پاسخنامه تمرین ۷ بخش ۲

سوال ۱)

policy iteration

زیرا در هر مرحله انتخاب‌ها را محدود می کند. در این روش تنها انتخاب‌هایی که مجاز هستند محاسبه می‌شوند ولی در روش دیگر تمامی حالت‌ها محاسبه می‌شوند.

سوال ۲)

با توجه به روش‌ها Q بهتر است. زیرا در روش value iteration در هر مرحله به اندازه $O(s^2*A)$ زمان می‌برد.در هر مرحله مقدار ماکزیمم مقدار اندکی تغییر می‌کند و شروط بسیار سریع تر از value ها همگرا می‌شوند.

در روش Q iteration فقط action ها مهم هستند و نه مقصد و به همین دلیل سریع تر است. همین طور با توجه به action ها می‌توان policy ها را توضیح داد.