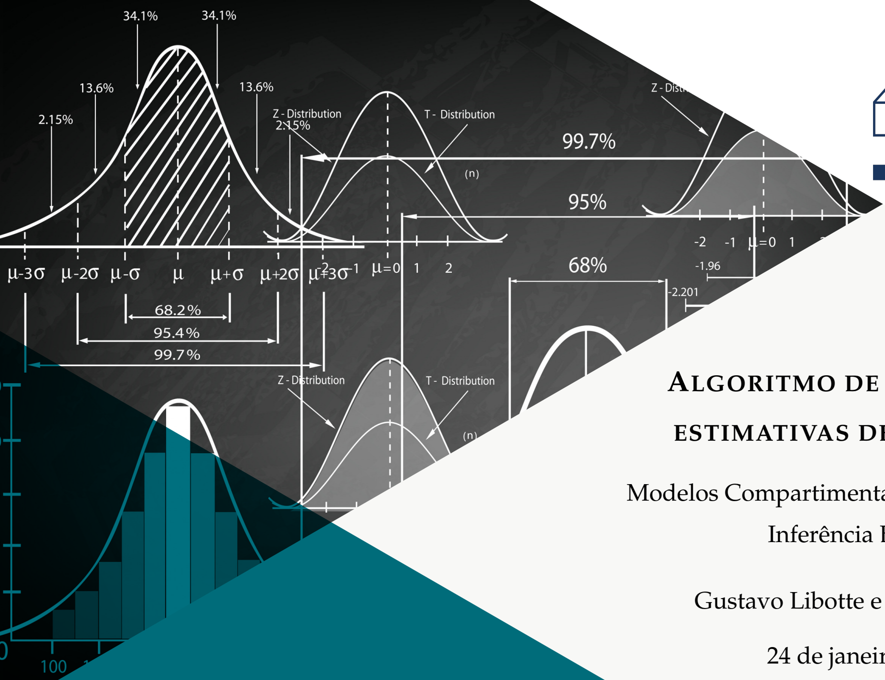



Laboratório Nacional de Computação Científica
UNIDADE DE PESQUISA DO MCTI

# Algoritmo de Metropolis e estimativas de parâmetros

Modelos Compartimentais em Epidemiologia e Inferência Bayesiana

Gustavo Libotte e Regina Almeida

24 de janeiro de 2022

# O Algoritmo de Metropolis

- Vamos considerar uma situação genérica onde temos um modelo de amostragem $Y \sim p(y \mid \theta)$ e uma distribuição *a priori* $p(\theta)$.
- Embora na maioria dos problemas $p(y \mid \theta)$ e $p(\theta)$ possam ser calculados para quaisquer valores de $y$ e $\theta$,
$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{\int p(\theta)\, p(y \mid \theta)\, \mathrm{d}\theta}$$
geralmente é difícil de calcular **devido à integral no denominador**.
- Se pudéssemos obter uma amostra de $p(\theta \mid y)$, poderíamos gerar $\theta^{(1)}, \ldots, \theta^{(S)} \overset{i.i.d.}{\sim} p(\theta \mid y)$ e obter aproximações de Monte Carlo para quantidades *a posteriori*, com
$$\mathrm{E}[g(\theta) \mid y] \approx \frac{1}{S} \sum_{s=1}^{S} g\left(\theta^{(s)}\right)$$
- **Mas e se não pudermos amostrar diretamente de $p(\theta \mid y)$?**

# O Algoritmo de Metropolis

- Em termos da aproximação da distribuição *a posteriori*, o crítico não é que tenhamos amostras i.i.d. de $p(\theta \mid y)$.
- Mas ao invés disso, somos capazes de construir uma grande coleção de $\theta$-valores, $\left\{\theta^{(1)}, \ldots, \theta^{(S)}\right\}$, cuja distribuição empírica se aproxima de $p(\theta \mid y)$.
- A grosso modo, para quaisquer dois valores diferentes $\theta_a$ e $\theta_b$, precisamos

$$\frac{\#\left\{\theta^{(s)}\text{ 's na coleção} = \theta_a\right\}}{\#\left\{\theta^{(s)}\text{ 's na coleção} = \theta_b\right\}} \approx \frac{p\left(\theta_a \mid y\right)}{p\left(\theta_b \mid y\right)}$$

- Vamos pensar **intuitivamente** sobre como podemos construir essa coleção.
- Suponha que temos uma coleção $\left\{\theta^{(1)}, \ldots, \theta^{(s)}\right\}$ à qual gostaríamos de adicionar um novo valor $\theta^{(s+1)}$.
- Vamos considerar a adição de um valor $\theta^*$ que está próximo a $\theta^{(s)}$.

# O Algoritmo de Metropolis

- Devemos incluir $\theta^*$ no conjunto ou não?
- Se $p\left(\theta^* \mid y\right) > p\left(\theta^{(s)} \mid y\right)$ então queremos mais que $\theta^*$ esteja no conjunto do que $\theta^{(s)}$. $\theta^*$ é **mais verossímil** do que $\theta^{(s)}$.
- Visto que $\theta^{(s)}$ já está no conjunto, então parece que **devemos** incluir $\theta^*$ também.
- Por outro lado, se $p\left(\theta^* \mid y\right) < p\left(\theta^{(s)} \mid y\right)$ então parece que **não devemos necessariamente** inclui $\theta^*$.
- Portanto, talvez nossa decisão de incluir $\theta^*$ ou não deva ser baseada em uma comparação de $p\left(\theta^* \mid y\right)$ com $p\left(\theta^{(s)} \mid y\right)$.
- Felizmente, essa comparação **pode ser feita** mesmo que não possamos calcular $p(\theta \mid y)$:

$$r = \frac{p\left(\theta^* \mid y\right)}{p\left(\theta^{(s)} \mid y\right)} = \frac{p\left(y \mid \theta^*\right) p\left(\theta^*\right)}{p(y)} \frac{p(y)}{p\left(y \mid \theta^{(s)}\right) p\left(\theta^{(s)}\right)} = \frac{p\left(y \mid \theta^*\right) p\left(\theta^*\right)}{p\left(y \mid \theta^{(s)}\right) p\left(\theta^{(s)}\right)}$$

# O Algoritmo de Metropolis

- Tendo calculado $r$, como devemos **proceder**?
- Se $r > 1$:
  - **Intuição**: Como $\theta^{(s)}$ já está em nosso conjunto, devemos incluir $\theta^*$, pois tem uma probabilidade maior do que $\theta^{(s)}$.
  - **Procedimento**: Aceite $\theta^*$ em nosso conjunto, ou seja, defina $\theta^{(s+1)} = \theta^*$.
- Se $r < 1$:
  - **Intuição**: A frequência relativa de $\theta$-valores em nosso conjunto igual a $\theta^*$ em comparação com aqueles iguais a $\theta^{(s)}$ deve ser $p\left(\theta^* \mid y\right) / p\left(\theta^{(s)} \mid y\right) = r$. Isso significa que para cada instância de $\theta^{(s)}$, devemos ter apenas uma “fração” de uma instância de um valor $\theta^*$.
  - **Procedimento**: Defina $\theta^{(s+1)}$ igual a $\theta^*$ ou $\theta^{(s)}$, com probabilidade $r$ e $1 - r$ respectivamente.
- Esta é a intuição básica por trás do famoso **algoritmo de Metropolis**.

# O Algoritmo de Metropolis

- O algoritmo de Metropolis procede amostrando um valor de proposta $\theta^*$ próximo ao valor atual $\theta^{(s)}$ usando uma **distribuição simétrica de proposta** $J\left(\theta^* \mid \theta^{(s)}\right)$.
- **Simétrico** aqui significa que $J(\theta_b \mid \theta_a) = J(\theta_a \mid \theta_b)$, ou seja, a probabilidade de propondo $\theta^* = \theta_b$ dado que $\theta^{(s)} = \theta_a$ é igual à probabilidade de propor $\theta^* = \theta_a$ dado que $\theta^{(s)} = \theta_b$.
- Normalmente $J\left(\theta^* \mid \theta^{(s)}\right)$ é muito simples, com amostras de $J\left(\theta^* \mid \theta^{(s)}\right)$ estando perto de $\theta^{(s)}$ com alta probabilidade.
- **Exemplos**:
  - $J\left(\theta^* \mid \theta^{(s)}\right) = \text{Uniforme}\left(\theta^{(s)} - \delta, \theta^{(s)} + \delta\right)$
  - $J\left(\theta^* \mid \theta^{(s)}\right) = \text{Normal}\left(\theta^{(s)}, \delta^2\right)$
- O valor do parâmetro $\delta$ é geralmente escolhido para fazer o algoritmo de aproximação funcionar de forma **eficiente**.
- **Como $\delta$ afeta a eficiência?**

# O Algoritmo de Metropolis
## Resumo

- Tendo obtido um valor de proposta $\theta^*$, nós o adicionamos ou uma cópia de $\theta^{(s)}$ ao nosso conjunto, dependendo da proporção $r = p(\theta^* \mid y) / p\left(\theta^{(s)} \mid y\right)$.
- Especificamente, dado $\theta^{(s)}$, o algoritmo de Metropolis gera um valor $\theta^{(s+1)}$ como segue:
  1. Gere uma amostra $\theta^* \sim J\left(\theta \mid \theta^{(s)}\right)$.
  2. Calcule a taxa de aceitação:
  $$r = \frac{p(\theta^* \mid y)}{p\left(\theta^{(s)} \mid y\right)} = \frac{p(y \mid \theta^*)\, p(\theta^*)}{p\left(y \mid \theta^{(s)}\right) p\left(\theta^{(s)}\right)}$$
  3. Tome
  $$\theta^{(s+1)} = \begin{cases} \theta^*, \text{ com probabilidade } \min(r, 1) \\ \theta^{(s)}, \text{ com probabilidade } 1 - \min(r, 1) \end{cases}$$
- A etapa 3 pode ser realizada amostrando $u \sim \text{Uniforme}(0, 1)$ e definindo $\theta^{(s+1)} = \theta^*$ se $u < r$, ou definindo $\theta^{(s+1)} = \theta^{(s)}$ caso contrário.

# O Algoritmo de Metropolis
## Exemplo prático

- A NFL tem 32 times e cada time joga 16 jogos da temporada regular por ano, para um total de $N = 256$ jogos.
- De acordo com o site Frontline, houve $Y_1 = 171$ concussões em 2012, $Y_2 = 152$ concussões em 2013, $Y_3 = 123$ concussões em 2014 e $Y_4 = 199$ concussões em 2015.
- O número de concussões é modelado por $Y_i \sim \text{Poisson}\left(N\lambda_i\right)$, onde $\lambda_i = \exp\left(\beta_1 + i\beta_2\right)$ é a taxa no ano $i$ .
- Para completar o modelo Bayesiano, $\beta_1, \beta_2 \sim \text{Normal}\left(0, \tau^2\right)$.
- O logaritmo da taxa média de concussão é linear no tempo, com $\beta_2$ determinando a inclinação. O **objetivo** é determinar se a taxa de concussão está aumentando, ou seja, $\beta_2 > 0$.

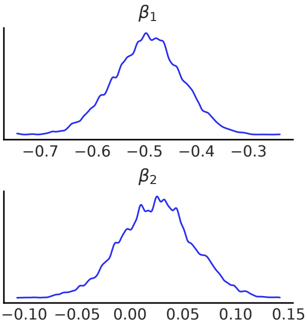
**Atividade prática**

Reproduza os resultados do próximo *slide*, usando as informações apresentadas aqui. Este problema é apresentado em [1], mas **você deve tentar implementar primeiro!**

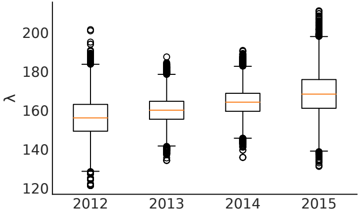
[1] S. K. Ghosh and B. J. Reich.
*Bayesian statistical methods*.
Chapman & Hall/CRC, Boca Raton, 1 edition, 2019.

# O Algoritmo de Metropolis
## Exemplo prático

Distribuição *a posteriori* e traço de $\beta_1, \beta_2$.




Os *boxplots* são a distribuição *a posteriori* de $N\lambda_i = N\exp(\beta_1 + i\beta_2)$, e os pontos são as amostras observadas.



# Quantis de uma variável aleatória
## Definição

**Quantis**

Para qualquer $p$ com $0 < p < 1$, o $p$-ésimo quantil da distribuição de uma variável aleatória $X$, denotado por $x_p$, é definido da seguinte forma:

- Se $X$ for contínua, então o $x_p$ (essencialmente) único é definido por:

$$P\left(X \leq x_p\right) = p \quad \text{e} \quad P\left(X \geq x_p\right) = 1 - p$$

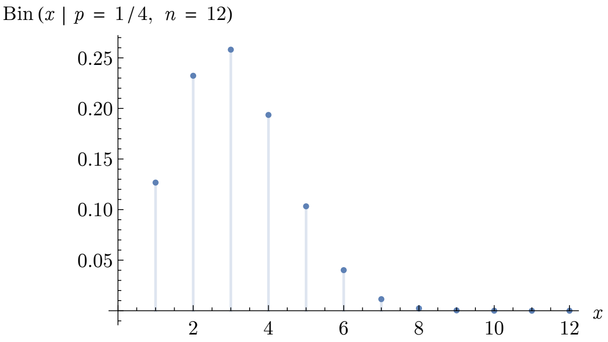
- Para o caso discreto, considere dois casos:
  - Seja $x_k$ o valor para o qual $P\left(X \leq x_k\right) = p$, se tal valor existir. Então o único quantil $p$ é definido como o ponto médio entre $x_k$ e $x_{k+1}$, ou seja, $x_p = \left(x_k + x_{k+1}\right)/2$.
  - Se não houver tal valor, o único $p$-ésimo quantil é definido pela relação $P\left(X < x_p\right) < p$ e $P\left(X \leq x_p\right) > p$ (ou $P\left(X \leq x_p\right) > p$ e $P\left(X \geq x_p\right) > 1 - p$).

# Quantis de uma variável aleatória
## Definição

- Assim, o $p$-ésimo quantil é um ponto $x_p$ que divide a distribuição de $X$ em duas partes, onde $(-\infty, x_p]$ contém exatamente $100p\%$ (ou pelo menos $100p\%$) da distribuição, e $[x_p, \infty)$ contém exatamente $100(1-p)\%$ (ou pelo menos $100(1-p)\%$) da distribuição de $X$.
- Para $p = 0,50$, obtemos a **mediana**.
- Podemo ver os quantis como qualquer separatriz que divide o intervalo de frequência de uma população, ou de uma amostra, **em partes iguais**:
  - Tercil: cada parte tem 33,3% dos dados;
  - Quartil: cada parte tem 25% dos dados;
  - Quintil: cada parte tem 20% dos dados;
  - Decil: cada parte tem 10% dos dados;
  - Duodecil: cada parte tem 8,33% dos dados;
  - Percentil: cada parte tem 1% dos dados;
- A **distância interquartil** é a diferença entre o primeiro e terceiro quartis.

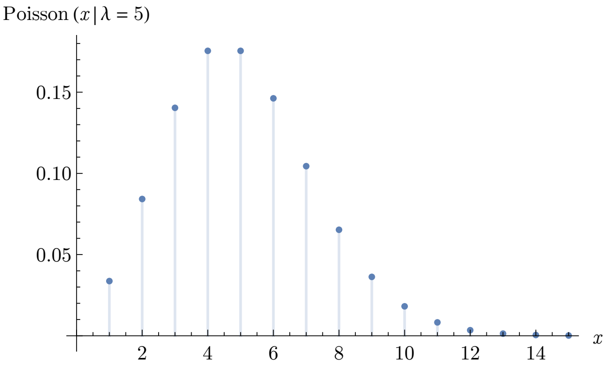
# Quantis de uma variável aleatória
## Alguns exemplos

- Considere que $X \sim \text{Bin}\,(x \mid p = 1/4, n = 12)$ e determine $x_{0,25}, x_{0,50}$, e $x_{0,75}$.




- $x_{0,25} = 2$ uma vez que $P(X < 2) = P(X = 0) + P(X = 1) = 0,1584 \leq 0,25$ e $P(X \leq 2) = 0,1584 + P(X = 2) = 0,3907 \geq 0,25$.
- Da mesma forma, $x_{0,50} = 3$ pois $P(X < 3) = 0,3907 \leq 0,50$ e $P(X \leq 3) = 0,6488 \geq 0,50$.
- Além disso, $x_{0,75} = 4$, já que $P(X < 4) = 0,6488 \leq 0,75$ e $P(X \leq 4) = 0,8424 > 0,75$.

# Quantis de uma variável aleatória
## Alguns exemplos




- Considere que $X \sim \text{Poisson}\,(x \mid \lambda = 5)$ e determine $x_{0,25}, x_{0,50}$, e $x_{0,75}$.
- Para este exemplo, $x_{0,25} = 2$, $x_{0,50} = 4$ e $x_{0,75} = 6$.
- **Verifique estes resultados!**

# Quantis de uma variável aleatória
## Alguns exemplos

1. Seja $X \sim U(0,1)$, tome $p \in \{0{,}10,\ 0{,}20,\ 0{,}30,\ 0{,}40,\ 0{,}50,\ 0{,}60,\ 0{,}70,\ 0{,}80,\ 0{,}90\}$ e determine os valores de $x_p$ correspondentes.
   - Aqui $F(x) = \int_0^x \mathrm{d}t = x, 0 \leq x \leq 1$. Portanto $F\left(x_p\right) = p$ resulta em $x_p = p$.
2. O tempo de vida útil (em anos) de um equipamento eletrônico de determinado tipo pode ser expresso por uma variável aleatória contínua $X$, cuja função de densidade é

$$f(x) = \begin{cases} \frac{1}{2}\exp(-x/2), & \text{para } x \geq 0 \\ 0, & \text{para } x < 0 \end{cases}$$

   - A função de distribuição acumulada, $F(x) = 1 - \exp(-x/2)$, para $x \geq 0$, é estritamente crescente.
   - Para obtermos o valor do segundo **quartil**, fazemos:

$$F\left(x_{0,5}\right) = 1 - \exp\left(-\frac{x_{0,5}}{2}\right) = 0,5 \Rightarrow -\frac{x_{0,5}}{2} = \ln(0,5) = -0,693 \Rightarrow x_{0,5} = 1,39$$

   - Analogamente encontramos: $x_{0,25} = 0,58, x_{0,75} = 2,77$.
   - Isso quer dizer que metade dos equipamentos desse tipo duram no máximo 1,39 anos (ou seja, aproximadamente um ano e cinco meses). Além disso, verifica-se também que 50% desses equipamentos têm seu tempo de vida entre 0,58 anos e 2,77 anos (ou seja, entre sete meses e dois anos e nove meses aproximadamente).

# Intervalos de credibilidade

- Dado $x$ e uma vez determinada uma distribuição *a posteriori*, um **intervalo de credibilidade** para um parâmetro $\theta$ (suponha, por enquanto, um escalar) é formado por dois valores em $\theta$, digamos $[\underline{\theta}(x), \bar{\theta}(x)]$, ou mais simples, $(\underline{\theta}, \bar{\theta})$, tal que

$$P(\underline{\theta} < \theta < \bar{\theta} \mid x) = \int_{\underline{\theta}}^{\bar{\theta}} h(\theta \mid x) d\theta = 1 - \alpha,$$

  onde $1 - \alpha$ (geralmente 0,90, 0,95 ou 0,99) é o nível de credibilidade desejado.

- Se $\Theta = (-\infty, +\infty)$, então uma maneira direta de construir um intervalo de credibilidade (neste caso, central) é baseado nas caudas da distribuição *a posteriori* tal que

$$\int_{-\infty}^{\underline{\theta}} h(\theta \mid x) d\theta = \int_{\bar{\theta}}^{+\infty} h(\theta \mid x) d\theta = \frac{\alpha}{2} .$$

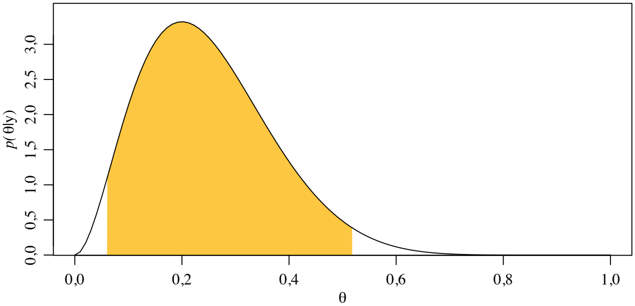
# Intervalos de credibilidade
## Intervalo baseado em quantil

- Talvez a maneira mais fácil de se obter um intervalo de credibilidade seja usar quantis da distribuição *a posteriori*.
- Para fazer um intervalo de credibilidade de $100 \times (1-\alpha)\%$ baseado em quantil, encontre os números $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ tais que
  - $\mathrm{P}\left(\theta < \theta_{\alpha/2} \mid Y = y\right) = \alpha/2$;
  - $\mathrm{P}\left(\theta > \theta_{1-\alpha/2} \mid Y = y\right) = \alpha/2$.
- Os números $\theta_{\alpha/2}, \theta_{1-\alpha/2}$ são os quantis $\alpha/2$ e $1-\alpha/2$ da distribuição *a posteriori* de $\theta$, e assim

$$
\begin{aligned}
\mathrm{P}\left(\theta \in [\theta_{\alpha/2}, \theta_{1-\alpha/2}] \mid Y = y\right) &= 1 - \mathrm{P}\left(\theta \notin [\theta_{\alpha/2}, \theta_{1-\alpha/2}] \mid Y = y\right) \\
&= 1 - \left[\mathrm{P}\left(\theta < \theta_{\alpha/2} \mid Y = y\right) + \mathrm{P}\left(\theta > \theta_{1-\alpha/2} \mid Y = y\right)\right] \\
&= 1 - \alpha.
\end{aligned}
$$

# Intervalos de credibilidade
## Intervalo baseado em quantil

- Suponha que de $N = 10$ sorteios condicionalmente independentes de uma variável aleatória binária, observemos $Y = 2$ ocorrências do evento "um".
- Usando uma distribuição *a priori* Uniforme para $\theta$, a distribuição *a posteriori* é $\theta \mid \{Y = 2\} \sim \text{Beta}(1 + 2, 1 + 8)$. **Veja a conjugação Beta-binomial.**
- Um intervalo de credibilidade de 95% pode ser obtido a partir dos quantis 0,025 e 0,975 desta distribuição Beta.
- Esses quantis valem 0,06 e 0,52 respectivamente, isto é, há 95% de chance da probabilidade *a posteriori* de $\theta \in [0{,}06,\ 0{,}52]$.



# Estimativa de máxima verossimilhança

- Considere o problema de estimar um conjunto de parâmetros $\theta$ de um modelo probabilístico, dado um conjunto de observações $x_1, x_2, \ldots, x_n$.
- As técnicas de máxima verossimilhança assumem que
  1. As amostras não dependem umas das outras, em que a ocorrência de um não tem efeito sobre os outros.
  2. Cada um deles pode ser modelado exatamente da mesma maneira.
- Isso significa que os eventos são independentes e identicamente distribuídos (i.i.d.).
- A suposição sobre i.i.d. implica que um modelo para a função densidade de probabilidade conjunta para todas as observações consiste no produto do mesmo modelo de probabilidade $p(x_i \mid \theta)$ aplicado a cada observação independentemente.
- Para $n$ observações, isso pode ser escrito como

$$p(x_1, x_2, \ldots, x_n \mid \theta) = p(x_1 \mid \theta)\, p(x_2 \mid \theta) \ldots p(x_n \mid \theta)$$

- Cada função $p(x_i \mid \theta)$ tem os mesmos valores de parâmetro $\theta$, e o objetivo da estimativa de parâmetro é maximizar um modelo de probabilidade conjunta desta forma.

# Estimativa de máxima verossimilhança

- Como as observações não mudam, este valor só pode ser alterado alterando a escolha dos parâmetros $\theta$.

$$L(\theta \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

- Como os dados são fixos, é sem dúvida mais útil pensar nisso como uma função de verossimilhança para os parâmetros, que somos livres para escolher.
- Multiplicar muitas probabilidades pode levar a números muito pequenos e, portanto, as pessoas geralmente trabalham com o logaritmo da probabilidade, ou log-verossimilhança:

$$\ln L(\theta \mid x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \ln p(x_i \mid \theta),$$

- Como os logaritmos são funções estritamente crescentes monotonicamente, maximizar a probabilidade logarítmica é o **mesmo** que maximizar a verossimilhança:

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \sum_{i=1}^{n} \ln p(x_i \mid \theta).$$

# Máximo *a posteriori*

- A máxima verossimilhança assume que todos os valores de parâmetros são **igualmente prováveis** *a priori*: não julgamos alguns valores de parâmetros como mais prováveis do que outros antes de considerarmos as observações.
- Em vez de simplesmente computar a estimativa de máxima verossimilhança, ainda podemos obter alguns dos benefícios da abordagem bayesiana ao permitir que a distribuição *a priori* **influencie** a escolha da estimativa pontual.
- Uma maneira racional de fazer isso é escolher a estimativa do ponto **máximo *a posteriori*** (MAP).
- A estimativa MAP escolhe o ponto de probabilidade *a posteriori* máxima (ou densidade de probabilidade máxima no caso mais comum de $\theta$ contínuo):

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta \mid x) = \arg\max_{\theta} \{\ln p(x \mid \theta) + \ln p(\theta)\}$$

- Observe que, para uma *priori* **uniforme**, o termo $\ln p(\theta)$ é uma constante e a expressão acima coincide então com a solução de máxima verossimilhança.
- Discussão: **Como computar $\theta_{\mathrm{ML}}$ e $\theta_{\mathrm{MAP}}$ de forma eficiente?**