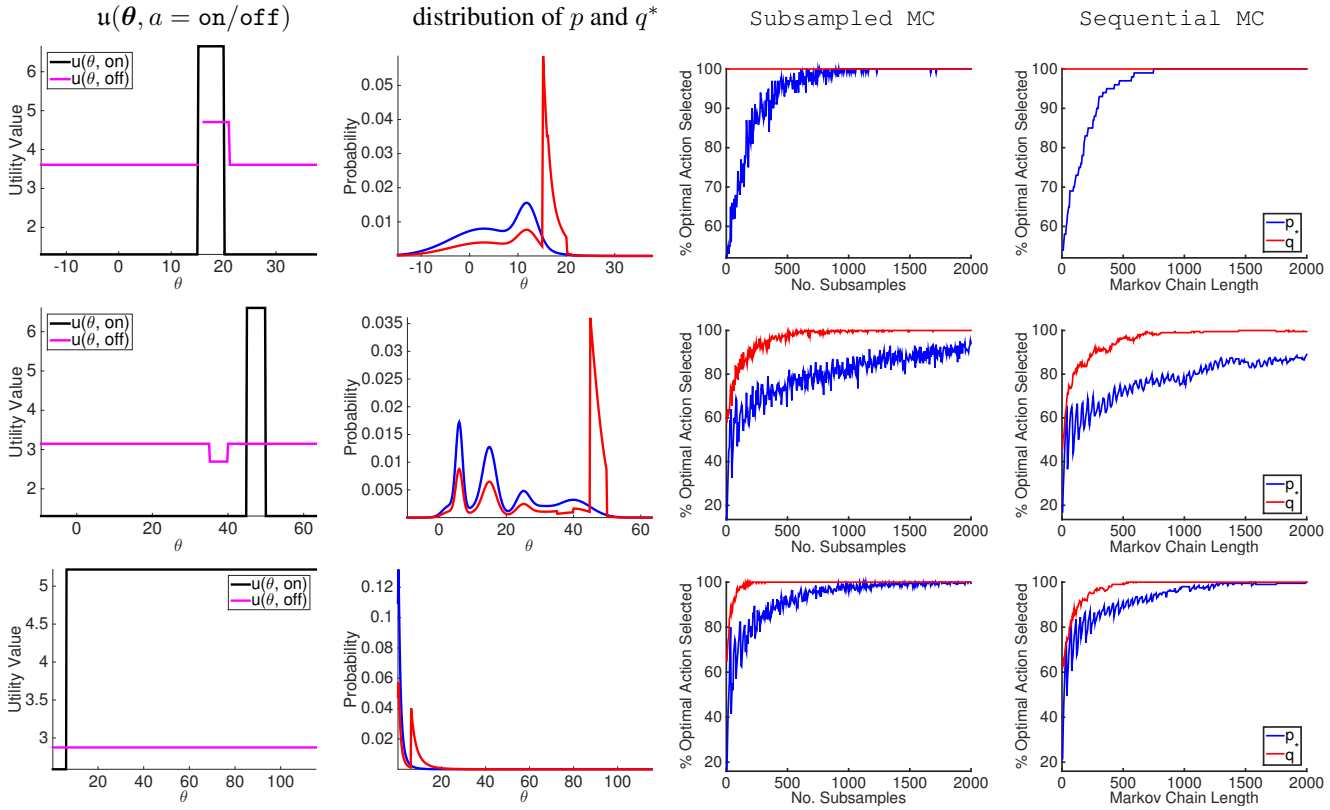
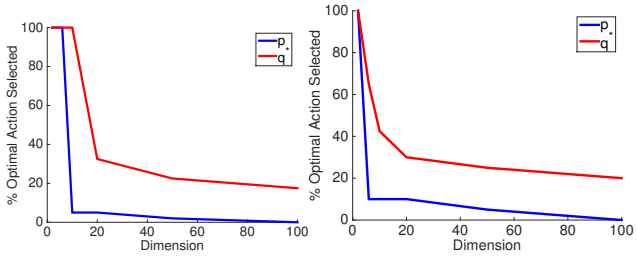


Figure 2: Power-plant simulations: the step-valued utility function (as in Equation 13) in the first column, the true distribution p (in blue) and q^* (in red) in the second column and in the third and forth columns the result of performing `Subsampled MC` and `Sequential MC` (as described in the text) are shown. In the two right-hand columns, note that q^* achieves the same percentage of optimal action selection performance as p in a mere fraction of the number of samples.

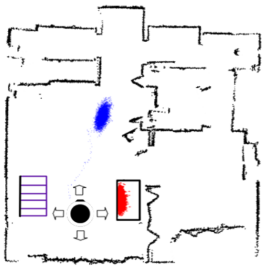
In Figure 2, the utility functions are shown in the first column (with black indicating the utility of on as detailed in Equation 13) and the temperature distributions (p in blue as described above and q^* in red) in the second column are shown. In the third and fourth columns the result of performing `Subsampled MC` and `Sequential MC` of the Metropolis-Hastings sampler for selecting the best action is shown such that the x-axis represents the number of samples and the y-axis shows the percentage of times the correct optimal action is selected. Here, in general, we observe that a significantly smaller number of samples from q^* is needed to select the best action in comparison to the number of samples from p required to achieve the same performance.

To investigate the performance of samples from p and q^* in higher dimensions, we use a d-dimensional Gaussian mixture corresponding to temperatures at each point in the plant as $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \mathbf{10}, \boldsymbol{\Sigma}) + \mathcal{N}(\boldsymbol{\theta}; \mathbf{20}, \boldsymbol{\Gamma})$ where $\mathbf{10}$ and $\mathbf{20}$ are d-dimensional vectors with constant value 10 and 20 as the mean and $\boldsymbol{\Sigma}_{i,j} = 5 + \mathbb{I}[i = j]$ and $\boldsymbol{\Gamma}_{i,j} = 3 + 7\mathbb{I}[i = j]$ as $d \times d$ covariance matrix. In addition, the utility function in Equation 13 is specified with $c^{(d)}_{\texttt{on},1} = 23, c^{(d)}_{\texttt{on},2} = 25, c^{(d)}_{\texttt{off},1} = 20, c^{(d)}_{\texttt{off},2} = 22, H_{\texttt{on}} = 50d, H_{\texttt{off}} = 13, L_{\texttt{on}} = 1.1, L_{\texttt{off}} = 1.5\log(d)$. In Figure 3 for $d \in \{2, 4, 10, 20, 50, 80, 100\}$, we observe that in an average of 100 runs of the MCMC with 200 samples, as the dimensions increase using q^* is more efficient. In fact, for a 100-dimensional bimodal Gaussian we are unable to find the optimal action using only 200 samples from p, which should be contrasted with the significantly improved performance given by sampling from q^*.

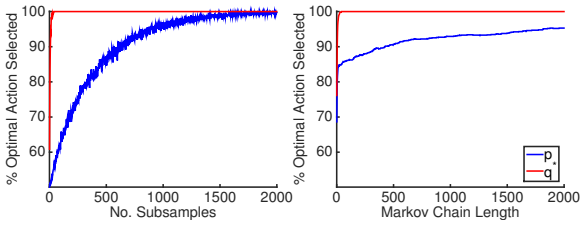


(a) `Subsampled MC` (b) `Sequential MC`

Figure 3: Performance of the decision maker in selecting the best action as the dimension of the problem increases in the power-plant. Note that at 100 dimensions, p is unable to select the optimal action whereas q still manages to select it a fraction of the time (and would do better if more samples were taken).



(a) Environment's Map



(b) `Subsampled MC` (c) `Sequential MC`

Figure 4: A robot's internal map showing the samples taken from its true belief distribution p (two modes are shown in blue, the second one is slightly obfuscated by the robot) and the optimal sampling distribution q^* derived by our loss calibrated Monte Carlo importance sampler in 4a. In 4b and 4c we see the performance (in terms of percentage of optimal action selected) of our loss-calibrated sampling method using q^* leads to near immediate detection of the optimal action in only a few samples.

Robotics

Another application where sampling has been commonly used is localization in robotics (Thrun 2000). It is a risk-sensitive-decision making problem where the robot is rewarded for navigating to the charger in order to maintain its power but wrong actions may lead to catastrophic incidents like falling down the stairs or crashing into obstacles. Due to minimal resources on-board a robot and the nature of the real-time localization problem, it is crucial for the robot to be able to select the optimal action rapidly, yet safely.

The state of the robot is the combination of its coordinates on a map and its heading direction. In our example for these experiments, we use a three dimensional Gaussian belief state distribution with two locations in a room intended to model that a robot's belief update has been confused by symmetries in the environment: one mode is at the robot's true location and the other at the opposite end of the room.

In this experiment, we consider a map as shown in Figure 4a where there is a flat in-door environment that the robot can move by selecting one of the four actions forward, backward, right or left. This action will lead to a movement step in robot from the current point on map with the heading direction towards the selected action. In doing so however, the robot has to avoid the stairs (low utility region) and select the charging source (high utility region).

Assuming a deterministic transition dynamics model $\boldsymbol{\theta}' = T(\boldsymbol{\theta}, a)$ and denoting $(T(\boldsymbol{\theta}_x, a), T(\boldsymbol{\theta}_y, a))$ as the location of the robot after taking action a from state $\boldsymbol{\theta}$ (that is, moving from the current location in the direction of the the selected action heading) and $\mathcal{R}_{\mathtt{r}}$ the set induced by region $\mathtt{r}$, we use the following utility function:

$$\mathrm{u}(\boldsymbol{\theta}, a) = \begin{cases} H & (T(\boldsymbol{\theta}_x, a), T(\boldsymbol{\theta}_y, a)) \in \mathcal{R}_{\mathtt{charger}} \\ L & (T(\boldsymbol{\theta}_x, a), T(\boldsymbol{\theta}_y, a)) \in \mathcal{R}_{\mathtt{stair}} \\ M & \text{otherwise} \end{cases}, \tag{14}$$

where $L < M < H$ (in our experiments: $L = 1, M = 10, H = 400$) and $a \in \{\mathtt{forward}, \mathtt{backward}, \mathtt{right}, \mathtt{left}\}$. Using distribution q^* from Theorem 5 as illustrated in Figure 4a, the samples from q^* (in red) concentrated on the charger's location which has higher utility value compared to the samples from p (in blue) that are from the mode of the distribution.

As shown in Figure 4b and 4c, using distribution q^* and running the same diagnostics as the previous experiment we see significant improvement in selection of the optimal action, requiring only a fraction of the samples of p to achieve the same optimal action selection percentage.

Conclusion and Future Work

We investigated the problem of loss-calibrated Monte Carlo importance sampling methods to improve the efficiency of optimal Bayesian decision-theoretic action selection in comparison to conventional loss-insensitive Monte Carlo methods. We derived an optimal importance sampling distribution to minimize the regret bounds on the expected utility for multiple actions. This, to the best of our knowledge, is the first result linking the utility function for actions and the optimal distribution for Monte Carlo importance sampling in Bayesian decision theory. We drew connections from regret to the probability of selecting non-optimal actions and from there to the variance. We showed using an alternative distribution as derived in Theorem 5 will sample more heavily from regions of significance as identified by their sum of utility differences.

Empirically, we showed that our loss-calibrated Monte Carlo method yields high-accuracy optimal action selections in a fraction of the number of samples required by loss-insensitive samplers in synthetic examples of up to 100 dimensions and robotics-motivated applications.

Future work should investigate the extension of the novel results in this work to the case of (a) continuously parameterized actions (Alessandro, Restelli, and Bonarini 2007), (b) imprecise utility functions (e.g, when the return of a state is not known precisely, but can be sampled) (Boutilier 2003), (c) uncontrollable sampling (where the utility partially depends on auxiliary variables that cannot be directly sampled from) and (d) applications in active learning and crowdsourcing (Beygelzimer, Dasgupta, and Langford 2009). Furthermore, the bounds obtained here are not tight in the multi-action setting and can be improved in future work.

Altogether, this work and the many avenues of further research it enables suggest a new class of state-of-the-art loss-calibrated Monte Carlo samplers for efficient online Bayesian decision-theoretic action selection.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Alessandro, L.; Restelli, M.; and Bonarini, A. 2007. Reinforcement learning in continuous action spaces through sequential monte carlo methods. In *Advances in Neural Information Processing Systems*.

Bartlett, P.; Jordan, I. M.; and McAaliffe, J. D. 2006. Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 101(473):138–156.

Berger, J. 2010. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition.

Beygelzimer, A.; Dasgupta, S.; and Langford, J. 2009. Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 49–56. New York, NY, USA: ACM.

Boutilier, C. 2003. On the foundations of expected expected utility. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, 285–290. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Gelman, A.; Robert, C.; Chopin, N.; and Rousseau, J. 1995. *Bayesian Data Analysis*. CRC press.

Geweke, J. 1989. Bayesian inference in econometric models using monte carlo integration. *Econometrica* 57(6):1317–1339.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. Applications of Mathematics. Springer, 1st edition.

Lacoste-Julien, S.; Huszar, F.; and Ghahramani, Z. 2011. Approximate inference for the loss-calibrated bayesian. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, volume 15, 416–424.

Neal, R. M. 1993. Probabilistic inference using markov chain monte carlo methods. Technical report, University of Toronto, University of Toronto.

Robert, C. 2001. *The Bayesian Choice*. Springer Texts in Statistics. Springer, 2nd edition.

Roberts, G. O.; Gelman, A.; and Gilks, W. R. 1997. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability* 7(1):110–120.

Rubinstein, R. Y. 1981. *Simulation and the Monte Carlo Method*. John Wiley & Sons, Inc., 1st edition.

Thrun, S. 2000. Probabilistic algorithms in robotics. *AI Magazine* 21:93–109.