Appendix A
Proof of Theorem 1

Theorem 1 (The Generalized Policy Gradient Theorem)**.** *The derivative of $J_s(\pi_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$ is the expectation of the product of the $\pi_{\boldsymbol{\theta}}$-induced trajectory's SOTA probability and the gradient of the* $\log$ *of policy $\pi_{\boldsymbol{\theta}}$, i.e., $\boldsymbol{\nabla_\theta} J_s(\pi_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau\sim\pi}[\mathbb{P}[R(\tau) \le T]\boldsymbol{\nabla_\theta} \log \pi_{\boldsymbol{\theta}}(\tau)]$, where τ is $\pi_{\boldsymbol{\theta}}$-induced trajectory, $R(\tau)$ is τ's total travel time (an RV), and $\pi_{\boldsymbol{\theta}}(\tau)$ refers to the probability of generating τ by $\pi_{\boldsymbol{\theta}}$.*

Proof.

$$\begin{aligned}
\boldsymbol{\nabla_\theta} J_s(\pi_{\boldsymbol{\theta}}) &= \boldsymbol{\nabla_\theta}\mathbb{P}[t^\pi(o,d) \le T] \\
&= \boldsymbol{\nabla_\theta}\int_\tau \mathbb{P}[R(\tau) \le T]\pi_{\boldsymbol{\theta}}(\tau)\,d\tau \\
&= \int_\tau \mathbb{P}[R(\tau) \le T]\boldsymbol{\nabla_\theta}\pi_{\boldsymbol{\theta}}(\tau)\,d\tau \\
&= \int_\tau \mathbb{P}[R(\tau) \le T]\frac{\boldsymbol{\nabla_\theta}\pi_{\boldsymbol{\theta}}(\tau)}{\pi_{\boldsymbol{\theta}}(\tau)}\pi_{\boldsymbol{\theta}}(\tau)\,d\tau \\
&= \int_\tau \mathbb{P}[R(\tau) \le T]\boldsymbol{\nabla_\theta}\log\pi_{\boldsymbol{\theta}}(\tau)\pi_{\boldsymbol{\theta}}(\tau)\,d\tau \\
&= \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}[\mathbb{P}[R(\tau) \le T]\boldsymbol{\nabla_\theta}\log\pi_{\boldsymbol{\theta}}(\tau)].
\end{aligned}$$

□

In the proof process, $\pi_{\boldsymbol{\theta}}(\tau)$ refers to the probability that trajectory τ is generated by $\pi_{\boldsymbol{\theta}}$.

Appendix B
Proof of Theorem 2

Theorem 2 (Off-Policy Generalized Policy Gradient Theorem)**.** *The gradient of $J_s(\pi_{\boldsymbol{\theta}})$ for behavior policy ($\mu_{\boldsymbol{\theta}}$) generated trajectory τ can be expressed as:*

$$\boldsymbol{\nabla} J_s(\pi_{\boldsymbol{\theta}}) = \mathbb{E}_{\tau\sim\mu_{\boldsymbol{\theta}}}[\mathbb{P}[R(\tau) \le T](\boldsymbol{\nabla_\theta}\log\pi_{\boldsymbol{\theta}}(\tau))\rho(\tau)], \tag{14}$$

where $\rho(\tau) = \pi_{\boldsymbol{\theta}}(\tau)/\mu_{\boldsymbol{\theta}}(\tau)$ is the importance sampling ratio of π to μ with respect to trajectory τ.

Proof.

$$\begin{aligned}
\boldsymbol{\nabla_\theta} J_s(\pi_{\boldsymbol{\theta}}) &= \boldsymbol{\nabla_\theta}\mathbb{P}[t^\pi(o,d) \le T] \\
&= \boldsymbol{\nabla_\theta}\int_\tau \mathbb{P}[R(\tau) \le T]\pi_{\boldsymbol{\theta}}(\tau)\,d\tau \\
&= \int_\tau \mathbb{P}[R(\tau) \le T]\boldsymbol{\nabla_\theta}\pi_{\boldsymbol{\theta}}(\tau)\,d\tau \\
&= \int_\tau \mathbb{P}[R(\tau) \le T]\frac{\boldsymbol{\nabla_\theta}\pi_{\boldsymbol{\theta}}(\tau)}{\pi_{\boldsymbol{\theta}}(\tau)}\frac{\pi_{\boldsymbol{\theta}}(\tau)}{\mu_{\boldsymbol{\theta}}(\tau)}\mu_{\boldsymbol{\theta}}(\tau)\,d\tau \\
&= \int_\tau \mathbb{P}[R(\tau) \le T]\boldsymbol{\nabla_\theta}\log\pi_{\boldsymbol{\theta}}(\tau)\rho(\tau)\mu_{\boldsymbol{\theta}}(\tau)\,d\tau \\
&= \mathbb{E}_{\tau\sim\mu_{\boldsymbol{\theta}}}[\mathbb{P}[R(\tau) \le T]\big(\boldsymbol{\nabla_\theta}\log\pi_{\boldsymbol{\theta}}(\tau)\big)\rho(\tau)].
\end{aligned}$$

□