# Heuristics


MY FIRST® TICONDEROGA®
2 HB
TICONDEROGA® Laddie®
2 HB
DIXON TICONDEROGA®
2 HB

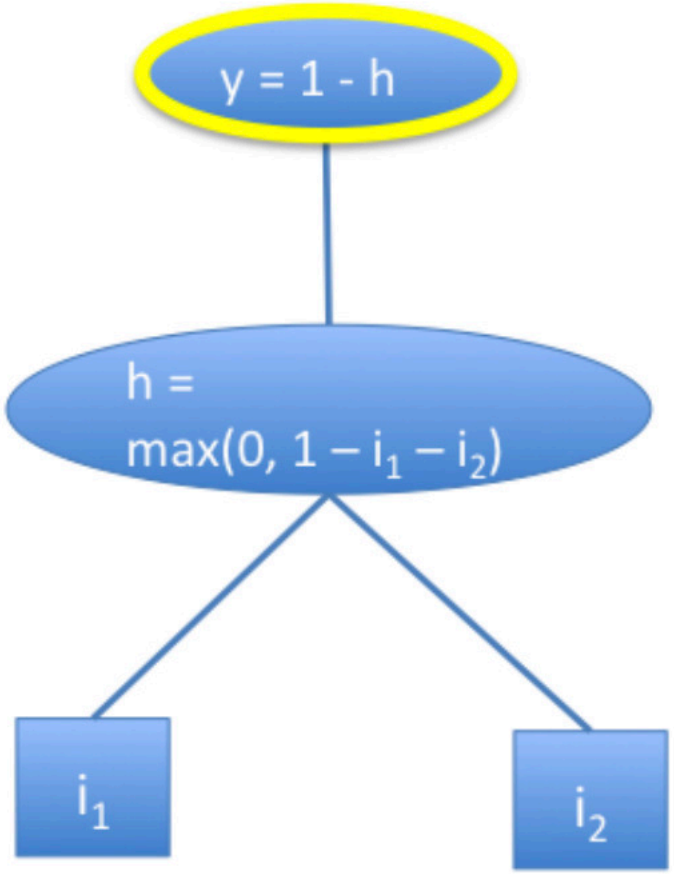
# Sensitivity Analysis

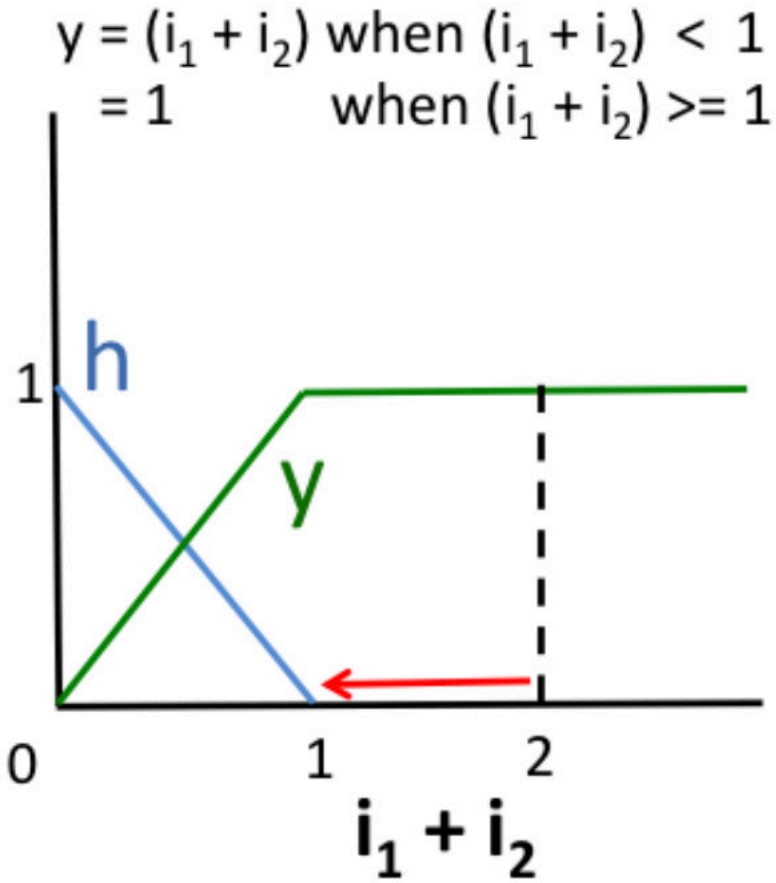
- Measure local change in estimate

$$s_f(x) = \partial_x f(x)$$

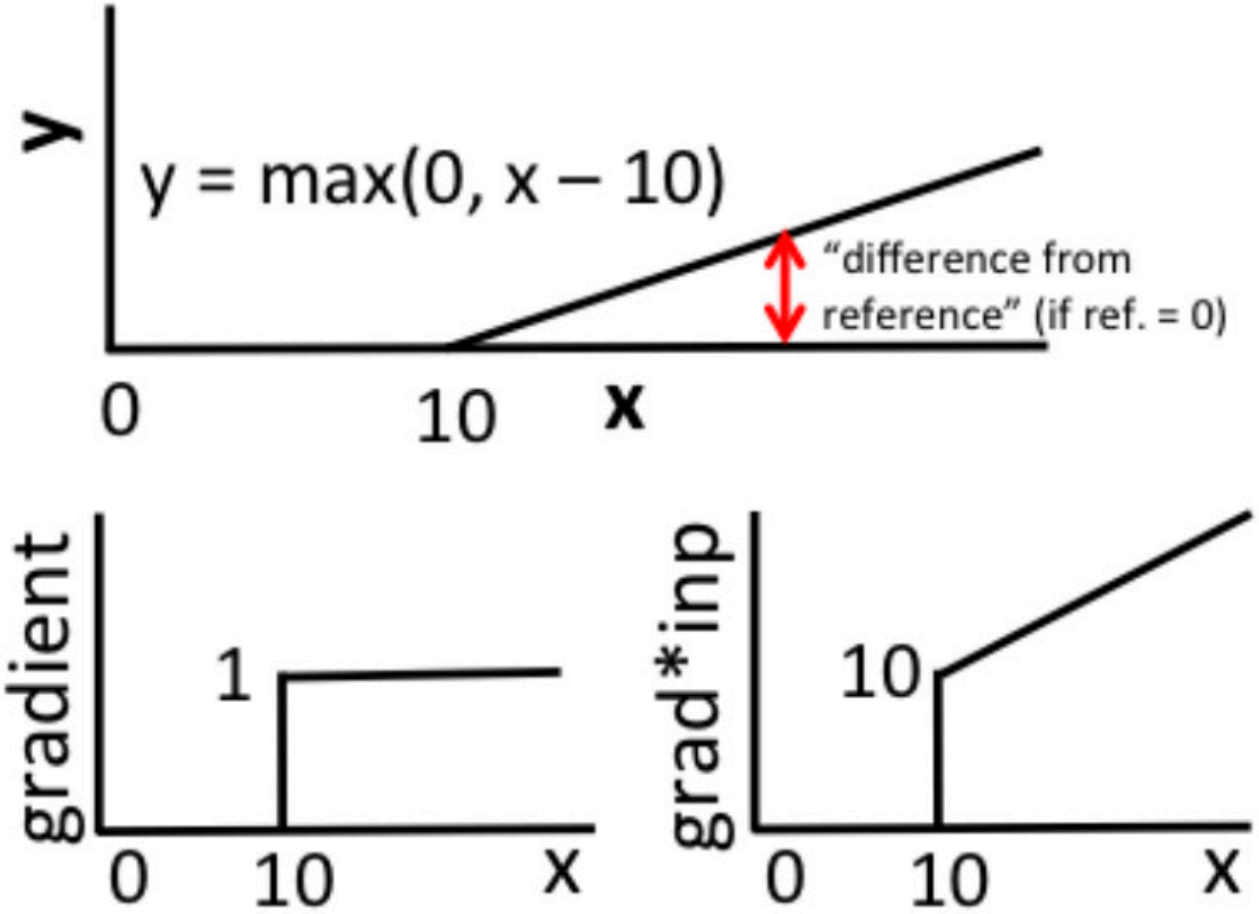
- Can be computed via back propagation
  - Easy support by DL frameworks
  - Often not so useful due to ReLU and other clipping operations (misses out on relevant changes)
  - Leads to weird results …
- Hack (Bach et al., 2015) - use grad * inp: $\Delta x \cdot \partial_x f(x)$

# Toy Problem (Example)



# Toy Problem (Example)



# Fixing It (DeepLIFT)

- **Key heuristic** - replace derivatives with finite differences

$$\partial_i f(x) \cdot \Delta x_i \Longrightarrow \frac{f(x' + \Delta x_i) - f(x')}{\Delta x_i} \cdot \Delta x_i$$

- Sundarajan & Najmi, 2019 design a lot of special rules for
  - ReLU (decompose into positive and negative changes)
  - Activation functions in general (use finite differences)
  - Linear operations (use as is)
  - Can use backprop to compute score

# Details (DeepLIFT)

github.com/kundajelab/deeplift

- Difference Decomposition

$$\sum_i C_{\Delta x_i, \Delta f} = \Delta f$$

- Gradient Approximation

$$m_{\Delta x_i, \Delta f} = \frac{C_{\Delta x_i, \Delta f}}{\Delta f}$$

- Chain Rule across Layers

$$m_{\Delta x_i, \Delta f} = \sum_j m_{\Delta x_i, \Delta y_j} m_{\Delta y_j, \Delta f}$$

Use backprop

# Lots more

- Guided backprop
- KernelSHAP
- Applications to text & images
  - Need to identify larger components
  - No longer possible to combine parts at random.
    What is the reference text $x_0$?
- Causality
  What we *really* want is to *explain why.*

# References


- Sundararajan & Najmi, 2020
  The Many Shapley Values for Model Explanation
  https://arxiv.org/pdf/1908.08474.pdf
- Ribeiro, Singh & Guestrin, 2016
  "Why Should I Trust You?": Explaining the Predictions of Any Classifier
  https://arxiv.org/abs/1602.04938 (LIME)
- Lundberg & Lee, 2017
  A Unified Approach to Interpreting Model Predictions
  https://arxiv.org/abs/1705.07874 (SHAP)
- Janzing, Minorics, Bloebaum, 2020
  Feature relevance quantification in explainable AI: A causal problem
  https://arxiv.org/abs/1910.13413
- Shrikumar, Greenside, Kundaje, 2019
  Learning Important Features Through Propagating Activation Differences
  https://arxiv.org/pdf/1704.02685.pdf
- Kevin Leyton Brown's Shapley Lecture (caution - typos!)
  https://www.cs.ubc.ca/~kevinlb/teaching/cs532l%20-%202007-8/lectures/lect23.pdf

# Summary

- Explainability
- Options
  - Simplicity
  - Approximate Simplicity
  - Local Simplicity
- Conditioning and Backdoors
- Axiomatic Approaches
  - SHAP
  - Integrated Gradient
- Heuristics