A. Proofs of Statements and Theorems

A.1. Equivalent Expression of the Density Function

In Section 3, we point out that the state density function ρ^π has two equivalent expressions, which we prove as follows:

$$
\begin{aligned}
\rho^\pi(s) &= \sum_{t=0}^{\infty} \gamma^t P(s^t = s|\pi, s^0 \sim \phi) \\
&= P(s^0 = s|\pi, s^0 \sim \phi) + \\
&\qquad \sum_{t=1}^{\infty} \gamma^t P(s^t = s|\pi, s^0 \sim \phi) \\
&= \phi(s) + \sum_{t=0}^{\infty} \gamma^{t+1} P(s^{t+1} = s|\pi, s^0 \sim \phi) \\
&= \phi(s) + \gamma \sum_{t=0}^{\infty} \gamma^{t} P(s^{t+1} = s|\pi, s^0 \sim \phi) \\
&= \phi(s) + \gamma \int_S \int_A \pi(a|s') P_a(s', s) \\
&\qquad \sum_{t=0}^{\infty} \gamma^t P(s^t = s'|\pi, s^0 \sim \phi) da ds' \\
&= \phi(s) + \gamma \int_S \int_A \pi(a|s') P_a(s', s) \rho^\pi(s') da ds'
\end{aligned}
$$

A.2. Proof of Lemma 1

The Lagrangian of (3) is

$$
\begin{aligned}
\mathcal{L} = &\int_S \int_A r(s,a) \bar{\rho}_s^\pi(s,a) da ds - \\
&\int_S \int_A \mu(s,a) \Big(\bar{\rho}_s^\pi(s,a) - \pi(a|s)(\phi(s) + \\
&\gamma \int_S \int_A P_{a'}(s', s) \bar{\rho}_s^\pi(s', a') da' ds') \Big) da ds
\end{aligned} \tag{8}
$$

where $\mu : S \times A \to \mathbb{R}$ is the Lagrange multiplier. The key step is by noting that

$$
\begin{aligned}
&\int_S \int_A \int_S \int_A \mu(s,a) \pi(a|s) P_{a'}(s', s) \bar{\rho}_s^\pi(s', a') da' ds' da ds \equiv \\
&\int_S \int_A \int_S \int_A \mu(s',a') \pi(a'|s') P_a(s, s') \bar{\rho}_s^\pi(s, a) da' ds' da ds
\end{aligned}
$$

Then by rearranging terms, the Lagrangian becomes

$$
\begin{aligned}
\mathcal{L} = &\int_S \phi(s) \int_A \mu(s,a) \pi(a|s) da ds - \\
&\int_S \int_A \bar{\rho}_s^\pi(s,a) \Big(\mu(s,a) - r(s,a) - \\
&\gamma \int_S P_a(s,s') \int_A \pi(a'|s') \mu(s',a') da' ds' \Big) da ds
\end{aligned} \tag{9}
$$

By the KKT condition and taking $Q = \mu$, the optimality condition satisfies (1) exactly.

A.3. Proof of Theorem 1

The solution π to the primal problem is the optimal policy for the modified MDP with reward $r + \sigma_- - \sigma_+$, which means π is the optimal solution to:

$$
\begin{aligned}
&\max_{Q,\pi} \int_S \phi(s) \int_A Q^\pi(s,a) \pi(a|s) da ds \\
&\text{s.t. } Q^\pi(s,a) = r(s,a) + \sigma_-(s) - \sigma_+(s) + \\
&\qquad \gamma \int_S P_a(s,s') \int_A \pi(a'|s') Q^\pi(s',a') da' ds'
\end{aligned}
$$

π is also the optimal solution to:

$$
\begin{aligned}
&\max_{\bar{\rho},\pi} \int_S \int_A \bar{\rho}^\pi(s,a)(r(s,a) + \sigma_-(s) - \sigma_+(s)) da ds \\
&\text{s.t. } \bar{\rho}^\pi(s,a) = \pi(a|s) \Big(\phi(s) + \\
&\qquad \gamma \int_S \int_A P_{a'}(s', s) \bar{\rho}^\pi(s', a') da' ds' \Big)
\end{aligned}
$$

Therefore, for any feasible policy π' the following inequality holds:

$$
\begin{aligned}
&\int_S \int_A \bar{\rho}^\pi(s,a)(r(s,a) + \sigma_-(s) - \sigma_+(s)) da ds \geq \\
&\int_S \int_A \bar{\rho}^{\pi'}(s,a)(r(s,a) + \sigma_-(s) - \sigma_+(s)) da ds
\end{aligned} \tag{10}
$$

By complementary slackness, if $\sigma_-(s) > 0$, then $\rho^\pi(s) = \rho_{min}(s)$. The same applies to σ_+ and ρ_{max}. Since $\rho^{\pi'}(s) \geq \rho_{min}(s)$ and $\rho^{\pi'}(s) \leq \rho_{max}(s)$, we have:

$$
\begin{aligned}
&\int_S \int_A \bar{\rho}^\pi(s,a)(\sigma_-(s) - \sigma_+(s)) da ds \leq \\
&\qquad \int_S \int_A \bar{\rho}^{\pi'}(s,a)(\sigma_-(s) - \sigma_+(s)) da ds
\end{aligned} \tag{11}
$$

Then we use (11) to eliminate the $\sigma_-(s) - \sigma_+(s)$ in (10) and derive:

$$
\int_S \int_A \bar{\rho}^\pi(s,a) r(s,a) da ds \geq \int_S \int_A \bar{\rho}^{\pi'}(s,a) r(s,a) da ds
$$

which means π is the optimal solution maximizing $J_d^\star$ among all the solutions satisfying the density constraints. As a result, π is the optimal solution to the DCRL problem.

A.4. Proof of Lemma 2

The proof of Lemma 2 follows from the proof of Theorem 2.2.7 in Rockafellar (1970). For simplicity, let $\hat{g} =$

$\hat{g}(\sigma_+, \alpha)$. For all $\sigma'_+ \in \mathcal{P}$, we have

$$
\begin{aligned}
& d(\sigma'_+) - \frac{\mu_d}{2}||\sigma'_+ - \sigma^k_+||^2 \\
& \leq d(\sigma^k_+) + \langle \nabla d(\sigma^k_+), \sigma'_+ - \sigma^k_+ \rangle \\
& = d(\sigma^k_+) + \langle \nabla d(\sigma^k_+), y^+ - \sigma^k_+ \rangle + \\
& \qquad \langle \nabla d(\sigma_+), \sigma'_+ - \sigma^{k+1}_+ \rangle \\
& \leq d(\sigma^k_+) + \langle \nabla d(\sigma^k_+), \sigma^{k+1}_+ - \sigma^k_+ \rangle + \\
& \qquad \langle \hat{g}, \sigma'_+ - \sigma^{k+1}_+ \rangle + \sqrt{\frac{2\epsilon}{\mu}}||\sigma'_+ - \sigma^{k+1}_+|| \\
& \leq d(\sigma^{k+1}_+) - \frac{\alpha}{2}||\hat{g}||^2 + \\
& \qquad \langle \hat{g}, \sigma'_+ - \sigma^k_+ \rangle + \sqrt{\frac{2\epsilon}{\mu}}||\sigma'_+ - \sigma^{k+1}_+||
\end{aligned}
$$

Taking $\sigma'_+ = \sigma^k_+$ and utilizing the fact that $\sigma^{k+1}_+ - \sigma^k_+ = \alpha\hat{g}$ gives the result.

A.5. Proof of Theorem 2

Let $\phi(\sigma_+) = \min_{\sigma'_+ \in \mathcal{P}^\star} ||\sigma_+(s) - \sigma'_+||$. Based on the Theorem 4.1 of Luo and Tseng (1993), there exists a constant τ such that

$$\phi(\sigma^k_+) + ||\rho^\star - \rho|| \leq \tau||\sigma^{k+1}_+ - \sigma^k_+|| \tag{12}$$

This shows that the distance to $\mathcal{P}^\star$ decreases linearly with $\hat{g}$. From Lemma 2, it is clear that $d(\sigma_+)$ monotonically increases for a sequence generated by Algorithm 1 when $||\hat{g}|| > 2\frac{1}{\alpha}\sqrt{\frac{2\epsilon}{\mu}}$, and the imperfect dual ascent would reach a $\hat{\sigma}_+$ satisfying $||\hat{g}(\hat{\sigma}_+, \alpha)|| < \frac{2}{\alpha}\sqrt{\frac{2\epsilon}{\mu}}$. From the fact that projection does not increase the distance between vectors, $||g(\sigma_+, \alpha)|| < (\frac{2}{\alpha}+1)\sqrt{\frac{2\epsilon}{\mu}}$. Taking $\lambda = \sqrt{2}\tau(2+\alpha)$, then it follows that $\min_{\sigma'_+ \in \mathcal{P}^\star} ||\hat{\sigma}_+ - \sigma'_+|| \leq \lambda\sqrt{\frac{\epsilon}{\mu}}$. Then we have

$$
\begin{aligned}
& d(\hat{\sigma}_+) = d(\sigma'_+) + \int_0^1 \nabla d(\sigma'_+ + t(\hat{\sigma}_+ - \sigma'_+))\mathrm{d}t(\hat{\sigma}_+ - \sigma'_+) \\
& \leq d(\sigma'_+) + \int_0^1 \nabla d(\sigma'_+) + \frac{1}{\mu}(\hat{\sigma}_+ - \sigma'_+)t\mathrm{d}t(\hat{\sigma}_+ - \sigma'_+) \\
& = d(\sigma'_+) + \int_0^1 \nabla d(\sigma'_+)(\hat{\sigma}_+ - \sigma'_+)\mathrm{d}t + \frac{1}{2\mu}||\hat{\sigma}_+ - \sigma'_+||^2 \\
& = d(\sigma'_+) + \frac{1}{2\mu}||\hat{\sigma}_+ - \sigma'_+||^2
\end{aligned}
$$

Also note that $d(\hat{\sigma}_+) - d(\sigma'_+) \geq 0$. Thus we have:

$$\min_{\sigma'_+ \in \mathcal{P}^\star} ||d(\hat{\sigma}_+) - d(\sigma'_+)|| \leq \lambda^2\frac{1}{2\mu}\frac{\epsilon}{\mu} = \frac{\lambda^2}{2}\frac{\epsilon}{\mu^2} \tag{13}$$

B. Supplementary Experiments

In this section, we provide additional case studies that are not covered in the main paper. We mainly compare with RCPO (Tessler et al., 2019) and the unconstrained DDPG (Lillicrap et al., 2016), which serves as the upper bound of the reward that can be achieved if the constraints are ignored.

B.1. Express Delivery Service

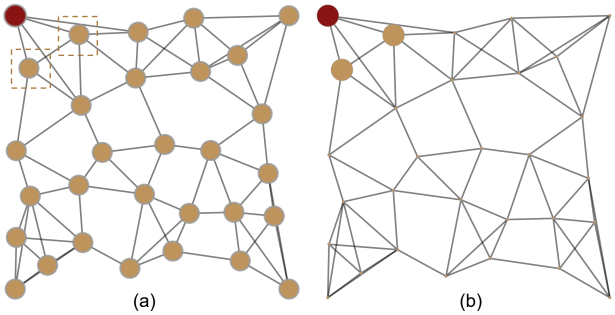


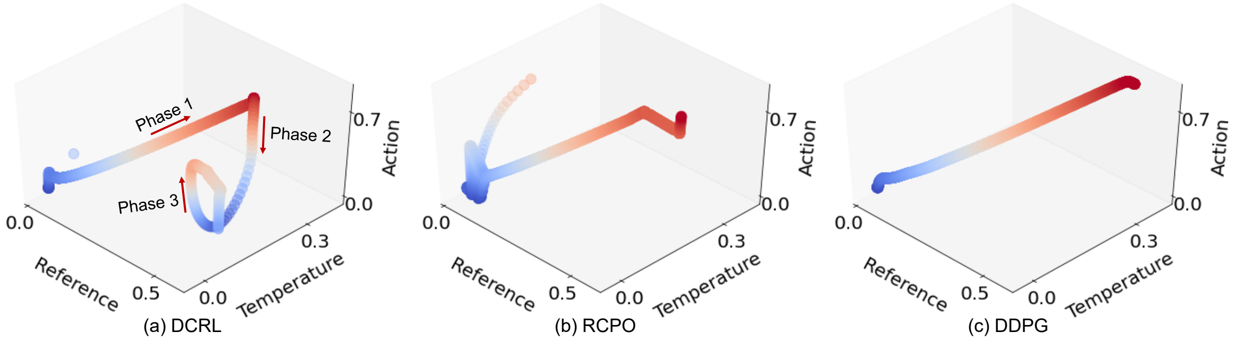
Figure 9. An example of the express delivery service company's transportation network with one ship center (red) and 29 service points (gold). (a) The vans start from the service points bounded by squares with equal probability, then visit other service points following a transition probability (policy), and finally reach the ship center (goal). (b) The standard Q-Learning method finds a policy that drives the vans directly to the goal without visiting any other service points, which minimizes the cost (traveling distance). The sizes of gold nodes represent the state density.

An express delivery service company has several service points and a ship center in a city. An example configuration is illustrated in Figure 9 (a). The company uses vans to transport the packages from each service point to the ship center. The vans start from some service points following an initial distribution, travel through some service points and finally reach the ship center. The cost is formulated as the traveling distance. The frequency that each service point is visited by vans should exceed a given threshold in order to transport the packages in the service points to the ship center. Such frequency constraints can be naturally viewed as density constraints. A policy is represented as the transition probability of the vans from one point to surrounding points. The optimal policy should satisfy the density constraints and minimize the transportation distance.

This case study is proposed to further understand Algorithm 1 and its key steps. In Algorithm 1, our approach adds Lagrange multipliers to the original reward in order to compute a policy that satisfies density constraints. The update of Lagrange multipliers follows the dual ascent, which is key to satisfying the KKT conditions. In this experiment, we try to update the Lagrange multipliers using an alternative approach and see how the performance changes. We replace the dual ascent with the cross-entropy method, where a set of Lagrange multipliers $\Sigma = [\sigma_1, \sigma_2, \sigma_3, \cdots]$ are drawn from an initial distribution $\sigma \sim Z(\sigma)$ and utilized to adjust the reward respectively, after which a set of policies

Table 1. Results of the express delivery service transportation task. The maximum allowed running time to solve for a feasible policy is 600s. The cost is the expectation of traveling distance from initial states to the goal.

Density space	Method	$\rho_{min} = 0.1$			$\rho_{min} = 0.3$			$\rho_{min} = 0.5$		
		Solved	Time (s)	Cost	Solved	Time (s)	Cost	Solved	Time (s)	Cost
$\rho_s \in \mathbb{R}^{10}$	CERS	True	163.69	3.85	True	183.81	5.36	True	466.32	5.12
	DCRL	True	1.35	2.91	True	2.05	2.66	True	4.31	4.28
$\rho_s \in \mathbb{R}^{20}$	CERS	True	227.15	5.80	True	527.26	6.00	False	Timeout	-
	DCRL	True	3.41	5.58	True	3.99	6.16	True	5.28	6.27
$\rho_s \in \mathbb{R}^{100}$	CERS	True	161.62	10.26	False	Timeout	-	False	Timeout	-
	DCRL	True	3.53	10.23	True	116.24	12.13	True	153.86	14.06



(a) DCRL (b) RCPO (c) DDPG

Figure 10. Visualization of the behavior of three methods in the safe electrical motor control task.

$[\pi_1, \pi_2, \pi_3, \cdots]$ are obtained following the same procedure in Algorithm 1. A small subset of Σ whose π has the least violation of the density constraints are chosen to compute a new distribution $Z(\sigma)$, which is utilized to sample a new Σ. The loop continues until we find a σ whose π completely satisfies the density constraints. We call this cross-entropy reward shaping (CERS). We experiment with 10D, 20D and 100D state spaces (corresponding to 10, 20 and 100 service points in the road network), whose density constraints lie in $\mathbb{R}^{10}$, $\mathbb{R}^{20}$ and $\mathbb{R}^{100}$ respectively. The density constraint vector $\rho_{min} : S \mapsto \mathbb{R}$ is set to identical values for each state (service point). For example, $\rho_{min} = 0.1$ indicates the minimum allowed density at each state is 0.1. In Algorithm 1, we use Q-Learning to update the policy for both DCRL and CERS since the state and action space are discrete.

From Table 1, there are two important observations. First, our computational time of finding the policy is significantly less than that of CERS. When $\rho_s \in \mathbb{R}^{10}$ and $\rho_{min} = 0.1$, our approach is at least 100 times faster than CERS on the same machine. When $\rho_s \in \mathbb{R}^{100}$ and $\rho_{min} = 0.5$, CERS cannot solve the problem (no policy found can completely satisfy the constraints) in the maximum allowed time (600s), while our approach can solve the problem in 153.86s. Second, the cost reached by our method is generally lower than that of CERS, which means our method can find better solutions in most cases.

B.2. Safe Electrical Motor Control (Section 5.3)

To gain more insight on the behavior of our DCRL agent, we visualize the trajectories and actions (duty cycles) taken at different temperatures and reference angular velocities in Figure 10. In Figure 10 (a), The trajectory using DCRL can be divided into three phases. In Phase 1, as the reference angular velocity grows, the duty cycle also increases, so the motor temperature goes up. When the temperature is too high, the algorithm enters Phase 2 where it reduces the duty cycle to control the temperature, even though the reference angular velocity remains high. As the temperature goes down, the algorithm enters Phase 3 and increases the duty cycle again to drive the motor angular velocity closer to the reference. In Figure 10 (b), when the temperature is high, the RCPO algorithm will stop increasing the duty cycle but will not decrease it as Algorithm 1 does. So the temperature remains high and thus the density constraints are violated. In Figure 10 (c), the unconstrained DDPG algorithm continues to increase the duty cycle in spite of the high temperature.

B.3. Agricultural Spraying Drone (Section 5.4)

In the agricultural pesticide spraying problem, we examine the methods with different pesticide density requirements and drone configurations to assess their capability of generalizing to new scenarios. In our main paper, from area 0 to 4, the minimum and maximum pesticide density are

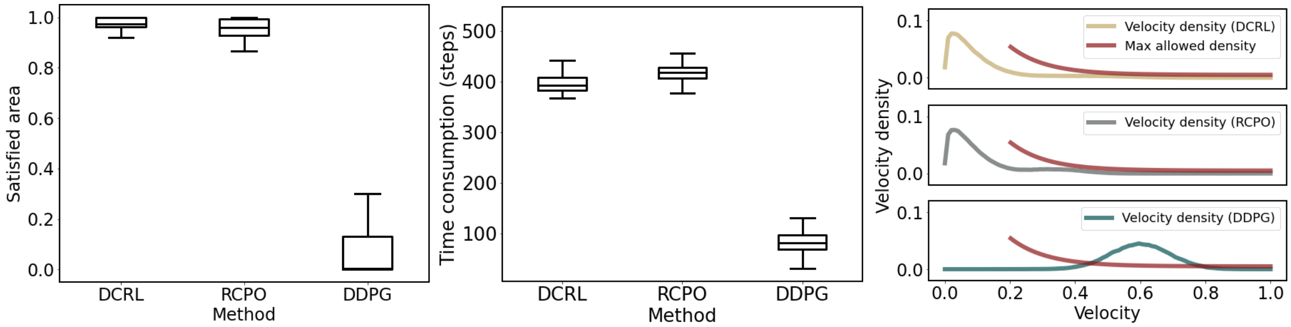


Figure 11. Results of the agricultural spraying problem with minimum pesticide density $(0, 1, 1, 0, 1)$ and maximum density $(0, 2, 2, 0, 2)$ from area 0 to 4. Left: Percentage of the entire area that satisfies the pesticide density requirement. Middle: Time consumption in steps. Whiskers in the left and middle plots denote confidence intervals. Right: visualization of the velocity densities using different methods.

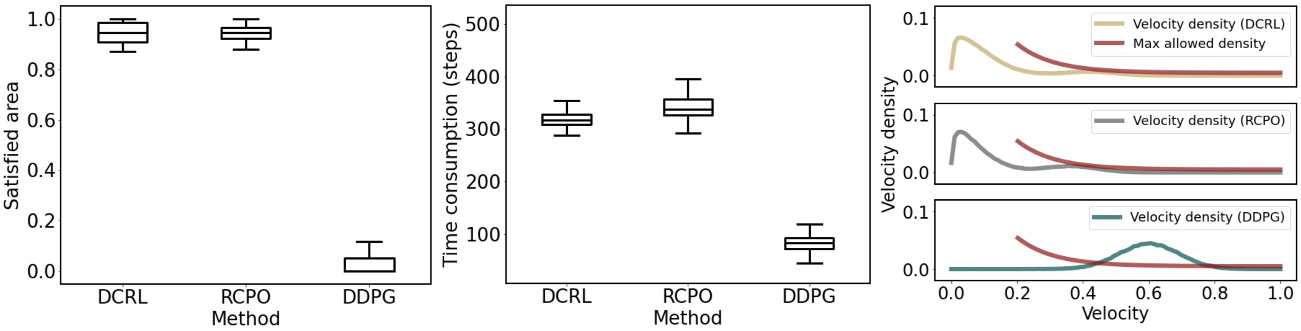


Figure 12. Results of the agricultural spraying problem with minimum pesticide density $(0, 0, 1, 1, 0)$ and maximum density $(0, 0, 2, 2, 0)$ from area 0 to 4. Left: Percentage of the entire area that satisfies the pesticide density requirement. Middle: Time consumption in steps. Whiskers in the left and middle plots denote confidence intervals. Right: visualization of the velocity densities using different methods.

$(1, 0, 0, 1, 1)$ and $(2, 0, 0, 2, 2)$ respectively. In this supplementary material, we evaluate with two new configurations. In Figure 11, the minimum and maximum density are set to $(0, 1, 1, 0, 1)$ and $(0, 2, 2, 0, 2)$ from area 0 to 4. In Figure 12, the minimum and maximum density are set to $(0, 0, 1, 1, 0)$ and $(0, 0, 2, 2, 0)$ from from area 0 to 4.

Although the settings are different from our main paper, the results convey consistent information. In Figure 11 and 12, DCRL and RCPO demonstrates similar performance in controlling pesticide densities to be within the minimum and maximum thresholds, while DCRL demands less time to finish the task. DDPG only minimizes the time consumption and thus requires the least time among the three methods, but cannot guarantee the pesticide density is satisfied. In terms of the velocity control, both DCRL and RCPO can avoid the high-speed movement. These observations suggest that when both DCRL and RCPO finds feasible policies satisfying density constraints, the policy found by DCRL can achieve lower cost or higher reward defined by the original unconstrained problem, which is the time consumption of executing the task in this case study.

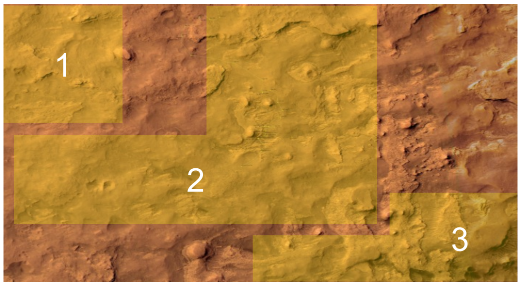


Figure 13. The mars rover environment. The agent starts from a random location in area 1 and is required to reach area 3. Area 2 is considered dangerous and the constraint is set on the total time that the agent is within area 2.

B.4. Mars Rover

We consider a mars rover task where the agent must constrain the amount of time within the dangerous region. The environment is shown in Figure 13, where there are three areas marked in yellow. The agent starts from a random location in area 1 and needs to reach area 3. At each timestep, the agent will receive a negative reward proportional to the energy consumption rate, and will receive a +10 reward after it reaches area 3. Area 2 is considered dangerous and the

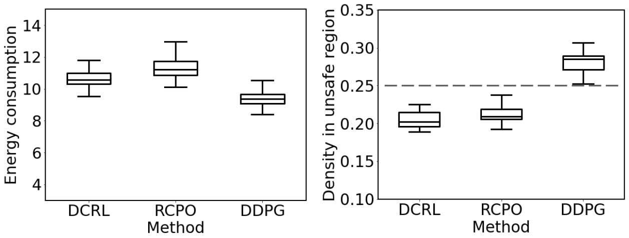


Figure 14. Results of the mars rover task.

amount of time that the agent stays in area 2 is constrained by an upper bound. In the RCPO method, the agent receives a negative reward if it is inside area 2, and the magnitude of this negative reward is automatically tuned by RCPO itself. In our DCRL method, the constraint is converted to an equivalent state density constraint in area 2.

The objective is to minimize the energy consumption while respecting the time constraint in area 2. Figure 14 shows the performance of the 3 methods. It is shown that DCRL and RCPO have similar energy consumption and both satisfy the density constraint. DDPG only minimizes the energy consumption and does not consider the constraint in area 2.