Accumulated Local Effects (ALE)

Unbiased Feature Effects with Correlated Features

⚠ PDP (Biased)

✓ ALE (Unbiased)

Key Features

ALE Advantages

- ✓ Unbiased with correlated features
- ✓ Local neighborhoods only
- ✓ Computationally efficient
- ✓ No marginalization needed

Local Effects

Accumulated across feature range

Reliability

More trustworthy for realistic datasets

Implementation

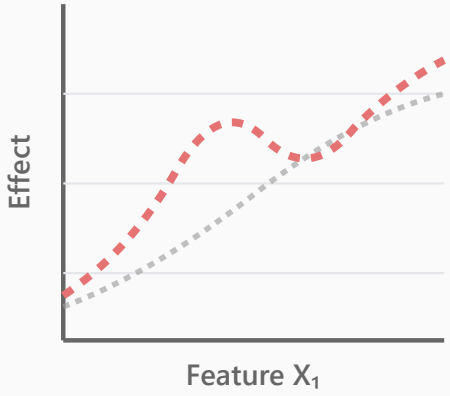
ALEPython package

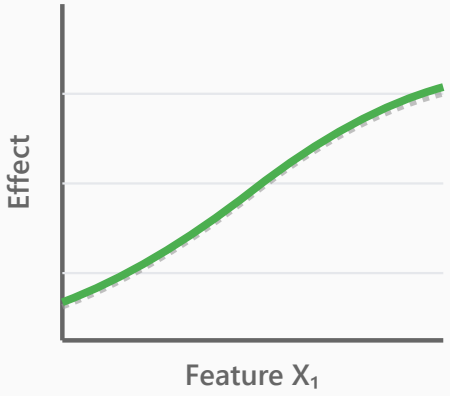
Quick Comparison

	PDP	ALE
Correlated	Biased	Unbiased
Speed	Slow	Fast

PDP (Biased)
ALE (Unbiased)
True Effect

How





ntervals. E
ally exists.

Differences

edictions when the feature value
nterval. This represents the "local

```
}) - f(z_j, X_{-j})
nts in each interval
```

cts

Average the local effects of all data points within each interval. This cancels out the effects of other features.

```
ALE_j = (1/n_j) Σ [f(z_{j+1}, X_i) - f(z_j, X_i)]
n_j: number of data points in interval j
```

ys i

Starting from the first interval, accumulate the local effects of each interval. This is what "Accumulated" in Accumulated Local Effects means.

```
ALE(z) = Σ(j=1 to k) ALE_j
Cumulative from left to right
```

5 Center the Plot

Center the entire ALE plot so the average is 0. This allows us to see the relative effect of the feature rather than its absolute effect.

```
ALE_centered(z) = ALE(z) - E[ALE(z)]
Interpretation: effect relative to average
```

6 Visualize & Interpret

The final ALE plot shows the pure effect on model predictions at each feature value. An upward slope indicates positive effect, while a downward slope indicates negative effect.

💡 Key Takeaway

ALE operates only within **local regions where actual data exists**, so it doesn't suffer from the bias caused by unrealistic data combinations like PDP does. Even when features are correlated, ALE measures pure effects while maintaining the natural distribution of other features by considering only small changes within each interval.