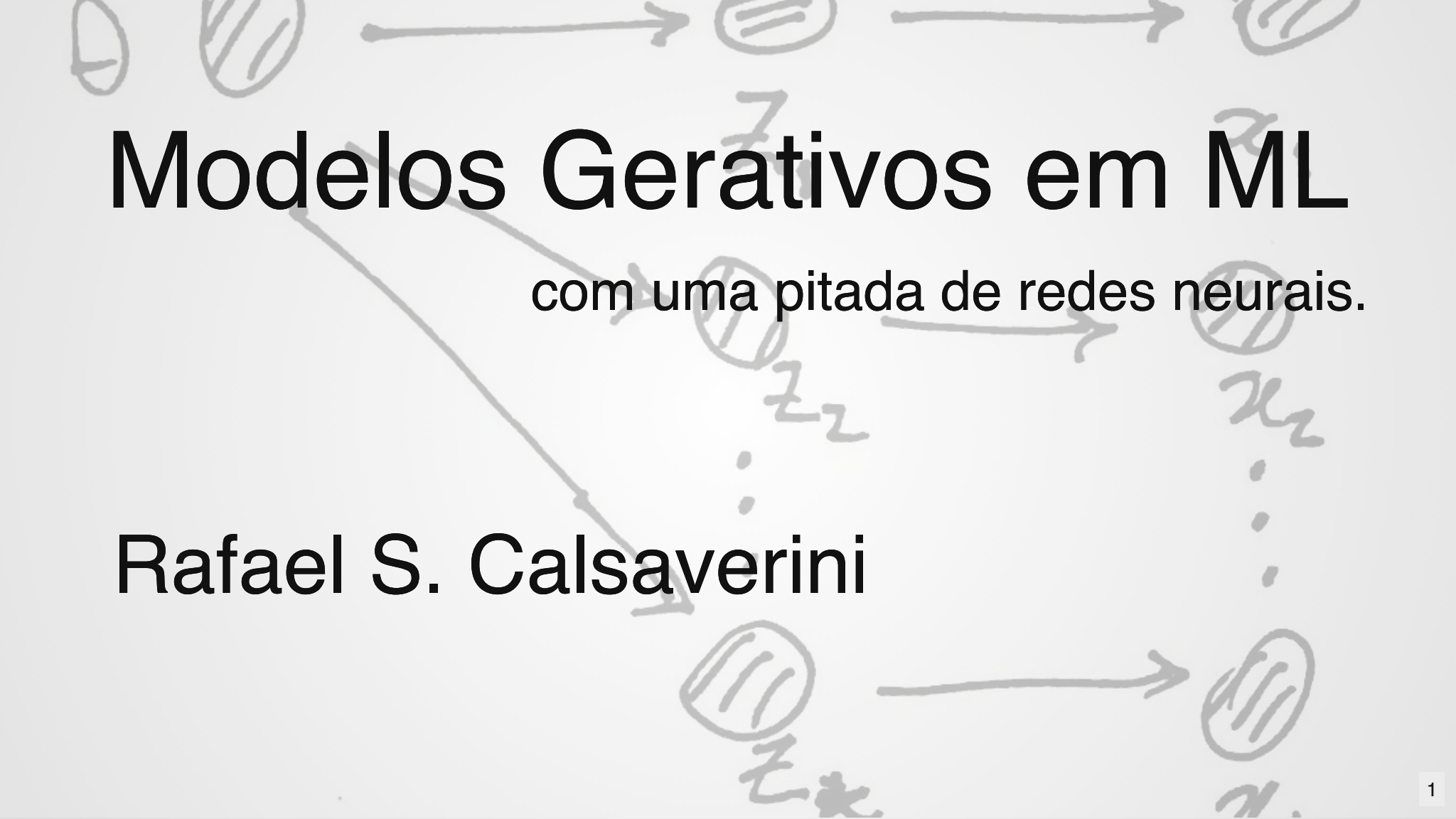


Modelos Gerativos em ML

com uma pitada de redes neurais.

Rafael S. Calsaverini

O que é um modelo de Aprendizagem de Máquina?

Modelos discriminativos

Entradas:

Dados: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$

Parâmetros: $\Theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$

Em que o algoritmo consiste?

Aprendizado: $\hat{\Theta} = \mathrm{Learn}(D)$

Inferência: $y = \mathrm{Infer}(X, \hat{\Theta})$

} Tipicamente tratado como um problema de otimização

Abordagem probabilística

Um modelo discriminativo versa a respeito do mecanismo pelo qual as entradas geram as saídas:

$$P(y|X,\theta)$$

- y target, saída, variável a ser predita
- X features, entradas, variáveis independentes
- θ parâmetros do modelo

Por exemplo, um modelo de regressão:

$$f(X,\theta)$$

$$y \sim \text{Normal}(f(X,\theta),\sigma)$$

$$P(y|X,\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma}\left(y - f(X,\theta)\right)^2\right)$$

Aprendizado vs. Inferência

$P(y|X,\theta)$

y target, saída, variável a ser predita

X features, entradas, variáveis independentes

θ parâmetros do modelo

Aprendizado: dado um conjunto de dados, desejamos encontrar os parâmetros do modelo[1]

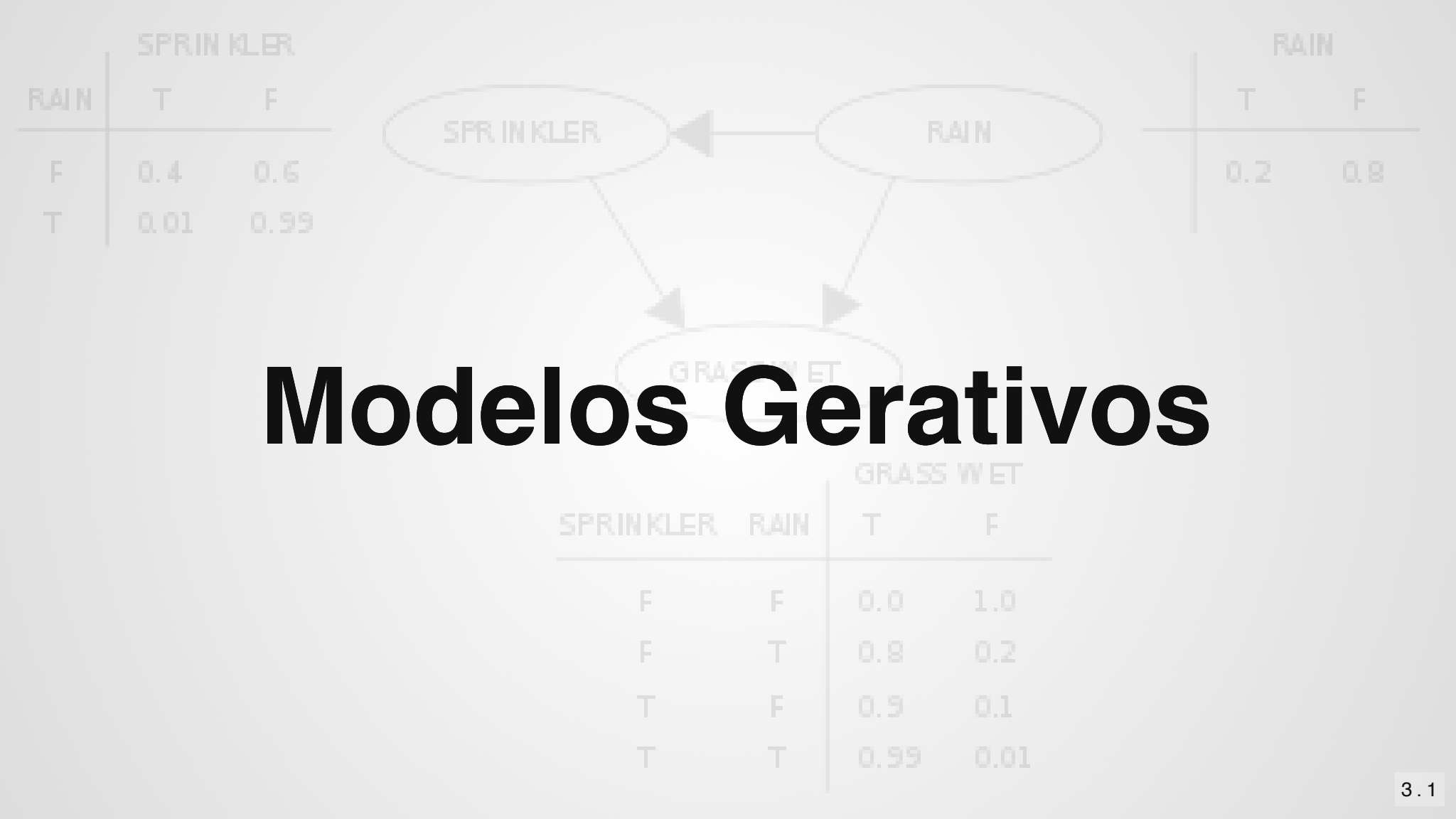
$$\theta^\star = \arg\max_\theta \log P(y|X,\theta)$$

Inferência: dado um conjunto de parâmetros conhecido e uma entrada, prever a saída.

$$\hat{y} = \arg\max_y P(y|X,\theta^\star)$$

$$\hat{y} = E[y|X,\theta^\star]$$

$$\ldots$$



Modelos Gerativos

Modelos generativos

Uma prescrição de como todas as variáveis (observáveis ou não) são gerados!!!

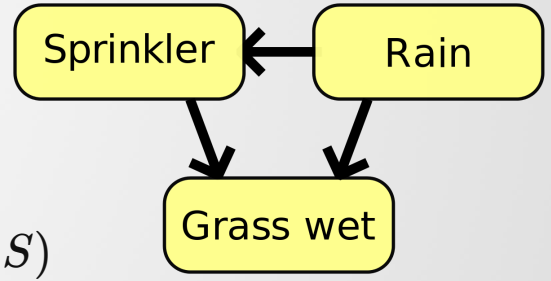
$$P(\text{todas as variáveis observáveis e não observáveis})$$

... e é sobre isso que vamos passar as próximas 3 horas falando.

Inferência

Inferência: dado um modelo probabilístico sobre todas as variáveis e medições das variáveis observadas, estimar as variáveis não-observadas.

Exemplo clássico:



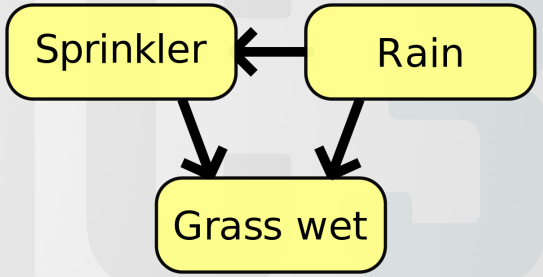
$$P(R,S,G) = P(G|R,S)P(S|R)P(R)$$

$$P(R,S|G) = \frac{P(G|R,S)P(R)P(S)}{P(G)} = \frac{P(G|R,S)P(R)P(S)}{\sum_{R,S} P(G|R,S)P(R)P(S)}$$

Aprendizado

Aprendizado: o modelo probabilistico pode ter parâmetros. Com um conjunto de dados externos esses parâmetros (ou suas distribuições) devem ser aprendidos para posterior inferência.

Exemplo clássico:



$$P(R, S, G|\theta) = P(G|R, S, \theta)P(S|R, \theta)P(R|\theta)$$

$$P(R, S, G, \theta) = P(G|R, S, \theta)P(S|R, \theta)P(R|\theta)P(\theta)$$

$$\theta, Z \sim P(\theta, Z|X)$$

Máxima verossimilhança

$P(X, Z|\theta)$

X variáveis observáveis

Z variáveis não-observáveis

θ parâmetros do modelo

Aprendizado:

$$\theta^\star = \arg\max_\theta \log \left(\sum_Z P(X, Z|\theta) \right)$$

Inferência:

$$X, Z \sim P(X, Z|\theta^\star)$$

ou

$$Z \sim P(Z|\theta^\star, X) = \frac{P(X,Z|\theta^\star)}{\sum_Z P(X,Z|\theta^\star)}$$

~~Máxima verossimilhança~~ Máximo a Posteriori

X variáveis observáveis

$P(X, Z|\theta)P(\theta)$ Z variáveis não-observáveis

θ parâmetros do modelo

Aprendizado:

$$\theta^\star = \arg\max_\theta \log \left(\sum_Z P(X, Z|\theta)P(\theta) \right)$$

Inferência:

$$X, Z \sim P(X, Z|\theta^\star)$$

ou

$$Z \sim P(Z|\theta^\star, X) = \frac{P(X,Z|\theta^\star)}{\sum_Z P(X,Z|\theta^\star)}$$

Inferência Bayesiana

$$P(X, Z|\theta)P(\theta)$$

X variáveis observáveis

Z variáveis não-observáveis

θ parâmetros do modelo

Aprendizado (teorema de Bayes):

$$P(\text{Não observado}|\text{Observado}) = \frac{P(\text{Observado}|\text{Não observado})P(\text{Não observado})}{P(\text{Observado})}$$

$$P(\theta, Z|X) = \frac{P(X|\theta,Z)P(\theta,Z)}{P(X)}$$

Inferência:

$$\theta, Z \sim P(\theta, Z|X)$$

Inferência Variacional

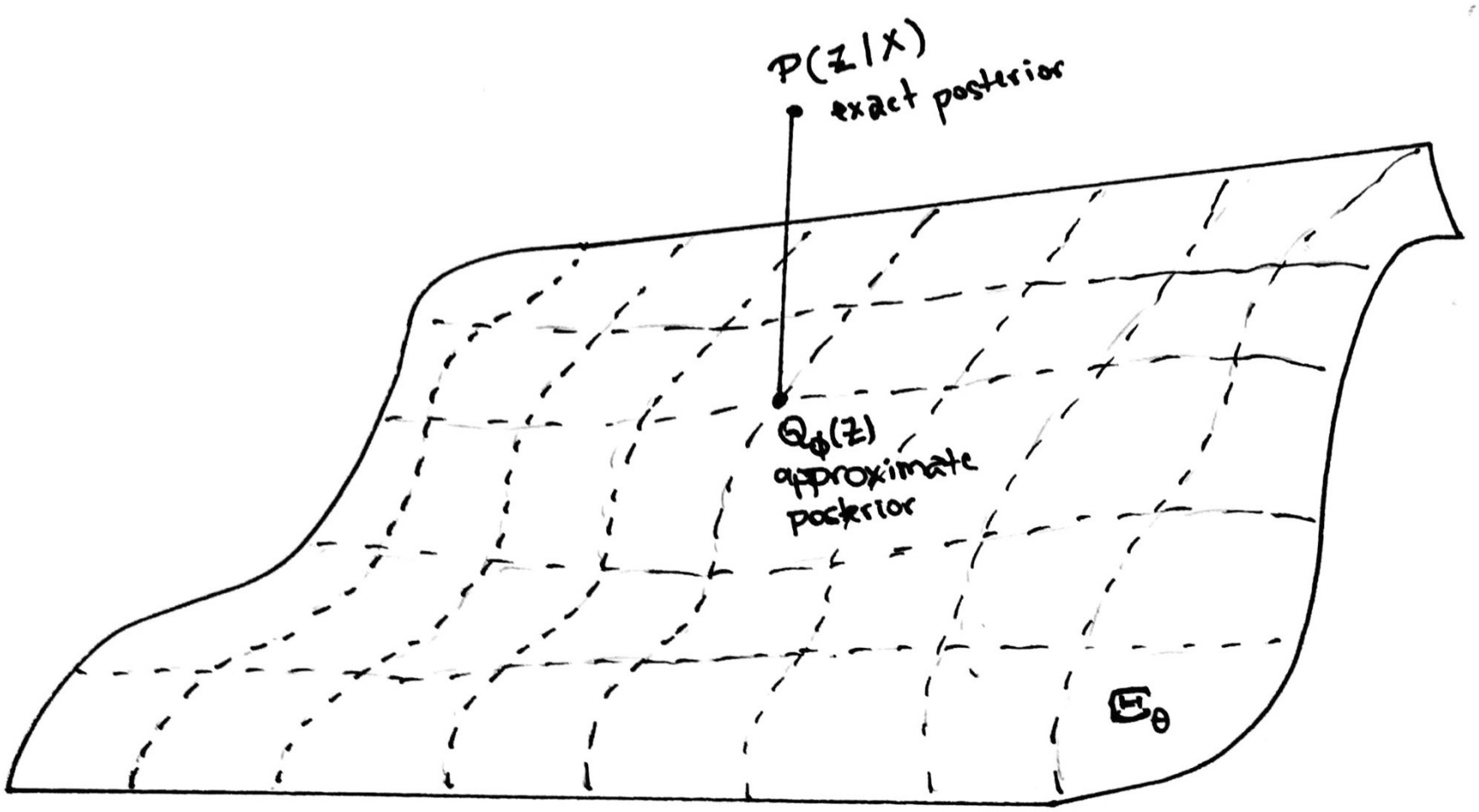
Aprendizado:

$$Q(\theta, Z) =$$ A distribuição mais parecida com P(θ, Z|X) que eu conseguir achar.

$$Q(\theta, Z) = \min_Q \text{Divergência}[Q(\theta, Z); P(\theta, Z|X)]$$

Inferência:

$$\theta, Z \sim Q(\theta, Z|X)$$


$p(z|x)$
exact posterior
$q_\phi(z)$
approximate
posterior
Ξ_θ

... e as redes neurais?

Modelos discriminativos

Nosso exemplo de regressão:

$$y \sim P(y|f(X,\theta))$$

$$\theta^\star = \arg\max_\theta \log P(y|X,\theta)$$

$$\theta^\star = \arg\min_\theta \text{Loss}\,(y, f(X,\theta))$$

Distribuição	**Loss equivalente**
Normal (gaussiana)	Erro quadrático médio
Exponencial	Erro absoluto médio
Binomial	Entropia cruzada

Modelos gerativos

Redes neurais podem ser usadas para representar **distribuições complexas em espaços de dimensionalidade alta**, e ainda assim prover **formas eficientes de aprendizado e inferência**.

- Máxima verossimilhança / máximo a posteriori
- Inferência variacional
- Modelos adversários

$$P(X|f(Z,\theta))P(Z)$$

Bibliotecas

Inferência bayesiana via MCMC:

- pymc & pymc3
- pystan
- emcee

Inferência Variacional:

- bayespy

Inferência Variacional com redes neurais:

- Edward (tensorflow)
- tf.probability ("Edward2")
- pyro (pytorch)


That's al

O que raios são probabilidades?

Considere as sentenças:

- *A **probabilidade** de obter cara em um lance de cara ou coroa é ½.*
- *Com a saída dos EUA, o acordo **provavelmente** será cancelado.*
- *A **probabilidade** de chuva amanhã é em torno de 80%.*
- *Vou chamar o Rafael para sair. Quais são as minhas **chances**?*
- *Diante das evidências é muito **improvável** que o réu seja inocente.*

Probabilidade vs. Frequência

É comum ver a definição *"probabilidade é o limite para a freqüência de um evento em um grande número de experimentos repetidos"*:

$$\text{Prob}\,(A) = \lim_{N \to \infty} \frac{\text{ocorrências do evento A}}{N}$$

Será que essa definição é de fato útil?
Será que sequer é uma definição?
Será que bate com o nosso conceito intuitivo de probabilidade?

Lógica e incerteza

Considere as seguintes frases:

- "Quando chove, a grama fica molhada. Choveu, **logo** a grama está molhada."
- "Quando chove, a grama fica molhada. A grama está molhada, **logo**..."

Probabilidades!

Probabilidades são uma forma de raciocinar sobre informação incompleta e incerteza:

- "Quando chove, a grama fica molhada. A grama está molhada, **isso aumenta minha suspeita** de que tenha chovido."

$$\text{Prob}\left(\text{chuva}|\text{grama molhada}\right) = \frac{\text{Prob}(\text{grama molhada}|\text{chuva})\text{Prob}(\text{chuva})}{\text{Prob}(\text{grama molhada})}$$

Em resumo...

Se $a \Rightarrow b$, e sabemos que a é verdadeiro.

- O lógico sabe que **b** é verdade.
- O probabilista atribui a **b** probabilidade 1.

Se $a \Rightarrow b$, e sabemos que b é verdadeiro.

- O lógico não sabe dizer nada sobre **a**.
- O probabilista sabe atualizar seu probabilidade de **a**.

$$P(b|a, \text{Model}) = \text{modelo causal conectando } a \text{ e } b$$

$$P(a|b, \text{Model}) = \frac{P(b|a, \text{Model})P(a|\text{Model})}{P(b|\text{Model})}$$