STAT 391

01/30/25

Lecture 8

Happy Lunar New Year!

Q1 Thu 2/6

LIV Cont distributions

HW2 due

Sol 2 Tue 2/4

HW 3 TB posted

Small probs

Continuous S

Lecture Notes III: Discrete probability in practice – Small Probabilities

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

January, 2025

The problem with estimating small probabilities

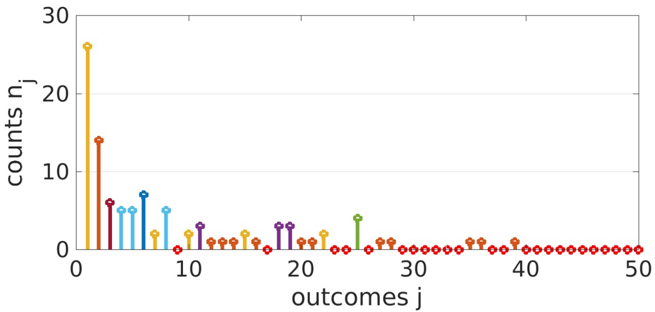
Definitions and setup

Additive methods (Laplace, Dirichlet, Bayesian, ELE)

Discounting (Ney-Essen)

Multiplicative smoothing: Estimating the next outcome (Witten-Bell, Good-Turing)

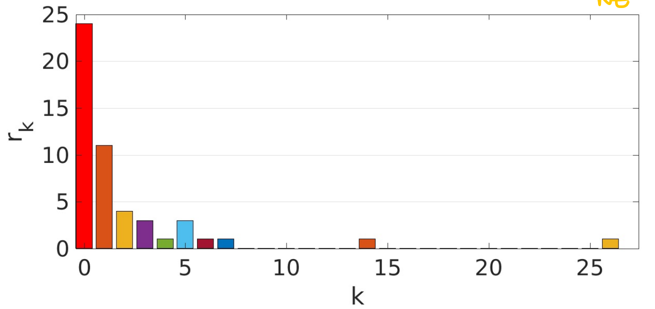
Back-off or shrinkage – mixing with simpler models



$R_k = \{ j$ that appear k times $\}$

$n_j = k$

$\theta_j^{ML} = \frac{k}{n} \to \tilde{\theta}_j$



NE $\underline{R_0} \cup \underline{R_1}$: $n'_j = 0$

R_k $k > 1$

WB R_0 special

R_k $k > 1$

GT R_k separate for all k

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** $\{n_j = \#j \text{ appears in } \mathcal{D},\ i = 1, \ldots n\}$. 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	$r_0 = 12 = \lvert\{\text{a,b,g,}\ldots\text{,y,z}\}\rvert$ $r_1 = 12 = \lvert\{\text{c,d,f,h,}\ldots\text{,u,x}\}\rvert$ $r_2 = 2 = \lvert\{\text{e,i}\}\rvert$ $r_3 = \ldots r_n = 0$
`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	$r_0 = 26 - 6 - 1 - 1 - 1 = 17$ $r_1 = 6 = \lvert\{\text{f,i,n,r,t,w}\}\rvert$ $r_2 = 1 = \lvert\{\text{s}\}\rvert$ $r_3 = 1 = \lvert\{\text{h}\}\rvert$ $r_4 = 1 = \lvert\{\text{o}\}\rvert$

- 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

Extra $\tilde{\theta}^{WB}|_{R_1} < \tilde{\theta}^{WB}|_{R_0}$?

- **the counts** $\{n_j = \#j$ 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

y^j

`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

Fingerprint r_k

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

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

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

$r_3 = \dots r_n = 0$

$r = m - r_0 = 14$

$m = 12 + 14 = 26$

Witten-Bell

$y^i = \begin{cases} 1 & \text{if } x^i \text{ "new"} \\ 0 & \text{otherwise} \end{cases}$

$\boxed{p_0 \text{"="} \Pr[y^i = 1] \text{"="} \frac{\#\text{new}}{n} = \frac{r}{n}}$

- no sense if $r \lesssim m$!
- some sense $r \ll m$

$j \in R_0 \Rightarrow \Pr[x^{n+1} = j] = \frac{p_0}{r_0}$

unobserved

$n_j > 0: \theta_j^{ML} = \frac{n_j}{n} \rightarrow \sum_j \theta_j^{ML} = 1$

$n_j = 0: \theta_j' = \frac{p_0}{r_0} \rightarrow \sum_{j \in R_0} \theta_j' = p_0$

$\rightarrow 1 + p_0 = Z$ normalization

n large $>> m >> r$

want $\Pr[x^{n+1}$ is new$] \approx \frac{r}{n}$

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_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	$r_0 = 12 = \lvert\{a,b,g,\ldots,y,z\}\rvert$ $r_1 = 12 = \lvert\{c,d,f,h,\ldots,u,x\}\rvert$ $r_2 = 2 = \lvert\{e,i\}\rvert$ $r_3 = \ldots r_n = 0$

$n_j > 0$: $\theta_j^{ML} = \dfrac{n_j}{n} \to \sum_j \theta_j^{ML} = 1 \longrightarrow \tilde{\theta}_j^{WB} = \dfrac{\theta_j^{ML}}{1+p_0}$ ← rescale Σ $\xrightarrow{}$ $\frac{1}{1+p_0}$ by $1+p_0$ $1+p_0$

$n_j = 0$: $\theta_j' = \dfrac{p_0}{r_0} \to \sum_{j \in R_0} \theta_j' = p_0 \longrightarrow \tilde{\theta}_j^{WB} = \dfrac{p_0}{r_0} \cdot \dfrac{1}{1+p_0}$ $\Sigma \rightarrow \dfrac{p_0}{1+p_0}$

TO CHECK: $\displaystyle\sum_{j=1}^{m} \tilde{\theta}_j^{WB} = 1$

- when ~~not~~ to use WB

  $\dfrac{n}{m} \approx 1$!!

Witten-Bell discounting – probability of a new value

- **Idea:**
  - Look at the sequence $(x_1, \ldots x_n)$ as a binary process: either we observe a value of X that was observed before, or we observe a new one.
  - Assume that of m possible values r were observed (and $m - r$ unobserved)
  - Then the probability of observing a new value is $p_0 = \frac{r}{n}$.
  - Hence, set the probability of all unseen values of X to p_0. The other probabiliy estimates are renormalized accordingly.

$$\theta_j^{WB} = \begin{cases} \frac{n_j}{n}\frac{1}{1+p_0} = & \frac{n_j}{n+r} & n_j > 0 \\ \frac{1}{m-r}\frac{p_0}{1+p_0} = & \frac{1}{m-r}\frac{r}{n+r} & n_j = 0 \end{cases} \tag{7}$$

Witten-Bell makes sense only when some n_j counts are zero. If all $n_j > 0$ then W-B smoothing has undefined results.

WB smoothing has no parameter to choose (GOOD!)

Comparison of the methods ← toy ex.

Numerical values to exemplify the results: $n = 1000$, $m = 1000$, $r = 100$

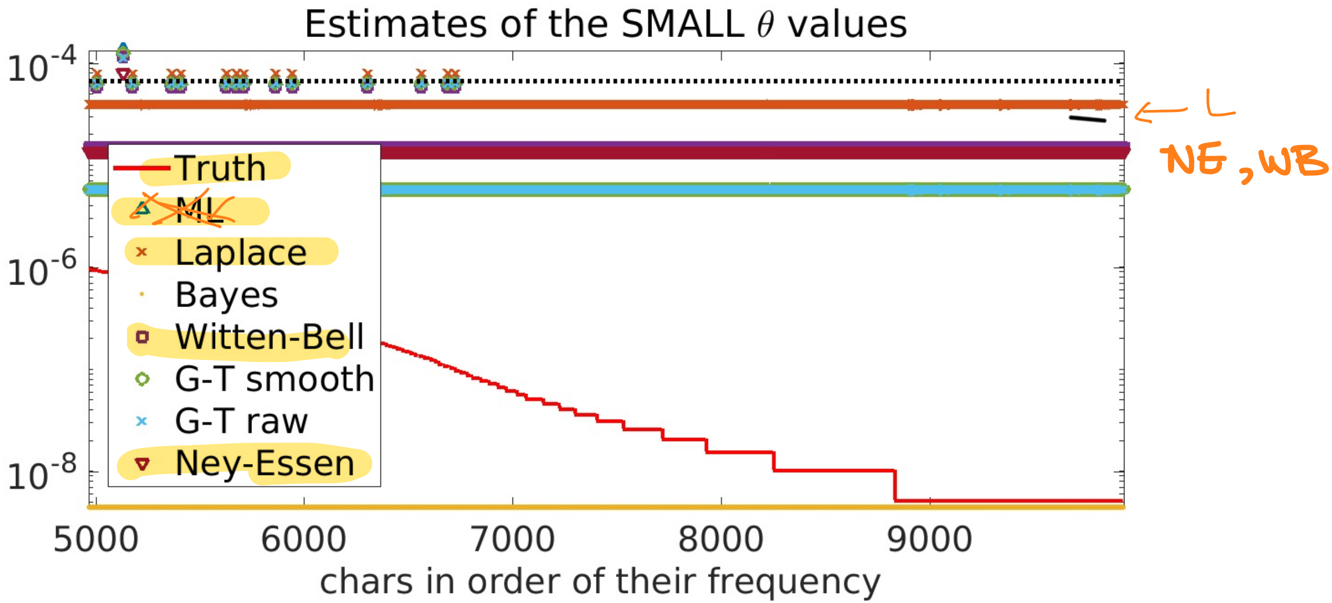
Count n_j	$n_j = 0$	$n_j = 1$	$n_j \gg 1$
θ_j^{ML}	0	$\frac{1}{n} = \frac{1}{1000}$	$\frac{n_j}{1000}$
$\theta_j^{Laplace}$	$\frac{1}{n+m} = \frac{1}{2000}$	$\frac{2}{n+m} = \frac{1}{1000}$	$\frac{n_j+1}{n+m} = \frac{n_j+1}{2000} = 0.5\,\theta_j^{ML}$!!
θ_j^{Bayes}, $n^0 = 1$, $n_j^0 = \frac{1}{m}$	$\frac{1}{m(n+1)} \approx \frac{1}{10^6}$	$\frac{1+1/m}{n+1} \approx \frac{1}{10^3}$	$\frac{n_j+1/m}{n+1} \approx \frac{n_j}{1000}$
θ_j^{NE}, $\delta = 1$ =tax	$\frac{r}{mn} = \frac{1}{10^4}$	$\frac{r}{mn} = \frac{1}{10^4}$	$\frac{n_j-1+r/m}{n} \approx \frac{n_j}{1000} \lesssim \theta_j^{ML}$
θ_j^{WB}	$\frac{1}{m-r}\,\frac{r}{n+r} = \frac{1}{9900}$ $\cdot\frac{1}{9}$	$\frac{1}{n+r} = \frac{1}{1100}$	$\frac{n_j}{n+r} = \frac{n_j}{1100}$

$1 + .1 = 1.1$

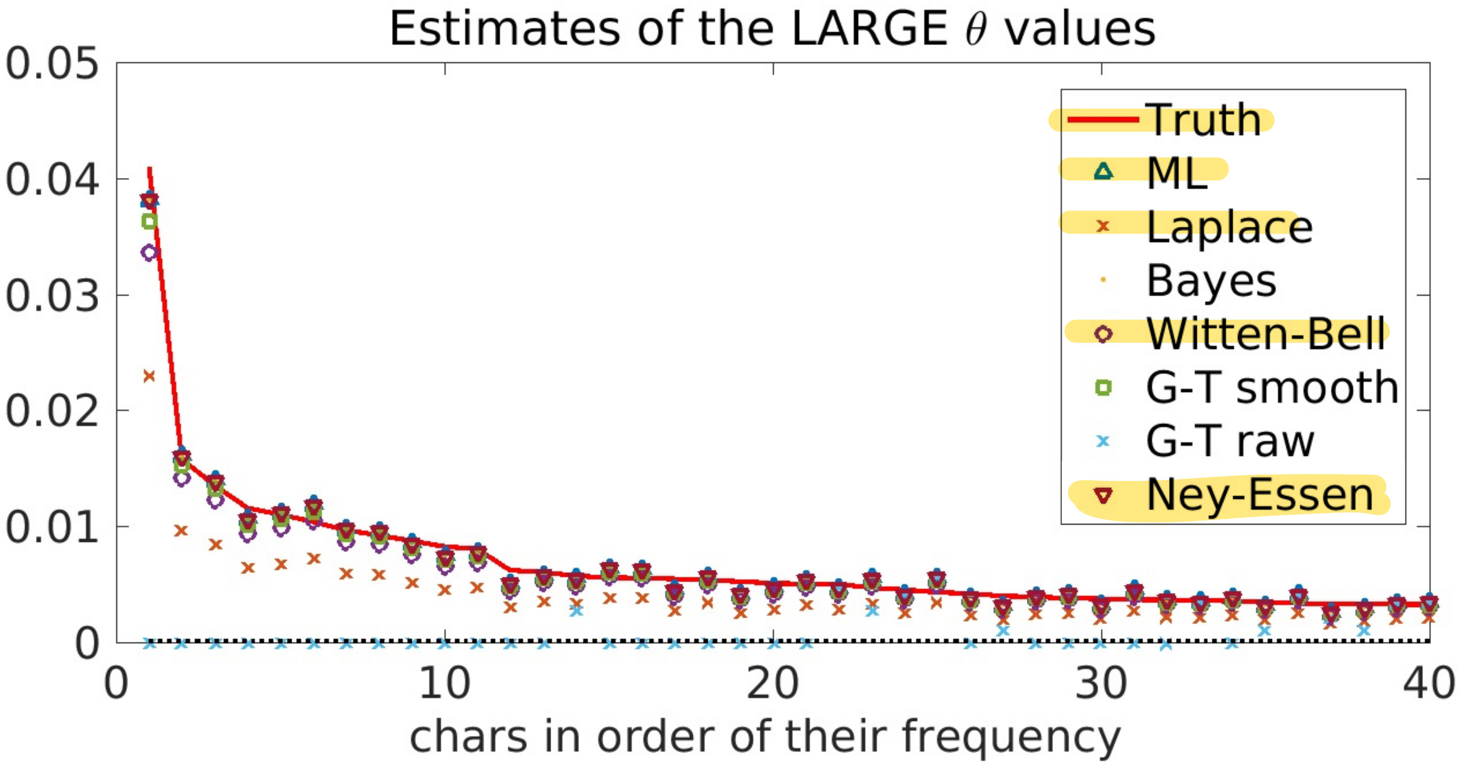
$\tilde{\theta}_j^{W} = \theta_j^{ML}/1.1$

Remarks

- Laplace shrinks ML estimates of large probabilities by factor of 2. Too much! (because large θ_j^{ML} are close to their true values)
- Bayes (with uninformative prior) affects large θ_j^{ML} much less than small ones. Good
- Ney-Essen smooths more when r is larger; any n_j is affected by less than δ.
- Ney-Essen estimates of θ^{NE} for counts of 0 and 1 are equal to a fraction of $\frac{r}{m}$ (this grows with n as r grows with n).
- In Witten-Bell, the large θ_j^{ML} are shrunk depending on r, but independently of m. Proportional, bad
- ...but, if we overestimate m grossly, the overestimation will only affect the θ_j^{WB} for the 0 counts, but none of the θ_j^{WB} for the values observed. (true for NE as well).



Estimates of the SMALL θ values
10^{-4}
10^{-6}
10^{-8}
5000
6000
7000
8000
9000
chars in order of their frequency
Truth
~~ML~~
Laplace
Bayes
Witten-Bell
G-T smooth
G-T raw
Ney-Essen
L
NE, WB



Estimates of the LARGE θ values
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30
40
chars in order of their frequency
Truth
ML
Laplace
Bayes
Witten-Bell
G-T smooth
G-T raw
Ney-Essen

sample spaces

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 Refresher

Examples of continuous distributions

ML estimation for continuous distributions

ML estimation by gradient ascent

Reading: Ch.5, 6

CDF and PDF refresher

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

$F(b) = \Pr[(-\infty, b]]$

$F(b) - F(a) = \Pr[(a, b]]$

Cumulative distribution function (CDF) ← Sampling

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

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)

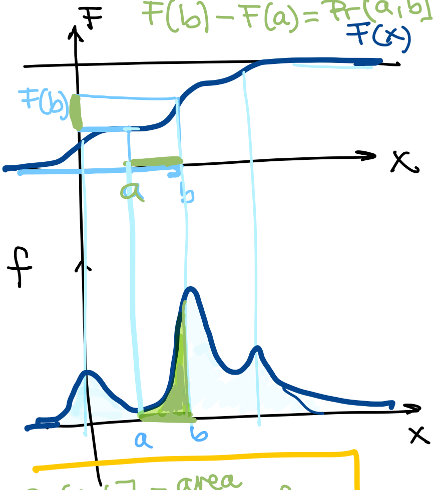
$$f \;=\; \boxed{\frac{dF}{dx}} \qquad (2)$$

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

Newton

normalization condition

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



$\Pr(a, b] =$ area under $f(x)$

$\Pr[S] = 1 \Rightarrow \int_{-\infty}^{\infty} f(x)\,dx = 1$

$\boxed{f(x) \ge 0}$

CDF and PDF refresher

Cumulative distribution function (CDF)

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

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

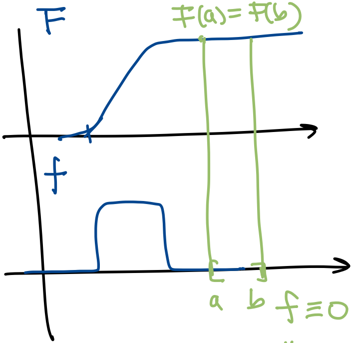
Probability density [function] (PDF)

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

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

normalization condition

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



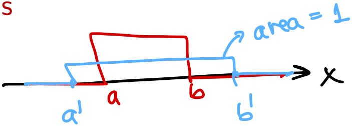
$f \equiv 0$ on $[a,b]$

$\Downarrow$

$P[a,b]=0$

Examples of continuous distributions

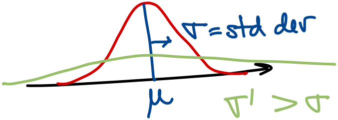
Model classes



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

$\rightarrow$ a, b = parameters $\in (-\infty, \infty)$
$a < b$

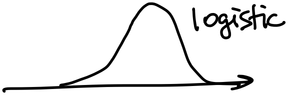
$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \tag{6}$$



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

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

$\rightarrow$ $\mu, \sigma^2 > 0$



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

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

$\rightarrow$ a, b
$a > 0$

exponential

$f(x; \lambda) = \lambda e^{-\lambda x}$ $\quad x \geq 0 \quad S = [0, \infty)$

$\rightarrow$ $\lambda > 0$

ML estimation for continuous distributions

STATISTICS

Given $x^{1,2,\dots n} \in (-\infty, \infty) = S \sim$ iid from P

What is ~~P~~?

$f(x) = ?$

Density estimation