For 1 isolated subsystem i: $Q_i^{\pi_i}(s,a) = R_{is}^a + \sum_{s'} \gamma_i P_{iss'}^a V_i^{\pi_i}(s')$ — For the start-state formulation.

$$V_i^{\pi_i}(s) \stackrel{def}{=} \sum_a \pi_i(s,a)\, Q_i^{\pi_i}(s,a)$$

$$\nabla \stackrel{def}{=} \frac{\partial}{\partial\theta}$$

For N interconnected subsystems, i.e., global value in linear relationship:

$$V_g^{\pi}(s) = \sum_{i=1}^{N} w_i V_i^{\pi_i}(s) = w_1 V_1^{\pi_1}(s) + w_2 V_2^{\pi_2}(s) + \dots + w_N V_N^{\pi_N}(s)$$

$$= w_1 \sum_a \pi_1(s,a)\, Q_1^{\pi_1}(s,a) + w_2 \sum_a \pi_2(s,a)\, Q_2^{\pi_2}(s,a) + \dots + w_N \sum_a \pi_N(s,a)\, Q_N^{\pi_N}(s,a)$$

$$\frac{\partial}{\partial\theta} V_g^{\pi}(s) = \frac{\partial}{\partial\theta}\Big[w_1 \sum_a \pi_1(s,a)\, Q_1^{\pi_1}(s,a)\Big] + \frac{\partial}{\partial\theta}\Big[w_2 \sum_a \pi_2(s,a)\, Q_2^{\pi_2}(s,a)\Big] + \dots + \frac{\partial}{\partial\theta}\Big[w_N \sum_a \pi_N(s,a)\, Q_N^{\pi_N}(s,a)\Big]$$

$$= \Big[\sum_a \pi_1(s,a)\, Q_1^{\pi_1}(s,a)\Big]\nabla w_1 + w_1 \nabla\Big[\sum_a \pi_1(s,a)\, Q_1^{\pi_1}(s,a)\Big] + \Big[\sum_a \pi_2(s,a)\, Q_2^{\pi_2}(s,a)\Big]\nabla w_2 + w_2 \nabla\Big[\sum_a \pi_2(s,a)\, Q_2^{\pi_2}(s,a)\Big] + \dots +$$

$$\Big[\sum_a \pi_N(s,a)\, Q_N^{\pi_N}(s,a)\Big]\nabla w_N + w_N \nabla\Big[\sum_a \pi_N(s,a)\, Q_N^{\pi_N}(s,a)\Big]$$

$$= \Big[\sum_a \pi_1(s,a)\, Q_1^{\pi_1}(s,a)\Big]\nabla w_1 + \Big[\sum_a \pi_2(s,a)\, Q_2^{\pi_2}(s,a)\Big]\nabla w_2 + \dots + \Big[\sum_a \pi_N(s,a)\, Q_N^{\pi_N}(s,a)\Big]\nabla w_N +$$

$$w_1 \nabla\Big[\sum_a \pi_1(s,a)\, Q_1^{\pi_1}(s,a)\Big] + w_2 \nabla\Big[\sum_a \pi_2(s,a)\, Q_2^{\pi_2}(s,a)\Big] + \dots + w_N \nabla\Big[\sum_a \pi_N(s,a)\, Q_N^{\pi_N}(s,a)\Big]$$

$$= \Big[\sum_a \pi_1(s,a)\, Q_1^{\pi_1}(s,a)\Big]\nabla w_1 + \Big[\sum_a \pi_2(s,a)\, Q_2^{\pi_2}(s,a)\Big]\nabla w_2 + \dots + \Big[\sum_a \pi_N(s,a)\, Q_N^{\pi_N}(s,a)\Big]\nabla w_N +$$

$$w_1 \nabla\Big[\sum_a \pi_1(s,a)\Big[R_{1s}^a + \sum_{s'} \gamma_1 P_{1ss'}^a V_1^{\pi_1}(s')\Big]\Big] + w_2 \nabla\Big[\sum_a \pi_2(s,a)\Big[R_{2s}^a + \sum_{s'} \gamma_2 P_{2ss'}^a V_2^{\pi_2}(s')\Big]\Big] +$$

$$\dots + w_N \nabla\Big[\sum_a \pi_N(s,a)\Big[R_{Ns}^a + \sum_{s'} \gamma_N P_{Nss'}^a V_N^{\pi_N}(s')\Big]\Big]$$

$$= \Big[\sum_a \pi_1(s,a)\, Q_1^{\pi_1}(s,a)\Big]\nabla w_1 + \Big[\sum_a \pi_2(s,a)\, Q_2^{\pi_2}(s,a)\Big]\nabla w_2 + \dots + \Big[\sum_a \pi_N(s,a)\, Q_N^{\pi_N}(s,a)\Big]\nabla w_N +$$

$$w_1 \sum_x \sum_{k=0}^{\infty} \gamma_1^k \Pr(s \to x, k, \pi_1) \sum_a Q_1^{\pi_1}(x,a)\, \nabla\pi_1(x,a) + w_2 \sum_x \sum_{k=0}^{\infty} \gamma_2^k \Pr(s \to x, k, \pi_2) \sum_a Q_2^{\pi_2}(x,a)\, \nabla\pi_2(x,a) +$$

$$\dots + w_N \sum_x \sum_{k=0}^{\infty} \gamma_N^k \Pr(s \to x, k, \pi_N) \sum_a Q_2^{\pi_2}(x,a)\, \nabla\pi_2(x,a)$$

Sutton
《Policy Gradient Methods for Reinforcement Learning with Function Approximation》

RLbook P325

$$\boxed{\sum_{k=0}^{\infty} \gamma_i^{k} \Pr(s \to x, k, \pi_i) = d^{\pi_i}(x)}$$

$$= \left[\sum_a \pi_1(s,a)\, Q_1^{\pi_1}(s,a)\right] \nabla W_1 + \left[\sum_a \pi_2(s,a)\, Q_2^{\pi_2}(s,a)\right] \nabla W_2 + \cdots + \left[\sum_a \pi_N(s,a)\, Q_N^{\pi_N}(s,a)\right] \nabla W_N +$$

$$W_1 \sum_x d^{\pi_1}(x) \sum_a Q_1^{\pi_1}(x,a)\, \nabla \pi_1(x,a) + W_2 \sum_x d^{\pi_2}(x) \sum_a Q_2^{\pi_2}(x,a)\, \nabla \pi_2(x,a) +$$

$$\cdots + W_N \sum_x d^{\pi_N}(x) \sum_a Q_N^{\pi_N}(x,a)\, \nabla \pi_N(x,a)$$

$$\nabla V_g^{\pi}(s) = \sum_{i=1}^{N} \left[\sum_a \pi_i(s,a)\, Q_i^{\pi_i}(s,a)\right] \nabla W_i + \sum_{i=1}^{N} W_i \left[\sum_x d^{\pi_i}(x) \sum_a Q_i^{\pi_i}(x,a)\, \nabla \pi_i(x,a)\right]$$

$$= \sum_{i=1}^{N} V_i^{\pi_i}(s) \cdot \nabla W_i + \sum_{i=1}^{N} W_i \left[\sum_x d^{\pi_i}(x) \sum_a Q_i^{\pi_i}(x,a)\, \nabla \pi_i(x,a)\right]$$

$$\eta(s) = \underbrace{h(s)}^{\text{start in } s} + \gamma \overbrace{\sum_{\bar{s}} \eta(\bar{s}) \sum_a \pi(a|\bar{s})\, p(s|\bar{s}, a)}^{\text{from } \bar{s} \text{ to } s}$$

$h(s)$: probability that an episode begins in each state s.

$\eta(s)$: the number of time steps spent, on average, in state s in a single episode.

$$\mu(s) = \frac{\eta(s)}{\sum_{s'} \eta(s')}$$: The on-policy distribution, the fraction of time spent in each state normalized to sum to 1.