Elementy teorii informacji.

ZNAKOMITY PODRĘCZNIK: T.M. COVER & J.A. THOMAS ELEMENTS OF INFORMATION THEORY

Całkowicie klasycznie: Mamy alfabet składający się z ośmiu symboli. A B C D E F G H. Wiadomości złożone z tych liter chcemy zakodować używając zer i jedynek. Jak to robić:

Trzy bity / symbol

Czy da się lepiej? TO ZALEŻY OD PRAWDOPODOBIEŃSTW

A	000	A	1/2	0
B	001	B	1/4	10
C	010	C	1/8	110
D	011	D	1/16	1110
E	100	E	1/64	111100
F	101	F	1/64	111101
G	110	G	1/64	111110
H	111	H	1/64	111111

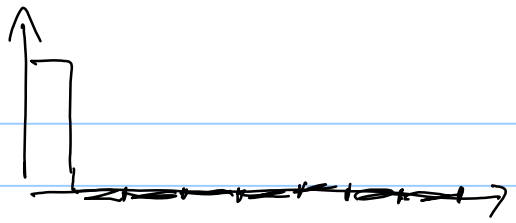
Ile bitów użyjemy średnio na kodowanie?

$$\frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 + \frac{1}{64}\cdot 4\cdot 6$$

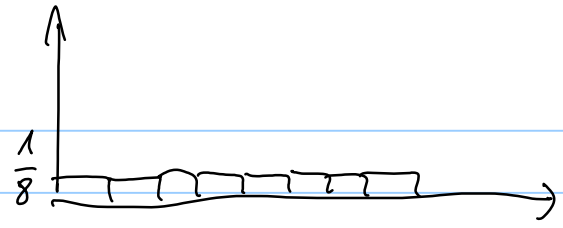
$$= \frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{1}{4} + \frac{3}{8} = 2$$

LEPIEJ.

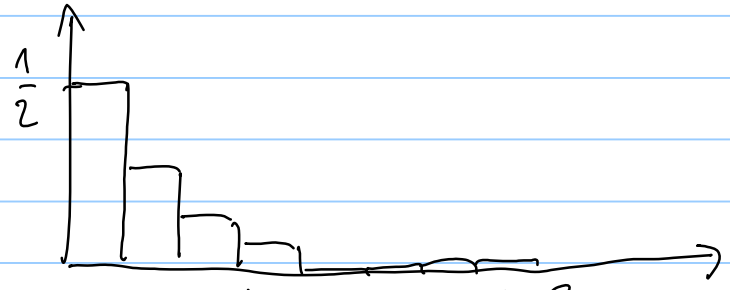
Popatrzmy z innej strony. To, ile bitów jest potrzebne do kodowania pewnej jakości zależy od niepewności statystycznej rozkładu prawdopodobieństwa.



MAŁA NIEOZNACZONOŚĆ



DUŻA NIEOZNACZONOŚĆ



GDZIEŚ POMIĘDZY.

Każdy wynik można traktować jako ZMIENNĄ LOSOWĄ, oznaczmy ją przez X.
Wprowadźmy ENTROPIĘ SHANNONA

$$H(X) = \sum_i p_i \log_2 \frac{1}{p_i} = -\sum_i p_i \log_2 p_i$$

Dla $x \to 0$ mamy $x \log_2 x \to 0$, więc wygodnie w wielu wzorach przyjmować, że $0 \log_2 0 = 0$.

$H = 0$ $\qquad$ $H = 3$

$H = 2$.

Oznacza to, że entropia Shannona jest DOLNYM OGRANICZENIEM na liczbę bitów (symboli) potrzebną do zakodowania wiadomości.
Jak to udowodnić? Nasz kod ma trzy własności: 1) binarny 2) jednoznaczny
3) natychmiastowy (więcej o tym będzie dalej).

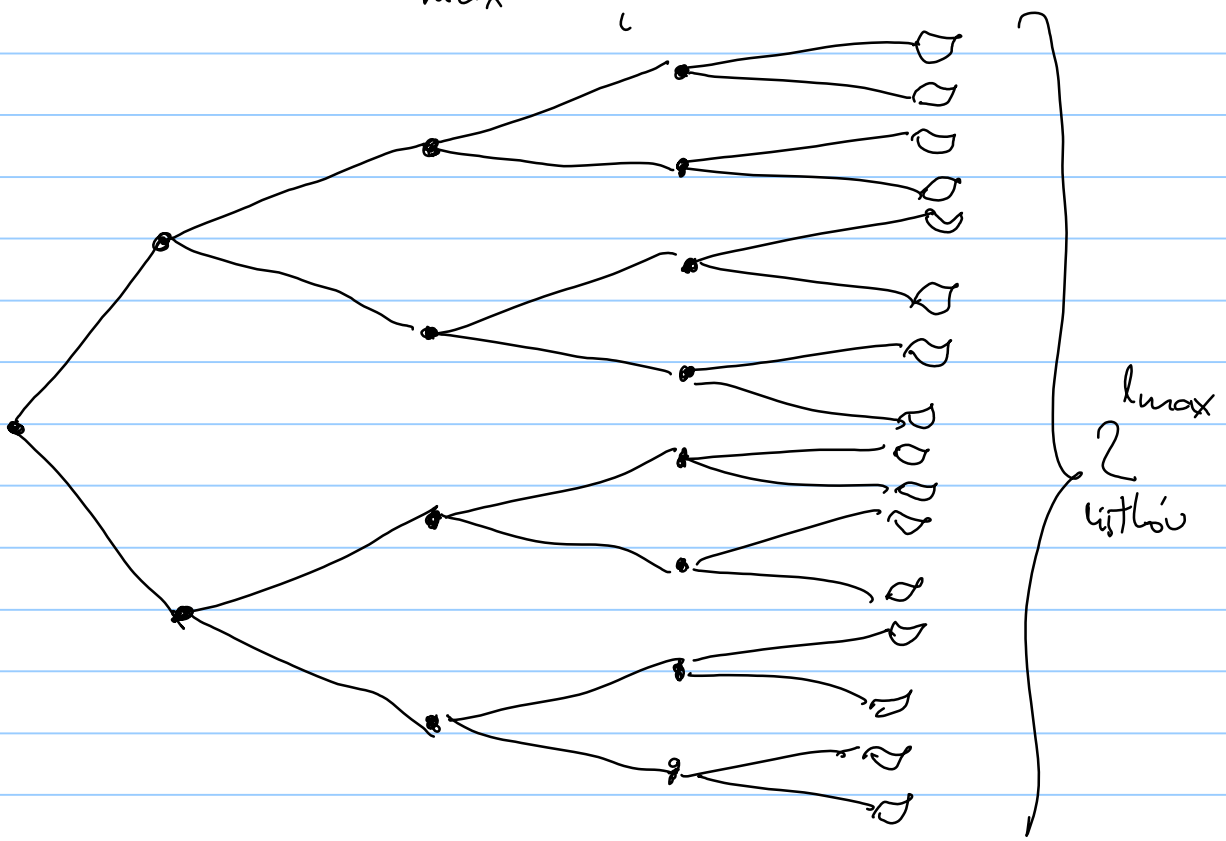
i tylko kodami będącymi permutacjami.
Przykład kodu nie mającego tej własności.

$$A \equiv 0 \quad , \quad B \equiv 01$$

Pytanie Czy istnieje kod o podanych własnościach złożony z K słów o długościach l_i, $i = 1, 2, \ldots, K$?

Odpowiedź istnieje wtedy i tylko wtedy, gdy $\sum_{i=1}^{K} \frac{1}{2^{l_i}} \leq 1$ NIERÓWNOŚĆ KRAFTA

Dowód: Niech $l_{max} = \max_i l_i$



Słowo o długości l_i "zajmuje nam" $2^{l_{max}-l_i}$ liści (inaczej kod nie będzie jednoznaczny / natychmiastowy).

W takim razie:

$$\sum_{i=1}^{K} 2^{l_{max}-l_i} \leq 2^{l_{max}}$$

co daje nam poszukiwaną nierówność. ■

Weźmy teraz zmienną losową X o wartościach x_i występujących z prawdopodobieństwami p_i. Załóżmy, że wartości x_i da się przypisać słowo o długości l_i danej przez

$$\log_2 \frac{1}{p_i} \leq l_i < \log_2 \frac{1}{p_i} + 1$$

Z dolnego ograniczenia mamy:

$\frac{1}{2^{l_i}} \leq p_i$, sumując po i widzimy, że nierówność Krafta jest spełniona.

Średnia długość słowa:

$$H(X) \leq \sum_i p_i l_i < H(X) + 1$$

Jak będziemy blokowo kodować N symboli, to po prawej stronie 1 zostanie zastąpione przez $1/N$, zatem asymptotycznie osiągamy $H(X)$.

ĆWICZENIE: Dla zmiennej losowej przyjmującej K wartości mamy:

$$H(X) \leq \log_2 K$$

A teraz dwie zmienne: X, która przyjmuje wartości x_i oraz Y przyjmuje wartości y_j opisane łącznym rozkładem prawdopodobieństwa $p(x_i, y_j)$.

Entropia Shannona dla pary (X, Y):

$$H(X,Y) = \sum_{ij} p(x_i, y_j) \log_2 \frac{1}{p(x_i, y_j)}$$

Przypuśćmy że zmienne są niezależne, tj. $p(x_i, y_j) = p(x_i) p(y_j)$ łatwo pokazać, że

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

W ogólności mamy oczywiście nierówność.

$$p(x_i) = \sum_j p(x_i, y_j) \qquad p(y_j) = \sum_i p(x_i, y_j)$$

i definiujemy $H(X) = \sum_i p(x_i) \log_2 \frac{1}{p(x_i)}$

$$H(Y) = \sum_j p(y_j) \log_2 \frac{1}{p(y_j)}$$

W ogólności zachodzi:

$$H(Y) \leq H(X,Y) \leq H(X) + H(Y)$$

↖ tu może być dokładnie $H(X)$

Dwa przykłady:

	x_1	x_2	...	x_k
y_1	$\frac{1}{k}$	0	...	0
y_2	0	$\frac{1}{k}$	...	0
⋮	⋮	⋮	⋱	0
y_k	0	0	...	$\frac{1}{k}$

$H(X) = H(Y) = H(X,Y) = \log_2 k$

	x_1	x_2	...	x_k
y_1	$\frac{1}{k^2}$	$\frac{1}{k^2}$		$\frac{1}{k^2}$
y_2	$\frac{1}{k^2}$	$\frac{1}{k^2}$		$\frac{1}{k^2}$
⋮				⋮
y_k	$\frac{1}{k^2}$	$\frac{1}{k^2}$	...	$\frac{1}{k^2}$

$H(X) = H(Y) = \log_2 k$

$H(X,Y) = 2\log_2 k$

Dowód:

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

$$= \sum_j p(y_j) \left(\sum_i p(x_i|y_j) \log_2 \frac{1}{p(x_i|y_j)} \right)$$

Widać, że ≥ 0. Co można wyrazić słownie inaczej jako ŚREDNIĄ ENTROPIĘ USZCZUPŁOCONEGO ROZKŁADU X JEŚLI ZNAMY Y. Oznacza $H(X|Y)$

2) $H(X) + H(Y) - H(X,Y) =$

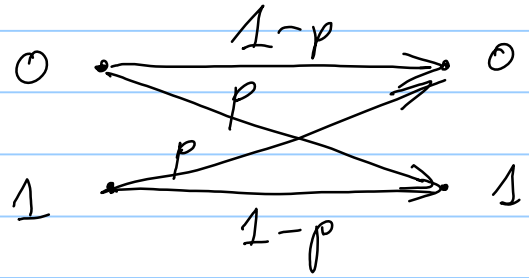
$$= -\sum_{ij} p(x_i, y_j) \log_2 \frac{p(x_i)p(y_j)}{p(x_i, y_j)}$$

$$\geq -\log_2 \left(\sum_{ij} p(x_i, y_j) \frac{p(x_i)p(y_j)}{p(x_i, y_j)} \right) = 0.$$

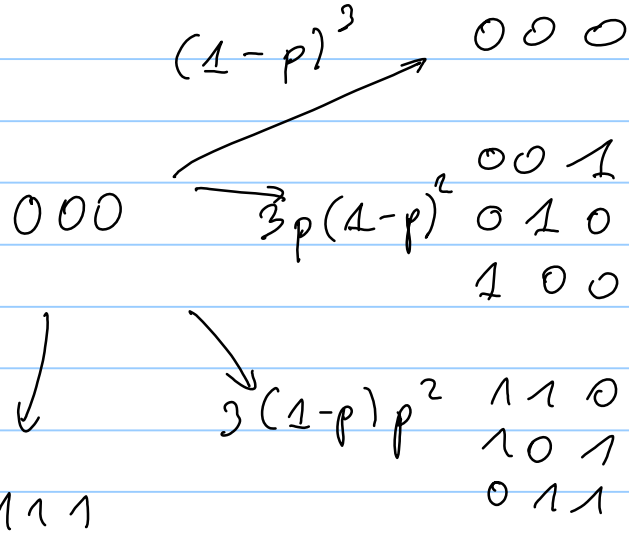
nierówność $-\log_2 t$

ZASZUMIONY KANAŁ

Inny problem: Alicja może wysłać zero i jedynkę, ale po drodze jest szum, który je zamienia z prawdopodobieństwem p.



Czy można zmniejszyć prawdopodobieństwo błędu?
Wysyłamy: 000 lub 111



Prawdopodobieństwo błędu:

$$3p^2 - 2p^3$$

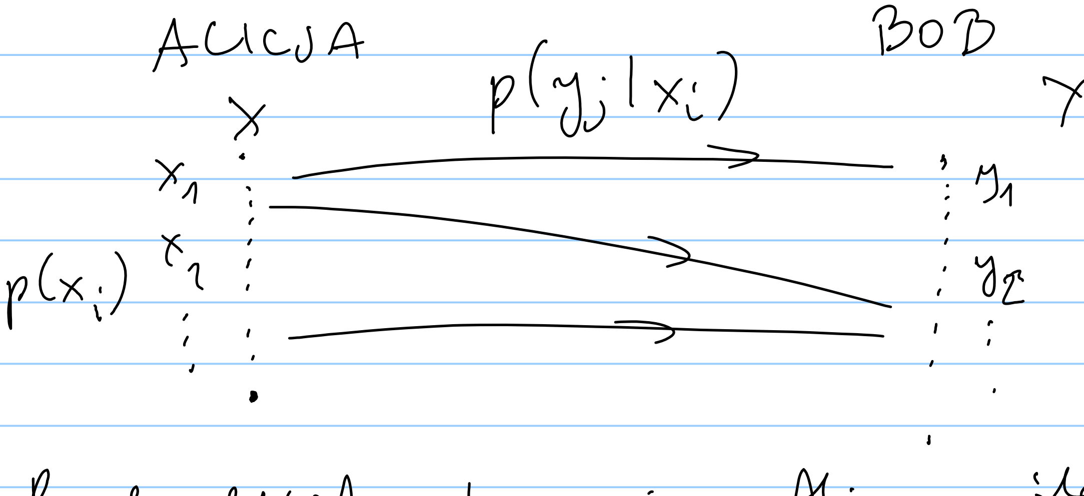
Jeśli p bardzo małe, to liczy się $3p^2$ to się opłaca, np. $p = 10^{-3}$, wtedy

$$3p^2 = 3\times 10^{-6}.$$

Mając do przesłania bardzo długą informację dzięki użyciu takiego kodowania ilość prawdopodobieństwo popełnienia błędu równo „przelatuje" do zera.

Ile informacji możemy przesłać?

Popatrzmy sobie w ten sposób:



Prawdopodobieństwo tego, że Alicja wyśle x_i a Bob dostanie y_j:

$$p(y_j, x_i) = p(y_j|x_i)\,p(x_i)$$

Popatrzmy od strony Boba. Pytanie, jak bardzo niepewność X redukuje w jego oczach znajomość Y?

Korzystamy wtedy:

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

Okazuje się, że to jest ilość informacji, którą do siebie przekazują nasze kanały. Jeśli symbole wejściowe występują z prawdopodobieństwami $\{p(x_i)\}$

Jeśli zmienne niezależne, to $I(X;Y) = 0$.

Jeśli

	x_1	x_2	...
y_1	$\frac{1}{k}$	0	
y_2	0	$\frac{1}{k}$	
⋮			

to

$$I(X;Y) = \log_2 K$$

bo wtedy mamy jednoznaczność.

Ogólnie mamy jeszcze dokonać optymalizacji po rozkładach różnych prawdopodobieństw wyjściowych.

Dojdę do mojej charakterystyki kanału
czyli PRZEPUSTOWOŚCI (CAPACITY)

$$C = \max_{\{p(x_i)\}} I(X;Y)$$

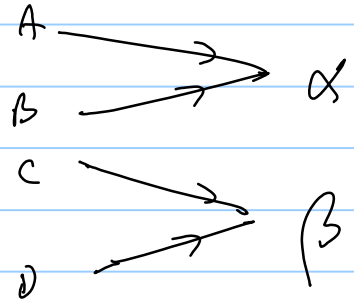
Pojawia przykład:

$p(y_j|x_i)$

	x_1	x_2	...	x_k
y_1	1	0	0	..
y_2	0	1	0	...
.	0	0	1	...
.	.	.	.	

Widać, że $H(X|Y) = 0$. Zatem
$I(X;Y) = H(X)$ i największa jest wtw
$H(X) = \log_2 K = C$.

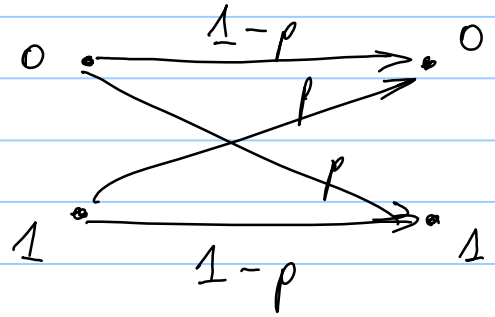
Inny przykład:



$C = 1$

Nie ma różnicy, że wzięliśmy $p_A = p_B = p_C = p_D = \frac{1}{4}$
czy $p_A = p_C = \frac{1}{2}$. Ale $p_A = p_B = \frac{1}{2}$ już
bez sensu do niczego nie nadawało.

Widmy do binarnego kanału symetrycznego:



Wyprowadź jest wzorem:

$$I(X;Y) = H(Y) - H(Y|X)$$

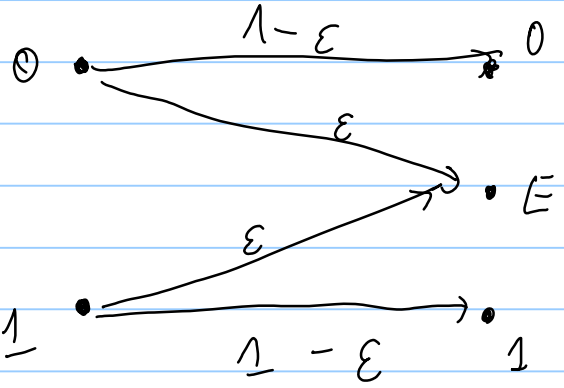
$$H(Y|X) \overset{?}{=} -p\log_2 p - (1-p)\log_2(1-p) = \underbrace{H(p)}_{\text{skalarna wersja wzoru na entropię binarną}}$$

entropia rozkładu binarnego z wartościami p, $1-p$

No a $H(Y) \leq 1$, stąd wynika:

$C = 1 - H(p)$ i symbole wyjściowe należą najwyżej z równym prawdopodobieństwem.

Inny przykład: ERASURE CHANNEL
(kanał z wymazywaniem)



KWANTOWA TEORIA INFORMACJI

Przypuśćmy, że kodujemy informację w stanach $|\psi_i\rangle$ z prawdopodobieństwami p_i

Skonstruujmy

$$\hat{\rho} = \sum_i p_i |\psi_i\rangle\langle\psi_i|$$

Entropię Shannona $H(\{p_i\})$ można zapisać jako:

$$H(\{p_i\}) = -\mathrm{Tr}\left[\hat{\rho}\log_2\hat{\rho}\right] = S(\hat{\rho})$$

ENTROPIA VON NEUMANNA

Widać, że

$$S(\hat{\rho}) = 0 \quad \Leftrightarrow \quad \hat{\rho} = |\psi\rangle\langle\psi|$$

(0 — MINIMALNA WARTOŚĆ)

$$S(\hat{\rho}) \leq \log_2 d$$

dla układu d-wymiarowego.

Entropia von Neumanna mierzy "ilość wiedzy" niezbędnej, która jest "zawarta" w mierzonej postaci.

Operacje pomiarowe: przypuśćmy, że mamy pomiar miary $\hat{M}_r = |u_r\rangle\langle u_r|$. Prawdopodobieństwo otrzymania wyniku r:

$$p_r = \mathrm{Tr}\left(\hat{M}_r\hat{\rho}\right)$$

Tw. $H(\{p_r\}) \geq S(\hat{\rho})$

Dowód:

$$p_r = \sum_i |\langle u_r | \psi_i \rangle|^2 p_i .$$

Funkcja $h(p) = -p \log_2 p$ jest wklęsła.

W takim razie:

$$h(p_r) \geq \sum_i |\langle u_r | \psi_i \rangle|^2 h(p_i)$$

Zatem:

$$H(\{p_r\}) \geq \sum_{ir} |\langle u_r | \psi_i \rangle|^2 h(p_i)$$

$$= \sum_i h(p_i) = S(\hat{\rho})$$

■

Rozważmy następujący problem. Alicja losuje symbole i z prawdopodobieństwem p_i i wysyła Bobowi stan $\hat{\rho}_i$. Bob dokonuje pomiaru $\{\hat{\Pi}_r\}$.

$$p(r, i) = \mathrm{Tr}(\hat{\Pi}_r \hat{\rho}_i) p_i$$

Niech $I(A:B)$ będzie informacją wzajemną dla $p(r, i)$. Niekwantowe protokoły, to jak najlepsze pomiary Boba.

Mamy ograniczenie górne:

$$I(A:B) \le S\left(\sum_i p_i \hat{\rho}_i\right) - p_i S(\hat{\rho}_i)$$

Dla układu d-poziomowego:

$$S\left(\sum_i p_i \hat{\rho}_i\right) \le \log_2 d \quad , \quad S(\hat{\rho}_i) \ge 0.$$

Stąd

$$I(A:B) \le \log_2 d$$

Najlepsze kodowanie informacji klasycznej to przy użyciu stanów ortogonalnych z jednakowymi prawdopodobieństwami.

Ćwiczenie: Oblicz prawą stronę nierówności Holevo dla dwóch stanów nieortogonalnych i porównaj wynik z informacją wzajemną dla pomiarów z minimalnym błędem i jednoznacznych.

Dodatek: Inne spojrzenie na entropię Shannona. Weźmy zmienną binarną $X = 0, 1$ z prawdopodobieństwami $p_0 = p$, $p_1 = 1 - p$.

Rozważmy bardzo długi ciąg N symboli. Oczekujemy, że będzie w nim z grubsza $k_0 \approx N p_0$ zer i $k_1 \approx N p_1$ jedynek.

Prawdopodobieństwo takiego ciągu (ale jedynie o ustalonej kolejności):

$$p(x_1, x_2, \dots, x_N) = p_0^{k_0} p_1^{k_1}$$

$$\approx p_0^{N p_0} p_1^{N p_1} = 2^{-N H(X)}$$

Zbiór ciągów typowych:

$$A_\varepsilon^{(N)} = \left\{ (x_1, \dots, x_N) \;\middle|\; H(X) - \varepsilon \le -\frac{1}{N} \log_2 p(x_1, \dots, x_N) \le H(X) + \varepsilon \right\}$$

Okazuje się, że dla dostatecznie dużych N:

* Prawdopodobieństwo $\left(A_\varepsilon^{(N)}\right) > 1 - \varepsilon$

* $(1-\varepsilon)\, 2^{n(H(X)-\varepsilon)} \leq |A_\varepsilon^{(N)}|$
$$\leq 2^{n(H(X)+\varepsilon)}$$

Ile potrzebujemy do zakodowania?

Typowa sekwencja? Piszemy 0 a potem skorzystamy, co możemy zrobić przy użyciu $n(H(X)+\varepsilon)+1$ bitów.

Nietypowa? Dajemy 1 a potem sekwencję n bitów. Średnio:

$$\overline{\ell^{(n)}} = 1 + \Pr(A_\varepsilon^{(N)})\left[n(H(X)+\varepsilon)+1\right]$$
$$+ \left[1 - \Pr(A_\varepsilon^{(N)})\right] n$$
$$\leq 2 + n(H(X)+\varepsilon) + \varepsilon n$$