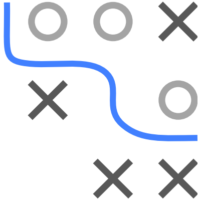
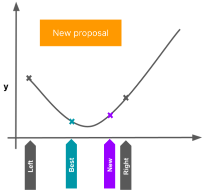
Optimization in Machine Learning



Univariate optimization
Golden ratio



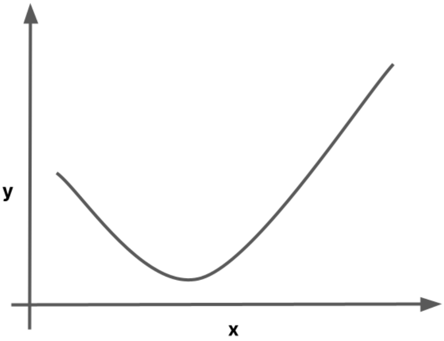
Learning goals

- Simple nesting procedure
- Golden ratio

UNIVARIATE OPTIMIZATION

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

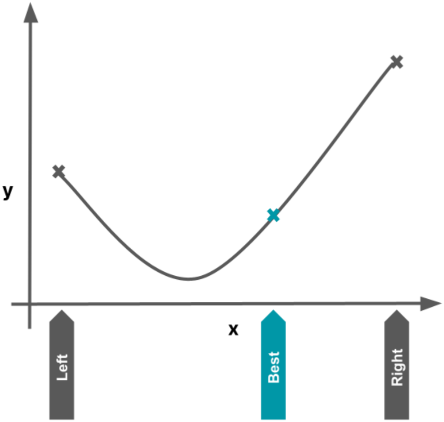
Goal: Iteratively improve eval points. Assume function is unimodal. Will not rely on gradients, so this also works for black-box problems.

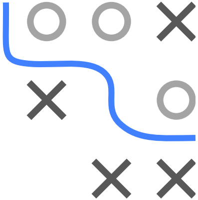


SIMPLE NESTING PROCEDURE

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

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

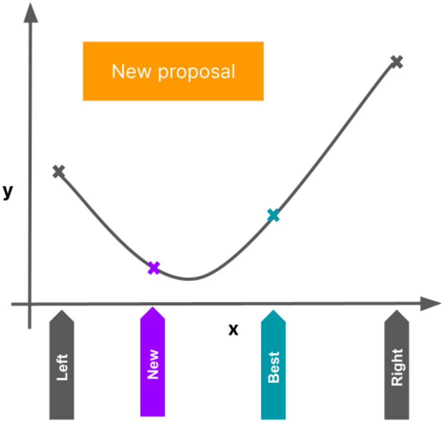




SIMPLE NESTING PROCEDURE

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

Propose random point in interval.

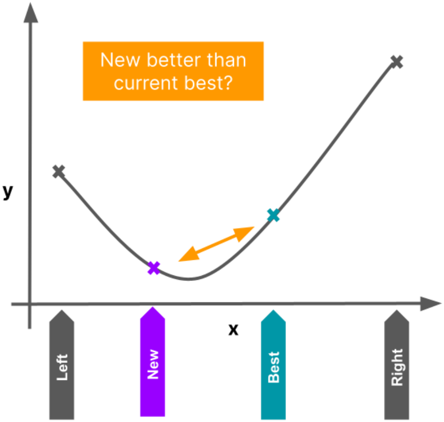


NB: Later we will define the optimal choice for a new proposal.

SIMPLE NESTING PROCEDURE

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

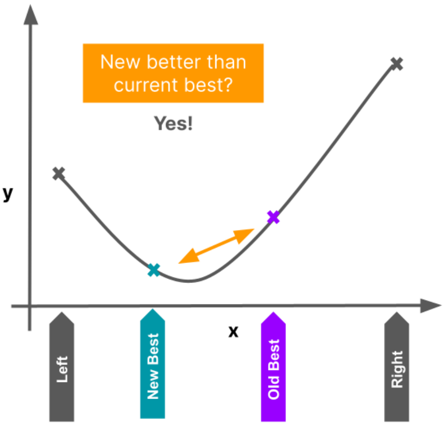
Compare proposal against current best.

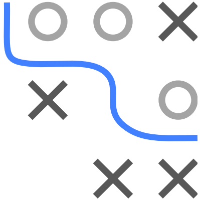


SIMPLE NESTING PROCEDURE

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

If it is better: proposal becomes current best.

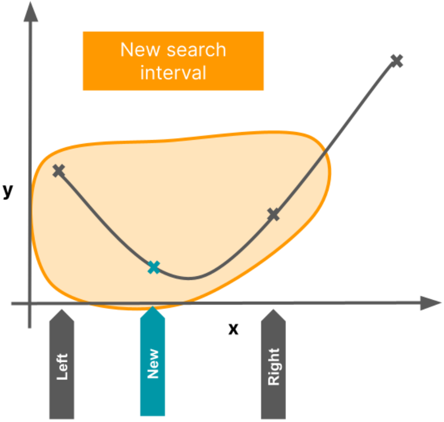


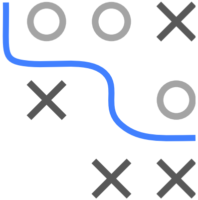


SIMPLE NESTING PROCEDURE

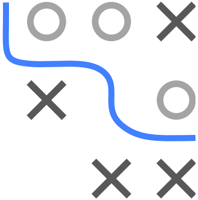
Let $f : \mathbb{R} \to \mathbb{R}$.

New search interval: around current best.



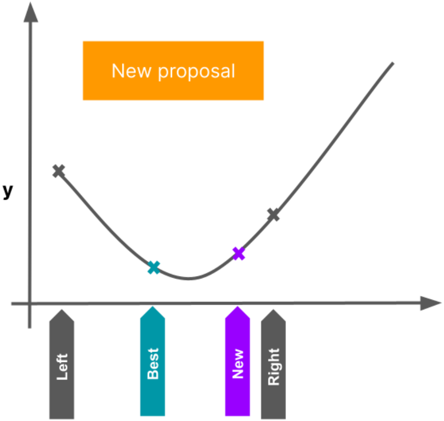


SIMPLE NESTING PROCEDURE

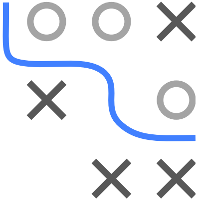


Let $f : \mathbb{R} \to \mathbb{R}$.

Propose a random point.

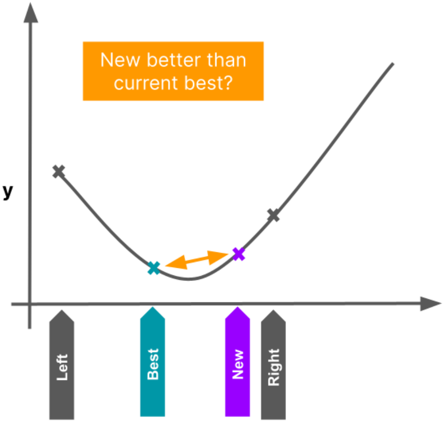


SIMPLE NESTING PROCEDURE



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

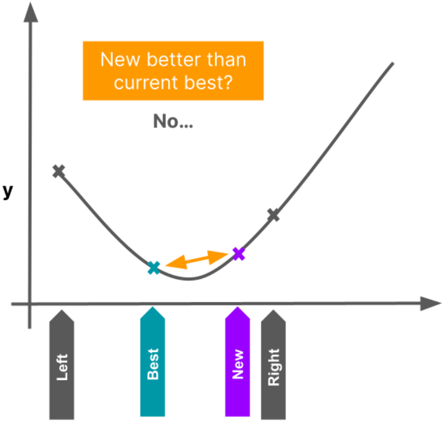
Compare proposal against current best.

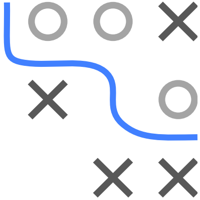


SIMPLE NESTING PROCEDURE

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

If it is better: proposal becomes current best.

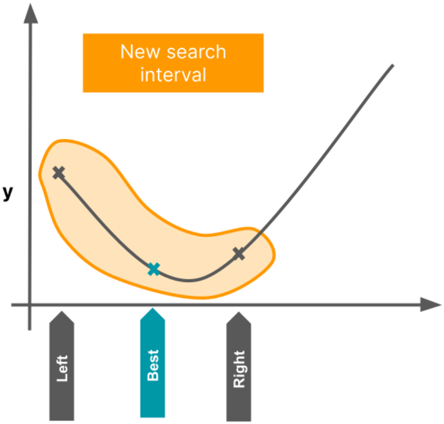




SIMPLE NESTING PROCEDURE

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

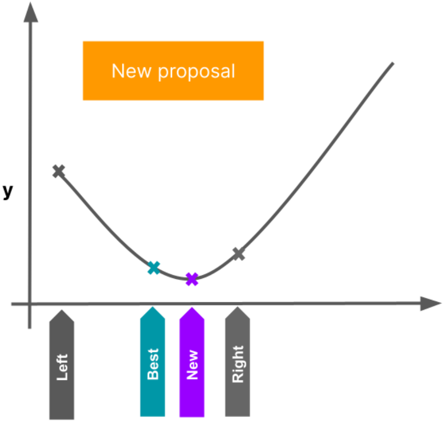
New search interval: around current best.



SIMPLE NESTING PROCEDURE

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

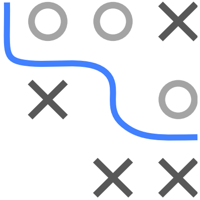
Propose a random point.



SIMPLE NESTING PROCEDURE

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

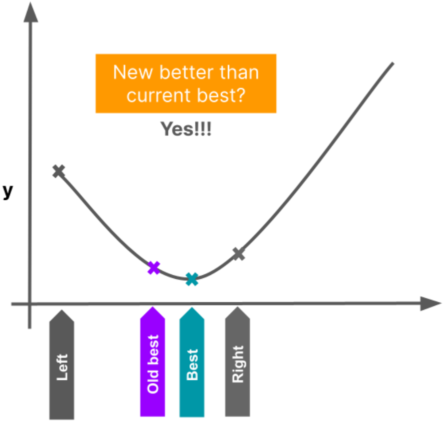
Compare proposal against current best.

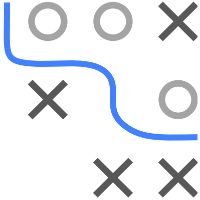


SIMPLE NESTING PROCEDURE

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

If it is better: proposal becomes current best.

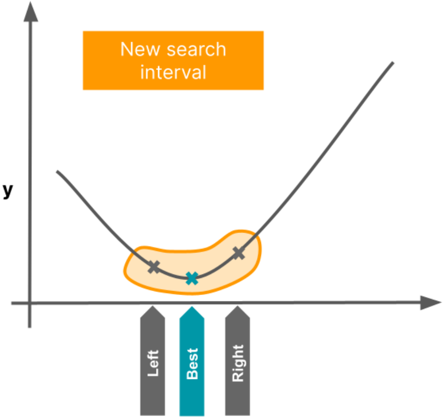


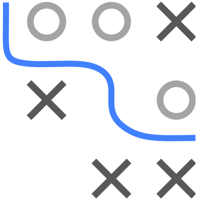


SIMPLE NESTING PROCEDURE

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

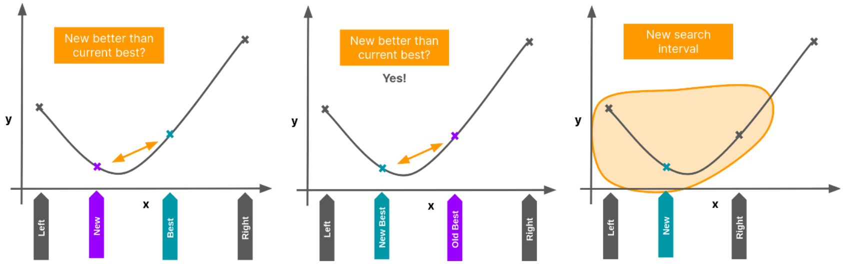
New search interval: around current best.





SIMPLE NESTING PROCEDURE

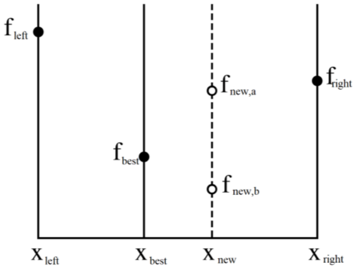
- **Initialization**: Search interval $(x^{\text{left}}, x^{\text{right}})$, $x^{\text{left}} < x^{\text{right}}$
- Choose x^{best} randomly.
- For $t = 0, 1, 2, \ldots$
  - Choose x^{new} randomly in $[x^{\text{left}}, x^{\text{right}}]$
  - If $f(x^{\text{new}}) < f(x^{\text{best}})$:
    - $x^{\text{best}}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

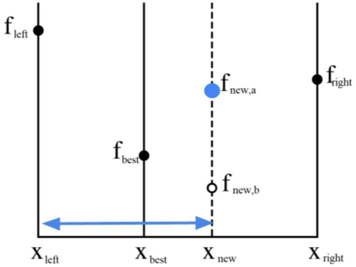


Consider two hypothetical outcomes x^{new}: $f_{new,a}$ and $f_{new,b}$.

GOLDEN RATIO / 2

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

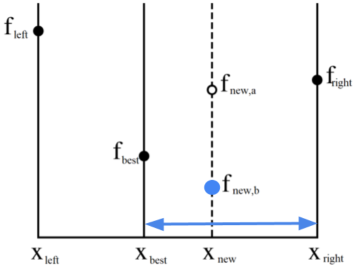
$$[x_{left}, x_{new}]$$



GOLDEN RATIO / 3

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

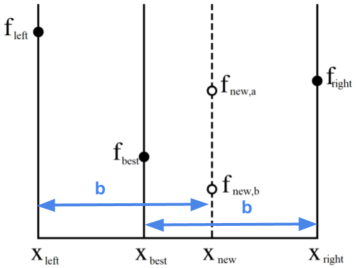
$$[x_{best}, x_{right}]$$



GOLDEN RATIO / 4

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

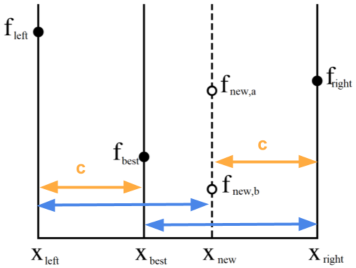
$$b := x_{right} - x_{best} = x_{new} - x_{left}$$



GOLDEN RATIO / 5

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

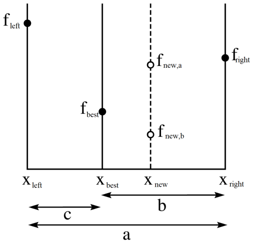
$$c := x_{best} - x_{left} = x_{right} - x_{new}$$



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:

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

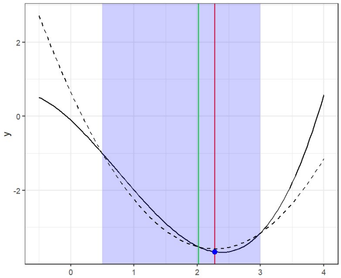
- In practice, more complicated termination criteria are usually applied, for example in *Numerical Recipes in C, 2017*

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

is proposed as a termination criterion.

Optimization in Machine Learning

Univariate optimization
Brent's method



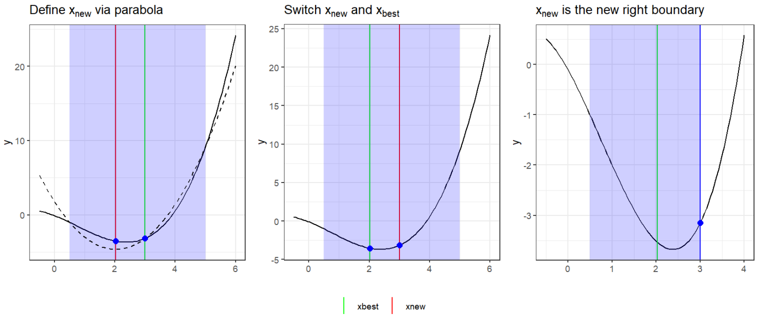
Learning goals

- 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

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

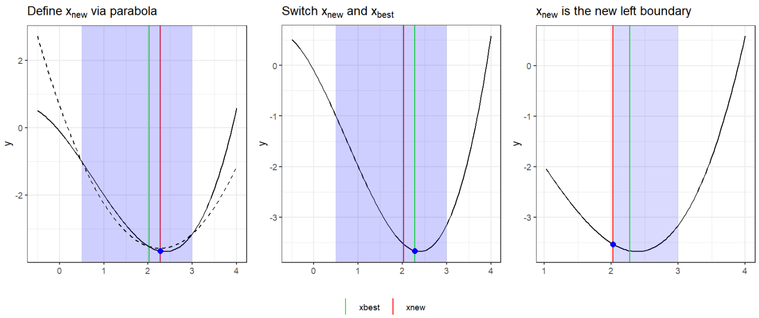


Left: Fit parabola (dashed) and propose minimum (red) as new point. Middle: Switch / not switch with x^{best}. Right: New interval.

QUADRATIC INTERPOLATION

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

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

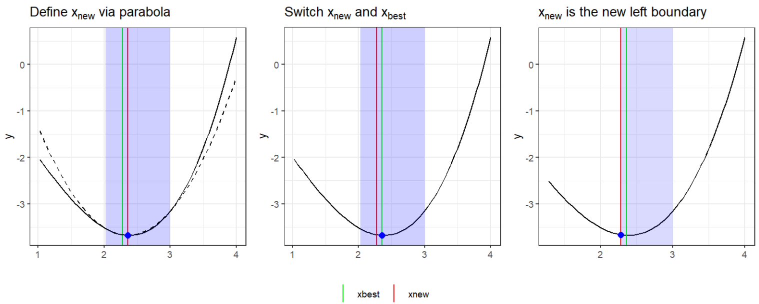


Left: Fit parabola (dashed) and propose minimum (red) as new point. Middle: Switch / not switch with x^{best}. Right: New interval.

QUADRATIC INTERPOLATION

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

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

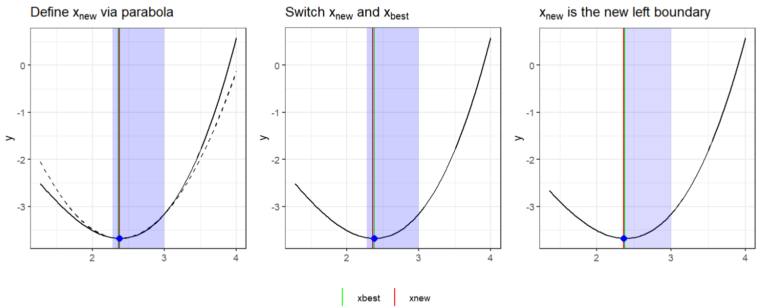


Left: Fit parabola (dashed) and propose minimum (red) as new point. Middle: Switch / not switch with x^{best}. Right: New interval.

QUADRATIC INTERPOLATION

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

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

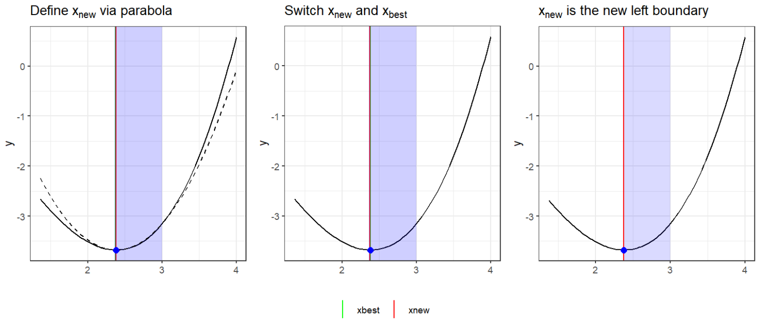


Left: Fit parabola (dashed) and propose minimum (red) as new point. Middle: Switch / not switch with x^{best}. Right: New interval.

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.

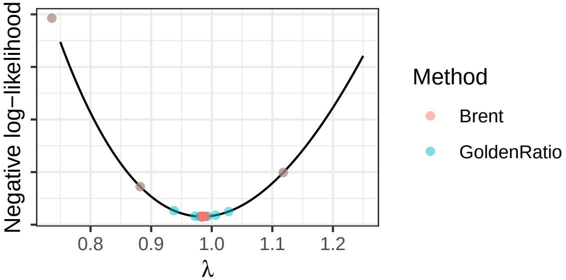
BRENT'S METHOD

- 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.