Optimism in Sequential Decision-Making under Uncertainty

Peter Bartlett
Department of Statistics and Division of Computer Science
UC Berkeley

Joint work with
Ambuj Tewari.

slides at http://www.stat.berkeley.edu/~bartlett

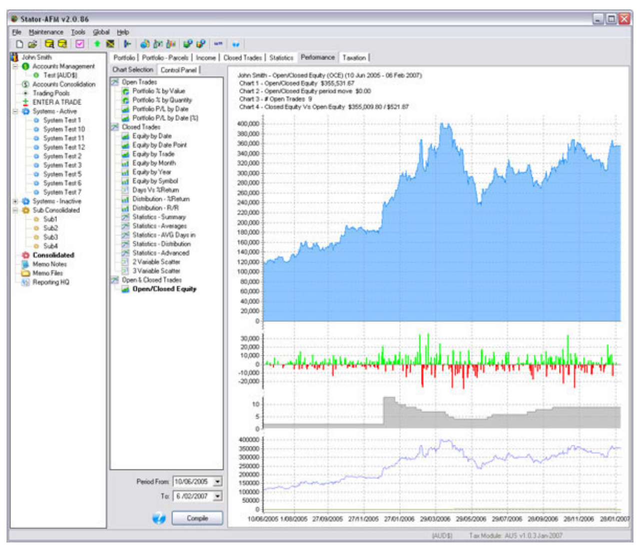
Sequential Decision-Making under Uncertainty

Robot navigation

Chess

Sequential Decision-Making under Uncertainty

Portfolio optimization



Dynamic treatment regimes

Sequential. Statistical.

Controlling a Markov Decision Process

$$
\begin{aligned}
S, A &= \text{set of states, set of actions} \\
p_s(a) &= \text{transition distribution} && \text{(unknown)} \\
r(s,a) &= \text{reward} && \text{(unknown)} \\
V_T^{\pi}(s_0) &= \mathbb{E} \sum_{t=0}^{T-1} r(s_t, a_t) && \text{total reward of policy } \pi : S \to A \\
R_T^{\pi}(s_0) &= \sup_{\tilde{\pi}} V_T^{\tilde{\pi}}(s_0) - V_T^{\pi}(s_0) && \text{regret}
\end{aligned}
$$

Aim: Minimize regret

Controlling a Markov Decision Process

$$
\begin{aligned}
S, A &= \text{set of states, set of actions} \\
p_s(a) &= \text{transition distribution} \\
r(s,a) &= \text{reward}
\end{aligned}
$$

$$\lambda + h(s) = \max_a \left(r(s,a) + \langle p_s(a), h \rangle \right) \qquad \text{optimality equations}$$

(linear program)

Minimizing regret: Exploration versus Exploitation

The Exploration/Exploitation Trade-off

How do we balance choosing actions that facilitate learning about the MDP (*exploring*, to gain more knowledge) with choosing actions that maximize average reward (*exploiting* knowledge we've gained)?

Approaches to Balancing Exploration and Exploitation

- Explicitly explore unknown territory (take the least tried actions) until we have an accurate model of the MDP. e.g., Kearns and Singh, 1998
- Implicitly explore: Be as *optimistic* about what we don't know as the data allows. e.g., Burnetas and Katehakis, 1997; Auer, 2003

Algorithm: Optimistic Linear Programming

1. Model MDP (excluding “undersampled” actions).
2. Compute solution $(\hat{h}_t, \hat{\lambda}_t)$ to optimality equations

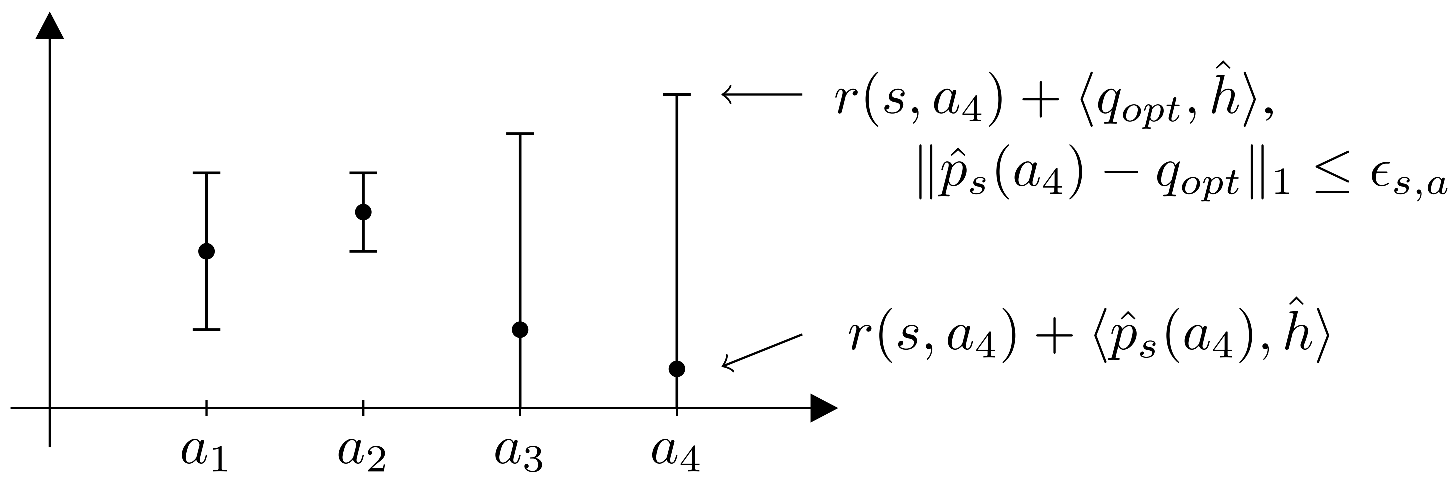
$$\lambda + h(s) = \max_a \left(r(s,a) + \langle p_s(a), h\rangle\right).$$

3. Choose action a_t to maximize the optimistic reward:

$$U(s_t,a) = \sup\left\{r(s_t,a) + \langle q, \hat{h}_t\rangle : \|\hat{p}_{s_t}(a) - q\|_1 \le \epsilon_{s_t,a,t}\right\},$$

where $\epsilon_{s,a,t}$ determines the size of a confidence set, which depends on how frequently (s,a) has been visited.

Optimistic Linear Programming



- Optimistic about the outcomes of actions, but not unreasonably so.
- Computation involves solving linear programs.

Regret Bound

For the optimistic linear programming approach, the regret grows logarithmically with time T:

$$\limsup_{T\to\infty} \frac{R_T(s_0)}{\log T} \leq \frac{|S||A|\tau^2}{\Phi},$$

where
τ is a hitting time of the MDP under the optimal policy, and
Φ measures the gap between optimal and suboptimal actions.

The rate ($\log T$) is optimal.

Confessions and Open Problems

- Regret rate hides large transient terms.
- Dependence on $|S|$ is problematic in applications.
- Optimality relative to a restricted class of policies?

Other areas of interest

- Prediction with high-dimensional data.
  - Classification and regression with ℓ_1 regularization.
  - Structured prediction: e.g., sequence classification, parsing.
  - Transfer learning.
- Prediction in adversarial settings.
  - Spam detection, portfolio optimization, web search.
  - Performance of statistical methods in these settings.