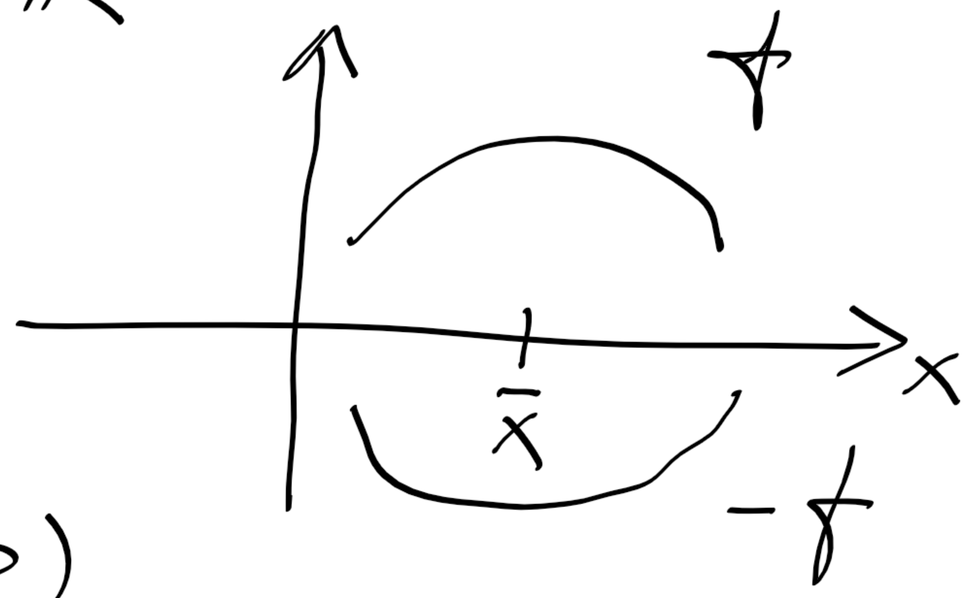
Optimization

Problem

(P) $\quad$ minimize $f(x)$
subject to $x \in S \subseteq \mathbb{R}^n$

where the objective function $f : S \to \mathbb{R}$

Note: $\boxed{\max f = -\min(-f)}$



* Unconstrained optimization (Ch. 2-3)
  $S = \mathbb{R}^n$ $\quad$ algorithms to find an approximate minimizer

* Constrained optimization (Ch. 4-8)
  $$S = \left\{ x \in \overline{X} \subseteq \mathbb{R}^n : \begin{array}{l} g_i(x) \le 0, \quad i = 1, \dots, m, \\ h_j(x) = 0, \quad j = 1, \dots, l \end{array} \right\}$$
  KKT theory, convexity, duality

* Penalty/barrier functions (Ch. 9):
  Convert constrained non-linear (P) to unconstrained and use algorithms.

Notation

- Def. $x \in \mathbb{R}^n$ is feasible iff $x \in S$.
  A constraint $g_i(x) \le 0$ is active at $\bar{x} \in \mathbb{R}^n$ iff $g_i(\bar{x}) = 0$.

- Vector $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$.

- Scalar product $x^T y = (x_1 \dots x_n) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$
  $= \sum_{i=1}^n x_i y_i = y^T x$

- length $\|x\| = \sqrt{x^T x} = \sqrt{\sum x_i^2}$

- gradient $\nabla f = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$

- Hessian

$$\nabla^2 f = \nabla \nabla^T f = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \left(\frac{\partial}{\partial x_1} \cdots \frac{\partial}{\partial x_n} \right) f$$

$$= \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Symmetric when $f \in C^2$

- we write $\nabla f^T = (\nabla f)^T = \nabla^T f$
- A line in $\mathbb{R}^n$ through $x_0 \in \mathbb{R}^n$ with direction d: $\boxed{x = x_0 + td}$

parameter

Review of calculus

- Behaviour of f along the line $\boxed{x(t) = x_0 + td}$ is given by

$$F(t) = f(x(t)) = f(x_0 + td)$$

$$F'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt} = \sum_i \frac{\partial f}{\partial x_i} d_i = \nabla f^T d = d^T \nabla f$$

$$F''(t) = \frac{d}{dt} F'(t) = d^T \nabla \nabla f^T d = d^T \nabla \nabla^T f d = d^T \nabla^2 f d$$

Hessian

- Directional derivative: $\|d\| = 1$

$$f'(x_0; d) = \lim_{t \to 0} \frac{f(x_0 + td) - f(x_0)}{t} = \lim_{t \to 0} \frac{F(t) - F(0)}{t}$$

$$= F'(0) = \nabla f(x_0)^T d$$

Property: $f'(x_0; d) = \nabla f(x_0)^T d \leq |\nabla f(x_0)^T d|$

$\leq \|\nabla f(x_0)\| \, \|d\| = \|\nabla f(x_0)\|$

Cauchy-Schwarz inequality

with equality iff $\nabla f(x_0) = td$, for some $t \geq 0$.

- Proposition: $\nabla f(a) \perp$ level surface $f(x) = C$ through $a \in \mathbb{R}^n$

Proof: True if $\nabla f(a) = 0$. Otherwise $\nabla f(a) \neq 0$, we $\frac{\partial f}{\partial x_n}(a) \neq 0$. Implicit function thm gives

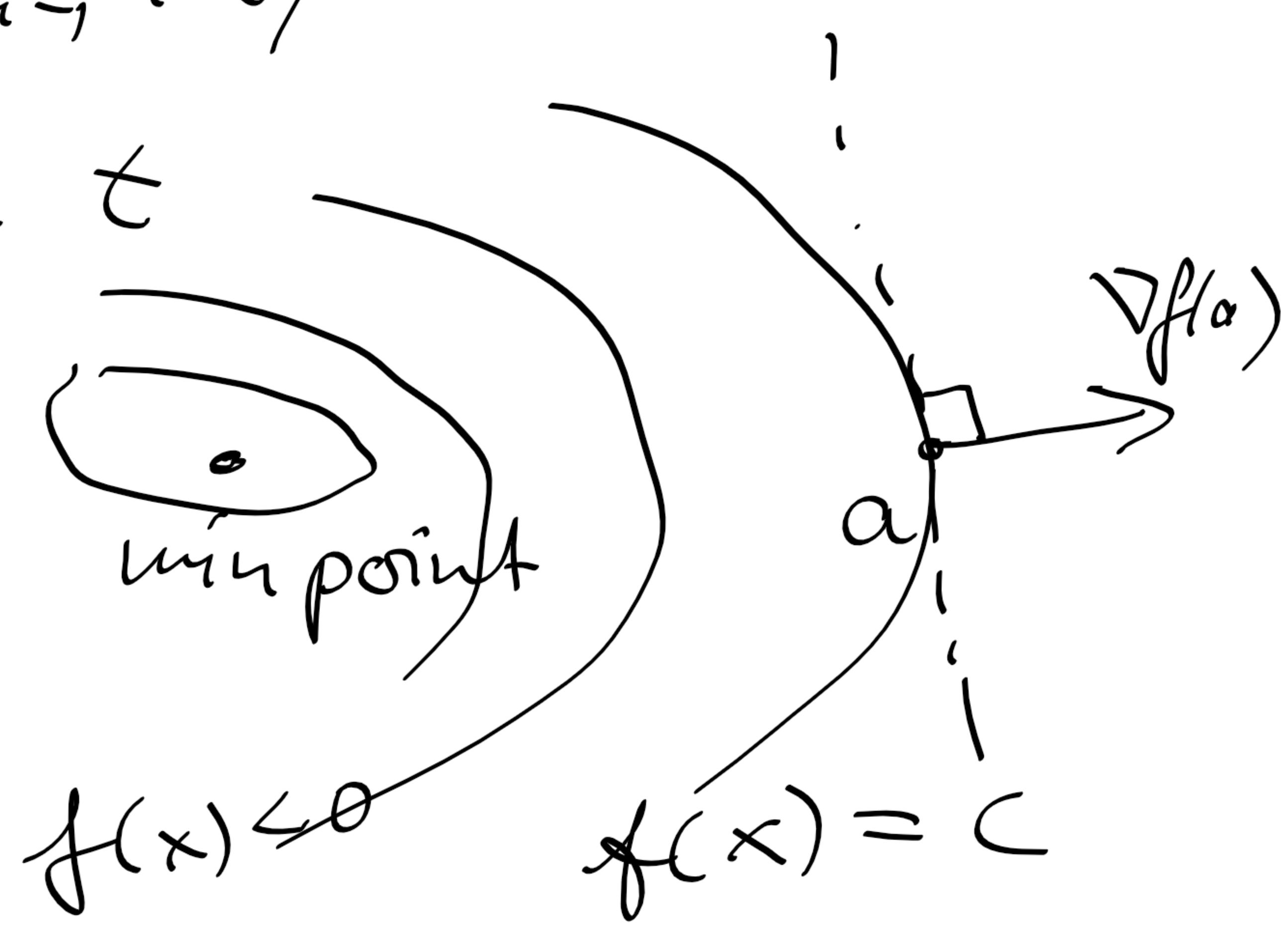
$g \in C^1$: $\quad f(x) = C \iff$

$$x_n = g(x_1, x_2, \ldots, x_{n-1}) \quad \text{near } a$$

$$\iff \begin{cases} x_1 = a_1 + t \\ x_2 = a_2 + t \\ \vdots \\ x_{n-1} = a_{n-1} + t \\ x_n = g(a_1 + t, \ldots, a_{n-1} + t) \end{cases} \quad \text{small } t$$

$\iff \quad x = \hat{x}(t) \quad$ small t

Then $f(\hat{x}(t)) = C \implies$



$$\nabla f(a)^T \underbrace{\hat{x}'(0)}_{\text{tangent}} = 0$$

- Taylor expansion of $f \in C^2$ at $x_0 \in \mathbb{R}^n$:

With $F(t) = f(x_0 + td)$

$$F(t) = F(0) + tF'(0) + \frac{1}{2} t^2 F''(0) + o(t^2)$$

where $\frac{o(t)}{t} \to 0$ as $t \to 0$

$$f(x_0+td) = f(x_0) + t\, d^T \nabla f(x_0) + \frac{1}{2} t^2 d^T \nabla^2 f(x_0) d + o(t^2)$$

set $x = x_0 + td \iff td = x - x_0$

$$f(x) = f(x_0) + (x-x_0)^T \nabla f(x_0) + \frac{1}{2}(x-x_0)^T \nabla^2 f(x_0)(x-x_0) + o(\|x-x_0\|^2)$$

— Conditions for local minima

(P) $\quad \underset{x \in S}{\text{minimize}} \; f(x)$

Def. $\bar{x} \in S$ is a

- global minimizer if $f(x) \geq f(\bar{x}) \quad \forall x \in S$
- local minimizer if —— " ——————
  and $\|x - \bar{x}\| < \delta$ for some δ
- strict local min. if $f(x) > f(\bar{x})$ when $x \neq \bar{x}$

Let $S = \mathbb{R}^n$

Thm (necessary conditions)

$\bar{x}$ local minimizer of $f \in C^2 \implies$

$$\begin{cases} \nabla f(\bar{x}) = 0 \\ \nabla^2 f(\bar{x}) \text{ positive semidef.} \end{cases}$$

Proof: Let $F(t) = f(\bar{x} + td)$, d arbitrary

Slope $\quad 0 = F'(0) = d^T \nabla f(\bar{x}) \quad \forall d \Rightarrow \nabla f(\bar{x}) = 0$

Curvature $\quad 0 \leq F''(0) = d^T \nabla^2 f(\bar{x}) d \quad \forall d \qquad$ #

Thm (sufficient conditions)

$f \in C^2$, $\quad \nabla f(\bar{x}) = 0$, $\quad \nabla^2 f(\bar{x})$ pos. def.

$\implies \bar{x}$ strict local minimizer.

Proof: Taylor expansion with $d \neq 0$:

$$f(\bar{x} + td) - f(\bar{x}) = \frac{1}{2} t^2 d^T \nabla^2 f(\bar{x}) d + o(t^2)$$

$$= t^2 \Big(\frac{1}{2} \underbrace{d^T \nabla^2 f(\bar{x}) d}_{\text{fixed scalar} > 0} + \underbrace{o(1)}_{\to 0,\ t \to 0} \Big) > 0 \text{ for small } t \quad \#$$

Very important to study optimization algorithms on a quadratic function:

$$q(x) = \frac{1}{2} x^T H x + c^T x + b$$

$$\nabla q(x) = Hx + c \qquad \text{(Exercise!)}$$

$$\nabla^2 q(x) = H$$

Unconstrained optimization

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ f(x)$$

Prototype algorithm: Given a starting point x_1, do for $k = 1, \dots$

- given x_k, choose a search direction d_k
- find $\lambda_k \in \mathbb{R}$ that minimizes
  $$F(\lambda) = f(x_k + \lambda d_k) \qquad \text{(line search)}$$
- $x_{k+1} = x_k + \lambda_k d_k$
- stop criterion (see p. 108)

Def. d_k is a descent direction at x_k iff $f(x_k + \lambda d_k) < f(x_k)$ when $0 < \lambda < \delta$ for some δ

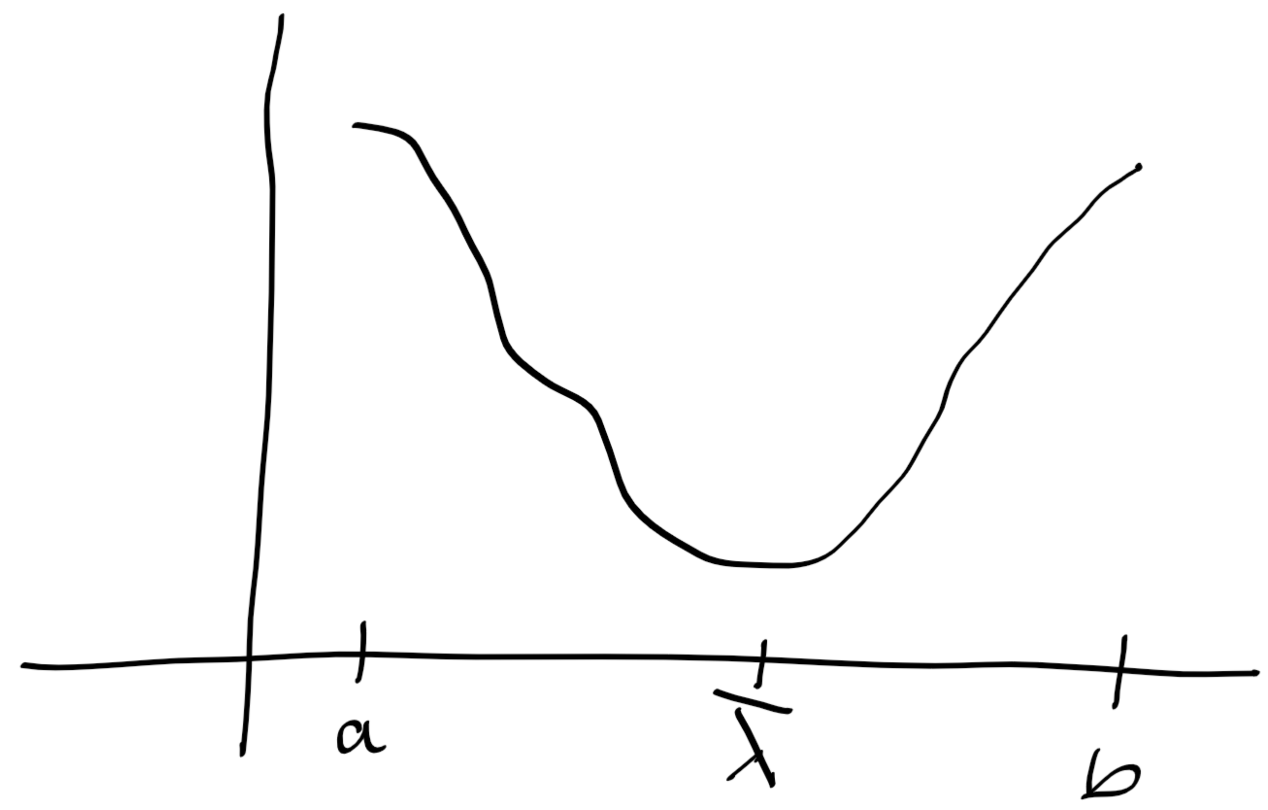
Subproblems

- choice of d_k – choice of method
- choice of line search and how accurate
- guarantee for convergence (Ch. 10) generally only to a stationary point
- convergence rate (App. 1, Ch. 3)

Line search (Chapter 2)

Problem: $\displaystyle\min_{a \le \lambda \le b} F(\lambda)$

Assume F is unimodal with minimizer $\bar{\lambda}$.



Algorithms:

- without derivatives
  - Dichotomous search
  - Golden section search
  - Quadratic fit $F(\lambda_k)$, $F(\lambda_{k-1})$, $F(\lambda_{k-2})$
- with derivatives
  - Bisection
  - Armijo's rule
  - Quadratic fit using $F(\lambda_k)$, $F(\lambda_{k-1})$, $F'(\lambda_k)$

Dichotomous search

Initial interval length $L_0 = b - a$.

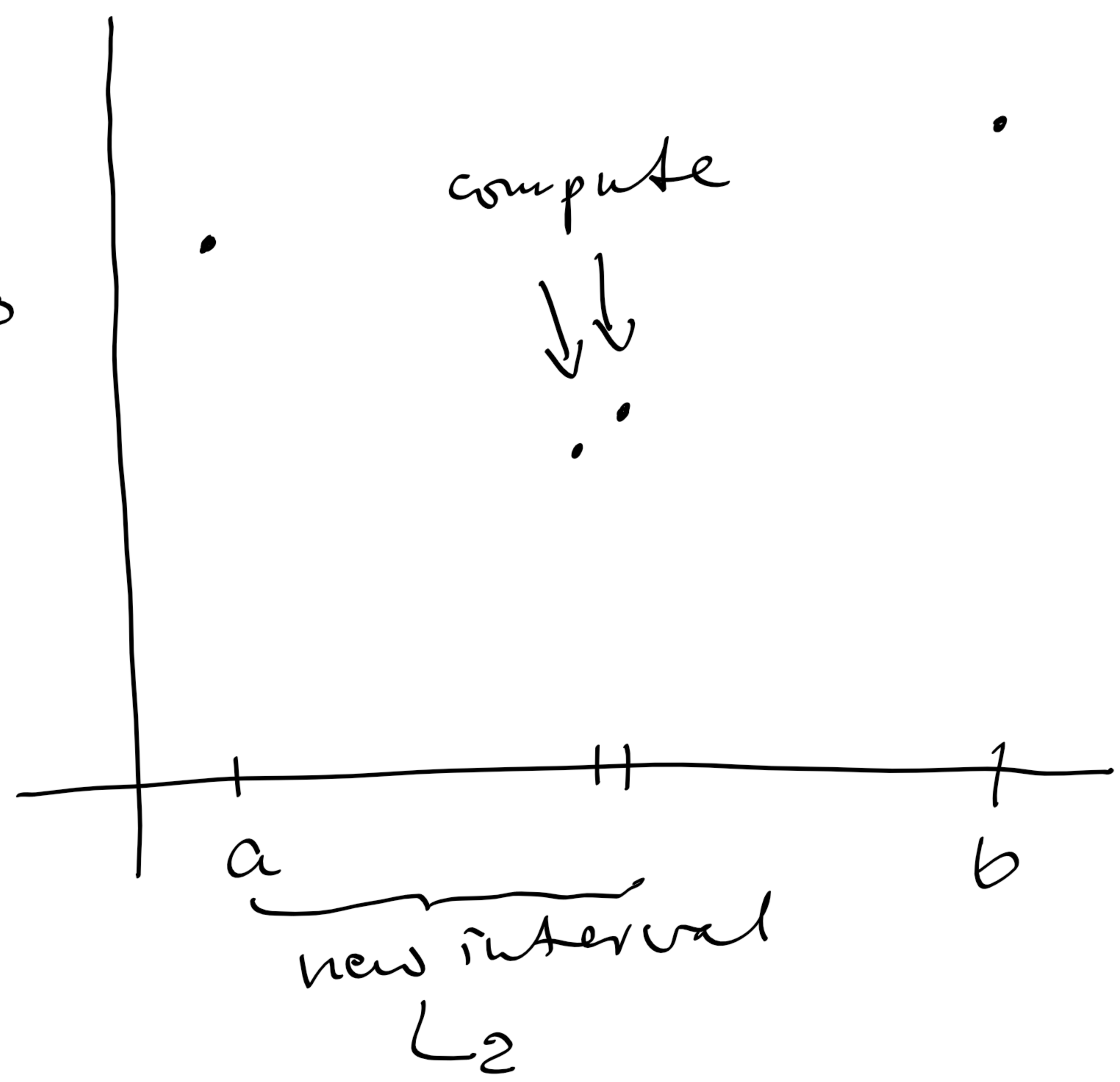
New interval after two function evaluations.

$$L_2 \approx \frac{1}{2} L_0$$

$$L_4 \approx \frac{1}{2} L_2 = \frac{1}{2^2} L_0$$

$$L_N \approx \frac{1}{2^{N/2}} L_0 = \Big(\underbrace{\frac{1}{\sqrt{2}}}_{\approx 0.71}\Big)^N L_0$$

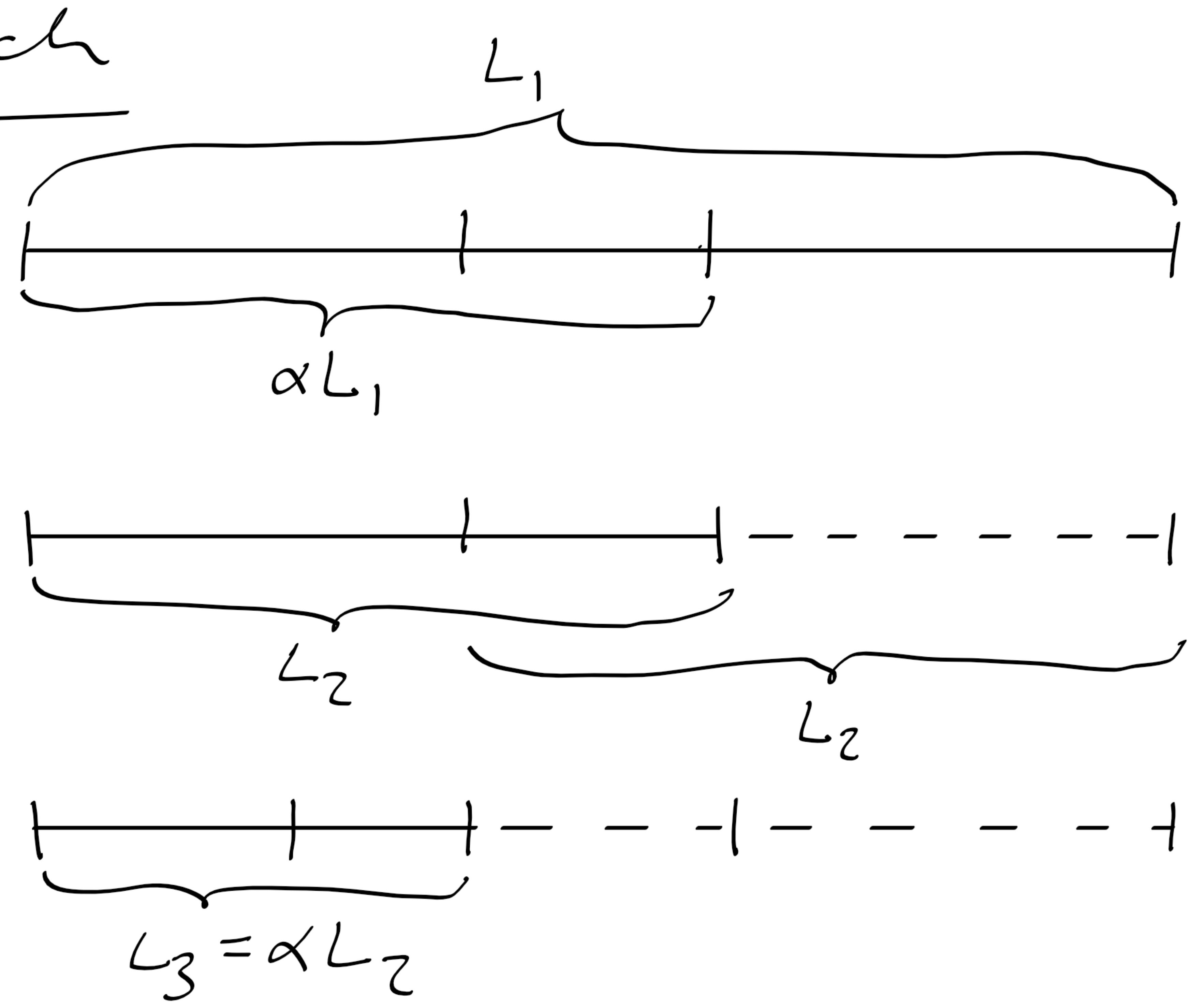
N even



Golden section search

Start (one function eval.)

One new function evaluation per iteration



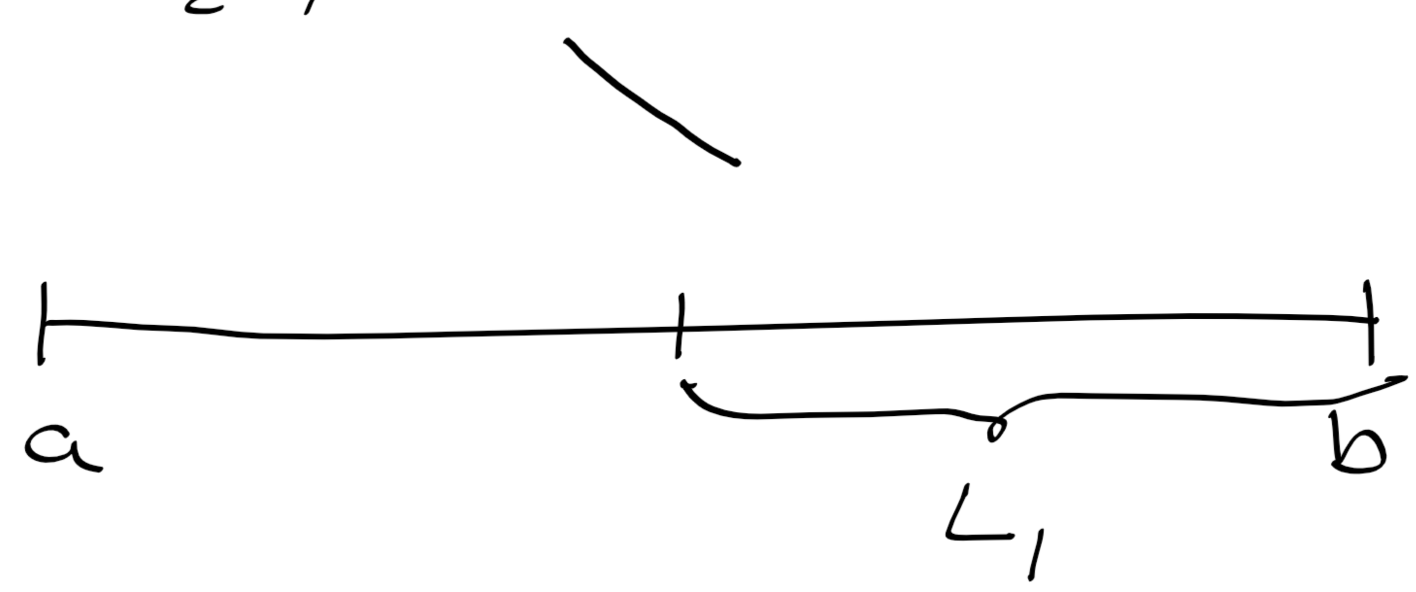
Symmetry gives

$$\alpha L_2 = L_1 - L_2 \iff$$
$$\alpha^2 L_1 = L_1 - \alpha L_1 \iff$$
$$\alpha^2 = 1 - \alpha \iff$$

$$\alpha^2 + \alpha - 1 = 0 \iff \alpha = -\frac{1}{2} \underset{(-)}{+} \sqrt{\frac{1}{4} + 1} = \frac{-1 + \sqrt{5}}{2} \approx 0.618$$

Thus $L_N = \alpha^{N-1} L_1$

Bisection: check $F'(\frac{a+b}{2})$



$L_0 = b - a$

$L_1 = \frac{1}{2} L_0$

$\vdots$

$L_N = \frac{1}{2^N} L_0$

Armijo's rule

λ is not too big
if $F(\lambda) \leq T(\lambda)$

λ is not too small
if $F(\alpha\lambda) \geq T(\alpha\lambda)$

$\alpha = 2$ typical