We'll consider the terms with Q and b in turn.

$$\begin{aligned}
&\mathbb{E}_{s_{0:\infty},a_{0:\infty}}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) Q_t(s_{0:\infty}, a_{0:\infty})\right] \\
&\quad= \mathbb{E}_{s_{0:t},a_{0:t}}\left[\mathbb{E}_{s_{t+1:\infty},a_{t+1:\infty}}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) Q_t(s_{0:\infty}, a_{0:\infty})\right]\right] \\
&\quad= \mathbb{E}_{s_{0:t},a_{0:t}}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) \mathbb{E}_{s_{t+1:\infty},a_{t+1:\infty}}\left[Q_t(s_{0:\infty}, a_{0:\infty})\right]\right] \\
&\quad= \mathbb{E}_{s_{0:t},a_{0:t-1}}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) A^\pi(s_t, a_t)\right]
\end{aligned}$$

Next,

$$\begin{aligned}
&\mathbb{E}_{s_{0:\infty},a_{0:\infty}}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) b_t(s_{0:t}, a_{0:t-1})\right] \\
&\quad= \mathbb{E}_{s_{0:t},a_{0:t-1}}\left[\mathbb{E}_{s_{t+1:\infty},a_{t:\infty}}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) b_t(s_{0:t}, a_{0:t-1})\right]\right] \\
&\quad= \mathbb{E}_{s_{0:t},a_{0:t-1}}\left[\mathbb{E}_{s_{t+1:\infty},a_{t:\infty}}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right] b_t(s_{0:t}, a_{0:t-1})\right] \\
&\quad= \mathbb{E}_{s_{0:t},a_{0:t-1}}\left[0 \cdot b_t(s_{0:t}, a_{0:t-1})\right] \\
&\quad= 0.
\end{aligned}$$

REFERENCES

Barto, Andrew G, Sutton, Richard S, and Anderson, Charles W. Neuronlike adaptive elements that can solve difficult learning control problems. *Systems, Man and Cybernetics, IEEE Transactions on*, (5):834–846, 1983.

Baxter, Jonathan and Bartlett, Peter L. Reinforcement learning in POMDPs via direct gradient ascent. In *ICML*, pp. 41–48, 2000.

Bertsekas, Dimitri P. *Dynamic programming and optimal control*, volume 2. Athena Scientific, 2012.

Bhatnagar, Shalabh, Precup, Doina, Silver, David, Sutton, Richard S, Maei, Hamid R, and Szepesvári, Csaba. Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems*, pp. 1204–1212, 2009.

Greensmith, Evan, Bartlett, Peter L, and Baxter, Jonathan. Variance reduction techniques for gradient estimates in reinforcement learning. *The Journal of Machine Learning Research*, 5:1471–1530, 2004.

Hafner, Roland and Riedmiller, Martin. Reinforcement learning in feedback control. *Machine learning*, 84 (1-2):137–169, 2011.

Heess, Nicolas, Wayne, Greg, Silver, David, Lillicrap, Timothy, Tassa, Yuval, and Erez, Tom. Learning continuous control policies by stochastic value gradients. *arXiv preprint arXiv:1510.09142*, 2015.

Hull, Clark. Principles of behavior. 1943.

Kakade, Sham. A natural policy gradient. In *NIPS*, volume 14, pp. 1531–1538, 2001a.

Kakade, Sham. Optimizing average reward using discounted rewards. In *Computational Learning Theory*, pp. 605–615. Springer, 2001b.

Kimura, Hajime and Kobayashi, Shigenobu. An analysis of actor/critic algorithms using eligibility traces: Reinforcement learning with imperfect value function. In *ICML*, pp. 278–286, 1998.

Konda, Vijay R and Tsitsiklis, John N. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

Lillicrap, Timothy P, Hunt, Jonathan J, Pritzel, Alexander, Heess, Nicolas, Erez, Tom, Tassa, Yuval, Silver, David, and Wierstra, Daan. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Marbach, Peter and Tsitsiklis, John N. Approximate gradient methods in policy-space optimization of markov reward processes. *Discrete Event Dynamic Systems*, 13(1-2):111–148, 2003.

Minsky, Marvin. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961.

Ng, Andrew Y, Harada, Daishi, and Russell, Stuart. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pp. 278–287, 1999.

Peters, Jan and Schaal, Stefan. Natural actor-critic. *Neurocomputing*, 71(7):1180–1190, 2008.