# Chapter 1

# Why Convex Optimization

## 1.1 Introduction

Optimization problems arise naturally in many computer vision and machine learning applications that estimate pixel values, motions, shapes, or model parameters from input images, videos, range sensors, or training data. By formalizing the problem into a concise mathematical form, we obtain an optimization problem whose solution are the model parameters that best fit the observed data and our prior knowledge of the physical world. The next step, finding a solution to the mathematical model, is far from trivial.

Most commercial optimization packages for general optimization are designed to interface with the vision or learning model as black boxes. They can thus fit a large class of problems, in a poor manner, because they use minimal assumptions on the structure or properties of the problem and results in poor optimizations that do not converge to global optima but stationary points or local minima. We must accept that, in general, optimization problems are unsolvable.

In many practical applications the process of creating a model takes a considerable amount of time and effort. Therefore, it is important to understand the properties of the model and the computational consequences of each decision. Very often we have to choose between a *good* model, which we cannot solve and a *bad* model, which can be solved efficiently. What is better? Convex optimization models are widespread, not because such models can describe our nonlinear world very well, but simply because practitioners prefer to deal with approximate solvable models, predict the effects of the approximation on the solution, and correct it rather than trying to solve a model without any guarantee for success.

The course discusses numerical methods for a large class of solvable optimization problems, namely convex optimization problems, that are common in computer vision and machine learning. To apply the optimization formulations successfully, it is necessary to be aware of some theory that explains what we can and what we cannot do with optimization problems.

## 1.2 Limitations in General Optimization

Let us start by describing our optimization problem. Let $u \in \mathbb{R}^n$ be an $n$-dimensional real vector, $C \subset \mathbb{R}^n$ be a subset of $\mathbb{R}^n$, and $E$ be a real-valued functions of $u$. We study different variants of the following general minimization problem:

$$\hat{u} \in \arg\min_{u \in C} E(u) \tag{1.1}$$

The function $E : \mathbb{R}^n \to \mathbb{R}$ is the objective function, while the set $C$ is the feasible set. We consider a minimization problem by convention, but we can also consider a maximization problem with $-E$ as objective function.

There is a natural classification of the types of minimization problems that we will study: unconstrained problems where $C = \mathbb{R}^n$, smooth problems where $E$ is differentiable, and non-smooth problems where $E$ is not differentiable. We also distinguish different types of solutions to the minimization problem

**Definition** $u^\star$ is a global solution of if $E(u^\star) \geq E(u)$ for all $u \in C$.

**Definition** $u^\star$ is a local solution if there exists a neighborhood of $u^\star$ of size $\epsilon > 0$, $B(u^\star, \epsilon) \subset C$ such that

$$E(u^\star) \geq E(u) \quad \forall u \in B(u^\star, \epsilon).$$

The simplest goal of general optimization is to find a local minimum of a differentiable function. In general, the global structure of such a function is not simpler than that one of a Lipschitz continuous function.

**Definition** A function $f : C \subset \mathbb{R}^n \to \mathbb{R}^m$ is **Lipschitz continuous** with *Lipschitz constant* $L$ if for all $x, y \in C$

$$\|f(x) - f(y)\|_2 \leq L\|x - y\|_2$$

A function $f : C \subset \mathbb{R}^n \to \mathbb{R}^m$ is **locally Lipschitz continuous** if for every $x \in C$ there exists $\epsilon > 0$ such that $f_{|B(\epsilon,x)}$ is Lipschitz continuous

The majority of optimization methods are based on the idea of relaxation:

**Definition** We call the sequence $\{a^k\}_{k=0}^{\infty}$ a relaxation sequence if $a^{k+1} \leq a^k \quad \forall k \geq 0$.

For instance, to solve an unconstrained minimization problem of the form

$$\min_u E(u)$$

we construct a relaxation sequence $\{E(u^k)\}$ that satisfies $E(u^{k+1}) \leq E(u^k)$. Such a relaxation sequence always improves the initial value of the objective function. Moreover, if $E$ is bounded below on $\mathbb{R}^n$, then the sequence $\{E(u^k)\}$ converges. Formally:

**Definition** We say that a sequence $\{a^k\} \subset \mathbb{R}^n$ converges to $\hat{a} \in \mathbb{R}^n$ if for all $\epsilon > 0$ there exists an $k_0 \in \mathbb{N}$ such that

$$\|a^k - \hat{a}\| < \epsilon \quad \forall k \geq k_0.$$

To implement the idea of relaxation we use another fundamental principle of numerical analysis, the approximation. The approximation in this case replace an initial complex objective function $E$ by a simplified one, which is close by its properties to the original. When the function is differentiable, we usually apply local approximations based on derivatives of the objective function to create first- and second- order Taylor approximations that results in linear and quadratic approximations. Let $E(u)$ be differentiable at $u^0$, then for $u \in \mathbb{R}^n$, we have

$$E(u) = E(u^0) + \langle \nabla E(u^0), u - u^0 \rangle + o(\|u - u^0\|) \quad \text{where} \quad \lim_{r \to 0} \frac{o(r)}{r} = 0.$$

Function $E(u; u^0) = E(u^0) + \langle \nabla E(u^0), u - u^0 \rangle$ is a linear approximation of $E$ in a neighborhood of $u$.

Given an initial estimate of the minimizer $u^0$, we can then use this linear approximation to reduce the value of $E(u)$ in a neighborhood of $u^0$. In particular we can decide to iteratively step in the direction of maximum descent of the approximation $E(u; u^0)$, that is

$$\begin{aligned} u^1 &= u^0 - \tau\nabla E(u^0) \\ u^2 &= u^1 - \tau\nabla E(u^1) \\ &\dots \\ u^{k+1} &= u^k - \tau\nabla E(u^k). \end{aligned}$$

To ensure convergence, the step size $\tau > 0$ depends on the Lipschitz constant of $E$. This gives us a very simple algorithm know as gradient descent. We will see in this course that under certain conditions, the algorithm creates a relaxation sequence that decreases the value of the objective function $E(u^{k+1}) < E(u^k)$ and converges to a point $\hat{u} \in \mathbb{R}^n$. This a point then satisfies $\hat{u} = \hat{u} - \tau\nabla E(\hat{u}) \Rightarrow \nabla E(\hat{u}) = 0$. This is a necessary condition for optimality, as the next theorem shows.

**Theorem 1. First-order optimality condition.** *Let $u^\star$ be a local minimum of differentiable function $E(u)$. Then $\nabla E(u^\star) = 0$.*

*Proof.* Since $u^\star$ is a local minimum of $E(u)$, then there exists $r > 0$ such that for all $v$ with $\|v - u^\star\| \leq r$, we have $E(v) \geq E(u^\star)$. Since $E$ is differentiable, this implies that

$$E(v) = E(u^\star) + \langle \nabla E(u^\star), v - u^\star \rangle + o(\|v - u^\star\|) \geq E(u^\star).$$

Thus, for all $s$, $\|s\| = 1$, we have $\langle \nabla E(u^\star), s \rangle \geq 0$. Consider the directions $s$ and $-s$, we get $\nabla E(u^\star) = 0$. $\square$

Note that we have proved only a necessary condition of a local minimum. The points satisfying this condition are called the stationary points of function. In order to see that such points are not always the local minima, it is enough to look at function $E(u) = u^3$. The optimality condition $E'(u) = 3u^2 = 0$ results in $u^\star = 0$ even though the function is decreasing for any $u < 0$ and can thus not have a minimum at zero. The point $u^\star = 0$ is a stationary point, not a maximum or minimum.

To discern between local minima and stationary points of a function, let us introduce the second-order approximation. Let function $E(u)$ be twice differentiable with Hessian $\nabla^2 E(u)$ at $u$. Then

$$E(v) = E(u) + \langle \nabla E(u), v - u \rangle + \frac{1}{2}\langle \nabla^2 E(u)(v - u), v - u \rangle + o(\|v - u\|^2).$$

The function $E(v; u) = E(u) + \langle \nabla E(u), v - u \rangle + \frac{1}{2}\langle \nabla^2 E(u)(v - u), v - u \rangle$ is the quadratic (or second-order) approximation of function $E$ at $u$. Note that the Hessian is a symmetric matrix that can be seen as a derivative of the vector function $\nabla E$. As a result, using a linear approximation to each component of $\nabla E$, we have

$$\nabla E(v) = \nabla E(u) + \nabla^2 E(u)(v - u) + o(\|v - u\|).$$

Using the second-order approximation, we can write down the second- order optimality conditions.

**Theorem 2. Second-order optimality condition** *Let $u^\star$ be a local minimum of twice differentiahte function $E(u)$. Then $\nabla E(u^\star) = 0$ and $\nabla^2 E(u^\star)$ is symmetric and positive semi-definite, that we denote by $\nabla^2 E(u^\star) \succeq 0$.*

*Proof.* Since $u^\star$ is a local minimum of function $E$, there exists $r > 0$ such that

$$E(u) \geq E(u^\star) \quad \forall u \quad \text{with} \quad \|u - u^\star\| < r.$$

The first order optimality condition gives us $\nabla E(u^\star) = 0$ and, as a result

$$E(u) = E(u^\star) + \langle \nabla^2 E(u^\star)(v - u^\star), v - u^\star \rangle + o(\|y - u^\star\|^2) \geq E(u^\star).$$

Thus, $\langle \nabla^2 E(u^\star)(v - u^\star), v - u^\star \rangle \geq 0$. Letting $s = v - u$ we have $\langle \nabla^2 E(u^\star)s, s \rangle \geq 0$, which implies positive semi-definiteness. $\square$

This second-order characteristic of a local minimum is also sufficient.

**Theorem 3.** *Let function $E(u)$ be twice differentiable on $\mathbb{R}^n$ and let $u^\star$ satisfy $\nabla E(u^\star) = 0$ and $\nabla^2 E(u^\star) \succ 0$. Then $u^\star$ is a strict local minimum of $E$.*

*Proof.* In a small neighborhood of $u^\star$, $E(u)$ can be represented as

$$E(u) = E(u^\star) + \langle \nabla^2 E(u^\star)(u - u^\star), u - u^\star \rangle + o(\|u - u^\star\|^2).$$

Since $\lim_{r \to 0} \frac{o(r)}{r} = 0$, there exists a value $\bar{r}$ such that for all $r \in [0, \bar{r}]$ we have

$$|o(r)| \leq \frac{r}{4} \lambda_1,$$

where $\lambda_1 > 0$ is the smallest eigenvalue of matrix $\nabla^2 E(u^\star)$. As $\nabla^2 E(u^\star)$ is symmetric and positive definite, it has positive eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n > 0$ and orthogonal eigenvectors $q_1, q_2, \ldots q_n$, such that $\nabla^2 E(u^\star) = \sum_{1 \leq i \leq n} \lambda_i q_i^T q_i$ and $\|q_i^T v\| = \|v\|$ for all $v \in \mathbb{R}^n$. As a result,

$$E(u) \geq E(u^\star) + \frac{\lambda_1}{2} \|u - u^\star\|^2 + o(\|u - u^\star\|^2) \geq E(u^\star) + \frac{\lambda_1}{4} \|u - u^\star\|^2 \geq E(u^\star). \tag{1.2}$$

□

For general optimization problems, we thus require second-order differentiablity to formulate necessary and sufficient optimality conditions. The optima described by these conditions is, moreover, only local.

In most applications of computer vision and machine learning the objective functions are not differentiable and the general optimality conditions that we have derived are meaningless. Even in the rare cases where second-order derivatives exists, computing the Hessian is not feasible because the size of the problem is too large. For these reasons, we resort to the field of convex optimization. In convex optimization the objective function does not need to be differentiable, optimality conditions that do not assume differentiablity are not only necessary but sufficient, and the algorithms scale well with the size of the problem.

# Chapter 2

# Convex Analysis

## 2.1 Convex Optimization

We start this section with the unconstrained minimization problem

$$\min_{u \in \mathbb{R}^n} E(u). \tag{2.1}$$

In the general situation we cannot do too much: even when the function is smooth, the gradient method converges only to a stationary point of function $E$, second-order differentiablity is necessary to discern local minima from stationary points, and the Hessian matrix necessary to detect if a point is a local optimum is usually computationally too expensive to compute. To make the problem tractable we introduce some assumptions on function $E$.

As the main cause of our trouble is the weakness of the first-order optimality condition, we call for the following additional property: for any $E$ differentiable, the first-order optimality condition is necessary and sufficient for a point to be a global solution to the unconstrained minimization problem. This is what convex functions guarantee.

**Definition** A function $E \colon \mathbb{R}^n \to \mathbf{R}$ is convex if and only if for any $u, v \in \mathbb{R}^n$ and $\theta \in [0, 1]$

$$E(\theta u + (1-\theta)v) \leq \theta E(u) + (1-\theta)E(v).$$

$E$ is strictly convex if the inequality is strict for all $\theta \in (0, 1)$, $v \neq u$.

The definition of convex functions implicitly assumes that it is possible to evaluate the function at any point of the segment

$$[u, v] = \{z = \theta u + (1-\theta)v \colon 0 \leq \theta \leq 1\}.$$

As a result, it is natural to consider a set that contains the whole segment between any two points in the set. Such sets are called convex.

**Definition Convex Sets.** The set $C$ is convex if for any $u, v \in C$ and $\theta \in [0, 1]$, $\theta u + (1-\theta)v \in C$.

We can then include this notion in the definition of convex functions with restricted domain.

**Definition** The domain of a function $E \colon \mathbb{R}^n \to \mathbb{R}$ is the set

$$\mathrm{dom}(E) = \{u \in \mathbb{R}^n \colon E(u) < \infty\}$$

We can now extend the definition of convexity to functions.

**Definition Convex Function.** The function $E : \mathbb{R}^n \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ is convex if

- its domain $\mathrm{dom}(E)$ is a convex set.
- For all $u, v \in \mathrm{dom}(E)$ and all $\theta \in [0, 1]$ it holds that

$$E(\theta u + (1-\theta)v) \le \theta E(u) + (1-\theta)E(v).$$

$E$ is **strictly convex** if the inequality is strict for all $\theta \in (0, 1)$, $v \neq u$.

In the following we assume that the domain of $E$ is not empty, that is, the function $E$ is proper.

**Definition** Function $E : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper if its domain is not empty.

This course will investigate convex minimization problems, they are characterized by the form

$$\hat{u} \in \arg\min_{u \in C} E(u), \tag{2.2}$$

where $C$ is a convex set and $E$ is a convex function. To write such a problem in our familiar unconstrained optimization form, we we define the **extended real-valued function** $\tilde{E}$ by introducing the constraint $u \in C$ into the domain of the original energy function $E$:

$$\begin{aligned} \tilde{E} :& \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\} \\ u \mapsto& \tilde{E}(u) = \begin{cases} E(u) & \text{if } u \in C, \\ \infty & \text{else.} \end{cases} \end{aligned}$$

and re-write (2.2) as

$$\hat{u} \in \arg\min_{u \in \mathbb{R}^n} \tilde{E}(u).$$

## 2.2 Convex Sets

We have already seen some convex sets as a result of convex functions

**Lemma 4.** *If $E$ is a convex function, then for any $\beta \in \mathbb{R}$, its level set*

$$\{u \quad : \quad E(u) \le \beta\}$$

*is either convex or empty.*

*Proof.* Let $u, v \in \mathrm{dom}(E)$ with $E(u) \le \beta$ and $E(v) \le \beta$, by convexity of $E$ we have $\theta u + (1-\theta)v \in \mathrm{dom}(E)$ and

$$E(\theta u + (1-\theta)v) \le \theta E(u) + (1-\theta)E(v) \le \theta\beta + (1-\theta)\beta = \beta.$$

$\square$

**Lemma 5.** *Let $E$ be a convex function, then its epigraph*

$$epi(E) = \{(u, \beta) \quad : \quad E(u) \le \beta\}$$

*is a convex set.*

*Proof.* Let $(u, \alpha), (v, \beta) \in \mathrm{epi}(E)$, then $u, v \in \mathrm{dom}(E)$ with $E(u) \le \alpha$ and $E(v) \le \beta$, by convexity of $E$ we have $\theta u + (1-\theta)v \in \mathrm{dom}(E)$ and

$$E(\theta u + (1-\theta)v) \le \theta E(u) + (1-\theta)E(v) \le \theta\alpha + (1-\theta)\beta$$

and $\theta(u, \alpha) + (1-\theta)(v, \beta) \in \mathrm{epi}(E)$. $\square$

To determine if a set is convex, a few properties are useful.

**Lemma 6.** *Let $C \subset \mathbb{R}^n, D \subset \mathbb{R}^m$ be convex sets and $\mathscr{A} : \mathbb{R}^n \to \mathbb{R}^m$ be a linear operator, then the following sets are convex*

- $C \cap D$
- $C + D = \{u = x + y \ : \ x \in C, \ y \in D\}$
- *Affine image* $\mathscr{A}(C) = \{u \in \mathbb{R}^m \ : \ u = \mathscr{A}(x), \ x \in C\}$
- *Inverse affine image* $\mathscr{A}^{-1}(D) = \{v \in \mathbb{R}^n \ : \ \mathscr{A}(v) \in D\}$

*Proof.* Left as exercise □

We can now show that the following sets are convex

- Half-space $\{u \in\in \mathbb{R}^n \ : \ \langle a, u\rangle \leq \}$ is convex since linear functions are convex.
- Polytope $\{u \in\in \mathbb{R}^n \ : \ \langle a_i, u\rangle \leq b_i\}$ is convex as an intersection of convex sets.
- Ellipsoid $\{u \in\in \mathbb{R}^n \ : \ \langle Au, u\rangle \leq 1 \ \text{with} \ A \succeq 0\}$ because the function $\langle Au, u\rangle$ is a convex function.

## 2.3 Convex Functions

In order to determine if a function is convex, it is useful to know some equivalent definitions of convexity.

**Theorem 7. Convexity and Epigraphs.** *A proper function $E : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex if and only if its epigraph is convex*

*Proof.* We have already seen one direction, the other is an exercise □

**Lemma 8. Jensen's Inequality.** *For any convex function $E$, $u_1, \ldots, u_m \in dom(E)$ and coefficients $\theta_1, \ldots, \theta_m \geq 0$ such that $\sum_{i=1}^m \theta_i u_i = 1$ it holds*

$$E(\sum_{i=1}^m \theta_i u_i) \leq \sum_{i=1}^m \theta_i E(u_i)$$

.

*Proof.* By induction on $m$. The case $m = 2$ is a result of the definition. □

**Corollary 9.** *For any $u$ a convex combination of $u_1, \ldots, u_m \in dom(E)$, $E(u) \leq \max_{1 \leq i \leq m} E(u_i)$*

**Corollary 10.** *Let $\Delta = \mathrm{Conv}\{u_1, \ldots, u_m\}$ be the convex hull of $u_1, \ldots, u_m$, then*

$$\max_{u \in \Delta} E(u) = \max_{1 \leq i \leq m} E(u_i)$$

**Lemma 11.** *Function $E : C \to \mathbb{R}$ is convex if and only if $C$ is convex and for all $u, v \in C$, $\beta \geq 0$ such that $u + \beta(u - v) \in C$ it holds that*

$$E(u + \beta(u - v)) \geq E(u) + \beta(E(u) - E(v))$$

*Proof.* Let $E$ be convex, we first prove the alternative definition. Given $\beta > 0$ define $\theta = \frac{\beta}{\beta+1} \in (0,1]$ and $x = u + \beta(u-v)$ such that

$$u = \frac{1}{1+\beta}(x + \beta v = (1-\theta)x + \theta v$$

by convexity of $E$,

$$E(u) \leq (1-\theta)E(x) + \theta E(v) = \frac{1}{1+\beta}E(u + \beta(u-v)) + \frac{\beta}{1+\beta}E(v)$$

$$(1+\beta)E(u) - \beta E(v) \leq E(u + \beta(u-v))$$

Let us now prove that this alternative definition implies convexity. Given any $u, v \in \mathrm{dom}E$, $\theta \in [0,1]$, define $\beta = \frac{1-\theta}{\theta}$ and $x = \theta u + (1-\theta)v$ such that

$$u = \frac{1}{\theta}(x - (1-\theta)v) = x + \beta(x-v)$$

the inequality reads

$$\begin{aligned}
&E(u) = E(x + \beta(x-v)) \geq E(x) + \beta[E(x) - E(v)] \\
&E(u) \geq (1+\beta)E(x) - \beta E(v) = \frac{1}{\theta}E(x) - \frac{1-\theta}{\theta}E(v) \\
&\theta E(u) + (1-\theta)E(v) \geq E(\theta u + (1-\theta)v)
\end{aligned}$$

□

**Theorem 12. Monotonicity of the gradient** *Let $E : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, and continuously differentiable at $u \in dom(E)$. Then $E$ is convex if and only if for any $u, v \in dom(E)$*

$$E(v) \geq E(u) + \langle \nabla E(u), v - u \rangle.$$

*Proof.* Given $u, x \in \mathrm{dom}(E)$, and $\theta \in [0,1]$, let $u_\theta = \theta u + (1-\theta)v$. If $E$ is continuously differentiable and satisfies the theorem's inequality, we have

$$\begin{aligned}
E(u_\theta) &\geq E(v) + \langle \nabla E(u_\theta), v - u_\theta \rangle = E(v) + \theta \langle \nabla E(u_\theta), v - u \rangle \\
E(u_\theta) &\geq E(u) + \langle \nabla E(u_\theta), u - u_\theta \rangle = E(u) - (1-\theta)\langle \nabla E(u_\theta), v - u \rangle.
\end{aligned}$$

Multiplying the first inequality by $1-\theta$, the second by $\theta$, and adding the results, we get the inequality that defines a convex function $E(\theta u + (1-\theta)v) \leq \theta E(u) + (1-\theta)E(v)$.
We now prove that a convex and continuously differentiable function satisfies the theorem's inequality. Given $u, v \in \mathrm{dom}(E)$, as $E$ is convex for any $\theta in[0,1]$

$$E(v) \geq \frac{1}{1-\theta}[E(u_\theta) - \theta E(u)] = E(u) + \frac{1}{1-\theta}[E(u_\theta) - E(u)] = E(u) + \frac{1}{1-\theta}[E(\theta u + (1-\theta)v) - \theta E(u)]1. \tag{2.3}$$

As $E$ is differentiable, the limit when $\theta$ tends to 1 exists and we get $E(v) \geq E(u) + \langle \nabla E(u), v - u \rangle$. □

### 2.3.1 Necessary and Sufficient Optimality Conditions

**Theorem 13.** *Let $E : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex. Any local minimum of $E$ is global.*

*Proof.* Let $u^\star$ be a global minimum of $E$ and $\bar{u}$ a local minimum that is not global, that is, $E(u^\star) < E(\bar{u})$. By definition of local minimum, there exists an $\epsilon$-ball centered at $\bar{u}$, $B(\bar{u}, \epsilon)$, such that $E(v) \geq E(\bar{u})$ for any $v \in B(\bar{u}, \epsilon)$. As $u^\star, \bar{u} \in \mathrm{dom}(E)$ convex, $\theta\bar{u} + (1-\theta)u^\star \in \mathrm{dom}(E)$ and

$$E(\theta\bar{u} + (1-\theta)u^\star) \leq \theta E(u^\star) + (1-\theta)E(\bar{u}) < E(\bar{u})$$

As $\theta$ tends to 1, $\theta\bar{u} + (1-\theta)u^\star \in B(\bar{u}, \epsilon)$ and this contradicts the definition of $\bar{u}$ as local minimum. □

When the function is differentiable, we can now prove that first-order optimality conditions are sufficient.

**Theorem 14.** *If $E\colon \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable function with $\nabla E(u^\star) = 0$ then $u^\star$ is the global minimum of $E(x)$.*

*Proof.* As $\nabla E(u^\star) = 0$, the inequality

$$E(v) \geq E(u^\star) + \langle \nabla E(u^\star), v - u^\star \rangle \quad \forall v \in \mathrm{dom}(E)$$

gives us the condition $E(v) \geq E(u^\star)$ that characterizes a global minimum. □

When the objective function is two-times differentiable, we can also characterize convexity in terms of the Hessian.

**Theorem 15.** *Two times continuously differentiable function $E\colon \mathbb{R}^n \to \mathbb{R}$ is convex if and only for any $u \in \mathbb{R}^n$ we have $\nabla^2 E(u) \succeq 0$.*

*Proof.* This is part of the exercise sheet. □

### 2.3.2 Analytic Properties of Convex Functions

The behavior of convex functions at the boundary of their domain can be out of control. To prevent this case, we ask the functions to be closed.

**Definition Closed convex function.** A convex function is closed if its epigraph is closed.

**Lemma 16.** *Property If $E$ is convex and closed, all its level sets are closed.*

*Proof.* For each $\beta$, the level-set $\{u \;:\; E(u) = \beta\} = \mathrm{epi}(E) \cap \{(x,t) \;:\; t = \beta\}$ can be described as the intersection of the epigraph of $E$, which is closed and convex, and the closed and convex set $\{(x,t) \;:\; t = \beta\}$. □

If $E$ is convex and continuous and its domain $\mathrm{dom}(E)$ is closed, then $E$ is a closed function. The converse is not true, a closed convex function is not necessarily continuous. Consider the following examples

- $E(u) = \frac{1}{u}$ is convex, has an open domain $\mathrm{dom}(E) = \mathbb{R}_{++} = \{u \in \mathbb{R} \;: u > 0\}$, but is closed because its epigraph is $\{(u,t) \in \mathbb{R} \times \mathbb{R}_{++} \;:\; \frac{1}{t} \leq u\}$
- Function $E(u) = \|u\|$, where $\|\cdot\|$ is any norm, is closed and convex as a result of the triangle inequality and homogeneity properties that define any norm:

$$\|\theta u + (1-\theta)v\| \leq \|\theta u\| + \|(1-\theta)\|v\| = |\theta|\|u\| + |1-\theta|\|\|v\| = \theta\|u\| + (1-\theta)\|\|v\|$$

The norms more common in computer vision and machine learning are the $\ell_p$ norms

$$\|u\| = \left(\sum_{i=1}^{n} |u_i|^p\right)^{\frac{1}{p}} \quad u \in \mathbb{R}^n$$

  - When we omit the subscript, we describe the Euclidean norm: $|u| = \sqrt{\sum_{i=1}^n u_i^2}$
  - the non-differentiable $\ell_1$ norm $\|u\|_1 = \sum_{i=1}^n |u_i|$
  - the $\ell_\infty$ norm $\|u\|_\infty = \max_{1\le i\le n} \ |u_i|$

- the function

$$E(x,y) = \begin{cases} 0 & \text{if } \ x^2+y^2<1 \\ \phi(x,y) & \text{if } \ x^2+y^2=1 \end{cases}$$

  with domain dom$E = B_2(0,1)$ is closed and convex for any $\phi(x,y)>0$ defined on the unit circle, the boundary of the function domain. Imposing that the function is closed, which results in $\phi(x,y)=0$, ensures that the function is well-behaved also on the boundary of its domain.

The behavior of convex function at the boundary of their domain can be disappointing, but their behavior in the interior of its domain is very simple.

**Theorem 17.** *Let function $E\colon C \subset \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex and $u \in \operatorname{int}$ dom$(E)$ then $E$ is locally bounded at $u$.*

*Proof.* Let us choose $\epsilon > 0$ such that $u \pm \epsilon e_i \in \operatorname{int}\operatorname{dom}(E) \ \ i = 1,\dots,n$, where $e_i$ is the $i$-th coordinate vector of $\mathbb{R}^n$ and define $\hat{\epsilon} = \frac{\epsilon}{\sqrt{n}}$. We have

$$B(u,\hat{\epsilon}) \subset \Delta = \operatorname{Conv}\{u \pm \epsilon e_i \ \ i=1,\dots,n\}.$$

From the corollary to Jensen's inequality, we find a local bound $M$ to $E$

$$M = \max_{v\in B(u,\hat{\epsilon})} \ E(v) \le \max_{v\in\Delta} \ E(v) \le \max_{1\le i\le n} \ E(u\pm\epsilon e_i)$$

□

**Theorem 18. Continuity of Convex Functions** *If $E: \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is convex, then $E$ is locally Lipschitz (and hence continuous) on int(dom$(E)$).*

*Proof.* Let $B(u_0,\epsilon) \subset \operatorname{dom}(E)$ and $M = \sup_{u\in B(u_0,\epsilon)} \ E(u) < \infty$.
Consider $v \in B(u_0,\epsilon)$, $v \ne u_0$ and define

$$\alpha = \frac{1}{\epsilon}\|v-u_0\| z = u_0 + \frac{1}{\alpha}(v-u_0)$$

It is clear that $\|z-u_0\| = \epsilon$, $\alpha \le 1$, and $v = \alpha z + (1-\alpha)u_0$. By convexity of $E$ then

$$E(v) \le \alpha E(z) + (1-\alpha)E(u_0) \le E(u_0) + \alpha(M - E(u_0)) = E(u_0) + \frac{M-E(u_0)}{\epsilon}\|v-u_0\|$$

Now define $y = u_0 + \frac{1}{\alpha}(u_0 - v)$ with $\|y-u_0\| = \epsilon$ and $v = u_0 + \alpha(u_0 - y)$. We have

$$E(v) \ge E(u_0) + \alpha(E(u_0) - E(y)) \ge E(u_0) - \alpha(M - E(u_0)) = E(u_0) - \frac{M-E(u_0)}{\epsilon}\|v-u_0\|$$

As a result of the 2 inequalities

$$|E(v) - E(u)| \le \frac{M-E(u_0)}{\epsilon}\|v-u_0\|$$

□

### 2.3.3 Examples of Convex Functions

The next statements significantly increases our possibilities of constructing convex functions.

**Lemma 19.** *Given a closed convex function $\phi$ and a linear operator $\mathscr{A} : \mathbb{R}^m \to \mathbb{R}^n$, then $E(u) = \phi(\mathscr{A}(u))$ is closed and convex with*

$$dom(E) = \{u \in \mathbb{R}^m \quad : \mathscr{A}(u) \in dom(\phi)\}.$$

*Proof.* Let $\mathscr{A}(u) = Au + b = x \in \mathrm{dom}(\phi)$ and $\mathscr{A}(v) = Av + b = y \in \mathrm{dom}(\phi)$, then by convexity of $\phi$ for any $\theta \in [0, 1]$ we have $\theta x + (1 - \theta)y \in \mathrm{dom}(\phi)$ and

$$E[\theta u+(1-\theta)v] = \phi[\mathscr{A}(\theta u+(1-\theta)v)] = \phi[\theta(Au+b)+(1-\theta)(Av+b)] \leq \theta\phi(Au+b)+(1-\theta)\phi(Av+b) = \theta E(u)+(1-\theta)E(v).$$

This proves convexity of $E$. The closedness of its epigraph follows from continuity of the linear operator $\mathscr{A}$. □

**Lemma 20.** *Given two closed and convex function $E_1, E_2$ and $\alpha_1, \alpha_2 > 0$, then $E = \alpha_1 E_1 + \alpha_2 E_2$ is closed and convex with $dom(E) = dom(E_1) \cap dom E_2$.*

*Proof.* Let $u, v \in \mathrm{dom}(E_1) \cap \mathrm{dom}E_2$ and $\theta \in [0, 1]$, by convexity of each $E_1, E_2$ we have

$$\begin{aligned}\alpha_1 E_1(\theta u + (1-\theta)v) + \alpha_2 E_2(\theta u + (1-\theta)v) &\leq \alpha_1\theta E_1(u) + \alpha_1(1-\theta)E(v) + \alpha_2\theta E_1(u) + \alpha_2(1-\theta)E(v) \\ &= \theta[\alpha_1 E_1(u) + \alpha_2 E_2(u)] + (1-\theta)[\alpha_1 E_1(v) + \alpha_2 E_2(v).\end{aligned} \tag{2.4}$$

This proves the convexity of $E$, to prove that it is closed we consider a sequence $\{(u_k, t_k)\} \in \mathrm{epi}(E)$ that satisfies

$$\alpha_1 E_1(u_k) + \alpha_2 E_2(u_k) \leq t_k \qquad \lim_{k\to\infty} u_k = \bar{u} \in \mathrm{dom}(E) \qquad \lim_{k\to\infty} t_k = \bar{t} \tag{2.5}$$

Since $E_1, E_2$ are closed, they are lower semi-continuous, and

$$\inf \lim_{k\to\infty} E_1(u_k) \geq E_1(\bar{u}) \qquad \inf \lim_{k\to\infty} E_2(u_k) \geq E_2(\bar{u}) \tag{2.6}$$

and

$$\bar{t} = \lim_k t_k \geq \inf \lim_{k\to\infty} \alpha_1 E_1(u_k) + \inf \lim_{k\to\infty} \alpha_2 E_2(u_k) \geq \alpha_1 E_1(\bar{u}) + \alpha_2 E_2(\bar{u}) \tag{2.7}$$

and $(\bar{u}, \bar{t}) \in \mathrm{epi}(E)$. □

**Lemma 21.** *Given two closed and convex function $E_1, E_2$, then $E(u) = \max\{E_1(u), E_1(u)\}$ is closed and convex with $dom(E) = dom(E_1) \cup dom E_2$.*

*Proof.* The epigraph

$$\mathrm{epi}(E) = \{(u, t) \quad : \quad u \in \mathrm{dom}(E_1) \cap \mathrm{dom}(E_2),\ E_1(u) \leq t,\ E_2(u) \leq t\} = \mathrm{epi}(E_1) \cap \mathrm{epi}(E_2)$$

is the intersection of two closed convex sets and it is thus closed and convex. □

**Theorem 22.** *Let $D$ be some set, not necessarily convex or finite dimensional, and $E(u) = \sup_{y\in C} \phi(u, y)$ such that for each $y \in D$, $\phi(u, y)$ is closed and convex in $u$. Then $E$ is a closed and convex function with domain*

$$dom(E) = \{u \in \cap_{y\in D} dom(\phi(\cdot, y)) \quad : \quad \exists\gamma \in \mathbb{R} \quad s.t. \quad \phi(u, y) \leq \gamma \quad \forall y \in D\}$$

*Proof.* If $u$ belongs to $\{u \in \cap_{y\in D}\mathrm{dom}(\phi(\cdot,y)) \ : \ \exists\gamma \in \mathbb{R} \text{ s.t. } \phi(u,y) \leq \gamma \ \ \forall y \in D\}$, then $E(u) < \infty$ and $u \in \mathrm{dom}(E)$. If $u$ does not belong to this set, then there exists a sequence $\{y_k\}$ such that $\phi(u,y_k) \to \infty$ and $u$ does not belong to $\mathrm{dom}(E)$.
$(u,t) \in \mathrm{epi}(E)$ if and only if for all $y \in D$, we have $u \in \mathrm{dom}(\phi(\cdot,y))$ and $\phi(u,y) \leq t$. As a results $\mathrm{epi}(E) = \cap_{y\in D}\mathrm{epi}(\phi(\cdot,y))$ is closed and convex as the intersection of closed and convex sets. □

We can now easily show that the next functions are convex:

- Linear function $E(u) = \alpha + \langle a, u\rangle$ is convex.
- For any matrix $A$ be symmetric and positive semi-definite, the quadratic function $E)u) = \alpha + \langle a,u\rangle + \langle a, Au\rangle$ is convex because $\nabla^2 E(u) = A \succeq 0$.
- The following 1-dimensional are convex

$$\begin{aligned} E(u) &= \exp(u) \\ E(u) &= |u|^p \quad p > 1 \\ E(u) &= |x| - \log(1 + |x|) \end{aligned}$$

- As a consequence, the following multi-dimensional functions are convex

$$\begin{aligned} E(u) &= \sum_{i=1}^{n} \exp(\alpha + \langle u, a_i\rangle) \\ E(u) &= |\langle u, a_i\rangle - b_i|^p \quad p > 1 \end{aligned}$$

- The function $E^*(y) = \sup_{u\in\mathrm{dom}(E)} \langle u,y\rangle - E(u)$ is convex.

## 2.4 Existence and Uniqueness of Minimizers

It only makes sense to try to solve an optimization problem if it has a solution. Specially, if the solution is the limit of a relaxation sequence that we need to compute with a costly iterative algorithm that might never converge. To show that a convex problem has a minimizer, we will see that it satisfies the necessary conditions to frame the problem in a more the general theoretical setting of lower semi-continuity. This section explains the tools from this general setting that we will need.

**Definition Lower semi-continuity.** A function $E : \mathbb{R}^n \to \mathbb{R}$ is lower semi-continuous (l.s.c.), if for all $u$ it holds that

$$\liminf_{v\to u} E(v) \geq E(u).$$

**Theorem 23.** *Let $E : \mathbb{R}^n \to \overline{\mathbb{R}}$ be l.s.c. and let there exist an $\alpha$ such that the sublevelset*

$$\{u \in \mathbb{R}^n \mid E(u) \leq \alpha\}$$

*is nonempty and bounded, then there exists*

$$\hat{u} \in \arg\min_u E(u).$$

*Proof.* Remember that the infimum is the largest lower bound on all possible values of $E(u)$ and consider a sequence $(u_k)_k$ such that $E(u_k) \to \inf_u E(u)$.
We distinguish two cases: For $\alpha = \inf_u E(u)$ the non-emptyness of $S_\alpha$ yields the assertion. For $\alpha >$

$\inf_u E(u)$ it holds that from some sufficiently large $k_0$ on, we will have $u_k \in S_\alpha$. Since $S_\alpha$ is bounded there exists a convergent subsequence $u_{k_l} \to \bar{u}$. Due to the lower semi-continuity we find

$$\inf_u E(u) = \lim_{k\to\infty} E(u_k) = \lim_{l\to\infty} E(u_{k_l}) \geq E(\bar{u}).$$

Since by definition $\inf_u E(u) \leq E(\bar{u})$ we obtain equality and hence there exists $\bar{u} \in \operatorname{argmin}_u E(u)$. ☐

**Theorem 24. Equivalence of l.s.c. and closedness.** *For $E : \mathbb{R}^n \to \overline{\mathbb{R}}$ the following two statements are equivalent*

- *$E$ is lower semi-continuous (l.s.c.)*
- *$E$ is closed (its epigraph is closed)*

*Proof.* Let $E$ be closed and assume that $E$ is not l.s.c.Then there exists a point $u^0$ and a sequence $(u_k)_k$ with $\lim_k u_k = u^0$ such that

$$\liminf_k E(u_k) < E(u^0).$$

In particular, there exists $\alpha \in \mathbb{R}$ and a subsequence $(u_{k_l})_{k_l}$ such that

$$E(u_{k_l}) \leq \alpha < E(u^0) \quad \forall k \tag{2.8}$$

Obviously, $(u_{k_l}, \alpha) \in \operatorname{epi}(E)$ for all $k_l$ and $(u_{k_l}, \alpha) \to (u^0, \alpha)$, but according to (**??**) $(u^0, \alpha) \notin \operatorname{epi}(E)$, which contradicts the closedness of $E$.
To prove the other direction of the claim, let $E$ be l.s.c. and assume that $E$ is not closed. Then there exists a sequence $(u_k, \alpha_k) \in \operatorname{epi}(E)$ with $(u_k, \alpha_k) \to (u^0, \alpha^0) \notin \operatorname{epi}(E)$. We find

$$\liminf_k E(u_k) \leq \lim_k \alpha_k = \alpha^0 < E(u^0).$$

On the other hand, due to $E$ being l.s.c. we have $E(u^0) \leq \liminf_k E(u_k)$, which is a contradiction. ☐

### 2.4.1 Existence of Minimizers of Convex Functions

**Definition Coercivity.** A function $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is called coercive if $E(v_n) \to \infty$ for all sequences $(v_n)_n$ with $\|v_n\| \to \infty$.

It is easy to proof by contradiction, that coercivity implies existence of a bounded sublevelset. We have now all the tools to prove existence of minimizers of convex functions.

**Theorem 25. Existence of a Minimizer** *Let $E : \mathbb{R}^n \to \mathbb{R}$ be convex and coercive, then an element $\hat{u} \in \arg\min_u E(u)$ exists.*

*Proof.* As $\operatorname{dom}(E) = \mathbb{R}^n$ and $E$ convex, $E$ is continuous. Similarly, as $E$ is coercive, there exists a non-empty bounded sublevelset, and we can apply the previous theorem on the existence of minimizers for general lower semi-continuous functions. ☐

**Theorem 26. Uniqueness.** *If $E : \mathbb{R}^n \to \overline{\mathbb{R}}$ is strictly convex, then there exists at most one local minimum which is the unique global minimum.*

*Proof.* Assume there are 2 global minima $u, v$ with $u \neq v$, $E(u) = E(v)$, then any $\theta \in [0, 1]$ we have

$$E(\theta u + (1-\theta)v) < \theta E(u) + (1-\theta)E(v),$$

which contradicts the definition of $u$, $v$ as global minima. ☐

## 2.5 Subdifferentials

### 2.5.1 Supporting Hyperplanes

Up to now we were describing properties of convex functions in terms of function values or their gradients. When the function is not differentiable, we need to define a direction that acts as the gradient of differentiable functions that points onto the direction of maximum ascent. In convex analysis such directions are defined by supporting hyperplanes.

**Definition** Let $C$ be a convex set. We say that hyperplane

$$\mathscr{H}(g,\gamma) = \{u \in \mathbb{R}^n \quad : \quad \langle g, u\rangle = \gamma, \quad g \neq 0\}$$

is supporting to $C$ if any $u \in C$ satisfies $\langle g, u\rangle \leq \gamma$.
We say that the hyperplane $\mathscr{H}(g,\gamma)$ separates a point $u_0$ from $C$ if

$$\langle g, u\rangle \leq \gamma \leq \langle g, u_0\rangle \quad \forall u \in C.$$

Now we can enunciate two separation theorems necessary to define gradient-like directions of maximum ascent or descent for non-differentiable functions.

**Theorem 27. Separating Hyperplane Theorem** *Let $C$ be a closed convex set and $u_0 \notin C$. Then there exists a hyperplane $\mathscr{H}(g,\gamma)$ that strictly separates $u_0$ from $C$.*

*Proof.* See Boyd and Vandenberghe, *Convex Optimization Theory*, pp 46–49. □

The next separation theorem deals with boundary points of convex sets.

**Theorem 28. Supporting Hyperplane Theorem** *Let $C$ be a closed convex set and $u_0$ in the boundary of $C$. Then there exists a hyperplane $\mathscr{H}(g,\gamma)$ supporting to $C$ and passing through $u_0$.*

*Proof.* See Boyd and Vandenberghe, *Convex Optimization Theory*, pp 50–51. □

### 2.5.2 The Subdifferential

We now have all the tools to introduce the notion of subdifferential that extends the gradient to non-differentiable functions.

**Definition Subdifferential.** Let $E : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex, the subdifferential of $E$ at $u$ is

$$\partial E(u) = \{p \in \mathbb{R}^n \mid E(v) - E(u) - \langle p, v-u\rangle \geq 0, \ \forall v \in \mathbb{R}^n\}$$

- Elements of $\partial E(u)$ are called subgradients.
- If $\partial E(u) \neq \emptyset$, we call $E$ subdifferentiable at $u$.
- By convention, $\partial E(u) = \emptyset$ for $u \notin \mathrm{dom}(E)$.

The subdifferential $\partial E$ is necessary because subgradients are not unique. Consider for example a function as friendly-looking as the absolute value at zero:

$$\forall g \in [-1,1], \quad E(u) = |u| \geq gu = E(0) + \langle g, u-0\rangle$$

As a result, the subdifferential at 0 is the interval $\partial E(0) = [-1,1]$. In general $\partial E(u)$ is a set. Form its definition as a set of linear constraints, we can easily see that it is closed and convex, in this case the interval $[-1,1]$.

### 2.5.3 Subdifferentiablity and Convexity

The subdifferentiability of a function is important because it implies its convexity.

**Theorem 29.** *If for any $u \in dom(E)$ the subdifferential $\partial E(u)$ is non-empty, then $E$ is a convex function.*

*Proof.* Given $u, v \in \mathrm{dom}(E)$, and $\theta \in [0,1]$, let $u_\theta = \theta u + (1-\theta)v$. As the subdifferential $\partial E(u_\theta$ is non-empty, we can pick $g \in \partial E(u_\theta$ satisfying

$$\begin{aligned} E(u_\theta) &\geq E(v) + \langle g, v - u_\theta \rangle = E(v) + \theta \langle g, v - u \rangle \\ E(u_\theta) &\geq E(u) + \langle g, u - u_\theta \rangle = E(u) - (1-\theta) \langle g, v - u \rangle. \end{aligned}$$

Multiplying the first inequality by $1-\theta$, the second by $\theta$, and adding the results, we get the inequality that defines a convex function $E(\theta u + (1-\theta)v) \leq \theta E(u) + (1-\theta)E(v)$. $\square$

The converse statement is also true.

**Theorem 30.** *If $E$ is a closed convex function and $u \in int(dom(E))$, then $\partial E(u)$ is a non-empty bounded set.*

*Proof.* Note that the point $(E(u), u)$ belongs to the boundary of $\mathrm{epi}(E)$, which is convex. As a result, there exists a hyperplane $\mathscr{H} = (g, \gamma)$ supporting to $\mathrm{epi}(E)$ at $(E(u), u)$:

$$\gamma\tau + \langle g, u \rangle \leq \gamma E(u) + \langle g, u \rangle \quad \forall (u, \tau) \in \mathrm{epi}(E)$$

Without loss of generality, we can assume $\|g\|^2 + \gamma^2 = 1$. We can determine the sign of $\gamma$ by checking the inequality for any point in the epigraph. In particular for any $\tau \geq E(u)$, we have $(u, \tau) \in \mathrm{epi}(E)$ that results in $\gamma > 0$.
To find a subgradient $p \in \partial E(u)$, we will use that a convex function is locally upper bounded in the interior of its domain. That is, there is some $\epsilon > 0, M > 0$ such that $B(u, \epsilon) \subset \mathrm{dom}(E)$ and

$$E(v) - E(u) \leq M\|v - u\| \quad \forall v \in B(u, \epsilon)$$

For any $v$ from this ball, the supporting hyperplane equation reads

$$\langle g, v - u \rangle \leq \gamma(E(v) - E(u)) \leq \gamma M \|v - u\|$$

In particular, if we choose $v = u + \epsilon g$ we get $\|g\|^2 \leq M\gamma\|d\|$. Plugging now the condition $\|g\|^2 + \gamma^2 = 1$ we get

$$\gamma \geq \frac{1}{\sqrt{1 + M^2}}.$$

If we choose $p = \frac{g}{\gamma}$ we obtain

$$E(v) \geq E(u) + \langle p, v - u \rangle \quad \forall v \in \mathrm{dom}(E)$$

and $p$ is a subgradient of $E$ at $u$. Finally, to show that the subdifferential is bounded we assume that $p \neq 0$ and consider the point $v = u + \epsilon \frac{p}{\|p\|}$ such that

$$\epsilon\|p\| = \langle p, v - u \rangle \leq E(v) - E(u) \leq M\|v - u\| = M\epsilon$$

Thus, $\partial E(u)$ is bounded by $M$. $\square$

The conditions of this theorem cannot be relaxed. For instance, the function $E(u) = -\sqrt{u}$ is convex and closed in its domain $\{u \colon u \geq 0\}$, but its subdifferential does not exists at the only point $(0)$ that is not in its interior. This is just another reminder that considering the interior of the domain for convex functions is important.

To conclude this section, let us point out to the property of the subgradients that makes it important for optimization.

**Theorem 31. Optimality Condition.** *Let* $0 \in \partial E(\hat{u})$, *then* $\hat{u} \in \arg\min_{u \in \mathbb{R}^n} E(u)$

*Proof.* If $0 \in \partial E(\hat{u})$, by definition of the subgradient

$$E(u) \geq E(\hat{u}) + \langle 0, u - \hat{u} \rangle = E(\hat{u}) \quad \forall u \in \mathrm{dom}(E)$$

and we conclude that $\hat{u}$ is a minimizer of $E$. On the other hand, if $E(u) \geq E(\hat{u})$ for all $u \in \mathrm{dom}(E)$, then 0 satisfies the condition of subgradient of $E$ at $\hat{u}$. □

### 2.5.4 Alternative Definitions of Subgradients

The supporting hyperplane theorem appears on the proof of the "subdifferentiability" theorem because subgradients can be interpreted in terms of supporting hyperplanes.

**Theorem 32. Geometric interpretation of Subgradients.** *Any subgradient* $p \in \partial E(u)$ *represents a non-vertical supporting hyperplane to epi*$(E)$ *at* $(u, E(u))$

*Proof.* Let $p \in \partial E(u)$. Then, by definition of subgradient,

$$\begin{aligned}
&E(v) - E(u) - \langle p, v - u \rangle \geq 0 && \forall v \in \mathbb{R}^n \\
&\alpha - E(u) - \langle p, v - u \rangle \geq 0 && \forall (v, \alpha) \in \mathrm{epi}(E) \\
&\left\langle \begin{bmatrix} -p \\ 1 \end{bmatrix}, \begin{bmatrix} v \\ \alpha \end{bmatrix} - \begin{bmatrix} u \\ E(u) \end{bmatrix} \right\rangle \geq 0 && \forall (v, \alpha) \in \mathrm{epi}(E).
\end{aligned}$$

As a result, the non-vertical hyperplane $\mathscr{H} = (g, \gamma)$ with $g = (-p, 1)$ and $\gamma = \langle p, u \rangle - E(u)$ supports epi$(E)$ at $(u, E(u))$. □

Apart from this geometric interpretation, it is useful to compute the subdifferential of a differentiable function to show how it is a generalization of the gradient. The next theorem does that.

**Theorem 33. Subdifferential of of Differentiable Functions.** *Let the convex function* $E : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *be differentiable at* $u \in$ *int*$($*dom*$(E))$. *Then*

$$\partial E(u) = \{\nabla E(u)\}.$$

*Proof.* The subdifferential $\partial E(u)$ of some convex $E$ at $u \in \mathrm{dom}(f)$ is given as

$$\{p \in \mathbb{R}^n : E(z) - E(u) - \langle p, z - u \rangle \geq 0, \forall\, z \in \mathrm{dom}(f)\}.$$

Since $u \in \mathrm{int}(\mathrm{dom}(E))$, we find that for all $v \in \mathbb{R}^n$, $z = u \pm \epsilon v \in \mathrm{dom}(E)$ for $\epsilon$ small enough. Therefore, it holds that

$$E(u + \epsilon v) \geq E(u) + \epsilon \langle p, v \rangle, \quad E(u - \epsilon v) \geq E(u) - \epsilon \langle p, v \rangle,$$

for all $v \in \mathbb{R}^n$ and $\epsilon$ small enough. This implies that

$$\lim_{\epsilon \to 0} \frac{E(u + \epsilon v) - E(u)}{\epsilon} \geq \langle p, v \rangle, \quad \lim_{\epsilon \to 0} \frac{E(u) - E(u - \epsilon v)}{\epsilon} \leq \langle p, v \rangle,$$

which means

$$\langle \nabla E(u), v \rangle \geq \langle p, v \rangle, \quad \langle \nabla E(u), v \rangle \leq \langle p, v \rangle,$$

i.e.

$$\langle \nabla E(u) - p, v \rangle = 0$$

for all $v \in \mathbb{R}^n$. For the particular choice of $v := \nabla E(u) - p$ we find $p = \nabla f(u)$. The above concludes the proof if we can show that $\partial f(u)$ is non-empty, which follows from the Theorem on Subdifferentiability. □

### 2.5.5 Subdifferential Rules

In the same way that the gradient of a differentiable function is only defined for points in the interior of the domain, the subdifferential of a proper convex function is always defined for points in the relative interior of its domain.

The relative interior of a set is a refinement of the concept of the interior that is useful when dealing with low-dimensional sets embedded in higher-dimensional spaces. Intuitively, the relative interior of a set contains all points that are not on the "edge" of the set, relative to the smallest subspace in which this set lies. When the set is convex, the definition takes the a simple form:

**Definition Relative Interior of Convex Sets** The relative interior of a convex set $C$ is defined as

$$\mathrm{ri}(C) := \{x \in C \mid \forall y \in C,\ \exists \lambda > 1,\ \text{s.t. } \lambda x + (1-\lambda)y \in C\}$$

As mentioned earlier, the subdifferentiability of convex functions can be guaranteed for points that are not necessarily in the interior of the domain, but that are in its relative interior. To better understand this difference, consider the line segment $I = [-1, 1]$ as a convex subset of the Euclidean plane $I \subset \mathbb{R}^2$. The interior of $I$ is empty with the Euclidean topology of $\mathbb{R}^2$, but its relative interior is the open line segment $\mathrm{ri}(I) = (0, 1)$.

One key property of the relative interior is that it is not empty for convex sets.

**Theorem 34.** *Let $C$ be a non-empty convex set, then $ri(C)$ is not empty.*

Now that we understand where subdifferentials exists, we can learn the rules that guide their computation.

**Theorem 35. Sum Rule.** *Let $E_1$, $E_2$ be convex functions, then $\partial(E_1 + E_2)(u) = \partial E_1(u) + \partial E_2(u)$ for all $u \in ri(dom(E_1)) \cap ri(dom(E_2))$.*

*Proof.* See Nesterov, *Introductory Lectures on Convex Optimization*, Lemma. 3.1.9. □

**Theorem 36. Chain Rule** *Given the linear operator $A \in \mathbb{R}^{m\times n}$ and the convex function $E : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, then $\partial(E \circ A)(u) = A^* \partial E(Au)$ for all $u \in ri(dom(E)) \cap range(A)$.*

*Proof.* See Nesterov, *Introductory Lectures on Convex Optimization*, Nesterov, Lemma. 3.1.8. □