Animação de Mãos como Controle de Cadeias Cinemáticas via Aprendizado por Reforço

Marcelo Augusto Salomão Ganem
Departamento de Ciência da Computação (UFMG)
Belo Horizonte, Minas Gerais
marceloganem@dcc.ufmg.br

Resumo — Este trabalho apresenta uma investigação de técnicas de reward shaping e curriculum learning no contexto de aprendizado por reforço para o controle puramente cinemático de uma mão, sem a necessidade de dados anotados ou aprendizado explícito da dinâmica do ambiente. Modelamos o problema como um Processo de Decisão de Markov em que a política, parametrizada por uma rede neural, gera diretamente as variações angulares das juntas da mão para tocar pontos-alvo posicionados aleatoriamente sobre um objeto. A função de recompensa é composta por três termos contínuos e diferenciáveis: contato, colisões e limites articulares, ponderados de forma a priorizar o estabelecimento de contato estável antes de refinar aproximações e penalizar movimentos inválidos. Aplicamos Proximal Policy Optimization (PPO) com respeito à função de recompensa, obtendo políticas capazes de realizar (em pelo menos 50% dos cenários apresentados) contatos precisos, respeitando limites anatômicos, com movimentações pouco sujeitas a jitter e orientações imprevisíveis. Demonstramos que o reward shaping e o curriculum learning melhoram a estabilidade e a eficiência do aprendizado, embora a alta dimensionalidade ainda imponha desafios de generalização.

Palavras-chave — reward shaping, reinforcement learning, animation, kinematics

I. Introdução

Soluções para a imitação adequada de movimentos observados em entidades complexas têm sido investigadas na literatura de aprendizado por reforço desde o início da última década e se desenvolvido desde então. Modelos como GAIL [1] e DeepMimic [2] são capazes de combinar objetivos circunstanciais e referências de movimento para produzir simulações com realismo expressivamente superior às primeiras soluções apresentadas.

Uma característica comum de modelos que consideram objetivos e referências é a demanda evidente por uma base de dados rotulados e a ruptura com um paradigma de aprendizado por reforço puro. A exploração de técnicas de reward shaping para produzir modelos igualmente eficazes mas independentes de extensos dados de treino ainda é uma tarefa corrente, como mostram os trabalhos de Harutyunyan et al. (2015) [3] e Zou et al. (2019) [4].

A. Contribuições

Este trabalho busca contribuir com a investigação e desenvolvimento de soluções puramente baseadas em aprendizado por reforço para controle de cadeias cinemáticas. Essas características são escolhidas justamente pela composição de um problema ainda não extensivamente investigado sob o prisma de aprendizado por reforço não-supervisionado.

Formulamos também, de maneira sucinta, a tarefa de controle de cadeias cinemáticas no espaço $\mathbb{R}^3$ como um Processo de Decisão de Markov. Essa definição e sua implementação, aqui dimensionadas para o controle de uma mão com 16 juntas, podem ser facilmente modificadas para atender cadeias cinemáticas arbitrárias. Semelhantemente, a metodologia proposta para o aprendizado de máquina da tarefa de controle é facilmente extensível para cadeias e requisitos diversos via customização das funções de recompensa e dos cenários de treino.

Para além do campo teórico, aplicações das soluções aqui estudadas incluem efeitos visuais para cinema, animação e jogos eletrônicos, planejamento de trajetórias em robótica e análises biométricas – em geral, cenários onde uma sequência temporal das diferentes posições de um agente executando uma atividade específica é desejável.

II. Trabalhos Relacionados

A. Reward Shaping

Wiewiora et al. (2003) [5] demonstram que, no contexto de aprendizado por reforço, é possível introduzir planejamento via termos adicionais na função de recompensa de maneira a preservar a política ótima (determinada por um objetivo central). Neste trabalho, exploramos a estratégia de reward shaping para induzir comportamentos desejáveis no agente. Especificamente, atendemos os requisitos do cenário: coerência anatômica, aproximação entre junta e ponto-alvo, baixo jitter[1] e coerência física (não-colisão) via construção intencional de múltiplas funções de recompensa, manipuladas aritmeticamente para favorecer o aprendizado via descida de gradiente e integradas por soma ponderada na função final, como descrito na Seção IV-D.

Espera-se, portanto, que a introdução dessas recompensas não interfira na política ótima – aqui definida como a prioridade primária no contato entre as juntas e os alvos.

B. Curriculum Learning

A premissa de que um agente é capaz de aprender uma representação útil da política ótima de um dado

[1] Em tradução literal: tremor, i.e., movimentos em direções alternantemente opostas.

problema final a partir de uma versão simplificada do problema é firmada no conceito de curriculum learning. Especificamente, Bengio et al.[6] formalizam a intuição de que apresentar tarefas de dificuldade crescente em ordem durante o treinamento de um agente é mais efetivo do que apresentar o problema completo.

Aqui, utilizamos esse princípio como fundamento para a estratégia de apresentar ao agente cenários inicialmente mais permissivos, restringindo gradativamente os requisitos do problema apresentado. Especificamente, construímos uma estratégia de anelamento do parâmetro ϵ que controla (não-linearmente) a magnitude da recompensa de contato. Como ilustrado na Figura 1, valores de ϵ mais altos aumentam a recompensa por contatos de menor precisão (i.e., a maior distância do alvo).

C. DeepMimic

Esse estudo utiliza aprendizado por reforço para desenvolver um agente capaz de produzir realistic physics-based character skills, i.e., desenvolver uma política capaz de considerar um objetivo geral (e.g., se locomover a um ponto x, y, z) e um objetivo de imitação – características chave para um modelo que deve ao mesmo tempo ser realista e responder de maneira dinâmica a problemas variados.

O ambiente é modelado a partir de uma simulação de física dinâmica, de maneira que o agente define valores-alvo para as velocidades lineares e angulares das juntas de uma entidade por meio de um controlador de derivada parcial – uma ferramenta útil para abstrair detalhes físicos do escopo do agente. O estado é representado como um conjunto de juntas e suas respectivas posições, orientações e velocidades lineares e angulares.

O DeepMimic também se diferencia pela estratégia de reward shaping, tal que a função de recompensa r utilizada é a média ponderada entre uma recompensa de objetivo r^G e uma recompensa de imitação r^I, definida como uma média ponderada de funções específicas para posição, velocidade, posição das mãos e pés e centro de massa.

O aprendizado é feito utilizando Proximal Policy Optimization para otimizar os pesos θ de uma rede neural que representa a política π. Semelhantemente, os pesos ψ de uma rede neural que representa a função de valor V são otimizados utilizando TD(λ). Os resultados atingidos ultrapassaram expressivamente o estado-da-arte à época da publicação.

O artigo de Peng et al. [2] introduz decisões de design na solução do problema proposto que atenuam a maior parte das dificuldades encontradas pela literatura até então. Especificamente, a representação formal do corpo do agente como um conjunto de juntas no espaço e suas propriedades específicas é valiosa na concepção deste trabalho de curso.

D. Learning Dexterous In-Hand Manipulation

Nesse trabalho, [7] PPO é aplicada ao cenário de uma mão robótica com 24 eixos para executar reorientações arbitrárias de um cubo puramente por meio de aprendizado por reforço. Executando milhares de simulações randomizadas em paralelo[2] e usando uma função de recompensa escalonada, seguindo uma estratégia de curriculum que inicialmente incentiva o alinhamento de uma única face antes de exigir a orientação completa do cubo, eles demonstram que o PPO consegue otimizar de forma estável uma política de 60 milhões de parâmetros mesmo sob alta dimensionalidade e contatos estocásticos.

Analogamente, em vez de mapear observações para torques, a política desenvolvida no caso puramente cinemático abordado no presente trabalho deve gerar diretamente os ângulos das juntas da mão e do objeto como uma sequência temporal. Sendo o caso aqui proposto um equivalente simplificado e com muito menos graus de liberdade, espera-se que a composição de uma única função de recompensa, junto à estratégia de curriculum learning, seja suficiente para garantir convergência. Por fim, o estudo providencia justificativas para o uso de PPO no problema abordado, destacando a estabilidade no aprendizado da política para sequências temporais de ações.

III. Objetivos

Genericamente, investigamos a aplicabilidade de soluções puramente baseadas em aprendizado por reforço, sem dados anotados e sem a construção de um modelo de transição do MDP (i.e., estratégias model-free) na tarefa de controle de cadeias cinemáticas no espaço tridimensional.

Para isso, definimos um cenário simulado onde temos o objetivo específico de aprender, para qualquer configuração do problema, a orientar uma mão no espaço $\mathbb{R}^3$ a partir do controle dos ângulos $\theta_j \in \mathbb{R}^3$ para cada uma de suas 16 juntas, de maneira a tocar um objeto alvo O em um ponto alvo τ_j posicionado aleatoriamente na superfície do objeto e correspondente a uma junta $j^\tau \in J$ aleatoriamente selecionada[3]. Essa tarefa promove uma investigação de proveito geral sobre a aplicação de aprendizado por reforço para produção de animações no contexto de cadeias cinemáticas.

IV. Metodologia

Sintetizamos a metodologia do seguinte trabalho em: modelar a tarefa do controle de uma mão em cinemática direta como um MDP; otimizar a política via Proximal Policy Optimization [8], considerando:

1) Objetivo no cenário: tocar um ponto τ_j com a junta j^τ no espaço
2) Reward shaping: manter anatomia e movimentos visualmente coerentes.

[2]Inspiração crucial para a paralelização de ambientes de treino implementada no presente trabalho, que viabiliza a execução de milhões de etapas de treino a baixo custo computacional.

[3]Dentre as pontas dos dedos.

e aplicando curriculum learning aleatorizado e em etapas. Então, devemos verificar a qualidade das animações produzidas:

1) O agente consegue alcançar o objeto?
2) Os limites anatômicos são obedecidos?
3) Os movimentos são bruscos ou imprevisíveis?
4) A manipulação da cadeia cinemática parece intencional e coerente?

A. Modelagem

Modelamos o problema da animação de uma mão como o controle dos ângulos $\theta_{j,i}$ para cada grau de liberdade i de cada junta j que faz parte da cadeia cinemática J. Incorporando um limite anatômico básico, limitamos as juntas (com exceção do pulso) a um único grau de liberdade, correspondente à flexão/extensão dos dedos.

A dinâmica que determina a posição e orientação de cada um dos componentes da mão no espaço tridimensional é o processo de cinemática direta – por sua vez, determinado pela relação hierárquica e posicional entre as juntas, representada pela função $\phi(j) : J \to J$ (indicando a junta antecessora) e pela matriz $D_{4\times4}$ (uma matriz de translação em $\mathbb{R}^3$ na forma homogênea), respectivamente.

B. Ações

Essa modelagem permite a definição de um MDP onde ações $A_t \in \mathcal{A}$ determinam a variação em θ a cada instante t, tal que

$$A_t = \{\Delta\theta_t\}.$$

C. Estados e dinâmica

Cada estado $S_t = \{W_t, O\} \in \mathcal{S}$ contém o conjunto das transformações espaciais $W_{j,t} \in \mathbb{R}^{4\times4}$ que representam a posição e orientação de cada $j \in J$, determinadas pela função $W(\theta_t) : \mathbb{R}^{16\times3} \to \mathbb{R}^{16\times4\times4}$ que implementa a cinemática direta:

$$p(S_{t+1}|S_t, A_t) = \begin{cases} 1 \text{ para } S_{t+1} = \{W(\theta_t),\ O\} \\ 0 \text{ caso contrário.} \end{cases}$$

A função $W(\theta_t)$ toma como parâmetros implícitos a configuração estática da mão representada por ϕ e D. Omitindo a notação temporal por conveniência, determinamos as transformações espaciais conforme:

$$W_j(\theta) = W_{\phi(j)}(\theta)\ D_j\ R(\theta_j),$$

sendo $R(\theta_j) : \mathbb{R}^3 \to \mathbb{R}^{4\times4}$ a função que implementa a fórmula de Rodrigues para rotações arbitrárias em cada um dos graus de liberdade da junta.

A parte do estado denotada por $O \in (\mathbb{R}^{4\times4}, \mathbb{R}^{16\times4})$ representa a transformação espacial homogênea correspondente ao objeto-alvo, bem como a especificação dos pontos de contato para cada junta. Assim, temos:

$$O = \{W^O, \tau_t\},$$

tal que $\tau_{j,t} \in \mathbb{R}^4$ é a posição relativa a W^O desejada para a junta j. Especificamente, τ_j tem a seguinte forma com relação a $\mathbb{R}^3$:

$$\tau_j = \begin{cases} (x, y, z, 1) \text{ caso } j \text{ deva tocar o objeto} \\ (x, y, z, 0) \text{ caso } j \text{ deva ignorar o objeto.} \end{cases}$$

Dessa maneira, a definição de S_t para cada t de um episódio completo é justamente o roteiro de uma animação onde o agente controla a cadeia cinemática com vistas a a aproximar j^τ de τ_{j^τ} (na prática, tocar o alvo com a ponta de dedo selecionada).

D. Recompensas

Determinamos as recompensas como uma soma ponderada entre três componentes, ilustrados na Figura 1, que representam aspectos relevantes para a naturalidade e qualidade da animação final. Especificamente, avaliamos: contato ao ponto-alvo, resposta a colisões e limites articulares conforme pesos empiricamente definidos por ρ, tal que:

$$r(S_t) = \sum_{r_i \in \{r_C,\, r_{\text{CD}},\, r_L\}} \rho_i r_i(S_t)$$

O componente r_C define a recompensa do agente por contato com pontos-alvo. Sendo p_j a posição absoluta da junta j e q_j a posição determinada pelos primeiros três índices de τ_j transformada por W^O, temos uma função de recompensa contínua e diferenciável:

$$r_C(S_t) = \sum_{j \in J^\tau} \frac{1}{1 + (\|p_j - q_j\|/\epsilon)^2}$$

onde J^τ é o conjunto das juntas $j \in J$ tal que o último índice de τ_j é 1 e ϵ é um parâmetro de escala que controla a sensibilidade da recompensa à distância. Essa formulação permite fornecer gradientes úteis para o aprendizado mesmo quando o agente está longe do alvo, viabilizando o aprendizado inicial – dada a dimensionalidade do problema, é extremamente improvável que o agente acerte a manipulação da cadeia cinemática de maneira a tocar o alvo por acaso (ou por exaustão de tentativas).

As colisões são detectadas assumindo que o objeto em W^O é uma esfera de raio $r_c = 0.3$. A penalidade por colisão utiliza uma função quártica que cresce rapidamente na direção negativa conforme a junta se aproxima do objeto:

$$r_{\text{CD}}(S_t) = -\sum_{\substack{j \in J \\ \|p_j - p_O\| < r_c}} \left[1 - \left(\frac{\|p_j - p_O\|}{r_c}\right)^4\right]$$

Sendo J^+ o conjunto de juntas com exceção do pulso, o componente r_L pune a violação de limites articulares. Especificamente, para cada junta dos dedos, induzimos o ângulo $\theta_{j,x}$ a obedecer os limites anatômicos $\theta_{j,x} \in [-\frac{2\pi}{3}, \frac{\pi}{2}]$ por meio de uma penalidade suave:

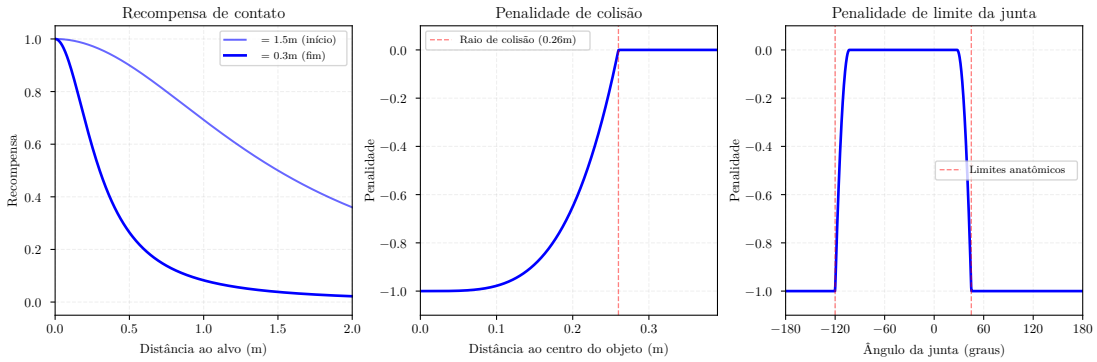

Figura 1. Funções de recompensa

$$r_L(S_t) = -\sum_{j \in J^+} \min\left(1, \left(\frac{\max(0, \theta_{j,x} - \frac{\pi}{2}, -\frac{2\pi}{3} - \theta_{j,x})}{\delta}\right)^2\right)$$

onde $\delta = 0.3$ controla a suavidade da transição entre as regiões válida e inválida do espaço de configuração.

E. Aprendizado

Cada episódio é truncado após, no máximo, $t_{\max} = 128$ passos (mais cedo, em caso de contato sustentado por ≥ 5 timesteps), e a recompensa total é calculada como soma ponderada dos componentes de contato ($\rho_{r_C} = 20$), colisão ($\rho_{r_{CD}} = 4$) e limites articulares ($\rho_{r_L} = 2$), refletindo prioridade máxima em estabelecer um contato estável antes de refinar a aproximação e impor penalidades crescentes para colisões ou violações de ângulo.

Para otimizar a política nesses cenários de alta dimensionalidade e espaço de ação contínuo, empregamos o método Proximal Policy Optimization (PPO) [8], assim como o DeepMimic [2]. A implementação do ambiente no contexto de aprendizado por reforço é essencialmente a função de cinemática direta descrita na Seção IV-C.

Duas redes neurais profundas, com duas camadas internas de 64 perceptrons, foram construídas para implementar a política $\pi : \mathcal{S} \to \mathcal{A}$ e a função de valor $V : \mathcal{S} \to \mathbb{R}$, que produzem a ação a ser tomada e o valor observado no estado (estimativas atualizadas a partir das recompensas observadas) a cada instante t.

O modelo recebe observações do ambiente contendo a posição da mão, posição do objeto e configuração do ponto-alvo (achatadas para um vetor em $\mathbb{R}^{77}$) como entrada de duas redes neurais. As duas redes distintas lêem o estado do ambiente e produzem, respectivamente, a ação a ser tomada no ambiente e o valor do estado observado. A ação interfere no ambiente por meio da dinâmica, determinando o próximo estado, e o valor do estado atual é utilizado para otimizar os pesos do actor em função do gradiente da política, determinado pela função de perda. Os pesos do critic, por sua vez, são semelhantemente otimizados para aproximar a função de valor real com base nas observações.

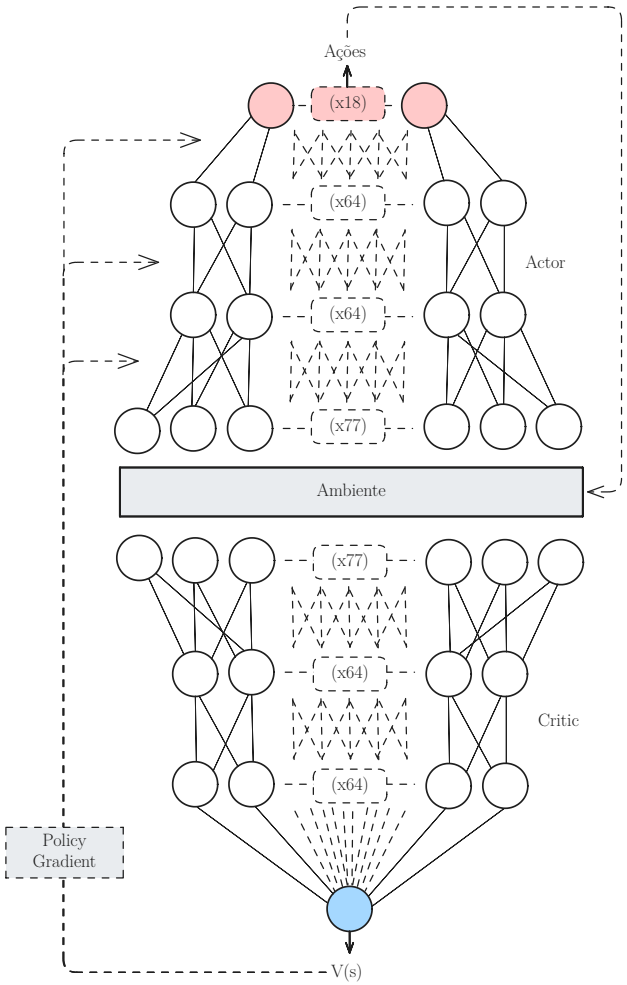

Figura 2. Ilustração simplificada da implementação de PPO [8].

Figura 3. Recompensas obtidas ao longo do treinamento.

Essa otimização é realizada executando iterativamente o processo ilustrado na Figura 2, utilizando o algoritmo Adam [9] para descida de gradiente. Especificamente, os resultados apresentados foram obtidos após 896.000 timesteps de treino, executando episódios de até 128 timesteps (terminados antes caso o contato seja atingido e mantido) em 16 ambientes paralelos, coletando batches com os estados e recompensas observados em 2048 timesteps.

Aplicamos o conceito de curriculum learning [6] por meio do anelamento do parâmetro ϵ que controla a sensibilidade da recompensa de contato. O valor de ϵ é reduzido linearmente de 1.5 para 0.3 ao longo do treinamento. Como evidenciado na Figura 3, valores maiores de ϵ aumentam a recompensa por contatos de menor precisão, tornando o problema inicialmente mais permissivo e facilitando o aprendizado inicial. Gradativamente, à medida que ϵ decresce, o requisito de precisão aumenta, forçando o agente a refinar sua política para alcançar contatos mais precisos.

V. Resultados

Mensuramos o progresso do aprendizado na tarefa proposta por meio da decomposição das recompensas em seus componentes ao longo do tempo; semelhantemente, acompanhamos a função de perda para verificar a ocorrência de otimização como um teste de sanidade. Adicionalmente, monitoramos a distância mínima entre o alvo τ_j e a junta j^τ (por episódio, em uma média móvel $n = 40$) e a taxa de sucesso cumulativa, i.e., a razão entre todos os episódios executados e aqueles terminados antes de $t = 128$ devido ao contato (dentro do threshold $\epsilon_{\text{min}} = 0.3$) sustentado por pelo menos 5 instantes.

A Figura 3 ilustra tendências características de aprendizado bem sucedido para o problema proposto. Na ausência de aprendizado do controle da cadeia cinemática, a recompensa de contato decresceria abaixo da linha tracejada que monitora ϵ, vide sua fórmula definida na Seção IV-D e ilustrada na Figura 1; semelhantemente, a distância mínima alcançada entre a ponta de dedo e o alvo não decresceria ao longo do treino.

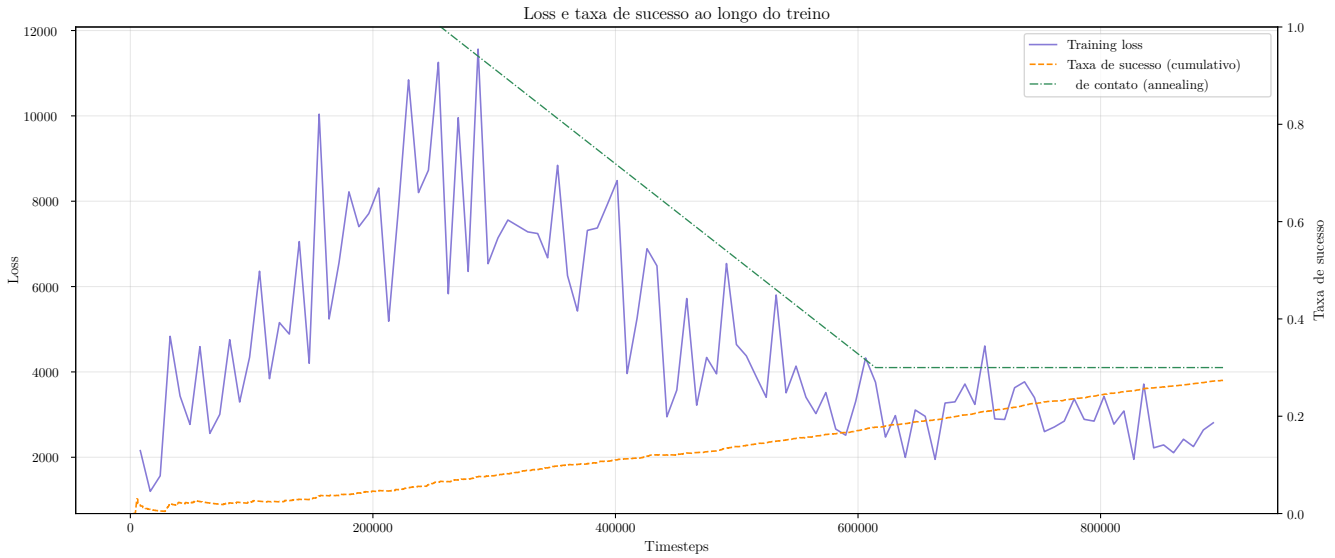

Figura 4. Função de perda e taxa de sucesso cumulativa ao longo do treinamento.

Observamos, de fato, que a recompensa de contato média segue crescendo e se estabiliza mesmo com o anelamento do parâmetro ϵ; observamos também a redução das penalidades por violação de limites anatômicos, quase totalmente evitados ao final do treino; a penalidade por colisão só aparece ocasionalmente durante o treino, mas sua magnitude também decresce com o tempo. Importantemente, a distância mínima atingida decresce consistentemente até $\approxeq 0.2$ o final do treino. Estas observações não indicam sucesso completo (especialmente com respeito à generalização para quaisquer instâncias do problema), mas sugerem pelo menos a manipulação direcionada da cadeia cinemática para atender os objetivos estabelecidos pelas funções de recompensa.

Analisando a Figura 4, observamos que a função de perda tem o caráter irregular típico de métodos baseados em gradiente de política, fortalecido pela recursividade estrutural da estratégia actor-critic (utilizamos as saídas de uma rede neural para treinar a outra) e pelo anelamento do parâmetro ϵ, que implica em uma função de recompensa não-estacionária e, portanto, mais difícil de se aproximar. No geral, entretanto, a função decresce até um mínimo próximo do início do treino, mas com $\epsilon = 0.3$ em vez de $\epsilon = 1.5$, adaptada ao requisito de maior precisão no contato.

Ainda, a taxa de sucesso cresce consistentemente – o que é esperado, naturalmente, para uma razão cumulativa – até 30% ao final do treino. Avaliado em 1000 episódios, o modelo obtido ao final do treino ilustrado nas Figuras 3 e 4 obteve uma taxa de sucesso de 50%. As animações produzidas estão disponíveis como sequências de transformadas, estruturadas para renderização via Blender[4], junto ao código-fonte do trabalho[5] e, para conveniência, também na plataforma YouTube[6].

Uma análise qualitativa das animações produzidas, técnica utilizada recorrentemente ao longo do desenvolvimento do trabalho para identificar direções de evolução e correção, revela a ausência do problema de jitter, comumente enfrentado em trabalhos estado-da-arte ([1], [2]); além disso, os movimentos produzidos são visualmente orgânicos em virtude da própria modelagem do problema, dispensando termos de eficiência de movimento ou adaptações específicas contra jitter. De fato, os limites anatômicos são raramente violados; por outro lado, não é difícil amostrar casos onde o modelo ainda obtém a penalidade de colisão[7]. No geral, a cadeia é manipulada com sucesso para aproximar a ponta de dedo selecionada do alvo aleatoriamente instanciado, mas em metade dos casos o modelo não atinge precisão completa. Atribuições dos sucessos e falhas observados nesta seção a decisões da metodologia proposta são elaboradas na Seção VI.

Tabela I
Hiperparâmetros utilizados na PPO

Hiperparâmetro	Valor
Taxa de aprendizado	1.6×10^{-3}
Fator de desconto (γ)	0.99
Suavização GAE (λ)	0.95
Ambientes paralelos	16
Tamanho do batch	2048
Intervalo de clipping	0.25
Coeficiente de entropia	0.005

[4]Software de modelagem e animação 3D de código aberto, disponível em https://github.com/blender/blender.

[5]Disponível em https://github.com/masganem/msi2.

[6]Disponível em https://youtu.be/ZsoqlpEgEbo.

[7]Intuitivamente, devido à inerente baixa probabilidade de colisão durante o treino.

VI. Considerações

A. Do problema de generalização

A tarefa de controle de cadeias cinemáticas no espaço $\mathbb{R}^3$ é um desafio de inerente alta dimensionalidade, proporcional ao número de graus de liberdade coordenados e à complexidade da estrutura da cadeia em si (vide $D_{4\times4\times|J|}$ e $\phi : J \to J$ definidas na Seção IV-A). Apesar da dinâmica constante, os cenários apresentados a cada episódio têm o objeto e o alvo aleatoriamente instanciados no espaço; além disso, a junta j^τ é amostrada aleatoriamente dentre as pontas dos dedos.

Essa estratégia, junto à execução de múltiplos ambientes de treino em paralelo (possível pela indiferença com respeito à ordenação dos episódios de treino[8]), é adotada com vistas aprender uma política genérica via amostragem do espaço total de possíveis configurações do problema. A taxa de sucesso examinada na Seção V sugere que tal generalização não é atingida; de fato, analisando qualitativamente as animações, observamos casos onde o agente só se aproxima parcialmente do alvo e para de se movimentar, ou atinge um ponto que equilibra penalidades e recompensas (e.g., alcançar o alvo com uma configuração anatômica inválida) – este último sendo um caso marginal.

A ordenação dos cenários apresentados, refletindo uma estratégia de curriculum [6], inviabiliza o treino paralelizado e não promove resultados melhores, mesmo coletando múltiplos batches[9] antes de cada atualização dos pesos; na verdade, o agente se enviesa para uma configuração específica do problema a cada novo episódio. A Seção VII elabora alternativas para a solução do problema de generalização sem comprometer a eficiência do processo de treino.

B. Da confecção das funções de recompensa

Um ponto limitante da abordagem adotada está relacionado à própria formulação manual das funções de recompensa e à escolha empírica de seus pesos. Embora tais decisões sejam essenciais para induzir propriedades desejáveis, elas também introduzem vieses no comportamento aprendido. Na prática, o agente pode estar otimizando particularidades da superfície multidimensional estabelecida pelo gradiente da função de recompensa, após somados todos os seus componentes, em vez de adquirir uma noção mais geral de controle da cadeia cinemática.

Além disso, alterações nos pesos ρ ou na forma aritmética dos termos de recompensa podem resultar em políticas significativamente distintas, indicando sensibilidade elevada à configuração de reward shaping. Esse fenômeno sugere que parte do desempenho obtido decorre mais do cenário de treinamento cuidadosamente construído do que de uma capacidade intrínseca de generalização do método.

[8]Refletindo a estratégia adotada por Andrychowicz et al. (2019) na tarefa de Dexterous In-Hand Manipulation [7].

[9]Conjuntos de pares $\{S_t, r(S_t)\}$ contemplando múltiplos episódios.

C. Da comparação com a cinemática inversa

Embora métodos tradicionais de cinemática inversa ofereçam soluções determinísticas e eficientes para o posicionamento de cadeias articuladas, sua formulação tende a se distanciar do comportamento orgânico que buscamos reproduzir neste trabalho. A abordagem proposta, ao operar diretamente sobre variações angulares sucessivas e propagar transformações via cinemática direta, aproxima-se mais de um modelo natural de movimento, no qual a trajetória emerge da interação temporal entre estados.

Em contraste, soluções de cinemática inversa frequentemente tratam o problema como uma busca instantânea por uma configuração final válida, sem considerar a evolução contínua das poses intermediárias. Deve-se a isso sua eficiência computacional, mas também o caráter robótico das animações produzidas por tais métodos. Além disso, a extensibilidade é uma contribuição central da metodologia proposta, visto que o uso de funções de recompensa configuráveis permite induzir comportamentos arbitrários, desde restrições anatômicas até preferências estilísticas de movimento – adaptações não triviais sob o paradigma da cinemática inversa.

VII. Trabalhos Futuros

Continuações do presente trabalho devem investigar estratégias que ampliem a capacidade do agente de generalizar sobre o espaço completo de configurações gerado pela aleatorização de alvos e seleção de juntas. Centralmente, pode-se adaptar o currículo de treino considerando a não equivalência de todas as configurações aleatórias, i.e., o fato de que alguns cenários são mais ou menos desafiadores a depender da estrutura da cadeia cinemática ou da junta j^τ selecionada. Amostragens mais estruturadas, currículos adaptativos que reajustem a dificuldade com base no desempenho e a exposição deliberada a configurações desafiadoras podem mitigar alguns dos problemas observados, especialmente aqueles em que o agente interrompe o movimento ou converge para poses insatisfatórias.

Além da forma como os dados são apresentados ao agente, avanços arquiteturais e algorítmicos podem contribuir de maneira substancial para a robustez do aprendizado. Abordagens como políticas hierárquicas, explicitação da dinâmica ou modelos com vieses espaciais explícitos podem capturar melhor as dependências entre juntas em cadeias cinemáticas. Do mesmo modo, algoritmos projetados para problemas contínuos de alta dimensionalidade como Twin Delayed Deep Deterministic Policy Gradient (TD3)[10] podem apresentar maior estabilidade no aprendizado do que o PPO neste cenário.

Por fim, sugere-se estudos de ablação com respeito aos diversos componentes das funções de recompensa propostas na Seção IV-D, bem como ao anelamento do parâmetro ϵ explicado na Seção IV-E.

Referências

[1] J. Ho and S. Ermon, "Generative adversarial imitation learning," in Advances in Neural Information Processing Systems, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/cc7e2b878868cbae992d1fb743995d8f-Paper.pdf

[2] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "Deepmimic: example-guided deep reinforcement learning of physics-based character skills," ACM Trans. Graph., vol. 37, no. 4, Jul. 2018. [Online]. Available: https://doi.org/10.1145/3197517.3201311

[3] A. Harutyunyan, S. Devlin, T. Vrancx, and A. Nowé, "Expressing arbitrary reward functions as potential-based advice," in Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015, pp. 2652–2658.

[4] H. Zou, T. Ren, D. Yan, H. Su, and J. Zhu, "Reward shaping via meta-learning," 2019.

[5] E. Wiewiora, G. Cottrell, and C. Elkan, "Principled methods for advising reinforcement learning agents," in Proceedings of the Twentieth International Conference on International Conference on Machine Learning, ser. ICML'03. AAAI Press, 2003, p. 792–799.

[6] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in Proceedings of the 26th Annual International Conference on Machine Learning, ser. ICML '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 41–48. [Online]. Available: https://doi.org/10.1145/1553374.1553380

[7] O. M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba, "Learning dexterous in-hand manipulation," The International Journal of Robotics Research, vol. 39, no. 1, pp. 3–20, 2020. [Online]. Available: https://doi.org/10.1177/0278364919887447

[8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[10] J. Wu, Q. M. J. Wu, S. Chen, F. Pourpanah, and D. Huang, "A-td3: An adaptive asynchronous twin delayed deep deterministic for continuous action spaces," IEEE Access, vol. PP, pp. 1–1, 01 2022.

O código implementado, as animações salvas, os gráficos produzidos e os checkpoints dos modelos treinados estão disponíveis em https://github.com/masganem/msi2.

DCC
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
UFMG