Apprentissage actif par la méthode de recherche de plan/politique

- Toute les méthodes vues jusqu'à maintenant dérivent leur plan/politique à partir d'une fonction de valeur ou action-valeur
  - les méthodes diffèrent seulement dans la façon d'estimer ces fonctions

- Pourquoi ne pas optimiser directement par rapport au plan π de l'agent
  - soit s_0 l'état initial
  - **problème d'optimisation à résoudre**: trouver π dont la valeur $V(s_0)$ est la plus grande
  - on ne connaît pas $P(s'|s, \pi(s))$, donc on ne peut pas calculer $V(s_0)$ directement
  - par contre, chaque essai donne une estimation stochastique $V(s_0)$

- C'est ce qu'on appelle la **recherche de plan/politique** (*policy search*)

Apprentissage actif par la méthode de recherche de plan/politique

- Exemple de la grille 3 x 4 ($\gamma = 1$)
  - soit l'essai (simulation) obtenu en suivant π

$(1,1)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (2,3)_{-.04} \rightarrow (3,3)_{-.04} \rightarrow (4,3)_{+1}$

  - alors on sait que $V(s_0) \approx 7 \times -0.04 + 1 = 0.72$

- Approche de recherche de politique par *hill-climbing*
  - répéter durant T itérations
    1. $V(s_0) \leftarrow$ résultat de l'essai obtenu en suivant π
    2. pour chaque politique π' voisine de π (successeur)
       a. $v \leftarrow$ résultat de l'essai (simulation) obtenu en suivant π'
       b. si $v > V(s_0)$
          I. $V(s_0), \pi \leftarrow v, \pi'$

Apprentissage actif par la méthode de recherche de plan/politique

- Pourrait générer les successeurs de π' en considérant tous les changements possibles d'une seule action, pour un seul état
  - ex.:

$\pi(s_0) = a_1, \pi(s_1) = a_3$ → successeurs →

$\boldsymbol{\pi'(s_0) = a_2}, \pi'(s_1) = a_3$

ou

$\pi'(s_0) = a_1, \boldsymbol{\pi'(s_1) = a_2}$

- Peut bien fonctionner seulement si l'espace des états et d'actions est petit

- L'apprentissage sera lent s'il y a beaucoup de variations stochastiques possibles
  - ex.: nos chances de gagner au Poker dépendent beaucoup des cartes que l'on pige

Apprentissage actif par la méthode de recherche de plan/politique

- Si on peut **contrôler les variations d'une simulation à l'autre**, on peut accélérer l'apprentissage
- Si on a accès au générateur de nombres aléatoires du simulateur, on peut l'initialiser au même état avant chaque simulation
- Dans le cas d'un jeu de carte, les mêmes cartes seraient pigées dans le même ordre
- C'est l'idée derrière l'algorithme **PEGASUS**, utilisé pour apprendre des manoeuvres acrobatiques à l'aide d'un hélicoptère
  - voir http://heli.stanford.edu/ pour des vidéos de démonstration