Weverthon Lobo de Oliveira

Acoplamentos: Uma primeira visão e algumas aplicações

Vitória - Espírito Santo, Brasil
24 de junho de 2016

Weverthon Lobo de Oliveira

Acoplamentos: Uma primeira visão e algumas aplicações

Dissertação apresentada ao Programa de Pós-graduação em Matemática da Universidade Federal do Espírito Santo PPGMAT/UFES, como parte dos requisitos para obtenção do título de Mestre em Matemática. Orientador: Profº Fábio Júlio da Silva Valentim .

Universidade Federal do Espírito Santo – UFES
Departamento de Matemática
Programa de Pós-Graduação em Matemática

Orientador: Fábio Júlio da Silva Valentim

Vitória - Espírito Santo, Brasil
24 de junho de 2016

Dados Internacionais de Catalogação-na-publicação (CIP)
(Biblioteca Central da Universidade Federal do Espírito Santo, ES, Brasil)

Oliveira, Weverthon Lobo de, 1991-
O48a Acoplamentos : uma primeira visão e algumas aplicações / Weverthon Lobo de Oliveira. – 2016.
82 p. : il.

Orientador: Fábio Júlio da Silva Valentim.
Dissertação (Mestrado em Matemática) – Universidade Federal do Espírito Santo, Centro de Ciências Exatas.

1. Probabilidades. 2. Acoplamentos. 3. Markov, Processos de. 4. Pesquisa operacional. I. Valentim, Fábio Júlio da Silva. II. Universidade Federal do Espírito Santo. Centro de Ciências Exatas. III. Título.

CDU: 51

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
Centro de Ciências Exatas
Programa de Pós-Graduação em Matemática

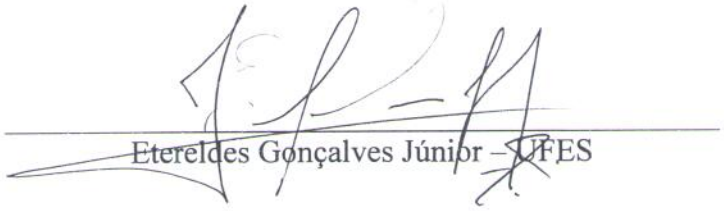
"Acoplamentos: Uma Primeira Visão e Algumas Aplicações"

Weverthon Lobo de Oliveira

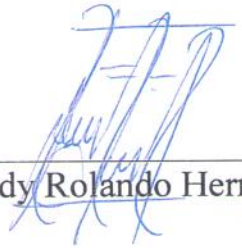
Dissertação submetida ao Programa de Pós-Graduação em Matemática da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do título de Mestre em Matemática.

Aprovada em 24/06/2016 por:

Fábio Júlio da Silva Valentim – UFES



Eterelides Gonçalves Júnior – UFES



Freddy Rolando Hernandez Romero - UFF

A Camilla, minha esposa.

Agradecimentos

A Deus, por ter me concedido a vida e forças para realizar este trabalho.

À minha esposa Camilla, Por ser companheira e a motivação para seguir em frente.

Aos meus pais Everton e Mara, pelos valores que me ensinaram e pelo carinho.

Aos meus irmãos Mayara e Emilio, pelo afeto.

Ao meu orientador, Prof. Dr. Fábio Júlio da Silva Valentim, por ter confiado na minha capacidade em fazer este trabalho.

Agradeço a CAPES pelo apoio financeiro.

Resumo

Abordaremos a Teoria de Acoplamentos mostrando aplicações em problemas de Probabilidade e Análise, mais especificamente, apresentaremos aplicações em dois contextos, que são as Cadeias de Markov e o Problema de Transporte Ótimo.
O Capítulo 1 apresenta algumas noções preliminares que servem como base para a compreensão deste trabalho, nele abordaremos noções de Probabilidade, Cadeias de Markov, Topologia, Funções Contínuas e Semicontínuas e Análise Convexa.
Nos dois Capítulos seguintes serão apresentados os contextos principais com as suas respectivas aplicações, mais precisamente, reservaremos o Capítulo 2 para apresentar Acoplamentos e algumas de suas aplicações em Cadeias de Markov, dando destaque para o cálculo do tempo de mistura de uma Cadeia de Markov e a demonstração do Teorema da Convergência via Acoplamentos.
No Capítulo 3 abordaremos o Problema de Transporte Ótimo, que pode ser divido em outros dois problemas, o Problema de Monge e o Problema de Kantorovich. Será apresentada a diferença entre os dois problemas, condições para existência de solução e a relação que existe entre os dois problemas e por fim apresentar uma sugestão de algoritmo que resolve o Problema de Kantorovich e demonstrar a Desigualdade Isoperimétrica via Problema de Monge.

Palavras-chaves: Acoplamentos. Cadeias de Markov. Problema de Monge. Problema de Kantorovich. Transporte Ótimo.

Abstract

We will cover the Couplings Theory showing applications in Probability and analysis of problems, more specifically, exhibit applications in two contexts, which are the Markov Chain and Optmal Transport Problem.
Chapter 1 presents some preliminary ideas which serve as a base for understanding this work, we will cover Probability notions, Markov Chains, Topology, Continuous Functions and semicontinuous and Convex Analysis.
In the next two chapters will be presented the main contexts with their respective applications, more precisely, reserve the Chapter 2 to display Couplings and some of its applications in Markov chains, highlighting the calculation of mixing time of a Markov Chain and the proof of Theorem of Convergence via couplings.
Chapter 3 will discuss the Transportation Problem Great, which can be divide in two problems, the Monge problem and Kantorovich problem. the difference will be presented between the two problems, conditions for solution of existence and the relationship between the two problems and finally present an suggestion algorithm that solves Kantorovich problem and demonstrate the isoperimetric inequality via Monge problem.

Key-words: Couplings. Markov chains. Monge problem. Kantorovich problem. Optmal transport.

Sumário

Introdução

Neste trabalho apresentaremos uma ferramenta bastante aplicável em diversas áreas, a teoria de acoplamentos. Dando destaques a dois tipos de aplicações: Cadeias de Markov e Problema do Transporte Ótimo.

Apresentaremos Cadeias de Markov (com espaço de estados finitos e homogênea no tempo) de um modo mais geométrico, como um passeio aleatório em um grafo, com o intuito de tornar mais claras as aplicações de acoplamentos nessas Cadeias, a referência básica se encontra em [13]. Uma das aplicações que será apresentada é o cálculo do tempo de mistura da Cadeia de Markov, em outras palavras, dada uma Cadeia de Markov $\{X_t\}_{\in\mathbb{N}}$ em Ω com a matriz de transição P e um única distribuição estacionária π, estimar o tempo que leva para a distância de variação total entre π e $P^t(x,\cdot)$ seja menor que um $\varepsilon > 0$ dado, para todo $x \in \Omega$. Outra aplicação que será feita nesse trabalho é a demonstração do Teorema da Convergência que mostra que se a Cadeia de Markov é Aperiódica e Irredutível, então para todo $x \in \Omega$ temos $P^t(x,\cdot)$ converge para sua distribuição estacionária π em tempo exponencial.

No Problema de Transporte Ótimo, dividiremos em dois problemas que são, problema de Monge e Problema de Kantorovich. O Problema do Transporte Ótimo informalmente pode ser descrito do seguinte modo. Considere $\mathcal{X}$ o conjunto das filiais de uma empresa que fabrica refrigerantes e $\mathcal{Y}$ uma grande franquia de lanchonetes, com filiais $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ e $\mathcal{Y} = \{y_1, y_2, \cdots, y_m\}$. Deve-se transportar produtos da empresa $\mathcal{X}$, na qual a proporção fabricada por cada filial é associada a uma medida de probabilidade μ, para $\mathcal{Y}$, cuja demanda de cada fábrica também está associada a uma medida de probabilidade ν. Cada filial x_i produz uma determinada quantidade de refrigerantes e cada filial y_j precisa de uma determinada quantidade do produto de $\mathcal{X}$.

Considere uma função custo $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, onde $c(x_i, y_j)$ significa o custo do transporte por unidade de x_i para y_j. O objetivo do problema é transportar os produtos da empresa $\mathcal{X}$ para $\mathcal{Y}$ com o menor custo.

O Problema de Monge é o Problema de Transporte Ótimo cujo o transporte é dado por uma aplicação $T : \mathcal{X} \to \mathcal{Y}$, onde $T(x_i, y_j)$ é a quantidade de refrigerantes transportados de x_i para y_j, e o Problema de Kantorovich, que é um versão fraca de Monge, consiste em encontrar um acoplamento de μ e ν que minimize o custo.

No decorrer do trabalho vamos fazer uso de algumas noções preliminares e apresentar alguns resultados de Análise Convexa baseando na referência [14]. O objetivo desse trabalho é apresentar a importância da teoria de acoplamentos mostrando algumas de suas aplicações com exemplos intuitivos mas sem perder o rigor matemático.

1 Preliminares

Neste capítulo faremos uma breve revisão dos conceitos e resultados que oferecem uma base para a compreensão deste trabalho. São relembrados alguns conceitos fundamentais da teoria de Probabilidade, Cadeias de Markov, Topologia e Funções Contínuas e Semicontínuas. Por se tratar de noções Preliminares, não demonstraremos alguns resultados, mas caso o leitor deseje ver as provas ou ter uma profundidade maior dos assuntos tratados, recomenda-se as referências [3] e [7].

1.1 Noções de Probabilidade

Definição 1.1.1. *(Álgebra de conjuntos) Considere Ω um conjunto não vazio e $\mathcal{K} \subset \wp(\Omega)$, sendo $\wp(\Omega)$ o conjunto das partes de Ω uma família de subconjuntos de Ω, dizemos que $\mathcal{K}$ é uma álgebra de conjuntos se cumpre as seguintes condições.*

a) $\Omega \in \mathcal{K}$.

b) Se $A \in \mathcal{K}$, *então* $A^c \in \mathcal{K}$.

c) Se $A \in \mathcal{K}$ *e* $B \in \mathcal{K}$ *então* $A \cup B \in \mathcal{K}$.

Caso $\mathcal{K}$ também satisfaça

d) Se $A_1, A_2, \cdots \in \mathcal{K}$ *então* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{K}$.

Diremos que $\mathcal{K}$ é uma σ-álgebra do conjunto Ω.

Intuitivamente, dado um conjunto Ω queremos verificar quais são os subconjuntos que podemos calcular sua probabilidade, ou ainda, queremos saber quais conjuntos somos capazes de medir. Como pode ser visto na teoria da medida, nem sempre é possível encontrar uma medida que seja capaz de medir qualquer subconjunto, para maiores detalhes veja o capítulo 3 da referência [3], o mesmo ocorre em teoria de probabilidade, nem sempre somos capazes de atribuir uma probabilidade a todos os subconjuntos de Ω mas pelo menos a família de subconjuntos $\mathcal{K}$ na qual atribui uma probabilidade (ou ainda, mensurável) forma uma álgebra ou σ-álgebra de conjuntos. Com esta motivação, estudar mais propriedades dessa família é fundamental para o restante da teoria.

Proposição 1.1.2. *Seja $\mathcal{K}$ uma álgebra de subconjuntos de Ω, então $\mathcal{K}$ possui as seguintes propriedades:*

a) $\emptyset \in \mathcal{K}$;

b) Para todo $n \in \mathbb{N}$ *e para todo* $A_1, A_2, \cdots, A_n \in \mathcal{K}$ *temos* $\bigcup_{i=1}^{n} A_i$ *e* $\bigcap_{i=1}^{n} A_i \in \mathcal{K}$.

c) Se $A \in \mathcal{K}$ *e* $B \in \mathcal{K}$, *então* $A - B \in \mathcal{K}$, *onde* $A - B = A \cap B^c = \{x \in A; x \notin B\}$.

Caso $\mathcal{K}$ satisfaça.

b') Para todo $\{A_n\}_{n\in\mathbb{N}} \in \mathcal{K}$ temos $\bigcup_{n=1}^{\infty} A_n$ e $\bigcap_{n=1}^{\infty} A_n \in \mathcal{K}$. Chamamos $\mathcal{K}$ de σ-álgebra.

Definição 1.1.3. *(Medida de probabilidade) Seja $\mathcal{F} \subset \Omega$ uma σ-álgebra de conjuntos e dizemos que uma função $P : \mathcal{F} \to [0, 1]$ é uma medida de probabilidade finitamente aditiva se cumpre as seguintes condições.*

a) $P(A) \geq 0$ *para todo* $A \in \mathcal{F}$.

b) $P(\Omega) = 1$;

c) (Aditividade finita). Se $A_1, \cdots, A_n \in \mathcal{F}$ *são dois a dois disjuntos, então*

$$P(\bigcup_{k=1}^{n} A_k) = \sum_{k=1}^{n} P(A_k).$$

Caso a função P satisfaça

d) (σ-Aditividade). Se $\{A_k\}_{k\in\mathbb{N}} \in \mathbb{A}$ *são dois a dois disjuntos então*

$$P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k).$$

Então a função de probabilidade P é σ-aditiva.

Definição 1.1.4. *(Espaço de probabilidade) Seguindo a notação acima, o trio $(\Omega, \mathcal{F}, P)$ é chamado espaço de probabilidade.*

Vale ressaltar que essa parte inicial da teoria de probabilidade tem uma relação muito forte com a teoria da medida, mais precisamente, a medida de probabilidade é um exemplo de medida, para maiores detalhes veja o capítulo 3 da referência [3]. Um objeto que será bastante explorado para os nossos estudos é a variável aleatória portanto segue a sua definição.

Definição 1.1.5. *(Variável aleatória) Seja $(\Omega, \mathcal{F}, P)$ um espaço de probabilidade, considere a função $X : \Omega \to \mathbb{R}$ e o conjunto $[X \leq x] := \{\omega \in \Omega; X(\omega) \leq x\}$. Dizemos que X é uma variável aleatória se para todo $x \in \mathbb{R}$, o conjunto $[X \leq x] \in \mathcal{F}$.*

Exemplo 1.1.6. *Lançar uma moeda n vezes. Seja X o número de caras realizadas então:*

$$\Omega = \{(\omega_1, \omega_2, \cdots, \omega_n); \omega_i = \textit{cara ou coroa }\}.$$

$$X : \Omega \to \mathbb{R}$$
$$X(\omega_1, \omega_2, \cdots, \omega_n) = \textit{número de } \{i; \omega_i = \textit{cara}\}.$$

Exemplo 1.1.7. *(Uniforme discreto) Dizemos que X é uma variável aleatória discreta que segue o modelo uniforme quando*

$$P(X = x_j) = \frac{1}{n} \quad j = 1, 2, \cdots, n. \tag{1.1}$$

As vezes é denotado por $X \sim U[x_1, x_2, \cdots, x_n]$.
Neste caso, por Ω ser finito, podemos considerar $\mathcal{F} = \mathcal{P}(\Omega)$.

Exemplo 1.1.8. *(Modelo de Bernoulli) Dizemos que uma variável aleatória X segue o modelo de Bernoulli se*

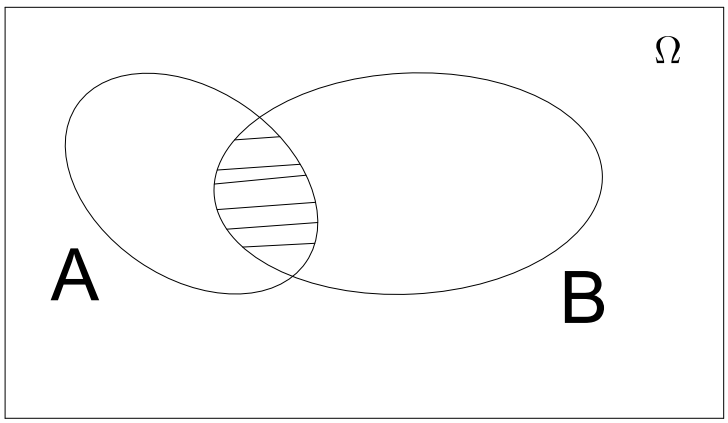
i) Para todo $w \in \Omega$, $X(w) = 0$ ou $X(w) = 1$.

ii) $P(X = 0) = 1 - p$ e $P(X = 1) = p$.

Em outras palavras, X determina o sucesso ou fracasso de um experimento é como lançar uma moeda com probabilidade p de sair cara e X expressa o resultado desse experimento.

Em muitas situações práticas a informação do que ocorreu em uma determinado evento pode influenciar nas probabilidades de outros eventos e é através dessa motivação que será apresentado o conceito de probabilidade condicional.

Definição 1.1.9. *(Probabilidade condicional) Seja $(\Omega, \mathcal{F}, P)$ um espaço de probabilidade. Se $B \in \mathcal{F}$ e $P(B) > 0$, a probabilidade condicional de A dado B é definida por*



$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad A \in \mathcal{F}.$$

No caso que $P(B) = 0$, definiremos $P(A|B) = P(A)$.

Como mostra a figura acima o símbolo $P(A|B)$ significa que queremos saber a probabilidade de ocorrer o evento A sabendo que B ocorre. Mas será que realmente para todo B evento aleatório, $P(A|B)$ define uma probabilidade? para responder esta pergunta basta verificar os axiomas:

i) $P(A|B) = \dfrac{P(A \cap B)}{P(B)} \geq 0.$

ii) $P(\Omega|B) = \dfrac{P(\Omega \cap B)}{P(B)} = \dfrac{P(B)}{P(B)} = 1.$

iii) Sejam $A_1, \cdots \in \mathcal{F}$ dois a dois disjuntos então pela definição de probabilidade condicional $P(\bigcup_n^\infty A_n|B) = \dfrac{P(\bigcup_n^\infty A_n \cap B)}{P(B)}$ se $(A_i)_{i=2}^\infty$ são dois a dois disjuntos é óbvio que $(A_i \cap B)_{i=1}^\infty$ também são e

$$\frac{P(\bigcup\limits_{n=1}^{\infty} A_n \cap B)}{P(B)} = \sum_{n=1}^{\infty} \frac{P(A_n \cap B)}{P(B)} = \sum_{n=1}^{\infty} P(A_n|B).$$

Com isto $P(\cdot|B)$ é uma medida de probabilidade e consequentemente temos

$$P(A^c|B) = 1 - P(A|B).$$

Se usar a definição de probabilidade condicional com o principio de indução será obtido o seguinte teorema.

Teorema 1.1.10. *Seja $(\Omega, \mathcal{F}, P)$ um espaço de probabilidade. Então*

i) $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$ *para todo* $A, B \in \mathcal{F}$.

ii) $P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap \cdots \cap A_{n-1})$.

Considere uma partição de Ω usando eventos $A_1, A_2, \cdots \in \mathcal{F}$, ou seja, $\Omega = \bigcup_{n\in\mathbb{N}} A_n$ com $A_i \cap A_j = \emptyset$ se $i \neq j$. Para todo $B \in \mathcal{F}$ temos $B = \bigcup_i (A_i \cap B)$ e $\{(A_i \cap B)\}_{i\in\mathbb{N}}$ são dois a dois disjuntos. A equação a seguir é chamada de *teorema da probabilidade total*.

$$P(B) = \sum_{i=1} P(A_i \cap B) = \sum_{i=1} P(A_i)P(B|A_i).$$

Usando a equação acima temos a *fórmula de Bayes* que é útil quando é conhecido as probabilidades dos A_i e a probabilidade B dado A_i. A fórmula de Bayes é descrita como:

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1} P(A_j)P(B|A_j)}.$$

1.2 Cadeias de Markov

O objetivo desta subseção é apresentar de forma introdutória, cadeias de Markov, um dos conceitos fundamentais deste trabalho. A ideia é apresentar cadeias de Markov de um modo mais intuitivo ou ilustrativo, para que sejamos capazes de criar bons exemplos e compreender com maior clareza a teoria apresentada neste trabalho. Por questão de objetividade, não vamos fazer todas as demonstrações, caso o leitor deseja se aprofundar neste assundo sugerimos que veja pelo menos os dois primeiros capítulos da referência [13].

Fixe um espaço de probabilidade $(\Omega, \mathcal{F}, P)$, um processo estocástico é definido como uma coleção de variáveis aleatórias X_t indexada por um parâmetro t pertencente a um conjunto T. Estamos interessados em processo com tempo discreto daí T será tomado como conjunto de inteiros não negativos e X_t representa uma característica mensurável de interesse (estado) no instante t.

O exemplo a seguir irá trazer uma noção intuitiva da cadeia de Markov e logo após será definido de modo mais preciso.

Exemplo 1.2.1. *Um certo sapo que mora em uma determinada lagoa com duas pétalas: Leste e Oeste. Em cada pétala existe uma moeda (não necessariamente moedas justas). Toda manhã o sapo joga a moeda, se der cara o sapo permanece na pétala e se der coroa o sapo se muda para a outra pétala.*

Digamos que se o sapo estiver na pétala leste, a probabilidade de jogar moeda e o resultado ser coroa é p *e se o sapo estiver na pétala oeste então a probabilidade de sair coroa é* q*, em outras palavras,* $\Omega = \{l, o\}$*, e seja* $(X_0, X_1, \cdots)$ *os estados de hoje, amanhã e assim por diante. O objetivo deste exemplo é estudar o movimento do sapo com o passar*

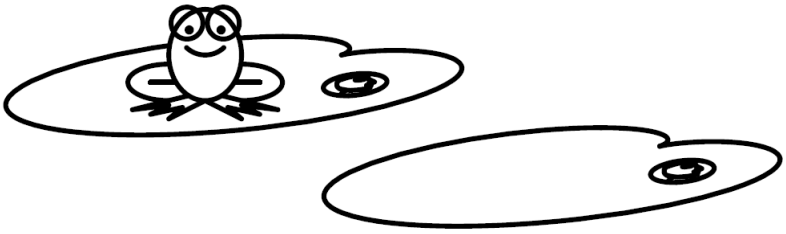


Figura 1 – Sapo na pétala oeste. Fonte: referência [13]

do tempo e tentar descobrir a probabilidade em que o sapo fica na pétala leste ou oeste. A matriz P de transição é dada por:

$$P = \begin{bmatrix} P(l,l) & P(l,o) \\ P(o,l) & P(o,o) \end{bmatrix} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}.$$

A primeira linha da matriz significa a distribuição da probabilidade do sapo ocupar os estados no instante $t+1$ dado que no estado t, o sapo estava na pétala leste. A segunda linha da matriz significa a distribuição da probabilidade do sapo ocupar os estados no instante $t+1$ dado que no estado t, o sapo estava na pétala oeste.

Vamos assumir que o sapo começa da pétala leste ($\mu_0 = (1,0)$) e que para todo i, $X_i = 0$ significa que no i-ésimo dia o sapo estava na pétala oeste e se $X_i = 1$ o sapo estava na pétala leste. No dia seguinte, como a probabilidade do sapo sair da pétala é p e a probabilidade de ficar é $1-p$ então:

$$P\{X_1 = 1 | X_0 = 1\} = 1 - p \quad e \quad P\{X_1 = 0 | X_0 = 1\} = p.$$

E o que acontece no segundo dia? se $X_2 = 1$ então pode ter acontecido de $X_1 = 1$ ou $X_1 = 0$ analogamente se $X_2 = 0$. Em outras palavras

$$P\{X_2 = 1 | X_0 = 1\} = (1-p)(1-p) \qquad P\{X_2 = 0 | X_0 = 1\} = (1-p)p + p(1-q).$$

Podemos ver pelo o que foi apresentado acima, que se a cadeia aomeça com uma distribuição inicial μ_0, então denotando a distribuição de probabilidade da cadeia após t passos com μ_t obtemos $\mu_1 = \mu_0 P$ e $\mu_2 = \mu_1 P$ mais geralmente, usando indução, temos que para todo t inteiro.

$$\mu_t = \mu_0 P^t. \tag{1.2}$$

A sequência $\{X_0, X_1, \cdots\}$ definida acima é uma cadeia de Markov. Note que no n-ésimo dia, para determinar o valor X_{n+1} basta somente conhecer X_n, e extraindo esta propriedade chegamos a definição da cadeia de Markov.

Definição 1.2.2. *(Cadeia de Markov homogénea no tempo) Uma sequência de variáveis aleatórias* $(X_0, X_1, \cdots)$ *é uma cadeia de Markov homogénea no tempo com espaço de estado finito* Ω *e matriz de transição* P *se para todo* $x, y \in \Omega$*, para todo* $t \geq 1$ *e para todo evento* $H_{t-1} = \bigcap_{s=0}^{t-1} \{X_s = x_s\}$ *que satisfaz* $P(H_{t-1} \cap \{X_t = x\}) > 0$ *se cumpre*

$$P\{X_{t+1} = y | (H_{t-1} \cap \{X_t = x\})\} = P\{X_{t+1} = y | X_t = x\} = P(x, y). \tag{1.3}$$

A expressão anterior nos diz que conhecendo o estado presente $\{X_t = x\}$ *o estado* $\{X_{t+1} = y\}$ *não depende de todo o seu passado, como se a cadeia perdesse a memoria e se preocupasse apenas com o instante* t*, além disso a probabilidade condicional do processo fazer a transição do estado* x *para o estado* y *pode ser descrita através de uma matriz de ordem* $| \Omega | \times | \Omega |$*.*

Vale ressaltar que o lado direito da equação (1.3) não depende de t, daí o nome homogénea no tempo. Além disso, a equação (1.3) nos permite concluir que a distribuição de probabilidade da cadeia iniciando no estado x após t passos é a x-ésima linha da matriz P^t que denotamos por $P^t(x, \cdot)$.

Uma maneira geométrica de interpretar cadeias de Markov, em espaço de estados finitos é via passeio aleatório sobre um grafo, indiretamente esta abordagem foi usada no exemplo do sapo. Será apresentada a definição de grafo e posteriormente exemplos de passeios aleatórios.

Definição 1.2.3. *(Grafo) Um grafo* $G = (V, A)$ *consiste em um conjunto de vértices* V *e um conjunto de arestas* A*. Onde* A *é formado por pares de vértices.*

$$A \subset \{\{x, y\}; x, y \in V, x \neq y\}.$$

Quando $\{x, y\} \in A$ usaremos o símbolo $x \sim y$ significa que x é vizinho de y, ou seja, que x e y formam uma aresta. O grau de x, denotado por $gr(x)$, é o números de vizinho de x.

Definição 1.2.4. *(Passeio aleatório) Dado um conjunto* Ω *e um grafo* $G = (\Omega, A)$*, definimos um passeio aleatório em um grafo por uma sequência de variáveis aleatórias* $(X_t)_{t=0}^{\infty}$ *onde para cada* t *temos* $X_t \in \Omega$ *e que* $P\{X_{t+1} = y | X_t = x\} > 0$ *se, e somente se,* $x \sim y$*.*

Exemplo 1.2.5. *Um passeio aleatório no* n*-ciclo é uma cadeia de Markov cujo estados são situados em* n *pontos do círculo com a transição ocorrendo apenas com os seus vizinhos (observe a figura).*

Definição 1.2.6. *(Passeio aleatório simples) Dado um grafo* $G = (V, A)$ *um passeio aleatório simples em* G *é uma cadeia de Markov em* V *tal que a matriz de transição é da forma*

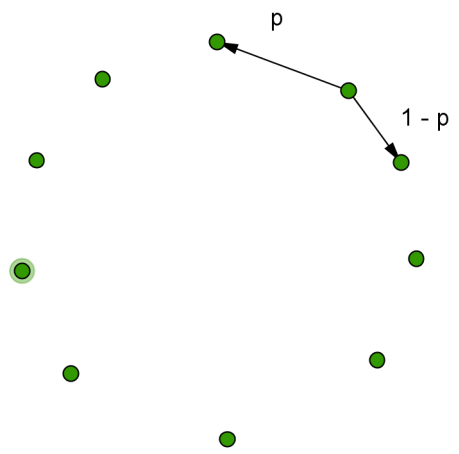


Figura 2 – Um exemplo de um 10-ciclo.

$$P(x,y) = \begin{cases} \dfrac{1}{gr(x)} & se \quad y \sim x \\ \\ 0 & Caso \;\; contrário \end{cases}$$

onde $gr(x)$ é o número de vizinhos do estado x.

Exemplo 1.2.7. *(n-ciclo) Seja $\mathbb{Z}_n = \{0, 1, \cdots, n-1\}$ e considere a seguinte matriz de transição*

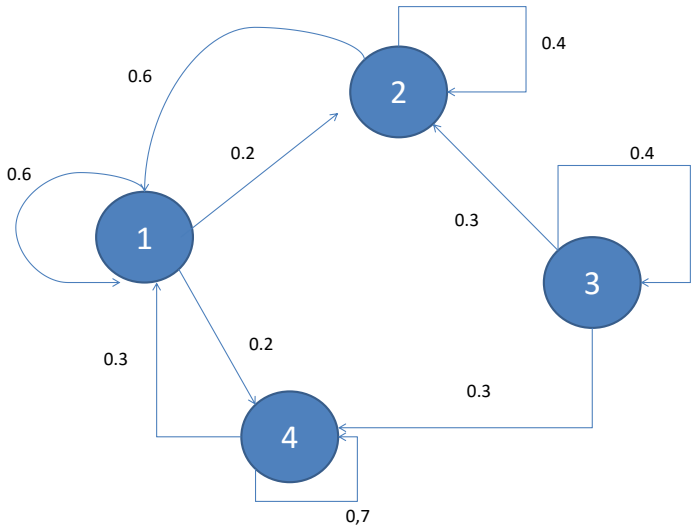
$$P(x,y) = \begin{cases} \dfrac{1}{2} & se \quad y = x + 1 (mod\ n) \\ \\ \dfrac{1}{2} & se \quad y = x - 1 (mod\ n). \\ \\ 0 & Caso \;\; contrário \end{cases}$$

O exemplo a seguir mostra que dado um grafo G é possível associa-lo a uma cadeia de Markov em V como um passeio aleatório em G. Reciprocamente, dada uma cadeia de Markov $(X_t)_{t=0}^{\infty}$ em Ω, podemos associar um passeio aleatório usando o fato de Ω ser um conjunto finito e usando a matriz de transição P para construir as arestas que são as entradas não nulas e assim construímos o grafo $G = (\Omega, V)$ e interpretamos a cadeia de Markov como um passeio aleatório em Ω.

Exemplo 1.2.8. *Considere uma cadeia de Markov em Ω cujo a matriz de transição P é dada por:*

$$P = \begin{bmatrix} 0,6 & 0,2 & 0 & 0,2 \\ 0,6 & 0,4 & 0 & 0 \\ 0 & 0,3 & 0,4 & 0,3 \\ 0,3 & 0 & 0 & 0,7 \end{bmatrix}$$

portanto podemos construir um grafo $G = (\Omega, A)$ e um passeio aleatório em que representa a matriz de transição P:



Exemplo 1.2.9. *Considere um grafo G em um n-ciclo, dizemos que a cadeia de Markov é preguiçosa, quando tem probabilidade $0,5$ de permanecer onde esta, com probabilidade $p \in [0, \frac{1}{2}]$ de mover no sentido horário e probabilidade $0,5 - p$ de mover no sentido anti-horário.*

Vimos pela seção anterior que toda cadeia de Markov pode ser vista como um passeio aleatório em um grafo G e a partir desta ilustração pretende-se estudar o comportamento da cadeia quando o tempo tende ao infinito, em outras palavras, a tentativa é agora tentar responder a seguinte pergunta: Dada uma cadeia de Markov $(X_t)_{t=0}^{\infty}$ com a matriz de transição P, existe alguma distribuição π tal que $\pi = \pi P$? π é única? essas perguntas motivam a nossa próxima definição.

Definição 1.2.10. *(Distribuição estacionária) Dada uma cadeia de Markov $(X_t)_{t=0}^{\infty}$ com a matriz de transição P, uma distribuição π é dita distribuição estacionaria se $\pi = \pi P$.*

Exemplo 1.2.11. *No exemplo do sapo temos que*

$$P = \begin{bmatrix} P(l,l) & P(l,o) \\ P(o,l) & P(o,o) \end{bmatrix} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}.$$

$\pi = (\frac{p}{p+q}, \frac{q}{p+q})$, é estacionária, pois $\pi P = \pi$.

Note que a distribuição π é estacionaria significa verificar o seguinte sistema linear com $|\Omega| + 1$ equações

$$\pi(y) = \sum_{x \in \Omega} \pi(x) \cdot P(x, y) \quad \forall y \in \Omega.$$

$$\sum_{x \in \Omega} \pi(x) = 1.$$

Dada uma cadeia de Markov $(X_t)_{t=0}^{\infty}$ não esta claro que esta cadeia possui uma única distribuição estacionária e nem garantimos a sua existência, para uma discussão com mais detalhes sobre esses questionamentos recomenda-se a referência [13]. Vamos apresentar o teorema, cuja a demonstração está na referência acima, de uma condição que garante a existência e unicidade da distribuição estacionária, mas antes segue algumas definições.

Definição 1.2.12. *(Irredutível) Uma cadeia de Markov $(X_t)_{t=0}^{\infty}$ com matriz de transição P é dita ser irredutível se para quaisquer dois estados x e $y \in \Omega$ existe algum número inteiro t (possivelmente dependente de x e y) tal que $P^t(x, y) > 0$.*

Intuitivamente significa que dados dois estados x e y sempre é possível sair do estado x e após t passos chegar no estado y para algum t inteiro.

Teorema 1.2.13. *Seja P uma matriz de transição irredutível da cadeia de Markov. Então existe e é única distribuição estacionária π.*

1.3 Topologia

Nesta seção vamos estudar de modo diagonal sobre topologia, apesar desse assunto ser bastante amplo, vamos apenas nos concentrar em conceitos que serão utilizados no decorrer deste texto. Para um aprofundamento dos tópicos apresentados veja as referências [12] e [5].

Definição 1.3.1. *(Topologia) Considere Ω um conjunto não vazio qualquer e $\tau \subset \Omega$ uma família de subconjuntos, então dizemos que τ é uma topologia de Ω se satisfaz as seguintes condições.*

a) $\Omega, \emptyset \in \tau$.

b) Se $V_1, V_2, \cdots, V_n \in \tau$, *então* $\bigcap_{j=1}^{n} V_j \in \tau$.

c) $\{V_\alpha; \alpha \in I\} \subset \tau$, *então* $\bigcup_{\alpha \in I} V_\alpha \in \tau$.

Os elementos da topologia são chamados de abertos, e o par (Ω, τ) é chamado de espaço topológico.

Exemplo 1.3.2. *$\tau = \{\emptyset, \Omega\}$ ou $\tau = \wp(\Omega)$, que é o conjunto das partes de Ω, são chamados topologias triviais.*

Exemplo 1.3.3. *$\Omega = \mathbb{R}$ e considere τ uma família de subconjuntos de $\mathbb{R}$ tal que*

$$U \in \tau \text{ se, e somente se, para todo } x \in U \text{ existe } r > 0; (x - r, x + r) \subset U.$$

A topologia definida acima é chamada de topologia da norma, as vezes é denotada por τ_{norma}. Note que de modo análogo se Ω é um espaço normado então τ_{norma} é bem definido.

Com o entendimento mais claro do que é uma topologia, é importante lembrar que algumas noções vistas em um curso de Análise na reta dependem da topologia, o que motiva revisitar esses conceitos.

Definição 1.3.4. *(Função contínua) Dados $\mathcal{X}$ e $\mathcal{Y}$ dois espaços topológicos defini-se $f : A \subset \mathcal{X} \to \mathcal{Y}$ contínua em $a \in A$ se para todo aberto G em $\mathcal{Y}$ tal que $f(a) \in G$, tem-se que $f^{-1}(G)$ é um aberto em A, ou seja, $f^{-1}(G) = A \cap B$ e B é um aberto.*

Exemplo 1.3.5. *Considere $f : \mathbb{R} \to \mathbb{R}$ uma função qualquer e considere o espaços topológico trivial $(\mathbb{R}, \wp(\mathbb{R}))$ no domínio da função, então f é contínua.*

Fica claro pelo exemplo anterior, que a noção de continuidade depende da topologia a ser considerada, mas ao dizer que f é continua sem fazer referencia da topologia, fica subtendido que a função é continua com respeito a topologia da norma.

Definição 1.3.6. *(Compacto) Considere $\mathcal{X}$ um espaço topológico, então dizemos que um subconjunto $K \subset \mathcal{X}$ é compacto se toda cobertura de K por abertos admite uma subcobertura finita.*

Observe que, quanto mais abertos uma topologia possuir, menor é a chance de um conjunto ser compacto e maior é a chance de uma função ser contínua nesse espaço. A ideia é retirar alguns abertos da topologia da norma, de forma que preserve a continuidade dos funcionais lineares e consequentemente aumentar a chance de se obter um conjunto compacto.

Fixada uma família de funções $\mathcal{F}$, existem topologias em um conjunto Ω para quais todos o elementos de $\mathcal{F}$ são contínuos, à saber basta considerar $(\Omega, \mathcal{P}(\Omega))$. Outro fato é que a interseção de topologias ainda é uma topologia, e assim ao considerar uma topologia τ, formada pela interseção de todas as topologias na qual todos o elementos de $\mathcal{F}$ são contínuos, obtemos que τ é a menor topologia que cumpre tal propriedade, em outras palavras, qualquer topologia τ' que cumpre a propriedade acima contém a topologia τ e com essas observações definimos.

Definição 1.3.7. *(Topologia gerada) Dada uma família de funções $\mathcal{F}$ em Ω, a topologia gerada por $\mathcal{F}$ é a menor topologia em Ω para qual todos os elementos de $\mathcal{F}$ são contínuos.*

Definição 1.3.8. *(Topologia fraca) A topologia fraca em um espaço normado Ω é a menor topologia relativamente à qual todo todos os funcionais lineares são contínuos. Denotaremos por τ_{fraca}.*

Para encerrar a seção, apresentaremos algumas noções de espaços separáveis e convergência fraca que são fundamentais para a compreensão refinada de um dos grandes resultados do texto, que é o Teorema Fundamental do Transporte Ótimo e o Teorema de Prokhorov.

Definição 1.3.9. *(Espaços separáveis) Dizemos que um espaço topológico $\mathcal{X}$ é separável se existe um conjunto enumerável D denso em $\mathcal{X}$, ou seja, $\overline{D} = \mathcal{X}$.*

Exemplo 1.3.10. *$\mathbb{R}$ é separável, pois possui $\mathbb{Q}$ como um subconjunto enumerável e denso.*

Exemplo 1.3.11. *Um espaço métrico discreto M é separável se, e somente se, é enumerável.*

Um fato interessante em espaços métricos separáveis, é que podemos supor sem perda de generalidade que toda cobertura aberta de $\mathcal{X}$ é enumerável, esta propriedade conhecida como propriedade de Linderlof se justifica na proposição a seguir.

Proposição 1.3.12. *Seja $\mathcal{X}$ um espaço métrico, então $\mathcal{X}$ é separável se, e somente se, toda cobertura aberta de $\mathcal{X}$ admite uma subcobertura enumerável.*

Demonstração. Suponha que $\mathcal{X}$ seja separável, então existe $D \subset \mathcal{X}$ enumerável e $\overline{D} = \mathcal{X}$. Considere $\mathcal{U}$, uma cobertura aberta de $\mathcal{X}$. Seja $\mathcal{B}$ uma família formada por abertos que contém algum ponto de D. Portanto para cada $\mathcal{B}' \in \mathcal{B}$ existe um conjunto $\mathcal{U}' \in \mathcal{U}$ de modo que $\mathcal{B}' \subset \mathcal{U}'$, os conjuntos $\mathcal{U}'$ formam uma coleção enumerável $\mathcal{U}' \subset \mathcal{U}$.

Resta mostrar que $\mathcal{U}'$ forma uma cobertura de $\mathcal{X}$. De fato, dado qualquer $x \in \mathcal{X}$, temos que $x \in \mathcal{U}$ para algum $\mathcal{U}$. Note que existe um aberto $\mathcal{B}'$ tal que $x \in \mathcal{B}' \subset \mathcal{U}$ e assim corresponde, na escolha feita acima que, existe $U' \in \mathcal{U}'$ tal que $x \in U'$ no que implica que $\mathcal{U}'$ cobre $\mathcal{X}$.

Reciprocamente, para cada $n \in \mathbb{N}$, considere uma cobertura aberta formada por bolas com raio $\frac{1}{n}$ centrada em cada elemento de $\mathcal{X}$, portanto temos uma cobertura aberta de $\mathcal{X}$, então podemos extrair uma subcobertura enumerável. Os centros de cada bola dessa subcobertura formam um conjunto enumerável E_n tal que todo ponto de $\mathcal{X}$ dista menos que $\frac{1}{n}$ de algum ponto de E_n. Segue que $E = \bigcup_n E_n$ é um subconjunto enumerável denso em $\mathcal{X}$.

□

Definição 1.3.13. *(Convergência fraca) Dizemos que uma sequência $\{\mu_n\}_{n=1}^{\infty} \subset \wp(\mathcal{X})$ converge fracamente para uma medida μ, se para toda função contínua e limitada ϕ, denotada por $\phi \in C_b(\mathcal{X})$, temos*

$$\int \phi \ \ d\mu_n \longrightarrow \int \phi \ \ d\mu.$$

Exemplo 1.3.14. *Considere nos conjuntos dos números reais uma sequência $\{x_n\}$ de forma que a mesma converge para um número real x. Então a medida δ_{x_n}, converge fracamente para a medida δ_x. De fato, seja $\phi \in C_b(\mathbb{R})$ qualquer, note que, pela propriedade da continuidade da ϕ temos que.*

$$\int \phi \ \ d\delta_{x_n} = \phi(x_n) \longrightarrow \phi(x) = \int \phi \ \ d\delta_x.$$

Definição 1.3.15. *(Convergência em geral de probabilidade) Seja $(\Omega, \mathcal{F})$ um espaço (métrico) mensurável. Uma sequência de medidas de probabilidade, $\{P_n\}$ converge em geral para medida de probabilidade P, denotado por $P_n \to P$, se*

$$P_n(A) \to P(A)$$

para todo $A \in \mathcal{F}$ com $P(\partial A) = 0$. Onde ∂A é a fronteira do conjunto A.

Proposição 1.3.16. *São equivalentes as seguintes sentenças*

a) P_n converge fracamente para P.

b) $\limsup P_n(A) \leq P(A)$, sempre que A é fechado.

c) $\liminf P_n(A) \geq P(A)$, sempre que A é aberto.

d) $P_n \to P$.

Demonstração. $(a) \Rightarrow (b)$ Sejam A um conjunto fechado qualquer, $f(x) = I_A(x)$, ρ uma métrica e para cada $\varepsilon > 0$ considere

$$f_\varepsilon(x) = g(\frac{1}{\varepsilon}\rho(x, A)), \varepsilon > 0$$

onde

$$\rho(x, A) = \inf\{\rho(x, y); y \in A\}.$$

$$g(t) = \begin{cases} 1, & t \leq 0 \\ 1 - t, & 0 \leq t \leq 1 \\ 0, & t \geq 1. \end{cases}$$

Considere também o conjunto $A_\varepsilon = \{x; \rho(x, A) < \varepsilon\}$. e note que por A ser fechado $A_\varepsilon \to A$ quando $\varepsilon \to 0$. Como $f_\varepsilon(x)$ é contínua e limitada, então

$$P_n(A) = \int_{\mathcal{X}} I_A(x) \ P_n(dx) \leq \int_{\mathcal{X}} f_\varepsilon(x) \ P_n(dx)$$

assim, obtemos usando a hipótese que

$$\limsup_n P_n(A) \leq \limsup_n \int_{\mathcal{X}} f_\varepsilon(x)\, P_n(dx) = \int_{\mathcal{X}} f_\varepsilon(x)\, P(dx) \leq P(A_\varepsilon) \to P(A) \quad \text{quando} \quad \varepsilon \to 0.$$

O que nos leva a prova da implicação.

$(b) \Longleftrightarrow (c)$ Considere A um aberto qualquer, então

$$\limsup_n P_n(A^c) \leq P(A^c) \Longleftrightarrow \limsup_n 1 - P_n(A) \leq 1 - P(A)$$

$$\liminf_n P_n(A) \geq P(A).$$

$(c) \Rightarrow (d)$ Seja A um conjunto que cumpre $P(\partial A) = 0$, temos

$$\limsup_n P_n(A) \leq \limsup_n P_n(\overline{A}) \leq P(\overline{A}) = P(A).$$

$$\liminf_n P_n(A) \geq \liminf_n P_n(int\ A) \geq P(int\ A) = P(A).$$

Conclusão $P_n(A) \to P(A)$ sempre que $P(\partial A) = 0$.

$(d) \Rightarrow (a)$ Seja f uma função contínua e limitada por M, ou seja, $|f(x)| \leq M$. Tome

$$D = \{t \in \mathbb{R}; P(x; f(x) = t) \neq 0\}.$$

E considere uma decomposição $T_k = (t_0, t_1, \cdots, t_k)$ de $[-M, M]$, onde

$$-M = t_0 < t_1 < \cdots < t_k = M$$

com $t_i \notin D$, $i \in \{0, 1, \cdots, k\}$ (Observe que D é enumerável, pois os conjuntos $f^{-1}(t)$ sejam disjuntos, P é finito e note que toda soma de quantidade não enumerável de positivo é infinita). Considere $B_i = \{x; t_i \leq f(x) < t_{i+1}\}$, como f é contínua temos $f^{-1}(t_i, t_{i+1})$ é aberto e $\partial B \subset f^{-1}(t_i) \bigcup f^{-1}(t_{i+1})$. Os pontos t_i e $t_{i+1} \notin D$, então $P(\partial B_i) = 0$, por hipótese temos

$$\sum_{i=0}^{k-1} t_i P_n(B_i) \to \sum_{i=0}^{k-1} t_i P(B_i).$$

Mas

$$|\int_{\mathcal{X}} f(x)\, P_n(dx) - \int_{\mathcal{X}} f(x)\, P(dx)| \leq |\int_{\mathcal{X}} f(x)\, P_n(dx) - \sum_{i=0}^{k-1} t_i P_n(B_i)| + |\sum_{i=0}^{k-1} t_i P_n(B_i) - \sum_{i=0}^{k-1} t_i P(B_i)| +$$

$$+|\sum_{i=0}^{k-1} t_i P(B_i) - \int_{\mathcal{X}} f(x)\ P(dx)| \leq 2 \max_{0\leq i\leq k-1}(t_{i+1}-t_i) + |\sum_{i=0}^{k-1} t_i P_n(B_i) - \sum_{i=0}^{k-1} t_i P(B_i)|.$$

Assim, para k e n suficientemente grande obtém-se

$$\lim_n \int_{\mathcal{X}} f(x)\ P(dx) = \int_{\mathcal{X}} f(x)\ P(dx).$$

O que completa a prova da proposição.

□

1.4 Função contínua e semicontínua

Vamos apresentar conceitos de função contínua e semicontínua. A importância destes conceitos é que um do principais resultados deste trabalho se baseia no fato de que a existência de solução do problema de Kantorovich ocorre se a função custo é semicontínua. Esta seção tem como a principal referência [3, Capítulo 2, na seção 6].

Definição 1.4.1. *(Função semicontínua) Seja $f : E \subset \mathbb{R} \to \mathbb{R}\bigcup\{-\infty, \infty\}$ uma função. Dizemos que f é semicontínua inferior em $x \in E \cap E'$ se,*

$$\liminf_{y\to x} f(y) \geq f(x). \tag{1.4}$$

Lema 1.4.2. *Se $f(x)$ é finito, então f é semicontínua inferior em x se, e somente se, dado $\varepsilon > 0$, existe $\delta > 0$ tal que $f(x) \leq f(y) + \varepsilon$ para todo $y \in E$ com $|x - y| < \delta$.*

Demonstração. Se f é semicontínua inferior em x, então $f(x) \leq \liminf_{y\to x} f(y)$, em outras palavras, dado $\varepsilon > 0$, existe $\delta > 0$ tal que se $\liminf_{y\to x} f(y) = L$, obtemos

- $f(x) \leq L$.
- $|L - f(y)| < \varepsilon$ sempre que $|y - x| < \delta$ e $y \in E$.

E assim conclui-se que $f(x) \leq f(y) + \varepsilon$.

□

Reciprocamente, suponha que dado $\varepsilon > 0$, existe $\delta > 0$ tal que $f(x) \leq f(y) + \varepsilon$ para todo $y \in E$ com $|x - y| < \delta$. O objetivo é mostrar que $f(x) \leq \liminf_{y\to x} f(y)$, ou seja, dado $\varepsilon_0 > 0$, devemos encontrar $\delta > 0$ de modo que

$$\textit{Para todo } y \in E \textit{ com } |x - y| < \delta, \textit{então } f(y) - f(x) > \varepsilon$$

E assim obtemos, $f(x) < f(y) - \varepsilon < f(y) + \varepsilon$. Portanto f é semicontínua inferior em x.

Teorema 1.4.3. *Seja $f : K \subset \mathbb{R}^n \to [-\infty, \infty]$ uma função semicontínua inferior com K compacto, então f assume mínimo em K.*

Demonstração. Dividiremos a prova nos seguintes passos

- Observe que $K \subset \bigcup_{z\in\mathbb{Z}} A_z$, onde $A_z = \{x \in K; f(x) > z\}$. Temos que.
- A_z é aberto em K para todo inteiro, pois basta verificar pela definição e usar o lema anterior.
- Como K é compacto e pelo primeiro passo, temos uma cobertura de abertos, então é possível extrair uma subcobertura finita de K. Daí conclui-se que f é limitada inferiormente.
- Considere $m = \inf_{x\in K} f(x)$. Resta construir uma sequência $\{x_n\} \subset K$ com $x_n \to x$ e $f(x_n) \to m$. Para a construção de tal sequência, basta notar que para cada n existe $x_n \in K$ tal que $x_n \leq m + \frac{1}{n}$ e como a sequência esta contida em um compacto, então existe uma subsequência convergente, digamos para x, e a compacidade de K garante que $x \in K$ e portanto $f(x) = m$.

□

Os próximos resultados são úteis para caracterização de funções semicontínuas e de associar sequência de funções contínuas com a função semicontínua. Estes resultados serão importantes na demonstração do teorema de existência de solução do Problema de Kantorovich.

Lema 1.4.4. *Seja $\{f_n\}_{n\in\mathbb{N}}$, com $f_n : E \to \mathbb{R}$ para todo $n \in \mathbb{N}$, uma sequência de funções semicontínuas inferiores. Então $f(x) = \sup_n f_n(x)$ é também semicontínua inferior.*

Demonstração. Seja $\{f_n\}_{n\in\mathbb{N}}$ uma sequência de funções semicontínua inferior e defina $f(x) = \sup_n f_n(x)$. Dado $\varepsilon > 0$, temos

$$f(x) - \frac{\varepsilon}{2} < f_{n_0}(x) \quad \text{Para algum} \quad n_0 \in \mathbb{N}.$$

Usando a hipótese, dado $\varepsilon > 0$ existe um $\delta_{n_0} > 0$ tal que

$$f_n(x) \leq f_n(y) + \frac{\varepsilon}{2} \quad \text{Para } y \in E, \ \ |x - y| < \delta_{n_0}.$$

Combinando as equações e usando a definição de f, obtemos $f(x) < f(y) + \varepsilon$. No que acarreta que f é semicontínua inferior.

□

Proposição 1.4.5. *Uma função real f definida em um compacto K é semicontínua inferior se, e somente se, existe uma sequência crescente de funções contínuas $\{\psi_n\}_{n\in\mathbb{N}}$ de modo que $f(x) = \lim_{n\to\infty} \psi_n(x)$ para cada $x \in K$.*

Demonstração. Suponha que exista uma sequência crescente de funções contínuas ψ_n de modo que $f(x) = \lim_{n\to\infty} \psi_n(x)$. Como para cada $n \in \mathbb{N}$, a função ψ_n é contínua, então ψ_n é também semicontínua inferior e pelo lema anterior, conclui-se que f é semicontínua inferior. Reciprocamente, suponha que a função f seja semicontínua inferior. Defina $\psi_n(x) = \inf_{t\in K}\{f(t) + n|t - x|\}$. Pela desigualdade triangular obtemos $\psi_n(x) \leq \inf_{t\in K}\{f(t) + n|t - y| + n|y - x|\}$, então obtemos $\psi_n(x) \leq \psi_n(y) + n|x - y|$, no que implica que para cada n, a função ψ_n é (uniformemente) contínua em K. Também obtemos $\psi_n \leq \psi_{n+1} \leq f$ para todo $n \in \mathbb{N}$. Em particular $f(x)$ é uma cota superior de $\{\psi_n(x); n \in \mathbb{N}\}$, agora se $\alpha < f(x)$, então existe $\delta > 0$ tal que

$$\alpha \leq f(y) \leq f(y) + n|y - x| \quad \text{Sempre que } \ |x - y| < \delta, y \in K.$$

Por outro lado, se $|x - y| \geq \delta$ tome $n \geq \frac{\alpha - m}{\delta}$, onde $m = \min_{t\in K} f(t)$, e assim $\psi_n(x) \geq \inf_{t\in K}\{f(t) + \alpha - m\} = \alpha$. Portanto conclui-se que $\alpha \leq \psi_n(x)$ para n consequentemente $f(x) = \sup_{n\in\mathbb{N}} \psi_n(x)$.

□

1.5 Tópicos de análise convexa

Nesta seção falaremos um pouco sobre Análise Convexa com o objetivo extrair alguns resultados dessa área para demonstração de um dos grandes resultados desse trabalho, que é o teorema de Brenier. Para isso vamos iniciar definindo função convexa e as referências desta seção são as referências [10], [11], [14].

Definição 1.5.1. *(Conjunto Convexo) Um conjunto $C \subset \mathbb{R}^n$ é dito convexo se para todo $x, y \in C$ e para qualquer $t \in [0, 1]$ temos que $(1 - t)x + ty \in C$.*

Definição 1.5.2. *(Epigrafo) O Epigrafo de uma função $f : S \subset \mathbb{R}^n \to (-\infty, \infty]$ é o conjunto $epi(f)$ definido por*

$$epi(f) = \{(x, v); x \in S, v \in R, v \geq f(x)\}.$$

Exemplo 1.5.3. *$f(x) = x^2$, então o epigráfico de f é o conjunto $epi(f) = \{(x, v); x \in \mathbb{R}, v \in R, v \geq x^2\}$ como mostra a figura 3.*

Definição 1.5.4. *(Função convexa) Uma função $f : S \subset \mathbb{R}^n \to (-\infty, \infty]$ é convexa, se $epi(f)$ é um conjunto convexo.*

No exemplo anterior podemos observar que $f(x) = x^2$ é uma função convexa, um exemplo de uma função que não é convexa é $f(x) = -x^2$. Um conceito muito importante que aparece nesse trabalho é o de subgradiente de uma função convexa, essa noção é importante pois é por meio dela que podemos verificar a existência de solução do problema

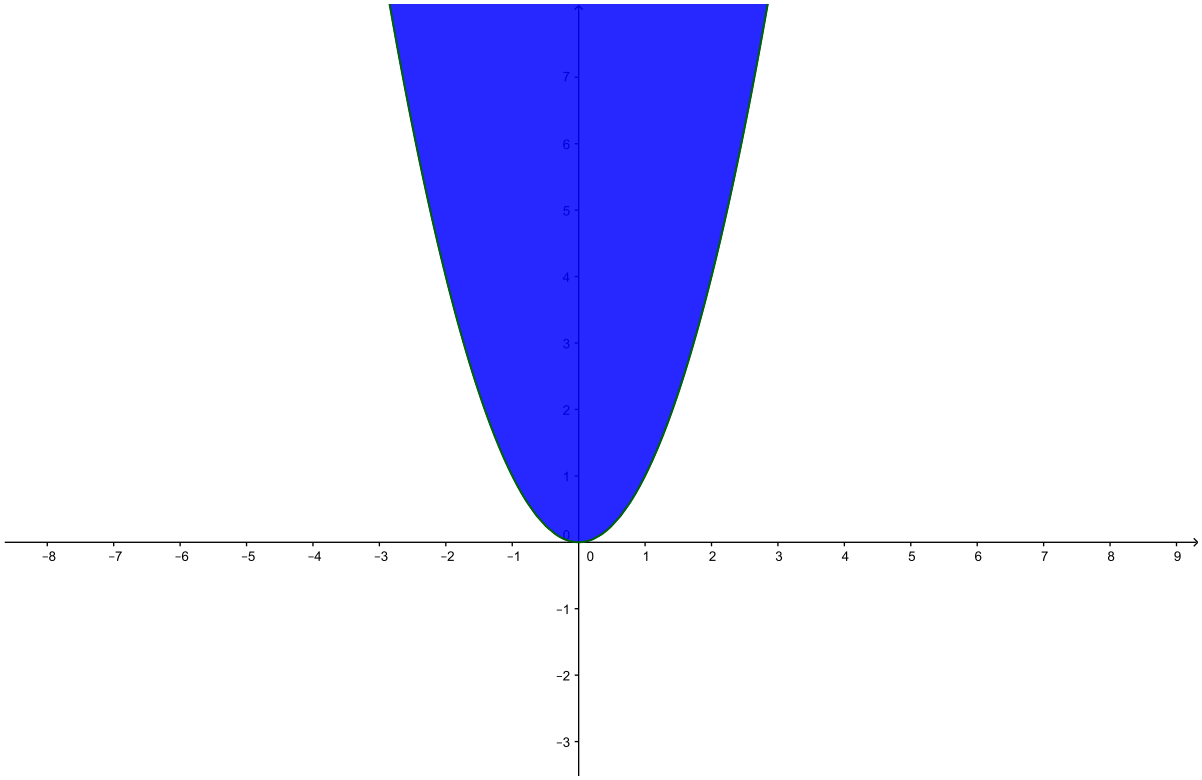


Figura 3 – A região azul é o epigráfico de f.

de Monge, ou seja, estudar o subgradiente de uma função convexa nos condicionará a determinar existência de acoplamentos ótimos induzidos por alguma aplicação T. A partir dessa motivação, segue a definição de subgradiente de um função convexa.

Definição 1.5.5. *(Subgradiente de uma função convexa) Um vetor $y \in \mathbb{R}^n$ pertence ao subgradiente de uma função convexa $f : S \subset \mathbb{R}^n \to \mathbb{R}$ em $x \in S$ se*

$$f(z) \geq f(x) + \langle y, z - x \rangle \quad \forall z \in S.$$

Denotaremos por $\partial^- f(x)$ o conjunto dos subgradientes de f em x e $\partial^- f$ o subgradiente de f, em outras palavras, $\partial^- f = \bigcup_{x \in S} \partial^- f(x)$.

Exemplo 1.5.6. *$f : \mathbb{R}^n \to \mathbb{R}$ com $f(x) = |x|$, note que f não é diferenciável apenas no ponto $x = 0$, mesmo assim podemos determinar $\partial^- f(0)$ e neste caso são os vetores y que cumprem*

$$|z| \geq \langle y, z \rangle \quad \forall z \in \mathbb{R}^n$$

portanto, $\partial^- f(0) = B[0, 1]$. No caso $\partial^- f(x)$ com $x \neq 0$, temos $\partial^- f(x) = \{\frac{x}{|x|}\}$, fato justificado pelo seguinte teorema.

Teorema 1.5.7. *Seja $f : S \subset \mathbb{R}^n \to (-\infty, \infty]$ uma função convexa com S aberto e considere x um ponto onde f é finita. Se f é diferenciável em x, então $\bigtriangledown f(x)$ é o único elemento do subgradiente de f em x, em particular*

$$f(z) \geq f(x) + \langle \bigtriangledown f(x), z - x \rangle \quad \forall z \in S.$$

Lema 1.5.8. *Nas mesmas condições do teorema acima, $\overline{x}$ pertence ao subgradiente de f em x se, e somente se,*

$$\frac{\partial f(x)}{\partial y} \geq \langle \overline{x}, y \rangle \quad \forall y \in S. \tag{1.5}$$

Demonstração. Tomando $z = x + \lambda y$, com λ positivo temos que se $\overline{x} \in \partial^- f(x)$ se, e somente se, para todo z no domínio, $f(z) \geq f(x) + \langle \overline{x}, z - x \rangle$, assim

$$\frac{f(x + \lambda y) - f(x)}{\lambda} \geq \langle \overline{x}, y \rangle \tag{1.6}$$

então se f é diferenciável em x o limite a esquerda, quando λ tende a zero, existe e portanto vale a equação (1.5). Reciprocamente se vale a equação acima para todo y, então temos primeiramente pelo teorema do valor médio que existe $0 \leq \theta \leq \lambda$ tal que

$$\frac{f(x + \lambda y) - f(x)}{\lambda} = \frac{\partial f(x + \theta y)}{\partial y}.$$

Como a função é convexa então a derivada é crescente, assim $\dfrac{\partial f(x + \theta y)}{\partial y} \geq \dfrac{\partial f(x)}{\partial y} \geq \langle \overline{x}, y \rangle$ para todo y e portanto vale a equação (1.6) e assim concluímos que $\overline{x}$ pertence ao subgradiente de f em x.

□

Demonstração. (Teorema 1.5.7) Pelo lema anterior e usando a diferenciabilidade de f em x obtemos

$$\langle \bigtriangledown f(x), y \rangle \geq \langle \overline{x}, y \rangle \quad \forall y \in S$$

daí, o único ponto $\overline{x}$ que satisfaz a inequação acima é $\overline{x} = \bigtriangledown f(x)$ e portando $\bigtriangledown f(x)$ é o único subgradiente de f em x. □

Vale ressaltar que a recíproca é verdadeira, ou seja, se y é o único subgradiente de f em x, então f é diferenciável em x. Para ver a demonstração da recíproca veja a referência [14, seção 23].

O teorema a seguir nos mostra uma forma de caracterizar uma função convexa, inclusive existem referências que definem funções convexas como a reciproca do próximo teorema, que segue.

Teorema 1.5.9. *Seja $f : U \to \mathbb{R}$ definida no aberto convexo $U \subset \mathbb{R}^n$, então são equivalentes.*

i) f é convexo.

ii) Para todo $x, y \in U$ e $t \in [0, 1]$

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

Demonstração. i) $\Rightarrow$ ii) Sejam $x, y \in U$ quaisquer, então $(x, f(x))$ e $(y, f(y)) \in epi(f)$. Por hipótese, todo $t \in [0, 1]$ obtemos $((1-t)x + ty, (1-t)f(x) + tf(y)) \in epi(f)$, ou seja,

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

ii) $\Rightarrow$ i) Considere (x_1, z_1) e $(x_2, z_2) \in epi(f)$. Queremos mostrar que para todo $t \in [0, 1]$, obtemos $((1-t)x_1 + tx_2, (1-t)z_1 + tz_2) \in epi(f)$. Como (x_1, z_1) e $(x_2, z_2) \in epi(f)$, então $f(x_1) \leq z_1$ e $f(x_2) \leq z_2$. No que acarreta $(1-t)f(x_1) + tf(x_2) \leq (1-t)z_1 + tz_2$ e usando a hipótese obtemos que

$$f((1-t)x_1 + tx_2) \leq (1-t)z_1 + tz_2.$$

Conclusão, $((1-t)x_1 + tx_2, (1-t)z_1 + tz_2) \in epi(f)$.

□

Um outro resultado importante para este trabalho é mostrar que se uma função convexa é duas vezes diferenciável, então a matriz Hessiana é uma forma quadrática não-negativa, consequentemente todos os seus autovalores são não negativos. Para mostrar este resultado vamos enunciar dois lemas úteis para demonstração, onde o primeiro lema omitiremos a prova, pois a prova é feita em detalhes na referência [9] e não é viável fazer a demonstração neste trabalho e o segundo lema, usaremos o primeiro para a sua demonstração. Portanto segue os seguintes lemas.

Lema 1.5.10. *As seguintes afirmações sobre a função $f : I \to \mathbb{R}$, derivável no intervalo I, são equivalentes:*

i) f é convexa.

ii) A derivada $f' : I \to \mathbb{R}$ é monótona não-decrescente.

iii) Para quaisquer $a, x \in I$ tem-se $f(x) \geq f(a) + f'(a)(x - a)$, ou seja, o gráfico de f está situado acima de qualquer de suas tangentes.

Demonstração. A prova completa está na referência [9, Capítulo 9, na seção 2].

□

Lema 1.5.11. *Seja $U \subset \mathbb{R}^m$ aberto convexo. Uma função $f : U \to \mathbb{R}$ diferenciável é convexa se, e somente se, para cada $x, x+v \in U$ quaisquer, tem-se $f(x+v) \geq f(x)+df(x)\cdot v$*

Demonstração. Se f é convexa e diferenciável, sabemos que para $x, x+v \in U$ quaisquer

$$f(x+v) = f(x) + df(x)\cdot v + r(v) \;\text{ com }\; \lim_{v\to 0}\frac{r(v)}{|v|} = 0.$$

e

$$f(x+tv) = f((1-t)x + t(x+v)) \leq (1-t)f(x) + tf(x+v) \;\; t \in (0,1).$$

Combinando as duas equações temos

$$tf(x+v) \geq f(x+tv) - (1-t)f(x) = f(x+tv) - f(x) + tf(x) = df(x)\cdot(tv) + r(tv) + tf(x).$$

Dividindo por t, obtemos

$$f(x+v) \geq df(x)\cdot v + f(x) + \frac{r(tv)}{t}.$$

Fazendo $t \searrow 0$, obtemos

$$f(x+v) \geq f(x) + df(x)\cdot v$$

Reciprocamente, se vale a desigualdade para qualquer $x, x+v \in U$. Considere uma função $\varphi : [0,1] \to \mathbb{R}$ como $\varphi(t) = f(x+tv)$. Assim $\varphi'(t) = df(x+tv)\cdot v$. Ora para qualquer $t, t_0 \in [0,1]$ tem-se $f(x+tv) = f(x+t_0v + (t-t_0)v) = f(x+t_0v+sv)$ com $s = t - t_0$, logo por hipótese

$$f(x+tv) \geq f(x+t_0v) + df(x+t_0v)\cdot sv = f(x+t_0v) + df(x+t_0v)\cdot v(t-t_0).$$

Que pode ser interpretado como $\varphi(t) \geq \varphi(t_0) + \varphi'(t_0)(t-t_0)$ e pelo lema anterior temos que φ é convexa e consequentemente obtemos que f é convexa.

□

Teorema 1.5.12. *Seja $U \subset \mathbb{R}^m$ aberto e convexo. Uma função duas vezes diferenciável $f : U \to \mathbb{R}$ é convexa se, e somente se, para cada $x \in U$, a Hessiana de f é uma forma quadrática não-negativa, ou seja, $H\cdot v^2 = \sum_{i,j=1}^m \frac{\partial^2 f(x)}{\partial x_i \partial x_j}\alpha_i\alpha_j \geq 0$ para todo $v = (\alpha_1, \alpha_2, \cdots, \alpha_m) \in \mathbb{R}^m$.*

Demonstração. Suponha f convexa e $v \in \mathbb{R}^m - \{0\}$ qualquer, então pelo fato de f ser duas vezes diferenciáveis temos

$$f(x+v) = f(x) + df(x)\cdot v + H\cdot v^2 + r(v)$$

onde $\lim_{v\to 0} \frac{r(v)}{|v|^2} = 0$. Usando o lema anterior, obtemos

$$H \cdot v^2 + r(v) \geq 0 \quad \text{para todo} \quad v \in \mathbb{R}^m - \{0\}$$

dividindo por $|v|^2$ temos

$$H \cdot (\frac{v}{|v|})^2 + \frac{r(v)}{|v|^2} \geq 0.$$

Tomando $v \to 0$, conclui-se $H \cdot u^2 \geq 0$, para todo $u \in S^{m-1} = \{x \in \mathbb{R}^m; |x| = 1\}$ e pela linearidade da Hessiana, concluímos que $H \cdot v^2 \geq 0$, para todo $v \in \mathbb{R}^m$.

Reciprocamente, suponha por absurdo que $f(x + tv) < f(x) + df(x) \cdot tv$ para algum $v \neq 0$. Então pelo fato de f ser duas vezes diferenciável temos

$$H \cdot (vt)^2 + r(tv) < 0 \quad \text{para algum} \quad v \in \mathbb{R}^m.$$

No que acarreta, dividindo ambos os lados por t^2 e fazendo $t \to 0$ temos

$$H \cdot v^2 < 0, \quad \text{para algum} \quad v \in \mathbb{R}^m.$$

O que nos leva em uma contradição.

□

2 Acoplamentos

A teoria de acoplamentos pode ser explorado em varios ramos da probabilidade. O objetivo deste capítulo é apresentar essa teoria e mostrar algumas aplicações de acoplamentos que podem ser usados em cadeias de Markov. Um dos principais resultados do capítulo é a prova do teorema da convergência, usando a técnica de acoplamentos e uma forma de calcular a distância de variação total entre duas distribuições de probabilidades via acoplamentos. Nossa principal referência foi o capítulo 3 e 5 de [13].

2.1 Introdução a tempo de mistura mistura de cadeias de Markov

O objetivo desta seção é discutir sobre a velocidade de convergência da cadeia de Markov. Primeiro, fixado um conjunto Ω finito, definiremos uma distância entre duas distribuições, e assim trabalhar em um espaço métrico no conjunto das distribuições de probabilidade em Ω e faremos algumas aplicações práticas sobre convergências das cadeias de Markov.

Definição 2.1.1. *(distância de variação total) A distância de variação total entre duas distribuições de probabilidades μ e ν em Ω é definido por*

$$\| \mu - \nu \|_{TV} = max_{A \subset \Omega} |\mu(A) - \nu(A)|. \tag{2.1}$$

Vamos mostrar uma proposição que facilitará o cálculo da distância de variação total.

Proposição 2.1.2. *Sejam μ e ν duas distribuições de probabilidade em Ω, então:*

$$\| \mu - \nu \|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|. \tag{2.2}$$

Demonstração. Sejam $B = \{x; \mu(x) \geq \nu(x)\}$ e $A \subset \Omega$ um conjunto qualquer.

$$\mu(A) - \nu(A) \leq \mu(A \cap B) - \nu(A \cap B) \leq \mu(B) - \nu(B).$$

De modo análogo obtemos $\nu(A) - \mu(A) \leq \nu(B^c) - \mu(B^c)$, note que o lado direito das equações são iguais. De fato;
$\mu(B) - \nu(B) = \mu(B) - 1 + 1 - \nu(B) = \nu(B^c) - \mu(B^c)$ assim

$$-(\mu(B) - \nu(B)) = -(\nu(B^c) - \mu(B^c)) \leq -(\nu(A) - \mu(A)) \leq \mu(B) - \nu(B).$$

portanto

$$\| \mu - \nu \|_{TV} = \frac{1}{2}(\mu(B) - \nu(B) + \nu(B^c) - \mu(B^c)) = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \nu(x)|.$$

□

Corolário 2.1.3. *Para quaisquer μ, ν, η vale a desigualdade triangular*

$$\| \mu - \nu \|_{TV} \leq \| \mu - \eta \|_{TV} + \| \eta - \nu \|_{TV} .$$

Demonstração. Segue do fato de que no lado direito da equação (2.2) cumpre a desigualdade triangular. □

Observação 2.1.4. *Nas mesmas condições da proposição anterior, ao observar sua prova, chegamos à seguinte expressão*

$$\| \mu - \nu \|_{TV} = \sum_{x\in\Omega, \mu(x)\geq\nu(x)} \mu(x) - \nu(x). \tag{2.3}$$

Nosso objetivo agora é delimitar a distância máxima (sobre $x_0 \in \Omega$) entre $P^t(x_0, \cdot)$ e π para isso, é conveniente definir.

$$d(t) = \max_{x\in\Omega} \| P^t(x, \cdot) - \pi \|_{TV} . \tag{2.4}$$

$$d'(t) = \max_{x,y\in\Omega} \| P^t(x, \cdot) - P^t(y, \cdot) \|_{TV} . \tag{2.5}$$

Nosso problema de maior interesse é estudar velocidade de convergência de uma determinada cadeia. Uma técnica para resolver este tipo de problema é a técnica do acoplamento. A importância do conceito de acoplamentos entre duas distribuições é que ao invés de comparar duas distribuições, comparamos duas variáveis aleatórias. Em outras palavras, queremos usar a teoria de acoplamentos para limitar $d(t)$. Reservaremos a próxima seção para apresentar este conceito.

2.2 Acoplamento

Nesta seção apresentaremos a definição e alguns exemplos de acoplamentos. Mostraremos a relação que este conceito possui com a distância de variação total e preparar esta técnica para aplicações envolvendo cadeias de Markov que serão abordados na próxima seção.

Definição 2.2.1. *(Acoplamentos) Sejam $(\mathcal{X}, \mathcal{F}_\mathcal{X}, \mu)$ e $(\mathcal{Y}, \mathcal{F}_\mathcal{Y}, \nu)$ dois espaços de probabilidades. Um acoplamento de μ e ν é uma variável aleatória $Z = (X, Y)$ em um espaço de probabilidade (Ω, P), tal que as marginais $X \sim \mu$ e $Y \sim \nu$.*

Se μ e ν são as únicas medidas do problema, então sem perda de generalidade podemos escolher $\Omega = \mathcal{X} \times \mathcal{Y}$. Mais ainda, um acoplamento de μ e ν passa por uma construção de uma medida π em $\mathcal{X} \times \mathcal{Y}$ tal que, satisfaz as seguintes condições, que são equivalentes.

a) $(Proj_{\mathcal{X}})_{\#}\pi = \mu$, $(Proj_{\mathcal{Y}})_{\#}\pi = \nu$, onde $Proj_{\mathcal{X}}$ e $Proj_{\mathcal{Y}}$ são respectivamente aplicações $(x, y) \to x$ e $(x, y) \to y$ e também $(Proj_{\mathcal{X}})_{\#}\pi = \pi(Proj_{\mathcal{X}}^{-1})$ e $(Proj_{\mathcal{Y}})_{\#}\pi = \pi(Proj_{\mathcal{Y}}^{-1})$ são as imagens da medida sobre uma aplicação.

b) Para todo conjunto mensurável $A \subset \mathcal{X}$ e $B \subset \mathcal{Y}$, vale $\pi(A \times \mathcal{Y}) = \mu(A)$ e $\pi(\mathcal{X} \times B) = \nu(B)$.

c) Para toda função mensurável ϕ e ψ em $\mathcal{X}, \mathcal{Y}$, respectivamente

$$\int_{\mathcal{X}\times\mathcal{Y}} (\phi(x) + \psi(y)) d\pi(x, y) = \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y).$$

A medida π das condições acima, por abuso de linguagem, é dita um acoplamento de μ e ν . Denotaremos o conjunto de todas as medidas que são acoplamentos de μ e ν por $ADM(\mu, \nu)$. Essas medidas possuem um papel muito importante na teoria de acoplamentos. Inclusive um dos problemas principais do texto é minimizar uma determinada função cujo o domínio é justamente o conjunto $ADM(\mu, \nu)$.

Exemplo 2.2.2. *Sejam $\Omega = \mathcal{X} = \mathcal{Y} = \{$cara, coroa$\}$, $\mu = \nu = b(1, \frac{1}{2})$, Bernoulli, a medida de probabilidade de em um lançamento de moeda justa com o resultado podendo ser cara e coroa.*

1) Defina duas variáveis aleatórias independentes tal que $\gamma(X = x, Y = y) = \frac{1}{4}$. Note que de fato (X, Y) formam um acoplamento de μ e ν pois

$$\sum_{y\in\mathcal{Y}} \gamma(X = x, Y = y) = \mu(x).$$

$$\sum_{x\in\mathcal{X}} \gamma(X = x, Y = y) = \nu(y).$$

Ou seja, conclui-se que $X \sim \mu$ e $Y \sim \nu$.

2) (X, X) é um outro exemplo de acoplamento de μ e ν.

Dado duas distribuições μ e ν como construir duas variáveis aleatórias, X e Y tal que (X, Y) de fato sejam um acoplamento? Um fato interessante sobre acoplamentos é que podemos fazer uma bijeção entre acoplamentos de duas variáveis aleatórias em Ω com uma única variável aleatória em $\Omega \times \Omega$.

Defina uma distribuição de probabilidade q em $\Omega \times \Omega$ com a seguinte propriedade

$$\sum_{y\in\Omega} q(x,y) = \mu(x) \quad e \quad \sum_{x\in\Omega} q(x,y) = \nu(y).$$

Nessas condições, (X,Y) forma um acoplamento ao considerar $q(x,y) = P(X = x, Y = y)$, em outras palavras, para construir um acoplamento de um par de distribuições, basta preencher as entradas de uma matriz de ordem $|\Omega| \times |\Omega|$ cujo a soma na $i-$ésima linha é a $i-$ésima entrada da primeira distribuição e a soma da j-ésima coluna é a j-ésima entrada da segunda distribuição e construir uma variável aleatória que está associada a distribuição q.

O próximo teorema relaciona acoplamentos de duas medidas com a distância de variação total. Podemos estimar a distância de variação usando variáveis aleatórias, ou seja, transportar um problema de distribuições para variáveis aleatórias. Este teorema é a base das aplicações na estimativa do tempo de mistura.

Teorema 2.2.3. *Sejam μ e ν duas distribuições de probabilidade em Ω, então*

$$\| \mu - \nu \|_{TV} = \inf\{P(X \neq Y); (X,Y) \text{ é acoplamento de } \mu \text{ e } \nu\}.$$

Demonstração. Considere (X,Y) um acoplamento qualquer de μ e ν e seja $A \subset \Omega$ um subconjunto qualquer, temos

$$\mu(A) - \nu(A) = P(X \in A) - P(Y \in A) \leq P(X \in A, Y \notin A) \leq P(X \neq Y).$$

$$\| \mu - \nu \| \leq P(X \neq Y) \quad \forall (X,Y).$$

Resta construir um acoplamento de modo que $\| \mu - \nu \| = P(X \neq Y)$.

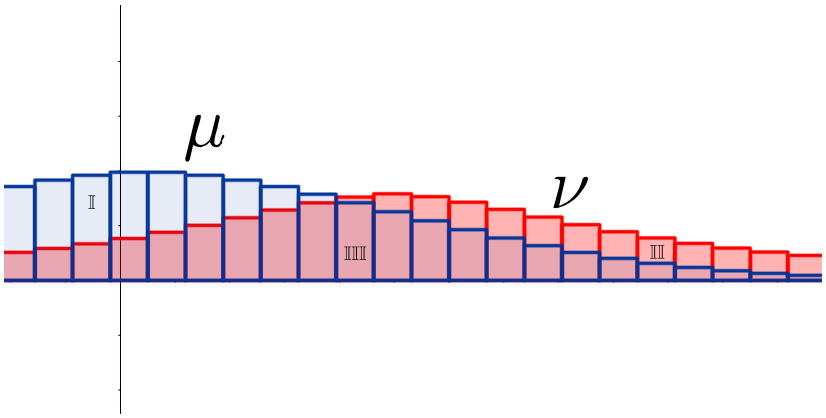


Figura 4 – Sendo $B = \{x; \mu(x) \geq \nu(x)\}$ a região $\mathbb{I}$ tem área $\mu(B) - \nu(B)$. A área da região $\mathbb{III}$ tem área $\nu(B^c) - \mu(B^c)$ e temos que a área da região $\mathbb{IIII}$ é igual a $1 - \| \mu - \nu \|_{TV}$.

A ideia é construir variáveis aleatórias X e Y tal que em $\mathbb{IIII}$, $X = Y$, veja a figura 4, daí considere

$$p = \sum_x \min\{\mu(x), \nu(x)\}.$$

Mas, usando a proposição (2.1.2) obtemos

$$\sum_x \min\{\mu(x), \nu(x)\} = \sum_{x;\mu(x)\geq\nu(x)} \nu(x) + \sum_{x;\nu(x)>\mu(x)} \mu(x) = 1 - \| \mu - \nu \|_{TV} .$$

Considere um lançamento de moeda com probabilidade p de ser cara.

i) Se sair cara, então determine o valor de uma variável aleatória Z com respeito a distribuição de probabilidade

$$\gamma_{\mathbb{III}}(x) = \frac{\min\{\mu(x), \nu(x)\}}{p}.$$

E seja $X = Y = Z$.

ii) Caso a moeda seja coroa, escolha X com distribuição

$$\gamma_{\mathbb{I}}(x) = \begin{cases} \dfrac{\mu(x) - \nu(x)}{\| \mu - \nu \|_{TV}} & \text{se } \mu(x) > \nu(x) \\ \\ 0 & \text{caso contrário.} \end{cases}$$

E de modo independente escolha Y com distribuição

$$\gamma_{\mathbb{II}}(x) = \begin{cases} \dfrac{\nu(x) - \mu(x)}{\| \mu - \nu \|_{TV}} & \text{se } \nu(x) > \mu(x) \\ \\ 0 & \text{caso contrário.} \end{cases}$$

Note que

$$p \cdot \gamma_{\mathbb{III}} + (1-p)\gamma_{\mathbb{I}} = \mu.$$

$$p \cdot \gamma_{\mathbb{III}} + (1-p)\gamma_{\mathbb{II}} = \nu.$$

Ou seja, $X \sim \mu$ e $Y \sim \nu$ com $X = Y$ se, e somente se, o lançamento da moeda sair cara, como $1 - p = \| \mu - \nu \|_{TV}$, obtemos

$$P(X \neq Y) = \| \mu - \nu \|_{TV} .$$

□

2.3 Acoplamentos em Cadeias de Markov

O objetivo desta seção é apresentar aplicações de acoplamentos em cadeias de Markov. Iniciaremos apresentando a noção de acoplamentos em cadeias de Markov e usaremos este conceito para calcular o tempo de mistura de algumas cadeias e encerraremos a seção com a demonstração do teorema da convergência via acoplamentos.

Definição 2.3.1. *(Acoplamentos em cadeias de Markov) Um acoplamento da cadeia de Markov com matriz de transição P é um processo (X_t, Y_t) tal que (X_t) e (Y_t) são cadeias de Markov com a mesma matriz de transição P.*

Todo acoplamento da cadeia de Markov pode-se modificar a cadeia para que as duas cadeias andem juntas após o primeiro encontro, em outras palavras,

$$\text{Se} \quad X_s = Y_s \quad \text{, então para todo} \quad t \geq s \text{ temos } X_t = Y_t. \tag{2.6}$$

Teorema 2.3.2. *Seja $\{(X_t, Y_t)\}$ acoplamentos da cadeia de Markov que cumpre (2.6) com $X_0 \sim \mu$ e $Y_0 \sim \nu$, considere τ_a o primeiro instante em que as cadeias se encontram, ou seja,*

$$\tau_a = \min\{t; X_t = Y_t\}.$$

Então

$$\| \mu P^t - \nu P^t \|_{TV} \leq P(\tau_a > t).$$

Demonstração. Basta observar que $P(X_t \neq Y_t) = P(\tau_a > t)$ e combinar com o teorema (2.2.3) da seção anterior. □

Corolário 2.3.3. *Nas mesmas condições do teorema anterior, com $X_0 \sim \delta_x$ e $Y_0 \sim \delta_y$*

$$d(t) \leq \max_{x,y} P_{x,y}(\tau_a > t).$$

Usaremos o corolário anterior para estimar o tempo de mistura de algumas cadeias. A essência do acoplamento de cadeias de Markov é bem simples. A ideia é usar os resultados dessa seção e construir um conjunto de processos em duas cópias da cadeia de Markov que tem tendência probabilística para andar rapidamente juntas.

Exemplo 2.3.4. *(Passeio aleatório no círculo)*

Considere um passeio aleatório em um n-ciclo ou em $\mathbb{Z}_n$ com o conjunto $\{1, 2, \cdots, n\}$ de vértices. E esse passeio é "preguiçoso", ou seja probabilidade $\frac{1}{2}$ de ficar parado e probabilidade $\frac{1}{4}$ de mover no sentido horário ou anti-horário.

Vamos construir um acoplamento (X_t, Y_t) de duas partículas nesse espaço, onde $X_0 = x$ e $Y_0 = y$. Suponha que as partículas não se movam simultaneamente para que não haja

um salto entre eles. Para cada movimento lançamos uma moeda, se o resultado for cara a cadeia (X_t) anda um passo e a direção será determinada em um lançamento de outra moeda, caso o resultado seja coroa o mesmo ocorre com (Y_t) e quando as partículas colidem eles fazem o mesmo movimento.
Seja D_x a distância no sentido anti-horário entre as duas partículas (de x para y). Note que D_x é passeio aleatório simples entre 0 e n. Usando o estudo da ruína do apostador temos, que se encontra no capítulo 2 da referência [13];

$E_{x,y}(\tau) = k(n-k)$, onde $\tau = min\{t \geq 0; D_x \in \{0, n\}\}$. Daí estudando o tempo de mistura, que denotaremos por $t_{mix}(\varepsilon) = \min\{t; d(t) \leq \varepsilon\}$ e usando a desigualdade de Markov resulta.

$$d(t) \leq max\{P_{x,y}\{\tau > t\}\} \leq max_{x,y}\frac{E_{x,y}(\tau)}{t} \leq \frac{n^2}{4t}.$$

Para $\epsilon = \dfrac{1}{4}$ denotaremos simplesmente por t_{mix}, neste caso obtemos $t = n^2$ onde $t_{mix} \leq n^2$.

Exemplo 2.3.5. *Um toro d-dimensional é um grafo cujo conjunto de vértices é o produto cartesiano*

$$\mathbb{Z}_n^d = \underbrace{\mathbb{Z}_n \times \mathbb{Z}_n \times \cdots \times \mathbb{Z}_n}_{d \text{ vezes}}.$$

Dizemos que os vértices $x = (x^1, x^2, \cdots, x^d)$ e $y = (y^1, y^2, \cdots, y^d)$ são vizinhos em $\mathbb{Z}_n^d$ se existe algum $j \in \{1, 2, \cdots, n\}$ de modo que $x^i = y^i$ para $i \neq j$ e $x^j \equiv y^j + 1 \mod n$ ou $x^j \equiv y^j - 1 \mod n$.

Quando n é par, o grafo $\mathbb{Z}_n^d$ é bipartido e associando a passeio aleatório vemos que é periódico. Para evitar essas complicações, vamos considerar o passeio preguiçoso em $\mathbb{Z}_n^d$, a ideia é usar a teoria de acoplamentos para limitar o tempo de mistura deste passeio.

Teorema 2.3.6. *Em um passeio aleatório preguiçoso no toro d-dimensional $\mathbb{Z}_n^d$*

$$t_{mix}(\varepsilon) \leq c(d) \cdot n^2 \log_2(\varepsilon^{-1}).$$

Onde $c(d)$ é uma constante que depende da dimensão do toro.

Demonstração. Vamos construir o seguinte acoplamento de Cadeias de Markov de (X_t, Y_t) com $X_0 = x$ e $Y_0 = y$ e considerar τ o tempo das cadeias ficarem acopladas, ou seja, $\tau = \min_t\{X_t = Y_t\}$. O passeio será construído do seguinte modo.

1. Escolha uma coordenada aleatoriamente.

2. Caso 1: Se as coordenadas de X_t e Y_t coincidirem, então elas andam juntas e com probabilidade $\frac{1}{4}, \frac{1}{4}$ e $\frac{1}{2}$ andam respectivamente $1, -1$ e 0.

   Caso 2: Se as coordenadas forem distintas, realize um lançamento de moeda para decidir qual cadeia se movimenta e qual fica parada, após a escolha, lance outra moeda para decidir qual será a direção do movimento.

Considere $X_t = (X_t^1, X_t^2, \cdots, X_t^d)$ e $Y_t = (Y_t^1, Y_t^2, \cdots, Y_t^d)$ e seja

$$\tau_i = \min\{t \geq 0; X_t^i = Y_t^i\}.$$

O tempo que leva para a i-ésima coordenada ser acoplada. Note que para cada coordenada, o comportamento da partícula coincide com o passeio aleatório no n-ciclo, como visto no exemplo anterior. Assim, definindo D_t^i como a distância no sentido horário entre X_t^i e Y_t^i. Então $E(D_t^i \in \{0, n\}) \leq \frac{n^2}{4}$.

Mais ainda, definindo $k_i = \min\{D_t^i \in \{0, d\} | \text{A coordenada i é escolhida}\}$, obtemos uma sequência de variáveis aleatórias independentes cuja a distribuição é a geométrica, ou seja, que conta a quantidade de tentativas até chegar ao primeiro sucesso, daí

$$E(\tau_i) = E(\sum_{j=1}^{k_i} X_j) = \sum_{u=1}^{\infty}\sum_{j=1}^{u} X_j P(k_i = u) = \sum_{u=1}^{\infty} E(\sum_{j=1}^{u} X_j) P(k_i = u) =$$

$$= \sum_{u=1}^{\infty} uE(X_1)P(k_i = u) = d\sum_{u=1}^{\infty} uP(k_i = u) = dE(k_i) \leq d\frac{n^2}{4}.$$

Portanto

$$E(\tau) \leq d^2\frac{n^2}{4}.$$

Usando a desigualdade de Markov com o teorema (2.3.2) concluímos que $P(\tau > t) \leq \frac{1}{t}d^2\frac{n^2}{4}$, no que acarreta , $t_{mix} < d^2n^2$ e usando a relação do t_{mix} com $t_{mix}(\varepsilon)$, para maiores detalhes veja na na seção 4.5 da referência [13], temos

$$t_{mix}(\varepsilon) \leq d^2n^2 \log_2(\varepsilon^{-1}).$$

□

Exemplo 2.3.7. *Uma árvore é um grafo conexo sem ciclos. A árvore começa enraizada por um único vértice que chamaremos de raiz. A profundidade do vértice v é a distância de v até a raiz. O nível da árvore consiste em todos os vértices de mesma profundidade. Os filhos de v são os vizinhos de v com a profundidade de v + 1. A folha são os vértices de grau 1. Uma árvore b-ária com profundidade k, denotado por $T_{b,k}$, é uma árvore com vértice v_0 sendo a raiz com.*

- v_0 *tem grau* b.
- *Todo vértice de profundidade* j, *com* $1 \leq j \leq k-1$, *tem grau* $b+1$.
- *Vértices de profundidades* k *são folhas.*

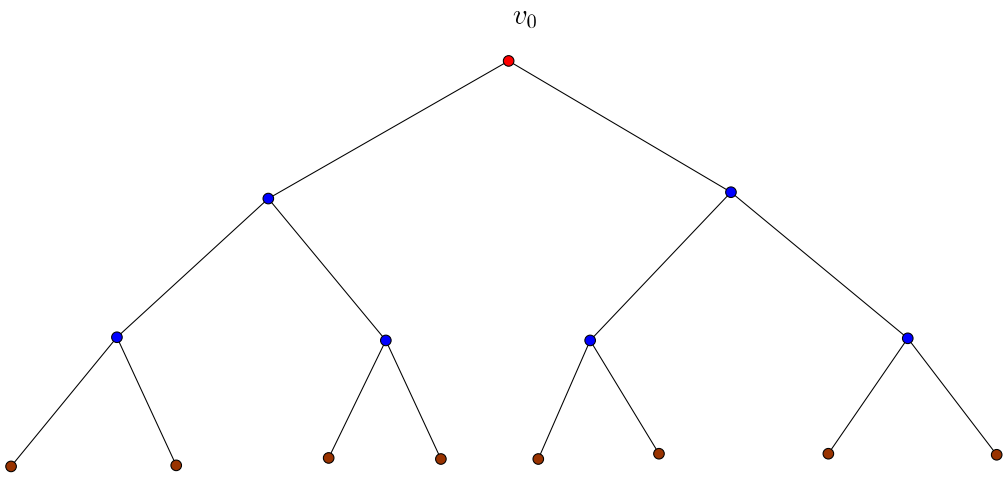


Figura 5 – Uma árvore binária com grau 3.

Assim o número total de vértices é $n = \dfrac{b^{k+1}-1}{b-1}$. *Vamos estudar o passeio preguiçoso em* $T_{b,k}$.

Considere o seguinte acoplamento (X_t, Y_t), *com* $X_0 = x$ *e* $Y_0 = y$. *Assuma, sem perda de generalidade, que* x_0 *está mais próximo da raiz do que* y_0. *A cada movimento lance uma moeda para decidir quem se move, se sair cara* $Y_{t+1} = Y_t$ *e* X_{t+1} *anda para um vizinho de* X_t *de modo aleatório e uniforme, se sair coroa, então* $X_{t+1} = X_t$ *e* Y_{t+1} *anda para um vizinho de* Y_t.

Quando as cadeias estiverem no mesmo nível, mude a dinâmica. Considere X_t *em um passeio preguiçoso e* X_t *se aproxima da raiz se, e somente se,* Y_t *se aproxima da raiz. Seja* L *o conjunto das folhas. Observe que se* X_t *faz uma visita na folha e retorna a raiz, então as cadeias se acoplam. A esperança do número de passos para sair da folha e chegar na raiz é a mesma de iniciar da raiz e chegar na folha e digamos que* τ *é o tempo para que este evento ocorra. Por argumentos de redes, veja o capítulo 9 da referência [13] temos* $E(\tau) \leq 4n$. *Então*

$$P_{xy}\{\tau_a > t\} \leq \frac{4n}{t}.$$

O que nos leva a concluir que $t_{mix} \leq 16n$.

Exemplo 2.3.8. *Seja V e C dois conjuntos finitos com $|V| = n$ e $|C| = k$, sendo V o conjunto que representa os vértices e C as cores. Considere o espaço de estados $\Omega \subset C^V$ e uma cadeia de Markov M em Ω com uma única distribuição estacionária π. Intuitivamente cada estado é uma coloração de vértices, cuja transição de estados ocorre da seguinte maneira.*

i) Selecione um vértice $v \in V$ de acordo com uma distribuição J fixada e uma cor $c \in C$ de acordo com uma distribuição $K_{X,v}$ em C que depende da configuração atual X e do vértice v

ii) O novo estado $X_{v\to c}$ será definido por $X_{v\to c}(w) = \begin{cases} c & se \quad w = v \\ X(w) & se \quad w \neq v \end{cases}$

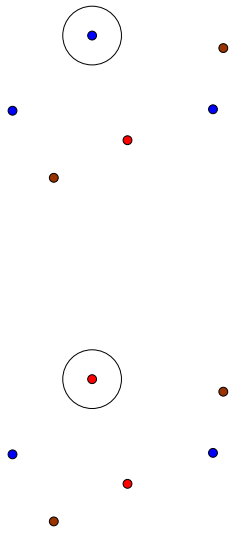


Figura 6 – Como funciona as transições da cadeia.

Considere pares de estados que são adjacentes em algum caminho e a métrica de Hamming com $H(X,Y) = |\{v \in V; X(v) \neq Y(v)\}|$.

Teorema 2.3.9. *Seja $\Omega = C^V$, e $\beta = max_{X,Y\in\Omega, i\in V}\{1 - J(i) + \sum_{j\in V} J(j) \parallel (K_{X_j} - K_{Y_j} \parallel_{TV}; Y = X_{i\to c} \quad para \quad algum \quad c \quad e \quad X \neq Y\}$. Então, se $\beta < 1$ e $t \geq \frac{ln(n\epsilon^{-1})}{ln(\beta^{-1})}$, conclui-se que $\parallel \mu_t - \pi \parallel_{TV} \leq \epsilon$.*

Uma explicação sobre β, é que este valor é uma cota superior da esperança da distância entre estados adjacentes após um passo.

Demonstração. Suponha que σ_1 e σ_2 são duas distribuições de probabilidades em C. Então definimos a seguinte distribuição

$$(\sigma_1 - \sigma_2)^+(c) = \frac{max\{0, \sigma_1(c) - \sigma_2(c)\}}{d_{TV}(\sigma_1, \sigma_2)}.$$

Para $X, Y \in C^V$, seja $H(X,Y)$ a distância de Hamming. Considere $h = H(X,Y)$ e uma sequência $X = Z_0, Z_1, \cdots, Z_h = Y$ em C^V tal que $H(Z_{a-1}, Z_a) = 1$ para $a \in \{1, 2, \cdots, h\}$.

Defina um acoplamento em M dos estados (X,Y) para (X',Y') de acordo com o seguinte experimento.

1) Escolha $j \in V$ de acordo com J e $c_0 \in C$ de acordo com $K_{Z_0,j}$.

2) Para cada $a \in \{1, 2, \cdots, h\}$, com probabilidade $1 - \parallel K_{Z_a,j} - K_{Z_{a-1},j} \parallel_{TV}$, seja $c_a = c_{a-1}$ caso contrário selecione c_a de acordo com $(K_{Z_a,j} - K_{Z_{a-1},j})^+$.

3) Assim o estado é atualizado para (X',Y'), onde $X' = X_{v \to c_0}$ e $Y' = Y_{v \to c_h}$.

Note que a marginal de escolher c_a de acordo com $K_{Z_a,j}$ é igual ao acoplamento ótimo entre $K_{Z_a,j}$ e $K_{Z_{a-1},j}$, em outras palavras, $P(c_a \neq c_{a-1}) = \parallel K_{Z_a,j} - K_{Z_{a-1},j} \parallel_{TV}$.

Usaremos o símbolo Z'_a para representar a atualização $(Z_a)_{v \to c_a}$ e suponha que Z_a e Z_{a-1} diferem no vértice i. Então

$$E(H(Z'_a, Z'_{a-1})) = 1 \cdot P(H(Z'_a, Z'_{a-1}) = 1) + 2 \cdot P(H(Z'_a, Z'_{a-1}) = 2)$$

$$= 1 - P(H(Z'_a, Z'_{a-1}) = 0) + P(H(Z'_a, Z'_{a-1}) = 2) =$$

$$1 - J(i)P(c_a = c_{a-1} | V = i) + \sum_{j \neq i} J(j) P(c_a \neq c_{a-1} | V = j) =$$

$$1 - J(i)[1 - \parallel (K_{Z_a,i} - K_{Z_{a-1},i} \parallel_{TV} + \sum_{j \neq i} J(j) \parallel K_{Z_a,j} - K_{Z_{a-1},j} \parallel_{TV} \leq \beta.$$

Assim conclui-se que

$$E(H(X',Y')) = E(\sum_{a=1}^{h} H(Z'_a, Z'_{a-1})) = \sum_{a=1}^{h} E(H(Z'_a, Z'_{a-1})) \leq \beta h = \beta H(X,Y). \quad (2.7)$$

Ou seja, temos que H é uma contração pela esperança. Ainda obtemos $E(X_2, Y_2) \leq \beta H(X',Y')$ e aplicando a esperança em ambos dos lados, usando as propriedades da esperança e a equação (2.7) obtém-se

$$E(H(X_2, Y_2)) \leq \beta^2 H(X,Y).$$

$$E(H(X_t, Y_t)) \leq \beta^t H(X,Y) \leq \beta^t n.$$

Logo $P(X_t \neq Y_t) \leq \beta^t n$ obtemos pelo Teorema 2.2.3 $\parallel \mu_t - \pi \parallel_{TV} \leq \beta^t n$ e através de manipulações com logaritmo conclui-se que $\parallel \mu_t - \pi \parallel_{TV} \leq \epsilon$.

□

O Teorema da Convergência por si só é visto como um resultado muito interessante na teoria de cadeias de Markov. Sabemos que ao considerar uma cadeia de Markov com transição P, se P for irredutível e aperiódica, então existe uma única distribuição estacionária. O Teorema da convergência estima o tempo que leva para que a distribuição convirja para a estacionária. Antes de enunciar e demonstrar o Teorema da convergência, vamos apresentar a noção de cadeia de Markov aperiódica

Definição 2.3.10. *(Cadeia aperiódica) Considere uma cadeia de Markov com matriz de transição P e seja $\psi(x) = \{t > 0; P^t(x,x)\}$ o conjunto dos tempos possíveis da cadeia voltar para o instante inicial x. O período do estado x é definido pelo máximo divisor comum de $\psi(x)$ e caso o período seja igual a 1 para todos os pontos $x \in \Omega$, então dizemos que a cadeia é aperiódica.*

Teorema 2.3.11. *(Teorema da Convergência) Suponha que P é irredutível e aperiódica com distribuição estacionária π. Então existe uma constante $\alpha \in (0,1)$ e $c > 0$ tal que*

$$max_{x\in\Omega} \parallel P^t(x,\cdot) - \pi \parallel_{TV} \leq c\alpha^t. \tag{2.8}$$

Demonstração. Usando o teorema (2.3.2) com $\nu = \pi$ e $\mu = \delta_x$ chegamos que

$$\parallel \pi - p^t(x,\cdot) \parallel \leq P_x(\tau_a > t). \tag{2.9}$$

Vamos mostrar que $P_x(\tau_a < \infty) = 1$ para todo $x \in \Omega$, para isso considere o acoplamento (X_t, Y_t) de duas cadeias com $X_0 \sim \delta_x$ e $Y_0 \sim \pi$. Como P é irredutível e aperiódica, usando a proposição 1.7 da referência [13] temos que existe $r, \beta = \min P^r(x,y) > 0$ fixando $x_0 \in \Omega$ e denotando $\{X_r \neq x_0, Y_r \neq x_0\} = A_r$, obtém-se que

$$P(A_r) \leq 1 - \beta.$$

Usando noções básicas de probabilidade condicional temos

$$P(A_{2r}|A_r) \leq 1 - \beta.$$

$$P(A_{2r}) \leq (1-\beta)^2.$$

E usando indução, conclui-se que

$$P(A_{kr}) \leq (1-\beta)^k.$$

Portanto quando k tende ao infinito, então $P(\tau > kr)$ tende a zero, ou seja, $P(\tau_a < \infty) = 1$ e ainda, usando a noção de divisão Euclidiana podemos escrever $t = kr + r_0$ e assim

$$P(\tau_a > t) = P(\tau_a > kr + r_0) \leq P(\tau_a > kr, \tau_a > r_0) =$$

$$= P(\tau_a > kr) \leq (1-\beta)^k.$$

$$P(\tau_a > t) \leq (1-\beta)^k = (1-\beta)^{\frac{t-r_0}{r}} = (1-\beta)^{\frac{t}{r}} \cdot (1-\beta)^{\frac{-r_0}{r}} \leq$$
$$\leq ((1-\beta)^{\frac{1}{r}})^t \cdot c.$$

Por (2.9) e substituindo $(1-\beta)^{\frac{1}{r}}$ por α, conclui-se que

$$\| \pi - P^t(x,\cdot) \| \leq c \cdot \alpha^t.$$

□

3 Problema de Monge Kantorovich

Os problemas de Monge e Kantorovich estão contextualizados na teoria do transporte ótimo, que nasceu na França em 1781, com o artigo de Monge. Essa teoria ganhou um destaque devido a diversos pesquisadores em diferentes áreas da Matemática ter relacionado em suas pesquisas. Estes problemas não são apenas importantes na Matemática, em 1975 Kantorovich e Tjalling Koopmans ganharam o prêmio Nobel da economia pelas suas contribuições nesses problemas. Neste capítulo vamos apresentar os problemas de Monge e de Kantorovich, verificar a relação entre os problemas, encontrar condições para que o problema tenha solução e mostrar uma aplicação destes problemas que é demonstrar a desigualdade Isoperimétrica. As principais referências desse capítulo são [1] e [16].

Neste capítulo faremos uso de conceitos estudados como Topologia, Função Semicontína e Análise Convexa no decorrer dos resultados, para demonstração dos principais resultados desse trabalho, como o Teorema Fundamental do Transporte Ótimo e o Teorema de Brenier.

3.1 Monge X Kantorovich

Relembre que, um acoplamento de μ e ν é a construção de duas variáveis aleatórias X e Y em um mesmo espaço de probabilidade (Ω, P), tal que $X \sim \mu$ e $Y \sim \nu$.

O objetivo deste capítulo é apresentar o problema de Monge e o problema de Kantorovich, que dados duas medidas, minimizar uma determinada função cujo domínio é o conjunto de todos os acoplamentos. Esses problemas nos levam a estudar um pouco mais sobre acoplamentos, retornaremos com o seguinte exemplo.

Exemplo 3.1.1. *Sejam $\mu = \nu = b(1, \frac{1}{2})$ a medida de probabilidade, que podemos associar a um lançamento de uma moeda justa e fixe os seguintes espaços de probabilidade $(\mathcal{X}, \mathcal{F}_1, \mu)$, $(\mathcal{Y}, \mathcal{F}_2, \nu)$ e $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_1 \times \mathcal{F}_2, \gamma)$.*

1) Defina duas variáveis aleatórias independentes tal que $\gamma(X = x, Y = y) = \frac{1}{4}$. Note que de fato (X, Y) formam um acoplamento de μ e ν pois

$$\sum_{y \in \mathcal{Y}} \gamma(X = x, Y = y) = \mu(x).$$

$$\sum_{x \in \mathcal{X}} \gamma(X = x, Y = y) = \nu(y).$$

Ou seja, conclui-se que $X \sim \mu$ e $Y \sim \nu$.

2) $X = Y$ é um outro exemplo de acoplamento de μ e ν.

Dada duas distribuições de probabilidades μ e ν sempre é possível realizar um acoplamento, por exemplo $\pi = \mu \times \nu$, o primeiro acoplamento do exemplo anterior, que é um tipo de acoplamento trivial.

Observando com mais detalhes, pode-se notar que no primeiro acoplamento do exemplo acima, o resultado da variável aleatória X não traz nenhuma informação nova sobre a variável Y. Já o segundo caso, a informação do resultado de X determina completamente o resultado de Y e esse fato motiva a nossa próxima definição.

Definição 3.1.2. *(Acoplamentos determinísticos) Um Acoplamento (X, Y) é dito determinístico se existe uma função mensurável $T : \mathcal{X} \to \mathcal{Y}$ tal que $T(X) = Y$.*

Exemplo 3.1.3. *Considere $\Omega = \{cara, coroa\}$, $\mu = \nu = ber(\frac{1}{2})$ e X, Y variáveis aleatórias em Ω tal que*

$$X(w) = \begin{cases} -1, & se\ w = cara. \\ 1, & se\ w = coroa. \end{cases} \qquad Y(w) = \begin{cases} 1, & se\ w = cara. \\ -1, & se\ w = coroa. \end{cases}$$

Neste caso (X, Y) é um acoplamento de μ e ν desde que $X \sim \mu$ e $Y \sim \nu$, a conclusão deste exemplo é que o acoplamento (X, Y) é determinístico com $Y = T(X) = -X$.

Dizer que um acoplamento (X, Y) é determinístico é equivalente a qualquer uma das condições abaixo

a) (X, Y) é um acoplamento de μ e ν cuja distribuição π é concentrada em um gráfico de uma função mensurável $T : \mathcal{X} \to \mathcal{Y}$.

b) $X \sim \mu$ e $Y = T(X)$, onde $T_{\#}\mu = \nu$. (Onde $T_{\#}\mu = \mu \circ T^{-1}$).

c) $X \sim \mu$ e $Y = T(X)$, onde T é a mudança de variáveis de μ para ν, ou seja, para toda função $\nu-$integrável vale:

$$\int_{\mathcal{Y}} \phi(y)\ d\nu(y) = \int_{\mathcal{X}} \phi(T(x))\ d\mu(x).$$

d) $\pi = (Id, T)_{\#}\mu$.

A aplicação T que aparece nas sentenças acima é chamada aplicação transporte induzido por T. Informalmente, a aplicação T transporta toda massa em x, representada com medida $\mu(x)$, para o compartimento $y = T(x)$ de medida $\nu(y)$.

O problema de Monge, informalmente, é fixado duas medidas de probabilidades, uma em um conjunto $\mathcal{X}$ e outro em um conjunto $\mathcal{Y}$, encontrar uma aplicação $T : \mathcal{X} \to \mathcal{Y}$ de modo que toda a massa em $\mathcal{X}$ seja transportada para $\mathcal{Y}$ com o menor custo possível. Segue abaixo o problema de Monge de modo formal.

Problema 3.1.4. *(Monge) Sejam $\mu \in \mathcal{P}(\mathcal{X})$ e $\nu \in \mathcal{P}(\mathcal{Y})$ duas distribuições de probabilidade e considere uma função, chamada custo, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$. O problema de Monge é minimizar a função*

$$T \to \int_{\mathcal{X}} c(x, T(x))\ d\mu(x). \tag{3.1}$$

Sobre todas as aplicações transporte T tal que $T_{\#}\mu = \nu$. Em outras palavras, queremos minimizar sob todos os acoplamentos determinísticos de μ e ν a função

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x, y)d\pi(x, y) \quad \textit{sendo} \quad \pi = (Id, T)_{\#}\mu. \tag{3.2}$$

Independente da escolha da função custo c, o problema de Monge pode não ser bem posto por que:

- A aplicação transporte, T, pode não existir, por exemplo, seja μ uma Dirac e ν uma não Dirac, ao supor que existe uma aplicação T que satisfaz o item (b) nas condições acima conclui-se que ν é uma Dirac o que é um absurdo.

- A restrição $T_{\#}\mu = \nu$ pode não ter uma sequência fracamente fechado, com respeito a topologia fraca.

Um fato importante na teoria de acoplamentos, nem todo par de distribuições μ e ν possui acoplamentos determinístico. Uma maneira de superar essas dificuldades é relaxar o problema, minimizando a função sob todos os acoplamentos ao invez de minimizar com respeito a todos os acoplamentos determinísticos. Com essa motivação será apresentado o problema de Kantorovich.

Problema 3.1.5. *:(Kantorovich) Nas mesmas condições do problema anterior, queremos minimizar a função em $ADM(\mu, \nu)$, que é o conjunto de todos os acoplamentos de μ e ν.*

$$\gamma \to \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)d\gamma(x, y) \quad \textit{sendo} \quad \gamma \in ADM(\mu, \nu). \tag{3.3}$$

Lembrando que, quando dizemos γ é um acoplamento de μ e ν temos.

$$\gamma(A \times \mathcal{Y}) = \mu(A) \quad \forall A \in \mathcal{B}(\mathcal{X}).$$

$$\gamma(\mathcal{X} \times B) = \nu(B) \quad \forall B \in \mathcal{B}(\mathcal{Y}).$$

Equivalentemente tem-se que $(Proj_{\mathcal{X}})_{\#}\gamma = \mu$ e $(Proj_{\mathcal{Y}})_{\#}\gamma = \nu$. O valor de $\gamma(A \times B)$ representa a quantidade de massa que é levada de A para B. Algumas das vantagens do problema de Kantorovich são:

- $ADM(\mu, \nu)$ nunca é vazio, pois contém a medida produto $\mu \times \nu$, ou seja, $\gamma(A \times B) = \mu(A) \cdot \nu(B)$.
- Plano de transporte, ou seja, acoplamentos inclui acoplamentos determinísticos, desde $T_{\#}\mu = \nu$ implica que $\gamma = (Id \times T)_{\#}\mu \in ADM(\mu, \nu)$.
- Mínimo sempre existe sob certas hipóteses na função custo, como veremos no Teorema (3.1.12).

Geometricamente, o que diferencia o problema de Monge para o de Kantorovich é que no problema de Monge cada massa $x \in \mathcal{X}$ é totalmente transportada para algum $T(x) = y \in \mathcal{Y}$, já no problema de Kantorovich é possível que exista algum $x \in \mathcal{X}$ na qual a sua produção é dividida e distribuída em mais filiais, digamos $y_1, y_2, \cdots y_n \in \mathcal{Y}$. Veja na figura abaixo a comparação dos problemas.

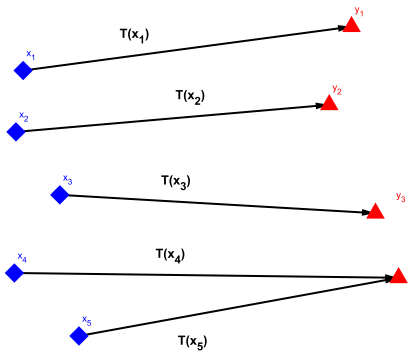


Figura 7 – Problema de Monge

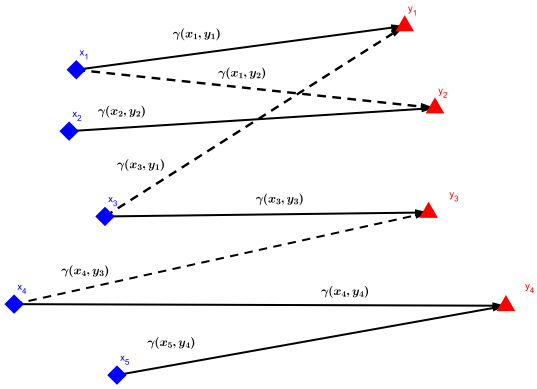


Figura 8 – Problema de Kantorovich

O teorema que será provado no final desta seção nos diz que, nas condições do problema de Kantorovich quando a função custo c é semicontínua inferior, então existe o acoplamento que satisfaz o mínimo em (3.3). A ideia da prova do teorema passa pelo fato de que uma função semicontínua inferior em um compacto assume mínimo. Daí, é preciso fazer um estudo sobre acoplamentos e espaços Polonês.

Uma observação importante é que pelo fato do espaço $\mathcal{X}$ ser separável, temos que a topologia fraca é metrizável, para mais detalhes veja [1].

Definição 3.1.6. *(Tight) Uma família $\mathcal{K} \subset \mathcal{P}(\mathcal{X})$ é dita tight se para todo $\epsilon > 0$ existe um conjunto compacto, que depende de ϵ, $K_\epsilon \subset \mathcal{X}$ tal que para toda medida $\mu \in \mathcal{K}$ tem-se*

$$\mu(\mathcal{X} - K_\epsilon) \leq \epsilon.$$

Exemplo 3.1.7. *Quando $\mathcal{X}$ é um espaço Polonês (ou seja, espaço métrico separável e completo). $\mathcal{K} = \{\mu\}$ é tight. De fato, como $\mathcal{X}$ é separável então existem elementos $\{x_1, x_2, \cdots\}$ tais que fixado $\varepsilon > 0$*

$$\mathcal{X} = \bigcup_{n=1}^{\infty} \overline{B}(x_n, \varepsilon).$$

Assim considerando $K_m = \bigcup_{n=1}^{m} \overline{B}(x_n, \epsilon)$ tem-se que a medida μ do complementar de K_m tende a zero. Como união finita de compactos é compacto usando a definição de convergência, conclui-se que μ é tight.

Exemplo 3.1.8. *Nas mesmas condições do exemplo anterior. $ADM(\mu, \nu)$ é tight, pois basta observar que para cada $\gamma \in ADM(\mu, \nu)$ vale*

$$\gamma(\mathcal{X} - K_1 \times \mathcal{Y} - K_2) \leq \gamma(\mathcal{X} - K_1 \times \mathcal{Y}) + \gamma(\mathcal{X} \times \mathcal{Y} - K_2) = \mu(\mathcal{X} - K_1) + \nu(\mathcal{Y} - K_2).$$

Então de sorte que μ e ν são tight, chegamos a conclusão de que $ADM(\mu, \nu)$ também o é.

Exemplo 3.1.9. *$\mathcal{X} = (0, 1)$ considere a família $\mathcal{K} = \{\delta_x; x \in \mathcal{X}\}$. Afirmamos que $\mathcal{K}$ não é tight. De fato,*

Para cada compacto K e para $0 < \epsilon < 1$, existe um elemento x_K no complementar de K pois $(0, 1)$ não é compacto e assim obtemos que $\delta_{x_k}(\mathcal{X} - K) = 1 \geq \epsilon$.

O teorema a seguir faz uma conexão entre uma família tight com conjuntos relativamente compactos (ou seja, o fecho do conjunto é compacto). Essa relação é um passo importante para nossa demonstração do teorema de existência de mínimo no problema de Kantorovich. O teorema enunciado a seguir por si só tem uma grande importância em Análise.

Teorema 3.1.10. *(Prohorov) Seja $(\mathcal{X}, d)$ um espaço métrico Polonês. Então uma família $\mathcal{K} \subset \mathcal{P}(\mathcal{X})$ é relativamente compacto com respeito a topologia fraca se, e somente se, é tight.*

Demonstração. Vamos provar apenas que se $\mathcal{K}$ é relativamente compacto, então $\mathcal{K}$ é tight, em outras palavras, para cada $\varepsilon > 0$, deve-se encontrar um conjunto compacto $K \subset \mathcal{X}$ tal que para todo $\mu \in \mathcal{K}$.

$$\mu(\mathcal{X} - K) < \varepsilon.$$

Para justificar este resultado, será usado o seguinte roteiro.

- Fixe $\varepsilon > 0$, então que para toda cobertura por abertos $\{U_i\}_{i\in\mathbb{N}}$ de $\mathcal{X}$, existe um inteiro k tal que para toda medida $\mu \in \mathcal{K}$.

$$\mu(\bigcup_{i=1}^{k} U_i) > 1 - \varepsilon.$$

Caso contrário, existe $\varepsilon > 0$ tal que para todo k, existe uma medida $\mu_k \in \mathcal{K}$ tal que

$$\mu_k(\bigcup_{i=1}^{k} U_i) \leq 1 - \varepsilon.$$

Assim tomando quando necessário uma subsequência de $\{\mu_k\}$, existe uma medida μ tal que $\mu_k \to \mu$ quando k tende ao infinito. Daí pela proposição (1.3.16)

$$\mu(\bigcup_{i=1}^{n} U_i) \leq \liminf_{j\to\infty} \mu_{k_j}(\bigcup_{i=1}^{n} U_i) \leq \liminf_{j\to\infty} \mu_{k_j}(\bigcup_{i=1}^{k_j} U_i) \leq 1 - \varepsilon \quad \text{para todo n} \in \mathbb{N}.$$

Mas Como $\mathcal{X} = \bigcup_{i=1}^{\infty} U_i$ chegamos em um absurdo, pois $1 = \mu(\mathcal{X}) = \lim_{n\to\infty} \mu(\bigcup_{i=1}^{n} U_i) \leq \lim_{n\to\infty} 1 - \varepsilon = 1 - \varepsilon$.

- Usando que $\mathcal{X}$ é separável, existe um conjunto enumerável $D = \{a_1, \cdots\}$ denso em $\mathcal{X}$. A partir do conjunto D, construiremos uma cobertura de $\mathcal{X}$, a saber, para cada $m \geq 1$ consideremos a cobertura formada por $\bigcup_{i=1}^{\infty} B(a_i, \frac{1}{m})$.

- Pela primeira afirmação obtemos para cada $m \geq 1$, existe k_m tal que para toda medida $\mu \in \mathcal{K}$ tem-se

$$\mu(\bigcup_{i=1}^{k_m} B(a_i, \frac{1}{m})) > 1 - \varepsilon 2^{-m}.$$

- Considere o conjunto $K = \bigcap_{m=1}^{\infty} \bigcup_{i=1}^{k_m} \overline{B}(a_i, \frac{1}{m})$ e observe que K é fechado, pois união finita de fechado é fechado e a interseção enumerável de fechados é fechados. Mais ainda, é possível concluir que K é limitado, pois para cada $\delta > 0$ podemos tomar $m > \frac{1}{\delta}$ e obter que $K \subset \bigcup_{i=1}^{k_m} B(a_i, \delta)$ no que acarreta que K é totalmente limitado e portanto é limitado e assim K é compacto com respeito a topologia fraca, obtemos assim

$$\mu(\mathcal{X} - K) = \mu(\bigcup_{m=1}^{\infty} [\bigcup_{i=1}^{k_m} \overline{B}(a_i, \frac{1}{m})]^c) \leq \sum_{m=1}^{\infty} \mu(\bigcup_{i=1}^{k_m} \overline{B}(a_i, \frac{1}{m})^c) =$$

$$= \sum_{m=1}^{\infty} 1 - \mu(\bigcup_{i=1}^{k_m} \overline{B}(a_i, \frac{1}{m})) < \sum_{m=1}^{\infty} \varepsilon 2^{-m} = \varepsilon.$$

Portanto, conclui-se que $\mathcal{K}$ é tight. Conclusão, se $\mathcal{K}$ é relativamente compacto, então $\mathcal{K}$ é tight.

A prova da reciproca pode ser encontrada na referência [4] na seção 5 do capítulo 1.

□

Lema 3.1.11. *Considere $\mathcal{X}$ e $\mathcal{Y}$ espaços Poloneses, então conjunto $ADM(\mu,\nu)$ é convexo e compacto com respeito a topologia fraca.*

Demonstração. Compacidade com respeito a topologia fraca segue do teorema de Prohorov combinado com o exemplo 3.1.8 e o fato de que o conjunto $ADM(\mu,\nu)$ ser fechado. Resta justificar a última afirmativa, seja $\{\gamma_n\}_{n\in\mathbb{N}}$ uma sequência em $ADM(\mu,\nu)$ com $\gamma_n \to \gamma$, note que

$$\gamma(A\times\mathcal{Y}) = \lim_{n\to\infty}\gamma_n(A\times\mathcal{Y}) = \lim_{n\to\infty}\mu(A)$$

$$\gamma(\mathcal{X}\times B) = \lim_{n\to\infty}\gamma_n(\mathcal{X}\times B) = \lim_{n\to\infty}\nu(b)$$

portanto $\gamma \in ADM(\mu,\nu)$. Para justificar a Convexidade considere γ_1 e $\gamma_2 \in ADM(\mu,\nu)$ arbitrários e para cada $t\in(0,1)$ defina σ_t por

$$\sigma_t(\cdot) = t\gamma_1(\cdot) + (1-t)\gamma_2(\cdot).$$

Vamos verificar se σ_t pertence ao $ADM(\mu,\nu)$ para todo t, qualquer que seja o conjunto mensurável $A\subset\mathcal{X}$ e $\mathcal{B}\subset\mathcal{Y}$ temos

$$\sigma_t(A\times\mathcal{Y}) = t\cdot\gamma_1(A\times\mathcal{Y}) + (1-t)\cdot\gamma_2(A\times\mathcal{Y}) = t\cdot\mu(A) + (1-t)\cdot\mu(A) = \mu(A).$$

De forma análoga conclui-se que $\sigma_t(\mathcal{X}\times B) = \nu(B)$. Portanto σ_t pertence a $ADM(\mu,\nu)$ para todo t em $(0,1)$.

□

Teorema 3.1.12. *Seja $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\cup\{\infty\}$ uma função semicontínua inferior e limitada inferiormente, então existe um acoplamento em $ADM(\mu,\nu)$ que minimiza a função $\psi(\gamma) = \int c(x,y)\ d\gamma$.*

Demonstração. Dividiremos a demonstração nos seguintes passos

- Usando o lema 3.1.11, temos que o conjunto $ADM(\mu,\nu)$ é compacto com respeito a topologia fraca.
- Pela proposição (1.4.5), existe uma sequência de funções contínuas $\{c_k\}$ tal que $c_k \nearrow c$.

- Considere $\psi_k(\gamma) = \int c_k(x,y)\ d\gamma$ em $ADM(\mu,\nu)$, note que ψ_k é contínua para todo $k \in \mathbb{N}$, usando o teorema da convergência monótona, obtemos $\psi_k \nearrow \psi$, que nos leva a concluir, também pela proposição (1.4.5) que ψ é uma função semicontínua inferior.

- Pelo item anterior temos que ψ é uma função semicontínua inferior, usando a compacidade de $ADM(\mu,\nu)$ e o teorema (1.4.3) nos garante que existe um γ^* que assume valor mínimo da função ψ.

□

Vale ressaltar no teorema acima que se a função custo for semicontínua, então existe um acoplamento que resolve o Problema de Kantorovich. O exemplo a seguir mostra um exemplo no problema de Kantorovich cujo o custo é uma função semicontínua inferior e não contínua e mesmo assim podemos garantir a existência do acoplamento ótimo.

Exemplo 3.1.13. *Seja $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, considere o Problema de Kantorovich para as medidas μ e ν com*

$$c(x,y) = \begin{cases} 0 & \text{se } (x,y) = (0,0) \\ 1 & \text{caso contrário} \end{cases}.$$

Então o teorema nos garante que existe um acoplamento ótimo e mais ainda basta considerar qualquer acoplamento que concentre a maior massa possível em $(0,0)$, ou seja, basta considerar γ um acoplamento de μ e ν com $\gamma(0,0) = \min\{\mu(0), \nu(0)\}$.

Um fato importante na teoria de acoplamentos é que todo sub-acoplamento (acoplamento restrito a um subconjunto de $\mathcal{X} \times \mathcal{Y}$) de um acoplamento ótimo é ainda ótimo, em outras palavras, para cada transporte ótimo, qualquer sub-plano é ainda ótima. Este fato decorre do teorema a seguir.

Teorema 3.1.14. *Sejam $(\mathcal{X},\mu)$ e $(\mathcal{Y},\nu)$ dois espaços poloneses, $a \in L^1(\mu)$, $b \in L^1(\nu)$, recorde que $L^1(\mu) = \{f : \mathcal{X} \to \mathbb{R}; f \text{ é mensurável}, \int_{\mathcal{X}} |f|\ d\mu < \infty\}$. Considere $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ função mensurável tal que $c(x,y) \geq a(x) + b(y)$. Seja $C(\mu,\nu)$ o custo do transporte ótimo de μ e ν, assumindo que este custo é finito, considere ainda $\pi \in ADM(\mu,\nu)$ um plano de transporte ótimo e π' uma medida não negativa em $\mathcal{X} \times \mathcal{Y}$, tal que $\pi' \leq \pi$ e $\pi'(\mathcal{X} \times \mathcal{Y}) > 0$. Então*

$$\pi\text{"} = \frac{\pi'}{\pi'(\mathcal{X} \times \mathcal{Y})}$$

é um plano de transporte ótimo entre suas marginais μ' e ν'. Mais ainda, se π é o único plano ótimo de μ e ν, então π" é o único plano ótimo entre μ' e ν'.

Demonstração. Suponha que π'' não seja um transporte ótimo, então existe uma medida γ tal que

$$\begin{cases} (Proj_{\mathcal{X}})_{\#}\gamma = (Proj_{\mathcal{X}})\pi" = \mu'. \\ (Proj_{\mathcal{Y}})_{\#}\gamma = (Proj_{\mathcal{Y}})\pi" = \nu'. \end{cases}$$

e

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\ d\gamma(x,y) < \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\ d\pi".$$

Considere $\overline{\pi} = (\pi - \pi') + Z'\gamma$, onde $Z' = \pi'(\mathcal{X}\times\mathcal{Y}) > 0$. Assim $\overline{\pi}$ tem a mesma marginal de π e ainda temos que o plano $\overline{\pi}$ tem um custo menor do que π, o que gera uma contradição. Conclusão π" é um plano ótimo e a unicidade se justifica de modo análoga. □

Exemplo 3.1.15. *Se (X,Y) é um acoplamento ótimo de μ e ν, e $\mathcal{Z} \subset \mathcal{X}\times\mathcal{Y}$ tal que $P[(X,Y)\in\mathcal{Z})] > 0$, então o par (X,Y) condicionado a $\mathcal{Z}$, é um acoplamento ótimo de (μ',ν'), onde μ' é a distribuição de X condicionado com o evento $"(X,Y)\in\mathcal{Z}$ "e ν' é a distribuição de Y condicionado ao mesmo evento.*

3.2 Ciclo monótono e a dualidade de Kantorovich

Para continuar o estudo de acoplamentos vamos fazer um estudo sobre dois conceitos na teoria de transporte ótimo. O primeiro é uma propriedade geométrica chamada c-ciclo monótono e a segunda é o problema dual de Kantorovich. Esses tais conceitos tem uma grande importância neste trabalho, pois um dos principais resultados envolvendo os problemas de Monge e Kantorovich está baseado no fato de um determinado conjunto ser ou não c-ciclo monótono. Iniciaremos estudando estes conceitos através do seguinte exemplo.

Exemplo 3.2.1. *Considere $\mathcal{X}$ o conjunto das filiais de uma empresa que fabrica refrigerantes e $\mathcal{Y}$ uma grande franquia de lanchonetes, com filiais $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ e $\mathcal{Y} = \{y_1, y_2, \cdots, y_m\}$. Deve-se transportar produtos da empresa $\mathcal{X}$, na qual a proporção fabricada por cada filial é associada a uma medida de probabilidade, para $\mathcal{Y}$, cuja demanda de cada fábrica também está associada a uma medida de probabilidade. Cada filial x_i produz uma determinada quantidade de refrigerantes e cada filial y_j precisa de uma determinada quantidade do produto de $\mathcal{X}$.*

Considere uma função custo $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, onde $c(x_i,y_j)$ significa o custo do transporte por unidade de x_i para y_j. O objetivo do problema é transportar os produtos da empresa $\mathcal{X}$ para $\mathcal{Y}$ com o menor custo. Escolha um plano de transporte γ, sendo $\gamma(x_i,y_j)$ o quanto de produto devemos levar de x_i para y_i.

Como existem reclamações de que o custo do transporte inicial é muito alto, então a tentativa é escolher outro plano com o intuito de reduzir o custo. Para tal propósito escolha uma empresa x_1 e envie a unidade da produção que era destinada a y_1 para y_2,

assim o lucro é de $c(x_1, y_1) - c(x_1, y_2)$. É claro que resulta excesso de produto de y_2, então repassa o restante para y_3 e assim sucessivamente. Com esta alteração na distribuição temos um novo plano de transporte, que é melhor do que o antigo se, e somente se,

$$\sum_{i=1}^{n} c(x_i, y_{i+1}) \leq \sum_{i=1}^{n} c(x_i, y_i)$$

onde convencionamos que $y_{n+1} = y_1$, se $m = n$. Então se encontrarmos ciclos (x_i, y_i) em seu plano de transferência que cumpre a desigualdade estrita acima, certamente este plano não é ótimo. Reciprocamente, se não encontrar um outro plano que cumpre a desigualdade estrita acima então seu plano não pode ser melhorado, em outras palavras, o plano inicial se torna ótimo. Esse fato motiva a seguinte definição.

Definição 3.2.2. *Sejam $\mathcal{X}, \mathcal{Y}$ conjuntos arbitrários e uma função custo $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \bigcup \{\infty\}$. O subconjunto $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ é dita c-ciclo monótono se, para todo $n \in \mathbb{N}$, toda permutação σ de $\{1, 2, \cdots, n\}$ e toda família $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ pontos de Γ, temos*

$$\sum_{i=1}^{n} c(x_i, y_i) \leq \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}).$$

O plano de transferência é dito c-ciclo monótono se a medida é concentrada em um conjunto c-ciclo monótono.

Apoiado na intuição da definição acima mostraremos que um plano é ótimo, se o suporte do plano é um c-ciclo monótono e vale ressaltar que a reciproca vale sob certas condições. Antes de justificar, vamos apresentar um outro conceito importante é o problema dual de Kantorovich. Enquanto o problema central do original é minimizar o custo, no dual é maximizar os lucros.

Imagine que uma transportadora ofereça um serviço para cuidar do seu problema de transporte, comprando o produto da empresa $\mathcal{X}$ e depois vendendo o mesmo para $\mathcal{Y}$. O que acontece no meio do caminho e o modo da distribuição não é mais relevante no problema. Seja $\psi(x_i)$ o preço da unidade do refrigerante na filial x_i e $\phi(y_j)$ o preço de venda da unidade do refrigerante para a padaria y_j.

Aceitando o serviço da transportadora, o preço pago pelo transporte é $\phi(y) - \psi(x)$, ao invés do custo original de $c(x, y)$. Claro que para cada unidade de refrigerante, se a quantidade tiver uma proporção $\mu(dx)$ de x, então o preço total do produto pode ser dado por $\psi(x)\mu(dx)$. De modo a ser competitiva, a transportadora precisa configurar os preços de tal forma que

$$\phi(y) - \psi(x) \leq c(x, y), \text{ Para todo} (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{3.4}$$

No problema anterior, queríamos minimizar o custo, agora o objetivo do problema é maximizar o lucro da transportadora e isto nos leva naturalmente ao problema dual de Kantorovich.

Problema 3.2.3. *(Problema dual de Kantorovich) Sejam $\mu \in \mathcal{P}(\mathcal{X})$ e $\nu \in \mathcal{P}(\mathcal{Y})$ duas distribuições de probabilidades. O problema dual de Kantorovich consiste em maximizar*

$$\sup_{\psi,\phi}\{\int_{\mathcal{Y}} \phi(y)\ d\nu(y) - \int_{\mathcal{X}} \psi(x)\ d\mu(x); \phi(y) - \psi(x) \leq c(x,y)\} \tag{3.5}$$

onde $\psi : \mathcal{X} \to \mathbb{R}$ e $\phi : \mathcal{Y} \to \mathbb{R}$ são duas funções reais com $\psi \in L^1(\mu)$ e $\phi \in L^1(\nu)$.

Com a intervenção da transportadora, o embarque de cada unidade de refrigerante não custa mais do que custava quando se lidava com o transporte; é natural que o supremo de (3.5) não é maior do que o custo do transporte ótimo, note que

$$\sup_{\psi,\phi}\{\int_{\mathcal{Y}} \phi(y)\, d\nu(y) - \int_{\mathcal{X}} \psi(x)\, d\mu(x); \phi(y) - \psi(x) \leq c(x,y)\} \leq \inf_{\pi \in ADM(\mu,\nu)}\{\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi(x,y)\}.$$

O par de funções (ψ, ϕ) é dita competitiva se cumpre (3.4).

Dado uma filial y, é claro que a transportadora deseja aumentar o máximo possível o preço de $\phi(y)$, ou seja, que o preço é igual ao ínfimo de $\psi(x) + c(x,y)$, dentre todas as filiais x em $\mathcal{X}$. Analogamente, dado x, o preço $\psi(x)$ deve ser supremo de $\phi(y) - c(x,y)$ dentre todas as filiais y em $\mathcal{Y}$. Definimos que o par de preço é tight se

$$\begin{cases} \phi(y) = \inf_{x\in\mathcal{X}}(\psi(x) + c(x,y) \\ \psi(x) = \sup_{y\in\mathcal{Y}}(\phi(y) - c(x,y). \end{cases} \tag{3.6}$$

Intuitivamente, um par de preços é tight se é impossível aumentar o preço de venda e baixar o preço de compra, sem perder a competitividade.

Considere o par de preços competitivos (ϕ, ψ) qualquer. Podemos sempre melhorar o par substituindo ϕ por $\phi_1(y) = \inf_{x\in\mathcal{X}}\{\psi(x) + c(x,y)\}$. Portanto podemos melhorar ainda mais substituindo ψ por $\psi_1(x) = \sup_{y\in\mathcal{Y}}\{\phi_1(y) + c(x,y)\}$, então o par (ϕ_1, ψ_1) é ainda competitivo e é melhor ou igual ao par anterior. Naturalmente a partir de (ϕ_1, ψ_1) podemos repetir o processo gerando o par (ϕ_2, ψ_2) e consequentemente através de n iterações podemos gerar o par $(\phi_n.\psi_n)$ melhor do que todos os pares anteriores, mas o que ocorre é que este processo além de ser estacionário, basta realizar uma iterada, ou seja, $\phi_1 = \phi_2$ e $\psi_1 = \psi_2$.

De fato, vamos mostrar que $\phi_1 = \phi_2$ e o outro caso será análogo. Sabemos que para todo $y \in \mathcal{Y}$ temos $\phi_1(y) \leq \phi_2(y)$, resta mostrar que vale a outra desigualdade, mas note que usando as definições de cada função e o fato de que o para (ϕ, ψ) ser competitivo temos

$$\phi_2(y) \leq \psi_1(x) + c(x,y) = \sup_{z\in\mathcal{Y}}\{\phi(z) - c(x,z)\} + c(x,y) \leq \psi(x) + c(x,y).$$

No que acarreta $\phi_2(y) \leq \phi_1(y)$ e consequentemente a igualdade.

A partir desse fato pode-se concluir, que usando apenas uma iteração no par obtemos o preço tight. Então quando consideramos o problema dual de Kantorovich, faz sentido concentrar nossa atenção no par de preços dados por (3.6). Dada uma função ψ podemos reconstruir ϕ em termos de ψ, assim podemos considerar que a única informação desconhecida do problema, digamos que seja, a função ψ.

Mas esta função desconhecida não pode ser qualquer função, pois é necessário que a ψ satisfaça (3.6), o fato que garantir que ψ cumpra tal condição é se, e somente se, ψ for uma função c-convexa, que motiva a próxima definição.

Definição 3.2.4. *(c-convexidade) Sejam $\mathcal{X}$ e $\mathcal{Y}$ conjuntos quaisquer e $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}$ uma função custo. Uma função $\psi : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ é dita c-convexa se não é constante ∞, e se existe uma função $\xi : \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ tal que*

$$\forall x \in \mathcal{X} \quad \psi(x) = \sup_{y \in \mathcal{Y}} (\xi(y) - c(x, y)). \tag{3.7}$$

Definimos sua c-transformada sendo uma função ψ_c onde

$$\forall y \in \mathcal{Y} \quad \psi_c(y) = \inf_{x \in \mathcal{X}} (\psi(x) + c(x, y)).$$

A definição da subdiferencial de uma c-convexa ψ é definida por

$$\partial_c \psi = \{(x, y) \in \mathcal{X} \times \mathcal{Y}; \psi_c(y) - \psi(x) = c(x, y)\}.$$

Ou equivalentemente, $(x, y) \in \partial_c \psi$ se

$$\forall z \in \mathcal{X} \quad \psi(x) + c(x, y) \leq \psi(z) + c(z, y).$$

Mais ainda,

$$\partial_c \psi(x) = \{y; (x, y) \in \partial_c \psi\}.$$

Note que pela definição apresentada acima, $\partial_c \psi$ é um conjunto c-ciclo monótono

Exemplo 3.2.5. *Seja $(\mathcal{X}, d)$ espaço métrico e considere $c(x, y) = d(x, y)$. Então uma função ψ é c-convexa se, e somente se, ψ é uma função lipschitziana com a constante igual a 1. Mais ainda, a sua c-transformada é a própria função. De fato, suponha que ψ seja uma função c-convexa, então existe uma função ξ que cumpre a condição (3.7), assim usando a desigualdade triangular, obtemos para quaisquer $x, y \in \mathcal{X}$.*

$$\psi(y) = \sup_{z \in \mathcal{X}} \xi(z) - d(z, y) \geq \sup_{z \in \mathcal{X}} \xi(z) - d(z, x) - d(x, y) = \psi(x) - d(x, y).$$

No que implica que $d(x, y) \geq \psi(x) - \psi(y)$, e portanto ψ é Lipschitziana. Reciprocamente, considere ψ Lipschitziana, então observe que $\psi(x) \leq \psi(y) + d(x, y)$ vale para

quaisquer $x, y \in \mathcal{X}$, em particular, fixando y obtemos que $\psi(x) - d(x,y) \leq \psi(y)$ e consequentemente $\psi(y) \geq \sup_{x\in\mathcal{X}} \psi(x) - d(x,y)$ e para $x = y$, obtemos que $\psi(y) = \sup_{x\in\mathcal{X}} \psi(x) - d(x,y)$ e portanto ψ é uma função c-convexa com sua c-transformada sendo ela mesma.

Exemplo 3.2.6. *$c(x,y) = - < x, y >$ em $\mathbb{R}^n \times \mathbb{R}^n$, a importância deste exemplo se deve ao fato de que uma função $\psi : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ é c-convexa, se e somente se, ψ é uma função semicontínua inferior e convexa, para maiores detalhes veja a referência [1], na seção 1.2.*

Dada uma função ψ definida em $\mathcal{X}$, então sua c-transformada é uma função definida em $\mathcal{Y}$ e reciprocamente, dada uma função em $\mathcal{Y}$, pode-se definir a sua c-transformada em $\mathcal{X}$. Essa simetria de funções entre $\mathcal{X}$ e $\mathcal{Y}$ e inspirado na equação (3.4), podemos definir a noção de função c-côncava.

Definição 3.2.7. *(c-concavidade) Com a mesma notação da definição anterior, uma função $\phi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ é dita c-côncava se não é identicamente $-\infty$ e se existe uma função $\psi : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$ tal que $\phi = \psi_c$.*

Definimos sua c-transformada, uma função ϕ^c definida por

$$\forall x \in \mathcal{X} \quad \phi^c(x) = sup_{y\in\mathcal{Y}}\{\phi(y) - c(x,y)\}.$$

E sua c-superdiferencial é um conjunto c-ciclo monótono definida por

$$\partial^c\phi = \{(x,y); \phi(y) - \phi^c(x) = c(x,y)\}.$$

Ou equivalentemente,$(x,y) \in \partial^c\phi$, então para todo $z \in \mathcal{X}$ temos

$$\phi(x) - c(x,y) \geq \phi(z) - c(z,y).$$

Uma observação importante sobre as definições acima, no caso $\mathcal{X} = \mathcal{Y}$, que pode ser justificado diretamente da definição, se uma função ψ é c-convexa, então $-\psi$ é c-côncava, $\partial_c\psi = \partial^c(-\psi)$.

Teorema 3.2.8. *(Fundamental do Transporte Ótimo) Seja $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ uma função contínua e limitada inferiormente e sejam $\mu \in \mathcal{P}(\mathcal{X})$ e $\nu \in \mathcal{P}(\mathcal{Y})$ tal que para algum $a \in L^1(\mu)$ e $b \in L^1(\nu)$ cumpre*

$$c(x,y) \leq a(x) + b(y). \tag{3.8}$$

Para $\gamma \in ADM(\mu,\nu)$ as seguintes condições são equivalentes

i) γ é um acoplamento ótimo.

ii) $supp(\gamma)$ é um conjunto ciclo monótono, onde

$$supp(\gamma) = \overline{\{(x,y); \gamma(V) \neq 0) \ \ \forall\ V, \ \ (x,y) \in V\}, V \ \acute{e}\ \gamma\ mensurável\}}.$$

iii) Existe uma função c-côncava ϕ tal que $max\{\phi, 0\} \in L^1(\nu)$ e $supp(\gamma) \subset \partial^c\phi$.

Demonstração. Primeiramente Note que por (3.8) temos que a função $max\{c, 0\}$ é integrável e $c \in L^1(\gamma')$ para todo acoplamento $\gamma' \in ADM(\mu, \nu)$. De fato,

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\ d\gamma' \leq \int_{\mathcal{X}} a(x)\ d\mu(x) + \int_{\mathcal{Y}} b(y)\ d\nu(y) < \infty.$$

$i) \Rightarrow ii)$

- Vamos supor por contradição, então existe um número natural N e pares $(x_1, y_1), \cdots, (x_N, y_N) \in supp(\gamma)$ e alguma permutação σ de $\{1, 2, \cdots, N\}$;

$$\sum_{j=1}^{N} c(x_j, y_j) > \sum_{j=1}^{N} c(x_j, y_{\sigma(j)}).$$

- Pela continuidade da função c existem abertos $U_i \ni x_i$ e $V_i \in y_i$ para cada $i \in \{1, 2, \cdots, N\}$ e ainda que para todo i temos que $U_i \times V_i \subset supp(\gamma)$ tal que

$$\sum_{j=1}^{N} c(u_j, v_j) > \sum_{j=1}^{N} c(u_j, v_{\sigma(j)}) \quad \forall (u_i, v_i) \subset U_i \times V_i. \tag{3.9}$$

- A ideia é construir uma "variação"$\gamma' = \eta + \gamma$ com o intuito de contrariar a minimidade de γ, para isto a medida η deve satisfazer

  A) $\eta^- \leq \gamma$, ou seja, a parte negativa da medida η é menor ou igual a γ.

  B) Nulo na primeira e segunda marginal, ou seja, $\eta(\mathcal{X}, B) = 0$ e $\eta(A, \mathcal{Y}) = 0$.

  C) $\int c\ d\eta$ é negativa.

- Considere $\Omega = \Pi_{j=1}^{N} U_j \times V_j$ e $P \in \mathcal{P}(\Omega)$ definido por

$$P(\cdot) = \Pi_{i=1}^{N} \frac{1}{m_i} \gamma(\cdot)|_{U_i \times V_j},$$

onde $m_i = \gamma(U_i \times V_i)$. Defina η por

$$\eta = \frac{min_i\ m_i}{N} \cdot \sum_{j=1}^{N} (\pi^{U_j}, \pi^{V_{\sigma(j)}})_{\#} P - (\pi^{U_j}, \pi^{V_j})_{\#} P.$$

Note que $\eta = \eta^+ - \eta^-$, cumpre as condições acima. De fato considere $A \times B \subset \Omega$ mensurável, então $A \cup B = \bigcup_{j=1}^{N} A_j \cup B_j$, onde $A_j \times B_j \subset U_j \times V_j$.

$(A)\eta^-(A\times B) = \frac{min_i\ m_i}{N}\cdot\sum_{j=1}^N(\pi^{U_j},\pi^{V_j})_\#(A\times B) = \frac{min_i\ m_i}{N}\cdot\sum_{j=1}^N(\pi^{U_j},\pi^{V_j})_\#(A_j\times B_j) = \frac{min_i\ m_i}{N}\cdot\sum_{j=1}^N P(U_1\times V_1,\cdots,A_j\times B_j,U_{j+1}\times V_{j+1},\cdots,U_N\times V_N) = \frac{min_i\ m_i}{N}\cdot\sum_{j=1}^N\frac{\gamma(A_j\times B_j)}{m_j}\leq\frac{\gamma(A\times B)}{N}\leq\gamma(A\times B).$

$(B)\eta(A\times\mathcal{Y}) = \frac{min_i\ m_i}{N}\cdot\sum_{j=1}^N(\pi^{U_j},\pi^{V_{\sigma(j)}})_\#P(A_j\times\mathcal{Y})-(\pi^{U_j},\pi^{V_j})_\#(A_j\times\mathcal{Y}) = 0$ pois note que não existe um conjunto em Ω tal que a imagem desse conjunto pelas projeções seja $A_j\times\mathcal{Y}$.

$(C)\ \int c(x,y)\ d\eta(x,y) = \int c(x,y)\ d\eta^+(x,y) - \int c(x,y)\ d\eta^-(x,y)\leq 0$, por (3.9). Portanto segue a prova.

$ii)\Rightarrow iii)$

- Vamos provar primeiramente que se um conjunto $\Gamma\subset\mathcal{X}\times\mathcal{Y}$ é ciclo monótono, então existe uma função c-côncava ϕ tal que $max\{\phi,0\}\in L^1(\nu)$ e $\Gamma\subset\partial^c\phi$.

- Fixemos $(\overline{x},\overline{y})\in\Gamma$ e observe que se existir uma função ϕ que cumpre as condições acima, então para toda escolha de $(x_i,y_i)\in\Gamma$ temos

$$\begin{cases}\phi(y)-\phi^c(x)\leq c(x,y) & \forall(x,y)\in\mathcal{X}\times\mathcal{Y}.\\ \phi(y)-\phi^c(x)=c(x,y) & se\ \ (x,y)\in\Gamma.\end{cases}\tag{3.10}$$

Usando várias vezes as expressões em (3.10), obtemos a possível candidata

$$\phi(y)\leq c(x_1,y)+\phi^c(x_1) = c(x_1,y)-c(x_1,y_1)+\phi(y_1)\leq$$
$$(c(x_1,y)-c(x_1,y_1))+c(x_2,y_1)+\phi^c(x_2)$$
$$=(c(x_1,y)-c(x_1,y_1))+(c(x_2,y_1)-c(x_2,y_2))$$
$$+\phi^c(x_2)\leq\cdots(c(x_1,y)-c(x_1,y_1))$$
$$+(c(x_2,y_1)-c(x_2,y_2))+\cdots+c(\overline{x},y_N)-c(\overline{x},\overline{y}))+\phi(\overline{y}).$$

- O que motiva a considerar a seguinte função candidata

$$\phi(y) = \inf_{N,(x_j,y_j)\in\Gamma}(c(x_1,y)-c(x_1,y_1))+(c(x_2,y_1)-c(x_2,y_2))+\cdots+c(\overline{x},y_N)-c(\overline{x},\overline{y})).\tag{3.11}$$

Por (3.11) sabemos que $\phi(\overline{y})\leq 0$, pois basta tomar $N=1$ e $(x_1,y_1)=(\overline{x},\overline{y})$ e ainda por Γ ser um conjunto ciclo monótono, obtemos que $\phi(\overline{y})\geq 0$, no que acarreta $\phi(\overline{y})=0$.

- ϕ é uma função c-concava, pois definido $\psi(x) = \inf_{N,(x_j,y_j)\in\Gamma} -c(x,y_1) + (c(x_2,y_1) - c(x_2,y_2)) + \cdots + (c(\overline{x},y_n) - c(\overline{x},\overline{y}))$, assim

$$\psi^c(y) = \inf_{x\in\mathcal{X}} c(x,y)+\psi(x) = \inf_{N,x\in\mathcal{X},(x_j,y_j)\in\Gamma}\{c(x,y)-c(x,y_1)+\cdots+(c(\overline{x},y_n)-c(\overline{x},\overline{y})) \leq$$

$$\leq \inf_{N,(x,y)\in\Gamma,(x_j,y_j)\in\Gamma}\{c(x,y) - c(x,y_1) + \cdots + (c(\overline{x},y_n) - c(\overline{x},\overline{y})) = \phi(y).$$

E ainda, para $x_1 = x$ temos que

$$\phi(y) \leq \inf_{N,(x_j,y_j)\in\Gamma}(c(x,y)-c(x,y_1))+(c(x_2,y_1)-c(x_2,y_2))+\cdots+c(\overline{x},y_N)-c(\overline{x},\overline{y})) =$$

$$= c(x,y)+\inf_{N,(x_j,y_j)\in\Gamma}(-c(x,y_1))+(c(x_2,y_1)-c(x_2,y_2))+\cdots+c(\overline{x},y_N)-c(\overline{x},\overline{y})) =$$

$$= c(x,y) + \psi(x) \leq \psi^c(y).$$

Daí obtemos que a função ϕ é c-côncava, pois $\phi = \psi^c$.

- $max\{\phi, 0\} \in L^1(\nu)$, de fato

$$\phi(y) \leq c(\overline{x},y) - c(\overline{x},\overline{y}) \leq a(\overline{x}) + b(\overline{y}) - c(\overline{x},\overline{y}) < \infty.$$

Ou seja, $max\{\phi, 0\} \in L^1(\nu)$.

- $\Gamma \subset \partial^c\phi$.

Seja $(x', y') \in \Gamma$, então para todo $y \in \mathcal{Y}$.

$$\phi(y) \leq c(x',y)-c(x',y')+\inf_{N,(x_j,y_j)\in\Gamma}(c(x_1,y')-c(x_1,y_1))+\cdots+(c(x',y_N)-c(x',y)).$$

O que implica que

$$\phi(y) \leq c(x',y) - c(x',y') + \phi(y').$$

O que nos leva a concluir que $(x', y') \in \partial^c\phi$, e portanto $\Gamma \subset \partial^c\phi$.

- Como temos por hipótese que $supp(\gamma)$ é ciclo monótono então segue o resultado.

iii) $\Rightarrow$ *i)* Seja $\gamma' \in ADM(\mu,\nu)$ um plano de transporte qualquer. Precisamos mostrar que $\int c\ d\gamma \leq \int c\ d\gamma'$, note que por (3.10) temos

$$\int c(x,y)\ d\gamma(x,y) = \int \phi(y)-\phi^c(x)\ d\gamma(x,y) = \int \phi(y)\ d\nu(y)-\int \phi^c(x)\ d\mu(x) =$$

$$= \int \phi(y) - \phi^c(x)\ d\gamma'(x,y) \leq \int c(x,y)\ d\gamma'(x,y).$$

Portanto γ é um acoplamento ótimo.

□

Teorema 3.2.9. *(Teorema da dualidade) Sejam $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$ e $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ contínua e limitada inferiormente. Suponha que $c(x, y) \leq a(x) + b(y)$, nas mesmas condições do teorema anterior. Então o mínimo do problema de Kantorovich é igual ao supremo do problema dual. Mais ainda, se (ϕ, ψ) é o par de preços ótimo, então ϕ é uma função c-côncava e $\psi = \phi^c$.*

Demonstração. Note que para todo par (ϕ, ψ) temos

$$\int c(x,y)\ d\gamma(x,y) \geq \int \phi(y) - \psi(x)\ d\gamma(x,y) = \int \phi(y)\ d\nu(y) - \int \psi(x)\ d\mu(x).$$

Seja γ' um acoplamento ótimo, então pelo teorema anterior, existe ϕ c-concavo tal que $supp(\gamma) \subset \partial^c\phi$, $\max\{0, \phi\} \in L^1(\nu)$ e $\max\{\phi^c\} \in L^1(\mu)$. Então por $iii) \Rightarrow i)$, temos

$$\int c(x,y)\ d\gamma'(x,y) = \int \phi(y) - \phi^c(y)\ d\gamma'(x,y) = \int \phi(y)\ d\nu(y) - \int \phi^c(x)\ d\mu(x).$$

O que acarreta $\phi \in L^1(\nu)$ e $\phi^c \in L^1(\mu)$ e ainda conclui-se que (ϕ^c, ϕ) é um acoplamento admissível o que mostra a igualdade das soluções do problema original com o dual de Kantorovich.

□

3.3 Uma sugestão de algoritmo

Considere o problema de Kantorovich com $\mathcal{X} = \{x_1, \cdots, x_n\}$ e $\mathcal{Y} = \{y_1, \cdots, y_m\}$ e medidas $\mu = (a_1, \cdots, a_n)$ e $\nu = (b_1, \cdots, b_m)$. Nesta seção iremos propor um algoritmo para obter o transporte ótimo, a principal referência desta seção é [8, Capítulo 14]. Considere duas matrizes, com n linhas e m colunas, sendo a primeira a tabela de custo, cuja a entrada ij é igual a $c(x_i, y_j)$ e a segunda matriz é o plano de transporte onde a entrada i, j é $\gamma(x_i, y_j)$.

$$\begin{pmatrix} c(x_1,y_1) & c(x_1,y_2) & \cdots & \cdots & c(x_1,y_m) \\ c(x_2,y_1) & c(x_2,y_2) & \cdots & \cdots & c(x_2,y_m) \\ \vdots & \vdots & \vdots & \vdots & c(x_1,y_m) \\ c(x_n,y_1) & c(x_n,y_2) & \cdots & \cdots & c(x_n,y_m) \end{pmatrix}$$

$$\begin{pmatrix} \gamma(x_1,y_1) & \gamma(x_1,y_2) & \cdots & \cdots & \gamma(x_1,y_m) \\ \gamma(x_2,y_1) & \gamma(x_2,y_2) & \cdots & \cdots & \gamma(x_2,y_m) \\ \vdots & \vdots & \vdots & \vdots & c(x_1,y_m) \\ \gamma(x_n,y_1) & \gamma(x_n,y_2) & \cdots & \cdots & \gamma(x_n,y_m) \end{pmatrix}$$

Por questão de simplificação chamaremos $\gamma_{ij} = \gamma(x_i, y_j)$. Observe que uma condição para $[\gamma_{ij}]_{n\times m}$ ser um acoplamento de $\mu = (a_1, \cdots, a_n)$ e $\nu = (b_1, \cdots, b_m)$ é

$$\sum_{i=1}^{n} \gamma_{ij} = a_j \tag{3.12}$$

$$\sum_{j=1}^{m} \gamma_{ij} = b_i \tag{3.13}$$

O algoritmo consiste basicamente em aplicar a regra do custo mínimo, ou seja, atribuir maior valor possível para γ_{ij} na entrada correspondente ao menor custo. Os passos do algoritmo são

a) Escolha a menor entrada da tabela de custo, digamos kl. Se existir mais de uma entrada, escolha uma delas aleatoriamente.

b) Atribua $\gamma_{kl} = \min\{a_k, b_l\}$. Cada γ_{kl} definido desse modo chamaremos de variável básica.

c) Temos os possíveis casos

- Se $a_k < b_l$, tome $\gamma_{kj} = 0$ para todo $j \neq l$ e observe que nessas condições a k-ésima coluna cumpre a equação (3.12). "Reduza o valor de b_l"para $(b_l - a_k)$.
- Se $a_k > b_l$, tome $\gamma_{il} = 0$ para todo $i \neq k$ e observe que nessas condições a l-ésima linha cumpre a equação (3.13). "Reduza o valor de a_k"para $(a_k - b_l)$.
- Se $a_k = b_l$, preencha com zeros as outras entradas da k-ésima linha ou a l-ésima coluna, mas não ambas. Entretanto, se a matriz possui algumas colunas e apenas uma linha para preencher, então escolha a k-ésima linha e se a matriz possui algumas linhas e uma coluna para preencher, então escolha a l-ésima coluna.

d) Repita os passos anteriores até preencher todas as entradas da matriz.

Vale ressaltar que o número total de variáveis básicas é $m + n - 1$, pois cada variável básica obtida "é deletado"uma linha ou uma coluna, exceto a última variável básica quando "ambos são deletados"e portanto conclui-se que são $m + n - 1$ variáveis básicas.

Esse algoritmo sugere que o acoplamento obtido pelo algoritmo acima seja ótimo, ou seja, esse acoplamento minimiza $\sum_{i=1}^{n}\sum_{j=1}^{m} c(x_i, y_j)\gamma(x_i, y_j)$. Apesar de não obtermos uma prova para esse fato de um modo geral, dado um exemplo conseguimos justificar essa afirmação usando o problema dual de Kantorovich.

Dado um acoplamento $[\gamma_{ij}]_{n\times m}$, tentaremos verificar se essa solução é ótima. A ideia é resolver o seguinte sistema

$$u_i + v_j = c(x_i, y_j) \;\text{ Para todo par }\; (i,j) \;\text{ com }\; \gamma_{ij} \;\text{ variável básica.} \tag{3.14}$$

Onde $u_i = -\psi(x_i)$ e $v_j = \phi(y_j)$, referente ao problema dual de Kantorovich. Note que são $m+n$ incógnitas e $m+n-1$ equações, pois o número de equações é o número de variáveis básicas. Portanto, uma das incógnitas pode ser fixada, digamos igual a zero, e usarmos o sistema (3.14) para determinar as outras. Uma vez que determinamos u_i e v_j devemos verificar se

$$u_i + v_j \leq c(x_i, y_j) \text{ Para todo } (i,j) \tag{3.15}$$

Se (3.15) é satisfeito, então podemos aplicar o teorema da dualidade de Kantorovich e isso sugere que o acoplamento $[\gamma_{ij}]$ seja ótimo, pois $\{u_i\}$ e $\{v_j\}$ formam uma solução ótima do problema dual, ou seja, usando o teorema da dualidade obtemos

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c(x_i, y_j)\gamma_{ij} = \sum_{i=1}^{m} u_i a_i + \sum_{j=1}^{n} v_j b_j.$$

Exemplo 3.3.1. *Considere* $c(x,y) = |x-y|$, $\mu = (\frac{1}{4}, 0, \frac{1}{4}, \frac{2}{4})$ *e* $\nu = (\frac{2}{4}, \frac{1}{4}, 0, \frac{1}{4})$. *Primeiramente vamos construir a tabela de custo*

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}$$

Começamos com $\gamma_{11} = \frac{1}{4}$, *assim*

$$\begin{pmatrix} \frac{1}{4} & 0 & 0 & 0 \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix}$$

Seguindo o roteiro apresentado acima chegamos ao acoplamento com as variáveis básicas em negrito

$$\begin{pmatrix} \mathbf{\frac{1}{4}} & 0 & 0 & 0 \\ 0 & \mathbf{0} & 0 & 0 \\ \mathbf{0} & \mathbf{\frac{1}{4}} & \mathbf{0} & 0 \\ \mathbf{\frac{1}{4}} & 0 & 0 & \mathbf{\frac{1}{4}} \end{pmatrix}.$$

Neste momento precisamos verificar se o acoplamento é ótimo para isso

$$u_i + v_j = c(x_i, y_j) \quad \textit{Para todo par} \quad (i,j) \quad \textit{com} \quad \gamma_{ij} \quad \textit{variável básica.}$$

Atribuindo $u_1 = 0$, *obteremos* $v_1 = 0$, $v_2 = -1, v_3 = -2, v_4 = -3, u_2 = 1, u_3 = 2$ *e* $u_4 = 3$. *Note que*

$$
\begin{aligned}
u_1 + v_1 &= 0 = c(x_1, y_1) \\
u_2 + v_1 &= 1 = c(x_2, y_1) \\
u_3 + v_1 &= 2 = c(x_3, y_1) \\
u_4 + v_1 &= 3 = c(x_4, y_1) \\
u_1 + v_2 &= -1 \leq c(x_1, y_2) \\
u_2 + v_2 &= 0 = c(x_2, y_2) \\
u_3 + v_2 &= 1 = c(x_3, y_2) \\
u_4 + v_2 &= 2 = c(x_4, y_2) \\
u_1 + v_3 &= -2 \leq c(x_1, y_3) \\
u_2 + v_3 &= -1 \leq c(x_2, y_3) \\
u_3 + v_3 &= 0 = c(x_3, y_3) \\
u_4 + v_3 &= 1 = c(x_4, y_3) \\
u_1 + v_4 &= -3 \leq c(x_1, y_4) \\
u_2 + v_4 &= -2 \leq c(x_2, y_4) \\
u_3 + v_4 &= -1 \leq c(x_3, y_4) \\
u_4 + v_4 &= 0 = c(x_4, y_4)
\end{aligned}
$$

E observe que

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c(x_i, y_j)\gamma_{ij} = 1 = \sum_{i=1}^{m} u_i a_i + \sum_{j=1}^{n} v_j b_j.$$

E portanto, o Teorema (3.2.9) *nos garante que o acoplamento obtido pelo algoritmo é ótimo e as funções* $\psi(x_i) = -u_i$ *e* $\phi(y_j) = v_j$ *resolvem o problema dual de Kantorovich.*

3.4 Existência de transporte ótimo para o Problema de Monge

O problema de Monge consiste em olhar a existência do transporte ótimo induzido por uma aplicação $T : \mathcal{X} \to \mathcal{Y}$, em outras palavras, planos γ que são da forma $(ID, T)_{\#}\mu$, onde $\mu = \pi^{\mathcal{X}}_{\#}\gamma$ e T alguma aplicação mensurável.

Foi discutido no inicio que, em geral o problema de Monge não possui solução. Daí podemos formular a seguinte questão: Fixadas duas medidas μ e ν e uma função custo c, é verdade que existe pelo menos um plano ótimo γ que é induzido por alguma aplicação?

A ideia inicial é usar o Teorema Fundamental do Transporte Ótimo, e analisar as propriedades de um conjunto c-ciclo monótono.

Lema 3.4.1. *Seja* $\gamma \in ADM(\mu, \nu)$*. Então* γ *é induzido por uma aplicação se, e somente se, existe um conjunto* $\Gamma \subset \mathcal{X} \times \mathcal{Y}$ γ*-mensurável onde* γ *é concentrado, tal que existe um único* $y = T(x) \in \mathcal{Y}$ μ*.q.s. tal que* $(x, y) \in \Gamma$*. Neste caso,* γ *é induzido pela aplicação* T*.*

Como foi visto no lema acima, um plano ótimo é concentrado em um conjunto c-ciclo monótono e que por sua vez um conjunto c-ciclo monótono está contido em uma

c-superdiferencial de uma função c-côncava φ, ou ainda, que o conjunto c-ciclo monótono está contido em uma c-subdiferencial de uma função c-convexa ψ. Portanto pelo lema, é necessário entender com que frequência uma c-subdiferencial da ψ é "valorizado em um ponto singular". Não existe uma resposta geral para esta questão, mas podemos estudar casos particulares. Vamos focar no seguinte caso, $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ e $c(x,y) = \dfrac{|x-y|^2}{2}$. Para isso, será preciso apresentar algumas definições e resultados que nos dará condições de responder esta pergunta.

Definição 3.4.2. *(c-c hipersuperfície) Um conjunto $E \subset \mathbb{R}^d$ é dita c-c hipersuperfície se, existem funções convexas $f, g : \mathbb{R}^{d-1} \to \mathbb{R}$ tal que*

$$E = \{(y,t) \in \mathbb{R}^d; y \in \mathbb{R}^{d-1}, t \in \mathbb{R}, t = f(y) - g(y)\}.$$

Exemplo 3.4.3. *O Gráfico de uma função convexa e o gráfico de uma função côncava são exemplos de c-c hipersuperfície. Em Particular, a esfera S^1 pode ser obtida por uma união de duas c-c hipersuperfície, a saber, $S^1 = E_1 \bigcup E_2$ onde E_1 é o gráfico da função $\sqrt{1-y^2}$ e E_2 é o gráfico da função $-\sqrt{1-y^2}$.*

Proposição 3.4.4. *Uma função $\varphi : \mathbb{R}^d \to \mathbb{R} \bigcup \{-\infty\}$ é c-côncavo se, e somente se, a função $\overline{\varphi}(x) = \frac{|x|^2}{2} - \varphi(x)$ é convexa e semicontínua inferior. Nesse caso $y \in \partial^c \varphi(x)$ se, somente se, $y \in \partial^- \overline{\varphi}(x)$.*

Demonstração. Suponha que φ seja uma função c-côncava, ou seja, existe uma função $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ tal que

$$\varphi(x) = \inf_{y \in \mathbb{R}^d} \frac{|x-y|^2}{2} - \psi(y).$$

Queremos mostrar que

$$\overline{\varphi}(x) = \sup_{y \in \mathbb{R}^d} < x, y > -(\frac{|y|^2}{2} - \psi(y))$$

lembrando que $c(x,y) = -\langle x, y \rangle$ e pelo exemplo (3.2.6) concluiremos que $\overline{\varphi}$ é convexo e semicontínua inferior. Observe que

$$\varphi(x) = \inf_{y \in \mathbb{R}^d} \frac{|x-y|^2}{2} - \psi(y) \Longleftrightarrow \varphi(x) = \inf_{y \in \mathbb{R}^d} \frac{|x|^2}{2} + < x, -y > + \frac{|y|^2}{2} - \psi(y) \Longleftrightarrow$$

$$\varphi(x) - \frac{|x|^2}{2} = \inf_{y \in \mathbb{R}^d} < x, -y > + \frac{|y|^2}{2} - \psi(y) \Longleftrightarrow \overline{\varphi}(x) = \sup_{y \in \mathbb{R}^d} < x, y > -(\frac{|y|^2}{2} - \psi(y)).$$

A conta acima nos mostra que φ é c-côncavo se, e somente se, a função $\overline{\varphi}(x) = \frac{|x|^2}{2} - \varphi(x)$ é convexo e semicontínua inferior. Vamos mostrar a segunda parte, que é $y \in \partial^c \varphi(x)$ se, somente se, $y \in \partial^- \overline{\varphi}(x)$. Por definição

$$y \in \partial^c\varphi(x) \Longleftrightarrow \begin{cases} \varphi(x) = \frac{|x-y|^2}{2} - \varphi^c(y). \\ \varphi(z) \leq \frac{|z-y|^2}{2} - \varphi^c(y) \quad \text{para todo} \quad z \in \mathbb{R}^d \end{cases} \Longleftrightarrow$$

e

$$y \in \partial^-(\varphi - \frac{|\cdot|^2}{2})(x) \Longleftrightarrow \overline{\varphi}(z) \geq \overline{\varphi}(x) + < z-x, y > \quad \text{para todo} \quad z \in \mathbb{R}^d.$$

Note que

$$y \in \partial^c\varphi(x) \Longleftrightarrow \begin{cases} \varphi(x) = \frac{|x-y|^2}{2} - \varphi^c(y). \\ \varphi(z) \leq \frac{|z-y|^2}{2} - \varphi^c(y). \end{cases} \Longleftrightarrow$$

$$\begin{cases} \varphi(x) - \frac{|x|^2}{2} = < x, -y > + \frac{|y|^2}{2} - \varphi^c(y). \\ \varphi(z) - \frac{|z|^2}{2} \leq < z, -y > + \frac{|y|^2}{2} - \varphi^c(y). \end{cases} \Longleftrightarrow$$

$$\varphi(z) - \frac{|z|^2}{2} \leq \varphi(x) - \frac{|x|^2}{2} + < z-x, -y > \Longleftrightarrow \overline{\varphi}(z) \geq \overline{\varphi}(x) + < z-x, y > \quad \text{para todo} \quad z \in \mathbb{R}^d \Longleftrightarrow$$

$$\Longleftrightarrow y \in \partial^-(\varphi - \frac{|\cdot|^2}{2})(x).$$

□

O teorema a seguir nos mostra a relação que existe entre funções convexas com c-c hipersuperfície, mais precisamente, fala da relação que existe entre o conjunto dos pontos onde a função convexa não é diferenciável com c-c hipersuperfície. A demonstração do teorema será omitida, mas para maiores informações veja a referência [1], no teorema 1.24.

Teorema 3.4.5. *(Estrutura do conjunto dos pontos não diferenciáveis de uma função convexa) Seja $A \subset \mathbb{R}^d$. Então existe uma função convexa $\overline{\varphi} : \mathbb{R}^d \to \mathbb{R}$ tal que A está contido em um conjunto dos pontos onde $\overline{\varphi}$ é não diferenciável se, e somente se, A possui uma cobertura enumerável de c-c hipersuperfície.*

Definição 3.4.6. *(Medida regular em $\mathbb{R}^d$) Uma medida $\mu \in \mathcal{P}(\mathbb{R}^d)$ é dita regular se $\mu(E) = 0$ para todo c-c hipersuperfície $E \subset \mathbb{R}^d$.*

Teorema 3.4.7. *(Brenier) Seja $\mu \in \mathcal{P}(\mathbb{R}^d)$ tal que $\int |x|^2 \, d\mu(x)$ é finito. Então são equivalentes.*

i) Para todo $\nu \in \mathcal{P}(\mathbb{R}^d)$ com $\int |x|^2 \, d\nu(x) < \infty$ existe um único plano de transporte de μ para ν e esse plano é induzido por uma aplicação T.

ii) μ é regular.

Se (i) ou (ii) são satisfeitos, a aplicação ótima T pode ser obtido tomando o gradiente de uma função convexa.

Demonstração. $(ii) \Rightarrow (i)$ Vamos seguir o seguinte roteiro.

- Considere $a(x) = b(x) = |x|^2$. Então sob as nossas hipóteses, podemos garantir que

$$c(x, y) \leq a(x) + b(y).$$

  e $a \in L^1(\mu)$ e $b \in L^1(\nu)$.

- Como a função custo é contínua, então existe um plano ótimo, digamos γ, e usando o teorema fundamental, existe uma função φ é c-convexa, onde $supp(\gamma) \subset \partial^c \varphi$.

- Pela proposição anterior temos que $\overline{\varphi} : \frac{|\cdot|^2}{2} - \varphi$ é convexo e $\partial^c \varphi = \partial^- \overline{\varphi}$.

- Usando os dois passos anteriores, obtemos $supp(\gamma) \subset \partial^- \overline{\varphi}$.

- Como μ é regular, então o conjunto E dos pontos não diferenciáveis de $\overline{\varphi}$ é μ-desprezível, fato justificado pelo teorema (3.4.5).

- Portanto $\bigtriangledown \overline{\varphi} : \mathbb{R}^d \to \mathbb{R}^d$ é bem definido a menos em um conjunto μ desprezível e assim, pelo teorema, $\partial^- \overline{\varphi}$ coincide com o gráfico do $\bigtriangledown \overline{\varphi}$ μ quase sempre.

- Desse modo $supp(\gamma)$ está contido no gráfico do gradiente de uma função convexa e pelo lema (3.4.1) podemos concluir que γ é induzido pela aplicação $\bigtriangledown \overline{\varphi}$.

$(i) \Rightarrow (ii)$

- Vamos supor por contradição que existe uma função convexa $\overline{\varphi} : \mathbb{R}^d \to \mathbb{R}$ de modo que o conjunto E dos pontos não diferenciáveis tenha uma medida μ positiva, em outras palavras, contrariando o teorema (3.4.5).

- Podemos modificar $\overline{\varphi}$ de modo que fora de um conjunto compacto fixado, podemos assumir que a função possui crescimento linear.

- Defina duas aplicações

$$T(x) := \text{ Elemento de menor norma em } \partial^- \overline{\varphi}(x).$$
$$S(x) := \text{ Elemento de maior norma em } \partial^- \overline{\varphi}(x).$$

  E um plano $\gamma = \frac{1}{2}[(Id, T)_{\#}\mu + (Id, S)_{\#}\mu$.

- Como $\overline{\varphi}$ tem crescimento linear, então $\nu = \pi^{\mathcal{Y}}_{\#}\gamma$ tem suporte compacto, pois fora do compacto fixado anteriormente a função é diferenciável e assim a subdiferencial de $\overline{\varphi}$ coincide com um gráfico do seu gradiente, veja o teorema (1.5.7), assim a subdiferencial de $\overline{\varphi}$ fora desse compacto tem uma medida γ desprezível.

- Note que $\gamma \in ADM(\mu, \nu)$ é c-ciclo monótono, pois pela proposição anterior combinado com o teorema fundamental do transporte ótimo, temos que $supp(\gamma)$ está contido no gráfico da subdiferencial de uma função convexa e semicontínua $\overline{\varphi}$.

- Portanto γ é uma aplicação ótima pelo teorema fundamental do transporte ótimo, mas γ não é uma aplicação induzida, pois $T \neq S$ em um conjunto de medida positiva em μ.

□

Vamos encerrar a seção mostrando uma aplicação em Geometria que é a demonstração da Desigualdade Isoperimétrica, via acoplamentos e o Teorema de Brenier. Este resultado mostra como a teoria pode ser utilizada em outras áreas da Matemática além de Probabilidade. Com estas considerações segue o Teorema da Desigualdade Isoperimétrica.

Teorema 3.4.8. *(Desigualdade Isoperimétrica) Considere B uma bola unitária no $\mathbb{R}^d$ e E um conjunto não vazio aberto qualquer também no $\mathbb{R}^d$ com perímetro $P(E)$. Então*

$$\mathcal{L}^d(E)^{1-\frac{1}{d}} \leq \frac{P(E)}{d\mathcal{L}^d(B)^{\frac{1}{d}}}.$$

Demonstração. Vamos provar via teorema de Brenier, desprezando todos os pontos onde a aplicação T não coincida com o gradiente de um função convexa. Seja

$$\mu = \frac{1}{\mathcal{L}^d(E)}\mathcal{L}^d \mid_E \quad \text{e} \quad \nu = \frac{1}{\mathcal{L}^d(B)}\mathcal{L}^d \mid_B .$$

Seja $T : E \to B$ uma aplicação ótima, com $c(x, y) = \frac{|x-y|^2}{2}$. Pela formula da mudança de variáveis obtemos

$$\frac{1}{\mathcal{L}^d(E)} = det \bigtriangledown T(x)\frac{1}{\mathcal{L}^d(B)} \quad \text{Para todo} \;\; x \in E. \tag{3.16}$$

De fato, observe que por $\nu = T_{\#}\mu$ temos

$$\int_B d\nu(y) = \int_E d\mu(x). \tag{3.17}$$

Mas pela fórmula de mudança de variáveis temos

$$\int_B d\mathcal{L}^d \mid_B (y) = \int_E det \bigtriangledown T(x) \; d\mathcal{L}^d \mid_E .$$

Desenvolvendo em (3.17) obtemos

$$\int_B \frac{1}{\mathcal{L}^d(B)}\, d\mathcal{L}^d \mid_B = \int_E \frac{1}{\mathcal{L}^d(E)}\, d\mathcal{L}^d \mid_E$$

Aplicando a formula no lado esquerdo temos

$$\int_E \frac{det \bigtriangledown T(x)}{\mathcal{L}^d(B)}\, d\mathcal{L}^d \mid_E = \int_E \frac{1}{\mathcal{L}^d(E)}\, d\mathcal{L}^d \mid_E$$

A equação acima nos leva a concluir que a expressão (3.16) vale para todo $x \in E$. Lembre-se que ainda pelo Teorema de Brenier, a aplicação T é igual ao gradiente de uma função convexa. Portanto $\bigtriangledown T(x)$ é uma matriz simétrica e é não negativa, pelo teorema (1.5.12), e assim todos os seus autovalores são não-negativos, para todo $x \in E$. Portanto o $det \bigtriangledown T(x)$ é o produto dos seus autovalores, pois $\bigtriangledown T(x)$ é diagonalizável, e o traço de $det \bigtriangledown T(x)$ é a soma dos seus autovalores. Usando a desigualdade entre as médias aritmética e geométrica temos

$$(det \bigtriangledown T(x))^{\frac{1}{d}} \leq \frac{\bigtriangledown \cdot T(x)}{d} \quad \text{Para todo} \;\; x \in E.$$

Juntando com a equação (3.16) obtemos

$$\frac{1}{\mathcal{L}^d(E)^{\frac{1}{d}}} \leq \frac{\bigtriangledown \cdot T(x)}{d} \frac{1}{\mathcal{L}^d(B)^{\frac{1}{d}}}.$$

Integrando sobre E e usando o Teorema do Divergente temos

$$\mathcal{L}^d(E)^{1-\frac{1}{d}} \leq \frac{1}{d\mathcal{L}^d(B)^{\frac{1}{d}}} \int_E \bigtriangledown \cdot T(x)\; dx = \frac{1}{d\mathcal{L}^d(B)^{\frac{1}{d}}} \int_{\partial E} \langle T(x), \eta(x) \rangle\; d\mathcal{L}^{d-1}(x).$$

Onde $\eta : \partial E \to \mathbb{R}^d$ é o vetor unitário normal exterior. Como $T(x) \in B$ para todo $x \in E$, temos $|T(x)| \leq 1$ para qualquer $x \in \partial E$ e $\langle T(x), \eta(x) \rangle \leq 1$ assim.

$$\mathcal{L}^d(E)^{1-\frac{1}{d}} \leq \frac{1}{d\mathcal{L}^d(B)^{\frac{1}{d}}} \int_{\partial E} d\mathcal{L}^{d-1}(x) = \frac{P(E)}{d\mathcal{L}^d(B)^{\frac{1}{d}}}.$$

Portanto segue o resultado.

□

3.5 Estabilidade do transporte ótimo

Assuma, como anteriormente, que estamos com o problema de transporte ótimo de Kantorovich entre produtores e consumidores, cujas as respectivas distribuições são modeladas por duas medidas de probabilidade.

Estamos interessados em construir e estudar o espaço formado pelas medidas de probabilidade em $\mathcal{X}$, cuja a distância entre duas medidas μ e ν é o ínfimo do problema de

Kantorovich para as duas medidas. Note que, o custo ótimo entre duas medidas é definido por:

$$C(\mu,\nu) = \inf_{\gamma \in ADM(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\ d\gamma(x,y). \tag{3.18}$$

Onde $c(x,y)$ é o custo da unidade de massa transportado de x para y. Neste contexto não estamos preocupados em determinar a medida γ que cumpre o ínfimo de (3.18), mas sim o custo ótimo entre μ e ν.

Um modo de pensar na equação (3.18) é como sendo uma distância entre μ e ν, mas em geral não é, devido ao fato do custo c não necessariamente uma métrica. Mas podemos definir o custo em termos da distância, o que motiva a definição a seguir.

Definição 3.5.1. *(Distância de Wasserstein) Sejam $(\mathcal{X}, d)$ um espaço Polonês e $p \in [1,\infty]$. Para quaisquer duas medidas de probabilidade μ e ν em $\mathcal{X}$, a distribuição de Wasserstein de ordem p entre μ e ν é definida pela fórmula*

$$W_p(\mu,\nu) = [\inf_{\gamma \in ADM(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} d(x,y)^p\ d\gamma(x,y)]^{\frac{1}{p}}.$$

ou equivalentemente,

$$W_p(\mu,\nu) = \inf_{X\sim\mu, Y\sim\nu} [E(d(X,Y)^p]^{\frac{1}{p}}.$$

Exemplo 3.5.2. *$W_p(\delta_x,\delta_y) = d(x,y)$. Neste exemplo, a distância não depende de p, mas isso não ocorre em geral.*

Proposição 3.5.3. *Nas condições acima W_p é uma métrica.*

Demonstração.

- $W_p(\mu,\nu) = W_p(\nu,\mu)$ pois considere duas variáveis aleatórias $X \sim \mu$ e $Y \sim \nu$ quaisquer , então $[E(d(X,Y)^p]^{\frac{1}{p}} = [E(d(Y,X)^p]^{\frac{1}{p}}$ no que acarreta que $W_p(\mu,\nu) = W_p(\nu,\mu)$.

- $W_p(\mu,\nu) = 0$ se , e somente se , $\mu = \nu$ decorre basicamente explorando propriedades de métrica e para justificar a reciproca basta usar o fato de que se $X \sim \mu$, então $X \sim \nu$ e explorar novamente a continuidade e propriedade da métrica.

- Resta mostrar que vale a desigualdade triangular, ou seja, dadas medidas μ_1, μ_2 e μ_3, queremos mostrar que

$$W_p(\mu_1,\mu_2) \leq W_p(\mu_1,\mu_3) + W_p(\mu_3,\mu_2).$$

  Sejam (X_1, X_2) acoplamento ótimo de (μ_1,μ_2) e (Z_2, Z_3) acoplamento ótimo de (μ_2,μ_3) , onde $c(x,y) = d(x,y)^p$. Considere γ um acoplamento de μ_1 e μ_2 e π um acoplamento de μ_2 e μ_3 e defina η uma medida de probabilidade em $\mathcal{X}^3$ como

$$\eta(x,y,z) = \frac{\gamma(x,y)\pi(y,z)}{\mu_2(y)}.$$

Sejam (X,Y,Z) um vetor aleatório com respeito a distribuição η e usando a propriedade da métrica d, temos que

$$d(X,Z) \leq d(X,Y) + d(Y,Z).$$

Então, tomando (X,Y) acoplamento ótimo de μ_1 e μ_2 e (Y,Z) acoplamento ótimo de μ_2 e μ_3, usando as propriedades da esperança e a desigualdade de Minkowski em L^p, ou seja $E[d(X,Z)^p]^{\frac{1}{p}} \leq E[d(X,Y)^p]^{\frac{1}{p}} + E[d(Y,Z)^p]^{\frac{1}{p}}$, conclui-se que

$$W_p(\mu_1,\mu_3) \leq E[d(X,Z)^p]^{\frac{1}{p}} \leq E[d(X,Y)^p]^{\frac{1}{p}} + E[d(Y,Z)^p]^{\frac{1}{p}} = W_p(\mu_1,\mu_2) + W_p(\mu_2,\mu_3).$$

□

Para completar a construção, ainda é necessário restringir a função W_p sob o subconjunto de $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ na qual assume valor finito, daí segue a definição do seguinte espaço.

Definição 3.5.4. *(Espaço de Wasserstein) Nas mesmas condições da definição anterior, o espaço de Wasserstein de ordem p é definido por*

$$\mathcal{P}_p(\mathcal{X}) = \{\mu \in \mathcal{P}(\mathcal{X}); \int_{\mathcal{X}} d(x_0,x)^p \ \mu(dx) < \infty\}.$$

Onde $x_0 \in \mathcal{X}$ é arbitrário.

Veremos que o espaço não dependerá da escolha de x_0.

Observe que W_p é finito em $\mathcal{P}_p$. De fato, seja π um acoplamento qualquer de μ e ν em $\mathcal{P}_p(\mathcal{X})$, então a desigualdade

$$d(x,y)^p \leq 2^{p-1}[d(x,x_0)^p + d(x_0,y)^p]$$

nos diz que $d(x,y)^p$ é π-integrável pois μ e $\nu \in \mathcal{P}_p(\mathcal{X})$. Agora estudaremos algumas propriedades topológicas do espaço $(\mathcal{P}_2(\mathcal{X}), W_2)$, iniciando com a noção de 2-uniforme integrável e logo após associar esta propriedade com o conceito de tight para extrair mais resultados e obter uma noção mais refinada de convergência no espaço de Wasserstein de ordem 2, portanto segue a nossa primeira definição.

Definição 3.5.5. *(2-uniforme integrável) Uma família $\mathcal{K} \subset \mathcal{P}_2(\mathcal{X})$ é dita ser 2-uniforme integrável se para todo $\varepsilon > 0$ e $x_0 \in \mathcal{X}$ existe $R_\varepsilon > 0$ de modo que*

$$\sup_{\mu \in \mathcal{K}} \int_{B(x_0,R_\varepsilon)^c} d^2(x,x_0) \ d\mu(x) \leq \varepsilon.$$

Exemplo 3.5.6. *Se $\mathcal{X}$ for um espaço Polonês infinito, ou seja, $*\mathcal{X} = \infty$ e $\mathcal{K} = \{\delta_x; x \in \mathcal{X}\}$, então $\mathcal{K}$ não é 2-uniforme integrável.*

Exemplo 3.5.7. *Se $\mathcal{K} = \{\mu\}$, então $\mathcal{K}$ é 2-uniforme integrável.*

Considere dois espaços métricos Poloneses $(\mathcal{X}, d_\mathcal{X})$ e $(\mathcal{Y}, d_\mathcal{Y})$ e adote em $\mathcal{X} \times \mathcal{Y}$ a distância $d^2((x_1, y_1), (x_2, y_2)) = d^2_\mathcal{X}(x_1, x_2) + d^2_\mathcal{Y}(y_1, y_2)$, então obtemos as seguintes desigualdades

$$\int_{(B(x_0,R)\times B(y_0,R))^c} d_\mathcal{X}(x, x_0)^2 \, d\gamma(x, y) = \int_{(B(x_0,R))^c\times\mathcal{Y}} d_\mathcal{X}(x, x_0)^2 \, d\gamma(x, y) + \int_{\mathcal{X}\times(B(y_0,R))^c} d_\mathcal{X}(x, x_0)^2 \, d\gamma(x, y) \leq$$

$$\leq \int_{(B(x_0,R)^c} d_\mathcal{X}(x, x_0)^2 \, d\mu(x) + \int_{\mathcal{X}\times(B(y_0,R))^c} R^2 \, d\gamma(x, y) \leq$$

$$\leq \int_{(B(x_0,R)^c} d_\mathcal{X}(x, x_0)^2 \, d\mu(x) + \int_{(B(y_0,R))^c} d_\mathcal{Y}(y, y_0)^2 \, d\nu(y).$$

Válido para quaisquer $\gamma \in ADM(\mu, \nu)$ e de modo análogo substituindo $d_\mathcal{Y}$ por $d_\mathcal{X}$, mostramos que se $\mathcal{K}_1 \subset \mathcal{P}_2(\mathcal{X})$ e $\mathcal{K}_2 \subset \mathcal{P}_2(\mathcal{Y})$ são 2-uniforme integrável, então o conjunto

$$\{\gamma \in \mathcal{P}_2(\mathcal{X} \times \mathcal{Y}); \pi^\mathcal{X}_\# \gamma \in \mathcal{K}_1, \pi^\mathcal{Y}_\# \gamma \in \mathcal{K}_2\}.$$

É 2-uniforme integrável.

Definição 3.5.8. *(Crescimento quadrático) Dizemos que uma função $f : \mathcal{X} \to \mathbb{R}$ tem crescimento quadrático se cumpre a seguinte condição*

$$|f(x)| \leq a \cdot (d^2(x, x_0) + 1).$$

Para algum $a \in \mathbb{R}$ e $x_0 \in \mathcal{X}$. Observe que se f possui crescimento quadrático e $\mu \in \mathcal{P}_2(\mathcal{X})$, então $f \in L^1(\mathcal{X}, \mu)$ pois

$$\int_\mathcal{X} f(x) \, d\mu(x) \leq a \int_\mathcal{X} d^2(x, x_0) \, d\mu(x) + a \int_\mathcal{X} 1 \, d\mu(x) < \infty.$$

O conceito de 2 - uniforme integráveis em relação a convergência de integrais de funções com crescimento quadrático, desempenha um papel similar ao do conceito de tight em relação a convergência de integrais das funções limitadas , o que mostra a proposição a seguir.

Proposição 3.5.9. *Seja $\{\mu_n\}_{n=1}^\infty \subset \mathcal{P}_2(\mathcal{X})$ uma sequência que converge fracamente para alguma medida μ. Então são equivalentes*

i) $\{\mu_n\}_{n=1}^\infty$ é 2-uniforme integrável.

ii) $\int f \, d\mu_n \to \int f \, d\mu$ para toda função f contínua com crescimento quadrático.

iii) $\int d^2(\cdot, x_0)\ d\mu_n \to \int d^2(\cdot, x_0)\ d\mu$ *para algum* $x_0 \in \mathcal{X}$.

Demonstração. i) $\Rightarrow$ *ii)* Sem perda de generalidade suponha que $f \geq 0$. Como f é contínua, então podemos considerar uma sequência $\{f_n\}$ de funções contínuas e limitadas, de modo que, $f_n \nearrow f$, explorando o lema de fatou, veja na referência [3], obtemos

$$\int f\ d\mu = \int \liminf_{k\to\infty} f_k\ d\mu \leq \liminf_{k\to\infty} \int f_k\ d\mu \leq$$

$$\leq \liminf_{k\to\infty} \int f_k\ d\mu_k \leq \liminf_{k\to\infty} \int f\ d\mu_k.$$

Ou seja,

$$\int f\ d\mu \leq \liminf_{k\to\infty} \int f\ d\mu_k.$$

Resta mostrar que $\limsup_{k\to\infty} \int f\ d\mu_k \leq \int f\ d\mu$, para isso fixe $x_0 \in \mathcal{X}$ e $\varepsilon > 0$ e encontre $R > 1$ tal que

$$\int_{B(x_0,R)^c} d^2(\cdot, x_0)\ d\mu_n \leq \varepsilon, \text{para todo } n \in \mathbb{N}.$$

Considere λ uma função contínua com suporte limitado , assumindo valores em $[0, 1]$, onde em $B(x_0, R)$ a função é identicamente igual a 1. Portanto para todo natural n obtemos

$$\int f\ d\mu_n = \int f \cdot \lambda\ d\mu_n + \int f \cdot (1-\lambda)\ d\mu_n \leq \int f \cdot \lambda\ d\mu_n + \int_{B(x_0,R)^c} f\ d\mu_n \leq$$

$$\leq \int f \cdot \lambda\ d\mu_n\ +\ a \cdot (\int_{B(x_0,R)^c} d^2(\cdot, x_0)\ d\mu_n + \int_{B(x_0,R)^c} 1\ d\mu_n) \leq$$

$$\int f \cdot \lambda\ d\mu_n\ +\ a \cdot (2\int_{B(x_0,R)^c} d^2(\cdot, x_0)\ d\mu_n) \leq \int f \cdot \lambda\ d\mu_n + 2a\varepsilon.$$

Como $f\lambda$ é contínua e limitada, temos que $\int f\lambda\ d\mu_n \to \int f\lambda\ d\mu$ mais ainda

$$\limsup_{n\in\mathbb{N}} \int f\ d\mu_n \leq \int f\lambda\ d\mu + 2\varepsilon a \leq \int f\ d\mu + 2\varepsilon a.$$

Como ε é arbitrário, concluímos que $\limsup \int f\ d\mu_n \leq \int f\ d\mu$ e portanto segue o resultado.

ii) $\Rightarrow$ *iii)* Imediato.

iii) $\Rightarrow$ *i)* Vamos supor por contradição e assumir que existe $\varepsilon > 0$ e $\overline{x}_0 \in \mathcal{X}$ tal que para todo $R > 0$ temos

$$\sup_{n\in\mathbb{N}} \int_{B(x_0,R)^c} d^2(\cdot,\overline{x}_0)\ d\mu_n > \varepsilon.$$

Para R suficiente grande, podemos ver que

$$\limsup_{n\in\mathbb{N}} \int_{B(x_0,R)^c} d^2(\cdot,\overline{x}_0)\ d\mu_n > \varepsilon.$$

Considere λ_R uma função contínua com valores em $[0,1]$ e suporte em $B(x_0,R)$ e identicamente igual a 1 em $B(x_0,\frac{R}{2})$. Explorando o fato de que $d^2(\cdot,x_0)\lambda_R$ é contínua e limitada temos que

$$\int d^2(\cdot,x_0)\lambda_R\, d\mu = \lim_{n\to\infty}\int d^2(\cdot,x_0)\lambda_R\, d\mu_n = \lim_{n\to\infty}(\int d^2(\cdot,x_0)\, d\mu_n - \int d^2(\cdot,x_0)(1-\lambda_R)\, d\mu_n) \leq$$

$$\leq \int d^2(\cdot,x_0)\, d\mu + \liminf_{n\to\infty} -\int_{B(x_0,R)^c} d^2(\cdot,x_0)\, d\mu_n = \int d^2(\cdot,x_o)\, d\mu - \limsup_{n\to\infty}\int_{B(x_0,R)^c} d^2(\cdot,x_0)\, d\mu_n \leq$$

$$\leq \int d^2(\cdot,x_0)\ d\mu - \varepsilon.$$

Assim concluímos que

$$\int d^2(\cdot,x_0)\ d\mu = \sup_R \int d^2(\cdot,x_0)\lambda_R\ d\mu \leq \int d^2(\cdot,x_0)\ d\mu - \varepsilon.$$

No que acarreta em uma contradição.

□

Proposição 3.5.10. *A distância W_2 é semicontínua inferior com respeito a convergência de medidas. Mais ainda, se $\{\gamma_n\} \subset \mathcal{P}_2(\mathcal{X}^2)$ é uma sequência de planos ótimos com a sequência convergindo para $\gamma \in \mathcal{P}_2(\mathcal{X}^2)$, então γ é plano ótimo também.*

Demonstração. Sejam $\{\mu_n\}$, $\{\nu_n\} \subset \mathcal{P}_2(\mathcal{X})$ com $\mu_n \to \mu$ e $\nu_n \to \nu$. Para cada natural n, considere γ_n acoplamento ótimo de μ_n e ν_n, pelo teorema de Prohorov temos que a sequência γ_n admite uma subsequência convergente para algum $\gamma \in \mathcal{P}(\mathcal{X}^2)$. É claro que $\pi^1_\#\gamma = \mu$ e $\pi^2_\#\gamma = \nu$, então

$$W_2^2(\mu,\nu) \leq \int d^2(x,y)\ d\gamma(x,y) \leq \liminf_{n\to\infty}\int d^2(x,y)\ d\gamma_n(x,y) = \liminf_{n\to\infty} W_2^2(\mu_n,\nu_n).$$

Considere $a(x) = b(x) = d^2(x,x_0)$ para algum x_0 e observe que $\mu,\nu \in \mathcal{P}_2(\mathcal{X})$. Pelo Teorema fundamental do transporte ótimo, fixe $N \in \mathbb{N}$ e selecione $(x^i,y^i) \in supp(\gamma)$, com $i = 1,2,\cdots,N$ e note que por $\gamma_n \to \gamma$ não é difícil de ver que $(x^i_n,y^i_n) \in supp(\gamma_n)$ tal que

$$\lim_{n\to\infty}(d(x^i_n,x^i) + d(y^i_n,y^i)) = 0.$$

Através da continuidade da função custo, conclui-se que o $supp(\gamma)$ é ciclo monótono.

□

Referências

[1] Ambrosio, L. ; Gigli, N. **A user's guide to optimal transport**

[2] Barry, J. **Probabilidade: Um Curso em Nível Intermediário**, 3ª edição, Rio de Janeiro: Sociedade Brasileira de Matemática, 2006

[3] Royden,H.L **Real Analysis**. Macmillan, N.York, 1968

[4] Billingsley, P. **Probability and Measure**, Wiley Series in Probability and Statistics, 2012

[5] Botelho,G; Pellegrino, D; Teixeira, E. **Fundamentos de Análise Funcional**. Rio de Janeiro, SBM, 2012

[6] Bubley, R ; Dyer, M.E. **Path Coupling: a Technique for Proving Rapid Mixing in Markov Chains**, Proceedings of the 38th Annual Symposium on Foundations of Computer Science, pp. 223–231, 1997

[7] Chung, K. L. **A Course in Probability Theory**, 3ª edição, San Diego: Academic Press, 2001

[8] Dantzig, G.B ; Ferguson, A.R. **Linear Programming and Extensions** , United States Air Force Project Rand, 1963

[9] Elon, Lages Lima **Análise Real, volume I**,11ª edição, Projeto Euclides, 2011

[10] Elon, Lages Lima **Curso de Análise, volume I**,12ª edição, Projeto Euclides, 2011

[11] Elon, Lages Lima **Curso de Análise, volume II**,12ª edição, Projeto Euclides,2011

[12] Elon, Lages Lima **Espaços métricos**, 5ª edição, Projeto Euclides, 2013

[13] Levin, David A.; Peres, Yurval; Wilmer, Elizabeth L. **Markov Chains and Mixing Times**, 1ª edição, Providence: American Mathematical Society, 2001

[14] Rockafellar, R. T **Convex Analysis**,Princeton University Press, Princeton, 1970

[15] Shiryaev, A.N. **Probability**, 2ª edition, Spriger, 1995

[16] Villani, C. **Optimal transport, old and new**, Springer Verlag, 2008