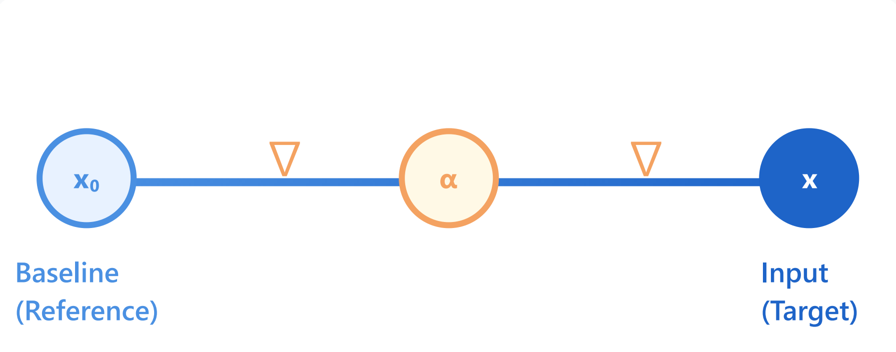
GradientSHAP: Gradient-Based Approximation

Combines Integrated Gradients with Shapley sampling

$$\varphi_i = E[\nabla f(x) \times (x_i - x_{baseline})]$$

Expectation over multiple baselines improves stability

Integration Path from Baseline to Input



Computation Steps

1. Sample multiple random baselines
2. Create interpolated points along path
3. Compute gradients at each point
4. Multiply by input difference
5. Average over all baselines

Efficiency

Fast computation for differentiable models

- Uses automatic differentiation
- GPU acceleration supported
- Scales well with features

Multiple Baselines

Random sampling improves robustness

- Reduces variance
- Better approximation
- More stable estimates

Non-linear Interactions

Captures complex feature relationships

- Gradient-based attribution
- Handles neural networks
- Integrated Gradients basis

📊 Calculation Principle & Examples

Example 1: Image Classification Model

🎯 Problem Setup

Cat image classification (features: ears, whiskers, eyes)

```
x = [0.8, 0.6, 0.9]
x₀ = [0, 0, 0] (black image)
```

📐 Step 1: Generate Path

Intermediate point at $\alpha = 0.5$:

```
x' = x₀ + 0.5(x - x₀)
= [0.4, 0.3, 0.45]
```

🔍 Step 2: Compute Gradients

```
∇f(x') = [0.7, 0.4, 0.8]
(importance of each feature)
```

✖️ Step 3: Multiply with Difference

```
φᵢ = ∇f(x') × (xᵢ - x₀ᵢ)
φ₁  = 0.7 × 0.8 = 0.56
```

Example 2: Multiple Baselines

🎲 Problem Setup

Credit score prediction (features: income, debt)

```
x = [80k, 20k]
Sample 3 baselines
```

📊 Calculate for Each Baseline

```
Baseline 1: x₀¹ = [30k, 10k]
→ φ¹ = [0.45, -0.15]

Baseline 2: x₀² = [50k, 15k]
→ φ² = [0.38, -0.12]

Baseline 3: x₀³ = [40k, 5k]
→ φ³ = [0.52, -0.18]
```

📈 Step 4: Average Results

```
φᵢ = E[φⁿ] = (φ¹ + φ² + φ³) / 3
```

```
φ2  = 0.4 × 0.6 = 0.24
φ3  = 0.8 × 0.9 = 0.72
```

Final Attribution

Eyes(0.72) > Ears(0.56) > Whiskers(0.24)

```
Income: (0.45+0.38+0.52)/3 = 0.45
Debt: (-0.15-0.12-0.18)/3 = -0.15
```

Stable Attribution

Income: +0.45 | Debt: -0.15

💡 Interpreting Results

✅ **Positive values:** Feature contributes to increasing the prediction (e.g., higher income → higher credit score)

❌ **Negative values:** Feature contributes to decreasing the prediction (e.g., higher debt → lower credit score)

📊 **Magnitude:** Larger absolute value indicates stronger influence (e.g., |0.72| > |0.56| → eyes more important than ears)

🎯 **Multiple baselines:** Using multiple reference points reduces variance and provides more stable results