Taming the Noise in Reinforcement Learning via Soft Updates

Roy Fox*[1] Ari Pakman*[2] Naftali Tishby[1]

[1]Hebrew University [2]Columbia University *Equal contribution

Abstract

Model-free reinforcement learning algorithms, such as Q-learning, perform poorly in the early stages of learning in noisy environments, because much effort is spent unlearning biased estimates of the state-action value function. The bias results from selecting, among several noisy estimates, the apparent optimum, which may actually be suboptimal. We propose G-learning, a new off-policy learning algorithm that regularizes the value estimates by penalizing deterministic policies in the beginning of the learning process. We show that this method reduces the bias of the value-function estimation, leading to faster convergence to the optimal value and the optimal policy. Moreover, G-learning enables the natural incorporation of prior domain knowledge, when available. The stochastic nature of G-learning also makes it avoid some exploration costs, a property usually attributed only to on-policy algorithms. We illustrate these ideas in several examples, where G-learning results in significant improvements of the convergence rate and the cost of the learning process.

Q-Learning

For a given policy $\pi(a|s)$, the state-action discounted cost-to-go function is

$$\begin{aligned} Q^\pi(s,a) &= \sum_{t\geq 0} \gamma^t\, \mathbb{E}[c_t|s_0=s, a_0=a] \\ &= \mathbb{E}[c|s,a] + \gamma\, \mathbb{E}_{p,\pi}[Q^\pi(s',a')|s,a]. \end{aligned}$$

It is optimized by the unique fixed point of the contractive Bellman operator

$$Q^*(s,a) = \mathbb{E}[c|s,a] + \gamma\, \mathbb{E}_p[\min_{a'} Q^*(s',a')|s,a].$$

In model-free learning, given samples (s_i, a_i, c_i, s'_i), Q-learning updates

$$Q_i(s_i,a_i) \leftarrow (1-\alpha_i)Q_{i-1}(s_i,a_i) + \alpha_i \left(c_i + \gamma \min_{a'} Q_{i-1}(s'_i,a')\right).$$

Q_i converges a.s. to Q^*, if each (s_i, a_i) is visited infinitely many times and

$$\sum_i \alpha_i = \infty; \qquad \sum_i \alpha_i^2 < \infty.$$

Winner's Curse

If $Q(s,a)$ is an unbiased noisy estimate of $Q^*(s,a)$, then by Jensen's inequality

$$\mathbb{E}[\min_a Q(s,a)] \leq \min_a Q^*(s,a),$$

with equality only when learning is completed, i.e. with probability 1

$$\arg\min_a Q(s,a) = \arg\min_a Q^*(s,a).$$

Even more bias can result from noise introduced by Q-function approximation.

Information Cost

Consider a simple stochastic prior policy $\rho(a|s)$, e.g. the uniform distribution. The *information cost* is

$$g^\pi(s,a) = \log \tfrac{\pi(a|s)}{\rho(a|s)},$$

and its expectation is the KL divergence of $\pi_s = \pi(\cdot|s)$ from $\rho_s = \rho(\cdot|s)$

$$\mathbb{E}_\pi[g^\pi(s,a)|s] = \mathbb{D}[\pi_s \| \rho_s].$$

The state-action discounted information-to-go function is

$$J^\pi(s,a) = \sum_{t>0} \gamma^t\, \mathbb{E}[g^\pi(s_t,a_t)|s_0=s,a_0=a].$$

g^π penalizes deviation of the learned policy from the prior policy. J^π penalizes deviation of the stochastic process from the prior distribution over trajectories.

Free-Energy Function

Trading off the cost and penalty gives the *state-action free-energy function*

$$\begin{aligned} G^\pi(s,a) &= Q^\pi(s,a) + \tfrac{1}{\beta} J^\pi(s,a) \\ &= \sum_{t\geq 0} \gamma^t\, \mathbb{E}[c_t + \tfrac{\gamma}{\beta} g^\pi(s_{t+1},a_{t+1}))|s_0=s,a_0=a] \\ &= \mathbb{E}_\theta[c|s,a] + \gamma\, \mathbb{E}_p[F^\pi(s')|s,a], \end{aligned}$$

with

$$F^\pi(s) = \sum_a \pi(a|s) \left[\tfrac{1}{\beta} \log \tfrac{\pi(a|s)}{\rho(a|s)} + G^\pi(s,a)\right].$$

Given G, the free energy is minimized by the soft-greedy policy

$$\pi(a|s) = \frac{\rho(a|s)e^{-\beta G(s,a)}}{\sum_{a'} \rho(a'|s)e^{-\beta G(s,a')}}.$$

- When the inverse-temperature β is small, the information cost is dominant, and π approaches the prior ρ.
- When β is large, we are willing to diverge much from the prior to reduce the external cost, and π approaches the deterministic greedy policy for G.

G is optimized by the unique fixed point of the contractive operator

$$G^*(s,a) = \mathbb{E}[c|s,a] - \tfrac{\gamma}{\beta}\, \mathbb{E}_p\left[\log \sum_{a'} \rho(a'|s')e^{-\beta G^*(s',a')}\right]$$

G-learning

We introduce G-learning, a model-free, off-policy TD-learning algorithm that learns G^* from samples by applying soft updates

$$G_i(s_i,a_i) \leftarrow (1-\alpha_i)G_{i-1}(s_i,a_i) + \alpha_i \left(c_i - \tfrac{\gamma}{\beta} \log\left(\sum_{a'} \rho(a'|s'_i)e^{-\beta G_{i-1}(s'_i,a')}\right)\right).$$

G_i converges a.s. to G^* under conditions similar to Q-learning.

Scheduling β

If $G_{i-1}(s,a)$ is an unbiased noisy estimate of $G^*(s,a)$, then for $\beta = 0$

$$c_i + \gamma \sum_{a'} \rho(a'|s'_i) G_{i-1}(s'_i,a'),$$

is a pessimistic (positively biased) estimate of $G^*(s,a)$, while for $\beta = \infty$

$$c_i + \gamma \min_{a'} G_{i-1}(s'_i,a'),$$

is an optimistic (negatively biased) estimate of $G^*(s,a)$. Soft-min is continuous and monotonic in β, so some β schedule gives unbiased updates. β should increase from 0 to ∞ as uncertainty about the value function decreases. We use the linear schedue

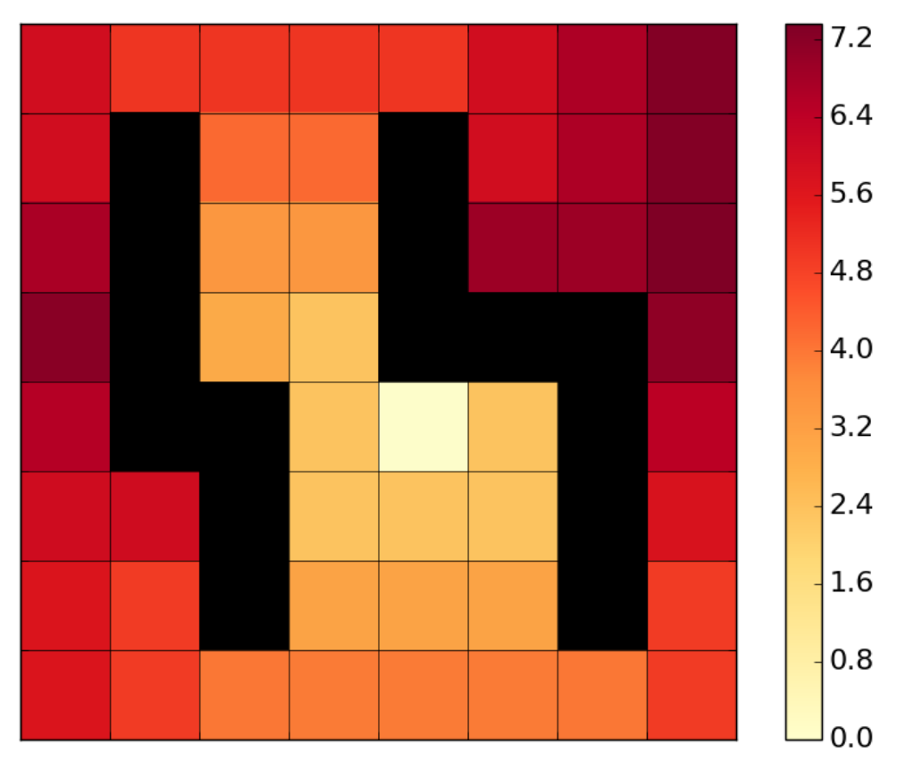
$$\beta_t = kt.$$

Examples

Empirical bias: $\frac{1}{Nn}\sum_{i=1}^{N}\sum_{s=1}^{n}(V_{i,t}(s) - V_i^*(s))$

Mean abs error: $\frac{1}{Nn}\sum_{i=1}^{N}\sum_{s=1}^{n}|V_{i,t}(s) - V_i^*(s)|$

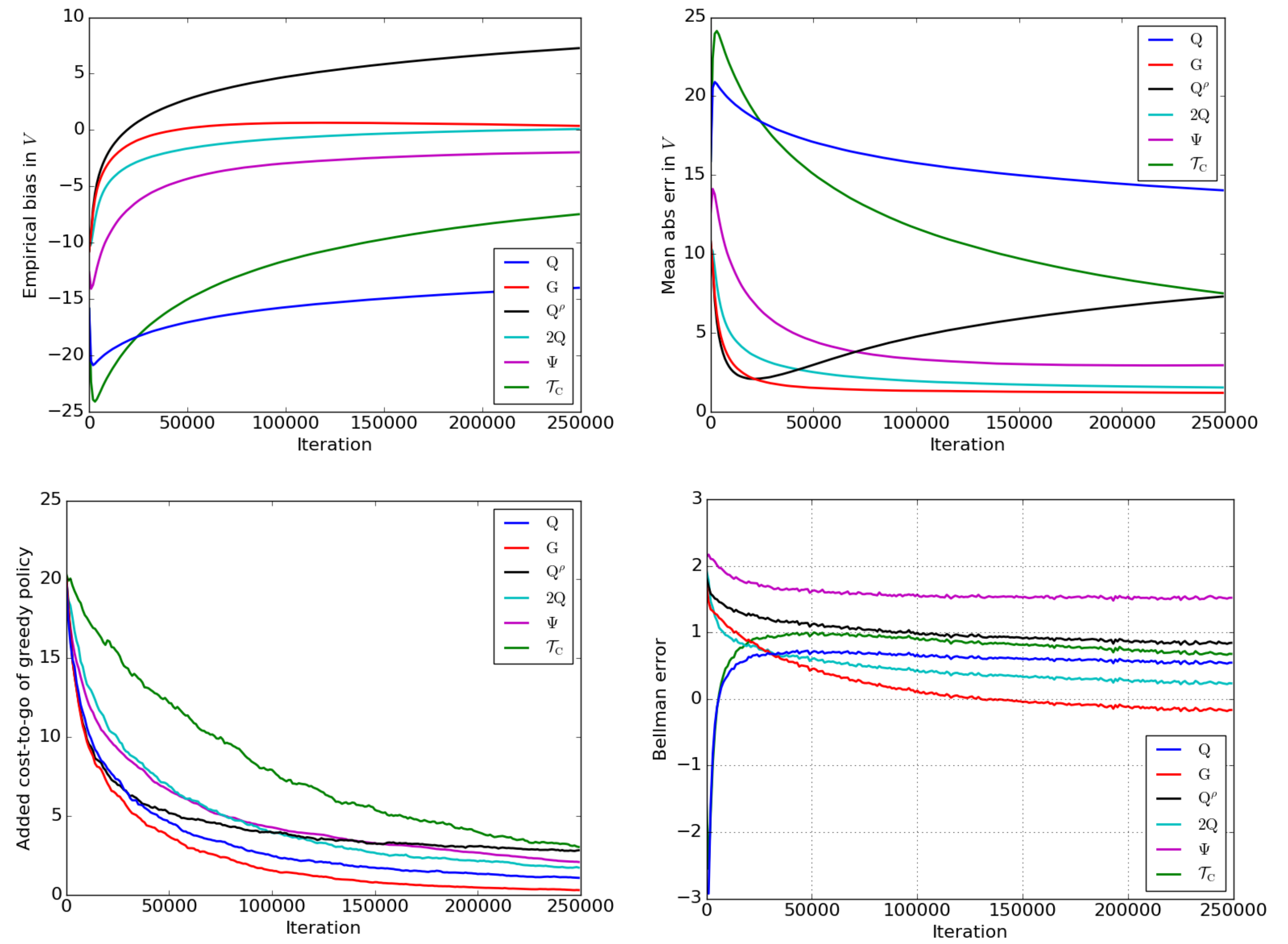
Policy value: $\frac{1}{Nn}\sum_{i=1}^{N}\sum_{s=1}^{n}(V^{\pi_{i,t}}(s) - V_i^*(s))$



$V(s)$ in gridworld example

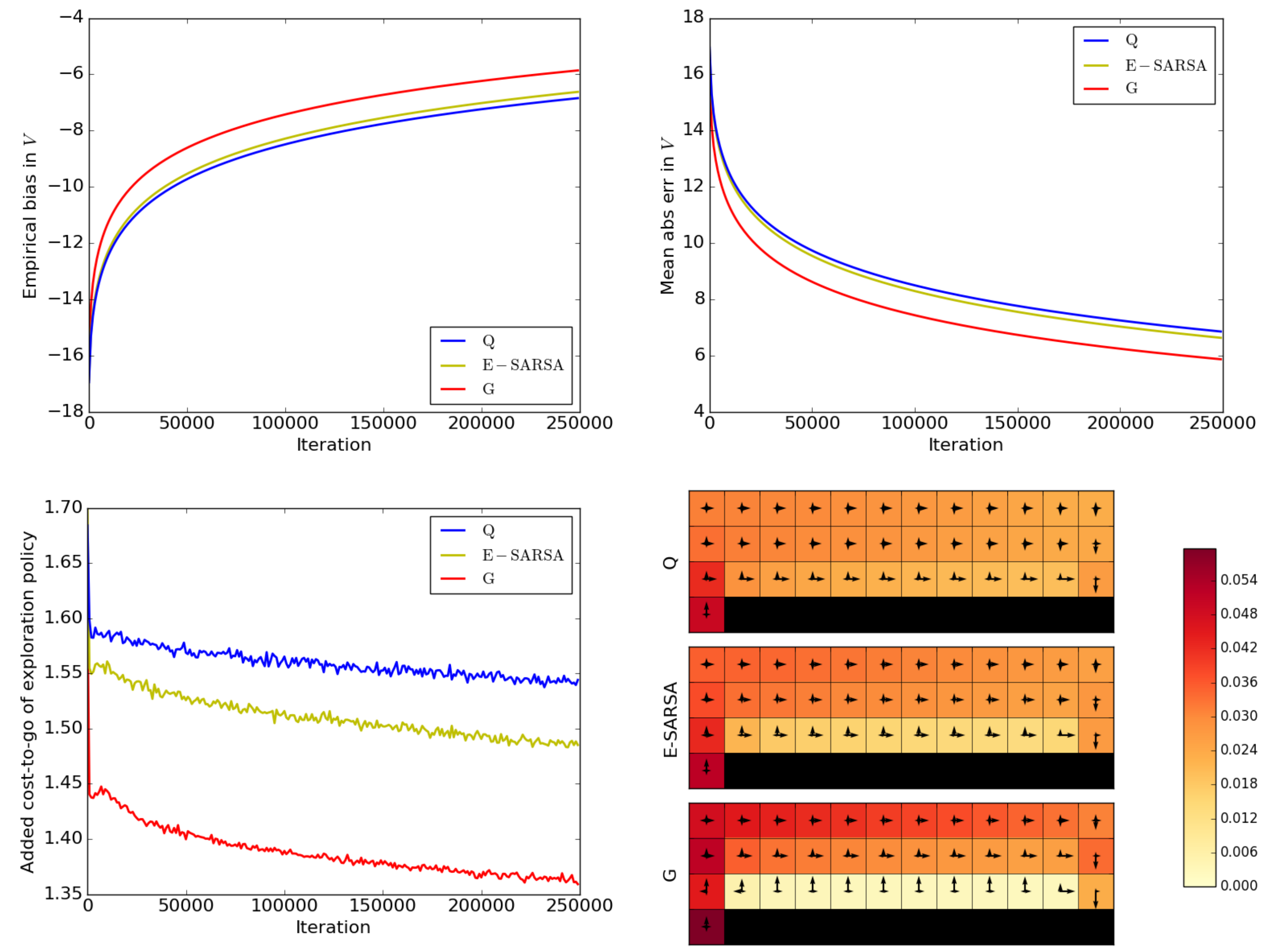
Gridworld Example

G-learning outperforms Q-, Double-Q-[1], Ψ-[2]and $\mathcal{T}_C$-learning[3], in 100 gridworld domains generated by $\mathbb{E}[c|s,a] \sim \mathcal{U}(1,3)$ i.i.d, where each step incurs cost $\mathbb{E}[c|s,a] + \mathcal{N}(0,4^2)$. G-learning enjoys early benefits of Q^ρ-learning ($\beta = 0$), and later benefits of Q-learning.



Cliff Example

G-learning outperforms Q-learning and Expected-SARSA, both in convergence rate and in exploration cost. Expected-SARSA avoids the cliff because ϵ-greedy cliff-walking sometimes falls. G-learning is off-policy, and it learns the value of its soft-greedy update policy, which requires information cost to prevent falling, thus better avoiding the cliff.



[1] van Hasselt, 2010 [2] Rawlik et al., 2010, Azar et al., 2012 [3] Bellemare et al., 2016