PID Accelerated Value Iteration Algorithm

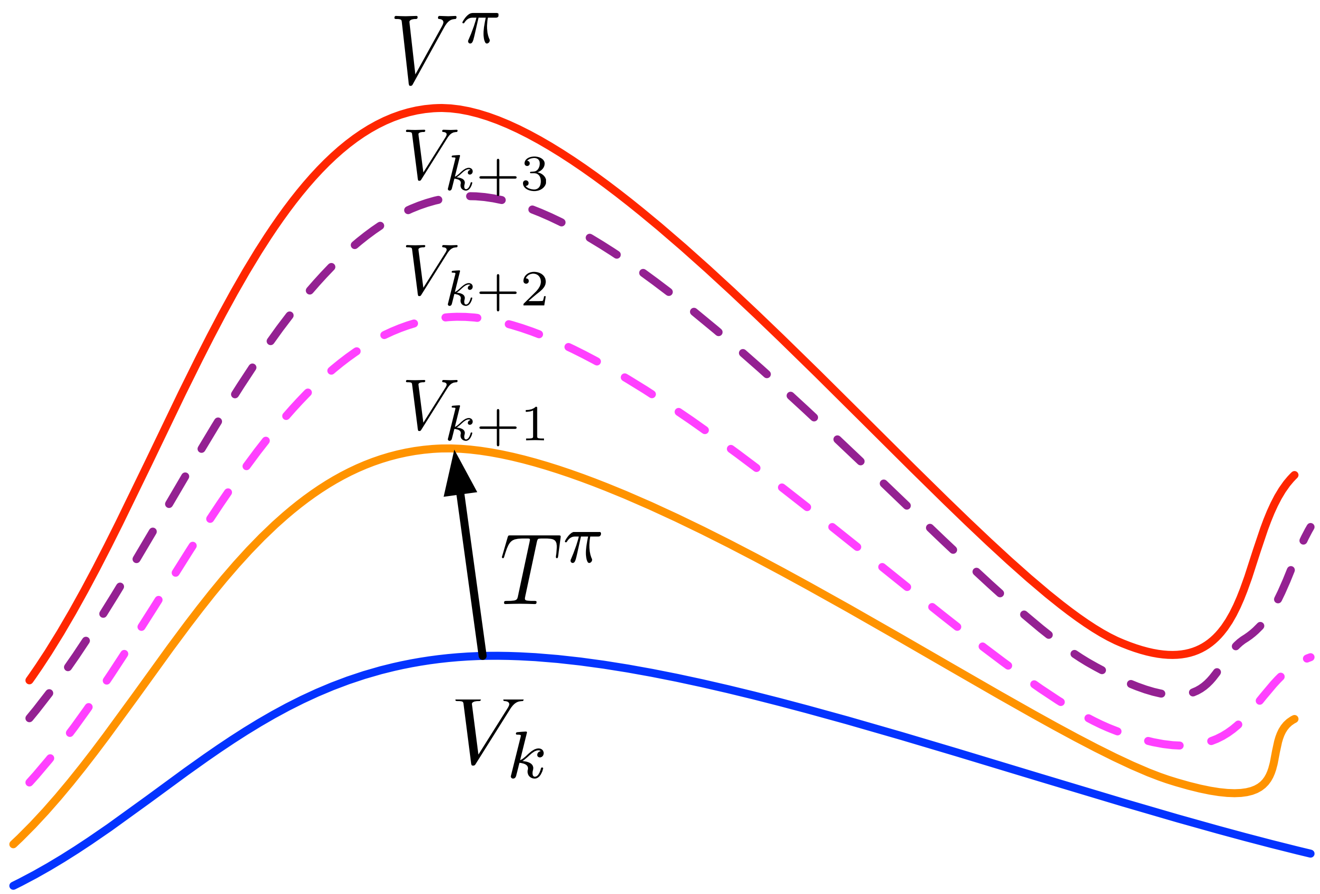
Amir-massoud Farahmand & Mohammad Ghavamzadeh

VECTOR INSTITUTE

Google Research

UNIVERSITY OF TORONTO

Value Iteration



$V_{k+1} = T^\pi V_k$ (policy evaluation)

$V_{k+1} = T^* V_k$ (control)

$$(T^\pi V)(x) \triangleq r^\pi(x) + \gamma \int \mathcal{P}^\pi(\mathrm{d}y|x) V(y),$$

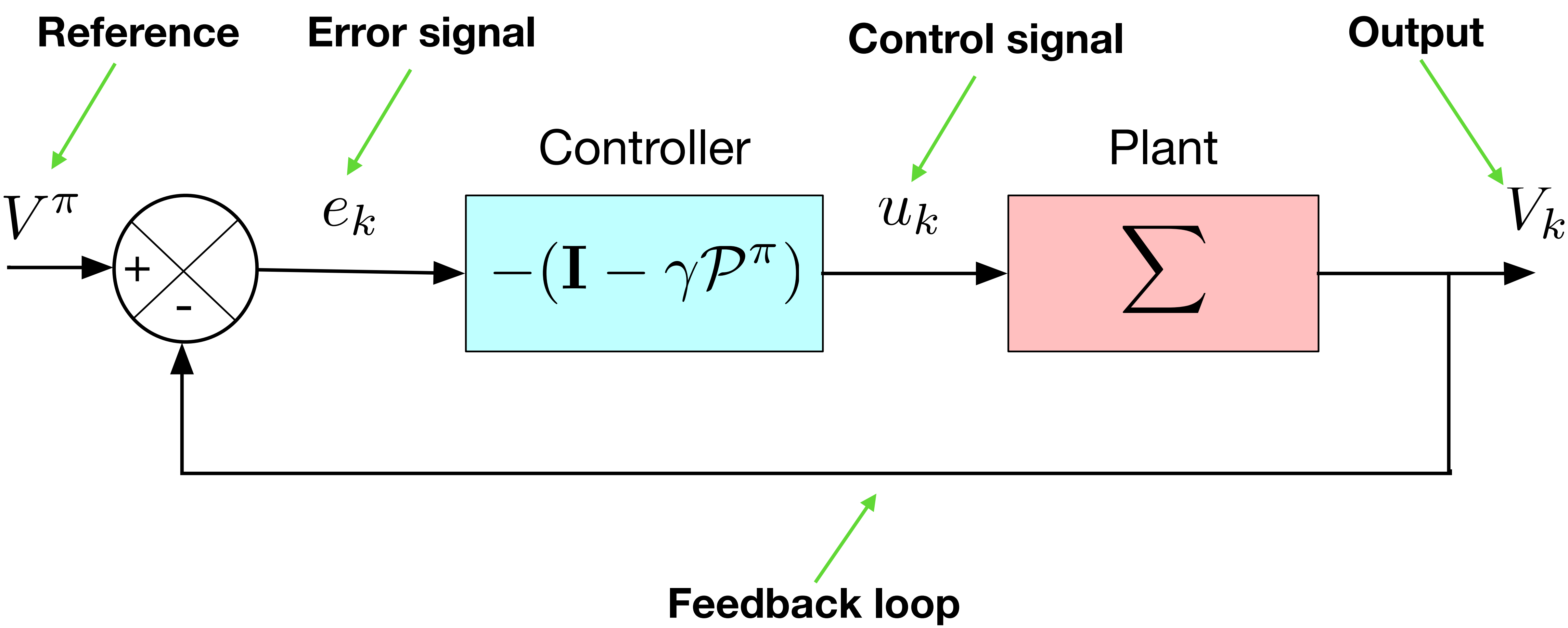
$$(T^* Q)(x, a) \triangleq r(x, a) + \gamma \int \mathcal{P}(\mathrm{d}y|x, a) \max_{a' \in \mathcal{A}} Q(y, a').$$

Value Iteration is Slow

- **Challenge:** The convergence rate is $O(\gamma^k)$ and is slow when $\gamma \approx 1$.
- **Solution:**
  - Interpreting the VI as a dynamical system.
  - Use simple controllers from *control theory* to accelerate it.

VI as a Dynamical System

$$V_{k+1} = T^{\pi} V_k = r^{\pi} + \gamma \mathcal{P}^{\pi} V_k$$



VI as a Proportional Controller

$$
\begin{aligned}
V_{k+1} = T^\pi V_k &= r^\pi + \gamma \mathcal{P}^\pi V_k \\
&= V_k + \mathrm{BR}(V_k) \qquad (\text{Bellman Residual } \mathrm{BR}(V_k) = T^\pi V_k - V_k) \\
&= V_k + u_k \qquad \text{with } u_k = -(\mathbf{I} - \gamma \mathcal{P}^\pi) e_k \text{ and } e_k = V_k - V^\pi
\end{aligned}
$$

Special case of Proportional (P) controller

$$
\begin{aligned}
V_{k+1} &= V_k + \kappa_p \mathrm{BR}(V_k) \qquad (\text{Bellman Residual } \mathrm{BR}(V_k) = T^\pi V_k - V_k) \\
&= V_k + \kappa_p u_k \qquad \text{with } u_k = -(\mathbf{I} - \gamma \mathcal{P}^\pi) e_k \text{ and } e_k = V_k - V^\pi
\end{aligned}
$$

Q: Can we use other controllers to modify (and accelerate) the dynamics?

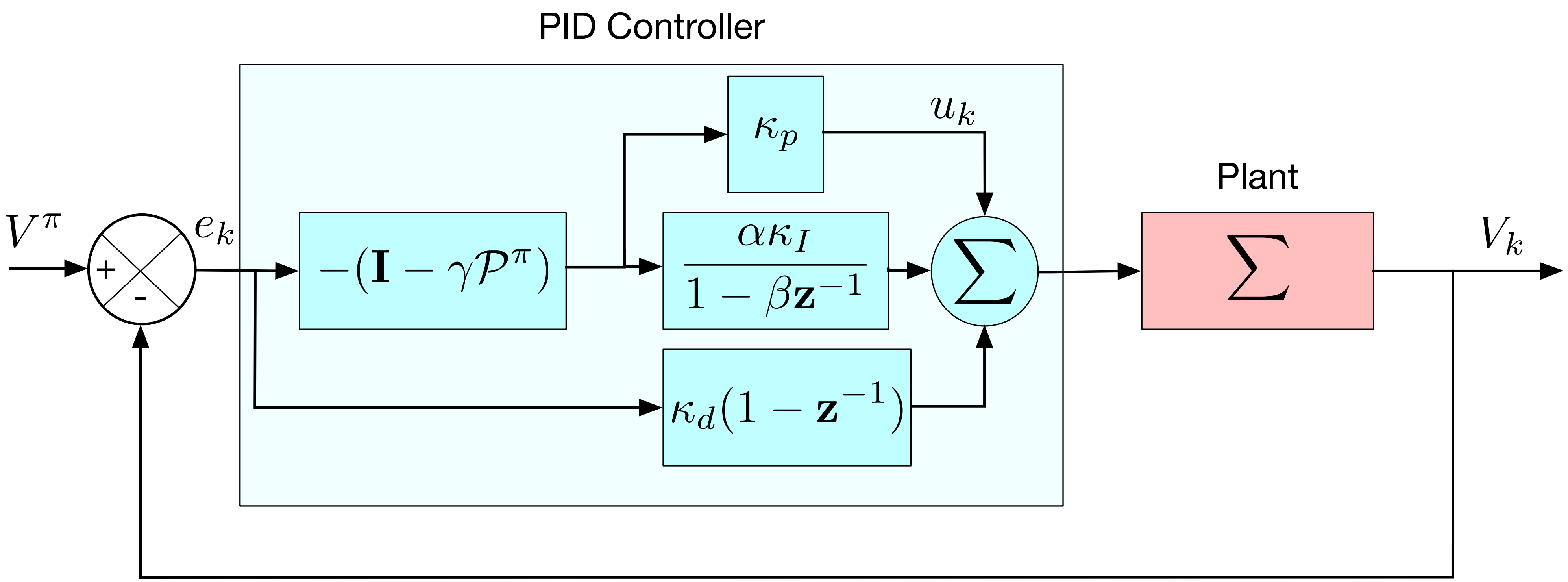
We use Proportional-Derivative-Integral (PID) controllers.

PID Value Iteration

$$z_{k+1} = \beta z_t + \alpha \mathrm{BR}(V_k),$$
$$V_{k+1} = (1 - \kappa_p)V_k + \kappa_p T^\pi V_k + \kappa_I \left[\beta z_k + \alpha \mathrm{BR}(V_k)\right] + \kappa_d(V_k - V_{k-1}).$$

The I term performs a low-pass filter (or smoothing) on the Bellman Residuals.

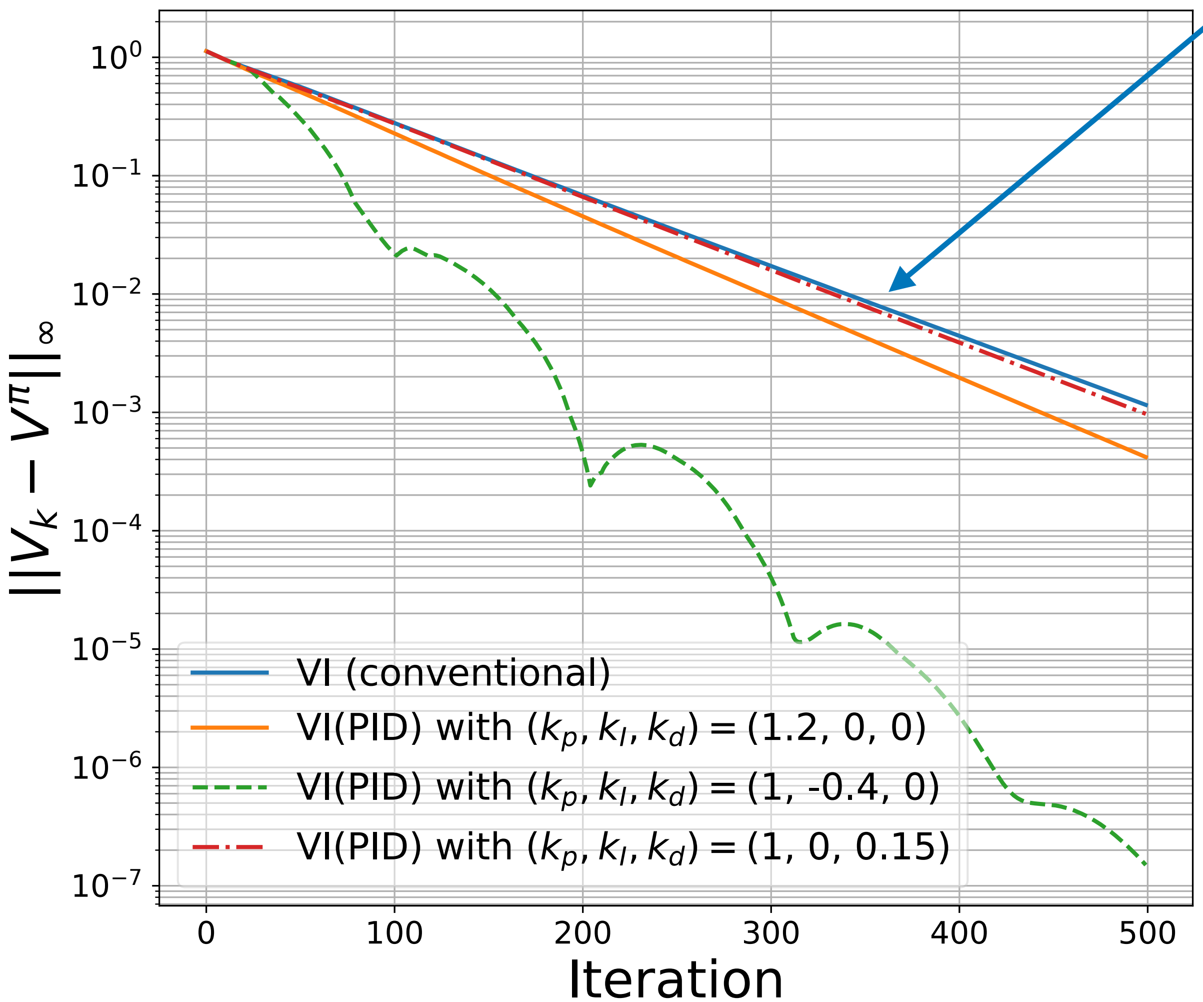
The D term extrapolates based on linear trends



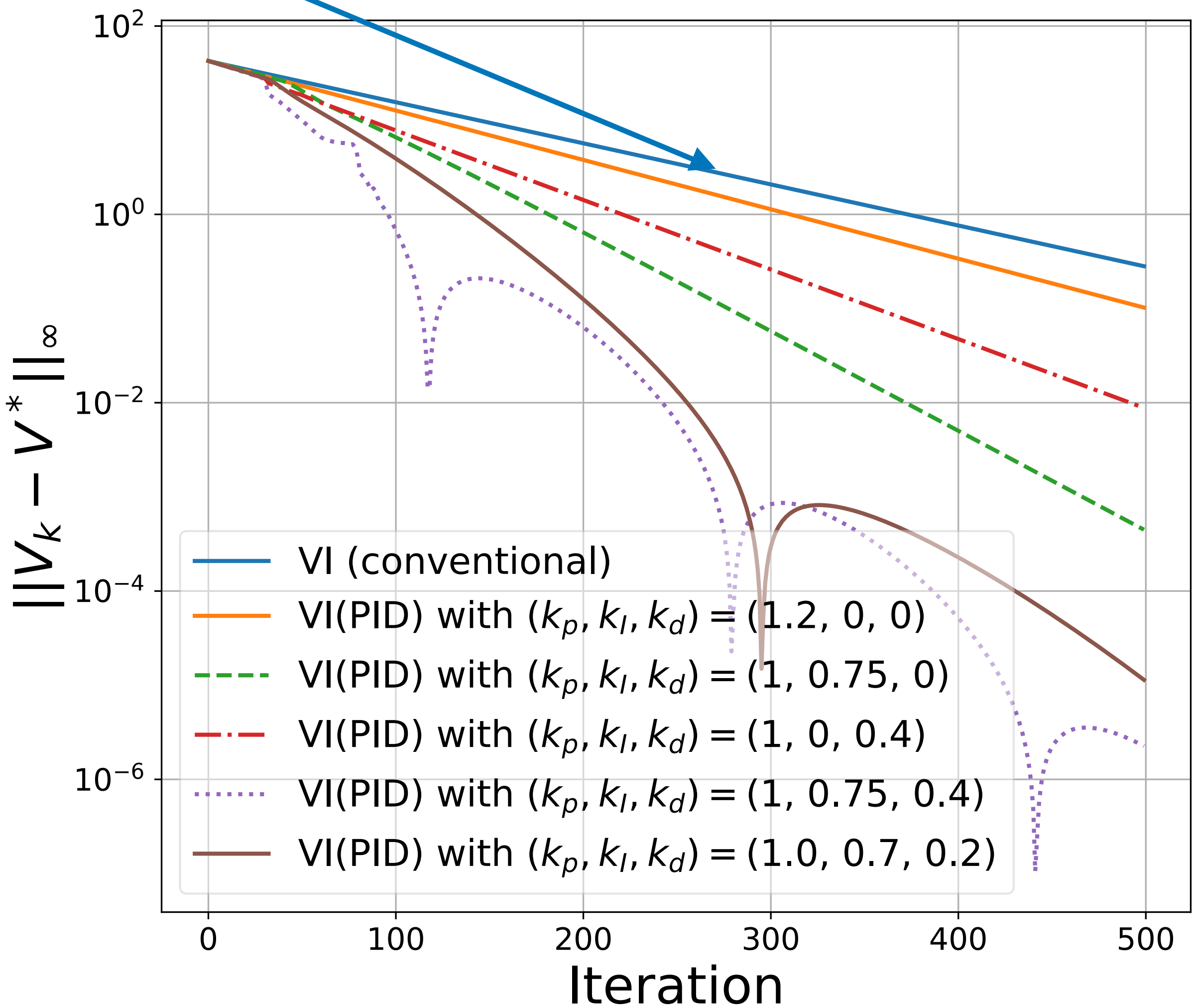
Sample Behaviour

Conventional VI

Policy Evaluation



Control

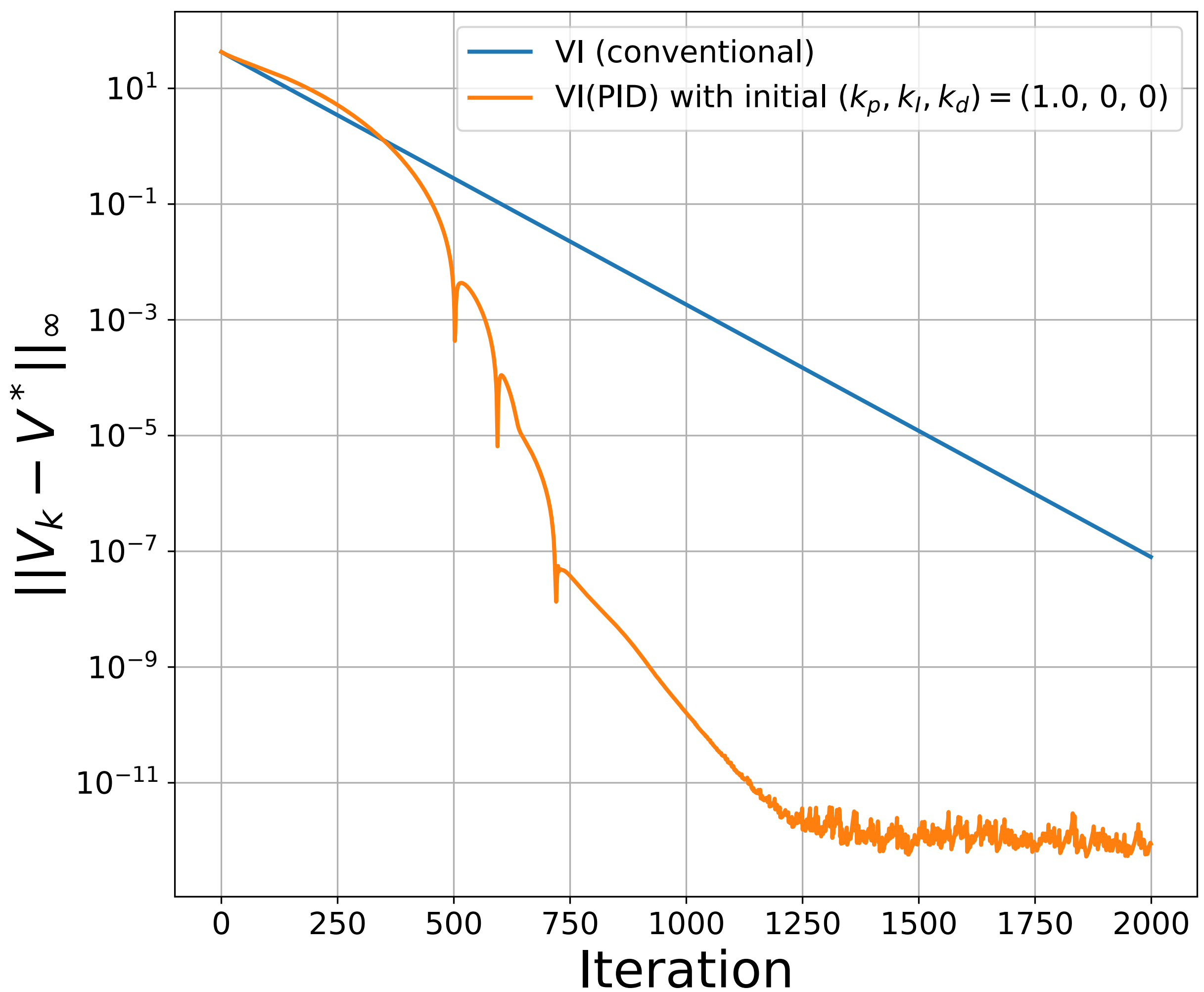


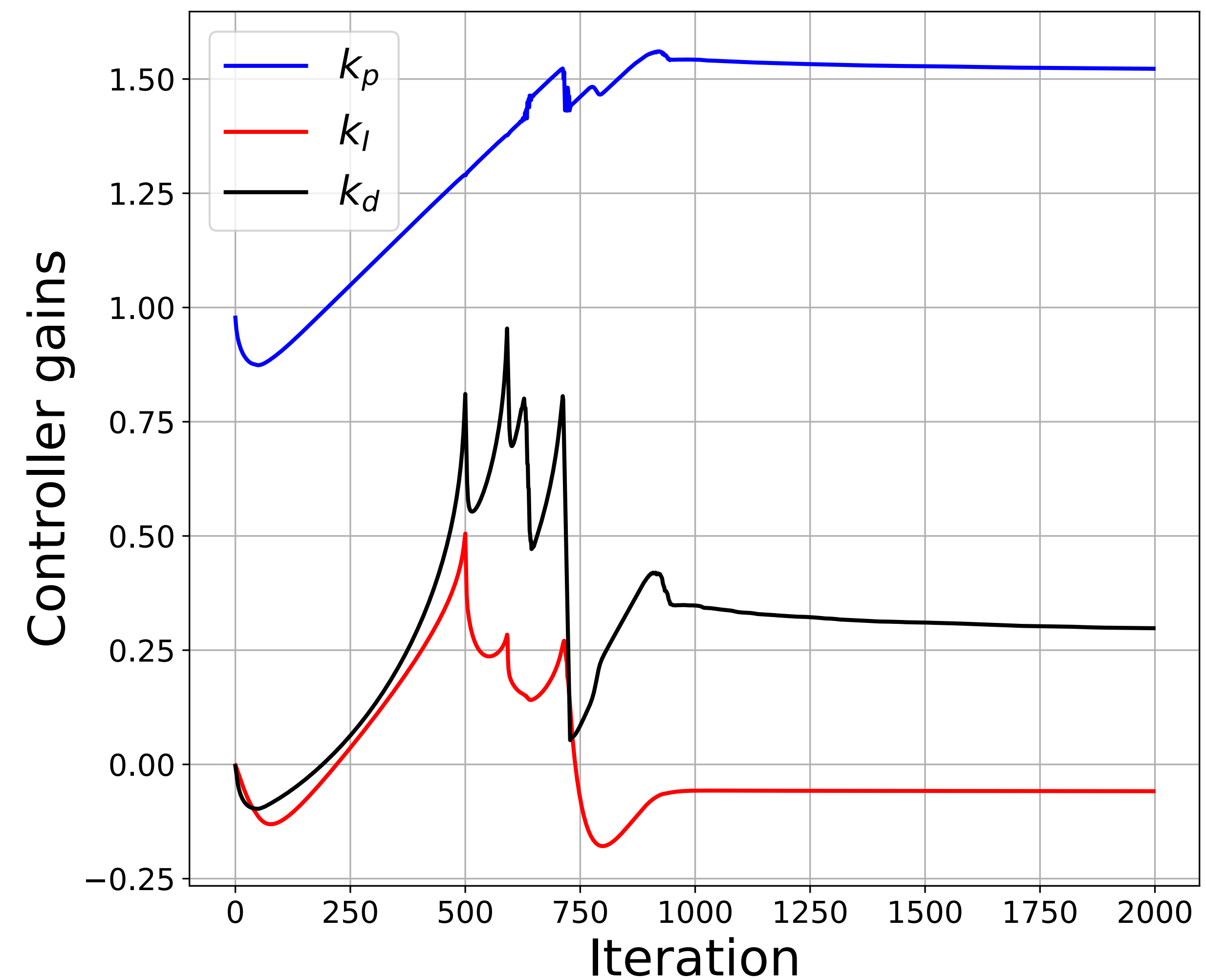
Domain: Chain Walk

Gain Adaptation

Q: How can we choose the controller gains?

We introduce a gain adaptation procedure.





Future Work

- ☐ Extension to RL
- ☐ Other controllers?

Thank you!