Projected Bellman equations (Track 3)

Shruti Bhanderi, 260724575, McGill University 15/03/2018

1 Question 1 (Oblique Projection)

1.1 What is oblique projection?

Let orthogonal projection be Π and the Bellman operator be T. Then we can consider the fixed point of the composition of Π and T. Let span(ϕ) be linear operator that satisfies $\Pi^2 = \Pi$ and whose range is $span(\Phi)$. We can consider using a projection Π which is not necessarily orthogonal onto span(ϕ). Then, oblique projection tuple (Φ, X) is defined as below,

The *oblique projection* Π_X is the projection defined as

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

where X is projection direction, and Π_X is the projection orthogonal to span(X) onto span (Φ).

1.2 Show that the projected Bellman operator gives rise to an oblique projection.

The advantage of the oblique projection viewpoint is that it enables extending the finite-sample analysis to Bellman Residual, Fixed Point, and other least-squares methods in a unified manner.

For the fixed point TD, solution can be unified under the framework of oblique projections with below properties, which will show us how projected Bellman operator gives rise to an oblique projection.

As proved in Preposition 2[Scherrer, 2010], the projected equation is,

$$\hat{v}_X = \Pi_X T\hat{v} = \Pi_{(I-\gamma F)'X}\hat{v}, \text{where } X = \Xi(I - \delta\gamma F)\Phi \tag{1}$$

Here, the fixed point solution $\hat{v}$ of equation $Z = \Pi_X TZ$. Secondly, the oblique projection of v onto subspace $span(\Phi)$ orthogonal to $span((I-\gamma F)'X)$, as in above equation $\hat{v}_X = \Pi_{(I-\gamma F)'X}\hat{v}$, where v is the solution of Bellman equation $Z = TZ$, and X specifies the direction of the projection as mentioned in **Section 1.1**, T is Bellman operator, F is state transition function, γ is discount factor, and let δ be the coefficient which controls the direction of the oblique projection. Hence, shown in equation (1) the direction of the oblique projection δ depends on the projected Bellman operator. We can see that by the change in the value of δ which will rise according to the projection of the Bellman operator. Initially, setting $\delta = 0$ and solving the above equations with projected Bellman operator gives rise to the oblique projection and results in $\delta = 1$.

1.3 How do you think the oblique projection changes with lambda ?

In $TD(\lambda)$, the eligibility trace vector is initialized to zero at the beginning of the episode, is incremented on each time step by the value gradient, and then fades away by $\gamma\lambda$,

$$z_{-1} \doteq 0,$$

$$z_t \doteq \gamma\lambda z_{t-1} + \Delta\hat{v}(S_t, w_t),$$

Here, for linear function approximation, $\Delta\hat{v}(S_t, w_t)$ is the feature vector, x_t, in which case the eligibility trace vector is just a sum of past, fading, input vectors. Considering the above definitions and **Section 1.2**, we can relate λ parameter to the projected equations. Talking about geometric representation, as λ will be included in the term $(I - \gamma F)'X$, the oblique projection of v onto subspace $span(\Phi)$ will be orthogonal to $span((I - \gamma F)'X)$ and hence to λ too.

The paper [Scherrer, 2010] discusses the oblique projection for $\lambda = 0$ only. I think for the values $\lambda > 0$, might solve the weakness of TD(0) and can reduce the stability issues for λ close to 1. In these case, for example $TD(\lambda)$ and $LSTD(\lambda)$ with $\lambda > 0$, even infinite asymptotic estimation variance can occur, despite the almost sure convergence of the algorithm. Hence, it makes the projected Bellman Equations to converge to the optimal value. And with the change in the value of λ to the high as $\lambda = 1$ will obtain the optimal projection which is the best.

2 Question 2 (Galerkin's method)

2.1 Explain what is Galerkin's method.

Galerkin's method is a method for converting continuous operator problem to a discrete problem. In principle, it is the equivalent of applying the method of variation of parameters to a function space, by converting the equation to a weak formulation. We then apply some constraints on the function space to characterize the space with a finite set of basis functions. Galerkin's method provides powerful numerical solution to differential equations and modal analysis. The definition of Galerkin's method for the most relevant abstract problem in weak formulation on a Hilbert space S says that find $x \in S$ such that for all $y \in S, a(x, y) = f(y)$, where $a(x, y)$ is a bilinear form and $f(y)$ is a bounded linear functional on S.

2.2 Explain why the projected Bellman operator can be interpreted as a Galerkin's method.

In the context of least square temporal difference learning, the value function V^* is the solution to the fixed point equation:

$$TV^* = V^*, \tag{2}$$

where T is *Bellman operator*:

$$TV(x) = \int (r + \gamma V(x')dP(x', r|x)) \tag{3}$$

As it can be seen from the definition, T is an affine linear operator, so Equation (3) is a linear fixed point equation. Also, T is a bounded and is in fact a γ-contraction. The projection method for solving above equation consists of choosing two finite dimensional linear subspaces F,G of a Banach space B of functions over X, F,G sharing a common dimension $d \in N$, and then solving the projected equation [Szepesvári, 2011]

$$\Pi_G TV = \Pi_G V$$

for $V \in F$, where $\Pi_G : B \to X$ is a projection operator as described in **Section 1**. It can be easily seen that this leads to a $d \times d$ linear system of equations once one fixes some bases for F and G. When B is a pre-Hilbert space, Π_G is the corresponding orthogonal projection and $T : F \to B$ is bounded, and

this is defined as Galerkin method in the context of this system. Hence, above explanation shows that the projected Bellman operator can be interpreted as a Galerkin's method.

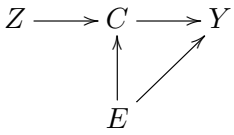
3 Question 3(Instrumental Variables Regression)

3.1 Explain what is instrumental variables regression.

An instrumental variable is defined as a variable Z that is correlated with the independent variable C and uncorrelated with the "error term" E in the linear equation,

$$Y = C\beta + E \tag{4}$$

More precisely, there exists an instrument variable Z that has the property whose changes associated with C but do not led to direct change in Y (indirect via C). This introduces the instrument variable Z which is associated with C but not with error E as shown in below path diagram.



It is still the case that Z and Y will be correlated, but the only source of such correlation is the indirect path of Z being correlated with C which in turn determines Y. The more direct path of Z being a regressor in the model for Y is ruled out.

3.2 Explain why LSTD can be interpreted as instrumental variables regression.

The TD algorithm used with a linear in the parameters function approximator addresses the problem of finding a parameter vector, θ^* that allows us to compute the value of a state x as $V(x) = \phi'_x\theta^*$.The value function satisfies the following consistency condition[Bradtke and Barto, 1996].

$$V(x) = r_x + \gamma \sum_{y \in X} P(x,y)V(y) \tag{5}$$

where r_x the expected immediate reward for any state transition from state x. We can rewrite this as,

$$r_x = V(x) - \gamma \sum_{y \in X} P(x,y)V(y) \tag{6}$$

$$= \phi'_x\theta^* - \gamma \sum_{y \in X} P(x,y)\phi'_y\theta^* \tag{7}$$

$$= (\phi_x - \gamma \sum_{y \in X} P(x,y)\phi_y)'\theta^* \tag{8}$$

for every state $x \in X$. Now, the problem with this is that substituting $\bar{r}_t$ directly for r_t has the effect of introducing the noise that is dependent upon the current state. This introduces a bias, and θ_t no longer converges to θ^*. For solving this problem least squares introduce instrumental variables. An instrumental variable, ρ_t, is a vector that is correlated with the true input vectors, but that is uncorrelated with the

observation noise. The scalar output,r_x, is the inner product of an input vector, $\phi_x - \gamma \sum_{y \in X} P(x,y)\phi_y$, and the true parameter vector, θ^*. For each time step t, we have following,

$$r_t = (\phi_x - \gamma \sum_{y \in X} P(x,y)\phi_y)'\theta^* + (r_t - \bar{r}_t) \tag{9}$$

where r_t is the reward received on the transition from x_t to x_{t+1}. $(r_t - \bar{r}_t)$ corresponds to the noise term mentioned above. This establishes that the noise term has zero mean and is uncorrelated with the input vector.

References

[Bradtke and Barto, 1996] Bradtke, S. J. and Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3):33–57.

[Scherrer, 2010] Scherrer, B. (2010). Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. *arXiv preprint arXiv:1011.4362*.

[Szepesvári, 2011] Szepesvári, C. (2011). Least squares temporal difference learning and galerkin's method.