Policy Iteration and Newton's Method

COMP 767 – Reinforcement Learning

Michael Noseworthy

Review

- Policy Iteration
  - Policy Evaluation

$$v^n = r_{d_n} + \lambda P_{d_n} v^n$$
$$v^n = (I - \lambda P_{d_n})^{-1} r_{d_n}$$

  - Policy Improvement

$$d_{n+1} \in \operatorname*{argmax}_{d \in D} \{ r_d + \lambda P_d v^n \}$$

  - Stop when the policy doesn’t change.

Some Notation

- Bellman Equation

$$Lv \equiv \max_{d \in D} \{r_d + \lambda P_d v\}$$

$$Lv = v$$

- Improvement Operator

$$\begin{aligned} Bv &\equiv \max_{d \in D} \{r_d + (\lambda P_d - I)v\} \\ &= \max_{d \in D} \{r_d + \lambda P_d v\} - v \\ &= Lv - v \end{aligned}$$

- Takeaway: $Lv = v \Longleftrightarrow Bv = 0$

Newton's Method

- Iterative method for finding a zero

$$x_{n+1} = x_n - \left[f'(x_n)\right]^{-1} f(x_n)$$

Policy Iteration (1)

- Has the same form as Newton's method!

$$
\begin{aligned}
v^{n+1} &= (I - \lambda P_{d_{n+1}})^{-1} r_{d_{n+1}} && \textit{Policy Evaluation} \\
&= (I - \lambda P_{d_{n+1}})^{-1} r_{d_{n+1}} - v^n + v^n \\
&= (I - \lambda P_{d_{n+1}})^{-1} \left[r_{d_{n+1}} + (\lambda P_{d_{n+1}} - I) v^n \right] + v^n && \textit{Factorization} \\
&= v^n - (\lambda P_{d_{n+1}} - I)^{-1} B v^n && \textit{Definition of B}
\end{aligned}
$$

- Generalization of Newton's method for Operator Equations

Policy Iteration (1)

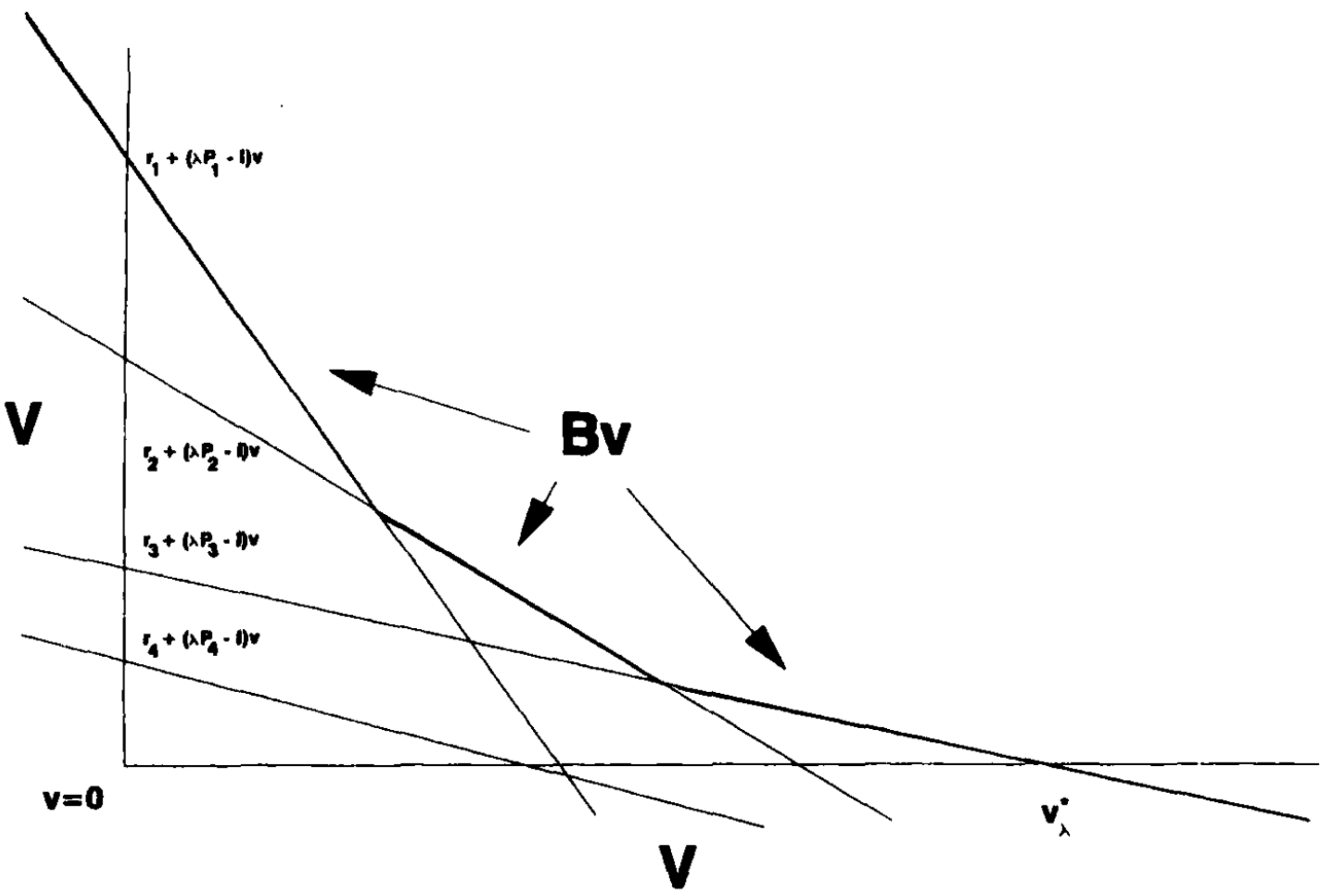
- Has the same form as Newton's method!

$$
\begin{aligned}
\boxed{v^{n+1}} &= (I - \lambda P_{d_{n+1}})^{-1} r_{d_{n+1}} && \textit{Policy Evaluation} \\
&= (I - \lambda P_{d_{n+1}})^{-1} r_{d_{n+1}} - v^n + v^n \\
&= (I - \lambda P_{d_{n+1}})^{-1} \left[r_{d_{n+1}} + (\lambda P_{d_{n+1}} - I) v^n \right] + v^n && \textit{Factorization} \\
&= \boxed{v^n} - \boxed{(\lambda P_{d_{n+1}} - I)^{-1}} \boxed{B v^n} && \textit{Definition of B}
\end{aligned}
$$

$$
\boxed{x_{n+1}} = \boxed{x_n} - \boxed{[f'(x_n)]^{-1}} \boxed{f(x_n)}
$$

- Generalization of Newton's method for Operator Equations

Geometry



$r_1 + (\lambda P_1 - I)v$
$r_2 + (\lambda P_2 - I)v$
$r_3 + (\lambda P_3 - I)v$
$r_4 + (\lambda P_4 - I)v$
V
Bv
v=0
V
v^*_λ

Policy Iteration (2)

- Newton's Method: If f is convex and a zero exists, we will find that zero (if we start at a non-negative value).

$$Bu \geq r_{d_v} + (\lambda P_{d_v} - I)u$$
$$Bv = r_{d_v} + (\lambda P_{d_v} - I)v$$

$$Bu \geq Bv + (\lambda P_{d_v} - I)(u - v)$$

Policy Iteration (2)

- Newton's Method: If f is convex and a zero exists, we will find that zero (if we start at a non-negative value).

$$Bu \geq r_{d_v} + (\lambda P_{d_v} - I)u$$

$$Bv = r_{d_v} + (\lambda P_{d_v} - I)v$$

$$\boxed{\begin{aligned} Bu &\geq Bv + (\lambda P_{d_v} - I)(u - v) \\ f(x) &\geq f(y) + f'(y)(x - y) \end{aligned}}$$

Other Interesting Results

- Like Newton's Method, under certain conditions, convergence will be quadratic *(A compact and convex, P affine in a, reward concave in a)*
- Iterates are bounded below by value iteration and above by the optimal value.

- *These results hold for compact action spaces.*
- *More details in Puterman (1994) - 6.4.2*