IE 8990

Spring 2022

Homework #4

Due Date: 05/05/2022 5PM CST

Submission: Please put your answer and code in a PDF file and upload on Canvas

Q1. Proof that any local solution of the convex problem is a global solution.

- Logically:
  - It makes sense that any local minimum is the global minimum. As demonstrated in the figure below, to be considered a convex problem all points must be able to have a line drawn between them at which no other point of the function intersects. This means no matter how flat the line/plane looks, if it considered a convex problem then there is only one spot where the derivative changes signs.
- Mathematically:
  - Assume convex problem where (x, f(x)) and (y, f(y)) are point on the function and (z, f(z)) is a point in between.
  - $z = tx + (1-t)y$
    - Where t is a percentage value to define where z is.
    - $0 \leq t \leq 1$
  - $f(tx + (1-t)y)) \leq tf(x) + (1-t)f(y)$
    - This is stating that the true value of z must fall below the line between x and y.
  - **Swapping Gears**
  - $f(x^*) \leq f(y)\ ;\ \forall y\ in\ (x^* - \alpha, x^* + \alpha)$
    - Let's say that we have a local minimum x* that is the minimum for α range.
  - $z \in \mathbb{R}; 0 \leq t \leq 1$
    - Let's say z is any real point in the convex problem.
    - t is the same value as before.
  - $tx^* + (1-t)z; is\ in\ \ (x^* - \alpha, x^* + \alpha)$
    - Let's prove that our random point z falls within the range of the local minimum.
  - $f(x^*) \leq f(tx^* + (1-t)z)$
    - All points between x* and z are greater than x*.
  - $f(tx^* + (1-t)z) \leq tf(x^*) + (1-t)f(z)$
    - This statement must be true if it is a convex problem.
  - $(1-t)f(x^*) \leq (1-t)f(z)$
    - Combining and rearranging the above two equations shows that f(z) must always be larger than f(x*)

Q2. Please discuss the convergence rate of ISTA and FISTA algorithms. Hint: please check this paper. Beck, Amir, and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." SIAM journal on imaging sciences 2, no. 1 (2009): 183-202.

- The ISTA function is used to minimize a function, the same idea as stochastic gradient descent. Both would be used to optimize the neural network. The origin of the ISTA is an improvement of the sub-gradient method, which is an improvement to L1 regularization (LASSO). The ISTA uses a proximal gradient that is assumed from the smooth function and applies that gradient to the complex function. This allows the function to applied to other uses such as image/pixel recovery where it is most often used. The ISTA method is very computationally expensive since it requires several hundred iterations to converge. The FISTA model stands for the Faster ISTA, because its goal is to reduce the number of epochs to converge. It does this in the same way momentum-based learning in gradient descent works. Taking previous gradients in mind to apply it to current/future iterations. This “lowers” the scope of search and allows the current iterations to find the convergence more quickly.

Q3. Bonus: Select one article from below and summarize it (1 page, single-space, font size 12pt)

- Choi, Dami, Christopher J. Shallue, Zachary Nado, Jaehoon Lee, Chris J. Maddison, and George E. Dahl. “On empirical comparisons of optimizers for deep learning.” arXiv preprint arXiv:1910.05446 (2019).
- Schmidt, Mark, Nicolas Roux, and Francis Bach. “Convergence rates of inexact proximal-gradient methods for convex optimization.” Advances in neural information processing systems 24 (2011).