E1 245 - Online Prediction and Learning, Aug-Dec 2014
Homework #4

1. *Stochastic Gradient Descent (6 points)*
Prove the following theorem. Suppose $f_t : \mathcal{K} \to \mathbb{R}$, $t = 1, 2, 3, \ldots$, is a sequence of convex, differentiable functions on the convex set $\mathcal{K} \subseteq \mathbb{R}^d$ with $0 \in \mathcal{K}$. Let $\eta > 0$, $w_1 := 0$, and[1] $w_{t+1} := \Pi_{\mathcal{K}}[w_t - \eta g_t]$, $t = 1, 2, 3, \ldots$ where g_t is a random variable satisfying $\mathbb{E}[g_t | w_t] = \nabla f_t(w_t)$ and $||g_t||_2 \leq G$ almost surely for some scalar constant G. Denote $D := \max_{x \in \mathcal{K}} ||x||_2$. Then, for any time horizon $T \geq 1$,
$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(w_t) - \min_{w \in \mathcal{K}} \sum_{t=1}^{T} f_t(w)\right] \leq \frac{\eta G^2 T}{2} + \frac{D^2}{2\eta}.$$

2. *Finite Time Planning (5 points)*
Consider a (time-homogeneous) Markov Decision Process (MDP) with two states and two actions, and finite time horizon N. Choose any non-trivial (non-zero and unequal) transition probabilities $\{T(s, a, s')\}_{s,a,s'}$ and rewards $\{R(s, a)\}_{s,a}$. Draw a state transition diagram for your model, write down explicitly the value iteration equation for this model, and compute the optimal value function and optimal policy for $N = 3$ (assuming zero terminal rewards).

3. *Risk-sensitive Control (15 points)*
Consider an MDP $(\mathcal{S}, \mathcal{A}, R, T)$, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, with the exponential (finite-time) reward objective $\max_{\pi \equiv (\pi_1, \ldots, \pi_N)} J^{\pi}_{\beta}(s)$. Here,
$$J^{\pi}_{\beta}(s) := \text{sign}(\beta) \cdot \mathbb{E}_{\pi}\left[\exp\left(\beta \sum_{k=1}^{N} r_k\right) \middle| s_1 = s\right],$$
where sign$(\cdot)$ is the sign function[2], the time horizon N is deterministic and known, and s_k, a_k and $r_k := R(s_k, a_k)$ are the state, action taken and reward obtained in round k respectively. This reward metric is called *risk-averse* or *risk-seeking* depending on the sign of β (i.e., -1 or $+1$).

   (a) *(3 points)* What is the optimal policy as $\beta \to 0$? (Hint: Use a Taylor series expansion.)

   (b) *(6 points)* Suggest a recursive planning algorithm that obtains the optimal value function $V^*_{\beta,k}(s)$, $1 \leq k \leq N$, $s \in \mathcal{S}$, and policy π^* for this problem. Express the optimal value function in terms of $v^*_{\beta,k} := \log V^*_k(s)$, and compare with the standard case.

   (c) *(6 points)* Explain what happens to the optimal policy for $\beta \to +\infty$ and $\beta \to -\infty$. Propose simple recursive algorithms for these two extreme regime cases of β.

[1] $\Pi_{\mathcal{K}}(\cdot)$ denotes projection with respect to the Euclidean norm onto $\mathcal{K}$.
[2] The sign of 0 is arbitrarily defined to be 0.

4. *Programming Exercise: Planning algorithms (30 points)*
   Generate a non-trivial (i.e., non-zero, non-1 transition probabilities) MDP randomly, using any reasonable scheme of your choice, with 10 states and 5 actions. Choose a discount factor $\gamma \in (0, 1)$, and find the optimal infinite-horizon discounted policy using both (a) Value iteration run until a suitably small convergence threshold (say 10^{-6}) and (b) Policy iteration. Record the number of value/policy iterations in (a) and (b), the per-iteration CPU time, and the total running time.

   Repeat this exercise for various generated MDPs and for a sequence of γ values gradually approaching 1. What happens to the performance of both these algorithms? Is one better than the other in practice?

   (Resource: If you use MATLAB, you might find the following MDP algorithms package convenient: `http://www7.inra.fr/mia/T/MDPtoolbox/Documentation.html`)