

- 小问 3 求解步骤：协同控制 (PPO 算法)
 - 1. 数据输入
 - 2. 参数初始化 (PPO 配置)
 - 3. 模型调用 (训练与推理)
 - 4. 结果输出

小问 3 求解步骤：协同控制 (PPO 算法)

1. 数据输入

- 加载物理模型：导入小问 1、2 生成的关节参数和机器人 URDF 文件。
- 定义目标轨迹：
 - 上肢：画圆轨迹坐标序列。
 - 下肢：维持重心的支撑多边形范围。
 - 整体：左转 45° 的目标朝向。

2. 参数初始化 (PPO 配置)

- 超参数设置：
 - 学习率 (Learning Rate): 3×10^{-4} (Adam 优化器标准值)。
 - 折扣因子 (γ): 0.99。
 - Clip 范围 (ϵ): 0.2 (PPO 核心参数，通常不需修改)。
 - Batch Size: 64 或 128。
 - 训练步数: 1,000,000 steps (视收敛情况而定)。
- 环境重置：将机器人置于初始站立姿态，所有关节归零。

3. 模型调用 (训练与推理)

- 步骤 3.1：环境交互 (Rollout)
 - Agent 观察当前状态 S_t (30+维)。
 - Policy 网络输出动作分布，采样得到动作 A_t (关节增量)。
 - 环境执行 A_t ，物理引擎计算下一帧状态 S_{t+1} 。
 - 计算奖励 R_t ：

$$R = 1.0(\text{存活}) - 0.5 \parallel \text{Error}_{\text{traj}} \parallel - 0.3 \parallel \text{Error}_{\text{balance}} \parallel$$

注意：奖励权重需精细调节，若重心惩罚过大，机器人可能“不敢动”；若轨迹奖励过大，可能导致摔倒。

- **步骤 3.2：网络更新**
 - 收集一定长度的轨迹数据 (Trajectory)。
 - 计算优势函数 (GAE)。
 - 最大化 PPO 目标函数，更新 Actor 和 Critic 网络参数。
- **步骤 3.3：策略验证**
 - 每训练 1000 次，进行一次测试（关闭动作噪声）。
 - 检查机器人是否在完成画圆的同时成功左转且未摔倒。

4. 结果输出

- **训练曲线：**绘制 Reward 随 Episode 变化的曲线，证明算法收敛。
- **关键数据：**输出左转完成时的总时间、重心最大偏移量。
- **可视化：**生成机器人动作的 3D 动画或关键帧截图（展示“手舞足蹈”的协同效果）。