Axiomatic Approaches

Concepts

- **Idea from LIME**
  Use an explanation model g to explain f.
- **Local accuracy**
  Explanation $g_x(x') = \phi_x^\top x' + \phi_x^0$ matches f for $x = x'$.
- **Missingness**
  Missing features $x_i = \emptyset$ have no influence, i.e. $\phi_i = 0$
- **Consistency**
  If for some function f' feature i always makes a bigger difference than for f, the explanation for f' is also bigger.

SHAP Theorem (Lundberg & Lee, 2017)

The only score satisfying the requirements of missingness, local accuracy and consistency is the Shapley value.

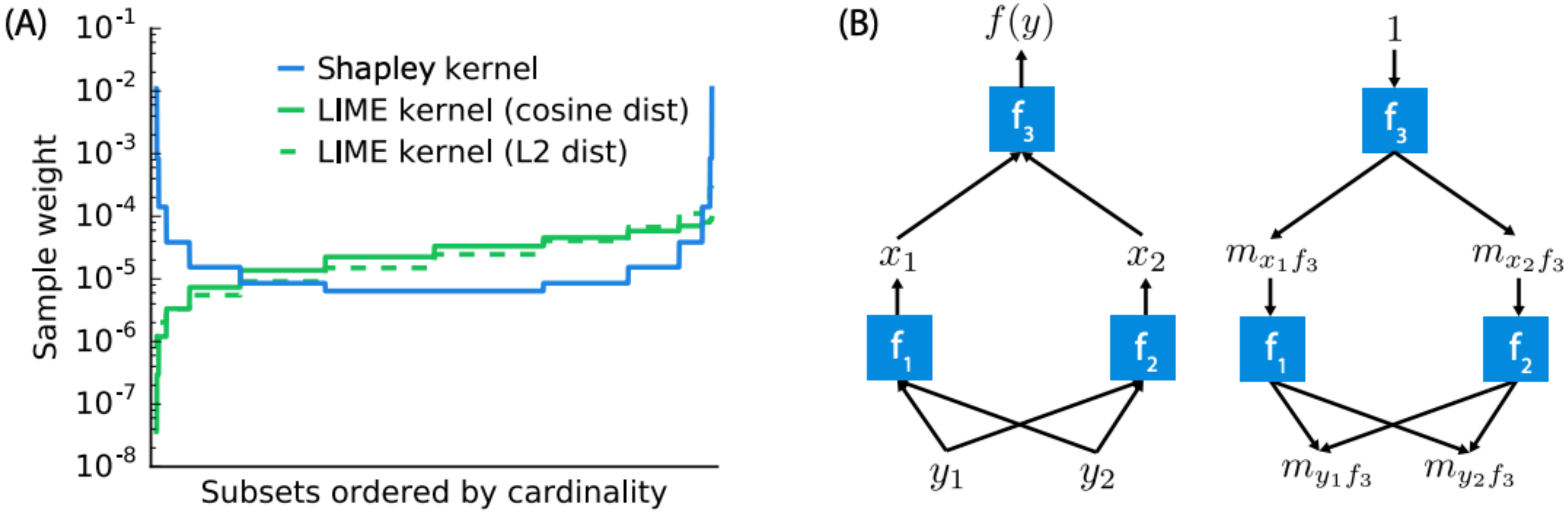
$$\phi(i, N) = \sum_{S \in N \setminus \{i\}} \frac{1}{|N|} \binom{|N| - 1}{|S|}^{-1} \left[f\left(x_{S \cup \{i\}}\right) - f\left(x_S\right) \right]$$

Good news

- For linear functions this returns the original function
- For LIME it gives us local weightings
- Linear model explainers as special cases

SHAP Theorem (Lundberg & Lee, 2017)

The only score satisfying the requirements of missingness, local accuracy and consistency is the Shapley value.



SHAP Theorem (Lundberg & Lee, 2017)

The only score satisfying the requirements of missingness, local accuracy and consistency is the Shapley value.

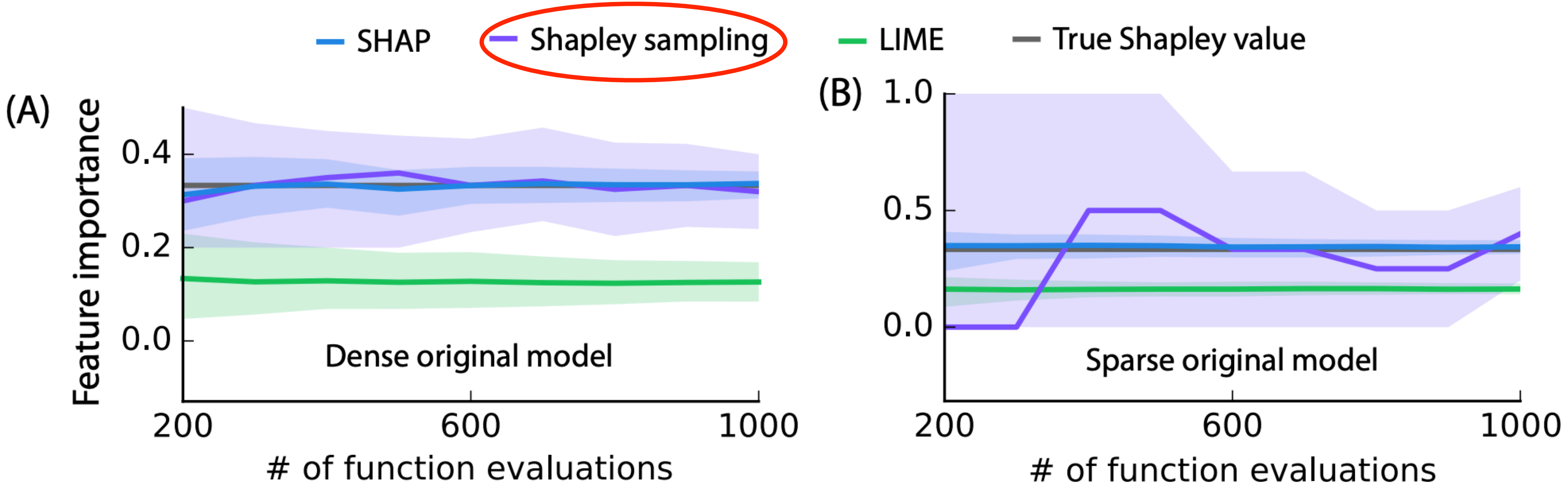
$$\phi(i, x) = \sum_{S \in N \setminus \{i\}} \frac{1}{|N|} \binom{|N| - 1}{|S|}^{-1} \left[f\left(x_{S \cup \{i\}}\right) - f\left(x_S\right) \right]$$

Devil in the detail

- How to define function on subset? What does leaving out a feature mean? Set to zero? Set to mean?
- How to compute/approximate this $O(2^{|N|})$ sum efficiently?

SHAP Theorem (Lundberg & Lee, 2017)

The only score satisfying the requirements of missingness, local accuracy and consistency is the Shapley value.



Lots of fast approximations

- Fast TreeShap - https://arxiv.org/abs/2109.09847
  (easy to compute since only affects few features)
- Approximate expansions $O(2^{|N|})$ to $O(|N|^k)$
  (only include the last few terms in the sum)
- Sample according to normalization weights
  (Shapley sampling)
- DeepShap and similar approximations

- Start with github.com/slundberg/shap

Reference Scores

- What to do with left-out features?
  - In general, **do not** try to model conditional distribution (a lot of the SHAP improvements do this)
  - Just use the approximation in original SHAP paper (see Janzing et al, 2020 and also previous discussion)
- **In practice**
  Draw unrelated values for the features that we are leaving out (works for tabular but more tricky for text & tabular data since context matters)

Toy example (from Janzing et al., 2020)

$$f(x_1, x_2) = x_1 \qquad p(x_1, x_2) = \begin{cases} 1/2 & \textit{for } x_1 = x_2 \\ 0 & \textit{otherwise} \end{cases}$$

(1) with conditional expectations:

$$\begin{aligned} f_\emptyset(\mathbf{x}) &= \mathbb{E}[f(X_1, X_2)] = 1/2 & (6) \\ f_{\{1\}}(\mathbf{x}) &= \mathbb{E}[f(x_1, X_2)|x_1] = x_1 & (7) \\ f_{\{2\}}(\mathbf{x}) &= \mathbb{E}[f(X_1, x_2)|x_2] = x_2 & (8) \\ f_{\{1,2\}}(\mathbf{x}) &= f(x_1, x_2) = x_1 & (9) \end{aligned}$$

Therefore,

$$\begin{aligned} C(2|\emptyset) &= f_{\{2\}}(\mathbf{x}) - f_\emptyset(\mathbf{x}) = x_1 - 1/2 \\ C(2|\{1\}) &= f_{\{1,2\}}(\mathbf{x}) - f_{\{1\}}(\mathbf{x}) = x_1 - x_1. \end{aligned}$$

Hence, the Shapley value for X_2 reads:

$$\phi_2 = \frac{1}{2}(x_1 - 1/2 + x_1 - x_1) = x_1/2 - 1/4 \neq 0.$$

(2) with marginal expectations:

$$\begin{aligned} f_\emptyset(\mathbf{x}) &= \mathbb{E}[f(X_1, X_2)] = 1/2 & (10) \\ f_{\{1\}}(\mathbf{x}) &= \mathbb{E}[f(x_1, X_2)] = x_1 & (11) \\ f_{\{2\}}(\mathbf{x}) &= \mathbb{E}[f(X_1, x_2)] = 1/2 & (12) \\ f_{\{1,2\}}(\mathbf{x}) &= f(x_1, x_2) = x_1. & (13) \end{aligned}$$

We then obtain

$$\begin{aligned} C(2|\emptyset) &= f_{\{2\}}(\mathbf{x}) - f_\emptyset(\mathbf{x}) = 0 \\ C(2|\{1\}) &= f_{\{1,2\}}(\mathbf{x}) - f_{\{1\}}(\mathbf{x}) = 0, \end{aligned}$$

which yields $\phi_2 = 0$.

Integrated Gradient Axioms

- **Completeness**

$$\sum_i \phi(i, x) = f(x) - f(x_0)$$

- **Sensitivity**

  If $f(x)$ does not depend on i then $\phi(i, x) = 0$.
- **Implementation Invariance**

  Scores do not depend on how f is implemented.
- **Linearity**

  For $f = \alpha_1 f_1 + \alpha_2 f_2$ the scores are $\phi(i, x) = \alpha_1 \phi_1(i, x) + \alpha_2 \phi_2(i, x)$.
- **Symmetry**

  If f is symmetric in inputs i, j then scores are identical.

Integrated Gradient Axioms

- **Theorem** (Sundarajan & Najmi, 2019)

$$\phi(i, x) = (x_i - x_i') \int_0^1 \partial_{x_i}(x' + \alpha(x - x'))d\alpha$$

  is the only representation that is admissible. Easy to check that the axioms are all satisfied.

- **Useful connection**
  This gives us a strategy to get the Shapley values more cheaply when IG can be computed.