Learning and Optimization in Multiagent Decision-Making Systems

Lecture Notes: Computational Methods for MDPs (based on lectures of S. Kakade and R. Srikant)
Instructor: Rasoul Etesami

Consider the infinite horizon MDP (S, A, P, r, γ) with discount factor $\gamma \in [0,1)$. Recall that there exists an optimal deterministic stationary policy π^* with value function $V^* : S \to \mathbb{R}_+$ which satisfies the Bellman optimality equations:

$$V^*(s) = \max_{a\in A}\left[\bar{r}(s,a) + \gamma \sum_{s'\in S} P_{s,s'}(a)V^*(s')\right] =: T(V^*)(s).$$

Recall that the Bellman optimality operator $T : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ satisfies the following properties:

(i) **Contraction:** T is a contraction in the ℓ_∞-norm with contraction factor γ, i.e., for any $V, V' \in \mathbb{R}^{|S|}$, we have

$$\|T(V) - T(V')\|_\infty \leq \gamma \|V - V'\|_\infty.$$

(ii) **Monotonicity:** T is monotone, i.e., for any $V, V' \in \mathbb{R}^{|S|}$, if $V(s) \leq V'(s)$ for all $s \in S$, then

$$T(V)(s) \leq T(V')(s) \quad \text{for all } s \in S.$$

(iii) **Shift:** For any $c \in \mathbb{R}$, we have

$$T(V + c\mathbf{1}) = T(V) + \gamma c\mathbf{1},$$

where $\mathbf{1} \in \mathbb{R}^{|S|}$ is the all-one vector.

All the above properties are also true for the Bellman operator $T^\pi : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ associated with any policy π.

Suppose that for any $a \in A$ and $s, s' \in S$, we have access to the expected reward $\bar{r}(s,a)$ and the transition probability $P_{s,s'}(a) = \mathbb{P}(S_1 = s' \mid S_0 = s, A_0 = a)$. Our goal is to develop efficient algorithms for finding an optimal stationary policy π^*. Recall that V^*, the value function of π^*, is the unique fixed point of the Bellman optimality operator, i.e.,

$$T(V^*) = V^*.$$

Using this as our key insight, we present two computational techniques to find the optimal policy.

Value Iteration Method

The value iteration algorithm is summarized in Algorithm 1.

Theorem 55. *Running the value iteration algorithm (Algorithm 1) for n iterations, we have:*

$$\|V_n - V^*\|_\infty \leq \frac{\gamma^n}{1-\gamma}\|V_1 - V_0\|_\infty, \quad \text{and} \quad \|V_{\pi_n} - V^*\|_\infty \leq \frac{2\gamma^n}{1-\gamma}\|V_1 - V_0\|_\infty.$$

In particular, to ensure that $\|V_{\pi_n} - V^*\|_\infty < \varepsilon$, it suffices to choose

$$n \geq \frac{\log(\|V_1 - V_0\|_\infty) + \log(2) - \log(\varepsilon(1-\gamma))}{\log\left(\frac{1}{\gamma}\right)}.$$

Algorithm 1 Value Iteration

```
Input: Infinite horizon MDP $(S, A, P, r, \gamma)$, number of iterations $n > 0$, initial estimate $V_0$
Output: Estimate $V_n \in \mathbb{R}^{|S|}$ and policy $\pi_n$
for $k \leftarrow 0$ to $n-1$ do
    $V_{k+1} \leftarrow T(V_k)$
end for
$\pi_n \leftarrow$ greedy policy with respect to $V_n$
```

Proof: Using the triangle inequality, we have:

$$\begin{aligned}\|V_n - V^*\|_\infty &\leq \|V_{n+1} - V_n\|_\infty + \|V_{n+2} - V_{n+1}\|_\infty + \cdots + \|V_{n+\ell} - V_{n+\ell-1}\|_\infty + \|V^* - V_{n+\ell}\|_\infty \\ &\leq (\gamma^n + \gamma^{n+1} + \gamma^{n+2} + \cdots + \gamma^{n+\ell-1})\|V_1 - V_0\|_\infty + \gamma^{n+\ell}\|V^* - V_0\|_\infty \\ &\leq \frac{\gamma^n}{1-\gamma}\|V_1 - V_0\|_\infty + \gamma^{n+\ell}\|V^* - V_0\|_\infty.\end{aligned}$$

Notice that the left-hand side of the above inequality does not depend on ℓ. Taking limit $\ell \to \infty$, we have

$$\|V_n - V^*\|_\infty \leq \frac{\gamma^n}{1-\gamma}\|V_1 - V_0\|_\infty.$$

Next, we bound the performance of the resulting greedy policy π_n. Notice that by definition, we have $T(V_n) = T^{\pi_n}(V_n)$ and $T^{\pi_n}(V_{\pi_n}) = V_{\pi_n}$. Recall that T^{π_n} is a contraction in the ℓ_∞-norm with contraction factor γ. We have:

$$\begin{aligned}\|V_n - V_{\pi_n}\|_\infty &\leq \|T(V_n) - T^{\pi_n}(V_{\pi_n})\|_\infty + \|V_n - T(V_n)\|_\infty \\ &= \|T^{\pi_n}(V_n) - T^{\pi_n}(V_{\pi_n})\|_\infty + \|T(V_n) - V_n\|_\infty \\ &\leq \gamma\|V_n - V_{\pi_n}\|_\infty + \gamma^n\|V_0 - V_1\|_\infty,\end{aligned}$$

which implies:

$$\|V_n - V_{\pi_n}\|_\infty \leq \frac{\gamma^n}{1-\gamma}\|V_0 - V_1\|_\infty.$$

Using the triangle inequality, we get:

$$\|V_{\pi_n} - V^*\|_\infty \leq \|V_n - V_{\pi_n}\|_\infty + \|V_n - V^*\|_\infty \leq \frac{2\gamma^n}{1-\gamma}\|V_1 - V_0\|_\infty.$$

□

Notice that the value iteration algorithm converges in finitely many iterations; however, the number of iterations required might be large. More specifically, if

$$\frac{2\gamma^n}{1-\gamma}\|V_1 - V_0\|_\infty < \delta,$$

then we have $V_{\pi_n} = V^*$, where π_n is a greedy deterministic policy with respect to V_n, and

$$\delta := \min_{\substack{\pi,\pi' \text{ deterministic} \\ V^\pi \neq V^{\pi'}}} \|V^\pi - V^{\pi'}\|_\infty.$$

Hence, the value iteration algorithm converges after at most

$$n = \left\lceil \frac{\log\left(\frac{\delta(1-\gamma)}{2\|V_1 - V_0\|_\infty}\right)}{\log(\gamma)} \right\rceil$$

many iterations.

Policy Iteration Method

The value iteration algorithm estimates the optimal value function and then uses this estimate to generate a policy. In the policy iteration algorithm, we directly estimate the policy at each iteration. Consider any policy π_0, and let V^{π_0} denote its value function. Recall that V^{π_0} is the unique fixed point of the Bellman operator T^{π_0}, and in particular,

$$V^{\pi_0} = (I - \gamma P^{\pi_0})^{-1} \bar{r}^{\pi_0}.$$

Let π_1 denote the greedy policy with respect to V^{π_0}, i.e., for any state $s \in S$ we have:

$$T(V^{\pi_0})(s) = \bar{r}(s, \pi_1(s)) + \gamma \sum_{s' \in S} P_{s,s'}(\pi_1(s)) V^{\pi_0}(s') = T^{\pi_1}(V^{\pi_0})(s).$$

One may expect π_1 to be closer to the optimal policy than π_0. We will show that this is indeed the case. The policy iteration algorithm uses π_1 as the input to the next iteration and repeats the same process. This is summarized in Algorithm 2.

Algorithm 2 Policy Iteration

```
Input: Infinite horizon MDP $(S, A, P, r, \gamma)$, number of iterations $n > 0$, initial policy $\pi_0$
Output: Policy $\pi_n$
for $k \leftarrow 0$ to $n - 1$ do
    $\pi_{k+1} \leftarrow$ greedy policy with respect to $V^{\pi_k}$, i.e., $T^{\pi_{k+1}}(V^{\pi_k}) = T(V^{\pi_k})$
end for
```

Theorem 56. *The policy iteration algorithm (Algorithm 2) generates a sequence of policies $\{\pi_k\}$ such that their value functions improve monotonically:*

$$V^{\pi_k} \leq V^{\pi_{k+1}}, \quad \forall k \geq 0.$$

If at any iteration k, we have $V^{\pi_k} = V^{\pi_{k+1}}$, then π_k is an optimal policy and $V^{\pi_k} = V^*$.

Proof: We prove the convergence of the above algorithm to the optimal policy. Notice that it is enough to show improvement at each iteration, i.e., $V^{\pi_k} \leq V^{\pi_{k+1}}$ for all k (component-wise). In particular, if $V^{\pi_k} = V^{\pi_{k+1}}$, then V^{π_k} is the fixed point of the Bellman optimality operator T, and π_k must be an optimal policy. This is because if $V^{\pi_k} = V^{\pi_{k+1}}$ for some k, we have

$$T(V^{\pi_k}) = T^{\pi_{k+1}}(V^{\pi_k}) = T^{\pi_{k+1}}(V^{\pi_{k+1}}) = V^{\pi_{k+1}} = V^{\pi_k}.$$

To show improvement, observe that

$$V^{\pi_k} = T^{\pi_k}(V^{\pi_k}) \leq T(V^{\pi_k}) = T^{\pi_{k+1}}(V^{\pi_k}).$$

By monotonicity of the Bellman operator $T^{\pi_{k+1}}$, we have:

$$V^{\pi_k} \leq T^{\pi_{k+1}}(V^{\pi_k}) \leq (T^{\pi_{k+1}})^2(V^{\pi_k}) \leq \cdots \leq (T^{\pi_{k+1}})^m(V^{\pi_k}).$$

Taking the limit as $m \to \infty$, we get:

$$V^{\pi_k} \leq \lim_{m\to\infty} (T^{\pi_{k+1}})^m(V^{\pi_k}) = V^{\pi_{k+1}}.$$

□

Policy Gradient Method

For a distribution μ over states, define:

$$V(\mu) := \mathbb{E}_{s_0 \sim \mu}[V(s_0)],$$

where we slightly overload notation. Consider a class of parametric policies $\{\pi_\theta \mid \theta \in \mathbb{R}^d\}$. The optimization problem of interest is:

$$\max_{\theta\in\mathbb{R}^d} V^{\pi_\theta}(\mu).$$

In many settings, one of the most practically effective methods is gradient ascent, which takes the form of

$$\theta^{k+1} = \theta^k + \eta_k \nabla_\theta V^{\pi_{\theta^k}}(\mu)$$

One immediate issue is that if the policy class $\{\pi_\theta\}$ consists of deterministic policies, then $V_\mu(\pi_\theta)$ will, in general, not be differentiable. This motivates us to consider policy classes that are stochastic, which permit differentiability.

Example (softmax policies): It is instructive to explicitly consider a "tabular" policy representation, given by the softmax policy:

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a'\in A} \exp(\theta_{s,a'})},$$

where the parameter space is $\theta \in \mathbb{R}^{|S||A|}$. Note that (the closure of) the set of softmax policies contains all stationary and deterministic policies.

Example (Linear softmax policies): For any state-action pair (s,a), suppose we have a feature mapping $\phi_{s,a} \in \mathbb{R}^d$. Let us consider the policy class:

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi_{s,a})}{\sum_{a'\in A} \exp(\theta^\top \phi_{s,a'})},$$

with $\theta \in \mathbb{R}^d$.

Advantages and the State-Action Visitation Distribution

Let us first introduce the concept of an advantage.

Definition 57. *The* advantage $A^\pi(s,a)$ *of a policy* π *is defined as:*

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

Note that for the optimal policy π^**, we have* $A^{\pi^*}(s,a) \leq 0, \forall s, a$.

Definition 58. *The discounted state visitation distribution $d^\pi_{s_0}$ is defined by:*

$$d^\pi_{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s \mid s_0),$$

where $\Pr(s_t = s \mid s_0)$ *is the state visitation probability using policy π starting from s_0. We also write $d^\pi_\mu(s) = \mathbb{E}_{s_0\sim\mu}[d^\pi_{s_0}(s)]$.*

By construction, observe that for any function $f : S \times A \to \mathbb{R}$,

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t, a_t)\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi_\mu}\mathbb{E}_{a\sim\pi(\cdot|s)}[f(s,a)].$$

The following lemma is a fundamental tool in the convergence analysis of direct policy search methods.

Lemma 59 (Performance Difference Lemma). *For all policies π, π' and distributions μ,*

$$V^\pi(\mu) - V^{\pi'}(\mu) = \mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi'}(s_t, a_t)\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi_\mu}\mathbb{E}_{a\sim\pi(\cdot|s)}\left[A^{\pi'}(s,a)\right].$$

Proof: Fix a state s as the initial state. We can write

$$\begin{aligned}
V^\pi(s) - V^{\pi'}(s) &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right] - V^{\pi'}(s) \\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s_t, a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)\right)\right] \\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t \left(r(s_t, a_t) + \mathbb{E}[V^{\pi'}(s_{t+1}) \mid s_t, a_t] - V^{\pi'}(s_t)\right)\right] \\
&= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi'}(s_t, a_t)\right],
\end{aligned}$$

where the second equality holds using a telescoping sum, and the last step uses the tower property of expectations and the definition of the Q-function function and the advantage function because $\mathbb{E}_{s'\sim P(\cdot|s,a)}[r(s,a) + \gamma V^{\pi'}(s') \mid s, a] = Q^{\pi'}(s,a)$. The proof is completed by applying linearity of expectation over the starting state distribution μ:

$$V^\pi(\mu) - V^{\pi'}(\mu) = \mathbb{E}_{s\sim\mu}\left[V^\pi(s) - V^{\pi'}(s)\right] = \mathbb{E}_{\mu,\pi}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi'}(s_t, a_t)\right].$$

□

In order to be able to use the policy gradient method, we need to be able to provide an expression for the gradient of the value function with respect to policy parametrization. The following lemma provides such characterizations.

Theorem 60 (Policy gradients). *The following are expressions for the policy gradient $\nabla_\theta V^\pi_\theta(\mu)$:*

REINFORCE:

$$\nabla_\theta V^{\pi_\theta}(\mu) = (1-\gamma)\mathbb{E}\left[\left(\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right)\left(\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right)\right]$$

.

Action-value expression:

$$\nabla_\theta V^{\pi_\theta}(\mu) = \mathbb{E}\left[\sum_{t=0}^{T}\nabla_\theta \log \pi_\theta(a_t \mid s_t) Q^{\pi_\theta}(s_t, a_t)\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s,a)\nabla_\theta \log \pi_\theta(a \mid s)]$$

Advantage expression:

$$\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}[A^{\pi_\theta}(s,a)\nabla_\theta \log \pi_\theta(a \mid s)]$$

Proof: Let $\tau = \{(s_0, a_0), (s_1, a_1), (s_2, a_2), \ldots\}$ denote a sample trajectory generated under the policy π_θ, whose unconditional distribution $\Pr(\tau)$ under policy π_θ with starting distribution μ is:

$$\Pr(\tau) = \mu(s_0)\pi_\theta(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi_\theta(a_1 \mid s_1)P(s_2 \mid s_1, a_1)\pi_\theta(a_2|s_2) \tag{14}$$

Define the discounted total reward of the trajectory τ as:

$$R(\tau) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t).$$

We have:

$$\nabla_\theta V^{\pi_\theta}(\mu) = \nabla_\theta \sum_\tau R(\tau)\Pr(\tau) = \sum_\tau R(\tau)\nabla_\theta \Pr(\tau) = \sum_\tau R(\tau)\Pr(\tau)\nabla_\theta \log \Pr(\tau)$$

Now using the trajectory probability (14) and taking log and derivative:

$$\nabla_\theta \log \Pr(\tau) = \sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t \mid s_t)$$

Thus:

$$\nabla_\theta V^{\pi_\theta}(\mu) = \sum_\tau R(\tau)\Pr(\tau)\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t \mid s_t) = \mathbb{E}\left[R(\tau)\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right],$$

which completes the proof of the first (REINFORCE) expression.

For the second claim, for any state s_0,

$$\begin{aligned}
\nabla_\theta V^{\pi_\theta}(s_0) &= \nabla_\theta \sum_{a_0} \pi_\theta(a_0 \mid s_0) Q^{\pi_\theta}(s_0, a_0) \\
&= \sum_{a_0} \nabla \pi_\theta(a_0 \mid s_0) Q^{\pi_\theta}(s_0, a_0) + \sum_{a_0} \pi_\theta(a_0 \mid s_0) \nabla Q^{\pi_\theta}(s_0, a_0) \\
&= \sum_{a_0} \pi_\theta(a_0 \mid s_0) \nabla \log \pi_\theta(a_0 \mid s_0) Q^{\pi_\theta}(s_0, a_0) \\
&\quad + \sum_{a_0} \pi_\theta(a_0 \mid s_0) \nabla \left[(1-\gamma) r(s_0, a_0) + \gamma \sum_{s_1} P(s_1 \mid s_0, a_0) V^{\pi_\theta}(s_1) \right] \\
&= \sum_{a_0} \pi_\theta(a_0 \mid s_0) \nabla \log \pi_\theta(a_0 \mid s_0) Q^{\pi_\theta}(s_0, a_0) \\
&\quad + \gamma \sum_{a_0, s_1} \pi_\theta(a_0 \mid s_0) P(s_1 \mid s_0, a_0) \nabla V^{\pi_\theta}(s_1) \\
&= \mathbb{E}_{\pi_\theta, P}\left[Q^{\pi_\theta}(s_0, a_0) \nabla \log \pi_\theta(a_0 \mid s_0) \right] + \mathbb{E}_{\pi_\theta, P}\left[\nabla V^{\pi_\theta}(s_1) \right]
\end{aligned}$$

By linearity of expectation and applying the same reasoning recursively:

$$\begin{aligned}
\nabla_\theta V^{\pi_\theta}(s_0) &= \mathbb{E}_{\pi_\theta, P}\left[Q^{\pi_\theta}(s_0, a_0) \nabla \log \pi_\theta(a_0 \mid s_0) \right] + \mathbb{E}_{\pi_\theta, P}\left[Q^{\pi_\theta}(s_1, a_1) \nabla \log \pi_\theta(a_1 \mid s_1) \right] + \cdots \\
&= \mathbb{E}_{\pi_\theta, P}\left[\sum_{t=0}^{\infty} Q^{\pi_\theta}(s_t, a_t) \nabla \log \pi_\theta(a_t \mid s_t) \right].
\end{aligned}$$

This completes the proof of the second claim.

The proof of the last claim is identical to the second claim after we realize that

$$\mathbb{E}_{\pi_\theta, P}\left[V^{\pi_\theta}(s_t) \nabla_\theta \log \pi_\theta(a_t \mid s_t) \right] = 0 \ \forall t,$$

and using the definition of the advantage function $A^{\pi_\theta}(s_t, a_t) = Q^{\pi_\theta}(s_t, a_t) - V^{\pi_\theta}(s_t)$. □

Natural Policy Gradient

Let us consider direct tabular parametrization, i.e., $\theta_{(s,a)} = \pi(a|s)$. Then, using the last expression for the gradient of the value function in the above theorem, we have $\left(\nabla_\theta V^\pi(\mu)\right)_{(s,a)} = \frac{1}{1-\gamma} d_\mu^\pi(s) A^\pi(s,a)$. Now, assume an estimate π^t of the optimal policy. Motivated by the mirror descent in static optimization, we update our policy in the next iteration as

$$\pi^{t+1} = \underset{\pi : \sum_a \pi(a|s) = 1 \forall s}{\arg\max} \left\{ (\pi - \pi^t)' \nabla V^{\pi^t}(\mu) + \frac{1}{\eta} \sum_s d_\mu^{\pi^t}(s) \mathrm{KL}\big(\pi(\cdot|s) \| \pi^t(\cdot|s)\big) \right\}.$$

Using $\left(\nabla_\theta V^{\pi^t}(\mu)\right)_{(s,a)} = \frac{1}{1-\gamma} d_\mu^{\pi^t}(s) A^{\pi^t}(s,a)$ in the above optimization problem, obtaining π^{t+1} reduces to solving

$$\begin{aligned}
&\max_{\pi : \sum_a \pi(a|s) = 1 \forall s} \left\{ \frac{1}{1-\gamma} \sum_{s,a} d_\mu^{\pi^t}(s) A^{\pi^t}(s,a) (\pi(a|s) - \pi^t(a|s)) + \frac{1}{\eta} \sum_s d_\mu^{\pi^t}(s) \mathrm{KL}\big(\pi(\cdot|s) \| \pi^t(\cdot|s)\big) \right\} \\
&= \sum_s d_\mu^{\pi^t}(s) \max_{\pi : \sum_a \pi(a|s) = 1} \left\{ \frac{1}{1-\gamma} \sum_a A^{\pi^t}(s,a) (\pi(a|s) - \pi^t(a|s)) + \frac{1}{\eta} \mathrm{KL}\big(\pi(\cdot|s) \| \pi^t(\cdot|s)\big) \right\}.
\end{aligned}$$

Now, for any fixed s, the solution to each inner maximization can be calculated in a closed-form to obtain $\pi^{t+1}(\cdot|s)$. Therefore, the *Natural Policy Gradient (NPG)* updates take the form:

$$\pi^{(t+1)}(a \mid s) = \frac{\pi^{(t)}(a \mid s) \exp\left(\eta \frac{A^{(t)}(s,a)}{1-\gamma}\right)}{Z_t(s)},$$

where the normalizing factor $Z_t(s)$ is given by:

$$Z_t(s) = \sum_{a \in \mathcal{A}} \pi^{(t)}(a \mid s) \exp\left(\eta \frac{A^{(t)}(s,a)}{1-\gamma}\right).$$

Lemma 61. *[Improvement lower bound for NPG] For the iterates $\pi^{(t)}$ generated by the NPG updates, we have for all starting state distributions μ:*

$$V^{\pi^{(t+1)}}(\mu) - V^{\pi^{(t)}}(\mu) \geq \frac{(1-\gamma)}{\eta} \mathbb{E}_{s\sim\mu} [\log Z_t(s)] \geq 0.$$

Proof: First, we show that $\log Z_t(s) \geq 0$. To see this, observe:

$$\log Z_t(s) = \log \sum_a \pi^{(t)}(a \mid s) \exp\left(\eta \frac{A^{(t)}(s,a)}{1-\gamma}\right).$$

By Jensen's inequality applied to the concave function $\log x$, we have:

$$\log Z_t(s) \geq \sum_a \pi^{(t)}(a \mid s) \log \exp\left(\eta \frac{A^{(t)}(s,a)}{1-\gamma}\right) = \frac{\eta}{1-\gamma} \sum_a \pi^{(t)}(a \mid s) A^{(t)}(s,a) = 0.$$

Now, apply the performance difference lemma:

$$V^{\pi^{(t+1)}}(\mu) - V^{\pi^{(t)}}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\mu^{(t+1)}} \left[\sum_a \pi^{(t+1)}(a \mid s) A^{(t)}(s,a) \right].$$

Using the NPG update expression, we rewrite this as:

$$= \frac{1}{\eta} \mathbb{E}_{s\sim d_\mu^{(t+1)}} \left[\sum_a \pi^{(t+1)}(a \mid s) \log\left(\frac{\pi^{(t+1)}(a \mid s) Z_t(s)}{\pi^{(t)}(a \mid s)} \right) \right].$$

This is:

$$= \frac{1}{\eta} \mathbb{E}_{s\sim d_\mu^{(t+1)}} \left[\mathrm{KL}(\pi_s^{(t+1)} \| \pi_s^{(t)}) + \log Z_t(s) \right].$$

Since KL divergence is non-negative:

$$\geq \frac{1}{\eta} \mathbb{E}_{s\sim d_\mu^{(t+1)}} [\log Z_t(s)].$$

Using the fact that $d_\mu^{(t+1)} \geq (1-\gamma)\mu$ componentwise and $\log Z_t(s) \geq 0$, we get:

$$\mathbb{E}_{s\sim d_\mu^{(t+1)}}[\log Z_t(s)] \geq (1-\gamma) \mathbb{E}_{s\sim\mu}[\log Z_t(s)].$$

Thus:

$$V^{\pi^{(t+1)}}(\mu) - V^{\pi^{(t)}}(\mu) \geq \frac{(1-\gamma)}{\eta} \mathbb{E}_{s\sim\mu}[\log Z_t(s)] \geq 0.$$

This concludes the proof. □

Theorem 62 (Global convergence for Natural Policy Gradient Ascent). *For the softmax policy class, suppose we run the NPG updates using a starting distribution $\mu \in \Delta(\mathcal{S})$ and with initial parameters $\theta^{(0)} = 0$. Fix $\eta > 0$. Then, for all $T > 0$, we have:*

$$V^{\pi^{(T)}}(\mu) \geq V^{\pi^*}(\mu) - \frac{\log|\mathcal{A}|}{\eta T} - \frac{1}{(1-\gamma)^2 T}.$$

Proof: Since π^* and μ are fixed, we use d^* as shorthand for $d_\mu^{\pi^*}$; we also use π_s as shorthand for the vector $\pi(\cdot \mid s)$. By the performance difference lemma, we have:

$$V^{\pi^*}(\mu) - V^{\pi^{(t)}}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^*}\left[\sum_a \pi^*(a \mid s) A^{(t)}(s,a)\right].$$

Using the closed-form update rule of NPG, we rewrite this as:

$$= \frac{1}{\eta} \mathbb{E}_{s\sim d^*}\left[\sum_a \pi^*(a \mid s) \log\left(\frac{\pi^{(t+1)}(a \mid s) Z_t(s)}{\pi^{(t)}(a \mid s)}\right)\right]$$

$$= \frac{1}{\eta} \mathbb{E}_{s\sim d^*}\left[\mathrm{KL}(\pi_s^* \| \pi_s^{(t)}) - \mathrm{KL}(\pi_s^* \| \pi_s^{(t+1)}) + \log Z_t(s)\right].$$

Hence:

$$V^{\pi^*}(\mu) - V^{\pi^{(t)}}(\mu) = \frac{1}{\eta} \mathbb{E}_{s\sim d^*}\left[\mathrm{KL}(\pi_s^* \| \pi_s^{(t)}) - \mathrm{KL}(\pi_s^* \| \pi_s^{(t+1)}) + \log Z_t(s)\right].$$

By applying Lemma 61 with d^* as the starting state distribution, we have:

$$\frac{1}{\eta} \mathbb{E}_{s\sim d^*}[\log Z_t(s)] \geq \frac{1}{1-\gamma}\left(V^{\pi^{(t+1)}}(d^*) - V^{\pi^{(t)}}(d^*)\right),$$

which gives us a bound on $\mathbb{E}_{s\sim d^*}[\log Z_t(s)]$.

Using the above equation and the fact that $V^{\pi^{(t+1)}}(\mu) \geq V^{\pi^{(t)}}(\mu)$ by Lemma 61, we have:

$$V^{\pi^*}(\mu) - V^{\pi^{(T-1)}}(\mu) \leq \frac{1}{T} \sum_{t=0}^{T-1} \left(V^{\pi^*}(\mu) - V^{\pi^{(t)}}(\mu)\right)$$

$$\leq \frac{1}{\eta T} \sum_{t=0}^{T-1} \mathbb{E}_{s\sim d^*}\left[\mathrm{KL}(\pi_s^* \| \pi_s^{(t)}) - \mathrm{KL}(\pi_s^* \| \pi_s^{(t+1)})\right] + \frac{1}{\eta T} \sum_{t=0}^{T-1} \mathbb{E}_{s\sim d^*}[\log Z_t(s)]$$

$$\leq \frac{\mathbb{E}_{s\sim d^*}\left[\mathrm{KL}(\pi_s^* \| \pi_s^{(0)})\right]}{\eta T} + \frac{1}{(1-\gamma)T} \sum_{t=0}^{T-1} \left(V^{\pi^{(t+1)}}(d^*) - V^{\pi^{(t)}}(d^*)\right)$$

$$= \frac{\mathbb{E}_{s\sim d^*}\left[\mathrm{KL}(\pi_s^* \| \pi_s^{(0)})\right]}{\eta T} + \frac{V^{\pi^{(T)}}(d^*) - V^{\pi^{(0)}}(d^*)}{(1-\gamma)T}$$

$$\leq \frac{\log|\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T}.$$

The proof is completed using the fact that $V^{\pi^{(T)}}(\mu) \geq V^{\pi^{(T-1)}}(\mu)$.

□

Remark 13. *Setting $\eta = (1-\gamma)^2 \log|\mathcal{A}|$, we see that NPG finds an ε-optimal policy in a number of iterations that is at most:*

$$T \leq \frac{2}{(1-\gamma)^2 \varepsilon},$$

which has no dependence on the number of states or actions, despite the non-concavity of the underlying optimization problem.

Multi-armed Bandit and Upper Confidence Bound (UCB) Algorithm

We consider the stochastic K-armed bandit problem:

- There are K arms.
- Each arm i yields i.i.d. rewards bounded in $[0,1]$ with unknown mean μ_i.
- Let $\mu^* = \max_i \mu_i$ be the best mean reward.
- Define the gap $\Delta_i = \mu^* - \mu_i \geq 0$.

Let T be the time horizon. The goal is to minimize the expected regret:

$$R(T) = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} r_{a_t}\right] = \sum_{i=1}^{K} \Delta_i \, \mathbb{E}[N_i(T)],$$

where:

- a_t is the arm selected at time t,
- r_{a_t} is the reward at time t,
- $N_i(T)$ is the number of times arm i is pulled up to time T.

UCB Algorithm

At each round t, the UCB algorithm selects the arm:

$$a_t = \arg\max_i \left[\hat{\mu}_i(t) + \sqrt{\frac{2\log t}{N_i(t)}}\right],$$

where:

- $N_i(t)$ is the number of times arm i has been played by time t.
- $\hat{\mu}_i(t) = \frac{\sum_{k=0}^{t} r_{a_k} \mathbf{1}_{\{a_k=i\}}}{N_i(t)}$ is the empirical mean of arm i up to time t,

Proof of Regret Bound

Step 1: Concentration Inequality

We use Hoeffding's inequality. For n i.i.d. samples $X_1, \ldots, X_n \in [0,1]$ with mean μ:

$$\Pr\left(\left|\frac{1}{n}\sum_{j=1}^{n} X_j - \mu\right| \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2).$$

Let the confidence radius be:

$$c_i(t) = \sqrt{\frac{2\log t}{N_i(t)}}.$$

By taking $n = N_i(t)$ and $\epsilon = c_i(t)$ in the above Hoeffding's inequality, with high probability $1 - \frac{2}{t^4}$, the empirical mean satisfies:

$$|\hat{\mu}_i(t) - \mu_i| \leq c_i(t).$$

Step 2: When Is a Suboptimal Arm Selected?

Suppose arm i is suboptimal ($\Delta_i > 0$), and it is selected at time t. Then:

$$\hat{\mu}_i(t) + c_i(t) \geq \hat{\mu}_{i^*}(t) + c_{i^*}(t),$$

where i^* is the optimal arm. Assuming the concentration bounds hold for both i and i^* (which by union bound happens with probability at least $1 - \frac{4}{t^4}$), we have:

$$\hat{\mu}_{i^*}(t) \geq \mu^* - c_{i^*}(t), \quad \hat{\mu}_i(t) \leq \mu_i + c_i(t).$$

So:

$$\mu_i + 2c_i(t) \geq \mu^* \Rightarrow 2c_i(t) \geq \Delta_i.$$

Solving:

$$2\sqrt{\frac{2\log t}{N_i(t)}} \geq \Delta_i \Rightarrow N_i(t) \leq \frac{8\log t}{\Delta_i^2}.$$

Thus, if $N_i(t) > \frac{8\log T}{\Delta_i^2}$, arm i will not be selected again (with high probability).

Step 3: Bounding Expected Pulls

We split $\mathbb{E}[N_i(T)]$ into:

$$\mathbb{E}[N_i(T)] \leq \frac{8\log T}{\Delta_i^2} + \sum_{t=1}^{T} \Pr(\text{concentration fails at time } t).$$

Each failure has probability $\leq \frac{4}{t^4}$ (from Hoeffding's inequality with union bound), so:

$$\sum_{t=1}^{T} \frac{4}{t^4} \leq \frac{\pi^4}{90} < 8.$$

So:

$$\mathbb{E}[N_i(T)] \leq \frac{8\log T}{\Delta_i^2} + 8.$$

Step 4: Total Regret

Substituting into the regret formula:

$$R(T) = \sum_{i:\Delta_i>0} \Delta_i \cdot \mathbb{E}[N_i(T)] \leq \sum_{i:\Delta_i>0} \left(\frac{8\log T}{\Delta_i} + 8\Delta_i \right).$$

Thus, the regret satisfies:

$$\boxed{R(T) = \mathcal{O}\left(\sum_{i:\Delta_i>0} \frac{\log T}{\Delta_i} \right)}$$

This is a logarithmic-in-T regret, which is order-optimal for stochastic bandits.

Episodic Reinforcement Learning (RL)

We consider a finite-horizon episodic MDP:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H),$$

with:

- State space $\mathcal{S}$, action space $\mathcal{A}$.
- Unknown transition kernel $P(s' \mid s, a)$.
- Known reward function $r(s,a) \in [0,1]$.
- Horizon length H, and K episodes.

Therefore, in each episode, we are dealing with a finite horizon MDP of the form

$$\max_{\{\pi_\ell\}_{\ell=0}^{H-1}} \mathbb{E}[\sum_{\ell=0}^{H-1} r(S_\ell, A_\ell)],$$

where for simplicity, we assume the reward function $r(s,a)$ is deterministic, and $\gamma = 1$. Note that in the finite horizon MDP, the optimal policy is not necessary stationary and is Markovian of the form $\{\pi_\ell\}_{\ell=0}^{H-1}$.

Recall how to solve finite horizon MDP by solving the Bellman optimality equations:

$$\begin{aligned} &V_H^*(s) = 0 \ \forall s \\ &V_k^*(s) = \max_a \left\{ r(s,a) + \mathbb{E}[V_{k+1}^*(s') | s_k = s, a_k = a] \right\} \ \ k = 0, 1, \ldots, H-1. \end{aligned}$$

Here, optimal value function is V_0^* and the optimal policy is a Markovian deterministic policy. Note that in RL, the goal is to solve this MDP without knowing the transition matrix $P(\cdot|\cdot,\cdot)$.

Each episode consists of running the MDP from 0 to H, and we repeat each episode with a new policy. Let k denote the kth episode started from some initial state s_0^k, and let $\pi^k = (\pi_0^k, \ldots, \pi_{H-1}^k)$ be the policy used in episode k. The goal is to achieve a regret defined by:

$$R(K) = \sum_{k=1}^{K} \left(V_0^*(s_0^k) - V_0^{\pi^k}(s_0^k) \right),$$

where V_0^* is the optimal value function.

Upper Confidence Reinforcement Learning (UCRL) Algorithm

At the end of each episode k, do the following:

1. Estimate Transitions

For each (s,a), compute empirical transition probability:

$$\hat{P}_k(s' \mid s, a) = \frac{N_k(s,a,s')}{\max\{1, N_k(s,a)\}},$$

where $N_k(s,a)$ is the total number of times (s,a) was visited during the fixed episode k, and $N_k(s,a,s')$ is the number of times during episode k that the MDP transitioned from s to s' when action a was taken in state s.

2. Construct Confidence Intervals

In episode $k+1$, compute a confidence set $C_{k,\delta}$ around $\hat{P}_k$ for transition probabilities as

$$C_{k,\delta} = \left\{P : \ \|P(\cdot \mid s,a) - \hat{P}_k(\cdot \mid s,a)\|_1 \leq f_\delta(N_k(s,a)) \ \forall s,a\right\},$$

where $f_\delta(N_k(s,a))$ is chosen such that $P^* \in C_{k,\delta} \forall k$, with probability at least $1-\delta$. In particular, one can take

$$f_\delta(N_k(s,a)) = \sqrt{\frac{2\log\left(2S^2 A N_k^2(s,a)/\delta\right)}{\max\{1, N_k(s,a)\}}}. \tag{15}$$

3. Optimistic Policy Update

Update the policy to $\pi^{k+1} = (\pi_0^{k+1}, \dots, \pi_{H-1}^{k+1})$, where $\pi^{k+1} = \arg\max_\pi \max_{P \in C_{k,\delta}} V_0^\pi(s_0^{k+1})$. Run policy π^{k+1} in episode $k+1$, collect trajectory, and update statistics.

Remark 14. *Note the similarity between UCRL and MAB, where we optimistically choose the best policy within a confidence set, which is called "optimism in the face of uncertainty".*

Regret Sketch

Step 1: Optimism

With high probability, the true MDP lies in the confidence set. More precisely:

Lemma 63. *Let P^* be the true (unknown) transition matrix of the MDP. By choosing $f_\delta(\cdot)$ as in* (15)*, with probability at least $1-\delta$, we have $P^* \in C_{k,\delta}, \ \forall k$.*

Thus, with high probability, the policy computed using the optimistic model achieves:

$$V_0^{\pi_k}(s_0^k) \geq V_0^*(s_0^k) - \text{error}_k.$$

Step 2: Decomposition

The regret decomposes into:

$$R(K) = \sum_{k=1}^{K} \left(V_0^*(s_0^k) - V_0^{\pi_k}(s_0^k)\right) \leq \sum_{k=1}^{K} \sum_{h=0}^{H-1} \text{bonus}_k(s_h^k, a_h^k),$$

where each bonus is controlled by $f_\delta(N_k(s,a))$.

Finally, using some martingale analysis, we can bound the total number of visits to uncertain state-action pairs, and in particular each $\text{bonus}_k(s_h^k, a_h^k)$ with high probability. Combining all the components, one can show the following result.

Theorem 64. *With high probability $1-\delta := 1 - \frac{1}{\sqrt{K}}$, the regret of the UCRL is bounded by*

$$R(K) = \tilde{\mathcal{O}}\left(H^{\frac{3}{2}} S \sqrt{AK}\right),$$

where $\tilde{\mathcal{O}}$ hides the logarithmic factors.