss the problem of prior distribution selection in statistics and build towards an analytic argument due to Wallis in 1962 that links maximizing entropy with statistical analysis.

1 Introduction: what is entropy?

Intuitively, the notion of entropy defines a measure of "disorder" or "expected surprise" given a probability distribution, as described by Claude Shannon in his seminal paper *A mathematical theory of communication [7].* This paper covers the principle of maximum entropy, an important tool with applications across various disciplines, including statistics, computer science, and physics.

First, we will define the entropy over probability distributions and explore basic properties. We will motivate the principle of maximum entropy and analyze the entropy of the Gaussian and exponential distributions. Having established these preliminaries, we will discuss the general case of maximizing entropy over moment constraints. Finally, we will discuss an analytic argument due to Wallis in 1962 that links maximizing entropy with statistical analysis.

As described by Shannon in [7], the entropy can be defined as follows.

Definition. *Let X be a discrete random variable on a space $\mathcal{X}$ with probability mass function $\mathbf{p}(x)$. We define the discrete (Shannon) entropy as follows, named after Boltzmann's H-theorem:*

$$H(X) = \mathbb{E}\left[\log \frac{1}{\mathbf{p}(x)}\right] = \sum_{x \in \mathcal{X}} \mathbf{p}(x) \log \frac{1}{\mathbf{p}(x)}$$

Proposition. *The discrete entropy satisfies the following basic properties:*

(a) $H(X) \geq 0$.

(b) $H(X) \leq \log |\mathcal{X}|$, *with equality if and only if X is distributed uniformly over $\mathcal{X}$.*

Proof. Part (a) follows from the fact that $\log \frac{1}{\mathbf{p}(x)} \geq 0$ for any $x \in \mathcal{X}$. Hence the expectation is nonnegative and $H(X) \geq 0$.

To show part (b), we will apply Jensen's inequality (see the appendix for a formal statement). First, note that $f(t) = \log(t)$ is concave. Applying Jensen's inequality, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\log \frac{1}{\mathbf{p}(x)}\right] &\leq \log \mathbb{E}\left[\frac{1}{\mathbf{p}(x)}\right] \\
&= \log \sum_{x \in \mathcal{X}} \mathbf{p}(x) \cdot \frac{1}{\mathbf{p}(x)} \\
&= \log |\mathcal{X}|.
\end{aligned}$$

□

A similar object to study is the so-called *relative entropy*, $D(\mathbf{p}||\mathbf{q})$, which serves as a measure of distance between two probability distributions. Importantly, $D(\mathbf{p}||\mathbf{q}) \neq D(\mathbf{q}||\mathbf{p})$ in general, so the relative entropy is not a metric.

Definition. *Let* $\mathbf{p}$ *and* $\mathbf{q}$ *be probability distributions on a space* $\mathcal{X}$*. Then the relative entropy*[1] $D(\mathbf{p}||\mathbf{q})$ *is defined as*

$$D(\mathbf{p}||\mathbf{q}) = \mathbb{E}_{x\sim\mathbf{p}(x)}\left[\log\frac{\mathbf{p}(x)}{\mathbf{q}(x)}\right] = \sum_{x\in\mathcal{X}}\mathbf{p}(x)\log\frac{\mathbf{p}(x)}{\mathbf{q}(x)}.$$

We will use the following basic fact in subsequent proofs:

Lemma. *The relative entropy is nonnegative. That is,* $D(\mathbf{p}||\mathbf{q}) \geq 0$ *with equality if and only if* $\mathbf{p}(x) = \mathbf{q}(x)$ *for all* $x \in \mathcal{X}$*.*

Proof. Applying Jensen's inequality to the concave function $f(t) = \log(t)$, we obtain:

$$\begin{aligned}
D(\mathbf{p}||\mathbf{q}) &= \mathbb{E}_{x\sim\mathbf{p}(x)}\left[\log\frac{\mathbf{p}(x)}{\mathbf{q}(x)}\right] \\
&= -\mathbb{E}_{x\sim\mathbf{p}(x)}\left[\log\frac{\mathbf{q}(x)}{\mathbf{p}(x)}\right] \\
&\geq -\log\mathbb{E}_{x\sim\mathbf{p}(x)}\left[\frac{\mathbf{q}(x)}{\mathbf{p}(x)}\right] \\
&= -\log\left(\sum_{x\in\mathcal{X}}\mathbf{q}(x)\right) \\
&= \log 1 \\
&= 0.
\end{aligned}$$

Since $f(t)$ is strictly concave, equality can be achieved in Jensen's inequality only when $\frac{\mathbf{q}(x)}{\mathbf{p}(x)}$ is constant. Since $\mathbf{p}$ and $\mathbf{q}$ are probability distributions, they must in fact be equal. □

We can define a version of the entropy for continuous random variables X.

Definition.

$$h(x) = \mathbb{E}\left[\log\frac{1}{\mathbf{p}(X)}\right] = \int_{-\infty}^{\infty}\mathbf{p}(x)\log\frac{1}{\mathbf{p}(x)}.$$

While these objects look similar, in fact the discrete and continuous definitions of entropy are rather different. For instance, while $H(X) \geq 0$, in fact the differential entropy does not satisfy this property.

Proposition. *The differential entropy of a Gaussian random variable* $X \sim \mathcal{N}(0, \sigma^2)$ *is* $\log\sqrt{2\pi e}\sigma$*.*

Proof. Recall that the density function $p(x)$ is given by $p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right)$. Therefore

$$\begin{aligned}
h(X) &= -\int_{-\infty}^{\infty}p(x)\log p(x)\,dx \\
&= \int_{-\infty}^{\infty}p(x)\log\sqrt{2\pi}\sigma\,dx + \int_{-\infty}^{\infty}p(x)\frac{x^2}{2\sigma^2}\,dx \\
&= \log\sqrt{2\pi}\sigma + \frac{\sigma^2}{2\sigma^2} \\
&= \log\sqrt{2\pi e}\sigma.
\end{aligned}$$

[1] This quantity is also known as the Kullback-Leibler (KL) divergence and has numerous applications in statistics and physics.

□

2 Prior probability distributions

Suppose we have some prior knowledge about a phenomenon we would like to observe. Given this knowledge, what is the best prior probability distribution?

The *principle of maximum entropy* states that we ought to choose the prior probability distribution that maximizes the entropy conditioned on our constraints. In particular, this prior will be "maximally disordered" given the constraints; in a technical sense this prior makes the fewest assumptions.

There is significant debate on whether the principle of maximum entropy is the best choice for prior selection in statistics. We do not concern ourselves here with these difficult issues, but we refer the interested reader to [3][4][6].

3 Examples of maximum entropy

We now present three simple examples of the maximum entropy principle.
Theorem. *Suppose* $\mathbf{p}$ *is a discrete probability distribution on a finite set* $\mathcal{X}$*. Then*

$$H(\mathbf{p}) \leq \log |\mathcal{X}|,$$

with equality if and only if $\mathbf{p}$ *is uniform on* $\mathcal{X}$*.*

This was proven in Section 1. Intuitively, this states that with no constraints on a discrete probability distribution, the maximum entropy prior is uniform on the event space $\mathcal{X}$.

Theorem. *Let* $\mathbf{p}$ *be a continuous probability distribution on* $\mathbb{R}$ *with variance* σ^2*. Then*

$$h(\mathbf{p}) \leq \log \sqrt{2\pi e}\sigma.$$

Equality holds if and only if p *is Gaussian with variance* σ^2*.*

Proof. We have already seen that the differential entropy of a Gaussian random variable $G \sim \mathcal{N}(0, \sigma^2)$ is $h(G) = \log \sqrt{2\pi e}\sigma$.

Let $\mathbf{p}$ be a probability density on with variance σ^2 and mean μ (which exists by the definition of variance). Let $\mathbf{q}$ be Gaussian with mean μ and variance σ^2.

Lemma. *Let* $\mathbf{p}, \mathbf{q}$ *be as defined above. Then*

$$\int_{-\infty}^{\infty} \mathbf{q}(x) \log \mathbf{p}(x)\, dx = \int_{-\infty}^{\infty} \mathbf{p}(x) \log \mathbf{p}(x)\, dx.$$

Proof. Since $\mathbf{p}(x)$ is Gaussian, we can write down the density for $\log \mathbf{p}(x)$ easily. Indeed,

$$\log \mathbf{p}(x) = -\log \sqrt{2\pi\sigma^2} - \frac{x^2}{2\sigma^2}.$$

Now, to compute the integral, we can write

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathbf{q}(x) \log \mathbf{p}(x) &= -\log \sqrt{2\pi\sigma^2} \int_{-\infty}^{\infty} \mathbf{q}(x) - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} \mathbf{q}(x) x^2\, dx \\
&= -\log \sqrt{2\pi\sigma^2} - \frac{\sigma^2}{2\sigma^2} \\
&= -\log \sqrt{2\pi e}\sigma,
\end{aligned}$$

which agrees with a computation we did earlier. □

Now, applying this lemma, we can easily prove the main theorem:

$$\begin{aligned} 0 \leq D(\mathbf{q}||\mathbf{p}) &= \int_{-\infty}^{\infty} \mathbf{q}(x) \log \frac{\mathbf{q}(x)}{\mathbf{p}(x)}\, dx \\ &= -h(\mathbf{q}) - \int_{-\infty}^{\infty} \mathbf{q}(x) \log \mathbf{p}(x)\, dx \\ &= -h(\mathbf{q}) - \int_{-\infty}^{\infty} \mathbf{p}(x) \log \mathbf{p}(x)\, dx \\ &= -h(\mathbf{q}) + h(\mathbf{p}), \end{aligned}$$

so $h(\mathbf{q}) \leq h(\mathbf{p})$ as claimed. □

Theorem. *Let* $\mathbf{p}$ *be a continuous probability density function on* $(0, \infty)$ *with mean* μ*. Then*

$$h(p) \leq 1 + \log \mu,$$

with equality if and only if $\mathbf{p}$ *is exponential with mean* μ*. That is,* $p(x) = \frac{1}{\mu} \exp(-\frac{x}{\mu})$.

This theorem has a natural interprettion in physics. Let X be a random variable describing the height of molecules in the atmosphere. The average potential energy of the molecules is fixed (mean λ), and the atmosphere tends to the distribution that has the maximum entropy.

Proof. We will prove this result using Lagrange multipliers. Let $\mathbf{p}$ be a probability distribution on $(0, \infty)$ with mean μ. Define

$$\begin{aligned} F(\mathbf{p}, \lambda_1, \lambda_2) &= -\int_0^{\infty} \mathbf{p}(x) \log \mathbf{p}(x)\, dx + \lambda_1 \left(\int_0^{\infty} \mathbf{p}(x)\, dx - 1 \right) + \lambda_2 \left(\int_0^{\infty} x\mathbf{p}(x)\, dx - \mu \right) \\ &= \int_0^{\infty} \mathcal{L}(x, \mathbf{p}(x), \lambda_1, \lambda_2)\, dx - \lambda_1 - \lambda_2 \mu, \end{aligned}$$

where $\mathcal{L}(x, \mathbf{p}, \lambda_1, \lambda_2) = -\mathbf{p} \log \mathbf{p} + \lambda_1 \mathbf{p} + \lambda_2 x \mathbf{p}$. Taking partials,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{p}} = -1 - \log p + \lambda_1 + \lambda_2 x,$$

and at a maximum entropy distribution we have $\frac{\partial}{\partial \mathbf{p}} = 0$, so that

$$\mathbf{p}(x) = \exp\left(\lambda_1 - 1 + \lambda_2 x\right),$$

for $x \geq 0$.

Now, since $\int_0^{\infty} \mathbf{p}(x)\, dx$ is finite, we must have $\lambda_2 < 0$. This implies

$$\begin{aligned} \int_0^{\infty} \mathbf{p}(x) &= 1 \\ &= e^{\lambda_1 - 1} \int_0^{\infty} e^{\lambda_2 x}\, dx \\ &= \frac{e^{\lambda_1 - 1}}{|\lambda_2|} \end{aligned}$$

so that $e^{\lambda_1 - 1} = |\lambda_2|$.

Now, since $\int_0^{\infty} x e^{\lambda_2 x}\, dx = \frac{1}{\lambda_2^2}$, the condition $\int_0^{\infty} x\, \mathbf{p}(x)\, dx = \mu$, so that $\lambda_2 = -\frac{1}{\mu}$.

Putting this together, we conclude $\mathbf{p}(x) = \frac{1}{\mu} e^{-x/\mu}$, which is indeed the exponential distribution. □

4 Generalizations

Is it possible to obtain a maximum entropy solution under more general constraints?

Problem statement. Maximize the entropy $h(f)$ over all probability densities f satisfying the moment constraints below, where S is the support set.

$$\int_S f(x) r_i(x) dx = \alpha_i; \text{ for } 1 \leq i \leq m. \qquad (*)$$

In particular, f is a density on support set S meeting moment constraints $\alpha_1, \alpha_2, \ldots, \alpha_m$.

Theorem. *Let $f^*(x) = f_\lambda(x) = \exp(\lambda_0 + \sum_{i=1}^m \lambda_i r_i(x))$, $x \in S$, where $\lambda_0, \ldots, \lambda_m$ are chosen so that f^* satisfies the constraints in (*). Then f^* uniquely maximizes $h(f)$ over all probability densities f that satisfy (*).*

Proof. (Sketch).

We first sketch the argument for why the λ_i can be chosen. The constant λ_0 and the n Lagrange multipliers $\lambda_1, \ldots, \lambda_n$ solve the constrained optimization problem below:

$$\max_{\lambda_i} \left\{ \sum_{i=0}^n \lambda_i \alpha_i - \int_S \exp\left(\sum_{i=0}^n \lambda_i f_i(x) \right) \right\}.$$

Under the Karush-Kuhn-Tucker (KKT) conditions, we can show that the optimization problem above has a unique solution, since the objective function is concave in the λ_i. The full argument is out of the scope of this article, but we refer the interested reader to [1].

Now, let g satisfy the constraints in (*). Then

$$\begin{aligned}
h(g) &= -\int_S g \ln g \\
&= -\int_S g \ln \frac{g}{f^*} f^* \\
&= -\mathrm{D}(g||f^*) - \int_S g \ln f^* \\
&\leq -\int_S g \ln f^* \\
&= -\int_S g \left(\lambda_0 + \sum_i \lambda_i r_i \right) \\
&= -\int_S f^* \left(\lambda_0 + \sum_i \lambda_i r_i \right) \\
&= -\int_S f^* \ln f^* \\
&= h(f^*)
\end{aligned}$$

Note that equality holds in (a) if and only if $g(x) = f^*(x)$ for all x, which demonstrates uniqueness. □

Example (Boltzmann's dice [3])**.** *Suppose that n dice are rolled and the total number of spots is $n\alpha$. What proportion of the dice are showing face i where $i \in \{1, 2, \ldots, 6\}$?*

We will count the number of ways that n dice can fall so that n_i dice show face i; this is just the multinomial coefficient $\binom{n}{n_1, n_2, \ldots, n_6}$.

To find the most probable state, we will maximize the multinomial coefficient $\binom{n}{n_1,n_2,\ldots,n_6}$ under the constraint $\sum_{i=1}^{6} in_i = n\alpha$.

A form of Stirling's approximation states $n! \approx \left(\frac{n}{e}\right)^n$. In particular, this implies

$$\begin{aligned}\binom{n}{n_1, n_2, \ldots, n_6} &\approx \frac{\left(\frac{n}{e}\right)^n}{\prod_{i=1}^{6}\left(\frac{n_i}{e}\right)^n} \\ &= \prod_{i=1}^{6}\left(\frac{n}{n_i}\right)^{n_i} \\ &= \exp\left(nH(\frac{n_1}{n}, \frac{n_2}{n}, \cdots, \frac{n_6}{n})\right).\end{aligned}$$

This shows that maximizing $\binom{n}{n_1,n_2,\ldots,n_6}$ under the given constraints is nearly equivalent to maximizing $H(p_1, \ldots, p_6)$ under the constraint $\sum_i ip_i = \alpha$. Applying the theorem, the maximum entropy distribution is

$$p_i^* = \frac{e^{\lambda_i}}{\sum_{i=1}^{6} e^{\lambda_i}},$$

where λ is chosen so that $\sum_i ip_i^* = \alpha$. Returning to the original question, the most probable state is $(np_1^*, np_2^*, \ldots, np_6^*)$ and we expect $n_i^* = np_i^*$ dice showing face i.

5 The Wallis experiment

This experiment is due to Graham Wallis who mentioned it to E.T. Jaynes in 1962 [5].

Suppose Alice is a researcher trying to understand some data. She is studying a phenomenon and wants to estimate a prior probability distribution among m possible outcomes. She has some prior information about what the phenomenon looks like.

To estimate the distribution, she runs the following experiment:

- Randomly distribute N quanta of probability, each worth $\frac{1}{N}$, among the m possibilities.
- Check if the probability assignment is consistent with her prior information. If inconsistent: reject and try again.
- If the assignment agrees with her prior information, her estimated prior distribution $\mathbf{p}$ is given by

$$\mathbf{p}_i = \frac{n_i}{N}; \qquad i \in \{1, 2, \ldots, m\}$$

  where $\mathbf{p}_i$ is the probability of the i-th outcome, and n_i is the number of quanta that were assigned to the i-th proposition.

As we'll see, this experiment has deep ties to the principle of maximum entropy. We now ask: what is the most probable prior distribution Alice will arrive at?

The probability of any particular probability distribution $\mathbf{p}$ is given by a multinomial coefficient.

$$\Pr(\mathbf{p}) = \frac{N!}{n_1!n_2!\cdots n_m!}m^{-N},$$

To find the most likely distribution $\mathbf{p}$, it suffices to maximize the term $A = \frac{N!}{n_1!n_2!\cdots n_m!}$, or a monotonically increasing function of A, e.g. $\frac{1}{N}\log A$.

$$\begin{aligned}
\operatorname{argmax}_{\mathbf{p}} A &= \operatorname{argmax}_{\mathbf{p}} \frac{1}{N}\log A \\
&= \operatorname{argmax}_{\mathbf{p}} \frac{1}{N}\log \frac{N!}{n_1!n_2!\cdots n_m!} \\
&= \operatorname{argmax}_{\mathbf{p}} \frac{1}{N}\log \frac{N!}{(N\mathbf{p}_1)!(N\mathbf{p}_2)!\cdots (N\mathbf{p}_m)!} \\
&= \operatorname{argmax}_{\mathbf{p}} \frac{1}{N}\left(\log N! - \sum_{i=1}^{m} \log(N\mathbf{p}_i)!\right)
\end{aligned}$$

What is the limit of this quantity as the number of trials $N \to \infty$? Applying Stirling's approximation:

$$\begin{aligned}
\operatorname{argmax}_{\mathbf{p}} \lim_{N\to\infty}\left(\frac{1}{N}\log A\right) &= \operatorname{argmax}_{\mathbf{p}} \frac{1}{N}\left(N\log N - \sum_{i=1}^{m} Np_i \log(Np_i)\right) \\
&= \operatorname{argmax}_{\mathbf{p}} \left(\log N - \sum_{i=1}^{m} p_i \log(Np_i)\right) \\
&= \operatorname{argmax}_{\mathbf{p}} \left(-\sum_{i=1}^{m} p_i \log p_i\right) \\
&= \operatorname{argmax}_{\mathbf{p}} H(\mathbf{p}).
\end{aligned}$$

In conclusion, Alice's experiment will most likely converge to the maximum entropy distribution as $N \to \infty$.

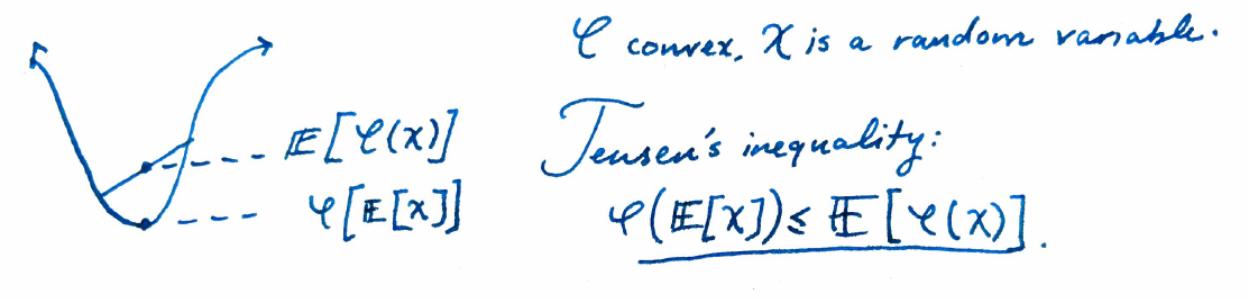
Appendix

Theorem (Jensen's inequality)**.** *Let X be a random variable, and let φ be a convex function. Then*

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}\left[\varphi(X)\right].$$

If φ is strictly convex, equality holds iff X is uniformly distributed.

Graphical intuition. The full proof is technical and out of the scope of this article. The following picture describes the intuition behind this proof.



References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] Keith Conrad. Probability distributions and maximum entropy. *Entropy*, 6(452):10, 2004.

[3] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[4] Edwin T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

[5] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[6] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[7] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.