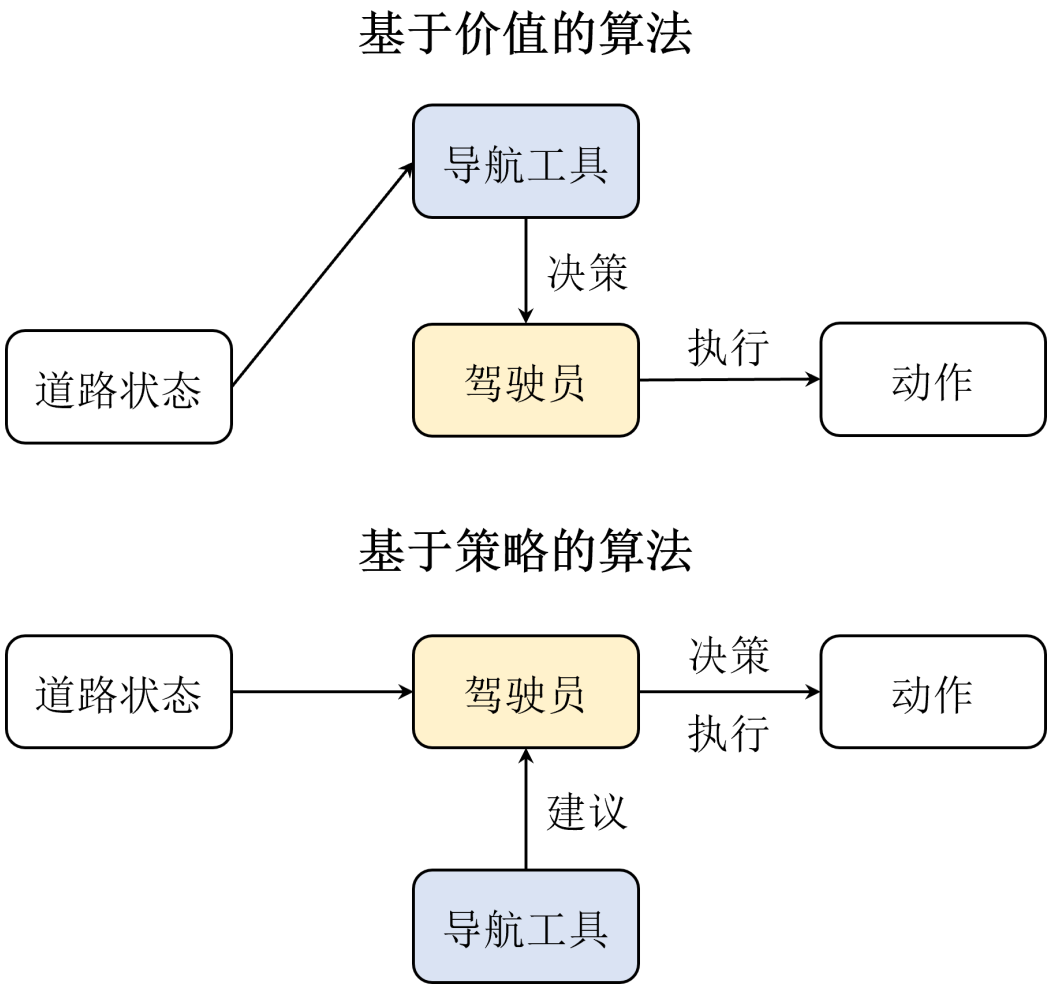
9 - 策略梯度

基于价值算法的缺点

- 无法表示连续动作
  - 由于智能体在学习状态和动作之间的关系时都是在离散时间点上进行的，因此无法表示连续时间轴上的动作判断
- 高方差
  - 价值算法会导致高方差，因此影响到了算法的收敛性
- 探索与利用的平衡问题
  - 无法很好地使用探索策略去模拟真实的环境

策略梯度算法

策略梯度算法是一类直接对策略进行优化的算法。



区别于之前学的算法——也就是基于价值的算法，基于策略的算法可以在训练的过程中得到一个很好的泛化效果。

REINFORCE算法

我们现在需要求解的是目标函数的梯度，REINFORCE 算法的做法是每次采样 N 条轨迹，然后对这 N 条轨迹的梯度求平均

$$\nabla J_\theta = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} G_t^n \nabla \log \pi_\theta(a_t^n|s_t^n)$$

其中 N 理论上越大越好，但实际上我们可能只采样几个回合的轨迹就能近似求解梯度了。