Further Extensions to Off-policy PI

- **Off-policy PI:** the key idea to model-free RL under arbitrary exploration.
- **Variants of the TD error in off-policy PI**:
  - Original TD error:

$$\delta_t(v_\pi) \doteq \int_t^{t'} \gamma^{\tau - t} R(X_\tau, U_\tau)\, d\tau + \gamma^{\Delta t} \cdot v_\pi(X_{t'}) - v_\pi(X_t)$$

  - TD error with on- and off-policy hybrid reward:

$$\delta_t^\pi(v_\pi) \doteq \int_t^{t'} \gamma^{\tau - t} R(X_\tau, \pi(X_\tau))\, d\tau + \gamma^{\Delta t} \cdot v_\pi(X_{t'}) - v_\pi(X_t)$$

  - TD error with a general discounting factor $\beta > 0$:

$$\delta_{t,\beta}(v_\pi) \doteq \int_t^{t'} \beta^{\tau - t} R(X_\tau, U_\tau)\, d\tau + \beta^{\Delta t} \cdot v_\pi(X_{t'}) - v_\pi(X_t)$$

  - $$\begin{cases} \delta_t(v_\pi) = \delta_t^\pi(v_\pi) = \delta_{t,\gamma}(v_\pi) = 0 \text{ if } \mu = \pi \text{ (on-policy case)} \\ \delta_t(v_\pi) \neq \delta_t^\pi(v_\pi) \neq \delta_{t,\beta}(v_\pi) \neq 0 \text{ in general off-policy case} \end{cases}$$

Off-policy Bellman Equation and Policy Evaluation

- **Off-policy Bellman equation:** for any admissible π,

$$0 = \mathbb{E}_\mu\big[\delta_t^{\mathsf{off}}(v_\pi) - \mathcal{E}_t^\pi \,\big|\, X_t = x \text{ (and } U_t = u)\big],$$

$$\begin{cases} \delta_t^{\mathsf{off}} : \text{one of the off-policy TD errors } \delta_t(v_\pi),\ \delta_t^\pi(v_\pi), \text{ and } \delta_{t,\beta}(v_\pi); \\ \mathcal{E}_t^\pi \ : \text{the residual determined depending on } \delta_t^{\mathsf{off}}. \end{cases}$$

- **Evaluation:** solve the off-policy Bellman equation over the spaces $\mathcal{X} \times \mathcal{U}$ (API, QPI), $\mathcal{X}$ (EPI), and $\mathcal{X} \times \mathcal{U}_{\mathsf{finite}}$ (CPI) with
  - δ_t^{off} equal to δ_t (API), δ^π (EPI, CPI), and $\delta_{t,\beta}$ (QPI).
  - $\mathcal{E}_t^\pi$ becomes zero when $\mu = \pi$ and contains the term:

$$a_\pi \text{ (API)}, \ \ q_\pi \text{ (QPI)}, \ \ \nabla v_\pi \cdot f_{\mathsf{c}} \text{ (EPI)}, \ \ c_\pi \text{ (CPI)}.$$

Explorized PI (EPI) / C-Policy-Iteration (CPI) from Control Discipline

- **EPI**, **the direct off-policy extension of the on-policy PI**, estimates the value function v_π under the behavior policy μ.
  - **Improvement** is exactly same to on-policy PI.

- **CPI**, **the model-free EPI under the u-AC setting**, estimates v_π and the C-function c_π defined by

$$c_\pi(x) \doteq F_{\mathsf{c}}^{\mathsf{T}}(x) \nabla v_\pi^{\mathsf{T}}(x).$$

  - In the u-AC setting: (with strictly convex S) $\left\{ \begin{array}{l} f_{\mathsf{c}}(x,u) = F_{\mathsf{c}}(x)u \\ R(x,u) = R_0(x) - S(u) \end{array} \right\}$,

$$\textbf{Improvement: } \pi'(x) = \sigma(c_\pi(x)) \text{ with } \sigma^{\mathsf{T}} \doteq \nabla S^{-1}.$$

  - $\mathcal{U}_{\mathsf{finite}} \doteq \{u_j\}_{j=0}^{m} \subset \mathcal{U}$, where u_j's are vectors in $\mathcal{U}$ s.t.

$$\operatorname{span}\{u_j - u_{j-1}\}_{j=1}^{m} = \mathbb{R}^m.$$

Advantage PI (API) / Q-Policy-Iteration (QPI) from RL Discipline

- **API**, **the ideal PI-form of advantage updating**,
  estimates v_π and the advantage function a_π defined by

$$a_\pi(x,u) \doteq \lim_{\Delta t \to 0} \mathbb{E}\big[\, \delta_t(v_\pi)/\Delta t \,\big|\, X_t = x, U_t = u \big]$$

  and then improves the policy using the estimate of a_π.
  - Normalization property: $a_\pi(x, \pi(x)) = 0$ for all $x \in \mathcal{X}$.
  - **Improvement:** $\pi'(x) \in \arg\max_{u \in \mathcal{U}} a_\pi(x,u)\ \forall x \in \mathcal{X}$.

- **QPI**, **the ideal PI-form of Q-learning in CTS**,
  estimates the Q-function q_π defined by

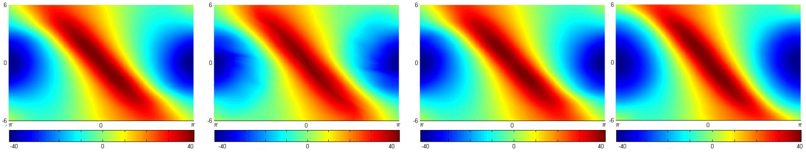
$$q_\pi(x,u) \doteq \kappa \cdot v_\pi(x) + a_\pi(x,u) \text{ for some } \kappa \neq 0$$

  under the different discounting $\beta \doteq \gamma e^\kappa \neq \gamma$.
  - Similarly to discrete case, $v_\pi(x) = q_\pi(x, \pi(x))/\kappa\ \forall x \in \mathcal{X}$.
  - **Improvement:** $\pi'(x) \in \arg\max_{u \in \mathcal{U}} q_\pi(x,u)\ \forall x \in \mathcal{X}$.

Inverted-Pendulum Simulations

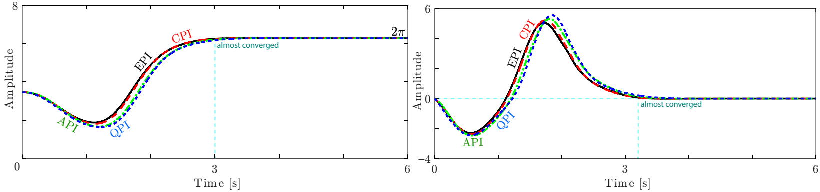
- Inverted-pendulum dynamics: $\ddot{\theta}_\tau = -0.01\dot{\theta}_\tau + 9.8\sin\theta_\tau + U_\tau$
  - State space $(n = 2)$: $\mathcal{X} = \mathbb{R}^2$ with $X_\tau = [\,\theta_\tau \;\; \dot{\theta}_\tau\,]^\mathsf{T}$
  - Action space $(m = 1)$: $\mathcal{U} = \{-5 \le U_\tau \le 5\} \subset \mathbb{R}$;
- Learning objective: swing-up and balance the pendulum at $\theta_\tau = 2k\pi$.
- VF parameters: $\gamma = 0.1$ and $R(x, u) = 10^2 \cos x_1 - S(u)$
  - $S(u) = \left(5^2/2\right) \cdot \ln\left(u_+^{u_+} \cdot u_-^{u_-}\right)$ with $u_\pm = 1 \pm u/5$
- Simulation methods:
  - $\Delta t = 10$ [ms], $\pi_0(x) = 0$, $\beta = 1$
  - the fncs all approximated by RBFNs in closed and bounded subsets:
    - $|\theta_\tau| \le \pi$, $|\dot{\theta}_\tau| \le 6$, $|U_\tau| \le 5$
  - RBF actor-network for policy improvement of IAPI.

Inverted-Pendulum Simulations



(a) API (b) QPI (c) EPI (d) CPI

Fig. 1. The value fnc $v_i(x)$ at $i = 10$ (position θ_τ versus velocity $\dot{\theta}_\tau$)



(a) Pos. trj. θ_τ (b) Vel. trj. $\dot{\theta}_\tau$

Fig. 2 The state trjs. generated under $\begin{cases} \text{1. the init. condition } X_0 = [1.1\pi \;\; 0]^\mathsf{T}; \\ \text{2. the obtained policy } \pi_i \text{ at } i = 10. \end{cases}$