Outline

• Sensitivity based optimization of Markov systems
• Optimal control of queueing systems
  – Service rate control
  – Admission control
• Course discussion and improvement
Markov decision process

• Markov chain + decision making
  – System dynamic is Markov chain
  – System evolution is controlled by actions
  – Performance criterion evaluate policies

• Solution approach
  – Dynamic programming, reinforcement learning

• Wide application
  – robotics, automated control, economics, and manufacturing
Markov decision process

- 4-tuple \{S,A,P,R\}
  - S: state, s
  - A: action, a
  - P: transition probability, \( p(s' | s, a) \)
  - R: reward, \( r(s, a) \)

- Category
  - Discrete/continuous time, discrete/continuous state MDP
Markov decision process

• Policy, \( L: S \rightarrow A \), a mapping from state space \( S \) to action space \( A \)

• Performance criterion
  – Long-run average, discount, finite stage

• Optimization problem

\[
L^* = \arg \max_{\eta} \arg \max_{all \ L} \ E_t \{r(s_t, L(s_t))\}
\]
MDP model (discrete case)

- $X=\{X_n, \ n=1,2,\ldots\}$, ergodic, $X_n$ in $S=\{1,2,\ldots,S\}$.
- Transition prob. Matrix $P=[p(i,j)]_{i,j=1,\ldots,S}$
- Steady-state probability: $\pi=(\pi(1), \pi(2),\ldots,\pi(S))$.
- Performance function: $f=(f(1), \ldots, f(S))^T$
- Steady-state performance: $\eta = \pi f = \sum_i \pi(i)f(i)$

$Pe=e, \quad \pi(I-P)=0, \quad \pi e=1. \quad e=(1,1,\ldots,1)^T, \quad I: \text{identity}$
Traditional theory of MDP

• Bellman optimality equation

\[ v^*(i) = \max_{a \in A} \left\{ r(i,a) - \eta + \sum_{j \in S} p(j \mid i,a) v^*(j) \right\} \]

– \( v^* \): value function of optimal policy

• Optimization algorithm

– Policy iteration (global optimal)
– Value iteration (\( \epsilon \)-optimal)
Sensitivity-based optimization of MDP

• Proposed by X. R. Cao (1997)
• A fundamental quantity
  – Performance potential
• Two sensitivity equations
  – Performance difference equation
  – Performance derivative equation
• New perspective to study optimization
Performance potential

• Definition, \( g(i), \ i \in S \)

\[
g(i) = \lim_{N \to \infty} E \left\{ \sum_{n=0}^{N} [f(X_n) - \eta] \mid X_0 = i \right\}
\]

\[
= f(i) - \eta + \lim_{N \to \infty} E \left\{ \sum_{j=1}^{S} p(j \mid i) \sum_{n=1}^{N} [f(X_n) - \eta] \mid X_1 = j \right\}
\]

• In matrix form, obtain the Poisson equation

\[(I - P + e \pi)g = f\]

• Physical meaning
  – Measure the long-term contribution of initial state to the average performance
Performance potentials, from the sample path viewpoint

\[ g(i) = \lim_{N \to \infty} E\left\{ \sum_{n=0}^{N} [f(X_n) - \eta] \mid X_0 = i \right\} \]

\[ \approx E\left\{ \sum_{n=0}^{N} [f(X_n)] \mid X_0 = i \right\} \]

\[ g(i) = f(X_0) + f(X_1) + \ldots + f(X_n) \]

\[ X_0 = i \quad X_1 \quad X_n \]
Two Sensitivity Equations

For two Markov chains $P, \eta, \pi$ and $P', \eta', \pi'$, let $Q = P' - P$

**Performance difference:**

$$\eta' - \eta = \pi' Q g = \pi' (P' - P) g$$

$g$: performance potentials

**Performance gradient:**

$$\frac{d\eta(\delta)}{d\delta} = \pi Q g = \pi P' g - \pi P g$$

set $P(\delta) = (1-\delta)P + \delta P' = P + \delta Q$ denote as $\pi(\delta)$ and $\eta(\delta)$
A Simple Derivation

• Poisson equation and potentials $g$ :
  \[(I - P + e\pi)g = f\]

• Two MCs: $P, P'\Rightarrow\pi,\pi'$ and $\eta, \eta'$
  \[\pi g = \pi f = \eta\quad\eta' = \pi' f\]
  \[\eta' - \eta = \pi'(P' - P)g = \pi'Qg\]

• Set $P(\delta)=P+Q$, Replacing $P'$ with $P(\delta)$, we have
  \[\eta(\delta) - \eta = \pi(\delta)[P(\delta) - P]g = \pi(\delta)Q\delta g\]

  \[\delta \to 0 \Rightarrow \quad \frac{d\eta(\delta)}{d\delta} = \pi Qg = \pi P'g - \pi Pg\]
Two Basic Approaches

- Continuous Parameters (policy gradient)
  \[ \frac{d \eta}{d \delta} = \pi Qg \]
  a. Only local information can be learned!
  b. PA⇒ estimate derivatives on a sample path
  c. \( \pi' > 0 \) ⇒ If \( Qg > 0 \), then \( \eta' > \eta \).

- Discrete Policy Space (policy iteration)
  \[ \eta' - \eta = \pi' Qg \]
Optimal control of queueing networks

• Service rate control of closed Jackson network
  – Identify the optimal service rates which attains the maximal average performance of the system

• Admission control of open Jackson network
  – Controller determines the admission of external arrivals (entrance or reject)
  – Identify the optimal admission policy which attains the maximal average performance
Service rate control in closed Jackson network

$M$ servers, $N$ customers, service rates: $\mu_i$, $i = 1, \cdots, M$

routing probability: $q_{ij}$, $i, j = 1, \cdots, M$

number of customers at server $i$ : $n_i$, system state: $n = (n_1, \cdots, n_M)$

cost function: $f(n(t), \mu)$

$$f(n, \mu) = \phi(n) + R \sum_{i=1}^{M} \mathbf{1}_{n_i > 0} \mu_i$$

time-average performance:

$$\eta = \lim_{T \to \infty} \frac{\int_0^T f(n(t))dt}{T}$$
MDP model for service rate control of queueing network

Markov model

- Process $X=\{X_t, t=1,2,\ldots\}$, ergodic, $X_t$ is state in the space $S=\{1,2,\ldots,S\}$. Here, $X_t$ is equivalent to previous $n = (n_1, \ldots, n_M)$
- State transition probability matrix $P = [p(i,j)]_{ij=1,\ldots,S}$, here $P$ is determined by $\mu_i, q_{ij}$, $i, j = 1, \ldots, M$. Policy is $U = \{\mu_i, i = 1, \ldots, M\}$
- Steady-state probability: $\pi = (\pi(1), \pi(2), \ldots, \pi(S))$.
- Cost function: $f = (f(1), \ldots, f(S))^T$
- Time-average performance: $\eta = \pi f = \sum_i \pi(i) f(i)$

$P e = e, \quad \pi(I-P)=0, \quad \pi e = 1. \quad e = (1,1,\ldots,1)^T, \quad I$: identity
Difference equation for queueing systems

Two sets of service rates $\mu'_{i,n}$ and $\mu_{i,n}$

and cost functions $f'(n)$ and $f(n)$

$$\eta' - \eta = \sum_{n \in S} \pi'(n) \sum_{i=1}^{M} [\mu'_{i,n} - \mu_{i,n}] \left\{ \sum_{j=1}^{M} q_{ij} [g(n_{-i+j}) - g(n)] + R \right\}$$

Since we always have $\pi' > 0$ \Rightarrow

If $[\mu'_{i,n} - \mu_{i,n}] \left\{ \sum_{j=1}^{M} q_{ij} [g(n_{-i+j}) - g(n)] + R \right\} > 0$

then $\eta' > \eta$

Policy Iteration
Max-Min optimality of service rate control

• Service rate control of Jackson networks
  – Find optimal $\mu_{i,n}^*$ leading to optimal performance

• Most general results in literature
  – C1: if $f(n, \mu)$ is linear to $\mu_{i,n}$, then $\eta$ is monotone, $\mu_{i,n}^*$ is either maximum or minimum
  – C2: if $f(n, \mu)$ is strictly convex (concave) to $\mu_{i,n}$, then $\mu_{i,n}^*$ for the performance maximization (minimization) problem is either maximum or minimum.
  
  • e.g., $f(n) = \sum_{i=1}^{M} n_i + \sum_{i=1}^{M} \mu_{i,n}^2$

• Reduce search space $|A|^{S \times M} \rightarrow |S| \times 2^M$

Optimization for service rate control

- Policy iteration, service rates as policy
  - Initialization – evaluation – improvement – evaluation – improvement – … stop

- Fast convergence
- Global optimum
- Special property
  - Max-Min service rates
- Example
  - 3-server, 5-customer
  - 4 iterations to get optimum
Admission control of open Jackson networks

- Problem formulation
  - customer Poisson arrival with rate $\lambda$
  - # of servers $M$
  - # of customers at server $i$, $n_i$
  - service rate $\mu_i$, exponential
  - system state $n=(n_1,...,n_M)$
  - only observe $n=n_1+...+n_M$ (# of total customers), not $n$
  - network capacity is $N$, $n \leq N$
  - routing prob., $q_{ij}$
  - first come first serve
Problem formulation (cont’d)

- Adm. Ctrl. in semi-open Jackson net.
  - observe: customer arrival event
    \[ e(n), n=0,1,\ldots, N \]
  - determine the admission prob.
    \[ a(n), n=0,1,\ldots, N \]
  - reward \( R \) per service
  - cost \( C \) per unit waiting time
  - reward function
    \[ f(n) = -C \sum_{i=1}^{M} n_i + R \sum_{i=1}^{M} 1_{n_i > 0} \mu_i \]
  - goal: optimal adm. prob.
    \[ a(n), n=0,1,\ldots, N \]
  - event-based policy
    \[ d: \{n=0,1,\ldots, N\} \to R[0,1] \]
Problem characteristics

• Observe partial information of the system
  – Only observe $n$, not system state $n$, $|n| << |\bar{n}|$

• Action is triggered by event, not state
  – When customer arrival event happens, make decision
  – When customers transit among servers, no need of decision
  – Event happening is much rarer than states

• Actions may be dependent of states
  – Admission probability $a(n)$ is identical for different states $n$ with the same $n$.
    E.g., states $n=(3,1,1),(2,1,2),(1,4,0)$ have the same $a(n=5)$.
  – Conflict with standard MDP theory

• Markovian, but not a standard MDP
  use Event Based Optimization (EBO) framework
Brief introduction of EBO theory

• Proposed by X.R. Cao (2005, 2007), etc.

• Definition of event

\[ e := \{ \langle i, j \rangle : i, j \in S \text{ and } \langle i, j \rangle \text{ has common properties} \} \]

– E.g., customer arrival event \( e := \{ \langle n, n_{+} \rangle, \langle n, n \rangle \} \)
– Event is much rarer
– Reveal future information of the system

• Action \( a \) at event \( e \) happening, determine trans. prob.,

\[ p^a (j \mid i, e), \quad i, j \in S, a \in A, e \in \mathbb{E}, \langle i, j \rangle \in e \]

• Event-based policy \( d : \mathbb{E} \rightarrow A \), and \( d \in D_e \)

• Goal:

\[ d^* = \arg \max_{d \in D_e} \{ \eta^d \} \]
Sample path of an EBO problem

\[ E_{l_k} \]: the event at the \( k \)-th event occurrence time (\( l_1=2, l_2=6, l_3=9 \))
Direct comparison for EBO systems

Direct Comparison:

performance difference equation under policy $h$ and $d$:

$$
\eta^h - \eta^d = \pi^h \left[ (P^h - P^d) g^d + (f^h - f^d) \right]
$$

$$
= \sum_{e \in \mathcal{E}} \pi^h(e) \sum_{i \in I(e)} \pi^h(i \mid e) \left\{ \sum_{j \in O_i(e)} \left[ p^h(j \mid i, e) - p^d(j \mid i, e) \right] g^d(j) + \left[ f^h(i) - f^d(i) \right] \right\}
$$

Conditional probability:

$$
\pi^h(i \mid e) = \pi^d(i \mid e), \quad \forall i \in I(e), \forall e \in \mathcal{E}, \forall h, d \in D_e
$$
Policy iteration for EBO

Proposition: Estimate $g^d(i)$’s and $\pi^d(i|e)$’s based on sample path, suppose the estimates are accurate enough. Update policy as:

$$h(e) := \arg \max_{a \in A} \sum_{i \in I(e)} \pi^d(i|e) \left\{ f(i,a) + \sum_{j \in O_i(e)} p^a(j|i,e)g^d(j) \right\}, \forall e \in \mathbb{E}$$

Then, we have $\eta^h > \eta^d$

- Require: $\pi^h(i|e) = \pi^d(i|e)$, for all policies $h$ and $d$
- Construct policy iteration for EBO

Repeat this process: Policy Iteration
Difference equation for admission control

• When event-based policy (admission prob. \( a(n) \), \( n=0,\ldots,N-1 \)) changes, performance difference is

\[
\eta' - \eta = \sum_{n=0}^{N-1} \pi'(n)[a'(n) - a(n)] \lambda \sum_{n \in S_n} \pi'(n \mid n) \left\{ \sum_{i=1}^{M} q_{0i} [g(n_{+i}) - g(n)] \right\}
\]

• Does the conditional probability has

\[
\pi^h (i \mid e) = \pi^d (i \mid e), \text{ i.e., } \pi(n \mid n) = \pi'(n \mid n) ?
\]
Property of conditional probability

• By product-form solution of Jackson network, we prove
  \[ \pi(n|n) = \pi'(n|n), \text{ for different } a(n), \ n=0,\ldots,N-1 \]

• Difference equation becomes
  \[ \eta' - \eta = \sum_{n=0}^{N-1} \pi'(n)[a'(n) - a(n)]D(n) \]

where \( D(n) := \lambda \sum_{n \in S_p} \pi(n|n) \left\{ \sum_{i=1}^{M} q_{0i}[g(n_{+i}) - g(n)] \right\} \), can be estimated from sample path
Property of admission control problem

### Monotonicity:

**Theorem 1:** System performance $\eta$ is monotonic w.r.t. the admission probabilities $a(n)$, $n=0,…,N-1$.

### Optimality of threshold-type policy:

**Theorem 2:** There exists a threshold $\theta^*$ in $\{0,1,…,N\}$, which has $a^*(n)=1$, if $n<\theta^*$; $a^*(n)=0$, if $n\geq\theta^*$. That is, threshold-type policy is an optimal policy.

### Difference equation:

performance difference under two threshold-type policies with $\theta$ and $\theta'$, and $\theta > \theta'$:

$$\eta' - \eta = - \pi'(\theta') D(\theta')$$

### Simplification:

$d : \{n=0,1,…,N\} \rightarrow R[0,1]$, simplified to find ONE para., $\theta^*$
Sufficient and necessary condition of optimal threshold

**Sufficient condition:**

| Theorem 3: $\theta^*$ is an optimal threshold if it satisfies, $D(n) \geq 0$, for $n=0,...,\theta^*-1$; $D(n) \leq 0$, for $n=\theta^*,...,N-1$. |

**Necessary condition:**

| Theorem 4: If $\theta^*$ is an optimal threshold, it has to satisfy, $D(n) \geq 0$, for $n=0,...,\theta^*-1$; $D(n) \leq 0$, for $n=\theta^*$. |
Iterative algorithm to find the optimal threshold

Algorithm:

1. **Initialization**: choose initial threshold as $\theta^{(0)}=N$, set $\Theta=\{0,1,\ldots,N-1\}$ and $k=0$.

2. **Evaluation**: for $\theta^{(k)}$, calculate or estimate $D(n)$.

3. **Reduction**: if $D(n) \geq 0$, remove $n$ from $\Theta$.

4. **Stopping**: if $\Theta=\emptyset$, set $\theta^*=\theta^{(k)}$ and stop; otherwise, choose $\theta^{(k+1)}=\max\{\Theta\}$ and remove $\theta^{(k+1)}$ from $\Theta$, set $k:=k+1$ and go to step 2.

Result on:

Simulation setting

- Parameter setting
  - $M=3$, $N=6$, $\lambda=50$, $\mu_1=40$, $\mu_2=50$, $\mu_3=80$
  - Routing prob. matrix $[[0,0.3,0.4,0.3; 0.2,0,0.4,0.4; 0.4,0.3,0,0.3; 0.3,0.2,0.5,0]]$
  - $R=10$, $C=100$
Simulation results

3 iterations to find the optimum
Optimal threshold $\theta^* = 4$
Maximal performance is $\eta^* = 576.1$
Comply to the numerical results
Bang-bang control is optimal

Fig 1. The curve of performance w.r.t. thresholds

Fig 2. The optimal admission probabilities