Lecture 4: 1/30/17

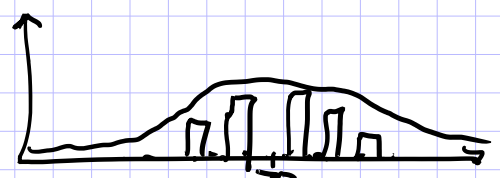
Natural Data

naturally occurring stochastic processes

Ex 1: Coin Flipping
H T H H T T H H ...

Ex 2: Dice Rolls
3 2 2 1 6 5 4 1 2 ...

Ex 3: Background Test taken by 601 Students



73, 100, 40, 68, 74, ...

Bernoulli($\phi = 0.5$)

Categorical($\vec{\theta}$)
$\theta_1 = 1/6$, $\theta_2 = 1/6$ ... $\theta_6 = 1/6$

Gaussian(μ, σ^2)
$\mu = 70$ $\sigma = 10$

Ex 2: Dice Rolls

$Pr(die = 1) = 1/6$, $Pr(die = 2) = 1/6$, ... $Pr(die = 6) = 1/6$

$Pr(die = 0) = 0$, $Pr(die = 7) = 0$

event: $die = 0$ probability: 0

Generate Synthetic Data

Why?

1. Plots for ML
2. Helpful input when debugging ML
3. Helpful output of (certain) ML algos
4. Inference (see Gibbs Sampling)
5. Tells about the dist.

Def: Random variable X: value is the outcome of a random experiment.

$Pr(X = x)$ — r.v.s by capital letters, values by lowercase

event: $X = x$

$X^{(1)}$ $X^{(2)}$ $X^{(N)}$

$D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$

If a r.v. X follows a Bernoulli w/parameter ϕ:

$X \sim Bernoulli(\phi)$

"is sampled from"

Categorical:

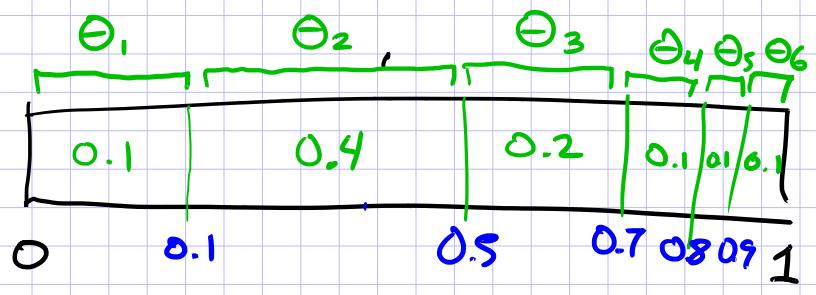
$X \sim \text{Categorical}(\vec{\theta})$ $\quad K = |\vec{\theta}|$ $\quad \vec{\theta}^T = [\theta_1, \theta_2 \cdots \theta_K]$

$\sum_{k=1}^{K} \theta_k = 1$, $\quad 0 \le \theta_k \le 1$, $\forall k$

$Pr(X = x) = \theta_x$

The ~~probability mass function~~ (pmf) for Cat:

$p(x) \triangleq Pr(X = x) = \theta_x$



rand() ← samples uniformly from [0, 1]

[1 1 1 ⋯ 1 | 2 2 2 ⋯ ⋯ 2 2 1 | 3 3 ⋯ 3 | 4 4 | 5 5 1 6 6]

10% 40% 20% ω ω

rand() = 0.4 $\Rightarrow x^{(1)} = 2$

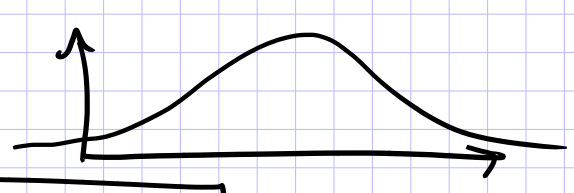
rand() = 0.6 $\Rightarrow x^{(2)} = 3$

rand() = 0.05 $\Rightarrow x^{(3)} = 1$

⋮

Bernoulli: Just Categorical $K = 2$ (coin flipping)

Gaussian:



$X \sim \text{Gaussian}(\mu, \sigma^2)$

↑ continuous r.v.

Data Likelihood

Assume: Sample in D are iid

↑ "independent and identically distributed"

① Indep. $Pr(X^{(1)} = 6, X^{(2)} = 3) = Pr(X^{(1)} = 6)\, Pr(X^{(2)} = 3)$

shorcut → $Pr(X^{(1)}, X^{(2)}) = Pr(X^{(1)})\, Pr(X^{(2)})$

$\forall a, b \quad Pr(X^{(1)} = a, X^{(2)} = b) = Pr(X^{(1)} = a)\, Pr(X^{(2)} = b)$

② Ident. Dist. each r.v. follows the same distribution

Ex: $X^{(i)} \sim \text{Categorical}(\vec{\theta})$

$Pr(X^{(i)} = x^{(i)}) = \theta_{x^{(i)}}$, $\forall i$

Mult($\vec{\theta}$, N) where $N = 1$ ≡ Cat($\vec{\theta}$)

Ex: Dice Rolls

Q: What is the data likelihood of D under $Cat(\vec{\theta})$?

A: $p(D|\vec{\theta}) = p(X^{(1)}=x^{(1)}, X^{(2)}=x^{(2)}, \dots, X^{(N)}=x^{(N)} | \vec{\theta})$

$= p(X^{(1)}=x^{(1)}|\vec{\theta})\, p(X^{(2)}=x^{(2)}|\vec{\theta}) \cdots p(X^{(N)}=x^{(N)}|\vec{\theta})$ ← by indep.

$= \prod_{i=1}^{N} p(X^{(i)}=x^{(i)})$ ← by indent. dist.

$= \prod_{i=1}^{N} \theta_{x^{(i)}}$

Learning from Data

MLE: The principle of Maximum likelihood estimation (MLE):

Choose parameters that make the data "most likely".

Assumptions: Data generated iid from distribution $p^*(x|\vec{\theta}^*)$ and comes from a family of distn parameterized $\theta \in \Theta$ ← set of possible parameters

Formally:

$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(D|\theta)$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}}\ \log p(D|\theta)$$

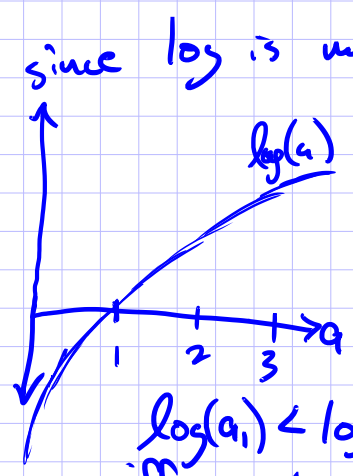
$$= \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell(\theta)$$

usually a continuous optimization

where $\ell(\theta) \triangleq \log p(D|\theta)$ "log-likelihood"

↳ treat as function of θ where D is constant

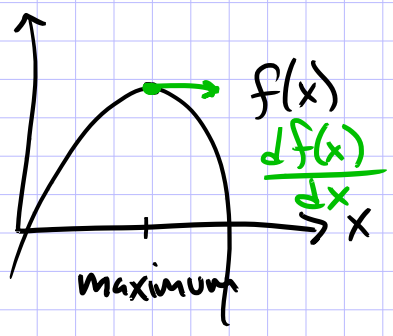
since log is monotonic

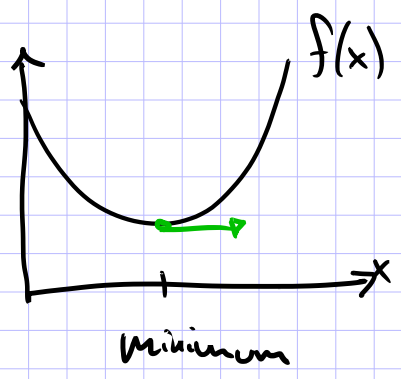


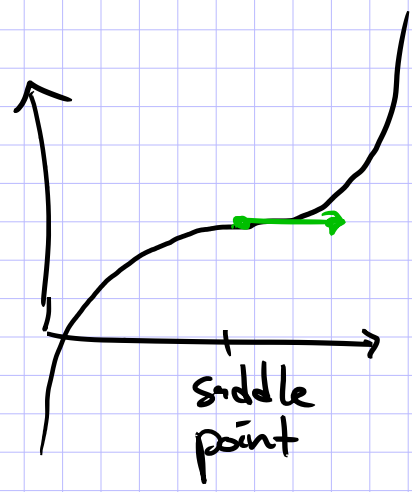
$\log(a_1) < \log(a_2)$ iff $a_1 < a_2$

$\Rightarrow \log(f(a_1)) < \log(f(a_2))$ iff $f(a_1) < f(a_2)$

Zero Derivatives



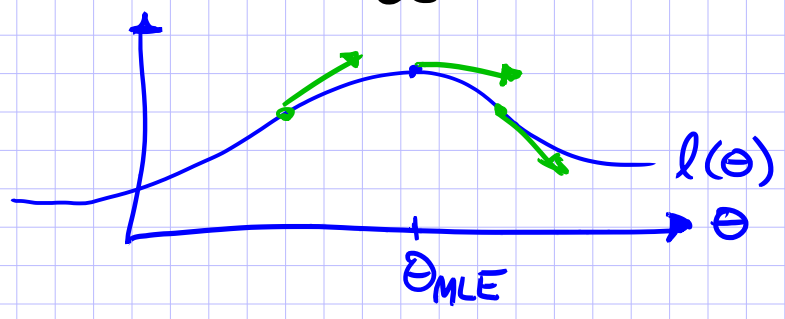




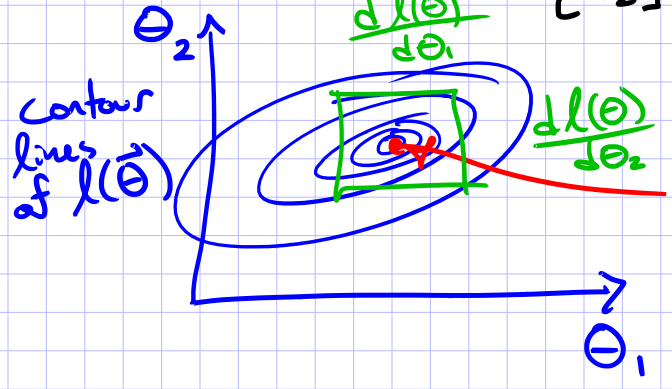
Optimization for MLE

Ex#1: Suppose $\Theta = \mathbb{R}$ ⇒ infinitely large

$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell(\theta)$$



Ex#2: $\Theta = \mathbb{R}^2 \Rightarrow \vec{\theta} = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix}$



$$\vec{\theta}_{MLE} = \underset{\theta_1}{\operatorname{argmax}}\ \underset{\theta_2}{\operatorname{argmax}}\ \ell(\theta_1, \theta_2)$$

$$= \underset{\vec{\theta}}{\operatorname{argmax}}\ \ell(\vec{\theta})$$

Key idea: the top of any hill is flat!

Formally: $\frac{d\ell(\vec{\theta})}{d\theta_k} = 0$ $\forall k$

Ex#3 MLE of Categorical

① Write $\ell(\theta)$

Let N_j = # times j appears in D

Likelihood: $p(D|\vec{\theta}) = \prod_{i=1}^{N} \theta_{x^{(i)}} = \prod_{k=1}^{K} \theta_k^{N_k}$ ← exponent!!!

Log-likelihood:

$$\ell(\theta) = \log \prod_{k=1}^{K} \theta_k^{N_k} = \sum_{k=1}^{K} N_k \log \theta_k$$

② Set deriv to 0

$$\frac{d\,\ell(\theta)}{d\theta_j} = \frac{d}{d\theta_j}\left(\sum_{k=1}^{K} N_k \log \theta_k\right) = \frac{d}{d\theta_j} N_j \log \theta_j = \frac{N_j}{\theta_j}$$

↑ $N_1 \log \theta_1 + N_2 \log \theta_2 + \cdots + N_k \log \theta_k$

$$\frac{d\,\ell(\theta)}{d\theta_j} = 0 \Rightarrow \frac{N_j}{\theta_j} = 0 \Rightarrow \theta_j = +\infty$$ ← Wrong!!!

WRONG

We forgot that $\sum_{k=1}^{K} \theta_k = 1$

The set of valid parameters

$$\Theta = \{\vec{\theta} : \vec{\theta} \in \mathbb{R}^K \text{ and } \sum_{k=1}^{K} \theta_k = 1\}$$

$$\theta_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell(\theta)$$