CS2822R Final Project Checkpoint 1

Group 21: Dennis Du, Sezim Yertanatov, Jared Ni

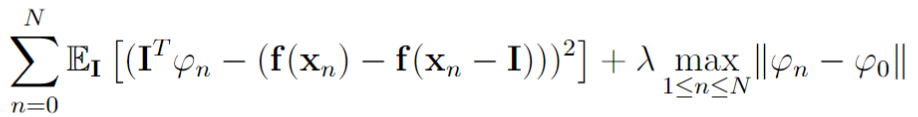
(a) the question you are asking:

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

[On the (In)fidelity and Sensitivity of Explanations](https://arxiv.org/pdf/1901.09392) provides measures of infidelity (INFD) and sensitivity (SENS), and discusses ways of minimizing them. To a certain degree they treat the minimization of the two metrics as two separate contributions in their paper: (1) they provide a closed-form solution for an explanation method that minimizes infidelity, and (2) they show that given some explanation method, a kernel can be applied to smooth the explanation and reduce sensitivity, with the happy consequence that infidelity could go down as well. But if you used (1) to generate some explanation Φ minimizing infidelity, there are no guarantees on Φ's sensitivity. And although applying (2) could help you reduce sensitivity by producing a smoothed explanation Φ\_smooth, you would definitely get worse infidelity than Φ, because by definition Φ was best for infidelity. So there is no clear procedure for optimizing both metrics.

(b) how do you think you 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 paper had 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 number (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 x\_1, ..., x\_N from the ball of radius r around x, and generate N explanations φ\_1, ..., φ\_N for each of these, in addition to an explanation φ\_0 for the original x (which we'll denote x\_0), so that the explanations optimize a combined infidelity and sensitivity metric of the form INFD + λ \* SENS:



Then the explanation for the original input x is the value of φ\_0.

By the next checkpoint, we will develop some more specific methods for finding an explanation that optimizes this metric. The ideal would be to develop a closed-form solution. If this does not work out, there are two further ideas to try: either we can try to tweak the metric into a form that captures the same idea but is conducive to a closed-form solution, or we can apply iterative optimization methods, such as gradient descent. We will also consider how to choose the hyperparameter λ. The max might end up being tricky to deal with, so we could use a sum or average instead. Another thought is to actually treat the x\_n as a random variable rather than as a set of N sampled points, in which case we would replace the sums with expectations, but that would probably complicate how to think about φ. In any case, once we've developed a method to produce explanations optimizing this metric, we can evaluate it on some actual models.

(c) how you plan to measure success:

We will measure our success by the following dimensions:

**Theory**: Formalizing the theory that combines multiple metrics, ideally a generalizable metric evaluation equation that includes most if not all metrics proposed in the literature. In all cases, we must include an equation for at least 2 metrics, but ideally more. The more metrics we can generalize the theory to include, the more successful we deem our theory research.

**Evaluation:** We will test our theory by comparing it to the optimizations proposed in the Infidelity paper. We will see if we can match their performance without any tradeoffs while optimizing for multiple metrics, such as fidelity 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.

**Time:** We aim to conclude our project within the designated timeline in the class, starting with the base cases of evaluating 2 metrics, 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.