COMS 4771 Lecture 20

1. Maximum entropy

Probabilistic modeling

Probabilistic modeling

We've encountered many different probability models
(e.g., Gaussian, Bernoulli, Binomial, Poisson, multinomial).

Probabilistic modeling

We've encountered many different probability models
(e.g., Gaussian, Bernoulli, Binomial, Poisson, multinomial).

- Is there a general approach for
  (i) picking a probability model, and
  (ii) parameter estimation?

Probabilistic modeling

We've encountered many different probability models
(e.g., Gaussian, Bernoulli, Binomial, Poisson, multinomial).

- Is there a general approach for
  (i) picking a probability model, and
  (ii) parameter estimation?
- How do familiar models (as above) fit into this approach?

Motivating example

Suppose you want to model a distribution over some discrete set $\mathcal{X}$ (e.g., $\mathcal{X}$ = all English words).

Motivating example

Suppose you want to model a distribution over some discrete set $\mathcal{X}$ (e.g., $\mathcal{X} =$ all English words).

What distribution over $\mathcal{X}$ should you pick?

Motivating example

Suppose you want to model a distribution over some discrete set $\mathcal{X}$ (e.g., $\mathcal{X}$ = all English words).

What distribution over $\mathcal{X}$ should you pick?

A default choice, before making any observations:

$$\text{uniform distribution over } \mathcal{X} \quad (\text{assuming } \mathcal{X} \text{ is finite}).$$

Motivating example

Suppose you want to model a distribution over some discrete set $\mathcal{X}$ (e.g., $\mathcal{X} =$ all English words).

What distribution over $\mathcal{X}$ should you pick?

A default choice, before making any observations:

uniform distribution over $\mathcal{X}$ (assuming $\mathcal{X}$ is finite).

Now you observe a random sample $x_1, x_2, \ldots, x_n$ **from** $\mathcal{X}$, and record some features $T_1, T_2, \ldots, T_k \colon \mathcal{X} \to \mathbb{R}$: e.g.,

- $T_1(x) = \mathbb{1}\{x \text{ ends with an 'e'}\}$
- $T_2(x) = \mathbb{1}\{x \text{ has more than five characters}\}$
- ...

Motivating example

Suppose you want to model a distribution over some discrete set $\mathcal{X}$ (e.g., $\mathcal{X}$ = all English words).

What distribution over $\mathcal{X}$ should you pick?

A default choice, before making any observations:

$$\text{uniform distribution over } \mathcal{X} \quad \text{(assuming } \mathcal{X} \text{ is finite).}$$

Now you observe a random sample $x_1, x_2, \ldots, x_n$ **from** $\mathcal{X}$, and record some features $T_1, T_2, \ldots, T_k \colon \mathcal{X} \to \mathbb{R}$: e.g.,

- $T_1(x) = \mathbb{1}\{x \text{ ends with an 'e'}\}$
- $T_2(x) = \mathbb{1}\{x \text{ has more than five characters}\}$
- ...

Say you observe

$$\frac{1}{n}\sum_{i=1}^{n} T_1(x_i) = 0.22, \quad \frac{1}{n}\sum_{i=1}^{n} T_2(x_i) = 0.32, \quad \ldots$$

Now what distribution should you pick?

A non-committal estimation principle

Idea: Pick the distribution that agrees with the empirical observations, but otherwise is as non-committal as possible . . .

A non-committal estimation principle

Idea: Pick the distribution that agrees with the empirical observations, but otherwise is as non-committal as possible . . .

Distribution should only be pinned down by the observations, but otherwise should express as much "uncertainty" / be as "random" as possible.

A non-committal estimation principle

Idea: Pick the distribution that agrees with the empirical observations, but otherwise is as non-committal as possible . . .

Distribution should only be pinned down by the observations, but otherwise should express as much "uncertainty" / be as "random" as possible.

How do we measure how "random" a distribution is?

Measuring randomness

Measuring randomness

Let X be a discrete $\mathcal{X}$-valued random variable. **How “random” is it?**

Measuring randomness

Let X be a discrete $\mathcal{X}$-valued random variable. **How "random" is it?**

- **Fair coin toss**: one unit of randomness (by definition).

Measuring randomness

Let X be a discrete $\mathcal{X}$-valued random variable. **How "random" is it?**

- **Fair coin toss**: one unit of randomness (by definition).
    - **Biased coin toss**?

Measuring randomness

Let X be a discrete $\mathcal{X}$-valued random variable. **How "random" is it?**

- **Fair coin toss**: one unit of randomness (by definition).
  - **Biased coin toss**? Less randomness.

Measuring randomness

Let X be a discrete $\mathcal{X}$-valued random variable. **How "random" is it?**

- **Fair coin toss**: one unit of randomness (by definition).
  - **Biased coin toss**? Less randomness.
- **Two independent fair coin tosses**?

Measuring randomness

Let X be a discrete $\mathcal{X}$-valued random variable. **How "random" is it?**

- **Fair coin toss**: one unit of randomness (by definition).
  - **Biased coin toss**? Less randomness.
- **Two independent fair coin tosses**? Two units of randomness.

Measuring randomness

Let X be a discrete $\mathcal{X}$-valued random variable. **How "random" is it?**

- **Fair coin toss**: one unit of randomness (by definition).
  - **Biased coin toss**? Less randomness.
- **Two independent fair coin tosses**? Two units of randomness.
  - **Two dependent fair coin tosses**?

Measuring randomness

Let X be a discrete $\mathcal{X}$-valued random variable. **How "random" is it?**

- **Fair coin toss**: one unit of randomness (by definition).
  - **Biased coin toss**? Less randomness.
- **Two independent fair coin tosses**? Two units of randomness.
  - **Two dependent fair coin tosses**? Less randomness.

Measuring randomness

Let X be a discrete $\mathcal{X}$-valued random variable. **How "random" is it?**

- **Fair coin toss**: one unit of randomness (by definition).
  - **Biased coin toss**? Less randomness.
- **Two independent fair coin tosses**? Two units of randomness.
  - **Two dependent fair coin tosses**? Less randomness.
- **Fair 32-sided die**?

Measuring randomness

Let X be a discrete $\mathcal{X}$-valued random variable. **How "random" is it?**

- **Fair coin toss**: one unit of randomness (by definition).
  - **Biased coin toss**? Less randomness.
- **Two independent fair coin tosses**? Two units of randomness.
  - **Two dependent fair coin tosses**? Less randomness.
- **Fair 32-sided die**? This is equivalent to five independent fair coin tosses, so five units of randomness.

Measuring randomness: desiderata

Natural desiderata for a measure of randomness H:

Measuring randomness: desiderata

Natural desiderata for a measure of randomness H:

1. (**Normalization**) If X is a fair coin toss, then

$$H(X) = 1.$$

Measuring randomness: desiderata

Natural desiderata for a measure of randomness H:

1. (**Normalization**) If X is a fair coin toss, then

$$H(X) = 1.$$

2. (**Small for small probs.**) If X is a p-biased coin toss with $p \to 1$, then

$$H(X) \to 0.$$

Measuring randomness: desiderata

Natural desiderata for a measure of randomness H:

1. (**Normalization**) If X is a fair coin toss, then
$$H(X) = 1.$$
2. (**Small for small probs.**) If X is a p-biased coin toss with $p \to 1$, then
$$H(X) \to 0.$$
3. (**Additivity**) If X and Y are independent RVs, and $Z := (X, Y)$, then
$$H(Z) = H(X) + H(Y).$$

Measuring randomness: desiderata

Natural desiderata for a measure of randomness H:

1. (**Normalization**) If X is a fair coin toss, then
$$H(X) = 1.$$
2. (**Small for small probs.**) If X is a p-biased coin toss with $p \to 1$, then
$$H(X) \to 0.$$
3. (**Additivity**) If X and Y are independent RVs, and $Z := (X, Y)$, then
$$H(Z) = H(X) + H(Y).$$
4. (**Subadditivity**) If X and Y are RVs, and $Z := (X, Y)$, then
$$H(Z) \leq H(X) + H(Y).$$

Measuring randomness: desiderata

Natural desiderata for a measure of randomness H:

1. (**Normalization**) If X is a fair coin toss, then
$$H(X) = 1.$$
2. (**Small for small probs.**) If X is a p-biased coin toss with $p \to 1$, then
$$H(X) \to 0.$$
3. (**Additivity**) If X and Y are independent RVs, and $Z := (X, Y)$, then
$$H(Z) = H(X) + H(Y).$$
4. (**Subadditivity**) If X and Y are RVs, and $Z := (X, Y)$, then
$$H(Z) \leq H(X) + H(Y).$$
5. (**Expansibility**) If $X \sim (p_1, p_2, \ldots, p_d)$ and $Y \sim (p_1, p_2, \ldots, p_d, 0)$, then
$$H(X) = H(Y).$$

Measuring randomness: desiderata

Natural desiderata for a measure of randomness H:

1. (**Normalization**) If X is a fair coin toss, then
$$H(X) = 1.$$
2. (**Small for small probs.**) If X is a p-biased coin toss with $p \to 1$, then
$$H(X) \to 0.$$
3. (**Additivity**) If X and Y are independent RVs, and $Z := (X, Y)$, then
$$H(Z) = H(X) + H(Y).$$
4. (**Subadditivity**) If X and Y $\boxed{\text{are RVs,}}$ and $Z := (X, Y)$, then
$$H(Z) \leq H(X) + H(Y).$$
never happen
5. (**Expansibility**) If $X \sim (p_1, p_2, \ldots, p_d)$ and $Y \sim (p_1, p_2, \ldots, p_d, \boxed{0})$, then
$$H(X) = H(Y).$$
6. (**Symmetry**) If $X \sim (p_1, p_2, \ldots, p_d)$ and $Y \sim (p_{\sigma(1)}, p_{\sigma(2)}, \ldots, p_{\sigma(d)})$ for some permutation σ on $\{1, 2, \ldots, d\}$, then
$$H(X) = H(Y).$$

Measuring randomness

The **only measure of randomness that satisfies the desiderata** is

$$H(X) = -\sum_{x \in \mathcal{X}} \Pr(X = x) \log_2 \Pr(X = x)$$

which is called (Shannon) **entropy**. (Note: $0 \log 0 = 0$ by convention.)

Measuring randomness

The **only measure of randomness that satisfies the desiderata** is

$$H(X) \;=\; -\sum_{x\in\mathcal{X}} \Pr(X=x)\log_2 \Pr(X=x)$$

which is called (Shannon) **entropy**. (Note: $0\log 0 = 0$ by convention.)

We equivalently refer to the **entropy of a discrete distribution P over $\mathcal{X}$**:

$$H(P) \;=\; -\sum_{x\in\mathcal{X}} P(x)\log_2 P(x),$$

which is the entropy of a RV $X \sim P$.

Measuring randomness

The **only measure of randomness that satisfies the desiderata** is

$$H(X) \;=\; -\sum_{x \in \mathcal{X}} \Pr(X = x) \log_2 \Pr(X = x)$$

which is called (Shannon) **entropy**. (Note: $0 \log 0 = 0$ by convention.)

We equivalently refer to the **entropy of a discrete distribution P over $\mathcal{X}$**:

$$H(P) \;=\; -\sum_{x \in \mathcal{X}} P(x) \log_2 P(x),$$

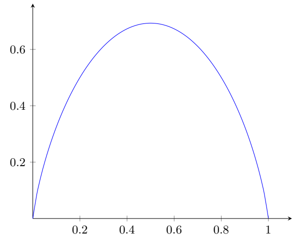
which is the entropy of a RV $X \sim P$.

Also may write this as

$$H(P) \;=\; \mathbb{E}_{X \sim P}\Big[-\log_2(P(X))\Big] \;=\; \mathbb{E}_{X \sim P}\left[\log_2 \frac{1}{P(X)}\right].$$

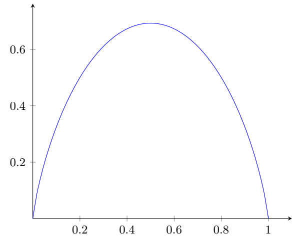
"Bits" $=$ units of entropy with $\log_2$. "Nats" $=$ units of entropy with $\ln$.
Different logarithm bases just change things by constant factors.

Entropy



Entropy $H(P)$ is a concave function of P.

Entropy

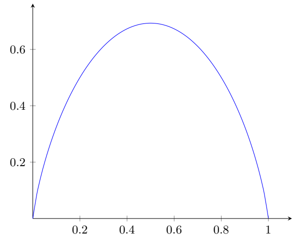


Entropy $H(P)$ is a concave function of P.

Distribution over $\mathcal{X}$ with highest entropy: **uniform distribution**

$$H(\text{uniform}) \;=\; -\sum_{x\in\mathcal{X}} \frac{1}{|\mathcal{X}|}\log\frac{1}{|\mathcal{X}|} \;=\; \log|\mathcal{X}|.$$

Entropy



Entropy $H(P)$ is a concave function of P.

Distribution over $\mathcal{X}$ with highest entropy: **uniform distribution**

$$H(\text{uniform}) \;=\; -\sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \log \frac{1}{|\mathcal{X}|} \;=\; \log |\mathcal{X}|.$$

Distribution over $\mathcal{X}$ with least entropy: **point mass at any** $x^* \in \mathcal{X}$

$$H(\delta_{x^*}) \;=\; -\sum_{x \in \mathcal{X}} \mathbb{1}\{x = x^*\} \log \mathbb{1}\{x = x^*\} \;=\; 0.$$

Interpretations of entropy

Surprise: $\log \frac{1}{P(x)}$ measures the amount of *surprise* you should feel if you observe $x \in \mathcal{X}$, according to the distribution P.

Interpretations of entropy

Surprise: $\log \frac{1}{P(x)}$ measures the amount of *surprise* you should feel if you observe $x \in \mathcal{X}$, according to the distribution P.

$H(P) = \mathbb{E}_{X \sim P}[\log \frac{1}{P(X)}]$ is the average surprise from a random draw from P.

Interpretations of entropy

Surprise: $\log \frac{1}{P(x)}$ measures the amount of *surprise* you should feel if you observe $x \in \mathcal{X}$, according to the distribution P.

$H(P) = \mathbb{E}_{X \sim P}[\log \frac{1}{P(X)}]$ is the average surprise from a random draw from P.

Shannon's source coding theorem: Any lossless compression of an i.i.d. sample from P must use $H(P)$ bits on average.

(This is essentially achieved via Huffman coding.)

Interpretations of entropy

Surprise: $\log \frac{1}{P(x)}$ measures the amount of *surprise* you should feel if you observe $x \in \mathcal{X}$, according to the distribution P.

$H(P) = \mathbb{E}_{X \sim P}[\log \frac{1}{P(X)}]$ is the average surprise from a random draw from P.

Shannon's source coding theorem: Any lossless compression of an i.i.d. sample from P must use $H(P)$ bits on average.

(This is essentially achieved via Huffman coding.)

Upshot: Entropy measures the average information content of a RV.

Asymptotic equipartition property

Let $(X_1, X_2, \ldots, X_n)$ be sequence of i.i.d. $\mathcal{X}$-valued RVs, with $X_i \sim P$.

Asymptotic equipartition property

Let $(X_1, X_2, \ldots, X_n)$ be sequence of i.i.d. $\mathcal{X}$-valued RVs, with $X_i \sim P$.

Asymptotic equipartition property:
For large n, we can divide all sequences in $\mathcal{X}^n$ into two sets:

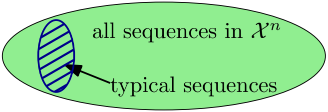
ASYMPTOTIC EQUIPARTITION PROPERTY

Let $(X_1, X_2, \ldots, X_n)$ be sequence of i.i.d. $\mathcal{X}$-valued RVs, with $X_i \sim P$.

Asymptotic equipartition property:
For large n, we can divide all sequences in $\mathcal{X}^n$ into two sets:

1. "**Typical sequences**":
   $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with

   $$P(x_1, x_2, \ldots, x_n) \approx 2^{-n(H(P) \pm \varepsilon)}.$$



2. **All other sequences.**

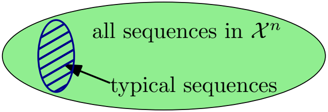
Asymptotic equipartition property

Let $(X_1, X_2, \ldots, X_n)$ be sequence of i.i.d. $\mathcal{X}$-valued RVs, with $X_i \sim P$.

Asymptotic equipartition property:
For large n, we can divide all sequences in $\mathcal{X}^n$ into two sets:

1. "**Typical sequences**":
   $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with

   $$P(x_1, x_2, \ldots, x_n) \approx 2^{-n(H(P)\pm\varepsilon)}.$$



2. **All other sequences.**

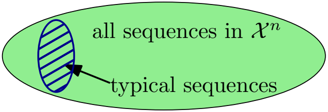
- **Typical sequences account for almost all the probability mass.**

ASYMPTOTIC EQUIPARTITION PROPERTY

Let $(X_1, X_2, \ldots, X_n)$ be sequence of i.i.d. $\mathcal{X}$-valued RVs, with $X_i \sim P$.

Asymptotic equipartition property:
For large n, we can divide all sequences in $\mathcal{X}^n$ into two sets:

1. "**Typical sequences**": ? ? ?
   $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with
   $$P(x_1, x_2, \ldots, x_n) \approx 2^{-n(H(P) \pm \varepsilon)}.$$



2. **All other sequences.**

- **Typical sequences account for almost all the probability mass.**
- **Number of typical sequences**:

$$\text{Between } (1-\varepsilon) 2^{n(H(P)-\varepsilon)} \text{ and } 2^{n(H(P)+\varepsilon)}.$$

Far fewer than $|\mathcal{X}|^n$ when $H(P) \ll \log_2 |\mathcal{X}|$.

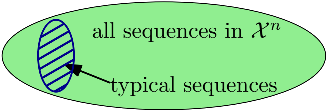
Asymptotic equipartition property

Let $(X_1, X_2, \ldots, X_n)$ be sequence of i.i.d. $\mathcal{X}$-valued RVs, with $X_i \sim P$.

Asymptotic equipartition property:
For large n, we can divide all sequences in $\mathcal{X}^n$ into two sets:

1. "**Typical sequences**":
   $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with

   $$P(x_1, x_2, \ldots, x_n) \approx 2^{-n(H(P)\pm\varepsilon)}.$$



2. **All other sequences.**

- **Typical sequences account for almost all the probability mass.**
- **Number of typical sequences**:

  Between $(1-\varepsilon)2^{n(H(P)-\varepsilon)}$ and $2^{n(H(P)+\varepsilon)}$.

  Far fewer than $|\mathcal{X}|^n$ when $H(P) \ll \log_2 |\mathcal{X}|$.

the larger the H(P), the larger the set of typical sequences

extract typical seqeunce

Upshot: $H(P)$ characterizes the number of typical i.i.d. sequences from P.

Entropy: recap

- Entropy is a fundamental measure of the
    - randomness
    - uncertainty
    - information content

  in a probability distribution.
- Quantifies achievable rates for data compression.
- Quantifies number of typical i.i.d. sequences.
- ...

Maximum entropy principle

Maximum entropy principle

Observe a random sample $x_1, x_2, \ldots, x_n$ **of words from** $\mathcal{X}$,
and record some features $T_1, T_2, \ldots, T_k \colon \mathcal{X} \to \mathbb{R}$: e.g.,

- $T_1(x) = \mathbb{1}\{x \text{ ends with an 'e'}\}$
- $T_2(x) = \mathbb{1}\{x \text{ has more than five characters}\}$
- ...

Observations:

$$\frac{1}{n}\sum_{i=1}^{n} T_1(x_i) = 0.22, \quad \frac{1}{n}\sum_{i=1}^{n} T_2(x_i) = 0.32, \quad \ldots$$

Maximum entropy principle

Observe a random sample $x_1, x_2, \ldots, x_n$ of words from $\mathcal{X}$,
and record some features $T_1, T_2, \ldots, T_k \colon \mathcal{X} \to \mathbb{R}$: e.g.,

- $T_1(x) = \mathbb{1}\{x \text{ ends with an 'e'}\}$
- $T_2(x) = \mathbb{1}\{x \text{ has more than five characters}\}$
- ...

more general way in depict whether a sample meets certain condition

Observations:

$$\frac{1}{n}\sum_{i=1}^{n} T_1(x_i) = 0.22, \quad \frac{1}{n}\sum_{i=1}^{n} T_2(x_i) = 0.32, \quad \ldots$$

Maximum entropy principle (Jaynes, 1957):

Pick the distribution that agrees with the empirical observations, but is otherwise as "random" as possible.

Our empirical observations from sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$:

Text

$$b_i \ := \ \frac{1}{n} \sum_{j=1}^{n} T_i(x_j) \ = \ \widehat{\mathbb{E}}[T_i(X)] \quad \text{for } i = 1, 2, \ldots, k$$

where $\widehat{\mathbb{E}}[\,\cdot\,]$ is expectation w.r.t. *empirical distribution based on the sample*.

Maximum entropy optimization problem

Our empirical observations from sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$:

$$b_i \ := \ \frac{1}{n} \sum_{j=1}^{n} T_i(x_j) \ = \ \widehat{\mathbb{E}}[T_i(X)] \quad \text{for } i = 1, 2, \ldots, k$$

where $\widehat{\mathbb{E}}[\,\cdot\,]$ is expectation w.r.t. *empirical distribution based on the sample*.

Maximum entropy optimization problem:

$$\begin{aligned} \max_{P \in \Delta(\mathcal{X})} \quad & H(P) \\ \text{s.t.} \quad & \mathbb{E}_{X \sim P}[T_i(X)] \ = \ b_i \quad \text{for all } i = 1, 2, \ldots, k \end{aligned}$$

(where $\Delta(\mathcal{X})$ is the space of probability distributions over $\mathcal{X}$).

Maximum entropy optimization problem

just for constrain purpose, has nothing to do with entropy,
which is the target!!

Our empirical observations from sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$:

$$b_i \; := \; \frac{1}{n}\sum_{j=1}^{n} T_i(x_j) \; = \; \widehat{\mathbb{E}}[T_i(X)] \quad \text{for } i = 1, 2, \ldots, k$$

the expectation
of Ti

X is random variable

where $\widehat{\mathbb{E}}[\,\cdot\,]$ is expectation w.r.t. *empirical distribution based on the sample*.

get the right P
through all possible
distributions

Maximum entropy optimization problem:

$$\begin{aligned} \max_{P \in \Delta(\mathcal{X})} \quad & H(P) \\ \text{s.t.} \quad & \mathbb{E}_{X \sim P}[T_i(X)] \; = \; b_i \quad \text{for all } i = 1, 2, \ldots, k \end{aligned}$$

must meet
expirical
distribution

(where $\Delta(\mathcal{X})$ is the space of probability distributions over $\mathcal{X}$).

Without the constraints (i.e., before observations are made), $\max_{P \in \Delta(\mathcal{X})} H(P)$ is achieved by the *uniform distribution* over $\mathcal{X}$.

Non-uniform base distributions

If $\mathcal{X}$ is discrete but infinite (e.g., $\mathcal{X} = \mathbb{N}$), **no uniform distribution over $\mathcal{X}$**.

Non-uniform base distributions

If $\mathcal{X}$ is discrete but infinite (e.g., $\mathcal{X} = \mathbb{N}$), **no uniform distribution over $\mathcal{X}$**.

Instead, consider different *default (base) distribution* π over $\mathcal{X}$ before making any observations.

Non-uniform base distributions

If $\mathcal{X}$ is discrete but infinite (e.g., $\mathcal{X} = \mathbb{N}$), **no uniform distribution over $\mathcal{X}$**.

Instead, consider different *default (base) distribution* π over $\mathcal{X}$ before making any observations.

Generalization of maximum entropy principle:

> *Pick the distribution that agrees with the empirical observations, but is otherwise as close to π as possible.*

Want $\pi =$ uniform $\Longrightarrow$ maximum entropy.

Non-uniform base distributions

If $\mathcal{X}$ is discrete but infinite (e.g., $\mathcal{X} = \mathbb{N}$), **no uniform distribution over $\mathcal{X}$**.

Instead, consider different *default (base) distribution π* over $\mathcal{X}$ before making any observations.

Generalization of maximum entropy principle:

> *Pick the distribution that agrees with the empirical observations, but is otherwise as close to π as possible.*

Want $\pi = \text{uniform} \implies$ maximum entropy.

How do we measure how close two probability distributions are?

Relative entropy

Entropy: expected information content measured by P, where expectation is w.r.t. random draw from P.

$$H(P) \;=\; \mathbb{E}_{X\sim P}\left[\ln \frac{1}{P(X)}\right].$$

Relative entropy

Entropy: expected information content measured by P, where expectation is w.r.t. random draw from P.

$$H(P) \;=\; \mathbb{E}_{X\sim P}\left[\ln \frac{1}{P(X)}\right].$$

Relative entropy: expected information content measured by Q, where expectation is w.r.t. random draw from P

$$\mathrm{RE}(P\|Q) \;:=\; \mathbb{E}_{X\sim P}\left[\ln \frac{1}{Q(X)}\right] - H(P).$$

(and we subtract off $H(P)$ so it is zero when $P = Q$).

Relative entropy

Entropy: expected information content measured by P, where expectation is w.r.t. random draw from P.

$$H(P) \;=\; \mathbb{E}_{X\sim P}\left[\ln \frac{1}{P(X)}\right].$$

Relative entropy: expected information $\boxed{\text{content measured by } Q,}$ where expectation is w.r.t. random draw from P

$$\mathrm{RE}(P\|Q) \;:=\; \mathbb{E}_{X\sim P}\left[\ln \frac{1}{Q(X)}\right] - H(P).$$

use Q to measure(compare) with P. all random draw from P

(and we subtract off $H(P)$ so it is zero when $P = Q$).

More typical form:

$$\mathrm{RE}(P\|Q) \;=\; \sum_{x\in\mathcal{X}} P(x)\ln\frac{P(x)}{Q(x)}.$$

note the expansion

Properties of relative entropy

$$\mathrm{RE}(P \| Q) \;=\; \sum_{x \in \mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)}.$$

Properties of relative entropy

$$\mathrm{RE}(P\|Q) \;=\; \sum_{x\in\mathcal{X}} P(x)\ln\frac{P(x)}{Q(x)}.$$

- $\mathrm{RE}(P\|Q) \geq 0$ for all P and Q.
  $\mathrm{RE}(P\|Q) = 0$ if and only if $P = Q$.

Properties of relative entropy

$$\mathrm{RE}(P \| Q) \;=\; \sum_{x \in \mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)}.$$

- $\mathrm{RE}(P\|Q) \geq 0$ for all P and Q.
  $\mathrm{RE}(P\|Q) = 0$ if and only if $P = Q$.
- $\mathrm{RE}(Q\|P) \neq \mathrm{RE}(P\|Q)$ in general, and triangle inequality does not hold.
  So RE is **not** a metric.

Properties of relative entropy

$$\mathrm{RE}(P\|Q) \;=\; \sum_{x\in\mathcal{X}} P(x)\ln\frac{P(x)}{Q(x)}.$$

- $\mathrm{RE}(P\|Q) \geq 0$ for all P and Q.
  $\mathrm{RE}(P\|Q) = 0$ if and only if $P = Q$.
- $\mathrm{RE}(Q\|P) \neq \mathrm{RE}(P\|Q)$ in general, and triangle inequality does not hold.
  So RE is **not** a metric.
- $\mathrm{RE}(P\|\text{uniform}) = \ln|\mathcal{X}| - H(P)$.

Properties of relative entropy

$$\mathrm{RE}(P\|Q) \;=\; \sum_{x\in\mathcal{X}} P(x)\ln\frac{P(x)}{Q(x)}.$$

- $\mathrm{RE}(P\|Q) \geq 0$ for all P and Q.
  $\mathrm{RE}(P\|Q) = 0$ if and only if $P = Q$.
- $\mathrm{RE}(Q\|P) \neq \mathrm{RE}(P\|Q)$ in general, and triangle inequality does not hold. So RE is **not** a metric.
- $\mathrm{RE}(P\|\text{uniform}) = \boxed{\ln|\mathcal{X}| - H(P).}$
- $\mathrm{RE}(P\|Q)$ is a *convex function* of (P, Q) (and hence also a convex function of P, by itself, and also of Q).

Properties of relative entropy

$$\mathrm{RE}(P \| Q) = \sum_{x \in \mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)}.$$

- $\mathrm{RE}(P\|Q) \geq 0$ for all P and Q.
  $\mathrm{RE}(P\|Q) = 0$ if and only if $P = Q$.
- $\mathrm{RE}(Q\|P) \neq \mathrm{RE}(P\|Q)$ in general, and triangle inequality does not hold.
  So RE is **not** a metric.

the larger the less randomness.
we don't care about constrain at this time

- $\mathrm{RE}(P\|\text{uniform}) = \ln|\mathcal{X}| - H(P)$.
- $\mathrm{RE}(P\|Q)$ is a *convex function* of (P, Q)
  (and hence also a convex function of P, by itself, and also of Q).
- Also called "Kullback-Leibler divergence".

Maximum entropy optimization problem

Maximum entropy optimization problem with base distribution π:

$$\begin{aligned} \min_{P \in \Delta(\mathcal{X})} \quad & \mathrm{RE}(P \| \pi) \\ \text{s.t.} \quad & \mathbb{E}_{X \sim P}[T_i(X)] \; = \; b_i \quad \text{for all } i = 1, 2, \ldots, k. \end{aligned}$$

Maximum entropy optimization problem

Maximum entropy optimization problem with base distribution π:

$$\begin{array}{rll} \min\limits_{P \in \Delta(\mathcal{X})} & \mathrm{RE}(P \| \pi) & \\ \text{s.t.} & \mathbb{E}_{X \sim P}[T_i(X)] \; = \; b_i & \text{for all } i = 1, 2, \ldots, k. \end{array}$$

(If $\pi =$ uniform, then objective is $\mathrm{RE}(P \| \pi) = \ln |\mathcal{X}| - H(P)$.)

Maximum entropy optimization problem

Maximum entropy optimization problem with base distribution π:

$$\begin{aligned} \min_{P \in \Delta(\mathcal{X})} \quad & \mathrm{RE}(P \| \pi) \\ \text{s.t.} \quad & \mathbb{E}_{X \sim P}[T_i(X)] \; = \; b_i \quad \text{for all } i = 1, 2, \dots, k. \end{aligned}$$

(If $\pi =$ uniform, then objective is $\mathrm{RE}(P \| \pi) = \ln |\mathcal{X}| - H(P)$.)

More explicitly, with $\boldsymbol{T}(x) := (T_1(x), T_2(x), \dots, T_k(x))$ and $\boldsymbol{b} := (b_1, b_2, \dots, b_k)$,

$$\begin{aligned} \min_{P \in \mathbb{R}^{\mathcal{X}}} \quad & \sum_{x \in \mathcal{X}} P(x) \ln \frac{P(x)}{\pi(x)} \\ \text{s.t.} \quad & \sum_{x \in \mathcal{X}} P(x) \boldsymbol{T}(x) \; = \; \boldsymbol{b} \\ & P(x) \geq 0 \quad \text{for all } x \in \mathcal{X} \\ & \sum_{x \in \mathcal{X}} P(x) = 1. \end{aligned}$$

Maximum entropy optimization problem

Maximum entropy optimization problem with base distribution π:

$$\begin{aligned} \min_{P \in \Delta(\mathcal{X})} \quad & \mathrm{RE}(P \| \pi) \\ \text{s.t.} \quad & \mathbb{E}_{X \sim P}[T_i(X)] = b_i \quad \text{for all } i = 1, 2, \ldots, k. \end{aligned}$$

both P(x) and Q(x) are kinds of probability arragement(distribution)

(If $\pi =$ uniform, then objective is $\mathrm{RE}(P \| \pi) = \ln |\mathcal{X}| - H(P)$.)

More explicitly, with $\boldsymbol{T}(x) := (T_1(x), T_2(x), \ldots, T_k(x))$ and $\boldsymbol{b} := (b_1, b_2, \ldots, b_k)$,

try to minimize

$$\begin{aligned} \min_{P \in \mathbb{R}^{\mathcal{X}}} \quad & \sum_{x \in \mathcal{X}} P(x) \ln \frac{P(x)}{\pi(x)} \\ \text{s.t.} \quad & \boxed{\sum_{x \in \mathcal{X}} P(x) \boldsymbol{T}(x) = \boldsymbol{b}} \\ & P(x) \geq 0 \quad \text{for all } x \in \mathcal{X} \\ & \sum_{x \in \mathcal{X}} P(x) = 1. \end{aligned}$$

for given x, each of its T(x) is meet with related b. (a list of b, with imposed feature requirement)

Convex objective function, with **linear** (in)equality constraints.

Entropy projection

Note that *any* feasible solution P must satisfy

$$\sum_{x \in \mathcal{X}} P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

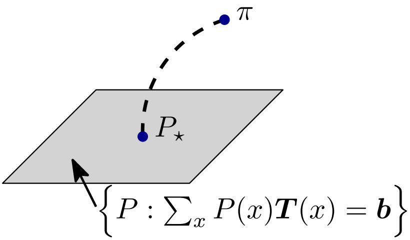
Entropy projection

Note that *any* feasible solution P must satisfy

$$\sum_{x \in \mathcal{X}} P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

These constraints define an affine hyperplane in $\mathbb{R}^{\mathcal{X}}$.

$$\min_{P \in \Delta(\mathcal{X})} \quad \mathrm{RE}(P \| \pi)$$
$$\text{s.t.} \quad \sum_{x \in \mathcal{X}} P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}$$



Entropy projection

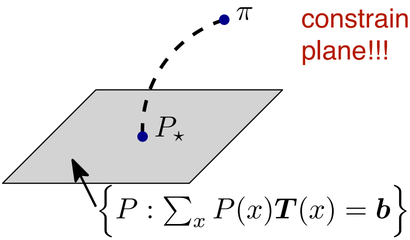
Note that *any* feasible solution P must satisfy

$$\sum_{x \in \mathcal{X}} P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

project the base into the constrain plane

These constraints define an affine hyperplane in $\mathbb{R}^{\mathcal{X}}$.

$$\begin{aligned} \min_{P \in \Delta(\mathcal{X})} \quad & \mathrm{RE}(P \| \pi) \\ \text{s.t.} \quad & \sum_{x \in \mathcal{X}} P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b} \end{aligned}$$



constrain plane!!!

Similar to the Euclidean projection of π onto an affine hyperplane, except we **use relative entropy instead of Euclidean distance**: an **entropy projection**.

Solution form

Maximum entropy optimization problem:

$$\begin{aligned} \min_{P \in \Delta(\mathcal{X})} \quad & \mathrm{RE}(P \| \pi) \\ \text{s.t.} \quad & \sum_{x \in \mathcal{X}} P(x) \boldsymbol{T}(x) \;=\; \boldsymbol{b}. \end{aligned}$$

Solution form

Maximum entropy optimization problem:

$$\begin{aligned} \min_{P \in \Delta(\mathcal{X})} \quad & \mathrm{RE}(P \| \pi) \\ \text{s.t.} \quad & \sum_{x \in \mathcal{X}} P(x) \boldsymbol{T}(x) \; = \; \boldsymbol{b}. \end{aligned}$$

Claim: A solution $P_\star$ to the optimization problem must have the form

$$P_\star(x) \; = \; \frac{1}{Z(\boldsymbol{\eta})} \cdot \exp\Big\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle \Big\} \cdot \pi(x)$$

for some $\boldsymbol{\eta} \in \mathbb{R}^k$, where

$$Z(\boldsymbol{\eta}) \; = \; \sum_{x \in \mathcal{X}} \exp\Big\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle \Big\} \cdot \pi(x)$$

is the normalizing constant that makes $P_\star$ a probability distribution.

Solution form

Maximum entropy optimization problem:

$$\begin{aligned} \min_{P \in \Delta(\mathcal{X})} \quad & \mathrm{RE}(P \| \pi) \\ \text{s.t.} \quad & \sum_{x \in \mathcal{X}} P(x) \boldsymbol{T}(x) \;=\; \boldsymbol{b}. \end{aligned}$$

Claim: A solution $P_\star$ to the optimization problem must have the form

$$P_\star(x) \;=\; \exp\Big\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle - \ln Z(\boldsymbol{\eta}) \Big\} \cdot \pi(x)$$

for some $\boldsymbol{\eta} \in \mathbb{R}^k$, where

$$Z(\boldsymbol{\eta}) \;=\; \sum_{x \in \mathcal{X}} \exp\Big\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle \Big\} \cdot \pi(x)$$

is the normalizing constant that makes $P_\star$ a probability distribution.

Proof

Consider any other feasible solution P—i.e., P is a valid probability distribution and (like $P_\star$) satisfies

$$\sum_x P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

Proof

Consider any other feasible solution P—i.e., P is a valid probability distribution and (like $P_\star$) satisfies

$$\sum_x P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

$$\mathrm{RE}(P\|\pi) - \mathrm{RE}(P_\star\|\pi)$$

Proof

Consider any other feasible solution P—i.e., P is a valid probability distribution and (like $P_\star$) satisfies

$$\sum_x P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

$$\begin{aligned}
&\mathrm{RE}(P\|\pi) - \mathrm{RE}(P_\star\|\pi) \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\ln\frac{P_\star(x)}{\pi(x)}
\end{aligned}$$

Proof

Consider any other feasible solution P—i.e., P is a valid probability distribution and (like $P_\star$) satisfies

$$\sum_x P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

$$\begin{aligned}
&\mathrm{RE}(P\|\pi) - \mathrm{RE}(P_\star\|\pi) \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\ln\frac{P_\star(x)}{\pi(x)} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\Big\{\langle\boldsymbol{\eta},\boldsymbol{T}(x)\rangle - \ln Z(\boldsymbol{\eta})\Big\}
\end{aligned}$$

Proof

Consider any other feasible solution P—i.e., P is a valid probability distribution and (like $P_\star$) satisfies

$$\sum_x P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

$$\begin{aligned}
&\mathrm{RE}(P\|\pi) - \mathrm{RE}(P_\star\|\pi) \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\ln\frac{P_\star(x)}{\pi(x)} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\Big\{\langle\boldsymbol{\eta},\boldsymbol{T}(x)\rangle - \ln Z(\boldsymbol{\eta})\Big\} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P(x)\Big\{\langle\boldsymbol{\eta},\boldsymbol{T}(x)\rangle - \ln Z(\boldsymbol{\eta})\Big\}
\end{aligned}$$

PROOF

Consider any other feasible solution P—i.e., P is a valid probability distribution and (like $P_\star$) satisfies

$$\sum_x P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

$$\begin{aligned}
&\mathrm{RE}(P\|\pi) - \mathrm{RE}(P_\star\|\pi) \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\ln\frac{P_\star(x)}{\pi(x)} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\Big\{\langle\boldsymbol{\eta},\boldsymbol{T}(x)\rangle - \ln Z(\boldsymbol{\eta})\Big\} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P(x)\Big\{\langle\boldsymbol{\eta},\boldsymbol{T}(x)\rangle - \ln Z(\boldsymbol{\eta})\Big\} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P(x)\ln\frac{P_\star(x)}{\pi(x)}
\end{aligned}$$

Proof

Consider any other feasible solution P—i.e., P is a valid probability distribution and (like $P_\star$) satisfies

$$\sum_x P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

$$\begin{aligned}
&\mathrm{RE}(P\|\pi) - \mathrm{RE}(P_\star\|\pi) \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\ln\frac{P_\star(x)}{\pi(x)} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\Big\{\langle\boldsymbol{\eta},\boldsymbol{T}(x)\rangle - \ln Z(\boldsymbol{\eta})\Big\} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P(x)\Big\{\langle\boldsymbol{\eta},\boldsymbol{T}(x)\rangle - \ln Z(\boldsymbol{\eta})\Big\} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P(x)\ln\frac{P_\star(x)}{\pi(x)} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{P_\star(x)}
\end{aligned}$$

Proof

Consider any other feasible solution P—i.e., P is a valid probability distribution and (like $P_\star$) satisfies

$$\sum_x P(x)\boldsymbol{T}(x) \;=\; \boldsymbol{b}.$$

$$\begin{aligned}
&\mathrm{RE}(P\|\pi) - \mathrm{RE}(P_\star\|\pi) \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\ln\frac{P_\star(x)}{\pi(x)} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P_\star(x)\Big\{\langle\boldsymbol{\eta},\boldsymbol{T}(x)\rangle - \ln Z(\boldsymbol{\eta})\Big\} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P(x)\Big\{\langle\boldsymbol{\eta},\boldsymbol{T}(x)\rangle - \ln Z(\boldsymbol{\eta})\Big\} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{\pi(x)} - \sum_x P(x)\ln\frac{P_\star(x)}{\pi(x)} \\
&\quad= \sum_x P(x)\ln\frac{P(x)}{P_\star(x)} \;=\; \mathrm{RE}(P\|P_\star) \;\ge\; 0 \quad \text{with equality iff } P = P_\star.
\end{aligned}$$

the minimization is achieved when p = p*

$\square$

Interpretation of the solution form

From our earlier example:

- $T_1(x) = \mathbb{1}\{x \text{ ends with an 'e'}\}$
- $T_2(x) = \mathbb{1}\{x \text{ has more than five characters}\}$
- ...

Maximum entropy solution is of the form

$$P_\star(x) = \frac{1}{Z(\boldsymbol{\eta})} \cdot \exp\Big\{\eta_1 T_1(x) + \eta_2 T_2(x) + \cdots\Big\} \cdot \pi(x).$$

Interpretation of the solution form

From our earlier example:

- $T_1(x) = \mathbb{1}\{x \text{ ends with an 'e'}\}$
- $T_2(x) = \mathbb{1}\{x \text{ has more than five characters}\}$
- ...

Maximum entropy solution is of the form

$$P_\star(x) = \frac{1}{Z(\boldsymbol{\eta})} \cdot \exp\Big\{\eta_1 T_1(x) + \eta_2 T_2(x) + \cdots\Big\} \cdot \pi(x).$$

Under $P_\star$, a word that ends with an 'e' is e^{η_1} times more likely than a word that doesn't end with an 'e'.

Interpretation of the solution form

From our earlier example:

- $T_1(x) = \mathbb{1}\{x \text{ ends with an 'e'}\}$
- $T_2(x) = \mathbb{1}\{x \text{ has more than five characters}\}$
- ...

Maximum entropy solution is of the form

$$P_\star(x) = \frac{1}{Z(\boldsymbol{\eta})} \cdot \exp\Big\{\eta_1 T_1(x) + \eta_2 T_2(x) + \cdots\Big\} \cdot \pi(x).$$

Under $P_\star$, a word that ends with an 'e' is e^{η_1} times more likely than a word that doesn't end with an 'e'.

How do we get these $\boldsymbol{\eta}$ parameters?

Exponential families

The $\boldsymbol{\eta}$ parameters for distributions of the form

$$P_{\boldsymbol{\eta}}(x) = \frac{1}{Z(\boldsymbol{\eta})} \cdot \exp\Big\{ \langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle \Big\} \cdot \pi(x)$$

are strongly related to a different parameterization of the distributions called the **expectation parameters**, which are easily estimated.

Exponential families

The $\boldsymbol{\eta}$ parameters for distributions of the form

$$P_{\boldsymbol{\eta}}(x) = \frac{1}{Z(\boldsymbol{\eta})} \cdot \exp\left\{ \boxed{\langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle} \right\} \cdot \pi(x)$$

are strongly related to a different parameterization of the distributions called the **expectation parameters**, which are easily estimated.

This relationship is revealed through the study of these distribution families, called **exponential families**.

Exponential families

The $\boldsymbol{\eta}$ parameters for distributions of the form

Zn is the normalizer!

$$P_{\boldsymbol{\eta}}(x) = \frac{1}{Z(\boldsymbol{\eta})} \cdot \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle\Big\} \cdot \pi(x)$$

are strongly related to a different parameterization of the distributions called the **expectation parameters**, which are easily estimated.

This relationship is revealed through the study of these distribution families, called **exponential families**.

Many familiar probability models are exponential families:

> Bernoulli, binomial, Poisson, exponential, Gaussian, gamma, categorical, multinomial, Dirichlet, . . .

Recap

- Maximum entropy approach to probabilistic modeling: choose the most non-committal distribution that agrees with the empirical observation.
- Solution must have the form

$$P_{\boldsymbol{\eta}}(x) = \frac{1}{Z(\boldsymbol{\eta})} \cdot \exp\Big\{\langle \boldsymbol{\eta}, \boldsymbol{T}(x) \rangle\Big\} \cdot \pi(x),$$

  corresponds to the entropy projection of the base distribution π onto an affine hyperplane.
- Extracting the $\boldsymbol{\eta}$ parameters: next time, via study of exponential families.