Math 254A Lecture 26 Notes

Daniel Raban

May 26, 2021

1 Basics of Shannon Entropy and Connection to Entropy Rate

1.1 Basic inequalities for Shannon entropy

Definition 1.1. Let A be a finite set with $p \in P(A)$, and let $\alpha \sim p$ be an A-valued random variable. Then

$$H(\alpha) := -\sum_{\alpha \in A} \mathbb{P}(\alpha = a) \log \underbrace{\mathbb{P}(\alpha = a)}_{p(a)} = H(p).$$

is the **Shannon entropy** of α (or of p).

The Shannon entropy quantifies how "uncertain" α is. We have seen that $H(p) \geq 0$ and is $\leq \log |A|$, with equalities achieved with a point mass and with the uniform distribution on $|A|$, respectively.

Next consider random variables α, β with values in A, B. Regard (α, β) as a random variable with values in $A \times B$. The joint distribution is $p_{\alpha,\beta} \in P(A \times B)$. Then

$$\begin{aligned}
H(\alpha, \beta) &= -\sum_{a,b} p_{\alpha,\beta}(a,b) \log p_{\alpha,\beta}(a,b) \\
&= -\sum_{a,b} p_{\alpha,\beta}(a,b) \log \left(p_\alpha(a) \underbrace{p_{\beta|\alpha}(b \mid a)}_{\mathbb{P}(\beta = b | \alpha = a)} \right) \\
&= -\sum_{a,b} p_{\alpha,\beta}(a,b) \log p_\alpha(a) - \sum_{a,b} p_\alpha(a) p_{\beta|\alpha}(b \mid a) \log p_{\beta|\alpha}(b \mid a) \\
&= -\sum_{a} p_\alpha(a) \log p_\alpha(a) + \sum_{a} p_\alpha(a) \cdot H(p_{\beta|\alpha}(\cdot \mid a)) \\
&= H(\alpha) + H(\beta \mid \alpha),
\end{aligned}$$

where $H(\beta \mid \alpha) := \sum_a p_\alpha(a) \cdot H(p_{\beta|\alpha}(\cdot \mid a))$.

Here is the generalization of this fact:

Theorem 1.1 (Chain rule)**.**

$$H(\alpha_1, \dots, \alpha_n) = H(\alpha_1) + H(\alpha_2 \mid \alpha_1) + H(\alpha_3 \mid \alpha_1, \alpha_2) + \cdots + H(\alpha_m \mid \alpha_1, \dots, \alpha_{m-1}).$$

We also have the following property.

Lemma 1.1.

$$H(\beta \mid \alpha) \leq H(\beta),$$

and equality holds iff α, β are independent, in which case

$$H(\alpha, \beta) \leq H(\alpha) + H(\beta)$$

Proof.

$$H(\beta \mid \alpha) = \sum_a p_\alpha(a) H(p_{\beta|\alpha}(\cdot \mid a)).$$

By the Law of Total Probability, for all $b \in B$,

$$p_\beta(b) = \sum_\alpha p_\alpha(a) p_{\beta|\alpha}(b \mid a).$$

Since H is strictly concave, Jensen's inequality gives that

$$H(\beta) = H(p_\beta) \geq \sum_\alpha p_\alpha(a) H(p_{\beta|\alpha}(\cdot \mid a)) = H(\beta \mid \alpha).$$

Equality holds in Jensen's inequality iff $P_{\beta|\alpha}(\cdot \mid a) = p_\beta$ whenever $p_\alpha(a) > 0$, i.e. α, β are independent. $\square$

Corollary 1.1.

$$H(\gamma \mid \alpha, \beta) \leq H(\gamma \mid \beta)$$

and similarly with more random variables. Equality holds iff α, γ are conditionally independent given β.

Here is a corollary of the chain rule:

Corollary 1.2. *Let A be a finite set, $p \in P(A)$, and $0 \leq \varepsilon < 1/2$. Suppose $A = B \sqcup C$ with $|B| \leq |C|$ and $p(C) \leq \varepsilon$. Then*

$$H(p) \leq H(\varepsilon, 1 - \varepsilon) + (1 - \varepsilon) \log |B| + \varepsilon \log |C|.$$

Proof. Let $\alpha \sim p$, and let

$$\beta = \mathbb{1}_B(\alpha) = \begin{cases} 1 & \alpha \in B \\ 0 & \alpha \in C. \end{cases}$$

So $H(\alpha) = H(\alpha) + H(\alpha \mid \beta) = H(\alpha, \beta)$. Expanding via β first, we get

$$\begin{aligned} H(\alpha) &= H(\alpha, \beta) \\ &= H(\beta) + H(\alpha \mid \beta) \\ &= H(\beta) + \mathbb{P}(\beta = 1) H(p(\cdot \mid B)) + \mathbb{P}(\beta = 0) H(p(\cdot \mid C)) \\ &\leq H(\varepsilon, 1-\varepsilon) + p(B) \cdot \log |B| + p(C) \cdot \log |C| \\ &= H(\varepsilon, 1-\varepsilon) + (1-\varepsilon) \log |B| + \varepsilon \log |C|. \end{aligned}$$

□

Here is the last information-theoretic inequality we need.

Theorem 1.2 (Shearer's inequality)**.** *Let $\alpha_1, \ldots, \alpha_m$ be valued in $A_1, \ldots, A_m$, let $\mathcal{S} \subseteq \mathscr{P}(\{1, \ldots, m\})$, and let $k \geq 1$. Assume that every $i \in \{1, \ldots, m\}$ is contained in $\geq k$ members of $\mathcal{S}$. Then*

$$H(\alpha_1, \ldots, \alpha_m) \leq \frac{1}{k} \sum_{S \in \mathcal{S}} H(\alpha_i : i \in S).$$

Proof. Here is the proof in the case $m = 3$ and $\mathcal{S} = \{\{1,2\}, \{1,3\}, \{2,3\}\}$ ($k = 2$); the argument generalizes well.

$$H(\alpha_1, \alpha_2) = H(\alpha_1) + H(\alpha_2 \mid \alpha_1)$$

$$H(\alpha_1, \alpha_3) = H(\alpha_1) + \qquad + H(\alpha_3 \mid \alpha_1)$$

$$H(\alpha_2, \alpha_3) = \qquad H(\alpha_2) + H(\alpha_3 \mid \alpha_2)$$

Adding together the columns, the first column is $H(\alpha_1$, the second column is $\geq 2H(\alpha_2 \mid \alpha_1)$, and the third column is $\geq 2H(\alpha_3 \mid \alpha_1, \alpha_2)$. So we get

$$\begin{aligned} H(\alpha_1, \alpha_2) + H(\alpha_1, \alpha_3) H(\alpha_2, \alpha_3) &= 2[H(\alpha_1) + H(\alpha_2 \mid \alpha_1) + H(\alpha_3 \mid \alpha_1, \alpha_2) \\ &= 2H(\alpha_1, \alpha_2, \alpha_3). \end{aligned}$$

□

1.2 Applying Shearer's inequality to lattice models

Here is a corollary of Shearer's inequality.

Corollary 1.3. *Let $W, B \subseteq \mathbb{Z}^d$ be finite with $0 < |A| < \infty$ and $\mu \in P(A^B)$. Then*

$$H(\mu) \leq \frac{1}{|W|} \sum_{v + W \subseteq B} H(\mu_{v+W}) + O\left(\frac{\log |A| \cdot |B| \cdot \operatorname{diam}(W)}{\text{min-side-length}(B)} \right).$$

Proof. Let $\mathcal{S}_0 = \{v + W : v + W \subseteq B\}$, and define $\mathcal{S}_1 = \{(v+W) \cap B : (v + W) \cap B \neq \varnothing\}$. Then $\mathcal{S}_0 \subseteq \mathcal{S}_1$, and $\mathcal{S}_1$ covers every element of B exactly $|W|$-many times. Apply Shearer's inequality to get

$$H(\mu) \leq \frac{1}{|W|} \sum_{(v+W) \cap B \in \mathcal{S}_1} H(\mu_{(v+W) \cap B}) = \frac{1}{|W|} \sum_{\mathcal{S}_0} H(\mu_{v+W}) + \text{error}.$$

The number of terms put into the error is $|\mathcal{S}_1 \setminus \mathcal{S}_0| = O(\frac{\text{diam}(W)\cdot|B|}{\text{min-side-length}(B)})$. Each of these terms is $\leq \log|A^W| = |W| \cdot \log|A|$. $\square$

Now return to shift-invariant measures $\mu \in P^T(A^{\mathbb{Z}^d})$.

Lemma 1.2. *The limit* $\lim_{B\uparrow\mathbb{Z}^d} \frac{1}{|B|}H(\mu_B)$ *exists, and*

$$\lim_{B\uparrow\mathbb{Z}^d} \frac{1}{|B|} H(\mu_B) = \inf_B \frac{1}{|B|} H(\mu_\beta).$$

Here is a proof using Shearer's inequality:

Proof. Apply the previous corollary to a shift-invariant measure μ, and observe $\mu_{v+W} = \mu_W$ (up to fixing indexing). Then

$$\begin{aligned}
\frac{1}{|B|} H(\mu_B) &\leq \frac{1}{|B|} \sum_{v+W\subseteq B} \frac{1}{|W|} H(\mu_W) + o(1) \\
&= \frac{|\{v : v + W \subseteq B\}|}{|B|} \cdot \frac{1}{|W|} H(\mu_W) + o(1) \\
&\leq \frac{1}{|W|} H(\mu_W) + o(1).
\end{aligned}$$

So in fact,

$$\lim_{B\uparrow\mathbb{Z}^d} \frac{1}{|B|} H(\mu_\beta) = \inf_{|w|<\infty} \frac{1}{|W|} H(\mu_W).$$

$\square$

Definition 1.2. The quantity

$$h(\mu) = \lim_{B\uparrow\mathbb{Z}^d} \frac{1}{|B|} H(\mu_B) \qquad (\mu \in P^T(A^{\mathbb{Z}^d}))$$

is called the **entropy rate** of μ.

The entropy rate satisfies

$$0 \leq h(\mu) \leq H(\mu_{\{0\}}).$$

Theorem 1.3. $s = h$ *on* $P^T(A^{\mathbb{Z}^d})$*, and so* $\{s > -\infty\} = \{s \geq 0\} = P^T(A^{\mathbb{Z}^d})$.