Extra questions for MLPR

Lecturers are often asked to provide more questions. At this level you are supposed to be moving towards doing independent research: you need to be able to come up with questions, *and answers*, yourself.

Strategy for 4th/5th year level courses: First, go over all the material. For each part, imagine trying to explain it to someone else. Can you say how it relates to other parts of the course? Where there are explanatory diagrams you should be able to label, explain, and reproduce them. In general try to create a small example, application or extreme case and play with it. If you have trouble, find the material in another reference, or ask a friend. If/when you understand the material, imagine what questions could be asked about it. Ask yourself how you might check the answers to those questions yourself. (Your dissertations should demonstrate these skills!)

That said, there are some extra questions that may be useful below. We will *not* provide detailed worked answers to these questions. Talk them over with your class-mates and if necessary ask specific questions on NB. Iain can give feedback on any written work that you send him before January (after which he is on leave).

Other textbook questions may of course be interesting. There are also questions in the lecture materials. It's surprising how few answers or questions about questions in the lecture materials tend to be posted to NB. You can get our feedback there…

Links to books

Murphy Free to view online via the University library web site. Search there, or *maybe* this link will work.

MacKay (free pdf online)

Barber (free pdf online, page refs won't match hard copy)

Bishop, not freely available in electronic form. Contains many informative but mathy exercises.

Monte Carlo

Murphy Ex 23.2, p835. Further questions I'd ask myself include: 1) Would other proposal distributions work (e.g., Gaussian?). 2) Why is this exercise possible, and what is the barrier to constructing rejection samplers in other cases? 3) If I was just interested in some expectation under the Gamma distribution, such as $E[x^3]$, what else might I do, and would it be better/worse?

For simplicity assume we have a situation where we can normalize both our target and proposal distributions, and we know $c = \max P(x)/Q(x)$. Describe how importance sampling works, and derive the probability of accepting a proposal. If proposals Q don't work well with importance sampling, will it necessarily be bad for rejection sampling and why?

MacKay Ex 29.2, p363. A question on importance sampling (with an answer).

MacKay Ex 29.3, p367. On speed of progress of Metropolis as a function of step-size. What's worse, getting a step size too big or too small?

(MacKay Ex 29.8 and 29.9, p374 are good for people wondering why I introduced the R reverse transition operator rather than just talking about detailed balance. Although this stuff is getting into technical minutiae.)

MacKay Ex 29.10, p377 on slice sampling.

Murphy Ex 24.1, p873 on Gibbs sampling. If you're prepared to do some work to understand a model, you could further check your understanding on Ex 24.2, Ex 24.3. Or MacKay Ex 29.17 p383.

If keen, there's a Summer School lab where you can compare MH and slice sampling on a more complicated example.

Gaussian approximations

Most of the textbook questions are far too mathy or involved for this course.

Have you done the extra Laplace approximation question from tutorial 6?

Why does the second derivative of the 'energy' with respect to each variable have to be positive to apply the Laplace approximation? Why does the Hessian matrix have to be positive definite (or at least semi-definite)? Might these constraints be violated given some target posterior distribution? If so, how?

The Monte Carlo methods rejection sampling, importance sampling and Metropolis–Hastings all contained distributions called Q. Which, if any, of the methods we've seen for fitting Gaussian approximations (Laplace, and KL-divergence each way around) might be useful for each of these algorithms? For each of the combinations, why or why not?

(Murphy Ex 21.2, p764 has a multivariate Laplace approximation if you want to do more maths.)

Imagine we want to choose between two models. Example 1: two logistic regression models where one has more fetures than the others. Example 2: logistic regression models with different priors on the parameters $p(w) = N(w; 0, 1)$ and $p(w) = N(w; 0, 100)$. Explain why we can't compare training set performance for the most probable (MAP) weights to pick models in these cases. How could we use the Laplace approximation to modify this idea (Answer: Eq 28.10, p350

MacKay)? What other ideas have you seen in this course that you could use to pick between the models?

Gaussian processes (GPs)

(I'm not the only one who thinks coming up with your own questions is useful. Compare to MacKay Ex 45.1, p534! I'm afraid neither MacKay nor Murphy has any useful GP exercises.)

Barber Chapter 19, and Bishop p320– have a lot of quite mathematical exercises related to kernels and GPs. The MLPR exam is likely to ask questions based on a more intuitive understanding of the model and how inference works, rather than some complex manipulation under exam conditions.

For questions more aligned with what is expected for MLPR, see the lecture log and the tutorial self-study sheet 8.

PCA

See tutorial self-study sheet 8, which uses PCA as a vehicle to revisit topics from earlier in the course.

Imagine a dataset containing the heights of children in nanometers (billionths of a metre) and their masses in kg. If we were to reduce this data to one dimension by PCA, what would happen and why? What could we do to fix the problem?

Fred points out that if we could scale up a child uniformly in all directions to be twice as tall, they'd have eight times the volume and mass. Would it be sensible to apply a non-linear transformation to the features before applying PCA?