

# Escuela de Invierno ADEEM - PUCP

## Regularización entrópica, algoritmo SISTA, estimación de costos y matching

Marcelo Gallardo Burga

PUCP

Julio 2023

# Índice

- 1 Regularización entrópica
- 2 Sinkhorn-Knopp
- 3 Convergencia del algoritmo Sinkhorn-Knopp
- 4 Logit

# El problema de regularización entrópica

Siguiendo a [Nenna, 2020], vamos a considerar dos conjuntos finitos  $\mathcal{X}, \mathcal{Y}$ , donde  $|\mathcal{X}| = |\mathcal{Y}| = N$  y establecemos las medidas de probabilidad sobre  $\mathcal{X}$  y  $\mathcal{Y}$  respectivamente

$$\mu = \sum_{x \in \mathcal{X}} \mu_x \delta_x, \quad \nu = \sum_{y \in \mathcal{Y}} \nu_y \delta_y.$$

Luego, denotamos el conjunto de acoplamientos por

$$\Pi(\mu, \nu) = \left\{ \pi_{x,y} : \pi_{x,y} \geq 0, \sum_{y \in \mathcal{Y}} \pi_{xy} = \mu_x, \forall x \in \mathcal{X} \wedge \sum_{x \in \mathcal{X}} \pi_{xy} = \nu_y, \forall y \in \mathcal{Y} \right\}.$$

El problema de transporte *original* es [Villani, 2009], [Luigi Ambrosio and Semola, 2021]

$$\min_{\pi_{xy} \in \Pi(\mu, \nu)} \left\{ \sum_{x,y} \pi_{xy} c(x, y) \right\}. \quad (1)$$

## Definición

Definimos la función de entropía (relativa)  $H(\pi)$  de la siguiente manera

$$H(\pi) = - \sum_{x,y} h(\pi_{x,y})$$

con

$$h(r) = \begin{cases} r(\ln(r) - 1), & \text{si } r > 0 \\ 0, & \text{si } r = 0 \\ +\infty, & \text{si } r < 0. \end{cases}$$

Trabajamos en  $\Pi : \pi_{xy} \geq 0$ .

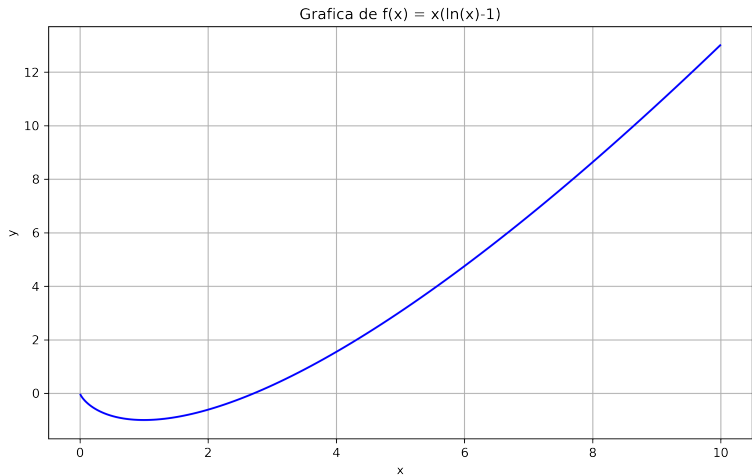


Figura  $h(x) = x(\ln x - 1)$

La entropía es un concepto vinculado al desorden, incertidumbre de la información. En concreto, dado un proceso con posibles resultados  $\{a_1, \dots, a_N\}$  y probabilidades  $\{p_1, \dots, p_N\}$ , la entropía es máxima cuando  $p_i = 1/N$ . En efecto, se resuelve

$$\begin{cases} \max & -\sum_{i=1}^N p_i (\ln p_i - 1) \\ \text{s.a.} & \sum_{i=1}^N p_i = 1. \end{cases}$$

Aplicando el método de los multiplicadores de Lagrange (dada la concavidad de la función objetivo y que solo hay una restricción lineal):

$$\mathcal{L}(p_1, \dots, p_N, \lambda) = -\sum_{i=1}^N p_i (\ln p_i - 1) + \lambda \left(1 - \sum_{i=1}^N p_i\right).$$

$$\frac{\partial \mathcal{L}}{\partial p_i} = -\ln p_i - \lambda = 0 \quad \forall i \implies p_i = p_j = \frac{1}{N}, \quad \forall i, j.$$

## Definición

El problema de regularización entrópica corresponde al siguiente problema de optimización

$$\mathcal{P}_\varepsilon : \inf \left\{ \sum_{x,y} \pi_{x,y} c(x,y) - \varepsilon H(\pi) \right\} \quad (2)$$

con  $\varepsilon > 0$  y sujeto a

$$\pi \in \Pi(\mu, \nu) = \left\{ \pi_{xy} \geq 0, \sum_{y \in \mathcal{Y}} \pi_{xy} = \mu_x, \forall x \in \mathcal{X} \wedge \sum_{x \in \mathcal{X}} \pi_{xy} = \nu_y, \forall y \in \mathcal{Y} \right\}.$$



## Teorema

El problema  $\mathcal{P}_\varepsilon$  tiene una única solución  $\pi^* \in \Pi(\mu, \nu)$  y, si  $\min\{\min_x \mu_x, \min_y \nu_y\} > 0$ ,  $\pi_{xy}^* > 0, \forall x \in \mathcal{X}, y \in \mathcal{Y}$ .

## Observación

La función  $-H : \pi \in (\mathbb{R}_{+*})^{N \times N} \rightarrow -\sum_{x,y} h(\pi_{xy})$  es estrictamente convexa.

A continuación, la prueba de que  $-H$  es convexa cuando  $\pi_{xy} > 0$ .

Prueba.

Como  $\pi_{xy} > 0$ ,  $h''(t) = \frac{1}{t} > 0$ . Luego,

$$H(-H(\pi)) = D \left( \frac{1}{\pi_{xy}} \right) = \begin{pmatrix} \frac{1}{\pi_{11}} & & \\ & \ddots & \\ & & \frac{1}{\pi_{NN}} \end{pmatrix}$$

la cual es definida positiva. □

## Prueba.

A continuación un sketch de la prueba:

➊ Definimos  $\gamma^\theta \in \Gamma(\mu, \nu) : \gamma^\theta = (1 - \theta)\pi^* + \theta(\mu \otimes \nu), (0, 1)$ .

➋ Por la convexidad de  $h(\cdot)$

$$h(\gamma_{xy}^\theta) \leq (1 - \theta)h(\pi_{xy}^*) + \theta h(\mu_x \nu_y) \leq h(\pi_{xy}^*) + \mathcal{O}(\theta). \quad (3)$$

➌ Definimos  $Z = \{(x, y) : \pi_{xy}^* = 0\} \neq \emptyset$  y  $C = \min\{\min_x \mu_x, \min_y \nu_y\} > 0$ .

➍ Para  $(x, y) \in Z$ :

$$h(\gamma_{xy}^\theta) = h(\theta \mu_x \nu_y) = \theta \mu_x \nu_y (\ln \theta + \ln \mu_x \nu_y - 1) \leq C \theta \ln \theta + \mathcal{O}(\theta). \quad (4)$$

➎ Sumando adecuadamente, se llega a

$$0 \leq \theta \left\{ \underbrace{\sum_{x,y} (c_{xy} \mu_x \nu_y + h(\mu_x \nu_y))}_{\text{cte}} + nC \underbrace{\ln \theta}_{< 0} \right\}, \quad \forall \theta \in (0, 1).$$

Esto fuerza a que  $n = 0$ . La unicidad de la solución es consecuencia directa de la convexidad estricta de  $-H(\cdot)$ .



El contexto es el siguiente

$$\left\{ \begin{array}{ll} \min & \sum_{x,y} \{ \pi_{x,y} c(x,y) - \varepsilon H(\pi_{x,y}) \} \\ \text{s.a.} & \sum_{x \in \mathcal{X}} \pi_{xy} = \nu_y \quad \forall y \\ & \sum_{y \in \mathcal{Y}} \pi_{xy} = \mu_x, \quad \forall x \\ & \pi_{xy} \geq 0. \end{array} \right. \quad (5)$$

- ❶ Condiciones de Slater
- ❷ Dualidad fuerte.
- ❸ Convexidad, continuidad.

**Condiciones de Slater.** En relación al siguiente problema de optimización:

$$\begin{cases} \min_x & f_0(x) \\ \text{s.a.} & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p. \end{cases}$$

si  $f_0$  es convexa, las restricciones son lineales y se cumple que  $\exists x : f_i(x) < 0, Ax = b$  (pues restricciones lineales) factible,

$$\max_{\lambda, \eta} \min_x \mathcal{L}(x, \lambda, \eta) = \min_{x \in S} f_0(x)$$

donde  $S$  es el conjunto determinado por las restricciones y

$$\mathcal{L}(x, \lambda, \eta) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \eta_i h_i(x).$$

Prueba.

Ver [Luneberger and Ye, 2021].



Con respecto a  $\mathcal{P}_\varepsilon$ , el Lagrangiano asociado es  $\mathcal{L}$  :

$$\begin{aligned}\mathcal{L}(\pi, \varphi, \psi) = & \sum_{x,y} \pi_{x,y} (c(x,y) + \varepsilon(\ln \pi_{xy} - 1)) \\ & + \sum_{x \in \mathcal{X}} \varphi(x) \left( \mu_x - \sum_{y \in \mathcal{Y}} \pi_{xy} \right) + \sum_{y \in \mathcal{Y}} \psi(y) \left( \nu_y - \sum_{x \in \mathcal{X}} \pi_{xy} \right).\end{aligned}$$

Las variables  $\{\varphi(x)\}_{x \in \mathcal{X}}$  y  $\{\psi(y)\}_{y \in \mathcal{Y}}$  son los multiplicadores de Lagrange. Una forma alternativa de expresar el Lagrangiano (y útil para aplicar condiciones de primer orden) es

$$\mathcal{L}(\pi, \varphi, \psi) = \sum_{x,y} \pi_{xy} \{c(x,y) - \psi(y) - \varphi(x) + \varepsilon(\ln(\pi_{xy}) - 1)\} \quad (6)$$

$$+ \sum_{x \in \mathcal{X}} \varphi(x) \mu_x + \sum_{y \in \mathcal{Y}} \psi(y) \nu_y \quad (7)$$

## Theorem

La solución al problema  $\mathcal{P}_\varepsilon$  es de la forma

$$\pi_{x,y} = e^{\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon}},$$

donde  $\varphi$  y  $\psi$  son a determinar.

## Prueba.

Aplicando condiciones de primer orden a (6) respecto a  $\pi_{x,y}$

$$c(x,y) - \varphi(x) - \psi(y) + \varepsilon(\ln \pi_{x,y} - 1) + \varepsilon = 0$$

$$\varepsilon \ln \pi_{x,y} = \varphi(x) + \psi(y) - c(x,y)$$

$$\pi_{x,y} = e^{\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon}}.$$



## Observación

A partir de  $\pi_{x,y} = e^{\frac{\varphi(x)+\psi(y)-c(x,y)}{\varepsilon}}$  tendremos

$$\pi = [\pi_{xy}] = D_{\varphi(x)} \underbrace{e^{-\frac{c(x,y)}{\varepsilon}}}_{K_\varepsilon} D_{\psi(y)} \quad (8)$$

donde  $D_{\varphi(x)}$  una matriz diagonal cuyos elementos en la diagonal son  $e^{\varphi(x)/\varepsilon}$ ,  $x \in \mathcal{X}$ .  
Análogo para  $D_{\psi(y)}$ ,  $y \in \mathcal{Y}$ .



Una vez  $\pi_{xy}$  determinado en función de  $\varphi_x$  y  $\psi_y$ , podemos determinar  $\min_{\pi} \mathcal{L}(\varphi, \psi)$  reemplazando (8) en  $\mathcal{L}(\pi, \varphi, \psi)$

$$\mathcal{L}(\varphi, \psi) = \sum_{x \in \mathcal{X}} \varphi_x \mu_x + \sum_{y \in \mathcal{Y}} \psi_y \mu_y - \varepsilon \sum_{x,y} \exp \left( \frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon} \right).$$

Finalmente,  $\max_{\varphi, \psi} \mathcal{L}(\varphi, \psi)$  conlleva por condiciones de primer orden (función objetivo estrictamente cóncava)

$$\begin{aligned} \mu_x &= \sum_{y \in \mathcal{Y}} \exp \left( \frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon} \right) \\ \nu_y &= \sum_{x \in \mathcal{X}} \exp \left( \frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon} \right). \end{aligned}$$

## Conclusiones.

- 1 Hemos presentado, para el caso de espacios finitos el problema de regularización entrópica.
- 2 Se han derivado diversas propiedades relativas a este problema: existencia de una solución, unicidad y como caracterizarla.
- 3 Hemos caracterizado el óptimo en términos de la relación entre  $\varphi$  y  $\psi$  y los parámetros  $\{\mu_x\}_{x \in \mathcal{X}}, \{\nu_y\}_{y \in \mathcal{Y}}, \{c_{xy}\}_{(x,y) \in \mathcal{X} \times \mathcal{Y}}$ .
- 4 A diferencia del problema clásico, hemos podido derivar  $\pi_{xy}$  en función de los multiplicadores y  $c_{xy}$ .
- 5 El objetivo a continuación es como obtener una solución aproximada.

# Sinkhorn-Knopp

- 1 Permite obtener una solución aproximada iterando sobre  $e^{\varphi^{(k)}}$  y  $e^{\psi^{(k)}}$ .
- 2 Complejidad computacional  $\mathcal{O}(N^2)$ .
- 3 El problema original se revuelve vía *Bertsekas' auction algorithm* en  $\mathcal{O}(N^3)$  [Merigot and Thibert, 2020].
- 4  $N = 100000$ , con Bertsekas son 115 días de ejecución mientras que con Sinkhorn-Knopp 0.12 días (teniendo en cuenta que  $\simeq 10^8$  operaciones se ejecutan por segundo).

- 1 Se de cumplir que

$$D_{\varphi} K_{\varepsilon} D_{\psi} 1_N = \mu \wedge D_{\psi} K_{\varepsilon}^T D_{\varphi} 1_N = \nu.$$

- 2 Expresado de otra forma

$$e^{\varphi/\varepsilon} \odot (K_{\varepsilon} e^{\psi/\varepsilon}) = \mu, \quad e^{\psi/\varepsilon} \odot (K_{\varepsilon}^T e^{\varphi/\varepsilon}) = \nu.$$

Acá  $e^{\varphi/\varepsilon}, e^{\psi/\varepsilon}$  son las respectivas diagonales de  $D_{\varphi}$  y  $D_{\psi}$  vectorizados. Usando estas ecuaciones, derivamos la siguiente situación:

$$e^{\varphi^{(k+1)}/\varepsilon} \doteq \frac{\mu}{K_{\varepsilon} e^{\psi^{(k)}/\varepsilon}}, \quad e^{\psi^{(k+1)}/\varepsilon} \doteq \frac{\nu}{K_{\varepsilon} e^{\varphi^{(k+1)}/\varepsilon}}$$

$e^{\psi^{(0)}/\varepsilon} = e^{\varphi^{(0)}/\varepsilon} = 1_N$ . Acá la división es componente a componente y  $\odot$  la multiplicación por entradas de un vector con otro vector.

---

**Algorithm** Sinkhorn-Knopp: algoritmo para el problema de transporte óptimo regularizado, caso discreto.

---

```
1: function Sinkhorn-Knopp( $K_\varepsilon, \mu, \nu$ )
2:    $e^{\varphi^{(0)}/\varepsilon} \leftarrow 1_N$ 
3:    $e^{\psi^{(0)}/\varepsilon} \leftarrow 1_N$ 
4:   for  $0 \leq k < k_{\max}$  do
5:      $e^{\varphi^{(k+1)}/\varepsilon} \leftarrow \frac{\mu}{K_\varepsilon e^{\psi^{(k)}/\varepsilon}}$ 
6:      $e^{\psi^{(k+1)}/\varepsilon} \leftarrow \frac{\nu}{K_\varepsilon^T e^{\varphi^{(k+1)}/\varepsilon}}$ 
7:   end for
8: end function
```

---

## Proposición

**Convergencia del Algoritmo Sinkhorn.** Se cumple que

- ❶  $(\varphi^{(k)}, \psi^{(k)}) \rightarrow (\varphi^*, \psi^*)$  con ratios de convergencia

$$d_{\mathcal{H}}(\varphi^{(k)}, \varphi^*) = \mathcal{O}(\lambda(K_{\varepsilon})^{2k})$$

$$d_{\mathcal{H}}(\psi^{(k)}, \psi^*) = \mathcal{O}(\lambda(K_{\varepsilon})^{2k}).$$

- ❷ Más aún, denotando  $\Pi^{(k)} = D_{\varphi^{(k)}} K_{\varepsilon} D_{\psi^{(k)}}$

$$d_{\mathcal{H}}(\varphi^{(k)}, \varphi^*) \leq \frac{d_{\mathcal{H}}(\Pi^{(k)} 1_N, \mu)}{1 - \lambda(K_{\varepsilon})}, \quad d_{\mathcal{H}}(\psi^{(k)}, \psi^*) \leq \frac{d_{\mathcal{H}}(\Pi^{(k)} 1_N, \nu)}{1 - \lambda(K_{\varepsilon})}.$$

## Proposición

La aplicación definida por:

$$\forall, x, y \in \mathbb{R}_{+*}^N : d_{\mathcal{H}}(x, y) = \ln \max_{i,j} \frac{x_i y_j}{x_j y_i}$$

es una métrica sobre el cono proyectivo  $\mathbb{R}_{+*}^N / \sim$  donde  $x \sim y$  si existe  $r > 0$  tal que  $x = ry$ . Más aún, el cono proyectivo es completo con la métrica  $d_{\mathcal{H}}$ .



## Prueba.

Ciertamente,  $d_{\mathcal{H}}$  es simétrica. Luego, si  $x = ry$ ,

$$d_{\mathcal{H}}(x, y) = d_{\mathcal{H}}(ry, y) = \ln \max_{i,j} \frac{ry_i y_j}{y_j r y_i} = \ln 1 = 0.$$

Por otro lado, si  $\frac{x_i y_j}{x_j y_i} < 1$ , entonces  $\frac{x_j y_i}{x_i y_j} > 1$ . Así,  $d_{\mathcal{H}} \geq 0$ . Luego, si  $d_{\mathcal{H}}(x, y) = 0$ , entonces  $\frac{x_i y_j}{x_j y_i} = 1$  para todo  $i, j$ . Esto es,  $\frac{x_i}{y_i} = \frac{x_j}{y_j} = r$ :  $x = ry$ . Finalmente, dados  $x, y, z \in \mathbb{R}_{+*}^N$ , existen  $\ell, k \in \{1, \dots, N\}$  tales que

$$\begin{aligned} d_{\mathcal{H}}(x, z) &= \ln \frac{x_k z_{\ell}}{x_{\ell} z_k} \\ &= \ln \left( \frac{x_k y_{\ell}}{x_{\ell} y_k} \cdot \frac{y_k z_{\ell}}{y_{\ell} z_k} \right) \\ &= \ln \frac{x_k y_{\ell}}{x_{\ell} y_k} + \ln \frac{y_k z_{\ell}}{y_{\ell} z_k} \\ &\leq d_{\mathcal{H}}(x, y) + d_{\mathcal{H}}(y, z). \end{aligned}$$



## Lema

Sea  $M \in \mathcal{M}_{N \times N}^{*+}$ . Entonces, dados  $x, y \in \mathbb{R}_{+*}^N$

$$d_{\mathcal{H}}(Mx, My) \leq \lambda(M) d_{\mathcal{H}}(x, y)$$

donde

$$\begin{cases} \lambda(M) &= \frac{\sqrt{\eta(M)}-1}{\sqrt{\eta(M)}+1} < 1 \\ \eta(M) &= \max_{i,j,k,\ell} \frac{M_{i,k}M_{j,\ell}}{M_{j,k}M_{i,\ell}} \geq 1. \end{cases}$$

## Prueba.

Ver [Birkhoff, 1957]. □

Para simplificar la notación, denotamos  $e^{\frac{\varphi}{\varepsilon}}$  por  $\varphi$  y  $e^{\frac{\psi}{\varepsilon}}$  por  $\psi$ . Luego, para la inicialización, en rigor, solo basta que los vectores tengan entradas positivas. Finalmente, cabe mencionar la siguiente situación; todo re-escalamiento  $c\varphi, \frac{1}{c}\psi$  sigue siendo solución,  $c > 0$ .

# Convergencia del algoritmo Sinkhorn-Knopp

## Prueba.

Para cualesquiera  $x, y \in \mathbb{R}_{+*}^N$  tenemos

$$d_{\mathcal{H}}(x, y) = d_{\mathcal{H}}(x/y, 1_N) = d_{\mathcal{H}}(1_N/x, 1_N/y)$$

donde  $/$  representa la división entrada por entrada. Luego, por la definición del algoritmo y el Lema (1)

$$\begin{aligned} d_{\mathcal{H}}(\varphi^{(k+1)}, \varphi^*) &= d_{\mathcal{H}}\left(\frac{\mu}{K_{\varepsilon}\psi^{(k)}}, \frac{\mu}{K_{\varepsilon}\psi^*}\right) \\ &= d_{\mathcal{H}}(K_{\varepsilon}\psi^{(k)}, K_{\varepsilon}\psi^*). \\ &\leq \lambda(K_{\varepsilon})d_{\mathcal{H}}(\psi^{(k)}, \psi^*). \end{aligned}$$



## Prueba.

De manera análoga

$$\begin{aligned}d_{\mathcal{H}}(\psi^{(k)}, \psi^*) &= d_{\mathcal{H}}\left(\frac{\nu}{K_{\varepsilon}^T \varphi^{(k)}}, \frac{\nu}{K_{\varepsilon}^T \varphi^*}\right) \\&= d_{\mathcal{H}}(K_{\varepsilon}^T \varphi^{(k)}, K_{\varepsilon}^T \varphi^*) \\&\leq \lambda(K_{\varepsilon}^T) d_{\mathcal{H}}(\varphi^{(k)}, \varphi^*).\end{aligned}$$

Como  $\lambda(K_{\varepsilon}) = \lambda(K_{\varepsilon}^T)$

$$\begin{aligned}d_{\mathcal{H}}(\varphi^{(k+1)}, \varphi^*) &\leq \lambda(K_{\varepsilon})^2 d_{\mathcal{H}}(\varphi^{(k)}, \varphi^*) \\d_{\mathcal{H}}(\psi^{(k)}, \psi^*) &\leq \lambda(K_{\varepsilon})^2 d_{\mathcal{H}}(\psi^{(k-1)}, \psi^*).\end{aligned}$$

Dado que  $\lambda(K)^2 < 1$  concluimos que  $\varphi^{(k)} \rightarrow \varphi^*$  y  $\psi^{(k)} \rightarrow \psi^*$ . □

# Las aplicaciones

Hoja de ruta:

- 1 El problema base.
- 2 Estimación de costos y objetivos [Dupuy et al., 2021].
- 3 Algunos resultados de subdiferencial y análisis convexo (enunciados, pruebas si el tiempo lo permite).
- 4 SISTA.
- 5 Convergencia del algoritmo SISTA.
- 6 Matching.

Nos situamos nuevamente en el contexto de conjuntos finitos de cardinalidad  $N$  y se busca resolver

$$\min_{\pi \in \Pi(\mu, \nu)} \left\{ \sum_{x,y} \pi_{xy} c_{xy} \right\}.$$

Por propósitos de notación, re-escribimos (pues es más apropiado para el contexto del problema), usando  $(i, j)$

$$\min_{\pi \in \Pi(\mu, \nu)} \left\{ \sum_{1 \leq i, j \leq N} \pi_{ij} c_{ij} \right\}.$$

En efecto,  $i \rightarrow j$  representa la migración entre países. Debido a los beneficios computacionales expuestos previamente, se incorpora el término de regularización entrópica y el problema de optimización se torna

$$\min_{\pi \in \Pi(\mu, \nu)} \left\{ \sum_{1 \leq i, j \leq N} \pi_{ij} c_{ij} + \varepsilon \pi_{ij} (\ln \pi_{ij} - 1) \right\} \quad (9)$$

con

$$\Pi(\mu, \nu) = \left\{ \pi_{ij} \geq 0 : \sum_{j=1}^N \pi_{ij} = \mu_i, \sum_{i=1}^N \pi_{ij} = \nu_j \right\}.$$

Ya sabemos que

$$\pi_{ij} = \exp\left(\frac{\varphi_i + \psi_j - c_{ij}}{\varepsilon}\right).$$

Vamos a considerar  $\varepsilon = 1$ . Luego, el algoritmo Sinkhorn, adaptado a este contexto es el siguiente:

$$\begin{aligned}\exp(\varphi_i^{(k+1)}) &= \frac{\mu_i}{\sum_{j=1}^N \exp(\psi_j^{(k)} - c_{ij})} \\ \exp(\psi_j^{(k+1)}) &= \frac{\nu_j}{\sum_{i=1}^N \exp(\varphi_i^{(k+1)} - c_{ij})}.\end{aligned}$$

En muchas situaciones, el problema es inverso: se posee el transport plan y se desea conocer qué costos generaron el transport plan. En este contexto, dicho problema inverso es atacado vía estimación paramétrica.



En este caso, la estructura de costos viene determinada completamente por un vector de parámetros  $\beta \in \mathbb{R}^K$  de forma que

$$c_{ij}^{\beta} = \sum_{k=1}^K \beta_k d_{ij}^k.$$

Matricialmente:

$$\begin{pmatrix} c_{11}^{\beta} & c_{12}^{\beta} & & \\ c_{21}^{\beta} & \ddots & & \\ & & \ddots & \\ & & & c_{NN}^{\beta} \end{pmatrix} = \sum_{k=1}^K \beta_k \left\{ \begin{pmatrix} d_{11}^k & d_{12}^k & & \\ d_{21}^k & \ddots & & \\ & & \ddots & \\ & & & d_{NN}^k \end{pmatrix} \right\}. \quad (10)$$

En (10)  $d_{ij}$  mide la disimilitud entre  $i$  y  $j$ . Por ejemplo, en caso de problemas espaciales, puede ser una distancia, diferencias de PBI, diferencias de población etc.

Lo que se busca es  $\beta$ ,

$$\pi_{ij}^{\beta} = \exp(\varphi_i + \psi_j - c_{ij}^{\beta})$$

tal que

$$\sum_i \pi_{ij}^{\beta} = \nu_j$$

$$\sum_j \pi_{ij}^{\beta} = \mu_i$$

$$\sum_{i,j} \pi_{ij}^{\beta} d_{ij}^k = \sum_{i,j} \underbrace{\hat{\pi}_{ij}}_{\text{conocido y } \in \Pi(\mu, \nu)} \underbrace{d_{ij}^k}_{\text{dato}}, \quad \forall k = 1, \dots, K.$$

El siguiente resultado es desarrollado en [Galichon and Salanié, 2022]. Para estimar  $\varphi, \psi$  y  $\beta$  se resuelve

$$\min_{\varphi, \psi, \beta} \underbrace{\left\{ \sum_{1 \leq i, j \leq N} \exp(\varphi_i + \psi_j - \underbrace{c_{ij}^\beta}_{=\sum_{k=1}^K \beta_k d_{ij}^k}) + \sum_{1 \leq i, j \leq N} \hat{\pi}_{ij} (c_{ij}^\beta - \varphi_i - \psi_j) \right\}}_{F(\varphi, \psi, \beta), \text{ convexa}}.$$

$$\textcircled{1} \quad \frac{\partial F}{\partial \varphi_i} = \sum_{j=1}^N \exp(\varphi_i + \psi_j - c_{ij}^\beta) - \underbrace{\sum_{j=1}^N \hat{\pi}_{ij}}_{=\mu_i} = 0, \quad \forall i.$$

$$\textcircled{2} \quad \frac{\partial F}{\partial \psi_j} = \sum_{i=1}^N \exp(\varphi_i + \psi_j - c_{ij}^\beta) - \underbrace{\sum_{i=1}^N \hat{\pi}_{ij}}_{=\nu_j} = 0, \quad \forall j.$$

$$\textcircled{3} \quad \frac{\partial F}{\partial \beta_k} = - \sum_{1 \leq i, j \leq N} d_{ij}^k \exp(\varphi_i + \psi_j - c_{ij}^\beta) + \sum_{1 \leq i, j \leq N} \hat{\pi}_{ij} d_{ij}^k.$$

Con el objetivo de capturar solo los efectos importantes, se agrega una penalización tipo Lasso:

$$\min_{\varphi, \psi, \beta} \Phi(\varphi, \psi, \beta) = \left\{ \sum_{1 \leq i, j \leq N} \exp(\varphi_i + \psi_j - c_{ij}^{\beta}) - \hat{\pi}_{ij}(\varphi_i + \psi_j - c_{ij}^{\beta}) + \gamma \|\beta\|_1 \right\}. \quad (11)$$

A continuación, brindamos algunos resultados preliminares del análisis convexo y optimización. Luego, presentamos la forma en la que se resuelve (11): el algoritmo SISTA.

# Preliminares

## Sub-diferenciales

Dada  $f : S \subset \mathbb{R}^n \rightarrow \mathbb{R}$  convexa y diferenciables, se cumple que

$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

¿Y si  $f$  no es diferenciable?

### Definition

Sea  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . El epigrafo de  $f$  es el conjunto

$$\mathcal{E} = \{(x, y) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq y\} \subset \mathbb{R}^{n+1}.$$

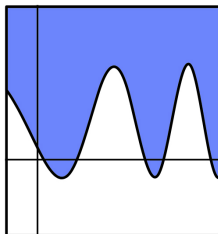


Figura Epigrafo.

## Definition

$g \in \mathbb{R}^n$  es un subgradiente de  $f : S \subset \mathbb{R}^n \rightarrow \mathbb{R}$  en  $x$  si

$$f(y) \geq f(x) + g^T(y - x), \quad \forall y \in S.$$

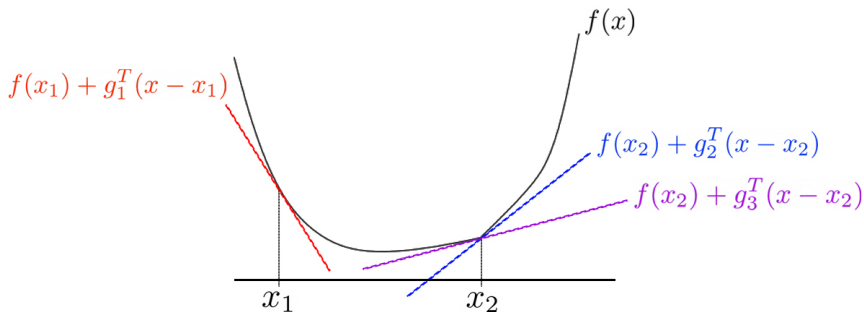


Figura Sub-gradients.

## Definition

El subdiferencial de  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  en  $x$ , denotado  $\partial f(x)$  es el conjunto de subgradiientes en  $x$ . Esto es,

$$\partial f(x) = \{y \in \mathbb{R}^d : f(z) \geq f(x) + \langle y, z - x \rangle, \forall z \in \mathbb{R}^d\}.$$

## Proposición.

El subdiferencial en un punto interior del dominio de finitud de  $f$  es no vacío cuando  $f$  es convexa.



## Prueba.

Esto es solo un sketch:

- 1 Tomamos  $x_0 \in D^\circ$  y definimos  $A = \{(x, y) \in D^\circ \times \mathbb{R} : y > f(x)\}$ .
- 2 Por el Teorema de Hahn-Banach, existe un hiperplano que separa  $(x_0, f(x_0))$  y  $A$ .
- 3 Por el Teorema de Representación de Riesz-Fréchet:  $[\varphi = c]$  es de la forma  $\langle u, (x, y) \rangle : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$ .
- 4 Suponiendo que  $u_{d+1} = 0$ , esto es,

$$[\varphi = c] = \{(x, y) : \langle w, x \rangle = c\}$$

se llega a una contradicción. Se usa que  $\varphi(C)$ , con  $C$  convexo abierto no vacío es un intervalo abierto.

- 5 De este modo,

$$[\varphi = c] = \{(x, y) : \langle w, x \rangle + y = c\}.$$

- 6 Concluimos que  $w \in \partial f(x_0)$ .



### Proposición.

Si  $f : \mathbb{X} \rightarrow \mathbb{R}$  es convexa y  $C^1$ , entonces  $\partial f(x) = \{\nabla f(x)\}$ .

### Prueba.

Ciertamente,  $\nabla f(x) \in \partial f(x)$ . Ahora, supongamos que existe otro elemento, digamos  $y \in \partial f(x)$ . Por definición  $f(x') \geq f(x) + \langle y, x' - x \rangle$ ,  $x' \in \mathbb{X}$ . Tomamos  $x' = x + tz$  para  $z \in \mathbb{X}$  y  $t > 0$ . Entonces,

$$\frac{f(x + tz) - f(x)}{t} \geq \langle y, z \rangle.$$

Haciendo  $t \rightarrow 0$ ,

$$\langle \nabla f(x), z \rangle \geq \langle y, z \rangle.$$

Pero entonces,  $\langle \nabla f(x) - y, z \rangle \geq 0$ . Al ser  $z$  arbitrario, tomamos  $z = -(\nabla f(x) - y)$ . Esto nos permite obtener

$$\|\nabla f(x) - y\| \leq 0 \implies y = \nabla f(x).$$



### Proposición.

$x^*$  minimiza  $f$  si y solamente si  $0 \in \partial f(x^*)$ .

### Prueba.

Tenemos:

$$f(x^*) = \min_x f(x)$$

$$\Leftrightarrow$$

$$f(x^*) \leq f(y), \forall y$$

$$\Leftrightarrow$$

$$f(y) \geq f(x^*) + 0^T(y - x^*), \forall y$$

$$\Leftrightarrow$$

$$0 \in \partial f(x^*).$$



## Lema

Sea  $f(x) = \max\{f_1(x), f_2(x)\}$ , con  $f_1, f_2$  diferenciables. Entonces,

$$\partial f(x) = \begin{cases} \nabla f_1(x), & \text{si } f_1(x) > f_2(x) \\ \nabla f_2(x), & \text{si } f_2(x) > f_1(x) \\ [\nabla f_1(x), \nabla f_2(x)] & \text{si } f_1(x) = f_2(x). \end{cases}$$

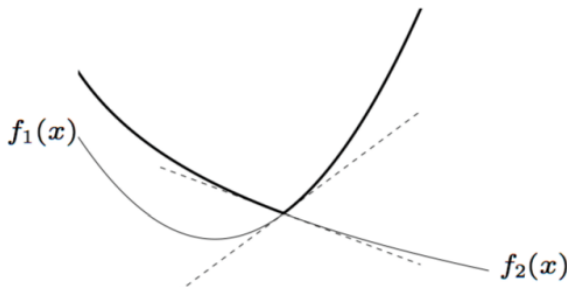


Figura Sub-diferencial  $\max\{f_1, f_2\}$ .

## Definition

**La función prox.** La función proximidad de una función convexa  $h(\cdot)$  es, para  $t > 0$

$$\text{prox}_{h,t}(x) = \underset{u}{\operatorname{argmin}} \left( h(u) + \frac{1}{2t} \|u - x\|_2^2 \right). \quad (12)$$

Si  $h(x) = \lambda \|x\|_1$ ,  $t = 1$

$$\text{prox}_h(x)_i = \begin{cases} x_i - \lambda, & \text{si } x_i \geq \lambda \\ 0, & \text{si } |x_i| \leq \lambda \\ x_i + \lambda, & \text{si } x_i \leq -\lambda. \end{cases} \quad (13)$$

En efecto,  $|x| = \max\{x, -x\}$ . Así pues, de (12), observando que el problema es separable, basta tener

$$0 \in \partial f(u_i^*).$$

Luego, por Moreau-Rockafellar,

$$\partial f(u_i^*) = \{u_i^* - x_i + \lambda \partial | \cdot |(u_i^*)\}.$$

Finalmente, dado que

$$\partial | \cdot |(u_i^*) = \begin{cases} -1, & \text{si } u_i^* < 0 \\ [-1, 1], & \text{si } u_i^* = 0 \\ 1, & \text{si } u_i^* > 0 \end{cases}$$

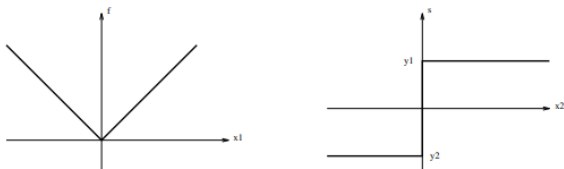


Figura Sub-diferencial  $|\cdot|$ .

Concluimos que

$$u_i^* = \begin{cases} x_i - \lambda, & \text{si } x_i \geq \lambda \\ 0, & \text{si } |x_i| \leq \lambda \\ x_i + \lambda, & \text{si } x_i \leq -\lambda. \end{cases}$$

Ahora bien, en el caso de tener

$$f(x) = g(x) + h(x) \quad (14)$$

con  $g$  convexa y diferenciable sobre su dominio y  $h$  convexa pero eventualmente no diferenciable, se ejecuta el siguiente algoritmo con la finalidad de minimizar (14):

$$x^{(k)} = \text{prox}_{t_k, h}(x^{(k-1)} - t_k \nabla g(x^{(k-1)})).$$

¿Cuál es la intuición?

$$\begin{aligned} x^{(k)} &= \operatorname{argmin}_u \left( h(u) + \frac{1}{2t_k} \left\| u - x^{(k-1)} - t_k \nabla g(x^{(k-1)}) \right\|_2^2 \right) \\ &= \operatorname{argmin}_u \left( h(u) + g(x) + \nabla g(x)^T (u - x) + \frac{1}{2t} \|u - x\|_2^2 \right). \end{aligned}$$

Lo que minimizamos es  $h(u)$  más la diferencia de  $u$  y el clásico gradient-descent.

# SISTA

El nombre SISTA es un algoritmo híbrido entre 2 algoritmos: **Sinkhorn** (S) y **Proximal Gradient Descent** (ISTA). Queda por presentar el Proximal Gradient Descent.



El algoritmo SISTA - Sinkhorn (S) y Proximal Gradient Descent (ISTA) - consiste en minimizar (11) en tres etapas:

- 1 Iterar  $\varphi$ , manteniendo  $\beta$  y  $\psi$  constantes.
- 2 Iterar  $\psi$ , manteniendo  $\beta$  y  $\varphi$  constantes.
- 3 Descenso de gradiente proximal respecto a  $\beta$  manteniendo  $\varphi$  y  $\psi$  constantes.

## Definition

**Función**  $\text{prox}_{\gamma \|\cdot\|_1}$ .

$$\text{prox}_{\rho, \gamma, \|\cdot\|_1}(z) = \begin{cases} z - \rho\gamma, & \text{si } z > \rho\gamma \\ 0, & \text{si } |z| \leq \rho\gamma \\ z + \rho\gamma, & \text{si } z < -\rho\gamma. \end{cases}$$

Acá  $\rho$  es el parámetro en el gradient-descent.

## Algorithm SISTA.

1:  $\beta^{(0)}, \rho, \mu, \nu, d_{ij}^k, \hat{\pi}_{ij}, \varphi^{(0)}, \psi^{(0)}$

2:

3: **while** not converged **do**

4:     Planteamos  $c_{ij}^{\beta^{(t)}} = \sum_{k=1}^K \beta_k^{(t)} d_{ij}^k$

$$\begin{cases} \exp(\varphi_i^{(t+1)}) &= \frac{\mu_i}{\sum_{j=1}^N \exp(\psi_j^{(t)} - c_{ij}^{\beta^{(t)}})} \\ \exp(\psi_j^{(t+1)}) &= \frac{\nu_j}{\sum_{i=1}^N \exp(\varphi_i^{(t+1)} - c_{ij}^{\beta^{(t)}})} \end{cases}$$

5:     Sea  $\pi_{ij}^{\beta^{(t)}} = \exp(\varphi_i^{(t+1)} + \psi_j^{(t+1)} - c_{ij}^{\beta^{(t)}})$ . Para  $k = 1, \dots, K$

$$\beta_k^{(t+1)} = \text{prox}_{\rho, \gamma, \|\cdot\|} \left( \beta_k^{(t)} - \rho \underbrace{\sum_{1 \leq j, i \leq N} (\hat{\pi}_{ij} - \pi_{ij}^{\beta^{(t)}}) d_{ij}^{(k)}}_{\nabla_{\beta} F} \right).$$

6: **end while**

7: **return**  $\beta$

# Convergencia del algoritmo SISTA

Supuestos:

$$\sum_{i=1}^N d_{ij}^{(k)} = 0, \quad \forall j = 1, \dots, N \quad (15)$$

$$\sum_{j=1}^N d_{ij}^{(k)} = 0, \quad \forall i = 1, \dots, N \quad (16)$$

$$\{d_1, \dots, d^K\} \quad (17)$$

son l.i. Además,

$$\hat{\pi}_{ij} > 0. \quad (18)$$

## Observación

Siempre podemos obtener  $\sum_{i,j} d_{ij} = 0$  re-definiendo

$$\tilde{d}_{ij}^k = d_{ij}^k - a_i^k - b_j^k$$

con  $a_i^k = \frac{1}{N} \sum_{j=1}^N d_{ij}^k$  y  $b_j^k = \frac{1}{N} \sum_{i=1}^N d_{ij}^k - \frac{1}{N^2} \sum_{1 \leq p, q \leq N} d_{pq}^k$ .

## Prueba.

Sumando:

$$\begin{aligned}\sum_{1 \leq i, j \leq N} \tilde{d}_{ij}^k &= \sum_{1 \leq i, j \leq N} (d_{ij}^k - a_i^k - b_j^k) \\&= \sum_{1 \leq i, j \leq N} d_{ij}^k - \sum_{1 \leq i, j \leq N} \left[ \frac{1}{N} \sum_{j=1}^N d_{ij}^k \right] - \sum_{1 \leq i, j \leq N} \left[ \frac{1}{N} \sum_{i=1}^N d_{ij}^k - \frac{1}{N^2} \sum_{1 \leq p, q \leq N} d_{pq}^k \right] \\&= \sum_{1 \leq i, j \leq N} d_{ij}^k - \sum_{1 \leq i \leq N} \sum_{1 \leq j \leq N} N \cdot \frac{1}{N} d_{ij}^k - \sum_{1 \leq j \leq N} \sum_{1 \leq i \leq N} N \cdot \frac{1}{N} d_{ij}^k \\&\quad + N^2 \frac{1}{N^2} \sum_{1 \leq p, q \leq N} d_{pq}^k \\&= 2 \sum_{1 \leq i, j \leq N} d_{ij}^k - 2 \sum_{1 \leq i, j \leq N} d_{ij}^k \\&= 0.\end{aligned}$$



Nuestro objetivo es resolver

$$\inf_{(\varphi, \psi, \beta) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^K} \Phi(\varphi, \psi, \beta) = G \circ \Lambda(\varphi, \psi, \beta) + \gamma \|\beta\|_1 \quad (19)$$

donde  $\Lambda : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^K \rightarrow \mathcal{M}_{N \times N}$  es una lineal definida de la siguiente forma:

$$(\Lambda(\varphi, \psi, \beta))_{ij} = \varphi_i + \psi_j - c_{ij}^\beta, \quad \forall (i, j)$$

y  $G$  una función  $C^\infty$  y convexa, con regla de correspondencia

$$G(\lambda) = \sum_{1 \leq i, j \leq N} (\exp(\lambda_{ij}) - \hat{\pi}_{ij} \lambda_{ij}), \quad \forall \lambda \in \mathcal{M}_{N \times N}.$$

Establecemos en  $\mathcal{M}_{N \times N}$  el producto interno  $\langle A, B \rangle_F$ :

$$\langle A, B \rangle_F = \sum_{i,j} a_{ij} b_{ij}.$$

## Definition

Sean  $X, Y$  espacios de Banach normados y  $f : X \rightarrow Y$ . Diremos que  $f$  es Fréchet diferenciable en  $x_0 \in X$  si existe una transformación lineal  $T : X \rightarrow Y$  tal que

$$\lim_{h \rightarrow 0} \frac{\|f(x_0 + h) - f(x_0) - T(h)\|_Y}{\|h\|_X} = 0.$$

En dicho caso,  $f'(x_0)$  es la derivada de Fréchet en  $x_0$ .

Si  $X$  es un espacio de Hilbert sobre  $\mathbb{R}$  y  $f : X \rightarrow \mathbb{R}$ , entonces, si la derivada de Fréchet en  $x \in X$  existe, es un funcional lineal continuo. Por el Teorema de Riesz Fréchet, existe  $u \in X$  tal que  $f'(x)(h) = \langle h, u \rangle$  para todo  $h \in X$ . Dicho vector  $u$  se conoce como el vector gradiente  $\nabla f(x)$ . De este modo,

$$f'(x)(h) = \langle h, \nabla f(x) \rangle, \quad \forall h \in X.$$



Calculamos ahora  $\nabla F(x)$ :

$$dF_x(v) : \mathbb{R}^{2N+K} \rightarrow \mathbb{R}$$

y  $dF_x(v) = \langle \nabla F(x), v \rangle$ . Por otro lado,

$$dF_x(v) = dG_{\Lambda(x)} \circ d\Lambda_x(v).$$

Ahora bien,

$$\nabla F(x)^T = df_x^T = \Lambda^T \nabla G(x) = \Lambda^T [e^{\Lambda(x)} - \hat{\pi}]_{ij}.$$

Respecto a la Hessiana,  $\mathbf{H}(F(\mathbf{x})) = \mathbf{J}(\nabla F(\mathbf{x}))^T$ . Primero,

$$\mathbf{H}(G(\lambda)) = \text{diag}(\exp(\lambda)).$$

En efecto,

$$\mathbf{H}(G(\lambda)) = \begin{pmatrix} \nabla \left[ \frac{\partial G}{\partial \lambda_{11}} \right] \\ \nabla \left[ \frac{\partial G}{\partial \lambda_{12}} \right] \\ \vdots \\ \nabla \left[ \frac{\partial G}{\partial \lambda_{NN}} \right] \end{pmatrix} = \begin{pmatrix} \nabla(e^{\lambda_{11}} - \pi_{11}) \\ \nabla(e^{\lambda_{12}} - \pi_{12}) \\ \vdots \\ \nabla(e^{\lambda_{NN}} - \pi_{NN}) \end{pmatrix} = \begin{pmatrix} e^{\lambda_{11}} & & & \\ & e^{\lambda_{12}} & & \\ & & \ddots & \\ & & & e^{\lambda_{NN}} \end{pmatrix}.$$

Aplicando la regla de la cadena:

$$\begin{aligned} J(\nabla F(x)) &= d(\nabla F)_x \\ &= d(\Lambda^T \circ \nabla G)_x \\ &= d(\Lambda^T(\exp(\Lambda(x)) - \hat{\pi})) \\ &= \Lambda^T d(\exp(\Lambda(x))) \\ &= \Lambda^T \text{diag}(\exp(\Lambda(x)))\Lambda. \end{aligned}$$

## Lema

Bajo los supuestos,  $\Lambda$  es inyectiva de

$$E = \{(\varphi, \psi, \beta) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^K : \varphi_1 = 0\} \simeq \mathbb{R}^{2N-1+K}.$$

a  $\mathcal{M}_{N \times N}$ .

## Prueba.

Sea  $(\varphi, \psi, \beta) \in E$ , perteneciendo al núcleo de  $\Lambda$ , entonces  $\forall i, j: \varphi_i + \psi_j = \sum_{k=1}^K \beta_k d_{ij}^k$ .  
Luego, por los supuestos

$$\begin{aligned} \sum_{j=1}^N \{\varphi_i + \psi_j\} &= \sum_{j=1}^N \sum_{k=1}^K \beta_k d_{ij}^k \\ &= \sum_{k=1}^K \underbrace{\left[ \sum_{j=1}^N \beta_k d_{ij}^k \right]}_{=0}. \end{aligned}$$

Esto es,  $N\varphi_i + \sum_{j=1}^N \psi_j = 0$ , para todo  $i$ . En particular, para  $i = 1$ ,  $\sum_{j=1}^N \psi_j = 0$ . Así,  $\varphi = 0$ . Pero entonces,  $\psi = 0$  también. Con lo cual,  $\sum_{k=1}^K \beta_k d^k = 0$ . Por la independencia lineal de  $\{d^k\}_{k=1}^K$ ,  $\beta = \mathbf{0}$ . Así,  $\{\mathbf{0}_{\mathbb{R}^N}, \mathbf{0}_{\mathbb{R}^N}, \mathbf{0}_{\mathbb{R}^K}\} = \text{Ker}(\Lambda)$ . □

## Lema

Para  $x, y \in E$  se cumple existen  $\alpha = \alpha(M)$  y  $v = v(M)$ , con

$$\max\{\|\Lambda(x)\|_\infty, \|\Lambda(y)\|_\infty\} \leq M,$$

tales que

$$F(x) \geq F(y) + \nabla F(y)(x - y) + \frac{v}{2}\|x - y\|_2^2$$

$$\|\nabla F(x) - \nabla F(y)\|_2 \leq \alpha\|x - y\|_2.$$

Proponemos  $v = e^{-M}\sigma_{\min}$  donde  $\sigma_{\min}$  es el menor autovalor de la matriz simétrica  $\Lambda^T\Lambda$ . Por el Teorema de Taylor, existe  $c \in [x, y]$  tal que

$$F(y) = F(x) + \nabla F(x)^T(y - x) + \frac{1}{2}(y - x)^T \mathbf{H}(F(c))(y - x).$$

Luego,  $c = \theta x + (1 - \theta)y$ ,  $\theta \in [0, 1]$ . Analicemos  $(y - x)^T \mathbf{H}(F(c))(y - x)$ :

$$\begin{aligned} (y - x)^T \mathbf{H}(F(c))(y - x) &= (y - x)^T \Lambda^T \text{diag}(\exp(\Lambda(\theta x + (1 - \theta)y))) \Lambda (y - x)^T \\ &\stackrel{\text{linealidad de } \Lambda}{=} (y - x)^T \Lambda^T \text{diag}(\exp(\theta \Lambda(x) + (1 - \theta) \Lambda(y))) \Lambda (y - x) \\ &\geq (y - x)^T \Lambda^T \text{diag}(\exp(-M)) \Lambda (y - x) \\ &\geq e^{-M} \sigma_{\min} \|y - x\|_2^2. \end{aligned}$$

La última desigualdad se explica a continuación.

$$\begin{aligned}
\min_{\|n\|=1} \{n^T A n\} &= \min_{\|n\|=1} \{n^T (UDU^T) n\} \\
&= \min_{\|n\|=1} \{(U^T n)^T D (U^T n)\} \\
&= \min_{\|U^T n\|=1} \{(U^T n)^T D (U^T n)\} \\
&= \min_{\|z\|=1} \{z^T D z\} \\
&= \min_{\|z\|=1} \sum_i z_i^2 D_{ii} \\
&= \min_{\|z\|=1} \sum_i z_i^2 \lambda_i \\
&\geq \min_i \{\lambda_i\} \underbrace{\sum_i z_i^2}_{=1}.
\end{aligned}$$

Así,

$$\frac{1}{\|y - x\|_2^2} (y - x)^T (\Lambda^T \Lambda) (y - x) \geq \min_i \lambda_i.$$



Con respecto a la segunda desigualdad:  $\|\nabla F(x) - \nabla F(y)\|_2 \leq \alpha \|x - y\|_2$ , esto es consecuencia de que

$$\|g(x) - g(y)\| \leq \sup_{z \in (x,y)} \|g'(z)\| \cdot \|x - y\|$$

y

$$\begin{aligned} \sup_{z \in (x,y)} \|g'(z)\| &= \sup_{z \in (x,y)} \|J(\nabla F(z))\| \\ &= \sup_{z \in (x,y)} \|H(z)\| \\ &= \sup_{z \in (x,y)} \|\Lambda^T \text{diag}(\exp(\Lambda(z))) \Lambda\| \\ &\leq e^M \|\Lambda^T \Lambda\| \\ &\leq e^M \|\Lambda^T\| \cdot \|\Lambda\|. \end{aligned}$$

## Proposición

Bajo los supuestos:

- 1  $\Phi$  es coerciva en  $\mathbb{R}^{2N-1+K}$ , i.e., los conjuntos  $\{x \in \mathbb{R}^{2N-1+K} : f(x) \leq c\}$  son acotados (compactos de hecho por la continuidad).
- 2 El problema (19) admite una única solución  $x^* = (\varphi^*, \psi^*, \beta^*)$ .
- 3 El óptimo  $x^*$  viene caracterizado por

$$\nabla_{\varphi} F(x^*) = 0, \nabla_{\psi} F(x^*) = 0, -\nabla_{\beta} F(x^*) \in \gamma \partial \|\cdot\|_1(\beta^*). \quad (20)$$

## Lema

Toda función coerciva  $f : X \subset \mathbb{R}^n \rightarrow \mathbb{R}$ , definida en un cerrado, que es continua, posee un mínimo.

# Sketch de la prueba de la convergencia

El Teorema en cuestión es:

### Theorem

Supongamos que tenemos 15, 15 17 y 18. Entonces, la sucesión  $x^{(t)} = (\varphi^{(t)}, \psi^{(t)}, \beta^{(t)})$  generada por el algoritmo SISTA converge a la solución del problema de optimización,  $\min_{(\varphi, \psi, \beta) \in K} \Phi(\varphi, \psi, \beta)$ , cuando  $t \rightarrow \infty$  (para un  $\rho$  suficientemente chico que será dado), con  $\varphi(0), \psi(0) \in \mathbb{R}_+^N$  y  $\beta^{(0)}$  un guess. Más aún, existe  $\delta > 0$  tal que

$$\Phi(x^{(t)}) - \Phi(x^*) \leq \frac{\Phi(x^{(0)}) - \Phi(x^*)}{(1 + \delta)^t}. \quad (21)$$

Teníamos

$$\begin{cases} \exp(\varphi_i^{(t+1)}) &= \frac{\mu_i}{\sum_{j=1}^N \exp(\psi_j^{(t)} - c_{ij}^{\beta^{(t)}})} \\ \exp(\psi_j^{(t+1)}) &= \frac{\nu_j}{\sum_{i=1}^N \exp(\varphi_i^{(t+1)} - c_{ij}^{\beta^{(t)}})}. \end{cases}$$

$$\beta^{(t+1)} = \text{prox}_{\rho, \gamma, \|\cdot\|} \left( \beta_k^{(t)} - \rho \underbrace{\sum_{1 \leq j, i \leq N} (\hat{\pi}_{ij} - \pi_{ij}^{\beta^{(t)}}) d_{ij}^{(k)}}_{\nabla_{\beta} F} \right).$$

Esto es lo mismo que minimizar respecto a  $\beta$

$$\rho\gamma\|\beta\|_1 + \frac{1}{2}\|\beta - (\beta^{(t)} - \rho\nabla_{\beta} F(\varphi^{(t+1)}, \psi^{(t+1)}, \beta^{(t)}))\|_2^2. \quad (22)$$

### Sketch de la prueba:

- ❶ Definimos  $C = \Phi(x^{(0)}) = \Phi(\varphi^{(0)}, \psi^{(0)}, \beta^{(0)})$ . Luego, como

$$\Phi(\varphi, \psi, \beta) \geq \sum_{1 \leq i, j \leq N} \hat{\pi}_{ij} |\Lambda_{ij}(\varphi, \psi, \beta)| - 2N^2 \ln 2 + \gamma \|\beta\|_1 \quad (23)$$

tendremos

$$\Phi(\varphi, \psi, \beta) \leq C \implies \|\beta\|_1 \leq \frac{C + 2N^2 \ln 2}{\gamma}.$$

- ❷ Proponemos  $\theta \in C^\infty(\mathbb{R}, \mathbb{R})$ , función no decreciente tal que

$$\theta(t) = \begin{cases} t, & \text{si } t \leq A \\ \theta(t) \geq t, & \text{si } t \in [A, 2A] \\ 2A, & \text{si } t \geq 2A. \end{cases}$$

Definamos  $\tilde{F} = \theta \circ F$ .  $\nabla \tilde{F}$  continua siendo Lipschitziana pues  $\theta$  lo es (globalmente).  
Luego,

$$\tilde{F}(x) \leq \tilde{F}(y) + \nabla \tilde{F}(y)(x - y) + \frac{\alpha}{2} \|x - y\|_2^2.$$

En efecto, si consideramos  $\phi(t) = \tilde{F}(y + t(x - y))$

$$\begin{aligned} \tilde{F}(x) - \tilde{F}(y) - \langle \nabla \tilde{F}(y), x - y \rangle &= \int_0^1 \langle \nabla \tilde{F}(y + t(x - y)), x - y \rangle dt - \langle \nabla \tilde{F}(y), x - y \rangle \\ &= \int_0^1 \langle \nabla \tilde{F}(y + t(x - y)) - \nabla \tilde{F}(y), x - y \rangle dt \\ &\leq \int_0^1 \|\nabla \tilde{F}(y + t(x - y)) - \nabla \tilde{F}(y)\|_2 \|x - y\|_2 dt \\ &\leq \int_0^1 \alpha t \|x - y\|_2^2 dt \\ &= \frac{\alpha}{2} \|x - y\|_2^2. \end{aligned}$$

Finalmente,

- ❶ Se prueba que, para  $\rho \in (0, \alpha^{-1}]$ :

$$C \geq \Phi(\varphi^{(t)}, \psi^{(t)}, \beta^{(t)}) \geq \Phi(\varphi^{(t+1)}, \psi^{(t+1)}, \beta^{(t)}).$$

y

$$C \geq \Phi(\varphi^{(t+1)}, \psi^{(t+1)}, \beta^{(t+1)}).$$

- ❷  $\Phi^{(t)}$  es decreciente y  $\Phi \in [-2N^2 \ln 2, C]$ .
- ❸  $\Phi^{(t)} \rightarrow \Phi^* = \Phi(\varphi^*, \psi^*, \beta^*)$ .
- ❹ Se establece que  $A_{t-1} - A_t \geq \frac{\nu}{2} \|x_t - x_{t-1}\|_2^2$ ,  $\forall t \geq 1$ , con  $A_t = \Phi(x_t) - \Phi(x^*)$ .
- ❺ Definiendo  $\delta = \frac{\nu^2 \rho^2}{2\alpha^2 \rho^2 + 1}$ , como  $A_t \leq A_{t-1} - \frac{\nu}{2} \|x_t - x_{t-1}\|_2^2$

$$A_t \leq \frac{A_{t-1}}{1 + \delta}. \quad (24)$$



## Conclusiones del estudio

El algoritmo permite obtener  $\beta$  y se concluye que las variables más importantes son: el idioma, la relación colonial, la distancia geográfica, una dummy del estado de los servicios públicos, la interacción entre la superficie geográfica de los países, y la esperanza de vida de las mujeres (la interacción de las variables según el criterio origen-destino).

# Matching

Un agente de decisión  $n$  debe elegir entre  $\{1, \dots, J\}$  opciones que le generan una utilidad  $U_{nj} = V_{nj} + \varepsilon_{nj}$ ,  $\forall j = 1, \dots, J$ . El término  $V_{nj}$  es conocido, mientras que  $\varepsilon_{nj}$  es un término estocástico. El supuesto cada elemento de la  $\{\varepsilon_{nj}\}_j$  es independiente y se distribuye según una Extreme Value de tipo 1:

$$F(\varepsilon_{nj}) = e^{-e^{-\varepsilon_{nj}}}$$

$$f(\varepsilon_{nj}) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}}.$$

Luego, siguiendo a [McFadden, 1974] y [Echenique and Chambers, 2016] la probabilidad de que el agente  $n$  escoja la alternativa  $i$  es

$$\mathbb{P}(V_{ni} + \varepsilon_{ni} \geq V_{nj} + \varepsilon_{nj}, \forall j \neq i).$$

Esto es,

$$\mathbb{P}(\varepsilon_{nj} < V_{ni} - V_{nj} + \varepsilon_{ni}, \forall j \neq i).$$

Usando la independencia,

$$\mathbb{P}_{ni}|\varepsilon_{ni} = \mathbb{P}(\varepsilon_{nj} < V_{ni} - V_{nj} + \varepsilon_{ni}, \forall j \neq i|\varepsilon_{ni}) = \prod_{j \neq i} e^{-e^{-(\varepsilon_{ni} + V_{ni} - V_{nj})}}.$$

Así,

$$\mathbb{P}_{ni} = \int_{\mathbb{R}} \left( \prod_{j \neq i} e^{-e^{-(\varepsilon_{ni} + V_{ni} - V_{nj})}} \right) e^{-\varepsilon_{ni}} e^{-e^{-\varepsilon_{ni}}} d\varepsilon_{ni}.$$

Como  $e^{-e^{-\varepsilon_{ni}}} = e^{-e^{-(\varepsilon_{ni} + V_{ni} - V_{ni})}}$

$$\mathbb{P}_{ni} = \int_{\mathbb{R}} \left( \prod_j e^{-e^{-(\varepsilon_{ni} + V_{ni} - V_{nj})}} \right) e^{-\varepsilon_{ni}} d\varepsilon_{ni}.$$

Luego,  $\prod_j e^{-e^{-(\varepsilon_{ni} + V_{ni} - V_{nj})}} = \exp \left\{ - \sum_j e^{-(V_{ni} - V_{nj} + \varepsilon_{ni})} \right\}$

$$\mathbb{P}_{ni} = \int_{\mathbb{R}} \exp \left( - \exp(\varepsilon_{ni}) \sum_j e^{-(V_{ni} - V_{nj})} \right) e^{-\varepsilon_{ni}} d\varepsilon_{ni}.$$

Sea  $t = -e^{-\varepsilon_{ni}}$ ,  $dt = e^{-\varepsilon_{ni}} d\varepsilon_{ni}$ . Así,

$$\mathbb{P}_{ni} = \int_{-\infty}^0 \exp \left( t \sum_j e^{-(V_{ni} - V_{nj})} \right) dt.$$

De ahí la situación es directa:

$$\begin{aligned} \mathbb{P}_{ni} &= \frac{\exp \left( t \sum_j e^{-(V_{ni} - V_{nj})} \right) \Big|_{-\infty}^0}{\sum_{j=1}^J e^{V_{nj} - V_{ni}}} \\ &= \frac{1}{\sum_{j=1}^J e^{V_{nj} - V_{ni}}} \\ &= \frac{e^{V_{ni}}}{\sum_{j=1}^J e^{V_{nj}}}. \end{aligned}$$

# Marriage market y estimación de beneficios y productividad laboral

Ahora se considera el siguiente problema de maximización

$$\max_{\pi_{xy} \in \Pi(\mu, \nu)} \left\{ \sum_{x,y} \pi_{xy} \Phi(x, y) \right\}.$$

Ahora  $\Phi(x, y)$  es una función de beneficios. Esta función  $\Phi$  depende de la información almacenada en  $x, y \in \mathbb{R}^L$ . Nuevamente nos situamos en el caso finito. Ahora bien, el problema de transporte óptimo puede ser interpretado como un problema de elección discreta en el contexto del Marriage Market:

$$\text{Utilidad del hombre : } U(x, y) + \frac{\sigma}{2} \varepsilon_m(y)$$

$$\text{Utilidad de la mujer : } V(x, y) + \frac{\sigma}{2} \eta_w(x).$$

Esta situación:

- 1 Conlleva a

$$\pi(y|x) = \frac{\exp \left[ \frac{U(x,y)}{\sigma/2} \right]}{\int_{\mathcal{Y}} \exp \left[ \frac{U(x,y')}{\sigma/2} \right] dy'}$$
$$\pi(x|y) = \frac{\exp \left[ \frac{U(x,y)}{\sigma/2} \right]}{\int_{\mathcal{X}} \exp \left[ \frac{U(x',y)}{\sigma/2} \right] dx'}.$$

- 2 El problema de origen de un problema de regularización entrópica.



## Beneficios laborales y productividad

Se define el salario  $w(x, y)$  del trabajador de tipo  $x$  en la firma  $y$ . El trabajador no valora únicamente su salario, pero también unas *comodidades* adicionales que se descomponen en  $\alpha(x, y)$ , un componente sistemático y  $\varepsilon(y)$  un componente estocástico.

Concretamente, el trabajador valora

$$\alpha(x, y) + w(x, y) + \sigma_1 \varepsilon(y).$$

Por el lado de la firma,

$$\gamma(x, y) - w(x, y) + \sigma_2 \eta(x),$$

donde  $\gamma$  es una medida de productividad. Análogo al caso del mercado matrimonial, se tiene el siguiente resultado:

$$\begin{aligned}\pi(y|x) &= \exp\left(\frac{\alpha(x, y) + w(x, y) - u(x)}{\sigma_1}\right) \\ \pi(x|y) &= \exp\left(\frac{\gamma(x, y) - w(x, y) - v(y)}{\sigma_2}\right),\end{aligned}$$

donde

$$\begin{aligned}u(x) &= \sigma_1 \ln \int_{\mathcal{Y}} \exp\left(\frac{\alpha(x, y') + w(x, y')}{\sigma_1}\right) dy' \\ v(y) &= \sigma_2 \ln \int_{\mathcal{X}} \exp\left(\frac{\gamma(x', y) - w(x', y)}{\sigma_2}\right) dx' .\end{aligned}$$

Gracias



Birkhoff, G. (1957).  
Extensions of jentzsch theorem.  
[American Mathematical Society](#), 85(1):219–227.



Dupuy, A., Carlier, G., Galichon, A., and Sun, Y. (2021).  
Sista learning optimal transport costos under sparsity constraints.  
[Discussion Paper Series](#).



Echenique, F. and Chambers, C. (2016).  
[Revealed Preference Theory](#).  
Cambridge University Press.



Galichon, A. and Salanié, B. (2022).  
Cupid's invisible hand: Social surplus and identification in matching models.  
[The Review of Economic Studies](#), 89(5):2600–2629.



Luigi Ambrosio, E. B. and Semola, D. (2021).  
[Lectures on Optimal Transport](#).  
Springer Verlag.




Luneberger, D. and Ye, Y. (2021).  
[Linear and Nonlinear Programming](#).  
Springer Verlag.



McFadden, D. (1974).  
Conditional logit analysis of qualitative choice behavior.

pages 105–142.

 Merigot, Q. and Thibert, B. (2020).  
Optimal transport, discretization and algorithms.

 Nenna, L. (2020).  
Lecture 4 entropic optimal transport and numerics.

 Villani, C. (2009).  
Optimal Transport Old and New.  
Springer Verlag.