Fast Multi-agent Temporal-difference Learning: Homotopy Stochastic Primal-dual Method

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, Mihailo R. Jovanović

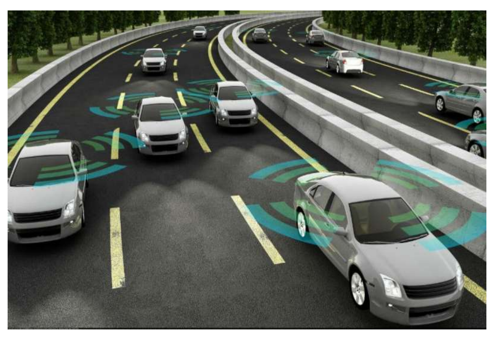
dongshed@usc.edu, xiaohanw@usc.edu, zy6@princeton.edu, zhaoran.wang@northwestern.edu, mihailo@usc.edu

Background and motivation

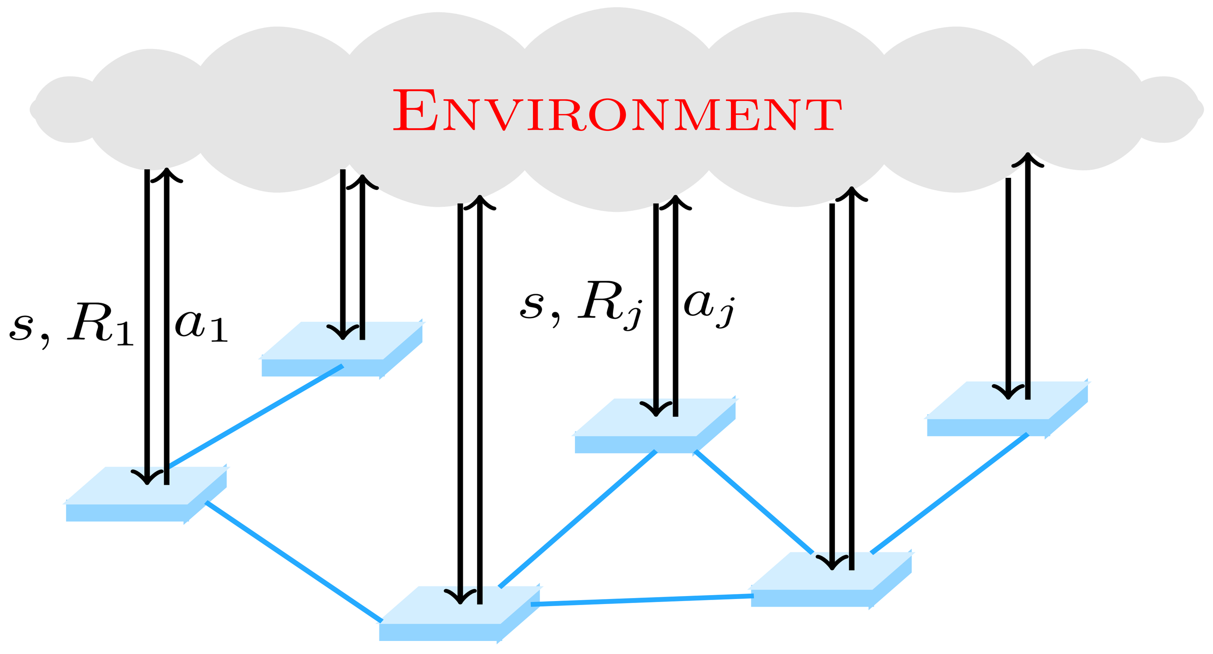
Multi-agent reinforcement learning

Power

Robotics



Transportation



$$R_c^\pi(s) \;=\; \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{a\sim\pi(\cdot\,|s)}[\,R_j(s,a)\,]$$

- Objective of policy evaluation

$$V^\pi(s) \;=\; \mathbb{E}[\,R_c^\pi(s_0)+\gamma R_c^\pi(s_1)+\gamma^2 R_c^\pi(s_2)+\cdots\,|\,s_0=s,\pi\,]$$

- Challenges
  - Distributed samples over a network
  - Markovian samples for a given policy

Multi-agent TD Learning

Bellman error minimization

- Bellman equation with linear function approximation

$$\mathbf{V}_x = \gamma\mathbf{P}^\pi\mathbf{V}_x + \mathbf{R}_c^\pi$$

  - $V_x(s) = \phi^T(s)x$ – linear approximation of $V^\pi(s)$
  - $\mathbf{V}_x, \mathbf{R}_c^\pi$ – vectors of $V_x(s), R_c^\pi(s)$ for all states s
  - $\mathbf{P}^\pi$ – probability transition marix

- Projected Bellman error minimization

Centralized problem	Decentralized problem
$\underset{x\in\mathcal{X}}{\text{minimize}}\;\frac{1}{2}\,\Vert Ax-b\Vert_{C^{-1}}^2$	$\underset{x\in\mathcal{X}}{\text{minimize}}\;\frac{1}{2N}\sum_{j=1}^N\Vert Ax-b_j\Vert_{C^{-1}}^2$
$b=\mathbb{E}_{s\sim\Pi}[\,\mathcal{R}_c^\pi(s)\phi(s)\,]$	$b_j=\mathbb{E}_{s\sim\Pi}[\,\mathcal{R}_j^\pi(s)\phi(s)\,]$

  - $A=\mathbb{E}_{s\sim\Pi}[\,\phi(s)(\phi(s)-\gamma\phi(s'))^T\,]$ and $C=\mathbb{E}_{s\sim\Pi}[\,\phi(s)\phi(s)^T\,]$
  - Π – unknown stationary distribution for a given policy

Decentralized stochastic saddle-point problem

- Dualization of the objective function

$$\Vert Ax-b_j\Vert_{C^{-1}}^2 \;=\; \max_{y_j\in\mathcal{Y}}\;\underbrace{y_j^T(Ax-b_j)-\tfrac{1}{2}\,y_j^TCy_j}_{\psi_j(x,y_j)\;=\;\mathbb{E}_{\xi\sim\Pi}[\,\Psi_j(x,y_j;\xi)\,]}$$

- Stochastic saddle-point problem

$$\min_{x\in\mathcal{X}}\;\max_{y_j\in\mathcal{Y}}\;\psi(x,y) := \frac{1}{N}\sum_{j=1}^N\mathbb{E}_{\xi\sim\Pi}[\,\Psi_j(x,y_j;\xi)\,]$$

- Dependent samples, unknown distribution
  - $\xi_t\sim P_t$ – samples from a Markov process P_t at time t
  - P_t – Markov process converging to the unknown Π

Homotopy primal-dual Algorithm

- Distributed dual averaging for aggregating local information
- Homotopy method for adaptive stepsize selection

Algorithm 1 Distributed Homotopy Primal-Dual (DHPD) Algorithm

Initialization: $x_{j,1}(1)=x'_{j,1}(1)=0,\,y_{j,1}(1)=y'_{j,1}(1)=0,\eta_1,T_1,K$

For $k=1$ to K **do** ▷ for all agents $j\in\mathcal{V}$

(1) **For** $t=1$ to T_k-1 **do**

– Primal update ▷ Distributed dual averaging

$$x'_{j,k}(t+1) \;=\; \textstyle\sum_{i=1}^N W_{ij}\,x'_{i,k}(t)\;-\;\eta_k\,\nabla_x\Psi_j(z_{j,k}(t);\xi_k(t))$$

$$x_{j,k}(t+1) \;=\; \mathcal{P}_\mathcal{X}(x'_{j,k}(t+1))$$

– Dual update ▷ Local update

$$y'_{j,k}(t+1) \;=\; y'_{j,k}(t)\;+\;\eta_k\,\nabla_y\Psi_j(z_{j,k}(t);\xi_k(t))$$

$$y_{j,k}(t+1) \;=\; \mathcal{P}_\mathcal{Y}(y'_{j,k}(t+1))$$

end for

(2) $(x_{j,k+1}(1),\,y_{j,k+1}(1)) = \left(\frac{1}{T_k}\sum_{t=1}^{T_k}x_{j,k}(t),\;\frac{1}{T_k}\sum_{t=1}^{T_k}y_{j,k}(t)\right)$

(3) $(x'_{j,k+1}(1),\,y'_{j,k+1}(1)) = (x_{j,k+1}(1),\,y_{j,k+1}(1))$

(4) $\eta_{k+1}=\eta_k/2,\;T_{k+1}=2T_k$ ▷ Adaptive stepsize

end for

Output: $(\hat{x}_{j,K},\hat{y}_{j,K}) = \left(\frac{1}{T_K}\sum_{t=1}^{T_K}x_{j,K}(t),\;\frac{1}{T_K}\sum_{t=1}^{T_K}y_{j,K}(t)\right)$

- Optimality gap induced by $\hat{x}_{i,k}$

$$\epsilon(\hat{x}_{i,k}) \;=\; \frac{1}{2N}\sum_{j=1}^N\left(\Vert A\hat{x}_{i,k}-b_j\Vert_{C^{-1}}^2-\Vert Ax^\star-b_j\Vert_{C^{-1}}^2\right)$$

Fast Convergence rate

- Assumptions
  - $\sup_{x\in\mathcal{X},y\in\mathcal{Y}}\Vert(x,y)\Vert^2\le R^2$ – convex compact domain
  - $\psi_j(x,y_j)$ – ρ_y-strongly concave; G-gradient bounded; L-gradient Lipschitz
  - $\max_{y_j\in\mathcal{Y}}\psi_j(x,y_j)$ – ρ_x-strongly convex
  - W – doubly stochastic communication matrix
- Claim: For any $\eta_1\ \ge 1/(4/\rho_y+2/\rho_x)$, any T_1,K satisfying $T_1\ge 1+\lceil\log(\Gamma T)/|\log\rho|\rceil := \tau$ where $T=\sum_{k=1}^K T_k$, we have

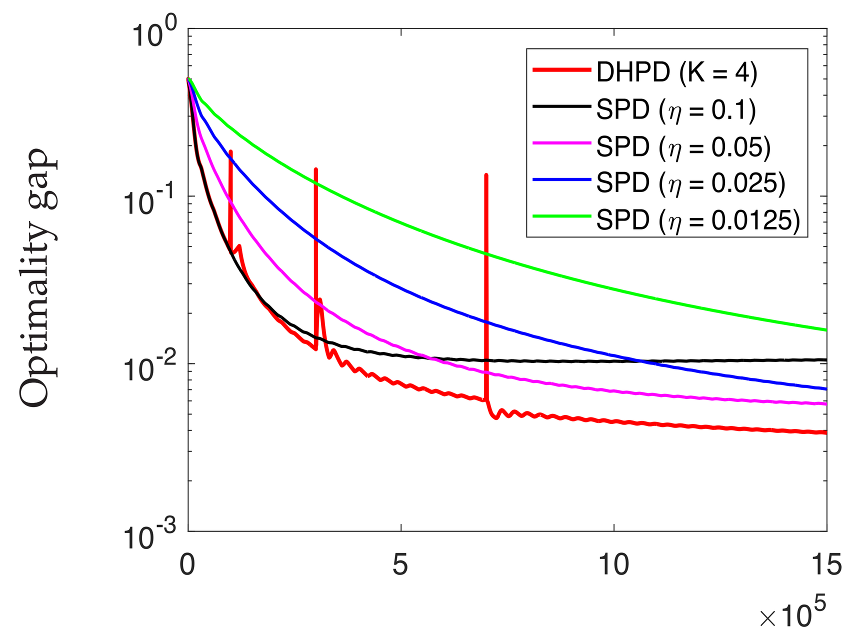
$$\frac{1}{N}\sum_{j=1}^N\mathbb{E}[\varepsilon(\hat{x}_{j,K})] \;=\; C_1\frac{G(RL+G)\log^2(\sqrt{N}T)}{T(1-\sigma_2(W))}+C_2\frac{G(G+RL)(1+T_1)}{T}$$

  - Fast convergence rate $O(\log(\sqrt{N}T)/T)$
  - Network dependence $\log^2(\sqrt{N}T)/(1-\sigma_2(W))$
  - Fast $1/T$-mixing fast convergence

Case study of Mountain Car Task

- SPD – stochastic primal-dual method

$N=10$



$N=100$