Math. review

Real vectors and matrices (§2.1)

- $\mathbb{R}$: set of real numbers
- $\mathbb{R}^n$: set of real column vectors

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad x_i \in \mathbb{R}$$

- $\mathbb{R}^{m \times n}$: set of $m \times n$ real matrices

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

- Treat $\mathbb{R}^{n \times 1}$ and $\mathbb{R}^n$ as equivalent
- $\boldsymbol{A}^T$: transpose of $\boldsymbol{A}$

$$[1, 2]^T = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Functions

- Function $f : X \to Y$
- f takes values in X and gives values in Y
  - f is Y-valued
  - $f(x)$ is the value of f at x, where $x \in X$
- Example: $f : \mathbb{R}^3 \to \mathbb{R}$

$$f(\boldsymbol{x}) = \frac{x_1^3 + 3\log(x_2 x_3)}{x_2}$$

Linear independence (§2.1)

- A set of vectors $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k\}$ is said to be *linearly independent* if the equality

$$\alpha_1 \boldsymbol{a}_1 + \alpha_2 \boldsymbol{a}_2 + \cdots + \alpha_k \boldsymbol{a}_k = \mathbf{0}$$

  implies that all the scalar coefficients α_i, $i = 1, \ldots, k$, are equal to zero.

- A set of vectors is linearly dependent if and only if one of the vectors from the set is a linear combination of the remaining vectors.

Rank of a matrix (§2.2)

- The maximal number of linearly independent columns of $\boldsymbol{A}$ is called the *rank* of the matrix $\boldsymbol{A}$, denoted rank $\boldsymbol{A}$.
- A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is *nonsingular* or *invertible* if rank $\boldsymbol{A} = n$ (full rank).
- A matrix is nonsingular if and only if its determinant is nonzero.

Inner product and norm (§2.4)

- Given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.
- Define the *inner product* of $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

- What are the properties of inner product?
- Define the *norm* of $\boldsymbol{x}$:

$$\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} = \sqrt{\boldsymbol{x}^T \boldsymbol{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

- What are the properties of norm?
- $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal* if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

Eigenvalues and eigenvectors (§3.2)

- Let $\boldsymbol{A}$ be an $n \times n$ square matrix.
- A scalar λ (possibly complex) and a nonzero vector $\boldsymbol{v}$ satisfying the equation $\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$ are said to be, respectively, an *eigenvalue* and *eigenvector* of $\boldsymbol{A}$.
- λ is an eigenvalue of $\boldsymbol{A}$ if and only if $\lambda \boldsymbol{I} - \boldsymbol{A}$ is singular (i.e., $\det[\lambda \boldsymbol{I} - \boldsymbol{A}] = 0$).
- $\det[\lambda \boldsymbol{I} - \boldsymbol{A}]$ is called the *characteristic polynomial* of $\boldsymbol{A}$.
- What are the zeros of the characteristic polynomial of $\boldsymbol{A}$?

Symmetric matrices (§3.4)

- $\boldsymbol{Q}$ is symmetric if $\boldsymbol{Q} = \boldsymbol{Q}^T$.
- A symmetric matrix $\boldsymbol{Q}$ is said to be *positive definite* if $\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} > 0$ for all nonzero vectors $\boldsymbol{x}$.
- It is *positive semidefinite* if $\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x}$.
- Similarly define *negative definite* and *negative semidefinite*.
- How is definiteness related to eigenvalues?

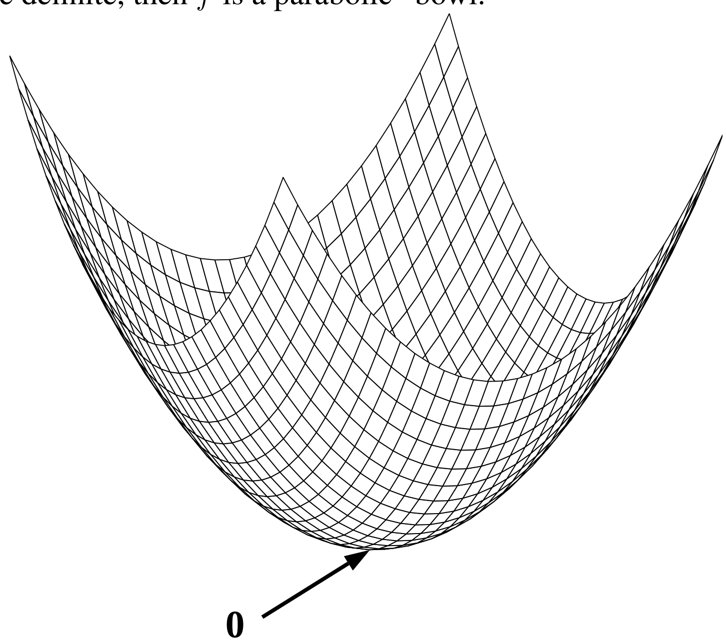
Quadratic functions (§3.4)

- $f : \mathbb{R}^n \to \mathbb{R}$ is a quadratic function if

$$f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c,$$

  where $\boldsymbol{Q}$ is symmetric.

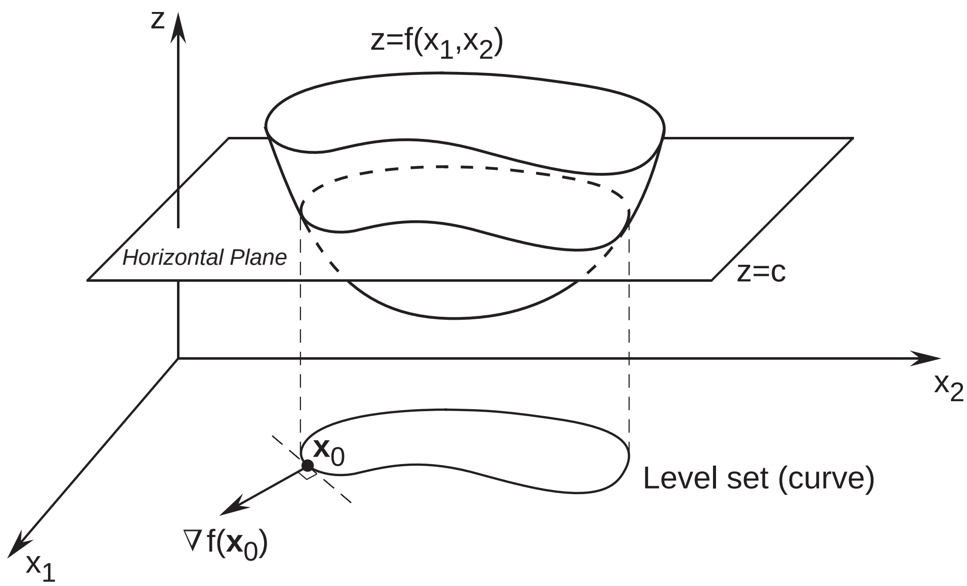
- If $\boldsymbol{Q}$ is positive definite, then f is a parabolic "bowl."



- Quadratics are useful in the study of optimization.
- Often, objective functions are "close to" quadratic near the solution.
- It is easier to analyze the behavior of algorithms when applied to quadratics.
- Analysis of algorithms for quadratics gives insight into their behavior in general.

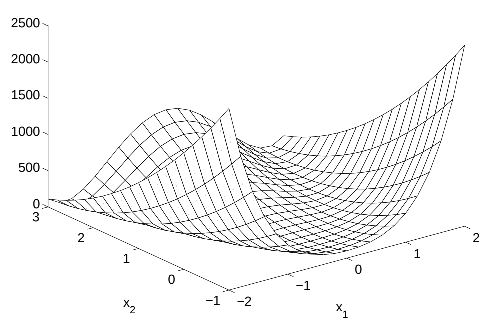
Level sets (§5.5)

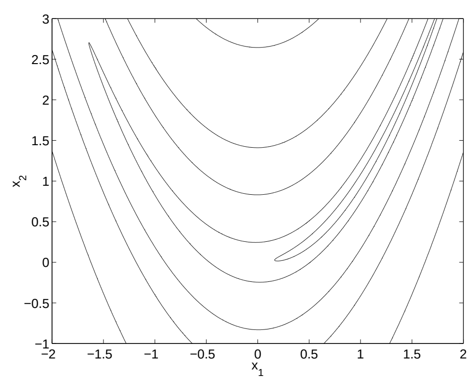
- The *level set* of a function $f : \mathbb{R}^n \to \mathbb{R}$ at level c is the set of points $S = \{\boldsymbol{x} : f(\boldsymbol{x}) = c\}$.
- The level set of f is a subset of $\mathbb{R}^n$.



- Example (Rosenbrock's function):

$$f(\boldsymbol{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2, \qquad \boldsymbol{x} = [x_1, x_2]^T.$$





Derivatives (§5.1–5.2)

- Given $f : \mathbb{R} \to \mathbb{R}$
- The *derivative* of f is a function $f' : \mathbb{R} \to \mathbb{R}$ given by

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$

if the limit exists.

- Also written $\dfrac{df}{dx}$
- If the derivative exists, we say that f is *differentiable*.
- If f' is continuous, we say that f is *continuously differentiable*.

- Given $f : \mathbb{R}^n \to \mathbb{R}$
- The *gradient* of f is a function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ given by

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}$$

- At each $\boldsymbol{x}$, $\nabla f(\boldsymbol{x})$ is a vector in $\mathbb{R}^n$.

- Given $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, $\boldsymbol{f} = [f_1, \ldots, f_m]^T$.
- The derivative of $\boldsymbol{f}$ is a function $D\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^{m \times n}$ given by

$$D\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}.$$

- Sometimes called *Jacobian*.

- At each $\boldsymbol{x}$, $D\boldsymbol{f}(\boldsymbol{x})$ is an $m \times n$ matrix.
- If $D\boldsymbol{f}$ is continuous, we say that $\boldsymbol{f}$ is *continuously differentiable*.
- We write $\boldsymbol{f} \in \mathcal{C}^1$.

- Note that for $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\nabla f(\boldsymbol{x}) = Df(\boldsymbol{x})^T.$$

- If the derivative of ∇f exists, we say that f is *twice differentiable*.

  Write the second derivative as $D^2 f$ (or $\boldsymbol{F}$), and call it the *Hessian* of f.

$$\boldsymbol{F} = D^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

- If $\boldsymbol{F}$ is continuous, we write $f \in \mathcal{C}^2$.

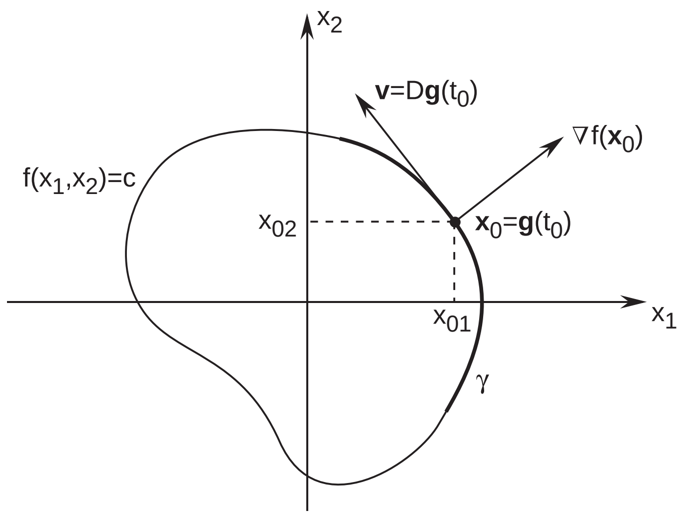
Chain rule (§5.4)

- Given $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R} \to \mathbb{R}^n$, both differentiable.
- Suppose $f \in \mathcal{C}^1$.
- Define the composite function $F : \mathbb{R} \to \mathbb{R}$ by $F(t) = f(\boldsymbol{g}(t))$.
- Then, F is differentiable, and

$$\begin{aligned} F'(t) &= Df(\boldsymbol{g}(t)) \cdot D\boldsymbol{g}(t) \\ &= \nabla f(\boldsymbol{g}(t))^T \boldsymbol{g}'(t) \\ &= \boldsymbol{g}'(t)^T \nabla f(\boldsymbol{g}(t)). \end{aligned}$$

Gradients and level sets (§5.5)

- Given $f : \mathbb{R}^n \to \mathbb{R}$.
- Fact: $\nabla f(\boldsymbol{x}_0)$ is orthogonal to the level set at $\boldsymbol{x}_0$



Proof of fact:

- Imagine a particle traveling along the level set.
- Let $\boldsymbol{g}(t)$ be the position of the particle at time t, with $\boldsymbol{g}(0) = \boldsymbol{x}_0$.
- Note that $f(\boldsymbol{g}(t))$ = constant for all t.
- Velocity vector $\boldsymbol{g}'(t)$ is tangent to the level set.

- Consider $F(t) = f(\boldsymbol{g}(t))$. We have $F'(0) = 0$. By the chain rule,

$$F'(0) = (\boldsymbol{g}'(0))^T \nabla f(\boldsymbol{g}(0)).$$

- Hence, $\nabla f(\boldsymbol{x}_0)$ and $\boldsymbol{g}'(0)$ are orthogonal.

Taylor's formula (§5.5)

- Suppose $f : \mathbb{R} \to \mathbb{R}$ is in $\mathcal{C}^1$.
- Then,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + o(x - x_0).$$

- $o(h)$ is a term such that $o(h)/h \to 0$ as $h \to 0$.
- Around x_0, f can be approximated by a linear function, and the approximation gets better the closer we are to x_0.

- Suppose $f : \mathbb{R} \to \mathbb{R}$ is in $\mathcal{C}^2$.
- Then,

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 \\ &\quad + o((x - x_0)^2). \end{aligned}$$

- Around x_0, f can be approximated by a quadratic function.

- Suppose $f : \mathbb{R}^n \to \mathbb{R}$.
- If $f \in \mathcal{C}^1$, then

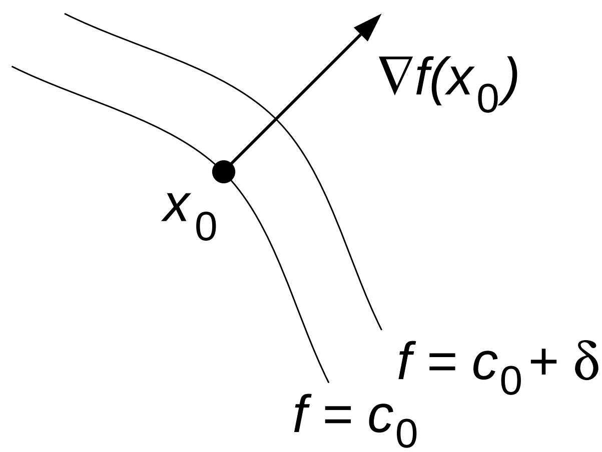
$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T (\boldsymbol{x} - \boldsymbol{x}_0) + o(\|\boldsymbol{x} - \boldsymbol{x}_0\|).$$

- If $f \in \mathcal{C}^2$, then

$$\begin{aligned} f(\boldsymbol{x}) &= f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0)^T (\boldsymbol{x} - \boldsymbol{x}_0) \\ &\quad + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^T \boldsymbol{F}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \\ &\quad + o(\|\boldsymbol{x} - \boldsymbol{x}_0\|^2). \end{aligned}$$

In what direction does a gradient point?

- Given $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{x}_0 \in \mathbb{R}^n$.
- We already know that $\nabla f(\boldsymbol{x}_0)$ is orthogonal to the level set at $\boldsymbol{x}_0$.
- Suppose $\nabla f(\boldsymbol{x}_0) \neq \mathbf{0}$.
- Fact: ∇f points in the direction of increasing f.



Proof of fact:

- Consider $\boldsymbol{x}_\alpha = \boldsymbol{x}_0 + \alpha \nabla f(\boldsymbol{x}_0)$, $\alpha > 0$.
- By Taylor's formula,

$$\begin{aligned} f(\boldsymbol{x}_\alpha) &= f(\boldsymbol{x}_0) + (\boldsymbol{x}_\alpha - \boldsymbol{x}_0)^T \nabla f(\boldsymbol{x}_0) + o(\|\boldsymbol{x}_\alpha - \boldsymbol{x}_0\|) \\ &= f(\boldsymbol{x}_0) + \alpha \|\nabla f(\boldsymbol{x}_0)\|^2 + o(\alpha). \end{aligned}$$

- Therefore, for sufficiently small α, $f(\boldsymbol{x}_\alpha) > f(\boldsymbol{x}_0)$.

