Ex 13-1

(a) stochastic policy pros (example):

① rock-scissors-paper: deterministic policy is predictable.

② aliased grid world: environment is partially observable

value based: $Q_\theta(s,a) = f(\phi(s,a), \theta)$

policy based: $\pi_\theta(s,a) = g(\phi(s,a)), \theta)$

③ continuous action spaces.

(b)

$$\pi_\theta(a|s) = \frac{\exp\{x(s,a)^T\theta\}}{\sum_{a' \in A} \exp\{x(s,a')^T\theta\}}$$

$$\log \pi_\theta(a|s) = \log \exp\{x(s,a)^T\theta\} - \log \sum_{a' \in A} \exp\{x(s,a')^T\theta\}$$

$$= x(s,a)^T\theta - \log \sum_{a' \in A} \exp\{x(s,a')^T\theta\}$$

$$\nabla_\theta \log \pi_\theta(a|s) = x(s,a)^T - \nabla_\theta \log \sum_{a' \in A} \exp\{x(s,a')^T\theta\}$$

$$= x(s,a)^T - \frac{\sum_{a' \in A} \left(\exp\{x(s,a')^T\theta\}\right) x(s,a')}{\sum_{a' \in A} \exp\{x(s,a')^T\theta\}}$$

$$= x(s,a)^T - \sum_{a' \in A} \pi_\theta(a'|s) \cdot x(s,a')^T$$

$$= x(s,a)^T - \mathbb{E}_{\pi_\theta}[x(s,\cdot)]$$