Today:

- sensitivity in nD
- methods in 1D
- methods in nD

Announcements:

- Examlet 2 pages, 90k
- 4CH ① released
- Broyden derivation
- Office hour capacity

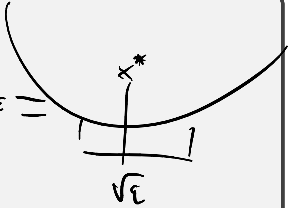
Sensitivity and Conditioning (1D)

How does optimization react to a slight perturbation of the minimum?

$$f(x^*+h) = f(x^*) + \cancel{f'(x^*)}h + f''(x^*)\frac{h^2}{2} \quad \varepsilon =$$

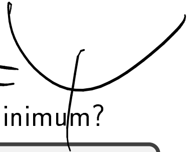
assumed known to ε

$$|\tilde{x} - x^*| < \sqrt{2\varepsilon / f''(x^*)}$$



Sensitivity and Conditioning (nD)

$f(\vec{x})$ $\epsilon :=$



How does optimization react to a slight perturbation of the minimum?

$$f(\vec{x}^* + h\vec{s}) = f(\vec{x}^*) + \cancel{h\nabla f(\vec{x}^*)^T\vec{s}} + \frac{h^2}{2}\vec{s}^T \underbrace{H_f(\vec{x}^*)\vec{s}}$$

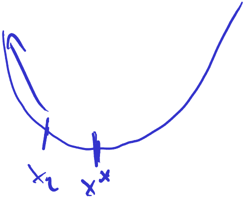
$\|\vec{s}\|_2 = 1$

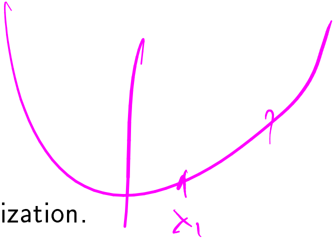
$$|h|^2 \le \frac{2\epsilon}{\lambda_{min}(H_f)}$$

$f(x,y) = x^2 + 10^{-5} y^2$ $\quad \nabla f = \begin{pmatrix} 2x \\ 2y \end{pmatrix}$

$H_f = \begin{pmatrix} 2 & 0 \\ 0 & 2\cdot 10^{-5} \end{pmatrix}$

Unimodality





Would like a method like bisection, but for optimization.

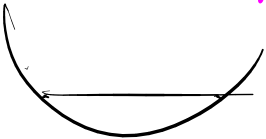
In general: No invariant that can be preserved.

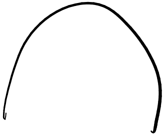
Need *extra assumption*.

f is called unimodal if for all $x_1 < x_2$

if $x_2 < x^* \Rightarrow f(x_1) > f(x_2)$

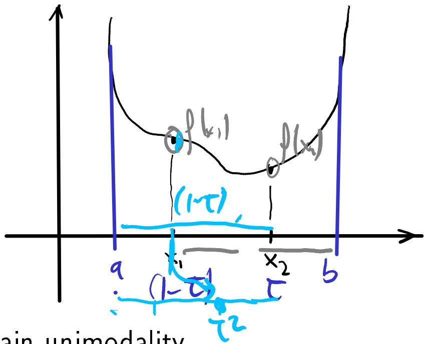
if $x^* < x_1 \Rightarrow f(x_1) < f(x_2)$





Golden Section Search

Suppose we have an interval with f unimodal:



$x_1 = a + (1-\tau)(b-a)$

$x_2 = a + \tau(b-a)$

$(1-\tau) = \tau^2$

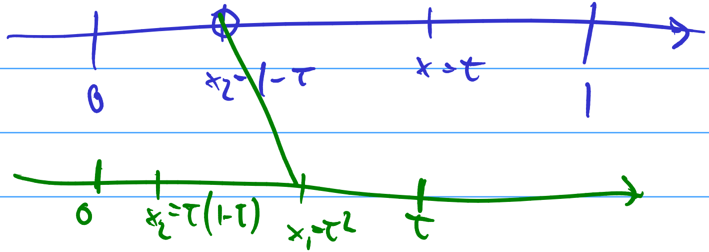
$\tau = (\sqrt{5}-1)/2$

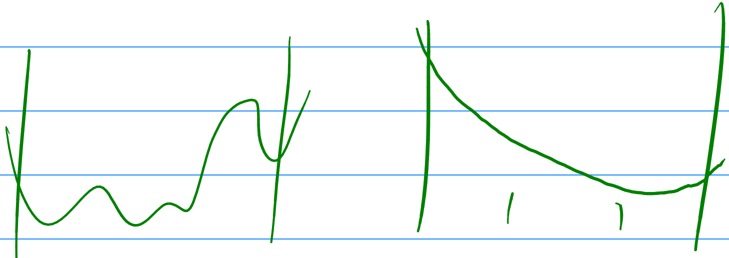
Would like to maintain unimodality.

Pick x_1, x_2:

if $f(x_1) > f(x_2) \longrightarrow$ shrink to (x_1, b)

if $f(x_1) \le f(x_2) \rightarrow$ shrink to (a, x_2)


0
$x_2 = 1-\tau$
$x = t$
1
0
$x_2 = \tau(1-\tau)$
$x_1 = \tau^2$
t



Golden Section Search: Efficiency

Where to put x_1, x_2?

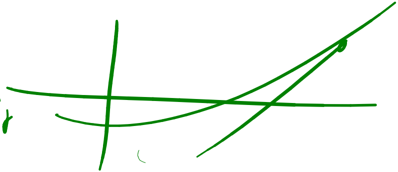
Convergence rate?

Linear w/ factor τ

Demo: Golden Section Proportions

Newton's Method

Eqn. solving



$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

Reuse the Taylor approximation idea, but for optimization.

$$f(x+h) \approx f(x) + f'(x)\cdot h + f''(x)\cdot\frac{h^2}{2} =: \hat{f}(h)$$

Demo: Newton's Method in 1D

$$\hat{f}'(h) = f'(x) + f''(x)\cdot h = 0$$

$$h = -\frac{f'(x)}{f''(x)}$$

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

Steepest Descent

Given a scalar function $f : \mathbb{R}^n \to \mathbb{R}$ at a point $\mathbf{x}$, which way is down?

Direction of SD: $-\nabla f$

$x_0 = $ <initial guess>

$$\underbrace{x_{k+1} = x_k + \alpha \underbrace{(-\nabla f(x_k))}_{s_k}}_{\varphi(\alpha)}$$

Find $\min_\alpha \varphi(\alpha)$ (e.g. using GS), use that as our next guess

Demo: Steepest Descent

Steepest Descent: Convergence

Consider quadratic model problem:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{c}^T\mathbf{x}$$

where A is SPD. (A good model of f near a minimum.)

Define error $\mathbf{e}_k = \mathbf{x}_k - \mathbf{x}^*$. Then

$$||\mathbf{e}_{k+1}||_A = \sqrt{\mathbf{e}_{k+1}^T A\mathbf{e}_{k+1}} = \underbrace{\frac{\sigma_{\max}(A) - \sigma_{\min}(A)}{\sigma_{\max}(A) + \sigma_{\min}(A)}}||\mathbf{e}_k||_A$$

$\to$ confirms linear convergence.

Convergence constant related to conditioning:

$$\frac{\sigma_{\max}(A) - \sigma_{\min}(A)}{\sigma_{\max}(A) + \sigma_{\min}(A)} = \frac{\kappa(A) - 1}{\kappa(A) + 1}.$$

Hacking Steepest Descent for Better Convergence

Extrapolation methods: Look back a step, maintain '*momentum*'.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1})$$

Heavy ball method: constant $\alpha_k = \alpha$ and $\beta_k = \beta$. Gives:

$$||\mathbf{e}_{k+1}||_A = \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} ||\mathbf{e}_k||_A$$

Conjugate gradient method:

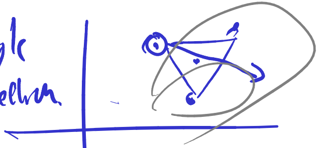
$$(\alpha_k, \beta_k) = \operatorname{argmin}_{\alpha_k, \beta_k} \left[f\left(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) + \beta_k(\mathbf{x}_k - \mathbf{x}_{k-1}) \right) \right]$$

- Will see in more detail later (for solving linear systems)
- Provably optimal first-order method for the quadratic model problem
- Turns out to be closely related to Lanczos (A-orthogonal search directions)

$(x, y) = x^T A y$

Nelder-Mead Method

triangle
tetrahedron



Idea:

> Take n-simplex, move "worst point" on a line through center/centroid of the simplex

Demo: Nelder-Mead Method

Newton's method (n D)

$\nabla f = 0$

What does Newton's method look like in n dimensions?

$$\underbrace{f(\vec{x}+\vec{h}) = f(\vec{x}) + \nabla f(\vec{x})\cdot\vec{h} + \frac{1}{2} h^T H_f(x) h + \dots}_{\hat{f}(\vec{h})}$$

$$\nabla \hat{f}(h) = \nabla f(\vec{x}) + H_f(x)\cdot\vec{h} = 0$$

$$H_f(x)\cdot\vec{h} = -\nabla f(x)$$

$$\underline{\vec{h} = -H_f^{-1}(x)\cdot\nabla f(x)}$$

equ solving

$x_{k+1} = x_k - J^{-1} f$

$$x_{k+1} = x_k + h$$
$$= x_k - H_f^{-1}(x_k)\cdot\nabla f(x_k)$$

Newton's method (n D): Observations

Drawbacks?

Demo: Newton's method in n dimensions