RL Multi-Objetivo Aplicado ao Problema Integrado de Patrulha e Despacho

Moacir Almeida Simões Júnior & Tobias de Abreu Kuse

Parte 1: O Problema e o Panorama das Soluções

O Problema Central: Um Conflito de Objetivos

- Operações policiais têm dois objetivos conflitantes e simultâneos.
- **1. Patrulhamento Proativo:** Maximizar a *presença* em áreas de alto risco ("hotspots") para dissuadir o crime.
- **2. Despacho Reativo:** Minimizar o *tempo de resposta* a novas chamadas de emergência dinâmicas.
- **O Desafio:** Cada vez que uma unidade é despachada *reativamente*, ela compromete o plano de patrulha *proativo*.
- Este é um problema **bi-objetivo**.

Um Panorama das Soluções

- Problema complexo: mistura previsão espaço-temporal, roteamento e eventos estocásticos em tempo real.
- Três filosofias de solução principais:
  1. Modelos **Híbridos Heurísticos** (Simulação + Meta-heurísticas)
  2. Modelos **MARL Descentralizados** (RL Multiagente)
  3. Modelos **MORL Centralizados** (RL Multi-Objetivo)

Abordagem 1: A Híbrida-Heurística

(Simões Júnior & Borenstein, 2025)

- **Conceito:** Modelar o sistema com simulação de eventos discretos.
- **Método:**
  - **Prever:** Usar ML (ex: XGBoost) para prever *hotspots dinâmicos*.
  - **Patrulhar:** Usar **ACO** para rotas de patrulha ideais.
  - **Testar:** Simular interrupções por chamadas aleatórias.
- **Conclusão:** Abordagem de *"otimizar e depois simular"*.

Abordagem 2: O Modelo Multiagente

(Repasky et al., 2024)

- **Conceito:** Cada viatura = um agente de RL.
- **Método:**
  - Sistema **heterogêneo** com $N + 1$ agentes.
  - N **Patrulheiros:** aprendem políticas próprias (DQN compartilhada).
  - **1 Despachante:** aprende política central (MIP + VFA).
- **Conclusão:** Abordagem **descentralizada**, com agentes aprendendo individualmente.

Abordagem 3: RL Multi-Objetivo (MORL)

- RL padrão → recompensa escalar R.
- MORL → vetor de recompensas $\vec{R} = \langle R_{\text{resposta}}, R_{\text{presença}} \rangle$.
- **Desafio:** "Ótimo" é subjetivo (quanto vale piorar patrulha pra melhorar resposta?).
- **Solução comum:** Soma ponderada

$$R_{total} = w_1 R_{resposta} + w_2 R_{presença}$$

- **Problema:** Altamente sensível aos pesos — arbitrários e difíceis de justificar.

Transição: Uma Abordagem MORL Superior

- As falhas da soma ponderada motivam novos métodos.
- Abordagens “conjuntas” (Joe et al., Repasky et al.) resolvem ambos problemas.
- Vamos analisar o artigo de **Joe, Lau & Pan (2022)**:
  - Modelo MORL centralizado que **evita soma ponderada**.

Parte 2: Estudo de Caso — Joe, Lau & Pan (2022)

Visão Geral e Formulação

- **Título:** *Reinforcement Learning Approach to Solve Dynamic Bi-objective Police Patrol Dispatching and Rescheduling Problem*
- **Formulação:** MDP Centralizado de Agente Único.
- **Agente de RL:** Planejador central / despachante.
- **Atores:** Unidades de patrulha (recursos controlados).
- Não é MARL — unidades não aprendem.

O MDP: Estado e Ação

- **Estado (S_k):** snapshot do sistema no incidente ω_k.
  - Hora atual (t_k), cronogramas ($\delta(k)$), status de patrulha ($\sigma(k)$).
- **Ação (x_k):** tupla de decisão $\langle x_k^i, x_k^t, \delta^x(k) \rangle$.
  - x_k^i: quem despachar
  - x_k^t: quando
  - $\delta^x(k)$: novo cronograma conjunto
- **Insight:** espaço de ação enorme — decisão global sobre todas as unidades.

A Solução: Híbrido de VFA + Heurística

- **Divisão de trabalho:**
  - **Offline (Cérebro):**
    - Rede de Função Valor $\hat{V}$ treinada offline.
    - Aprende o valor futuro esperado de cada estado.
  - **Online (Gerador de Ação):**
    - Heurística de Reagendamento (Ejection Chains) gera ações viáveis.
    - $\hat{V}$ avalia cada ação e escolhe:

$$x_k^* = \text{argmax}_{x_k}\{R_{imediata} + \gamma \hat{V}(S_k^x)\}$$

Inovação Bi-Objetivo (Sem Soma Ponderada)

- Recompensa **multiplicativa e diferencial**:

$$R(S_k, x_k) = f_r(x_k) \times f_p(\delta^x(k)) - f_p(\delta(k))$$

- **Componentes:**
  - $f_r(x_k)$ → sucesso da resposta (1.0, 0.5, 0)
  - $f_p(\delta^x(k)) - f_p(\delta(k))$ → mudança na presença da patrulha
- **Inteligência:**
  - A recompensa pondera *resposta* × *impacto na patrulha*.
  - Uma falha que destrói o plano de patrulha → fortemente penalizada.

Resultados Chave

- **Joint Learning > Two-Stage**
  - O híbrido ($\hat{V}$ + heurística) supera o método "Two-Stage".
- **Computacionalmente Viável:**
  - Despacho + reagendamento em tempo quase real (< 10s).

Parte 3: Conclusão

Resumo das Abordagens

- **Problema:** Patrulha Conjunta (Proativa) & Despacho (Reativo).
- **Híbrida/Heurística:** poderosa, mas depende de heurísticas (ACO).
- **MARL:** baseado em agentes, mas decompõe o problema.
- **MORL Centralizado:** abordagem holística (VFA + heurística inteligente).

Principais Conclusões

- Problema real e ideal para RL avançado.
- Arquitetura de Joe, Lau & Pan:
  - **Gerador-Heurístico + Avaliador-VFA** → ótimo para espaços de ação complexos.
- Recompensa multiplicativa → resolve bi-objetivo sem pesos arbitrários.

Obrigado! 🙌

Perguntas?