Introduction to Entropy

Stochastic Processes Seminar

Erik Stern

University of Passau

December 10, 2025

Outline

Entropy and Mutual Information

Definition 1: Entropy

Definition

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{X}$ a countable set, and $X : \Omega \to \mathcal{X}$ a discrete random variable.

$$\textbf{Entropy wrt. base:} \quad H_b(X) = \mathbb{E}\left(-\log_b p_X(X)\right) = -\sum_{x \in \mathrm{supp}(X)} p_X(x) \log_b p_X(x)$$

$$\textbf{Entropy conventionally:} \quad H(X) = H_2(X)$$

Remark 1: Notation Conventions (1/2)

Remark

1. Let $\mathcal{X}, \mathcal{Y}$ be countable sets and $X : \Omega \to \mathcal{X}$, $Y : \Omega \to \mathcal{Y}$ be discrete random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. From now on, the random variables X, Y are always available for use.
2. We do **not** use the shorthand notations $p(x) = \mathbb{P}[X = x]$ and $p(y) = \mathbb{P}[Y = y]$, to keep the notation easily understandable.
3. We use the convention $\log = \log_2$, as the entropy H is defined wrt. base 2.

Remark 1: Notation Conventions (2/2)

Remark

4. We also use the following convention and justify it through a continuity argument:

$$0 \log 0 = \lim_{x \to 0^+} x \log x = \lim_{x \to 0^+} \frac{\log x}{\frac{1}{x}} = \lim_{x \to 0^+} \frac{\frac{1}{\ln(2)x}}{-\frac{1}{x^2}} = \lim_{x \to 0^+} \frac{-x}{\ln(2)} = 0$$

   This choice is sensible, as $\log x$ is not defined for negative x.

5. The conventions, definitions and theorems are from Definitions, Theorems, Remarks and Exercises in *Elements of Information Theory, second edition*.

Remark 2: Existence of Entropy

Remark (Existence of Entropy)

Note that if $|\mathcal{X}|$ is finite, $(\forall p \in \mathbb{R}_+ : H_b(X)$ finite) and $H(X) \leq |\mathcal{X}|$ (see Theorem 5).

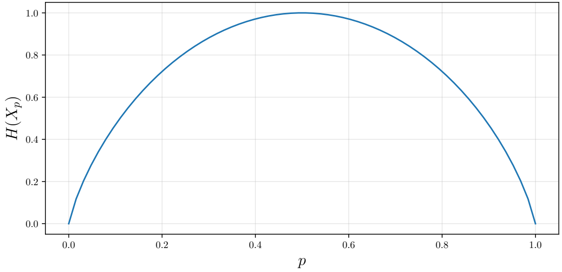
For $|\mathcal{X}|$ countably infinite, there are counterexamples where $H_b(X) = \infty$.

From now on, we will assume that entropy is finite.

Example 1: Entropy of Bernoulli Variable

Example (Entropy of bernoulli variable)

Let $p \in (0, 1)$ and $X_p \sim B(1, p)$ be a weighted coin flip.
We can calculate the Entropy of X_p: $H(X_p) = -p \log p - (1 - p) \log(1 - p)$.
A visual inspection reveals that $H(X_p)$ seems to be maximised for $p = 0.5$ and minimised for $p \in \{0, 1\}$. An increase in uncertainty about the result of the coin flip seems to correspond with an increase in entropy.



Example 2: Entropy of Geometric Variable (1/3)

Example (Entropy of geometric variable)

Let $p \in (0,1)$ and $X_p \sim G(p)$ be the number of times a weighted coin is flipped, until the first head occurs. We will calculate the Entropy of X_p. It will require the two well-known series:

$$\forall r \in (0,1) : \sum_{n \in \mathbb{N}_0} r^n = \frac{1}{1-r} \tag{1}$$

$$\forall r \in (0,1) : \sum_{n \in \mathbb{N}_0} n r^n = \frac{r}{(1-r)^2} \tag{2}$$

Example 2: Entropy of Geometric Variable (2/3)

We can now directly calculate the Entropy of X_p:

$$
\begin{aligned}
H(X_p) &= \sum_{x \in \mathbb{N}} -p(x) \log p(x) && \text{(Def. of Entropy)} \\
&= -\sum_{x \in \mathbb{N}} (1-p)^{x-1} p \log\left((1-p)^{x-1} p\right) && \text{(Geometric mass func.)} \\
&= -\sum_{x \in \mathbb{N}} (1-p)^{x-1} p \left((x-1)\log(1-p) + \log p\right) && \text{(Log rules)} \\
&= -p \log(1-p) \sum_{x \in \mathbb{N}_0} (1-p)^x x - p \log p \sum_{x \in \mathbb{N}_0} (1-p)^x && \text{(Factor out)} \\
&= -p \log(1-p) \frac{1-p}{p^2} - p \log p \frac{1}{p} && \text{(Use series 1, 2)} \\
&= \frac{-(1-p)\log(1-p) - p \log p}{p} && \text{(Simplify)}
\end{aligned}
$$

Example 2: Entropy of Geometric Variable (3/3)

For $p = 0.5$ we get $H(X_{0.5}) = \frac{-0.5 \log 0.5 - 0.5 \log 0.5}{0.5} = -2 \log 0.5 = 2$.

We can visually inspect $(0, 1) \to \mathbb{R}, p \mapsto H(X_p)$ to get a feeling for the entropy of X_p. An increase in p is linked to lower variance and more concentration of the distribution towards zero. Based on the plot, that increase looks to be linked to a lower entropy and vice-versa.

A simple strategy to determine X is to ask "Is $X = 1$?", "Is $X = 2$?" and so on. The number of questions required is exactly X. Thus, the average number of questions is $\mathbb{E}(X_{0.5}) = \frac{1}{0.5} = 2$. This matches $H(X_{0.5}) = 2$.

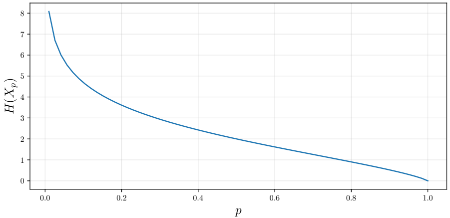


Figure 2: Geometric Entropy $H(X_p)$

Definition 2: Conditional, Joint, Relative Entropy, Mutual Information

Definition

Conditional Entropy: $H(X \mid Y) = -\mathbb{E}\left(\log p_{(X|Y)}(X \mid Y)\right)$

Joint Entropy: $H(X, Y) = -\mathbb{E}\left(\log p_{(X,Y)}(X, Y)\right)$

Relative Entropy: $D(p\|q) = \mathbb{E}_{X\sim p}\left(\log \frac{p(X)}{q(X)}\right)$ (also: KL-Divergence)

Mutual Information: $I(X; Y) = D(p_{(X,Y)}\|p_X\, p_Y)$

where p, q are probability mass functions on the same set $\mathcal{Z}$, with conventions from Remark 3.

Remark 3: Relative Entropy Conventions

Remark

We have $D(p\|q) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$. To understand the conventions, we look at limit cases:

1. Case $p \in (0,1], q = 0$: $\lim_{q\to 0^+} p \log \frac{p}{q} = \lim_{q\to 0^+}(p \log p - p \log q) = \infty$.
2. Case $p = 0, q \in (0,1]$: $0 \log \frac{0}{q} = 0$.
3. Case $p = q = 0$: Case 1 logic yields ∞ and Case 2 logic yields 0.
   As we want $\sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$ to sum over $x \in \mathcal{X}, p(x) > 0$, we choose $0 \log \frac{0}{0} = 0$.

Example 3: Relative Entropy of Binomial Distributions (1/2)

Example

To understand the concept, we can now calculate the relative entropies for an example. Let $X \sim B(20, \alpha)$ and $Y \sim B(20, \beta)$ with $(\alpha, \beta) \in [0, 1]^2$.

$$D(p_X \| p_Y) = \sum_{x=0}^{20} p_X(x) \log \frac{p_X(x)}{p_Y(x)}$$

$$\alpha = 0, \beta = 1: \quad D(p_X \| p_Y) = 1 \log \frac{1}{0} + 0 \log \frac{0}{1} = \infty + 0 = \infty$$

$$\alpha = 0.1, \beta = 0.9: \quad D(p_X \| p_Y) \approx 50.7$$

$$\alpha = 0.3, \beta = 0.7: \quad D(p_X \| p_Y) \approx 9.8$$

$$\alpha = 0.5, \beta = 0.5: \quad D(p_X \| p_Y) = \sum_{x=0}^{20} p(x) \log 1 = 0$$

Example 3: Relative Entropy of Binomial Distributions (2/2)

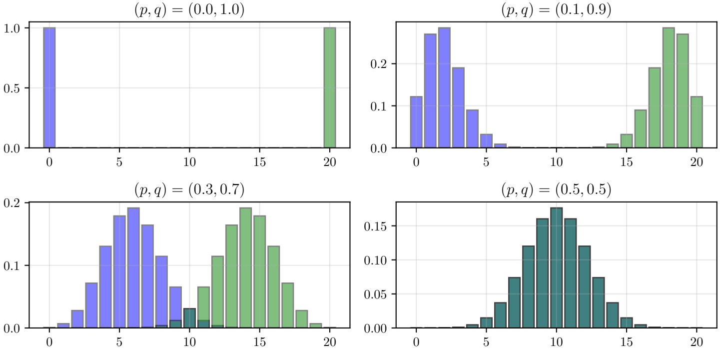


Figure 3: Relative Entropies of Binomial Distributions

Intuitively, the more overlap the distributions have, the closer to zero the relative entropy is.

Mutual Information and Chain Rules

Theorem 1: Chain Rule for Entropy

Theorem (Chain Rule for Entropy)

We have $H(X, Y) = H(X) + H(Y \mid X)$

Proof.

$$
\begin{aligned}
H(X, Y) &= -\mathbb{E}\left(\log p(X, Y)\right) = -\mathbb{E}\left(\log p(Y \mid X)p(X)\right) \\
&= -\mathbb{E}\left(\log p(Y \mid X)\right) - \mathbb{E}\left(\log p(X)\right) = H(X) + H(Y \mid X)
\end{aligned}
$$

□

Theorem 2: Mutual Information Equivalences

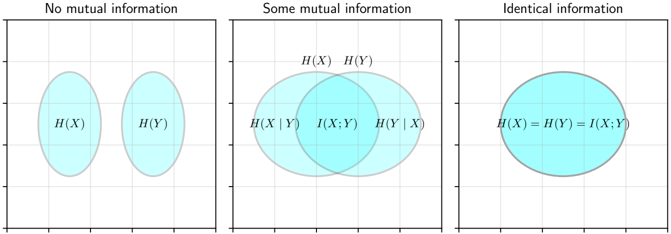


Figure 4: Relationship between Entropy, Conditional Entropy and Mutual Information

Theorem 2: Mutual Information Equivalences (Statement)

Theorem

There are multiple equivalent ways to express Mutual Information:

1. $I(X;Y) = H(Y) - H(Y|X)$
2. $I(X;Y) = I(Y;X)$
3. $I(Y;X) = H(X) - H(X|Y)$
4. $I(X;Y) = H(X) + H(Y) - H(X,Y)$
5. $I(X;X) = H(X)$

Theorem 2: Proof of (1)

Proof of (1).

We can use the definition of mutual information and relative entropy to obtain:

$$
\begin{aligned}
I(X;Y) &= D(p_{(X,Y)} \| p_X p_Y) && \text{(by def. of mutual info.)} \\
&= \mathbb{E}_{p_{(X,Y)}}\left(\log \frac{p_{(X,Y)}(X,Y)}{p_X(X)p_Y(Y)}\right) && \text{(by def. relative entropy)} \\
&= \mathbb{E}_{p_{(X,Y)}}\left(\log \frac{p_X(X)p_{(Y|X)}(Y \mid X)}{p_X(X)p_Y(Y)}\right) && \text{(using cond. probability)} \\
&= \mathbb{E}_{p_{(X,Y)}}\left(\log \frac{p_{(Y|X)}(Y \mid X)}{p_Y(Y)}\right) && \text{(simplify fraction)} \\
&= \mathbb{E}_{p_{(X,Y)}}(\log p_{(Y|X)}(Y \mid X)) - \mathbb{E}_{p_{(X,Y)}}(\log p_Y(Y)) && \text{(simplify logarithm)} \\
&= -H(Y \mid X) + H(Y) && \text{(by def. of entropy)}
\end{aligned}
$$

$\square$

Theorem 3: Chain Rules

Theorem (Chain Rules)

Let $n \in \mathbb{N}, n \geq 2$, $(X_1, \cdots, X_n) \sim p(x_1, \cdots, x_n)$ and Y a random variable. The following statements about Entropy and Mutual Information are called Chain Rules:

1. $H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \cdots, X_1)$
2. $D(p(x) \| q(x)) = D(p(x \mid y) \| q(x \mid y)) + D(p(y) \| q(y))$
3. $I(X_1, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, \cdots, X_1)$

Theorem 3: Proof of (1)

Proof of (1).

Prove this result using induction by n.
Base case $n = 2$:

$$\begin{aligned} H(X_1, X_2) &= -\mathbb{E}\left(\log p(X_1, X_2)\right) \\ &= -\mathbb{E}\left(\log p(X_2 \mid X_1) p(X_1)\right) \\ &= -\mathbb{E}\left(\log p(X_2 \mid X_1)\right) + -\mathbb{E}\left(\log p(X_1)\right) \\ &= H(X_1) + H(X_2 \mid X_1) \end{aligned}$$

Assume the theorem holds for $n - 1$. Induction case $n - 1$ to n:

$$\begin{aligned} H(X_1, \cdots, H_n) &= H(X_n | X_1, \cdots, H_{n-1}) + H(X_1, \cdots, H_{n-1}) && \text{(apply base case)} \\ &= H(X_n | X_1, \cdots, H_{n-1}) + \sum_{i=1}^{n-1} H(X_i \mid X_{i-1}, \cdots, X_1) && \text{(induction hypothesis)} \end{aligned}$$

Inequalities for Entropy and Mutual Information

Remark 4: Convexity

Remark

We use the common definition of convex functions and concave functions from analysis.

Theorem 4: Jensen's Inequality

Theorem (Jensen's Inequality)

Let $\mathcal{X} \subset \mathbb{R}$ and $f : \mathcal{X} \to \mathbb{R}$ a function.

1. *If f is convex, we have $\mathbb{E}f(X) \geq f(\mathbb{E}X)$.*
2. *If f is concave, we have $\mathbb{E}f(X) \leq f(\mathbb{E}X)$.*
3. *If the inequality is strict we have $\mathbb{E}(X) = X$ almost surely.*

Corollary 1: Non-negativity (Statement)

Corollary

Entropy and Mutual Information are non-negative:

1. $0 \leq H(X)$
2. $0 \leq D(p(x) \| q(x))$
3. $0 \leq I(X; Y)$

Corollary 1: Proof of (1)

Proof of (1).

Note that $\log(\frac{1}{[0,1]}) = \log([1, \infty]) = [0, \infty]$ and $p(X)(\mathcal{X}) \in [0, 1]$.
Using the monotonicity of the expected value, we obtain

$$0 \leq \mathbb{E}\left(\log\left(\frac{1}{p(X)}\right)\right) = -\mathbb{E}\left(\log\left(p(X)\right)\right) = H(X)$$

$\square$

Corollary 1: Proof of (2)

Proof of (2).

We can prove this using Jensens Inequality on a *concave* function:

$$
\begin{aligned}
-D(p(x)\|q(x)) &= -\mathbb{E}_p\left(\log\frac{p(X)}{q(X)}\right) = \mathbb{E}_p\left(\log\frac{q(X)}{p(X)}\right) && \text{(def. of relative entropy)}\\
&\leq \log\left(\mathbb{E}_p\frac{q(X)}{p(X)}\right)\\
&= \log\left(\sum_{x\in\mathcal{X}} p(x)\frac{q(x)}{p(x)}\right) = \log\left(\sum_{x\in\mathcal{X}} q(x)\right) && \text{(def. of exp. value, simplify expr.)}\\
&= \log(1) = 0 && \text{(q is a prob. function)}
\end{aligned}
$$

So equivalently, we have $D(p(x)\|q(x)) \geq 0$. □

Corollary 1: Proof of (3)

Proof of (3).

Follows from part (2): $I(X;Y) = D(p(x,y) \| p(x)p(y)) \geq 0$. $\square$

Advanced Properties of Entropy

Remark 5: Natural Questions about Entropy

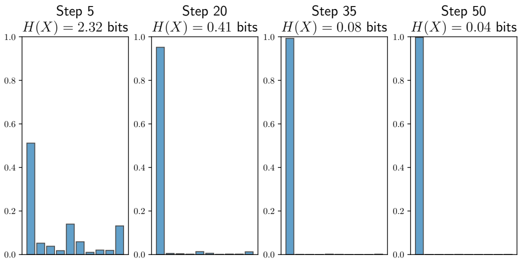


Figure 5: Peak Distribution minimises Entropy

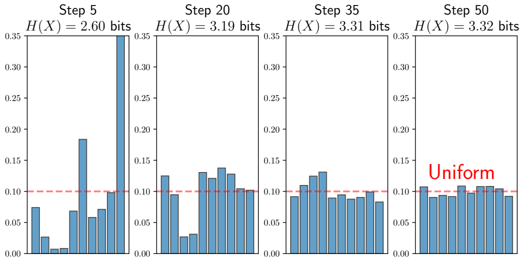


Figure 6: Uniform Distribution maximises Entropy

Remark 5: Question 2 – What distributions extremize entropy?

Remark

Question 2: What distribution minimises and what distribution maximises the value of Entropy?

Let $X_p \sim B(p_1, \cdots, p_n)$ be a categorical distribution. We can write the question as an optimisation problem: $\min_{p \in \mathbb{R}^n, \|p\|_1 = 1, p \geq 0} H(X_p)$.

From the figures, it seems plausible that the uniform distribution maximises entropy and a peaked distribution minimises entropy.

Theorem

- *A peaked distribution minimises entropy:* $H(X) = 0 \iff \exists x \in \mathcal{X} : p_X(x) = 1.$
- *The uniform distribution maximizes entropy:* $H(X) = \log |\mathcal{X}| \iff X \sim U(\mathcal{X}).$

Theorem 5: Proof – Peaked distribution minimises entropy

Proof.

Let X be a discrete random variable with an entropy. Then we have

$$\begin{aligned} H(X) = 0 &\iff \forall x \in \mathcal{X} : p_X(x) \log p_X(x) = 0 \\ &\iff \forall x \in \mathcal{X} : p_X(x) = 0 \oplus p_X(x) = 1 \end{aligned}$$

□

Theorem 5: Proof – Uniform distribution maximises entropy

Proof.

Let $Y \sim U(\mathcal{X})$ st. $\forall x \in \mathcal{X} : q(x) = \frac{1}{|\mathcal{X}|}$.

$$
\begin{aligned}
0 \leq D(p(x) \| q(x)) &= \mathbb{E}_p \left(\log \frac{p(X)}{q(X)} \right) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in \mathcal{X}} p(x) \log \left(p(x) |\mathcal{X}| \right) = \sum_{x \in \mathcal{X}} p(x) \log p(x) + \log |\mathcal{X}| \sum_{x \in \mathcal{X}} p(x) \\
&= -H(X) + \log |\mathcal{X}| = \log |\mathcal{X}| - H(X)
\end{aligned}
$$

This is equivalent to $H(X) \leq \log |\mathcal{X}|$. Lastly, Jensen's Inequality (3) yields the equivalence. □

UNIVERSITÄT PASSAU

The Log-Sum Inequality and Convexity

Theorem 6: Log-Sum Inequality

Theorem (Log-Sum Inequality)

Let $n \in \mathbb{N}$, $a_1, \cdots, a_n, b_1, \cdots, b_n \geq 0$. *Then we have*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

Theorem 7: Convexity of Relative Entropy

Theorem (Convexity of Relative Entropy)

$D(.\|.)$ *is a convex function. This means that*

$$\forall p_1, q_1, p_2, q_2, \forall \lambda \in [0,1]: \quad \begin{aligned} &D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \\ &\leq \lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2) \end{aligned}$$

Theorem 7: Proof

Proof.

Let p_1, q_1, p_2, q_2 be probability mass functions on $\mathcal{X}$. Let $\lambda \in (0,1)$.
First of all, the set of probability densities is convex. So the above statement is well-defined.
Secondly, the inequality needs to be verified:

$$
\begin{aligned}
&D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \\
&= \sum_{x\in\mathcal{X}} (\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} && \text{(by definition)} \\
&\leq \sum_{x\in\mathcal{X}} \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda)p_2(x) \log \frac{p_2(x)}{q_2(x)} && \text{(log-sum inequality)} \\
&= \lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2) && \text{(by definition)}
\end{aligned}
$$

□

Application to Optimisation

Example 5: MNIST Digit Classification

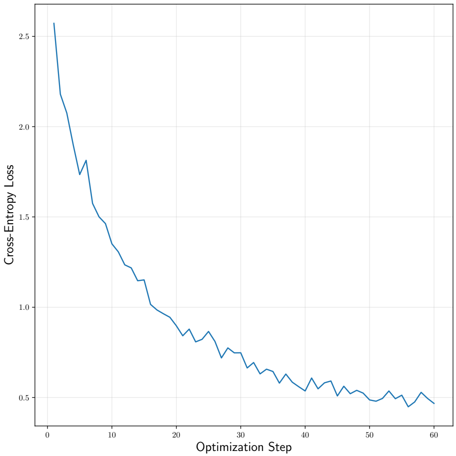


Figure 7: MNIST objective function

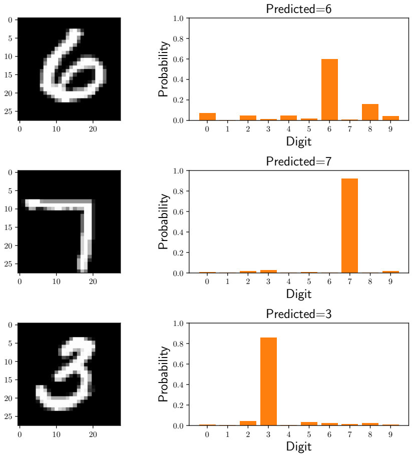


Figure 8: MNIST images and pred. digit distr.

Example 5: MNIST Digit Classification (1/2)

Example (MNIST Digit Classification)

We can now apply the concept of Relative Entropy to solve a common classification problem from machine learning. Let Ω be the set of 28x28 pixel images that contain exactly one handwritten digit from 0 to 9. The task is to predict the digit 0 to 9, based on the input image $\omega \in \Omega$.

In order to accomplish this task, we define a model $f : \mathbb{R}^n \to \mathbb{R}_+^{10}$ with $n \in \mathbb{N}$. This function f takes parameters as inputs that allow it to output a probability mass function that indicates the digit in the image. So $\forall \omega \in \Omega : f(\omega) \geq 0 \wedge \sum_{i=1}^{10} f(\omega)_i = 1$.

Example 5: MNIST Digit Classification (2/2)

The optimisation objective is

$$\Phi = \text{argmin}_{\phi \in \mathbb{R}^n} D(d \| f_\phi)$$

As we have shown in Theorem 7, the Relative Entropy is a convex function.

For f, we can use a two-layer convolutional neural network with dropout. In this case, the model f is not a convex function itself. So the optimisation objective is not convex either. But at least the loss function $D(d \| .)$ is convex. We can now use Stochastic Gradient Descent to optimise for 60 steps with step size 5e-3 and batch size 512.

Conclusion

Summary

- Defined **Entropy**, **Mutual Information**, and **Relative Entropy**
- Established **Chain Rules** for entropy and mutual information
- Proved **Jensen's Inequality** and non-negativity of entropy measures
- Showed which distributions **extremize entropy**
- Proved the **Log-Sum Inequality** and convexity of relative entropy
- Connected theory to practice via **Relative Entropy** loss in neural networks

References

T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, 2006.

Thank you!

Questions?