Summary of report – Modernizing Model-Based Reinforcement Learning:

This paper introduces an improved (or as the author puts it, modernized) approach to the forward propagation component of policy search algorithms in generalized control problems. Beginning with the popular PILCO reinforcement learning algorithm, the author first describes how a Gaussian process (GP) model can be used to evaluate the results of some control policy as the PILCO algorithm’s “forward model”, using trajectory samples instead of moment-matching for state distribution approximation. Then recent developments in fast inference for scalable GPs, namely KISS-GP and LOVE, are described along with other approximations made in the proposed approach, which improve the speed of inference enough to yield a near-real time feasible learning process. The report concludes with a benchmark example which shows significant improvement in training time over leading approaches, but as of yet does not converge to a robust stable policy, which would be a future benchmark.

Review:

This report is quite complete, especially with respect to the introductory material. The introduction well-motivates the developed algorithm, and the background section describes in reasonable detail the inference involved in the underlying algorithm developed. The associated proposal lists two goals which have been achieved in this report, namely the application of scalable GP inference to algorithms such as PILCO, and the use of gradient-based policy search without the use of the moment-matching approximation described above (instead using trajectory samples obtained via GP models). Only the third task, scaling this approach to previously intractable problems, remains to be undertaken for the final report, as well as an analysis of failure conditions for methods which eschew the moment-matching approximation in favor of only trajectory samples (as proposed by the author in the midterm report) remain.

This paper scores well in all four criteria in the NIPS instruction outline. It appears thorough and careful to my eye, is readable and reasonably clear, and while it may be a case of “let’s apply this new fast algorithm to previously intractable problems” it certainly strikes me as original enough and then some. As for significance, the results reported for the benchmark test appear quite promising, and with a few more experiments this could prove a true contribution to the field. As far from an expert in reinforcement learning or control problems it strikes me that this is a significant result, and that if robust policies can be extracted from this method it may become a standard PILCO variant in a number of applications.

One concerning aspect of this paper is that it seems the benchmark experiment results were included in the proposal. While this isn’t inherently an issue it strikes me as a little disingenuous to state goals which have already been achieved (goals 1 and 2 of the proposal). Still, the results are of high enough quality and the final paper (when it is finished) will no doubt be a solid enough piece of work that I personally don’t mind if a significant bit of effort went into the project before the proposal was due – I just wanted to mention this because I did a small double take when comparing the proposal with the midterm report, noticing that figure 4 of the report was in fact included in the proposal itself.

The only other high-level gripe I have with this report is that it’s technically too long (between 5 and 6 pages), but given the detailed nature of the background section it seems like a case of “this project is farther along than required” rather than inability to pare down content.

Suggestions:

1. Figure placement and referencing – in general it is my opinion that figures should be reasonably close to the sections they are relevant to, and that no figure should go un-referenced. Currently the order of the word “figure” in this report is: Figure 1, page break, figure 2, reference to figure 2, reference to figure 1, page break, figure 3 (unreferenced!), reference to figure 4, page break, and then figure 4. I’m certain that by strategically placing a few “\floatbarrier”s or fixing figure positions the flow of information could be significantly improved.

If figure 3 is necessary then it should be necessary to talk about it somewhere in your results section. What is the significance of either plot? Why do I care what they look like and what do they say about your method?

Also it’s my opinion that the plots in figure 4 could be combined into a single plot, but that’s a judgement call based on space and is likely influenced by the length of the report.

1. In your description of approximations for model inference, section 4.1, you transition rather suddenly from talking about derivatives of your expected cumulative reward $J^\pi(\theta)$ and the (unnamed) objective function $R(\pi)$. I can see that they’re related, but not much more. Can you (a) give a name to the function $R$ to afford it physical intuition like you did for $J^\pi$, and (b) either explicitly describe why finding derivatives of $J^\pi$ would necessitate derivatives of $R$, or better describe their relationship when introducing $J^\pi$ so that it doesn’t feel like you’re pivoting between two discussions.
2. There is a small typo on the bottom of page 4 in the left column – close the parentheses on $p(\Delta^{(t)})$

Technical Questions:

1. Is there, and if so what is, a metric for judging a control scheme to be a “robust stable policy” as mentioned at the end of section 5.1? It seems to me that this is somewhat subjective, but I’d be interested to see what criteria are usually used to decide that quality. Is it perhaps that without perturbation the system remains in the desired state indefinitely? Or does “robust” refer to being robust to small perturbations? I ask because it seems to me that almost any policy that can swing a cartpole upright would likely know what to do if it were deflecting to one side slightly, unless there were deterministic initial conditions to the problem (and even so, observations take the form of random actions which ought to help in my mind). As a follow-up, is there anything that can be said about the density of these non-robust solutions as compared with robust solutions in the case of the cartpole problem? (I imagine the answer is no for more general problems)
2. Looking at figure 2, is there a principled way to choose $m$? It strikes me that this choice is key in realizing the on-line potential of this approach, so perhaps a look at choice of $m$ and its effect on overall training time and solution quality (with respect to robustness, if indeed there is some measure of it, see question 1) is in order.

Review Summary:

This report is clearly poised to become a high-quality paper with potentially serious impact. It leverages recent advancements in scalable GP regression techniques (KISS-GP and LOVE) to replace moment-matching with trajectory sampling for state estimation in the popular PILCO algorithm, improving upon both past attempts to leverage GP techniques in this context and past attempts to eliminate moment-matching entirely. It provides a promising benchmark and ambitious future work directions to finish fleshing out the algorithm. Aside from a few potential improvements in flow, the composition is generally good, the background is sufficient and to someone in the field of control and reinforcement learning this would no doubt be an excellent halfway mark for a paper.