Computational optimal transport

Bernhard Schmitzer

Uni Göttingen, winter term 2021, January 19, 2022

1 Introduction

[for some 'global announcements' see StudIP]

1.1 Literature

- Gabriel Peyré, Marco Cuturi: Computational Optimal Transport, Foundations and Trends in Machine Learning, 2019, 11, 355-607, available online: https://optimaltransport.github.io/book/
  Introduction with a focus on computational aspects, avoiding mathematical details.
- Filippo Santambrogio: Optimal Transport for Applied Mathematicians, Birkhäuser Boston, 2015
  Introduction aimed at applied mathematicians, 'harder' than the other reference, but maybe more suitable to analytically inclined students.

1.2 Tentative table of contents

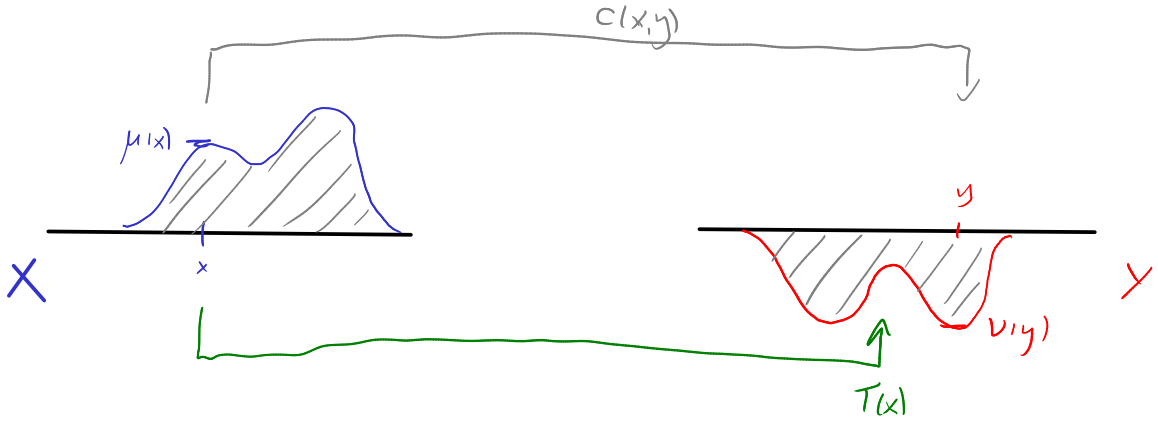
- Introduction: Monge and Kantorovich problem, motivation for application in data analysis
- Kantorovich duality
  - dual problem, primal-dual optimality conditions
- OT in one dimension
- mini-interlude: measures and weak* convergence
- Wasserstein distances
  - triangle inequality, displacement interpolation
- W1 on graphs
  - Kantorovich–Rubinstein duality
  - equivalence with min-cost-flow problem
- classical algorithms
  - Hungarian method
  - auction algorithm

- entropic regularization
  - regularized primal and dual
  - Sinkhorn algorithm (derivation, epsilon scaling, numerical stability)
- mini-interlude: basics of convex analysis
  - subdifferentials, Fenchel–Legendre conjugation
  - duality in optimization
- unbalanced transport
- Wasserstein barycenters
- prototypical data analysis / machine learning applications

2 First contact with optimal transport: the principle of least effort

2.1 Gaspard Monge: piles of sand

- 1746 - 1818, French mathematician, engineer and politician
- prototypical optimal transport problem: sand piles and holes
  - $\mu(x)$: height of pile at x, $\nu(y)$: depth of hole at y, volumes must be equal $\int_X \mu(x)\,\mathrm{d}x = \int_Y \nu(y)\,\mathrm{d}y$



  - we want to use the sand to fill the hole in most efficient way
  - cost of moving one unit of sand from x to y: $c(x, y)$ (e.g. distance, maybe taking into account obstacles, difficulty of path,...)
  - transport map $T : X \to Y$, take grains of sand from x to $T(x)$, requirement: moving all grains of sand in μ along T must result in distribution ν to fill the hole. For now think of this intuitively, by splitting μ into small (infinitesimal) grains of sand that are then moved individually. We will later give a more rigorous definition. We call the transformation that T induces on μ the 'push-forward' and denote it by $T_\sharp\mu$.
  - total transport cost associated with map T: $\int_X c(x, T(x))\mu(x)\,\mathrm{d}x$

– optimal transport problem according to Monge:

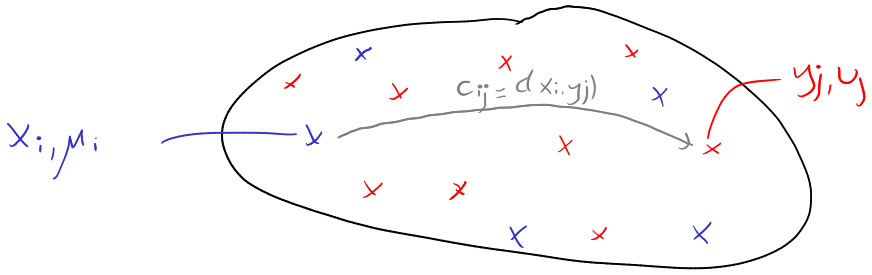
$$\inf\left\{\int_X c(x,T(x))\mu(x)\,\mathrm{d}x \,\middle|\, T : X \to Y \text{ such that } T_\sharp\mu = \nu\right\}$$

– rigorous mathematical analysis (do minimal T even exist? what properties do they have?) was not possible at the time. Properly solved only at the end of the 20th century.

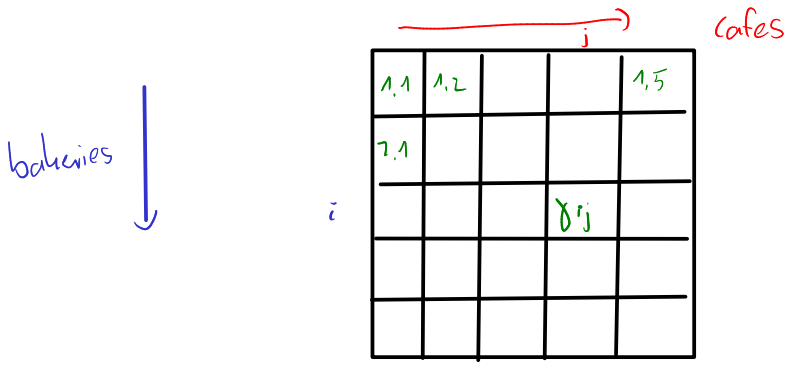
– intuitively, sometimes a map T that satisfies $T_\sharp\mu = \nu$ may not exist. Maybe there are two grains of sand at x, one needs to go to y_1, the other to y_2.

2.2 Leonid Kantorovich: cafes in Paris

- 1912 - 1986, Soviet mathematician, inventor of linear programming, his work was applied to optimization of industrial production efficiency by the Soviet government
- toy problem: bakeries and cafes in Paris $= \Omega \subset \mathbb{R}^2$.
  – let $(x_i)_{i=1}^M$ be locations of bakeries in Ω, each morning, each bakery i produces an amount $\mu_i \geq 0$ of bread



  – locations of cafes given by $(y_j)_{j=1}^N$ in Ω, each morning, each cafe orders amount $\nu_j \geq 0$ of bread for sale during the day
  – we work for the Parisian bakery-cafe-commission, need to work out which bakery delivers to what cafes
  – total amounts of bread have been coordinated: $\sum_{i=1}^M \mu_i = \sum_{j=1}^N \nu_j$
  – as in Monge's case: cost function $c : \Omega \to \Omega$, $c(x,y)$ describes how much effort is required to transport one unit of bread from location x to y.
  – problem: transport-map ansatz of Monge will not work. Most bakeries deliver bread to more than one cafe.
- need new description for bread allocation: transport plan.



- intuitively, a big table of size $M \times N$, stored as $\gamma \in \mathbb{R}_+^{M \times N}$, where each entry $\gamma_{i,j}$ specifies, how much bread goes from bakery i to cafe j.
- need to make sure that all bakeries 'get rid of all their bread': $\sum_{j=1}^{N} \gamma_{i,j} = \mu_i$ for $i = 1, \dots, M$. likewise, each cafe gets the ordered amount of bread: $\sum_{i=1}^{M} \gamma_{i,j} = \nu_j$ for $j = 1, \dots, N$.
- set of all transport plans

$$\Gamma(\mu, \nu) := \left\{ \gamma \in \mathbb{R}_+^{M \times N} \,\middle|\, \sum_{j=1}^{N} \gamma_{i,j} = \mu_i \text{ for } i = 1, \dots, M, \; \sum_{i=1}^{M} \gamma_{i,j} = \nu_j \text{ for } j = 1, \dots, N \right\}$$

- cost associated with transport plan: $\langle c, \gamma \rangle := \sum_{i=1}^{M} \sum_{j=1}^{N} c_{i,j} \cdot \gamma_{i,j}$ where $c_{i,j} := c(x_i, y_j)$ are entries of cost matrix
- Kantorovich optimal transport problem: $\inf \{ \langle c, \gamma \rangle | \gamma \in \Gamma(\mu, \nu) \}$
- this is a linear program: $\gamma \mapsto \langle c, \gamma \rangle$ is linear, inequality constraints $\gamma_{i,j} \geq 0$ are linear, 'amount-of-bread'-constraints are linear
- the set $\Gamma(\mu, \nu)$ is bounded and non-empty (exercise) $\Rightarrow$ minimizing γ in Kantorovich problem exists

- outlook: generalization of Kantorovich problem
  - we only considered finite, discrete locations $(x_i)_{i=1}^M$ and $(y_j)_{j=1}^N$ with discrete mass distributions μ and ν
  - Kantorovich problem can be generalized to arbitrary measures, e.g. diffuse distributions of sand as in Monge's example
  - then $\Gamma(\mu, \nu)$ becomes a set of measures on the product space Ω^2 (or $X \times Y$)
  - can also compare discrete distribution (locations of large shops with product reserves, or schools with teaching capacities) with approximately diffuse distributions such as locations of customers or pupils

2.3 Motivation: applications in data analysis

Example 2.1 (Matching point clouds)**.** see python example `001_PointCloudMatching` and notes on 'serializing' the linear program (i.e. transforming minimization over matrix into minimization over vector) in `2021-04-14_ComputationalOT_sketches`.

Example 2.2 (Matching histograms)**.** see doodles in `2021-04-14_ComputationalOT_sketches`

3 The Kantorovich optimal transport problem

3.1 Primal problem

We state the Kantorovich problem with slightly more formal care than in the previous section.

Probability simplex. Throughout the whole section let $M, N \in \mathbb{N}$ be fixed. Denote by

$$\Sigma_M := \left\{ \mu \in \mathbb{R}_+^M \middle| \sum_{i=1}^M \mu_i = 1 \right\}$$

the simplex of discrete probabilities over M points (and likewise use Σ_N).

Marginal projection operators. We have seen that the Kantorovich problem is a linear program and have 'serialized' (i.e. transformed the matrix γ into a vector) it in Example 2.1 so that the row and column sums could be written by matrix-vector multiplications. It will be convenient (and more general) to adopt a slightly more abstract view in the following. Introduce row-sum operators:

$$\mathrm{P}_X : \mathbb{R}^{M\times N} \to \mathbb{R}^M, \qquad (\mathrm{P}_X\gamma)_i := \sum_{j=1}^N \gamma_{i,j} \quad \text{for} \quad i = 1, \dots, M,$$

$$\mathrm{P}_Y : \mathbb{R}^{M\times N} \to \mathbb{R}^N, \qquad (\mathrm{P}_Y\gamma)_j := \sum_{i=1}^M \gamma_{i,j} \quad \text{for} \quad j = 1, \dots, N.$$

Clearly, both are linear operators. They take a matrix to a vector.

Definition 3.1 (Primal Kantorovich problem)**.** For $\mu \in \Sigma_M$, $\nu \in \Sigma_N$ the set of optimal transport plans between them is given by

$$\Gamma(\mu, \nu) := \left\{ \gamma \in \mathbb{R}_+^{M\times N} \middle| \mathrm{P}_X\gamma = \mu, \mathrm{P}_Y\gamma = \nu \right\}.$$

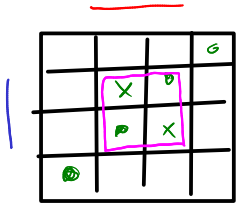
For $c \in \mathbb{R}^{M\times N}$ the associated primal Kantorovich optimal transport problem is given by

$$\mathcal{C}(\mu, \nu) := \inf \left\{ \langle c, \gamma \rangle | \gamma \in \Gamma(\mu, \nu) \right\}.$$

Remark 3.2. We have shown that the set $\Gamma(\mu, \nu)$ is non-empty, closed and bounded (and therefore compact). The function $\gamma \mapsto \langle c, \gamma \rangle$ is continuous. Hence, minimal γ exist.

Proposition 3.3 (Restrictions of optimal plans are optimal)**.** Let $\mu \in \Sigma_M$, $\nu \in \Sigma_N$, $\gamma \in \Gamma(\mu, \nu)$ minimal for $\mathcal{C}(\mu, \nu)$ and $\tilde{\gamma} \in \mathbb{R}_+^{M\times N}$ with $\tilde{\gamma} \leq \gamma$ (inequality holds for all entries). Set $\tilde{\mu} = \mathrm{P}_X\tilde{\gamma}$, $\tilde{\nu} = \mathrm{P}_Y\tilde{\gamma}$. Then $\tilde{\gamma}$ is optimal for $\mathcal{C}(\tilde{\mu}, \tilde{\nu})$.

Proof in exercise.



Proposition 3.4 (Convexity of optimal total cost)**.** The function $\Sigma_M \times \Sigma_N \ni (\mu, \nu) \mapsto \mathcal{C}(\mu, \nu)$ is convex.

Proof.

- Let $\mu_0, \mu_1 \in \Sigma_M$, $\nu_0, \nu_1 \in \Sigma_N$. Let γ_0, γ_1 be corresponding optimal plans.
- For $\lambda \in [0, 1]$ set

$$\tilde{\mu} := (1-\lambda) \cdot \mu_0 + \lambda \cdot \mu_1, \qquad \tilde{\nu} := (1-\lambda) \cdot \nu_0 + \lambda \cdot \nu_1, \qquad \tilde{\gamma} := (1-\lambda) \cdot \gamma_0 + \lambda \cdot \gamma_1.$$

- It is clear that $(\tilde{\mu}, \tilde{\nu}) \in \Sigma_M \times \Sigma_N$. Need to show that

$$\mathcal{C}(\tilde{\mu}, \tilde{\nu}) \leq (1-\lambda) \cdot \mathcal{C}(\mu_0, \nu_0) + \lambda \cdot \mathcal{C}(\mu_1, \nu_1).$$

- Show first that $\tilde{\gamma} \in \Gamma(\tilde{\mu}, \tilde{\nu})$: By construction $\tilde{\gamma} \geq 0$. Check row sums:

$$\mathrm{P}_X \tilde{\gamma} = \mathrm{P}_X[(1-\lambda) \cdot \gamma_0 + \lambda \cdot \gamma_1] = (1-\lambda) \cdot \mathrm{P}_X \gamma_0 + \lambda \cdot \mathrm{P}_X \gamma_1 = (1-\lambda)\, \mu_0 + \lambda\, \mu_1 = \tilde{\mu}.$$

  Here we used linearity of P_X. Column sums follow analogously.

- Consequently:

$$\mathcal{C}(\tilde{\mu}, \tilde{\nu}) \leq \langle c, \tilde{\gamma} \rangle = (1-\lambda)\ \langle c, \gamma_0 \rangle + \lambda\ \langle c, \gamma_1 \rangle = (1-\lambda) \cdot \mathcal{C}(\mu_0, \nu_0) + \lambda \cdot \mathcal{C}(\mu_1, \nu_1). \qquad \square$$

Remark 3.5 (Consequences)**.** Important for subsequent results. Convex functions are 'almost differentiable' (sub-differentiable, more details later) and have only global minimizers. $\mathcal{C}(\mu, \nu)$ can be used as building block in more complicated problems, and by convexity we still have a chance to solve them numerically.

3.2 Dual problem

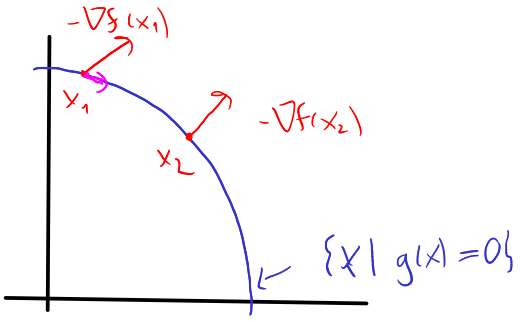
Heuristic derivation of the dual problem.
Step 1: Lagrangian (mini-recap).

- Assume we want to solve

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{subject to} \qquad g(x) = 0$$

  for some (differentiable) $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$.

- Necessary optimality condition for minimizer x: $\nabla f(x)$ orthogonal to feasible set $\{x' \in \mathbb{R}^n | g(x') = 0\}$ at x.



- Normal vector given by $\nabla g(x)$, so $\nabla f(x) = \lambda \cdot \nabla g(x)$ for some $\lambda \in \mathbb{R}$. λ is called Lagrange multiplier.
- Introduce Lagrangian $L(x, \lambda) := f(x) + \lambda \cdot g(x)$. Then necessary optimality condition: $\nabla_x L(x, \lambda) = 0$ for some $\lambda \in \mathbb{R}$. (Will not hold in our case, since our f is not differentiable.)
- In addition, at minimizer have $0 = g(x) = \nabla_\lambda L(x, \lambda)$.
- Alternative interpretation: $\lambda \cdot g(x)$ is a penalty term for the constraint. Write constrained minimization as
$$\inf_{x \in \mathbb{R}^n} \sup_{\lambda \in \mathbb{R}} f(x) + \lambda \cdot g(x)$$
Intuitively, whenever $g(x) \neq 0$, the 'inner' supremum will yield $+\infty$, so the 'outer' infimum must obey the constraint.
- In general have $\inf_x \sup_\lambda L(x, \lambda) \geq \sup_\lambda \inf_x L(x, \lambda)$.
- In special cases: have equality, will lead to dual problem. (More details later.)
- Lagrangian for Kantorovich problem: Have $M + N$ equality constraints: row and column sums. So need $M + N$ Lagrange multipliers. Call them $\alpha \in \mathbb{R}^M$ and $\beta \in \mathbb{R}^N$. Lagrangian given by
$$L(\gamma, \alpha, \beta) := \langle c, \gamma \rangle + \langle \alpha, \mu - \mathrm{P}_X \gamma \rangle + \langle \beta, \nu - \mathrm{P}_Y \gamma \rangle$$
(We will handle the non-negativity constraint of γ separately (which is why we cannot merely consider the derivative $\nabla_\gamma L$.)

Step 2: adjoint operators.

- Recall: P_X is linear operator $\mathbb{R}^{M \times N} \to \mathbb{R}^M$. $\Rightarrow$ there will be an adjoint operator P_X^* from $\mathbb{R}^M$ to $\mathbb{R}^{M \times N}$ such that
$$\langle \alpha, \mathrm{P}_X \gamma \rangle = \langle \mathrm{P}_X^* \alpha, \gamma \rangle$$
for all $\alpha \in \mathbb{R}^M$, $\gamma \in \mathbb{R}^{M \times N}$ where from now on $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in any finite-dimensional Euclidean vectors space.
- Usually, if P_X is a matrix, the adjoint is simply given by transpose matrix. But we want to avoid serialization. (Notion of adjoint also generalizes naturally to infinite-dimensional setting.)
- Determine explicit expression for adjoint:
$$\langle \mathrm{P}_X^* \alpha, \gamma \rangle = \langle \alpha, \mathrm{P}_X \gamma \rangle = \sum_{i=1}^{M} \alpha_i \cdot (\mathrm{P}_X \gamma)_i = \sum_{i=1}^{M} \sum_{j=1}^{N} \alpha_i \cdot \gamma_{i,j} = \sum_{i=1}^{M} \sum_{j=1}^{N} (\mathrm{P}_X^* \alpha)_{i,j} \cdot \gamma_{i,j}$$
so $(\mathrm{P}_X^* \alpha)_{i,j} = \alpha_i$.
- Likewise: $(\mathrm{P}_Y^* \beta)_{i,j} = \beta_j$.
- Now re-write Lagrangian:
$$\begin{aligned} L(\gamma, \alpha, \beta) &:= \langle c, \gamma \rangle + \langle \alpha, \mu - \mathrm{P}_X \gamma \rangle + \langle \beta, \nu - \mathrm{P}_Y \gamma \rangle \\ &= \langle c, \gamma \rangle + \langle \alpha, \mu \rangle - \langle \mathrm{P}_X^* \alpha, \gamma \rangle + \langle \beta, \nu \rangle - \langle \mathrm{P}_Y^* \beta, \gamma \rangle \end{aligned}$$

Step 3: minimax theorem.

- Constrained formulation with Lagrangian:

$$\mathcal{C}(\mu,\nu) = \inf_{\gamma\in\mathbb{R}_+^{M\times N}} \sup_{\alpha,\beta\in\mathbb{R}^M,\mathbb{R}^N} \langle\alpha,\mu\rangle + \langle\beta,\nu\rangle + \langle c - \mathrm{P}_X^*\alpha - \mathrm{P}_Y^*\beta, \gamma\rangle$$

- Now assume/pretend that we may flip order of infimum and supremum. This is usually allowed by using so-called minimax theorems. Here it can be deduced rigorously from duality of finite-dimensional linear programs. We will later provide another argument. With this get:

$$\mathcal{C}(\mu,\nu) = \sup_{\alpha,\beta\in\mathbb{R}^M,\mathbb{R}^N} \langle\alpha,\mu\rangle + \langle\beta,\nu\rangle + \left[\inf_{\gamma\in\mathbb{R}_+^{M\times N}} \langle c - \mathrm{P}_X^*\alpha - \mathrm{P}_Y^*\beta, \gamma\rangle\right]$$

  This looks like a Lagrangian where α, β are the variables and γ is the Lagrange multiplier.

- Since $\gamma \geq 0$, we will find it is a multiplier for inequality constraints: If

$$(c - \mathrm{P}_X^*\alpha - \mathrm{P}_Y^*\beta)_{i,j} = c_{i,j} - \alpha_i - \beta_j < 0$$

  for some i, j, then by sending $\gamma_{i,j} \to \infty$ we can send the inner infimum to $-\infty$.

- If $(\ldots)_{i,j} \geq 0$, then $\gamma_{i,j} \to 0$. Summarize:

$$\inf_{\gamma\in\mathbb{R}_+^{M\times N}} \langle c - \mathrm{P}_X^*\alpha - \mathrm{P}_Y^*\beta, \gamma\rangle = \begin{cases} 0 & \text{if } c - \mathrm{P}_X^*\alpha - \mathrm{P}_Y^*\beta \geq 0, \\ -\infty & \text{else.} \end{cases}$$

- Hence, for the outer supremum, the infimum acts as a constraint. Can now state dual problem.

Proposition 3.6 (Dual Kantorovich problem)**.**

$$\mathcal{C}(\mu,\nu) = \sup\Big\{ \langle\alpha,\mu\rangle + \langle\beta,\nu\rangle \,\Big|\, \alpha\in\mathbb{R}^M,\ \beta\in\mathbb{R}^N, \\ \alpha_i + \beta_j \leq c_{i,j} \text{ for all } i\in\{1,\ldots,M\},\ j\in\{1,\ldots,N\}\Big\}$$

Proof follows rigorously from duality for finite-dimensional linear programs (or with methods introduced later in lecture).

Corollary 3.7.

- Let $\gamma \in \Gamma(\mu,\nu)$, $(\alpha,\beta) \in (\mathbb{R}^M,\mathbb{R}^N)$, $\mathrm{P}_X^*\alpha + \mathrm{P}_Y^*\beta \leq c$. Then have:

$$\langle c,\gamma\rangle \geq \mathcal{C}(\mu,\nu) \geq \langle\alpha,\mu\rangle + \langle\beta,\nu\rangle$$

  with equality if and only if γ and (α,β) are primal-dual optimal.

- $\langle c,\gamma\rangle - \langle\alpha,\mu\rangle - \langle\beta,\nu\rangle$ is called the primal-dual gap and it is a bound on the sub-optimality of γ and (α,β).

Proposition 3.8. Let $\gamma \in \Gamma(\mu,\nu)$, (α,β) dual feasible (i.e. $\mathrm{P}_X^*\alpha + \mathrm{P}_Y^*\beta \le c$). Then they are primal and dual optimal if and only if for all i,j one has

$$[\gamma_{i,j} > 0] \qquad \Rightarrow \qquad [\alpha_i + \beta_j = c_{i,j}].$$

Proof. • [γ, (α,β) both optimal] $\Leftrightarrow$ [primal-dual gap is zero] $\Leftrightarrow$

$$0 = \langle c,\gamma\rangle - \langle \alpha,\mu\rangle - \langle \beta,\nu\rangle = \langle c,\gamma\rangle - \langle \alpha,\mathrm{P}_X\gamma\rangle - \langle \beta,\mathrm{P}_Y\gamma\rangle = \langle \underbrace{c - \mathrm{P}_X^*\alpha - \mathrm{P}_Y^*\beta}_{\ge 0},\gamma\rangle$$

where the last expression is therefore zero if and only if $c_{i,j} = \alpha_i + \beta_j$ for all i,j where $\gamma_{i,j} > 0$. □

Looking at dual problem, note that $\mu \ge 0$, $\nu \ge 0$. If we fix some α and only try to maximize over β, then we want to make each entry as large as the constraint allows, and likewise for fixed β and maximizing over α. Introduce notation for this 'partial maximization'.

Definition 3.9 (c-transform)**.** For $\alpha \in \mathbb{R}^M$ and $\beta \in \mathbb{R}^N$ introduce $\alpha^c \in \mathbb{R}^N$ and $\beta^{\overline{c}}$ via

$$\alpha_j^c := \min_i c_{i,j} - \alpha_i, \qquad\qquad \beta_i^{\overline{c}} := \min_j c_{i,j} - \beta_j.$$

We call these the c-transforms of α and β (added overline in second one, since they are not strictly identical).

Remark 3.10. Now let us establish the existence of dual maximizers. This is not entirely trivial, since the dual feasible set is unbounded. Indeed, let (α,β) be dual feasible. Then for any $\lambda \in \mathbb{R}$ have that

$$(\alpha_i + \lambda) + (\beta_j - \lambda) = \alpha_i + \beta_j \le c_{i,j}$$

and thus $(\alpha+\lambda, \beta-\lambda)$ is also dual feasible, and has the same dual objective:

$$\langle \alpha+\lambda,\mu\rangle + \langle \beta-\lambda,\nu\rangle = \langle \alpha,\mu\rangle + \langle \beta,\nu\rangle + \lambda \cdot \underbrace{\left[\sum_{i=1}^M \mu_i - \sum_{j=1}^N \nu_j\right]}_{=0}$$

We therefore need to show that we can restrict maximization to a bounded (and thus compact) subset of the dual feasible set. This can be done with the c-transform.

Lemma 3.11. Let $\beta \in \mathbb{R}^N$, $\alpha := \beta^{\overline{c}}$. Then

$$\max_i \alpha_i - \min_i \alpha_i \le 2\|c\|_\infty$$

where $\|c\|_\infty := \max_{i,j} |c_{i,j}|$.

Proof. • Introduce $B := \max_j \beta_j$.

- Then for all i:

$$\alpha_i = \min_j \underbrace{c_{i,j}}_{\ge -\|c\|_\infty} - \underbrace{\beta_j}_{\le B} \ge -\|c\|_\infty - B$$

- Likewise, by picking some j' such that $B = \beta_{j'}$ have

$$\alpha_i = \min_j c_{i,j} - \beta_j \leq c_{i,j'} - \beta_{j'} \leq \|c\|_\infty - B$$

- These two bounds now obviously also hold for the maximal and minimal value of α. Together get:

$$\underbrace{\max_i \alpha_i}_{\leq \|c\|_\infty - B} - \underbrace{\min_i \alpha_i}_{\geq -\|c\|_\infty - B} \leq 2\|c\|_\infty$$

□

Proposition 3.12. Dual maximizers exist.

Proof.

- Recall dual problem:

$$\mathcal{C}(\mu,\nu) = \sup\Big\{ \langle \alpha, \mu\rangle + \langle \beta, \nu\rangle \,\Big|\, \alpha \in \mathbb{R}^M,\, \beta \in \mathbb{R}^M, \\ \alpha_i + \beta_j \leq c_{i,j} \text{ for all } i \in \{1,\dots,M\},\, j \in \{1,\dots,N\}\Big\}$$

- For every feasible candidate have that $\alpha^c \geq \beta$ and $\beta^{\overline{c}} \geq \alpha$:

$$\alpha_j^c = \min_i \underbrace{c_{i,j} - \alpha_i}_{\geq \beta_j}$$

- So replacing β by α^c, and then α by $\beta^{\overline{c}} = (\alpha^c)^{\overline{c}}$ will not give a worse score.
- So we can constrain maximization to variables that can be written as c-transforms.
- By previous Lemma can confine maximization over α to those that satisfy

$$\max_i \alpha_i - \min_i \alpha_i \leq 2\|c\|_\infty.$$

- As discussed in Remark 3.10 we can add a constant $\lambda \in \mathbb{R}$ to α, while subtracting it from β and still have dual candidates, with the same score. Thus we can impose the additional constraint $\min_i \alpha_i = 0$. And so entries of α must lie between 0 and $2\|c\|_\infty$.
- Simple similar argument: β can be constrained to lie between $\|c\|_\infty$ and $-3\|c\|_\infty$.
- A feasible candidate is given by $\alpha_i = 0$, $\beta_j = -\|c\|_\infty$. Thus the feasible set is non-empty.
- Objective $(\alpha, \beta) \mapsto \langle \alpha, \mu\rangle + \langle \beta, \nu\rangle$ is continuous.
- Maximization of continuous function over compact, non-empty set has a maximizer. □

Proposition 3.13 (Limits of optimal primal and dual solutions are optimal)**.**

- Let $(\mu_n)_n$ and $(\nu_n)_n$ be sequences in Σ_M and Σ_N with limits $\mu \in \Sigma_M$ and $\nu \in \Sigma_N$.
- Let $(\gamma_n)_n$ be a sequence of corresponding primal minimizers and let $(\alpha_n, \beta_n)_n$ be a sequence of dual maximizers.

- Then there is a subsequence $(n_k)_k$ such that

$$\lim_{k\to\infty} \gamma_{n_k} = \gamma, \qquad \lim_{k\to\infty} \alpha_{n_k} = \alpha, \qquad \lim_{k\to\infty} \beta_{n_k} = \beta$$

  for suitable limits $\gamma \in \mathbb{R}^{M\times N}$, $\alpha \in \mathbb{R}^M$ and $\beta \in \mathbb{R}^N$ which are primal and dual optimizers of the limit problems between the limits μ and ν.

Remark 3.14. Two main motivations:

- First: stability. If the marginals μ and ν change a little bit, it is possible to also only change γ a little bit to keep it optimal.
- Second: numerical approximation. Assume M and N are very large. We can approximate the original μ and ν by increasingly more non-zero entries (every zero entry simplifies numerical solution of the problem, because we can remove the row or column from γ) and get increasingly better approximations of an optimal γ.

Proof of Proposition 3.13.

- The sequence of optimal $(\gamma_n)_n$ is bounded (all entries lie between 0 and 1).
- Arguing as in the proof of Proposition 3.12 we can shift the sequence of dual maximizers until all their entries lie in $[-3\|c\|_\infty, 2\|c\|_\infty]$. Hence the sequence of dual maximizers is also bounded.
- By Bolzano–Weierstrass there exists a subsequence $(n_k)_k$ and limits $\gamma \in \mathbb{R}^{M\times N}$, $\alpha \in \mathbb{R}^M$ and $\beta \in \mathbb{R}^N$ such that each subsequence converges to the respective limit.
- Now let us see that the limits are still feasible for the primal and dual problems.
- The set $\mathbb{R}_+^{M\times N}$ is closed. Hence $\gamma_n \in \mathbb{R}_+^{M\times N}$ for all n implies that also the limit γ lies in this set.
- The operator P_X is linear and thus continuous (in finite dimensions). Therefore:

$$\mathrm{P}_X\gamma = \mathrm{P}_X(\lim_k \gamma_{n_k}) = \lim_k \mathrm{P}_X\gamma_{n_k} = \lim_k \mu_{n_k} = \mu$$

  The same argument applies for the second marginal. Hence $\gamma \in \Gamma(\mu,\nu)$.
- The Euclidean inner product is continuous. Hence $\langle c, \gamma_{n_k}\rangle \to \langle c, \gamma\rangle$.
- Therefore:

$$\mathcal{C}(\mu,\nu) = \inf_{\gamma'\in\Gamma(\mu,\nu)} \langle c, \gamma'\rangle \le \langle c, \gamma\rangle = \lim_k \langle c, \gamma_{n_k}\rangle = \lim_k \mathcal{C}(\mu_{n_k}, \nu_{n_k})$$

- Now, dual feasibility: The set $S = \{\psi \in \mathbb{R}^{M\times N} \mid \psi \le c\}$ is closed.
- We have that the sequence $\psi_n := \mathrm{P}_X^*\alpha_n + \mathrm{P}_Y^*\beta_n$ lies in S.
- By continuity of the operators P_X^* and P_Y^*, we have

$$\psi := \mathrm{P}_X^*\alpha + \mathrm{P}_Y^*\beta = \lim_k \mathrm{P}_X^*\alpha_{n_k} + \mathrm{P}_Y^*\beta_{n_k} = \lim_k \psi_{n_k}$$

  and thus $\psi \in S$. So α and β are dual feasible.

- Again, the dual objective $(\alpha, \beta, \mu, \nu) \mapsto \langle \alpha, \mu \rangle + \langle \beta, \nu \rangle$ is continuous (simultaneously in dual variables and marginals). So:

$$\begin{aligned} \mathcal{C}(\mu, \nu) &= \sup_{\substack{(\alpha', \beta'): \\ \text{dual feasible}}} \langle \alpha', \mu \rangle + \langle \beta', \nu \rangle \geq \langle \alpha, \mu \rangle + \langle \beta, \nu \rangle \\ &= \lim_k \langle \alpha_{n_k}, \mu_{n_k} \rangle + \langle \beta_{n_k}, \nu_{n_k} \rangle = \lim_k C(\mu_{n_k}, \nu_{n_k}). \end{aligned}$$

- Combining upper and lower bound we get $\mathcal{C}(\mu, \nu) = \lim_k C(\mu_{n_k}, \nu_{n_k})$ and thus γ and (α, β) must be primal and dual optimal. □

Remark 3.15. The proposition can easily be adapted to account for a sequence of cost functions $(c_n)_n$ with $c_n \to c$ for some limit cost $c \in \mathbb{R}^{M \times N}$.

4 Mini-introduction: Measures and weak convergence

Non-negative measures.

- Piles of sand: mass is distributed over continuum, $\mu(x)$: density of mass at point x, mass in region $A \subset X$ given by $\int_A \mu(x)\,\mathrm{d}x$. The mass located at any single point x is zero.
- Bakeries and cafes: mass is concentrated on a discrete set of points, μ_i is mass at single point x_i, density would be $+\infty$.
- Both can also be seen as limits of each other:
  - If the pile of sand is very high and concentrated on a small region, if we look at it on a map, it may seem as a single point.
  - Conversely, if we look at the large number of cafes in Paris, it may be impractical to consider each individually and therefore compute an approximate 'cafe-density' or 'bread-density' over each block of buildings.
- Both concepts can be described in a mathematically unified way by means of measures.
- A (non-negative) measure on X is a function from subsets of X to $\mathbb{R} \cup \{\infty\}$ that satisfies certain axioms which are consistent with the notion of mass or volume:
  - non-negative
  - $\mu(\emptyset) = 0$
  - σ-additivity: $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ for a sequence of pairwise disjoint sets $(A_i)_i$.

  The fact that (countable) infinite sequences of sets are allowed in the last part is important for technical reasons.
- Denote by $\mathcal{M}_+(X)$ the (non-negative) measures over X, the probability measures by $\mathcal{P}(X) := \{\mu \in \mathcal{M}_+(X) \,|\, \mu(X) = 1\}$.

Measurable sets. Another technical aspect: measures are not necessarily defined for all subsets $A \subset X$. By choosing very sophisticated, evil sets, one can arrive at very counter-intuitive situations such as the Banach–Tarski paradox which showed that a sphere in three dimensions can be decomposed into a finite number of sets, which can then be shifted to form two spheres. So the intuitive notion of volume is violated. To allow for such an intuitive notion of volume, measures can only be defined on a sub-family of sets, which are called 'measurable' sets. The set of measurable sets must of course be closed under standard operations such as intersection, union, taking the closure or interior, et cetera. Sets that are not measurable for the Lebesgue measure cannot be constructed explicitly. Their existence can only be shown via the infamous axiom of choice. Measurability will not be an issue for us.

Example 4.1 (Some examples)**.**

- Lebesgue measure in 1d, denoted by $\mathcal{L}$: assigns $b - a$ to intervals $[a, b]$ where $a \leq b$, this fully characterizes the measure on all measurable sets. Generalization to higher dimensions works via the assignment of volumes to cuboids.

- Scaled Lebesgue measure: if f is a (sufficiently regular) function $\mathbb{R} \to \mathbb{R}_+$, then one can use it to re-scale the Lebesgue measure by the rule

$$(f \cdot \mathcal{L})(A) := \int_A f(x)\,\mathrm{d}\mathcal{L}(x)$$

for measurable $A \subset \mathbb{R}$. In the Monge-example, f would be the height of the sand pile at each point. Prominent example: Gaussian distribution with mean z and standard deviation σ has density

$$f(x) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\tfrac{(x-z)^2}{2\sigma^2}\right).$$

- Dirac measure: for $x \in X$, the Dirac measure over X at point x is defined as

$$\delta_x(A) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

- Measures can be combined by summing and scaling. The distribution of bread in bakeries could be written as $\mu := \sum_{i=1}^{M} \mu_i \cdot \delta_{x_i}$, where now μ denotes the whole measure and $(\mu_i)_i$ are the mass-coefficients for the individual locations. We can also combine (scaled) Lebesgue measure with Dirac measures, et cetera.

Signed measures.

- Measures can also assign negative values (or even vector values), as long as we can rule out inconsistencies that could arise by adding $+\infty$ and $-\infty$ in the additivity-rule. In the easiest case, we simply demand that a signed measure is the sum (difference) of two non-negative measures with bounded total mass.
- Example: charge density in physics.
- Denote set of signed measures by $\mathcal{M}(X)$
- We can assign a norm to (signed) measures by summing up the (absolute values of the) total positive mass and the total negative mass. Denote the norm by $\|\cdot\|_{\mathcal{M}}$. This yields a Banach space.

Integration.

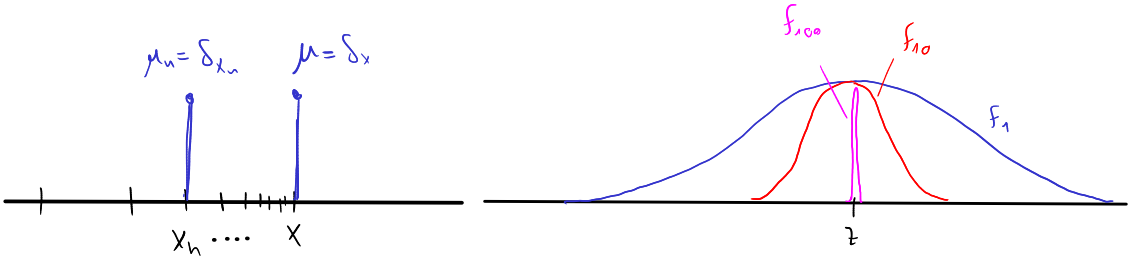
- For a given measure, we can integrate functions against it, e.g. write $\int_X f(x)\,\mathrm{d}\mu(x)$
- If $\mu = \delta_z$, then $\int_X f(x)\,\mathrm{d}\mu(x) = f(z)$. So if we change f in a single point, the value of the integral changes.
- Conversely, if $\mu = \mathcal{L}$, if we change f in a single point, the integral does not change.

Measures and continuous functions.

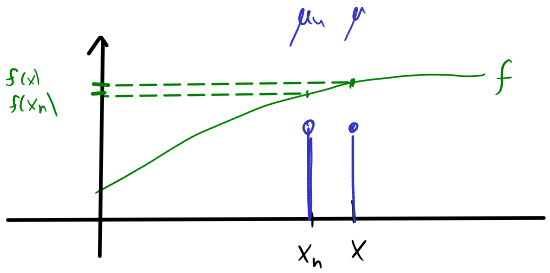
- Let X be a compact (closed, bounded) subset of $\mathbb{R}^d$. Then the set of continuous functions $X \to \mathbb{R}$, denoted by $C(X)$, is a vector space. It is a Banach space when equipped with the norm
$$\|f\|_\infty := \sup_{x \in X} |f(x)| \qquad \text{for} \qquad f \in C(X).$$
- A finite measure $\mu \in \mathcal{M}(X)$ induces a linear map from continuous functions to $\mathbb{R}$ by $C(X) \ni f \mapsto \int_X f(x)\,\mathrm{d}\mu(x)$.
- This map is *bounded* in the following sense: $\left|\int_X f(x)\,\mathrm{d}\mu(x)\right| \leq \|f\|_\infty \cdot \|\mu\|_{\mathcal{M}}$. So if the function goes to zero, so will the integral (with a uniform bound on the rate).
- One can show: *any* bounded linear map from $C(X)$ to $\mathbb{R}$ can be expressed as integration against some measure in $\mathcal{M}(X)$ (Riesz representation theorem).
- This means: two measures $\mu, \nu \in \mathcal{M}(X)$ are identical if and only if their integral against all functions in $C(X)$ is the same, i.e.
$$\int_X f\,\mathrm{d}\mu = \int_X f\,\mathrm{d}\nu \quad \text{for all} \quad f \in C(X).$$
- Comparing measures by integration against continuous 'test functions' is sometimes more convenient than comparing their values on all measurable sets.

Weak* convergence.

- Now consider sequence of measures $(\mu_n)_n$ with $\mu_n := \delta_{x_n}$ where sequence of points $(x_n)_n$ converges to some x (but $x_n \neq x$ for all n). Intuitively, we see that μ_n converges to $\mu := \delta_x$ in some sense. The mass moves to the right limit location, even though it never really reaches it.
- But: $\mu_n - \mu = \delta_{x_n} - \delta_x$ and so $\|\mu_n - \mu\|_{\mathcal{M}} = 2$. So the convergence is not in the norm.



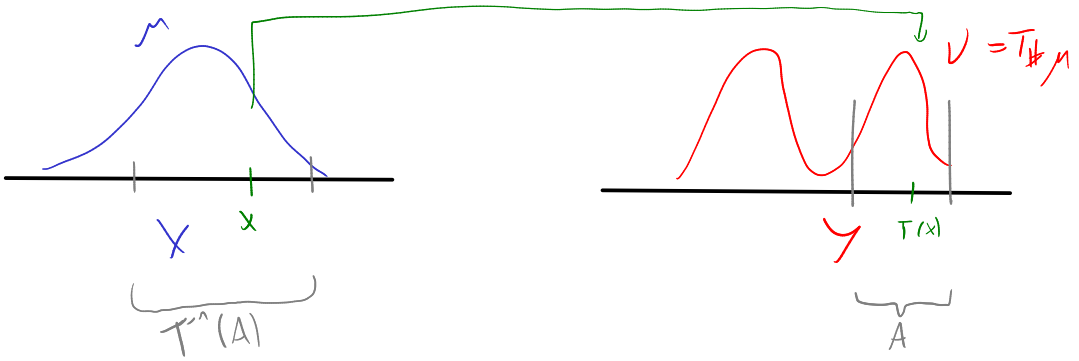
- Similar example: let $\mu_n := f_n \cdot \mathcal{L}$ with $f_n(x) := \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(x-z)^2}{2\sigma_n^2}\right)$ where $(\sigma_n)_n$ is a sequence of variances with $\sigma_n > 0$, $\sigma_n \to 0$ as $n \to \infty$. Then intuitively, μ_n converges to $\mu := \delta_z$ in some sense, but not in norm since $\|\mu_n - \delta_z\|_{\mathcal{M}} = 2$ for all n. The Gaussians become increasingly concentrated, in the limit all the mass will sit at z.
- In both examples we find: for any continuous function $f \in C(X)$ (X bounded, closed subset of $\mathbb{R}^d$) one has
$$\lim_{n \to \infty} \int_X f\,\mathrm{d}\mu_n = \int_X f\,\mathrm{d}\mu.$$



- This is a suitable notion of convergence for our (and most numerical / engineering) purposes. We say the sequence $\mu_n \in \mathcal{M}(X)$ converges weak* to μ if the integrals converge for each $f \in C(X)$.

Push-forward.

- In Monge example we discussed the transformation that a map $T : X \to Y$ can induce on a measure $\mu \in \mathcal{M}_+(X)$ by (intuitively) picking up all the small mass atoms of μ at each $x \in X$ and dropping them at $T(x) \in Y$.
- Now we can formalize this definition. The push-forward of μ under T is a measure in $\mathcal{M}_+(Y)$ denoted by $T_\sharp\mu$, which is given by $(T_\sharp\mu)(A) := \mu(T^{-1}(A))$ for all (measurable) $A \subset Y$. Here $T^{-1}(A) := \{x \in X \,|\, T(x) \in A\}$ is the pre-image of A under T.
- Intuitively: $(T_\sharp\mu)(A)$ contains the mass of μ on all points x that are taken to A under T.



- The push-forward can also be characterized by integration with continuous functions:

$$\int_Y f(y)\,\mathrm{d}(T_\sharp\mu)(y) = \int_X f(T(x))\,\mathrm{d}\mu(x) \qquad \text{for all} \qquad f \in C(Y)$$

This is called the change-of-variables formula.

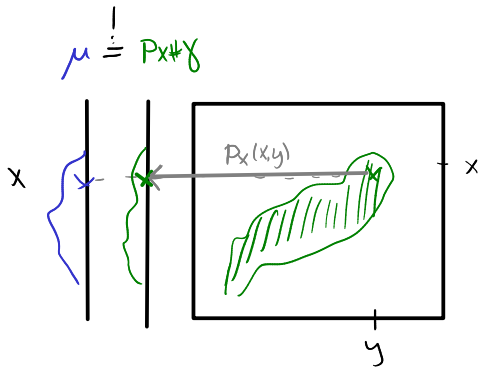
Transport plans as measures and Kantorovich primal.

- Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$. A transport plan will now be a measure $\gamma \in \mathcal{M}_+(X \times Y)$. $\gamma(A \times B)$ describes mass that goes from $A \subset X$ to $B \subset Y$.
- What are row sums? Mass that goes from $A \subset X$ to any point in Y (which is $\gamma(A \times Y)$) has to equal total available mass in A (which is $\mu(A)$). So:

$$\gamma(A \times Y) = \mu(A) \quad \text{for all (measurable)} \quad A \subset X.$$

Same for column sums.

- Let $\mathrm{p}_X : X \times Y \to X$, $(x, y) \mapsto x$. Find: $\mathrm{p}_{X\sharp}\gamma(A) = \gamma(\mathrm{p}_X^{-1}(A)) = \gamma(A \times Y)$. Formally define row sum operator: $\mathrm{P}_X\gamma := \mathrm{p}_{X\sharp}\gamma$. Same for column sums.



- So set of transport plans can be written as

$$\Gamma(\mu, \nu) := \{\gamma \in \mathcal{M}_+(X \times Y) \,|\, \mathrm{P}_X\gamma = \mu,\, \mathrm{P}_Y\gamma = \nu\}\,.$$

- General formulation of Kantorovich primal:

$$\mathcal{C}(\mu, \nu) := \inf\left\{ \int_{X\times Y} c(x, y)\,\mathrm{d}\gamma(x, y) \,\middle|\, \gamma \in \Gamma(\mu, \nu) \right\}$$

Kantorovich dual.

- Dual is given by:

$$\mathcal{C}(\mu, \nu) = \sup\left\{ \int_X \alpha(x)\,\mathrm{d}\mu(x) + \int_Y \beta(y)\,\mathrm{d}\nu(y) \,\middle|\, \alpha \in C(X),\, \beta \in C(Y), \right.$$
$$\left. \alpha(x) + \beta(y) \le c(x, y) \text{ for all } (x, y) \in X \times Y \right\}$$

- Sketch of derivation: Slight generalization of the discrete derivation.

- Equality of measures can be tested by integration against continuous functions. So the Lagrange multipliers for row and column sums become functions in $C(X)$ and $C(Y)$. Lagrangian saddle point problem:

$$\inf_{\gamma\in\mathcal{M}_+(X\times Y)} \sup_{\substack{\alpha\in C(X),\\ \beta\in C(Y)}} \int_{X\times Y} c\,\mathrm{d}\gamma + \left(\int_X \alpha\,\mathrm{d}\mu - \int_X \alpha\,\mathrm{d}(\mathrm{P}_X\gamma)\right) + \left(\int_Y \beta\,\mathrm{d}\nu - \int_Y \beta\,\mathrm{d}(\mathrm{P}_Y\gamma)\right)$$

- Now use notion of adjoint operator: $\mathrm{P}_X : \mathcal{M}(X \times Y) \to \mathcal{M}(X)$, define adjoint $\mathrm{P}_X^* : C(X) \to C(X \times Y)$ by condition

$$\int_{X\times Y} (\mathrm{P}_X^*\alpha)(x, y)\,\mathrm{d}\gamma(x, y) := \int_X \alpha(x)\,\mathrm{d}(\mathrm{P}_X\gamma)(x) = \int_X \alpha(x)\,\mathrm{d}(\mathrm{p}_{X\sharp}\gamma)(x)$$
$$= \int_X \alpha(\mathrm{p}_X(x, y))\,\mathrm{d}\gamma(x, y)$$

So in analogy to discrete case: $(\mathrm{P}_X^*\alpha)(x, y) = \alpha(x)$.

- As before, assume we can flip the order of inf and sup:

$$\sup_{\substack{\alpha\in C(X),\\ \beta\in C(Y)}} \int_X \alpha \,\mathrm{d}\mu + \int_Y \beta \,\mathrm{d}\nu + \inf_{\gamma\in\mathcal{M}_+(X\times Y)} \int_{X\times Y} (c - \mathrm{P}_X^*\alpha - \mathrm{P}_Y^*\beta)\,\mathrm{d}\gamma$$

- Intuitively: inf-term becomes inequality constraint:

$$0 \leq (c - \mathrm{P}_X^*\alpha - \mathrm{P}_Y^*\beta)(x,y) = c(x,y) - \alpha(x) - \beta(y) \quad \text{for all} \quad (x,y) \in X \times Y$$

Generalization. All of the previous statements can be subsumed into the new case by using discrete measures. All of the statements in this lecture can be proved for the measure setting. The arguments are often almost the same, sometimes slightly trickier.

5 Wasserstein spaces

5.1 Wasserstein metric

Definition 5.1 (Metric)**.** A metric on a set X is a function $d : X \times X \to \mathbb{R}_+$ that satisfies:

- $[d(x, y) = 0] \Leftrightarrow [x = y]$ (separates objects)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, y) + d(y, z) \geq d(x, z)$ (triangle inequality)

for all $x, y, z \in X$.

Remark 5.2.

- A metric formalizes the natural notion of distances. Metrics are ubiquitous in mathematics. Norms in vector spaces induce metrics: $d(x, y) = \|x - y\|$.
- In data science a metric can be used to express similarity between samples. To be of practical use, the metric must be well-suited for the problem at hand.
- As shown in Example 2.2 (Problem sheet 2), the L^1-norm $\|\cdot\|_1$ is not a very robust metric for histograms. Optimal transport can induce a more meaningful family of metrics. These are the so-called Wasserstein distances.

Proposition 5.3 (Wasserstein distances)**.** Let $X = \{x_1, \ldots, x_M\}$, $d : X \times X \to \mathbb{R}_+$ a metric on X, $p \in [1, \infty)$. Then for $\mu, \nu \in \mathcal{P}(X) \equiv \Sigma_M$, set

$$W_p(\mu, \nu) := \inf \left\{ \sum_{i,j=1}^{M} d(x_i, x_j)^p \, \gamma_{i,j} \,\middle|\, \gamma \in \Gamma(\mu, \nu) \right\}^{1/p}.$$

The function $W_p : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}_+$ is called the p-Wasserstein distance on $\mathcal{P}(X)$ and it is indeed a metric on $\mathcal{P}(X)$.

Proof that W_p is a metric.
Part 0: $W_p(\mu, \nu) \geq 0$

- In the following denote $c_{i,j} = d(x_i, x_j)^p$.
- This follows directly from $\gamma \geq 0$ and $c_{i,j} \geq 0$. So $\langle c, \gamma \rangle \geq 0$ for all $\gamma \in \Gamma(\mu, \nu)$ and consequently also the infimum / minimum is non-negative.

Part 1: Symmetry $W_p(\mu, \nu) = W_p(\nu, \mu)$

- Let $\gamma \in \Gamma(\mu, \nu)$. We find $\gamma^\top \in \Gamma(\nu, \mu)$ (rows and columns exchanged). And since $c^\top = c$ (since d is symmetric), have $\langle c, \gamma^\top \rangle = \langle c, \gamma \rangle$. So from each $\gamma \in \Gamma(\mu, \nu)$ can construct one in $\Gamma(\nu, \mu)$ with same objective value (and vice versa). So both problems have the same infimal value.

Part 2: Separation $[W_p(\mu, \nu) = 0] \Leftrightarrow [\mu = \nu]$

- $\Leftarrow$: If $\mu = \nu$, then $\gamma_{i,j} = \mu_i \cdot \delta_{i,j}$ is a valid coupling in $\Gamma(\mu, \nu = \mu)$. We find

$$\sum_{i,j=1}^{M} c_{i,j}\, \gamma_{i,j} = \sum_{i,j=1}^{M} c_{i,j}\, \mu_i\, \delta_{i,j} = \sum_{i}^{M} c_{i,i}\, \mu_i = 0$$

- But since $c_{i,j} \geq 0$ we must have $W_p(\mu, \nu) \geq 0$ (which is also a requirement for a metric) and therefore this γ must be minimal and we have $W_p(\mu, \mu) = 0$.

- $\Rightarrow$: Assume $W_p(\mu, \nu) = 0$, let $\gamma \in \Gamma(\mu, \nu)$ be optimal transport plan (for existence see Remark 3.2). Since $c_{i,j} \geq 0$ with $[c_{i,j} = 0] \Leftrightarrow [i = j]$ we must have that $[\gamma_{i,j} > 0] \Rightarrow [i = j]$.

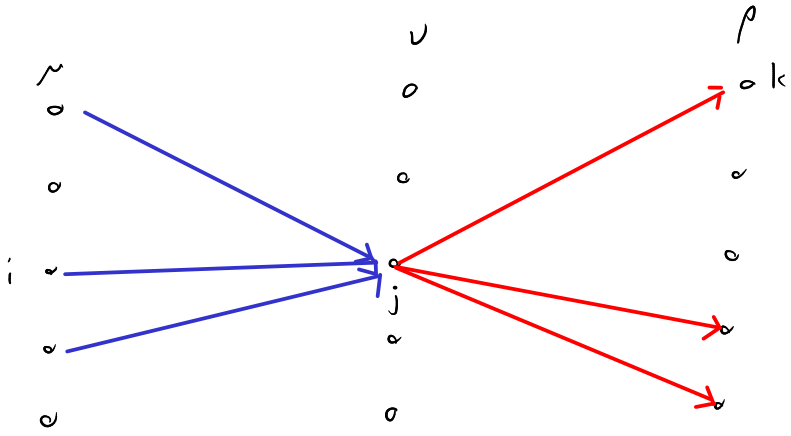
- Therefore, γ can be written as $\gamma_{i,j} = \rho_i \cdot \delta_{i,j}$ for some $\rho \in \mathbb{R}_+^M$ (or even Σ_M). Row and col sum constraints yield $\rho_i = \mu_i = \nu_i$ for $i = 1, \ldots, M$. Therefore, $\mu = \nu$.

Part 3: Triangle inequality

- Let μ, ν, $\rho \in \mathcal{P}(X)$. Let $\gamma \in \Gamma(\mu, \nu)$ and $\lambda \in \Gamma(\nu, \rho)$ be optimal plans for $W_p(\mu, \nu)$ and $W_p(\nu, \rho)$.

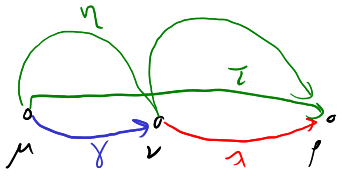
- Now we glue them together: γ describes rearrangement of mass from μ to ν, λ describes rearrangement from ν to ρ. Together, they should be able to describe a rearrangement from μ to ρ.

- We introduce a table $\eta \in \mathbb{R}_+^{M \times M \times M}$ where $\eta_{i,j,k}$ will denote the amount of mass traveling from i via j to k.



- For $j \in \{1, \ldots, M\}$ with $\nu_j > 0$, interpret $\lambda_{j,k}/\nu_j$ as conditional probability that particle arriving at j continues to travel to k. Combine this with influx from i to j, described by γ and set

$$\eta_{i,j,k} := \begin{cases} \frac{\gamma_{i,j}\, \lambda_{j,k}}{\nu_j} & \text{if } \nu_j > 0, \\ 0 & \text{else.} \end{cases}$$



- Quickly verify: $\sum_{i=1}^M \eta_{i,j,k} = \lambda_{j,k}$, $\sum_{k=1}^M \eta_{i,j,k} = \gamma_{i,j}$. Total mass of η equals total mass of γ, λ, equals 1.
- Extract transport plan from μ to ρ by summing over j in $\eta_{i,j,k}$: $\tau_{i,k} := \sum_{j=1}^M \eta_{i,j,k}$. Find $\tau \in \Gamma(\mu, \rho)$. For example:

$$\sum_{k=1}^M \tau_{i,k} = \sum_{j=1}^M \sum_{k=1}^M \eta_{i,j,k} = \sum_{j=1}^M \gamma_{i,j} = \mu_i$$

- Now plug τ into problem for $W_p(\mu, \rho)$ to obtain upper bound:

$$\begin{aligned}
W_p(\mu,\rho) &\le \left(\sum_{i,k=1}^M d(x_i,x_k)^p \, \tau_{i,k} \right)^{1/p} = \left(\sum_{i,j,k=1}^M d(x_i,x_k)^p \, \eta_{i,j,k} \right)^{1/p} \\
&\le \left(\sum_{i,j,k=1}^M \underbrace{d(x_i,x_k)^p}_{\le (d(x_i,x_j)+d(x_j,x_k))^p} \eta_{i,j,k} \right)^{1/p} \\
&\le \left(\sum_{i,j,k=1}^M d(x_i,x_j)^p \, \eta_{i,j,k} \right)^{1/p} + \left(\sum_{i,j,k=1}^M d(x_j,x_k)^p \, \eta_{i,j,k} \right)^{1/p} \\
&= \left(\sum_{i,j=1}^M d(x_i,x_j)^p \, \gamma_{i,j} \right)^{1/p} + \left(\sum_{j,k=1}^M d(x_j,x_k)^p \, \lambda_{j,k} \right)^{1/p} = W_p(\mu,\nu) + W_p(\nu,\rho).
\end{aligned}$$

- Here we have used the Minkowski inequality: For $\mu \in \Sigma_M$, $f, g \in \mathbb{R}^M$ have

$$\left(\sum_{i=1}^M |f_i + g_i|^p \mu_i \right)^{1/p} \le \left(\sum_{i=1}^M |f_i|^p \mu_i \right)^{1/p} + \left(\sum_{i=1}^M |g_i|^p \mu_i \right)^{1/p}. \qquad \square$$

Example 5.4 (Dirac measures)**.**

- Let $\mu := \delta_{x_i}$, a Dirac measure at x_i, or equivalently $\mu_k := \delta_{i,k}$ for $k = 1, \dots, M$ with Kronecker-delta notation. And $\nu := \delta_{x_j}$.
- Then a quick computation yields $\Gamma(\mu, \nu) = \{\delta_{(x_i, x_j)}\}$. There is only a single transport plan. All mass from x_i must go to x_j.
- Therefore: $W_p(\delta_{x_i}, \delta_{x_j}) = d(x_i, x_j)$.
- So the Wasserstein metric, restricted to Dirac measures, equals the original metric over single points. The Wasserstein metric can be seen as a lifting / extension of the original metric from Dirac measures to general probability measures.

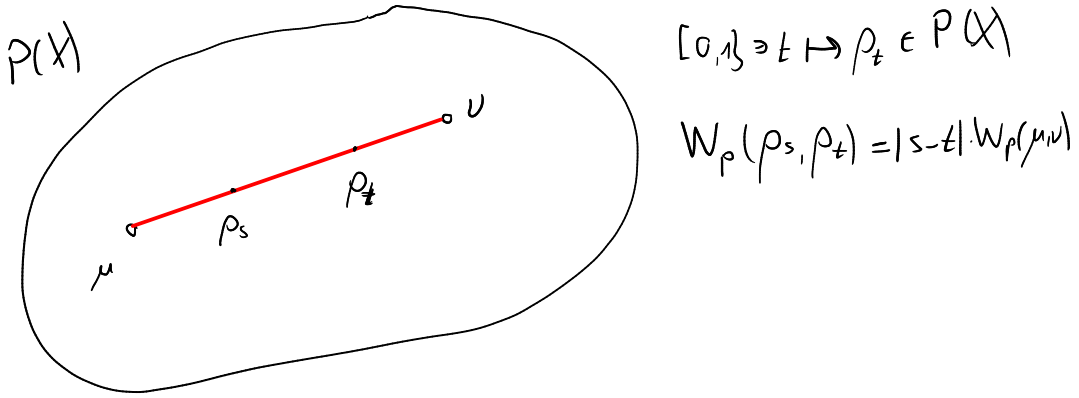
Example 5.5 (Rearranging a bookshelf)**.**

Remark 5.6 (Wasserstein distances metrize weak* convergence)**.** On a compact metric space (X, d) a sequence of probability measures $(\mu_n)_n$ converges to some limit μ if and only if $W_p(\mu_n, \mu) \to 0$ for $n \to \infty$. Slightly more attention must be paid on non-compact spaces.

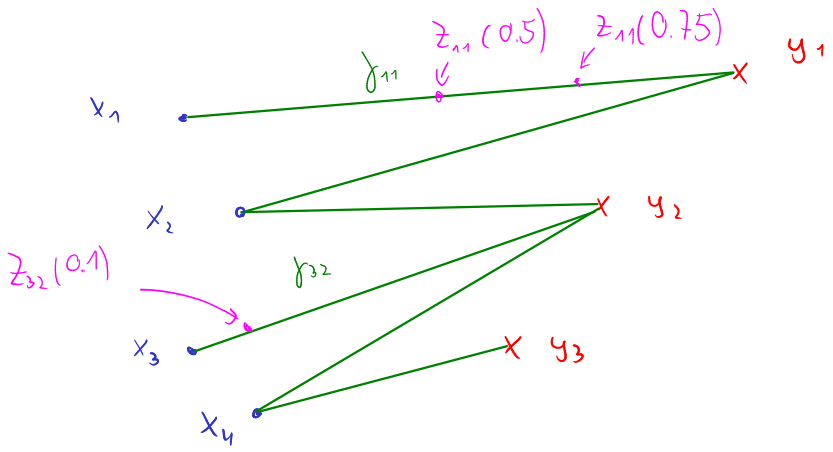
5.2 Displacement interpolation

Remark 5.7.

- More generally as in Section 5.1, Wasserstein distances can be defined for general probability measures on (bounded) subsets $X \subset \mathbb{R}^d$.
- If X is convex, then there are even shortest paths. This means, for two probabilities $\mu_0, \mu_1 \in \mathcal{P}(X)$, there is a curve $[0,1] \ni t \mapsto \mu(t)$ with $\mu(0) = \mu_0$, $\mu(1) = \mu_1$ that satisfies
$$W_p(\mu(s), \mu(t)) = |s-t| \cdot W_p(\mu_0, \mu_1) \qquad \text{for} \qquad s,t \in [0,1].$$
- This is essentially brand new mathematics: these curves were introduced in the 1990s.[1]



Proposition 5.8. Let $X \subset \mathbb{R}^d$ be closed, bounded, and convex. Let $\mu = \sum_{i=1}^M \mu_i \cdot \delta_{x_i}$, $\nu = \sum_{j=1}^N \nu_j \cdot \delta_{y_j}$ be two discrete probability measures on X and let $\gamma \in \Gamma(\mu,\nu) \subset \mathbb{R}_+^{M \times N}$ be an optimal transport plan between μ and ν for the p-Wasserstein distance problem for some $p \in [1,\infty)$. For $t \in [0,1]$ set
$$\rho(t) := \sum_{i=1}^M \sum_{j=1}^N \gamma_{i,j} \cdot \delta_{z_{i,j}(t)} \quad \text{with} \quad z_{i,j}(t) := (1-t) \cdot x_i + t \cdot y_j.$$
Then $[0,1] \ni t \mapsto \rho(t)$ is a shortest curve between μ and ν in $\mathcal{P}(X)$ with respect to the p-Wasserstein distance.



[1] R. McCann: A Convexity Principle for Interacting Gases, Advances in Mathematics 128, pages 153–179, 1997

Proof.

- $\rho(t)$ can be interpreted as (non-negative) vector in $\mathbb{R}_+^{M\times N}$. Transport plans between μ and $\rho(t)$ will live in $\mathbb{R}_+^{M\times(M\times N)}$ where $\hat{\gamma}_{i,(j,k)}$ denotes the mass that is sent from x_i to $z_{j,k}(t)$.
- Claim: A feasible transport plan $\hat{\gamma} \in \Gamma(\mu, \rho(t))$ is given by $\hat{\gamma}_{i,(j,k)} := \delta_{i,j} \cdot \gamma_{i,k}$.
- Column sums:

$$\sum_{j=1}^{M}\sum_{k=1}^{N} \hat{\gamma}_{i,(j,k)} = \sum_{j=1}^{M}\sum_{k=1}^{N} \delta_{i,j} \cdot \gamma_{i,k} = \sum_{k=1}^{N} \gamma_{i,k} = \mu_i$$

- Row sums:

$$\sum_{i=1}^{M} \hat{\gamma}_{i,(j,k)} = \sum_{i=1}^{M} \delta_{i,j} \cdot \gamma_{i,k} = \gamma_{j,k}$$

- What cost is associated with $\hat{\gamma}$? The corresponding cost is given by

$$\hat{c}_{i,(j,k)} := \|x_i - z_{j,k}(t)\|^p = \|x_i - (1-t)\, x_j + t\, y_k\|^p$$

  Note that $\hat{\gamma}_{i,(j,k)} > 0 \Rightarrow i = j$. So it is sufficient to look at the corresponding entries of the cost only. We find:

$$\hat{c}_{i,(i,k)} = t^p \cdot \|x_i - y_k\|^p.$$

- Now plug $\hat{\gamma}$ into objective for $W_p(\mu, \rho(t))$. This gives an upper bound:

$$\begin{aligned} W_p(\mu, \rho(t)) &\leq \left(\sum_{i=1}^{M}\sum_{k=1}^{N} \hat{c}_{i,(i,k)} \cdot \hat{\gamma}_{i,(i,k)} \right)^{1/p} = \left(\sum_{i=1}^{M}\sum_{k=1}^{N} t^p \cdot \|x_i - y_k\|^p \cdot \gamma_{i,k} \right)^{1/p} \\ &= t \cdot W_p(\mu, \nu) \end{aligned}$$

- By same argument show $W_p(\rho(t), \nu) \leq (1-t) \cdot W_p(\mu, \nu)$.
- Now use triangle inequality:

$$W_p(\mu, \nu) \leq W_p(\mu, \rho(t)) + W_p(\rho(t), \nu) \leq [t + (1-t)] \cdot W_p(\mu, \nu)$$

  So both inequalities for $W_p(\mu, \rho(t))$ and $W_p(\rho(t), \nu)$ must be equalities. So the coupling $\hat{\gamma}$ constructed above must be optimal.

- Could now re-use same argument to show that $\hat{\gamma}_{(i,j),(k,l)} := \delta_{i,k} \cdot \delta_{j,l} \cdot \gamma_{i,j}$ is optimal between $\rho(s)$ and $\rho(t)$ for $s, t \in [0,1]$. Finally arrive at $W_p(\rho(s), \rho(t)) = |s-t| \cdot W_p(\mu, \nu)$.

□

6 1-Wasserstein problems on graphs

Remark 6.1 (Motivation)**.**

- Kantorovich optimal transport problem between $\mu \in \Sigma_M$, $\nu \in \Sigma_N$ has $M \cdot N$ primal variables, run-time of standard linear solvers is (empirically) polynomial in number of variables (e.g. cubic). If $M = N \approx 10^6$ (representing a mega-pixel image), then $M \cdot N \approx 10^{12}$ and applying a standard linear solver will be problematic both in terms of run-time and memory.
- If $(X = Y, d)$ is a metric graph (e.g. a grid graph) with $O(M = N)$ edges, we will find that the W_1 problem on (X, d) can be written as 'flow problem' on the graph with $O(M)$ variables and constraints, hence reducing the complexity.
- In some models $p > 1$ (mostly $p = 2$) is more natural (discussion will follow), but in some cases $p = 1$ is exactly what is needed.

6.1 Kantorovich–Rubinstein duality

Definition 6.2 (Metric space and 1-Lipschitz functions)**.**

- Let $X = Y = \{x_1, \dots, x_M\}$ be a metric space with metric $d : X \times X \to \mathbb{R}_+$.
- A function $\alpha : X \to \mathbb{R}$ is 1-Lipschitz if for any $x, y \in X$ one has $|\alpha(x) - \alpha(y)| \leq d(x, y)$.
- Denote the set of 1-Lipschitz functions over X by $\mathrm{Lip}_1(X)$.

Remark 6.3 (c-transform for metric cost)**.** Set $c = d$. Recall c-transform:

$$\alpha^c(x) := \min_z c(x, z) - \alpha(z)$$

In this case do not need 'reverse transform' $\alpha^{\bar{c}}$ since c is symmetric.

Lemma 6.4.

- If $\beta = \alpha^c$ for some $\alpha \in \mathbb{R}^X$, then β is 1-Lipschitz.
- If α is 1-Lipschitz, then $\beta = \alpha^c = -\alpha$.
- Therefore, set of functions on X that can be written as α^c for some $\alpha : X \to \mathbb{R}$ is precisely $\mathrm{Lip}_1(X)$.

Proof.

- **Part 1:** Let $x, z \in X$.

$$\beta(x) = \min_{y \in X} \underbrace{d(x, y)}_{\leq d(x,z)+d(z,y)} - \alpha(y) \leq d(x, z) + \beta(z)$$

- Swap roles of x and z to get $\beta(z) \leq d(x, z) + \beta(x)$. Combine both inequalities to get $|\beta(x) - \beta(z)| \leq d(x, z)$.

- **Part 2:** Let $x \in X$.

$$\beta(x) = \min_{y \in X} d(x,y) - \alpha(y) \leq -\alpha(x) \quad (\text{set } y = x \text{ in min})$$
$$\beta(x) = \min_{y \in X} d(x,y) - \underbrace{\alpha(y)}_{\leq \alpha(x) + d(x,y)} \geq \min_{y \in X} d(x,y) - \alpha(x) - d(x,y) = -\alpha(x)$$

□

Proposition 6.5 (Kantorovich–Rubinstein formula)**.**

$$W_1(\mu,\nu) = \sup\{\langle \varphi, \mu - \nu\rangle | \varphi \in \mathrm{Lip}_1(X)\}$$

Proof.

- general Kantorovich dual:

$$W_1(\mu,\nu) = \sup\left\{\langle \varphi,\mu\rangle + \langle \psi,\nu\rangle | \varphi,\psi \in \mathbb{R}^X, \varphi(x) + \psi(y) \leq d(x,y)\, \forall (x,y) \in X^2\right\}$$

- Recall: can add constraints $\varphi = \psi^c$, $\psi = \varphi^c$. By above Lemma this is equivalent to $\varphi, \psi \in \mathrm{Lip}_1(X)$ and $\psi = -\varphi$. □

Remark 6.6.

- inherits shift invariance from general Kantorovich dual: can add constant to φ and still get feasible candidate with same objective
- also inherit existence from Kantrovich dual

Example 6.7 (Two Diracs)**.**

6.2 Min-cost flow problem

Definition 6.8 (Metric graph)**.**

- vertices X, edge list $E \subset X \times X$, edge lengths $\ell : E \to \mathbb{R}_{++} = (0,\infty)$
- symmetric: $[(x,y) \in E] \Leftrightarrow [(y,x) \in E]$, $\ell(x,y) = \ell(y,x)$.
- path in X is tuple $(x_1,\ldots,x_K)$ with $(x_{i+1},x_i) \in E$ for $i = 1,\ldots,K-1$.
- assume graph is connected, i.e. exists path between any two vertices
- length of path $L(x_1,\ldots,x_K) := \sum_{i=1}^{K-1} \ell(x_{i+1},x_i)$
- graph metric: induced by length of shortest paths

$$d(x,y) := \min\{L(x_1,\ldots,x_K) | (x_1,\ldots,x_K) \text{ path in } (X,E) \text{ with } x_1 = x,\, x_K = y\}$$

(feasible set of paths is non-empty, since graph is connected. minimizer exists, since only finite number of paths exists)

- for $x = y$ we say that $(x_1 = x = y)$ is a path, which contains no edges and hence $L(x_1) = 0$

- d is clearly metric: $d(x,y) = d(y,x)$ by symmetry of E, ℓ (any path can be reversed without changing its length). $d(x,x) = 0$. $d(x,y) > 0$ for $x \neq y$ since $\ell(\ldots) > 0$. Triangle inequality by concatenation of optimal paths.

Definition 6.9 (Gradient operator on graph)**.**

$$\operatorname{grad} : \mathbb{R}^X \to \mathbb{R}^E, \qquad \operatorname{grad}\varphi(x,y) := \frac{\varphi(x) - \varphi(y)}{\ell(x,y)}$$

- to avoid redundancy with 'double-edges': on a symmetric graph could select arbitrary edge orientations, keep only one of the two edges.
- many different conventions possible, could also keep both edges

Lemma 6.10 (1-Lipschitz functions on graphs)**.**

- For $\psi \in \mathbb{R}^E$ set $\|\psi\|_\infty := \max_{(x,y)\in E} |\psi(x,y)|$.
- Then $\varphi \in \operatorname{Lip}_1(X) \Leftrightarrow \|\operatorname{grad}\varphi\|_\infty \leq 1$.

Proof.

- $\Leftarrow$: Let $(x,y) \in X$, let $x_1 = x, \ldots, x_K = y$ be a shortest path from x to y. Then

$$\begin{aligned} |\varphi(y) - \varphi(x)| &= \left| \sum_{i=1}^{K-1} \varphi(x_{i+1}) - \varphi(x_i) \right| \leq \sum_{i=1}^{K-1} |\varphi(x_{i+1}) - \varphi(x_i)| \\ &= \sum_{i=1}^{K-1} |\operatorname{grad}\varphi(x_{i+1}, x_i) \cdot \ell(x_{i+1}, x_i)| \leq \sum_{i=1}^{K-1} \ell(x_{i+1}, x_i) \\ &= L(x_1, \ldots, x_K) = d(x,y) \end{aligned}$$

- $\Rightarrow$:

$$|\operatorname{grad}\varphi(x,y)| = \frac{|\varphi(x) - \varphi(y)|}{\ell(x,y)} \leq \frac{d(x,y)}{\ell(x,y)} \leq 1. \qquad \square$$

Remark 6.11 (Deriving the min-cost flow problem)**.**

- Start with Kantorovich–Rubinstein formula:

$$\begin{aligned} W_1(\mu,\nu) &= \sup\{\langle\varphi, \mu - \nu\rangle | \varphi \in \operatorname{Lip}_1(X)\} \\ &= \sup\left\{\langle\varphi, \mu - \nu\rangle | \varphi \in \mathbb{R}^X, |\operatorname{grad}\varphi(x,y)| \leq 1 \,\forall\, (x,y) \in E\right\} \end{aligned}$$

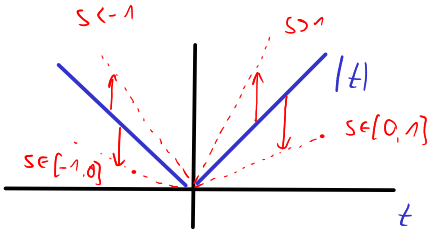
- Rewrite constraint. Let

$$H : \mathbb{R} \to \mathbb{R} \cup \{\infty\}, \qquad H(s) = \begin{cases} 0 & \text{if } |s| \leq 1, \\ +\infty & \text{else.} \end{cases}$$

  Then:

$$W_1(\mu,\nu) = \sup\left\{ \langle\varphi, \mu - \nu\rangle - \sum_{(x,y)\in E} H(\operatorname{grad}\varphi(x,y)) \,\middle|\, \varphi \in \mathbb{R}^X \right\}$$

- **Generalize notion of Lagrange multiplier.** $H(s) = \sup_{t \in \mathbb{R}} s \cdot t - |t|$



$$W_1(\mu, \nu) = \sup_{\varphi \in \mathbb{R}^X} \inf_{t \in \mathbb{R}^E} \langle \varphi, \mu - \nu \rangle - \sum_{(x,y) \in E} [\operatorname{grad} \varphi(x,y) \cdot t(x,y) - |t(x,y)|]$$

- As earlier: constraints have been expressed as optimization over an 'adversarial' variable.
- **Adjoint and minimax.** As before, use adjoint of linear operator and pretend that we can swap the order of optimization. grad^* maps $\mathbb{R}^E \to \mathbb{R}^X$.

$$\begin{aligned} W_1(\mu, \nu) &= \inf_{t \in \mathbb{R}^E} \sum_{(x,y) \in E} |t(x,y)| + \sup_{\varphi \in \mathbb{R}^X} \langle \varphi, \mu - \nu - \operatorname{grad}^* t \rangle \\ &= \inf \left\{ \sum_{(x,y) \in E} |t(x,y)| \,\middle|\, t \in \mathbb{R}^E, \operatorname{grad}^* t = \mu - \nu \right\} \end{aligned}$$

- **Explicit form of** grad^***.**

$$\begin{aligned} \langle \operatorname{grad} \varphi, t \rangle_E &= \sum_{(x,y) \in E} \frac{\varphi(x) - \varphi(y)}{\ell(x,y)} \cdot t(x,y) \\ &= \sum_{x \in X} \varphi(x) \cdot \underbrace{\left[\sum_{\substack{y \in X: \\ (x,y) \in E}} \frac{t(x,y)}{\ell(x,y)} - \sum_{\substack{y \in X: \\ (y,x) \in E}} \frac{t(y,x)}{\ell(y,x)} \right]}_{= \operatorname{grad}^* t(x)} \end{aligned}$$

- **Change of variables.** $\omega \in \mathbb{R}^E$, $\omega(x,y) := -\frac{t(x,y)}{\ell(x,y)}$.
- Introduce divergence:

$$\operatorname{div} : \mathbb{R}^E \to \mathbb{R}^X, \qquad \operatorname{div} \omega(x) := \sum_{\substack{y \in X: \\ (y,x) \in E}} \omega(y,x) - \sum_{\substack{y \in X: \\ (x,y) \in E}} \omega(x,y)$$

Then $\operatorname{div} \omega = \operatorname{grad}^* t$. Interpretation: $\omega(x,y)$ is flow on edge (x,y) from y to x (if $\omega(x,y) > 0$, in opposite direction otherwise).

(signed) flows on these edges

$$\operatorname{div}\omega(x) := \sum_{\substack{y\in X:\\(y,x)\in E}} \omega(y,x) - \sum_{\substack{y\in X:\\(x,y)\in E}} \omega(x,y)$$

= total flow leaving (+) or coming into (−) x

edges leaving x ↑ ↖ edges coming into x

- **New primal problem.**

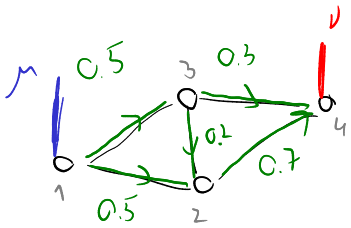
$$W_1(\mu,\nu) = \inf\left\{ \sum_{(x,y)\in E} |\omega(x,y)|\cdot \ell(x,y) \,\middle|\, \omega\in\mathbb{R}^E, \operatorname{div}\omega = \mu-\nu \right\}$$

- ω is (signed) flow on edges, transforming μ into ν, we pay for amount of flow $\times$ length of edges
- If we keep symmetric edges, then can impose constraint $\omega(x,y)\geq 0$ (since negative flow can be 'delegated' to reverse edge with positive sign) and get a linear program:

$$W_1(\mu,\nu) = \inf\left\{ \sum_{(x,y)\in E} \omega(x,y)\cdot \ell(x,y) \,\middle|\, \omega\in\mathbb{R}^E_+, \operatorname{div}\omega = \mu-\nu \right\}$$

- This problem is called the **min-cost flow problem**. (The continuous version is called Beckmann's problem.)

Example 6.12 (A simple graph example)**.**



Remark 6.13 (Existence of minimizers and relation to primal Kantorovich problem)**.**

- Let $\gamma\in\Gamma(\mu,\nu)$ be minimal in Kantorovich primal.
- For $(x,y)\in X^2$ let $(x_1=x,\dots,x_K=y)$ be shortest path from x to y. Set

$$\omega_{x,y}(a,b) = \begin{cases} 1 & \text{if } (a,b)=(x_{k+1},x_k) \text{ for some } k=1,\dots,K-1,\\ -1 & \text{if } (a,b)=(x_k,x_{k+1}) \text{ for some } k=1,\dots,K-1,\\ 0 & \text{else.}\end{cases}$$

(need second line only if we deleted the symmetric edges)

- Find $\operatorname{div} \omega_{x,y} = \delta_x - \delta_y$.
- Set now $\omega := \sum_{x,y \in X^2} \gamma(x,y) \cdot \omega_{x,y}$

$$\begin{aligned} \operatorname{div} \omega &= \sum_{x,y} \gamma(x,y) \operatorname{div} \omega_{x,y} = \sum_{x,y} \gamma(x,y) \, (\delta_x - \delta_y) \\ &= \sum_x \mu(x) \, \delta_x - \sum_y \nu(y) \, \delta_y = \mu - \nu. \end{aligned}$$

- Cost of ω: Set $G(\omega) := \sum_{(x,y) \in E} |\omega(x,y)| \cdot \ell(x,y)$.
- Find: $G(\omega_{x,y}) = d(x,y)$. Note: Since $|a+b| \leq |a| + |b|$, also G is sub-additive. Also: G is positively one-homogeneous: $G(\lambda \cdot \omega) = \lambda G(\omega)$ for $\lambda \geq 0$. Therefore:

$$G(\omega) = G\left(\sum_{x,y} \gamma(x,y) \cdot \omega_{x,y} \right) \leq \sum_{x,y} \gamma(x,y) \cdot G(\omega_{x,y}) = \sum_{x,y} \gamma(x,y) \cdot d(x,y) = W_1(\mu,\nu)$$

- Since infimal values of Kantorovich primal and min-cost flow problems are identical, the constructed ω must be optimal.
- Reverse construction (γ from ω) is also possible, by 'following' flows, but a bit more tedious.
- Alternatively: existence of minimizers can also be shown by existence of feasible candidates (use connectedness) and coercivity of G. (And closedness of feasible set, continuity of objective.)

Remark 6.14 (Alternative proof for metric properties of W_1)**.**

- $W_1(\mu,\nu) \geq 0$ since $G(\omega) \geq 0$. $W_1(\mu,\mu) = 0$ since $\omega = 0$ is feasible. $W_1(\mu, \nu \neq \mu) > 0$ since we need $\omega \neq 0$ for $\operatorname{div} \omega \neq 0$.
- Symmetry: if $\operatorname{div} \omega = \mu - \nu$ then $\operatorname{div}(-\omega) = \nu - \mu$ and $G(\omega) = G(-\omega)$.
- Triangle inequality: Let ω be optimal flow for $W_1(\mu,\nu)$, η optimal flow for $W_1(\nu,\rho)$. Then

$$\operatorname{div}(\omega + \eta) = \operatorname{div} \omega + \operatorname{div} \eta = \mu - \nu + \nu - \rho = \mu - \rho$$

  So $\omega + \eta$ is feasible flow from μ to ρ.
- Subadditivity of G:

$$W_1(\mu,\rho) \leq G(\omega + \eta) \leq G(\omega) + G(\eta) = W_1(\mu,\nu) + W_1(\nu,\rho).$$

Proposition 6.15 (Primal-dual optimality condition)**.** $\varphi \in \mathbb{R}^X$ with $\| \operatorname{grad} \varphi \|_\infty \leq 1$ and $\omega \in \mathbb{R}^E$ with $\operatorname{div} \omega = \mu - \nu$ are dual-primal optimal for Kantorovich–Rubinstein formula and min-cost flow problem if and only if

$$\operatorname{grad} \varphi(x,y) = -\operatorname{sign} \omega(x,y) \quad \text{for all } (x,y) \in E \text{ with } \omega(x,y) \neq 0.$$

Proof.

- Set $t(x,y) = -\ell(x,y) \cdot \omega(x,y)$. Then $\operatorname{grad}^* t = \mu - \nu$. Get

$$\begin{aligned} G(\omega) &= \sum_{(x,y)\in E} |t(x,y)| \geq \sum_{(x,y)\in E} t(x,y) \cdot \operatorname{grad}\varphi(x,y) \\ &= \sum_{x\in X} \operatorname{grad}^* t(x) \cdot \varphi(x) = \langle \mu - \nu, \varphi \rangle \end{aligned}$$

  with equality in the second step if and only if

$$\operatorname{grad}\varphi(x,y) = \operatorname{sign} t(x,y) \quad \text{for all } (x,y) \in E \text{ with } t(x,y) \neq 0.$$

  (This is equivalent to the above condition on ω.)

- Since $\langle \mu - \nu, \varphi \rangle \leq W_1(\mu,\nu) \leq G(\omega)$, equality of $G(\omega) = \langle \mu - \nu, \varphi \rangle$ is equivalent to optimality of ω and φ. □

Example 6.16.

- Intuition: φ wants to be large on μ, small on ν. ω flows 'against' the gradient of φ from μ to ν. ω acts indeed as Lagrange multiplier for the gradient constraint on φ.

7 Optimal transport in one dimension

7.1 Monge property, monotonous couplings and north-west corner rule

Remark 7.1.

- Get some intuition on simple problems.
- Much simpler numerically and also theoretically.

Definition 7.2 (Monge property and monotonous plans)**.**

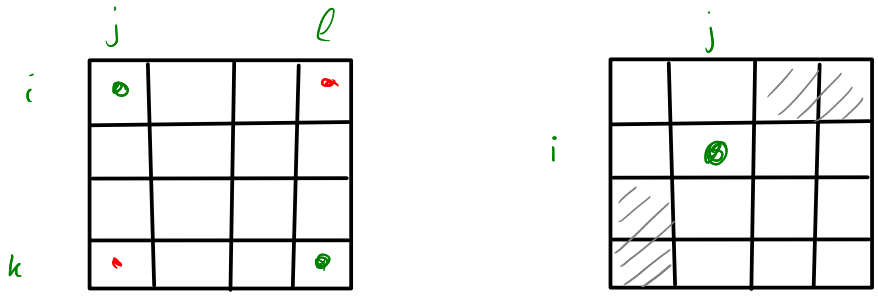
- A cost matrix $c \in \mathbb{R}^{M\times N}$ satisfies the *Monge property* if

$$c_{i,j} + c_{k,l} \leq c_{i,l} + c_{k,j} \quad \text{when} \quad i \leq k,\, j \leq l$$

  .

- A transport plan $\gamma \in \mathbb{R}_+^{M\times N}$ is *monotonous* if

$$\gamma_{i,j} > 0 \quad \Rightarrow \quad \gamma_{i',j'} = 0 \quad \text{for } [i' > i \wedge j' < j] \text{ or } [i' < i \wedge j' > j]$$



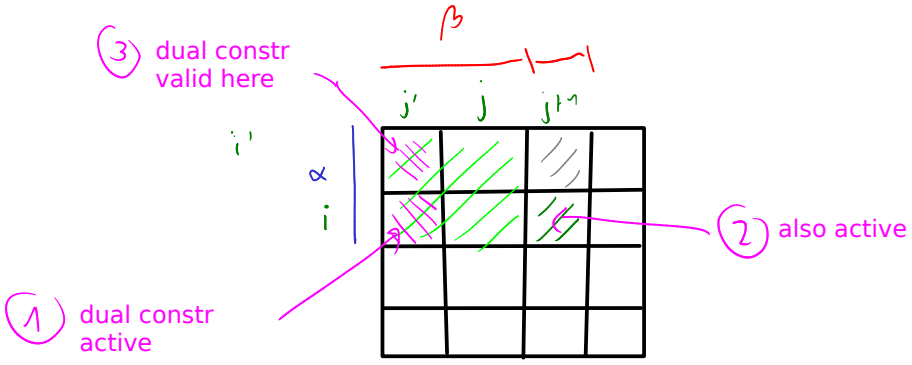
(Monge cost and monotonous coupling)

Proposition 7.3. If $c \in \mathbb{R}^{M\times N}$ satisfies the Monge property and $\gamma \in \mathbb{R}_+^{M\times N}$ is monotonous, then γ is an optimal transport plan for $\mu = \mathrm{P}_X\gamma$ and $\nu = \mathrm{P}_Y\gamma$ with respect to cost c.

Proof.

- For simplicity the proof assume that all entries of μ and ν are strictly positive. Extension is simple but tedious.
- Proof by induction. Assume γ is optimal on $\{1,\dots,i\}\times\{1,\dots,j\}$ for some $i \in \{0,\dots,M-1\}$ and $j \in \{0,\dots,N-1\}$.
- Assume $\gamma_{i,j+1} > 0$.
- (Other case, $\gamma_{i+1,j} > 0$, $\gamma_{i+1,j'} = 0$ for $j' = 1,\dots,j-1$ is analogous. The diagonal case, $\gamma_{i+1,j+1} > 0$, and both 'sides' zero, can be subsumed in either of the two cases by re-indexing.)
- [Monotonicity of γ] + [$\mu_i > 0$] + [$\nu_j > 0$] implies $\gamma_{i,j} > 0$. With this, monotonicity implies $\gamma_{i',j+1} = 0$ for $i' \in \{1,\dots,i-1\}$.

- Let α, β be optimal duals on $\{1, \dots, i\} \times \{1, \dots, j\}$. Now we need to extend β to $j+1$.
- Without loss of generality can assume: $\alpha_i = \min_{j' \in \{1,\dots,j\}} c_{i,j'} - \beta_{j'}$, let j' be a minimizing index. So: $\alpha_i + \beta_{j'} = c_{i,j'}$ (1).
- By PD optimality condition need: $\alpha_i + \beta_{j+1} = c_{i,j+1}$ (2).
- Optimality of α and β on previous set: $\alpha_{i'} + \beta_{j'} \le c_{i',j'}$ (3).



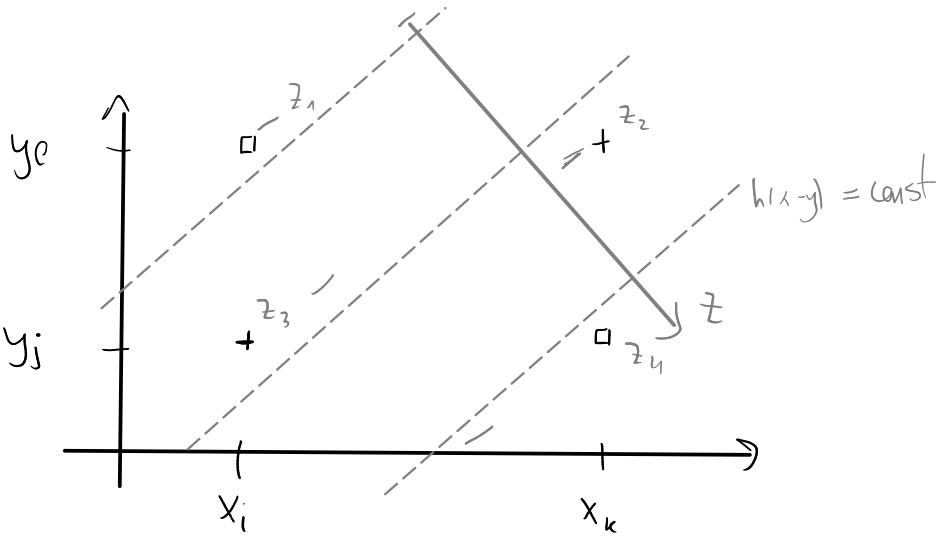
- Now combine (1), (2), (3):

$$\underbrace{\alpha_{i'}}_{\le c_{i',j'} - \beta_{j'} \ (3)} + \underbrace{\beta_{j+1}}_{= c_{i,j+1} - \alpha_i \ (2)} \le c_{i',j'} - \beta_{j'} + c_{i,j+1} - \alpha_i = c_{i',j'} + c_{i,j+1} - \underbrace{c_{i,j'}}_{(1)} \le c_{i',j+1}$$

  where we used the Monge property in the last step with $i' \le i$ and $j' \le j+1$.
- So α and β are dual feasible on $\{1, \dots, i\} \times \{1, \dots, j+1\}$ and satisfy the PD optimality relation. □

Lemma 7.4. If $h : \mathbb{R} \to \mathbb{R}$ is convex, $X = \{x_1, \dots, x_M\} \subset \mathbb{R}$, $Y = \{y_1, \dots, y_N\} \subset \mathbb{R}$ with $x_i \le x_{i+1}$ and $y_j \le y_{j+1}$ for $i = 1, \dots, M-1$, $j = 1, \dots, N-1$, then $c_{i,j} := h(x_i - y_j)$ satisfies the Monge property.

Proof.



- Let $i \le k$, $j \le l$. Set

$$z_1 := x_i - y_l, \qquad z_2 := x_k - y_l, \qquad z_3 := x_i - y_j, \qquad z_4 := x_k - y_j.$$

- Then $z_1 \le z_2 \le z_4$, $z_1 \le z_3 \le z_4$.
- $z_2 - z_1 = x_k - x_i = z_4 - z_3$.
- So there exists $\lambda \in [0,1]$ such that

$$z_2 = \lambda\, z_1 + (1-\lambda)\, z_4, \qquad z_3 = (1-\lambda)\, z_1 + \lambda\, z_4.$$

- Now:

$$\begin{aligned} c_{i,j} + c_{k,l} &= h(z_3) + h(z_2) \le (1-\lambda)\, h(z_1) + \lambda\, h(z_4) + \lambda\, h(z_1) + (1-\lambda)\, h(z_4) \\ &= h(z_1) + h(z_4) = c_{i,l} + c_{k,j}. \end{aligned}$$

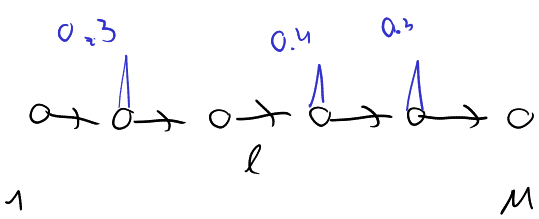
□

Remark 7.5. The north-west corner rule generates monotonous transport plans. (See exercises.)

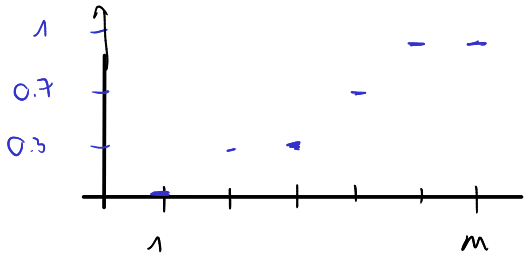
7.2 The cumulative distribution formula for W_1 on chain graphs.

Definition 7.6 (Chain graph, metric and cumulative distributions)**.**

- Let $X = Y = \{1, \dots, M\}$ with metric d induced by a chain graph with edges $(i+1, i)$, $i = 1, \dots, M-1$ and edge lengths $\ell(i+1, i)$.



(masses on a chain graph)



(cumulative distribution)

- For uniform edge lengths $\ell(i+1, i) = \Delta x$ one has $d(i, j) = \Delta x \cdot |i - j|$.
- For $\mu \in \mathbb{R}^M$ define the cumulative distribution function $F^\mu \in \mathbb{R}_+^{M-1}$ by

$$F_i^\mu := \sum_{i'=1}^{i} \mu_{i'} \quad \text{for } i = 1, \dots, M-1.$$

- We can formally extend F^μ by $F_0^\mu := 0$ and $F_M^\mu := \sum_{i=1}^M \mu_i$ (the latter is done by `numpy.cumsum`).
- Claim: $W_1(\mu, \nu) = \sum_{i=1}^{M-1} \ell(i+1, i) \cdot |F_i^\mu - F_i^\nu|$.

Proof.

- We prove the result by constructing a primal-dual feasible pair for the min-cost flow problem and the Kantorovich–Rubinstein formula.
- Set $\omega(i+1,i) := F_i^\mu - F_i^\nu$ (difference of mass of μ and ν on vertices $\{1,\dots,i\}$ is precisely the mass that needs to flow on edge from i to $i+1$).

  $$\operatorname{div}\omega(i) = \omega(i+1,i) - \omega(i,i-1) = F_i^\mu - F_i^\nu - F_{i-1}^\mu + F_{i-1}^\nu = \mu_i - \nu_i.$$

  (Need to be a bit careful at first and last vertex.)
- $G(\omega) = \sum_{i=1}^{M-1} \ell(i+1,i)|F_i^\mu - F_i^\nu|$.
- construct matching dual:

  $$\varphi(1) = 0, \qquad \varphi(i+1) = \begin{cases} \varphi(i) - \ell(i+1,i) & \text{if } \omega(i+1,i) > 0, \\ \varphi(i) + \ell(i+1,i) & \text{if } \omega(i+1,i) < 0, \\ \in [\varphi(i) - \ell(i+1,i), \varphi(i) + \ell(i+1,i)] & \text{else.} \end{cases}$$

- clear: $\|\operatorname{grad}\varphi\|_\infty \le 1$, φ and ω satisfy PD optimality condition $\Rightarrow$ both are optimal

□

Example 7.7 (Two Diracs)**.**

Example 7.8 (One Dirac (middle) splits into two (left and right))**.**

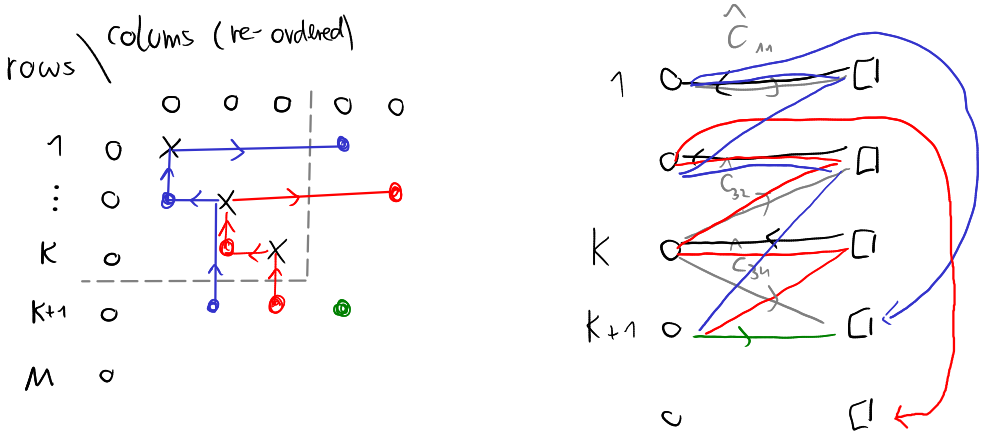
8 The Hungarian method

8.1 Intuition and description of the algorithm

Remark 8.1 (Outline, intuition and relation do Dijkstra's algorithm)**.**

- We want to solve the linear assignment problem, which can be seen as a special case of the Kantorovich optimal transport problem. For $M \in \mathbb{N}$, let $\mu = \nu = (1, \dots, 1) \in \mathbb{R}_+^M$ and let cost $c \in \mathbb{R}^{M \times M}$ be given. Solve optimal transport problem $\min\{\langle c, \gamma \rangle \,|\gamma \in \Gamma(\mu, \nu)\}$. In this case, $\Gamma(\mu, \nu)$ are the bi-stochastic matrices. We will show that solving the problem over the bi-stochastic matrices is equivalent to solving it over permutation matrices.
- The Hungarian method can be seen as iteratively constructing an optimal assignment by starting from an empty matrix and then add and move ones such that at each step the current matrix is optimal for its marginals.
- Search for updates (and proof of optimality) is obtained by simultaneously constructing a dual solution.
- Upon convergence an optimal primal and dual solution are available.
- Originally introduced by Kuhn in 1955[2], named in honor of Hungarian mathematicians Dénes Kőnig and Jenő Egerváry upon the work of which it builds. Many variants exist, lots of heuristics about initialization and ordering in for loops, do not change asymptotic worst case complexity. We will focus on the main loop and basic idea.
- Before each iteration of the main loop of the algorithm assume that the following holds:
  - rows $1, \dots, K$ are assigned to some columns,
  - some current dual values are given such that: $\alpha_i + \beta_j \leq c_{i,j}$, equality on current assignments
  - (remark: then this current partial assignment is optimal between its marginals)
- now consider new row $K + 1$. want to extend assignment, also need a new previously unassigned column
  - consider various 'changes' and extensions to the original assignment, and the 'inflicted cost'
  - interpret each extension as a path
  - edge from row i to col j has cost $c_{i,j}^{\text{eff}} := c_{i,j} - \alpha_i - \beta_j$
  - edge from col j to row i: cost $c_{i,j}^{\text{eff}} := 0$ if (i, j) is part of current partial assignment, otherwise no edge
  - then best extension corresponds to shortest path from row $K + 1$ to any currently unassigned column on this directed graph

[2]H. W. Kuhn: The Hungarian method for the assignment problem, Naval Research Logistics 2, pages 83–97, 1955

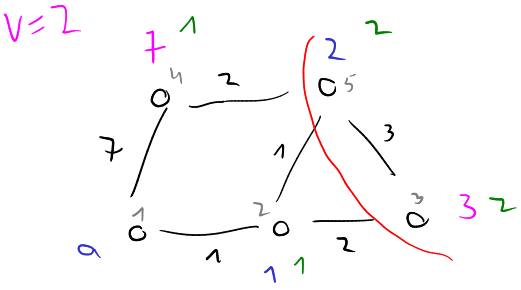


- when shortest path was found:
  - update assignment and update dual variables
  - will find: satisfies the original assumptions, now for $K+1$
  - therefore, the algorithm terminates after M main iterations with an optimal assignment

Algorithm 8.1 (Recall: Dijkstra's shortest path algorithm on general graph with non-negative edge weights)**.**

```
1:  function DIJKSTRA(c,a,label)
2:      // find shortest path from a to region where label==0 with edge weights c ≥ 0
3:      // set current vertex to root
4:      i=a
5:      // current lower bounds on distance from initial node a
6:      d=[∞ for j ∈ X]; d[i]=0                                   // X = {1,...,M}
7:      // indices of predecessors in shortest paths from root, 0 means 'empty'
8:      pred=[0 for j ∈ X]
9:      v=0
10:     // vertices that still need to be scanned
11:     scan=X \ {i}
12:     while label[i]!=0 do                                      // while not reached target region
13:         // rowscan to get new lower bound on distances
14:         for j ∈ scan do
15:             if c[i,j]+v<d[j] then
16:                 d[j]=c[i,j]+v
17:                 pred[j]=i
18:             end if
19:         end for
20:         // find smallest current lower bound
21:         vNew=∞
22:         for j ∈ scan do
23:             if d[j]<vNew then
24:                 vNew=d[j]
25:                 i=j
```

```
            end if
        end for
        v=vNew
        scan=scan\{i}
    end while
    b=i
    return b,d,pred
end function
```



- throughout the algorithm (and upon termination): `d[i] >= dist[a,i]`, equality when $i \notin$ `scan`.
- `dist` is length of shortest paths between nodes, it is almost metric: non-negative, satisfies triangle inequality, (but may be asymmetric or non-separating, if c is asymmetric or not strictly positive away from diagonal, may also be $+\infty$ when graph is not connected, i.e. $c[i,j] = \infty$ for some entries).

Algorithm 8.2 (Adjustment to 'partial assignment graph')**.**

- assignment graph is bipartite, directed with two disjoint vertex sets $X = Y$, with edges only going from X to Y and from Y to X, but not from X or Y to themselves.
- from each Y there is at most one edge to X, if we reach a node in Y with no outgoing edge, we have found the sought-after shortest path / extension, otherwise we automatically take the single outgoing edge with weight zero
- partial assignments are stored in arrays γ`C2R` and γ`R2C` which store for each column or row the currently assigned row or column (and zero if the column or row is still unassigned)
- slightly adjust the general algorithm to this special setting

```
function FindPath(c,α,β,γC2R,a)
    // find shortest path from a to region where γC2R==0
    // with edge weights c[i,j]-α[i]-β[j] >= 0
    i=a
    d=[∞ for j ∈ X]
    predC=[0 for j ∈ X]
    predR=[0 for j ∈ X]
    v=0
    scanCols=X
    while i!=0 do
```

```
        // rowscan to get new lower bound on distances
        for j ∈ scanCols do
            if c[i,j]-α[i]-β[j]+v<d[j] then
                d[j]=c[i,j]-α[i]-β[j]+v
                predC[j]=i
            end if
        end for
        // find smallest current lower bound
        vNew=∞
        for j ∈ scanCols do
            if d[j]<vNew then
                vNew=d[j]
                jMin=j
            end if
        end for
        v=vNew
        scanCols=scanCols\{jMin}
        i=γC2R[jMin]
        if i!=0 then
            predR[i]=jMin
        end if
    end while
    b=jMin
    return b,d,predC,predR
end function
```

Algorithm 8.3.

- With the path search in place, we can now formulate the full Hungarian method

```
function HungarianMethod(c)
    γR2C=[0 for i ∈ X]
    γC2R=[0 for i ∈ X]
    α=[0 for i ∈ X]
    β=[0 for i ∈ X]
    for a ∈ X do
        // find next shortest path
        b,d,predC,predR=findPath(c,α,β,γC2R,a)
        // update dual variables
        for i ∈ X do
            j=predR[i]
            if j!=0 then
                α[i]+=d[b]-d[j]
                β[j]-=d[b]-d[j]
            end if
        end for
        α[a]+=d[b]
        // update primal variable
        j=b
```

```
20:         while j!=0 do
21:             i=predC[j]
22:             jPred=γR2C[i]                    // predecessor, we are back-tracking the path
23:             γC2R[j]=i
24:             γR2C[i]=j
25:             j=jPred
26:         end while
27:     end for
28:     return γR2C,γC2R,α,β
29: end function
```

Remark 8.2. The algorithm can be extended to discrete optimal transport problems with μ, ν being general probability vectors in Σ_M, Σ_N.

8.2 Proof of termination and optimality

Lemma 8.3. When α,β are dual feasible, γ`C2R` and γ`R2C` are valid partial assignments of K elements at beginning of an iteration, then after iteration γ`C2R` and γ`R2C` are valid partial assignment of $K+1$ elements.

Proof.

- shortest path returned by FINDPATH leads from row a (which is not part of initial partial assignment) to column b (which is not part of initial partial assignment).
- such a path always exists, since we can simply use a direct edge c`[`a`,`b`]` for some b
- the path is bipartite, going from row to col to row etc, containing a number n of edges col-to-row which are part of assignment and $n+1$ edges row-to-col, that are not part of assignment, change in assignment corresponds to 'flipping' assignment along this path
- therefore get valid assignment with one more entry □

Lemma 8.4. When α,β are dual feasible, γ`C2R` and γ`R2C` are valid partial assignments of K elements, satisfying the complementary slackness condition α`[`i`]+`β`[`j`]=`c`[`i`,`j`]` on assigned pairs at beginning of an iteration, then after iteration α and β are dual feasible and satisfy complementary slackness on updated assignment.

Proof.

- let $A_R, A_C \subset X$ be rows and columns of intermediate nodes visited in search for shortest path from a to b (including a,b).
  A_R are those indices where `predR` was changed and a.
  A_C are those indices where `predC` was changed (which includes b).
- complements $I_R = X \setminus A_R$, $I_C = X \setminus A_C$
- Dijkstra: $d[j] \leq d[b]$ for all $j \in A_C$
- α is increased on A_R, β is decreased on A_C
- $I_R \times I_C$: no dual feasibility issues (no change, so feasibility and slackness are preserved)

- $I_R \times A_C$: no issues (only decrease beta, no assignments in this set, since all assignments of A_C go into A_R)
- $A_R \times I_C$:
  - let $i \in A_R$, $j \in I_C$: no assignments in this set, since all assignments of A_C go into A_R.
  - so only need to check dual feasibility. Assume i is successor of j`Pred` in path from a to b. Then change in $\alpha[i]$ is given by

$$\begin{aligned}\Delta\alpha[i] = d[b] - d[j\texttt{Pred}] &= \mathrm{dist}[\mathrm{row}_a, \mathrm{col}_b] - \underbrace{\mathrm{dist}[\mathrm{row}_a, \mathrm{col}_{j\texttt{Pred}}]}_{=\mathrm{dist}[\mathrm{row}_a, \mathrm{row}_i]} \\ &\leq \mathrm{dist}[\mathrm{row}_a, \mathrm{col}_j] - \mathrm{dist}[\mathrm{row}_a, \mathrm{col}_i] \leq c^{\mathrm{eff}}[i,j]\end{aligned}$$

    where we use complementary slackness in the third step and the triangle inequality for dist in the last step. If i is the initial node a, with $\mathrm{dist}[\mathrm{row}_a, \mathrm{row}_a] = 0$ the same expression holds. With this we obtain

$$\Delta\alpha[i] + \underbrace{\Delta\beta[j]}_{=0} \leq c^{\mathrm{eff}}[i,j]$$

  - Now add the old α and β on both sides to obtain dual feasibility.
- $A_R \times A_C$:
  - let $i \in A_R$, $j \in A_C$: arguing as above, get

$$\begin{aligned}\Delta\alpha[i] + \Delta\beta[j] &= \mathrm{dist}[\mathrm{row}_a, \mathrm{col}_b] - \mathrm{dist}[\mathrm{row}_a, \mathrm{row}_i] - (\mathrm{dist}[\mathrm{row}_a, \mathrm{col}_b] - \mathrm{dist}[\mathrm{row}_a, \mathrm{col}_j]) \\ &= \mathrm{dist}[\mathrm{row}_a, \mathrm{col}_j] - \mathrm{dist}[\mathrm{row}_a, \mathrm{row}_i] \leq c^{\mathrm{eff}}[i,j]\end{aligned}$$

    (note: if $j = b$, then $\Delta\beta[j] = 0$, which is also true since $\beta[b]$ is not changed). So we have dual feasibility.
  - Now check slackness: First on old assignments (which might no longer be in use). Let (i,j) be an old assigned pair in $A_R \times A_R$. Then `predR[i]=j` (see FINDPATH, line 30). Therefore, dual changes to α and β in HUNGARIANMETHOD, line 13 cancel each other. Complementary slackness is preserved from previous iteration.
  - Now updated assignments: when (i,j) is edge in shortest paths from a to b, j being successor of i, then edge length between them is precisely distance between them, i.e. have equality $\mathrm{dist}[\mathrm{row}_a, \mathrm{col}_j] - \mathrm{dist}[\mathrm{row}_a, \mathrm{row}_i] = c^{\mathrm{eff}}[i,j]$ and therefore, complementary slackness holds. □

Proposition 8.5. The Hungarian method terminates and solves the linear assignment problem.

Proof. The proof is a result of the two preceding lemmas. □

Remark 8.6.
The worst-case complexity of the Hungarian method is $O(M^3)$:

- The main loop in HUNGARIANMETHOD, line 6 runs for M iterations.

- At each iteration the dual update loop, line 10, runs for M iterations. (This could be made a bit shorter in practice by using a dynamic list (stack, queue) for keeping track of relevant rows. But it would not change asymptotic worst-case bound, since the length of the list would be $O(M)$.)
- The primal update loop, 20, retraces the shortest path. Its length is therefore also $O(M)$.
- We need to check the complexity of FINDPATH.
- The main loop at line 10 runs at most over every row, i.e. M times.
- Within the main loop, the col-scans run at most over every column, i.e. M times.
- So the total complexity of FINDPATH is $O(M^2)$.
- The total complexity of HUNGARIANMETHOD is $O(M^3)$.
- Precise bound is not so important. Crucial: not exponential! There are $M!$ potential assignments. Hungarian method proves that we do not need to check all of them to find the best. This is possible by the relation to convex optimization.
- Still: strictly super-linear! So can still become quickly impractical on large instances.

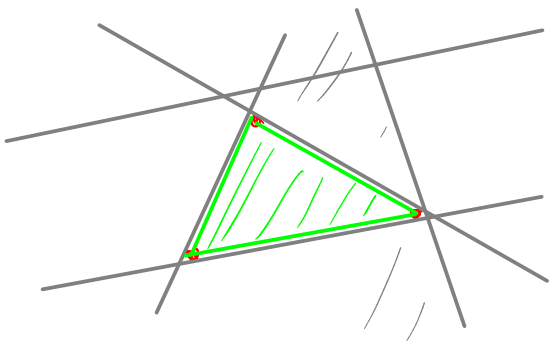
8.3 The Birkhoff-von Neumann theorem

Definition 8.7 (Convex hull and vertices)**.**

- Let $A \subset \mathbb{R}^d$. The *convex hull* of A is given by

$$\operatorname{conv} A := \left\{ \sum_{i=1}^{k} \lambda_i \cdot x_i \middle| k \in \mathbb{N},\, x_1, \ldots, x_k \in A,\, \lambda_1, \ldots, \lambda_k > 0,\, \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

- Intuitively, this is set set of all points that lie 'between' the points of A.
- Can also show: $\operatorname{conv} A$ is intersection of all convex sets that contain A. So $\operatorname{conv} A$ is 'smallest convex set which contains A.'
- Even more: If A is closed, then $\operatorname{conv} A$ is the intersection of all closed half-spaces that contain A.
- The vertices of a convex set are those that cannot be written as convex combinations of other points in the set.



Proposition 8.8 (Birkhoff-von Neumann)**.** Let $M \in \mathbb{N}$. Let P_M be the set of $M \times M$ permutation matrices and D_M be the set of $M \times M$ bi-stochastic matrices. Then $\operatorname{conv} P_M = D_M$ and P_M are the vertices of D_M.

Proof.

- Clearly, $P_M \subset D_M$. Any convex combination of permutation matrices is bi-stochastic. For instance: Let $A = \sum_{l=1}^{k} \lambda_l P^l$ where each $P^l \in P_M$. Then:

$$\sum_{j=1}^{M} A_{i,j} = \sum_{j=1}^{M} \sum_{l=1}^{k} \lambda_l P^l_{i,j} = \sum_{l=1}^{k} \lambda_l \underbrace{\sum_{j=1}^{M} P^l_{i,j}}_{=1} = \sum_{l=1}^{k} \lambda_l = 1$$

  Therefore $\operatorname{conv} P_M \subset D_M$.

- We will prove $D_M \subset \operatorname{conv} P_M$ at the end (because it is the most work). Assume for now that it holds and consider the vertices:
  - Consider some permutation matrix $P \in P_M$ and any bi-stochastic matrix $D \in D_M$. If $D \neq P$, then D must have a non-zero entry somewhere where P is zero (a permutation matrix is completely specified by the list of non-zero entries). Therefore, P cannot be written as convex combination of any other bi-stochastic matrices, since only positive coefficients λ_k are allowed. So P must be a vertex of D_M.
  - Let D be a vertex of D_M. So, since $D \in D_M$, by assuming $D_M \subset \operatorname{conv} P_M$, D can be written as convex combination of some P_M. But since it is a vertex, it cannot be written as convex combination of multiple matrices. Therefore, D must be in P_M.

- For the last step: we will use the Hungarian method to show that any $D \in D_M$ can be written as $D = \lambda\, P + (1 - \lambda)\, D'$ for some $\lambda \in (0, 1]$, $P \in P_M$ and $D' \in D_M$ and if $\lambda < 1$, then D' has at least one less strictly positive entry than D (next Lemma). Then we apply the same again on D' and so on. At each step k we get a decomposition as follows:

$$D = \sum_{i=1}^{k} \lambda_i\, P^i + D'$$

  where all $\lambda_i > 0$ and all D' has at least k less strictly positive entries than D (or the sum has terminated before). Since the number of positive entries in D is finite, the sum must eventually terminate and we have found the sought-after convex combination. □

Now provide the missing Lemma.

Lemma 8.9. Let $D \in D_M$. Then there are $\lambda \in (0, 1]$, $P \in P_M$ and $D' \in D_M$, such that $D = \lambda\, P + (1 - \lambda)\, D'$. If $\lambda < 1$, D' has at least one less strictly positive entry than D.

Proof.

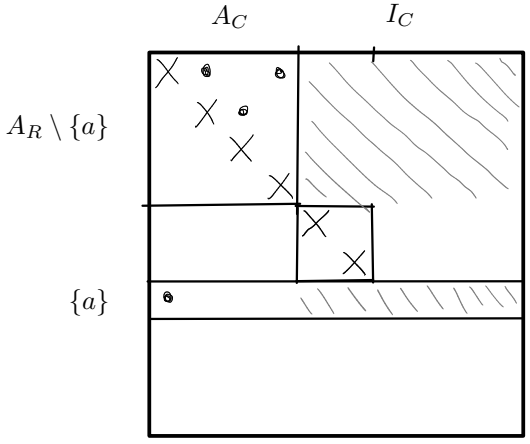
- We will use the Hungarian method to find some P such that $[P_{i,j} > 0] \Rightarrow [D_{i,j} > 0]$ for all $i, j \in \{1, \ldots, M\}$. Then set

$$\lambda := \min\{D_{i,j} \,|\, i, j \in \{1, \ldots, M\},\, P_{i,j} > 0\}.$$

- If $\lambda < 1$, set $D' := (D - \lambda P)/(1 - \lambda)$. By construction, $D \geq \lambda P$ and $D - \lambda P$ has at least one less positive entry than D.
- If $\lambda = 1$, the choice of D' is irrelevant.
- Now focus on finding P: define the cost

$$c_{i,j} := \begin{cases} 0 & \text{if } D_{i,j} > 0, \\ 1 & \text{else.} \end{cases}$$

- Run the Hungarian method on this cost and let P be the returned optimal assignment. If $\langle c, P \rangle = 0$, then $[P_{i,j} > 0] \Rightarrow [c_{i,j} = 0] \Rightarrow [D_{i,j} > 0]$.
- Now need to show that $\langle c, P \rangle = 0$. We will show that this holds throughout the algorithm. Clearly, it holds upon initialization, since $P = 0$.
- As long as $d[b] = 0$ in the result of FINDPATH, no edges with $c_{i,j} = 1$ have been used for shortest paths, the duals α and β remain unchanged at zero and thus P must live solely on entries with $c_{i,j} = 0$.
- We must therefore make sure that no edge with $c_{i,j} = 1$ becomes part of a shortest path in FINDPATH.
- Assume now for some $a \in \{1, \ldots, M\}$, within FINDPATH, we have explored the sub-graph reachable from a with zero distance in FINDPATH and no connection to an unassigned column has been found so far. We have $\#(A_R) = \#(A_C) + 1$.



- This means that D is zero on $A_R \times I_C$ (no edges of length zero in this set). Using that D is bi-stochastic we obtain the following contradiction:

$$\#(A_R) = \sum_{\substack{i \in A_R \\ j \in \{1,\ldots,M\}}} D_{i,j} = \sum_{\substack{i \in A_R \\ j \in A_C}} D_{i,j} \leq \sum_{\substack{i \in \{1,\ldots,M\} \\ j \in A_C}} D_{i,j} = \#(A_C) = \#(A_R) - 1$$

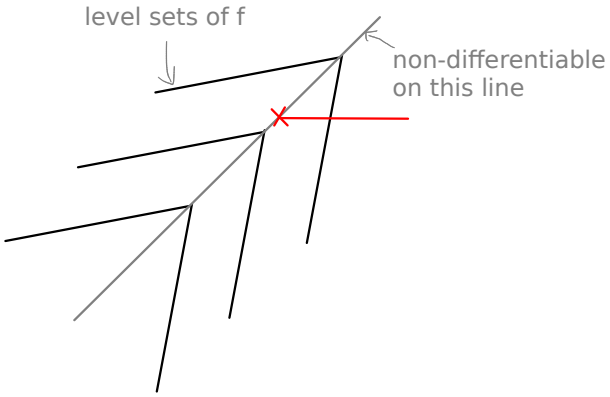
- Therefore, this situation cannot occur and we must always find a path to an unassigned column with length zero. □

9 The auction algorithm

9.1 Intuition and description of the algorithm

Remark 9.1 (Context, motivation)**.**

- issue with Hungarian method: long, explicit search for extensions (augmenting paths) is expensive and cannot be parallelized
- dual feasible set seems a bit easier to handle than primal feasible set, can we approach the problem from the dual side?
- problem: alternating maximization on the dual does not converge, since problem is non-smooth (due to constraints).
- can show $\alpha^{ccc} = \alpha^c$, so alternating maximization on dual becomes stationary after two iterations, without reaching a minimizer in general
- sketch: 2d example, getting caught in a ridge



- one interpretation of auction algorithm: local alternating dual optimization, but 'overstep' local optimum by a little, controlled by a parameter ε
- use partial primal assignment to coordinate the 'overstepping'

Remark 9.2 (alternative interpretation (from which the name derives))**.**

- elements in X compete for elements in Y by means of an auction
- interpretation of variables:
  $\alpha[x]$: how much does x need to pay,
  $c[x, y]$: transport cost (assumed non-negative),
  $-\beta[y]$: how much money does y receive?
- if x wants to buy y, they need to pay $c[x, y] - \beta[y]$, i.e. transport cost + what y demands on top of that.
- we start with $\alpha = \beta = 0$, and no buy-assignments
- each iteration has two stages: bidding stage and assignment stage

- during the bidding stage, all x that do currently not have an assigned y, make a bid for the y that is currently the cheapest one for them, i.e. $\operatorname{argmin}_y c[x,y] - \beta[y]$.
- during the assignment stage, each y reviews the received bids and chooses the one for which $-\beta[y]$ is maximal. the bidding x is then temporarily assigned to that y.
- a key step: when y adjusts their demanded price $-\beta[y]$, they add an additional step ε, so that if one x wants to out-bid another, they have to offer at least ε more
- the auction ends when all buyers and objects are assigned

Remark 9.3 (on the role of ε)**.**

- the algorithm only terminates if $\varepsilon > 0$
- we will show: the number of bidding stages is bounded by $\approx M \cdot \|c\|_\infty/\varepsilon$
- the algorithm always returns a feasible primal and dual pair, but they are not necessarily globally optimal.
- the usual complementary slackness condition (primal-dual optimality condition) is only satisfied up to ε:
$$[\gamma_{i,j} > 0] \qquad \Rightarrow \qquad [c_{i,j} - \alpha_i - \beta_j \leq \varepsilon]$$
- so the PD gap is at most:
$$\sum_{i,j} c_{i,j}\,\gamma_{i,j} - \sum_i \alpha_i - \sum_j \beta_j = \sum_{i,j} [c_{i,j} - \alpha_i - \beta_j]\;\gamma_{i,j} \leq \varepsilon \sum_{i,j} \gamma_{i,j} = \varepsilon \cdot M$$
- trade-off: runtime vs solution precision
- in principle: if values of c are integer, then among permutation matrices the cost $\langle c, \gamma\rangle$ takes integer values. a non-optimal assignment the has to be sub-optimal by at least 1. so if $\varepsilon < 1/M$, the algorithm must return the optimal assignment
- we will also show: it is possible to reduce the value of ε iteratively, reducing the total number of bidding stages to at most $O(M^2 \cdot \log \|c\|_\infty/\varepsilon)$
- also will discuss: a heuristic adaptation of the bidding behaviour that is more effective in practice

Remark 9.4.
auction algorithm features several important paradigms for large scale optimization

- the Hungarian method seeks for sophisticated, exact, non-local updates
- the auction algorithm makes only local updates that are cheap/easy to generate
- it can be parallelized (at least locally)
- by choosing ε one can balance between run-time and accuracy
- gradually reducing ε yields a more efficient method

Algorithm 9.1 (statement of the algorithm)**.**

```
1: function submitBids(c,α,β,γRow)
2:    bidLists=[[] for i in range(M)] # generate empty bid lists
3:    nBids=0
4:    for x in range(M) do
5:       if γRow[x]<0 then
6:          # iterate over unassigned x
7:          # find most attractive y
8:          y=argmin(c[x,:]-β)
9:          # update dual
10:         α[x]=c[x,y]-β[y]
11:         # submit bid
12:         bidLists[y].append(x)
13:         # for convenience:  total bid counter
14:         nBids+=1
15:      end if
16:   end for
17:   return nBids,bidLists
18: end function
19:
20: function acceptBids(c,α,β,bidLists,γRow,γCol,ε)
21:   for y in range(M) do
22:      # iterate over all y that received at least one bid
23:      if len(bidLists[y])>0 then
24:         # find best bid
25:         idx=argmin(c[bidLists[y],y]-α[bidLists[y]])
26:         x=bidLists[y][idx]
27:         # update assignment and duals
28:         # make dual a bit smaller than ultimately necessary
29:         β[y]=c[x,y]-α[x]-ε
30:         # update assignment:  remove old assigned x, add new one
31:         xOld=γCol[y]
32:         if xOld>=0 then
33:            γRow[xOld]=-1
34:         end if
35:         γCol[y]=x
36:         γRow[x]=y
37:      end if
38:   end for
39: end function
40:
41: function auction(c,ε)
42:   α=zeros(M)
43:   β=zeros(M)
44:   γRow=full(M,-1)
45:   γCol=full(M,-1)
46:   loop
47:      nBids,bidLists=submitBids(c,α,β,γRow)
```

```
48:         if nBids==0 then
49:             break
50:         end if
51:         acceptBids(c,α,β,bidLists,γRow,γCol,ε)
52:     end loop
53:     return γRow,γCol,α,β
54: end function
```

9.2 Convergence analysis

Lemma 9.5 (Monotonicity and increments)**.**

- Throughout the algorithm, β is non-increasing, α is non-decreasing.
- When y accepts a bid, $\beta(y)$ decreases by ε.

Proof.

- if x submits bid to y, then $\alpha^n(x) = c(x, y) - \beta^{n-1}(y)$ (here n denotes the iteration number in the main loop, to keep track of different values at different iterations)
- if y receives best bid from x then

$$\beta^n(y) = c(x, y) - \alpha^n(x) - \varepsilon = \beta^{n-1}(y) - \varepsilon.$$

- so β non-increasing $\Rightarrow$ α non-decreasing. □

Lemma 9.6 (Primal-dual relation)**.**

- after each assignment phase, primal is consistent partial assignment (i.e. each row and column is assigned at most once).
- duals are feasible throughout algorithm.
- primal and duals satisfy ε-complementary slackness (ε-CS for short), which means:

$$[\gamma_{i,j} > 0] \qquad \Rightarrow \qquad [c_{i,j} - \alpha_i - \beta_j \leq \varepsilon]$$

Proof.

- if y accepts bid from some x, x was previously unassigned. if y was not previously unassigned, the unique x' that was previously assigned to it becomes unassigned. then x and y are assigned to each other. thus, no double assignments happen.
- dual feasibility:
- since we assume $c \geq 0$, $\alpha = 0$, $\beta = 0$ upon initialization, the duals are feasible initially.
- upon submitting a bid, $\alpha(x)$ is set to the value of $\beta^c(x)$, i.e. all constraints in row x are satisfied.
- currently assigned x do not submit a bid and therefore do not change their dual variable.
- so if duals were feasible before bidding stage, they are satisfied after bidding stage.

- β is only decreasing, so if duals were feasible before assignment stage, their are feasible after assignment stage.
- if y accepts bid from some x (and the two are assigned to each other), $\beta(y)$ is set to $c(x,y) - \alpha(x) - \varepsilon$, hence the entry satisfies the ε-CS condition. $\alpha(x)$ and $\beta(y)$ remain unchanged as long as the assignment is intact (x submits no bids, the assignment is removed when y accepts another bid). □

Lemma 9.7 (Iteration bound)**.** The algorithm terminates after at most $(M-1)\cdot\lceil\|c\|_\infty/\varepsilon+1\rceil+1$ iterations.

Proof.

- during each iteration that does not lead to termination, a bid by at least one x is submitted.
- so at least one bid is accepted ⇒ one entry of β is decreased by (at least) ε (see previous Lemma)
- if y has been assigned once, it never gets unassigned again, only re-assigned
- for unassigned y, $\beta(y)$ is still zero, since no bid was accepted so far
- if algorithm does not terminate, then at least one y must remain unassigned, and another y' must receive an infinite amount of bids, the corresponding $\beta(y')$ tends to $-\infty$.
- eventually we will have: $c(x,y') - \beta(y') > c(x,y) - \beta(y)$; y' will be no longer be competitive and receive no more bids.
- this is a contradiction. so each y must eventually receive bid. ⇒ all y assigned ⇒ all x assigned ⇒ termination
- get a bound on the number of iterations: denote $C := \|c\|_\infty$ for simplicity. let y be the entry that last receives a bid.
- let y' be some other element. if $C < -\beta(y')$ then $c(x,y) - \beta(y) < c(x,y') - \beta(y')$ for all x, and thus y' will be no longer preferred over y during bid submission.
- so y' can accept at most $\lceil C/\varepsilon\rceil + 1$ bids before losing competitiveness to y
- so after at most $(M-1)\cdot\lceil C/\varepsilon + 1\rceil$ iterations all $y' \neq y$ will have lost competitiveness to y. y will receive a bid in the next iteration and the algorithm terminates.

□

Remark 9.8 (total complexity bound)**.**

- submitting a bid has a complexity $O(M)$ since one row-scan must be performed
- receiving a bid has a complexity of $O(M^2)$, since up to M bids can be received, each of which cost $O(M)$ to generate
- by the previous lemma $O(MC/\varepsilon)$ bids are received
- so the total complexity is $O(M^3C/\varepsilon)$ (here we are a bit sloppy about the meaning of O with more than one variable)

Example 9.9 (return to the economic interpretation)**.**

- Let again X denote cafes, Y bakeries in Paris. Assume all bakeries and cafes produce and buy a unit amount of bread, so the allocation boils down to an assignment problem.
- Assume $c(x, y)$ includes the transport cost as well as the production cost, i.e. this is the minimal amount that cafe x has to pay. Then each cafe x wants to buy from the bakery with the lowest $c(x, y)$ (for x fixed). But this may not yield a valid assignment.
- Assume the cafes perform an auction to resolve the conflict.
- Some extreme, but prototypical situations: all cafes and bakeries are very close, except for a few outliers. How are the duals chosen?

Remark 9.10 (aggressive bidding strategy)**.**

- the determination of the dual variable α`[`x`]` can be adjusted as follows:
- in addition to computing the minimal y, compute the 'next best minimizer':
  y_2`=argmin(`c`[`x`,:]-`β`|`y_2 `!=` y`)`
- then the dual is set to:
  α`[`x`]=`c`[`x`,`y_2`]-`β`[`y_2`]`
- this may temporarily violate the constraint at (x, y), but this is fixed upon acceptance of a bid at y (even if it is one not coming from x)
- in practice this may faster resolve certain 'almost-ties' where two x compete for the same y for several iterations. it can sometimes lead to a faster increase in α
- the duals still remain monotonously increasing / decreasing, all of the above lemmas remain valid

9.3 Epsilon scaling

Remark 9.11.

- To obtain a good accuracy of the resulting assignment, a small value of ε is required.
- But this implies a potentially large number of iterations.
- We will now show: the algorithm can be run with a large value of ε first. Then decrease ε, delete the primal assignments but keep the current duals. The algorithm then converges faster.
- By choosing a suitable schedule for ε, this will effectively lead to replacing the factor C/ε in the complexity estimate by $M \log(C/\varepsilon)$ where ε is now the desired final value.
- This strategy is called epsilon scaling.
- Similar to the standard iteration bound, for the epsilon scaling bound we need to bound the maximal decrease of the dual variables β. This will then imply a bound on the maximal number of accepted bids. The new bound no longer depends on c, but on the fact that we already have found a solution for a larger value $\hat{\varepsilon}$.

Lemma 9.12.

- Let $\hat{\alpha}$ and $\hat{\beta}$ be dual feasible variables and let $\hat{\gamma}$ be a permutation matrix, such that $\hat{\gamma}$ and $(\hat{\alpha}, \hat{\beta})$ satisfy $\hat{\varepsilon}$-complementary slackness for some $\hat{\varepsilon} \geq 0$.
- If the auction algorithm is initialized with $(\alpha, \beta) = (\hat{\alpha}, \hat{\beta})$ (and $\gamma = 0$) and with some step parameter $\varepsilon > 0$, then

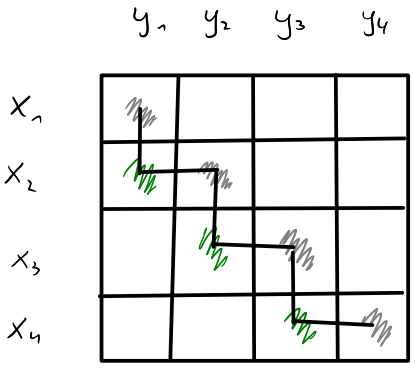
$$\alpha[x] \leq \hat{\alpha}[x] + M\,(\hat{\varepsilon} + \varepsilon), \qquad\qquad \beta[y] \geq \hat{\beta}[y] - M\,(\hat{\varepsilon} + \varepsilon)$$

  for all entries of α and β throughout the algorithm.

Proof.

- At any point during the algorithm let γ be the current partial assignment. Define a bipartite directed graph with vertex sets $X = \{x_1, \ldots, x_M\}$ and $Y = \{y_1, \ldots, y_M\}$ with an edge from x_i to y_j if $\hat{\gamma}_{i,j} = 1$, and an edge from y_j to x_i if $\gamma_{i,j} = 1$.
- (The notation $\{x_1, \ldots, x_M\}$ and $\{y_1, \ldots, y_M\}$ makes it easier to distinguish the two vertex sets.)
- Observe: there must always be a path from an unassigned x (unassigned under γ) to some unassigned y, because every x (unassigned and assigned) has an outgoing edge induced by $\hat{\gamma}$, and every assigned y has an outgoing edge induced by γ (which however can never lead to the initial x or some previously visited x).
- For some unassigned x, let $(x = x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$ be such a path, i.e.

$$\hat{\gamma}(x_i, y_i) = 1 \text{ for } i = 1, \ldots, n, \qquad\qquad \gamma(x_{i+1}, y_i) = 1 \text{ for } i = 1, \ldots, n-1.$$



$\hat{\gamma}$: $\hat{\gamma}_{ij} > 0 \Rightarrow c_{ij} - \hat{\alpha}_i - \hat{\beta}_j \leq \hat{\varepsilon}$ (old assignment)

γ : $\gamma_{ij} > 0 \Rightarrow c_{ij} - \alpha_i - \beta_j \leq \varepsilon$ (new, partial assignment)

- Recall: α and β are dual feasible throughout algorithm and satisfy ε-CS with γ. Use this to bound:

$$\alpha(x_i) \leq c(x_i, y_i) - \beta(y_i) \text{ for } i = 1, \ldots, n,$$
$$\beta(y_i) \geq c(x_{i+1}, y_i) - \alpha(x_{i+1}) - \varepsilon \text{ for } i = 1, \ldots, n-1$$

  combine these to get:

$$\alpha(x_1) \leq \sum_{i=1}^{n-1} [c(x_i, y_i) - c(x_{i+1}, y_i)] + c(x_n, y_n) - \beta(y_n) + (n-1) \cdot \varepsilon$$

- Similarly, use that $\hat{\alpha}$, $\hat{\beta}$ are dual feasible and satisfy $\hat{\varepsilon}$-CS with $\hat{\gamma}$ to bound:

$$\hat{\alpha}(x_i) \geq c(x_i, y_i) - \hat{\beta}(y_i) - \hat{\varepsilon} \text{ for } i = 1, \ldots, n,$$
$$\hat{\beta}(y_i) \leq c(x_{i+1}, y_i) - \hat{\alpha}(x_{i+1}) \text{ for } i = 1, \ldots, n-1$$

  combine these to get:

$$\hat{\alpha}(x_1) \geq \sum_{i=1}^{n-1} [c(x_i, y_i) - c(x_{i+1}, y_i)] + c(x_n, y_n) - \hat{\beta}(y_n) - n \cdot \hat{\varepsilon}$$

- Now use that y_n is still unassigned, i.e. it has not yet received a bid during the current run of the auction algorithm. Therefore, $\beta(y_n) = \hat{\beta}(y_n)$.

- Combining now the two bounds on $\hat{\alpha}(x_1)$ and $\alpha(x_1)$ we get:

$$\alpha(x_1) \leq \hat{\alpha}(x_1) + n \cdot \hat{\varepsilon} + (n-1) \cdot \varepsilon$$

- Here n is at most M. This bound holds as long as x_1 is unassigned, even after x_1 has submitted its final bid (we only used dual feasibility and complementary slackness, which still hold in this moment). Therefore, the bound holds also when the last bid was accepted and therefore at all stages throughout the algorithm (by monotonicity of α, see previous section). We get:

$$\alpha(x) \leq \hat{\alpha}(x) + M\hat{\varepsilon} + (M-1) \cdot \varepsilon$$

  This implies the claimed bound on α.

- For β: For $y \in Y$, let $x \in X$ be the partner that it is eventually assigned to after completion of the current run of the auction algorithm. By initial dual feasibility and final ε-CS we find:

$$\hat{\beta}(y) \leq c(x, y) - \hat{\alpha}(x), \qquad \beta(y) \geq c(x, y) - \alpha(x) - \varepsilon$$

  and by combination:

$$\beta(y) \geq \hat{\beta}(y) - [\alpha(x) - \hat{\alpha}(x)] - \varepsilon$$

  With the bound on α this implies the claimed bound. □

Lemma 9.13 (Bound on accepted bids)**.**

- With the initialization of the previous lemma, at most $(M-1)(\lfloor M\hat{\varepsilon}/\varepsilon \rfloor + M) + 1$ bids are accepted.

- The total complexity of the re-run of the auction algorithm is $O(M^4\hat{\varepsilon}/\varepsilon)$.

Proof.

- As in the last section: β is non-increasing, every accepted bid decreases the corresponding β by (at least) ε. Therefore $(M-1)$ elements of Y can accept at most $\lfloor M\hat{\varepsilon}/\varepsilon \rfloor + M$ bids before violating the given bound on β. At least one element of Y accepts only a single bid (the last one to be assigned).

- As before: for every accepted bid there may be $O(M)$ submitted ones. The cost of a submitted bid is $O(M)$ (this includes the cost of accepting). Thus, the total bound is $O(M^4\hat{\varepsilon}/\varepsilon)$. □

Remark 9.14 (ε-scaling)**.**

- If we want to solve an assignment problem with $c \in \mathbb{R}_+^{M\times M}$ with $C := \|c\|_\infty$ up to a complementary slackness precision of $\varepsilon > 0$, this can be done with a complexity of $O(M^4 \cdot \log(C/\varepsilon))$ by repeatedly using the auction algorithm (again, we are a bit sloppy in the notation of O and multiple variables).

- To achieve this, we initialize $\alpha = \beta = 0$ and repeatedly re-apply the auction algorithm with $\varepsilon^k = q^k \cdot C$ starting at $k = 1$, with some scaling factor $q \in (0, 1)$, until $\varepsilon^k \leq \varepsilon$.

- At $k = 0$, we know that any assignment is C-CS with $\alpha = \beta = 0$ by the bound on c, where $C = \varepsilon^0$. At any subsequent k we know that the previous γ,α,β satisfy ε^{k-1}-CS. So the complexity bound from the previous lemma applies at each stage with $\hat{\varepsilon}/\varepsilon = q^{-1}$.

- We need approximately $\lceil \log(C/\varepsilon)/\log(1/q) \rceil$ stages.

10 Entropic regularization

10.1 Regularized primal and dual problems

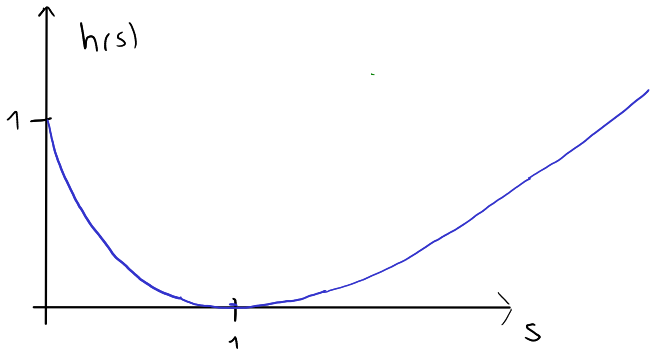
Remark 10.1 (Motivation)**.**

- add (negative) entropy of transport plan to objective
- advantage 1: leads to a very simple numerical algorithm that can be interpreted as a smooth auction algorithm. Can be parallelized and implemented on GPUs.
- advantage 2: primal optimal transport plan becomes unique, minimal value becomes differentiable as function of marginals:
  improved robustness in more complicated data analysis pipelines where OT is only a part
- advantage 3: related to improved sample complexity in higher dimensions, see for instance: Feydy et al. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences, AISTATS 2019

Definition 10.2 (Negative entropy)**.**

$$H(\gamma) := \sum_{i,j} h(\gamma_{i,j}), \qquad h(s) : \mathbb{R}_+ \to \mathbb{R}, \quad s \mapsto s\,\log(s) - s + 1$$

with the convention $0 \log 0 = 0$ such that $h(0) = 1$



- definition consistent at zero: h is continuous on $\mathbb{R}_+$ and therefore H is continuous on $\mathbb{R}_+^{M\times N}$
- motivation for 'extra terms' in h: non-negative, minimal value is $h(1) = 0$
- strictly convex:

$$h'(s) = \log s \text{ for } s > 0, \qquad h''(s) = 1/s \text{ for } s > 0$$

- high when γ is concentrated on few entries, low when diffuse (maximal values at the vertices of Γ, follows from strict convexity)
- careful about continuum limit, related notion: Kullback–Leibler divergence:

$$\mathrm{KL}(\rho|\sigma) := \begin{cases} \int h\left(\frac{\mathrm{d}\rho}{\mathrm{d}\sigma}\right) \mathrm{d}\sigma & \text{if } \rho \ll \sigma, \rho \geq 0, \\ +\infty & \text{else.} \end{cases}$$

or in discrete setting:

$$\mathrm{KL}(\rho|\sigma) := \begin{cases} \sum_i h\left(\rho_i/\sigma_i\right)\,\sigma_i & \text{if } ([\sigma_j = 0] \Rightarrow [\rho_j = 0] \text{ for all } j), \rho \geq 0, \\ +\infty & \text{else.} \end{cases}$$

KL is jointly convex and weak* lower-semicontinuous in both its arguments.

- Our choice above corresponds to $\sigma_{i,j} = 1$. $\sigma_{i,j} = \mu_i \cdot \nu_j$ is also common.

Definition 10.3 (Entropic primal problem)**.**

$$\min\left\{\langle c, \gamma\rangle + \varepsilon\, H(\gamma) | \gamma \in \Gamma(\mu, \nu)\right\}$$

where H is the negative entropy and $\varepsilon \geq 0$ is the regularization strength.

Proposition 10.4. The entropic primal problem has a unique minimizer.

Proof.

- H is continuous on $\mathbb{R}_+^{M\times N}$, so the primal objective is continuous on $\Gamma(\mu, \nu)$.
- $\Gamma(\mu, \nu)$ is compact (closed, bounded, see earlier arguments, Bolzano–Weierstrass) $\Rightarrow$ existence of minimizers
- uniqueness: assume γ_1 and γ_2 were two distinct minimizers. Then both have the same objective value. Let now $\gamma := \frac{1}{2}\gamma_1 + \frac{1}{2}\gamma_2$.
- Denote by E the objective, which is strictly convex since the linear term is convex and H is strictly convex. Then:

$$E(\gamma) < \tfrac{1}{2}E(\gamma_1) + \tfrac{1}{2}E(\gamma_2) = E(\gamma_1) = E(\gamma_2)$$

Hence, γ_1, γ_2 cannot be minimal. □

Remark 10.5 (Derivation of the dual problem)**.**

- as before, argue via Lagrangian: primal problem is equivalent to

$$\inf_{\gamma\in\mathbb{R}_+^{M\times N}} \sup_{\alpha\in\mathbb{R}^M, \beta\in\mathbb{R}^N} \langle c, \gamma\rangle + \varepsilon\, H(\gamma) + \langle \alpha, \mu - \mathrm{P}_X\gamma\rangle + \langle \beta, \nu - \mathrm{P}_Y\gamma\rangle$$

- swap order of minimization, reorder terms

$$\sup_{\alpha\in\mathbb{R}^M, \beta\in\mathbb{R}^N} \langle \alpha, \mu\rangle + \langle \beta, \nu\rangle + \inf_{\gamma\in\mathbb{R}_+^{M\times N}} \left[\langle c - \mathrm{P}_X^*\alpha - \mathrm{P}_Y^*\beta, \gamma\rangle + \varepsilon\, H(\gamma)\right]$$

- since H is acting 'entry-wise' on γ, the min can be performed for each entry of γ separately. let us solve the following sub-problem:

$$\inf_{s\geq 0} \psi \cdot s + \varepsilon\, h(s)$$

- try first order optimality condition:

$$0 = \partial_s[\psi \cdot s + \varepsilon\, h(s)] = \psi + \varepsilon \log(s) \quad \Rightarrow \quad s = \exp(-\psi/\varepsilon) > 0$$

this value lies in the region where h' is defined. by strict convexity of h this must be the unique minimizer. we get:

$$\inf_{s\geq 0} \psi \cdot s + \varepsilon \underbrace{h(s)}_{=s\log s - s + 1} = \psi\, \exp(-\psi/\varepsilon) + \varepsilon\, [\exp(-\psi/\varepsilon)\,(-\psi/\varepsilon) - \exp(-\psi/\varepsilon) + 1]$$
$$= -\varepsilon \cdot [\exp(-\psi/\varepsilon) - 1]$$

- back to full problem, arrive at regularized dual:

$$(\ldots) = \sup_{\alpha\in\mathbb{R}^M, \beta\in\mathbb{R}^N} \langle \alpha, \mu\rangle + \langle \beta, \nu\rangle - \varepsilon \sum_{i,j} \left[\exp\left(-\tfrac{c_{i,j}-\alpha_i-\beta_j}{\varepsilon}\right) - 1\right]$$

- discussion: the term $-\varepsilon \exp\left(-\frac{c_{i,j}-\alpha_i-\beta_j}{\varepsilon}\right)$ acts like a smooth approximation of the constraint $c_{i,j} - \alpha_i - \beta_j \geq 0$. if the constraint is violated, the term tends to $+\infty$ as $\varepsilon \to 0$. if the constraint is satisfied, the penalty tends to 0.
- the last term, $-\varepsilon \cdot (-1)$ is constant and tends to 0 as $\varepsilon \to 0$.
- so we have a smooth, unconstrained approximation of the original dual problem
- observe: still have the invariance under constant shifts of α and β, but by convexity of exp the objective is strictly concave up to these constant shifts, i.e. dual maximizers are unique up to these shifts

Proposition 10.6. Dual maximizers exist and are unique up to constant shifts.

Proof.

- assume for simplicity, $\mu \in \Sigma_M$, $\nu \in \Sigma_N$. if they do not have equal mass, the problem is not well-defined (or the optimal value is $+\infty$). if they do not have unit mass, the problem can be rescaled accordingly. assume μ and ν have strictly positive entries. zero entries can be eliminated beforehand.
- then we can rewrite the dual objective as:

$$\sup_{\alpha\in\mathbb{R}^M, \beta\in\mathbb{R}^N} \sum_{i,j} f_{i,j}(\alpha_i + \beta_j) \quad \text{with} \quad f_{i,j}(z) := z \cdot \mu_i \cdot \nu_j - \varepsilon \left[\exp\left(-\tfrac{c_{i,j}-z}{\varepsilon}\right) - 1\right]$$

- each $f_{i,j}$ is bounded from above and tends to $-\infty$ as $z \to \pm\infty$. hence, in a maximizing sequence of (α, β), all entries $\alpha_i + \beta_j$ must remain bounded.
- thus we can extract a subsequence where $\mathrm{P}_X^*\alpha + \mathrm{P}_Y^*\beta$ converges. by the invariance under constant shifts we can, for instance, fix $\alpha_1 = 0$ in the whole sequence. then, by convergence of $\alpha_1 + \beta_j$, β must converge, and then by the same argument also all entries of α
- since the objective is continuous, this limit must be a maximizer.

- uniqueness up to constant shifts: assume (α_1, β_1) and (α_2, β_2) are two maximizers that do not just differ by a constant shift. therefore, $\mathrm{P}_X^*\alpha_1 + \mathrm{P}_Y^*\beta_1 \neq \mathrm{P}_X^*\alpha_2 + \mathrm{P}_Y^*\beta_2$. since $f_{i,j}$ introduced above is strictly concave, this implies that the midpoint between the two maximizers must have a better score, which is a contradiction. □

Proposition 10.7 (Primal-dual optimality condition)**.** $\gamma \in \Gamma(\mu, \nu)$ and $(\alpha, \beta) \in \mathbb{R}^M \times \mathbb{R}^N$ are primal-dual optimal if and only if

$$\gamma_{i,j} = \exp\left(-\frac{c_{i,j} - \alpha_i - \beta_j}{\varepsilon}\right) \text{ for } i = 1, \dots, M,\ j = 1, \dots, N.$$

Proof.

- Recall the derivation of the dual problem, when we explicitly minimized over the entries of γ. We obtained:

  $$s \cdot \psi + \varepsilon\, h(s) \geq -\varepsilon[\exp(-\psi/\varepsilon) - 1]$$

  with equality if and only if $s = \exp(-\psi/\varepsilon)$.

- Now apply this to the primal dual gap:

  $$[\langle c, \gamma\rangle + \varepsilon\, H(\gamma)] - \left[\langle \alpha, \mu\rangle + \langle \beta, \nu\rangle - \varepsilon \sum_{i,j} \left[\exp\left(-\tfrac{c_{i,j}-\alpha_i-\beta_j}{\varepsilon}\right) - 1\right]\right]$$

  (now use $\gamma \in \Gamma(\mu, \nu)$, i.e. $\mathrm{P}_X\gamma = \mu$,...)

  $$= \sum_{i,j} \left[[(c_{i,j} - \alpha_i - \beta_j) \cdot \gamma_{i,j} + \varepsilon\, h(\gamma_{i,j})] + \varepsilon \left[\exp\left(-\tfrac{c_{i,j}-\alpha_i-\beta_j}{\varepsilon}\right) - 1\right]\right]$$

  (now for each i, j the corresponding term is of the above form, so we get:)

  $$\geq 0$$

  with equality if and only if $\gamma_{i,j} = \exp\left(-\frac{c_{i,j}-\alpha_i-\beta_j}{\varepsilon}\right)$ for all i, j. □

Proposition 10.8. As $\varepsilon \to 0$, minimizers of the primal entropic problem converge to a minimizer of the unregularized problem.

Proof.

- At $\varepsilon > 0$ denote by γ_ε the unique minimizer of the entropic problem with regularization strength ε.
- Since all γ_ε lie in $\Gamma(\mu, \nu)$, which is compact, we can extract a converging subsequence from $\{\gamma_\varepsilon | \varepsilon > 0\}$ with a cluster point γ_0.
- Since $\Gamma(\mu, \nu)$ is closed, $\gamma_0 \in \Gamma(\mu, \nu)$.
- Now recall that H is continuous on $\mathbb{R}_+^{M\times N}$ and bounded on $\Gamma(\mu, \nu)$. The former means that $\lim_{n\to\infty} H(\gamma_n) = H(\gamma)$ for converging sequences $(\gamma_n)_n$ with limit γ. The latter means that for any $\gamma \in \Gamma(\mu, \nu)$ we have $\lim_{\varepsilon\to 0} \varepsilon H(\gamma) = 0$.

- Let now $(\varepsilon_n)_n$ be a sequence of strictly positive regularization parameters, $\varepsilon_n \to 0$, such that $\gamma_{\varepsilon_n} \to \gamma_0$. (Exists by above compactness argument.) For any $\gamma \in \Gamma(\mu, \nu)$ we find:

$$\langle \gamma, c \rangle = \lim_{n\to\infty} \langle \gamma, c \rangle + \varepsilon_n H(\gamma) \geq \lim_{n\to\infty} \langle \gamma_{\varepsilon_n}, c \rangle + \varepsilon_n H(\gamma_{\varepsilon_n}) = \langle \gamma_0, c \rangle$$

where we used boundedness of entropy in the first equality, optimality of γ_{ε_n} in the second step, and boundedness of entropy in the third step. □

Remark 10.9 (Convergence as $\varepsilon \to 0$)**.**

- In the continuous setting this convergence is harder to prove, since H is unbounded and often $H(\gamma_0) = \infty$. Proof can be done via Γ-convergence.
- If the minimizer at $\varepsilon = 0$ is not unique, one can show that γ_0 is the minimizer which has the lowest (negative) entropy. So entropy regularization always selects a unique well-characterized minimizer which is useful in many applications.
- One can also show convergence of the optimal dual variables.

10.2 Sinkhorn algorithm

Remark 10.10 (Derivation as alternating dual maximization)**.**

- Consider now the dual objective for fixed β and optimize over α. The objective can be written as:

$$\sum_i \left[\mu_i \cdot \alpha_i - \varepsilon \exp(\alpha_i/\varepsilon) \cdot \sum_j \exp\left(-\tfrac{c_{i,j}-\beta_j}{\varepsilon}\right) \right] + \langle \beta, \nu \rangle + \varepsilon \sum_{i,j} 1$$

- So we can optimize over each α_i individually. Take partial derivative and set to zero:

$$0 = \mu_i - \exp(\alpha_i/\varepsilon) \cdot \sum_j \exp\left(-\tfrac{c_{i,j}-\beta_j}{\varepsilon}\right)$$

- Resolve for α, analogous formula for optimization over β:

$$\alpha_i = \varepsilon \log \left[\mu_i / \sum_j \exp\left(-\tfrac{c_{i,j}-\beta_j}{\varepsilon}\right) \right], \qquad \beta_j = \varepsilon \log \left[\nu_j / \sum_i \exp\left(-\tfrac{c_{i,j}-\alpha_i}{\varepsilon}\right) \right]$$

- Now, if we start with some initial $\alpha^{(0)}$, $\beta^{(0)}$, then generate $\alpha^{(1)}$ by optimizing over α, then $\beta^{(1)}$ by optimizing over β and keep on iterating, the update rule is given by:

$$\alpha_i^{(\ell)} := \varepsilon \log \left[\mu_i / \sum_j \exp\left(-\tfrac{c_{i,j}-\beta_j^{(\ell-1)}}{\varepsilon}\right) \right], \quad \beta_j^{(\ell)} := \varepsilon \log \left[\nu_j / \sum_i \exp\left(-\tfrac{c_{i,j}-\alpha_i^{(\ell)}}{\varepsilon}\right) \right]$$

for $\ell \geq 1$.

Remark 10.11 (Reformulation with scaling factors)**.**

- Define the matrix $k \in \mathbb{R}_{++}^{M \times N}$ via $k_{i,j} := \exp(-c_{i,j}/\varepsilon)$. Introduce the scaling factors $u^{(\ell)} \in \mathbb{R}_+^M$, $v^{(\ell)} \in \mathbb{R}_+^N$ via

$$u_i^{(\ell)} := \exp(\alpha_i^{(\ell)}/\varepsilon), \qquad v_j^{(\ell)} := \exp(\beta_j^{(\ell)}/\varepsilon).$$

- Then the above iterations for $\alpha^{(\ell)}$ and $\beta^{(\ell)}$ can be equivalently rewritten as

$$u_i^{(\ell)} := \frac{\mu_i}{\sum_j k_{i,j}\, v_j^{(\ell-1)}}, \qquad v_j^{(\ell)} := \frac{\nu_j}{\sum_i k_{i,j}\, u_i^{(\ell)}}.$$

- This can be compactly written as

$$u^{(\ell)} := \frac{\mu}{k.v^{(\ell-1)}}, \qquad v^{(\ell)} := \frac{\nu}{k^\top.u^{(\ell)}}.$$

  where the . denotes matrix-vector multiplication and the fraction of to vectors is two be understood entry-wise. This is the famous Sinkhorn algorithm and its main loop can be written in two lines in most scientific computing environments.

- Note that since $\alpha, \beta \in \mathbb{R}^M \times \mathbb{R}^N$, one has $u, v = \exp(\alpha/\varepsilon), \exp(\beta/\varepsilon) \in \mathbb{R}_{++}^M \times \mathbb{R}_{++}^N$ and also $k = \exp(-c/\varepsilon) \in \mathbb{R}_{++}^{M \times N}$, the division is always well-defined. However, numerically this may become an issue. We will address this later.

Proposition 10.12. The iterates generated by the Sinkhorn algorithm converge to a dual maximizer (up to constant shifts).

Proof.

- As long as the iterates change (by more than constant shifts), the dual objective is strictly increasing. If they do not change, by virtue of the first-order optimality conditions, we have found a dual maximizer.

- Argue as in the dual existence proof: the dual objective is bounded from above, its super-level sets are compact (up to constant shifts). Hence, up to constant shifts, the iterates must have converging subsequences.

- Since μ and ν are assumed to be strictly positive (otherwise, eliminate those rows and columns), the entries of α and β are always $> -\infty$ (or u and v are strictly positive). For such values, the iteration maps $\beta^{(\ell-1)} \mapsto \alpha^{(\ell)}$ and $\alpha^{(\ell)} \mapsto \beta^{(\ell)}$ are continuous. Hence, the cluster point must be a fixed-point of the iteration maps and therefore dual optimal.

- This must hold for all cluster points of the dual iterates. But since there is only one dual optimizer (after discarding constant shifts), the whole sequence must converge (up to shifts). □

Remark 10.13 (Corresponding primal sequence)**.**

- Recall the primal-dual optimality condition:

$$\gamma_{i,j} = \exp\left(-\frac{c_{i,j} - \alpha_i - \beta_j}{\varepsilon}\right) = u_i \cdot k_{i,j} \cdot v_j$$

So we can associate a sequence of primal iterates with the dual iterates. Note the following:

$$\sum_j u_i^{(\ell)}\, k_{i,j}\, v_j^{(\ell-1)} = \sum_j \frac{\mu_i}{\sum_{j'} k_{i,j'}\, v_{j'}^{(\ell-1)}} k_{i,j}\, v_j^{(\ell-1)} = \mu_i$$
$$\sum_i u_i^{(\ell)}\, k_{i,j}\, v_j^{(\ell)} = \sum_i \frac{\nu_j}{\sum_{i'} k_{i',j}\, u_{i'}^{(\ell)}} k_{i,j}\, u_i^{(\ell)} = \nu_j$$

- So after a u-update, the primal iterate satisfies the row-constraints, after a v-update it satisfies the column-constraints.
- The updates can be interpreted as re-scaling each row or column such that those constraints are satisfied.
- Since the map from dual to primal iterates is continuous, convergence of dual iterates implies convergence of primal iterates.
- By stationarity of the optimal duals (under further iterations), the limit of the primal iterates satisfies row and column constraints and therefore, by the primal-dual optimality condition, must be the unique primal minimizer.

Remark 10.14 (Stopping criterion)**.**

- The Sinkhorn algorithm virtually never converges exactly. When do we stop in practice?
- Various stopping criteria possible. Simplest choices: maximum number of iterations, L^1-error of marginals (or KL), step-size of dual variables.
- Numerical evaluation of primal-dual gap is unfortunately not possible, since primal candidate $\gamma_{i,j}$ does not lie in $\Gamma(\mu,\nu)$. (We will later discuss examples where this is possible.)

Remark 10.15 (Relation to auction algorithm)**.**

- For simplicity, let $M = N$, $\mu_i = \nu_j = 1$, i.e. we will solve a regularized assignment problem.
- The update for α is then given by:

$$\alpha_i^{(\ell)} = -\varepsilon \log \left[\sum_j \exp\left(-\tfrac{c_{i,j} - \beta_j^{(\ell-1)}}{\varepsilon} \right) \right]$$

- Now, for a vector $\psi \in \mathbb{R}^M$ consider the following operation:

$$\begin{aligned} C &:= -\varepsilon \log\big(\sum_i \exp(- \underbrace{\psi_i}_{\geq \min \psi} /\varepsilon)\big) \\ &\geq -\varepsilon \log\big(M \cdot \exp(-\min\psi/\varepsilon)\big) = \min\psi - \varepsilon \log M \end{aligned}$$

and similarly

$$C \leq -\varepsilon \log\big(\exp(-\min\psi/\varepsilon)\big) = \min\psi$$

- So we have: $\min\psi - \varepsilon \log M \leq C \leq \min\psi$

- C is therefore often called 'soft-min' if ψ with regularization strength ε.
- We find: Sinkhorn iterates are computing soft-min of $(c_{i,j} - \beta_j)_j$, i.e. a smooth version of the c-transform. For $\varepsilon \to 0$, the iteration map converges to the c-transform.
- We know: alternating c-transform (i.e. alternating maximization) of unregularized dual does not converge to minimizer. But it works on smoothed version. This is a bit similar to auction algorithm, where we avoid getting stuck. But this time not by explicit 'overstepping' the ridge, but by smoothing the ridge.
- Note: Sinkhorn algorithm is 'symmetric' in X, Y, whereas auction is not.
- Can anticipate: convergence may get slow as $\varepsilon \to 0$.

Remark 10.16 (Speed of convergence)**.**

- Franklin, Lorenz, Linear Algebra and its Applications, (1989): linear convergence of dual iterates to maximizer in Hilbert's projective metric. But: contraction ratio approaches 1 like $1 - \exp(-\|c\|_\infty/\varepsilon)$ as $\varepsilon \to 0$.
- Knight, SIAM. J. Matrix Anal. & Appl. (2008): local linearization of Sinkhorn iterations around dual solution, get better linear rates, but these are not so relevant in practice, since (at least for small ε) one usually has to start far from minimizer
- Schmitzer, SIAM J. Sci. Comput. (2019): convergence of an asymmetric ('auction-like') Sinkhorn algorithm in $O(1/\varepsilon)$ iterations (measured in L^1-error of primal iterate marginal constraints)
- Berman, Numerische Mathematik (2020): convergence of the Sinkhorn algorithm for the W_2 distance on the Torus in $O(1/\varepsilon)$ iterations, by showing that the iterates asymptotically follow a non-linear PDE
- ε-scaling very efficient in practice (at least on 'normal problems') but no proof for its efficiency yet (as far as I am aware)

10.3 Numerical tweaks for the Sinkhorn algorithm

Remark 10.17 (Rolling max for log-sum-exp)**.**

- Recurring problem in scientific computing / machine learning: compute

$$C = \varepsilon \cdot \log\left(\sum_j \exp(\psi_j/\varepsilon)\right)$$

  for some vector $\psi \in \mathbb{R}^M$ and $\varepsilon > 0$. this operation is sometimes called 'log-sum-exp'. problem: naive evaluation can become numerically unstable for small ε.
- absolute max and re-scaling: if $\max \psi$ is known, set $\Delta\psi_j := \psi_j - \max \psi$ and proceed as follows:

$$C = \varepsilon \cdot \log\left(\sum_j \exp((\max \psi + \Delta\psi_j)/\varepsilon)\right) = \max \psi + \varepsilon \cdot \log\left(\sum_j \exp(\Delta\psi_j/\varepsilon)\right)$$

since $\Delta\psi_j \leq 0$, the worst that can happen is 'underflow' and $\exp(\Delta\psi_j)$ becomes numerically zero for some j. But then, contributions of that index to the result is indeed negligible. And there is always at least one j (where the maximum is attained) where $\exp(\Delta\psi_j) = 1$ and thus the argument of the log remains a numerically stable number.
Main problem: if $\max \psi$ is not known, an additional pass through ψ is required to determine the maximum.

- alternative: rolling max. numerically represent positive (potentially very large) real number (redundantly) as $s_i = b_i \cdot \exp(e_i/\varepsilon)$. can compute addition in a stable way as follows:

$$s_1 + s_2 = b_1 \cdot \exp(e_1/\varepsilon) + b_2 \cdot \exp(e_2/\varepsilon) = b_3 \cdot \exp(e_3/\varepsilon)$$

b_3 and e_3 are not uniquely determined. we choose them as follows:

$$\begin{cases} \text{if } e_1 \geq e_2: & e_3 = e_1,\ b_3 = b_1 + b_2 \cdot \exp((e_2 - e_1)/\varepsilon), \\ \text{else}: & e_3 = e_2,\ b_3 = b_1 \cdot \exp((e_1 - e_2))/\varepsilon) + b_2 \end{cases}$$

in short: $e_3 = \max\{e_1, e_2\}$, $b_3 = \sum_{i=1}^{2} b_i \exp((e_i - e_3)/\varepsilon)$ (but numerically one should use the if-formulation above, uses less calls to exp)

- now apply this to exp sum. let

$$S_j := b_j \cdot \exp(e_j/\varepsilon) := \sum_{k=1}^{j} \exp(\psi_j/\varepsilon).$$

then

$$S_{j+1} = b_{j+1} \cdot \exp(e_{j+1}/\varepsilon) = S_j + \exp(\psi_{j+1}/\varepsilon)$$

and we set

$$e_{j+1} = \max\{e_j, \psi_{j+1}\}, \qquad b_{j+1} = b_j \cdot \exp((e_j - e_{j+1})/\varepsilon) + \exp((\psi_{j+1} - e_{j+1})/\varepsilon).$$

Finally, the result $C = \varepsilon \log(S_M)$ is given by $e_M + \varepsilon \log(b_M)$.

The following tweaks are described in detail in [Schmitzer, SIAM J. Sci. Comput. (2019)]. Code is available at https://bernhard-schmitzer.github.io/MultiScaleOT/.

Remark 10.18 (Alternative: 'stabilized' kernel matrix)**.**

- Entries of $k_{i,j} = \exp(-c_{i,j}/\varepsilon)$ are bounded from above (if $c_{i,j} \geq 0$), but they can be very small/close to zero.
- Upon convergence entries of $\gamma_{i,j} = u_i\, k_{i,j}\, v_j$ are bounded from above, but we can have $\gamma_{i,j} \approx 1$ where $k_{i,j}$ is exponentially small. So typically entries of u_i, v_j can be both very large and very small.
- Recall: optimal assignment is invariant under changing the cost row and column-wise. $\hat{c}_{i,j} := c_{i,j} - \hat{\alpha}_i - \hat{\beta}_j$ has same optimal coupling as $c_{i,j}$. This also holds with entropic regularization.

- Now set $\hat{\alpha}$ and $\hat{\beta}$ to optimal (entropic) dual solutions. Set $\hat{u} := \exp(\hat{\alpha}/\varepsilon)$, $\hat{v} := \exp(\hat{\beta}/\varepsilon)$. Then for the primal optimal solution γ we have:

$$\gamma_{i,j} = \hat{u}_i \, k_{i,j} \, \hat{v}_j = \exp\left(-\tfrac{c_{i,j}-\hat{\alpha}_i-\hat{\beta}_j}{\varepsilon}\right) = \exp(-\hat{c}_{i,j}/\varepsilon) = u_i \, \hat{k}_{i,j} \, v_j$$

  where we set $u_i = v_j = 1$ and $\hat{k}_{i,j} = \exp(-\hat{c}_{i,j})$.

- This means: if we knew optimal duals, by re-weighting the cost with them, the necessary scaling factors become just 1.

- Of course: chicken-and-egg problem: to know optimal duals we need to have solved problem already.

- Practical suggestion: estimate the re-weighting duals iteratively during the Sinkhorn algorithm:
  - Start with $\hat{\alpha}_i = \hat{\beta}_j = 0$, $u_i = v_j = 1$ and iterate u and v with respect to $\hat{k} = k$.
  - When entries of u or v become larger than some threshold, set

$$\hat{\alpha} \leftarrow \hat{\alpha} + \varepsilon \log(u), \qquad \hat{\beta} \leftarrow \hat{\beta} + \varepsilon \log(v), \qquad u \leftarrow 1, \qquad v \leftarrow 1,$$

    recompute $\hat{k}$ (and $\hat{c}$) from new $(\hat{\alpha}, \hat{\beta})$ and keep iterating with the re-set u,v.

- Advantage over rolling max: preserves matrix-vector multiplication structure of algorithm, can still use standard matrix libraries

- Disadvantage: not guaranteed to be stable. Even a single iteration might lead to numerical overflow in extreme cases.

Remark 10.19 (Epsilon-scaling)**.**

- Observe in numerical experiments: convergence gets slow as $\varepsilon \to 0$. But in many applications small ε is desirable (not in all!).

- Effective heuristic, similar to auction algorithm: start with large value of ε and gradually reduce it (for instance, whenever the stopping criterion is reached).

- Unfortunately no full proof for the effectiveness available yet.

- Small but relevant detail: when changing ε, do keep α and β constant, not u and v.

Remark 10.20 (Kernel truncation)**.**

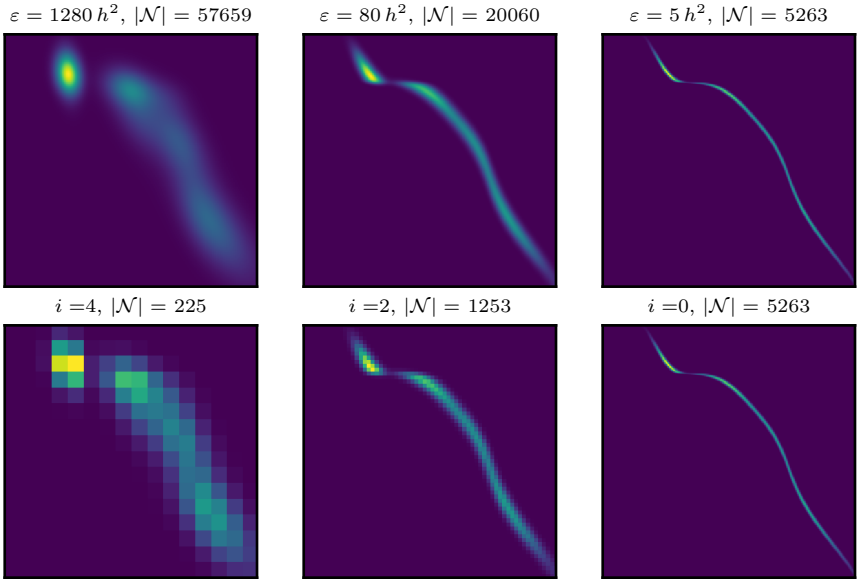
- On large problems storing the matrix k can be prohibitive, and matrix-vector multiplications $k.v$ become slow.

- Naive idea: many entries of k are exponentially small. Can we drop them and approximate k by a sparse matrix?

- Problem: entries of u, v can become exponentially large. Removing small entries of k corresponds to removing large entries in c. These could still be highly relevant in optimal transport plan.

- Solution: combine with kernel stabilization (see above).
- When u and v are re-set and $\hat{k}$ is re-computed, apply truncation. As long as entries in u and v are small, discarding small entries in $\hat{k}$ is harmless. When entries in u or v become large, re-set; recompute and re-truncate $\hat{k}$.
- If one works with truncated $\hat{k}$, it is advisable to use sparse matrix format, such as CSR, which are efficient for matrix-vector multiplication. Since we need to compute multiplications with $\hat{k}$ and $\hat{k}^\top$, it may be advisable to store two copies of the sparse matrix, one in transposed form. (This will still be more memory efficient than a single dense matrix.)

Remark 10.21 (Coarse-to-fine scheme)**.**



11 Mini-introduction: Convex optimization and duality

11.1 Convex functions

Remark 11.1 (Motivation)**.**

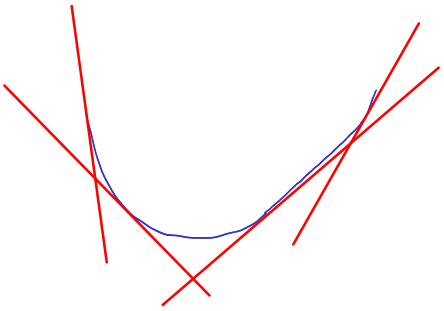
- In general: optimization is 'hard', can only do it locally, e.g. by following a gradient (in continuous problems).
- Will typically converge to some local minimum, little to no information about how well we have done in a global sense.
- Combinatorial optimization is also hard, a priori not even gradient descent is possible.
- Convex problems are the notable exception: every local minimum is also a global minimum.
- But there is more: convexity is a very strong global structural property, notion of convex duality: sub-optimality bounds.
- So: for convex problems there are efficient large scale algorithms
- But: convexity has its limits in modelling interesting behaviour. Probably not surprising that modern machine learning is largely built beyond convex methods. (But convexity can still be used for the analysis of some systems.)
- Many of the tricks and derivations, such as (generalized) Lagrange multipliers, we have used so far seem like 'individual clever tricks'. We will now learn that there is a system behind them.

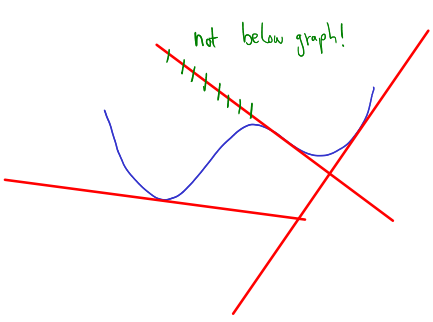
Definition 11.2 (Convex function)**.**

- A function $f : \Omega \to \mathbb{R}$ is convex if $\Omega \subset \mathbb{R}^d$ is convex and
$$f\big((1-\lambda)\cdot x + \lambda \cdot y\big) \leq (1-\lambda)\cdot f(x) + \lambda \cdot f(y)$$
for all $x, y \in \Omega$, $\lambda \in [0,1]$.
- Common convention: $f(x) := \infty$ where it is not defined (e.g. entropy for negative measures). Then convexity implies that the region where $f < \infty$ (usually called effective domain) is convex (which corresponds to Ω above).

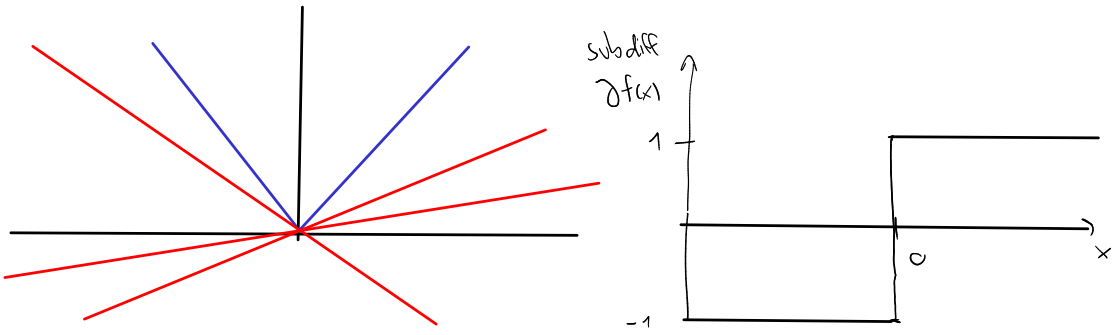
Definition 11.3 (Subdifferential)**.**

- Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. A vector $v \in \mathbb{R}^d$ is a *subgradient* at x if
$$f(y) \geq f(x) + \langle y - x, v\rangle \quad \text{for all} \quad y \in \mathbb{R}^d.$$
- The set of all subgradients at x is denoted by $\partial f(x)$.

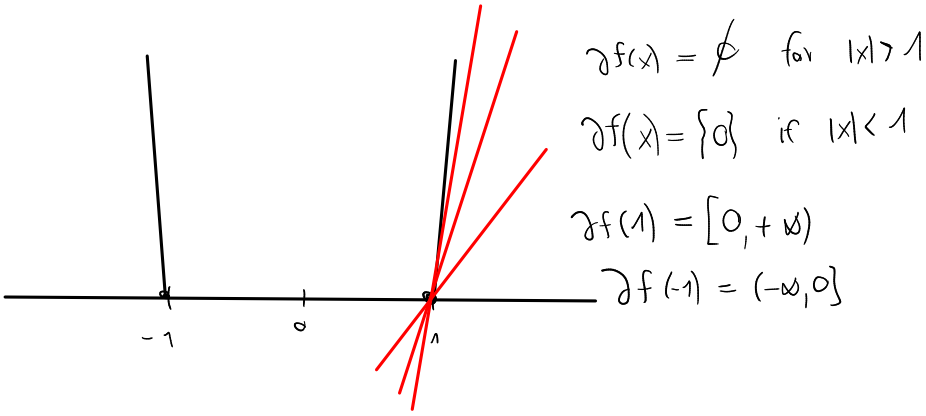




Example 11.4 (absolute value).



Example 11.5 (indicator of $[-1, 1]$).

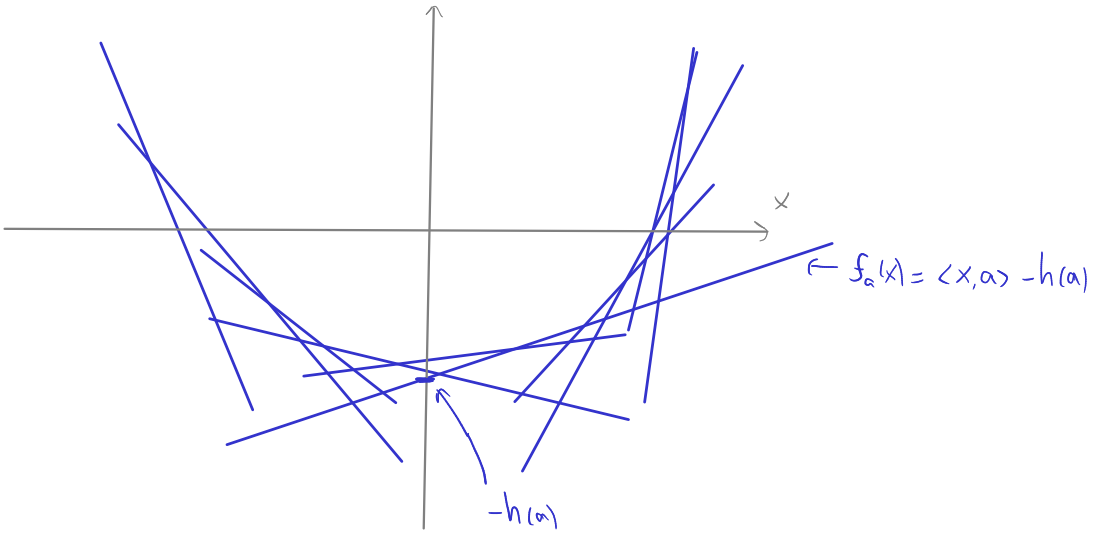


Remark 11.6 (Pointwise supremum over affine functions).

- Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a function that assigns to each slope a an offset $h(a)$. Then $f_a : x \mapsto \langle x, a \rangle - h(a)$ is an affine function with slope a, shifted by $-h(a)$.
- We can now define a function f as follows: for every x we take the supremum (maximum) over all functions f_a:

$$f(x) := \sup_{a \in \mathbb{R}^d} \langle x, a \rangle - h(a)$$

  We can set $h(a) = +\infty$ if we want to 'forbid' some slope a.



- One can show: functions built this way are always convex and lower-semicontinuous (lsc). And any convex, lower-semicontinuous function can be written this way, for a suitable h. How can we find this h?

$$\begin{aligned} -h(a) &= \max\{y \in \mathbb{R} \mid \langle x, a\rangle + y \le f(x) \,\forall\, x\} \\ &= \max\{y \in \mathbb{R} \mid \langle x, a\rangle - f(x) \le -y \,\forall\, x\} \\ h(a) &= \min\{y \in \mathbb{R} \mid \langle x, a\rangle - f(x) \le y \,\forall\, x\} \\ &= \min\{y \in \mathbb{R} \mid \sup_x \langle x, a\rangle - f(x) \le y\} \\ &= \sup_x \langle x, a\rangle - f(x) := f^*(a) \end{aligned}$$

- Note: if we have found $h = f^*$, then reconstructing f is done by $f = h^* = f^{**}$. This is a surprising function transformation which seems to be its own inverse (on the set of convex lsc functions).
- The transformation is called the 'Fenchel–Legendre conjugation'.

Example 11.7.

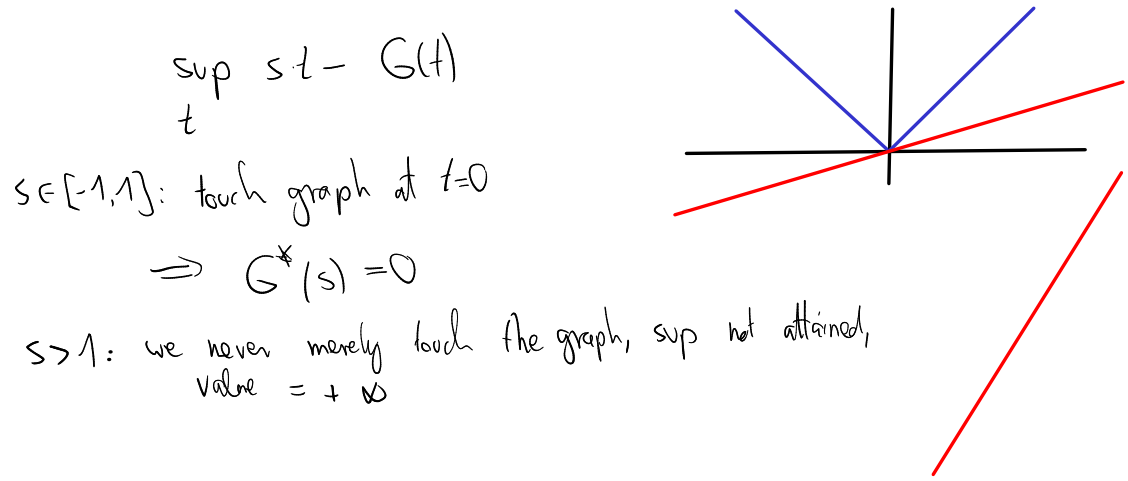
- Recall in min-cost flow problem: we used the function

$$H(s) = \begin{cases} 0 & \text{if } |s| \le 1, \\ +\infty & \text{else.} \end{cases}$$

  and we observed that it can be written as

$$H(s) = \sup_{t \in \mathbb{R}} s \cdot t - G(t) \quad \text{where} \quad G(t) = |t|.$$

- Now we can write this as $H = G^*$. And since G is convex and lsc we must have $H^* = G$.
- So the Fenchel–Legendre conjugation gives us a systematic way to derive the suitable G for given H.



Proposition 11.8 (Fenchel–Young inequality)**.**
Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex and lsc.

- $f(x) + f^*(y) \geq \langle x, y \rangle$ for all $x, y \in \mathbb{R}^d$.
- $[f(x) + f^*(y) = \langle x, y \rangle] \Leftrightarrow [y \in \partial f(x)] \Leftrightarrow [x \in \partial f^*(y)]$

Proof.

- By definition: $f^*(y) = \sup_{x'} \langle x', y \rangle - f(x') \geq \langle x, y \rangle - f(x)$. So we have the desired inequality with equality iff when x is optimal in the supremum. By symmetry, y is then optimal when computing $f = f^{**}$ from f^* by conjugation.
- Let now x be a point that attains the supremum for computing $f^*(y)$. Then:

  $$f^*(y) = \langle x, y \rangle - f(x) \geq \langle x', y \rangle - f(x') \quad \forall x'$$

  The inequality is equivalent to

  $$f(x') \geq f(x) + \langle x' - x, y \rangle \quad \forall x'$$

  and this is equivalent to $y \in \partial f(x)$.
- The relation $x \in \partial f^*(y)$ follows by the same argument with roles of f and f^* swapped. □

11.2 Fenchel–Rockafellar duality

Theorem 11.9 (Fenchel–Rockafellar)**.** Let $F : \mathbb{R}^e \to \mathbb{R} \cup \{\infty\}$, $G : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex, $A \in \mathbb{R}^{e \times d}$. Assume that there is a point $x_0 \in \mathbb{R}^d$ such that $G(x_0) < \infty$, $F(Ax_0) < \infty$, and F continuous at Ax_0. Then:

$$\inf_{x \in \mathbb{R}^d} F(Ax) + G(x) = \max_{y \in \mathbb{R}^e} -F^*(-y) - G^*(A^\top y)$$

and in particular there is a maximizer of the dual problem.

Remark 11.10 (Intuition for the form of the dual problem)**.**

$$\inf_{x\in\mathbb{R}^d} F(Ax) + G(x)$$

now write F as FL conjugate (pretend that F is lsc):

$$= \inf_{x\in\mathbb{R}^d} \sup_{y\in\mathbb{R}^e} \langle Ax, -y\rangle - F^*(-y) + G(x)$$

now as before, pretend that we can flip order of optimization (one last time)

$$\begin{aligned}
&= \sup_{y\in\mathbb{R}^e} -F^*(-y) + \inf_{x\in\mathbb{R}^d} [\langle Ax, -y\rangle + G(x)] \\
&= \sup_{y\in\mathbb{R}^e} -F^*(-y) - \sup_{x\in\mathbb{R}^d} [\left\langle x, A^\top y\right\rangle - G(x)] \\
&= \sup_{y\in\mathbb{R}^e} -F^*(-y) - G^*(A^\top y)
\end{aligned}$$

Sketch of proof for Theorem 11.9.

- For simplicity we only consider the case $e = d$, $A = \mathrm{id}$. Extension to general A can be done afterwards as additional step.
- If f and g were differentiable, we could look for a point x such that $f'(x) + g'(x) = 0$. By convexity this must be a global minimizer.
- In the non-differentiable, but convex setting, we could look for a point x such that there exists some y with

  $$-y \in \partial f(x), \qquad\qquad y \in \partial g(x).$$

  In this case we find for any $z \in \mathbb{R}^d$:

  $$f(z) + g(z) \geq f(x) + \langle z - x, -y\rangle + g(x) + \langle z - x, y\rangle = f(x) + g(x)$$

  $\Rightarrow x$ is primal optimal. And:

  $$f(x) + f^*(-y) = \langle x, -y\rangle, \qquad\qquad g(x) + g^*(y) = \langle x, y\rangle$$

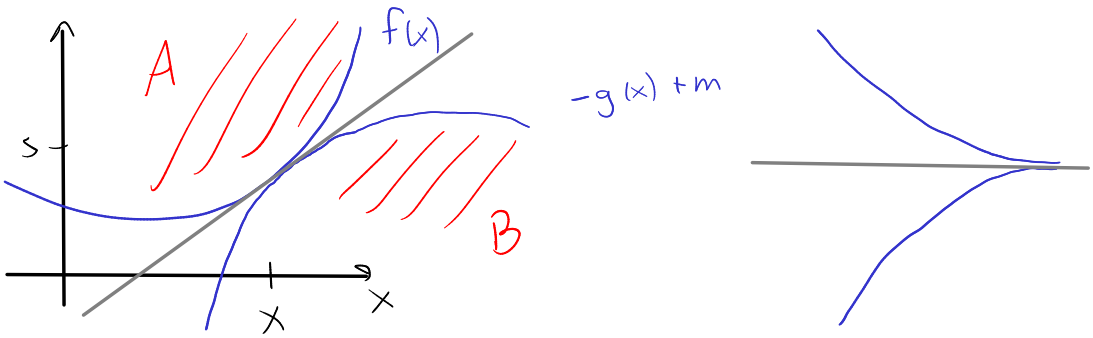
  and therefore

  $$-f^*(-y) - g^*(y) = f(x) - \langle x, -y\rangle + g(x) - \langle x, y\rangle = f(x) + g(x) \geq -f^*(-y') - g^*(y')$$

  for all $y' \in \mathbb{R}^d$. $\Rightarrow y$ is dual optimal.
- But: primal minimizer might not exist, and how do we find the right slope y? Let us prepare an idea:
- Let $m := \inf_x[f(x) + g(x)]$. If $m = -\infty$, then we must have $-f^*(-y) - g^*(y) = -\infty$ for all y and thus any y is a dual maximizer and the duality gap is zero. So assume $m > -\infty$.
- Consider now the following sets:

  $$A := \{(x, s) \in \mathbb{R}^{d+1} | s > f(x)\}, \qquad B := \{(x, s) \in \mathbb{R}^{d+1} | s < -g(x) + m\}$$

- Intuition: A is set above graph of f. B is set below graph of $-g$, then shifted upwards by m.



- We must have: $A \cap B = \emptyset$, since otherwise there would be some (x, s) such that

$$f(x) < s < -g(x) + m \qquad \text{and so} \qquad f(x) + g(x) < m$$

  which violates the definition of the infimum.

- On the other hand: cannot shift B any further upwards, without the two sets intersecting, since for any $m' > m$ we can find some x such that $f(x) + g(x) < m'$.

- A and B are disjoint convex sets, so there must exist a hyperplane that separates them (i.e. goes between them). (In infinite dimensions this is more complicated.)

- Now show that this provides the dual maximizer:

- Intuition: if surfaces of A and B are 'smooth', then slope of hyperplane must be given by $f'(x)$ and $-g'(x)$ at some point x where the sets 'almost touch'. By the above considerations this is then our dual maximizer.

- Rest of proof: use separation theorem for disjoint convex sets to get existence of separating hyperplane, use continuity assumption on F, G to show that it is not vertical, so that the 'slope' intuition is valid, and can then show by above estimates that the hyperplane gives a dual maximizer.

- Note: it can happen that no primal minimizer exists, when the closures of A and B do not touch. □

Example 11.11 (Kantorovich transport problem)**.**

Example 11.12 (Entropic transport problem)**.**

Proposition 11.13 (Generalized primal-dual optimality conditions)**.** $x \in \mathbb{R}^d$, $y \in \mathbb{R}^e$ are minimizers of primal-dual problem

$$\inf_{x \in \mathbb{R}^d} F(Ax) + G(x) = \sup_{y \in \mathbb{R}^e} -F^*(-y) - G^*(A^\top y)$$

(F, G convex) if and only if

$$[-y \in \partial F(Ax) \Leftrightarrow Ax \in \partial F^*(-y)] \quad \text{and} \quad [A^\top y \in \partial G(x) \Leftrightarrow x \in \partial G^*(A^\top y)]$$

Proof.

- We use the primal-dual gap and the Fenchel–Young inequality:

$$\begin{aligned}&[F(Ax)+G(x)]+[F^*(-y)+G^*(A^\top y)]\\&\qquad=\underbrace{[F(Ax)+F^*(-y)-\langle Ax,-y\rangle]}_{\geq 0}+\underbrace{[G(x)+G^*(A^\top y)-\left\langle x,A^\top y\right\rangle]}_{\geq 0}\geq 0\end{aligned}$$

- So we have zero if and only if both brackets become zero. By Fenchel–Young this is equivalent to the above conditions. □

Example 11.14 (Kantorovich transport problem)**.**

Example 11.15 (Entropic transport problem)**.**