IV. Approximate Dynamic Programming (ADP)

In this section, we arrive at methods for online adaptive optimal control, i.e. RL using data measured along state trajectories
We will use supervised learning, which can apply to model-based or model-free algos.
We call these methods approx. dynamic programming (ADP) [Werbos 1991, 1992] or "neuro DP" (NDP) [Bertsekas 1996]

We require two concepts:

① the temporal difference error ←

② value function approx.

IV.A. Temporal Difference (TD) Error

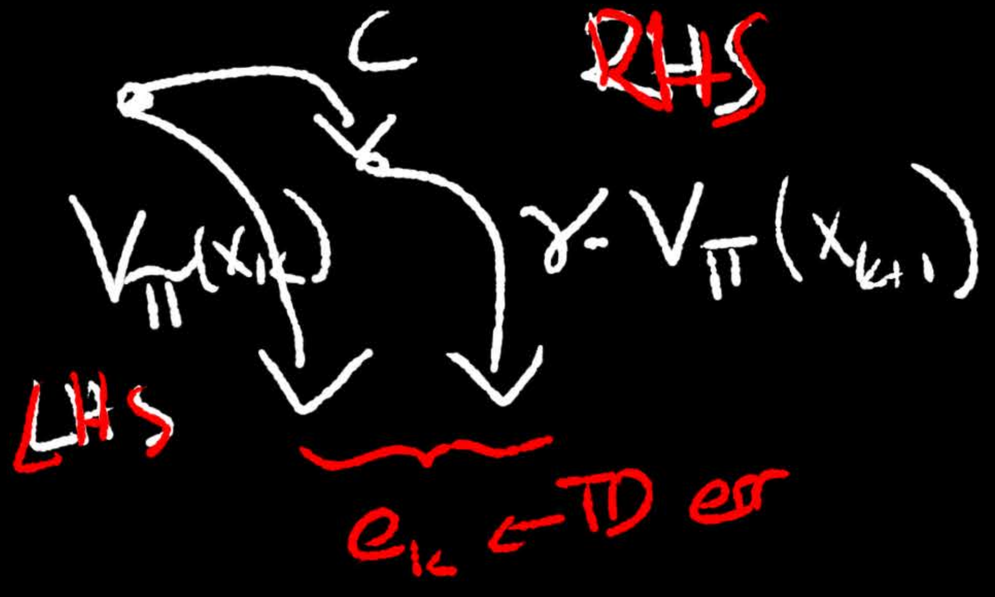
Recall Bellman eqn:

$$\rightarrow V_\pi(x_k) = c(x_k, \pi(x_k)) + \gamma \cdot V_\pi(x_{k+1}) \quad \forall x_k \in X$$

We can interpret this eqn as a consistency eqn

Construct the time-varying residual:

TD err → $e_k = c(x_k, \pi(x_k)) + \gamma \cdot V_\pi(x_{k+1}) - V_\pi(x_k)$



Conceptual visualization of TD error

Suppose at each time step k, we collect data $(x_k, x_{k+1}, c(x_k, \pi(x_k))$. This data can be used to fit a regression model for $V_\pi(\cdot)$ such that it minimizes, e.g. the sum of squared residuals.

IV.B. Value Fcn Approx.

To perform supervised learning on $V_\pi(\cdot)$, we must parameterize it.

Consider Weierstrass higher order approx thm:
There exists a (dense) basis set $\{\phi_i(x)\}$ s.t.

$$V_\pi(x) = \sum_{i=0}^{\infty} w_i \phi_i(x) = \sum_{i=1}^{L} w_i \phi_i(x) + \underbrace{\sum_{i=L+1}^{\infty} w_i \phi_i(x)}_{= \epsilon_L}$$

$$= W^T \phi(x) + \epsilon_L$$

where $W = [w_1, w_2, \cdots, w_L]^T$, $\phi(x) = [\phi_1(x), \phi_2(x), \cdots]$

$\varepsilon_L \to 0$ uniformly in x as $L \to \infty$.

One of main contributions of Werbos & Bertsekas was using this for ADP/NDP.

Idea: $V_\pi(x) \approx \underbrace{W^T}_{\text{params}} \underbrace{\phi(x)}_{\text{basis fcns}}$