Note: ongoing workshop : "Bayesian Inference in
this week   Subatomic Physics"
                                in Gothenburg, Sweden

9/18/19

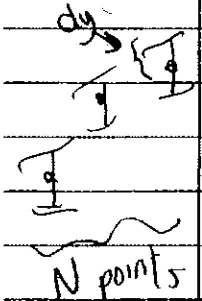
Physics 8805 : Learning from Data : Lecture 8

Notebooks for today :

- Assignment_2.ipynb
- mini-project-I_toy_model_of_EFT.ipynb
- Metropolis_Poisson_example_with_results.ipynb
- MCMC-random-walk-and-sampling.ipynb
- Why_bayes_is_better_I.ipynb and bayes_billiard.ipynb

Comments about assignments

• Assignment 1 : I got notebooks from everyone registered (and some from those sitting in, which are welcome). Email worked fine but still working on something better. Feedback will be brief. Grade is check, plus, minus. For this assignment plus was for extra answers above and beyond the minimum.



always verify !

• Assignment 2 : (ask for help on coding as needed)

- For A.1 and B.1, just add any questions you have.
- For A.2, the data set depends on the number of points N and the error bar for each dy. Determine empirically how the fluctuations change when you (say) reduce dy by a factor of five, or quadruple N. [Make hypotheses for the answers before you check!] Suggestion: add a loop to make repeated runs ⇒ see example. Important to do repeated trials ⇒ fluctuations!
- A.3-A.5 are straight questions.
- B.2 ⇒ see sample plots. Ask me if you get stuck on proof.
- B.3 ⇒ see discussion in Sivia Chapter 3 (Carmen under 2. Bayesian parameter estimation) for clues on the analysis.

Comments on Mini-project I. ⇒ now available on Carmen as well as Github.

• Overview slides : Bayesian Statistics for EFTs (MSU statistics conference 2018_Furnstahl.pdf

- Do up to slide 17. Note BUQEYE github page.
- Look at mini-project-I_toy_model_of_EFT.ipynb and arXiv : 1511.03618. You'll find a lot of now-familiar Bayesian concepts discussed.

9/18/19

Recap of Poisson MCMC example (Metropolis_Poisson_example_with_results.ipynb

Recall Metropolis algorithm for this example and the next:

```
• start with $\vec{\theta}_0$
• given $\vec{\theta}_i$, propose $\vec{\phi}$ from $q(\vec{\phi}|\vec{\theta}_i)$
• calculate $r = \dfrac{p(\vec{\phi}|D,I)}{p(\vec{\theta}_i|D,I)}\left[\dfrac{q(\vec{\theta}_i|\vec{\phi})}{q(\vec{\phi}|\vec{\theta}_i}\right]$ ← correction factor for r if q pdf is not symmetric: $q(\vec{\phi}|\vec{\theta}_i) \neq q(\vec{\theta}_i|\vec{\phi})$
  compare posteriors →
• decide whether to keep $\vec{\phi}$:
    if $r \geq 1$, set $\vec{\theta}_{i+1} = \vec{\phi}$   (accept)
    else draw $u \sim$ uniform$(0,1)$
        if $u \leq r$, $\vec{\theta}_{i+1} = \vec{\phi}$   (accept)
        else $\vec{\theta}_{i+1} = \vec{\theta}_i$   (add another copy of $\theta_i$ to the chain)
• repeat until "converged"
```

Key questions: When are you converged?
How many "warm-up" or "burn-in" steps to skip?

Poisson take-aways:

① It works! Sampled histogram agrees with (scaled) exact Poisson pdf. But not normalized! Compare 1000 to 100,000

This is success →

② Burn in time is (apparently) seen from trace.
Moral: always check traces!

③ Trace also shows that the space is explored.

④ What if the $\vec{\theta}_{i+1} = \vec{\theta}_i$ step is not implemented? (So the chain is only incremented if the step is accepted.) Not obviously intuitive!
See Metropolis_Poisson_example_with_results_no_repeats.ipynb
⇒ compare 100,000 ⇒ invalidates Markov chain ⇒ wrong stationary distribution

⑤ The spread of the means decreases as $1/\sqrt{N_{steps}}$, as expected.

9/18/19

- Before considering another example, some summary points from arXiv: 1710.06068 "Data Analysis Recipes: Using Markov Chain Monte Carlo" by David Hogg and Daniel Foreman Mackey.
  - Both computational astrophysicists (or cosmologists or astronomers)
  - DFM wrote emcee.
  - in physics scenarios → Highly experienced, highly opinionated, not statisticians but interact with them.

- Selected comments:
  - MCMC is good for <u>sampling</u>, not optimizing. If you want to find modes, use an optimizer.

  - For MCMC, you only have to calculate <u>ratios of pdfs</u>
    ⇒ don't need analytic normalized pdf
    ⇒ great for sampling posterior pdfs

$$p(\theta|D) = \frac{1}{Z} p(D|\theta)\, p(\theta)$$

↖ really difficult, because need <u>global</u> information

  - Extremely easy to implement, without requiring derivatives or integrals of the function (but see later discussion of HMC)

* - Success! A histogram of the samples looks like the pdf

$$E_{p(\theta)}[\theta] \approx \frac{1}{N}\sum_{k=1}^{N} \theta_k$$

$$E_{p(\theta)}[g(\theta)] \approx \frac{1}{N}\sum_{k=1}^{N} \theta_k \rightarrow \frac{\int d\theta\, g(\theta) f(\theta)}{\int d\theta\, f(\theta)}$$

↙ $f(\theta) = Z p(\theta)$ unnormalized

  - Works for expectation values even though we don't know Z

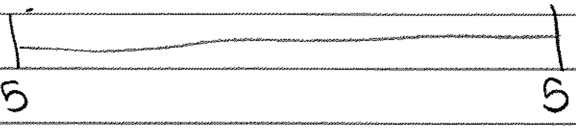
* - Nuisance parameters are very easy to marginize: just sum that column
  - Autocorrelation is important to monitor, Tune to minimize it. More later.
  - How do you know when to stop? Heuristics and diagnostics to come!
  - Practical advice for initialization and burn-in.

9/18/19

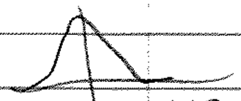
Look at MCMC-random-walk-and-sampling.ipynb

- Let's do some of the first part together

Part 1: Random walk in the [-5, 5] region.



proposal from



normal with proposal_width

Algorith: always accept unless across boundary.

Class: Map this problem onto Metropolis algorithm!

- What is $q(\phi | \theta_i)$? → $\phi \sim \mathcal{N}(\theta_i, (\text{proposal\_width})^2)$

- Note: rvs means random variates; 1 by default

use shift-tab-tab to check whether call of norm function takes σ or σ^2.

* - What is $p(\theta_i | D, I)$?

⇒ must be constant except for borders ⇒ $\sim U(-5,5)$!

- Check the answer. Change np.random.seed(10) to np.random.seed().

do Questions using Copy1 and Copy2 notebooks

- Note fluctuations
- Try 200 then 20000 samples.
- Try changing to uniform proposal

- Try not adding rejected position.
  - move samples.append(current_position) under if statement
  - doesn't fill full space (try different runs — trouble with edges)

- Decrease width to 0.2 2000 steps

⇒ samples too correlated.

- Look at definition and correlation time for 0.2 and 2
- Trend in autocorrelation plot: 1 at lag h=0, how fast to fluctuate around 0?

9/18/19

Why Bayes is Better I

- These examples were developed by Christian Forssen for the 2019 TALENT course at York, UK, (See nucleartalent.github.io/Bayes2019

- We'll use the notebooks:
  - (A) why_bayes_is_better_I.ipynb
  - (B) bayes_billiard.ipynb
  - (C) parameter_estimation_fitting_straight_line_I.ipynb

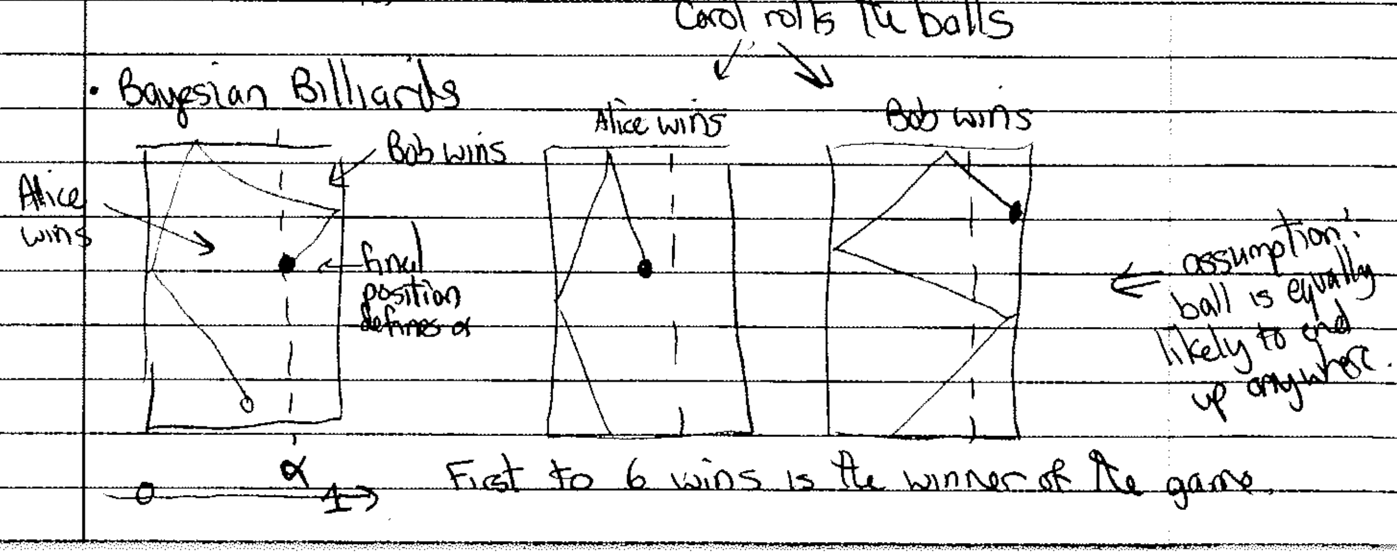
- Start with (A)
  - Review: Advantages of the Bayesian approach
    - Step through 1-6 as a review.
    - Occam's Razor (picture is from Wikipedia of a leprechauns - could always be added but are not necessary)

      ⇒ more about this later (evidence)

- Nuisance parameters I: A Bayesian Billiard game ⇒ go to (B)
  - Nuisance parameters are parameters we introduce to characterize a situation but which we don't care about or know in detail: "auxiliary variables"

    ⇒ integrate over them: "marginalize"
  - The procedure here is quite generic, so it is worth looking at in detail.

- Bayesian Billiards



First to 6 wins is the winner of the game.

Do not rely on the discussion here ⇒ go through the notebook where there are more details

9/18/19

Capsule summary: Carol knows α but Alice and Bob don't.

- Alice and Bob are betting on various outcomes.
- After 8 throws, the score is Alice 5, Bob 3.

They now are going to bet on Bob pulling out overall win. Alice is most likely to win, she only needs 1 winning roll out of 3, and there is already some indication she is favored. What odds should Bob accept?

[Note: This is obviously not a physics problem but you can map it onto many possible physics experimental or theoretical situations. In this case α could be a normalization in an experiment (not between 0 and 1, but α_{min} and α_{max}) or a model parameter in a theory that we don't know.]

Naive frequentist approach

- Here we think about the best estimate for α, call it $\hat{\alpha}$. If B is the statement "Bob wins", then what is $p(B)$?
  - Bob winning a given roll is $1-\hat{\alpha}$, and he must win 3 in a row:

$$\Rightarrow p(B) = (1-\hat{\alpha})^3$$

Find the maximum likelihood estimate for $\hat{\alpha}$. The likelihood

$$\mathcal{L}(\alpha) = \binom{8}{5}\alpha^5(1-\alpha)^3$$ for the result Alice 5, Bob 3

$$\Rightarrow \left.\frac{\partial\mathcal{L}}{\partial\alpha}\right|_{\hat{\alpha}} = 0 \Rightarrow 5\hat{\alpha}^4(1-\hat{\alpha})^3 - 3\hat{\alpha}^5(1-\hat{\alpha})^2 = 0$$

$$\Rightarrow 5(1-\hat{\alpha}) - 3\hat{\alpha} = 0$$

$$\Rightarrow \boxed{\hat{\alpha} = \frac{5}{8}}$$

This yields $p(B) \doteq .053$ or 18 to 1 odds.

9/18/19

Bayesian approach ← class fill in details

- Goal: find $p(B|D,I)$ where $D=\{n_A=5,\ n_B=3\}$ and I are the details of the game
- Plan: introduce α as a "nuisance parameter". If we know α, the calculation is straightforward. If we only know it with some probability, then marginalize
- Different equivalent paths to the same result

↙ drop I's mostly

a. $p(B|D,I)=\int_0^1 d\alpha\, p(B,\alpha|D)=\int_0^1 d\alpha\, p(B|\alpha,D)\, p(\alpha|D)$

b. $p(B,\alpha|D,I)\Rightarrow$ marginalize over α

c. $p(B|\alpha,D,I)\Rightarrow$ marginalize, weighting by $p(\alpha|D)$

↑ assume we know α

- What to do about $p(\alpha|D)$? Bayes theorem to convert to quantities we know.

binomial prop. ↙ ↙ uniform by assumption (no bias toward any value from 0 to 1)

class! Which is in numerator, $p(\alpha)$ or $p(D)$?

$$p(\alpha|D)=\frac{p(D|\alpha)\,p(\alpha)}{p(D)}$$ ← we need this here

$$\Rightarrow p(D|I)=\int_0^1 d\alpha\, p(D|\alpha,I)\,p(\alpha,I)$$

$$\Rightarrow p(B|D)=\frac{\int_0^1 d\alpha\, p(B|\alpha,D)\,p(D,\alpha)\,p(\alpha)}{\int_0^1 p(D|\alpha)\,p(\alpha)}$$

Does this integral make sense to you?

$$=\frac{\int_0^1 d\alpha\,(1-\alpha)^3\cancel{\binom{8}{5}}\alpha^5(1-\alpha)^3\cdot 1}{\int_0^1\cancel{\binom{8}{5}}\alpha^5(1-\alpha)^3\cdot 1}$$

$p(B|\alpha,D)=(1-\alpha)^3$

$p(D|\alpha)=\binom{8}{5}\alpha^5(1-\alpha)^3$ from binomial probability.

Don't need $\binom{8}{5}$!

$$p(\alpha|I)=\begin{cases}1 & \text{if } 0\le\alpha\le 1\\ 0 & \text{otherwise}\end{cases}$$

9/18/19

$$\Rightarrow \quad p(B|D,I) = \frac{\int_0^1 (1-\alpha)^6 \alpha^5 \, d\alpha}{\int_0^1 (1-\alpha)^3 \alpha^5 \, d\alpha} \doteq 0.091 \text{ or } 10 \text{ to } 1 \text{ odds.}$$

[We can use the beta function $\beta(n,m) = \int_0^1 (1-t)^{n-1} t^{m-1} \, dt$]

- So very different results!

- How do we check? In many cases, we can do a Monte Carlo simulation (at least to validate test cases).

  - See the notebook for implementation of this simulation
  - Result ⇒ Bayes win! Using an MLE estimate incorrectly predicts the likelihood of the result B.

Note the discussion points:

- Introducing α is straightforward in Bayesian approach, with all assumptions clear.
- In general one introduces many such variables, which is how we end up with posterior integrals we need to sample to do marginalization.

- The problem with the "naive frequentist" approach is "naive", not "frequentist." But it is hard to see how to proceed to take into account the need to sum over possibilities for α, while it is natural for Bayes.