
A Multi-Agent, Policy-Gradient approach to Network Routing

Nigel Tao

Department of Computer Science, The Australian National University, Canberra 0200, Australia

NIGEL.TAO@CS.ANU.EDU.AU

Jonathan Baxter

WhizBang! Labs. 4616 Henry Street Pittsburgh, PA 15213

JBAXTER@WHIZBANG.COM

Lex Weaver

Department of Computer Science, The Australian National University, Canberra 0200, Australia

LEX.WEAVER@CS.ANU.EDU.AU

Abstract

Network routing is a distributed decision problem which naturally admits numerical performance measures, such as the average time for a packet to travel from source to destination. OLPOMDP, a policy-gradient reinforcement learning algorithm, was successfully applied to simulated network routing under a number of network models. Multiple distributed agents (routers) learned cooperative behavior without explicit inter-agent communication, and they avoided behavior which was individually desirable, but detrimental to the group's overall performance. Furthermore, shaping the reward signal by explicitly penalizing certain patterns of sub-optimal behavior was found to dramatically improve the convergence rate.

1. Introduction

Network routing is the problem of efficiently using communication paths. The standard example is Internet packet routing, but a broader view incorporates messaging in multi-processor computers, and the design of transport networks (such as vehicle traffic, airline scheduling and water distribution).

Current approaches to routing model the network as a graph with nodes representing routers and weighted edges representing links and their associated costs. Finding good approaches to routing is thus reduced to analytically solving various optimization problems for weighted graphs (Deo & Pang, 1984). Although elegant solutions to such problems have been found (Dijkstra, 1956; Bellman, 1958; Cormen et al., 1990; Tanenbaum, 1996), the algorithms can become intractable, as the models approach realism. In one example, Orda et al. (1993) consider the problem of graphs where the

edge weights can change probabilistically (as a Markov process). In this case, they conclude that the minimal expected delay routing problem is NP-hard.

Even if the problem posed by the model is tractable, unrealistic modeling may mean that the derived solution is sub-optimal. Real world networks exhibit a variety of features which are not captured in the standard weighted-graph model. Such complexities include links having limited bandwidth or capacity, limited buffering at either or both ends of channels, packet collisions, mobile hosts (i.e. changing link delays), deadlock concerns, unreliable links, non-uniform traffic (e.g. Poisson distributed, or bursty), minimum Quality-of-Service requirements, and prioritized traffic (Tanenbaum, 1996).

There is also the problem of co-ordination, whether between routers or between packets from one router. Synchronization problems can occur where two routers sharing two links simultaneously notice that one of the links is not congested and both attempt to use it, causing congestion. This problem is worsened when the two routers notice the other link is now free and both choose to use it, causing congestion again. This situation repeats, with the two routers oscillating between the two links — a phenomenon known as flapping (Tanenbaum, 1996).

In view of the potential difficulties with model-based approaches to network routing, in this paper we pursue an alternative, model-free approach in which the routing problem is treated as a multi-agent reinforcement learning problem. Each router is viewed as a single independent agent, and makes its routing decisions according to a local parameterized stochastic policy. The negative of the trip time of each packet is used as an instantaneous reward signal delivered to all agents (routers) upon arrival of the packet at its destination. Each agent uses a policy-gradient reinforcement learn-

ing algorithm to adjust the parameters of its policy in the direction of the gradient of the average reward, leading to convergence of all router parameters to a local maximum of the average reward, or equivalently to a local minimum of the average packet trip time. The policy-gradient algorithm we use is a multi-agent variant of the OLPOMDP algorithm introduced in Baxter et al. (1999), and analyzed in (Bartlett & Baxter, 2000b).

The key feature of multi-agent OLPOMDP that makes it attractive for the network routing problem is that the only non-local information each router needs is the instantaneous reward signal distributed upon a packet’s arrival at its destination. In particular, routers do not need to know the network topology or any other information about the network in order to climb the gradient of the global average reward.

We describe experiments on several different network models in which multiple distributed routers learned to successfully minimize average packet trip-time. The networks were designed such that routers had to avoid individually desirable behaviour (minimizing trip-time for the single router’s packets alone) in order to achieve co-operative behaviour that maximized the global objective. We also show that shaping the reward signal by explicitly penalizing certain patterns of sub-optimal behavior can dramatically improve the convergence rate, alleviating one of the major difficulties with the use of stochastic gradient algorithms.

1.1 Reinforcement Learning and Routing

The idea of applying reinforcement learning to network routing is not new. Boyan and Littman (1993) introduced Q-routing, an adaptation of Q-learning to network routing. In this application, the destination node is given, and the value of a state (i.e. node) is the expected time-to-arrival from that node to the destination node. Value estimates are communicated between nodes, i.e. state evaluations are updated based on values of future states.

This approach is very similar to Distance Vector routing, and the two methods have the same roots. Distance Vector routing is also known as Bellman-Ford routing (Bellman, 1958), and Q-learning has its roots in dynamic programming, also due to Bellman (1957).

Although Q-routing adapted well to rising link costs (i.e. queueing times), it has a number of weaknesses. First, it is a deterministic algorithm, in that at a particular time, there is only one link chosen for traffic from a node. If this link has limited bandwidth, then routing all traffic through it may not be an op-

timal choice. Second, Q-routing does not maintain exploratory behavior — when a link is continually valued highest, then other links are never tried. Network changes (e.g. new hosts, faster links) may not be discovered, since a sub-optimal policy can be continually followed whilst the value of a better policy is under-estimated. Third, the “Count-to-Infinity” problem may occur (Tanenbaum, 1996), when node A’s value is calculated from node B’s value, which is calculated from node A’s value. This indirect self-reference can lead to long convergence times.

Stone (2000) later applied TPOT-RL to network routing. His algorithm is explicitly multi-agent, and efficiently incorporates domain-specific features such as link utilization rates into the agents’ observation space. However, TPOT-RL learns value functions, like Q-routing, and may suffer from the first two weaknesses outlined above. It does avoid the “Count-to-Infinity” problem, by using a TD(1) style update rather than Q-routing’s simpler TD(0) approach.

2. Network Routing as a Multi-Agent, Partially Observable Markov Decision Process

Each router in a network may be viewed as an agent that receives as input packets destined for some other node in the network. The agent (router) must decide on which of the possible outgoing links to send the packet. Assuming a stochastic agent, let $\mu_u^i(y)$ denote the probability that router i chooses link u for a packet routed to destination node y . The decision problem is only partially observable, since each agent sees only a subset of the packets, and not the state of the entire network. It is a Markov process, because the next state of the network is dependent only upon the current state and the decisions of all the routers. Finally, multiple nodes makes it a multi-agent problem. Thus, network routing is a multi-agent, partially observable, Markov decision process (POMDP).

To complete the picture we need to define a performance measure for the network. A natural performance measure at the packet level is trip time, or the time taken for a packet to get from source node to destination node. Longer trip times are less desirable, so at time t we set the reward signal to be the negative sum of the trip times of all packets which arrived at time t :

$$r_t := - \sum_p \text{trip_time_of}(p) \quad (1)$$

where the sum is over all packets that arrive at their

destination at time t .¹ Note that provided the probability of each routing decision, $\mu_u^i(y)$, is non-zero, every packet will eventually arrive at its destination with probability 1.

To make the routers trainable, we parameterize each with a set of real-valued parameters θ_{yu}^i , one for each possible destination / outgoing link pair (y, u) . To keep the computations simple, we use a Gibbs distribution so that a packet arriving at router i , destined for node y , is sent on link u with probability

$$\mu_u^i(y, \theta) := \frac{e^{\theta_{yu}^i}}{\sum_{u'} e^{\theta_{yu'}^i}} \quad (2)$$

where the sum is over all outgoing links at router i .

2.1 Multi-Agent Policy-Gradient Using OLPOMDP

Our aim is to find parameter settings for all the routers that maximizes the expected long-term average reward:

$$\eta(\theta) := \mathbb{E} \left[\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T r_t \right] \quad (3)$$

where θ is the concatenation of the parameters from all routers. Provided that the system is ergodic (i.e. it converges to a unique steady state, given θ), the right-hand-side of (3) is independent of the network starting state, and converges with probability 1 over all possible reward sequences $\{r_t\}$. Note that maximizing the long-term average reward $\eta(\theta)$ is equivalent to minimizing the average packet trip-time.

We use stochastic gradient ascent to find local optima of $\eta(\theta)$. That is, at each time step t the parameters θ_t of all the routers are updated by

$$\theta_{t+1} = \theta_t + \Delta\theta_t \quad (4)$$

where the long-term average of the updates $\Delta\theta_t$ lie in the gradient direction $\nabla\eta(\theta)$. Since we are adjusting the policy parameters θ to climb the gradient of

¹This is unrealistic, in that information at a node (i.e. the trip times of all arriving packets) has to be instantaneously communicated to all other nodes. Alternatively, each node i can calculate its component of the reward signal $r_t^i := -\sum_p \text{trip_time_of}(p)$ where now the sum is only over packets destined for i . This component can be regularly broadcast via the network, and each router's reward signal is merely the sum of all components received in that time step. This means that rewards are delayed, but this does not affect a policy's *long-term* average reward, although it may affect the training algorithm in that the delay between an action and its effect on the reward signal is increased. In practice, we found that modeling the broadcast and receipt of reward components via the network does not significantly affect results (Tao, 2000).

the average reward $\eta(\theta)$, this approach is known as a *policy-gradient* algorithm.

Policy-gradient algorithms were introduced into Machine Learning by Williams (1992) with his REINFORCE algorithm (similar algorithms were described earlier in the Monte-Carlo literature by Glynn (1986); Reiman and Weiss (1986)). A practical difficulty with REINFORCE, and a number of subsequent algorithms (Marbach & Tsitsiklis, 1998; Baird & Moore, 1999; Meuleau et al., 1999; Meuleau et al., 2000) is that they estimate the gradient from a specified recurrent state, which requires perfect, rather than partial, observability (at least in that state). In addition, the variance of the gradient estimate is related to the time to return to this recurrent state, which can be large, particularly in multiple-agent environments. For example, with network routing, the time between identical configurations of the network can grow exponentially with the size of the network.

Removing the need to identify a particular recurrent state, Baxter and Bartlett (1999) introduced GPOMDP, an algorithm for generating a biased estimate of the gradient $\nabla_{\beta}\eta(\theta)$. GPOMDP takes one free parameter $\beta \in [0, 1)$ which controls both the bias and the variance of the estimates produced by the algorithm. In particular, the bias of GPOMDP is small provided $\tau_{\text{ALG}} := 1/(1 - \beta)$ is small compared to a certain mixing time τ of a Markov process associated with the POMDP. The relevant mixing time is essentially the time constant in the decay of the influence of an action on the reward signal. τ is always shorter than the recurrence time, and is often considerably so. The variance of the estimate produced by GPOMDP after T steps is proportional to τ_{ALG}/T , illustrating the bias/variance trade-off in the selection of β : large β gives small bias but large variance and vice-versa. See (Bartlett & Baxter, 2000a) for a detailed exposition.

OLPOMDP is an online adaptation of GPOMDP which applies policy parameter updates as soon as they are calculated (Baxter et al., 1999). If θ_t^i is the parameter vector of the i -th router at time t , the updated parameter vector θ_{t+1}^i is given by

$$\theta_{t+1}^i := \theta_t^i + \gamma_t \cdot r_t \cdot z_t^i \quad (5)$$

where r_t is the sum of the rewards (negative trip times) for all packets arriving at time t , γ_t are suitable step-sizes², and z_t^i is an eligibility trace of the same dimen-

²Sufficient conditions on γ_t are $\gamma_t > 0$, $\sum \gamma_t = \infty$, $\sum \gamma_t^2 < \infty$, which is satisfied, for example, by $\gamma_t = 1/t$, although to speed convergence in this paper we set γ_t to be a suitably small constant.

sionality as θ_t^i , which is updated according to

$$z_{t+1}^i := \beta \cdot z_t^i + \frac{\nabla \mu_{u_t}^i(y_t, \theta^i)}{\mu_{u_t}^i(y_t, \theta^i)} \quad (6)$$

Here y_t is the destination node of the packet at router i at time t , and u_t is the outgoing link on which it was routed. The gradient operator ∇ is with respect to the parameters of the i -th router, θ^i .

With the Gibbs parameterization in (2),

$$\frac{\nabla \mu_{u_t}^i(y_t, \theta^i)}{\mu_{u_t}^i(y_t, \theta^i)} = -[\mu_1^i(y_t, \theta^i), \dots, \mu_{u_t}^i(y_t, \theta^i) - 1, \dots, \mu_n^i(y_t, \theta^i)] \quad (7)$$

Since OLPOMDP computes a biased estimate of the gradient, it cannot be guaranteed to converge to a local optimum of the average reward. However, provided the bias is sufficiently small (i.e β is sufficiently large), OLPOMDP will converge to a region of near-zero gradient (Bartlett & Baxter, 2000b). This can be extended to the multi-agent case.

Apart from the globally distributed reward signal r_t , each router uses only local information in its parameter updates (the router’s parameters, the packet destination, and the link on which the packet is routed). Thus, multi-agent OLPOMDP is particularly attractive as an algorithm for optimizing the co-operative behaviour of all routers. For similar observations in the context of episodic tasks, see (Peshkin et al., 2000).

3. Experiments

3.1 A Simple Network

Figure 1 depicts a simple network with three nodes, and three links. The numbers on the links are the link delay; the time between when a packet is placed on the link and when it arrives at the other end. Note that the shortest path from A to C is actually via B. In this simulation, traffic was generated at a rate of one packet per time step at each of the three nodes, with the destination node chosen at random.

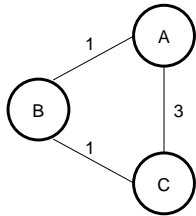


Figure 1. The triangle network.

Figure 2 shows a typical run of the OLPOMDP routing algorithm (with parameters $\beta = 0.99$ and $\gamma = 10^{-5}$) on the simple network. The top chart shows that the reward signal³ improves, on average, over time to an optimum. The bottom chart shows the probability of the agent at A routing packets destined for C on link AB. Initially, it learns to avoid that link, since the AC link (at a cost of 3 for certain) is preferable to going via B (which chooses the wrong link half the time). However, as B learns to route packets destined for C, the AB link (followed by the BC link) becomes a better option. The family of agents have learned the best collective policy, even though it requires the agent at A to rely on the agent at B.

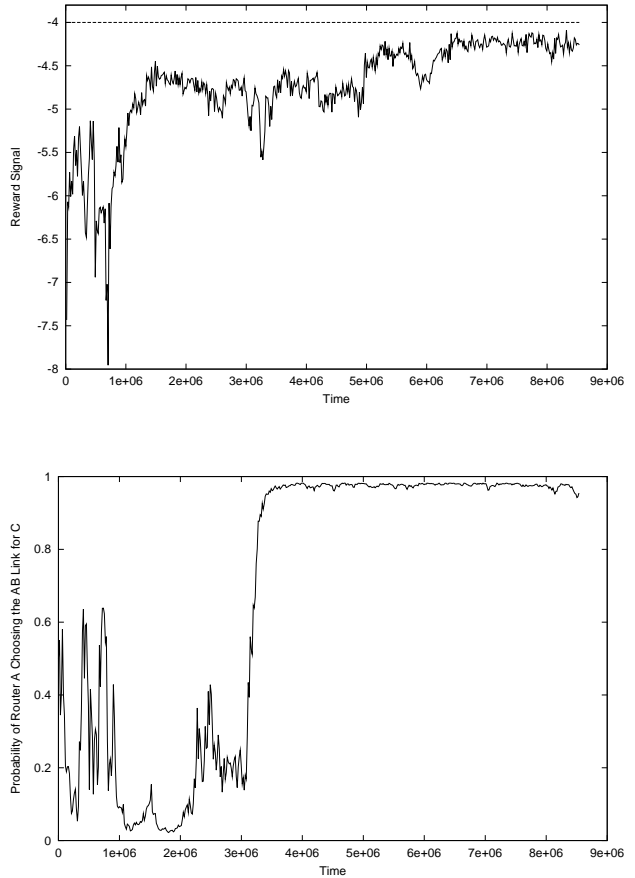


Figure 2. Reward signal r_t and probability $\mu_{AB}^A(C)$ for OLPOMDP on the triangle network.

³The chart actually shows a moving average, since the reward signal is quite volatile. Rewards are generated only when packets arrive at their destination, whilst traffic is generated probabilistically, and routed probabilistically.

3.2 Mixed Strategies

Figure 3 shows a two-node network, with two links from A to B. Unlike the previous example, these links have a limited capacity, and packets placed on a saturated link are simply dropped. To capture the undesirability of losing packets, a penalty term d is subtracted from the reward signal every time a packet is dropped.

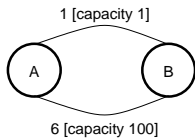


Figure 3. The contention network.

In this experiment, traffic was only generated at A (and destined for B), at a rate of two packets every time step. If the top link is chosen every time, then the average reward signal is $-1 - d$, since one packet gets through (for a trip time of 1) and one packet is dropped. If the bottom link is chosen every time, then the average reward signal is -12 , since two packets get through (for a trip time of 6 each). If a probabilistic policy is pursued, then the long-term expected average reward can be calculated based on the probabilities of each possible outcome occurring: both packets placed on the top link, both on the bottom link, or one on each link.

If the drop penalty d is set to 21, then a simple calculation shows that the optimal probability of choosing the top link is $\frac{1}{4}$. Figure 4 shows the performance of the OLPOMDP routing algorithm (with $\beta = 0.99$ and $\gamma = 10^{-7}$) on such a situation. The routing agent has learned the optimal non-deterministic policy, for a long-term average performance of $-10\frac{3}{4}$. This outperforms either deterministic strategy (-22 for the top link only, and -12 for the bottom link only).

3.3 Penalizing Cycles

The routing algorithm thus far does not include any domain knowledge. In general, no shortest path can contain a cycle⁴. This fact has been exploited by Caro and Dorigo (1998), who drop all links of a cycle out of their eligibility trace equivalent, and Stone (2000), who immediately downgrades an appropriate Q-value by some proportion of the duration of the cycle.

⁴This is because each extra link in a cycle taken incurs a further cost, but traversing a cycle has no net effect. However, this result breaks down if links chosen for one packet can affect another packet, e.g. with limited bandwidth.

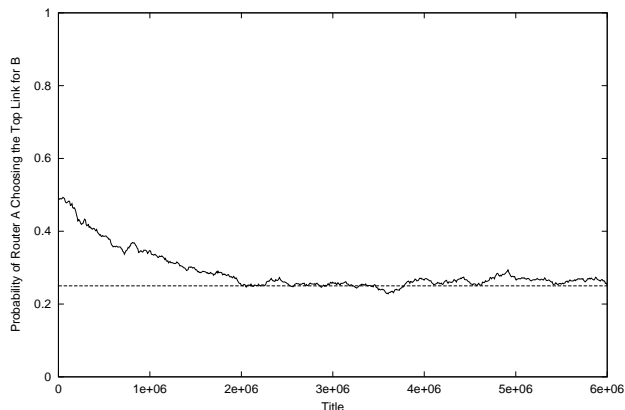
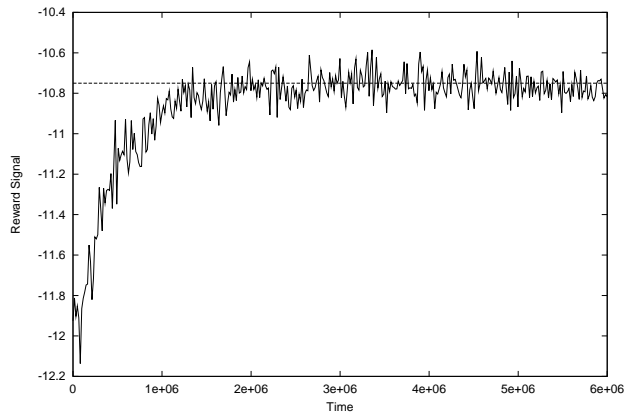


Figure 4. Reward signal r_t and probability $\mu_{top_link}^A(B)$ for OLPOMDP on the 2-node contention network.

In this paper, we incorporate an explicit global penalty term for cycles. The penalty is delivered as a negative component to the reward signal. This an example of reward shaping, where the reinforcement signal r consists of both an underlying performance measure $r^{underlying}$ and a shaping term $r^{shaping}$. The two components are simply summed:

$$r_t := r_t^{underlying} + r_t^{shaping}$$

where the underlying performance measure is as before, given by equation 1.

Note that an optimal policy (under the original performance measure) remains optimal even when cycles are penalized, since such a policy cannot contain cycles

Figure 5 depicts a simple network with six nodes with each link having delay 1 and unlimited capacity. Packets maintained a history of the last 2 nodes visited, and

signalled a cycle if it arrived at a node it had visited before. The shaping term was set to -100 multiplied by the number of cycles detected in that time step.

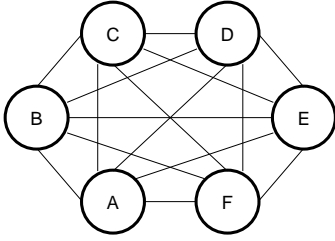


Figure 5. The complete six-node network. All links have delay 1 and unlimited capacity.

Figure 6 shows the underlying performance measure for two typical runs of the OLPOMDP algorithm (with $\beta = 0.9$, and $\gamma = 10^{-6}$). Clearly, the introduction of cycle penalties has dramatically improved the convergence rate. Note that this occurred even without explicit credit assignment. All agents were penalized even if only one agent made the wrong decision. Nonetheless, this scheme was very effective.

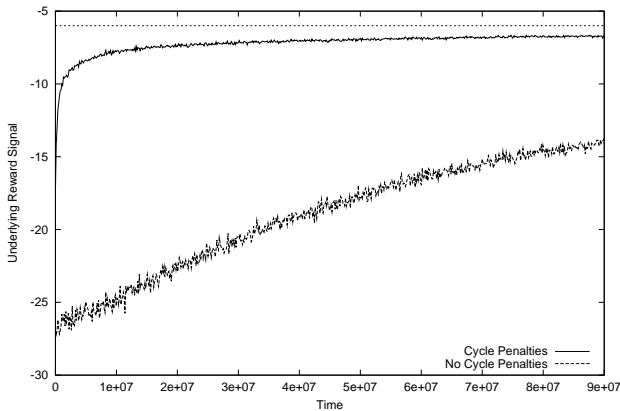


Figure 6. Improvement in the underlying reward signal, with and without cycle penalties, for OLPOMDP on the complete six-node network.

3.4 Braess' Network Paradox

Braess (1968) discovered a routing problem in which adding an extra path had the paradoxical effect of worsening overall performance. This paradox has been constructed for queueing networks (Cohen & Kelly, 1990), but we take our formulation from Tumer and Wolpert (1999), which is simpler to analyze.

The two networks are shown in figure 7. The network model here differs from the previous experiments, in that costs are incurred at nodes, rather than travelling through links. The formulae (e.g. $50 + x$) marking each node denote the cost per packet of using that node, where x is the traffic flow through that node. The motivation for this model comes from vehicular traffic, where each additional packet (i.e. car) causes a cost (i.e. congestion and waiting time) for each other packet through that node (i.e. cars at that junction).

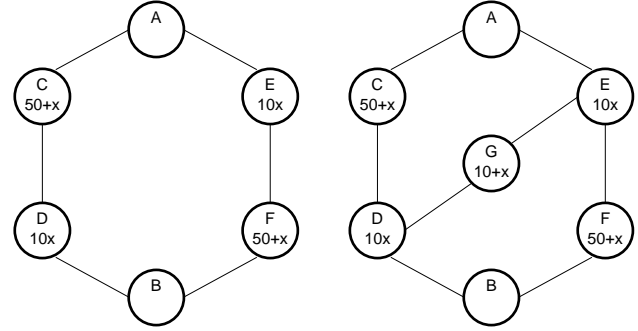


Figure 7. Braess' networks: the left hand network is $Braess_0$, and the right hand network is $Braess_1$.

Consider the left hand network, which we denote $Braess_0$. Suppose that there are six packets to be routed from A to B. If all packets go down the left hand path, nodes C and D record a flow of $x = 6$, and so the cost per packet is $(50 + 6) + (10 \times 6) = 116$. Similarly, if all packets are routed down the right hand side, the total cost is 6×116 . In fact, the lowest cost routing strategy is three packets down each path, for a cost per packet of $(50 + 3) + (10 \times 3) = 83$.

Now consider the right hand network, denoted $Braess_1$. Again, let there be six packets to be routed from A to B. Any valid routing strategy for $Braess_0$ remains a valid routing strategy for $Braess_1$, since $Braess_1$ is simply $Braess_0$ augmented by the node G and the two links EG and GD. In particular, it is possible to achieve an average cost of 83.

However, packets at E now have two paths to B: either EFB or EGDB. A greedy strategy (such as Q-routing) will not stay at three packets down each path ACDB and AEFB (and none down AEGDB). This is because the cost from E to B via F is an extra 53 at F. However, the cost from E to B via G and D is 11 at G (with flow 1), and 40 at D (with flow 4), which totals to 51. Thus, the lowest cost path from E to B for the marginal packet is actually through G. In $Braess_1$, equilibrium is achieved when two packets go down each path:

ACDB, AEFB, and AEGDB. However, in this case, the average cost per packet is 92, which is higher than the best strategy for $Braess_0$.

The OLPOMDP routing algorithm (with $\beta = 0.99$ and $\gamma = 10^{-5}$) was run on $Braess_1$, the augmented network. In this simulation, traffic was only generated at A for B (at a rate of six packets per time step), and links were one-way and downward (i.e. C has only one link to choose from, and E has two). Figure 8 shows a typical run, with the top chart depicting the probability of packets at A taking the left link, and the bottom chart depicting the probability of packets at E taking the downward link. Note that the first number hovers around 50%, whilst the second converges to 100%. This represents the optimal policy, with half of the traffic (i.e. three packets) going through ACDB, and of the remaining three packets through E, zero go through G, with three going through AEFB.

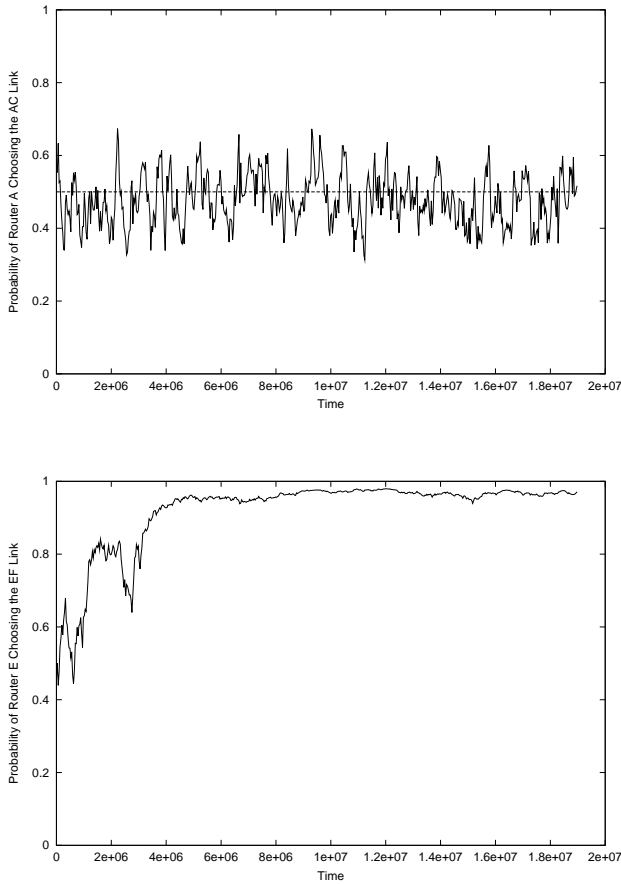


Figure 8. Probabilities $\mu_{AC}^A(B)$ and $\mu_{EF}^E(B)$ for OLPOMDP in Braess' second network.

4. Conclusion

We have shown that OLPOMDP, a policy-gradient reinforcement learning algorithm, can successfully tackle a distributed, delayed reward, partially observable, multiple agent learning problem. Optimum solutions were found, and this was achieved without an explicit system model (i.e. a set of system states and the state transition function).

The algorithm was robust, performing well under a number of network models: the link-delay-as-cost model (with and without link capacities) of §3.1-3, and the node-flow-determines-cost model of §3.4. This is in contrast to most routing algorithms that are tailored to a specific network model and may fail if their assumptions are broken.

In particular, we have shown that OLPOMDP, and policy-gradient methods in general, can

- learn co-operative behavior without explicit credit assignment, or even inter-agent communication other than the reward signal. (See §3.1)
- learn mixed (or non-deterministic) strategies, when they are preferable to any deterministic strategy. (See §3.2)
- learn at a much faster rate, when the reward signal is altered to explicitly penalize certain patterns of sub-optimal behavior. (See §3.3)
- learn to avoid behavior which is individually desirable, but detrimental to the group's performance. (See §3.4)

These strengths are not limited to the domain of network routing. Policy-gradient reinforcement learning is an effective approach with wide applicability.

Acknowledgements

Nigel Tao received financial support from the Australian National University and Burgmann College. This research was supported by the Australian Research Council.

References

- Baird, L., & Moore, A. (1999). Gradient Descent for General Reinforcement Learning. *Advances in Neural Information Processing Systems 11*. MIT Press.
- Bartlett, P. L., & Baxter, J. (2000a). Estimation and Approximation Bounds for Gradient-Based Re-

- inforcement Learning. *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*. To Appear.
- Bartlett, P. L., & Baxter, J. (2000b). Stochastic Optimization of Controlled Partially Observable Markov Decision Processes. *Proceedings of the 39th IEEE Conference on Decision and Control (CDC00)*. To Appear.
- Baxter, J., & Bartlett, P. L. (1999). *Direct Gradient-Based Reinforcement Learning: I. Gradient Estimation Algorithms* (Technical Report). Research School of Information Science and Engineering, Australian National University.
- Baxter, J., Weaver, L., & Bartlett, P. L. (1999). *Direct Gradient-Based Reinforcement Learning: II. Gradient Ascent Algorithms and Experiments* (Technical Report). Research School of Information Science and Engineering, Australian National University.
- Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, Princeton.
- Bellman, R. E. (1958). On a Routing Problem. *Quarterly of Applied Mathematics*, 16, 87–90.
- Boyan, J. A., & Littman, M. L. (1993). Packet Routing in Dynamically Changing Networks: A Reinforcement Learning Approach. *Advances in Neural Information Processing Systems*, 6, 671–678.
- Braess, D. (1968). Uber ein Paradoxon aus der Verkehrsplanung [A paradox of traffic assignment problems]. *Unternehmensforschung*, 12, 258–268.
- Caro, G. D., & Dorigo, M. (1998). Ant colonies for adaptive routing in packet-switched communications networks. *Proceedings of the Fifth International Conference on Parallel Problem Solving from Nature*. Amsterdam, Holland: Springer-Verlag.
- Cohen, J. E., & Kelly, F. P. (1990). A Paradox of Congestion in a Queueing Network. *Journal of Applied Probability*, 27, 730–734.
- Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (1990). *Introduction to Algorithms*. MIT Press, Cambridge, MA.
- Deo, N., & Pang, C. (1984). Shortest Path Algorithms: Taxonomy and Annotation. *Networks*, 14, 275–323.
- Dijkstra, E. (1956). A Note on Two Problems in Connection with Graphs. *Canadian Journal of Mathematics*, 8, 419–433.
- Glynn, P. W. (1986). Stochastic Approximation for Monte-Carlo Optimization. *Proceedings of the 1986 Winter Simulation Conference* (pp. 356–365).
- Marbach, P., & Tsitsiklis, J. N. (1998). *Simulation-Based Optimization of Markov Reward Processes* (Technical Report). MIT.
- Meuleau, N., Peshkin, L., Kaelbling, L. P., & Kim, K.-E. (2000). *Off-Policy Policy Search* (Technical Report TR-1713). MIT Artificial Intelligence Laboratory.
- Meuleau, N., Peshkin, L., Kim, K.-E., & Kaelbling, L. P. (1999). Learning Finite-State Controllers for Partially Observable Environments. *Proceedings of the Fifteenth International Conference on Uncertainty in Artificial Intelligence*.
- Orda, A., Rom, R., & Sidi, M. (1993). Minimum Delay Routing in Multisatellite Networks. *IEEE/ACM Transactions on Networking*, 1, 187–198.
- Peshkin, L., Kim, K.-E., Meuleau, N., & Kaelbling, L. P. (2000). Learning to Cooperate via Policy Search. *Proceedings of the Sixteenth International Conference on Uncertainty in Artificial Intelligence*.
- Reiman, M. I., & Weiss, A. (1986). Sensitivity Analysis via Likelihood Ratios. *Proceedings of the 1986 Winter Simulation Conference*.
- Stone, P. (2000). TPOT-RL Applied to Network Routing. *Proceedings of the Seventeenth International Conference on Machine Learning*.
- Tanenbaum, A. S. (1996). *Computer Networks*. Prentice Hall International. 3rd edition.
- Tao, N. (2000). *Walking the Path: An Application of Policy-Gradient Reinforcement Learning to Network Routing*. Unpublished Honours Thesis. Department of Computer Science, Australian National University. <http://mehta.anu.edu.au/~taonchon/thesis.ps>.
- Tumer, K., & Wolpert, D. H. (1999). *Avoiding Braess' Paradox through Collective Intelligence* (Technical Report). NASA Ames Research Center.
- Williams, R. J. (1992). Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8, 229–256.