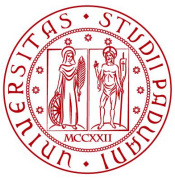

Università degli Studi di Padova

Università degli Studi di Padova

Dipartimento di Matematica "Tullio Levi-Civita"
Corso di Laurea Triennale in Matematica

Distanza di Wasserstein Adattata Proprietà e un'Applicazione

Relatore:
Prof. Markus Fischer

Laureando: Lorenzo Vigolo
Matricola: 2000417

Anno Accademico 2022/2023

22 Settembre 2023

Ringraziamenti

Vorrei in primo luogo ringraziare il mio relatore, il prof. Markus Fisher, inanzitutto per avermi proposto un argomento interessante e stimolante come quello presentato in questa tesi, per la grande disponibilità e pazienza che ha mostrato nei mei confronti e in generale per tutto l'aiuto fondamentale che mi ha dato durante tutta la stesura dell'elaborato. Ringrazio poi la mia famiglia per il supporto che mi ha dato durante questi tre anni e infine un grazie speciale a tutti gli amici, sia vecchi che nuovi, per avermi accompagnato durante questo percorso importante e aver sempre creduto in me.

Abstract

In questa tesi si studia la distanza di Wasserstein adattata, si tratta di una variante della usuale distanza di Wasserstein tra misure di probabilità definita con lo scopo di tener conto del flusso di informazione codificato negli spazi di probabilità filtrati e più in particolare nei processi stocastici. Ne diamo la definizione e ne studiamo le principali proprietà, in particolare evidenziamo come esse si differenziano da quelle della distanza classica. Infine forniamo una applicazione allo studio della sensitività di una classe di problemi di ottimizzazione stocastica.

Indice

Introduzione

In questa tesi viene trattata la distanza di Wasserstein adattata una distanza tra misure di probabilità, la definizione che si usa nell'elaborato è quella data da George Ch. Pflug e Alois Pichler nel libro *Multistage Stochastic Optimization* del 2014. Essa è definita sulla base della usuale distanza di Wasserstein tra misure di probabilità su spazi Polacchi. Ricordiamo infatti che dato uno spazio polacco (X, d), pensato dotato della sua σ-algebra dei Boreliani, è possibile, date due misure di probabilità $\mu, \nu \in \mathcal{P}(X)$ definire la distanza tra esse come

$$\mathcal{W}_r(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \iint_{X\times X} \mathrm{d}(x,y)^r \, d\pi(x,y) \right)^{1/r},$$

ove $r \geq 1$ è un parametro e $\Pi(\mu,\nu)$ è l'insieme delle misure di probabilità su $X \times X$ con marginali μ e ν rispettivamente. Tuttavia si vede, anche tramite esempi molto semplici, che, nel caso in cui gli spazi di probabilità siano dotati di filtrazioni, ovvero nel caso in cui si consideri anche l'informazione disponibile ad ogni istante, tale definizione di distanza non è in grado di captare la differenza nei flussi di informazione. Si contruiscono infatti esempi di processi molto simili come valori assunti, ma molto diversi dal punto di vista dell'informazione che racchiudono. Ciò nonostante in tali casi la distanza $\mathcal{W}_r$ è arbitrariamente piccola e quindi non è adatta per differenziare i processi sotto quel punto di vista. Per ovviare al problema si definisce una nuova metrica, ancora come il valore ottimo di un problema di ottimizzazione, simile a quello che definisce $\mathcal{W}_r$. In particolare $\mathcal{AW}_r(\mu,\nu)^r$ sarà l'estremo inferiore dello stesso funzionale considerato sopra, fatto però sull'insieme delle misure di probabilità π su $X \times X$ che soddisfano le condizioni

$$\pi(A \times X \mid \mathcal{M}_t \otimes \mathcal{N}_t) = \mu(A \mid \mathcal{M}_t)$$
$$\pi(X \times B \mid \mathcal{M}_t \otimes \mathcal{N}_t) = \nu(B \mid \mathcal{N}_t)$$

per ogni $0 \leq t \leq T$, con $T \in \mathbb{N}$. Dove $(\mathcal{M}_t)_{t=0,1,\dots,T}$ e $(\mathcal{N}_t)_{t=0,1,\dots,T}$ sono le due filtrazioni su X che modellizzano il flusso di informazione relativo alle due misure. Calcolare questa distanza sugli stessi esempi che mostrano come mai $\mathcal{W}_r$ fallisce, mostra che invece $\mathcal{AW}_r$ con questa definizione è in grado di cogliere la differenza in informazione, e quindi processi che sono arbitrariamente vicini dal punto di vista dei valori assunti, ma diversi dal punto di vista dell'informazione, non sono arbitrariamente vicini nella metrica $\mathcal{AW}_r$, come voluto.

Questa tesi consta di tre capitoli, il primo del quale è dedicato allo studio delle principali proprietà della distanza di Wasserstein usuale $\mathcal{W}_r$, vengono trattate alcune proprietà classiche della metrica $\mathcal{W}_r$, ad esempio il suo legame con il problema di trasporto ottimo di Kantorovich, il teorema di Kantorovich-Rubenstein, il legame con la convergenza debole delle misure e la completezza di tale metrica.

Nel secondo capitolo si introduce la distanza adattata e si spiega perchè quella classica fallisce come metrica in informazione, si mostrano inoltre alcune proprietà della distanza $\mathcal{AW}_r$ ad esempio il fatto che la nuova metrica definisce uno spazio non completo. Si spiega anche come estendere la nozione di misura di probabilità allo scopo di identificare il completamento di tale spazio. Per entrambi questi capitoli l'ultima sezione riguarderà cosa accade nel caso in cui si considerino misure discrete a supporto finito, ovvero le uniche per cui è possibile una implementazione numerica. Si mostra in particolare il legame tra i problemi di ottimizzazione che definiscono le distanze suddette e la programmazione lineare. Nel caso della distanza adattata si mostra anche il modo di modellizzare il flusso di informazione contenuto in una filtrazione su uno spazio finito e il legame con gli alberi finiti, ovvero dei grafi orientati aciclici con una sola radice. Nell'appendice B si forniscono per queste casistiche delle implementazioni numeriche, con MATLAB, per il calcolo di queste distanze.

Infine nell'ultimo capitolo vediamo come la distanza $\mathcal{AW}_r$ può essere utilizzata per studiare la sensitività di una classe di problemi di ottimizzazione stocastica multiperiodale. Questo è possibile perchè si può mostrare che la topologia indotta dalla distanza adattata è la più grossolana topologia rispetto a cui i problemi di ottimizzazione stocastica multiperiodali sono continui. In questa tesi trattiamo il problema di minimizzazione del valore atteso di una funzione convessa, $f\colon \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}$, sotto controlli predicibili, ossia vettori $a = (a_1, \dots, a_T)$ tali che $a_t\colon \mathbb{R}^T \to \mathbb{R}$ dipende solamente da $x_1, \dots, x_{t-1}$. Indichiamo l'insieme dei controlli predicibili con $\mathcal{A}$. Siamo quindi interessati al problema

$$v(\mathbb{P}) := \inf_{a \in \mathcal{A}} \mathbb{E}^{\mathbb{P}}\big[f(\xi, a(\xi))\big],$$

ove $\mathbb{P} \in \mathcal{P}(\mathbb{R}^T)$ e $\xi\colon \mathbb{R}^T \to \mathbb{R}^T$ è il processo canonico. Si noti che il problema di massimizzazione dell'utilità attesa, uno dei problemi fondamentali della finanza matematica, può essere scritto in tale forma e verrà trattato come un caso speciale. In particolare il risultato centrale che viene provato in questo ultimo capitolo asserisce che, fissato un modello con misura $\mathbb{P}$ è possibile ottenere uno sviluppo al primo ordine dell'errore che si commette nel valore ottimo del problema quando si usa $\mathbb{Q}$, anzichè $\mathbb{P}$, ove $\mathbb{Q}$ varia in un intorno di $\mathbb{P}$ rispetto $\mathcal{AW}_r$. Più precisamente definito, per ogni $\delta > 0$, l'insieme

$$B_\delta(\mathbb{P}) := \{\mathbb{Q} \in \mathcal{P}(\mathbb{R}^T) \mid \mathcal{AW}_r(\mathbb{P}, \mathbb{Q}) \leq \delta\},$$

si prova che, per $\delta \to 0$

$$\sup_{\mathbb{Q} \in B_\delta(\mathbb{P})} v(\mathbb{Q}) = v(\mathbb{P}) + C\delta + o(\delta),$$

con $C > 0$ una costante. In realtà viene data un'espressione esplicita alla costante C, che nel caso del problema di massimizzazione dell'utilità attesa può essere interpretata come la variazione quadratica attesa dell'ottimizzante al problema ottenuto con la misura $\mathbb{P}$ ma distorta da il valore atteso condizionato della derivata prima di $\ell = -U$ dove U è la funzione di utilità.

Capitolo 1

La Distanza di Wasserstein

1.1 Problema e Dualità di Kantorovich

In questo capitolo definiamo la distanza di Wasserstein tra misure di probabilità che nel prossimo capitolo verrà adattata ad una *metrica in informazione* tra processi stocatistici. Vedremo che essa è definita come un caso particolare di un problema di ottimizzazione più generale il *problema di Kantorovich*, in questa sezione formuliamo quindi tale problema e ne dimostriamo un teorema di dualità. Per la maggiorparte per la stesura di questo capitolo abbiamo utilizzato come referenza i capitoli 1 e 7 di [12] ad esclusione del teorema 1.12 che è il risultato principale presentato in [4].

Definizione 1.1. Siano $(X, \mathcal{M}, \mu)$ e $(Y, \mathcal{N}, \nu)$ due spazi di probabilità, definiamo allora l'insieme:

$$\Pi(\mu,\nu) := \big\{\pi \in \mathcal{P}(X\times Y) \mid \pi(A\times Y) = \mu(A), \pi(X\times B) = \nu(B) \quad \text{per ogni} A \in \mathcal{M}, B \in \mathcal{N}\big\}.$$

Gli elementi di $\Pi(\mu,\nu)$ sono detti *couplings* tra μ e ν.

Lemma 1.1. *Siano $(X, \mathcal{M}, \mu)$ e $(Y, \mathcal{N}, \nu)$ due spazi di probabilità, allora le seguenti affermazioni sono equivalenti:*

1. $\pi \in \Pi(\mu,\nu)$

2. *π è una misura non negativa su $X \times Y$ e per ogni funzione misurabile $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$ vale:*

$$\iint_{X\times Y} \big[\varphi(x) + \psi(y)\big]\, d\pi(x,y) = \int_X \varphi\, d\mu + \int_Y \psi\, d\nu. \tag{1.1}$$

3. *π è una misura non negativa su $X \times Y$ e per ogni funzione misurabile $(\varphi, \psi) \in L^\infty(\mu) \times L^\infty(\nu)$ vale:*

$$\iint_{X\times Y} \big[\varphi(x) + \psi(y)\big]\, d\pi(x,y) = \int_X \varphi\, d\mu + \int_Y \psi\, d\nu. \tag{1.2}$$

Dimostrazione. Dimostriamo prima l'equivalenza tra 1. e 2., supponiamo π sia una misura non negativa su $X \times Y$ per cui valga (1.1). Osserviamo che allora π è necessariamente una misura di probabilità, infatti siccome $(1,0) \in L^1(\mu) \times L^1(\nu)$:

$$\pi(X \times Y) = \iint_{X\times Y} (1+0)\, d\pi = \int_X d\mu = 1,$$

inoltre fissati $A \in \mathcal{M}, B \in \mathcal{N}$ possiamo scegliere le coppie: $(\mathbf{1}_A, 0) \in L^1(\mu) \times L^1(\nu)$ e $(0, \mathbf{1}_B) \in L^1(\mu) \times L^1(\nu)$. Applicando la proprietà (1.1) alla prima otteniamo:

$$\pi(A \times Y) = \iint_{X\times Y} \mathbf{1}_{A\times Y}\, d\pi = \iint_{X\times Y} (\mathbf{1}_A(x) + 0)\, d\pi(x,y) = \int_X \mathbf{1}_A\, d\mu = \mu(A),$$

mentre con la seconda coppia otteniamo:

$$\pi(X\times B) = \iint_{X\times Y} \mathbf{1}_{X\times B}(x,y)\, d\pi(x,y) = \iint_{X\times Y} (0+\mathbf{1}_B(y))\, d\pi(x,y) = \int_Y \mathbf{1}_B\, d\nu = \nu(A).$$

Quindi $\pi \in \Pi(\mu,\nu)$. Supponiamo ora $\pi \in \Pi(\mu,\nu)$ allora dalle uguaglianze sopra otteniamo che la proprietà (1.1) è vera per le coppie: $(\mathbf{1}_A,0), (0,\mathbf{1}_B) \in L^1(\mu)\times L^1(\nu)$, con $A \in \mathcal{M}, B \in \mathcal{N}$. Da questo si deduce poi per linearità che la (1.1) vale per le coppie: $(f,g) \in L^1(\mu)\times L^1(\nu)$ con f e g funzioni semplici non negative, infatti date tali f e g si ha che:

$$f = \sum_{j=1}^n c_j \mathbf{1}_{A_j}, g = \sum_{k=1}^m d_k \mathbf{1}_{B_k} \quad \text{per certi} \quad c_j, d_k \geq 0, A_j \in \mathcal{M}, B_k \in \mathcal{N},$$

allora:

$$\begin{aligned}
\iint_{X\times Y} \big(f(x)+g(y)\big)\, d\pi(x,y) &= \iint_{X\times Y} f(x)\, d\pi(x,y) + \iint_{X\times Y} g(y)\, d\pi(x,y) \\
&= \iint_{X\times Y} \sum_{j=1}^n c_j \mathbf{1}_{A_j}(x)\, d\pi(x,y) + \iint_{X\times Y} \sum_{k=1}^m d_k \mathbf{1}_{B_k}(y)\, d\pi(x,y) \\
&= \sum_{j=1}^n c_j \int_X \mathbf{1}_{A_j}\, d\mu + \sum_{k=1}^m d_k \int_Y \mathbf{1}_{B_k}\, d\nu \\
&= \sum_{j=1}^n c_j \mu(A_j) + \sum_{k=1}^m d_k \nu(B_k) = \int_X f\, d\mu + \int_Y g\, d\nu.
\end{aligned}$$

Presa poi una coppia: $(\varphi,\psi) \in L^1(\mu)\times L^1(\nu)$ di funzioni integrabili non negative è noto che esistono due successioni *crescenti* di funzioni semplici non negative: $(f_n)_{n\in\mathbb{N}} \subseteq L^1(\mu)$ e $(g_n)_{n\in\mathbb{N}} \subseteq L^1(\nu)$ tali che: $f_n \xrightarrow[n\to\infty]{} \varphi$ e $g_n \xrightarrow[n\to\infty]{} \psi$ puntualmente. Allora $f_n(x) + g_n(y) \xrightarrow[n\to\infty]{} \varphi(x) + \psi(y)$ crescendo, per ogni $(x,y) \in X \times Y$. Segue

allora dal teorema di convergenza monotona che:

$$\begin{aligned}\iint_{X\times Y}\big(\varphi(x)+\psi(y)\big)\,d\pi(x,y) &= \iint_{X\times Y}\lim_{n\to\infty}\big(f_n(x)+g_n(y)\big)\,d\pi(x,y)\\ &= \lim_{n\to\infty}\iint_{X\times Y}\big(f_n(x)+g_n(y)\big)\,d\pi(x,y)\\ &= \lim_{n\to\infty}\int_X f_n\,d\mu+\lim_{n\to\infty}\int_Y g_n\,d\nu\\ &= \int_X\lim_{n\to\infty} f_n\,d\mu+\int_Y\lim_{n\to\infty} g_n\,d\nu\\ &= \int_X\varphi\,d\mu+\int_Y\psi\,d\nu.\end{aligned}$$

Infine per $(\varphi,\psi)\in L^1(\mu)\times L^1(\nu)$ generiche è sufficiente passare alle parti positiva e negativa:

$$\begin{aligned}&\iint_{X\times Y}\big(\varphi(x)+\psi(y)\big)\,d\pi(x,y)\\ &= \iint_{X\times Y}\big(\varphi^+(x)+\psi^+(y)\big)\,d\pi(x,y)-\iint_{X\times Y}\big(\varphi^-(x)+\psi^-(y)\big)\,d\pi(x,y)\\ &= \int_X\varphi^+\,d\mu+\int_Y\psi^+\,d\nu-\int_X\varphi^-\,d\mu-\int_Y\psi^-\,d\nu\\ &= \int_X\varphi\,d\mu+\int_Y\psi\,d\nu.\end{aligned}$$

Proviamo ora l'equivalenza tra 2. e 3., il fatto che 2. implichi 3. segue subito in quanto poichè μ e ν sono misure di probabilità sono, in particolare, misure finite, quindi: $L^\infty(\mu)\subseteq L^1(\mu)$ e $L^\infty(\nu)\subseteq L^1(\nu)$. Per mostrare che 3. implica 2. è sufficiente osservare che 3. implica 1. e quindi 2. per i calcoli sopra. Infatti se vale 3. come sopra: $\pi\in\mathcal{P}(X\times Y)$ e, per ogni $A\in\mathcal{M}, B\in\mathcal{N}$, si ha: $\mathbf{1}_A\in L^\infty(\mu)$ e $\mathbf{1}_B\in L^\infty(\nu)$, dunque:

$$\pi(A\times Y)=\iint_{X\times Y}\mathbf{1}_{A\times Y}\,d\pi=\iint_{X\times Y}\mathbf{1}_A(x)\,d\pi(x,y)=\int_X\mathbf{1}_A\,d\mu=\mu(A),$$

e analogamente:

$$\pi(X\times B)=\iint_{X\times Y}\mathbf{1}_{X\times B}\,d\pi=\iint_{X\times Y}\mathbf{1}_B(y)\,d\pi(x,y)=\int_Y\mathbf{1}_B\,d\nu=\nu(B).\quad\square$$

Definizione 1.2 (Problema di Kantorovich)**.** Siano $(X,\mathcal{M},\mu)$ e $(Y,\mathcal{N},\nu)$ due spazi di probabilità e $c\colon X\times Y\to[0,+\infty]$ una funzione misurabile rispetto alla σ-algebra $\mathcal{M}\otimes\mathcal{N}$. Definiamo il funzionale:

$$I\colon\mathcal{P}(X\times Y)\to[0,+\infty]\quad\text{tramite:}\quad\pi\mapsto I[\pi]:=\iint_{X\times Y}c\,d\pi.$$

Si dice allora *problema di trasporto ottimo di Kantorovich* il problema di minimo:

$$\inf_{\pi\in\Pi(\mu,\nu)} I[\pi].$$

Ne indichiamo il valore ottimo con: $\mathcal{I}_c(\mu,\nu)$.

Osservazione 1.1. Usando la notazione probabilistica il problema di Kantorovich può essere scritto come

$$\inf_{U,V} \mathbb{E}^{\mathbb{P}}\big[c(U,V)\big],$$

ove l'estremo inferiore è fatto su tutte le possibili v.a. U, V definite su uno spazio di probabilità $(\Omega,\mathcal{F},\mathbb{P})$ e a valori rispettivamente in $(X,\mathcal{M})$ e $(Y,\mathcal{N})$ tali che $U\sim\mu, V\sim\nu$. Da questa osservazione otteniamo la conseguenza importante che date le misure μ e ν se riusciamo a costruire delle v.a. U, V con leggi μ e ν abbiamo la validità della stima fondamentale

$$\mathcal{I}_c(\mu,\nu)\leq\mathbb{E}^{\mathbb{P}}\big[c(U,V)\big].$$

Prima di enunciare il teorema di dualità ricordiamo la definizione e una proprietà importante delle funzioni *semicontinue inferiormente*.

Definizione 1.3. Sia (E,d) uno spazio metrico e $F\colon E\to\mathbb{R}$ una funzione, F si dice essere *semicontinua inferiormente* se vale che: per ogni successione $(x_n)_{n\in\mathbb{N}}\subseteq E$ convergente ad un punto $x\in E$ si ha:

$$F(x)\leq\liminf_{n\to\infty} F(x_n).$$

Proposizione 1.2. *Sia (E,d) uno spazio metrico e $F\colon E\to\mathbb{R}$ una funzione semicontinua inferiormente e tale che $F\geq 0$. Allora esiste una successione crescente $(\varphi_n)_{n\in\mathbb{N}}$ di funzioni definite su E a valori in $\mathbb{R}$, uniformemente continue e tali che*

$$\lim_{n\to\infty}\varphi_n(x)=F(x)\quad\forall x\in E.$$

Dimostrazione. È sufficiente porre, per ogni $n\in\mathbb{N}$

$$\varphi_n(x):=\inf_{z\in E}\{F(z)+n\mathrm{d}(x,z)\},$$

e verificare che tale successione verifica le proprietà volute. Infatti ovviamente

$$0\leq\varphi_n\leq\varphi_{n+1},$$

inoltre fissato $(x,y)\in E\times E$ ed $n\in\mathbb{N}$ abbiamo

$$F(z)+n\mathrm{d}(x,z)\leq F(z)+n\mathrm{d}(y,z)+n\mathrm{d}(x,y),$$

per ogni $z\in E$. Quindi passando all'estremo inferiore otteniamo

$$\varphi_n(x)\leq\varphi_n(y)+n\mathrm{d}(x,y).$$

Invertendo i rouli di x e y troviamo allo stesso modo

$$\varphi_n(y) \leq \varphi_n(x) + n\mathrm{d}(x,y),$$

e quindi

$$|\varphi_n(x) - \varphi_n(y)| \leq n\mathrm{d}(x,y).$$

Disuguaglianza che implica subito l'uniforme continuità (Lipschitzianità). Per provare che F ne è il limite puntuale, fissiamo $z \in E$ e $\varepsilon > 0$. Siccome F è semicontinua inferiormente, abbiamo che esiste $\delta > 0$ tale che per ogni $x \in E$ tale che $\mathrm{d}(x,z) < \delta$ si ha

$$F(z) - \frac{\varepsilon}{2} < F(x).$$

Sia $\overline{n} \in \mathbb{N}$ tale che

$$\frac{F(z) - \varepsilon/2}{\delta} < \overline{n}$$

e sia $n \geq \overline{n}$. Allora per ogni $x \in E$ se $\mathrm{d}(x,z) \geq \delta$, risulta

$$F(z) - \frac{\varepsilon}{2} < n\delta \leq F(x) + n\delta \leq F(x) + n\mathrm{d}(x,z).$$

Se invece $\mathrm{d}(x,y) < \delta$, risulta

$$F(z) - \frac{\varepsilon}{2} < F(x) \leq F(x) + n\mathrm{d}(x,z).$$

In ogni caso vale

$$F(z) - \frac{\varepsilon}{2} < F(x) \leq F(x) + n\mathrm{d}(x,z),$$

da cui per arbitrarietà di x segue

$$F(z) - \varepsilon < \varphi_n(z),$$

ed ora per arbitrarietà di $\varepsilon > 0$ otteniamo

$$F(z) = \sup_{n\in\mathbb{N}} \varphi_n(z) = \lim_{n\to\infty} \varphi_n(z)$$

e si conclude. □

Teorema 1.3 (Dualità di Kantorovich)**.** *Siano X e Y due spazi Polacchi e $\mu \in \mathcal{P}(X), \nu \in \mathcal{P}(Y)$ e $c\colon X \times Y \to [0, +\infty]$ una funzione misurabile e semicontinua inferiormente. Allora per ogni $\pi \in \mathcal{P}(X \times Y)$ e $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$ definiamo rispettivamente i due funzionali:*

$$I[\pi] := \iint_{X\times Y} c(x,y)\, d\pi(x,y), \qquad J[\varphi,\psi] := \int_X \varphi\, d\mu + \int_Y \psi\, d\nu.$$

Definiamo l'insieme:

$$\Phi_c := \big\{(\varphi,\psi) \in L^1(\mu)\times L^1(\nu) \mid \varphi(x)+\psi(y) \leq c(x,y) \quad \textit{per } \mu\textit{-q.o. } x \in X \textit{ e } \nu\textit{-q.o. } y \in Y\big\}.$$

Allora:

$$\inf_{\pi\in\Pi(\mu,\nu)} I[\pi] = \sup_{(\varphi,\psi)\in\Phi_c} J[\varphi,\psi]. \tag{1.3}$$

Inoltre, l'estremo inferiore nel membro di sinistra di (1.3) *è raggiunto, in più il valore dell'estremo superiore a destra in* (1.3) *non cambia se si restringe la definizione di* Φ_c *alle funzioni* $(\varphi,\psi)\in\mathcal{C}_b(X)\times\mathcal{C}_b(Y)$.

Osservazione 1.2. Quando sarà necessario distinguere i casi per la definzione dell'insieme Φ_c useremo le notazioni: $\Phi_c\cap L^1$ e $\Phi_c\cap\mathcal{C}_b$ rispettivamente. È utile inoltre prima di effettuare la dimostrazione completa del teorema, osservare che una delle disuguaglianze, che danno l'uguaglianza (1.3) è relativamente semplice.

Proposizione 1.4. *Sotto le stesse ipotesi e con le stesse notazioni del teorema 1.3 si ha:*

$$\sup_{(\varphi,\psi)\in\Phi_c\cap\mathcal{C}_b} J[\varphi,\psi] \le \sup_{(\varphi,\psi)\in\Phi_c\cap L^1} J[\varphi,\psi] \le \inf_{\pi\in\Pi(\mu,\nu)} I[\pi]. \tag{1.4}$$

Dimostrazione. Poichè $\mathcal{C}_b(X)\times\mathcal{C}_b(Y)\subseteq L^1(\mu)\times L^1(\nu)$ la prima disuguaglianza in (1.4) segue subito, siccome il primo sup è fatto su un insieme più piccolo. Per provare la seconda disuguaglianza consideriamo: $(\varphi,\psi)\in\Phi_c\cap L^1, \pi\in\Pi(\mu,\nu)$ allora:

1. Esistono $N_1\in\mathcal{M}, N_2\in\mathcal{N}$ tali che: $\mu(N_1)=\nu(N_2)=0$ e per ogni $(x,y)\in N_1^c\times N_2^c$ si ha:
$$\varphi(x)+\psi(y)\le c(x,y)$$
ma allora poichè:
$$\pi\big((N_1^c\times N_2^c)^c\big)\le\pi(N_1\times Y)+\pi(X\times N_2)=\mu(N_1)+\nu(N_2)=0,$$
abbiamo che $\varphi+\psi\le c$ π-q.o.

2. Inoltre per la proprietà sulle marginali si ha:
$$J[\varphi,\psi]=\int_X\varphi\,d\mu+\int_Y\psi\,d\nu=\iint_{X\times Y}\big[\varphi(x)+\psi(y)\big]\,d\pi(x,y).$$

Mettendo insieme 1. e 2. otteniamo:

$$J[\varphi,\psi]\le\iint_{X\times Y}c(x,y)\,d\pi(x,y)$$

e ora il membro di destra dell'equazione precedente è indipendente da π, mentre quello di sinistra è indipendente da (φ,ψ). Quindi posso passare al sup a poi all'inf nella disuguaglianza precedente, la quale viene mantenuta, ottenendo esattamente la (1.4). □

Osservazione 1.3. Si noti che questa proposizione implica in particolare che per dimostrare il teorema di dualità e sufficiente dimostrare l'altra disuguaglianza solo per $\Phi_c\cap\mathcal{C}_b$.

Necessitiamo ancora di una ultima definizione e di un ultimo risultato prima di poter effettuare la prova del teorema di dualità.

Definizione 1.4 (Trasformata di Legendre-Fenchel)**.** Sia E uno spazio normato, e $\Theta\colon E \to \mathbb{R} \cup \{+\infty\}$ una funzione convessa, ossia:

$$\forall x, y \in E, \lambda \in [0,1] \quad \Theta(\lambda x + (1-\lambda)y) \leq \lambda\Theta(x) + (1-\lambda)\Theta(y)$$

con le ovvie estensioni delle operazioni di $\mathbb{R}$ a $\mathbb{R} \cup \{+\infty\}$. Si dice allora *trasformata di Legendre-Fenchel* di Θ la funzione:

$$\Theta^*\colon E^* \to \mathbb{R} \quad \text{definita tramite} \quad \Theta^*(f) := \sup_{x \in E}\big[\langle f, x\rangle - \Theta(x)\big].$$

Teorema 1.5 (Frenchel-Rockfellar)**.** *Sia E uno spazio normato, $\Theta, \Lambda\colon E \to \mathbb{R} \cup \{+\infty\}$ due funzioni convesse tali che esiste $x_0 \in E$ per cui:*

$$\Theta(x_0), \Lambda(x_0) < +\infty \quad \textit{e } \Theta \textit{ è continua in } x_0.$$

Siano poi Θ^*, Λ^* le loro trasformate di Legendre-Fenchel. Allora:

$$\inf_{x \in E}\big[\Theta(x) + \Lambda(x)\big] = \max_{f \in E^*}\big[-\Theta^*(-f) - \Lambda^*(f)\big]. \tag{1.5}$$

Dimostrazione. Notiamo che dalla definizione di trasformata di Legendre-Fenchel segue che:

$$\begin{aligned}
\sup_{f \in E^*}\big[-\Theta^*(-f) - \Lambda^*(f)\big] &= \sup_{f \in E^*}\Big[-\sup_{x \in E}\big[\langle -f, x\rangle - \Theta(x)\big] - \sup_{y \in E}\big[\langle f, y\rangle - \Lambda(y)\big]\Big] \\
&= \sup_{f \in E^*}\Big[\inf_{x \in E}\big[\Theta(x) + \langle f, x\rangle\big] + \inf_{y \in E}\big[\Lambda(y) - \langle f, y\rangle\big]\Big] \\
&= \sup_{f \in E^*}\inf_{x,y \in E}\big[\Theta(x) + \Lambda(y) + \langle f, x - y\rangle\big],
\end{aligned}$$

quindi l'equazione (1.5) è equivalente a:

$$\sup_{f \in E^*}\inf_{x,y \in E}\big[\Theta(x) + \Lambda(y) + \langle f, x - y\rangle\big] = \inf_{x \in E}\big[\Theta(x) + \Lambda(x)\big].$$

Osserviamo allora che nell'estremo inferiore nel membro di sinistra è possibile prendere $x = y$ si ha allora che per ogni forma lineare $f \in E^*$:

$$\inf_{x,y \in E}\big[\Theta(x) + \Lambda(y) + \langle f, x - y\rangle\big] \leq \inf_{x \in E}\big[\Theta(x) + \Lambda(x)\big]$$

poichè l'inf è fatto su un insieme più grande di numeri reali. Quindi passando al sup in f si ottiene:

$$\sup_{f \in E^*}\inf_{x,y \in E}\big[\Theta(x) + \Lambda(y) + \langle f, x - y\rangle\big] \leq \inf_{x \in E}\big[\Theta(x) + \Lambda(x)\big].$$

È quindi sufficiente dimostrare che esiste una forma lineare $f \in E^*$ tale che:

$$\forall x, y \in E \quad \Theta(x) + \Lambda(y) + \langle f, x - y \rangle \geq m := \inf_{x \in E} \big[\Theta(x) + \Lambda(x)\big].$$

Siccome per ipotesi $\Theta(x_0) + \Lambda(x_0) < +\infty$ l'estremo inferiore m è finito. Definiamo gli insiemi:

$$C := \{(x, \theta) \in E \times \mathbb{R} \mid \Theta(x) < \theta\}$$
$$C' := \{(y, \lambda) \in E \times \mathbb{R} \mid \lambda \leq m - \Lambda(y)\}$$

siccome Θ e Λ sono funzioni convesse allora C e C' sono insiemi convessi, in quanto ne sono i rispettivi epigrafici. Dall'ipotesi che Θ sia continua in x_0 segue che: $(x_0, \Theta(x_0) + 1) \in \mathrm{Int}(C)$, in particolare quindi C ha interno non vuoto e di conseguenza: $\overline{C} = \overline{\mathrm{Int}(C)}$. Notiamo poi che i due insiemi sono necessariamente disgiunti, se infatti esistesse $(x, \eta) \in E \times \mathbb{R}$ tale che:

$$\eta > \Theta(x), \eta + \Lambda(x) \leq m \quad \text{otterremmo che:} \quad \Theta(x) + \Lambda(x) < \eta + \Lambda(x) \leq m$$

che è assurdo per definizione di m. Segue allora dal teorema di Hahn-Banach in forma geometrica che esistono $f \in E^*, \alpha \in \mathbb{R}$ non nulli tali che, per ogni $(x, \theta) \in C, (y, \lambda) \in C'$ si abbia:

$$\langle f, x \rangle + \alpha\theta \geq \langle f, y \rangle + \alpha\lambda.$$

Proviamo che ciò è possibile solamente se $\alpha > 0$, infatti se fosse: $\alpha < 0$ si avrebbe, per ogni $(x, \theta) \in C, (y, \lambda) \in C'$:

$$\langle f, x \rangle + \alpha\Theta(x) > \langle f, x \rangle + \alpha\theta \geq \langle f, y \rangle + \alpha\lambda \geq \langle f, y \rangle + \alpha(m - \Lambda(y)),$$

da cui:

$$\langle f, x - y \rangle + \alpha(\Theta(x) + \Lambda(y)) > \alpha m$$

Scegliendo ora due coppie in C e C' tali che $x = y$ otteniamo dividendo l'equazione sopra per α:

$$\Theta(x) + \Lambda(x) > m$$

che è assurdo per definizione di m (è l'inf di tali quantità). Defininendo allora $\hat{f} := f/\alpha$ abbiamo:

$$\langle \hat{f}, x \rangle + \theta \geq \langle \hat{f}, y \rangle + \lambda,$$

scegliendo allora per $\varepsilon > 0$ le coppie: $(x, \Theta(x) + \varepsilon) \in C, (y, m - \Lambda(y) - \varepsilon) \in C'$ otteniamo:

$$\langle \hat{f}, x \rangle + \Theta(x) \geq \langle \hat{f}, y \rangle + m - \Lambda(y) - 2\varepsilon.$$

Per arbitrarietà di $\varepsilon > 0$ otteniamo (passando al limite $\varepsilon \to 0^+$) che:

$$\langle \hat{f}, x \rangle + \Theta(x) \geq \langle \hat{f}, y \rangle + m - \Lambda(y)$$

per ogni $x, y \in E$. Riordinando otteniamo la tesi. □

Dimostrazione teorema 1.3. Prima di iniziare con la effettiva dimostrazione del teorema 1.3, ricordiamo alcune proprietà delle misure su spazi polacchi:

1. Una misura di Borel μ su uno spazio Polacco X è *regolare*, quindi in particolare è di *Radon*.
2. Le misure di Borel finite (quindi, in particolare, le misure di probabilità) su uno spazio Polacco sono *serrate* (risultato A.1).
3. Dal teorema di Prohorov A.3, famiglie *uniformemente serrate* di misure di probabilità sono *debolmente relativamente sequenzialmente compatte*, ovvero da ogni successione in tali famiglie è possibile estrarre una sottosuccessione convergente debolmente.

Cominciamo ora con l'effetiva dimostrazione del teorema di Dualità. Dividiamo la dimostrazione in tre passi, aumentando il livello di generalità.

Assumiamo inizialmente che X, Y siano spazi *compatti* e che la funzione c sia *continua* su $X \times Y$. Consideriamo $\mathcal{C}_b(X \times Y)$ dotato della norma della convergenza uniforme: $\|\cdot\|_\infty$. Dal teorema di Riesz, il suo duale topologico può essere identificato con lo spazio delle misure di Radon con segno, ossia $(\mathcal{C}_b(X \times Y))^* = \mathcal{M}(X \times Y)$, ma di più una forma lineare non negativa su $\mathcal{C}_b(X \times Y)$ è identificata da una misura non negativa in $\mathcal{M}(X \times Y)$. Introduciamo poi le seguenti funzioni:

$$\Theta\colon \mathcal{C}_b(X \times Y) \to \mathbb{R} \cup \{+\infty\}, u \mapsto \begin{cases} 0 & \text{se } u(x,y) \geq -c(x,y) \text{ per ogni } \mu\text{-q.o. } x \text{ e } \nu\text{-q.o. } y, \\ +\infty & \text{altrimenti,} \end{cases}$$

$$\Lambda\colon \mathcal{C}_b(X \times Y) \to \mathbb{R} \cup \{+\infty\}, u \mapsto \begin{cases} \displaystyle\int_X \varphi \, d\mu + \int_Y \psi \, d\nu & \begin{array}{c}\text{se } u(x,y) = \varphi(x) + \psi(y) \\ \text{per ogni } \mu\text{-q.o. } x \text{ e } \nu\text{-q.o. } y,\end{array} \\ +\infty & \text{altrimenti,} \end{cases}$$

per certe funzioni $\varphi \in \mathcal{C}_b(X), \psi \in \mathcal{C}_b(Y)$. Si noti che Λ è ben definita, infatti se $\tilde{\varphi}, \tilde{\psi}$ è un'altra coppia di funzioni che verifica: $u(x,y) = \tilde{\varphi}(x) + \tilde{\psi}(y)$, questo implica che: $\varphi = \tilde{\varphi} + s, \psi = \tilde{\psi} - s$, con $s \in \mathbb{R}$, e quindi:

$$\int_X \varphi \, d\mu + \int_Y \psi \, d\nu = \int_X \tilde{\varphi} \, d\mu + \int_Y \tilde{\psi} \, d\nu.$$

Ora vediamo che le ipotesi del teorema 1.5 sono verificate per Θ e Λ, infatti esse sono convesse, per ogni $u, v \in \mathcal{C}_b(X \times Y), t \in [0,1]$ si ha che: se una tra u, v verifica: $u < -c$ o $v < -c$, allora o $\Theta(u) = +\infty$ oppure $\Theta(v) = +\infty$ da cui segue subito:

$$\Theta(tu + (1-t)v) \leq t\Theta(u) + (1-t)\Theta(v) = +\infty.$$

Se invece $u, v \geq -c$ allora: $tu + (1-t)v \geq -tc - (1-t)c = -c$, quindi la disuguaglianza diviene: $0 \leq 0$ che è vera. Analogamente per Λ, siano $u, v \in \mathcal{C}_b(X \times Y), t \in [0,1]$, allora: se $u(x,y) \neq \varphi(x) + \psi(y)$ o $v(x,y) \neq \tilde{\varphi}(x) + \tilde{\psi}(y)$ su insiemi di misura non nulla, si ha: $\Lambda(u) = +\infty$ oppure $\Lambda(v) = +\infty$, da cui segue banalmente

la disuguaglianza di convessità come sopra. Se invece: $u(x,y) = \varphi(x) + \psi(y)$ o $v(x,y) = \tilde{\varphi}(x) + \tilde{\psi}(y)$, per μ-q.o. x e ν-q.o. y e per opportune funzioni: $\varphi, \tilde{\varphi} \in \mathcal{C}_b(X), \psi, \tilde{\psi} \in \mathcal{C}_b(Y)$ allora per μ-q.o. x e ν-q.o. y:

$$\begin{aligned} tu(x,y) + (1-t)v(x,y) &= t\varphi(x) + t\psi(y) + (1-t)\tilde{\varphi}(x) + (1-t)\tilde{\psi} \\ &= t\varphi(x) + (1-t)\tilde{\varphi}(x) + t\psi(y) + (1-t)\tilde{\psi}(y) \end{aligned}$$

ed ora $t\varphi(x) + (1-t)\tilde{\varphi}(x) \in \mathcal{C}_b(X), t\psi(y) + (1-t)\tilde{\psi}(y) \in \mathcal{C}_b(Y)$, dunque:

$$\begin{aligned} \Lambda(tu + (1-t)v) &= \int_X \left[t\varphi + (1-t)\tilde{\varphi}\right] d\mu + \int_Y \left[t\psi + (1-t)\tilde{\psi}\right] d\nu \\ &= t\left(\int_X \varphi \, d\mu + \int_Y \psi \, d\nu\right) + (1-t)\left(\int_X \tilde{\varphi} \, d\mu + \int_Y \tilde{\psi} \, d\nu\right) \\ &= t\Lambda(u) + (1-t)\Lambda(v). \end{aligned}$$

Quindi Θ, Λ sono convesse, ed è inoltre ovvio che $1 \in \mathcal{C}_b(X \times Y)$ verifica che $\Theta(1), \Lambda(1) < +\infty$, e Θ è continua in 1 poichè è costante in un intorno sufficientemente piccolo di 1 in $(\mathcal{C}_b(X \times Y), \|\cdot\|_\infty)$. È allora possibile applicare la formula (1.5). Calcoliamone entrambi i membri. Quello di sinistra è relativamente facile e risulterà:

$$\begin{aligned} \inf_{u \in \mathcal{C}_b(X \times Y)} \left\{ \int_X \varphi \, d\mu + \int_Y \psi \, d\nu \mid u(x,y) = \varphi(x) + \psi(y) \geq -c(x,y) \quad \mu\text{-q.o. } x \text{ e } \nu\text{-q.o. } y \right\} \\ = - \sup_{(\varphi,\psi) \in \Phi_c \cap \mathcal{C}_b} J[\varphi, \psi] \end{aligned}$$

con il significato visto in precedenza di $\Phi_c \cap \mathcal{C}_b$ e J. Per il membro di destra ci servono le trasformate di Legendre-Fenchel di Θ e Λ, osserviamo allora che, per ogni $\pi \in \mathcal{M}(X \times Y)$ si ha:

$$\begin{aligned} \Theta^*(-\pi) &= \sup_{u \in \mathcal{C}_b(X \times Y)} \left\{ -\iint_{X \times Y} u \, d\pi \mid u \geq -c \right\} \\ &= \sup_{u \in \mathcal{C}_b(X \times Y)} \left\{ \iint_{X \times Y} u \, d\pi \mid u \leq c \right\} \end{aligned}$$

ed ora notiamo che:

- Se π non è una misura non negativa, allora certamente esiste una funzione non positiva: $v \in \mathcal{C}_b(X \times Y)$ tale che:

  $$\iint_{X \times Y} v \, d\pi > 0.$$

  Ma allora la scelta: $u = tv$, con $t \in \mathbb{R}$, mostra che, per $t \to +\infty$ l'estremo superiore è $+\infty$.

- Se invece, π è non negativa è chiaro che per convergenza monotona tale estremo superiore è:

$$\iint_{X\times Y} c\, d\pi$$

Dunque, otteniamo:

$$\Theta^*(-\pi) = \begin{cases} \iint_{X\times Y} c\, d\pi & \text{se } \pi \in \mathcal{M}_+(X\times Y), \\ +\infty & \text{altrimenti.} \end{cases}$$

Ragionando in modo simile, abbiamo, per ogni $\pi \in \mathcal{M}(X\times Y)$:

$$\begin{aligned} \Lambda^*(\pi) &= \sup_{u\in\mathcal{C}_b(X\times Y)} \left\{ \iint_{X\times Y} u\, d\pi - \Lambda(\pi) \right\} \\ &= - \inf_{u\in\mathcal{C}_b(X\times Y)} \left\{ \Lambda(u) - \iint_{X\times Y} u\, d\pi \right\} \end{aligned}$$

ora è sufficiente, per calcolare tale estremo inferiore, considereare solo le $u \in \mathcal{C}_b(X\times Y)$ tali per cui esistono $\varphi \in \mathcal{C}_b(X), \psi \in \mathcal{C}_b(Y)$ tali che: $u(x,y) = \varphi(x) + \psi(y)$ per μ-q.o. x e ν-q.o. y, se così non è si ha infatti: $\Lambda(u) = +\infty$. Adesso distinguiamo due casi per π:

- Se π è tale che, esistono $(\varphi,\psi) \in \mathcal{C}_b(X)\times\mathcal{C}_b(Y)$, per cui:

$$\iint_{X\times Y} \big[\varphi(x)+\psi(y)\big]\, d\pi(x,y) \neq \int_X \varphi\, d\mu + \int_Y \psi\, d\nu,$$

  e senza perdere di generalità supponiamo valga il $>$ (altimenti è sufficiente prendere gli opposti), allora con la scelta: $u(x,y) = t[\varphi(x)+\psi(y)]$, per $t \in \mathbb{R}_+$, abbiamo per $t \to +\infty$, che:

$$t\left(\int_X \varphi\, d\mu + \int_Y \psi\, d\nu - \iint_{X\times Y} \big[\varphi(x)+\psi(y)\big]\, d\pi(x,y) \right) \to -\infty.$$

  Quindi in questo caso:

$$\Lambda^*(\pi) = - \inf_{u\in\mathcal{C}_b(X\times Y)} \left\{ \Lambda(u) - \iint_{X\times Y} u\, d\pi \right\} = +\infty.$$

- Se invece π è tale che, per ogni $(\varphi,\psi) \in \mathcal{C}_b(X)\times\mathcal{C}_b(Y)$, si ha:

$$\iint_{X\times Y} \big[\varphi(x)+\psi(y)\big]\, d\pi(x,y) = \int_X \varphi\, d\mu + \int_Y \psi\, d\nu,$$

  allora in questo caso:

$$\Lambda(u) - \iint_{X\times Y} u\, d\pi = \int_X \varphi\, d\mu + \int_Y \psi\, d\nu - \iint_{X\times Y} \big[\varphi(x)+\psi(y)\big]\, d\pi(x,y) = 0$$

per ogni u, che si scrive come somma di (φ, ψ), quindi in questo caso l'inf, e di conseguenza $\Lambda^*(\pi)$ sono 0.

Riassumendo abbiamo calcolato che:

$$\Lambda^*(\pi) = \begin{cases} 0 & \text{se } \overset{\forall(\varphi,\psi)\in\mathcal{C}_b(X)\times\mathcal{C}_b(Y)}{\iint_{X\times Y}[\varphi(x)+\psi(y)]\,d\pi(x,y)=\int_X \varphi\,d\mu+\int_Y \psi\,d\nu} \text{ si ha} \\ +\infty & \text{altrimenti.} \end{cases}$$

Da questo ricaviamo in particolare che:

$$\Theta^*(-\pi) + \Lambda^*(\pi) = \begin{cases} \displaystyle\iint_{X\times Y} c\,d\pi & \text{se } \pi \in \Pi(\mu,\nu), \\ +\infty & \text{altrimenti.} \end{cases}$$

Applicando dunque la formula (1.5), otteniamo:

$$\begin{aligned} -\sup_{(\varphi,\psi)\in\Phi_c\cap\mathcal{C}_b} J[\varphi,\psi] &= \max_{\pi\in\mathcal{M}(X\times Y)} \big[-\Theta^*(-\pi) - \Lambda^*(\pi)\big] \\ &= -\min_{\pi\in\mathcal{M}(X\times Y)} \big[\Theta^*(-\pi) + \Lambda^*(\pi)\big] \\ &= -\min_{\pi\in\Pi(\mu,\nu)} \big[\Theta^*(-\pi) + \Lambda^*(\pi)\big] \\ &= -\min_{\pi\in\Pi(\mu,\nu)} \iint_{X\times Y} c\,d\pi = -\min_{\pi\in\Pi(\mu,\nu)} I[\pi]. \end{aligned}$$

Cambiando il segno a destra e sinistra si ottiene la tesi, si noti che abbiamo anche mostrato che, in questo caso, il minimo è raggiunto. Questo termina il primo passo della dimostrazione.

Per il secondo passo, tralasciamo l'ipotesi di compattezza per X e Y, e manteniamo le ipotesi di limitatezza e uniforme continuità della funzione di costo c. Definiamo:

$$\|c\|_\infty := \sup_{(x,y)\in X\times Y} c(x,y).$$

Ora per prima cosa dimostriamo che si raggiungie il minimo. A tale scopo osserviamo che poichè X, Y sono spazi polacchi, le misure di probabilità μ, ν sono serrate, ossia per ogni $\varepsilon > 0$, esistono insiemi compatti $K_\varepsilon \subseteq X, L_\varepsilon \subseteq Y$ tali che:

$$\mu(K_\varepsilon^c) \le \varepsilon/2, \quad \nu(L_\varepsilon^c) \le \varepsilon/2.$$

Questo implica che $\Pi(\mu,\nu)$ è una famiglia di misure *uniformemente stirate*, infatti: per ogni $\varepsilon > 0$ possiamo considerare l'insieme: $K_\varepsilon \times L_\varepsilon \subseteq X \times Y$, allora esso è compatto dal teorema di Tychonoff ed inoltre per ogni $\pi \in \Pi(\mu,\nu)$ si ha:

$$\pi((K_\varepsilon\times L_\varepsilon)^c) = \pi(K_\varepsilon^c\times L_\varepsilon^c) \le \pi(K_\varepsilon^c\times Y)+\pi(X\times L_\varepsilon^c) = \mu(K_\varepsilon^c)+\nu(L_\varepsilon^c) \le \varepsilon/2+\varepsilon/2 = \varepsilon.$$

Segue allora dal teorema di Prohorov che $\Pi(\mu,\nu)$ è relativamente sequenzialmente compatta per la topologia debole* in $\mathcal{M}_f(X\times Y)$. Osserviamo anche che $\Pi(\mu,\nu)$

è anche debolmente* chiuso, e quindi debolmente* compatto. Se infatti $(\pi_n)_{n\in\mathbb{N}} \subseteq \Pi(\mu,\nu)$ è una successione tale che $\pi_n \xrightarrow[n\to\infty]{} \pi$ (debolmente), abbiamo certamente che $\pi \in \mathcal{P}(X\times Y)$, infatti: $1 \in \mathcal{C}_b(X\times Y)$ e quindi:

$$\pi(X\times Y) = \int_{X\times Y} d\pi = \lim_{n\to\infty} \int_{X\times Y} d\pi_n = 1.$$

Inoltre è facile osservare che $\pi \in \Pi(\mu,\nu)$. Questo segue dal fatto che se una successione di misure su uno spazio prodotto $(\pi_n)_{n\in\mathbb{N}}$ converge debolmente a π, allora anche le marginali di $(\pi_n)_{n\in\mathbb{N}}$ convergono debolmente a quelle di π. Infatti siano μ_n, μ le marginali di π_n, π sul primo spazio, allora detta $p\colon X\times Y \to X$ la proiezione canonica, abbiamo che data $f \in \mathcal{C}_b(X)$, si ha: $f\circ p \in \mathcal{C}_b(X\times Y)$. Dunque:

$$\begin{aligned}\int_X f\,d\mu &= \iint_{X\times Y} f\circ p\,d\pi \\ &= \lim_{n\to\infty} \iint_{X\times Y} f\circ p\,d\pi_n = \lim_{n\to\infty} \int_X f\,d\mu_n,\end{aligned}$$

cioè $\mu_n \xrightarrow[n\to\infty]{} \mu$ (debolmente). Allo stesso modo per l'altra marginale. Nel nostro caso la successione $(\pi_n)_{n\in\mathbb{N}}$ ha marginali costanti: μ e ν, che quindi sono anche le marginali di π, e quindi $\pi \in \Pi(\mu,\nu)$. Quindi $\Pi(\mu,\nu)$ è debolmente compatto. Ora l'integrale di una funzione continua e limitata su $X\times Y$, come c, è ovviamente continuo rispetto a la topologia debole*, per come essa è definita. Dunque segue dal teorema di Weistrass che il minimo è raggiunto.

Sia allora $\pi_* \in \Pi(\mu,\nu)$ che realizza il minimo e $\delta > 0$ arbitrariamente piccolo. Allora come conseguenza della uniforme serratezza di $\Pi(\mu,\nu)$, si ha che esistono $X_0 \subseteq X, Y_0 \subseteq Y$ tali che: $\mu(X_0), \nu(Y_0) \leq \delta$, ragionando come sopra otteniamo allora:

$$\pi_*((X_0\times Y_0)^c) \leq 2\delta.$$

Sia π_{*0} la probabilità condizionata rispetto a $X_0\times Y_0$ associata a π_*, ovvero:

$$\pi_{*0} = \frac{\pi_*\big(\cdot \cap (X_0\times Y_0)\big)}{\pi_*(X_0\times Y_0)},$$

Siano inoltre μ_0, ν_0 le sue marginali su X_0, Y_0 rispettivamente. Costruiamo un secondo problema di Kantorovich su $X_0\times Y_0$, allo scopo di usare il primo passo della dimostrazione ($X_0\times Y_0$ è infatti compatto). Definiamo quindi l'insieme:

$$\Pi_0(\mu_0,\nu_0) := \{\pi_0 \in \mathcal{P}(X_0\times Y_0) \mid \pi_0(A\times Y_0) = \mu_0(A), \pi(X_0\times B) = \nu_0(B) \quad \forall A \in \mathcal{M}_{|X_0}, B \in \mathcal{N}_{|Y_0}\},$$

ed il funzionale:

$$I_0\colon \Pi_0(\mu_0,\nu_0) \to [0,+\infty), \quad \pi_0 \mapsto \iint_{X_0\times Y_0} c\,d\pi_0.$$

Sia $\tilde{\pi}_0 \in \Pi_0(\mu_0, \nu_0)$ tale che:

$$I_0[\tilde{\pi}_0] = \inf_{\pi_0 \in \Pi_0(\mu_0,\nu_0)} I_0[\pi_0].$$

L'esistenza di $\tilde{\pi}_0$ è garantita dal primo passo della dimostrazione. Costruiamo ora $\tilde{\pi}$ usando $\tilde{\pi}_0$ e π_* nel seguente modo:

$$\tilde{\pi} := \pi_*(X_0 \times Y_0)\tilde{\pi}_0 + \pi_*\big(\cdot \cap (X_0 \times Y_0)^c\big),$$

Allora $\tilde{\pi} \in \Pi(\mu, \nu)$, infatti è certamente una misura non negativa su $X \times Y$, inoltre:

$$\begin{aligned}\tilde{\pi}(X \times Y) &= \pi_*(X_0 \times Y_0)\tilde{\pi}_0(X \times Y) + \pi_*\big((X \times Y) \cap (X_0 \times Y_0)^c\big) \\ &= \pi_*(X_0 \times Y_0) + \pi_*((X_0 \times Y_0)^c) = 1,\end{aligned}$$

e per ogni $A \in \mathcal{M}, B \in \mathcal{N}$ si ha:

$$\begin{aligned}\tilde{\pi}(A \times Y) &= \pi_*(X_0 \times Y_0)\tilde{\pi}_0(A \times Y) + \pi_*\big((A \times Y) \cap (X_0 \times Y_0)^c\big) \\ &= \pi_*(X_0 \times Y_0)\tilde{\pi}_0\big((A \cap X_0) \times Y_0\big) + \pi_*\big((A \cap X_0^c) \times Y_0^c\big) \\ &= \pi_*(X_0 \times Y_0)\mu_0(A \cap X_0) + \pi_*\big((A \cap X_0^c) \times Y_0^c\big) \\ &= \pi_*(X_0 \times Y_0)\pi_{*0}\big((A \cap X_0) \times Y_0\big) + \pi_*\big((A \cap X_0^c) \times Y_0^c\big) \\ &= \pi_*\big((A \cap X_0) \times Y_0\big) + \pi_*\big((A \cap X_0^c) \times Y_0^c\big) \\ &= \pi_*(A \times Y) = \mu(A),\end{aligned}$$

e

$$\begin{aligned}\tilde{\pi}(X \times B) &= \pi_*(X_0 \times Y_0)\tilde{\pi}_0(X \times B) + \pi_*\big((X \times B) \cap (X_0 \times Y_0)^c\big) \\ &= \pi_*(X_0 \times Y_0)\tilde{\pi}_0\big(X_0 \times (B \cap Y_0)\big) + \pi_*\big(X_0^c \times (B \cap Y_0^c)\big) \\ &= \pi_*(X_0 \times Y_0)\nu_0(B \cap Y_0) + \pi_*\big(X_0^c \times (B \cap Y_0^c)\big) \\ &= \pi_*(X_0 \times Y_0)\pi_{*0}\big(X_0 \times (B \cap Y_0)\big) + \pi_*\big(X_0^c \times (B \cap Y_0^c)\big) \\ &= \pi_*\big(X_0 \times (B \cap Y_0)\big) + \pi_*\big(X_0^c \times (B \cap Y_0^c)\big) \\ &= \pi_*(X \times B) = \nu(B).\end{aligned}$$

Allora:

$$\begin{aligned}I[\tilde{\pi}] &= \pi_*(X_0 \times Y_0)I_0[\tilde{\pi}_0] + \iint_{(X_0 \times Y_0)^c} c \, d\pi_* \\ &\leq I_0[\tilde{\pi}_0] + 2\|c\|_\infty \delta = \inf_{\pi_0 \in \Pi_0(\mu_0,\nu_0)} I_0[\pi_0] + 2\|c\|_\infty \delta.\end{aligned}$$

Da cui segue:

$$\inf_{\pi \in \Pi(\mu,\nu)} I[\pi] \leq \inf_{\pi_0 \in \Pi_0(\mu_0,\nu_0)} I_0[\pi_0] + 2\|c\|_\infty \delta.$$

Introduciamo ora il funzionale:

$$J_0 \colon L^1(\mu_0) \times L^1(\nu_0) \to \mathbb{R}, \quad (\varphi_0, \psi_0) \mapsto J_0[\varphi_0, \psi_0] := \int_{X_0} \varphi_0 \, d\mu_0 + \int_{Y_0} \psi_0 \, d\nu_0.$$

Dal primo passo della dimostrazione sappiamo che:

$$\inf_{\pi_0 \in \Pi_0(\mu_0,\nu_0)} I_0[\pi_0] = \sup_{(\varphi_0,\psi_0)\in\Phi_{0c}} J_0[\varphi_0,\psi_0],$$

ove:

$$\Phi_{0c} := \{(\varphi_0,\psi_0) \in L^1(\mu_0)\times L^1(\nu_0) \mid \varphi_0(x)+\psi_0(y) \le c(x,y) \quad \text{per } \mu_0\text{-q.o. } x \in X_0 \text{ e } \nu_0\text{-q.o. } y \in Y_0\}.$$

In particolare quindi, esistono $(\tilde\varphi_0, \tilde\psi_0) \in \Phi_{0c}$ tali che:

$$J_0[\tilde\varphi_0, \tilde\psi_0] \ge \sup_{(\varphi_0,\psi_0)\in\Phi_{0c}} J_0[\varphi_0,\psi_0] - \delta.$$

Il nostro scopo è ora costruire da $(\tilde\varphi_0, \tilde\psi_0)$ una coppia $(\varphi,\psi) \in \Phi_c$, che sia efficiente per il problema di massimizzazione di J. A tale scopo sarà utile che la disuguaglianza: $\tilde\varphi_0(x) + \tilde\psi_0(y) \le c(x,y)$ sia valida per ogni $(x,y) \in X \times Y$, non solo quasi ovunque. Questo si può fare fissando due versioni di $\tilde\varphi_0, \tilde\psi_0$ che valgono $-\infty$ negli insiemi di misura nulla dove la disuguaglianza non è vera. Senza perdita di generalità, possiamo assumere $\delta \le 1$. Poichè $J_0(0,0) = 0$, si ha: $\sup_{(\varphi_0,\psi_0)\in\Phi_{0c}} J_0[\varphi_0,\psi_0] \ge 0$, e quindi $J_0[\tilde\varphi_0, \tilde\psi_0] \ge -\delta \ge -1$. Ora scrivendo:

$$J_0[\tilde\varphi_0, \tilde\psi_0] = \iint_{X\times Y} \left[\tilde\varphi_0(x) + \tilde\psi_0(y)\right] d\pi_0(x,y),$$

ove π_0 è un qualsiasi elemento di $\Pi_0(\mu_0,\nu_0)$, deduciamo che esiste $(x_0,y_0) \in X_0 \times Y_0$ tale che:

$$\tilde\varphi_0(x_0) + \tilde\psi_0(y_0) \ge -1.$$

Adesso è sempre possibile scegliere $s \in \mathbb{R}$ tale che sostituendo la coppia: $(\tilde\varphi_0, \tilde\psi_0)$ con $(\tilde\varphi_0 + s, \tilde\psi_0 - s)$, che si noti è ancora ammissibile e non altera il valore di J_0, si abbia:

$$\tilde\varphi_0(x_0) \ge -\frac{1}{2}, \qquad \tilde\psi_0(y_0) \ge -\frac{1}{2}.$$

Questo implica in particolare che, per ogni $(x,y) \in X_0 \times Y_0$ si ha:

$$\tilde\varphi_0(x) \le c(x,y_0) - \tilde\psi_0(y_0) \le c(x,y_0) + \frac{1}{2},$$
$$\tilde\psi_0(y) \le c(x_0,y) - \tilde\varphi_0(x_0) \le c(x_0,y) + \frac{1}{2}.$$

Ora definiamo, per $x \in X$:

$$\overline{\varphi}_0(x) = \inf_{y\in Y_0} \left\{c(x,y) - \tilde\psi_0(y)\right\}.$$

Dalla disuguaglianza $\tilde\varphi_0(x) \le c(x,y) - \tilde\psi_0(y)$ abbiamo che $\tilde\varphi_0 \le \overline{\varphi}_0$ su X_0, e ciò implica: $J_0[\overline{\varphi}_0, \tilde\psi_0] \ge J_0[\tilde\varphi_0, \tilde\psi_0]$. Ma di più, per ogni $x \in X$ abbiamo controlli

da sotto e sopra di $\overline{\varphi}_0$ tramite la funzione di costo c, grazie alle disuguaglianze precedenti:

$$\overline{\varphi}_0(x) \geq \inf_{y\in Y_0}\{c(x,y) - c(x_0,y)\} - \frac{1}{2},$$
$$\overline{\varphi}_0(x) \leq c(x,y_0) - \tilde{\psi}_0(y_0) \leq c(x,y_0) + \frac{1}{2}.$$

Infine definiamo, per $y \in Y$:

$$\overline{\psi}_0(y) = \inf_{x\in X}\{c(x,y) - \overline{\varphi}_0(x)\}.$$

Allora per ogni $(x,y) \in X \times Y$, si ha:

$$\begin{aligned}\overline{\varphi}_0(x) + \overline{\psi}_0(y) &= \overline{\varphi}_0(x) + \inf_{z\in X}\{c(z,y) - \overline{\varphi}_0(z)\}\\ &\leq \overline{\varphi}_0(x) + c(x,y) - \overline{\varphi}_0(x) = c(x,y),\end{aligned}$$

quindi: $(\overline{\varphi}_0, \overline{\psi}_0) \in \Phi_c$. Ora non è difficile controllare che: $J_0[\overline{\varphi}_0, \overline{\psi}_0] \geq J_0[\overline{\varphi}_0, \tilde{\psi}_0] \geq J_0[\tilde{\varphi}_0, \tilde{\psi}_0]$. Infatti per ogni $y \in Y_0$ si ha:

$$\begin{aligned}\overline{\psi}_0(y) &:= \inf_{x\in X}\{c(x,y) - \overline{\varphi}_0(x)\}\\ &= \inf_{x\in X}\Big\{c(x,y) - \inf_{z\in Y_0}\{c(x,z) - \tilde{\psi}_0(z)\}\Big\}\\ &= \inf_{x\in X}\Big\{c(x,y) + \sup_{z\in Y_0}\{-c(x,z) + \tilde{\psi}_0(z)\}\Big\}\\ &\geq \inf_{x\in X}\{c(x,y) - c(x,y) + \tilde{\psi}_0(y)\} = \tilde{\psi}_0(y).\end{aligned}$$

Inoltre, per ogni $y \in Y$,

$$\overline{\psi}_0(y) \geq \inf_{x\in X}\{c(x,y) - c(x,y_0)\} - \frac{1}{2},$$
$$\overline{\psi}_0(y) \leq c(x_0,y) - \overline{\varphi}_0(x_0) \leq c(x_0,y) + \tilde{\varphi}_0(x_0) \leq c(x_0,y) + \frac{1}{2}.$$

In particolare, per ogni $(x,y) \in X \times Y$:

$$\overline{\varphi}_0(x) \geq -\|c\|_\infty - \frac{1}{2}, \overline{\psi}_0(y) \geq -\|c\|_\infty - \frac{1}{2}.$$

Ed ora grazie a tali disuguaglianze segue:

$$\begin{aligned}
J[\overline{\varphi}_0,\overline{\psi}_0] &= \int_X \overline{\varphi}_0\,d\mu + \int_Y \overline{\psi}_0\,d\nu = \iint_{X\times Y}\big[\overline{\varphi}_0(x)+\overline{\psi}_0(y)\big]\,d\pi_*(x,y)\\
&= \pi_*(X_0\times Y_0)\iint_{X_0\times Y_0}\big[\overline{\varphi}_0(x)+\overline{\psi}_0(y)\big]\,d\pi_{*0}(x,y) + \iint_{(X_0\times Y_0)^c}\big[\overline{\varphi}_0(x)+\overline{\psi}_0(y)\big]\,d\pi_*\\
&\geq (1-2\delta)\bigg(\int_{X_0}\overline{\varphi}_0\,d\mu_0 + \int_{Y_0}\overline{\psi}_0\,d\nu_0\bigg) - (2\|c\|_\infty+1)\pi_*\big((X_0\times Y_0)^c\big)\\
&\geq (1-2\delta)J_0[\overline{\varphi}_0,\overline{\psi}_0] - 2(2\|c\|_\infty+1)\delta\\
&\geq (1-2\delta)J_0[\tilde{\varphi}_0,\tilde{\psi}_0] - 2(2\|c\|_\infty+1)\delta\\
&\geq (1-2\delta)\big(\inf_{\pi_0\in\Pi_0(\mu_0,\nu_0)} I_0[\pi_0]-\delta\big) - 2(2\|c\|_\infty+1)\delta\\
&\geq (1-2\delta)\big(\inf_{\pi\in\Pi(\mu,\nu)} I[\pi] - (2\|c\|_\infty+1)\delta\big) - 2(2\|c\|_\infty+1)\delta.
\end{aligned}$$

Per arbitrarietà di $\delta\in(0,1]$ otteniamo:

$$\sup_{(\varphi,\psi)\in\Phi_c} J[\varphi,\psi] \geq \inf_{\pi\in\Pi(\mu,\nu)} I[\pi],$$

ricordando le disuguaglianze (1.4), abbiamo la tesi. Si noti che le funzioni $\overline{\varphi}_0,\overline{\psi}_0$ sono uniformemente continue su tutti X e Y siccome c è uniformemente continua (sono estremi inferiori di funzioni u.c.). Quindi sono in particolare misurabili e non è necessario specificare se si prende l'estremo superiore su $\Phi_c\cap\mathcal{C}_b$ o $\Phi_c\cap L^1$.

Passiamo ora al terzo passo, in cui consideriamo il caso più generale dell'enunciato del teorema. Dalla proposizione 1.2 sappiamo che possiamo scrivere: $c=\sup_{n\in\mathbb{N}} c_n$, ove c_n è una successione crescente di funzioni di costo uniformemente continue. Si ossevi inoltre che a patto di sostituire c_n, con $\min\{c_n,n\}$ (scelta che preserva la convergenza a c e la uniforme continuità), possiamo assumere che le c_n siano anche limitate.

Definiamo allora per ogni $n\in\mathbb{N}$ definiamo su $\Pi(\mu,\nu)$ il funzionale:

$$I_n\colon \Pi(\mu,\nu)\to[0,+\infty),\quad \pi\mapsto I_n[\pi] := \iint_{X\times Y} c_n\,d\pi.$$

Allora dal secondo passo sappiamo che:

$$\inf_{\pi\in\Pi(\mu,\nu)} I_n[\pi] = \sup_{(\varphi,\psi)\in\Phi_{c_n}} J[\varphi,\psi]. \tag{1.6}$$

Concluderemo la dimostrazione facendo vedere che:

$$\inf_{\pi\in\Pi(\mu,\nu)} I[\pi] = \sup_{n\in\mathbb{N}}\inf_{\pi\in\Pi(\mu,\nu)} I_n[\pi], \tag{1.7}$$

e che per ogni $n\in\mathbb{N}$, vale:

$$\sup_{(\varphi,\psi)\in\Phi_{c_n}} J[\varphi,\psi] \leq \sup_{(\varphi,\psi)\in\Phi_c} J[\varphi,\psi]. \tag{1.8}$$

Difatti utilizzando in combinazione (1.6), (1.7) e (1.8), otteniamo:

$$\begin{aligned}\inf_{\pi\in\Pi(\mu,\nu)} I[\pi] &= \sup_{n\in\mathbb{N}} \inf_{\pi\in\Pi(\mu,\nu)} I_n[\pi] \\ &= \sup_{n\in\mathbb{N}} \sup_{(\varphi,\psi)\in\Phi_{c_n}} J[\varphi,\psi] \\ &\leq \sup_{n\in\mathbb{N}} \sup_{(\varphi,\psi)\in\Phi_c} J[\varphi,\psi] = \sup_{(\varphi,\psi)\in\Phi_c} J[\varphi,\psi].\end{aligned}$$

ossia:

$$\inf_{\pi\in\Pi(\mu,\nu)} I[\pi] \leq \sup_{(\varphi,\psi)\in\Phi_c} J[\varphi,\psi],$$

sappiamo poi da (1.4) che l'altra disuguaglianza è sempre verificata e quindi possiamo concludere. Proviamo le due equazioni (1.7) ed (1.8). Osserviamo che per costruzione: $c_n \leq c$ per ogni $n \in \mathbb{N}$, quindi $\Phi_{c_n} \subseteq \Phi_c$ per ogni $n \in \mathbb{N}$, segue allora che (1.8) è banalmente vera, perchè stiamo facendo l'estremo superiore di un insieme più grande di numeri reali (non è inoltre necessario specificare se stiamo considerando $\Phi_c \cap \mathcal{C}_b$ o $\Phi_c \cap L^1$). Per (1.7) osserviamo preliminarmente che la crescenza di c_n implica, per monotonia dell'integrale, che I_n è una successione crescente di funzionali, limitata dall'alto da I. Ma allora anche:

$$\left(\inf_{\pi\in\Pi(\mu,\nu)} I_n[\pi]\right)_{n\in\mathbb{N}}$$

è una successione crescente, limitata dall'alto da:

$$\inf_{\pi\in\Pi(\mu,\nu)} I[\pi]$$

Per provare (1.7) è allora sufficiente dimostrare che:

$$\lim_{n\to\infty} \inf_{\pi\in\Pi(\mu,\nu)} I_n[\pi] \geq \inf_{\pi\in\Pi(\mu,\nu)} I[\pi]. \tag{1.9}$$

A tale scopo abbiamo già notato che $\Pi(\mu,\nu)$ è debolmente compatto, quindi ogni successione in esso ammette una sottosuccessione convergente ad un elemento in $\Pi(\mu,\nu)$. Per ogni $n \in \mathbb{N}$ sia $(\pi_n^k)_{k\in\mathbb{N}}$ una successione minimizzante per:

$$\inf_{\pi\in\Pi(\mu,\nu)} I_n[\pi].$$

Allora ognuna di queste successioni ammette una sottosuccessione convergente ad una misura $\pi_n \in \Pi(\mu,\nu)$ che realizza il minimo. Ma ora, sempre grazie alla compattezza di $\Pi(\mu,\nu)$, abbiamo che $(\pi_n)_{n\in\mathbb{N}}$, ammette una sottosuccessione convergente ad un elemento $\pi_* \in \Pi(\mu,\nu)$. Adesso per la crescenza di c_n, abbiamo già osservato che se $n \geq m$ allora:

$$I_n[\pi_n] \geq I_m[\pi_n],$$

da cui, per la continuità di I_m rispetto alla convergenza debole abbiamo:

$$\lim_{n\to\infty} I_n[\pi_n] \geq \limsup_{n\to\infty} I_m[\pi_n] \geq I_m[\pi_*].$$

Ora per convergenza monotona di c_m a c, abbiamo: $I_m[\pi_*] \xrightarrow[m\to\infty]{} I[\pi_*]$, quindi:

$$\lim_{n\to\infty} I_n[\pi_n] \geq \lim_{m\to\infty} I_m[\pi_*] = I[\pi_*] \geq \inf_{\pi\in\Pi(\mu,\nu)} I[\pi],$$

che prova la (1.9) per la minimalità di π_n, per ogni $n \in \mathbb{N}$. L'unica cosa che rimane da provare per terminare il terzo passo della dimostrazione è provare che il minimo è raggiunto. Per farlo sia $(\pi_n)_{n\in\mathbb{N}}$ una successione minimizzante per:

$$\inf_{\pi\in\Pi(\mu,\nu)} I[\pi],$$

e sia $\pi_* \in \Pi(\mu,\nu)$ il limite di una sottosuccessione convergente, $(\pi_{n_k})_{k\in\mathbb{N}}$, di $(\pi_n)_{n\in\mathbb{N}}$. Allora sfruttando che, c_n, I_n sono successioni crescenti ed il teorema di convergenza monotona otteniamo:

$$\begin{aligned} I[\pi_*] &= \lim_{n\to\infty} I_n[\pi_*] \leq \lim_{n\to\infty} \limsup_{k\to\infty} I_n[\pi_{n_k}] \\ &\leq \limsup_{k\to\infty} I[\pi_{n_k}] = \lim_{k\to\infty} I[\pi_{n_k}] = \inf_{\pi\in\Pi(\mu,\nu)} I[\pi], \end{aligned}$$

quindi:

$$I[\pi_*] = \min_{\pi\in\Pi(\mu,\nu)} I[\pi],$$

e si conclude. □

Osservazione 1.4 (Funzioni c-concave coniugate). Si noti che dalla dimostrazione, segue in particolare che, quando c è limitata e uniformemente continua, è possibile restringere l'insieme su cui si effettua l'estremo superiore nel lato destro di (1.3) alle coppie: $(\varphi^{cc}, \varphi^c)$, ove φ è limitata su X e:

$$\varphi^c(y) := \inf_{x\in X}\big\{c(x,y) - \varphi(x)\big\}, \qquad \varphi^{cc}(x) := \inf_{y\in Y}\big\{c(x,y) - \varphi^c(y)\big\}. \tag{1.10}$$

La coppia $(\varphi^{cc}, \varphi^c)$ si dice *coppia delle funzioni c-concave coniugate associate a* φ. Questa proprietà può essere estesa al caso più generale in cui c sia semicontinua inferiormente e limitata. Infatti esiste allora una successione crescente di funzioni u.c. c_k che convergono puntualmente a c, e in questo caso si ha

$$\varphi^c(y) = \lim_{k\to\infty} \inf_{x\in X}\big\{c_k(x,y) - \varphi(x)\big\},$$

quindi φ^c è limite puntuale di funzioni u.c. quindi in particolare misurabili e quindi è anch'essa misurabile. Similmente φ^{cc} è misurabile.

Ci concentriamo ora al caso specifico di $X = Y$ e di $c = \mathrm{d}$, dove d indica una distanza. In questo contesto il teorema di dualità può essere raffinato come segue:

Teorema 1.6 (Teorema di Kantorovich-Rubenstein)**.** *Sia X uno spazio polacco, μ e ν due misure di probabilità definite sulla σ-algebra dei Boreliani di X e* d *una*

distanza inferiormente semicontinua su X. Sia poi $\mathcal{I}_{\mathrm{d}}(\mu,\nu)$ il valore ottimo del relativo problema di Kantorovich, ossia:

$$\mathcal{I}_{\mathrm{d}}(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \iint_{X\times Y} \mathrm{d}\, d\pi.$$

Allora:

$$\mathcal{I}_{\mathrm{d}}(\mu,\nu) = \sup\left\{ \int_X \varphi\, d\mu - \int_X \varphi\, d\nu \mid \varphi \in L^1(|\mu-\nu|)\cap Lip(X), \|\varphi\|_{Lip} \le 1 \right\}. \quad (1.11)$$

Dove $Lip(X)$ indica l'insieme delle funzioni Lipschitziane su X e:

$$\|\varphi\|_{Lip} := \sup_{\substack{x,y\in X\\ x\neq y}} \frac{|\varphi(x)-\varphi(y)|}{\mathrm{d}(x,y)}$$

è la costante di Lipschitz di φ.

Dimostrazione. Sia per $n\in\mathbb{N}$:

$$\mathrm{d}_n := \frac{\mathrm{d}}{1+n^{-1}\mathrm{d}} \le n.$$

Allora per ogni n, d_n è una distanza su X che verifica: $\mathrm{d}_n \le \mathrm{d}$, inoltre per ogni $x,y\in X$ la quantità $\mathrm{d}_n(x,y)$ converge crescendo a $\mathrm{d}(x,y)$ per $n\to\infty$. In particolare l'insieme delle funzioni 1-Lipschitziane rispetto a d_n è contenuto in quello delle funzioni 1-Lipschitziane rispetto a d. Ragionando allora come nell'ultimo passo del teorema 1.3, è sufficiente provare l'affermazione solo per d_n. Possiamo dunque assumere che d sia limitata.

In tal caso tutte le funzioni Lipschitziane sono limitate, e quindi integrabili rispetto a μ e ν. Inoltre in virtù del teorema 1.3, l'unica cosa da provare è che:

$$\sup_{(\varphi,\psi)\in\Phi_{\mathrm{d}}} J[\varphi,\psi] = \sup\left\{ \int_X \varphi\, d\mu - \int_X \varphi\, d\nu \mid \varphi\in L^1(|\mu-\nu|)\cap \mathrm{Lip}(X), \|\varphi\|_{\mathrm{Lip}} \le 1 \right\},$$

ove si ricordi: $J[\varphi,\psi] := \int_X \varphi\, d\mu + \int_X \psi\, d\nu$. Dall'osservazione 1.4 sappiamo che:

$$\sup_{(\varphi,\psi)\in\Phi_{\mathrm{d}}} J[\varphi,\psi] = \sup_{\varphi \text{ limitata}} J[\varphi^{\mathrm{dd}},\varphi^{\mathrm{d}}],$$

ove:

$$\varphi^{\mathrm{d}}(y) := \inf_{x\in X}\big[\mathrm{d}(x,y)-\varphi(x)\big], \qquad \varphi^{\mathrm{dd}}(x) := \inf_{y\in X}\big[\mathrm{d}(x,y)-\varphi^{\mathrm{d}}(y)\big].$$

Ora, φ^{d}, essendo l'estremo inferiore di funzioni 1-Lipshitziane, limitate dal basso in qualche punto $x_0\in X$, è 1-Lipshitziana. Dunque:

$$-\varphi^{\mathrm{d}}(x) \le \inf_{y\in X}\big[\mathrm{d}(x,y)-\varphi^{\mathrm{d}}(y)\big] \le -\varphi^{\mathrm{d}}(x),$$

ove la disuguaglianza di destra segue scegliendo $y = x$ nell'estremo inferiore, mentre quella di sinistra segue dalla 1-Lipschitzianità. Questo significa che: $\varphi^{\mathrm{dd}} = -\varphi^{\mathrm{d}}$, e dunque:

$$\begin{aligned}\sup_{(\varphi,\psi)\in\Phi_{\mathrm{d}}} J[\varphi,\psi] &= \sup_{\varphi\in L^1(\mu)} J[\varphi^{\mathrm{dd}},\varphi^{\mathrm{d}}] = \sup_{\varphi\in L^1(\mu)} J[-\varphi^{\mathrm{d}},\varphi^{\mathrm{d}}] \\ &\leq \sup_{\|\varphi\|_{\mathrm{Lip}}\leq 1} J[\varphi,-\varphi] \leq \sup_{(\varphi,\psi)\in\Phi_{\mathrm{d}}} J[\varphi,\psi].\end{aligned}$$

Quindi le precedenti sono tutte uguaglianze, ed il teorema è dimostrato. □

1.2 La Distanza di Wasserstein

In questa sezione introduciamo la distanza di Wasserstein tra misure di probabilità su uno spazio Polacco X, per il resto della sezione assumiamo che le distanze considerate siano sempre inferiormente semicontinue di modo tale che siano verificate le ipotesi del teorema di dualità. Prima di dare l'effettiva definizione della distanza è tuttavia opportuno fissare una notazione:

Definizione 1.5. Sia (X,d) uno spazio Polacco, dotato della sua σ-algebra dei Boreliani: $\mathcal{M}$. Per $r \geq 1$ definiamo il seguente insieme:

$$\mathcal{P}_r(X,\mathrm{d}) := \left\{\mu \in \mathcal{P}(X) \mid \int_X \mathrm{d}(x,x_0)^r \, d\mu(x) < +\infty\right\} \tag{1.12}$$

per qualche (e quindi per ogni) $x_0 \in X$.

Osservazione 1.5. Il "e quindi per ogni" nella definizione sopra è implicato dalla seguente disuguaglianza, valida per ogni $r \geq 1$ e per ogni $x, y \in X$:

$$\mathrm{d}(x,y)^r \leq 2^{r-1}\big(\mathrm{d}(x,x_0)^r + \mathrm{d}(y,x_0)^r\big).$$

Si osservi anche che se d è limitata allora ovviamente $\mathcal{P}_r(X,\mathrm{d}) = \mathcal{P}(X)$ per ogni $r \geq 1$.

Definizione 1.6 (Distanza di Wasserstein)**.** Sia (X,d) come nella definizione precedente, siano poi $r \geq 1$ e $\mu,\nu \in \mathcal{P}_r(X,\mathrm{d})$. Si dice *distanza di Wasserstein di ordine* r tra μ e ν la quantità $\mathcal{W}_r(\mu,\nu) := (\mathcal{I}_{\mathrm{d}^r}(\mu,\nu))^{1/r}$, ossia:

$$\mathcal{W}_r(\mu,\nu) := \left(\inf_{\pi\in\Pi(\mu,\nu)} \iint_{X\times X} \mathrm{d}(x,y)^r \, d\pi(x,y)\right)^{1/r}. \tag{1.13}$$

con l'usuale significato di $\Pi(\mu,\nu)$.

Nel seguito dimostriamo che effettivamente $\mathcal{W}_r$ definisce una distanza su $\mathcal{P}_r(X,\mathrm{d})$, per farlo ci occorre però il seguente lemma:

Lemma 1.7 (Gluing Lemma)**.** *Siano μ_1, μ_2, μ_3 tre misure di probabilità sugli spazi Polacchi: X_1, X_2, X_3 rispettivamente. Siano allora: $\pi_{12} \in \Pi(\mu_1, \mu_2)$ e $\pi_{23} \in \Pi(\mu_2, \mu_3)$. Allora esiste una misura di probabilità $\pi \in \mathcal{P}(X_1 \times X_2 \times X_3)$ avente per marginali π_{12} su $X_1 \times X_2$ e π_{23} su $X_2 \times X_3$.*

Per la dimostrazione del lemma si riferisca a quella contenuta in [12].

Teorema 1.8. *Sia (X, d) come nella definizione 1.5, allora per ogni $r \geq 1$, si ha che $\mathcal{W}_r$ definisce una metrica su $\mathcal{P}_r(X, \mathrm{d})$.*

Dimostrazione. Sia $r \geq 1$ fissato. Proviamo preliminarmente che $\mathcal{W}_r$ su $\mathcal{P}_r(X, \mathrm{d})$ è finita. Siano $\mu, \nu \in \mathcal{P}_r(X, \mathrm{d})$ allora, fissato $x_0 \in X$, per ogni $x, y \in X$ usiamo nuovamente la disuguaglianza:

$$\mathrm{d}(x, y)^r \leq 2^{r-1}\big(\mathrm{d}(x, x_0)^r + \mathrm{d}(y, x_0)^r\big)$$

e che la misura prodotto: $\mu \otimes \nu$ ha le marginali richieste abbiamo:

$$\begin{aligned}
\mathcal{W}_r(\mu, \nu)^r &\leq \iint_{X \times X} \mathrm{d}(x, y)^r \, d\mu \otimes \nu(x, y) \\
&\leq 2^{r-1} \iint_{X \times X} \big(\mathrm{d}(x, x_0)^r + \mathrm{d}(y, x_0)^r\big) \, d\mu \otimes \nu(x, y) \\
&= 2^{r-1} \left[\int_X \int_X \mathrm{d}(x, x_0)^r \, d\mu(x) \, d\nu(y) + \int_X \int_X \mathrm{d}(y, x_0)^r \, d\mu(x) \, d\nu(y) \right] \\
&= 2^{r-1} \left[\int_X \mathrm{d}(x, x_0)^r \, d\mu(x) + \int_Y \mathrm{d}(y, x_0)^r \, d\nu(y) \right] < \infty.
\end{aligned}$$

Dove nella seconda uguaglianza abbiamo utilizzato la linearità dell'integrale ed il teorema di Tonelli (l'integranda è misurabile e ≥ 0). È chiaro poi che $\mathcal{W}_r$ è non negativa. Supponiamo $\mu, \nu \in \mathcal{P}_r(X, \mathrm{d})$ tali che: $\mathcal{W}_r(\mu, \nu) = 0$. Per dimostrare che allora $\mu = \nu$, notiamo che per ogni $1 \leq s \leq r$ si ha: $\mathcal{W}_s(\mu, \nu) \leq \mathcal{W}_r(\mu, \nu)$. Infatti, per ogni $\pi \in \Pi(\mu, \nu)$, usando la disuguaglianza di Hölder e che:

$$\frac{1}{\frac{r}{s}} + \frac{1}{\frac{r}{r-s}} = 1,$$

otteniamo:

$$\begin{aligned}
\iint_{X \times Y} \mathrm{d}^s \, d\pi &= \iint_{X \times Y} \mathrm{d}^s \cdot 1 \, d\pi \\
&\leq \left(\iint_{X \times Y} \mathrm{d}^{s\frac{r}{s}} \, d\pi \right)^{\frac{s}{r}} \cdot \left(\iint_{X \times Y} 1^{\frac{r}{r-s}} \, d\pi \right)^{\frac{r-s}{r}} \\
&= \left(\iint_{X \times Y} \mathrm{d}^r \, d\pi \right)^{\frac{s}{r}}.
\end{aligned}$$

da cui:

$$\left(\iint_{X\times Y} \mathrm{d}^s \, d\pi\right)^{1/s} \leq \left(\iint_{X\times Y} \mathrm{d}^r \, d\pi\right)^{1/r}.$$

Passando allora all'inf in π otteniamo: $\mathcal{W}_s(\mu,\nu) \leq \mathcal{W}_r(\mu,\nu)$. Questo implica che è sufficiente dimostrare che $\mathcal{W}_r(\mu,\nu) = 0$, implica $\mu = \nu$, solo per $r = 1$. Dal teorema di Kantorovich-Rubestein otteniamo che se $\mathcal{W}_1(\mu,\nu) = 0$ allora:

$$\sup\left\{\int_X \varphi \, d\mu - \int_X \varphi \, d\nu \mid \varphi \in L^1(|\mu-\nu|) \cap \mathrm{Lip}(X), \|\varphi\|_{\mathrm{Lip}} \leq 1\right\} = 0.$$

Da cui: per ogni $\varphi \in L^1(|\mu-\nu|) \cap \mathrm{Lip}(X)$, con $\|\varphi\|_{\mathrm{Lip}} \leq 1$, vale:

$$\int_X \varphi \, d\mu = \int_X \varphi \, d\nu,$$

ma poichè tale famiglia è separante per $\mathcal{P}(X)$ (proposizione A.2), otteniamo: $\mu = \nu$. $\mathcal{W}_r$ è poi ovviamente simmetrica, infatti date $\mu,\nu \in \mathcal{P}_r(X,\mathrm{d})$, gli elementi di $\Pi(\mu,\nu)$ sono in corrispondenza biunivoca con gli elementi di $\Pi(\nu,\mu)$. Infatti presa $\pi \in \Pi(\mu,\nu)$ possiamo definire $\hat{\pi} \in \Pi(\nu,\mu)$ ponendo, per ogni $A, B \in \mathcal{M}$

$$\hat{\pi}(A \times B) := \pi(B \times A),$$

e viceversa. È sufficiente ora osservare che per ogni $\pi \in \Pi(\mu,\nu)$ si ha:

$$\iint_{X\times X} \mathrm{d}(x,y)^r \, d\pi(x,y) = \iint_{X\times X} \mathrm{d}(y,x)^r \, d\pi(x,y) = \iint_{X\times X} \mathrm{d}(y,x)^r \, d\hat{\pi}(y,x).$$

Dunque $\mathcal{W}_r(\mu,\nu) = \mathcal{W}_r(\nu,\mu)$. L'unica cosa che rimane quindi da dimostrare è la disuguaglianza triangolare. Per farlo usiamo il lemma 1.7. Siano: $\mu_1,\mu_2,\mu_3 \in \mathcal{P}_r(X,\mathrm{d})$ e siano $\pi_{12} \in \Pi(\mu_1,\mu_2)$ e $\pi_{23} \in \Pi(\mu_2,\mu_3)$ che realizzano il minimo per la distanza tra μ_1,μ_2 e μ_2,μ_3 rispettivamente. Definiamo $X_j := \mathrm{supp}\,\mu_j \subseteq X$, per $j =$ 1,2,3 (il supporto di una misura μ è definito come il più piccolo sottoinsieme chiuso $F \subseteq X$ tale che $\mu(X \setminus F) = 0$). Sia allora π come nell'enunciato del lemma 1.7, e sia π_{13} la marginale di π su $X_1 \times X_3$, è chiaro allora che $\pi_{13} \in \Pi(\mu_1,\mu_3)$. Usando successivamente la definizione di $\mathcal{W}_r$, la proprietà sulle marginali e la disuguaglianza

di Minkovski per le funzioni L^r otteniamo:

$$\begin{aligned}
\mathcal{W}_r(\mu_1,\mu_3) &\le \left(\iint_{X_1\times X_3} \mathrm{d}(x_1,x_3)^r\, d\pi_{13}(x_1,x_3)\right)^{1/r}\\
&= \left(\iiint_{X_1\times X_2\times X_3} \mathrm{d}(x_1,x_3)^r\, d\pi(x_1,x_2,x_3)\right)^{1/r}\\
&\le \left(\iiint_{X_1\times X_2\times X_3} \big[\mathrm{d}(x_1,x_2)+\mathrm{d}(x_2,x_3)\big]^r\, d\pi(x_1,x_2,x_3)\right)^{1/r}\\
&\le \left(\iiint_{X_1\times X_2\times X_3} \mathrm{d}(x_1,x_2)^r\, d\pi(x_1,x_2,x_3)\right)^{1/r}\\
&+ \left(\iiint_{X_1\times X_2\times x_3} \mathrm{d}(x_2,x_3)^r\, d\pi(x_1,x_2,x_3)\right)^{1/r}\\
&= \left(\iint_{X_1\times X_2} \mathrm{d}(x_1,x_2)^r\, d\pi_{12}(x_1,x_2)\right)^{1/r} + \left(\iint_{X_2\times X_3} \mathrm{d}(x_2,x_3)^r\, d\pi_{23}(x_2,x_3)\right)^{1/r}\\
&= \mathcal{W}_r(\mu_1,\mu_2)+\mathcal{W}_r(\mu_2,\mu_3).
\end{aligned}$$

□

Osservazione 1.6. Osserviamo come dalla dimostrazione abbiamo mostrato un'importante proprietà della famiglia di distanze $\{\mathcal{W}_r\}_{r\ge1}$, ovvero che sono ordinate, nel senso che:

$$r_1 \ge r_2 \quad \text{implica} \quad \mathcal{W}_{r_1} \ge \mathcal{W}_{r_2}.$$

In generale una stima nell'altro senso non si può ottenere, a meno che d sia limitata. Infatti, supponiamo $r_1 \ge r_2$, allora per limitatezza:

$$\mathrm{diam}(X) := \sup_{x,y\in X} \mathrm{d}(x,y) < +\infty,$$

e quindi, per ogni $\mu,\nu \in \mathcal{P}(X)$:

$$\begin{aligned}
\mathcal{W}_{r_1}(\mu,\nu)^{r_1} &= \inf_{\pi\in\Pi(\mu,\nu)} \iint_{X\times X} \mathrm{d}(x,y)^{r_1}\, d\pi(x,y)\\
&= \inf_{\pi\in\Pi(\mu,\nu)} \iint_{X\times X} \mathrm{d}(x,y)^{r_2}\mathrm{d}(x,y)^{r_1-r_2}\, d\pi(x,y)\\
&\le \mathrm{diam}(X)^{r_1-r_2} \inf_{\pi\in\Pi(\mu,\nu)} \iint_{X\times X} \mathrm{d}(x,y)^{r_2}\, d\pi(x,y)\\
&= \mathrm{diam}(X)^{r_1-r_2}\mathcal{W}_{r_2}(\mu,\nu)^{r_2}.
\end{aligned}$$

Da cui otteniamo la stima:

$$\mathcal{W}_{r_1}(\mu,\nu) \le \mathrm{diam}(X)^{1-\frac{r_2}{r_1}}\mathcal{W}_{r_2}(\mu,\nu)^{\frac{r_2}{r_1}},$$

questo ci dice che se d è limitata allora tutte le distanze $\{\mathcal{W}_r\}_{r\ge1}$ sono equivalenti, cioè generano la stessa topologia su $\mathcal{P}(X)$. Vedremo in una successiva sessione che

essa coincide con la traccia della topologia debole* che è possibile considerare su $\mathcal{P}(X)$ visto come sottoinsieme di $\mathcal{M}(X)$.

Proposizione 1.9. *Sia* (X, d) *uno spazio Polacco, allora per ogni* $r \geq 1, x_0 \in X, \mu \in \mathcal{P}_r(X, \mathrm{d})$ *si ha:*

$$\mathcal{W}_r(\mu, \delta_{x_0})^r = \int_X \mathrm{d}(x, x_0)^r \, d\mu(x),$$

e quindi la mappa:

$$i \colon (X, \mathrm{d}) \to (\mathcal{P}_r(X, \mathrm{d}), \mathcal{W}_r) \quad x \mapsto \delta_x$$

che assegna ad ogni punto la misura di Dirac concentrata in quel punto è una immersione isometrica.

Dimostrazione. La prima osservazione segue dal fatto che c'è un'unica misura π sullo spazio prodotto con marginali μ e δ_{x_0}, ossia: $\pi = \mu \otimes \delta_{x_0}$, dunque:

$$\mathcal{W}_r(\mu, \delta_{x_0}) = \int_X \int_X \mathrm{d}(x, y)^r \, d\delta_{x_0}(y) \, d\mu(x) = \int_X \mathrm{d}(x, x_0)^r \, d\mu(x).$$

Con la scelta particolare di $\mu = \delta_{x_1}$ la formula sopra diviene:

$$\mathcal{W}_r(\delta_{x_0}, \delta_{x_1})^r = \int_X \mathrm{d}(x_0, x)^r \, d\delta_{x_1}(x) = \mathrm{d}(x_0, x_1)^r,$$

da cui: $x \mapsto \delta_x$ è un'isometria. □

Proposizione 1.10 (r-Convessità)**.** *Sia* (X, d) *uno spazio Polacco,* $r \geq 1$ *e* $(a_n)_{n \in \mathbb{N}} \subseteq [0, +\infty)$ *una successione tale che* $\sum_{n=1}^{\infty} a_n = 1$*. Allora data* $(\mu_n)_{n \in \mathbb{N}} \subseteq \mathcal{P}_r(X, \mathrm{d})$ *e* $\mu \in \mathcal{P}_r(X, \mathrm{d})$ *si ha:*

$$\mathcal{W}_r\Big(\mu, \sum_{n=1}^{\infty} a_n \mu_n\Big)^r \leq \sum_{n=1}^{\infty} a_n \mathcal{W}_r(\mu, \mu_n)^r. \tag{1.14}$$

Dimostrazione. Per ogni n, sia π_n la misura di probabilità in $\Pi(\mu, \mu_n)$ che realizza il minimo per la definizione di $\mathcal{W}_r(\mu, \mu_n)$. Consideriamo allora la misura π definita da:

$$\pi := \sum_{n=1}^{\infty} a_n \pi_n,$$

allora è ovvio che π ha come marginali le misure μ e $\sum_{n=1}^{\infty} a_n \mu_n$ e dunque:

$$\begin{aligned} \mathcal{W}_r\Big(\mu, \sum_{n=1}^{\infty} a_n \mu_n\Big)^r &\leq \iint_{X \times X} \mathrm{d}(x, y)^r \, d\pi(x, y) \\ &= \sum_{n=1}^{\infty} a_n \iint_{X \times X} \mathrm{d}(x, y)^r \, d\pi_n(x, y) \\ &= \sum_{n=1}^{\infty} a_n \mathcal{W}_r(\mu, \mu_n)^r. \end{aligned}$$

Ove l'ultima uguaglianza segue dalla definizione di π_n. □

1.3 Proprietà topologiche della Distanza di Wasserstein

In questa sezione studiamo le proprietà topologiche della distanza di Wasserstein, in particolare il suo legame con la *convergenza debole* delle misure.

Teorema 1.11. *Sia (X, d) uno spazio Polacco, dotato della sua σ-algebra dei Boreliani $\mathcal{M}$. Siano poi $r \geq 1, (\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}_r(X, \mathrm{d}), \mu \in \mathcal{P}(X)$. Allora, le seguenti affermazioni sono equivalenti:*

1. $\mathcal{W}_r(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0$;
2. $\mu_n \xrightarrow[n\to\infty]{} \mu$ *(debolmente), e $(\mu_n)_{n\in\mathbb{N}}$ soddisfa la seguente condizione di serratezza per qualche (e quindi ogni) $x_0 \in X$:*
$$\lim_{R\to\infty} \limsup_{n\to\infty} \int_{\mathrm{d}(x_0,x)\geq R} \mathrm{d}(x_0, x)^r \, d\mu_n(x) = 0; \tag{1.15}$$
3. $\mu_n \xrightarrow[n\to\infty]{} \mu$ *(debolmente), e vale la seguente condizione di convergenza dei momenti per qualche (e quindi ogni) $x_0 \in X$:*
$$\int_X \mathrm{d}(x_0, x)^r \, d\mu_n(x) \xrightarrow[n\to\infty]{} \int_X \mathrm{d}(x_0, x)^r \, d\mu(x). \tag{1.16}$$
4. *Per ogni funzione $\phi \in \mathcal{C}(X)$ che soddisfa la condizione di crescita sublineare: $|\phi(x)| \leq C(1 + \mathrm{d}(x_0, x)^r)$, per qualche $x_0 \in X, C > 0$, vale:*
$$\int_X \phi \, d\mu_n \xrightarrow[n\to\infty]{} \int_X \phi \, d\mu \tag{1.17}$$

Dimostrazione. Supponiamo siano soddisfatte le ipotesi del teorema. Proviamo inanzitutto 2, 3 e 4 sono equivalenti. Ovviamente 4 implica 3 poichè 4 implica banalmente la convergenza debole di μ_n a μ e la funzione $x \mapsto \mathrm{d}(x_0, x)^r$ è continua per $x_0 \in X$ fissato e soddisfa banalmente la proprietà di crescita sublineare, quindi è anche verificata la proprietà di convergenza dei momenti. Mostriamo che 2 implica 4, sia $x_0 \in X$ che soddisfa la condizione in 2 e supponiamo μ_n converga debolmente a μ. Sia ϕ una arbitraria funzione continua su X che soddisfa la condizione di crescita richiesta in 4, per ogni $R > 1$ scriviamo:

$$\phi = \phi_R + \psi_R \quad \text{ove} \quad \phi_R(x) := \min\{\phi(x), C(1 + R^r)\} \quad \text{e} \quad \psi_R = \phi - \phi_R.$$

Si noti che ψ_R è puntualmente limitata da $C\mathrm{d}(x_0, x)^r \mathbf{1}_{\mathrm{d}(x_0,x)\geq R}$. Infatti se $x \in X$ è tale che $\mathrm{d}(x, x_0) < R$ allora $\phi_R(x) = \phi(x)$ per l'ipotesi di crescita, e quindi $\psi(x) = 0$

e la disuguaglianza è banale. Se invece $\mathrm{d}(x, x_0) \geq R$ allora $\phi_R(x) = C(1 + R^r)$ e dunque

$$\psi_R(x) = \phi(x) - C(1 + R^r) \leq C(1 + \mathrm{d}(x_0, x)^r) - C - CR^r \leq C\mathrm{d}(x_0, x)^r.$$

Allora:

$$\begin{aligned}\left|\int_X \phi\, d\mu_n - \int_X \phi\, d\mu\right| &\leq \left|\int_X \phi_R\, d(\mu_n - \mu)\right| + C\int_{\mathrm{d}(x_0,x)\geq R} \mathrm{d}(x_0, x)^r\, d\mu_n(x) \\ &\quad + C\int_{\mathrm{d}(x_0,x)\geq R} \mathrm{d}(x_0, x)^r\, d\mu(x) \\ &= \left|\int_X \phi_R\, d(\mu_n - \mu)\right| + C\int_{\mathrm{d}(x_0,x)\geq R} \mathrm{d}(x_0, x)^r\, d(\mu_n + \mu)(x).\end{aligned}$$

Dunque,

$$\begin{aligned}\limsup_{n\to\infty}\left|\int_X \phi\, d\mu_n - \int_X \phi\, d\mu\right| &\leq \lim_{n\to\infty}\left|\int_X \phi_R\, d(\mu_n - \mu)\right| + \limsup_{n\to\infty} C\int_{\mathrm{d}(x_0,x)\geq R} \mathrm{d}(x_0, x)^r\, d(\mu_n + \mu)(x) \\ &= \limsup_{n\to\infty} C\int_{\mathrm{d}(x_0,x)\geq R} \mathrm{d}(x_0, x)^r\, d(\mu_n + \mu)(x).\end{aligned}$$

Mandando poi $R \to \infty$ l'ultimo termine ottenuto va a 0 per la proprietà 2, questo termina la dimostrazione del fatto che 2 implichi 4. Proviamo ora che 3 implica 2, nel corso della dimostrazione usiamo la notazione $a \wedge b := \min\{a, b\}$. Per la convergenza debole, siccome fissato $x_0 \in X$ la finzione $x \mapsto \mathrm{d}(x_0, x)$ è continua, abbiamo che per ogni $R > 0$ vale:

$$\int_X \big[\mathrm{d}(x_0, x) \wedge R\big]^r\, d\mu_n(x) \xrightarrow[n\to\infty]{} \int_X \big[\mathrm{d}(x_0, x) \wedge R\big]^r\, d\mu(x),$$

d'altra parte, per il teorema di convergenza monotona:

$$\lim_{R\to\infty} \int_X \big[\mathrm{d}(x_0, x) \wedge R\big]^r\, d\mu(x) = \int_X \mathrm{d}(x_0, x)^r\, d\mu(x),$$

infine per 3. vale:

$$\int_X \mathrm{d}(x_0, x)^r\, d\mu_n(x) \xrightarrow[n\to\infty]{} \int_X \mathrm{d}(x_0, x)^r\, d\mu(x).$$

Quindi possiamo concludere che:

$$\begin{aligned}
&\lim_{R\to\infty}\lim_{n\to\infty}\int_X \left[\mathrm{d}(x_0,x)^r - \left(\mathrm{d}(x_0,x)\wedge R\right)^r\right] d\mu_n(x)\\
&= \lim_{R\to\infty}\lim_{n\to\infty}\int_X \mathrm{d}(x_0,x)^r\, d\mu_n(x)\\
&- \lim_{R\to\infty}\lim_{n\to\infty}\int_X \left[\mathrm{d}(x_0,x)\wedge R\right]^r d\mu_n(x)\\
&= \lim_{n\to\infty}\int_X \mathrm{d}(x_0,x)^r\, d\mu_n(x) - \lim_{R\to\infty}\int_X \left[\mathrm{d}(x_0,x)\wedge R\right]^r d\mu(x)\\
&= \int_X \mathrm{d}(x_0,x)^r\, d\mu(x) - \int_X \mathrm{d}(x_0,x)^r\, d\mu(x) = 0.
\end{aligned}$$

Osservando ora che quando $\mathrm{d}(x_0,x)\geq 2R$ si ha:

$$\mathrm{d}(x_0,x)^r - R^r \geq (1-2^{-r})\mathrm{d}(x_0,x)^r,$$

segue dall'uguaglianza sopra che:

$$\lim_{R\to\infty}\limsup_{n\to\infty}\int_{\mathrm{d}(x_0,x)\geq 2R} \mathrm{d}(x_0,x)^r\, d\mu_n(x) = 0,$$

che è esattamente la 2 come si voleva. Questo termina la dimostrazione che 2, 3 e 4 sono equivalenti. Rimane dunque da mostrare che 1 è equivalente ad una qualsiasi tra le affermazioni 2, 3 o 4. Mostriamo che 1 è equivalente a 3. A tale scopo notiamo che se μ_n converge debolmente a μ per semicontinuità inferiore vale:

$$\int_X \mathrm{d}(x_0,x)^r\, d\mu(x) \leq \liminf_{n\to\infty}\int_X \mathrm{d}(x_0,x)^r\, d\mu_n(x),$$

quindi la proprietà di convergenza dei momenti enunciata in 3 è equivalente a mostrare che:

$$\limsup_{n\to\infty}\int_X \mathrm{d}(x_0,x)^r\, d\mu_n(x) \leq \int_X \mathrm{d}(x_0,x)^r\, d\mu(x). \tag{1.18}$$

Mostriamo inanzitutto che la convergenza rispetto a $\mathcal{W}_r$ implica la (1.18). Per il seguente fatto: per ogni $\varepsilon>0$ esiste una costante $C_\varepsilon>0$ tale che per ogni coppia di numeri reali non negativi a,b si ha:

$$(a+b)^r \leq (1+\varepsilon)a^r + C_\varepsilon b^r.$$

Combinando questa disuguaglianza con la disuguaglianza triangolare otteniamo che per ogni scelta di $x_0,x,y\in X$ si ha:

$$\mathrm{d}(x_0,x)^r \leq (1+\varepsilon)\mathrm{d}(x_0,y)^r + C_\varepsilon \mathrm{d}(x,x)^r.$$

Sia ora $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{P}_r(X,\mathrm{d})$ tale che: $\mathcal{W}_r(\mu_n,\mu) \xrightarrow[n\to\infty]{} 0$. Per ogni $n\in\mathbb{N}$ definiamo π_n come la misura che realizza l'ottimo per $\mathcal{W}_r(\mu_n,\mu)$. Intengrando allora l'ultima disuguaglianza rispetto a π_n e usando la proprietà sulle marginali otteniamo:

$$\begin{aligned}\int_X \mathrm{d}(x_0,x)^r\,d\mu_n(x) &\le (1+\varepsilon)\int_X \mathrm{d}(x_0,y)^r\,d\mu(y) + C_\varepsilon \int_X \mathrm{d}(x,y)^r\,d\pi_n(x,y)\\ &= (1+\varepsilon)\int_X \mathrm{d}(x_0,y)^r\,d\mu(y) + \mathcal{W}_r(\mu_n,\mu)^r.\end{aligned}$$

da cui passando al limite $n\to\infty$ otteniamo, usando l'ipotesi 1,

$$\limsup_{n\to\infty}\int_X \mathrm{d}(x_0,x)^r\,d\mu_n(x) \le (1+\varepsilon)\int_X \mathrm{d}(x_0,y)^r\,d\mu(y)$$

da cui mandando $\varepsilon\to 0$ otteniamo esattamente la (1.18) come voluto. A questo punto per dimostrare il teorema rimane da verificare che la convergenza in distanza di Wasserstein implica la convergenza debole (questo concluderà la dimostrazione del fatto che 1 implica 3) e che 3 implica 1. Osserviamo che per farlo possiamo assumere che d sia limitata, se non lo fosse infatti potremmo costruire una metrica limitata (da 1) $\tilde{\mathrm{d}}$ da d ponendo: $\tilde{\mathrm{d}} := \min\{\mathrm{d},1\}$ e costruire da $\tilde{\mathrm{d}}$ la distanza di Wassertein associata: $\tilde{\mathcal{W}}_r$, la quale certamente verificherà: $\mathcal{W}_r \ge \tilde{\mathcal{W}}_r$. Quindi la convergenza in metrica $\mathcal{W}_r$ implicherebbe quella in metrica $\tilde{\mathcal{W}}_r$ e per controllare che la convergenza in metrica $\tilde{\mathcal{W}}_r$ implichi quella debole, sarebbe sufficiente farlo per $\tilde{\mathcal{W}}_r$. Viceversa supponiamo che sia soddisfatta l'affermazione 3 e che μ_n converga rispetto a $\tilde{\mathcal{W}}_r$, mostriamo che effettivamente allora μ_n converge anche rispetto a $\mathcal{W}_r$. Usiamo la disuguaglianza:

$$\mathrm{d}(x,y) \le \mathrm{d}(x,y)\wedge R + 2\mathrm{d}(x,x_0)\mathbf{1}_{\mathrm{d}(x,x_0)\ge R/2} + 2\mathrm{d}(y,x_0)\mathbf{1}_{\mathrm{d}(y,x_0)\ge R/2},$$

in particolare la sua generalizzazione:

$$\mathrm{d}(x,y)^r \le C_r\Big(\big[\mathrm{d}(x,y)\wedge R\big]^r + 2\mathrm{d}(x,x_0)^r\mathbf{1}_{\mathrm{d}(x,x_0)\ge R/2} + 2\mathrm{d}(y,x_0)^r\mathbf{1}_{\mathrm{d}(y,x_0)\ge R/2}\Big),$$

valide per ogni $x,y\in X, R>0$ e dove $C_r>0$ è una costante dipendente solo da $r\ge 1$. Definiamo, per ogni $n\in\mathbb{N}$, π_n come la misura per cui si ottiene l'ottimo in $\mathcal{W}(\mu_n,\mu)$. Allora dalla disuguaglianza precedente, se $R\ge 1$ abbiamo:

$$\begin{aligned}\mathcal{W}_r(\mu_n,\mu)^r &= \iint_{X\times X}\mathrm{d}(x,y)^r\,d\pi_n(x,y) \le C_r\iint_{X\times X}\big[\mathrm{d}(x,y)\wedge R\big]^r\,d\pi_n(x,y)\\ &+ C_r\iint_{\mathrm{d}(x,x_0)\ge R/2}\mathrm{d}(x,x_0)^r\,d\pi_n(x,y) + C_r\iint_{\mathrm{d}(y,x_0)\ge R/2}\mathrm{d}(y,x_0)^r\,d\pi_n(x,y)\\ &= R^r\mathcal{W}_r(\mu_n,\mu)^r + C_r\int_{\mathrm{d}(x,x_0)\ge R/2}\mathrm{d}(x,x_0)^r\,d\mu_n(x)\\ &+ C_r\int_{\mathrm{d}(y,x_0)\ge R/2}\mathrm{d}(y,x_0)^r\,d\mu(y).\end{aligned}$$

Da cui concludiamo che vale 1, mandando $n \to \infty$, poi $R \to \infty$ e usando la proprietà 2, che è vera siccome abbiamo supposto valesse 3. Possiamo quindi assumere che d sia limitata, questo implica che tutte le distanze $\{\mathcal{W}_r\}_{r\geq 1}$ sono equivalenti. Quindi è sufficiente mostrare che 1 implica la convergenza debole e che 3 implica 1 solo per il caso $r = 1$. Allora usando il teorema di Kantorovich-Rubenstein la affermazione 1, è equivalente a:

$$\sup_{\substack{\varphi\in L^1(|\mu_n-\mu|)\\ \|\varphi\|_{\mathrm{Lip}}\leq 1}} \int_X \varphi\, d(\mu_n - \mu) \xrightarrow[n\to\infty]{} 0. \tag{1.19}$$

Supponiamo allora che $\mathcal{W}_1(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0$, mostriamo che $\mu_n \xrightarrow[n\to\infty]{} \mu$ (debolmente). Dobbiamo cioè mostrare che per ogni $\varphi \in \mathcal{C}_b(X)$ vale:

$$\int_X \varphi\, d\mu_n \xrightarrow[n\to\infty]{} \int_X \varphi\, d\mu. \tag{1.20}$$

Dalla (1.19) sappiamo che la (1.20) è vera se la funzione φ considerata è 1-Lipschitziana. Rimpiazzando φ con $\varphi/\|\varphi\|_{\mathrm{Lip}}$ otteniamo che ciò è vero anche solo se φ è Lipschitziana. Concludiamo la dimostrazione ricordando che in uno spazio metrico, ogni funzione continua e limitata, può essere approssimata da sotto e sopra da funzioni Lipschitziane. Più precisamente esistono due successioni di funzioni Lipschtziane $(a_k)_{k\in\mathbb{N}}, (b_k)_{k\in\mathbb{N}}$, uniformemente limitate, e tali che a_k sia crescente, b_k sia decrescente e per ogni $x \in X$ vale:

$$\lim_{k\to\infty} a_k(x) = \lim_{k\to\infty} b_k(x) = \varphi(x).$$

Ma allora:

$$\limsup_{n\to\infty} \int_X \varphi\, d\mu_n \leq \liminf_{k\to\infty} \limsup_{n\to\infty} \int_X b_k\, d\mu_n = \liminf_{k\to\infty} \int_X b_k\, d\mu = \int_X \varphi\, d\mu,$$

ove l'ultima equazione segue dal teorema di convergenza dominata. Similmente:

$$\liminf_{n\to\infty} \int_X \varphi\, d\mu_n \geq \limsup_{k\to\infty} \liminf_{n\to\infty} \int_X a_k\, d\mu_n = \limsup_{k\to\infty} \int_X a_k\, d\mu = \int_X \varphi\, d\mu.$$

Questo dimostra che la (1.20) vale per tutte le funzioni $\varphi \in \mathcal{C}_b(X)$ e quindi $\mu_n \xrightarrow[n\to\infty]{} \mu$ (debolmente), come voluto. Viceversa supponiamo che $\mu_n \xrightarrow[n\to\infty]{} \mu$ (debolmente), proviamo la (1.19). Per farlo, fissiamo $x_0 \in X$ e sia $\mathrm{Lip}_{1,x_0}(X)$ lo spazio di tutte le funzioni Lipschitziane φ su X, con costante di Lipschitz al più 1, e tali che $\varphi(x_0) = 0$. Allora è sufficiente mostrare che:

$$\sup_{\varphi\in \mathrm{Lip}_{1,x_0}(X)} \int_X \varphi\, d(\mu_n - \mu) \xrightarrow[n\to\infty]{} 0. \tag{1.21}$$

Ora la (1.21) implica la (1.19) poichè ad ogni elemento $\varphi \in \mathrm{Lip}(X)$ con $\|\varphi\|_{\mathrm{Lip}} \leq 1$, possiamo associare un elemento $\psi \in \mathrm{Lip}_{1,x_0}$ che verifica: $\int_X \varphi\, d(\mu_n - \mu) =$

$\int_X \psi\, d(\mu_n - \mu)$ ponendo: $\psi = \varphi - \varphi(x_0)$. Ora consideriamo la famiglia di misure $\{\mu_n \mid n \in \mathbb{N}\} \cup \{\mu\}$. Essa è ovviamente compatta debolmente. Quindi dal teorema di Prohorov essa è uniformemente serrata e pertanto esiste una successione di insiemi compatti $(K_m)_{m\geq 1}$ tale che per ogni $m \geq 1$, si ha:

$$\sup_{n\in\mathbb{N}} \mu_n(K_m^c) \leq \frac{1}{m} \quad \text{e} \quad \mu(K_m^c) \leq \frac{1}{m}.$$

Senza perdere di generalità assumiamo che $x_0 \in K_1$. Allora, per ogni $m \geq 1$, l'insieme:

$$C_m := \{\varphi_n \mathbf{1}_{K_m} \mid \varphi \in \mathrm{Lip}_{1,x_0}(X)\}$$

è un sottoinsieme di $\mathrm{Lip}_{1,x_0}(K_m)$. In particolare C_m è equilimitato perchè d è limitata ed è equicontinuo siccome è formato da funzioni lipschitziane tutte con costante di Lipschitz al più 1. Dal teorema di ascoli Arzelà segue quindi che C_m è un sottoinsieme relativamente compatto di $\mathcal{C}_b(K_m)$ (dotato della norma della convergenza uniforme). Osserviamo che però C_m è anche chiuso, e quindi compatto. Infatti se $(\varphi_n)_{n\in\mathbb{N}}$ è una successione in C_m che converge a $\varphi \in \mathcal{C}_b(K_m)$ allora per convergenza uniforme:

$$\varphi(x_0) = \lim_{x\to x_0} \varphi(x) = \lim_{x\to x_0} \lim_{n\to\infty} \varphi_n(x) = \lim_{n\to\infty} \lim_{x\to x_0} \varphi_n(x) = 0$$

e similmente dati $x, y \in K_m$ si ha:

$$|\varphi(x) - \varphi(y)| = \lim_{n\to\infty} |\varphi_n(x) - \varphi_n(y)| \leq \mathrm{d}(x, y),$$

da cui anche φ è Lipschitziana con $\|\varphi\|_{\mathrm{Lip}} \leq 1$, e dunque $\varphi \in C_m$. Quindi da ogni succhessione in C_m è possibile estrarre una sottosuccessione che converge uniformemente su K_m. Con un argomento diagonale otteniamo che da ogni successione $(\varphi_n)_{n\in\mathbb{N}} \subseteq \mathrm{Lip}_{1,x_0}(X)$ possiamo estrarre una sottosuccessione che converge uniformemente su ogni K_m a qualche funzione misurabile φ_∞, definita su $S = \bigcup_{m\geq 1} K_m$, che sarà limitata e Lipschitziana, perchè la successione $(\varphi_n)_{n\in\mathbb{N}}$ è uniformemente limitata e uniformemente Lipschitziana.

Applichiamo questo risultato ad una successione $(\varphi_n)_{n\in\mathbb{N}}$ che soddisfa:

$$\sup_{\varphi\in\mathrm{Lip}_{1,x_0}} \int_X \varphi\, d(\mu_n - \mu) \leq \int_X \varphi_n\, d(\mu_n - \mu) + \frac{1}{n}.$$

Esiste dunque una sottosuccessione $(\varphi_{n_h})_{h\in\mathbb{N}}$, che converge uniformemente su ogni K_m, ad una funzione 1-Lipschitziana φ_∞ su $S = \bigcup_{m\geq 1} K_m$. Ora usando il teorema di estensione per le funzioni Lipschitziane otteniamo che φ_∞ si può estendere ad un elemento di $\mathrm{Lip}_{1,x_0}(X)$, in particolare quindi φ_∞ è continua e limitata (perchè d lo è). Per concludere la dimostrazione è sufficiente mostrare che:

$$\int_X \varphi_n\, d(\mu_n - \mu) \xrightarrow[n\to\infty]{} 0.$$

Per farlo scriviamo:

$$\left|\int_X \varphi_n \, d(\mu_n - \mu)\right| \leq \left|\int_{K_m} (\varphi_n - \varphi_\infty) \, d(\mu_n - \mu)\right| + \left|\int_{K_m^c} (\varphi_k - \varphi_\infty) \, d(\mu_n - \mu)\right| + \left|\int_X \varphi_\infty \, d(\mu_n - \mu)\right|.$$

Ed ora è necessario osservare che, per m fissato il primo termine a destra va a 0 per $n \to \infty$, per convergneza uniforme. Poi siccome sia φ che φ_∞ sono uniformemente limitate, il secondo termine è stimato da $C\big(\mu_n(K_m^c) + \mu(K_m^c)\big) \leq 2C/m$, quindi va a zero per $m \to \infty$, uniformemente in n. L'ultimo termine converge a zero per $n \to \infty$ per convergenza debole di μ_n a μ. Quindi possiamo concludere la dimostrazione mandando prima $n \to \infty$ e dopo $m \to \infty$. □

Osservazione 1.7. Si noti che se d è limitata allora la condizione di serratezza presente nell'affermazione 2 del teorema è triviale, quindi $\mathcal{W}_r$ metricizza la topologia debole* su $\mathcal{P}(X)$, ma poichè si può sempre sostituire d con una metrica limitata equivalente otteniamo come corollario del teorema che la topologia debole * su $\mathcal{P}(X)$ è sempre metrizzabile, quando X soddisfa le ipotesi del teorema.

Teorema 1.12. *Sia* (X, d) *uno spazio Polacco e* $r \geq 1$. *Allora:* $(\mathcal{P}_r(X, \mathrm{d}), \mathcal{W}_r)$ *è uno spazio Polacco.*

Dimostrazione. Cominciamo con la prova della separabilità, a tale scopo siccome (X, d) è separabile, esiste $(x_n)_{n\in\mathbb{N}} \subseteq X$ una successione densa. Definiamo per ogni $N \in \mathbb{N}$ l'insieme:

$$L_N := \left\{ \sum_{n=0}^{N} b_n \delta_{x_n} \mid b_n \in \mathbb{Q}, b_n \geq 0 \quad \text{e} \quad \sum_{n=0}^{N} b_n = 1 \right\},$$

esso è ovviamente numerabile, quindi posto:

$$L := \bigcup_{N\in\mathbb{N}} L_N$$

anche L è numerabile. Proviamo che L è denso in $(\mathcal{P}_r(X, \mathrm{d}), \mathcal{W}_r)$. Sia $\mu \in \mathcal{P}_r(X, \mathrm{d})$, procediamo in due passi:

1. Approssimiamo μ con una misura μ_1 della forma:

$$\mu_1 := \sum_{n=1}^{\infty} a_n \delta_{x_n},$$

ove gli $a_n \in \mathbb{R}, a_n \geq 0$ verificano: $\displaystyle\sum_{n=1}^{\infty} a_n = 1$;

2. Approssimiamo μ_1 con una misura $\mu_2 \in L$, e concludiamo con la disuguaglianza triangolare.

Per mostrare 1. osserviamo che fissato $\varepsilon > 0$, siccome gli x_n sono densi, X è ricoperto dagli insiemi:

$$B_n := B(x_n, \varepsilon),$$

e quindi partizionato dagli insiemi:

$$\tilde{B}_n := B_n \setminus \bigcup_{k \leq n-1} B_k.$$

Ma allora ponendo $a_n := \mu(\tilde{B}_n) \geq 0$, per ogni $n \in \mathbb{N}$, abbiamo che gli a_n hanno somma unitaria. Definiamo allora:

$$\mu_1 := \sum_{n=1}^{\infty} a_n \delta_{x_n}.$$

Stimiamo la distanza tra μ e μ_1, a tale scopo osserviamo che per ogni $n \in \mathbb{N}$ si ha (dalla proposizine 1.9):

$$\begin{aligned}
\mathcal{W}_r(\mu, \delta_{x_n})^r &= \int_X \mathrm{d}(x, x_n)^r \, d\mu(x) \\
&= \sum_{n=1}^{\infty} \int_{\tilde{B}_n} \mathrm{d}(x, x_n)^r \, d\mu(x) \\
&< \sum_{n=1}^{\infty} a_n \varepsilon^r = \varepsilon^r,
\end{aligned}$$

conseguentemente usando la convessità di $\mathcal{W}_r$:

$$\begin{aligned}
\mathcal{W}_r(\mu, \mu_1)^r &= \mathcal{W}_r(\mu, \sum_{n=1}^{\infty} a_n \delta_{x_n}) \\
&\leq \sum_{n=1}^{\infty} a_n \mathcal{W}_r(\mu, \delta_{x_n})^r \\
&< \varepsilon^r \sum_{n=1}^{\infty} a_n = \varepsilon^r.
\end{aligned}$$

Da cui $\mathcal{W}_r(\mu, \mu_1) < \varepsilon$.

Passiamo alla prova di 2, per prima cosa notiamo che μ_1 appartiene a $\mathcal{P}_r(X, \mathrm{d})$ siccome è a distanza di Wasserstein finita da μ. Segue che allora anche la distanza tra μ_1 e δ_{x_1} è finita, siccome ovviamente anche δ_{x_1} sta in $\mathcal{P}_r(X, \mathrm{d})$, ossia:

$$\mathcal{W}_r(\mu_1, \delta_{x_1})^r = \sum_{n=1}^{\infty} a_n \mathrm{d}(x_n, x_1)^r < +\infty.$$

In particolare dunque, considerando $\epsilon > 0$ fissato in precedenza, esiste $N \in \mathbb{N}$ tale che:

$$\sum_{n=N+1}^{\infty} a_n \mathrm{d}(x_n, x_1)^r < \varepsilon^r.$$

Ora, per ogni $2 \leq n \leq N$, scegliamo dei numeri $b_n \in \mathbb{Q}$ non negativi tali che:

$$0 \leq a_n - b_n \leq \frac{\varepsilon^r}{\sum_{j=1}^{N} a_j \mathrm{d}(x_j, x_1)^r} a_n,$$

definiamo poi b_1 come:

$$b_1 := a_1 + \sum_{n=2}^{N} (a_n - b_n) + \sum_{n=N+1}^{\infty} a_n,$$

notiamo che $b_1 \in \mathbb{Q}$, siccome i b_n sono razionali e gli a_n hanno somma 1. In particolare la definizione di b_1 implica che anche i b_n hanno somma 1. Definiamo:

$$\mu_2 := \sum_{n=1}^{N} b_n \delta_{x_n}$$

e stimiamo $\mathcal{W}_r(\mu_1, \mu_2)$. Per farlo consideriamo uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ e su di esso costruiamo due variabili aleatorie Y e Z, a valori in (X, d) e tali che $Y \sim \mu_1$ e $Z \sim \mu_2$. A tale scopo siano $(A_n)_{n \in \mathbb{N}} \subseteq \mathcal{F}$ eventi disgiunti tali che $\mathbb{P}(A_n) = a_n$ per ogni $n \in \mathbb{N}$. Allora

$$Y := \sum_{n=1}^{\infty} x_n \mathbf{1}_{A_n} \sim \mu_1.$$

Allo stesso modo consideriamo $B_n \subseteq A_n$ per $n = 1, \ldots, N$ tali che $\mathbb{P}(B_n) = b_n$ e definiamo

$$Z := \sum_{n=1}^{N} x_n \mathbf{1}_{B_n} + x_1 \mathbf{1}_{\left(\bigcup_{n=1}^{N} B_n\right)^c},$$

allora $Z \sim \mu_2$. Segue

$$\begin{aligned}
\mathcal{W}_r(\mu_1, \mu_2)^r &\leq \mathbb{E}^{\mathbb{P}}\left[\mathrm{d}(Y, Z)^r\right] \\
&= \sum_{n=1}^{N} (a_n - b_n) \mathrm{d}(x_n, x_1)^r + \sum_{n=N+1}^{\infty} a_n \mathrm{d}(x_n, x_1)^r \\
&\leq \sum_{n=1}^{N} \frac{\varepsilon^r}{\sum_{j=1}^{N} a_j \mathrm{d}(x_j, x_1)^r} a_n \mathrm{d}(x_n, x_1)^r + \sum_{n=N+1}^{\infty} a_n \mathrm{d}(x_n, x_1)^r \\
&< \varepsilon^r + \varepsilon^r = 2\varepsilon^r.
\end{aligned}$$

Da cui:

$$\mathcal{W}_r(\mu_1, \mu_2) < 2^{1/r} \varepsilon.$$

Quindi dalla disuguaglianza triangolare:

$$\mathcal{W}_r(\mu,\mu_2) \le \mathcal{W}_r(\mu,\mu_1) + \mathcal{W}_r(\mu_1,\mu_2) < \varepsilon + 2^{1/r}\varepsilon = (1+2^{1/r})\varepsilon.$$

Questo termina la dimostrazione della separabilità.

Passiamo alla prova della completezza, sia $(\mu_n)_{n\in\mathbb{N}}$ una successione di Chauchy in $(\mathcal{P}_r(X,\mathrm{d}),\mathcal{W}_r)$. Per provare che converge procediamo, come sopra, in due passi:

1. Proviamo inanzitutto che $(\mu_n)_{n\in\mathbb{N}}$ è uniformemente serrata.
2. Deduciamo dal passo 1 che μ_n converge, rispetto a $\mathcal{W}_r$.

Per dimostrare 1, osserviamo che $(\mu_n)_{n\in\mathbb{N}}$ è di Cauchy rispetto a $\mathcal{W}_1$, siccome $\mathcal{W}_1 \le \mathcal{W}_r$. Fissato $\varepsilon > 0$, esiste allora $N \in \mathbb{N}$, tale che per ogni $n \ge N$ si ha: $\mathcal{W}_r(\mu_n,\mu_N) < \varepsilon^2$, così che, per ogni $n \in \mathbb{N}$, esiste $j \le N$ tale che:

$$\mathcal{W}_1(\mu_n,\mu_j) < \varepsilon^2. \tag{1.22}$$

Ora la famiglia finita $(\mu_j)_{j\le N}$ è uniformemente serrata, infatti per ogni $j \le N$ esiste un compatto K_j, tale che:

$$\mu_j(K_j^c) < \varepsilon \quad \text{o equivalentemente} \quad \mu_j(K_j) > 1-\varepsilon.$$

Definendo allora $K = \bigcup_{j=1}^N K_j$, abbiamo che K è chiuso (unione finita di chiusi) quindi completo (sottoinsieme chiuso di un completo) e totalmente limitato (perchè tali sono i K_j), quindi dal teorema di Heine-Borel K è compatto e verifica:

$$\mu_j(K) \ge \mu_j(K_j) > 1-\varepsilon \quad \forall j \le N.$$

Per la totale limitatezza, esistono inoltre $x_1,\dots,x_q \in K$, tali che: $U := \bigcup_{k=1}^q B(x_k,\varepsilon)$ verifica:

$$\mu_j(U) > 1-\varepsilon \quad \forall j \le N. \tag{1.23}$$

Sia poi $\phi\colon X \to [0,+\infty)$ la funzione definita da:

$$\phi(x) := \Big(1 - \frac{\mathrm{d}(x,U)}{\varepsilon}\Big)^+.$$

Notiamo allora che ϕ è $\frac{1}{\varepsilon}$-Lipschitziana, infatti, distiguendo i casi:

1. Se $x,y \in U$, allora $|\phi(x)-\phi(y)| = 0$, e la Lipschitzianità è ovvia.
2. Se $x \in U, y \notin U$ allora:

$$\begin{aligned} |\phi(x)-\phi(y)| &= \Big|1 - \Big(1 - \frac{\mathrm{d}(y,U)}{\varepsilon}\Big)^+\Big| \\ &= \begin{cases} 1 \le \dfrac{1}{\varepsilon}\mathrm{d}(x,y) & \text{se } \mathrm{d}(y,U) \ge \varepsilon, \\ \dfrac{\mathrm{d}(y,U)}{\varepsilon} \le \dfrac{1}{\varepsilon}\mathrm{d}(x,y) & \text{se } \mathrm{d}(y,U) < \varepsilon. \end{cases} \end{aligned}$$

Un ragionamento analogo funziona per il caso: $x \notin U, y \in U$.

3. se $x, y \notin U$ allora dobbiamo distinguere tre sottocasi:
   (a) se vale: $\mathrm{d}(x,U), \mathrm{d}(y,U) \geq \varepsilon$, allora $|\phi(x) - \phi(y)| = 0$ e la Lipschitzianità è ovvia,
   (b) se vale: $\mathrm{d}(x,U), \mathrm{d}(y,U) \leq \varepsilon$, allora:
   $$|\phi(x) - \phi(y)| = \frac{|\mathrm{d}(y,U) - \mathrm{d}(x,U)|}{\varepsilon} \leq \frac{\mathrm{d}(x,y)}{\varepsilon}$$
   ove la seconda disuguaglianza segue dal fatto che $\mathrm{d}(y,U) \leq \mathrm{d}(x,y) + \mathrm{d}(x,U)$;
   (c) se invece vale che: $\mathrm{d}(x,U) < \varepsilon, \mathrm{d}(y,U) \geq \varepsilon$ allora:
   $$\begin{aligned} |\phi(x) - \phi(y)| &= \left|1 - \frac{d(x,U)}{\varepsilon}\right| \\ &\leq \frac{1}{\varepsilon}\Big(\mathrm{d}(y,U) - \mathrm{d}(x,U)\Big) \leq \frac{\mathrm{d}(x,y)}{\varepsilon} \end{aligned}$$
   ove l'ultima stima è ottenuta ragionando come sopra, il caso: $\mathrm{d}(x,U) \geq \varepsilon, \mathrm{d}(y,U) < \varepsilon$ si ottiene in modo analogo.

Dati allora $j, n \in \mathbb{N}$ come in (1.22) e (1.23), se $\pi \in \Pi(\mu_j, \mu_n)$, allora:

$$\begin{aligned} \int_X \phi(x)\, d\mu_j(x) - \int_X \phi(y)\, d\mu_n(y) &= \iint_{X\times X} \big[\phi(x) - \phi(y)\big]\, d\pi(x,y) \\ &\leq \frac{1}{\varepsilon} \iint_{X\times X} \mathrm{d}(x,y)\, d\pi(x,y), \end{aligned}$$

da cui:

$$\int_X \phi(x)\, d\mu_j(x) - \int_X \phi(y)\, d\mu_n(y) \leq \frac{1}{\varepsilon}\mathcal{W}_1(\mu_j, \mu_n).$$

D'altra parte si ha anche: $\mathbf{1}_U \leq \phi \leq \mathbf{1}_{U^\varepsilon}$, ove:

$$U^\varepsilon := \{x \in X \mid \mathrm{d}(x,U) < \varepsilon\}$$

dunque:

$$\int_X \phi(x)\, d\mu_j(x) \geq \mu_j(U) \qquad \text{e} \qquad \int_X \phi(y)\, d\mu_n(y) \leq \mu_n(U^\varepsilon).$$

Conseguentemente:

$$\mu_n(U^\varepsilon) \geq \mu_j(U) - \frac{1}{\varepsilon}\mathcal{W}_1(\mu_j, \mu_n) \tag{1.24}$$

Notiamo ora che: $U^\epsilon \subseteq \bigcup_{k=1}^q B(x_k, 2\varepsilon)$, quindi usando (1.22) (1.23) e (1.24) otteniamo:

$$\begin{aligned} \mu_n\Big(X \setminus \bigcup_{k=1}^q B(x_k, 2\varepsilon)\Big) &\leq \mu_n(X \setminus U^\varepsilon) \\ &= 1 - \mu_n(U^\varepsilon) \leq 1 - \mu_j(U) + \frac{1}{\varepsilon}\mathcal{W}_1(\mu_j, \mu_n) \\ &< 1 - 1 + \varepsilon + \varepsilon = 2\varepsilon. \end{aligned}$$

Da questo deduciamo che sostituendo ε con $\varepsilon 2^{-m-1}$, dove m è un qualsiasi intero positivo, esistono $q(m)$ punti $x_1^m, \dots, x_{q(m)}^m \in X$ tali che:

$$\mu_n\Big(X \setminus \bigcup_{k=1}^{q(m)} B(x_k^m, \varepsilon 2^{-m})\Big) < \varepsilon 2^{-m}$$

per ogni $n \in \mathbb{N}$. In particolare l'insieme:

$$S := \bigcap_{m=1}^{+\infty} \bigcup_{k=1}^{q(m)} B(x_k^m, \varepsilon 2^{-m})$$

verifica

$$\mu_n(X \setminus S) \leq \sum_{m=1}^{+\infty} \mu_n\Big(X \setminus \bigcup_{k=1}^{q(m)} B(x_k^m, \varepsilon 2^{-m})\Big) < \sum_{m=1}^{+\infty} \varepsilon 2^{-m} = \varepsilon$$

per ogni $n \in \mathbb{N}$. È sufficiente infine osservare che, per ogni $\eta > 0$, scelto $m \in \mathbb{N}$ tale che $\varepsilon 2^{-m} \leq \eta$, l'insieme S è ricoperto dalle $q(m)$ palle $B(x_k^m, \varepsilon 2^{-m})$ le quali hanno raggio minore di η, ossia S è totalmente limitato, segue che la sua chiusura $\overline{S}$ è compatta in X. Ma allora $\overline{S}$ verifica:

$$\mu_n(X \setminus \overline{S}) \leq \mu_n(X \setminus S) < \varepsilon \quad \forall n \in \mathbb{N}.$$

Quindi $(\mu_n)_{n\in\mathbb{N}}$ è uniformemente serrata come si voleva.

Ci concentriamo ora sul passo 2, mostriamo come la uniforme serratezza della successione $(\mu_n)_{n\in\mathbb{N}}$ ne implichi la convergenza. Dal teorema di Prohorov sappiamo che esiste una sottosuccessione: $(\mu_{n_k})_{k\in\mathbb{N}}$ che converge debolmente ad una misura di probabilità μ su X. Mostriamo che:

$$\mathcal{W}_r(\mu, \mu_{n_k}) \xrightarrow[k\to\infty]{} 0, \tag{1.25}$$

da cui seguirà la convergenza dell'intera successione, infatti se tale convergenza è vera, per ogni $\varepsilon > 0$ esistono $n, k \in \mathbb{N}$ sufficientemente grandi tali che:

$$\mathcal{W}_r(\mu, \mu_n) \leq \mathcal{W}_r(\mu, \mu_{n_k}) + \mathcal{W}_r(\mu_{n_k}, \mu_n) < \varepsilon + \varepsilon = 2\varepsilon,$$

e si conclude per l'arbitrarietà di $\varepsilon > 0$. Per provare la (1.25), dati $k, k' \in \mathbb{N}$ sia $\pi_{n_k n_{k'}}$ che realizza il minimo per $\mathcal{W}_r(\mu_{n_k}, \mu_{n_{k'}})$. Ora la successione $(\mu_{n_k})_{k\in\mathbb{N}}$ è uniformemente stirata, dunque lo è anche $(\pi_{n_k n_{k'}})_{k\in\mathbb{N}}$, per $k' \in \mathbb{N}$ fissato. Applicando di nuovo il teorema di Prohorov segue che esiste una ulteriore sottosuccessione $(\pi_{n_{k_h} n_{k'}})_{h\in\mathbb{N}}$ di $(\pi_{n_k n_{k'}})_{k\in\mathbb{N}}$ convergente debolmente ad una misura di probabilità $\pi_{n_{k'}}$ su $X \times X$. Ma allora per semicontinuità inferiore della convergenza debole:

$$\begin{aligned} \iint_{X\times Y} \mathrm{d}(x,y)^r \, d\pi_{n_{k'}}(x,y) &\leq \liminf_{h\to\infty} \iint_{X\times X} \mathrm{d}(x,y)^r \, d\pi_{n_{k_h} n_{k'}}(x,y) \\ &= \liminf_{h\to\infty} \mathcal{W}_r(\mu_{n_{k_h}}, \mu_{n_{k'}})^r. \end{aligned} \tag{1.26}$$

Ora però è noto che $\pi_{n_{k_h} n_{k'}}$ ha marginali $\mu_{n_{k_h}}$ e $\mu_{n_{k'}}$, quindi il limite per $h \to \infty$, cioè $\pi_{n_{k'}}$ ha marginali μ e $\mu_{n_{k'}}$ rispettivamente. Dunque:

$$\mathcal{W}_r(\mu, \mu_{n_{k'}}) \leq \iint_{X \times Y} \mathrm{d}(x,y)^r \, d\pi_{n_{k'}}(x,y) \tag{1.27}$$

per ogni $k' \in \mathbb{N}$. D'altra parte la successione $(\mu_{n_k})_{k \in \mathbb{N}}$ come già osservato è anch'essa di Cauchy per $\mathcal{W}_r$, quindi per ogni $\varepsilon > 0$ fissato e $h, k' \in \mathbb{N}$ sufficientemente grandi si ha:

$$\mathcal{W}_r(\mu_{n_{k_h}}, \mu_{n_{k'}}) < \varepsilon. \tag{1.28}$$

Segue allora da (1.26) (1.27) e (1.28) che:

$$\mathcal{W}_r(\mu, \mu_{n_{k'}}) < \varepsilon$$

per $k' \in \mathbb{N}$ sufficientemente grande. Da cui per arbitrarietà di $\varepsilon > 0$ segue la (1.25) come voluto, e si conclude. □

1.4 Il caso discreto

Spesso nelle applicazioni le misure di probabilità considerate sono discrete finite, ossia combinazioni convesse finite di misure di Dirac. In questa sezione vediamo cosa accade al problema di minimo che definisce $\mathcal{W}_r$ quando entrambe le misure sono della forma appena menzionata.

Teorema 1.13. *Sia (X, d) uno spazio Polacco, $r \geq 1$ e $\mu = \sum_{i=1}^n \mu_i \delta_{x_i}, \nu = \sum_{j=1}^m \nu_j \delta_{y_j}$ due misure di probabilità discrete su X. Allora la distanza $\mathcal{W}_r(\mu,\nu)^r$ eguaglia il valore ottimo del seguente Programma Lineare:*

$$\begin{aligned}
&\textit{MIN (in } \pi) && \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} \mathrm{d}_{ij}^r \\
&\textit{soggetto a} && \sum_{j=1}^m \pi_{ij} = \mu_i \quad \forall i = 1, \dots, n \\
& && \sum_{i=1}^n \pi_{ij} = \nu_j \quad \forall j = 1, \dots, m \\
& && \pi_{ij} \geq 0 \qquad \forall i, j
\end{aligned}$$

Ove $\mathrm{d}_{ij} := \mathrm{d}(x_i, y_j)$ è una matrice $n \times m$ che ha come entrate le distanze tra i punti del supporto di μ e ν e la matrice $n \times m$ $(\pi_{ij})_{ij}$ corrisponde invece alla misura di probabilità bivariata sullo spazio prodotto $X \times X$, definita da:

$$\pi = \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} \delta_{(x_i, y_j)}.$$

Dimostrazione. È sufficiente osservare che ogni coupling di μ e ν ha necessariamente supporto contenuto in:

$$S := \text{supp}\mu \times \text{supp}\nu = \{(x_i, y_j) \in X \times X \mid i = 1, \dots, n, j = 1, \dots, m\}.$$

Infatti se $\pi \in \Pi(\mu, \nu)$ e $A \in \mathscr{B}(X)$ è tale che $A \cap S = \emptyset$, allora:

$$\pi(A) \leq \pi(p^1(A) \times X) + \pi(X \times p^2(A)) = \mu(p^1(A)) + \nu(p^2(A)) = 0,$$

ove abbiamo indicato con p^1, p^2 le proiezioni sulla prima e sulla seconda coordinata rispettivamente. Quindi anche tutti i coupling di μ e ν hanno supporto finito e sono quindi rappresentati da una matrice $n \times m$, che indichiamo con $(\pi_{ij})_{ij}$. Da questo deduciamo che la quantità da minimazzare per la definizione di $\mathcal{W}_r$ è:

$$\iint_{X \times X} \mathrm{d}(x, y)^r \, d\pi(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \mathrm{d}_{ij}^r,$$

ove $\mathrm{d}_{ij} = \mathrm{d}(x_i, y_j)$ per ogni $i = 1, \dots, n, j = 1, \dots, m$. Infine per dimostrare l'equivalenza dei vincoli è sufficiente osservare che siccome i coupling di μ e ν sono discreti basta richiedere la condizione sulle marginali solo sui singoletti, quella generale seguirà per additività, ossia è sufficiente chiedere:

$$\pi(\{x_i\} \times X) = \mu(\{x_i\}) = \mu_i \quad \forall i = 1, \dots, n$$

ma ora per ogni $i = 1, \dots, n$:

$$\pi(\{x_i\} \times X) = \sum_{j=1}^{m} \pi_{ij}$$

otteniamo quindi esattamente i primi m vincoli del programma lineare. Un ragionamento analogo per ν porta all'ottenimento dei seguenti n vincoli. Per fare in modo che le soluzioni ammissibili dei vicoli ottenuti siano effettivamente la densità discreta di una misura di probabilità è sufficiente aggiungiere la condizione che $\pi_{ij} \geq 0$ per ogni i, j. Il fatto che i π_{ij} sommino a 1 è conseguenza dei vincoli di equazioni, infatti:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} = \sum_{i=1}^{n} \mu_i = 1. \qquad \square$$

Osservazione 1.8. Si noti che la matrice dei vincoli nel Programma Lineare presente nell'enunciato del teorema ha sempre la stessa forma, in particolare i vincoli di

equazioni possono essere scritti in notazione vettoriale come:

$$\begin{bmatrix} 1 & & 0 & 1 & & 0 & & 1 & & 0 \\ & \ddots & & & \ddots & & \dots & & \ddots & \\ 0 & & 1 & 0 & & 1 & & 0 & & 1 \\ 1 & \dots & 1 & 0 & \dots & 0 & & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & & 0 & \dots & 0 \\ & & & & & & \ddots & & & \\ 0 & \dots & 0 & 0 & \dots & 0 & & 1 & \dots & 1 \end{bmatrix} \begin{pmatrix} \pi_{11} \\ \vdots \\ \pi_{n1} \\ \pi_{12} \\ \vdots \\ \pi_{n2} \\ \vdots \\ \vdots \\ \pi_{nm} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \\ \nu_1 \\ \vdots \\ \nu_m \end{pmatrix}.$$

Più compattamente introducendo il vettore: $\boldsymbol{I}_n := \underbrace{(1, 1, \dots, 1)}_{\text{n volte}}$, la matrice indicata sopra si può scrivere:

$$\begin{bmatrix} \boldsymbol{I}_m \otimes \mathbf{1}_n \\ \mathbf{1}_m \otimes \boldsymbol{I}_n \end{bmatrix},$$

ove $\mathbf{1}_n$ indica la matrice identica di ordine n e $\otimes$ indica in prodotto di Kronecker tra matrici, cioè ad esempio $\boldsymbol{I}_m \otimes \mathbf{1}_n$ è una matrice formata da una riga lunga m di matrici identiche di ordine n. Questo permette una facile implementazione numerica, quando si considerano misure discrete su $\mathbb{R}^T$ e si conoscono le loro densità discrete. Nell'appendice B è fornita una implementazione MATLAB, che risolve il programma lineare associato a $\mathcal{W}_r$.

Capitolo 2

La Distanza di Wasserstein Adattata

In questo capitolo cerchiamo di generalizzare la distanza di Wasserstein al caso non di singole variabili aleatorie (o meglio, delle loro leggi) ma al caso *di processi stocastici* a tempo discreto. In particolare vogliamo costruire una distanza che tenga conto dell'*informazione* contenuta nel processo ad ogni istante. Vedremo che la ovvia generalizzazione della distanza di Wasserstein non è in grado di "captare" questa informazione, e questo è dovuto al fatto che essa "vede" solamente le probabilità "finali" delle *traiettorie* del processo, ma non considera le *probabilità condizionate*.

2.1 Introduzione

Consideriamo un processo stocastico $\xi = (\xi_t)_{t=1,\dots,T}$, con:

$$\xi_t \colon (X, \mathcal{M}, \mu) \to (E, \mathrm{d}), \quad t = 1, \dots, T.$$

Ove $T \in \mathbb{N}, T \geq 1$, (E, d) è uno *spazio polacco* che pensiamo dotato della sua σ-algebra dei Boreliani: $\mathscr{B}(E)$. Le v.a. $(\xi_t)_{t=1,\dots,T}$ possono essere pensate come una singola v.a.:

$$\begin{aligned} \xi \colon (X, \mathcal{M}, \mu) &\to E^T \\ x &\mapsto (\xi_1(x), \xi_2(x), \dots, \xi_T(x)), \end{aligned}$$

che mappa ogni x nella traiettoria corrispondente nello spazio prodotto $(E^T, \mathscr{B}(E^T))$, ove $\mathscr{B}(E^T) = \bigotimes_{j=1}^{T} \mathscr{B}(E)$, siccome E è polacco e quindi *separabile*. Allora è possibile considerare la *legge del processo*: $\mathcal{L}(\xi) = \mu \circ \xi^{-1}$, che è una misura di probabilità su $(E^T, \mathscr{B}(E^T))$. Ora poichè E è dotato di una distanza d, ci sono vari modi di costruire una metrica sullo spazio prodotto, ad esempio:

- *Distanza ℓ^1*: definita ponendo $\forall x, y \in E^T$:

$$\mathrm{d}_{E^T}(x, y) = \sum_{t=1}^{T} \mathrm{d}(x_t, y_t);$$

- *Distanza* ℓ^2: data in modo analogo tramite, $\forall x, y \in E^T$:

$$\mathrm{d}_{E^T}(x, y) = \left(\sum_{t=1}^{T} \mathrm{d}(x_t, y_t)^2 \right)^{1/2};$$

- *Distanza* ℓ^∞: definita ponendo $\forall x, y \in E^T$:

$$\mathrm{d}_{E^T}(x, y) = \max_{t=1,\dots,T} \mathrm{d}(x_t, y_t).$$

Per ognuna delle precedenti scelte possiamo allora costruire la distanza di Wasserstein tra misure di probabilità su $(E^T, \mathscr{B}(E^T))$, in particolare dati due processi $\xi = (\xi_t)_{t=1,\dots,T}$ e $\eta = (\eta_t)_{t=1,\dots,T}$ entrambi a valori in (E, d) possiamo calcolare la distanza tra i due processi come:

$$\mathcal{W}_r(\mathcal{L}(\xi), \mathcal{L}(\eta)).$$

L'esempio seguente chiarisce perchè questa definizione di distanza tra processi non è soddisfaciente dal punto di vista dell'informazione contenuta negli stessi.

Esempio 2.1. Consideriamo i due processi stocastici $(\xi^1_t)_{t=0,1,2}$ e $(\xi^2_t)_{t=0,1,2}$ rappresentati dai primi due alberi nella figura 2.1 nella pagina seguente. La distanza di Wasserstein fra le leggi dei due processi, può essere agevolmente calcolata risolvendo il programma lineare associato, come visto nella sezione. Usando la distanza ℓ^1 su $\mathbb{R}^3$, con pesi tutti uguali a 1, risulta:

$$\mathcal{W}_r(\mathcal{L}(\xi^1), \mathcal{L}(\xi^2)) = \varepsilon \xrightarrow[\varepsilon \to 0^+]{} 0,$$

quindi per la definizione classica di distanza di Wasserstein, ξ^1 e ξ^2 sono *arbitrariamente "vicini"*, ossia *simili*, questo è dovuto al fatto che le traiettorie dei due processi sono molto simili e hanno la stessa probabilità totale. Tuttavia i due processi sono estremamente diversi dal punto di vista dell'informazione che racchiudono, infatti se osserviamo che $\xi^2_1 = 2 - \varepsilon$ o che $\xi^2_1 = 2 + \varepsilon$ possiamo già concludere se ξ^2_2 sarà uguale a 2 o a 1, questa informazione non è disponibile con ξ^1, quando osserviamo $\xi^1_1 = 2$ non sappiamo con certezza che valore assumerà ξ^1_2. Matematicamente questo corrisponde al fatto che se consideriamo le filtrazioni degli spazi di probabilità sottostanti ai due processi queste differiscono in $t = 1$, vale infatti: $\mathcal{M}_1 \subsetneq \mathcal{N}_1$, come si vede dalla fig. 2.1 nella pagina successiva. Il motivo per cui la distanza definita sopra non *"vede"* questa differenza è perchè essa considera solo le probabilità totali delle traiettorie, ma ignora le probabilità condizionali, che in $t = 1$ sono molto diverse per ξ^1 e ξ^2.

Questo ci dice che se vogliamo costruire una *metrica in informazione* tra processi stocastici dobbiamo introdurre delle condizioni sulle probabilità condizionali, rispetto alle σ-algebre che vanno a comporre le filtrazioni da essi generati. Nella prossima sezione spieghiamo come fare.

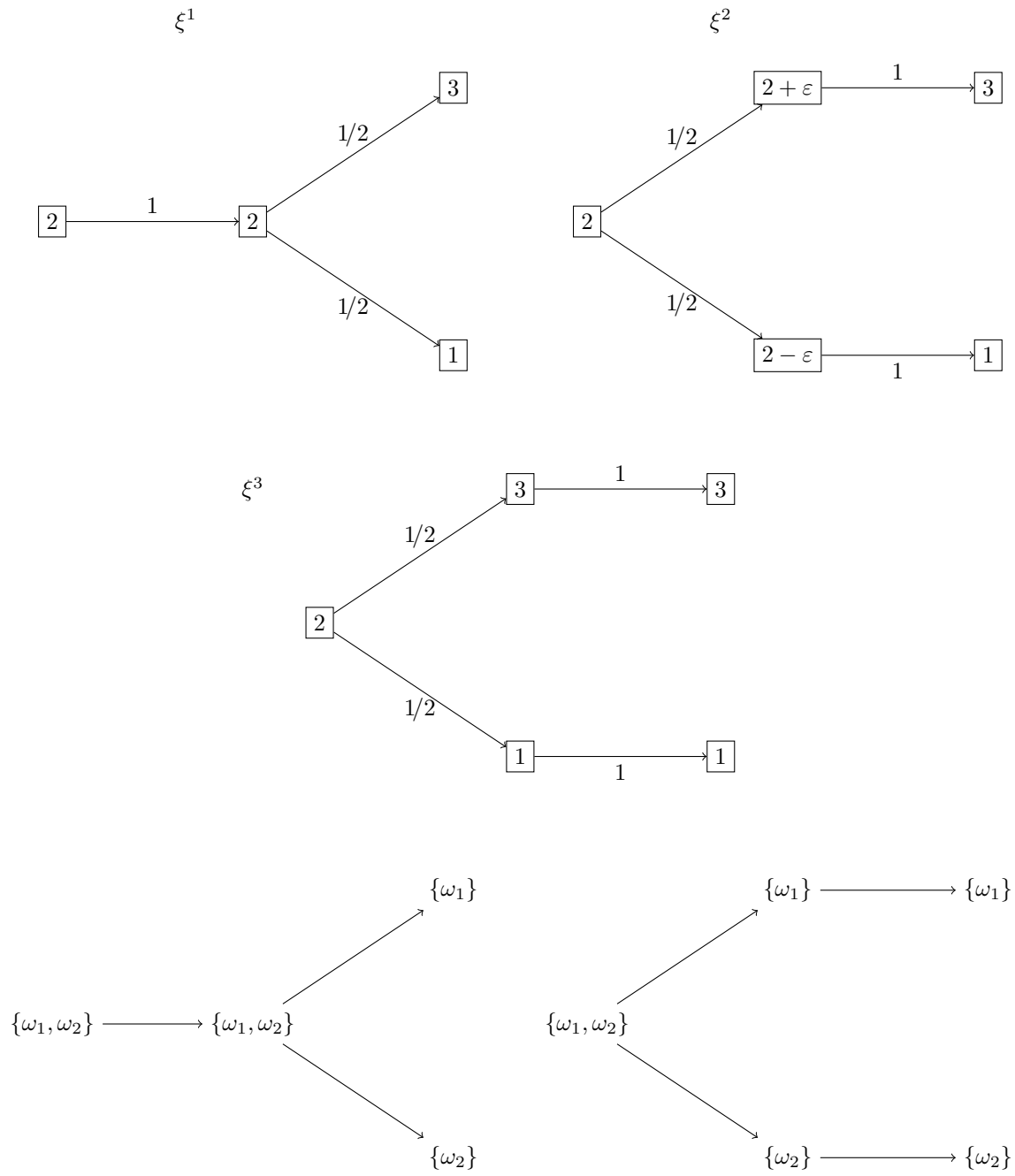

Figura 2.1: Alberi dei processi ξ^1, ξ^2 e ξ^3 e spazi di probabilità sottostanti a ξ^1 e ξ^2. Come si vede al tempo $t = 1$ la σ-algebra relativa a $\xi^1, \mathcal{M}_1 = \sigma(\{\omega_1, \omega_2\})$, è ancora la sigma algebra banale, mentre nel caso di ξ^2, la σ-algebra al tempo $t = 1$ è $\mathcal{N}_1 = \sigma(\{\omega_1\}, \{\omega_2\})$, cioè è già uguale alla sigma algebra al tempo $t = 2$. Infatti in ξ_2 tutta l'informazione è già disponibile al tempo $t = 1$.

2.2 La Distanza di Wasserstein adattata

Nel seguito $\mathbf{T}$ indicherà l'insieme dei tempi, ossia: $\mathbf{T} := \{0,1,\dots,T\}$, ed assumeremo che negli spazi filtrati la σ-algebra al tempo $t = 0$ coincida con quella banale, e che quella al tempo T coincida con quella data. Per tener conto delle filtrazioni presenti negli spazi di probabilità che si vanno a considerare, quando si vuole calcolare la distanza tra due misure μ e ν, raffiniamo l'insieme dei coupling su cui andremo a prendere l'estremo inferiore. In particolare introduciamo il concetto di *causalità* e *bicausalità*.

Definizione 2.1. Siano $(X, \mathcal{M}, (\mathcal{M}_t)_{t\in\mathbf{T}}, \mu)$ e $(Y, \mathcal{N}, (\mathcal{N}_t)_{t\in\mathbf{T}}, \nu)$ due spazi di probabilità filtrati, diciamo che un coupling $\pi \in \Pi(\mu,\nu)$ è *causale da μ a ν* se vale la seguente condizione:

$$\pi(A \times Y \mid \mathcal{M}_t \otimes \mathcal{N}_t) = \mu(A \mid \mathcal{M}_t) \quad \forall A \in \mathcal{M}, t \in \mathbf{T}. \tag{2.1}$$

Se poi vale anche la sua simmetrica:

$$\pi(X \times B \mid \mathcal{M}_t \otimes \mathcal{N}_t) = \nu(B \mid \mathcal{N}_t) \quad \forall B \in \mathcal{N}, t \in \mathbf{T}, \tag{2.2}$$

allora diremmo che il coupling π è *bicausale*. Indichiamo gli insiemi dei coupling causali e bicausali tra μ e ν rispettivamente con $\Pi_{\mathrm{C}}(\mu,\nu)$ e $\Pi_{\mathrm{BC}}(\mu,\nu)$.

Osservazione 2.1. Ricordiamo che le probabilità condizionali, rispetto ad una σ-algebra, sono definite tramite il valore atteso condizionato:

$$\forall A \in \mathcal{M}_T \quad \mu(A \mid \mathcal{M}_t) := \mathbb{E}^{\mu}[\mathbf{1}_A \mid \mathcal{M}_t],$$

ed sono pertanto esse stesse v.a. : $\mu(A \mid \mathcal{M}_t)\colon X \to [0,1]$. La loro proprietà caratterizzante è:

$$\mathbb{E}^{\mu}\big[\mu(A \mid \mathcal{M}_t)\mathbf{1}_B\big] = \mathbb{E}^{\mu}\big[\mathbf{1}_A\mathbf{1}_B\big] = \mu(A \cap B) \quad \forall B \in \mathcal{M}_t.$$

Dunque la condizione in (2.1) significa che per ogni $A \in \mathcal{M}_T$ deve valere:

$$\pi(A \times Y \mid \mathcal{M}_t \otimes \mathcal{N}_t)(x,y) = \mu(A \mid \mathcal{M}_t)(x) \quad \text{per } \pi\text{-q.o. } (x,y).$$

Analogamente per la seconda condizione. Si noti in particolare che il membro di destra dell'equazione precedente è indipendente da y, talvolta è utile esplicitare questa indipendenza usando le proiezioni $p^1\colon X\times Y \to X, (x,y) \mapsto x$ e $p^2\colon X\times Y \to Y, (x,y) \mapsto y$, cioè scrivendo ad esempio:

$$\pi(A \times Y \mid \mathcal{M}_t \otimes \mathcal{N}_t) = \mu(A \mid \mathcal{M}_t) \circ p^1.$$

Possiamo osservare anche che per $t = 0$ sia $\pi(A \times Y \mid \mathcal{M}_0 \otimes \mathcal{N}_0) = \pi(A \times Y \mid \{\emptyset, X \times Y\})$ sia $\mu(A \mid \mathcal{M}_0)$ sono quantità deterministiche, uguali in particolare a $\pi(A\times Y)$ e $\mu(A)$ rispettivamente. Dunque per $t = 0$ le condizioni (2.1) e (2.2) sono equivalenti a:

$$\pi(A \times Y) = \mu(A) \quad \text{e} \quad \pi(X \times B) = \nu(B) \quad \forall A \in \mathcal{M}, B \in \mathcal{N}.$$

Questo ci dice che: $\Pi_{\mathrm{BC}}(\mu,\nu) \subseteq \Pi(\mu,\nu)$. Osserviamo poi che per $t = T$ abbiamo invece che $\mu(A \mid \mathcal{M}_T) = \mathbb{E}^{\mu}[\mathbf{1}_A \mid \mathcal{M}_T] = \mathbf{1}_A$ e $\pi(A \times Y \mid \mathcal{M}_T \otimes \mathcal{N}_T) = \mathbf{1}_{A\times Y}$. Ma siccome $\mathbf{1}_A \circ p^1 = \mathbf{1}_{A\times Y}$ è sempre vero capiamo che le condizioni in (2.1) e (2.2) sono ridondanti per $t = T$ e possono essere tralasciate.

Lemma 2.1. *Siano $(X,\mathcal{M},(\mathcal{M}_t)_{t\in\mathbf{T}},\mu)$ e $(Y,\mathcal{N},(\mathcal{N}_t)_{t\in\mathbf{T}},\nu)$ due spazi di probabilità filtrati, allora $\mu\otimes\nu \in \Pi_{BC}(\mu,\nu) \subseteq \Pi_C(\mu,\nu)$, in particolare dunque $\Pi_{BC}(\mu,\nu) \neq \emptyset$.*

Dimostrazione. Basta osservare che per ogni $t \in \mathbf{T}$, fissati $A \in \mathcal{M}_T$ e $B \in \mathcal{N}_T$, per ogni $C \in \mathcal{M}_t$ e per ogni $D \in \mathcal{N}_t$ si ha:

$$\begin{aligned}
&\iint_{C\times D} (\mu(A \mid \mathcal{M}_t)\circ p^1)\cdot(\nu(B \mid \mathcal{N}_t)\circ p^2)\, d\,\mu\otimes\nu \\
&= \left(\int_C \mu(A \mid \mathcal{M}_t)\, d\mu\right)\cdot\left(\int_D \nu(B \mid \mathcal{N}_t)\, d\nu\right) \\
&= \mu(A\cap C)\cdot\nu(B\cap D) \\
&= \mu\otimes\nu\big((A\cap C)\times(B\cap D)\big) \\
&= \mu\otimes\nu\big((A\times B)\cap(C\times D)\big) \\
&= \iint_{C\times D} \mu\otimes\nu(A\times B \mid \mathcal{M}_t\otimes\mathcal{N}_t)\, d\,\mu\otimes\nu.
\end{aligned}$$

ora siccome entrambe le funzioni: $(\mu(A \mid \mathcal{M}_t)\circ p^1)\cdot(\nu(B \mid \mathcal{N}_t)\circ p^2)$ che $\mu\otimes\nu(A\times B \mid \mathcal{M}_t\otimes\mathcal{N}_t)$ sono misurabili rispetto a $\mathcal{M}_t\otimes\mathcal{N}_t$ e non negative, siccorme i loro integrali sono uguali su ogni generatore di $\mathcal{M}_t\otimes\mathcal{N}_t$ otteniamo che sono versioni di se stesse, ossia:

$$\big(\mu(A \mid \mathcal{M}_t)\circ p^1\big)\cdot\big(\nu(B \mid \mathcal{N}_t)\circ p^2\big) = \mu\otimes\nu(A\times B \mid \mathcal{M}_t\otimes\mathcal{N}_t) \quad \mu\otimes\nu\text{-q.c.}$$

con le particolari scelte di $A = X$ oppure $B = Y$ otteniamo che $\mu\otimes\nu$ verifica le condizioni (2.1) e (2.2) e quindi sta in $\Pi_{\mathrm{BC}}(\mu,\nu)$. □

Proposizione 2.2. *Siano $(X,\mathcal{M},(\mathcal{M}_t)_{t\in\mathbf{T}},\mu)$ e $(Y,\mathcal{N},(\mathcal{N}_t)_{t\in\mathbf{T}},\nu)$ due spazi di probabilità filtrati, allora le seguenti affermazioni sono equivalenti.*

1. *$\pi \in \Pi_C(\mu,\nu)$.*
2. *Per ogni funzione $\lambda \in L^1(\mu)$ e per ogni $t \in \mathbf{T}$ si ha*
$$\mathbb{E}^{\pi}\big[\lambda\circ p^1 \mid \mathcal{M}_t\otimes\mathcal{N}_t\big] = \mathbb{E}^{\mu}\big[\lambda \mid \mathcal{M}_t\big]\circ p^1, \tag{2.3}$$
3. *Per ogni funzione $\lambda \in L^\infty(\mu)$ e per ogni $t \in \mathbf{T}$ si ha*
$$\mathbb{E}^{\pi}\big[\lambda\circ p^1 \mid \mathcal{M}_t\otimes\mathcal{N}_t\big] = \mathbb{E}^{\mu}\big[\lambda \mid \mathcal{M}_t\big]\circ p^1, \tag{2.4}$$

Osservazione 2.2. Ovviamente per simmetria avremmo che $\pi \in \Pi_{\mathrm{BC}}(\mu,\nu)$ se e solo se valgono anche proprietà analoghe a 2 e 3 per la misura ν, in particolare se vale anche che, per ogni $\zeta \in L^1(\nu), t \in \mathbf{T}$ (o $L^\infty(\nu)$) si ha

$$\mathbb{E}^{\pi}\big[\zeta\circ p^2 \mid \mathcal{M}_t\otimes\mathcal{N}_t\big] = \mathbb{E}^{\nu}\big[\zeta \mid \mathcal{N}_t\big]\circ p^2.$$

Dimostrazione. Ovviamente la 2 implica la 1 perchè se vale la seconda affermazione allora essa vale in particolare per le funzioni caratteristiche $\mathbf{1}_A \in L^1(\mu)$ per ogni $A \in \mathcal{M}$. Ma allora

$$\begin{aligned}\pi(A \times Y \mid \mathcal{M}_t \otimes \mathcal{N}_t) &= \mathbb{E}^\pi[\mathbf{1}_{A\times Y} \mid \mathcal{M}_t \otimes \mathcal{N}_t] = \mathbb{E}^\pi[\mathbf{1}_A \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t] \\ &= \mathbb{E}^\mu[\mathbf{1}_A \mid \mathcal{M}_t] \circ p^1 = \mu(A \mid \mathcal{M}_t) \circ p^1,\end{aligned}$$

Dunque $\pi \in \Pi_{\mathrm{C}}(\mu, \nu)$.

Viceversa supponiamo $\pi \in \Pi_{\mathrm{C}}(\mu, \nu)$, allora i calcoli sopra mostrano che la proprietà (2.3) vale per le funzioni caratteristiche. Ma allora per linearità del valore atteso condizionato vale anche per le funzioni semplici non negative. Scelta infatti

$$f = \sum_{i=1}^n a_i \mathbf{1}_{A_i} \quad \text{ove } a_i \geq 0 \text{ e } A_i \in \mathcal{M},$$

si ha, per ogni $t \in \mathbf{T}$

$$\begin{aligned}\mathbb{E}^\pi[f \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t] &= \sum_{i=1}^n a_i \mathbb{E}^\pi[\mathbf{1}_{A_i} \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t] \\ &= \sum_{i=1}^n a_i \big(\mathbb{E}^\mu[\mathbf{1}_{A_i} \mid \mathcal{M}_t] \circ p^1\big) \\ &= \mathbb{E}^\mu\Big[\sum_{i=1}^n a_i \mathbf{1}_{A_i} \mid \mathcal{M}_t\Big] \circ p^1 = \mathbb{E}^\mu[f \mid \mathcal{M}_t] \circ p^1,\end{aligned}$$

Presa allora φ funzione misurabile non negativa, esiste una successione di funzioni semplici non negative $(f_n)_{n\in\mathbb{N}}$ tale che

$$f_n(x) \xrightarrow[n\to\infty]{} \varphi(x) \quad \text{crescendo, per ogni } x \in X$$

Che è equivalente a dire che puntualmente su $X \times Y$, $f_n \circ p^1 \xrightarrow[n\to\infty]{} \varphi \circ p^1$ crescendo. Ma allora per convergneza monotona per ogni $t \in \mathbf{T}$

$$\begin{aligned}\mathbb{E}^\pi[\varphi \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t] &= \lim_{n\to\infty} \mathbb{E}^\pi[f_n \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t] \\ &= \lim_{n\to\infty} \mathbb{E}^\mu[f_n \mid \mathcal{M}_t] \circ p^1 = \mathbb{E}^\mu[\varphi \mid \mathcal{M}_t] \circ p^1,\end{aligned}$$

Infine presa $\lambda \in L^1(\mu)$ è sufficiente passare alla parte positiva e negativa, infatti per ogni $t \in \mathbf{T}$ si ha

$$\begin{aligned}\mathbb{E}^\pi[\lambda \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t] &= \mathbb{E}^\pi[\lambda^+ \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t] + \mathbb{E}^\pi[\lambda^- \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t] \\ &= \mathbb{E}^\mu[\lambda^+ \mid \mathcal{M}_t] \circ p^1 + \mathbb{E}^\mu[\lambda^- \mid \mathcal{M}_t] \circ p^1 = \mathbb{E}^\mu[\lambda \mid \mathcal{M}_t] \circ p^1,\end{aligned}$$

Quindi π soddisfa la condizione (2.3). Questo termina la dimostrazione che 1 è equivalente a 2. Per mostrare anche l'equivalenza con 3 è sufficiente osservare come prima cosa che 2 implica 3 in virtù del fatto che $L^\infty(\mu) \subseteq L^1(\mu)$ e poi che 3 implica 1, e quindi 2, siccome le funzioni caratteristiche sono elemnti di $L^\infty(\mu)$. □

Lemma 2.3 (Tower Property)**.** *Nella definizione 2.1, la condizione:*

$$\pi\big(A \times Y \mid \mathcal{M}_t \otimes \mathcal{N}_t\big) = \mu(A \mid \mathcal{M}_t) \circ p^1 \quad \forall A \in \mathcal{M}_T, \forall t \in \mathbf{T} \tag{2.5}$$

è equivalente a:

$$\pi\big(A \times Y \mid \mathcal{M}_t \otimes \mathcal{N}_t\big) = \mu(A \mid \mathcal{M}_t) \circ p^1 \quad \forall A \in \mathcal{M}_{t+1}, \forall t \in \{0,1,\dots,T-1\} \tag{2.6}$$

Un risultato analogo vale ovviamente per le eventuali condizioni su π rispetto a ν se stiamo considerando i coupling bicausali.

Dimostrazione. Ovviamente, poichè $\mathcal{M}_{t+1} \subseteq \mathcal{M}_T$, si ha che (2.5) implica (2.6). Supponiamo poi che valga (2.6) e dimostriamo che vale (2.5), per farlo osserviamo che per $t = T-1$ (2.5) è diretta applicazione di (2.6), osserviamo anche preliminarmente che per ogni $t \in \{0,1,...,T-1\}$ e per ogni $A \in \mathcal{M}_{t+1}$ si ha che:

$$\begin{aligned}\mathbb{E}^\pi\big[\mathbf{1}_A \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t\big] &= \mathbb{E}^\pi\big[\mathbf{1}_{A\times Y} \mid \mathcal{M}_t \otimes \mathcal{N}_t\big] = \pi\big(A \times Y \mid \mathcal{M}_t \otimes \mathcal{N}_t\big) \\ &= \mu(A \mid \mathcal{M}_t) \circ p^1 = \mathbb{E}^\mu\big[\mathbf{1}_A \mid \mathcal{M}_t\big] \circ p^1\end{aligned}$$

per linearità deduciamo che ciò vale per tutte le $\varphi \geq 0$ semplici φ misurabili rispetto a $\mathcal{M}_{t+1}$ e per convergenza monotona otteniamo che ciò vale per tutte le funzioni $\lambda \geq 0$ tali che λ sia misurabile rispetto a $\mathcal{M}_{t+1}$, ossia:

$$\mathbb{E}^\pi\big[\lambda \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t\big] = \mathbb{E}^\mu\big[\lambda \mid \mathcal{M}_t\big] \circ p^1.$$

Ora fissiamo $A \in \mathcal{M}_T$, allora grazie alla *tower property* del valore atteso condizionato abbiamo:

$$\begin{aligned}\pi\big(A \times Y \mid \mathcal{M}_t \otimes \mathcal{N}_t\big) &= \mathbb{E}^\pi\big[\mathbf{1}_{A\times Y} \mid \mathcal{M}_t \otimes \mathcal{N}_t\big] \\ &= \mathbb{E}^\pi\big[\mathbf{1}_A \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t\big] \\ &= \mathbb{E}^\pi\Big[\mathbb{E}^\pi\big[\mathbf{1}_A \circ p^1 \mid \mathcal{M}_{T-1} \otimes \mathcal{N}_{T-1}\big] \mid \mathcal{M}_t \otimes \mathcal{N}_t\Big] \\ &= \mathbb{E}^\pi\Big[\mathbb{E}^\mu\big[\mathbf{1}_A \mid \mathcal{M}_{T-1}\big] \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t\Big].\end{aligned}$$

Poichè ora $\mathbb{E}^\mu\big[\mathbf{1}_A \mid \mathcal{M}_{T-1}\big] \geq 0$ e $\mathbb{E}^\mu\big[\mathbf{1}_A \mid \mathcal{M}_{T-1}\big]$ è misurabile rispetto a $\mathcal{M}_{T-1}$ i passaggi sopra possono essere ripetuti ottenendo:

$$\begin{aligned}\pi\big(A \times Y \mid \mathcal{M}_t \otimes \mathcal{N}_t\big) &= \mathbb{E}^\pi\Big[\mathbb{E}^\mu\big[\mathbf{1}_A \mid \mathcal{M}_{T-1}\big] \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t\Big] \\ &= \mathbb{E}^\pi\Big[\mathbb{E}^\pi\big[\mathbb{E}^\mu[\mathbf{1}_A \mid \mathcal{M}_{T-1}] \circ p^1 \mid \mathcal{M}_{T-2} \otimes \mathcal{N}_{T-2}\big] \mid \mathcal{M}_t \otimes \mathcal{N}_t\Big] \\ &= \mathbb{E}^\pi\Big[\mathbb{E}^\mu\big[\mathbb{E}^\mu[\mathbf{1}_A \mid \mathcal{M}_{T-1}] \mid \mathcal{M}_{T-2}\big] \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t\Big] \\ &= \mathbb{E}^\pi\Big[\mathbb{E}^\mu\big[\mathbf{1}_A \mid \mathcal{M}_{T-2}\big] \circ p^1 \mid \mathcal{M}_t \otimes \mathcal{N}_t\Big]\end{aligned}$$

iterando il procedimento otteniamo:

$$\begin{aligned}\pi\big(A\times Y \mid \mathcal{M}_t\otimes\mathcal{N}_t\big) &= \mathbb{E}^{\pi}\Big[\mathbb{E}^{\mu}\big[\mathbf{1}_A \mid \mathcal{M}_{T-2}\big]\circ p^1 \mid \mathcal{M}_t\otimes\mathcal{N}_t\Big]\\ &= \mathbb{E}^{\pi}\Big[\mathbb{E}^{\mu}\big[\mathbf{1}_A \mid \mathcal{M}_{T-3}\big]\circ p^1 \mid \mathcal{M}_t\otimes\mathcal{N}_t\Big]\\ &= \dots\\ &= \mathbb{E}^{\pi}\Big[\mathbb{E}^{\mu}\big[\mathbf{1}_A \mid \mathcal{M}_t\big]\circ p^1 \mid \mathcal{M}_t\otimes\mathcal{N}_t\Big].\end{aligned}$$

Ed ora $\mathbb{E}^{\mu}\big[\mathbf{1}_A \mid \mathcal{M}_t\big]\circ p^1 \sim \mathcal{M}_t\otimes\mathcal{N}_t$ si ha:

$$\begin{aligned}\pi\big(A\times Y \mid \mathcal{M}_t\otimes\mathcal{N}_t\big) &= \mathbb{E}^{\pi}\Big[\mathbb{E}^{\mu}\big[\mathbf{1}_A \mid \mathcal{M}_t\big]\circ p^1 \mid \mathcal{M}_t\otimes\mathcal{N}_t\Big]\\ &= \mathbb{E}^{\mu}\big[\mathbf{1}_A \mid \mathcal{M}_t\big]\circ p^1 = \mu\big(A \mid \mathcal{M}_t\big)\circ p^1\end{aligned}$$

che è esattamente la condizione generale (2.5), come si voleva. □

Definizione 2.2 (Distanza Adattata)**.** Sia (X,d) uno spazio Polacco e $(X,\mathcal{M},(\mathcal{M}_t)_{t\in\mathbf{T}},\mu)$ e $(X,\mathcal{N},(\mathcal{N}_t)_{t\in\mathbf{T}},\nu)$ due spazi di probabilità filtrati. Si dice allora *distanza di Wasserstein adattata* (o *distanza annidata*) di ordine $r\geq 1$ tra μ e ν la quantità:

$$\mathcal{AW}_r(\mu,\nu) := \left(\inf_{\pi\in\Pi_{\mathrm{BC}}(\mu,\nu)}\iint_{X\times Y}\mathrm{d}(x,y)^r\,d\pi(x,y)\right)^{1/r} \tag{2.7}$$

Osservazione 2.3. La funzione $\mathcal{AW}_r$ risulta effettivamente una distanza tra misure di probabilità. La dimostrazione è simile a quella effettuata per $\mathcal{W}_r$ in quanto essa fa uso di una versione del 1.7 adattata al caso dei coupling bicausali.

Teorema 2.4 (Upper e Lower Bounds per la distanza adattata)**.** *Sia (X,d) uno spazio Polacco e $(X,\mathcal{M},(\mathcal{M}_t)_{t\in\mathbf{T}},\mu)$ e $(X,\mathcal{N},(\mathcal{N}_t)_{t\in\mathbf{T}},\nu)$ due spazi di probabilità filtrati, allora per ogni $r\geq 1$:*

$$\mathcal{W}_r(\mu,\nu)^r \leq \mathcal{AW}_r(\mu,\nu)^r \leq \mathbb{E}^{\mu\otimes\nu}\big[\mathrm{d}^r\big]. \tag{2.8}$$

Osservazione 2.4. Da questo teorema deduciamo che ogni distanza adattata può essere interpretata come la somma di due termini, uno interpretabile come la differenza tra le misure di probabilità, $\mathcal{W}_r(\mu,\nu)$ e la rimanente quantità $\mathcal{AW}_r(\mu,\nu) - \mathcal{W}_r(\mu,\nu)$ interpretabile come la differenza tra le filtrazioni dei due spazi, ovvero la differenza tra i due flussi di informazione.

Dimostrazione. La seconda disuguaglianza segue subito dal lemma 2.1 a pagina 53, infatti esso afferma che $\mu\otimes\nu$ è ammissibile per il problema di minimo che definisce $\mathcal{AW}_r(\mu,\nu)$, la prima segue dall'osservazione 2.1 sulle probabilità condizinali quando $t=0$, abbiamo infatti visto che $\Pi_{\mathrm{BC}}(\mu,\nu)\subseteq\Pi(\mu,\nu)$, quindi:

$$\mathcal{W}_r(\mu,\nu)^r = \inf_{\pi\in\Pi(\mu,\nu)}\iint_{X\times Y}\mathrm{d}^r\,d\pi \leq \inf_{\pi\in\Pi_{\mathrm{BC}}(\mu,\nu)}\iint_{X\times Y}\mathrm{d}^r\,d\pi = \mathcal{AW}_r(\mu,\nu)^r. \quad \square$$

Abbiamo introdotto la distanza adattata tra misure su spazi di probabilità filtrati generali, ora siamo invece interessati a specializzare la definizione al caso di leggi di processi stocastici a tempo discreto a valori nello stesso spazio Polacco (E, d). Come già osservato esse sono leggi sullo spazio prodotto E^T. È importante dunque capire che filtrazioni scegliere su tale spazio. In questo contesto si utilizza la cosidetta filtrazione canonica.

Definizione 2.3. Dato uno spazio misurabile $(E, \mathcal{E})$ si definisce sullo spazio prodotto $(E^T, \bigotimes_{t=1}^T \mathcal{E})$ la *filtrazione canonica*, come quella generata dalle proiezioni, ossia posto $\mathcal{E}_0 = \{\emptyset, E^T\}$ definiamo:

$$\mathfrak{E} = (\mathcal{E}_t)_{t\in\mathbf{T}} \quad \text{ove} \quad \mathcal{E}_t := \sigma(Y_s \mid s \leq t) \quad \forall t \in \mathbf{T} \setminus \{0\},$$
$$\text{e} \quad Y_t \colon E^T \to E \quad \text{è definita da} \quad (x_1, \dots, x_T) \mapsto x_t \quad \forall t \in \mathbf{T} \setminus \{0\}.$$

Con questa definizione la specializzazione della definizione di $\mathcal{AW}_r$ segue subito.

Definizione 2.4. Consideriamo due processi stocastici: $(\xi_t)_{t\in\mathbf{T}}$, $(\eta_t)_{t\in\mathbf{T}}$ definiti sullo spazio di probabilità $(X, \mathcal{M}, \mu)$ e a valori nello spazio polacco (E, d). Si definisce *distanza adattata tra ξ e η*, di ordine $r \geq 1$, la distanza adattata tra le misure dei due spazi filtrati: $(E^T, \mathscr{B}(E^T), (\mathcal{E}_t)_{t\in\mathbf{T}}, \mathcal{L}(\xi))$ e $(E^T, \mathscr{B}(E^T), (\mathcal{E}_t)_{t\in\mathbf{T}}, \mathcal{L}(\eta))$, ossia:

$$\mathcal{AW}_r(\xi, \eta) := \mathcal{AW}_r(\mathcal{L}(\xi), \mathcal{L}(\eta)), \tag{2.9}$$

ove $(\mathcal{E}_t)_{t\in\mathbf{T}}$ è la filtrazione naturale dello spazio prodotto.

Esempio 2.2. Considerando nuovamente i processi rappresentati nella figura 2.1, si calcola con i metodi spiegati nella sezione 4 di questo capitolo che la distanza $\mathcal{AW}_1$ tra le leggi di ξ^1 e ξ^2 risulta $1 + \varepsilon$, che è strettamente maggiore di 1 anche per $\varepsilon \to 0$, quindi i due processi non sono arbitrariamente vicini come si voleva. Se invece consideriamo ξ^2 e ξ^3 la distanza adattata tra le due leggi risulta uguale a $1 - \varepsilon$, che tende a 0 per $\varepsilon \to 1$, cosa coerente perchè in tal caso i due processi hanno le stesse traiettorie, ma anche la stessa struttura di informazione. Questo mostra come la definizione data per $\mathcal{AW}$ è quella corretta allo scopo di differenziare i processi considerando anche i flussi di informazione.

In vista del prossimo capitolo torna utile avere nel caso in cui si considerano leggi di processi, quindi spazi con filtrazioni canoniche, avere delle caratterizzazioni equivalenti per la causalità e bicausalità. Ciò è dato dal prossimo risultato, abbiamo tuttavia bisogno di fissare prima alcune notazioni.

Notazioni: Indichiamo il pushforward di una misura γ tramite una mappa M con $M_*\gamma := \gamma \circ M^{-1}$. Per misure γ su uno spazio prodotto $E \times F$ indichiamo con γ^x, γ^y i nuclei regolari di γ rispetto alla prima e alla seconda coordinata, ottenuti per disintegrazione, ovvero definiti da

$$\gamma(A \times B) = \int_A \gamma^x(B)\, d\gamma^1(x) \quad \text{e} \quad \gamma(A \times B) = \int_B \gamma^y(A)\, d\gamma^2(y),$$

ove $\gamma^1 := p^1_*\gamma$ e $\gamma^2 := p^2_*\gamma$. Per risultati di esistenza e alcune proprietà si riferisca a A.2 a pagina 91. La notazione si estende in modo analogo al prodotto di due o più spazi. Per una misura di probabilità γ su E^T indichiamo con $\gamma^{x_1,\dots,x_t}$ la misura uno dimensionale sulla $(t+1)$-esima coordinata ottenuta per disintegrazione di γ rispetto alle prime t coordinate, ovvero stiamo decomponendo γ in nuclei successivi

$$d\gamma(x_1,\dots,x_T) = dp^1_*\gamma(x_1)d\gamma^{x_1}(x_2)d\gamma^{x_1,x_2}(x_3)\cdot\dots\cdot d\gamma^{x_1,\dots,x_{T-1}}(x_T).$$

Su $E^T \times E^T$ denotiamo con $(x_1,\dots,x_T)$ le prime T coordinate e con $(y_1,\dots,y_T)$ le seconde T coordinate. Per una misura di probabilità γ su $E^T \times E^T$ indichiamo con $\gamma^{x_1,\dots,x_t,y_1,\dots,y_t}$ la misura due dimensionale sulle coordinate (x_{t+1}, y_{t+1}) ottenuta per disintegrazione di γ rispetto a $(x_1,\dots,x_t,y_1,\dots,y_t)$.

Proposizione 2.5. *Sia (E,d) uno spazio Polacco, $T \in \mathbb{N}$ e $\mu,\nu \in \mathcal{P}(E^T)$, dove dotiamo E^T della filtrazione canonica. Allora le seguenti affermazioni sono equivalenti.*

1. *$\pi \in \Pi_C(\mu,\nu)$.*
2. *Per ogni $t = 1,\dots,T$ e per ogni $B \in \mathcal{E}_t$ si ha che la funzione*
$$x \mapsto \pi^x(B) \quad \text{è } \mathcal{E}_t\text{-misurabile.} \tag{2.10}$$
3. *Dati, uno spazio di probabilità $(\Omega,\mathcal{F},\mathbb{P})$ e $(\xi_t)_{t=1,\dots,T}, (\eta_t)_{t=1,\dots,T}$ due processi stocastici definiti su Ω e a valori in E tali che $\xi \sim \mu, \eta \sim \nu$. Se $\pi = (\xi,\eta)_*\mathbb{P}$ allora per ogni $B_t \in \mathscr{B}(E^t)$ vale*
$$\mathbb{P}\big((\eta_1,\dots,\eta_t) \in B_t \mid \xi_1,\dots,\xi_T\big) = \mathbb{P}\big((\eta_1,\dots,\eta_t) \in B_t \mid \xi_1,\dots,\xi_t\big). \tag{2.11}$$

Osservazione 2.5. Per simmetria si ha poi ovviamente che $\pi \in \Pi_{\mathrm{BC}}(\mu,\nu)$ se e solo se oltre a quelle dell'enunciato valgono anche condizioni analoghe a (2.10) per ν ed a (2.11) invertendo i ruoli di ξ ed η. In particolare la proprietà (2.11) corrisponde alla indipendenza condizionale di $(\eta_1,\dots,\eta_t)$ e $(\xi_{t+1},\dots,\xi_T)$ dato $(\xi_1,\dots,\xi_t)$ che può essere parafrasata dicendo che "dato il passato di ξ, il passato di η ed il futuro di ξ sono indipendenti".

Dimostrazione. Dimostriamo prima l'equivalenza tra 1 e 2. Per farlo è sufficiente ricordare che dalla proposizione 2.2 la causalità di π è equivalente a chiedere che, per ogni $f \in L^\infty(\mu)$ valga

$$\mathbb{E}^\pi\big[f\circ p^1 \mid \mathcal{E}_t \otimes \mathcal{E}_t\big] = \mathbb{E}^\mu\big[f \mid \mathcal{E}_t\big]\circ p^1, \tag{2.12}$$

quindi è sufficiente mostrare l'equivalenza tra la 2 e la proprietà precedente. A tale scopo osserviamo che per ogni $f \in L^\infty(\mu)$ e per ogni $A, B \in \mathcal{E}_t$ vale

$$\mathbb{E}^\pi\Big[\big(f\circ p^1 - \mathbb{E}^\mu[f \mid \mathcal{E}_t]\circ p^1\big)\mathbf{1}_{A\times B}\Big] = \mathbb{E}^\mu\Big[f\mathbf{1}_A\big(\pi(B) - \mathbb{E}^\mu[\pi(B) \mid \mathcal{E}_t]\big)\Big]. \tag{2.13}$$

Infatti dalla definizione di nucleo regolare abbiamo

$$\begin{aligned}\mathbb{E}^\pi\big[(f\circ p^1)\mathbf{1}_{A\times B}\big] &= \iint_{A\times B} f\circ p^1\, d\pi \\ &= \int_A f(x)\pi^x(B)\, d\mu(x) = \mathbb{E}^\mu\big[f\mathbf{1}_A\pi(B)\big].\end{aligned}$$

ma d'altra parte

$$\begin{aligned}\mathbb{E}^\pi\Big[\big(\mathbb{E}^\mu[f\mid\mathcal{E}_t]\circ p^1\big)\mathbf{1}_{A\times B}\Big] &= \mathbb{E}^\mu\Big[\mathbb{E}^\mu[f\mid\mathcal{E}_t]\mathbf{1}_A\pi(B)\Big] \\ &= \mathbb{E}^\mu\Big[\mathbb{E}^\mu\big[f\mathbf{1}_A\mid\mathcal{E}_t\big]\pi(B)\Big] \\ &= \mathbb{E}^\mu\Big[\mathbb{E}^\mu\big[f\mathbf{1}_A\mid\mathcal{E}_t\big]\mathbb{E}^\mu\big[\pi(B)\mid\mathcal{E}_t\big]\Big] \\ &= \mathbb{E}^\mu\Big[f\mathbf{1}_A\mathbb{E}^\mu[\pi(B)\mid\mathcal{E}_t]\Big].\end{aligned}$$

Vediamo come dalla (2.13) segue subito l'equivalenza tra 1 e 2. Se infatti vale 2 allora la funzione $x\to\pi^x(B)$ è $\mathcal{E}_t$-misurabile per ogni $B\in\mathcal{E}_t$. Ma allora dalla (**??**) otteniamo

$$\mathbb{E}^\pi\big[(f\circ p^1)\mathbf{1}_{A\times B}\big] = \mathbb{E}^\pi\Big[\big(\mathbb{E}^\mu[f\mid\mathcal{E}_t]\circ p^1\big)\mathbf{1}_{A\times B}\Big],$$

da cui per definizione

$$\mathbb{E}^\pi\big[f\circ p^1\mid\mathcal{E}_t\otimes\mathcal{E}_t\big] = \mathbb{E}^\mu\big[f\mid\mathcal{E}_t\big]\circ p^1$$

che come già osservato è equivalente a 1. Viceversa se vale 1 e quindi (2.12) abbiamo che fissato $B\in\mathcal{E}_t$ per ogni $f\in L^\infty(\mu)$ e per ogni $A\in\mathcal{E}_t$ vale

$$\mathbb{E}^\mu\big[f\mathbf{1}_A\pi(B)\big] = \mathbb{E}^\mu\Big[f\mathbf{1}_A\mathbb{E}^\mu\big[\pi(B)\mid\mathcal{E}_t\big]\Big].$$

Da cui scegliendo $A=E^T$ otteniamo

$$\mathbb{E}^\mu\Big[f\big(\pi(B)-\mathbb{E}^\mu[\pi(B)\mid\mathcal{E}_t]\big)\Big] = 0,$$

con la scelta poi di $f=\pi(B)-\mathbb{E}^\mu[\pi(B)\mid\mathcal{E}_t]\in L^\infty(\mu)$ ricaviamo

$$\mathbb{E}^\mu\Big[\big|\pi(B)-\mathbb{E}^\mu[\pi(B)\mid\mathcal{E}_t]\big|^2\Big] = 0$$

quindi

$$\pi(B) = \mathbb{E}^\mu[\pi(B)\mid\mathcal{E}_t]\quad\mu\text{-q.c.}$$

che è equivalente a dire che $x\mapsto\pi^x(B)$ è $\mathcal{E}_t$-misurabile come si voleva.

Proviamo ora l'equivalenza tra 2 e 3. Per farlo osserviamo preliminarmente che fissato $B_t\in\mathscr{B}(E^t)$ si ha $\mathbb{P}$-q.c.

$$\mathbb{P}\big((\eta_1,\dots,\eta_t)\in B_t\mid\xi_1,\dots,\xi_T\big)(\omega) = \pi^{(\xi_1(\omega),\dots,\xi_T(\omega))}(B_t\times E^{T-t}).$$

Infatti il nucleo $x \mapsto \pi^x(B_t \times E^{T-t})$ è $\mathcal{E}_T$-misurabile, e quindi la funzione $\omega \mapsto \pi^{(\xi_1(\omega),\dots,\xi_T(\omega))}(B_t \times E^{T-t})$ è $\sigma(\xi_1,\dots,\xi_T)$-misurabile, inoltre per ogni $A \in \mathscr{B}(E^T)$ si ha

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}\Big[\mathbf{1}_{B_t\times E^{T-t}}(\eta)\mathbf{1}_A(\xi)\Big] &= \pi\big(A \times (B_t \times E^{T-t})\big)\\ &= \int_{E^T} \mathbf{1}_A(x)\pi^x(B_t \times E^{T-t})\, d\mu(x)\\ &= \int_{\Omega} \mathbf{1}_A(\xi)\pi^\xi(B_t \times E^{T-t})\, d\mathbb{P}\\ &= \mathbb{E}^{\mathbb{P}}\Big[\pi^\xi(B_t \times E^{T-t})\mathbf{1}_A(\xi)\Big].\end{aligned}$$

Supponendo allora valga la 2 otteniamo che $\omega \mapsto \pi^{\xi(\omega)}(B_t \times E^{T-t})$ è $\sigma(\xi_1,\dots,\xi_t)$-misurabile inoltre per ogni $A_t \in \mathscr{B}(E^t)$ si ha

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}\Big[\mathbf{1}_{B_t\times E^{T-t}}(\eta)\mathbf{1}_{A_t\times E^{T-t}}(\xi)\Big] &= \pi\big((A_t \times E^{T-t}) \times (B_t \times E^{T-t})\big)\\ &= \int_{E^T} \mathbf{1}_{A_t\times E^{T-t}}(x)\pi^x(B_t \times E^{T-t})\, d\mu(x)\\ &= \int_{\Omega} \mathbf{1}_{A_t\times E^{T-t}}(\xi)\pi^\xi(B_t \times E^{T-t})\, d\mathbb{P}\\ &= \mathbb{E}^{\mathbb{P}}\Big[\mathbf{1}_{A_t\times E^{T-t}}(\xi)\pi^\xi(B_t \times E^{T-t})\Big].\end{aligned}$$

Quindi $\mathbb{P}$-q.c.

$$\begin{aligned}\mathbb{P}\big((\eta_1,\dots,\eta_t) \in B_t \mid \xi_1,\dots,\xi_T\big)(\omega) &= \pi^{(\xi_1(\omega),\dots,\xi_T(\omega))}(B_t \times E^{T-t})\\ &= \mathbb{P}\big((\eta_1,\dots,\eta_t) \in B_t \mid \xi_1,\dots,\xi_t\big)(\omega).\end{aligned}$$

Viceversa se vale la 3 allora dall'osservazione preliminare otteniamo che $\mathbb{P}$-q.c.

$$\omega \mapsto \pi^{\xi(\omega)}(B_t \times E^{T-t}) = \mathbb{P}\big((\eta_1,\dots,\eta_t) \in B_t \mid \xi_1,\dots,\xi_t\big)(\omega),$$

dunque la mappa $\omega \mapsto \pi^{\xi(\omega)}(B_t \times E^{T-t})$ è $\sigma(\xi_1,\dots,\xi_t)$-misurabile, che implica che la mappa $x \mapsto \pi^x(B_t \times E^{T-t})$ è $\mathcal{E}_t$-misurabile come si voleva. □

Osservazione 2.6. In vista del prossimo capitolo sarà importante anche il caso in cui esiste una mappa Borel misurabile $\psi\colon E^T \to E^T$ tale che $\nu = \psi_*\mu$, ossia in termini di ξ e η tale che $\eta = \psi(\xi)$. Allora la proprietà (2.10) è equivalente a chiedere che ψ sia adattata, ossia che ψ_t sia $\mathcal{E}_t$-misurabile per ogni $t = 1,\dots,T$. Infatti supponiamo $\pi \in \Pi_C(\mu,\nu)$, e scriviamo $d\pi(x,y) = d\delta_{\psi(x)}(y)d\mu(x)$ e fissiamo $t \le T$. Allora per ogni $B \in \mathcal{E}_t$, la mappa

$$x \mapsto \phi_B(x) := \delta_{\psi(x)}(B)$$

è $\mathcal{E}_t$-misurabile. Ora poichè $\psi^{-1}(B) = \phi_B^{-1}(\{1\})$, questo implica che ψ è $\mathcal{E}_t/\mathcal{E}_t$-misurabile, segue $x \mapsto (\psi_1(x),\dots,\psi_t(x))$ è $\mathcal{E}_t$-misurabile. Dal lemma di fattorizzazione otteniamo che esiste una funzione boreliana $\tilde{\psi}\colon E^t \to E^t$ tale che $(\psi_1(x),\dots,\psi_t(x)) =$

$\tilde{\psi}(x_1, \dots, x_t)$, ma allora $\psi_t = p^t \circ \tilde{\psi}$ da cui la conclusione. Viceversa supponiamo $d\pi(x, y) = d\delta_{\psi(x)}(y)d\mu(x)$ e $\nu = \psi_*\mu$ con ψ adattata. Fissiamo $t \leq T$ e $B \in \mathcal{E}_t$, dobbiamo mostrare che ϕ_B è $\mathcal{E}_t$-misurabile. Tuttavia siccome ϕ_B assume solo i valori 0 e 1, è sufficiente provare che $\phi_B^{-1}(\{1\}) \in \mathcal{E}_t$. A tale scopo osserviamo che $B = B_t \times E^{T-t}$ per qualche $B_t \in \mathscr{B}(E^t)$ e che

$$\phi_B^{-1}(\{1\}) = \psi^{-1}(B) = (\psi_1, \dots, \psi_t)^{-1}(B_t) \times E^{T-t} \in \mathcal{E}_t$$

perchè ψ è adattata, quindi $\pi \in \Pi_C(\mu, \nu)$ come si voleva.

Ora enunciamo un teorema che tornerà utile più volte in seguito, che afferma che la distanza adattata può essere anche calcolata ricorsivamente risolvendo iterativamente altri problemi di Kantorovich di dimensioni ridotte.

Teorema 2.6 (Computazione Ricorsiva)**.** *Sia* (E, d) *uno spazio Polacco,* $r \geq 1$*,* $T \in \mathbb{N}$ *e* $\mu, \nu \in \mathcal{P}_r(E^T, \mathrm{d}_{E^T})$*. Definiamo ricorsivamente le quantità* $(V_t^r)_{t \in \mathbf{T}}$ *tramite*

$$V_T^r := 0$$

e poi per $0 \leq t \leq T-1$

$$\begin{aligned} &V_t^r(x_1, \dots, x_t, y_1, \dots, y_t) \\ &:= \inf_{\pi^{t+1} \in \Pi(\mu^{x_1,\dots,x_t}, \nu^{y_1,\dots,y_t})} \iint_{E \times E} \left[V_{t+1}^r(x_1, \dots, x_{t+1}, y_1, \dots, y_{t+1}) + \mathrm{d}(x_{t+1}, y_{t+1})^r \right] d\pi^{t+1}(x_1, x_2). \end{aligned} \tag{2.14}$$

Allora si ha che

$$\mathcal{AW}_r(\mu, \nu)^r = V_0^r = \inf_{\pi \in \Pi(p_*^1\mu, p_*^1\nu)} \iint_{E \times E} \left[V_1^r(x_1, y_1) + \mathrm{d}(x_1, y_1)^r \right] d\pi^1(x_1, y_1). \tag{2.15}$$

Osservazione 2.7. Notiamo che siccome $V_T^r = 0$ per definizione i termini $V_t^r(x_1, \dots, x_t, y_1, \dots, y_t)$ che compaiono nella formulazione ricorsiva possono essere interpretati come la distanza di Wasserstein tra le due misure di disintegrazione $\mu^{x_1,\dots,x_t}, \nu^{y_1,\dots,y_t}$, ovvero corrisponde a $\mathcal{W}_r(\mu^{x_1,\dots,x_t}, \nu^{y_1,\dots,y_t})^r$.

Per la dimostrazione del risultato si fa riferimento alla sezione 4 di [2].

2.3 Proprietà topologiche della distanza adattata

In questa sezione studiamo alcune proprietà topologiche della distanza adattata $\mathcal{AW}_r$ in particolare il fatto che dato uno spazio Polacco (E, d) e un orizzonte temporale $T \in \mathbb{N}$, lo spazio delle leggi dei processi stocastici a valori in (E, d) non è uno spazio metrico completo se lo dotiamo della metrica $\mathcal{AW}_r$ e spieghiamo come identificarne il completamento. L'incompletezza si vede anche su controesempi molto semplici come il seguente.

Esempio 2.3. Per semplicità consideriamo $E = \mathbb{R}, T = 2$ e $r = 1$. Per ogni $n \in \mathbb{N}$ consideriamo le misure:

$$\mu_n := \frac{1}{2}\big(\delta_{(\frac{1}{n},1)} + \delta_{(-\frac{1}{n},-1)}\big).$$

Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità e $A \in \mathcal{F}$ con $\mathbb{P}(A) = 1/2$. Definiamo le variabili aleatorie $(\xi^n)_{n\in\mathbb{N}}$ a valori in $\mathbb{R}^2$ tramite

$$\xi^n := \begin{pmatrix} 1/n \\ 1 \end{pmatrix} \mathbf{1}_A + \begin{pmatrix} -1/n \\ -1 \end{pmatrix} \mathbf{1}_{A^c}.$$

Allora è evidente che $\xi^n \sim \mu_n$. Ora osserviamo che fissati $n, m \in \mathbb{N}$ la legge di (ξ^n, ξ^m) è un coupling bicausale tra μ_n e μ_m, infatti

$$\mathbb{P}\Big(\xi_1^n = \frac{1}{n} \mid \xi_1^m, \xi_2^m\Big) = \mathbb{P}\Big(\xi_1^n = \frac{1}{n} \mid \xi_1^m\Big),$$

e

$$\mathbb{P}\Big(\xi_1^n = -\frac{1}{n} \mid \xi_1^m, \xi_2^m\Big) = \mathbb{P}\Big(\xi_1^n = -\frac{1}{n} \mid \xi_1^m\Big).$$

Questo è vero siccome $\{\xi^m = (1/m, 1)\} = \{\xi_1^m = 1/m\} = A$, dunque

$$\begin{aligned}
\mathbb{E}^{\mathbb{P}}\Big[\mathbf{1}_{\{\xi_1^n=\frac{1}{n}\}}\mathbf{1}_{\{\xi^m=(1/m,1)\}}\Big] &= \mathbb{E}^{\mathbb{P}}\Big[\mathbf{1}_{\{\xi_1^n=\frac{1}{n}\}}\mathbf{1}_{\{\xi_1^m=1/m\}}\Big] \\
&= \mathbb{E}^{\mathbb{P}}\Big[\mathbb{P}\Big(\xi_1^n = \frac{1}{n} \mid \xi_1^m\Big)\mathbf{1}_{\{\xi_1^m=1/m\}}\Big] \\
&= \mathbb{E}^{\mathbb{P}}\Big[\mathbb{P}\Big(\xi_1^n = \frac{1}{n} \mid \xi_1^m\Big)\mathbf{1}_{\{\xi^m=(1/m,1)\}}\Big].
\end{aligned}$$

I calcoli precedenti possono essere ripetuti in modo simile ancora per l'evento $\{\xi_1^n = -1/n\}$, e condizionando rispetto a $\{\xi^m = (-1/m, -1)\}$ e successivamente invertendo i rouli di ξ^n e ξ^m. Otteniamo quindi che $(\xi^n, \xi^m)_*\mathbb{P} \in \Pi_{\mathrm{BC}}(\mu_n, \mu_m)$, da cui osservato che

$$\begin{aligned}
\mathbb{E}^{\mathbb{P}}\big[\|\xi^n - \xi^m\|_1\big] &= \mathbb{E}^{\mathbb{P}}\big[|\xi_1^n - \xi_1^m| + |\xi_2^n - \xi_2^m|\big] \\
&= \frac{1}{2}\left|\frac{1}{n} - \frac{1}{m}\right| + \frac{1}{2}\left|-\frac{1}{n} + \frac{1}{m}\right| \\
&= \left|\frac{1}{n} - \frac{1}{m}\right|,
\end{aligned}$$

ricaviamo

$$\mathcal{AW}_1(\mu_n, \mu_m) \le \mathbb{E}^{\mathbb{P}}\big[\|\xi^n - \xi^m\|_1\big] \le \left|\frac{1}{n} - \frac{1}{m}\right| \xrightarrow[n,m\to\infty]{} 0,$$

quindi la successione $(\mu_n)_{n\in\mathbb{N}}$ è di Cauchy per $\mathcal{AW}_1$. Tuttavia l'unico possibile limite di μ_n è la misura

$$\mu = \frac{1}{2}\big(\delta_{(0,1)} + \delta_{(0,-1)}\big),$$

siccome essa è il limite della successione rispetto a $\mathcal{W}_1$. Infatti definendo su $(\Omega, \mathcal{F}, \mathbb{P})$ la v.a.

$$\zeta := \begin{pmatrix} 0 \\ 1 \end{pmatrix} \mathbf{1}_A + \begin{pmatrix} 0 \\ -1 \end{pmatrix} \mathbf{1}_{A^c},$$

si ha $\zeta \sim \mu$, quindi $(\xi^n, \zeta)_* \mathbb{P} \in \Pi(\mu_n, \mu)$, e ragionando come sopra abbiamo

$$\mathcal{W}_1(\mu_n, \mu) \leq \mathbb{E}^{\mathbb{P}}\big[\|\xi^n - \zeta\|_1\big] = \frac{1}{n} \xrightarrow[n\to\infty]{} 0.$$

Però si calcola, con i metodi spiegati nella sezione 4 di questo capitolo, che

$$\mathcal{AW}_1(\mu_n, \mu) = \left(1 + \frac{1}{n}\right) > 1.$$

Quindi la successione non converge rispetto a $\mathcal{AW}_1$ che dunque non è una metrica completa.

Per identificare il completamento di $\mathcal{P}_r(E^T, \mathrm{d}_{E^T})$ dobbiamo estendere il concetto di misura di probabilità su E^T. Ciò si fa introducendo le cosiddette distribuzioni annidate.

Definizione 2.5. Sia (E, d) uno spazio Polacco, $r \geq 1$ e $T \in \mathbb{N}$. Definiamo ricorsivamente i seguenti spazi Polacchi:

- $\mathcal{X}_{T:T} := E$ dotato della distanza $\mathrm{d}_{T:T} := \mathrm{d}$.
- $\mathcal{X}_{T-1:T} := E \times \mathcal{P}_r(E, \mathrm{d})$ dotato della distanza $\mathrm{d}_{T-1:T} := \big[\mathrm{d}^r + \mathcal{W}^r_{\mathrm{d},r}\big]^{1/r}$, ove $\mathcal{W}_{\mathrm{d},r}$ indica la distanza di Wasserstein di ordine r costruita a partire da d.
- ...
- $\mathcal{X}_{t-1:T} := \mathcal{X}_{t:T} \times \mathcal{P}_r(\mathcal{X}_{t:T}, \mathrm{d}_{t:T})$ dotato della distanza $\mathrm{d}_{t-1:T} := \big[\mathrm{d}^r_{t:T} + \mathcal{W}^r_{\mathrm{d}_{t:T},r}\big]^{1/r}$, ove $\mathcal{W}_{\mathrm{d}_{t:T},r}$ indica la distanza di Wasserstein di ordine r costruita a partire da $\mathrm{d}_{t:T}$.
- ...
- $\mathcal{X}_{1:T} := \mathcal{X}_{2:T} \times \mathcal{P}_r(\mathcal{X}_{2:T}, \mathrm{d}_{2:T})$ dotato della distanza $\mathrm{d}_{1:T} := \big[\mathrm{d}^r_{2:T} + \mathcal{W}^r_{\mathrm{d}_{2:T},r}\big]^{1/r}$.

Si definisce *spazio delle distribuzioni annidate di profondità* T l'insieme: $\mathcal{P}_r(\mathcal{X}_{1:T}, \mathrm{d}_{1:T})$, che dotiamo della metrica completa $\mathcal{W}_{\mathrm{d}_{1:T},r}$.

Esempio 2.4. Nel caso in cui $T = 2$, abbiamo che $\mathcal{X}_{1:2} = E \times \mathcal{P}_r(E, \mathrm{d})$ e per $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_r(\mathcal{X}_{1:2}, \mathrm{d}_{1:2})$ la distanza tra $\mathbb{P}$ e $\mathbb{Q}$ è:

$$\mathcal{W}_{\mathrm{d}_{1:2},r}(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\Gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \iint_{\mathcal{X}_{1:2} \times \mathcal{X}_{1:2}} \Big(\mathrm{d}(x, y)^r + \mathcal{W}_{\mathrm{d},r}(\mu, \nu)^r \Big)\, d\Gamma(x, y, \mu, \nu) \right)^{1/r}$$

Proviamo ora che le distribuzioni annidate introdotte sopra estendono la nozione di misura di probabilità su $\hat{E}$ in un modo metricamente significativo. Introduciamo la seguente funzione:

$$\mathscr{I}\colon \mathcal{P}_r(E^T, \mathrm{d}_{E^T}) \to \mathcal{P}_r(\mathcal{X}_{1:T}, \mathrm{d}_{1:T}) \quad \text{definita tramite}$$
$$\mu \mapsto \mathscr{I}[\mu] := \mathcal{L}\Big(\xi_1, \mathcal{L}^{\xi_1}\Big(\xi_2, \dots, \mathcal{L}^{\xi_{1:T-2}}\big(\xi_{T-1}, \mathcal{L}^{\xi_{1:T-1}}(\xi_T)\big)\dots\Big)\Big) \tag{2.16}$$

ove $(\xi_t)_{t=1,\dots,T}$ è un processo stocastico a valori in E e con legge μ. Abbiamo usato la notazione abbreviata $\mathcal{L}^{\xi_{1:t}}$ per indicare la legge condizionale dato $(\xi_1, \dots, \xi_t)$. Per esempio, $\mathcal{L}^{\xi_{1:T-1}}(\xi_T)$ è la distribuzione condizionale di ξ_T dato il passato fino a $T-1$, poi $\mathcal{L}^{\xi_{1:T-2}}\big(\xi_{T-1}, \mathcal{L}^{\xi_{1:t-1}}(\xi_T)\big)$ è la legge congiunta di $\mathcal{L}^{\xi_{1:T-1}}(\xi_T)$ e ξ_{T-1} dato il passato fino a $T-2$. La distribuzione annidata $\mathscr{I}[\mu]$ è ottenuta ripetendo questa procedura indietro nel tempo. Per il caso significativo $T = 2$, abbiamo

$$\mathscr{I}[\mu] = \mathcal{L}\big(\xi_1, \mathcal{L}^{\xi_1}(\xi_2)\big).$$

Osserviamo che effettivamente si ha che $\mathscr{I}[\mu] \in \mathcal{P}(\mathcal{X}_{1:2}, \mathrm{d}_{1:2})$, infatti $\mathcal{L}^{\xi_1}(\xi_2)$ è effetivamente una variabile aleatoria a valori in $\mathcal{P}_r(E, \mathrm{d})$. In termini di μ possiamo dire, per come è definita $\mathcal{L}^{\xi_1}(\xi_2)$ che $\mathscr{I}[\mu]$ è la legge rispetto a $p^1_*\mu$ della variabile aleatoria $x_1 \mapsto (x_1, \mu^{x_1})$. Mostriamo che la funzione $\mathscr{I}$ è un'isometria.

Teorema 2.7. *Nel contesto precedente, la funzione $\mathscr{I}$ definita in* (2.16) *è una immersione isometrica di $(\mathcal{P}_r(E^T, \mathrm{d}_{E^T}), \mathcal{AW}_r)$ nello spazio polacco $(\mathcal{P}_r(\mathcal{X}_{1:T}, \mathrm{d}_{1:T}), \mathcal{W}_{\mathrm{d}_{1:T},r})$. In particolare quindi $(\mathcal{P}_r(E^T, \mathrm{d}_{E^T}), \mathcal{AW}_r)$ è uno spazio separabile.*

Dimostrazione. Svolgiamo la dimostrazione solo per il caso $T = 2$. Fissiamo $\mu \in \mathcal{P}_r(E^2, \mathrm{d}_{E^2})$. Mostriamo che $\mathscr{I}[\mu] \in \mathcal{P}_r(\mathcal{X}_{1:2}, \mathrm{d}_{1:2})$. Per farlo disintegriamo μ rispetto alla prima coordinata, ovvero scriviamo per ogni $A, B \in \mathscr{B}(E)$

$$\mu(A \times B) = \int_A \mu^{x_1}(B)\, dp^1_*\mu(x_1),$$

indichiamo con T_μ la funzione misurabile

$$T_\mu\colon E \to \mathcal{P}(E),$$
$$x_1 \mapsto \mu^{x_1}.$$

Abbiamo allora visto che nel caso $T = 2$, $\mathscr{I}[\mu]$ è definita, dati $A \in \mathscr{B}(E), B \in \mathscr{B}(\mathcal{P}(E))$, da:

$$\mathscr{I}[\mu](A \times B) = p^1_*\mu\big(A \cap T_\mu^{-1}(B)\big).$$

Scegliamo $x_0 \in E$ tale che

$$\iint_{E\times E} \mathrm{d}_{E\times E}\big((x,y),(x_0,x_0)\big)^r\, d\mu(x,y) < \infty.$$

Allora, siccome $\mathrm{d}(x,x_0),\mathrm{d}(y,x_0)\le \mathrm{d}_{E\times E}((x,y),(x_0,x_0))$, si ha

$$\int_E \mathrm{d}(x,x_0)^r\,dp^1_*\mu(x)=\iint_{E\times E}\mathrm{d}(x,x_0)^r\,d\mu(x,y)\le\iint_{E\times E}\mathrm{d}_{E\times E}((x,y),(x_0,x_0))^r\,d\mu(x,y)<\infty$$

e

$$\int_E \mathrm{d}(y,x_0)^r\,dp^2_*\mu(y)=\iint_{E\times E}\mathrm{d}(y,x_0)^r\,d\mu(x,y)\le\iint_{E\times E}\mathrm{d}_{E\times E}((x,y),(x_0,x_0))^r\,d\mu(x,y)<\infty$$

da cui segue

$$\iint_{E\times E}\left[\mathrm{d}(x_1,x_0)^r+\mathrm{d}(x_2,x_0)^r\right]d\mu(x_1,x_2)<\infty.$$

Quindi

$$\begin{aligned}\int_{\mathcal{X}_{1:2}\times\mathcal{X}_{1:2}}\left[\mathrm{d}(x,x_0)^r+\mathcal{W}_r(\nu,\delta_{x_0})^r\right]d\mathscr{I}[\mu](x,\nu)&=\int_E\left[\mathrm{d}(x_1,x_0)^r+\mathcal{W}_r(\mu^{x_1},\delta_{x_0})^r\right]dp^1_*\mu(x_1)\\&=\int_E\left[\mathrm{d}(x_1,x_0)^r+\int_E\mathrm{d}(x_2,x_0)^r\,d\mu^{x_1}(x_2)\right]dp^1_*\mu(x_1)\\&=\iint_{E\times E}\left[\mathrm{d}(x_1,x_0)^r+\mathrm{d}(x_2,x_0)^r\right]d\mu(x_1,x_2)<\infty.\end{aligned}$$

Ove l'ultima uguaglianza segue dalla proprietà sulle marginali applicata al primo addendo nell'integrale e dalla definizione di nucleo regolare, infatti

$$\int_E\int_E\mathrm{d}(x_2,x_0)^r\,d\mu^{x_1}(x_2)\,dp^1_*\mu(x_1)=\iint_{E\times E}\mathrm{d}(x_2,x_0)^r\,d\mu(x_1,x_2).$$

Dunque $\mathscr{I}[\mu]\in\mathcal{P}_r(\mathcal{X}_{1:2},\mathrm{d}_{1:2})$. Proviamo ora che $\mathscr{I}$ è un'isometria. A tale scopo è necessario osservare che date $\mu,\nu\in\mathcal{P}_r(E^2,\mathrm{d}_{E^2})$ ogni coupling tra $\mathscr{I}[\mu]$ e $\mathscr{I}[\nu]$, ovvero ogni elemento $\Gamma\in\Pi(\mathscr{I}[\mu],\mathscr{I}[\nu])$ è della forma $d\delta_{\mu^{x_1}}(M)d\delta_{\nu^{y_1}}(N)d\overline{\gamma}(x_1,y_1)$ per qualche $\overline{\gamma}\in\Pi(p^1_*\mu,p^1_*\nu)$ e viceversa. Questo è vero poichè sappiamo che $\mathscr{I}[\mu]$ e $\mathscr{I}[\nu]$ sono le leggi, rispetto a $p^1_*\mu,p^1_*\nu$, delle variabili aleatorie $x_1\mapsto(x_1,\mu^{x_1})$ e $y_1\mapsto(y_1,\nu^{y_1})$ rispettivamente, quindi i coupling di $\mathscr{I}[\mu],\mathscr{I}[\nu]$ sono in corrispondenza biunivoca con le misure immagine degli elementi $\Pi(p^1_*\mu,p^1_*\nu)$ tramite la variabile aleatoria $(x_1,y_1)\mapsto((x_1,\mu^{x_1}),(y_1,\nu^{y_1}))$, ovvero detta S tale variabile aleatoria, abbiamo che per ogni $\Gamma\in\Pi(\mathscr{I}[\mu],\mathscr{I}[\nu])$ esiste $\overline{\gamma}\in\Pi(p^1_*\mu,p^1_*\nu)$ tale che

$$\Gamma=\overline{\gamma}\circ S^{-1}.$$

Questo ci dice in particolare che per ogni scelta di $A,B\in\mathscr{B}(E),M,N\in\mathscr{B}(\mathcal{P}_r(E,\mathrm{d}))$ si ha

$$\Gamma((A\times M)\times(B\times N))=\iint_{E\times E}\delta_{\mu^{x_1}}(M)\delta_{\nu^{y_1}}(N)\,d\overline{\gamma}(x_1,x_2).$$

Dunque si ha

$$\begin{aligned}\mathcal{W}_{\mathrm{d}_{1:T},r}(\mathscr{I}[\mu],\mathscr{I}[\nu])^r&=\inf_{\overline{\gamma}\in\Pi(p^1_*\mu,p^1_*\nu)}\iint_{E\times E}\left[\mathrm{d}(x_1,y_1)^r+\mathcal{W}_r(\mu^{x_1},\nu^{y_1})^r\right]d\overline{\gamma}(x_1,y_1)\\&=\mathcal{AW}_r(\mu,\nu)^r,\end{aligned}$$

ove l'ultimo passaggio è ottenuto usando (2.15). Quindi $\mathscr{I}$ è un'isometria. Infine per concludere la dimostrazione è sufficiente osservare che, da questo, otteniamo che $(\mathcal{P}_r(E^T, \mathrm{d}_{E^T}), \mathcal{AW}_r)$ è isometrico, quindi in particolare omeomorfo, ad un sottoinsieme di uno spazio separabile ed è quindi esso stesso separabile. □

Teorema 2.8 (Completamento)**.** *Se (E, d) è uno spazio polacco senza punti isolati, $r \geq 1$, allora lo spazio $(\mathcal{P}_r(\mathcal{X}_{1:T}, \mathrm{d}_{1:T}), \mathcal{W}_{\mathrm{d}_{1:T},r})$ è il completamento metrico dello spazio $(\mathcal{P}_r(E^T, \mathrm{d}_{E^T}), \mathcal{AW}_r)$.*

Dimostrazione. Come al solito per semplicità notazionale limitiamo la dimostrazione al caso $T = 2$. Per dimostrare il teorema è sufficiente mostrare che l'immagine dell'isometria considerata in precedenza $\mathscr{I}$ è densa. Dal capitolo 1 sappiamo che le combinazioni convesse finite di misure di Dirac sono dense in $(\mathcal{P}_r(\mathcal{X}_{1:2}, \mathrm{d}_{1:2}), \mathcal{W}_{\mathrm{d}_{1:2},r})$, quindi è sufficiente mostrare che tale insieme è contenuto nella chiusura dell'immagine di $\mathscr{I}$. A tale scopo fissiamo una k-upla di punti in E, $A := (a_1, \ldots, a_k)$ e k misure su E: $m_1, \ldots, m_k \in \mathcal{P}_r(E, \mathrm{d})$. Fissati dei coefficienti $(\lambda_j)_{j=1,\ldots,k}$ tali che $0 \leq \lambda_j \leq 1$ e $\sum_{j=1}^k \lambda_j = 1$ consideriamo la misura su $\mathcal{X}_{1:2}$ definita da

$$d\mathbb{P}(x, m) := \sum_{j=1}^k \lambda_j d\delta_{(a_j, m_j)}(x, m).$$

È sufficiente allora scegliere una qualsiasi successione $(A_n)_{n\in\mathbb{N}}$ tale che $A_n := (a_1^n, \ldots, a_k^n) \xrightarrow[n\to\infty]{} A$ e per cui, per ogni n fissato, tutte le componenti di A_n siano distinte. Definiamo allora, per ogni $n \in \mathbb{N}$, le misure $\mu_n \in \mathcal{P}_r(E \times E, \mathrm{d}_{E\times E})$ in modo che la loro prima marginale sia $\sum_{j=1}^k \lambda_j \delta_{a_j^n}$ e tali che (usando la notazione della precedente dimostrazione) $T_{\mu_n}(a_j^n) = m_j$. Allora grazie al fatto che gli a_j^n sono distinti per n fissato, abbiamo

$$\mathscr{I}[\mu_n] = \sum_{j=1}^k \lambda_j \delta_{(a_j^n, m_j)}.$$

Dobbiamo ora stimare la distanza

$$\mathcal{W}_{\mathrm{d}_{1:2},r}(\mathscr{I}[\mu_n], \mathbb{P})$$

Per farlo facciamo un ragionamento analogo a quello svolto nella dimostrazione del teorema 1.12 e nell'esempio 2.3. Consideriamo delle variabili aleatorie su uno spazio di probabilità $(\Omega, \mathcal{G}, \mathbb{Q})$ a valori in $\mathcal{X}_{1:2}$ tali che abbiano leggi $\mathbb{P}$ e $\mathscr{I}[\mu_n]$ rispettivamente. A tale scopo scegliamo degli eventi disgiunti $A_1, \ldots, A_k \in \mathcal{G}$ tali che

$$\mathbb{Q}(A_j) = \lambda_j \quad \forall j = 1, \ldots, k.$$

Definiamo allora

$$Y_n := \sum_{j=1}^k a_j^n \mathbf{1}_{A_j}, \quad Y := \sum_{j=1}^k a_j \mathbf{1}_{A_j}$$

e

$$Z = Z_n = \sum_{j=1}^{k} m_j \mathbf{1}_{A_j},$$

allora ovviamente $(Y, Z) \sim \mathbb{P}$ e $(Y_n, Z_n) \sim \mathscr{I}[\mu_n]$. Ma allora

$$\begin{aligned}\mathcal{W}_{\mathrm{d}_{1:2},r}(\mathscr{I}[\mu_n], \mathbb{P}) &\leq \mathbb{E}^{\mathbb{Q}}\big[\mathrm{d}_{1:2}((Y_n, Z_n), (Y, Z))^r\big] \\ &= \mathbb{E}^{\mathbb{Q}}\Big[\mathrm{d}(Y_n, Y)^r + \mathcal{W}_r(Z_n, Z)^r\Big] = \mathbb{E}^{\mathbb{Q}}\big[\mathrm{d}(Y_n, Y)^r\big] \\ &= \sum_{j=1}^{k} \lambda_j \mathrm{d}(a_j^n, a_j) \xrightarrow[n\to\infty]{} 0\end{aligned}$$

Ovvero $\mathscr{I}[\mu_n] \xrightarrow[n\to\infty]{} \mathbb{P}$ rispetto a $\mathcal{W}_{\mathrm{d}_{1:2},r}$. Cioè $\mathbb{P} \in \overline{\mathrm{Im}(\mathscr{I})}$ come si voleva e si conclude. □

2.4 Il caso discreto e i processi ad albero.

Nella sezione 4 del capitolo 1 abbiamo visto come il problema di ottimizzazione che definisce $\mathcal{W}_r$ si riduce a un programma lineare, quando si considerano misure discrete con supporto finito. Per ottenere un risultato simile per la distanza adattata usiamo degli alberi di scenari per modellare gli spazi filtrati finiti che considereremo. Questo è possibile in quanto si può sempre associare ad uno spazio finito dotato di una filtrazione, un albero di scenari e viceversa. Nella prima parte di questa sezione spieghiamo cosa intendiamo per albero di scenari e poi mostriamo l'equivalenza suddetta.

Definizione 2.6. Un *albero di scenari*, su $T \in \mathbb{N}$ periodi, è un grafo orientato finito $(\mathcal{V}, \mathcal{A})$ e aciclico con una sola radice, indicata con 0, in cui tutte le foglie sono a profondità T.

Notazioni: Dato un albero di scenari $(\mathcal{V}, \mathcal{A})$ indichiamo, per ogni $t = 0, \ldots, T$, $\mathcal{V}_t$ come l'insieme dei nodi a profondità t, diciamo che un nodo $m \in \mathcal{V}$ è un *diretto predecessore* del nodo $n \in \mathcal{V}$, o che n è un *diretto successore* di m, se $(n, m) \in \mathcal{A}$, in tal caso denotiamo m con $n-$. L'insieme di tutti i *diretti successori* di un nodo $n \in \mathcal{V}$ si indica con $n+$. Se dati due nodi $n, m \in \mathcal{V}$ esiste una sequenza (finita) di nodi $(n_j)_{j=1,\ldots,k}$ tale che $n_1 \in m+, n_2 \in n_1+, \ldots, n \in n_j+$ allora diciamo che n è un *successore* di m e che m è un *predecessore* di n, in tal caso scriviamo $n \succ m$.

Proposizione 2.9. *Sia Ω un'insieme finito dotato di una filtrazione $\mathfrak{F} = (\mathcal{F}_t)_{t=0,1,\ldots,T}$, in cui $\mathcal{F}_0 = \{\emptyset, \Omega\}$ e $\mathcal{F}_T = P(\Omega)$. Allora a $(\Omega, \mathfrak{F})$ è canonicamente associato un albero di scenari $(\mathcal{V}, \mathcal{A})$. Viceversa dato un albero di scenari $(\mathcal{V}, \mathcal{A})$ è sempre possibile associargli canonicamente uno spazio filtrato $(\Omega, \mathfrak{F})$.*

Dimostrazione. Supponiamo sia dato uno spazio filtrato $(\Omega, (\mathcal{F}_t)_{t=0,1,\dots,T})$ finito. Per ogni $0 \leq t \leq T$ definiamo l'insieme A_t come l'insieme contenente tutti i sottoinsiemi di Ω che generano $\mathcal{F}_t$. Si definiscano allora i nodi

$$\mathcal{V} := \{(a,t) \mid a \in A_t\}$$

e gli archi

$$\mathcal{A} := \{((a,t),(b,t+1)) \mid a \in A_t, b \in A_{t+1}, b \subseteq a\}$$

allora $(\mathcal{V}, \mathcal{A})$ è un albero di scenari che rappresenta la filtrazione $\mathfrak{F}$. Viceversa dato un albero di scenari su T periodi, $(\mathcal{V}, \mathcal{A})$ definiamo

$$\Omega := \mathcal{V}_T$$

e per ogni nodo $n \in \mathcal{V}$ gli insiemi

$$a(n) := \begin{cases} \{n\} & \text{se } n \in \mathcal{V}_T, \\ \bigcup_{j \in n+} a(j) & \text{altrimenti.} \end{cases}$$

e poi per ogni $0 \leq t \leq T$ poniamo

$$\mathcal{F}_t = \sigma(a(n) \mid n \in \mathcal{V}_t).$$

Dalla costruzione degli insiemi $a(n)$ risulta allora evidente che $\mathcal{F}_0 = \{\emptyset, \Omega\}, \mathcal{F}_T = P(\Omega)$ e che $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$, quindi $(\Omega, (\mathcal{F}_t)_{t=0,1,\dots,T})$ è uno spazio filtrato con le proprietà volute. □

Osservazione 2.8. Si noti in particolare che gli elementi di Ω sono in biiezione con le foglie dell'albero $(\mathcal{V}, \mathcal{A})$, mentre i nodi intermedi corrispondono ai sottoinsiemi di Ω formati dall'unione dei singoletti che corrispondono alle foglie dell'albero che sono successori del nodo intermedio considerato.

Supponiamo ora che sullo spazio filtrato finito $(\Omega, (\mathcal{F}_t)_{t \in \mathbf{T}})$ sia assegnata una misura di probabilità $\mathbb{P}$, che quindi è univocamente determinata dal suo valore p_ω sui singoletti $\{\omega\}$. Grazie alla proposizione precedente siamo in grado da $(\Omega, (\mathcal{F}_t)_{t \in \mathbf{T}})$ di costruire un albero degli scenari $(\mathcal{V}, \mathcal{A})$ che come visto verifica $\mathcal{V}_T = \Omega$. In virtù del fatto che i nodi intermedi corrispondono all'unione degli insiemi corrispondenti ai loro diretti successori, possiamo definire la probabilità di un nodo intermedio m come

$$\mathbb{P}(m) := \sum_{\substack{n \succ m \\ n \in \mathcal{V}_T}} \mathbb{P}(\{n\}).$$

Ove $\mathbb{P}(\{n\})$ è definita come la probailità del singoletto $\{\omega\}$ che corrisponde alla foglia n. Da questa osservazione possiamo definire allora le probabilità condizionali di un nodo m dato un suo predecessore h, tramite

$$\mathbb{P}(m \mid h) := \frac{\mathbb{P}(m)}{\mathbb{P}(h)}.$$

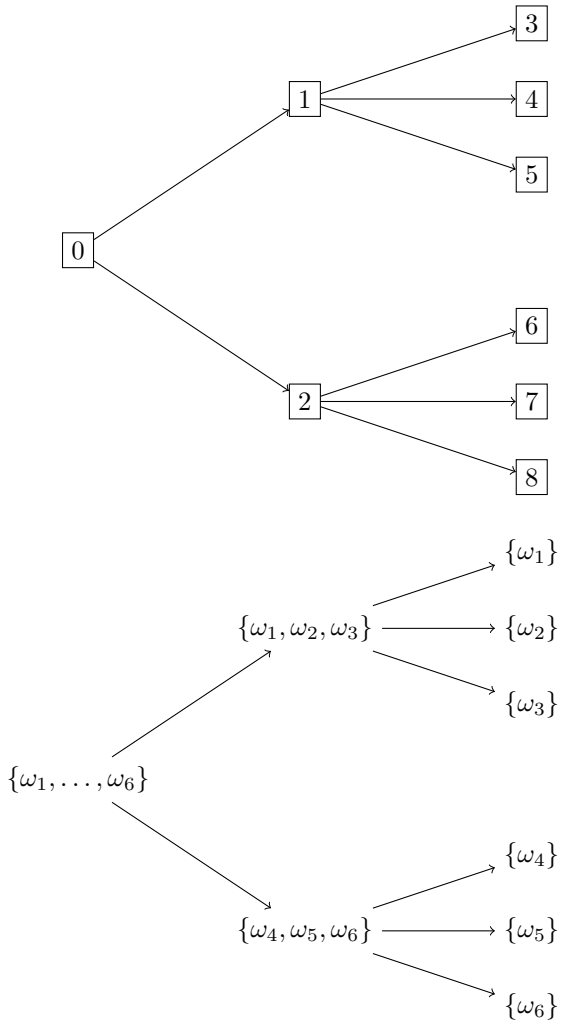

Figura 2.2: Esempio di un albero di scenari e il suo spazio filtrato equivalente. Come si vede le foglie corrispondono ai singoletti di Ω, mentre i nodi intermedi sono associati all'unione degli insiemi che corrispondono ai loro diretti successori. La filtrazione associata è per definizione $\mathcal{F}_0 = \sigma(\Omega) = \{\emptyset, \Omega\}, \mathcal{F}_1 = \sigma(\{\omega_1, \omega_2, \omega_3\}, \{\omega_4, \omega_5, \omega_6\}), \mathcal{F}_2 = \sigma(\{\omega_1\}, \ldots, \{\omega_6\}) = P(\Omega)$, vediamo quindi anche come i nodi a profondità $0 \leq t \leq 2$, corrispondono ai generatori di $\mathcal{F}_t$.

Si noti che ciò è coerente, perchè tale probabilità corrisponde esattamente alla probabilità condizionale dell'evento che corrisponde a m, dato quello corrispondente a h. Infatti se m è un successore di h allora $a(m) \subseteq a(h)$. Se poi $h = m-$, la probabilità $\mathbb{P}(m \mid m-)$ può essere interpretata come la probabilità di percorrere l'arco $(m-, m)$. Possiamo quindi pensare a tale probabilità come assegnata a tale arco. Se invece, a ciasun arco $(m-, m) \in \mathcal{A}$, assegnamo le probabilità condizionali $\mathbb{P}(m \mid m-)$ la probabilità non condizionale di ciascun nodo m è data da

$$\mathbb{P}(m) = \prod_{k \prec m} \mathbb{P}(k \mid k-).$$

Possiamo quindi indifferentemente assegnare le probabilità solamente agli archi (faremo così) o solamente ai nodi.

Osservazione 2.9. Si noti che poichè ogni processo stocastico a tempo discreto e con stati discreti $(\xi_t)_{t\in\mathbf{T}}$, a valori in uno spazio Polacco (E, d) può essere visto come uno spazio di probabilità finito, il ragionamento precedente mostra anche l'equivalenza tra processi stocastici a tempo discreto e stati discreti e gli alberi di scenari.

Abbiamo introdotto gli alberi di scenari per capire come formulare il problema del calcolo di $\mathcal{AW}_r$ nel caso discreto, dobbiamo quindi estendere la nozione di distanza adattata agli alberi di scenari. Poichè sappiamo che le foglie di tali alberi corrispondono agli elementi dello spazio filtrato associato è necessario capire come definire una distanza tra le foglie dei due alberi. Un primo modo è quello di partire da (X, d), uno spazio Polacco e considerare i due spazi di probabilità filtrati $(X, \mathcal{M}, (\mathcal{M}_t)_{t\in\mathbf{T}}, \mu)$ e $(X, \mathcal{N}, (\mathcal{N}_t)_{t\in\mathbf{T}}, \nu)$, ove μ e ν sono misure date da combinazioni convesse finite di misure di Dirac. Possiamo allora restringere il problema ai supporti di μ e ν ottenendo due spazi di probabilità finiti. Da essi possiamo dunque costruire due alberi di scenari $(\mathcal{V}, \mathcal{A}), (\mathcal{V}', \mathcal{A}')$, per cui è possibile definire una distanza tra le foglie (sarà la distanza tra i corrispondenti elemnti di X). Un altro modo, consiste invece nel supporre che agli alberi di scenari considerati siano associati dei processi ad albero a valori in uno spazio polacco (E, d), ovvero intuitivamente che sia assegnato, a ciascun nodo dell'albero, un valore $x \in E$. Per capire il ragionamento formale diamo inanzitutto la definizione di processo ad albero.

Definizione 2.7. Un processo stocastico $(\eta_t)_{t\in\mathbf{T}}$ definito su uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ si dice un *processo ad albero*, se vale

$$\mathcal{F}_t = \sigma(\eta_s \mid s \leq t) = \sigma(\eta_t) \quad \forall t \in \mathbf{T},$$

ovvero equivalentemente $(\sigma(\eta_t))_{t\in\mathbf{T}}$ è una filtrazione.

Per costruire un processo ad albero a valori in E a partire da un albero di scenari $(\mathcal{V}, \mathcal{A})$, associamo prima di tutto ad ogni nodo $n \in \mathcal{V}$ un valore $\xi(n) \in E$ e definiamo un processo ad albero a valori in $\mathcal{V}$. Questo si fa ponendo

$$\begin{aligned} \eta\colon \{0, \dots, T\} \times \mathcal{V}_T &\to \mathcal{V} \\ (t, i) &\mapsto n \quad \text{se } n \in \mathcal{V}_t \text{ e } i \succ n \end{aligned}$$

allora

$$\begin{aligned} \eta_t\colon \mathcal{V}_T &\to \mathcal{V}_t \\ i &\mapsto \eta(t, i) \end{aligned}$$

è un processo adattato a $(\mathcal{F}_t)_{t\in\mathbf{T}}$, la filtrazione associata a $(\mathcal{V}, \mathcal{A})$, ma di più verifica esattamente

$$\sigma(\eta_t) = \mathcal{F}_t,$$

è quindi un processo ad albero. Per ottenerne uno a valori in E è sufficiente ora definire per ogni $t \in \mathbf{T}$

$$\begin{aligned} \xi_t \colon \mathcal{V}_T &\to E \\ i &\mapsto \xi_t(i) := (\xi \circ \eta_t)(i). \end{aligned}$$

È allora possibile dati due alberi di scenari $(\mathcal{V}, \mathcal{A})$ e $(\mathcal{V}', \mathcal{A}')$, associati a due processi ad albero ξ, ξ' a valori nello stesso spazio Polacco E, definire la distanza tra due coppie di foglie $i \in \mathcal{V}_T, j \in \mathcal{V}'_T$ come la distanza tra le traiettorie ad esse associate dei due processi ad albero considerati. Ovvero

$$\mathrm{d}(i,j) := \mathrm{d}_{E^T}\big((\xi_1(i), \xi_2(i), \dots, \xi_T(i)), (\xi'_1(j), \xi'_2(j), \dots, \xi'_T(j))\big).$$

La distanza adattata tra i due alberi di scenari sarà in questo caso la distanza adattata tra le due leggi dei processi ξ e ξ', indichiamole con μ e ν. In ogni caso indichiamo con $\mathrm{d} := (d_{i,j})_{i,j}$ la matrice contenete tali distanze tra le foglie degli alberi. Ora in virtù del fatto che $\Pi_{\mathrm{BC}}(\mu,\nu) \subseteq \Pi(\mu,\nu)$ sappiamo che i coupling bicausali tra le due misure μ e ν hanno supporto finito e in particolare contenuto nel prodotto cartesiano dei supporti. Possono quindi essere visti come misure discrete sul prodotto cartesiano degli insiemi delle foglie, ossia $\mathcal{V}_T \times \mathcal{V}'_T$. Sono pertanto rappresentati da una matrice $\pi = (\pi_{i,j})_{i,j}$ delle stesse dimensioni di d, in cui

$$\pi_{i,j} := \pi(\{(i,j)\}) \quad \forall (i,j) \in \mathcal{V}_T \times \mathcal{V}'_T$$

Come visto per i singoli alberi, possiamo poi definire la probabilità di una coppia di nodi intermedi $m \in \mathcal{V}_t, n \in \mathcal{V}'_t$ come

$$\pi_{m,n} := \sum_{\substack{i \in \mathcal{V}_T \\ i \succ m}} \sum_{\substack{j \in \mathcal{V}'_t \\ j \succ n}} \pi_{i,j},$$

e date due coppie di nodi $i, m \in \mathcal{V}, j, n \in \mathcal{V}'$, con $i \succ m, j \succ n$, definiamo la probabilità condizionale della coppia (i,j) data (m,n) come

$$\pi(i, j \mid m, n) = \frac{\pi_{i,j}}{\pi_{m,n}}.$$

Teorema 2.10. *Dati due spazi filtrati finiti $(X, \mathcal{M}, (\mathcal{M}_t)_{t\in\mathbf{T}}, \mu)$ e $(X, \mathcal{N}, (\mathcal{N}_t)_{t\in\mathbf{T}}, \nu)$, con (X, d) uno spazio Polacco, detti $(\mathcal{V}, \mathcal{A})$ e $(\mathcal{V}', \mathcal{A}')$ i due alberi di scenari associati. O equivalentemente dati i due alberi di scenari, e due processi ad albero ξ e η costruiti su di essi. Usando le notazioni introdotte in questa sezione, il valore di*

$\mathcal{AW}_r(\mu,\nu)^r$ eguaglia il valore ottimo del seguente problema di ottimizzazione

$$\begin{aligned}
&\textit{MIN (in } \pi\textit{)} && \sum_{i\in\mathcal{V}_T}\sum_{j\in\mathcal{V}'_T}\pi_{i,j}\mathrm{d}_{i,j}^r && \\
&\textit{soggetto a} && \sum_{n\prec j}\pi(i,j\mid m,n)=\mu(i\mid m) && \forall m\prec i,\\
& && \sum_{m\prec i}\pi(i,j\mid m,n)=\nu(j\mid n) && \forall n\prec i,\\
& && \pi_{h,k}\geq 0 && \forall h\in\mathcal{V}_T, k\in\mathcal{V}'_T,\\
& && \sum_{h\in\mathcal{V}_T}\sum_{k\in\mathcal{V}'_T}\pi_{h,k}=1. &&
\end{aligned}$$

Dove $i\in\mathcal{V}_t, j\in\mathcal{V}'_t$ sono arbitrari nodi intermedi, per $t=1,\dots,T$.

Dimostrazione. La prova del fatto che la funzione da minimizzare è quella nell'enunciato è uguale a quella fatta per $\mathcal{W}_r$, come anche gli ultimi vincoli del problema corrispondono alla richiesta che la matrice $\pi=(\pi_{h,k})_{h,k}$ rappresenti una probabilità sull'insieme $\mathcal{V}_T\times\mathcal{V}'_T$. Per dimostrare il teorema è sufficiente provare l'equivalenza tra la condizione

$$\pi(A\times X\mid\mathcal{M}_t\otimes\mathcal{N}_t)=\mu(A\mid\mathcal{M}_t)\quad\forall A\in\mathcal{M}, t=0,1,\dots,T-1,$$

e

$$\sum_{n\prec j}\pi(i,j\mid m,n)=\mu(i\mid m)\quad\forall m\prec i, i\in\mathcal{V}_t, j\in\mathcal{V}'_t\quad\forall t=1,\dots,T,$$

tuttavia questo è facile in quanto basta osservare che siccome le σ-algebre considerate sono finitamente generate i valori attesi condizionali sopra assumono un unico valore su ogni generatore, pari al valore atteso calcolato usando le probabilità condizionali rispetto a tale generatore. Quindi basta uguagliare tali valori e si ottengono esattamente le condizioni dell'enunciato. La prova dell'equivalenza per le condizioni su ν sono analoghe. □

Osservazione 2.10. Si noti che la condizione $\sum_{i\in\mathcal{V}_T}\sum_{j\in\mathcal{V}'_T}\pi_{i,j}=1$, non può essere omessa al fine che la matrice π rappresenti una probabilità, a differenza del caso di $\mathcal{W}_r$, in quanto i vincoli sono lineari in $\pi(i,j\mid m,n)$, che per definizione sono

$$\pi(i,j\mid m,n)=\frac{\pi_{i,j}}{\pi_{m,n}}=\frac{\pi_{i,j}}{\displaystyle\sum_{\substack{i\in\mathcal{V}_T\\ i\succ m}}\sum_{\substack{j\in\mathcal{V}'_t\\ j\succ n}}\pi_{i,j}}$$

quindi senza la condizione di somma a 1, dato un vettore ammissibile π, anche $\lambda\pi$ sarebbe ammissibile, per ogni $\lambda>0$. Questa osservazione ci fa notare anche che il problema di ottimizzazione enunciato nel teorema precedente non è un programma lineare perchè nei vincoli sono appunto coinvolti dei quozienti.

Osserviamo però che, in virtù della possibilità di calcolare $\mathcal{AW}_r$ ricorsivamente, come spiegato in 2.6, possiamo anzichè risolvere il problema (2.10), adottare un algoritmo ricorsivo per calcolare la distanza adattata tra due alberi. Infatti siano dati due alberi di scenari $(\mathcal{V}, \mathcal{A})$, $(\mathcal{V}', \mathcal{A}')$ associati ai processi ad albero ξ e ξ', di leggi μ e ν rispettivamente. Definiamo ricorsivamente le quantità d_t, per $t \in \mathbf{T}$ come segue. Per ogni coppia di foglie $i \in \mathcal{V}_T, j \in \mathcal{V}'_T$ poniamo

$$\mathrm{d}_T(i,j) := \mathrm{d}(i,j) = \mathrm{d}_{E^T}\big((\xi_1(i), \xi_2(i), \dots, \xi_T(i)), (\xi'_1(j), \xi'_2(j), \dots, \xi'_T(j))\big).$$

Poi per $0 \leq t \leq T-1$, date le quantità $\mathrm{d}_{t+1}(n,m)$, per ogni coppia di nodi $n \in \mathcal{V}_{t+1}, m \in \mathcal{V}'_{t+1}$, definiamo, per ogni $h \in \mathcal{V}_t, k \in \mathcal{V}'_t$, la quantità $\mathrm{d}_t(h,k)^r$ come il valore ottimo del seguente programma lineare

$$\begin{aligned}
\text{MIN (in } \pi_t(\cdot,\cdot \mid h,k)) \quad & \sum_{l \in h+} \sum_{u \in k+} \pi_t(l,u \mid h,k) \mathrm{d}_{t+1}(l,u)^r & \\
\text{soggetto a} \quad & \sum_{u \in k+} \pi_t(l,u \mid h,k) = \mu(l \mid h) & \forall l \in h+ \\
& \sum_{l \in h+} \pi_t(l,u \mid h,k) = \nu(u \mid k) & \forall u \in k+ \\
& \pi_t(l,u \mid h,k) \geq 0 & \forall l \in h+, u \in k+ .
\end{aligned} \tag{2.17}$$

Allora abbiamo che

$$\mathcal{AW}_r(\mu,\nu) = \mathrm{d}_0(0,0'),$$

ove 0 e $0'$ sono le radici dei due alberi. Le quantità $\mathrm{d}_t(n,m)$ si possono interpretare come la distanza adattata tra le leggi dei processi ad albero associati ai sottoalberi dei due dati che hanno n ed m come radici. Indicando con $\pi^*_t(\cdot,\cdot \mid h,k)$ l'ottimizzante del problema (2.17), la misura ottima π^* per $\mathcal{AW}_r(\mu,\nu)$ si ricostruisce per ogni coppia di foglie $i \in \mathcal{V}_T, j \in \mathcal{V}'_T$ tramite

$$\pi^*(i,j) = \pi^*_0(i_1, j_1 \mid 0, 0') \cdot \dots \cdot \pi^*_{T-1}(i_{T-1}, j_{T-1} \mid i_{T-2}, j_{T-2}) \pi^*_T(i,j \mid i_{T-1}, j_{T-1}),$$

ove abbiamo usato le notazioni, per ogni $1 \leq t \leq T$

$$i_t := \underbrace{(\dots((i-)-)\dots)-}_{(T-t)-volte}$$

e analogamente per j. Nell'appendice B è fornita una implementazione MATLAB del precedente algoritmo per il calcolo di $\mathcal{AW}_r$ tra due processi ad albero dati.

Capitolo 3

Un'applicazione

In questo capitolo presentiamo un'applicazione della *distanza adattata* definita nel precedente. In particolare mostriamo come essa può essere utilizzata per ottenere degli sviluppi al primo ordine per il massimo errore che si commette nel calcolare il valore ottimo del problema di ottimizzazione stocastica di minimizzazione del valore atteso di una funzione convessa sotto controlli predicibili, applicando piccole variazioni al modello, ovvero quando ci si muove in un intorno della misura data. Il problema di massimizzazione dell'utilità attesa, cruciale in finanza matematica, si può scrivere in tale forma e verrà trattato come un caso speciale. Tutti i risultati presentati in questo capitolo sono tratti dalla sezione centrale di [3] (Teorema 2.4 e Corollario 2.7).

3.1 Contesto e Assunzioni

Contesto e notazioni: sia $T \in \mathbb{N}$, indichiamo come al solito con $\mathbf{T} = \{0, 1, \dots, T\}$ l'insieme dei tempi, con $\mathcal{P}_p(\mathbb{R}^T)$ l'insieme delle misure di probabilità su $\mathbb{R}^T$ con momento p-esimo finito, ove $p \geq 1$. Indichiamo con $\xi \colon \mathbb{R}^T \to \mathbb{R}^T$ il processo canonico e allo stesso modo con $\xi, \eta \colon \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}^T$ le proiezioni sulla prima e sulla seconda coordinata. Indichiamo poi con:

$$\mathfrak{F} = (\mathcal{F}_t)_{t \in \mathbf{T}} \quad \text{ove} \quad \mathcal{F}_0 = \{\emptyset, \mathbb{R}^T\}, \quad \mathcal{F}_t = \sigma(\xi_s \mid s \leq t) \quad \forall t = 1, \dots, T$$

la filtrazione canonica su $\mathbb{R}^T$ e similmente, quando consideriamo lo spazio prodotto $\mathbb{R}^T \times \mathbb{R}^T$, indicheremo con: $\tilde{\mathfrak{F}} = (\tilde{\mathcal{F}}_t)_{t \in \mathbf{T}}$ la filtrazione canonica sulla seconda coordinata. In questo contesto dunque il nostro spazio Polacco di riferimento è $\mathbb{R}^T$ pensato dotato della distanza indotta dalla norma ℓ^p, dunque prese $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_p(\mathbb{R}^T)$, abbiamo

$$\mathcal{AW}_p(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\pi \in \Pi_{\mathrm{BC}}(\mu,\nu)} \mathbb{E}^\pi \left[\|\xi - \eta\|_p^p \right] \right)^{1/p} = \left(\inf_{\pi \in \Pi_{\mathrm{BC}}(\mu,\nu)} \sum_{t=1}^{T} \mathbb{E}^\pi \left[|\xi_t - \eta_t|^p \right] \right)^{1/p}.$$

Per le funzioni $f\colon \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}$ indichiamo poi con $\partial_{x_t} f$ la derivata parziale di f rispetto alla t-esima coordinata in x, cioè

$$\partial_{x_t} f(x,a) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\big(f(x+\varepsilon e_t, a) - f(x,a)\big),$$

ove e_t è il t-esimo vettore della base canonica in $\mathbb{R}^T$. Indichiamo poi con $\nabla_a f$ e $\nabla_a^2 f$ rispettivamente il gradiente e l'Hessiana di f nelle seconde T coordinate a. Per le funzioni univariate $\ell\colon \mathbb{R} \to \mathbb{R}$ scriviamo semplicemente ℓ', ℓ'' per le derivate prima e seconda.

Formulazione del problema: sia $f\colon \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}^T$ una funzione convessa nel suo secondo argomento e sia $\mathbb{P} \in \mathcal{P}_p(\mathbb{R}^T)$ fissata. Denotiamo, per $r > 0$,

$$B_r(\mathbb{P}) := \{\mathbb{Q} \in \mathcal{P}_p(\mathbb{R}^T) \mid \mathcal{AW}_p(\mathbb{P}, \mathbb{Q}) \leq r\},$$

Indichiamo con $\mathcal{A}$ l'insieme di tutti i *controlli predicibili uniformemente limitati*, ovvero l'insieme delle T-uple $a = (a_t)_{t=1,\dots,T}$, tali che, per ogni t:

$$a_t\colon \mathbb{R}^T \to \mathbb{R}^T \quad \text{dipende solo da } x_1, \dots, x_{t-1} \text{ e } |a_t| \leq L,$$

ove $L > 0$ è una costante dipendente solo da $\mathcal{A}$. Siamo interessati a studiare la sensitività del problema di ottimizzazione stocastica:

$$\inf_{a \in \mathcal{A}} \mathbb{E}^{\mathbb{Q}}\big[f(\xi, a(\xi))\big],$$

quando $\mathbb{Q}$ varia in $B_r(\mathbb{P})$, per ogni $\mathbb{Q}$ ne indichiamo il valore ottimo con: $v(\mathbb{Q})$. Si noti che il problema di *massimizzazione dell'utilità attesa*, cruciale in finanza matematica, è di tale forma. Infatti data $\ell\colon \mathbb{R} \to \mathbb{R}$ una *funzione di perdita*, ovvero una funzione convessa e limtata dal basso (ℓ è da pensare come la negativa di una funzione di utilità $-U$), e $g\colon \mathbb{R}^T \to \mathbb{R}$ una certa funzione di payoff (o meglio il suo opposto), allora fissato $\xi_0 \in \mathbb{R}$ abbiamo che:

$$u(\mathbb{Q}) := \inf_{a \in \mathcal{A}} \mathbb{E}^{\mathbb{Q}}\Big[\ell\big(g(\xi) + \sum_{t=1}^{T} a_t(\xi)(\xi_t - \xi_{t-1})\big)\Big],$$

corrisponde all'utilità massima ottenibile tramite i controlli $\mathcal{A}$, ovvero è il valore ottimo del problema di massimizzazione dell'utilità con funzione di payoff g ottenuto usando la misura $\mathbb{Q}$.

Assunzioni: Supponiamo che per ogni $x \in \mathbb{R}^T$, la funzione $f(x,\cdot)$ sia due volte differenziabile con continuità, e che soddisfi:

$$\nabla_a^2 f(\xi, \cdot) \succ \varepsilon(\xi)\mathbf{1}_{T\times T} \quad \text{su } [-L, L]^T,$$

con $\mathbb{P}(\varepsilon(\xi) > 0) = 1$. Dove per due matrici $T \times T$ A e B, scriviamo $A \succ B$ se $A - B$ è semidefinita positiva, ossia se per ogni $z \in \mathbb{R}^T \setminus \{0\}$ vale $\langle z, Az\rangle \geq \langle z, Bz\rangle$.

Supponiamo inoltre che $f(\cdot, a)$ sia differenziabile con continuità per ogni $a \in \mathbb{R}^T$, ed esista una costante $C > 0$ tale che:

$$\|\nabla_x f(x,a)\|_q^q = \sum_{t=1}^{T} |\partial_{x_t} f(x,a)|^q \leq C \quad \text{per ogni } x \in \mathbb{R}^T \text{ e } a \in [-L, L]^T.$$

3.2 Risultato principale

Teorema 3.1. *Se, nel contesto precedente, valgono le assunzioni fatte sopra allora esiste unico $a^* \in \mathcal{A}$ tale che:*

$$v(\mathbb{P}) = \mathbb{E}^{\mathbb{P}}\big[f(\xi, a^*(\xi))\big].$$

Inoltre per $r \to 0$, vale:

$$\sup_{\mathbb{Q} \in B_r(\mathbb{P})} v(\mathbb{Q}) = v(\mathbb{P}) + r\left(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\Big[\big|\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big]\big|^q\Big]\right)^{1/q} + o(r),$$

ove $q = \dfrac{p}{p-1}$ è l'esponente coniugato di p.

Per effettuare la dimostrazione abbiamo bisogno di alcuni lemmi preliminari:

Lemma 3.2. *Siano $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_p(\mathbb{R}^T)$ e sia $\pi \in \Pi_C(\mathbb{P}, \mathbb{Q})$. Allora per ogni $\delta > 0$ esiste $\eta^\delta \colon \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}^T$ tale che η_t^δ sia $\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t$-misurabile, ξ_t sia $\sigma(\eta_t^\delta)$-misurabile e $|\eta_t^\delta - \eta_t| \leq \delta$ per ogni $t = 1, \dots, T$. In particolare, $\pi^\delta := (\xi, \eta^\delta)_* \pi$ è un coupling bicausale tra $\mathbb{P}$ e $\mathbb{Q}^\delta := \eta_*^\delta \pi$.*

Dimostrazione. Per $\delta > 0$ fissato consideriamo le mappe Boreliane

$$\psi_\delta \colon \mathbb{R} \to (0, \delta) \quad \text{e} \quad \phi_\delta \colon \mathbb{R} \to \delta\mathbb{Z} := \{\delta k \mid k \in \mathbb{Z}\},$$

ove ψ_δ è un isomorfismo e $\phi_\delta(x) := \max\{\delta k \mid \delta k \leq x\}$. Per $t = 1, \dots, T$ definiamo

$$\eta_t^\delta := \phi_\delta(\eta_t) + \psi_\delta(\xi_t).$$

Per definizione allora ξ_t è $\sigma(\eta_t^\delta)$-misurabile, η_t^δ è $\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t$-misurabile, e $|\eta_t^\delta - \eta_t| \leq \delta$. Sono inoltre chiaramente soddisfatte le condizioni di bicausalità (osservazione 2.6). □

Lemma 3.3. *Nel contesto e con le assunzioni precedenti, sia $(\mathbb{Q}_n)_{n \in \mathbb{N}} \subseteq \mathcal{P}_p(\mathbb{R}^T)$ una successione tale che: $\mathcal{AW}_p(\mathbb{Q}_n, \mathbb{P}) \xrightarrow[n \to \infty]{} 0$, allora:*

$$v(\mathbb{Q}_n) \xrightarrow[n \to \infty]{} v(\mathbb{P}).$$

Dimostrazione. Sia $\mathbb{Q} \in B_r(\mathbb{P})$ e sia $\pi \in \Pi_{\mathrm{BC}}(\mathbb{P}, \mathbb{Q})$ tale che:

$$\Big(\sum_{t=1}^{T} \mathbb{E}^{\pi}\big[|\eta_t - \xi_t|^p\big]\Big)^{1/p} \leq \mathcal{AW}_p(\mathbb{P}, \mathbb{Q}) + o(r) \leq r + o(r).$$

Fissiamo $\varepsilon > 0$ e sia $a \in \mathcal{A}$ tale che:

$$\mathbb{E}^{\mathbb{P}}\big[f(\xi, a(\xi))\big] \leq v(\mathbb{P}) + \varepsilon.$$

Definiamo poi $b \in \mathbb{R}^T$ tramite:

$$b_t := \mathbb{E}^{\pi}\big[a(\xi)\big] \quad \forall t = 1, \dots, T.$$

Allora certamente $|b_t| \leq L$ e siccome b_t è costante, per ogni t, si ha che $b \in \mathcal{A}$. Segue:

$$\begin{aligned} v(\mathbb{Q}) &\leq \mathbb{E}^{\mathbb{Q}}\big[f(\eta, b(\eta))\big] = \mathbb{E}^{\pi}\big[f(\eta, b(\eta))\big] \\ &= \mathbb{E}^{\pi}\Big[f\big(\eta, \mathbb{E}^{\pi}[a(\xi)]\big)\Big] \leq \mathbb{E}^{\pi}\big[f(\eta, a(\xi))\big]. \end{aligned}$$

Dove la seconda uguaglianza segue, poichè in particolare $\pi \in \Pi(\mathbb{P}, \mathbb{Q})$, mentre l'ultima disuguaglianza segue per laconvessità di f applicando la disuguagliaza di Jensen. Troviamo quindi $v(\mathbb{Q}) \leq \mathbb{E}^{\pi}\big[f(\eta, a(\xi))\big]$. Ora dal teorema fondamentale del calcolo e dal teorema di Fubini:

$$\begin{aligned} \mathbb{E}^{\pi}\big[f(\eta, a(\xi))\big] - \mathbb{E}^{\mathbb{P}}\big[f(\xi, a(\xi))\big] &= \mathbb{E}^{\pi}\big[f(\eta, a(\xi))\big] - \mathbb{E}^{\pi}\big[f(\xi, a(\xi))\big] \\ &= \mathbb{E}^{\pi}\big[f(\eta, a(\xi)) - f(\xi, a(\xi))\big] = \mathbb{E}^{\pi}\left[\int_0^1 \nabla_x f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) \bullet (\eta - \xi)\, d\lambda\right] \\ &= \mathbb{E}^{\pi}\left[\int_0^1 \sum_{t=1}^{T} \Big[\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big)(\eta_t - \xi_t)\Big]\, d\lambda\right] \\ &= \sum_{t=1}^{T} \int_0^1 \mathbb{E}^{\pi}\Big[\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big)(\eta_t - \xi_t)\Big]\, d\lambda. \end{aligned}$$

Adesso usando la tower-property del valore atteso condizionato, la misurabilità di $\eta_t - \xi_t$ rispetto a $\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t$ e la disuguaglianza di Hölder otteniamo:

$$\begin{aligned} \mathbb{E}^{\pi}\big[f(\eta, a(\xi))\big] - \mathbb{E}^{\mathbb{P}}\big[f(\xi, a(\xi))\big] &= \sum_{t=1}^{T} \int_0^1 \mathbb{E}^{\pi}\Big[\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big)(\eta_t - \xi_t)\Big]\, d\lambda \\ &= \sum_{t=1}^{T} \int_0^1 \mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}\Big[\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\Big](\eta_t - \xi_t)\right] d\lambda \\ &\leq \sum_{t=1}^{T} \int_0^1 \mathbb{E}^{\pi}\left[\Big|\mathbb{E}^{\pi}\Big[\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\Big]\Big|^q\right]^{1/q} \cdot \mathbb{E}^{\pi}\big[|\eta_t - \xi_t|^p\big]^{1/p} d\lambda. \end{aligned}$$

Ora per ogni $1 \leq t \leq T$ poniamo

$$F_t := \mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a(\xi)) \mid \mathcal{F}_t\big],$$

e mostriamo che per ogni $\lambda \in [0,1]$ vale:

$$\mathbb{E}^{\pi}\left[\left|\mathbb{E}^{\pi}\Big[\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\Big]\right|^q\right]^{1/q} \xrightarrow[r\to 0]{} \mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big]^{1/q}.$$

Infatti poichè $\pi \in \Pi_{\mathrm{BC}}(\mathbb{P}, \mathbb{Q})$ si ha:

$$F_t = \mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a(\xi)) \mid \mathcal{F}_t\big] = \mathbb{E}^{\pi}\big[\partial_{x_t} f(\xi, a(\xi)) \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\big],$$

e anche:

$$\mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big]^{1/q} = \mathbb{E}^{\pi}\big[|F_t|^q\big]^{1/q}$$

ma allora usando la disuguaglianza triangolare e la disuguaglianza di Jensen condizionata abbiamo:

$$\begin{aligned}
&\left|\mathbb{E}^{\pi}\left[\left|\mathbb{E}^{\pi}\Big[\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\Big]\right|^q\right]^{1/q} - \mathbb{E}^{\pi}\big[|F_t|^q\big]^{1/q}\right|^q \\
&\leq \mathbb{E}^{\pi}\left[\left|\mathbb{E}^{\pi}\Big[\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\Big] - F_t\right|^q\right] \\
&= \mathbb{E}^{\pi}\left[\left|\mathbb{E}^{\pi}\Big[\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) - \partial_{x_t} f\big(\xi, a(\xi)\big) \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\Big]\right|^q\right] \\
&\leq \mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}\Big[\left|\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) - \partial_{x_t} f\big(\xi, a(\xi)\big)\right|^q \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\Big]\right] \\
&= \mathbb{E}^{\pi}\Big[\left|\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) - \partial_{x_t} f\big(\xi, a(\xi)\big)\right|^q\Big]
\end{aligned}$$

che converge a 0 per $r \to 0$. Questo segue dal fatto che, se $r \to 0$ allora anche $\sum_{t=1}^{T} \mathbb{E}^{\pi}\big[|\eta_t - \xi_t|^p\big] \to 0$, ossia

$$\eta \xrightarrow[r\to 0]{L^1(\pi)} \xi,$$

che implica $\eta \to \xi$ in π-probabilità. Poichè le funzioni continue preservano la convergenza in probabilità otteniamo

$$\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) \xrightarrow[r\to 0]{} \partial_{x_t} f\big(\xi, a(\xi)\big) \quad \text{in } \pi\text{-probabilità},$$

per ogni $\lambda \in (0,1)$. Ed ora dalla assunzione fatta su $\nabla_x f$ otteniamo

$$|\partial_{x_t} f\big(\xi, a(\xi)\big)|^q \leq C \in L^1(\pi).$$

Quindi dal teorema di convergenza dominata

$$\mathbb{E}^{\pi}\Big[\left|\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) - \partial_{x_t} f\big(\xi, a(\xi)\big)\right|^q\Big] \xrightarrow[r\to 0]{} 0,$$

come voluto. Dalla disuguaglianza traingolare, e dalle scelte per a e π segue allora che, per $r \to 0$:

$$\begin{aligned}
v(\mathbb{Q}) - v(\mathbb{P}) &\leq \mathbb{E}^{\pi}\big[f(\eta, a(\xi))\big] - \mathbb{E}^{\pi}\big[f(\xi, a(\xi))\big] + \varepsilon \\
&\leq \sum_{t=1}^{T} \int_0^1 \mathbb{E}^{\pi}\left[\left|\mathbb{E}^{\pi}\left[\partial_{x_t} f\big(\xi + \lambda(\eta - \xi), a(\xi)\big) \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right]\right|^q\right]^{1/q} \cdot \mathbb{E}^{\pi}\big[|\eta_t - \xi_t|^p\big]^{1/p} d\lambda + \varepsilon \\
&\leq \sum_{t=1}^{T} \Big(\mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big]^{1/q} + o(1)\Big)\mathbb{E}^{\pi}\big[|\eta_t - \xi_t|^p\big]^{1/p} + \varepsilon \\
&\leq \Big(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big] + o(1)\Big)^{1/q} \Big(\sum_{t=1}^{T} \mathbb{E}^{\pi}\big[|\eta_t - \xi_t|^p\big]\Big)^{1/p} + \varepsilon \\
&\leq \Big(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big]\Big)^{1/q} \mathcal{AW}_p(\mathbb{P}, \mathbb{Q}) + o(r) + \varepsilon \\
&\leq r\Big(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big]\Big)^{1/q} + o(r) + \varepsilon.
\end{aligned}$$

Questo dimostra, per arbitrarietà di $\varepsilon > 0$, che

$$v(\mathbb{Q}) - v(\mathbb{P}) \leq O(r),$$

invertendo i ruoli di $\mathbb{P}$ e $\mathbb{Q}$ otteniamo con lo stesso ragionamento

$$|v(\mathbb{Q}) - v(\mathbb{P})| \leq O(r)$$

da cui segue subito la tesi del lemma. □

Lemma 3.4. *Nel contesto e con le assunzioni precedenti esiste unico $a^* \in \mathcal{A}$ tale che $v(\mathbb{P}) = \mathbb{E}^{\mathbb{P}}\big[f(\xi, a^*(\xi))\big]$.*

Questo è un risultato classico, l'esistenza segue dal Lemma di Komlos enunciato in [7] e l'unicità segue dalla stretta convessità.

Dimostrazione teorema 3.1. Sia $a^* \in \mathcal{A}$ l'unico ottimizzante per $v(\mathbb{P})$, e per semplicità di notazione, definiamo per ogni $t = 1, \ldots, T$

$$F_t := \mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big].$$

Proviamo prima l'*upper bound*, ovvero

$$\sup_{\mathbb{Q} \in B_r(\mathbb{P})} v(\mathbb{Q}) - v(\mathbb{P}) \leq r\left(\sum_{t=1}^{T} \mathbb{E}^{\pi}\big[|F_t|^q\big]\right)^{1/q} + o(r).$$

A tale scopo definiamo $\mathbb{Q}^r \in B_r(\mathbb{P})$ tale che

$$v(\mathbb{Q}^r) \geq \sup_{\mathbb{Q} \in B_r(\mathbb{P})} v(\mathbb{Q}) - o(r)$$

e sia $\pi \in \Pi_{\mathrm{BC}}(\mathbb{P}, \mathbb{Q}^r)$ tale che

$$\left(\sum_{t=1}^{T} \mathbb{E}^{\pi}\big[|\eta_t - \xi_t|^p\big]\right)^{1/p} \leq \mathcal{AW}_p(\mathbb{P}, \mathbb{Q}^r) + o(r) \leq r + o(r).$$

Analogamente a quanto fatto per la precedente dimostrazione definiamo $b^r \in \mathcal{A}$ come

$$b^r = \mathbb{E}^{\pi}\big[a^*(\xi)\big].$$

Allora usando la disuguaglianza di Jensen, valida grazie alla convessità di f, otteniamo

$$\begin{aligned} v(\mathbb{Q}^r) &\leq \mathbb{E}^{\mathbb{Q}^r}\big[f(\eta, b^r(\eta))\big] = \mathbb{E}^{\pi}\big[f(\eta, b^r(\eta))\big] \\ &\leq \mathbb{E}^{\pi}\big[f(\eta, a^*(\xi))\big]. \end{aligned}$$

Ragionando allora allo stesso modo della dimostrazione di 3.3 otteniamo dalla disuguaglianza sopra

$$v(\mathbb{Q}^r) - v(\mathbb{P}) \leq r\left(\sum_{t=1}^{T} \mathbb{E}^{\pi}\big[|F_t|^q\big]\right)^{1/q} + o(r).$$

Ricordando ora la scelta di $\mathbb{Q}^r$ otteniamo esattamente

$$\sup_{\mathbb{Q} \in B_r(\mathbb{P})} v(\mathbb{Q}) - v(\mathbb{P}) \leq v(\mathbb{Q}^r) - v(\mathbb{P}) + o(r) \leq r\left(\sum_{t=1}^{T} \mathbb{E}^{\pi}\big[|F_t|^q\big]\right)^{1/q} + o(r).$$

Questo termina la dimostrazione della prima disuguaglianza. Passiamo alla dimostrazione del *lower bound*, che possiamo scrivere

$$\liminf_{r \to 0} \frac{1}{r}\Big(\sup_{\mathbb{Q} \in B_r(\mathbb{P})} v(\mathbb{Q}) - v(\mathbb{P})\Big) \geq \Big(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big]\Big)^{1/q}$$

A tale scopo osserviamo che usando la dualità tra $\ell^q(\mathbb{R}^T)$ e $\ell^p(\mathbb{R}^T)$ abbiamo che esiste $a \in [0 + \infty)^T$ tale che

$$\left(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big]\right)^{1/q} = \sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big] a_t \quad \text{e} \quad \sum_{t=1}^{T} a_t^p = 1.$$

Infatti per ogni $x \in \ell^q(\mathbb{R}^T) \setminus \{0\}$ si ha che esiste $y \in \big(\ell^q(\mathbb{R}^T)\big)^*$ tale che $y(x) = \|x\|_q$ e $\|y\|_{(\ell^q)^*} = 1$. Ma ora usando il teorema di rappresentazione di Riesz abbiamo che esiste unico $a \in \ell^p(\mathbb{R}^T)$ tale che

$$y(x) = \sum_{t=1}^{T} x_t a_t \quad \text{e} \quad \|a\|_p^p = \sum_{t=1}^{T} a_t^p = 1.$$

Per ottenere l'affermazione sopra è sufficiente applicare il precedente risultato al vettore $x \in \mathbb{R}^T$ definito da $x_t = \mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big]^{1/q}$ per ogni $t = 1, \dots, T$. In modo del tutto analogo, applicando il risultato visto sopra ma agli spazi $L^q(\mathbb{P}), L^p(\mathbb{P})$ e alle funzioni $a_t F_t$ abbiamo che esistono delle v.a. $(\zeta_t)_{t=1,\dots,T}$ tali che

$$\mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big]^{1/q} a_t = \mathbb{E}^{\mathbb{P}}\big[F_t \zeta_t\big] \quad \text{e} \quad \mathbb{E}^{\mathbb{P}}\big[|\zeta_t|^p\big]^{1/p} = a_t,$$

per ogni $t = 1, \dots, T$. Combinando i due risultati otteniamo che esistono v.a. $(\zeta_t)_{t=1,\dots,T}$ tali che

$$\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[F_t \zeta_t\big] = \Big(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[|F_t|^q\big]\Big)^{1/q} \quad \text{e} \quad \sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[|\zeta_t|^p\big] = 1.$$

Notiamo che siccome F_t è $\mathcal{F}_t$-misurabile, lavorando sullo spazio di probabilità $(\mathbb{R}^T, \mathcal{F}_t, \mathbb{P})$ la v.a. ζ_t può essere scelta a sua volta $\mathcal{F}_t$-misurabile.

Ora fissato $r > 0$ definiamo $\mathbb{P}^r$ come la legge della v.a. $\xi + r\zeta$, ovvero $\mathbb{P}^r := (\xi + r\zeta)_* \mathbb{P}$ e π^r come la legge della v.a. congiunta $(\xi, \xi + r\zeta)$ cioè $\pi^r := (\xi, \xi + r\zeta)_* \mathbb{P}$. Sappiamo allora siccome ζ_t è $\mathcal{F}_t$-misurabile che certamente $\pi^r \in \Pi_{\mathrm{C}}(\mathbb{P}, \mathbb{P}^r)$, ma in generale non è bicausale. Usiamo allora il lemma 3.2 con $\delta = r^2$. Otteniamo che esistono dei processi $\eta^r \colon \mathbb{R}^T \times \mathbb{R}^T$ tali che per ogni $t = 1, \dots, T$

- η_t^r è $\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t$-misurabile;
- ξ_t è $\sigma(\eta_t^r)$-misurabile;
- $|\eta_t^r - \eta_t| \leq r^2$.

In particolare $\gamma^r := (\xi, \eta^r)_* \pi^r$ è un coupling bicausale tra $\mathbb{P}$ e $\mathbb{Q}^r := \eta_*^r \pi^r$. Dunque

$$\begin{aligned}
\mathcal{AW}_p(\mathbb{P}, \mathbb{Q}^r) &\leq \Big(\sum_{t=1}^{T} \mathbb{E}^{\gamma^r}\big[|\xi_t - \eta_t|^p\big]\Big)^{1/p} \\
&= \Big(\sum_{t=1}^{T} \mathbb{E}^{\pi^r}\big[|\xi_t - \eta_t^r|^p\big]\Big)^{1/p} \\
&\leq \Big(\sum_{t=1}^{T} \mathbb{E}^{\pi^r}\big[|\xi_t - \eta_t|^p\big]\Big)^{1/p} + \Big(\sum_{t=1}^{T} \mathbb{E}^{\pi^r}\big[|\eta_t - \eta_t^r|^p\big]\Big)^{1/p} \\
&= \Big(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[|r\zeta_t|^p\big]\Big)^{1/p} + Tr^2 = r + Tr^2.
\end{aligned} \tag{3.1}$$

Dai calcoli precedenti otteniamo che, per ogni $\varepsilon > 0$, per r sufficientemente piccolo abbiamo $\mathbb{Q}^r \in B_{r+\varepsilon}(\mathbb{P})$, da cui

$$\sup_{\mathbb{L} \in B_{r+\varepsilon}(\mathbb{P})} v(\mathbb{L}) \geq v(\mathbb{Q}^r). \tag{3.2}$$

Ora per ogni $r > 0$ sia $a^r \in \mathcal{A}$ un controllo quasi ottimale per $v(\mathbb{Q}^r)$, ossia

$$\begin{aligned} v(\mathbb{Q}^r) &\geq \mathbb{E}^{\mathbb{Q}^r}\big[f(\eta, a^r(\eta))\big] - o(r) \\ &= \mathbb{E}^{\pi^r}\big[f(\eta^r, a^r(\eta^r))\big] - o(r). \end{aligned}$$

Adesso siccome η_t^r è $\mathcal{F}_t \otimes \tilde{\mathcal{F}}_t$-misurabile e ζ_t è $\mathcal{F}_t$-misurabile otteniamo che η^r è funzione adattata della sola ξ, dunque esiste $b^r \in \mathcal{A}$ tale che

$$\mathbb{P}\big(a^r(\eta^r) = b^r(\xi)\big) = 1$$

e con un abuso di notazione

$$\mathbb{E}^{\pi^r}\big[f(\eta^r, a^r(\eta^r))\big] = \mathbb{E}^{\mathbb{P}}\big[f(\eta^r, a^r(\eta^r))\big].$$

Mettendo insieme le due cose ricaviamo

$$v(\mathbb{Q}^r) \geq \mathbb{E}^{\mathbb{P}}\big[f(\eta^r, b^r(\xi))\big] - o(r).$$

Inoltre poichè $b^r \in \mathcal{A}$ si ha

$$v(\mathbb{P}) \leq \mathbb{E}^{\mathbb{P}}\big[f(\xi, b^r(\xi))\big],$$

disuguaglianza dalla quale, usando il teorema fondamentale del calcolo e il teorema di Fubini otteniamo

$$\begin{aligned} v(\mathbb{Q}^r) - v(\mathbb{P}) &\geq \mathbb{E}^{\mathbb{P}}\big[f(\eta^r, b^r(\xi)) - f(\xi, b^r(\xi))\big] - o(r) \\ &= \sum_{t=1}^{T} \int_0^1 \mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi + \lambda(\eta^r - \xi), b^r(\xi))(\eta_t^r - \xi_t)\big]\, d\lambda - o(r) \\ &= \sum_{t=1}^{T} \int_0^1 \mathbb{E}^{\mathbb{P}}\Big[\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi + \lambda(\eta^r - \xi), b^r(\xi)) \mid \mathcal{F}_t\big](\eta_t^r - \xi_t)\Big]\, d\lambda - o(r). \end{aligned}$$

Da cui segue

$$\frac{v(\mathbb{Q}^r) - v(\mathbb{P})}{r} \geq \sum_{t=1}^{T} \int_0^1 \mathbb{E}^{\mathbb{P}}\Big[\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi + \lambda(\eta^r - \xi), b^r(\xi)) \mid \mathcal{F}_t\big]\frac{\eta_t^r - \xi_t}{r}\Big]\, d\lambda - o(1). \tag{3.3}$$

Sia ora $(r_n)_{n\in\mathbb{N}}$ una arbitraria successione che converge a zero. Mostriamo che

$$b^{r_n}(\xi) \xrightarrow[n\to\infty]{} a^*(\xi) \quad \text{in } \mathbb{P}\text{-probabilità.}$$

Per farlo ricordiamo che b^{r_n} è stato scelto quasi ottimale per $\mathbb{Q}^{r_n}$, ossia

$$v(\mathbb{Q}^{r_n}) \geq \mathbb{E}^{\mathbb{P}}\big[f(\eta^{r_n}, b^{r_n}(\xi))\big] - o(r_n) \geq \mathbb{E}^{\mathbb{P}}\big[f(\xi, b^{r_n}(\xi))\big] - O(r_n),$$

ove l'ultima disuguaglianza segue per il teorema fondamentale del calcolo e dall'ipotesi fatta su $\nabla_x f$. Infatti da uno sviluppo al primo ordine abbiamo

$$\begin{aligned}
&\mathbb{E}^{\mathbb{P}}\big[f(\eta^r, b^r(\xi))\big] = \mathbb{E}^{\mathbb{P}}\big[f(\xi, b^r(\xi))\big] + \mathbb{E}^{\mathbb{P}}\bigg[\int_0^1 \langle \nabla_x f\big(\xi + \theta(\eta^r - \xi), b^r(\xi)\big), \eta^r - \xi\rangle \, d\theta\bigg] \\
&= \mathbb{E}^{\mathbb{P}}\big[f(\xi, b^r(\xi))\big] + \mathbb{E}^{\mathbb{P}}\bigg[\int_0^1 \sum_{t=1}^T \partial_{x_t} f\big(\xi + \theta(\eta^r - \xi), b^r(\xi)\big)(\eta_t^r - \xi_t) \, d\theta\bigg] \\
&\geq \mathbb{E}^{\mathbb{P}}\big[f(\xi, b^r(\xi))\big] - \mathbb{E}^{\mathbb{P}}\bigg[\int_0^1 \Big(\sum_{t=1}^T |\partial_{x_t} f\big(\xi + \theta(\eta^r - \xi), b^r(\xi)\big)|^q\Big)^{1/q} \Big(\sum_{t=1}^T |\eta_t^r - \xi_t|^p\Big)^{1/p} d\theta\bigg] \\
&= \mathbb{E}^{\mathbb{P}}\big[f(\xi, b^r(\xi))\big] - \mathbb{E}^{\mathbb{P}}\bigg[\Big(\sum_{t=1}^T |\eta_t^r - \xi_t|^p\Big)^{1/p} \int_0^1 \big\|\nabla_x f\big(\xi + \theta(\eta^r - \xi), b^r(\xi)\big)\big\|_q \, d\theta\bigg] \\
&\geq \mathbb{E}^{\mathbb{P}}\big[f(\xi, b^r(\xi))\big] - C^{1/q}\Big(\sum_{t=1}^T \mathbb{E}^{\mathbb{P}}\big[|\eta_t^r - \xi_t|^p\big]\Big)^{1/p} \\
&= \mathbb{E}^{\mathbb{P}}\big[f(\xi, b^r(\xi))\big] - O(r).
\end{aligned}$$

Allora facendo un'espansione in serie di Taylor al primo ordine, con resto in forma di Lagrange otteniamo

$$\begin{aligned}
v(\mathbb{Q}^{r_n}) - v(\mathbb{P}) &\geq \mathbb{E}^{\mathbb{P}}\big[f(\xi, b^{r_n}(\xi)) - f(\xi, a^*(\xi))\big] - O(r_n) \\
&= \mathbb{E}^{\mathbb{P}}\Big[\langle \nabla_a f(\xi, a^*(\xi)), b^{r_n}(\xi) - a^*(\xi)\rangle\Big] \\
&\quad + \mathbb{E}^{\mathbb{P}}\Big[\frac{1}{2}\int_0^1 \langle b^{r_n}(\xi) - a^*(\xi), \nabla_a^2 f\big(\xi, b^{r_n}(\xi) + \theta(a^*(\xi) - b^{r_n}(\xi))\big)\big(b^{r_n}(\xi) - a^*(\xi)\big)\rangle \, d\theta\Big] - O(r_n).
\end{aligned}$$

Ora però ricordiamo che per ipotesi $\nabla_a^2 f(\xi, a) \succ \varepsilon(\xi)\mathbf{1}_{T\times T}$ per $a \in [-L, L]^T$ e con $\mathbb{P}(\varepsilon(\xi) > 0) = 1$. Dunque

$$\begin{aligned}
v(\mathbb{Q}^{r_n}) - v(\mathbb{P}) &\geq \mathbb{E}^{\mathbb{P}}\Big[\langle \nabla_a f(\xi, a^*(\xi)), b^{r_n}(\xi) - a^*(\xi)\rangle\Big] \\
&\quad + \mathbb{E}^{\mathbb{P}}\Big[\frac{\varepsilon(\xi)}{2}\|b^{r_n}(\xi) - a^*(\xi)\|_{\ell^2}^2\Big] - O(r_n),
\end{aligned}$$

da cui deduciamo che per $n \to \infty$ deve essere

$$\mathbb{E}^{\mathbb{P}}\Big[\langle \nabla_a f(\xi, a^*(\xi)), b^{r_n}(\xi) - a^*(\xi)\rangle\Big] + \mathbb{E}^{\mathbb{P}}\Big[\frac{\varepsilon(\xi)}{2}\|b^{r_n}(\xi) - a^*(\xi)\|_{\ell^2}^2\Big] \to 0$$

poichè per $n \to \infty$ si ha $\mathcal{AW}_p(\mathbb{P}, \mathbb{Q}^{r_n}) \to 0$ e dunque dal lemma 3.3 abbiamo $v(\mathbb{Q}^{r_n}) \to v(\mathbb{P})$. Adesso osservando che nella somma sopra il primo addendo è sempre non negativo (perchè a^* è ottimo) otteniamo che deve essere

$$\mathbb{E}^{\mathbb{P}}\Big[\frac{\varepsilon(\xi)}{2}\|b^{r_n}(\xi) - a^*(\xi)\|_{\ell^2}^2\Big] \xrightarrow[n\to\infty]{} 0,$$

che è possibile solamente se $b^{r_n}(\xi) \to a^*(\xi)$ in $\mathbb{P}$-probabilità, poichè $\varepsilon(\xi)$ è $\mathbb{P}$-q.c. positivo. Per arrivare alla conclusione è ora necessario fare le seguenti osservazioni.

1. Per $n \to \infty$ abbiamo per ogni $t = 1, \dots, T$ $\eta_t^{r_n} \to \xi_t$ in $L^1(\mathbb{P})$, infatti

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}\big[|\eta_t^{r_n} - \xi_t|\big] &\leq \mathbb{E}^{\mathbb{P}}\big[|\eta_t^{r_n} - \xi_t - r_n\zeta_t|\big] + \mathbb{E}^{\mathbb{P}}\big[|\xi_t + r_n\zeta_t - \xi_t|\big] \\ &\leq r_n^2 + r_n\mathbb{E}^{\mathbb{P}}\big[|\zeta_t|\big] \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

2. Similmente per ogni $t = 1, \dots, T$ per $n \to \infty$ abbiamo anche

$$\frac{\eta_t^{r_n} - \xi_t}{r_n} \to \zeta_t \quad \text{in } L^1(\mathbb{P})$$

infatti

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}\left[\left|\frac{\eta_t^{r_n} - \xi_t}{r_n} - \zeta_t\right|\right] &= \mathbb{E}^{\mathbb{P}}\left[\left|\frac{\eta_t^{r_n} - \xi_t - r_n\zeta_t}{r_n} + \frac{r_n\zeta_t}{r_n} - \zeta_t\right|\right] \\ &= \mathbb{E}^{\mathbb{P}}\left[\left|\frac{\eta_t^{r_n} - \xi_t - r_n\zeta_t}{r_n}\right|\right] \leq r_n \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

Dalla 1 otteniamo che per ogni $t = 1, \dots, T$ $\eta_t^{r_n} \xrightarrow[n\to\infty]{} \xi_t$ in $\mathbb{P}$-probabilità, quindi combinandola con la convergenza di b^{r_n} abbiamo per ogni $\lambda \in (0,1)$, siccome $\partial_{x_t} f$ è continua

$$\partial_{x_t} f(\xi + \lambda(\eta^{r_n} - \xi), b^{r_n}(\xi)) \xrightarrow[n\to\infty]{} \partial_{x_t} f(\xi, a^*(\xi)) \quad \text{in } \mathbb{P}\text{-probabilità.}$$

Il che implica che esiste una sottosuccessione $(r_{n_k})_{k\in\mathbb{N}}$ tale che

$$\partial_{x_t} f(\xi + \lambda(\eta^{r_{n_k}} - \xi), b^{r_{n_k}}(\xi)) \xrightarrow[k\to\infty]{} \partial_{x_t} f(\xi, a^*(\xi)) \quad \mathbb{P}\text{-q.c.}$$

ed ora dalla assunzione su $\nabla_x f$ la successione ha una dominante L^1 e quindi possiamo usare il teorema di convergenza dominata per il valore atteso condizionato, in particolare otteniamo

$$\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi + \lambda(\eta^{r_{n_k}} - \xi), b^{r_{n_k}}(\xi)) \mid \mathcal{F}_t\big] \xrightarrow[k\to\infty]{} \mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big] \quad \mathbb{P}\text{-q.c.}$$

però da 2 a meno di passare a sottosuccessioni otteniamo

$$\frac{\eta_t^{r_{n_k}} - \xi_t}{r_{n_k}} \to \zeta_t \quad \mathbb{P}\text{-q.c.}$$

e quindi

$$\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi+\lambda(\eta^{r_{n_k}}-\xi), b^{r_{n_k}}(\xi)) \mid \mathcal{F}_t\big]\frac{\eta_t^{r_{n_k}} - \xi_t}{r_{n_k}} \xrightarrow[k\to\infty]{} \mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big]\zeta_t \quad \mathbb{P}\text{-q.c.}$$

da cui applicando nuovamente il teorema di convergenza dominata otteniamo

$$\mathbb{E}^{\mathbb{P}}\Big[\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi+\lambda(\eta^{r_{n_k}}-\xi), b^{r_{n_k}}(\xi)) \mid \mathcal{F}_t\big]\frac{\eta_t^{r_{n_k}} - \xi_t}{r_{n_k}}\Big] \xrightarrow[k\to\infty]{} \mathbb{E}^{\mathbb{P}}\Big[\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big]\zeta_t\Big].$$

Dunque combinando la convergenza precedente con la (3.3) otteniamo

$$\liminf_{k\to\infty} \frac{v(\mathbb{Q}^{r_{n_k}}) - v(\mathbb{P})}{r_{n_k}} \geq \sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\Big[\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big]\zeta_t\Big],$$

da cui per arbitrarietà della successione $(r_n)_{n\in\mathbb{N}}$ otteniamo

$$\liminf_{r\to 0} \frac{v(\mathbb{Q}^{r}) - v(\mathbb{P})}{r} \geq \sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\Big[\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big]\zeta_t\Big].$$

Ora usando la (3.2) con $\varepsilon = o(r)$ otteniamo

$$\begin{aligned}\liminf_{r\to 0} \frac{1}{r}\Big(\sup_{\mathbb{Q}\in B_r(\mathbb{P})} v(\mathbb{Q}) - v(\mathbb{P})\Big) &\geq \sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\Big[\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big]\zeta_t\Big] \\ &= \left(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\big[|\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big]|^q\big]\right)^{1/q}\end{aligned}$$

e si conclude, l'ultima uguaglianza segue dalla definizione delle v.a. $(\zeta_t)_{t=1,\dots,T}$. □

3.3 Il caso di massimizzazione dell'utilità attesa

Abbiamo già osservato che il problema di *massimizzazione dell'utilità attesa* può essere scritto nella stessa forma dei problemi di ottimizzazione stocastica considerati in questo capitolo. In questa sezione ne ricordiamo il contesto e cerchiamo di specializzare il risultato 3.1 ottenuto nella sezione precedente a questo caso. Ricordiamo che supponiamo di conescere una funzione $\ell\colon \mathbb{R} \to \mathbb{R}$ convessa e limitata dal basso, e una funzione $g\colon \mathbb{R}^T \to \mathbb{R}$ (la negativa di una funzione di payoff) e consideriamo il problema

$$u(\mathbb{P}) := \inf_{a\in\mathcal{A}} \mathbb{E}^{\mathbb{P}}\Big[\ell\Big(g(\xi) + \sum_{t=1}^{T} a_t(\xi)(\xi_t - \xi_{t-1})\Big)\Big],$$

ove $\xi_0 \in \mathbb{R}$ è un valore fissato. Supponiamo inoltre che ℓ sia due volte differenziabile con continuità con $\ell'(v) \leq C$, con $C > 0$ una costante, $\ell'' > 0$, che g sia differenziabile con continuità con derivate uniformemente limitate e infine assumiamo che $\mathbb{P}(\xi_{t-1} = \xi_t) = 0$ per ogni $t = 1, \dots, T$.

Corollario 3.5. *Nel contesto precedente, sia a^* l'unico ottimizzante per $u(\mathbb{P})$. Si ponga $a_{T+1} = 0$,*

$$\zeta := g(\xi) + \sum_{t=1}^{T} a^*(\xi)(\xi_t - \xi_{t-1}),$$

e

$$V := \left(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}} \Big[\big| (a^*_{t+1}(\xi) - a^*_t(\xi)) \mathbb{E}^{\mathbb{P}} \big[\ell'(\zeta) \mid \mathcal{F}_t \big] - \mathbb{E}^{\mathbb{P}} \big[\ell'(\zeta) \partial_{x_t} g(\xi) \mid \mathcal{F}_t \big] \big|^q \Big] \right)^{1/q}.$$

Allora per $r \to 0$*, si ha*

$$\sup_{\mathbb{Q} \in B_r(\mathbb{P})} u(\mathbb{Q}) = u(\mathbb{P}) + r \cdot V + o(r). \tag{3.4}$$

Osservazione 3.1. Si noti che nel caso in cui $p = q = 2$ e $g = 0$ la quantità V è la variazione quadratica di a^* distorta dal valore atteso condizionato di ℓ'.

Dimostrazione. Per brevità di notazione scriviamo, per ogni $a, x \in \mathbb{R}^T$

$$(a \cdot x)_T := \sum_{t=1}^{T} a_t (x_t - x_{t-1}).$$

L'obiettivo è di applicare il teorema 3.1 alla funzione

$$f(x, a) := \ell\big(g(x) + (a \cdot x)_T\big)$$

per $(x, a) \in \mathbb{R}^T$. Per farlo dobbiamo controllare che le ipotesi del teorema 3.1 sono verificate. Partiamo con l'osservare che grazie alle ipotesi di differenziabilità fatte su ℓ e g otteniamo che f verifica le proprietà volute. Infatti siccome ℓ è due volte differenziabile con continuità e $(a \cdot x)_T$ è $\mathcal{C}^\infty$ in a, abbiamo banalmente che $f(x, \cdot)$ è due volte differenziabile con continuità. Inoltre per gli stessi motivi e poichè anche g è differenziabile con continuità allora $f(\cdot, a)$ lo è per ogni $a \in \mathbb{R}^T$. In particolare per ogni $t, s = 1, \dots, T$ si ha

$$\partial_{a_t} f(x, a) = \ell'\big(g(x) + (a \cdot x)_T\big)(x_t - x_{t-1})$$

e

$$\begin{aligned} \partial_{a_s} \partial_{a_t} f(x, a) &= \partial_{a_s} \ell'\big(g(x) + (a \cdot x)_T\big)(x_t - x_{t-1}) \\ &= \ell''\big(g(x) + (a \cdot x)_T\big)(x_t - x_{t-1})(x_s - x_{s-1}). \end{aligned}$$

Quindi per ogni $u \in \mathbb{R}^T \setminus \{0\}$ si ha

$$\begin{aligned} \langle u, \nabla_a^2 f(x, a) u \rangle &= \ell''\big(g(x) + (a \cdot x)_T\big) \Bigg[\sum_{t=1}^{T} u_t^2 (x_t - x_{t-1})^2 \\ &\quad + 2 \sum_{1 \le t < s \le T} u_t u_s (x_t - x_{t-1})(x_s - x_{s-1}) \Bigg] \\ &= \ell''\big(g(x) + (a \cdot x)_T\big) \Big[\sum_{t=1}^{T} u_t (x_t - x_{t-1}) \Big]^2. \end{aligned}$$

Dai precedenti calcoli otteniamo che in questo caso non è verificata l'ipotesi su $\nabla_a^2 f(\xi, a)$, tuttavia una comunicazione personale del primo degli autori di [3] assicura che mediante una diversa dimostrazione, usando che, per convessità di f, abbiamo

$$\langle u, \nabla_a^2 f(x,a)u\rangle \geq 0 \quad \forall u \in \mathbb{R}^T \setminus \{0\}.$$

e $\mathbb{P}(\xi_t = \xi_{t-1}) = 0$, si riesce anche in questo caso a provare che $b^r(\xi) \to a^*(\xi)$ in $\mathbb{P}$-probabilità per $r \to 0$ (si veda la dimostrazione del teorema 3.1). Infatti ragionando come nella dimostrazione del teorema, presa una arbitraria successione $(r_n)_{n\in\mathbb{N}}$ infinitesima, da uno sviluppo in serie di Taylor al primo ordine con resto in forma di Lagrange si ottiene

$$\mathbb{E}^{\mathbb{P}}\Big[\int_0^1 \ell''\Big(g(\xi)+\big((a^*(\xi)+\lambda(b^{r_n}(\xi)-a^*(\xi))\cdot\xi\big)_T\Big)\Big[\sum_{t=1}^T \big(a_t^*(\xi)-b_t^{r_n}(\xi)\big)(\xi_t-\xi_{t-1})\Big]^2\Big] \xrightarrow[n\to\infty]{} 0.$$

Da ciò segue che

$$\mathbb{E}^{\mathbb{P}}\Big[\Big(\sum_{t=1}^T \big(a_t^*(\xi) - b_t^{r_n}(\xi)\big)(\xi_t - \xi_{t-1})\Big)^2\Big] \xrightarrow[n\to\infty]{} 0,$$

ed ora è sufficiente utilizzare l'isometria di Ito a tempo discreto, essa infatti afferma che se $(Y_t)_{t=0,1,\dots,T}$ è una martingala e $(H_t)_{t=1,\dots,T}$ è un processo predicibile, allora

$$\mathbb{E}^{\mathbb{P}}\Big[\Big(\sum_{t=1}^T H_t(Y_t - Y_{t-1})\Big)^2\Big] = \sum_{t=1}^T \mathbb{E}^{\mathbb{P}}\big[H_t^2(Y_t - Y_{t-1})^2\big].$$

Nel nostro caso ponendo per ogni $t = 0, \dots, T-1$, $a_t = \mathbb{E}^{\mathbb{P}}\big[(\xi_{t+1}-\xi_t)^2 \mid \mathcal{F}_t\big]$, quantità che per l'ipotesi fatta è $\mathbb{P}$-q.c. positiva, l'isometria di Ito e la tower property del valore atteso condizionato implicano

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}\Big[\Big(\sum_{t=1}^T \big(a_t^*(\xi) - b_t^{r_n}(\xi)\big)(\xi_t - \xi_{t-1})\Big)^2\Big] &= \sum_{t=1}^T \mathbb{E}^{\mathbb{P}}\Big[\big(a_t^*(\xi) - b_t^{r_n}(\xi)\big)^2(\xi_t - \xi_{t-1})^2\Big] \\ &= \sum_{t=1}^T \mathbb{E}^{\mathbb{P}}\Big[\big(a_t^*(\xi) - b_t^{r_n}(\xi)\big)^2 a_{t-1}\Big] \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

Da cui per la positività quasi certa di a_{t-1} otteniamo $b_t^{r_n}(\xi) \xrightarrow[n\to\infty]{} a_t^*(\xi)$ in $\mathbb{P}$-probabilità, per ogni $t = 1, \dots, T$ come voluto. Dunque se dimostriamo l'assunzione su $\nabla_x f(x,a)$ possiamo utilizzare il teorema. Per farlo osserviamo che

$$\partial_{x_t} f(x,a) = \ell'\big(g(x) + (a\cdot x)_T\big)(\partial_{x_t} g(x) + a_t - a_{t+1}),$$

da cui, usando le assunzioni fatte su ℓ' e $\partial_{x_t} g$ otteniamo che per ogni $x \in \mathbb{R}^T, a \in [-L, L]^T$ si ha:

$$|\partial_{x_t} f(x,a)| \leq C(\tilde{C} + 2L),$$

e dunque

$$\|\nabla_x f(x,a)\|_q^q \leq TC^q(\tilde{C}+2L)^q = \overline{C}.$$

Quindi possiamo applicare il teorema 3.1. Per concludere la dimostrazione dobbiamo solamente notare che nel nostro caso, per ogni $t = 1, \dots, T$ abbiamo

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big] &= \mathbb{E}^{\mathbb{P}}\big[\ell'(\zeta)(\partial_{x_t} g(\xi) + a_t^*(\xi) - a_{t+1}^*(\xi)) \mid \mathcal{F}_t\big] \\ &= \mathbb{E}^{\mathbb{P}}\big[a_t^*(\xi) - a_{t+1}^*(\xi)\ell'(\zeta) + \ell'(\zeta)\partial_{x_t} g(\xi) \mid \mathcal{F}_t\big] \\ &= (a_t^*(\xi) - a_{t+1}^*(\xi))\mathbb{E}^{\mathbb{P}}\big[\ell'(\zeta) \mid \mathcal{F}_t\big] + \mathbb{E}^{\mathbb{P}}\big[\ell'(\zeta)\partial_{x_t} g(\xi) \mid \mathcal{F}_t\big]\end{aligned}$$

Da cui

$$\begin{aligned}&\left(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\Big[\big|\mathbb{E}^{\mathbb{P}}\big[\partial_{x_t} f(\xi, a^*(\xi)) \mid \mathcal{F}_t\big]\big|^q\Big]\right)^{1/q} \\ &= \left(\sum_{t=1}^{T} \mathbb{E}^{\mathbb{P}}\Big[\big|(a_{t+1}^*(\xi) - a_t^*(\xi))\mathbb{E}^{\mathbb{P}}\big[\ell'(\zeta) \mid \mathcal{F}_t\big] - \mathbb{E}^{\mathbb{P}}\big[\ell'(\zeta)\partial_{x_t} g(\xi) \mid \mathcal{F}_t\big]\big|^q\Big]\right)^{1/q} = V,\end{aligned}$$

quindi l'enunciato del teorema 3.1 diviene esattamente

$$\sup_{\mathbb{Q}\in B_r(\mathbb{P})} u(\mathbb{Q}) = u(\mathbb{P}) + r \cdot V + o(r)$$

e si conclude. □

Appendice A

Risultati utili di Probabilità

A.1 Convergenza di misure e teorema di Prohorov

Definizione A.1. Uno spazio topologico (E, τ) si dice essere uno *spazio Polacco* se è *separabile* ed esiste una metrica completa d su E che induce la topologia τ.

Nel seguito sia (E, τ) uno spazio topologico completamente metrizzabile e pensato dotato della σ-algebra di Borel: $\mathcal{E} = \mathscr{B}(E)$. Per le misure definite su $(E, \mathcal{E})$ introduciamo le seguenti nozioni di regolarità.

Definizione A.2. Una misura σ-finita μ su $(E, \mathcal{E})$ si dice:

1. *misura di Borel* se, per ogni $x \in E$, esiste un intorno aperto U di x tale che $\mu(U) < \infty$;
2. *regolare dall'interno* se, per ogni $A \in \mathcal{E}$ si ha:
$$\mu(A) = \sup\{\mu(K) \mid K \subseteq A \quad \text{è compatto}\};$$
3. *regolare dall'esterno* se, per ogni $A \in \mathcal{E}$ si ha:
$$\mu(A) = \inf\{\mu(U) \mid A \subseteq U \quad \text{è aperto}\};$$
4. *regolare* se μ è regolare dall'interno e dall'esterno;
5. *misura di Radon* se è una misura di Borel regolare dall'interno.

Definizione A.3. Introduciamo i seguenti spazi di misure su E:

$$\begin{aligned} \mathcal{M}_+(E) &:= \{\mu \mid \mu \quad \text{è una misura di Radon su } (E, \mathcal{E})\}; \\ \mathcal{M}_f(E) &:= \{\mu \mid \mu \quad \text{à una misura } \sigma\text{-finita su } (E, \mathcal{E})\}; \\ \mathcal{M}_1(E) = \mathcal{P}(E) &:= \{\mu \in \mathcal{M}_f(E) \mid \mu(E) = 1\}. \end{aligned}$$

Fissiamo inoltre le seguenti notazioni per le funzioni continue su E:

$$\begin{aligned} \mathcal{C}(E) &:= \{f \colon E \to \mathbb{R} \mid f \quad \text{è continua}\}; \\ \mathcal{C}_b(E) &:= \{f \in \mathcal{C}(E) \mid f \quad \text{è limitata}\}; \\ \mathcal{C}_c(E) &:= \{f \in \mathcal{C}(E) \mid f \quad \text{è a supporto compatto}\} \subset \mathcal{C}_b(E). \end{aligned}$$

Proposizione A.1. *Se E è Polacco e $\mu \in \mathcal{M}_f(E)$, allora μ è* serrata*, ossia per ogni $\varepsilon > 0$, esiste un insieme compatto $K \subseteq E$ tale che: $\mu(E \setminus K) < \varepsilon$.*

Dimostrazione. Sia $\varepsilon > 0$, allora poichè E è separabile: $\forall n \in \mathbb{N}$, esiste una successione: $(x_j^n)_{j\in\mathbb{N}} \subseteq E$ tale che: $E = \bigcup_{j=1}^{\infty} B_{1/n}(x_j^n)$. Fissato allora $N_n \in \mathbb{N}$ tale che $\mu(E \setminus \bigcup_{j=1}^{N_n} B_{1/n}(x_j^n)) < \frac{\varepsilon}{2^n}$, definiamo:

$$A := \bigcap_{n=1}^{\infty} \bigcup_{j=1}^{N_n} B_{1/n}(x_j^n).$$

Allora per costruzione, A è totalmente limitato. Ma allora siccome E è Polacco, $\overline{A}$ è compatto, inoltre:

$$\mu(E \setminus \overline{A}) \leq \mu(E \setminus A) < \sum_{n=1}^{\infty} \frac{\varepsilon}{2^n} = \varepsilon. \qquad \square$$

Definizione A.4. Sia $\mathcal{F} \subset \mathcal{M}(E)$ una famiglia di misure di Radon su E. Una famiglia $\mathcal{C}$ di funzioni misurabili da E in $\mathbb{R}$ si dice *separante per* $\mathcal{F}$ se, $\forall \mu, \nu \in \mathcal{F}$ vale che:

$$\int_E f\, d\mu = \int_E f\, d\nu \qquad \forall f \in \mathcal{C} \cap L^1(\mu) \cap L^1(\nu)$$

implica: $\mu = \nu$.

Proposizione A.2. *Sia (E, d) uno spazio metrico. Allora $Lip_1(E)$ (funzioni Lipschitziane con costante di Lipschitz al più 1) è separante per $\mathcal{M}_+(E)$ e dunque anche per ogni suo sottoinsieme.*

Si veda [6] teorema 13.11.

Definizione A.5 (Convergenza debole di misure)**.** Sia E uno spazio metrico e $(\mu_n)_{n\in\mathbb{N}} \subseteq \mathcal{M}_f(E)$, diciamo che $(\mu_n)_{n\in\mathbb{N}}$ *converge debolmente* a μ, in simboli $\mu_n \xrightarrow[n\to\infty]{} \mu$ (debolmente), se:

$$\int f\, d\mu_n \xrightarrow{n\to\infty} \int f\, d\mu \qquad \forall f \in \mathcal{C}_b(E).$$

Osservazione A.1. Su $\mathcal{M}_f(E)$ definizamo la *topologia debole* τ_w come la più piccola topologia tale che, per ogni $f \in \mathcal{C}_b(E)$, la mappa:

$$\Phi_f \colon \mathcal{M}_f(E) \to \mathbb{R}, \mu \mapsto \int f\, d\mu$$

è continua. Con questa definizione risulta allora che $\mu_n \xrightarrow[n\to\infty]{\tau_w} \mu$ se e solo se: $\mu_n \xrightarrow[n\to\infty]{} \mu$ (debolmente). Bisogna tuttavia osservare che questa nozione di topologia e convergenza debole è diversa da quella dell'analisi funzionale, quella data qui

corrisponde infatti alla cosiddetta *topologia debole**. Dato uno spazio normato X, qui $X = \mathcal{C}_b(E)$ dotato della norma del sup: $\|\cdot\|_\infty$, si definisce la topologia debole* sul suo *duale topologico* X' dicendo che, data $(T_n)_{n\in\mathbb{N}} \subseteq X'$, essa *converge debole** a T, se e solo se: $T_n(x) \xrightarrow[n\to\infty]{} T(x)$ per ogni $x \in X$. È allora ovvio che ogni $\mu \in \mathcal{M}_f(E)$ definisce un funzionale lineare e continuo su $\mathcal{C}_b(E)$:

$$f \mapsto \langle \mu, f \rangle := \int f \, d\mu.$$

Quindi $\mathcal{M}_f(E) \subseteq (\mathcal{C}_b(E))'$, quindi la topologia τ_w definita sopra è la *topologia debole* indotta* da $(\mathcal{C}_b(E))'$ su $\mathcal{M}_f(E)$.

Definizione A.6. Sia (E, τ) uno spazio topologico. Una famiglia di misure $\mathcal{F} \subseteq \mathcal{M}(E)$ si dice *uniformemente serrata* se, per ogni $\varepsilon > 0$ esiste un insieme compatto $K_\varepsilon \subseteq E$ tale che

$$\mu(K_\varepsilon^c) < \varepsilon \quad \text{per ogni } \mu \in \mathcal{F}.$$

Teorema A.3 (Prohorov)**.** *Sia* (E, d) *uno spazio metrico e* $\mathcal{F} \subseteq \mathcal{M}_1(E)$. *Allora*

1. *se* $\mathcal{F}$ *è* uniformemente serrata *allora* $\mathcal{F}$ *è* relativamente debolmente sequenzialmente compatta*, ossia* $\overline{\mathcal{F}}$ *è debolmente sequenzialmente compatta;*
2. *se* (E, d) *è uno spazio Polacco allora anche il viceversa è vero, ossia se* $\mathcal{F}$ *è* relativamente debolmente sequenzialmente compatta*, allora* $\mathcal{F}$ *è* uniformemente serrata*.*

Per la dimostrazione del teorema si fa riferimento ad esempio a [6].

A.2 Nuclei Markoviani e Distribuzioni Condizionali

Definizione A.7 (Nuclei di Transizione e Nuclei Markoviani)**.** Siano $(S, \mathcal{S}), (T, \mathcal{T})$ due spazi misurabili, una mappa $\kappa\colon S \times \mathcal{T} \to [0, +\infty]$ si dice un *nucleo regolare di transizione* se vale che

1. per ogni $B \in \mathcal{T}$ la mappa $s \mapsto \kappa(s, B)$ è $\mathcal{S}$-misurabile,
2. per ogni $s \in S$ la mappa $B \mapsto \kappa(s, B)$ è una misura su $(T, \mathcal{T})$.

Se poi la misure $\kappa(s, \cdot)$ su $(T, \mathcal{T})$ è una misura di probabilità, allora il nucleo κ è detto *nucleo di probabilità* o *nucleo markoviano.*

Osservazione A.2. Se $\mathcal{T} = \sigma(\mathcal{C})$ con $\mathcal{C}$ un sistema $\cap$-stabile tale che contiene T, oppure una successione crescente di insiemi E_n che converge a T, allora per controllare se κ sia un nucleo di transizione, la proprietà 1 può essere verificata solo per gli insiemi $A \in \mathcal{C}$. Infatti, in questo caso,

$$\mathcal{D} := \{A \in \mathcal{T} \mid s \mapsto \kappa(s, A) \quad \text{è } \mathcal{S}\text{-misurabile}\}$$

è un sistema di Dynkin. Se quindi $\mathcal{C} \subseteq \mathcal{D}$ otteniamo che $\mathcal{D} = \sigma(\mathcal{C}) = \mathcal{T}$.

Le seguenti caratterizzazione dei nuclei di transizione sono spesso molto utili. Per semplicità limitiamo lo studio solo al caso dei nuclei markoviani, ossia gli unici considerati in questa tesi.

Lemma A.4. *Siano $(S,\mathcal{S}),(T,\mathcal{T})$ due spazi misurabili tali che esiste $\mathcal{C}\subseteq P(S)$ un sistema $\cap$-stabile tale che $\sigma(\mathcal{C})=\mathcal{S}$, e sia $\kappa=\{\kappa_s \mid s\in S\}$ una famiglia di misure di probabilità su $(T,\mathcal{T})$. Allora le seguenti affermazioni sono equivalenti*

1. *$\mu\colon S\times\mathcal{T}\to[0,1], \quad (s,B)\mapsto\kappa_s(B)$ è un nucleo markoviano da S a T;*
2. *la funzione $s\mapsto\kappa_s$ è misurabile da S a $\mathcal{P}(T)$;*
3. *$s\mapsto\kappa_s(B)$ è una funzione misurabile da S a $[0,1]$ per ogni $B\in\mathcal{T}$.*

Per la dimostrazione si veda il Lemma 1.37 in [5].

Nel seguito sia $(\Omega,\mathcal{A},\mathbb{P})$ uno spazio di probabilità fissato, $(S,\mathcal{S})$ e $(T,\mathcal{T})$ due spazi misurabili. È noto che data una sotto σ-algebra $\mathcal{F}\subseteq\mathcal{A}$ la probabilità condizionata di un evento $A\in\mathcal{A}$ data $\mathcal{F}$ è definita tramite il valore atteso condizionato. Più precisamente

$$\mathbb{P}(A\mid\mathcal{F}):=\mathbb{E}^{\mathbb{P}}\big[\mathbf{1}_A\mid\mathcal{F}\big],$$

essa è pertanto la v.a. $\mathbb{P}$-q.c. unica tale che

$$\mathbb{E}^{\mathbb{P}}\big[\mathbb{P}(A\mid\mathcal{F})\mathbf{1}_B\big]=\mathbb{E}^{\mathbb{P}}\big[\mathbf{1}_A\mathbf{1}_B\big]=\mathbb{P}(A\cap B),\quad\forall B\in\mathcal{F}.$$

Si noti che $\mathbb{P}(A\mid\mathcal{F})=\mathbb{P}(A)$ $\mathbb{P}$-q.c. se e solo se A è indipendente da $\mathcal{F}$ e $\mathbb{P}(A\mid\mathcal{F})=\mathbf{1}_A$ $\mathbb{P}$-q.c. se e solo se A è, a meno di insiemi di misura nulla, uguale ad un elemento di $\mathcal{F}$. Dalla monotonia del valore atteso condizionato otteniamo $0\le\mathbb{P}(A\mid\mathcal{F})\le 1$ q.c., e dal teorema di convergenza monotona otteniamo che per una sequenza di eventi disgiunti $(A_n)_{n\in\mathbb{N}}\subseteq\mathcal{A}$ si ha

$$\mathbb{P}\Big(\bigcup_{n\in\mathbb{N}}A_n\mid\mathcal{F}\Big)=\sum_{n\in\mathbb{N}}\mathbb{P}(A_n\mid\mathcal{F})\quad\mathbb{P}\text{-q.c.}$$

Tuttavia nella precedente uguaglianza l'insieme di misura nulla su cui essa non vale dipende in generale dalla sequenza $(A_n)_{n\in\mathbb{N}}$, quindi in generale $\mathbb{P}(\cdot\mid\mathcal{F})$ non è una misura di probabilità su Ω. Questo problema si può aggirare usando i nuclei markoviani e definendo le *distribuzioni condizionali*.

Definizione A.8. Sia $(\Omega,\mathcal{A},\mathbb{P})$ uno spazio di probabilità, $\mathcal{F}\subseteq\mathcal{A}$ una sotto σ-algebra, $(S,\mathcal{S})$ uno spazio misurabile e ξ una v.a. su Ω a valori in S. Diciamo che un nucleo markoviano $\kappa_{\xi,\mathcal{F}}$ è una *distribuzione condizionale regolare di ξ data $\mathcal{F}$* se per $\mathbb{P}$-q.o. $\omega\in\Omega$ e per ogni $B\in\mathcal{S}$ vale

$$\kappa_{\xi,\mathcal{F}}(\omega,B)=\mathbb{P}(\xi\in B\mid\mathcal{F}).$$

Ossia deve valere per ogni $A\in\mathcal{F},B\in\mathcal{S}$

$$\mathbb{P}(\{\xi\in B\}\cap A)=\int_\Omega\mathbf{1}_B(\xi)\mathbf{1}_A\,d\mathbb{P}=\int_\Omega\kappa_{\xi,\mathcal{F}}(\omega,B)\mathbf{1}_A(\omega)\,d\mathbb{P}(\omega).$$

Se poi η è un'altra v.a. su Ω a valori in un altro spazio misurabile $(T, \mathcal{T})$, una distribuzione condizionale regolare di ξ data η è un nucleo markoviano κ da T in S che verifica

$$\kappa(\omega, B) = \mathbb{P}(\xi \in B \mid \eta) \quad \mathbb{P}\text{-q.c.}, \quad \forall B \in \mathcal{S}.$$

Teorema A.5 (Distribuzioni condizionali in $\mathbb{R}$.)**.** *Sia $(\Omega, \mathcal{A}, \mathbb{P})$ uno spazio di probabilità, $\mathcal{F} \subseteq \mathcal{A}$ una sotto σ-algebra e sia $\xi\colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ una v.a., allora esiste una* distribuzione condizionale regolare *$\kappa_{\xi,\mathcal{F}}$ di ξ data $\mathcal{F}$.*

Dimostrazione. Per $r \in \mathbb{Q}$, sia $F(r, \cdot)$ una versione della probabilità condizionale $\mathbb{P}(\xi \in (-\infty, r] \mid \mathcal{F})$. Per $r \leq s$ si ha chiaramente $\mathbf{1}_{\{\xi\in(-\infty,r]\}} \leq \mathbf{1}_{\{\xi\in(-\infty,s]\}}$, quindi dalla monotonia del valore atteso condizionato si ha che esiste un evento di probabilità nulla $A_{r,s} \in \mathcal{F}$ tale che

$$F(r, \omega) \leq F(s, \omega) \quad \forall \omega \in A_{r,s}. \tag{A.1}$$

Dal teorema di convergenza dominata abbiamo inoltre che esistono eventi di probabilità nulla $(B_r)_{r\in\mathbb{Q}} \subseteq \mathcal{F}, C \in \mathcal{F}$ tali che

$$\lim_{n\to\infty} F\Big(r + \frac{1}{n}, \omega\Big) = F(r, \omega) \quad \forall \omega \in \Omega \setminus B_r \tag{A.2}$$

e anche

$$\inf_{n\in\mathbb{N}} F(-n, \omega) = 0 \quad \text{e} \quad \sup_{n\in\mathbb{N}} F(n, \omega) = 1 \quad \forall \omega \in \Omega \setminus C. \tag{A.3}$$

Definiamo allora

$$N := \Big(\bigcup_{r,s\in\mathbb{Q}} A_{r,s}\Big) \cup \Big(\bigcup_{r\in\mathbb{Q}} B_r\Big) \cup C,$$

e per $\omega \in \Omega \setminus N$

$$\tilde{F}(z, \omega) := \inf_{\substack{r\in\mathbb{Q}\\ r>z}} F(r, \omega) \quad \forall z \in \mathbb{R}.$$

Per costruzione $\tilde{F}(\cdot, \omega)$ è crescente e continua a destra. Inoltre, da (A.1) e (A.2), abbiamo

$$\tilde{F}(z, \omega) = F(z, \omega) \quad \forall z \in \mathbb{Q}, \omega \in \Omega \setminus N. \tag{A.4}$$

Quindi da (A.3), $\tilde{F}(\cdot, \omega)$ è una funzione di ripartizione per ogni $\omega \in \Omega \setminus N$. Per $\omega \in N$ definiamo $\tilde{F}(\cdot, \omega) = F_0$, ove F_0 è una funzione di ripartizione su $\mathbb{R}$ arbitrariamente fissata. Per ogni $\omega \in \Omega$ sia ora $\kappa(\omega, \cdot)$ la misura di probabilità su $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ con funzione di distribuzione $\tilde{F}(\cdot, \omega)$. In più per $r \in \mathbb{Q}$ e $B = (-\infty, r]$ la mappa

$$\omega \mapsto \kappa(\omega, B) = F(r, \omega)\mathbf{1}_{N^c}(\omega) + F_0(r)\mathbf{1}_N(\omega) \tag{A.5}$$

è $\mathcal{F}$-misurabile. Ora $\{(-\infty, r] \mid r \in \mathbb{Q}\}$ è un sistema $\cap$-stabile che genera $\mathscr{B}(\mathbb{R})$. Allora dall'osservazione A.2, otteniamo allora ch ela misurabilità vale per ogni $B \in \mathscr{B}(\mathbb{R})$ e quindi κ è un nucleo markoviano. Rimane da verificare che effitivamente κ

è una distribuzione condizionale regolare di ξ data $\mathcal{F}$. Per farlo osserviamo che per ogni $A \in \mathcal{F}, r \in \mathbb{Q}$, detto $B = (-\infty, r]$, dalla (A.4) e dalla definizione di F segue

$$\int_A \kappa(\omega, B)\, d\mathbb{P}(\omega) = \int_A \mathbb{P}(\xi \in B \mid \mathcal{F})\, d\mathbb{P} = \mathbb{P}(A \cap \{\xi \in B\}).$$

Ora come funzione di B, sia il primo che l'ultimo membro dell'equazione precedente sono misure finite su $\mathscr{B}(\mathbb{R})$ che coincidono sul generatore $\cap$-stabile $\{(-\infty, r] \mid r \in \mathbb{Q}\}$. Dall'unicità dell'estensione di misure otteniamo le uguaglianze sopra per ogni $B \in \mathscr{B}(\mathbb{R})$. Dunque $\mathbb{P}$-q.c. $\kappa(\cdot, B) = \mathbb{P}(\xi \in B \mid \mathcal{F})$ e quindi $\kappa = \kappa_{\xi,\mathcal{F}}$. □

Definizione A.9. Due spazi misurabili $(S, \mathcal{S}), (T, \mathcal{T})$ si dicono *isomorfi*, se esiste una funzione biiettiva $\phi\colon S \to T$ tale che ϕ sia $\mathcal{S}/\mathcal{T}$-misurabile e la funzione inversa, ϕ^{-1}, sia $\mathcal{T}/\mathcal{S}$-misurabile. Diremo anche che ϕ è un *isomorfismo di spazi misurabili*. Uno spazio misurabile $(S, \mathcal{S})$ si dice *di Borel* se esiste un boreliano $C \in \mathcal{B}(\mathbb{R})$ tale che $(S, \mathcal{S})$ e $(C, \mathcal{C})$ siano isomorfi.

Teorema A.6. *Uno spazio Polacco (E, τ) con la sua σ-algebra dei boreliani è sempre uno spazio di Borel.*

Teorema A.7. *Sia $(\Omega, \mathcal{A}, \mathbb{P})$ uno spazio di probabilità, $\mathcal{F} \subseteq \mathcal{A}$ una sotto-σ-algebra e ξ una v.a. a valori in uno spazio di Borel $(S, \mathcal{S})$. Allora esiste una distribuzione condizionale regolare di ξ data $\mathcal{F}$*

Dimostrazione. Sia $B \in \mathscr{B}(\mathbb{R})$ e sia $\phi\colon S \to B$ un isomorfismo di spazi misurabili. Dal teorema A.5, otteniamo che esiste una distribuzione condizionale regolare di ξ' data $\mathcal{F}$, ove ξ' è la v. a. reale $\xi' := \phi \circ \xi$. Allora defininendo per ogni $A \in \mathcal{S}$

$$\kappa(\omega, A) := \kappa_{\xi',\mathcal{F}}(\omega, \phi(A)),$$

abbiamo che κ è banalmente una distribuzione condizionale regolare di ξ data $\mathcal{F}$. Infatti è ovviamente un nucleo markoviano ma inoltre per ogni $A \in \mathcal{F}, C \in \mathcal{S}$

$$\begin{aligned}\int_A \kappa(\omega, S)\, d\mathbb{P}(\omega) &= \int_A \kappa_{\xi',\mathcal{F}}(\omega, \phi(S))\, d\mathbb{P}(\omega) \\ &= \mathbb{P}(A \cap \{\xi' \in \phi(S)\}) = \mathbb{P}(A \cap \{\xi \in S\}).\end{aligned}$$ □

Teorema A.8 (Disintegrazione)**.** *Sia ξ una v.a. sullo spazio di probabilità $(\Omega, \mathcal{A}, \mathbb{P})$ a valori in uno spazio di Borel $(S, \mathcal{S})$. Sia poi $\mathcal{F} \subseteq \mathcal{A}$ una sotto-σ-algebra e $\kappa_{\xi,\mathcal{F}}$ una distribuzione condizionale regolare di ξ data $\mathcal{F}$. Sia inoltre $f\colon S \to \mathbb{R}$ misurabile e tale che $\mathbb{E}\big[|f(\xi)|\big] < \infty$. Allora*

$$\mathbb{E}\big[f(\xi) \mid \mathcal{F}\big](\omega) = \int_S f\, d\kappa_{\xi,\mathcal{F}}(\omega, \cdot) \quad \textit{per } \mathbb{P}\textit{-q.o. } \omega \in \Omega. \tag{A.6}$$

Dimostrazione. Per provare il teorema mostriamo che il membro di destra in (A.6) verifica le proprietà del valore atteso condizionato. Per linearità, potendo passare

alle parti positiva e negativa possiamo assumere $f \geq 0$. Allora esistono degli insiemi $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{S}$ e dei numeri $(\alpha_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}_+$ tali che

$$g_n := \sum_{j=1}^{n} \alpha_j \mathbf{1}_{A_j} \xrightarrow[n\to\infty]{} f.$$

Ora, per ogni $n \in \mathbb{N}$ e $B \in \mathcal{F}$,

$$\begin{aligned}
\mathbb{E}\big[g_n(\xi)\mathbf{1}_B\big] &= \sum_{j=1}^{n} \alpha_j \mathbb{P}\big(\{\xi \in A_j\} \cap B\big) \\
&= \sum_{j=1}^{n} \alpha_j \int_B \mathbb{P}\big(\{\xi \in A_j\} \mid \mathcal{F}\big)\, d\mathbb{P} \\
&= \sum_{j=1}^{n} \alpha_j \int_B \kappa_{\xi,\mathcal{F}}(\omega, A_j)\, d\mathbb{P}(\omega) \\
&= \int_B \sum_{j=1}^{n} \alpha_j \int_B \kappa_{\xi,\mathcal{F}}(\omega, A_j)\, d\mathbb{P}(\omega) \\
&= \int_B \left(\int_S g_n(x)\, d\kappa_{\xi,\mathcal{F}}(\omega, x) \right) d\mathbb{P}(\omega).
\end{aligned}$$

Per il teorema di convergneza monotona, per $\mathbb{P}$-q.o. ω, l'integrale interno converge a

$$\int_S f(x)\, d\kappa_{\xi,\mathcal{F}}(\omega, x).$$

E ora applicando nuovamente il teorema di convergneza monotona, otteniamo

$$\mathbb{E}\big[f(\xi)\mathbf{1}_B\big] = \lim_{n\to\infty} \mathbb{E}\big[g_n(\xi)\mathbf{1}_A\big] = \int_B \int_S f(x)\, d\kappa(\omega, x)\, d\mathbb{P}(\omega). \qquad \square$$

Osservazione A.3. In questa tesi siamo particolarmente interessati al caso in cui consideriamo leggi su spazi prodotto $\gamma \in \mathcal{P}(E \times E)$, con E uno spazio Polacco, pensiamo a γ come la a legge di una variabile aleatoria congiunta (ξ, η) definita su uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ a valori in $E \times E$, ove ξ ed η sono v. a. sullo stesso spazio di probabilità a valori in E, con leggi μ e ν rispettivamente. Le distribuzioni condizionali di η data ξ e di ξ data η si dicono misure ottenute per disintegrazione di γ sulla prima e sulla seconda coordinata. Le indichiamo con

$$\kappa_{\xi,\sigma(\eta)}(\omega, B) = \gamma^x(B) \quad \text{e} \quad \kappa_{\eta,\sigma(\xi)}(\omega, A) = \gamma^y(A)$$

ove $x = \xi(\omega), y = \eta(\omega)$. Allora la loro proprietà caratterizzante diviene

$$\mathbb{P}(\{\xi \in A\} \cap \{\eta \in B\}) = \gamma(A \times B) = \int_A \gamma^x(B)\, d\mu(x) = \int_B \gamma^y(A)\, d\nu(y).$$

Appendice B

Script MATLAB

In questa appendice sono riportati degli script MATLAB che costituiscono implementazioni per il calcolo numerico delle distanze $\mathcal{W}_r$ e $\mathcal{AW}_r$ quando si considerano misure discrete. In particolare si è cercato di definire delle funzioni MATLAB, che calcolano tali distanze tra distribuzioni discrete. Nel caso di $\mathcal{AW}_r$ la funzione definita nel relativo script utilizza l'algoritmo ricorsivo visto nell'ultima sezione del capitolo 2. In ogni caso per tutti i codici presentati in questa sezione i programmi lineari che emergono vengono sempre risolti con il comando predefinito nell'Optimization ToolBox di MATLAB linprog.

Partiamo dalla distanza di Wasserstein usuale. Come visto nella sezione 4 del Capitolo 1, il calcolo di $\mathcal{W}_r$, per il caso di misure discrete a supporto finito su $\mathbb{R}^N$, si riduce ad un programma lineare che ha sempre la stessa matrice dei vincoli, le cui dimensioni dipendono dalla cardinalità dei due supporti. Il codice seguente definisce una funzione matlab che prende come imput, nel seguente ordine, quattro arrays riga contenenti rispettivamente

1. gli elementi del supporto della prima misura;
2. gli elementi del supporto della seconda misura;
3. la densità discreta della prima misura;
4. la densità discreta della seconda;

inoltre il codice prende anche in input il parametro $r \geq 1$, l'ordine della distanza $\mathcal{W}_r$ che vogliamo calcolare e $p \in [1, +\infty)$ che fornisce al programma l'informazione di usare su $\mathbb{R}^N$ la distanza ℓ^p. Dati gli input la funzione semplicemente costruisce la matrice dei vincoli delle giuste dimensioni e calcola le distanze tra gli elementi dei supporti, infine risolve il programma lineare suddetto e ne restituisce la radice r-esima del valore ottimo, ossia esattamente $\mathcal{W}_r(\mu, \nu)$.

```
1 function d_W_r = dWasserstein(omega_1, omega_2, prob_1, prob_2, r, p)
2 n_1 = length(omega_1(1,:));
3 n_2 = length(omega_2(1,:));
4 i_n_1 = ones(1, n_1);
```

```
5  i_n_2 = ones(1, n_2);
6  id_1 = eye(n_1);
7  id_2 = eye(n_2);
8  Aeq = [kron(id_1, i_n_2);
9         kron(i_n_1, id_2)];
10 prob = [prob_1 prob_2]';
11 lb = zeros(n_1*n_2, 1);
12 f = zeros(n_1*n_2,1);
13 for j = 1: n_2
14     for k = 1: n_1
15         f((j - 1)*n_1 + k, 1) = norm((omega_1(:,k)-omega_2(:,j)), p);
16     end
17 end
18 [x, fval, exitflag, output] = linprog(f.^r, [], [], Aeq, prob, lb)
19 d_W_r = fval^(1/r);
20 end
```

Passiamo ora al caso di $\mathcal{AW}_r$. Abbiamo visto che nel caso discreto per modellizzare il flusso di informazione, ovvero le filtrazioni, possiamo usare degli alberi. Per tale motivo i codici da qui in avanti prenderanno come input degli oggetti MATLAB di tipo digraph ovvero appunto dei grafi orientati che rappresentano gli alberi considerati. Supponiamo che essi siano costruiti con il comando G = digraph(A), ove A è la matrice di incidenza dell'albero, in cui le entrate non nulle corrispondono alla presenza del relativo arco nel grafo ed il valore nell'entrata corrisponde al peso associato a quell'arco, ossia la probabilità condizionale del nodo in cui l'arco termina dato il nodo da cui l'arco esce. Assumerememo inoltre che ai nodi degli oggetti digraph considerati siano associati i seguenti attributi:

1. G.Nodes.Names, i nomi dei nodi, che come canonicamente si fa, sono dati tramite stringhe di un carattere ad esempio "1", "2" ecc.

2. G.Nodes.Number che assegna ad ogni nodo un numero, la radice sarà sempre l'1, i suoi diretti successori saranno $2, 3$ ecc. (si noti che per il calcolo di $\mathcal{AW}_r$ la numerazione dei nodi è irrelevante, l'unica cosa che conta è la topologia del grafo, ciò nonostante torna utile numerare i nodi ai fini implementativi).

3. G.Nodes.Value i valori del processo considerato assunto nei vari nodi.

4. G.Nodes.Depht le profondità dei nodi nell'albero.

Mentre agli archi, oltre al peso già assegnato tramite la matrice di incidenza, associamo un numero con l'attributo G.Edges.Number. Fatte queste ipotesi sugli oggetti digraph che daremo in input alla funzione per il calcolo di $\mathcal{AW}_r$ notiamo che per implementare l'algoritmo ricorsivo è necessario risolvere iterativamente l'usuale problema di Wasserstein tra sottoalberi dei due dati usando come distanze quelle calcolate al passo precedente. Quindi è più conveniente definire un'altra funzione che come input prenda due alberi e un vettore di distanze delle giuste dimensioni che restituisca come output la distanza di Wasserstein tra i due alberi dati ottenuta usando tale vettore di distanze. Per definire questa funzione è necessario essere in

grado di costruire dato un albero, un vettore contenente le probabilità totali di ciascuna traiettoria, e tornerà utile all'inizio dell'algoritmo ricorsivo, per calcolare tutte le possibili distanze tra traiettorie, avere una funzione che prende in input un albero e restituisce una riga di vettori colonna, ove ogni colonna è una possibile traiettoria del processo. I seguenti due script definiscono due funzioni che rispondono esattamente alle due esigenze suddette. Per il calcolo delle probabilità totali definiamo prima la funzione costo che prende in input un albero e un suo cammino e calcola il costo del cammino, ovvero nel nostro caso la probabilità della traiettoria.

```
function c = costo(G, path)
c = 1;
num_archi = zeros(1, length(path)-1);
Tavola = table2array(cell2table(G.Edges.EndNodes, "VariableNames", %
                                    ["SourceNode", "EndNode"]));
for i = 1:length(G.Edges.Number)
    for j = 1:length(path)-1
        if Tavola(i,:) == path(j:j+1)
            num_archi(j) = G.Edges.Number(i);
        end
    end
end             %questa prima parte di codice serve
% ad individuare quali archi nell'albero fanno parte del cammino
for k = 1: length(path)-1
    c = c*G.Edges.Weight(num_archi(k));
end
```

Poi grazie a costo definiamo la funzione prob che ha come input un albero e restituisce un array con le probabilità di ciascuna traiettoria.

```
function p = prob(G)
T = max(G.Nodes.Depht);
N_T = [];
for i = 1: length(G.Nodes.Number)
    if G.Nodes.Depht(i) == T
        N_T = [N_T G.Nodes.Name(i)];
    end
end
p = ones(1, length(N_T));
paths = [];
for i = 1:length(N_T)
    paths = [paths shortestpath(G, "1", N_T(i))];
end
for i = 1: length(N_T)
    p(i) = costo(G, [paths((1 + (T+1)*(i-1)):(T+1)*i)]);
end
```

In modo del tutto analogo per il calcolo delle traietorie definiamo prima la funzione traj, che dato un albero e un cammino, calcola la traiettoria associata al cammino.

```
function v = traj(G, path)
v = zeros(length(path),1);
for i = 1:length(path)
    for j = 1:length(G.Nodes.Number)
        if G.Nodes.Name(j) == path(i)
```

```
            v(i) = G.Nodes.Value(j);
        end
    end
end
```

E poi definiamo la funzione `trajectories` che ha come input un albero e restituisce in output un array contenente le traiettorie possibili del processo associato all'albero.

```
function Omega = trajectories(G)
T = max(G.Nodes.Depht);
N_T = [];
for i = 1: length(G.Nodes.Number)
    if G.Nodes.Depht(i) == T
        N_T = [N_T G.Nodes.Name(i)];
    end
end
paths = [];
for i = 1:length(N_T)
    paths = [paths shortestpath(G, "1", N_T(i))];
end
Omega = zeros(T+1, length(N_T));
for i = 1:length(N_T)
    Omega(:,i) = traj(G, [paths((1 + (T+1)*(i-1)):(T+1)*i)]);
end
```

Con le funzioni definite negli script precedenti è facile scrivere l'implementazione per il calcolo di $\mathcal{W}_r$ per le distribuzioni multivariate associate a dei processi ad albero. Il codice si scrive facilmente modificando leggermente quello iniziale.

```
function d_A_r = dWassTreeProcesses(G_1, G_2, D, r)
prob_1 = prob(G_1);
prob_2 = prob(G_2);
n_1 = length(prob_1);
n_2 = length(prob_2);
i_n_1 = ones(1, n_1);
i_n_2 = ones(1, n_2);
id_1 = eye(n_1);
id_2 = eye(n_2);
Aeq = [kron(id_1, i_n_2);
       kron(i_n_1, id_2)];
PROBAB = [prob_1 prob_2]';
lb = zeros(n_1*n_2, 1);
[x, fval] = linprog(D.^r, [], [], Aeq, PROBAB, lb);
d_A_r = fval^(1/r);
end
```

Per il calcolo numerico di $\mathcal{AW}_r$ tornerà utile anche la seguente funzione, la quale costruisce dato un albero, un cell array di lunghezza pari alla profondità dell'albero che ha come compenente t-esima un array contenente i nomi dei nodi di G a profondità t, per ogni $0 \leq t \leq T$.

```
function N = nodi(G)
T = max(G.Nodes.Depht);
N = cell(1,T+1);
```

```
for i = 1:T+1
    N{i} = [];
end
for i = 1:T+1
    for j = 1:length(G.Nodes.Number)
        if G.Nodes.Depht(j) == i-1
            N{i} = [N{i} G.Nodes.Name(j)];
        end
    end
end
```

Infine lo script che definisce la funzione per il calcolo di $\mathcal{AW}_r$ è il seguente. Oltre ai due alberi la funzione accetta in input anche i parametri $p \in [1, \infty)$, che da l'informazione di calcolare le distanze tra le traiettorie con la metrica ℓ^p e $r \geq 1$ l'ordine della distanza $\mathcal{AW}_r$.

```
function AW = NestedDistance(G_1, G_2,p, r)
T = max(G_1.Nodes.Depht);
E_1 = trajectories(G_1);
E_2 = trajectories(G_2);
n_1 = length(E_1(1,:));
n_2 = length(E_2(1,:));
D_T = zeros(n_1*n_2, 1);
d = cell(1,T+1);
for j = 1: n_1
    for k = 1: n_2
       D_T((j - 1)*n_2+ k, 1) = norm((E_1(:,k)-E_2(:,j)), p);
    end
end
N_1 = nodi(G_1);
N_2 = nodi(G_2);
for i = 1:T
    d{i} = zeros(length(N_1{i})*length(N_2{i}),1);
end                    %questa prima parte di codice inizializza l'algoritmo
d{T+1} = D_T; %calcola le distanze tra tutte le traiettorie
for i = 0:T-1       %qui inizia la ricorsione
    for j = 1:length(N_1{T - i})
        for k = 1:length(N_2{T - i})
            s_1 = successors(G_1, N_1{T - i}(j));
            s_2 = successors(G_2, N_2{T - i}(k));
            g_1_temp = subgraph(G_1, [N_1{T - i}(j); s_1]);
            g_2_temp = subgraph(G_2, [N_2{T - i}(k); s_2]);
            g_1_temp.Nodes.Name = ["1" g_1_temp.Nodes.Name(2:end)']';
            g_2_temp.Nodes.Name = ["1" g_2_temp.Nodes.Name(2:end)']';
            numero_nodi_1 = 1:numnodes(g_1_temp);
            numero_nodi_2 = 1:numnodes(g_2_temp);
            g_1_temp.Nodes.Number = numero_nodi_1';
            g_2_temp.Nodes.Number = numero_nodi_2';
            g_1_temp.Nodes.Depht = [0 ones(1, numnodes(g_1_temp)-1)]';
            g_2_temp.Nodes.Depht = [0 ones(1, numnodes(g_2_temp)-1)]';
            numero_edges_1 = 1:numedges(g_1_temp);
            numero_edges_2 = 1:numedges(g_2_temp);
            g_1_temp.Edges.Number = numero_edges_1';
            g_2_temp.Edges.Number = numero_edges_2';
            prima_1 = 0;
```

```
40                  prima_2 = 0;
41                  for l = 1:j-1
42                      prima_1 = prima_1 + length(successors(G_1, N_1{T - i}(l)));
43                  end
44                  for l = 1:k-1
45                      prima_2 = prima_2 + length(successors(G_2, N_2{T - i}(l)));
46                  end
47                  d_temp = d{T - i + 1}(prima_1*length(N_2{T - i + 1}) + prima_2 + 1 : %
48                     prima_1*length(N_2{T - i + 1}) + prima_2 + length(s_1)*length(s_2),1);
49                  d{T - i}((j - 1)*length(N_2{T - i}) + k, 1) = %
50                                      dWassTreeProcesses(g_1_temp, g_2_temp, d_temp, r);
51              end
52          end
53      end                 % qui termina il calcolo ricorsivo e si assegna il valore a AW
54      AW = d{1};
55      end
```

Bibliografia

[1] Backhoff Veraguas, Julio; Beiglböck, Mathias; Eder, Manu; Pichler, Alois (2020): Fundamental Properties of Process Distances, *Stochastic Processes and their Applications* 130, 5575-5591.

[2] Backhoff Veraguas, Julio; Beiglböck, Mathias; Yiqing, Lin; Zalashko, Anastasiia: Causal Transport in Discrete Time and Applications, *Society for Industrial and Applied Mathematematics Journal on Optimization* 27 (4), 2528-2562, 2017.

[3] Bartl, Daniel; Wiesel, Johannes: Sensitivity of Multiperiod Optimization Problems with respect to the Adapted Wasserstein Distance, *Society for Industrial and Applied Mathematics Journal on Financial Mathematics*, 14(2):704-720, 2023.

[4] Bolley, François (2008): Separability and Completeness for the Wasserstein Distance, *Séminaire de probabilités XLI*, 371-377.

[5] Kallenberg, Olav (1997): *Foundations of Modern Probability*, Springer.

[6] Klemke, Achim (2014): *Probability Theory, a Comprihensive Course*, Second Edition, Springer.

[7] Komlos, Janos: A generalization of a problem of Steinhaus. *Acta Mathematica Academiae Scientiarium Hungaricae*, 18(1-2):217-229, 1967.

[8] Kovacevic, Raimund M.; Pichler, Alois (2016): Tree Approximation for Discrete Time Stochastic Processes - A Process Distance Approach, *Annals of Operation Research* 235 (1), 395-421.

[9] Lassalle, Rémi (2013): Causal Transference Plans and their Monge-Kantorovich Problems.

[10] Plug, George Ch.; Pichler, Alois (2012): A Distance for Multistage Stochastic Optimization Models, *Society for Industrial and Applied Mathematics Journal on Optimization* Vol. 22, No. 1, pp. 1-23.

[11] Pflug, George Ch.; Pichler, Alois (2014): *Multistage Stochastic Optimization*, Springer, Springer Series in Operation Research and Financial Engineering.

[12] Villani, Cedric (2003): *Topics in Optimal Transportation*, American mathematical society, Graduate studies in mathematics, vol.58.