Emphatic temporal-difference learning

Rupam Mahmood, Huizhen (Janey) Yu, Martha White, Rich Sutton

Reinforcement Learning and Artificial Intelligence Laboratory
Department of Computing Science
University of Alberta
Canada




What everybody should know about

Temporal-difference (TD) learning

- A general method for learning to make multi-step predictions, e.g., value functions in reinforcement learning
- Learns a guess from a guess
- Applied by Samuel to play Checkers (1959), by Tesauro to beat humans at Backgammon (1992-5) and Jeopardy! (2011), and by Deepmind to play Atari games (2015)
- Explains (accurately models) the brain reward systems of primates, rats, bees, and many other animals (see Schultz, Dayan & Montague 1997)
- Arguably solves Bellman's "curse of dimensionality"

Milestones in TD research

On-policy

1959 – First TD-like algorithm (Samuel)

1974 – First TD algorithm (Witten)

1988 – Linear TD(λ) & first convergence theory (Sutton)

1992 – General convergence theory for linear TD(λ) (Dayan)

1992 – TD-gammon (Tesauro)

1994 – Sarsa(λ) (TD for control) (Rummery)

1997 – Asymptotic bound for TD(λ) (Tsitsiklis & Van Roy)

1995-9 – LSTD(λ) (Barto & Bradtke, Boyan)

2014 – True online TD(λ) (van Seijen)

Off-policy

1989 – Q-learning (TD for control) (Watkins)

1995 – Counterexamples for convergence of linear off-policy TD learning (Baird)

1999 – Residual gradient methods (Baird)

2003 – LSPI (Lagoudakis & Parr)

2009 – Gradient-TD methods (Sutton, Maei…)

2010 – Off-policy LSTD (Yu)

2014 – Proximal-gradient TD (Mahadevan)

Context: my focus on core model-free TD learning algorithms

- TD(λ), Sarsa(λ), actor-critic, and descendants
- I see them as the key building blocks of large-scale AI architectures
  - not just for value functions and reward, but for everything (GVFs)
- I have challenging requirements that i nevertheless see as “modest”
  - Compatible with scaleable function approximation
  - Computationally congenial — extremely(?) low per-step computational complexity, O(thing being learned)
  - Sound and reasonably data efficient with off-policy training

State weightings are important, powerful, even magical,

when using “genuine function approximation”

(i.e., when the optimal solution can’t be approached)

- They are the difference between convergence and divergence in on-policy and off-policy TD learning
- They are needed to make the problem well-defined
- We can change the weighting by *emphasizing* some steps more than others in learning

Often some time steps are more important

- Early time steps of an *episode* may be more important
    - Because of *discounting*
    - Because the control objective is to maximize the value of the *starting state*
- In general, function approximation resources are limited
    - Not all states can be accurately valued
    - The accuracy of different state must be traded off!
    - You may want to control the tradeoff

Bootstrapping interacts with state importance

- In the Monte Carlo case (λ=1) the values of different states (or time steps) are estimated independently, and their importances can be assigned independently

- But with bootstrapping (λ<1) each state's value is estimated based on the estimated values of later states; if the state is important, then it becomes important to accurately value the later states even if they are not important on their own

Two kinds of importance

- Intrinsic and derived, primary and secondary
  - The one you specify, and the one that follows from it because of bootstrapping
- Our terms: *Interest* and *Emphasis*
  - Your intrinsic *interest* in valuing accurately on a time step
  - The total resultant *emphasis* that you place on each time step

Real-time off-policy prediction learning with linear function approximation

Problem

- Data

$\phi : \mathcal{S} \to \Re^n$
feature function

$$\cdots \; \boldsymbol{\phi}(S_t) \; A_t \; R_{t+1} \; \boldsymbol{\phi}(S_{t+1}) \; A_{t+1} \; R_{t+2} \; \cdots$$

- State distribution

behavior policy

$$d_\mu(s) = \lim_{t\to\infty} \Pr\left[S_t = s \mid A_{0:t-1} \sim \mu\right]$$

- Objective to minimize

parameter vector

true value function

transpose (inner product)

$$\mathrm{MSE}(\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}} d_\mu(s) i(s) \left(v_\pi(s) - \boldsymbol{\theta}^\top \boldsymbol{\phi}(s)\right)^2$$

interest function
$i : \mathcal{S} \to \Re^+$

target policy

Solution

- Emphatic TD(0)

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha M_t \rho_t \left(R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t\right) \boldsymbol{\phi}_t$$

emphasis
$M_t > 0$

importance sampling ratio

$$\rho_t = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)} \qquad \mathbb{E}[\rho_t] = 1$$

$$\boldsymbol{\phi}_t = \boldsymbol{\phi}(S_t)$$

$$\cdots\ \boldsymbol{\phi}(S_t)\ A_t\ R_{t+1}\ \boldsymbol{\phi}(S_{t+1})\ A_{t+1}\ R_{t+2}\ \cdots$$

Problem

- State distribution

behavior policy

$$d_\mu(s) = \lim_{t\to\infty} \Pr\big[S_t = s \mid A_{0:t-1} \sim \mu\big]$$

- Objective to minimize

parameter vector

true value function

transpose (inner product)

$$\mathrm{MSE}(\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}} d_\mu(s) i(s) \Big(v_\pi(s) - \boldsymbol{\theta}^\top \boldsymbol{\phi}(s)\Big)^2$$

interest function
$i : \mathcal{S} \to \Re^+$

target policy

Solution

- Emphatic TD(0)

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha M_t \rho_t \left(R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t\right) \boldsymbol{\phi}_t$$

emphasis
$M_t > 0$

importance sampling ratio

$$\rho_t = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)} \quad \mathbb{E}[\rho_t] = 1$$

$$\boldsymbol{\phi}_t = \boldsymbol{\phi}(S_t)$$

- Emphatic LSTD(0)

$$\mathbf{A}_t = \sum_{k=0}^{t} M_k \rho_k \boldsymbol{\phi}_k \big(\boldsymbol{\phi}_k - \gamma \boldsymbol{\phi}_{k+1}\big)^\top \quad \mathbf{b}_t = \sum_{k=1}^{t} M_k \rho_k R_k \boldsymbol{\phi}_k$$

$$\boldsymbol{\theta}_{t+1} = \mathbf{A}_t^{-1} \mathbf{b}_t$$

Emphasis algorithm

(Sutton, Mahmood & White 2015)

- Derived from analysis of general bootstrapping relationships (Sutton, Mahmood, Precup & van Hasselt 2014)
- Emphasis is a scalar signal $M_t \geq 0$

$$M_t = \lambda_t \, i(S_t) + (1 - \lambda_t) F_t$$

- Defined from a new scalar *followon trace* $F_t \geq 0$

$$F_t = \rho_{t-1} \gamma_t F_{t-1} + i(S_t)$$

Off-policy implications

- The emphasis weighting is *stable under off-policy TD(λ)* (like the on-policy weighting) (Sutton, Mahmood & White 2015)
  - It is the *followon* weighting, from the interest weighted behavior distribution ($d_\mu(s)i(s)$), under the target policy
- Learning is *convergent* (though not necessarily of finite variance) under the emphasis weighting for arbitrary target and behavior policies (with coverage) (Yu 2015)
- There are error bounds analogous to those for on-policy TD(λ) (Munos)
- Emphatic TD is the simplest convergent off-policy TD algorithm (one parameter, one learning rate)

<u>On</u>-policy implications

- The emphasis weighting is still special, even in the on-policy case (and even for LSTD)
  - It weights states according to their effect (including via bootstrapping) on states of high interest
  - *This may be key to optimizing interest-weighted MSE*
- Emphasis is uniform in the classical continuing case — constant λ, γ, i, and ρ
  - It makes a difference *iff* any of these are non-constant
- Let's now consider some simple episodic cases

What <u>should</u> the emphasis be?

Consider 4 simple episodic cases

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

	Interest only in the start state	**Uniform interest in all states**
No bootstrapping λ=1	?	?
Full bootstrapping λ=0	?	?

How should emphasis be distributed over the time steps of an episode?

Equally?
To the start state only?
Some other way?

Case 1

- No bootstrapping, λ=1
- Interest only in the start state

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

	time	0	1	2	3	4	5	6
boot-strapping	λ	1	1	1	1	1	1	1
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	?	?	?	?	?	?	?

How should the emphasis be distributed over time steps??

Case 1

- No bootstrapping, λ=1
- Interest only in the start state

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

	time	0	1	2	3	4	5	6
boot-strapping	λ	1	1	1	1	1	1	1
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	0	0	0	0	0	0

How should the emphasis be distributed over time steps??

Answer: All on the start state

anything else will reduce the asymptotic MSVE

Case 2

$\gamma_t = 1, \forall t$
$\rho_t = 1, \forall t$

- No bootstrapping, λ=1
- Interest in all states (to the extent that they occur)

	time	0	1	2	3	4	5	6
boot-strapping	λ	1	1	1	1	1	1	1
intrinsic interest	I	1	1	1	1	1	1	1
emphasis	M	1	?	?	?	?	?	?

How should the emphasis be distributed over time steps??

Case 2

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

- No bootstrapping, λ=1
- Interest in all states (to the extent that they occur)

	time	0	1	2	3	4	5	6
boot-strapping	λ	1	1	1	1	1	1	1
intrinsic interest	I	1	1	1	1	1	1	1
emphasis	M	1	1	1	1	1	1	1

How should the emphasis be distributed over time steps??

Answer: Equally

which is the same as what TD(λ) and LSTD(λ) would do

What should the emphasis be?

Consider 4 simple episodic cases

$\gamma_t = 1, \forall t$

$\rho_t = 1, \forall t$

	Interest only in the start state	**Uniform interest in all states**
No bootstrapping λ=1	All on the start state M_0=1, others 0	Equally M_t=1
Full bootstrapping λ=0	?	?

How should emphasis be distributed over the time steps of an episode?

Equally?
To the start state only?
Some other way?

Case 3

$\gamma_t = 1, \forall t$
$\rho_t = 1, \forall t$

- Complete bootstrapping, λ=0
- Interest only in the start state

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	?	?	?	?	?	?	?

How should the emphasis be distributed over time steps??

Case 3

$\gamma_t = 1, \forall t$
$\rho_t = 1, \forall t$

- Complete bootstrapping, λ=0
- Interest only in the start state

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	1	1	1	1	1	1

How should the emphasis be distributed over time steps??

Answer: Equally

which is the same as what TD(λ) and LSTD(λ) would do

Case 4

$\gamma_t = 1, \forall t$
$\rho_t = 1, \forall t$

- Complete bootstrapping, λ=0
- Interest in all states (to the extent that they occur)

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
intrinsic interest	I	1	1	1	1	1	1	1
emphasis	M	1	?	?	?	?	?	?

How should the emphasis be distributed over time steps??

Case 4

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

- Complete bootstrapping, λ=0
- Interest in all states (to the extent that they occur)

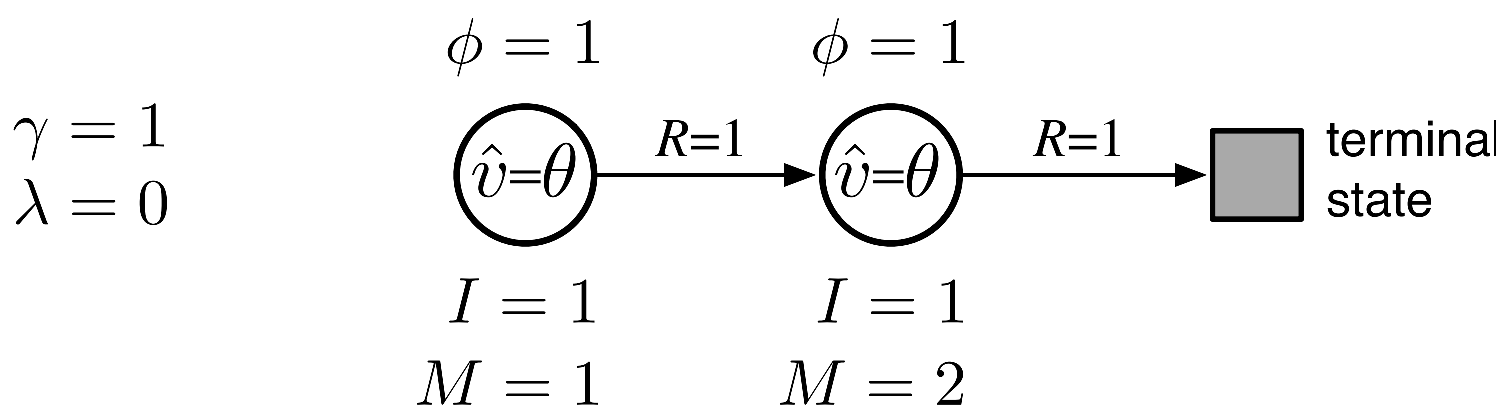
	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
intrinsic interest	I	1	1	1	1	1	1	1
emphasis	M	1	2	3	4	5	6	7

How should the emphasis be distributed over time steps??

Answer: Increasing linearly through the episode

a surprising prediction

2-state scalar example



	Solution	MSVE
Conventional TD	θ = 2	½
Emphatic TD	θ = 1.5	¼
Optimal	θ = 1.5	¼

- Increasing emphasis is not so crazy after all…
- Maybe emphasis, or something like it, can provide a uniform improvement in the asymptotic error of TD methods (λ<1)

What should the emphasis be?

Consider 4 simple episodic cases

$\gamma_t = 1, \forall t$

$\rho_t = 1, \forall t$

	Interest only in the start state	**Uniform interest in all states**
No bootstrapping λ=1	All on the start state M_0=1, others 0	Equally M_t=1
Full bootstrapping λ=0	Equally M_t=1	Increasing M_t=t

How should emphasis be distributed over the time steps of an episode?

Equally?
To the start state only?
Some other way?

The right distribution seems to depend on…everything

Derivation of the emphasis algorithm

From the general forward view of TD(λ) (Sutton et al ICML2014), the update at step k bootstraps from (and thus relies on the accuracy of) the estimate at later time $t > k$, with coefficient

$$\rho_k \left(\prod_{i=k+1}^{t-1} \gamma_i \lambda_i \rho_i \right) \gamma_t (1 - \lambda_t), \text{ where } \rho_i = \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}.$$

The degree M_t to which we should *emphasize* the update at time t is the sum of these coefficients for times $k < t$, each times the emphasis M_k for those times, plus any interest $I_t = i(S_t)$ in time t:

$$\begin{aligned} M_t &= \sum_{k=0}^{t-1} \rho_k \left(\prod_{i=k+1}^{t-1} \gamma_i \lambda_i \rho_i \right) \gamma_t (1 - \lambda_t) M_k \; + \; I_t \\ &= \lambda_t I_t + (1 - \lambda_t) I_t + (1 - \lambda_t) \gamma_t \sum_{k=0}^{t-1} \rho_k M_k \prod_{i=k+1}^{t-1} \gamma_i \lambda_i \rho_i \\ &= \lambda_t I_t + (1 - \lambda_t) F_t \end{aligned}$$

$$
\begin{aligned}
M_t &= \sum_{k=0} \rho_k \left(\prod_{i=k+1} \gamma_i \lambda_i \rho_i \right) \gamma_t (1-\lambda_t)\, M_k \; + \; I_t \\
&= \lambda_t I_t + (1-\lambda_t) I_t + (1-\lambda_t)\, \gamma_t \sum_{k=0}^{t-1} \rho_k M_k \prod_{i=k+1}^{t-1} \gamma_i \lambda_i \rho_i \\
&= \lambda_t I_t + (1-\lambda_t) F_t
\end{aligned}
$$

The scalar random variable F_t, called the *followon trace*, can be written and updated recursively by

$$
\begin{aligned}
F_{t+1} &= I_{t+1} + \gamma_{t+1} \sum_{k=0}^{t} \rho_k M_k \prod_{i=k+1}^{t} \gamma_i \lambda_i \rho_i \qquad \text{(by def'n)} \\
&= I_{t+1} + \gamma_{t+1} \left(\rho_t M_t + \sum_{k=0}^{t-1} \rho_k M_k \prod_{i=k+1}^{t} \gamma_i \lambda_i \rho_i \right) \\
&= I_{t+1} + \gamma_{t+1} \left(\rho_t \left(\lambda_t I_t + (1-\lambda_t) F_t \right) + \rho_t \lambda_t \gamma_t \sum_{k=0}^{t-1} \rho_k M_k \prod_{i=k+1}^{t-1} \gamma_i \lambda_i \rho_i \right) \\
&= I_{t+1} + \gamma_{t+1} \left(\rho_t F_t - \rho_t \lambda_t F_t + \rho_t \lambda_t I_t + \rho_t \lambda_t \gamma_t \sum_{k=0}^{t-1} \rho_k M_k \prod_{i=k+1}^{t-1} \gamma_i \lambda_i \rho_i \right)
\end{aligned}
$$

The scalar random variable F_t, called the *followon trace*, can be written and updated recursively by

$$
\begin{aligned}
F_{t+1} &= I_{t+1} + \gamma_{t+1} \sum_{k=0}^{t} \rho_k M_k \prod_{i=k+1}^{t} \gamma_i \lambda_i \rho_i \qquad \text{(by def'n)} \\
&= I_{t+1} + \gamma_{t+1} \left(\rho_t M_t + \sum_{k=0}^{t-1} \rho_k M_k \prod_{i=k+1}^{t} \gamma_i \lambda_i \rho_i \right) \\
&= I_{t+1} + \gamma_{t+1} \left(\rho_t \left(\lambda_t I_t + (1-\lambda_t) F_t \right) + \rho_t \lambda_t \gamma_t \sum_{k=0}^{t-1} \rho_k M_k \prod_{i=k+1}^{t-1} \gamma_i \lambda_i \rho_i \right) \\
&= I_{t+1} + \gamma_{t+1} \left(\rho_t F_t - \rho_t \lambda_t F_t + \rho_t \lambda_t I_t + \rho_t \lambda_t \gamma_t \sum_{k=0}^{t-1} \rho_k M_k \prod_{i=k+1}^{t-1} \gamma_i \lambda_i \rho_i \right) \\
&= I_{t+1} + \gamma_{t+1} \left(\rho_t F_t - \rho_t \lambda_t F_t + \rho_t \lambda_t F_t \right) \\
&= I_{t+1} + \gamma_{t+1} \rho_t F_t,
\end{aligned}
$$

or

$$
F_t = \gamma_t \rho_{t-1} F_{t-1} + I_t \quad \text{with } F_{-1} = 0.
$$

Case 1except

- No bootstrapping, λ=1, *except at one step*
- Interest only in the start state

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

	time	0	1	2	3	4	5	6
boot-strapping	λ	1	1	1	0	1	1	1
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	?	?	?	?	?	?	?

How should the emphasis be distributed over time steps??

Case 1except

- No bootstrapping, λ=1, *except at one step*
- Interest only in the start state

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

	time	0	1	2	3	4	5	6
boot-strapping	**λ**	1	1	1	0	1	1	1
intrinsic interest	**I**	1	0	0	0	0	0	0
emphasis	**M**	1	0	0	1	0	0	0

How should the emphasis be distributed over time steps??

Bootstrapping gives emphasis

Case 2except

- No bootstrapping, λ=1, *except at one step*
- Interest in all states (to the extent that they occur)

$\gamma_t = 1, \forall t$

$\rho_t = 1, \forall t$

	time	0	1	2	3	4	5	6
boot-strapping	**λ**	1	1	1	0	1	1	1
intrinsic interest	**I**	1	1	1	1	1	1	1
emphasis	**M**	1	1	1	?	?	?	?

How should the emphasis be distributed over time steps??

Case 2except

$\gamma_t = 1, \forall t$
$\rho_t = 1, \forall t$

- No bootstrapping, λ=1, *except at one step*
- Interest in all states (to the extent that they occur)

	time	0	1	2	3	4	5	6
boot-strapping	λ	1	1	1	0	1	1	1
intrinsic interest	I	1	1	1	1	1	1	1
emphasis	M	1	1	1	4	1	1	1

How should the emphasis be distributed over time steps??

The followon trace accumulates with interest
It just needs bootstrapping to bring it out

Case 2except-twice

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

- No bootstrapping, λ=1, *except at <u>two</u> steps*
- Interest in all states (to the extent that they occur)

	time	0	1	2	3	4	5	6
boot-strapping	λ	1	1	1	0	1	0	1
intrinsic interest	I	1	1	1	1	1	1	1
emphasis	M	1	1	1	4	1	6	1

How should the emphasis be distributed over time steps??

Case 2 except-twice

- No bootstrapping, λ=1, *except at <u>two</u> steps*
- Interest in all states (to the extent that they occur)

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

	time	0	1	2	3	4	5	6
boot-strapping	λ	1	1	1	0	1	0	1
intrinsic interest	I	1	1	1	1	1	1	1
emphasis	M	1	1	1	4	1	6	1

How should the emphasis be distributed over time steps??

The followon trace accumulates with interest
It just needs bootstrapping to bring it out

Case 3except

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

- Complete bootstrapping, λ=0, *except at one step*
- Interest only in the start state

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	1	0	0	0
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	1	1	?	?	?	?

How should the emphasis be distributed over time steps??

Case 3except

$\gamma_t = 1, \forall t$
$\rho_t = 1, \forall t$

- Complete bootstrapping, λ=0, *except at one step*
- Interest only in the start state

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	1	0	0	0
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	1	1	0	1	1	1

How should the emphasis be distributed over time steps??

The state is ignored, skipped over,
but bootstrapping continues afterwards

Case 3 the-other-except

$$\gamma_t = 1, \forall t$$
$$\rho_t = 1, \forall t$$

- Complete bootstrapping, λ=0
- Interest only in the start state, *and at one other step*

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
intrinsic interest	I	1	0	0	1	0	0	0
emphasis	M	1	1	1	?	?	?	?

How should the emphasis be distributed over time steps??

Case 3the-other-except

$\gamma_t = 1, \forall t$
$\rho_t = 1, \forall t$

- Complete bootstrapping, λ=0
- Interest only in the start state, *and at one other step*

	time	0	1	2	3	4	5	6
boot-strapping	**λ**	0	0	0	0	0	0	0
intrinsic interest	**I**	1	0	0	1	0	0	0
emphasis	**M**	1	1	1	2	2	2	2

How should the emphasis be distributed over time steps??

Again, interest accumulates in the followon trace
and is revealed by bootstrapping

Case 4except

$\gamma_t = 1, \forall t$
$\rho_t = 1, \forall t$

- Complete bootstrapping, λ=0, *except at one step*
- Interest in all states (to the extent that they occur)

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	1	0	0	0
intrinsic interest	I	1	1	1	1	1	1	1
emphasis	M	1	2	3	?	?	?	?

How should the emphasis be distributed over time steps??

Case 4except

$\gamma_t = 1, \forall t$
$\rho_t = 1, \forall t$

- Complete bootstrapping, λ=0, *except at one step*
- Interest in all states (to the extent that they occur)

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	1	0	0	0
intrinsic interest	I	1	1	1	1	1	1	1
emphasis	M	1	2	3	1	5	6	7

How should the emphasis be distributed over time steps??

Weird, but it kinda makes sense…

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	½	½	½	½	½	½	½
importance sampling	ρ	1	1	1	1	1	1	1
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	?	?	?	?	?	?	?

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	½	½	½	½	½	½	½
importance sampling	ρ	1	1	1	1	1	1	1
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	½	¼	⅛	1/16	1/32	1/64

$M_t = \gamma^t$ Phil Thomas, 2014

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	½	½	½	½	½	½	½
importance sampling	ρ	1	1	1	1	1	1	1
intrinsic interest	I	0	1	0	0	1	0	0
emphasis	M	?	?	?	?	?	?	?

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	½	½	½	½	½	½	½
importance sampling	ρ	1	1	1	1	1	1	1
intrinsic interest	I	0	1	0	0	1	0	0
emphasis	M	0	1	½	¼	1+⅛	½+1/16	¼+1/32

Off-policy examples…

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	1	1	1	1	1	1	1
importance sampling	ρ	1	1	½	1	1	1	1
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	1	?	?	?	?	?

$$\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$$

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	1	1	1	1	1	1	1
importance sampling	ρ	1	1	½	1	1	1	1
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	1	1	½	½	½	½

$$\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$$

if there is a deviation, it affects the next emphasis

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	1	1	1	1	1	1	1
importance sampling	ρ	1	1	0	1	1	1	1
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	1	?	?	?	?	?

$$\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$$

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	1	1	1	1	1	1	1
importance sampling	ρ	1	1	0	1	1	1	1
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	1	1	0	0	0	0

$$\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$$

if there is a deviation, then nothing after matters
(until the next intrinsically interesting thing happens)

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	1	1	1	1	1	1	1
importance sampling	ρ	1	2	2	2	¼	1	0
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	?	?	?	?	?	?

$$\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$$

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	1	1	1	1	1	1	1
importance sampling	ρ	1	2	2	2	¼	1	0
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	1	2	4	8	?	?

$$\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$$

You must scale by the product of importance sampling ratios

What should the emphasis be?

	time	0	1	2	3	4	5	6
boot-strapping	λ	0	0	0	0	0	0	0
dis-counting	γ	1	1	1	1	1	1	1
importance sampling	ρ	1	2	2	2	¼	1	0
intrinsic interest	I	1	0	0	0	0	0	0
emphasis	M	1	1	2	4	8	2	2

$$\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$$

You must scale by the product of importance sampling ratios

Conclusions

- The allocation of function approximation resources by state weightings is important
  - It can make off-policy learning stable
  - Our emphasis algorithm makes some surprising predictions about optimal allocation of FA resources
  - It *may* be able to improve error bounds for *on-policy* learning
- We have treated only policy evaluation (prediction); the control case will bring its own surprises
- There is still a lot to learn about bootstrapping, state weightings, and function approximation

Thank you for your attention

Rupam Mahmood

Janey Yu

Martha White