Gradients, Linearizations, and Optimality

CPSC 406 – Computational Optimization

Gradients, linearizations, and optimality

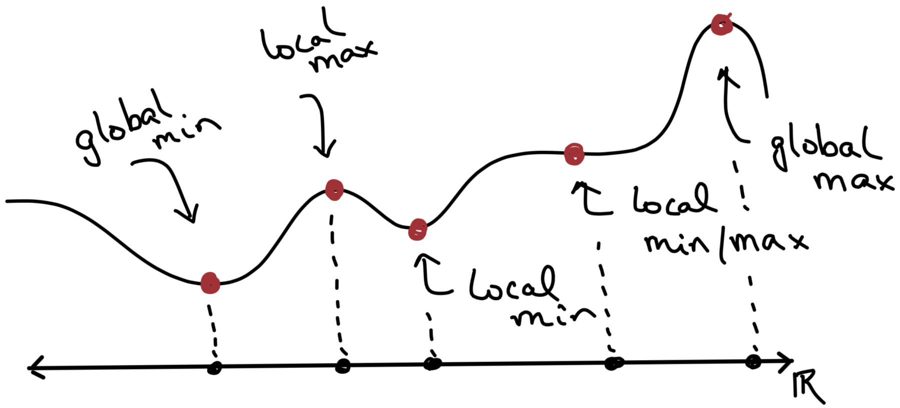
- directional derivatives
- gradients
- first-order expansions
- necessary conditions for optimality

Optimality

$$\min_x \; f(x) \quad \text{where} \quad f : \mathbb{R}^n \to \mathbb{R}$$

$x^* \in \mathbb{R}^n$ is a

- **global minimizer** if $f(x^*) \leq f(x)$ for all x
- **strict global minimizer** if $f(x^*) < f(x)$ for all x
- **local minimizer** if $f(x^*) \leq f(x)$ for all $x \in \epsilon\mathbf{B}(x^*)$
- **strict local minimizer** if $f(x^*) < f(x)$ for all $x \in \epsilon\mathbf{B}(x^*)$



Maximizers

- flip inequalities for analogous *maximizer* def's
- $\underset{x}{\operatorname{argmin}}\{f(x)\} = \underset{x}{\operatorname{argmax}}\{-f(x)\}$

The ϵ-ball centered at $\bar{x}$ is the set of points $\epsilon\mathbf{B}(\bar{x}) = \{x \in \mathbb{R}^n \mid \|x - \bar{x}\| < \epsilon\}$. If $\bar{x} = 0$, we use the shorthand $\mathbf{B}(0) = \mathbf{B}$.

Optimal attainment

- an optimal value may not be **attained**, eg,
  - $\inf_x \, e^{-x}$ is not attained for **any** $x \in \mathbb{R}$
- an optimal value may not **exist**, eg,
  - $\min_x \, -x^2$ has no minimizer (unbounded below)
- global solution set (may be empty / unique element / many elements)

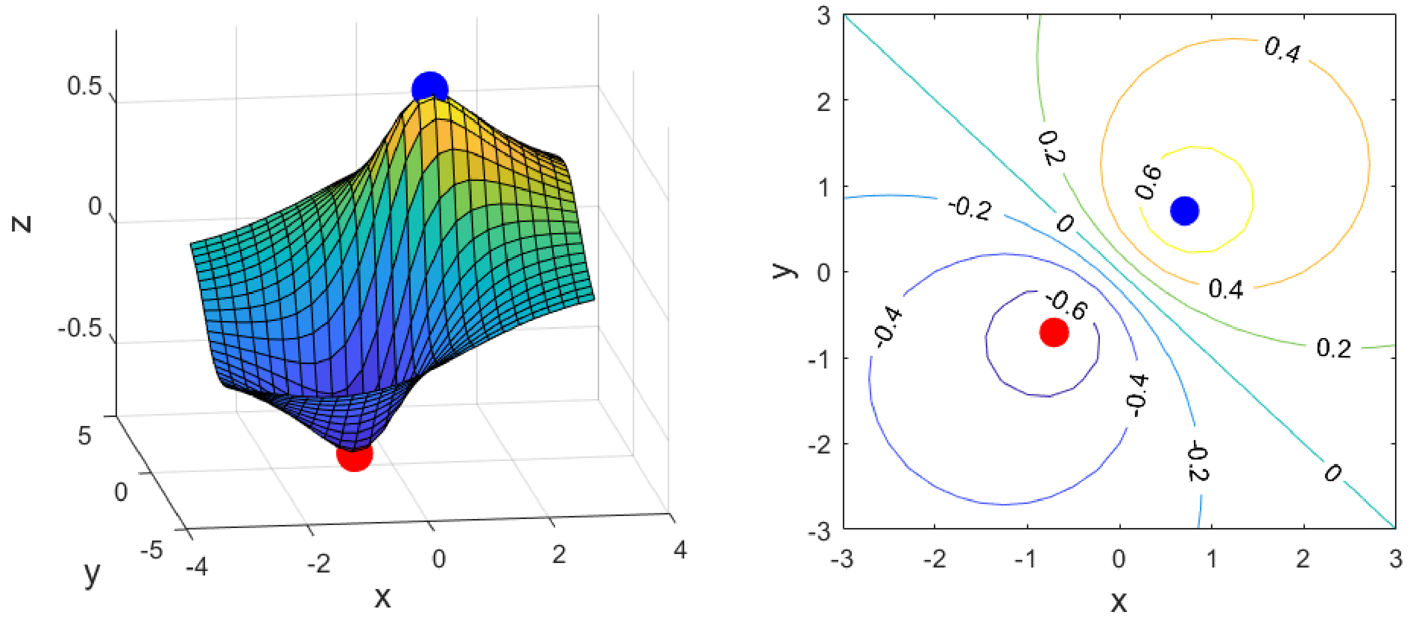
$$\operatorname*{argmin}_x f(x) = \{\bar{x} \mid f(\bar{x}) \le f(x) \text{ for all } x\}$$

- optimal values are unique even if an optimal point is not unique

Theorem 1 (Coercivity implies existence of minimizer) If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and $\lim_{\|x\|\to\infty} f(x) = \infty$ (**coercive**), then $\min_x \, f(x)$ has a global minimizer.

Example

$$\min_{x \in \mathbb{R}^2} \frac{x_1 + x_2}{x_1^2 + x_2^2 + 1}$$



- global minimizer at $-\frac{1}{\sqrt{2}}(1,1)$
- global maximizer at $\frac{1}{\sqrt{2}}(1,1)$

scalar variable (n)

Local optimality (1-D)

Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable. The point $x = x^*$ is a

- **local minimizer** if

$$\underbrace{f'(x) = 0}_{\text{stationary at } x} \quad \text{and} \quad \underbrace{f''(x) > 0}_{\text{(strictly) convex at } x}$$

- **local maximizer** if

$$\underbrace{f'(x) = 0}_{\text{stationary at } x} \quad \text{and} \quad \underbrace{f''(x) < 0}_{\text{(strictly) concave at } x}$$

- if $f'(x) = 0$ and $f''(x) = 0$, not enough information, eg,
  - $f(\bar{x}) = x^3 \quad \implies \quad x = 0$ in **not** a local minimizer or maximizer even though $f'(0) = 0$
  - $f(\bar{x}) = x^4 \quad \implies \quad x = 0$ is the unique global **minimizer** even though $f''(0) = 0$

Local optimality (1-D): motivation

- suppose $f'(x^*) = 0$ and $f''(x^*) > 0$ at some x^*
- Taylor series, where remainder term $o(\alpha)/\alpha \to 0$ as $\alpha \to 0^+$:

$$f(x^* + \Delta x) = f(x^*) + \underbrace{f'(x^*)\Delta x}_{=0} + \underbrace{\tfrac{1}{2} f''(x^*)(\Delta x)^2}_{>0} + o((\Delta x)^2)$$

- divide both sides by $(\Delta x)^2$; for Δx small enough, right-hand side is positive:

$$\frac{f(x^* + \Delta x) - f(x^*)}{(\Delta x)^2} = \tfrac{1}{2} f''(x^*) + \frac{o((\Delta x)^2)}{(\Delta x)^2} > 0$$

- implies $f(x^* + \Delta x) > f(x^*)$ for Δx small enough

multivariable (n>1)

Directional derivative

- restrict $f : \mathbb{R}^n \to \mathbb{R}$ to the ray $\{x + \alpha d \mid \alpha \in \mathbb{R}_+\}$:

$$\phi(\alpha) = f(x + \alpha d) \qquad \phi'(0) = \lim_{\alpha \to 0^+} \frac{\phi(\alpha) - \phi(0)}{\alpha}$$

Definition 1 The **directional derivative** of f at $x \in \mathbb{R}^n$ in the direction $d \in \mathbb{R}^n$ is

$$f'(x; d) = \lim_{\alpha \to 0^+} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$

- partial derivatives are directional derivatives along each canonical basis vector e_i:

$$\frac{\partial f}{\partial x_i}(x) = f'(x; e_i) \quad \text{with} \quad e_i(j) = \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases}$$

Descent directions

- a nonzero vector d is a **descent direction** of f at x if

$$f(x + \alpha d) < f(x) \quad \forall \alpha \in (0, \epsilon) \text{ for some } \epsilon > 0$$

- equivalently, the directional derivative is negative:

$$f'(x; d) := \lim_{\alpha \to 0^+} \frac{f(x + \alpha d) - f(x)}{\alpha} < 0$$

Gradients

- if $f : \mathbb{R}^n \to \mathbb{R}$ is **continuously differentiable** (ie, differentiable at all x and ∇f is continuous) the **gradient** of f at x is the vector

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix} \in \mathbb{R}^n$$

- **gradient and directional derivative** related via

$$f'(x; d) = \nabla f(x)^\intercal d$$

- direction derivative gives
  - the rate of change of f at x in the direction d
  - (if $\|d\| = 1$) the projection of $\nabla f(x)$ onto d

Example

$$f(x) = x_1^2 + 8x_1x_2 - 2x_3^2$$

What is $f'(x; d)$ for $x = (1, 1, 2)$ and $d = (1, 0, 1)$?

a. 1

b. 2

c. 3

d. 4

e. 5

Automatic differentiation

$$f(x) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$$

gradient

```
using ForwardDiff
f(x) = (1 - x[1])^2 + 100*(x[2] - x[1]^2)^2
∇f(x) = ForwardDiff.gradient(f, x)
x = [1.0, 1.0]
@show ∇f(x);
```

```
∇f(x) = [-0.0, 0.0]
```

directional derivative

```
fp(x, d) = ForwardDiff.derivative(α->f(x + α*d), 0.)
d = [1.0, 0.0]
fp(x, d)
fp(x, d) == ∇f(x)'d
```

```
true
```

Visualizing the gradient

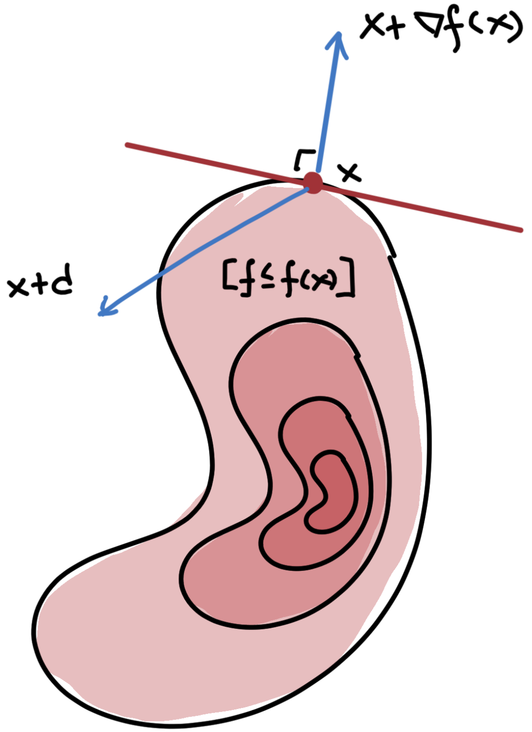
Definition 2 (Level set) The α-level set of f is the set of points x where the function value is at most α:

$$[f \leq \alpha] = \{x \mid f(x) \leq \alpha\}$$

- a direction d points "into" the level set $[f \leq f(x)]$ if

$$f'(x; d) := \nabla f(x)^\intercal d < 0$$

- the gradient $\nabla f(x)$ is orthogonal to the level set defined by $f(x)$



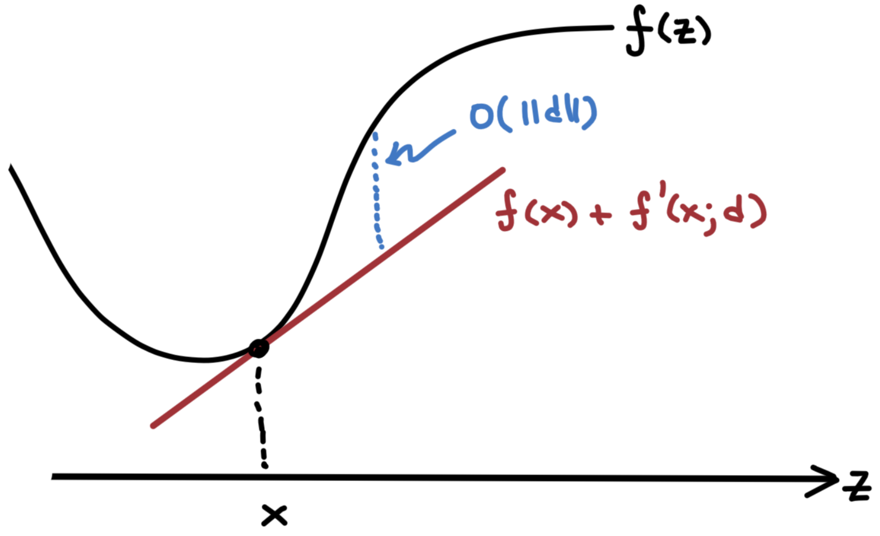
Linear approximation

- if $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at x, then for any direction d

$$f(x+d) = f(x) + \nabla f(x)^\intercal d + o(\|d\|) = f(x) + f'(x;d) + o(\|d\|)$$

- the *remainder* $o : \mathbb{R}_+ \to \mathbb{R}$ decays faster than $\|d\|$

$$\lim_{\alpha \to 0+} \frac{o(\alpha)}{\alpha} = 0$$



1st-order conditions

Theorem 2 (Necessary first-order conditions) For $f : \mathbb{R}^n \to \mathbb{R}$ differentiable, x^* is a local minimizer *only if* it is a **stationary point**:

$$\nabla f(x^*) = 0$$

- up to first order, for any direction d

$$\begin{aligned} f(x^* + \alpha d) - f(x^*) &= \nabla f(x^*)^\intercal(\alpha d) + o(\alpha\|d\|) \\ &= \alpha f'(x^*; d) + o(\alpha\|d\|) \end{aligned}$$

- because f is (locally) minimal at x^*

$$0 \leq \lim_{\alpha \to 0^+} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha} = f'(x^*; d) = \nabla f(x^*)^\intercal d$$

- because this holds for all d, necessarily $\nabla f(x^*) = 0$

Example: Quadratic

$$f(x) = \tfrac{1}{2}x^\intercal Hx - c^\intercal x + \gamma, \quad H = H^\intercal \in \mathbb{R}^n, \quad c \in \mathbb{R}^n$$

- x^* is a local minimizer *only if* $\nabla f(x^*) = 0$, ie,

$$0 = \nabla f(x^*) = Hx^* - c \quad \Longrightarrow \quad Hx^* = c$$

- if $\mathbf{null}(H) \neq \emptyset$ and $c \in \mathbf{range}(H)$, then there exists x_0 such that $Hx_0 = b$ and

$$\operatorname*{argmin}_x \; f(x) = \{\, x_0 + z \mid z \in \mathbf{null}(H) \,\}$$

Example: Least squares

$$f(x) = \frac{1}{2}\|Ax - b\|^2 = \frac{1}{2}(Ax - b)^\top (Ax - b) = \frac{1}{2} x^\top \underbrace{(A^\top A)}_{=H} x - \underbrace{(b^\top A)}_{=c^\top} x + \underbrace{\frac{1}{2} b^\top b}_{=\gamma}$$

- x^* is a least-squares solution if and only if it satisfies the **normal equations**

$$0 = \nabla f(x^*) = A^\top A x^* - A^\top b \quad \Longleftrightarrow \quad A^\top A x^* = A^\top b$$

Example: Nonlinear least squares

$$f(x) = \tfrac{1}{2}\|r(x)\|^2 = \tfrac{1}{2} r(x)^\intercal r(x) = \tfrac{1}{2}\sum_{i=1}^{m} r_i(x)^2$$

where

$$r(x) = \begin{bmatrix} r_1(x) \\ \vdots \\ r_m(x) \end{bmatrix} \quad \text{where} \quad r_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, \dots, m$$

gradient

$$\nabla f(x) = \nabla \left[\tfrac{1}{2} \sum_{i=1}^{m} r_i(x)^2 \right] = \sum_{i=1}^{m} \nabla r_i(x) r_i(x)$$

$$= \underbrace{\left[\, \nabla r_1(x) \mid \cdots \mid \nabla r_m(x) \,\right]}_{\nabla r(x) \equiv J(x)^\intercal} \begin{bmatrix} r_1(x) \\ \vdots \\ r_m(x) \end{bmatrix} = J(x)^\intercal r(x)$$

Gradients and convergence

```
using Plots
using Optim: g_norm_trace, f_trace, iterations, LBFGS, optimize

f(x) = (1 - x[1])^2 + 100 * (x[2] - x[1]^2)^2

x0 = zeros(2)
res = optimize(f, x0, method=LBFGS(), autodiff=:forward, store_trace=true)
fval, gnrm, itns = f_trace(res), g_norm_trace(res), iterations(res)
plot(0:itns, [fval gnrm], yscale=:log10, lw=3, label=["f(x)" "||∇f(x)||"], size=(5
```