Model-Free Reinforcement Learning as Mixture Learning

Nikos Vlassis
TU Crete, Greece

Marc Toussaint
TU Berlin, Germany

ICML-09, Montreal, Canada

Menu

- Stochastic optimal control as mixture learning.
- Model-free Reinforcement Learning via stochastic EM.
- A new mixture representation and a more efficient EM algorithm.
- Relationships to optimistic policy iteration (Sarsa(1)).
- Demonstration.

The final algorithm in the tabular case

Choose $\delta \in (0, 1)$, initialize policy π, and run until convergence:

1. Sample a batch of **random-length** trajectories using policy π, where the length T of each trajectory is drawn from $T \sim \delta^T$.

2. Estimate $Q(x, u)$ for all x, u by batch **every-visit** Monte Carlo.

3. Update policy: $\pi_{xu} \propto Q(x, u)$.

Menu

- Stochastic optimal control as mixture learning.
- Model-free Reinforcement Learning via stochastic EM.
- A new mixture representation and a more efficient EM algorithm.
- Relationships to optimistic policy iteration (Sarsa(1)).
- Demonstration.

Background

- Probabilistic inference/learning techniques for optimal control.
  (Cooper '88; Dayan & Hinton '97; Attias '03; Doucet & Tadic '04; Verma & Rao '06; **Toussaint & Storkey '06**; Todorov '08; Kappen et al. '08; Peters & Schaal '08; Hoffman et al. '08, '09; **Kober & Peters '09**)

- What it buys us: Fresh look at the problem; leverage existing algorithms for inference/learning; new analysis tools; natural extensions to structured domains/policies.

- Here we address the infinite-horizon **model-free** RL case.

Setup: Infinite-horizon MDP

- Discrete MDP with states x_t, actions u_t, and rewards $r_t \in [0, 1]$, for $t \geq 0$.
- Assume a starting distribution $x_0 \sim p(x_0)$.
- Find a **stochastic** policy π that maximizes the value function

$$V(\pi) = E\Big[\sum_{t=0}^{\infty} \gamma^t r_t \,;\pi\Big]$$

for discount factor $\gamma \in (0, 1)$.

Value function as mixture likelihood

- Toussaint & Storkey (2006) showed that V can be expressed as the likelihood function of an **infinite mixture** model.
- Key idea 1: Treat the discounting factor γ^t as a **geometric distribution** over time steps, $p(t) \propto \gamma^t$.
- Key idea 2[†]: Treat each reward r_t as the **Bernoulli parameter** of a fictitious binary random variable R, i.e., $r_t \equiv p(R = 1|t)$.

[†]Cooper (1988); Dayan & Hinton (1997).

Value function as mixture likelihood (2)

Let $\xi = (x_0, u_0, \ldots)$ be a state-action trajectory through the MDP. Then $V(\pi)$ reads:

$$V(\pi) = \sum_{t=0}^{\infty} \gamma^t E\big[r_t; \pi\big] = \sum_{t=0}^{\infty} \gamma^t \sum_{\xi} p(\xi|t;\pi)\ r_t,$$

where $p(\xi|t;\pi)$ is the distribution of t-length trajectories:

$$p(\xi|t;\pi) = p(x_0)\ p(\xi|t;\text{MDP}) \prod_{\tau=0}^{t} \pi(u_\tau|x_\tau).$$

Value function as mixture likelihood (3)

- Using $p(t) \propto \gamma^t$, the value function $V(\pi)$ reads:

$$V(\pi) \propto \sum_{t=0}^{\infty} p(t) \sum_{\xi} p(\xi|t;\pi)\; p(R=1|\xi),$$

where $p(R=1|\xi)$ is the **terminal reward** of ξ.

- $V(\pi)$ is proportional to the mixture likelihood $p(R=1;\pi)$ of a joint model $p(t,\xi,R;\pi)$.

- Here t, ξ are the 'latent' variables and R is the 'observed' variable.

Menu

- Stochastic optimal control as mixture learning.
- Model-free Reinforcement Learning via stochastic EM.
- A new mixture representation and a more efficient EM algorithm.
- Relationships to optimistic policy iteration (Sarsa(1)).
- Demonstration.

Maximize $V(\pi)$ using Stochastic EM

- E-step: Sample t_i, ξ_i from the distribution of the latent data given the observed data and the previous-step policy π_{k-1}:

$$p(t, \xi | R = 1) \propto \; p(t)\; p(\xi | t; \pi_{k-1})\; p(R = 1 | \xi).$$

- M-step: Maximize over π the expected joint log-likelihood:

$$F(\pi) = \sum_{t=0}^{\infty} p(t) \sum_{\xi} p(\xi | t; \pi_{k-1})\; p(R = 1 | \xi)\; \log p(\xi | t; \pi).$$

Forward simulation with importance sampling

- Consider a forward-simulation implementation of the E-step:

  Sample $t_i \sim p(t)$ and $\xi_i \sim p(\xi|t_i;\pi_{k-1})$.

  Assign to each ξ_i **importance weight** $w_i = p(R=1|\xi_i)$.

- This scheme uses only the terminal reward of each sampled trajectory.

- This is very inefficient; would require huge sample complexity.

- We need a new mixture representation.

Menu

- Stochastic optimal control as mixture learning.
- Model-free Reinforcement Learning via stochastic EM.
- A new mixture representation and a more efficient EM algorithm.
- Relationships to optimistic policy iteration (Sarsa(1)).
- Demonstration.

A new mixture model for $V(\pi)$

Theorem 1. For any scalar δ with $\gamma < \delta < 1$, and corresponding geometric distributions $a(t) \propto \delta^t$ and $b(t) = (\gamma/\delta)^t$, holds:

$$V(\pi) \propto \sum_{T=0}^{\infty} a(T) \sum_{t=0}^{\infty} b(t) \sum_{\xi} p(\xi|t;\pi)\ p(R=1|\xi,T),$$

where $p(R=1|\xi,T)$ is the terminal reward of ξ if $t \leq T$ and 0 otherwise.

- In this model the latent variables are T, t, ξ.

Special case: The limit $\delta \longrightarrow \gamma$

- When $\delta \longrightarrow \gamma$ the value function reads:[†]

$$V(\pi) \approx \sum_{T=0}^{\infty} p(T) \sum_{\xi} p(\xi|T;\pi) \sum_{t=0}^{T} r_t$$

- This is the **stochastic shortest path** formulation of an infinite-horizon value function (Bertsekas & Tsitsiklis, 1996).

- Our theorem can be viewed as a generalization of this formulation.

[†]Hoffman, Kueck, de Freitas, Doucet (UAI'09).

Stochastic EM in the new model

- In the new model the E-step allows more efficient sampling:

  Sample $T_i \sim a(T)$ and $\xi_i \sim p(\xi|T_i;\pi_{k-1})$, for $i = 1,\ldots,m$.

  Reuse all **sub-trajectories** ξ_{it} of ξ_i, weighted by $w_{it} = b(t)\ r_{it}$.

- In the M-step maximize over π the function

$$F(\pi) = \sum_{i=1}^{m} \frac{1}{|\xi_i|} \sum_{t=0}^{|\xi_i|} w_{\xi_{it}} \log p(\xi_{it}|t;\pi).$$

Menu

- Stochastic optimal control as mixture learning.
- Model-free Reinforcement Learning via stochastic EM.
- A new mixture representation and a more efficient EM algorithm.
- Relationships to optimistic policy iteration (Sarsa(1)).
- Demonstration.

The tabular case: Discrete x, u, multinomial π

- The probability of the t-length sub-trajectory of ξ is:

$$\log p(\xi_t | t; \pi) = \text{const.} + \sum_{xu} c_{t\xi}^{xu} \log \pi_{xu},$$

where $c_{t\xi}^{xu}$ are **counts** of occurences of (x, u) in ξ up to time t.

- The M-step function $F(\pi)$ reads:

$$F(\pi) = \sum_{xu} (\log \pi_{xu}) \sum_{i=1}^{m} \frac{1}{|\xi_i|} \sum_{t=0}^{|\xi_i|} b(t)\ r_{\xi_{it}}\ c_{t\xi_i}^{xu}.$$

Stochastic EM = Optimistic policy iteration

- Define a function $Q(x, u)$ as

$$Q(x,u) = \sum_{i=1}^{m} \frac{1}{|\xi_i|} \sum_{t=0}^{|\xi_i|} b(t)\; r_{\xi_{it}}\; c_{t\xi_i}^{xu}.$$

- This is policy evaluation with batch **every-visit** Monte Carlo!
- Maximization of $F(\pi)$ then gives:

$$\pi_{xu} \propto Q(x,u).$$

The algorithm (again)

Choose $\delta \in (0, 1)$, initialize policy π, and run until convergence:

1. Sample a batch of **random-length** trajectories using policy π, where the length T of each trajectory is drawn from $T \sim \delta^T$.

2. Estimate $Q(x, u)$ for all x, u by batch **every-visit** Monte Carlo.

3. Update policy: $\pi_{xu} \propto Q(x, u)$.

Application on POMDPs

The proposed algorithm is **non-bootstrapping** hence it is also applicable on POMDPs. Related algorithms:

Jaakkola et al. (1995): optimizes average reward; ad-hoc derivation; requires learning rate.

Perkins (2002): ad-hoc definition of $Q(x, u)$; requires learning rate.

Perkins & Precup (2003): must fully evaluate a policy before updating it.

Continuous states-actions, finite horizon

- Linear-Gaussian controller $u_t = (\theta + \varepsilon_t)\phi(x_t)$, where $\phi(\cdot)$ are basis functions, and $\varepsilon_t \sim \mathcal{N}(\varepsilon_t; 0, \sigma^2)$ is injected noise.
- The M-step gives the PoWER algorithm (Kober & Peters, 2009):

$$\theta = \theta_{old} + \frac{\left\langle \sum_{t=0}^{T} Q_{\xi t} \varepsilon_{\xi t} \right\rangle_{\xi}}{\left\langle \sum_{t=0}^{T} Q_{\xi t} \right\rangle_{\xi}}, \qquad Q_{\xi t} = \sum_{\tau=t}^{T} r_{\xi \tau}.$$

- Our framework extends PoWER to the infinite-horizon setting.

Menu

- Stochastic optimal control as mixture learning.
- Model-free Reinforcement Learning via stochastic EM.
- A new mixture representation and a more efficient EM algorithm.
- Relationships to optimistic policy iteration (Sarsa(1)).
- Demonstration.

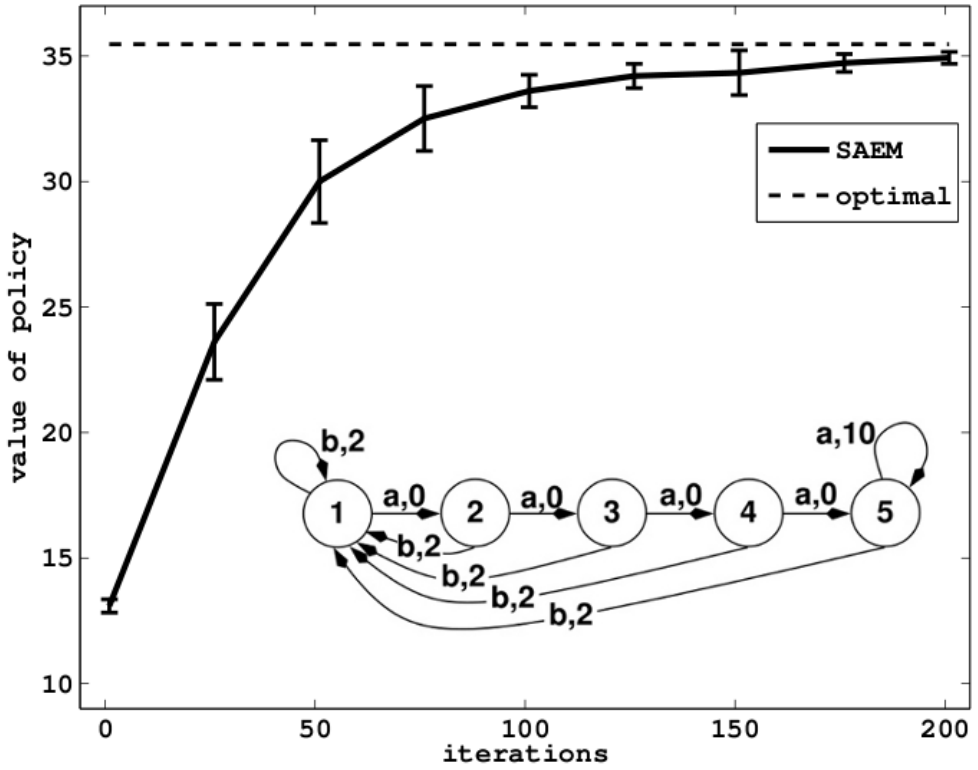
Experiment 1: The 'chain' toy MDP

Multinomial π.

Initialized π uniform.

Here we used $\delta = \gamma$ and $m = 50$.

All runs converged to the optimal policy.



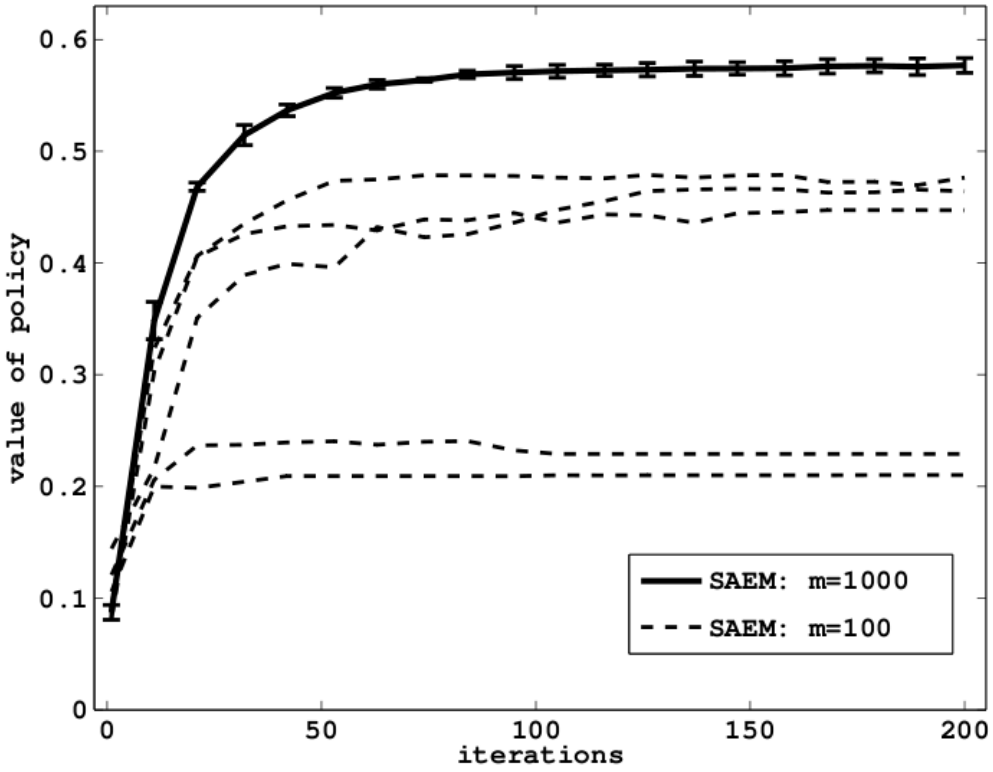
Experiment 2: The Hallway POMDP

Used stochastic **memoryless** policy π.

Initialized π uniform.

Here we used $\delta = 0.99$ (and $\gamma = 0.95$).

For $m = 1000$ Stochastic EM converged to an **optimal policy**.†



†Thanks to Chris Amato for sharing software.

Summary

- We have cast infinite-horizon model-free RL as a mixture learning problem.
- We proposed a new mixture model and an efficient EM algorithm.
- Interesting links between stochastic EM and optimistic policy iteration (Monte Carlo ES, Sarsa(1)).
- The proposed algorithm is non-bootstrapping, hence it is also applicable on POMDPs.
- It found an optimal memoryless policy in the Hallway POMDP.

[Abbeel et al., 2007] P. Abbeel, A. Coates, M. Quigley, and Ng. A. Y. An application of reinforcement learning to aerobatic helicopter flight. In B. Schölkopf, J. Platt, and T. Hoffman, editors, Advances in Neural Information Processing Systems 19, pages 1–8. MIT Press, Cambridge, MA, 2007.

[Bertsekas and Tsitsiklis, 1996] D. P. Bertsekas and J. N. Tsitsiklis. Neuro-dynamic Programming. Athena Scientific, 1996.

[Celeux and Diebolt, 1985] G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Comp. Statis. Quaterly, 2:73–82, 1985.

[Cooper, 1988] G. F. Cooper. A method for using belief networks as influence diagrams. In Proc. 4th Workshop on Uncertainty in Artificial Intelligence, pages 55–63, Minneapolis, Minnesota, USA, 1988.

[Dayan and Hinton, 1997] P. Dayan and G. E. Hinton. Using Expectation-Maximization for reinforcement learning. Neural Computation, 9(2):271–278, 1997.

[Dearden et al., 1998] R. Dearden, N. Friedman, and S. Russell. Bayesian Q-learning. In Proc. 15th National Conf. on Artificial Intelligence, pages 761–768, Madison, Wisconsin, USA, 1998.

[Delyon et al., 1999] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. The Annals of Statistics, 27:94–128, 1999.

[Dempster et al., 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of The Royal Statistical Society B, 39:1–38, 1977.

[Gordon, 1996] G. Gordon. Chattering in Sarsa(λ). Technical report, CMU Learning Lab internal report, 1996.

[Hansen, 1998] E. Hansen. Solving POMDPs by searching in policy space. In Proc. 14th Int. Conf. on Uncertainty in Artificial Intelligence, pages 211–219, Madison, Wisconsin, USA, 1998.

[Hoffman et al., 2008] M. Hoffman, A. Doucet, N. De Freitas, and A. Jasra. Bayesian policy learning with trans-dimensional MCMC. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, Advances in Neural Information Processing Systems 20, pages 665–672. MIT Press, Cambridge, MA, 2008.

[Jaakkola et al., 1995] T. Jaakkola, S. P. Singh, and M. I. Jordan. Reinforcement learning algorithm for partially observable Markov decision problems. In Advances in Neural Information Processing Systems 7, pages 345–352. MIT Press, 1995.

[Kober and Peters, 2009] J. Kober and J. Peters. Policy search for motor primitives in robotics. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, Advances in Neural Information Processing Systems 21, pages 849–856. 2009.

[Littman et al., 1995] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling. Learning policies for partially observable environments: Scaling up. In Proc. 12th Int. Conf. on Machine Learning, pages 362–370, 1995.

[Loch and Singh, 1998] J. Loch and S. P. Singh. Using eligibility traces to find the best memoryless policy in partially observable Markov decision processes. In Proc. 15th Int. Conf. on Machine Learning, pages 323–331, Madison, Wisconsin, USA, 1998.

[McCullagh and A., 1989] P. McCullagh and Nelder. J. A. Generalized Linear Models. Chapman & Hall, 2nd edition, 1989.

[Melo et al., 2008] F. S. Melo, S. P. Meyn, and M. I. Ribeiro. An analysis of reinforcement learning with function approximation. In Proc. 25th Int. Conf. on Machine Learning, pages 664–671, Helsinki, Finland, 2008.

[Neal and Hinton, 1998] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, Learning in graphical models, pages 355–368. Kluwer Academic Publishers, 1998.

[Perkins and Pendrith, 2002] T. J. Perkins and M. D. Pendrith. On the existence of fixed points for Q-learning and Sarsa in partially observable domains. In Proc. 19th Int. Conf. on Machine Learning, pages 490–497, 2002.

[Perkins and Precup, 2003] T. J. Perkins and D. Precup. A convergent form of approximate policy iteration. In S. Thrun S. Becker and K. Obermayer, editors, Advances in Neural Information Processing Systems 15, pages 1595–1602. MIT Press, Cambridge, MA, 2003.

[Peters and Schaal, 2008] J. Peters and S. Schaal. Learning to control in operational space. International Journal of Robotics Research, 27:197–212, 2008.

[Pineau et al., 2006] J. Pineau, G. Gordon, and S. Thrun. Anytime point-based approximations for large POMDPs. Journal of Artificial Intelligence Research, 27:335–380, 2006.

[Poupart, 2005] P. Poupart. Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes. PhD thesis, Dept. of Computer Science, University of Toronto, 2005.

[Shani et al., 2007] G. Shani, R. I. Brafman, and S. E. Shimony. Forward search value iteration for POMDPs. In In Int. Joint Conf. on Artificial Intelligence, pages 2619–2624, 2007.

[Spaan and Vlassis, 2005] M. T. J. Spaan and N. Vlassis. Perseus: Randomized point-based value iteration for POMDPs. Journal of Artificial Intelligence Research, 24:195–220, 2005.

[Sutton and Barto, 1998] R. S. Sutton and A. G. Barto. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA, 1998.

[Toussaint and Storkey, 2006] M. Toussaint and A. Storkey. Probabilistic inference for solving discrete and continuous state Markov decision processes. In Proc. 23rd Int. Conf. on Machine Learning, pages 945–952, Pittsburgh, Pennsylvania, USA, 2006.

[Tsitsiklis, 2002] J. N. Tsitsiklis. On the convergence of optimistic policy iteration. Journal of Machine Learning Research, 3:59–72, 2002.

[Wei and Tanner, 1990] G. Wei and M. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. J. Amer. Statist. Assocation, 85:699–704, 1990.