Appendix A. Proofs on the performance bound

For the following proof, we define the greedy policy and the Bellman operator regularized by Shannon entropy as well as KL divergence as $\mathcal{G}_{\mu}^{\lambda,\tau}(q) = \mathrm{argmax}_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}}\left(\langle\pi, q\rangle - \lambda\,\mathrm{KL}(\pi||\mu) + \tau\mathcal{H}(\pi)\right)$ and $T_{\pi|\mu}^{\lambda,\tau}q = r + \gamma P\left(\langle\pi, q\rangle - \lambda\,\mathrm{KL}(\pi||\mu)\right) + \tau\mathcal{H}(\pi)$, respectively. We also note the following fact about the greedy policy (Vieillard et al., 2020a):

$$\mathcal{G}_{\mu}^{\lambda,\tau}(q) = \operatorname*{argmax}_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}}\left(\langle\pi, q\rangle - \lambda\,\mathrm{KL}(\pi||\mu) + \tau\mathcal{H}(\pi)\right) \propto \mu^{\frac{\lambda}{\lambda+\tau}} \exp\frac{1}{\lambda+\tau}q, \tag{14}$$

and we have the following maximum:

$$\max_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}}\left(\langle\pi, q\rangle - \lambda\,\mathrm{KL}(\pi||\mu) + \tau\mathcal{H}(\pi)\right) = (\lambda+\tau)\ln\langle\mathbf{1}, \mu^{\frac{\lambda}{\lambda+\tau}} \exp\frac{q}{\lambda+\tau}\rangle. \tag{15}$$

Before going to the proof of Theorem 2, we provide the following proposition.

Proposition 4 *Define $Z_k = \sum_{j=0}^{k}\eta_j$, $h_0 = q_0$, and h_k for $k \geq 1$ as the average of past smoothed q-functions: $h_k = \frac{1}{Z_k}\sum_{j=0}^{k}\eta_j q_j = \frac{Z_{k-1}}{Z_k}h_{k-1} + \frac{\eta_k}{Z_k}q_k$. If $\lambda_k > 0$ for all k, GVI is equivalent to the following iteration:*

$$\begin{cases} \pi_{k+1} = \mathcal{G}^{0,\frac{1}{Z_k}}(h_k) \\ q_{k+1} = (T_{\pi_{k+1}|\pi_k}^{\frac{1}{\eta_k},0})^m q_k + \epsilon_{k+1} \\ h_{k+1} = \frac{1}{Z_{k+1}}\sum_{j=0}^{k+1}\eta_j q_j = \frac{Z_k}{Z_{k+1}}h_k + \frac{\eta_{k+1}}{Z_{k+1}}q_{k+1} \end{cases}. \tag{16}$$

Proof Using Eq. (14) and by direct induction, we have $\pi_{k+1} \propto \pi_k \exp\eta_k q_k \propto \cdots \propto \exp\sum_{j=0}^{k}\eta_j q_j = \exp Z_k h_k$. Eq. (14) also provides $\mathrm{argmax}_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}}\left(\langle\pi, q\rangle + \tau\mathcal{H}(\pi)\right) \propto \exp(\frac{1}{\tau}q)$. Hence, π_{k+1} satisfies $\pi_{k+1} = \mathrm{argmax}_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}}\left(\langle\pi, h_k\rangle + \frac{1}{Z_k}\mathcal{H}(\pi)\right) = \mathcal{G}^{0,\frac{1}{Z_k}}(h_k)$. ■

We now prove the error-bound of GVI using Eq. (16).

Proof. We first transform $q_* - q_{\pi_{k+1}}$, the difference between the optimal value function and the value function computed by Eq. (16), using the following useful lemma:

Lemma 5 (Kakade and Langford (2002)) *For any $q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, we have $q_\pi - q = (I - \gamma P_\pi)^{-1}(T_\pi q - q)$.*

Using Lemma 5, $q_* - q_{\pi_{k+1}}$ can be transformed as

$$\begin{aligned} q_* - q_{\pi_{k+1}} &= q_* - h_k + h_k - q_{\pi_{k+1}} \\ &= (I - \gamma P_{\pi_*})^{-1}(T_{\pi_*}h_k - h_k) - (I - \gamma P_{\pi_{k+1}})^{-1}(T_{\pi_{k+1}}h_k - h_k). \end{aligned} \tag{17}$$

Since the KL regularization vanishes after the iteration converges, the optimal policy must be deterministic, and hence $\mathcal{H}(\pi_*) = 0$. Since π_{k+1} is the regularized greedy policy, we have

$$\begin{aligned} \pi_{k+1} = \mathcal{G}^{0,\frac{1}{Z_k}}(h_k) &\Rightarrow \langle\pi_{k+1}, h_k\rangle + \frac{1}{Z_k}\mathcal{H}(\pi_{k+1}) \geq \langle\pi_*, h_k\rangle + \frac{1}{Z_k}\mathcal{H}(\pi_*) \\ &\Rightarrow T_{\pi_{k+1}}^{0,\frac{1}{Z_k}}h_k = T_{\pi_{k+1}}h_k + \gamma\frac{1}{Z_k}P\mathcal{H}(\pi_{k+1}) \geq T_{\pi_*}h_k. \end{aligned} \tag{18}$$

Using this with Eq. (5) and the fact that for any π the matrix $(I-\gamma P_\pi)^{-1} = \sum_{t\geq 0}\gamma^t P_\pi^t$ is positive, we have the following inequality:

$$q_* - q_{\pi_{k+1}} \leq (I-\gamma P_{\pi_*})^{-1}(T_{\pi_{k+1}}^{0,\frac{1}{Z_k}} h_k - h_k) - (I-\gamma P_{\pi_{k+1}})^{-1}(T_{\pi_{k+1}}^{0,\frac{1}{Z_k}} h_k - h_k - \gamma \frac{1}{Z_k} P\mathcal{H}(\pi_{k+1})). \tag{19}$$

As for the residual $T_{\pi_{k+1}}^{0,\frac{1}{Z_k}} h_k - h_k$, we have the following useful lemma:

Lemma 6 *For any $k \geq 1$, we have $\eta_k T_{\pi_{k+1}|\pi_k}^{\frac{1}{\eta_k},0} q_k = Z_k T_{\pi_{k+1}}^{0,\frac{1}{Z_k}} h_k - Z_{k-1} T_{\pi_k}^{0,\frac{1}{Z_{k-1}}} h_{k-1}$. For $k=0$, we have $\eta_0 T_{\pi_1|\pi_0}^{\frac{1}{\eta_0},0} q_0 = Z_0 T_{\pi_1}^{0,\frac{1}{\eta_0}} h_0 - \gamma P\mathcal{H}(\pi_0)$.*

Proof Using the definition of π_k and h_k, the following equation holds.

$$\eta_k q_k + \ln \pi_k = \eta_k q_k + (Z_{k-1}h_{k-1} - \ln\langle 1, \exp Z_{k-1}h_{k-1}\rangle) = Z_k h_k - \ln\langle 1, \exp Z_{k-1}h_{k-1}\rangle. \tag{20}$$

Therefore, we have $\langle \pi, \eta_k q_k\rangle - \mathrm{KL}(\pi||\pi_k) = \langle \pi, Z_k h_k\rangle - \langle \pi, \ln\pi\rangle - \ln\langle \mathbf{1}, \exp Z_{k-1}h_{k-1}\rangle$. From Eq. (15), the maximum of $\langle \pi, Z_k h_k\rangle - \langle \pi, \ln \pi\rangle$ is $\ln\langle \mathbf{1}, \exp Z_k h_k\rangle$, and the maximizer is π_{k+1} from the definition. By substituting π_{k+1} to π, the following equation holds:

$$\langle \pi_{k+1}, \eta_k q_k\rangle - \mathrm{KL}(\pi_{k+1}||\pi_k) = Z_k \frac{1}{Z_k}\ln\langle \mathbf{1}, \exp Z_k h_k\rangle - Z_{k-1}\frac{1}{Z_{k-1}}\ln\langle \mathbf{1}, \exp Z_{k-1}h_{k-1}\rangle. \tag{21}$$

From Eq. (15), $\frac{1}{Z_k}\ln\langle 1, \exp Z_k h_k\rangle$ is the maximum of $\langle \pi, h_k\rangle + \frac{1}{Z_k}\mathcal{H}(\pi)$, and the associated maximizer is again π_{k+1}. Hence, the following equation holds:

$$\langle \pi_{k+1}, \eta_k q_k\rangle - \mathrm{KL}(\pi_{k+1}||\pi_k) = Z_k\left(\langle \pi_{k+1}, h_k\rangle + \frac{1}{Z_k}\mathcal{H}(\pi_{k+1})\right) - Z_{k-1}\left(\langle \pi_k, h_{k-1}\rangle + \frac{1}{Z_{k-1}}\mathcal{H}(\pi_k)\right). \tag{22}$$

Observing that $\eta_k r = Z_k r - Z_{k-1} r$, we have the first part of the result: $\eta_k T_{\pi_{k+1}|\pi_k}^{\frac{1}{\eta_k},0} q_k = Z_k T_{\pi_{k+1}}^{0,\frac{1}{Z_k}} h_k - Z_{k-1} T_{\pi_k}^{0,\frac{1}{Z_{k-1}}} h_{k-1}$. For $k=0$, using the fact that $h_0 = q_0$,

$$\eta_0 T_{\pi_1|\pi_0}^{\frac{1}{\eta_0},0} q_0 = \eta_0 r + \gamma P(\langle \pi_1, \eta_0 h_0\rangle + \eta_0 \frac{1}{\eta_0}\mathcal{H}(\pi_1) + \eta_0\frac{1}{\eta_0}\langle \pi_1, \ln \pi_0\rangle) = \eta_0 T_{\pi_1}^{0,\frac{1}{\eta_0}} h_0 - \gamma P\mathcal{H}(\pi_0), \tag{23}$$

where we use in the last line the fact that π_0, being uniform, $\langle \pi_1, \ln\pi_0\rangle = -\ln|\mathcal{A}| = -\mathcal{H}(\pi_0)$. This concludes the proof. ■

Using Lemma 6, we can provide induction on h_k.

Lemma 7 *Define $E_k = -\sum_{j=1}^k \eta_j \epsilon_j$ and $X_k = \sum_{j=0}^k (\eta_{j+1}-\eta_j) T_{\pi_{j+1}|\pi_j}^{\frac{1}{\eta_j},0} q_j$. For any $k \geq 1$, we have $h_{k+1} = \frac{Z_k}{Z_{k+1}} T_{\pi_{k+1}}^{0,\frac{1}{Z_k}} h_k + \frac{1}{Z_{k+1}}\left(\eta_0 q_0 - E_{k+1} + X_k - \gamma P\mathcal{H}(\pi_0)\right)$.*

Proof Using the definition of h_k, Lemma 6, and the fact that $q_{k+1} = T^{\frac{1}{\eta_k},0}_{\pi_{k+1}|\pi_k} q_k + \epsilon_{k+1}$, we have

$$
\begin{aligned}
&Z_{k+1}h_{k+1} = \sum_{j=0}^{k+1} \eta_j q_j = \eta_0 q_0 + \eta_1 q_1 + \sum_{j=1}^{k} \eta_{j+1} q_{j+1} \\
&= \eta_0 q_0 + \left((\eta_1 - \eta_0) + \eta_0\right) T^{\frac{1}{\eta_0},0}_{\pi_1|\pi_0} q_0 + \eta_1 \epsilon_1 + \sum_{j=1}^{k} \left(\left((\eta_{j+1} - \eta_j) + \eta_j\right) T^{\frac{1}{\eta_j},0}_{\pi_{j+1}} q_j + \eta_{j+1}\epsilon_{j+1} \right) \\
&= \eta_0 q_0 + \left(Z_0 T^{0,\frac{1}{\eta_0}}_{\pi_1} h_0 - \gamma P \mathcal{H}(\pi_0) \right) + \sum_{j=1}^{k} \left(Z_j T^{0,\frac{1}{Z_j}}_{\pi_{j+1}} h_j - Z_{j-1} T^{0,\frac{1}{Z_{j-1}}}_{\pi_j} h_{j-1} \right) + X_k - E_{k+1} \\
&= \eta_0 q_0 + X_k - E_{k+1} - \gamma P \mathcal{H}(\pi_0) + Z_k T^{0,\frac{1}{Z_k}}_{\pi_{k+1}} h_k \qquad (24) \\
&\Leftrightarrow h_{k+1} = \frac{Z_k}{Z_{k+1}} T^{0,\frac{1}{Z_k}}_{\pi_{k+1}} h_k + \frac{1}{Z_{k+1}} \left(\eta_0 q_0 - E_{k+1} + X_k - \gamma P \mathcal{H}(\pi_0) \right). \qquad (25)
\end{aligned}
$$

■

Using Lemma 7 and the fact that $Z_{k+1}h_{k+1} = Z_k h_k + \eta_{k+1} q_{k+1}$, we have $T^{0,\frac{1}{Z_k}}_{\pi_{k+1}} h_k - h_k = \frac{1}{Z_k} \left(\eta_{k+1} q_{k+1} - \eta_0 q_0 + E_{k+1} - X_k + \gamma P \mathcal{H}(\pi_0) \right)$. Injecting this last result into decomposition (19), we get

$$
\begin{aligned}
q_* - q_{\pi_{k+1}} &\le (I - \gamma P_{\pi_*})^{-1} (T^{0,\frac{1}{Z_k}}_{\pi_{k+1}} h_k - h_k) - (I - \gamma P_{\pi_{k+1}})^{-1} (T^{0,\frac{1}{Z_k}}_{\pi_{k+1}} h_k - h_k - \gamma P \mathcal{H}(\pi_{k+1})) \\
&\le (I - \gamma P_{\pi_*})^{-1} \left(\frac{1}{Z_k} \left(Y_k + \gamma P \mathcal{H}(\pi_0) \right) \right) - (I - \gamma P_{\pi_{k+1}})^{-1} \left(\frac{1}{Z_k} \left(Y_k - \gamma P \mathcal{H}(\pi_{k+1}) \right) \right), \quad (26)
\end{aligned}
$$

where we write $Y_k = \eta_{k+1} q_{k+1} - \eta_0 q_0 + E_{k+1} - X_k$ for the uncluttered notation and the last inequality holds, since $-(I - \gamma P_{\pi_{k+1}})^{-1} P \mathcal{H}(\pi_0) \le 0$. Next, using the fact that $q_* - q_{\pi_{k+1}} \ge 0$ and rearranging terms, we have

$$
\begin{aligned}
q_* - q_{\pi_{k+1}} \le &\left| \left((I - \gamma P_{\pi_*})^{-1} - (I - \gamma P_{\pi_{k+1}})^{-1} \right) \frac{E_{k+1}}{Z_k} \right| \\
&+ (I - \gamma P_{\pi_*})^{-1} \left| \frac{1}{Z_k} \left(\eta_{k+1} q_{k+1} - \eta_0 q_0 - X_k + \gamma P \mathcal{H}(\pi_0) \right) \right| \\
&+ (I - \gamma P_{\pi_{k+1}})^{-1} \left| \frac{1}{Z_k} \left(\eta_{k+1} q_{k+1} - \eta_0 q_0 - X_k + \gamma P \mathcal{H}(\pi_{k+1}) \right) \right|. \qquad (27)
\end{aligned}
$$

From the assumptions $\|q_k\|_\infty \le q_{\max}$ for all k, we have $\|X_k\|_\infty = \|\sum_{j=0}^{k} (\eta_{j+1} - \eta_j) T^{\frac{1}{\eta_j},0}_{\pi_{j+1}|\pi_j} q_j\|_\infty \le q_{\max} \sum_{j=0}^{k} |\eta_{j+1} - \eta_j|$. Combined with Eq. (27), we have

$$
\|q_* - q_{\pi_{k+1}}\|_\infty \le \frac{2}{(1-\gamma) Z_k} \left(\left\| \sum_{j=1}^{k} \eta_j \epsilon_j \right\|_\infty + (\eta_{k+1} + \eta_0 + \sum_{j=0}^{k} |\eta_{j+1} - \eta_j|) q_{\max} + \gamma \ln |A| \right). \qquad (28)
$$

Table 1: Hyperparameters of algorithms in deep RL experiments

Parameter	Value
Shared	
optimizer	Adam
learning rate	10^{-4}
discount factor (γ)	0.99
replay buffer size	10^6
number of hidden layers	2
number of hidden units per layer	256
number of samples per minibatch	32
activations	ReLU

Appendix B. Proof of Theorem 3

Define for any $k \geq 0$ the term ${q'}_k = \lambda_{k+1} (q_k - \ln \pi_k)$. By basic calculus, the evaluation step of Eq. 8 can be transformed as

$$\begin{aligned} & q_{k+1} = \frac{r}{\lambda_{k+1}} + \ln \pi_{k+1} + \frac{\lambda_k}{\lambda_{k+1}} \gamma P \left\langle \pi_{k+1}, q_k - \ln \pi_{k+1} \right\rangle \\ \Leftrightarrow\ & \lambda_{k+1} (q_{k+1} - \ln \pi_{k+1}) = r + \gamma P \left\langle \pi_{k+1}, \lambda_k (q_k - \ln \pi_k) \right\rangle - \lambda_k \left\langle \pi_{k+1}, \ln \pi_{k+1} - \ln \pi_k \right\rangle \\ \Leftrightarrow\ & {q'}_{k+1} = r + \gamma P \left\langle \pi_{k+1}, {q'}_k \right\rangle - \lambda_k \, \mathrm{KL}(\pi_{k+1} \| \pi_k). \end{aligned} \tag{29}$$

For the greedy step, we have

$$\begin{aligned} \underset{\pi}{\mathrm{argmax}} \left\langle \pi, q_k \right\rangle + \mathcal{H}(\pi) \propto \exp(q_k) = \pi_k \exp\left(\frac{{q'}_k}{\lambda_k} \right) \\ \propto \underset{\pi}{\mathrm{argmax}} \left\langle \pi, {q'}_k \right\rangle + \mathrm{KL}(\pi \| \pi_k). \end{aligned} \tag{30}$$

Therefore, we have shown that

$$\begin{aligned} & \begin{cases} \pi_{k+1} = \mathrm{argmax}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \left\langle \pi, q_k \right\rangle + \mathcal{H}(\pi) \\ q_{k+1} = \ln \pi_{k+1} + \frac{r}{\lambda_{k+1}} + \frac{\lambda_k}{\lambda_{k+1}} \gamma P \left\langle \pi_{k+1}, q_k - \ln \pi_{k+1} \right\rangle \end{cases} \\ \Leftrightarrow & \begin{cases} \pi_{k+1} = \mathrm{argmax}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \left\langle \pi, {q'}_k \right\rangle - \lambda_k \, \mathrm{KL}(\pi \| \pi_k) \\ {q'}_{k+1} = r + \gamma P \left\langle \pi_{k+1}, {q'}_k - \lambda_k \, \mathrm{KL}(\pi_{k+1} \| \pi_k) \right\rangle \end{cases} . \end{aligned} \tag{31}$$

Appendix C. Hyperparameters

Table. 1 lists the hyperparameters used in the comparative evaluation in Section. 5.

Appendix D. Maze Environment Details

For the tabular experiments, we use randomly generated 5×5 mazes. Figure D shows a sample maze used in the experiment. The agent starts from a fixed position marked with S and can move to any of its neighboring states with success probability 0.9, or to a different random direction with probability 0.1. The agent receives +1 reward when it reaches the goal marked with G, and the environment terminates after 25 steps. The agent cannot enter the black tiles.

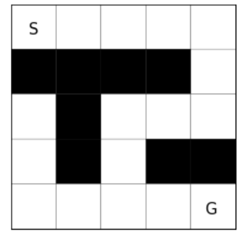

Figure 6: Example of a generated maze.