Vorlesung 14b

Relative Entropie

Zur Wiederholung:

Sei S eine endliche oder abzählbare Menge
und ρ eine Verteilung auf S.

Die binäre Entropie von ρ ist

$$\mathbf{H}_2[\rho] := -\sum_{a \in S} \rho(a) \log_2 \rho(a)$$

(die bis auf maximal ein Bit kleinstmögliche erwartete Länge
eines binären Präfixcodes unter der Verteilung ρ)

Anstatt an binäre Präfixcodes kann man auch an trinäre, oder allgemeiner (für $b \in \mathbb{R}_+$) an b-äre Präfixcodes denken.

Die Entropie zur Basis b der Verteilung ρ ist

$$\mathbf{H}_b[\rho] := -\sum_{a \in S} \rho(a) \log_b \rho(a).$$

Die prominentesten Wahlen sind $b = 2$ und $b = e$. Im Folgenden denken wir uns ein $b > 0$ fest gewählt und lassen das Subskript b weg.

1. Definition und Interpretation der relativen Entropie

Definition: Seien ρ und π Wahrscheinlichkeitsverteilungen mit Gewichten $\rho(a)$ und $\pi(a)$, $a \in S$. Dann ist die *relative Entropie* von ρ bzgl. π definiert als

$$D(\rho\|\pi) := \sum_{a\in S} \rho(a) \log \frac{\rho(a)}{\pi(a)}$$

$$= -\sum_{a\in S} \rho(a) \log \pi(a) - \mathbf{H}[\rho] \ ,$$

wobei die Summanden mit $\rho(a) = 0$ gleich 0 gesetzt werden.

Interpretation der relativen Entropie:

Man denke sich einen zufälligen Buchstaben mit Verteilung ρ

mit einem Shannon-Code codiert, der nicht der Verteilung ρ,

sondern der Verteilung π angepasst ist,

also mit Codewortlängen

$$-\log \pi(a) \leq \ell(a) < -\log \pi(a) + 1.$$

Dann ändert sich die erwarteten Codelänge

im Vergleich zu dem an ρ angepassten Shannon-Code

(bis auf höchstens 1) um

$$-\sum_a \rho(a) \log \pi(a) - \Big(-\sum_a \rho(a) \log \rho(a) \Big) = D(\rho \| \pi).$$

2. Die Informationsungleichung

Satz: („Informationsungleichung") $D(\rho\|\pi) \geq 0$.

Beweis: Wieder verwenden wir die Abschätzung

$\log x \leq c \cdot (x-1)$ mit $c := \log'(1)$:

$$D(\rho\|\pi) = -\sum_{a:\rho(a)>0} \rho(a) \log \frac{\pi(a)}{\rho(a)}$$

$$\geq -\sum_{a:\rho(a)>0} \rho(a)\, c \cdot \left(\frac{\pi(a)}{\rho(a)} - 1\right)$$

$$= -c\Big(\sum_{a:\rho(a)>0} \pi(a) - \sum_{a} \rho(a)\Big) \geq 0. \quad \square$$

Bemerkung: Aus $D(\rho\|\pi) = 0$ folgt $\rho = \pi$.

In der Tat: In der Ungleichung $\quad \log x \le c(x-1)$

besteht (abgesehen für $x = 1$) *strikte* Ungleichung.

Also folgt aus

$$- \sum_{a:\rho(a)>0} \rho(a) \log \frac{\pi(a)}{\rho(a)} = -c \sum_{a:\rho(a)>0} \rho(a)\Big(\frac{\pi(a)}{\rho(a)} - 1\Big),$$

dass $\quad \pi(a) = \rho(a)$ für alle a mit $\rho(a) > 0$.

Daraus folgt $\displaystyle\sum_{a:\rho(a)>0} \pi(a) = 1$, also $\displaystyle\sum_{a:\rho(a)=0} \pi(a) = 0$,

somit auch $\quad \pi(a) = \rho(a)$ für alle a mit $\rho(a) = 0$. $\square$

Zusammenfassend ergibt sich der

Satz (von der relativen Entropie):

Die relative Entropie $D(\rho\|\pi)$ ist nichtnegativ,

und verschwindet genau für $\rho = \pi$.

3. Entropieschranken

In den folgenden Beispielen

benutzen wir den eben bewiesenen Satz in der Gestalt

$$(*) \qquad -\sum_a \rho(a) \log \rho(a) \le -\sum_a \rho(a) \log \pi(a)$$

mit Gleichheit genau für $\rho = \pi$.

Wir sehen:

Jede Wahl von π liefert in $(*)$ eine Schranke für $\mathbf{H}[\rho]$,

mit Gleichheit genau für $\rho = \pi$.

Für jede Wahl von π wird die rechte Seite von $(*)$
zum Erwartungswert der Zufallsvariablen $g(X) := -\log \pi(X)$.

$$(*) \qquad -\sum_a \rho(a) \log \rho(a) \leq -\sum_a \rho(a) \log \pi(a)$$ mit Gleichheit genau für $\rho = \pi$.

Beispiel 1: Vergleich mit der uniformen Verteilung:

Sei S endlich mit n Elementen

und sei $\pi(a) = 1/n$ für alle $a \in S$.

Dann folgt aus $(*)$ für jede Verteilung ρ auf S:

$$\mathbf{H}[\rho] \leq -\sum_a \rho(a) \log \left(\frac{1}{n}\right) = \log n\ .$$

$$\boxed{\mathbf{H}[\rho] \leq \log n.}$$

Gleichheit gilt genau im Fall der uniformen Verteilung,

sie maximiert auf S die Entropie. $\qquad \square$

$$(*) \qquad -\sum_a \rho(a) \log \rho(a) \le -\sum_a \rho(a) \log \pi(a) \text{ mit Gleichheit genau für } \rho = \pi.$$

Beispiel 2: Vergleich mit verschobener geometr. Verteilung:

Sei nun $S = \{0, 1, 2, \ldots\}$, und $\pi(k) := 2^{-k-1}$.

Dann folgt aus $(*)$ für alle Verteilungen ρ mit EW $\mu(\rho)$:

$$\mathbf{H}_2[\rho] \le -\sum_k \rho(k) \log_2(2^{-k-1})$$

$$= \sum_{k=0}^{\infty} \rho(k)(k+1) = \mu(\rho) + 1 \, .$$

Gleichheit gilt für $\rho = \pi$, dann ist $\mathbf{H}_2[\rho] = 2$. Also:

Unter allen Verteilungen auf $\mathbb{N}_0$ mit EW ≤ 1

hat die Verteilung π die größte binäre Entropie, nämlich 2.$\square$

Im nächsten Beispiel betrachten wir

(für eine Abbildung $u : S \to \mathbb{R}$)

die Frage:

Wie sieht unter allen Verteilungen von X

mit vorgegebenem Wert η für $\mathbf{E}[u(X)]$

diejenige mit der größten Entropie aus?

(Das obige Beispiel 2 passt in diesem Rahmen mit $u(k) := k$)

Wieder verwenden wir die Informationsungleichung

in der Form

$$(*) \qquad -\sum_a \rho(a) \log \rho(a) \le -\sum_a \rho(a) \log \pi(a)$$

mit Gleichheit genau für $\rho = \pi$.

Beispiel 3: Vergleich mit einer “Boltzmann-Gibbs-Verteilung”:

Gegeben sei $u : S \to \mathbb{R}$, $\beta \geq 0$.

Wir definieren die Gewichte $\pi(a) := e^{-\beta u(a)}/z$ mit

$$z := \sum_{a \in S} e^{-\beta u(a)} \quad \text{(Annahme: } z < \infty.\text{)}$$

Sei $\eta := \sum_a u(a)\pi(a)$.

Die Abschätzung ($*$) ergibt für alle ρ mit $\sum \rho(a)u(a) = \eta$

$$\mathbf{H}_e[\rho] \leq -\sum \rho(a) \ln \pi(a) = \beta \sum \rho(a)u(a) + \ln z$$

$$= \beta\eta + \ln z$$

mit Gleichheit genau für $\rho = \pi$.

Anders gewendet:

Unter allen Zufallsvariablen X mit

vorgegebenem Erwartungswert $\eta = \mathbf{E}[u(X)]$

hat diejenige die größte Entropie,

die die Verteilungsgewichte $e^{-\beta u(a)}/z$ hat

wobei β so eingerichtet ist, dass $\sum_a u(a) e^{-\beta u(a)}/z = \eta$ gilt.

Die Verteilung mit den Gewichten $e^{-\beta u(a)}/z$

heißt *Boltzmann-Gibbsverteilung*

zum Potenzial u mit Parameter β.

4. Relative Entropie und große Abweichungen

Beim n-fachen Würfeln mit Gewichten $\mathbf{p} := (p_1, \dots, p_g)$

sind die relativen Häufigkeiten $K_1/n, \dots, K_g/n$

für großes n mit großer W'keit nahe bei $p_1, \dots, p_g$.

Wie wahrscheinlich ist ein "atypischer Ausgang" $(a_1, \dots, a_g)$

mit $a_j \sim nt_j, \;\; \mathbf{t} := (t_1, \dots, t_g) \neq \mathbf{p}$?

$$\mathbf{P}_{\mathbf{p}}(K_1 = a_1, \dots, K_g = a_g) = \binom{n}{a_1, \dots, a_g} p_1^{a_1} \cdots p_g^{a_g}$$

Aus der Stirling-Formel folgt: Bis auf einen Faktor $f = f(n)$, der (nur) wie eine Potenz in n wächst, ist $\binom{n}{a_1, \dots, a_g} \asymp \frac{1}{t_1^{a_1} \cdots t_g^{a_g}}$. Also:

$$\mathbf{P}(K_1 = a_1, \dots, K_g = a_g) \asymp \left(\left(\frac{p_1}{t_1}\right)^{t_1} \cdots \left(\frac{p_g}{t_g}\right)^{t_g} \right)^n$$

Nach dem Logarithmieren fällt der Faktor $f(n)$ nicht mehr ins Gewicht:

$$\ln \mathbf{P_p}(K_1 = a_1, \ldots, K_g = a_g) \sim n \sum_{j=1}^{g} t_j \ln \frac{p_j}{t_j}$$

$$= -nD(\mathbf{t}||\mathbf{p}))$$

mit $\mathbf{t} := (t_1, \ldots, t_g)$ und $\mathbf{p} := (p_1, \ldots, p_g)$
aufgefasst als W-Verteilungen auf $\{1, \ldots, g\}$

Fazit: Unter der Annahme $a_j \sim nt_j$ mit $n \to \infty$ gilt:

$$\mathbf{P_p}(K_1 = a_1, \ldots, K_g = a_g)$$

fällt exponentiell in n mit Rate $D(\mathbf{t}||\mathbf{p})$.

$\mathsf{S} = \mathsf{k}\ \log W$

Entropie =
k mal
Logarithmus der
Wahrscheinlichkeit

Ludwig Boltzmann
1844-1906

Grabmal am
Wiener
Zentralfriedhof

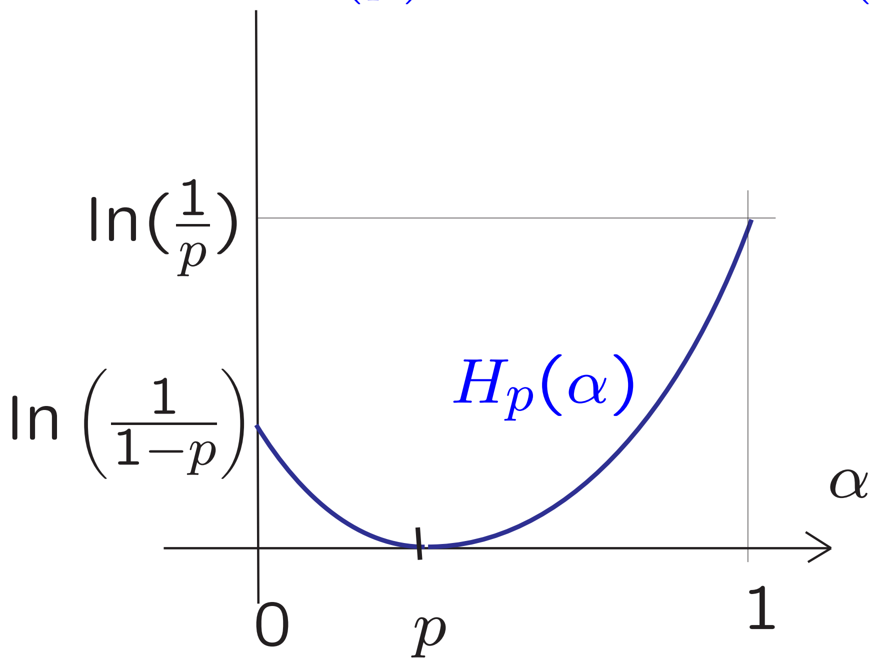
5. Eine Beziehung zur Chernoff-Ungleichung

In V 7b hatten wir für Binomial(n, p)-verteiltes X_n

(und $\alpha \geq p$) die Chernoff-Ungleichung bewiesen:

$$\mathbf{P}(X_n > \alpha\, n) \leq e^{-nH_p(\alpha)}$$

mit $H_p(\alpha) := \alpha \ln\left(\frac{\alpha}{p}\right) + (1-\alpha)\ln\left(\frac{1-\alpha}{1-p}\right) > 0.$



In V 7b hatten wir für Binomial(n, p)-verteiltes X_n

(und $\alpha \geq p$) die Chernoff-Ungleichung bewiesen:

$$\mathbf{P}(X_n > \alpha\, n) \leq e^{-nH_p(\alpha)}$$

mit $H_p(\alpha) := \alpha \ln\left(\frac{\alpha}{p}\right) + (1-\alpha)\ln\left(\frac{1-\alpha}{1-p}\right) > 0.$

Ist π die Verteilung auf $\{1, 0\}$ mit Gewichten p und $1 - p$

und ρ die Verteilung auf $\{1, 0\}$ mit Gewichten α und $1 - \alpha$

(also: π = Bernoulli(p), ρ = Bernoulli (α)),

so hat man

$$H_p(\alpha) = D(\rho \| \pi).$$