STAT 391

2/6/2025

Lecture 10

Q1 on Tue 2/11

HW3 due Fri 2/7
at 11:59 pm

LV nonparametric
density

LIII supplement
uniform

HW IV t.b. posted

Parametric density est
- gradient ascent

Lecture Notes IV – Continuous distributions. Parametric density estimation.

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

January 2025

CDF and PDF. Sampling ✓

Examples of continuous distributions ✓

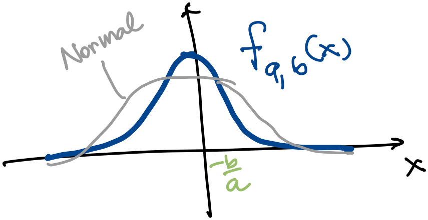
ML estimation for continuous distributions ✓

ML estimation by gradient ascent ←

when no analytic
solution exists

Reading: Ch.5, 6

Examples of continuous distributions



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

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

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

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

CDF

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

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

$b=0$
$a=1$

$$F = \frac{1}{1+e^{-x}}$$

$$f = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$ax+b = a\left(x - \frac{-b}{a}\right) = z$$

like $\frac{1}{\sigma^2}$

mean

Ex: symmetric $f(x) = f(-x)$

$\arg\max_x f(x) = 0$

ML estimation by gradient ascent

Logistic density

$\mathcal{D} = \{x^1, \dots x^n\} \subset (-\infty, \infty)$ data

$\mathcal{F} = \{ f(x \mid a, b),\ a > 0,\ b \}$ model class

params

1. Likelihood $L = \prod_{i=1}^{n} f(x^i \mid a, b)$

Want: $a^{ML}, b^{ML} = ?$

1.' $$l(a,b) \;=\; n \ln a - a \sum_i x_i - nb - 2 \sum_{i=1}^{n} \ln(1 + e^{-ax_i - b})$$

STAT

CALC

$$\frac{\partial l}{\partial a} \;=\; \frac{n}{a} - \sum_{i=1}^{n} x_i \frac{1 - e^{-ax_i - b}}{1 + e^{-ax_i - b}} \;=\; 0$$

$$\frac{\partial l}{\partial b} \;=\; -\sum_{i=1}^{n} \frac{1 - e^{-ax_i - b}}{1 + e^{-ax_i - b}} \;=\; 0$$

solve for a, b

NUMERICALLY

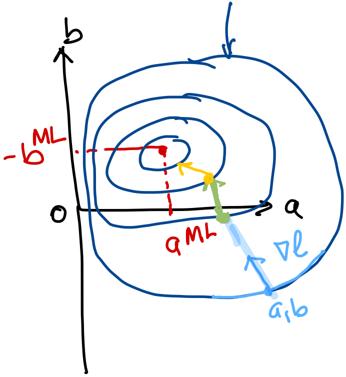
by Gradient Ascent

ML estimation by gradient ascent

level curves
$\ell(a,b) = \text{constant}$

$\ell(a,b) \dashrightarrow \frac{1}{n}\ell(a,b)$

$$l(a,b) = n \ln a - a \sum_i x_i - nb - 2\sum_{i=1}^{n} \ln(1+e^{-ax_i - b})$$



$$\frac{\partial l}{\partial a} = \frac{n}{a} - \sum_{i=1}^{n} x_i \frac{1-e^{-ax_i-b}}{1+e^{-ax_i-b}} = 0$$

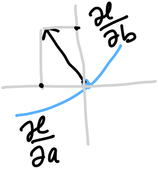
$$\frac{\partial l}{\partial b} = -\sum_{i=1}^{n} \frac{1-e^{-ax_i-b}}{1+e^{-ax_i-b}} = 0$$

for any $a,b \rightarrow \ell(a,b)$

initial $(\overset{0}{a},\overset{0}{b})$: $\nabla \ell(\overset{0}{a},\overset{0}{b}) = \begin{bmatrix} \frac{\partial \ell}{\partial a} \\ \frac{\partial \ell}{\partial b} \end{bmatrix} \neq 0$

(a^1, b^1) : $\nabla \ell(a^1, b^1) \neq 0$

(a^2, b^2) : $\nabla \ell(a^2, b^2) \neq 0$ ...



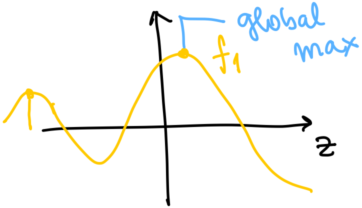
logistic density estimation

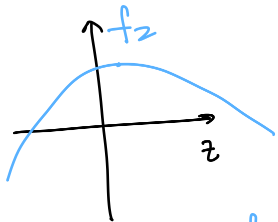
$$\nabla \ell(a,b) = \begin{bmatrix} \frac{\partial \ell}{\partial a} \\ \frac{\partial \ell}{\partial b} \end{bmatrix} \in \mathbb{R}^2$$

$$\nabla^2 \ell(a,b) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial a^2} & \frac{\partial^2 \ell}{\partial a \partial b} \\ \frac{\partial^2 \ell}{\partial a \partial b} & \frac{\partial^2 \ell}{\partial b^2} \end{bmatrix} \in \mathbb{R}^{2\times 2} \quad \text{symmetric}$$

- if $\underline{\nabla^2 \ell(a,b) \prec 0}$ for all a,b $\Rightarrow$ ℓ has at most 1 maximum

- $A \prec 0$ if $\lambda_1, \lambda_2 < 0$





$f_1' = 0$

$f_1'' < 0$

f_1 has 2 local maxima

f_2 has 1 local max $\equiv$ global max

$f_2' = 0$

$\boxed{f_2'' < 0 \text{ for all } z} \Rightarrow$ at most 1 maximum

Gradient Ascent Algorithm

1. Initialize with some a^0, b^0, $a^0 > 0$
2. For $t = 1, 2, \dots T$

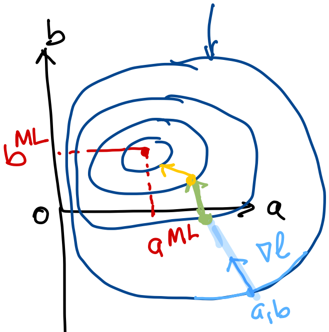
   compute $\nabla \ell(a^{t-1}, b^{t-1}) = \underbrace{g^t}_{\text{gradient}} \in \mathbb{R}^2$

   update $a^t \leftarrow a^{t-1} + \eta \cdot \frac{\partial \ell}{\partial a}(a^{t-1}, b^{t-1})$

   $b^t \leftarrow b^{t-1} + \eta \cdot \frac{\partial \ell}{\partial b}(a^{t-1}, b^{t-1})$

   if "$|g^t|$" $<$ tol STOP and output a^t, b^t

$$\begin{bmatrix} a^t \\ b^t \end{bmatrix} \leftarrow \begin{bmatrix} a^{t-1} \\ b^{t-1} \end{bmatrix} + \eta g^t$$



Details

- init $a^0, b^0 \leftarrow$ random (or best of knowledge)

  Alg alway converges to a^{ML}, b^{ML}

- T number of steps

  Safety against ∞ loops

  corectly run should exit with convergence

- CONVERGENCE TEST

  Rule 1 $\|g^t\| \leq$ tol ? NOT RELIABLE

  $$\left\| \begin{matrix} z_1 \\ z_2 \end{matrix} \right\| = \sqrt{z_1^2 + z_2^2}$$

  Rule 2 rel change $\leq$ tol YES

  $$0 < \left| \frac{\ell(a^t, b^t) - \ell(a^{t-1}, b^{t-1})}{\ell(a^{t-1}, b^{t-1})} \right| < \varepsilon$$

  use other diagnoses

  tol $> \sqrt{\varepsilon_{machine}}$

  $\varepsilon = 10^{-3, 4, 6}$

- choice of <u>step size</u> $\eta > 0$

  usually $\eta \ll 1$

i) η small

$\nabla \ell(a^1, b^1) \propto \nabla \ell(a^0, b^0)$

advance too slow!

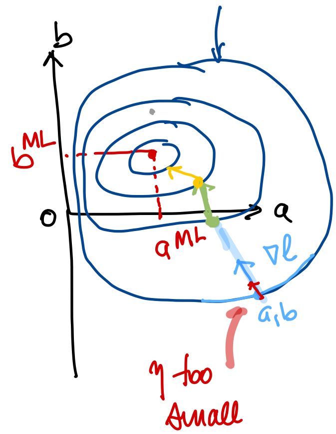
ii) η too large

instability

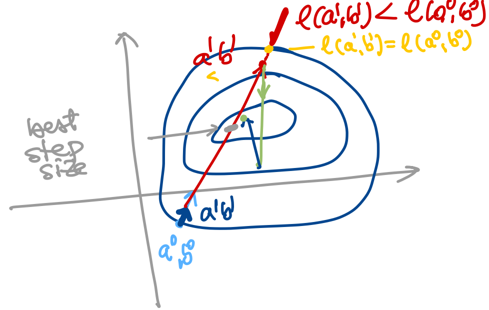
increase of ℓ too slow

OR divergence

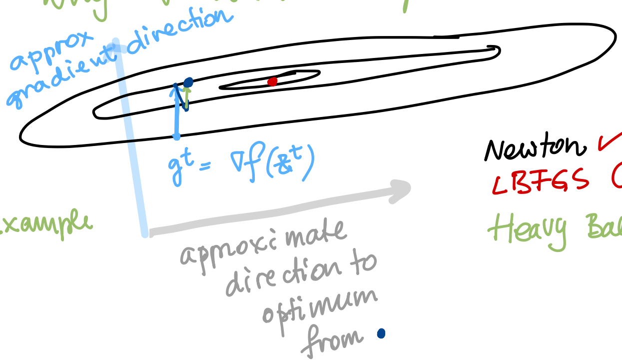
$\ell \nearrow \searrow$



η too large:



Why ∇ ascent not optimal



Example

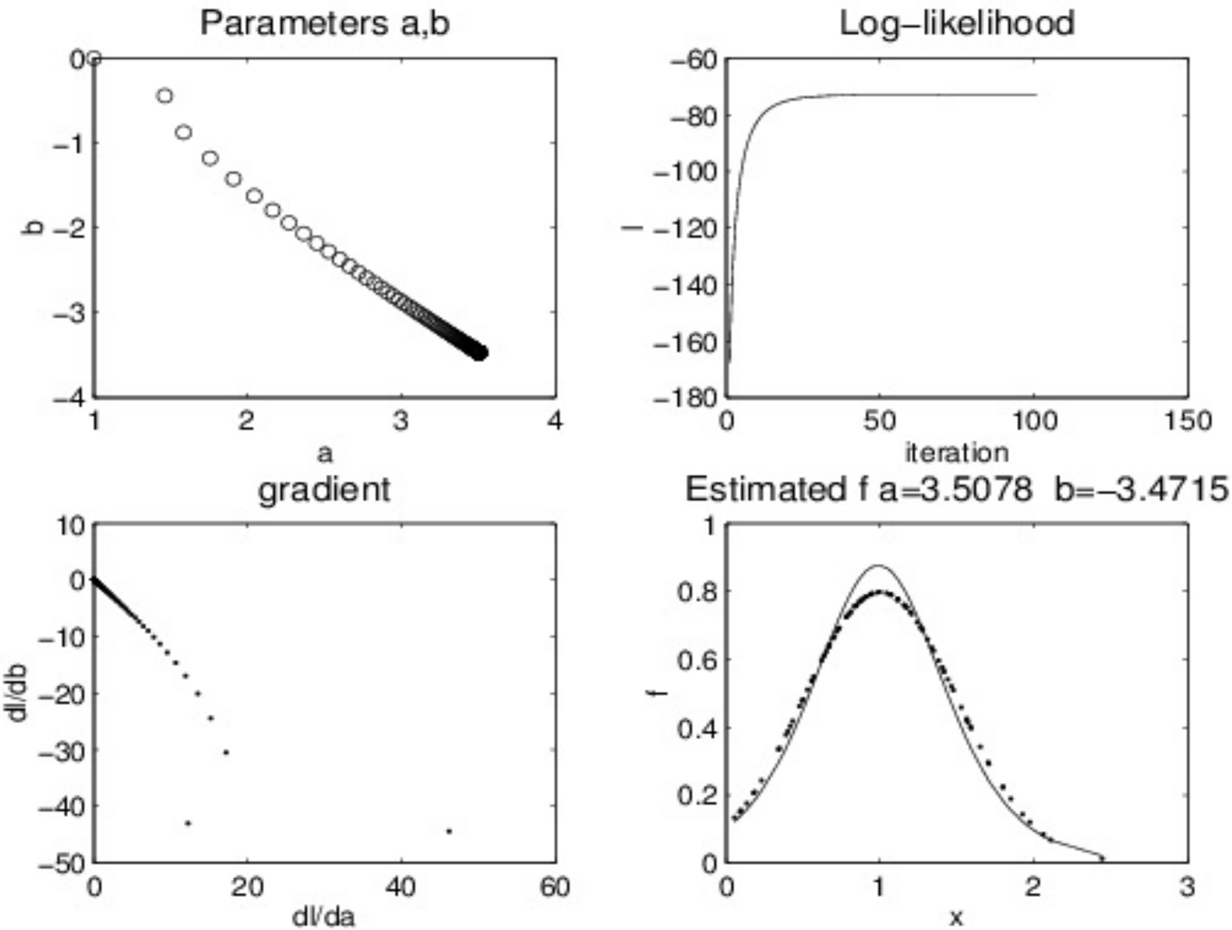
Newton ✓

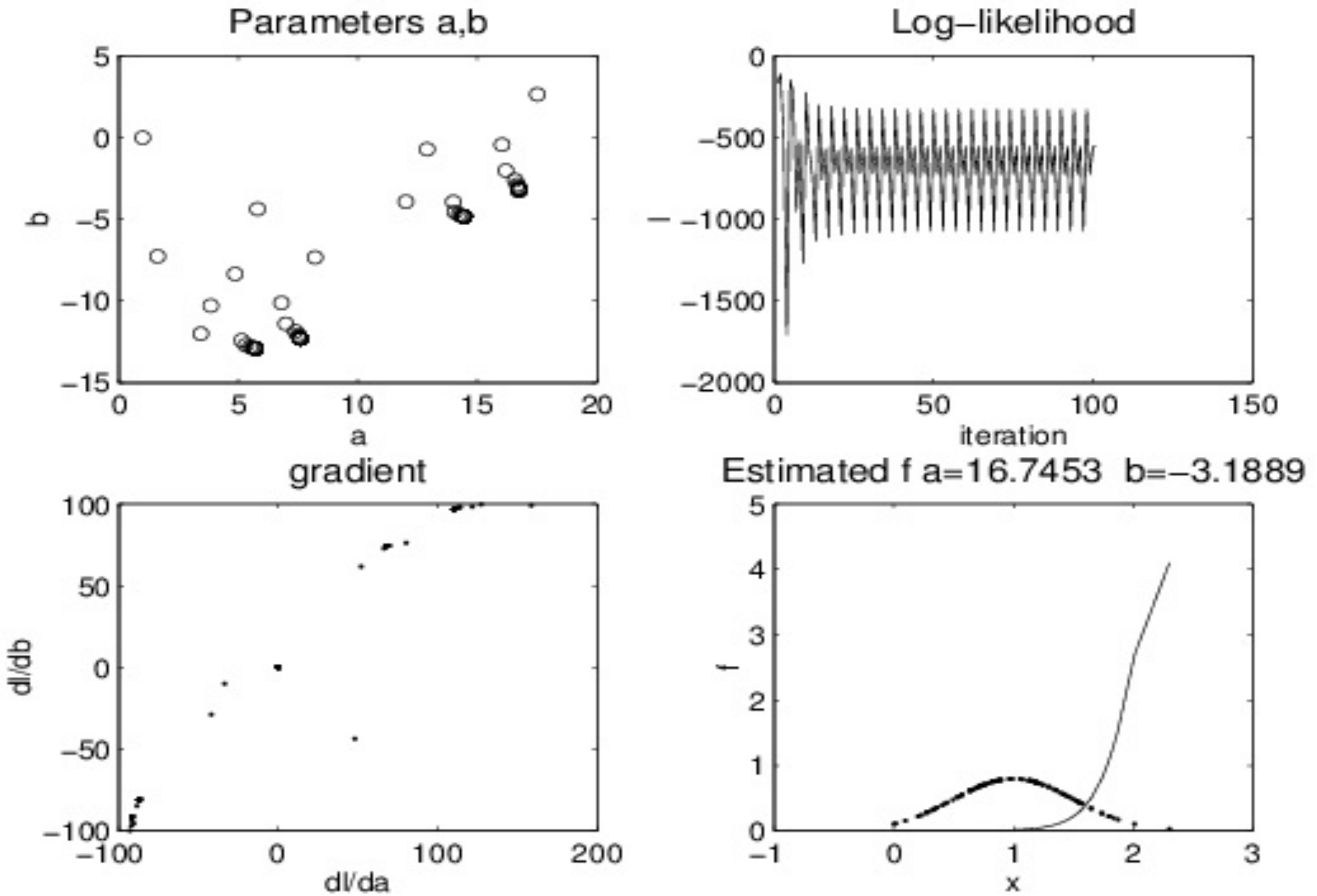
LBFGS (quasi-Newton) ✓

Heavy Ball ✓

LBFGS optimization routine

ML estimation for continuous distributions


Parameters a,b
b
a
0
−1
−2
−3
−4
1
2
3
4
Log−likelihood
l
iteration
−60
−80
−100
−120
−140
−160
−180
0
50
100
150
gradient
dl/db
dl/da
10
0
−10
−20
−30
−40
−50
0
20
40
60
Estimated f a=3.5078 b=−3.4715
f
x
1
0.8
0.6
0.4
0.2
0
0
1
2
3


Parameters a,b
b
a
Log−likelihood
iteration
gradient
dl/db
dl/da
Estimated f a=16.7453 b=−3.1889
f
x