

# CMPT 419/983: Theoretical Foundations of Reinforcement Learning

## Lecture 10

---

Sharan Vaswani

November 10, 2023

# Recap

- Given a policy parameterization s.t.  $\pi = h(\theta)$  and a step-size  $\eta$ , policy gradient methods have the following update:  $\theta_{t+1} = \theta_t + \eta \nabla_{\theta} J(\theta_t)$  where  $J(\theta) := v^{\pi_{\theta}}(\rho) = \mathbb{E}_{s_0 \sim \rho} v^{\pi_{\theta}}(s_0)$ .
- Policy Gradient Theorem:**  $\nabla_{\theta} J(\theta) = \frac{\partial v^{\pi_{\theta}}(\rho)}{\partial \theta} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_{\theta}}} \left[ \sum_{a \in \mathcal{A}} \frac{\partial \pi_{\theta}(a|s)}{\partial \theta} q^{\pi_{\theta}}(s, a) \right]$ .
- Consider function  $h : \mathbb{R}^A \rightarrow \mathbb{R}^A$  such that  $h(\theta) = \pi_{\theta}$  where  $\pi_{\theta}(a) = \frac{\exp(\theta(a))}{\sum_{a'} \exp(\theta(a'))}$ . The Jacobian of  $h$  is given by  $H(\pi_{\theta}) \in \mathbb{R}^{A \times A} = \text{diag}(\pi_{\theta}) - \pi_{\theta} \pi_{\theta}^T$ .
- Tabular softmax policy parameterization:** There are **SA parameters** such that  $\pi_{\theta}(\cdot|s) = h(\theta(s, \cdot))$ . In this case,  $[\nabla J(\theta)]_{s,a} = \frac{\partial v^{\pi_{\theta}}(\rho)}{\partial \theta(s,a)} = \frac{d^{\pi_{\theta}}(s)}{1-\gamma} \pi_{\theta}(a|s) a^{\pi_{\theta}}(s, a)$ , where  $a^{\pi}(s, a) = q^{\pi}(s, a) - v^{\pi}(s)$  is the advantage function.
- For the **bandit setting with deterministic rewards**,  $J(\theta) = \mathbb{E}_{a \sim \pi_{\theta}} [r(a)] = \langle \pi_{\theta}, r \rangle$  and  $[\nabla J(\theta)]_a = \frac{\partial v^{\pi_{\theta}}(\rho)}{\partial \theta(a)} = \pi_{\theta}(a) [r(a) - \langle \pi_{\theta}, r \rangle]$ . Hence, the corresponding policy gradient update is:  $\theta_{t+1} = \theta_t + \eta \pi_{\theta_t}(a) [r(a) - \langle \pi_{\theta_t}, r \rangle]$ .

# Softmax Policy Gradient for Bandits

**Claim:** For the tabular softmax policy parameterization where  $\pi_\theta(a) = \frac{\exp(\theta(a))}{\sum_{a'} \exp(\theta(a'))}$ , the objective  $J(\theta) = \langle \pi_\theta, r \rangle$  can be non-concave w.r.t  $\theta$ .

*Proof:* Recall that a function  $f : \mathcal{D} \rightarrow \mathbb{R}$  is concave if for all  $\theta, \theta' \in \mathcal{D}$  and  $\alpha \in [0, 1]$ ,  $f(\alpha\theta + (1 - \alpha)\theta') \geq \alpha f(\theta) + (1 - \alpha)f(\theta')$ . Consider a multi-armed bandit problem where  $A = 3$ , and  $r = [1, 9/10, 1/10]$ ,  $\theta = [0, 0, 0]$  and  $\theta' = [\ln(9), \ln(16), \ln(25)]$ . Choosing  $\alpha = \frac{1}{2}$ ,

$$\pi = h(\theta) = [1/3, 1/3, 1/3] \implies J(\theta) = \frac{1}{3} + \frac{3}{10} + \frac{1}{30} = \frac{2}{3}$$

$$\pi' = h(\theta') = [9/50, 16/50, 25/50] \implies J(\theta') = \frac{90}{500} + \frac{144}{500} + \frac{25}{500} = \frac{259}{500}$$

$$\implies \text{RHS} = \alpha J(\theta) + (1 - \alpha)J(\theta') = \frac{1}{2} \left( \frac{2}{3} + \frac{259}{500} \right) = \frac{1777}{3000}$$

$$\alpha\theta + (1 - \alpha)\theta' = [\ln(3), \ln(4), \ln(5)] \implies h(\alpha\theta + (1 - \alpha)\theta') = [3/12, 4/12, 5/12]$$

$$\implies \text{LHS} = J(\alpha\theta + (1 - \alpha)\theta') = \frac{3}{12} + \frac{36}{120} + \frac{5}{120} = \frac{71}{120}$$

$\text{RHS} = \frac{1777}{3000} = \frac{14216}{24000} > \frac{14200}{24000} = \text{LHS}$ , meaning that  $J(\theta)$  is non-concave for this example.

## Digression – Smooth functions

**Smooth functions:** For smooth functions that are differentiable everywhere, the gradient is Lipschitz-continuous i.e. it can not change arbitrarily fast.

- Formally, the gradient  $\nabla f$  is  $L$ -Lipschitz continuous if for all  $x, y \in \mathcal{D}$ ,

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$$

where  $L$  is the Lipschitz constant of the gradient (also called the smoothness constant of  $f$ ).

- If  $f$  is twice-differentiable and smooth, then for all  $x \in \mathcal{D}$ ,  $\nabla^2 f(x) \preceq L I_d$  i.e.  $\sigma_{\max}[\nabla^2 f(x)] \leq L$  where  $\sigma_{\max}$  is the maximum singular value.

- For  $L$ -smooth functions, for all  $x, y \in \mathcal{D}$ ,

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2$$

Hence the function  $f(y)$  is upper and lower-bounded by quadratics:

$f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$  and  $f(x) + \langle \nabla f(x), y - x \rangle - \frac{L}{2} \|y - x\|^2$  respectively.

These bounds are *global* and hold for all  $y \in \mathcal{D}$ .

# Softmax Policy Gradient

**Fact:** For the tabular softmax policy parameterization where  $\pi_\theta = h(\theta)$  i.e.

$\pi_\theta(a) = \frac{\exp(\theta(a))}{\sum_{a'} \exp(\theta(a'))}$ , the objective  $J(\theta) = \langle \pi_\theta, r \rangle$  is  $\frac{5}{2}$ -smooth.

See [MXSS20, Lemma 2] for a proof. Such a smoothness property also holds for general MDPs (see [MXSS20, Lemma 7]).

- By putting together these results, we conclude that for the tabular softmax policy parameterization, the objective  $J(\theta)$  is a smooth, non-concave function.
- Hence, in general (without additional properties), softmax PG is not guaranteed to converge to the optimal policy, but only to a stationary point where  $\|\nabla_\theta J(\theta)\| = 0$ . Assuming that we can exactly calculate  $\nabla_\theta J(\theta)$ , we can prove the following result from non-convex optimization.

**Claim:** For the tabular policy parameterization where  $J(\theta)$  is  $L$ -smooth w.r.t  $\theta$ , softmax PG with  $\eta = \frac{1}{L}$  returns  $\hat{\theta}_T$  such that  $\left\| \nabla J(\hat{\theta}_T) \right\|^2 \leq \epsilon$  and requires  $T = \frac{2L}{(1-\gamma)\epsilon}$  iterations.

# Stationary point Convergence of Softmax Policy Gradient

*Proof:* Using the  $L$ -smoothness of  $J$  with  $x = \theta_t$  and  $y = \theta_{t+1} = \theta_t + \frac{1}{L} \nabla J(\theta_t)$  in the quadratic bound (also referred to as the *ascent lemma*),

$$\begin{aligned} J(\theta_{t+1}) &\geq J(\theta_t) + \left\langle \nabla J(\theta_t), \frac{1}{L} \nabla J(\theta_t) \right\rangle - \frac{L}{2} \left\| \frac{1}{L} \nabla J(\theta_t) \right\|^2 \\ \implies J(\theta_{t+1}) &\geq J(\theta_t) + \frac{1}{2L} \|\nabla J(\theta_t)\|^2 \end{aligned}$$

By moving from  $\theta_t$  to  $\theta_{t+1}$ , the algorithm has increased the value of  $J$ . Rearranging the inequality, for every iteration  $t$ ,

$$\frac{1}{2L} \|\nabla J(\theta_t)\|^2 \leq J(\theta_{t+1}) - J(\theta_t)$$

Summing up from  $t = 0$  to  $T - 1$ ,

$$\frac{1}{2L} \sum_{t=0}^{T-1} \|\nabla J(\theta_t)\|^2 \leq \sum_{t=0}^{T-1} [J(\theta_{t+1}) - J(\theta_t)] = J(\theta_T) - J(\theta_0)$$

## Stationary point Convergence of Softmax Policy Gradient

Recall that  $\frac{1}{2L} \sum_{t=0}^{T-1} \|\nabla J(\theta_t)\|^2 \leq J(\theta_T) - J(\theta_0)$ . Since  $J(\theta) \in \left[0, \frac{1}{1-\gamma}\right]$  for all  $\theta$ ,

$$\frac{\sum_{t=0}^{T-1} \|\nabla J(\theta_t)\|^2}{T} \leq \frac{2L}{(1-\gamma) T}$$

Define  $\hat{\theta}_T := \arg \min_{t \in \{0, 1, \dots, T-1\}} \|\nabla J(\theta_t)\|^2$ .

$$\|\nabla J(\hat{\theta}_T)\|^2 \leq \frac{2L}{(1-\gamma) T}$$

If the RHS equal to  $\frac{2L}{(1-\gamma) T} \leq \epsilon$ , this would guarantee that  $\|\nabla J(\hat{\theta}_T)\|^2 \leq \epsilon$  and we would achieve our objective. Hence, we need to run the algorithm for  $T \geq \frac{2L}{(1-\gamma)\epsilon}$  iterations.

Next, we will see that for the tabular softmax policy parameterization,  $J(\theta)$  satisfies an additional gradient domination property that allows us to prove convergence to the optimal policy.

# Non-uniform Lojasiewicz condition for Bandits

**Claim:** For a **bandit problem** with **deterministic rewards**, where  $J(\theta) = \langle \pi_\theta, r \rangle$ , assuming that there is a unique optimal action  $a^*$  and  $\pi^* := \arg \max_\pi \langle \pi, r \rangle$  is the optimal policy, then,

$$\left\| \frac{\partial J(\theta)}{\partial \theta} \right\| \geq \pi_\theta(a^*) [\langle \pi^*, r \rangle - \langle \pi_\theta, r \rangle] = \pi_\theta(a^*) [\langle \pi^*, r \rangle - J(\theta)]$$

- The result implies that if  $\pi_\theta(a^*) > 0$ , as  $\|\nabla_\theta J(\theta)\| \rightarrow 0$ ,  $J(\theta) \rightarrow \langle \pi^*, r \rangle$ . Hence, decreasing the gradient norm of  $J(\theta)$  is sufficient for global convergence to the optimal value function.
- The property does not rely on the concavity of the objective, and hence characterizes a special class of non-concave functions that can be maximized to the optimum.
- The inequality is an instance of the *Lojasiewicz* or *gradient domination condition*. Function  $f$  satisfies a gradient domination with parameters  $(C, \zeta)$  if:  $\|\nabla_\theta f(\theta)\| \geq C [f^* - f(\theta)]^\zeta$ .
- For the above inequality,  $C = \pi_\theta(a^*)$ . Since the condition depends on  $\theta$ , it is non-uniform. The dependence on  $\pi_\theta(a^*)$  is necessary [MXSS20, Remark 1].
- $\zeta = \frac{1}{2}$  is more common in non-convex optimization, and is referred to as the *Polyak Lojasiewicz condition*.



# Non-uniform Lojasiewicz condition for Bandits

*Proof:* Recall that for any action  $a$ ,  $\frac{\partial J(\theta)}{\partial \theta(a)} = \pi_\theta(a) [r(a) - \langle \pi_\theta, r \rangle]$ . Hence,

$$\begin{aligned} \left\| \frac{\partial J(\theta)}{\partial \theta} \right\|^2 &= \sum_a [\pi_\theta(a)]^2 [r(a) - \langle \pi_\theta, r \rangle]^2 \geq [\pi_\theta(a^*)]^2 [r(a^*) - \langle \pi_\theta, r \rangle]^2 \\ &= [\pi_\theta(a^*)]^2 [r(a^*) - J(\theta)]^2 = [\pi_\theta(a^*)]^2 [\langle \pi^*, r \rangle - J(\theta)]^2 \\ \implies \left\| \frac{\partial J(\theta)}{\partial \theta} \right\| &\geq \pi_\theta(a^*) [\langle \pi^*, r \rangle - J(\theta)] \quad \square \end{aligned}$$

• Recall the stationary point convergence – tabular softmax PG returns a point  $\hat{\theta}_T$  such that  $\left\| \nabla J(\hat{\theta}_T) \right\|^2 \leq \frac{2L}{(1-\gamma)T}$ . Combining with the above Lojasiewicz condition,

$$\pi_{\hat{\theta}_T}(a^*) [\langle \pi^*, r \rangle - J(\hat{\theta}_T)] \leq \sqrt{\frac{2L}{(1-\gamma)T}} \implies \langle \pi^*, r \rangle - J(\hat{\theta}_T) \leq \frac{1}{\pi_{\hat{\theta}_T}(a^*)} \sqrt{\frac{2L}{(1-\gamma)T}}$$

• Hence, softmax PG (with the tabular parameterization) will converge to the optimal arm at an  $O(1/\sqrt{T})$  rate if  $\pi_{\hat{\theta}_T}(a^*) \neq 0$ .

# Global Convergence of Softmax Policy Gradient

**Fact:** For tabular softmax PG with step-size  $\eta = \frac{1}{L}$  and a uniform initialization ( $\forall a, \pi_0(a) = \frac{1}{A}$ ) ensures that  $\min_{t \geq 0} \pi_{\theta_t}(a^*) \geq \frac{1}{A} > 0$  (see [MXSS20, Lemma 5] for a proof).

- We have established that for the multi-armed bandit setting, tabular softmax PG (with exact gradients) can converge to the optimal arm at an  $O(1/\sqrt{T})$  rate.

**Q:** Where is the exploration? **Ans:** All arms are pulled to construct the gradient. So no exploration is required.

**Fact:** For general MDPs, if  $\pi^*$  is the optimal policy corresponding to taking action  $a^*(s)$  in state  $s$ , then the objective  $J(\theta) = \mathbb{E}_{s_0 \sim \rho} v^{\pi_\theta}(s_0)$  satisfies a non-uniform Lojasiewicz condition:

$$\left\| \frac{\partial J(\theta)}{\partial \theta} \right\| \geq \frac{\min_{s \in \mathcal{S}} \pi_\theta(a^*(s)|s)}{\sqrt{S} \left\| d^{\pi^*} / d^{\pi_\theta} \right\|_\infty} [v^{\pi^*}(\rho) - J(\theta)]$$

- Similar to the bandit case, there is a dependence on  $\pi_\theta(a^*(s)|s)$ , but now for each state.
- There is a dependence on the *distribution mismatch coefficient*  $\left\| d^{\pi^*} / d^{\pi_\theta} \right\|_\infty$ .

# Global Convergence of Softmax Policy Gradient

Recall that  $\left\| \frac{\partial J(\theta)}{\partial \theta} \right\| \geq \frac{\min_{s \in \mathcal{S}} \pi_{\theta}(a^*(s)|s)}{\sqrt{S} \|d^{\pi^*}/d^{\pi_{\theta}}\|_{\infty}} [v^{\pi^*}(\rho) - J(\theta)]$ .

- Define  $\mathcal{S}^* = \{s \in \mathcal{S} | d^{\pi^*}(s) \neq 0\}$ . For the distribution mismatch coefficient to be bounded, we want that  $d^{\pi_{\theta}}(s) \neq 0$  for all  $s \in \mathcal{S}^*$ . Hence, the algorithm needs to have a non-zero probability of visiting states in  $\mathcal{S}^*$  and requires sufficient exploration to ensure this. The distribution mismatch coefficient thus captures the need for policy gradient algorithms to explore the state space.
- The dependence on the mismatch coefficient is necessary for the non-uniform Lojasiewicz condition and hence for global convergence to the optimal policy [MXSS20, Proposition 3].
- A practical way to guarantee that the distribution mismatch coefficient is bounded is to ensure that  $\rho(s) \neq 0$  for all  $s \in \mathcal{S}$ . This may not always be feasible, and exploration with policy gradient is problematic. See [AHKS20, CYJW20, LWG<sup>+</sup>23] for some recent attempts to handle this.
- Using a uniform distribution over actions for each state also ensures that  $\min_s \pi_{\theta}(a^*(s)|s) > 0$ . With these settings, softmax PG can be shown to converge to the optimal policy  $\pi^*$  at an  $O(1/T)$  rate [MXSS20, Theorem 4].

# Global Convergence of Softmax Policy Gradient

**Claim:** Assuming  $J(\theta)$  is  $L$ -smooth and satisfies the Lojasiewicz condition with constant  $\mu$  i.e.  $\|\nabla J(\theta)\| \geq \mu [v^{\pi^*}(\rho) - J(\theta)]$ , softmax PG with the tabular policy parameterization, uniform initialization,  $\eta = \frac{1}{L}$  and  $T$  iterations converges as:  $\delta_T \leq \frac{2L}{\mu^2 T}$ , where  $\delta_t := v^{\pi^*}(\rho) - J(\theta_t)$ .

*Proof:* Using the  $L$ -smoothness of  $J(\theta)$  and the update as before,

$$J(\theta_{t+1}) \geq J(\theta_t) + \frac{1}{2L} \|\nabla J(\theta_t)\|^2 \geq J(\theta_t) + \frac{\mu^2}{2L} [v^{\pi^*}(\rho) - J(\theta_t)]^2 \quad (\text{Lojasiewicz condition})$$

$$\implies \delta_{t+1} \leq \delta_t - \frac{\mu^2}{2L} \delta_t^2 \implies \frac{1}{\delta_t} \leq \frac{1}{\delta_{t+1}} - \frac{\mu^2}{2L} \frac{\delta_t}{\delta_{t+1}} \implies \frac{\mu^2}{2L} \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}$$

(Dividing by  $\delta_t \delta_{t+1}$ , and using that  $\delta_t \geq \delta_{t+1}$ )

$$\implies \frac{\mu^2 T}{2L} \leq \sum_{t=0}^{T-1} \left[ \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \right] \leq \frac{1}{\delta_T} \implies \delta_T \leq \frac{2L}{\mu^2 T} \quad \square$$

- For bandit problems,  $\mu = \min_{t \geq 0} \pi_{\theta_t}(a^*)$  and for general MDPs,  $\mu = \min_{t \geq 0} \frac{\min_{s \in \mathcal{S}} \pi_{\theta_t}(a^*(s)|s)}{\sqrt{S} \|d^{\pi^*}/d^{\pi_{\theta_t}}\|_{\infty}}$ .
- The  $O(1/T)$  rate is tight for softmax PG and cannot be improved [MXSS20, Theorems 9, 10].

## Natural Policy Gradient

# Natural Policy Gradient

- Softmax PG has a slow  $O(1/\tau)$  rate of convergence (even when using exact gradients). On the other hand, policy iteration has a linear  $O(\exp(-T))$  convergence rate.
- Natural Policy Gradient (NPG) overcomes this shortcoming of softmax PG, and achieves a linear rate of convergence. NPG is an instantiation of *preconditioned gradient ascent*.
- For a general symmetric, positive definite matrix  $Q$ , preconditioned gradient ascent on  $J(\theta)$  can be written as:  $\theta_{t+1} = \theta_t + \eta Q \nabla_{\theta} J(\theta)$ .
- Preconditioned gradient ascent is equivalent to the update that “follows” the direction of the gradient, but stays “close” to the previous iterate  $\theta_t$  in the norm induced by  $Q^{-1}$ , i.e.

$$\theta_{t+1} = \arg \max_{\theta} \left[ \langle \nabla_{\theta} J(\theta), \theta \rangle - \frac{1}{2\eta} \|\theta - \theta_t\|_{Q^{-1}}^2 \right].$$

- Preconditioning is equivalent to reparameterizing the space so that the maximum remains the same, but the function becomes easier to optimize, enabling gradient ascent to converge faster.

# Natural Policy Gradient

NPG chooses the preconditioner  $Q$  to be the (pseudo)-inverse of the *Fisher information matrix*:  $F_\theta \in \mathbb{R}^{d \times d}$  (where  $d$  is the dimension of the parameter  $\theta$ ):

F是二阶Taylor展开的二次项

$$F_\theta := \mathbb{E}_{s \sim d^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} [\nabla_\theta \log(\pi_\theta(a|s)) \nabla_\theta \log(\pi_\theta(a|s))^\top] = \mathbb{E}_{s \sim d^{\pi_\theta}} \left[ \frac{\partial^2 \text{KL}(\pi_\theta || \pi_{\theta'})}{\partial \theta'^2} \right]_{\theta'=\theta}$$

- $F_\theta$  is symmetric, positive semi-definite and corresponds to the Hessian of the KL divergence.
- $F_\theta$  is also the covariance of the score function  $\frac{\partial \ln(\pi_\theta(a|s))}{\partial \theta}$  and determines the amount of information the observed data has about the true (unknown) parameter generating the data.
- The NPG update can be written as:  $\theta_{t+1} = \theta_t + \eta F_{\theta_t}^\dagger \nabla J(\theta_t)$ .
- Next, we will instantiate the NPG update for the tabular softmax policy parameterization, and prove that preconditioning by  $F_\theta^\dagger$  enables NPG to converge at a faster  $\exp(-T)$  rate, compared to softmax PG.

# Natural Policy Gradient for Softmax Parametrization

For the tabular softmax policy parameterization,  $\theta \in \mathbb{R}^{SA}$  and  $\pi_\theta(\cdot|s) = h(\theta(s, \cdot))$  such that  $\pi_\theta(a|s) = \frac{\exp(\theta(a|s))}{\sum_{a'} \exp(\theta(a'|s))}$ . Recall that  $\frac{\partial \pi_\theta(\cdot|s)}{\partial \theta(s, \cdot)} = H(\pi_\theta(\cdot|s)) = \text{diag}(\pi_\theta) - \pi_\theta \pi_\theta^T$ .

We can calculate the Jacobian element-wise, and derive the following relation: for any  $s', a'$ ,  $\frac{\partial \log(\pi_\theta(a'|s'))}{\partial \theta(s, a)} = \mathcal{I}\{s' = s\} [\mathcal{I}\{a' = a\} - \pi_\theta(a|s)]$ . (Prove in Assignment 4!).

**Claim:** For the tabular softmax policy parameterization,  $[F_\theta^\dagger \nabla J(\theta)]_{s,a} = \frac{\alpha^{\pi_\theta(s,a)}}{1-\gamma}$ .

*Proof:* Define  $w_\theta := \arg \min_w \|F_\theta w - \nabla J(\theta)\|^2 = F_\theta^\dagger \nabla J(\theta)$ . First, let us calculate  $F_\theta w$  for a general  $w \in \mathbb{R}^{SA}$ .

$$\begin{aligned} F_\theta w &= \mathbb{E}_{s' \sim d^{\pi_\theta}} \mathbb{E}_{a' \sim \pi_\theta(\cdot|s')} [\nabla_\theta \log(\pi_\theta(a'|s')) \nabla_\theta \log(\pi_\theta(a'|s'))^T] w \\ &= \sum_{s'} d^{\pi_\theta}(s') \sum_{a'} \pi_\theta(a'|s') [\nabla_\theta \log(\pi_\theta(a'|s')) \nabla_\theta \log(\pi_\theta(a'|s'))^T] w \\ F_\theta w &= \sum_{s'} d^{\pi_\theta}(s') \sum_{a'} \pi_\theta(a'|s') \underbrace{\langle \nabla_\theta \log(\pi_\theta(a'|s')), w \rangle}_{:= C(s', a')} \nabla_\theta \log(\pi_\theta(a'|s')) \end{aligned}$$



# Natural Policy Gradient for Softmax Parametrization

$F_{\theta} w = \sum_{s'} d^{\pi_{\theta}}(s') \sum_{a'} \pi_{\theta}(a'|s') C(s', a') \nabla_{\theta} \log(\pi_{\theta}(a'|s'))$  where  $C(s', a') = \langle \nabla_{\theta} \log(\pi_{\theta}(a'|s')), w \rangle$ .

Recall that,

$$[\nabla_{\theta} \log(\pi_{\theta}(a'|s'))]_{s,a} = \frac{\partial \log(\pi_{\theta}(a'|s'))}{\partial \theta(s, a)} = \mathcal{I} \{s' = s\} [\mathcal{I} \{a' = a\} - \pi_{\theta}(a|s)]$$

$$\implies [F_{\theta} w]_{s,a} = d^{\pi_{\theta}}(s) \sum_{a'} \pi_{\theta}(a'|s) C(s, a') [\mathcal{I} \{a' = a\} - \pi_{\theta}(a|s)]$$

$$C(s', a') = \langle \nabla_{\theta} \log(\pi_{\theta}(a'|s')), w \rangle = \sum_{\tilde{s}, \tilde{a}} \frac{\partial \log(\pi_{\theta}(a'|s'))}{\partial \theta(\tilde{s}, \tilde{a})} w(\tilde{s}, \tilde{a})$$

$$= \sum_{\tilde{s}, \tilde{a}} \mathcal{I} \{s' = \tilde{s}\} [\mathcal{I} \{a' = \tilde{a}\} - \pi_{\theta}(\tilde{a}|\tilde{s})] w(\tilde{s}, \tilde{a}) = \sum_{\tilde{a}} w(s', \tilde{a}) [\mathcal{I} \{a' = \tilde{a}\} - \pi_{\theta}(\tilde{a}|s')]$$

$$\implies C(s', a') = w(s', a') - \underbrace{\langle \pi_{\theta}(\cdot|s'), w(s', \cdot) \rangle}_{:=c(s')}$$

$$\implies [F_{\theta} w]_{s,a} = d^{\pi_{\theta}}(s) \sum_{a'} \pi_{\theta}(a'|s) \left[ [w(s, a') - c(s)] [\mathcal{I} \{a' = a\} - \pi_{\theta}(a|s)] \right]$$

# Natural Policy Gradient for Softmax Parametrization

Recall that  $[F_\theta w]_{s,a} = d^{\pi_\theta}(s) \sum_{a'} \pi_\theta(a'|s) \left[ [w(s, a') - c(s)] [\mathcal{I}\{a' = a\} - \pi_\theta(a|s)] \right]$  where  $c(s) := \langle \pi_\theta(\cdot|s), w(s, \cdot) \rangle$ . Simplifying,

$$\begin{aligned} [F_\theta w]_{s,a} &= d^{\pi_\theta}(s) \sum_{a'} \pi_\theta(a'|s) [w(s, a') \mathcal{I}\{a' = a\} - c(s) \mathcal{I}\{a' = a\} - w(s, a') \pi_\theta(a|s) + c(s) \pi_\theta(a|s)] \\ &= d^{\pi_\theta}(s) \left[ \pi_\theta(a|s) w(s, a) - \pi_\theta(a|s) c(s) - \pi_\theta(a|s) \sum_{a'} \pi_\theta(a'|s) w(s, a') + c(s) \pi_\theta(a|s) \right] \\ &= d^{\pi_\theta}(s) \pi_\theta(a|s) [w(s, a) - c(s)] \quad (\text{Since } \sum_{a'} \pi_\theta(a'|s) w(s, a') = c(s)) \end{aligned}$$

$$\begin{aligned} \|F_\theta w - \nabla J(\theta)\|^2 &= \sum_s \sum_a [[F_\theta w]_{s,a} - [\nabla J(\theta)]_{s,a}]^2 \\ &= \sum_s \sum_a \left[ d^{\pi_\theta}(s) \pi_\theta(a|s) \left( w(s, a) - c(s) - \frac{\alpha^{\pi_\theta}(s, a)}{1 - \gamma} \right) \right]^2 \end{aligned}$$

(Using the expression for the policy gradient for the tabular softmax parameterization)

# Natural Policy Gradient for Softmax Parametrization

Recall that  $F_\theta^\dagger \nabla J(\theta) = w_\theta = \arg \min_w \|F_\theta w - \nabla J(\theta)\|^2$

$$= \arg \min_w \sum_s \sum_a \left[ d^{\pi_\theta}(s) \pi_\theta(a|s) \left( w(s, a) - \sum_{a'} \pi_\theta(a'|s) w(s, a') - \frac{\alpha^{\pi_\theta}(s, a)}{1 - \gamma} \right) \right]^2$$

Setting  $w_{s,a} = \frac{\alpha^{\pi_\theta}(s,a)}{1-\gamma}$  ensures that each  $(s, a)$  term is zero since  $\sum_{a'} \alpha^{\pi_\theta}(s, a') \pi_\theta(a'|s) = 0$

$$\implies [F_\theta^\dagger \nabla J(\theta)]_{s,a} = \frac{\alpha^{\pi_\theta}(s, a)}{1 - \gamma}$$

Comparing the preconditioned gradient to the softmax policy gradient  $d^\pi(s) \pi_\theta(a|s) \frac{\alpha^{\pi_\theta}(s,a)}{1-\gamma}$ ,

- The preconditioned gradient does not depend on  $d^\pi(s)$  or  $\pi_\theta(a|s)$ .
- As  $\pi_\theta \rightarrow \pi^*$ , for  $a \neq a^*(s)$ ,  $\pi_\theta(a|s) \rightarrow 0$ . Consequently,  $\|\nabla_\theta J(\theta)\|$  becomes smaller with increasing number of iterations and the resulting method becomes slower.
- Since the NPG update does not depend on  $\pi_\theta(a|s)$ , it does not suffer from the above problem, resulting in faster convergence.

-  Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun, *Pc-pg: Policy cover directed exploration for provable policy gradient learning*, Advances in neural information processing systems **33** (2020), 13399–13412.
-  Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang, *Provably efficient exploration in policy optimization*, International Conference on Machine Learning, PMLR, 2020, pp. 1283–1294.
-  Qinghua Liu, Gellért Weisz, András György, Chi Jin, and Csaba Szepesvári, *Optimistic natural policy gradient: a simple efficient policy optimization framework for online rl*, arXiv preprint arXiv:2305.11032 (2023).
-  Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans, *On the global convergence rates of softmax policy gradient methods*, International Conference on Machine Learning, PMLR, 2020, pp. 6820–6829.