Computational Optimal Transport for Machine and Deep Learning

Corentin Moumard
Issu du cours de *Mathurin Massias*, *Titouan Vayer* et *Quentin Bertrand*,
et rédigé pour ce même cours

Contents

Chapter 1

The Monge-Kantorovich Problem

1.1 The Historical Problem of Monge

In the *Mémoire sur la théorie des déblais et des remblais* from 1781, Gaspard Monge proposes a fundamental problem in optimal transport, which involves moving source points x_i to target points y_j while minimising the total cost (see Figure 1.1). For each source-target pair, a cost $C_{i,j}$ is associated, based on the Euclidean distance. The objective is to find the most economical matching between the sources and the targets. This is formalised by searching for the permutation σ that minimises the expression $\sum_i C_{i,\sigma(i)}$, that is:

$$\min_{\sigma \in \Sigma_n} \sum_{i=1}^{n} C_{i,\sigma(i)}$$

where Σ_n represents the set of all $n!$ possible permutations of the indices $\{1, 2, \ldots, n\}$.

The challenge is to determine this optimal permutation among the $n!$ possibilities, thus ensuring that the total cost is minimised. In this problem, each unit of mass from a source point must be fully transferred to a single target point.

Monge's Problem

Data:
- Source points $\{x_i\}_{i=1}^n$, target points $\{y_j\}_{j=1}^n$.
- Distributions $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\beta = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$, with $\delta_x(A) = 1$ if $x \in A$, 0 otherwise.
- Cost matrix C, $C_{i,j} = c(x_i, y_j)$.

Objective:

$$\min_{\sigma \in \mathrm{Perm}(n)} \sum_{i=1}^{n} C_{i,\sigma(i)}$$

under the constraint that σ associates each x_i with a unique y_j.

Monge's formulation laid the groundwork for the field of optimal transport, posing essential questions regarding existence, uniqueness, and optimality of solutions. However, the restrictive

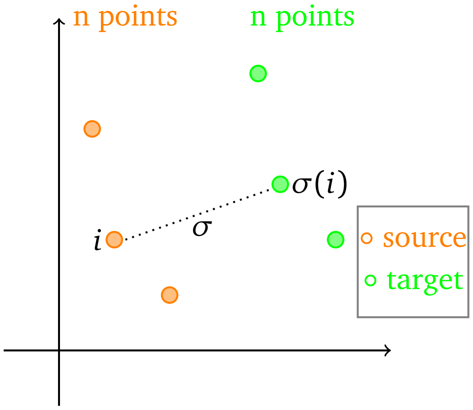

Figure 1.1: Distribution of Monge's Problem

nature of requiring a one-to-one mapping between sources and targets introduces complexities that can complicate both theoretical analysis and practical computation.

1.2 Introduction to the *Kantorovich* Problem

Building upon Monge's foundational work, *Leonid Kantorovich* introduced a relaxed version of the optimal transport problem in the 1940s. Recognizing the challenges inherent in Monge's formulation, Kantorovich's approach allows for the distribution of mass from source points to multiple target points, thereby broadening the applicability and enhancing the mathematical tractability of optimal transport theory. This generalization not only addresses some of the limitations of Monge's original problem but also paved the way for significant advancements in both theoretical and applied aspects of optimal transport.

1.2.1 Definition of Mass Distributions

Consider two vectors $a \in \sigma_n$ and $b \in \sigma_m$, where $\sigma_n = \{a \in \mathbb{R}_+^n : \sum_{i=1}^n a_i = 1\}$ represents the set of probability distributions over n points. Here, a represents the initial mass distribution over the source points, and b describes how the mass should be distributed to the target points (see Figure 1.2).

1.2.2 Transport Plan: Design and Constraints

The goal is to determine a transport plan P, a matrix of dimension $n \times m$, which distributes the mass from the source points x_i to the target points y_j. Each element $P_{i,j}$ of the matrix indicates the amount of mass transferred from the source point x_i to the target point y_j. The plan P belongs to the set $U(a,b) \subseteq \mathbb{R}_+^{n \times m}$, which is defined by:

$$U(a,b) = \{P \in \mathbb{R}_+^{n \times m} : P\mathbb{1}_m = a,\ P^\top \mathbb{1}_n = b\}.$$

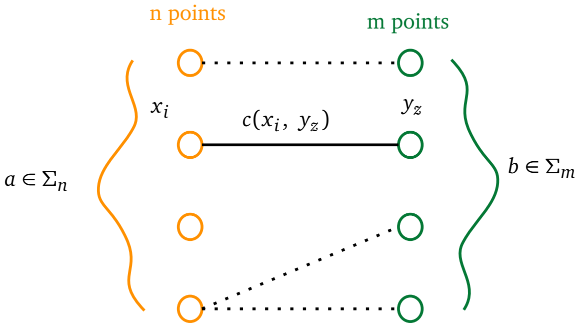

Figure 1.2: Distribution of Monge's Problem

The conditions $P\mathbb{1}_m = a$ and $P^\top \mathbb{1}_n = b$ ensure conservation of mass, meaning that the sum of the elements of each row of P corresponds to the initial mass a_i at each source, and the sum of the elements of each column (after transposition) corresponds to the final mass b_j at each target.

1.2.3 Optimisation Problem

The objective of the Kantorovich problem is to minimise the total cost of transport defined by:

$$\min_{P \in U(a,b)} \sum_{i,j} P_{i,j} C_{i,j} = \langle C, P \rangle$$

This problem is formulated as a linear optimisation, seeking the transport plan P that minimises the costs.

Definition 1. *The Wasserstein distance W_p between two arbitrary probability distributions $\alpha \in \mathscr{P}(\Omega)$ and $\beta \in \mathscr{P}(\Omega)$ can be transported by a measure π such that:*

$$W_p(\alpha, \beta) = \left(\min_{\pi \in U(\alpha,\beta)} \int_{\Omega \times \Omega} \|x - y\|^p \, d\pi(x, y) \right)^{1/p} = \left(\mathbb{E}_{(x,y)\sim\pi}[\|x - y\|^p] \right)^{1/p}$$

Where $U(\alpha, \beta) = \{\pi \in \mathscr{P}(X \times Y) : \pi(x, \cdot) = \alpha,\ \pi(\cdot, y) = \beta\}$.

The Wasserstein distance can be seen as a generalisation of the L^p distance, which measures the distance between two points in Euclidean space. Here, it measures the "minimal rearrangement" needed to transform one probability distribution into another, taking into account both the individual distances between all points x and y and the transport plan π that optimises this cost.

Kantorovich Problem

Data:

- Source points $\{x_i\}_{i=1}^n$, target points $\{y_j\}_{j=1}^m$.

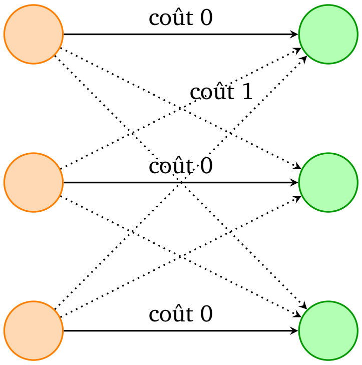

Figure 1.3: Illustration of an example of the application of the Kantorovich algorithm for $n = m$

- Distributions $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$, $\beta = \sum_{j=1}^{m} b_j \delta_{y_j}$, with $\delta_x(A) = 1$ if $x \in A$, 0 otherwise.
- Cost matrix C, $C_{i,j} = c(x_i, y_j)$.

Objective:

$$\min_{P \in U(\alpha,\beta)} \langle P, C \rangle_F = \sum_{i,j} P_{i,j} C_{i,j}$$

where $U(\alpha, \beta)$ is defined by:

$$U(\alpha,\beta) = \left\{ P \in \mathbb{R}_+^{n \times m} \;\middle|\; P\mathbf{1}_m = a,\ P^T\mathbf{1}_n = b \right\}.$$

1.2.4 Example of Application: Special Case $n = m$

Consider a special case of the Kantorovich problem where $n = m$ and the cost matrix $C_{i,j}$ is defined as $C_{i,j} = 0$ if $i = j$ and $C_{i,j} = 1$ otherwise (see Figure 1.3). Our goal is to determine the best strategy to solve this problem by maximising the amount transported from i to j.

Example 1. *Optimal Strategy For each pair (i, j), we apply the following strategies:*

1. *If $i = j$, the optimal transport plan P_{ii} is given by:*

$$P_{ii} = \min(a_i, b_i) = c_{i,i}$$

2. *If $i \neq j$, the amount of mass to be transferred is:*

$$P_{ij} = \frac{(a_i - \min(a_i, b_i))(b_j - \min(a_j, b_j))}{1 - \sum_{k=1}^{n} \min(a_k, b_k)}$$

Proposition 1 (Optimality of the Strategy)**.** *We will show that this strategy is the best possible by using two key results:*

1. *The total cost of transport for the plan $\hat{P}$ is given by:*

$$\langle C, \hat{P} \rangle = 1 - \sum_{i=1}^{n} c_{i,i}$$

2. For any admissible plan $P \in U(a,b)$, we have:

$$\langle C, \hat{P} \rangle \leq \langle C, P \rangle$$

Moreover, for any pair (i,j), $P_{i,j} \leq \min(a_i, b_j)$.

Proof. 1. For the first point, since $C_{i,j} = 0$ for $i = j$, the cost only depends on transports $i \neq j$. Thus, the total cost of transport is directly related to the quantity unsatisfied by the minimum, that is $1 - \sum_{i=1}^{n} c_{i,i}$.

2. For the second point, consider an admissible plan $P \in U(a,b)$. Each element $P_{i,j}$ is constrained by the minimum possible quantity, thus $P_{i,j} \leq \min(a_i, b_j)$. Since the strategy maximises P_{ii}, it follows that the cost $\langle C, \hat{P} \rangle$ reaches its lower bound, proving its optimality. □

1.2.5 The Monge-Mather Shortening Principle

Consider the support of the optimal transport plan P, defined as:

$$\operatorname{supp}(P) = \left\{(i,j) \in [n] \times [m] : P_{ij} > 0\right\}$$

Theorem 1 (Monge-Mather Shortening Principle). *If $(i_1, j_1), (i_2, j_2) \in \operatorname{supp}(P)^2$, then the segments $[\mathbf{x}_{i_1}, \mathbf{y}_{j_1}]$ and $[\mathbf{x}_{i_2}, \mathbf{y}_{j_2}]$ do not intersect.*

Proof. Suppose we have $(i_1, j_1) = (1,1)$ and $(i_2, j_2) = (2,2)$ in $\operatorname{supp}(P)^2$.

If the segments $[\mathbf{x}_1, \mathbf{y}_1]$ and $[\mathbf{x}_2, \mathbf{y}_2]$ intersect at point z, then:

$$\|\mathbf{x}_1 - \mathbf{y}_1\| = \|\mathbf{x}_1 - z\| + \|z - \mathbf{y}_1\|$$

$$\|\mathbf{x}_2 - \mathbf{y}_2\| = \|\mathbf{x}_2 - z\| + \|z - \mathbf{y}_2\|$$

This would imply:

$$\|\mathbf{x}_1 - \mathbf{y}_1\| + \|\mathbf{x}_2 - \mathbf{y}_2\| > \|\mathbf{x}_1 - \mathbf{y}_2\| + \|\mathbf{x}_2 - \mathbf{y}_1\|$$

In terms of cost:

$$c(\mathbf{x}_1, \mathbf{y}_1) + c(\mathbf{x}_2, \mathbf{y}_2) > c(\mathbf{x}_1, \mathbf{y}_2) + c(\mathbf{x}_2, \mathbf{y}_1)$$

Construct a new plan $\hat{P}$ by adjusting P:

$$\hat{P} = \begin{bmatrix} P_{11} - \epsilon & P_{12} + \epsilon & \dots \\ P_{21} + \epsilon & P_{22} - \epsilon & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Where $\epsilon = \min(P_{11}, P_{22}) > 0$ because $(1,1)$ and $(2,2)$ are in $\operatorname{supp}(P)^2$.

Thus, $\hat{P} \in U(a,b)$ still respects the mass conservation constraints.

The cost difference is:

$$\langle C, \hat{P} \rangle - \langle C, P \rangle = \epsilon(c(\mathbf{x}_2, \mathbf{y}_1) + c(\mathbf{x}_1, \mathbf{y}_2) - c(\mathbf{x}_1, \mathbf{y}_1) - c(\mathbf{x}_2, \mathbf{y}_2)) < 0$$

This implies that $\langle C, \hat{P} \rangle < \langle C, P \rangle$, which contradicts the optimality of P. Therefore, the segments cannot intersect except at their endpoints. □

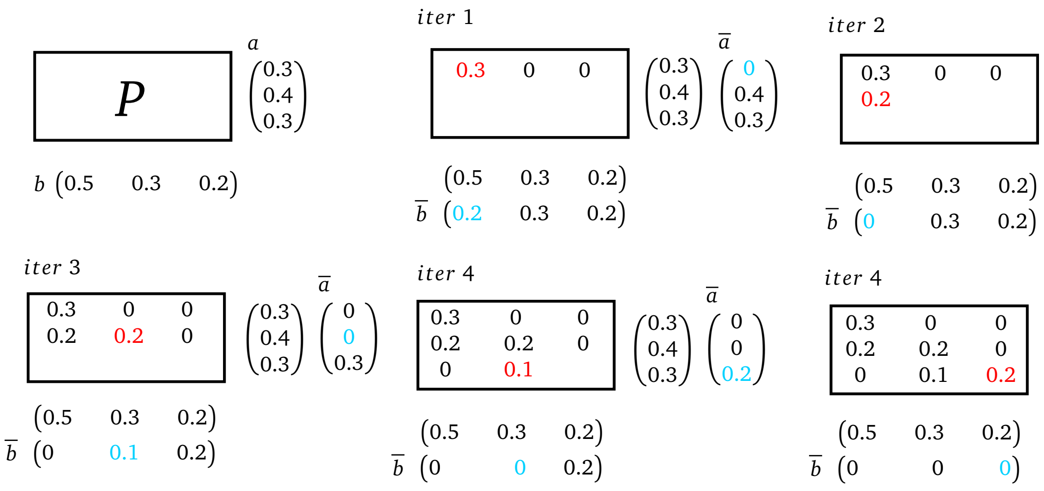

Figure 1.4: Example Execution of the Northwest Corner Rule Algorithm

1.2.6 The Northwest Corner Rule

The Northwest Corner Rule is an algorithm for initializing a transport plan between two mass distributions, used to solve the optimal transport problem.

Northwest Corner Rule Algorithm

Let **a** and **b** be vectors of size n and m, respectively.

```
1. Initialise $\overline{\mathbf{a}} = \mathbf{a}$, $\overline{\mathbf{b}} = \mathbf{b}$, and $(i, j) = (1, 1)$.
2. While $i \leq n$ and $j \leq m$, do:
    • $P_{ij} = \min\{\overline{a}_i, \overline{b}_j\}$.
    • $\overline{a}_i \leftarrow \overline{a}_i - P_{ij}$, $\overline{b}_j \leftarrow \overline{b}_j - P_{ij}$.
    • If $\overline{a}_i = 0$, then $i \leftarrow i + 1$.
    • If $\overline{b}_j = 0$, then $j \leftarrow j + 1$.
3. Return $P$.
```

This process performs mass transfers in a systematic order, starting from the northwest "corner" of the cost matrix and moving either south or east once one of the margins is saturated (see Figure 1.4).

Theorem 2. *The northwest corner rule algorithm operates in $\mathcal{O}(n + m)$ operations.*

Proof. The main loop executes at most n + m iterations. □

Theorem 3. *The northwest corner rule algorithm produces a valid solution to the transport problem.*

Proof. Let P be the solution obtained by the northwest corner rule algorithm. Suppose, for contradiction, that P is not optimal. Consider Q an optimal solution such that $Q \neq P$.

Identify the first differing index: let (i_l, j_l) be the first index such that $P_{i_l,j_l} \neq Q_{i_l,j_l}$.

1. According to the algorithm, we have $P_{i,j} = \min(\bar{a}_i, \bar{b}_j)$. This implies $P_{i_l,j_l} > Q_{i_l,j_l}$, meaning that mass is missing in Q at this location.

2. Since P has saturated the position with the minimum of the margins at (i_l, j_l), we have:

$$Q_{i_l,j_l} < \min(\bar{a}_{i_l}, \bar{b}_{j_l}) \leq \min(a_{i_l}, b_{j_l})$$

3. To compensate for this deficit, mass in Q must be added after (i_l, j_l). Therefore, there exist indices $r > i_l$ and $s > j_l$ such that $Q_{r,j_l} > 0$ and $Q_{i_l,s} > 0$.

Define a slight modification of Q:

$$\epsilon = \min(Q_{r,j_l}, Q_{i_l,s}) > 0$$

By constructing $\hat{P}$ from Q:

$$\hat{P}_{i_l,j_l} = Q_{i_l,j_l} + \epsilon$$

$$\hat{P}_{i_l,s} = Q_{i_l,s} - \epsilon$$

$$\hat{P}_{r,s} = Q_{r,s} + \epsilon$$

$$\hat{P}_{r,j_l} = Q_{r,j_l} - \epsilon$$

Thus, $\hat{P} \in U(a, b)$.

The cost difference is:

$$\langle C, \hat{P} \rangle - \langle C, Q \rangle = \epsilon(c_{i_l,j_l} + c_{r,s} - c_{i_l,s} - c_{r,j_l}) \leq 0$$

This means that $\langle C, \hat{P} \rangle \leq \langle C, Q \rangle$, which contradicts the optimality of Q since we have constructed a better solution.

Conclusion: P must be optimal, proving that the algorithm produces a valid solution. □

Corollary 1. *If $x_i, y_j \in \mathbb{R}$, and $c(x, y) = h(x, y)$ with h convex, then the transport can be resolved in $O(n \log n + m \log m)$.*

Proof. Exercise left to the reader. □

1.2.7 Monge Matrix

Definition 2. *In one-dimensional distributions, a matrix C is said to be a Monge matrix if it satisfies:*

$$\forall(i,j),\ C_{i,j} + C_{i+1,j+1} \leq C_{i+1,j} + C_{i,j+1}$$

In this case, applying the Northwest Corner Rule ensures an optimal solution for the transport problem.

Example 2. *Consider the points $x_1 \leq \cdots \leq x_n$ and $y_1 \leq \cdots \leq y_m$ are ordered. Then, the matrix $C = \left(\left|x_i - y_j\right|^2\right)_{i,j}$ is a Monge matrix. More generally, if $C = \left(h(x_i - y_j)\right)_{i,j}$ where h is a convex function, then C is also a Monge matrix.*

Proof. Let's demonstrate the property when h is convex.

For two consecutive indices (i,j) and $(i+1,j+1)$, with the points x and y sorted, consider the following expression:

$$C(i,j) + C(i+1,j+1) - C(i,j+1) - C(i+1,j)$$

Substituting C with $h(x_i - y_j)$, we obtain:

$$h(x_i - y_j) + h(x_{i+1} - y_{j+1}) - h(x_i - y_{j+1}) - h(x_{i+1} - y_j)$$

For the quadratic function $h(u) = u^2$, we calculate this expression:

$$\begin{aligned}&(x_i - y_j)^2 + (x_{i+1} - y_{j+1})^2 - (x_i - y_{j+1})^2 - (x_{i+1} - y_j)^2 \\ &= 2\left[x_i(y_{j+1} - y_j) - x_{i+1}(y_{j+1} - y_j)\right] \\ &= 2(y_{j+1} - y_j)(x_i - x_{i+1})\end{aligned}$$

Given that the points are ordered, $y_{j+1} \geq y_j$ and $x_i \leq x_{i+1}$. Thus, the product $(y_{j+1} - y_j)(x_i - x_{i+1})$ is less than or equal to zero, implying that:

$$C(i,j) + C(i+1,j+1) - C(i,j+1) - C(i+1,j) \leq 0$$

This satisfies the Monge property.

Now, consider a general convex function h. To show that C is a Monge matrix, we must demonstrate that for all $i < k$ and $j < l$:

$$C(i,j) + C(k,l) \leq C(i,l) + C(k,j)$$

Applying the convexity of h, which implies that for all $a \leq b \leq c \leq d$:

$$h(a) + h(d) \leq h(b) + h(c)$$

In our context, setting $a = x_i - y_j$, $b = x_i - y_{j+1}$, $c = x_{i+1} - y_j$, and $d = x_{i+1} - y_{j+1}$, we have:

$$h(x_i - y_j) + h(x_{i+1} - y_{j+1}) \leq h(x_i - y_{j+1}) + h(x_{i+1} - y_j)$$

Which confirms that C satisfies the Monge inequality for any convex function h. Thus, C is a Monge matrix. □

Chapter 2

Some optimisation tools

This section is taken entirely from the course notes of *Mathurin Massias*, *Optimization for large scale machine learning* [1]

2.1 Calculus - Reminder

Definition 3 (Gradient vector). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Then for any $x \in \mathbb{R}^d$, there exists a vector of $\mathbb{R}^d$, denoted $\nabla f(x)$ and called the gradient of f at x, such that for all $h \in \mathbb{R}^d$,*

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|) \tag{2.1}$$

This vector is unique.

This is a linear approximation at f, generalizing the well-known Taylor formula in 1D.

Definition 4 (Hessian matrix). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable. Then for any $x \in \mathbb{R}^d$, there exists a matrix in $\mathbb{R}^{d \times d}$, denoted $\nabla^2 f(x)$ and called the Hessian of f at x, such that for all $h \in \mathbb{R}^d$,*

$$f(x+h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} h^\top \nabla^2 f(x) h + o\left(\|h\|^2\right) \tag{2.2}$$

As the gradient, if the Hessian exists it is uniquely defined.

Much like Equation 2.1, Equation 2.2 is a local approximation of f, but this time quadratic (Taylor of order 2). The uniqueness property is very useful to compute gradients and Hessians: if $f(x+h) = f(x) + \langle v, h \rangle + o(\|h\|)$, then v must be equal to $\nabla f(x)$. Hence a technique to compute $\nabla f(x)$ is to write $f(x+h)$, try to isolate $f(x)$ and a linear term, and then identify the slope of the linear term with $\nabla f(x)$.

When f is vector-valued, there is a generalization of the gradient, called the Jacobian.

[1] You can find the course notes here : mathurinm.github.io/class.pdf.

Definition 5 (Jacobian)**.** *The Jacobian of* $f : x \in \mathbb{R}^n \mapsto \begin{pmatrix} f_1(x) \\ \cdots \\ f_m(x) \end{pmatrix}$ *is the matrix of partial derivatives:*

$$\left(\frac{\partial f_i}{\partial x_j}(x)\right)_{ij} \in \mathbb{R}^{m\times n} \tag{2.3}$$

For real-valued functions, the Jacobian is the transposed gradient. Else, the Jacobian consists of stacked transposed gradients:

$$\mathscr{J}_f(x) = \begin{pmatrix} -\nabla f_1(x)^\top - \\ \cdots \\ -\nabla f_n(x)^\top - \end{pmatrix} \tag{2.4}$$

The Jacobian of the gradient is the Hessian (transposed, equal if f is twice differentiable by Clairaut/Schwarz theorem; holds if second partial derivatives are continuous or more weakly if partial derivatives are continuous):

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \tag{2.5}$$

$$\mathscr{J}_{\nabla f}(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1}(x) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{pmatrix} \tag{2.6}$$

Proposition 2 (Chain rule)**.** *Let* $f : \mathbb{R}^d \to \mathbb{R}^n$ *and* $g : \mathbb{R}^n \to \mathbb{R}^p$ *be differentiable functions. Then their composition* $g \circ f$ *is differentiable and*

$$\mathscr{J}_{g\circ f}(x) = \mathscr{J}_g(f(x))\mathscr{J}_f(x) \tag{2.7}$$

2.2 Minimization and Convexity

Definition 6 (Global and local minimizers)**.** *Let* $f : \mathbb{R}^d \to \mathbb{R}$*. A global minimizer of* f *is a point* x^* *such that* $f(x^*) \leq f(x)$ *for all* $x \in \mathbb{R}^d$*. A local minimizer of* f *is a point* x^* *such that there exists a neighborhood* V *of* x^* *such that* $f(x^*) \leq f(x)$ *for all* $x \in V$ *(See Figure 2.1).*

Definition 7 (Convexity)**.** *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is convex if and only if it lies below its chords:*

$$\forall x, y \in \mathbb{R}^d, \forall \lambda \in]0,1[, \quad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \tag{2.8}$$

It is strictly convex if the inequality (2.8) holds strictly.

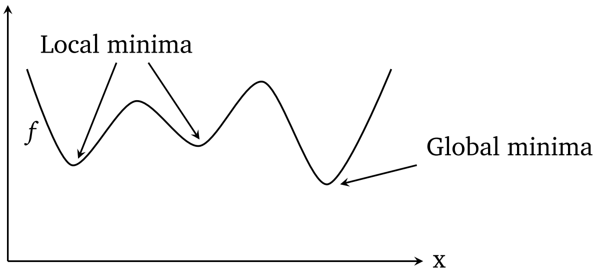

Figure 2.1: Example of a function f with local minimisers and a global minimum

Remark 1. *Equation (2.8) trivially holds for $\lambda = 0$ and $\lambda = 1$, and so we could equivalently define it for $\lambda \in [0,1]$. However, we will later work with functions that can take the value $+\infty$, and in that case, we could end up with $\lambda f(x) = 0 \times (+\infty)$ which is not defined. So it is more rigorous to require the convexity inequality to hold only for $\lambda \in]0,1[$.*

Definition 8 (Strong convexity)**.** *Let $\mu > 0$ and $\|\cdot\|$ be a norm. A function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex with respect to $\|\cdot\|$ if for all $x, y \in \mathbb{R}^d$ and $\lambda \in]0,1[$,*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2 \tag{2.9}$$

When $\|\cdot\|$ is the Euclidean norm, then f is μ-strongly convex if and only if $f - \frac{\mu}{2}\|\cdot\|^2$ is convex. In this sense, a strongly convex function is so convex that when you subtract a parabola with positive curvature, it remains convex.

2.3 Lagrangian Duality and KKT Conditions

Lagrangian duality is a powerful framework used to address optimization problems, especially those involving convex functions with affine constraints. These problems can be stated in the following form:

$$\min_{x \in \mathbb{R}^d} f(x) \quad \text{s.t.} \quad g_i(x) \leq 0 \forall i \in [m], \quad h_i(x) = 0 \forall i \in [p] \tag{2.10}$$

We write the inequality constraints as ≤ 0, because we like working with convex functions and convex sets; if g is convex, its sublevel sets, and in particular $\{x : g(x) \leq 0\}$, are convex—this would not be true for $\{x : g(x) \geq 0\}$.

Definition 9 (Lagrangian, dual function, Lagrangian dual)**.** *The Lagrangian associated with Problem (2.10) is the function*

$$\begin{aligned} \mathscr{L} : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m_+ &\to \mathbb{R} \\ (x, \lambda, \mu) &\mapsto f(x) + \sum_{i=1}^{n} \lambda_i h_i(x) + \sum_{i=1}^{p} \mu_i g_i(x), \end{aligned} \tag{2.11}$$

where λ and μ are called the Lagrange multipliers. Note that the multipliers μ_i associated with the inequality constraints are positive; the rationale is that as soon as x does not satisfy one constraint, maximizing $\mathscr{L}$ in λ_i or μ_i will make the Lagrangian go to $+\infty$, so that Problem (2.10) is equivalent to:

$$\min_{x\in\mathbb{R}^n} \max_{\lambda\in\mathbb{R}^p,\mu\in\mathbb{R}^m_+} \mathscr{L}(x,\lambda,\mu) \tag{2.12}$$

Also note that for any x feasible for Problem (2.10), the Lagrangian value for any $\lambda, \mu \geq 0$ upper bounds $f(x)$.

Finally, the dual function is:

$$\begin{aligned} g : \mathbb{R}^p \times \mathbb{R}^m_+ &\to \mathbb{R} \\ (\lambda,\mu) &\mapsto \min_{x\in\mathbb{R}^n} \mathscr{L}(x,\lambda,\mu) \end{aligned} \tag{2.13}$$

and the Lagrange dual problem of Problem (2.10) is:

$$\max_{\lambda\in\mathbb{R}^p,\mu\in\mathbb{R}^m_+} g(\lambda,\mu), \tag{2.14}$$

that is, Problem (2.12) but with the order of the min and the max inverted.

We now introduce a very powerful tool to solve constrained optimization problems: the Karush-Kuhn-Tucker (KKT) conditions.

Definition 10 (KKT Conditions). *The triplet (x,λ,μ) is said to satisfy the KKT conditions for Problem (2.10) if it is a saddle point for the Lagrangian, meaning that it satisfies:*

$$\begin{aligned} \nabla_x \mathscr{L}(x,\lambda,\mu) &= 0 & & \textit{(stationarity)} \\ g_i(x) &\leq 0 & \forall i \in [m] & \quad \textit{(primal feasibility)} \\ h_i(x) &= 0 & \forall i \in [p] & \quad \textit{(primal feasibility)} \\ \mu &\geq 0 & & \textit{(dual feasibility)} \\ \mu_i g_i(x) &= 0 & \forall i \in [m] & \quad \textit{(complementary slackness)} \end{aligned} \tag{2.15}$$

For the sake of clarity, we have used a gradient for stationarity (thus implying that everything is differentiable); but the same results exist with subgradients. Actually, the complementary slackness together with dual feasibility is easier understood as $\mu_i \in \partial \iota_{\mathbb{R}^-}(g_i(x))$.

For many simple problems, it is easier to solve the above system. Under reasonable conditions, solutions of the minimization problem are amongst KKT points.

Proposition 3 (KKT are sufficient in the convex + affine equality case). *Let the h_i's be affine. Let f and the g_i's all be convex (and differentiable for easier proof, but everything works with subgradients). Then the KKT are sufficient: if they hold at x^*, λ^*, μ^*, then x^* is a global minimizer of the constrained problem.*

Proof. By the requirement on f, g_i, the h_i and the fact that $\mu^* \geq 0$, the Lagrangian $\mathscr{L}(\cdot,\lambda^*,\mu^*)$ is convex (because we do not know the sign of λ^*, we must require the h_i's to be affine for $\lambda_i h_i$ to be convex).

Because of this convexity, the KKT condition $\nabla_x \mathscr{L}(\cdot, \lambda^*, \mu^*) = 0$ implies that x^* is a minimizer of $\mathscr{L}(\cdot, \lambda^*, \mu^*)$. Therefore for any x, $\mathscr{L}(x, \lambda^*, \mu^*) \geq \mathscr{L}(x^*, \lambda^*, \mu^*) = f(x^*)$ (because of feasibility of x^* and complementary slackness).

Now if we take x feasible, the LHS is $f(x) + \sum_i \mu_i^* g_i(x^*)$, which is smaller than $f(x)$ ($\mu^* \geq 0$ and x feasible so $g_i(x) \leq 0$). So $f(x) \geq f(x^*)$ for any feasible x, which concludes the proof. □