Topic Course on Probabilistic Methods

(Week 14)

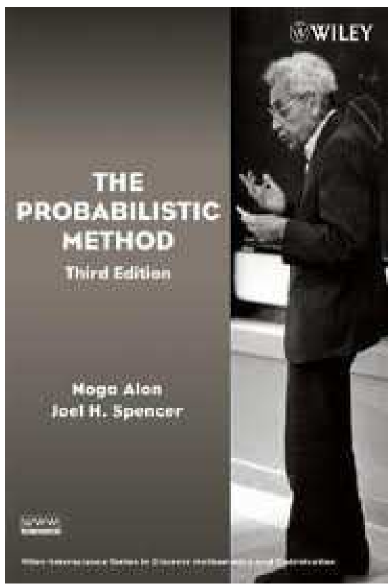
Entropy

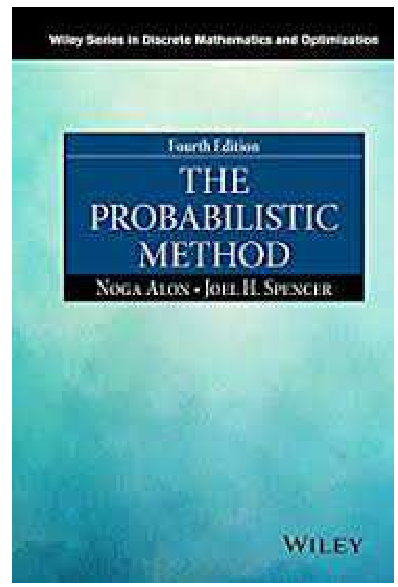
Linyuan Lu

University of South Carolina

Introduction

The topic course is mostly based the textbook "The probabilistic Method" by Noga Alon and Joel Spencer (third edition 2008, John Wiley & Sons, Inc. ISBN 9780470170205 or fourth edition ISBN-13: 978-1119061953.)





Selected topics

- Linearity of Expectation (2 weeks)
- Alterations (1 week)
- The second moment method (1 week)
- The Local Lemma (1-2 weeks)
- Correlation Inequalities (1 week)
- Large deviation inequalities (3 weeks)
- Poisson Paradigm (1 week)
- Random graphs (2 weeks)
- Discrepancy (1 week)
- Entropy (1 week)

Subtopics

Entropy

- Motivation
- Entropy
- Properties
- Applications
- Shannon's theorem

Motivation

Estimate binary coefficients: For fixed $\alpha \in (0,1)$,

$$\begin{aligned}
\binom{n}{\alpha n} &= \frac{n!}{(\alpha n)!((1-\alpha)n)!} \\
&\approx \frac{\sqrt{2\pi n}\frac{n^n}{e^n}}{\sqrt{2\pi\alpha n}\frac{(\alpha n)^{\alpha n}}{e^{\alpha n}}\sqrt{2\pi(1-\alpha)n}\frac{((1-\alpha)n)^{(1-\alpha)n}}{e^{(1-\alpha)n}}} \\
&= \frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}}\left(\alpha^{-\alpha}(1-\alpha)^{-(1-\alpha)}\right)^n \\
&= 2^{(1+o(1))H(\alpha)n},
\end{aligned}$$

where $H(\alpha) = -\alpha\log_2\alpha - (1-\alpha)\log_2(1-\alpha)$.

Motivation

Estimate binary coefficients: For fixed $\alpha \in (0,1)$,

$$\binom{n}{\alpha n} = \frac{n!}{(\alpha n)!((1-\alpha)n)!}$$
$$\approx \frac{\sqrt{2\pi n}\frac{n^n}{e^n}}{\sqrt{2\pi\alpha n}\frac{(\alpha n)^{\alpha n}}{e^{\alpha n}}\sqrt{2\pi(1-\alpha)n}\frac{((1-\alpha)n)^{(1-\alpha)n}}{e^{(1-\alpha)n}}}$$
$$= \frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}}\left(\alpha^{-\alpha}(1-\alpha)^{-(1-\alpha)}\right)^n$$
$$= 2^{(1+o(1))H(\alpha)n},$$

where $H(\alpha) = -\alpha\log_2\alpha - (1-\alpha)\log_2(1-\alpha)$.

For $\alpha < \frac{1}{2}$, we also have $\sum_{i<\alpha n}\binom{n}{i} = 2^{(1+o(1))H(\alpha)n}$.

Entropy

Let X be a random variable taking values in some range S. The **binary entropy** of X, denoted by $H(X)$ is defined by

$$H(X) = \sum_{x \in S} \Pr(X = x) \log_2 \frac{1}{\Pr(X = x)}.$$

Entropy

Let X be a random variable taking values in some range S. The **binary entropy** of X, denoted by $H(X)$ is defined by

$$H(X) = \sum_{x \in S} \Pr(X = x) \log_2 \frac{1}{\Pr(X = x)}.$$

Example 1: If $X = 0$ with probability α and $X = 1$ with probability $1 - \alpha$, then

$$H(X) = -\alpha \log_2 \alpha - (1 - \alpha) \log_2 (1 - \alpha) = H(\alpha).$$

Entropy

Let X be a random variable taking values in some range S. The **binary entropy** of X, denoted by $H(X)$ is defined by

$$H(X) = \sum_{x \in S} \Pr(X = x) \log_2 \frac{1}{\Pr(X = x)}.$$

Example 1: If $X = 0$ with probability α and $X = 1$ with probability $1 - \alpha$, then

$$H(X) = -\alpha \log_2 \alpha - (1 - \alpha) \log_2 (1 - \alpha) = H(\alpha).$$

Example 2: If X takes n values with equal probability, then

$$H(X) = \log_2 n.$$

Property I

Property 1: Among all random variables taking values in S, the variable with uniform distribution has the largest entropy.

Property I

Property 1: Among all random variables taking values in S, the variable with uniform distribution has the largest entropy.

Proof: Note that $z \log_2 z$ is concave upward. We have

$$
\begin{aligned}
H(X) &= \sum_{x \in S} \Pr(X = x) \log_2 \frac{1}{\Pr(X = x)} \\
&\leq \log_2 \sum_{x \in S} \Pr(X = x) \frac{1}{\Pr(X = x)} \\
&\leq \log_2 |S|.
\end{aligned}
$$

The equality holds if and only if $\Pr(X = x) = \frac{1}{|S|}$ for any $x \in S$.

Property II

Property 2: $H(X,Y) \geq H(X)$.

Property II

Property 2: $H(X,Y) \geq H(X)$.

Proof:

$$
\begin{aligned}
H(X,Y) &= \sum_{x\in S, y\in T} \Pr(X=x, Y=y)\log_2 \frac{1}{\Pr(X=x,Y=y)} \\
&\geq \sum_{x\in S, y\in T} \Pr(X=x, Y=y)\log_2 \frac{1}{\Pr(X=x)} \\
&= \sum_{x\in S} \Pr(X=x)\log_2 \frac{1}{\Pr(X=x)} \\
&= H(X).
\end{aligned}
$$

Property III

Property 3: $H(X, Y) \leq H(X) + H(Y)$.

Property III

Property 3: $H(X,Y) \le H(X) + H(Y)$.

Proof:

$$
\begin{aligned}
& H(X) + H(Y) - H(X,Y) \\
& \quad = \sum_{i \in S} \sum_{j \in T} \Pr(X=i, Y=j) \log_2 \frac{\Pr(X=i, Y=j)}{\Pr(X=i)\Pr(Y=j)} \\
& \quad = \sum_{i \in S} \sum_{j \in T} \Pr(X=i)\Pr(Y=j) f(z_{ij}),
\end{aligned}
$$

where $f(z) = z \log_2 z$ and $z_{ij} = \frac{\Pr(X=i, Y=j)}{\Pr(X=i)\Pr(Y=j)}$. By the convexity inequality of $f(z)$, we have

$$H(X) + H(Y) - H(X,Y) \ge f(1) = 0. \qquad \square$$

Conditional entropy

The **conditional entropy** of X given Y is

$$H(X|Y) = H(X,Y) - H(Y)$$
$$= \sum_{i\in S}\sum_{j\in T} \Pr(X=i, Y=j) \log_2 \frac{\Pr(Y=j)}{\Pr(X=i, Y=j)}.$$

Conditional entropy

The **conditional entropy** of X given Y is

$$H(X|Y) = H(X,Y) - H(Y)$$

$$= \sum_{i\in S}\sum_{j\in T} \Pr(X=i, Y=j)\log_2 \frac{\Pr(Y=j)}{\Pr(X=i, Y=j)}.$$

By the definition, we have

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X).$$

Conditional entropy

The **conditional entropy** of X given Y is

$$\begin{aligned} H(X|Y) &= H(X,Y) - H(Y) \\ &= \sum_{i \in S} \sum_{j \in T} \Pr(X=i, Y=j) \log_2 \frac{\Pr(Y=j)}{\Pr(X=i, Y=j)}. \end{aligned}$$

By the definition, we have

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X).$$

Mutual information:

$$I(X;Y) = H(X) + H(Y) - H(X,Y).$$

Property IV

Property 4: $H(X|Y,Z) \leq H(X|Y)$.

Property IV

Property 4: $H(X|Y,Z) \le H(X|Y)$.

Proof : $H(X|Y) - H(X|Y,Z)$

$$= \sum_{i\in S}\sum_{j\in T}\sum_{k\in U} \Pr(X=i,Y=j,Z=k)$$

$$\log_2 \frac{\Pr(Y=j)\Pr(X=i,Y=j,Z=k)}{\Pr(X=i,Y=j)\Pr(Y=j,Z=k)}$$

$$= \sum_{i\in S}\sum_{j\in T}\sum_{k\in U} \frac{\Pr(X=i,Y=j)\Pr(Y=j,Z=k)}{\Pr(Y=j)} f(z_{irk})$$

$$\le f(1) = 0.$$

Here $f(z) = z\log z$ and $z_{ijk} = \frac{\Pr(Y=j)\Pr(X=i,Y=j,Z=k)}{\Pr(X=i,Y=j)\Pr(Y=j,Z=k)}$. □

Applications in set theory

Proposition: Let $X = (X_1, X_2, \ldots, X_n)$ be a random variable taking values in the set $S = S_1 \times \cdots S_n$ where each of the coordinates X_i of X is a random variable taking values in S_i. Then

$$H(X) \leq \sum_{i=1}^{n} H(X_i).$$

Applications in set theory

Proposition: Let $X = (X_1, X_2, \ldots, X_n)$ be a random variable taking values in the set $S = S_1 \times \cdots S_n$ where each of the coordinates X_i of X is a random variable taking values in S_i. Then

$$H(X) \leq \sum_{i=1}^{n} H(X_i).$$

Corollary: Let $\mathcal{F}$ be a family of subsets of $[n]$ and let p_i denote the fraction of sets that contain i. Then

$$|\mathcal{F}| \leq 2^{\sum_{i=1}^{n} H(p_i)}.$$

Extension

For any subset $I \subset [n]$, let $X(I)$ denote the random variable $(X_i)_{i \in I}$.

Proposition [Shearer 1986]: If $\mathcal{G}$ is a family of subsets of $[n]$ and each $i \in [n]$ belongs to at least k members of $\mathcal{G}$ then

$$kH(X) \leq \sum_{G \in \mathcal{G}} H(X(G)).$$

Extension

For any subset $I \subset [n]$, let $X(I)$ denote the random variable $(X_i)_{i \in I}$.

Proposition [Shearer 1986]: If $\mathcal{G}$ is a family of subsets of $[n]$ and each $i \in [n]$ belongs to at least k members of $\mathcal{G}$ then

$$kH(X) \leq \sum_{G \in \mathcal{G}} H(X(G)).$$

Proof: We allow $\mathcal{G}$ to be multisets. Now induction on k.

Extension

For any subset $I \subset [n]$, let $X(I)$ denote the random variable $(X_i)_{i \in I}$.

Proposition [Shearer 1986]: If $\mathcal{G}$ is a family of subsets of $[n]$ and each $i \in [n]$ belongs to at least k members of $\mathcal{G}$ then

$$kH(X) \leq \sum_{G \in \mathcal{G}} H(X(G)).$$

Proof: We allow $\mathcal{G}$ to be multisets. Now induction on k.

For $k = 1$, shrink the sets in $\mathcal{G}$ to obtain a family $\mathcal{G}'$ whose members forms a partition of $[n]$.

$$\sum_{G \in \mathcal{G}} H(X(G)) \geq \sum_{G' \in \mathcal{G}'} H(X(G')) \geq H(X).$$

continue

For $k \geq 2$, if $[n] \in \mathcal{G}$, then $\mathcal{G} \setminus \{[n]\}$ covers each point at least $k-1$. By inductive hypothesis,

$$(k-1)H(X) \leq \sum_{G \in \mathcal{G} \setminus \{[n]\}} H(X(G)).$$

It follows

$$\sum_{G \in \mathcal{G}} H(X(G)) = H(X([n])) + \sum_{G \in \mathcal{G} \setminus \{[n]\}} H(X(G)) \geq kH(X).$$

In general, we will replace a pair of G and G' by $G \cap G'$ and $G \cup G'$ first until we get a $[n]$. We claim

$$H(X(G)) + H(X(G')) \geq H(X(G \cup G')) + H(X(G \cap G')).$$

continue

Recall Property IV:

$$H(X'|Y,Z) \le H(X'|Y).$$

This is equivalent to

$$H(X',Y,Z) + H(Y) \le H(X',Y) + H(Y,Z).$$

continue

Recall Property IV:

$$H(X'|Y,Z) \le H(X'|Y).$$

This is equivalent to

$$H(X',Y,Z) + H(Y) \le H(X',Y) + H(Y,Z).$$

Let $X = X(G \setminus G')$, $Y = X(G \cap G')$, and $Z = X(G' \setminus G)$. Note that $(X',Y,Z) = X(G \cup G')$, $(X',Y) = X(G)$, and $(Y,Z) = X(G')$. We get

$$H(X(G \cup G')) + H(X(G \cap G')) \le H(X(G)) + H(X(G')).$$

This finishes the proof of claim and the inductive step. $\square$

Application I

Corollary: Let $\mathcal{F}$ be a family of vectors in $S_1 \times \cdots, S_n$ and $\mathcal{G} := \{G_1, G_2, \ldots, G_m\}$ be a family of subsets of $[n]$ such that each $i \in [n]$ belongs to at least k members of $\mathcal{G}$. For $1 \leq i \leq m$, let $\mathcal{F}_i$ be the set of all projections of the members of $\mathcal{F}$ on G_i. Then

$$|\mathcal{F}|^k \leq \prod_{i=1}^{m} |\mathcal{F}_i|.$$

Application I

Corollary: Let $\mathcal{F}$ be a family of vectors in $S_1 \times \cdots, S_n$ and $\mathcal{G} := \{G_1, G_2, \ldots, G_m\}$ be a family of subsets of $[n]$ such that each $i \in [n]$ belongs to at least k members of $\mathcal{G}$. For $1 \leq i \leq m$, let $\mathcal{F}_i$ be the set of all projections of the members of $\mathcal{F}$ on G_i. Then

$$|\mathcal{F}|^k \leq \prod_{i=1}^{m} |\mathcal{F}_i|.$$

Proof: Let $X = (X_1, \ldots, X_n)$ be the uniform random variable taking values in $\mathcal{F}$. We have

$$kH(X) \leq \sum_{i=1}^{m} H(X(G_i)).$$

But $H(X) = \log_2 |F|$ and $H(X(G_i)) \leq \log_2 |F_i|$, implying the desired result. □

Corollary

Theorem [Loomis, Whitney, 1949]: Let B be a measurable body in the n-dimensional Euclidean space, let $\mathrm{Vol}(B)$ denote its volume, and let $\mathrm{Vol}_i(B)$ denote the $(n-1)$-dimensional volume of the projection of B on the hyperplane orthogonal to i-th axis. Then

$$(\mathrm{Vol}(B))^{n-1} \leq \prod_{i=1}^{n} \mathrm{Vol}(B_i).$$

Corollary

Theorem [Loomis, Whitney, 1949]: Let B be a measurable body in the n-dimensional Euclidean space, let $\mathrm{Vol}(B)$ denote its volume, and let $\mathrm{Vol}_i(B)$ denote the $(n-1)$-dimensional volume of the projection of B on the hyperplane orthogonal to i-th axis. Then

$$(\mathrm{Vol}(B))^{n-1} \leq \prod_{i=1}^{n} \mathrm{Vol}(B_i).$$

Proof: Approximate the volume of a body by the number of standard grid points if the grid is fine enough. The apply the previous corollary. □

Shannon's theorem

The entropy $H(X)$ is also known as Shannon's entropy.

Shannon's theorem

The entropy $H(X)$ is also known as Shannon's entropy.

- A: a set of alphabet.
- $\mathcal{A}$: a probability distribution over A.

Shannon's theorem

The entropy $H(X)$ is also known as Shannon's entropy.

- A: a set of alphabet.
- $\mathcal{A}$: a probability distribution over A.

To encode a file that contain $n|A|$ symbols, the number of bits are required so that the file can be encoded without loss of information is roughly $n \log_2 |A|$.

Shannon's theorem

The entropy $H(X)$ is also known as Shannon's entropy.

- A: a set of alphabet.
- $\mathcal{A}$: a probability distribution over A.

To encode a file that contain $n|A|$ symbols, the number of bits are required so that the file can be encoded without loss of information is roughly $n \log_2 |A|$.

Now we allow an error δ. We seek to encode only files that fall in a set $B \subset A^n$ with $\Pr(B) \geq 1 - \delta$. Then then the number of bits needed is

$$H_\delta(A^n) := \inf_{B \subset A^n, \Pr(B) \geq 1-\delta} \log_2 |B|.$$

Shannon's theorem

The entropy $H(X)$ is also known as Shannon's entropy.

- A: a set of alphabet.
- $\mathcal{A}$: a probability distribution over A.

To encode a file that contain $n|A|$ symbols, the number of bits are required so that the file can be encoded without loss of information is roughly $n \log_2 |A|$.

Now we allow an error δ. We seek to encode only files that fall in a set $B \subset A^n$ with $\Pr(B) \geq 1 - \delta$. Then then the number of bits needed is

$$H_\delta(A^n) := \inf_{B \subset A^n, \Pr(B) \geq 1-\delta} \log_2 |B|.$$

Shannon's theorem: $\forall \delta$, $\lim_{n \to \infty} \frac{1}{n} H_\delta(A^n) = H(\mathcal{A})$.

Proof

Proof: Apply the law of large numbers to the random variable $\log_2 p(a)$: for any $\epsilon > 0$ and a sequence $a_1 a_2, \ldots, a_n \in A^n$,

$$\lim_{n\to\infty} \Pr\left(\left| \frac{1}{n} \sum_{i=1}^{n} \log_2 p(a_i) - \mathrm{E}(\log_2 p(a)) \right| > \epsilon \right) = 0.$$

Proof

Proof: Apply the law of large numbers to the random variable $\log_2 p(a)$: for any $\epsilon > 0$ and a sequence $a_1 a_2, \ldots, a_n \in A^n$,

$$\lim_{n\to\infty} \Pr\left(\left| \frac{1}{n} \sum_{i=1}^{n} \log_2 p(a_i) - \mathrm{E}(\log_2 p(a)) \right| > \epsilon \right) = 0.$$

With probability $1 - o(1)$, $a_1, \ldots, a_n$ satisfies

$$2^{-n(H(\mathcal{A})+\epsilon)} \leq p(a_1, \ldots, p_n) \leq 2^{-n(H(\mathcal{A})-\epsilon)}.$$

Proof

Proof: Apply the law of large numbers to the random variable $\log_2 p(a)$: for any $\epsilon > 0$ and a sequence $a_1 a_2, \ldots, a_n \in A^n$,

$$\lim_{n\to\infty} \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} \log_2 p(a_i) - \mathrm{E}(\log_2 p(a))\right| > \epsilon\right) = 0.$$

With probability $1 - o(1)$, $a_1, \ldots, a_n$ satisfies

$$2^{-n(H(\mathcal{A})+\epsilon)} \leq p(a_1, \ldots, p_n) \leq 2^{-n(H(\mathcal{A})-\epsilon)}.$$

Let $A_{n,\epsilon}$ be the above event. Note that

$$1 \geq p(A_{N,\epsilon}) \geq |A_{n,\epsilon}| 2^{-n(H(\mathcal{A})+\epsilon)}.$$

We get $|A_{n,\epsilon}| \leq 2^{n(H(\mathcal{A})+\epsilon)}$.

continue

Thus $$H_\delta(\mathcal{A}^n) \le \log_2 |A_{n,\epsilon}| \le n(H(\mathcal{A}) + \epsilon).$$

It follows that
$$\lim_{n\to\infty} \limsup \frac{1}{n} H_\delta(\mathcal{A}^n) \le H(\mathcal{A}).$$

continue

Thus $$H_\delta(\mathcal{A}^n) \leq \log_2 |A_{n,\epsilon}| \leq n(H(\mathcal{A}) + \epsilon).$$

It follows that
$$\lim_{n\to\infty} \limsup \frac{1}{n} H_\delta(\mathcal{A}^n) \leq H(\mathcal{A}).$$

Now we prove the lower bound. Let $B_{n,\delta}$ be the minimizer for H_δ; that is, $p(B_{n,\delta}) \geq 1 - \delta$ and $H_\delta(\mathcal{A}^n) = \log_2 |B_{n,\delta}|$.

continue

Thus $$H_\delta(\mathcal{A}^n) \le \log_2 |A_{n,\epsilon}| \le n(H(\mathcal{A}) + \epsilon).$$

It follows that

$$\lim_{n\to\infty} \limsup \frac{1}{n} H_\delta(\mathcal{A}^n) \le H(\mathcal{A}).$$

Now we prove the lower bound. Let $B_{n,\delta}$ be the minimizer for H_δ; that is, $p(B_{n,\delta}) \ge 1-\delta$ and $H_\delta(\mathcal{A}^n) = \log_2 |B_{n,\delta}|$.

For sufficiently large n, we have

$$p(B_{n,\delta} \cap A_{n,\delta}) \ge p(B_{n,\delta}) - \delta \ge 1 - 2\delta.$$

Then

$$|B_{n,\delta} \cap A_{n,\delta}| \ge (1-2\delta) 2^{n(H(\mathcal{A})-\epsilon)}.$$

We have

$$\frac{1}{n} H_\delta(\mathcal{A}^n) \ge \frac{1}{n}\log_2(1-2\delta) + H(\mathcal{A}) - \epsilon. \qquad \square$$