- We want to represent counts
  Poisson R.V.'s capture the idea of uniformity
  - The count in an interval
    - depends on length of interval
    - in fact, $\propto$ length
  - Events are independent.

(Idea works in multiple dimensions).

If X is poisson, intensity λ

- $E(X) = \lambda$
- $Var(X) = \lambda$

Notice λ <u>has units</u> (eg $\frac{\#}{s}$, etc

and <u>depends on scale of interval</u>.

PMF :

$$P\left(X = n \,\middle|\, \begin{matrix}\text{unit interval,} \\ \lambda\end{matrix}\right) = \frac{e^{-\lambda} \cdot \lambda^{n}}{n!}$$

straightforward series manipulation gives
- this is a PMF
- expectation
- variance.

Now assume whatever we're watching
is · Poisson

- types are independent

[These assumptions are a stretch; words aren't like this; neither are animals or objects; but they're simple + generic]

Notice :

1) If I observe a Poisson RV for an interval longer than 1

$$P(X = n \mid \text{interval length } t, \lambda) = e^{-\lambda t}\frac{(\lambda t)^x}{x!}$$

2) If I observe Poisson RV for N unit intervals, seeing count n_i in the i'th, Max likelihood est of λ is

$$\lambda^* = \frac{1}{N}\sum_i n_i$$

3) If I observe poisson RV for N ~~unit~~ intervals, i'th has length L_i, and see n_i in i'th, Max likelihood gives

$$\lambda^* = \frac{1}{N}\sum_i \left(\frac{n_i}{L_i}\right)$$

Now assume we see $\left\{\begin{array}{l}\text{words}\\ \text{animals}\\ \text{objects}\end{array}\right\}$ for an interval (which could be time or space). We *choose* a scale so this interval is $[-1, 0]$.

Each word *type* has intensity

$$\lambda_w \qquad \lambda \leftarrow \text{intensity}, \quad w \leftarrow \text{type}$$

which is *unknown*

Natural to try and draw conclusions from word type counts.

$$n_x = \left[\begin{array}{l}\text{number of word types}\\ \text{that appear } x \text{ times}\end{array}\right]$$

Natural because Max likelihood on individual words is no help.

~~→ ML~~

- Also, assume future is like the past.

  i.e. if a word has λ_w in $[-1, 0]$
  it has this λ later.

- E + T phrase this as.
  "conditionally binomial"

- ML est. of λ for words we <u>haven't</u> seen is 0
- But n_1 large compared to n_2, etc suggests there are word types where
  - we haven't seen them
  - $\lambda_w > 0$
- So we should be looking at $G(\lambda)$

$$G(\lambda) = P(\lambda_w \leq \lambda)$$

- clearly, a discrete distribution
- represents CDF (cumulative dist. function)

$$dG(\lambda) \equiv p(\lambda_w = \lambda)\, d\lambda$$

δ functions or atoms

Now

$$P(a \overset{\text{given}}{\uparrow} \text{ word type has count } x \mid \lambda_w)$$

$$= e^{-\lambda_w} \frac{\lambda_w^x}{x!}.$$

$$P(a \overset{\text{given}}{\uparrow} \text{ word type has count } x)$$

$$= \int e^{-\lambda} \frac{\lambda^x}{x!} \, dG(\lambda)$$

<u>or</u> $\int e^{-\lambda} \frac{\lambda^x}{x!} \, p(\lambda) \, d\lambda$

$$E[\# \text{ of word types w/ count } x]$$

$$= \sum_{i \in \text{word types}} P(\text{word type } i \text{ has count } x)$$

$$= C \int e^{-\lambda} \frac{\lambda^x}{x!} \, dG(\lambda) = \eta_x$$

$\uparrow$ total # of word types, <u>unknown</u>!

- Another way to look at this is that

$$d\Gamma(\lambda) = C\,dG(\lambda) \quad \text{is a } \underline{\text{measure}}$$

(like a PDF, +ve, but <u>doesn't</u> $\int$ to 1)

- Notice that C could be hard to get, because we could have support for $G(\lambda)$ at, say,

$$\lambda = 10^{-12}$$

$$\Updownarrow$$

- there is a word we see about once in 10^{12} intervals.
- affects C, but not a significant effect on observations.

Now

- η_x is the observed value of an RV
- call it r_x
- we have
  - $E(r_x) = \eta_x$
- reasonable approx ?
  - $Var(r_x) = \eta_x$

Sum of Poisson

$$r_x \sim N(\eta_x, \sqrt{\eta_x})$$

sum of random variables

- This will come in useful.

Now consider

$$\cancel{E}\ \Delta(t) = E\left[\begin{array}{l}\#\text{ of types seen in } [0;t], \\ \text{but } \underline{\text{not}} \text{ in } [-1;0]\end{array}\right]$$

then

$$\Delta(t) = C\int_0^\infty \underbrace{\left[e^{-\lambda}\frac{\lambda^0}{0!}\right]}_{\text{seen 0 times in } -1,0} \underbrace{\left[1 - e^{-\lambda t}\right]}_{\underline{\text{not}}\text{ seen 0 times in } 0,t.} dG(\lambda)$$

Notice you can derive expressions for

$$E\left[\begin{array}{l}\#\text{ seen } a \text{ times in } [-1,0] \text{ and} \\ \qquad b \text{ times in } [0,t]\end{array}\right]$$

in the straightforward way.

- Notice also that you <u>dont</u> need C to evaluate this, just $C\,dG(\lambda)$.
- Q: <u>estimate $\Delta(t)$ given n_x</u>

- <u>Notice</u>

$$1-e^{-\lambda t} = \lambda t - \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^3}{3!} - \frac{(\lambda t)^4}{4!} \ldots$$

<u>So:</u>

$$\Delta(t) = \eta_1 t - \eta_2 t^2 + \eta_3 t^3 - \eta_4 t^4 \ldots$$

(assuming convergence, etc).

<u>Natural estimator:</u>

assume $\eta_1 = n_1$, $\eta_2 = n_2$, etc.
substitute

Paper gives word counts, word type counts

884647 words,
14376 only once
31534 word types

Q: if we saw another 884647 words, how many new words would there be?

A: $t = \frac{\text{\# of words in new}}{\text{\# in ref}}$

$= 1$

$\Delta(t) = 11430 = \Delta(1)$

$\mathrm{Var}(\Delta(1))$ by assuming the n_x are indep, poisson

$$\mathrm{Var}\{\Delta(1)\} \approx \sum_{i=1}^{\infty} n_i t^i = 31534$$

std = 178

(skip Fisher model)

Notice that, for $t > 1$

$$n_1 t - n_2 t^2 + n_3 t^3 - \cdots$$

is a series that oscillates savagely.

You could interpret this several ways

- it doesn't converge. — panic
- the oscillations "cancel", and we need some way to accelerate this cancellation

quite plausible, as $n_x \to 0$ as $x \to \infty$.

this gives §4 — Euler's transform.

Now skip to §7.

Recall that $\Delta(t)$ may be hard to estimate for large t (because there may be low frequency words).

Instead, try for a lower bound on $\Delta(t)$ — call this bound $b(t)$

Problem becomes:

$$b(t) = \inf_{C\,dG(\lambda)} \left[\int_0^\infty e^{-\lambda} \left[1 - e^{-\lambda t} \right] \left[C\,dG(\lambda) \right] \right]$$

Subject to

$$\eta_x = \mathcal{E} \int_0^\infty \left[e^{-\lambda} \frac{\lambda^x}{x!} \right] \left[C\,dG(\lambda) \right]$$

This would be an LP in $C\,dG(\lambda)$, IF we knew η_x.

Strategy 1:

- assume $\eta_x = n_x$.
- discretize η_x
- $\rightarrow$ then we have an honest LP and can solve.

- This works for $E+T$, but failed on object data (infeasible - ?!?)

Strategy 2:

- assume $\eta_x \leq n_x + \gamma\sqrt{n_x}$

$$\eta_x \geq n_x - \gamma\sqrt{n_x}$$

(i.e. γ stds away from mean)

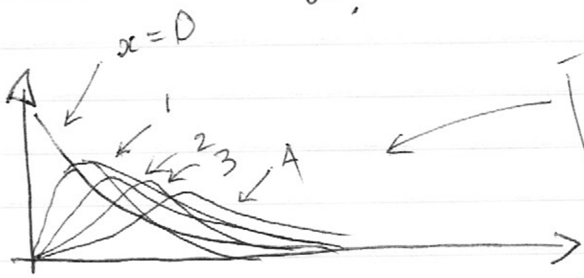
- This might slacken the bound, but is probably better practice

(15)

But S_2 isn't all that reliable either
(on objects, get feasibility ~~ss~~ only if you apply big γ AND use only n_x for $x \in [1..5]$ which is worrying.)

<u>What is going on here?</u>

Consider $\frac{e^{-\lambda} \lambda^x}{x!}$ <u>as a function of λ</u>



[Smooth bumps with wide support

This means we cant have

$$\eta_1 = 100; \; \eta_2 = 0; \; \eta_3 = 100; \; \text{etc.}$$

because these functions overlap so strongly — $\int \frac{e^{-\lambda} \lambda^1}{1!} C\, dG(\lambda)$ is similar to $\int \frac{e^{-\lambda} \lambda^2}{2!} C\, dG(\lambda)$

Alternative view:

$$\text{feasible } CdG(\lambda) \xrightarrow[\text{(integrals)}]{\text{linear map}} \eta_x\text{'s}$$

↑ ∞-dim; Convex;

finite dim

Since $CdG(\lambda) \geq 0$, this isn't the whole space

-ve η_x's can't occur.

~~This picture strongly implies that~~

use this picture with plots;

there are vectors of η_x

that are (a) non-negative

(b) infeasible

and if you use $\eta_x = \eta_x$,

this gets infeasibility

Also explains why large r is required,
AND why large x creates problems
(the n_x estimates are poor).

What to do?
we actually know quite a lot about r_x,
which is what we should be working with
- approximately Gaussian
- var approx = mean.

Strategy 3

assume $r_x \sim N(n_x, \sqrt{n_x})$

and $\eta_x = r_x$.

Now we must (a) estimate the counts and (b) estimate $G(t)$

$$\inf_{C\,dG(\lambda)} \int e^{-\lambda}\left[1-e^{-\lambda t}\right]\left[C\,dG(\lambda)\right] + \mu \sum_{i=1}^{k} S_i^2$$

weight

$$S_i^2 = \frac{(r_i - n_i)^2}{2\, n_i^2}$$ ← Standard error of ith count.

$$r_i = \int \frac{e^{-\lambda}\lambda^i}{i!}\, C\, dG(\lambda)$$ ← Count estimates

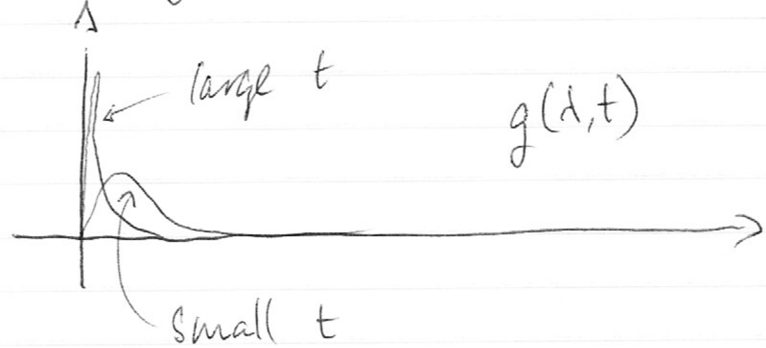
Q: how to choose μ?

A: cross validation

Now a QP, but it's convex, so no worries

Notice also that

$$g(\lambda, t) = e^{-\lambda}(1 - e^{-\lambda t}) \quad \text{is important}$$



- This (basically) looks for weight in small λ's (of $CdG(\lambda)$) which makes sense.

<u>Issue</u>: · lower bounds are helpful, but we want more (estimates, etc).

- <u>Options</u>
    - work w/ continuous $CdG(\lambda)$ models
    - make models explicitly discrete.

- Notice one attractive feature of this ~~proble~~ formulation

  - η_x, $b(t)$, $\Delta(t)$ are quite insensitive to "small" changes in $Cd(G(t))$
  - because
$$\frac{e^{-\lambda}\lambda^x}{x!}$$ is pretty smooth.

- This means that big shifts in where the weight is in a prob model are required to change counts

- In turn, we could use quite rough discretizations (but finer near 0).