Optimization - Minitests

Joris LIMONIER

January - March 2022

Contents

1 Minitest 1

1.1 Question 1

Submission No, it is not possible to pursue directly this goal because we don't know the true distribution $\mathcal{D}$. This is a fundamental difference between Machine Learning and Statistics. We know however that our data was sampled from $\mathcal{D}$, and we know by the law of large numbers that our empirical loss will converge towards the expected value of the loss, as the number of samples increases.

Correction No because you don't know the underlying distribution $\mathcal{D}$.

1.2 Question 2

Submission One way to learn a model is by performing a train-test split in order to verify that our function (that we train on the train set) performs well on a set that is never seen before (*i.e.* the test set). We need to find the right model with not too many parameters (otherwise we over-fit our training set), and not too few parameters (otherwise we under-fit and do not learn enough

from data).
Other solutions, especially in case with small data sets, include K-fold cross-validation. One of its variations consists in disregarding a fold of the data set, while looking only at the $K-1$ other folds. Then repeat this step with the other folds.

Correction Instead of working with the true loss, work with the empirical loss.

2 Minitest 2

2.1 Question 1

What is a surrogate loss?

Correction

- A loss that we use instead of the natural loss.
- It is greater than the surrogate loss.
- It is convex in the number of parameters.

2.2 Question 2

Why do we use it in Machine Learning?

Correction Because convex optimization problems are easier to solve.

3 Minitest 3

4 Minitest 4

Consider an L-smooth and c-strongly convex function. Explain how the ratio $\frac{L}{c}$ (the condition number) affect the minimization process.

Correction Call our function F. Since $\kappa = \frac{\beta}{k+\gamma}$, small condition number means there is less space to search.
Consider the example of a very stretched ellipse. You will waste a lot of time going in one direction before starting to go (slowly) in the other direction.

5 Minitest 5

Explain how the convergence results are expressed for general non-convex functions in comparison to the case of strongly convex functions.

Sumbmission For convex functions, we usually have access to some nice formulas with guarantees at each step, so we know that after k steps, we will have reached some "improvement" (*e.g.* in terms of the loss function). For non-convex functions however, convergence results are expressed in the form of an infimum limit ($\liminf$) as the number of steps goes to infinity. This means that we don't know **for sure** that we will move towards a better solution at each step, but on average (*i.e.* in expectation), as the number of steps goes to infinity, we will achieve some result.
We saw that the sum of the squared gradients (its expectation) is finite, therefore the norm of the gradient of F (as per the lecture notation) goes to 0, which allows us to keep some of the results we had previously learned for the convex case.
A difference with the convex case is that with decreasing learning rate, we cannot bound the expected optimality gap.
We have also shown that SG brings us to regions where the gradient of F is small, but the noise may prevent us from getting the the absolute optimal solution.
Using second order moments can help us go through the roughness of non-convexity (intuitively, if the loss function goes down then up, the momentum may help us making it through the uphill slope).

Correction Expectation of the norm and $\liminf$