Variational Bayesian Optimization for Runtime Risk-Sensitive Control

Scott Kuindersma
Computer Science Department
University of Massachusetts Amherst
scottk@cs.umass.edu

Roderic Grupen
Computer Science Department
University of Massachusetts Amherst
grupen@cs.umass.edu

Andrew Barto
Computer Science Department
University of Massachusetts Amherst
barto@cs.umass.edu

***Abstract*—We present a new Bayesian policy search algorithm suitable for problems with policy-dependent cost variance, a property present in many robot control tasks. We extend recent work on variational heteroscedastic Gaussian processes to the optimization case to achieve efficient minimization of very noisy cost signals. In contrast to most policy search algorithms, our method explicitly models the cost variance in regions of low expected cost and permits runtime adjustment of risk sensitivity without relearning. Our experiments with artificial systems and a real mobile manipulator demonstrate that flexible risk-sensitive policies can be learned in very few trials.**

I. Introduction

Experiments on physical robot systems are typically associated with significant practical costs, such as experimenter time, money, and robot wear and tear. However, such experiments are often necessary due to the extreme difficulty associated with constructing simulated systems of sufficiently high fidelity that behaviors translate to hardware without performance loss. For many nonlinear systems, it can even be infeasible to perform simulations or construct a reasonable model.

For this reason, model-free policy search methods have become one of the standard tools for constructing controllers for robot systems [27, 23, 12, 29, 17, 11]. These algorithms are designed to minimize the expected value of a noisy cost signal, $\hat{J}(\boldsymbol{\theta})$, by adjusting policy parameters, $\boldsymbol{\theta}$, for a fixed class of policies. By considering only the expected cost of a policy and ignoring cost variance, the solutions found by these algorithms are by definition *risk-neutral*, where the term *risk* is equivalent to *cost variance*. However, for systems that operate in a variety of contexts, it can be advantageous to have a more flexible attitude toward risk. For example, a humanoid otherwise capable of a fast and energy efficient gait might adopt a more predictable, possibly less energy efficient gait when operating near a large crater. Indeed, studies in human motor control and animal behavior suggest that variable risk sensitivity may be pervasive in nature [2, 1].

Recently there has been increased interest in applying Bayesian optimization algorithms to solve model-free policy search problems [17, 19, 20, 33, 28, 13]. In contrast to well-studied policy gradient methods [23], Bayesian optimization algorithms perform policy search by building a distribution of cost in policy parameter space and applying a selection criterion to *globally* select the next policy. Selection criteria are typically designed to balance exploration and exploitation with the intention of minimizing the total number of policy evaluations. These properties make Bayesian optimization attractive for robotics since cost functions often have multiple local minima and policy evaluations are typically expensive. It is also straightforward to incorporate approximate prior knowledge about the distribution of cost (such as could be obtained from simulation) and enforce hard constraints on the policy parameters.

Previous implementations of Bayesian optimization have assumed the variance of the cost signal is the same for all policies in the search space, which is not true in general. In this work, we propose a new type of Bayesian optimization algorithm that relaxes this assumption and efficiently captures both the expected cost and cost variance in regions of low cost. Specifically, we extend recent work developing a variational Gaussian process model for problems with input-dependent noise (or *heteroscedasticity*) [15] to the optimization case by deriving an expression for expected improvement [22], a commonly used criterion for selecting the next policy, and incorporating log priors into the optimization to improve numerical performance. We also consider the use of confidence bounds to produce *runtime* changes to risk sensitivity and derive a generalized expected risk improvement criteria that balance exploration and exploitation in risk-sensitive setting. We evaluate the algorithm in experiments with synthetic systems and a dynamic stabilization task using a real mobile manipulator.

II. Background

A. Bayesian Optimization

Bayesian optimization algorithms are a family of general stochastic optimization techniques that are well suited to problems where noisy samples of a cost function, $\hat{J}(\boldsymbol{\theta})$, are expensive to obtain [17, 6, 3, 28, 13]. In the control context, Bayesian optimization methods use data from previous policy evaluations to compute a nonparameteric distribution of cost in policy parameter space. Given this distribution, a decision-theoretic *selection criterion* is used to globally select policy parameter values, $\boldsymbol{\theta}$, that, e.g., have a high probability of having low cost or have high cost uncertainty.

Most Bayesian optimization implementations represent the prior over cost functions as a *Gaussian process* (GP). To fully specify the GP prior, $J(\boldsymbol{\theta}) \sim \mathcal{GP}(m(\boldsymbol{\theta}), k_f(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, one must define a mean function and a covariance (kernel) function, $m(\boldsymbol{\theta}) = \mathbb{E}[J(\boldsymbol{\theta})]$ and $k_f(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathbb{E}[(J(\boldsymbol{\theta}) - m(\boldsymbol{\theta}'))(J(\boldsymbol{\theta}) - m(\boldsymbol{\theta}'))]$. Typically, we set $m(\boldsymbol{\theta}) = 0$ and let $k_f(\boldsymbol{\theta}, \boldsymbol{\theta}')$ take on one of several standard forms. A common choice is the anisotropic squared exponential kernel,

$$k_f(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sigma_f^2 \exp(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^\top M(\boldsymbol{\theta} - \boldsymbol{\theta}')), \qquad (1)$$

where σ_f^2 is the signal variance and $M = \mathrm{diag}(\boldsymbol{\ell}_f^{-2})$ is a diagonal matrix of length scale hyperparameters. Intuitively, the signal variance captures the overall magnitude of the cost function variation and the length scales capture the sensitivity of the cost with respect to changes in each policy parameter. If prior information regarding the shape of the cost distribution is available, e.g., from simulation experiments, the mean function and kernel hyperparameters can be set accordingly [17]. However, in many cases such information is not available, so these quantities are optimized using maximum likelihood or maximum a posteriori techniques [24].

Samples of the latent cost function are assumed to have additive i.i.d. noise:

$$\hat{J}(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_n^2). \qquad (2)$$

Given a GP prior and data, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N]^\top \in \mathbb{R}^{N \times |\boldsymbol{\theta}|}$, $\mathbf{y} = [\hat{J}(\boldsymbol{\theta}_1), \hat{J}(\boldsymbol{\theta}_2), \ldots, \hat{J}(\boldsymbol{\theta}_N)]^\top \in \mathbb{R}^N$, one can compute the posterior (predictive) cost distribution for a policy parameterized by $\boldsymbol{\theta}_*$ as $\hat{J}_* \equiv \hat{J}(\boldsymbol{\theta}_*) \sim \mathcal{N}(\mathbb{E}[\hat{J}_*], s_*^2)$,

$$\begin{aligned} \mathbb{E}[\hat{J}_*] &= \mathbf{k}_{f*}^\top (\mathbf{K}_f + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y} & (3) \\ s_*^2 &= k_f(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) - \mathbf{k}_{f*}^\top (\mathbf{K}_f + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_{f*}, & (4) \end{aligned}$$

where $\mathbf{k}_{f*} = [k_f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_*), k_f(\boldsymbol{\theta}_2, \boldsymbol{\theta}_*), \ldots, k_f(\boldsymbol{\theta}_N, \boldsymbol{\theta}_*)]^\top$ and $\mathbf{K}_f$ is the positive-definite kernel matrix, $[\mathbf{K}_f]_{ij} = k_f(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$. When the hyperparameters are unknown, the log marginal likelihood, $\log p(\mathbf{y}|\boldsymbol{\Theta}, \sigma_f, \boldsymbol{\ell}_f)$, is commonly used to perform an optimization before computing the posterior [24]. It is straightforward to compute $\log p(\mathbf{y}|\boldsymbol{\Theta}, \sigma_f, \boldsymbol{\ell}_f)$ and its gradients, so we are free to choose from existing nonlinear optimization methods to perform the optimization.

To select the $(N+1)^{\text{th}}$ policy parameters, we optimize a selection criterion computed on the posterior. A common choice is *expected improvement* (EI) [22, 3], which is defined as the expected value of the improvement, I, over the expected cost of the best policy previously evaluated. Since the predictive distribution under the GP model is Gaussian, the improvement for policy, $\boldsymbol{\theta}_*$, is also Gaussian, $I_* \sim \mathcal{N}(\mu_{\text{best}} - \mathbb{E}[\hat{J}_*], s_*^2)$, where $\mu_{\text{best}} = \min_{i=1,\ldots,N} \mathbb{E}[\hat{J}(\boldsymbol{\theta}_i)]$. Considering only nonnegative improvements, the expected improvement is

$$\begin{aligned} \mathrm{EI}(\boldsymbol{\theta}_*) &= \int_0^\infty I_* p(I_*) dI_* \\ &= s_*(u_* \Phi(u_*) + \phi(u_*)), & (5) \end{aligned}$$

where $u_* = (\mu_{\text{best}} - \mathbb{E}[\hat{J}_*])/s_*$, and $\Phi(\cdot)$ and $\phi(\cdot)$ are the CDF and PDF of the normal distribution, respectively. If $s_* = 0$, the expected improvement is defined to be 0. Both (5) and its gradient, $\partial \mathrm{EI}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$, are efficiently computable, so we can apply standard nonlinear optimization methods to maximize EI to select the next policy. In practice, a parameter ξ is often used to adjust the balance of exploration and exploitation, $u_* = (\mu_{\text{best}} - \mathbb{E}[\hat{J}_*] + \xi)/s_*$, where $\xi > 0$ leads to an optimistic estimate of improvement and tends to encourage exploration of regions of high uncertainty. Cost scale invariance can be achieved by multiplying ξ by the signal standard deviation, σ_f [18].

B. Variational Heteroscedastic Gaussian Process Regression

One limitation of the standard regression model is the assumption of i.i.d. noise over the input space (see equation (2)). Many data do not adhere to this simplification and models capable of capturing heteroscedasticity are required. The heteroscedastic regression model takes the form

$$\hat{J}(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \varepsilon(\boldsymbol{\theta}), \quad \varepsilon(\boldsymbol{\theta}) \sim \mathcal{N}(0, r(\boldsymbol{\theta})), \qquad (6)$$

where the noise variance, $r(\boldsymbol{\theta})$, is dependent on the input. In the Bayesian setting, a second GP prior, $g(\boldsymbol{\theta}) \sim \mathcal{GP}(\mu_0, k_g(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, is placed over the unknown log variance function, $g(\boldsymbol{\theta}) \equiv \log r(\boldsymbol{\theta})$ [7, 10, 15]. This heteroscedastic Gaussian process (HGP) model has the unfortunate property that the computations of the posterior distribution and the marginal likelihood are intractable, thus making hyperparameter optimization and prediction difficult.

In the variational heteroscedastic Gaussian process (VHGP) model [15], a variational lower bound on the marginal likelihood of the HGP model serves as a tractable surrogate function for optimizing the hyperparameters. Let $\mathbf{g} = [g(\boldsymbol{\theta}_1), g(\boldsymbol{\theta}_2), \ldots, g(\boldsymbol{\theta}_N)]^\top$ be the vector of latent log noise variances for the N data points. By defining a normal variational probability density, $q(\mathbf{g}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the marginal variational bound can be derived [15],

$$\begin{aligned} F(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_f + \mathbf{R}) - \tfrac{1}{4}\mathrm{tr}(\boldsymbol{\Sigma}) \\ &\quad - \mathrm{KL}(\mathcal{N}(\mathbf{g}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) || \mathcal{N}(\mathbf{g}|\mu_0 \mathbf{1}, \mathbf{K}_g)), & (7) \end{aligned}$$

where $\mathbf{R}$ is a diagonal matrix with elements $[\mathbf{R}]_{ii} = e^{[\boldsymbol{\mu}]_i - [\boldsymbol{\Sigma}]_{ii}/2}$. Intuitively, by maximizing equation (7) with respect to the variational parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we maximize the log marginal likelihood under the variational approximation while minimizing the distance (in the Kullback-Leibler sense) between the variational distribution and the distribution implied by the GP prior. By exploiting properties of $F(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ at its maximum, one can write $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in terms of N variational parameters [15]:

$$\boldsymbol{\mu} = \mathbf{K}_g(\boldsymbol{\Lambda} - \frac{1}{2}\mathbf{I})\mathbf{1} + \mu_0 \mathbf{1}, \qquad \boldsymbol{\Sigma}^{-1} = \mathbf{K}_g^{-1} + \boldsymbol{\Lambda},$$

where $\boldsymbol{\Lambda}$ is a positive semidefinite diagonal matrix of variational parameters. $F(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be simultaneously maximized with respect to the variational parameters and the HGP model hyperparameters, Ψ_f and Ψ_g. If the kernel functions $k_f(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and $k_g(\boldsymbol{\theta}, \boldsymbol{\theta}')$ are squared exponentials (1), then $\Psi_f = \{\sigma_f, \boldsymbol{\ell}_f\}$ and $\Psi_g = \{\mu_0, \sigma_g, \boldsymbol{\ell}_g\}$.

The VHGP model yields a non-Gaussian variational predictive density,

$$q(\hat{J}_*) = \int \mathcal{N}(\hat{J}_*|a_*, c_*^2 + e^{g_*})\mathcal{N}(g_*|\mu_*, \sigma_*^2)dg_*, \tag{8}$$

where

$$\begin{aligned}
a_* &= \mathbf{k}_{f*}^\top(\mathbf{K}_f + \mathbf{R})^{-1}\mathbf{y}, \\
c_*^2 &= k_f(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) - \mathbf{k}_{f*}^\top(\mathbf{K}_f + \mathbf{R})^{-1}\mathbf{k}_{f*}, \\
\mu_* &= \mathbf{k}_{g*}^\top(\mathbf{\Lambda} - \frac{1}{2}\mathbf{I})\mathbf{1} + \mu_0, \\
\sigma_*^2 &= k_g(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) - \mathbf{k}_{g*}^\top(\mathbf{K}_g + \mathbf{\Lambda}^{-1})^{-1}\mathbf{k}_{g*}.
\end{aligned}$$

Although this predictive density is still intractable, its mean and variance can be calculated in closed form [15]:

$$\begin{aligned}
\mathbb{E}_q[\hat{J}_*] &= a_*, \qquad (9) \\
\mathbb{V}_q[\hat{J}_*] &= c_*^2 + \exp(\mu_* + \sigma_*^2/2) \equiv s_*^2. \qquad (10)
\end{aligned}$$

III. Variational Bayesian Optimization

There are two practical motivations for capturing policy-dependent variance structure during optimization. First, metrics computed on the predictive distribution, such as EI and probability of improvement, will return more meaningful values for the system under consideration. Second, it creates the opportunity to employ policy selection criteria that take cost variance into account, i.e., that are risk-sensitive.

We extend the VHGP model to the optimization case by deriving the expression for expected improvement and its gradient and show that both can be efficiently approximated to several decimal places using Gauss-Hermite quadrature (as is the case for the predictive distribution itself [15]). We show how efficiently computable confidence bound selection criteria can be used to select risk-sensitive policies and generalize the expected improvement criterion. To address numerical issues that arise when N is small (i.e. in the early stages of optimization), we incorporate independent log priors into the marginal variational bound and identify heuristic sampling strategies that perform well empirically.

A. Expected Improvement

The improvement, I_*, of a policy, $\boldsymbol{\theta}_*$, under the variational predictive distribution (8) is

$$q(I_*) = \int \mathcal{N}(I_*|\mu_{\text{best}} - a_*, v_*^2)\mathcal{N}(g_*|\mu_*, \sigma_*^2)dg_*, \tag{11}$$

where $v_*^2 = c_*^2 + e^{g_*}$. The expression for EI then becomes

$$\begin{aligned}
\text{EI}(\boldsymbol{\theta}_*) &= \int_0^\infty I_* q(I_*)dI_*, \\
&= \int v_*\left[u_*\Phi(u_*) + \phi(u_*)\right]\mathcal{N}(g_*|\mu_*, \sigma_*^2)dg_*, \qquad (12)
\end{aligned}$$

where $u_* = (\mu_{\text{best}} - a_*)/v_*$. Although this expression is not analytically tractable, it can be efficiently approximated using Gauss-Hermite quadrature. This can be made clear by setting $w = (g_* - \mu_*)/\sqrt{2}\sigma_*$ and replacing all occurrences of g_* in the expressions for v_* and u_*,

$$\begin{aligned}
\text{EI}(\boldsymbol{\theta}_*) &= \int e^{-w^2}\frac{v_*}{\sqrt{2\pi}\sigma_*}\left[u_*\Phi(u_*) + \phi(u_*)\right]dw, \\
&\equiv \int e^{-w^2}h(w)dw. \qquad (13)
\end{aligned}$$

Similarly, the gradient $\partial\text{EI}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ can be computed under the integral (12) and the result is of the desired form:

$$\frac{\partial\text{EI}(\boldsymbol{\theta}_*)}{\partial\boldsymbol{\theta}} = \int e^{-w^2}z(w)dw, \tag{14}$$

where

$$\begin{aligned}
z(w) &= \frac{1}{\sqrt{2\pi}\sigma_*}\Bigg[\frac{1}{\sigma_*}v_*\left(u_*\Phi(u_*) + \phi(u_*)\right) \\
&\times \left(-\frac{\partial\sigma_*}{\partial\boldsymbol{\theta}} + 2w^2\frac{\partial\sigma_*}{\partial\boldsymbol{\theta}} + \sqrt{2}w\frac{\partial\mu_*}{\partial\boldsymbol{\theta}}\right) \\
&+ \frac{\partial v_*}{\partial\boldsymbol{\theta}}(u_*\Phi(u_*) + \phi(u_*)) + v_*\frac{\partial u_*}{\partial\boldsymbol{\theta}}\Phi(u_*)\Bigg].
\end{aligned}$$

As in the standard Bayesian optimization setting, we can easily incorporate an exploration parameter, ξ, by setting $u_* = (\mu_{\text{best}} - a_* + \xi)/v_*$. EI can be maximized using standard nonlinear optimization algorithms. Since flat regions and multiple local maxima may be present, it is common practice to perform random restarts during EI optimization to avoid low-quality solutions. In our experiments, we use the NLOPT [8] implementation of sequential quadratic programming (SQP) with 25 random restarts to optimize EI.

B. Confidence Bound Selection

In order to exploit cost variance information for policy selection, we must consider selection criteria that flexibly take cost variance into account. Although EI performs well during learning by balancing exploration and exploitation, it falls short in this regard since it always favors high variance or high uncertainty among solutions with equivalent expected cost. In contrast, *confidence bound* (CB) selection criteria allow one to directly specify the sensitivity to cost variance.

The family of confidence bound selection criteria have the general form

$$\text{CB}(\boldsymbol{\theta}_*, \kappa) = \mathbb{E}[\hat{\mathrm{J}}_*] + \mathrm{b}(\mathbb{V}[\hat{\mathrm{J}}_*], \kappa), \tag{15}$$

where $b(\cdot, \cdot)$ is a function of the cost variance and a constant κ that controls the system's sensitivity to risk. Such criteria have been extensively studied in the context of statistical global optimization [5, 26] and economic decision making [16]. Favorable regret bounds for sampling with CB criteria of the form $b(\mathbb{V}[J_*], \kappa) = \kappa\sqrt{\mathbb{V}[J_*]} \equiv \kappa s_*$ have also been derived for certain types of Bayesian optimization problems [26].

Interestingly, CB criteria have a strong connection to the exponential utility functions of risk-sensitive optimal control (RSOC) [32, 31]. Consider the standard RSOC objective function,

$$\gamma(\kappa) = -2\kappa^{-1}\log\mathbb{E}[e^{-\frac{1}{2}\kappa\hat{J}_*}]. \tag{16}$$

Taking the second order Taylor expansion of $e^{-\frac{1}{2}\kappa\hat{J}_*}$ about $\mathbb{E}[\hat{J}_*]$ yields

$$\gamma(\kappa) \approx \mathbb{E}[\hat{J}_*] - \frac{1}{4}\kappa\mathbb{V}[\hat{J}_*]. \tag{17}$$

This approximation exposes the role of the parameter κ in determining the risk sensitivity of the system: $\kappa < 0$ is *risk-averse*, $\kappa > 0$ is *risk-seeking*, and $\kappa = 0$ is *risk-neutral* [31]. Thus, policies selected according to a CB criterion with $b(\mathbb{V}[\hat{J}_*], \kappa) = -\frac{1}{4}\kappa\mathbb{V}[\hat{J}_*]$ can be viewed as approximate RSOC solutions. Furthermore, since the selection is performed with resect to the predictive distribution (8), policies with different risk characteristics can be selected *on-the-fly*, without having to perform separate optimizations that require additional policy executions on the system.

We can also apply confidence bound criteria to generalize EI to an *expected risk improvement* (ERI) criterion. We define risk improvement for a policy $\boldsymbol{\theta}_*$ as $I_*^\kappa = \mu_{\text{best}} + \kappa s_{\text{best}} - \hat{J}_* - \kappa s_*$, where μ_{best} and s_{best} are found by minimizing $\mathbb{E}_q[\hat{J}_{\boldsymbol{\theta}_i}] + \kappa\sqrt{\mathbb{V}_q[\hat{J}_{\boldsymbol{\theta}_i}]}$ over all previously evaluated policies. Thus, ERI can be viewed as a generalization of EI where $u_* = (\mu_{\text{best}} - a_* + \kappa(s_{\text{best}} - s_*))/v_*$ and $\text{ERI} = \text{EI}$ if $\kappa = 0$.

C. Coping with Small Sample Sizes

1) Log Hyperpriors: To avoid numerical instability in the hyperparameter optimization when N is small, we augment $F(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with independent log-normal priors [18] for each hyperparameter in the VHGP model,

$$\hat{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = F(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \sum_{\psi_k \in \Psi} \log \mathcal{N}(\log \psi_k | \mu_k, \sigma_k^2), \tag{18}$$

where $\Psi = \Psi_f \cup \Psi_g$ is the set of all hyperparameters. In practice, these priors can be quite vague and thus do not require significant experimenter insight. For example, in our experiments we set the log prior on length scales so that the width of the 95% confidence region is at least 20 times the actual policy parameter range.

As is the case with standard marginal likelihood maximization, $\hat{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ may have several local optima. In practice, performing random restarts helps avoid low-quality solutions (especially when N is small). In our experiments, we perform 10 random restarts using SQP for policy selection.

2) Sampling: It is well known [9] that selecting policies based on distributions fit using very little data can lead to myopic sampling and premature convergence. Incorporating external randomization is one way to help alleviate this problem. For example, it is common to obtain a random sample of N_0 initial policies prior to performing optimization. We have found that sampling according to EI with probability $1-\epsilon$ and randomly otherwise performs well empirically. In the standard Bayesian optimization setting with model selection, ϵ-random EI selection has been shown to yield near-optimal global convergence rates [4]. Randomized CB selection with, e.g., $\kappa \sim \mathcal{N}(0, 1)$ can also be applied when the policy search is aimed at identifying a spectrum of policies with different risk sensitivities. However, since this technique relies completely on the estimated cost distribution, it is most appropriate to apply after a reasonable initial estimate of the cost distribution has been obtained.

The Variational Bayesian Optimization (VBO) algorithm is outlined in Box 1.

Algorithm 1 Variational Bayesian Optimization

```
Input: Previous evaluations: $\boldsymbol{\Theta}, \mathbf{y}$, Iterations: $n$
1) for $i := 1 : n$
    a) Maximize equation (18) given $\boldsymbol{\Theta}, \mathbf{y}$
       $\Psi_f^+, \Psi_g^+, \boldsymbol{\Lambda}^+ := \arg\max \hat{F}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
    b) Optimize selection criterion, EI, ERI, or CB, w.r.t. optimized model
       $\boldsymbol{\theta}' := \arg\max_{\boldsymbol{\theta}} S(\boldsymbol{\theta}, \Psi_f^+, \Psi_g^+, \boldsymbol{\Lambda}^+)$
    c) Execute $\boldsymbol{\theta}'$, observe cost, $\hat{J}(\boldsymbol{\theta}')$
    d) Append $\boldsymbol{\Theta} := [\boldsymbol{\Theta}; \boldsymbol{\theta}']$, $\mathbf{y} := [\mathbf{y}; \hat{J}(\boldsymbol{\theta}')]$
2) Return $\boldsymbol{\Theta}, \mathbf{y}$
```

IV. Experiments

A. Synthetic Data

As an illustrative example, in Figure 1 we compare the performance of the VBO to standard Bayesian optimization in a simple 1-dimensional noisy optimization task. For this task, the true underlying cost distribution (Figure 1(a)) has two global minima with different cost variances. Both algorithms begin with the same $N_0 = 10$ random samples and perform 10 iterations of EI selection ($\xi = 1.0$, $\epsilon = 0.25$). In Figure 1(b), we see that Bayesian optimization succeeds in identifying the regions of low cost, but it cannot capture the policy-dependent variance characteristics.

In contrast, VBO reliably identifies the minima *and* approximates the local variance characteristics. Figure 1(d) shows the result of applying two different confidence bound selection criteria to vary risk sensitivity. Here we maximized $-\text{CB}(\boldsymbol{\theta}_*, \kappa)$, where

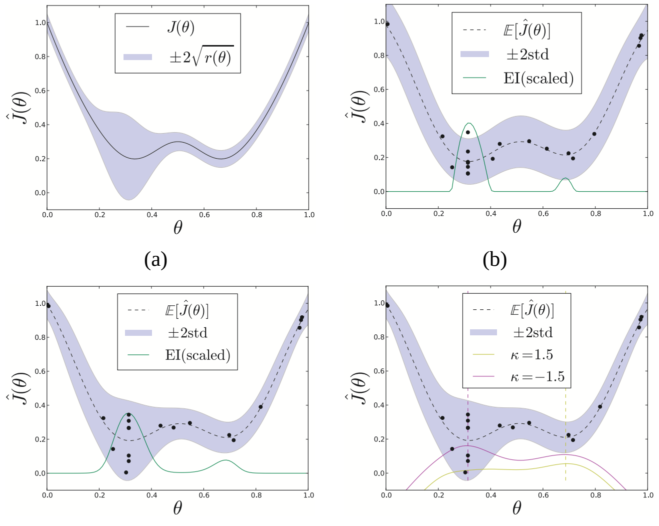
$$\text{CB}(\boldsymbol{\theta}_*, \kappa) = \mathbb{E}_q[\hat{J}_*] + \kappa s_* \tag{19}$$

and $\kappa = -1.5$ and $\kappa = 1.5$ were used to select a risk-seeking and risk-averse policy parameters, respectively.

B. Noisy Pendulum

As another simple example, we considered a swing-up task for a noisy pendulum system. In this task, the maximum torque output of the pendulum actuator is unknown and is drawn from a normal distribution at the beginning of each episode. As a rough physical analogy, this might be understood as fluctuations in motor performance that are caused by unmeasured changes in temperature. The policy space consisted of “bang-bang” policies in which the maximum torque is applied in the positive or negative direction, with switching times specified by two parameters, $0 \leq t_1, t_2 \leq 1.5$ sec. Thus, $\boldsymbol{\theta} = [t_1, t_2]$. The cost function was defined as

$$J(\boldsymbol{\theta}) = \int_0^T 0.01\alpha(t) + 0.0001u(t)^2 dt, \tag{20}$$



(a) (b)

(c) (d)

Fig. 1. (a) A noisy cost function sampled during 10 iterations ($N_0 = 10$) of (b) Bayesian optimization and (c) the VBO algorithm. Bayesian optimization succeeded in identifying the minima, but it cannot distinguish between high and low variance solutions. (d) Confidence bound selection criteria are used to select risk-seeking and risk-averse policy parameter values.

where $0 \leq \alpha(t) \leq \pi$ is the pendulum angle measured from upright vertical, $T = 3.5$ sec, and $u(t) = \tau_{\max}$ if $0 \leq t \leq t_1$, $u(t) = -\tau_{\max}$ if $t_1 < t \leq t_1 + t_2$, and $u(t) = \tau_{\max}$ if $t_1 + t_2 < t \leq T$. The system always started in the downward vertical position with zero initial velocity and the episode terminated if the pendulum came within 0.1 radians of the upright vertical position. The parameters of the system were $l = 1.0$ m, $m = 1.0$ kg, and $\tau_{\max} \sim \mathcal{N}(4, 0.3^2)$ Nm. With these physical parameters, the pendulum must (with probability ≈ 1.0) perform at least two swings to reach vertical in less than T seconds.

The cost function (20) suggests that policies that reach vertical as quickly as possible (i.e., using the fewest swings) are preferred. However, the success of an aggressive policy depends on the torque generating capability of the pendulum. With a noisy actuator, we expect aggressive policies to have higher variance. An approximation of the cost distribution obtained via discretization ($N = 40000$) is shown in Figure 2. Here we indeed see that regions around policies that attempt two-swing solutions ($\boldsymbol{\theta} = [0.0, 1.0]$, $\boldsymbol{\theta} = [1.0, 1.5]$) have low expected cost, but high cost variance.

Figure 3 shows the results of 25 iterations of VBO using EI selection ($N_0 = 15, \xi = 1.0, \epsilon = 0.2$) in the noisy pendulum task. After $N = 40$ total evaluations, the expected cost and cost variance are sensibly represented in regions of low cost. Figure 4 illustrates two policies selected by minimizing the CB criterion (19) on the learned distribution with $\kappa = \pm 2.0$. The risk-seeking policy ($\boldsymbol{\theta} = [1.03, 1.5]$) makes a large initial swing, attempting to reach the vertical position in two swings. The risk-averse policy ($\boldsymbol{\theta} = [0.63, 1.14]$) always produces three swings and exhibits low cost variance, though it has higher cost than the risk-seeking policy when the maximum torque is large.

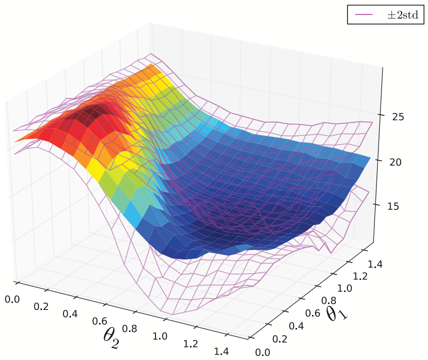


Fig. 2. The cost distribution for the simulated noisy pendulum system obtained by a 20x20 discretization of the policy space. Each policy was evaluated 100 times to estimate its mean and variance ($N = 40000$).

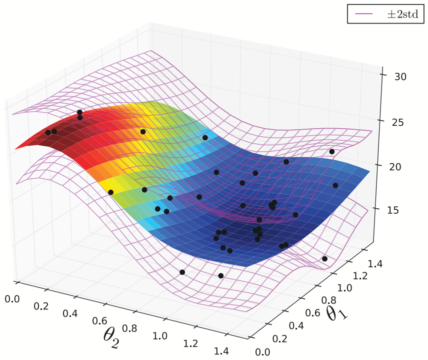


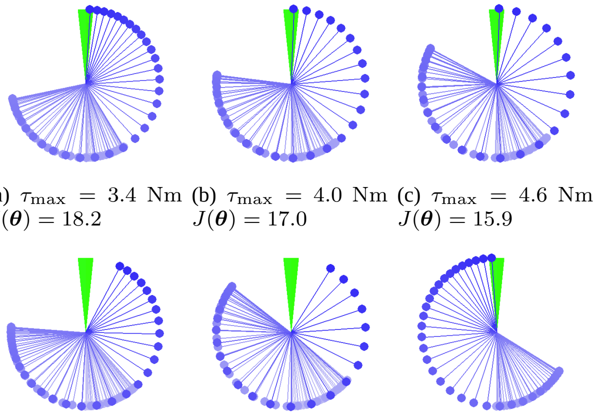
Fig. 3. Estimated cost distribution after 25 iterations of VBO ($N = 40$). The optimization algorithm focuses modeling effort to regions of low cost.

It is often easy to understand the utility of risk-averse and risk-neutral policies, but the motivation for selecting risk-seeking policies may be less clear. The above result suggests one possibility: the acquisition of specialized, high-performance policies. For example, in some cases risk-seeking policies could be chosen in an attempt to identify observable initial conditions that lead to rare low-cost events. Subsequent optimizations might then be performed to direct the system to these initial conditions. One could also imagine situations where the context might demand performance that lower risk policies are very unlikely to generate. For example, if the minimum time to goal was reduced so that only two swing policies had a reasonable chance of succeeding. In such instances it may be desirable to select higher risk policies, even if the probability of succeeding is quite low.

C. Balance Recovery with the uBot-5

The uBot-5 is an 11-DoF mobile manipulator that has two 4-DoF arms, a rotating trunk, and two wheels in a differential drive configuration. The robot has a mass of 19 kg and stands 60 cm from the ground with a torso that is roughly similar to a small adult human in scale and geometry (Figure 5). The robot balances using a linear-quadratic regulator (LQR) with feedback from an onboard inertial measurement unit (IMU).

In our previous experiments [13], the energetic and stabilizing effects of rapid arm motions on the LQR stabilized system were evaluated in the context of recovery from impact



(a) $\tau_{\max} = 3.4$ Nm $J(\boldsymbol{\theta}) = 18.2$ (b) $\tau_{\max} = 4.0$ Nm $J(\boldsymbol{\theta}) = 17.0$ (c) $\tau_{\max} = 4.6$ Nm $J(\boldsymbol{\theta}) = 15.9$

(d) $\tau_{\max} = 3.4$ Nm $J(\boldsymbol{\theta}) = 19.9$ (e) $\tau_{\max} = 4.0$ Nm $J(\boldsymbol{\theta}) = 17.7$ (f) $\tau_{\max} = 4.6$ Nm $J(\boldsymbol{\theta}) = 13.0$

Fig. 4. Performance of risk-averse (a)-(c) and risk-seeking (d)-(f) policies as the maximum pendulum torque is varied.

perturbations. One observation we made was that high energy impacts caused a subset of possible recovery policies to have high cost variance: successfully stabilizing in some trials, while failing to stabilize in others. We extend these experiments by considering larger impact perturbations, increasing the set of arm initial conditions, defining a policy space that permits more flexible, asymmetric arm motions.

The robot was placed in a balancing configuration with its upper torso aligned with a 3.3 kg mass suspended from the ceiling (Figure 5). The mass was pulled away from the robot to a fixed angle and released, producing a controlled impact between the swinging mass and the robot. The pendulum momentum prior to impact was 9.9 ± 0.8 Ns and the resulting impact force was approximately equal to the robot's total mass in earth gravity. The robot was consistently unable to recover from this perturbation using only the wheel LQR (see the rightmost column of Figure 6).

This problem is well suited for model-free policy optimization since there are several physical properties, such as joint friction, wheel backlash, and tire slippage, that make the system difficult to model accurately. In addition, although the underlying state and action spaces are high dimensional (22 and 8, respectively), low-dimensional policy spaces that contain high-quality solutions are relatively straightforward to identify.

The parameterized policy controlled each arm joint according to an exponential trajectory, $\tau_i(t) = e^{-\lambda_i t}$, where $0 \leq \tau_i(t) \leq 1$ is the commanded DC motor power for joint i at time t. The λ parameters were paired for the shoulder/elbow pitch and the shoulder roll/yaw joints. This pairing allowed the magnitude of dorsal and lateral arm motions to be independently specified. We commanded the pitch (dorsal) motions separately for each arm and mirrored the lateral motions, which reduced the number of policy parameters to 3. The range of each λ_i was constrained: $1 \leq \lambda_i \leq 15$. At time t, if $\forall_i \; \tau_i(t) < 0.25$, the arms were retracted to a nominal configuration (the mean of the initial configurations) using a fixed, low-gain linear position controller.

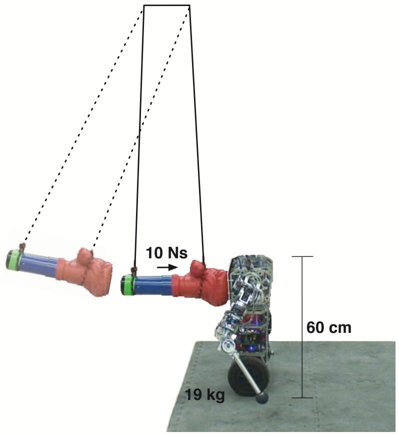


Fig. 5. The uBot-5 situated in the impact pendulum apparatus.

The cost function was designed to encourage energy efficient solutions that successfully stabilized the system,

$$J(\boldsymbol{\theta}) = h(\mathbf{x}(T)) + \int_0^T \frac{1}{10} I(t)V(t)dt, \tag{21}$$

where $I(t)$ and $V(t)$ are the total absolute motor current and voltage at time t, respectively, $T = 3.5$ sec, and $h(\mathbf{x}(T)) = 5$ if $\mathbf{x}(T) \in \mathit{FailureStates}$, otherwise $h(\mathbf{x}(T)) = 0$.

After 15 random initial trials, we applied VBO with EI selection ($\xi = 1.0, \epsilon = 0.2$) for 15 episodes and randomized CB selection ($\kappa \sim \mathcal{N}(0, 1)$) for 15 episodes resulting in a total of $N = 45$ policy evaluations. Since the left and right pitch parameters are symmetric with respect to cost, we imposed an arbitrary ordering constraint, $\lambda_{\text{left}} \geq \lambda_{\text{right}}$, during policy selection.

After training, we evaluated four policies with different risk sensitivity selected by minimizing the CB criterion (19) with $\kappa = 2$, $\kappa = 0$, $\kappa = -1.5$, and $\kappa = -2$. Each selected policy was evaluated 10 times and the results are shown in Figure 6. The sample statistics confirm the algorithmic predictions about the relative riskiness of each policy. In this case, the risk-averse and risk-neutral policies were very similar (no statistically significant difference between the mean or variance), while the two risk-seeking policies had higher variance (for $\kappa = -2$, the differences in both the sample mean and variance were statistically significant).

For $\kappa = -2$, the selected policy produced an upward laterally-directed arm motion that failed approximately 50% of the time. In this case, the standard deviation of cost was sufficiently large that the second term in equation (19) dominated, producing a policy with high variance and poor average performance. A slightly less risk-seeking selection ($\kappa = -1.5$) yielded a policy with conservative low-energy arm movements that was more sensitive to initial conditions than the lower risk policies. This exertion of minimal effort could be viewed as a kind of gamble on initial conditions. Figure 7 shows two successful trials executing risk-averse and risk-seeking policies.

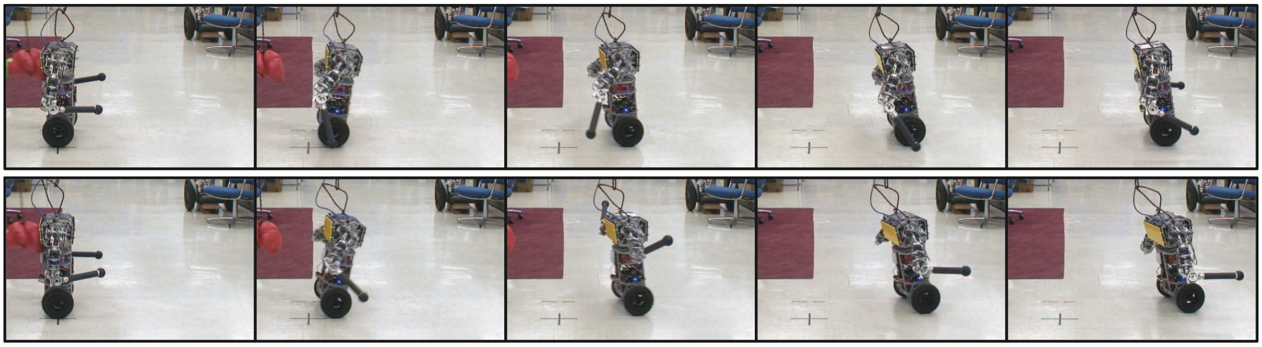


Fig. 7. Time series (duration: 1 second) showing two successful trials executing low-risk (top, $\kappa = 2$) and high-risk (bottom, $\kappa = -2$) policies selected using confidence bound criteria on the learned cost distribution. The low-risk policy produced an asymmetric dorsally-directed arm motion with reliable recovery performance. The high-risk policy produced an upward laterally-directed arm motion that failed approximately 50% of the time.

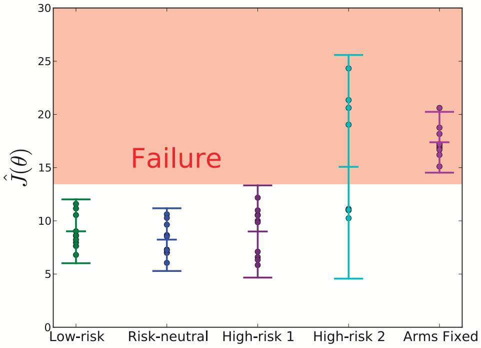


Fig. 6. Data collected over 10 trials using policies identified as risk-averse, risk-neutral, and risk-seeking after performing VBO. The policies were selected using confidence bound criteria with $\kappa = 2$, $\kappa = 0$, $\kappa = -1.5$, and $\kappa = -2$, from left to right. The sample means and two times sample standard deviations are shown. The shaded region on the top part of the plot contains all trials that resulted in failure to stabilize. Ten trials with a fixed-arm policy are plotted on the far right to serve as a baseline level of performance for this impact magnitude.

V. Discussion and Future Work

In many systems, it may be advantageous to adjust risk sensitivity based on runtime context. For example, systems whose environments change in ways that make failures more or less costly (such as operating around catastrophic obstacles or in a safety harness) or when the context demands that the system seek out a low-probability high-performance event. Perhaps not surprisingly, this variable risk property has been observed in a variety of animal species, from simple motor tasks in humans to foraging birds and bees [2, 1].

However, most methods for learning policies by interaction focus on the risk-neutral minimization of expected cost. Extending Bayesian optimization methods to capture policy-dependent cost variance creates the opportunity to select policies with different risk sensitivity. Furthermore, the ability to change risk sensitivity at runtime offers an advantage over existing risk-sensitive control techniques, e.g., [21, 30], that require separate optimizations and policy executions to produce policies with different risk.

There are several properties of VBO that should be considered when determining its suitability for a particular problem. First, although the computational complexity is the same as Bayesian optimization, $\mathcal{O}(N^3)$, the greater flexibility of the VHGP model means that VBO tends to require more initial policy evaluations than standard Bayesian optimization. In addition, many model-free policy search algorithms, such as Bayesian optimization, VBO, and stochastic gradient descent [25], are sensitive to the number of policy parameters—high-dimensional policies can require many trials to optimize. Thus, these algorithms are most effective in problems where low-dimensional policy spaces are available, but accurate system models are not. However, there is evidence policy spaces at least up to 15 dimensions can be efficiently explored with Bayesian optimization if estimates of the GP hyperparameters can be obtained *a priori* [17].

In contrast to local methods, such as policy gradient, Bayesian optimization and VBO can produce large changes in policy parameters between episodes, which could be undesirable in some situations. One approach to alleviating this potential problem is to combine VBO with local gradient methods. For example, one could imagine collecting data by performing gradient descent, rather than randomly selecting policies initially. In this case, both the samples obtained and the gradient estimates could be used to constrain the posterior cost distribution. In turn, the learned local cost distribution could act as a critic structure to reduce the variance of the policy update. Local offline optimization could be interweaved with the local policy updates to select greedy policies or change risk sensitivity using CB criteria. Some of these ideas have been explored in our recent work [14].

Another important consideration is the choice of kernel functions in the GP priors. In this work, we used the anisotropic squared exponential kernel to encode our prior assumptions regarding the smoothness and regularity of the underlying cost function. However, for many problems the underlying cost function is not smooth or regular; it contains flat regions and sharp discontinuities that can be difficult to represent. An interesting direction for future work is the use kernel functions with *local support*, i.e. kernels that are not invariant to shifts in policy space [24].

VI. Conclusion

Varying risk sensitivity based on runtime context is a potentially powerful way to generate flexible control in robot systems. We considered this problem in the context of model-free policy search, where risk-sensitive policies can be selected based on an efficiently learned cost distribution. Our experimental results suggest that VBO is an efficient and plausible method for achieving risk-sensitive control.

Acknowledgments

The authors would like to thank George Konidaris and the anonymous reviewers for their helpful comments. Scott Kuindersma was supported by a NASA GSRP Fellowship from Johnson Space Center. Roderic Grupen was supported by the ONR MURI award N00014-07-1-0749. Andrew Barto was supported by the AFOSR under grant FA9550-08-1-0418.

References

[1] Melissa Bateson. Recent advances in our understanding of risk-sensitive foraging preferences. *Proceedings of the Nutrition Society*, 61:1–8, 2002.

[2] Daniel A. Braun, Arne J. Nagengast, and Daniel M. Wolpert. Risk-sensitivity in sensorimotor control. *Frontiers in Human Neuroscience*, 5:1–10, January 2011.

[3] Eric Brochu, Vlad Cora, and Nando de Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical Report TR-2009-023, University of British Columbia, Department of Computer Science, 2009.

[4] Adam D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12: 2879–2904, October 2011.

[5] Dennis D. Cox and Susan John. A statistical method for global optimization. In *Systems, Man and Cybernetics, 1992., IEEE International Conference on*, volume 2, pages 1241–1246, 1992. doi: 10.1109/ICSMC.1992.271617.

[6] Marcus Frean and Phillip Boyle. Using Gaussian processes to optimize expensive functions. In *AI 2008: Advances in Artificial Intelligence*, pages 258–267, 2008.

[7] Paul W. Goldberg, Christopher K. I. Williams, and Christopher M. Bishop. Regression with input-dependent noise: A Gaussian process treatment. In *Advances in Neural Information Processing Systems 10 (NIPS)*, pages 493–499, 1998.

[8] Steven G. Johnson. The NLopt nonlinear-optimization package. http://ab-initio.mit.edu/nlopt, 2011.

[9] Donald R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21:345–383, 2001.

[10] Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic Gaussian process regression. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 393–400, 2010.

[11] Jens Kober and Jan Peters. Policy search for motor primitives in robotics. In *Advances in Neural Information Processing Systems 21*. MIT Press, 2009.

[12] J. Zico Kolter and Andrew Y. Ng. Policy search via the signed derivative. In *Robotics: Science and Systems V (RSS)*, 2010.

[13] Scott Kuindersma, Roderic Grupen, and Andrew Barto. Learning dynamic arm motions for postural recovery. In *Proceedings of the 11th IEEE-RAS International Conference on Humanoid Robots*, pages 7–12, Bled, Slovenia, October 2011.

[14] Scott Kuindersma, Roderic Grupen, and Andrew Barto. Variable risk dynamic mobile manipulation. In *RSS 2012 Workshop on Mobile Manipulation*, Sydney, Australia, 2012.

[15] Miguel Lázaro-Gredilla and Michalis K. Titsias. Variational heteroscedastic Gaussian process regression. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[16] H. Levy and H. M. Markowitz. Approximating expected utility by a function of mean and variance. *The American Economic Review*, 69(3):308–317, June 1979.

[17] Daniel Lizotte, Tao Wang, Michael Bowling, and Dale Schuurmans. Automatic gait optimization with Gaussian process regression. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[18] Daniel James Lizotte. *Practical Bayesian Optimization*. PhD thesis, University of Alberta, Edmonton, Alberta, 2008.

[19] Ruben Martinez-Cantin, Nando de Freitas, Arnaud Doucet, and José A. Castellanos. Active policy learning for robot planning and exploration under uncertainty. In *Proceedings of Robotics: Science and Systems*, 2007.

[20] Ruben Martinez-Cantin, Nando de Freitas, Eric Brochu, José A. Castellanos, and Arnaud Doucet. A Bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot. *Autonomous Robots*, 27: 93–103, 2009.

[21] Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49:267–290, 2002.

[22] J. Močkus, V. Tiesis, and A. Žilinskas. The application of Bayesian methods for seeking the extremum. In *Toward Global Optimization*, volume 2, pages 117–128. Elsevier, 1978.

[23] Jan Peters and Stefan Schaal. Policy gradient methods for robotics. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 2219–2225, 2006.

[24] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[25] John W. Roberts and Russ Tedrake. Signal-to-noise ratio analysis of policy gradient algorithms. In *Advances of Neural Information Processing Systems 21 (NIPS)*, 2009.

[26] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

[27] Russ Tedrake, Teresa Weirui Zhang, and H. Sebastian Seung. Stochastic policy gradient reinforcement learning on a simple 3D biped. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 2849–2854, Sendai, Japan, September 2004.

[28] Matthew Tesch, Jeff Schneider, and Howie Choset. Using response surfaces and expected improvement to optimize snake robot gait parameters. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Francisco, CA, 2011.

[29] Evangelos Theodorou, Jonas Buchli, and Stefan Schaal. Reinforcement learning of motor skills in high dimensions: A path integral approach. In *Proceedings of the IEEE International Conference on Robotics and Automation*, Anchorage, Alaska, May 2010.

[30] Bart van den Broek, Wim Wiegerinck, and Bert Kappen. Risk sensitive path integral control. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 615–622, 2010.

[31] Peter Whittle. Risk-sensitive linear/quadratic/Gaussian control. *Advances in Applied Probability*, 13:764–777, 1981.

[32] Peter Whittle. *Risk-Sensitive Optimal Control*. John Wiley & Sons, 1990.

[33] Aaron Wilson, Alan Fern, and Prasad Tadepalli. A behavior based kernel for policy search via Bayesian optimization. In *Proceedings of the ICML 2011 Workshop: Planning and Acting with Uncertain Model*, Bellevue, WA, 2011.