COMPSCI590.02 Algorithmic Aspects of Machine Learning
Assignment 3

Due Date: November 16, 2015 in class.

Problem 1 (Stochastic Gradient Descent)**.** In this problem we will try to analyze stochastic gradient descent algorithm for strongly convex functions.

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a L-smooth, μ-strongly convex function with optimal point at x^*. In particular

$$\langle \nabla f(x), x - x^* \rangle \geq \frac{\mu}{2} \|x - x^*\|_2^2 + \frac{1}{2L} \|\nabla f(x)\|_2^2.$$

We will try to optimize this function by running a stochastic gradient descent algorithm:

Algorithm 1 Stochastic Gradient Descent

for $t = 0$ **to** $k - 1$ **do**
  $x^{(t+1)} = x^{(t)} - \eta_t(\nabla f(x^{(t)}) + \epsilon_t)$.
end for

In the algorithm, η_t is a step size that we will choose later. The vector $\nabla f(x^{(t)}) + \epsilon_t$ is a *stochastic gradient* for f at $x^{(t)}$, in particular, ϵ_t is a random variable that only depends on $x^{(t)}$, and for every x

$$\mathbb{E}[\epsilon | x] = 0, \mathbb{E}[\|\epsilon\|_2^2 | x] \leq \sigma^2. \tag{1}$$

(a) (5 points) Let $r_t = \mathbb{E}[\|x^{(t)} - x^*\|_2^2]$, show that when $\eta \leq \frac{1}{L}$,

$$r_{t+1} \leq (1 - \eta\mu)\, r_t + \eta^2 \sigma^2.$$

(Hint: Consider $r_{t+1} = \mathbb{E}[\|(x^{(t)} - x_*) - \eta(\nabla f(x^{(t)}) + \epsilon_t)\|_2^2]$, and expand out the square.)

(b) (5 points) Show that when $r_t \geq \frac{2\sigma^2}{\mu L}$, we can choose $\eta_t = \frac{1}{L}$, and get $r_{t+1} \leq (1 - \frac{\mu}{2L}) r_t$.

(c) (10 points) Suppose $r_{t_0} = \frac{4\sigma^2}{\mu^2 k}$ for some integer k, and $k \geq \frac{2L}{\mu}$. Show that we can choose η_t appropriately to ensure $r_{t_0+t} \leq \frac{4\sigma^2}{\mu^2(k+t)}$ for all integer $t > 0$.
(Hint: The bound in (b) is quadratic in η, optimize that to get a good choice of step size.)

Problem 2 (Saddle Points)**.** We would like to find a tensor decomposition via optimization. Consider an orthogonal tensor

$$T = \sum_{i=1}^{n} u_i \otimes u_i \otimes u_i \otimes u_i.$$

Here $\{u_i\}$'s are orthonormal vectors. We would like to maximize

$$T(x,x,x,x) - \|x\|^6 = \sum_{i=1}^{n} \langle x, u_i \rangle^4 - \|x\|^6.$$

For simplicity, we can express x in the basis of $\{u_i\}$'s. Let $y_i = \langle x_i, u_i \rangle$, then the maximization problem becomes

$$\max f(y) = \sum_{i=1}^{n} y_i^4 - \|y\|^6.$$

Our goal now is to prove all local maxima of this function corresponds to directions $y = \pm e_i$ (where e_i is the i-th basis vector). As a result $x = \pm u_i$.

(a) (10 points) For any basis direction e_i, show that there are exactly two local maxima for $f(y)$ along this direction (one in the $+e_i$ direction and the other in the $-e_i$ direction).

(b) (10 points) For any subset $S \subseteq \{1, 2, ..., n\}$, if $1 < |S| \leq n$, show that there is a saddle point in the direction 1_S. Here 1_S is the indicator vector for S: $1_S[i] = 1$ if and only if $i \in S$.
(Hint: For both (a) (b) you can look at the gradient and Hessian, and apply second order optimality condition.)