Desigualdades de Concentração

Thiago Ramos

1 de outubro de 2024

Licença

1 de outubro de 2024

Sumário

Legenda das Caixas

Caixas desta cor representam resultados importantes, como teoremas, exemplos, etc.

Caixas desta cor representam observações importantes, como intuição dos problemas, conexões com outros resultados, etc.

Capítulo 1

Desigualdades de Concentração

As principais referências foram:

1. [?]
2. [?]
3. [?]

1.1 Desigualdades Básicas

1.1.1 Markov e Chebyshev

Seja $f \geq 0$ monotonamente crescente e $X \geq 0$, é fácil ver que

$$\mathbb{P}(X > \varepsilon) \leq \mathbb{P}(f(X) > f(\varepsilon)) = \mathbb{E}1_{f(X)>f(\varepsilon)} \leq \frac{\mathbb{E}(f(X))}{f(\varepsilon)}. \tag{1.1.1}$$

Desse simples fato, conseguimos derivar várias desigualdades interessantes, como por exemplo:

- (Markov) Para qualquer X, não necessariamente positiva, podemos fazer o seguinte
$$\mathbb{P}(|X| > \varepsilon) \leq \frac{\mathbb{E}(|X|)}{\varepsilon}.$$
- (Chebyshev) Tomando $|X - \mathbb{E}X|$ como v.a. e $f(x) = x^2$, temos que
$$\mathbb{P}(|X - \mathbb{E}X| > \varepsilon) \leq \frac{\mathbb{E}(|X - \mathbb{E}X|^2)}{\varepsilon^2} = \frac{\mathbb{V}\text{ar}X}{\varepsilon^2}.$$

Observação 1.1.1. Perceba que essas desigualdades de concentração servem para construirmos intervalos de confiança. Por exemplo, por Markov,

$$\mathbb{P}(X > t) \leq f(t)^{-1}\mathbb{E}f(X),$$

isto é,

$$\mathbb{P}(X \leq t) \geq 1 - f(t)^{-1}\mathbb{E}f(X),$$

portanto, encontrando $t > 0$ tal que $f(t)^{-1}\mathbb{E}f(X) = \epsilon$, temos que com probabilidade $1 - \epsilon$

$$X \leq t.$$

Exemplo 1.1.2. Suponha que $\mathbb{E}(X) < \infty$ e que $\sigma^2 = \mathbb{V}\text{ar}(X) < \infty$. Considere

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n},$$

onde $X,\ X_i$ são i.i.d. para todo i. Note que

$$\mathbb{E}(\bar{X}_n) = \frac{n}{n}\mathbb{E}(X) = \mathbb{E}(X)$$

e que

$$\mathbb{V}\text{ar}(\bar{X}_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Logo, pela desigualdade de Chebyshev,

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(X)| > \varepsilon) \leq \frac{\mathbb{V}\text{ar}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2},$$

então, X_n converge em probabilidade para $\mathbb{E}(X)$ quando n cresce. Ou seja, provamos a lei fraca dos grandes números para o caso específico em X_i são i.i.d. com variância finita.

1.1.2 Chernoff

Seja X uma v.a. e $M_x(t)$ sua função geradora de momento. Utilizando **??** com $f(x) = e^{tx}$, temos que

$$\mathbb{P}(X > \varepsilon) \leq e^{-\varepsilon t}\mathbb{E}(e^{Xt}) = e^{-\varepsilon t + \ln M_X(t)}. \tag{1.1.2}$$

O método de Chernoff consiste em minimizar o lado direito da expressão acima em t. Note que a escolhe de f dessa forma é bem interessante, primeiro porque o decaimento de $e^{-t\varepsilon}$ é bem rápido em t. Além disso, muitas vezes temos uma fórmula explícita para $\mathbb{E}(e^{Xt})$ já que esta é a função geradora de momento de X.

Exemplo 1.1.3 (Bernoulli)**.** Temos que

$$\begin{aligned}\mathbb{P}(X > \varepsilon) &\leq e^{-\varepsilon t}\mathbb{E}(e^{Xt})\\ &\leq e^{-\varepsilon t + \ln(pe^t + 1 - p)}\end{aligned}$$

e o lado direito é mínimo quando

$$t = \ln\left[\left(\frac{\varepsilon}{1-\varepsilon}\right)\left(\frac{1-p}{p}\right)\right]$$

Exemplo 1.1.4 (Normal)**.** Temos que

$$\begin{aligned}\mathbb{P}(X > \varepsilon) &\leq e^{-\varepsilon t}\mathbb{E}(e^{Xt})\\ &\leq e^{-\varepsilon t + \ln(e^{t\mu + t^2\sigma^2/2})}\end{aligned}$$

e o lado direito é mínimo quando

$$t_0 = \frac{\varepsilon - \mu}{\sigma^2}.$$

E portanto, o método de Chernoff nos dá o seguinte bound:

$$\begin{aligned}\mathbb{P}(X > \varepsilon) &\leq e^{-\varepsilon t_0}\mathbb{E}(e^{Xt_0})\\ &= e^{-\varepsilon t_0} e^{t_0\mu + t_0^2\sigma^2/2}\\ &= e^{-\varepsilon(\varepsilon-\mu)/\sigma^2} e^{(\varepsilon-\mu)/\sigma^2\mu + ((\varepsilon-\mu)/\sigma^2)^2\sigma^2/2}\\ &= e^{-(\varepsilon-\mu)^2/2\sigma^2}.\end{aligned}$$

No caso em que $\mu = 0$, temos que

$$\mathbb{P}(X > \varepsilon) \leq e^{-\varepsilon^2/2\sigma^2}.$$

Observação 1.1.5. O bound encontrado acima

$$\mathbb{P}(X > \varepsilon) \leq e^{-\varepsilon^2/2\sigma^2}$$

é ótimo a menos de uma constante 1/2, isto é, temos que

$$\sup_{\varepsilon>0} \mathbb{P}(X > \varepsilon) e^{\varepsilon^2/2\sigma^2} = 1/2.$$

Para provarmos isso, note que

$$\begin{aligned}
\mathbb{P}(X > \varepsilon) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_\varepsilon^\infty e^{-x^2/2\sigma^2} dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty e^{-(x+\varepsilon)^2/2\sigma^2} dx \\
&\leq \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty e^{-(x^2+\varepsilon^2)/2\sigma^2} dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^\infty e^{-x^2/2\sigma^2} dx e^{-\varepsilon^2} \\
&= \mathbb{P}(X > 0) e^{-\varepsilon^2} = 1/2 e^{-\varepsilon^2},
\end{aligned}$$

e como $\mathbb{P}(X > 0) e^{0^2/2\sigma^2} = 1/2$, temos o resultado.
Ou seja, mostrarmos que apesar da técnica aplicada ser bem simples, conseguimos um resultado sharp para o caso Normal.

Definição 1.1.6. Seja X centrada, dizemos que X é sub-Gaussiana com fator v se $M_X(t) \leq M_N(t)$, onde N tem distribuição $N(0, v)$, isto é,

$$M_X(t) \leq e^{t^2 v/2}.$$

Corolário 1.1.7. *Se X é sub-Gaussiana com fator v, entaõ $\mathbb{V}ar(X) \leq v$.*

Demonstração. Temos que

$$\mathbb{V}\text{ar}(X) = (M_X(t)'')|_{t=0} = (e^{t^2 v/2}(tv)^2 + e^{t^2 v/2} v)|_{t=0} = v.$$

■

Exemplo 1.1.8 (Sub-Gaussianas)**.** Seja X uma v.a. centrada sub-gaussiana com fator v, e Y uma v.a. normal $N(0, v)$, então

$$\begin{aligned}
\mathbb{P}(X > \varepsilon) &\leq e^{-\varepsilon t} M_X(t) \\
&\leq e^{-\varepsilon t} M_Y(t). \\
&\leq e^{-\varepsilon t} M_Y(t_0) \\
&\leq e^{-\varepsilon^2/2v}.
\end{aligned}$$

Isto é, a calda de uma v.a. sub-gaussiana é limitada pela gaussiana de mesma variância, já que o mesmo vale para $\mathbb{P}(-X > \varepsilon)$

Exemplo 1.1.9 (Soma i.i.d)**.** Suponha que $X_1, \ldots, X_n$ é i.i.d. e defina

$$S_n = X_1 + \cdots + X_n.$$

Temos então que

$$\begin{aligned}
\mathbb{P}(S_n > \varepsilon) &\leq e^{-\varepsilon t}\mathbb{E}(e^{S_n t}) = e^{-\varepsilon t}\mathbb{E}(e^{t(X_1+\cdots+X_n)}) \\
&= e^{-\varepsilon t}\mathbb{E}(e^{t(X_1)})\mathbb{E}(e^{t(X_2)})\cdots\mathbb{E}(e^{t(X_n)}) \\
&= e^{-\varepsilon t}\mathbb{E}(e^{t(X_1)})^n \\
&= e^{-\varepsilon t + n\ln(M_X)}.
\end{aligned}$$

E daí conseguimos encontrar bounds superiores muito facilmente.

1.1.3 Chebyshev-Cantelli

Por Chebyshev, temos que

$$\mathbb{P}(X - \mathbb{E}X > t) = \mathbb{P}(X - \mathbb{E}X + a > t + a) \leq \frac{\mathbb{V}\mathrm{ar}(X) + a^2}{(t+a)^2},$$

e minimizando em a, temos que a expressão do lado direito é miníma quando

$$a' = \frac{\mathbb{V}\mathrm{ar}X}{t}.$$

Substituindo, temos que

$$\begin{aligned}
\mathbb{P}(X - \mathbb{E}X > t) &\leq \frac{\mathbb{V}\mathrm{ar}(X) + a^2}{(t+a)^2} \\
&\leq \frac{\mathbb{V}\mathrm{ar}(X) + (\mathbb{V}\mathrm{ar}X/t)^2}{(t + \mathbb{V}\mathrm{ar}X/t)^2} \\
&= \frac{\mathbb{V}\mathrm{ar}X}{t^2 + \mathbb{V}\mathrm{ar}X}.
\end{aligned}$$

Isto é,

$$\boxed{\mathbb{P}(X - \mathbb{E}X > t) = \frac{\mathbb{V}\mathrm{ar}X}{t^2 + \mathbb{V}\mathrm{ar}X}.}$$

1.1.4 Paley-Zygmund

Suponha que $\mathbb{E}(X^2) < \infty$. Sabemos que

$$\mathbb{E}(X) = \mathbb{E}(X 1_{X<s}) + \mathbb{E}(X 1_{X\geq s}),$$

logo,

$$\begin{aligned}\mathbb{E}(X) &\leq \mathbb{E}(X1_{X\geq a\mathbb{E}(X)}) + a\mathbb{E}(X)\cdot P(X\geq 0)\\ &\leq \mathbb{E}(X1_{X\geq a\mathbb{E}(X)}) + a\mathbb{E}(X)\cdot 1,\end{aligned}$$

isto é,

$$(1-a)\mathbb{E}(X) \leq \mathbb{E}(X1_{X\geq a\mathbb{E}(X)}).$$

Por Holder, temos que

$$(1-a)\mathbb{E}(X) \quad \leq \mathbb{E}(X1_{X\geq a\mathbb{E}(X)}) \quad \leq \sqrt{\mathbb{E}(X^2)\mathbb{P}(X\geq a\mathbb{E}(X))}$$

isto é,

$$(1-a)^2\mathbb{E}(X)^2 \leq \mathbb{E}(X^2)\mathbb{P}(X\geq a\mathbb{E}(X))$$

que é equivalente a

$$\boxed{\mathbb{P}(X\geq a\mathbb{E}(X)) \geq (1-a)^2\frac{\mathbb{E}(X)^2}{\mathbb{E}(X^2)}}$$

1.1.5 Média-Mediana

Definimos a mediana MX de uma v.a. X como o valor tal que $\mathbb{P}(X\geq MX)\geq 1/2$ e $\mathbb{P}(X\leq MX)\geq 1/2$, isto é, MX divide a distribuição em duas partes iguais.

Lema 1.1.10. *Temos que MX minimiza a norma L_1, isto é*

$$MX = \arg\min_c \mathbb{E}|X-c|.$$

Demonstração. Basta notar que

$$\mathbb{E}(|X-c|) = \int_c^{+\infty}\mathbb{P}(X\geq t)dt + \int_{-\infty}^{c}\mathbb{P}(X\leq t)dt,$$

portanto, derivando em c temos que

$$\begin{aligned}\frac{d\mathbb{E}(|X-c|)}{dc} &= \mathbb{P}(X\geq +\infty) - \mathbb{P}(X\geq c) + \mathbb{P}(X\leq c) - \mathbb{P}(X\leq -\infty)\\ &= -\mathbb{P}(X\geq c) + \mathbb{P}(X\leq c),\end{aligned}$$

que é nulo quando $c = MZ$, positivo se $c > MZ$ e negativo se $c < MZ$. ■

Utilizando o lema acima e Cauchy-Schwarz, temos o seguinte resultado:

$$\boxed{\begin{aligned}|\mathbb{E}X - MX| &= |\mathbb{E}(X-MX)|\\ &\leq \mathbb{E}|X-MX|\\ &\leq \mathbb{E}|X-\mathbb{E}X|\\ &\leq \sqrt{\mathbb{E}[(X-\mathbb{E}X)^2]}\\ &= \sqrt{\mathbb{V}\mathrm{ar}X}.\end{aligned}}$$

1.1.6 Hoeffding

Para provarmos Hoeffding, faremos uso da técnica para limitantes de Chernoff.

Lema 1.1.11. *Seja X uma v.a. com $\mathbb{E}X = 0$* [1] *e tal que $X \in [a, b]$, temos então que:*

$$\mathbb{E}e^{tX} \leq \exp \frac{t^2(b-a)^2}{8}, \; t > 0.$$

Isso significa que toda v.a. limitada é Sub-Gaussiana!

Demonstração. Vamos usar o fato de que $f(x) = e^{tx}$ é convexa. Temos então que

$$e^{tx} \leq \frac{b-x}{b-a}e^{at} + \frac{-a+x}{b-a}e^{bt},$$

tomando $\mathbb{E}$ na expressão acima, temos que

$$\mathbb{E}(e^{tX}) \leq \frac{b}{b-a}e^{at} - \frac{a}{b-a}e^{bt},$$

já que $\mathbb{E}X = 0$. Defina então

$$\begin{aligned}
\phi(t) &= \ln\left(\frac{b}{b-a}e^{at} - \frac{a}{b-a}e^{bt}\right) \\
&= \ln\left(e^{at}\left(\frac{b}{b-a} - \frac{a}{b-a}e^{(b-a)t}\right)\right) \\
&= \ln\left(\frac{b}{b-a} - \frac{a}{b-a}e^{(b-a)t}\right) + at \\
&= \ln\left(B - Ae^{(b-a)t}\right) + at
\end{aligned}$$

e então temos que

$$\mathbb{E}(e^{tX}) \leq e^{\phi(t)}.$$

Com um pouco de cálculo, conseguimos mostrar que

$$\begin{aligned}
\phi'(t) &= a + \frac{-A(b-a)e^{t(b-a)}}{B - Ae^{t(b-a)}} \\
&= a + \frac{-ae^{t(b-a)}}{B - Ae^{t(b-a)}}
\end{aligned}$$

e que

$$\phi''(t) \leq \frac{(b-a)^2}{8},$$

logo por Taylor, existe $\theta \in [0, t]$ tal que

[1] Lembre-se que sempre podemos nos reduzir a esse caso fazendo $X' = X - \mathbb{E}X$.

$$\phi(t) = 0 + t0 + t^2\phi(\theta) \le t^2\frac{(b-a)^2}{8},$$

e portanto

$$\mathbb{E}(e^{tX}) \le \exp\left(t^2\frac{(b-a)^2}{8}\right).$$

■

Teorema 1.1.12 (Desigualdade de Hoeffding)**.** *Sejam $X_1, X_2, \ldots, X_n$ independentes, tal que $X_i \in [a_i, b_i]$, então dado $\varepsilon > 0$, temos que*

$$\mathbb{P}(S_m - \mathbb{E}(S_m) > \varepsilon) \le \exp(-2\varepsilon^2/\sum_{i=1}^{n}(b_i - a_i)^2)$$

Demonstração. Por **??** e o lema anterior, temos que

$$\begin{aligned}
\mathbb{P}(S_m - \mathbb{E}(S_m) > \varepsilon) &\le \exp(t\varepsilon)^{-1}\mathbb{E}(\exp(S_m - \mathbb{E}(S_m))) \\
&= \exp(t\varepsilon)^{-1}\mathbb{E}(\exp\sum_{i=1}^{n} X_i - \mathbb{E}X_i) \\
&\le \exp(t\varepsilon)^{-1}\prod_{i=1}^{n} e^{t^2(b_i-a_i)^2/8} \\
&\le e^{-t\varepsilon+\sum_{i=1}^{n} t^2(b_i-a_i)^2/8}
\end{aligned}$$

por um simples argumento de cálculo, é fácil ver que o mínimo em t da expressão do lado direito é atingido em

$$t_0 = \frac{\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2/4},$$

ficando assim com

$$\begin{aligned}
\mathbb{P}(S_m - \mathbb{E}(S_m) > \varepsilon) &\le e^{-t_0\varepsilon + t^2\sum_{i=1}^{n}(b_i-a_i)^2/8} \\
&= e^{-2\varepsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2}
\end{aligned}$$

■

Observação 1.1.13. Suponha que X_i são i.i.d.'s centradas e que $0 \leq X_i \leq 1$ para todo i. Um chute lógico para o comportamento de S_n seria dizer que $S_n \in O(n)$, já que estamos somando n v.a. com tamanho máximo de 1. Porém, o teorema acima nos diz que

$$\mathbb{P}(S_n > \alpha\sqrt{n}) \leq e^{-2\alpha^2},$$

ou seja, temos que na verdade S_n é da ordem de $O(\sqrt{n})$. Assim como dito em [**?**],
"The basic intuition here is that it is difficult for a large number of independent variables $X_1, ..., X_n$ to "work together" to simultaneously pull a sum $X_1 + \cdots + X_n$ [...] too far away from its mean. Independence here is the key; concentration of measure results typically fail if the X_i are too highly correlated with each other."

Exemplo 1.1.14 (Classificação)**.** Considere $(X, Y) \sim P$ com $Y \in \{0, 1\}$ e $h : x \mapsto \{0, 1\}$ um classificador. Definimos o risco de h como

$$R(h) = \mathbb{E}(1_{h(X) \neq Y}) = \mathbb{P}(h(X) \neq Y).$$

Considere agora uma amostra $(X_1, Y_1), \ldots, (X_n, Y_n)$ e defina o risco empírico de h como

$$\hat{R}_n(h) = \frac{1}{n}\sum_{i}^{n} 1_{h(X_i) \neq Y_i}.$$

Note que $1_{h(X_i) \neq Y_i} \in [0, 1]$, então por Hoeffding:

$$\mathbb{P}(|\hat{R}_n(h) - R(h)| > \varepsilon) \leq \exp(-2(ne)^2/n),$$

então dado $\delta > 0$, com probabilidade $1 - \delta$, temos que

$$|\hat{R}_n(h) - R(h)| \leq \sqrt{\frac{\ln(1/\delta)}{2n}}$$

Exemplo 1.1.15 (Problema com Hoeffding)**.** Seja $X \in \{-1, 0, 1\}$ uma v.a. tal que

$$\begin{aligned} \mathbb{P}(X = 1) &= p/2 \\ \mathbb{P}(X = -1) &= p/2 \\ \mathbb{P}(X = 0) &= 1 - p \end{aligned}$$

É fácil ver que $\mathbb{E}(X) = 0$ e que $\mathbb{V}\text{ar}(X) = p$.
Tomando $S_n = X_1 + \cdots + X_n$ com X_i i.i.d. a X, temos por Hoeffding,

$$\mathbb{P}(S_n > \varepsilon) \leq e^{-\varepsilon^2/2n}.$$

Note que o bound encontrado não depende de p, entretanto, intuitivamente esperamos que quanto menor for a variância $\mathbb{V}\text{ar}(X) = p$, mais concentrada seja S_n.
Note que esse problema com Hoeffding surge porque essa desigualdade não leva em consideração momentos de ordem maior.

1.1.7 Bernstein

A desigualdade de Bernstein nos ajuda a resolver o problema que vimos no exemplo **??**, adicionando informação extra sobre a variância das v.a.s para melhorar nossos bounds.

Lema 1.1.16. *Suponha que* $|X| < c$, $\mathbb{E}(X) = 0$ *e que* $\mathbb{V}ar(X) = \sigma^2$. *Então*

$$\mathbb{E}(e^{tX}) \leq \exp \frac{\sigma^2}{c^2}(e^{tc} - 1 - tc).$$

Demonstração. Temos que

$$\begin{aligned} \mathbb{E}(e^{tX}) &= \mathbb{E}\left(1 + Xt + \sum_{k=2}^{\infty} \frac{t^k X^k}{k!}\right) \\ &= 1 + 0 + \sum_{k=2}^{\infty} \frac{t^k \mathbb{E}(X^k)}{k!} \\ &= 1 + \sum_{k=2}^{\infty} \frac{t^k \mathbb{E}(X^2 X^{k-2})}{k!} \\ &\leq 1 + \sum_{k=2}^{\infty} \frac{t^k \sigma^2 c^{k-2}}{k!} \\ &\leq 1 + \frac{\sigma^2}{c^2} \sum_{k=2}^{\infty} \frac{t^k c^k}{k!} \\ &= 1 + \frac{\sigma^2}{c^2}(e^{tc} - 1 - tc) \\ &\leq \exp \frac{\sigma^2}{c^2}(e^{tc} - 1 - tc) \end{aligned}$$

■

A primeira vista, pode parecer estranho a passagem

$$1 + \frac{\sigma^2}{c^2}(e^{tc} - 1 - tc) \leq \exp \frac{\sigma^2}{c^2}(e^{tc} - 1 - tc),$$

já que o lado esquerdo já nos dá um bound e perdemos alguma coisa nessa desigualdade. Porém, lembre-se que o que precisamos estudar é $\ln \mathbb{E}(e^{tX})$, por isso é interessante fazer aparecer um termo exponencial que vai ser simplificado depois utilizando o logaritmo.

Teorema 1.1.17 (Bernstein)**.** *Sejam X_i independentes, com $|X_i| \leq c$, $\mathbb{E}(X_i) = 0$ e $\mathbb{V}ar(X_i) = \sigma_i^2$. Então*

$$\mathbb{P}(S_n > \varepsilon) \leq \exp\left(\frac{-\varepsilon^2/2}{s + \varepsilon c/3}\right).$$

Demonstração. Basta fazer a mesma coisa de sempre. Temos que

$$\mathbb{P}(S_n > \varepsilon) \leq \exp(-\varepsilon t + (e^{tc} - 1 - tc)(\sum \sigma_i^2)/c^2).$$

O mínimo é atingido quando

$$t = \frac{1}{c} \ln\left(1 + \varepsilon c / \sum \sigma_i^2\right).$$

Logo, se $s_n = \sum \sigma_i^2$

$$\begin{aligned} \mathbb{P}(S_n > \varepsilon) &\leq \exp(-\frac{\varepsilon}{c}\ln(1 + \varepsilon c/s_n) + \frac{\varepsilon}{c} - \frac{s_n}{c^2}\ln(1 + \varepsilon c/s_n)) \\ &\leq \exp(-h(\varepsilon c/s_n)s_n/c^2), \end{aligned}$$

onde

$$h(x) = (1+x)\ln(1+x) - x.$$

Mas temos a seguinte relação:

$$h(x) \geq \frac{x^2}{2(1 + x/3)},$$

logo,

$$\begin{aligned} \mathbb{P}(S_n > \varepsilon) &\leq \exp\left(-\frac{s_n}{2c^2}\left(\frac{(\varepsilon c/s_n)^2}{1 + (\varepsilon c/s_n)/3}\right)\right) \\ &= \exp\left(-\frac{1}{2s_n}\left(\frac{\varepsilon^2}{1 + \varepsilon c/3s_n}\right)\right) \\ &= \exp\left(\frac{-\varepsilon^2}{2s_n + 2\varepsilon c/3}\right). \end{aligned}$$

■

Escrevendo a expressão acima em termos de $\sqrt{s_n}$, ficamos com

$$\mathbb{P}(S_n > \varepsilon\sqrt{s_n}) \leq \exp\left(\frac{-\varepsilon^2}{2 + 2\varepsilon c/3\sqrt{s_n}}\right).$$

Logo, escolhendo ε tal que $\varepsilon \in O(\sqrt{s_n})$, temos um decaimento ε^{-2}. Caso contrário, temos um decaimento pior do que o encontrado em Hoeffding, isto é, da ordem ε^{-1}.

Exemplo 1.1.18 (Problema com Hoeffding)**.** Seja $X \in \{-1, 0, 1\}$ uma v.a. tal que

$$\begin{aligned} \mathbb{P}(X = 1) &= p/2 \\ \mathbb{P}(X = -1) &= p/2 \\ \mathbb{P}(X = 0) &= 1 - p \end{aligned}$$

Já sabemos que $|X| \leq 1$, $\mathbb{E}(X) = 0$ e que $\mathbb{V}\mathrm{ar}(X) = p$. Por Bernstein, temos que

$$\mathbb{P}(S_n > \varepsilon) \leq \exp\left(\frac{-\varepsilon^2/2}{np + \varepsilon/3}\right),$$

da mesma forma,

$$\mathbb{P}(S_n > \varepsilon\sqrt{np}) \leq \exp\left(\frac{-\varepsilon^2/2}{1 + \varepsilon/3\sqrt{np}}\right).$$

1.1.8 Hoeffding Novamente

Note que poderíamos ter usado a mesma técnica que utilizamos em Bernstein para provarmos Hoeffding. É claro que o esperado é que conseguimos um bound pior (pense por que). Mas vamos fazer as contas por curiosidade.

É fácil ver que, se $|X| \leq c$ e $\mathbb{E}(X) = 0$, então utilizando a técnica de Bernstein, temos que

$$\mathbb{E}(e^{tX}) \leq 1 + e^{tc} - tx - 1 \leq \exp(e^{tc} - tc - 1),$$

logo,

$$\begin{aligned} \mathbb{P}(S_n > \varepsilon) &\leq \exp(-\varepsilon t + \ln \mathbb{E}(e^{tS_n})) \\ &\leq \exp(-\varepsilon t + n(e^{tc} - tc - 1). \end{aligned}$$

O valor acima atinge o mínimo quando

$$t = \frac{1}{c}\ln(1 + \varepsilon/cn),$$

e então temos que

$$\begin{aligned}\mathbb{P}(S_n > \varepsilon) &\leq \exp(-n[(1+\varepsilon/cn)\ln(1+\varepsilon/cn) - \varepsilon/cn]) \\ &\leq \exp \frac{-\varepsilon^2}{(2nc^2)(1+\varepsilon/3cn)} \\ &\leq \exp \frac{-\varepsilon^2}{2nc^2 + \varepsilon c2/3},\end{aligned}$$

logo,

$$\mathbb{P}(S_n > \varepsilon\sqrt{n}) \leq \exp\left(\frac{-\varepsilon^2/2}{2c^2 + \varepsilon/3\sqrt{n}}\right),$$

1.1.9 Desigualdade de Azuma e das Diferenças Limitadas

Definição 1.1.19. Seja X_n uma sequência de v.a.'s adaptada à uma filtração $\mathcal{F}_n$. Dizemos que X_n possui a propriedade da diferença de Martingales se

$$\mathbb{E}(X_n|\mathcal{F}_{n-1}) = 0.$$

Lema 1.1.20. *Seja X v.a. tal que $\mathbb{E}(X|\mathcal{G}) = 0$ e $X \in [a,b]$. Então*

$$\mathbb{E}(e^{tV}|\mathcal{G}) \leq e^{t^2(b-a)/8}.$$

Demonstração. Basta fazer o mesmo que fizemos na desigualdade de Hoeffiding, porém dessa vez usamos a condicional $\mathbb{E}(\,\cdot\,|\mathcal{G})$ ao invés de apenas $\mathbb{E}$. ■

Teorema 1.1.21 (Desigualdade de Azuma)**.** *Seja X_n um Martingale com respeito a uma filtração $\mathcal{F}_n$. Considere $Y_i := X_i - X_{i-1}$ e suponha que para todo i exista $c_i \geq 0$ tal que*

$$|X_i - X_{i-1}| \leq c_i.$$

Então, para todo $\varepsilon > 0$ e m temos que

$$\mathbb{P}\left(X_n - X_0 \geq \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum c_i}\right).$$

Demonstração. Note que

$$\mathbb{E}(X_i - X_{i-1}|\mathcal{F}_{i-1}) = 0,$$

já que X_{i-1} é $\mathcal{F}_{i-1}$ mensurável e também é um Martingale. Logo, Y_i se encaixa no lema anterior se tomarmos $\mathcal{G} = \mathcal{F}_{i-1}$. Vamos utilizar a técnica de Chernoff novamente. Temos que

$$\begin{aligned}\mathbb{P}(X_n - X_0 \geq \varepsilon) &\leq e^{-t\varepsilon}\mathbb{E}(e^{X_n - X_0}) \\ &= e^{-t\varepsilon}\mathbb{E}(e^{\sum_{i=1}^n Y_i}) \\ &= e^{-t\varepsilon}\mathbb{E}\mathbb{E}((e^{\sum_{i=1}^n Y_i})|\mathcal{F}_{n-1}) \\ &= e^{-t\varepsilon}\mathbb{E}(e^{\sum_{i=1}^{n-1} Y_i}\mathbb{E}(e^{Y_n}|\mathcal{F}_{n-1})) \\ &\leq e^{-t\varepsilon}\mathbb{E}(e^{\sum_{i=1}^{n-1} Y_i}e^{t^2 c_n/8}) \\ &\leq e^{-t\varepsilon}e^{t^2(\sum_{i=1}^n c_i)/8}\end{aligned}$$

que é minimizado quando $t = 4\varepsilon / \sum c_i$.

∎

Note que a mesma prova anterior é verdadeira para:

Teorema 1.1.22 (Desigualdade de Azuma - 2)**.** *Seja (Y_n) uma sequência de v.a. e $\mathcal{F}_n$ uma filtração para Y_n tal que*

$$\mathbb{E}(Y_i|\mathcal{F}_{i-1}) = 0.$$

Suponha que para todo i exista $c_i \geq 0$ tal que

$$|Y_i| \leq c_i.$$

Então, para todo $\varepsilon > 0$ e m temos que

$$\mathbb{P}\left(|S_n| \geq \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum c_i^2}\right),$$

onde $S_n = Y_1 + \cdots + Y_n$.

Isto é, assim como Hoeffding, estamos estudando o comportamento de de somas de v.a. limitadas, porém agora ao invés de pedirmos independência, estamos pedindo a propriedade da diferença de Martingales.

Definição 1.1.23. Dizemos que uma função $f(x_1, \ldots, x_n)$ possui a propriedade das diferenças limitadas se existem constantes $c_1, \ldots, c_n$ tal que fixadas $x_1, \ldots, x_n$

$$\sup_{x_i'} |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i.$$

Teorema 1.1.24 (Desigualdade das diferenças limitadas/McDiarmid)**.** *Suponha que $f : \mathcal{X} \to \mathbb{R}$ satisfaça a propriedade das diferenças limitadas com constantes c_i. Então dado X_i independentes*

$$\mathbb{P}(|Z - \mathbb{E}Z| > \varepsilon) \leq \exp\left(\frac{-2\varepsilon^2}{\sum c_i^2}\right),$$

onde

$$Z = f(X_1, \ldots, X_n).$$

Se X_i são independentes, então $f(X_1, \ldots, X_n)$ está próxima de sua média se $f(x_1, \ldots, x_n)$ não é sensível a mudanças em suas coordenadas isoladamente.

Demonstração. Considere o Martingale

$$W_i = \mathbb{E}(Z|X_1, \ldots, X_i),$$

e a diferença entre Martingales

$$Y_i = W_i - W_{i-1}.$$

Note que para pontos $a_1, \dots, a_i$, temos que

$$
\begin{aligned}
&|\mathbb{E}(f(a_1, \dots, a_i, X_{i+1}, \dots, X_n)) - \mathbb{E}(f(a_1, \dots, a_{i-1}, X_i, \dots, X_n))| \\
&\leq \int |f(a_1, \dots, a_{i-1}, a_i, x_{i+1}, \dots, x_n) - f(a_1, \dots, a_{i-1}, x_i x_{i+1}, \dots, x_n)| f_{X_i}(x_i) \dots f_{X_n}(x_n) dx_i \dots dx_n \\
&\leq \int c_i f_{X_i}(x_i) \dots f_{X_n}(x_n) dx_i \dots dx_n = c_i.
\end{aligned}
$$

Logo, o resultado saí pela Desigualdade de Azuma.

■

Perceba que o Teorema **??** é uma generalização da de desigualdade de Hoeffding, porém ao invés de encontrarmos bounds para a calda da distribuição de X (com X limitada), agora a quantia de nosso interesse é uma função da amostra

$$Z = f(X_1, \dots, X_n),$$

limitada em cada coordenada.

Note que para:

- Para Hoeffding, precisamos que a **v.a.** X **seja limitada**;
- Para Azuma, precisamos que a **diferença do Martingal** X_n seja limitada, mas não necessariamente o X_n é limitado;
- Para McDiarmid, precisamos que a **diferença da função das v.a. independentes** seja limitada coordenada a coordenada e não precisamos da limitação das X_n.

1.1.10 Desigualdade de Concentração Gaussiana

Teorema 1.1.25 (DCG - Funções Lipschitz). *Sejam $X_1, \dots, X_n$ v.a.'s i.i.d. com $X_1 \sim N(0,1)$ e $F : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável e Lipschitz com constante L com relação à norma Euclidiana, isto é,*

$$|F(X) - F(Y)| \leq L\|X - Y\|,$$

Então, se $X = (X_1, \dots, X_n)$, existem constantes $c, C > 0$ tal que

$$\mathbb{P}(|F(X) - \mathbb{E}(F(X))| > \varepsilon) \leq Ce^{-ct^2}.$$

Intuição: Temos que

$$F(X) - F(Y) = F'(Y)(X - Y) + O(|X - Y|^2),$$

isto é,

$$F(X) - F(Y) \approx F'(Y)(X - Y).$$

Pela hipótese de F ser L-Lipschitz, é fácil ver que $|F'(Y)| \leq L$. Além disso, como $x \mapsto e^{-tx}$ é convexa, temos que

$$e^{-t\mathbb{E}_Y(Y)} \leq \mathbb{E}_Y(e^{-tY}).$$

Portanto, tomando Y i.i.d. com relação à X e supondo que X é centrada, temos que

$$\begin{aligned} \mathbb{E}_X(e^{tF(X)}) \cdot 1 &= \mathbb{E}_X(e^{tF(X)})e^{-t\mathbb{E}_Y(F(Y))} \\ &\leq \mathbb{E}_X\mathbb{E}_Y(e^{t|F(X)-F(Y)|}) \\ &\leq \mathbb{E}_X\mathbb{E}_Y(e^{tL|X-Y|}) \leq Ce^{ct^2}. \end{aligned}$$

Para formalizarmos a intuição acima, precisamos dar um sentido para

$$F(X) - F(Y) \approx F'(Y)(X - Y).$$

Para isso, vamos usar o seguinte lema:

Lema 1.1.26. *Suponha $F : \mathbb{R}^n \to \mathbb{R}$ diferenciável, então para qualquer função convexa $\varphi : \mathbb{R} \to \mathbb{R}$, temos que*

$$\mathbb{E}\left(\varphi[f(X) - \mathbb{E}(f(X)]\right) \leq \mathbb{E}_X\mathbb{E}_Y\left(\varphi\left(\frac{\pi}{2}\langle F'(X), Y\rangle\right)\right),$$

onde $X, Y \sim N(0, I_n)$ independentes.

Prova do Lema **??**. Considere $Z(\theta) \in \mathbb{R}^n$ com coordenadas

$$Z_i(\theta) = X_i \sin\theta + Y_i \cos\theta,$$

como $Z_i(\theta)$ é uma combinação linear de X_i, Y_i, temos que

$$Z_i(\theta) \sim N(0, \sin^2\theta + \cos^2\theta) = N(0, 1),$$

e é fácil ver que o mesmo vale para $Z_i'(\theta)$, isto é,

$$Z_i'(\theta) \sim N(0, \sin^2\theta + \cos^2\theta) = N(0, 1).$$

Mas então, supondo $f(X)$ centrada e Y i.i.d. com relação à X, temos que

$$\begin{array}{rcl}
\mathbb{E}_X(\varphi(f(X))) & = & \mathbb{E}_X(\varphi(f(X)-\mathbb{E}_Y(f(Y)))) \\
& \underset{\text{Jensen}}{\leq} & \mathbb{E}_X\mathbb{E}_Y(\varphi[f(X)-f(Y)]) \\
& \underset{\text{T.Fund.Calc}}{\leq} & \mathbb{E}_X\mathbb{E}_Y\varphi\left(\int_0^{\pi/2}\frac{dF(Z_\theta)}{d\theta}d\theta\right) \\
& \underset{\text{Normalizando}}{\leq} & \mathbb{E}_X\mathbb{E}_Y\varphi\left(\int_0^{\pi/2}\pi/2\frac{dF(Z_\theta)}{d\theta}\frac{d\theta}{\pi/2}\right) \\
& \underset{\text{Jensen}}{\leq} & \mathbb{E}_X\mathbb{E}_Y\frac{1}{\pi/2}\int_0^{\pi/2}\varphi\left(\pi/2\frac{dF(Z_\theta)}{d\theta}\right)d\theta \\
& = & \mathbb{E}_X\mathbb{E}_Y\frac{1}{\pi/2}\int_0^{\pi/2}\varphi\left(\frac{\pi}{2}\langle F'(Z(\theta)),Z'(\theta)\rangle\right)d\theta \\
& \underset{\tilde{X},\tilde{Y}\sim N(0,I_n)}{=} & \mathbb{E}_X\mathbb{E}_Y\frac{1}{\pi/2}\int_0^{\pi/2}\varphi\left(\frac{\pi}{2}\langle F'(\tilde{X}),\tilde{Y}\rangle\right)d\theta \\
& \leq & \mathbb{E}_X\mathbb{E}_Y\varphi\left(\frac{\pi}{2}\langle F'(\tilde{X}),\tilde{Y}\rangle\right)
\end{array}$$

■

Prova do Teorema **??**. Utilizando o lema anterior, temos que

$$\begin{array}{rcl}
\mathbb{E}_X(e^{tF(X)}) & = & \mathbb{E}_X(e^{tF(X)})e^{-t\mathbb{E}_Y(F(Y))} \\
& \leq & \mathbb{E}_X\mathbb{E}_Y(e^{t(F(X)-F(Y))}) \\
& \leq & \mathbb{E}_X\mathbb{E}_Y\exp(\pi t/2\langle F'(X),Y\rangle) \\
& \leq & \mathbb{E}_X\mathbb{E}_Y\exp\left(\frac{\pi t}{2}\sum_{i=1}^n\frac{dF(X)}{x_i}Y_i\right) \\
& \underset{\text{Independência}}{\leq} & \mathbb{E}_X\mathbb{E}_Y\exp\left(\frac{\pi t}{2}\sum_{i=1}^n\frac{dF(X)}{x_i}Y_i\right) \\
& \underset{\text{Independência}}{\leq} & \mathbb{E}_X\prod_{i=1}^n\mathbb{E}_Y\exp\left(\frac{\pi t}{2}\frac{dF(X)}{x_i}Y_i\right) \\
& \underset{\text{FGdM de Gaussiana}}{\leq} & \mathbb{E}_X\prod_{i=1}^n\exp\left(\frac{\pi^2t^2}{8}\left(\frac{dF(X)}{x_i}\right)^2\right) \\
& \underset{\text{Independência}}{\leq} & \mathbb{E}_X\exp\left(\frac{\pi^2t^2}{8}\left(\sum_{i=1}^n\frac{dF(X)}{x_i}\right)^2\right) \\
& = & \mathbb{E}_X\exp\left(\frac{\pi^2t^2}{8}\|F'(X)\|^2\right) \\
& \leq & \exp\left(\frac{\pi^2t^2}{8}L^2\right).
\end{array}$$

E daí basta minimizar utilizar a técnica de Chernoff. ■

Teorema 1.1.27 (Desigualdade de Concentração Gaussiana)**.** *Seja $A\subset\mathbb{R}^d$ mensurável e $X\sim N(0,1)^d$. Então*

$$\mathbb{P}(X\in A)\mathbb{P}(X\notin A_t)\leq\exp(-ct^2),$$

para todo $t > 0$ e para alguma constante absoluta $c > 0$

Demonstração. É claro que

$$f(x) = \text{dist}(x, A)$$

é 1-Lipschitz, onde $\text{dist}(a, b) = \|a - b\|_2, \ a, b \in \mathbb{R}^d$.

Agora note que, vamos separar nosso problema em dois casos:

1. Suponha que $t/2 > \mathbb{E}(f(X))$. Então,

$$\begin{aligned}\mathbb{P}(X \notin A_t) &= \mathbb{P}(f(X) > t) \\ &= \mathbb{P}(f(X) - \mathbb{E}(X) > t - \mathbb{E}(X)) \\ &= \mathbb{P}(f(X) - \mathbb{E}(X) > t/2) \\ &\leq \mathbb{P}(|f(X) - \mathbb{E}(X)| > t/2) \\ &\underset{??}{\leq} Ce^{-ct^2/4}.\end{aligned}$$

2. Suponha que $t/2 \leq \mathbb{E}(f(X))$. Note que se ω é tal que $f(X(\omega)) = 0$, então

$$0 + \mathbb{E}(f(X)) = -f(X(\omega)) + \mathbb{E}(f(X)) \geq t/2,$$

ou seja

$$\{f(X) = 0\} \subset \{-f(X) + \mathbb{E}f(X) \geq t/2\}.$$

Portanto,

$$\begin{aligned}\mathbb{P}(X \in A) &\leq \mathbb{P}(|f(X) - \mathbb{E}(X)| > t/2) \\ &\underset{??}{\leq} Ce^{-ct^2/4}.\end{aligned}$$

Mas então, se vale o caso 1:

$$\mathbb{P}(X \in A)\mathbb{P}(X \notin A_t) \leq 1 \cdot \mathbb{P}(X \notin A_t) \leq \tilde{C}\exp(-\tilde{c}t^2),$$

e se vale o caso 2:

$$\mathbb{P}(X \in A)\mathbb{P}(X \notin A_t) \leq 1 \cdot \mathbb{P}(X \in A) \leq \tilde{C}\exp(-\tilde{c}t^2)$$

∎

Com um pouco mais de esforço, conseguimos mostrar que, de fato, as constantes encontradas nas Desigualdades de Concentração Gaussiana (**??** e **??**) são absolutas se fixado a constante de Lipschitz L.

1.2 Aplicações

1.2.1 Melhorando Algoritmos Quase Aleatórios

Considere um problema de decisão **P** com resposta binária verdadeira ou falso e suponha que tenhamos acesso à um algoritmo que retorne a resposta certa para **P** com probabilidade $1/2+\delta$, isto é, apenas um pouco melhor que um algoritmo totalmente aleatório.

Uma possível ideia para tentar melhorar a performance do nosso algoritmo seria rodar o mesmo N vezes e então escolher a resposta que saiu com maior frequência como resposta final. Utilizando Hoeffding, somos capazes de provar que de fato tal procedimento funciona com probabilidade alta, se tomarmos

$$n \geq \frac{1}{\delta^2}\ln\left(\frac{1}{\varepsilon}\right).$$

Para provarmos tal resultado, seja X a função que é 1 se escolhemos a resposta errada e 0 caso escolhemos a resposta certa, isto é, X é a função indicadora de escolhermos a resposta errada. Temos que

$$\mathbb{E}(X) = \mathbb{P}(\{\text{escolher resposta errada}\}) = 1/2-\delta,$$

então, se

$$S_n = X_1 + \cdots + X_n,$$

onde X_i são i.i.d. com relação a X, queremos mostrar que para n apropriado

$$\mathbb{P}(S_n/n > 1/2) < \varepsilon,$$

já que a expressão acima representa a frequência com que o nosso algoritmo escolheu a resposta errada em n lançamentos. Por Hoeffding, temos que

$$\begin{aligned}
\mathbb{P}(S_n/n > 1/2) &= \mathbb{P}(S_n/n + \delta > 1/2 + \delta) \\
&= \mathbb{P}(S_n/n - 1/2 + \delta > \delta) \\
&= \mathbb{P}(S_n/n - \mathbb{E}(S_n/n) > \delta) \\
&= \mathbb{P}(S_n - \mathbb{E}(S_n) > n\delta) \leq e^{-2(\delta n)^2/n} \\
&\leq e^{-2\delta^2 n}.
\end{aligned}$$

Mas note que

$$e^{-2\delta^2 n} < \varepsilon \Leftrightarrow n \geq \frac{1}{\delta^2}\ln\left(\frac{1}{\varepsilon}\right).$$

1.2.2 Johnson-Lindenstrauss

1.3 Exercícios Resolvidos

Capítulo 2

Limitando a Variância via Efron-Stein

As principais referências foram:

1. [?]

Considere (X_n) uma sequência de v.a.'s e defina

$$Z = f(X_1, \dots, X_n),$$

para uma função f em n variáveis. Supondo que Z possua variância finita, o nosso objetivo nessa seção é encontrar bounds para

$$\mathbb{V}\text{ar}(Z).$$

A priori, não precisamos que (X_n) sejam i.i.d., e como veremos, a única hipótese essencial será independência.

Note que uma vez que conseguimos um limitante para a variância de Z, conseguimos facilmente encontrar desigualdades de concentração para Z, utilizando por exemplo Chebyshev. Antes de continuar, recomendamos a leitura do Capítulo **??** sobre Martingales para melhor entendimento desta seção.

2.1 Provando a Desigualdade de Efron-Stein

Definição 2.1.1. Seja $X_1, \dots, X_n$ uma sequência de v.a.'s. Então:

1. Definimos $\mathbb{E}_i$, o operador tomar condicional até i, como

   $$\mathbb{E}_i(X) = \mathbb{E}(X|X_1, \dots, X_i),\ i = 1, \dots, n$$

   e $\mathbb{E}_0 X = \mathbb{E}X$.

2. Definimos $\mathbb{E}_{(i)}$, o operador tomar condicional exceto em i, como

   $$\mathbb{E}^{(i)}(X) = \mathbb{E}(X|X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n).$$

3. Definimos Δ_i, o operador tomar a diferença condicional, como

$$\Delta_i(X) = \mathbb{E}_i X - \mathbb{E}_{i-1} X.$$

Teorema 2.1.2 (Desigualdade de Efron-Stein)**.** *Sejam $X_1, \ldots, X_n$ v.a.'s independentes e $Z = f(X_1, \ldots, X_n)$. Então*

1. *Vale a seguinte desigualdade:*

$$\mathbb{V}ar(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathbb{E}^{(i)} Z\right)^2\right] := v.$$

2. *Sejam $X'_1, \ldots, X'_n$ cópias independentes de $X_1, \ldots, X_n$ e*

$$Z'_i = f(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n),$$

então

$$v = \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z'_i)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z'_i)_+^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z'_i)_-^2\right].$$

3. *Temos que*

$$v = \sum_{i=1}^{n} \inf_{Z_i} \mathbb{E}\left[(Z - Z_i)^2\right],$$

onde o ínfimo é escolhido sobre todas as funções Z_i que são $X^{(i)}$-mensuráveis e quadrado integráveis.

Demonstração. Vamos provar os itens separadamente.

1. Note que

$$\begin{aligned}
\mathbb{V}\text{ar}(Z) &= \mathbb{E}\left(\sum_{i=1}^{n} \Delta_i Z\right)^2 \\
&= \mathbb{E}\left(\sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_i Z \Delta_j Z\right) \\
&= \mathbb{E}\left(\sum_{i=1}^{n} (\Delta_i Z)^2\right) + 2\mathbb{E}\left(\sum_{i>j} \Delta_i Z \Delta_j Z\right) \\
&= \sum_{i=1}^{n} \mathbb{E}(\Delta_i Z)^2 + 0,
\end{aligned}$$

já que, para $i > j$, como já vimos que $\Delta_i Z$ possui a propriedade das diferenças de Martingales, então

$$\begin{aligned}
\mathbb{E}\left(\Delta_i Z \Delta_j Z\right) &= \mathbb{E}\mathbb{E}_j\left(\Delta_i Z \Delta_j Z\right) \\
&= \mathbb{E}\Delta_j Z \mathbb{E}_j\left(\Delta_i Z\right) = \mathbb{E}\Delta_j Z \cdot 0,
\end{aligned}$$

Ou seja,

$$\mathbb{V}\mathrm{ar}(Z) = \sum_{i=1}^{n} \mathbb{E}(\Delta_i Z)^2.$$

Note que a relação

$$\mathbb{V}\mathrm{ar}(Z) = \sum_{i=1}^{n} \mathbb{E}(\Delta_i Z)^2$$

vale mesmo sem independência!

Agora, como X_i são independentes, é fácil ver que

$$\mathbb{E}_i(\mathbb{E}^{(i)}Z) = \mathbb{E}_{i-1}Z,$$

e portanto

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(Z) &= \sum_{i=1}^{n} \mathbb{E}(\Delta_i Z)^2 \\
&= \sum_{i=1}^{n} \mathbb{E}(\mathbb{E}_i Z - \mathbb{E}_{i-1}Z)^2 \\
&= \sum_{i=1}^{n} \mathbb{E}(\mathbb{E}_i Z - \mathbb{E}_i(\mathbb{E}^{(i)}Z))^2 \\
&= \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}_i(Z - \mathbb{E}^{(i)}Z)]^2 \\
&\underset{\text{Jensen}}{\leq} \sum_{i=1}^{n} \mathbb{E}\mathbb{E}_i(Z - \mathbb{E}^{(i)}Z)^2 \\
&= \sum_{i=1}^{n} \mathbb{E}(Z - \mathbb{E}^{(i)}Z)^2.
\end{aligned}$$

2. Primeiro, vamos provar o seguinte lema:

Lema 2.1.3. *Suponha que X, Y tenham a mesma distribuição e que sejam independentes condicionadas à $\mathcal{F}$, então*

$$\mathbb{V}ar(X|\mathcal{F}) = \frac{\mathbb{E}((X-Y)^2|\mathcal{F})}{2}$$

Demonstração. Utilizando **??**,

$$\begin{aligned}
\mathbb{E}((X-Y)^2|\mathcal{F}) &= \mathbb{E}(X^2|\mathcal{F}) + \mathbb{E}(Y^2|\mathcal{F}) - 2\mathbb{E}(XY|\mathcal{F}) \\
&= \mathbb{E}(X^2|\mathcal{F}) + \mathbb{E}(Y^2|\mathcal{F}) - 2\mathbb{E}(X|\mathcal{F})\mathbb{E}(Y|\mathcal{F}) \\
&= \mathbb{E}(X^2|\mathcal{F}) + \mathbb{E}(X^2|\mathcal{F}) - 2\mathbb{E}(X|\mathcal{F})^2 \\
&= 2\mathbb{V}\mathrm{ar}(X|\mathcal{F}).
\end{aligned}$$

∎

Mas então, temos que

$$\begin{aligned}
v &= \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathbb{E}^{(i)}Z\right)^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}^{(i)}\left(\left(Z - \mathbb{E}^{(i)}Z\right)^2\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left(\mathbb{V}\mathrm{ar}^{(i)}(Z)\right) \\
&= \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left(\mathbb{E}^{(i)}\left[(Z - Z_i')^2\right]\right) \\
&= \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')^2\right] \\
&= \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')_+^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')_-^2\right],
\end{aligned}$$

onde a última igualdade é verdade já que $Z - Z_i'$ é simétrica. Além disso, é fácil ver que Z e Z_i' são independentes com respeito à $X^{(i)}$.

3. Como $\mathbb{E}^{(i)}(Z)$ é a projeção ortogonal de Z em $Z^{(i)}$ com o produto interno de L^2, temos que

$$\mathbb{E}((Z - \mathbb{E}^{(i)}Z)^2) = \inf_{Z_i} \mathbb{E}\left[(Z - Z_i)^2\right],$$

mas então, pelo item 1

$$\begin{aligned}
v &= \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathbb{E}^{(i)}Z\right)^2\right] \\
&= \sum_{i=1}^{n} \inf_{Z_i} \mathbb{E}\left[(Z - Z_i)^2\right]
\end{aligned}$$

■

Sejam $X_1, \ldots, X_n$ independentes e considere

$$Z = f(X_1, \ldots, X_n) = X_1 + \cdots + X_n.$$

Então, por Efron-Stein **??**:

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(X_i) = \mathbb{V}\mathrm{ar}(Z) \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathbb{E}^{(i)} Z\right)^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[(X_i - \mathbb{E}(X_i))^2\right] \\
&= \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(X_i).
\end{aligned}$$

Isso nos diz, que de certa forma, o bound encontrado via Efron-Stein não pode ser melhorado.

2.2 Aplicações

2.2.1 Desigualdade de Poincaré: Caso Convexo

Definição 2.2.1 (Função Separadamente Convexa)**.** Uma função f em n variáveis é separadamente convexa se fixados $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ então f é convexa na variável i, isso para $i = 1, \dots, n$.

Teorema 2.2.2 (Desigualdade de Poincaré: caso Convexo)**.** *Sejam $X_1, \dots, X_n$ v.a.'s independentes com valores em $[0,1]$ e considere $f : [0,1]^n \to \mathbb{R}$ separadamente convexa e cujas derivadas parciais existem. Então se $f(X) = (X_1, \dots, X_n)$*

$$\mathbb{V}ar(f(X)) \leq \mathbb{E}\left[\|\nabla f(X)\|^2\right].$$

Demonstração. Pelo item 3 de Efron-Stein **??**, temos que para Z_i quadrado-integrável e $X^{(i)}$-mensurável,

$$\mathbb{V}\mathrm{ar}(f(X)) \leq \sum_{i=1}^{n} \inf_{Z_i} \mathbb{E}\left[(f(X) - Z_i)^2\right],$$

então, basta estudarmos o comportamento de

$$(f(X) - Z_i)^2$$

para alguma Z_i apropriada. Considere então

$$Z_i = \min_{x_i} f(X_1, \dots, x_i, \dots, X_n),$$

e a_i o valor para qual o mínimo acima é atingido [1]. Se $X_i' = (X_1, \dots, a_i, \dots, X_n)$, então

$$\begin{aligned}
\mathbb{E}\left[(Z - Z_i)^2\right] &= \mathbb{E}\left[(f(X) - f(X_i'))^2\right] \\
&\underset{\text{Sep.Convx.}}{\leq} \mathbb{E}\left[\left(\frac{df}{dx_i}(X)\right)^2 (X_i - a_i)^2\right] \\
&\leq \mathbb{E}\left[\left(\frac{df}{dx_i}(X)\right)^2\right].
\end{aligned}$$

E portanto,

$$\mathbb{V}\mathrm{ar}(f(X)) \leq \sum_{i=1}^{n} \inf_{Z_i} \mathbb{E}\left[(Z - Z_i)^2\right] \leq \mathbb{E}\left[\|\nabla f(X)\|^2\right].$$

■

[1] O mínimo existe já que f é contínua e defina num compacto.

2.2.2 Desigualdade de Poincaré: Caso Gaussiano

2.2.3 Diferenças Limitadas

Lema 2.2.3. *Suponha que f satisfaça a propriedade das diferenças limitadas* **??**, *então se $Z = f(X_1, \ldots, X_n)$*

$$\mathbb{V}ar(Z) \leq \frac{1}{4} \sum_{i=1}^{n} c_i^2.$$

Demonstração. Já sabemos que

$$\mathbb{V}\text{ar}(Z) \leq \inf_{Z_i} \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i)^2\right],$$

onde Z_i são quadrado-integráveis em $X^{(i)}$. Defina então

$$Z_i = \frac{1}{2}\left(\sup_{a_i} f(X_1, \ldots, a_i, \ldots, X_n) + \inf_{a_i} f(X_1, \ldots, a_i, \ldots, X_n)\right),$$

isto é, Z_i é a média ponderada do sup e inf de Z na coordenada i. Então é claro que

$$(Z - Z_i)^2 \leq \frac{c_i^2}{4}.$$

■

Note que, utilizando Chebyshev, temos que

$$\mathbb{P}(|Z - \mathbb{E}(Z)| > \varepsilon) \leq \frac{1}{\varepsilon^2} \sum_{i=1}^{n} \frac{c_i^2}{4}.$$

Porém, já vimos utilizando a Desigualdade das Diferenças Limitadas **??**, que

$$\mathbb{P}(|Z - \mathbb{E}Z| > \varepsilon) \leq \exp\left(\frac{-2\varepsilon^2}{\sum c_i^2}\right).$$

Ou seja, nesse caso, não ganhamos em nada utilizando a estimativa de Efron-Stein com Chebyshev.

2.2.4 Maior Subsequência Comum

Considere duas sequências $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$ binárias. Defina Z como o tamanho da maior subsequência com respeito às duas sequências, isto é

$$Z = \max\{k : X_{i_1} = Y_{j_1}, \ldots, X_{i_k} = X_{j_k}.$$

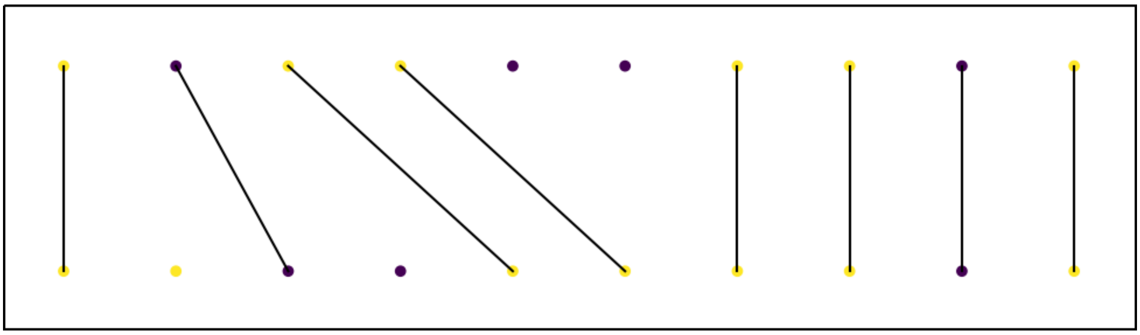

Figura 2.1: Cada linha horizontal representa uma sequência binária. Dois pontos conectador por uma semi-reta fazem parte de uma maior subsequência.

Note que mudar um elemento de uma das duas subsequências, altera em no máximo ± 1 o valor Z, logo, por Efron-Stein **??**:

$$\mathbb{V}\text{ar}(Z) \leq \frac{n}{2}.$$

Logo, por Chebyshev, temos que

$$\mathbb{P}(|Z - \mathbb{E}(Z)| > \varepsilon) \leq \frac{n}{2\varepsilon^2},$$

ou seja, a menos de uma constante, Z está a uma taxa de $\sqrt{n}$ do seu valor esperado.

2.2.5 Autovalores de Matrizes Aleatórias Simétricas

Seja A uma matriz simétrica, que tem como entradas v.a.'s $a_{i,j}$, com $1 \leq i \leq j \leq n$ e com valor absoluto limitado por 1.

O nosso objetivo é estudar $Z(A) = \lambda_1$, onde λ_1 representa o maior autovalor da matriz aleatória A.

Lema 2.2.4. *Seja A uma matriz simétrica e λ_1 seu maior autovalor, então*

$$\lambda_1 = v^t A v = \sup_{\|x\|=1} x^t A x,$$

onde v é um autovetor associada á λ, com $\|v\| = 1$.

Demonstração. Como A é simétrica, vamos supor que A é diagonalizável. É claro que se v é um autovalor unitário para λ_1, então

$$v^t A v = \lambda_1 \|v\|^2 = \lambda_1,$$

e portanto

$$\sup_{\|x\|=1} x^t A x \geq \lambda_1.$$

Como $f(x) = x^t A x$ é contínua, então restrita à bola unitária, ela deve atingir máximo. Suponha que f atinga o máximo em y, com $\|y\| = 1$.

Então, se $\lambda_1 \geq \ldots, \lambda_n$, temos que

$$\begin{aligned} y^t A y &= \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2 \\ &\leq \lambda_1 y_1^2 + \cdots + \lambda_1 y_n^2 \\ &\leq \lambda_1 (y_1^2 + \cdots + y_n^2) = \lambda_1. \end{aligned}$$

■

Agora, vamos utilizar o Teorema **??** da seguinte maneira. Considere $A'_{i,j}$ uma matriz idêntica à A, porém substituímos a coordenada $a_{i,j}$ por uma cópia $a'_{i,j}$ independente de $a_{i,j}$. Então, por Efron-Stein, se $v_{i,j}$ é um autovalor unitário para o maior autovalor de $A'_{i,j}$, então

$$\begin{aligned} (Z - Z'_{i,j})_+ &= (v^t A v - v_{i,j}^t A'_{i,j} v_{i,j})_+ \\ &\leq (v^t A v - v^t A'_{i,j} v)_+ \\ &\leq (v^t (A - A'_{i,j}) v)_+ \\ &= (2 v_i v_j (a_{i,j} - a'_{i,j}))_+ \\ &\leq 4 |v_i v_j|. \end{aligned}$$

Portanto,

$$\begin{aligned} \sum_{1 \leq 1 \leq j \leq n} (Z - Z'_{i,j})_+^2 &\leq \sum_{1 \leq 1 \leq j \leq n} 16 |v_i v_j|^2 \\ &\leq 16 \sum_{1 \leq 1 \leq j \leq n} |v_i v_j|^2 \\ &\leq 16 \sum_{i=1}^n \sum_{i=1}^n |v_i v_j|^2 \\ &\leq 16 \left(\sum_{i=1}^n |v_i|^2 \right)^2 = 16. \end{aligned}$$

Ou seja, mostramos que

$$\mathbb{V}\mathrm{ar}(Z) \leq 16,$$

independente do tamanho da matriz e da distribuição das entradas!

2.2.6 Valores Singulares de Matrizes Aleatórias

2.3 Exercícios Resolvidos

Capítulo 3

Teoria da Informação

As Principais referências usadas foram:

1. [?].

3.1 Entropia de Shannon e Entropia Relativa

Definição 3.1.1 (Entropia de Shannon)**.** Seja X uma v.a. tomando valores num conjunto enumerável com distribuição p_X E entropia de Shannon de X é definida como

$$H(X) = \mathbb{E}(-\ln p_X(X)) = -\sum_x p_X(x) \ln p_X(x),$$

com a convenção de que $0 \ln 0 = 0$.

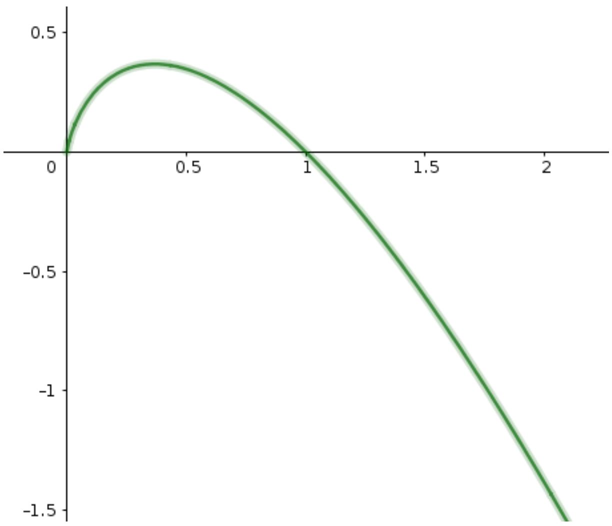

Figura 3.1: Gráfico de $x \mapsto -x \ln x$.

Definição 3.1.2 (Divergência de Kullback-Leibler/ Entropia Relativa)**.** Sejam P, Q duas probabilidades num conjunto enumerável, com distribuições p, q, respectivamente. Então a divergência de Kullback-Leibler ou Entropia Relativa de P e Q é definida como

$$D(P\|Q) = \sum_x p(x) \ln \frac{p(x)}{q(x)},$$

se P é absolutamente contínua com respeito à Q e infinito caso contrário.

Lema 3.1.3. *Temos que $D(P\|Q) \geq 0$, valendo a igualdade se, e só se, $P = Q$.*

Demonstração. Lembrando que $\ln x \leq x - 1,\ x > 0$, temos que

$$\begin{aligned} D(P\|Q) &= -\sum_x p(x) \ln \frac{q(x)}{p(x)} \\ &\geq \sum_x p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \\ &= \sum_x (q(x) - p(x)) \\ &= \sum_x q(x) - \sum_x p(x) = 0. \end{aligned}$$

E é claro que a igualdade vale se, e só se, $P = Q$. ■

Vamos analisar o caso específico quando Q é a probabilidade uniforme no espaço base Ω. Temos então que

$$\begin{aligned} D(P\|Q) &= \sum_x p(x) \ln \frac{p(x)}{q(x)} \\ &= \sum_x p(x) \ln (|\Omega| p(x)) \\ &= \ln |\Omega| - H(X), \end{aligned}$$

onde X é uma v.a. com distribuição p. Do cálculo acima, podemos concluir duas coisas, a primeira é que encontramos uma fórmula explícita para o caso em que Q é uniforme. A segunda é que, como $D(P\|Q) \geq 0$, temos que

$$H(X) \leq \ln |\Omega|,$$

valendo a igualdade se, e só se, X também tem distribuição uniforme sobre Ω.
Isto é, estamos dizendo que a entropia de Shannon é maximizada quando X é uniforme! De certa forma, isso significa que dentre todas as distribuições possíveis, a uniforme é a que menos nos dá 'informação'.

Exemplo 3.1.4. ** fazer ex 4.1 e 4.2

3.2 Entropia em Produtos e Regra da Cadeia

Proposição 3.2.1. *Sejam X, Y v.a.'s tomando valores num conjunto enumerável. Se $p(x, y)$ é a probabilidade conjunta de X, Y e p_X, p_Y são as probabilidades marginais de X, Y, respectivamente, então*

$$H(X) + H(Y) - H((X, Y)) = \sum_{x,y} p(x, y) \ln \frac{p(x, y)}{p_X(x) p_Y(y)}.$$

Demonstração. Basta lembrar que, por exemplo, para p_X, vale

$$p_X(x) = \sum_y p(x, y).$$

■

Note que

$$\sum_{x,y} p(x, y) \ln \frac{p(x, y)}{p_X(x) p_Y(y)}$$

nada mais é que a entropia relativa da probabilidade conjunta P de (X, Y) e da medida produto $P_X \otimes P_y$, logo

$$D(P|P_X \otimes P_y) = H(X) + H(Y) - H((X, Y)) \geq 0,$$

valendo a igualdade se, e só se, X, Y são independentes e portanto $P = P_X \otimes P_y$.

Note que a entropia relativa, ou divergência de Kullback-Leible, funciona como uma distância entre probabilidades. Por exemplo, podemos dizer que

$$H(X) + H(Y) - H((X, Y)) \geq 0$$

mede o quão 'independentes' são X, Y.
O valor

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

é usualmente conhecido como *informação mutua entre X e Y*. Note que a entropia de Shannon nada mais é que a informação mutua entre X e X.

Definição 3.2.2 (Entropia Condicional)**.** Sejam X, Y duas v.a.'s discretas. Então, a entropia condicional de $H(X|Y)$ é definida como

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

Lembrando que $p(x, y) = p(x|y)p_y(y)$, é fácil ver que a definição acima satisfaz

$$H(X|Y) = -\mathbb{E}(\ln p(X|Y)),$$

e então, é claro que $H(X|Y) \geq 0$.

Além disso, como

$$D(P|P_X \otimes P_y) = H(X) + H(Y) - H(X,Y),$$

temos que

$$D(P|P_X \otimes P_y) = H(X) - H(X|Y) \geq 0,$$

ou seja,

$$H(X) \geq H(X|Y),$$

ou seja, a operação de tomar condicional reduz a entropia, o que faz sentido com a nossa intuição de que condicionar uma v.a. X com respeito à uma σ-álgebra $\mathcal{F}$, nos dá mais informação sobre X.

Proposição 3.2.3 (Regra da Cadeia)**.** *Sejam* $X_1, \ldots, X_n$ *v.a.'s, então*

$$\begin{aligned} H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots \\ + H(X_n|X_1, \ldots, X_{n-1}). \end{aligned}$$

Demonstração. Façamos o caso em que temos três v.a.'s, o caso geral sai por indução. Através de alguns cálculos simples, conseguimos mostrar que a definição de entropia condicional continua sendo verdade condicionalmente. Temos então que que

$$H(X_3, X_2|X_1) = H(X_2|X_1) + H(X_3|X_1, X_2),$$

somando $H(X_1)$ dos dois lados, temos que

$$H(X_1) + H(X_3, X_2|X_1) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2),$$

mas, utilizando condicionalmente a definição de entropia condicional,

$$H(X_1) + H(X_3, X_2|X_1) = H(X_1, X_2, X_3).$$

■

Para lembrar da relação

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

basta lembrar que

$$P(A,B) = P(A|B)P(B)$$

e que na definição de H há um ln, ou seja, transformamos os produtos em somas.

3.3 Desigualdade de Han

Teorema 3.3.1 (Desigualdade de Han)**.** *Seja* $X_1, \ldots, X_n$ *uma sequência de v.a.'s discretas. Então*

$$H(X_1, \ldots, X_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Demonstração. Note que, dado $0 \leq i \leq n$, usando o fato de que $H(X,Y) = H(X|Y) + H(Y)$, temos que

$$\begin{aligned} H(X_1, \ldots, X_n) &= H(X_1, \ldots, X_i, \ldots, X_n) \\ &= H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n). \end{aligned}$$

Logo, somando em i, temos que

$$nH(X_1, \ldots, X_n) = \sum_{i=1}^{n} \left(H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \right).$$

Utilizando o fato de que condicionar reduz a entropia, temos que

$$H(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \leq H(X_i|X_1, \ldots, X_{i-1}),$$

mas note que, pela regra da cadeia

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}),$$

e assim concluímos o teorema. ■

3.4 Exercícios Resolvidos