Reinforcement Learning: An Overview

Shalabh Bhatnagar

Department of Computer Science and Automation
Indian Institute of Science
Bangalore 560 012
shalabh@csa.iisc.ernet.in

January 28, 2011

Outline

An Introduction to Reinforcement Learning

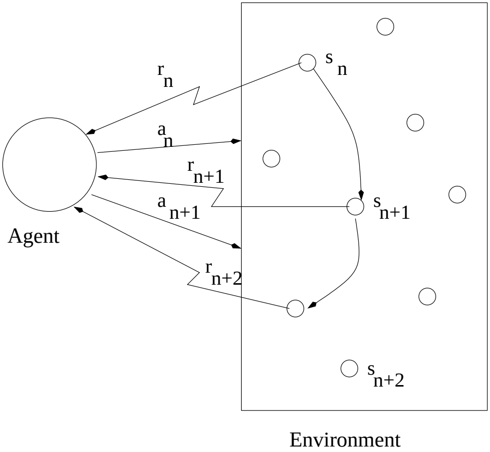


Figure: Agent-Environment Interaction

Markov Decision Processes

- Puterman [1994], Bertsekas [2005,2007]
- A Markov Decision Process (MDP) is a controlled random process $\{s_t\}$ that depends on a control-valued sequence $\{a_t\}$ and satisfies the controlled Markov property (below)
- Let S denote the state space and A the action space. Assume S and A are finite sets
- In general, when state is $i \in S$, feasible action space is $A(i)$. Here $A = \cup_{i \in S} A(i)$
- Let $k(s_t, a_t, s_{t+1})$ be the cost incurred when state at time t is s_t, action chosen is a_t and the next state is s_{t+1}.

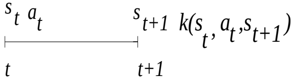


Figure: State, Action and Single-Stage Cost

The Controlled Markov Property

- For all $i_0, i_1, \ldots, s, s', b_0, b_1 \ldots, a$ in appropriate sets,

$$P(s_{t+1} = s' \mid s_t = s, a_t = a, \ldots, s_0 = i_0, a_0 = b_0)$$

$$= P(s_{t+1} = s' \mid s_t = s, a_t = a) = P^a_{ss'}$$

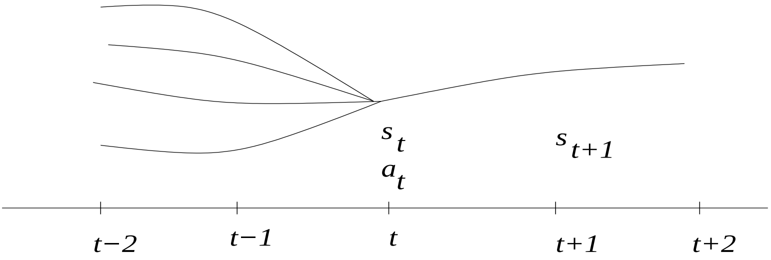


Figure: The Controlled Markov Behaviour

The Finite Horizon Problem

- Here horizon length $= N < \infty$
- By an admissible policy π, we mean a sequence of functions $\pi = \{\mu_0, \mu_1, \ldots, \mu_{N-1}\}$ such that each $\mu_n : S \to A$ with $\mu_n(j) \in A(j), j \in S$. At instant n, actions under π are selected according to μ_n.
- Let Π be set of all admissible policies
- Objective: Find a $\pi^* \in \Pi$ that minimizes

$$J_\pi(i) = E\left[\sum_{j=0}^{N-1} k(X_j, \mu_j(X_j), X_{j+1}) + h(X_N) \mid X_0 = i\right],$$

  where $h(l)$ is the cost incurred when the 'terminal state' is $l \in S$.
- Let $J^*(i) = \min_{\pi \in \Pi} J_\pi(i) = J_{\pi^*}(i)$

The Principle of Optimality

- Let $\pi^* = \{\mu_0^*, \mu_1^*, \ldots, \mu_{N-1}^*\}$ be an optimal policy. Suppose that when using π^*, a state x_i occurs at time i with positive probability. Consider the subproblem – minimize from time i to N,

$$E\left[\sum_{j=i}^{N-1} k(X_j, \mu_j(X_j), X_{j+1}) + h(X_N) \mid X_i = x_i\right].$$

  Then the truncated policy $\{\mu_i^*, \mu_{i+1}^*, \ldots, \mu_{N-1}^*\}$ is optimal for this subproblem.
- Thus optimal policy can be constructed by going backwards in time i.e., construct optimal policy for tail subproblem involving last stage, then extending optimal policy to tail subproblem involving last two stages and continuing till optimal policy for full problem is constructed

The Dynamic Programming Algorithm

- For every initial state i_0, the optimal cost $J^*(i_0)$ of the basic problem equals $J_0(i_0)$, given by the last step of the following algorithm, that proceeds backwards in time from period $N-1$ to period 0:

$$J_N(i_N) = h(i_N),$$

$$J_l(i_l) = \min_{u_l \in A(i_l)} E\left[k(X_l, u_l, X_{l+1}) + J_{l+1}(X_{l+1}) \mid X_l = i_l\right],$$

$$= \min_{u_l \in A(i_l)} \sum_{j \in S} p_{i_l j}^{u_l} \left(k(i_l, u_l, j) + J_{l+1}(j)\right),$$

$\forall l = 0, 1, \ldots, N-1$, $\forall i_0, \ldots, i_N \in S$

- If $u_l^* = \mu_l^*(i_l)$ minimizes RHS above for each i_l and l, then the policy $\pi^* = \{\mu_0^*, \ldots, \mu_{N-1}^*\}$ is optimal

Example – Control of a Queue

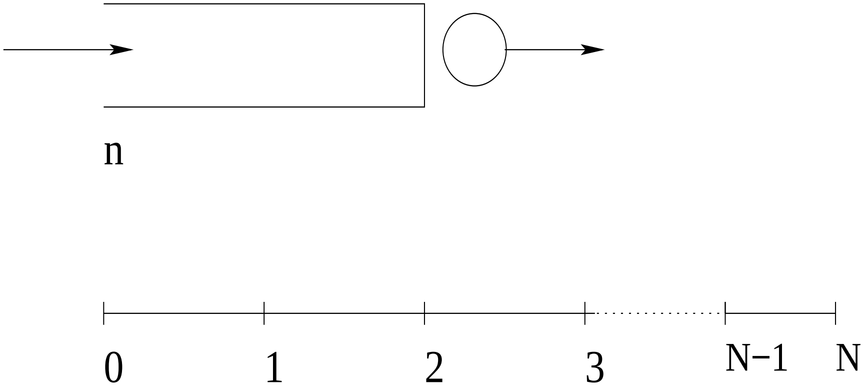


Figure: A Discrete Time Queue

Example – The Setting

- Assume arrivals/departures at discrete instants. Only one customer can be served in a period. A customer can take multiple periods for service
- Two types of service – fast (u_f with cost c_f per period) and slow (u_s with cost c_s per period)
- Let $p_m =$ probability of m arrivals in a period ($m \geq 0$)
- With fast (slow) service, customer in service at beginning of period will finish service w.p. q_f (q_s) independent of number of periods a customer has been in service for and the number of customers in system. Assume $q_f > q_s$
- Assume cost $r(i)$ is incurred in each period for which i customers are in system. Also, let $R(i)$ be terminal cost if i customers are left at time N in system

Example – Transition Probabilities

- $p_{0j}^{u_f} = p_{0j}^{u_s} = p_j \; (j = 0, 1, \ldots, n-1)$
- $p_{0n}^{u_f} = p_{0n}^{u_s} = \sum_{m=n}^{\infty} p_m \; (j = n)$
- $p_{ij}^{u_f} = p_{ij}^{u_s} = 0 \; (j < i-1, i > 0)$
- $p_{ij}^{u_f} = q_f p_0 \; (j = i-1, i = 0)$
- $p_{ij}^{u_s} = q_s p_0 \; (j = i-1, i = 0)$
- $p_{ij}^{u_f} = q_f p_{j-i+1} + (1-q_f) p_{j-i} \; (i-1 < j < n-1)$
- $p_{ij}^{u_s} = q_s p_{j-i+1} + (1-q_s) p_{j-i} \; (i-1 < j < n-1)$
- $p_{i(n-1)}^{u_f} = q_f \sum_{m=n-i}^{\infty} p_m + (1-q_f) p_{n-1-i}$
- $p_{i(n-1)}^{u_s} = q_s \sum_{m=n-i}^{\infty} p_m + (1-q_s) p_{n-1-i}$
- $p_{in}^{u_f} = (1-q_f) \sum_{m=n-i}^{\infty} p_m$
- $p_{in}^{u_s} = (1-q_s) \sum_{m=n-i}^{\infty} p_m$

Example – DP Algorithm

- Single stage cost $= r(i) + c_f$ if fast service is used; else $r(i) + c_s$ if slow service is used
-
$$J_N(i) = R(i),$$
$$J_k(i) = \min(c_f + r(i) + \sum_{j \in S} p_{ij}^{u_f} J_{k+1}(j), c_s + r(i) + \sum_{j \in S} p_{ij}^{u_s} J_{k+1}(j)),$$
$\forall k = 1, \ldots, N-1, i \in S$
- For $i = 0$, no service is required. Thus $A(0) = \phi$, while $A(i) = \{u_f, u_s\}$, for all $i > 0$. Thus, $J_N(0) = R(0)$ while $J_k(0) = r(0) + \sum_{j \in S} p_{0j}^{u_f} J_{k+1}(j)$ for $k = 1, \ldots, N-1$. Note here that $p_{0j}^{u_f} = p_{0j}^{u_s}$ (shown before) and in fact equals p_{0j} i.e., no action

The Infinite Horizon Discounted Cost Problem

- Here $N = \infty$
- An admissible policy π is a sequence of functions $\pi = \{\mu_0, \mu_1, \ldots,\}$ such that each $\mu_n : S \rightarrow A$ and $\mu_n(j) \in A(j)$, $\forall j \in S$. At instant n, actions under π are selected according to μ_n.
- Let Π be set of all admissible policies
- Objective: Find a $\pi^* \in \Pi$ that minimizes the cost-to-go or the value function

$$V_\pi(i) = E\left[\sum_{j=0}^{\infty} \gamma^k k(X_j, \mu_j(X_j), X_{j+1}) \mid X_0 = i\right]$$

- Let $V^*(i) = \min_{\pi \in \Pi} V_\pi(i) = V_{\pi^*}(i)$

Stationary Policies

- A stationary deterministic policy (SDP) π is one for which $\mu_i \equiv \mu$ for all $i = 0, 1, 2, \ldots$. Many times we just call μ an SDP.
- A stationary randomized policy ϕ can be characterized by distributions $\phi(i) = (\phi(i, a), a \in A(i))$, $i \in S$.
- It can be shown that the optimal policy (i.e., the one that attains the minimum) is an SDP and so also an SRP
- Let $T, T_\mu : \mathcal{R}^{|S|} \to \mathcal{R}^{|S|}$ be the maps

$$TJ(i) = \min_{a \in A(i)} \sum_{j \in S} P_{ij}^{a}(k(i, a, j) + \gamma J(j)),$$

$$T_\mu J(i) = \sum_{j \in S} P_{ij}^{\mu(i)}(k(i, \mu(i), j) + \gamma J(j)),$$

$i \in S$.

Finite Horizon Operators

- Let $T^k J(i) = T(T^{k-1}J(i))$, $T^k_\mu J(i) = T_\mu(T^{k-1}_\mu J(i))$, $i \in S$, $k \geq 0$. Here $T^0 J = T^0_\mu J = J$.
- Note that

$$T^2 J(i) = \min_{a \in A(i)} \sum_j P^a_{ij}(k(i,a,j) + \gamma TJ(j))$$

$$= \min_{a \in A(i)} (\sum_j P^a_{ij}(k(i,a,j) + \gamma \min_{u \in A(j)} \sum_l P^u_{jl}(k(j,u,l) + \gamma J(l)))$$

$$= \min_{a \in A(i)} (\sum_j P^a_{ij}(k(i,a,j) + \min_{u \in A(j)} \sum_l P^u_{jl}(\gamma k(j,u,l) + \gamma^2 J(l)))$$

- The above corresponds to DP algorithm for a two-stage γ-discounted problem with initial state i, cost per stage k and terminal cost function $\gamma^2 J$.

Monotonicity of Operators T^k

- Proposition 1: For any functions $J, J' : S \rightarrow \mathcal{R}$, $J(i) \leq J'(i), \forall i \in S$ implies $T^k J(i) \leq T^k J'(i)$ and $T_\mu^k J(i) \leq T_\mu^k J(i)$ for all $i \in S$, $k = 1, 2, \ldots$.
- Proof: Since $T^k J$ can be viewed as a k-stage problem cost with terminal cost function $\gamma^k J$, $J \leq J'$ implies $T^k J \leq T^k J'$. $\square$
- Proposition 2: $\forall k \geq 0, i \in S$,

$$T^k(J + re)(i) = T^k J(i) + \gamma^k r,$$

$$T_\mu^k(J + re)(i) = T_\mu^k J(i) + \gamma^k r,$$

where $e = (1, \ldots, 1)^T$ is a $|S|$-dimensional unit vector.

Convergence of DP

- We assume that $|k(i, a, j)| \leq M < \infty$, for all $i, j \in S$, $a \in A(i)$.
- Proposition 3(a): For any bounded function $J : S \to \mathcal{R}$,

$$V^*(i) = \lim_{N \to \infty} T^N J(i), \quad \forall i \in S.$$

- Proposition 3(b): For any SDP μ and bounded J,

$$V_\mu(i) = \lim_{N \to \infty} T_\mu^N J(i), \quad \forall i \in S,$$

where $V_\mu = V_\pi$ with $\pi = \{\mu, \mu, \ldots\}$.

The Bellman Equation

- Proposition 4 – The Bellman equation: The optimal cost function V^* satisfies

$$V^*(i) = \min_{a \in A(i)} \sum_j P_{ij}^a (k(i, a, j) + \gamma V^*(j)), \quad i \in S, \text{ or}$$

$$V^* = TV^*$$

Further, V^* is the unique solution of this equation within the class of bounded functions.

- Proposition 5 – The Poisson Equation: For every stationary policy μ, the associated cost function V_μ satisfies

$$V_\mu(i) = \sum_j P_{ij}^{\mu(i)} (k(i, \mu(i), j) + \gamma V_\mu(j)), \quad i \in S, \text{ or}$$

$$V_\mu = T_\mu V_\mu$$

Further, V_μ is the unique solution of this equation within the class of bounded functions.

The Optimality Condition

- Proposition 6 – Necessary and Sufficient Condition for Optimality: A stationary policy μ is optimal if and only if $\mu(i)$ attains the minimum in the Bellman equation for each $i \in S$, i.e., $TV^* = T_\mu V^*$
- Proof: Suppose $TV^* = T_\mu V^*$. Then by the Bellman equation,

$$V^* = TV^* = T_\mu V^*.$$

  Now since the operator T_μ has a unique fixed point V_μ (Result 4), we have $V^* = V_\mu$ i.e., μ is optimal
  Suppose now that μ is optimal. Then $V^* = V_\mu$. Hence $V^* = T_\mu V^*$ (Proposition 5) $= TV^*$ (Proposition 4). $\square$

Numerical Approaches

- Value Iteration:
- Recall that (Propositions 3(a)-3(b)) $V^*(i) = \lim_{N\to\infty} T^N J_1(i)$ and $V_\mu(i) = \lim_{N\to\infty} T_\mu^N J_2(i)$ for any bounded $J_1, J_2 : S \to \mathcal{R}$.
- Start with initial estimate $V_0 = J_1$ and iterate

$$V_{n+1} = TV_n \text{ i.e.,}$$

$$V_{n+1}(i) = \min_a \sum_j P_{ij}^a (k(i,a,j) + \gamma V_n(j))$$

  Then $V_n \to V^*$.
- Similarly W_n, $n \geq 0$ with $W_0 = J_2$ and $W_{n+1} = T_\mu W_n$ satisfies $W_n \to V_\mu$.

- The following is a key result on which policy iteration is based.
- Proposition 7: Let μ and $\bar{\mu}$ be SDPs such that $T_{\bar{\mu}}V_{\mu} = TV_{\mu}$, i.e.,

$$\sum_j P_{ij}^{\bar{\mu}(i)}(k(i,\bar{\mu}(i),j) + \gamma V_{\mu}(j))$$

$$= \min_{a \in A(i)} \left(\sum_j P_{ij}^{a}(k(i,a,j) + \gamma V_{\mu}(j) \right).$$

  Then $V_{\bar{\mu}}(i) \leq V_{\mu}(i)$, $\forall i \in S$. Further, if μ is not optimal, strict inequality holds for at least one state i.

The Policy Iteration Algorithm

- (Initialize:) Start with a given stationary policy μ_0.
- (Policy Evaluation:) Let $K_{\mu_n} = (\sum_j P_{ij}^{\mu_n(i)} k(i,\mu_n(i),j), i \in S)^T$, $P_{\mu_n} = [[P_{ij}^{\mu_n(i)}]]_{i,j\in S}$ and $V_{\mu_n} = (V_{\mu_n}(i), i \in S)^T$. Solve the linear system of equations $V_{\mu_n} = K_{\mu_n} + \gamma P_{\mu_n} V_{\mu_n}$.
- If $V_{\mu_n} = V_{\mu_{n-1}}$, terminate procedure, else go to next step.
- (Policy Improvement:) Find a stationary policy μ_{n+1} such that

$$\sum_j P_{ij}^{\mu_{n+1}(i)}(k(i,\mu_{n+1}(i),j) + \gamma V_{\mu_n}(j))$$

$$= \min_{a\in A(i)} \sum_j P_{ij}^{a}(k(i,a,j) + \gamma V_{\mu_n}(j))$$

- Set $n := n+1$ and go to the second step (PE) above.

Long-run Average Cost Problems

- $N = \infty$
- Objective: Find a $\pi^* \in \Pi$ that minimizes over all $\pi \in \Pi$, the average cost-per-stage starting from a given initial state $i \in S$ i.e.,

$$\lambda_\pi(i) = \limsup_{N \to \infty} \frac{1}{N} E\left[\sum_{j=0}^{N-1} k(X_j, \mu_j(X_j), X_{j+1}) \mid X_0 = i\right]$$

- Note that limit may not exist in general (hence we use limsup). Limit can be shown to exist under any stationary policy μ if the underlying Markov chain $\{X_n\}$ under that policy is ergodic.

The Bellman Optimality Equation

- Assume that $\{X_n\}$ is ergodic under all stationary policies
- Poisson Equation: For all $i \in S$ and given a stationary policy μ,

$$\lambda_\mu + h_\mu(i) = P_{ij}^{\mu(i)}(k(i,\mu(i),j) + h_\mu(j)),$$

where $h_\mu(i)$ is the differential cost under μ in state i defined as

$$h_\mu(i) = E_\mu\left[\sum_{l=0}^{\infty}(k(X_l,\mu(X_l),X_{l+1}) - \lambda_\mu) \mid X_0 = i\right]$$

- Bellman Equation: For all $i \in S$,

$$\lambda^* + h(i) = \min_{a \in A(i)} P_{ij}^{a}(k(i,a,j) + h(j)),$$

where λ^* is the optimal average cost and $h(i)$ is the differential cost in state i i.e., $h(i) = \min_\mu h_\mu(i)$

Relation between Average and Discounted Cost

- Let $\lambda_\mu(i)$ and $V_{\gamma,\mu}(i)$, $i \in S$ denote the average and γ-discounted costs from state i. Then

$$\lambda_\mu(i) = \limsup_{N\to\infty} \frac{1}{N} E\left[\sum_{l=0}^{N-1} k(X_l, \mu(X_l), X_{l+1}) \mid X_0 = i\right]$$

$$= \limsup_{N\to\infty} \lim_{\gamma\to 1} \frac{E[\sum_{l=0}^{N-1} \gamma^l k(X_l, \mu(X_l), X_{l+1}) \mid X_0 = i]}{\sum_{l=0}^{N-1} \gamma^l}$$

- Assuming an interchange of limits (see Bertsekas (2007) for a rigorous argument),

$$\lambda_\mu(i) = \lim_{\gamma\to 1} \limsup_{N\to\infty} \frac{E[\sum_{l=0}^{N-1} \gamma^l k(X_l, \mu(X_l), X_{l+1}) \mid X_0 = i]}{\sum_{l=0}^{N-1} \gamma^l}$$

$$= \lim_{\gamma\to 1} (1-\gamma) V_{\gamma,\mu}(i)$$

Value Iteration

- VI–version 1:
- Define operator $T : \mathcal{R}^{|S|} \rightarrow \mathcal{R}^{|S|}$ by $Th = \min_\mu (K_\mu + P_\mu h)$. Here $K_\mu = (\sum_{j \in S} P_{ij}^{\mu(i)} k(i, \mu(i), j), i \in S)^T$. Then one can show that $T^r h / r \rightarrow \lambda^*$ as $r \rightarrow \infty$.
- VI–version 2 or relative value iteration:
- Fix a state $i_0 \in S$ arbitrarily. Select a function $h_0 : S \rightarrow \mathcal{R}$. Iterate over $n \geq 0$,

$$h_{n+1}(i) = \min_{a \in A(i)} \sum_j P_{ij}^a (k(i, a, j) + h_n(j)) - h_n(i_0).$$

Then it can be shown that $h_n(i_0) \rightarrow \lambda^*$ as $n \rightarrow \infty$.

Policy Iteration

- Let μ_0 be an estimate of the optimal policy. Fix a state $i_0 \in S$ arbitrarily.
- Policy Evaluation: In the nth stage, $n \geq 0$, solve $\forall i \in S$,

$$h^{\mu_n}(i) = \sum_j P_{ij}^{\mu_n(i)}(k(i, \mu_n(i), j) + h^{\mu_n}(j)) - h^{\mu_n}(i_0)$$

  If $h^{\mu_n} = h^{\mu_{n-1}}$, terminate, else go to the next step.
- Policy Improvement: For all $i \in S$,

$$\mu_{n+1}(i) = \arg\min_{a \in A(i)} \left(\sum_j P_{ij}^a (k(i, a, j) + h^{\mu_n}(j)) \right).$$

- It can be shown that $\mu_n \to \mu^*$ for a stationary policy μ^* such that $h^{\mu^*}(i_0) = \lambda^*$.

Limitations of Numerical Methods for Exact Schemes

- For solving Bellman optimality equations (in various cases) using numerical methods, one requires complete knowledge of transition probabilities P_{ij}^{a}, $i, j \in S$, $a \in A(i)$. (*lack of model information*)
- The amount of computation required to solve Bellman equation grows exponentially in the cardinality of the state and action spaces. (*curse of dimensionality*)
- Hence, one resorts to approaches that use a combination of “simulation” and “feature-based approximations”

- Bertsekas [2010]
- Consider the discounted cost case. Let

$$V_\mu(i) \approx \tilde{V}_\theta(i) = \theta^T \phi_i,$$

where $\phi_i = (\phi_i(1), \ldots, \phi_i(d))^T$ is a state-feature associated with state i and $\theta = (\theta_1, \ldots, \theta_d)^T$ is the associated parameter

- Let $\Phi = [[\phi_i^T]]_{i \in S}$ be the $(|S| \times d)$-feature matrix. Let $\tilde{V}_\theta = (\tilde{V}_\theta(i), i \in S)^T$. Then $\tilde{V}_\theta = \Phi\theta = \sum_{j=1}^{d} \phi(j)\theta_j$, where $\phi(j) = (\phi_i(j), i \in S)^T$ (the jth column of the Φ matrix).
- The aim is to find the best approximation of V_μ within the space $S_0 = \{\Phi\theta \mid \theta \in \mathcal{R}^d\}$, i.e., the subspace spanned by columns of Φ.

Assumptions

- **Assumption (A1):** The Markov chain $\{X_n\}$ under the given stationary policy is aperiodic and irreducible
- **Assumption (A2):** The basis functions $\{\phi(k), k = 1, \ldots, d\}$ are linearly independent. Further, $d \leq |S|$ and Φ has full rank.
- Let $d^\mu = (d^\mu(1), \ldots, d^\mu(|S|))^T$ denote the stationary distribution of $\{X_n\}$ under the stationary policy μ. Let D^μ be a diagonal matrix with diagonal entries $d^\mu(i)$, $i \in S$.
- For $x \in \mathcal{R}^{|S|}$, define $\| x \|_D$ according to $\| x \|_D = (x^T D^\mu x)^{1/2}$.

The Projection Operator

- Let Π be the projection operator from $\mathcal{R}^{|S|}$ to S_0 w.r.t. $\| \cdot \|_D$. Thus given $V_\mu \in \mathcal{R}^{|S|}$, $\Pi V_\mu = \arg\min_{\hat{V} \in S_0} \| V_\mu - \hat{V} \|_D^2$. Since Φ has rank d, $\hat{V} = \Phi\theta$ for a unique $\theta \in \mathcal{R}^d$.
- Thus $\| V_\mu - \hat{V} \|_D^2 = \| V_\mu - \Phi\theta \|_D^2 = (V_\mu - \Phi\theta)^T D^\mu (V_\mu - \Phi\theta)$. Thus, $\Pi V_\mu = \Phi\theta_V$ where $\theta_V = \arg\min_{\theta \in \mathcal{R}^d} \| V_\mu - \Phi\theta \|_D^2$.
- Setting $\nabla_\theta(\| V_\mu - \Phi\theta \|_D^2) = 0$, one gets $\theta_V = (\Phi^T D^\mu \Phi)^{-1} \Phi^T D^\mu V_\mu$. Thus

$$\Pi = \Phi(\Phi^T D^\mu \Phi)^{-1} \Phi^T D^\mu V_\mu.$$

The Projected Poisson Equation

- Let ΠT_μ be a composition of Π with T_μ. Then
- Projected Poisson Equation:

$$\Phi\theta = \Pi T_\mu(\Phi\theta).$$

- Proposition 8: The mappings T_μ and ΠT_μ are contractions of modulus γ with respect to $\| \cdot \|_D$ i.e.,

$$\| T_\mu V - T_\mu \bar{V} \|_D \leq \gamma \| V - \bar{V} \|_D,$$

$$\| \Pi T_\mu V - \Pi T_\mu \bar{V} \|_D \leq \gamma \| V - \bar{V} \|_D,$$

$\forall V, \bar{V} \in \mathcal{R}^{|S|}$.

- Proposition 9: Let $\Phi\theta^*$ be the fixed point of ΠT. Then

$$\| V_\mu - \Phi\theta^* \|_D \leq \frac{1}{\sqrt{1-\gamma^2}} \| V_\mu - \Pi V_\mu \|_D$$

Numerical Solution of Projected Poisson Equation

- Use value iteration: start from an initial estimate $\theta_0 \in \mathcal{R}^d$ and iterate
$$\Phi\theta_{k+1} = \Pi T_\mu(\Phi\theta_k), \quad k = 0, 1, \ldots$$
- From Proposition 9, ΠT_μ is a contraction. Hence $\Phi\theta_k \to \Phi\theta^*$ as $k \to \infty$, where $\Phi\theta^*$ is the unique fixed point of ΠT_μ.
- Note that one can write $\theta_{k+1} = \arg\min_{\theta \in \mathcal{R}^d} \parallel \Phi\theta - (K_\mu + \gamma P_\mu \Phi\theta_k) \parallel_D^2$.

  Thus set
$$\nabla_\theta(\Phi\theta - K_\mu - \gamma P_\mu \Phi\theta_k)^T D^\mu(\Phi\theta - K_\mu - \gamma P_\mu \Phi\theta_k)) = 0 \quad i.e.,$$
$$\Phi^T D^\mu(\Phi\theta_{k+1} - K_\mu - \gamma P_\mu \Phi\theta_k)^T = 0.$$
- Thus $\theta_{k+1} = \theta_k - (\Phi^T D^\mu \Phi)^{-1}(C\theta_k - d)$, where $C = \Phi^T D^\mu(I - \gamma P_\mu)\Phi$ and $d = \Phi^T D^\mu K_\mu$.

Positive Definiteness of $\Phi^T D^\mu (I - \gamma P_\mu)\Phi$

- Note that $\| x \|_D^2 = x^T D^\mu x = \| (D^\mu)^{1/2} x \|^2$. Now for any function $V \in \mathcal{R}^{|S|}$, we have

$$\| P_\mu V \|_D^2 = V^T {P_\mu}^T D^\mu P_\mu V = \sum_{i \in S} d^\mu(i) E_\mu^2 [V(X_{n+1}) \mid X_n = i]$$

$$\leq \sum_{i \in S} d^\mu(i) E_\mu [V^2(X_{n+1}) \mid X_n = i] = \sum_{j \in S} d^\mu(j) V^2(j) = \| V \|_D^2 \text{ . Now,}$$

$$\begin{aligned} V^T D^\mu \gamma P_\mu V &= \gamma V^T (D^\mu)^{1/2} (D^\mu)^{1/2} P_\mu V \\ &\leq \gamma \| (D^\mu)^{1/2} V \| \| (D^\mu)^{1/2} P_\mu V \| \\ &= \gamma \| V \|_D \| P_\mu V \|_D \leq \gamma \| V \|_D^2 = \gamma V^T D^\mu V. \end{aligned}$$

- Thus, $D^\mu (I - \gamma P_\mu)$ is positive definite as

$$V^T D^\mu (I - \gamma P_\mu) V \leq (1 - \gamma) \| V \|_D^2 > 0, \ \forall V \neq 0.$$

- Hence $\Phi^T D^\mu (I - \gamma P_\mu)\Phi$ is positive definite as well since Φ is full rank

Stochastic Approximation

- Objective: Solve the equation $F(\theta) = 0$ when analytical form of F is not known, however, noisy measurements $F(\theta(n)) + M_{n+1}$ can be obtained, where $\theta(n)$, $n \geq 0$ are the input parameters and M_{n+1}, $n \geq 0$ are zero-mean i.i.d. random variables

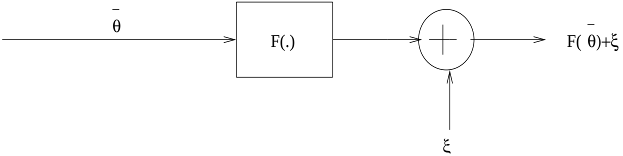


Figure: Noisy System with $E[\xi] = 0$

- More generally, the noise random variables M_{n+1}, $n \geq 0$ may depend on the 'system state' and may not be i.i.d.

- Algorithm (Robbins and Monro [1951])

$$\theta(n+1) = \theta(n) + a(n)(F(\theta(n)) + M_{n+1})$$

- Algorithm closes the loop

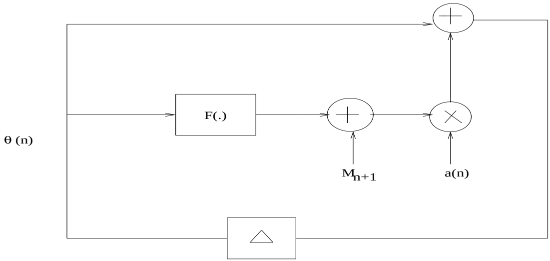


Figure: Robbins-Monro Algorithm

A More General Case

- Assume noise enters as an argument of the objective i.e., the available observations are $f(\theta(n), \eta_n)$ with i.i.d. η_n, $n \geq 0$, where $E[f(\theta, \eta_n) \mid \theta] = F(\theta)$
- Then

$$
\begin{aligned}
\theta(n+1) &= \theta(n) + a(n) f(\theta(n), \eta_n) \\
&= \theta(n) + a(n)(F(\theta(n)) + M_{n+1}),
\end{aligned}
$$

where $M_{n+1} = f(\theta(n), \eta_n) - F(\theta(n))$, $n \geq 0$ is a martingale difference sequence since $E[M_{n+1} \mid \theta(n)] = 0$, $\forall n$

Convergence of the Algorithm

- Step-size conditions: Assume

$$\sum_n a(n) = \infty; \quad \sum_n a(n)^2 < \infty$$

- Show stability of the iterates i.e., that $\sup_n \parallel \theta(n) \parallel < \infty$ w.p.1, or alternatively, $\sum_n a(n)(F(\theta(n) + M_{n+1}) < \infty$ w.p.1.
- Consider the associated ODE

$$\dot{\theta}(t) = F(\theta(t))$$

Let $K \stackrel{\triangle}{=} \{\theta \mid F(\theta) = 0\}$ denote the set of 'asymptotically stable equilibria' of this ODE (assuming they exist)

- One then argues that $\theta(n) \to K$ as $n \to \infty$ with probability one

The Borkar and Meyn Stability Theorem

- Borkar and Meyn [2000] analyze the recursion

$$X_{n+1} = X_n + a(n)(h(X_n) + M_{n+1}),$$

  under the following assumptions:
- **Assumption (B1):** (i) $h : \mathcal{R}^d \rightarrow \mathcal{R}^d$ is Lipschitz continuous and $h_c(x) \stackrel{\triangle}{=} h(cx)/c$, $c \geq 1$ satisfies $h_c \rightarrow h_\infty$, for some $h_\infty : \mathcal{R}^d \rightarrow \mathcal{R}^d$ uniformly on compacts.
  (ii) The origin in $\mathcal{R}^d$ is a unique globally asymptotically stable equilibrium for the ODE $\dot{x}(t) = h_\infty(x(t))$.
  (iii) There is a unique globally asymptotically stable equilibrium $x^* \in \mathcal{R}^d$ for the ODE $\dot{x}(t) = h(x(t))$.

B-M Stability (Contd)

- **Assumption (B2):** $\{M_n, \mathcal{G}_n, n \geq 1\}$ with $\mathcal{G}_n = \sigma(X_i, M_i, i \leq n)$ is a martingale difference sequence. Further for some constant $C_0 < \infty$ and any $X_0 \in \mathcal{R}^d$,

$$E[\| M_{n+1} \|^2 | \mathcal{G}_n] \leq C_0(1 + \| X_n \|^2), \; n \geq 0.$$

- **Assumption (B3):** $\{a(n)\}$ is a step-size sequence that satisfies $a(n) > 0$ for all n and

$$\sum_n a(n) = \infty, \quad \sum_n a(n)^2 < \infty.$$

- The Borkar-Meyn Theorem: Under (B1)-(B3), for any initial condition $X_0 \in \mathcal{R}^d$, $\sup_n \| X_n \| < \infty$ almost surely (a.s.). Further, $X_n \to x^*$ a.s. as $n \to \infty$.

Temporal Difference Learning - Full State Representation

- Cost-to-go for a given stationary policy μ is

$$V_\mu(s_j) = E\left[\sum_{m=0}^{\infty} \gamma^m k(s_{j+m}, \mu(s_{j+m}), s_{j+m+1})\right]$$

- Hence Poisson equation becomes

$$V_\mu(s_j) = E[k(s_j, \mu(s_j), s_{j+1}) + \gamma V_\mu(s_{j+1})]$$

- Alternatively, consider l-step Poisson equation

$$V_\mu(s_j) = E\left[\sum_{m=0}^{l} \gamma^m k(s_{j+m}, \mu(s_{j+m}), s_{j+m+1}) + \gamma^{l+1} V_\mu(s_{j+l+1})\right]$$

- Since l is arbitrary, consider the following weighted average of multi-step Poisson equations

- Suppose $0 \leq \lambda < 1$. Then

$$V_\mu(s_j) = (1-\lambda)E[\sum_{l=0}^{\infty} \lambda^l (\sum_{m=0}^{l} \gamma^m k(s_{j+m}, \mu(s_{j+m}), s_{j+m+1})$$

$$+\gamma^{l+1} V_\mu(s_{j+l+1}))]$$

- Since $(1-\lambda)\sum_{l=m}^{\infty} \lambda^l = \lambda^m$,

$$V_\mu(s_j) = E\left[(1-\lambda)\sum_{m=0}^{\infty} \gamma^m k(s_{j+m}, \mu(s_{j+m}), s_{j+m+1}) \sum_{l=m}^{\infty} \lambda^l\right]$$

$$+(1-\lambda)E\left[\sum_{l=0}^{\infty} \lambda^l \gamma^{l+1} V_\mu(s_{j+l+1})\right]$$

TD - FS (Contd.)

- Upon simplification, one obtains

$$V_\mu(s_j) = E\left[\sum_{m=0}^{\infty} \lambda^m \gamma^m d_{j+m}\right] + V_\mu(s_j)$$

where

$$d_{j+m} = k(s_{j+m}, \mu(s_{j+m}), s_{j+m+1}) + \gamma V_\mu(s_{j+m+1}) - V_\mu(s_{j+m})$$

- Stochastic Approximation Version:

$$J_{n+1}(s_j) = J_n(s_j) + a(n)\sum_{m=j}^{\infty} (\gamma\lambda)^{m-j} d_m$$

TD Learning with Function Approximation: TD(0)

- As described in the case of projection based methods, let

$$V_\mu(s) \approx \tilde{V}_\theta(s) = \theta^T \phi_s,$$

where $\phi_s = (\phi_s(1), \ldots, \phi_s(d))^T$ is a state-feature and $\theta = (\theta_1, \ldots, \theta_d)^T$ is the associated parameter

- Note that

$$\nabla \tilde{V}_\theta(s) = \phi_s.$$

- Define temporal difference term

$$\delta_n = k(s_n, \mu(s_n), s_{n+1}) + \gamma \theta_n^T \phi_{s_{n+1}} - \theta_n^T \phi_{s_n}$$

- The TD(0) Algorithm

$$\theta_{n+1} = \theta_n + a(n)\delta_n \phi_{s_n}, \; n \geq 0$$

Convergence of TD(0)

- Tsitsiklis and Van Roy [1997] give the first proof of convergence
- We present an alternative proof based on the B-M theorem
- Theorem: TD(0) Convergence
  Under Assumptions (A1), (A3) and (B3), $\{\theta_n, n \geq 0\}$ governed by TD(0) satisfy $\theta_n \to \theta^*$ with probability one, where θ^* is the unique solution to the system of equations

$$\Phi^T D^\mu \Phi \theta^* = \Phi^T D^\mu T_\mu(\Phi\theta^*). \tag{1}$$

In particular,

$$\theta^* = -(\Phi^T D^\mu(\gamma P - I)\Phi)^{-1}\Phi^T D^\mu K_\mu. \tag{2}$$

Proof of TD(0) Convergence

- Proof of TD(0) Convergence: The ODE associated with TD(0) recursion is the following:

$$\dot{\theta}(t) = \Phi^T D^\mu (T_\mu(\Phi\theta(t)) - \Phi\theta(t)) \stackrel{\triangle}{=} h(\theta(t)). \qquad (3)$$

  Note that $h(\cdot)$ is Lipschitz continuous. Let $h_\infty(\theta) \stackrel{\triangle}{=} \lim_{r\to\infty} \frac{h(r\theta)}{r}$. Consider also the ODE

$$\dot{\theta}(t) = h_\infty(\theta(t)) = \Phi^T D^\mu (\gamma P_\mu - I)\Phi\theta(t). \qquad (4)$$

- We have previously shown that $\Phi^T D^\mu (I - \gamma P_\mu)\Phi$ is positive definite. Hence, $\Phi^T D^\mu (\gamma P_\mu - I)\Phi$ is negative definite.

Proof of TD(0) Convergence (Contd.)

- From the foregoing, the ODE $\dot{\theta} = h_\infty(\theta) = \Phi^T D^\mu(\gamma P_\mu - I)\Phi\theta$ has the origin as its unique globally asymptotically stable equilibrium. Next, define M_n, $n \geq 0$ according to

$$M_{n+1} = (k(s_n, \mu(s_n), s_{n+1}) + \gamma\theta_n^T \phi_{s_{n+1}} - \theta_n^T \phi_{s_n})\phi_{s_n}$$

$$-E[(k(s_n, \mu(s_n), s_{n+1}) + \gamma\theta_n^T \phi_{s_{n+1}} - \theta_n^T \phi_{s_n})\phi_{s_n} \mid \mathcal{G}(n)],$$

where $\mathcal{G}(n) = \sigma(\theta_r, s_r, r \leq n)$. It is easy to see that

$$E[\| M_{n+1} \|^2 | \mathcal{G}(n)] \leq C_1(1 + \| \theta_n \|^2), \; n \geq 0, \tag{5}$$

for some constant $0 < C_1 < \infty$.

Proof of TD(0) Convergence (Contd.)

- Finally, consider the system of equations

$$h(\theta) = \Phi^T D^\mu (T_\mu(\Phi\theta) - \Phi\theta) = 0, \tag{6}$$

that can be alternatively written as

$$\Phi^T D^\mu K_\mu + \Phi^T D^\mu (\gamma P_\mu - I)\Phi\theta = 0. \tag{7}$$

Now since $\Phi^T D^\mu (\gamma P_\mu - I)\Phi$ is negative definite, it is of full rank and invertible. Hence θ^* (below) is the unique solution to (7)

$$\theta^* = -(\Phi^T D^\mu (\gamma P_\mu - I)\Phi)^{-1} \Phi^T D^\mu K_\mu.$$

Assumptions (A1)-(A3) are now satisfied and the claim follows from the Borkar-Meyn theorem. $\square$

TD Learning with Function Approximation: TD(λ)

- Sutton [1988], Tsitsiklis and Van Roy [1997]
- As before, we let

$$V_\mu(s) \approx V_\theta(s) = \theta^T \phi_s$$

- Define *eligibility trace*

$$z_n = \sum_{k=0}^{n} (\alpha\lambda)^{n-k} \nabla V_\theta(s_k)$$

$$= \sum_{k=0}^{n} (\alpha\lambda)^{n-k} \phi_{s_k}$$

- The TD(λ) Algorithm: Let $z_{-1} = 0$ and update

$$\theta_{n+1} = \theta_n + \gamma_n \delta_n z_n$$

$$z_{n+1} = \gamma\lambda z_n + \phi_{s_{n+1}}$$

Q-Value Iteration

- Define the action-value function or Q-value function associated with a stationary policy μ as

$$Q^{\mu}(i,a) = E_{\mu}\left\{\sum_{t=0}^{\infty} \gamma^t k(X_t, \mu(X_t), X_{t+1}) \mid X_0 = i, Z_0 = a\right\} \quad (8)$$

- Let $Q^*(i,a) = \min_{\mu} Q^{\mu}(i,a)$. Then

$$V^*(i) = \min_{a \in A(i)} Q^*(i,a)$$

Further, the Q-Bellman Equation holds.

$$Q^*(i,a) = \sum_j P_{ij}^a [k(i,a,j) + \gamma \min_{a' \in A(j)} Q^*(j,a')] \quad (9)$$

- VI for Q-Bellman equation or QVI: Start from an initial Q_0 and iterate $Q_{n+1}(i,a) = \sum_j P_{ij}^a (k(i,a,j) + \gamma \min_{a' \in A(j)} Q_n(j,a'))$

Q-learning with Full State Representation

- Watkins and Dayan [1992]
- It can be shown that $Q_n(i,a)$ obtained according to QVI satisfy $Q_n(i,a) \to Q^*(i,a)\ \forall (i,a)$, $i \in S$, $a \in A(i)$ as $n \to \infty$
- Stochastic Approximation Version of QVI: Let $\eta_n(i,a)$, $n \geq 0$ be independent random variables (simulation samples) having the common distribution $P^a_{i\cdot}$
- Let $c(n)$, $n \geq 0$ satisfy (A3).
- The QL-FS Algorithm: For every feasible state-action tuple (i,a), iterate

$$Q_{n+1}(i,a) = Q_n(i,a) + c(n)(k(i,a,\eta_n(i,a)) + \gamma \min_{v \in A(\eta_n(i,a))} Q_n(\eta_n(i,a),v) - Q_n(i,a)) \tag{10}$$

- Convergence of QL-FS can be shown using the Borkar-Meyn stability theorem.

Q-learning with Function Approximation

- Let $Q(i,a) \approx \theta^T \sigma_{i,a}$, where
  - $\sigma_{i,a}$: $\hat{d}$-dimensional feature vector corresponding to (i,a), with $\hat{d} << |S \times A(S)|$. Here
  $$S \times A(S) = \{(i,a) \mid i \in S, a \in A(i)\}$$
  - θ is a tunable $\hat{d}$-dimensional parameter
- Q-learning with FA: Let $\{s_n\}$ denote a sample trajectory of states of the MDP $\{X_n\}$. Also, let a_n be the action chosen at time n. Then,
$$\theta_{n+1} = \theta_n + c(n)\sigma_{s_n,a_n}(k(s_n,a_n,s_{n+1}) + \gamma \min_{v \in A(s_{n+1})} \theta_n^T \sigma_{s_{n+1},v} - \theta_n^T \sigma_{s_n,a_n})$$
- This algorithm suffers from the "off-policy" problem and hence it is difficult to prove its convergence in general. However, see Melo and Ribeiro [2007] for its convergence under some conditions.

Finite Difference Gradient Approximation

- Kiefer and Wolfowitz [1952]
- Problem: Estimate $\nabla J(\theta)$ when form of $J : \mathcal{R}^d \rightarrow \mathcal{R}$ is not known
- $\nabla J(\theta) = (\nabla_1 J(\theta), \ldots, \nabla_d J(\theta))^T$, where $\nabla_i J(\theta) = \dfrac{\partial J(\theta)}{\partial \theta_i}$, $i = 1, \ldots, d$.
- Finite Difference Balanced Estimate:

$$\nabla_i J(\theta) \approx (J(\theta + \delta e_i) - J(\theta - \delta e_i))/2\delta, \; i = 1, \ldots, d$$

  Requires $2d$ parallel simulations to estimate gradient once i.e., with parameters $\theta \pm \delta e_i$, $i = 1, \ldots, d$
- Finite Difference Unbalanced Estimate:

$$\nabla_i J(\theta) \approx (J(\theta + \delta e_i) - J(\theta))/\delta, \; i = 1, \ldots, d$$

  Requires $(d + 1)$ parallel simulations to estimate gradient once i.e., with parameters θ, $\theta + \delta e_i$, $i = 1, \ldots, d$

Simultaneous Perturbation Gradient Estimates

- Spall [1992]
- Unbalanced SP Gradient Estimate:

$$\nabla_i J(\theta) \approx (J(\theta + \delta\Delta) - J(\theta))/\delta\Delta_i, \; i = 1, \ldots, d$$

where $\Delta = (\Delta_1, \ldots, \Delta_d)^T$ is such that $\Delta_i = \pm 1$ w.p.1/2 and Δ_i are independent

- Using Taylor's argument, observe that

$$\frac{J(\theta + \delta\Delta) - J(\theta)}{\delta\Delta_i} \approx \nabla_i J(\theta) + \sum_{j=1, j \neq i}^{d} \frac{\nabla_j J(\theta)\Delta_j}{\Delta_i} + O(\delta)$$

Thus $E\left[(J(\theta + \delta\Delta) - J(\theta))/(\delta\Delta_i) \mid \theta\right] \approx \nabla_i J(\theta) + O(\delta)$

- Balanced SP Gradient Estimate:

$$\nabla_i J(\theta) \approx (J(\theta + \delta\Delta) - J(\theta - \delta\Delta))/2\delta\Delta_i, \; i = 1, \ldots, d$$

where Δ, $\Delta_1, \ldots, \Delta_d$ are as above.

Actor-Critic Algorithm with Full State Representation

- Bhatnagar and Kumar [2004]
- Assume $A(i)$ are compact sets for each $i \in S$ of type $\prod_{l=1}^{N}[\check{L}_i, \hat{L}_i]$.
  Let $a_i = (a_i^1, \ldots, a_i^N)^T$ be action taken in state i
- Run two parallel simulations with policies $\pi^1(n)$ and $\pi^2(n)$ at nth update where $\pi^1(n) = (P_i(a_i(n) - \delta\triangle_i(n)), i \in S)^T$ and $\pi^2(n) = (P_i(a_i(n) + \delta\triangle_i(n)), i \in S)^T$.
- Let $\{b(n)\}$ and $\{c(n)\}$ be two step-size schedules that satisfy
- **Assumption (C1):** $\sum_n b(n) = \sum_n c(n) = \infty$,
  $\sum_n b(n)^2, \sum_n c(n)^2 < \infty$ and $c(n) = o(b(n))$

- Actor recursion:

$$a_i^j(n+1) = P_i^j\left(a_i^j(n) + c(n)\left(\frac{V_{nL}^1(i) - V_{nL}^2(i)}{2\delta\triangle_i^j(n)}\right)\right),$$

where, for $m = 0, 1, \ldots, L-1$,

- Critic recursions:

$$V_{nL+m+1}^1(i) = V_{nL+m}^1(i) + b(n)(k(i, \pi_i^1(n)), \eta_{nL+m}^1(i, \pi_i^1(n)))$$

$$+\gamma V_{nL+m}^1(\eta_{nL+m}^1(i, \pi_i^1(n))) - V_{nL+m}^1(i)),$$

$$V_{nL+m+1}^2(i) = V_{nL+m}^2(i) + b(n)(K(i, \pi_i^2(n)), \eta_{nL+m}^2(i, \pi_i^2(n)))$$

$$+\gamma V_{nL+m}^2(\eta_{nL+m}^2(i, \pi_i^2(n))) - V_{nL+m}^2(i)).$$

Actor-Critic with FA for Average Cost

- Bhatnagar et al. [2009]
- Recall that for a given policy π (assume SRP),

$$\lambda_\pi = \lim_{N\to\infty} \frac{1}{N} E\left[\sum_{j=0}^{N-1} k(X_j, \mu_j(X_j), X_{j+1}) \mid \pi\right]$$

Further, for all $i \in S, a \in A(i)$,

$$Q^\pi(i,a) = \sum_{n=0}^{\infty} E[(k(X_n, \pi(X_n), X_{n+1}) - \lambda_\pi) \mid X_0 = i, Z_0 = a, \pi]$$

$$V^\pi(i) = \sum_{a\in A(i)} \pi(i,a) Q^\pi(i,a)$$

- The Poisson Equation:

$$\lambda_\pi + V^\pi(i) = \sum_{a\in A(i)} \pi(i,a) \sum_{j\in S} P_{ij}^{\pi(i)}(k(i,\pi(i),j) + V^\pi(j))$$

Policy Gradient Methods

- Let $\pi(i,a) \triangleq \pi^\theta(i,a) = Pr(Z_n = a \mid X_n = i, \theta)$.
- Goal: Find
$$\theta^\star = \arg\min_\theta \lambda_\pi.$$
- **Assumption (A3):** $\pi^\theta(i,a)$ is continously differentiable in θ for any $i \in S$, $a \in A(i)$
- An Important Result (Marbach-Tsitsiklis 2001, Sutton et al 2000, Baxter-Bartlett 2001): Under (A1) and (A3),
$$\nabla_\theta \lambda_\pi = \sum_{i \in S} d^\pi(i) \sum_{a \in A(i)} \nabla_\theta \pi(i,a) Q^\pi(i,a).$$

Compatible Features

- Suppose $\pi(i,a) = \dfrac{\exp(\theta^T \phi_{ia})}{\sum_{b \in A(i)} \exp(\theta^T \phi_{ib})}, \forall i \in S, a \in A(i)$, where each ϕ_{ia} is a $\hat{d}$-dimensional feature vector. Note that

$$\frac{\partial \pi(i,a)}{\partial \theta} = \pi(i,a)(\phi_{ia} - \sum_{b \in A(i)} \pi(i,b)\phi_{ib}) = \pi(i,a)\psi_{ia}$$

  Also note that $\displaystyle\sum_{a \in A(i)} \pi(i,a)\psi_{ia} = 0$

- In general, features ψ_{ia} derived from $\pi(i,a)$ according to $\psi_{ia} = \nabla_\theta \log \pi(i,a)$ are called compatible features.

A Generalization of Policy Gradient Theorem

- Generalization of PGT (Greensmith et al. [2004]):

$$\nabla_\theta \lambda_\pi = \sum_{i \in S} d^\pi(i) \sum_{a \in A(i)} \nabla_\theta \pi(i,a)(Q^\pi(i,a) - b(i)),$$

  for any *baseline* $b(i)$

- The Fisher information matrix (Amari [1998], Kakade [2002], Peters et al. [2003])

$$G(\theta) = E_{i \sim d^\pi, a \sim \pi}[\nabla_\theta \log \pi(i,a) \nabla_\theta \log \pi(i,a)^T]$$

$$= \sum_{i \in S} d^\pi(i) \sum_{a \in A(i)} \pi(i,a) \frac{\nabla_\theta \pi(i,a)(\nabla_\theta \pi(i,a))^T}{\pi(i,a)\pi(i,a)}$$

$$= \sum_{i \in S} d^\pi(i) \sum_{a \in A(i)} \pi(i,a) \psi_{ia} \psi_{ia}^T = E_{i \sim d^\pi, a \sim \pi}[\psi_{ia} \psi_{ia}^T].$$

Results for a Fixed SRP π

- Let $\mathcal{E}^\pi(w) = \sum_{i \in S} d^\pi(i) \sum_{a \in A(i)} \pi(i,a) \, [(w^T \psi_{ia} - Q^\pi(i,a) + b(i))^2]$ be the *mean squared error* of a parameterized (compatible) approximation to $Q^\pi(i,a)$ and $b(i)$ be an arbitrary baseline.
- Lemma 1: For given θ,

$$w^\star = \arg\min_w \mathcal{E}^\pi(w) = G(\theta)^{-1} E_{i \sim d^\pi, a \sim \pi}[Q^\pi(i,a)\psi_{ia}]$$

- Let $b^\star(i) = \arg \min_{b=(b(i), i \in S)} \mathcal{E}^\pi(w^\star)$.
- Lemma 2: For any given policy π, the minimum variance baseline $b^\star(i)$ corresponds to the value function $V^\pi(i)$.
- From Lemmas 1-2, $w^{\star T} \psi_{ia}$ serves as a least squares optimal parametric representation for the advantage $A^\pi(i,a) = Q^\pi(i,a) - V^\pi(i,a)$ as well, and not just $Q^\pi(i,a)$.

Results for a Fixed SRP π (Contd.)

- Let $\bar{\delta}_n = k(s_n, \pi(s_n), s_{n+1}) - J_n + \hat{V}_{s_{n+1}} - \hat{V}_{s_n}$ where $E[\hat{V}_{s_n} \mid s_n, \pi] = V^\pi(s_n)$, $E[J_n \mid s_n, \pi] = \lambda_\pi$. Then
- Lemma 3: Under given policy π with actions a_n chosen according to it, we have
$$E[\bar{\delta}_n \mid s_n, a_n] = A^\pi(X_n, a_n) \text{ a.s.}$$
- Let ϕ_i, $i \in S$ be a d-dimensional feature vector for state i. Let $V^\pi(i) \approx v^T \phi_i$, where v is a d-dimensional weight vector. Now suppose

$$\delta_n \stackrel{\triangle}{=} k(s_n, \pi(s_n), s_{n+1}) - J_n + v_n^T \phi_{s_{n+1}} - v_n^T \phi_{s_n},$$

and

$$\bar{V}^\pi(i) \stackrel{\triangle}{=} \sum_{a \in A(i)} \pi(i,a) \sum_{j \in S} P_{ij}^{\pi(i,a)}(k(i,\pi(i,a),j) - \lambda_\pi + v^{\pi T} \phi_j) \quad (11)$$

Function Approximation Version of Policy Gradient Theorem

- Lemma 4 (Function Approximation Analog of PGT (Bhatnagar et al. [2009])):

$$E[\delta_n \psi_{s_n,a_n} \mid \theta] = \nabla_\theta \lambda_\pi + \sum_{i \in S} d^\pi(i)(\nabla_\theta \bar{V}^\pi(i) - \nabla_\theta {v^\pi}^T \phi_i).$$

- Corollary 1:

$$\sum_{i \in S} d^\pi(i)(\bar{V}^\pi(i) - {v^\pi}^T \phi_i) = 0.$$

In what follows, we also assume the following in addition to (A1)-(A3) and (C1):

- **Assumption (A4):** For every $v \in \mathcal{R}^d$, $\Phi v \neq e$, where e is the n-dimensional vector with all entries equal to one.

- Let $\xi(n) = cb(n)$ for some $c > 0$. Then

$$J_{n+1} = (1 - \xi(n))J_n + \xi(n)k(s_n, \pi(s_n), s_{n+1}), \tag{12}$$

$$\delta_n = k(s_n, \pi(s_n), s_{n+1}) - J_{n+1} + v_n^T \phi_{s_{n+1}} - v_n^T \phi_{s_n}, \tag{13}$$

$$v_{n+1} = v_n + b(n)\delta_n \phi_{s_n}, \tag{14}$$

$$\theta_{n+1} = \theta_n - c(n)\delta_n \psi_{x_n z_n}. \tag{15}$$

- The recursions (12)-(14) correspond to TD(0) for long-run average cost. Also, observe that the TD term δ_n is used in both actor and critic recursions.

An Application

- Traffic Signal Control (Prashanth and Bhatnagar [2010])
- AIM: Maximize traffic flow across intersections through adaptive control of traffic lights
  - State: $s_n = (q_1, \ldots, q_N, t_1, \ldots, t_N)$
  - Action: $A_n = \{\text{feasible sign configurations in state } s_n\}$
  - Cost:

$$\begin{aligned} k(s_n, a_n) = \quad & r_1 * (\textstyle\sum_{i \in I_p} r_2 * q_i(n) + \sum_{i \notin I_p} s_2 * q_i(n)) \\ + \quad & s_1 * (\textstyle\sum_{i \in I_p} r_2 * t_i(n) + \sum_{i \notin I_p} s_2 * t_i(n)), \end{aligned} \tag{16}$$

  - where $r_i, s_i \geq 0$ and $r_i + s_i = 1, i = 1, 2$
  - We set $r_1 = s_1 = 0.5$ and $r_2 = 0.6, s_2 = 0.4$ in experiments

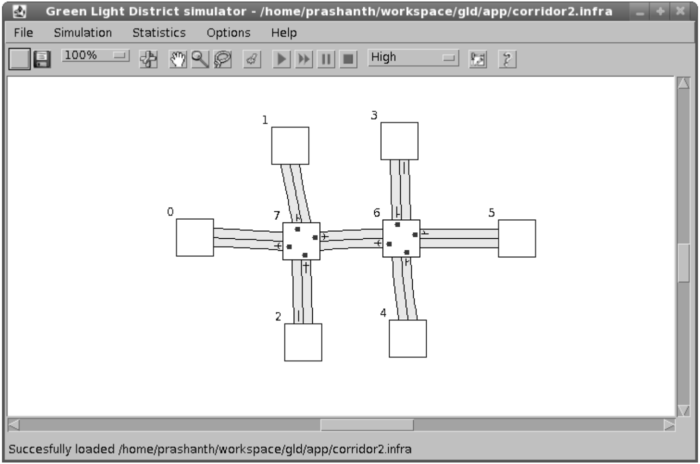
- State-action features

$$\begin{aligned}
\sigma_{s_n,a_n} &= (\sigma_{q_1(n)},\ldots,\sigma_{q_N(n)},\sigma_{t_1(n)},\ldots,\sigma_{t_N(n)},\\
&\quad \sigma_{a_1(n)},\ldots,\sigma_{a_M(n)})^T\\
\text{where}\\
\sigma_{q_i(n)} &= \begin{cases} 0 & \text{if } q_i(n) < L1\\ 0.5 & \text{if } L1 \le q_i(n) \le L2\\ 1 & \text{if } q_i(n) > L2 \end{cases} \qquad (17)\\
\sigma_{t_i(n)} &= \begin{cases} 0 & \text{if } t_i(n) \le T1\\ 1 & \text{if } t_i(n) > T1 \end{cases}\\
\sigma_{a_i(n)} &= \text{sign config chosen at junction } i
\end{aligned}$$

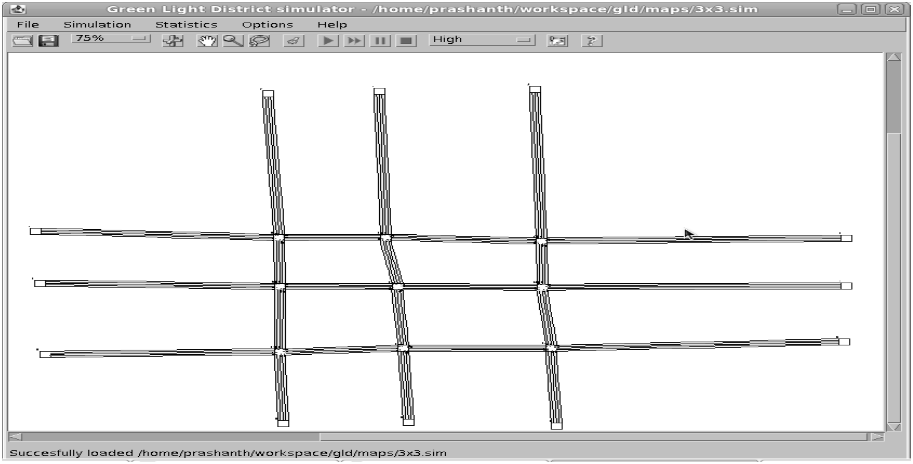
Other Algorithms Implemented

- **Fixed Timing TLC**
  - cycle periodically through feasible sign configurations
- **Self Organizing TLC** (SOTL) (Cools et al. [2008])
  - switch lane to green if elapsed time crosses a threshold, provided the # of vehicles crosses another threshold
- **Longest Queue TLC** (LTLC)
  - switch lane to green if it has the longest queue
- **Q-learning with Full State Representation** (QTLC-FS)
- **Q-learning with No Priority** (QTLC-NP) (Abdulhai et al. [2003])
  - similar to QTLC-FS, but no prioritization of traffic

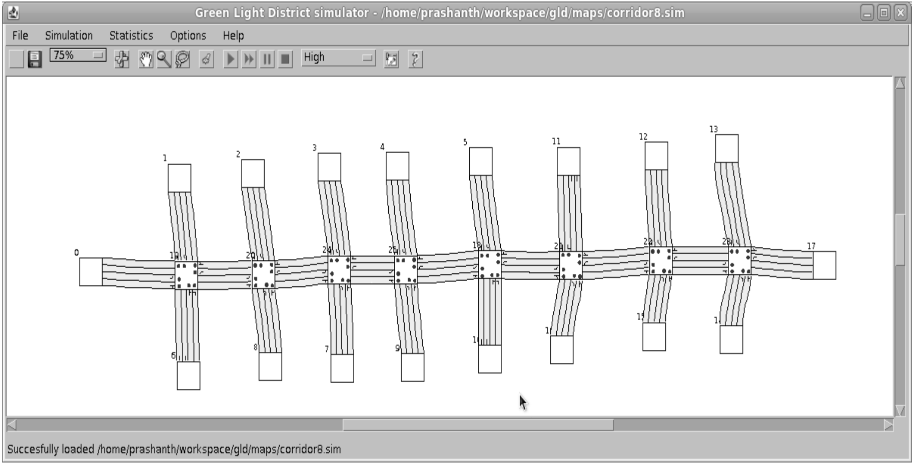
A Two-Junction Corridor Setting (1)



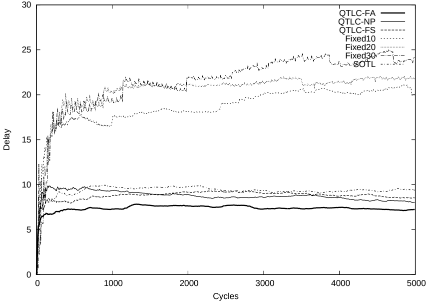
A 3 $\times$ 3–Grid Network (2)



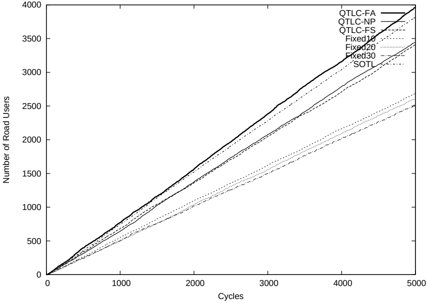
An Eight-Junction Corridor (3)



Setting (1)



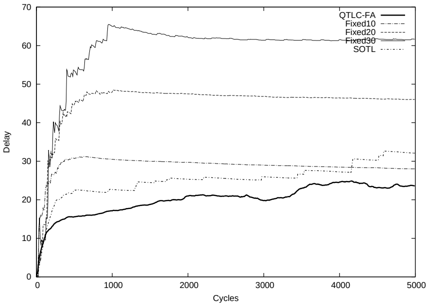
(a) Average junction waiting time



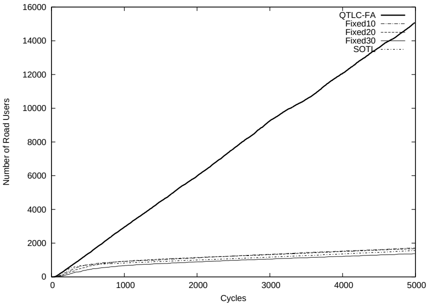
(b) Total Arrived Road Users

- LTLC: traffic invariably entered a deadlock situation
- It is interesting to note that QTLC-FA is better than both QTLC-FS and QTLC-NP

Setting (2)



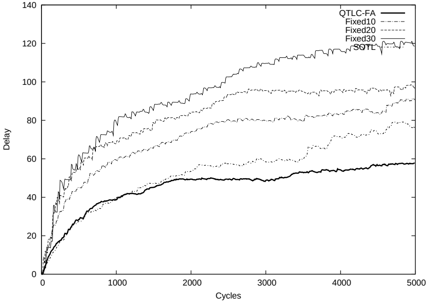
(c) Average junction waiting time



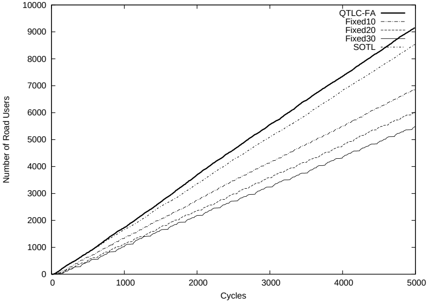
(d) Total Arrived Road Users

- QTLC-FS and QTLC-NP are not even implementable on a 3x3-grid because the size of state-action space $|S \times A(S)| \sim 10^{101}$
- On the other hand, in QTLC-FA, the number of features (i.e., clusters from the above state-action space over which the algorithm works) is about 200

Setting (3)



(e) Average junction waiting time



(f) Total Arrived Road Users

- Here also sizes of state-action spaces are large. Hence, QTLC-FS and QTLC-NP are not implementable
- QTLC-FA shows the best results as in previous settings

Important Topics Not Covered in this Tutorial

- Non-incremental methods (LSTD, LSPE etc.)
- RL for constrained MDPs
- Algorithms with Bellman error objectives
- Algorithms with off-policy and nonlinear function approximation
- Feature adaptation methods
- POMDPs
- $\dots$

References - 1

- Abdulhai, B., Pringle, R. and Karakoulas, G.J. (2003) “Reinforcement learning for true adaptive traffic signal control”, *Journal of Transportation Engineering*, 129: 278-285.
- Amari, S. (1998) “Natural gradient works efficiently in learning”, *Neural Computation*, 10(2):251-276.
- Baxter, J. and Bartlett, P. L. (2001) “Infinite-horizon policy-gradient estimation”, *Journal of Artificial Intelligence Research*, 15:319-350.
- Bertsekas, D.P. (2005) *Dynamic Programming and Optimal Control, Vol.I, 3rd Ed.*, Athena Scientific, Belmont, MA.
- Bertsekas, D.P. (2007) *Dynamic Programming and Optimal Control, Vol.II, 3rd Ed.*, Athena Scientific, Belmont, MA.

References - 2

- Bertsekas, D.P. (2010) Approximate Dynamic Programming, Chapter 6 of *Dynamic Programming and Optimal Control, Vol.II, 3rd Ed.*, http://web.mit.edu/dimitrib/www/dpchapter.pdf.
- Bertsekas, D.P. and Tsitsiklis J.N. (1996) *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA.
- Bhatnagar, S. and Kumar, S. (2004) "A simultaneous perturbation stochastic approximation based actor–critic algorithm for Markov decision processes", *IEEE Transactions on Automatic Control*, 49(4):592-598.
- Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M. and Lee, M. (2009) "Natural actor-critic algorithms", *Automatica*,45: 2471–2482.
- Borkar, V. S. and Meyn, S. P. (2000) "The O.D.E. method for convergence of stochastic approximation and reinforcement learning", *SIAM Journal of Control and Optimization*, 38(2):447-469.

References - 3

- Cools, S.B., Gershenson, C. and DHooghe, B. (2008) “Self-organizing traffic lights: A realistic simulation”, *Advances in Applied Self-organizing Systems*, pp. 41–50.
- Greensmith, E., Bartlett, P. L. and Baxter, J. (2004) “Variance reduction techniques for gradient estimates in reinforcement learning”, *Journal of Machine Learning Research*, 5:1471-1530.
- Kakade, S. (2002) “A Natural Policy Gradient”, *Advances in Neural Information Processing Systems*, 14.
- Kiefer, E. and Wolfowitz, J. (1952) “Stochastic estimation of the maximum of a regression function”, *Ann. Math. Statist.*, 23:462466.
- Marbach, P. and Tsitsiklis J.N. (2001) “Simulation-based optimization of Markov reward processes”, *IEEE Transactions on Automatic Control*, 46(2):191-209.

References - 4

- Melo, F. and Ribeiro, M. (2007) “Q-learning with linear function approximation”, *Learning Theory*, pp. 308–322.
- Peters, J., Vijayakumar, S. and Schaal, S. (2003) “Reinforcement learning for humanoid robotics”, *Proceedings of the Third IEEE-RAS International Conference on Humanoid Robots*.
- Prashanth, L.A. and Bhatnagar, S. (2010) “Reinforcement learning with function approximation for traffic signal control”, *IEEE Transactions on Intelligent Transportation Systems*, (to appear) (DOI: 10.1109/TITS.2010.2091408).
- Puterman, M.L. (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley, New York.
- Robbins, H. and Monro, S. (1951) “A stochastic approximation method” *Ann. Math. Statist.*, 22:400–407.

References - 5

- Sutton, R. and Barto, A. (1998) *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA.
- Sutton, R., McAllester, D., Singh, S. and Mansour, Y. (2000) “Policy gradient methods for reinforcement learning with function approximation”, *Advances in Neural Information Processing Systems*, 12:1057-1063.
- Watkins, C. and Dayan, P. (1992) “Q-learning”, *Machine Learning*, 8:279-292.
- Tsitsiklis, J. N. and Van Roy, B. (1997) “An analysis of temporal difference learning with function approximation”, *IEEE Transactions on Automatic Control*, 42(5):674-690.
- Sutton, R. (1988) “Learning to predict by the methods of temporal differences”, *Machine Learning*, 3:835-846.
- Spall, J.C. (1992) “Multivariate stochastic approximation using a simultaneous perturbation gradient approximation”, *IEEE Trans. Autom. Contr.*, 37(3):332-341.