第 3 章　强化学习的抽象视角

本章将用几何构造来洞察贝尔曼方程、值迭代算法、策略迭代算法、值空间近似以及对应的单步或者多步前瞻策略 $\tilde{\mu}$ 的一些性质。为了理解这些构造，我们需要使用贝尔曼方程中涉及的算子的抽象符号框架。

3.1 贝尔曼算子

我们用 TJ 标记出现在贝尔曼方程右侧的 x 的函数，其在状态 x 的取值给定如下

$$(TJ)(x) = \min_{u \in U(x)} E\left\{g(x,u,w) + \alpha J\left(f(x,u,w)\right)\right\}, \forall x \tag{3.1}$$

对每个策略 μ，引入对应的函数 $T_\mu J$，其在 x 的取值给定如下

$$\left(T_\mu J\right)(x) = E\left\{g\left(x,\mu(x),w\right) + \alpha J\left(f\left(x,\mu(x),w\right)\right)\right\}, \forall x \tag{3.2}$$

所以 T 和 T_μ 可以被视作算子（宽泛地称为贝尔曼算子），将函数 J 映射成其他的函数（分别是 TJ 或者 $T_\mu J$）。[①]

算子 T 和 T_μ 的一个重要性质是单调性，意思是：如果 J 和 J' 是 x 的两个函数且满足

$$J(x) \geqslant J'(x), \forall x$$

则有

$$(TJ)(x) \geqslant (TJ')(x), \left(T_\mu J\right)(x) \geqslant \left(T_\mu J'\right)(x), \forall x \text{和} \mu \tag{3.3}$$

在式 (3.1) 和式 (3.2) 中 J 的取值与非负数相乘，于是式 (3.3) 的单调性是显而易见的。

另一个重要的性质是贝尔曼算子 T_μ 是线性的，即其具有 $T_\mu J = G + A_\mu J$ 的形式，其中 $G \in R(X)$ 是某个函数，$A_\mu : R(X) \mapsto R(X)$ 是一个算子且满足对任意函数 J_1，J_2 和标量 γ_1，γ_2，我们有[②]

$$A_\mu(\gamma_1 J_1 + \gamma_2 J_2) = \gamma_1 A_\mu J_1 + \gamma_2 A_\mu J_2$$

进一步，从式 (3.1) 和式 (3.2) 的定义，我们有

$$(TJ)(x) = \min_{\mu \in \mathcal{M}} (T_\mu J)(x), \forall x$$

其中 $\mathcal{M}$ 是平稳策略构成的集合。因为对任意策略 μ，对应于两个不同状态 x 和 x' 的控制 $\mu(x)$ 和 $\mu(x')$ 之间没有耦合约束，于是上式成立。于是有对每个 $x(TJ)(x)$ 是 J 的凹函数（线性函数的逐点最小化是凹函数）。这对于将单步和多步前瞻最小化解释为求解贝尔曼方程 $J = TJ$ 的牛顿迭代是很重要的。

---

① 在本书中，T 和 T_μ 运算的函数 J 是 x 的实值函数，记为 $J \in R(X)$。我们将自始至终假设当 J 是实值时，式 (3.1) 和式 (3.2) 中的期望值定义良好且有限。这意味着 $T_\mu J$ 也将是 x 的实值函数。另一方面因为式 (3.1) 中的最小化，所以 $(TJ)(x)$ 可能取值 $-\infty$。我们允许这一可能性，尽管我们的解释将主要描述当 TJ 为实值的情形。注意抽象动态规划的一般性理论采用拓展的实值函数；见 [Ber22a]。

② 具有这一性质的算子 T_μ 通常被称为“仿射”，但是在本书中我们只是称其为“线性”。我们也使用简化的符号来表示逐点的等式和不等式，这样我们用 $J = J'$ 或者 $J \geqslant J'$ 来分别表示 $J(x) = J'(x)$ 或者 $J(x) \geqslant J'(x)$ 对所有的 x 成立。

例 3.1.1（一个包括两个状态和两个控制的例子）

假设存在两个状态 1 和 2，两个控制 u 和 v。考虑策略 μ，在状态 1 施加控制 u，在状态 2 施加控制 v。那么算子 T_μ 的形式为

$$(T_\mu J)(1) = \sum_{y=1}^{2} p_{1y}(u)\left(g(1,u,y) + \alpha J(y)\right) \tag{3.4}$$

$$(T_\mu J)(2) = \sum_{y=1}^{2} p_{2y}(v)\left(g(2,v,y) + \alpha J(y)\right) \tag{3.5}$$

其中 $p_{xy}(u)$ 和 $p_{xy}(v)$ 分别是当前状态为 x 且控制为 u 或者 v 时下一个状态是 y 的概率。显然，$(T_\mu J)(1)$ 和 $(T_\mu J)(2)$ 是 J 的线性函数。贝尔曼方程 $J = TJ$ 的算子 T 的形式为

$$\begin{aligned}(TJ)(1) = \min\Bigg[&\sum_{y=1}^{2} p_{1y}(u)\left(g(1,u,y) + \alpha J(y)\right), \\ &\sum_{y=1}^{2} p_{1y}(v)\left(g(1,v,y) + \alpha J(y)\right)\Bigg]\end{aligned} \tag{3.6}$$

$$\begin{aligned}(TJ)(2) = \min\Bigg[&\sum_{y=1}^{2} p_{2y}(u)\left(g(2,u,y) + \alpha J(y)\right), \\ &\sum_{y=1}^{2} p_{2y}(v)\left(g(2,v,y) + \alpha J(y)\right)\Bigg]\end{aligned} \tag{3.7}$$

于是，$(TJ)(1)$ 和 $(TJ)(2)$ 是两维向量 J 的凹的分片线性函数（线性片数有两片；更一般地，与控制的数量一样多）。这一凹性一般性地成立，因为 $(TJ)(x)$ 是 J 的线性函数集合中的最小值，每个对应于一个 $u \in U(x)$。图 3.1.1 展示了 $\mu(1) = u$ 且 $\mu(1) = v$ 情形下的 $(T_\mu J)(1)$，$\mu(2) = u$ 且 $\mu(2) = v$ 情形下的 $(T_\mu J)(2)$，$(TJ)(1)$ 和 $(TJ)(2)$ 作为 $J = (J(1), J(2))$ 的函数。

从动态规划的视角至关重要的性质是 T 和 T_μ 是否存在不动点；等价地，贝尔曼方程 $J = TJ$ 和 $J = T_\mu J$ 是否在实值函数类中有解，以及解集是否分别包括 J^* 和 J_μ。于是验证 T 或者 T_μ 是压缩映射是重要的。这对于每阶段费用分别有界的折扣问题的良好情形成立。然而，对于无折扣问题，确定 T 或者 T_μ 的压缩性质可能要复杂许多，甚至不可能；抽象动态规划书 [Ber22a] 详细处理了这类问题，以及关于贝尔曼方程的解集的有关问题。

几何解释

我们将用几何的方式解释贝尔曼算子，从 T_μ 开始。图 3.1.2 展示了其形式。这里注意函数 J 和 $T_\mu J$ 是多维的。它们分别拥有与状态 x 的数量一样多的标量成员 $J(x)$ 和 $(T_\mu J)(x)$，但是它们只能被投影到一维来展示。对每个策略 μ，函数 $T_\mu J$ 是线性的。费用函数 J_μ 满足 $J_\mu = T_\mu J_\mu$，所以当 J_μ 是实值时，从 $T_\mu J$ 的图和 45 度线的交集中可得 J_μ。稍后，我们将 J_μ 非实值的情形解释为系统在 μ 之下缺乏稳定性 [我们对某些初始状态 x 有 $J_\mu(x) = \infty$]。

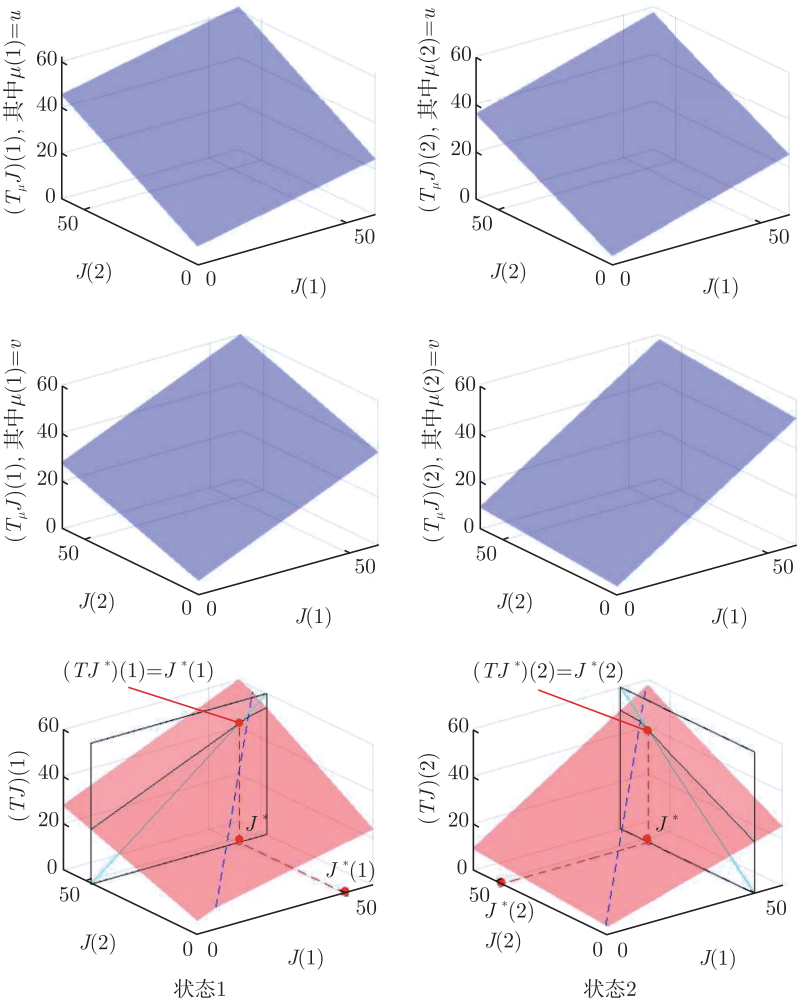

图 3.1.1　例 3.1.1 中状态 1 和 2 的贝尔曼算子 T_μ 和 T 的几何图示；参见式 (3.4)∼ 式 (3.7)。问题的转移概率是：$p_{11}(u)=0.3, p_{12}(u)=0.7, p_{21}(u)=0.4, p_{22}(u)=0.6, p_{11}(v)=0.6, p_{12}(v)=0.4, p_{21}(v)=0.9, p_{22}(v)=0.1$。各阶段费用为 $g(1,u,1)=3, g(1,u,2)=10, g(2,u,1)=0, g(2,u,2)=6, g(1,v,1)=7, g(1,v,2)=5, g(2,v,1)=3, g(2,v,2)=12$。折扣因子是 $\alpha=0.9$，最优费用是 $J^*(1)=50.59$ 和 $J^*(2)=47.41$。最优策略是 $\mu^*(1)=v$ 和 $\mu^*(2)=u$。图中也展示了与 $J(1)$ 和 $J(2)$ 轴平行且经过 J^* 的 T 的两个一维切片。

贝尔曼算子 T 的形式示于图 3.1.3 中。再次说明，函数 J，J^*，TJ，$T_\mu J$ 等是多维的，但是它们被投影到一维之上（即所展示的是这些函数在特定系统上的情形，该系统有单个状态且可能还有一个终止状态）。贝尔曼方程 $J=TJ$ 可能有一个或者许多实值解。它可能在特殊的情形之下没有实值解，正如我们将稍后讨论的那样（见 3.8 节）。图中假设贝尔曼方程 $J=TJ$ 和 $J=T_\mu J$ 有唯一实值解，如果 T 和 T_μ 是压缩映射的话这是成立的，正如对于每阶段费用有界的折扣问题那样。否则，这些方程可能在实值函数中没有解或者有多个解（见 3.8 节）。方程 $J=TJ$ 通常以 J^* 为解，但是在 $\alpha=1$ 或者 $\alpha<1$ 且每阶段费用无界的情形下拥有多于一个解。

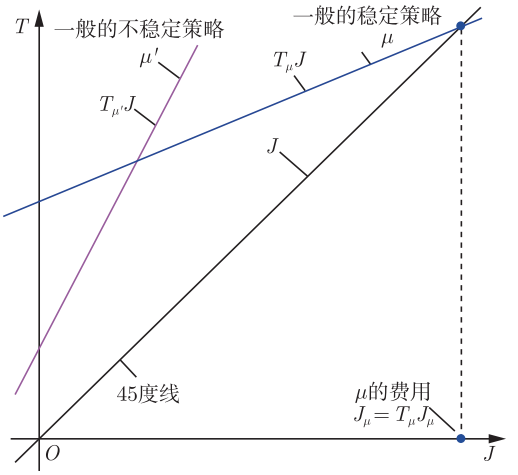

图 3.1.2　线性贝尔曼算子 T_μ 及对应的贝尔曼方程的几何解释。T_μ 的图是在 $R(X) \times R(X)$ 空间中的平面，并且当投影到对应于单个状态且经过 J_μ 的一维空间的时候，它变成了一条线。那么存在三种情形：（a）线的坡度小于 45 度，于是与 45 度线有唯一交点，该交点等于 J_μ，为贝尔曼方程 $J = T_\mu J$ 的解。如果 T_μ 是一个压缩映射，那么这是成立的，与在每阶段费用有界的折扣问题的情形中一样。（b）线的坡度大于 45 度。于是与 45 度线相交于唯一点，这是贝尔曼方程 $J = T_\mu J$ 的解，但不等于 J_μ。那么 J_μ 不是实值的；我们在 3.2 节中称这样的 μ 是*不稳定的*。（c）线的坡度正好等于 45 度。这是一种特殊的情形，其中贝尔曼方程 $J = T_\mu J$ 要么拥有无穷多的实值解，要么根本没有实值解；我们将在 3.8 节中提供出现这种情形的例子。

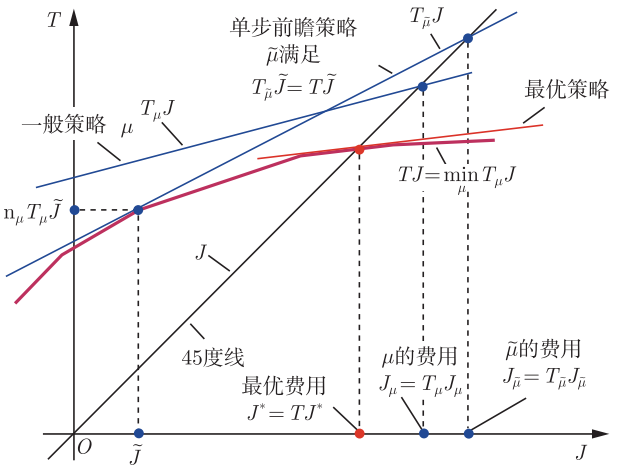

图 3.1.3　贝尔曼算子 T 和对应的贝尔曼方程的几何解释。对于固定的 x，函数 $(TJ)(x)$ 可以写成 $\min_\mu (T_\mu J)(x)$，所以它是 J 的凹函数。最优费用函数 J^* 满足 $J^* = TJ^*$，所以它是从 TJ 的图与所示的 45 度线的交集产生的，假设 J^* 是实值的。

注意 T 的图位于每个算子 T_μ 的图之下，而且事实上随着 μ 在策略集合 $\mathcal{M}$ 上变化取值时 T 的图是 T_μ 的图的下包络线。特别地，对于任意给定的函数 $\tilde{J}$，对于每个 x，值 $(T\tilde{J})(x)$ 通过找到凹函数 $(TJ)(x)$ 的图在 $J = \tilde{J}$ 的支撑超平面/次梯度从而获得，正如在图中所示。这一支撑超平面由在 μ 上达到 $(T_\mu \tilde{J})(x)$ 最小值的策略 $\tilde{\mu}$ 的控制 $\mu(x)$ 而定义

$$\tilde{\mu}(x) \in \arg\min_{\mu \in \mathcal{M}} (T_\mu \tilde{J})(x)$$

（可能存在多个策略达到这一最小值，从而定义多个支撑超平面）。

例 3.1.2 （一个有两个状态和无限控制的问题）

让我们考虑对于一个涉及两个状态 1 和 2，但是无限多控制的问题的映射 T。特别地，两个状态的控制空间都是单位区间，$U(1) = U(2) = [0,1]$。这里 $(TJ)(1)$ 和 $(TJ)(2)$ 给定如下

$$(TJ)(1) = \min_{u\in[0,1]} \left\{ g_1 + r_{11}u^2 + r_{12}(1-u)^2 + \alpha u J(1) + \alpha(1-u)J(2) \right\}$$

$$(TJ)(2) = \min_{u\in[0,1]} \left\{ g_2 + r_{21}u^2 + r_{22}(1-u)^2 + \alpha u J(1) + \alpha(1-u)J(2) \right\}$$

在每个状态 $x = 1, 2$，控制 u 的含义是我们在那个状态必须选择的概率。特别地，我们控制移动到状态 $y = 1$ 和 $y = 2$ 的概率 u 和 $(1-u)$，对应的控制费用分别是 u 和 $(1-u)$ 的二次型形式。对于这个问题，$(TJ)(1)$ 和 $(TJ)(2)$ 可以闭式计算，所以易于绘制和理解。它们是分片二次的，与图 3.1.1 中分片线性的图示不同，见图 3.1.4。

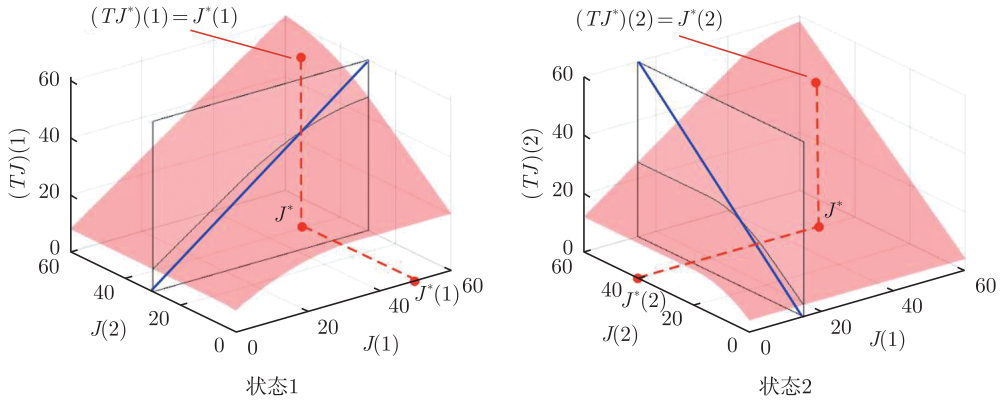

图 3.1.4 例 3.1.2 的状态 1 和 2 的贝尔曼算子 T 的示意图。参数取值是 $g_1 = 5, g_2 = 3, r_{11} = 3, r_{12} = 15, r_{21} = 9, r_{22} = 1$，折扣因子是 $\alpha = 0.9$。最优费用是 $J^*(1) = 49.7$ 和 $J^*(2) = 40.0$，最优策略是 $\mu^*(1) = 0.59$ 和 $\mu^*(2) = 0$。该图也展示了与 $J(1)$ 和 $J(2)$ 轴平行的算子在 $J(1) = 15$ 和 $J(2) = 30$ 的两个一维切片。

值迭代的可视化

算子符号简化了与强化学习有关的算法描述、推导和证明。例如，我们可以将值迭代算法写成如下的紧凑形式

$$J_{k+1} = TJ_k, k = 0, 1, \cdots$$

正如在图 3.1.5 中所示。进一步，给定策略 μ 的值迭代算法可以写成

$$J_{k+1} = T_\mu J_k, k = 0, 1, \cdots$$

且可以类似地解释，这里函数 $T_\mu J$ 的图是线性的。我们也可以很快看到存在策略迭代算法的类似的紧凑描述。

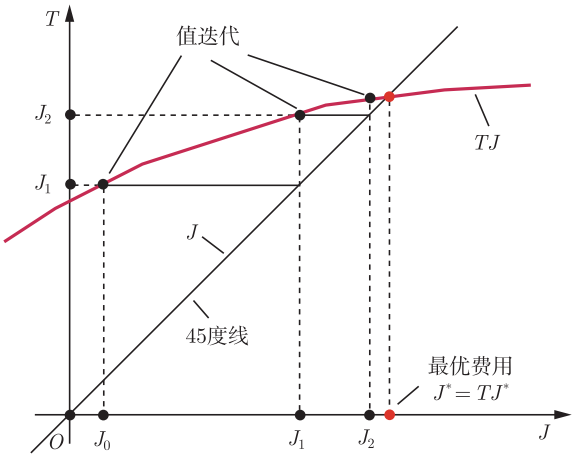

图 3.1.5　从某个初始函数 J_0 开始的值迭代算法 $J_{k+1} = TJ_k$ 的几何解释。后续迭代通过在图中所示的阶梯状构造获得。对于给定策略 μ 的值迭代算法 $J_{k+1} = T_\mu J_k$ 可以类似地解释，除了函数 $T_\mu J$ 的图是线性的。

为了让表述简单，我们将集中关注抽象动态规划框架，因为其适用于 2.1 节中的最优控制问题。特别地，我们假设 T 和 T_μ 具有式 (3.3) 的单调性，$T_\mu J$ 对所有的 μ 是线性的，（结果）对每个状态 x 元素 $(TJ)(x)$ 是 J 的凹函数。然而我们指出抽象符号方便了将无限时段动态规划理论推广到本书所讨论的范围之外的模型。这样的模型包括半马尔可夫问题、极小化极大控制问题、风险敏感问题、马尔可夫博弈以及其他（见动态规划教材 [Ber12] 和抽象动态规划学术专著 [Ber22a] ）。

3.2　值空间近似和牛顿法

现在考虑值空间近似和抽象的几何解释，这是在作者的书 [Ber20a] 中首次给出的。通过对给定的 $\tilde{J}$ 使用算子 T 和 T_μ，单步前瞻策略 $\tilde{\mu}$ 通过方程 $T_{\tilde{\mu}}\tilde{J} = T\tilde{J}$ 描述，或者等价地

$$\tilde{\mu}(x) \in \arg\min_{u \in U(x)} E\Big\{g(x,u,w) + \alpha \tilde{J}\left(f(x,u,w)\right)\Big\} \tag{3.8}$$

正如在图 3.2.1 中所示。进一步，这一方程意味着 $T_{\tilde{\mu}}J$ 的曲线只是在 $\tilde{J}$ 碰到了 TJ 的曲线，如图 3.2.1 所示。

在数学形式上，对每个状态 $x \in X$，超平面 $H_{\tilde{\mu}}(x) \in R(X) \times \Re$

$$H_{\tilde{\mu}}(x) = \{(J,\xi) \,|\, (T_{\tilde{\mu}}J)(x) = \xi\} \tag{3.9}$$

从上面支撑了凹函数 $(TJ)(x)$ 的亚图，即，凸集合

$$\{(J,\xi) \,|\, (TJ)(x) \geqslant \xi\}$$

支撑点是 $\left(\tilde{J}, \left(T_{\tilde{\mu}}\tilde{J}\right)(x)\right)$，并且将函数 $\tilde{J}$ 与对应的单步前瞻最小化策略 $\tilde{\mu}$ 联系在一起，策略 $\tilde{\mu}$ 满足 $T_{\tilde{\mu}}\tilde{J} = T\tilde{J}$。式 (3.9) 的超平面 $H_{\tilde{\mu}}(x)$ 定义了 $(TJ)(x)$ 在 $\tilde{J}$ 的次梯度。注意单步

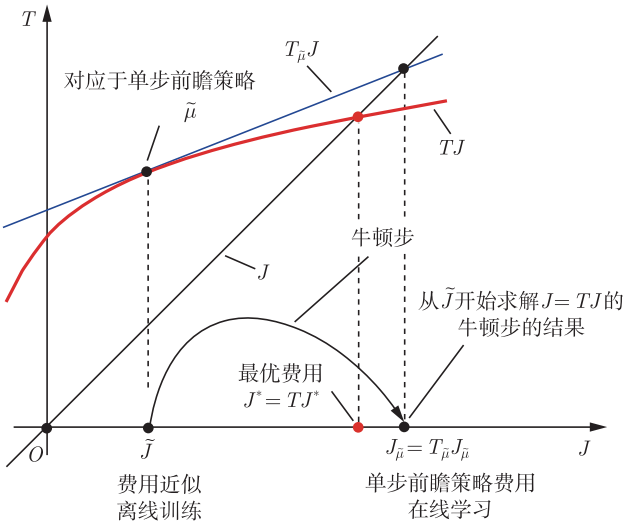

图 3.2.1　值空间近似即单步前瞻策略 $\tilde{\mu}$ 作为一步牛顿法的几何解释 [参见式 (3.8)]。给定 $\tilde{J}$，我们找到在关系式

$$T\tilde{J}=\min_{\mu}T_{\mu}\tilde{J}$$

中达到最小值的策略 $\tilde{J}$。该策略满足 $T\tilde{J}=T_{\tilde{\mu}}\tilde{J}$，所以 TJ 和 $T_{\tilde{\mu}}J$ 的图在 $\tilde{J}$ 相触碰，正如图所示。它可能不唯一。因为 TJ 有凹的元素，方程 $J=T_{\tilde{\mu}}J$ 是方程 $J=TJ$ 在 $\tilde{J}$ 的线性化 [对每个 x，式 (3.9) 的超平面 $H_{\tilde{\mu}}(x)$ 定义了 $(TJ)(x)$ 在 $\tilde{J}$ 的次梯度]。在牛顿法的典型一步中通过求解该线性化方程获得下一个迭代，这就是 $J_{\tilde{\mu}}$。

前瞻策略 $\tilde{\mu}$ 未必唯一，因为 T 未必是可微的，于是可能存在多个超平面在 $\tilde{J}$ 支撑。这一构造仍然展示了线性算子 $T_{\tilde{\mu}}$ 是算子 T 在 $\tilde{J}$ 点（对每个 x 逐点）的线性化。

等价地，对每个 $x\in X$，线性标量方程 $J(x)=(T_{\tilde{\mu}}J)(x)$ 是非线性方程 $J(x)=(TJ)(x)$ 在 $\tilde{J}$ 点的线性化。结果，线性算子方程 $J=T_{\tilde{\mu}}J$ 是方程 $J=TJ$ 在 $\tilde{J}$ 的线性化，其解 $J_{\tilde{\mu}}$ 可视作牛顿迭代在 $\tilde{J}$ 点的结果（我们在这里采用了牛顿迭代的扩展视角，适用于可能不可微的不动点方程组；见附录）。小结一下，在 $\tilde{J}$ 的牛顿迭代是 $J_{\tilde{\mu}}$，这是线性化方程 $J=T_{\tilde{\mu}}J$ 的解。[①]

式 (3.1) 和式 (3.2) 的贝尔曼算子的结构，及其单调性与凹性，倾向于增强牛顿法的收

① 求解 $y=G(y)$ 形式的不动点问题的经典的牛顿法，其中 y 是 n 维向量，按如下方式操作：在当前迭代 y_k，我们将 G 线性化并且找到对应的线性不动点问题的解 y_{k+1}。假设 G 可微，线性化通过使用一阶泰勒展开获得

$$y_{k+1}=G(y_k)+\frac{\partial G(y_k)}{\partial y}(y_{k+1}-y_k)$$

其中，$\partial G(y_k)/\partial y$ 是 G 在向量 y_k 处评价的 $n\times n$ 雅可比矩阵。牛顿法最常见的收敛速率性质是二次收敛性。这一性质指出在解 y^* 的附近，我们有

$$\|y_{k+1}-y^*\|=O(\|y_k-y^*\|^2)$$

其中，$\|\cdot\|$ 是欧氏范数，并且在假设雅可比矩阵存在、可逆、李普希兹连续的条件下成立（见 Ortega 和 Rheinboldt 的书 [OrR70] 和作者的 [Ber16] 第 1.4 节）。

牛顿法存在许多推广形式。在这些推广形式中，仍然采用在当前迭代中求解线性化系统的思想，但是放松了可微性假设条件，如放松为分片可微性、B-可微性和半连续性。这些推广形式的方法依然保持了超线性收敛性质。特别地，在 [Ber16] 一书中命题 1.4.1 分析了可微 G 的二次收敛速率，这一分析是对 [KoS86] 论文中对分片可微 G 的分析的直接且直观的推广；见附录，其中包含了参考文献。

敛性及收敛速率，即便没有可微性时也是如此，支持这一点的证据包括对策略迭代的与牛顿法有关的良好的收敛性分析，以及滚动、策略迭代和模型预测控制的大量良好体验。事实上，在影响牛顿法的收敛性中单调性与凹性扮演的角色已在数学文献中处理。①

正如之前所注意到的，使用 $\tilde{J}$ 进行 l 步前瞻的值空间近似与使用 $\tilde{J}$ 的 $(l-1)$ 重 T 操作（$T^{l-1}\tilde{J}$）的值空间近似相同。所以这可以解释为从在 $\tilde{J}$ 上施加 $l-1$ 次值迭代所得结果，$T^{l-1}\tilde{J}$，开始的一步牛顿步，示于图 3.2.2 中。②

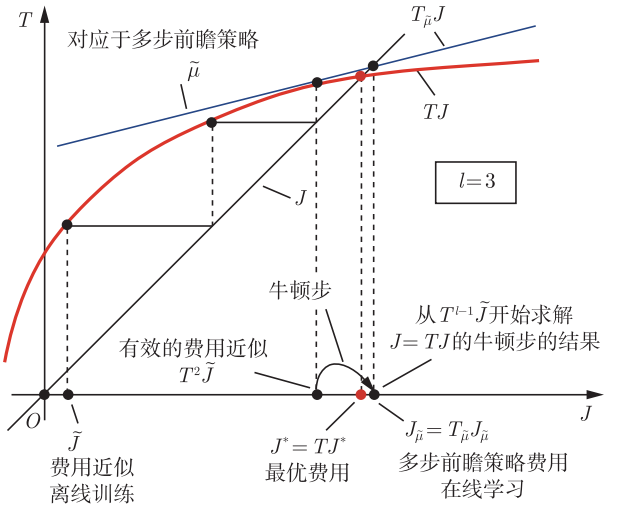

图 3.2.2　采用 l 步前瞻的值空间近似的几何解释（图中 $l=3$）。这与使用 $T^{l-1}\tilde{J}$ 作为费用近似的单步前瞻值空间近似相同。可以被视作在 $T^{l-1}\tilde{J}$ 点的牛顿步，这是对 $\tilde{J}$ 使用 $l-1$ 次值迭代的结果。注意随着 l 增加，l 步前瞻策略 $\tilde{\mu}$ 的费用函数 $J_{\tilde{\mu}}$ 更加接近最优的 J^* 满足 $\lim_{l\to\infty} J_{\tilde{\mu}}=J^*$。

我们也注意到 l 步前瞻最小化涉及 l 轮连续的值迭代，但这些迭代中只有第一次有牛顿步的解释：作为一个例子，考虑具有末端费用近似 $\tilde{J}$ 的两步前瞻最小化。第二步最小化是一次值迭代从 $\tilde{J}$ 开始产生 $T\tilde{J}$。第一步最小化是一次值迭代从 $T\tilde{J}$ 开始产生 $T^2\tilde{J}$，但是也做了其他一些更加重要的事情：这通过 $T_{\tilde{\mu}}(T\tilde{J})=T(T\tilde{J})$ 产生一个两步前瞻最小化策略 $\tilde{\mu}$，且从 $T\tilde{J}$ 到 $J_{\tilde{\mu}}$（$\tilde{\mu}$ 的费用函数）的步骤是牛顿步。所以，仅产生了一个策略（即 $\tilde{\mu}$）且仅有单个牛顿步（从 $T\tilde{J}$ 到 $J_{\tilde{\mu}}$）。在单步前瞻最小化中，牛顿步从 $\tilde{J}$ 开始并止于 $J_{\tilde{\mu}}$。类似地，在 l 步前瞻最小化情形中，第一步前瞻是牛顿步（从 $T^{l-1}\tilde{J}$ 到 $J_{\tilde{\mu}}$），且第一步前瞻之后接下来的不论是什么都是对牛顿步的准备。

最后，值得指出的是，值空间近似算法计算 $J_{\tilde{\mu}}$ 的方式既不同于策略迭代法也不同于经典形式的牛顿法。它并不显式地计算 $J_{\tilde{\mu}}$ 的任意值，而是施加控制到系统上，费用相应

① 见 Ortega 和 Rheinboldt[OrR67] 和 Vandergraft[Van67] 的论文，Ortega 和 Rheinboldt[OrR70] 和 Argyros[Arg08] 的书以及其中引用的参考文献。关于这一联系，值得指出的是，在马尔可夫博弈中，凹性不再成立，策略迭代法可能振荡，正如 Pollatshek 和 Avi-Itzhak[PoA69] 所展示的那样，且需要修订方可恢复其全局收敛性；见作者的论文 [Ber21c] 及其中所引用的参考文献。

② 我们注意到几种牛顿法的变形在数值分析中很著名，设计一阶迭代方法的组合，例如高斯-赛德尔和雅可比算法以及牛顿法。这些方法属于牛顿-SOR 方法的广泛家族（SOR 表示“逐次超松弛”）；见 Ortega 和 Rheinboldt[OrR70] 一书（13.4 节）。只要涉及一步单纯的牛顿步以及一阶步骤，它们的收敛速率是超线性的，与牛顿法类似。

累计。所以 $J_{\bar{\mu}}$ 的值仅对于那些在线生成的系统轨迹中碰到的那些 x 隐式地计算。

确定性等价近似与牛顿步

我们之前注意到对于随机动态规划问题，由于问题的随机特征导致前瞻树随着 l 增加而迅速增长，l 步前瞻可能在计算上是昂贵的，*确定性等价方法*是处理这一难点的一种重要的近似思想。在这一方法的典型形式中，一些随机扰动 w_k 被替换为确定性量，例如它们的期望值。然后对于所得到的确定性问题离线计算出一个策略，并在线应用于真实的随机问题。

确定性等价方法也可用于加速计算 l 步前瞻最小化。一种实现方法是简单地将不确定的 l 个量 $w_k, w_{k+1}, \cdots, w_{k+l-1}$ 中的每一个都替换为确定值 $\bar{w}$。从概念上，这将贝尔曼算子 T 和 T_μ，

$$(TJ)(x) = \min_{u \in U(x)} E\left\{g(x,u,w) + \alpha J\left(f(x,u,w)\right)\right\}$$

$$(T_\mu J)(x) = E\left\{g(x,\mu(x),w) + \alpha J\left(f(x,\mu(x),w)\right)\right\}$$

[参阅式 (3.1) 和式 (3.2)] 替换为算子 $\bar{T}$ 和 $\bar{T}_\mu$，给定如下

$$\left(\bar{T}J\right)(x) = \min_{u \in U(x)} \left[g(x,u,\bar{w}) + \alpha J\left(f(x,u,\bar{w})\right)\right]$$

$$\left(\bar{T}_\mu J\right)(x) = g\left(x,\mu(x),\bar{w}\right) + \alpha J\left(f(x,\mu(x),\bar{w})\right)$$

所得到的 l 步前瞻最小化于是变得更简单了。例如，在有限控制空间的问题中，这是一个确定性最短路径计算，涉及一个无环的 l 阶段图并在每阶段按因数 n 扩展，其中 n 是控制空间的大小。然而，这一方法获得的策略 $\bar{\mu}$ 满足

$$\bar{T}_{\bar{\mu}}\left(\bar{T}^{l-1}\tilde{J}\right) = \bar{T}\left(\bar{T}^{l-1}\tilde{J}\right)$$

且这一策略的费用函数 $J_{\bar{\mu}}$ 由牛顿步生成，旨在从 $\bar{T}^{l-1}\tilde{J}$ 开始找到 $\bar{T}$（而非 T）的一个不动点。于是牛顿步现在的目标是 $\bar{T}$ 的不动点，这不等于 J^*。结果牛顿步的优势在相当大的程度上丧失了。

然而，我们通过仅对 l 步前瞻的*最后* $l-1$ *阶段*使用确定性等价，可显著纠正前述难点，并且保持本质上的简洁。这可以通过如下方式实现：在 l 步前瞻机制中仅将不确定量 $w_{k+1}, w_{k+2}, \cdots, w_{k+l-1}$ 替换为确定性值 $\bar{w}$，而 w_k 被视作随机量。通过这一方式我们获得一个策略 $\bar{\mu}$ 满足

$$T_{\bar{\mu}}\left(\bar{T}^{l-1}\tilde{J}\right) = T\left(\bar{T}^{l-1}\tilde{J}\right)$$

这一策略的费用函数 $J_{\bar{\mu}}$ 再一次从 $\bar{T}^{l-1}\tilde{J}$ 开始由牛顿步生成，其目标是找到 T（而非 $\bar{T}$）的不动点。于是牛顿法的快速收敛性的优势得以恢复。事实上基于从这一牛顿步解释得到的启发，看起来当 $\bar{T}^{l-1}\tilde{J}$ “接近” J^* 时将 l 步前瞻的最后 $l-1$ 阶段变成确定性带来的性能损失是较小的。同时，l 步最小化 $T\left(\bar{T}^{l-1}\tilde{J}\right)$ 仅涉及一个随机步，即第一步，于是与不涉及任何确定性等价近似的 l 步最小化 $T^l\tilde{J}$ 相比可能有更“瘦”的前瞻树。

前述讨论也指向了一种更一般的近似思想，可以处理长程多步前瞻最小化的繁重计算需求。我们可以将 l 步前瞻最小化的后 $(l-1)$ 步的部分 $T^{l-1}\tilde{J}$ 近似为任意产生近似 $\hat{J} \approx T^{l-1}\tilde{J}$ 的简化计算，然后使用最小化

$$T_{\tilde{\mu}}\hat{J} = T\hat{J}$$

获得前瞻策略 $\tilde{\mu}$。这类简化仍将涉及牛顿步（从 $\hat{J}$ 到 $J_{\tilde{\mu}}$），并且从对应的快速收敛性质中受益。

局部与全局性能估计对比

前述对于从 $\tilde{J}$（末端费用函数近似）到 $J_{\tilde{\mu}}$（前瞻策略 $\tilde{\mu}$ 的费用函数）的移动的牛顿步解释建立了超线性的性能估计

$$\max_x |J_{\tilde{\mu}}(x) - J^*(x)| = o\left(\max_x |\tilde{J}(x) - J^*(x)|\right)$$

然而，这一估计在特征上是*局部的*。它仅当 $\tilde{J}$ “接近” J^* 时有意义。当 $\tilde{J}$ 远离 J^* 时，当 $\tilde{\mu}$ 不稳定时，$\max_x |J_{\tilde{\mu}}(x) - J^*(x)|$ 可能很大甚至无穷大（见下一节的讨论）。

对于几类问题存在差分

$$\max_x |J_{\tilde{\mu}}(x) - J^*(x)|$$

的全局估计，包括对于 l 步前瞻以及当贝尔曼算子 T_μ 是压缩映射时的 α 折扣问题，存在差分的上界

$$\max_x |J_{\tilde{\mu}}(x) - J^*(x)| \leqslant \frac{2\alpha^l}{1-\alpha} \max_x |\tilde{J}(x) - J^*(x)|$$

见神经动态规划一书 [BeT96]（6.1 节命题 6.1），或者强化学习一书 [Ber20a]（5.4 节命题 5.4.1）。这些书中也包含其他相关的全局估计，对所有的 $\tilde{J}$ 成立，不论距离 J^* 近或者远。然而，这些全局估计过于保守，且当 $\tilde{J}$ 接近 J^* 时对于值空间近似机制的性能并不具有代表性。在本书中将不考虑，因为它们对于我们想集中关注的启示与方法论没有贡献。例如，对于有限空间 α 折扣 MDP，当 $\max_x |\tilde{J}(x) - J^*(x)|$ 充分小时可证明 $\tilde{\mu}$ 是最优的；这也可以从另一事实看出，即贝尔曼算子的元素 $(TJ)(x)$ 不仅是凹的而且是分片线性的，所以牛顿法在有限步之内收敛。

3.3 稳定域

对于任意的控制系统设计方法，所获得策略的稳定性至关重要。于是调查并验证通过值空间近似机制获得的控制器的稳定性颇为关键。历史上，控制理论中出现过几种关于稳定性的定义。在本书中，我们对稳定性的关注将主要针对具有免费的终止状态 t 且在终止状态之外每阶段费用为正的问题，例如之前介绍的无折扣正费用确定性问题（参见 2.1 节）。进一步，对我们的目的而言最好采用基于优化的定义。特别地，如果 $J_\mu(x) = \infty$ 对某个状态 x 成立，我们说策略 μ 是*不稳定的*。等价地，如果对所有的状态 x 有 $J_\mu(x) < \infty$，那

么我们说策略 μ 是稳定的。这一定义的优点是适用于一般的状态空间和控制空间。自然地，这可以在特定的问题实例中更加具体化。[①]

在值空间近似的上下文中我们感兴趣的是稳定域，这是由近似费用函数 $\tilde{J} \in R(X)$ 构成的集合，要求其对应的单步或者多步前瞻策略 $\tilde{\mu}$ 稳定。对于每阶段费用有界的折扣问题，所有的策略具有实值费用函数，所以没有出现稳定性的问题。然而，一般而言，稳定域可能是实值函数的一个严格子集；这将在稍后对于 2.1 节的线性二次型问题的无折扣确定性情形中展示（参见例 2.1.1）。图 3.3.1 展示了采用单步前瞻的值空间近似的稳定域和不稳定域。

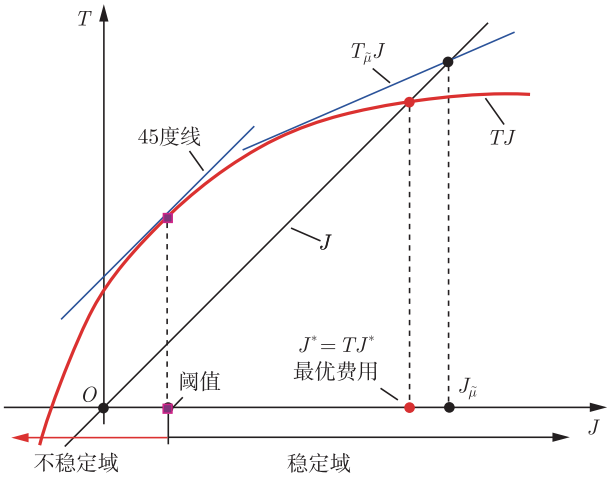

图 3.3.1　采用单步前瞻的值空间近似的稳定域和不稳定域示意图。稳定域是让从单步前瞻最小化 $T_{\tilde{\mu}}\tilde{J} = T\tilde{J}$ 中获得的策略 $\tilde{\mu}$ 对所有 x 满足 $J_{\tilde{\mu}}(x) < \infty$ 的 $\tilde{J}$ 构成的集合。

从图 3.3.1 中看到的一个有趣的现象是，如果 $\tilde{J}$ 不属于稳定域且 $\tilde{\mu}$ 是对应的单步前瞻不稳定策略，那么贝尔曼方程 $J = T_{\tilde{\mu}}J$ 可能拥有实值解。然而，这些解将不等于 $J_{\tilde{\mu}}$，因为这将违反稳定域的定义。一般而言，如果 T_μ 不是压缩映射，T_μ 可能拥有实值不动点，但任何一个都不等于 J_μ。

图 3.3.2 展示了采用 l 步前瞻最小化的值空间近似的稳定域和不稳定域。这张图的启发与图 3.3.1 的单步前瞻情形类似。然而，该图展示了 l 步前瞻控制器 $\tilde{\mu}$ 的稳定域依赖于 l，并且随着 l 增加倾向于变得更大。原因是采用末端费用 $\tilde{J}$ 的 l 步前瞻等于采用末端费用 $T^{l-1}\tilde{J}$ 的单步前瞻，后者倾向于比 $\tilde{J}$ 更加接近最优费用函数 J^*（假设值迭代方法收敛）。

我们如何在稳定域内获得近似函数 $\tilde{J}$？

自然地，识别并获得位于稳定域之内的近似费用函数 $\tilde{J}$ 并采用单步或者多步前瞻对于我们非常重要。我们将对于每阶段期望费用非负的特殊情形关注这一问题

$$E\{g(x,u,w)\} \geqslant 0, \forall x, u \in U(x)$$

① 对于之前介绍的无折扣正费用确定性问题（参见 2.1 节），可以证明，如果策略 μ 是稳定的，那么 J_μ 是贝尔曼方程 $J = T_\mu J$ 在非负实值函数中的“最小”解，并且在宽松的假设条件下，这也是 $J = T_\mu J$ 在满足 $J(t) = 0$ 的非负实值函数 J 中的唯一解；见作者的论文 [Ber17b]。进一步，如果 μ 是不稳定的，那么贝尔曼方程 $J = T_\mu J$ 在非负实值函数中无解。

并且假设 J^* 是实值的。这是模型预测控制中最有趣的情形，但是也在其他有趣的问题中出现，包括涉及终止状态的随机最短路问题。

从图 3.3.2 中可以推断，如果由值迭代算法产生的序列 $\{T^k\tilde{J}\}$ 对所有满足 $0 \leqslant \tilde{J} \leqslant J^*$ 的 $\tilde{J}$ 都收敛到 J^*（这一点在非常一般性的条件下成立；见 [Ber12]、[Ber22a]），那么 $T^{l-1}\tilde{J}$ 对于充分大的 l 属于稳定域。相关的想法已经在 Liu 及其合作者的自适应动态规划的文献 [HWL21]、[LXZ21]、[WLL16] 以及 Heydari[Hey17]、[Hey18] 中讨论过，他们提供了详细的参考文献；也见 Winnicki 等 [WLL21]。我们将在线性二次型问题中重新回到这一问题。这一断言一般是正确的，但是需要在 J^* 的某个邻域内的所有函数 $\tilde{J}$ 都属于稳定域。我们的后续讨论将旨在解决这一困难。

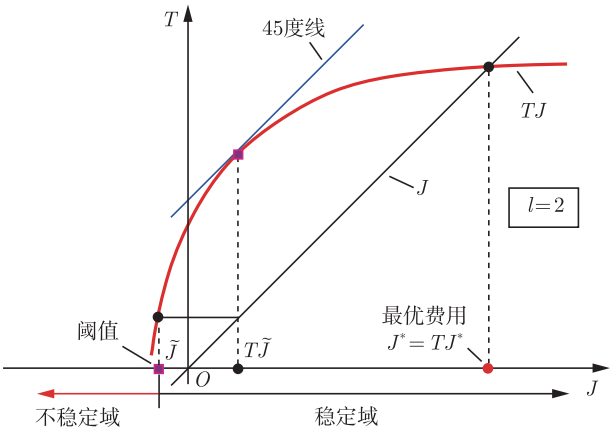

图 3.3.2　采用 l 步前瞻最小化的值空间近似的稳定域和不稳定域示意图。稳定域是从 $T^l\tilde{J} = T_{\tilde{\mu}}T^{l-1}\tilde{J}$ 中获得的，策略 $\tilde{\mu}$ 对所有 x 满足 $J_{\tilde{\mu}}(x) < \infty$ 的所有 $\tilde{J}$ 构成的集合（图中展示的是 $l = 2$ 的情形）。随着 l 的增加，不稳定域倾向于减小。

在我们的上下文中一条重要的事实是稳定域包括所有满足

$$T\tilde{J} \leqslant \tilde{J} \tag{3.10}$$

的实值非负函数 $\tilde{J}$。事实上如果 $\tilde{\mu}$ 是对应的单步前瞻策略，我们有

$$T_{\tilde{\mu}}\tilde{J} = T\tilde{J} \leqslant \tilde{J}$$

以及从非负费用无限时段问题的一个众所周知的结论 [见 [Ber12] 命题 4.1.4（a）]，于是有

$$J_{\tilde{\mu}} \leqslant \tilde{J}$$

（用一句话证明这一结论：如果有 $T_{\tilde{\mu}}\tilde{J} \leqslant \tilde{J}$，那么对所有的 k 有 $T_{\tilde{\mu}}^{k+1}\tilde{J} \leqslant T_{\tilde{\mu}}^{k}\tilde{J}$，再使用 $0 \leqslant \tilde{J}$ 的事实，于是 $T_{\tilde{\mu}}^{k}\tilde{J}$ 的极限，称为 J_∞，满足 $J_{\tilde{\mu}} \leqslant J_\infty \leqslant \tilde{J}$）。所以如果 $\tilde{J}$ 是非负且实值的，$J_{\tilde{\mu}}$ 也是实值的，所以 $\tilde{\mu}$ 是稳定的。于是有 $\tilde{J}$ 属于稳定域。在特定的背景下这是已知的结论，例如模型预测控制（见 Rawlings、Mayne 和 Diehl[RMD17] 一书的 2.4 节，其中包含了此前大量有关稳定性问题的参考文献）。

满足条件 $T\tilde{J} \leqslant \tilde{J}$ 的一个重要的特殊情形是，当 $\tilde{J}$ 是稳定策略的费用函数，即 $\tilde{J} = J_\mu$，那么我们有 J_μ 是实值的且满足 $T_\mu J_\mu = J_\mu$，所以有 $TJ_\mu \leqslant J_\mu$。这一情形关联到滚动算法并且展示了采用稳定策略的滚动获得稳定的前瞻策略。这也意味着如果 μ 是稳定的，那么对于充分大的 m，$T_\mu^m \tilde{J}$ 属于稳定域。

除了与稳定的 μ 对应的 J_μ 和 J^*，存在其他有趣的函数 $\tilde{J}$ 满足条件 $T\tilde{J} \leqslant \tilde{J}$。特别地，令 β 为满足 $\beta > 1$ 的标量，且对于稳定的策略 μ，考虑如下定义的 β 放大算子 $T_{\mu,\beta}$

$$(T_{\mu,\beta}J)(x) = E\left\{\beta g\left(x, \mu(x), w\right) + \alpha J\left(f(x, \mu(x), w)\right)\right\}, \forall x$$

于是可以看出函数

$$J_{\mu,\beta} = \beta J_\mu$$

是 $T_{\mu,\beta}$ 的一个不动点且满足 $TJ_{\mu,\beta} \leqslant J_{\mu,\beta}$。这通过写出

$$J_{\mu,\beta} = T_{\mu,\beta}J_{\mu,\beta} \geqslant T_\mu J_{\mu,\beta} \geqslant TJ_{\mu,\beta} \tag{3.11}$$

可以得出。所以 $J_{\mu,\beta}$ 位于稳定域之中，且位于 J_μ 的“更靠其右侧的位置”。所以我们可以推测在 m 步截断滚动的上下文中 $T_{\mu,\beta}^m \tilde{J}$ 可以比 $T_\mu^m \tilde{J}$ 更可靠地近似 J_μ。

为了说明这一事实，考虑稳定策略 μ，并假设除去终止状态 t（如果存在）之外的所有状态下的每阶段期望费用与 0 隔离开，即

$$C = \min_{x \neq t} E\left\{g\left(x, \mu(x), w\right)\right\} > 0$$

那么我们宣称，给定标量 $\beta > 1$，对于任意满足 $\hat{J}(t) = 0$ 的函数 $\hat{J} \in R(X)$，如果 $\hat{J}$ 满足

$$\max_x |\hat{J}(x) - J_{\mu,\beta}(x)| \leqslant \delta, \forall x \tag{3.12}$$

其中

$$\delta = \frac{(\beta - 1)C}{1 + \alpha}$$

那么 $\hat{J}$ 也满足稳定性条件 $T\hat{J} \leqslant \hat{J}$。从这一点于是有对于给定的非负实值 $\tilde{J}$ 和让函数 $\hat{J} = T_{\mu,\beta}^m \tilde{J}$ 满足式 (3.12) 的充分大的 m，有 $\hat{J}$ 位于稳定域中。

为了理解这一点，注意对所有的 $x \neq t$，我们有

$$J_{\mu,\beta}(x) = \beta E\left\{g\left(x, \mu(x), w\right)\right\} + \alpha E\left\{J_{\mu,\beta}\left(f\left(x, \mu(x), w\right)\right)\right\}$$

于是通过使用式 (3.12)，我们有

$$\hat{J}(x) + \delta \geqslant \beta E\left\{g\left(x, \mu(x), w\right)\right\} + \alpha E\left\{\hat{J}\left(f\left(x, \mu(x), w\right)\right)\right\} - \alpha\delta$$

于是有

$$\begin{aligned}\hat{J}(x) &\geqslant E\left\{g\left(x,\mu(x),w\right)\right\}+\alpha E\left\{\hat{J}\left(f(x,\mu(x),w)\right)\right\}+(\beta-1)\,E\left\{g\left(x,\mu(x),w\right)\right\}-(1+\alpha)\delta\\ &\geqslant E\left\{g\left(x,\mu(x),w\right)\right\}+\alpha E\left\{\hat{J}\left(f\left(x,\mu(x),w\right)\right)\right\}+(\beta-1)C-(1+\alpha)\delta\\ &=(T_\mu\hat{J})(x)\\ &\geqslant (T\hat{J})(x)\end{aligned}$$

于是稳定性条件 $T\hat{J}\leqslant\hat{J}$ 被满足。

类似地，如下函数

$$J_\beta^*=\beta J^*$$

是如下定义的算子 T_β 的不动点

$$\left(T_\beta J\right)(x)=\min_{u\in U(x)}E\left\{\beta g(x,u,w)+\alpha J\left(f(x,u,w)\right)\right\},\forall x$$

可以看出，使用与式 (3.11) 类似的论述，J_β^* 满足 $TJ_\beta^*\leqslant J_\beta^*$，所以它位于稳定域之内。进一步，与之前讨论的截断滚动的情形类似，我们可以断言比起在 l 步前瞻上下文中 J^* 由 $T^{l-1}\tilde{J}$ 替代的情形，J_β^* 可以用 $T_\beta^{l-1}\tilde{J}$ 更稳定地近似。

3.4 策略迭代、滚动和牛顿法

无限时段算法的另一个主要类别基于策略迭代（简称为 PI），这涉及重复使用策略改进，与在第 1 章中所描述的在阿尔法零和时序差分西洋双陆棋离线训练算法类似。PI 算法的每轮迭代从稳定策略开始（我们称之为当前或者基础策略），并生成另一个稳定策略（我们将分别称之为新的或者滚动的策略）。对于 2.1 节的无限时段问题，给定基础策略 μ，迭代包括两个阶段（见图 3.4.1）。

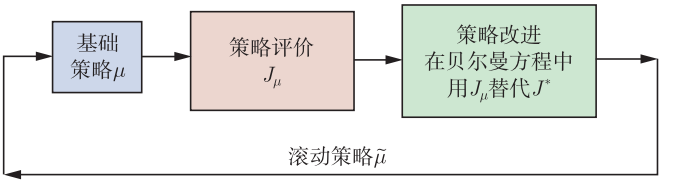

图 3.4.1 PI 作为重复滚动的示意图。它生成了一系列策略，序列中的每个策略 μ 作为基础策略生成序列中的下一个策略 $\tilde{\mu}$ 作为对应的滚动策略。

（a）策略评价，计算费用函数 J_μ。一种可能性是求解对应的贝尔曼方程

$$J_\mu(x)=E\left\{g\left(x,\mu(x),w\right)+\alpha J_\mu\left(f\left(x,\mu(x),w\right)\right)\right\},\forall x \tag{3.13}$$

然而，也可以通过蒙特卡洛仿真来计算与任意 x 对应的 $J_\mu(x)$ 值，即通过在许多从 x 开始由策略随机生成的轨迹上取平均。其他更复杂的可能性包括使用定制的基于仿真的方法，例如时序差分方法，对此存在大量的文献（如 [BeT96]、[SuB98]、[Ber12] 等书）。

（b）策略改进，使用单步前瞻最小化计算滚动策略 $\tilde{\mu}$

$$\tilde{\mu}(x) \in \arg\min_{u \in U(x)} E\left\{g(x,u,w) + \alpha J_\mu\left(f(x,u,w)\right)\right\}, \forall x \tag{3.14}$$

通常期待（而且可以在宽松的条件下证明）滚动策略改进了，即对所有的 x 有

$$J_{\tilde{\mu}} \leqslant J_\mu(x)$$

可以在大部分动态规划书籍中找到在多种语境中对这一事实的证明，包括作者的 [Ber12]、[Ber18a]、[Ber19a]、[Ber20a]、[Ber22a]。

所以 PI 生成一系列的稳定策略 $\{\mu^k\}$，通过使用式 (3.13) 对之前的策略 μ^k 进行策略评价获得 J_{μ^k}。通过在式 (3.14) 中使用 J_{μ^k} 替代 J_μ，并经过策略改进运算获得 μ^{k+1}。众所周知（精确的）策略迭代具有坚实的收敛性质；见之前引用的动态规划教材，以及作者的强化学习一书 [Ber19a]。即使当该方法在涉及异步分布式计算的非传统计算环境中实现时（经过恰当的改造）策略迭代的收敛性质也成立，正如由 Bertsekas 和 Yu 发表的一系列论文 [BeY10]、[BeY12]、[YuB13] 中所示。

使用我们的抽象符号，PI 算法可以写成紧凑的形式。对于所生成的策略序列 $\{\mu^k\}$，策略评价阶段通过如下方程获得 J_{μ^k}

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k} \tag{3.15}$$

而策略改进阶段通过如下方程获得 μ^{k+1}

$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k} \tag{3.16}$$

正如图 3.4.2 所示，PI 可以被视作在费用函数 J 的函数空间中求解贝尔曼方程的牛顿法。特别地，式 (3.16) 的策略改进是从 J_{μ^k} 开始的牛顿步，并且获得 μ^{k+1} 作为对应的单步前瞻（滚动策略）。图 3.4.3 展示了滚动算法，这只是 PI 的首轮迭代。

与值空间的近似相比，策略迭代基于牛顿法的解释有较长的历史。我们推荐 Kleiman 对线性二次型问题的最初工作 [Klei68][①]，以及由 Pollatschek 和 Avi-Itzhak 关于有限状态无限时段折扣和马尔可夫博弈问题的工作 [PoA69]（他们也证明了该方法可能在博弈情形下振荡）。后续的工作，讨论了算法变形和近似，包括 Hewer[Hew71]，Puterman 和 Brumelle[PuB78] 和 [PuB79]，Santos 和 Rust[SaR04]，Bokanowski、Maroso 和 Zidani [BMZ09]，Hylla[Hyl11]，Magirou、Vassalos 和 Barakitis[MVB20]，Bertsekas[Ber21c]，Kundu 和 Kunitsch[KuK21]。这些论文中的一些处理了更广类型的问题（例如连续时间最优控制、极小化极大问题和马尔可夫博弈），并包括了在多种（经常受限制的）假设下的超线性收敛速率的结论，以及策略迭代的变形。早期与控制系统设计相关的工作包括 Saridis 和 Lee[SaL79]、Beard[Bea95] 以及 Beard、Saridis 和 Wen[BSW99]。

① 这是 Kleinman 在 MIT 的博士论文 [Kle67]，由 M. Athans 指导。Kleinman 将一维版本的结论归功于 Bellman 和 Kalaba[BeK65]。还要注意策略迭代方法由 Bellman 在他的经典书 [Bel57] 中首次给出，并使用了“策略空间的近似”这一名称。

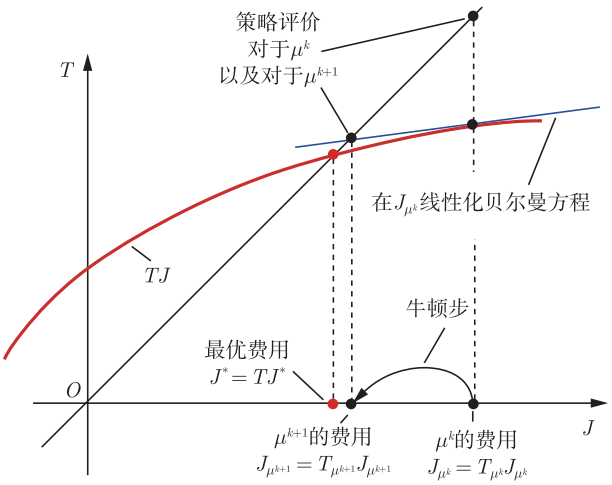

图 3.4.2　策略迭代的几何解释。从稳定的当前策略 μ^k 出发，策略迭代评价对应的费用函数 J_{μ^k} 并按照

$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}$$

计算下一个策略 μ^{k+1}。对应的费用函数 $J_{\mu^{k+1}}$ 是线性方程 $J = T_{\mu^{k+1}} J$ 的解，于是这是从 J_{μ^k} 出发求解贝尔曼方程 $J = TJ$ 的牛顿步的结果。注意在策略迭代中，牛顿步总是从一个函数 J_μ 开始，该函数满足 $J_\mu \geqslant J^*$ 以及 $TJ_\mu \leqslant J_\mu$（参阅 3.3 节中关于稳定性的讨论）。

滚动

一般说来，采用稳定的基础策略 μ 的滚动可以被视作从贝尔曼方程的解 J_μ 开始的牛顿法的单轮迭代（见图 3.4.3）。注意在系统的实时操作中滚动、策略改进仅施加在当前的状态之上。这让在线实现变成可能，即使问题的状态空间非常大，只要可以按照需要在线地评价基础策略的性能即可。为此，我们经常需要对系统实时产生的每一个状态 x_k 进行在线的确定性或者随机的仿真。

正如图 3.4.3 所示，滚动策略 $J_{\tilde{\mu}}$ 的费用函数通过构建贝尔曼方程在 J_μ 的线性化近似并求解来获得。如果函数 TJ 是接近线性的（即有小的“曲率”），即使基础策略 μ 与最优解相去甚远，滚动策略性能 $J_{\tilde{\mu}}(x)$ 与最优的 $J^*(x)$ 仍然非常接近。这解释了为何在实际中采用滚动算法时通常能观察到大的费用改进。

一个有趣的问题是如何对于给定的初始状态 x 将滚动性能 $J_{\tilde{\mu}}(x)$ 与基础策略性能 $J_\mu(x)$ 进行比较。显然，我们希望 $J_\mu(x) - J_{\tilde{\mu}}(x)$ 很大，但这不是看待费用改进的正确方法。原因是如果其上界 $J_\mu(x) - J^*(x)$ 小，即，如果基础策略接近最优，那么 $J_\mu(x) - J_{\tilde{\mu}}(x)$ 也小。因此，更重要的是误差率

$$\frac{J_{\tilde{\mu}}(x) - J^*(x)}{J_\mu(x) - J^*(x)} \tag{3.17}$$

小。确实，由于牛顿法的超线性收敛速率，上述误差率随着 $J_\mu(x) - J^*(x)$ 接近 0 而变得更小（参见图 3.4.3）。可是，因为不知道 $J^*(x)$，所以评价这一比例是困难的。另一方面，如果我们观察到小的性能改进 $J_\mu(x) - J_{\tilde{\mu}}(x)$，那么也不应不知所措，原因可能是基础策略已经近优，事实上可能按照式 (3.17) 的误差率来说做得相当好。

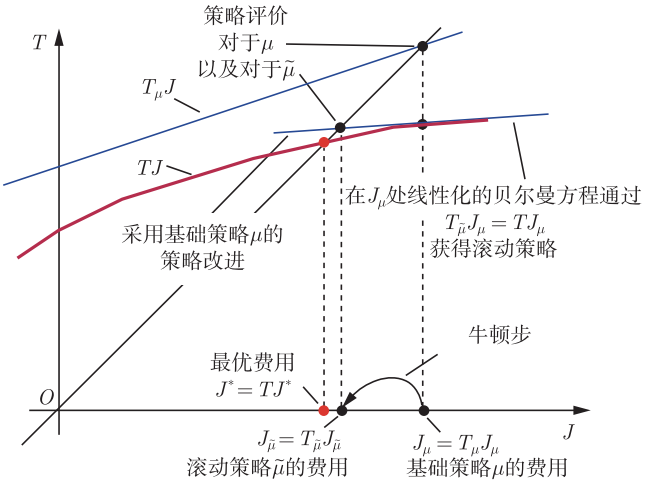

图 3.4.3　滚动的几何解释。每个策略 μ 定义了由式 (3.2) 给出的 J 的线性函数 $T_\mu J$，TJ 是由式 (3.1) 给出的函数，也可写成 $TJ = \min_\mu T_\mu J$。本图展示了从基础策略 μ 开始的一步策略迭代。由策略评价（通过如图所示求解线性方程 $J = T_\mu J$）计算 J_μ。然后使用 μ 作为基础策略进行一步策略改进产生滚动策略 $\tilde{\mu}$，如图所示，通过求解在点 J_μ 线性化版本的贝尔曼方程获得滚动策略的费用函数 $J_{\tilde{\mu}}$，正如在牛顿法中那样。

截断滚动和乐观策略迭代

滚动的变形可能涉及多步前瞻、截断和末端费用函数近似，正如在阿尔法零和时序差分西洋双陆棋程序中那样，参见第 1 章。这些变形的几何解释与之前所给的类似，见图 3.4.4。截断滚动采用基础策略 μ 进行 m 次值迭代，并使用近似末端费用函数 $\tilde{J}$ 来近似费用函数 J_μ。

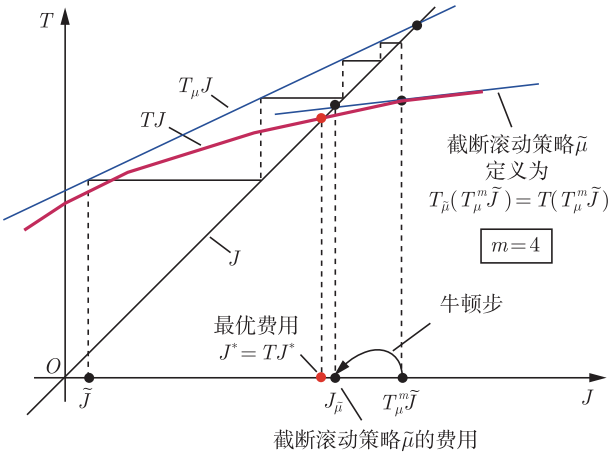

图 3.4.4　采用单步前瞻最小化、使用基础策略 μ 和末端费用函数近似 $\tilde{J}$ 的 m 次值迭代（这里 $m = 4$）的截断滚动的几何解释。

在单步前瞻的情形中，截断滚动策略 $\tilde{\mu}$ 定义为

$$T_{\tilde{\mu}}(T_\mu^m \tilde{J}) = T(T_\mu^m \tilde{J}) \tag{3.18}$$

即，当贝尔曼算子 T 施加到函数 $T_{\mu}^{m}\tilde{J}$（通过用基础策略 μ 进行 m 步之后用近似末端费用 $\tilde{J}$ 获得的费用）上时 $\tilde{\mu}$ 达到最小值；见图 3.4.4。在 l 步前瞻的情形中，截断滚动策略 $\tilde{\mu}$ 定义为

$$T_{\tilde{\mu}}\left(T^{l-1}T_{\mu}^{m}\tilde{J}\right)=T(T^{l-1}T_{\mu}^{m}\tilde{J}) \tag{3.19}$$

截断滚动与策略迭代的一种乐观的变形有关。这一变形通过使用基础策略 μ 进行 m 次值迭代近似策略评价步骤；见 [BeT96]、[Ber12]、[Ber19a] 中关于这一关系的更详细的讨论。与乐观策略迭代有关的方法是 λ 策略迭代方法，这与凸分析的近似算法有关，并在作者的其他几本书（[BeT96]、[Ber12]、[Ber20a]、[Ber22a]）和论文（[BeI96]、[Ber15]、[Ber18d]）中进行了讨论，且也可以用于替换式 (3.18) 定义单步前瞻策略。特别地，论文 [Ber18d] 的第 6 节集中关注 λ 策略迭代算法，作为对有限状态折扣和随机最短路问题的常规策略迭代和牛顿法的近似。

正如之前注释的，在每次牛顿步骤前后进行多步不动点迭代且不使用截断滚动，即

$$T_{\tilde{\mu}}(T^{l-1}\tilde{J})=T(T^{l-1}\tilde{J}) \tag{3.20}$$

这样的牛顿法变形众所周知。Ortega 和 Rheinboldt 的经典数值分析一书 [OrR70]（13.3 节和 13.4 节）提供了多种收敛性结论，所用的假设条件包括 T 的元素的可微性和凸性，以及 T 的逆雅可比的非负性。这些假设条件，特别是可微性，可能在我们的动态规划上下文中不满足。进一步，对式 (3.20) 形式的方法，初始点必须满足额外的假设条件，其保证了 J^{*} 的凸性是上单调的（在这一情形下，如果还有 T 的雅可比是保序的，则可以构造一条辅助序列由下单调地收敛到 J^{*}），见 [OrR70]、13.3.4 节、13.4.2 节。这与动态规划中乐观策略迭代方法的已有收敛结果类似，见 [BeT96]、[Ber12]。

如图 3.4.4 所示的几何解释还建议：

（a）从基础到滚动策略的费用改进 $J_{\mu}-J_{\tilde{\mu}}$ 倾向于随着前瞻长度 l 增加变得更大；

（b）采用 l 步前瞻最小化，之后是 m 步的基础策略 μ，再然后是末端费用函数近似 $\tilde{J}$ 的截断滚动可以在适当的条件下被视作使用 $\tilde{J}$ 作为末端费用函数近似的 $(l+m)$ 步前瞻最小化的经济的替代。

图 3.4.5 总结解释了采用 l 步前瞻最小化和 m 步截断滚动 [参见式 (3.19)] 的值空间近似机制及其与牛顿法的关联。本图标出的部分通常与在线对弈和离线训练相关，并且与此前的图 1.2.1 并列，后者适用于阿尔法零、时序差分西洋双陆棋程序及相关的在线机制。

截断滚动中的前瞻长度问题

实际中感兴趣的一个问题是如何在截断滚动机制中选择前瞻长度 l 和 m。显然前瞻最小化中 l 取值大是有好处的（在产生改进的前瞻策略费用函数 $J_{\tilde{\mu}}$ 的意义下），因为额外的值迭代让牛顿步的起点 $T^{l-1}\tilde{J}$ 更接近 J^{*}。然而请注意尽管长程前瞻最小化的计算量大（其复杂性随着 l 增加按指数速度增大），多步前瞻中仅有第一阶段对牛顿步有贡献，而剩余的 $l-1$ 步是效果差许多的一阶值迭代。

关于 m 的取值，长程截断滚动让牛顿步的起点更接近 J_{μ}，但未必更接近 J^{*}，正如图 3.4.4 所示。确实在计算实验中看到增加 m 的取值在超过某个阈值后可能起反作用，同

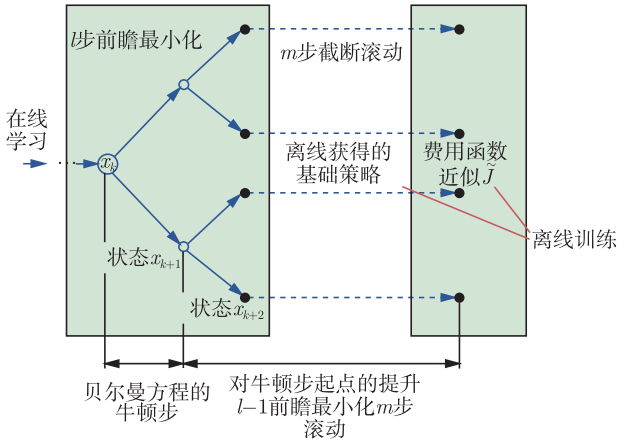

图 3.4.5 采用 l 步前瞻最小化（图中 $l=2$）和 m 步截断滚动 [参见式 (3.19)] 的值空间近似机制及其与牛顿法关联的示意图。求解贝尔曼方程 $J^*=TJ^*$ 的牛顿步对应于（l 步）前瞻最小化中的第一步。剩余的 $l-1$ 步前瞻最小化（值迭代）及 m 步截断滚动（采用基础策略的值迭代），从其离线获得的费用函数近似 $\tilde{J}$ 基础上改进了牛顿步的起点。

时看到这一阈值通常依赖于问题和末端费用近似 $\tilde{J}$；也见 4.6 节中我们对线性二次型问题的后续讨论。正如之前注释的，这也与长期以来对乐观策略迭代的经验一致，后者与截断滚动紧密相关。然而，遗憾的是，暂时没有分析可以解释这一问题，截断滚动的可用的误差界（见 [Ber19a]、[Ber20a]）是保守的且对这个问题只能提供有限的指导。

另一个需要记住的重要事实是截断滚动步骤比前瞻最小化步骤需要更少的计算量。所以与其他顾虑平等考虑之后，在计算上倾向于让 m 取值大而不是让 l 取值大（这是 Tesauro 的时序差分西洋双陆棋程序的截断滚动的启示 [TeG96]）。另一方面，尽管 m 取值大可能在计算上是可以忍受的，但是即使是相对较小的 l 的取值在计算上也是相当困难的。对于前瞻树的宽度倾向于快速增长的随机问题尤其是这样。

一条有趣的性质，且有一定的普遍性，是采用稳定策略的截断滚动对前瞻策略的稳定性有好处。原因是基础策略 μ 的费用函数 J_μ 位于稳定域内部，正如 3.2 节指出的。进一步 μ 的值迭代（即截断滚动）倾向于将牛顿步骤的起点推向 J_μ。所以这些值迭代在充分次数之后将牛顿步骤的起始点推进稳定域。

前述讨论引出了如下的定性问题：基于滚动的前瞻是基于最小化的前瞻的经济的替代吗？对此问题的回答似乎应是一个合格的肯定：对于给定的计算资源，小心地权衡 m 和 l 的取值之后倾向于比简单地尽可能增加 l 而设定 $m=0$（这对应于没有滚动）获得更好的前瞻策略性能。这与通过例如图 3.4.4 所展示的几何解释获得的直观保持一致，但是难以建立结论。我们之后在 4.6 节中对于线性二次型问题进一步讨论这一点。

3.5 在线对弈对于离线训练过程有多敏感？

在值空间近似中需要考虑的一个重要问题是单步或者多步最小化中，或者末端费用近似 $\tilde{J}$ 中的误差的影响。因为控制约束集合 $U(x)$ 是无限的，或者因为计算的权宜之计简化了最小化过程（见我们后续对多智能体问题的讨论）上述误差经常是不可避免的。进一步，