Computational Optimal Transport

James Bannon

July 2020

Contents

1 Introduction

These are notes on optimal transport for my depth qualifying exam (DQE). They cover

1. The basics of optimal transport problems, including the Monge and Kantorovich formulations and their motivations.
2. The geometry of the problems and existence and uniqueness of solutions to these problems.
3. Numerical solutions to these problems for large-scale data science applicatons.
4. Extensions/Applications of the OT framework.

Obviously these concepts are highly interrelated and serve to illuminate the other. Main sources for the material are [11] and [5]. There is an appendix with foundational mathematical concepts included for reference. For measure theoretic probability we consult [14, 3, 4]. For functional analysis select chapters of [12, 13]. For background on optimization we consult [2, 8].

It is worth noting that we do not go too deeply into the theoretical side of optimal transport. Extensive mathematical treatises that go into significantly more depth can be found in [16, 1].These dependent on deeper convex analysis and would stray too far from focusing on algorithms and applications. Also, these full treatments show deeper mathematical connections and uses of the theory, as well as fairly general characterizations of the results. We treat them in as much generality as is required and as such some facts are to be taken on faith.

2 Fundamentals

Conceptualization

Optimal transport, in its basic form, is an alignment problem. Given two probability distributions a "source" μ and a reference or "target" ν we want a way to align μ with ν in a sense which we'll make precise.

A more intuitive view is something like this. You have a collection of m factories each with some amount of widgets, and n stores. Transporting widgets from a factory to a store incurs some cost c. You want to find a way to transport widgets such that:

- Each widget goes to a store
- The cost of the transportation is minimized.

Similar to how the computational tractability of linear programs depends on if we require integral solutions, the tractability and feasibility of the OT problem depends critically on how we are allowed to transport the widgets. The two formulations of what kind of "transport" is allowed are due to Monge and Kantorovich.

Before we dive deeper: an second intuitive approach due to [7] is worth addressing. If you had a pile of bricks in some arrangement that you want to move to a different location, you need two things: 1) a

destination for each brick and 2) a sense of how much effort (the cost) it would take to move each piece of dirt to the target location.

Before going in to the heart of the course it's worth noting some of the applications.

In its most general form the problem has the following ingredients:

- A pair of measure spaces $((X, \sigma(X), \mu), (Y, \sigma(Y), \nu)$ where $\sigma(\cdot)$ is a sigma algebra (see appendix A). We suppress the sigma algebra going forward.
- A cost function $c : X \times Y \to \mathbb{R}$ on which we make no assumptions at the present time.

we can work on three levels of generality. one one we have the abstract measures which for technical reasons may be Radon

The Monge Problem

The originator of the optimal transport problem was Gaspard Monge [9] took the brick analogy as inspiration. Mathematically, the problem was formulated as follows. Let $(X, \mathcal{E}, \mu)$ be a measure space. We see to solve the following optimization problem

$$\begin{aligned} \inf_{T:X\to Y} \quad & \int_X c(x, T(x)) f(x) d\mu \\ \text{subject to} \quad & T_\sharp \mu = \nu \end{aligned} \tag{M}$$

Troubles With Monge

A few obvious problems come from looking at the formulation of (M)

Kantorovich Relaxation

$$\begin{aligned} \inf_{T:X\to Y} \quad & \int_X c(x, T(x)) f(x) d\mu \\ \text{subject to} \quad & T_\sharp \mu = \nu \end{aligned} \tag{M}$$

2.0.1 Kantorovich Duality

2.1 Connections Between The Two

Suppose f is a transport *map*. Then we can get a transport plan by constructing:

$$\gamma = (\iota \times f)_\sharp \mu$$

aka

$$\gamma(A, B) = \mu(\{x \in A : f(x) \in B\})$$

where

$$\begin{aligned} (\iota \times f) : X &\to X \times Y \\ x &\mapsto (x, f(x)) \end{aligned}$$

this let's us write for any function $h(x, y)$ on $X \times Y$ given by

$$\int h(x, y) d\gamma(x, y) = \int h(x, f(x)) d\mu(x)$$

if f^* an optimal *map* then $(Id \times f^*)$ is an optimal plans and we have the expected lower bound $K(\mu, \nu) \leq M(\mu, \nu)$

$$\begin{aligned} \min_{\gamma} \quad & \sum_{i,j} c(x_i, y_j)\gamma_{ij} \\ \text{subject to} \quad & \gamma_{ij} \geq 0 \\ & \sum_j \gamma_{ij} = p_i \\ & \sum_i \gamma_{ij} = q_j \end{aligned} \tag{M}$$

From [16] we have that if $c(x, y)$ is l.c.s. (see Appendix D) and bounded from below, then the Kantorovic problem.

or equivalently

$$\begin{aligned} \min_{\gamma} \quad & \langle C, \gamma \rangle \\ \text{subject to} \quad & \gamma \geq 0 \\ & 1^T\gamma = p \\ & \gamma 1 = q \end{aligned} \tag{K}$$

Gangbo and McCann gives a result that if $c(x, y) = h(x - y)$ it's good and we have a correct solution

3 Algorithms: Sinkhorn Iterations

As pointed out above in the general case these are linear programs which can be plugged in to generic solvers (see appendix C)

4 Metric Properties

5 Miscellaneous Extensions & Variants

5.1 Barycenters & Clustering

The typical Barycenter problem is to take a collection of elements in a metric space and set of weights that sum to 1 and are non-negative computing the centroid $\min_{x \in \mathcal{X}} \sum_i \lambda_i d(x, x_s)^p$. For the Wasserstein case we have a group of measures (histograms) $\{b_s\}_{s=1}^S$ and cost matrices $C_s \in \mathbb{R}^{n \times n_s}$ we minimize the convex combination

$$\min_{a \in \Sigma_n} \sum_{s=1}^{S} \lambda_s L_{C_s}(a, b_s) \tag{1}$$

For $n_s = n$ and $C_s = D^p$ with D a distance matrix, then the barycenter problem becomes

$$\min_{a \in \Sigma_n} \sum_{s=1}^{S} \lambda_s W_p^P(a, b_s) \tag{2}$$

For histograms the general barycenter is an LP but too large to solv eiwth generic solvers. Generic subgradient descent on dual problem (Carlier et. al 2015). It's also possible to use the entropically smoothed L_C^ϵ which becomes a smooth convex minimization problem where gradient descent can be used.

5.2 Continuum Formulation

We can consider a continuum formulation of the 2 Wasserstein

5.3 Gradient Flows

This is essentialy a gradient-descent style scheme but with the Wasserstein metric:

$$a^{\ell} = \arg\min_{a} W_p(a, a^{(\ell-1)})^p + \tau F(a)$$

where F a function on histograms, e.g. entropy. Can be thought of as proximal point in Wasserstein space.

5.4 Unbalanced OT

This addresses cases where the two measures do not have the same total mass. In Liero 2018 we relax by penalizing the kantorovich using a Divergence D_ϕ

$$L_c^{\tau}(a,b) = \min_{(\tilde{a},\tilde{b})} L_C(a,b) + \tau_1 D_\phi(a,\tilde{a}) + \tau_2 D_\phi(b,\tilde{b})$$

Again we can generaliize the barycenter problems to L_C^{τ}

5.5 Gromov-Wasserstein Distances

If we can't pre-register spaces or if the histograms have supports on different metric spaces (e.g. suppose the histograms a, b have support on two graphs with the same set of vertices but different edge sets such that the shortest path metrics aren't the same. In tis instance for two matrices D, D' quantfifying simlarity between the two we have

$$GW\left((a,D),(b,D')\right) = \min_{\Pi(a,b)} E_{D,D'}(P) = \min_{\Pi(a,b)} \sum_{i,j,i',j'} |D_{i,i'} - D'_{j,j'}|^2 P_{i,j} P_{i'j'} \tag{3}$$

This is NP hard for arbitrary histograms/memasures. As hinted above, it is a graph matching problem. In generic settings we care about computing couplings between metric measure spaces $(X, d_x, \alpha_x), (Y, d_y, \alpha_y)$ here we have

$$GW((\alpha_x, d_x)(\alpha_y, d_y)) = \min_{\gamma \in \Pi(\alpha_x, \alpha_y)} \int_{X^2 \times Y^2} |d_x(x,x') - d_y(y,y')| d\gamma(x,y) \tag{4}$$

This is a metric space up to isometries where an isometry is ϕ bijection such that $\phi_\sharp \alpha_x = \alpha_y$ and $d(\phi(x), \phi(x')) = d(x,x')$

Also not numerically too stable and/or tractable.

5.6 Sliced Wassertein Distances

The idea behind these is to take one-parameter slices and aggregate. Specifically

$$SW(\alpha,\beta)^2 = \int_{S^d} W_2(P_{\theta,\sharp}\alpha, P_{\theta,\sharp}\beta) d\theta \tag{5}$$

As the 1-D Case is tractable for discrete measures we sort the points and compute

$$SW(\alpha,\beta)^2 = \int_{S^d} \left(\sum_i |\langle x_{\sigma(i)}, \theta\rangle - \langle y_{\kappa(i)}, \theta\rangle|^2 \right) d\theta \tag{6}$$

This can also be used to define distances and kernels for kpca etc. Kolouri papers, e.g. 2016

6 A Biological Case Study

7 Ongoing & Future Research Efforts

7.1 Clustering Immune Infiltrates

Current ongoing standard of care for the cancer patients is moving increasingly towards immune-based therapies. For solid tumors the nature of infiltrating immune cells is predictive of

7.2 Imputing Immune Response

In a pending collaboration with colleagues at the Hudson Alpha institute,

References

[1] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

[2] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[3] Marek Capinski and Peter E Kopp. *Measure, integral and probability*. Springer Science & Business Media, 2013.

[4] Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

[5] Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[6] Victor Klee and George J Minty. How good is the simplex algorithm. *Inequalities*, 3(3):159–175, 1972.

[7] Bruno Lévy and Erica L Schwindt. Notions of optimal transport theory and how to implement them on a computer. *Computers & Graphics*, 72:135–148, 2018.

[8] Jiri Matousek and Bernd Gärtner. *Understanding and using linear programming*. Springer Science & Business Media, 2007.

[9] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[10] Frank Nielsen. Legendre transformation and information geometry, 2010.

[11] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[12] James C Robinson. *An Introduction to Functional Analysis*. Cambridge University Press, 2020.

[13] Amol Sasane. *A friendly approach to functional analysis*. World Scientific, 2017.

[14] René L Schilling. *Measures, integrals and martingales*. Cambridge University Press, 2017.

[15] Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

[16] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

A Measure & Probability

A.1 Abstract Measure Theory

Definition A.1 (σ-algebra)**.** *For a set X a* **sigma algebra** *or* **σ-algebra** *on X is a collection $\mathcal{M}$ of subsets of x that satisfies:*

- *$\mathcal{M}$ is closed under countable unions*
- *$\mathcal{M}$ is closed under complementation*

It can be shown that this implies $X, \emptyset \in \mathcal{M}$.

Definition A.2 (σ-finite)**.** *A measure space (X, μ) is* **σ-finite** *if X can be expressed as a countable union of measurable sets of finite measure.*

Example A.3. *$\mathbb{R}$ is σ finite, as*

$$\mathbb{R} = \cup_{n\in\mathbb{N}}[-n, n]$$

Theorem A.4 (Product Measures)**.** *For two measure spaces $(X, \mathcal{E}, \mu), (Y, \mathcal{F}, \nu)$ we can define the product sigma algebra as the sigma algebra $\mathcal{E} \otimes \mathcal{F}$ as the σ algebra generated by*

$$\{E \times F : E \in \mathcal{E}, F \in \mathcal{F}\}$$

then there is a unique measure $\xi : \mathcal{E} \otimes \mathcal{F} \to [0, \infty]$ such that

$$\xi(E \otimes F) = \mu(E)\nu(F)$$

we sometimes write these as $\xi = \mu \times \nu$ or $d\xi = d\mu \times d\nu$ and the corresponding integrals are

$$\int_{X\times Y} f d\xi = \int_X \int_Y f(x, y)d\mu(x)d\nu(y)$$

in order to exchange the order of integration we need fubini's theorem:

Theorem A.5 (Fubini)**.** *Let $(X, \mathcal{E}, \mu), (Y, \mathcal{F}, \nu)$ be measure spaces and let f be measurable on $X \times Y$ and suppose either:*

- $f \geq 0$
- $f \in L^1(\mathbb{R}^2)$

then

$$\int_{X\times Y} f d\xi = \int_X \int_Y f(x, y)d\mu(x)d\nu(y) = \int_Y \int_X f(x, y)d\nu(x)d\mu(y)$$

Definition A.6 (Weighted Measure)**.** *For a non-negative measurable f the resulting weighted measure on $(X, \mathcal{E}, \mu)$ is given by*

$$d\nu = f d\mu$$

with

$$\nu(E) = \int_E f d\mu$$

Definition A.7 (Absolute Continuity/Radon Derivative)**.** *If $(X, \mathcal{E}, \mu)$ is a measure space and $\nu : \mathcal{E} \to [0, \infty]$ another measure. We say that ν is absolutely continuous with respect to μ if there exists a non-negative measurable function f on X such that*

$$d\nu = f d\mu$$

If this occurs we call f the **Radon-Nikodym Derivative** *and write*

$$f = \frac{d\nu}{d\mu}$$

Theorem A.8 (Radon Nikodym Theorem)**.** *$(X, \mathcal{E}, \mu)$ is a measure space and let $\nu : \mathcal{E} \to [0, \infty]$ be a measure. Then ν is absolutely continuous with respect to μ if andonly if for all $E \in \mathcal{E}$*

$$\mu(E) = 0 \implies \nu(E) = 0$$

Definition A.9 (Pushforward of a Measure)**.** *Let $(X, \mathcal{E}, \mu)$ b a measure space and let Y any set with $f : X \to Y$. The pushforward sgma algebra is*

1. *$f(\mathcal{E})$ is the collection*

$$\{S \subseteq Y : f^{-1}(S) \in \mathcal{E}\}$$

2. *The pushfoward of μ by f is the measure*

$$\nu(S) = \mu(f^{-1}(S))$$

Theorem A.10. *If $(X, \mathcal{E}, \mu)$ is a measure space and $f : X \to Y$ and ν the pushforward of μ by f then*

$$\int g d\nu = \int (g \circ f) d\mu$$

Definition A.11 (L^p spaces)**.** *For a measure space $(X, \sigma(x), \mu)$*

Definition A.12 (Inner/Outer Regular)**.** *A Borel measure μ on X with a Borel subset E. We say mu is* **outer regular** *if*

$$\mu(E) = \inf\{\mu(U) : E \subset U, U \text{ open}\}$$

and **inner regular** *if*

$$\mu(E) = \sup\{\mu(K) : K \subset E, K \text{ compact}\}$$

we say it is just **regular** *if it is both*

Definition A.13 (Radon Measures)**.** *A* **Radon Measure** *is a Borel measure that is*

1. *finite on all compact sets*
2. *outer regular on all borel sets, and inner regular on all open sets.*

that is a Radon measure is finite

B Functional Analysis

B.1 Metric Space Concepts

Let (X, d) be a metric space and $A \subseteq X$. We have

Definition B.1. *the interior of A is the union of all open sets contained in A, written A°. The closure is the intersection of all closed sets containing A, written $\overline{A}$*

We say a set in a metric space is **dense** if $\overline{A} = X$. The idea of density is that there's one "every where you look", a-la $\mathbb{Q} \in \mathbb{R}$. We say a metric space is **separable** if it has a countable dense subset.

The intuition of separability s that for a countable collection $\{x_j\}$ we can approximate any element of X arbitrarily well. $\mathbb{R}$ provides some intuitive. $\mathbb{Q}$ is a countable dense subset of $\mathbb{R}$, so we can think of $\mathbb{R}$ as being split into $\mathbb{Q}$ and "things that can be arbitrarily closely approximated by $\mathbb{Q}$" aka irrational numbers. A separable space has a countable dense set of wheat, and then some chaff.

B.2 Normed & Banach Spaces

any separable banach space is polish

C Optimization

C.1 Basics

The typical optimization problem has the form

$$\begin{aligned} \inf_{x \in \mathbb{R}^n} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0 \quad \forall i = 1, \ldots, n \\ & h_i(x) = 0 \quad \forall i = 1, \ldots, p \end{aligned} \tag{7}$$

where various constraints on the functional form yield different classes of problems which can be solved with varying degrees of algorithmic diffiiculty.

There problems have similar forms, including a change of variables and the introduction of slack variables.

We can always optimize over a subset of the variables $\inf_{x,y} = \inf_x \inf_y f(x,y)$.

$$\begin{aligned} \text{minimize} \quad & t \\ \text{s.t.} \quad & f_0(x) - t \leq 0 \\ & f_i(x) \leq 0 \quad \forall i = 1, \ldots, n \\ & h_i(x) = 0 \quad \forall i = 1, \ldots, p \end{aligned}$$

Another interesting point of view is the epigraph form

C.2 Convex Optimization

A convex optinmization problem is a problem in one of the many forms above where $f_0, \ldots, f_n$ are convex and the equality constraints are linear: $a_i^T x = b_i$. This implies that the problem domain $\mathcal{D} = \cap_0^n \text{dom} f_i$ convex.

C.3 Optimality Conditions for Diff'able convex objectives

A feasible point x is optimal iff for $\nabla f_0(x)^T (y - x) \geq 0$ for all feasible y.

Proof. If $x \in X$ satisfies the above then if $y \in X$ that $f_0(y) \geq f_0(x)$ by taylor expanding around x. COnversly if it's optimal and the condition does NOT hold then if we look at $z(t) = (ty + (1-t)x$ with $t \in [0,1]$. This whole line must b e feasible. And the claim is that we have $f_0(z(t)) < f_0(x)$ for some t. Differentiating $f(z(t))$ with respect to t and setting to zero gives $\nabla f_0(x)^T (y - x) < 0$ Thus, by the limiting definition of the derivative, we have for some small t that $f_0(z(t)) = f_0(x)$ □

C.4 Geometric Programming

Definition C.1 (Monomial, Posynomiial)**.** *A Monomial in the geometric programming sense is a function $f : \mathbb{R}^n \to R$ given by*

$$x \mapsto c x_1^{a_1} x_2^{a_2} \ldots x_n^{a_n}$$

and a posynomial is a sum of monomals

$$f(x) = \sum_i c_i x_1^{a_{i1}} x_2^{a_{i2}} \ldots x_n^{a_{in}}$$

This can be given a standard optimization form

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 1 \quad \forall i = 1, \ldots, n \\ & h_i(x) = 1 \quad \forall i = 1, \ldots, p \end{aligned} \tag{8}$$

where f_i are posynomials and h_i are monomials. This is called a geometryc program.

Example C.2.

$$\begin{aligned} \textit{minimize} \quad & x/y \\ \textit{s.t.} \quad & 2 \leq x \leq 3 \\ & x^2 + 3y/z \leq \sqrt{y} \\ & x/y = z^2 \end{aligned}$$

A series of transformations can easily turn this into standard form. Specifically if you have $f(x) \leq h(x)$ divide to get $f(x)/h(x) \leq 1$. If h_1, h_2 are non-zero monomials thatn $h_1(x) = h_2(x) \implies h_1(x)/h_2(x) = 1$ We can also maximize by minimizing the inverse.

C.4.1 Transforming to Convex Program

We can convert geometric prrograms to convex ones with a change of variables. Let $y_i = \ln x_i$ so a monomial objective becomes

$$f(x) = c(e^{y_1})^{a_1}(e^{y_2})^{a_2}\dots(e^{y_n})^{a_n} = e^{a^T y+b}$$

where $b = \ln c$. Extending to posynomials and repeating the we a convex form for geometric programs:

$$\begin{aligned} \text{minimize} \quad & \ln\left(\sum_{k=1}^{K_0} e^{a_{0k}^T y+b_{0k}}\right) \\ \text{s.t.} \quad & f_i(x) = \ln\left(\sum_{k=1}^{K_i} e^{a_{ik}^T y+b_{ik}}\right) \le 0 \quad \forall i = 1,\dots,n \\ & h_i(y) = g_i^T y + h_i = 0 \end{aligned} \tag{9}$$

where g_i contains the monomial constraints

C.4.2 Examples

Minimum Scaling. Suppose we want to scale a matrix M with a diagonal matrix D with $D_{ii} > 0$. If we want to minimize $||DMD^{-1}||_F^2$ we expand it out to a posynomial $\sum_{ij} M_{ij}^2 d_i^2/d_j^2$

C.5 Lagrangian Duality

Consider an optimization problem of the form (7). We can construct a function by adding a weighted sum of the constraint functions to get the **Lagrangian** associated to the problem as

$$\begin{aligned} L :& \mathbb{R}^N \times \mathbb{R}^m \times \mathbb{R}^p :\to \mathbb{R} \\ (x,\lambda,\nu) &\mapsto f_0(x) + \sum_j \lambda_j f_j(x) + \sum_k \nu_k h_k(x) \\ &= f_0(x) + \lambda^T \boldsymbol{f} + \nu \boldsymbol{h} \end{aligned} \tag{10}$$

To get a grounding intuition, note that if we're feasble, the h_i's don't impact minimizing L and for $\lambda_i \ge 0$ for feasible x the second term in the sum is at most zero. Another way to view this is that we impose a COST to violating the constraints rather than making it a forced impossibility. Specifically, if we satisfy the constraints and $\lambda_i \ge 0$ then minimizing L, for any feasible x it lets us get an essentially equivalent problem to the orginal. To clarify define the lagrangian dual function as

$$g(\lambda,\nu) = \inf_{x\in\mathcal{D}} L(x,\lambda,\nu) \tag{11}$$

There is a connection to conjugate functions (see Appendix D). Consider minimizing $f(x)$ subject to $x = 0$ then $g(\nu) = -\sup f^*(-\nu)$. Doin the same for a linear inequality and equality constraint $Ax \le b, Cx = d$ we get

$$g(\lambda,\nu) = -b^T\lambda - d^T\nu - f_0^*(-A^T\lambda - c^T\nu)$$

. Thus looking at the domain of f_0^* helps us think through the domain of g Suppose we have $\lambda \ge 0$ then we we have $g(\lambda,\nu \le p^*$, where p^* is the optimal value of (7).

Proof. For feasible x, non-neg λ we have

$$\lambda^T \boldsymbol{f} + \nu \boldsymbol{h} \le 0$$

so $g(\lambda,\nu) \le f(x)$ for any feasible x □

The dual problem becomes

$$\begin{aligned} \sup \quad & g(\lambda,\nu) \\ \text{s.t.} \quad & \lambda \ge 0s \end{aligned}$$

Weak duality gives us that for optimal d^* for the dual we have

$$d^* \leq p^*$$

Strong duality holds when this is an equality. One way to ensure strong dualty holds is called constraint qualificaton. We have Slater's condtions for convex problems:

Theorem 1 (Slater condition)**.** *for convex problems if there is a strictly feasible x we have strong duality*

and a more relaxed one

Theorem 2 (Relaxed Slater condition)**.** *for convex problems if there is a strictly feasiblif the first k in equality constraints are affine, then we get strong duality if the feasibility is strict only on $f_{k+1} \ldots, f_m$*

Thhis has that strong dualty always holds for LPs

C.5.1 Optimality Conditions

Suppose we have a dual feasible pair (λ, ν) and a primal feasible x. By weak duality we can certify a lower bound. Strong duality gives arbitrarily good certificates. If we set $f_0(x) - p^* \leq f_0(x) - p^*$. We call $f_0(x) - p^*$ the **dualty gap**. In the case of strong duality, we have optimal $x^*, (\lambda^*, \nu^*)$ so

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, \nu^*) \\ &\leq \inf L(x, \lambda^*, \nu^*) \\ &\leq f_0(x^*) \end{aligned}$$

Since we have strong duality, the last inequality is an equality so we have

$$\sum \lambda_i^* f_i(x^*) = 0$$

which means we have

$$\begin{aligned} \lambda_i > 0 &\implies f_i(x^*) \\ f_i(x^*) < 0 &\implies \lambda_i^* = 0 \end{aligned}$$

which means that the i'th multiplier is zero unless the ith constraint is active.

another option is the KKT conditions for differentiable non-convex which is that x^* feasible we need $\lambda_i^* f_i(x^*) = 0$ and that the gradient of L at x^*, λ^*, ν^* is zero

C.6 Linear Programming

can be thought of as a case where the constraint and objective functions all have only one term then we get a linear program. Suppose we want to optimize the function $x_1 + x_2$ on $\mathbb{R}^2$. This clearly has no fixed maximum because we haven't imposed boundaries on (x_1, x_2), that is, we can *always* increase the sum. Suppose we add linear inequality constraints, e.g. $x_1 + x_2 \leq 3$, $x_1 + 5x_2 \leq 19$ and require that $x_1, x_2 \geq 0$. Suppose in addition we add additional weights c_1, c_2 for x_1 and x_2 respectively. This gives us the following optimization problem

$$\begin{aligned} \text{minimize} \quad & c_1 x_1 + c_2 x_2 \\ \text{s.t.} \quad & x_1 + x_2 \leq 3 \\ & x_1 + 5x_2 \leq 19 \end{aligned}$$

which can be rewritten

$$\begin{aligned} \text{minimize} \quad & c^T x \\ \text{s.t.} \quad & Ax \leq b \end{aligned} \tag{12}$$

with $A = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix}, b = (3, 19)$ and the inequality is understood to be component-wise. Equation (12) is an example of a class of optimization programs called **linear programs** (LPs), so named because the objective

function $f(x) = c^T x$ and the constraints $Ax \leq b$ are all linear. In [8] they focus on the so-called equational form LPs

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & c^T x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned} \tag{13}$$

which is different from the so-called 'standard form'

$$\begin{aligned} \max_{x \in \mathbb{R}^n} \quad & c^T x \\ \text{s.t.} \quad & Ax \leq b \end{aligned} \tag{14}$$

It is worth noting that in [2] they approach linear programming from the point of view of minimization, which is equivalent (we can negate the objective) Thus look at

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & c^T x \\ \text{s.t.} \quad & Ax \leq 0 \\ & x \geq 0 \end{aligned} \tag{15}$$

C.7 Two Routes to LP Duality

C.7.1 Lagrangian

If we look at eq. (15) we can define the **Lagrangian** as the function L given by

$$L(x, \lambda, \eta) := c^T x + \lambda^T A x + \nu^T x$$

taking the lagrange dual gives

$$g(\lambda, \nu) = \begin{cases} -b^T \boldsymbol{\nu} & A^T \boldsymbol{\nu} - \boldsymbol{\lambda} + c = 0 \\ -\infty & \text{o.w.} \end{cases}$$

Thus creating the dual problem gives

$$\begin{aligned} \max_{\boldsymbol{\nu}} \quad & -b^T \boldsymbol{\nu} \\ \text{s.t.} \quad & A^T \boldsymbol{\nu} + c \geq 0 \end{aligned}$$

C.7.2 A Simpler LP-Specific Path

Consider the following forms of the Farkas lemma

Theorem 3 (Farkas Lemma)**.** *Let A be an $m \times n$ matrix and $b \in \mathbb{R}^m$. We have the following equivalent formulations:*

- *Exactly one of the following possibilities occurs:*
  1. *There exists a vector $x \in \mathbb{R}^n$ such that $Ax = b, x \geq 0$*
  2. *There exists a vector $y \in \mathbb{R}^m$ st. $y^T A \geq 0, y^T b \geq 0$*
- *The system $Ax = b$ has a non-negative solution iff every $y \in \mathbb{R}^m$ with $y^T A \geq 0$ also satisfies $y^T b \geq 0$*
- *The system $Ax \leq b$ has a non-negative solution iff every non-negative $y \in \mathbb{R}^m$ with $y^T A \geq 0$ also satisfies $y^T b \geq 0$*
- *The system $Ax \leq b$ has a non-negative solution iff every non-negative $y \in \mathbb{R}^m$ with $y^T A = 0$ also satisfies $y^T b \geq 0$*

We use this as follow.s. Suppose z is optimal and $c^T x = \gamma$. Then $Ax \leq b, c^T x \geq \gamma$ has a non-negative solution. Then $Ax \leq b, c^T x \geq \gamma + \epsilon$ has NO non-negative solutions. If we make the block matrices/vectors

$$\hat{A} = \begin{pmatrix} A \\ -c^T x \end{pmatrix}$$

and

$$\hat{b} = \begin{pmatrix} b \\ -\gamma - \epsilon \end{pmatrix}$$

Thus for $\epsilon > 0$ $\hat{A}x \leq \hat{b}$ has no non-negative solutons, so by the farkas lemma we have a non-negative vector $y = (u, z) \in \mathbb{R}^{m+1}$ such that $y^T \hat{A} \geq 0$. But $y^T \hat{b} < 0$ so $A^T u \geq zc$ and $b^T u < z(\gamma + \epsilon)$ For $\epsilon = 0$ we have that the same vector must have $y^T \hat{b} \geq 0$ which is the same as $b^T u \geq z\gamma$. Thus $z > 0$ bc $z = 0$ would contradict the strict equality $b^T u < z(\gamma + \epsilon)$. But then $v = \frac{1}{z} u \geq 0$ which means $A^T u \geq 0$ and $b^T v < \gamma + \epsilon$, So V is dual feasible with a value smaler than $\gamma + \epsilon$. By weak duality, we have $b^T y^*$ is between γ and $\gamma + \epsilon$ for all epsilon, so it holds.

C.8 LP Duality and Algorithms

C.8.1 Duality

Strong duality always holds for linear programs with one pathological exception. Exhaustively our optons are:

1. Neither the primal nor the dual has a feasible solution.
2. the primal is unbounded and the dual has no feasible.
3. the primal has no feasible solution and the dual is unbounded.
4. both have feasible solution and strong duality holds

Exactly one of the above occurs.

C.8.2 Algorithms

There are three algorithms typically used to solve LPs:

1. Simplex method. Exponential in worst case for deterministic rules [6] but if we look at randomized implementation/smoothed analysis we get expected polytime [15]
2. Interor point (Karmarkar) $O(n^{3.5})$
3. Ellipsoid method. Not numerically stable but polytime $O(n^3)$

C.9 ILPs & Relaxations

In many real world applications arbitrarily divisible objects are not available. Thus we might might want to enforce the solutions to eq. (13) or eq. (14) to be integral. This additional constraint results in so-called **Integer Programs** or **Integer Linear Programs** (ILP) which are of the form

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t.} \quad & Ax \leq b \\ & x \in \mathbb{Z}^n \end{aligned} \tag{16}$$

which are in general NP-hard. One approach to solving these is to construct the so-called LP relaxation which involves changing the constraint $x \in \mathbb{Z}^n$ to $x \in \mathbb{R}^n$ which just spits out a linear program. It's easy to see that the solution to the linear program is an upper bound on the solution to the integer program as it is

maximizing over a larger set. *The step from ILP to LP is relaxing the optimization variables to allow them to be infinitely divisible.*

It's sometimes possible that the LP will have an optimal integral solution or that one can be constructed, however that's not always possible and simple schemes like rounding may fail catastrophically. However, the following lemma is true due to the fact that when we solve an LP relaxation we solve over a larger set.

Theorem C.3. *If p^* is a solution the the linear programming relaxation of an integer program and q^* is the optimal solution to the integer program, then*

$$q^* \leq p^*$$

Theorem C.4 (Weak Duality for LPs)**.** *if we have the two problems*

$$\begin{aligned} \min \quad & c^T x \\ s.t. \quad & Ax \leq b \\ & x \geq 0 \end{aligned}$$

and

$$\begin{aligned} \max \quad & b^T y \\ s.t. \quad & A^T y \leq c \\ & y \geq 0 \end{aligned}$$

Then we have $b^T y \leq c^T x$

Proof. Note that $y \geq 0, Ax - b \geq 0$ so

$$y^T b \leq y^T Ax = (A^T y)^T x \leq c^T x$$

□

D Convexity, Conjugacy, & Continuity

D.1 Convexity

We assume familiarity with the basic notions of convexity:

Definition D.1 (Convex, Strictly Convex)**.** *A function $f : \mathcal{X} \to \mathbb{R}$ is convex if*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \tag{17}$$

for all $x, y \in \mathcal{X}$ and $t \in [0, 1]$

as well as related first- and second-order definitions for differentiable functions. Convex optimization is often focused on minimizing problems like Equation (13) where the linearity of the objective and constraints is derived from

D.2 Continuity

Aside from classical (analytic, topological) concepts of continuity it's worth noting the following two notions of *semi-continuity*

Definition D.2 (Semi-Continuity)**.** *A function $f : D \to \mathbb{R}$ is said to be* **lower semi-continuous (lsc)** *at a point $x_0 \in D$ if for every $\epsilon > 0$ we have a $\delta > 0$ such that*

$$f(x_0) - \epsilon < f(x)$$

for all $x \in B_\delta(x_0) \cap D$. Similarly f is **upper semi-continuous (usc)** *if*

$$f(x) < f(x_0) + \epsilon$$

for all $x \in B_\delta(x_0) \cap D$.

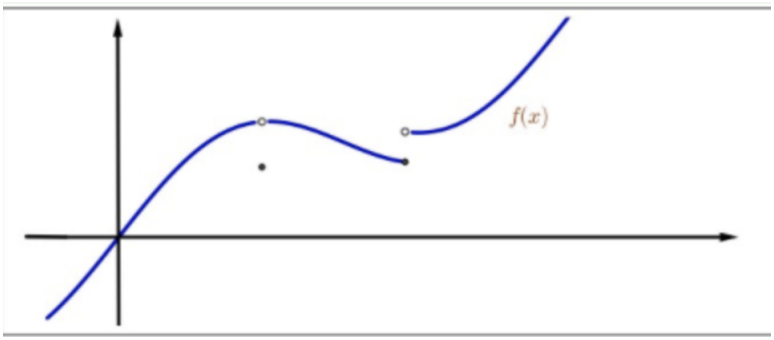

Figure 1: Lower semi contiuous at filled in points. The function completely upper bounds the values at the filled in poiints but with arbitrary jumps.

The intuition for this definition is something like this: if f is lsc (or usc) at some point x_0 then it only keeps one half of the definition of continuity. Specifically, for any real number $v < f(x_0)$ we can re-write this as $v = f(x_0) - \epsilon$ and then we can over-approximate v if we get close enough to x_0. Having $f : \mathbb{R} \to \mathbb{R}$ makes this clearer. If f is continuous at x_0 then for all $\epsilon > 0$ we have

$$x_0 - \delta < x < x_0 + \delta \implies f(x_0) - \epsilon < f(x) < f(x_0) + \epsilon$$

thus for lsc/usc we relax one of the inequalities on the right. Thus lsc functions are always "close enough" only from below. The "upward jump" of f near $f(x_0)$ can be as large as we please, but we can get a 'floor' very near $f(x_0)$. The following lemma makes this more precise:

Theorem D.3. *f is LSC at x_0 iff*

$$\liminf_{x \to x_0} f(x) \geq f(x_0)$$

and f is USC at x_0 iff

$$\limsup_{x \to x_0} f(x) \leq f(x_0)$$

Proof. fix $\epsilon > 0$. For all $x \in B_\delta(x_0)$ we have $f(x_0) - \epsilon$ as a lower bound for $f(x)$ and thus

$$f(x_0) - \epsilon < \inf_{x \in B_\delta(x_0)} f(x)$$

by definition

$$\liminf_{x \to x_0} f(x) = \sup_{\delta > 0} \inf_{x \in B_\delta(x_0)} f(x) \geq \inf_{x \in B_\delta(x_0)} f(x) \geq f(x_0) - \epsilon$$

so the result holds bc ϵ is arbitrary. To show the converse suppose

$$\liminf_{x \to x_0} f(x) = \sup_{\delta > 0} \inf_{x \in B_\delta(x_0)} f(x) \geq f(x_0)$$

By the definition of sup for all $\epsilon > 0$ we have a δ such that

$$\inf_{x \in B_\delta(x_0)} > f(x_0) - \epsilon$$

so $f(x) > f(x_0) - \epsilon$ for all $x \in B_\delta(x_0)$. □

D.3 Fenchel Conjugacy

Definition D.4 (Fenchel Transform/Conjugate)**.** *For $f : \mathbb{R}^n \to \mathbb{R}$ the* **fenchel conjugate** *$f^*(y) : \mathbb{R}^n \to \mathbb{R}$ is given by*

$$f^*(y) = \sup_{x \in dom(f)} \left\{ x^T y - f(x) \right\}$$

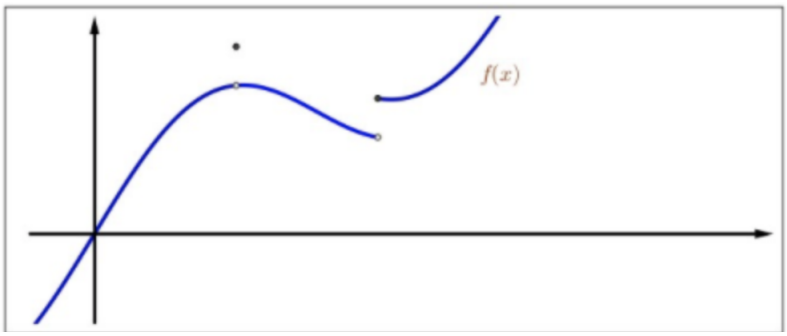

Figure 2: upper semi continuos

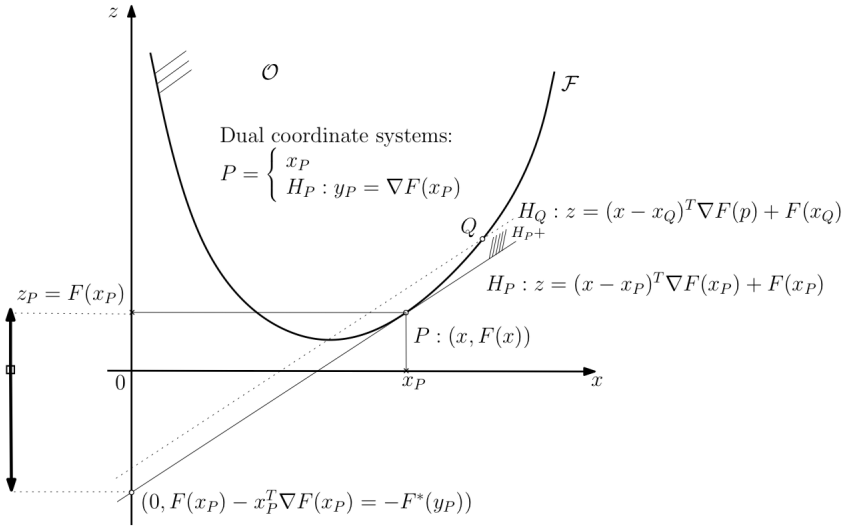

Figure 3: Illustration of the Legendre-Fenchel Transform as Dual Coordinates from [10]

f^* is always convex as it's pointwise sup of affine functions of y. Examples: for $f(x) = ax+b$, $f^*(y) = -b$ with singleton domain a. For $f(x) = -\log x$, $f^*(y) = -\log(-y) - 1$ for $y < 0$. For $f(x) = e^x$, $f^*(y) = y - y$. $f(x) = 1/x$ gives $f^*(y) = -2(-y)^{1/2}$ for $y \leq 0$ Suppose f is a strictly convex function with domain $\mathcal{X}$. We want to somehow create a description of the function f. One option is just its graph

$$\mathcal{F} := \{(x, f(x)) : x \in \mathcal{X}\}$$

but since

Example D.5 (C)**.**

D.4 c-Transforms

This is a way to generalize fenchel transforms to metric spaces:

Definition D.6 (c-transform)**.** *For any $f : Y \to \mathbb{R} \cup \{-\infty\}$, the c-transform is defined by*

$$f^c(x) = \inf_{y \in Y} \{c(x, y) - f(y)\}$$

for $x \in X$