

研究报告

题目：足式机器人强化学习控制调研

研 究 生 彭道杰

摘 要

足式机器人相对于常见的轮式和履带式机器人有着更加灵活，环境适应能力强的优势。与启发它们的人类或者动物一样，这类机器人拥有着更强的越野能力，能够适应复杂的地形环境，到达普通轮式或履带式机器人无法到达的地方。然而，这类机器人往往具有更多的自由度，这使得对它们的控制变得具有挑战性。如何让这类足式机器人像它们的动物模仿对象一样能够灵活、优雅地运动成为一个重要的研究课题。在这其中，模仿四足动物的四足机器人（机械狗）是一类被广泛研究的对象。

机械狗是一个多自由度运动系统，其典型自由度数量为 12 个。每条腿有三个自由度分别实现小腿伸缩、大腿伸缩和整腿伸缩。由于它的求解空间巨大而可用的解空间相对较小，因此机械狗的运动控制常常是启发式的——预先定义一些运动，以缩小求解时的搜索空间。经典的基于运动动力学建模和优化问题的控制方式主要有轨迹优化 (*Trajectory Optimization, TO*) 和模型预测控制 (*Model Predictive Control, MPC*)。由于求解需要涉及大量的非线性优化问题，这类控制方法求解算力开销较大往往不能够在线完成，需要上位机的辅助。相比较而言近年来发展起来的基于深度学习的强化学习控制方法在运行期间的算力开销很小。强化学习控制是一种通过构建神经网络并依靠大量训练优化神经网络参数以得到预期的控制效果的控制方法，可以完全不需要动力学建模。其训练过程往往在仿真环境中完成，因此需要对机械狗的物理模型进行建模，这些建模包括：机械狗的躯干及四肢的刚体模型、电机驱动关节的响应模型。

不过由于神经网络的黑盒性，单纯的强化学习控制往往只能针对特定的训练目标而很难与其它模型融合，缺乏跨模式泛化能力。因此经典控制和强化学习控制的融合控制变得十分重要。

关键词：机器人控制；轨迹优化；深度学习；强化学习控制；Isaac 仿真

目 录

摘 要	I
第 1 章 基于质心动力学 (Centroidal Dynamics) 模型的 TO 控制	1
1.1 机械狗的运动模型 (Motion Model)	1
1.2 分层次优化 (Hierarchical Optimization) 计算	3
1.2.1 基体运动方程	3
1.2.2 地面接触部分运动约束	3
1.2.3 接触力限制	4
1.2.4 扭矩限制	4
1.2.5 目标运动跟随控制	4
1.2.6 接触力最小化	5
1.2.7 根据优化结果计算电机扭矩	5
1.3 质心运动优化 (Motion Optimization)	5
1.3.1 足态保持生成	5
1.3.2 支撑多边形序列生成	6
1.3.3 质心运动优化问题描述	6
1.3.4 规划初始化	7
1.4 质心运动优化 QP 问题的建立	8
1.4.1 成本函数	8
1.4.2 等式约束	10
1.4.3 不等式约束	11
1.4.4 约束松弛	11
第 2 章 基于直接刚体动力学模型的 TO 控制	13
2.1 基于直接刚体动力学模型的 TO 控制	13
第 3 章 基于 MPC 的控制	14
3.1 基于 MPC 的控制	14
第 4 章 基于强化学习的机械狗控制	15
4.1 强化学习控制	15
4.1.1 强化学习的基本概念	15
4.1.2 基于深度强化学习的控制器的部署和优化流程	15

4.2	单层神经网络直接驱动关节方式	16
4.3	双层神经网络本体感知方式	17
4.3.1	总体概况	17
4.3.2	策略训练	18
4.3.3	地形定义	19
4.3.4	控制架构	19
4.4	双层网络本体和外部感知融合方式	20
4.4.1	问题描述	22
4.4.2	教师策略训练	22
4.4.3	观测和行动	23
4.4.4	策略构架	24
4.4.5	奖励函数	25
4.4.6	课程	28
4.4.7	学生策略训练	28
4.4.8	高度采样随机化	29
4.4.9	信念状态寄存器	30
4.4.10	部署	31
4.5	电机驱动关节点模型	31
第 5 章	强化学习部署实践和关键点	32
5.1	强化学习平台 Isaac	32
5.2	强化学习案例 legged_gym 库	32
5.3	PyTorch 的使用	32
第 6 章	几种构想的机器人类型草图	33
6.1	弹跳轮足 + 机械手机器人	33
6.2	轮鞋足式 + 机械手机器人	34
6.3	弹跳轮足 + 飞行器机器人	34
结 论		36
参考文献		37

第 1 章 基于质心动力学 (Centroidal Dynamics) 模型的 TO 控制

足式机器人的运动规划是比较困难的，不仅因为它的自由度较多，更因为它的机体运动不能被直接得出，而是要通过四肢状态及四肢与环境的接触产生。

经典控制的基础是对机器人的运动学和动力学建模。经典控制主要方式有轨迹优化 (Trajectory Optimization, TO) 和模型预测控制 (Model Predictive Control, MPC) 两类。常用的轨迹优化机械狗动力模型有两种^{[1]p2}：1. 基于质心动力学的模型；2. 基于直接刚体动力学的模型。基于这些模型的优化控制都被描述为非线性优化问题，这些问题的求解算力开销较大往往不能够在线完成，需要上位机的辅助。

1.1 机械狗的运动模型 (Motion Model)^{[2-3]p2}

机器人与环境接触的机械系统的运动模型描述方程可以描述如下：

$$\mathbf{M}(\mathbf{q})\dot{\mathbf{u}} + \mathbf{h}(\mathbf{q}, \mathbf{u}) = \mathbf{S}^T \boldsymbol{\tau} + \mathbf{J}_S^T \boldsymbol{\lambda} \quad (1-1)$$

这其中 \mathbf{q} 是一个描述机器人主体及各个节点的广义位置矢量：

$$\mathbf{q} = \begin{bmatrix} {}_I \mathbf{r}_{IB} \\ \mathbf{q}_{IB} \\ \mathbf{q}_j \end{bmatrix} \in SE(3) \times \mathbb{R}^{n_j} \quad (1-2)$$

它里面 ${}_I \mathbf{r}_{IB} \in \mathbb{R}^3$ 是机器人主体相对于惯性系的三维位置矢量； $\mathbf{q}_{IB} \in SO(3)$ 是机器人主体相对于惯性系的转动描述，用哈密顿单位四元数表示的； $\mathbf{q}_j \in \mathbb{R}^{n_j}$ 是一个储存机器人所有节点角度的 n_j 维矢量。

这其中 \mathbf{u} 是一个描述机器人主体及各个节点的广义速度矢量：

$$\mathbf{u} = \begin{bmatrix} {}_I \mathbf{v}_B \\ {}_B \boldsymbol{\omega}_{IB} \\ \dot{\mathbf{q}}_j \end{bmatrix} \in \mathbb{R}^{n_u} \quad (1-3)$$

它里面的 ${}_I \mathbf{v}_B$ 描述了机器人主体相对于惯性系的速度； ${}_B \boldsymbol{\omega}_{IB}$ 描述了机器人主体相对于自身的角速度； $\dot{\mathbf{q}}_j$ 描述了机器人的各个关节转动的速度。这其中 \mathbf{M} 是一

个关于机器人整体的质量矩阵，它是一个 $n_u \times n_u$ 的矩阵：

$$\mathbf{M} \in \mathbb{R}^{n_u \times n_u} \quad (1-4)$$

这个质量矩阵的具体数值跟机器人的机械系统的状态（各个节点的位置 \mathbf{q} ）相关，可以通过通用的方式计算出它的表达式。实际计算的时候只需要带入 \mathbf{q} 的值，就可以计算出 \mathbf{M} 矩阵的各个元素具体数值。这其中 \mathbf{h} 是一个跟机器人的机械位置和速度都有关的量，包含了机械系统产生的克里奥利力、离心力和重力的作用，它是一个 n_u 维的矢量：

$$\mathbf{h} \in \mathbb{R}^{n_u} \quad (1-5)$$

这其中 \mathbf{S} 是一个选择矩阵，可以用来选择整个公式中哪些自由度被激活，它是一个 $n_\tau \times n_u$ 的矩阵：

$$\mathbf{S} = \begin{bmatrix} \mathbf{0}_{n_\tau \times (n_u - n_\tau)} & \mathbf{I}_{n_\tau \times n_\tau} \end{bmatrix} \in \mathbb{R}^{n_\tau \times n_u} \quad (1-6)$$

它其中包含的参数 n_τ 表示被激活的自由度数量，如果机器人的所有自由度都被激活，则 $n_\tau = n_j$ 。这其中 $\boldsymbol{\tau}$ 是机器人各个关节的电机提供的扭矩，它是一个 n_j 维的向量：

$$\boldsymbol{\tau} \in \mathbb{R}^{n_j} \quad (1-7)$$

它的节点成员是否产生作用受到 \mathbf{S}^T 矩阵的选择。这其中 \mathbf{J}_S 是一些列雅可比矩阵的集合：

$$\mathbf{J}_S = \begin{bmatrix} \mathbf{J}_{C_1}^T & \dots & \mathbf{J}_{C_{n_c}}^T \end{bmatrix}^T \in \mathbb{R}^{3n_c \times n_u} \quad (1-8)$$

它是接触点的支撑力 $\boldsymbol{\lambda}$ 向节点力转换的矩雅可比矩阵，包含了 n_c 个雅可比矩阵， n_c 为接触地面的肢体个数。

如果是接触建模为点接触的话，在稳定接触的情况下，每个接触点会产生三个相应的约束条件：

$$\begin{cases} {}_I\mathbf{r}_{IC}(t) = \text{const} \\ {}_I\dot{\mathbf{r}}_{IC} = \mathbf{J}_C \mathbf{u} = 0 \\ {}_I\ddot{\mathbf{r}}_{IC} = \mathbf{J}_C \dot{\mathbf{u}} + \dot{\mathbf{J}}_C \mathbf{u} = 0 \end{cases} \quad (1-9)$$

其含义就是在惯性系下看接触点的位置是固定的为定值，并且其速度和加速，也即其一阶导数和二阶导数为零。

1.2 分层次优化 (Hierarchical Optimization) 计算^{[2]p2}

整个优化计算过程的最终目的是得到一个扭矩参考目标 ξ_d 用来作为输入控制电机， ξ_d 包含两部分信息：关节点的运动加速度 $\dot{\mathbf{u}}_d^T$ 和接触力 λ_d^T 。控制方法采用接触力控制，优化化计算的策略是采取分层次优化的方式。在这一部分我们将从1.1节中描述的运动模型出发，拆接触其中的内容进行分层次优化。我们将介绍运动方程、接触部分的运动约束、接触力和扭矩限制、运动跟随、接触力最小化、根据优化结果计算电机扭矩等内容。

1.2.1 基体运动方程

通过利用选择矩阵 \mathbf{S}^T 诱导的分解，我们可以将运动和接触力限制在浮动基系统动力学描述的流形上，有如下式：

$$\begin{bmatrix} \mathbf{M}_{fb} & -\mathbf{J}_{sfb}^T \end{bmatrix} \xi_d = -\mathbf{h}_{fb} \quad (1-10)$$

这里 \mathbf{M}_{fb} : 下标的意思是浮动的主体-'floating base'，各个参数的具体含义解释如下：

- (1) ξ_d 是一个一个 $n_u + n_c$ 行 1 列的向量: $\xi_d = \begin{bmatrix} \mathbf{u}_d^T & \lambda_d^T \end{bmatrix} \in \mathbb{R}^{n_u+n_c}$ ，其中：
 - ① \mathbf{u}_d^T 是目标关节点的加速度；
 - ② λ_d^T 是目标接触力；
- (2) \mathbf{M}_{fb} 是复合型惯性矩阵的前六行；
- (3) \mathbf{J}_{sfb}^T 是雅克比阵的前六行，它将接触力转换到节点的扭矩；
- (4) \mathbf{h}_{fb} 是非线性项的前六行，包括克里奥利力、离心力和重力项；

1.2.2 地面接触部分运动约束

控制器找到的解决方案不应违反(1-9)中定义的接触约束。因此，我们通过设置在接触点施加空加速度：

$$\begin{bmatrix} \mathbf{J}_s & \mathbf{0}_{3n_c \times 3n_c} \end{bmatrix} \xi_d = -\dot{\mathbf{J}}_s \mathbf{u} \quad (1-11)$$

它将 ξ_d 带进去计算后就得到了: $\mathbf{J}_s \xi_d = -\dot{\mathbf{J}}_s \mathbf{u}$ 其实就是公式(1-9)的第二个式子 $\mathbf{J}_s \xi_d + \dot{\mathbf{J}}_s \mathbf{u} = 0$ 。

1.2.3 接触力限制

$$\begin{cases} ({}_I\mathbf{h} - {}_I\mathbf{n}_\mu)^T {}_I\boldsymbol{\lambda}_k \leq 0 \\ -({}_I\mathbf{h} + {}_I\mathbf{n}_\mu)^T {}_I\boldsymbol{\lambda}_k \leq 0 \\ ({}_I\mathbf{l} - {}_I\mathbf{n}_\mu)^T {}_I\boldsymbol{\lambda}_k \leq 0 \\ -({}_I\mathbf{l} + {}_I\mathbf{n}_\mu)^T {}_I\boldsymbol{\lambda}_k \leq 0 \end{cases} \quad (1-12)$$

${}_I\mathbf{n}$ 是接触面的法向量； μ 是摩擦系数；有了这两个再乘上受力 ${}_I\boldsymbol{\lambda}$ ，就可以得出最大静摩擦力 ${}_I\mathbf{n}_\mu^T \boldsymbol{\lambda}$ 。实际在接触点平行于地面方向的两个分力 ${}_I\mathbf{h}$ ， ${}_I\mathbf{l}$ 在正反方向上都不应该比这个值大，否则就会发生滑动。这就是公式(1-12)约束的由来。

1.2.4 扭矩限制

对于扭矩有如下约束：

$$\boldsymbol{\tau}_{min} - \mathbf{h}_j \leq [\mathbf{M}_j \quad -\mathbf{J}_{s_j}^T] \leq \boldsymbol{\tau}_{max} - \mathbf{h}_j \quad (1-13)$$

这里的 \mathbf{h}_j 就是科里奥利力那一堆东西。而 $[\mathbf{M}_j \quad -\mathbf{J}_{s_j}^T]$ 就是加速度力 $\mathbf{M}_j \dot{\mathbf{u}}_d^T$ 和传递力 $-\mathbf{J}_{s_j}^T \boldsymbol{\lambda}_d^T$ 两项的和。它们计算的结果就是电机提供的扭矩力 $\boldsymbol{\tau}$ 克服完 $\mathbf{h}(\mathbf{q}, \mathbf{u})$ 剩下的力。

1.2.5 目标运动跟随控制

为了能跟随浮动主体和摆动腿的目标运动。我们通过实现具有前馈参考加速度和运动相关状态反馈状态的操作空间控制器来约束关节加速度。对于主体的线性运动：

$$[{}_c\mathbf{J}_P \quad 0] \boldsymbol{\xi}_d = {}_c \ddot{\mathbf{r}}_{IB}^d + \mathbf{k}_D^{pos} ({}_c \dot{\mathbf{r}}_{IB}^d - {}_c \mathbf{v}_B) + \mathbf{k}_P^{pos} ({}_c \mathbf{r}_{IB}^d - {}_c \mathbf{r}_B) \quad (1-14)$$

对于主体的角度运动：

$$[{}_c\mathbf{J}_R \quad 0] \boldsymbol{\xi}_d = -\mathbf{k}_D^{ang} {}_c \boldsymbol{\omega}_B + \mathbf{k}_P^{ang} (\mathbf{q}_{CB}^d \boldsymbol{\varepsilon} \mathbf{q}_{CB}) \quad (1-15)$$

具体参数解释如下：

- 雅可比矩阵 ${}_c\mathbf{J}_P$ 和 ${}_c\mathbf{J}_R$ 是与控制坐标系 C （这是一个与地形局部估计和机器人航向方向对齐的帧）中表达的基相关的平移和旋转雅可比矩阵。
- $\boldsymbol{\varepsilon}$ 这个算子产生欧拉向量，表示期望姿态 \mathbf{q}_{CB}^d 和估计姿态 \mathbf{q}_{CB} 之间的相对方向。
- 这里面 $\mathbf{k}_P^{pos}, \mathbf{k}_D^{pos}, \mathbf{k}_P^{ang}, \mathbf{k}_D^{ang}$ 是用来控制增益的对角正定矩阵。
- 参考的运动 ${}_c \mathbf{r}_{IB}$ 和它的导数是运动规划的结果。

1.2.6 接触力最小化

可以通过下面的方法将接触力设置为最小值：

$$\begin{bmatrix} 0_{3n_c \times n_u} & \mathbb{I}_{3n_c \times 3n_c} \end{bmatrix} \xi_d = 0 \quad (1-16)$$

这个式子的含义是让每个接触力的目标值都为零。

1.2.7 根据优化结果计算电机扭矩

如果给定了一个优化的关节运动加速度和接触力， $\xi_d = [\dot{\mathbf{u}}_d^T \quad \lambda_d^T]^T$ 我们可以用以下公式计算各个电机的扭矩：

$$\boldsymbol{\tau}_d = \begin{bmatrix} \mathbf{M}_j & -\mathbf{J}_{s_j}^T \end{bmatrix} \xi_d + \mathbf{h}_j$$

其中 \mathbf{M}_j , $-\mathbf{J}_{s_j}^T$, \mathbf{h}_j 在公式(1-13)中定义。

因此，所有规划的目的就是给出关节运动加速度 $\dot{\mathbf{u}}_d^T$ 和接触力 λ_d^T 目标，也即 ξ_d^* ，从它出发直接计算得到控制各个关节电机的扭矩目标。

1.3 质心运动优化 (Motion Optimization)^{[2]p3}

在机械狗的运动模型描述1.1节部分我们定义了一些关键参数，拥有了对机械狗运动的描述；在分层次优化1.2节部分我们分解了运动方程的成分并给出了一些硬性约束条件和底层控制方法，知道了如何从关节运动加速度的目标值和接触力的目标值这两个变量得到对电机的扭矩 $\boldsymbol{\tau}_d$ 的控制目标值。下面就要进行实际的运动优化问题描述和求解，来计算出关节运动加速度的目标值和接触力的目标值。

整个机器人的控制遵循一下流程：首先由预定义的方式设定机器人应该遵循的步态1.3.1节；为了保持机器人的稳定性，这些步态的生成需要满足 ZMP 约束，也即机械狗的步态是由一系列的支撑多边形序列具体定义的1.3.2节；在机器人获得稳定支撑的基础上，我们根据外部的指令对机器人的具体运动方式调整以获得指令要求的目标运动，这被描述为一个运动优化问题1.3.3；在优化前首先要对初始状态进行规划初始化1.3.4节；优化过程采用质心运动优化的方式进行，这是一个二次优化问题，它的具体建立在1.4节给出。

1.3.1 足态保持生成

外部高级速度命令用于通过调整参考立足点以特定方向和速度驱动运动，这些立足点是针对每个新控制回路的每条腿计算的。根据腿的接触状态，这些是通过两种不同的方式计算的：当腿接触时，命令速度将被投影到计算的立足点位置，

以便平均躯干速度与所需的运动速度相匹配；当腿摆动时，参考立足点以相同的方式计算，但添加了一个速度反馈项，该项用来在机器人受到会导致速度控制误差的外部推力时稳定机器人。相关细节参考文献^[4]。

这块儿随后要看看，补充到下面...

1.3.2 支撑多边形序列生成

我们在相域中 (phase domain) 为每条腿定义升降事件 ϕ_{lo} 和触下事件 ϕ_{td} 。这也定义了所有腿的接触时间表。给定所有腿的触下和升降事件，以及当前脚位置和所需立足点的集合，我们可以计算一系列支持多边形 (定义为顶点元组和以秒为单位的持续时间) 用于运动规划器。

当一个新的运动计划可用时，我们都会执行这样的操作。这样新的解决方案就可以适应接触计划的变化、参考立足点的变化以及高级操作员速度的变化。

为了计算每个多边形，我们从当前阶段 ϕ 开始，并存储接触的脚的 $x - y$ 坐标。我们搜索 $k = 1$ 的下一阶段事件 ϕ_k 以获得第一个多边形 $t_0 = t_s(\phi_k - \phi)$ 的持续时间， t_s 是以秒为单位的步幅持续时间。这样，我们就通过顶点及其持续时间在几秒钟内完全定义了一个支持多边形。我们通过将 ϕ 更新为 ϕ_k 并搜索下一阶段相位事件来不断迭代。由于接触计划是周期性的，我们重复这些步骤，直到相位事件 $\phi_0 + 1$ ，对应于起始相位 ϕ_0 。

1.3.3 运动优化问题描述^{[3]p3}

与文献中类似，每一个坐标方向的质心运动规划都被描述成一个系列的五次样条。比如沿着 x 方向的其中第 $i - th$ 曲线可以描述为：

$$\begin{aligned} x(t) &= a_{i5}^x t^5 + a_{i4}^x t^4 + a_{i3}^x t^3 + a_{i2}^x t^2 + a_{i1}^x t^1 + a_{i0}^x t^0 \\ &= \begin{bmatrix} t^5 & t^4 & t^3 & t^2 & t^1 & t^0 \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} a_{i5} & a_{i4} & a_{i3} & a_{i2} & a_{i1} & a_{i0} \end{bmatrix} \\ &= \boldsymbol{\eta}^T(t) \boldsymbol{\alpha}_i^x \end{aligned} \quad (1-17)$$

这其中 $t \in [t, t + \Delta t_i]$ 是第 i 个曲线段前所有 $(i - 1)$ 个曲线持续时间长度的总和， Δt_i 是第 i 个曲线持续的时长。基于(1-17)的关于位置的表述，可以很容易地将关于速度和加速度的表述写出来：

$$\dot{x}(t) = \dot{\boldsymbol{\eta}}^T(t) \boldsymbol{\alpha}_i^x \quad (1-18)$$

$$\ddot{x}(t) = \ddot{\boldsymbol{\eta}}^T(t) \boldsymbol{\alpha}_i^x \quad (1-19)$$

这其中：

$$\dot{\boldsymbol{\eta}}^T(t) = \begin{bmatrix} 5t^4 & 4t^3 & 3t^2 & 2t^1 & 1 & 0 \end{bmatrix}^T \quad (1-20)$$

$$\ddot{\boldsymbol{\eta}}^T(t) = \begin{bmatrix} 20t^3 & 12t^2 & 6t^1 & 2 & 0 & 0 \end{bmatrix}^T \quad (1-21)$$

对于质心在 y, z 方向上的描述是一样的。每一条质心曲线都由 x, y, z 三部分分量 $\boldsymbol{\alpha}_i = [\boldsymbol{\alpha}_i^x \quad \boldsymbol{\alpha}_i^y \quad \boldsymbol{\alpha}_i^z]^T$ 组成，可以将 n_s 条质心曲线的 $3n_s$ 条曲线参数写到一起，优化的参数矢量可以写成： $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_0^T \quad \dots \quad \boldsymbol{\alpha}_i^T \quad \dots \quad \boldsymbol{\alpha}_{n_s}^T]^T$ 。这样一来，质心的位置可以表示为：

$$\mathbf{p}_{CoM}(t) = \mathbf{T}(t)\boldsymbol{\alpha}_i \in \mathbb{R}^3, \quad \mathbf{T}(t) = \begin{bmatrix} \boldsymbol{\eta}^T(t) & 0 & 0 \\ 0 & \boldsymbol{\eta}^T(t) & 0 \\ 0 & 0 & \boldsymbol{\eta}^T(t) \end{bmatrix} \quad (1-22)$$

同样质心的速度和加速度也可以得到了：

$$\dot{\mathbf{p}}_{CoM}(t) = \dot{\mathbf{T}}(t)\boldsymbol{\alpha}_i \in \mathbb{R}^3 \quad (1-23)$$

$$\ddot{\mathbf{p}}_{CoM}(t) = \ddot{\mathbf{T}}(t)\boldsymbol{\alpha}_i \in \mathbb{R}^3 \quad (1-24)$$

这样一来整个优化问题就变为了通过求解二次优化问题 (*Quadratic Problem, QP*) 问题寻找优化的参数 $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_0^T \quad \dots \quad \boldsymbol{\alpha}_i^T \quad \dots \quad \boldsymbol{\alpha}_{n_s}^T]^T$ 。接下来将逐步介绍这个优化问题的建立：

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{Q} \boldsymbol{\alpha} + \mathbf{c}^T \boldsymbol{\alpha} \quad (1-25)$$

$$s.t. \quad \mathbf{A} \boldsymbol{\alpha} = \mathbf{b}, \quad \mathbf{D} \boldsymbol{\alpha} \leq \mathbf{f}. \quad (1-26)$$

简单起见，接下来的描述中除特殊说明外都采用 x 轴方向来表述代指各条曲线段 $\boldsymbol{\alpha}_i = [\boldsymbol{\alpha}_i^x \quad \boldsymbol{\alpha}_i^y \quad \boldsymbol{\alpha}_i^z]^T$ 。

1.3.4 规划初始化

运动计划使用扩展状态（位置、速度和加速度）进行初始化，该扩展状态用作样条序列初始状态的硬约束。这是通过使用一个 *alpha* 滤波器来融合前一次目标位置 $\mathbf{c} \mathbf{r}_{IB}^{des}$ 和当前测量 $\mathbf{c} \mathbf{r}_{IB}$ 到的位置实现的：

$$\mathbf{c} \mathbf{r}_{IB}^{ref} = \alpha \mathbf{c} \mathbf{r}_{IB}^{des} + (1 - \alpha) \mathbf{c} \mathbf{r}_{IB} \quad (1-27)$$

其中 $\alpha = 0.5e^{-\lambda t_c}$ ，其中 $\lambda \in \mathbb{R}$ 是一个用来调节前一次目标位置随其已完成时间长度 t_c 的权重。类似的过滤器也用于设置初始速度，而加速度使用最后可用的参考

值进行初始化。

1.4 质心运动优化 QP 问题建立

如1.3.3节所述，质心运动优化问题是可以转化成一个多维二次优化问题，本节将具体描述这问题的建立。一个二次优化问题主要由三个部分的定义：成本函数、等式约束、不等式约束。另外，如果约束太紧可能求解不出结果，因此要进行一定的约束松弛

1.4.1 成本函数

在我们的问题描述中，在公式(1-25)中出现的 Hessian 矩阵 $Q \in \mathbb{R}^{12n \times 12n}$ 和线性项 $c \in \mathbb{R}^{12n}$ 都有若干组成部分。我们通过下面式子来惩罚实际位置 $p(t) = T(t)\alpha$ 与 p_r 的偏差：

$$\begin{cases} \frac{1}{2} \|T\alpha - p_r\|_W^2 \\ = \frac{1}{2} (T\alpha - p_r)^T W (T\alpha - p_r) \\ = \frac{1}{2} \alpha^T T^T W T \alpha - p_r^T W T \alpha + \frac{1}{2} p_r^T W p_r \end{cases} \quad (1-28)$$

最小化公式(1-28)中的最小化平方范数相当于用公式(1-25)的形式求解 QP 问题：

$$Q = T^T W T \quad c = -T^T W^T p_r \quad (1-29)$$

为了惩罚速度参考的偏差，只需在公式(1-28)中使用 \dot{T} 和 \dot{p}_r 。可以对加速度偏差进行类似的推理。

一般来说为了使得控制更平滑、节能、稳定，成本函数会包括一些加速度最小化、软最终约束、过大偏差抑制等项目^{[2]p5}。

1.4.1.1 加速度最小化

如文献^{[5]p}所示，我们可以通过写作来最小化加速度：

$$Q_k^{acc} = \begin{bmatrix} (400/7)\Delta t_k^7 & 40\Delta t_k^6 & 24\Delta t_k^5 & 10\Delta t_k^4 & & \\ & 40\Delta t_k^6 & 28.8\Delta t_k^5 & 18\Delta t_k^4 & 8\Delta t_k^3 & \\ & 24\Delta t_k^5 & 18\Delta t_k^4 & 12\Delta t_k^3 & 6\Delta t_k^2 & 0_{4 \times 2} \\ & 10\Delta t_k^4 & 8\Delta t_k^3 & 6\Delta t_k^2 & 4\Delta t_k & \\ & & 0_{2 \times 4} & & & 0_{2 \times 2} \end{bmatrix} \quad (1-30)$$

此时非线性项 $c_k^{acc} = 0$ 。这里请注意，如果这是添加到成本函数的唯一项，则

不会有与每个样条的 α_{1k} 和 α_{2k} 系数相关联的成本, 从而导致正半定 Hessian 矩阵。当使用诸如 *Active Set one*^[6] 之类的方法时, 这是有问题的, 该方法要求 Hessian 矩阵是正定的。在这种情况下, 可以添加一个正则化项:

$$\mathbf{Q}_k^{acc\rho} = \begin{bmatrix} 0_{4 \times 4} & 0_{4 \times 2} \\ 0_{2 \times 4} & \rho \mathbb{I}_{2 \times 2} \end{bmatrix} \quad (1-31)$$

其中 $\rho = 10e^{-8}$, 线性项为空。

从文献^[6]中看得, 由于两次求导, α_{1k} , α_{2k} 将都是零。这样整个成本函数就从(1-30)简化成了(1-31), 这在对矩阵初始化的时候很有意义因为非零项会被初始化成很小的数 $\rho = 10e^{-8}$, 而恒零项就可以直接赋值为 0。

1.4.1.2 软最终约束

我们将最终位置 $\mathbf{p}_f \in \mathbb{R}^2$ 设置为由计划立足点定义的多边形的中心 $\mathbf{p}_f^{ref} \in \mathbb{R}^2$, 这是如果机器人支持多边形序列的末尾停止, 它将支持机器人的多边形。为了最小化这个范数 $\|\mathbf{p}_f - \mathbf{p}_f^{ref}\|_{\mathbf{W}_f}^2 = \|\mathbf{A}_f \mathbf{s}_f - \mathbf{p}_f^{ref}\|_{\mathbf{W}_f}^2$, 给出如下式子:

$$\mathbf{Q}_f = \mathbf{A}_f^T \mathbf{W}_f \mathbf{A}_f \quad \mathbf{c}_f = -\mathbf{A}_f^T \mathbf{W}_f \mathbf{p}_f^{ref} \quad (1-32)$$

其中 $\mathbf{W}_f \in \mathbb{R}^{2 \times 2}$ 是一个正权重的对称对角矩阵。为了避免优化器将最终状态放置在远离参考位置的解决方案, 我们在位置上添加不等式约束, 使其被限制在以参考位置为中心的框中。

1.4.1.3 偏离之前目标的解决方案

由于一旦前一个优化成功, 我们就计算一个新的优化, 为了避免连续运动计划之间的较大偏差, 我们惩罚当前解 ξ 得到的位置、速度和加速度与前一个解 ξ_{i-1} 产生的偏差。

记 t_{pre} 是前一个结果给出后消逝的时间长度, 我们通过下式惩罚两者的偏差:

$$\|\mathbf{p}(\bar{t}) - \mathbf{p}_{i-1}(t_{pre} + \bar{t})\|_{\mathbf{W}_f}^2, \bar{t} \in [0, t_f] \quad (1-33)$$

其中 $t_f = \sum_{k=0}^{n-1} f_{fk}$ 是之前所有 n 段曲线持续时间的总和。我们使用样本时间 dt 离散化优化范围 $[0, t_f]$ 。我们采用类似的成本函数来惩罚上一个解决方案获得的速度和加速度的偏差。

1.4.1.4 轨迹正则化

运动计划的持续更新可能会导致躯干相对于参考立足点的运动漂移。这可能是由于控制误差的累积造成的，这改变了解决方案，使运动变得不可行。为了避免这个问题，我们添加一个与参考正则化器路径的偏差比较的成本。这个正则化的路径可以近似表示成曲线段 $\boldsymbol{\pi}(t), \dot{\boldsymbol{\pi}}(t), \ddot{\boldsymbol{\pi}}(t)$ ，这个路径正则化器的样条系数是从最小化问题设置中获得的，使得：

- $\boldsymbol{\pi}(t)$ 的初始位置与初始化时支撑多边形的中心位置重合， $\dot{\boldsymbol{\pi}}(t), \ddot{\boldsymbol{\pi}}(t)$ 与初始速度和加速度重合；

- $\boldsymbol{\pi}(t), \dot{\boldsymbol{\pi}}(t), \ddot{\boldsymbol{\pi}}(t)$ 的结束状态由 1.4.1.2 节中定义。

- 加速度最小化

- 通过在样条结点上的状态设置等式约束来平滑地连接样条线

通过对 1.4.1.3 节中所做的整个运动进行采样，我们惩罚从路径正则化器产生的运动的偏差。

根据实际策略的不同，成本函数可能会对一些项目进行调整，还可能会包含一些其它的项目，如文献中表述^{[3]p4}。

1.4.2 等式约束

等式约束主要是从规划曲线段的连续性上出发的。整体上就是要求曲线段的位置、速度、加速度前后连续。

由于整个运动由一系列样条组成，我们设置运动优化问题以确保它们前后相连接。由于初始状态不能被运动规划器修改，我们设置了一个硬性等式约束，使得第一个样条 s_0 的初始状态与计划初始化中的一个集合重合。这个初始化的硬性等式约束可以写作 $\boldsymbol{p}(0) = \boldsymbol{p}^r, \dot{\boldsymbol{p}}(0) = \dot{\boldsymbol{p}}^r, \ddot{\boldsymbol{p}}(0) = \ddot{\boldsymbol{p}}^r$ ，于是有：

$$\begin{bmatrix} \boldsymbol{\eta}(0)^T \\ \dot{\boldsymbol{\eta}}(0)^T \\ \ddot{\boldsymbol{\eta}}(0)^T \end{bmatrix} \boldsymbol{\alpha}_0^x = \begin{bmatrix} \boldsymbol{p}_x^r \\ \dot{\boldsymbol{p}}_x^r \\ \ddot{\boldsymbol{p}}_x^r \end{bmatrix} \quad (1-34)$$

为了保证曲线段 s_k 和 s_{k+1} 是连续的，引入如下约束：

$$\begin{bmatrix} \boldsymbol{\eta}(t_{fk})^T & -\boldsymbol{\eta}(0)^T \\ \dot{\boldsymbol{\eta}}(t_{fk})^T & -\dot{\boldsymbol{\eta}}(0)^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_k^x \\ \boldsymbol{\alpha}_{k+1}^x \end{bmatrix} \quad (1-35)$$

其中 t_{fk} 表示以秒为单位的曲线 s_k 的持续时间。当连接两个样条的相等连接位于两个不相交的支持多边形之间时，我们不能再保证平滑度或运动，而是必须允许 ZMP 跳跃，这将导致加速度参考的跳跃。尽管这对控制器有负面影响，但它

确实允许优化器在遍历不相交的支持多边形时找到解决方案。我们使用文献^[1]中描述的分离轴定理 (SAT) 来检查两个支撑多边形是否相交。如果两者相交，那我们再多添加一条约束条件：

$$\begin{bmatrix} \ddot{\eta}(t_{fk})^T & -\ddot{\eta}(0)^T \end{bmatrix} \begin{bmatrix} \alpha_k^x \\ \alpha_{k+1}^x \end{bmatrix} = 0 \quad (1-36)$$

1.4.3 不等式约束

不等约束的主要内容就是将 ZMP 约束到支撑多边形内部。

如 [11] 和 [19] 所示，通过约束 Zero-Moment Point(ZMP) 位于支撑多边形内，可以保证运动过程中的平衡。从运动计划调用时使用计划脚的位置测量的脚配置开始，我们通过使用 [5] 中定义的接触时间表来计算一系列支持多边形及其持续时间。通过始终假设地面上至少有两只脚，四足机器人的支撑多边形可以是一条线、三角形或四边形。

正如 [16] 和 [19] 中所讨论的，ZMP 的位置是质心运动的函数。ZMP 的 x 坐标位置定义为：

$$x_{zmp} = x_{com} - \frac{z_{com}\ddot{x}_{com}}{g + \ddot{z}_{com}} \quad (1-37)$$

其中 g 是重力加速度。我们通过计算每个支撑多边形的顶点与中心点的极坐标并比较他们的相位来对其顶点进行顺时针排列。这样一来就可以通过添加下式约束来使得 ZMP 点满足要求：

$$ax_{zmp} + by_{zmp} + c \geq 0 \quad (1-38)$$

其中， a, b, c 是经过支撑多边形各个边的直线方程的系数。这条直线的法向量是 $\mathbf{n} = [a, b]^T$ ，其方向定义为指向约束多边形内部。

如 1.4.1 节所述，我们对运动的最终位置 \mathbf{p}_f 施加了额外的不等式约束。这是通过在 \mathbf{p}_f 上添加如公式(1-37)形式的四个约束来获得的。

1.4.4 约束松弛

约束松弛主要涉及三个方面的处理：

- ZMP 约束初始化的问题；
- 支撑多边形的缩小和逐步扩大来适应实际情况；
- 去除持续时间过短的约束多边形情况；

如果约束太紧，优化器可能会失败。首先，初始 ZMP 可能不位于当前支持多边形中。出于这个原因，我们在运动计划的第一个样本上排除了 ZMP 约束。这样优化器仍然能够找到解决方案，我们的实验表明这适用于真实系统。其次，支撑

多边形安全裕度可能太高。为了抵消这一点，每当优化失败时，我们都会以固定量迭代地减少边距，并以较慢的速度将其增加以确保优化的成功。最后，我们检查与每个多边形相关的持续时间 t_{fk} 。如果它与用于离散化运动并设置约束的样本时间更小或相当，我们从序列中删除支持多边形。

第 2 章 基于直接刚体动力学模型的 TO 控制

2.1 基于直接刚体动力学模型的 TO 控制^{[7-8]p2-6}

基于参考文献阐述基于刚体动力学模型的控制方法...

第 3 章 基于 MPC 的控制

3.1 基于 MPC 的控制^{[9]p2-4}

对于 MPC 还不太熟悉，等看完了补充下 MPC 描述...

基于参考文献阐述基于 MPC 的控制方法...

第4章 基于强化学习的机械狗控制

在这部分我将阐述关于如何使用强化学习来进行运动控制的内容。相比于传统的基于模型的控制方法，该方法通过在模拟环境中训练控制策略，可以更好地适应复杂的非线性系统动力学和环境不确定性，并且能够实时响应用户命令和环境变化。在这里我们主要关注如何使用强化学习来训练机械狗类型的足式和轮式机器人控制器。

4.1 强化学习控制

4.1.1 强化学习的基本概念^[10]

在强化学习中，控制问题被建模为一个马尔可夫决策过程（Markov Decision Process）。MDP 是一个 RL 中常用的用于随机控制过程建模的数学框架，它定义了一个包含状态空间、动作空间、奖励函数和状态转移函数的元组，描述了一个决策过程的基本组成部分。在 MDP 中，每个时间步骤代理从周围环境中观察到某个状态 $s_t \in \mathcal{S}$ ，并输出一个动作 $a_t \in \mathcal{A}$ ，接着环境通过状态转移函数 $p(s_{t+1}|s_t, a_t)$ 演化到新的状态 s_{t+1} ，并且根据奖励函数给出相应的奖励 $r_t \in \mathcal{R} : \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ 和对更新后的环境状态的观察结果作为下一周期的 s_t 。代理可以根据 θ 参数化的随机策略 $\pi_\theta(a_t|s_t)$ 来采取行动。RL 通过与环境交互来更新参数 θ 以最大化累计折扣奖励（cumulative discounted rewards） $\mathbb{E}[\sum_{t=k}^{\infty} \gamma^t r_t]$ ，其中 k 是当前时间步长（timestep）， γ 是折扣因子（discount factor）。

4.1.2 基于深度强化学习的控制器的部署和优化流程

（1）在仿真环境中训练控制器：使用深度强化学习算法在仿真环境中训练控制器，使其能够完成所需的任务。在训练过程中，可以使用特权训练方法来提高训练效率和性能。

（2）部署控制器到实际环境中：将训练好的控制器部署到实际环境中，例如机器人或移动设备。在实际环境中，控制器将接收传感器数据并输出动作命令。

（3）优化控制器：在实际环境中，可以使用在线学习方法来进一步优化控制器的性能。例如，可以使用模型预测控制方法来对控制器进行在线微调。

需要注意的是，在实际环境中，传感器数据可能会受到噪声和不确定性的影响，因此需要设计鲁棒性强的控制器来应对这些问题。此外，还需要考虑实际环

境中的安全性和可靠性问题。

4.2 单层神经网络直接驱动关节方式^{[11]p3}

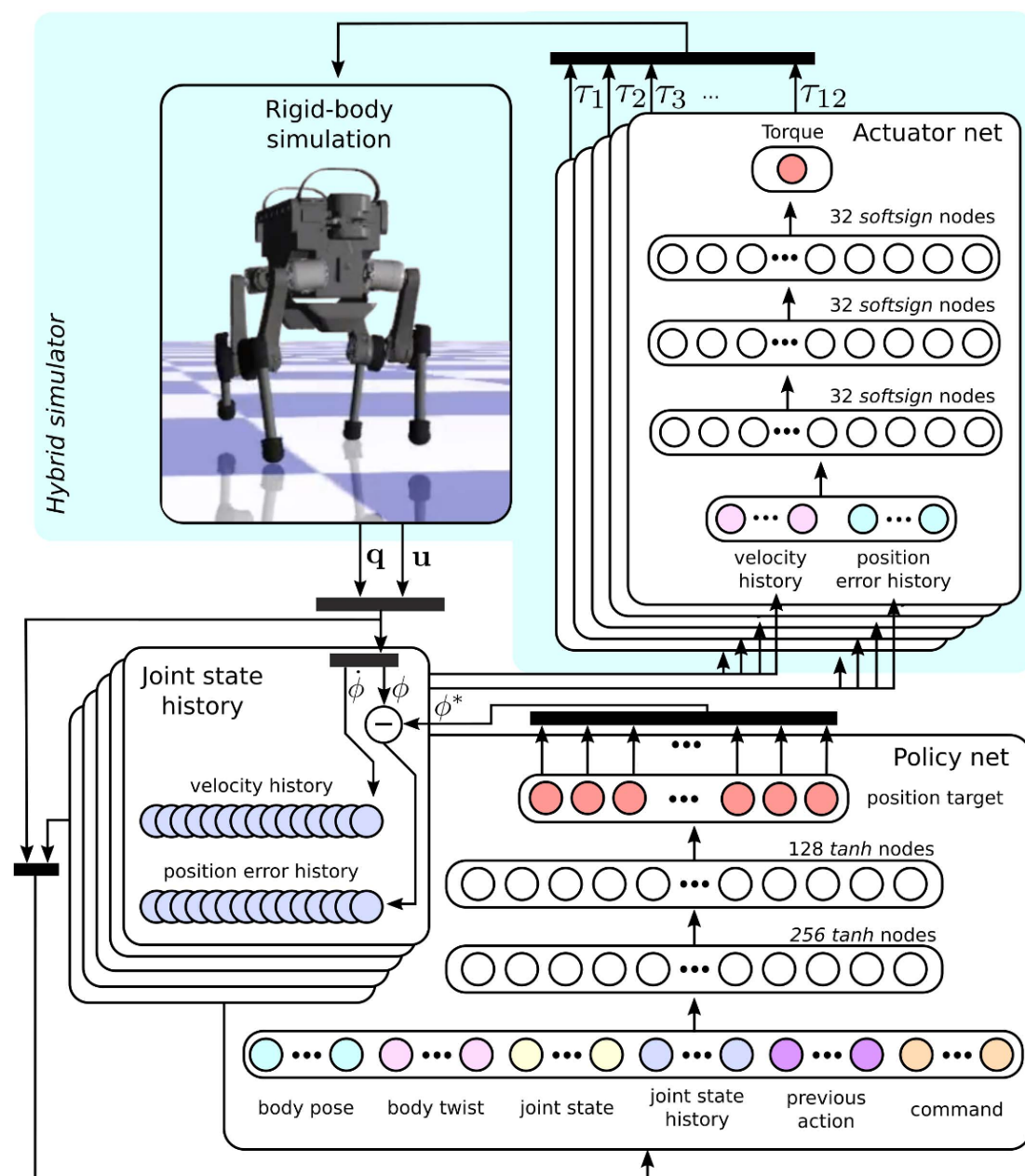


Fig. 5 Training control policies in simulation. The policy network maps the current observation and the joint state

图 4-1 The policy network maps the current joint state and position history to the joint target.

如图4-1所示，整个训练循环包含下面几个部分：

- 首先，一个刚体仿真模拟器运用给定的关节节点扭矩和电流状态计算输出机器人下一状态。与此同时关节的速度和位置误差在有限时间窗口内被缓冲在节点状态历史中以备网络层使用。

- 接着，由两个隐藏层的 MLP 神经网络构成控制策略，它将输入的电流状态和节点状态历史的观测信息映射成节点位置目标。
- 最后，关节神经网络将关节状态历史和关节位置目标映射到 12 个关节的扭矩值上，然后继续整个循环。

4.3 双层神经网络本体感知方式^{[12]p8}

4.3.1 总体概况

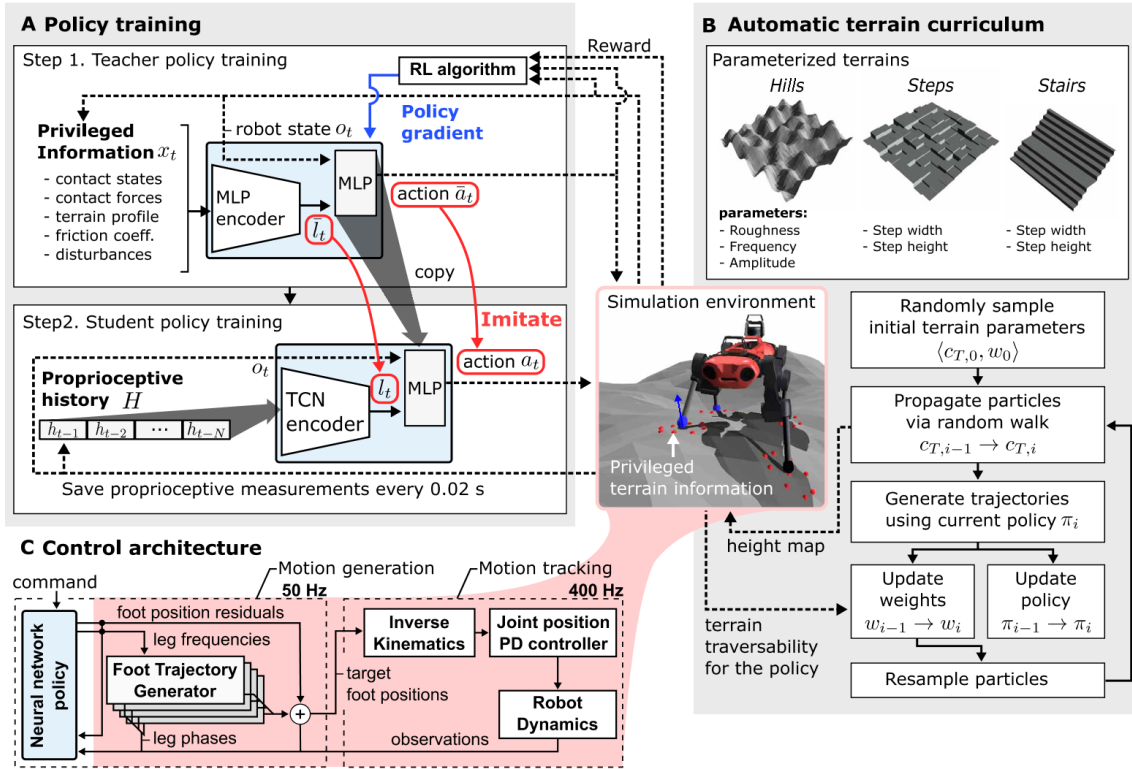


图 4-2 Creating a proprioceptive locomotion controller^{[12]p7}.

图4-2是一个只采用本体感知进行机器狗控制的 RL 策略实现架构。它包括训练定义、地形定义、控制架构。

- 策略训练：整个 RL 训练采用特权训练 (*privileged training*)^{[13]p} 的理念，训练过程分为两个阶段，如图4-2A 所示。第一阶段用完善的特权信息作为输入训练出一个教师策略 (*teacher policy*)；第二阶段仅用本体感知作为输入训练出一个学生策略 (*student policy*)，学生策略的训练过程通过模仿教师策略实现。实际部署到机器人上的策略是学生策略。

- 地形定义：RL 训练过程是在一定的地形条件下完成的，如图4-2B 所示。对于同样的控制命令，不同的地形条件会产生不同的任务难度。为了更好地促进训练进步，训练过程中的地形采用自适应地形课程 (*adaptive terrain curriculum*)，它

会从简单地形开始根据训练控制器的性能提升而不断提高地形难度，使得 RL 控制器时钟面对相对于当前状况的中等地形难度。

- 控制架构: 整个控制架构采用调节轨迹生成器 (*Policies Modulating Trajectory Generators, PMTG*)^{[14]p} 架构来提供运动生成的先验。神经网络策略通过合成残差位置命令来调节腿相和运动原语，如图4-2C 所示。

4.3.2 策略训练

4.3.2.1 教师策略

整个控制问题被构建成一个马尔科夫决策过程 (*Markov Decision Process, MDP*)。MDP 是一种用于状态和结果部分随机的离散时间控制过程的数学框架。MDP 被定义为一个状态空间 \mathcal{S} ，动作空间 \mathcal{A} ，一个标量奖励函数 $\mathcal{R}(s_t, s_{t+1})$ ，一个转移概率 $P(s_{t+1}|s_t, a_t)$ 。一个学习代理从策略 $\pi(a_t|s_t)$ 中选择一个动作 a_t 并从环境中获得奖励 r_t 。RL 的目的是找到一个优化的策略 π^* 使得无限时间范围内的折扣奖励总和最大。

状态空间 s_t 定义为: $\langle o_t, x_t \rangle$ ，其中 o_t 是机器人自身的观测信息， x_t 是特权信息。

$$s_t \begin{cases} o_t \begin{cases} \text{proprioceptive sensors} \\ \text{state estimator} \begin{cases} \text{base velocity} \\ \text{orientation} \end{cases} \end{cases} \\ x_t \begin{cases} \text{only for simulation} \end{cases} \end{cases} \quad (4-1)$$

动作空间 \bar{a}_t 是一个 16 维向量，由腿的频率和交的位置残差构成。

奖励函数 $\mathcal{R}(s_t|s_{t+1})$ 定义。

reward function definition details ...

如图4-2A 所示，策略网络由两个 MLP 模块构成。第一个 MLP 模块将 x_t 信息转化成潜在向量 \bar{l}_t ，由于 x_t 中并不含有机器人的状态或指令信息，也就是说他只包含地形信息和接触相关的信息。假设 \bar{l}_t 的作用是驱动自适应行为，比如根据地形轮廓改变脚步间隙大小。然后 \bar{l}_t 和 o_t 再提供给第二个 MLP 网络来计算动作。

训练采用信任区域策略优化 (*Trust Region Policy Optimization, TRPO*)^{[15]p}。

4.3.2.2 学生策略

学生策略仅能获取 o_t 信息。这里的一个关键假设是潜在特征 \bar{l}_t 可以从本体感觉观察的时间序列 h_t 中恢复 (部分), $h_t :=$

$o_t\{f_0, \text{joint history}, \text{previous foot position targets}\}$ 。

学生策略采用 TCN^{[16]p} 编码器。使用 TCN 架构的原因是它对输入历史长度提供透明的控制，可以容纳长历史，并且已知对超参数设置具有鲁棒性^{[16]p}。

学生策略采用监督学习的方式进行训练，损失函数定义为：

$$\mathcal{L} := (\bar{a}_t(o_t, x_t) - a_t(o_t, H))^2 + (\bar{l}_t(o_t, x_t) - l_t(H))^2 \quad (4-2)$$

带 $(\bar{\cdot})$ 标识的项是来自教师策略的生成。对于每个访问的状态，教师策略计算其嵌入和动作向量 $(\bar{\cdot})$ ，这些教师策略的输出随即被用作与相应状态相关的监督信号。

4.3.3 地形定义

Adaptive terrain curriculum ...

4.3.4 控制架构

整个控制架构结构如图4-2C所示。它被分为两大类：运动生成、跟随。整个系统的输入只有控制指令和本体感知，输出为关节点位置目标。

运动生成策略是一种基于周期性腿相位的策略，之前的一些工作中通常利用预定义的脚步接触时间表^{[3,17]p7}。为每条腿都定义一个周期的相位变量 $\phi_i \in [0.0, 2\pi)$ 。在每一个时间步长 t 里， $\phi_i = (\phi_{i,0} + (f_0 + f_i)t) \pmod{2\pi}$ ，其中 $\phi_{i,0}$ 是初始相位， f_0 是一般基础频率， f_i 是第 i 条腿的频率偏移。我们希望腿在 $f_0 + f_i \neq 0$ 时表现出周期性运动，并在接触阶段与地面接触。这里基础频率参考之前开发的传统控制^{[3]p7}中鸚鵡步态的值， $f_0 = 1.25\text{Hz}$ 。

target foot position...

我们采用 PMTG 架构来将神经网络集成系统中用来调节控制器的输出。整体实现由四个完全相同的足迹生成器 (*foot trajectory generators, FTGs*) 和一个神经网络策略 (*neural network policy, NNP*) 构成。足迹生成器是一个输出每条腿的脚位置的函数： $F(\phi) : [0.0, 2\pi) \rightarrow \mathbb{R}^3$ 。这个 FTG 在 f_i 不为零时驱动垂直方向的踱步运动。

Addition 4.3.1. $F(\phi)$ 的定义如下：

$$F(\phi_i) = \begin{cases} (h(-2k^3 + 3k^2) - 0.5)^{H_i z} & k \in [0, 1] \\ (h(2k^3 - 9k^2 + 12k - 4) - 0.5)^{H_i z} & k \in [1, 2] \\ -0.5^{H_i z} & otherwise \end{cases} \quad (4-3)$$

其中 $k = 2(\phi_i - \pi)/\pi$, h 是一个表示最大步高的参数。在摆动相位的每一个阶段都是一个连接最高点和最低点的三次 Hermite 样条曲线, 在连接点处曲线具有一阶连续性 (一阶导数为零)。其它的周期函数, 如 $h_i \sin(\phi_i)$ 可以用于 FTG。给定一组合适的 $f_0, h, \phi_{i,0}$ 参数值, 机器人就能稳定地在地面上踱步了。比如, 文献^{[12]p7}中使用的值: $f_0 = 1.25, h = 0.2, \phi_{i,0}$ 从 $U(0, 2\pi)$ 中采样得到。

神经网络策略输出 f_{iS} 和脚的目标位置残差 ($\Delta \mathbf{r}_{f_i, T}$)。这样, 第 i 个脚的目标位置是 $\mathbf{r}_{f_i, T} := F(\phi_i) + \Delta \mathbf{r}_{f_i, T}$ 。

跟随控制是使用解析逆运动学 (*inverse kinematics, IK*) 和关节点位置控制 (*position control*) 实现的。在 H_i 中定义脚位置首先表述在机器人身体系中, 然后用 IK 计算关节点的位置目标。而关节点的位置目标依靠关节点的 PD 控制器来跟随。使用 IK 的好处是可以最大化计算效率同时在仿真到实际 (*sim-to-real*) 的过程中可以复用已有的位置控制驱动器模型^{[18-19]p}。

4.4 双层网络本体和外部感知融合方式^{[20]p7}

整个神经网络的训练是在仿真环境中完成的, 然后采用 *zero-shot sim-to-real* 的转换部署到实际的额机器人上。整个方法分为三个阶段, 如图4-3所示。

(1) 首先, 使用 RL 训练教师策略, 以在随机生成的具有随机干扰的地形上遵循随机目标速度。该策略可以访问特权信息, 例如无噪声地形测量、地面摩擦和引入的扰动。

(2) 在第二阶段, 训练学生策略重现教师策略的动作, 而不使用这种特权信息。学生策略构造一个信念状态来使用循环编码器捕获未观察到的信息, 并根据该信念状态输出一个动作。在训练期间, 我们利用两个损失: 行为克隆损失和重建损失。行为克隆损失旨在模仿教师策略。重新构造损失鼓励编码器产生信息丰富的内部表示。

(3) 最后, 我们将学习到的学生策略转移到物理机器人上, 并将其与机载传感器在现实世界中部署。机器人通过整合来自板传感器的深度数据和从构建的高程图中采样高度读数来构建高程图, 以形成策略的外部感知输入。这种外部感知

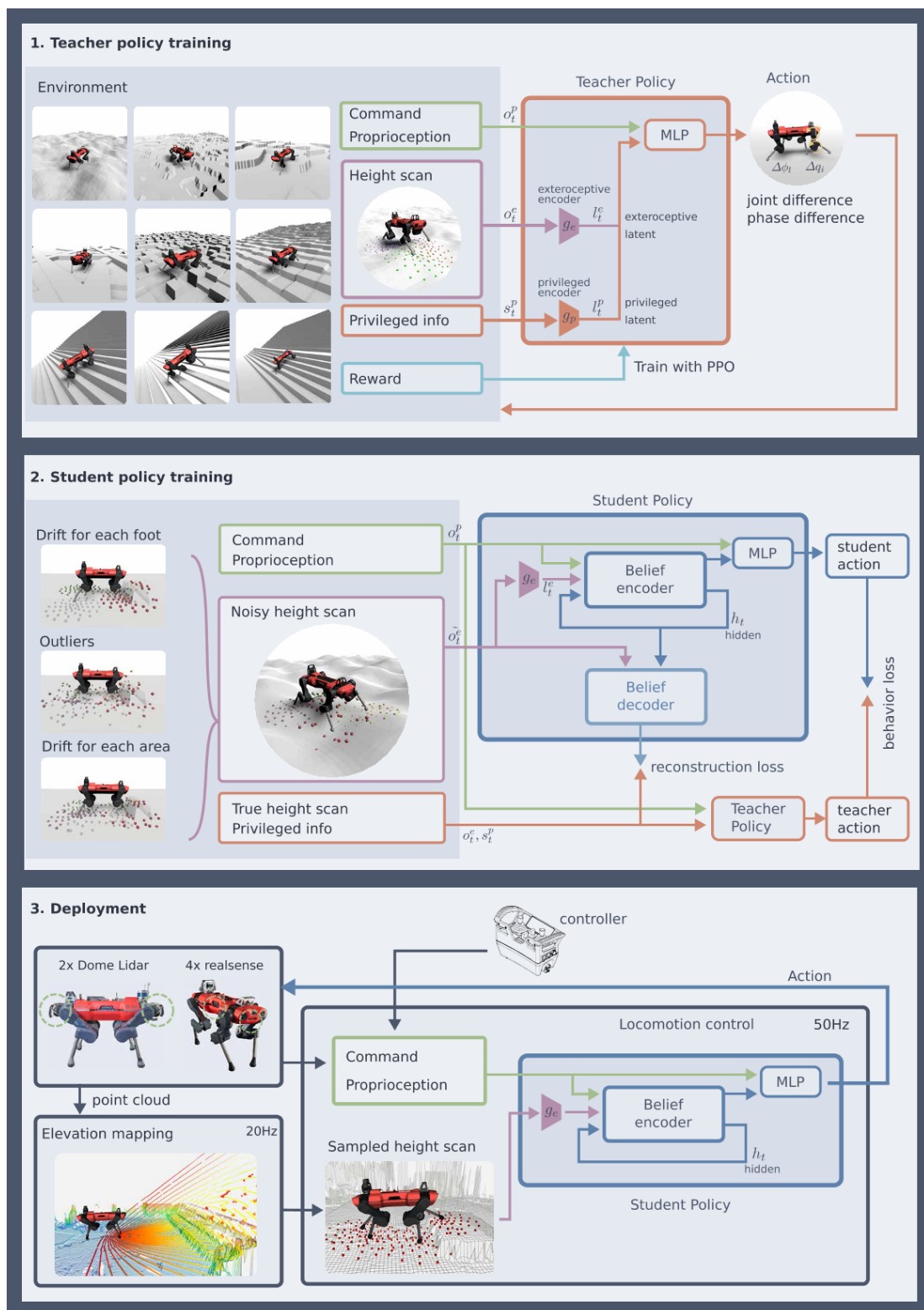


图 4-3 RL implementation process^{[20]p9}.

输入与本体感觉数据相结合，并提供给神经网络，该网络产生执行器命令。板上传感器和自身构建等高图融合

4.4.1 问题描述

我们在离散时间动力学中制定了我们的控制问题，其中环境完全由时间步 t 的状态 s_t 定义。该策略实施一个动作 a_t 并且从环境中获得一个观测结果 o_t ，这个测量结果来自观测模型 $\mathcal{O}(o_t|s_t, a_t)$ 。接着环境以 $P(s_{t+1}|s_t, a_t)$ 的概率转移到下一个状态 s_{t+1} 并给出一个奖励 r_{t+1} 。

当所有状态都可以被观测的实况下，也即 $o_t = s_t$ 时，整个问题可以被看做一个马尔可夫决策过程 *Markov decision process(MDP)*。然而，当存在不可观察性的信息时，例如外力或完整的地形信息，动力学被建模为部分可观察的马尔可夫决策过程 *partially observable Markov decision process(POMDP)*。

RL 的目标是找到一个能够使得未来轨迹的预期折扣奖励最大化的策略 π^* ，以使得：

$$\pi^* = \underset{a}{\operatorname{argmax}} E\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \quad (4-4)$$

已经开发了许多 RL 算法来解决完全可观察的 MDP 问题，并且易于用于训练。然而，POMDP 问题的情况更具挑战性，因为状态不能完全观察到。这通常通过从历史的观测结果 o_0, \dots, o_t 中构建一个信念状态 b_t belief state 以尝试构建完全状态的方式来解决。在深度 RL 中，这通常是通过堆叠一系列先前的观测结果^{[21]p}或使用可以压缩过去信息的架构来完成的，例如循环神经网络 (RNN)^{[22]p}或时间卷积网络^{[12,16]p}。

从头开始训练一个天真地处理序列数据的复杂的神经网络策略可能很耗时^{[12]p}。因此，我们使用特权学习 (45)，我们首先训练一个具有特权信息的教师策略，然后通过监督学习将教师策略提炼为学生策略^{[13]p}。

- (1) 训练环境：
- (2) 地形：
- (3) 域随机化：
- (4) 片段终止条件：

4.4.2 教师策略训练

在训练的第一阶段，我们的目标是找到一个可以访问完美、特权信息的最佳参考控制策略，并使 ANYmal 在随机生成的地形上遵循所需的命令速度。命令需

求随机生成并构成一个向量 $\mathbf{v}_{des} \in \mathbb{R}^3 = (v_x, v_y, w)$ ，其中 v_x, v_y 分别表示在机器人自身坐标系下的纵向速度和横向速度， w 表示自转速度。

我们采用近端策略优化 *proximal policy optimization(PPO)*^[23] 来训练教师策略。教师策略被建模为一个高斯策略， $\mathbf{a}_t \sim \mathcal{N}(\pi_\theta(o_t = s_t), \sigma I)$ ，其中 π_θ 由用 θ 参数化的多层感知器 *multilayer perceptron(MLP)* 实现， σ 表示每个动作之间的方差。

4.4.3 观测和行动

教师策略的观察定义为 $\mathbf{o}_t^{teacher} = (\mathbf{o}_t^p, \mathbf{o}_t^e, \mathbf{s}_t^p)$ ，其中 \mathbf{o}_t^p 表示本体感觉观察 *proprioception observation*， \mathbf{s}_t^p 表示外部观察 *exteroceptive observation*， \mathbf{s}_t^p 表示特权状态 *privileged state*。

- \mathbf{o}_t^p 包含本体速度、转动、节点位置和速度历史、动作历史、每条腿的相位；
- \mathbf{o}_t^e 是每只脚周围高度样本的向量，包括五种不同的半径；
- \mathbf{s}_t^p 包括接触状态、接触力、接触法线、摩擦系数、大腿和小腿接触状态、施加到身体上外部力和力矩、摆动阶段持续时间；

我们的动作空间受到中心模式生成器的启发^[12]。每条腿 $l = 1, 2, 3, 4$ 保存一个相位变量 ϕ_l 并定义了基于相位的标称轨迹。这个标称轨迹是一个脚尖的步伐运动，我们使用逆运动学来计算每个关节执行器 $i = 1, \dots, 12$ 的标称关节目标 $q_i(\phi_l)$ 。来自策略的动作是相差 $\Delta\phi_l$ 和关节位置目标残差 Δq_i 。

更详细相关内容看相关附件的S5。

观测向量定义如4-1表所示。本体感知包括命令、节点、身体信息、腿部相位信息。中央模式生成器 *central pattern generator(CPG)* 的相位信息包含 $\Delta\phi_l, \cos \phi_l, \sin \phi_l$ 和每条腿 l 的基础频率。对于外部感知我们采用每个脚周围的高度采样来代替局部高程图。圆形采样模式包括每只脚周围的 $\{6, 8, 10, 12, 16\}$ 个点，半径分别为 $\{0.08, 0.16, 0.26, 0.36, 0.48\}$ m。

运动被定义为 $\langle \Delta\phi_l, \Delta q_i \rangle$ ，其中 $\Delta\phi_l$ 和 Δq_i 分别表示每个条腿 ($l \in \{legs\}$) 的相位偏移和节点目标位置残差 ($i \in \{1, \dots, 12\}$)。我们有一个标称轨迹 $\mathbf{p}(\phi) : \mathbb{R} \rightarrow \mathbb{R}^3$ ，它将各个 ϕ_l 映射到目标脚位置，随着 ϕ 在 $[0, 2\pi)$ 范围内生成周期性步进运动。从动作中，每条腿 l 的节点目标位置用逆运动学 $IK(\cdot)$ 和基础相位频率 $\Delta\phi_0$ 定义为 $q_{i \in l}^{target} = IK(\mathbf{p}(\phi_l + \Delta\phi_l + \Delta\phi_0)) + \Delta q_{i \in l}$ 。

标称足迹定义如下。如果相位是处于上摆动期间 ($0 \leq \phi_l \leq \pi/2$) 则：

$$\mathbf{p}_l(\phi_l) = \langle x_l^n, y_l^n, z_l^n + 0.2 \cdot (-2t_l^3 + 3t_l^2) \rangle, \text{ where } t_l = 2/\pi \cdot \phi_l \quad (4-5)$$

$\{x, y, z\}_l^n$ 是默认姿态配置处的标称脚位置。三次 Hermite 样条在 $\phi_l = 0$ 处连接 $z = z_l^n$ 在 $\phi_l = \pi/2$ 处连接 $z = z_l^n + 0.2$ 。

在下摆动期间 ($\pi/2 < \phi_l \leq \pi$), 足高计算如下:

$$\mathbf{p}_l(\phi_l) = \langle x_l^n, y_l^n, z_l^n + 0.2 \cdot (2t_l^3 - 3t_l^2 + 1) \rangle, \text{ where } t_l = 2/\pi \cdot \phi_l - 1 \quad (4-6)$$

它与前面的函数对称。在驻立阶段 ($\pi < \phi_l \leq 2\pi$), $\mathbf{p}_l(\phi_l) = \langle x_l^n, y_l^n, z_l^n \rangle$ 。

表 4-1 Observations. Proprioception is used for both teacher and student training. Exteroception is given in the form of height samples. The privileged information is used only for teacher training.

Observation type	Input	Dimensionality
Proprioception	command	3
	body orientation	3
	body velocity	6
	joint position	12
	joint velocity	12
	joint position history (3 time steps)	36
	joint velocity history (2 time steps)	24
	joint target history (2 time steps)	24
	CPG phase information	13
Exteroception	height samples	208
Privileged info.	contact states	4
	contact forces	12
	contact normals	12
	friction coefficients	4
	thigh and shank contact	8
	external forces and torques	6
	airtime	4

4.4.4 策略构架

我们将教师策略 π_θ 建模为一个 MLP。它包括三个 MLP 组成部分: 外部感知编码器、特权编码器、主网络, 如图4-3所示。

- (1) 外部感知编码器 g_e 接收来自 o_t^e 的信息, 然后输出一个小的潜在表示

$$l_t^e = g_e(o_t^e)$$

- (2) 特权编码器 g_p 接收来自特权状态 s_t^p 的信息, 然后输出一个潜在表示

$$l_t^{priv} = g_p(s_t^p)$$

(3)

这些编码器将每个输入压缩为更紧凑的表示，并使得学生策略能更方便地重用一些教师策略组件。

更详细相关内容看相关附件的 S6。

策略网络由多层 MLP 组成。用基于 MLP 的编码器 (g_e, g_p)，将高度采样结果首先编码成为一个 $24 \times 4 = 96$ 维的潜在向量；将特权信息编码成为一个 24 维的潜在变量。每个编码器有两层分别是 $\{80, 60\}$ 和 $\{64, 32\}$ 的隐藏单元。高度样本首先分别针对每只脚分别输入到编码器中，然后连接成一个特征向量。然后将这些特征与本体感受观察连接起来，并馈送到另一个具有三个隐藏层 $\{256, 160, 128\}$ 的 MLP。所有 MLP 的激活函数为 LeakyReLU (72)。

我们使用具有外部感受门的 GRU 作为信念编码器 (图4-4C)。GRU 由 2 个堆叠的层组成，每个层有 50 个隐藏单元。信念编码器和外部感知门 g_b, g_a 用于计算 $96 + 24 = 120$ 维信念状态 b_t 和 96 维注意力向量 α 。每个编码器有两个隐藏层，每个隐藏层有 $\{64, 64\}$ 和 $\{64, 64\}$ 个隐藏单元。过滤后的外部感受信息 $l_t^e \odot \alpha$ 被添加到 $g_b(b_t')$ ，使用零填充来匹配维度差异。

4.4.5 奖励函数

针对速度控制命令的跟随，我们定义正奖励；针对违反约束的情况，我们定义负奖励。指令跟随奖励定义如下：

$$r_{command} = \begin{cases} 1.0, & \text{if } \mathbf{v}_{des} \cdot \mathbf{v} > |\mathbf{v}_{des}| \\ \exp(-(\mathbf{v}_{des} \cdot \mathbf{v})^2), & \text{otherwise} \end{cases} \quad (4-7)$$

其中 $\mathbf{v}_{des} \in \mathbb{R}^2$ 是所需的水平速度， $\mathbf{v} \in \mathbb{R}^2$ 是身体坐标系下当前身体速度。同样的奖励机制也被应用与转动速度情况。

我们惩罚与期望速度正交的速度分量以及横摇、俯仰和偏航周围的身体速度。此外，我们使用整形奖励进行身体方向、关节扭矩、关节速度、关节加速度和脚滑以及小腿和膝盖碰撞。身体方向奖励用于避免身体的奇怪姿势。联合相关奖励术语用于避免过于激进的运动。脚滑和碰撞奖励术语用于避免它们。我们通过模拟中查看策略的行为来调整奖励术语。除了遍历性能外，我们还检查了运动的平滑度。

更详细相关内容看相关附件的 S7。

奖励函数定义为：

$$r = 0.75(r_{lv} + r_{av} + r_{lvo}) + r_b + 0.003r_{fc} + 0.1r_{co} + 0.001r_j + 0.08r_{jc} + 0.003r_s + 1.0 \cdot 10^{-6}r_t + 0.0 \quad (4-8)$$

各项的具体定义如下：

(1) 线速度奖励 *Linear Velocity Reward*(r_{lv})：这项鼓励策略去跟随需要的水平 (x, y plane) 速度指令：

$$r_{lv} = \begin{cases} \exp(-|\mathbf{v}|^2), & \text{if } |\mathbf{v}_{des}| = 0 \\ 1.0, & \text{elseif } \mathbf{v}_{des} \cdot \mathbf{v} > |\mathbf{v}_{des}| \\ \exp(-(\mathbf{v}_{des} \cdot \mathbf{v} - |\mathbf{v}_{des}|)^2), & \text{otherwise} \end{cases} \quad (4-9)$$

其中 $\mathbf{v}_{des} \in \mathbb{R}^2$ 是目标水平速度， $\mathbf{v} \in \mathbb{R}^2$ 是身体坐标系下当前身体的速度。

(2) 角速度奖励 *Angular Velocity Reward*(r_{av})：这一项鼓励策略去跟随需要的转动速度指令：

$$r_{av} = \begin{cases} \exp(-w_z^2), & \text{if } w_{des} = 0 \\ 1.0, & \text{elseif } w_{des} \cdot \mathbf{w}_z > w_{des} \\ \exp(-(w_{des} \cdot \mathbf{w}_z - |w_{des}|)^2), & \text{otherwise} \end{cases} \quad (4-10)$$

(3) 线性正交速度奖励 *Linear Orthogonal Velocity Reward*(r_{lvo})：这项惩罚与目标速度指令正交的速度：

$$r_{lvo} = \exp(-3.0|\mathbf{v}_o|^2), \text{ where } \mathbf{v}_o = \mathbf{v} - (\mathbf{v}_{des} \cdot \mathbf{v})\mathbf{v}_{des} \quad (4-11)$$

(4) 身体运动奖励 *Body Motion Reward*(r_b)：这项惩罚身体速度中不符合指令要求的部分：

$$r_{bm} = -1.25v_z^2 - 0.4|\omega_x| - 0.4|\omega_y| \quad (4-12)$$

(5) 脚感奖励 *Foot Clearance Reward*(r_{fc})：当一条腿处于摆动相位时，比如 $\phi_i \in [0, \pi)$ ，机器人应该将这条腿上对应的脚抬起到比障碍物高的程度。但是，为了防止机器人做出不必要的抬高间隙，我们通过给出惩罚 r_{fcl} 来正则化腿轨迹。 $H_{sample,l}$ 是第 l 只脚的高度采样。这样，间隙成本定义为：

$$r_{fcl} = \begin{cases} -1.0, & \text{if } \max(H_{sample,l}) < -0.2 \\ 0.0, & \text{otherwise} \end{cases} \quad (4-13)$$

$$r_{fc} = \sum_{l=1}^4 r_{fcl} \quad (4-14)$$

注意高度样本是相对于脚高度进行采样，因此 -0.2 表示地形高度比脚低 0.2m ；脚比采样的地形高度高 0.2m 。

(6) 小腿和膝关节碰撞奖励 *Shank and Knee Collision Reward*(r_{co})：我们希望

惩罚脚以外的地形和机器人部件之间的不良接触，以避免硬件损坏：

$$r_{co} = \begin{cases} -c_k, & \text{if shank or knee is in collision} \\ 0.0, & \text{otherwise} \end{cases}$$

这里 c_k 是课程因子，它单调增加并收敛到 1。

(7) 节点运动奖励 *Joint Motion Reward*(r_j)：该项惩罚关节速度和加速度以避免振动：

$$r_s = -c_k \sum_{i=1}^{12} (0.01 \dot{q}_i^2 + \ddot{q}_i^2) \quad (4-15)$$

其中 \dot{q}_i, \ddot{q}_i 分别是节点的速度和加速度。

(8) 节点约束奖励 *Joint Constraint Reward*(r_{jc})：该项在联合空间中引入了一个软约束。为了避免相反方向的膝关节翻转，我们对超过阈值的惩罚：

$$r_{jc,i} = \begin{cases} -(q_i - q_{i,th})^2, & \text{if } q_i > q_{i,th} \\ 0.0, & \text{otherwise} \end{cases} \quad (4-16)$$

$$r_{jc} = \sum_{i=1}^{12} r_{jc,i} \quad (4-17)$$

$$(4-18)$$

其中 $q_{i,th}$ 是第 i 个节点的阈值。我们只设置膝关节的阈值。

(9) 目标平滑奖励 *Target Smoothness Reward*(r_s)：通过对目标位置一阶和二阶有限差分导数的大小进行惩罚，使生成的足部轨迹变得更平滑：

$$r_s = c_k \sum_{i=1}^{12} ((q_{i,t}^{des} - q_{i,t-1}^{des})^2 + (q_{i,t}^{des} - 2q_{i,t-1}^{des} + q_{i,t-2}^{des})^2) \quad (4-19)$$

其中 $q_{i,t}^{des}$ 是 t 步时间里第 i 个节点的目标位置。

(10) 扭矩奖励 *Torque Reward*(r_τ)：我们惩罚节点的扭矩里节省能耗 ($\tau \propto electriccurrent$)：

$$r_\tau = -c_k \sum_{i=1}^{12} \tau_i^2 \quad (4-20)$$

其中 τ_i 是执行器网络计算出的第 i 个节点的扭矩。

(11) 滑动奖励 *Slip Reward*(r_{slip})：我们惩罚与地面接触的速度以减少滑

动情况：

$$r_{slip} = -c_k \sum_{l \in \{footincontact\}} v_{f,l}^2 \quad (4-21)$$

其中 $v_{f,l}^2$ 是与地面接触了的第 l 只脚的速度。

4.4.6 课程

随着策略性能的提高，我们使用两个课程来提高难度。一个课程使用自适应方法^{[12]p}调整地形难度，另一个改变元素，如奖励或使用逻辑函数^{[11]p}应用干扰。

对于地形课程，粒子滤波更新地形参数，使它们仍然具有挑战性，但在策略训练期间的任何时候都可以实现^{[12]p}。

第二个课程将域随机化的幅度和一些奖励项（关节速度、关节加速度、方向、滑移和大腿和小腿接触）乘以单调递增且渐近趋势为 1 的因子：

$$c_{k+1} = (c_k)^d$$

其中 c_k 是第 k 次迭代的课程因子， $d \in (0, 1)$ 是收敛率。

4.4.7 学生策略训练

在我们训练好一个可以在特权信息的帮助下穿越各种地形教师策略后，我们就可以将其提炼成一个学生策略，该策略只能访问真实 robot 上可用的信息。我们使用与教师策略相同的训练环境，但在学生高度样本观察中添加额外的噪声： $o_t^{student} = (o_t^p, n(o_t^e))$ ，其中 $n(o_t^e)$ 是一个用于高度样本输入的噪声模型。该噪声模型模拟了现场部署过程中经常遇到的外部感觉的不同失败案例，具体如下。

当外部感觉中存在较大的噪声时，它变得不可观察；因此，动力学被认为是 POMDP。此外，由于缺乏直接测量的传感器，特权状态是不可观察的。因此，该策略需要考虑顺序相关性来估计不可观察的状态。我们建议使用循环信念状态编码器来组合外部感知和本体感觉的序列，以估计不可观察的状态作为信念状态。

学生策略由循环信念状态编码器和 MLP 组成，如图4-3所示。我们用 h_t 表示循环网络的隐藏状态。信念状态编码器接收 $o_t^{student}, h_t$ 并输出一个潜在向量 b_t ，它表示信念状态。我们的目标是将信念状态 b_t 与编码所有运动相关信息的教师策略的特征向量 (l_t^e, l_t^{priv}) 进行匹配。接下来我们将 o_t^p 和 b_t 传入 MLP，由它计算出最终的动作。MLP 结构与教师策略相同，这样我们就可以重用教师策略的学习权重来初始化学生网络并加快训练速度。

学生策略的训练通过最小化两个损失以有监督的方式进行训练：行为克隆损失 *behavior cloning loss* 和重建损失 *reconstruction loss*。

• 克隆损失定义为给定相同状态和指令的学生动作和教师动作之间的平方距离。

• 重建损失定义为无噪声高度样本和特权信息 (o_t^e, s_t^p) 及其与信念状态 b_t 的重建之间的平方距离。

我们通过推出学生策略来生成样本，以提高鲁棒性^{[24-25]p}。

4.4.8 高度采样随机化

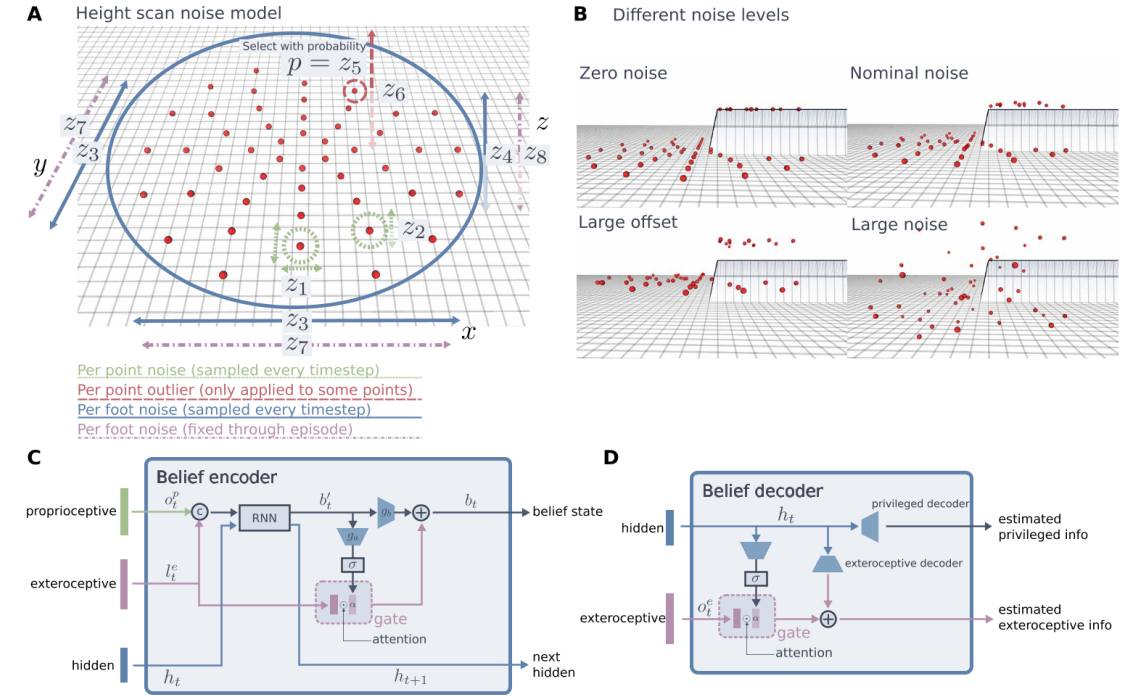


Fig. 7. Details of robust terrain perception components. (A) During student training, random noise is added to the height samples. The noise is sampled from a Gaussian

图 4-4 Robust terrain perception^{[20]p10}。

在学生训练过程中，我们使用参数化噪声模型 $n(\sigma_t^e | o_t^e, z), z \in \mathbb{R}^{8 \times 4}$ 将随机噪声注入到高度样本中。我们在对高度进行采样时应用了两种不同类型的测量噪声，如图4-4A所示：

- (1) 横向移动扫描点。
- (2) 扰动高度值。

每个噪声值都是从高斯分布中采样的，噪声参数 z 定义方差。这两种类型的噪声都应用于三个不同的范围，所有这些都有自己的噪声方差：每个扫描点、每只脚和每一集。每个扫描点和每只脚的噪声值在每个时间步重新采样，而所有扫描点的表观轮廓噪声保持不变。

此外，我们定义了三个具有相关噪声参数 z 的映射条件来模拟不断变化的地图质量和误差源，如图4-4B所示。

- (1) 名义噪声假设常规操作期间具有良好的地图质量。

(2) 大偏移量噪声来模拟由于姿态估计漂移或可变形地形造成的地图偏移。

(3) 幅度较大的噪声来模拟由于遮挡或映射失败导致完全缺乏地形信息的情况。

这三个映射条件在每个训练集的开头以 60%、30% 和 10% 的比例选择。

最后，我们将每个训练地形划分为单元格，并向高度样本添加附加偏移量，具体取决于它采样的单元格。这模拟了不同地形特征的区域之间的转换，如植被和深度雪。参数向量 z 也是学习课程的一部分，其幅度随训练持续时间线性增加。

更详细相关内容看相关附件的 S8。

在学生训练期间，我们随机化每只脚周围绘制的高度样本（图4-4A）。我们扰动每个样本的位置，并将噪声添加到测量的高度值中，如下所示。

$$x_p = r_p \cos(\theta_p) + \epsilon_{px} + \epsilon_{fx} + w_x \quad (4-22)$$

$$y_p = r_p \sin(\theta_p) + \epsilon_{py} + \epsilon_{fy} + w_y \quad (4-23)$$

$$h_p = h(x_p, y_p) + \epsilon_{pz} + \epsilon_{fz} + w_z + \epsilon_{outlier} \quad (4-24)$$

其中 $h(x_p, y_p)$ 表示在点 (x_p, y_p) 处的高度， r_p 是 p 点的径向距离， θ_p 是 p 在脚周围的极坐标中的方位角。 $\epsilon_{px}, \epsilon_{py}, \epsilon_{pz}$ 表示每个时间步长里各个点的采样噪声。 $\epsilon_{fx}, \epsilon_{fy}, \epsilon_{fz}$ 表示每个时间步长里各个脚的采样噪声。 w_x, w_y, w_z 表示每个片段里各个脚的采样噪声。 $\epsilon_{outlier}$ 是间歇性添加的大噪声来模拟异常值。

使用参数 z 从正态分布中采样每个噪声。 $\epsilon_{px}, \epsilon_{py} \sim \mathcal{N}(0, z_0), \epsilon_{pz} \sim \mathcal{N}(0, z_1), \epsilon_{fx}, \epsilon_{fy} \sim \mathcal{N}(0, z_2), \epsilon_{fz} \sim \mathcal{N}(0, z_3), \epsilon_{outlier} \sim \mathcal{N}(0, z_4)$ 概率分别为 $p = z_5, w_x, w_y \sim \mathcal{N}(0, z_6), w_z \sim \mathcal{N}(0, z_7)$ 。

我们为学生训练定义三种情况：*nominal*, *offset*, *noisy*。每种情况的参数 z 定义如下：

$$z_{nominal} = \langle 0.004, 0.005, 0.01, 0.04, 0.03, 0.05, 0.1 \rangle \quad (4-25)$$

$$z_{offset} = \langle 0.004, 0.005, 0.01, 0.1c_{sk}, 0.1c_{sk}, 0.02, 0.1 \rangle \quad (4-26)$$

$$z_{noisy} = \langle 0.004, 0.1c_{sk}, 0.1c_{sk}, 0.3c_{sk}, 0.3c_{sk}, 0.3c_{sk}, 0.1 \rangle \quad (4-27)$$

其中 c_{sk} 是学生课程因子，会随着训练片段增加而不断线性增加。我们在轨迹的开头和中间随机选择其中一个条件。选择的概率分别是 60%、30% 和 10%。

4.4.9 信念状态寄存器

循环信念状态编码器编码不能直接观察到的状态。为了整合本体感受和外部环境数据，我们引入了一个门控编码器，如图 7C 所示，灵感来自门控 RNN 模型^[26-27]和多模态信息融合 (4-66)。

信念状态编码器学习使用一个可变的门控因子来控制外部感知信息通过的量。首先，内部感知 s_t^p 、从含噪声观测提取的外部感知 $l_t^e = g_e(\tilde{o}_t^e)$ 以及隐藏状态 s_t 被 RNN 模型编码成为一个中间信念状态 $b_{t'}^e$ 。它控制最终进入 b_t 的外部感知信息量：

$$b_{t'}, h_{t+1} = RNN(o_t^p, l_t^e, h_t) \quad (4-28)$$

$$\alpha = \sigma(g_a(b_{t'})) \quad (4-29)$$

$$b_t = g_b(b_{t'}) + l_t^e \odot \alpha \quad (4-30)$$

这里 g_a, g_b 是全连接的神经网络， $\sigma(\cdot)$ 是 sigmoid 函数。

解码器使用相同的门，用于重建特权信息和高度样本（图 7D）。这用于计算重建损失，它鼓励信念状态捕获有关环境的真实信息。我们使用 GR^{[26]p} 作为我们的 RNN 架构。

结构有效性的评估见第 S9 节。

4.4.10 部署

我们在 PyTorch^{[28]p} 中训练策略，并在没有任何微调的情况下部署在机器人 zero-shot 上。我们通过估计机器人的姿态，并相应地从传感器中调节点云读数，构建了一个以机器人为中心的 2.5D 高程图刷新率为 20 赫兹。该策略以 50Hz 运行，并从最新的高程图中映射采样高度；如果查询位置没有可用的地图信息，则填充随机采样的值。

我们开发了一个高程映射管道，用于在图形处理单元上快速地形映射，以并行化点云处理。我们遵循与 Fankhauser 等人^{[29]p} 使用的类似方法，以卡尔曼滤波的方式更新地图，另外按漂移补偿 *drift compensation* 和光线投射 *ray casting* 以获得更吻合的地图。这种快速映射实现对于保持快速处理速率和跟上我们的控制器实现的快速运动速度至关重要。

4.5 电机驱动关节点模型^{[4]p4}

基于参考文献阐述电机驱动关节点模型描述...

第 5 章 强化学习部署实践和关键点

Addition 5.0.1. 两个参考的 RL 深度学习案例库：

isaacgym 库

legged_gym 库

5.1 关于强化学习平台 Isaac^[30]p6-10

这部分参考文献介绍对 Isaac 平台的基本理解，主要是它的并行工作流程和提供的读取和设置 API...

5.2 强化学习案例 legged_gym 库^[31]p1

这部分展示 legged_gym 库关于 ANYmal-c 的训练代码，主要关注整个训练流程、奖励函数的设计、对 Isaac-API 的使用、地形设计、训练课程设计等与论文中提到的概念对应...

5.3 PyTorch 的使用

这部分结合 Pytorch 的官方文档和 legged_gym 库提取其中关于 .pth 模型的构建和使用流程和实例...

第6章 几种构想的机器人类型草图

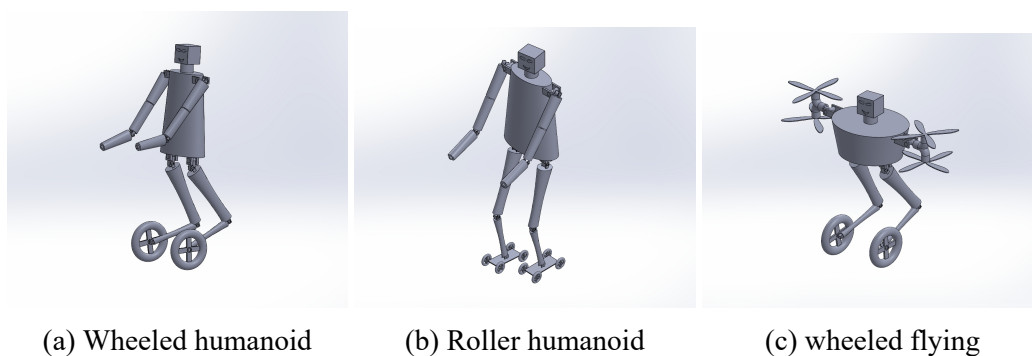


图 6-1 三种构想机器人类型的示意图汇总

如图6-1所示，这里给出几种近来构想的机器人结构草图和简单介绍，可以作为随后研究的机器人备选平台参考。

6.1 弹跳轮足 + 机械手机器人

如图6-2所示，该类型的机器人由类似于 Ascento^{[32]p1} 类型机器人的双轮足，在此基础上添加两个机械臂构成类人形机器人。该类型的机器人可以轮式地行走和跳跃，同时可以用机械臂模仿双手进行各种操作。

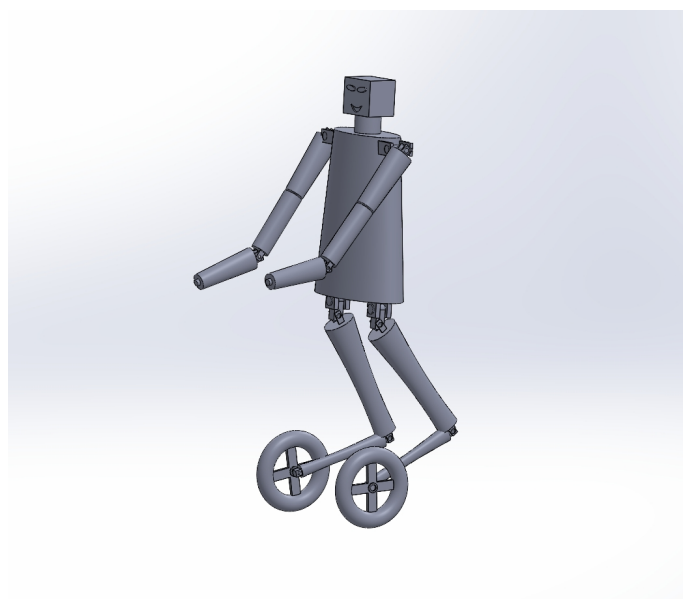


图 6-2 Schematic view of wheeled-humanoid robot

该类型机器人的控制基础是倒立摆模型和机械臂运动学，它相对于现有双轮

机器人，如 Ascento^{[32]p1} 等，新的控制问题难点是：

- (1) 在机械手运动的同时保持双轮足的稳定，比如从空手到搬起重物的过程；
- (2) 如何让手臂配合轮足实现更加灵巧和优雅的动作；

6.2 轮鞋足式 + 机械手机器人

如图6-3所示，该类型的机器人由类似一般人形机器人的基本结构，在此基础上为每个脚添加四个轮子

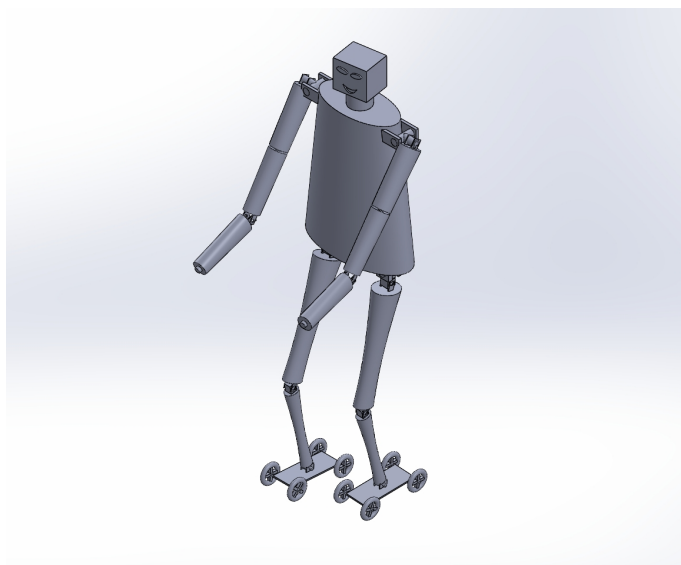


图 6-3 Schematic view of roller humanoid.

该类型机器人的控制基础是 ZMP 约束和机械臂运动学，它相对于现有人形机器人，如^{[33-34]p}等，新的控制问题难点是：

- (1) 如何实现步态和轮式之间的运动切换；
- (2) 如何实现步态和轮式的联合运动平衡；
- (3) 在上面两点的基础上，如何结合机械臂实现更加灵巧、优雅、节能的动作；

6.3 弹跳轮足 + 飞行器机器人

如图6-4所示，该类型的机器人由类似于 Ascento^{[32]p1} 类型机器人的双轮足，在此基础上添加一对旋翼使其具备飞行能力。

该类型机器人的控制基础是倒立摆模型和飞行器控制，它相对于现有双轮机器人，如 Ascento^{[32]p1} 等，新的控制问题难点是：

- (1) 飞行器的矢量控制；

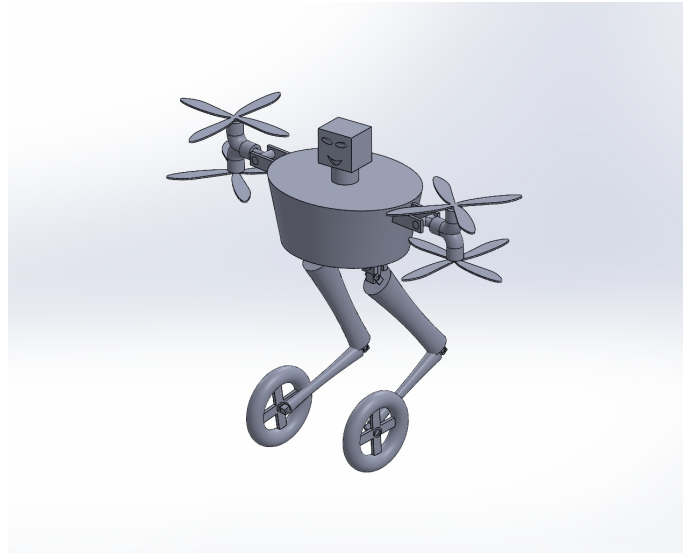


图 6-4 Schematic view of wheeled-flying robot.

- (2) 如何优化设计使得旋翼提供的升力可以满足飞行控制
- (3) 续航安全问题;
- (4) 如何让旋翼配合轮足实现更加灵巧、优雅、节能、稳定的动作;

结 论

对于经典控制的实现,涉及到复杂的运动模型和动力学建模,难点在于选择优化目标构建优化策略和优化问题的求解。经典控制的优点在于控制过程清晰,可以方便地进行迭代优化和泛化部署;缺点在于建模和实际机器调试复杂,运行时算力开销大。对于强化学习控制的实现,涉及到机器人的仿真模型建模和神经网络设计,难点在于设计合适的优化目标和训练策略以使得训练过程能较快收敛且结果达到预期目标。强化学习控制的优点在于模型设计和调试应用简单迅速,运行时算力开销小;缺点在于控制过程不清晰,导致模型的泛化能力较弱。

经典控制与深度学习控制各有优缺点,在实践应用中两者的有机结合有助于实现更好的控制。在面对一种机器人设计控制策略时,如何分配经典控制和深度学习控制是一个十分重要的问题。两者的融合也是未来的重要发展方向之一。

对于接下来的研究,就当前的情况来看,实现强化学习到实际机器人 *ANYmal* 上的部署还欠缺的是:

- (1) 一套 *ANYmal* 机器人硬件;
- (2) 可用的例程代码;
- (3) 电机、控制手柄、各类传感器的驱动和软件接口。
 - 在有直接可用例程的情况下,可以采用例程的架构和电机接口直接部署如4.2节所述的神经网络策略。
 - 在没有直接可用例程有类似参考例程的情况下,可以花一些时间阅读相关代码,逐步将电机、控制手柄、各类传感器等驱动重构到 *ANYmal* 上,然后在此基础上设计控制结构实现策略部署。
 - 在没有任何软件接口的情况下至少应该有各种器件相应的手册。会消耗比较长的时间来完成驱动和信息格式的整理和构建。完成后设计控制结构进行策略部署。
- (4) 对 *Isaac* 仿真环境的深入学习;
- (5) 对 *PyTorch* 框架的深入学习;

对于后续的研究,在上面的基础上需要进一步学习和尝试将经典策略应用到机器人上,以更加深入的理解经典控制方法。与此同时加强对强化学习理论的理解和仿真环境的使用能力,在接下来的研究中构思经典-强化学习融合的控制策略。在此基础上,考虑设计和实现新结构的机器人。

参考文献

- [1] WINKLER A W, BELLICOSO C D, HUTTER M, et al. Gait and Trajectory Optimization for Legged Systems Through Phase-Based End-Effector Parameterization[J/OL]. IEEE Robotics and Automation Letters, 2018: 1560–1567. <http://dx.doi.org/10.1109/lra.2018.2798285>.
- [2] BELLICOSO C, JENELTEN F, FANKHAUSER P, et al. Dynamic locomotion and whole-body control for quadrupedal robots[C/OL]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2017. <http://dx.doi.org/10.1109/iros.2017.8206174>.
- [3] BELLICOSO C D, JENELTEN F, GEHRING C, et al. Dynamic Locomotion Through Online Nonlinear Motion Optimization for Quadrupedal Robots[J/OL]. IEEE Robotics and Automation Letters, 2018: 2261–2268. <http://dx.doi.org/10.1109/lra.2018.2794620>.
- [4] GEHRING C, COROS S, HUTTER M, et al. An Optimization-based Approach to Controlling Agile Motions for a Quadruped Robot[J/OL]. IEEE Robotics and Automation, 2016. DOI: [10.3929/ethz-a-010644954](https://doi.org/10.3929/ethz-a-010644954).
- [5] KALAKRISHNAN M, BUCHLI J, PASTOR P, et al. Fast, robust quadruped locomotion over challenging terrain[C/OL]//2010 IEEE International Conference on Robotics and Automation. 2010. <http://dx.doi.org/10.1109/robot.2010.5509805>.
- [6] GOLDFARB D, IDNANI A. A numerically stable dual method for solving strictly convex quadratic programs[J/OL]. Mathematical Programming, 1983, 27(1): 1–33. <http://dx.doi.org/10.1007/bf02591962>.
- [7] DARIO BELLICOSO C, GEHRING C, HWANGBO J, et al. Perception-less terrain adaptation through whole body control and hierarchical optimization[C/OL]//2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids). 2016. <http://dx.doi.org/10.1109/humanoids.2016.7803330>.
- [8] PARDO D, NEUNERT M, WINKLER A, et al. Hybrid direct collocation and control in the constraint-consistent subspace for dynamic legged robot locomotion[C/OL]//Robotics: Science and Systems XIII. 2017. <http://dx.doi.org/10.15607/rss.2017.xiii.042>.
- [9] NEUNERT M, STAUBLE M, GIFTTHALER M, et al. Whole-Body Nonlinear Model Predictive Control Through Contacts for Quadrupeds[J/OL]. IEEE Robotics and Automation Letters, 2018: 1458–1465. <http://dx.doi.org/10.1109/lra.2018.2800124>.
- [10] LEE J, BJELONIC M, HUTTER M. Control of Wheeled-Legged Quadrupeds Using Deep Reinforcement Learning[M/OL]. 2023: 119–127. http://dx.doi.org/10.1007/978-3-031-15226-9_14.
- [11] HWANGBO J, LEE J, DOSOVITSKIY A, et al. Learning agile and dynamic motor skills for legged robots[J/OL]. Science Robotics, 2019. <http://dx.doi.org/10.1126/scirobotics.aau5872>.
- [12] LEE J, HWANGBO J, WELLHAUSEN L, et al. Learning Quadrupedal Locomotion over Challenging Terrain[J/OL]. Science Robotics, 2020. <http://dx.doi.org/10.1126/scirobotics.abc5986>.

-
- [13] CHEN D, ZHOU B, KOLTUN V, et al. Learning by Cheating[J]. Conference on Robot Learning, Conference on Robot Learning, 2019.
- [14] ISCEN A, CALUWAERTS K, TAN J, et al. Policies Modulating Trajectory Generators[J]. Conference on Robot Learning, Conference on Robot Learning, 2018.
- [15] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust Region Policy Optimization[A]. 2015.
- [16] BAI S, KOLTER J, KOLTUN V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling[A]. 2018.
- [17] BARASUOL V, BUCHLI J, SEMINI C, et al. A reactive controller framework for quadrupedal locomotion on challenging terrain[C/OL]//2013 IEEE International Conference on Robotics and Automation. 2013. <http://dx.doi.org/10.1109/icra.2013.6630926>.
- [18] LEE J, HWANGBO J, HUTTER M. Robust Recovery Controller for a Quadrupedal Robot using Deep Reinforcement Learning.[A]. 2019.
- [19] HWANGBO J, BELLICOSO C, FANKHAUSER P, et al. Probabilistic foot contact estimation by fusing information from dynamics and differential/forward kinematics[J]. Intelligent Robots and Systems, Intelligent Robots and Systems, 2016.
- [20] MIKI T, LEE J, HWANGBO J, et al. Learning robust perceptive locomotion for quadrupedal robots in the wild[J/OL]. Science Robotics, 2022(VOL. 7, NO. 62). DOI: DOI:, [10.1126/scirobotics.abk2822](https://doi.org/10.1126/scirobotics.abk2822).
- [21] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with Deep Reinforcement Learning[A]. 2013.
- [22] ZHU P, LI X, POUPART P, et al. On Improving Deep Reinforcement Learning for POMDPs [A]. 2017.
- [23] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal Policy Optimization Algorithms [A]. 2017.
- [24] ROSS S, GORDON G, BAGNELL J. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning[A]. 2010.
- [25] CZARNECKI W, PASCANU R, OSINDERO S, et al. Distilling Policy Distillation[A]. 2019.
- [26] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation[C/OL]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. <http://dx.doi.org/10.3115/v1/d14-1179>.
- [27] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J/OL]. Neural Computation, 1997: 1735–1780. <http://dx.doi.org/10.1162/neco.1997.9.8.1735>.
- [28] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library: Article No.: 721[M/OL]. 2019: Pages 8026-8037. DOI: <https://dl.acm.org/doi/10.5555/3454287.3455008>.
- [29] FANKHAUSER P, BLOESCH M, HUTTER M. Probabilistic Terrain Mapping for Mobile Robots With Uncertain Localization[J/OL]. IEEE Robotics and Automation Letters, 2018: 3019–3026. <http://dx.doi.org/10.1109/lra.2018.2849506>.

-
- [30] MAKОВIYCHUK V, WAWRZYNIAK L, GUO Y, et al. Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning[A]. 2021.
 - [31] RUDIN N, HOELLER D, REIST P, et al. Learning to Walk in Minutes Using Massively Parallel Deep Reinforcement Learning[A]. 2021.
 - [32] KLEMM V, MORRA A, SALZMANN C, et al. Ascento: A Two-Wheeled Jumping Robot [C/OL]//2019 International Conference on Robotics and Automation (ICRA). 2019. <http://dx.doi.org/10.1109/icra.2019.8793792>.
 - [33] HERZOG A, ROTELLA N, SCHAAAL S, et al. Trajectory generation for multi-contact momentum control[C/OL]//2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids). 2015. <http://dx.doi.org/10.1109/humanoids.2015.7363464>.
 - [34] DIETRICH A, BUSSMANN K, PETIT F, et al. Whole-body impedance control of wheeled mobile manipulators Stability analysis and experiments on the humanoid robot Rollin' Justin [J/OL]. Springer Science+Business, 2015, Media New York 2015. DOI: [10.1007/s10514-015-9438-z](https://doi.org/10.1007/s10514-015-9438-z).