STAT 391

4/18/23

Lecture 8

Quiz 1
on 4/20
at 12:30

Smoothing example
Continuous S +
parametric

Sol 1, 2
L IV posted

Lecture Notes III: Discrete probability in practice – Small Probabilities

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

April, 2023

Definitions and setup

We will look at estimating categorical distributions from samples, when the number of outcomes m is large.

- Let $S = \{1, \ldots m\}$ be the sample space, and $P = (\theta_1, \ldots \theta_m)$ a distribution over S.
- We draw n independent samples from P, obtaining the **data set** $\mathcal{D}$
- Define **the counts** $\{n_j = \#j$ appears in $\mathcal{D},\ i = 1, \ldots n\}$. The counts are also called **sufficient statistics** or **histogram**.
- Define the **fingerprint** (or **histogram of histogram**) of $\mathcal{D}$ as the counts of the counts, i.e $\{\, r_k = \#\text{counts}\, n_j = k,\ \text{for } k = 0, 1, 2 \ldots\}$

Example $m = 26$ alphabet letters

Data	**Counts** n_i	**Fingerprint** r_k
`the red fox is quick` $n = 16$ letters	$n_j = 0$: `a,b,g,j,l,m,n,p,v,w,y,z` $n_j = 1$: `c,d,f,h,k,o,q,r,s,t,u,x` $n_j = 2$: `e,i`	$r_0 = 12 = \lvert\{\texttt{a,b,g,...,y,z}\}\rvert$ $r_1 = 12 = \lvert\{\texttt{c,d,f,h,...,u,x}\}\rvert$ $r_2 = 2 = \lvert\{\texttt{e,i}\}\rvert$ $r_3 = \ldots r_n = 0$
`ho ho who s on first` $n = 15$ letters	$n_j = 0$: `a,b,c...,x,z` $n_j = 1$: `f,i,n,r,t,w` $n_j = 2$: `s` $n_j = 3$: `h` $n_j = 4$: `o`	$r_0 = 26 - 6 - 1 - 1 - 1 = 17$ $r_1 = 6 = \lvert\{\texttt{f,i,n,r,t,w}\}\rvert$ $r_2 = 1 = \lvert\{\texttt{s}\}\rvert$ $r_3 = 1 = \lvert\{\texttt{h}\}\rvert$ $r_4 = 1 = \lvert\{\texttt{o}\}\rvert$

- It is easy to verify that $n_j \in 0 : n$, hence $r_{0:n}$ may be non-zero (but $r_{n+1,n+2,\ldots} = 0$), and that

$$m \;=\; r_0 + r_1 + \ldots r_n \quad n \;=\; 0 \times r_0 + 1 \times r_1 + \ldots k \times r_k + \ldots \tag{1}$$

Smoothing on an example

$n =$ sample size

$\boxed{m} =$ alphabet size $= |S|$

$(a)_+ = \begin{cases} a & \text{if } a \geq 0 \\ 0 & \text{otherwise} \end{cases}$

- **the counts** $\{n_j = \#j$ appears in $\mathcal{D}, \ i = 1, \ldots n\}$ (or **sufficient statistics** or **histogram**)
- **fingerprint** (or **histogram of histogram**) of $\mathcal{D}$ as the counts of the counts $\{ r_k = \#\text{counts } n_j = k, \text{ for } k = 0, 1, 2 \ldots\}$, and $R_k = \{ j, \ n_j = k, \}$

$m = r + r_0$

obs, unobs

Example $\boxed{m = 26}$ alphabet letters

Data

`the red fox is quick`

$n = 16$ letters

Counts n_i

$n_j = 0$: a,b,g,j,l,m,n, p,v,w,y,z

$n_j = 1$: c,d,f,h,k,o,q,r,s,t,u,x

$n_j = 2$: e,i

$\Rightarrow \theta^{ML}_{a,b} = 0$

Fingerprint r_k

$r_0 = 12 = |\{$a,b,g,...,y,z$\}|$

$r_1 = 12 = |\{$c,d,f,h,...,u,x$\}|$

$r_2 = 2 = |\{$e,i$\}|$

$r_3 = \ldots r_n = 0$

$n_j = 0$ $\qquad$ $n_j > 0$ $\qquad$ $r = r_1 + r_2 + \ldots$

(Lap) $n_j \leftarrow n_j + 1$

NE $\quad \theta_j^{NE} = \dfrac{r/m}{n}$ $\qquad$ $\theta_j^{NE} = \dfrac{(n_j - 1)_+ + r/m}{n}$ $\quad n_j^{NE}$

WB $\quad P[NEW] = \dfrac{r}{n}$ $\qquad$ $\theta_j^{WB} = \dfrac{1}{r_0} \dfrac{\frac{r}{n}}{1 + \frac{r}{n}}$ $\qquad$ $\theta_j^{WB} = \dfrac{n_j}{n} \cancel{\left(1 - \dfrac{r}{n}\right)} \dfrac{1}{1 + \frac{r}{n}}$

CORRECTION !!

GT $\quad P_1 = \dfrac{r_1}{n}$ $\qquad$ $\theta_j^{GT} = \dfrac{1}{r_0} \dfrac{r_1}{n}$ $\qquad$ $\theta_j^{GT} = \dfrac{n_j}{n} \left(1 - \dfrac{r_1}{n}\right)$

Smoothing on an example

- **the counts** $\{n_j = \#j \text{ appears in } \mathcal{D},\ i = 1, \ldots n\}$ (or **sufficient statistics** or **histogram**)
- **fingerprint** (or **histogram of histogram**) of $\mathcal{D}$ as the counts of the counts $\{\ r_k = \#\text{counts}\ n_j = k,\ \text{for}\ k = 0, 1, 2 \ldots\}$, and $R_k = \{\ j,\ n_j = k, \}$

Example $m = 26$ alphabet letters

Data	**Counts** n_j	**Fingerprint** r_k
`the red fox is quick`	$n_j = 0$: `a,b,g,j,l,m,n,p,v,w,y,z`	$r_0 = 12 = \lvert\{$`a,b,g,...,y,z`$\}\rvert$
$n = 16$ letters	$n_j = 1$: `c,d,f,h,k,o,q,r,s,t,u,x`	$r_1 = 12 = \lvert\{$`c,d,f,h,...,u,x`$\}\rvert$
	$n_j = 2$: `e,i`	$r_2 = 2 = \lvert\{$`e,i`$\}\rvert$
		$r_3 = \ldots r_n = 0$

$n_j = 0$

$r = m - r_0 = 26 - 12 = 14$

$$\theta_j^{NE} = \frac{r/m}{n} = \frac{7}{13 \cdot 16}$$

$r/m = \frac{14}{26} = 7/13$

a, b, g ... z

$$\theta_j^{WB} = \frac{1}{r_0}\frac{r}{n} = \frac{7}{\cancel{8}\ 15} \cdot \frac{1}{12} > \theta_j^{NE}$$

$r/n = \frac{14}{16} = \frac{7}{8}$

$r_0 = 12$

$$\theta_j^{GT} = \frac{1}{r_0}\frac{r_1}{n} = \frac{3}{4} \cdot \frac{1}{12} = \frac{1}{16} < \theta_j^{WB}$$

always

$\frac{r_1}{n} = \frac{12}{16} = \frac{3}{4}$

$$\frac{r/n}{1 + r/n} = \frac{r}{n+r} = \frac{14}{30} = \frac{7}{15}$$

Smoothing on an example

- **the counts** $\{n_j = \#j \text{ appears in } \mathcal{D},\ i = 1, \dots n\}$ (or **sufficient statistics** or **histogram**)
- **fingerprint** (or **histogram of histogram**) of $\mathcal{D}$ as the counts of the counts $\{\, r_k = \#\text{counts}\, n_j = k,\ \text{for } k = 0, 1, 2 \dots\}$, and $R_k = \{\, j,\ n_j = k, \}$

Example $m = 26$ alphabet letters

Data	Counts n_j	Fingerprint r_k
`the red fox is quick`	$n_j = 0$: `a,b,g,j,l,m,n,p,v,w,y,z`	$r_0 = 12 = \lvert\{\texttt{a,b,g,...,y,z}\}\rvert$
$n = 16$ letters	$n_j = 1$: `c,d,f,h,k,o,q,r,s,t,u,x`	$r_1 = 12 = \lvert\{\texttt{c,d,f,h,...,u,x}\}\rvert$
	$n_j = 2$: `e,i`	$r_2 = 2 = \lvert\{\texttt{e,i}\}\rvert$
		$r_3 = \dots r_n = 0$

$n_j=1$

$r = m - r_0 = 26 - 12 = 14$

$$\theta^{NE}_j = \frac{(n_j-1)_+ + r/m}{n}$$

$\Rightarrow \theta^{NE}_{c,d,\dots} = \frac{7}{13 \cdot 16}$

$\Rightarrow \theta^{NE}_{e,i} = \frac{1 + 7/13}{16}$

$r/m = \frac{14}{26} = 7/13$

$r/n = \frac{14}{16} = \frac{7}{8}$

$$\theta^{WB}_j = \frac{n_j}{n}\left(1 - \frac{r}{n}\right)$$

$\Rightarrow \theta^{WB}_{c,d,\dots} = \frac{1}{16} \cdot \cancel{\frac{1}{8}}\ \frac{8}{15}$

$\Rightarrow \theta^{WB}_{e,i} = \frac{2}{16} \cdot \cancel{\frac{1}{8}}\ \frac{8}{15}$

$r_0 = 12$

$\frac{r_1}{n} = \frac{12}{16} = \frac{3}{4}$

$$\theta^{GT}_j = \frac{n_j}{n}\left(1 - \frac{r_1}{n}\right)$$

$\frac{1}{1+\frac{r}{n}} = \frac{n}{n+r} = \frac{16}{30} = \frac{8}{15}$

$\cancel{1 - \frac{r}{n} = \frac{1}{8}}$

(Estimating)

Lecture Notes IV – Continuous distributions. Parametric density estimation.

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

April, 2023

CDF and PDF. Sampling

Examples of continuous distributions

ML estimation for continuous distributions

ML estimation by gradient ascent

Reading: Ch.5, 6

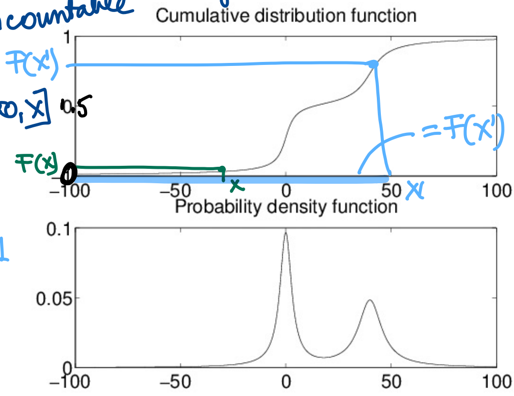
CDF and PDF refresher

$S = (-\infty, \infty)$ or uncountable e.g. $S = (a,b)$

Cumulative distribution function (CDF)

$$F(x) \;=\; P[X \le x] = Pr(-\infty, x] \qquad (1)$$

1. $F \;\ge\; 0$ positivity.
2. $\lim\limits_{x \to -\infty} F = 0$ $= Pr[-\infty]$
3. $\lim\limits_{x \to \infty} F = 1$ $= Pr(-\infty, \infty) = 1$
4. F is an increasing function



Probability density [function] (PDF)

$$f \;=\; \frac{dF}{dx} \qquad (2)$$

$$P(a,b) \;=\; P[a,b] \;=\; F(b) - F(a) \;=\; \int_a^b f(x)dx \qquad (3)$$

$\Rightarrow F(b) \ge F(a)$

normalization condition

$$\int_{-\infty}^{\infty} f(x)dx \;=\; 1 \qquad (4)$$

CDF and PDF refresher

Cumulative distribution function (CDF)

$$F(x) = P[X \le x] \qquad (1)$$

1. $F \ge 0$ positivity.
2. $\lim_{x \to -\infty} F = 0$
3. $\lim_{x \to \infty} F = 1$
4. F is an increasing function

Probability density [function] (PDF)

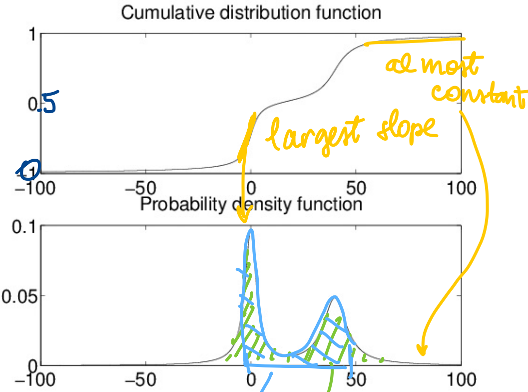
$$\boxed{f = \frac{dF}{dx}} \qquad (2)$$

$$P(a,b) = \underline{P[a,b]} = \underline{F(b) - F(a)} = \underbrace{\int_a^b f(x)dx}_{\text{area over }(a,b)} \qquad (3)$$

normalization condition

2. $$\boxed{\int_{-\infty}^{\infty} f(x)dx = 1} \qquad (4)$$ area under $f(x)$

1. $f(x) \ge 0$ for all x



area = 1

$f(x) \approx 0$

Notation

$P[E]$
↑ event

$[a, \infty)$, $(-\infty, b)$, ...

in $(-\infty, \infty)$, events are $[a,b)$, (a,b), $[a,b]$, ...

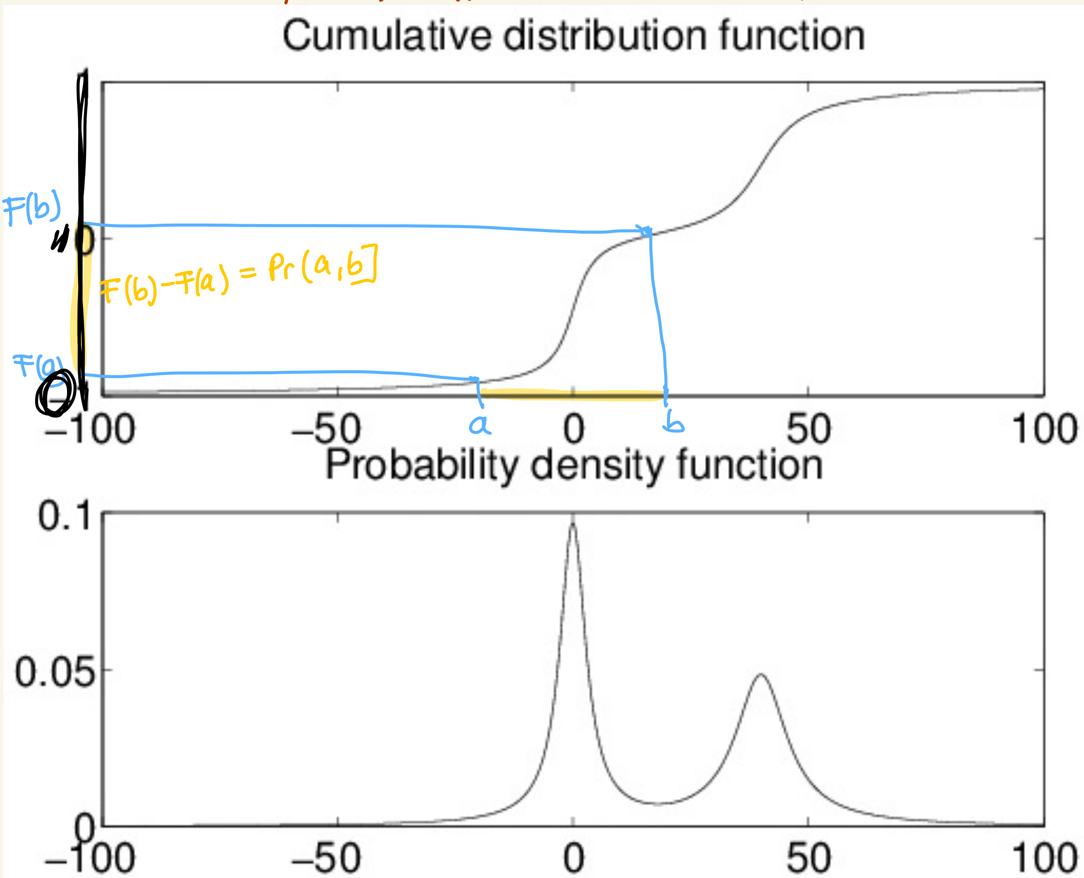
or unions of intervals

$P[\,[a,b)\,] \equiv P[a,b)$

simplification

Assume 1) $F(x)$ continuous $\Rightarrow$ $Pr[a] \equiv Pr[a,a] = F(a) - F(a) = 0$

2) $F(x)$ differentiable $\Rightarrow$ $f(x)$ exists



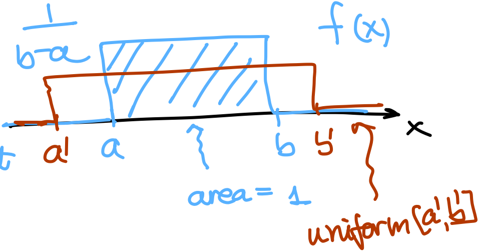
Examples of continuous distributions

family of distributions

$$\mathcal{F}_1 = \{u_{[a,b]},\ a < b\} \quad \text{uniform} \tag{5}$$

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

constant



$$\mathcal{F}_2 = \{N(.; \mu, \sigma^2)\} \quad \text{normal} \tag{7}$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

$$F(x; a, b) = \frac{1}{1+e^{-ax-b}}, \quad a > 0 \quad \text{logistic} \tag{9}$$

$$f(x; a, b) = \frac{ae^{-ax-b}}{(1+e^{-ax-b})^2} \tag{10}$$

~~uniform~~ $(-\infty, \infty)$

$$\int_{-\infty}^{\infty} c\,dx = \infty$$

for any $c > 0$

$\mathcal{F}_1$ has parameters a, b, $a < b$

Examples of continuous distributions

Normal: μ = mean $\in (-\infty, \infty)$ } parameters for $\mathcal{F}_2$
σ^2 = variance > 0

$$\mathcal{F}_1 = \{u_{[a,b]},\ a < b\} \quad \text{uniform} \tag{5}$$

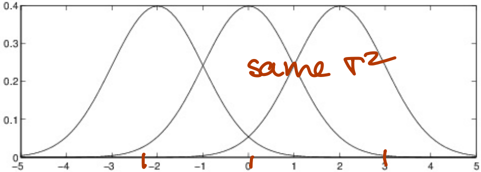
$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

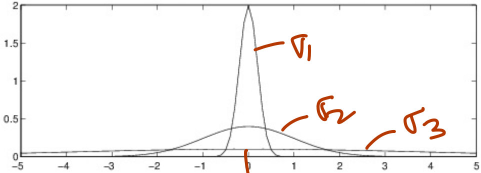
$$\mathcal{F}_2 = \{N(.; \mu, \sigma^2)\} \quad \text{normal} \tag{7}$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

$$F(x; a, b) = \frac{1}{1 + e^{-ax-b}}, \quad a > 0 \quad \text{logistic} \tag{9}$$

$$f(x; a, b) = \frac{ae^{-ax-b}}{(1 + e^{-ax-b})^2} \tag{10}$$





$\sigma_3 > \sigma_2 > \sigma_1$

Examples of continuous distributions

a, b parameters for $\mathcal{F}_3$

a > 0

$$\mathcal{F}_1 = \{u_{[a,b]},\ a < b\} \quad \text{uniform} \qquad (5)$$

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \qquad (6)$$

$$\mathcal{F}_2 = \{N(.; \mu, \sigma^2)\} \quad \text{normal} \qquad (7)$$

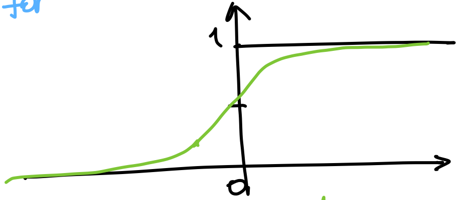
$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (8)$$

$$F(x; a, b) = \frac{1}{1 + e^{-ax-b}}, \quad a > 0 \quad \text{logistic} \qquad (9)$$

$$f(x; a, b) = \frac{ae^{-ax-b}}{(1 + e^{-ax-b})^2} \qquad (10)$$

$\mathcal{F}_3$

Logistic



$a = 1$, $b = 0$ $\Rightarrow F(x) = \dfrac{1}{1+e^{-x}}$

$x = 0 \Rightarrow F(x) = \frac{1}{2}$

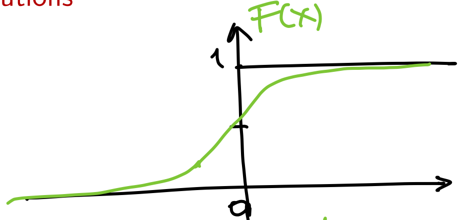
$x \to \infty \Rightarrow F(x) \to 1$

$x \to -\infty \Rightarrow F(x) \to 0$

$F(x) = 1 - F(-x)$

Exercise

Examples of continuous distributions



$$\mathcal{F}_1 = \{u_{[a,b]},\ a < b\} \quad \text{uniform} \tag{5}$$

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

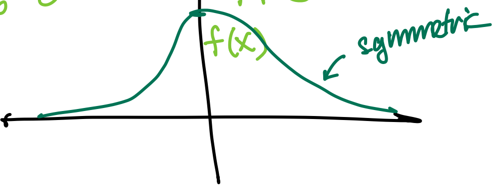
$$\mathcal{F}_2 = \{N(.; \mu, \sigma^2)\} \quad \text{normal} \tag{7}$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

$$F(x; a, b) = \frac{1}{1 + e^{-ax-b}}, \quad a > 0 \quad \text{logistic} \tag{9}$$

$$f(x; a, b) = \frac{ae^{-ax-b}}{(1 + e^{-ax-b})^2} \tag{10}$$

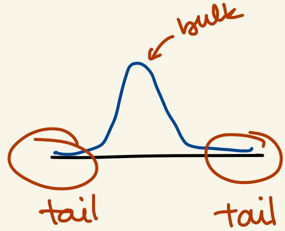
$a=1$, $b=0$ $\Rightarrow F(x) = \frac{1}{1+e^{-x}}$



$f(x) = f(-x)$ for $a=1, b=0$

$\arg\max_x f(x) \Leftrightarrow ax+b=0 \Leftrightarrow x = -\frac{b}{a}$

About "tails"



regions $x \to \pm\infty$

$$N(0,1) \to f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \approx 0$$

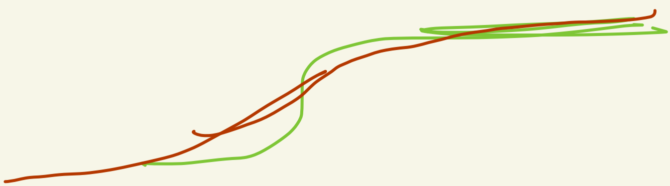
$$\text{Logistic } (a=1, b=0) \to f_2(x) = \frac{e^{-x}}{(1+\underbrace{e^{-x}}_{\approx 1})^2}$$

$(x \to \infty)$ Large

$$\frac{f_1}{f_2} \approx \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{x^2}{2}}}{e^{-x}} = \frac{1}{\sqrt{2\pi}} e^{x-\frac{x^2}{2}} \longrightarrow 0 \; !!$$

x large $\quad x - \frac{x^2}{2} \longrightarrow -\infty$

$$f_1(x) \ll f_2(x) \quad \text{for } |x| \text{ large}$$



ML estimation for continuous distributions = density estimation

$S = (-\infty, \infty)$ or $S \subset (-\infty, \infty)$

Sample space

Data $\mathcal{D} = \{x^1, x^2, \ldots x^n\} \subset S$ iid from true unknown f

Model family $\mathcal{F} = \{f(x|\theta), \theta = \text{parameter(s)}\}$

Normal, Uniform, ...

Chosen

Problem estimate θ from $\mathcal{D}$

Solution : Max likelihood $\to \theta^{ML} \to f(x|\theta^{ML})$

estimated density

ML estimation for continuous distributions

Likelihood of θ $F'(x$

$$L(\theta) = \prod_{i=1}^{n} f(x^i|\theta)$$

$P(x^i|\theta)$ for S discrete

log-likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \ln f(x^i|\theta)$$

Max Likelihood Principle

Choose $\theta^{ML} = \operatorname{argmax} \ell(\theta)$

x^i observed

$$P[x^i] = 0$$

$$P[x^i - \Delta, x^i + \Delta] \cong 2\Delta \cdot f(x^i)$$

$$L(\theta) = \prod_i \left(\quad \right) \cong \prod_i 2\Delta f(x^i) = \underbrace{(2\Delta)^n}_{\text{independent of } \theta} \prod_i f(x^i \mid \theta)$$

Deriving $L(\theta)$

by analogy with discrete case