Machine Learning, Spring 2019

Homework 3

Due on **April 7, 11:59 PM**

---

1. Submit your solutions to Gradescope (www.gradescope.com). Homework of this week contains two part, **theoretical part** and **programming part**. So there are two assignments in gradescope, the assignment titled with programming part will require you to submit your code.
2. **Make sure each solution page is assigned to the corresponding problems** when you submit your homework.
3. Any programming language is allowed for your code, but **make sure it is clear and readable with necesary comments.**

---

1 Subdifferential

For the following functions, verify whether the function is subdifferentiable everywhere, if it is, calculate a subgradient at a given x, if not, give your proof.

(a) $f(x) = \max_{i=1,\dots,m} |a_i^T x + b_i|$.(5 points)

(b) $f(x) = x_{[1]} + \cdots + x_{[k]}$, where $x_{[i]}$ is the ith largest element of $x \in \mathbb{R}^n$.(5 points)

(c) $f : \mathbb{R} \to \mathbb{R}$ with dom $f = \mathbb{R}_+$:(5 points)

$$f(x) = \begin{cases} 1 & x = 0, \\ 0 & x > 0. \end{cases}$$

2 Backtracking line search

Consider the following optimization problem:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and differentiable function, and its gradient is Lipschitz continuous, i.e., there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n.$$

The gradient descent method for solving problem (1) updates as

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)\,. \tag{2}$$

Here we apply backtracking line search to determine the stepsize α_k in (2). The gradient descent method using backtracking line search updates according to:

- Initialization: Fix parameter $\gamma \in (0, 1)$.
- At the kth iteration:
  - Starts with $\alpha_k = 1$.
  - If

$$f(x_k - \alpha_k \nabla f(x_k)) > f(x_k) - \frac{\alpha_k}{2} \|\nabla f(x_k)\|_2^2 ,$$

    update $\alpha_k = \gamma \alpha_k$ and repeat this sub-step. Otherwise, $k = k + 1$.

(a) Show that $\|\nabla f(x_k)\|_2 \to 0$ when $k \to \infty$.(10 points)

> Hint: The proof can be taken in three steps and we give the hint of the first two steps
> - Step 1: find a lower bound of step size such that $\underline{\alpha} \leq \alpha_k$ for each k.
> - Step 2: find a upper bound of $f(x_k) - f(x_{k+1})$ in terms of $\|\nabla f(x_k)\|_2$.
> - Step 3: (No hint).

(b) Instead of using the negative gradient as the descent direction, we consider a more general direction d_k, where $\|d_k\| = \|\nabla f(x_k)\|$ and the angle between d_k and $-\nabla f(x_k)$ satisfies $0^\circ \leq \angle(d_k, -\nabla f(x_k)) \leq \theta < 90^\circ$ for some θ. The new update follows

$$x_{k+1} = x_k + \alpha_k d_k. \tag{3}$$

Here we also consider using backtracking line search to determine the stepsize α_k. At the kth iteration, (**??**) with backtracking line search check whether the following condition holds:

$$f(x_k + \alpha_k d_k) > f(x_k) + \frac{\alpha_k}{2} \langle \nabla f(x_k), d_k \rangle . \tag{4}$$

The remaining part of the algorithm is the same as it for gradient descent. Show that $\|\nabla f(x_k)\| \to 0$ when $k \to \infty$.(10 points)

3 Stationary Points

3.1 Finding Stationary Points

Find the stationary points for the following functions $f(x)$:
(Hint: notice whether the domain of the function includes the stationary points you find.)

(a) $f(x) = (x - 1)/(x^2 + 5x + 3)$.(3 points)

(b) $f(x) = \ln(x^3 - 6x^2 - 15x + 1)$.(3 points)

(c) $f(x) = 7 + (2x^2 - 10x)\sqrt{x}$.(3 points)

3.2 Testing Stationary Points

You are given the function $f(x)$ and you have to find its stationary points. For each stationary point determine if it is a local maximum or local minimum(or neither) using the second derivative information.

(a) $f(x) = \ln(x) + 1/x$.(3 points)

(b) $f(x) = 2x^3 - 3x^2 - 12x + 5$.(3 points)

4 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.

(a) 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.)(10 points)

(b) 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: $\|w_t - w_*\| \leq (1 - \frac{\mu}{L})^t \|w_0 - w_*\|$.(10 points)

(c) 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 $\|\nabla F(w)\| \leq B$ and $\|w_0 - w_*\| \leq R$. Set $\eta = \frac{R}{B}\sqrt{\frac{2}{T}}$, then please show that (10 points)

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

5 Programming Problem

(a) Given the following basic least squares formulation,

$$\min \quad f(x) = \|Ax - b\|_2^2 \tag{5}$$

where, $A \in \mathbb{R}^{80\times 3}, x \in \mathbb{R}^3, b \in \mathbb{R}^{80}$. Please implement the gradient descent method using the given data set (**A.txt** and **b.txt**) and plot the iteration process.(5 points)

(b) Let's look deeper into above problem and use the optimal value of x you get (approximate value is fine). Generate matrix A, b by yourself to make the Hessian matrix have different condition number, verify the relation between convergence rate and condition number and plot your results.(5 points)

(c) Given the following function, please compare the performance of different stepsize/learning rate iteration methods (at least two) and plot your results.(10 points)

$$f(x, y) = (1 - x)^2 + 100\left(y - x^2\right)^2$$