

# Using Simulation to Improve Sample-Efficiency of Bayesian Optimization for Bipedal Robots

**Akshara Rai\***

ARAI@CS.CMU.EDU

*Robotics Institute, School of Computer Science  
Carnegie Mellon University, PA, USA*

**Rika Antonova\***

ANTONOVA@KTH.SE

*Robotics, Perception and Learning, CSC  
KTH Royal Institute of Technology, Stockholm, Sweden*

**Franziska Meier**

FRANZI.MEIER@GMAIL.COM

*Paul G. Allen School of Computer Science & Engineering  
University of Washington, Seattle, WA, USA*

**Christopher G. Atkeson**

CGA@CS.CMU.EDU

*Robotics Institute, School of Computer Science  
Carnegie Mellon University, PA, USA*

\*These authors contributed equally.

**Editor:** Bayesian Optimization Special Issue

## Abstract

Learning for control can acquire controllers for novel robotic tasks, paving the path for autonomous agents. Such controllers can be expert-designed policies, which typically require tuning of parameters for each task scenario. In this context, Bayesian optimization (BO) has emerged as a promising approach for automatically tuning controllers. However, sample-efficiency can still be an issue for high-dimensional policies on hardware. Here, we develop an approach that utilizes simulation to learn structured feature transforms that map the original parameter space into a domain-informed space. During BO, similarity between controllers is now calculated in this transformed space. Experiments on the ATRIAS robot hardware and simulation show that our approach succeeds at sample-efficiently learning controllers for multiple robots. Another question arises: What if the simulation significantly differs from hardware? To answer this, we create increasingly approximate simulators and study the effect of increasing simulation-hardware mismatch on the performance of Bayesian optimization. We also compare our approach to other approaches from literature, and find it to be more reliable, especially in cases of high mismatch. Our experiments show that our approach succeeds across different controller types, bipedal robot models and simulator fidelity levels, making it applicable to a wide range of bipedal locomotion problems.

**Keywords:** Bayesian Optimization, Bipedal Locomotion, Transfer Learning

## 1. Introduction

Machine learning can provide methods for learning controllers for robotic tasks. Yet, even with recent advances in this field, the problem of automatically designing and learning controllers for robots, especially bipedal robots, remains difficult. It is expensive to do learning

experiments that require a large number of samples with physical robots. Specifically, legged robots are not robust to falls and failures, and are time-consuming to work with and repair. Furthermore, commonly used cost functions for optimizing controllers are noisy to evaluate, non-convex and non-differentiable. In order to find learning approaches that can be used on real robots, it is thus important to keep these considerations in mind.

Deep reinforcement learning approaches can deal with noise, discontinuities and non-convexity of the objective, but they are not data-efficient. These approaches could take on the order of a million samples to learn locomotion controllers (Peng et al., 2016), which would be infeasible on a real robot. For example, on the ATRIAS robot, 10,000 samples would take 7 days, in theory. But practically, the robot needs to be “reset” between trials and repaired in case of damage. Using structured expert-designed policies can help minimize damage to the robot and make the search for successful controllers feasible. However, the problem is black-box, non-convex and discontinuous. This eliminates approaches like PI<sup>2</sup> (Theodorou et al., 2010) which make assumptions about the dynamics of the system and PILCO (Deisenroth and Rasmussen, 2011) which assumes a continuous cost landscape. Evolutionary approaches like CMA-ES (Hansen, 2006) can still be prohibitively expensive, needing thousands of samples (Song and Geyer, 2015).

In comparison, Bayesian optimization (BO) is a sample-efficient optimization technique that is robust to non-convexity and noise. It has been recently used in a range of robotics problems, such as Calandra et al. (2016b), Marco et al. (2017) and Cully et al. (2015). However, sample-efficiency of conventional BO degrades in high dimensions, even for dimensionalities commonly encountered in locomotion controllers. Because of this, hardware-only optimization becomes intractable for flexible controllers and complex robots. One way of addressing this issue is to utilize simulation to optimize controller parameters. However, simulation-only optimization is vulnerable to learning policies that exploit the simulation and, because of that, perform well in simulation but poorly on the actual robot. This motivates the development of hybrid approaches that can incorporate simulation-based information into the learning method and then optimize with few samples on hardware.

Towards this goal, our previous work in Antonova, Rai, and Atkeson (2016), Antonova et al. (2017), Rai et al. (2018) presents a framework that uses information from high-fidelity simulators to learn sample-efficiently on hardware. We use simulation to build informed feature transforms that are used to measure controller similarity during BO. Thus, during optimization on hardware, the similarity between controller parameters is informed by how they perform in simulation. With this, it becomes possible to quickly infer which regions of the input space are likely to perform well on hardware. This method has been tested on the ATRIAS biped robot (Figure 1) and shows considerable improvement in sample-efficiency over traditional BO.

In this article, we present in-depth explanations and empirical analysis of our previous work. Furthermore, for the first time, we present a procedure for systematically evaluating



Figure 1: ATRIAS robot.

robustness of such approaches to simulation-hardware mismatch. We extend our previous work incorporating mismatch estimates (Rai et al., 2018) to this setting. We also conduct extensive comparisons with competitive baselines from prior work, e.g. Cully et al. (2015).

The rest of this article is organized as follows: Section 2 provides background for BO, then gives an overview of related work on optimizing locomotion controllers. Section 3.1 describes the idea of incorporating simulation-based transforms into BO; Section 3.2 explains how we handle simulation-hardware mismatch. Sections 4.1-4.5 describe the robot and controllers we use for our experiments; Section 4.6 explains the motivation and construction of simulators with various kinds of simulation-hardware mismatch. Section 5 summarizes hardware experiments on the ATRIAS robot. Section 5.2 shows generalization to a different robot model in simulation. Section 5.3 presents empirical analysis of the impact of simulator quality on the performance of the proposed algorithms and alternative approaches.

## 2. Background and Related Work

This section gives a brief overview of Bayesian optimization (BO), the state-of-the-art research on optimizing locomotion controllers, and utilizing simulation information in BO.

### 2.1. Background on Bayesian Optimization

Bayesian optimization (BO) is a framework for online, black-box, gradient-free global search (Shahriari et al. (2016) and Brochu et al. (2010) provide a comprehensive introduction). The problem of optimizing controllers can be interpreted as finding controller parameters  $\mathbf{x}^*$  that optimize some cost function  $f(\mathbf{x})$ . Here  $\mathbf{x}$  contains parameters of a pre-structured policy; the cost  $f(\mathbf{x})$  is a function of the trajectory induced by controller parameters  $\mathbf{x}$ . For brevity, we will refer to ‘controller parameters  $\mathbf{x}$ ’ as ‘controller  $\mathbf{x}$ ’. We use BO to find controller  $\mathbf{x}^*$ , such that:  $f(\mathbf{x}^*) = \min_{\mathbf{x}} f(\mathbf{x})$ .

BO is initialized with a prior that expresses the *a priori* uncertainty over the value of  $f(\mathbf{x})$  for each  $\mathbf{x}$  in the domain. Then, at each step of optimization, based on data seen so far, BO optimizes an auxiliary function (called *acquisition function*) to select the next  $\mathbf{x}$  to evaluate. The acquisition function balances exploration vs exploitation. It selects points for which the posterior estimate of the objective  $f$  is promising, taking into account both mean and covariance of the posterior. A widely used representation for the cost function  $f$  is a Gaussian process (GP):

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}_i, \mathbf{x}_j))$$

The prior mean function  $\mu(\cdot)$  is set to 0 when no domain-specific knowledge is provided. The kernel function  $k(\cdot, \cdot)$  encodes similarity between inputs. If  $k(\mathbf{x}_i, \mathbf{x}_j)$  is large for inputs  $\mathbf{x}_i, \mathbf{x}_j$ , then  $f(\mathbf{x}_i)$  strongly influences  $f(\mathbf{x}_j)$ . One of the most widely used kernel functions is the Squared Exponential (SE):

$$k_{SE}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_k^2 \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \text{diag}(\boldsymbol{\ell})^{-2}(\mathbf{x}_i - \mathbf{x}_j)\right),$$

where  $\sigma_k^2$ ,  $\boldsymbol{\ell}$  are signal variance and a vector of length scales respectively.  $\sigma_k^2$ ,  $\boldsymbol{\ell}$  are referred to as ‘hyperparameters’ in the literature.

The SE kernel belongs to a broader class of Matérn kernels, which in general have more free parameters. In some cases, carefully choosing kernel parameters improves performance of BO (Snoek et al., 2012). However, domain-informed kernels can easily out-perform even

well-tuned Matérn kernels (Cully et al., 2015). SE and Matérn kernels are stationary:  $k(\mathbf{x}_i, \mathbf{x}_j)$  depend only on  $r = \mathbf{x}_i - \mathbf{x}_j \forall \mathbf{x}_{i,j}$ . We aim to remove this limitation in a manner informed by simulation. In future work, incorporating other approaches that relax stationarity (e.g. Martinez-Cantin (2018)) could further improve BO.

## 2.2. Optimizing Locomotion Controllers

Parametric locomotion controllers can be represented as  $\mathbf{u} = \pi_{\mathbf{x}}(\mathbf{s})$ , where  $\pi$  is a policy structure that depends on parameters  $\mathbf{x}$ . For example,  $\pi$  can be parameterized by feedback gains on the center of mass (CoM), reference joint trajectories, etc. Vector  $\mathbf{s}$  is the state of the robot, such as joint angles and velocities, used in closed-loop controllers. Vector  $\mathbf{u}$  represents the desired control action, for example: torques, angular velocities or positions for each joint on the robot. The sequence of control actions yields a sequence of state transitions, which form the overall ‘trajectory’  $[\mathbf{s}_0, \mathbf{u}_1, \mathbf{s}_1, \mathbf{u}_2, \mathbf{s}_2, \dots]$ . This trajectory is used in the cost function to judge the quality of the controller  $\mathbf{x}$ . In our work, we use structured controllers designed by experts. State of the art research on walking robots featuring such controllers includes Feng et al. (2015) and Kuindersma et al. (2016). The overall optimization then includes manually tuning the parameters  $\mathbf{x}$ . An alternative to manual tuning is to use evolutionary approaches, like CMA-ES, as in Song and Geyer (2015). However, these require a large number of samples and can usually be conducted only in simulation. Optimization in simulation can produce controllers that perform well in simulation, but not on hardware. In comparison, BO is a sample-efficient technique which has become popular for direct optimization on hardware. Recent successes include manipulation (Englert and Toussaint, 2016) and locomotion (Calandra et al., 2016b).

BO for locomotion has been previously explored for several types of mobile robots. These include: snake robots (Tesch et al., 2011), AIBO quadrupeds (Lizotte et al., 2007), and hexapods (Cully et al., 2015). Tesch et al. (2011) optimize a 3-dimensional controller for a snake robot in 10-40 trials (for speeds up to  $0.13m/s$ ). Lizotte et al. (2007) use BO to optimize gait parameters for a AIBO robot in 100-150 trials. Cully et al. (2015) learn 36 controller parameters for a hexapod. Even with hardware damage, they can obtain successful controllers for speeds up to  $0.4m/s$  in 12-15 trials.

Hexapods, quadrupeds and snakes spend a large portion of their gaits with their center of mass within their support polygon (convex hull of the feet on the ground). Ignoring velocity, this is statically stable. In contrast, bipedal walking can be highly dynamic, especially for point-feet robots like ATRIAS. ATRIAS, like most bipeds, spends a significant time of its gait being “unstable”, or dynamically stable. In our experiments on hardware, ATRIAS goes up to speeds of  $1m/s$ . All of this leads to a challenging optimization setting and discontinuous cost function landscape. Calandra et al. (2016b) use BO for optimizing gaits of a dynamic biped on a boom, needing 30-40 samples for finding walking gaits for a 4-dimensional controller. While this is promising, optimizing a higher-dimensional controller needed for complex robots would be even more challenging. If significant number of samples lead to unstable gaits and falls, they could damage the robot. Hence, it is important to develop methods that can learn complex controllers fast, without damaging the robot.

### 2.3. Incorporating Simulation Information into Bayesian Optimization

The idea of using simulation to speed up BO on hardware has been explored before. Information from simulation can be added as a prior to the GP used in BO, such as in Cully et al. (2015). While such methods can be successful, one needs to carefully tune the influence of simulation points over hardware points, especially when simulation is significantly different from hardware.

More generally, approaches such as Kandasamy et al. (2017) and Marco et al. (2017) address trading off computation vs simulation accuracy when selecting the source for the next trial/evaluation (with real world viewed as the most expensive source). Instead, we consider a setting with ample compute resources for simulation, but an extremely small number of experiments on a real robot. This is appropriate for bipedal locomotion with full-scale robots, since these can be expensive to run. Hence, our work does not fall into the ‘multi-fidelity’ BO paradigm, and we instead take a two step approach: pre-computing information needed for kernel transforms in the first stage, then running an ultra-sample efficient version of BO in the second stage on a real robot.

Recently, several approaches proposed incorporating Neural Networks (NNs) into the Gaussian process (GP) kernels (Wilson et al. (2016), Calandra et al. (2016a)). The strength of these approaches is that they can jointly update the GP and the NN. Calandra et al. (2016a) demonstrated how this added flexibility can handle discontinuities in the cost function landscape. However, these approaches do not directly address the problem of incorporating a large amount of data from simulation in hardware BO experiments.

Wilson et al. (2014) explored enhancing GP kernel with trajectories. Their Behavior Based Kernel (BBK) computes an estimate of a symmetric variant of the KL divergence between trajectories induced by two controllers, and uses this as a distance metric in the kernel. However, getting an estimate would require samples for each controller  $\mathbf{x}_i, \mathbf{x}_j$  whenever  $k(\mathbf{x}_i, \mathbf{x}_j)$  is needed. This can be impractical, as it involves an evaluation of every controller considered. To overcome this, the authors suggest combining BBK with a model-based approach. But building an accurate model from hardware data might be an expensive process in itself.

Cully et al. (2015) utilize simulation by defining a behavior metric and collecting best performing points in simulation. This behavior metric then guides BO to quickly find controllers on hardware, and can even compensate for damage to the robot. The search on hardware is conducted in behavior space, and limited to pre-selected “successful” points from simulation. This helps make their search faster and safer on hardware. However, if an optimal point was not pre-selected, BO cannot sample it during optimization.

In our work, we utilize trajectories from simulation to build feature transforms that can be incorporated in the GP kernel used for BO. Our approaches incorporate trajectory/behavior information, but ensure that  $k(\mathbf{x}_i, \mathbf{x}_j)$  is computed efficiently during BO. Our work is related, in part, to input space warping, as described in Snoek et al. (2014), but we construct our transforms from simulated data. Our simulation-informed kernels bias the search towards regions that look promising, but are able to ‘recover’ and search in other parts of the space if simulation-hardware mismatch becomes apparent.

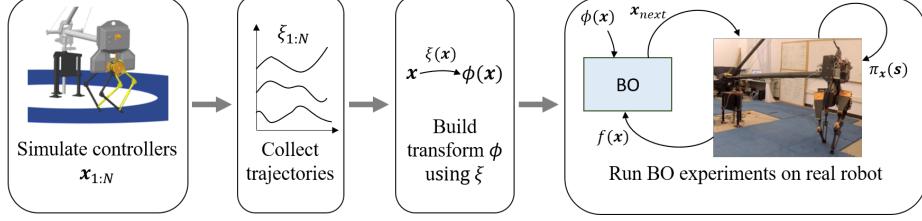


Figure 2: Overview of our proposed approach. Here,  $\pi_{\mathbf{x}}(\mathbf{s})$  is the policy (Section 2.2);  $\mathbf{x}$  is a vector of controller parameters;  $\mathbf{s}$  is the state of the robot;  $\xi(\mathbf{x})$  is a trajectory observed in simulation for  $\mathbf{x}$ ;  $\phi(\cdot)$  is the transform built using  $\xi(\cdot)$ ;  $f(\mathbf{x})$  is the cost of  $\mathbf{x}$  evaluated on hardware. BO uses  $\phi(\mathbf{x})$  and evaluated costs  $f(\mathbf{x})$  to propose next promising controller  $x_{next}$ .

### 3. Proposed Approach: Bayesian Optimization with Informed Kernels

In this section, we offer in-depth explanation of approaches from our work in Antonova, Rai, and Atkeson (2016), Antonova et al. (2017), and Rai et al. (2018). This work proposes incorporating domain knowledge into BO with the help of simulation. We evaluate locomotion controllers in simulation, and collect their induced trajectories, which are then used to build an informed transform. This can be achieved by using a domain-specific feature transform (Section 3.1.1) or by learning to reconstruct short trajectory summaries (Section 3.1.2). This feature transform is used to construct an informed distance metric for BO, and helps BO discover promising regions faster. Figure 2 gives an overview. In Section 3.2 we discuss how to incorporate simulation-hardware mismatch in to the transform, ensuring that BO can benefit from inaccurate simulations as well.

#### 3.1. Constructing Flexible Kernels using Simulation-based Transforms

High dimensional problems with discontinuous cost functions are very common with legged robots, where slight changes to some parameters can make the robot unstable. Both of these factors can adversely affect BO’s performance, but informed feature transforms can help BO sample high-performing controllers even in such scenarios.

In this section, we demonstrate how to construct such transforms  $\phi(\mathbf{x})$  utilizing simulations for a given controller  $\mathbf{x}$ . We then use  $\phi$  to create an informed kernel  $k_{\phi}(\mathbf{x}_i, \mathbf{x}_j)$  for BO on hardware:

$$\begin{aligned} \mathbf{t}_{ij} &= \phi(\mathbf{x}_i) - \phi(\mathbf{x}_j) \\ k_{\phi}(\mathbf{x}_i, \mathbf{x}_j) &= \sigma_k^2 \exp \left( -\frac{1}{2} \mathbf{t}_{ij}^T \text{diag}(\boldsymbol{\ell})^{-2} \mathbf{t}_{ij} \right) \end{aligned} \quad (1)$$

Note that the functional form above is the same as that of Squared Exponential kernel, if considered from the point of view of the transformed space, with  $\phi(\mathbf{x})$  as input. Using a low-dimensional  $\phi$  could improve sample efficiency by reducing dimensionality. More notably, crucial gains arise when  $\phi$  brings controllers that perform similar in simulation closer together, as compared to the original parameter space. For locomotion, this could bring failing controllers close together to occupy only a small portion of the transformed space, as illustrated in Rai et al. (2018). In essence, this means that the resultant kernel, though stationary in  $\phi$ , is non-stationary in  $\mathbf{x}$ .

### 3.1.1. THE DETERMINANTS OF GAIT TRANSFORM

We propose a feature transform for bipedal locomotion derived from physiological features of human walking called Determinants of Gaits (DoG) (Inman et al., 1953).  $\phi_{DoG}$  was originally developed for human-like robots and controllers (Antonova, Rai, and Atkeson, 2016), and then generalized to be applicable to a wider range of bipedal locomotion controllers and robot morphologies (Rai et al., 2018). It is based on the features in Table 1.

---



$M_1$ (Swing leg retraction) – If the maximum ground clearance of the swing foot $x_f$ is more than a threshold, $M_1=1$ (0 otherwise); ensures swing leg retraction.
$M_2$ (Center of mass height) – If CoM height $z$ stays about the same at the start and end of a step, $M_2=1$ (0 otherwise); checks that the robot is not falling.
$M_3$ (Trunk lean) – If the average trunk lean $\theta$ is the same at the start and end of a step, $M_3=1$ (0 otherwise); ensures that the trunk is not changing orientation.
$M_4$ (Average walking speed) – Average forward speed $v$ of a controller per step, $M_4=v_{avg}$ ; $M_4$ helps distinguish controllers that perform similar on $M_{1-3}$ .

---

Table 1: Illustration of the features used to construct DoG transform.

$\phi_{DoG}$  combines features  $M_{1-4}$  per step and scales them by the normalized simulation time to obtain the DoG score of controller  $\mathbf{x}$ :

$$score_{DoG} = \frac{t_{sim}}{t_{max}} \cdot \sum_{s=1}^N \sum_{k=1}^4 M_{ks} \quad (2)$$

Here, subscript  $k$  identifies the feature metric,  $s$  is the step number,  $N$  is the total number of steps taken in simulation,  $t_{sim}$  is time at which simulation terminated (possibly due to a fall),  $t_{max}$  is total time allotted for simulation. Since larger number of steps lead to higher DoG, some controllers that chatter (step very fast before falling) could get misleadingly high scores; we scale the scores by  $\frac{t_{sim}}{t_{max}}$  to prevent that.  $\phi_{DoG}(\mathbf{x})$  for controller parameters  $\mathbf{x}$  now becomes the computed  $score_{DoG}$  of the resulting trajectories when  $\mathbf{x}$  is simulated.  $\phi_{DoG}$  essentially aids in (soft) clustering of controllers based on their behaviour in simulation. High scoring controllers are more likely to walk than low scoring ones. Since  $M_{1-4}$  are based on intuitive gait features, they are more likely to transfer between simulation and hardware, as compared to direct cost. The thresholds in  $M_{1-3}$  are chosen according to values observed in nominal human walking from Winter and Yack (1987).

### 3.1.2. LEARNING A FEATURE TRANSFORM WITH A NEURAL NETWORK

While domain-specific feature transforms can be extremely useful and robust, they might be difficult to generate when a domain expert is not present. This motivates directly learning such feature transforms from trajectory data. In this section we describe our approach to train neural networks to reconstruct trajectory summaries (Antonova et al., 2017) that achieves this goal of minimizing expert involvement.

Trajectory summaries are a convenient choice for reparametrizing controllers into an easy to optimize space. For example, controllers that fall would automatically be far away from controllers that walk. If these trajectories can be extracted from a high-fidelity sim-

ulator, we would not have to evaluate each controller on hardware. However, conventional implementations of BO evaluate the kernel function for a large number of points per iteration, requiring thousands of simulations each iteration. To avoid this, a Neural Network (NN) can be trained to reconstruct trajectory summaries from a large set of pre-sampled data points. NN provides flexible interpolation, as well as fast evaluation (controller  $\rightarrow$  trajectory summary). Furthermore, trajectories are agnostic to the specific cost used during BO. Thus the data collection can be done offline, and there is no need to re-run simulations in case the definition of the cost is modified.

We use the term ‘trajectory’ in a general sense, referring to several sensory states recorded during a simulation. To create trajectory summaries for the case of locomotion, we include measurements of: walking time (time before falling), energy used during walking, position of the center of mass and angle of the torso. With this, we construct a dataset for NN to fit: a Sobol grid of controller parameters ( $\mathbf{x}_{1:N}$ ,  $N \approx 0.5$  million) along with trajectory summaries  $\xi_{\mathbf{x}_i}$  from simulation. NN is trained using mean squared loss:

NN input:  $\mathbf{x}$  – a set of controller parameters

NN output:  $\phi_{trajNN}(\mathbf{x}) = \hat{\xi}_{\mathbf{x}}$  – reconstructed trajectory summary

NN loss:  $\frac{1}{2} \sum_{i=1}^N \|\hat{\xi}_{\mathbf{x}_i} - \xi_{\mathbf{x}_i}\|^2$

The outputs  $\phi_{trajNN}(\mathbf{x})$  are then used in the kernel for BO:

$$k_{trajNN}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_k^2 \exp\left(-\frac{1}{2} \mathbf{t}_{ij}^T \text{diag}(\boldsymbol{\ell})^{-2} \mathbf{t}_{ij}\right), \quad \mathbf{t}_{ij} = \phi_{trajNN}(\mathbf{x}_i) - \phi_{trajNN}(\mathbf{x}_j) \quad (3)$$

Appendix A describes our data collection and training. We did not carefully select the sensory traces used in the trajectory summaries. Instead, we used the most obvious states, aiming for an approach that could be easily adapted to other domains. To apply this approach to a new setting, one could simply include information that is customarily tracked, or used in costs. For example, for a manipulator, the coordinates of the end effector(s) could be recorded at relevant points. Force-torque measurements could be included, if available.

### 3.2. Kernel Adjustment for Handling Simulation-Hardware Mismatch

Approaches described in previous sections could provide improvement for BO when a high-fidelity simulator is used in kernel construction. In Rai et al. (2018) we presented promising results of experimental evaluation on hardware. However, it is unclear how the performance changes when simulation-hardware mismatch becomes apparent.

In Rai et al. (2018), we also proposed a way to incorporate information about simulation-hardware mismatch into the kernel from the samples evaluated so far. We augment the simulation-based kernel to include this additional information about mismatch, by expanding the original kernel by an extra dimension that contains the predicted mismatch for each controller  $\mathbf{x}$ .

A separate Gaussian process is used to model the mismatch experienced on hardware, starting from an initial prior mismatch of 0:  $g(\mathbf{x}) \sim \mathcal{GP}(0, k_{SE}(\mathbf{x}_i, \mathbf{x}_j))$ . For any evaluated controller  $\mathbf{x}_i$ , we can compute the difference between  $\phi(\mathbf{x}_i)$  in simulation and on hardware:  $d_{\mathbf{x}_i} = \phi^{sim}(\mathbf{x}_i) - \phi^{hw}(\mathbf{x}_i)$ . We can now use mismatch data  $\{d_{\mathbf{x}_i} | i = 1 \dots n\}$  to construct a model for the expected mismatch:  $\bar{g}(\mathbf{x})$ . In the case of using a GP-based model,  $\bar{g}(\cdot)$  would denote the posterior mean. With this, we can predict simulation-hardware mismatch in the original space of controller parameters for unevaluated controllers. Combining this with

kernel  $k_\phi$  we obtain an adjusted kernel:

$$\begin{aligned} \boldsymbol{\phi}_{\mathbf{x}}^{adj} &= \begin{bmatrix} \phi^{sim}(\mathbf{x}) \\ \bar{g}(\mathbf{x}) \end{bmatrix}, & \mathbf{t}_{ij}^{adj} &= \boldsymbol{\phi}_{\mathbf{x}_i}^{adj} - \boldsymbol{\phi}_{\mathbf{x}_j}^{adj} \\ k_{\phi_{adj}}(\mathbf{x}_i, \mathbf{x}_j) &= \sigma_k^2 \exp \left( -\frac{1}{2} (\mathbf{t}_{ij}^{adj})^T \text{diag} \left( \begin{bmatrix} \boldsymbol{\ell}_1 \\ \boldsymbol{\ell}_2 \end{bmatrix} \right)^{-2} \mathbf{t}_{ij}^{adj} \right) \end{aligned} \quad (4)$$

The similarity between points  $\mathbf{x}_i, \mathbf{x}_j$  is now dictated by two components: representation in  $\phi$  space and expected mismatch. This construction has an intuitive explanation: Suppose controller  $\mathbf{x}_i$  results in walking when simulated, but falls during hardware evaluation.  $k_{\phi_{adj}}$  would register a high mismatch for  $\mathbf{x}_i$ . Controllers would be deemed similar to  $\mathbf{x}_i$  only if they have both similar simulation-based  $\phi(\cdot)$  and similar estimated mismatch. Points with similar simulation-based  $\phi(\cdot)$  and low predicted mismatch would still be ‘far away’ from the failed  $\mathbf{x}_i$ . This would help BO sample points that still have high chances of walking in simulation, but are in a different region of the original parameter space. In the next section, we present a more mathematically rigorous interpretation for  $k_{\phi_{adj}}$ .

### 3.2.1. INTERPRETATION OF KERNEL WITH MISMATCH MODELING

Let us consider a controller  $\mathbf{x}_i$  evaluated on hardware. The difference between simulation-based and hardware-based feature transform for  $\mathbf{x}_i$  is  $d_{\mathbf{x}_i} = \phi^{sim}(\mathbf{x}_i) - \phi^{hw}(\mathbf{x}_i)$ . The ‘true’ hardware feature transform for  $\mathbf{x}_i$  is  $\phi^{hw}(\mathbf{x}_i) = \phi^{sim}(\mathbf{x}_i) - d_{\mathbf{x}_i}$ . After  $n$  evaluations on hardware,  $\{d_{\mathbf{x}_i} | i = 1 \dots n\}$  can serve as data for modeling simulation-hardware mismatch. In principle, any data-efficient model  $g(\cdot)$  can be used, such as GP (a multi-output GP in case  $\phi(\cdot) > 1D$ ). With this, we can obtain an adjusted transform:  $\hat{\phi}^{hw}(\mathbf{x}) = \phi^{sim}(\mathbf{x}) - \bar{g}(\mathbf{x})$ , where  $\bar{g}(\cdot)$  is the output of the model fitted using  $\{d_{\mathbf{x}_i} | i = 1 \dots n\}$ .

Suppose  $\mathbf{x}_{new}$  has not been evaluated on hardware. We can use  $\hat{\phi}^{hw}(\mathbf{x}_{new}) = \phi^{sim}(\mathbf{x}_{new}) - \bar{g}(\mathbf{x}_{new})$  as the adjusted estimate of what the output of  $\phi$  should be, taking into account what we have learned so far about simulation-hardware mismatch.

Let’s construct kernel  $k_{\phi_{adj}}^{v_2}(\mathbf{x}_i, \mathbf{x}_j)$  that uses these hardware-adjusted estimates directly:

$$\begin{aligned} \mathbf{q}_{ij}^{adj} &= \hat{\phi}^{hw}(\mathbf{x}_i) - \hat{\phi}^{hw}(\mathbf{x}_j) \\ &= (\phi^{sim}(\mathbf{x}_i) - \bar{g}(\mathbf{x}_i)) - (\phi^{sim}(\mathbf{x}_j) - \bar{g}(\mathbf{x}_j)) \\ &= \underbrace{(\phi^{sim}(\mathbf{x}_i) - \phi^{sim}(\mathbf{x}_j))}_{\mathbf{v}_\phi} - \underbrace{(\bar{g}(\mathbf{x}_i) - \bar{g}(\mathbf{x}_j))}_{\mathbf{v}_g} \\ k_{\phi_{adj}}^{v_2}(\mathbf{x}_i, \mathbf{x}_j) &= \sigma_{k_{v_0}}^2 \exp \left( -\frac{1}{2} (\mathbf{q}_{ij}^{adj})^T \text{diag}(\boldsymbol{\ell})^{-2} \mathbf{q}_{ij}^{adj} \right) \\ &= \sigma^2 \exp \left( -\frac{1}{2} [(\mathbf{v}_\phi - \mathbf{v}_g)^T \text{diag}(\boldsymbol{\ell})^{-2} (\mathbf{v}_\phi - \mathbf{v}_g)] \right) \\ &= \sigma^2 \exp \left( -\frac{1}{2} [\mathbf{v}_\phi^T \text{diag}(\boldsymbol{\ell})^{-2} \mathbf{v}_\phi + \mathbf{v}_g^T \text{diag}(\boldsymbol{\ell})^{-2} \mathbf{v}_g - prod_{ij}] \right) \\ &\quad \text{where } prod_{ij} = 2\mathbf{v}_\phi^T \text{diag}(\boldsymbol{\ell})^{-2} \mathbf{v}_g \end{aligned}$$

Using  $\exp(a + b + c) = \exp(c) \cdot \exp(a + b)$ , we have:

$$k_{\phi_{adj}}^{v_2}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp(prod_{ij}) \exp \left( -\frac{1}{2} [\mathbf{v}_\phi^T \text{diag}(\boldsymbol{\ell})^{-2} \mathbf{v}_\phi + \mathbf{v}_g^T \text{diag}(\boldsymbol{\ell})^{-2} \mathbf{v}_g] \right)$$

Which can be written as:

$$k_{\phi_{adj}}^{v_2}(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp(-\text{prod}_{ij}) \exp\left(-\frac{1}{2}(\mathbf{t}_{ij}^{adj})^T \text{diag}\left(\begin{bmatrix} \boldsymbol{\ell} \\ \boldsymbol{\ell} \end{bmatrix}\right)^{-2} \mathbf{t}_{ij}^{adj}\right)$$

$$\mathbf{t}_{ij}^{adj} = \begin{bmatrix} \mathbf{v}_\phi \\ \mathbf{v}_g \end{bmatrix} = \begin{bmatrix} \phi^{\text{sim}}(\mathbf{x}_i) - \phi^{\text{sim}}(\mathbf{x}_j) \\ \bar{g}(\mathbf{x}_i) - \bar{g}(\mathbf{x}_j) \end{bmatrix}$$

Compare this to  $k_{\phi_{adj}}$  from Equation 4:

$$k_{\phi_{adj}}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_k^2 \exp\left(-\frac{1}{2}(\mathbf{t}_{ij}^{adj})^T \text{diag}\left(\begin{bmatrix} \boldsymbol{\ell}_1 \\ \boldsymbol{\ell}_2 \end{bmatrix}\right)^{-2} \mathbf{t}_{ij}^{adj}\right) \quad (5)$$

Now we see that  $k_{\phi_{adj}}^{v_2}$  and  $k_{\phi_{adj}}$  have a similar form. Hyperparameters  $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2$  provide flexibility in  $k_{\phi_{adj}}$  as compared to having only vector  $\boldsymbol{\ell}$  in  $k_{\phi_{adj}}^{v_2}$ . They can be adjusted manually or with Automatic Relevance Determination. For  $k_{\phi_{adj}}^{v_2}$ , the role of signal variance is captured by  $\sigma^2 \exp(-\text{prod}_{ij})$ . This makes the kernel nonstationary in the transformed  $\phi$  space. Since  $k_{\phi_{adj}}$  is already non-stationary in  $\mathbf{x}$ , it is unclear whether non-stationarity of  $k_{\phi_{adj}}^{v_2}$  in the transformed  $\phi$  space has any advantages.

The above discussion shows that  $k_{\phi_{adj}}$  proposed in Rai et al. (2018) is motivated both intuitively and mathematically. It aims to use a transform that accounts for the hardware mismatch, without adding extra non-stationarity in the transformed space. While the above analysis shows the connection between  $k_{\phi_{adj}}$  and  $k_{\phi_{adj}}^{v_2}$ , a more systematic empirical analysis would be beneficial as part of future work.

## 4. Robots, Simulators and Controllers Used

In this section we give a concise description of the robots, controllers and simulators used in experiments with BO for bipedal locomotion. Our approach is applicable to a wide range of bipedal robots and controllers, including state-of-the-art controllers (Feng et al., 2015). We work with two different types of controllers – a reactively stepping controller and a human-inspired neuromuscular controller (NMC). The reactively stepping controller is model-based: it uses inverse-dynamics models of the robot to compute desired motor torques. In contrast, NMC is model-free: it computes desired torques using hand-designed policies, created with biped locomotion dynamics in mind. These controllers exemplify two different and widely used ways of controlling bipedal robots. In addition to this, we show results on two different robot morphologies – a parallel bipedal robot ATRIAS, and a serial 7-link biped model. Our hardware experiments are conducted on ATRIAS; the 7-link biped is only used in simulation. Our success on both robots shows that the approaches developed in this paper are widely applicable to a range of bipedal robots and controllers.

### 4.1. ATRIAS Robot

Our hardware platform is an ATRIAS robot (Figure 1). ATRIAS is a parallel bipedal robot, weighing  $\approx 64\text{kg}$ . The legs are 4-segment linkages actuated by 2 Series Elastic Actuators (SEAs) in the sagittal plane and a DC motor in the lateral plane. Details can be found in Hubicki et al. (2016). In this work we focus on planar motion around a boom. ATRIAS is a highly dynamic system due to its point feet, with static stability only in double stance on the boom.

## 4.2. Planar 7-link Biped

The second robot used in our experiments is a 7-link biped (Thatte and Geyer, 2016). It has a trunk and segmented legs with ankles, with actuators in the hip, knees and ankles. The inertial properties of its links are similar to an average human (Winter and Yack, 1987). This 7-link model is a canonical simulator for testing bipedal walking algorithms, for example in Song and Geyer (2015). It is a simplified two-dimensional simulator for a large range of humanoid robots, like Atlas (Feng et al., 2015). We use this simulator to study the generalizability of our proposed approaches to systems different from ATRIAS.

## 4.3. Feedback Based Reactive Stepping Policy

We design a parametric controller for controlling the CoM height, torso angle and the swing leg by commanding desired ground reaction forces and swing foot landing location.

$$F_x = K_{pt}(\theta_{des} - \theta) - K_{dt}\dot{\theta} \quad F_z = K_{pz}(z_{des} - z) - K_{dz}\dot{z} \quad x_p = k(v - v_{tgt}) + Cd + 0.5vT$$

Here,  $F_x$  is the desired horizontal ground reaction force (GRF),  $K_{pt}$  and  $K_{dt}$  are the proportional and derivative feedback gains on the torso angle  $\theta$  and velocity  $\dot{\theta}$ .  $F_z$  is the desired vertical GRF,  $K_{pz}$  and  $K_{dz}$  are the proportional and derivative gains on the CoM height  $z$  and vertical velocity  $\dot{z}$ .  $z_{des}$  and  $\theta_{des}$  are the desired CoM height and torso lean.  $x_p$  is the desired foot landing location for the end of swing;  $v$  is the horizontal CoM velocity,  $k$  is the feedback gain that regulates  $v$  towards the target velocity  $v_{tgt}$ .  $C$  is a constant and  $d$  is the distance between the stance leg and the CoM;  $T$  is the swing time.

The desired GRFs are sent to ATRIAS inverse dynamics model that generates desired motor torques  $\tau_f, \tau_b$ . Details can be found in Rai et al. (2018). This controller assumes no double-support, and the swing leg takes off as soon as stance is detected. This leads to a highly dynamic gait, posing a challenging optimization problem.

To investigate the effects of increasing dimensionality on our optimization, we construct two controllers with different number of free parameters:

- *5-dimensional controller* : optimizing 5 parameters  $[K_{pt}, K_{dt}, k, C, T]$   
( $z_{des}, \theta_{des}$  and the feedback on  $z$  are hand tuned)
- *9-dimensional controller* : optimizing all 9 parameters of the high-level policy  
 $[K_{pt}, K_{dt}, \theta_{des}, K_{pz}, K_{dz}, z_{des}, k, C, T]$

## 4.4. 16-dimensional Neuromuscular Controller

We use neuromuscular model policies, as introduced in Geyer and Herr (2010), as our controller for the 7-link planar human-like biped model. These policies use approximate models of muscle dynamics and human-inspired reflex pathways to generate joint torques, producing gaits that are similar to human walking.

Each leg is actuated by 7 muscles, which together produce torques about the hip, knee and ankle. Most of the muscle reflexes are length or force feedbacks on the muscle state aimed at generating a compliant leg, keeping knee from hyperextending and maintaining torso orientation in stance. The swing control has three main components – target leg angle, leg clearance and hip control due to reaction torques. Together with the stance control, this leads to a total of 16 controller parameters, described in details in Antonova et al. (2016).

Though originally developed for explaining human neural control pathways, this controller has recently been applied to prosthetics and bipeds, for example Thatte and Geyer (2016) and Van der Noot et al. (2015). This controller is capable of generating a variety of locomotion behaviours for a humanoid model – walking on rough ground, turning, running, and walking upstairs, making it a very versatile controller (Song and Geyer, 2015). This is a model-free controller as compared to the reactive-stepping controller, which was model-based.

#### 4.5. 50-dimensional Virtual Neuromuscular Controller

Another model-free controller we use on ATRIAS is a modified version of Batts et al. (2015). VNMC maps a neuromuscular model, similar to the one described in Section 4.4 to the ATRIAS robot’s topology and emulates it to generate desired motor torques. We adapt VNMC by removing some biological components while preserving its basic functionalities. The final version of the controller consists of 50 parameters including low-level control parameters, such as feedback gains, as well as high level parameters, such as desired step length and desired torso lean. Details can be found in Rai et al. (2018). When optimized using CMA-ES, it can control ATRIAS to walk on rough terrains with height changes of  $\pm 20$  cm in planar simulation (Batts et al., 2015).

#### 4.6. Increasingly Inaccurate Simulators

To compare the performance of different methods that can be used to transfer information from simulation to hardware, we create a series of increasingly approximate simulators. These simulators emulate increasing mismatch between simulation and hardware and its effect on the information transfer. In this setting, the high-fidelity ATRIAS simulator (Martin et al., 2015), which was used in all the previous simulation experiments becomes the “*hardware*”. Next we make dynamics approximations to the original simulator, which are used commonly in simulators to decrease fidelity and increase simulation speed. For example, the complex dynamics of harmonic drives are approximated as a torque multiplication, and the boom is removed from the simulation, leading to a two-dimensional simulator. These approximate simulators now become the “*simulators*”. As the approximations in these simulators are increased, we expect the performance of methods that utilize simulation for optimization on hardware to deteriorate.

The details of the approximate simulators are described in the two paragraphs below:

**1. Simulation with simplified gear dynamics :** We replace the original high-fidelity gear model with a commonly used approximation for geared systems – multiplying the rotor torque by the gear ratio. This reduces the simulation time to about a third of the original simulator, but leads to an approximate gear dynamics model.

**2. Simulation with no boom and simplified gear dynamics :** The ATRIAS robot walks on a boom in our hardware experiments. In our second approximation, we remove the boom from the original simulator and constraint the motion of the robot to a 2-dimensional plane. This is a common approximation for two-dimensional robots. However, the boom leads to lateral forces on the robot, which have vertical force components that are not modelled anymore.



Figure 3: ATRIAS during BO with DoG-based kernel (video: <https://youtu.be/hpXNFREgaRA>)

The advantage of such an arrangement is that we can test the effect of un-modelled and wrongly modelled dynamics on information transfer between simulation and hardware. Even in our high-fidelity original simulator, there are several un-modelled components of the actual hardware. For example, the non-rigidness of the robot parts, misaligned motors and relative play between joints. In our experiments, we find that the 50-dimensional VNMC is a sensitive controller, with little hope of directly transferring from simulation to hardware. Anticipating this, we can now test several methods of compensating for this mismatch using our increasingly approximate simulators.

## 5. Experiments

We will now present our experiments on optimizing controllers that are 5, 9, 16 and 50 dimensional. We split our experiments into three categories: hardware experiments on the ATRIAS robot, simulation experiments on the 7-link biped and experiments using simulators with different levels of mismatch. We demonstrate that our proposed approach is able to generalize to different controllers and robot structures and is also robust to simulation inaccuracies. The sections below present experimental results, while Appendix A gives further details on data collection, BO implementation and kernel generation.

### 5.1. Hardware Experiments on the ATRIAS Robot

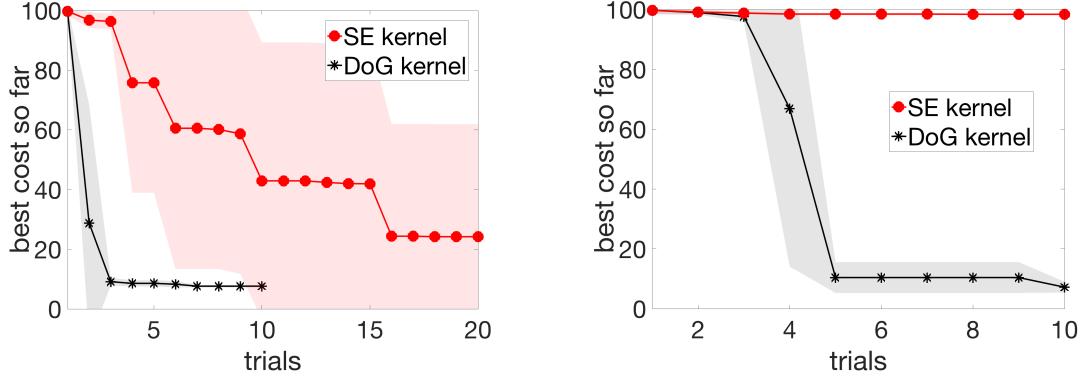
In this section we describe experiments conducted on the ATRIAS robot (hardware from Section 4.1). These experiments were conducted around a boom. The cost function used in our experiments is a slight modification of the cost used in Song and Geyer (2015):

$$\text{cost} = \begin{cases} 100 - x_{fall}, & \text{if fall} \\ ||v_{avg} - v_{tgt}||, & \text{if walk} \end{cases} \quad (6)$$

where  $x_{fall}$  is distance covered before falling,  $v_{avg}$  is average speed per step and  $v_{tgt}$  contains target velocity profile, which can be variable. This cost function heavily penalizes falls, and encourages walking controllers to track target velocity.

We do multiple runs of each algorithm on the robot. Each run typically consists of 10 experiments on the robot. All BO runs start from scratch, with an uninformed GP prior. At the end of the run, the GP posterior has 10 data points, depending on the experiment. Each robot trial is designed to be between 30s to 60s long and the robot is reset to its “home” position between trials.

We will present two sets of hardware experiments in the following subsections. First we describe experiments with the DoG-based kernel on the 5 and 9 dimensional controllers



(a) BO for 5D controller. BO with SE finds walking points in 4/5 runs within 20 trials. BO with DoG-based kernel finds walking points in 5/5 runs within 3 trials.

(b) BO for 9D controller. BO with SE doesn't find walking points in 3 runs. BO with DoG-based kernel finds walking points in 3/3 runs within 5 trials.

Figure 4: BO for 5D and 9D controller on ATRIAS robot hardware. Plots show mean best cost so far. Shaded region shows  $\pm$  one standard deviation. From Rai et al. (2018).

(first reported in Rai et al. (2018)). The second set describes a new set of experiments for optimizing a 9-dimensional controller using a Neural Network based kernel on hardware.

### 5.1.1. EXPERIMENTS WITH A 5-DIMENSIONAL CONTROLLER AND DOG-BASED KERNEL

In our first set of experiments on the robot, we investigated optimizing the 5-dimensional controller from Section 4.3. For these experiments we picked a challenging variable target speed profile:  $0.4m/s$  (15 steps) -  $1.0m/s$  (15 steps) -  $0.2m/s$  (15 steps) -  $0m/s$  (5 steps). The controller was stopped after the robot took 50 steps.

To evaluate the difficulty of this setting, we sampled 100 random points on hardware. 10% of these randomly sampled controllers could walk. In contrast, in simulation the success rate of random sampling was  $\approx 27.5\%$ . This indicates that the simulation was easier, which could be potentially detrimental to algorithms that rely heavily on simulation, because a large portion of controllers that walk in simulation fall on hardware. Nevertheless, using a DoG-based kernel offered significant improvements over a standard SE kernel (Figure 4a).

We conducted 5 runs of BO with DoG-based kernel and 5 runs of BO with SE, 10 trials for DoG-based kernel per run, and 20 for SE kernel. In total, this led to 150 experiments on the robot (excluding the 100 random samples, which were not used during BO). BO with DoG-based kernel found walking points in 100% of runs within 3 trials. In comparison, BO with SE found walking points in 10 trials in 60% runs, and in 80% runs in 20 trials (Figure 4a). Although BO could find walking controllers as early as the second trial (with no prior hardware information), it is worth noting that the optimization did not converge after only a few trials. Sampling more could possibly lead to better walking controllers, but our goal was to find the best controller with a budget of only 10 to 20 trials.

### 5.1.2. EXPERIMENTS WITH A 9-DIMENSIONAL CONTROLLER AND DOG-BASED KERNEL

Our next set of experiments optimized the 9-dimensional controller from Section 4.3. First, we sampled 100 random points for the variable speed profile described above, but this led

to no walking points. To ensure that we have a reasonable baseline we decided to simplify the speed profile for this setting:  $0.4m/s$  for 30 steps. We evaluated 100 random points on hardware, and 3 walked for the easier speed profile. In comparison, the success rate in simulation was 8% for the tougher variable-speed profile, implying an even greater mismatch between hardware and simulation than the 5-dimensional controller. For experiments in Sections 5.1.1 and 5.1.2 the inertial measurement unit (IMU) of the robot was broken, and we replaced its functionality with external boom sensors. These were lower resolution than the IMU, leading to noisier readings and larger time delays. As a result, the system did not have a good estimation of vertical height of the CoM, leading to poor control authority. However, the IMU on ATRIAS is a very expensive fiber-optic IMU that is not commonly used on humanoid robots, and most robots use simple state estimation methods. So, this is a common setting for humanoid robots, even if it presents a challenge for the optimization methods.

We conducted 3 runs of BO with DoG-based kernel and BO with SE, 10 trials for DoG-based kernel per run, and 10 for SE. In total, this led to 60 experiments on the hardware (excluding the random samples, which were not used for BO). BO with DoG-based kernel found walking points in 5 trials in 3/3 runs. BO with SE did not find any walking points in 10 trials in all 3 runs. These results are shown in Figure 4b.

Based on these results, we concluded that BO with DoG-based kernel was indeed able to extract useful information from simulation and speed up learning on hardware, even when there was mismatch between simulation and hardware.

### 5.1.3. EXPERIMENTS WITH A 9-DIMENSIONAL CONTROLLER AND NN-BASED KERNEL

In the next set of experiments, we evaluated performance of the NN-based kernel described in Section 3.1.2. We optimize the 9-dimensional controller from Section 4.3.

The target of hardware experiments was to walk for 30 steps at  $0.4m/s$ , similar to Section 5.1.2.

For these experiments the IMU was repaired, leading to better state estimation on the robot. For a fair comparison, we re-ran experiments with the baseline for this setting and the baseline performed slightly better than the baseline of earlier experiments (because of improved sensing).

Figure 5 shows comparison of BO with NN-based kernel and SE kernels. We conducted 5 runs of both algorithms with 10 trials in each run, leading to a total of 100 robot trials. BO with the NN-based kernel found walking points in all 5 runs within 6 trials, while BO with SE kernel only found walking points in 2 of 5 runs in 10 trials. Hence, even without explicit hand-designed domain knowledge, like the DoG-based kernel, the NN-based kernel is able to extract useful information from simulation and successfully guide hardware experiments.

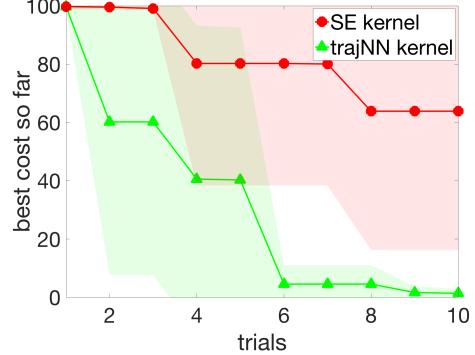


Figure 5: BO for 9D controller on ATRIAS robot hardware.

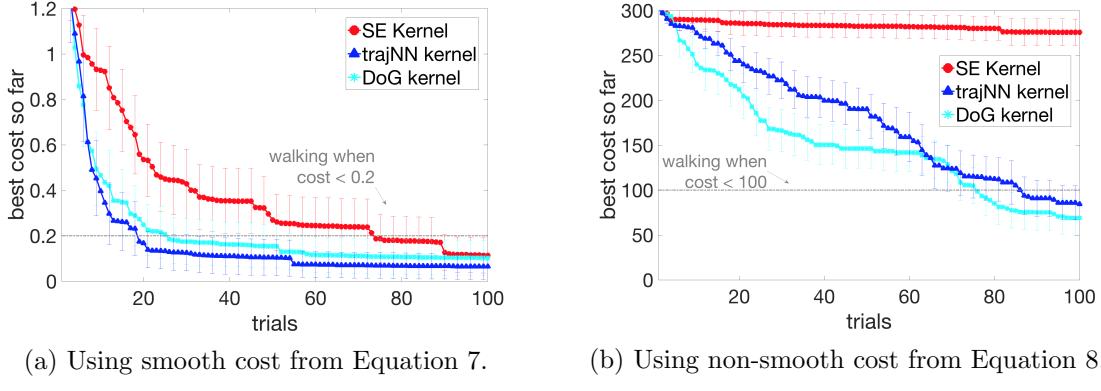


Figure 6: BO for the Neuromuscular controller. *trajNN* and *DoG* kernels were constructed with undisturbed model on flat ground. BO is run with mass/inertia disturbances on different rough ground profiles. Plots show means over 50 runs, 95% CIs. From Antonova et al. (2017).

### 5.2. Simulation Experiments on a 7-link Biped

This section describes simulation experiments with a 16-dimensional Neuromuscular controller (Section 4.4) on a 7-link biped model. These experiments, from Antonova et al. (2017), highlight the cost-agnostic nature of our approach by optimizing two very different costs.

Figure 6 shows BO with DoG-based kernel, NN-based kernel and SE kernel for two different costs from prior literature. The first cost promotes walking further and longer before falling, while penalizing deviations from the target speed (Antonova et al., 2016):

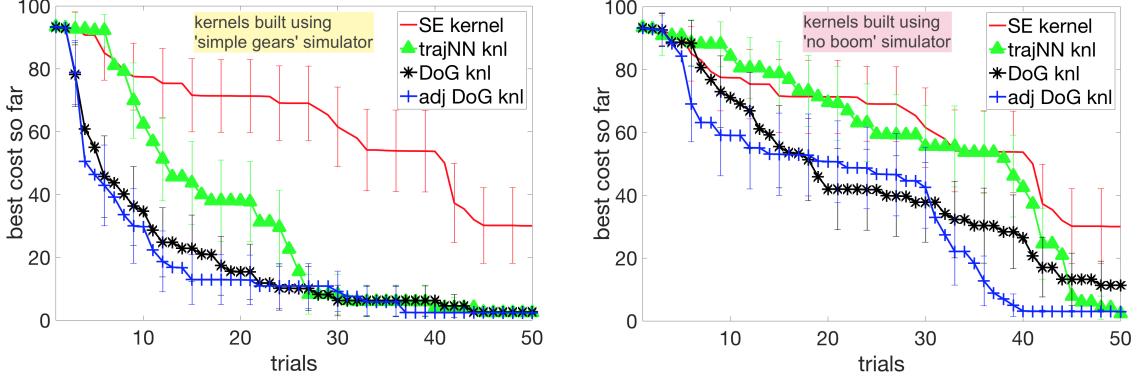
$$cost_{smooth} = 1/(1+t) + 0.3/(1+d) + 0.01(s - s_{tgt}), \quad (7)$$

where  $t$  is seconds walked,  $d$  is the final CoM position,  $s$  is speed and  $s_{tgt}$  is the desired walking speed ( $1.3m/s$  in our case). The second cost function is similar to the cost used in Section 5. It penalizes falls explicitly, and encourages walking at desired speed and with lower cost of transport:

$$cost_{non-smooth} = \begin{cases} 300 - x_{fall}, & \text{if fall} \\ 100\|v_{avg} - v_{tgt}\| + c_{tr}, & \text{if walk} \end{cases} \quad (8)$$

where  $x_{fall}$  is the distance covered before falling,  $v_{avg}$  is the average speed of walking,  $v_{tgt}$  is the target velocity, and  $c_{tr}$  captures the cost of transport. The changed constants is to account for a longer simulation time. Figure 6a shows that the NN-based kernel and the DoG-based kernel offer a significant improvement over BO with the SE kernel in sample efficiency when using the  $cost_{smooth}$ , with more than 90% of runs achieving walking after 25 trials. BO with the SE kernel takes 90 trials to get 90% success rate. Figure 6b shows that similar performance by the two proposed approaches is observed on the non-smooth cost. With the NN-based kernel, 70% of the runs find walking solutions after 100 trials, similar to the DoG-based kernel. However, optimizing non-smooth cost is very challenging for BO with the SE kernel: a walking solution is found only in 1 out of 50 runs after 100 trials.

We attribute the difference in performance of the SE kernel on the two costs to the nature of the costs. If a point walks some distance  $d$ , Equation 7 reduces in terms of  $\frac{1}{d}$  and Equation 8 reduces by  $-d$ . A sharper fall in the first cost encourages BO to exploit around points that walk some distance, quickly finding points that walk forever. BO with the second cost continues to explore, as the signal is too weak. However the success of both



(a) Informed kernels generated using simulator with simplified gear dynamics.

(b) Informed kernels generated using simplified gear dynamics, without boom model.

Figure 7: BO is run on the original simulator. Informed kernels perform well despite significant mismatch, when kernels are generated using simulator with simplified gear dynamics (left). In the case of severe mismatch, when the boom model is also removed, informed kernels still improve over baseline SE (right). Plots show best cost for mean over 50 runs for each algorithm, 95% CIs.

NN-based and DoG-based kernels on both costs shows that the same kernel can indeed be used for optimizing multiple costs robustly, without any further tuning needed.

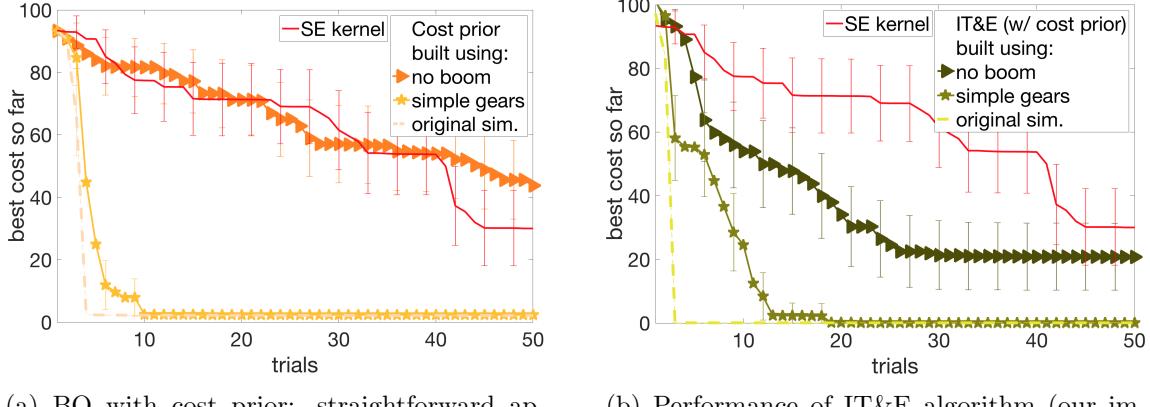
### 5.3. Experiments with Increasing Simulation-Hardware Mismatch

In this section, we describe our experiments with increasing simulation-hardware mismatch and its effect on approaches that use information from simulation during hardware optimization. The quality of information transfer between simulation and hardware depends not only on the mismatch between the two, but also on the controller used. For a robust controller, small dynamics errors would not cause a significant deterioration in performance, while for a sensitive controller this might be much more detrimental.

In the rest of this section, we provide experimental analyses of settings with increasing simulated mismatch and their effect on optimization of the 50-dimensional VNMC from Section 4.5. We compare several approaches that improve sample-efficiency of BO and investigate if the improvement they offer is robust to mismatch between the simulated setting used for constructing kernel/prior and the setting on which BO is run.

First, we examine the performance of our proposed approaches with informed kernels:  $k_{DoG}$ ,  $k_{trajNN}$  and  $k_{DoG_{adj}}$ . Figure 7a shows the case when informed kernels are generated using the simulator with simplified gear dynamics while BO is run on the original simulator. After 50 trials, all runs with informed kernels find walking solutions, while for SE only 70% have walking solutions.

Next, Figure 7b shows performance of  $k_{DoG}$ ,  $k_{trajNN}$  and  $k_{DoG_{adj}}$  when the kernels are constructed using a simulator with simplified dynamics and without a the boom. In this case the mismatch with the original simulator is larger than before and we see the advantage of using adjustment for DoG-based kernel:  $k_{DoG_{adj}}$  finds walking points in all runs after 35 trials.  $k_{trajNN}$  also achieves this, but after 50 trials.  $k_{DoG}$  finds walking points in  $\approx 90\%$



(a) BO with cost prior: straightforward approach useful for low-to-medium mismatch; but no improvement if mismatch is severe.

(b) Performance of IT&E algorithm (our implementation of Cully et al. (2015), adapted to bipedal locomotion).

Figure 8: BO using prior-based approaches. Mean over 50 runs for each algorithm, 95% CIs.

of the runs after 50 trials. The performance of SE stays the same, as it uses no prior information from any simulator.

This illustrates that while the original DoG-based kernel can recover from slight simulation-hardware mismatch, the adjusted DoG-kernel is required for higher mismatch.  $k_{\text{trajNN}}$  seems to recover from the mismatch, but might benefit from an adjusted version. We leave this to future work.

### 5.3.1. COMPARISONS OF PRIOR-BASED AND KERNEL-BASED APPROACHES

In this section, we classify approaches that use simulation information in hardware optimization as prior-based or kernel-based. Prior-based approaches use costs from the simulation in the prior of the GP used in the BO. This can help BO a lot if the costs are similar between simulation and hardware, and the cost function is fixed. However, in the presence of large mismatch, controllers that perform well in simulation might fail on hardware. A prior-based method can be biased towards sampling promising points from simulation, resulting in worse performance than uninformed BO. Kernel-based approaches consist of methods that incorporate information from simulation into the kernel of the GP. These can be less sample-efficient as compared to prior-based method, but not as likely to be biased towards unpromising regions in the presence of mismatch. They also easily generalize to multiple costs, so that there is no need to re-run simulations for data collection if the cost is changed. This is important because a lot of these approaches can take several days of computation to generate a cost prior or informed kernel. For example, Cully et al. (2015) report taking 2 weeks on a 16-core computer to generate their map.

It is possible to also combine both prior-based and kernel-based methods, as in Cully et al. (2015). We classify these as ‘prior-based’ methods, since in our experiments prior outweighs the kernel effects for such cases. In our comparison with Cully et al. (2015), we implement a version with and without the prior points. We do not add a cost prior to BO using DoG-based kernel, as this would limit us to a particular cost and high-fidelity simulators, and both of these can be major obstacles in real robot experiments.

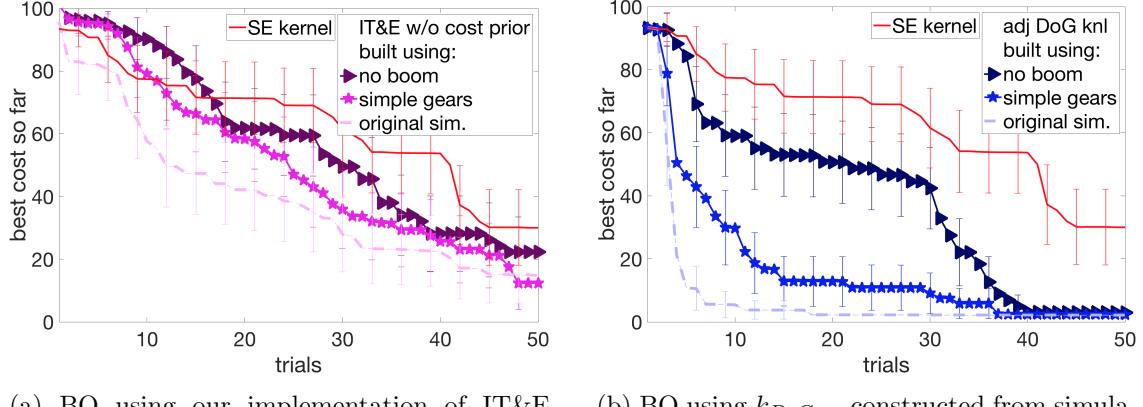
Figure 8a shows the performance when using simulation cost in the prior during BO. BO with a cost prior created using the original version of the simulator illustrates what would happen in the best case scenario. When the simulator with simplified gear dynamics is used for constructing the prior, we observe significant improvements over uninformed BO prior. However, when the prior is constructed from simplified gear dynamics and no boom setting, the approach performs slightly worse than uninformed BO. This shows that while an informed prior can be very helpful when created from a simulator close to hardware, it can hurt performance if simulator is significantly different from hardware.

Next, we discuss experiments with our implementation of Intelligent Trial and Error (IT&E) algorithm from Cully et al. (2015). This algorithm combines adding a cost prior from simulated evaluations with adding simulation information into the kernel. IT&E defines a behavior metric and tabulates best performing points from simulation on their corresponding behavior score. The behavior metric used in our experiments is duty-factor of each leg, which can go from 0 to 1.0. We discretize the duty factor into 21 cells of 0.05 increments, leading to a  $21 \times 21$  grid. We collect the 5 highest performing controllers for each square in the behavior grid, creating a  $21 \times 21 \times 5$  grid. Next, we generate 50 random combinations of a  $21 \times 21$  grid, selecting 1 out of the 5 best controllers per grid cell. Care was taken to ensure that all 5 controllers had comparable costs in the simulator used for creating the map. Cost of each selected controller is added to the prior and BO is performed in the behavior space, like in Cully et al. (2015).

Figure 8b shows BO with IT&E constructed using different versions of the simulator. IT&E constructed using simplified gear dynamics simulator is slightly less sample-efficient than the straightforward ‘cost prior’ approach. When constructed with the simulator with no boom, IT&E is able to improve over uninformed BO. However, it only finds walking points in 77% of the runs in 50 trials in this case, as some of the generated maps contained no controllers that could walk on the ‘hardware’. This is a shortcoming of the IT&E algorithm, as it eliminates a very large part of the search space and if the pre-selected space does not contain a walking point, no walking controllers can be sampled with BO. This problem could possibly be avoided by using a finer grid, or a different behavior metric. However tuning such hyper-parameters can turn out to be expensive, in computation and hardware experiment time.

To separate the effects of using simulation information in prior mean vs kernel, we evaluated a kernel-only version of IT&E algorithm (Figure 9a). It shows that the cost prior is crucial for the success of IT&E and performance deteriorates without it. Hence, it is not practical to use IT&E on a cost different than what it was generated for. Nonetheless, Figure 7 showed that BO with adjusted DoG kernel is able to handle both moderate and severe mismatch with kernel-only information, collected in Figure 9b.

Summarizing this section, we created two simulators with increasing modelling approximations, and studied the effect of using these to aid optimization on the original simulator. We found that while methods that use cost in the prior of BO can be very sample-efficient in low mismatch, their performance worsens as mismatch increases. IT&E, introduced in Cully et al. (2015), uses simulation information in both prior mean and kernel, and is very sample-efficient in cases of low mismatch. Even with high mismatch, it performed better than just prior-based BO but doesn’t find walking controllers reliably. In comparison, ad-



(a) BO using our implementation of IT&E without cost prior (from Cully et al. (2015)).

(b) BO using  $k_{DoG_{adj}}$  constructed from simulators with various levels of mismatch.

Figure 9: BO using kernel-based approaches. Mean over 50 runs for each algorithm, 95% CIs.

justed DoG-based kernel performed well in all the tested scenarios, and can reliably improve sample-efficiency of BO even when the mismatch between simulation and hardware is high.

## 6. Conclusion

In this paper, we presented and analyzed in detail our work from Antonova et al. (2016), Antonova et al. (2017) and Rai et al. (2018). These works introduce domain-specific feature transforms that can be used to optimize locomotion controllers on hardware efficiently. The feature transforms project the original controller space into a space where BO can discover promising regions quickly. We described a transform for bipedal locomotion designed with the knowledge of human walking and a neural network based transform that uses more general information from simulated trajectories. Our experiments demonstrate success at optimizing controllers on the ATRIAS robot. Further simulation-based experiments also indicate potential for other bipedal robots. For optimizing sensitive high-dimensional controllers, we proposed an approach to adjust simulation-based kernel using data seen on hardware. To study the performance of this, as well as compare our approach to other methods, we created a series of increasingly approximate simulators. Our experiments show that while several methods from prior literature can perform well with low simulation-hardware mismatch (sometimes even better than our proposed approach), they suffer when this mismatch increases. In such cases, our proposed kernels with hardware adjustment can yield reliable performance across different costs, simulators and robots.

## Acknowledgments

This research was supported in part by National Science Foundation grant IIS-1563807, the Max-Planck-Society, & the Knut and Alice Wallenberg Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding organizations.

## Appendix A: Implementation Details

In this appendix we provide a summary of data collection and implementation details. Our implementation of BO was based on the framework in Gardner et al. (2014). We used Expected Improvement (EI) acquisition function (Mockus et al., 1978). We also experimented with Upper Confidence Bound (UCB) (Srinivas et al., 2010), but found that performance was not sensitive to the choice of acquisition function. Hyper-parameters for BO were initialized to default values: 0 for mean offset, 1.0 for kernel length scales and signal variance, 0.1 for  $\sigma_n$  (noise parameter). Hyperparameters were optimized using the marginal likelihood (Shahriari et al. (2016), Section V-A). For all algorithms, we started optimizing hyperparameters after a low-cost controller was found (to save compute resources and avoid premature hyperparameter optimization). The search space boundaries for controller parameters was designed with physical constraints of the ATRIAS robot in mind.

Kernel type	Controller dim	# Sim points	Sim duration	Kernel dim	Features in kernel
$k_{DoG}$	5	20K	3.5s	1	$score_{DoG}$
	9	100K	5s	1	$score_{DoG}$
	50	200K	5s	1	$score_{DoG}$
$k_{trajNN}$	9	100K	5s	4	$t_{walk}, x_{end}, \theta_{avg}, v_{x,avg}$
	16	100K	5s	8	$t_{walk}, x_{end}, \theta_{end}, v_{x,end}$ $c_\tau, y_{end}, v_{y,end}, \dot{\theta}_{end}$
	50	200K	5s	13	$t_{walk}, x_{end}, c_\tau, \mathbf{traj}_x, \mathbf{traj}_\theta$

Table 2: Simulation data collection details.  $score_{DoG}$  was described in Section 3.1.1. For  $k_{trajNN}$ :  $t_{walk}$  is time walked in simulation before falling,  $x_{end}$  and  $y_{end}$  are the  $x$  and  $y$  positions of Center of Mass (CoM) at the end of the short simulation,  $\theta$  is the torso angle,  $\dot{\theta}$  is the torso velocity,  $v$  is the CoM speed ( $v_x$  is the horizontal and  $v_y$  is the vertical component),  $c_\tau$  is the squared sum of torques applied;  $\mathbf{traj}_x$ ,  $\mathbf{traj}_\theta$  denote vectors with mean CoM and  $\theta$  measurements every second.

$k_{trajNN}$  discussed in Section 3.1.2 was constructed by training a fully connected neural network (NN) with 3 hidden layers, using L1 loss to reconstruct features encoded by trajectory summaries. For experiments with 16D controller in Section 5.2, for example, the hidden layers contained 512, 128, 32 units; NN was trained on 100K simulated examples to reconstruct 8D trajectory summaries (see next-to-last row of Table 2). We experimented with various hidden layer sizes for 9D and 50D controllers, but did not find the overall BO performance to be sensitive to size or other training parameters, like NN learning rate. This was likely because NN used to construct an informed kernel only needs to approximately learn to separate well-performing parts from failing parts of the control parameter space. This is the benefit of our choice not to put the output of NN directly into a GP posterior.

To create cost prior for experiments in Section 5.3 we collected 50,000 evaluations of 30s trials for a range of controller parameters. Then we conducted 50 runs, using random subsets of 35,000 evaluations to construct the prior. The numbers were chosen such that this approach used similar amount of computation as our kernel-based approaches. To accommodate GP prior with a large number of points we used a sparse GP construction provided by Rasmussen and Nickisch (2010).

## References

- Rika Antonova, Akshara Rai, and Christopher G Atkeson. Sample efficient optimization for learning controllers for bipedal locomotion. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 22–28. IEEE, 2016.
- Rika Antonova, Akshara Rai, and Christopher G Atkeson. Deep Kernels for Optimizing Locomotion Controllers. In *Conference on Robot Learning (CoRL), PMLR 78*, pages 47–56, 2017.
- Zachary Batts, Seungmoon Song, and Hartmut Geyer. Toward a virtual neuromuscular control for robust walking in bipedal robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6318–6323. IEEE, 2015.
- Eric Brochu, Vlad M Cora, and Nando De Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv preprint arXiv:1012.2599*, 2010.
- Roberto Calandra, Jan Peters, Carl Edward Rasmussen, and Marc Peter Deisenroth. Manifold Gaussian processes for regression. In *International Joint Conference on Neural Networks (IJCNN)*, pages 3338–3345. IEEE, 2016a.
- Roberto Calandra, André Seyfarth, Jan Peters, and Marc Peter Deisenroth. Bayesian Optimization for Learning Gaits Under Uncertainty. *Annals of Mathematics and Artificial Intelligence*, 76(1-2):5–23, 2016b.
- Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. Robots that can adapt like animals. *Nature*, 521(7553):503–507, 2015.
- Marc Deisenroth and Carl E Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In *International Conference on Machine Learning (ICML)*, pages 465–472, 2011.
- Peter Englert and Marc Toussaint. Combined Optimization and Reinforcement Learning for Manipulation Skills. In *Robotics: Science and Systems*, 2016.
- Siyuan Feng, Eric Whitman, X Xinjilefu, and Christopher G Atkeson. Optimization-based Full Body Control for the DARPA Robotics Challenge. *Journal of Field Robotics*, 32(2):293–312, 2015.
- Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kilian Q Weinberger, and John Cunningham. Bayesian Optimization with Inequality Constraints. In *International Conference on Machine Learning (ICML)*, pages 937–945, 2014.
- Hartmut Geyer and Hugh Herr. A Muscle-reflex Model that Encodes Principles of Legged Mechanics Produces Human Walking Dynamics and Muscle Activities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(3):263–273, 2010.
- Nikolaus Hansen. The CMA evolution strategy: a comparing review. In *Towards a new evolutionary computation*, pages 75–102. Springer, 2006.

- Christian Hubicki, Jesse Grimes, Mikhail Jones, Daniel Renjewski, Alexander Spröwitz, Andy Abate, and Jonathan Hurst. ATRIAS: Design and validation of a tether-free 3D-capable spring-mass bipedal robot. *The International Journal of Robotics Research (IJRR)*, 35(12):1497–1521, 2016.
- Verne T Inman, Howard D Eberhart, et al. The major determinants in normal and pathological gait. *Journal of Bone and Joint Surgery (JBJS)*, 35(3):543–558, 1953.
- Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabás Póczos. Multi-fidelity Bayesian Optimisation with Continuous Approximations. In *International Conference on Machine Learning (ICML)*, pages 1799–1808, 2017.
- Scott Kuindersma, Robin Deits, Maurice Fallon, Andrés Valenzuela, Hongkai Dai, Frank Permenter, Twan Koolen, Pat Marion, and Russ Tedrake. Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot. *Autonomous Robots*, 40(3):429–455, 2016.
- Daniel J Lizotte, Tao Wang, Michael H Bowling, and Dale Schuurmans. Automatic Gait Optimization with Gaussian Process Regression. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 7, pages 944–949, 2007.
- Alonso Marco, Felix Berkenkamp, Philipp Hennig, Angela P Schoellig, Andreas Krause, Stefan Schaal, and Sebastian Trimpe. Virtual vs. real: Trading off simulations and physical experiments in reinforcement learning with Bayesian optimization. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1557–1563. IEEE, 2017.
- William C Martin, Albert Wu, and Hartmut Geyer. Robust spring mass model running for a physical bipedal robot. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6307–6312. IEEE, 2015.
- Ruben Martinez-Cantin. Funneled Bayesian optimization for design, tuning and control of autonomous systems. *IEEE transactions on cybernetics*, (99):1–12, 2018.
- J Mockus, V Tiesis, and A Zilinskas. Toward Global Optimization, Volume 2, Chapter: Bayesian Methods for Seeking the Extremum. 1978.
- Xue Bin Peng, Glen Berseth, and Michiel van de Panne. Terrain-adaptive locomotion skills using deep reinforcement learning. *ACM Transactions on Graphics (TOG)*, 35(4):81, 2016.
- Akshara Rai, Rika Antonova, Seungmoon Song, William Martin, Hartmut Geyer, and Christopher Atkeson. Bayesian optimization using domain knowledge on the ATRIAS biped. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1771–1778. IEEE, 2018.
- Carl Edward Rasmussen and Hannes Nickisch. Gaussian processes for machine learning (gpml) toolbox. *J. Mach. Learn. Res.*, 11:3011–3015, December 2010. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=1756006.1953029>.

- Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando de Freitas. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.
- Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems (NIPS)*, pages 2951–2959, 2012.
- Jasper Snoek, Kevin Swersky, Rich Zemel, and Ryan Adams. Input warping for Bayesian optimization of non-stationary functions. In *International Conference on Machine Learning (ICML)*, pages 1674–1682, 2014.
- Seungmoon Song and Hartmut Geyer. A Neural Circuitry that Emphasizes Spinal Feedback Generates Diverse Behaviours of Human Locomotion. *The Journal of Physiology*, 593(16):3493–3511, 2015.
- Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *International Conference on Machine Learning (ICML)*, pages 1015–1022. Omnipress, 2010.
- Matthew Tesch, Jeff Schneider, and Howie Choset. Using response surfaces and expected improvement to optimize snake robot gait parameters. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1069–1074. IEEE, 2011.
- Nitish Thatte and Hartmut Geyer. Toward Balance Recovery with Leg Prostheses Using Neuromuscular Model Control. *IEEE Transactions on Biomedical Engineering*, 63(5):904–913, 2016.
- Evangelos Theodorou, Jonas Buchli, and Stefan Schaal. A generalized path integral control approach to reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 11 (Nov):3137–3181, 2010.
- Nicolas Van der Noot, Luca Colasanto, Allan Barrea, Jesse van den Kieboom, Renaud Ronsse, and Auke J Ijspeert. Experimental validation of a bio-inspired controller for dynamic walking with a humanoid robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 393–400. IEEE, 2015.
- Aaron Wilson, Alan Fern, and Prasad Tadepalli. Using Trajectory Data to Improve Bayesian Optimization for Reinforcement Learning. *The Journal of Machine Learning Research (JMLR)*, 15(1):253–282, 2014.
- Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378, 2016.
- DA Winter and HJ Yack. EMG profiles during normal human walking: stride-to-stride and inter-subject variability. *Electroencephalography and clinical neurophysiology*, 67(5):402–411, 1987.