

## Addendum

In [1], we proposed two hybrid least-squares algorithms for approximate policy evaluation. These algorithms, which were referred to as  $H_1$  and  $H_2$ , combine the optimization criterion of two least-squares algorithms: the Bellman residual method (which minimizes the Bellman residual) and the fixed point method (which minimizes the projection of the Bellman residual). Algorithm  $H_1$  was well motivated, but we proposed  $H_2$  in a more ad-hoc manner. Here we show that  $H_2$  actually has a much more principled foundation. The derivation of  $H_2$  is inspired by Kolter and Ng’s recent paper [2].

## Derivation of Hybrid Algorithm $H_2$

We use the following terminology:  $P^\pi$  is a transition matrix encoding the effects of policy  $\pi$ ,  $R^\pi$  is the reward function,  $\gamma \in [0, 1)$  is a discount factor,  $T^\pi(x) \equiv R^\pi + \gamma P^\pi x$  is the Bellman operator,  $\rho$  is a distribution over states,  $\Phi$  is a basis function matrix, and  $\beta \in [0, 1]$  is a parameter trading off between the Bellman residual method ( $\beta = 1$ ) and the fixed point method ( $\beta = 0$ ). An approximate value function  $\hat{V} = \Phi w$  linearly combines the basis functions in  $\Phi$  with an adjustable vector of coefficients,  $w$ . For a complete description of the terminology, please see the full paper [1].

Following [2], we introduce the function  $f(w)$ :

$$f(w) = \underset{u}{\operatorname{argmin}} \left[ \frac{\beta}{2} \|T^\pi(\Phi u) - \Phi u\|_\rho^2 + \frac{1 - \beta}{2} \|T^\pi(\Phi w) - \Phi u\|_\rho^2 \right].$$

The first norm in  $f(w)$  is the optimization criterion for the Bellman residual method and the second norm is the criterion for the fixed point method. Given a vector  $w$ ,  $f(w)$  returns the argument  $u$  that minimizes the combination of the two norms. The approximate policy evaluation algorithms attempt to find a vector  $w$  such that  $w = f(w)$ .

Differentiating  $f(w)$  with respect to  $u$ , we obtain:

$$\begin{aligned} \frac{\partial f(w)}{\partial u} &= \beta \left[ \frac{\partial}{\partial u} (T^\pi(\Phi u) - \Phi u) \right]^T D_\rho (T^\pi(\Phi u) - \Phi u) + \\ &\quad (1 - \beta) \left[ \frac{\partial}{\partial u} (T^\pi(\Phi w) - \Phi u) \right]^T D_\rho (T^\pi(\Phi w) - \Phi u) \\ &= \beta (\gamma P^\pi \Phi - \Phi)^T D_\rho (T^\pi(\Phi u) - \Phi u) + (1 - \beta) (-\Phi)^T D_\rho (T^\pi(\Phi w) - \Phi u) \\ &= \beta (\Phi - \gamma P^\pi \Phi)^T D_\rho (\Phi u - \gamma P^\pi \Phi u - R^\pi) + (1 - \beta) \Phi^T D_\rho (\Phi u - \gamma P^\pi \Phi w - R^\pi) \end{aligned}$$

where  $D_\rho$  is a diagonal matrix with elements  $\rho$ .

To find an extrema, we set  $\frac{\partial f(w)}{\partial u}$  to 0 and solve for  $u$ :

$$\begin{aligned} \left[ \beta (\Phi - \gamma P^\pi \Phi)^T D_\rho (\Phi - \gamma P^\pi \Phi) + (1 - \beta) \Phi^T D_\rho \Phi \right] u = \\ \left[ \beta (\Phi - \gamma P^\pi \Phi)^T D_\rho + (1 - \beta) \Phi^T D_\rho \right] R^\pi + (1 - \beta) \Phi^T D_\rho (\gamma P^\pi \Phi) w. \end{aligned}$$

By adding and subtracting  $(1 - \beta)\Phi^T D_\rho(\gamma P^\pi \Phi)u$  from the left-hand side of the equation, we get:

$$\begin{aligned} \left[ \beta (\Phi - \gamma P^\pi \Phi)^T D_\rho (\Phi - \gamma P^\pi \Phi) + (1 - \beta)\Phi^T D_\rho (\Phi - \gamma P^\pi \Phi) + (1 - \beta)\Phi^T D_\rho (\gamma P^\pi \Phi) \right] u = \\ \left[ \beta (\Phi - \gamma P^\pi \Phi)^T D_\rho + (1 - \beta)\Phi^T D_\rho \right] R^\pi + (1 - \beta)\Phi^T D_\rho (\gamma P^\pi \Phi)w. \end{aligned}$$

The equation can be simplified using  $\beta (\Phi - \gamma P^\pi \Phi) + (1 - \beta)\Phi = (\Phi - \beta\gamma P^\pi \Phi)$ :

$$\begin{aligned} \left[ (\Phi - \beta\gamma P^\pi \Phi)^T D_\rho (\Phi - \gamma P^\pi \Phi) + (1 - \beta)\Phi^T D_\rho (\gamma P^\pi \Phi) \right] u = \\ (\Phi - \beta\gamma P^\pi \Phi)^T D_\rho R^\pi + (1 - \beta)\Phi^T D_\rho (\gamma P^\pi \Phi)w. \end{aligned}$$

Finally, we enforce  $w = f(w) = u$  to get the final result:

$$(\Phi - \beta\gamma P^\pi \Phi)^T D_\rho (\Phi - \gamma P^\pi \Phi)w = (\Phi - \beta\gamma P^\pi \Phi)^T D_\rho R^\pi.$$

This least-squares problem coincides with the  $H_2$  least-squares problem  $A_{H_2}w = b_{H_2}$  where:

$$\begin{aligned} A_{H_2} &= (\Phi - \beta\gamma P^\pi \Phi)^T D_\rho (\Phi - \gamma P^\pi \Phi) \\ b_{H_2} &= (\Phi - \beta\gamma P^\pi \Phi)^T D_\rho R^\pi. \end{aligned}$$

## References

- [1] J. Johns, M. Petrik, and S. Mahadevan. Hybrid least-squares algorithms for approximate policy evaluation. *Machine Learning*, 76(2):243–256, 2009.
- [2] J. Kolter and A. Ng. Regularization and feature selection in least-squares temporal difference learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 521–528, 2009.