CS2822R Final Project Checkpoint 2

Group 21: Dennis Du, Sezim Yertanatov, Jared Ni  
  
(a) the question we are asking:

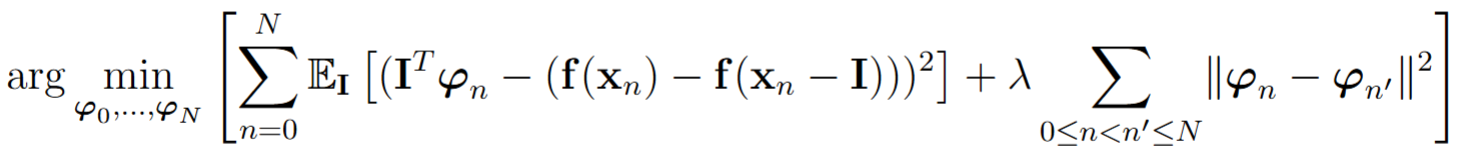
**How would we generalize metrics optimization for explanation methods where there are two or more target metrics to optimize?**

The research question of our team has remained largely consistent since the last checkpoint. As we discussed earlier, [On the (In)fidelity and Sensitivity of Explanations](https://arxiv.org/pdf/1901.09392) sheds light on the metrics of infidelity and sensitivity, and discusses the methodologies to derive an explanation that has low values for both of the properties. Nevertheless, the process outlined in this paper involves a two-step approach: first optimizing it for infidelity and then applying a smoothing kernel on it to achieve the smaller values of sensitivity. Thus, optimizing an explanation for the two metrics is considered as separate instances. We, in turn, posit that despite yielding surprisingly promising results, this approach is subject to two major drawbacks: 1) solely minimizing infidelity often hurts the sensitivity 2) the subsequent optimization for sensitivity tends to compromise the good infidelity results we have previously achieved. In light of these observations, our team is aiming to carry out a research on multi-objective optimization that takes into consideration several explanation properties while trying to derive a high-quality explanation that is balanced.

(b) how do we think we will address it:

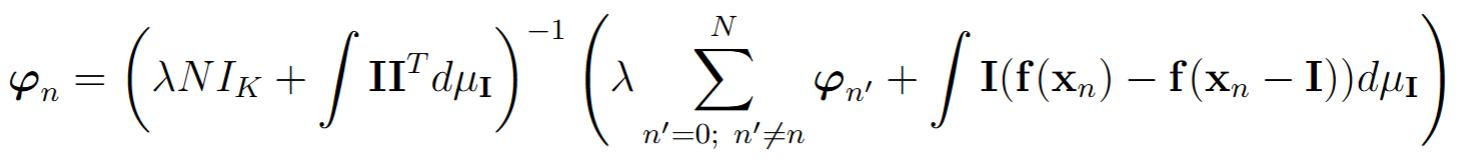
The contribution of our project would be to fill in this gap. The general idea is to optimize for a metric of the form INFD + λ \* SENS, where λ is a hyperparameter. The original paper proposed a way of optimizing INFD using the following logic: Suppose you are given a model f and an input x (which are treated as fixed). Then you find some vector (say φ) that minimizes the infidelity , and simply let Φ(f,x) = φ. However, sensitivity adds an additional wrinkle: we can't just optimize a single explanation φ, because sensitivity is a measure over the nearby explanations within a ball of radius r around the input x. So our idea would be to sample N points x1, ..., xN from a region around x, and generate N explanations φ1, ..., φN for each of these, in addition to an explanation φ0 for the original x (which we'll denote x0), so that the explanations optimize a combined infidelity and sensitivity metric of the form INFD + λ \* SENS. Note that *each* φn is a K-dimensional vector, where K is the dimension of the inputs xn. Over the course of the last several weeks, we have made some progress on defining the optimization objectives and the optimization methods.

The objective function that we define for the optimal explanation is



The term on the left represents the sums of the infidelities (as defined in Yeh et al.) across the explanations φ0, ..., φN. The term on the right represents the sum of the pairwise squared distances between explanations. This is a reasonable measure of sensitivity because the smaller this value is, the closer the φn are to each other. Once the optimal solution is found, we return the value of φ0, which is our generated explanation for the original input x0.

By setting the objective function’s gradient to 0, we find that the optimal explanations satisfy



IK is the K×K identity matrix and μI is the PDF of the distribution of the perturbations I. This equation relates each explanation φn to each of the other explanations φn’ for n’ ≠ n. This motivates optimization by coordinate descent: we initialize φ0, ..., φN to some value (currently, we use zero vectors), and then we iteratively update the value of each φn according to this equation, treating the other φn’ as fixed. We cycle through all of φ0, ..., φN as many times as is needed until the maximum amount by which any individual feature explanation changes is lower than a specified tolerance threshold.

(c) how you plan to measure success:

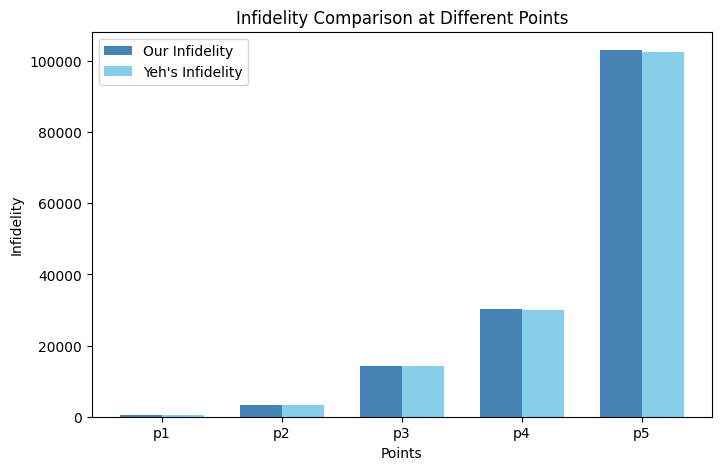
In the previous checkpoint, we defined success by the following three metrics; Theory, Evaluation, and Time. We will further extend the definition of these metrics for checkpoint 2.

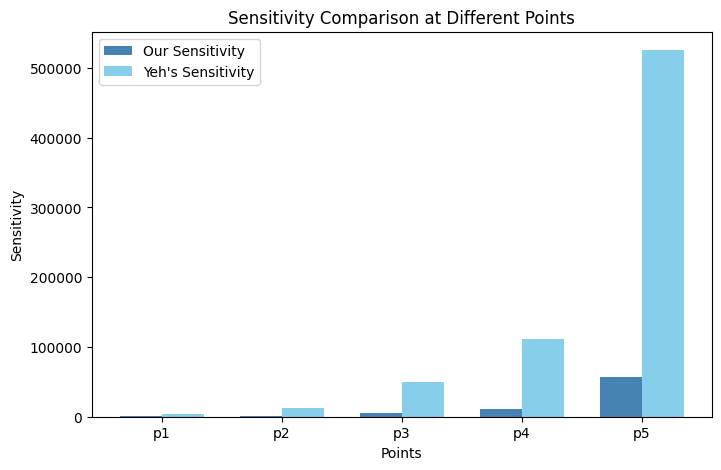
**Theory**: We have successfully formalized the theory that combines multiple metrics (sensitivity and infidelity) into a closed-form, iterative explanation that prioritizes both. So far, we have developed the closed-form equation for a new explanation method, where we prioritize both metrics iteratively. The iterative approach, where we optimize one explanation per iteration, is a system that can potentially extend into a generalizable form that can incorporate more metrics to optimize. Next step for theory, we will try to develop our equation into a more generalizable form given this approach we discovered.

**Evaluation:** We have evaluated our theory on simple toy example models (f(x) = x3, for example), and next up we will visualize it for more complicated models and equations while benchmarking the performance on the infidelity and sensitivity front, and then measure this performance against other explanation methods found in the literature. We will see if we can match their performance without any tradeoffs while optimizing for multiple metrics, such as infidelity and sensitivity. Since we won’t be able to predict the results, we deem the evaluation successful as long as we run the metrics on some common models and datasets (e.g. those used in the infidelity paper), and we record our results. Ideally, we have time to expand to metrics proposed in other papers (for example, the paper that discusses complexity: [Evaluating and Aggregating Feature-based Model Explanations](https://arxiv.org/pdf/2005.00631)). We would also try to optimize for the metrics proposed in those papers if time and scope allow. Moreover, another potential addon would be to try various perturbation distributions to generate explanations, which would allow us to investigate the effectiveness of multi-object optimization in different scenarios.

Specifically, we evaluated this on a toy model of the form f(a, b, c) =a3 + b3 + c3. We calculated the infidelity and sensitivity of our explanation method and compared this to Yeh et al’s explanation method. (Infidelity and sensitivity are based loosely on Yeh et al’s definitions: We used standard normal perturbations in the infidelity metric, and sensitivity was defined as the maximum squared L2 norm of the distance between φ0 and any of the φn.) The results below show that with our method, we can significantly decrease sensitivity while allowing infidelity to increase only slightly.

| **Input point**  **x = (a, b, c)** | **Our**  **infidelity** | **Yeh et al.’s infidelity** | **Our sensitivity** | **Yeh et al.’s sensitivity** |
| --- | --- | --- | --- | --- |
| (2, 1, 3) | 532.17 | 518.28 | 337.83 | 3495.66 |
| (3, 5, 7) | 3304.62 | 3275.27 | 1305.66 | 12871.97 |
| (9, 11, 12) | 14230.14 | 14139.57 | 5237.15 | 49488.68 |
| (11, 14, 21) | 30214.63 | 30041.73 | 11696.58 | 111637.35 |
| (19, 42, 21) | 103094.17 | 102546.38 | 57499.28 | 525702.00 |



****

**Time:** We aim to conclude our project within the designated timeline in the class, starting with the base cases of evaluating 2 metrics, which we are currently in the process of doing, and expanding to more evaluations if time allows. We will deem it a success to have a complete project where we can develop meaningful research insights about generalizing metrics optimization, and that we learn in the process.