Optimal learning of transition probabilities in the two-agent newsvendor problem

Ilya O. Ryzhov[1] Martin R. Valdez-Vivas[2] Warren B. Powell[1]

[1]Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

[2]Management Science and Engineering
Stanford University
Stanford, CA 94305, USA

Winter Simulation Conference
December 6, 2010

Outline

Outline

The two-agent newsvendor problem: a repeating game

- Two players with different cost functions coordinate to meet an uncertain demand
- In every time period, the requesting agent submits a request for funding
- The oversight agent decides how much of the request to satisfy
- Both players incur costs and play again

IMF LOANS AMID FINANCIAL CRISIS 2008-9

Country	IMF Loans (US$ Millions)	IMF Loan as % of GDP	Status
Armenia	$540	4.47%	Approved
Belarus	$2,460	4.26%	Approved
El Salvador	$800	3.59%	Approved
Georgia	$750	5.65%	Approved
Hungary*	$15,700	11.00%	Approved
Iceland	$2,100	11.04%	Approved
Kenya	$100	0.32%	
Latvia**	$2,130	6.00%	
Malawi	$77	1.89%	
Mongolia	$224	4.49%	
Pakistan	$7,600	4.72%	
Romania	$17,500	8.18%	
Serbia	$520	1.00%	
Seychelles	$26	3.34%	
Sri Lanka	$1,900	4.51%	
Turkey	$25,000	3.13%	
Ukraine	$16,400	8.28%	
Zambia	$100 to $150	0.66% to 0.98%	

Applications: printing industry

- Planners determine how much time is required by each job on a printing press
- If too little time is allocated, the job runs over or is incomplete (costly for the customer)
- If too much time is allocated, it slows down the presses (costly for the planner)
- A customer may ask for more time than is necessary, to make sure the job finishes

Applications

- Pricing IPOs: a bank solicits information from investment banks, who indicate a price lower than they are willing to pay
- Project scheduling: an IT department requests 3000 hours to complete a programming assignment
- Marketing: a marketing department requests a budget for a new advertising campaign
- Academic budgeting: the Vice Provost at a university receives funding requests from departments and academic programs

The two-agent newsvendor problem

- The requesting agent can submit misleading requests, exaggerating the amount needed
- Through repeated play, the oversight agent can gradually learn the bias behaviour of the requesting agent
- The oversight agent needs to balance exploitation (minimizing costs) with exploration (learning the biases)
- We use optimal learning techniques to achieve a good trade-off

Outline

Generation of the request

- D is a demand process that is unknown to both agents
- However, at time n, the requesting agent has access to an unbiased estimate $\hat{D}^n$ of D^n
- Assuming normal distributions, costs are minimized by ordering at the critical quantile (Arrow et al. 1951)

$$q^{order,n} = \hat{D}^n + \sigma\Phi^{-1}\left(\frac{c_r^u}{c_r^u + c_r^o}\right)$$

- For underage/overage costs c_r^u, c_r^o and allocation quantity $q^{alloc,n}$, the requesting agent receives the newsvendor payoff

$$C^R\left(D^n, q^{alloc,n}\right) = c_r^u\left(D^n - q^{alloc,n}\right)^+ + c_r^o\left(q^{alloc,n} - D^n\right)^+$$

The bias in the request

- The oversight agent has a different cost structure $\frac{c_a^u}{c_a^o} < \frac{c_r^u}{c_r^o}$, for which the optimal quantity is

$$q^{alloc,n} = \hat{D}^n + \sigma\Phi^{-1}\left(\frac{c_a^u}{c_a^u + c_a^o}\right) < q^{order,n}$$

- The requesting agent may deliberately give an inflated request

$$Q^n = q^{order,n} + \beta^n$$

- The oversight agent's decision can be written in terms of β^n as

$$q^{alloc,n} = Q^n + \sigma\left(\Phi^{-1}\left(\frac{c_a^u}{c_a^u + c_a^o}\right) - \Phi^{-1}\left(\frac{c_r^u}{c_r^u + c_r^o}\right)\right) - \beta^n$$

Casting the problem in terms of cost differentials

- The decision can now be viewed as a difference

$$x^n = q^{alloc,n} - \left(Q^n - \sigma \Phi^{-1} \left(\frac{c_r^u}{c_r^u + c_r^o} \right) \right)$$

  representing how much the request is under-funded, rather than the allocation quantity
- The only uncertain quantity in the problem is now the bias β^n:

$$C^O(\beta^n, x^n) = c_a^u [-(x^n + \beta^n)]^+ + c_a^o [x^n + \beta^n]^+ .$$

- The request Q^n is now contained in the decision x^n

The cost history

- The bias β^n is drawn randomly from a finite set $\{b_1, ..., b_K\}$ where $b_1 < ... < b_K$
- The requesting agent's bias behaviour (the pmf of β^n) depends on the past history of the game,

$$s^n = h^n\left(\beta^0, x^0, ..., \beta^{n-1}, x^{n-1}\right)$$

- For a given history s^n, the pmf of β^n is given by

$$P\left(\beta^n = b_k \,|\, s^n\right) = \rho_{s^n,k}$$

- We discretize the space of possible cost histories into S values, so bias distributions are characterized by an $S \times K$ matrix ρ

The cost history

- The bias β^n is drawn randomly from a finite set $\{b_1, ..., b_K\}$ where $b_1 < ... < b_K$
- The requesting agent's bias behaviour (the pmf of β^n) depends on the past history of the game,

$$s^n = \left(\beta^{n-1}, x^{n-1}\right)$$

- For a given history s^n, the pmf of β^n is given by

$$P\left(\beta^n = b_k \,|\, s^n\right) = \rho_{s^n,k}$$

- We discretize the space of possible cost histories into S values, so bias distributions are characterized by an $S \times K$ matrix ρ

Markov decision process model

- The physical state of the problem is the game history s^n
- The decision is a quantity $x^n \in \mathscr{X}$ representing the amount under- or over-allocated relative to the request Q^n
- The transition from s^n to s^{n+1} given x^n is a function of the bias β^n, whose distribution is given by ρ

Objective function

The oversight agent must choose an allocation policy π to minimize total discounted cost:

$$\inf_{\pi} \mathbb{E}^{\pi} \sum_{n=0}^{\infty} \gamma^n C^O(s^n, X^{\pi,n}(s^n), \beta^n)$$

Learning uncertain transition probabilities

- Because the state space is discrete, we can solve the problem using value iteration:

$$V(s) = \max_{x \in \mathscr{X}} \sum_k \rho_{s,k} \left[C^O(s, x, b_k) + \gamma V\left(s'(s, x, b_k)\right) \right]$$

- We find V by iteratively solving Bellman's equation for all states s until it appears to converge
- Suppose now that the transition probabilities $\rho_{s,k}$ are unknown

Learning uncertain transition probabilities

- Let ρ_s be the vector of transition probabilities $\rho_{s,k} = P(\beta = b_k|s)$ given state s
- Assume ρ_s follows a Dirichlet distribution with parameter vector $\alpha_s^0 \geq 0$:

$$P(\rho_s = (p_1, ..., p_K)) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_{s,k}^0\right)}{\prod_{k=1}^K \Gamma\left(\alpha_{s,k}^0\right)} \prod_{k=1}^K p_k^{\alpha_{s,k}^0 - 1}, \quad p_k \geq 0, \sum_{k=1}^K p_k = 1$$

- The parameter vector measures our confidence in the likelihood of each transition:

$$\mathbb{E}^0(\rho_{s,k}) = \frac{\alpha_{s,k}^0}{\sum_{j=1}^K \alpha_{s,j}^0}$$

Learning uncertain transition probabilities

- Suppose that we make a decision and observe $\beta = b_k$ from the true underlying probability distribution
- After this observation, ρ_s is Dirichlet with parameter vector α_s^1, where

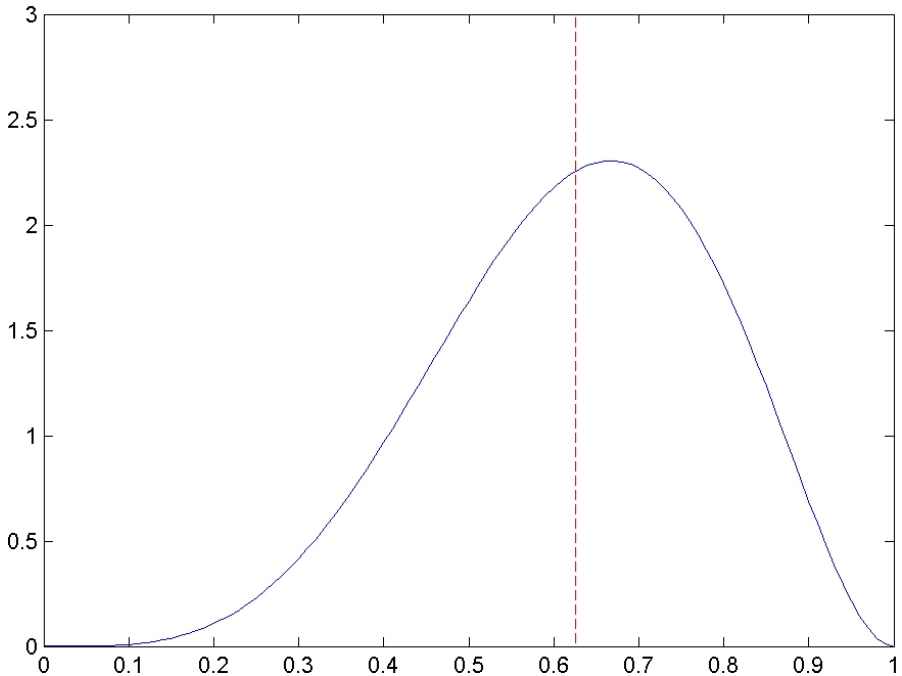
$$\begin{aligned} \alpha_{s,k}^1 &= \alpha_{s,k}^0 + 1 \\ \alpha_{s,k'}^1 &= \alpha_{s,k'}^0, \qquad k' \neq k \end{aligned}$$

- We now believe that outcome k is more likely to happen:

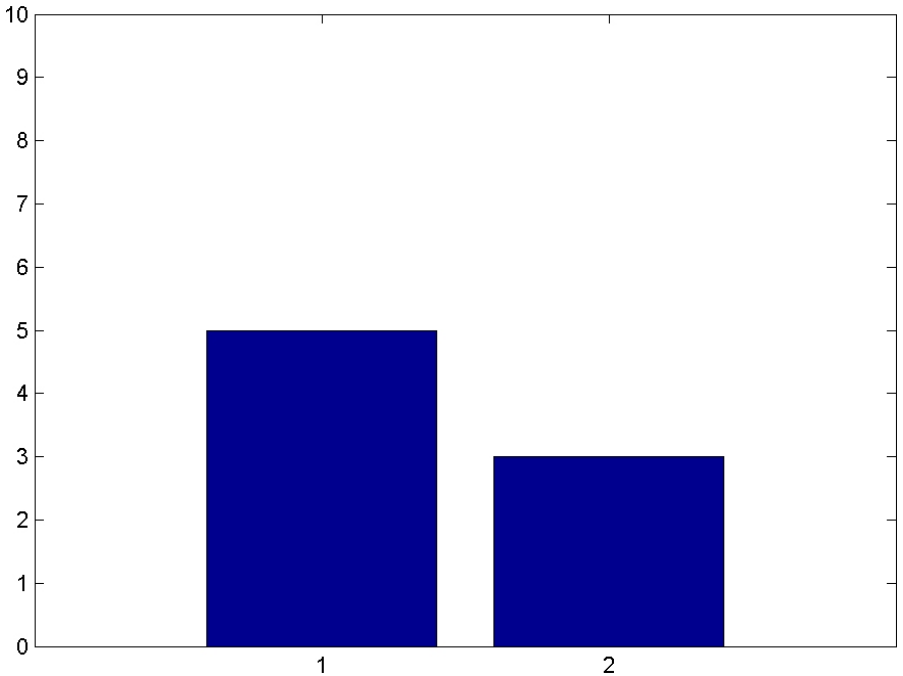
$$\mathbb{E}^1(\rho_{s,k}) = \frac{\alpha_{s,k}^1}{\sum_{j=1}^K \alpha_{s,j}^1}$$

Learning with Dirichlet distributions

Example: Two-dimensional problem with $\alpha_1^0 = 5$, $\alpha_2^0 = 3$ and outcome 1 observed



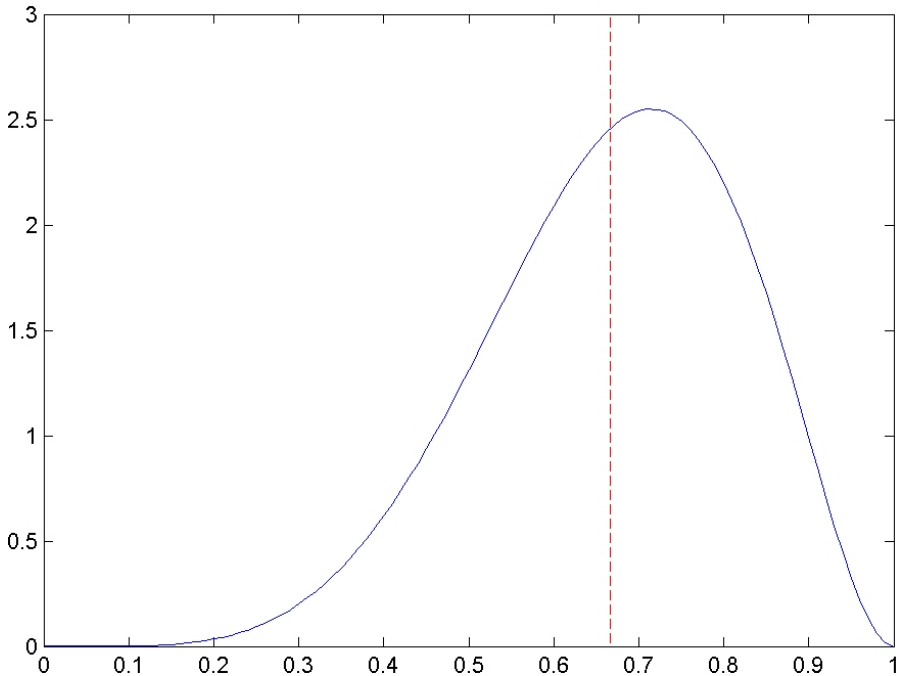
(a) Density of ρ_1 at time 0.



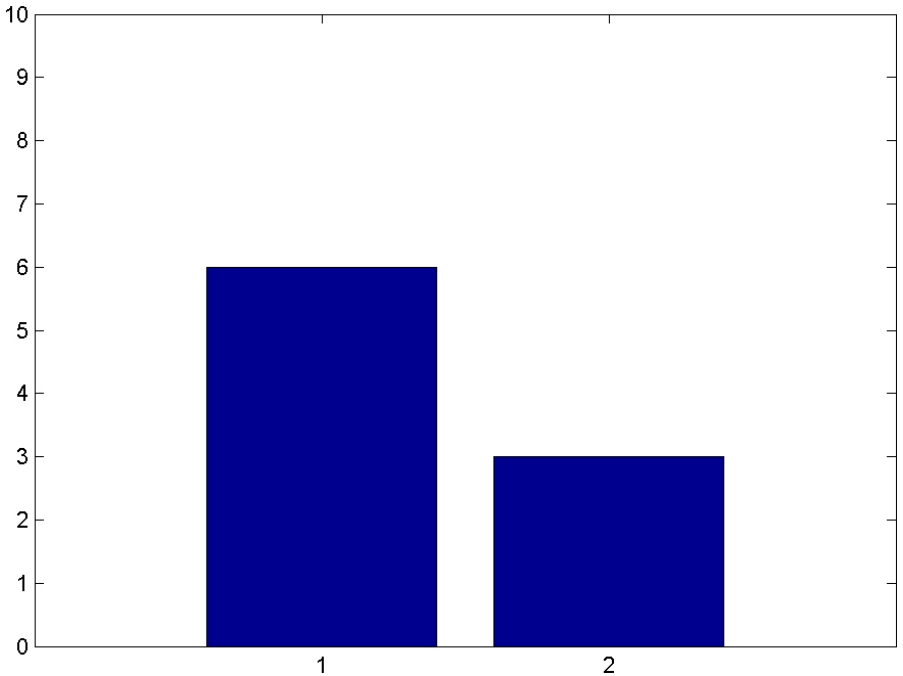
(b) Values of α_i^0, $i = 1, 2$.

Learning with Dirichlet distributions

Example: Two-dimensional problem with $\alpha_1^0 = 5$, $\alpha_2^0 = 3$ and outcome 1 observed



(c) Density of ρ_1 at time 1.



(d) Values of α_i^1, $i = 1, 2$.

Using our beliefs to make decisions

- The numbers

$$\rho_{s,k}^n = \mathbb{E}^n(\rho_{s,k}) = \frac{\alpha_{s,k}^n}{\sum_{j=1}^K \alpha_{s,k}^n}$$

  represent our beliefs about the transition probabilities at the beginning of time period n
- If we suppose that our beliefs are accurate (pure exploitation), we can solve a value-iteration problem,

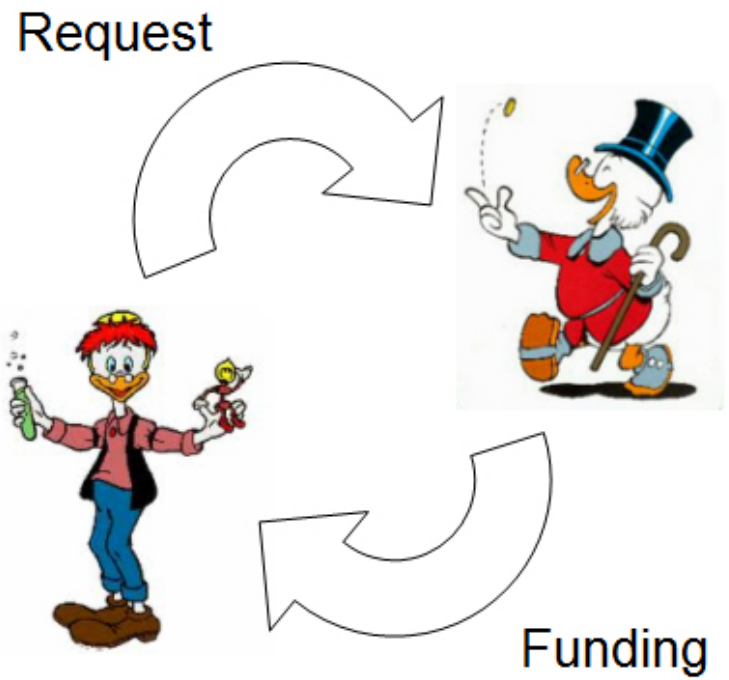
$$V(s;\alpha^n) = \max_{x\in\mathscr{X}} \sum_k \rho_{s,k}^n \left[C^O(s,x,b_k) + \gamma V\left(s'(s,x,b_k);\alpha^n\right)\right]$$

Summary of two-agent newsvendor problem

In every time step,

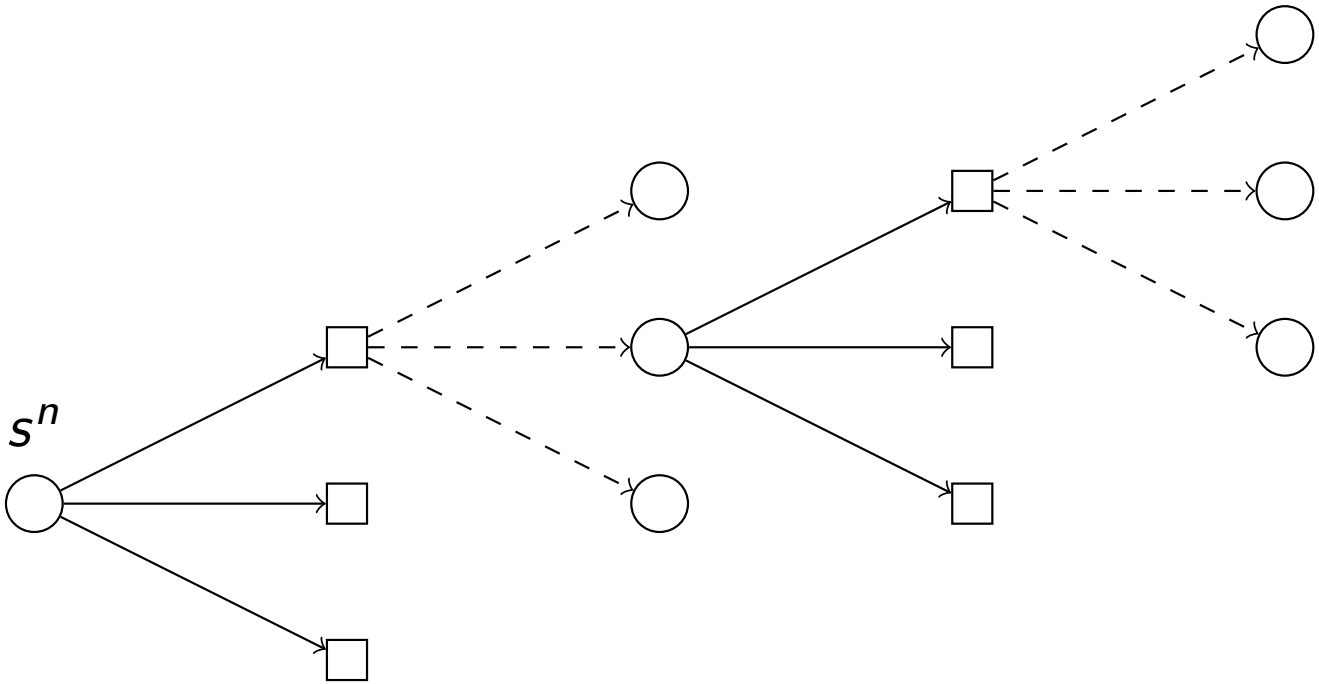
1. The requesting agent submits a request Q^n containing a bias β^n
2. The oversight agent decides how much to underfund the request (x^n represents the difference)
3. The oversight agent incurs a cost, learns β^n and uses this information to update the beliefs

The cost history is then updated and a new request is made.

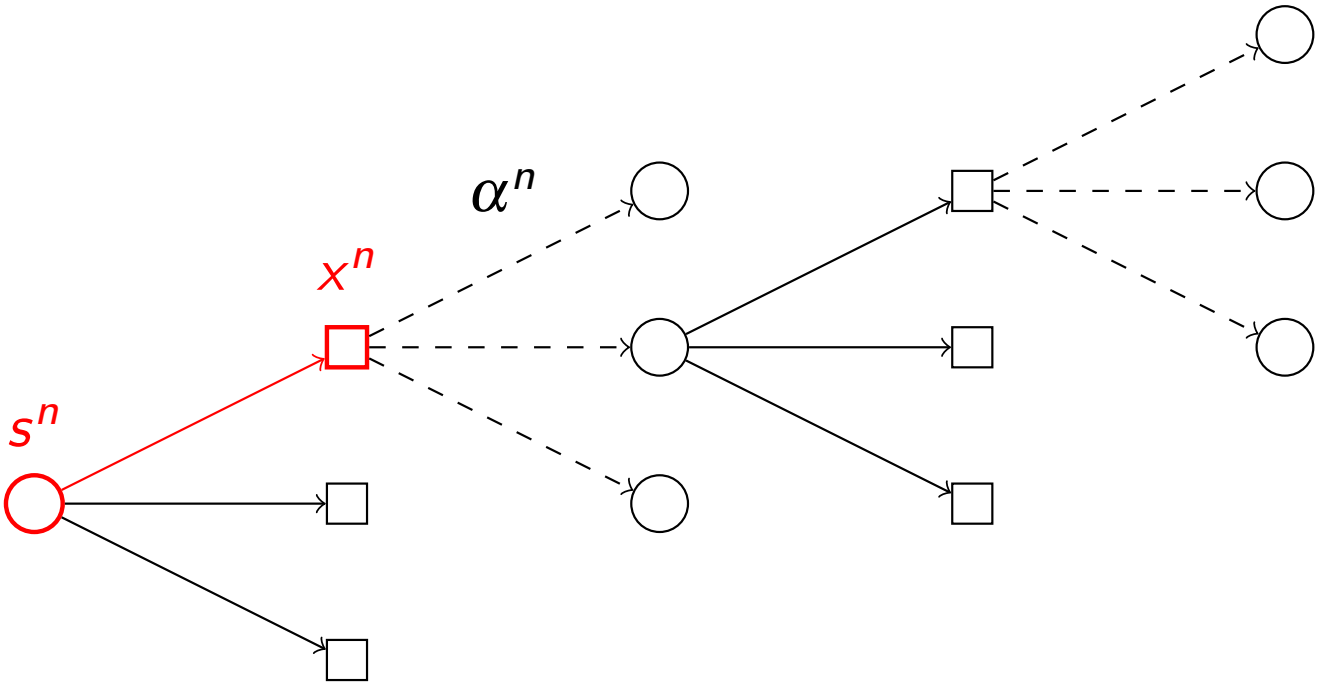


Outline

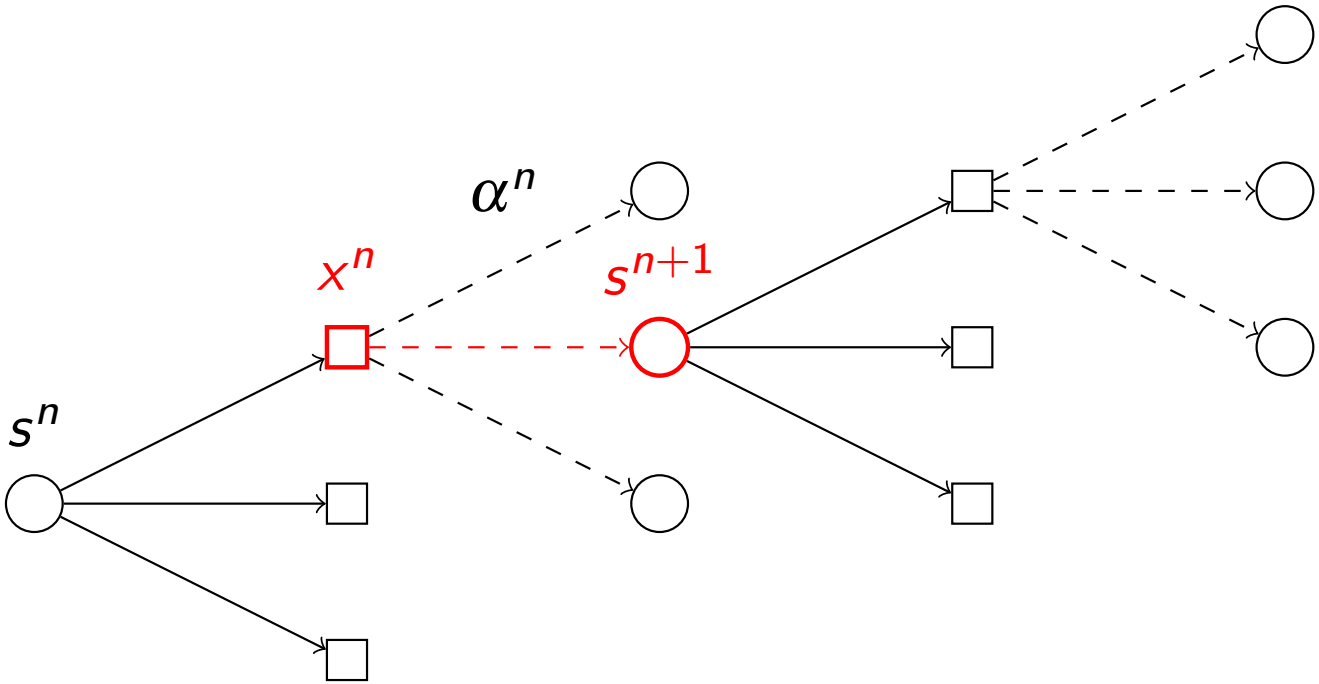
Optimal learning of transition probabilities



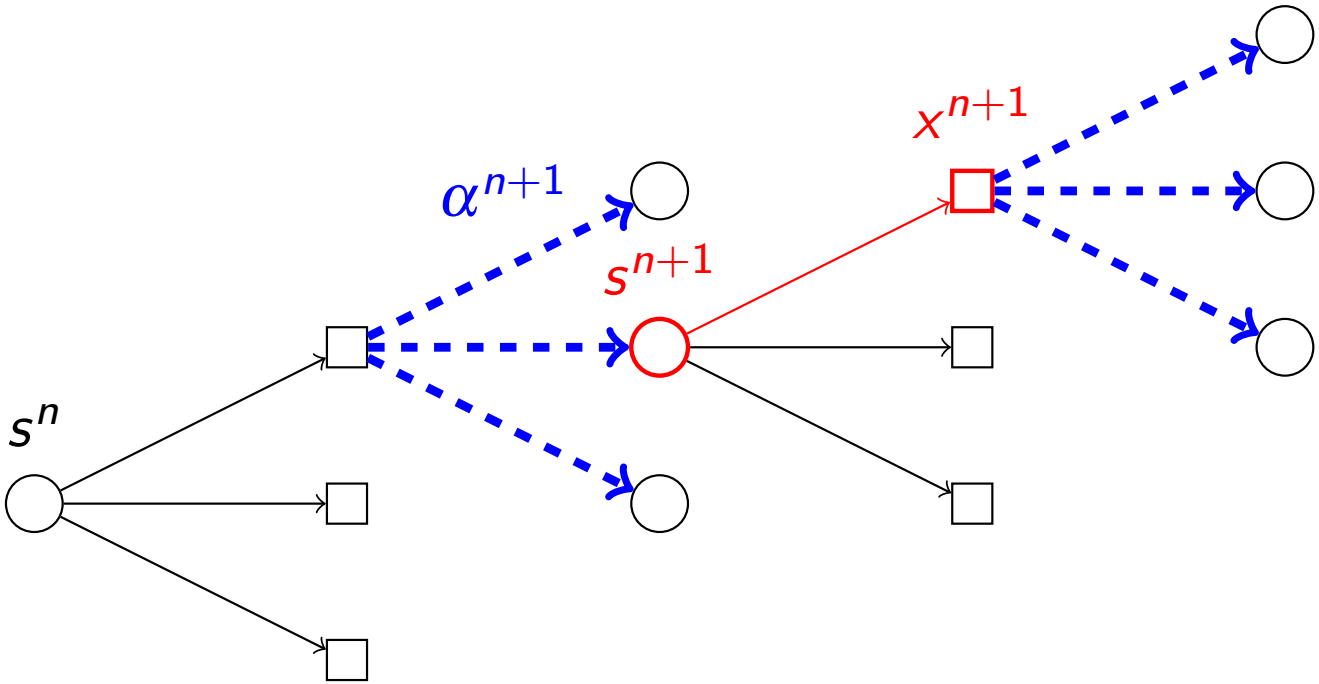
Optimal learning of transition probabilities



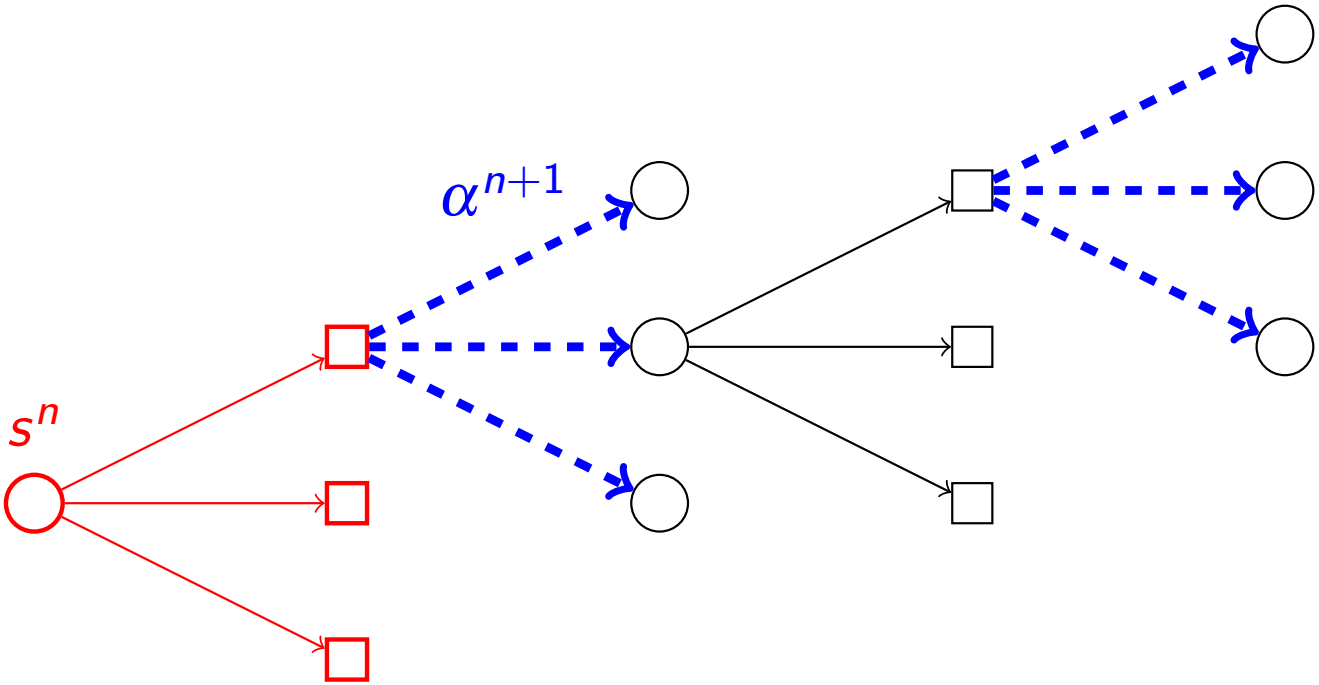
Optimal learning of transition probabilities



Optimal learning of transition probabilities



Optimal learning of transition probabilities



Can we account for the change from α^n to α^{n+1} before we make decision x^n?

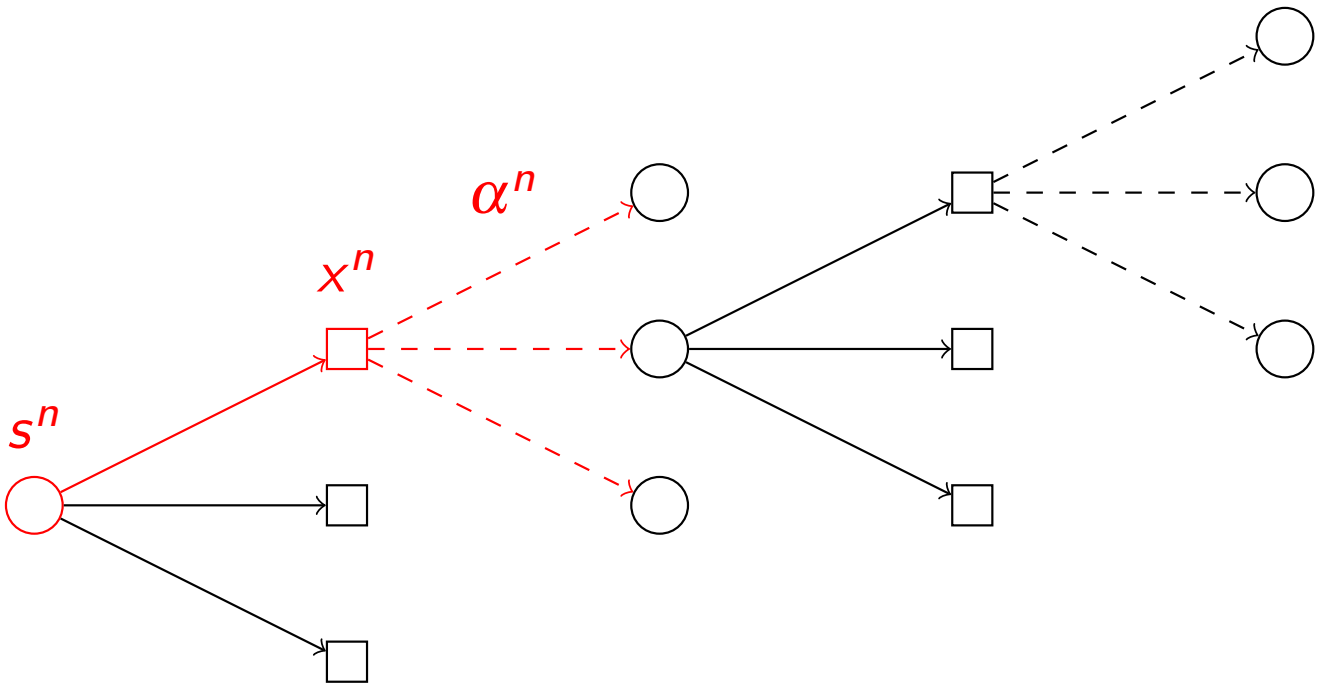
The knowledge gradient approach

- One-step look-ahead (Gupta & Miescke 1996; Ryzhov et al. 2010): if we stop learning at time $n+1$, what is the optimal decision at time n?
- Our beliefs at time $n+1$ depend on the bias observed in the nth time period:

$$\begin{aligned} \alpha_{s^n,k}^{n+1,k} &= \alpha_{s^n,k}^{n} + 1 \\ \alpha_{s^n,k'}^{n+1,k} &= \alpha_{s^n,k'}^{n}, \qquad k' \neq k \end{aligned}$$

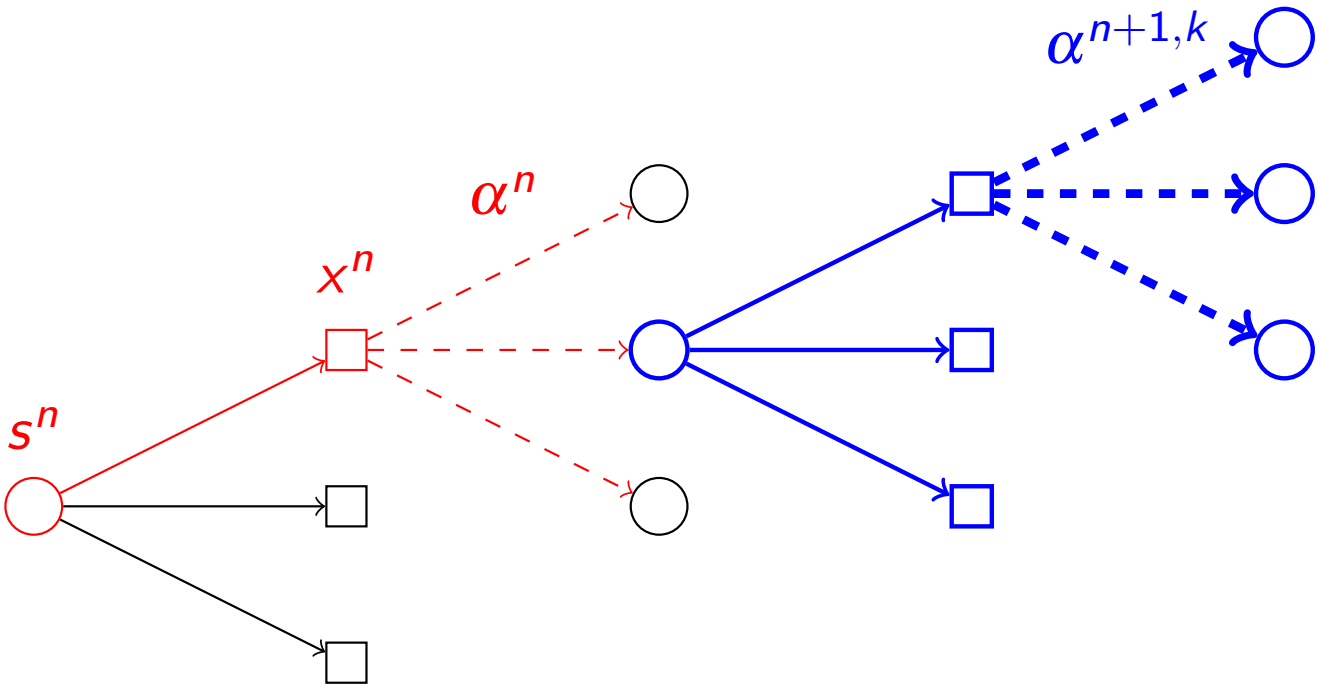
- Because the bias is discrete, we can compute a vector $\alpha_{s^n}^{n+1,k}$ of *possible* future beliefs for each outcome k

The knowledge gradient approach



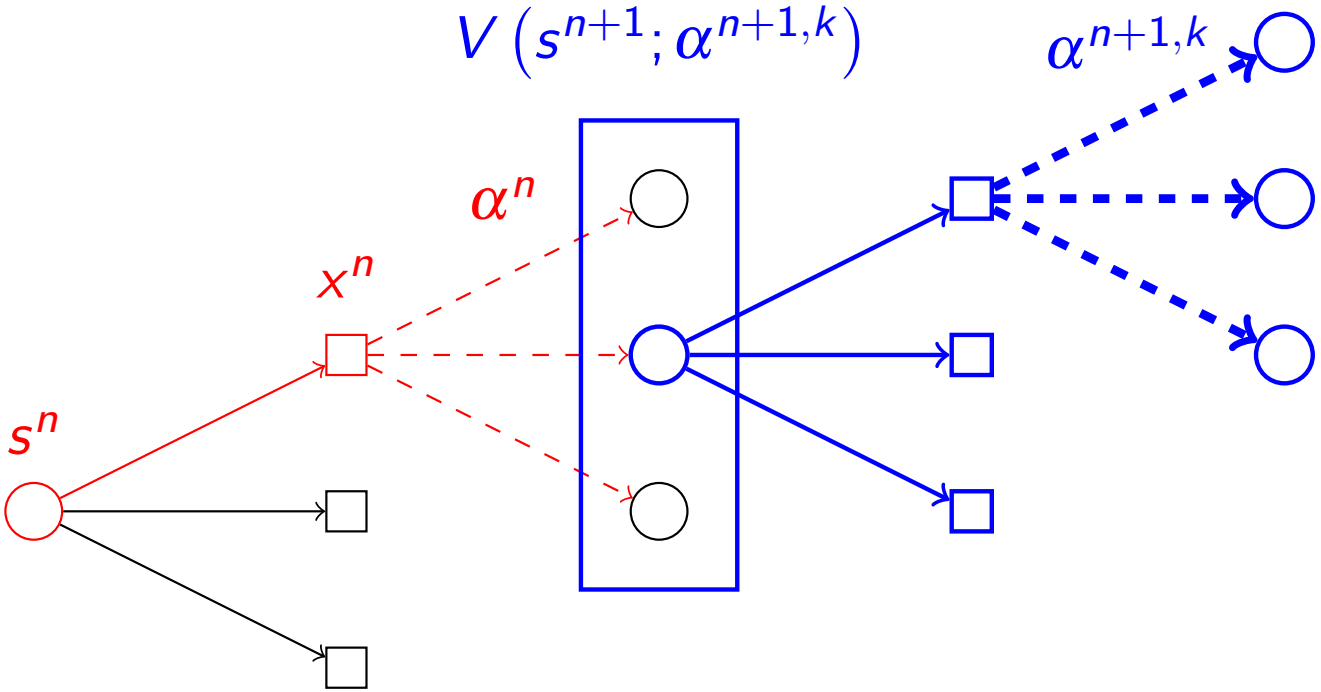
We use $\alpha^{n+1,k}$ to compute the possible downstream values, then take an expectation over α^n

The knowledge gradient approach



We use $\alpha^{n+1,k}$ to compute the possible downstream values, then take an expectation over α^n

The knowledge gradient approach



We use $\alpha^{n+1,k}$ to compute the possible downstream values, then take an expectation over α^n

Implementation of KG

- For each outcome k, solve a value-iteration problem

$$V\left(s;\alpha^{n+1,k}\right)=\max_{x\in\mathscr{X}}\sum_{k'}\rho_{s,k'}^{n+1,k}\left[C^{O}\left(s,x,b_{k'}\right)+\gamma V\left(s';\alpha^{n+1,k}\right)\right]$$

- Use the solution as the downstream value and solve for x^n:

$$x^{n}=\arg\max_{x}\sum_{k}\rho_{s^{n},k}^{n}\left[C^{O}\left(s^{n},x,b_{k}\right)+\gamma V\left(s^{n+1}\left(s^{n},x,b_{k}\right);\alpha^{n+1,k}\right)\right]$$

- Computational cost: fairly high, need to solve $K\cdot|\mathscr{X}|$ value-iteration problems in every time step

Other learning policies

- **Local bandit approximation** (Duff & Barto 1996): view the decision problem at time n as a multi-armed bandit problem, where each possible decision is an "arm" leading to a reward process
  - Computational cost: solve one value iteration problem, then $2 \cdot |\mathscr{X}|$ systems of $S \times S$ linear equations
- **Value of information exploration** (Dearden et al. 1999): generate samples from the distribution of belief, solve a value iteration problem for each one
  - Computational cost: grows with the number of samples

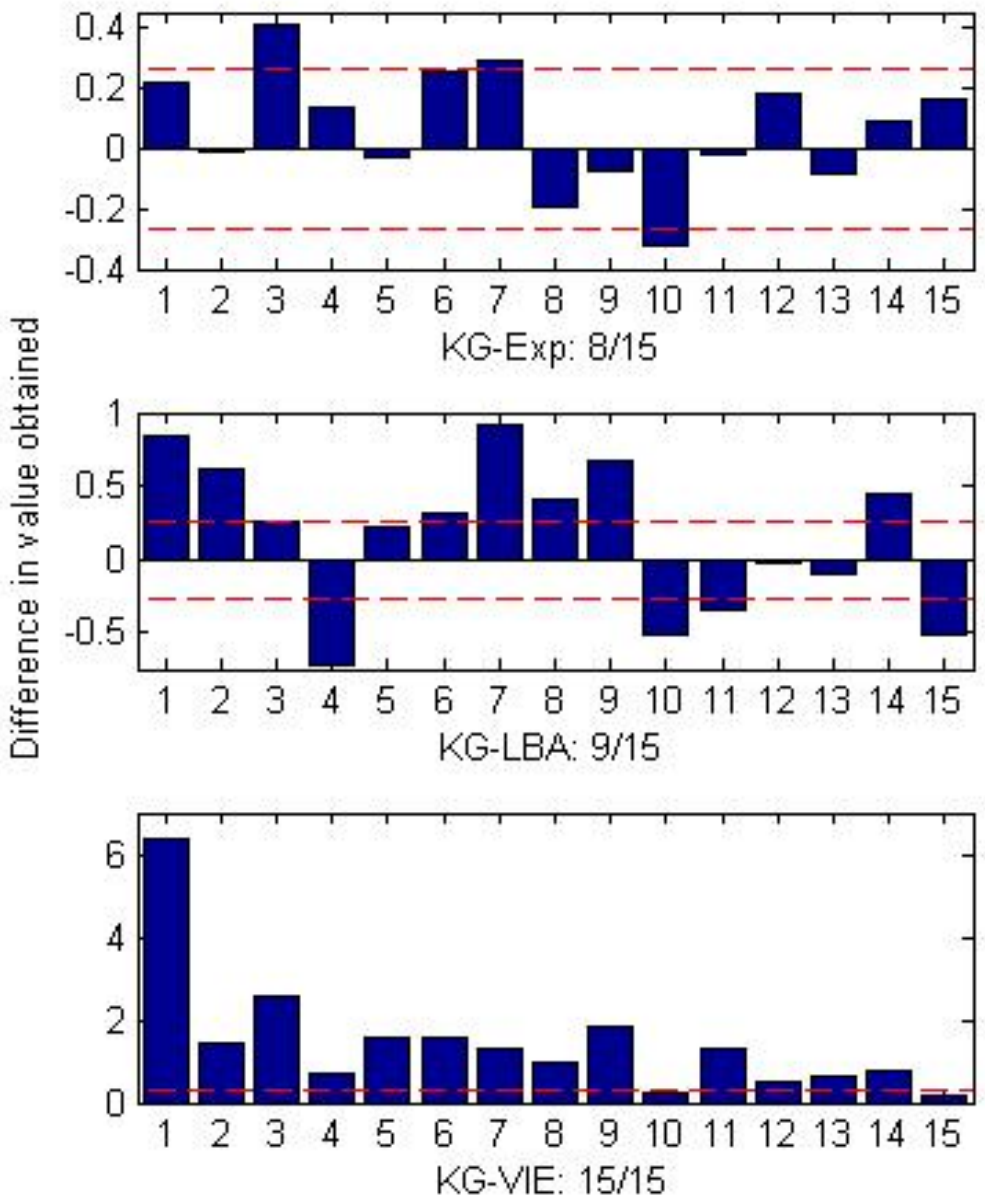
Outline

Experimental results: speeding up KG

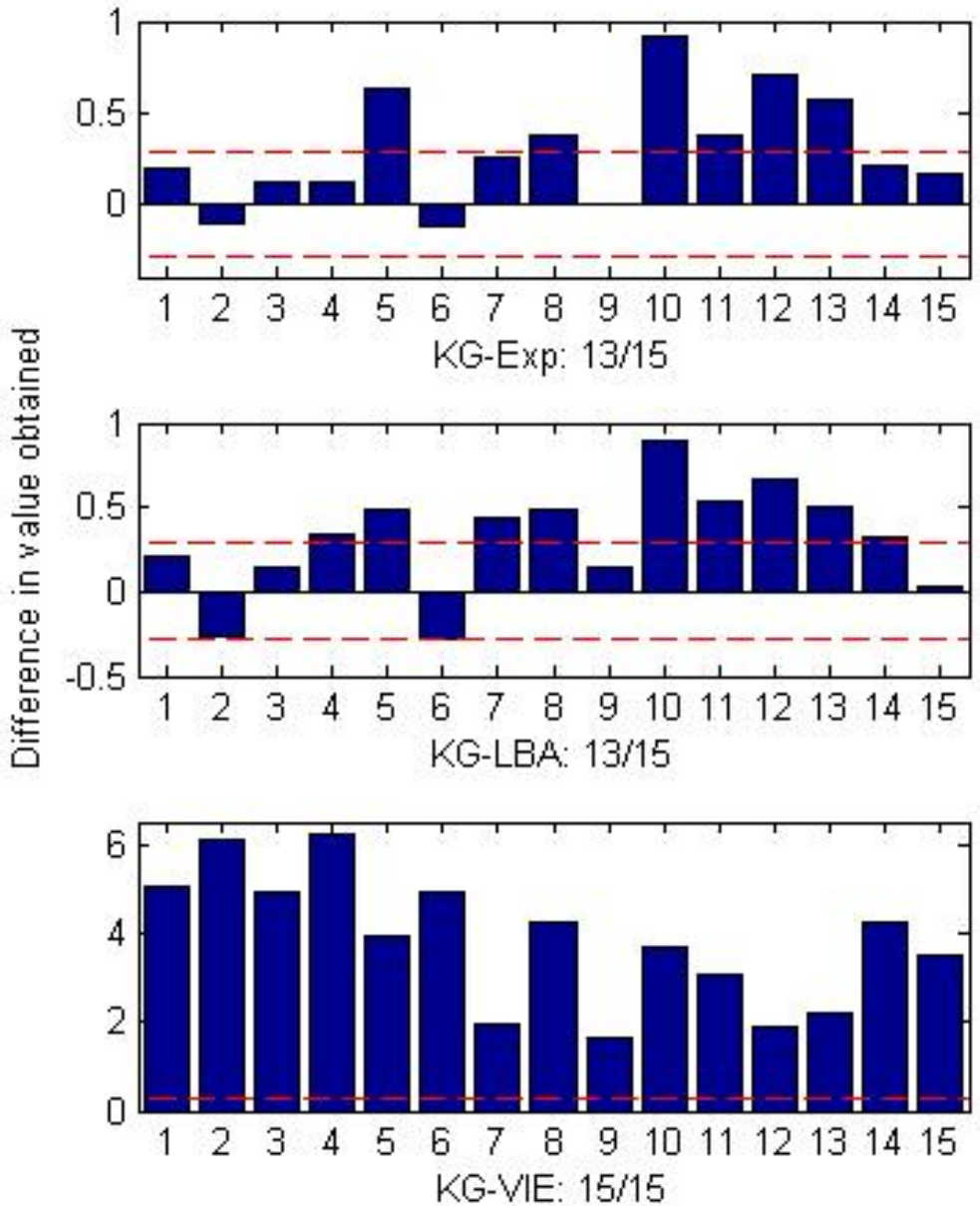
- We can speed up KG using **fast-starting**: initialize value iteration at time $n+1$ with the results from time n
- The first iteration is expensive, but subsequent iterations are much faster

Size ($K \cdot \lvert\mathscr{X}\rvert$)	$n=0$	$n=1$	$n=2$
20	0.068s	0.010s	0.019s
72	0.164s	0.057s	0.054s
272	0.613s	0.185s	0.170s
1056	2.688s	0.755s	0.610s
4160	15.432s	3.784s	1.977s
16512	2m 2s	26.131s	8.041s
65792	19m 31s	3m 49s	3m 42s
262656	4h 30m	46m 48s	43m 45s

Experimental results: accurate priors



Experimental results: equal priors



Outline

Conclusions

- We have formulated the two-agent newsvendor problem as a Markov decision process with unknown transition probabilities
- The dimension of optimal learning allows us to learn the probabilities as we go, and incorporate the value of learning into our decision-making
- A one-period look-ahead approach performs competitively, especially when we start with less prior information

References

- Arrow, K., Harris, T. & Marschak, K. (1951) "Optimal inventory policy." *Econometrica* **19**:3, 250–272.
- Dearden, R., Friedman, N. & Andre, D. (1999) "Model-based Bayesian Exploration." *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 150–159.
- Duff, M. & Barto, A. (1996) "Local bandit approximation for optimal learning problems." *Advances in Neural Processing Systems* **9**, 1019–1025.
- Gupta, S. & Miescke, K. (1996) "Bayesian look ahead one stage sampling allocation for selecting the best population." *J. on Statistical Planning and Inference* **54**:229-244.
- Puterman, M.L. (1994) *Markov decision processes (2nd ed).* John Wiley & Sons, NY.
- Ryzhov, I.O., Powell, W.B. & Frazier, P.I. (2010) "The knowledge gradient algorithm for a general class of online learning problems." Submitted for publication.