Question 1

(a) The density is $f(x) = \frac{1}{2\theta}$ for $-\theta \le x \le \theta$, $f(x) = 0$ otherwise.

i.e. $f(x) = \frac{1}{2\theta} 1_{-\theta \le x \le \theta}$

The likelihood is

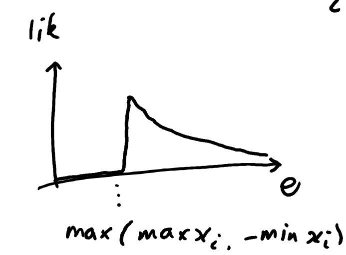
$$lik(\theta | x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) = \frac{1}{2^n \theta^n} \prod_{i=1}^{n} 1_{-\theta \le x_i \le \theta}$$

$$= \frac{1}{2^n \theta^n} 1_{-\theta \le x_i \text{ for all } i} \, 1_{x_i \le \theta \text{ for all } i}$$

$$= \frac{1}{2^n \theta^n} 1_{-\theta \le \min x_i} \, 1_{\max x_i \le \theta}$$

$$= \frac{1}{2^n \theta^n} 1_{\theta \geq -\min x_i} \, 1_{\theta \geq \max x_i}$$

$$= \frac{1}{2^n \theta^n} 1_{\theta \geq \max(\max x_i, \, -\min x_i)}.$$



The MLE is thus $\hat{\theta} = \max\left(\max_i x_i, \, -\min_i x_i\right) = \max_i |x_i|$

(b) The Resampling Method is based on the approximation

$$\mathbb{P}(\theta \in [l(\underline{X}), u(\underline{X})]) \approx \mathbb{P}(\hat{\theta}(\underline{X}) \in [l(\underline{X}^*), u(\underline{X}^*)])$$

where θ is the unknown parameter,

$[l(\underline{X}), u(\underline{X})]$ is the confidence interval, a function of the sample

$\hat{\theta}(\underline{X})$ is the MLE for θ, computed on the sample

$\underline{X}^*$ is a resampled version of the sample.

It's up to us to invent $[l, u]$.

In this case, based on the shape of the likelihood function, a reasonable choice is $[M, M(1+\varepsilon)]$ where $M = \max_i |x_i|$.

Thus,

$$\mathbb{P}(\theta \in [M, M(1+\varepsilon)]) \approx \mathbb{P}(\hat\theta \in [M^*, M^*(1+\varepsilon)])$$

where $\hat\theta$ is the MLE computed on the actual data.

It's up to us to decide how to resample the data to get M^*. The goal is to simulate "How else might this experiment have turned out?"

Idea 1: let X_i^* = one of $X_1, \cdots, X_n$, chosen at random with replacement. (resample from the empirical distribution)

Idea 2: let $X_i^* \sim \text{Uniform}[-\hat\theta, \hat\theta]$ (parametric resampling)

Idea 3: let $X_1^*, \cdots, X_n^*$ be a shuffled version of $X_1, \cdots, X_n$. (permutation resampling).

In this course we've seen 1 and 2, and your supervisor might have told you about 3. For this question, 3 is no good — it would always yield exactly the same M^*, for any random shuffle. The other two will give plausible answers — we'd see a range of outcomes for M^*. As far as this course is concerned, you should just pick a method and explain why it gives a plausible answer ie a probability that isn't automatically 0 or 1.

Finally, the ε parameter can be tuned to achieve a target probability. To do this slickly, we want e.g.

$$\mathbb{P}(\hat\theta \in [M^*, M^*(1+\varepsilon)]) = 95\%$$

$$\Rightarrow \mathbb{P}\left(\frac{\hat\theta}{M^*} \in [1, 1+\varepsilon]\right) = 95\%$$

$$\Rightarrow \mathbb{P}\left(\frac{\hat\theta}{M^*} \leq 1+\varepsilon\right) = 95\%$$ since $M^* \leq \hat\theta$ using any of the resampling methods described above

$$\Rightarrow \mathbb{P}\left(M^* \geq \frac{\hat\theta}{1+\varepsilon}\right) = 95\%$$

$$\Rightarrow \frac{\hat\theta}{1+\varepsilon} = q_{0.05}$$ where q is the 5%ile of M^*, ie $\mathbb{P}(M^* \leq q_{0.05}) = 0.05$.

$$\Rightarrow \quad \frac{\hat{\theta}}{1+\varepsilon} = q_{0.05} \quad \text{where } q \text{ is the 5\%ile of } M^*, \text{ ie } \mathbb{P}(M^* \le q_{0.05}) = 0.05.$$

$$\Rightarrow \quad 1+\varepsilon = \hat{\theta} / q_{0.05}$$

This gives us the following code:

```
def confint(xs, p):
    θ̂ = mle(xs)
    M*s = [mle([random.uniform(-θ̂, θ̂) for i in range(n)])
           for rep in range(10 000)]          generate 10,000 samples of M*
    q = quantile(M*s, 1-p)                    Sanity check: each M* is in the range [0, θ̂]
                                              so q ≤ θ̂. The bigger p is, the smaller q is.
    return [θ̂, θ̂ × (θ̂/q)]

def mle(xs):
    return max_i |xs_i|
```

(c) Let $Y_i = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{else} \end{cases}$, $X_i = \begin{cases} \text{Uniform}[-\theta,\theta] & \text{if } Y_i = 1 \\ \text{Normal}(0,\sigma^2) & \text{if } Y_i = 0. \end{cases}$

Then

$$\mathbb{P}(X_i \le x) = \mathbb{P}(X_i \le x \mid Y_i = 1)\,\mathbb{P}(Y_i = 1) + \mathbb{P}(X_i \le x \mid Y_i = 0)\,\mathbb{P}(Y_i = 0)$$

$$= \mathbb{P}(\text{Uniform}[-\theta,\theta] \le x)\; p + \mathbb{P}(\text{Normal}(0,\sigma^2) \le x)\;(1-p).$$

The density of X_i is

$$f(x) = \frac{d}{dx}\mathbb{P}(X_i \le x) = p\,\frac{d}{dx}\mathbb{P}(\text{Uniform}[-\theta,\theta] \le x) + (1-p)\,\frac{d}{dx}\mathbb{P}(\text{Normal}(0,\sigma^2) \le x)$$

$$= p \times f_{\text{Uniform}[-\theta,\theta]}(x) + (1-p)\, f_{\text{Normal}(0,\sigma^2)}(x).$$

$$= \frac{p}{2\theta} 1_{-\theta \le x \le \theta} + (1-p)\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2}$$

You could have written this out straight away, if you're confident enough! Personally, I like to go via the distribution function $\mathbb{P}(X_i \le x)$ and

You could have written this out straight ...

Personally, I like to go via the distribution function $\mathbb{P}(X_i \le x)$ and then differentiate. It's a more general method. See e.g. Q2 (a).

$$\text{loglik}(\theta, p, \sigma) = \sum_i \log f(x_i)$$

$$= \sum_i \log \left(\frac{p}{2\theta} 1_{-\theta \le x_i \le \theta} + (1-p) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} x^2} \right).$$

(d) let $q \in \mathbb{R}$, $p = \dfrac{e^q}{1+e^q} \in [0,1]$

see section 3.1, 'natural parameters'.

let $\rho \in \mathbb{R}$, $\sigma = e^\rho \in (0, \infty)$.

```
def loglik2(θ, q, ρ):
    p = e^q/(1+e^q)
    σ = e^ρ
    return loglik(θ, p, σ).
```

Question 2

(a) $\mathbb{P}(T \le t) = \mathbb{P}(\max_i h(X_i) \le t)$

$= \mathbb{P}(h(X_i) \le t \text{ for all } i)$

$= \mathbb{P}(h(Y_j) \le t \text{ for all } j)$ where $Y_1, \dots, Y_m$ are the distinct items

$= \mathbb{P}(h(Y_1) \le t) \times \dots \times \mathbb{P}(h(Y_m) \le t)$
since distinct items yield independent hashes

$= t^m$ since $h(Y_j) \sim \text{Uniform}[0,1]$

density $f(t)$ is $\frac{d}{dt}\mathbb{P}(T \le t) = \frac{d}{dt} t^m = m\, t^{m-1}$ (for $0 \le t \le 1$)

(b). This question might throw you...
In the course, we only worked through the MLE calculation for a random sample (of independent, identically distributed random variables).
There's no reason why the sample size has to be larger than 1!
Generally speaking, the MLE is whatever value of the parameter maximizes the likelihood (ie probability density) of the observed data.
Here, the observed data consists of a single value of T.

$\text{lik}(m) = m\, t^{m-1}$

$\text{loglik}(m) = \log m + (m-1)\log t$.

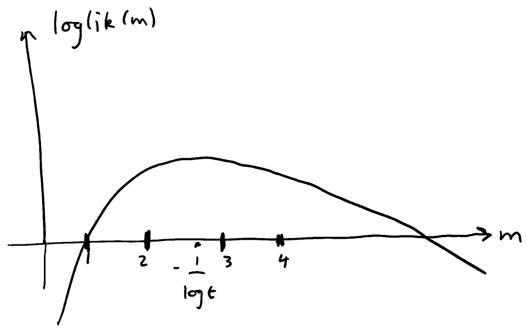
sanity check: $t < 1$ so $\log t < 0$
so this function has a maximum at finite m

To maximize this: $\frac{d}{dm}\text{loglik}(m) = \frac{1}{m} + \log t = 0 \Rightarrow \hat{m} = \frac{-1}{\log t}$.

Sanity check: $0 < t < 1$, so $-\infty < \log t < 0$,
so $\frac{-1}{\log t} > 0$

To be pedantic, m must be an integer, and $\text{loglik}(m)$ is concave (since $\frac{d^2}{dm^2}\text{loglik}(m) < 0$), so the actual $\hat{m}$ is whichever yields the greater loglik out of floor$\left(\frac{-1}{\log t}\right)$, ceiling$\left(\frac{-1}{\log t}\right)$.

e.g.



(c) Step 1. Invent a probability we want to estimate.

In this question, we're asked to produce a confidence interval for m,

so let's estimate

$$\mathbb{P}\left(m \in \left[\hat{m}(1-\varepsilon),\ \hat{m}(1+\varepsilon) \right] \right) \quad \text{where} \quad \hat{m} = \frac{-1}{\log T}.$$

(and then we can tune ε to make this probability equal to say 95%).

A confidence interval has the general form

$$\mathbb{P}\left(\text{unknown parameter} \in \text{interval, which depends on the data} \right) = p.$$

It's up to us to invent the shape of the interval to use. Here I chose $[\hat{m}(1-\varepsilon),\ \hat{m}(1+\varepsilon)]$ because it looks reasonably sensible given the loglik function, and it's simple to write down. It could equally well be $[\hat{m}-\varepsilon,\ \hat{m}+\varepsilon]$ or $[\hat{m}-\alpha,\ \hat{m}+\beta]$, or anything else you want.

Step 2. Replace random variables by their resampled versions; replace unknown parameters by their MLE.

$$\approx \mathbb{P}\left(\hat{m} \in \left[m^*(1-\varepsilon),\ m^*(1+\varepsilon) \right] \right) \qquad (1)$$

where $\hat{m}$ is obtained from the actual data we've seen and m^* is resampled.

Step 3. Decide a plausible way to resample.

See discussion from Q1.

Let's compute $\hat{m}$ from the data as observed, then generate $\hat{m}$ independent Uniform $[0,1]$ random variables, and compute T^* and m^* from them.

There are more sophisticated answers! This answer is good enough for the exam.

Step 4. Tune ε so that the confidence interval has the right probability.

Let's say we're trying to produce a 95% confidence interval

Rewrite (1) as

$$\mathbb{P}\left(\frac{\hat{m}}{m^*} \in [1-\varepsilon,\ 1+\varepsilon] \right)$$

$$= \mathbb{P}\left(\left| \frac{\hat{m}}{m^*} - 1 \right| < \varepsilon \right) = 0{\cdot}95$$

So compute $\hat{m}$ from the data, compute say 10 000 samples of

So, compute $\hat{m}$ from the data, compute say 10 000 samples of m^*, compute $\left|\frac{\hat{m}}{m^*} - 1\right|$ for each sample, and pick

$$\varepsilon = \text{quantile}\left(\left|\frac{\hat{m}}{m^*} - 1\right|, 0.95 \right).$$

Then, we report the confidence interval

$$\left[\hat{m}(1-\varepsilon),\ \hat{m}(1+\varepsilon) \right].$$

Question 3

(a) A Markov chain is <u>stationary</u> if its distribution does not change over time,

ie if X_1 has the same distribution as X_2

has the same distribution as X_3 See Section 4.3.1

has the same distribution as X_4

$\vdots$

(b) $$\mathbb{E} X_{n+1} = \mathbb{E}(\alpha X_n + \sigma \varepsilon_n)$$ See Section 2.1

$$= \alpha \mathbb{E} X_n + \sigma \mathbb{E} \varepsilon_n$$

$$= \alpha \mathbb{E} X_n \quad \text{since } \varepsilon_n \sim N(0,1) \text{ so } \mathbb{E}\varepsilon_n = 0.$$

$$\operatorname{Var} X_{n+1} = \operatorname{Var}(\alpha X_n + \sigma \varepsilon_n)$$

$$= \alpha^2 \operatorname{Var} X_n + \sigma^2 \operatorname{Var} \varepsilon_n \quad \text{since } \varepsilon_n \text{ are all independent}$$

$$= \alpha^2 \operatorname{Var} X_n + \sigma^2 \quad \text{since } \varepsilon_n \sim N(0,1) \text{ so } \operatorname{Var} \varepsilon_n = 1$$

If the sequence is stationary, $\mathbb{E} X_{n+1} = \mathbb{E} X_n$ hence $\mathbb{E} X_n = 0$.

And $\operatorname{Var} X_{n+1} = \operatorname{Var} X_n = \rho^2$ say, where

$$\rho^2 = \alpha^2 \rho^2 + \sigma^2 \Rightarrow \rho^2 = \frac{\sigma^2}{1-\alpha^2}.$$

(c) We know that there is a link between the stationary distribution and the limiting distribution. Let's try to calculate the limiting distribution, ie the distn. of X_n for large n — it may give us a hint about the stationary distn.

$$X_0$$

$$X_1 = \alpha X_0 + \sigma \varepsilon_0$$

$$X_2 = \alpha^2 X_0 + \alpha \sigma \varepsilon_0 + \sigma \varepsilon_1$$

$$X_3 = \alpha^3 X_0 + \alpha^2 \sigma \varepsilon_0 + \alpha \sigma \varepsilon_1 + \sigma \varepsilon_2$$

$\vdots$

$$X_3 = \alpha^3 X_0 + \alpha^2 \sigma \varepsilon_0 + \alpha\sigma\varepsilon_1 + \sigma\varepsilon_2$$

$$\vdots$$

So $X_n = \alpha^n X_0 + \sigma(\varepsilon_{n-1} + \alpha\varepsilon_{n-2} + \cdots + \alpha^{n-1}\varepsilon_0)$

By the rules for $\mathbb{E}$ and Var, and using the formula $1 + r + \cdots + r^{n-1} = \frac{1-r^n}{1-r}$,

$$\sigma(\varepsilon_{n-1} + \cdots + \alpha^{n-1}\varepsilon_0) \sim N\left(0, \sigma^2 \frac{1-\alpha^{2n}}{1-\alpha^2}\right).$$

For large n, we expect

X_n is approx. dist. like $N\left(0, \frac{\sigma^2}{1-\alpha^2}\right)$.

Does $N\left(0, \frac{\sigma^2}{1-\alpha^2}\right)$ make sense as a stationary distribution? It certainly has the right mean and variance (from part (b)) — but there are many other distributions with these parameters, not just the Normal. Let's verify, using the definition of stationarity.

- Suppose $X_n \sim N(0, \sigma^2/1-\alpha^2)$
- Then, $X_{n+1} = \alpha X_n + \sigma\varepsilon_n \sim N\left(0, \alpha^2 \frac{\sigma^2}{1-\alpha^2} + \sigma^2\right)$

  ie $N\left(0, \sigma^2\left(\frac{\alpha^2 + 1 - \alpha^2}{1-\alpha^2}\right)\right)$

  ie $N(0, \sigma^2/1-\alpha^2)$.

We conclude that $N(0, \sigma^2/1-\alpha^2)$ is a stationary distribution.

(It takes more maths to show that it is <u>the</u> stationary distribution, much more maths than would fit into this course.)

This question requires you to understand what a stationary distribution is, and the answer given here involves a deep understanding of the relationship between stationary distributions and limiting distributions. You haven't seen any worked examples or example sheet questions quite like this. To do well in data science, you need an agile mind that can apply a fairly limited collection of concepts in all sorts of new ways. Exam questions will test your mental agility.

Question 4

(a) Prior distribution for α_1 is

$$\Pr(\alpha_1 = x) = K\, x^{\delta-1}(1-x)^{\delta-1} \quad \text{for some constant } K$$

Posterior dist. is

$$\Pr(\alpha_1 = x \mid w_1 \text{ wins}, \ell_1 \text{ losses}) \propto \Pr(\alpha_1 = x)\, \mathbb{P}(w_1 \text{ wins}, \ell_1 \text{ losses} \mid \alpha_1 = x)$$

using Bayes' rule

$$= K\, x^{\delta-1}(1-x)^{\delta-1} \binom{w_1+\ell_1}{w_1} x^{w_1}(1-x)^{\ell_1}$$

number of wins ~ Binomial (number of plays, probability of win)

$$\propto x^{\delta+w_1-1}(1-x)^{\delta+\ell_1-1}$$

ignoring terms that don't involve x

This is the density of a Beta distribution; specifically

posterior dist. of α_1 ~ Beta $(\delta+w_1, \delta+\ell_1)$.

For reference (to be used in part (d)):

$$\text{posterior mean} = \frac{\delta+w_1}{2\delta+w_1+\ell_1}$$

(b)

```
def nextmove(w1, l1, w2, l2):
    # use Monte Carlo method to estimate P(α1 > α2)
    α1s = [rbeta(δ+w1, δ+l1) for i in range(10000)]
    α2s = [rbeta(δ+w2, δ+l2) for i in range(10000)]

    p = (number of cases where α1 > α2, from these two lists) / 10000

    if random() < p:
        return "Play machine 1"
    else:
        return "Play machine 2".
```

w_1, ℓ_1: wins + losses on machine 1; w_2, ℓ_2: wins + losses on machine 2

The point of the question is: the probabilistic strategy requires that we compute a probability. You could calculate it with an that we compute a probability. You could calculate it with an integral, if you know enough maths. But it's pretty easy to approximate it directly, using Monte Carlo integration, as above. The examiner will be looking for the word "Monte Carlo" in your answer.

For the probabilistic strategy,

$$\mathbb{P}(\text{"Play machine 1"}) = \mathbb{P}(A_1 > A_2)$$

where $A_1 \sim \text{Beta}(\delta + w_1, \delta + \ell_1)$ and $A_2 \sim \text{Beta}(\delta + w_2, \delta + \ell_2)$.

This is exactly what the Thompson sampling strategy produces! In other words, the probabilistic strategy and the Thompson sampling strategy will produce exactly the same distribution of outcomes. The only difference is that the probabilistic strategy (in the form given above) is way way less efficient.

(c) Stop and think before answering! In the course, we've seen two sorts of confidence interval:

$$\mathbb{P}(X \in [\mu - 1.96\sigma, \mu + 1.96\sigma]) \approx 95\%$$

using the central limit theorem
where $\mu = \mathbb{E}X$, $\sigma^2 = \text{Var}\,X$,
as in Section 2.3.

$$\mathbb{P}(\alpha_1 \in [x, y]) = \int_x^y \Pr(\alpha_1 = s)\, ds$$

where α_1 is an unknown parameter
which we're analysing with Bayesianism,
(or we could use a resampling approximation)

In the first case, the term on the left is an observable quantity, and we are working out the range of likely outcomes of a trial.

and we are working out the range of likely outcomes of a trial.

In the second case, the term on the left is an unknown parameter which we never actually observe directly.

In this question, we're asked for a 95% confidence interval for the number of wins, which is an observable quantity, which hints that we want an answer of the first type:

no. wins in n_1 plays ~ Binomial (n_1, α_1) where α_1 is machine 1's true (unknown) success probability.

So $\mathbb{E}$ no.wins $= n_1 \alpha_1$, Var no.wins $= n_1 \alpha_1 (1-\alpha_1)$

In the exam, you'd be told the mean and variance of the binomial.

So no.wins $\approx$ Normal $(n_1\alpha_1, \; n_1 \alpha_1 (1-\alpha_1))$

So $\mathbb{P}\left(\text{no.wins} \in \left[n_1 \alpha_1 - 1.96\sqrt{n_1\alpha_1(1-\alpha_1)}, \; n_1\alpha_1 + 1.96\sqrt{n_1\alpha_1(1-\alpha_1)}\right]\right)$

$\approx 95\%$

You could in principle combine the two types of confidence interval: use this approximation for no.wins $| \alpha_1$, and combine it with $\Pr(\alpha_1)$, using Bayes' rule. This takes a lot more work, and in the exam the question would be phrased more tightly to nudge you towards the easier answer.

(d) Suppose, for the sake of argument, that machine 1 truly has the larger payout probability. Write $\bar{\alpha}_1$ for the true (unknown) payout probability, and α_1 for the posterior distribution from (a).

- Greedy might have an early run of bad luck on machine 1, leading to a low posterior mean for α_1.

  Then it would keep playing machine 2. By part (c), the no. of wins on machine 2 would likely be around the no. of wins on machine 2 would likely be around

$$n_2 \left(\bar{\alpha}_2 \pm \frac{1}{\sqrt{n_2}} k \right) \quad \text{for some constant } k.$$

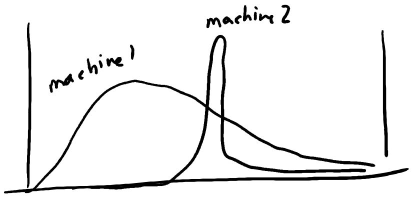
so the posterior distribution will have mean around

$$\frac{\delta + n_2 \left(\bar{\alpha}_2 \pm \frac{1}{\sqrt{n_2}} k \right)}{n_2 + 2\delta} \approx \bar{\alpha}_2 \pm \frac{1}{\sqrt{n_2}} k.$$

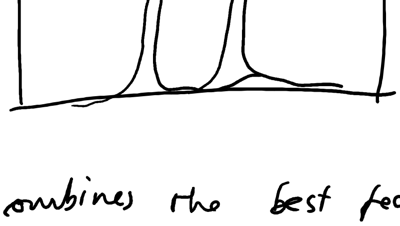
So, it can get stuck with an unduly low α_1, and α_2 that gets closer and closer to the true value $\bar{\alpha}_2$.

- ε-greedy keeps on playing both machines, so it keeps on refining its posterior distributions, so it gets closer and closer to learning $\bar{\alpha}_1$ and $\bar{\alpha}_2$, so it never gets stuck on the wrong machine. The price it pays is that it never settles entirely on the right machine – it always throws away a fraction ε of its plays.

- The probabilistic / sampling strategies will play both machines while there is uncertainty



will occasionally play machine 1, because A_1 will sometimes be $> A_2$, thanks to the uncertainty of α_1



will hardly ever play machine 2

ie it combines the best features of the other two

By a similar argument to Greedy, the posterior distributions will approach the true probabilities $\bar{\alpha}_1$ and $\bar{\alpha}_2$.

Question 5

(a) $\text{lik}(\xi) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i) = \prod_{i=1}^{n} \begin{cases} e^{\xi}/1+e^{\xi} & \text{if } y_i = 1 \\ 1/1+e^{\xi} & \text{if } y_i = 0 \end{cases}$

$$= \frac{(e^{\xi})^m}{(1+e^{\xi})^n}$$ where n = number of samples, m = number of samples where $y_i = 1$

$$= \frac{e^{m\xi}}{(1+e^{\xi})^n}.$$

$$\log \text{lik}(\xi) = m\xi - n\log(1+e^{\xi})$$

MLE solves $\frac{d}{d\xi} = m - \frac{n e^{\xi}}{1+e^{\xi}} = 0$

$$\Rightarrow \frac{e^{\xi}}{1+e^{\xi}} = \frac{m}{n}$$

$$\Rightarrow e^{\xi} n = m + e^{\xi} m$$

$$\Rightarrow e^{\xi} = \frac{m}{n-m}$$

$$\Rightarrow \hat{\xi} = \log \frac{m}{n-m}.$$

(a_2) Let's assume that, given skill levels, matches are independent.

The tricky thing here is getting all the notation to work together. Hopefully, the first line of the answer to (a) will set you in the right direction.

$$\log \text{lik}(\mu_A, \mu_B, \mu_C) = \sum_{i=1}^{30} \left(\xi_i - \log(1+e^{\xi_i}) \right)$$

where $\xi_i = \mu_{\text{winner of match } i} - \mu_{\text{loser of match } i}$

If you feel masochistic, or if you don't read ahead to part (b), you could go through all the algebra:

let n_{AB} = # matches where A beats B, etc.

$\mathbb{P}(A \text{ beats } B \text{ in an } A+B \text{ match}) = \frac{e^{\xi_{AB}}}{1+e^{\xi_{AB}}}$ where $\xi_{AB} = \mu_A - \mu_B$

$$\mathbb{P}(A \text{ beats } B \text{ in an } A+B \text{ match}) = \frac{e^{\xi_{AB}}}{1+e^{\xi_{AB}}} \qquad \text{where } \xi_{AB} = \mu_A - \mu_B$$

$$\mathbb{P}\left(\text{observe all the outcomes } n_{AB}, n_{BA}, \dots\right) = \left(\frac{e^{\xi_{AB}}}{1+e^{\xi_{AB}}}\right)^{n_{AB}} \left(\frac{1}{1+e^{\xi_{AB}}}\right)^{n_{BA}}$$

$$\times \left(\frac{e^{\xi_{AC}}}{1+e^{\xi_{AC}}}\right)^{n_{AC}} \left(\frac{1}{1+e^{\xi_{AC}}}\right)^{n_{AC}}$$

$$\times \left(\frac{e^{\xi_{BC}}}{1+e^{\xi_{BC}}}\right)^{n_{BC}} \left(\frac{1}{1+e^{\xi_{BC}}}\right)^{n_{BC}}$$

$$\begin{aligned} \text{loglik}(\mu_A, \mu_B, \mu_C) &= 7\,\xi_{AB} - 10 \log(1+\xi^{AB}) \\ &+ 9\,\xi_{AC} - 10 \log(1+\xi^{AC}) \\ &+ 6\,\xi_{BC} - 10 \log(1+\xi^{BC}) \end{aligned}$$

(b) A <u>linear model</u> is a model with unknown parameters, in which the parameters are weighted by features and combined linearly. A <u>feature</u> is any measurable property of the objects being studied.

$$\begin{aligned} \underline{\xi} &= \mu_A \left(\underline{1}_{A\text{ won}} - \underline{1}_{A\text{ lost}}\right) \\ &+ \mu_B \left(\underline{1}_{B\text{ won}} - \underline{1}_{B\text{ lost}}\right) \\ &+ \mu_C \left(\underline{1}_{C\text{ won}} - \underline{1}_{C\text{ lost}}\right) \end{aligned}$$

An example feature: $\underline{x}_B = \underline{1}_{B\text{ won}} - \underline{1}_{B\text{ lost}}$.

This is a vector of length 30, one entry for each match, where

$$[x_B]_i = \begin{cases} 1 & \text{if } B \text{ played match } i \text{ and won} \\ -1 & \text{if } B \text{ played match } i \text{ and lost} \\ 0 & \text{otherwise} \end{cases}$$

(c) A collection of features $\underline{v}_1, \dots, \underline{v}_n$ is <u>linearly independent</u> if

$$\sum_{i=1}^{n} \lambda_i \underline{v}_i = \underline{0} \quad \Rightarrow \quad \lambda_i = 0 \text{ for all } i$$

i.e. if there is no non-trivial linear combination of them that adds up to $\underline{0}$.

i.e. if there is no non-trivial linear combination of them ...

In this case, $\underline{x}_A + \underline{x}_B + \underline{x}_C = \underline{0}$.

Example: if the match is between A and C and A wins

$x_A = 1, \quad x_B = 0, \quad x_C = -1, \qquad x_A + x_B + x_C = 0.$

Similarly for all other five match types.

So they are not linearly independent.

(d) What's the relevance of linear dependence? Writing $\underline{x}_C$ in terms of $\underline{x}_A$ and $\underline{x}_B$,

$$\underline{\xi} = \mu_A \underline{x}_A + \mu_B \underline{x}_B + \mu_C(-\underline{x}_A - \underline{x}_B)$$

$$= (\mu_A - \mu_C)\underline{x}_A + (\mu_B - \mu_C)\underline{x}_C$$

Thus, if we add +10 to μ_A and μ_B and μ_C, we won't change $\underline{\xi}$, so the likelihood will be unchanged. In other words there is no unique maximizer for the likelihood. (This will trip up most numerical optimization routines.) The problem is called <u>non-identifiability</u> — see Section 3.1.

First, note that $\underline{\xi} = (\mu_A - \mu_C)\underline{x}_A + (\mu_B - \mu_C)\underline{x}_B$, which implies the model is non-identifiable. To fix this, we may (without loss of generality) arbitrarily set $\mu_C = 0$. The model is then

$$\underline{\xi} = \mu_A \underline{x}_A + \mu_B \underline{x}_B$$

and it's easy to see that $\underline{x}_A$ and $\underline{x}_B$ are linearly independent.

Now we can simply use numerical optimization:

def negloglik(μ_A, μ_B):

$\quad \underline{\xi} = \mu_A \underline{x}_A + \mu_B \underline{x}_B$

$\quad$ return $-\left(\text{sum}(\underline{\xi}) - \text{sum}(\log(1 + e^{\underline{\xi}}))\right)$

the negative of the log likelihood function

$\mu_C = 0$ — without loss of generality, $\mu_C = 0$

$(\mu_A, \mu_B) = \text{fmin}(\text{negloglik}, (0,0))$. — find μ_A, μ_B to maximize loglik

A more intelligent initial guess, instead of $(0,0)$, might be to use part (a), restricting attention to AC and BC games respectively. We have set $\mu_C = 0$, so $\mathbb{P}(A \text{ wins an } A+C \text{ game}) = \dfrac{e^{\mu_A}}{1+e^{\mu_A}} \Rightarrow \hat{\mu}_A = \log \frac{9}{1}$

and $\hat{\mu}_B = \log \frac{6}{4}$. These initial guesses ignore the extra information about AB matches, which is why we need numerical optimization — but nonetheless it's always good to do some thinking before rushing in with numerical optimization; and the examiners will be delighted to see you link back to part (a).