Learning Theory of Optimal Decision Making
Part II: Batch Learning in Markovian Decision Processes

Csaba Szepesvári

1Department of Computing Science
University of Alberta

Machine Learning Summer School, Ile de Re, France, 2008
with thanks to: RLAI group, SZTAKI group, Jean-Yves Audibert, Remi Munos
OUTLINE

1 HIGH LEVEL OVERVIEW OF THE TALKS

2 MOTIVATION
   - What is it?
   - Why should we care?
   - The challenge

3 MARKOVIAN DECISION PROBLEMS
   - Basics
   - Dynamic programming

4 REGRESSION

5 BATCH RL

6 CONCLUSION
High level overview of the talks

- **Day 1**: Online learning in stochastic environments
- **Day 2**: Batch learning in Markovian Decision Processes
- **Day 3**: Online learning in adversarial environments
Union bound: $P(A \cup B) \leq P(A) + P(B)$

Inclusion: $A \subset B \Rightarrow P(A) \leq P(B)$

Inversion: If $P(X > t) \leq F(t)$ holds for all $t$ then for any $0 < \delta < 1$, w.p. $1 - \delta$, $X \leq F^{-1}(\delta)$

Expectation: If $X \geq 0$, $E[X] = \int_{0}^{\infty} P(X \geq t) \, dt$

Jensen: If $f$ is convex then $f(E[X]) \leq E[f(X)]$
Linearity: \( \mathbb{E} [X + Y] = \mathbb{E} [X] + \mathbb{E} [Y] \)

Law of total expectation:
- \( \mathbb{E} [Z] = \sum_x \mathbb{E} [Z|X = x] \mathbb{P} (X = x) \)
- \( \mathbb{E} [Z|U = u] = \sum_x \mathbb{E} [Z|U = u, X = x] \mathbb{P} (X = x|U = x) \)
- \sim \) tower rule: \( \mathbb{E} [Z|Y] \overset{\text{def}}{=} f(Y) \), where \( f(y) = \mathbb{E} [Z|Y = y] \).
  Then

\[
\mathbb{E} [Z] = \mathbb{E} [\mathbb{E} [Z|X]], \quad \mathbb{E} [Z|U] = \mathbb{E} [\mathbb{E} [Z|U, X]|U].
\]

A corollary of the Markov property: Let \( X_1, X_2, \ldots \) be a Markov process. Then

\[
\mathbb{E} [f(X_1, X_2, \ldots)|X_1 = y, X_0 = x] = \mathbb{E} [f(X_1, X_2, \ldots)|X_1 = y]
\]
**What is it?**

**Protocol of Learning**

**Concepts:** Experimenter, Learner, Environment

states, actions, rewards

**One-shot learning:**

1. Experimenter generates training data
   \[ D = \{(X_1, A_1, X'_1, R_1), \ldots, (X_n, A_n, X'_n, R_n)\} \]
   by following some policy \( \pi_b \) in the Environment

2. Learner computes a policy \( \pi \) based on \( D \)

3. Policy is implemented in the Environment
   (it’s performance \( V_\pi \) is compared with that of \( \pi_b \))

**Goal:** maximize \( V_\pi \rightarrow \max \)
Why should we care?

- Batch – a frequent situation
- learning – you know!
- in Markovian Decision Processes – convenience, ..
THE CHALLENGE

- State space $\mathcal{X}$ is ..
  - large
  - infinite
  - continuous
- Action space $\mathcal{A}$ is ..
  - large
  - infinite
  - continuous
- Model based approaches require
  - efficient learners (Markov kernel?)
  - efficient planners
- Direct approach?

NOTE

In this talk $\mathcal{A}$ is finite.
Markovian Decision Problems

\((\mathcal{X}, \mathcal{A}, p, r, \gamma)\)

- \(\mathcal{X}\) – set of states
- \(\mathcal{A}\) – set of actions
- \(p\) – transition kernel
  \(p(\cdot | x, a)\) – next state distribution
  \(p(y | x, a)\) – prob. of \(y\) after taking \(a\) in state \(x\)
- \(r\) – reward function
  \(r(x, a, y)\), or \(r(x, a)\), or \(r(x)\)
- \(0 < \gamma \leq 1\) – discount factor
THE PROCESS VIEW

$(X_t, A_t, R_t)$ – controlled Markov process

- $X_t \in \mathcal{X}$ – state at time $t$
- $A_t \in \mathcal{A}$ – action at time $t$
- $R_t \in \mathbb{R}$ – reward at time $t$

Laws:

- $X_{t+1} \sim p(\cdot | X_t, A_t)$
- $A_t \sim \pi(\cdot | X_t, A_{t-1}, R_{t-1}, \ldots, A_1, R_1, X_0)$
- $\pi$ – policy, mapping histories to $M(\mathcal{A})$
- $R_t = r(X_t, A_t, X_{t+1}) + W_t$
- $W_t$ – reward noise (can depend on transition)
π, π, π, π, π, π...
The control problem

- Value functions:

\[ V_\pi(x) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t R_t \mid X_0 = x \right], \quad x \in \mathcal{X} \]

- Optimal value function:

\[ V^*(x) = \sup_\pi V_\pi(x), \quad x \in \mathcal{X} \]

- Optimal policy \( \pi^* \):

\[ V_{\pi^*}(x) = V^*(x), \quad x \in \mathcal{X} \]
Applications of MDPs

Operations research, econometrics control, statistics, games, AI, ...

- Optimal investments
- Replacement problems
- Option pricing
- Logistics, inventory management
- Active vision
- Production scheduling
- Dialogue control
- Bioreactor control
- Robotics (e.g., Robocup Soccer)
- Driving
- Real-time load balancing
- Design of experiments (Medical tests)
Bellman Operators

Definition (Supremum norm)
\[ \| V \|_\infty = \max_{x \in X} |V(x)| \] (or \[ \| V \|_\infty = \sup_{x \in X} |V(x)| \] in infinite spaces).
We let \( B(X) = (X, \| \cdot \|_\infty) \).

- Let \( \pi : X \to A \) be a stationary policy
- \( B(X) = \{ V : X \to \mathbb{R} | \| V \|_\infty < +\infty \} \) – value functions
- \( T_\pi : B(X) \to B(X) \) – Bellman operator of \( \pi \):
  \[ (T_\pi V)(x) = \sum_{y \in X} p(y|x, \pi(x)) \{ r(x, \pi(x), y) + \gamma V(y) \} \]

Theorem
\( V_\pi \) is the fixed point of \( T_\pi \)
\[ T_\pi V_\pi = V_\pi \]
and is unique.
**Bellman Operators II**

**Note**

\[ T_\pi V_\pi = V_\pi \] is a linear system of equations!

E.g., \( \mathcal{X} = \{1, 2, \ldots, n\} \)

- \( P_\pi \in \mathbb{R}^{n \times n} : (P_\pi)_{ij} = p(j|i, \pi(i)) \)
- \( r_\pi \in \mathbb{R}^n : (r_\pi)_i = \sum_j p(j|i, \pi(i))r(i, \pi(i), j) \)

Let \( V_\pi \in \mathbb{R}^n \). Then \( T_\pi V_\pi \in \mathbb{R}^n \), \( (T_\pi V_\pi)_i \overset{\text{def}}{=} (T_\pi V_\pi)(i) \),

\[
(T_\pi V_\pi)_i = \sum_{j=1}^{n} p(j|i, \pi(i)) \{ r(i, \pi(i), j) + \gamma V_\pi(j) \} \\
= (r_\pi)_i + \gamma (P_\pi V_\pi)_i, \quad \text{hence} \\
(T_\pi V_\pi) = V_\pi
\]

\[ r_\pi + \gamma P_\pi V_\pi = V_\pi \]

Also: \( V_\pi = (I - \gamma P_\pi)^{-1} r_\pi \).
Define $R_{t:∞} = \sum_{s=0}^{∞} \gamma^s R_{t+s}$. Then $R_{0:∞} = R_0 + \gamma R_{1:∞}$.

$V_\pi(x) = \mathbb{E}_\pi [R_{0:∞} \mid X_0 = x]$

$= \sum_{y \in\mathcal{X}} \mathbb{P}(X_1 = y \mid x, \pi(x)) \mathbb{E}_\pi [R_{0:∞} \mid X_0 = x, X_1 = y]$

$= \sum_{y \in\mathcal{X}} p(y \mid x, \pi(x)) \mathbb{E}_\pi [R_{0:∞} + \gamma R_{1:∞} \mid X_0 = x, X_1 = y]$

$= \sum_{y \in\mathcal{X}} p(y \mid x, \pi(x)) \{ r(x, \pi(x), y) + \gamma V_\pi(y) \}$

$= (T_\pi V_\pi)(x)$
**The Banach Fixed-Point Theorem**

**Definition (Contraction)**
Let $T : \mathbb{R}^n \to \mathbb{R}^n$. $T$ is a contraction if $\exists \gamma < 1$ s.t. for any $U, V \in \mathbb{R}^n$,

$$\|TU - TV\| \leq \gamma \|U - V\| .$$

**Theorem (Banach fixed-point theorem)**
Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a contraction with factor $\gamma$. Then $\exists! V \in \mathbb{R}^n$ s.t. $TV = V$. Further, $\forall V_0 \in \mathbb{R}^n$, the sequence $V_{k+1} = TV_k$ converges to $V$ and $\|V_k - V\| = O(\gamma^k)$.

**Note**
This all holds when $(\mathbb{R}^n, \| \cdot \|)$ is replaced by a Banach-space (e.g., $(\mathbb{R}^X, \| \cdot \|_{\infty})$).
Policy evaluations are contractions

Let $\| \cdot \| = \| \cdot \|_\infty$.

**Theorem**

Let $\pi$ be any stationary policy. Then $T_\pi$ is a $\gamma$-contraction.

**Corollary**

The function $V_\pi$ is the unique fixed point of $T_\pi$ and $V_{k+1} = T_\pi V_k \rightarrow V_\pi$ for any $V_0 \in B(\mathcal{X})$ and $\| V_k - V_0 \| = O(\gamma^k)$. 
**The Bellman Optimality Operator**

**Definition (BOO)**

Let $T : B(\mathcal{X}) \to B(\mathcal{X})$, 

$$(TV)(x) \overset{\text{def}}{=} \max_{a \in A} \sum_{y \in \mathcal{X}} p(y|x, a) \left\{ r(x, a, y) + \gamma V(y) \right\}.$$ 

**Definition (Greedy Policy)**

Policy $\pi : \mathcal{X} \to A$ is greedy w.r.t. $V$ if $T_{\pi} V = TV$, or 

$$\sum_{y \in \mathcal{X}} p(y|x, \pi(x)) \left\{ r(x, \pi(x), y) + \gamma V(y) \right\} = \max_{a \in A} \sum_{y \in \mathcal{X}} p(y|x, a) \left\{ r(x, a, y) + \gamma V(y) \right\}.$$ 

**Proposition**

$T$ is a $\gamma$-contraction.
**Theorem**

\[ TV^* = V^*. \]

**Definition**

Let \( T_1, T_2 : B(\mathcal{X}) \to B(\mathcal{X}) \). We say that \( T_1 \leq T_2 \) if \( \forall V \in B(\mathcal{X}), T_1 V \leq T_2 V \).

**Proof.**

Let \( V \) be the fixed point of \( T \).

\[ T_\pi \leq T \Rightarrow V^* \leq V. \]

Let \( \pi \) be greedy w.r.t. \( V \): \( T_\pi V = TV \Rightarrow V_\pi = V \)

\[ \Rightarrow V = V_\pi \leq \max_\pi V_\pi = V^* \Rightarrow V = V^*. \]
VALUE ITERATION

Theorem
For any \( V_0 \in B(\mathcal{X}) \), let \( V_{k+1} = TV_k \), \( k = 0, 1, \ldots \). Then \( V_k \to V^* \) and \( \|V_k - V^*\| = O(\gamma^k) \).

Theorem
Let \( V \in B(\mathcal{X}) \) arbitrary and \( \pi \) be greedy w.r.t. \( V \). Then
\[
\|V_\pi - V^*\| \leq \frac{2\|TV - V\|}{1 - \gamma}.
\]

Proof.
\[
\|V_\pi - V^*\| \leq \|V_\pi - V\| + \|V - V^*\|.
\]
\[
T_\pi V = TV \Rightarrow V_\pi - V = TV_\pi - TV + TV - V.
\]
\[
V^* - V = TV^* - TV + TV - V.
\]
Use triangle inequalities.
**Policy Iteration (Howard, 1960)**

**Definition**

For $U, V \in B(\mathcal{X})$ we say that $U \geq V$ if $\forall x \in \mathcal{X}, U(x) \geq V(x)$. For $U, V \in B(\mathcal{X})$ we say that $U > V$ if $U \geq V$ and $\exists x \in \mathcal{X}$ s.t. $U(x) > V(x)$.

**Theorem (Policy Improvement)**

Let $\pi$ be any policy. Let $\pi'$ be greedy w.r.t. $V_\pi$. Then $V_{\pi'} \geq V_\pi$. If $TV_\pi > V_\pi$ then $V_{\pi'} > V_\pi$. 
**Policy Iteration Algorithm**

**Policy-Iteration**\((\pi)\)

1. \(V := V_{\pi}\)
2. do
3. \(V' := V\)
4. Let \(\pi\) be greedy w.r.t. \(V\)
5. Evaluate \(\pi: V := V_{\pi}\)
6. while \((V > V')\)
7. return \(\pi\)

**Theorem**

Consider a finite MDP. Policy-Iteration stops after a finite number of steps, returning an optimal policy. Further, it is at least as fast as value iteration.
**Value Iteration on Large State Spaces**

**Approximate Value Iteration**

\[ V_{k+1} = TV_k + \epsilon_k. \]

**Theorem**

Let \( \epsilon = \max_k \|\epsilon_k\| \). Then

\[
\limsup_{k \to \infty} \|V_k - V^*\| \leq \frac{2\gamma \epsilon}{1 - \gamma}.
\]

**Proof.**

Let \( a_k = \|V_k - V^*\| \).

\[
a_{k+1} = \|V_{k+1} - V^*\| = \|TV_k - TV^* + \epsilon_k\| \leq \gamma \|V_k - V^*\| + \epsilon
\]

\[
= \gamma a_k + \epsilon.
\]

Hence, \( a_k \) is bounded.

Take \( \limsup \) of both sides \( \Rightarrow a \leq \gamma a + \epsilon \), reorder. \( \square \)
**VALUE ITERATION ON LARGE STATE SPACES**

**Definition (Non-expansion)**

\[ A : B(\mathcal{X}) \to B(\mathcal{X}) \] is a non-expansion if \( \forall U, V \in B(\mathcal{X}), \) \( \|AU - AV\| \leq \|U - V\|. \)

**Fitted Value Iteration [Gordon, 1995]**

\[ V_{k+1} = ATV_k. \]

**Theorem**

Let \( U, V \in B(\mathcal{X}) \) s.t. \( ATU = U, TV = V. \) Then

\[ \|U - V\| \leq \frac{\|AV - V\|}{1 - \gamma}. \]

**Proof.**

Let \( U' \) be the fixed point of \( TA. \) Then \( \|U' - V\| \leq \frac{\gamma\|AV - V\|}{1 - \gamma}. \)

Since \( AU' = AT(AU'), \) thus \( U = AU'. \)

Hence, \( \|U - V\| = \|AU' - V\| \leq \|AU' - AV\| + \|AV - V\|. \) \( \square \)
**Definition (Action values)**

Let $A_0 = a$, from step 1 follow policy to get $\pi$ to obtain $R_0, R_1, \ldots$.

$$Q_\pi(x, a) = \mathbb{E}[R_{0:\infty}|X_0 = x, A_0 = a].$$

**Definition (Optimal action-value function)**

$Q^*: B(\mathcal{X} \times \mathcal{A}) \rightarrow B(\mathcal{X} \times \mathcal{A})$:

$$Q^*(x, a) = \sup_{\pi} Q_\pi(x, a), \quad (x, a) \in \mathcal{X} \times \mathcal{A}.$$

**Definition (Operators)**

$T_\pi, T: B(\mathcal{X} \times \mathcal{A}) \rightarrow B(\mathcal{X} \times \mathcal{A})$:

$$(T_\pi Q)(x, a) = \sum_{y \in \mathcal{X}} p(x, a, y) \left\{ r(x, a, y) + Q(y, \pi(y)) \right\},$$

$$(TQ)(x, a) = \sum_{y \in \mathcal{X}} p(x, a, y) \left\{ r(x, a, y) + \max_{a' \in \mathcal{A}} Q(y, a') \right\}.$$
**Definition (Greedy Policy)**

Policy \( \pi : \mathcal{X} \to \mathcal{A} \) is greedy w.r.t. \( Q \) if

\[
Q(x, \pi(x)) = \max_{a \in \mathcal{A}} Q(x, a).
\]

**Algorithms**

- Value iteration for policy evaluation: 
  \[ Q_{k+1} = T_\pi Q_k \to Q_\pi \]
- Value iteration for computing \( Q^* \):
  \[ Q_{k+1} = TQ_k \to Q^* \]
- Policy iteration: \( \pi_{k+1} \) greedy w.r.t. \( Q_{\pi_k} \).
  
  Policy Iteration Theorem continues to hold.

**Note**

Bounds for approximate procedures still hold.
**Problem**

Let \((X_i, Y_i) \sim P_\ast(\cdot, \cdot), X_i \in \mathcal{X}, \ Y_i \in \mathbb{R}, \ i = 1, 2, \ldots, n.\)

Find \(f : \mathcal{X} \rightarrow \mathbb{R}\) s.t.

\[
L(f) = \mathbb{E} \left[ (f(X) - Y)^2 \right]
\]

is minimal, where \((X, Y) \sim P_\ast(\cdot, \cdot).\)

**Theorem (Optimal Solution \equiv Conditional Expectation)**

Let \(f^*(x) = \mathbb{E}[Y|X = x].\) Then for any \(f, L(f^*) \leq L(f).\)
**Procedures**

**Definition (Empirical Risk Functional)**

\[ L_n(f) = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2. \]

**Note (Law of Large Numbers)**

Fix \( f \). As \( n \to \infty \), \( L_n(f) \to L(f) = \mathbb{E} [(f(X) - Y)^2] \).

**Empirical Risk Minimization**

Fix \( \mathcal{F} \subset \mathbb{R}^X \), \( f_n := \arg\min_{f \in \mathcal{F}} L_n(f) \).

Note: if \( \mathcal{F} = \mathbb{R}^X \), \( L_n(f_n) = 0 \): “overfitting”.

**Structural Risk Minimization**

Fix \( (\mathcal{F}_d)_{d \in \mathbb{N}} \), an infinite sequence of increasing set of functions.

\( f_n = \arg\min_{f \in \mathcal{F}_d, d \in \mathbb{N}} L_n(f) + \text{pen}(n, d) \).

**Regularization**

Fix \( \mathcal{F} \subset \mathbb{R}^X \). \( f_n = \arg\min_{f \in \mathcal{F}} L_n(f) + \lambda \| f \| \).
**Definition (Consistency)**

An algorithm $A$ is **consistent** if for any probability distribution $P_*$,

$$
\mathbb{E} \left[ L(f_n^A) \right] \to L^*.
$$

**Theorem (Slow Rates; Theorem 3.1 of [Györfi et al., 2002])**

Pick $A$ consistent, $(a_n)_n$, $a_n \to 0$.

Then $\exists P_*$ s.t. $L^* = 0$ and asymptotically $\mathbb{E} \left[ L(f_n^A) \right] \geq a_n$.

**Theorem (Lower Bound, “Curse of Dim”; Theorem 3.3 of [Györfi et al., 2002])**

Let $\mathcal{X} = [0, 1]^d$. Let $D$ be the class of $(p, C)$-smooth distributions over $\mathcal{X} \times \mathbb{R}$. Then for any $A$, any slow $(b_n)_{n \in \mathbb{N}}$, $b_n \to 0$, there exists a distribution $P \in D$ s.t. on $A$ on data from $P$ gives

$$
\mathbb{E} \left[ L(f_n^A) \right] \geq b_n n^{-\frac{2p}{2p+d}} 
$$

asymptotically.
**ERROR DECOMPOSITIONS**

\[ L(f_n) = L_n(f_n) + \{L(f_n) - L_n(f_n)\} \]

Best regressor in class \( \mathcal{F} \): \( f^*_\mathcal{F} := \arg\min_{f \in \mathcal{F}} L(f) \).

\[ L(f_n) - L^* = \{L(f^*_\mathcal{F}) - L^*\} + \{L(f_n) - L(f^*_\mathcal{F})\} \]

loss to best = approximation error + estimation error.

- **Estimation error bound:**
  \[ \mathbb{E} [L(f_n)] \leq L_n(f_n) + B(n, \mathcal{F}) \]

- **Oracle bound:**
  \[ \mathbb{E} [L(f_n)] \leq \inf_{f \in \mathcal{F}} L(f) + B(n, \mathcal{F}) \]

- **Error bound relative to the minimum loss:**
  \[ \mathbb{E} [L(f_n)] \leq L^* + B(n, \mathcal{F}) \]
**Empirical Processes**

**Definition (Loss class)**
\[ \mathcal{L} = \{ \ell_f \in \mathbb{R}^X \times \mathbb{R} : \ell_f(x, y) = (f(x) - y)^2, f \in \mathcal{F} \}. \]

**Empirical Measure**
\[ L(f) = \mathbb{E} [(f(X) - Y)^2] = \mathbb{E} [\ell_f(X, Y)] \overset{\text{def}}{=} P\ell_f. \]
\[ L_n(f) = 1/n \sum_{i=1}^n (f(X_i) - Y_i)^2 = \mathbb{E}_n [\ell_f(X, Y)] \overset{\text{def}}{=} P_n\ell_f. \]

**Hoeffding Bound**
Assume \( \ell(X_i, Y_i) \in [a, a + b] \). Then
\[ |P\ell_f - P_n\ell_f| \leq b \sqrt{\frac{\log(2/\delta)}{2n}}. \]
For any fixed $f \in \mathcal{F}$,

$$P\ell_f \leq P_n \ell_f + b \sqrt{\frac{\log(2/\delta)}{2n}}.$$
Empirical Processes: Uniform Deviations

\[ L(f_n) - L_n(f_n) \leq \sup_{f \in \mathcal{F}} (L(f) - L_n(f)). \]

Let \( \mathcal{F} = \{f_1, \ldots, f_N\} \).

When is \( \sup_{f \in \mathcal{F}} (L(f) - L_n(f)) \) large (bad event)?

If it is large for some \( f_i \):

\[
\mathbb{P} \left( \sup_{f \in \mathcal{F}} (L(f) - L_n(f)) > \epsilon \right) \leq \sum_{i=1}^{N} \mathbb{P} (L(f_i) - L_n(f_i) > \epsilon) 
\leq N \exp\left( -2n\epsilon^2 / b^2 \right).
\]
Empirical Processes: Uniform Deviations II

**Inversion:** for any $\delta > 0$, w.p. $1 - \delta$, for any $f \in \mathcal{F}$ simultaneously it holds that

$$L(f) \leq L_n(f) + b\sqrt{\frac{\log N + \log(1/\delta)}{2n}}.$$ 

**Estimation error bound:**

$$L(f_n) \leq L(f^*_\mathcal{F}) + 2b\sqrt{\frac{\log N + \log(1/\delta)}{2n}}.$$ 

**Perspective:** Key is how $P\ell_f - P_n\ell_f$ varies as $f \in \mathcal{F}$. For $\mathcal{F}$ infinite, cover $\mathcal{F}$ with balls + tones of tricks to get

$$L(f) \leq L_n(f) + C\sqrt{\frac{D_\mathcal{F} + \log(1/\delta)}{2n}},$$

where $D_\mathcal{F}$ is characteristic of the size (metric-entropy $\approx$ dimension) of $\mathcal{F}$. 
**What is it?**

**Protocol of Learning**

**Concepts:** Experimenter, Learner, Environment

states, actions, rewards

**One-shot learning:**

1. Experimenter generates training data
   \[ D = \{ (X_1, A_1, X'_1, R_1), \ldots, (X_n, A_n, X'_n, R_n) \} \] by following some policy \( \pi_b \) in the Environment

2. Learner computes a policy \( \pi \) based on \( D \)

3. Policy is implemented in the Environment
   (it’s performance \( V_\pi \) is compared with that of \( \pi_b \))

**Goal:** maximize \( V_\pi \rightarrow \max \)
EVALUATING A POLICY WITH FITTED VALUE ITERATION

TRAINING DATA
Training data \( D = \{ (X_1, A_1, R_1, X_2, A_2, R_2, \ldots, X_n, A_n, R_n, X_{n+1}) \} \) generated by following some policy \( \pi \)

FACT
\[ \forall Q \in B(\mathcal{X} \times \mathcal{A}), \pi : \mathcal{X} \rightarrow \mathcal{A} \text{ it holds that} \]
\[ \mathbb{E} \left[ R_t + \gamma Q(X_{t+1}, \pi(X_{t+1})) \bigg| X_t = x, A_t = a \right] = (T_\pi Q)(x, a) \]

FITTED VALUE ITERATION FOR POLICY EVALUATION
Let \( L_n(f; Q) = \sum_{t=1}^{n} \{ R_t + \gamma Q(X_{t+1}, \pi(X_{t+1})) - f(X_t, A_t) \}^2 \).

1. Pick \( Q_0, m := 0 \).
2. do
3. \[ Q_{m+1} := \arg\min_{Q \in \mathcal{F}} L_n(Q; Q_m). \]
4. \[ m := m + 1 \]
5. while \( (\| Q_m - Q_{m-1} \| > \epsilon) \)
**Fitted value iteration II**

\[
L_n(f; Q) = \sum_{t=1}^{n} \left\{ R_t + \gamma Q(X_{t+1}, \pi(X_{t+1})) - f(X_t, A_t) \right\}^2
\]

\[
Q_{m+1} = \arg\min_{Q \in \mathcal{F}} L_n(Q; Q_m).
\]

**Error Analysis**

Define \( \epsilon_m = Q_{m+1} - T_\pi Q_m \). Then \( Q_{m+1} = T_\pi Q_m + \epsilon_m \).

Plan:

- Show that \( \epsilon_m \) is small if \( n \) is big and \( \mathcal{F} \) is rich enough.
- Show that \( \| Q_M - Q_\pi \|_\nu^2 \) is small if all the errors are small
  “Error propagation”
Let $Q_{m+1} = T_\pi Q_m + \epsilon_m$, $\epsilon_{-1} = Q_0 - Q_\pi$.

\[
U_{m+1} = Q_{m+1} - Q_\pi \\
= T_\pi Q_m - Q_\pi + \epsilon_m \\
= T_\pi Q_m - T_\pi Q_\pi + \epsilon_m \\
= \gamma P_\pi U_m + \epsilon_m.
\]

Hence

\[
U_M = \sum_{m=0}^{M} (\gamma P_\pi)^{M-m} \epsilon_{m-1}.
\]
ERROR PROPAGATION II.

Notation:
for ρ measure, \( \rho f = \int f(x)\rho(dx) \); \((Pf)(x) = \int f(y)P(dy|x)\).

\[
U_M = \sum_{m=0}^{M} \left(\gamma P_\pi\right)^{M-m} \epsilon_{m-1}.
\]

\[
\mu|U_M|^2 \leq \left(\frac{1}{1-\gamma}\right)^2 \frac{1-\gamma}{1-\gamma^{M+1}} \sum_{m=0}^{M} \gamma^m \mu \left((P_\pi)^m \epsilon_{M-m-1}\right)^2
\]
(Jensen twice)

\[
\leq C_1 \left(\frac{1}{1-\gamma}\right)^2 \frac{1-\gamma}{1-\gamma^{M+1}} \sum_{m=0}^{M} \gamma^m \nu|\epsilon_{M-m-1}|^2
\]
(Jensen for operators, \(\forall \rho : \rho P_\pi \leq C_1 \nu, \nu \leq C_1 \mu\))

\[
\leq C_1 \left(\frac{1}{1-\gamma}\right)^2 \frac{1-\gamma}{1-\gamma^{M+1}} \left(\gamma^M \nu|\epsilon_{-1}|^2 + \sum_{m=0}^{M} \gamma^m \epsilon^2\right)
\]
\(\epsilon := \max_m \nu|\epsilon_m|^2\)

\[
= C_1 \left(\frac{1}{1-\gamma}\right)^2 \epsilon^2 + C_1 \gamma^M \nu|\epsilon_{-1}|^2
\]
Result: If the regression errors $\nu|\epsilon_m|^2$, $\epsilon_m = Q_{m+1} - T_\pi Q_m$, are small and the system is noisy ($\forall \rho : \rho P_\pi \leq C_1 \nu$), and the test and train distributions are similar ($\nu \leq C_1 \mu$), then $\mu|U_M|^2$ is small.

How to make the regression errors small?

Regression error decomposition:

$$\| Q_{m+1} - T_\pi Q_m \|_2^2 = \| Q_{m+1} - \Pi_F T_\pi Q_m \|_2^2$$

$$+ \| \Pi_F T_\pi Q_m - T_\pi Q_m \|_2^2.$$

Bias-variance dilemma:

- If $F$ is big, the estimation error will be big
- If $F$ is small, the approximation error will be big
More algorithms

Algorithms

- Policy iteration
  - Evaluate policies with fitted value iteration
  - Solve the projected fixed point equation $\Pi_{\mathcal{F}} T_\pi Q = Q$ [Lagoudakis and Parr, 2003, Antos et al., 2008]
  - Modified Bellman-error minimization [Antos et al., 2008]
- Value iteration [Munos and Szepesvári, 2008]

Results

- Consistency
- Oracle inequalities
- Rate of convergence
- Regularization
**Conclusions**

- Merging regression and RL ⇒ efficient algorithms
- Works in practice! [Ernst et al., 2005, Riedmiller, 2005]
- Works in theory!
- Difficulty: How can you prove performance improvement?
- How to take advantage of other regularities?
  - Factored dynamics
  - Relevant features
Let’s switch to that policy – after all the paper says that learning converges at an optimal rate!
Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.

Tree-based batch mode reinforcement learning.

Stable function approximation in dynamic programming.

*A distribution-free theory of nonparametric regression*.
Springer-Verlag, New York.

Least-squares policy iteration.

Finite-time bounds for fitted value iteration.

Neural fitted Q iteration – first experiences with a data efficient neural reinforcement learning method.