Response to thesis comments

Jack Kennedy

June 6, 2023

Two versions of the updated thesis have been provided. The content of each is identical, but in `thesis-highlight.pdf`, changes are marked in blue. In `thesis.pdf`, changes are not highlighted.

This response document is structured as a point-by-point response to each comment from the examiners. Comments from the examiners are *in italics* and preceded by the page number that is quoted in the examiners' report. My changes are in standard typeface, and preceded by 'JK:'.

1 Chapter 1

P3 *A clear statement of which elements of this thesis are Jack's and which are contributions from co-authors needs to be included, otherwise I cannot assess the originality of Jack's PhD*

- JK: I have added a statement just after the title page which explains that the work was my own, apart from some stated exceptions.

2 Chapter 2

P15 *Elicitation of probability distributions. I did not understand what "least squares" refers to in the following sentence: "More generally, if there are more p_i than parameters of the chosen distribution, or the form of the distribution does not lend itself to a tractable solution, then a numerical least squares approach may be appropriate."*

- JK: I have clarified this statement by including an appropriate expression to be minimised so that the distribution can be fitted.

P23 *The $<$ (or is it \prec?) symbol is undefined in item (2) — only the symbol with the equals underneath was defined*

- JK: The symbol is indeed `\prec`. The meaning of $\prec$ is now explained at the bottom of P22.

P23 *Under these conditions there is a utility function, u, satisfying" — where is a proof, or a reference?*

- JK: Start of Section 2.3 gives multiple references to this result: "Unless explicitly stated otherwise, this section is based upon Keeney & Raiffa (1976), Smith (2010) and chapters of Dias *et al* (2018); in particular González-Ortega *et al* 2018."

P25 *The usage of the ">=" symbol in the first stand out equation has not been defined.*

- JK: The meaning of $\succeq$ is now explained immediately after this equation.

3 Chapter 3

P30 *covariance functions allow us to specify beliefs about local behaviour" — I disagree; the covariance function is defined for all inputs x, x', and is therefore capturing global properties. Local properties are things like values f(x), derivatives f'(x), that are defined at a single point (or neighbourhood of a single point)*

- JK: Changed to "... covariance functions allow us to concisely specify complex structure such as highly nonlinear behaviour."

P31 *Grammar: "The first covariance function we consider a squared exponential covariance function."*

- JK: This sentence has been revised.

P31 *On this page, and also later, the Euclidean norm* $||.||$ *is applied to scalars. Whilst not incorrect, this usage is unconventional, and usually the modulus* $|.|$ *is preferred.*

- JK: Modulus is now used for scalars.

P31 *"For larger* θ*, the correlation decays quite slowly" — well, convergence is still at an exponential rate, which we could contrast to the slower polynomial decay of a Matern covariance function*

- JK: This has been corrected to explain that the correlation decays at a reduced rate (rather than "slowly").

P37 *"it is computationally desirable to set* λ *to be some small value(say* $\lambda = 10^{-6}$*) to avoid computational issues surrounding matrix inversion". It is a major bugbear of mine that nuggets are so widely used in statistics, when there is literally a sub-field of machine learning dedicated to fast approximate GP calculations. The value* $\lambda = 10^{-6}$ *is entirely arbitrary, and ought to depend on the scale of the data at least (if the data are on scales less than* $\lambda = 10^{-6}$ *then the nugget will swamp the signal). I would suggest removing this "rule of thumb", or at least acknowledging its arbitrariness and pointing out that there is a rich literature of better alternatives that one could use instead.*

- JK: I have changed the word "desirable" to "convenient" and at the end of this paragraph, and cited some alternative methods.

P38 *"In a Bayesian framework we could elicit a prior* $\pi(\beta, \theta)$ *then use MCMC" — one need not use MCMC to perform a non-conjugate Bayesian analysis, there are lots of different computational methods that could be used.*

- JK: This has been changed to "In a Bayesian framework we could elicit a prior $\pi(\boldsymbol{\beta}, \Theta)$ then use numerical methods to obtain either exact or approximate samples from $\pi(\boldsymbol{\beta}, \Theta \mid \mathcal{D})$"

P39 *"The Student's t process is more complex to work with, thus the GP is often preferred." Why is it more complex?*

- JK: This has been changed to "The Student's t process arises only from a certain families of prior specification, thus a GP is often preferred" The t process is I guess not much more complicated, but rather the resulting posterior under a limited collection of prior specifications.

P39 *The notation for covariance matrices of the form K(X,X) and C(X,X) is undefined.*

- JK: the notion $C(X, X)$ is now defined at the start of P38 (where it is first used)

P39 *"Unless explicitly stated, $C(\cdot, \cdot)$ is used for squared exponential covariance functions" — why? This choice is unfavourable from a computational perspective (since the eigenvalues of covariance matrices decay exponentially quickly, making matrices numerically ill-conditioned, much worse so that for any Matern), and experts in GPs such as Michael Stein repeatedly advise against its use (due to the problems associated with singularity of Gaussian measures). I would therefore hope there is a strong argument in favour of using it, but what is it*

- JK: Numerical conditioning is not a problem in this work; the simulations we emulate are stochastic and the (estimated) nugget term/measurement error has always been large enough to mitigate issues around numerical conditioning. Further, the methodologies proposed are not reliant on a squared exponential covariance, thus a Matérn (or anything else) could be used where desired.

P43 *Best not to use the term "nugget" when referring to measurement error, since "nugget" is usually understood as being a cheap hack, rather than part of the statistical model*

- JK: This has been rephrased as 'measurement error/nugget term' since within much of the emulation literature, the terms are used interchangeably

P43 *Inconsistent use of x and $\boldsymbol{x}$*

- JK: The inconsistencies have been corrected

P44 *"The previous figures in used"*

- JK: This has been changed to "The previous figures in the chapter used..."

P44 *"uniform designs can often lead to design points being very close together in space" — perhaps you mean "random designs"?*

- JK: I have changed "uniform designs" to "randomly generated designs"

P44 *"In general, there is no globally superior choice as to whether we go with one-shot or sequential designs" — I disagree. If there is non-stationary to be learned then we need a sequential design, while on the other hand there is also theory for the use of sequential designs in the stationary context that guarantees they are still space filling. See the work of Luc Pronzato, for example his SIAM JUQ paper from 2020"*

- JK: I was taking a more pragmatic view here. One shot designs are easy to use and generate, thus are computationally much simpler than a sequential design. This is a benefit. The Pronzato 2020 reference has been added just before the start of section (3.4), on P47

P45 *"The final advantage of one-shot design is that they are agnostic to choice of covariance and mean function". Isn't this a bad thing? For a non-stationary covariance function, a non-space-filling design will be optimal in general. E.g. consider the covariance function* $C(x, x') = 1_{x>0}1_{x'>0}\exp(-(x-y)^2)$*, which generates functions such that* $f(x)$ *is identically* 0 *for* $x < 0$*. There would be no need to explore* $x < 0$ *in that example.*

- JK: I've decided to remove the word 'advantage', but explained that one-shot designs can be useful as an exploratory tool when little is known about $Cov(f(x), f(x'))$. We then go on to explain the use cases and advantages of one shot and sequential designs.

P46 *What you have called "Active Learning McKay" is called "uncertainty sampling" in the machine learning literature. See e.g. "Nguyen VL, Shaker MH, Hüllermeier E. How to measure uncertainty in uncertainty sampling for active learning. Machine Learning. 2022 Jan;111(1):89-122.*

- JK: A comment has been included mentioning that ALM is called uncertainty sampling in the ML literature; the Nguyen *et al* (2022) reference is given at the end of the comment.

P47 *In (3.29), isn't y random and unobserved - in which case how do we evaluate(3.29)?*

- JK: Good point. A simple approximation, based on a sum and some observed data, is now given below this equation.

P48 *In (3.31) the conditioning variables should be removed, since this is a mathematical equality and not a distributional statement*

- JK: The conditioning variables have been removed.

P49 *"it then follows that" - only if* f_t *and* $\delta_{(t+1)}$ *are also assumed to be independent. Several omissions of independence assumptions of this kind can be found in this chapter and chapter 4*

- JK: This has been reworded to include a statement about independence.

P50 *(3.38) is missing a factor of 1/T.*

- JK: 1/T now pre-multiplies the sum

P51 *Emulating a computer model with fixed fidelity T is not the true goal. What if T cannot be made large enough to reduce the discretisation error in the computer model to a negligible level? See e.g. "Teymur, O., Foley, C., Breen, P., Karvonen, T. and Oates, C.J., 2021. Black box probabilistic numerics. Advances in Neural Information Processing Systems, 34, pp.23452-23464", which formulates the problem of predicting the limit as* $T \rightarrow$ *infinity.*

- JK: At the end of Section 3.4 a comment is included which states that Teymur *et al* 2021 explores the idea of $T \to \infty$, which can be useful when T can not be made sufficiently large/the most accurate simulator level is too coarse.

P52 *I don't understand why the joint predictive likelihood of the test data was not used –why are we assuming the predictions are statistically independent, when we constructed them deliberately to be dependent?*

- JK: When the number of test data points is small and the points are well spaced, they should be approximately independent. This is added just below equation (3.41)

P58 *"Linear regression (Bayesian or frequentist) offers a much faster inference and prediction framework than GPs (Rougier et al., 2009)." Please rephrase, as Gaussian linear regression is a special instance of GP regression – as explained in detail in the opening chapters of Rasmussen and Williams (2006)."*

- JK: I've now stated "Linear regression (Bayesian or frequentist) offers a much faster inference and prediction framework than GPs with, for example, squared exponential or Matérn covariance structures, (Rougier et al., 2009).". A few lines down I explain that in the case of Bayesian linear regression, the posterior may be tractable thus complex computations such as matrix operations only have to be done once.

P60 *"the Bayes linear approach works exceptionally well when B is approximately Normal, but offers a valid and robust analysis when B is non-Normal"* - *in what sense is it "robust"? If this claim cannot be made precise, it should be removed*

- JK: The word "robust" has been removed

4 Chapter 4

P67 *"independent, statistical" → "independent statistical"*

- JK: The comma has been removed.

P69 *I do not understand how you are training your HetGP model. Equation (4.7) involves $\lambda^2(X)$, but this is not something that we can directly observe, so how do we make use of (4.7)? I understand that you want to first estimate lambda and then plug in an estimate into the GP for f, but I do not see any explanation of how you estimate lambda in the first instance. This lack of understanding prevented me from thoroughly assessing this part.*

- JK: Since $\lambda^2(X)$ is a latent variable, it can be estimated via a point estimate in an empirical Bayes or MAP framework. Further details of estimation are in the section 4.5.3 and section 4.6. A comment is added on P69 to explain this, and the reader is directed to sections 4.5.3 and 4.6 for more details.

P78 *In (4.22) we see the covariance of Z^C but then we are told that Z^C is an expectation and therefore wasn't random in the first place. Which is it?*

- JK: Z^C is the mean/expectation of a stochastic process. This mean is uncertain, it therefore has its own mean and (co)variance

P79 *Could an objective prior, such as the g-prior of Zellner, be a more useful and automatic alternative to the normal inverse gamma prior that you have used (whose hyperparameters need to be specified)?*

- JK: I would argue that there is no such thing as an objective prior.

P79 *"Note the absence of λ_E^2" — but it appears in (4.31) when * = E*

- JK: λ_*^2 is now outside the block of equations which allows us to state separately that $\lambda_*^2 \sim Inv - Gamma(e_{j,*}, f_{j,*})$ for $* \in \{C, V\}$

P79 *"For β_j^*we adopt independent N(0, 1) priors.Because our GP is on the probit scale this prior covers a wide range of observable values" — but the choice of prior on regression coefficients should reflect also the scale of the covariance that enter into the regression model, not just the scale of the response variable. This prior is surely not appropriate in general*

- JK: A comment about the scale is now included in the sentence: "Because our GP is on the probit scale — and our inputs are mean centred and scaled to have unit variance — this prior covers a wide range of observable values"

P79 *Here and elsewhere the term "MAP" is used incorrectly (in my opinion), and what you are doing would be called either Empirical Bayes or Type II Maximum Likelihood (since beta are always integrated out)*

- JK: I’ve now used more precise/conventional language. Empirical Bayes (EB) is used when β are integrated out, MAP is reserved for when *all* parameters are estimated via a point estimate.

P84 *”It took 5.7 seconds to fit HetGP and 29.6 seconds to fit SML on a laptop with 4 × 2.40 GHz processors and 8 GB RAM” — what is the relevance of the number of cores?*

- JK: Specification has been included for completeness.

P84 *No details are given on how the Empirical Bayes parameters were numerically approximated. What numerical method was used? Did it work well? How can we be sure? Were all the parameters even identifiable? Could the challenge of doing numerical optimisation over a higher-dimensional parameter space have contributed in some way to the observed decline in performance in the “full MAP” method that you report on P89?*

- JK: We have expanded on the details on P77: “MAP estimates are found via a numerical optimisation of the log-posterior (up to an additive constant) using the optimizing function from rstan (Stan Development Team, 2020), which uses the L-BFGS algorithm for numerical optimisation. To prevent a local mode being chosen as the maximiser, we recommend running the `optimizing()` function multiple times and selected the best result. In our work, we run optimizing three times”. For the comment about the challenge of numerical optimisation possibly contributing to worse results - yes this is a valid point. However the empirical Bayes method provides us with enhanced uncertainty quantification (we have a tractable posterior on β) *and* it is faster, so the full MAP approach feels like a non-starter anyway. Of course, we would not have known that the full MAP was slower if we had not implemented it In fact, we actually tried this approach before implementing the EB approach.

P91 *As I understand it, the claimed novelty in this method is that we are analytically integrating out beta in HetGP. But, as we discussed earlier in the thesis, giving Gaussian priors to the coefficients in the linear regression part is equivalent to using a GP with mean 0 and a covariance function of the form (3.11). So one could argue that this contribution, of integrating out beta, is not actually novel in the context of HetGP. Is there any other novelty in the proposed method, and if so what?*

- JK: The main novel feature of this chapter is the SML emulator (which is an extension of HetGP; we use the mean function from a cheap version of a stochastic simulator to aid the emulation of an expensive, stochastic simulator). From a “pure” HetGP perspective, we have considered different algorithms for estimating a HetGP and provided guidance about their relative merits. For example, the EB approach is — in our case — faster than the MAP approach, and offers more (i.e. uncertainty quantification about β parameters is automatic). The abstract of the thesis has been updated to better reflect the way in which the reader should interpret the novelty of the approach. I have also included a couple of sentences at the start of chapter 4, which states the novel contributions of this chapter.

5 Chapter 5

P98 *"hypersphere of radius 1/2" — shouldn't this be radius 1?*

- JK: I think $\frac{1}{2}$ is correct. The diameter would then be 1 meaning that each $x_i \in [0, 1]$ and the hypersphere has the same width as the hypercube on $[0, 1]^k$ in each margin

P102 *How can we interpret S_J and S_{T_J} in general?*

- JK: The interpretations are now included.

P104 *"unbiased unbiased"*

- JK: The 2nd unbiased has been removed.

P105 *Better methods to approximate nested expectations are called "nested Monte Carlo", and there is some theory on these in the machine learning literature that can be cited.*

- JK: On P106 a reference to Rainforth *et. al.* 2018 has been added with some comments about the use cases of nested monte carlo and applicability to this thesis.

P106 *Algorithm 1 seems to implicitly assume that G(x) is the uniform distribution, without remark. The line $\bar{X} < -X$ need not be inside a loop, as I understand*

- JK: Directly above the algorithm it is now stated that $G_i(x_i)$ is uniform. A comment about non-uniform distributions is now given below the algorithm. The line $\tilde{X} < -X$ is now outside the inner loop.

P106 *The approach of Marrel et al. (2012) formulates a stochastic computer model y() as a function of x, the simulator inputs, but also x_ε, the state of the pseudo random number generator of the simulator". To me this seems like a bad idea — why would we assume there is any continuity or smoothness in the map from x_ε to simulator output? Why would we go so far as to assume this relationship is infinitely smooth by using a Gaussian covariance? For example, simulators may involve some form of logic based on the outcome of discrete random events, which renders this mapping extremely complex. And the choice of parameterisation for x_ε is certainly arbitrary — how do we select a parameterisation to use?*

- JK: I don't think this does assume that $y(\cdot)$ is smooth w.r.t. x_ε. We simply quantify the amount of variability in the observed response induced by x_ε via the nugget term (i.e. modelled by a non-differentiable process).

P108 *"where $E\{\ell(x)\}$ is the 'true' value of $\ell(x)$" — but $E\{\ell(x)\}$ does not appear in (5.36)...?*

- JK: I have changed $\text{E}[\ell(\boldsymbol{x}) \mid \boldsymbol{x}]$ in 5.36 to $\text{E}[\ell(\boldsymbol{x})]$.

P133 *"The data plotted in Figure 5.6 is" — data ... are*

- JK: The "is" has been changed to "are".

6 Chapter 6

P124 *Please avoid using "y" as a dummy variable, as we have already committed to using "y" for data*

- JK: The dummy variable for this integration has been changed to v.

P124 *Here and elsewhere it is being assumed that minima/maxima are unique, e.g. in (6.7). But this either needs to be explicitly assumed or the "=" should be changed to "\**in**".*

- JK: On the line above Eq 6.7 I have now been explicit that we assume a single maximum exists.

P124 *"When $f(\cdot)$ is a deterministic,"*

- JK: This has been changed to "When $f(\cdot)$ is a deterministic function,"

P127 *derivatives incorrect at top of p127*

- JK: I've double checked these and they appear to be correct. Intermediate calculations are reasonably complex but there is a lot of cancellation.

P136 *"...Figure 6.4) The two ..."* t

- JK: This has been changed to "...Figure 6.4). The two ..." (i.e. added full stop before "The")

P136 *"Suppose that there exists a 'best input' (or optimal decision) x^*" — but x^* does not seem to appear anywhere in the sequel?*

- JK: Some context has been added after Eq (6.24): "The goal of history matching is to find inputs that return outputs which are "close" to $y(\boldsymbol{x}^*)$, relative to all the uncertainties in the problem that we are willing to specify."

P137 *After (6.25) we read ".. where ε_{MD} is a mean-zero error-like term which accounts for model discrepancy .Systematic model discrepancy has not been included in Equation (6.25)." This sounds like a contradiction?*

- JK: The intention was to communicate how you could adjust for model discrepancy when the DM can specify more structure. Hence, I have changed 'systematic' to 'structured'.

P139 *For the NROY to always be bounded we would need to assume that the domain $\mathcal{X}$ itself is bounded, wouldn't we? I cannot see where this was assumed.*

- JK: In this particular instance, $\mathcal{X}$ is bounded by the definition of (6.25). However, to be clear in the general case it is now stated that $\mathcal{X}$ is a bounded set when setting up the history matching problem at the start of Section 6.4 (P135).

P140 *The weaknesses of history matching were not discussed. Are data retained or discarded between waves? If the NROY region is disconnected, then we can certainly fit independent emulators in different regions, but why would we expect this to work well from the perspective of statistical efficiency? If we rule out good values in any wave, we can never get them back: Due to stochasticity in the model, it may well be that almost surely the NROY region is empty in the limit as the number of waves k is increased, since there is always a small chance that any given point will be deemed implausible. Is this not a problem with the use of history matching in the context of a stochastic model?*

- JK: A subsection (6.4.2) has been added to address some drawbacks of history matching (HM). If the NROY region is disconnected, independent emulators will be statistically more efficient in that, a simple way to emulate a non-stationary function is to partition the function's domain up in such a way that within each sub-domain, the function is (approximately) stationary across each subdomain. This is the approach of Gramacy & Lee 2008, which is the reference at the end of the sentence in question. It is possible that, as $k \to \infty$, the NROY region will be empty. However, HM is a finite process (and our analysis only performed 2 waves), so this is not an issue for the thesis.

P144 *"the reduction in the NROY volume is small, which suggests we should terminate the HM procedure". What do you mean by "suggests" here? Are you explicitly advocating this as a termination criterion, and if so how would you precisely formulate it? On P162 you write that we should stop when "the reduction in the size of the NROY volume is negligible", but this is also a vague statement*

- JK: I think this is a necessarily vague statement. I do not think there is a good rule of thumb (or theoretical stopping rule) which could be implemented. The analyst must make their best judgement about whether more waves are worthwhile.

P149 *I believe Algorithm 3 will fail in general, and not just for pathological instances of history matching. In general, Gibbs sampling (which is all this Algorithm is, with rejection sampling used to facilitate sampling from each conditional in turn) fails when the support of the target distribution decomposes into two or more sets such that there is zero probability of the sampler transitioning between these sets (we say the Markov chain is "reducible"). Consider for example a 2-dimensional NROY of the form* $[0, 0.4]x[0, 0.4] \cup [0.6, 1]x[0.6, 1]$. *Initialising Algorithm 3 in the set* $[0, 0.4]x[0, 0.4]$ *will mean that it never leaves this set, since according to the algorithm we must first sample* x_1 *from [0,0.4], and then sample x2 from [0,0.4], and repeat. In high dimensions it is quite reasonable to expect to see two or more disconnected sets in the NROY space, and no reason to think that these sets should be lined up in just the right way to ensure a Gibbs sampler will work. The fact that disconnected sets are not seen in the plots in the thesis is not evidence to rule out this possibility, since the Gibbs sampler may have simply missed other regions of the NROY space. I imagine that this point has some major ramifications for chapter 6 and chapter 7.* ***Recommended resolution: Instead,try several random initialisations and collect together all of the NROY points found.***

- JK: For chapter 6, the algorithm has remained with a statement explaining that the sampler is not guaranteed to be ergodic. New algorithm (Algorithm 4) has been included to include the resolution. For chapter 7, 30 chains, of length 1000 were constructed to form a single run. We investigated the chains (in chapter 7) to look for (non)-ergodicity in the following way:

  1. Run the corrected algorithm. Take note of the chain from which a sample originated from.
  2. Perform PCA on the pooled chains and return the rotated samples.
  3. Plot PC1 against PC2 to observe if clusters are present and highlight just one chain on the plot. If the highlighted chain does not move between any evident clusters, this suggests that individual chains may not be ergodic.

  In our case, this check did not provide any evidence that our particular sampling scheme is not ergodic. In particular, the chains all seem to lie in the same region and the highlighted chain covers this region well. The results of a simulation experiment which checks for multiple regions has been added to Section 7.6.2. In particular, see Fig 7.13 which shows no evidence that the NROY region is a union of disconnected sub-regions.

7 Chapter 7

P160 *3 (ii) has a grammatical problem*

- JK: This has been rephrased as "Since $x_{1:9} \in \mathcal{S}_n^8 \subset \mathbb{N}^9$, with $n \in \mathbb{N}$, the use of gradient-based optimiser is difficult.".

P165 *"by the Central limit theorem (CLT), a large enough number of replicates means y(x) should be approximately Normal" — y(x) will be constant, not normal (by the laws of large numbers). What point are you trying to make here?*

- JK: The point that I'm trying to make is that the distribution of $y(\boldsymbol{x})$ will, for sufficiently large levels of replication, be unimodal (as required by Pukelsheim's 3σ rule). For large, but still finite, n, $y(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} u_i(\boldsymbol{x})$ should be unimodal but still stochastic since $u_i(\boldsymbol{x})$ are assumed to be stochastic.

P172 *Algorithm 4 will fail in the same way as Algorithm 3, I think*

- JK: Algorithm 4 (now 5) is still included, but it is explained that it is not ergodic. A new algorithm (Algorithm 6) is included which is ergodic — it repeats the algorithm for many random initialisations.

P193 *"Firstly, we are not aware of any simulation experiments which address how often a history matching based approach will rule out the 'true' maximiser" — why do you think this is? Is this not a problem for the thesis, since we are using something with no idea if it works well or not?*

- JK: I think in our case this is not an issue as we achieved our aims - to reduce the decision space to a smaller set of 'good' decisions (e.g. those with high expected utility) with an aim to support decision making. I think part of the reason that there is little theoretical investigation is the complexity of the NROY space which is (typically) a region which is defined non-parametrically. A comment communicating these ideas/points has been added directly after the sentence of interest.