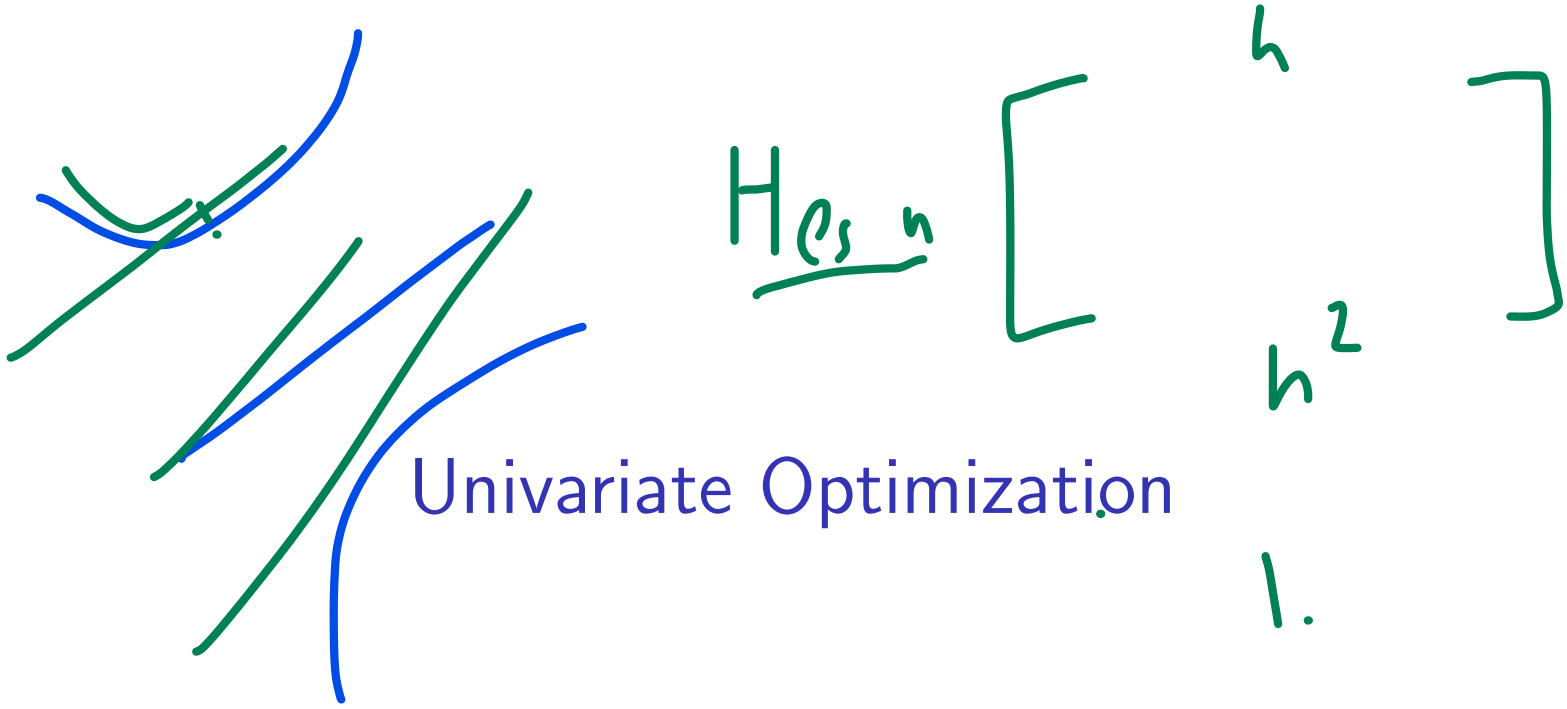


Univariate Optimization

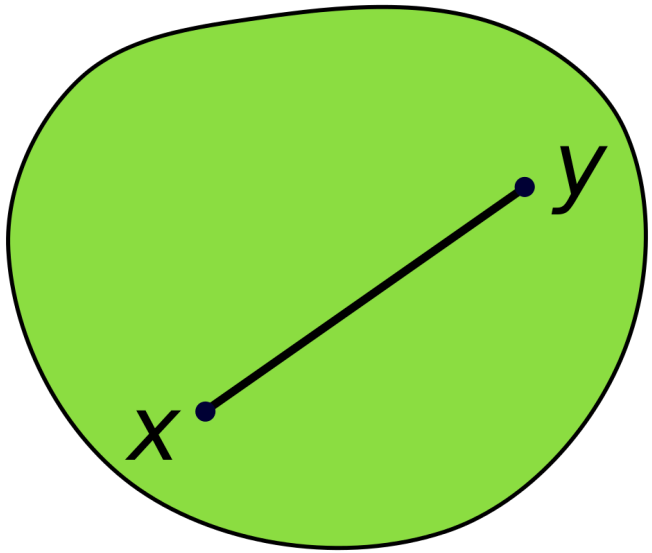
October 20, 2025

Content Overview

- Convex Sets / Convex Functions
- Unimodal Functions
- Univariate Optimization Methods (Golden Section Search, Brent's Method)
- Stopping Criteria

Optimization in Machine Learning

Mathematical Concepts: Convexity



Learning goals

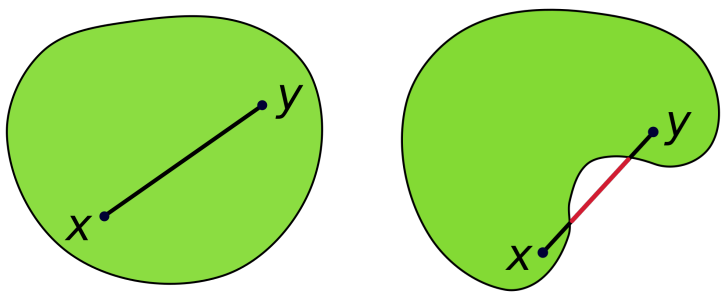
- Convex sets
- Convex functions

CONVEX SETS

A set of $\mathcal{S} \subseteq \mathbb{R}^d$ is **convex**, if for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and all $t \in [0, 1]$ the following holds:

$$\mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in \mathcal{S}$$

Intuitively: Connecting line between any $\mathbf{x}, \boldsymbol{y} \in \mathcal{S}$ lies completely in $\mathcal{S}$.



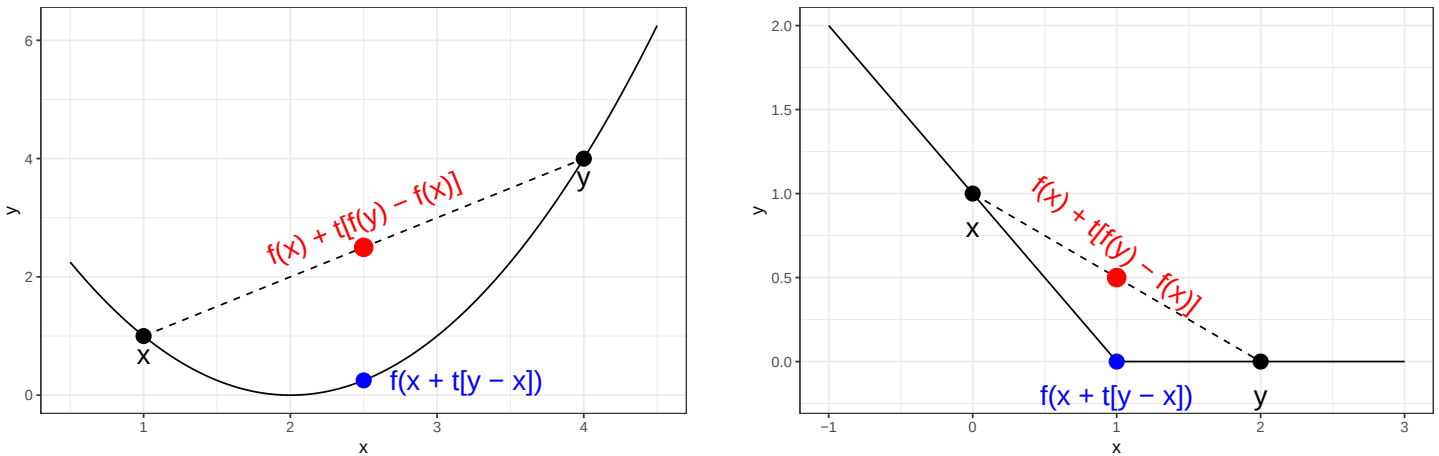
Left: convex set. **Right:** not convex. (Source: Wikipedia)

CONVEX FUNCTIONS

Let $f : \mathcal{S} \to \mathbb{R}$, $\mathcal{S}$ convex. f is **convex** if for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and all $t \in [0, 1]$

$$f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) \leq f(\mathbf{x}) + t(f(\mathbf{y}) - f(\mathbf{x})).$$

Intuitively: Connecting line lies above function.



Left: Strictly convex function. **Right:** Convex, but not strictly.

Strictly convex if "$<$" instead of "$\leq$". **Concave** (strictly) if the inequality holds with "$\geq$" ("$>$"), respectively.

Note: f (strictly) concave $\Leftrightarrow$ $-f$ (strictly) convex.

EXAMPLES

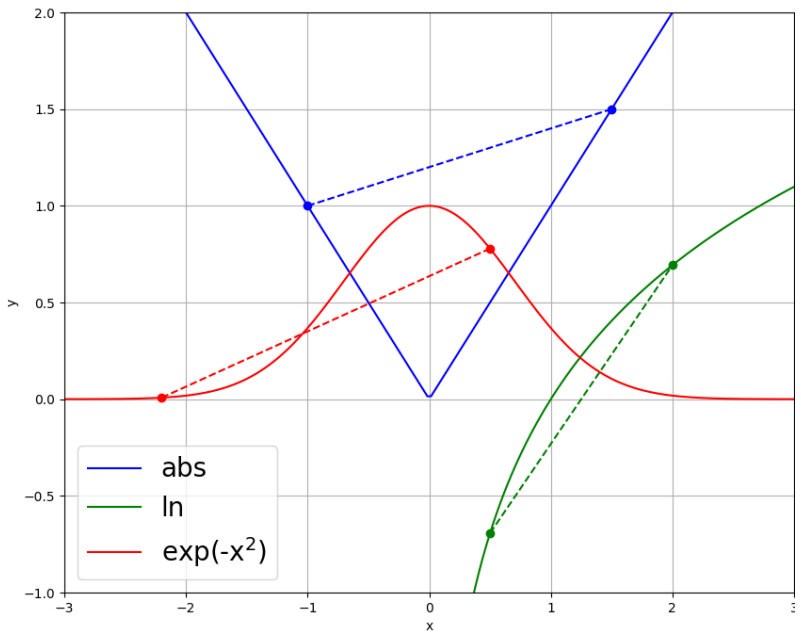
Convex function: $f(x) = |x|$

Proof:

$$\begin{aligned} f\left(x + t(y - x)\right) &= |x + t(y - x)| = |(1 - t)x + t \cdot y| \\ &\leq |(1 - t)x| + |t \cdot y| = (1 - t)|x| + t|y| \\ &= |x| + t \cdot (|y| - |x|) = f(x) + t \cdot (f(y) - f(x)) \end{aligned}$$

Concave function: $f(x) = \log(x)$

Neither nor: $f(x) = \exp(-x^2)$ (but log-concave)



OPERATIONS PRESERVING CONVEXITY

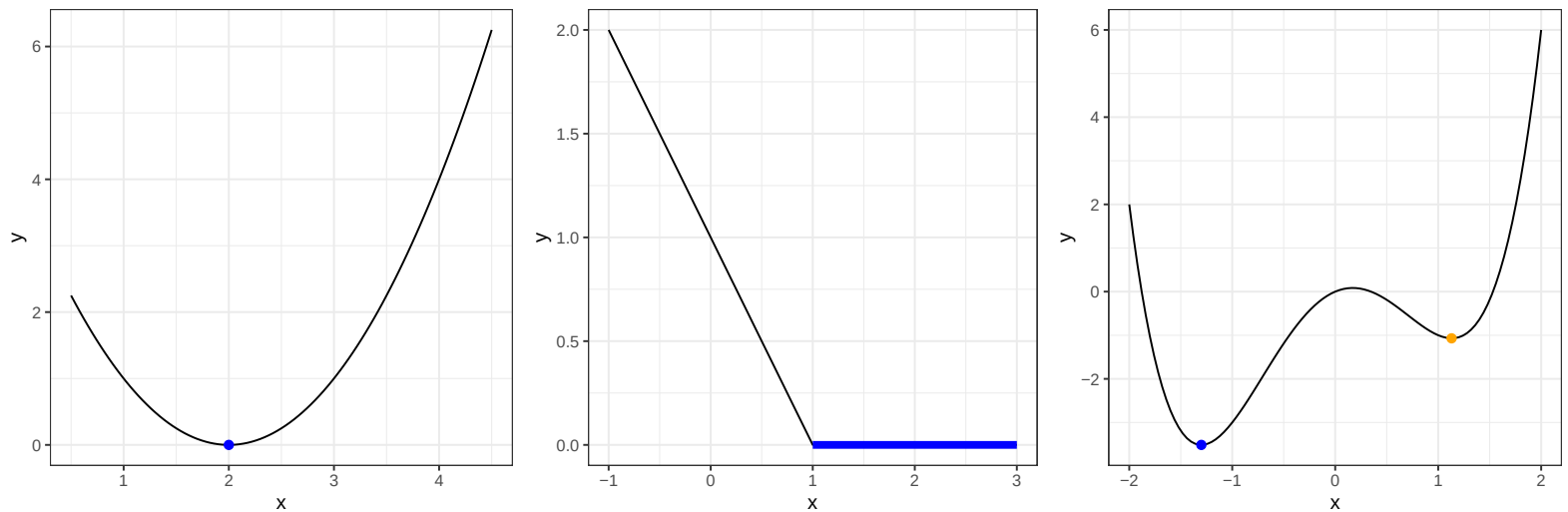
- **Nonnegatively weighted summation:** Weights $w_1, \ldots, w_n \geq 0$, convex functions $f_1, \ldots, f_n$: $w_1 f_1 + \cdots + w_n f_n$ also convex
  In particular: Sum of convex functions also convex
- **Composition:** g convex, f linear: $h = g \circ f$ also convex
  Proof:

$$\begin{aligned} h(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) &= g(f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))) \\ &= g(f(\mathbf{x}) + t(f(\mathbf{y}) - f(\mathbf{x}))) \\ &\leq g(f(\mathbf{x})) + t(g(f(\mathbf{y})) - g(f(\mathbf{x}))) \\ &= h(\mathbf{x}) + t(h(\mathbf{y}) - h(\mathbf{x})) \end{aligned}$$

- **Elementwise maximization:** $f_1, \ldots, f_n$ convex functions: $g(\mathbf{x}) = \max\{f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})\}$ also convex

CONVEX FUNCTIONS IN OPTIMIZATION

- For a convex function, every local optimum is also a global one
  $\Rightarrow$ No need for involved global optimizers, local ones are enough
- A strictly convex function has at most one optimal point
- Example for strictly convex function without optimum: exp on $\mathbb{R}$



Left: Strictly convex; exactly one local minimum, which is also global. **Middle:** Convex, but not strictly; all local optima are also global ones but not unique. **Right:** Not convex.

Unimodal Functions: Definition

Definition

A function $f : \mathbb{R} \to \mathbb{R}$ is called **unimodal** if it has exactly one local optimum (minimum or maximum), which is also the global optimum.

Formal Definition:

- For a unimodal function with minimum at x^*:
  - f is strictly decreasing on $(-\infty, x^*]$
  - f is strictly increasing on $[x^*, \infty)$
- For a unimodal function with maximum at x^*:
  - f is strictly increasing on $(-\infty, x^*]$
  - f is strictly decreasing on $[x^*, \infty)$

Unimodal Functions: Key Properties

Important Properties:

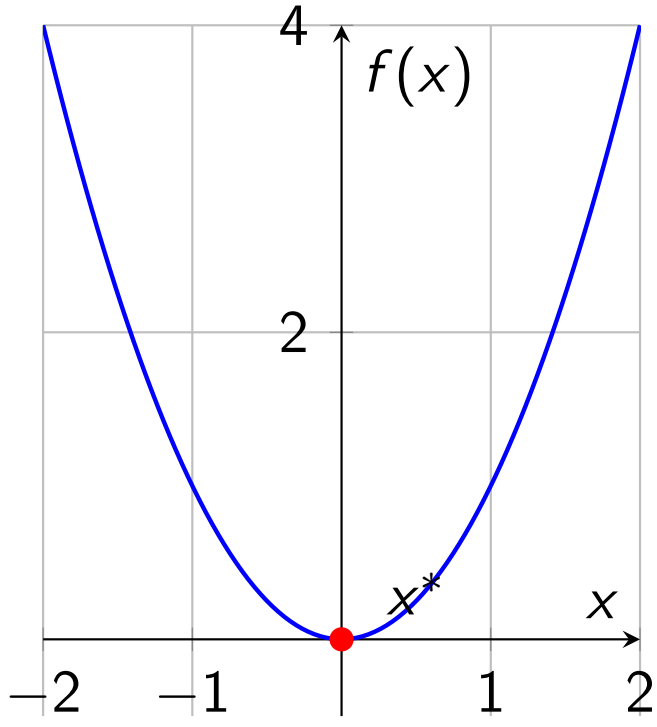
- Any local optimum is also the global optimum
- Optimization is easier: no risk of getting stuck in local optima

Example 1: Quadratic Function (Convex & Unimodal)

$$f(x) = x^2$$

Properties:

- Convex
- Unimodal
- Global minimum at $x^* = 0$
- Strictly decreasing on $(-\infty, 0]$
- Strictly increasing on $[0, \infty)$

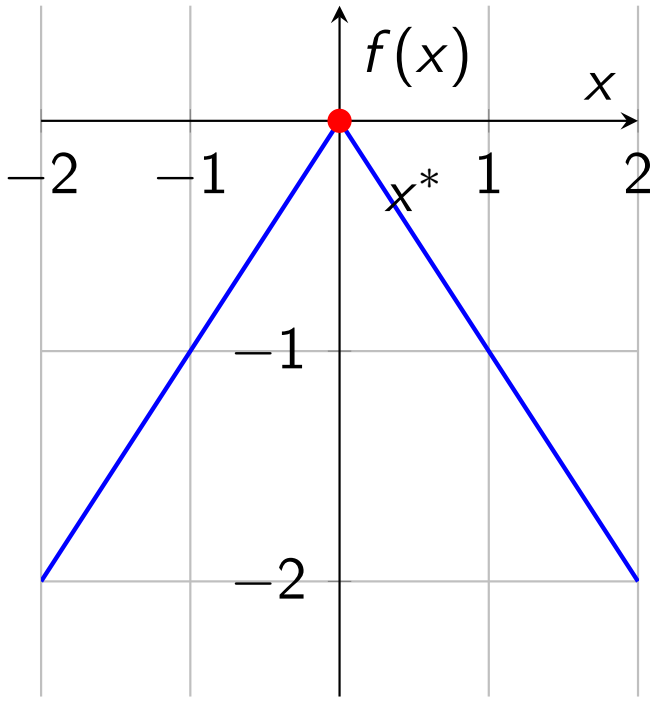


Example 2: Absolute Value

$$f(x) = -|x|$$

Properties:

- NOT convex (concave)
- Unimodal
- Global maximum at $x^* = 0$
- Strictly increasing on $(-\infty, 0]$
- Strictly decreasing on $[0, \infty)$

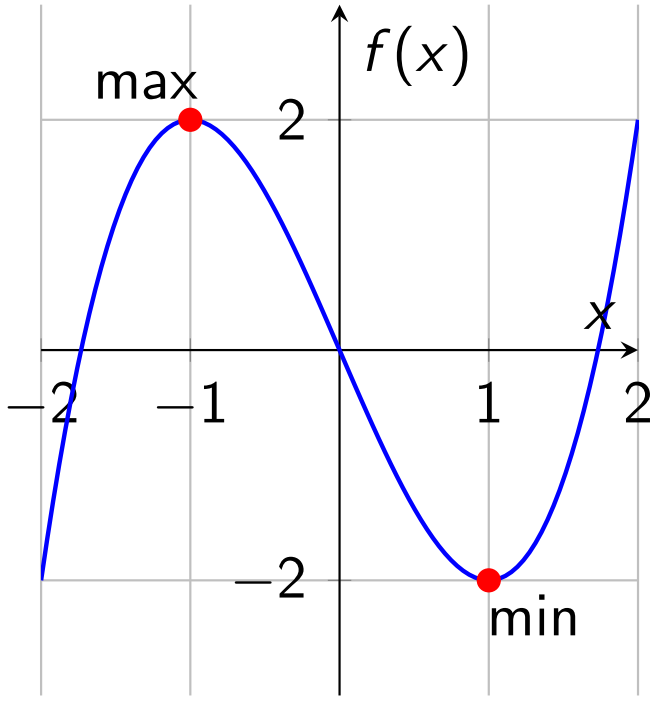


Counter-Example: Non-Unimodal Function

$$f(x) = x^3 - 3x$$

Properties:

- NOT unimodal
- Has two local optima:
  - Local max at $x = -1$
  - Local min at $x = 1$
- Neither is global
- More challenging to optimize



Univariate Optimization: Overview

Definition: Univariate optimization involves finding the minimum or maximum of a function $f : \mathbb{R} \to \mathbb{R}$.
Assumptions: For algorithms discussed here, we assume that f is unimodal.
Common Methods:

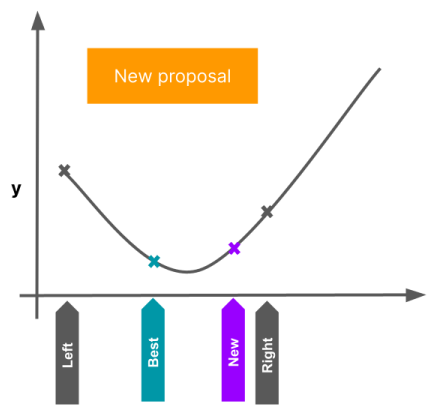
- Golden Section Search
- Brent's Method

Golden section - `https://sketchplanations.com/the-golden-ratio`
Scipy optimize - `https://docs.scipy.org/doc/scipy/reference/optimize.html`

Optimization in Machine Learning

Univariate optimization

Golden ratio



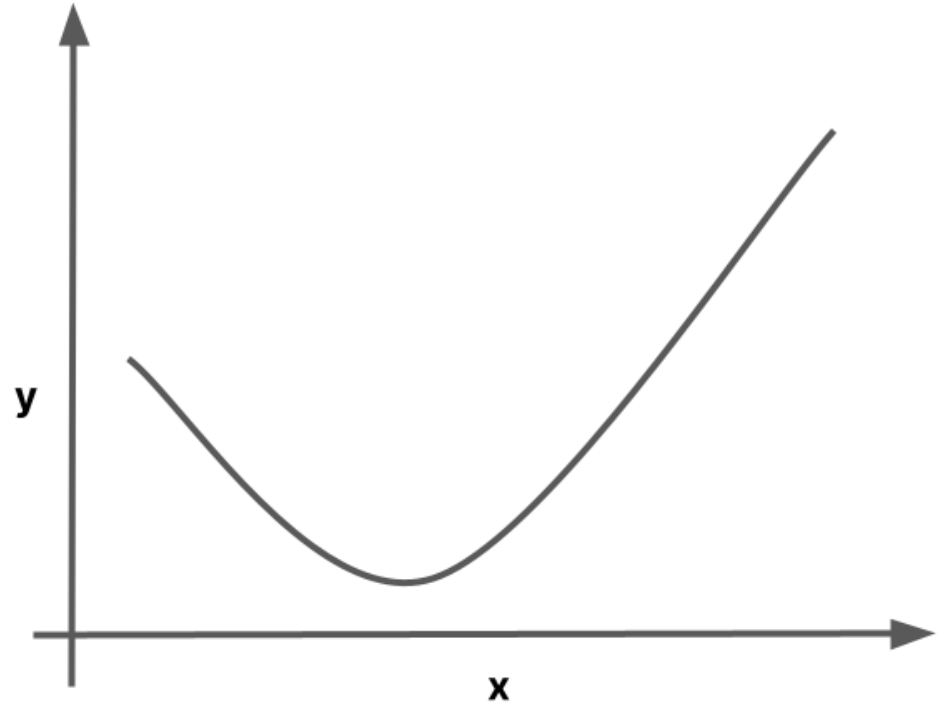
Learning goals

- Simple nesting procedure
- Golden ratio

UNIVARIATE OPTIMIZATION

Let $f : \mathbb{R} \to \mathbb{R}$.

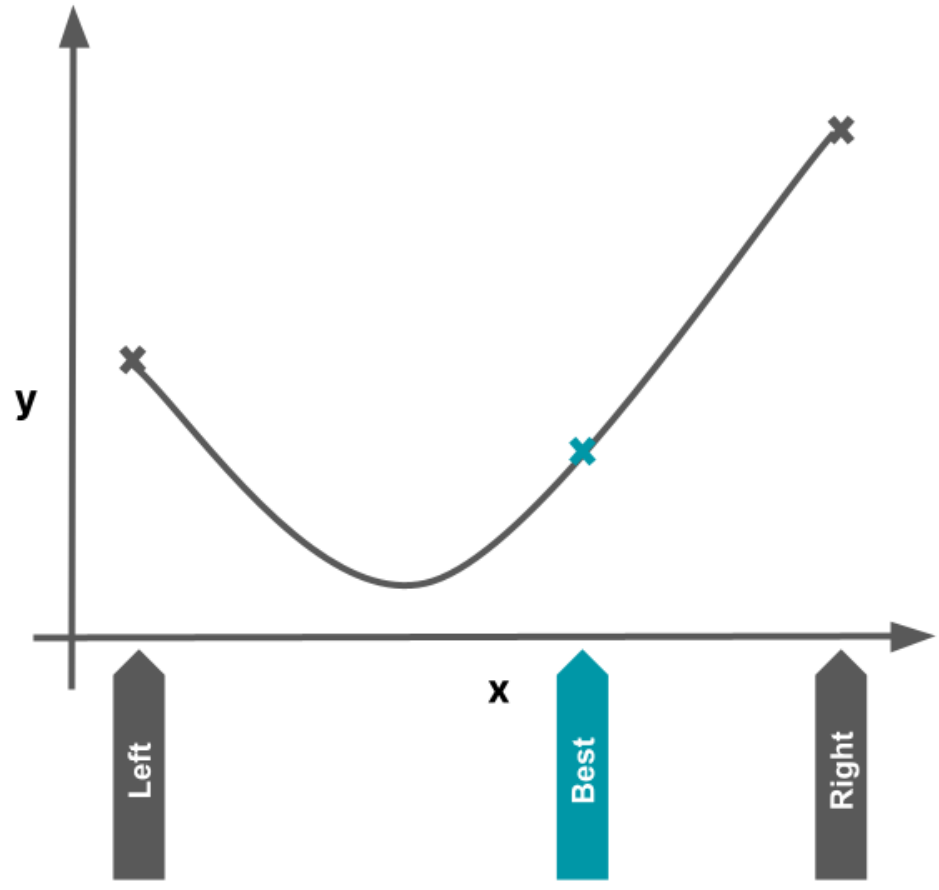
Goal: Iteratively improve eval points. Assume function is unimodal. Will not rely on gradients, so this also works for black-box problems.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

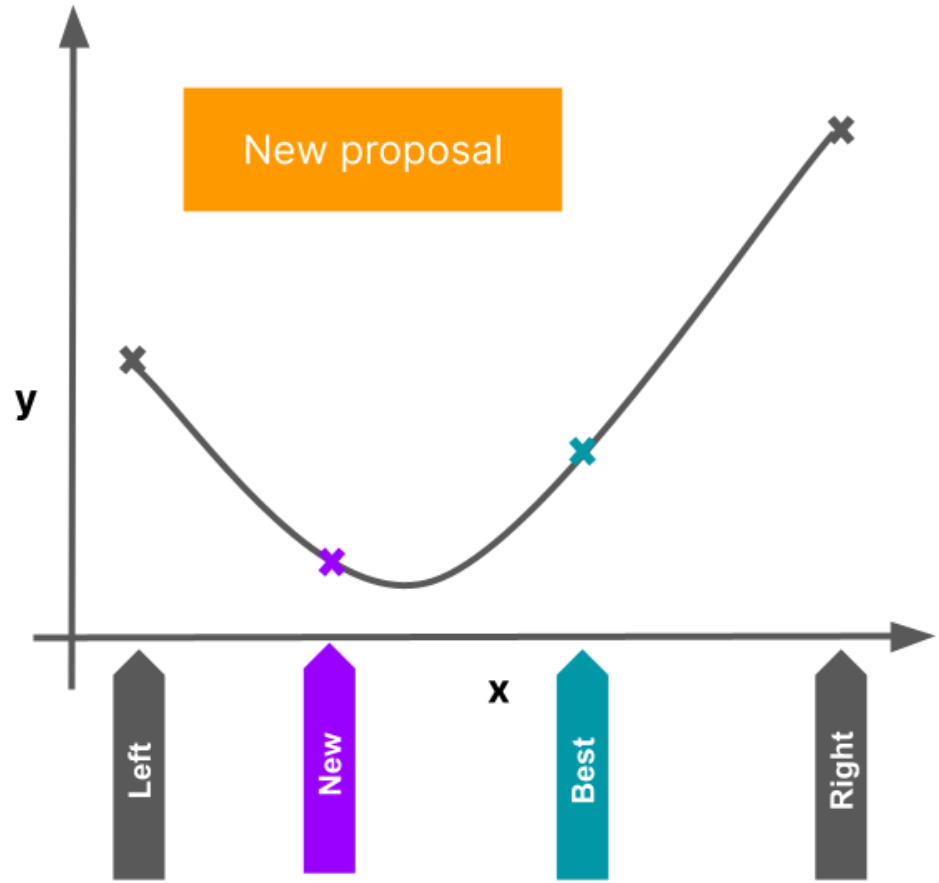
Always maintain three points: left, right, and current best.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

Propose random point in interval.

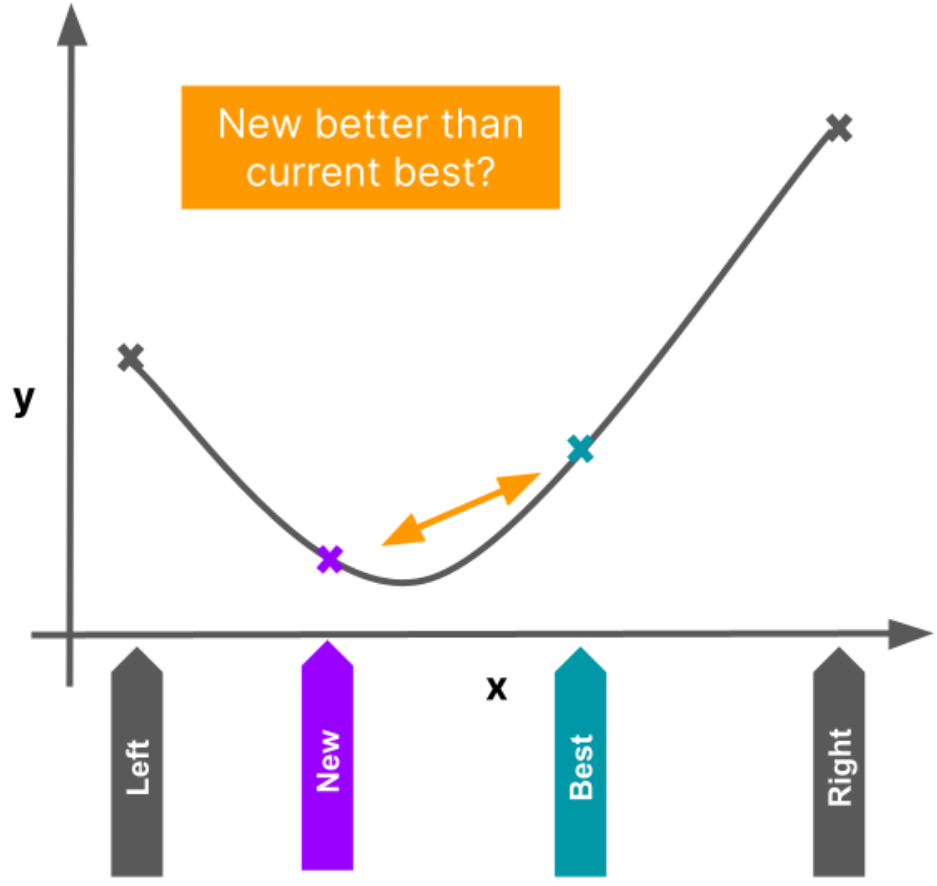


NB: Later we will define the optimal choice for a new proposal.

SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

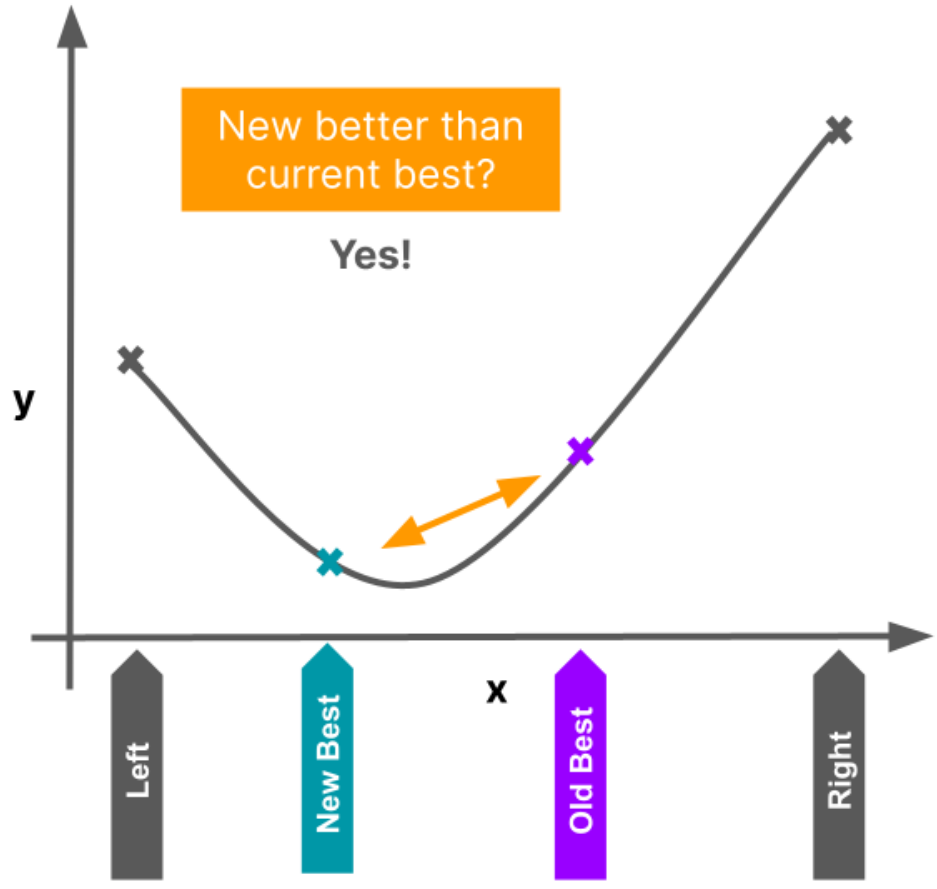
Compare proposal against current best.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

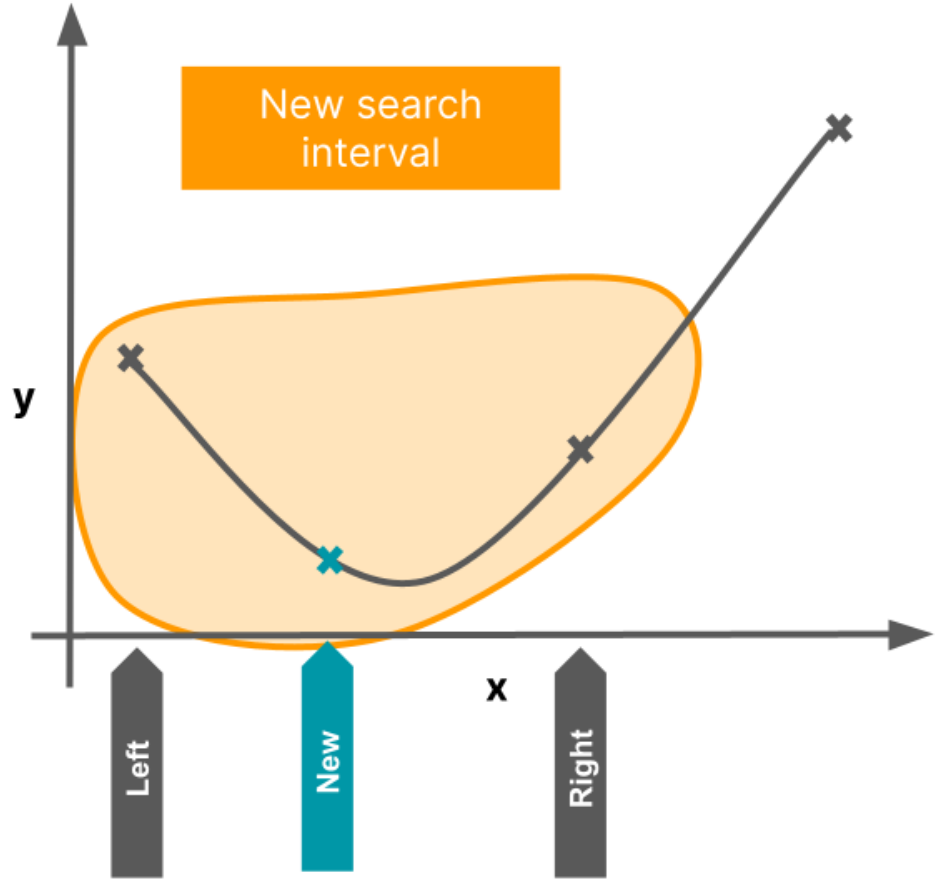
If it is better: proposal becomes current best.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

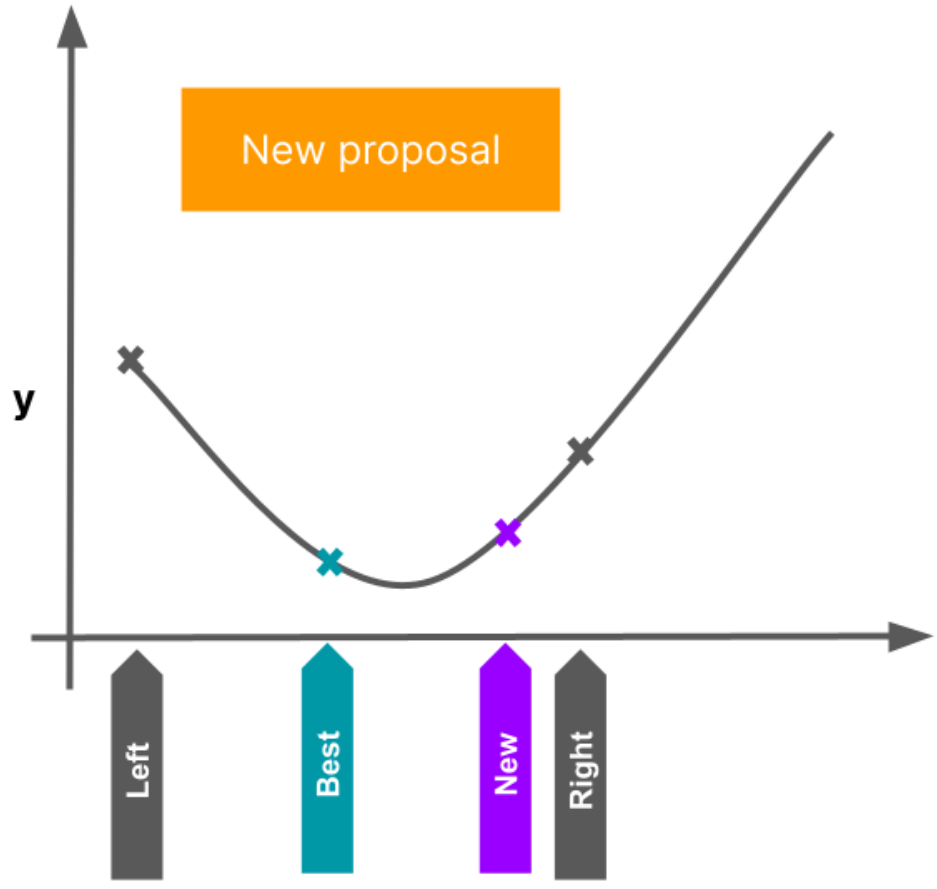
New search interval: around current best.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

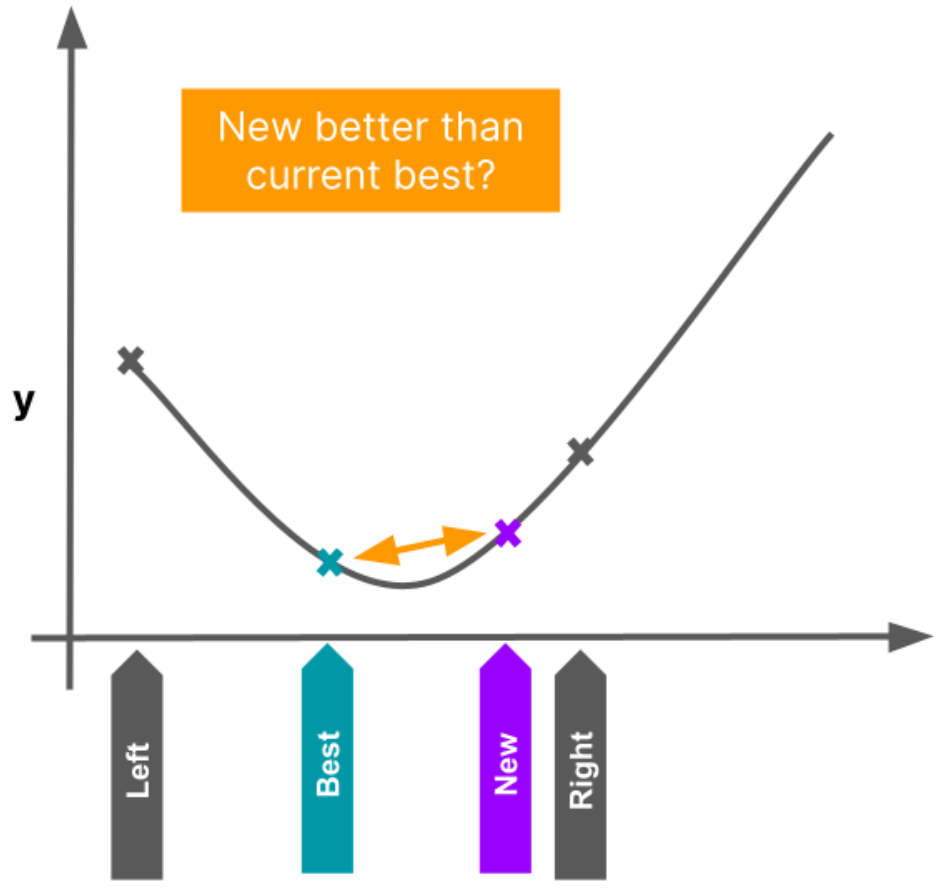
Propose a random point.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

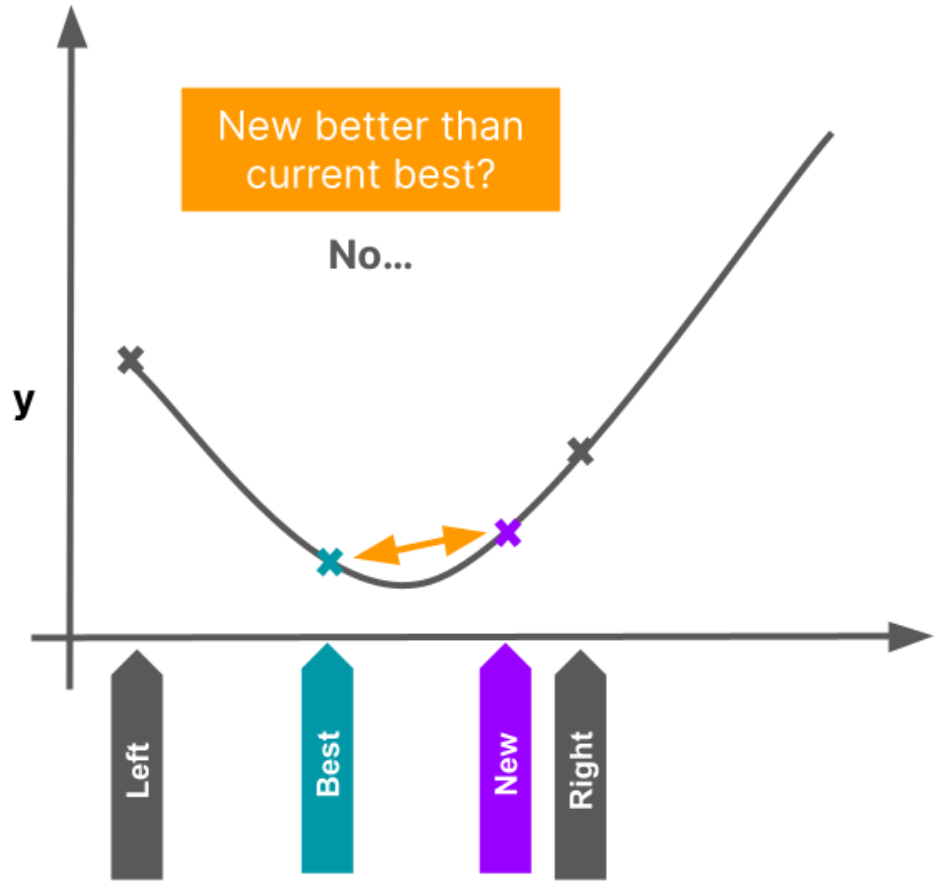
Compare proposal against current best.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

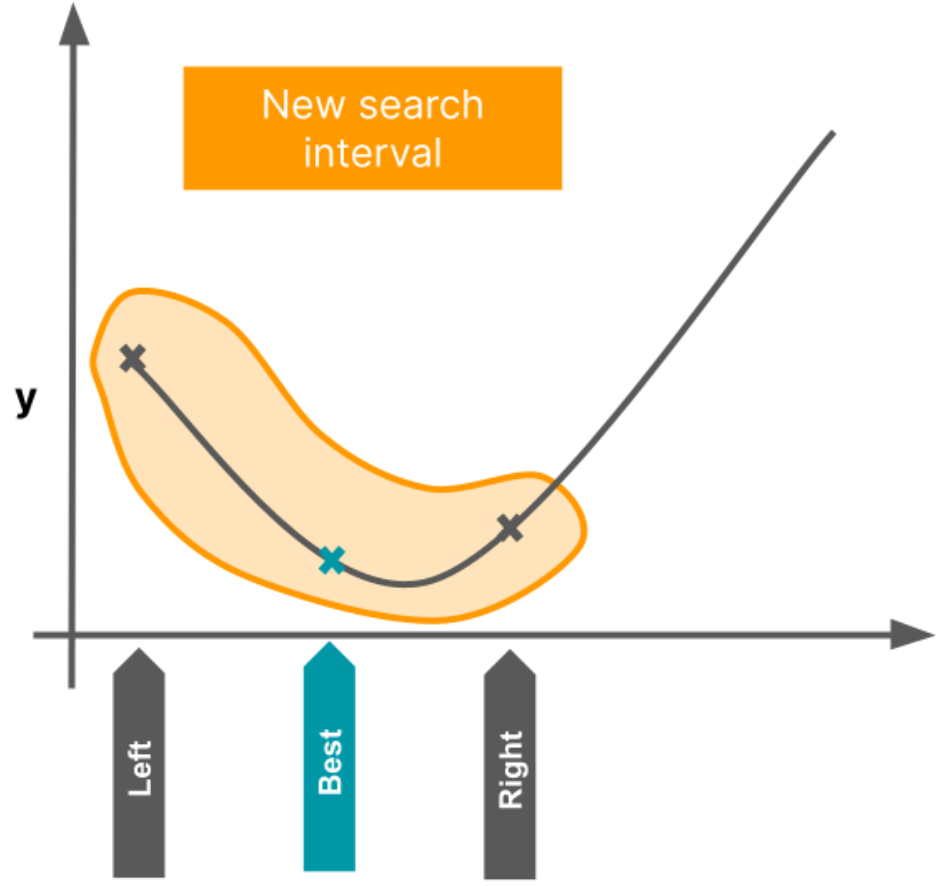
If it is better: proposal becomes current best.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

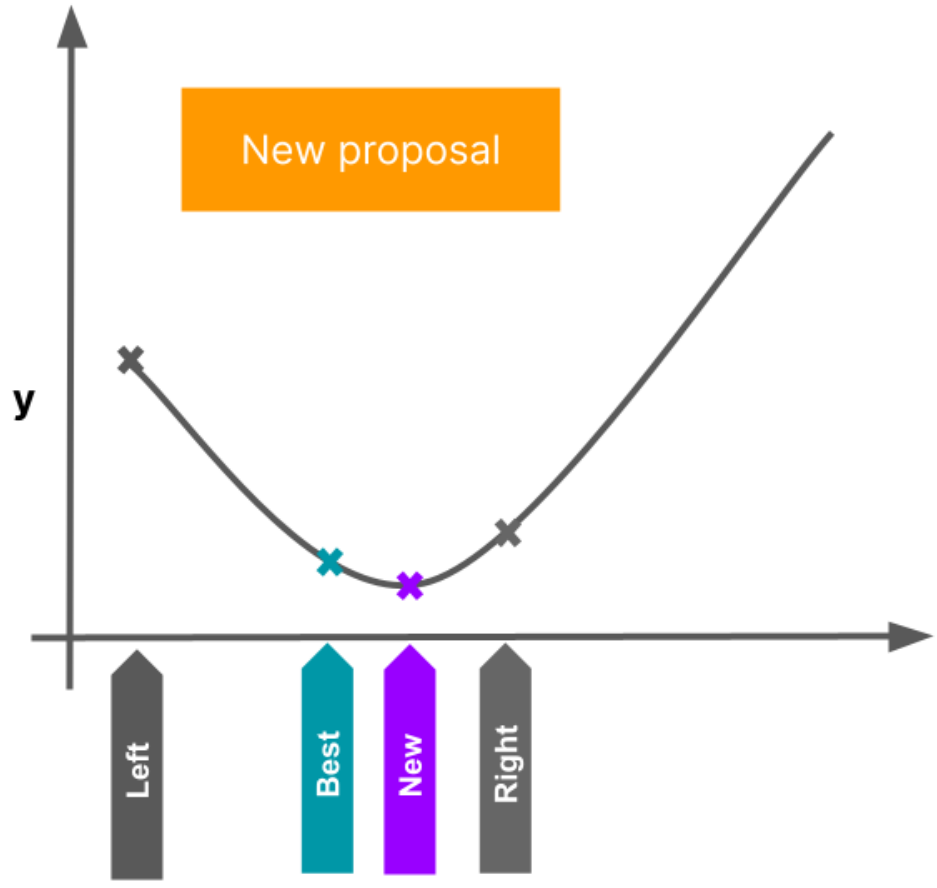
New search interval: around current best.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

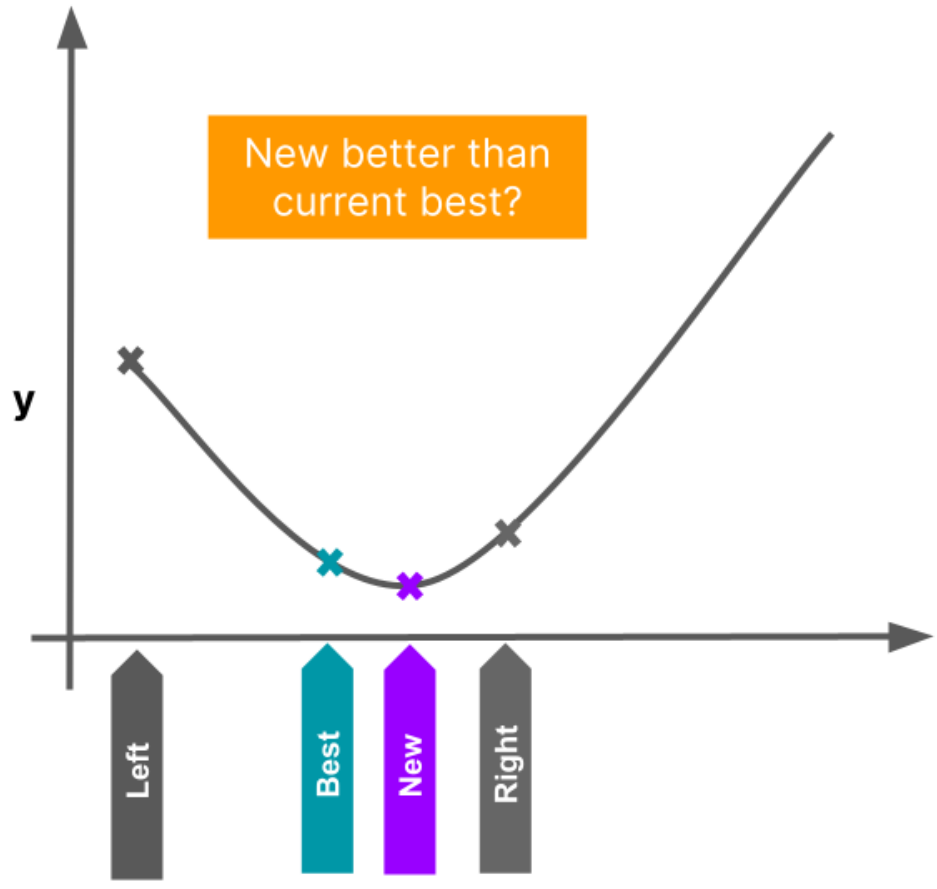
Propose a random point.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

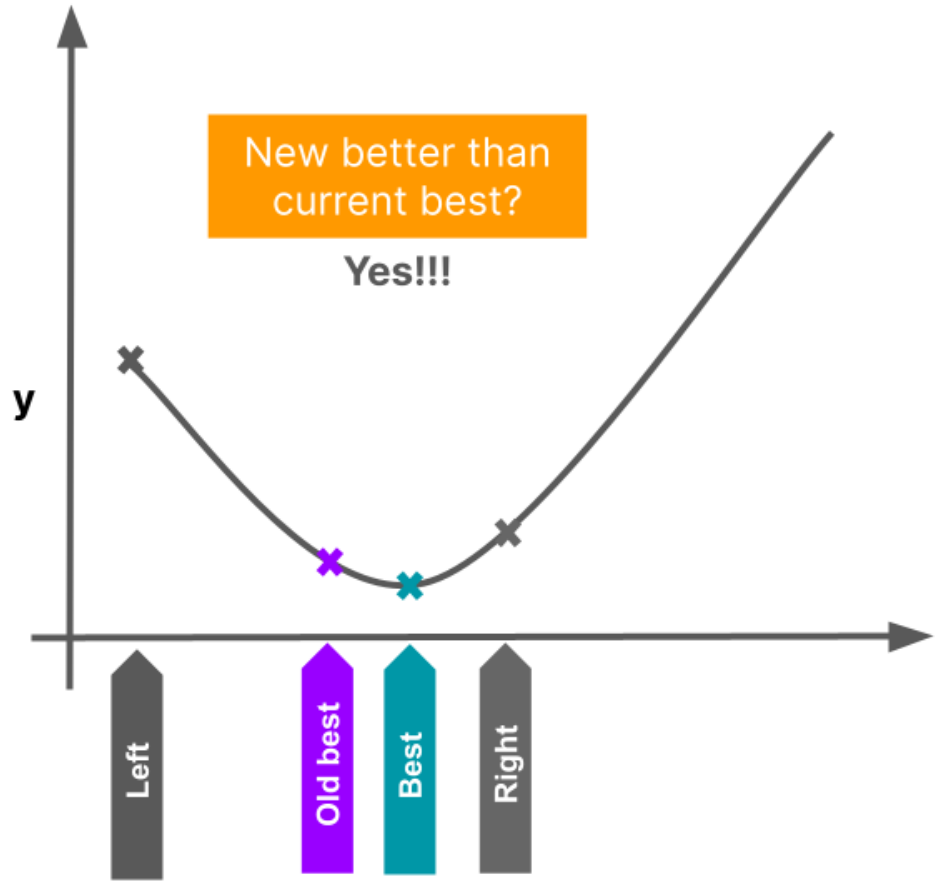
Compare proposal against current best.



SIMPLE NESTING PROCEDURE

Let $f : \mathbb{R} \to \mathbb{R}$.

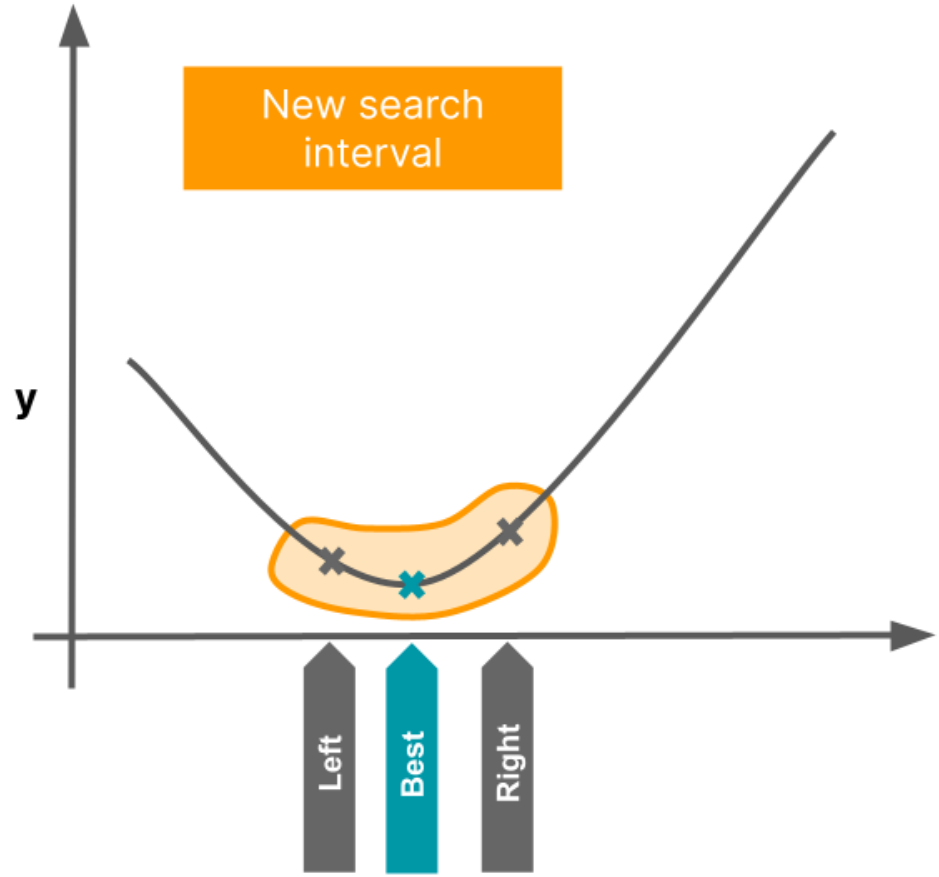
If it is better: proposal becomes current best.



SIMPLE NESTING PROCEDURE

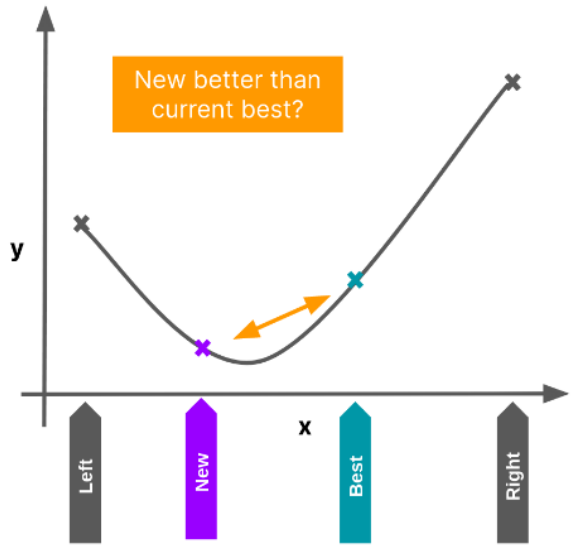
Let $f : \mathbb{R} \to \mathbb{R}$.

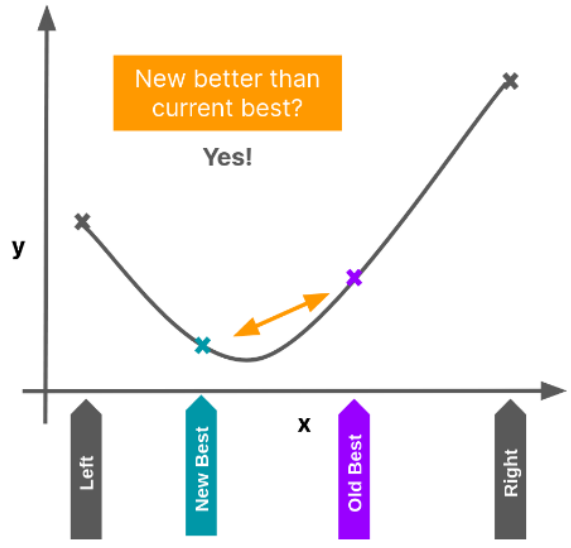
New search interval: around current best.

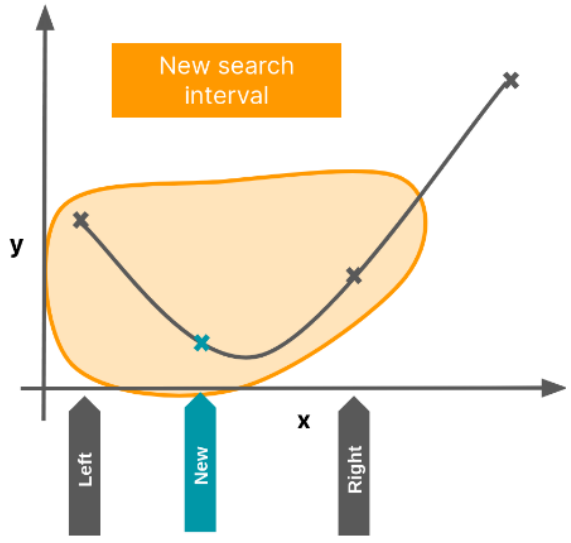


SIMPLE NESTING PROCEDURE

- **Initialization**: Search interval $(x^{\text{left}}, x^{\text{right}})$, $x^{\text{left}} < x^{\text{right}}$
- Choose x^{best} randomly.
- For $t = 0, 1, 2, \ldots$
  - Choose x^{new} randomly in $[x^{\text{left}}, x^{\text{right}}]$
  - If $f(x^{\text{new}}) < f(x^{\text{best}})$:
    - $x^{\text{best}} \leftarrow x^{\text{new}}$
  - New interval: Points around x^{best}



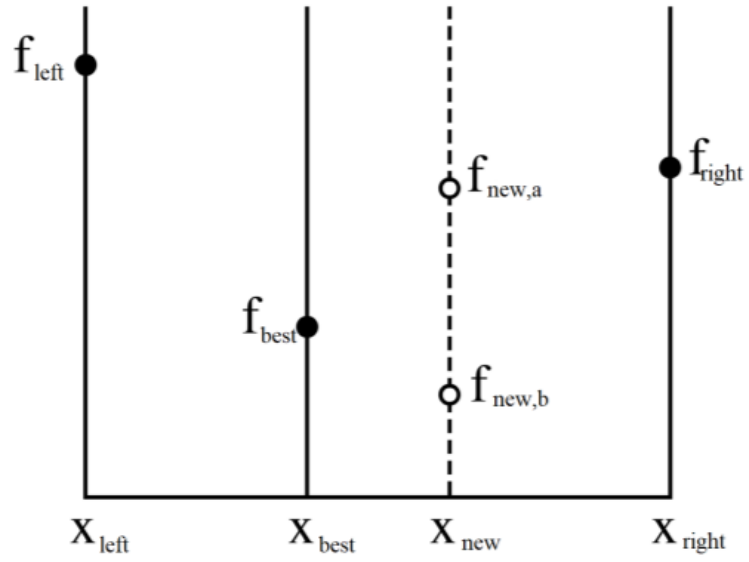




GOLDEN RATIO

Key question: How can x^{new} be chosen better than randomly?

- **Insight 1:** Always in bigger subinterval to maximize reduction.
- **Insight 2:** x^{new} symmetrically to x^{best} for uniform reduction.

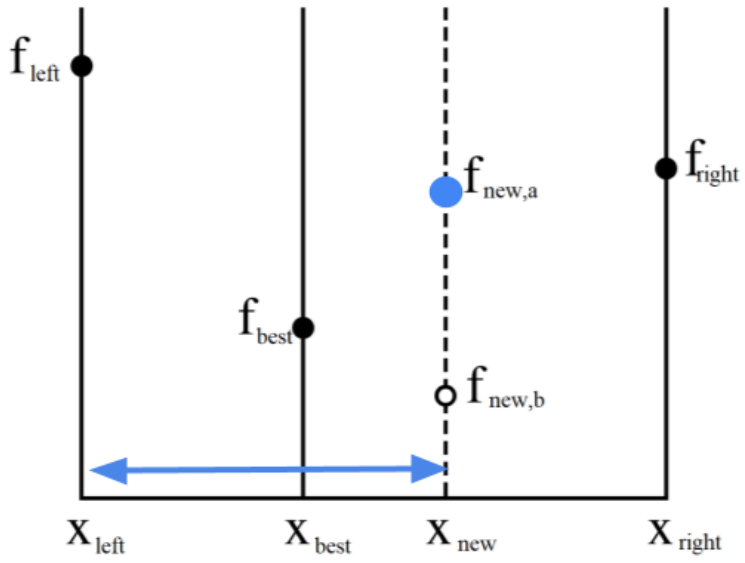


Consider two hypothetical outcomes x^{new}: $f_{new,a}$ and $f_{new,b}$.

GOLDEN RATIO / 2

If $f_{new,a}$ is the outcome, x_{best} stays best and we search around x_{best} :

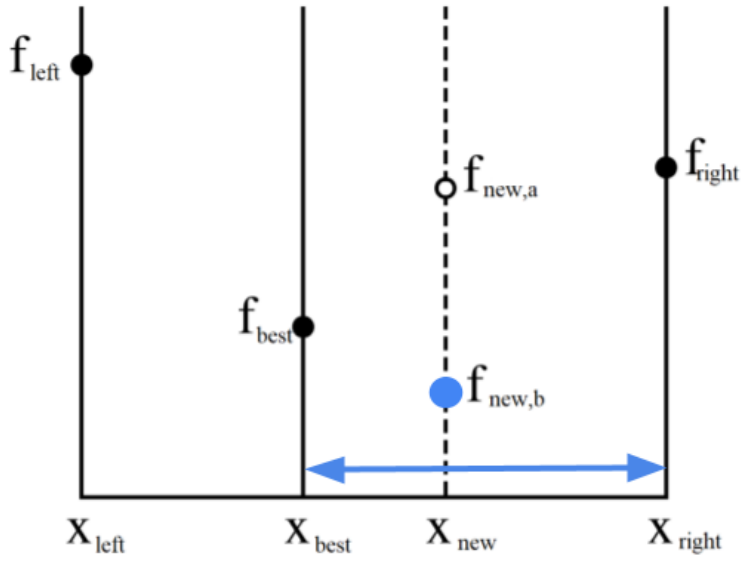
$$[x_{left}, x_{new}]$$



GOLDEN RATIO / 3

If $f_{new,b}$ is outcome, x_{new} becomes best point and search around x_{new} :

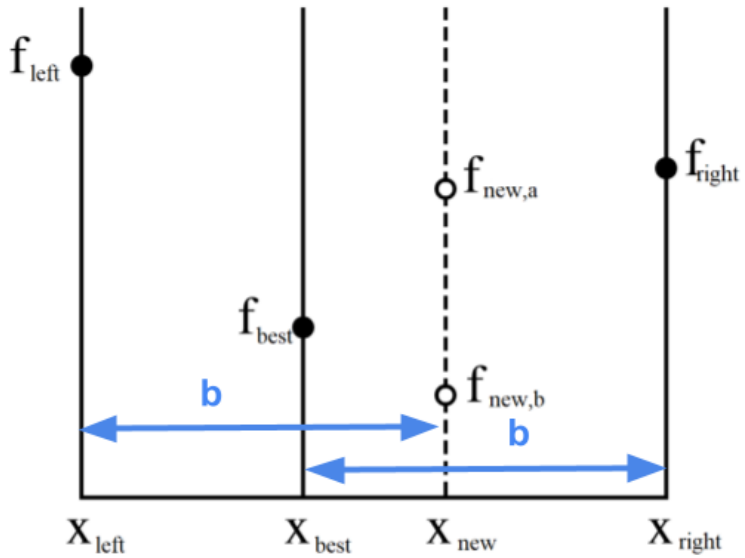
$$[x_{best}, x_{right}]$$



GOLDEN RATIO / 4

For uniform reduction, require the two potential intervals equal sized:

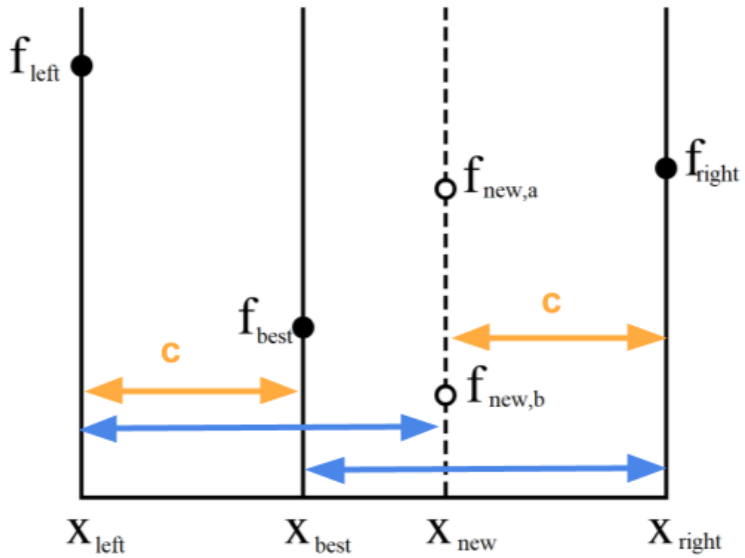
$$b := x_{right} - x_{best} = x_{new} - x_{left}$$



GOLDEN RATIO / 5

One iteration ahead: require again the intervals to be of same size.

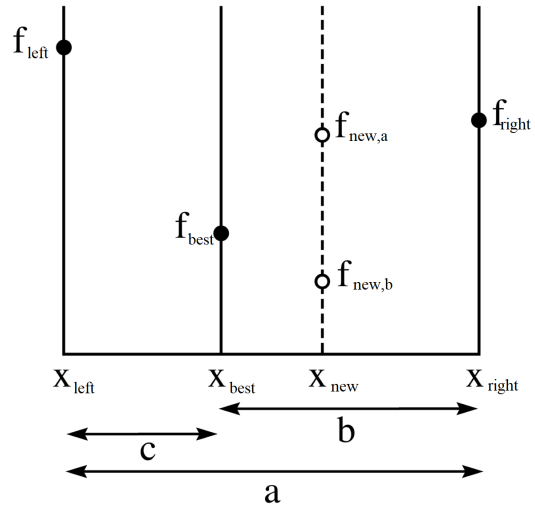
$$c := x_{best} - x_{left} = x_{right} - x_{new}$$



GOLDEN RATIO / 6

To summarize, we require:

$$
\begin{aligned}
a &= x^{right} - x^{left}, \\
b &= x_{right} - x_{best} = x_{new} - x_{left} \\
c &= x_{best} - x_{left} = x_{right} - x_{new}
\end{aligned}
$$



GOLDEN RATIO / 7

- We require the same percentage improvement in each iteration
- For φ reduction factor of interval sizes (a to b, and b to c)

$$\varphi := \frac{b}{a} = \frac{c}{b}$$

$$\varphi^2 = \frac{b}{a} \cdot \frac{c}{b} = \frac{c}{a}$$

- Divide $a = b + c$ by a:

$$\begin{aligned} \frac{a}{a} &= \frac{b}{a} + \frac{c}{a} \\ 1 &= \varphi + \varphi^2 \\ 0 &= \varphi^2 + \varphi - 1 \end{aligned}$$

- Unique positive solution is $\varphi = \frac{\sqrt{5}-1}{2} \approx 0.618$.

GOLDEN RATIO / 8

- With x^{new} we always go φ percentage points into the interval.
- Given x^{left} and x^{right} it follows

$$\begin{aligned} x^{best} &= x^{right} - \varphi(x^{right} - x^{left}) \\ &= x^{left} + (1-\varphi)(x^{right} - x^{left}) \end{aligned}$$

and due to symmetry

$$\begin{aligned} x^{new} &= x^{left} + \varphi(x^{right} - x^{left}) \\ &= x^{right} - (1-\varphi)(x^{right} - x^{left}). \end{aligned}$$

GOLDEN RATIO / 9

Termination criterion:

- A reasonable choice is the absolute error, i.e. the width of the last interval:

$$|x^{best} - x^{new}| < \tau$$

- In practice, more complicated termination criteria are usually applied, for example in *Numerical Recipes in C, 2017*

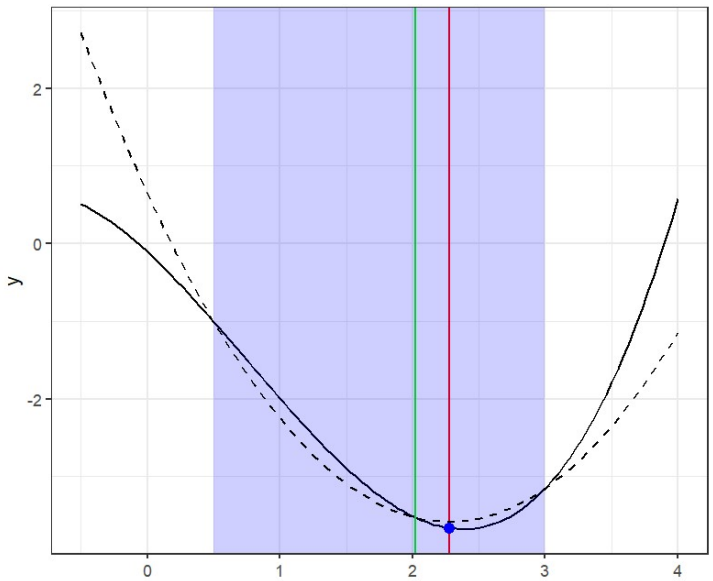
$$|x^{right} - x^{left}| \leq \tau(|x^{best}| + |x^{new}|)$$

  is proposed as a termination criterion.

Optimization in Machine Learning

Univariate optimization

Brent's method



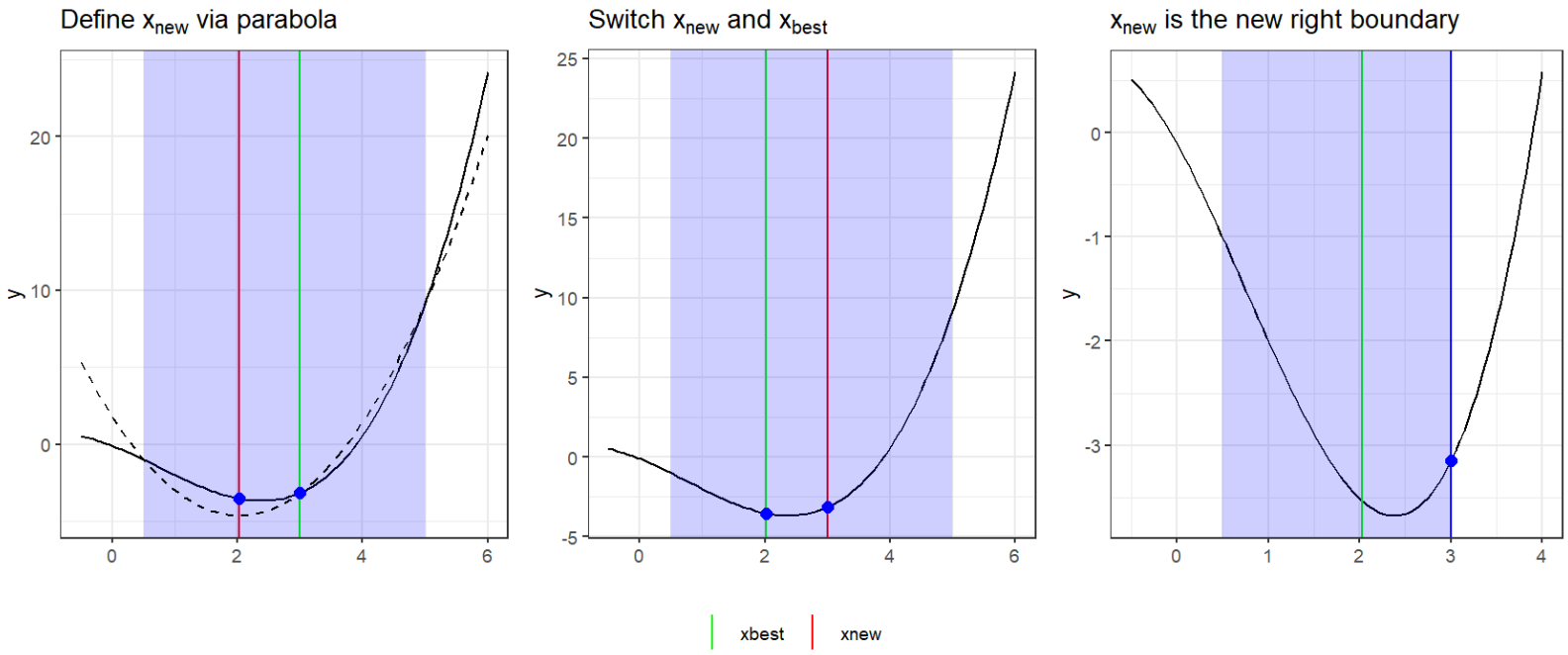
Learning goals

- Quadratic interpolation
- Brent's procedure

QUADRATIC INTERPOLATION

Similar to golden ratio procedure but select x^{new} differently: x^{new} as minimum of a parabola fitted through

$$(x^{\text{left}}, f^{\text{left}}), (x^{\text{best}}, f^{\text{best}}), (x^{\text{right}}, f^{\text{right}}).$$

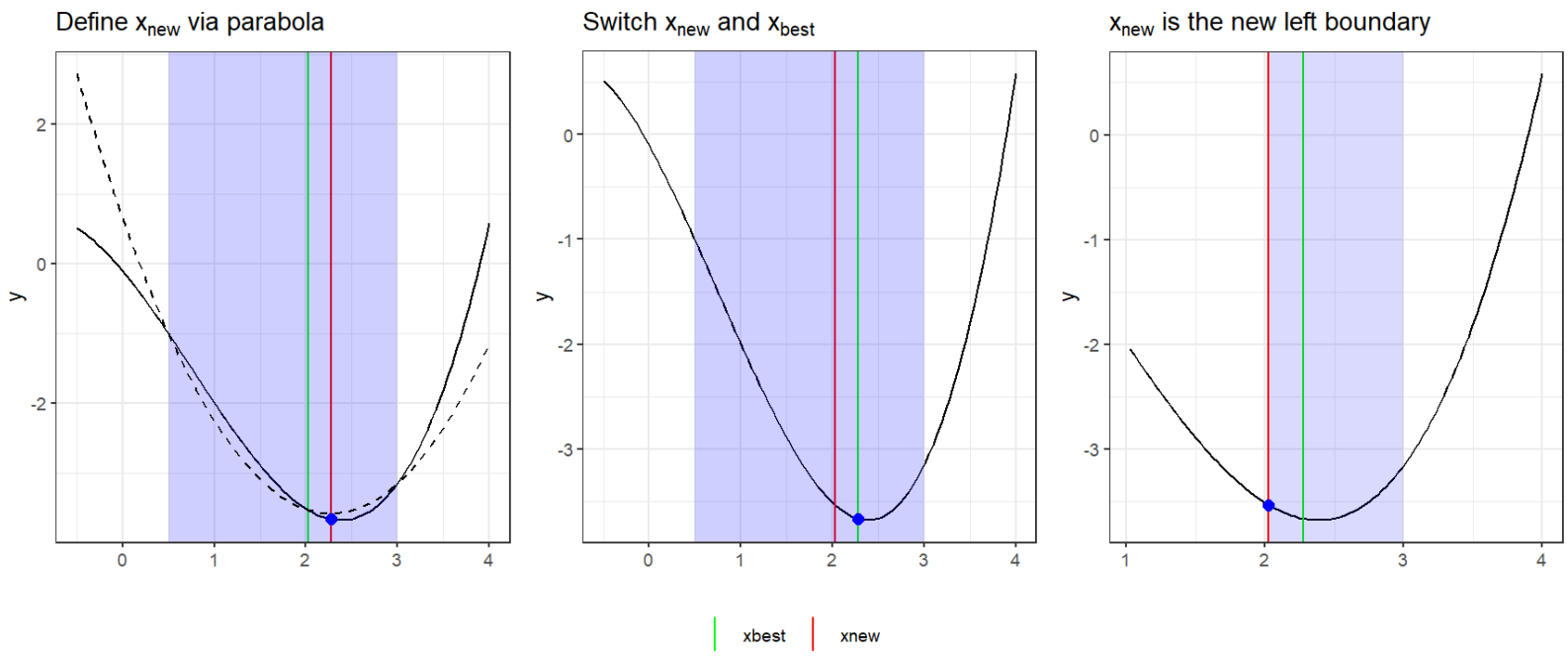


Left: Fit parabola (dashed) and propose minimum (red) as new point. Middle: Switch / not switch with x^{best}. Right: New interval.

QUADRATIC INTERPOLATION

Similar to golden ratio procedure but select x^{new} differently: x^{new} as minimum of a parabola fitted through

$$(x^{\text{left}}, f^{\text{left}}), (x^{\text{best}}, f^{\text{best}}), (x^{\text{right}}, f^{\text{right}}).$$

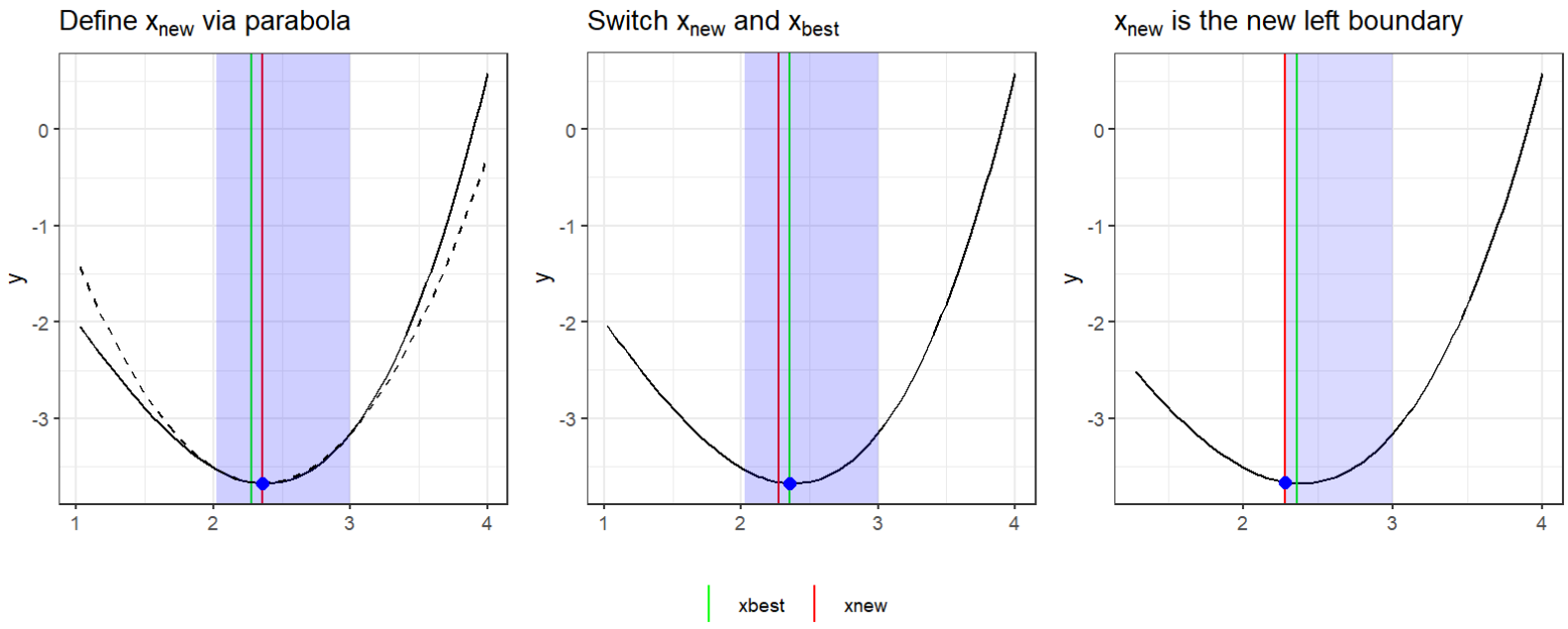


Left: Fit parabola (dashed) and propose minimum (red) as new point. Middle: Switch / not switch with x^{best}. Right: New interval.

QUADRATIC INTERPOLATION

Similar to golden ratio procedure but select x^{new} differently: x^{new} as minimum of a parabola fitted through

$$(x^{\text{left}}, f^{\text{left}}), (x^{\text{best}}, f^{\text{best}}), (x^{\text{right}}, f^{\text{right}}).$$

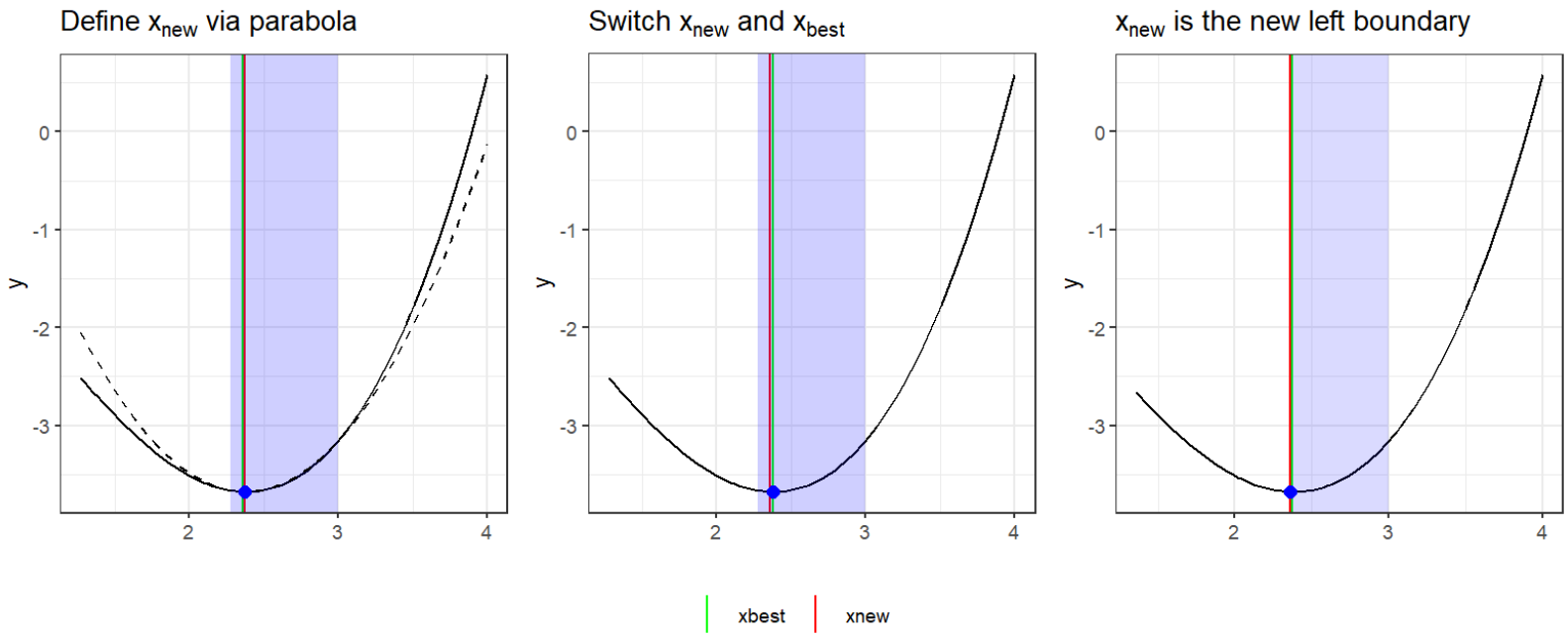


Left: Fit parabola (dashed) and propose minimum (red) as new point. Middle: Switch / not switch with x^{best}. Right: New interval.

QUADRATIC INTERPOLATION

Similar to golden ratio procedure but select x^{new} differently: x^{new} as minimum of a parabola fitted through

$$(x^{\text{left}}, f^{\text{left}}), (x^{\text{best}}, f^{\text{best}}), (x^{\text{right}}, f^{\text{right}}).$$

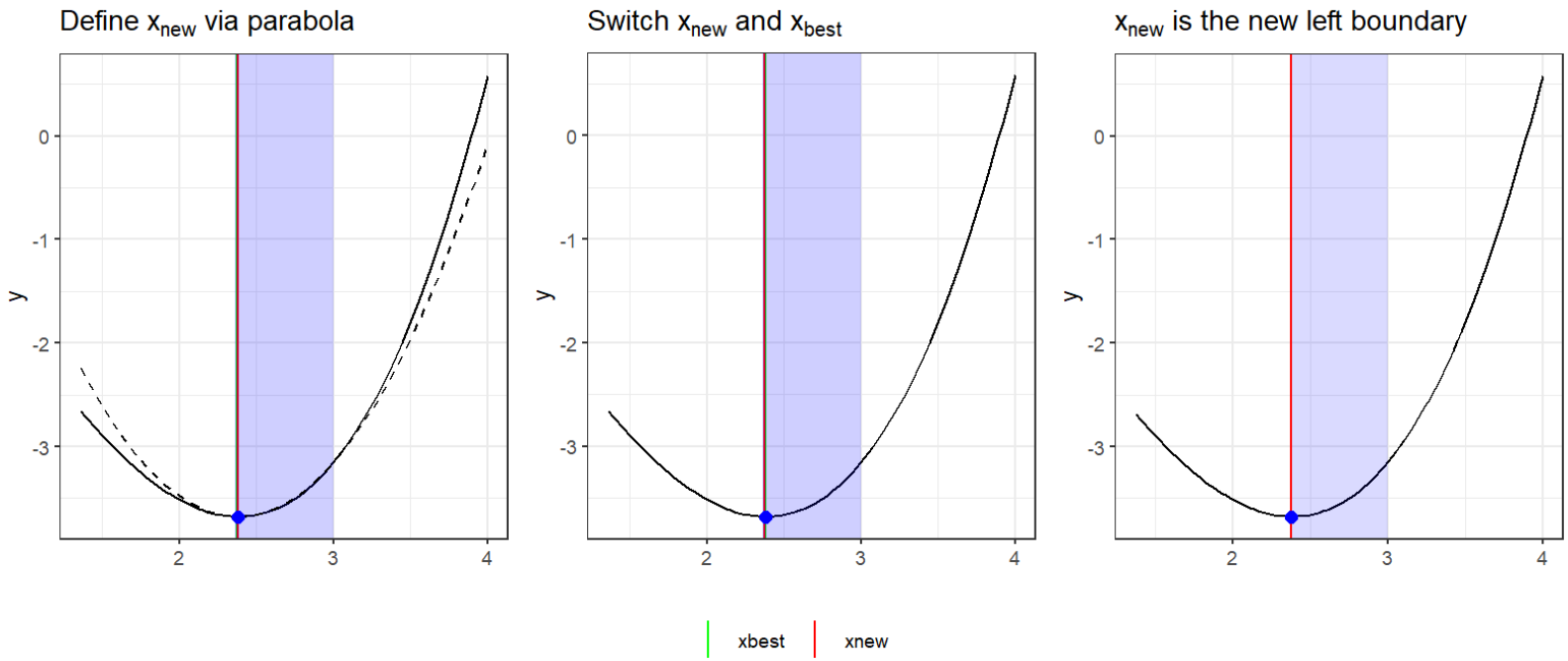


Left: Fit parabola (dashed) and propose minimum (red) as new point. Middle: Switch / not switch with x^{best}. Right: New interval.

QUADRATIC INTERPOLATION

Similar to golden ratio procedure but select x^{new} differently: x^{new} as minimum of a parabola fitted through

$$(x^{\text{left}}, f^{\text{left}}), (x^{\text{best}}, f^{\text{best}}), (x^{\text{right}}, f^{\text{right}}).$$



Left: Fit parabola (dashed) and propose minimum (red) as new point. Middle: Switch / not switch with x^{best}. Right: New interval.

QUADRATIC INTERPOLATION COMMENTS

- Quadratic interpolation **not robust**. The following may happen:
  - Algorithm suggests the same x^{new} in each step,
  - x^{new} outside of search interval,
  - Parabola degenerates to line and no real minimum exists
- Algorithm must then abort, finding a global minimum is not guaranteed.

BRENT'S METHOD

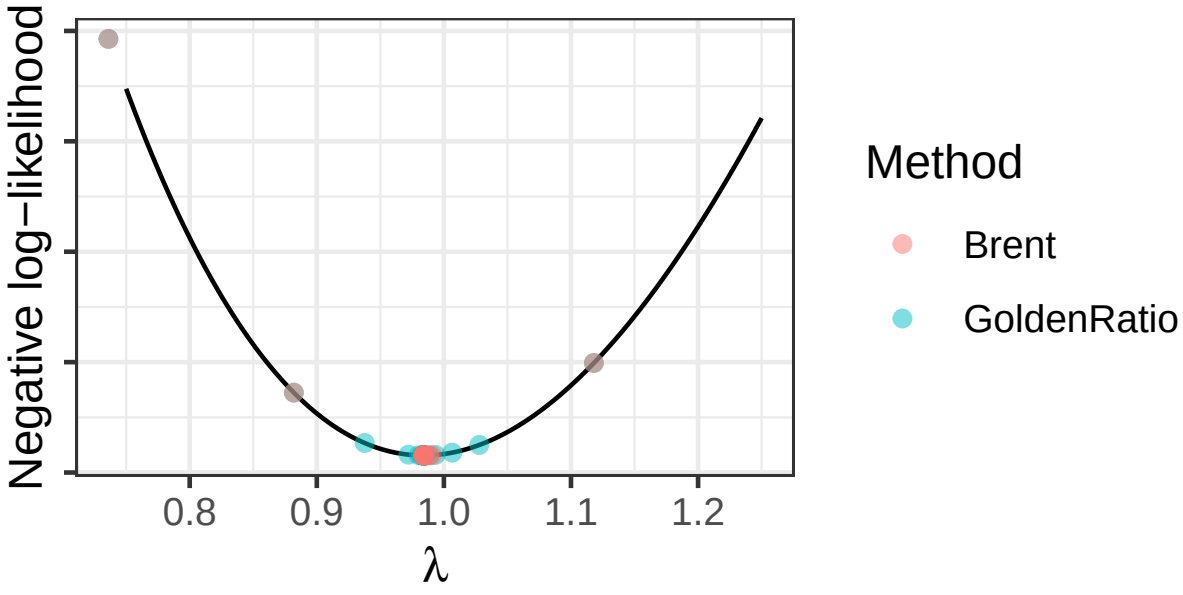
- Brent proposed an algorithm (1973) that alternates between golden ratio search and quadratic interpolation as follows:
  - Quadratic interpolation step acceptable if: (i) x^{new} falls within $[x^{\text{left}}, x^{\text{right}}]$ (ii) x^{new} sufficiently far away from x^{best}

    (Heuristic: Less than half of movement of step before last)
  - Otherwise: Proposal via golden ratio
- Benefit: Fast convergence (quadratic interpolation), unstable steps (e.g. parabola degenerated) stabilized by golden ratio search
- Convergence guaranteed if the function f has a local minimum
- Used in R-function `optimize()`

EXAMPLE: MLE POISSON

- Poisson density: $f(k \mid \lambda) := \mathbb{P}(x = k) = \frac{\lambda^k \cdot \exp(-\lambda)}{k!}$
- Negative log-likelihood for n observations:

$$-\ell(\lambda, \mathcal{D}) = -\log \prod_{i=1}^{n} f\left(x^{(i)} \mid \lambda\right) = -\sum_{i=1}^{n} \log f\left(x^{(i)} \mid \lambda\right)$$



GR and Brent converge to minimum at $x^* \approx 1$.
But: GR needs $\approx$ 45 it., Brent only needs $\approx$ 15 it. for same tolerance.

Why Stopping Criteria?

Challenge: Optimization algorithms are iterative and theoretically converge to the optimum as iterations $\to \infty$.

In Practice:

- We cannot run algorithms forever
- Need to decide when to stop
- Balance between accuracy and computational cost
- Different criteria for different problems

Goal: Find a "good enough" solution in reasonable time.

Types of Stopping Criteria

1. **Absolute Function Value Change**

$$|f(x^{(k+1)}) - f(x^{(k)})| < \varepsilon_f$$

2. **Relative Function Value Change**

$$\frac{|f(x^{(k+1)}) - f(x^{(k)})|}{|f(x^{(k)})| + \delta} < \varepsilon_f$$

3. **Absolute Parameter Change**

$$|x^{(k+1)} - x^{(k)}| < \varepsilon_x$$

4. **Gradient Norm** (if available)

$$\|f'(x^{(k)})\| < \varepsilon_g$$

5. **Maximum Iterations**

$$k > k_{\text{max}}$$

where k is the iteration number and ε values are tolerance thresholds

Absolute vs Relative Criteria

Absolute Criteria:

$$|f(x^{(k+1)}) - f(x^{(k)})| < \varepsilon$$

Pros:

- Simple to implement
- Direct interpretation

Cons:

- Scale-dependent
- Same ε may be too tight for small values, too loose for large values

Relative Criteria:

$$\frac{|f(x^{(k+1)}) - f(x^{(k)})|}{|f(x^{(k)})| + \delta} < \varepsilon$$

Pros:

- Scale-invariant
- Works across different magnitudes

Cons:

- Slightly more complex
- Need to handle $f(x) \approx 0$
- (δ is a small constant)

Gradient-Based Stopping Criteria

For differentiable functions, we can use gradient information:

First-Order Optimality Condition:

$$\|f'(x^{(k)})\| < \varepsilon_g$$

At a local optimum: $f'(x^*) = 0$

Advantages:

- Directly related to optimality conditions
- Works well near the optimum
- Independent of function scale (if normalized)

Disadvantages:

- Requires gradient computation
- May be slow near saddle points
- Can give false positives at local optima

Practical Considerations

Common Practice: Use multiple criteria simultaneously

$$\text{STOP if } \begin{cases} |f(x^{(k+1)}) - f(x^{(k)})| < \varepsilon_f & \text{AND} \\ |x^{(k+1)} - x^{(k)}| < \varepsilon_x & \text{OR} \\ k > k_{\max} \end{cases}$$

Typical Values:

- $\varepsilon_f = 10^{-6}$ to 10^{-8} (function tolerance)
- $\varepsilon_x = 10^{-6}$ to 10^{-8} (parameter tolerance)
- $\varepsilon_g = 10^{-5}$ to 10^{-7} (gradient tolerance)
- $k_{\max} = 100$ to 10000 (depends on problem)

Note: Values depend heavily on problem scale and required precision!