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 Professor Sebastian Bubeck, enhanced by Claire Boyer's comprehensive notes, and structured according to Tomer Koren's course on Optimization for Computer Science.

Winter 2024

Contents

Chapter 1

Mathematical Background

1.1 Multivariable Calculus

Definition 1.1.1. *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 1.1.2. *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 1.1.3. *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 1.1.4. *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 1.1.5. *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 1.1.6. *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 1.1.1. *(Diffrentiability vs. Partial Derivatives)*
If f is differentiable at x, then all partial derivatives of f exist at x and:

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

- 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 1.1.7. *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 1.1.2. *If f is continuously differentiable, then f is differentiable.*

Definition 1.1.8. *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 1.1.3. *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 1.1.4. *Fermat's Theorem*
If f is differentiable at x and x is a local minimum of f, then $\nabla f(x) = 0$.

Theorem 1.1.5. *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 1.1.6. *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 1.1.9. *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 1.1.10. *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 1.1.7. *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 1.1.11. *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

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

1.2 Taylor series

Definition 1.2.1. *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 1.2.2. *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:

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

1.3 Convexity

1.3.1 Definitions and Fundamental Theorems

Definition 1.3.1. *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 $\lambda t + (1-\lambda)y \in S$.

Definition 1.3.2. *(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 $\theta \in [0,1]$,*

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

Theorem 1.3.1. *(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.*

claim 1.3.1. *(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 $t \in \mathbb{R}$.*

1.3.2 Inequalities and Characterizations

Theorem 1.3.2. *(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 1.3.3. *(First-order characterization, aka "the gradient inequality"): If f is a differentiable convex function on an open set S, then for all $x, y \in S$,*

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

Theorem 1.3.4. *(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 1.3.5. *(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.*

1.3.3 Optimization and Projection

Definition 1.3.3. *(Convex optimization): The problem of minimizing a convex function over a convex set.*

Theorem 1.3.6. *(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 1.3.7. *(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 $x \in C$.*

Definition 1.3.4. *Normal cone*
The normal cone to a set S at a point x is defined as

$$N_S(x) = \{v \in \mathbb{R}^n : \langle v, y - x \rangle \leq 0 \text{ for all } y \in S\} \tag{1.15}$$

Corollary 1.3.1. *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 1.3.5. *(Projection): The projection of a point x onto a convex set C is defined as $Proj_C(x) = \arg\min_{y \in C} \|y - x\|$.*

Theorem 1.3.8. *(Generalized cosine theorem): In the context of convex sets, this theorem often relates to angles between vectors in normed spaces and can be used to derive or prove geometric properties about projections and distances in convex optimization.*

1.4 Properties of Convex Functions

1.5 Important Inequalities

1.5.1 $\mathbf{1 + x \leq e}^x$