Adapting Control Policies for Expensive Systems to Changing Environments

Matthew Tesch, Jeff Schneider, and Howie Choset
Robotics Institute, Carnegie Mellon University
{mtesch,schneide,choset}@cs.cmu.edu

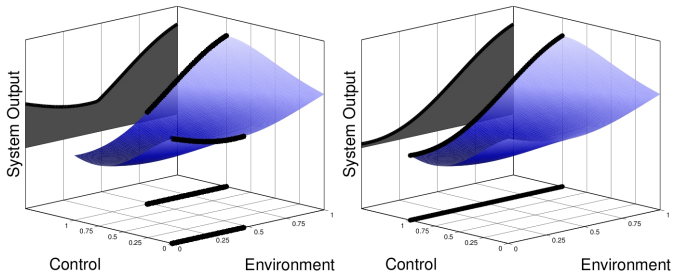
Abstract

Many controlled systems operate over a range of external conditions. In this work, we focus on the problem of learning a globally optimal policy to adapt a system's controller based on the value of these external conditions in order to maximize expected performance, even when the system output and policy score functions have local optima. In addition, we are concerned with systems for which it is expensive to run experiments, restricting the number that can be run during training. We formally define the problem setup and the notion of an optimal control policy. We consider myopic search methods using metrics based on the expected improvement in an objective, and propose two such algorithms. We present results comparing these algorithms with various other approaches and discuss the inherent tradeoffs in the proposed algorithms. Finally, we use these methods to train both simulated and physical snake robots to automatically adapt to changing terrain, and demonstrate improved performance on test courses with changing environments.

During typical operation of many robotic and industrial systems, the environment can change significantly. For example, a locomoting humanoid robot may move over gently up-sloped terrain, traverse a slightly bumpy horizontal area, and move downhill through many large obstacles. Assume there is a parameterized controller which can be tuned to perform well in each of these environments. Obviously, a static set of parameters for this controller would be a suboptimal method for controlling the system in multiple environments, as one would expect the controller parameters for uphill motion to be different than those for downhill. Although the robot's performance may be nonlinear with many local maxima, one would expect some continuity – similar control parameters should lead to similar performance, and similar environments should engender similar optimal control parameters. In this work, we seek to intelligently generate *control policies* that adapt to changes in the environment by selecting the best controller parameters for a given environment (Fig. 1).

Unfortunately, for some systems there is no known analytic expression or approximation for performance, and it is infeasible to test every possible controller in every possible environment. In particular, we focus on systems for which evaluation of a single controller/environment pair may take significant effort, and therefore we must minimize the necessary number of these evaluations. The choice of experiments (parameters at which to evaluate the system) can significantly affect the quality of a generated policy. The goal of *experiment selection* is to select parameters at which to evaluate in a manner that enables generation of globally optimal policies.

Effective policy generation is made possible by the assumption of continuous system output with respect to the controller parameters and the environment, allowing us to infer reasonable values for an unsampled parameter based on nearby sampled values. One of the keys to this work is the idea of using a *surrogate function* to represent a function which is expensive to evaluate, and basing search methods on this cheap model. These ideas have been extensively explored in the global optimization community [1, 2], often relying on stochastic processes to create a surrogate function [3, 4, 5]. Given a surrogate, the goal of expensive global optimization is to choose subsequent true function evaluations to minimize the number of total evaluations while maximizing global performance (avoiding



(a) (b)

Figure 1: We are interested in problems for which the optimal control parameter changes significantly depending on the environmental conditions. Ideally similar environment/control combinations lead to similar system output; we therefore assume this function is continuous, although these methods can still operate with some discontinuities. Here we show an example f for one dimensional environment and control spaces. The resulting policy tends to be piecewise continuous; it is shown below the function and its performance is projected to the left. Good policies can be estimated from a low-cost model of the true expensive system output, requiring only a handful of carefully chosen points. An optimal policy is illustrated in (a), presenting the best control parameter for every environment parameter. Note that the policy shown in (b), which also maps to a control parameter that maximizes f at $x_e = 1$, results in a significantly lower overall score because of its poorer choices in other regions of the environment space.

local maxima). In cases where the function evaluations are costly (hours to days), computational requirements are not a significant issue; careful choice of the sample is more important.

A number of heuristics and statistical methods have been derived to use information from the surrogate function to choose this sample location ([2] provides a survey of many existing methods). These metrics include the upper confidence bound of the predicted function [6, 7], the probability of improvement [8, 9], and the expected improvement [10, 11]. This last quantity has been shown to effectively trade off exploration of the parameter space and exploitation of the known good areas without requiring algorithm parameters to be carefully tuned.

However, none of these methods are directly applicable to the addition of environment parameters. In contrast, our algorithms explicitly account for these parameters through a metric based on myopic expected improvement, and evaluated on a learned surrogate function. Perhaps the most relevant related work is that done in the field of robust controller selection [12, 13]; this work also explicitly breaks the parameter space into control and environment subspaces, but seeks to find a *robust* controller, rather than an *adaptive* controller.

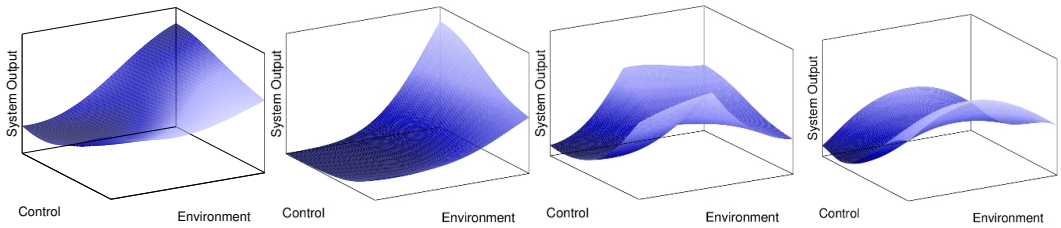
To formalize this problem, we define the notions of a *control policy* $\gamma\colon X_e \to X_c$ and the *score* $\mathcal{S}(\gamma)$ of that policy. A control policy is a mapping from the environment parameter space X_e to the control parameter space X_c; its score is the expected performance of that policy over X_e:

$$\mathcal{S}(\gamma) = \int_{X_e} \omega(x_e) f(x_e, \gamma(x_e))\, dx_e \tag{1}$$

Here ω is an optional probability distribution over environments, and $f\colon X_e \times X_c \to \mathbb{R}$ is the system output. The optimal policy γ^* is defined as $\operatorname{argmax} \mathcal{S}(\gamma)$. These ideas are illustrated in Fig. 1(a); γ^* is shown projected onto the control-environment plane, and $\mathcal{S}(\gamma^*)$ is visualized on the system output-environment plane; Fig. 1(b) shows a suboptimal policy.

The goal of this work is to find the highest scoring policy γ after a number of system output evaluations. In other words, the quantity of interest is not a single point, but a mapping from environment to control parameters. We break this problem into two subproblems:

1. **Policy Generation:** Given n evaluations of f, choose the best estimate for γ^*.
2. **Experiment Selection:** Choose the sequence of points $X = \{x^1, x^2, \ldots x^n\}$, $x^i \in X_e \times X_c$, where the choice of x^{k+1} is informed by $\{f(x^i)\ |i \leq k\}$, which maximizes the score of the policy produced by the chosen policy generation algorithm.



(a) (b) (c) (d)

Figure 2: **(a)** A surrogate for f; the color represents the confidence (dark = high, light = low). **(b)** The value of a potential experiment using expected improvement. The expected improvement metric biases point selection towards environments with high values of f. **(c)** The value of a potential experiment using expected improvement over the best predicted system output value for that environment. This reduces this bias toward easy environments and more directly optimizes the policy score. **(d)** The estimated policy score improvement as a result of sampling each point. This selection criterion is computationally intensive, but results in better performance than (c).

Proposed Methods

One of the difficulties in solving this problem lies in the fact that the true objective function we are maximizing during policy generation is $\mathcal{S}$. Unfortunately, as evaluating the score of a single policy involves an integral of evaluations of the expensive f, it is impossible to calculate $\mathcal{S}(\gamma)$ for any single policy γ, let alone apply a standard optimization technique to $\mathcal{S}$. Inspired by the success of the expected improvement metric for single-environment problems, the methods proposed in this work attempt to provide a tractable optimization method which maximizes a statistical quantity related to the score function: approximate expectation of improvement above the current predicted policy score.

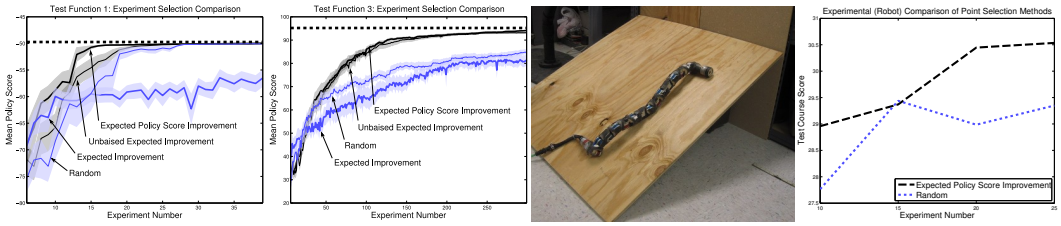
We use Gaussian processes ([14]) as a nonlinear function approximation method that generates a predictive distribution $p_x(y)$ at each $x \in X_e \times X_c$; techniques for the selection of the kernel and hyperparameters of such a function via likelihood maximization is detailed in [15]. By using this surrogate $\hat{f}$ to model f, we can address the policy generation and experiment selection subproblems. Policy generation involves selecting the policy which maximizes our best estimate of the score, as given by our surrogate function. Although more efficient approximations could be applied, in low dimensional spaces $\hat{\gamma}^*$ can be estimated via a dense sampling of the (relatively) cheap surrogate $\hat{f}$.

One method we propose for experiment selection adapts the basic idea of expected improvement to the explicit separation of the environment and control spaces. Instead of measuring improvement at a test point over the best evaluation of f so far, we restrict this notion to consider the improvement over the maximum predicted value of f for the *same environment* as the test point. This reduces the draw towards "easy" environments, giving the *unbiased expected improvement* (UEI) (Fig. 2(c)):

$$\mathrm{UEI}(x) = \omega(x_e) \int_{y^*_{x_e}}^{\infty} (y - y^*_{x_e}) p_x(y)\, dy, \text{ where } y^*_{x_e} = \max_{x_c \in X_c} (\hat{f}(x_c, x_e)) \tag{2}$$

Although UEI begins to approximate improvement of the true policy score function, it only considers improvement at one environment parameter. To measure the true expected improvement of the policy score at a point x, the expectation must be computed over the predictive distribution $p_x(y)$, where each potential y involves regression of a new surrogate $\hat{f}_y$ conditioned on the addition of this potential sampled value. This second proposed method produces better results, but requires additional computation (an expensive numeric integral) that may be impractical for some applications; it is termed the *expected policy score improvement*, or EPSI (Fig. 2(d)):

$$\mathrm{EPSI}(x) = \int_{-\infty}^{\infty} p_x(y) \int_{X_e} \omega(x_e) \max\left(\hat{f}_y(x_e, \gamma^*_y(x_e)) - \hat{f}(x_e, \gamma^*(x_e)), 0\right) dx_e\, dy \tag{3}$$



(a) (b) (c) (d)

Figure 3: **(a), (b):** Comparison of algorithms on analytic test functions. Each line represents the mean of 20 trials, the shaded regions indicate ± 1 standard error, and the dotted black line represents the best possible policy for that function. **(c):** The physical experimental setup: system performance is a measure of locomotive energy efficiency for a physical snake robot climbing up an incline. Higher amplitudes work well for flat ground, but smaller amplitudes allow the robot to climb steeper inclines without slipping backwards. **(d):** Performance of policies generated from points selected randomly versus using EPSI for the physical experiment trial.

Experimental Results and Concluding Remarks

To evaluate the performance of the two proposed algorithms we first used analytic test functions in leiu of a physical system, which allowed the completion of enough tests to enable one to draw reasonable conclusions. As standard search methods do not directly apply to this problem, for a baseline we compared both against the expected improvement global search algorithm EGO [11] applied directly to f, as well as random point selection. A summary of the results of these methods is shown in Fig. 3(a) and (b). Unsurprisingly, standard global optimization methods directly optimizing f initially have good performance, but the policy score tends to stagnate quickly. This is because these methods are optimizing f rather than $\mathcal{S}$ (Fig. 2(b)), and hence the resulting policy is very weak in "difficult" environments (ones with low values of f). As expected, random point selection also performs suboptimally, showing that it is important to carefully select experiments.

Unbiased expected improvement and expected policy score improvement both perform a better global search; interestingly the latter outperforms the former only slightly, indicating that UEI provides a much simpler and quicker method which produces similarly high quality results. The final choice between these methods involves several factors, and is largely application dependent. We note that while not strictly algorithm parameters, implementation decisions can have significant effects on performance of these algorithms. Experiments comparing the effects of sampling density and numeric integral resolution were also run; these results can be found in [16].

Although a complete analysis could not be run on physical systems due to the expensive nature of evaluating f and the resulting inability to compute a true policy score, we set up a range of environmental conditions in a "test course", and then used the above algorithms to generate policies which were scored on this test course. These policies map environment parameters (slope) into a 2-D gait parameter control space (see [17]). Results of evaluation of the policies generated during testing are shown in Fig. 3(d). Again, expected policy score improvement caused superior policies to be generated; differences can be noted even after only 10 samples of the space (the first 5 of which are the randomly generated initial sampling).

Although demonstrated here on snake robots, this framework is applicable to a rich set of problems. Locomoting systems, industrial processes, and prescription drugs all operate in changing environmental conditions, are expensive to test, and could benefit from optimal adaptive control policies. We have described two potential approaches for this experiment selection, both inspired by the statistical notion of expected improvement. One approach provides a fast, efficient computation that performs reasonably well, while the other is more complex and computationally intensive, but produces better results overall. We have also proposed a simple method for policy generation, given the experiments chosen by such an algorithm. We have demonstrated the efficacy of these algorithms, and presented a summary of results on analytic test functions as well as on a physical snake robot system. As a follow on to the work presented here, we are interested in relaxing the assumption that any environment/control combination can be tested during training, as well as investigating optimality bounds and completeness proofs for these algorithms.

References

[1] G. E. P. Box and N. R. Draper, *Empirical model-building and response surfaces*. Wiley, 1987.

[2] D. R. Jones, "A taxonomy of global optimization methods based on response surfaces," *Journal of Global Optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[3] R. G. Regis and C. A. Shoemaker, "A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions," *INFORMS Journal on Computing*, vol. 19, no. 4, 2007.

[4] H.-M. Gutmann, "A Radial Basis Function Method for Global Optimization," *Journal of Global Optimization*, vol. 19, 1999.

[5] K. Holmström, "An adaptive radial basis algorithm (ARBF) for expensive black-box global optimization," *Journal of Global Optimization*, vol. 41, no. 3, 2008.

[6] A. W. Moore and J. Schneider, "Memory-based stochastic optimization," *Advances in Neural Information Processing Systems*, pp. 1066–1072, 1996.

[7] D. Cox and S. John, "A statistical method for global optimization," in *1992 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1241–1246, Ieee, 1992.

[8] H. J. Kushner, "A new method for locating the maximum point of an arbitrary multipeak curve in the presence of noise.," *Journal of Basic Engineering*, vol. 86, pp. 97–106, 1964.

[9] A. Žilinskas, "A review of statistical models for global optimization," *Journal of Global Optimization*, vol. 2, pp. 145–153, June 1992.

[10] J. Mockus, V. Tiesis, and A. Zilinskas, "The application of Bayesian methods for seeking the extremum," *Towards Global Optimization*, vol. 2, pp. 117–129, 1978.

[11] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, vol. 13, no. 4, 1998.

[12] J. Lehman, *Sequential Design of Computer Experiments for Robust Parameter Design*. PhD thesis, Ohio State University, 2002.

[13] B. Williams, T. Santner, and W. Notz, "Sequential design of computer experiments to minimize integrated response functions," *Statistica Sinica*, vol. 10, no. 4, pp. 1133–1152, 2000.

[14] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[15] M. Tesch, J. Schneider, and H. Choset, "Using Response Surfaces and Expected Improvement to Optimize Snake Robot Gait Parameters," in *International Conference on Intelligent Robots and Systems*, (San Francisco), IEEE/RSJ, 2011.

[16] M. Tesch, J. Schneider, and H. Choset, "Adapting Control Policies for Expensive Systems to Changing Environments," in *International Conference on Intelligent Robots and Systems*, (San Francisco), IEEE/RSJ, 2011.

[17] M. Tesch, K. Lipkin, I. Brown, R. Hatton, A. Peck, J. Rembisz, and H. Choset, "Parameterized and Scripted Gaits for Modular Snake Robots," *Advanced Robotics*, vol. 23, pp. 1131–1158, June 2009.