

## Reinforcement Learning

Lecturer: Jiaming Liang

November 13, 2025

## 1 Markov decision process

The finite Markov decision process is abstracted by a quintuple  $M = (\mathbb{S}, \mathbb{A}, \mathbb{P}, c, \gamma)$ , where

- $\mathbb{S}$  is a finite state space,
- $\mathbb{A}$  is a finite action space,
- $\mathbb{P} : \mathbb{S} \times \mathbb{S} \times \mathbb{A} \rightarrow \mathbb{R}$  is transition model,
- $c : \mathbb{S} \times \mathbb{A} \rightarrow \mathbb{R}$  is the cost function, and
- $\gamma \in (0, 1)$  is the discount factor.

A policy  $\pi : \mathbb{A} \times \mathbb{S} \rightarrow \mathbb{R}$  determines the probability of selecting a particular action at a given state.

Consider an environment where: one follows a fixed policy  $\pi$  to take action  $a_t$  given state  $s_t$ , and the environment transitions according to  $\mathbb{P}(s_{t+1} \mid s_t, a_t)$ . Then the resulting dynamics

$$s_{t+1} \sim \mathbb{P}_\pi(\cdot \mid s_t) \quad \text{where } \mathbb{P}_\pi(s' \mid s) = \sum_a \pi(a \mid s) \mathbb{P}(s' \mid s, a)$$

define a Markov chain over states. A Markov chain is stationary if its distribution no longer changes over time. A distribution  $v$  over states is a stationary distribution if

$$v(s') = \sum_s v(s) \mathbb{P}_\pi(s' \mid s),$$

i.e., if one starts in distribution  $v$ , one stays in  $v$  after arbitrary transition steps.

For a given policy  $\pi$ , we measure its performance by the action-value function ( $Q$ -function)  $Q^\pi : \mathbb{S} \times \mathbb{A} \rightarrow \mathbb{R}$  defined as

$$Q^\pi(s, a) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t [c(s_t, a_t)] \mid s_0 = s, a_0 = a, a_t \sim \pi(\cdot \mid s_t), s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t) \right].$$

We also define the state-value function  $V^\pi : \mathbb{S} \rightarrow \mathbb{R}$  associated with  $\pi$  as

$$V^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t [c(s_t, a_t)] \mid s_0 = s, a_t \sim \pi(\cdot \mid s_t), s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t) \right].$$

These functions are often called un-regularized or classic value functions for MDP. In order to impose certain properties of the policy, one may consider their regularized counterparts

$$Q^\pi(s, a) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t [c(s_t, a_t) + h^\pi(s_t)] \mid s_0 = s, a_0 = a, a_t \sim \pi(\cdot \mid s_t), s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t) \right]$$

and

$$V^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t [c(s_t, a_t) + h^\pi(s_t)] \mid s_0 = s, a_t \sim \pi(\cdot \mid s_t), s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t) \right]. \quad (1)$$

Here  $h^\pi$  is a closed convex function w.r.t. the policy  $\pi$ , i.e., there exist some  $\mu \geq 0$  s.t.

$$h^\pi(s) - \left[ h^{\pi'}(s) + \left\langle (h')^{\pi'}(s, \cdot), \pi(\cdot \mid s) - \pi'(\cdot \mid s) \right\rangle \right] \geq \mu D_{\pi'}^\pi(s),$$

where  $\langle \cdot, \cdot \rangle$  denotes the inner product over the action space  $\mathbb{A}$ ,  $(h')^{\pi'}(s, \cdot)$  denotes a subgradient of  $h(s)$  at  $\pi'$ , and  $D_{\pi'}^\pi(s)$  is the Bregman divergence of  $\omega(\pi(\cdot \mid s)) := \sum_{a \in \mathbb{A}} \pi(a \mid s) \log \pi(a \mid s)$  or KL divergence between  $\pi$  and  $\pi'$ , i.e.,

$$D_{\pi'}^\pi(s) = \text{KL}(\pi(\cdot \mid s) \parallel \pi'(\cdot \mid s)) = \sum_{a \in \mathbb{A}} \pi(a \mid s) \log \frac{\pi(a \mid s)}{\pi'(a \mid s)}.$$

It can be easily seen from the definitions of  $Q^\pi$  and  $V^\pi$  that

$$\begin{aligned} V^\pi(s) &= \sum_{a \in \mathbb{A}} \pi(a \mid s) Q^\pi(s, a) = \langle Q^\pi(s, \cdot), \pi(\cdot \mid s) \rangle, \\ Q^\pi(s, a) &= c(s, a) + h^\pi(s) + \gamma \sum_{s' \in \mathbb{S}} \mathbb{P}(s' \mid s, a) V^\pi(s'). \end{aligned} \quad (2)$$

The main objective in MDP/RL is to find an optimal policy  $\pi^* : \mathbb{S} \times \mathbb{A} \rightarrow \mathbb{R}$  s.t.

$$V^{\pi^*}(s) \leq V^\pi(s), \quad \forall \pi(\cdot \mid s) \in \Delta_{|\mathbb{A}|}, s \in \mathbb{S}.$$

Here  $\Delta_{|\mathbb{A}|}$  denotes the simplex constraint over the action space  $\mathbb{A}$ . Hence, we can formulate the above problem as an optimization problem with a single objective by taking the weighted sum of  $V^\pi$  over  $s$  (with weights  $\rho_s > 0$  and  $\sum_{s \in \mathbb{S}} \rho_s = 1$ ):

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}_{s \sim \rho} [V^\pi(s)] \\ \text{s.t.} \quad & \pi(\cdot \mid s) \in \Delta_{|\mathbb{A}|}, \forall s \in \mathbb{S}. \end{aligned}$$

While the weights  $\rho$  can be arbitrarily chosen, a reasonable selection of  $\rho$  would be the stationary state distribution induced by the optimal policy  $\pi^*$ , denoted by  $v^* \equiv v(\pi^*)$ . As such, the above problem reduces to

$$\begin{aligned} \min_{\pi} \quad & \{f(\pi) := \mathbb{E}_{s \sim v^*} [V^\pi(s)]\} \\ \text{s.t.} \quad & \pi(\cdot \mid s) \in \Delta_{|\mathbb{A}|}, \forall s \in \mathbb{S}. \end{aligned} \quad (3)$$

As we will also see later, even though the definition of the objective  $f$  depends on  $v^*$  and hence the unknown optimal policy  $\pi^*$ , the algorithms for solving  $\min_{\pi} f(\pi)$  do not really require the input of  $\pi^*$ .

For a given policy  $\pi$ , we define the discounted state visitation distribution by

$$d_{s_0}^{\pi}(s) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr_{\pi}^t(s \mid s_0),$$

where  $\Pr_{\pi}^t(s \mid s_0)$  denotes the state visitation probability of  $s_t = s$  after we follow the policy  $\pi$  starting at state  $s_0$ . Let  $\mathbb{P}_{\pi}$  denote the transition probability matrix associated with policy  $\pi$ , i.e.,  $\mathbb{P}_{\pi}(i, j) = \sum_{a \in \mathbb{A}} \pi(a \mid i) \mathbb{P}(j \mid i, a)$ , and  $e_i$  be the  $i$ -th unit vector. Then,

$$\begin{aligned} \Pr_{\pi}^t(s \mid s_0) &= e_{s_0}^{\top} (\mathbb{P}_{\pi})^t e_s, \\ d_{s_0}^{\pi}(s) &= (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t e_{s_0}^{\top} (\mathbb{P}_{\pi})^t e_s. \end{aligned} \tag{4}$$

The visitation distribution will be used frequently in the analysis of RL algorithms.

## 2 Policy gradient and optimality conditions

It is well-known that the value function  $V^{\pi}(s)$  in (1) is highly nonconvex w.r.t.  $\pi$ , because the components of  $\pi(\cdot \mid s)$  are multiplied by each other in their definitions.

**Lemma 1** (performance difference lemma). *For any two feasible policies  $\pi$  and  $\pi'$ , we have*

$$V^{\pi'}(s) - V^{\pi}(s) = \frac{1}{1 - \gamma} \mathbb{E}_{s' \sim d_s^{\pi'}} \left[ \langle A^{\pi}(s', \cdot), \pi'(\cdot \mid s') \rangle + h^{\pi'}(s') - h^{\pi}(s') \right],$$

where  $d_s^{\pi'}$  is the discounted state visitation distribution and the advantage function  $A^{\pi}$  is

$$A^{\pi}(s', a) := Q^{\pi}(s', a) - V^{\pi}(s').$$

The proof is left as a homework problem.

It is interesting to note the relation between the advantage function  $A^{\pi}$  and the value function  $Q^{\pi}$ . Let  $e$  denote the vector of all 1's. Then, we have

$$\begin{aligned} \langle A^{\pi}(s', \cdot), \pi'(\cdot \mid s') \rangle &= \langle Q^{\pi}(s', \cdot) - V^{\pi}(s') e, \pi'(\cdot \mid s') \rangle \\ &= \langle Q^{\pi}(s', \cdot), \pi'(\cdot \mid s') \rangle - V^{\pi}(s') \\ &= \langle Q^{\pi}(s', \cdot), \pi'(\cdot \mid s') \rangle - \langle Q^{\pi}(s', \cdot), \pi(\cdot \mid s') \rangle \\ &= \langle Q^{\pi}(s', \cdot), \pi'(\cdot \mid s') - \pi(\cdot \mid s') \rangle \end{aligned} \tag{5}$$

where the first identity follows from the definition of  $A^\pi(s', \cdot)$ , the second equality follows from the fact that  $\langle e, \pi'(\cdot | s') \rangle = 1$ , and the third equality follows from the observation (2).

Thanks to the performance difference lemma, i.e., Lemma 1, we have the following alternative optimality condition.

**Lemma 2.** *For any feasible policy  $\pi$ , we have*

$$\mathbb{E}_{s \sim v^*} \left[ (1 - \gamma) \left( V^\pi(s) - V^{\pi^*}(s) \right) \right] = \mathbb{E}_{s \sim v^*} \left[ \langle Q^\pi(s, \cdot), \pi(\cdot | s) - \pi^*(\cdot | s) \rangle + h^\pi(s) - h^{\pi^*}(s) \right].$$

*Proof.* It follows from Lemma 1 with  $\pi' = \pi^*$  that

$$(1 - \gamma) \left[ V^{\pi^*}(s) - V^\pi(s) \right] = \mathbb{E}_{s' \sim d_s^{\pi^*}} \left[ \langle A^\pi(s', \cdot), \pi^*(\cdot | s') \rangle + h^{\pi^*}(s') - h^\pi(s') \right].$$

Combining the above relation with (5) and taking expectation w.r.t.  $v^*$ , we obtain

$$\begin{aligned} (1 - \gamma) \mathbb{E}_{s \sim v^*} \left[ V^{\pi^*}(s) - V^\pi(s) \right] &= \mathbb{E}_{s \sim v^*, s' \sim d_s^{\pi^*}} \left[ \langle Q^\pi(s', \cdot), \pi^*(\cdot | s') - \pi(\cdot | s') \rangle + h^{\pi^*}(s') - h^\pi(s') \right] \\ &= \mathbb{E}_{s \sim v^*} \left[ \langle Q^\pi(s, \cdot), \pi^*(\cdot | s) - \pi(\cdot | s) \rangle + h^{\pi^*}(s) - h^\pi(s) \right] \end{aligned}$$

where the second identity follows from the fact that  $v^*$  is the steady state distribution induced by  $\pi^*$ . The result then follows by rearranging the terms.  $\square$

Noting that  $f(\pi) = \mathbb{E}_{s \sim v^*} [V^\pi(s)]$  from (3), Lemma 2 implies that for any feasible policy  $\pi$ ,

$$\mathbb{E}_{s \sim v^*} \left[ \langle Q^\pi(s, \cdot), \pi(\cdot | s) - \pi^*(\cdot | s) \rangle + h^\pi(s) - h^{\pi^*}(s) \right] = (1 - \gamma)[f(\pi) - f(\pi^*)] \geq 0. \quad (6)$$

This inequality is an optimality condition of (3), which can be understood as a weak variational inequality (VI). Conceptually, this is close to the optimality condition of composite optimization. Next, we provide an intuitive explanation by formally defining the policy gradient.

**Definition 1.** *For any function of policy  $f : \Pi \rightarrow \mathbb{R}$ , the policy gradient of  $f$  with respect to  $\pi$ , denoted by  $\nabla f(\pi)$ , is the vector satisfying the following,*

$$\lim_{\delta \rightarrow 0, \pi + \delta \in \Pi} |f(\pi + \delta) - f(\pi) - \langle \nabla f(\pi), \delta \rangle| / \|\delta\|_2 \rightarrow 0.$$

**Lemma 3.** *Given a state  $s \in \mathbb{S}$ , then the policy gradient of  $V^\pi(s)$  with respect to  $\pi$  is given by*

$$\nabla V^\pi(s) [s', a] = \frac{1}{1 - \gamma} d_s^{\pi, u}(s') [Q^\pi(s', a) + \nabla h^\pi(s')(a)], \quad \forall (s', a) \in \mathbb{S} \times \mathbb{A},$$

where  $\nabla V^\pi(s) [s', a]$  denotes the entry of  $\nabla V^\pi(s)$  corresponding to the  $(s', a)$  state-action pair.

In view of Lemma 2, the gradient of the objective  $f(\pi)$  in (3) at the optimal policy  $\pi^*$  is given by

$$\begin{aligned}
\nabla f(\pi^*)(s, a) &= \mathbb{E}_{s_0 \sim v^*} \left[ \nabla V^{\pi^*}(s_0)(s, a) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s_0 \sim v^*} \left[ d_{s_0}^{\pi^*}(s) \left[ Q^{\pi^*}(s, a) + \nabla h^{\pi^*}(s)(a) \right] \right] \\
&= \sum_{t=0}^{\infty} \gamma^t (v^*)^T \left( \mathbb{P}^{\pi^*} \right)^t e_s \left[ Q^{\pi^*}(s, a) + \nabla h^{\pi^*}(s, a) \right] \\
&= \frac{1}{1-\gamma} (v^*)^T e_s \left[ Q^{\pi^*}(s, a) + \nabla h^{\pi^*}(s, a) \right] \\
&= \frac{1}{1-\gamma} v^*(s) \left[ Q^{\pi^*}(s, a) + \nabla h^{\pi^*}(s, a) \right],
\end{aligned}$$

where the third identity follows from (4) and the last one follows from the fact that  $(v^*)^\top (\mathbb{P}^{\pi^*})^t = (v^*)^\top$  for any  $t \geq 0$  since  $v^*$  is the steady state distribution of  $\pi^*$ . Therefore, the optimality condition of (3) suggests us to solve the following strong VI

$$\mathbb{E}_{s \sim v^*} \left[ \left\langle Q^{\pi^*}(s, \cdot) + \nabla h^{\pi^*}(s, \cdot), \pi(\cdot | s) - \pi^*(\cdot | s) \right\rangle \right] \geq 0.$$

However, the above VI requires  $h^\pi$  to be differentiable. In order to handle the possible non-smoothness of  $h^\pi$ , we instead solve the following strong VI

$$\mathbb{E}_{s \sim v^*} \left[ \left\langle Q^{\pi^*}(s, \cdot), \pi(\cdot | s) - \pi^*(\cdot | s) \right\rangle + h^\pi(s) - h^{\pi^*}(s) \right] \geq 0.$$

The above strong VI differs from the weak VI in (6) in  $Q^{\pi^*}$  and  $Q^\pi$ , which is the difference in standard strong and weak VI. In general, they are equivalent if monotonicity holds. In RL, this means

$$\mathbb{E}_{s \sim v^*} \left[ \left\langle Q^\pi(s, \cdot) - Q^{\pi^*}(s, \cdot), \pi(\cdot | s) - \pi^*(\cdot | s) \right\rangle \right] \geq 0.$$

The key is to understand optimality condition of RL as VI, and then design first-order methods for solving the (weak or strong) VI instead of optimization (3).

### 3 Policy mirror descent

We are ready to present the policy mirror descent (PMD) and establish its convergence results. In the proposed PMD method, we will update a given policy  $\pi$  to  $\pi^+$  through the following proximal mapping

$$\pi^+(\cdot | s) = \operatorname{argmin}_{p(\cdot | s) \in \Delta_{|\mathcal{A}|}} \eta [\langle Q^\pi(s, \cdot), p(\cdot | s) \rangle + h^p(s)] + D_\pi^p(s).$$

When  $h^p(s) = 0$  or  $h^p(s) = \tau D_{\pi_0}^p(s)$  for some  $\tau > 0$  and given  $\pi_0$ , the above update scheme has closed-form solutions, and it boils down to a multiplicative update as in standard mirror descent.

For the sake of simplicity, we assume that

$$\pi_0(a \mid s) = 1/|\mathbb{A}|, \quad \forall a \in \mathbb{A}, s \in \mathbb{S}.$$

It can be shown that

$$D_{\pi_0}^\pi(s) \leq \log |\mathbb{A}|, \quad \forall \pi(\cdot \mid s) \in \Delta_{|\mathbb{A}|}.$$

---

**Algorithm 1** Policy mirror descent

---

**Input:** initial points  $\pi_0$  and stepsizes  $\eta_k \geq 0$ .

**for**  $k = 0, 1, \dots$  **do** **do**

$$\pi_{k+1}(\cdot \mid s) = \operatorname{argmin}_{p(\cdot \mid s) \in \Delta_{|\mathbb{A}|}} \{ \eta_k [\langle Q^{\pi_k}(s, \cdot), p(\cdot \mid s) \rangle + h^p(s)] + D_{\pi_k}^p(s) \}, \forall s \in \mathbb{S}.$$

**end for**

---

The following result characterizes the optimality condition of the PMD update. Its proof is standard as in Lecture 5 on mirror descent and hence is skipped.

**Lemma 4.** *For any  $p(\cdot \mid s) \in \Delta_{|\mathbb{A}|}$ , we have*

$$\eta_k [\langle Q^{\pi_k}(s, \cdot), \pi_{k+1}(\cdot \mid s) - p(\cdot \mid s) \rangle + h^{\pi_{k+1}}(s) - h^p(s)] + D_{\pi_k}^{\pi_{k+1}}(s) \leq D_{\pi_k}^p(s) - (1 + \eta_k \mu) D_{\pi_{k+1}}^p(s).$$

**Lemma 5.** *For any  $s \in \mathbb{S}$ , we have*

$$V^{\pi_{k+1}}(s) \leq V^{\pi_k}(s), \quad (7)$$

$$\langle Q^{\pi_k}(s, \cdot), \pi_{k+1}(\cdot \mid s) - \pi_k(\cdot \mid s) \rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s) \geq V^{\pi_{k+1}}(s) - V^{\pi_k}(s). \quad (8)$$

*Proof.* It follows from Lemma 1 with  $\pi' = \pi_{k+1}$  and  $\pi = \pi_k$  that

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) = \frac{1}{1 - \gamma} \mathbb{E}_{s' \sim d_s^{\pi_{k+1}}} [\langle A^{\pi_k}(s', \cdot), \pi_{k+1}(\cdot \mid s') \rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s')].$$

Similarly to (5), we can show that

$$\begin{aligned} \langle A^{\pi_k}(s', \cdot), \pi_{k+1}(\cdot \mid s') \rangle &= \langle Q^{\pi_k}(s', \cdot) - V^{\pi_k}(s') e, \pi_{k+1}(\cdot \mid s') \rangle \\ &= \langle Q^{\pi_k}(s', \cdot), \pi_{k+1}(\cdot \mid s') \rangle - V^{\pi_k}(s') \\ &= \langle Q^{\pi_k}(s', \cdot), \pi_{k+1}(\cdot \mid s') - \pi_k(\cdot \mid s') \rangle. \end{aligned}$$

Combining the above two identities, we then obtain

$$V^{\pi_{k+1}}(s) - V^{\pi_k}(s) = \frac{1}{1 - \gamma} \mathbb{E}_{s' \sim d_s^{\pi_{k+1}}} [\langle Q^{\pi_k}(s', \cdot), \pi_{k+1}(\cdot \mid s') - \pi_k(\cdot \mid s') \rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s')]. \quad (9)$$

Now we conclude from Lemma 4 with  $p(\cdot \mid s') = \pi_k(\cdot \mid s')$  that

$$\langle Q^{\pi_k}(s', \cdot), \pi_{k+1}(\cdot \mid s') - \pi_k(\cdot \mid s') \rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s') \leq -\frac{1}{\eta_k} \left[ (1 + \eta_k \mu) D_{\pi_{k+1}}^{\pi_k}(s') + D_{\pi_k}^{\pi_{k+1}}(s') \right].$$

The previous two conclusions then clearly imply the result in (7). It also follows from the above inequality that

$$\begin{aligned} & \mathbb{E}_{s' \sim d_s^{\pi_{k+1}}} [\langle Q^{\pi_k}(s', \cdot), \pi_{k+1}(\cdot | s') - \pi_k(\cdot | s') \rangle + h^{\pi_{k+1}}(s') - h^{\pi_k}(s')] \\ & \leq d_s^{\pi_{k+1}}(s) [\langle Q^{\pi_k}(s, \cdot), \pi_{k+1}(\cdot | s) - \pi_k(\cdot | s) \rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s)] \\ & \leq (1 - \gamma) [\langle Q^{\pi_k}(s, \cdot), \pi_{k+1}(\cdot | s) - \pi_k(\cdot | s) \rangle + h^{\pi_{k+1}}(s) - h^{\pi_k}(s)]. \end{aligned}$$

where the last inequality follows from the fact that  $d_s^{\pi_{k+1}}(s) \geq (1 - \gamma)$  due to the definition of  $d_s^{\pi_{k+1}}$  in (4). The result in (8) then follows immediately from (9) and the above inequality.  $\square$

We are ready to present the key recursion.

**Theorem 1.** *For any  $k \geq 0$  in the PMD method, we have*

$$f(\pi_{k+1}) - f(\pi^*) + \left( \frac{1}{\eta_k} + \mu \right) \mathbb{D}(\pi_{k+1}, \pi^*) + \frac{1}{\eta_k} \mathbb{D}(\pi_k, \pi_{k+1}) \leq \gamma [f(\pi_k) - f(\pi^*)] + \frac{1}{\eta_k} \mathbb{D}(\pi_k, \pi^*)$$

where

$$\mathbb{D}(\pi, \pi') := \mathbb{E}_{s \sim v^*} [D_{\pi}^{\pi'}(s)].$$

*Proof.* By Lemma 4 with  $p = \pi^*$ , we have

$$\eta_k [\langle Q^{\pi_k}(s, \cdot), \pi_{k+1}(\cdot | s) - \pi^*(\cdot | s) \rangle + h^{\pi_{k+1}}(s) - h^{\pi^*}(s)] + D_{\pi_k}^{\pi_{k+1}}(s) \leq D_{\pi_k}^{\pi^*}(s) - (1 + \eta_k \mu) D_{\pi_{k+1}}^{\pi^*}(s)$$

which, in view of (8), then implies that

$$\begin{aligned} & \eta_k [\langle Q^{\pi_k}(s, \cdot), \pi_k(\cdot | s) - \pi^*(\cdot | s) \rangle + h^{\pi_k}(s) - h^{\pi^*}(s)] + \eta_k [V^{\pi_{k+1}}(s) - V^{\pi_k}(s)] + D_{\pi_k}^{\pi_{k+1}}(s) \\ & \leq D_{\pi_k}^{\pi^*}(s) - (1 + \eta_k \mu) D_{\pi_{k+1}}^{\pi^*}(s). \end{aligned}$$

Taking expectation w.r.t.  $v^*$  on both sides of the above inequality and using Lemma 2, we arrive at

$$\begin{aligned} & \mathbb{E}_{s \sim v^*} [\eta_k (1 - \gamma) (V^{\pi_k}(s) - V^{\pi^*}(s))] + \eta_k \mathbb{E}_{s \sim v^*} [V^{\pi_{k+1}}(s) - V^{\pi_k}(s)] + \mathbb{E}_{s \sim v^*} [D_{\pi_k}^{\pi_{k+1}}(s)] \\ & \leq \mathbb{E}_{s \sim v^*} [D_{\pi_k}^{\pi^*}(s) - (1 + \eta_k \mu) D_{\pi_{k+1}}^{\pi^*}(s)]. \end{aligned}$$

Rearranging the terms in the above inequality, we have

$$\begin{aligned} & \mathbb{E}_{s \sim v^*} [\eta_k (V^{\pi_{k+1}}(s) - V^{\pi^*}(s)) + (1 + \eta_k \mu) D_{\pi_{k+1}}^{\pi^*}(s)] + \mathbb{E}_{s \sim v^*} [D_{\pi_k}^{\pi_{k+1}}(s)] \\ & \leq \gamma \mathbb{E}_{s \sim v^*} [\eta_k (V^{\pi_k}(s) - V^{\pi^*}(s))] + \mathbb{E}_{s \sim v^*} [D_{\pi_k}^{\pi^*}(s)] \end{aligned}$$

which, in view of the definitions of  $f$  and  $\mathbb{D}$ , then implies the result.  $\square$

Finally, we conclude the lecture by presenting convergence results with different stepsize rules.

**Corollary 1.** *The following statements hold:*

(a) *suppose  $\eta_k = \eta$  and  $1 + \eta\mu \geq \frac{1}{\gamma}$ , then for every  $k \geq 0$ ,*

$$f(\pi_k) - f(\pi^*) + \left(\frac{1}{\eta} + \mu\right) \mathbb{D}(\pi_k, \pi^*) \leq \gamma^k \left[ f(\pi_0) - f(\pi_\tau^*) + \frac{1}{\gamma\eta} \log |\mathbb{A}| \right];$$

(b) *suppose  $\eta_k = \eta$ , then for every  $k \geq 0$ ,*

$$f(\pi_{k+1}) - f(\pi^*) \leq \frac{\eta\gamma [f(\pi_0) - f(\pi^*)] + \log |\mathbb{A}|}{\eta(1 - \gamma)(k + 1)};$$

(c) *suppose  $\eta_k = 1/\gamma^k$ , then for every  $k \geq 0$ ,*

$$f(\pi_k) - f(\pi^*) + \gamma^k \mathbb{D}(\pi_k, \pi^*) \leq \gamma^k [f(\pi_0) - f(\pi_\tau^*) + \log |\mathbb{A}|].$$