Should one compute the Temporal Difference fix point or minimize the Bellman Residual ?

The unified oblique projection view

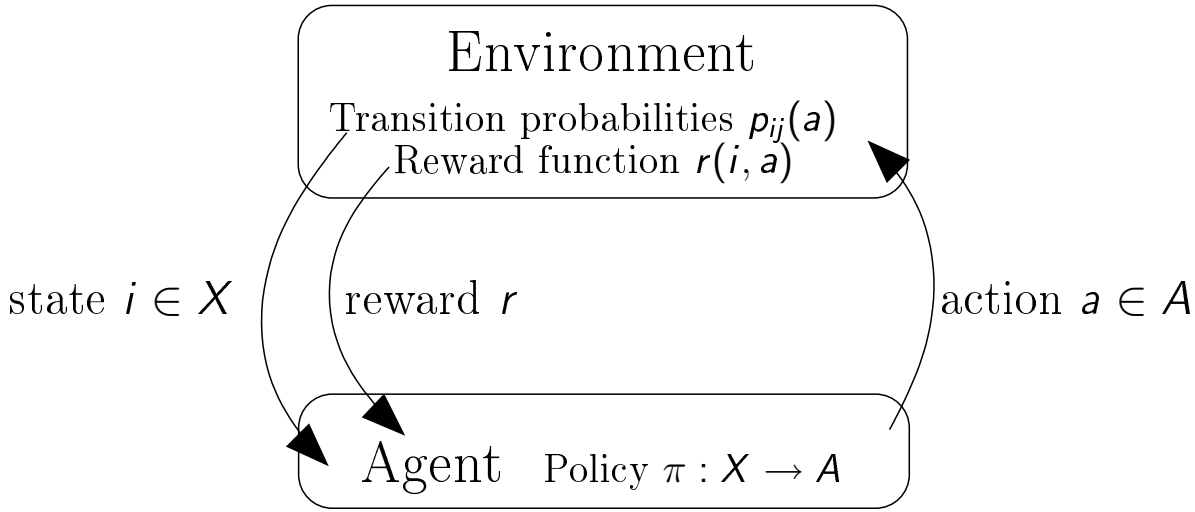
Bruno Scherrer

INRIA - LORIA - Maia Team

June 2010

Markov Decision Processes

(Puterman, 1994; Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 1998)



- **Goal :** Given a policy $\pi : X \to A$, compute

$$v(i) = E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r(i_k, \pi(i_k)) \middle| i_0 = i \right] \qquad (0 < \gamma < 1)$$

- The value v satisfies

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

- Thus

$$v = L^{-1} r \quad \text{with } L = I - \gamma P$$

- Look for a linear approximation $\hat{v}(i) = \sum_{j=1}^{m} w_j \phi_j(i)$ or $\hat{v} = \Phi w$

$$\Phi = \begin{pmatrix} \phi(1)' \\ \vdots \\ \phi(N)' \end{pmatrix} = (\underbrace{\phi_1 \quad \dots \quad \phi_m}_{\text{linearly independent}}) \text{ and } w = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix}$$

- Projection onto span $(\Phi) = \{\Phi w; w \in \mathbb{R}^m\}$
  - Let $\xi > 0$ be a distribution on the state space $\{1, \cdots, N\}$
  - Quadratic weighted norm : $\|v\|_\xi = \sqrt{\sum_i \xi(i) v(i)^2}$
  - Orthogonal projection : $\Pi(v) = \arg\min_{\hat{v} \in \text{span}(\Phi)} \|\hat{v} - v\|_\xi$
  - Linear projection in closed form : $\Xi = \text{diag}(\xi)$

$$\Pi = \Phi\pi \quad \text{with} \quad \pi = (\Phi'\Xi\Phi)^{-1}\Phi'\Xi$$

Properties : $\Phi w = \Pi v \Leftrightarrow w = \pi v, \quad \pi\Phi = I_m, \quad \pi\Pi = \pi, \quad \Pi\Pi = \Pi$

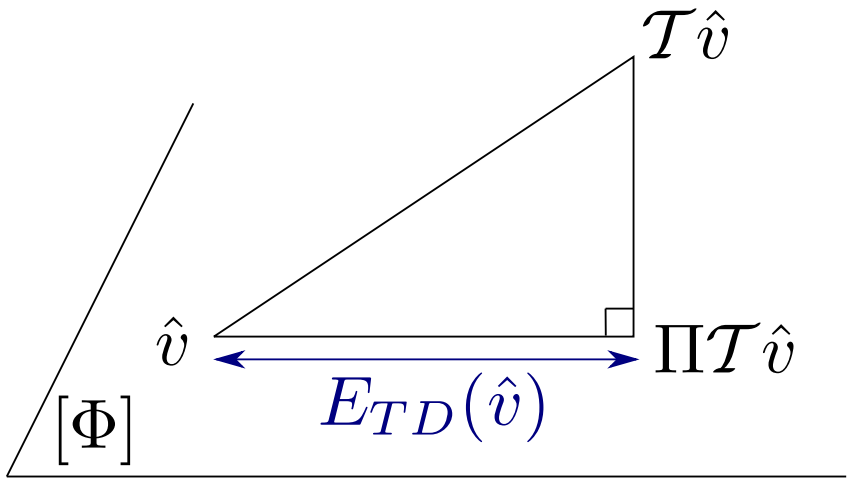
- Ideally, one would like to compute the "best" approximation

$$\hat{v}_{best} = \Phi w_{best} \text{ with } w_{best} = \pi v = \pi L^{-1} r.$$

Solution : TD(1), full trajectories, high variance
Alternatives based on one-step samples : $\hat{v} \simeq \mathcal{T}\hat{v}$

TD(0) fix point method

One looks for $\hat{v}_{TD} \in \text{span}(\Phi)$ satisfying $\hat{v}_{TD} = \Pi \mathcal{T} \hat{v}_{TD}$.



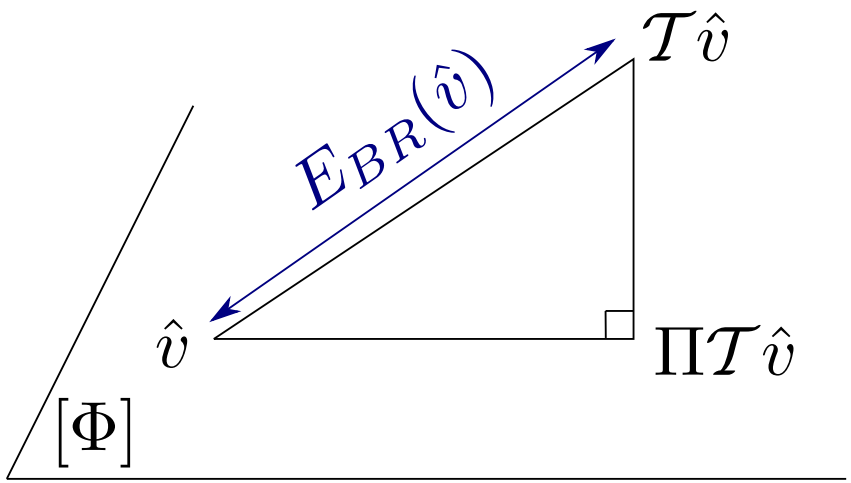
When the inverse exists (Schoknecht, 2002), it can be proved that $\hat{v}_{TD} = \Phi w_{TD}$ with

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

This is equivalent to minimizing for $\hat{v} \in \text{span}(\Phi)$ the TD error $E_{TD}(\hat{v}) := \|\hat{v} - \Pi \mathcal{T} \hat{v}\|_\xi$ down to 0. (Antos *et al.*, 2008; Farahmand *et al.*, 2008; Sutton *et al.*, 2009)

Bellman Residual minimization method

One looks for $\hat{v} \in \text{span}(\Phi)$ minimizing $E_{BR}(\hat{v}) := \|\hat{v} - \mathcal{T}\hat{v}\|_\xi$.



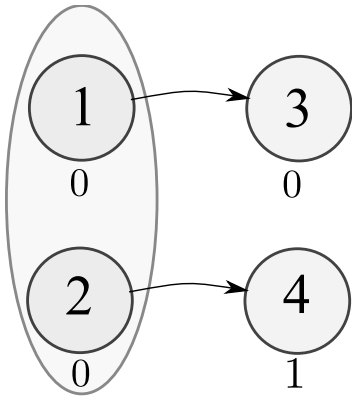
Since $E_{BR}(\Phi w) = \| \underbrace{\Phi w - \gamma P \Phi w}_{\Psi_w} - r\|_\xi$, $\quad \Psi = (I - \gamma P)\Phi = L\Phi$,

it can be seen that $\hat{v}_{BR} = \Phi w_{BR}$ with

$$w_{BR} = (\Psi' \Xi \Psi)^{-1} \Psi' \Xi r = (\Phi' L' \Xi L \Phi)^{-1} \Phi' L' \Xi r.$$

The above inverse always exists (Schoknecht, 2002).

Example 1 (Sutton *et al.*, 2009)

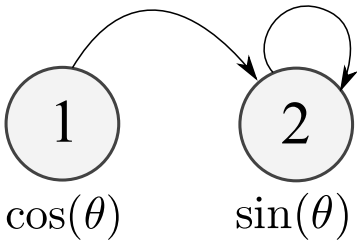


$$\Phi = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \xi = \begin{pmatrix} .25 \\ .25 \\ .25 \\ .25 \end{pmatrix},$$

$$v(1) = w_1,\ v(2) = w_1,\ v(3) = w_2,\ v(4) = w_3$$

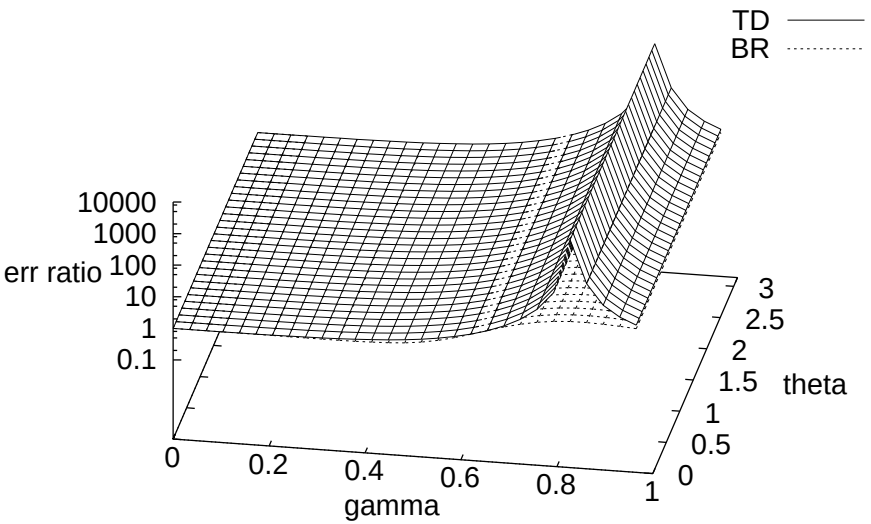
We have $\hat{v}_{TD} = \hat{v}_{best}$ while $\hat{v}_{BR} \neq \hat{v}_{best}$.

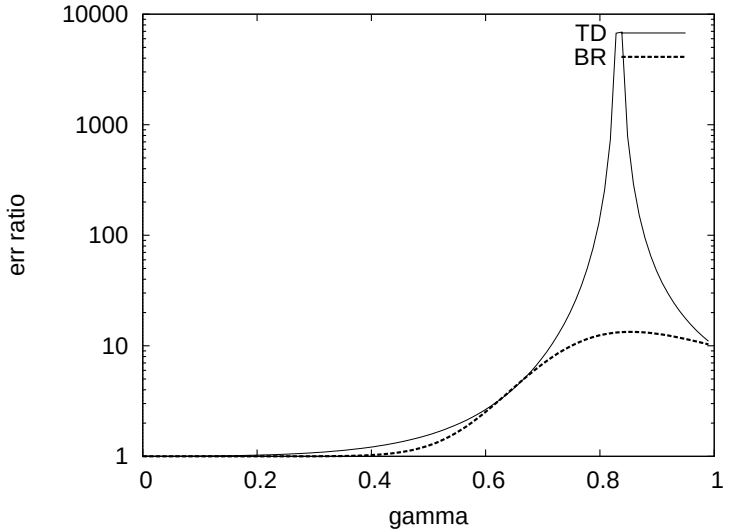
Example 2



$$\Phi = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \xi = \begin{pmatrix} .5 \\ .5 \end{pmatrix}, \quad v(1) = w,\ v(2) = 2w,$$

$$v \Rightarrow \hat{v}_{best}, \hat{v}_{TD}, \hat{v}_{BR} \quad \Rightarrow e(\hat{v}) = \frac{\|\hat{v} - v\|_\xi}{\|\hat{v}_{best} - v\|_\xi}$$





Guarantee when minimizing BR

Proposition (Williams & Baird, 1993)

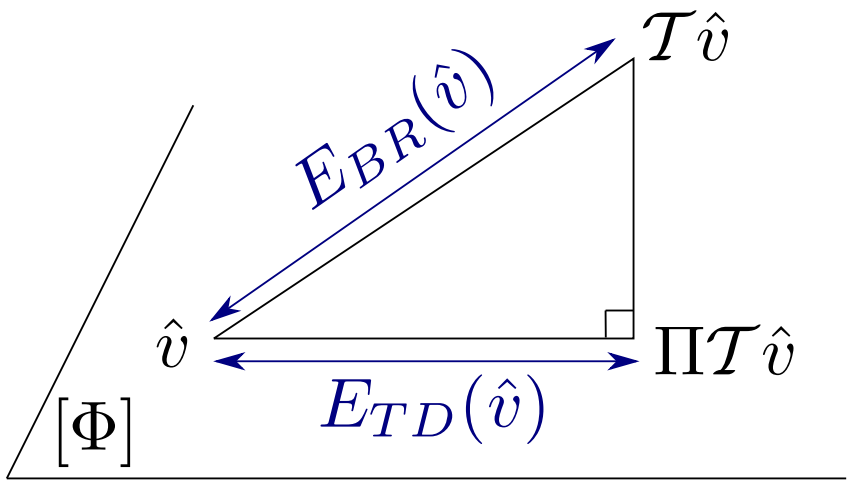
$$\forall \hat{v},\ \|v - \hat{v}\|_\infty \leq \frac{1}{1-\gamma}\|\mathcal{T}\hat{v} - \hat{v}\|_\infty.$$

Proposition

$$\forall \hat{v},\ \|v - \hat{v}\|_\xi \leq \frac{\sqrt{C(\xi)}}{1-\gamma}\|\mathcal{T}\hat{v} - \hat{v}\|_\xi$$

where $C(\xi) := \max_{i,j} \frac{p_{ij}}{\xi_i}$ is a *concentration coefficient* (Munos, 2003)

Relation between the 2 criteria



Proposition

The BR is an upper bound of the TD error, and more precisely :

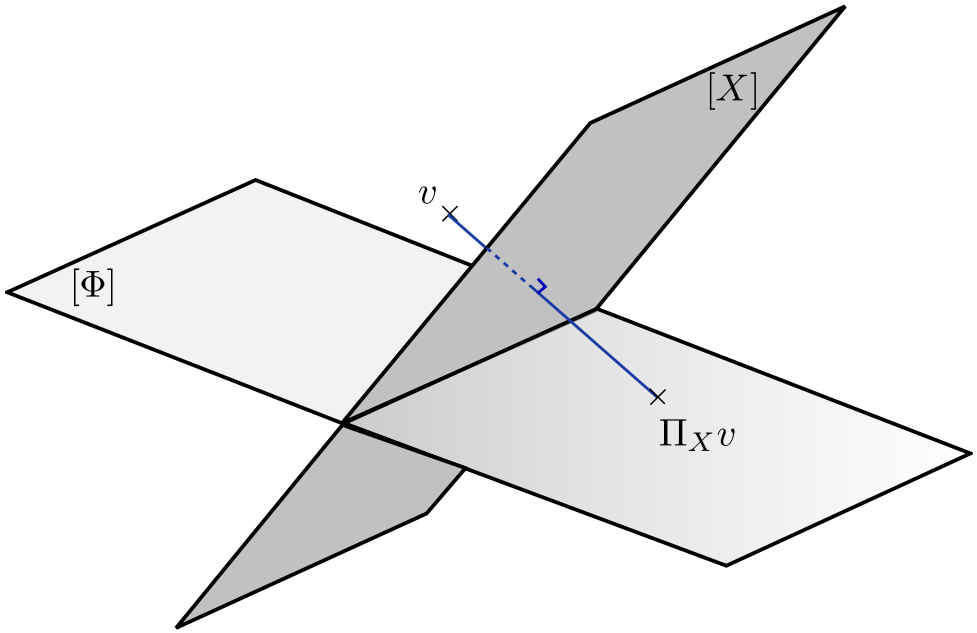
$$\forall \hat{v} \in \text{span}(\Phi), E_{BR}(\hat{v})^2 = E_{TD}(\hat{v})^2 + \|\mathcal{T}\hat{v} - \Pi\mathcal{T}\hat{v}\|_\xi^2.$$

- The adequacy term $\|\mathcal{T}\hat{v} - \Pi\mathcal{T}\hat{v}\|_\xi$ matters !

Oblique projections

Projection onto span (Φ) orthogonally to span (X)

A linear projection ($\Pi\Pi = \Pi$) is defined by its range span (Φ) (dim. m) and its null space (dim. $N - m$), of which the orthogonal complement (dim. m) is span (X).



$$\Pi_X = \Phi\pi_X, \quad \pi_X = (X'\Phi)^{-1}X'$$

Properties :

$$\Phi w = \Pi_X v \Leftrightarrow w = \pi_X v,$$
$$\pi_X\Phi = I_m,$$
$$\pi_X\Pi_X = \pi_X$$

When $X = \Phi$: Euclidean orthogonal projection

When $X = \Xi\Phi$: Orthogonal projection w.r.t. $\|\cdot\|_\xi$.

Main result

Proposition

- For any X, the solution of the projected equation

$$\hat{v}_X = \Pi_X \mathcal{T} \hat{v}_X$$

is the oblique projection of v onto $\text{span}(\Phi)$ orthogonally to $\text{span}(L'X)$, i.e. $\hat{v}_X = \Pi_{L'X}\ v$.

- The TD fix point and the BR minimizer respectively corredpond to the cases $X = X_{TD} = \Xi\Phi$ and $X = X_{BR} = \Xi L \Phi$.

- TD/BR are oblique projections of the value
- BR is a fixed point method
- Neither TD nor BR is optimal for $\|\cdot\|_\xi$ ($X^* = L'^{-1}\Xi\Phi$ is)

Proof

One solves : $\Phi w_X = \Pi_X(r + \gamma P \Phi w_X)$.

Multiplying on both sides by π_X, one obtains :

$$\begin{aligned} w_X &= \pi_X(r + \gamma P \Phi w_X) \\ w_X &= (I - \gamma \pi_X P \Phi)^{-1} \pi_X r. \end{aligned}$$

$$\begin{aligned} \text{Hence, } w_X &= (I - \gamma (X'\Phi)^{-1} X' P \Phi)^{-1} (X'\Phi)^{-1} X' r \\ &= \left[(X'\Phi)(I - \gamma (X'\Phi)^{-1} X' P \Phi) \right]^{-1} X' r \\ &= (X'(I - \gamma P)\Phi)^{-1} X' r \\ &= (X' L \Phi)^{-1} X' L v \\ &= \pi_{L'X}\, v \end{aligned}$$

$$w_{TD} = (\underbrace{\Phi'\Xi}_{X_{TD'}} L\Phi)^{-1} \underbrace{\Phi'\Xi}_{X_{TD'}} Lv \text{ and } w_{BR} = (\underbrace{\Phi' L' \Xi}_{X_{BR'}} L\Phi)^{-1} \underbrace{\Phi' L' \Xi}_{X_{BR'}} Lv.$$

Related works

Proposition (Schoknecht, 2002)

The TD fix point and the BR minimizer are orthogonal projections of the value v respectively induced by the seminorm $\|\cdot\|_{Q_{TD}}$ with $Q_{TD} = L'\Xi\Phi\Phi'\Xi L$ and by the norm $\|\cdot\|_{Q_{BR}}$ with $Q_{BR} = L'\Xi L$.

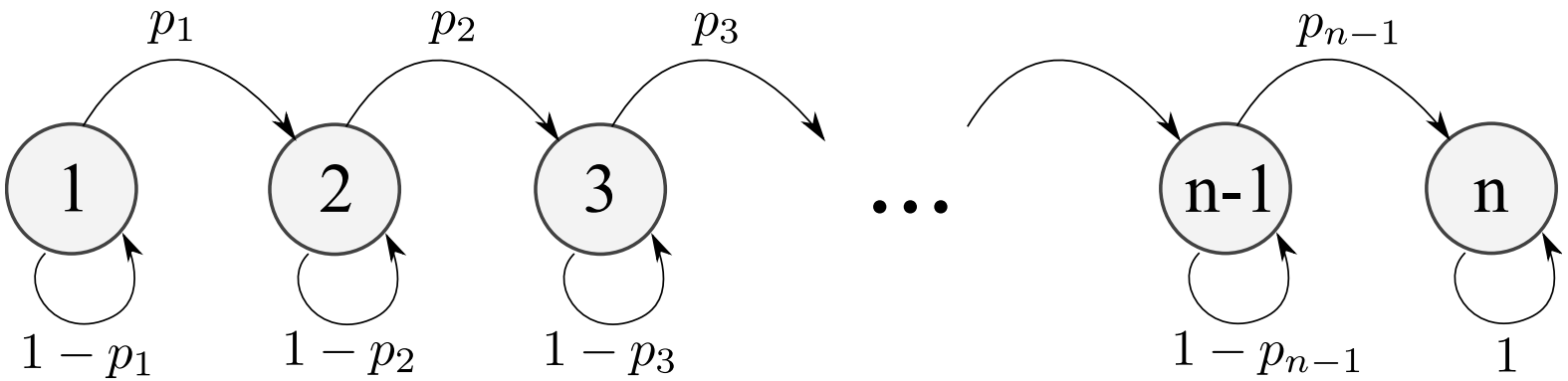
Proposition (revisiting (Yu & Bertsekas, 2008))

For any choice of X, the approximation error satisfies :

$$\begin{aligned} \|v - \hat{v}_X\|_\xi &\leq \|\Pi_{L'X}\|_\xi \|v - \hat{v}_{best}\|_\xi \\ &= \sqrt{\sigma(ABCB')}\|v - \hat{v}_{best}\|_\xi \end{aligned}$$

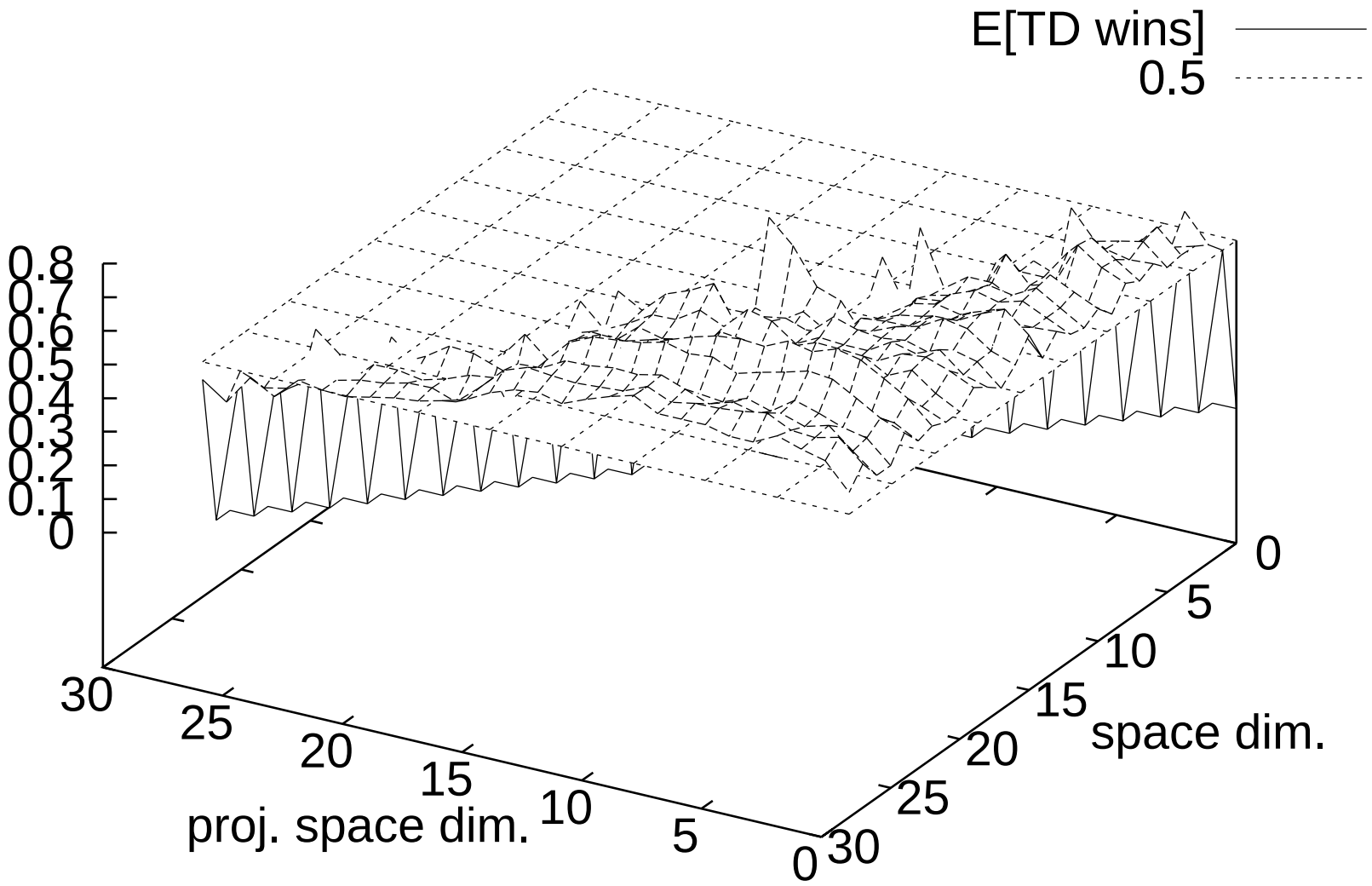
where $A = \Phi'\Xi\Phi$, $B = (X'L\Phi)^{-1}$ and $C = XL\Xi^{-1}L'X$ are matrices of size $m \times m$.

Model

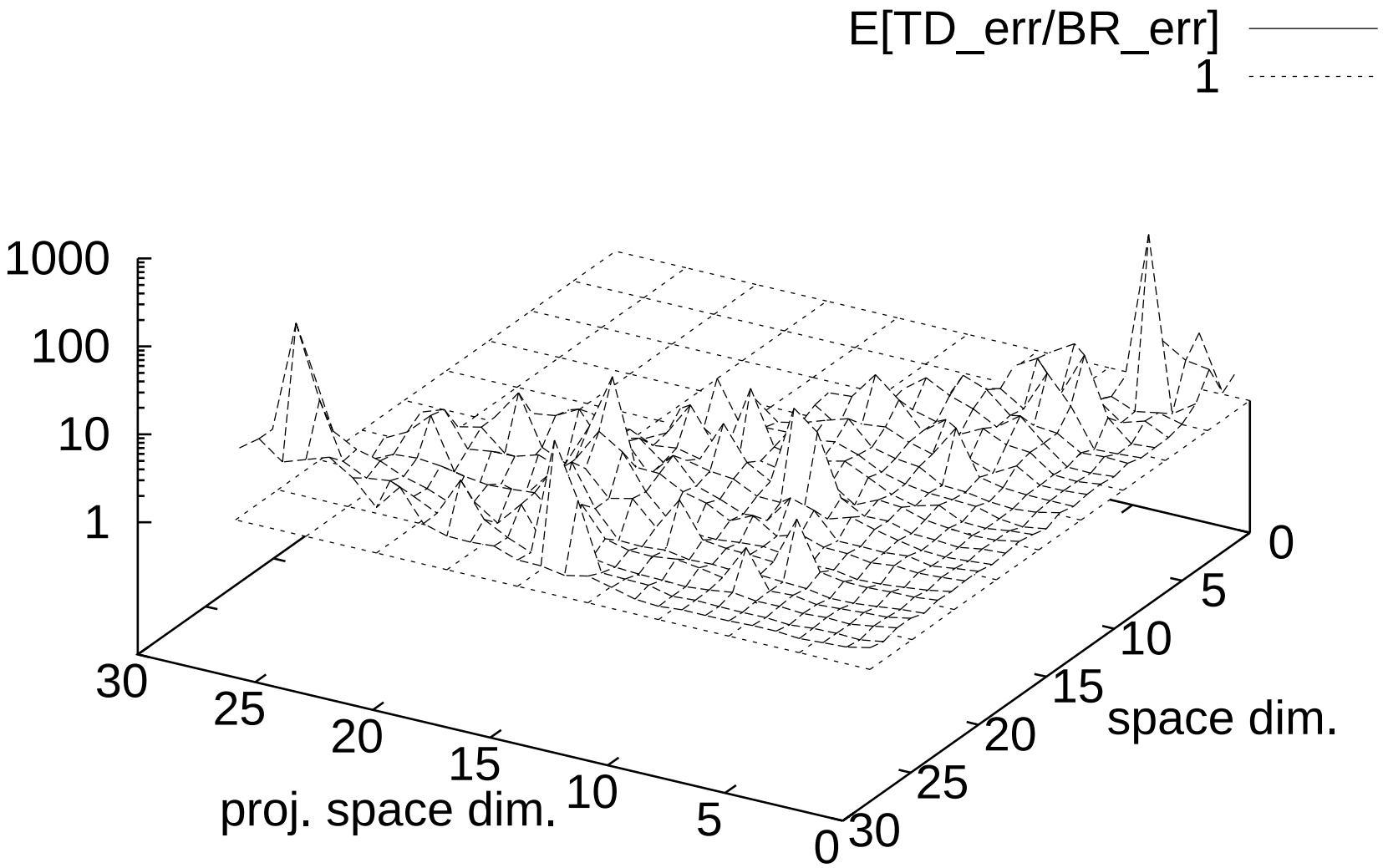


- Spaces of dimension $n = 2, 3, ..., 30$
- Projections of dimension $k = 1, 2, ..., n$
- For each (n, k)
  - 20 random projections (Φ, ξ) $\times$ 20 random MDPs $(r,\ p_i)$
  - $\Rightarrow$ 400 data points $(v, \hat{v}_{best}, \hat{v}_{TD}, \hat{v}_{BR}, b_{TD}, b_{BR})$

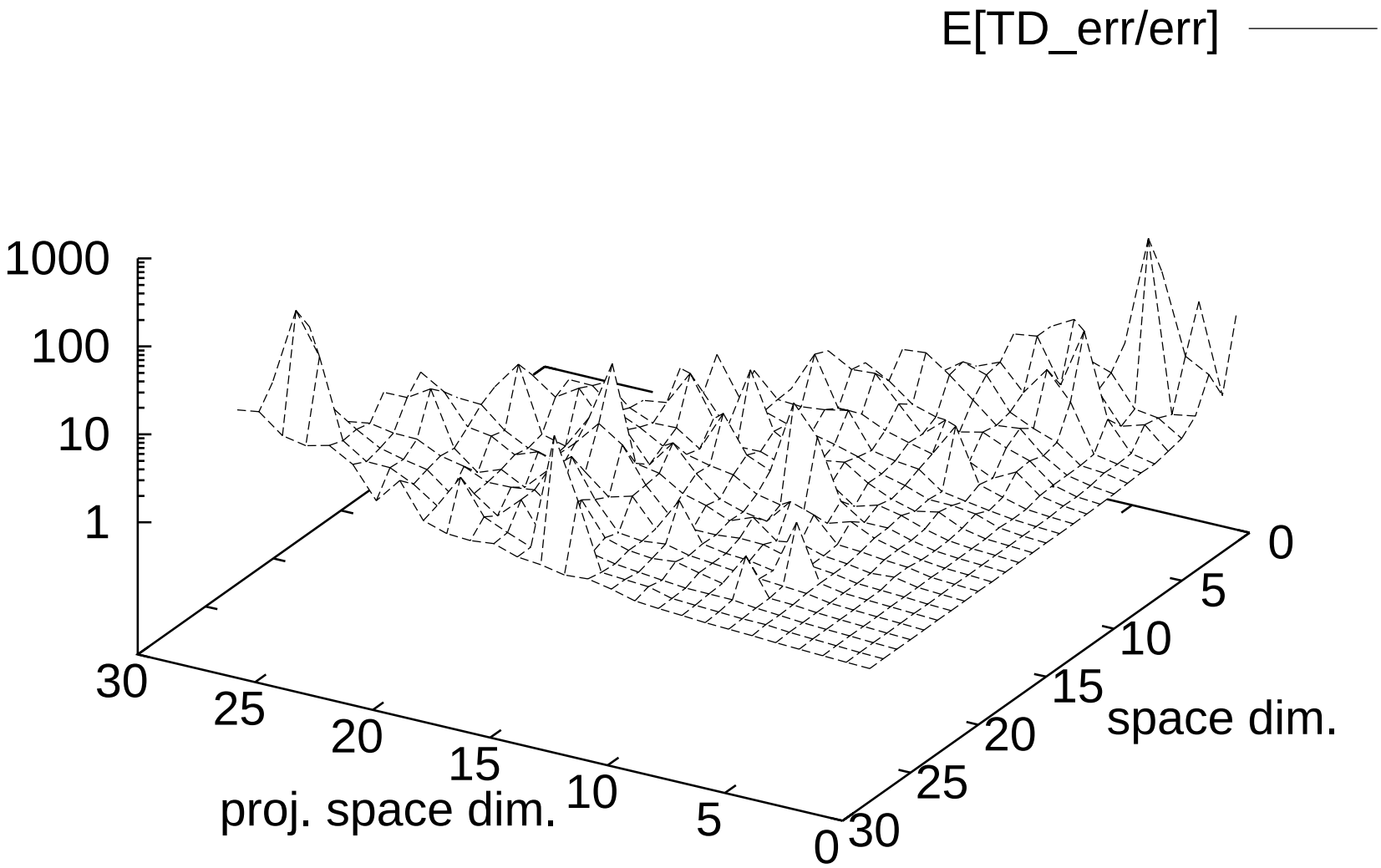
TD win ratio


E[TD wins]
0.5
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
30
25
20
15
10
5
0
proj. space dim.
0
5
10
15
20
25
30
space dim.

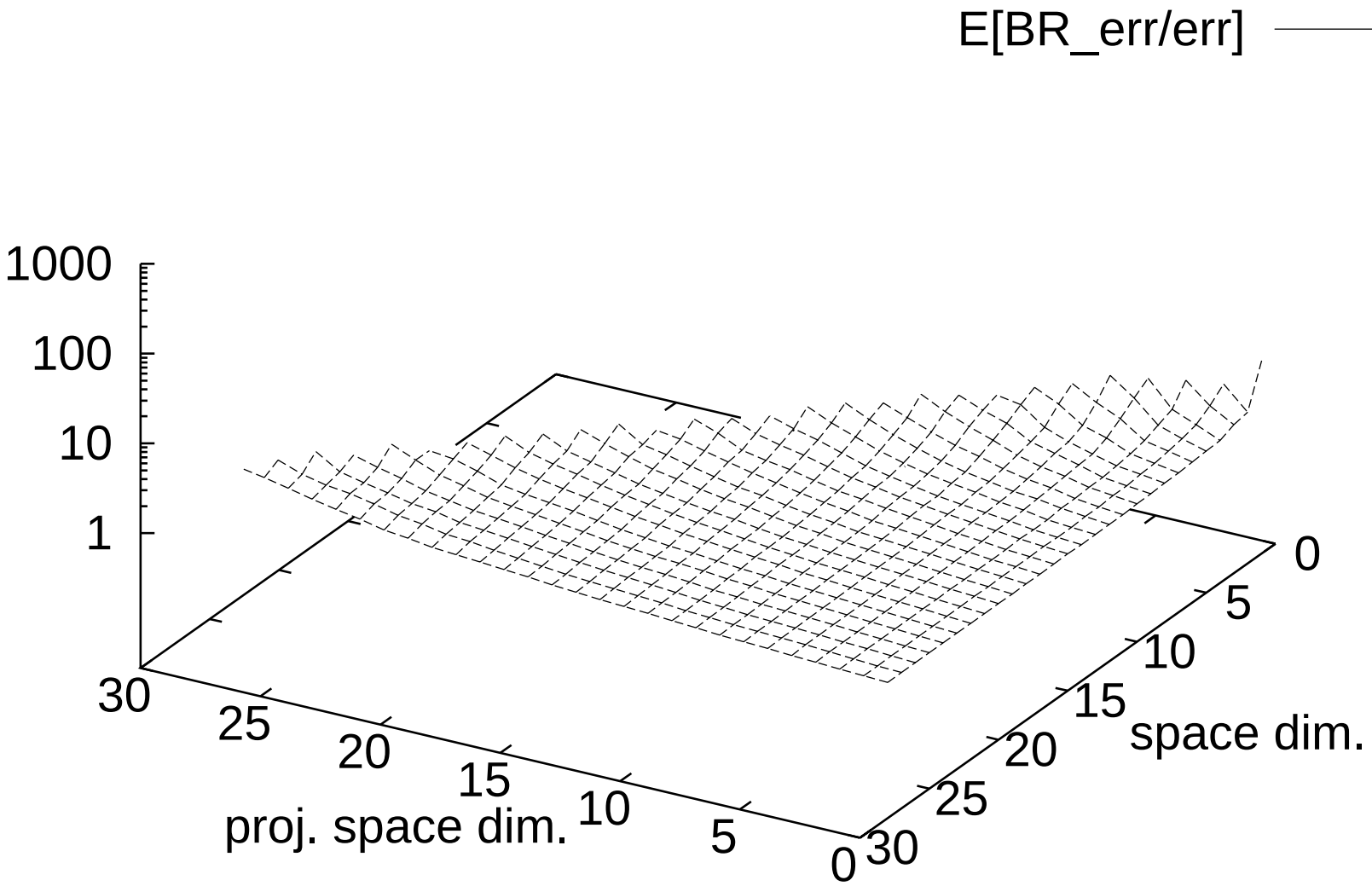
Expectation of $e(\hat{V}_{TD})/e(\hat{V}_{BR})$



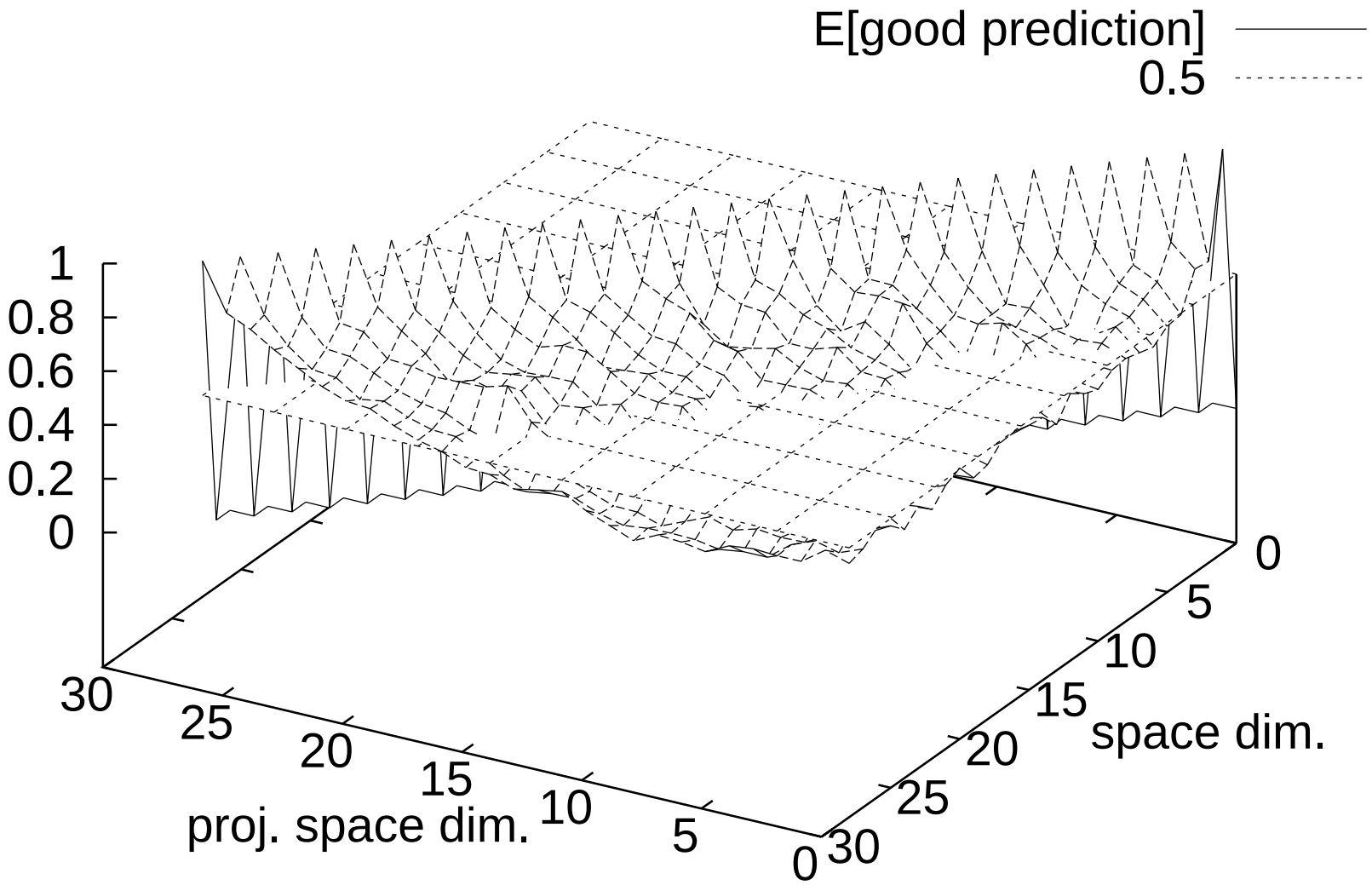
Expectation of $e(\hat{V}_{TD})$



Expectation of $e(\hat{V}_{BR})$



Prediction of the best method through bounds



Conclusion and future work

- TD fix point and BR methods
- Two examples, where each method outperforms the other
- A unified view in terms of oblique projection :
  - Both methods solve a fix point equation (new for BR)
  - Both amounts to do an oblique projection of v
  - Related to/extends (Schoknecht, 2002) and (Yu & Bertsekas, 2008)

Conclusion and future work

- Which method ?
  - BR is *sounder* than TD
  - Extensive simulations suggest :
    - TD is *more often* better than BR
    - Sometimes, TD fails dramatically
    - BR is better *on average*
  - Some reasons to use TD :
    - When sampling, BR requires double samples
    - TD(λ) with λ big enough solves the stability issue
- Future work :
  - TD(λ) vs BR(λ) for $\lambda > 0$
  - New to RL, the idea of oblique projection is well established in the Numerical Analysis community (Saad, 2003)

References I

Antos, A., Szepesvári, C., & Munos, R. 2008.
Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path.
Machine Learning, 71(1), 89–129.

Bertsekas, D.P., & Tsitsiklis, J.N. 1996.
Neurodynamic Programming.
Athena Scientific.

Farahmand, A.M., Ghavamzadeh, M., Szepesvári, C., & Mannor, S. 2008.
Regularized Policy Iteration.
In : NIPS.

Munos, R. 2003.
Error Bounds for Approximate Policy Iteration.
In : ICML.

Puterman, M. 1994.
Markov Decision Processes.
Wiley, New York.

Saad, Y. 2003.
Iterative Methods for Sparse Linear Systems, 2nd edition.
Philadelpha, PA : SIAM.

References II

Schoknecht, R. 2002.
Optimality of Reinforcement Learning Algorithms with Linear Function Approximation.
Pages 1555–1562 of : NIPS.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., & Wiewiora, E. 2009.
Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation.
In : ICML.

Sutton, R.S., & Barto, A.G. 1998.
Reinforcement Learning, An introduction.
BradFord Book. The MIT Press.

Williams, R. J., & Baird, L. C. 1993.
Tight performance bounds on greedy policies based on imperfect value functions.
Tech. rept. College of Computer Science, Northeastern University.

Yu, H., & Bertsekas, D.P. 2008 (July).
New Error Bounds for Approximations from Projected Linear Equations.
Tech. rept. C-2008-43. Dept. Computer Science, Univ. of Helsinki.