# The Average-Reward Policy Gradient Theorem

The Bellman Route

**Sai Sampath Kedari**
sampath@umich.edu

---

## 1 Average-Reward Policy Gradient Theorem

The objective is the long-run reward rate,

$$J_{avR}(\theta) = r(\pi_\theta) = \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}^{\pi_\theta}[R_t \mid S_0] = \sum_s \mu_{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s)\, r(s,a),$$

the time-average of the reward, which under ergodicity equals the one-step reward averaged over the stationary state-action distribution. The chain induced by $\pi_\theta$ is ergodic with stationary distribution $\mu_{\pi_\theta}$, and the differential action-value is $q_{\pi_\theta}$. We assume finite $\mathcal{S}, \mathcal{A}$, $\pi_\theta(a \mid s)$ differentiable in $\theta$, ergodicity, and bounded reward, so that $\mu_{\pi_\theta}$ and $q_{\pi_\theta}$ are differentiable in $\theta$. The target is

$$\nabla_\theta J_{avR}(\theta) = \nabla_\theta\, r(\pi_\theta) = \sum_s \mu_{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(a \mid s)\, q_{\pi_\theta}(s,a).$$

**Step 1: rearranged differential Bellman identity.** The differential Bellman equation with the rate isolated on the left, valid for every $(s,a)$:

$$r(\pi_\theta) = r(s,a) - q_{\pi_\theta}(s,a) + \sum_{s'} p(s' \mid s,a) \sum_{a'} \pi_\theta(a' \mid s')\, q_{\pi_\theta}(s',a').$$

**Step 2: differentiate.** The reward $r(s,a)$ and transition $p$ carry no $\theta$, so $\nabla_\theta r(s,a) = 0$ and $\nabla_\theta$ passes through $p$. The product rule on $\sum_{a'} \pi_\theta\, q_{\pi_\theta}$ gives

$$\nabla_\theta r(\pi_\theta) = -\nabla_\theta q_{\pi_\theta}(s,a) + \sum_{s'} p(s' \mid s,a) \left[ A(s') + B(s') \right],$$

$$A(s') = \sum_{a'} \nabla_\theta \pi_\theta(a' \mid s')\, q_{\pi_\theta}(s',a'), \qquad B(s') = \sum_{a'} \pi_\theta(a' \mid s')\, \nabla_\theta q_{\pi_\theta}(s',a').$$

$A$ is the reinforce term; $B$ is the ripple term.

**Step 3: average against $\mu_{\pi_\theta}(s)\pi_\theta(a \mid s)$.** The left side is constant in $(s,a)$ and the weights sum to one, so it stays $\nabla_\theta r(\pi_\theta)$. The right side splits into three:

$$\nabla_\theta r(\pi_\theta) = T_1 + T_2 + T_3,$$

$$T_1 = -\sum_s \mu_{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s)\, \nabla_\theta q_{\pi_\theta}(s,a),$$

$$T_2 = \sum_s \mu_{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s) \sum_{s'} p(s' \mid s,a)\, A(s'),$$

$$T_3 = \sum_s \mu_{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s) \sum_{s'} p(s' \mid s, a)\, B(s').$$

**Step 4: the cancellation.** Sum the $(s, a)$ part of $T_3$ first, using the stationarity identity

$$\sum_s \mu_{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s)\, p(s' \mid s, a) = \mu_{\pi_\theta}(s').$$

Then

$$T_3 = \sum_{s'} \mu_{\pi_\theta}(s')\, B(s') = \sum_{s'} \mu_{\pi_\theta}(s') \sum_{a'} \pi_\theta(a' \mid s')\, \nabla_\theta q_{\pi_\theta}(s', a') = -\, T_1,$$

after renaming $s' \to s, a' \to a$. So $T_1 + T_3 = 0$: the $\nabla_\theta q_{\pi_\theta}$ terms cancel, because $\mu_{\pi_\theta}$ is invariant.

**Step 5: what survives.** Only $T_2$ remains. Collapse its $(s, a)$ sums by the same identity:

$$\nabla_\theta r(\pi_\theta) = \sum_{s'} \mu_{\pi_\theta}(s')\, A(s') = \sum_{s'} \mu_{\pi_\theta}(s') \sum_{a'} \nabla_\theta \pi_\theta(a' \mid s')\, q_{\pi_\theta}(s', a').$$

Renaming $s' \to s, a' \to a$ gives the theorem.

$$\nabla_\theta J_{avR}(\theta) = \sum_s \mu_{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(a \mid s)\, q_{\pi_\theta}(s, a) = \mathbb{E}_{S \sim \mu_{\pi_\theta},\, A \sim \pi_\theta(\cdot \mid S)} \big[ \nabla_\theta \log \pi_\theta(A \mid S)\, q_{\pi_\theta}(S, A) \big],$$

the second form by $\nabla_\theta \pi_\theta = \pi_\theta \nabla_\theta \log \pi_\theta$. ■

**Two things to notice.** There is no $\frac{1}{1-\gamma}$ factor: the rate is already per-step, and $\mu_{\pi_\theta}$ has total mass 1 (the discounted visitation $\eta$ had mass $\frac{1}{1-\gamma}$). And $\nabla_\theta \mu_{\pi_\theta}$ never appears: differentiating $r(\pi_\theta) = \sum_s \mu_{\pi_\theta}(s) \sum_a \pi_\theta\, r$ directly would have forced it, but starting from the Bellman identity dodged it. The visitation gradient vanishes silently in the Step 4 cancellation, exactly as $\nabla_\theta \eta$ vanished in the discounted recursive route. The one assumed step is that $\nabla_\theta q_{\pi_\theta}$ is well-defined and finite, which the regularity assumptions provide.