Machine Learning, Spring 2019
Homework 1

Due on 23:59 Mar 15, 2019
Send to *cs*282_01@163.*com*
with subject "Chinese name+student number+HW1"

1 Problem 5: Convergence rates of Gradient Descent

We analysis the convergence rate from three possible cases, and suppose that:

$$w_* \in \arg\min_w F(w)$$

where F is convex.

1. The Smooth Case:
   Suppose F is L smooth and we can obtain:

   $$F(w') \leq F(w) + \nabla F(w) \cdot (w' - w) + \frac{L}{2}\|w - w'\|^2$$

   We consider the update rule :

   $$w_{t+1} = w_t - \eta \nabla F(w_t)$$

   Please try to show that the gradient descent converges at rate of $1/t$ in this case. (Hint: $\eta = 1/L$ and it equals to proof that $F(w_t) - F(w_*) < \frac{L}{t}\|w_0 - w_*\|^2$, where w_* is a new point.)

2. The Smooth and Strongly Convex Case:
   A function F is μ strongly convex if

   $$F(w') \geq F(w) + \nabla F(w) \cdot (w' - w) + \frac{\mu}{2}\|w - w'\|^2$$

   Similarly, we suppose that :

   $$w_{t+1} = w_t - \eta \nabla F(w_t)$$

And we know the supporting lemma that:

$$\|\nabla F(w)\|^2 \leq 2L(F(w) - F(w_*))$$

Please try to show that the GD algorithm has a constant learning rate. (Hint: In $\frac{L}{\mu} log(\|w_0 - w_*\|/\epsilon)$ iterations our distance to the optimal point is $\mathcal{O}(\epsilon)$, and try to proof that: $\|w_t - w_*\| \leq (1 - \frac{\mu}{L})^t \|w_0 - w_*\|$.

3. Non-smooth optimization and (sub-)gradient descent:
We denote that the update rule is :

$$w_{t+1} = w_t - \eta \nabla F(w_t)$$

where $\nabla F(w_t)$ is the sub-gradient at w_t and it satisfies:

$$F(w') \geq F(w) + \nabla F(w) \cdot (w' - w)$$

Suppose that for all w we have that $\|F(w)\| \leq B$ and $\|w_0 - w_*\| \leq R$. Set $\eta = \frac{R}{B}\sqrt{\frac{2}{T}}$, then please show that

$$F(\frac{1}{T}\sum_t w_t) - F(w_*) \leq \frac{RB}{\sqrt{T}}$$

1. (1) Since F is L-smooth, a natural idea is to simply choose η to minimize the upper bound. The following proof shows GD converges at rate of $1/t$.

Proof. By smoothness:

$$\begin{aligned} F(w_{t+1}) - F(w_t) &\leq F(w_t - \eta \nabla F(w_t)) - G(w_t) \\ &\leq -\eta \|\nabla F(w_t)\|^2 - \frac{L\eta^2}{2}\|\nabla F(w_t)\|^2 \\ &= -\frac{1}{2L}\|\nabla F(w_t)\|^2 \end{aligned}$$

we denote that:

$$\Delta_t = F(w_t) - F(w_*)$$

We have shown that :

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{2L}\|\nabla F(w_t)\|^2$$

Also, by convexity, we have that:

$$\Delta_t \leq \nabla F(w_t) \cdot (w_t - w_*) \leq \|\nabla F(w_t)\|\|w_t - w_*\|$$

Hence,

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{2L\|w_t - w_*\|^2}\Delta_t^2$$

Also, it is not difficult to show that $\|w_t - w_*\|^2$ decrease with t. Thus,

$$\Delta_{t+1} \leq \Delta_t - \frac{1}{2L\|w_1 - w_*\|^2}\Delta_t^2$$

where we have w_1 in the above expression. The above implies the rate (which one can show through induction).

□

2. *Proof.* Note that by strong convexity we have:

$$\nabla F(w) \cdot (w - w_*) \geq F(w) - F(w_*) + \frac{\mu}{2}\|w - w_*\|^2$$

Using these, we have that:

$$\begin{aligned}
\|w_{t+1} - w_*\| &= \|w_t - \eta\nabla F(w_t) - w_*\|^2 \\
&= \|w_t - w_*\|^2 - 2\eta\nabla F(w_t) \cdot (w_t - w_*) + \eta^2\|\nabla F(w_t)\|^2 \\
&\leq \|w_t - w_*\|^2 - 2\eta\left(F(w) - F(w_*) + \frac{\mu}{2}\|w_t - w_*\|^2\right) + \eta^2\|\nabla F(w_t)\|^2 \\
&\leq \|w_t - w_*\|^2 - 2\eta(F(w) - F(w_*)) - \eta\mu\|w_t - w_*\|^2 + 2\eta^2 L(F(w) - F(w_*)) \\
&\leq \|w_t - w_*\|^2 - \eta\mu\|w_t - w_*\|^2 + 2\eta(\eta L - 1)(F(w) - F(w_*)) \\
&\leq (1 - \frac{\mu}{L})\|w_t - w_*\|^2
\end{aligned}$$

where we used the setting of η in the last step.

□

3. *Proof.* First, note that the We have that:

$$\begin{aligned}
\|w_{t+1} - w_*\|^2 &= \|w_t - \nabla F(w_t) - w_*\|^2 \\
&= \|w_t - w_*\|^2 - 2\eta\nabla F(w_t) \cdot (w_t - w_*) + \eta^2\|\nabla F(w_t)\|^2 \\
&\leq \|w_t - w_*\|^2 - \eta\nabla F(w_t) \cdot (w_t - w_*) + \eta^2 B^2
\end{aligned}$$

using the definition of B. Hence,

$$\nabla F(w_t) \cdot (w_t - w_*) = \frac{1}{2\eta}\|w_t - w_*\|^2 - \|w_{t+1} - w_*\|^2 + \frac{\eta}{2}B^2$$

and so:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\nabla F(w_t) \cdot (w_t - w_*) &= \frac{1}{2\eta}(\|w_1 - w_*\|^2 - \|w_{T+1} - w_*\|^2) + \frac{\eta T}{2}B^2 \\
&\leq \frac{\|w_1 - w_*\|^2}{2\eta} + \frac{\eta T}{2}B^2 \\
&\leq \frac{RB}{\sqrt{T}}
\end{aligned}$$

where the last step uses our choice of η. The proof is completed since:

$$F(\frac{1}{T}\sum_t w_t) \leq \frac{1}{T}\sum_t F(w_t) \leq \frac{1}{T}\sum_{t=1}^{T} \nabla F(w_t) \cdot (w_t - w_*)$$

where both steps follow from convexity. □