Choice of objective for approximate policy evaluation

Arthur Guez

Tea talk

Should one compute the Temporal Difference fix point or minimize the Bellman Residual ? The unified oblique projection view

---

Bruno Scherrer SCHERRER@LORIA.FR

Reminder: Exact Policy Evaluation

Value for policy π at state i:

$$v_\pi(i) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r(i_k) | i_0 = i]$$

Let $v_\pi \in \mathbb{R}^N$, P is a matrix $N \times N$ containing transition prob (dynamics + policy). Then v_π is the unique **fixed point** of the Bellman operator:

$$\mathcal{T}v := r + \gamma P v$$

$$v_\pi = \mathcal{T}v_\pi \implies v_\pi = (I - \gamma P)^{-1} r$$

Approximate Policy Evaluation

(Note: π fixed, dropping π subscripts)

Suppose N is very large (or infinite), parametrize v with low-dim vector w as:

$$\hat{v}(i) = \sum_{j=1}^{m} w_j \phi_j(i)$$

with $m << N$ and ϕ_j the feature vectors.
Denote by $\Phi = (\phi_1 \dots \phi_m)$ the $N \times m$ feature matrix, then:

$$\hat{v} = \Phi w$$

Approximate Policy Evaluation

Which $\hat{v}$ should we compute to approximate v?

- Ideal $\hat{v}$: minimize $\|\hat{v} - v\|$ according to some norm.

Approximate Policy Evaluation

Which $\hat{v}$ should we compute to approximate v?

- Ideal $\hat{v}$: minimize $\|\hat{v} - v\|$ according to some norm.
- Usual norm in DP/RL: ξ-weighted quadratic norm ($\|x\|_\xi = \sqrt{\sum \xi_i x_i^2} = \sqrt{x' \Xi x}$), where ξ is a distribution on the states.

Approximate Policy Evaluation

Which $\hat{v}$ should we compute to approximate v?

- Ideal $\hat{v}$: minimize $\|\hat{v} - v\|$ according to some norm.
- Usual norm in DP/RL: ξ-weighted quadratic norm ($\|x\|_\xi = \sqrt{\sum \xi_i x_i^2} = \sqrt{x'\Xi x}$), where ξ is a distribution on the states.
- $$\begin{aligned} \hat{v}_{\text{best}} &= \Phi w_{\text{best}} \\ &= \underbrace{\Phi(\Phi'\Xi\Phi)^{-1}\Phi'\Xi}_{\Pi} v \\ &= \Pi(I - \gamma P)^{-1} r \end{aligned}$$

  Can't compute directly! Direct Monte-Carlo estimates are possible but high-variance.

Approximate Policy Evaluation: TD

Tractable objective: TD(0) fixpoint

- Look for fixed point of $\Pi\mathcal{T}$ operator.
- Want $\hat{v}_{\text{TD}} = \Pi\mathcal{T}\hat{v}_{\text{TD}}$. Closed-form for weights (if inverse exists):

$$w_{\text{TD}} = (\Phi'\Xi L\Phi)^{-1}\Phi'\Xi r \quad (1)$$

Approximate Policy Evaluation: TD

Tractable objective: TD(0) fixpoint

- Look for fixed point of $\Pi\mathcal{T}$ operator.
- Want $\hat{v}_{\text{TD}} = \Pi\mathcal{T}\hat{v}_{\text{TD}}$. Closed-form for weights (if inverse exists):

$$w_{\text{TD}} = (\Phi'\Xi L\Phi)^{-1}\Phi'\Xi r \tag{1}$$

- By far the most popular objective, both for incremental (online) methods (TD(0), gradient TDs) and batch (LSTD, LSPE, and some iterative methods).
- Example: Gradient TD methods minimize error $E_{\text{TD}}(\hat{v}) = \|\hat{v} - \Pi\mathcal{T}\hat{v}\|_\xi$

Approximate Policy Evaluation: BR

Tractable objective: minimize Bellman Residual

- Find $\hat{v}$ that minimize $E_{\text{BR}}(\hat{v}) = \|\hat{v} - \mathcal{T}\hat{v}\|_\xi$
- Closed-form solution for the weights (always exists):

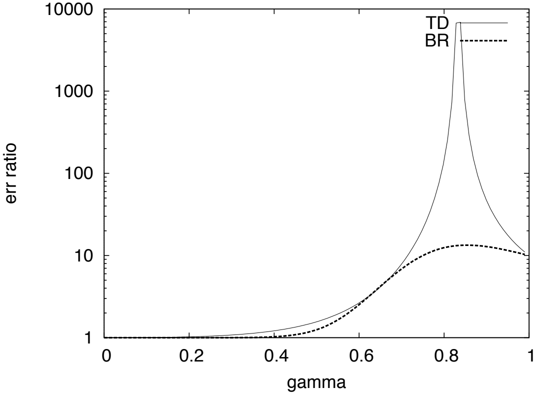
$$w_{\text{BR}} = (\Psi'\Xi\Psi)^{-1}\Psi'\Xi r, \tag{2}$$

with $\Psi = L\Phi$.

Picture and examples

Result

Quality of TD fixpoint and BR solutions on the small example:



y axis = $\frac{e(w_{TD})}{e(w_{\text{best}})}$ and $\frac{e(w_{BR})}{e(w_{\text{best}})}$

Theoretical guarantees

- **TD:** Yes but only for on-policy sampling ($\xi = p_\pi$). (Tsitsiklis & Van Roy, 1996)
  The fixpoint might exist nonetheless and most methods will converge to it. (see (Kolter 2011) for quality of solution in that case)

- **BR:** Yes in all cases (can bound the error relative to BR).
  (See (Williams & Baird, 1993) and (Munos, 2003))

BR not really popular for these reasons:

- Sample-based BR slower to converge (plus might require double-sampling).
- TD finds v_{best} but not BR in some cases.

Oblique projection view

TD and BR are both oblique projections onto span(Φ) and orthogonal to subspace spanned by $X_{TD} = \Xi\Phi$ or $X_{BR} = \Xi L\Phi$.

Can prove bound for any oblique projection. Not predictive of empirical performance according to results.

Empirical result

Randomly generated Φ and P for different N and m. More situations where TD is better:

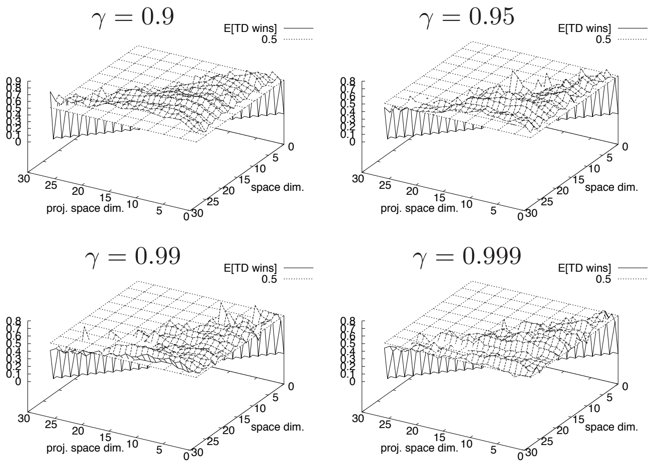
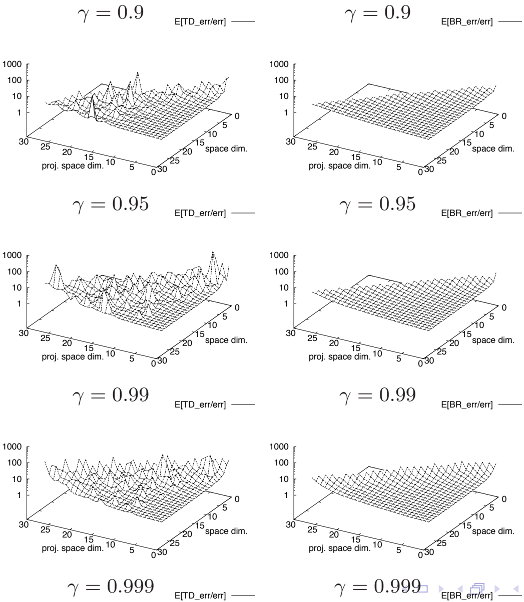


Figure 2. TD win ratio.

Empirical result

but TD fails badly with the instabilities:



Conclusion

- TD(0) objective can be unstable, has advantages in practice.
- What if ξ and p_π are not too different? Or if TD(λ) is used?
- In the end, mostly after the results of approximate policy **iteration**, with a lot more instabilities to deal with (e.g. policy oscillations).