# Combining Reinforcement Learning and Trajectory Optimization for Multi-Contact Motion Planning & Control of Quadrupedal Locomotion

Vassilios Tsounis, Ruben Grandia, Farbod Farshidian, and Marco Hutter

***Abstract*—This work addresses the problem of multi-contact motion planning for quadrupedal legged robots on non-flat terrain. A common challenge in this domain is the selection of a contact schedule (i.e. contact configurations and contact switching event times) and respective sequences of foothold positions. Our approach is centered on formulating Markov decision processes using the evaluation of kinematic and dynamic feasibility criteria in the form of linear programs and are used in place of physical simulation. The resulting MDPs are solved using a policy-gradient-based reinforcement learning algorithm. Specifically, we train neural-network policies which generate reference foothold positions, Center-of-Mass poses and velocities as well contact switching timings. We evaluate our method in unstructured environments using proprioceptive and exteroceptive sensory input and on a suite of relevant terrain scenarios such as stairs, gaps and stepping stones.**

***Index Terms*—Legged Robots; Deep Learning in Robotics and Automation; Motion and Path Planning; Reinforcement Learning; Quadrupedal Robots;**

***Paper Type* – Original Work**

## I. Introduction

In recent years, significant progress has been made in the development of techniques for solving the problem of perceptive locomotion on unstructured terrain for legged robots. Operating autonomously in such environments requires addressing the problem of multi-contact motion planning and control.

This work deals specifically with foothold and gait planning for terrain-ware quadrupedal locomotion on rigid non-flat terrain. Given a certain parameterization of the terrain, the objective is to plan when and where to establish contact between the robot's end-effectors and the terrain and then subsequently generate motion trajectories for the swing-feet and torso. The challenge here, however, lies in that foothold selection is directly tied to the selection of the gait, i.e. contact configuration and contact switching event times. As the gait can heavily impact the overall locomotion performance, selecting one that is suitable for the coincident terrain, and with least modelling assumptions as possible, is of paramount importance if legged robots such as ANYmal [1] are to operate autonomously in complex environments.

Solving for both gait and footholds simultaneously necessitates performing hybrid continuous-discrete optimizations, which, can become prohibitively computationally expensive for executing on-line and on-board during operation. Moreover, such transcriptions of such problems must also incorporate elements which ensure the selected footholds and respective feet and torso trajectories are both kinematically and dynamically feasible.

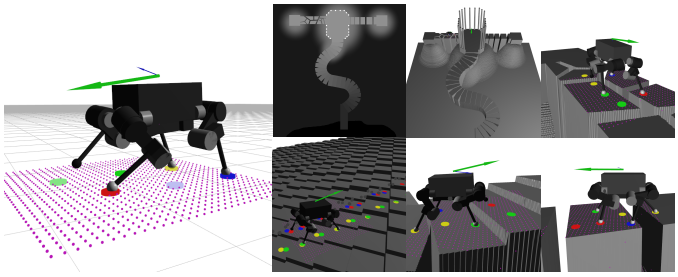

Fig. 1. The suite of terrains: the baseline Flat-World scenario (left), the Random-Stairs scenario (bottom center), and composite Temple-Ascent (right) scenario comprising a set of winding stairs and two derelict bridges containing stepping-stones and gaps of varying size.

Within present literature, two broad families of approaches have become most prominent: 1) Model-based Trajectory Optimization (TO) and 2) Deep Reinforcement Learning (DRL). Indeed, past works addressing terrain-aware locomotion problems have predominantly used model-based approaches, such as those employing deterministic optimization techniques [2], [3], in conjunction with other heuristics [4], to plan motions for both the base and feet. However newer trends employing DRL [5], [6] tend to relax many of the modelling assumptions made but require accurate physical simulation in order to transfer well to real systems.

We propose a new method that combines state-of-the-art model-based TO and model-free DRL methods to enable quadrupedal systems to traverse complex non-flat terrain. Our formulation consists of a terrain-aware Gait Planner (GP) that generates sequences of footholds, gait parameters and base motions that direct the robot towards a target heading. The GP is realized as stochastic policy parameterized using Neural-Network (NN) function approximation, and policy search is performed using state-of-the-art DRL algorithms.

*Contributions:* a) We introduce a novel method for training kinodynamic motion planners, which employs a Trajectory Optimization (TO) technique for determining so-called *transition feasibility* between discrete support phases using a coarse model of the robot. This removes the need for a planner to interact with both physics and a motion controller during training, allows the two policies to be trained independently, and leads to a significant reduction in overall sample complexity. b) We present results from combining our GP with a state-of-the-art Nonlinear Model Predictive Controller (NMPC) and

Whole Body Controller (WBC) and demonstrate sim-to-real transfer of both elements on a real quadrupedal robot.

This workshop paper partially summarizes the work presented in Tsounis et al [7]. However, several extensions to the original work have been made in order to explicitly account for 3D terrain such as stairs etc, as well as use NMPC instead of RL for realizing the motion planning and control parts. Please refer to the former for further details regarding all technical pertaining to the GP that are not included here.

## II. Methodology

The GP serves as a local terrain-aware planner, and uses both *exteroceptive* and *proprioceptive* measurements to generate a finite sequence of support phases, i.e. a *phase plan*. It operates by sequentially querying the planning policy $\pi_{\boldsymbol{\theta}_P}$ to generate the aforementioned phase plan. We thus formulate an MDP in order to train $\pi_{\boldsymbol{\theta}_P}$ using DRL, and our objective is to ensure that the resulting policy learns to respect the kinodynamic properties and limits of the robot, as well as contact constraints, when proposing phase transitions. Moreover, we aim to avoid direct interaction with the physics of the system, and instead craft the transition dynamics of the MDP by employing a *transition feasibility* criterion realized as a TO problem using the frameworks defined in [8] and [9]. Lastly, we avoid explicitly modeling or qualifying the terrain, as done in [4], [10], and instead directly use measurements of local terrain elevation. The resulting MDP, allows us to train $\pi_{\boldsymbol{\theta}_P}$ to infer a distribution over phase transitions. The high-level command to the system is provided as a desired Cartesian pose and desired step length.

*Support Phases:* In order to reason precisely about gaits and transitions between contact supports, we define a parameterization thereof that encompasses all necessary information. We thus parameterize a gait as a sequence of so-called *support phases*. Each phase is defined by the tuple

$$\boldsymbol{\Phi} \coloneqq \langle \mathbf{R}_B,\, \mathbf{r}_B,\, \mathbf{v}_B,\, \mathbf{r}_F,\, \mathbf{c}_F,\, t_S \rangle \;\in \Phi \tag{1}$$

where $\mathbf{c}_F \in \{0,1\}^4$ is a vector indicating for each of the feet a closed, 1, or open, 0, contact w.r.t the terrain, $\mathbf{r}_F \in \mathbb{R}^{3\times 4}$ are the stacked absolute positions of the feet, and $t_S \in \mathbb{R}_{\geq 0}$ defines the time at which the switch to the contact configuration of the respective phase has occurred.

*Support Phase Transition Feasibility:* Transition feasibility amounts to evaluating if a feasible motion exists between a pair of support phases $\boldsymbol{\Phi}_t, \boldsymbol{\Phi}^*_{t+1}$, where the former is assumed while the latter is a candidate successor. In our previous work in Tsounis et al [7] we employed the general framework defined in [8] to design a convex LP using the Convex Resolution Of Centroidal dynamics trajectories (CROC) formulation. We use a specific variant of CROC to derive a set of linear equality and inequality constraints, a trivial cost, a time-discretization of the CoM trajectory, and incorporates the parameterization of the contact forces into the decision variables of the optimization. However, one limitation of using CROC as the feasibility TO is that only static gaits where able to be generated. In order to overcome this, we also implemented the convex LP described in Dai et al [9],

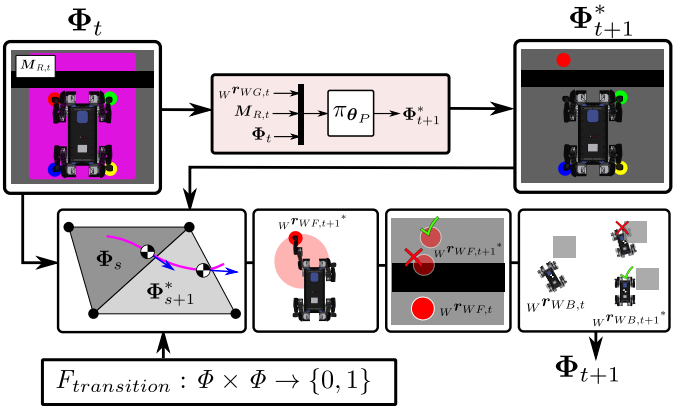


Fig. 2. An overview of the transition dynamics of the MDP used to train the GP policy. Both states and actions of the MDP are defined as stance phases $\boldsymbol{\Phi}_t$ and transitions proposed by the policy are evaluated only on their feasibility given the TO scheme as well as checks for kinematic limits, body/torso collisions and validity of footholds.

which, allows for the inclusion of the angular momentum in the decision variables whilst also directly parameterizing the centroid's linear and angular trajectories using a uniform time discretization instead of employing Bezier curves. Thus, this alternate solver, henceforth referred as Dai's LP, allows for the generation of dynamic transitions and gaits. We defer the reader to [8] for further details regarding CROC and [9] for Dai's LPs respectively. The resulting formulation, either using CROC or Dai's LP, allows us to realize the transition feasibility mapping $F_{feasibility} : \Phi \times \Phi \rightarrow \{0,1\}$. Therefore, we evaluate the LP for pairs of phases $\boldsymbol{\Phi}_t, \boldsymbol{\Phi}^*_{t+1}$, to determine if the corresponding phase transition is feasible, 1, or not, 0.

*Transition Dynamics:* We define state transition dynamics for this MDP employing a formalism defining so-called *termination condition* functions $T(\boldsymbol{s}_{P,t}, \boldsymbol{a}_{P,t}, \boldsymbol{s}_{P,t+1})$, which determine if an episode terminates. By formulating an episode termination as a transition into an absorbing terminal state, we can say that, an episode under this MDP, terminates whenever $\boldsymbol{s}_{P,s+t} = \boldsymbol{s}_{P,s}$, $\forall t > 0$. In this MDP, in particular, we employ the following termination conditions:

1) $T_{footholds}$: Checks for obstacles or gaps within the vicinity of each foothold using an fixed eight-point grid surrounding each foot.
2) $T_{collisions}$: Checks for external collisions between the base and terrain as well as internal ones between the footholds and the base.
3) $T_{kinematics}$: Checks if new candidate footholds are kinematically reachable for the given joint morphology.
4) $T_{feasibility}$: Evaluates $F_{feasibility}(\boldsymbol{\Phi}_t, \boldsymbol{\Phi}^*_{t+1})$.

Thus, each step of this MDP proceeds as follows: (1) Given a state $\boldsymbol{s}_{P,t}$, the MDP computes the corresponding observation $\boldsymbol{o}_{P,t}$ and is passed to the agent to select an appropriate action according to $\pi_{\boldsymbol{\theta}_P}$. (2) The selected action $\boldsymbol{a}_{P,t}$, is used to compute the candidate phase $\boldsymbol{\Phi}^*_{t+1}$. (3) The aforementioned termination conditions are used to assert if the phase transition is feasible. This formulation therefore allows the agent to propose the phase transition directly, while the MDP only checks if it is feasible and otherwise terminates the episode.

Fig. 2 provides an overview of the transition dynamics.

*Policy Definition:* We parameterize the GP's policy as a Gaussian distribution with a diagonal covariance matrix. The mean is output by a NN which inputs both exteroceptive and proprioceptive measurements into a series of NN layers, similar to those proposed in [6]. The policy is trained $\pi_{\boldsymbol{\theta}_P}$ with a variant of Proximal Policy Optimization (PPO) using clipped loss and a Generalized Advantage Estimation (GAE) critic [11].

*Policy Evaluation:* In order to evaluate our approach, we crafted a suit of terrain scenarios for training and testing the GP policies, as depicted in Fig. 1. The first and most basic scenario consists of an infinite flat plane we refer to as *Flat-World*, which we use to establish a baseline for performance and behavior. Secondly, the *Random-Stairs* terrain presents a $20 \times 20\,\mathrm{m}^2$ square area consisting of $1 \times 1\,\mathrm{m}^2$ flat regions of randomly selected elevation. The elevation changes were generated in a way that results in an effective inclination diagonally across the map. The third terrain scenario is that which we call *Temple-Ascent*, and is a composite terrain consisting of gaps, stepping stones, stairs as well as flat regions.

## III. Motion Planning & Control

In Tsounis et al [7], we realized the generation and execution of motion using an DRL-based neural-network Gait Control (GC) policy. However, to demonstrate the truly modular and decoupled nature of our GP policy and respective training thereof, we instead use the combination of state-of-the-art NMPC and WBC modules as described in [12] and [13] respectively.

In brief, the setup works as follows: The GP is first queried recursively on its own output however many times until the generated phase sequence fills the time horizon of the NMPC. The generated contact configurations and foothold positions are provided to the NMPC, which, subsequently computes and generates optimized trajectories for the base, swing-feet, and end-effector contact forces at 20Hz. All quantities from the head of each trajectory are provided to the WBC in order to compute optimized joint torques as well as respective joint position and velocity references at 400Hz. These final joint references are the ones transmitted to the actuators.

Essentially, we have only replaced the Raibert-like foothold heuristic used in [12] with the footholds generated by the GP. In fact, the NMPC and WBC were used mostly unmodified w.r.t. how they were used in [12] and [13], and only a mere handful of parameters affected by the gait type were tuned. Fig. 3 provides a visual depiction of preliminary experiments conducted on the quadrupedal robot ANYmal overcoming a test obstacle.

## IV. Discussion

In this short summary of our most recent and ongoing work, we have presented a general framework for training neural network policies for centroidal pose and foothold planning for quadrupedal locomotion on unstructured terrain.

Fig. 3. Experimental verification of the GP+NMPC+WBC combination performed on the quadrupedal robot ANYmal in a laboratory setting. All frames are extracted from a single continuous trial, however, the left column corresponds to first stepping up onto the obstacle, while the right column to the respective descent.

Our experimental verification[1] on a real robotic platform, thus far however, only includes results from training GPs only using the CROC LP, whilst extending these to dynamic gaits using Dai's LP is part of ongoing work and only works in simulation. Moreover, one notable difference to our previous work in Tsounis et al [7] is the use of NMPC+WBC in place of an RL-based GC. Work on the latter has been attempted, but preliminary results did not prove promising enough to justify pursuing this direction further at present time. Our main objectives in extending the work in [7] are to: a) extend to 3D environments such as stairs, b) experimentally verify sim-to-real transfer of GP policies, and c) overcome the limitations

[1] https://photos.app.goo.gl/1kYhyZWQvussAmQC6

of CROC in only being able to generate static gaits. The use of NMPC+WBC or RL-based for motion control exists as a separate and independent research direction which is to be pursued in future work. We thus view the decoupling of GP and GC realizations as an advantage, in that the former has the potential to be used in applications where the use of neural-networks for motion control might indeed prove to be unnecessary.

## REFERENCES


[1] M. Hutter, C. Gehring, D. Jud, A. Lauber, C. D. Bellicoso, V. Tsounis, J. Hwangbo, K. Bodie, P. Fankhauser, M. Bloesch *et al.*, "ANYmal-a highly mobile and dynamic quadrupedal robot," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*. IEEE, 2016, pp. 38–44.

[2] S. Kuindersma, R. Deits, M. Fallon, A. Valenzuela, H. Dai, F. Permenter, T. Koolen, P. Marion, and R. Tedrake, "Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot," *Autonomous Robots*, pp. 429–455, 2016.

[3] A. W. Winkler, C. D. Bellicoso, M. Hutter, and J. Buchli, "Gait and trajectory optimization for legged systems through phase-based end-effector parameterization," *IEEE Robotics and Automation Letters*, pp. 1560–1567, 2018.

[4] P. Fankhauser, M. Bjelonic, C. D. Bellicoso, T. Miki, and M. Hutter, "Robust rough-terrain locomotion with a quadrupedal robot," in *IEEE Int. Conf. on Robotics and Automation*. IEEE, 2018, pp. 1–8.

[5] N. Heess, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. Eslami, M. Riedmiller, and D. Silver, "Emergence of locomotion behaviours in rich environments," *arXiv preprint arXiv:1707.02286*, 2017.

[6] X. B. Peng, G. Berseth, K. Yin, and M. Van De Panne, "DeepLoco: Dynamic Locomotion Skills Using Hierarchical Deep Reinforcement Learning," *ACM Trans. Graph.*, pp. 41:1–41:13, 2017.

[7] V. Tsounis, M. Alge, J. Lee, F. Farshidian, and M. Hutter, "DeepGait: Planning and Control of Quadrupedal Gaits using Deep Reinforcement Learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3699–3706, 2020.

[8] P. Fernbach, S. Tonneau, O. Stasse, J. Carpentier, and M. Taix, "C-CROC: Continuous and Convex Resolution of Centroidal Dynamic Trajectories for Legged Robots in Multicontact Scenarios," *IEEE Transactions on Robotics*, 2020.

[9] H. Dai and R. Tedrake, "Planning robust walking motion on uneven terrain via convex optimization," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 579–586.

[10] S. Tonneau, A. Del Prete, J. Pettré, C. Park, D. Manocha, and N. Mansard, "An efficient acyclic contact planner for multiped robots," *IEEE Transactions on Robotics*, vol. 34, no. 3, pp. 586–601, 2018.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[12] R. Grandia, F. Jenelten, S. Yang, F. Farshidian, and M. Hutter, "Perceptive locomotion through nonlinear model predictive control," *(submitted to) IEEE Transactions on Robotics*, 2022.

[13] F. Jenelten, R. Grandia, F. Farshidian, and M. Hutter, "Tamols: Terrain-aware motion optimization for legged systems," *(submitted to) IEEE Transactions on Robotics*, 2021.