The Five Miracles of Mirror Descent

Hadar Tal
hadar.tal@mail.huji.ac.il

This paper is a summary of the educational materials and lectures from

- **The Five Miracles of Mirror Descent** by Professor Sebastien Bubeck (Claire Boyer's notes)
- **Optimization for Computer Science** by Professor Tomer Koren
- **Manifolds 1** by "The Bright Side of Mathematics" youtube channel
- **Wikipedia**

Winter 2024

Contents

Chapter 1

Multivariable Calculus

Definition 1.0.1. *Diffrentiability, single variable*
Let $f : (a, b) \to \mathbb{R}$ be a function. We say that f is differentiable at $x_0 \in (a, b)$ if

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{1.1}$$

exists. If f is differentiable at x_0, then $f'(x_0)$ is the derivative of f at x_0.

Definition 1.0.2. *Diffrentiability, single variable (alternative)*
Let $f : (a, b) \to \mathbb{R}$ be a function. We say that f is differentiable at $x_0 \in (a, b)$ if there exists a number m such that:

$$f(x_0 + h) = f(x_0) + m \cdot h + E(h) \text{ where } \lim_{h \to 0} \frac{E(h)}{h} = 0 \tag{1.2}$$

If f is differentiable at x_0, then $f'(x_0) = m$ is the derivative of f at x_0.

Definition 1.0.3. *Diffrentiability, multivariable*
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function. We say that f is differentiable at x_0 if there exists a vector m $\in \mathbb{R}^n$ such that:

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0) - m \cdot h}{||h||} = 0 \tag{1.3}$$

If f is differentiable at x_0, then m is the gradient of f at x_0, denoted $\nabla f(x_0)$.

Suppose the $S \subseteq \mathbb{R}^n$ and $f : S \to \mathbb{R}$ is a function.

Definition 1.0.4. *Limit, multivariate function*
We say that the limit of f at x_0 is L if for all $\epsilon > 0$, there exists $\delta > 0$ such that for all x such that $||x - x_0|| < \delta$, we have $|f(x) - L| < \epsilon$.

Definition 1.0.5. *Diffrentiability, multivariable (alternative)*
We say that f is differentiable at x_0 if there exists a vector $m \in \mathbb{R}^n$ such that:

$$f(x_0 + h) = f(x_0) + m^T \cdot h + E(h) \text{ where } \lim_{h \to 0} \frac{E(h)}{||h||} = 0 \tag{1.4}$$

If f is differentiable at x_0, then m is the gradient of f at x_0, denoted $\nabla f(x_0)$.

Definition 1.0.6. *Partial Derivative*
The partial derivative of f with respect to the i-th variable at x is:

$$\frac{\partial f}{\partial x_i}(x) = \lim_{h \to 0} \frac{f(x + h \cdot e_i) - f(x)}{h} \tag{1.5}$$

where e_i is the i-th standard basis vector.

Theorem 1.0.1. *(Diffrentiability vs. Partial Derivatives)*
If f is differentiable at x, then all partial derivatives of f exist at x and:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right) \tag{1.6}$$

- If any partial derivative of f does not exist at x, then f is not differentiable at x.
- If all partial derivatives of f exist at x, then f may still not be differentiable at x and the vector $m = \nabla f(x)$ is the only possible vector that satisfies the definition of differentiability.

Definition 1.0.7. *Continuously Differentiable*
We say that f is continuously differentiable or of class C^1 if all partial derivatives of f exist and are continuous at every point in S.

Theorem 1.0.2. *If f is continuously differentiable, then f is differentiable.*

Definition 1.0.8. *The directional derivative*
For a given $x \in S$ and a unit vector $u \in \mathbb{R}^n$, the directional derivative of f at x in the direction of u is:

$$\partial_u f(x) = \lim_{h \to 0} \frac{f(x + h \cdot u) - f(x)}{h} \tag{1.7}$$

Equivalently, $\partial_u f(x) = g'(0)$ where $g(h) = f(x + h \cdot u)$.

Theorem 1.0.3. *If f is differentiable at x, then for all $u \in \mathbb{R}^n$, the directional derivative of f at x in the direction of u exists and is given by:*

$$\partial_u f(x) = \nabla f(x) \cdot u \tag{1.8}$$

Theorem 1.0.4. *Fermat's Theorem*
If f is differentiable at x and x is a local minimum of f, then $\nabla f(x) = 0$.

Theorem 1.0.5. *Suppose that $f : S \to \mathbb{R}$ is differentiable at x. Then $\nabla f(x)$ is orthogonal to the level set of f that passes through x.*

Theorem 1.0.6. *The mean value theorem*
If $f : S \to \mathbb{R}$ is differentiable on the open interval between a and b, then there exists $c \in [a, b]$ such that:

$$f(b) - f(a) = \nabla f(c) \cdot (b - a) \tag{1.9}$$

where $[a, b] = a + t(b - a) | t \in [0, 1]$.

Definition 1.0.9. *Second-order partial derivatives*
Suppose that f is a C^1 function. If the partial derivatives of f are differentiable, then the second-order partial derivatives of f are:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right) \tag{1.10}$$

Equivalently, $\frac{\partial^2 f}{\partial i \partial j} = \partial_j \partial_j f$. If $i = j$ we denote $\frac{\partial^2 f}{\partial x_i^2}$ or $(\partial_i^2 f$

Definition 1.0.10. *The C^2 class*
We say that f is of class C^2 if all second-order partial derivatives of f exist and are continuous.

Theorem 1.0.7. *Clairaut's Theorem*
If f is of class C^2, then $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$.

Definition 1.0.11. *Hessian Matrix*
The Hessian matrix of f at x is the matrix of second-order partial derivatives of f at x:

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{1.11}$$

Corollary. *The interpretation of the Hessian matrix*
Let $u \in \mathbb{R}^n$ be a unit vector. then

$$\partial_{uu}^2 f(x) = \sum_{i,j=1}^{n} \partial_{ij} f(x) u_i u_j = u^T \nabla^2 f(x) u \tag{1.12}$$

1.1 Taylor series

Definition 1.1.1. *Taylor Series*
Let $f : \mathbb{R} \to \mathbb{R}$ be a function that is k times differentiable at x_0. Then the Taylor series of f at x_0 is given by:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots + \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + R_k(x) \tag{1.13}$$

where $R_k(x) = \frac{f^{(k+1)}(c)}{(k+1)!}(x - x_0)^{k+1}$ for some c between x and x_0.

Definition 1.1.2. *Taylor Series for Multivariable Functions (k=2)*
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that is C^2 at x_0. Then for any h such that $x_0 + h \in S$, there exists $\theta \in [0, 1]$ such that:

$$f(x_0 + h) = f(x_0) + \nabla f(x_0) \cdot h + \frac{1}{2} h^T \nabla^2 f(x_0 + \theta h) h \tag{1.14}$$

Chapter 2

Algebraic Structures

2.0.1 Properties

Definition 2.0.1. *Closure (sgirot)*
An operation $*$ on a set G is said to have the property of closure if for every $a, b \in G$, the result $a * b$ is also in G.

Definition 2.0.2. *Commutativity (hilofiot)*
An operation $*$ on a set G is commutative if for every $a, b \in G$, we have $a * b = b * a$.

Definition 2.0.3. *Associativity*
An operation $*$ on a set G is associative if for every $a, b, c \in G$, we have $(a * b) * c = a * (b * c)$.

Definition 2.0.4. *Distributivity*
An operation $*$ on a set G is distributive if for every $a, b, c \in G$, we have $a*(b+c) = a*b+a*c$.

Definition 2.0.5. *Identity (zehot)*
An operation $*$ on a set G has an identity element if there exists an element $e \in G$ such that for every $a \in G$, $a * e = e * a = a$.

Definition 2.0.6. *Inverse (ofchiot)*
An operation $*$ on a set G has inverses if for every $a \in G$, there exists an element $b \in G$ such that $a * b = b * a = e$, where e is the identity element.

2.0.2 Structures

Group

Definition 2.0.7. ***Group*** *(havura)*
A group is a set G along with an operation $*$ such that $\forall a, b, c \in G$ the following properties hold:

1. *$a * b \in G$ (closure)*
2. *$(a * b) * c = a * (b * c)$ (associativity)*
3. *There exists an element $e \in G$ such that $a * e = e * a = a$ (identity)*
4. *For each $a \in G$ there exists $b \in G$ such that $a * b = b * a = e$ (inverse)*

Example. *Examples of groups:*

1. *$(\mathbb{R}, +)$ is a group.*
2. *$(\mathbb{Z}, +)$ is a group.*
3. *Non-zero reals, complex, and rational numbers are groups under multiplication.*

Definition 2.0.8. ***Abelian Group***
An abelian group is a group $(G, *)$ in which the binary operation $*$ is commutative, meaning that for all $a, b \in G$, $a * b = b * a$.

Ring

Definition 2.0.9. ***Ring*** *(hug)*
A ring is a set R equipped with two binary operations $+$ (addition) and $\times$ (multiplication) satisfying the following three sets of axioms:

1. *R is an **abelian group** under addition, meaning that:*
   - $(a + b) + c = a + (b + c)$ *for all* $a, b, c \in R$ *(associativity).*
   - $a + b = b + a$ *for all* $a, b \in R$ *(commutativity).*
   - *There is an element* $0 \in R$ *such that* $a + 0 = a$ *for all* $a \in R$ *(additive identity).*
   - *For each* $a \in R$ *there exists* $-a \in R$ *such that* $a + (-a) = 0$ *(additive inverse).*
2. *R is a **monoid** under multiplication, meaning that:*
   - $(a \times b) \times c = a \times (b \times c)$ *for all* $a, b, c \in R$ *(associativity).*
   - *There is an element* $1 \in R$ *such that* $a \times 1 = a$ *and* $1 \times a = a$ *for all* $a \in R$ *(multiplicative identity).*
3. *Multiplication is distributive with respect to addition, meaning that:*
   - $a \times (b + c) = (a \times b) + (a \times c)$ *for all* $a, b, c \in R$ *(left distributivity).*
   - $(b + c) \times a = (b \times a) + (c \times a)$ *for all* $a, b, c \in R$ *(right distributivity).*

Example. *Examples of rings:*

1. $(\mathbb{Z}, +, \times)$ *is a ring.*
2. $(\mathbb{R}, +, \times)$ *is a ring.*
3. *The set of odd integers is not a ring because it is not closed under addition.*

Field

Definition 2.0.10. ***Field*** *(sadeh)*
A field is a set F with two operations, addition $+$ and multiplication $\times$, such that:

1. $(F, +)$ *is an **abelian group** with the identity element* 0 *(additive identity).*
2. $(F \setminus \{0\}, \times)$ *is an **abelian group** with the identity element* 1 *(multiplicative identity).*
3. *Multiplication is distributive with respect to addition, meaning that:*
   - $a \times (b + c) = (a \times b) + (a \times c)$ *for all* $a, b, c \in R$ *(left distributivity).*
   - $(b + c) \times a = (b \times a) + (c \times a)$ *for all* $a, b, c \in R$ *(right distributivity).*

Example. *Examples of fields:*

1. $(\mathbb{R}, +, \times)$ *is a field.*
2. $(\mathbb{Q}, +, \times)$ *is a field.*
3. $(\mathbb{C}, +, \times)$ *is a field.*
4. $(\mathbb{Z}_p, +, \times)$ *for a prime p is a field.*

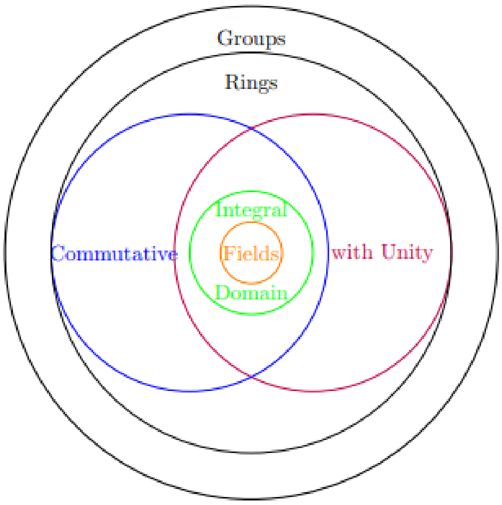


Figure 2.1: Algebraic Structures

2.0.3 Spaces

Vector Space

A major difference between a field and a vector space is that the operations on a field $\mathbb{F}$ are

- $+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$
- $\times : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$

but the operations on a vector space $\mathbb{V}$ over a field $\mathbb{F}$ are

- $+ : \mathbb{V} \times \mathbb{V} \to \mathbb{V}$
- $\cdot : \mathbb{F} \times \mathbb{V} \to \mathbb{V}$

Definition 2.0.11. ***Vector Space***
A vector space over a field F is a non-empty set V together with two operations: vector addition $+$ and scalar multiplication $\cdot$, satisfying the following axioms for every $u, v, w \in V$ and $a, b \in F$:

1. *Associativity of vector addition: $u + (v + w) = (u + v) + w$*
2. *Commutativity of vector addition: $u + v = v + u$*
3. *Identity element of vector addition: There exists an element $0 \in V$, called the **zero vector**, such that $v + 0 = v$ for all $v \in V$.*
4. *Inverse elements of vector addition: For every $v \in V$, there exists an element $-v \in V$, called the **additive inverse** of v, such that $v + (-v) = 0$.*
5. *Compatibility of scalar multiplication with field multiplication: $a(bv) = (ab)v$*
6. *Identity element of scalar multiplication: $1v = v$, where 1 denotes the multiplicative identity in F.*
7. *Distributivity of scalar multiplication with respect to vector addition: $a(u + v) = au + av$*
8. *Distributivity of scalar multiplication with respect to field addition: $(a + b)v = av + bv$*

Example. *Examples of vector spaces:*

1. $\mathbb{R}^n$ *is a vector space.*
2. *The space* $M_{m\times n}(\mathbb{F})$ *of* $m \times n$ *matrices over a field* $\mathbb{F}$ *is a vector space.*
3. *The set of all continuous functions over some interval is a vector space.*
4. *The space of all differentiable functions over a certain interval is a vector space.*

Definition 2.0.12. ***Complex conjugate***
The complex conjugate of a complex number $z = a + bi$ *is the number* $\overline{z} = a - bi$*.*

Inner Product Space

Definition 2.0.13. ***Inner Product Space***
An inner product space is a vector space V *over a field* F *equipped with an inner product, which is a function that associates each pair of vectors* u, v *in* V *with a scalar in* F*, denoted* $\langle u, v\rangle$*, and satisfies the following properties for all* $u, v, w \in V$ *and* $a \in F$*:*

1. *Linearity in the first argument:* $\langle au + v, w\rangle = a\langle u, w\rangle + \langle v, w\rangle$
2. *Conjugate symmetry:* $\langle u, v\rangle = \overline{\langle v, u\rangle}$
3. *Positive-definiteness:* $\langle u, u\rangle \geq 0$ *and* $\langle u, u\rangle = 0$ *if and only if* $u = 0$

Definition 2.0.14. ***Hermetian adjoint***
Let V *be an inner product space over a field* F*. The Hermetian adjoint of a linear operator* $T : V \to V$ *is the unique linear operator* $T^* : V \to V$ *such that for all* $u, v \in V$*, we have* $\langle Tu, v\rangle = \langle u, T^*v\rangle$*.*

A classic exmaple of an inner product space is the **Euclidean space** $\mathbb{E}^n$, which is a vector space equipped with the inner product $\langle u, v\rangle = u^T v$. The geometry of Euclidean space follows the familiar rules of Euclidean geometry, which include notions such as angles, lengths, and the Pythagorean theorem. It is always complete, meaning that every Cauchy sequence in Euclidean space converges to a point within the space. Euclidean space can be thought of as the "standard" n-dimensional space that conforms to our intuitive geometric concepts.

Metric and complete metric spaces

Definition 2.0.15. ***Metric Space***
A metric space is a set X *equipped with a metric, which is a function that defines a distance between each pair of elements in* X*, satisfying the following properties for all* $x, y, z \in X$*:*

1. *Non-negativity:* $d(x, y) \geq 0$ *and* $d(x, y) = 0$ *if and only if* $x = y$
2. *Symmetry:* $d(x, y) = d(y, x)$
3. *Triangle inequality:* $d(x, y) + d(y, z) \geq d(x, z)$

Inner Producy Space vs Metric Space

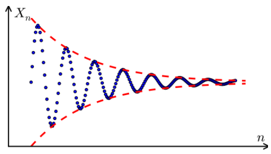
- An inner product space is a vector space equipped with an inner product, which is a function that associates each pair of vectors with a scalar.
- A metric space is a set equipped with a metric, which is a function that defines a distance between each pair of elements in the set.

- Every inner product space is a metric space, but not every metric space is an inner product space.

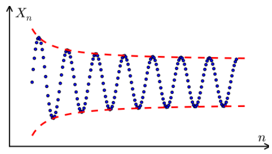
  Example. *Each inner product space must satisfy the Parallelogram Law, which states that for all u, v in the space, $2\|u\|^2 + 2\|v\|^2 = \|u + v\|^2 + \|u - v\|^2$. A clasic example of a metric space that is not an inner product space is the space of continuous functions on the interval $[0, 1]$ with the metric $d(f, g) = \max_{x\in[0,1]} |f(x) - g(x)|$.*

Definition 2.0.16. ***Cauchy Sequence***
In a metric space (X, d), a sequence $\{x_1, x_2, x_3, \ldots\}$ is said to be Cauchy *if, for every positive real number $\varepsilon > 0$, there exists a positive integer N such that for all positive integers $m, n > N$, the distance*

$$d(x_m, x_n) < \varepsilon.$$



(a) The plot of a Cauchy sequence (x_n) shown in blue, as x_n versus n. If the space is complete, then the sequence has a limit.



(b) A sequence that is not Cauchy. The elements of the sequence do not get arbitrarily close to each other as the sequence progresses.

Figure 2.2: Cauchy Sequence

Definition 2.0.17. ***Complete Metric Space***
A metric space is complete if every Cauchy sequence in the space converges to a limit in the space.

Hilbert Space

Definition 2.0.18. ***Hilbert Space***
A Hilbert space *is a complete inner product space. That is, it is an inner product space $\mathcal{H}$ that is also a complete metric space with respect to the metric induced by its inner product. The metric is given by $d(x, y) = \sqrt{\langle x - y, x - y\rangle}$ for all $x, y \in \mathcal{H}$.*

A Hilbert space generalizes the notion of Euclidean space to an infinite-dimensional context. In a Hilbert space, one can still use concepts like angle, orthogonality, and projection, which are crucial in many areas of mathematics and physics. Completeness is a key feature of Hilbert spaces, meaning that every Cauchy sequence in the space converges to a limit within the space.

2.0.4 Topological Spaces

Definition 2.0.19. ***Topological Space***
A topological space is a set X equipped with a collection of open sets $\mathcal{T}$ that satisfy the following properties:

1. *The empty set and the entire space are open: $\emptyset, X \in \mathcal{T}$.*
2. *The intersection of any finite number of open sets is open.*
3. *The union of any number of open sets is open.*

Example. *Examples of topological spaces:*

1. *$X = 1, 2, 3$ with the topology $T = \{\varnothing, 1, 1, 2, 1, 2, 3\}$.*
2. *Given any set X, the discrete topology on X is the power set of X - $\mathcal{T} = P(X)$ and is the largest possible topology on X.*
3. *The indiscrete topology on X is $\mathcal{T} = \{\emptyset, X\}$ and is the smallest possible topology on X - $\forall \mathcal{T}$, $\{\emptyset, X\} \subseteq \mathcal{T} \subseteq P(X)$.*
4. *The real line $\mathbb{R}$ with the standard topology.*
5. *The set of integers $\mathbb{Z}$ with the discrete topology.*
6. *The set of real numbers $\mathbb{R}$ with the lower limit topology.*

Chapter 3

Convexity and Optimization

3.1 Important subsets of $\mathbb{R}^n$

Definition 3.1.1. *Open set*
A set $S \subseteq \mathbb{R}^n$ is open if for all $x \in S$, there exists $\epsilon > 0$ such that $B(x, \epsilon) \subseteq S$.

Definition 3.1.2. *Closed set*
A set $S \subseteq \mathbb{R}^n$ is closed if its complement is open.

Definition 3.1.3. *Interior point*
A point $x \in S$ is an interior point of S if there exists $\epsilon > 0$ such that $B(x, \epsilon) \subseteq S$.

Corollary 3.1.1. *Open set characterization*
A set $S \subseteq \mathbb{R}^n$ is open if and only if every point in S is an interior point of S.

Definition 3.1.4. *Boundary point*
A point $x \in S$ is a boundary point of S if for all $\epsilon > 0$, $B(x, \epsilon) \cap S \neq \emptyset$ and $B(x, \epsilon) \cap S^c \neq \emptyset$.

Definition 3.1.5. *Half-space*
A half-space in $\mathbb{R}^n$ is a set of the form $\{x \in \mathbb{R}^n : a^T x \leq b\}$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

Definition 3.1.6. *Hyperplane*
A hyperplane in $\mathbb{R}^n$ is a set of the form $\{x \in \mathbb{R}^n : a^T x = b\}$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

Definition 3.1.7. *Polyhedron (Polyhedra)*
A polyhedron in $\mathbb{R}^n$ is a set of the form $\{x \in \mathbb{R}^n : Ax \leq b\}$ for some $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Equivalently, a polyhedron is the intersection of finitely many half-spaces.

Definition 3.1.8. *Polytope*
A polytope in $\mathbb{R}^n$ is a bounded polyhedron - i.e., there exists $r > 0$ such that $\forall x \in \{x \in \mathbb{R}^n : Ax \leq b\} \implies ||x|| \leq r$. Equivalently, a polytope is the convex hull of finitely many points.

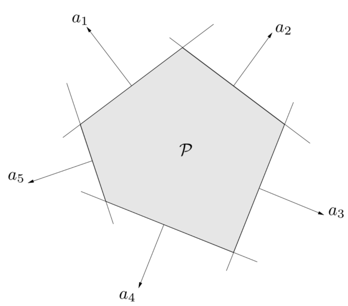


Figure 3.1: Polytope

Definition 3.1.9. *Convex set*
A set $S \subseteq \mathbb{R}^n$ is convex if for all $x, y \in S$ and $\lambda \in [0,1]$, we have $\lambda t + (1-\lambda)y \in S$.

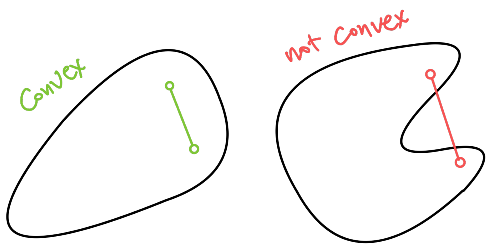


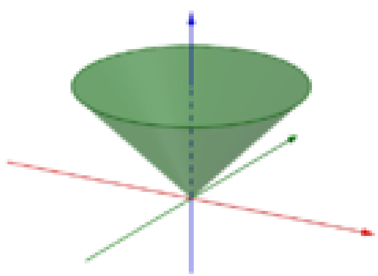
Figure 3.2: Convex set

Definition 3.1.10. *Convex hull*
The convex hull of a set $S \subseteq \mathbb{R}^n$ is the smallest convex set that contains S.

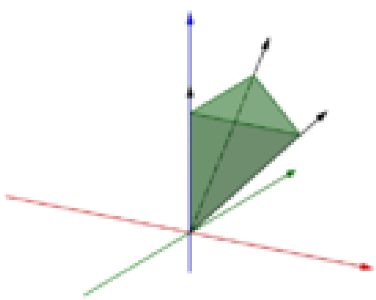
Definition 3.1.11. *Conic combination*
A point $x \in \mathbb{R}^n$ is a conic combination of $y_1, \ldots, y_k \in \mathbb{R}^n$ if there exist $\lambda_1, \ldots, \lambda_k \geq 0$ such that $x = \sum_{i=1}^k \lambda_i y_i$.

Definition 3.1.12. *Conic hull*
The conic hull of a finite set $S \subseteq \mathbb{R}^n$ is the set of all conic combinations of points in S.

Definition 3.1.13. *Convex cone*
A set $S \subseteq \mathbb{R}^n$ is a convex cone if for all $x \in S$ and $\lambda \geq 0$, we have $\lambda x \in S$.



(a) Convex cone that is not a conic hull of finitely many generators.



(b) Convex cone genrated by the conic combination of three black vectors (conic hull).

Definition 3.1.14. *Normal cone*
The normal cone to a set S at a point x is defined as

$$N_S(x) = \{v \in \mathbb{R}^n : \langle v, y - x \rangle \leq 0 \text{ for all } y \in S\} \tag{3.1}$$

Definition 3.1.15. *Tangent cone*
The tangent cone to a set S at a point x is defined as

$$T_S(x) = \{v \in \mathbb{R}^n : \lim_{t \to 0^+} \frac{x + tv - x}{t} \in S\} \tag{3.2}$$

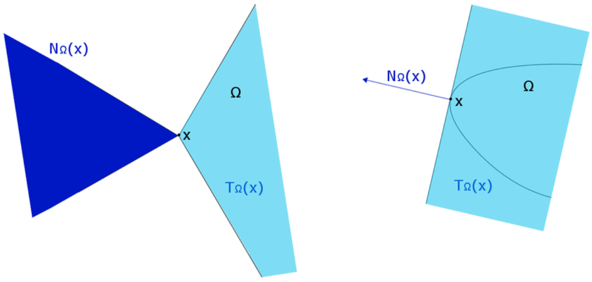


Figure 3.4: Normal and tangent cones

Theorem 3.1.1. *Normal cone of polyhedron*
The normal cone to a polyhedron $S = \{x \in \mathbb{R}^n : \forall j \in [m] \quad a_j \cdot x \leq b_j\}$ *at a point* x *is given by*

$$N_S(x) = \{\sum_j \lambda_j a_j : \lambda_j \geq 0 \text{ and } a_j \cdot x = b_j\} \tag{3.3}$$

3.2 Definitions and Fundamental Theorems

Definition 3.2.1. *(Convex function): A function* $f : S \to \mathbb{R}$ *defined on a convex set* S *is convex if, for all* $x, y \in S$ *and* $\lambda \in [0, 1]$*,*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

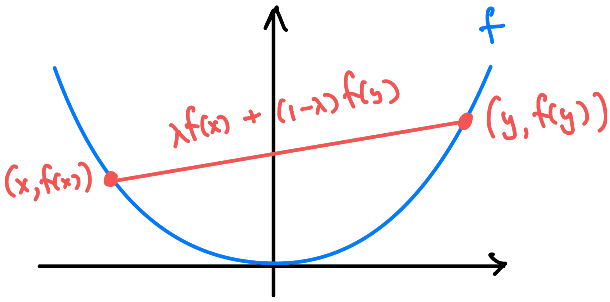


Figure 3.5: Convex function

Theorem 3.2.1. *(Characterization via epigraph): A function* $f : S \to \mathbb{R}$ *is convex if and only if its epigraph* $\{(x, t) \in S \times \mathbb{R} : f(x) \leq t\}$ *is a convex set.*

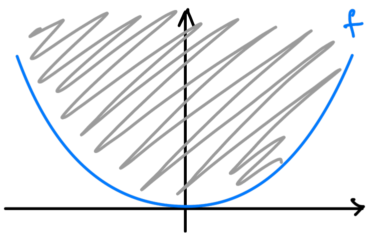


Figure 3.6: Epigraph of a convex function

claim 3.2.1. *(Convexity of sublevel sets): If* $f : S \to \mathbb{R}$ *is convex, then the sublevel set* $S_t = \{x \in S : f(x) \leq t\}$ *is convex for any* $t \in \mathbb{R}$*.*

3.3 Inequalities and Characterizations

Theorem 3.3.1. *(Jensen's inequality): If f is a convex function, then for any $x_1, x_2, \ldots, x_n \in S$ and any non-negative weights α_i such that $\sum_{i=1}^n \alpha_i = 1$,*

$$f\left(\sum_{i=1}^n \alpha_i x_i\right) \leq \sum_{i=1}^n \alpha_i f(x_i).$$

Theorem 3.3.2. *(First-order characterization, aka "the gradient inequality"): If f is a differentiable convex function on an open set S, then for all $x, y \in S$,*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

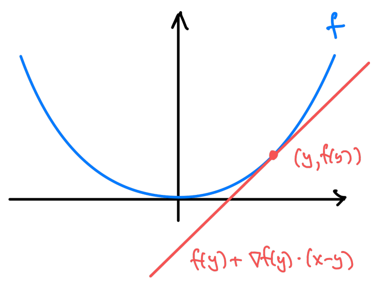


Figure 3.7: First-order characterization of convexity

Definition 3.3.1. *Bergman divergence (distance)*
The Bergman divergence between two points $x, y \in \mathbb{R}^n$ is defined as

$$D_f(x, y) = f(x) - f(y) - \nabla f(y)^\top (x - y) \tag{3.4}$$

Theorem 3.3.3. *(Jensen's inequality, generalized for expectation): If f is a convex function and X is a random variable over S, then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Theorem 3.3.4. *(Second-order characterization of convexity): A twice differentiable function f is convex on an open set S if and only if the Hessian matrix of f is positive semidefinite at every point in S.*

3.4 Optimization and Projection

Definition 3.4.1. *(Convex optimization): The problem of minimizing a convex function over a convex set.*

Theorem 3.4.1. *(Optimality conditions, unconstrained): If f is convex and differentiable, x^* is a local minimum of f $\Leftrightarrow$ x^* is a global minimum of f $\Leftrightarrow$ $\nabla f(x^*) = 0$.*

Theorem 3.4.2. *(Optimality conditions, constrained): If f is differentiable and C is a convex set, x^* is a local minimum of f on C if and only if $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all $x \in C$.*

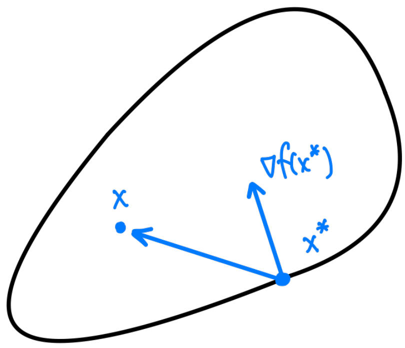


Figure 3.8: Optimality conditions, constrained

Corollary 3.4.1. *Optimality conditions, constrained (alternative)*
If f is differentiable and C is a convex set, then x^* is a local minimum of f on C if and only if $-\nabla f(x^*) \in N_C(x^*)$.

Definition 3.4.2. *(Projection): The projection of a point x onto a convex set S is defined as* $\Pi_S(x) = \arg\min_{y \in S} \|y - x\|$.

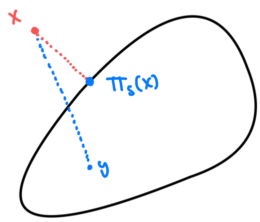


Figure 3.9: Projection

Theorem 3.4.3. *Generalized cosine theorem*
Let $S \subseteq \mathbb{R}^d$ be convex and $x \in \mathbb{R}^d$. Then the projection $\Pi_S[x]$ is unique and satisfies:

$$\|x - \Pi_S[x]\|^2 + \|\Pi_S[x] - y\|^2 \leq \|x - y\|^2, \quad \forall y \in S. \tag{3.5}$$

In particular:

$$\|\Pi_S[x] - y\| \leq \|x - y\|, \quad \forall y \in S. \tag{3.6}$$

3.5 Properties of Convex Functions

Definition 3.5.1. *L - Lipschitz continuous*
A function $f : S \to \mathbb{R}$ is L-Lipschitz continuous if for all $x, y \in S$,

$$|f(x) - f(y)| \leq L||x - y|| \tag{3.7}$$

Theorem 3.5.1. *Convexity and Lipschitz continuity*
If f is convex, differentiable and L-Lipschitz continuous, then $||\nabla f(x)|| \leq L$ for all $x \in S$.

Definition 3.5.2. *Smooth function*
A differentiable function f is β-smooth over $S \subseteq domf$ if for all $x, y \in S$:

$$-\frac{\beta}{2}\|y-x\|^2 \leq f(y) - f(x) - \nabla f(x) \cdot (y-x) \leq \frac{\beta}{2}\|y-x\|^2.$$

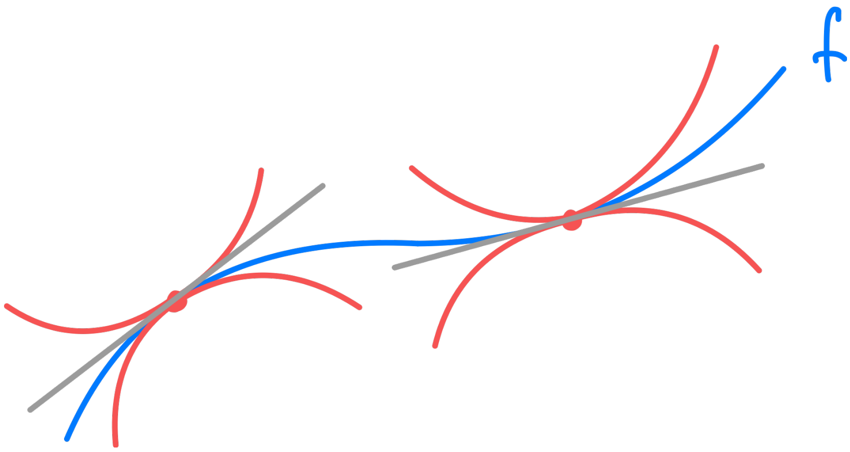


Figure 3.10: Smooth function

Theorem 3.5.2. *Lipschitz gradient interpretation*
Let f be differentiable and let $S \subseteq domf$ be convex and closed. Suppose that

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x-y\|, \quad \forall x, y \in S.$$

Then f is β-smooth over S.

Theorem 3.5.3. *Second-order characterization of smoothness*
Let f be C^2 and let $S \subseteq domf$ be convex and closed. Then f is β-smooth over S if and only if

$$-\beta I \preceq \nabla^2 f(x) \preceq \beta I, \quad \forall x \in S.$$

Lemma 3.5.1. *The Descent Lemma*
Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth, and let $x \in \mathbb{R}^d$.

- For $\eta \leq \frac{1}{\beta}$, $x^+ = x - \eta\nabla f(x)$, we have

$$f(x^+) - f(x) \leq -\frac{\eta}{2}\|\nabla f(x)\|^2.$$

- For $x^* \in \arg\min_x f(x)$, we have

$$\frac{1}{2\beta}\|\nabla f(x)\|^2 \leq f(x) - f(x^*).$$

Basic Facts:

- An affine function $f : \mathbb{R}^d \to \mathbb{R}$, $f(x) = a^\top x + b$, is 0-smooth.
- A quadratic function $f : \mathbb{R}^d \to \mathbb{R}$, $f(x) = \frac{1}{2}x^\top Ax + b^\top x + c$, is $\lambda_{\max}(A)$-smooth.

- A linear combination of smooth functions is smooth with an appropriate parameter.
- A convex combination of β-smooth functions is β-smooth.

Definition 3.5.3. *Strong convexity*
A function f is α-strongly convex (for $\alpha \geq 0$) over a convex and closed set $S \subseteq domf$ if for any $x \in S$, there exists $g_x \in \partial f(x)$ such that:

$$\forall y \in S, \quad f(y) \geq f(x) + g_x \cdot (y-x) + \frac{\alpha}{2}\|y-x\|^2.$$

In particular, a differentiable f is α-strongly convex over S if for any $x \in S$,

$$\forall y \in S, \quad f(y) \geq f(x) + \nabla f(x) \cdot (y-x) + \frac{\alpha}{2}\|y-x\|^2.$$

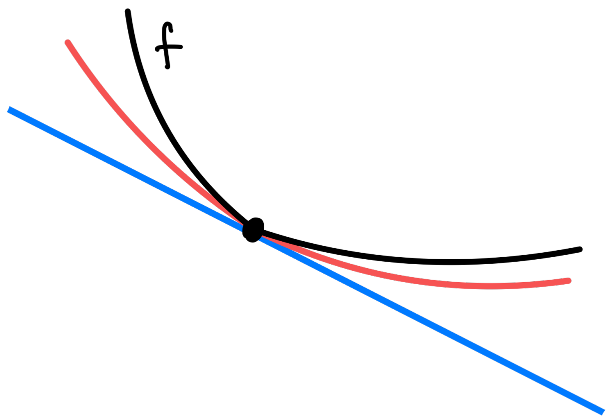


Figure 3.11: Strongly convex function

Theorem 3.5.4. *Strong convexity, second-order characterization*
Let f be C^2 and let $S \subseteq domf$ be convex and closed. Then f is α-strongly convex over S if and only if

$$\forall x \in S, \quad \nabla^2 f(x) \succeq \alpha I.$$

Theorem 3.5.5. *Usage of strong convexity*
If a differentiable f is α-strongly convex over a convex and closed $S \subseteq domf$ with a minimum at $x^* \in S$, then

$$\forall x \in S, \quad \frac{\alpha}{2}\|x-x^*\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2.$$

In particular, the minimum of a strongly convex function is unique.

3.6 Important Inequalities

Theorem 3.6.1. $1 + x \leq e^x$
For all $x \in \mathbb{R}$, we have $1 + x \leq e^x$.

Proof. Let $f(x) = e^x - 1 - x$. Then $f'(x) = e^x - 1$ and $f''(x) = e^x > 0$. Thus, f is convex and $f(0) = 0$. Therefore, $f(x) \geq 0$ for all $x \in \mathbb{R}$. □

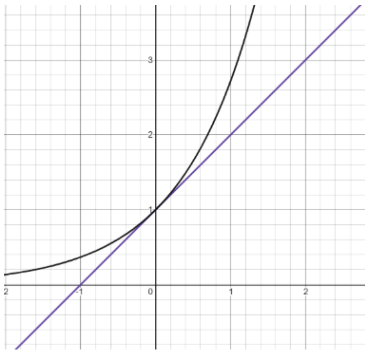


Figure 3.12: $1 + x \leq e^x$

Chapter 4

The First Miracle: Robustness

Let f be a convex function, and let x^* be a minimizer of f.

4.1 Gradient Descent

Definition 4.1.1. *Gradient Descent*

$$x_{t+1} = x_t - \eta \nabla f(x_t) \tag{4.1}$$

It holds that:

$$f(x^*) \geq f(x_t) + \nabla f(x_t) \cdot (x^* - x_t) \tag{4.2}$$

$$0 \leq f(x_t) - f(x^*) \leq \nabla f(x_t) \cdot (x_t - x^*) \tag{4.3}$$

4.1.1 Analysis of the Gradient Descent Algorithm

$$\begin{aligned} &\|a\|^2 = \|b\|^2 + \|a-b\|^2 \\ &\|b\|^2 = \|a\|^2 - \|a-b\|^2 = \|a\|^2 - (\|a\|^2 - 2a \cdot b + \|b\|^2) = 2a \cdot b - \|b\|^2 \end{aligned}$$

Then we have:

$$\begin{aligned} \|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2 &= -2\eta(x^* - x_t) \cdot \nabla f(x_t) - \eta^2 \|\nabla f(x_t)\|^2 \\ &= 2\eta(x_t - x^*) \cdot \nabla f(x_t) - \eta^2 \|\nabla f(x_t)\|^2 \\ &\geq 2(f(x_t) - f(x^*)) - \eta^2 L^2 \end{aligned}$$

Where the last inequality follows from the convexity and the Lipschitz continuity of f.

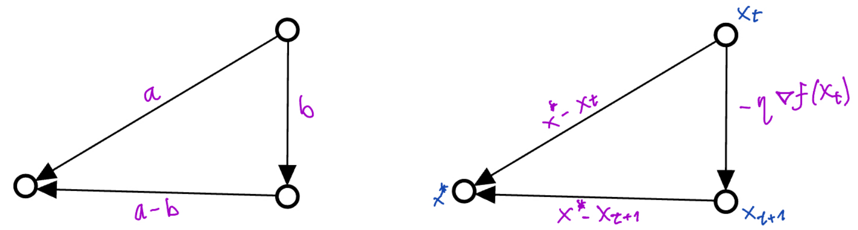


Figure 4.1: Gradient Descent

Then if we sum the above inequality from $t = 1$ to T, we get:

$$\sum_{t=1}^{T}(f(x_t) - f(x^*)) \leq \frac{\|x_1 - x^*\|^2}{2\eta} + \frac{\eta L^2}{2}T$$

In fact, this is a specific case of the Fundamental Inequality of Optimization.

Theorem 4.1.1. *Fundamental Inequality of Optimization (unconstrained version)*
Suppose $x_{t+1} = x_t - \eta g_t$ *for all* t*, where* $g_1, \ldots, g_T \in \mathbb{R}^d$ *are arbitrary vectors. Then for all* $x^* \in \mathbb{R}^d$ *it holds that*

$$\sum_{t=1}^{T} g_t \cdot (x_t - x^*) \leq \frac{\|x_1 - x^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|^2.$$

Proof. Fundamental Inequality of Optimization
The proof tracks $\|x_t - x^*\|^2$ as a "potential". First write

$$\|x_{t+1} - x^*\|^2 = \|(x_t - x^*) - \eta g_t\|^2 = \|x_t - x^*\|^2 - 2\eta g_t \cdot (x_t - x^*) + \eta^2\|g_t\|^2,$$

that is,

$$\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 = 2\eta g_t \cdot (x_t - x^*) - \eta^2\|g_t\|^2.$$

Summing over $t = 1, \ldots, T$ and telescoping terms, we obtain

$$\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2 = 2\eta\sum_{t=1}^{T} g_t \cdot (x_t - x^*) - \eta^2\sum_{t=1}^{T}\|g_t\|^2.$$

Organizing terms, we conclude:

$$\sum_{t=1}^{T} g_t \cdot (x_t - x^*) \leq \frac{\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|^2.$$

$\square$

osing eta

Chapter 5

The Second Miracle: Potential Based

5.1 Experts Problem

At each time step, the player picks an action $I_t \in [n]$ (we have n experts) and the adversary picks a loss vector $l_t \in 0, 1^n$. The player incurs loss $l_t(I_t)$ and the goal is to minimize the regret:

$$\text{Regret}_T(i) = \sum_{t=1}^{T} (l_t(I_t) - l_t(i)) \tag{5.1}$$

We consider the case where in each time step the player chooses an action from a distribution $\vec{p}$ over the n experts (a vector from the simplex):

$$\vec{p} \in \triangle_n = \{\vec{p} \in \mathbb{R}^n_+ : p_i \geq 0, \sum_{i=1}^{n} p_i = 1\}$$

Approach 1: Gradient Descent

We can use gradient descent on $f_t(\vec{p_t}) = \vec{l_t} \cdot \vec{p}$, where $\vec{l_t}$ is the loss vector at time t. It holds that $\nabla f_t(\vec{p_t}) = \vec{l_t}$. We can use the analysis of the gradient descent algorithm for gradient descent of convex functions varying in time.
Let $q \in \triangle_n$ be any distribution. Then we have:

$$\begin{aligned} f_t(q) \geq f_t(\vec{p_t}) + \nabla f_t(q) \cdot (q - \vec{p_t}) \Longrightarrow \\ f_t(\vec{p_t}) - f_t(q) \leq \nabla f_t(q) \cdot (\vec{p_t} - q) \end{aligned}$$

Then:

$$\begin{aligned} \|q - p_t\|^2 - \|q - p_{t+1}\|^2 &= -2\eta(q - p_t) \cdot \nabla f_t(p_t) - \eta^2 \|\nabla f_t(p_t)\|^2 \Longrightarrow \\ f_t(\vec{p_t}) - f_t(q) &\leq \nabla f_t(q) \cdot (\vec{p_t} - q) = \frac{1}{2\eta}\left(\|q - \vec{p_t}\|^2 - \|q - \vec{p}_{t+1}\|^2\right) + \frac{\eta}{2}\|\nabla f_t(\vec{p_t})\|^2 \Longrightarrow \\ \sum_{t=1}^{T} (f_t(\vec{p_t}) - f_t(q)) &\leq \frac{1}{2\eta}\left(\|q - \vec{p}_1\|^2 - \|q - \vec{p}_{T+1}\|^2\right) + \frac{\eta}{2}\sum_{t=1}^{T} \|\nabla f_t(\vec{p_t})\|^2 \\ &\leq \frac{1}{2\eta}\|q - \vec{p}_1\|^2 + \frac{\eta}{2}\sum_{t=1}^{T} \|\nabla f_t(\vec{p_t})\|^2 \\ &\leq \frac{1}{\eta} + \frac{\eta}{2} Tn = \mathbf{O}(\sqrt{Tn}) \end{aligned}$$

We have used the facts that:

- Both q and $\vec{p}_1$ are distributions, so $\|q - \vec{p}_1\|^2 \leq 2$.
- $\|\nabla f_t(\vec{p}_t)\|^2 \leq n$ (as the loss vector is in $0, 1^n$).

we can see that in this case, the rate of convergence DO depend on the dimension of the problem, in contrast to the non-varying case. The fact that the rate of convergence DO NOT depend on the dimension of the problem in GD is one of the reasons why GD is so useful in practice.

2: M

Approach 2: Multiplicative Weights Update (MWU)

5.2 Mirror Descent

Endow K with a Riemannian structure: $< \cdot, \cdot >_x$ for each $x \in K$. Before:

$$x_{t+1} = x_t - \eta \nabla f(x_t) \rightarrow f(x + dx) \approx f(x) + \nabla f(x) \cdot dx \tag{5.2}$$

Chapter 6

The Third Miracle:

Chapter 7

The Fourth Miracle:

Chapter 8

The Fifth Miracle: