Mehryar Mohri
Advanced Machine Learning 2015
Courant Institute of Mathematical Sciences
Homework assignment 2
April 13, 2015
Due: April 27, 2015

A. RWM and FPL

Let RWM(β) denote the RWM algorithm described in class run with parameter $\beta > 0$. Consider the version of the FPL algorithm FPL(β) defined using the perturbation:

$$\mathsf{p}_1 = \left[\frac{\log(-\log(u_1))}{\beta}, \ldots, \frac{\log(-\log(u_N))}{\beta}\right]^\top.$$

where, for $j \in [1, N]$, u_j is drawn from the uniform distribution over $[0, 1]$. At round $t \in [1, T]$, $\mathbf{w}_t$ is found via $\mathbf{w}_t = M(\mathbf{x}_{1:t-1} + \mathsf{p}_1) = \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \mathbf{w} \cdot (\mathbf{x}_{1:t-1} + \mathsf{p}_1)$ using the notation adopted in the class lecture for FPL, with $\mathcal{W}$ the set of coordinate vectors. Show that FPL(β) coincides with RWM(β).

Solution: Let $\mathsf{p}_{1i} = \frac{\log(-\log(u_i))}{\beta}$ denote the i-th coordinate of vector p_1. The distribution of p_{1i} is given by

$$\Pr(\mathsf{p}_{1i} \geq x) = \Pr(u_i \leq e^{-e^{\beta x}}) = e^{-e^{\beta x}},$$

where we used the fact that u_i is a uniform random variable. Fix t and let $L_i = \sum_{k=1}^{t-1} x_{ik}$ be the cumulative loss of coordinate i and $\widetilde{L}_i = L_i + \mathsf{p}_{1i}$. Since the set $\mathcal{W}$ consists of coordinate vectors , the minimization problem is equivalent to finding i^* such that $i^* = \operatorname{argmin}_i \widetilde{L}_i$. Let

$$F_i(x) = \Pr(\widetilde{L}_i \leq x) = 1 - e^{-e^{\beta(x - L_i)}} \qquad \text{and}$$
$$f_i(x) = \beta e^{\beta(x - L_i)} e^{-e^{\beta(x - L_i)}}$$

be the cumulative distribution function and density function respectively for the random variable $\widetilde{L}_i$ and let $G_i(x) = 1 - F_i(x)$. If $p_i = \Pr(i^* = i)$, then

$$
\begin{aligned}
p_i = \Pr(\widetilde{L}_i = \min_j \widetilde{L}_j) &= \Pr(\widetilde{L}_i \leq \widetilde{L}_j \ \forall\ j \neq i) \\
&= \operatorname*{E}_{\widetilde{L}_i} \big[\Pr(\widetilde{L}_i \leq \widetilde{L}_j \ \forall\ j \neq i | \widetilde{L}_i) \big] \\
&= \operatorname*{E}_{\widetilde{L}_i} \Big[\prod_{i \neq j} G_j(\widetilde{L}_i) \Big] \\
&= \int_{-\infty}^{\infty} f_i(x) \prod_{i \neq j} G_j(x) dx.
\end{aligned}
$$

Using the definition of f_i and G_j we see that the above expression is given by

$$
\begin{aligned}
\int_{-\infty}^{\infty} \beta e^{\beta(x - L_i)} \prod_{j=1}^{n} e^{-e^{\beta(x - L_j)}} &= \int_{-\infty}^{\infty} e^{-\beta L_i} \beta e^{\beta x} e^{-e^{\beta x} \sum_{j=1}^{n} e^{-\beta L_j}} \\
&= \frac{e^{-\beta L_i}}{\sum_{j=1}^{n} e^{-\beta L_j}}.
\end{aligned}
$$

Therefore, the probability of choosing coordinate i is the same as the one given by RWM(β).

B. Zero-sum games

For all the questions that follow, we consider a zero-sum game with payoffs in $[0, 1]$.

1. Show that the time complexity of the RWM algorithm to determine an ϵ-approximation of the value of the game is in $O(\log N/\epsilon^2)$.

   Solution: Let $\mathbf{p}_t$ be the probability vectors associated with RWM. From the proof of von Neumman's theorem we know that the mixed strategy $\mathbf{p}_{\text{RWM}} = \frac{1}{T} \sum \mathbf{p}_t$ satisfies:

   $$\max_{\mathbf{q} \in \Delta_N} \mathbf{p}_{\text{RWM}} \mathbf{M} \mathbf{q} \leq \min_{\mathbf{p} \in \Delta_N} \max_{\mathbf{q} \in \Delta_N} \mathbf{p}^\top \mathbf{M} \mathbf{q} + \frac{R_T}{T}.$$

   Furthermore, we know that the regret of RWM is in $O(\sqrt{\log N}/\sqrt{T})$. Therefore, after only $O(\log N/\epsilon^2)$ iterations of RWM we can obtain an ϵ-approximation of the value of the game.

2. Use the proof given in class for von Neumann's theorem to show that both players can come up with a strategy achieving and ϵ-approximation of the value of the game (or Nash equilibrium) that are sparse: the support of each mixed strategy is in $O(\log N/\epsilon^2)$. What fraction of the payoff matrix does it suffice to consider to compute these strategies?

   Solution: We let the strategy used by the row player be given by the RWM algorithm and denote by p_t be the probabilities used by this algorithm. The column player, plays the best response strategy. That is, given p_t he selects $q_t = \operatorname{argmax}_q p_t^\top Mq$. By von Neumann's minimax theorem we know that the value v of the game satisfies

$$v \leq \min_p p^\top M\Big(\frac{1}{T}\sum_{t=1}^{T} q_t\Big) \leq v + \frac{R_T}{T}.$$

   Since R_T is in $O(\sqrt{\log NT})$, then the strategies $q^* = \frac{1}{T}\sum_{t=1}^{T} q_t$ and $p^* = \operatorname{argmin}_p p^\top Mq^*$ form an ϵ-approximation to an equilibrium. Furthermore, q_t has only one non-zero entry, therefore the vector q^* can have at most $O(\log N/\epsilon^2)$ non-zero entries.

   By definition of the RWM algorithm, at every time t, to find p_t we need to evaluate $\sum_{s=1}^{t} p_s^\top Mq_s$. Moreover, since q_s has only one non-zero entry, it follows that $p_s^\top Mq_s$ can be calculated using only N entries of the matrix. Therefore, to calculate all vectors p_t we need only to inspect $O(N\log N/\epsilon^2)$ entries of the matrix M.

C. Bregman divergence

1. Given an open convex set C, provide necessary and sufficient conditions for a differentiable function $G\colon C \to \mathbb{R}$ to be a Bregman divergence. That is, give conditions for the existence of a convex function $F\colon C \to \mathbb{R}$ such that $G(x,y) = F(x) - F(y) - \nabla F(y)(x-y)$.
   Hint: Show that a Bregman divergence satisfies the following identity

$$B_F(x||y) + B_F(y||z) = B_F(x||z) + \langle x-y, \nabla F(z) - \nabla F(y)\rangle.$$

   Solution: A function $G(x,y)$ is a Bregman divergence if and only if

   - $G(x,y) + G(y,z) = G(x,z) + \langle x-y, \nabla_z G(z,y)\rangle$
   - $\nabla_z G(z,y) = \nabla F(z) - \nabla F(y)$,

where F is a convex function. Furthermore, $G(z, y) = B_F(z||y)$. Notice that these properties can be easily verified for any function G. We first prove the necessity of these conditions. The second bullet point follows directly from the definition of Bregman divergence. To prove the first bullet point, let $G(x, y) = B_F(x||y)$, then

$$\begin{aligned}
&G(x, y) + G(y, z) \\
&= F(x) - F(y) - \langle \nabla F(y), x - y\rangle + F(y) - F(z) - \langle \nabla F(z), y - z\rangle \\
&= F(x) - F(z) - \langle \nabla F(z), x - z\rangle + \langle \nabla F(z), x - y\rangle - \langle \nabla F(y), x - y\rangle \\
&= G(x, z) + \langle x - y, \nabla F(z) - \nabla F(y)\rangle \\
&= G(x, z) + \langle x - y, \nabla_z G(z, y)\rangle.
\end{aligned}$$

In order to show sufficiency, notice that the second condition implies $G(z, y) = F(z) - \langle \nabla F(y), z\rangle + g(y)$ for some function g. By plugging this into the equation defining the first condition we get

$$\begin{aligned}
&G(x, y) + G(y, z) - G(x, z) \\
&= F(x) - \langle \nabla F(y), x\rangle + g(y) + F(y) - \langle \nabla F(z), y\rangle + g(z) - F(x) + \langle \nabla F(z), x\rangle - g(z) \\
&= \langle \nabla F(z), x - y\rangle - \langle \nabla F(y), x\rangle + F(y) + g(y).
\end{aligned}$$

In order for G to satisfy the first condition, the following equality must then hold:

$$\begin{aligned}
g(y) + F(y) - \langle \nabla F(y), x\rangle &= -\langle x - y, \nabla F(y)\rangle \\
\Rightarrow \qquad g(y) &= -F(y) + \langle \nabla F(y), y\rangle
\end{aligned}$$

Replacing this expression in the equation defining G gives $G(z, y) = F(z) - F(y) - \langle \nabla F(y), z\rangle + \langle \nabla F(y), y\rangle = B_F(z||y)$.

2. Using the results of the previous exercise, decide whether or not the following functions are a Bregman divergence.

   - The KL-divergence: the function $G\colon \mathbb{R}^n_+ \times \mathbb{R}^n_+ \to \mathbb{R}$ defined for $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ by $G(x, y) = \sum_{i=1}^n x_i \log\left(\frac{x_i}{y_i}\right)$.
   - The function $G\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ given by $G(x, y) = x(e^x - e^y) - ye^y(x - y)$.

Solution:

- By definition we have $\frac{\partial}{\partial z_i} G(z, y) = \log z_i - \log y_i + 1$. Therefore, there exists no function F such that $\nabla_z G(z, y) = \nabla F(z) - \nabla F(y)$ and G cannot be a Bregman divergence.

- By definition we have $\frac{\partial}{\partial z} G(z, y) = e^z + z e^z - e^y - y e^y = F'(z) - F'(y)$, where $F(z) = z e^z$. Furthermore,

$$\begin{aligned} F(x) - F(y) - F'(y)(x - y) &= x e^x - y e^y - (y e^y + e^y)(x - y) \\ &= x e^x - x y e^y - x e^y + y^2 e^y \\ &= x(e^x - e^y) - y e^y (x - y) = G(x, y). \end{aligned}$$

  Therefore G is a Bregman divergence.