



HAL
open science

Asymptotical estimates for some algorithms for data and image processing: a study of the Sinkhorn algorithm and a numerical analysis of total variation minimization

Corentin Caillaud

► To cite this version:

Corentin Caillaud. Asymptotical estimates for some algorithms for data and image processing: a study of the Sinkhorn algorithm and a numerical analysis of total variation minimization. Optimization and Control [math.OC]. Institut Polytechnique de Paris, 2020. English. NNT: 2020IPPAX023. tel-02926037

HAL Id: tel-02926037

<https://theses.hal.science/tel-02926037>

Submitted on 31 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2020IPPAX023

Thèse de doctorat



Asymptotical estimates for some algorithms for data and image processing: a study of the Sinkhorn algorithm and a numerical analysis of total variation minimization

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 20 juillet 2020, par

CORENTIN CAILLAUD

Composition du Jury :

Julie Delon Professeure, Université de Paris (MAP5)	Présidente
Nicolas Papadakis Chargé de recherche, CNRS et Université de Bordeaux (IMB)	Rapporteur
Gabriele Steidl Professeure, Université de Kaiserslautern	Rapporteuse
Jalal Fadili Professeur, ENSICAEN (GREYC)	Examineur
Stéphane Gaubert Directeur de recherche, INRIA et École polytechnique (CMAP)	Examineur
Antonin Chambolle Directeur de recherche, CNRS et École polytechnique (CMAP)	Directeur de thèse

Version du 27 juillet 2020 pour le second dépôt en ligne.

Titre en Français : Estimations asymptotiques pour des algorithmes de traitement d'images et de données : une étude de l'algorithme de Sinkhorn et une analyse numérique de la minimisation de la variation totale.

REMERCIEMENTS

Je tiens d'abord à remercier vivement mon directeur de thèse, Antonin Chambolle, qui a su me guider dans des domaines mathématiques qui m'étaient inconnus, toujours avec bienveillance, en bouillonnant d'idées, et en me laissant beaucoup de liberté. Je remercie Gabriele Steidl et Nicolas Papadakis pour avoir accepté de rapporter cette thèse, et Julie Delon et Stéphane Gaubert pour leur participation au jury. Merci aussi à Jalal Fadili qui en plus de faire partie du jury a assuré mon suivi à mi-parcours.

Je remercie ensuite toute l'équipe du CMAP pour son accueil durant ces trois années. Un grand merci à Nasséra Naar et Alexandra Noiret pour leur efficacité face aux subtilités administratives, et à Sylvain Ferrand et Pierre Straeblers toujours si gentiment disponibles pour régler les tracas informatiques. Merci également aux autres doctorants du CMAP et d'ailleurs avec qui j'ai partagé mes interrogations : Julia, Mathilde, Thomas, Léo, Isao ; des parties de badminton : Perle, Tristan, Hadrien, Paul, Jean-Bernard ; et le bureau 2015 pour la plupart du temps : Aude, Eugénie, Luca, Rémi, Frédéric et Corentin.

Pour finir, je souhaite remercier très particulièrement Anne-Elisabeth, Baptiste, Clément, Thibaut ainsi que toute ma famille pour leur soutien sans faille.

RÉSUMÉ

Cette thèse traite de problèmes discrets d'optimisation convexe et s'intéresse à des estimations de leurs taux de convergence. Elle s'organise en deux parties indépendantes.

Dans la première partie, nous étudions le taux de convergence de l'algorithme de Sinkhorn et de certaines de ses variantes. Cet algorithme apparaît dans le cadre du Transport Optimal (TO) par l'intermédiaire d'une régularisation entropique. Ses itérations, comme celles de ses variantes, s'écrivent sous la forme de produits composante par composante de matrices et de vecteurs positifs. Pour les étudier, nous proposons une nouvelle approche basée sur des inégalités de convexité simples et menant au taux de convergence linéaire observé en pratique. Nous étendons ce résultat à un certain type de variantes de l'algorithme que nous appelons algorithmes de Sinkhorn équilibrés de dimension 1. Nous présentons ensuite des techniques numériques traitant le cas de la convergence vers zéro du paramètre de régularisation des problèmes de TO. Enfin, nous menons l'analyse complète du taux de convergence en dimension 2.

Dans la deuxième partie, nous donnons des estimations d'erreur pour deux discrétisations de la variation totale (TV) dans le modèle de Rudin, Osher et Fatemi (ROF). Ce problème de débruitage d'image, qui revient à calculer l'opérateur proximal de la variation totale, bénéficie de propriétés d'isotropie assurant la conservation de discontinuités nettes dans les images débruitées, et ce dans toutes les directions. En discrétisant le problème sur un maillage carré de taille h et en utilisant une variation totale discrète standard dite TV isotrope, cette propriété est perdue. Nous démontrons que dans une direction particulière l'erreur sur l'énergie est d'ordre $h^{2/3}$, ce qui est relativement élevé face aux attentes pour de meilleures discrétisations. Notre preuve repose sur l'analyse d'un problème équivalent en dimension 1 et de la TV perturbée qui y intervient. La deuxième variation totale discrète que nous considérons copie la définition de la variation totale continue en remplaçant les champs duaux habituels par des champs discrets dits de Raviart-Thomas. Nous retrouvons ainsi le caractère isotrope du modèle ROF discret. Pour conclure, nous prouvons, pour cette variation totale et sous certaines hypothèses, une estimation d'erreur en $O(h)$.

ABSTRACT

This thesis deals with discrete optimization problems and investigates estimates of their convergence rates. It is divided into two independent parts.

The first part addresses the convergence rate of the Sinkhorn algorithm and of some of its variants. This algorithm appears in the context of Optimal Transportation (OT) through entropic regularization. Its iterations, and the ones of the Sinkhorn-like variants, are written as componentwise products of nonnegative vectors and matrices. We propose a new approach to analyze them, based on simple convex inequalities and leading to the linear convergence rate that is observed in practice. We extend this result to a particular type of variants of the algorithm that we call 1D balanced Sinkhorn-like algorithms. In addition, we present some numerical techniques dealing with the convergence towards zero of the regularizing parameter of the OT problems. Lastly, we conduct the complete analysis of the convergence rate in dimension 2.

In the second part, we establish error estimates for two discretizations of the total variation (TV) in the Rudin-Osher-Fatemi (ROF) model. This image denoising problem, that is solved by computing the proximal operator of the total variation, enjoys isotropy properties ensuring the preservation of sharp discontinuities in the denoised images in every direction. When the problem is discretized into a square mesh of size h and one uses a standard discrete total variation – the so-called isotropic TV – this property is lost. We show that in a particular direction the error in the energy is of order $h^{2/3}$ which is relatively large with respect to what one can expect with better discretizations. Our proof relies on the analysis of an equivalent 1D denoising problem and of the perturbed TV it involves. The second discrete total variation we consider mimics the definition of the continuous total variation replacing the usual dual fields by discrete Raviart-Thomas fields. Doing so, we recover an isotropic behavior of the discrete ROF model. Finally, we prove a $O(h)$ error estimate for this variant under standard hypotheses.

CONTENTS

Notations	7
I Linear convergence rate of Sinkhorn-like algorithms	10
Introduction de la partie I	11
1 Emergence of the Sinkhorn algorithm	19
1.1 Historical example of Telephone Forecasting	19
1.2 Appearance in Optimal Transportation	21
2 Linear convergence of the classical Sinkhorn algorithm	26
2.1 First linear rates for positive matrices	26
2.1.1 Min-max analysis	26
2.1.2 Hilbert distance analysis	28
2.2 General linear rate	30
2.2.1 Convergence conditions	30
2.2.2 A linear convergence rate	34
2.3 Proof of general theorem	40
3 Linear convergence of Sinkhorn-like algorithms	43
3.1 Barycenter of two measures	43
3.1.1 Simple barycenter	43
3.1.2 Multiple barycenters	50
3.2 Graph labelling problem	54
3.2.1 General setting	54
3.2.2 Case of 1D graphs	59

3.3	Final remarks	65
3.3.1	Limits of the analysis: 1D balanced Sinkhorn-like algorithms	65
3.3.2	Degenerate direction of Sinkhorn algorithm	67
4	Issues of the setting when $\varepsilon \rightarrow 0$	74
4.1	Theoretical issues	74
4.2	Numerical issues	78
4.2.1	Log domain computation	78
4.2.2	Bethe entropy for barycenter	79
4.2.3	Iterated process	82
4.3	Behavior of rates in dimension 2	87
4.3.1	Doubly stochastic setting	87
4.3.2	General setting	88
	Conclusion of part I	92
	 II Error estimates for discretizations of the ROF model	 93
	 Introduction de la partie II	 94
5	The ROF model with isotropic total variation	99
5.1	Continuous setting of the ROF model	99
5.2	Isotropic total variation	101
6	Reduction to a 1D total variation denoising problem	105
6.1	Setting on periodic domain	105
6.2	Continuous solution	108
6.3	Form of discrete solution	109
7	Upper bound estimate	111
7.1	General strategy	112
7.2	Approach for a general function	113
7.3	Result for a particular function	116
7.4	Modifications for Neumann boundary conditions	118
8	Lower bound estimate	120
8.1	Dual problem	120
8.2	Change of variables	121
8.3	Study of the limit problem	123
8.3.1	A dual of the dual	123
8.3.2	Strong duality result	126

8.4	Return to the discrete problem	127
8.4.1	Compact support case	128
8.4.2	Positive case	129
8.5	Modifications for Neumann boundary conditions	132
9	The ROF model with Raviart-Thomas total variation	135
9.1	Definitions	135
9.2	Error estimate	139
9.2.1	Primal estimate	140
9.2.2	Dual estimate	142
9.2.3	Combination of the two estimates	145
10	Implementation and results	148
10.1	A united framework	148
10.2	Resolution by a primal-dual algorithm	149
10.3	Numerical results	151
	Conclusion of Part II	154
	Bibliography	155

NOTATIONS

General notations

$\llbracket a, b \rrbracket$	the set of integers $\{a, \dots, b\}$ for $a, b \in \mathbb{Z}$
$\mathbb{R}^+, \mathbb{R}^-$	respectively the sets $[0, +\infty)$ and $(-\infty, 0]$ of nonnegative and nonpositive reals
$\mathbb{R}_*^+, \mathbb{R}_*^-$	respectively the sets $(0, +\infty)$ and $(-\infty, 0)$ of positive and negative reals
$\{x\}_+$	the positive part of $x \in \mathbb{R}$ given by $\{x\}_+ = \max(0, x)$
$Jf(x)$	the Jacobian matrix of the function f at point x
$Df(x)$	the differential of the function f at point x
$f _{\mathcal{X}}$	the restriction of the function f to the smaller domain \mathcal{X}
$f(x^+), f(x^-)$	respectively the right and left limits of the function f at point x
$f'(x^+), f'(x^-)$	respectively the right and left derivatives of the function f at point x
1_E	the indicator function of a set E given by $1_E(x) = 1$ if $x \in E$, 0 otherwise
$ S , E $	the number of elements of a finite set S or the volume of a measurable set E
$\mathcal{X}^{\mathbb{N}}$	the set of sequences of elements of \mathcal{X}
$\partial\Omega$	the boundary of a domain $\Omega \subset \mathbb{R}^d$

Vectors and matrices

$(\mathbb{R}^+)^{d \times d}$	the set of nonnegative square matrices of size d
$(\mathbb{R}_*^+)^{d \times d}$	the set of positive square matrices of size d
$d(x)$	the diagonal matrix of $\mathbb{R}^{d \times d}$ with main diagonal given by $x \in \mathbb{R}^d$
$\mathbf{1}$	the vector $(1, \dots, 1)$ of size given by the context
\underline{X}	the smallest entry of the matrix or vector X
\overline{X}	the largest entry of the matrix or vector X
A^T	the transpose of the matrix A
$Sp(A)$	the spectrum of the matrix A
$\rho(A)$	the spectral radius of the matrix A

Componentwise operations

For vectors $x = (x_1, \dots, x_d), y = (y_1, \dots, y_d) \in \mathbb{R}^d$ and matrices $A, B \in \mathbb{R}^{d \times d}$

$x \leq y$	all the inequalities $x_1 \leq y_1, \dots, x_n \leq y_n$
xy	the vector $(x_1y_1, \dots, x_dy_d) \in \mathbb{R}^d$
$\frac{x}{y}$	the vector $(\frac{x_1}{y_1}, \dots, \frac{x_d}{y_d}) \in \mathbb{R}^d$
x^θ	for $\theta \in \mathbb{R}$, the vector $(x_1^\theta, \dots, x_d^\theta) \in \mathbb{R}^d$
$\log(x)$	the vector $(\log(x_1), \dots, \log(x_d)) \in \mathbb{R}^d$
$\exp(x)$	the vector $(\exp(x_1), \dots, \exp(x_d)) \in \mathbb{R}^d$
$\exp(A)$	the matrix of $\mathbb{R}^{d \times d}$ such that $\exp(A)_{i,j} = \exp(A_{i,j})$
Ax	the usual matrix-vector product
AB	the usual matrix-matrix product
$A \odot B$	the matrix of $\mathbb{R}^{d \times d}$ such that $(A \odot B)_{i,j} = A_{i,j}B_{i,j}$
$A^{\odot p}$	the matrix $A \odot \dots \odot A$ (p times)
$x \oplus y$	the matrix of $\mathbb{R}^{d \times d}$ such that $(x \oplus y)_{i,j} = x_i + y_j$
$x \otimes y$	the matrix of $\mathbb{R}^{d \times d}$ such that $(x \otimes y)_{i,j} = x_i y_j$

Norms and scalar products on \mathbb{R}^d

For a positive vector $s \in (\mathbb{R}_*^+)^d$

$\langle \cdot \cdot \rangle$	the standard scalar product on \mathbb{R}^d or $\mathbb{R}^{d \times d}$ seen as \mathbb{R}^{d^2}
$\ \cdot\ _2$	the standard Euclidean norm on \mathbb{R}^d
$ \cdot $	the standard Euclidean norm on \mathbb{R}^2
$\ \cdot\ _\infty$	the standard ℓ^∞ norm on \mathbb{R}^d given by $\ x\ _\infty = \max_k x_k $
$\ \cdot\ _{\ell^1}$	the standard ℓ^1 norm on \mathbb{R}^d given by $\ x\ _{\ell^1} = \sum_k x_k $
$\langle \cdot \cdot \rangle_s$	the scalar product on \mathbb{R}^d given by $\langle x y \rangle_s = \sum_k s_k x_k y_k$
$\ \cdot\ _s$	the Euclidean norm on \mathbb{R}^d given by $\ x\ _s = \sqrt{\sum_k s_k x_k^2}$
$\langle x \rangle_s$	the value of $\langle x \mathbf{1} \rangle_s$

Function spaces and norms

For an open subset $\Omega \subset \mathbb{R}^d$ and a function $f : \Omega \rightarrow \mathbb{R}$

$\mathcal{C}^k(\Omega, \mathbb{R}^n)$	the space of k times continuously differentiable functions from Ω to \mathbb{R}^n , smooth functions for $k = \infty$
$\mathcal{C}^k(\Omega)$	the space $\mathcal{C}^k(\Omega, \mathbb{R})$
$\mathcal{C}_c^k(\Omega)$	the space of functions of $\mathcal{C}^k(\Omega)$ with compact support
$H^1(\Omega)$	the space of functions of $L^2(\Omega)$ whose distributional derivatives belong to $L^2(\Omega)$
$\ f\ _{L^p(\Omega)}$	the standard L^p norm given by $\ f\ _{L^p(\Omega)} = (\int_\Omega f ^p)^{1/p}$ for $p \in [1, \infty)$
$\ f\ _\infty$	the standard L^∞ norm given by $\ f\ _\infty = \sup_{x \in \Omega} f(x) $
$\ f\ _{\infty, \mathcal{X}}$	the value of $\ f _{\mathcal{X}}\ _\infty$

Part I

Linear convergence rate of Sinkhorn-like algorithms

INTRODUCTION DE LA PARTIE I

Dans la première partie de cette thèse nous nous intéressons à un algorithme qu'il est possible de décrire en seulement quelques lignes de code Matlab¹ :

```
1 a = ones (N, 1) ;
2 b = ones (N, 1) ;
3 for k=1:100 % nombre d'itérations souhaitées
4     a = f ./ A*b;
5     b = g ./ A'*a;
6 end
7 X = diag (a) *A*diag (b) ;
```

Étant donné une matrice positive $A \in (\mathbb{R}^+)^{N \times N}$ et deux vecteurs $f, g \in (\mathbb{R}_*^+)^N$, l'objectif de cet algorithme est de trouver des réels α_i, β_j tels que la matrice X donnée par $X_{i,j} = \alpha_i A_{i,j} \beta_j$ ait f pour somme de ses lignes et g pour somme de ses colonnes. Comme ces conditions se traduisent par les deux équations $\alpha = f/A\beta$ et $\beta = g/A^T\alpha$ (où la division s'entend composante par composante), la démarche proposée par cet algorithme est assez naturelle : après avoir initialisé α et β à $\mathbf{1}$, modifier α pour que la première équation soit satisfaite, puis β pour que la deuxième le soit, et ainsi de suite. Il s'agit donc de la procédure suivante :

Algorithme I.1. *Étant donné une matrice positive $A \in (\mathbb{R}^+)^{N \times N}$ et deux vecteurs $f, g \in (\mathbb{R}_*^+)^N$, depuis $\alpha^{(0)} = \beta^{(0)} = \mathbf{1}$, faire pour $n = 0, 1, \dots$*

$$\alpha^{(n+1)} = \frac{f}{A\beta^{(n)}} ; \beta^{(n+1)} = \frac{g}{A^T\alpha^{(n+1)}} \quad (1)$$

¹Rappelons que $\text{ones}(N, 1)$ désigne le vecteur de \mathbb{R}^N dont toutes les composantes sont 1, que $./$ réalise la division composante par composante, que A' désigne la transposée de A et enfin que la fonction diag forme la matrice diagonale dont les coefficients sont ceux de son argument.

Un examen très complet des occurrences de cet algorithme et des différentes approches envisagées pour son étude est proposé par Idel dans [Ide16]. On y comprend que la simplicité des itérations (1) a conduit à leur redécouverte successive par des communautés mathématiques indépendantes. La multiplicité des dénominations de cet algorithme nous paraît illustrer la diversité des domaines où il apparaît; sans volonté d’exhaustivité citons : “méthode des doubles facteurs” pour Kruithof aux Pays-Bas dans les années 30 (voir [Kru37, dB]), nous développerons cet exemple en guise d’introduction à l’algorithme; “Iterative proportional fitting procedure (IPFP)” ou “RAS method”, noms plus communs en économie ou en traitement de données statistiques; ainsi qu’“algorithme de Sinkhorn” du nom d’un mathématicien américain ayant donné les premières preuves de convergence dans les années 60 dans [Sin64, Sin67] puis dans [SK67] en collaboration avec Knopp. Ce dernier nom semble finalement s’être relativement imposé dans le domaine de l’optimisation – même si on note encore aujourd’hui l’existence de deux pages Wikipedia indépendantes “Iterative Fitting Procedure” [Wika] et “Sinkhorn’s theorem” [Wikb]. Nous désignerons donc l’algorithme I.1 par “algorithme de Sinkhorn”.

Dans cette thèse, le cadre dans lequel nous étudions l’algorithme de Sinkhorn est celui fourni par la théorie du Transport Optimal. Le problème de Transport Optimal trouve son origine dans les travaux de Monge [Mon81] à la fin du XVIII^{ème} siècle puis est reformulé sous la forme qui nous intéresse par Kantorovitch [Kan42] dans les années 40. Dans le cadre discret, il s’agit de trouver un plan de transport $x^* \in \Pi(f, g) := \{x \in (\mathbb{R}^+)^{N \times N} \text{ tel que } x\mathbf{1} = f, x^T\mathbf{1} = g\}$ entre deux mesures de probabilité discrètes $f, g \in \Sigma^N := \{p \in (\mathbb{R}_*^+)^N \text{ tel que } \sum_k p_k = 1\}$ minimisant le coût de transport correspondant $\sum_{i,j} w_{i,j} x_{i,j}$ associé à un coût $w \in \mathbb{R}^{N \times N}$. Il s’agit donc du problème linéaire suivant :

$$W_w(f, g) = \min_{x \in \Pi(f, g)} \langle w | x \rangle \quad (2)$$

Les algorithmes usuels pour résoudre ce problème linéaire (notamment la méthode dite hongroise ou des enchères, voir [Wal17]) se révèlent difficiles à mettre en œuvre quand la dimension N devient élevée. Une solution alternative, popularisée notamment par Cuturi dans [Cut13], consiste à calculer uniquement une approximation $W_w^\varepsilon(f, g)$ de $W_w(f, g)$ via une méthode de point intérieur par perturbation entropique :

$$W_w^\varepsilon(f, g) = \min_{x \in \Pi(f, g)} \langle w | x \rangle + \varepsilon \langle x | \log x - \mathbf{1} \rangle \quad (3)$$

Le plan de transport approché $x(\varepsilon)$ solution de ce problème s’obtient alors comme le résultat de l’algorithme de Sinkhorn appliqué à la matrice $A = \exp(-\frac{w}{\varepsilon})$ et peut donc être calculé par cette méthode très efficacement (voir [PC19]).

Lorsque le poids w régissant le problème de transport optimal est donné selon une distance $d(p_i, p_j)$ entre des points p_1, \dots, p_N de \mathbb{R}^d , la valeur $W_w(f, g)$ de (2) définit une distance sur l’espace des mesures discrètes Σ^N appelée distance de Wasserstein

(voir encore [PC19]). Cette métrique donne naissance à de nouveaux problèmes, similaires à (2). Nous étudions celui du barycentre de Wasserstein entre deux mesures $f^0, f^1 \in \Sigma^N$. Suivant [AC11], un tel barycentre est défini pour $\theta \in [0, 1]$ comme la moyenne de Fréchet selon la distance W_w :

$$\begin{aligned} f^\theta &\in \arg \min_{f \in \Sigma^N} \theta W_w(f^0, f) + (1 - \theta) W_w(f, f^1) \\ &= \arg \min_{f \in \Sigma^N} \min_{\substack{x^0 \in \Pi(f^0, f) \\ x^1 \in \Pi(f, f^1)}} \theta \langle w | x^0 \rangle + (1 - \theta) \langle w | x^1 \rangle \end{aligned}$$

En introduisant une perturbation entropique, on obtient le problème approché suivant :

$$\begin{aligned} f^\theta(\varepsilon) &= \arg \min_{f \in \Sigma^N} \theta W_w^\varepsilon(f^0, f) + (1 - \theta) W_w^\varepsilon(f, f^1) \\ &= \arg \min_{f \in \Sigma^N} \min_{\substack{x^0 \in \Pi(f^0, f) \\ x^1 \in \Pi(f, f^1)}} \theta \langle w | x^0 \rangle + (1 - \theta) \langle w | x^1 \rangle \\ &\quad + \varepsilon \theta \langle x^0 \log x^0 - \mathbf{1} \rangle + \varepsilon (1 - \theta) \langle x^1 \log x^1 - \mathbf{1} \rangle \end{aligned}$$

dont la solution peut se calculer via une variante de l'algorithme de Sinkhorn proposée dans [BCC⁺15] :

Algorithme I.2. *Étant données deux matrices $A^0, A^1 \in (\mathbb{R}_*^+)^{N \times N}$, et deux marginales $f^0, f^1 \in \Sigma^N$, depuis $\alpha^{(0)} = \beta^{(0)} = \gamma^{(0)} = \mathbf{1}$, faire pour $n = 0, 1, \dots$*

$$\alpha^{(n+1)} = \frac{f^0}{A^0 \left(\frac{1}{\gamma^{(n)}}\right)^{1-\theta}} ; \beta^{(n+1)} = \frac{f^1}{A^1 \left(\gamma^{(n)}\right)^\theta} ; \gamma^{(n+1)} = \frac{A^{0T} \alpha^{(n+1)}}{A^1 \beta^{(n+1)}}$$

Appliquer cet algorithme avec les matrices $A^0 = A^1 = A = \exp\left(-\frac{w}{\varepsilon}\right)$ permet d'obtenir le barycentre approché $f^\theta(\varepsilon)$ sous la forme $f^\theta(\varepsilon) = \gamma^\theta(A\beta)$. On peut également rechercher simultanément les points $f^{\frac{k}{K}}$ pour $k \in \llbracket 1, K-1 \rrbracket$ de la géodésique $(f^\theta)_{\theta \in [0,1]}$ en résolvant

$$\arg \min_{f^1, \dots, f^{K-1} \in \Sigma^N} \sum_{k=0}^{K-1} W_w(f^k, f^{k+1})$$

À nouveau, l'introduction d'une perturbation entropique permet de résoudre une version approchée de ce problème et d'obtenir une approximation des barycentres sous la forme $f^k(\varepsilon) = \gamma_k(A\gamma_{k+1}^{-1})$, où $A = \exp\left(-\frac{w}{\varepsilon}\right)$ et où les variables γ^k sont obtenues en prenant les matrices $A^k = A$ pour tout k dans la variante suivante de l'algorithme de Sinkhorn (décrite ici pour K pair) :

Algorithme I.3. *Étant donnés $K = 2P$ matrices $A^0, \dots, A^{K-1} \in (\mathbb{R}_*^+)^{N \times N}$ et deux marginales $f^0, f^K \in \Sigma^N$, depuis $\gamma_k^{(0)} = \mathbf{1}$ pour tout $k \in \llbracket 0, K \rrbracket$, faire pour $n = 0, 1, \dots$*

Étape A :

$$\gamma_0^{(n+1)} = \frac{f^0}{A^0 \frac{1}{\gamma_1^{(n)}}}; \quad \gamma_K^{(n+1)} = \frac{A^{K-1 T} \gamma_{K-1}^{(n)}}{f^K};$$

$$\forall k \in \llbracket 1, P-1 \rrbracket, \quad \gamma_{2k}^{(n+1)} = \sqrt{\frac{A^{2k-1 T} \gamma_{2k-1}^{(n)}}{A^{2k} \frac{1}{\gamma_{2k+1}^{(n)}}}}$$

Étape B :

$$\forall k \in \llbracket 0, P-1 \rrbracket, \quad \gamma_{2k+1}^{(n+1)} = \sqrt{\frac{A^{2k T} \gamma_{2k}^{(n+1)}}{A^{2k+1} \frac{1}{\gamma_{2k+2}^{(n+1)}}}}$$

Par sa structure fondée sur un graphe “ligne” à K points, l’algorithme I.3 est relativement proche de ceux étudiés dans les domaines du Transert de Message (ou Propagation des Croyances, en anglais “Message Passing” ou “Belief Propagation”) [Pea88] et des Champs Aléatoires Conditionnels (“Conditional Random Fields” ou “Markov Random Fields”) [WJ08]. La dernière variante du problème de transport optimal (2) que nous étudions est un problème d’affectation apparaissant notamment dans ces domaines et formulé sur un graphe (biparti) non orienté $(\mathcal{V}, \mathcal{E})$. Il consiste à attribuer à chaque sommet i de \mathcal{V} un label ℓ d’un ensemble fini L fixé (correspondant par exemple à des objets à identifier dans une image dans le contexte de Segmentation Sémantique, ou à des niveaux de gris, voir [KSS⁺20, AZJ⁺18]) de manière à résoudre le problème suivant :

$$\min_{(\ell_i) \in L^{\mathcal{V}}} \sum_{i \in \mathcal{V}} \theta_i(\ell_i) + \sum_{\{i,j\} \in \mathcal{E}} \theta_{i,j}(\ell_i, \ell_j)$$

où θ sont des fonctions codant une énergie se décomposant en un terme centré sur les sommets du graphe (θ_i pour $i \in \mathcal{V}$) et un terme centré sur les liens entre les pixels ($\theta_{i,j}$ pour $i \sim j$ c’est-à-dire tels que $\{i, j\} \in \mathcal{E}$). Après ajout d’un terme entropique dit entropie de Bethe [PA02, MJGF09], on obtient le problème suivant :

$$\min \sum_{i,\ell} \theta_i(\ell) v_i^\ell + \beta \varepsilon v_i^\ell \log v_i^\ell + \sum_{\{i,j\}, \ell, m} \theta_{i,j}(\ell, m) w_{i,j}^{\ell, m} + \varepsilon w_{i,j}^{\ell, m} \log w_{i,j}^{\ell, m}$$

où le minimum porte sur les couples $(v, w) \in \mathbb{R}^{\mathcal{V} \times L} \times \mathbb{R}^{\mathcal{E} \times L \times L}$ tels que

$$\forall i \in \mathcal{V}, \sum_{\ell} v_i^\ell = 1; \quad \forall \{i, j\} \in \mathcal{E}, \quad \forall m \in L, \sum_{\ell} w_{i,j}^{\ell, m} = v_j^m$$

Une variante de l'algorithme de Sinkhorn associée à ce problème est :

Algorithme I.4. *Étant donné un graphe non orienté biparti $(\mathcal{V} = \mathcal{V}_1 \sqcup \mathcal{V}_2, \mathcal{E})$, des matrices $X_{i,j} \in (\mathbb{R}_*^+)^{N \times N}$ pour $\{i, j\} \in \mathcal{E}$, des vecteurs $y_i \in (\mathbb{R}_*^+)^N$ et $\beta > -\min_{i \in \mathcal{V}} n_i$ où n_i désigne le nombre d'arêtes issues de $i \in \mathcal{V}$, depuis $\alpha_{i,j}^{(0)} = \mathbf{1}$ pour tout $\{i, j\} \in \mathcal{E}$, faire pour $n = 0, 1, \dots$*

Étape A :

$$\forall i \in \mathcal{V}_1, \forall j \sim i, \alpha_{j,i}^{(n+1)} = \frac{\left(y_i \prod_{j': i \sim j'} X_{i,j'} \alpha_{i,j'}^{(n)} \right)^{\frac{1}{n_i + \beta}}}{X_{i,j} \alpha_{i,j}^{(n)}}$$

Étape B :

$$\forall i \in \mathcal{V}_2, \forall j \sim i, \alpha_{j,i}^{(n+1)} = \frac{\left(y_i \prod_{j': i \sim j'} X_{i,j'} \alpha_{i,j'}^{(n+1)} \right)^{\frac{1}{n_i + \beta}}}{X_{i,j} \alpha_{i,j}^{(n+1)}}$$

La première partie de cette thèse s'intéresse plus précisément aux taux de convergence linéaires des algorithmes I.1, I.2, I.3, I.4 décrits ci-dessus. Il s'agit donc de montrer qu'il existe pour chaque algorithme une valeur $\lambda \in [0, 1)$ telle que l'erreur e_k qu'il commet après k itérations soit contrôlée sous la forme $e_k \leq c\lambda^k$ pour une constante $c > 0$. Si l'on revient à [Ide16], on constate encore la diversité des approches possibles de l'algorithme de Sinkhorn I.1 à travers les différentes preuves de convergence proposées. Alors que dès les premières démonstrations, avec celle fournie par Sinkhorn dans [Sin64] – citons aussi les taux en distance de Hilbert [FL89] que nous présenterons – des taux linéaires sont donnés, il faut attendre Knight dans [Kni08] pour que soit énoncé (dans le cas particulier $f = g = \mathbf{1}$) la valeur du taux de convergence observé en pratique. Nous fournissons une nouvelle preuve et une généralisation de ce résultat au cas f et g quelconques :

Théorème I.1. *Soient $A \in (\mathbb{R}^+)^{N \times N}$ et $f, g \in \Sigma^N$. Les itérés $\alpha^{(n)}, \beta^{(n)}$ définis par l'algorithme I.1 convergent vers $\alpha^*, \beta^* \in (\mathbb{R}_*^+)^N$ tels que la matrice associée $X = d(\alpha^*) A d(\beta^*)$ appartienne à $\Pi(f, g)$ dès que de tels vecteurs existent. De plus il existe deux suites de réels $(u_n), (v_n) \in (\mathbb{R}_*^+)^N$ convergeant vers 1, et une norme euclidienne $\|\cdot\|_s$ sur \mathbb{R}^{2N} telles que : $\forall \delta > 0, \exists n_\delta$ tel que $\forall n \geq n_\delta$,*

$$\left\| \begin{pmatrix} \alpha^{(n+1)} \\ \beta^{(n+1)} \end{pmatrix} - \begin{pmatrix} u_{n+1} \alpha^* \\ v_{n+1} \beta^* \end{pmatrix} \right\|_s \leq (\lambda_2 + \delta) \left\| \begin{pmatrix} \alpha^{(n)} \\ \beta^{(n)} \end{pmatrix} - \begin{pmatrix} u_n \alpha^* \\ v_n \beta^* \end{pmatrix} \right\|_s$$

où $\lambda_2 = \max\{|\lambda|, \lambda \in Sp(M) \text{ tel que } |\lambda| < 1\}$ est la deuxième valeur propre de $M = d(\frac{1}{g}) X^T d(\frac{1}{f}) X$.

Notre preuve se base sur l'obtention d'inégalités de convexité pour les itérations (1), et passe par le théorème suivant :

Théorème I.2. Soit $M \in (\mathbb{R}^+)^{N \times N}$ une matrice primitive stochastique symétrique pour un produit scalaire $\langle \cdot | \cdot \rangle_s$ donné par $s \in \Sigma^N$. Si $(x^n) \in ((\mathbb{R}_*^+)^N)^{\mathbb{N}}$ est une suite telle que :

$$\forall n \in \mathbb{N}, x^{n+1} \leq Mx^n \text{ et } \frac{1}{x^{n+1}} \leq M \frac{1}{x^n}$$

Alors il existe $x^* \in \mathbb{R}_*^+$ tel que $x^n \xrightarrow{n \rightarrow +\infty} x^* \mathbf{1}$ avec l'estimation suivante : pour tout $\delta > 0$ il existe $n_\delta \in \mathbb{N}$ tel que,

$$\forall n \geq n_\delta, \|x^{n+1} - \langle x^{n+1} \rangle_s \mathbf{1}\|_s \leq (\lambda_2 + \delta) \|x^n - \langle x^n \rangle_s \mathbf{1}\|_s$$

où $\lambda_2 = \max\{|\lambda|, \lambda \in \text{Sp}(M) \text{ tel que } |\lambda| < 1\}$ est la deuxième valeur propre de M .

Le cadre fourni par ce théorème nous permet alors de généraliser le résultat de convergence de l'algorithme de Sinkhorn I.1 aux variantes I.2, I.3, I.4 que nous avons décrites. Nous démontrons successivement que toutes les suites $\alpha^{(n)}, \beta^{(n)}, \gamma^{(n)}, \gamma_k^{(n)}$ introduites convergent vers des points fixes des itérations $\alpha^*, \beta^*, \gamma^*, \gamma_k^*$ (qui fournissent les solutions des problèmes initiaux), et qu'il existe des suites $(u_n), (v_n)$ convergeant vers 1, des normes euclidiennes $\|\cdot\|_s$ et des valeurs $\lambda_2 < 1$ apparaissant comme deuxièmes valeurs propres de matrices bistochastiques symétriques fournissant les estimations suivantes : $\forall \delta > 0, \exists n_\delta \in \mathbb{N}$ tel que $\forall n \geq n_\delta$,

- Pour l'algorithme I.2 du barycentre simple :

$$\left\| \gamma^{(n+1)} - u_{n+1} \gamma^* \right\|_s \leq (\lambda_2 + \delta) \left\| \gamma^{(n)} - u_n \gamma^* \right\|_s$$

- Pour l'algorithme I.3 des barycentres multiples :

$$\left\| \gamma_{[1]}^{(n+1)} - u_{n+1} \gamma_{[1]}^* \right\|_s \leq (\lambda_2 + \delta) \left\| \gamma_{[1]}^{(n)} - u_n \gamma_{[1]}^* \right\|_s$$

- Pour l'algorithme I.4 sur un graphe "ligne" ou "cercle" (c'est-à-dire d'arité maximale 2) et avec $\beta = 0$:

$$\left\| \begin{pmatrix} \gamma_{[0]}^{(n+1)} \\ \gamma_{[1]}^{(n+1)} \end{pmatrix} - \begin{pmatrix} u_{n+1} \gamma_{[0]}^* \\ v_{n+1} \gamma_{[1]}^* \end{pmatrix} \right\|_s \leq (\lambda_2 + \delta) \left\| \begin{pmatrix} \gamma_{[0]}^{(n)} \\ \gamma_{[1]}^{(n)} \end{pmatrix} - \begin{pmatrix} u_n \gamma_{[0]}^* \\ v_n \gamma_{[1]}^* \end{pmatrix} \right\|_s$$

où $\gamma_{[0]}$ (respectivement $\gamma_{[1]}$) désigne la concaténation des variables γ_k pour k pair (respectivement impair).

Si pour certaines applications l'ajout du terme entropique $\varepsilon \langle x | \log x - \mathbf{1} \rangle$ avec ε relativement grand permet de régulariser les données en jeu (voir [BCP19]), il est généralement d'usage de chercher à faire tendre ε vers 0 pour approcher au mieux la solution exacte des problèmes considérés. Cominetti et San Martin montrent dans [CSM94] que le problème perturbé (3) converge vers le problème exact (2) en $O(\exp(-\frac{c}{\varepsilon}))$ pour une constante $c > 0$. Ce résultat positif est contrebalancé par le fait que les matrices sur lesquelles l'algorithme de Sinkhorn et ses variantes sont appliqués s'obtiennent sous la

forme $A = \exp(-\frac{w}{\varepsilon})$, particulièrement instable numériquement quand ε tend vers 0. Une stratégie permettant d'accéder à de plus petites valeurs de ε consiste à travailler avec des variables logarithmiques (voir [Sch19, SGG11]). Nous proposons également d'utiliser l'entropie de Bethe pour atténuer la régularisation entropique et explicitons cette technique dans le cas du barycentre de deux mesures. Nous présentons enfin une méthode itérative, proposée dans un article, [XWWZ18], alors que ce travail de thèse était en cours, consistant à recentrer le terme de perturbation entropique sur l'itéré précédent en posant $x_\varepsilon^1 = x(\varepsilon)$ puis pour tout $n \geq 1$,

$$x_\varepsilon^{n+1} = \arg \min_{x \in \Pi(f,g)} \langle w|x \rangle + \varepsilon \langle x | \log \frac{x}{x_\varepsilon^n} \rangle$$

Nous démontrons qu'appliquer n pas de cette méthode correspond à diviser le paramètre ε par n : $x_\varepsilon^n = x(\frac{\varepsilon}{n})$, et proposons d'autres variantes.

Enfin, même si les taux de convergence “en λ_2 ” des algorithmes de Sinkhorn que nous démontrons sont effectivement ceux observés en pratique, ils ne sont malheureusement pas directement accessibles avant d'avoir effectué l'algorithme (ou d'en connaître la valeur limite); notre étude reste donc à ce titre relativement théorique. Des études plus récentes de l'algorithme de Sinkhorn [KLR08, CK18, DGK18] semblent d'ailleurs se détourner d'une description du taux de convergence linéaire effectif pour obtenir des garanties de convergence *a priori* dépendant de la matrice A , des marginales f, g ou de la dimension N . Ainsi Altschuler et ses co-auteurs démontrent dans [ANWR17] le théorème suivant :

Théorème I.3. *L'algorithme I.1 fournit des valeurs $\alpha^{(n)}$ et $\beta^{(n)}$ telles que la matrice $X^{(n)} = d(\alpha^{(n)})A d(\beta^{(n)})$ satisfasse*

$$\|X^{(n)}\mathbf{1} - f\|_{\ell^1} + \|X^{(n)T}\mathbf{1} - g\|_{\ell^1} \leq \eta$$

après un nombre d'itérations $n = 4\eta^{-2} \log \frac{S}{m}$ où $S = \sum_{i,j} A_{i,j}$ et $m = \min_{i,j} A_{i,j}$.

Si sa dépendance en η^{-2} peut sembler sous-optimale au regard de la convergence linéaire observée en pratique, il est remarquable que ce taux ne dépende de la dimension N qu'à travers la somme des coefficients de A . Dans le but d'approcher le problème de transport optimal exact (2) par sa version perturbée (3), il est intéressant de savoir comment le taux de convergence de l'algorithme de Sinkhorn évolue pour une matrice A donnée par $A = \exp(-\frac{w}{\varepsilon})$ lorsque ε tend vers 0. Le théorème précédent fournit donc un taux de convergence en $O(\eta^{-2}\varepsilon^{-1}(\bar{w} - \underline{w}) \log N)$ où $\bar{w} = \max_{i,j} w_{i,j}$ et $\underline{w} = \min_{i,j} w_{i,j}$.

La comparaison avec le taux linéaire que nous proposons au théorème I.1 semble difficile à mener en toute généralité. Nous traitons uniquement le cas de la dimension 2 où des calculs explicites peuvent être menés, démontrant pour ε tendant vers 0 la convergence de λ_2 vers une valeur limite en $O(\exp(-\frac{c}{\varepsilon}))$ pour une constante $c > 0$.

Cette partie s'organise comme suit. Dans le premier chapitre, nous introduisons l'algorithme de Sinkhorn en présentant un exemple historique concret issu de [Kru37] et en détaillant son apparition dans le cadre du transport optimal. Le chapitre 2 est consacré à l'étude de sa convergence linéaire et démontre notamment les théorèmes I.1 et I.2. Le chapitre 3 décrit et étudie les variantes I.2, I.3, I.4 de l'algorithme de Sinkhorn. Enfin, le chapitre 4 s'intéresse à l'évolution du problème perturbé (3) quand le paramètre ε tend vers 0 et décrit les stratégies envisagées pour surmonter les difficultés engendrées ; il se termine par l'étude complète du cas de la dimension 2.

Le travail présenté dans cette partie fait l'objet d'un article actuellement en préparation.

EMERGENCE OF THE SINKHORN ALGORITHM

1.1 Historical example of Telephone Forecasting

In his review [Ide16], Idel emphasizes the fact that tracing back the Sinkhorn algorithm to its origins is a difficult task. Following his research, one can say that one of the first mentions of the algorithm dates back to the 30's in a telephone traffic Dutch article by Kruithof [Kru37]. Relying on the translation of this untraceable Dutch-written article made by de Boer [dB], we can give an interpretation of its motivations for introducing the Sinkhorn algorithm in the context of Telephone Traffic Forecasting (under the name “method of double factors”).

Model the telephone traffic as a collection of exchanges between originating nodes $i \in \llbracket 1, N \rrbracket$ and terminating nodes $j \in \llbracket 1, M \rrbracket$. At a given time, say the year 1937, the situation of the telephone network is known, that is we have the number $a_{i,j}$ of calls from node i to node j for all $(i, j) \in \llbracket 1, N \rrbracket \times \llbracket 1, M \rrbracket$. For the following year, experts on the extension of the telephone network predict that the total of calls emanating from node i will move from its current value $\sum_j a_{i,j}$ to a certain value f_i . They also predict that the total of calls terminating to node j , $\sum_i a_{i,j}$ will change to g_j . What prediction of the telephone traffic for the year 1938 can we make? and how should we modify the infrastructures of the originating and terminating nodes to support this evolution? Motivated by this second question, one can try to predict the new number $b_{i,j}$ of calls from i to j under the form $b_{i,j} = \alpha_i a_{i,j} \beta_j$. The factors α_i and β_j will then be interpreted as growth factors for nodes i and j . To fulfill the predictions of the

experts, one would want for $b_{i,j}$ to satisfy the equations

$$\begin{aligned} \forall i \in \llbracket 1, N \rrbracket, \sum_{j=1}^M b_{i,j} &= \alpha_i \sum_{j=1}^M a_{i,j} \beta_j = f_i \\ \forall j \in \llbracket 1, M \rrbracket, \sum_{i=1}^N b_{i,j} &= \beta_j \sum_{i=1}^N a_{i,j} \alpha_i = g_j \end{aligned}$$

Several procedures could be considered to find – if they exist – such α_i and β_j . The one we are interested in, and that is probably the most natural, is the following. First set $\alpha_i = \beta_j = 1$ so that our first guess for $b_{i,j}$ simply is to make no change to the current situation $b_{i,j} = a_{i,j}$. Then chose α_i such that the first equations are fulfilled, namely

$$\forall i \in \llbracket 1, N \rrbracket, \alpha_i = \frac{f_i}{\sum_{j=1}^M a_{i,j} \beta_j} \quad (1.1)$$

Then $b_{i,j}$ is such that the total number of emanating calls from i is coherent with the prediction. But the number of terminating calls to j is not, so we now set β_j so that the second equations are fulfilled, namely

$$\forall j \in \llbracket 1, M \rrbracket, \beta_j = \frac{g_j}{\sum_{i=1}^N a_{i,j} \alpha_i} \quad (1.2)$$

But then of course the first equations are no longer satisfied for this value of β , so one repeats step (1.1), and then step (1.2) and so on. If this procedure converges, we would have found in α_i , β_j and $b_{i,j} = \alpha_i a_{i,j} \beta_j$ a solution to our problem. Finally, rephrasing it with matrix-vector multiplications and componentwise divisions, we obtain the Sinkhorn algorithm that is under study:

Algorithm I.1. *Given a nonnegative matrix $A \in (\mathbb{R}^+)^{N \times M}$ with no zero line or column, and given two marginals $f \in (\mathbb{R}^+)^N$, $g \in (\mathbb{R}^+)^M$, starting from $\forall i, j \alpha_i^{(0)} = \beta_j^{(0)} = 1$, do for $n = 0, 1, \dots$*

$$\alpha^{(n+1)} = \frac{f}{A\beta^{(n)}}; \beta^{(n+1)} = \frac{g}{A^T \alpha^{(n+1)}} \quad (1.3)$$

One can first make an obvious remark: for this procedure to converge, one must have $\sum_i f_i = \sum_j g_j$ – that is the predictions of the experts on originating and terminating calls give the same total amount of communications. Without loss of generality, we can suppose f and g belong to the simplex $\Sigma^N = \{p \in (\mathbb{R}_*^+)^N, \sum_k p_k = 1\}$. Note that later on we will focus on the case $M = N$ for simplicity of notations. However, all the presented results (with exception of the doubly stochastic case for which $M = N$ is mandatory) adapt with no modification to the setting $M \neq N$.

This telephone forecasting problem we described is of course a modest insight of the large area where the Sinkhorn algorithm, and more broadly questions of matrix scaling, appear. Matrix scaling problems are problems where one wants to obtain a new

matrix satisfying some property by multiplying the lines and columns of a given matrix by some quantities, or in other words by multiplying the given matrix by diagonal matrices on the right and on the left hand sides. The resulting matrix is called a *scaling* of the original matrix. From that point of view, the matrix scaling problem related to the Sinkhorn algorithm is the following:

Problem I.1. *Given a nonnegative matrix $A \in (\mathbb{R}^+)^{N \times N}$ and $f, g \in \Sigma^N$, can one find diagonal matrices D_1 and D_2 with positive diagonal entries such that $B = D_1 A D_2$ has f and g as marginals, meaning that $B \mathbf{1} = f$ and $B^T \mathbf{1} = g$?*

Such issues of matrix scaling can arise in many practical situations; in [BR97], Bapat and Raghavan mention for instance budget allocation or scaling contingency tables problems (which are problems similar to our telephone forecasting example), scaling in Gaussian elimination (which consists of applying an appropriate scaling improving condition numbers before solving a linear system), and Markov chain issues (estimating the transition matrix from observations of the states). These issues also admit generalizations to continuous settings and to positive maps, see for instance [Rus95] and [Gur03]. Furthermore, they arise naturally in the analysis of transportation problems through entropic regularization.

1.2 Appearance in Optimal Transportation

In the 2000's, a renewed interest for the Sinkhorn algorithm aroused from the Optimal Transport community. Especially with the success of [Cut13], it became a popular way of computing the optimal transport cost $W_w(f, g)$ between two discrete measures $f, g \in \Sigma^N$ with respect to a ground cost $w \in \mathbb{R}^{N \times N}$. In finite dimension, Optimal Transport deals with solving the following linear problem (see for instance [PC19] for a large background on the subject):

$$W_w(f, g) = \min_{x \in \Pi(f, g)} \langle w | x \rangle := \sum_{i, j=1}^N w_{i, j} x_{i, j} \quad (1.4)$$

where $\Pi(f, g) = \{x \in (\mathbb{R}^+)^{N \times N}, x \mathbf{1} = f, x^T \mathbf{1} = g\}$ is the admissible transport plans polytope.

Under some hypothesis on the ground cost w (typically $w_{i, j} = \|p_i - p_j\|_2^2$ for some fixed points $p_1, \dots, p_N \in \mathbb{R}^d$), the (square root of the) optimal transport cost $W_w(f, g)$ defines a distance on Σ^N known as the Wasserstein distance. This distance enjoys nice properties (such as taking into account the geometry of the underlying space) which make it a valuable tool for comparing measures, see again [PC19]. However, as the transport plan variable x has dimension N^2 , classical linear solvers fail to tackle this problem when N becomes too large. Introducing an entropic regularization to the linear problem (1.4) will make the dimension of the variable at stake drop from N^2 to $2N$. By strong convexity of the regularization, this interior point method also ensures uniqueness of the minimizer (which we did not have in (1.4)). The perturbed problem

under study is, for a (small) $\varepsilon > 0$:

$$W_w^\varepsilon(f, g) = \min_{x \in \Pi(f, g)} \langle w | x \rangle + \varepsilon \langle x | \log x - \mathbf{1} \rangle := \sum_{i,j=1}^N w_{i,j} x_{i,j} + \varepsilon x_{i,j} (\log(x_{i,j}) - 1) \quad (1.5)$$

To see the link with the matrix scaling problem I.1 and the Sinkhorn algorithm, one has to focus on the corresponding dual problem in the sense of convex duality. We first associate Lagrange multipliers $\lambda, \mu \in \mathbb{R}^N$ respectively to the equality constraints $x\mathbf{1} = f$ and $x^T\mathbf{1} = g$ to form the Lagrangian

$$\mathcal{L}(x, \lambda, \mu) := \langle w, x \rangle + \varepsilon \langle x | \log x - \mathbf{1} \rangle + \langle \lambda | f - x\mathbf{1} \rangle + \langle \mu | g - x^T\mathbf{1} \rangle$$

Then we need the following well-known calculus of the Legendre-Fenchel conjugate of the entropy. Together with its variant Proposition I.5 below, we will extensively exploit this relation in the following sections.

Proposition I.1. *For any $a \in \mathbb{R}^{N \times N}$ and $\eta > 0$ one has*

$$\min_{x \in (\mathbb{R}^+)^{N \times N}} \langle x | a \rangle + \eta \langle x | \log x - \mathbf{1} \rangle = -\eta \langle \exp\left(\frac{-a}{\eta}\right) | \mathbf{1} \rangle$$

Proof. The optimality condition of this convex problem writes $a + \eta \log x = 0$ so that the optimizer is $x = \exp\left(\frac{-a}{\eta}\right)$ which gives the announced value. \square

Using this proposition to calculate the value of $F(\lambda, \mu) = \min_x \mathcal{L}(x, \lambda, \mu)$ for any λ, μ , we form the dual problem $\max_{\lambda, \mu} F(\lambda, \mu)$ of (1.5). One concludes that the minimization problem (1.5) is equivalent to the following maximization problem:

$$\max_{\lambda, \mu \in \mathbb{R}^N} \langle \lambda | f \rangle + \langle \mu | g \rangle - \varepsilon \sum_{i,j} \exp\left(\frac{-w_{i,j} + \lambda_i + \mu_j}{\varepsilon}\right) \quad (1.6)$$

It is noteworthy to mention that this dual problem is *unconstrained*. This is a crucial advantage when compared to the dual of the original unregularized optimal transport problem (1.4) which writes as:

$$\max_{\substack{\lambda, \mu \in \mathbb{R}^N \\ \forall i,j, \lambda_i + \mu_j \leq w_{i,j}}} \langle \lambda | f \rangle + \langle \mu | g \rangle \quad (1.7)$$

While enjoying the same reduction of the number of variables from N^2 to $2N$, problem (1.6) alleviates through an exponential penalty the N^2 constraints of (1.7) that make it difficult to tackle. Optimality conditions $\nabla F(\lambda, \mu) = 0$ for problem (1.6) can be expressed using the variables

$$\alpha = \exp\left(\frac{\lambda}{\varepsilon}\right); \quad \beta = \exp\left(\frac{\mu}{\varepsilon}\right); \quad A = \exp\left(\frac{-w}{\varepsilon}\right)$$

and one recovers that λ, μ are optimal values in problem (1.6) if and only if

$$\alpha = \frac{f}{A\beta}; \beta = \frac{g}{A^T\alpha} \quad (1.8)$$

Hence one can apply the Sinkhorn algorithm to the matrix $A = \exp(\frac{-w}{\varepsilon})$ to find a solution of the dual problem. The primal solution is then recovered through $x_{i,j} = \exp\left(\frac{\lambda_i + \mu_j - w_{i,j}}{\varepsilon}\right)$ as $X = d(\alpha)A d(\beta) \in \Pi(f, g)$ is indeed the transport plan formed by this procedure.

Remark I.1. *Depending on the context, some forms of the entropy term may appear more natural than others. One can for instance replace $\langle x | \log x - \mathbf{1} \rangle$ by $\langle x | \log x \rangle$ or $\langle x | \log \frac{x}{f \otimes g} \rangle$. One can check that this simply leads to multiplying the matrix $A = \exp(-\frac{w}{\varepsilon})$ by a constant or by diagonal matrices. As a consequence this has no effect on the resulting Sinkhorn algorithm as we will see below (however one should keep in mind that the value of the optimization problem is modified).*

In this context, the Sinkhorn algorithm can be interpreted in several ways. In terms of the variables λ and μ , the iterates (1.3) correspond to

$$\begin{aligned} \lambda_i^{(n+1)} &= \varepsilon \log(f_i) - \varepsilon \log \left(\sum_{j=1}^N \exp \left(\frac{\mu_j^{(n)} - w_{i,j}}{\varepsilon} \right) \right) \\ \mu_j^{(n+1)} &= \varepsilon \log(g_j) - \varepsilon \log \left(\sum_{i=1}^N \exp \left(\frac{\lambda_i^{(n+1)} - w_{i,j}}{\varepsilon} \right) \right) \end{aligned} \quad (1.9)$$

which can be seen as alternate maximizations on the dual (1.6). Another interesting point of view is provided by Benamou and co-authors in [BCC⁺15]. They remark that the corresponding primal iterates of this process actually achieve alternate projections with respect to the so-called Kullback-Leiber divergence onto the constraints $x\mathbf{1} = f$ and $x^T\mathbf{1} = g$ as follows. For any positive matrices $x, y > 0$ define

$$\mathcal{KL}(x, y) = \langle x | \log \frac{x}{y} - \mathbf{1} \rangle := \sum_{i,j=1}^N x_{i,j} (\log \frac{x_{i,j}}{y_{i,j}} - 1)$$

then the primal formulation (1.5) writes as

$$x^* = \arg \min_{x \in \Pi(f, g)} \mathcal{KL}(x, \exp\left(-\frac{w}{\varepsilon}\right))$$

Interpreting \mathcal{KL} as a distance on $(\mathbb{R}_*^+)^{N \times N}$ (or more precisely as a Bregman distance, see Definition I.5), this corresponds to finding the projection of $x^{(0)} = \exp(-\frac{w}{\varepsilon})$ onto the intersection $\Pi(f, g) = \Pi(f) \cap \Pi^T(g)$ where $\Pi(f) = \{x \in (\mathbb{R}^+)^{N \times N}, \text{ s.t. } x\mathbf{1} = f\}$ and $\Pi^T(g) = \{x \in (\mathbb{R}^+)^{N \times N}, \text{ s.t. } x^T\mathbf{1} = g\}$. A natural strategy, following for instance Karczmarz's method for solving linear systems [Kar37], is to perform

alternate projections on $\Pi(f)$ and $\Pi^T(g)$ defining for all $n \geq 0$:

$$\begin{aligned} x^{(n+\frac{1}{2})} &= \arg \min_{x \in \Pi(f)} \mathcal{KL}(x, x^{(n)}) \\ x^{(n+1)} &= \arg \min_{x \in \Pi^T(g)} \mathcal{KL}(x, x^{(n+\frac{1}{2})}) \end{aligned}$$

One easily computes that this corresponds to setting

$$\begin{aligned} x^{(n+\frac{1}{2})} &= d\left(\frac{f}{x^{(n)}\mathbf{1}}\right)x^{(n)} \\ x^{(n+1)} &= x^{(n+\frac{1}{2})} d\left(\frac{g}{x^{(n+\frac{1}{2})T}\mathbf{1}}\right) \end{aligned} \tag{1.10}$$

so that the Sinkhorn algorithm is recovered noticing that $x^{(n)} = d(\alpha^{(n)})A d(\beta^{(n)})$ and $x^{(n+\frac{1}{2})} = d(\alpha^{(n+\frac{1}{2})})A d(\beta^{(n+\frac{1}{2})})$. In addition, this point of view ensures convergence of such a process thanks to a general theorem on iterated projections on affine sets proved in [Bre67] (but we also refer to [BL00], where this is proved in a more general setting).

The historical setting of the Sinkhorn algorithm deals with an interesting special case: the doubly-stochastic setting. It corresponds to taking $f = g = \mathbf{1}$ (or $f = g = \frac{1}{N}\mathbf{1}$ if one wants to keep $f, g \in \Sigma^N$; in the following we do not use this no effect renormalization). In that setting, the polytope $\Pi(\mathbf{1}, \mathbf{1}) =: \mathbf{B}_N$ of doubly stochastic matrices has a simple structure. In particular, its extreme points correspond exactly to the permutation matrices so that the optimal transport problem (1.4) in that setting:

$$\arg \min_{x \in \mathbf{B}_N} \sum_{i,j=1}^N w_{i,j} x_{i,j} \tag{1.11}$$

appears as the convex relaxation of the so called *assignment problem*:

$$\arg \min_{\sigma \in \mathbf{S}_N} \sum_{i=1}^N w_{i,\sigma(i)} \tag{1.12}$$

where \mathbf{S}_N denotes the set of permutations of $\{1, \dots, N\}$. This famous problem (see [BDM09] for more background on this subject) aims at obtaining an optimal allocation when distributing some tasks with costs $w_{i,j}$ to agents: one wants to find a way of assigning each task $i \in \{1, \dots, N\}$ to an agent $j \in \{1, \dots, N\}$ through a one to one correspondence that minimizes the global cost of the operation. Classical solvers for this problem include the celebrated Hungarian (or auction) algorithm, first introduced by Kuhn in [Kuh55], and that can be extended to the general optimal transport setting (see for instance [BC89, BE88, Wal17]). The equality of problems (1.11) and (1.12) is the classical equivalence between Monge and Kantorovitch formulations of the Optimal Transportation problem.

The context of Optimal Transport gives us a first result on the matrix scaling problem I.1. If one can write the matrix A as $A = \exp(-\frac{w}{\varepsilon})$ for some $w \in \mathbb{R}^{N \times N}$, in other words if A is positive, then the existence and uniqueness of a minimizer of problem (1.5), is a proof to the existence and uniqueness of the matrix $B = D_1 A D_2$ such that $B\mathbf{1} = f$ and $B^T \mathbf{1} = g$. Furthermore, the strict concavity with respect to $\lambda_i + \mu_j$ of the objective in the dual problem (1.6) is a proof that the scaling diagonal matrices are in that case unique up to a scalar multiple: if (D_1, D_2) and (D'_1, D'_2) are two couples of diagonal matrices such that $B = D_1 A D_2 = D'_1 A D'_2$ then $D'_1 = r D_1$ and $D'_2 = \frac{1}{r} D_2$ for some $r > 0$.

However, this existence and uniqueness result does not prove (linear) convergence of the Sinkhorn algorithm; and one may ask what happens if A is no longer positive but only nonnegative. We answer these questions in the following chapter.

**LINEAR CONVERGENCE OF THE CLASSICAL
SINKHORN ALGORITHM**

2.1 First linear rates for positive matrices

2.1.1 Min-max analysis

When the matrix A is positive, Sinkhorn proved convergence of the algorithm, first for constant marginals $f = g = \mathbf{1}$ in his first paper on the subject [Sin64], then in the general case in [Sin67]. His proofs rely on the monotonicity of the minimum and maximum values of the intermediate matrices – the matrices $d(\alpha^{(n)})A d(\beta^{(n)})$ and $d(\alpha^{(n+1)})A d(\beta^{(n)})$. As such they are very close to the first convergence proof we give below. We wish to reproduce this proof here for three reasons: first, it is a simple and straightforward proof often omitted to our knowledge in the literature; second, it introduces the rescaled variables we will use in our convergence rate analysis; third, like the historical proof in [Sin67], it gives a first linear convergence rate that we will refine later on.

First, knowing that a solution to the scaling problem I.1 exists, one can reformulate the Sinkhorn algorithm in terms of variables rescaled by this solution.

Lemma I.1. *Let $(\alpha^{(n)}, \beta^{(n)})$ be the Sinkhorn iterates defined by (1.3), denote by α^*, β^* a solution of the scaling problem I.1 and by $X = d(\alpha^*)A d(\beta^*) \in \Pi(f, g)$ the corresponding scaling, then the rescaled variables*

$$a^{(n)} = \frac{\alpha^{(n)}}{\alpha^*} ; b^{(n)} = \frac{\beta^{(n)}}{\beta^*}$$

satisfy the following iteration rule:

$$a^{(n+1)} = \frac{1}{d(\frac{1}{f})Xb^{(n)}}; \quad b^{(n+1)} = \frac{1}{d(\frac{1}{g})X^T a^{(n+1)}} \quad (2.1)$$

Proof. Just compute

$$a_i^{(n+1)} = \frac{1}{\alpha_i^* \sum_j A_{i,j} \beta_j^{(n)}} = \frac{1}{\sum_j \frac{1}{f_i} \alpha_i^* A_{i,j} \beta_j^{(n)}} = \frac{1}{(d(\frac{1}{f})Xb^{(n)})_i}$$

and similarly for $b^{(n+1)}$. □

This rescaling process forces the appearance of the stochastic matrices $d(\frac{1}{f})X$ and $d(\frac{1}{g})X^T$ that make our estimates easier to understand. We now make use of the following lemma (appearing also in [BL18]):

Lemma I.2. *Let $M \in \mathbb{R}_{+*}^{N \times N}$ be a stochastic matrix: $M\mathbf{1} = \mathbf{1}$, and let $m = \underline{M}$. Then for any $x \in \mathbb{R}^N$,*

$$\underline{x} + m(\bar{x} - \underline{x}) \leq \underline{Mx} \leq \overline{Mx} \leq \bar{x} - m(\bar{x} - \underline{x})$$

Proof. Write $Mx = \underline{x}\mathbf{1} + M(x - \underline{x}\mathbf{1})$. For any i and any j one has

$$(M(x - \underline{x}\mathbf{1}))_i = \sum_k m_{i,k}(x_k - \underline{x}) \geq m(x_j - \underline{x})$$

The first inequality comes from taking j such that $\bar{x} = x_j$. Similarly, writing that $Mx = \bar{x}\mathbf{1} - M(\bar{x}\mathbf{1} - x)$ and

$$(M(\bar{x}\mathbf{1} - x))_i = \sum_k m_{i,k}(\bar{x} - x_k) \geq m(\bar{x} - x_j)$$

for j such that $\underline{x} = x_j$ gives the second inequality. □

We are now able to state our first linear convergence rate of the classical Sinkhorn algorithm:

Theorem I.1. *When A is a positive matrix, the rescaled variables $a^{(n)}$ and $b^{(n)}$ converge to constant vectors $a^\infty\mathbf{1}$ and $b^\infty\mathbf{1}$ and one has $\forall n \geq 0$,*

$$\begin{aligned} \bar{b}^{(n+1)} - \underline{b}^{(n+1)} &\leq (1 - 2\underline{x}_a) (\bar{b}^{(n)} - \underline{b}^{(n)}) \\ \bar{a}^{(n+1)} - \underline{a}^{(n+1)} &\leq (1 - 2\underline{x}_b) (\bar{a}^{(n)} - \underline{a}^{(n)}) \end{aligned}$$

where $\underline{x}_b = \underline{d(\frac{1}{g})X^T}$ and $\underline{x}_a = \underline{d(\frac{1}{f})X}$.

Proof. On one hand, from $\frac{1}{a^{(n+1)}} = d(\frac{1}{f})Xb^{(n)}$ we derive thanks to our lemma:

$$\underline{b}^{(n)} + \underline{x}_a(\bar{b}^{(n)} - \underline{b}^{(n)}) \leq \frac{1}{\underline{a}^{(n+1)}} \leq \frac{1}{\bar{a}^{(n+1)}} \leq \bar{b}^{(n)} - \underline{x}_a(\bar{b}^{(n)} - \underline{b}^{(n)})$$

On the other hand, from $b^{(n+1)} = \frac{1}{d(\frac{1}{g})X^T a^{(n+1)}}$ we get, thanks to a weaker version of our lemma:

$$\frac{1}{\bar{a}^{(n+1)}} \leq \underline{b}^{(n+1)} \leq \bar{b}^{(n+1)} \leq \frac{1}{\underline{a}^{(n+1)}}$$

Combining these two inequalities gives

$$\underline{b}^{(n)} + \underline{x}_a(\bar{b}^{(n)} - \underline{b}^{(n)}) \leq \underline{b}^{(n+1)} \leq \bar{b}^{(n+1)} \leq \bar{b}^{(n)} - \underline{x}_a(\bar{b}^{(n)} - \underline{b}^{(n)})$$

which first shows that $(\bar{b}^{(n)})$ is nonincreasing and bounded below, while $(\underline{b}^{(n)})$ is non-decreasing and bounded above so these sequences converge. Next, it also shows that the announced inequality is true. As, for $N \geq 2$, one has $\underline{x}_a \in (0, \frac{1}{2}]$ because the matrix $d(\frac{1}{f})X$ is stochastic, this proves that $(\bar{b}^{(n)})$ and $(\underline{b}^{(n)})$ have the same limit $b^\infty \in \mathbb{R}$ and consequently $b^{(n)} \rightarrow b^\infty \mathbf{1}$. The proof for $a^{(n)}$ is similar. \square

Note that the limit values a^∞, b^∞ depend on the choice of the first vectors $a^{(0)}, b^{(0)}$ that depend themselves on the choice of the solution α^*, β^* used to rescale the variables $\alpha^{(n)}, \beta^{(n)}$. Of course picking another solution $r\alpha^*, \frac{1}{r}\beta^*$ would lead to the limits $\frac{a^\infty}{r}$ and rb^∞ . At this point, we also notice that the convergence of the Sinkhorn algorithm is also guaranteed for any initial vectors $\alpha^{(0)}, \beta^{(0)}$ (that is, not just for $\alpha^{(0)} = \beta^{(0)} = \mathbf{1}$).

2.1.2 Hilbert distance analysis

The fact that multiplying all vectors by positive constants has no effect on convergence, as well as the componentwise divisions that are characteristic to the Sinkhorn algorithm, make the Hilbert distance very appropriate for its study. This projective metric is defined on the quotient space $(\mathbb{R}_*^+)^N / \sim$ where for any $x, y \in (\mathbb{R}_*^+)^N$ we have $x \sim y$ if and only if $y = rx$ for some $r > 0$. We add some straightforward results to the definition below:

Definition-Proposition I.1. (Hilbert distance) *The function d_H given by*

$$\forall x, y \in (\mathbb{R}_*^+)^N, d_H(x, y) = \log \max_{1 \leq i, j \leq N} \frac{x_i y_j}{x_j y_i}$$

defines a distance on $(\mathbb{R}_^+)^N / \sim$ that is such that:*

1. $\forall f, u, v \in (\mathbb{R}_*^+)^N, d_H(fu, fv) = d_H(u, v)$
2. $\forall f, u, v \in (\mathbb{R}_*^+)^N, d_H(\frac{f}{u}, \frac{f}{v}) = d_H(u, v)$
3. $\forall u, v \in (\mathbb{R}_*^+)^N, \forall \theta \in \mathbb{R}, d_H(u^\theta, v^\theta) = |\theta|d_H(u, v)$
4. $\forall a, b, c, d \in (\mathbb{R}_*^+)^N, d_H(ab, cd) \leq d_H(a, c) + d_H(b, d)$

Also note that $d_H(x, y) = \log \bar{z} - \log \underline{z}$ where $z = \frac{x}{y}$, so that the analysis of Sinkhorn's iterations in Hilbert distance is really close to the analysis of Theorem I.1. We refer to [BR97] for more background on such distances that can be defined on any pointed convex cone similar to $(\mathbb{R}_*^+)^N$ (precisely, K is a pointed convex cone when $x, y \in K$, $\lambda, \mu \in \mathbb{R}^+ \Rightarrow \lambda x + \mu y \in K$ and $x, y \in K$, $x + y = 0 \Rightarrow x = y = 0$). The fundamental result we need to prove convergence of the Sinkhorn algorithm is originally due to Birkhoff [Bir57] (but we refer to [BR97] and [Car04] for the proof¹). It states that the linear operator A – which is indeed defined from $(\mathbb{R}_*^+)^N / \sim$ to $(\mathbb{R}_*^+)^N / \sim$ as A is positive – is a contraction in the Hilbert metric:

Lemma I.3. *When A is a positive matrix, one has*

$$\forall x, y \in (\mathbb{R}_*^+)^N, d_H(Ax, Ay) \leq \kappa(A)d_H(x, y)$$

$$\text{where } \theta(A) = \max_{1 \leq i, j, k, l \leq N} \frac{a_{i,k}a_{j,l}}{a_{j,k}a_{i,l}} \text{ and } \kappa(A) = \frac{\sqrt{\theta(A)} - 1}{\sqrt{\theta(A)} + 1}.$$

Furthermore, this constant is optimal in the sense that $\kappa(A) = \sup_{\substack{x, y \in (\mathbb{R}_*^+)^N \\ x \not\sim y}} \frac{d_H(Ax, Ay)}{d_H(x, y)}$.

Note that $\kappa(A) = \kappa(A^T)$. As first noticed by Franklin and Lorenz in [FL89], this result directly shows linear convergence of the Sinkhorn algorithm in the Hilbert metric:

Theorem I.2. *When A is a positive matrix, the Sinkhorn iterates $\alpha^{(n)}, \beta^{(n)}$ converge in $(\mathbb{R}_*^+)^N / \sim$ to the fixed point α^*, β^* with,*

$$\forall n \geq 0, \begin{cases} d_H(\alpha^{(n+1)}, \alpha^*) \leq \kappa(A)^2 d_H(\alpha^{(n)}, \alpha^*) \\ d_H(\beta^{(n+1)}, \beta^*) \leq \kappa(A)^2 d_H(\beta^{(n)}, \beta^*) \end{cases}$$

Proof. Make use of point 1 of Definition-Proposition I.1 and Lemma I.3 to get

$$\begin{aligned} d_H(\beta^{(n+1)}, \beta^*) &= d_H\left(\frac{g}{A^T \alpha^{(n+1)}}, \frac{g}{A^T \alpha^*}\right) = d_H(A^T \alpha^{(n+1)}, A^T \alpha^*) \\ &\leq \kappa(A) d_H(\alpha^{(n+1)}, \alpha^*) \\ &= \kappa(A) d_H\left(\frac{f}{A \beta^{(n)}}, \frac{f}{A \beta^*}\right) = \kappa(A) d_H(A \beta^{(n)}, A \beta^*) \\ &\leq \kappa(A)^2 d_H(\beta^{(n)}, \beta^*) \end{aligned}$$

and similarly for $\alpha^{(n)}$. □

Note that this linear convergence is only expressed in the quotient space $(\mathbb{R}_*^+)^N / \sim$. One would need additional arguments – typically, the ones we gave in Theorem I.1 – to ensure that the actual Sinkhorn iterates do not crash to 0, become unbounded or even oscillate between multiples of the solution.

¹Historical proofs and [BR97] only show the inequality given below. [Car04] gives the remark on the optimality of the constant $\kappa(A)$.

Both Theorems I.1 and I.2 express a linear convergence rate of the Sinkhorn iterates which is indeed what we observe in practice. However, the proposed rates in $1 - 2\underline{x}$ and $\kappa(A)^2$ are pessimistic as one can witness from Figure 2.1. In this numerical experiment, we compare these rates to the actual error evolution (represented here by $\|\beta^{(n)} - \beta^*\|_2$, identical plots being obtained for the variable α) of the Sinkhorn algorithm run in dimension $N = 100$ on random marginals f, g and $A = \exp(-\frac{w}{\varepsilon})$ with the cost w given by an L^1 distance on equidistant points in $[0, 1]$ and $\varepsilon = 0.2$. Unless otherwise specified, this setting will be the one we use, with additional parameters or variations, for the subsequent experiments on classical and generalized Sinkhorn algorithms. In this particular experiment, we get that the Sinkhorn iterates converge linearly with rate $\lambda = 0.338$ while $1 - 2\underline{x}_\alpha = 0.863$ and $\kappa(A)^2 = 0.999$. In addition, these rates can only be written when A is positive. We give in the following section a more precise estimation of the linear convergence rate.

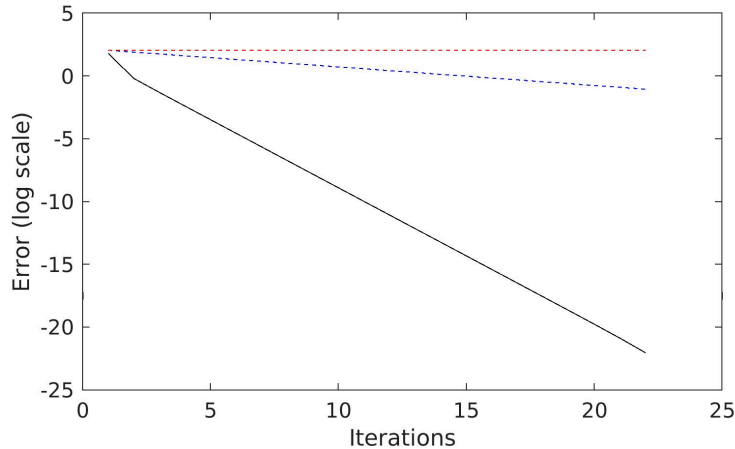


Figure 2.1 – Error evolution of the classical Sinkhorn algorithm (plain black) vs Hilbert distance (dotted red) and min-max (dotted blue) rates

2.2 General linear rate

We now turn to the general situation. We place ourselves in the weakest setting one can ask to consider the Sinkhorn’s iteration: A is a nonnegative matrix with no zero line or column.

2.2.1 Convergence conditions

Before any consideration on linear convergence, we must first precise what we mean exactly by the words “the Sinkhorn algorithm converges”, and state when this situation occurs. Indeed, in our introduction to the algorithm we deliberately presented Sinkhorn’s iterations as acting on the variables $\alpha^{(n)}$ and $\beta^{(n)}$, that is on the coefficients

of the diagonal matrices D^1 and D^2 for which D^1AD^2 would belong to $\Pi(f, g)$. This is of course what is done in practice, as a key advantage of the Sinkhorn algorithm is to be led on $2N$ rather than N^2 variables. However, the historical point of view upon the Sinkhorn algorithm (for instance in our telephone forecasting example [Kru37] and Sinkhorn's papers [Sin64, Sin67]) focuses on the *matrix* itself. For instance in [Sin64], the corresponding convergence theorem is stated the following way:

The iterative process of alternately normalizing the rows and columns of a strictly positive $N \times N$ matrix is convergent to a strictly positive doubly stochastic matrix.

meaning that the author is actually concerned by the sequences of matrices $A^{(k)}$ obtained as $A^{(0)} = A$ and $A^{(n+\frac{1}{2})} = d(\alpha^{(n+1)})Ad(\beta^{(n)})$ (whose rows sum to f), $A^{(n+1)} = d(\alpha^{(n+1)})Ad(\beta^{(n+1)})$ (whose columns sum to g). It turns out that this notion of convergence is actually strictly weaker than the convergence of the variables $\alpha^{(n)}, \beta^{(n)}$ that we considered previously: for instance taking

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \text{ with } f = g = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

one checks that the sequences $(\alpha^{(n)})$ and $(\beta^{(n)})$ are divergent (with $\alpha_1^{(n)}, \beta_2^{(n)} \rightarrow 0$ and $\alpha_2^{(n)}, \beta_1^{(n)} \rightarrow +\infty$) while the corresponding matrices

$$A^{(n)} = \begin{pmatrix} 1 & \frac{1}{2n+1} \\ 0 & \frac{2n}{2n+1} \end{pmatrix} \text{ and } A^{(n+\frac{1}{2})} = \begin{pmatrix} \frac{2n+1}{2n+2} & \frac{1}{2n+2} \\ 0 & 1 \end{pmatrix}$$

converge to the identity matrix.

Actually, in this example one couldn't expect the sequences $(\alpha^{(n)}), (\beta^{(n)})$ to converge as there is no positive diagonal matrices D^1, D^2 such that D^1AD^2 is doubly stochastic. However, the convergence of $A^{(k)}$ reveals that the matrix A is such that for any $\varepsilon > 0$, there exist positive diagonal matrices D^1, D^2 such that $B = D^1AD^2$ satisfies $\|B\mathbf{1} - f\| \leq \varepsilon$ and $\|B^T\mathbf{1} - g\| \leq \varepsilon$. This weaker scaling notion is known as *approximate scaling*, we refer again to the survey [Ide16] for more background on this subject. In the following we will not study this setting but focus on the case where $(\alpha^{(n)})$ and $(\beta^{(n)})$ converge. Still, we wish to mention that the convergence speed of the sequence of matrices $(A^{(k)})$ is deeply affected by the convergence of $(\alpha^{(n)})$ and $(\beta^{(n)})$. Indeed, when $(\alpha^{(n)})$ and $(\beta^{(n)})$ converge we will prove below that their convergence – hence the one of $(A^{(k)})$ – is linear; however, when $(\alpha^{(n)})$ and $(\beta^{(n)})$ diverge the convergence of $(A^{(k)})$ is always worse than linear as proved by Achilles in [Ach93] (for the case $f = g = \mathbf{1}$). In the following, the phrase “the Sinkhorn algorithm converges” will always refer to the situation where the sequences $(\alpha^{(n)})$ and $(\beta^{(n)})$ converge.

We have seen that a necessary condition for the Sinkhorn algorithm to converge is the existence of diagonal matrices D^1, D^2 with positive entries such that the matrix D^1AD^2 belongs to $\Pi(f, g)$. We will prove in the following section that this condition

is sufficient to obtain linear convergence. The question of when such diagonal matrices exist has already been answered in the literature: we rely on the articles [Men68], [Bru68] and [HRS88] which prove that it depends on the pattern of A , that is the position of its non-zero entries.

Definition I.1. We say that two nonnegative matrices $A, B \in (\mathbb{R}^+)^{N \times N}$ have the same pattern when

$$\forall i, j \in \llbracket 1, N \rrbracket, a_{i,j} > 0 \Leftrightarrow b_{i,j} > 0$$

Of course having the same pattern as a matrix of $\Pi(f, g)$ is a necessary condition for a scaling to exist. Actually it is also sufficient as stated by Menon in [Men68]:

Theorem I.3. Given $A \in (\mathbb{R}^+)^{N \times N}$ and $f, g \in (\mathbb{R}_*^+)^N$, there exist diagonal matrices D^1, D^2 such that $D^1 A D^2 \in \Pi(f, g)$ if and only if there exists a nonnegative matrix $B \in \Pi(f, g)$ having the same pattern as A .

This existence condition was turned into a criteria one can “test” directly (either theoretically, or numerically if A does not have too many zero entries) when given A , f and g by Brualdi in [Bru68]. This condition is quite natural: suppose that A takes the form $A = \begin{pmatrix} A_1 & X \\ 0 & A_2 \end{pmatrix}$ and admits a diagonal scaling $B = D^1 A D^2 = \begin{pmatrix} B_1 & Y \\ 0 & B_2 \end{pmatrix}$ such that $B \in \Pi(f, g)$. Denote by $I = \llbracket N_1 + 1, N \rrbracket$ and $J = \llbracket 1, N_2 \rrbracket$ the sets such that $a_{i,j} = b_{i,j} = 0$ for $(i, j) \in I \times J$. Looking at column sums, the sum of all the coefficients of B_1 must be equal to $\sum_{j \in J} g_j$. However, looking at row sums, the sum of all these coefficients added to those of Y must be equal to $\sum_{i \notin I} f_i$. This can only be achieved if $\sum_{i \notin I} f_i \geq \sum_{j \in J} g_j$. In [Bru68], Brualdi shows that this condition is actually enough to ensure the existence of a diagonal scaling:

Theorem I.4. Let $A \in (\mathbb{R}^+)^{N \times N}$ and $f, g \in (\mathbb{R}_*^+)^N$. There exists a nonnegative matrix $B \in \Pi(f, g)$ having the same pattern as A if and only if for every subsets $I, J \subset \llbracket 1, N \rrbracket$ one has

$$\left(\forall i \in I, \forall j \in J, a_{i,j} = 0 \right) \Rightarrow \left(\begin{array}{c} \sum_{i \notin I} f_i > \sum_{j \in J} g_j \\ \text{or} \\ \left(\sum_{i \notin I} f_i = \sum_{j \in J} g_j \text{ and } \right. \\ \left. \forall i \notin I, \forall j \notin J, a_{i,j} = 0 \right) \end{array} \right)$$

As a consequence, we will state our convergence theorem under the hypothesis of Theorem I.4 that we denote (I.4). Historically, the first case studied is the doubly stochastic setting $f = g = \mathbf{1}$. In that case, a more specific characterization of the support can be given.

First, we need the following definition:

Definition I.2. Let $A \in (\mathbb{R}^+)^{N \times N}$ be a nonnegative matrix. For $(i, j) \in \llbracket 1, N \rrbracket^2$, we say that the entry $a_{i,j}$ of A lies on a positive diagonal of A when there exists a permutation σ of $\llbracket 1, N \rrbracket$ such that $j = \sigma(i)$ and $\forall k \in \llbracket 1, N \rrbracket$, $a_{k, \sigma(k)} > 0$.

We say that A has support when it admits at least one positive diagonal, and total support when every non zero entry of A lies on a positive diagonal.

In [SK67], Sinkhorn and Knopp give the following criteria:

Theorem I.5. A nonnegative matrix $A \in (\mathbb{R}^+)^{N \times N}$ admits diagonal matrices D^1, D^2 with positive entries such that $D^1 A D^2$ is doubly stochastic if and only if A has total support.

Remark I.2. Still concerning the doubly stochastic setting, [SK67] also proves that a necessary and sufficient condition for the sequence of matrices $(A^{(k)})$ to converge is that A has support.

To finish with, as the diagonal matrices D^1 and D^2 will appear as the limit values of the sequences $(\alpha^{(n)}), (\beta^{(n)})$ we are not only interested in their existence, but also wonder about their uniqueness. We find in [Men68] and in [HRS88] the following result:

Theorem I.6. Let $A \in (\mathbb{R}^+)^{N \times N}$ and $f, g \in (\mathbb{R}_*^+)^N$. Suppose there exist diagonal matrices D^1, D^2 such that $D^1 A D^2 \in \Pi(f, g)$. Then:

1. The matrix $D^1 A D^2$ belonging to $\Pi(f, g)$ is unique.
2. If there is no permutation matrices P and Q such that $PAQ^T = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$ (with A_1 and A_2 rectangular matrices), then D^1 and D^2 are unique up to scalar multiple, meaning that if $D^1 A D^2 = D'^1 A D'^2$ then $D'^1 = r D^1$ and $D'^2 = \frac{1}{r} D^2$ for some $r > 0$.

The additional condition required on the matrix A in the second point is actually natural. Indeed, applying the Sinkhorn algorithm to any matrix A always reduces to applying it in parallel to several matrices A_1, \dots, A_k which satisfy this condition. More precisely, suppose A is such that $PAQ^T = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$ for some permutation matrices P, Q corresponding to permutations σ, ρ and some rectangular matrices $A_1 \in (\mathbb{R}^+)^{N_1 \times N_2}$, $A_2 \in (\mathbb{R}^+)^{N'_1 \times N'_2}$. Then denoting $\tilde{f}_i = f_{\sigma(i)}$, $\tilde{g}_j = g_{\rho(j)}$, one checks that the Sinkhorn algorithm on A, f, g is split into the Sinkhorn algorithms on $A_1, (\tilde{f}_i)_{1 \leq i \leq N_1}, (\tilde{g}_j)_{1 \leq j \leq N_2}$ and $A_2, (\tilde{f}_i)_{N_1+1 \leq i \leq N}, (\tilde{g}_j)_{N_2+1 \leq j \leq N}$. Note that one can have $P \neq Q$ which makes it tricky to catch such a decomposition at first glance on a given matrix. Note also that the smaller Sinkhorn algorithms that appear are led on the rectangular matrices A_1, A_2 . As we wanted to keep our analysis with square matrices, we do not ask this uniqueness hypothesis in our theorems but lead another equivalent reduction instead, see Remark I.3.

In the doubly stochastic setting, this reduction hypothesis is simplified thanks to the following lemma that we present here because we will use similar arguments in our result I.6 concerning reduction of symmetric stochastic matrices.

Lemma I.4. *Suppose A is a nonnegative matrix admitting positive diagonal scalings D^1, D^2 such that $D^1 A D^2$ is doubly stochastic. Then we have equivalence between:*

1. *There exist permutation matrices P, Q such that $PAQ^T = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$ for rectangular matrices A_1, A_2 , and*
2. *There exist permutation matrices P, Q such that $PAQ^T = \begin{pmatrix} A_1 & X \\ 0 & A_2 \end{pmatrix}$ for square matrices A_1, A_2 .*

Proof. Suppose we have 1., as A admits a doubly stochastic scaling there exist a doubly stochastic matrix $B = \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix}$ with B_1 and B_2 having the same sizes as A_1 and A_2 . But then B_1 and B_2 ought to be doubly stochastic as well, hence square. Suppose now we have 2., then there exist a doubly stochastic matrix $B = \begin{pmatrix} B_1 & Y \\ 0 & B_2 \end{pmatrix}$ with the same pattern as PAQ^T . Denote N_1 the size of B_1 . From $B\mathbf{1} = B^T\mathbf{1} = \mathbf{1}$ one deduces for appropriate sizes of the unit vector $\mathbf{1}$: $B_1\mathbf{1} + Y\mathbf{1} = \mathbf{1}$ and $B_1^T\mathbf{1} = \mathbf{1}$ hence the sum of the coefficients of Y is equal to $\langle Y\mathbf{1}|\mathbf{1} \rangle = \langle \mathbf{1} - B_1\mathbf{1}|\mathbf{1} \rangle = N_1 - \langle \mathbf{1}|B_1^T\mathbf{1} \rangle = 0$, hence $Y = 0$, which implies $X = 0$. \square

Therefore the previous theorem about uniqueness of the diagonal matrices D^1, D^2 reduces in that setting to the following result due to [SK67]:

Proposition I.2. *Let A be a nonnegative matrix with total support. The diagonal matrices D^1, D^2 such that $D^1 A D^2$ is doubly stochastic are unique up to scalar multiple if and only if A is fully indecomposable, meaning that there is no permutation matrices P, Q such that $PAQ^T = \begin{pmatrix} A_1 & X \\ 0 & A_2 \end{pmatrix}$ for square matrices A_1, A_2 .*

2.2.2 A linear convergence rate

As we have already mentioned, although the convergence of the Sinkhorn iterates is indeed linear, the rates we gave in subsection 2.1 are pessimistic. To obtain the exact convergence rate, one must study the spectrum of the iterated function $T : (\mathbb{R}_*^+)^N \rightarrow (\mathbb{R}_*^+)^N$ such that $T(\alpha^{(n)}, \beta^{(n)}) = (\alpha^{(n+1)}, \beta^{(n+1)})$ near a fixed point (α^*, β^*) . This strategy was already suggested by Menon and Schneider in [MS69], but it is not until [Kni08] (see also [Sou91]) that a linear convergence rate involving the subdominant eigenvalue of the appropriate matrix was stated by Knight in the doubly stochastic setting.

His proof is based on the study of the distance of the iterates to the line of multiples of the solution, and relies on his Lemma 4.3. When iterating the function T near a multiple $(s\alpha^*, \frac{1}{s}\beta^*)$ of the fixed point (α^*, β^*) , one wants to estimate the value of

$$\min_{r>0} \|T(s\alpha^* + d_\alpha, \frac{1}{s}\beta^* + d_\beta) - (r\alpha^*, \frac{1}{r}\beta^*)\|$$

for small d_α, d_β and in a certain norm $\|\cdot\|$. Knight obtains a bound written as

$$|\lambda_2|^2 \|(d_\alpha, d_\beta)\| + o(\|(d_\alpha, d_\beta)\|)$$

for some $\lambda_2 \in (0, 1)$. However, we do not clearly see how the dependence on s of the $o(\|(d_\alpha, d_\beta)\|)$ term is avoided in order to lead to the concluding Theorem 4.4. In the following we give a different proof of this convergence result that furthermore extends to arbitrary marginals f, g .

Remember from Lemma I.1 that provided the scaling problem I.1 admits a solution one can study the Sinkhorn iterates on the rescaled variables $a^{(n)}, b^{(n)}$. The iterations write

$$a^{(n+1)} = \frac{1}{d(\frac{1}{f})X b^{(n)}} ; b^{(n+1)} = \frac{1}{d(\frac{1}{g})X^T a^{(n+1)}}$$

where $X = d(\alpha^*)A d(\beta^*) \in \Pi(f, g)$ so that the matrices $d(\frac{1}{f})X$ and $d(\frac{1}{g})X^T$ are stochastic. As in Lemma I.2, we use the stochasticity of these matrices to obtain our bounds. More precisely, we now exploit the following lemma that simply derives from the convexity of the inverse function:

Lemma I.5. *Let $M \in (\mathbb{R}^+)^{N \times N}$ be a stochastic matrix. Then for any positive vector $x \in (\mathbb{R}_*^+)^N$, one has*

$$\frac{1}{Mx} \leq M \frac{1}{x}$$

Using this lemma with $M = d(\frac{1}{f})X$ and $M = d(\frac{1}{g})X^T$ finally gives us:

Fact I.1. *On the rescaled variables $a^{(n)}, b^{(n)}$ of the Sinkhorn iterates, one has $\forall n \geq 0$,*

$$\begin{aligned} a^{(n+1)} &\leq M_a a^{(n)} ; \frac{1}{a^{(n+1)}} \leq M_a \frac{1}{a^{(n)}} \\ b^{(n+1)} &\leq M_b b^{(n)} ; \frac{1}{b^{(n+1)}} \leq M_b \frac{1}{b^{(n)}} \end{aligned} \tag{2.2}$$

where $M_a = d(\frac{1}{f})X d(\frac{1}{g})X^T$ and $M_b = d(\frac{1}{g})X^T d(\frac{1}{f})X$.

What can we say about the matrices M_a and M_b ? First, $d(f)M_a$ and $d(g)M_b$ are symmetric, or in other words M_a is symmetric for the scalar product $\langle \cdot, \cdot \rangle_f$, and M_b for $\langle \cdot, \cdot \rangle_g$. This shows that M_a and M_b are diagonalizable, and their spectrum, which is the same as $Sp(YZ) = Sp(ZY)$ for any square matrices Y, Z , is real. Next, as $X \in \Pi(f, g)$, M_a and M_b are stochastic matrices, hence their spectral radius

$\rho(M_a) = \max\{|\lambda|, \lambda \in Sp(M_a)\}$ is 1. In the objective of obtaining a contraction result for a convergence rate, one would also like 1 to be a strictly dominant eigenvalue, meaning that $\forall \lambda \in Sp(M_a) \setminus \{1\}, |\lambda| < 1$. This last point is not true in general, but we will reduce our setting to it thanks to a result about reduction of stochastic symmetric matrices. To do so, let us first introduce a standard definition (see [Min88] chapters I.2 and III):

Definition I.3. A nonnegative matrix $A \in (\mathbb{R}^+)^{N \times N}$ is called reducible when there exists a permutation matrix P such that

$$PAP^T = \begin{pmatrix} B & X \\ 0 & C \end{pmatrix}$$

where B and C are square submatrices. Otherwise A is called irreducible.

When dealing with symmetric stochastic matrices, the natural reduction of any nonnegative matrix to an upper triangular matrix with irreducible matrices on the diagonal can be made into diagonal form:

Lemma I.6. Let $M \in (\mathbb{R}^+)^{N \times N}$ be a stochastic matrix which is symmetric for some scalar product $\langle \cdot, \cdot \rangle_s$ given by a positive vector $s \in (\mathbb{R}_*^+)^N$. Then there exist a permutation matrix P and stochastic irreducible square matrices M_1, \dots, M_r which are symmetric for scalar products $\langle \cdot, \cdot \rangle_{s_i}$ given by positive vectors s_i of appropriate sizes such that :

$$PMP^T = \begin{pmatrix} M_1 & & 0 \\ & \ddots & \\ 0 & & M_r \end{pmatrix}$$

Proof. Suppose M is reducible and write $PMP^T = \begin{pmatrix} M_1 & X \\ 0 & M_2 \end{pmatrix}$ for some square matrices $M_1 \in (\mathbb{R}^+)^{k \times k}$, $M_2 \in (\mathbb{R}^+)^{(N-k) \times (N-k)}$ and $X \in (\mathbb{R}^+)^{k \times (N-k)}$. First, the fact that $M\mathbf{1} = \mathbf{1}$ gives in particular (with appropriate lengths for $\mathbf{1}$) $M_1\mathbf{1} + X\mathbf{1} = \mathbf{1}$. Second, the fact that $M^T s = s$ (which comes from $M\mathbf{1} = \mathbf{1}$ and the symmetry of $d(s)M$) gives in particular $M_1^T s_1 = s_1$ with $s_1 \in (\mathbb{R}_*^+)^k$ being the first k components of Ps . Then one has $\langle s_1 | X\mathbf{1} \rangle = \langle s_1 | \mathbf{1} - M_1\mathbf{1} \rangle = \langle s_1 | \mathbf{1} \rangle - \langle M_1^T s_1 | \mathbf{1} \rangle = 0$ so that $X = 0$ (because $s_1 > 0$). The result follows by induction. \square

On the later, we will apply this reduction to the matrices M_a, M_b ; it is important to notice that this does not affect the relations (2.2). We mean that if M_a is reduced through the permutation matrix P_a then the relation (2.2) on the vector $a^{(n)}$ splits into smaller identical relations on partitions of the vector $P_a a^{(n)}$. We will denote $a_s^{(n)}$ such partitions: the vectors $a_s^{(n)}$ simply consist of reorderings of the components of the vector $a^{(n)}$ that satisfy the relations

$$a_s^{(n+1)} \leq M_{a,s} a_s^{(n)}, \quad \frac{1}{a_s^{(n+1)}} \leq M_{a,s} \frac{1}{a_s^{(n)}}$$

for some irreducible matrices $M_{a,s}$. The process is of course similar for $b^{(n)}$ with M_b . We now wish to have results about the spectrum of these smaller irreducible matrices $M_{a,s}$ and $M_{b,s}$ and introduce the notion of primitivity:

Definition I.4. A nonnegative matrix $M \in (\mathbb{R}^+)^{N \times N}$ is called primitive when it is irreducible and such that $\forall \lambda \in Sp(M) \setminus \{\rho(M)\}$, $|\lambda| < \rho(M)$.

Many characterizations of primitive matrices can be found in [Min88, BP79]. We will need the two following practical results (found in [BP79] (Theorem 2.7 and Corollary 2.28)):

Proposition I.3. For a nonnegative matrix $M \in (\mathbb{R}^+)^{N \times N}$,

1. M is primitive if and only if there exists $p \in \mathbb{N}$ such that M^p is positive.
2. If M is irreducible and if $Tr(M) > 0$ then² M is primitive.

In our setting, every diagonal coefficient of the matrices M_a and M_b is positive. Indeed, remember that $M_a = d(\frac{1}{f})X d(\frac{1}{g})X^T$ with $X = D^1AD^2$ being the diagonal scaling of A , hence M_a has the same pattern as AA^T . But $(AA^T)_{i,i} = \sum_k a_{i,k}^2 > 0$ because A has no zero line. Similarly, M_b has the pattern of $A^T A$ hence its main diagonal entries are positive because A has no zero column. This proves that the irreducible matrices that appear in the block diagonal decomposition of M_a and M_b have positive trace hence are primitive. Finally, the Sinkhorn iterates on $a^{(n)}, b^{(n)}$ write as parallel iterates on $a_s^{(n)}, b_s^{(n)}$ satisfying (2.2) with primitive stochastic symmetric matrices $M_{a,s}$ and $M_{b,s}$. This is what we needed to conclude with our general linear convergence theorem:

Theorem I.7. Let $M \in (\mathbb{R}^+)^{N \times N}$ be a primitive stochastic matrix which is symmetric for some scalar product $\langle \cdot, \cdot \rangle_s$ given by $s \in \Sigma^N$. Suppose $(x^n) \in ((\mathbb{R}_*^+)^N)^{\mathbb{N}}$ is a sequence satisfying:

$$\forall n \in \mathbb{N}, x^{n+1} \leq Mx^n \text{ and } \frac{1}{x^{n+1}} \leq M \frac{1}{x^n}$$

Then there exists $x^* \in \mathbb{R}_*^+$ such that $x^n \xrightarrow[n \rightarrow +\infty]{} x^* \mathbf{1}$ with the following estimate:
 $\forall \delta > 0, \exists n_\delta \in \mathbb{N}$ such that $\forall n \geq n_\delta$

$$\|x^{n+1} - \langle x^{n+1} \rangle_s \mathbf{1}\|_s \leq (\lambda_2 + \delta) \|x^n - \langle x^n \rangle_s \mathbf{1}\|_s$$

where $1 > \lambda_2 = \max Sp(M) \setminus \{1\}$ is the subdominant eigenvalue of M .

We postpone the proof of this theorem to section 2.3. Now let us conclude: for each smaller variable $a_s^{(n)}, b_s^{(n)}$ we have convergence towards a certain multiple $x^* \mathbf{1}$ of the unit vector $\mathbf{1}$, which corresponds to a solution of the scaling problem I.1. Had we rescaled the initial variables $\alpha^{(n)}, \beta^{(n)}$ by this solution instead of a random α^*, β^* ,

²The converse is wrong: $M = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$ is primitive but $Tr(M) = 0$.

all the variables $a_s^{(n)}, b_s^{(n)}$ would have converged towards $\mathbf{1}$. With this new rescale, the matrices M_a and M_b remain unchanged (as the diagonal scaling $X \in \Pi(f, g)$ is unique, Theorem I.6) and one still has asymptotically

$$\begin{aligned} \left\| a^{(n+1)} - \langle a^{(n+1)} \rangle_f \mathbf{1} \right\|_f &\leq (\lambda_2 + \delta) \left\| a^{(n)} - \langle a^{(n)} \rangle_f \mathbf{1} \right\|_f \\ \left\| b^{(n+1)} - \langle b^{(n+1)} \rangle_g \mathbf{1} \right\|_g &\leq (\lambda_2 + \delta) \left\| b^{(n)} - \langle b^{(n)} \rangle_g \mathbf{1} \right\|_g \end{aligned}$$

with $\lambda_2 = \max\{|\lambda|, \lambda \in Sp(M_a) \text{ s.t. } |\lambda| < 1\}$ being the subdominant eigenvalue of M_a and M_b . Then $u_n = \langle a^{(n)} \rangle_f$ and $v_n = \langle b^{(n)} \rangle_g$ converge to 1 and one checks that the linear convergence of $a^{(n)}, b^{(n)}$ in the $\|\cdot\|_f$ and $\|\cdot\|_g$ norms reformulates in the norm given by $s = (\frac{f}{\alpha^{*2}}, \frac{g}{\beta^{*2}})$ into the following theorem:

Theorem I.8. *Let $A \in (\mathbb{R}^+)^{N \times N}$ and $f, g \in \Sigma^N$ satisfying hypothesis (I.4). The Sinkhorn iterates $(\alpha^{(n)}, \beta^{(n)})$ corresponding to A, f, g and defined by (1.3) converge to $\alpha^*, \beta^* \in (\mathbb{R}_*^+)^N$ which are solutions of the scaling problem I.1. In addition there exist sequences of real numbers $(u_n), (v_n) \in (\mathbb{R}_*^+)^N$ converging to 1, and a norm $\|\cdot\|_s$ on \mathbb{R}^{2N} such that: $\forall \delta > 0, \exists n_\delta$ such that $\forall n \geq n_\delta$,*

$$\left\| \begin{pmatrix} \alpha^{(n+1)} \\ \beta^{(n+1)} \end{pmatrix} - \begin{pmatrix} u_{n+1} \alpha^* \\ v_{n+1} \beta^* \end{pmatrix} \right\|_s \leq (\lambda_2 + \delta) \left\| \begin{pmatrix} \alpha^{(n)} \\ \beta^{(n)} \end{pmatrix} - \begin{pmatrix} u_n \alpha^* \\ v_n \beta^* \end{pmatrix} \right\|_s$$

where $\lambda_2 = \max\{|\lambda|, \lambda \in Sp(M) \text{ s.t. } |\lambda| < 1\}$ is the subdominant eigenvalue of the matrix $M = d(\frac{1}{g})X^T d(\frac{1}{f})X$ and $X = d(\alpha^*)A d(\beta^*) \in \Pi(f, g)$ is the limit matrix of the Sinkhorn iterations.

Remark I.3. *During the course of the proof, we dealt in Lemma I.6 with the reduction as diagonal blocks of irreducible matrices of the matrix $M = d(\frac{1}{f})X d(\frac{1}{g})X^T$. This step could have been overlooked by previously reducing the matrix A itself to the form described in Theorem I.6 that ensures uniqueness of the scaling diagonal matrices D^1, D^2 . Remember that this reduction writes $PAQ^T = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$ with rectangular matrices A_1, A_2 and that the Sinkhorn iterates split into smaller blocks so that the convergence of the algorithm is obtained by the convergence of each block. Actually, the two decompositions coincide so that reducing A is enough to ensure the irreducibility of M . This result relies on the fact that M has the pattern of AA^T and writes as the following proposition. However, we chose to present our reduction on the matrix M to keep all the Sinkhorn iterates on square matrices.*

Proposition I.4. *For a nonnegative matrix A having no zero row or column, we have equivalence between:*

1. *There is no permutation matrices P, Q such that $PAQ^T = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$ with A_1, A_2 being rectangular matrices, and*
2. *There is no permutation matrix P such that $PAA^T P^T = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix}$ with M_1, M_2 being square matrices.*

Proof. The fact that 2. implies 1. is easy. For the converse, suppose AA^T admits the decomposition $AA^T = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix}$ with $M_1 \in \mathbb{R}^{N_1 \times N_1}$ (the case $P \neq Id$ falls into this setting just replacing A by PA). One wants to find a permutation matrix Q such that $AQ^T = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$. In other words one wants to split the columns of A into two groups: one group E such that $\forall k \in E, \forall i > N_1, a_{i,k} = 0$ and one group F such that $\forall k \in F, \forall j \leq N_1, a_{j,k} = 0$. The hypothesis on AA^T gives that:

$$\forall i > N_1, \forall j \leq N_1, (AA^T)_{i,j} = \sum_k a_{i,k}a_{j,k} = 0$$

This implies that for all $k \in \llbracket 1, N \rrbracket, \forall i > N_1, \forall j \leq N_1, a_{i,k} = 0$ or $a_{j,k} = 0$. Then setting $E = \{k \text{ s.t. } \exists i > N_1 \text{ s.t. } a_{i,k} > 0\}$ and $F = \{k \text{ s.t. } \exists j \leq N_1 \text{ s.t. } a_{j,k} > 0\}$ gives the desired result. \square

To conclude, we note that numerical computations confirm that the second eigenvalue λ_2 of the matrix $M = d(\frac{1}{g})X^T d(\frac{1}{f})X$ is the actual linear convergence rate of the Sinkhorn algorithm, see Figure 2.2. In this experiment, we took the same setting than for Figure 2.1, and run it for different the values of the parameter ε . Sadly, this convergence rate is given according to the limit matrix X of the iterates, which is unknown *a priori*. Future works should try to understand how this rate depends on the algorithm data A, f, g . In section 4.3 we explore the case $N = 2$ where everything can be computed explicitly enlightening this dependency.

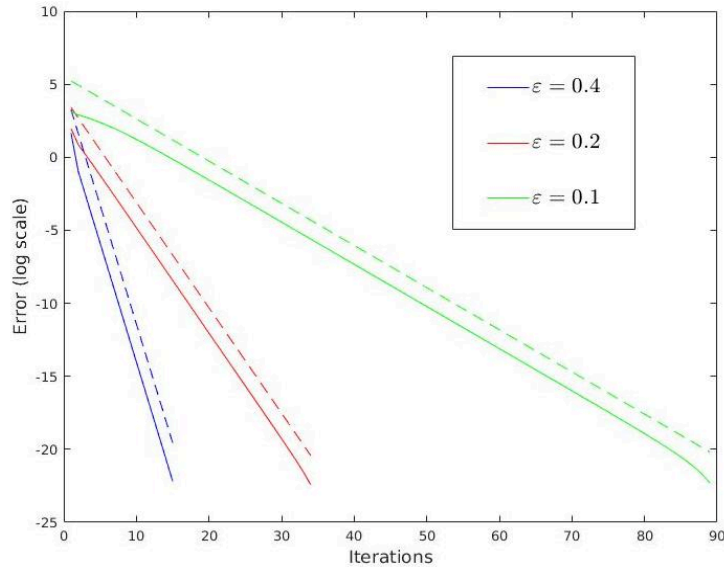


Figure 2.2 – Error evolution of the classical Sinkhorn algorithm (plain lines) and predicted λ_2 rate (dotted lines) for different values of ε

2.3 Proof of general theorem

To prepare the proof of Theorem I.7, we need the following lemma:

Lemma I.7. *Let $M \in (\mathbb{R}^+)^{N \times N}$ be a stochastic matrix and let $s \in (\mathbb{R}_*^+)^N$. For any $c > 0$, there exists $\mu > 0$ such that for any positive vector $x \in (\mathbb{R}_*^+)^N$ satisfying $\langle x \rangle_s \geq c$ and $x \geq \frac{1}{2} \langle x \rangle_s \mathbf{1}$ one has*

$$\frac{1}{M_x} \geq Mx - \mu \|x - \langle x \rangle_s \mathbf{1}\|_s^2$$

Proof. First use the fact that $\forall t \geq -\frac{1}{2}$, $\frac{1}{1+t} \leq 1 - t + 2t^2$, so that for $x \in (\mathbb{R}_*^+)^N$ and $h \in \mathbb{R}^N$ such that $h \geq -\frac{1}{2} \langle x \rangle_s \mathbf{1}$ one has componentwise:

$$\frac{1}{\langle x \rangle_s \mathbf{1} + h} = \frac{1}{\langle x \rangle_s} \frac{1}{\mathbf{1} + \frac{h}{\langle x \rangle_s}} \leq \frac{1}{\langle x \rangle_s} \left(\mathbf{1} - \frac{h}{\langle x \rangle_s} + 2 \frac{h^2}{\langle x \rangle_s^2} \right)$$

As $M \geq 0$ and $M\mathbf{1} = \mathbf{1}$ we derive successively:

$$\begin{aligned} M \frac{1}{\langle x \rangle_s \mathbf{1} + h} &\leq \frac{1}{\langle x \rangle_s} \left(\mathbf{1} - \frac{1}{\langle x \rangle_s} Mh + \frac{2}{\langle x \rangle_s^2} M(h^2) \right) \\ \frac{1}{M \frac{1}{\langle x \rangle_s \mathbf{1} + h}} &\geq \langle x \rangle_s \frac{1}{\mathbf{1} + \left(-\frac{1}{\langle x \rangle_s} Mx + \frac{2}{\langle x \rangle_s^2} M(h^2) \right)} \end{aligned}$$

Using now that $\frac{1}{1+t} \geq 1 - t$ we get:

$$\frac{1}{M \frac{1}{\langle x \rangle_s \mathbf{1} + h}} \geq \langle x \rangle_s \left(\mathbf{1} + \frac{1}{\langle x \rangle_s} Mh - \frac{2}{\langle x \rangle_s^2} M(h^2) \right)$$

Taking $h = x - \langle x \rangle_s \mathbf{1}$, we obtain, if $h \geq -\frac{1}{2} \langle x \rangle_s \mathbf{1}$ i.e. if $x \geq \frac{1}{2} \langle x \rangle_s \mathbf{1}$:

$$\begin{aligned} \frac{1}{M_x} &\geq \langle x \rangle_s \mathbf{1} + Mx - \langle x \rangle_s M\mathbf{1} - \frac{2}{\langle x \rangle_s} M((x - \langle x \rangle_s \mathbf{1})^2) \\ &\geq Mx - \frac{2}{\langle x \rangle_s} \frac{1}{s} \|x - \langle x \rangle_s \mathbf{1}\|_s^2 \mathbf{1} \end{aligned}$$

where the last inequality uses the fact that all the coefficients of M are bounded by 1 to state that:

$$\forall x \in (\mathbb{R}_*^+)^N, M(x^2) \leq \frac{1}{s} \|x\|_s^2 \mathbf{1}$$

The result dealing with the case where $\langle x \rangle_s \geq c$ follows by taking $\mu = \frac{2}{cs}$. \square

We now turn to the proof of our main theorem that we recall below:

Theorem I.9. *Let $M \in (\mathbb{R}^+)^{N \times N}$ be a primitive stochastic matrix which is symmetric for some scalar product $\langle \cdot, \cdot \rangle_s$ given by $s \in \Sigma^N$. Suppose $(x^n) \in ((\mathbb{R}_*^+)^N)^\mathbb{N}$ is a sequence satisfying:*

$$\forall n \in \mathbb{N}, x^{n+1} \leq Mx^n \text{ and } \frac{1}{x^{n+1}} \leq M \frac{1}{x^n}$$

Then there exists $x^ \in \mathbb{R}_*^+$ such that $x^n \xrightarrow{n \rightarrow +\infty} x^* \mathbf{1}$ with the following estimate:
 $\forall \delta > 0, \exists n_\delta \in \mathbb{N}$ such that $\forall n \geq n_\delta$*

$$\|x^{n+1} - \langle x^{n+1} \rangle_s \mathbf{1}\|_s \leq (\lambda_2 + \delta) \|x^n - \langle x^n \rangle_s \mathbf{1}\|_s$$

where $1 > \lambda_2 = \max \text{Sp}(M) \setminus \{1\}$ is the subdominant eigenvalue of M .

Proof. Step 1: We give here a first estimate which is valid for all iterations and constitutes a convergence proof.

First, note that since $x^{n+1} \leq Mx^n$ and $M\mathbf{1} = \mathbf{1}$, applying the nonnegative matrix M to the inequality $x^n \leq \bar{x}^n \mathbf{1}$ leads to $\bar{x}^{n+1} \leq \bar{x}^n$. Using similarly that $\frac{1}{x^{n+1}} \leq M \frac{1}{x^n}$ leads to $\underline{x}^{n+1} \geq \underline{x}^n$. Consequently the sequences (\bar{x}^n) and (\underline{x}^n) both converge. We denote $x^* \in \mathbb{R}_*^+$ the limit of (\underline{x}^n) and we will show that $x^n \xrightarrow{n \rightarrow +\infty} x^* \mathbf{1}$.

Second, since M is stochastic and primitive, it has 1 as a dominant and simple eigenvalue. Exploiting the fact that M is also symmetric for $\langle \cdot, \cdot \rangle_s$, we obtain that

$$1 > \lambda_2 = \sup_{\substack{x \neq 0 \text{ s.t.} \\ \langle x \rangle_s = 0}} \frac{\|Mx\|_s}{\|x\|_s}$$

which we will use by saying that for any vector x one has

$$\|M(x - \langle x \rangle_s \mathbf{1})\|_s^2 \leq \lambda_2^2 \|x - \langle x \rangle_s \mathbf{1}\|_s^2 \quad (2.3)$$

Our analysis starts with saying that $x^{n+1} \leq Mx^n$ implies that

$$x^{n+1} - \underline{x}^{n+1} \mathbf{1} \leq M(x^n - \langle x^n \rangle_s \mathbf{1}) + (\langle x^n \rangle_s - \underline{x}^n) \mathbf{1}$$

As these vectors are nonnegative, this componentwise inequality transfers to a norm inequality

$$\|x^{n+1} - \underline{x}^{n+1} \mathbf{1}\|_s^2 \leq \|M(x^n - \langle x^n \rangle_s \mathbf{1}) + (\langle x^n \rangle_s - \underline{x}^n) \mathbf{1}\|_s^2$$

Then, we develop these two square norms using orthogonality: for $x \in \mathbb{R}^N$ such that $\langle x \rangle_s = 0$ and $r \in \mathbb{R}$ one has $\|x + r\mathbf{1}\|_s^2 = \|x\|_s^2 + r^2$ because $\|\mathbf{1}\|_s^2 = 1$. Applying this result to $x = x^{n+1} - \langle x^{n+1} \rangle_s \mathbf{1}$ and to $x = M(x^n - \langle x^n \rangle_s \mathbf{1})$ (which satisfies $\langle x \rangle_s = 0$ because M is symmetric), we get:

$$\|x^{n+1} - \langle x^{n+1} \rangle_s \mathbf{1}\|_s^2 + |\langle x^{n+1} \rangle_s - \underline{x}^{n+1}|^2 \leq \lambda_2^2 \|x^n - \langle x^n \rangle_s \mathbf{1}\|_s^2 + |\langle x^n \rangle_s - \underline{x}^n|^2$$

To obtain a geometric convergence from this inequality, we compare the two terms by noting that for any vector x one has $|\langle x \rangle_s - \underline{x}|^2 \leq \frac{1}{s} \|x - \langle x \rangle_s \mathbf{1}\|_s^2$. It follows that for any $\eta > 0$:

$$\begin{aligned} & \|x^{n+1} - \langle x^{n+1} \rangle_s \mathbf{1}\|_s^2 + |\langle x^{n+1} \rangle_s - \underline{x}^{n+1}|^2 \\ & \leq (\lambda_2^2 + \eta) \|x^n - \langle x^n \rangle_s \mathbf{1}\|_s^2 + (1 - \eta s) |\langle x^n \rangle_s - \underline{x}^n|^2 \end{aligned}$$

Taking $\eta > 0$ such that $\lambda_2^2 + \eta = 1 - \eta s$ i.e. $\eta = \frac{1 - \lambda_2^2}{1 + s} > 0$ we obtain

$$\begin{aligned} & \|x^{n+1} - \langle x^{n+1} \rangle_s \mathbf{1}\|_s^2 + |\langle x^{n+1} \rangle_s - \underline{x}^{n+1}|^2 \\ & \leq (\lambda_2^2 + \eta) (\|x^n - \langle x^n \rangle_s \mathbf{1}\|_s^2 + |\langle x^n \rangle_s - \underline{x}^n|^2) \end{aligned}$$

with $\lambda_2^2 + \eta = \frac{1 + s \lambda_2^2}{1 + s} < 1$ so that $(\langle x^n \rangle_s - \underline{x}^n)$ and $(x^n - \langle x^n \rangle_s \mathbf{1})$ are converging sequences. As we already know that $\underline{x}^n \xrightarrow{n \rightarrow +\infty} x^*$ we finally get that $x^n \xrightarrow{n \rightarrow +\infty} x^* \mathbf{1}$.

Step 2: We now derive the sharper asymptotic rate.

As $x^n \xrightarrow{n \rightarrow +\infty} x^* \mathbf{1}$ for some real number $x^* > 0$, we are in the context of Lemma I.7:

$$\exists c > 0, \exists N \in \mathbb{N} \text{ such that } \forall n \geq N, x^n \geq \frac{1}{2} \langle x^n \rangle_s \mathbf{1} \geq \frac{1}{2} c \mathbf{1}$$

hence

$$\exists \mu > 0 \text{ such that } \forall n \geq N, \frac{1}{M \frac{1}{x^n}} \geq M x^n - \mu \|x^n - \langle x^n \rangle_s \mathbf{1}\|_s^2$$

and we can then write:

$$\begin{aligned} M(x^n - \langle x^n \rangle_s \mathbf{1}) - \mu \|x^n - \langle x^n \rangle_s \mathbf{1}\|_s^2 \mathbf{1} & \leq \frac{1}{M \frac{1}{x^n}} - \langle x^n \rangle_s M \mathbf{1} \\ & \leq x^{n+1} - \langle x^n \rangle_s \mathbf{1} \leq M(x^n - \langle x^n \rangle_s \mathbf{1}) \end{aligned}$$

Denoting $e^n = x^n - \langle x^n \rangle_s \mathbf{1}$ and $f^n = x^{n+1} - \langle x^n \rangle_s \mathbf{1}$ the previous inequalities can be written as $M e^n - \mu \|e^n\|_s^2 \mathbf{1} \leq f^n \leq M e^n$ or $0 \leq M e^n - f^n \leq \mu \|e^n\|_s^2 \mathbf{1}$. Passing to the norm on these nonnegative vectors yields $\|M e^n - f^n\|_s \leq \mu \|e^n\|_s^2$ and finally:

$$\|f^n\|_s \leq \|M e^n\|_s + \mu \|e^n\|_s^2$$

But one easily shows that for $z \in \mathbb{R}^N$ the minimum value of $\|z - t \mathbf{1}\|_s$ is obtained for $t = \langle z \rangle_s$ so that $\|f^n\|_s \geq \|e^{n+1}\|_s$, and we obtain:

$$\|e^{n+1}\|_s \leq \|M e^n\|_s + \mu \|e^n\|_s^2$$

To finish with, we use the symmetry of M through property (2.3) and get:

$$\exists \mu > 0 \exists N \in \mathbb{N} \text{ such that } \forall n \geq N, \|e^{n+1}\|_s \leq \lambda_2 \|e^n\|_s + \mu \|e^n\|_s^2$$

from which the conclusion of the theorem follows by taking $n_\delta \geq N$ large enough to ensure $\|e^n\|_s \leq \frac{\delta}{\mu}$. \square

CHAPTER 3

LINEAR CONVERGENCE OF SINKHORN-LIKE ALGORITHMS

The Sinkhorn algorithm computes an approximate Wasserstein distance between two measures. The growing use of Wasserstein distances in problems involving measure comparisons (see [PC19]) induced the emergence of variants of the Sinkhorn iterates. These variants arise most of the time from an entropic regularization of a linear program with marginal constraints. We describe and analyze below some of the Sinkhorn-like algorithms that appear in this context. By “Sinkhorn-like” we mean that all these algorithms involve componentwise products of nonnegative matrices and vectors. Some of these variants also come with a matrix scaling problem similar to I.1.

3.1 Barycenter of two measures

3.1.1 Simple barycenter

The first variant of the Sinkhorn algorithm we study deals with finding the Wasserstein barycenter of two discrete measures $f^0, f^1 \in \Sigma^N$. Following for instance the work of Benamou and co-authors in [BCC⁺15], one introduces this notion of barycenter as a Fréchet mean in the Wasserstein metric space, meaning that for $\theta \in (0, 1)$, one seeks to find $f^\theta \in \Sigma^N$ minimizing the weighted sum $\theta W_w(f^0, f^\theta) + (1 - \theta) W_w(f^\theta, f^1)$ where W_w denotes, as in (1.4), the Wasserstein distance according to some ground cost $w \in \mathbb{R}^{N \times N}$. This notion actually coincides with the McCann interpolation of f^0 and f^1 , first introduced in [McC97], so that the path $(f^\theta)_{\theta \in (0,1)}$ is the geodesic between f^0 and f^1 in the Wasserstein metric space.

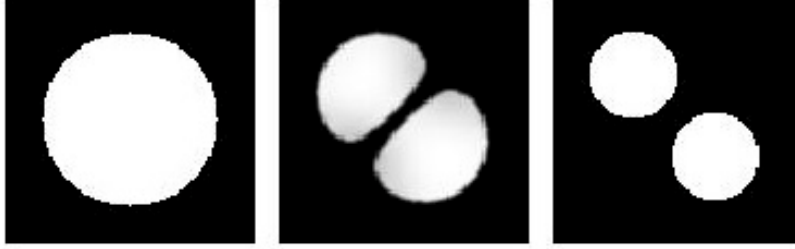


Figure 3.1 – A Wasserstein barycenter $f^{\frac{1}{2}}$ (middle) of two marginals f^0 (left) and f^1 (right) computed through the Sinkhorn-like algorithm I.2

Again, our approach deals with the entropic regularizations of Wasserstein distances, so that our problem actually consists of finding for a small $\varepsilon > 0$ the minimizer of $\theta W_w^\varepsilon(f^0, f^\theta) + (1 - \theta)W_w^\varepsilon(f^\theta, f^1)$ where W_w^ε is the regularization (1.5) of W_w . Although this is not what one does in practice to find barycenters, our following analysis does not require that the two Wasserstein distances appearing in this problem be calculated according to the same ground cost w . As a consequence we state our problem with two ground costs $w^0, w^1 \in \mathbb{R}^{N \times N}$ as:

$$\min_{x^0, x^1} \theta \langle w^0 | x^0 \rangle + \varepsilon \theta \langle x^0 | \log x^0 - \mathbf{1} \rangle + (1 - \theta) \langle w^1 | x^1 \rangle + \varepsilon (1 - \theta) \langle x^1 | \log x^1 - \mathbf{1} \rangle \quad (3.1)$$

where the minimum runs over $x^0 \in \Pi(f^0, f^\theta), x^1 \in \Pi(f^\theta, f^1)$ where $f^\theta = x^{0T} \mathbf{1} = x^1 \mathbf{1}$ or in other words $x^0, x^1 \in (\mathbb{R}^+)^{N \times N}$ such that $x^0 \mathbf{1} = f^0, x^{1T} \mathbf{1} = f^1$ and $x^{0T} \mathbf{1} = x^1 \mathbf{1}$. Introducing the Lagrange multipliers $\lambda, \mu, \nu \in \mathbb{R}^N$ for, respectively, $x^0 \mathbf{1} = f^0, x^{1T} \mathbf{1} = f^1$ and $x^{0T} \mathbf{1} = x^1 \mathbf{1}$, we obtain using Proposition I.1 the following dual formulation:

$$\begin{aligned} & \sup_{\lambda, \mu, \nu \in \mathbb{R}^N} \langle \lambda | f^0 \rangle + \langle \mu | f^1 \rangle \\ & - \varepsilon \sum_{i,j=1}^N \theta \exp \left(\frac{-\theta w_{i,j}^0 + \lambda_i + \nu_j}{\varepsilon \theta} \right) + (1 - \theta) \exp \left(\frac{-(1 - \theta) w_{i,j}^1 + \mu_j - \nu_i}{\varepsilon (1 - \theta)} \right) \end{aligned} \quad (3.2)$$

Making the variable change

$$A^i = \exp \left(\frac{-w^i}{\varepsilon} \right); \alpha = \exp \left(\frac{\lambda}{\varepsilon \theta} \right); \beta = \exp \left(\frac{\mu}{\varepsilon (1 - \theta)} \right); \gamma = \exp \left(\frac{-\nu}{\varepsilon \theta (1 - \theta)} \right)$$

the optimality conditions for this unconstrained problem constitute a scaling problem for the matrices A^0 and A^1 that is finding positive vectors α, β, γ such that:

$$\alpha = \frac{f^0}{A^0 (\frac{1}{\gamma})^{1-\theta}}; \beta = \frac{f^1}{A^{1T} \gamma^\theta}; \gamma = \frac{A^{0T} \alpha}{A^1 \beta} \quad (3.3)$$

The link between primal and dual problems indicates that the optimal transport plans solving (3.1) are given by $X^0 = d(\alpha^*)A^0 d(\gamma^{*\theta-1}) \in \Pi(f^0, f^\theta)$ and $X^1 = d(\gamma^{*\theta})A^1 d(\beta^*) \in \Pi(f^\theta, f^1)$ where $\alpha^*, \beta^*, \gamma^*$ are solutions of the equations (3.3), and where the barycenter is finally given by $f^\theta = (\gamma^*)^{\theta-1}A^0 \alpha^* = (\gamma^*)^\theta A^1 \beta^*$. The strict concavity of problem (3.2) with respect to $\lambda_i + \nu_j$ and $\mu_j + \nu_i$ ensures the existence of a solution $(\alpha^*, \beta^*, \gamma^*)$ to this scaling problem as well as its uniqueness up to scalar multiple, meaning that all the other solutions are of the form $(r^{1-\theta}\alpha^*, r^\theta\beta^*, r\gamma)$ for some $r > 0$. As in the classical Sinkhorn algorithm, one can solve these equations by alternately updating α, β and γ according to one equation only. This procedure corresponds to the alternate maximizations on the dual problem (3.2), and constitutes the following Sinkhorn-like algorithm:

Algorithm I.2. Given two positive matrices $A^0, A^1 \in (\mathbb{R}_*^+)^{N \times N}$, and two marginals $f^0, f^1 \in \Sigma^N$, starting from $\alpha^{(0)} = \beta^{(0)} = \gamma^{(0)} = \mathbf{1}$, do for $n = 0, 1, \dots$

$$\alpha^{(n+1)} = \frac{f^0}{A^0(\frac{1}{\gamma^{(n)}})^{1-\theta}}; \beta^{(n+1)} = \frac{f^1}{A^1 \gamma^{(n)\theta}}; \gamma^{(n+1)} = \frac{A^0 \alpha^{(n+1)}}{A^1 \beta^{(n+1)}} \quad (3.4)$$

Remark I.4. We only consider this variant of the Sinkhorn algorithm in the case where A^0 and A^1 are positive matrices. However, one could consider this algorithm as well as the scaling problem (3.3) for general nonnegative matrices and wonder what pattern conditions are needed in that context. We wish to emphasize that thanks to the reduction Lemma I.6, our main result I.12 would extend to the context where the scaling problem (3.3) admits a solution.

From the point of view of primal iterates x^0, x^1 , one checks that these iterations still write as alternate \mathcal{KL} -projections. More precisely, iterations on α, β correspond to projections on the constraints $x^0 \mathbf{1} = f^0$ and $x^1 \mathbf{1} = f^1$ while iteration on γ corresponds to $x^0 \mathbf{1} = x^1 \mathbf{1}$. Similarly to what we did for the classical Sinkhorn algorithm in section 2.1, we can first analyze these iterations in the Hilbert metric using the properties given in Definition-Proposition I.1:

Theorem I.10. The Sinkhorn-like iterates $\alpha^{(n)}, \beta^{(n)}, \gamma^{(n)}$ converge in $(\mathbb{R}_*^+)^N / \sim$ to the fixed point $\alpha^*, \beta^*, \gamma^*$ with the estimate

$$\forall n \geq 0, d_H(\gamma^{(n+1)}, \gamma^*) \leq ((1-\theta)\kappa(A^0)^2 + \theta\kappa(A^1)^2)d_H(\gamma^{(n)}, \gamma^*)$$

Proof. Using in particular the third point of Definition-Proposition I.1 one gets

$$\begin{aligned} d_H(\alpha^{(n+1)}, \alpha^*) &\leq (1-\theta)\kappa(A^0)d_H(\gamma^{(n)}, \gamma^*) \\ d_H(\beta^{(n+1)}, \beta^*) &\leq \theta\kappa(A^1)d_H(\gamma^{(n)}, \gamma^*) \end{aligned}$$

And the fourth point gives

$$d_H(\gamma^{(n+1)}, \gamma^*) \leq \kappa(A^0)d_H(\alpha^{(n+1)}, \gamma^*) + \kappa(A^1)d_H(\beta^{(n+1)}, \beta^*)$$

which leads to the result. \square

To state our following convergence rates, we rely again on the rescaled variables that satisfy the same iteration rules with A^0, A^1 replaced by the limit transport plan matrices X^0, X^1 . The fact that $X^0 \neq X^1$ even when $A^0 = A^1$ motivated the choice of taking two different ground costs w^0, w^1 .

Lemma I.8. *Let $\alpha^{(n)}, \beta^{(n)}, \gamma^{(n)}$ be the Sinkhorn-like iterates defined by (3.4), and let $\alpha^*, \beta^*, \gamma^*$ be a solution of the scaling problem (3.3). Denote the corresponding transport plan matrices by $X^0 = d(\alpha^*)A^0 d(\frac{1}{\gamma^{*\theta}}) \in \Pi(f^0, f^\theta)$ and $X^1 = d(\gamma^{*\theta})A^1 d(\beta^*) \in \Pi(f^\theta, f^1)$, then the rescaled variables*

$$a^{(n)} = \frac{\alpha^{(n)}}{\alpha^*}; \quad b^{(n)} = \frac{\beta^{(n)}}{\beta^*}; \quad c^{(n)} = \frac{\gamma^{(n)}}{\gamma^*}$$

satisfy the iteration rules

$$a^{(n+1)} = \frac{1}{d(\frac{1}{f^0})X^0(\frac{1}{c^{(n)}})^{1-\theta}}; \quad b^{(n+1)} = \frac{1}{d(\frac{1}{f^1})X^1(c^{(n)})^\theta}; \quad c^{(n+1)} = \frac{X^0 a^{(n+1)}}{X^1 b^{(n+1)}} \quad (3.5)$$

The appearance of the stochastic matrices $M_a = d(\frac{1}{f^0})X^0$ and $M_b = d(\frac{1}{f^1})X^1$ leads to a min-max analysis similar to Theorem I.1:

Theorem I.11. *The rescaled variables $a^{(n)}, b^{(n)}, c^{(n)}$ converge to constant vectors $a^\infty \mathbf{1}, b^\infty \mathbf{1}, c^\infty \mathbf{1}$ and one has the following estimate:*

$$\forall n \geq 0, \quad \bar{c}^{(n+1)} - \underline{c}^{(n+1)} \leq (1 - \theta m_b - (1 - \theta) \frac{\underline{c}^{(0)}}{\bar{c}^{(0)}} m_a) (\bar{c}^{(n)} - \underline{c}^{(n)})$$

where $m_a = \underline{M}_a$ and $m_b = \underline{M}_b$.

Proof. Denote $p^{(n)} = (c^{(n)})^\theta$ and $q^{(n)} = (c^{(n)})^{\theta-1}$. Applying Lemma I.2 to the iterates on a and b yields:

$$\begin{aligned} \underline{q}^{(n)} + m_a(\bar{q}^{(n)} - \underline{q}^{(n)}) &\leq \left(\frac{1}{a^{(n+1)}} \right) \leq \overline{\left(\frac{1}{a^{(n+1)}} \right)} \leq \bar{q}^{(n)} - m_a(\bar{q}^{(n)} - \underline{q}^{(n)}) \\ \underline{p}^{(n)} + m_b(\bar{p}^{(n)} - \underline{p}^{(n)}) &\leq \left(\frac{1}{b^{(n+1)}} \right) \leq \overline{\left(\frac{1}{b^{(n+1)}} \right)} \leq \bar{p}^{(n)} - m_b(\bar{p}^{(n)} - \underline{p}^{(n)}) \end{aligned}$$

Besides, as $X^0 \mathbf{1} = X^1 \mathbf{1} = f^\theta$ one has

$$c^{(n+1)} = \frac{X^0 a^{(n+1)}}{X^1 b^{(n+1)}} \leq \frac{X^0 \bar{a}^{(n+1)} \mathbf{1}}{X^1 \underline{b}^{(n+1)} \mathbf{1}} = \frac{\bar{a}^{(n+1)}}{\underline{b}^{(n+1)}} \frac{X^0 \mathbf{1}}{X^1 \mathbf{1}} = \frac{\bar{a}^{(n+1)}}{\underline{b}^{(n+1)}} \mathbf{1}$$

so that $\bar{c}^{(n+1)} \leq \frac{\bar{a}^{(n+1)}}{\underline{b}^{(n+1)}}$ and similarly $\underline{c}^{(n+1)} \geq \frac{\underline{a}^{(n+1)}}{\bar{b}^{(n+1)}}$. Together with the previous

inequalities, this gives:

$$\bar{c}^{(n+1)} \leq \frac{\bar{p}^{(n)} - m_b(\bar{p}^{(n)} - \underline{p}^{(n)})}{\bar{q}^{(n)} + m_a(\bar{q}^{(n)} - \underline{q}^{(n)})} = \bar{c}^{(n)} \frac{1 - m_b \left[1 - \left(\frac{\underline{c}^{(n)}}{\bar{c}^{(n)}} \right)^\theta \right]}{1 + m_a \left[\left(\frac{\bar{c}^{(n)}}{\underline{c}^{(n)}} \right)^{1-\theta} - 1 \right]} \quad (3.6)$$

$$\underline{c}^{(n+1)} \geq \frac{\underline{p}^{(n)} + m_b(\bar{p}^{(n)} - \underline{p}^{(n)})}{\bar{q}^{(n)} - m_a(\bar{q}^{(n)} - \underline{q}^{(n)})} = \underline{c}^{(n)} \frac{1 + m_b \left[\left(\frac{\bar{c}^{(n)}}{\underline{c}^{(n)}} \right)^\theta - 1 \right]}{1 - m_a \left[1 - \left(\frac{\underline{c}^{(n)}}{\bar{c}^{(n)}} \right)^{1-\theta} \right]} \quad (3.7)$$

Then first exploit (3.6) just saying that for any $u, v \geq 0$ one has $\frac{1-u}{1+v} \leq 1-u$ to get:

$$\begin{aligned} \bar{c}^{(n+1)} &\leq \bar{c}^n - m_b(\bar{c}^{(n)} - (\bar{c}^{(n)})^{1-\theta}(\underline{c}^{(n)})^\theta) \\ &\leq \bar{c}^n - m_b(\bar{c}^{(n)} - (1-\theta)\bar{c}^{(n)} - \theta\underline{c}^{(n)}) \\ &\leq \bar{c}^{(n)} - \theta m_b(\bar{c}^{(n)} - \underline{c}^{(n)}) \end{aligned} \quad (3.8)$$

where we used that $x^\theta y^{1-\theta} \leq \theta x + (1-\theta)y$ for any $x, y > 0$. Second, exploit similarly (3.7) using $\frac{1+u}{1-v} \geq 1+v$ to get:

$$\begin{aligned} \underline{c}^{(n+1)} &\geq \underline{c}^{(n)} + m_a(\underline{c}^{(n)} - (\underline{c}^{(n)})^{2-\theta}(\bar{c}^{(n)})^{\theta-1}) \\ &= \underline{c}^{(n)} + \frac{\underline{c}^{(n)}}{\bar{c}^{(n)}} m_a(\bar{c}^{(n)} - (\underline{c}^{(n)})^{1-\theta}(\bar{c}^{(n)})^\theta) \\ &\geq \underline{c}^{(n)} + \frac{\underline{c}^{(n)}}{\bar{c}^{(n)}} m_a(1-\theta)(\bar{c}^{(n)} - \underline{c}^{(n)}) \\ &\geq \underline{c}^{(n)} + \frac{\underline{c}^{(0)}}{\bar{c}^{(0)}} m_a(1-\theta)(\bar{c}^{(n)} - \underline{c}^{(n)}) \end{aligned} \quad (3.9)$$

To obtain the last inequality (3.9) we used the fact that $(\bar{c}^{(n)})$ is nonincreasing while $(\underline{c}^{(n)})$ is nondecreasing. This indeed comes from equations (3.6) and (3.7); moreover it first provides that these sequences converge so that the estimate stated in the theorem – that derives from equations (3.8) (3.9) – proves that $(c^{(n)})$ converges to some constant $c^\infty \mathbf{1}$. The convergence of $(a^{(n)})$ and $(b^{(n)})$ towards constant vectors follows. \square

Again, these two linear rates are pessimistic. Similar to the observations we made in Figure 2.1 and following the same setting with $\theta = 0.5$, we show in Figure 3.2 an example in which the evolution of the error made by the algorithm on the variable γ is governed by $\lambda = 0.320$ while the Hilbert distance and min-max guaranties are respectively of 0.862 and 0.998.

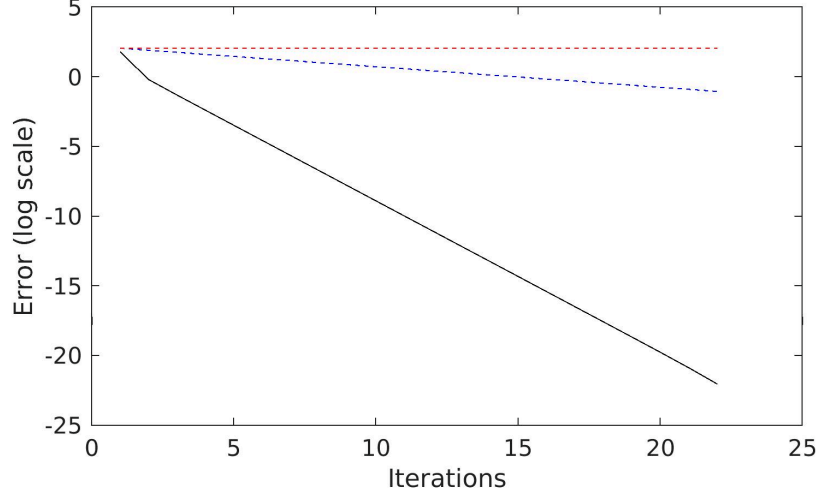


Figure 3.2 – Error evolution of the Sinkhorn algorithm for the barycenter (plain black) vs Hilbert distance (dotted red) and min-max (dotted blue) rates

To obtain a rate similar to Theorem I.8, we will fit into the general Theorem I.7 thanks to another lemma, similar to I.5, that expresses the concavity of $t \mapsto t^\theta$ for $\theta \in (0, 1)$:

Lemma I.9. *Let $M \in (\mathbb{R}^+)^{N \times N}$ be a stochastic matrix. Then for any positive vector $x \in (\mathbb{R}_*^+)^N$, one has for all $\theta \in (0, 1)$*

$$M(x^\theta) \leq (Mx)^\theta$$

We also force the appearance of the stochastic matrices $d(\frac{1}{f^\theta})X^{0T}$ and $d(\frac{1}{f^\theta})X^1$ in the update rule of variable c :

$$c^{(n+1)} = \frac{X^{0T} a^{(n+1)}}{X^1 b^{(n+1)}} = \frac{d(\frac{1}{f^\theta})X^{0T} a^{(n+1)}}{d(\frac{1}{f^\theta})X^1 b^{(n+1)}}$$

so that making use of Lemmas I.5 and I.9, we get:

$$\begin{aligned} c^{(n+1)} &\leq \left(d\left(\frac{1}{f^\theta}\right)X^{0T} a^{(n+1)} \right) \left(d\left(\frac{1}{f^\theta}\right)X^1 \frac{1}{b^{(n+1)}} \right) \\ &\leq \left(d\left(\frac{1}{f^\theta}\right)X^{0T} d\left(\frac{1}{f^0}\right)X^0 (c^{(n)})^{1-\theta} \right) \left(d\left(\frac{1}{f^\theta}\right)X^1 d\left(\frac{1}{f^1}\right)X^{1T} (c^{(n)})^\theta \right) \\ &\leq \left(d\left(\frac{1}{f^\theta}\right)X^{0T} d\left(\frac{1}{f^0}\right)X^0 c^{(n)} \right)^{1-\theta} \left(d\left(\frac{1}{f^\theta}\right)X^1 d\left(\frac{1}{f^1}\right)X^{1T} c^{(n)} \right)^\theta \\ &\leq \left((1-\theta) d\left(\frac{1}{f^\theta}\right)X^{0T} d\left(\frac{1}{f^0}\right)X^0 + \theta d\left(\frac{1}{f^\theta}\right)X^1 d\left(\frac{1}{f^1}\right)X^{1T} \right) c^{(n)} \end{aligned}$$

i.e. $c^{(n+1)} \leq M c^{(n)}$ with $M = (1 - \theta) d(\frac{1}{f^\theta}) X^{0T} d(\frac{1}{f^0}) X^0 + \theta d(\frac{1}{f^\theta}) X^1 d(\frac{1}{f^1}) X^{1T}$ being a stochastic positive (hence primitive) matrix. One also sees that $d(f^\theta)M$ is symmetric; and similar reasoning shows that $\frac{1}{c^{(n+1)}} \leq M \frac{1}{c^{(n)}}$. Finally, we are exactly in the context of Theorem I.7, which allows to conclude the following theorem concerning the convergence of the variable γ with $s = \frac{f^\theta}{\gamma^{*2}}$:

Theorem I.12. *Let $A^0, A^1 \in (\mathbb{R}_*^+)^{N \times N}$ be two positive matrices and $f^0, f^1 \in \Sigma^N$. The Sinkhorn iterates $(\alpha^{(n)}, \beta^{(n)}, \gamma^{(n)})$ corresponding to (A^0, A^1, f^0, f^1) and defined by (3.4) converge to $\alpha^*, \beta^*, \gamma^* \in (\mathbb{R}_*^+)^N$ which are solutions of the scaling problem (3.3). In addition there exist a sequence of real numbers $(u_n) \in (\mathbb{R}_*^+)^N$ converging to 1, and a norm $\|\cdot\|_s$ on \mathbb{R}^N such that: $\forall \delta > 0, \exists n_\delta$ such that $\forall n \geq n_\delta$,*

$$\left\| \gamma^{(n+1)} - u_{n+1} \gamma^* \right\|_s \leq (\lambda_2 + \delta) \left\| \gamma^{(n)} - u_n \gamma^* \right\|_s$$

where $0 < \lambda_2 < 1$ is the subdominant eigenvalue of the matrix $M = (1 - \theta)M^0 + \theta M^1$ with $M^0 = d(\frac{1}{f^\theta}) X^{0T} d(\frac{1}{f^0}) X^0$ and $M^1 = d(\frac{1}{f^\theta}) X^1 d(\frac{1}{f^1}) X^{1T}$; and where we also denoted the transport plans between the marginals f^0, f^1 and the barycenter f^θ by $X^0 = d(\alpha^*) A^0 d(\frac{1}{\gamma^{*1-\theta}}) \in \Pi(f^0, f^\theta)$, $X^1 = d(\gamma^{*\theta}) A^1 d(\beta^*) \in \Pi(f^\theta, f^1)$.

Remark I.5. *During our min-max study of the algorithm, we saw in particular that*

$$\begin{aligned} (\underline{c}^{(n)})^{1-\theta} \mathbf{1} &\leq a^{(n+1)} \leq (\bar{c}^{(n)})^{1-\theta} \mathbf{1} \\ (\underline{c}^{(n)})^{-\theta} \mathbf{1} &\leq b^{(n+1)} \leq (\bar{c}^{(n)})^{-\theta} \mathbf{1} \end{aligned}$$

These inequalities allows one to expect the same convergence speed for α and β as stated for γ in the theorem.

This rate provided by Theorem I.12 is actually observed in practice as shown in Figure 3.3 where we display it together with the evolution of the error made on the variable γ for different parameter values.

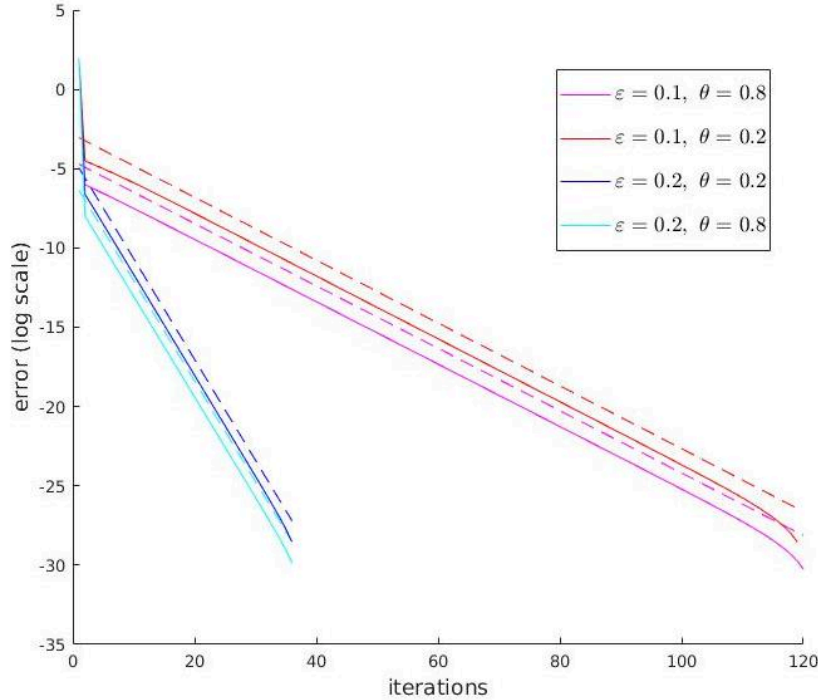


Figure 3.3 – Error evolution of the Sinkhorn algorithm for the barycenter (plain lines) and predicted λ_2 rate (dotted lines) for different values of ε and θ

3.1.2 Multiple barycenters

As the path of the barycenters $(f^\theta)_{\theta \in (0,1)}$ describes the Wasserstein geodesic between f^0 and f^1 , one might be interested in getting simultaneously several barycenters such as all the $f^{\frac{k}{K}}$ for $k \in \llbracket 1, K-1 \rrbracket$. For notational convenience, we now denote by f^0 and f^K the fixed marginals and by f^1, \dots, f^{K-1} the intermediate barycenters. A way of computing these barycenters¹ simultaneously is to solve the following problem:

$$\min_{f^1, \dots, f^{K-1} \in \Sigma^N} \sum_{k=0}^{K-1} W_{w^k}^\varepsilon(f^k, f^{k+1})$$

for some ground cost matrices $w^k \in \mathbb{R}^{N \times N}$. Expanding the minimas corresponding to the Wasserstein distances, this problem writes:

$$\min_{x^0, \dots, x^{K-1}} \sum_{k=0}^{K-1} \langle w^k | x^k \rangle + \varepsilon \langle x^k | \log x^k - \mathbf{1} \rangle$$

¹To be exact, even for $\varepsilon = 0$ the barycenter $f^{\frac{k}{K}}$ defined in this section's way is *not* equal to the barycenter f^θ for $\theta = \frac{k}{K}$ as defined “individually” above by $f^{\frac{k}{K}} = \arg \min_f \frac{k}{K} W_{w^1}(f^1, f) + \frac{K-k}{K} W_{w^2}(f, f^2)$. However, this two barycenters are close because those definitions coincide in a continuous setting (in the case where a transport map exists for instance).

where the minimum runs over $x^0, \dots, x^{K-1} \in (\mathbb{R}^+)^{N \times N}$ such that

$$x^0 \mathbf{1} = f^0; x^{K-1T} \mathbf{1} = f^K; \forall k \in \llbracket 0, K-2 \rrbracket, x^{k+1} \mathbf{1} = x^k T \mathbf{1}$$

Introducing the Lagrange multipliers μ^0 for $x^0 \mathbf{1} = f^0$, μ^K for $x^{K-1T} \mathbf{1} = f^K$ and μ^k for $x^{k+1} \mathbf{1} = x^k T \mathbf{1}$, one derives the dual problem using again Proposition I.1:

$$\sup_{\mu^0, \dots, \mu^K \in \mathbb{R}^N} \langle \mu^0 | f^0 \rangle - \langle \mu^K | f^K \rangle - \varepsilon \sum_{k=0}^{K-1} \sum_{i,j} \exp \left(\frac{-w_{i,j}^k + \mu_i^k - \mu_j^{k+1}}{\varepsilon} \right)$$

Making our usual variable change

$$A^k = \exp \left(\frac{-w^k}{\varepsilon} \right); \gamma_k = \exp \left(\frac{\mu^k}{\varepsilon} \right)$$

the optimality conditions for this unconstrained problem constitute a scaling problem of the matrices A^k that is finding positive vectors γ_k such that:

$$\gamma_0 = \frac{f^0}{A^0 \frac{1}{\gamma_1}}; \gamma_K = \frac{A^{K-1T} \gamma_{K-1}}{f^K}; \forall k \in \llbracket 1, K-1 \rrbracket, \gamma_k = \sqrt{\frac{A^{k-1T} \gamma_{k-1}}{A^k \frac{1}{\gamma_{k+1}}}} \quad (3.10)$$

For simplicity we now suppose that $K = 2P$ is even. We conduct the iterations of the associated Sinkhorn-like algorithm in the following order:

Algorithm I.3. Given $K = 2P$ positive matrices $A^0, \dots, A^{K-1} \in (\mathbb{R}_*^+)^{N \times N}$ and two marginals $f^0, f^K \in \Sigma^N$, starting from $\gamma_k^{(0)} = \mathbf{1}$ for all $k \in \llbracket 0, K \rrbracket$, do for $n = 0, 1, \dots$

Step A:

$$\gamma_0^{(n+1)} = \frac{f^0}{A^0 \frac{1}{\gamma_1^{(n)}}}; \gamma_K^{(n+1)} = \frac{A^{K-1T} \gamma_{K-1}^{(n)}}{f^K};$$

$$\forall k \in \llbracket 1, P-1 \rrbracket, \gamma_{2k}^{(n+1)} = \sqrt{\frac{A^{2k-1T} \gamma_{2k-1}^{(n)}}{A^{2k} \frac{1}{\gamma_{2k+1}^{(n)}}}}$$

Step B:

$$\forall k \in \llbracket 0, P-1 \rrbracket, \gamma_{2k+1}^{(n+1)} = \sqrt{\frac{A^{2kT} \gamma_{2k}^{(n+1)}}{A^{2k+1} \frac{1}{\gamma_{2k+2}^{(n+1)}}}}$$

Again, step A and B appear as Bregman projections on primal transport plans. It is also still possible to analyze these iterations in terms of Hilbert distance (showing that $\max_k d_H(\gamma_{2k+1}^{(n)}, \gamma_{2k+1}^*)$ converges linearly to 0 with rate κ^2) as well as in terms of min-max. However we focus on the asymptotic theorem:

Theorem I.13. Let $A^0, \dots, A^{K-1} \in (\mathbb{R}_*^+)^{N \times N}$ be positive matrices, and let $f^0, f^K \in \Sigma^N$. Let $\gamma^{(n)} = (\gamma_k^{(n)})_{0 \leq k \leq K}$ be the sequence obtained by algorithm I.3, and denote $\gamma_{[0]}^{(n)}$ (respectively $\gamma_{[1]}^{(n)}$) the concatenation of the $\gamma_k^{(n)}$ for k even (respectively odd). Then $(\gamma^{(n)})$ converges to a solution $\gamma^* = (\gamma_{[0]}^*, \gamma_{[1]}^*)$ of the scaling problem (3.10) and there exist $s \in (\mathbb{R}_*^+)^{NP}$, a sequence of real numbers (u_n) converging to 1, and $\lambda_2 \in (0, 1)$ such that

$$\forall \delta > 0, \exists n_\delta \in \mathbb{N} : \forall n \geq n_\delta, \left\| \gamma_{[1]}^{(n+1)} - u_{n+1} \gamma_{[1]}^* \right\|_s \leq (\lambda_2 + \delta) \left\| \gamma_{[1]}^{(n)} - u_n \gamma_{[1]}^* \right\|_s$$

Proof. The rescaled variables $c_k^{(n)} = \frac{\gamma_k^{(n)}}{\gamma_k^*}$ – where $(\gamma_k^*)_k$ is a solution of the scaling problem (3.10) (such a solution exists by strict concavity of the dual problem) – satisfy the same iterations than the initial variables γ_k , but with the matrices A^k being replaced by the transport plan matrices:

$$X^k = d(\gamma_k^*) A^k d\left(\frac{1}{\gamma_{k+1}^*}\right) \in \Pi(f^k, f^{k+1}) \text{ for } 0 \leq k \leq K-1$$

Lemma I.5 gives

$$\begin{aligned} c_0^{(n+1)} &\leq d\left(\frac{1}{f^0}\right) X^0 c_1^{(n)} ; c_K^{(n+1)} \leq d\left(\frac{1}{f^K}\right) X^{K-1T} c_{K-1}^{(n)} \\ \forall k \in \llbracket 1, P-1 \rrbracket, c_{2k}^{(n+1)} &\leq \frac{1}{2} \left(d\left(\frac{1}{f^{2k}}\right) X^{2k-1T} c_{2k-1}^{(n)} + d\left(\frac{1}{f^{2k}}\right) X^{2k} c_{2k+1}^{(n)} \right) \\ \forall k \in \llbracket 0, P-1 \rrbracket, c_{2k+1}^{(n+1)} &\leq \frac{1}{2} \left(d\left(\frac{1}{f^{2k+1}}\right) X^{2kT} c_{2k}^{(n+1)} + d\left(\frac{1}{f^{2k+1}}\right) X^{2k+1} c_{2k+2}^{(n+1)} \right) \end{aligned} \quad (3.11)$$

Denote by $c_{[0]}^{(n)}$ the concatenation of the $c_{2k}^{(n)}$ for $k = 0, \dots, P$ and by $c_{[1]}^{(n)}$ the concatenation of the $c_{2k+1}^{(n)}$ for $k = 0, \dots, P-1$. We deduce from the inequalities (3.11) that $c_{[1]}^{(n+1)} \leq M c_{[1]}^{(n)}$ where $M \in (\mathbb{R}^+)^{NP \times NP}$ is the block tridiagonal matrix defined by

$$M = \begin{pmatrix} M^0 & N^0 & & & \\ L^1 & M^1 & N^1 & & \\ & \ddots & \ddots & \ddots & \\ & & L^{P-2} & M^{P-2} & N^{P-2} \\ & & & L^{P-1} & M^{P-1} \end{pmatrix}$$

with

$$\begin{aligned}
 M^0 &= \frac{1}{2} d\left(\frac{1}{f^1}\right) X^{0T} d\left(\frac{1}{f^0}\right) X^0 + \frac{1}{4} d\left(\frac{1}{f^1}\right) X^1 d\left(\frac{1}{f^2}\right) X^{1T} \\
 N^0 &= \frac{1}{4} d\left(\frac{1}{f^1}\right) X^1 d\left(\frac{1}{f^2}\right) X^2 \\
 L^{P-1} &= \frac{1}{4} d\left(\frac{1}{f^{K-1}}\right) X^{K-2T} d\left(\frac{1}{f^{K-2}}\right) X^{K-3T} ; \\
 M^{P-1} &= \frac{1}{4} d\left(\frac{1}{f^{K-1}}\right) X^{K-2T} d\left(\frac{1}{f^{K-2}}\right) X^{K-2} + \frac{1}{2} d\left(\frac{1}{f^{K-1}}\right) X^{K-1} d\left(\frac{1}{f^K}\right) X^{K-1T}
 \end{aligned}$$

as well as $\forall 1 \leq k \leq P - 2$,

$$\begin{aligned}
 L^k &= \frac{1}{4} d\left(\frac{1}{f^{2k+1}}\right) X^{2kT} d\left(\frac{1}{f^{2k}}\right) X^{2k-1T} ; \\
 M^k &= \frac{1}{4} \left(d\left(\frac{1}{f^{2k+1}}\right) X^{2kT} d\left(\frac{1}{f^{2k}}\right) X^{2k} + d\left(\frac{1}{f^{2k+1}}\right) X^{2k+1} d\left(\frac{1}{f^{2k+2}}\right) X^{2k+1T} \right) ; \\
 N^k &= \frac{1}{4} d\left(\frac{1}{f^{2k+1}}\right) X^{2k+1} d\left(\frac{1}{f^{2k+2}}\right) X^{2k+2}
 \end{aligned}$$

Similarly one also obtains $\frac{1}{c_{[1]}^{(n+1)}} \leq M \frac{1}{c_{[1]}^{(n)}}$. Denote by s the concatenation of the f^{2k+1} for $k = 0, \dots, P - 1$ (divided by P to get $s \in \Sigma^{PN}$). One checks that $d(s)M$ is symmetric. Furthermore, M is tridiagonal hence primitive and M is stochastic because $X^k \in \Pi(f^k, f^{k+1})$.

To conclude, the convergence of $\gamma_{[1]}$ and the announced rate are obtained using again theorem I.7. For the convergence of $\gamma_{[0]}$, note that inequalities (3.11) also easily lead to $c_{[0]}^{(n+1)} \leq \overline{c_{[1]}^{(n)}} \mathbf{1}$ and similarly one has $\frac{1}{c_{[0]}^{(n+1)}} \leq \frac{1}{\overline{c_{[1]}^{(n)}}} \mathbf{1}$ hence

$$\underline{c_{[1]}^{(n)}} \mathbf{1} \leq c_{[0]}^{(n+1)} \leq \overline{c_{[1]}^{(n)}} \mathbf{1}$$

so that the convergence of $\gamma_{[0]}$ follows from the one of $\gamma_{[1]}$. □

Numerical experiments confirm once again the validity of this rate in practice, see Figure 3.4 where we use our usual setting for different values of P .

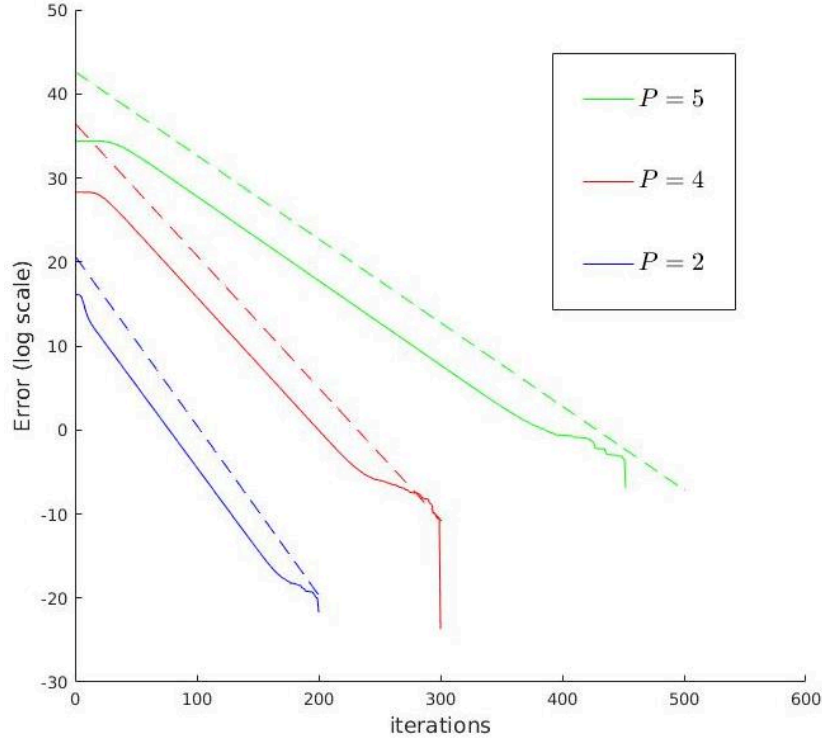


Figure 3.4 – Error evolution of the Sinkhorn algorithm for multiple barycenters (plain lines) and predicted λ_2 rates (dotted lines) for different values of P

3.2 Graph labelling problem

In this section we consider an assignment problem, close to problem (1.12), that will lead to a Sinkhorn-like algorithm on a undirected graph. This section relies on an unpublished work of Caillaud, Chambolle, Pock, for which the connection with Message Passing algorithms has to be investigated further.

3.2.1 General setting

Consider an undirected graph $(\mathcal{V}, \mathcal{E})$ – \mathcal{V} is a finite set of vertices and $\mathcal{E} \subset \mathcal{P}_2(\mathcal{V})$ (the set of 2-elements subsets of \mathcal{V}) a set of edges – and a set of labels $L = \{1, \dots, N\}$. We want to assign a label $\ell_i \in L$ to each vertex $i \in \mathcal{V}$ while solving the following minimization problem:

$$\min_{(\ell_i) \in L^{\mathcal{V}}} \sum_{i \in \mathcal{V}} \theta_i(\ell_i) + \sum_{\{i,j\} \in \mathcal{E}} \theta_{i,j}(\ell_i, \ell_j) \quad (3.12)$$

where $\theta_i : L \rightarrow \mathbb{R}$, $\theta_{i,j} : L \times L \rightarrow \mathbb{R}$ are some cost functions such that $\theta_{i,j}(\ell, m) = \theta_{j,i}(m, \ell)$ for our notations to make sense. We will denote $j \sim i$ when $\{i, j\} \in \mathcal{E}$, and $n_i = |\{j \in \mathcal{V}, j \sim i\}|$ will stand for the number of edges at vertex i .

Such energy minimization problems appear widely in Image Processing tasks. In that context, one takes $\mathcal{V} = \llbracket 1, N_1 \rrbracket \times \llbracket 1, N_2 \rrbracket$ the set of pixels of an image, and chooses for \mathcal{E} a set of edges corresponding to a notion of neighbor or linked pixels. Then through the choice of the functions θ one can express any minimization problem involving a term centered on the pixels (with the θ_i for $i \in \mathcal{V}$) and a term centered on the links between pixels (with the $\theta_{i,j}$ for $\{i, j\} \in \mathcal{E}$). In the context of Semantic Segmentation, the label set L can consist in classes of objects that one must identify in the picture; for Image Restoration, $L \subset [0, 1]$ can represent gray levels of the pixels with 0 corresponding to “black” and 1 to “white”. We refer to [AZJ⁺18] for an overview on applications.

To illustrate the expressiveness of problem (3.12), we detail one choice of functions $\theta_i, \theta_{i,j}$ that encodes a famous image processing model: the ROF image denoising problem, introduced by Rudin, Osher and Fatemi in [ROF92]. We refer to our introduction of part II of this manuscript for more details on the ROF model. In this setting, the label set $L = \{1, \dots, N\}$ corresponds to the possible values of the pixels: for instance one can choose that pixel $i \in \mathcal{V}$ is set to label $\ell \in L$ when it has value $\frac{\ell}{N} \in [0, 1]$. The ROF model aims at denoising a noisy image $f \in L^{\mathcal{V}}$ by minimization of a combination of two terms: a fidelity term to the data which is the pixel-centered term, given by

$$\forall i \in \mathcal{V}, \forall \ell \in L, \theta_i(\ell) = \frac{1}{2} \left| \frac{\ell}{N} - f_i \right|^2$$

and a regularity term known as the total variation which penalizes the differences between neighbor pixels, given by:

$$\forall \{i, j\} \in \mathcal{E}, \forall (\ell, k) \in L^2, \theta_{i,j}(\ell, k) = \frac{1}{\lambda} \left| \frac{\ell}{N} - \frac{k}{N} \right|$$

In this setting, one can use for \mathcal{E} the standard 4-neighborhood configuration, meaning that each vertex is linked to its four closest neighbors. In addition, the parameter $\lambda > 0$ describes the trade-off between fidelity and regularization; when it is set to an appropriate value, solving (3.12) allows one to achieve the denoising of the noisy image f , see Figure 3.5.

Similarly to the relaxation process for problem (1.12), one can reformulate problem (3.12) as a linear program with marginal constraints by introducing probability densities matrices. Indeed, defining new variables (v_i^ℓ) (respectively $w_{i,j}^{\ell,m}$) that equals 1 when $\ell_i = \ell$ (respectively $(\ell_i, \ell_j) = (\ell, m)$) and 0 otherwise, one sees that our problem writes as

$$\min \sum_{i,\ell} \theta_i(\ell) v_i^\ell + \sum_{\{i,j\}, \ell, m} \theta_{i,j}(\ell, m) w_{i,j}^{\ell, m} \quad (3.13)$$

where the minimum runs over variables $v \in (\mathbb{R}^+)^{\mathcal{V} \times L}$ and $w \in (\mathbb{R}^+)^{\mathcal{E} \times L \times L}$ (with the convention $w_{i,j}^{\ell,m} = w_{j,i}^{m,\ell}$) such that

$$\forall i \in \mathcal{V}, \sum_{\ell} v_i^{\ell} = 1; \forall \{i, j\} \in \mathcal{E}, \forall m \in L, \sum_{\ell} w_{i,j}^{\ell,m} = v_j^m \quad (3.14)$$

From the solution v^* of this new minimization problem, we will finally recover the optimal assignment of our original problem (3.12), by letting $\ell_i = \arg \max_{\ell} v_i^{*\ell}$ (or $\ell_i \in \arg \max_{\ell} v_i^{*\ell}$ if there exist several optimal assignments).

Once again, we deal with an entropic regularization of this linear problem. We introduce two regularizing terms that may be of different amplitudes, one for the variable w and one for the variable v , hence we write for two parameters $\varepsilon, \beta > 0$:

$$\min_{\substack{(v,w) \\ \text{s.t. (3.14)}}} \sum_{i,\ell} \theta_i(\ell) v_i^{\ell} + \beta \varepsilon v_i^{\ell} \log v_i^{\ell} + \sum_{\{i,j\},\ell,m} \theta_{i,j}(\ell, m) w_{i,j}^{\ell,m} + \varepsilon w_{i,j}^{\ell,m} \log w_{i,j}^{\ell,m} \quad (3.15)$$

We introduce the Lagrange multiplier $\lambda_{i,j}^m$ for the constraint $\sum_{\ell} w_{i,j}^{\ell,m} = v_j^m$. Unlike the previous dualization processes, we replace the use of Proposition I.1 to keep our computations on the spaces of variables v and w such that $\sum_{\ell} v_i^{\ell} = 1$ for all $i \in \mathcal{V}$ and $\sum_{\ell,m} w_{i,j}^{\ell,m} = 1$ for all $\{i, j\} \in \mathcal{E}$. Formally, we use the following well-known fact:

Proposition I.5. *For any $a \in \mathbb{R}^d$ and $\eta > 0$ one has:*

$$\min_{x \geq 0 \text{ s.t. } \langle x | \mathbf{1} \rangle = 1} \langle a | x \rangle + \eta \langle x | \log x \rangle = -\eta \log \sum_k \exp\left(-\frac{a_k}{\eta}\right)$$

Proof. Adding the Lagrange multiplier ν for the constraint $\langle x | \mathbf{1} \rangle = 1$, one finds that the value of this problem equals

$$\sup_{\nu \in \mathbb{R}} \nu - \eta \exp\left(\frac{\nu}{\eta}\right) e^{-1} \sum_k \exp\left(-\frac{a_k}{\eta}\right)$$

via $x = \exp\left(-1 + \frac{\nu - a}{\eta}\right)$. Then the optimal value of ν is $\nu = \eta - \eta \log \sum_k \exp\left(-\frac{a_k}{\eta}\right)$; this leads to the announced result which corresponds to $x = \frac{\exp\left(-\frac{a}{\eta}\right)}{\langle \exp\left(-\frac{a}{\eta}\right) | \mathbf{1} \rangle}$. \square

Doing so, we are led to the following dual problem of (3.15):

$$\begin{aligned} -\varepsilon \min_{\lambda \in \mathbb{R}^{\mathcal{E} \times L}} \beta \sum_i \log \sum_{\ell} \exp\left(-\frac{\theta_i(\ell) + \sum_{j:j \sim i} \lambda_{j,i}^{\ell}}{\varepsilon \beta}\right) \\ + \sum_{\{i,j\}} \log \sum_{\ell,m} \exp\left(\frac{\lambda_{i,j}^m + \lambda_{j,i}^{\ell} - \theta_{i,j}(\ell, m)}{\varepsilon}\right) \end{aligned}$$

Note that this dual energy has the same value when one replaces any $\lambda_{j,i}^\ell$ by $\lambda_{j,i}^\ell + c_{j,i} \mathbf{1}$ for some $c_{j,i} \in \mathbb{R}$. This will be reflected in the optimality conditions that we will express in our usual exponential variables

$$y_i = \exp\left(-\frac{\theta_i}{\varepsilon}\right); \alpha_{i,j} = \exp\left(\frac{\lambda_{i,j}}{\varepsilon}\right); X_{i,j} = \exp\left(-\frac{\theta_{i,j}}{\varepsilon}\right)$$

Note for future reasoning that as $\theta_{i,j}(\ell, m) = \theta_{j,i}(m, \ell)$, the matrices $X_{i,j}$ are such that: $X_{i,j}^T = X_{j,i}$. When differentiating the dual objective with respect to $\lambda_{j,i}^\ell$, one computes that $\alpha_{j,i}^\ell$ is optimal if and only if

$$\frac{\left(\frac{y_i}{\prod_{j':i \sim j'} \alpha_{j',i}}\right)^{\frac{1}{\beta}}}{\left\langle \left(\frac{y_i}{\prod_{j':i \sim j'} \alpha_{j',i}}\right)^{\frac{1}{\beta}} \mid \mathbf{1} \right\rangle} = \frac{\alpha_{j,i}(X_{i,j} \alpha_{i,j})}{\langle \alpha_{j,i}(X_{i,j} \alpha_{i,j}) \mid \mathbf{1} \rangle}$$

which in other words states that $\alpha_{j,i}(X_{i,j} \alpha_{i,j})$ and $Z_i^{\frac{1}{\beta}}$ where $Z_i = y_i \prod_{j':j' \sim i} \alpha_{j',i}^{-1}$ are multiple vectors. As any scalar normalization of the $\alpha_{j,i}$ remains an optimum, we chose to seek for α such that $Z_i^{\frac{1}{\beta}} = \alpha_{j,i}(X_{i,j} \alpha_{i,j})$. Taking the product over $j \sim i$ leads to $Z_i^{\frac{n_i}{\beta}} = \frac{y_i}{Z_i} \prod_{j:j \sim i} (X_{i,j} \alpha_{i,j})$ where we recall that $n_i = |\{j \in \mathcal{V}, j \sim i\}|$ is the number of edges at vertex i . This allows one to express Z_i in terms of the $\alpha_{i,j}$ for $j \sim i$ and we obtain finally the following optimality conditions that constitute a scaling problem for the matrices $X_{i,j}$:

$$\forall \{i, j\} \in \mathcal{E}, \alpha_{j,i} X_{i,j} \alpha_{i,j} = \left(y_i \prod_{j':i \sim j'} X_{i,j'} \alpha_{i,j'} \right)^{\frac{1}{n_i + \beta}} \quad (3.16)$$

This calculation being done, let us stop briefly for some remarks. First, the entropic regularization we introduced in (3.15) is known as the Bethe (free) energy in communities of Conditional (or Markov) Random Fields. We refer the reader to the second chapter of the book [WJ08], where Wainwright and Jordan explain in several ways how the minimization problem (3.13) arises in these settings, see also [Hes06, YFW05]. Second, and quite surprisingly at first sight, this energy can actually be considered not only for $\beta > 0$ but for $\beta > -n_i$ for all $i \in \mathcal{V}$, while remaining strictly convex. Indeed, this derives from writing the entropy terms under the following form thanks to (3.14):

$$\beta v_i^\ell \log v_i^\ell + \sum_{j \sim i, m} w_{i,j}^{\ell, m} \log w_{i,j}^{\ell, m} = \sum_{j \sim i, m} w_{i,j}^{\ell, m} \log \frac{w_{i,j}^{\ell, m}}{v_i^\ell} + (\beta + n_i) v_i^\ell \log v_i^\ell$$

and noticing that the function $(x, y) \mapsto x \log \frac{x}{y}$ is strictly convex in the domain $0 < x < y$. One can find more conditions for convexity of the general Bethe energy on graphs in [PA02, MJGF09]. Second, although our previous calculations required $\beta > 0$

to be rigorous, our final optimality conditions (3.16) still make sense for $\beta > -n_i$. Moreover, one can check that they express the cancellation of the derivatives of the Lagrangian associated to (3.15), so that solving these equations still lead to the solution of (3.15) for admissible values of $\beta \leq 0$. Finally, different algorithms can be proposed to find the $\alpha_{i,j}$ factors satisfying (3.16). In the context of Belief Propagation or Message Passing introduced in [Pea88], the current values of $\alpha_{i,j}$ are seen as pieces of information that can be transmitted from a vertex to its neighbors by updating $\alpha_{j,i}$ thanks to $\alpha_{i,j'}$ for $j' \sim i$. The corresponding update rule

$$\alpha_{j,i} = \frac{\mathbf{1}}{X_{i,j}\alpha_{i,j}} \left(y_i \prod_{j':i \sim j'} X_{i,j'} \alpha_{i,j'} \right)^{\frac{1}{n_i + \beta}} \quad (3.17)$$

leads to so-called Sum-Product algorithms (the “sum” standing for the matrix-vector product $X_{i,j'} \alpha_{i,j'}$), and the underlying question is now to decide how to organize the traversal of the graph $(\mathcal{V}, \mathcal{E})$, that is, in which order iterations (3.17) should be run, to reach a fixed point solving (3.16). In this view, some standard schemes resemble dynamic programming procedures relying on the progression of “solved” parts of the graph for which values do not change after some time. On the contrary, the update rules we propose below, inspired by Sinkhorn-like algorithms and alternate Bregman projection, impacts half the vertices of the graph at every step. It is close to the method proposed by Kushinsky and co-authors in [KMDL19].

We want to perform alternate iterations solving (3.16) that correspond to alternate Bregman projections on two groups of constraints of the perturbed linear program (3.15). To do so we make the further assumption that our graph $(\mathcal{V}, \mathcal{E})$ is bipartite, meaning that its vertices can be split into two sets, $\mathcal{V} = \mathcal{V}_1 \sqcup \mathcal{V}_2$ such that for every edge $\{i, j\} \in \mathcal{E}$, either $(i, j) \in \mathcal{V}_1 \times \mathcal{V}_2$ or $(i, j) \in \mathcal{V}_2 \times \mathcal{V}_1$. This allows to alternately solve equations (3.16) by deducing the variables $(\alpha_{j,i})_{j \sim i}$ for $i \in \mathcal{V}_p$ from the variables $(\alpha_{j',i'})_{j' \sim i'}$ for $i' \in \mathcal{V} \setminus \mathcal{V}_p$. This procedure constitutes the following Sinkhorn-like algorithm:

Algorithm I.4. *Given a bipartite undirected graph $(\mathcal{V} = \mathcal{V}_1 \sqcup \mathcal{V}_2, \mathcal{E})$, positive matrices $X_{i,j} \in (\mathbb{R}_*^+)^{N \times N}$ for $\{i, j\} \in \mathcal{E}$, $y_i \in (\mathbb{R}_*^+)^N$ and $\beta > -\min_{i \in \mathcal{V}} n_i$, where n_i denotes the number of edges at vertex $i \in \mathcal{V}$, starting from vectors $\alpha_{i,j}^{(0)} = \mathbf{1}$ for all $\{i, j\} \in \mathcal{E}$, do for $n = 0, 1, \dots$*

Step A:

$$\forall i \in \mathcal{V}_1, \forall j \sim i, \alpha_{j,i}^{(n+1)} = \frac{\left(y_i \prod_{j':i \sim j'} X_{i,j'} \alpha_{i,j'}^{(n)} \right)^{\frac{1}{n_i + \beta}}}{X_{i,j} \alpha_{i,j}^{(n)}}$$

Step B:

$$\forall i \in \mathcal{V}_2, \forall j \sim i, \alpha_{j,i}^{(n+1)} = \frac{\left(y_i \prod_{j':i \sim j'} X_{i,j'} \alpha_{i,j'}^{(n+1)} \right)^{\frac{1}{n_i + \beta}}}{X_{i,j} \alpha_{i,j}^{(n+1)}}$$

Following this procedure on the 2D grid graph with the 4-neighbor configuration (which is indeed bipartite: we split pixels $(i_x, i_y) \in \mathcal{V} = \llbracket 1, N_1 \rrbracket \times \llbracket 1, N_2 \rrbracket$ into sets $\mathcal{V}_1, \mathcal{V}_2$ depending on the parity of $i_x + i_y$), one can solve the ROF denoising model we introduced above. We present in Figure 3.5 the results on a test image with a label set L of size $N = 50$ using trade-off parameter $\lambda = 10$ and regularization parameters $\varepsilon = 10^{-4}, \beta = 0$.



Figure 3.5 – Denoised version (right) of a noisy image (left) obtained through the Sinkhorn-like algorithm I.4

3.2.2 Case of 1D graphs

In this section, we present the linear convergence results for the Sinkhorn-like algorithm I.4 on one dimensional graphs, that is when each vertex has at most two neighbors: $\forall i \in \mathcal{V}, n_i \in \{1, 2\}$. With this arity restriction, the graph is then either a “cycle” or a “straight line” on a set of vertices $\mathcal{V} = \{1, \dots, K\}$, meaning that the set of edges is given by $\mathcal{E} = \{\{i, i+1\}, i \in \mathcal{V}\}$ with indices taken modulo K (this is the cycle case) or by $\mathcal{E} = \{\{i, i+1\}, i \in \{1, \dots, K-1\}\}$ (this is the straight line case). The parameter β must also satisfy $\beta > -2$ or $\beta > -1$. In addition, we assume (as it is mandatory in the cycle case for our graph to be bipartite) the number of vertices is even and write $K = 2P$. We present our results in the case of the cycle and leave to remarks the case of the line which consists only of slight adjustments.

In the case of the cycle, algorithm I.4 simplifies into:

Algorithm I.5. Starting from $\alpha_{i,j}^{(0)} = \mathbf{1}$ for all $\{i, j\} \in \mathcal{E}$, do for $n = 0, 1, \dots$

$$\begin{aligned} \text{Step A: for } i \in \mathcal{V} \text{ even, } & \left\{ \begin{aligned} \alpha_{i+1,i}^{(n+1)} &= y_i^{\frac{1}{2+\beta}} \frac{(X_{i,i-1} \alpha_{i,i-1}^{(n)})^{\frac{1}{2+\beta}}}{(X_{i,i+1} \alpha_{i,i+1}^{(n)})^{\frac{1+\beta}{2+\beta}}} \\ \alpha_{i-1,i}^{(n+1)} &= y_i^{\frac{1}{2+\beta}} \frac{(X_{i,i+1} \alpha_{i,i+1}^{(n)})^{\frac{1}{2+\beta}}}{(X_{i,i-1} \alpha_{i,i-1}^{(n)})^{\frac{1+\beta}{2+\beta}}} \end{aligned} \right. \\ \\ \text{Step B: for } i \in \mathcal{V} \text{ odd, } & \left\{ \begin{aligned} \alpha_{i+1,i}^{(n+1)} &= y_i^{\frac{1}{2+\beta}} \frac{(X_{i,i-1} \alpha_{i,i-1}^{(n+1)})^{\frac{1}{2+\beta}}}{(X_{i,i+1} \alpha_{i,i+1}^{(n+1)})^{\frac{1+\beta}{2+\beta}}} \\ \alpha_{i-1,i}^{(n+1)} &= y_i^{\frac{1}{2+\beta}} \frac{(X_{i,i+1} \alpha_{i,i+1}^{(n+1)})^{\frac{1}{2+\beta}}}{(X_{i,i-1} \alpha_{i,i-1}^{(n+1)})^{\frac{1+\beta}{2+\beta}}} \end{aligned} \right. \end{aligned}$$

Remark I.6. In the case of the straight line, iterations for “inside” variables remain the same while on the “boundary” we get the following modifications. First note that there is no variable $\alpha_{0,1}, \alpha_{1,0}$ (or $\alpha_{K,1}, \alpha_{1,K}$); second the iterates for $i = P$ in step A and B are replaced by

$$\alpha_{K-1,K}^{(n+1)} = y_K^{\frac{1}{1+\beta}} \left(\frac{1}{X_{K,K-1} \alpha_{K,K-1}^{(n)}} \right)^{\frac{\beta}{1+\beta}} ; \alpha_{2,1}^{(n+1)} = y_1^{\frac{1}{1+\beta}} \left(\frac{1}{X_{1,2} \alpha_{1,2}^{(n+1)}} \right)^{\frac{\beta}{1+\beta}}$$

We now turn to the convergence analysis of these iterations in the Hilbert metric, and prove the following:

Theorem I.14. For $\beta > -1$, the Sinkhorn-like iterates $\alpha_{i,j}^{(n)}$ defined in algorithm I.5 converge in $(\mathbb{R}_*^+)^N / \sim$ to the fixed point $\alpha_{i,j}^*$ with the estimate

$$\forall n \geq 0, \left\{ \begin{aligned} d_{[1]}^{(n+1)} &\leq \kappa^2 d_{[1]}^{(n)} \\ d_{[0]}^{(n+1)} &\leq \kappa^2 d_{[0]}^{(n)} \end{aligned} \right. \quad (3.18)$$

where for $\tau \in \{0, 1\}$, $d_{[\tau]}^{(n)} = \max_{j,i} d_H(\alpha_{j,2i+\tau}^{(n)}, \alpha_{j,2i+\tau}^*)$ denotes the largest error on even vertices for $\tau = 0$ and odd vertices for $\tau = 1$ after n iterations, and where $\kappa = \max_{i,j} \kappa(X_{i,j}) < 1$ is the worst contraction rate of the positive matrices $X_{i,j}$.

Proof. Using properties of the Hilbert distance I.1, we get from step A for $i \in \mathcal{V}$ even:

$$\begin{aligned}
 d_H(\alpha_{i+1,i}^{(n+1)}, \alpha_{i+1,i}^*) &= d_H \left(y_i^{\frac{1}{2+\beta}} \frac{(X_{i,i-1} \alpha_{i,i-1}^{(n)})^{\frac{1}{2+\beta}}}{(X_{i,i+1} \alpha_{i,i+1}^{(n)})^{\frac{1+\beta}{2+\beta}}}, y_i^{\frac{1}{2+\beta}} \frac{(X_{i,i-1} \alpha_{i,i-1}^*)^{\frac{1}{2+\beta}}}{(X_{i,i+1} \alpha_{i,i+1}^*)^{\frac{1+\beta}{2+\beta}}} \right) \\
 &= d_H \left(\frac{(X_{i,i-1} \alpha_{i,i-1}^{(n)})^{\frac{1}{2+\beta}}}{(X_{i,i+1} \alpha_{i,i+1}^{(n)})^{\frac{1+\beta}{2+\beta}}}, \frac{(X_{i,i-1} \alpha_{i,i-1}^*)^{\frac{1}{2+\beta}}}{(X_{i,i+1} \alpha_{i,i+1}^*)^{\frac{1+\beta}{2+\beta}}} \right) \\
 &\leq \frac{1}{2+\beta} d_H(X_{i,i-1} \alpha_{i,i-1}^{(n)}, X_{i,i-1} \alpha_{i,i-1}^*) \\
 &\quad + \frac{1+\beta}{2+\beta} d_H(X_{i,i+1} \alpha_{i,i+1}^{(n)}, X_{i,i+1} \alpha_{i,i+1}^*) \\
 &\leq \frac{1}{2+\beta} \kappa(X_{i,i-1}) d_H(\alpha_{i,i-1}^{(n)}, \alpha_{i,i-1}^*) \\
 &\quad + \frac{1+\beta}{2+\beta} \kappa(X_{i,i+1}) d_H(\alpha_{i,i+1}^{(n)}, \alpha_{i,i+1}^*) \\
 &\leq \kappa d_{[1]}^{(n)}
 \end{aligned}$$

We also get similarly $d_H(\alpha_{i-1,i}^{(n+1)}, \alpha_{i-1,i}^*) \leq \kappa d_{[1]}^{(n)}$ so that $d_{[0]}^{(n+1)} \leq \kappa d_{[1]}^{(n)}$ and iterations of step B lead to $d_{[1]}^{(n+1)} \leq \kappa d_{[0]}^{(n+1)}$ which gives the desired rate. \square

Remark I.7. In the straight line case, one has to impose $\beta > -\frac{1}{2}$ to get

$$\begin{aligned}
 d_H(\alpha_{2,1}^{(n+1)}, \alpha_{2,1}^*) &\leq \frac{|\beta|}{1+\beta} \kappa(X_{1,2}) d_H(\alpha_{1,2}^{(n+1)}, \alpha_{1,2}^*) \leq \kappa d_{[0]}^{(n+1)} \\
 d_H(\alpha_{K-1,K}^{(n+1)}, \alpha_{K-1,K}^*) &\leq \frac{|\beta|}{1+\beta} \kappa(X_{K,K-1}) d_H(\alpha_{K,K-1}^{(n)}, \alpha_{K,K-1}^*) \leq \kappa d_{[1]}^{(n)}
 \end{aligned}$$

and finally, the same rate holds. Note that this estimate indeed vanishes for $\beta = 0$ as $\alpha_{2,1}$ and $\alpha_{K-1,K}$ are then prescribed to be multiples of y_1 and y_K .

Second, for our “ λ_2 ” analysis we now suppose $\beta = 0$. As a consequence, our iterates in algorithm I.5 satisfy at every step $\alpha_{i-1,i} \alpha_{i+1,i} = y_i$ and denoting $\gamma_i = \alpha_{i-1,i}$ it reduces to

$$\begin{aligned}
 \text{Step A : for } i \in \mathcal{V} \text{ even, } \gamma_i^{(n+1)} &= \sqrt{\frac{X_{i,i+1} \gamma_{i+1}^{(n)}}{X_{i,i-1} \gamma_{i-1}^{(n)}}} \\
 \text{Step B : for } i \in \mathcal{V} \text{ odd, } \gamma_i^{(n+1)} &= \sqrt{\frac{X_{i,i+1} \gamma_{i+1}^{(n+1)}}{X_{i,i-1} \gamma_{i-1}^{(n+1)}}}
 \end{aligned} \tag{3.19}$$

Remark I.8. In the case of the straight line with $\beta = 0$, iterations on $\alpha_{2,1}$ and $\alpha_{K-1,K}$ state that these variables are respectively constant to y_1 and y_K so that we do not compute them in practice. However it is appropriate to keep them for our convergence analysis, and we also introduce $\gamma_1 = \frac{y_1}{\alpha_{2,1}}$, $\gamma_K = \alpha_{K,K-1}$. Doing so, the iterations (3.19) in the case of the line remain the same on the variables $\gamma_i = \alpha_{i-1,i}$ for all $i \in \{2, \dots, K-1\}$, while we just keep at every step $\gamma_1 = \gamma_K = \mathbf{1}$.

We can now proceed to our usual analysis of such iterations, and get the following theorem:

Theorem I.15. Let $X_{i,j} \in (\mathbb{R}_*^+)^{N \times N}$ for $i \in \mathcal{V} = \llbracket 1, K \rrbracket$ and $j \in \{i-1, i+1\}$ be positive matrices such that $X_{i,j}^T = X_{j,i}$, and let $y^1, \dots, y^K \in (\mathbb{R}_*^+)^N$. Let $\gamma^{(n)} = (\gamma_i^{(n)})_{1 \leq i \leq K}$ be obtained by iterations (3.19), and denote $\gamma_{[0]}^{(n)}$ (respectively $\gamma_{[1]}^{(n)}$) the concatenation of the $\gamma_i^{(n)}$ for i even (respectively odd). Then $(\gamma^{(n)})$ converges to a fixed point $\gamma^* = (\gamma_{[0]}^*, \gamma_{[1]}^*)$ and there exist $s \in (\mathbb{R}_*^+)^{NK}$, sequences of real numbers $(u_n), (v_n)$ converging to 1, and $\lambda_2 \in (0, 1)$ such that $\forall \delta > 0, \exists n_\delta \in \mathbb{N}$ such that:

$$\forall n \geq n_\delta, \left\| \begin{pmatrix} \gamma_{[0]}^{(n+1)} \\ \gamma_{[1]}^{(n+1)} \end{pmatrix} - \begin{pmatrix} u_{n+1} \gamma_{[0]}^* \\ v_{n+1} \gamma_{[1]}^* \end{pmatrix} \right\|_s \leq (\lambda_2 + \delta) \left\| \begin{pmatrix} \gamma_{[0]}^{(n)} \\ \gamma_{[1]}^{(n)} \end{pmatrix} - \begin{pmatrix} u_n \gamma_{[0]}^* \\ v_n \gamma_{[1]}^* \end{pmatrix} \right\|_s$$

Proof. Denote by γ^* a fixed point of iterations (3.19) and define:

$$Y_{i,i+1} = d\left(\frac{y_i}{\gamma_i^*}\right) X_{i,i+1} d(\gamma_{i+1}^*); \quad Y_{i,i-1} = d(\gamma_i^*) X_{i,i-1} d\left(\frac{y_{i-1}}{\gamma_{i-1}^*}\right)$$

as well as the matrices W (which constitute in fact the solution of the original labelling problem (3.15)) given by $W_{i,j} = d\left(\frac{1}{Y_{i,j} \mathbf{1}}\right) Y_{i,j}$ for all $i \in \mathcal{V}, j \in \{i-1, i+1\}$. Since

$Y_{i,i-1} \mathbf{1} = Y_{i,i+1} \mathbf{1} = \sqrt{y_i (X_{i,i+1} \gamma_{i+1}^*) (X_{i,i-1} \frac{y_{i-1}}{\gamma_{i-1}^*})}$, one checks that the rescaled variables $c_i^{(n)} = \frac{\gamma_i^{(n)}}{\gamma_i^*}$ satisfy in fine:

$$\begin{aligned} \text{Step A : for } i \in \mathcal{V} \text{ even, } c_i^{(n+1)} &= \sqrt{\frac{W_{i,i+1} c_{i+1}^{(n)}}{W_{i,i-1} \frac{1}{c_{i-1}^{(n)}}}} \\ \text{Step B : for } i \in \mathcal{V} \text{ odd, } c_i^{(n+1)} &= \sqrt{\frac{W_{i,i+1} c_{i+1}^{(n+1)}}{W_{i,i-1} \frac{1}{c_{i-1}^{(n+1)}}}} \end{aligned}$$

with matrices W being stochastic. Consequently, Lemma I.5 yields for i even:

$$c_i^{(n+1)} \leq \sqrt{\left(W_{i,i+1}c_{i+1}^{(n)}\right)\left(W_{i,i-1}c_{i-1}^{(n)}\right)} \leq \frac{1}{2}\left(W_{i,i+1}c_{i+1}^{(n)} + W_{i,i-1}c_{i-1}^{(n)}\right)$$

$$\frac{1}{c_i^{(n+1)}} \leq \sqrt{\left(W_{i,i-1}\frac{1}{c_{i-1}^{(n)}}\right)\left(W_{i,i+1}\frac{1}{c_{i+1}^{(n)}}\right)} \leq \frac{1}{2}\left(W_{i,i+1}\frac{1}{c_{i+1}^{(n)}} + W_{i,i-1}\frac{1}{c_{i-1}^{(n)}}\right)$$

and similarly for i odd with $(n+1)$ instead of (n) .

Combining these inequalities, one gets for i even:

$$c_i^{(n+1)} \leq \frac{1}{4}(W_{i,i-1}W_{i-1,i-2})c_{i-2}^{(n)} +$$

$$\frac{1}{4}(W_{i,i+1}W_{i+1,i} + W_{i,i-1}W_{i-1,i})c_i^{(n)} + \frac{1}{4}(W_{i,i+1}W_{i+1,i+2})c_{i+2}^{(n)}$$

which can be wrapped into $c_{[0]}^{(n+1)} \leq M c_{[0]}^{(n)}$ with $c_{[0]}$ being the concatenation of the c_i for even i , and M being the block matrix defined by:

$$M = \frac{1}{4} \begin{pmatrix} M_{1,1} & M_{1,2} & \dots & M_{1,P} \\ M_{2,1} & M_{2,2} & M_{2,3} & \dots \\ \vdots & & \ddots & \vdots \\ M_{P,1} & \dots & M_{P,P-1} & M_{P,P} \end{pmatrix}$$

where the square submatrices $M_{i,j}$ are defined by

$$M_{i,i} = W_{2i,2i+1}W_{2i+1,2i} + W_{2i,2i-1}W_{2i-1,2i}$$

$$M_{i,i+1} = W_{2i,2i+1}W_{2i+1,2i+2}; \quad M_{i,i-1} = W_{2i,2i-1}W_{2i-1,2i-2}$$

and all other entries of M are zeros. One similarly gets $\frac{1}{c_{[0]}^{(n+1)}} \leq M \frac{1}{c_{[0]}^{(n)}}$.

At this point, note that since the W matrices are stochastic so is M . In addition, the zero pattern of M is smaller than the one of a tridiagonal matrix, but such a matrix is primitive according to point 1 of Proposition I.3: finally, M is primitive. In the objective of applying the main theorem I.7, we just need to find some positive vector $s \in \Sigma^{PN}$ such that $d(s)M$ is symmetric. To do so, just remember that by assumption we have $\theta_{i,j}(\ell, m) = \theta_{j,i}(m, \ell)$ so that $X_{i,j}^T = X_{j,i}$ and $Y_{i,j}^T = Y_{j,i}$, then defining the

vectors $s_i = Y_{2i,2i+1}\mathbf{1} = Y_{2i,2i-1}\mathbf{1}$ we get:

$$\begin{aligned} d(s_i)M_{i,i} &= Y_{2i,2i+1}\mathbf{1}W_{2i,2i+1}W_{2i+1,2i} + Y_{2i,2i-1}\mathbf{1}W_{2i,2i-1}W_{2i-1,2i} \\ &= Y_{2i,2i+1}\frac{1}{Y_{2i+1,2i}\mathbf{1}}Y_{2i+1,2i} + Y_{2i,2i-1}\frac{1}{Y_{2i-1,2i}\mathbf{1}}Y_{2i-1,2i} \\ &= Y_{2i,2i+1}^T\frac{1}{Y_{2i+1,2i}\mathbf{1}}Y_{2i,2i+1} + Y_{2i,2i-1}^T\frac{1}{Y_{2i-1,2i}\mathbf{1}}Y_{2i,2i-1} \end{aligned}$$

and one can consequently take for s the concatenation of the s_i (divided by the appropriate constant to ensure $s \in \Sigma^{PN}$). The same analysis holds for $c_{[1]}$ with a similar matrix M' . To conclude just take λ_2 to be the smallest subdominant eigenvalue of M and M' . □

Remark I.9. *The same estimates remain valid in the case of the straight line as the inside iterations are unchanged and one can write, for instance for the variable γ_2 :*

$$c_2^{(n+1)} \leq \frac{1}{4}W_{2,3}W_{3,4}c_4^{(n)} + \frac{1}{4}W_{2,3}W_{3,2}c_2^{(n)} + \frac{1}{2}W_{2,1}c_0^{(n)}$$

for a fictional variable $c_0^{(n)} = \mathbf{1}$. No modification is needed however for c_{K-2} when we use the fictional variable $c_K^{(n)} = \mathbf{1}$. Finally, we obtain that $c_{[0]}^{(n+1)} \leq Mc_{[0]}^{(n)}$ where M is the matrix given by

$$M = \frac{1}{4} \begin{pmatrix} 4I & 0 & \dots & \dots & \dots & \dots & 0 \\ 2W_{2,1} & W_{2,3}W_{3,2} & W_{2,3}W_{3,4} & 0 & \dots & \dots & 0 \\ 0 & M_{2,1} & M_{2,2} & M_{2,3} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & M_{P-1,P-2} & M_{P-1,P-1} & M_{P-1,P} \\ 0 & \dots & \dots & \dots & \dots & 0 & 4I \end{pmatrix}$$

where we use the matrices $M_{i,j}$ defined above, and one obtains similar results.

Remark I.10. *The hypothesis we made that the graph $(\mathcal{V}, \mathcal{E})$ is bipartite only aims at giving a natural way of conducting the iterations (3.17), and hence the corresponding inequalities of our proofs. This setting also provides great similarity with our previous Sinkhorn-like algorithms that always lead iterations on two separate groups of variables. However, it seems that the key argument of the proofs we give in the case of the circle and the line graphs is more the fact that each iteration involves at most 2 other variables than the fact that we ultimately only have 2 steps (A and B) in our procedure. As such, the bipartite hypothesis could be released, but not the arity restriction, so that this would not lead to more complex graphs.*

3.3 Final remarks

3.3.1 Limits of the analysis: 1D balanced Sinkhorn-like algorithms

While we stated our last Sinkhorn-like algorithm I.4 on a general bipartite graph $(\mathcal{V}, \mathcal{E})$, we only presented its convergence analysis in a 1D setting. This is because the analysis supported by the general Theorem I.7 does not seem to extend directly to a general setting. To apprehend why, first consider the analysis in Hilbert distance we presented in Theorem I.14. The analogous estimate for algorithm I.4 writes

$$\begin{aligned} d_H(\alpha_{j,i}^{(n+1)}, \alpha_{j,i}^*) &\leq \sum_{j' \sim i, j' \neq j} \frac{1}{n_i + \beta} \kappa(X_{i,j'}) d_H(\alpha_{i,j'}^{(n)}, \alpha_{i,j'}^*) \\ &\quad + \left(1 - \frac{1}{n_i + \beta}\right) \kappa(X_{i,j}) d_H(\alpha_{i,j}^{(n)}, \alpha_{i,j}^*) \\ &\leq (n_i - 1) \frac{1}{n_i + \beta} \kappa d_{[j]}^{(n)} + \left(1 - \frac{1}{n_i + \beta}\right) \kappa d_{[j]}^{(n)} \\ &= \frac{2n_i - 2 + \beta}{n_i + \beta} \kappa d_{[j]}^{(n)} \end{aligned}$$

To prove convergence with the same argument one would then need $\frac{2n_i - 2 + \beta}{n_i + \beta} \leq 1$ which only occurs for $n_i \leq 2$. This is why we requested our graph to have maximal arity 2. The appearance of all the $(n_i - 1)$ terms in the above calculation is similar to the obstruction we obtain when trying to write an estimate of the type $a^{(n+1)} \leq M a^{(n)}$ with a stochastic matrix M on the rescaled variables $a = \frac{\alpha}{\alpha^*}$.

The same phenomenon prevents us from generalizing the analysis of the Sinkhorn-like algorithm for the barycenter we led in section 3.1.1. The Fréchet mean definition of the barycenter of two measures can of course be generalized to an arbitrary number of measures f^1, \dots, f^K for $K > 2$, see [AC11, BCC⁺15]. Given a family of weights $\theta_k \in (0, 1)$ such that $\sum_{k=1}^K \theta_k = 1$, one defines the corresponding Wasserstein barycenter $f \in \Sigma^N$ as the minimizer of $\sum_{k=1}^K \theta_k W_w(f^k, f)$. One can also study this procedure for different cost matrices w^k . Adding entropic regularizations, that is replacing W_{w^k} by $W_{w^k}^\varepsilon$, and dualizing the constraints leads to the dual problem:

$$\begin{aligned} \max_{\lambda, \mu \in (\mathbb{R}^N)^K} \quad & \sum_k \langle \lambda^k | f^k \rangle - \varepsilon \sum_k \theta_k \sum_{i,j} \exp\left(\frac{-\theta_k w_{i,j}^k + \mu_j^k + \lambda_i^k}{\varepsilon \theta_k}\right) \\ \text{s.t.} \quad & \sum_{k=1}^K \mu^k = 0 \end{aligned}$$

When making the following change of variables

$$A^k = \exp\left(-\frac{w^k}{\varepsilon}\right); \quad \alpha^k = \exp\left(\frac{\lambda^k}{\varepsilon}\right); \quad \beta^k = \exp\left(\frac{\mu^k}{\varepsilon}\right)$$

optimality conditions write as the scaling problem:

$$\forall k \in \llbracket 1, K \rrbracket, \quad \alpha^k = \frac{f^k}{A^k \beta^k}; \quad \beta^k = \frac{\prod_{\ell=1}^K (A^\ell T \alpha^\ell)^{\theta_\ell}}{A^k T \alpha^k}$$

This gives rise to the following Sinkhorn-like barycenter algorithm, that generalizes algorithm I.2:

Algorithm I.6. *Given positive matrices $A^1, \dots, A^K \in (\mathbb{R}_*^+)^{N \times N}$, marginals $f^1, \dots, f^K \in \Sigma^N$ and weights $\theta_1, \dots, \theta_K \in (0, 1)$ such that $\sum_k \theta_k = 1$, starting from $\alpha^{k(0)} = \beta^{k(0)} = \mathbf{1}$ for all $k \in \llbracket 1, N \rrbracket$, do for $n = 0, 1, \dots$*

$$\forall k \in \llbracket 1, K \rrbracket, \alpha^{k(n+1)} = \frac{f^k}{A^k \beta^{k(n)}}; \beta^{k(n+1)} = \frac{\prod_{\ell=1}^K (A^\ell T \alpha^{\ell(n+1)})^{\theta_\ell}}{A^k T \alpha^{k(n+1)}} \quad (3.20)$$

As noted by [BCC⁺15], one can also interpret these iterates as Bregman projections on the primal variables $x^k \in \Pi(f^k, f)$ that form the transport plans between the marginals and the barycenter. Iterates on α^k correspond to a projection onto the constraints $x^k \mathbf{1} = f^k$ while iterates on β^k correspond to $x^k T \mathbf{1} = x^\ell T \mathbf{1}$ for all k, ℓ . These transport plans will finally be obtained, after convergence of this procedure towards positive vectors α^{k*}, β^{k*} , through $X^k = d(\alpha^{k*}) A^k d(\beta^{k*}) \in \Pi(f^k, f)$ where $f = \prod_{\ell=1}^K (A^\ell T \alpha^{\ell(n+1)})^{\theta_\ell}$ is the desired barycenter. To pursue with an analysis close to what we did for the case $K = 2$, one notices that at every step of algorithm I.6, $\prod_k (\beta^k)^{\theta_k} = \mathbf{1}$ so that one can get rid of the variable β^K . Doing so, we still have $K - 1$ variables β^k in the second step of the iterations written as a product over the K variables α^k . If $K \geq 3$, it appears that we are not able to introduce a new variable γ to simplify these products as we did for algorithms I.2, I.3, I.4.

Finally, these two examples show us that our analysis is restricted to the case of Sinkhorn-like algorithms in which all the iterations involve at most two other variables. We name these types of Sinkhorn variants the *1D Sinkhorn-like algorithms*.

Actually, this “1D” setting is not enough to guarantee the success of our analysis. For instance it does not work either on the natural over-relaxation of the classical Sinkhorn algorithm proposed by Thibault and co-authors in [TCDP17]²:

Algorithm I.7. *Given a positive matrix $A \in (\mathbb{R}_*^+)^{N \times N}$, marginals $f, g \in \Sigma^N$ and a parameter $\omega \in (1, 2)$, starting from $\alpha^{(0)} = \beta^{(0)} = \mathbf{1}$, do for $n = 0, 1, \dots$*

$$\begin{aligned} \alpha^{(n+1)} &= \frac{(\hat{\alpha}^{(n)})^\omega}{(\alpha^{(n)})^{\omega-1}} \text{ where } \hat{\alpha}^{(n)} = \frac{f}{A \beta^{(n)}} \\ \beta^{(n+1)} &= \frac{(\hat{\beta}^{(n+1)})^\omega}{(\beta^{(n)})^{\omega-1}} \text{ where } \hat{\beta}^{(n+1)} = \frac{g}{A^T \alpha^{(n+1)}} \end{aligned} \quad (3.21)$$

If we first try to analyze these iterations in Hilbert distance, we get stuck by the appearing exponents and can only write:

$$d_H(\alpha^{(n+1)}, \alpha^*) \leq \omega \kappa d_H(\beta^{(n)}, \beta^*) + (\omega - 1) d_H(\alpha^{(n)}, \alpha^*)$$

and similar expression for $\beta^{(n+1)}$, which ultimately does not lead to a contraction rate. This is because the Hilbert distance does not take into account the simplification of

²In this article, the authors prove convergence of a slightly altered version of this scheme and show it indeed improves the local convergence rate of the Sinkhorn algorithm.

the exponents, introducing their absolute value in point 3 of Proposition-Definition I.1. The same reason imposed bounds on the parameter β in our Hilbert analysis on 1D graphs ($\beta > -1$ for the circle, and $\beta > -\frac{1}{2}$ for the line) in Theorem I.14.

We also note that our analysis of the Sinkhorn-like algorithm for the barycenter section 3.1.1 deeply relies on the use of the convex inequality $x^\theta y^{1-\theta} \leq \theta x + (1-\theta)y$ for $x, y > 0$ and $\theta \in (0, 1)$. Consequently, the exponents ω and $\omega - 1$ with $\omega \in (1, 2)$ appearing in algorithm I.7 do not seem to be suited for a similar analysis because the sum of their absolute values is not 1.

Finally, our method applies to *1D balanced Sinkhorn-like algorithms*, meaning that each iteration involves at most two other variables and that the absolute values of the appearing exponents sum to 1.

3.3.2 Degenerate direction of Sinkhorn algorithm

However, numerical experiments confirm that the two algorithms I.6, I.7 do converge linearly and that their convergence rate is also given by the subdominant eigenvalue of a matrix with rows summing to 1. This matrix is the Jacobian matrix of the iterated function at the fixed point. To understand what is at stake here, let us first go back to the classical Sinkhorn algorithm. Focus on the variable b of the Sinkhorn iterates (2.1):

$$a^{(n+1)} = \frac{f}{Xb^{(n)}}; b^{(n+1)} = \frac{g}{X^T a^{(n+1)}}$$

where $X = d(\alpha^*)A d(\beta^*)$. One can write this iteration rule as $b^{(n+1)} = \phi(b^{(n)})$ for the function $\phi : (\mathbb{R}_*^+)^N \rightarrow (\mathbb{R}_*^+)^N$ defined by

$$\phi(b) = \frac{g}{X^T \frac{f}{Xb}}$$

In that view, $(b^{(n)})$ is the sequence obtained by iterating ϕ from $b^{(0)} = \frac{1}{\beta^*}$ where β^* is the limit value of the Sinkhorn algorithm. We showed in the previous sections that this sequence converges to $b^* = \mathbf{1}$ which is a fixed point of ϕ . It is then natural to look at $J\phi(\mathbf{1})$, the Jacobian matrix of ϕ at $\mathbf{1}$. An easy computation shows that

$$J\phi(\mathbf{1}) = d\left(\frac{1}{g}\right)X^T d\left(\frac{1}{f}\right)X = M_b$$

is the stochastic matrix involved in our analysis³. Unfortunately, the spectral radius of this Jacobian matrix is 1 as it posses the eigenvector $\mathbf{1}$. This prevents us from using directly the following classical linear convergence result:

³One can also lead this computation directly on the variable β , that is on the sequence given by $\beta^{(n+1)} = \psi(\beta^{(n)})$ where ψ is the function ϕ with X being replaced by A . At the fixed point β^* one finds $J\psi(\beta^*) = d(\beta^*)M_b d(\beta^*)^{-1}$ so that the spectrum analysis remains the same.

Theorem I.16. *Let E be a real finite dimensional vector space and let $\phi : E \rightarrow E$ be a \mathcal{C}^1 function admitting $x^* \in E$ as a fixed point: $\phi(x^*) = x^*$. Suppose that $D\phi(x^*)$ is diagonalizable and that its spectral radius is less than 1: $\rho(D\phi(x^*)) = \rho < 1$, then there exists a norm $\|\cdot\|$ on E and $R > 0$ such that: if $x^0 \in E$ satisfies $\|x^0 - x^*\| \leq R$ then the sequence (x^n) given by $x^{n+1} = \phi(x^n)$ converges to x^* with the following estimate: $\forall \delta > 0, \exists n_\delta \in \mathbb{N}$ such that*

$$\forall n \geq n_\delta, \|x^{n+1} - x^*\| \leq (\rho + \delta)\|x^n - x^*\|$$

Proof. Write that

$$x^{n+1} - x^* = \phi(x^n) - \phi(x^*) = \int_0^1 D\phi(x^* + t(x^n - x^*))(x^n - x^*) dt \quad (3.22)$$

As $D\phi(x^*)$ is diagonalizable, defining for instance $\|\cdot\|$ to be the infinite norm in a diagonalization basis of $D\phi(x^*)$ we have:

$$\forall h \in E, \|D\phi(x^*)h\| \leq \rho\|h\|$$

Then, as ϕ is \mathcal{C}^1 , for any $\rho' \in (\rho, 1)$ one can find $R > 0$ such that

$$\|x - x^*\| \leq R \Rightarrow \left(\forall h \in E, \|D\phi(x)h\| \leq \rho'\|h\| \right)$$

Using this bound in (3.22) shows that if $\|x^0 - x^*\| \leq R$ then one has for all n , $\|x^n - x^*\| \leq R\rho'^n$. Hence (x^n) converges to x^* , and the convergence rate follows using the same argument. \square

Remark I.11. *The hypothesis that $D\phi(x^*)$ is diagonalizable is mandatory to recover exactly the spectral radius through the operator norm $\sup_{h \neq 0} \frac{\|D\phi(x^*)h\|}{\|h\|}$. However as this spectral radius equals the supremum of all the operator norms of $D\phi(x^*)$, one can get a similar, but norm-dependant, result if $D\phi(x^*)$ is no longer diagonalizable: the sequence (x^n) converges to x^* and for all $\delta > 0$, there exist a norm $\|\cdot\|_\delta$ and $n_\delta \in \mathbb{N}$ such that*

$$\forall n \geq n_\delta, \|x^{n+1} - x^*\|_\delta \leq (\rho + \delta)\|x^n - x^*\|_\delta$$

In our setting, the differential of ϕ represented by its Jacobian matrix M_b has $\mathbf{1}$ as a degenerate direction. This is related to the non-uniqueness of the scaling diagonal matrices D^1, D^2 such that $X = D^1 A D^2$, which can be replaced by $D'^1 = r D^1, D'^2 = \frac{1}{r} D^2$ for any $r > 0$. Going back to the logarithmic variables $\lambda = \varepsilon \log \alpha$ and $\mu = \varepsilon \log \beta$ that appeared in the Optimal Transportation formulation of the scaling problem section 1.2, remember that the Sinkhorn algorithm corresponds to the iterates (1.9):

$$\begin{aligned} \lambda_i^{(n+1)} &= \varepsilon \log(f_i) - \varepsilon \log \left(\sum_{j=1}^N \exp \left(\frac{\mu_j^{(n)} - w_{i,j}}{\varepsilon} \right) \right) \\ \mu_j^{(n+1)} &= \varepsilon \log(g_j) - \varepsilon \log \left(\sum_{i=1}^N \exp \left(\frac{\lambda_i^{(n+1)} - w_{i,j}}{\varepsilon} \right) \right) \end{aligned} \quad (3.23)$$

This procedure actually achieves the alternate maximizations on the dual objective of the perturbed Optimal Transportation problem

$$\max_{\lambda, \mu \in \mathbb{R}^N} \langle \lambda | f \rangle + \langle \mu | g \rangle - \varepsilon \sum_{i,j} \exp\left(\frac{-w_{i,j} + \lambda_i + \mu_j}{\varepsilon}\right) =: F(\lambda, \mu)$$

As noticed by Bezdek and co-authors in a general setting in [BHH⁺87], when searching for the maximum (λ^*, μ^*) of F through alternate maximizations, one iterates the functions $S \circ T$ and $T \circ S$ where

$$T(\lambda) = \arg \max_{\mu \in \mathbb{R}^N} F(\lambda, \mu) \quad ; \quad S(\mu) = \arg \max_{\lambda \in \mathbb{R}^N} F(\lambda, \mu)$$

and the spectral radius of $D(S \circ T)(\lambda^*)$ and $D(T \circ S)(\mu^*)$ (which is the same) is less than 1 if and only if the Hessian matrix of F at (λ^*, μ^*) is negative definite. In our case, we stress again that we are not in this situation because the maximum (λ^*, μ^*) of F is not unique as $F(\lambda^* + r\mathbf{1}, \mu^* - r\mathbf{1}) = F(\lambda^*, \mu^*)$ for any $r \in \mathbb{R}$. However, this setting with variables λ and μ emphasizes the linear structure of this obstruction, and calls for an interpretation in terms of quotient variables. We develop this point of view below.

The Sinkhorn algorithm can be written as $(\lambda^{(n+1)}, \mu^{(n+1)}) = \varphi(\lambda^{(n)}, \mu^{(n)})$ where $\varphi : \mathbb{R}^{2N} \rightarrow \mathbb{R}^{2N}$ is given by

$$\varphi(\lambda, \mu) = \left(\varepsilon \log \frac{f}{A \left(\exp \left(\frac{\mu}{\varepsilon} \right) \right)}, \varepsilon \log \frac{g}{A^T \frac{f}{A \left(\exp \left(\frac{\mu}{\varepsilon} \right) \right)}} \right)$$

At a fixed point $(\lambda^*, \mu^*) \in \mathbb{R}^{2N}$ of φ , one computes the following Jacobian matrix:

$$J\varphi(\lambda^*, \mu^*) = \begin{pmatrix} 0 & -d(\frac{1}{f})X \\ 0 & d(\frac{1}{g})X^T d(\frac{1}{f})X \end{pmatrix}$$

where $X = D^1 A D^2 \in \Pi(f, g)$. Then one checks that the function φ satisfies:

$$\forall (\lambda, \mu) \in \mathbb{R}^{2N}, \forall r \in \mathbb{R}, \varphi(\lambda + r\mathbf{1}, \mu - r\mathbf{1}) = \varphi(\lambda, \mu) + r(\mathbf{1}, -\mathbf{1})$$

which shows that it can be considered on the quotient space $E = \mathbb{R}^{2N} / \mathbb{R}(\mathbf{1}, -\mathbf{1})$. We denote the elements of this space by $(\overline{\lambda}, \overline{\mu}) = (\lambda, \mu) + \mathbb{R}(\mathbf{1}, -\mathbf{1})$, and define the function $\phi : E \rightarrow E$ given by $\phi((\overline{\lambda}, \overline{\mu})) = \varphi(\lambda, \mu)$. The space E comes with the Euclidean structure inherited from \mathbb{R}^{2N} that correspond to the distance on the orthogonal space of $\mathbb{R}(\mathbf{1}, -\mathbf{1})$ and given by

$$\|(\overline{\lambda}, \overline{\mu})\|_E = \min_{r \in \mathbb{R}} \|(\lambda + r\mathbf{1}, \mu - r\mathbf{1})\|_2$$

where $\|\cdot\|_2$ denotes the Euclidean norm on \mathbb{R}^{2N} . Now the differential of ϕ is of course obtained from the one of φ thanks to the following proposition:

Proposition I.6. *Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and let $z \in \mathbb{R}^d$ such that $\forall x \in \mathbb{R}^d, \forall r \in \mathbb{R}, \varphi(x + rz) \in \varphi(x) + \mathbb{R}z$; consider $\phi : E \rightarrow E$ given by $\phi(\bar{x}) = \overline{\varphi(x)}$ on the quotient space $E = \mathbb{R}^d / \mathbb{R}z$. Suppose φ is differentiable at $x \in \mathbb{R}^d$ then ϕ is differentiable at $\bar{x} \in E$ and $D\phi(\bar{x})$ is given by: $\forall \bar{h} \in E, D\phi(\bar{x})(\bar{h}) = \overline{D\varphi(x)(h)}$.*

Proof. First, the fact that $\varphi(x + z) = \varphi(x) + rz$ for some $r \in \mathbb{R}$ implies that $D\varphi(x)(z) = rz$ so that the proposed differential is correctly defined. Second, define $\varepsilon : \mathbb{R}^d \rightarrow \mathbb{R}^d$ given by $\varepsilon(h) = \varphi(x + h) - \varphi(x) - D\varphi(x)(h)$. For $\delta > 0$ fixed, by definition of the differential of φ there exists $\eta > 0$ such that :

$$\forall h \in \mathbb{R}^d, \|h\|_2 \leq \eta \Rightarrow \|\varepsilon(h)\|_2 \leq \delta \|h\|_2$$

Then if $\bar{h} \in E$ is such that $\|\bar{h}\|_E \leq \eta$, write that $\|\bar{h}\|_E = \|h - sz\|_2$ for some $s \in \mathbb{R}$ to get that $\|\varepsilon(h - sz)\|_2 \leq \delta \|\bar{h}\|_E$. Now notice that $\varepsilon(h - sz) = \varepsilon(h)$ so that $\|\varepsilon(\bar{h})\|_E \leq \|\varepsilon(h - sz)\|_2$. Finally we showed that $\varepsilon(\bar{h}) = o(\bar{h})$ in E which gives our result. \square

As a consequence, the differential of ϕ at its unique fixed point $(\overline{\lambda^*, \mu^*})$ is given by the action of $J\varphi(\lambda^*, \mu^*)$ on the space $E = \mathbb{R}^{2N} / \mathbb{R}(\mathbf{1}, -\mathbf{1})$. At this point, we see that the quotienting achieves exactly the desired operation on the spectrum, and that we get

$$Sp(D\phi(\overline{(\lambda^*, \mu^*)})) = Sp(J\varphi(\lambda^*, \mu^*)) \setminus \{1\} \quad (3.24)$$

Indeed, the matrix $J\varphi(\lambda^*, \mu^*)$ admits precisely $2N$ linear independent eigenvectors which are: the $(e^{(k)}, 0)$ for $k = 1, \dots, N$ where $e^{(k)}$ is k^{th} basis vector of \mathbb{R}^N , associated to the eigenvalue 0; and $(-\text{d}(\frac{1}{f})Xv^{(k)}, v^{(k)})$ for $k = 1, \dots, N$ where $v^{(k)}$ are the eigenvectors of the stochastic matrix $M_b = \text{d}(\frac{1}{g})X^T \text{d}(\frac{1}{f})X$. Among these last vectors is $(-\mathbf{1}, \mathbf{1})$, associated to the eigenvalue 1 and that reduces to 0 in the quotient. All the other eigenvectors being associated to the other eigenvalues of M_b , which are smaller than 1 in magnitude, we finally get that the spectral radius of $D\phi(\overline{(\lambda^*, \mu^*)})$ is less than 1. Searching for these eigenvalues, we also proved that $D\phi(\overline{(\lambda^*, \mu^*)})$ is diagonalizable, so that one can use Theorem I.16 and get the following convergence result, expressed in the quotient space and under a condition on the initial value $(\lambda^{(0)}, \mu^{(0)})$:

Theorem I.17. *If $\min_{r \in \mathbb{R}} \|(\lambda^{(0)} - \lambda^* + r\mathbf{1}, \mu^{(0)} - \mu^* - r\mathbf{1})\|_2$ is small enough, then the iterates $(\lambda^{(n)}, \mu^{(n)})$ of the Sinkhorn algorithm defined by (3.23) converge in the quotient space $E = \mathbb{R}^{2N} / \mathbb{R}(\mathbf{1}, -\mathbf{1})$ towards $(\overline{\lambda^*, \mu^*})$. In addition, there exists a norm $\|\cdot\|$ on E such that $\forall \delta > 0, \exists n_\delta \in \mathbb{N}$ such that $\forall n \geq n_\delta$:*

$$\|(\overline{(\lambda^{(n+1)}, \mu^{(n+1)})} - \overline{(\lambda^*, \mu^*)})\| \leq (\lambda_2 + \delta) \|(\overline{(\lambda^{(n)}, \mu^{(n)})} - \overline{(\lambda^*, \mu^*)})\|$$

where $\lambda_2 = \max\{|\lambda|, \lambda \in Sp(M_b) \text{ s.t. } |\lambda| < 1\}$.

Remark I.12. *It is also possible to obtain a result on “true” variables rather than on the quotient space using a natural representation of E such as $(\mathbb{R}(\mathbf{1}, -\mathbf{1}))^\perp = \{(\lambda, \mu) \in \mathbb{R}^{2N} \text{ s.t. } \sum_i \lambda_i = \sum_j \mu_j\}$. However in that setting, the iterated function $\phi : (\mathbb{R}(\mathbf{1}, -\mathbf{1}))^\perp \rightarrow (\mathbb{R}(\mathbf{1}, -\mathbf{1}))^\perp$ leads to a slightly different version of the Sinkhorn algorithm. Indeed, as the constraint $\sum_i \lambda_i^{(n)} = \sum_j \mu_j^{(n)}$ is satisfied at each step, the*

corresponding exponential variables will be updated in such a way that the quantities $\prod_i \alpha_i^{(n)}$ and $\prod_j \beta_j^{(n)}$ are constant. In other words, we then get a renormalization step and actually are studying the Sinkhorn variant:

Algorithm I.8. Given a nonnegative matrix $A \in (\mathbb{R}^+)^{N \times N}$ and two marginals $f, g \in \Sigma^N$, starting from $\alpha^{(0)} = \beta^{(0)} = \mathbf{1}$, do for $n = 0, 1, \dots$

$$\alpha^{(n+1)} = \frac{\hat{\alpha}^{(n+1)}}{\prod_i \hat{\alpha}_i^{(n+1)}} \text{ where } \hat{\alpha}^{(n+1)} = \frac{f}{A\beta^{(n)}}$$

$$\beta^{(n+1)} = \frac{\hat{\beta}^{(n+1)}}{\prod_j \hat{\beta}_j^{(n+1)}} \text{ where } \hat{\beta}^{(n+1)} = \frac{g}{A^T \alpha^{(n+1)}}$$

Of course the directions of such iterates are the same as the ones of the classical Sinkhorn algorithm. However to transfer the convergence analysis we would obtain for these iterates by Theorem I.16 to the initial iterates, one would need to control the growth of $\prod_{k=1}^n \prod_i \alpha_i^{(k)}$. These types of rescaling constants appear widely in convergence analyses of the Sinkhorn algorithm.

The result of Theorem I.17 could possibly be adapted to prove that the local convergence rates of algorithms I.6 and I.7 are, in some sense, given by the subdominant eigenvalues of their Jacobian matrices at their fixed points. These matrices take the following form. For algorithm I.6, using rescaled variables $a^{k(n)}, b^{k(n)}$, one rephrases iterations (3.20) as $b^{(n+1)} = \phi(b^{(n)})$ where $b^{(n)} = (b^{k(n)})_{1 \leq k \leq K-1}$ and $\phi : (\mathbb{R}_*^+)^{N(K-1)} \rightarrow (\mathbb{R}_*^+)^{N(K-1)}$ is given by

$$\phi(b) = \left(d^{K\theta_K} c^{k\theta_k-1} \prod_{\ell \neq k} c^{\ell\theta_\ell} \right)_{1 \leq k \leq K-1} \text{ where } \begin{cases} \forall k \in \llbracket 1, K-1 \rrbracket, c^k = X^{kT} \frac{f^k}{X^k b^k} \\ d^K = X^{KT} f^K \left(X^K \prod_k (b^k)^{-\frac{\theta_k}{\theta_K}} \right)^{-1} \end{cases}$$

The Jacobian matrix, which generalizes the matrix $M = (1 - \theta)M^0 + \theta M^1$ we introduced in Theorem I.12 is:

$$J\phi(\mathbf{1}) = \begin{pmatrix} \theta_1 M^K + (1 - \theta_1)M^1 & \theta_2(M^K - M^2) & \dots & \theta_{K-1}(M^K - M^{K-1}) \\ \theta_1(M^K - M^1) & \theta_2 M^K + (1 - \theta_2)M^2 & \dots & \vdots \\ \vdots & \theta_2(M^K - M^2) & \dots & \vdots \\ \vdots & \vdots & \dots & \theta_{K-1}(M^K - M^{K-1}) \\ \theta_1(M^K - M^1) & \theta_2(M^K - M^2) & \dots & \theta_{K-1}M^K + (1 - \theta_{K-1})M^{K-1} \end{pmatrix}$$

where M^k are the stochastic matrices given by $M^k = d(\frac{1}{f})X^{kT} d(\frac{1}{f^k})X^k$. Concerning algorithm I.7, iterations (3.21) on rescaled variables $a^{(n)}, b^{(n)}$ can be written as $(a^{(n+1)}, b^{(n+1)}) = \phi(a^{(n)}, b^{(n)})$ with $\phi : (\mathbb{R}_*^+)^{2N} \rightarrow (\mathbb{R}_*^+)^{2N}$ defined by

$$\phi(a, b) = \left(a^{1-\omega} \left(\frac{f}{Xb} \right)^\omega, b^{1-\omega} \left(\frac{g}{a^{1-\omega} \left(\frac{f}{Xb} \right)^\omega} \right)^\omega \right)$$

It admits the following Jacobian matrix

$$J\phi(\mathbf{1}) = \begin{pmatrix} (1 - \omega)I & -\omega d(\frac{1}{f})X \\ \omega(\omega - 1) d(\frac{1}{g})X^T & (1 - \omega)I + \omega^2 d(\frac{1}{g})X^T d(\frac{1}{f})X \end{pmatrix}$$

For both cases, we observe in Figures 3.6, 3.7 (for which we used our usual setting) that the iterates converge linearly with rate given by the second eigenvalue of $J\phi(\mathbf{1})$.

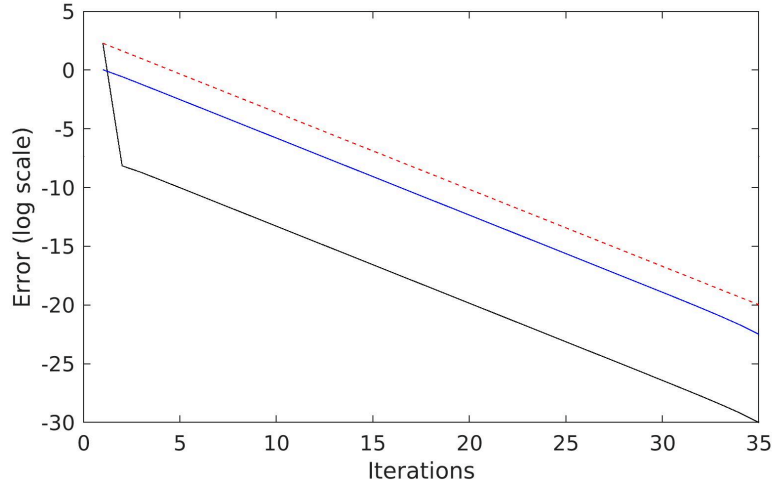


Figure 3.6 – Error evolution of the Sinkhorn-like algorithm I.6 with $K = 3$ (α in plain black, β in plain blue) and λ_2 rate of the Jacobian matrix (dotted red)

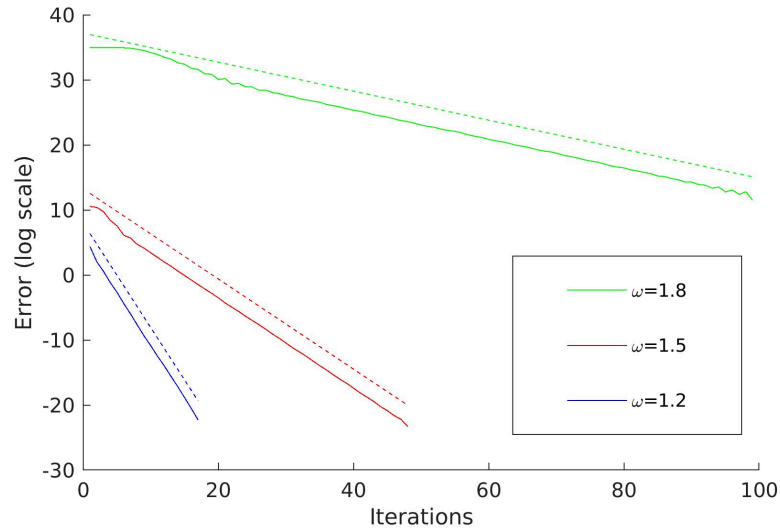


Figure 3.7 – Error evolution of the Sinkhorn-like algorithm I.7 (plain lines) and λ_2 rate of the Jacobian matrix (dotted lines) for different values of ω

CHAPTER 3. LINEAR CONVERGENCE OF SINKHORN-LIKE ALGORITHMS

Finally, the use of the quotient space E in Theorem I.17 is a natural way to get rid of the degenerate direction associated to the eigenvalue 1. The cancellation of this eigenvalue in (3.24) was also the objective of the so-called Wielandt deflation used by Knight in [Kni08]. We can say that dealing with this degenerate direction is precisely the point of any convergence analysis of the Sinkhorn algorithm. The method we proposed in the previous sections through the use of Theorem I.7 achieves this goal in an original way; however, it is possible that this last point of view could be generalized to a much wider class of Sinkhorn-like algorithms.

CHAPTER 4

ISSUES OF THE SETTING WHEN $\varepsilon \rightarrow 0$

Remember from section 1.2 that the Sinkhorn algorithm was introduced in the context of Optimal Transport to solve the *approximate* transportation problem:

$$x(\varepsilon) = \arg \min_{x \in \Pi(f,g)} \langle c|x \rangle + \varepsilon \langle x | \log x - \mathbf{1} \rangle \quad (4.1)$$

We saw that $x(\varepsilon)$ can be obtained by performing the Sinkhorn iterates on the matrix $A = \exp(\frac{-c}{\varepsilon})$. However, what is actually sought for is a solution of the exact transportation problem:

$$x^* \in \arg \min_{x \in \Pi(f,g)} \langle c|x \rangle \quad (4.2)$$

In this chapter, we are interested in the behavior of $x(\varepsilon)$, as well as the behavior of its computation through the Sinkhorn algorithm, when ε tends to 0.

We first present the theoretical results about the function $\varepsilon \mapsto x(\varepsilon)$ next turn to computational issues occurring for small values of ε . In the last subsection we extensively treat the case of the dimension $N = 2$ for which all computations can be led explicitly, enlightening the dependency of the convergence rates to ε .

4.1 Theoretical issues

First of all, the strategy of tackling the transportation problem (4.2) through a perturbed problem (4.1) is of course appropriate in the sense that one recovers x^* as the limit of $x(\varepsilon)$ when ε tends to 0. To be precise, as the exact transportation problem (4.2) may have several solutions, $x(\varepsilon)$ converges to the solution with the highest entropy:

Proposition I.7. *As $\varepsilon \rightarrow 0$, one has $x(\varepsilon) \rightarrow x^*$ where*

$$x^* = \arg \min_{x \in \mathcal{X}} \langle x | \log(x) \rangle \text{ and } \mathcal{X} = \arg \min_{x \in \Pi(f,g)} \langle c | x \rangle$$

Proof. The result follows from the fact that $x \mapsto \langle x | \log x - \mathbf{1} \rangle$ is continuous on $\Pi(f, g)$. Indeed, as $x(\varepsilon)$ belongs to the compact set $\Pi(f, g)$ for all $\varepsilon > 0$ one can suppose $x(\varepsilon) \rightarrow x^0$ for some $x^0 \in \Pi(f, g)$. First, passing to the limit $\varepsilon \rightarrow 0$ in

$$\forall x \in \Pi(f, g), \langle c | x(\varepsilon) \rangle + \varepsilon \langle x(\varepsilon) | \log x(\varepsilon) - \mathbf{1} \rangle \leq \langle c | x \rangle + \varepsilon \langle x | \log x - \mathbf{1} \rangle$$

gives $x^0 \in \mathcal{X}$. Second, for any $x \in \mathcal{X}$ one has

$$\begin{aligned} \langle x | \log x - \mathbf{1} \rangle &= \frac{1}{\varepsilon} (\langle c | x \rangle + \varepsilon \langle x | \log x - \mathbf{1} \rangle) - \frac{1}{\varepsilon} \langle c | x \rangle \\ &\geq \frac{1}{\varepsilon} (\langle c | x(\varepsilon) \rangle + \varepsilon \langle x(\varepsilon) | \log x(\varepsilon) - \mathbf{1} \rangle) - \frac{1}{\varepsilon} \langle c | x \rangle \end{aligned} \quad (4.3)$$

$$\begin{aligned} &= \langle x(\varepsilon) | \log x(\varepsilon) - \mathbf{1} \rangle + \frac{1}{\varepsilon} (\langle c | x(\varepsilon) \rangle - \langle c | x \rangle) \\ &\geq \langle x(\varepsilon) | \log x(\varepsilon) - \mathbf{1} \rangle \end{aligned} \quad (4.4)$$

where we used the definition of $x(\varepsilon)$ in (4.3) and the fact that $x \in \mathcal{X}$ in (4.4). As $\langle x(\varepsilon) | \mathbf{1} \rangle = \langle x | \mathbf{1} \rangle = 1$, taking $\varepsilon \rightarrow 0$ in that last inequality finally gives $x^0 = x^*$. \square

The strong convexity of the entropic regularization implies that the solution x^* belongs to the relative interior of the set of solutions of the optimal transport problem. In the case of the assignment problem – that is $f = g = \mathbf{1}$, see equations (1.11), (1.12) – this has an important consequence: although there always exists an optimal transport plan that takes the form of a permutation matrix, as soon as we do not have uniqueness of this solution the limit x^* of $x(\varepsilon)$ is *not* a permutation matrix. More generally, the limit x^* will always have the worse sparsity among the solutions of the optimal transport problem.

Another interesting question to be raised is the regularity of the function $\varepsilon \mapsto x(\varepsilon)$. For this study, it is useful to remember that the primal problem (4.1) defining $x(\varepsilon)$ admits a *unconstrained* dual problem:

$$(\lambda(\varepsilon), \mu(\varepsilon)) \in \arg \max_{\lambda, \mu \in \mathbb{R}^N} \langle \lambda | f \rangle + \langle \mu | g \rangle - \varepsilon \sum_{i,j} \exp \left(\frac{-w_{i,j} + \lambda_i + \mu_j}{\varepsilon} \right) \quad (4.5)$$

The solution $(\lambda(\varepsilon), \mu(\varepsilon))$ is defined up to a constant, meaning that only the quantities $\lambda(\varepsilon)_i + \mu(\varepsilon)_j$ are uniquely defined and that if (λ, μ) and $(\hat{\lambda}, \hat{\mu})$ are two solutions of equation (4.5) then there exists a constant $r \in \mathbb{R}$ such that $\hat{\lambda} = \lambda + r\mathbf{1}$ and $\hat{\mu} = \mu - r\mathbf{1}$. Solutions of the primal and dual problems are linked through

$$x(\varepsilon)_{i,j} = \exp \left(\frac{\lambda(\varepsilon)_i + \mu(\varepsilon)_j - w_{i,j}}{\varepsilon} \right) = a_{i,j} \exp \left(\frac{\lambda(\varepsilon)_i + \mu(\varepsilon)_j}{\varepsilon} \right) \quad (4.6)$$

To state precisely our results, we define $(\lambda(\varepsilon), \mu(\varepsilon))$ to be the unique solution of (4.5) such that $\lambda(\varepsilon)_N = 0$. Doing so we can make use of the inverse function theorem to state the regularity of $\varepsilon \mapsto x(\varepsilon)$:

Proposition I.8. $\lambda(\varepsilon)$, $\mu(\varepsilon)$ and $x(\varepsilon)$ are C^∞ functions of ε .

Proof. Optimality conditions in problem (4.5) state that $(\lambda(\varepsilon), \mu(\varepsilon)) = (\ell(\varepsilon), 0, \mu(\varepsilon))$ where $(\ell(\varepsilon), \mu(\varepsilon))$ is the only zero of the function $f_\varepsilon : \mathbb{R}^{N-1} \times \mathbb{R}^N \rightarrow \mathbb{R}^{N-1} \times \mathbb{R}^N$ defined by

$$f_\varepsilon(\ell, \mu) = \left(\left\{ f_i - \sum_{j=1}^N a_{i,j} \exp\left(\frac{\ell_i + \mu_j}{\varepsilon}\right) \right\}_{1 \leq i \leq N-1}, \right. \\ \left. \left\{ g_j - a_{N,j} \exp\left(\frac{\mu_j}{\varepsilon}\right) - \sum_{i=1}^{N-1} a_{i,j} \exp\left(\frac{\ell_i + \mu_j}{\varepsilon}\right) \right\}_{1 \leq j \leq N} \right)$$

As a result the function

$$f : \mathbb{R}_*^+ \times \mathbb{R}^{N-1} \times \mathbb{R}^N \rightarrow \mathbb{R}^{N-1} \times \mathbb{R}^N \\ : (\varepsilon, \ell, \mu) \mapsto f_\varepsilon(\ell, \mu)$$

satisfies $f(\varepsilon, \ell, \mu) = 0 \Leftrightarrow (\ell, 0, \mu) = (\lambda(\varepsilon), \mu(\varepsilon))$. Provided the partial derivative of f with respect to (ℓ, μ) is invertible, the inverse function theorem will transfer the regularity of f with respect to ε to $(\lambda(\varepsilon), \mu(\varepsilon))$ and, as a consequence of (4.6) to $x(\varepsilon)$. Writing $(f_i^\ell)_{1 \leq i \leq N-1}$ et $(f_j^\mu)_{1 \leq j \leq N}$ for the components of f , one has

$$\frac{\partial f}{\partial(\ell, \mu)} = \begin{pmatrix} \left(\frac{\partial f_i^\ell}{\partial \ell_{i_2}} \right)_{\substack{1 \leq i_1, i_2 \leq N-1}} & \left(\frac{\partial f_i^\ell}{\partial \mu_j} \right)_{\substack{1 \leq i \leq N-1 \\ 1 \leq j \leq N}} \\ \left(\frac{\partial f_j^\mu}{\partial \ell_i} \right)_{\substack{1 \leq j \leq N \\ 1 \leq i \leq N-1}} & \left(\frac{\partial f_{j_1}^\mu}{\partial \mu_{j_2}} \right)_{1 \leq j_1, j_2 \leq N} \end{pmatrix}$$

and one computes $\frac{\partial f_i^\ell}{\partial \mu_j} = \frac{\partial f_j^\mu}{\partial \ell_i} = \frac{-x_{i,j}}{\varepsilon}$ as well as

$$\frac{\partial f_{i_1}^\ell}{\partial \ell_{i_2}} = \begin{cases} \frac{-1}{\varepsilon} \sum_{j=1}^N x_{i,j} & \text{if } i_1 = i_2 = i \\ 0 & \text{otherwise} \end{cases} ; \quad \frac{\partial f_{j_1}^\mu}{\partial \mu_{j_2}} = \begin{cases} \frac{-1}{\varepsilon} \sum_{i=1}^N x_{i,j} & \text{if } j_1 = j_2 = j \\ 0 & \text{otherwise} \end{cases}$$

where $\forall j \in \llbracket 1, N \rrbracket$, $x_{i,j} = a_{i,j} \exp\left(\frac{\ell_i + \mu_j}{\varepsilon}\right)$ if $i \in \llbracket 1, N \rrbracket$ and $x_{N,j} = a_{N,j} \exp\left(\frac{\mu_j}{\varepsilon}\right)$ are positive.

Doing so, one sees that the matrix $M = -\varepsilon \frac{\partial f}{\partial(\ell, \mu)}$ is symmetric positive definite as for any $U = (u, v) \neq 0 \in \mathbb{R}^{N-1} \times \mathbb{R}^N$:

$$U^T M U = \sum_{i=1}^{N-1} u_i \left(u_i \sum_{j=1}^N x_{i,j} + \sum_{j=1}^N x_{i,j} v_j \right) + \sum_{j=1}^N v_j \left(v_j \sum_{i=1}^N x_{i,j} + \sum_{i=1}^{N-1} x_{i,j} u_i \right) \\ = \sum_{j=1}^N v_j^2 x_{N,j} + \sum_{\substack{i \in \llbracket 1, N-1 \rrbracket \\ j \in \llbracket 1, N \rrbracket}} x_{i,j} (u_i + v_j)^2 > 0$$

hence $\frac{\partial f}{\partial(\ell, \mu)}$ is invertible. \square

In [SGG11], Sharify and co-authors even obtain the analyticity of x : making use of the theory of real ordered fields they establish that each coefficient of the matrix $x(\varepsilon)$ writes as a generalized Dirichlet series in $\exp(-\frac{1}{\varepsilon})$ meaning that there exist real numbers c_n such that

$$x(\varepsilon)_{i,j} = \sum_{n=0}^{\infty} c_n \exp\left(-\frac{\alpha_n}{\varepsilon}\right)$$

where the sequence of real numbers (α_n) is such that $\alpha_n \rightarrow +\infty$.

Remark I.13. *Note that we only stated here the regularity of $x(\varepsilon)$ according to ε . The question of the regularity of the limit matrix X of the Sinkhorn algorithm according to the iterated matrix A may be more subtle. We ought to mention [Sin72] where Sinkhorn proves the continuity of X with respect to A in the doubly stochastic case.*

We now address the convergence rate of $x(\varepsilon)$ towards x^* . It is deeply linked to the convergence of the dual solution $(\lambda(\varepsilon), \mu(\varepsilon))$ towards a solution (λ^*, μ^*) of the unregularized dual problem (1.7)

$$\max_{\substack{\lambda, \mu \in \mathbb{R}^N \text{ s.t.} \\ \forall i, j, \lambda_i + \mu_j \leq c_{i,j}}} \langle \lambda | f \rangle + \langle \mu | g \rangle$$

Cominetti and San Martin gave in [CSM94] the proof of the following result:

Theorem I.18. *There exist vectors $\lambda^*, \mu^* \in \mathbb{R}^N$ and $d^* \in \mathbb{R}^{2N}$ such that for all $\varepsilon > 0$,*

$$(\lambda(\varepsilon), \mu(\varepsilon)) = (\lambda^*, \mu^*) + \varepsilon d^* + \eta(\varepsilon)$$

for a function $\eta : \mathbb{R}_*^+ \rightarrow \mathbb{R}^{2N}$ converging to 0 exponentially fast, meaning that there exist constants $K, c > 0$ such that for ε small enough

$$\|\eta(\varepsilon)\|_{\infty} \leq K \exp\left(-\frac{c}{\varepsilon}\right)$$

Remark I.14. *The limit value (λ^*, μ^*) is called the centroid of the polytope of solutions of the dual problem (1.7). It is defined by a shrinkage of constraints that remain unsaturated by the solutions of this linear problem.*

From this point we can use the relation between primal and dual solutions (4.6) and first understand the following: if (i, j) is such that $\lambda_i^* + \mu_j^* < w_{i,j}$ then we will have $x_{i,j}^* = 0$ and $x(\varepsilon)_{i,j}$ will tend to 0 exponentially fast. However if $\lambda_i^* + \mu_j^* = w_{i,j}$ then $x_{i,j}^* > 0$ will be given by $\exp(d^*)$ and again convergence of $x(\varepsilon)_{i,j}$ towards $x_{i,j}^*$ is exponential. The study in [CSM94] is a bit more precise and gives values for the constants of the exponential convergence:

Theorem I.19. *Let λ^*, μ^* be the limits of $\lambda(\varepsilon), \mu(\varepsilon)$ when $\varepsilon \rightarrow 0$. Denote $I = \{(i, j) \in \llbracket 1, N \rrbracket^2 \text{ s.t. } \lambda_i^* + \mu_j^* < w_{i,j}\}$, $C = \min_{(i,j) \in I} w_{i,j} - \lambda_i^* - \mu_j^*$ and $K = \sum_{(i,j) \in I} w_{i,j} - \lambda_i^* - \mu_j^*$ then for any $c < C$ there exists $\varepsilon_0 > 0$ such that*

$$\forall \varepsilon \leq \varepsilon_0, \|x(\varepsilon) - x^*\|_{\infty} \leq \frac{2\sqrt{K}}{c} \exp\left(-\frac{c}{2\varepsilon}\right)$$

Remark I.15. *This theorem remains correct in the case where $I = \emptyset$ taking $K = 0$ (and $C = +\infty$). Indeed, in this situation, the cost matrix w splits into $\lambda^* \oplus \mu^*$ hence any transport plan $x \in \Pi(f, g)$ is optimal. Then according to the Sinkhorn algorithm on the splitted matrix $A = \exp(-\frac{\lambda^*}{\varepsilon}) \otimes \exp(-\frac{\mu^*}{\varepsilon})$, $x(\varepsilon)$ is equal for any ε to $x^* = f \otimes g$ (which is as a consequence the element of $\Pi(f, g)$ with highest entropy).*

Remark I.16. *We focused on the simple optimal transport setting. However the results of this section rely only on the fact that we considered an entropic regularization of a linear problem. As such, they remain correct for the barycenter and graph variants we presented above.*

4.2 Numerical issues

In the context of Optimal Transport, one wants to perform the Sinkhorn algorithm on the matrix $A = \exp(-\frac{w}{\varepsilon})$ where $w \in \mathbb{R}^{N \times N}$ is the ground cost matrix. When taking ε close to zero, the values appearing in the matrix A become very small and numerical issues arise. In addition, it is worth noticing that the Sinkhorn algorithm suffers from a great dependency of the zero pattern of the matrix it is applied to, see the general theorems of section 2.2.1 As a consequence, if some entries of the matrix A are numerically set to 0 the pattern of the matrix is changed and divergence of the iterations may occur. This phenomenon is added to the general computational errors that arise to any procedure involving calculus with $\exp(-\frac{w}{\varepsilon})$ terms.

4.2.1 Log domain computation

A first numerical trick that helps improving the range of the parameter ε one can access to is to work on the “log domain”, meaning that one computes the Sinkhorn iterates on the variables $\lambda = \varepsilon \log \alpha, \mu = \varepsilon \log \beta$ rather than on the usual α, β . This method is presented for instance in [Sch19, SGG11] and is widely used in practice (see [PC19]). Remember we saw in equation (1.9) that the Sinkhorn iterates on these variables write:

$$\begin{aligned} \lambda_i^{(n+1)} &= \varepsilon \log(f_i) - \varepsilon \log \left(\sum_{j=1}^N \exp \left(\frac{\mu_j^{(n)} - w_{i,j}}{\varepsilon} \right) \right) \\ \mu_j^{(n+1)} &= \varepsilon \log(g_j) - \varepsilon \log \left(\sum_{i=1}^N \exp \left(\frac{\lambda_i^{(n+1)} - w_{i,j}}{\varepsilon} \right) \right) \end{aligned} \quad (4.7)$$

The $\varepsilon \log \sum \exp(\frac{\cdot}{\varepsilon})$ operator that appears in these expressions is often called a *softmax* operator as it approaches the maximum of a vector when ε goes to 0:

Proposition I.9. *For any $x \in \mathbb{R}^N$ one has when $\varepsilon \rightarrow 0$*

$$\varepsilon \log \sum_{k=1}^N \exp \left(\frac{x_k}{\varepsilon} \right) \rightarrow \max_{1 \leq k \leq N} x_k$$

Proof. Denote $x_{k^*} = \max_k x_k$. First, $\varepsilon \log \sum_k \exp\left(\frac{x_k}{\varepsilon}\right) \geq x_{k^*}$. Second, as for any k , one has $x_k - x_{k^*} \leq 0$, we can write

$$\varepsilon \log \sum_k \exp\left(\frac{x_k}{\varepsilon}\right) - x_{k^*} = \varepsilon \log \sum_k \exp\left(\frac{x_k - x_{k^*}}{\varepsilon}\right) \leq \varepsilon \log N$$

which concludes the proof. \square

In the case of the assignment problem, that is $f = g = \mathbf{1}$, the limit as $\varepsilon \rightarrow 0$ of the iterations (4.7) is consequently

$$\lambda_i^{(n+1)} = \min_{1 \leq j \leq N} w_{i,j} - \mu_j^{(n)}; \mu_j^{(n+1)} = \min_{1 \leq i \leq N} w_{i,j} - \lambda_i^{(n+1)}$$

which is exactly the “bidding” step of the auction algorithm (or Hungarian method, see [Kuh55, BE88, BC89, Wal17, PC19]). In the general setting, it seems fundamental to keep the appearance of the marginals f and g . We therefore are interested in the first order approximation of (4.7). Taking care of the number of times the maximum is reached, that is:

$$\begin{aligned} \lambda_i^{(n+1)} &= w_{i,j^*} - \mu_{j^*}^{(n)} + \varepsilon \log\left(\frac{f_i}{|J^*|}\right) \text{ where } j^* \in J^* := \arg \min_{1 \leq j \leq N} w_{i,j} - \mu_j^{(n)} \\ \mu_j^{(n+1)} &= w_{i^*,j} - \lambda_{i^*}^{(n+1)} + \varepsilon \log\left(\frac{g_j}{|I^*|}\right) \text{ where } i^* \in I^* := \arg \min_{1 \leq i \leq N} w_{i,j} - \lambda_i^{(n+1)} \end{aligned} \quad (4.8)$$

Equations (4.8) are consequently a first order approximation of the Sinkhorn iterates that one can use while treating the remaining $o(\varepsilon)$ terms such as

$$-\varepsilon \log \left(1 + \frac{1}{|J^*|} \sum_{j \notin J^*} \exp\left(\frac{\mu_j^{(n)} - w_{i,j} - \mu_{j^*}^{(n)} + w_{i,j^*}}{\varepsilon}\right) \right)$$

with the Taylor expansion of $\log(1+x)$. This leads to greater stability of the iterates and allows one to reach smaller values of ε . Similar formulas could be obtained for Sinkhorn-like algorithms.

4.2.2 Bethe entropy for barycenter

The quality of the approximation $x(\varepsilon)$ of x^* is governed by the size of the chosen regularization $\varepsilon \langle x | \log x - \mathbf{1} \rangle$. To lower down this perturbation while keeping the benefits of the entropic regularizer – that is its convexity and the Sinkhorn-like structure it provides – one can use the Bethe entropy presented section 3.2 about Sinkhorn algorithm on a graph. In this section, we present this strategy in the case of the simple barycenter setting described section 3.1.1.

Remember the form of the linear program defining the barycenter of $f^0, f^1 \in \Sigma^N$:

$$\begin{aligned} \arg \min & \theta \langle w^0 | x^0 \rangle + (1 - \theta) \langle w^1 | x^1 \rangle \\ & x^0 \mathbf{1} = f^0, x^1 \mathbf{1} = f^1 \\ & x^0 \mathbf{1} = x^1 \mathbf{1} = f^\theta \end{aligned}$$

To lower our usual entropic perturbation acting on x^0, x^1 , it is possible to substract the term $(1 - \delta)\varepsilon\langle f^\theta | \log f^\theta - \mathbf{1} \rangle$ while keeping the convexity of the energy. Indeed, under the constraint $x^{0T}\mathbf{1} = x^1\mathbf{1} = f^\theta$ one can write

$$\begin{aligned} & \theta\langle x^0 | \log x^0 - \mathbf{1} \rangle + (1 - \theta)\langle x^1 | \log x^1 - \mathbf{1} \rangle - (1 - \delta)\langle f^\theta | \log f^\theta - \mathbf{1} \rangle \\ &= \theta\langle x^0 | \log \frac{x^0}{f^\theta \otimes \mathbf{1}} - \mathbf{1} \rangle + (1 - \theta)\langle x^1 | \log \frac{x^1}{\mathbf{1} \otimes f^\theta} - \mathbf{1} \rangle + \delta\langle f^\theta | \log f^\theta - \mathbf{1} \rangle \end{aligned}$$

and invoque the strict convexity of $(x, y) \mapsto x \log \frac{x}{y}$ for $0 < x < y$. Finally, we write for $\varepsilon, \delta > 0$ the following optimization problem

$$\begin{aligned} & \min_{\substack{x^0\mathbf{1}=f^0, x^{1T}\mathbf{1}=f^1 \\ x^{0T}\mathbf{1}=x^1\mathbf{1}=f^\theta}} \theta\langle w^0 | x^0 \rangle + (1 - \theta)\langle w^1 | x^1 \rangle \\ & + \varepsilon\theta\langle x^0 | \log \frac{x^0}{f^\theta \otimes \mathbf{1}} - \mathbf{1} \rangle + \varepsilon(1 - \theta)\langle x^1 | \log \frac{x^1}{\mathbf{1} \otimes f^\theta} - \mathbf{1} \rangle \\ & + \varepsilon\delta\langle f^\theta | \log f^\theta - \mathbf{1} \rangle \end{aligned}$$

which is a lower-regularized version of (3.1) when $\delta < 1$. We introduce the Lagrange multipliers $\lambda^0, \lambda^1, \mu^0, \mu^1$ for respectively $x^0\mathbf{1} = f^0, x^{1T}\mathbf{1} = f^1, x^{0T}\mathbf{1} = f^\theta, x^1\mathbf{1} = f^\theta$ followed by our usual changes of variables

$$\begin{aligned} \alpha &= \exp\left(-\frac{\lambda^0}{\varepsilon\theta}\right); \quad \beta = \exp\left(-\frac{\lambda^1}{\varepsilon(1-\theta)}\right) \\ \gamma^k &= \exp\left(\frac{\mu^k}{\varepsilon}\right); \quad A^k = \exp\left(-\frac{w^k}{\varepsilon}\right) \text{ for } k \in \{0, 1\} \end{aligned}$$

Then one checks that (x^0, x^1, f^θ) is solution of our problem if and only if it writes $x^0 = d(\alpha)A^0 d(\gamma^0)$, $x^1 = d(\gamma^1)A^1 d(\beta)$ and $f^\theta = ((A^{0T}\alpha)^\theta (A^1\beta)^{1-\theta})^{\frac{1}{\delta}}$ for $(\alpha, \beta, \gamma^0, \gamma^1)$ such that

$$\begin{aligned} \alpha &= \frac{f^0}{A^0\gamma^0}; \quad \beta = \frac{f^1}{A^{1T}\gamma^1} \\ \gamma^0 &= (A^{0T}\alpha)^{\frac{\theta}{\delta}-1} (A^1\beta)^{\frac{1-\theta}{\delta}}; \quad \gamma^1 = (A^{0T}\alpha)^{\frac{\theta}{\delta}} (A^1\beta)^{\frac{1-\theta}{\delta}-1} \end{aligned}$$

This leads to our last Sinkhorn-like algorithm:

Algorithm I.9. Given two positive matrices $A^0, A^1 \in (\mathbb{R}_*^+)^{N \times N}$, two marginals $f^0, f^1 \in \Sigma^N$ and $\delta > 0$, from $\alpha^{(0)} = \beta^{(0)} = \gamma^{0(0)} = \gamma^{1(0)} = \mathbf{I}$ do for $n = 0, 1, \dots$

$$\begin{aligned} \alpha^{(n+1)} &= \frac{f^0}{A^0\gamma^{0(n)}}; \quad \beta^{(n+1)} = \frac{f^1}{A^{1T}\gamma^{1(n)}} \\ \gamma^{0(n+1)} &= \left(A^{0T}\alpha^{(n+1)}\right)^{\frac{\theta}{\delta}-1} \left(A^1\beta^{(n+1)}\right)^{\frac{1-\theta}{\delta}} \\ \gamma^{1(n+1)} &= \left(A^{0T}\alpha^{(n+1)}\right)^{\frac{\theta}{\delta}} \left(A^1\beta^{(n+1)}\right)^{\frac{1-\theta}{\delta}-1} \end{aligned} \tag{4.9}$$

This algorithm coincides with the classical Sinkhorn-like algorithm for the barycenter I.2 when $\delta = 1$ through the variable $\gamma = \gamma^0 \frac{1}{\delta-1} = \gamma^1 \frac{1}{\delta}$. Taking $\delta < 1$ actually improves the quality of the approximation for simple barycenter problems. We present in Figure 4.1 the result obtained for a barycenter of two shapes: f^0 and f^1 are given as indicator functions of circles, $A^0 = A^1 = \exp(-\frac{w}{\varepsilon})$ for an L^2 cost w on the grid $[0, 1]^2$ discretized into a 60×60 image size and $\varepsilon = 10^{-3}$. We compute the barycenter for $\theta = 0.5$ and compare the case $\delta = 1$ corresponding to the classical Sinkhorn-like algorithm I.2 to the case $\delta = 0.51$ (values of δ lower than 0.5 being theoretically excluded, and numerically unstable as well).

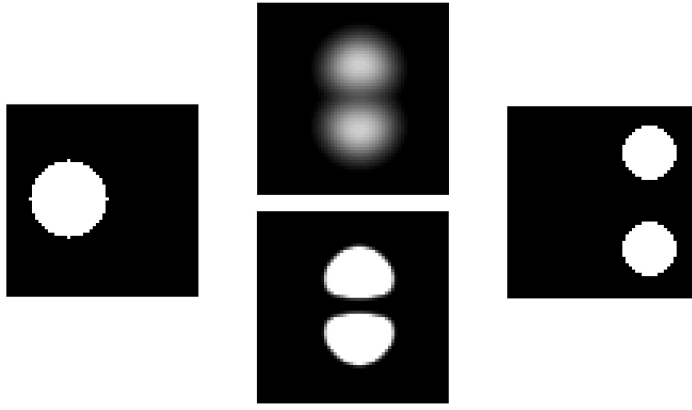


Figure 4.1 – A Wasserstein barycenter $f^{\frac{1}{2}}$ of two marginals f^0 (left) and f^1 (right) computed through algorithm I.9 with $\delta = 1$ (top middle) and $\delta = 0.51$ (bottom middle)

Finally, we are able to conduct an analysis of these iterations in Hilbert metric:

Theorem I.20. *For $\delta \geq \max(\theta, 1 - \theta)$, the Sinkhorn iterates $\alpha^{(n)}, \beta^{(n)}, \gamma^{0(n)}, \gamma^{1(n)}$ defined by (4.9) converge in $(\mathbb{R}_*^+)^N / \sim$ to the fixed point $\alpha^*, \beta^*, \gamma^{0*}, \gamma^{1*}$ with the following estimate: $\forall n \geq 0$,*

$$d_H(\gamma^{0(n+1)}, \gamma^{0*}) + d_H(\gamma^{1(n+1)}, \gamma^{1*}) \leq \kappa^2 (d_H(\gamma^{0(n)}, \gamma^{0*}) + d_H(\gamma^{1(n)}, \gamma^{1*}))$$

where $\kappa = \max(\kappa(A^0), \kappa(A^1)) < 1$.

Proof. Using Proposition I.1 we get

$$\begin{aligned} d_H(\gamma^{0(n+1)}, \gamma^{0*}) &\leq \frac{1-\theta}{\delta} \kappa d_H(\beta^{(n+1)}, \beta^*) + \left| \frac{\theta}{\delta} - 1 \right| \kappa d_H(\alpha^{(n+1)}, \alpha^*) \\ &\leq \kappa^2 \left(\frac{1-\theta}{\delta} d_H(\gamma^{1(n+1)}, \gamma^{1*}) + \left| \frac{\theta}{\delta} - 1 \right| d_H(\gamma^{0(n+1)}, \gamma^{0*}) \right) \end{aligned}$$

And similarly

$$d_H(\gamma^{1(n+1)}, \gamma^{1*}) \leq \kappa^2 \left(\left| \frac{1-\theta}{\delta} - 1 \right| d_H(\gamma^{1(n+1)}, \gamma^{1*}) + \frac{\theta}{\delta} d_H(\gamma^{0(n+1)}, \gamma^{0*}) \right)$$

The hypothesis $\delta \geq \max(\theta, 1 - \theta)$ allows one to get rid of the absolute values and the result follows summing the two estimates. \square

Once again, numerical experiments (see Figure 4.2) show that the actual convergence rate is governed by the subdominant eigenvalue of the matrix M below, that is the Jacobian matrix of the iterations (calculated at the fixed point and on rescaled variables):

$$M = \begin{pmatrix} \left(1 - \frac{\theta}{\delta}\right) M^0 & \frac{\theta - 1}{\delta} M^1 \\ -\frac{\theta}{\delta} M^0 & \left(1 + \frac{\theta - 1}{\delta}\right) M^1 \end{pmatrix}$$

where $M^0 = d(\frac{1}{f^\theta})X^{0T} d(\frac{1}{f})X^0$ and $M^1 = d(\frac{1}{f^\theta})X^1 d(\frac{1}{g})X^{1T}$.

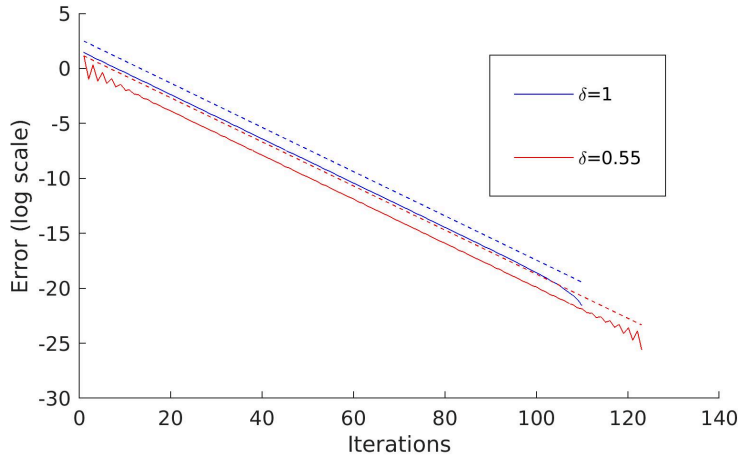


Figure 4.2 – Error evolution of the Sinkhorn-like algorithm I.9 (plain lines) and λ_2 rates of the Jacobian matrix (dotted lines) for two extreme values of δ

4.2.3 Iterated process

If working in the log domain or dealing with more subtle entropic regularization can allow one to reach smaller values of ε , one still has to perform operations involving $\exp(\frac{-w}{\varepsilon})$, and the computation still remains uncertain when ε is very small. In [XWWZ18], Xie and co-authors state that this is a real issue in the sense that some problems (such as the Wasserstein barycenter) actually demand to obtain the true optimal transport plan x^* at the limit $\varepsilon \rightarrow 0$. We explain below a way to witness in practice the convergence of $x(\varepsilon)$ towards x^* . This method was presented in [XWWZ18] (which appeared as we were studying it) together with applications to Wasserstein barycenters. It tackles the problem differently by fixing a value of ε for which the matrix $\exp(\frac{-w}{\varepsilon})$ is computable and building a sequence (x_ε^n) converging to x^* when $n \rightarrow +\infty$.

Remember we perturbed the transportation problem 1.4 into

$$x(\varepsilon) = \arg \min_{x \in \Pi(f, g)} \langle w | x \rangle + \varepsilon \langle x | \log x - \mathbf{1} \rangle$$

The entropy term $\langle x | \log x - \mathbf{1} \rangle$ can be interpreted as a Bregman divergence, a type of function introduced by [Bre67] and for which we take the definition of [CT93]:

Definition I.5. (Bregman divergence) *Let Ω be an open and convex subset of \mathbb{R}^d , and let $\psi : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ such that:*

- ψ is strongly convex with constant 1
- The domain of ψ is $\bar{\Omega}$
- ψ is continuous on $\bar{\Omega}$ and continuously differentiable on Ω
- The subdifferential of ψ is empty on the boundary of Ω : $\forall x \in \partial\Omega, \partial\psi(x) = \emptyset$

We define the Bregman divergence associated to ψ as

$$\begin{aligned} D_\psi &: \mathbb{R}^d \times \Omega \longrightarrow \mathbb{R} \cup \{+\infty\} \\ &: (x, y) \longmapsto \psi(x) - \psi(y) - \langle \nabla\psi(y) | x - y \rangle \end{aligned}$$

In our setting, the Bregman divergence is given by the function $\psi(x) = \langle x | \log x \rangle$ on $\Omega = (\mathbb{R}_*^+)^{N \times N}$, which leads to $D_\psi(x, y) = \langle x | \log \frac{x}{y} \rangle + \langle y - x | \mathbf{1} \rangle$. Then the perturbation we add to the transportation problem is $D_\psi(x, \mathbf{1})$.

One can actually interpret the Bregman divergence as a sort of nonsymmetric distance on Ω . This leads to saying that in our search of a first approximation $x_\varepsilon^1 = x(\varepsilon)$ of the exact optimal transport solution x^* we chose to penalize the distance from $x(\varepsilon)$ to $\mathbf{1}$. In the search of x^* it is interesting to now consider a second approximation x_ε^2 obtained by penalizing the distance to x_ε^1 , and pursue this process setting for all $n \geq 1$:

$$x_\varepsilon^{n+1} = \arg \min_{x \in \Pi(f, g)} \langle w | x \rangle + \varepsilon D_\psi(x, x_\varepsilon^n) \quad (4.10)$$

This method referred to as nonlinear proximal iteration (as it generalizes the Euclidean proximity operator given for $\psi(x) = \|x\|_2^2$) has been introduced in [CZ92]. Chen and Teboulle proved its convergence in a very general setting in [CT93]. We explain below the iterates to which it leads and what it can achieve.

It turns out that the update rule (4.10) rewrites as our original regularized transportation problem (1.5) up to a change of the cost matrix w . Indeed, suppose that $x_\varepsilon^n \in \Pi(f, g)$, then for any $x \in \Pi(f, g)$, one has $\langle x | \mathbf{1} \rangle = \langle x_\varepsilon^n | \mathbf{1} \rangle$ so that

$$\langle w | x \rangle + \varepsilon D_\psi(x, x_\varepsilon^n) = \langle w | x \rangle + \varepsilon \langle x | \log \frac{x}{x_\varepsilon^n} \rangle = \langle w - \varepsilon \log x_\varepsilon^n | x \rangle + \varepsilon \langle x | \log x \rangle$$

As a consequence, computing x_ε^n is achieved by leading the Sinkhorn algorithm on the matrix $A^n = \exp(-\frac{w - \varepsilon \log x_\varepsilon^n}{\varepsilon}) = A^0 \odot x_\varepsilon^n$ where $A^0 = \exp(-\frac{w}{\varepsilon})$, and where we recall that \odot stands for the componentwise matrix multiplication (and $\cdot^{\odot p}$ for the componentwise exponentiation). As several Sinkhorn iterations will be led, we take the following notation:

Notation. *For fixed marginals $f, g \in \Sigma^N$, and for a matrix $A \in (\mathbb{R}^+)^{N \times N}$ satisfying hypothesis (I.4), let D^1, D^2 be diagonal matrices such that $X = D^1 A D^2 \in \Pi(f, g)$. We denote $SK(A) := X$.*

Finally the proximal iterated algorithm (4.10) writes as:

Algorithm I.10. Given two marginals $f, g \in \Sigma^N$, a cost matrix $w \in \mathbb{R}^{N \times N}$ and $\varepsilon > 0$, compute $A^0 = \exp(-\frac{w}{\varepsilon})$, $x_\varepsilon^1 = \mathcal{SK}(A^0)$, and then for all $n \geq 1$,

$$A^n = A^0 \odot x_\varepsilon^n \text{ and } x_\varepsilon^{n+1} = \mathcal{SK}(A^n)$$

Remark I.17. When implementing such iterated Sinkhorn algorithms, one should keep in mind that they can be lead from any initial values of $\alpha^{(0)}, \beta^{(0)} \in (\mathbb{R}_*^+)^N$, that is not just from $\alpha^{(0)} = \beta^{(0)} = \mathbf{I}$. As the whole process converges, the calculation of $\mathcal{SK}(A^{n+1})$ would need less Sinkhorn steps if one starts with α and β set to the ending values obtained when calculating $\mathcal{SK}(A^n)$.

Actually, this n step process achieves the same result as dividing ε by n . To see this we make use of the following lemma which is a direct consequence of the uniqueness part of Theorem I.6:

Lemma I.10. Let A be a nonnegative matrix, and let D_1, D_2 be any diagonal positive matrices. Then

$$\mathcal{SK}(D_1 A D_2) = \mathcal{SK}(A)$$

We then have the announced result:

Proposition I.10. The sequence $(x_\varepsilon^n) \in (\mathbb{R}^{N \times N})^{\mathbb{N}}$ defined by algorithm I.10 is such that:

$$\forall n \geq 1, x_\varepsilon^n = x\left(\frac{\varepsilon}{n}\right)$$

where for all $\eta > 0$, $x(\eta) = \arg \min_{x \in \Pi(f, g)} \langle w|x \rangle + \eta \langle x | \log x - \mathbf{I} \rangle$.

Proof. In other words, we want to prove that $x_\varepsilon^n = \mathcal{SK}(\exp(-n\frac{w}{\varepsilon})) = \mathcal{SK}((A^0)^{\odot n})$. This is true for $n = 1$ and supposing the result is true for some $n \geq 1$, there exist diagonal positive matrices D_1, D_2 such that $x_\varepsilon^n = D_1(A^0)^{\odot n} D_2$ so that applying the previous lemma gives:

$$\begin{aligned} x_\varepsilon^{n+1} &= \mathcal{SK}(A^0 \odot x_\varepsilon^n) = \mathcal{SK}(A^0 \odot (D_1(A^0)^{\odot n} D_2)) \\ &= \mathcal{SK}(D_1(A^0)^{\odot n+1} D_2) \\ &= \mathcal{SK}((A^0)^{\odot n+1}) \end{aligned}$$

□

Remark I.18. Another proof of this result consists in writing the optimality conditions for equation (4.10). If we work in the affine space V spanned by $\Pi(f, g)$, one sees that they write $w|_V + \varepsilon (\nabla_x (D_\psi(x_\varepsilon^{n+1}, x_\varepsilon^n))) = w|_V + \varepsilon (\nabla \psi(x_\varepsilon^{n+1}) - \nabla \psi(x_\varepsilon^n)) = 0$ where $w|_V$ is the orthogonal projection of w in V and ∇ denotes de gradient operator in V . Then it leads to $\nabla \psi(x_\varepsilon^n) = \nabla \psi(x_\varepsilon^0) - \frac{nw}{\varepsilon}$ just as $x(\frac{\varepsilon}{n}) = \arg \min_x \langle w|x \rangle + \varepsilon D_\psi(x, x_\varepsilon^0)$ would. So finally $x_\varepsilon^n = x(\varepsilon/n)$ by uniqueness of the solution to this equation. One sees that the key argument here is that the original problem $\min_{x \in \Pi(f, g)} \langle x|x \rangle$ is linear.

Remember an issue of the direct calculation of $\mathcal{SK}(\exp(-\frac{w}{\varepsilon}))$ was the appearance of numerical zero entries modifying the pattern of the matrix to which the Sinkhorn algorithm is applied, hence possibly compromising its convergence. To this extent, calculating $x(\frac{\varepsilon}{n})$ through algorithm I.10 is more stable than the direct computation of $\mathcal{SK}(\exp(-n\frac{w}{\varepsilon}))$. Indeed, considering that $\mathcal{SK}(A)$ and A have the same pattern (which is true theoretically but can be lost numerically), the only way for entry (i, j) of x_ε^n to be numerically set to 0 is when it gets multiplied by $A_{i,j}^0 = \exp(-\frac{w_{i,j}}{\varepsilon})$. If ε is large enough, this will only happen after a repeated number of multiplications by A^0 , in other words only if n is large enough. In that view it becomes more likely that this entry is set to 0 because its corresponding limit value $x_{i,j}^*$ indeed vanishes.

We note that the relation between x_ε^n and $x(\varepsilon)$ simply relies on the good behavior of the Sinkhorn operator \mathcal{SK} with respect to diagonal scaling and componentwise multiplication. As such one can propose many other procedures similar to algorithm I.10 of the form $x_\varepsilon^{n+1} = \mathcal{SK}(y_\varepsilon^n \odot A^0)$ for some matrix y_ε^n depending on the previous iterates. This will lead to relations of the form $x_\varepsilon^n = x(\frac{\varepsilon}{u_n})$ for some real numbers $u_n \rightarrow +\infty$. One can suggest many schemes to get a sequence (u_n) that grows very fast to infinity; however, the efficiency of these procedures highly depends on the ability to compute without numerical errors the matrices y_ε^n . We give two methods below. The first one is an over-relaxed version of algorithm I.10:

Algorithm I.11. *Given two marginals $f, g \in \Sigma^N$, a cost matrix $w \in \mathbb{R}^{N \times N}$ and $\varepsilon > 0$, compute $A^0 = \exp(-\frac{w}{\varepsilon})$, $x_\varepsilon^0 = x_\varepsilon^1 = \mathcal{SK}(A^0)$, and then for all $n \geq 0$,*

$$A^{n+1} = A^0 \odot \frac{(x_\varepsilon^{n+1})^{\odot 2}}{x_\varepsilon^n} \quad \text{and} \quad x_\varepsilon^{n+2} = \mathcal{SK}(A^{n+1})$$

Achieving n steps of this algorithm theoretically corresponds to dividing the parameter ε by a $O(n^2)$ factor:

Proposition I.11. *The sequence $(x_\varepsilon^n) \in (\mathbb{R}^{N \times N})^{\mathbb{N}}$ defined by algorithm I.11 is such that:*

$$\forall n \geq 1, \quad x_\varepsilon^n = x\left(\frac{2\varepsilon}{n^2 - n + 2}\right)$$

Proof. The same reasoning as before shows that $x_\varepsilon^n = \mathcal{SK}(\exp(-u_n \frac{w}{\varepsilon}))$ for a sequence (u_n) such that $u_0 = u_1 = 1$ and $u_{n+2} = 2u_{n+1} - u_n + 1$ which is solved into $u_n = \frac{1}{2}n^2 - \frac{1}{2}n + 1$. \square

The second method we propose does componentwise exponentiation of the current approximation:

Algorithm I.12. *Given two marginals $f, g \in \Sigma^N$, a cost matrix $w \in \mathbb{R}^{N \times N}$, $p > 1$ and $\varepsilon > 0$, compute $A^0 = \exp(-\frac{w}{\varepsilon})$, $x_\varepsilon^1 = \mathcal{SK}(A^0)$ and then for all $n \geq 1$,*

$$A^n = A^0 \odot (x_\varepsilon^n)^{\odot p} \text{ and } x_\varepsilon^{n+1} = \mathcal{SK}(A^n)$$

Achieving n steps of this algorithm theoretically corresponds to dividing the parameter ε by a $O(p^n)$ factor:

Proposition I.12. *The sequence $(x_\varepsilon^n) \in (\mathbb{R}^{N \times N})^{\mathbb{N}}$ defined by algorithm I.12 is such that:*

$$\forall n \geq 1, x_\varepsilon^n = x \left(\frac{\varepsilon(p-1)}{p^n - 1} \right)$$

Proof. The same reasoning as before shows that $x_\varepsilon^n = \mathcal{SK}(\exp(-u_n \frac{w}{\varepsilon}))$ for a sequence (u_n) such that $u_1 = 1$ and $u_{n+1} = pu_n + 1$ which is solved into $u_n = \frac{p^n - 1}{p - 1}$. \square

In the doubly stochastic setting, and if the optimal assignment problem admits a unique optimal permutation x^* , methods like algorithm I.12 can achieve quick convergence to the solution at least in some cases (and provided one knows the appropriate size for the parameter ε , for instance by previously testing different parameters for similar ground costs w). Indeed, in this particular setting the entries of $x(\varepsilon)$ converge to either 0 or 1. As a consequence, the exponentiation step somehow realizes a thresholding of the entries of x_ε^n enhancing the gap between higher entries – that may converge to 1 – and smaller ones – that may converge to 0.

The disadvantage of all these iterated algorithms is that they rely on the computation of the matrices A^n on which the Sinkhorn algorithm is performed at every step. One must consequently successively store N^2 variables rather than the $2N$ variables needed to perform a classical Sinkhorn algorithm. In practice, to speed up computation one can stop the inside Sinkhorn algorithms before convergence happens. In [XWWZ18], the authors chose for instance to perform only one Sinkhorn iterate (that is, one step on the rows and one step on the columns) when following algorithm I.10. They also provide an error analysis to guarantee convergence of this inexact scheme.

Remark I.19. *Similar algorithms can be implemented for Sinkhorn-like variants. For instance, the same reasoning in the case of the barycenter of two measures, corresponds to first perform the usual Sinkhorn-like algorithm I.2*

$$\alpha^{(n+1)} = \frac{f^0}{A^0 (\frac{1}{\gamma^{(n)}})^{1-\theta}} ; \beta^{(n+1)} = \frac{f^1}{A^{1T} (\gamma^{(n)})^\theta} ; \gamma^{(n+1)} = \frac{A^{0T} \alpha^{(n+1)}}{A^1 \beta^{(n+1)}}$$

which leads to the first approximate transport plans $X^0 = d(\alpha^*) A^0 d(\gamma^{*\theta-1})$, $X^1 = d(\gamma^{*\theta}) A^1 d(\beta^*)$ and then repeat the same algorithm replacing the matrices A^0 and A^1 respectively by $A^0 \otimes X^0$ and $A^1 \otimes X^1$ and so on. In particular, this also motivates the differentiation of the matrices A^0 and A^1 although one uses the same ground cost w .

4.3 Behavior of rates in dimension 2

We believe the complete treatment of the dimension $N = 2$ enlightens the possible behavior of the convergence rates we described with respect to the coefficients of the matrix A or to ε . We start with the doubly stochastic setting where formulas are quite simple and then give explicit formulas, a bit more tedious to analyze, in the general case. We denote $A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}$ the nonnegative matrix on which the Sinkhorn algorithm is applied.

4.3.1 Doubly stochastic setting

As mentioned in section 2.2.1, in the doubly stochastic setting the matrix A must have total support for the Sinkhorn algorithm to converge. Excluding the trivial cases where the iterations split into two iterations on matrices of size $N = 1$, this only leaves us with the positive case $A > 0$. The setting in dimension 2 is easy to treat because the doubly stochastic matrices are precisely those of the form

$$B(t) = \begin{pmatrix} t & 1-t \\ 1-t & t \end{pmatrix}$$

for $t \in [0, 1]$. For $B(t)$ to be a scaling of the positive matrix A one more precisely has to have $t \in (0, 1)$. The simplest way of computing the unique $t \in (0, 1)$ such that $B(t)$ writes as $d(\alpha)A d(\beta)$ for positive vectors α, β is probably to use the following fact (extensively used in the analysis of the Sinkhorn iterates in particular [Sin64]):

Fact I.2. *If two positive matrices $A, B \in (\mathbb{R}_*^+)^{N \times N}$ are such that $B = d(\alpha)A d(\beta)$ for positive vectors $\alpha, \beta \in (\mathbb{R}_*^+)^N$, then for any permutations σ_1, σ_2 of $\llbracket 1, N \rrbracket$ one has*

$$\prod_{k=1}^N \frac{b_{k, \sigma_1(k)}}{b_{k, \sigma_2(k)}} = \prod_{k=1}^N \frac{a_{k, \sigma_1(k)}}{a_{k, \sigma_2(k)}}$$

When $N = 2$ this reduces to saying that the ratio $\Delta = \frac{a_{1,2}a_{2,1}}{a_{1,1}a_{2,2}}$ must be preserved after scaling. Hence one has $\frac{(1-t)^2}{t^2} = \Delta$ which leads to $t = \frac{1}{1 + \sqrt{\Delta}}$. To conclude, the convergence of the Sinkhorn iterates is governed by the spectrum of the matrix $B(t)B(t)^T$ which is $\{1, (2t-1)^2\}$ so that we get the following conclusion:

Proposition I.13. *In the doubly stochastic setting $f = g = \mathbf{1}$, the Sinkhorn algorithm applied to the matrix $A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}$ converges linearly with rate $\left(\frac{\sqrt{\Delta} - 1}{\sqrt{\Delta} + 1} \right)^2$ where $\Delta = \frac{a_{1,2}a_{2,1}}{a_{1,1}a_{2,2}}$.*

Note that the expression of this rate is invariant when replacing Δ by Δ^{-1} and as a consequence when replacing Δ by $\theta(A) = \max_{i,j,k,l} \frac{a_{i,k}a_{j,l}}{a_{j,k}a_{i,l}}$: finally in the doubly stochastic setting we recover the Hilbert distance contraction rate of Theorem I.2. One can also check that the case $\Delta = 1$, which corresponds to a linear speed of 0, occurs for matrices on which convergence is achieved in a finite number of Sinkhorn iterations.

To finish with, one can apply this result to the setting of the approximated assignment problem (1.12), that is making the assumption that $A = \exp\left(-\frac{w}{\varepsilon}\right)$ for some cost matrix $w \in \mathbb{R}^{2 \times 2}$. Then one finds $\Delta = \exp\left(\frac{\bar{w}}{\varepsilon}\right)$ for $\bar{w} = w_{1,1} + w_{2,2} - w_{1,2} - w_{2,1}$, so that the solution of the perturbed assignment problem is

$$x(\varepsilon) = \frac{1}{1 + e^{-\bar{w}/2\varepsilon}} \begin{pmatrix} e^{-\bar{w}/2\varepsilon} & 1 \\ 1 & e^{-\bar{w}/2\varepsilon} \end{pmatrix}$$

and of course the limit matrix $x^* = \lim_{\varepsilon \rightarrow 0} x(\varepsilon)$ is given by the sign of \bar{w} :

- If $\bar{w} > 0$ then $x^* = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\|x(\varepsilon) - x^*\| \underset{\varepsilon \rightarrow 0}{\sim} c e^{-\bar{w}/2\varepsilon}$.
- If $\bar{w} < 0$ then $x^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\|x(\varepsilon) - x^*\| \underset{\varepsilon \rightarrow 0}{\sim} c e^{\bar{w}/2\varepsilon}$.
- If $\bar{w} = 0$ then all the doubly stochastic matrices are solution and for all $\varepsilon > 0$, $x(\varepsilon) = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = x^*$ is the solution with the highest entropy (and lowest sparsity).

To finish with, expect when $\bar{w} = 0$, the convergence rate of the Sinkhorn algorithm converges to 1 when $\varepsilon \rightarrow 0$ as more precisely it is equal to

$$\lambda_2(\varepsilon) = \left(\frac{\exp(\frac{\bar{w}}{2\varepsilon}) - 1}{\exp(\frac{\bar{w}}{2\varepsilon}) + 1} \right)^2 = \left(1 - \frac{2}{\exp(\frac{\bar{w}}{2\varepsilon}) + 1} \right)^2$$

And $1 - \lambda_2(\varepsilon) \underset{\varepsilon \rightarrow 0}{\sim} 2 \exp\left(\frac{-|\bar{w}|}{2\varepsilon}\right)$. In particular, the speed of the Sinkhorn algorithm deteriorates as $\varepsilon \rightarrow 0$.

4.3.2 General setting

We now turn to the case of general marginals $f = \begin{pmatrix} f_1 \\ 1 - f_1 \end{pmatrix}$ and $g = \begin{pmatrix} g_1 \\ 1 - g_1 \end{pmatrix}$ for some $f_1, g_1 \in (0, 1)$. To avoid trivial cases, one can only allow up to one entry of A to vanish, for instance $a_{2,1} = 0$. We first treat this case and then focus on the positive case.

If $a_{2,1} = 0$, then according to section 2.2.1 one must also suppose that $f_1 > g_1$ for A to admit a scaling belonging to $\Pi(f, g)$. Respecting the zero pattern of A the limit matrix of the Sinkhorn iterates must be

$$X = \begin{pmatrix} g_1 & f_1 - g_1 \\ 0 & 1 - f_1 \end{pmatrix}$$

Then one checks that the matrix $M = d(\frac{1}{f})X d(\frac{1}{g})X^T$ is given by

$$M = \begin{pmatrix} \frac{g_1}{f_1} + \frac{(f_1 - g_1)^2}{f_1(1 - g_1)} & \frac{(f_1 - g_1)(1 - f_1)}{f_1(1 - g_1)} \\ \frac{f_1 - g_1}{1 - g_1} & \frac{1 - f_1}{1 - g_1} \end{pmatrix}$$

and its subdominant eigenvalue is $\lambda_2 = Tr(M) - 1 = \frac{g_1(1 - f_1)}{f_1(1 - g_1)}$. Note that as $1 > f_1 > g_1 > 0$ one has indeed $\lambda_2 \in (0, 1)$. It is noteworthy to mention that both the limit matrix X and this convergence rate do not depend on the entries of A .

In the positive setting, one looks for a limit matrix of the form

$$X(t) = \begin{pmatrix} t & f_1 - t \\ g_1 - t & 1 - f_1 - g_1 + t \end{pmatrix}$$

for some t such that $0 < t < f_1$ and $f_1 + g_1 - 1 < t < g_1$. Applying Fact I.2, one finds that t must satisfy

$$\frac{(g_1 - t)(f_1 - t)}{t(1 - f_1 - g_1 + t)} = \Delta = \frac{a_{1,2}a_{2,1}}{a_{1,1}a_{2,2}} \quad (4.11)$$

Solving this equation to find t and computing the subdominant eigenvalue of the matrix $M = d(\frac{1}{f})X(t) d(\frac{1}{g})X(t)^T$ finally gives the following conclusions:

Proposition I.14. *The Sinkhorn algorithm applied with marginals $f = \begin{pmatrix} f_1 \\ 1 - f_1 \end{pmatrix}$ and $g = \begin{pmatrix} g_1 \\ 1 - g_1 \end{pmatrix}$ to the positive matrix $A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}$ converges linearly with rate given by:*

$$\lambda_2 = \begin{cases} \frac{(t - f_1 g_1)^2}{f_1(1 - f_1)g_1(1 - g_1)} & \text{if } \Delta \neq 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{where } t = \begin{cases} \frac{1}{2} \left(f_1 + g_1 + \frac{\Delta}{1 - \Delta} - \sqrt{\delta} \right) & \text{if } \Delta < 1 \\ \frac{1}{2} \left(f_1 + g_1 + \frac{\Delta}{1 - \Delta} + \sqrt{\delta} \right) & \text{if } \Delta > 1 \end{cases}$$

$$\text{with } \delta = \left(f_1 + g_1 + \frac{\Delta}{1 - \Delta} \right)^2 - \frac{4f_1 g_1}{1 - \Delta} \text{ and } \Delta = \frac{a_{1,2}a_{2,1}}{a_{1,1}a_{2,2}}.$$

Proof. We start from equation (4.11). The case where $\Delta = 1$ is easily treated as the limit matrix of the Sinkhorn iterates is $X = f \otimes g$. Then $M = f \otimes \mathbf{1}$ has second eigenvalue 0.

Suppose now $\Delta \neq 1$. Then, the value of the discriminant of equation (4.11) is given by δ stated in the proposition. Note that one always has $\delta > 0$ (if $\Delta < 1$, just notice that $\delta = h(\frac{1}{1-\Delta})$ for a degree 2 polynomial h whose minimum occurs in \mathbb{R}^- , hence $\delta > h(1) > 0$). To decide which sign in front of $\sqrt{\delta}$ is appropriate to obtain the correct value of t , we use that $t < \frac{1}{2}(f_1 + g_1)$ for $\Delta < 1$ and $t > f_1 + g_1 - 1$ for $\Delta > 1$. \square

In the context of Optimal Transport, that is when $A = \exp(-\frac{w}{\varepsilon})$, this convergence rate is still given according to $\Delta = \exp(\frac{\bar{w}}{\varepsilon})$ for $\bar{w} = w_{1,1} + w_{2,2} - w_{1,2} - w_{2,1}$. Furthermore, the value of its limit is given by the marginals f, g in the following way:

Proposition I.15. *Denote $\lambda_2(\varepsilon)$ the convergence rate of the Sinkhorn algorithm applied to the matrix $A = \exp(-\frac{w}{\varepsilon})$ described in Theorem I.14, and $\bar{w} = w_{1,1} + w_{2,2} - w_{1,2} - w_{2,1}$. If $\bar{w} = 0$ then $\lambda_2(\varepsilon) \equiv 0$, otherwise $\lambda_2(\varepsilon)$ converges to a limit $\lambda_2^0 \in (0, 1]$ given by*

$$\lambda_2^0 = \begin{cases} \min\left(\frac{f_1(1-g_1)}{g_1(1-f_1)}, \frac{g_1(1-f_1)}{f_1(1-g_1)}\right) & \text{if } \bar{w} < 0 \\ \min\left(\frac{f_1 g_1}{(1-f_1)(1-g_1)}, \frac{(1-f_1)(1-g_1)}{f_1 g_1}\right) & \text{if } \bar{w} > 0 \end{cases}$$

Furthermore, there exists a constant $K > 0$ depending only on f, g such that for ε small enough:

$$|\lambda_2(\varepsilon) - \lambda_2^0| \leq \begin{cases} K \exp\left(\frac{-|\bar{w}|}{\varepsilon}\right) & \text{if } \bar{w} < 0 \text{ and } f_1 \neq g_1, \text{ or if } \bar{w} > 0 \text{ and } f_1 \neq 1 - g_1 \\ K \exp\left(\frac{-|\bar{w}|}{2\varepsilon}\right) & \text{otherwise} \end{cases}$$

Proof. We only treat the case $\bar{w} < 0$, the case $\bar{w} > 0$ is similar. With notations of Theorem I.14, $t \rightarrow t^0 = \frac{1}{2}(f_1 + g_1 - |f_1 - g_1|)$. If $f_1 \geq g_1$, then $t^0 = g_1$ and $\lambda_2^0 = \frac{(t^0 - f_1 g_1)^2}{f_1(1-f_1)g_1(1-g_1)} = \frac{g_1(1-f_1)}{f_1(1-g_1)} \leq 1$ while if $g_1 \geq f_1$ then $t^0 = f_1$ and $\lambda_2^0 = \frac{f_1(1-g_1)}{g_1(1-f_1)} \leq 1$. The further distinction on $f_1 \neq g_1$ comes from the appearance of $\sqrt{\delta}$ with $\delta \rightarrow (f_1 - g_1)^2$. In the case where $f_1 \neq g_1$, $\lambda_2(\varepsilon)$ writes as a \mathcal{C}^1 function h of $\exp(\frac{\bar{w}}{\varepsilon})$ and the announced rate follows taking $K > |h'(0)|$ (for instance when $f_1 > g_1$ one computes $h'(0) = -\frac{2g_1(1-f_1)}{f_1(1-g_1)(f_1-g_1)}$). In the case where $f_1 = g_1$, which is the case such that $\lambda_2^0 = 1$, then λ_2 is no longer a regular function of Δ . However writing that $\delta = (2f_1 + \frac{\Delta}{1-\Delta})^2 - \frac{4f_1^2}{1-\Delta} = \frac{\Delta}{1-\Delta}(4f_1(1-f_1) + \frac{\Delta}{1-\Delta}) \leq \frac{\Delta}{(1-\Delta)^2}$ (because $f_1(1-f_1) \leq \frac{1}{4}$) we see that

$$1 \geq \lambda_2(\varepsilon) \geq \left(\frac{f_1(1-f_1) - \frac{\sqrt{\Delta-\Delta}}{2(1-\Delta)}}{f_1(1-f_1)}\right)^2$$

The result follows as $1 - \lambda_2(\varepsilon)$ is bounded by a regular function of $\sqrt{\Delta} = \exp(\frac{\bar{w}}{2\varepsilon})$. \square

Remark that a limit value $\lambda_2^0 < 1$ indicates the situations where the convergence rate is more favorable. In dimension 2 it is when $|f_1 - g_1|$ (respectively $|f_1 - (1 - g_1)|$) is large in the case $\bar{w} < 0$ (respectively $\bar{w} > 0$). In that perspective, the doubly stochastic setting where $f_1 = 1 - f_1 = g_1 = 1 - g_1 = \frac{1}{2}$ corresponds to (one of) the worst possible marginals because $\lambda_2^0 = 1$. This last fact is also true in any dimension, provided the corresponding assignment problem has a unique solution, as noticed by Sharify and co-authors in [SGG11]. Indeed, in that case the result of the Sinkhorn algorithm $x(\varepsilon)$ converges when $\varepsilon \rightarrow 0$ to the optimal permutation matrix x^* . As a consequence the subdominant eigenvalue $\lambda_2(\varepsilon)$ of $x(\varepsilon)^T x(\varepsilon)$ converges to the one of $x^{*T} x^*$ which is again a permutation matrix hence only has eigenvalues of modulus 1. Authors of [SGG11] also show that the $O(\exp(-\frac{c}{\varepsilon}))$ convergence speed of $\lambda_2(\varepsilon)$ towards 1 showed in Proposition I.15 for dimension 2 holds in that setting.

The example of dimension 2 shows that there exist situations where the limit value of the convergence rate λ_2^0 is strictly less than 1. We do not know however how to characterize these situations in higher dimension. Furthermore, the typical behavior of the convergence rate $\lambda_2(\varepsilon)$ we obtained in the numerical experiments we realized is the one plotted in Figure 4.3 (here in dimension $N = 100$ with random marginals f, g and cost w), leading to believe that for larger values of the dimension the limit value λ_2^0 often approaches 1.

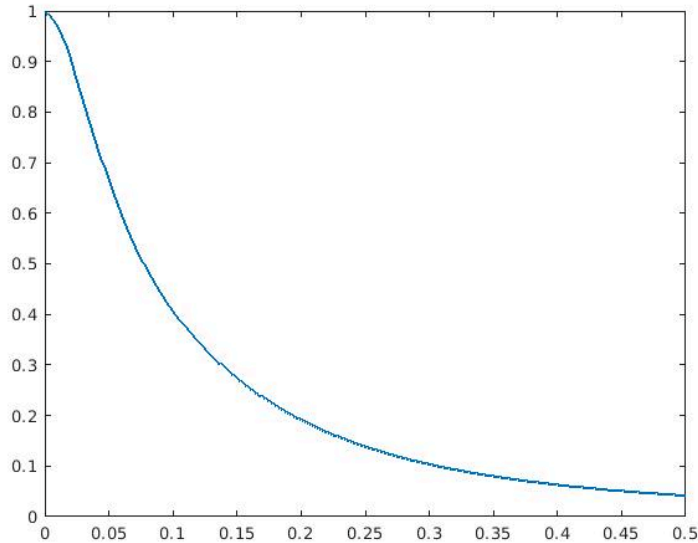


Figure 4.3 – Value of the convergence rate $\lambda_2(\varepsilon)$ of the classical Sinkhorn algorithm as a function of ε

CONCLUSION OF PART I

In the first part of this thesis, we built a new systematic approach to analyze the Sinkhorn-like algorithms we presented. This method leads to the linear convergence rate which is observed in practice, and can be summarized as follows:

1. Rescale the variables of the algorithm by dividing them by solutions of the associated scaling problem. The convergence of these new variables towards constant vectors is equivalent to the convergence of the initial variables towards a fixed point of the iterations.
2. Check that the update rules for these rescaled variables are similar to the initial iterations but also make appear some stochastic matrices.
3. Apply convexity inequalities using these stochastic matrices to obtain a linear convergence rate expressed as the subdominant eigenvalue of some stochastic symmetric primitive matrix.

This study seems to be limited to 1D balanced Sinkhorn-like algorithms meaning that each iteration must involve at most two other variables, and that the absolute values of the appearing exponents must sum to 1.

We also presented some techniques to reach smaller values of the regularizing parameter of Optimal Transportation problems, among which a proximal iterated process and the use of the Bethe entropy. In addition we gave the complete analysis of the convergence rate of the Sinkhorn algorithm in dimension 2. Future works could generalize our results to more complicated Sinkhorn-like iterates, showing how the Jacobian matrix of the iterated function relates to the linear convergence through its subdominant eigenvalue. More investigations could also be undertaken to understand precisely the behavior of these convergence rates with respect to the entries of the matrices and marginals at stake, as well as when the regularizing parameter tends to zero.

Part II

Error estimates for discretizations of the ROF model

INTRODUCTION DE LA PARTIE II

Dans la première partie, nous avons étudié des algorithmes permettant de calculer des objets de la métrique de Wasserstein comme des barycentres. Cette métrique s'avère particulièrement adéquate pour comparer deux mesures d'un point de vue "géométrique". Par exemple dans le contexte continu, pour le coût $w(x, y) = \|x - y\|_2^2$ dans \mathbb{R}^d , le barycentre de Wasserstein de deux mesures gaussiennes centrées en des points p_0 et p_1 est encore une mesure gaussienne dont le centre parcourt le segment joignant p_0 à p_1 . Ceci en fait un outil approprié pour étudier des problèmes dont l'inconnue est un ensemble E (régulier) de \mathbb{R}^d représenté par la mesure uniforme 1_E . Par exemple, étant donnés deux ensembles E^0 et E^1 de même volume $|E^0| = |E^1| = 1$ on peut s'intéresser au problème de barycentre suivant :

$$\arg \min_{E \text{ tel que } |E|=1} \theta W_w(1_{E^0}, 1_E) + (1 - \theta) W_w(1_E, 1_{E^1})$$

Malheureusement ce problème est mal posé au sens où il n'admet pas toujours de minimiseur, le barycentre de 1_{E^0} et 1_{E^1} ne s'écrivant pas forcément sous la forme 1_E . Après avoir relâché cette contrainte, et dans le but de retrouver des mesures proches d'indicatrices, on peut envisager de pénaliser le périmètre $P(E)$ des ensembles en jeu. Pour cela on utilise le fait que $P(E) = \text{TV}(1_E)$ où TV désigne la variation totale sur un domaine $\Omega \subset \mathbb{R}^d$ définie par :

$$\text{TV}(u) = \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi, \phi \in C_c^1(\Omega, \mathbb{R}^2) \text{ tel que } \|\phi\|_{\infty} \leq 1 \right\}$$

Cette quantité, finie pour toute fonction $u : \Omega \rightarrow \mathbb{R}$ dont la dérivée distributionnelle Du est une mesure de Radon (voir [AFP00]), mesure en effet les sauts de la fonction u puisque $\text{TV}(u) = \int_{\Omega} |Du|$ (identité qui prolonge au cas où u n'est pas régulière la relation $\text{TV}(u) = \int_{\Omega} |\nabla u|$). Dès lors on peut envisager d'étudier des problèmes du type :

$$\arg \min_u W(u) + \text{TV}(u)$$

où $W(u)$ est un terme faisant intervenir la distance de Wasserstein, comme par exemple $W(u) = \theta W_w(1_{E^0}, u) + (1 - \theta)W_w(u, 1_{E^1})$ ou plus simplement $W(u) = W_w(u, u^0)$ pour un certain u^0 fixé. Ce dernier problème, appelé flot de gradient Wasserstein de la variation totale [AGS08] et qu'on peut étudier avec ou sans la contrainte $u = 1_E$, est lié à la discrétisation d'équations aux dérivées partielles d'évolutions [CP19, CL19, Ott98]. Pour résoudre ce type de problèmes par des techniques classiques d'optimisation convexe (comme par exemple l'algorithme primal dual [CP11]), il est utile de pouvoir calculer l'opérateur proximal associé à la variation totale, c'est-à-dire de savoir résoudre pour tout g et tout $\lambda > 0$:

$$\arg \min_u \frac{1}{2\lambda} \|u - g\|_{L^2(\Omega)}^2 + \text{TV}(u)$$

Ce problème est en fait un problème de débruitage d'image bien connu, introduit par Rudin, Osher et Fatemi dans [ROF92] et appelé modèle de ROF. Dans la deuxième partie de cette thèse nous étudions des discrétisations de ce modèle.

Dans sa version continue, le modèle ROF repose sur le fait que $\text{TV}(u)$ est une mesure de la quantité de "bruit" d'une image donnée $u : \Omega \rightarrow \mathbb{R}$ (nous prendrons $\Omega = (0, 1)^2$). La variation totale pénalise en effet les oscillations, et sa sensibilité aux perturbations aléatoires est telle qu'une version bruitée d'une image est susceptible d'avoir une variation totale bien plus grande que sa version nette. Le modèle ROF propose donc de rechercher une version débruitée $u : \Omega \rightarrow \mathbb{R}$ d'une image donnée $g : \Omega \rightarrow \mathbb{R}$ sous la forme d'un minimiseur de la variation totale relativement proche de la donnée initiale g . Dans sa première version, le problème de minimisation correspondant est :

$$\arg \min \left\{ \text{TV}(u), u \text{ tel que } \|u - g\|_{L^2(\Omega)} \leq \sigma \right\} \quad (1)$$

où le paramètre $\sigma > 0$ indique à quelle distance L^2 maximale on s'autorise à s'éloigner de g pour trouver u . Dans [CL97], Chambolle et Lions montrent que (1) admet une reformulation équivalente où la contrainte devient second terme de l'objectif : à chaque valeur de $\sigma > 0$ correspond une valeur de $\lambda > 0$ telle que la solution de (1) soit la même que celle de :

$$\bar{u} = \arg \min_u \frac{1}{2\lambda} \|u - g\|_{L^2(\Omega)}^2 + \text{TV}(u) \quad (2)$$

Nous étudions cette version du problème, qui s'inscrit dans la famille des méthodes variationnelles à travers la présence d'un terme de fidélité aux données $\|u - g\|_{L^2(\Omega)}^2$ et d'un terme de régularisation $\text{TV}(u)$. L'équilibre entre ces deux termes est contrôlé par le paramètre $\lambda > 0$: quand $\lambda \rightarrow 0$ le terme de fidélité l'emporte et $\bar{u} \rightarrow g$, quand $\lambda \rightarrow +\infty$ le terme de régularisation est le plus important et \bar{u} converge vers une constante.

Il peut sembler surprenant d'avoir choisi d'utiliser une norme L^1 du gradient de u pour définir la variation totale alors qu'une norme L^2 semble tout aussi apte à capter les oscillations de l'image, tout en offrant un cadre théorique et numérique plus simple. Cette version L^2 (et finalement H^1) du problème porte le nom de régularisation de Tychonov et il est connu (voir [VKV16, CCC⁺10]) qu'elle possède un effet

régularisant se traduisant in fine par un flou au niveau des discontinuités de l'image débruitée. Cet écueil est résolu par l'utilisation d'une norme L^1 permettant d'obtenir des discontinuités nettes. La variation totale possède également dans sa version continue l'avantage d'être isotrope, c'est-à-dire de ne privilégier aucune direction. Ainsi peu importe l'orientation des discontinuités de g , celles-ci pourront être restituées de manière nette dans le minimiseur \bar{u} .

Cette dernière propriété est toutefois difficile à transposer au contexte discret, qui est celui qu'on utilise en pratique face à une image g se présentant sous la forme d'un tableau de pixels. Dans ce cadre, il faut en effet redéfinir le problème (2) pour l'adapter à des fonctions discrètes c'est-à-dire, dans notre cas où $\Omega = (0, 1)^2$, données par $u^h, g^h : \llbracket 1, N \rrbracket \times \llbracket 1, N \rrbracket \rightarrow \mathbb{R}$ où $h = 1/N$ désigne la largeur du maillage. Au problème continu (2) correspond donc le problème discret suivant :

$$\bar{u}^h = \arg \min_{u^h} \frac{1}{2\lambda} \|u^h - g^h\|_2^2 + \text{TV}^h(u^h) \quad (3)$$

où le terme de fidélité L^2 correspond naturellement à une norme euclidienne $\|\cdot\|_2$ et où TV^h désigne une variation totale discrète. De nombreux choix de définitions de TV^h s'offrent alors à nous. Citons notamment la variation totale dite anisotrope TV_a^h , les variations totales dites isotropes $\text{TV}_{i,\oplus,\ominus}^h$ pour $\oplus, \ominus \in \{+, -\}$ et leur version moyennée $\text{TV}_{i,*}^h$ introduite dans [WL11], la variation totale centrée TV_c^h étudiée dans [LLW09], ou encore la variation totale "upwind" TV_u^h introduite dans [OS88] dont les définitions sont les suivantes :

$$\begin{aligned} \text{TV}_a^h(u^h) &= h \sum_{i,j} |u_{i+1,j}^h - u_{i,j}^h| + |u_{i,j+1}^h - u_{i,j}^h| \\ \text{TV}_{i,\oplus,\ominus}^h(u^h) &= h \sum_{i,j} \sqrt{(u_{i\oplus 1,j}^h - u_{i,j}^h)^2 + (u_{i,j\oplus 1}^h - u_{i,j}^h)^2} \\ \text{TV}_{i,*}^h(u^h) &= \frac{1}{4} (\text{TV}_{i,++}^h(u^h) + \text{TV}_{i,+ -}^h(u^h) + \text{TV}_{i,- +}^h(u^h) + \text{TV}_{i,--}^h(u^h)) \\ \text{TV}_c^h(u^h) &= h \sum_{i,j} \sqrt{\left(\frac{u_{i+1,j}^h - u_{i-1,j}^h}{2}\right)^2 + \left(\frac{u_{i,j+1}^h - u_{i,j-1}^h}{2}\right)^2} \\ \text{TV}_u^h(u^h) &= h \sum_{i,j} \sqrt{\{u_{i,j}^h - u_{i+1,j}^h\}_+^2 + \{u_{i,j}^h - u_{i-1,j}^h\}_+^2 \\ &\quad + \{u_{i,j}^h - u_{i,j+1}^h\}_+^2 + \{u_{i,j}^h - u_{i,j-1}^h\}_+^2} \end{aligned}$$

Chaque variation totale discrète possède ses avantages et ses inconvénients. La variation totale anisotrope est relativement simple à implémenter par des méthodes efficaces de programmation linéaire mais privilégie les directions horizontales et verticales (voir [Cha05] et le chapitre 4 de [LG12]). Les variations totales isotropes induisent un flou dans la direction correspondant au quadrant donné par leur signe (nous développons ce point ci-dessous). Les variations totales moyennées, centrées et "upwind" sont plus isotropes mais aussi plus complexes à mettre en œuvre numériquement.

Afin de comparer toutes ces variations totales, une méthode systématique consiste à évaluer la vitesse de convergence du problème discret (3) vers le problème continu (2). Au sens de la gamma-convergence [Bra02], par exemple dans $L^2(\Omega)$, toutes les discrétisations TV^h suggérées, à l'exception de TV_a^h , tendent en effet vers TV quand h tend vers 0 (voir [CTV17]). Par conséquent, le minimiseur \bar{u}^h de (3) converge vers \bar{u} , minimiseur de (2), et on peut chercher à estimer l'ordre de grandeur de $\|\bar{u}^h - \bar{u}\|_{L^2(\Omega)}$. Grâce à la forte convexité des énergies en jeu, on peut estimer l'erreur commise sur les minimiseurs à partir de celle commise sur les valeurs minimales des problèmes correspondants, notées \bar{E}^h pour (3) et \bar{E} pour (2). On cherchera donc à obtenir des estimations du type $|\bar{E}^h - \bar{E}| = O(h^\theta)$ ou $|\bar{E}^h - \bar{E}| = o(h^\theta)$ pour un certain paramètre $\theta > 0$. Ainsi, Lai et ses co-auteurs montrent dans [LLW09] que sous certaines hypothèses $|\bar{E}^h - \bar{E}| \leq c\sqrt{h}$ pour le choix $\text{TV}^h = \text{TV}_c^h$. De même, Wang et Lucier prouvent dans [WL11] que $|\bar{E}^h - \bar{E}| \leq ch^{\frac{\alpha}{\alpha+1}}$ (où α est l'ordre de Lipschitz de g) pour le choix $\text{TV}^h = \text{TV}_{i,*}^h$. Dans ce même article, les auteurs obtiennent également que $\|\bar{u}^h - \bar{u}\|_{L^2}^2 \leq ch^{\frac{\alpha}{\alpha+1}}$. Les résultats que nous présentons dans cette partie II sont de ce type.

Tout d'abord nous donnons un premier taux de convergence dans le cas de la variation totale isotrope $\text{TV}_i^h = \text{TV}_{i,++}^h$. Nous nous plaçons dans le cas particulier où g présente une discontinuité orientée selon une certaine direction $\nu \in \mathbb{R}^2$ (choisie telle que $|\nu| = 1$). Plus précisément, $g = g_\nu$ est donné par :

$$g_\nu(x) = \begin{cases} 1 & \text{si } \langle x | \nu \rangle \geq a \\ 0 & \text{sinon} \end{cases}$$

où a est une valeur fixée par exemple $a = \langle (\frac{1}{2}, \frac{1}{2}) | \nu \rangle$. On constate que la variation totale dite isotrope TV_i^h se comporte moins bien dans les directions proches de $\nu = \frac{1}{\sqrt{2}}(-1, 1)$ que dans les directions proches de $\nu = \frac{1}{\sqrt{2}}(1, 1)$, engendrant un flou dans le premier cas là où les sauts sont nets dans le deuxième. Nous quantifions ce phénomène en démontrant le résultat suivant :

Théorème II.1. *Sur un domaine approprié $\Omega = \Omega_{per}$, et pour $\text{TV}^h = \text{TV}_i^h$ on a :*

1. *Pour $\nu = \frac{1}{\sqrt{2}}(1, 1)$, le débruitage est exact au sens où la solution du problème discret \bar{w}^h est l'image la plus proche de la solution du problème continu \bar{u} .*
2. *Pour $\nu = \frac{1}{\sqrt{2}}(-1, 1)$, le débruitage commet une erreur d'ordre $O(h^{2/3})$ au sens où il existe des constantes $\underline{h}, c, c' > 0$ dépendant seulement de λ telles que*

$$\forall h \leq \underline{h}, \quad ch^{2/3} \leq \bar{E}^h - \bar{E} \leq c'h^{2/3}$$

Alors que l'estimée $\bar{E}^h - \bar{E} \leq c'h^{2/3}$ est du même type que les résultats cités précédemment pour d'autres variations totales, signalons l'originalité du résultat "négatif" $\bar{E}^h - \bar{E} \geq ch^{2/3}$. Cette estimation inférieure, plus difficile à obtenir que la borne d'erreur, constitue le résultat principal de cette partie.

Dans un deuxième temps, nous étudions une variation totale discrète TV_{RT}^h basée sur des champs discrets introduits par Raviart et Thomas dans [RT77], et qui apparaît également dans ce contexte dans [DJS07, DJS12]. Cette partie est largement basée sur le travail de Chambolle et Pock dans [CP20], mais adapte les résultats obtenus sur un maillage triangulaire à notre contexte de maillage carré. Le principe de cette variation totale est de gagner en isotropie en utilisant la définition duale de la variation totale continue que nous avons déjà mentionnée :

$$TV(u) = \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi, \phi \in C_c^1(\Omega, \mathbb{R}^2) \text{ tel que } \|\phi\|_{\infty} \leq 1 \right\}$$

Dans cette définition, nous substituons à l'espace des champs continûment différentiables à support compact $C_c^1(\Omega, \mathbb{R}^2)$ l'espace des champs Raviart-Thomas d'ordre 0 s'annulant au bord $RT0_0$. Comme cet espace de champs discrets contient tous les champs constants (si on relâche la contrainte d'annulation au bord), on récupère une propriété d'isotropie pour la variation totale discrète suivante :

$$TV_{RT}^h(u^h) = \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi^h, \phi^h \in RT0_0 \text{ tel que } \|\phi^h\|_{\infty} \leq 1 \right\}$$

Nous sommes enfin en mesure de démontrer notre deuxième estimation d'erreur :

Théorème II.2. *Pour le choix $TV^h = TV_{RT}^h$, si $g \in BV(\Omega)$, et si le problème dual associé au problème continu (2) admet une solution lipschitzienne, alors le débruitage commet une erreur d'ordre $O(h)$ au sens où il existe une constante $c > 0$ ne dépendant que de λ et de la valeur de l'optimum \bar{E} telle que*

$$\forall h > 0, |\bar{E}^h - \bar{E}| \leq ch$$

Cette partie s'organise comme suit. Le chapitre 5 donne les définitions précises des variations totales continues et isotropes dans les contextes Neumann et Dirichlet et formule le constat de l'anisotropie de la variation totale isotrope. Les chapitres 6, 7 et 8 s'enchaînent pour démontrer le théorème II.1. Ils reposent sur la réduction à un problème de débruitage en dimension 1 et sur l'étude du problème dual associé. Le chapitre 9 présente la variation totale "Raviart-Thomas" étudiée puis démontre le théorème II.2 par des estimées sur les énergies primale et duale. Enfin le chapitre 10 présente des tests numériques comparant ces différentes variations totales sur des tâches simples.

Le travail présenté dans cette partie a fait l'objet d'un article, [CC20], actuellement soumis pour publication.

THE ROF MODEL WITH ISOTROPIC TOTAL VARIATION

5.1 Continuous setting of the ROF model

We begin with the presentation of the ROF model, introduced by Rudin, Osher and Fatemi in [ROF92]. In the continuous setting, we place ourselves in an open subset Ω of \mathbb{R}^2 and use the total variation TV as a regularizer. This quantity is given by $\text{TV}(u) = \int_{\Omega} |\nabla u|$ when u is regular (recall that $|\cdot|$ denotes the Euclidean norm in \mathbb{R}^2), and extends to functions of weaker regularity through the formula

$$\text{TV}(u) = \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi, \phi \in \mathcal{C}_c^1(\Omega, \mathbb{R}^2) \text{ s.t. } \|\phi\|_{\infty} \leq 1 \right\}$$

where $\mathcal{C}_c^1(\Omega, \mathbb{R}^2)$ is the space of continuously differentiable and compactly supported fields from Ω to \mathbb{R}^2 , and $\|\phi\|_{\infty} = \sup_{x \in \Omega} |\phi(x)|$. We are consequently interested in the space of functions of bounded total variation $BV(\Omega)$ given by

$$BV(\Omega) = \{u \in L^1(\Omega) \text{ s.t. } \text{TV}(u) < +\infty\}$$

This space coincides with the set of functions $u \in L^1(\Omega)$ such that the distributional derivative Du is a (vector valued) Radon measure, see [AFP00, CCC⁺10] for more details. The ROF model then writes as

$$\bar{u} = \arg \min_{u \in BV \cap L^2(\Omega)} \frac{1}{2\lambda} \|u - g\|_{L^2}^2 + \text{TV}(u) =: E(u) \quad (5.1)$$

where the regularizing parameter $\lambda > 0$ is fixed and where g is our noisy image, for which we will suppose $g \in L^\infty(\Omega)$ as well as $g \in BV(\Omega)$ when needed.

To be more precise, we consider both Neumann and Dirichlet boundary conditions to this setting. This will result in two different variants of (5.1): in the Neumann setting we study:

$$\begin{aligned} \bar{u}_N &= \arg \min_{u \in BV \cap L^2(\Omega)} \frac{1}{2\lambda} \|u - g\|_{L^2}^2 + \text{TV}_N(u) =: E_N(u) \text{ where} & (5.2) \\ \text{TV}_N(u) &= \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi, \phi \in \mathcal{C}_c^1(\Omega, \mathbb{R}^2) \text{ s.t. } \|\phi\|_{\infty} \leq 1 \right\} \end{aligned}$$

while in the Dirichlet setting, we add the constraint that $u = b$ on $\partial\Omega$ for some $b \in BV \cap L^\infty(\partial\Omega)$ (naturally, one usually takes $b = g|_{\partial\Omega}$), and replace TV_N by

$$\sup \left\{ - \int_{\Omega} u \operatorname{div} \phi + \int_{\partial\Omega} u \langle \phi | \vec{n} \rangle, \phi \in \mathcal{C}^1(\Omega, \mathbb{R}^2) \text{ s.t. } \|\phi\|_{\infty} \leq 1 \right\}$$

where \vec{n} denotes the outer normal unit vector. Equivalently, we formulate the Dirichlet problem as:

$$\begin{aligned} \bar{u}_D &= \arg \min_{u \in BV \cap L^2(\Omega)} \frac{1}{2\lambda} \|u - g\|_{L^2}^2 + \text{TV}_D(u) =: E_D(u) \text{ where} & (5.3) \\ \text{TV}_D(u) &= \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi + \int_{\partial\Omega} b \langle \phi | \vec{n} \rangle, \phi \in \mathcal{C}^1(\Omega, \mathbb{R}^2) \text{ s.t. } \|\phi\|_{\infty} \leq 1 \right\} \end{aligned}$$

In the following, we will denote for $B \in \{N, D\}$ the optimal value of the continuous problems $\bar{E}_B = E_B(\bar{u}_B)$. When no subscript (N or D) is used, it means our statement is valid under both boundary conditions. From now on, we also focus on the case where $\Omega = (0, 1) \times (0, 1)$.

The continuous ROF model enjoys the property that its solution \bar{u} behaves equally according to the ‘‘orientation’’ of the source term g . This isotropy result can be stated rigorously in the Dirichlet setting the following way. Given a direction $\nu \in \mathbb{R}^2$ with $|\nu| = 1$, take $g = g_\nu$ defined by $g_\nu(x) = 1$ if $\langle x | \nu \rangle \geq a$ and $g_\nu(x) = 0$ otherwise where a is some fixed real number (for instance $a = \langle (1/2, 1/2) | \nu \rangle$). Then, problem (5.3) with boundary condition $b = g_\nu|_{\partial\Omega}$ has solution $\bar{u}_D = g_\nu$, no matter the orientation of ν . This comes from the following important fact¹:

Fact II.1. *Fix $\nu \in \mathbb{R}^2$ with $|\nu| = 1$. When using the boundary condition $b = g_\nu|_{\partial\Omega}$, the value of $\text{TV}_D(g_\nu)$ is reached for $\phi \equiv \nu$ so that $\text{TV}_D(g_\nu) = \int_{\partial\Omega} g_\nu \langle \nu | \vec{n} \rangle$.*

Our claim is indeed a direct consequence: if $u \in BV \cap L^2(\Omega)$ is such that we have $u|_{\partial\Omega} = g_\nu|_{\partial\Omega}$, then taking the admissible field $\phi \equiv \nu$ gives

$$\text{TV}_D(u) \geq \int_{\partial\Omega} g_\nu \langle \nu | \vec{n} \rangle = \text{TV}_D(g_\nu)$$

and the result follows.

¹that just derives from the fact that $-\int_{\Omega} g_\nu \operatorname{div} \phi + \int_{\partial\Omega} g_\nu \langle \phi | \vec{n} \rangle = \int_{\{x \text{ s.t. } \langle x | \nu \rangle = a\}} \langle \nu | \phi \rangle$

5.2 Isotropic total variation

In practice, Ω is discretized into $N \times N$ square pixels of size $h = 1/N$, namely $\Omega = \cup_{1 \leq i, j \leq N} C_{i,j}$ with $C_{i,j} = [(i-1)h, ih] \times [(j-1)h, jh]$. Images are now elements of $P0 = \{u : \Omega \rightarrow \mathbb{R} \text{ s.t. } \forall i, j \in \llbracket 1, N \rrbracket, \exists u_{i,j} \in \mathbb{R} \text{ s.t. } u = u_{i,j} \text{ in } C_{i,j}\}$. One introduces the projection of the continuous image $g^h = \Pi_{P0}(g)$ given by $(g^h)_C = \frac{1}{h^2} \int_C g$ for every square pixel C , and the discrete counterpart of (5.1) is the following:

$$\bar{u}^h = \arg \min_{u^h \in P0} \frac{1}{2\lambda} \|u^h - g^h\|_{L^2}^2 + \text{TV}^h(u^h) =: E^h(u^h) \quad (5.4)$$

where TV^h is some discretization of the total variation defined on $P0$. In the Dirichlet setting, TV^h can involve the discretization b^h of b given by $(b^h)_e = \frac{1}{h} \int_e b$ for every boundary edge e .

A widely used choice for TV^h is the so called ‘‘isotropic’’ total variation which discretizes the expression $\text{TV}(u) = \int_{\Omega} |\nabla u|$ using a finite difference operator D to approximate the continuous ∇ operator. It is given by

$$\text{TV}_i^h(u^h) = h \sum_{1 \leq i, j \leq N} |(Du^h)_{i,j}| \quad \text{where } (Du^h)_{i,j} = \begin{pmatrix} u_{i+1,j}^h - u_{i,j}^h \\ u_{i,j+1}^h - u_{i,j}^h \end{pmatrix} \quad (5.5)$$

(with either $u_{N+1,j}^h = b_{N+\frac{1}{2},j}^h$, $u_{i,N+1}^h = b_{i,N+\frac{1}{2}}^h$ in the Dirichlet boundary conditions or $u_{N+1,j}^h - u_{N,j}^h = u_{i,N+1}^h - u_{i,N}^h = 0$ in the Neumann boundary conditions). The term ‘‘isotropic’’ refers to the behavior of this functional as the mesh size h tends to zero. One can indeed show that TV_i^h gamma converges in $L^2(\Omega)$ to TV when $h \rightarrow 0$. More precisely, if we define on $L^2(\Omega)$ the functionals

$$F_h(u) = \begin{cases} \text{TV}_i^h(u^h) & \text{if } u = u^h \in P0 \\ +\infty & \text{otherwise} \end{cases} \quad \text{and } F(u) = \begin{cases} \text{TV}(u) & \text{if } u \in BV \\ +\infty & \text{otherwise} \end{cases}$$

then we have the following proposition (see [Bra02] for more background on gamma convergence):

Proposition II.1. *When $h \rightarrow 0$, F_h gamma converges to F in $L^2(\Omega)$, that is:*

1. $\forall u \in L^2(\Omega), \forall u^h \rightarrow u, F(u) \leq \liminf F_h(u^h)$
2. $\forall u \in L^2(\Omega), \exists u^h \rightarrow u, F(u) \geq \limsup F_h(u^h)$

Proof. We emphasize below the main arguments of the proof following [CTV17] where this is proved for a more complicated total variation. We begin with the second point that simply derives from approximation. The result is true if $u \in C^\infty(\Omega)$ by just taking $u^h = \Pi_{P0}$; to extend it to $u \in L^2(\Omega)$ we use the following lemma (obtained through convolution, see Theorem 3.9. in [AFP00]):

Lemma II.1. *For any $u \in BV(\Omega)$, there exists a sequence $(u^n) \in (C^\infty(\Omega))^{\mathbb{N}}$ such that $u^n \rightarrow u$ in $L^2(\Omega)$ and $\text{TV}(u^n) \rightarrow \text{TV}(u)$.*

For the first point, if $u^h \rightarrow u$, then we can in fact suppose that $F_h(u^h) \rightarrow \ell < +\infty$ and $u^h \in P_0$, and we need to prove that $\text{TV}(u) \leq \ell$.

For $\phi = (\phi^1, \phi^2) \in C_c^\infty(\Omega, \mathbb{R}^2)$ with $|\phi| \leq 1$ on Ω , we integrate by parts:

$$\begin{aligned} \int_{\Omega} u^h \operatorname{div} \phi &= \sum_{i,j} \int_{C_{i,j}} u_{i,j}^h \operatorname{div} \phi = \sum_{i,j} \int_{\partial C_{i,j}} u_{i,j}^h \langle \phi | \vec{n} \rangle \\ &= \sum_{i,j} u_{i,j}^h \left(\phi_{i+\frac{1}{2},j}^1 - \phi_{i-\frac{1}{2},j}^1 + \phi_{i,j+\frac{1}{2}}^2 - \phi_{i,j-\frac{1}{2}}^2 \right) \\ &= \sum_{i,j} \phi_{i+\frac{1}{2},j}^1 (u_{i,j}^h - u_{i+1,j}^h) + \phi_{i,j+\frac{1}{2}}^2 (u_{i,j}^h - u_{i,j+1}^h) \\ &\leq \sum_{i,j} \sqrt{\left(\phi_{i+\frac{1}{2},j}^1 \right)^2 + \left(\phi_{i,j+\frac{1}{2}}^2 \right)^2} \sqrt{(u_{i,j}^h - u_{i+1,j}^h)^2 + (u_{i,j}^h - u_{i,j+1}^h)^2} \end{aligned}$$

where for $k = 1, 2$, $\phi_e^k = \int_e \phi^k$ is the flux of ϕ^k along the edge e oriented from bottom to top (respectively from right to left) for vertical (respectively horizontal) edges. As ϕ is smooth, if $p_{i,j}$ denotes the center of $C_{i,j}$ we have

$$\phi_{i+\frac{1}{2},j}^1 = h\phi^1(p_{i,j}) + O(h^2) \quad \text{and} \quad \phi_{i,j+\frac{1}{2}}^2 = h\phi^2(p_{i,j}) + O(h^2)$$

with $O(h^2)$ being uniform in i, j . Then

$$\sqrt{\left(\phi_{i+\frac{1}{2},j}^1 \right)^2 + \left(\phi_{i,j+\frac{1}{2}}^2 \right)^2} = h|\phi(p_{i,j})| + O(h^2) \leq h + O(h^2)$$

and finally,

$$\int_{\Omega} u^h \operatorname{div} \phi \leq \text{TV}_i^h(u^h) (1 + O(h)) = F_h(u^h) (1 + O(h)) \rightarrow \ell$$

But as $u^h \rightarrow u$ in L^1 we also have $\int_{\Omega} u^h \operatorname{div} \phi \rightarrow \int_{\Omega} u \operatorname{div} \phi$ so $\int_{\Omega} u \operatorname{div} \phi \leq \ell$. As this is true for any ϕ we obtain $F(u) = \text{TV}(u) \leq \ell$, which concludes the proof. \square

A classical consequence of Proposition II.1 is that the minimizers \bar{u}^h converge in L^2 to \bar{u} , the minimizer of (5.1). This leads to saying that TV_i^h inherits of the isotropy of TV for denoising problems such as ROF.

However, this convergence result does not guarantee the isotropy of the discrete isotropic total variation itself. In fact $\text{TV}_i^h(g_\nu^h)$ can be quite far from $\text{TV}(g_\nu)$ which equals the length of the line drawn by g_ν . What is worse is that the value of $\text{TV}_i^h(g_\nu^h)$ actually depends on the orientation of ν . The case of the 45° diagonal is eloquent: as noted for instance in [CP20], the choice of the finite difference operator D induces a difference of roughly 40% between the main diagonal, that is $\nu = \frac{1}{\sqrt{2}}(1, 1)$, and its flipped version $\nu = \frac{1}{\sqrt{2}}(-1, 1)$. In Figure 5.1, we represented the vector $(Du^h)_{i,j}$ at each pixel (i, j) where it does not vanish. When $\nu = \frac{1}{\sqrt{2}}(1, 1)$ (on the left image), there are approximately N such pixels (that is, when not addressing the issue of border

pixels), and each one contributes $\sqrt{2}$ to TV_i^h . When $\nu = \frac{1}{\sqrt{2}}(-1, 1)$ (on the right image), these pixels are about $2N$ and each one contributes 1 to TV_i^h . Finally, flipping the diagonal changes the value of TV_i^h from $\sqrt{2}N$ to $2N$.

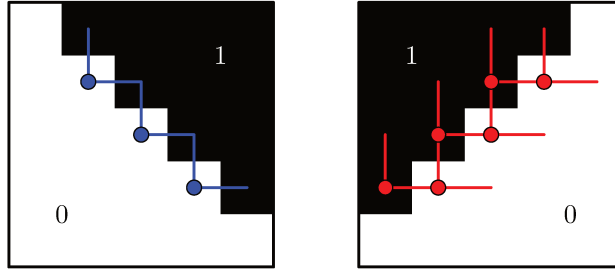


Figure 5.1 – On the left image $\text{TV}_i^h \simeq \sqrt{2}N$ while on the right $\text{TV}_i^h \simeq 2N$

This differentiation breaks the isotropy of TV_i^h for a fixed $h > 0$ leading to artefacts depending on the direction in denoising problems such as the denoising of a circle: the edges oriented along the more penalized diagonal are blurred (see Figure 5.2). Going back to the case $g = g_\nu$, even if one always has $\bar{u}_h \rightarrow g_\nu$, the speed of this convergence may vary with ν . We take again the example of the two mirror diagonals for which images “denoised” with TV_i^h are shown in Figure 5.2 for different step sizes h . One notices that the denoising is achieved correctly for the 135° diagonal \triangleleft (which we will now call consequently the “good” diagonal) whereas one needs to take h very small before obtaining a sharp looking discontinuity with the other diagonal \triangleleft (the “bad” one).

The purpose of the following chapters is to study the error made by the isotropic total variation in the “bad” diagonal denoising problem. To this end, we estimate the convergence of the optimal discrete energy of problem (5.4) towards the optimal continuous energy in (5.1). Up to a slight change of the domain ($\Omega = \Omega_{per}$ that we will define soon), we show that it is of order $O(h^{2/3})$. More precisely we will prove the following theorem:

Theorem II.1. *On an appropriate domain $\Omega = \Omega_{per}$ we have:*

1. For $\nu = \frac{1}{\sqrt{2}}(1, 1)$ the denoising is exact, meaning that $\bar{u}^h = \Pi_{P_0}(\bar{u})$.
2. For $\nu = \frac{1}{\sqrt{2}}(-1, 1)$, there exist $\underline{h}, c, c' > 0$ depending only on λ such that

$$\forall h \leq \underline{h}, ch^{2/3} \leq \bar{E}^h - \bar{E} \leq c'h^{2/3}$$

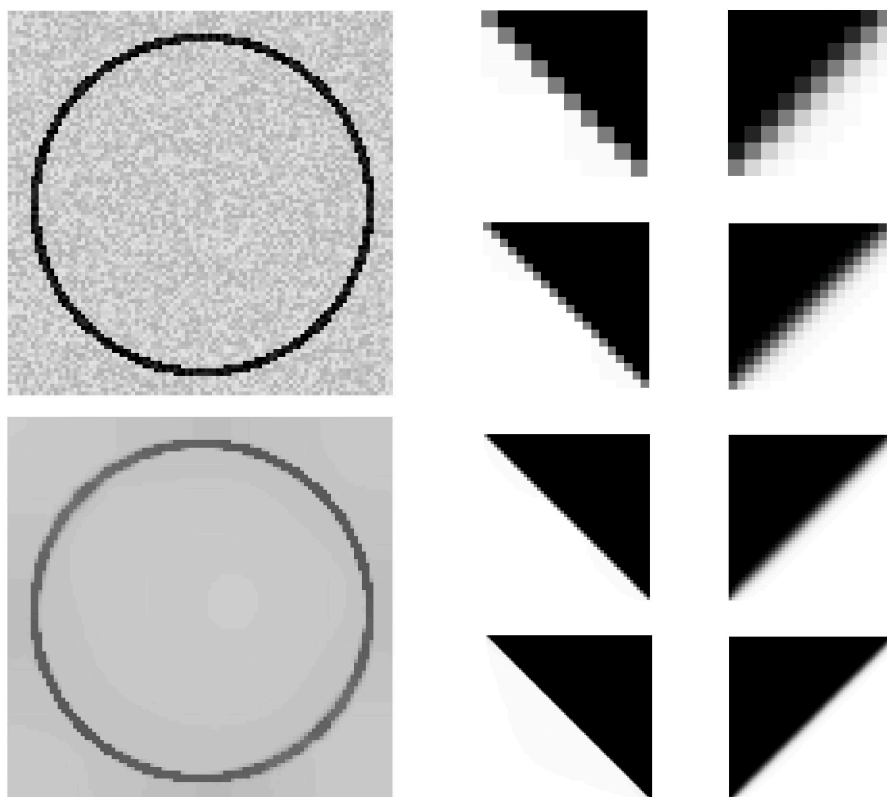


Figure 5.2 – Denoising with TV_i^h : noisy and denoised circle with Neumann boundary conditions; “good” (2nd column) and “bad” (3rd column) diagonals with Dirichlet boundary conditions and $N = 10, 20, 50, 100$.

CHAPTER 6

REDUCTION TO A 1D TOTAL VARIATION DENOISING PROBLEM

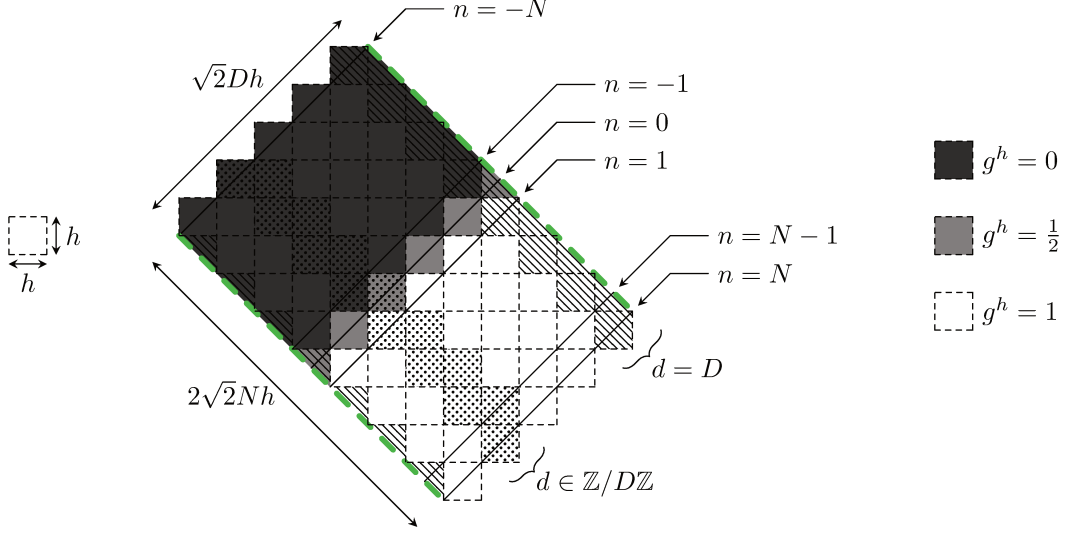
To study the orientation dependent error of the isotropic total variation, we introduce the following experiment. Placing ourselves in a well-chosen periodic domain $\Omega = \Omega_{per}$, we reduce the 2D TV_i^h denoising problem in the case of a diagonal image $g = g_\nu$ with $\nu = \frac{1}{\sqrt{2}}(-1, 1)$ to a 1D problem. In the following, we will denote respectively TV and tv the 2D and 1D total variations. The first point of Theorem II.1, which is the case $\nu = \frac{1}{\sqrt{2}}(1, 1)$, will be quickly obtained. We next present some general results about the case $\nu = \frac{1}{\sqrt{2}}(-1, 1)$ that will be useful to prove the second point of Theorem II.1 in the following chapters.

6.1 Setting on periodic domain

We actually do not consider the ROF model (5.4) on a square domain, but on a periodic strip oriented along the diagonal at stake, see the figure below in which each square pixel is of size $h = 1/N$ and where the green dotted lines are to be identified. For $\nu = \frac{1}{\sqrt{2}}(-1, 1)$, we now work with a variable $u_{i,j}^h$ defined for $(i, j) \in \mathbb{Z}^2$ such that $-N \leq i - j \leq N$; $0 \leq i + j \leq D$ and satisfying $u_{i+D, j+D}^h = u_{i,j}^h$ for any (i, j) . Making the change of variables $n = i - j$; $d = \lfloor \frac{i+j}{2} \rfloor$, one checks that our domain can be represented by

$$\Omega_{per} = \{(n, d), -N \leq n \leq N, d \in \mathbb{Z}/D\mathbb{Z}\}$$

Our source term $g^h : \Omega_{per} \rightarrow \mathbb{R}$ is given by $g^h(n, d) = 1$ for $n > 0$, $g^h(n, d) = 0$ for $n < 0$ and $g^h(0, d) = 1/2$, for all $d \in \mathbb{Z}/D\mathbb{Z}$.



Then the problem (5.4) is to solve

$$\bar{u}^h = \arg \min_{u^h: \Omega_{per} \rightarrow \mathbb{R}} \frac{h^2}{2\lambda} \sum_{(n,d) \in \Omega_{per}} |u^h(n,d) - g^h(n,d)|^2 + \text{TV}_i^h(u) := E^h(u^h)$$

where TV_i^h stands for the isotropic total variation on this particular domain. Suppose $u^h(n,d)$ codes for the value of $u_{i,j}^h$, that is that $n = i - j$ and $d = \lfloor \frac{i+j}{2} \rfloor$, then one finds that $u_{i+1,j}^h$ (respectively $u_{i,j+1}^h$) is represented by $u^h(n+1, d')$ (respectively $u^h(n-1, d')$) with $d' = d$ for n even and $d' = d+1$ for n odd. Following the definition (5.5), this leads to the following expression of the isotropic total variation:

$$\begin{aligned} \text{TV}_i^h(u^h) &= h \sum_{\substack{d \in \mathbb{Z}/D\mathbb{Z} \\ n \text{ even}}} \left| \begin{pmatrix} u^h(n+1, d) - u^h(n, d) \\ u^h(n-1, d) - u^h(n, d) \end{pmatrix} \right| \\ &\quad + h \sum_{\substack{d \in \mathbb{Z}/D\mathbb{Z} \\ n \text{ odd}}} \left| \begin{pmatrix} u^h(n+1, d+1) - u^h(n, d) \\ u^h(n-1, d+1) - u^h(n, d) \end{pmatrix} \right| \end{aligned}$$

We will first study the case of Dirichlet boundary conditions meaning that we impose (both on the definition of TV_i^h and on the optimization problem) that for all $d \in \mathbb{Z}/D\mathbb{Z}$:

$$\begin{cases} u^h(N+1, d) = u^h(N, d) = g^h(N, d) = 1 \\ u^h(-N-1, d) = u^h(-N, d) = g^h(-N, d) = 0 \end{cases}$$

Later on we will deduce from the Dirichlet setting the same rate for the Neumann boundary conditions:

$$\begin{cases} u^h(N+1, d) = u^h(N, d) \\ u^h(-N-1, d) = u^h(-N, d) \end{cases}$$

The benefit of this periodic setting is to reduce the problem from 2D to 1D as at the optimum one has $\bar{u}^h(n, d) = \bar{u}^h(n, d')$ for all n and $d, d' \in \mathbb{Z}/D\mathbb{Z}$. Indeed, as all the terms in the objective are invariant when changing d to $d+1$, the shifted image $\tilde{u}^h : (n, d) \mapsto \bar{u}^h(n, d+1)$ has the same energy E^h , hence $\tilde{u}^h = \bar{u}^h$ by uniqueness of the optimizer.

We keep the letter u for this now 1D variable, and divide our energy by a factor $\sqrt{2}Dh$ which is the width of our 2D domain. The problem then rewrites as:

$$\bar{u}^h = \arg \min_{\substack{u^h \in \mathbb{R}^{2N+1} \\ \text{s.t. BC}}} E^h(u^h) := \frac{h}{2\sqrt{2}\lambda} \|u^h - g^h\|_2^2 + \text{tv}_i^h(u^h) \quad (6.1)$$

where we defined

$$\begin{cases} \|u^h - g^h\|_2^2 = \sum_{n=-N}^N (u_n^h - g_n^h)^2 \\ \text{tv}_i^h(u^h) = \frac{1}{\sqrt{2}} \sum_{n=-N}^N \sqrt{(u_{n+1}^h - u_n^h)^2 + (u_n^h - u_{n-1}^h)^2} \end{cases}$$

with $g_n^h = 0$ for $n < 0$, $g_n^h = 1$ for $n > 0$ and $g_0^h = 1/2$ and where BC stands for the following boundary conditions:

$$\begin{cases} u_{N+1}^h = u_N^h = 1 \text{ and } u_{-N-1}^h = u_{-N}^h = 0 \text{ for Dirichlet} \\ u_{N+1}^h = u_N^h \text{ and } u_{-N-1}^h = u_{-N}^h \text{ for Neumann} \end{cases}$$

This problem is therefore a 1D signal denoising relying on a biased 1D total variation $\frac{1}{\sqrt{2}} \sum_n \sqrt{(u_{n+1}^h - u_n^h)^2 + (u_n^h - u_{n-1}^h)^2}$. This bias is responsible for the bad behavior of TV_i^h on this diagonal.

As a comparison, when dealing with the ‘‘good’’ diagonal one introduces the symmetric domain of Ω_{per} , similar in all aspects but oriented along the direction $\frac{1}{\sqrt{2}}(1, 1)$. Then, doing the same analysis, one checks that this leads to a 1D denoising with the classical 1D discrete total variation $\text{tv}^h(u^h) = \sum_n |u_{n+1}^h - u_n^h|$. As a consequence, the denoising is exact: $\bar{u}^h = g^h$. Indeed, the problem (in the Dirichlet setting) is to minimize $\|u^h - g^h\|_2^2 + c \text{tv}^h(u^h)$ for some constant $c > 0$ and under the constraint that $u_{N+1}^h = u_N^h = 1$ and $u_{-N}^h = 0$. This constraint gives $\text{tv}^h(u^h) \geq \left| \sum_{n=-N}^N u_{n+1}^h - u_n^h \right| = 1 = \text{tv}^h(g^h)$, hence we obtain the first point of Theorem II.1.

6.2 Continuous solution

In this section we investigate the continuous 1D denoising problem obtained when passing to the limit $h \rightarrow 0$ in problem (6.1). Assuming u^h is the discretization of a smooth function u defined on $[-1, 1]$, we write:

$$E^h(u^h) = \frac{1}{N\sqrt{2}} \sum_{n=-N}^N \frac{1}{2\lambda} (u(nh) - g_n^h)^2 + \sqrt{\left(\frac{u(nh+h) - u(nh)}{h}\right)^2 + \left(\frac{u(nh) - u(nh-h)}{h}\right)^2}$$

and we see that this converges as $h \rightarrow 0$ to

$$E(u) = \int_{-1}^1 \frac{1}{2\sqrt{2}\lambda} (u - g)^2 + |u'| \quad (6.2)$$

with $\int_{-1}^1 |u'| =: \text{tv}(u)$ being the continuous 1D total variation. It is easily shown that (6.2) is also the gamma-limit of the discrete problem, so that the minimizers \bar{u}^h of (6.1) will converge to the minimizer of (6.2).

For the Dirichlet setting, we enforce the constraint $u = g$ at the boundary of the domain i.e. $u(-1) = 0$ and $u(1) = 1$. In that situation, for any admissible u we have:

$$\int_{-1}^1 |u'| \geq \left| \int_{-1}^1 u' \right| = |u(1) - u(-1)| = 1 = \int_{-1}^1 |g'|$$

which directly shows that the energy (6.2) is minimal for $u = g$ with value $\bar{E}_D = 1$.

In the Neumann setting however, no boundary condition is required. To find the solution, one can write the optimality conditions given by duality theory (see [CCC⁺10]):

$$\text{tv}(u) = - \int_{-1}^1 uz' \quad \text{and} \quad \frac{1}{\sqrt{2}\lambda} (u - g) - z' = 0$$

for some function z such that $|z| \leq 1$ and $z(-1) = z(1) = 0$. If these equations are met for some couple (u, z) then u is optimal in problem (6.2). We search for u of the form $u = u_a$ for some $a \in \mathbb{R}$ with $u_a(x) = a$ if $x \in (-1, 0)$ and $u_a(x) = 1 - a$ if $x \in (0, 1)$. This leads to taking $z(x) = \frac{a}{\sqrt{2}\lambda}(x + 1)$ if $x \in (-1, 0)$ and $z(x) = \frac{a}{\sqrt{2}\lambda}(1 - x)$ if $x \in (0, 1)$. Then one must try to fulfill the equations $\text{tv}(u_a) = - \int_{-1}^1 u_a z'$ that is $|1 - 2a| = \frac{1}{\sqrt{2}\lambda} a(1 - 2a)$ and $|z| \leq 1$ that is $|a| \leq \sqrt{2}\lambda$. These two equations on a always give rise to a unique solution: if $\lambda \leq \lambda^* := \frac{\sqrt{2}}{4}$ then u_a is optimal with $a = a_{opt} := \sqrt{2}\lambda$ and the minimal energy is $\bar{E}_N^{\leq} := 1 - \sqrt{2}\lambda$. If $\lambda > \lambda^*$ then u_a is optimal with $a = \frac{1}{2}$ and the minimal energy is $\bar{E}_N^{>} := \frac{1}{4\sqrt{2}\lambda}$. In the following, we will see that in the case $\lambda > \lambda^*$ the discrete problem is exact ($\bar{u}^h \equiv \frac{1}{2}$), therefore we will always place ourselves in the case $\lambda \leq \lambda^*$, and we denote $\bar{E}_N := \bar{E}_N^{\leq} = 1 - \sqrt{2}\lambda$.

6.3 Form of discrete solution

Before turning to the proof of the $O(h^{2/3})$ bounds, we make some general remarks on the form of the solution of (6.1):

Proposition II.2. *The solution \bar{u}^h of problem (6.1) (either with Dirichlet or Neumann boundary conditions) satisfies:*

1. $\forall n \in \llbracket -N, N \rrbracket$, $\bar{u}_{-n}^h = 1 - \bar{u}_n^h$, in particular $\bar{u}_0^h = \frac{1}{2}$.
2. $\forall n \in \llbracket 1, N \rrbracket$, $1 \geq \bar{u}_n^h \geq \frac{1}{2}$, hence $\forall n \in \llbracket -N, -1 \rrbracket$, $0 \leq \bar{u}_n^h \leq \frac{1}{2}$.
3. \bar{u}^h is nondecreasing: $\forall n \in \llbracket -N, N-1 \rrbracket$, $\bar{u}_{n+1}^h \geq \bar{u}_n^h$.

Proof. For the first point, the symmetry of g^h and tv_i^h yields that $\tilde{u}_n^h = 1 - \bar{u}_{-n}^h$ satisfies $E^h(\tilde{u}^h) = E^h(\bar{u}^h)$. By uniqueness of the minimizer, $\tilde{u}^h = \bar{u}^h$.

For the second point, the truncated variable $\hat{u}_n^h = \max(g_n^h, \min(\bar{u}_n^h, \frac{1}{2}))$ satisfies $|\hat{u}_n^h - g_n^h| \leq |\bar{u}_n^h - g_n^h|$ and $|\hat{u}_{n+1}^h - \hat{u}_n^h| \leq |\bar{u}_{n+1}^h - \bar{u}_n^h|$ for any n , hence $E^h(\hat{u}^h) \leq E^h(\bar{u}^h)$ and $\bar{u}^h = \hat{u}^h$.

For the third point, finally consider the staircase version of \bar{u}^h given by: $\check{u}_n^h = \max\{\bar{u}_k^h, 0 \leq k \leq n\}$ if $n > 0$, $\check{u}_0^h = \frac{1}{2}$ and $\check{u}_n^h = \min\{\bar{u}_k^h, n \leq k \leq 0\}$ if $n < 0$. As $\bar{u}_n^h \in [0, 1]$ we have $|\check{u}_n^h - g_n^h| \leq |\bar{u}_n^h - g_n^h|$, and again $|\check{u}_{n+1}^h - \check{u}_n^h| \leq |\bar{u}_{n+1}^h - \bar{u}_n^h|$ for any n , hence $E^h(\check{u}^h) \leq E^h(\bar{u}^h)$ and $\bar{u}^h = \check{u}^h$. \square

Proposition II.3. *We denote $\lambda^* = \frac{\sqrt{2}}{4}$. The solution \bar{u}^h of problem (6.1) is such that:*

1. *With Dirichlet boundary conditions, $\bar{u}_1^h > \frac{1}{2}$ for any λ .*
2. *With Neumann boundary conditions, $\bar{u}^h \equiv \frac{1}{2}$ for any $\lambda \geq \lambda^*$ and $\bar{u}_1^h > \frac{1}{2}$ for any $\lambda < \lambda_h^*$ where λ_h^* is such that $|\lambda_h^* - \lambda^*| \leq ch^{1/3}$ for some constant $c > 0$. In particular, for any $\lambda < \lambda^*$ one has $\bar{u}_1^h > \frac{1}{2}$ for h small enough.*

Proof. For $u \in \mathbb{R}^{2N+1}$ satisfying the three properties of Proposition II.2 and such that $u_1 = \frac{1}{2}$, we define $k \in \llbracket 1, N \rrbracket$ such that $u_{-1} = u_0 = \dots = u_k = \frac{1}{2}$ and $u_{k+1} > \frac{1}{2}$. Suppose first that $k \leq N-2$ then the energy of u can be written

$$E^h(u) = \frac{h}{2\sqrt{2}\lambda}(u_k - 1)^2 + \frac{1}{\sqrt{2}}|u_k - \frac{1}{2}| + \frac{1}{\sqrt{2}}\sqrt{(u_{k+1} - u_k)^2 + (u_k - \frac{1}{2})^2} \\ + \frac{1}{\sqrt{2}}\sqrt{(u_{k+2} - u_{k+1})^2 + (u_{k+1} - u_k)^2} + R(u)$$

where $R(u)$ does not depend on u_k . As $u_{k+1} > \frac{1}{2}$, we have the following derivatives

or subgradients where we denote $\text{tv}_k = \sqrt{(u_{k+1} - u_k)^2 + (u_k - \frac{1}{2})^2}$ and $\text{tv}_{k+1} = \sqrt{(u_{k+2} - u_{k+1})^2 + (u_{k+1} - u_k)^2}$:

$$\begin{aligned} \frac{\partial}{\partial u_k} (\text{tv}_k)_{|u_k=\frac{1}{2}} &= \left(\frac{(u_k - u_{k+1}) + (u_k - \frac{1}{2})}{\sqrt{(u_{k+1} - u_k)^2 + (u_k - \frac{1}{2})^2}} \right)_{|u_k=\frac{1}{2}} = -1 \\ \frac{\partial}{\partial u_k} (\text{tv}_{k+1})_{|u_k=\frac{1}{2}} &= \frac{\frac{1}{2} - u_{k+1}}{\sqrt{(\frac{1}{2} - u_{k+1})^2 + (u_{k+2} - u_{k+1})^2}} = -d < 0 \\ \frac{\partial}{\partial u_k} ((u_k - 1)^2)_{|u_k=\frac{1}{2}} &= -1 \text{ and } \frac{\partial}{\partial u_k} (|u_k - \frac{1}{2}|)_{|u_k=\frac{1}{2}} = [-1, 1] \end{aligned}$$

Finally $\frac{\partial E^h}{\partial u_k} |_{u_k=\frac{1}{2}} = -\frac{h}{2\sqrt{2}\lambda} + \frac{1}{\sqrt{2}}[-1, 1] - \frac{1}{\sqrt{2}}(1+d) \subset \mathbb{R}_*^-$ so that $0 \notin \frac{\partial E^h}{\partial u_k} |_{u_k=\frac{1}{2}}$ hence u is not optimal. For $k = N - 1$ the same reasoning is correct in the Dirichlet setting noting that $u_{k+2} = 1$ whereas in the Neumann setting it is changed to

$$\begin{aligned} E^h(u) &= \frac{h}{2\sqrt{2}\lambda} (u_k - 1)^2 + \frac{1}{\sqrt{2}} |u_k - \frac{1}{2}| \\ &\quad + \frac{1}{\sqrt{2}} \sqrt{(u_{k+1} - u_k)^2 + (u_k - \frac{1}{2})^2} + \frac{1}{\sqrt{2}} |u_{k+1} - u_k| + R(u) \end{aligned}$$

for which one computes $\frac{\partial E^h}{\partial u_k} |_{u_k=\frac{1}{2}} = -\frac{h}{2\sqrt{2}\lambda} + \frac{1}{\sqrt{2}}[-1, 1] - \frac{2}{\sqrt{2}} \subset \mathbb{R}_*^-$ and gets the same conclusion. This concludes the proof in the Dirichlet setting as in this case $k < N$.

In the Neumann setting, the case $k = N$ corresponds to our alternative $\bar{u}^h \equiv \frac{1}{2}$ so that we only have to exhibit an admissible u^h such that $E^h(u^h) < E^h(\frac{1}{2})$ to prove that $\bar{u}_1^h > \frac{1}{2}$. We postpone this construction to section 7.4 where, provided that $\lambda < \lambda^*$, we will explicitly build a u^h such that $E^h(u^h) \leq 1 - \lambda\sqrt{2} + ch^{2/3}$ for some constant $c > 0$. In comparison the energy of the constant $u^h \equiv \frac{1}{2}$ is $E^h(\frac{1}{2}) = \frac{h}{2\sqrt{2}\lambda} \times 2N \times (\frac{1}{2})^2 = \frac{\sqrt{2}}{8\lambda}$. The conclusion comes from studying when $1 - \lambda\sqrt{2} + ch^{2/3} < \frac{\sqrt{2}}{8\lambda}$.

Finally, suppose now that $\lambda \geq \lambda^*$, we want to prove that $\bar{u}^h \equiv \frac{1}{2}$. For any $u \in \mathbb{R}^{2N+1}$ satisfying the three properties of Proposition II.2, denoting $a = u_{-N} \in [0, \frac{1}{2}]$ we form the following estimate. On one hand, as u is nondecreasing, the L^2 term $\|u - g^h\|^2$ is bounded below by $\|u^a - g^h\|^2$ where $u_n^a = a$ for $n < 0$, $u_0^a = \frac{1}{2}$ and $u_n^a = 1 - a$ for $n > 0$. On the other hand, we write that $\sqrt{(u_n - u_{n+1})^2 + (u_n - u_{n-1})^2} \geq \sqrt{2}|u_{n+1} - u_{n-1}| = \sqrt{2}(u_{n+1} - u_{n-1})$. We obtain:

$$E^h(u) \geq \frac{h}{2\sqrt{2}\lambda} \times 2Na^2 + \frac{1}{2}(u_{N+1} + u_N - u_{-N-1} - u_{-N}) = \frac{\sqrt{2}}{2\lambda} a^2 + 1 - 2a$$

As $\lambda \geq \lambda^* = \frac{\sqrt{2}}{4}$, minimizing this quantity over $a \in [0, \frac{1}{2}]$ leads to taking $a = \frac{1}{2}$, and we get $E^h(u) \geq \frac{\sqrt{2}}{8\lambda} = E^h(\frac{1}{2})$, hence $\bar{u}^h \equiv \frac{1}{2}$. \square

CHAPTER 7

UPPER BOUND ESTIMATE

In this chapter we prove the upper bound estimate of the point 2 of Theorem II.1, that is: $\exists \underline{h}, c > 0$ such that

$$\forall h \leq \underline{h}, \bar{E}^h - \bar{E} \leq ch^{2/3}$$

We first focus on the Dirichlet setting and will present later on the modifications needed for Neumann boundary conditions. As no reference to the continuous problem will appear in this chapter (except from its value \bar{E}) we drop the exponent h and denote the variables u^h, g^h simply by $u, g \in \mathbb{R}^{2N+1}$. Recall that the primal problem in the Dirichlet setting is:

$$\bar{u} = \arg \min_{\substack{(u_n)_{-2N \leq n \leq 2N} \\ u_{2N+1} = u_{2N} = 1 \\ u_{-2N-1} = u_{-2N} = 0}} \frac{h}{2\sqrt{2}\lambda} \|u - g\|_2^2 + \text{tv}_i^h(u) := E^h(u)$$

$$\text{with } \text{tv}_i^h(u) = \frac{1}{\sqrt{2}} \sum_{n=-2N}^{2N} \sqrt{(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2}$$

and where $g_n = 0$ for $n < 0$, $g_n = 1$ for $n > 0$ and $g_0 = 1/2$. The limit continuous energy is $\bar{E} = \bar{E}_D = 1$.

In the following we propose admissible functions u of a particular form to get upper bound estimates of the type

$$\bar{E}^h \leq E^h(u) \leq \bar{E} + ch^\theta$$

for $0 < \theta < 1$.

7.1 General strategy

The idea is to take a function u such that $u - g$ has a compact support of vanishing size but containing a number of points going to infinity. This is achieved by taking u_n , for $-N \leq n \leq N$, of the form (remember that $N = \frac{1}{h}$):

$$u_n = f\left(\frac{n}{N_\alpha}\right) \text{ with } N_\alpha = \lceil h^{-\alpha} \rceil \text{ and } 0 < \alpha < 1$$

where f is some continuous function increasing from $f(x) = 0$ for $x \leq -1$ to $f(x) = 1$ for $x \geq 1$. We also suppose in all what follows that f satisfies $f(-x) = 1 - f(x)$ for any $x \in \mathbb{R}$ to fulfill the conclusions of Proposition II.2.

As $u = g$ is constant for $|n| \geq N_\alpha$, one only has to consider what is happening in the transition phase, that is for $|n| < N_\alpha$ for the L^2 terms, and for $|n| \leq N_\alpha$ for the total variation terms. To understand what is at stake, let us first try with the piecewise affine function

$$f(x) = \begin{cases} 0 & \text{if } x < -1 \\ \frac{x+1}{2} & \text{if } -1 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

First compute the fidelity term:

$$\begin{aligned} \frac{h}{2} \|u - g\|_2^2 &= h \sum_1^{N_\alpha-1} (f(\frac{n}{N_\alpha}) - 1)^2 \\ &= \frac{h}{4N_\alpha^2} \sum_1^{N_\alpha-1} n^2 \\ &= \frac{hN_\alpha}{12} - \frac{h}{8} + \frac{h}{24N_\alpha} \end{aligned}$$

and then the total variation term:

$$\begin{aligned} \text{tv}_i^h(u) &= \frac{1}{\sqrt{2}} \sum_{-N_\alpha+1}^{N_\alpha-1} \sqrt{\left(\frac{n+1}{2N_\alpha} - \frac{n}{2N_\alpha}\right)^2 + \left(\frac{n}{2N_\alpha} - \frac{n-1}{2N_\alpha}\right)^2} \\ &\quad + \frac{1}{\sqrt{2}} \left| 1 - \frac{1}{2} \left(\frac{N_\alpha-1}{N_\alpha} + 1 \right) \right| + \frac{1}{\sqrt{2}} \left| \frac{1}{2} \left(\frac{-N_\alpha+1}{N_\alpha} + 1 \right) \right| \\ &= \frac{1}{\sqrt{2}} \left((2N_\alpha - 1) \times \sqrt{2 \times \frac{1}{4N_\alpha^2} + \frac{1}{N_\alpha}} \right) \\ &= 1 + \frac{\sqrt{2} - 1}{2N_\alpha} \end{aligned}$$

Note that the value of the limit energy appears in the above expression as $1 = \overline{E}$.

Combining the two terms finally leads to:

$$\begin{aligned}
 E^h(u) - \bar{E} &= \frac{\sqrt{2}-1}{2N_\alpha} + \frac{hN_\alpha}{12\sqrt{2}\lambda} - \frac{h}{8\sqrt{2}\lambda} + \frac{h}{24\sqrt{2}\lambda N_\alpha} \\
 &\leq \frac{\sqrt{2}-1}{2}h^\alpha + \frac{h(h^{-\alpha}+1)}{12\sqrt{2}\lambda} - \frac{h}{8\sqrt{2}\lambda} + \frac{h^{\alpha+1}}{24\sqrt{2}\lambda} \\
 &\leq \frac{\sqrt{2}-1}{2}h^\alpha + \frac{h^{1-\alpha}}{12\sqrt{2}\lambda} + \frac{h^{\alpha+1}}{24\sqrt{2}\lambda}
 \end{aligned}$$

The optimal choice of α is then to make the two dominant terms in h^α and $h^{1-\alpha}$ of the same order, hence $\alpha = 1/2$. We conclude that, for any $c > \frac{\sqrt{2}-1}{2} + \frac{1}{12\sqrt{2}\lambda}$, one has for h small enough

$$E^h(u) - \bar{E} \leq c\sqrt{h}$$

In the following we show that with a cubic function f , realizing a smoother transition, this procedure leads to the announced better result: there exist constants $c > 0$ and $\underline{h} > 0$ depending only on λ such that:

$$\forall h \leq \underline{h}, E^h(u) - \bar{E} \leq ch^{2/3} \quad (7.1)$$

7.2 Approach for a general function

In fact for any function f that is regular enough (\mathcal{C}^1), we have the following convergence when $h \rightarrow 0$: u^h converges to g in L^2 so $h\|u^h - g^h\|_2^2 \rightarrow 0$, and $\text{tv}_i^h(u) \rightarrow \text{tv}(g) = 1$. So $E(u) \rightarrow \bar{E}$. We want to estimate the speed of this convergence.

The L^2 term is easy to estimate:

$$\begin{aligned}
 \frac{h}{2}\|u - g\|_2^2 &= h \sum_{n=1}^{N_\alpha-1} (f(\frac{n}{N_\alpha}) - 1)^2 \\
 &= hN_\alpha \frac{1}{N_\alpha} \sum_1^{N_\alpha-1} (f(\frac{n}{N_\alpha}) - 1)^2 \\
 &\sim h^{1-\alpha} \int_0^1 (f-1)^2 \text{ when } N_\alpha \rightarrow \infty
 \end{aligned}$$

hence for any $c_1 > \frac{1}{\sqrt{2}\lambda} \int_0^1 (f-1)^2$, we have for h small enough:

$$\frac{h}{2\sqrt{2}\lambda}\|u - g\|_2^2 \leq c_1 h^{1-\alpha} \quad (7.2)$$

Manipulations on the total variation term are trickier, it is given by:

$$\begin{aligned} \text{tv}_i^h(u) &= \frac{1}{\sqrt{2}} \sum_{n=-N_\alpha+1}^{N_\alpha-1} \sqrt{\left(f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n}{N_\alpha}\right)\right)^2 + \left(f\left(\frac{n}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right)\right)^2} \\ &\quad + \frac{1}{\sqrt{2}} \left|1 - f\left(\frac{N_\alpha-1}{N_\alpha}\right)\right| + \frac{1}{\sqrt{2}} \left|f\left(\frac{-N_\alpha+1}{N_\alpha}\right)\right| \end{aligned}$$

The boundary terms simplify into

$$\frac{1}{\sqrt{2}} \left|1 - f\left(\frac{N_\alpha-1}{N_\alpha}\right)\right| + \frac{1}{\sqrt{2}} \left|f\left(\frac{-N_\alpha+1}{N_\alpha}\right)\right| = \sqrt{2} \left(1 - f\left(1 - \frac{1}{N_\alpha}\right)\right)$$

For the middle terms, we use the following lemma with $u_n = f\left(\frac{n}{N_\alpha}\right)$:

Lemma II.2. *If (u_n) is an increasing sequence, then for any n :*

$$\frac{1}{\sqrt{2}} \sqrt{(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2} \leq \frac{1}{2} (u_{n+1} - u_{n-1}) + d_n$$

with

$$d_n = \frac{1}{8} (u_{n+1} - u_{n-1}) (2u_n - u_{n+1} - u_{n-1}) \left(\frac{1}{u_{n+1} - u_n} - \frac{1}{u_n - u_{n-1}} \right) \quad (7.3)$$

$$= \frac{(u_{n+1} - u_{n-1}) (2u_n - u_{n+1} - u_{n-1})^2}{8(u_{n+1} - u_n)(u_n - u_{n-1})} \quad (7.4)$$

Proof. Denote $A = \sqrt{(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2}$ the quantity we want to estimate. Using $\sqrt{x+h} \leq \sqrt{x} + \frac{1}{2\sqrt{x}}h$ we get:

$$\begin{aligned} A &= \sqrt{2(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2 - (u_{n+1} - u_n)^2} \\ &= \sqrt{2(u_{n+1} - u_n)^2 + (u_{n+1} - u_{n-1})(2u_n - u_{n+1} - u_{n-1})} \\ &\leq \sqrt{2}(u_{n+1} - u_n) + \frac{1}{2\sqrt{2}(u_{n+1} - u_n)} (u_{n+1} - u_{n-1})(2u_n - u_{n+1} - u_{n-1}) \end{aligned}$$

And similarly

$$\begin{aligned} A &= \sqrt{2(u_n - u_{n-1})^2 + (u_{n+1} - u_n)^2 - (u_n - u_{n-1})^2} \\ &\leq \sqrt{2}(u_n - u_{n-1}) - \frac{1}{2\sqrt{2}(u_n - u_{n-1})} (u_{n+1} - u_{n-1})(2u_n - u_{n+1} - u_{n-1}) \end{aligned}$$

The result is obtained as the average of these two estimates. \square

Using this lemma leads to splitting the total variation terms under study into two terms. The term in $\frac{1}{2}(u_{n+1} - u_{n-1}) = \frac{1}{2}(f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right))$ is responsible for the convergence towards 1 as

$$\begin{aligned} \sum_{-N_\alpha+1}^{N_\alpha-1} \frac{1}{2} \left(f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right) \right) &= \frac{1}{2} \left(f(1) + f\left(1 - \frac{1}{N_\alpha}\right) - f(-1) - f\left(-1 + \frac{1}{N_\alpha}\right) \right) \\ &= f\left(1 - \frac{1}{N_\alpha}\right) \end{aligned}$$

For the term in d_n note that the symmetry of f gives $d_0 = 0$ and $d_{-n} = d_n$ so that the sum is reduced to $n \in \llbracket 1, N_\alpha - 1 \rrbracket$ and we get the expression:

$$\mathrm{tv}_i^h(u) \leq 1 + (\sqrt{2} - 1)(1 - f(1 - \frac{1}{N_\alpha})) + \sum_1^{N_\alpha-1} d_n \quad (7.5)$$

Now a general study of the sum $\sum_1^{N_\alpha-1} d_n$ using only assumptions on the derivatives of f seems to fail: we explain why below. However our result will come from its exact computation in the case of a cubic function f : we give this calculation in the next section.

To keep it with a general f , in order to bound d_n in its expression (7.3) one would assume (and these assumptions are indeed satisfied by our successful cubic f below):

1. f' and f'' are bounded on $(0, 1)$
2. $f' > 0$ and $f'' < 0$ on $(0, 1)$ (that is, f is increasing to 1 f' is decreasing to 0)

This allows one to write:

$$\begin{aligned} \bullet 0 &\leq u_{n+1} - u_{n-1} = f(\frac{n+1}{N_\alpha}) - f(\frac{n-1}{N_\alpha}) \leq \frac{2}{N_\alpha} \|f'\|_\infty \\ \bullet 0 &\leq 2u_n - u_{n-1} - u_{n+1} = 2f(\frac{n}{N_\alpha}) - f(\frac{n-1}{N_\alpha}) - f(\frac{n+1}{N_\alpha}) \leq \frac{1}{N_\alpha^2} \|f''\|_\infty \\ \bullet 0 &\leq \frac{1}{u_{n+1} - u_n} - \frac{1}{u_n - u_{n-1}} = \frac{1}{f(\frac{n+1}{N_\alpha}) - f(\frac{n}{N_\alpha})} - \frac{1}{f(\frac{n}{N_\alpha}) - f(\frac{n-1}{N_\alpha})} \end{aligned}$$

and we therefore obtain:

$$\begin{aligned} \sum_1^{N_\alpha-1} d_n &\leq \frac{1}{4N_\alpha^3} \|f'\|_\infty \|f''\|_\infty \sum_1^{N_\alpha-1} \frac{1}{f(\frac{n+1}{N_\alpha}) - f(\frac{n}{N_\alpha})} - \frac{1}{f(\frac{n}{N_\alpha}) - f(\frac{n-1}{N_\alpha})} \\ &\leq \frac{1}{4N_\alpha^3} \|f'\|_\infty \|f''\|_\infty \left(\frac{1}{f(1) - f(1 - \frac{1}{N_\alpha})} - \frac{1}{f(\frac{1}{N_\alpha}) - f(0)} \right) \\ &\leq \frac{1}{4N_\alpha^3} \|f'\|_\infty \|f''\|_\infty \frac{1}{1 - f(1 - \frac{1}{N_\alpha})} \end{aligned}$$

(we drop the term in $-\frac{1}{f'(0)}$ as it will not count against the unbounded one in $\frac{1}{f'(1)}$)

Finally, from (7.5) we have

$$\mathrm{tv}_i^h(u) \leq 1 + (\sqrt{2} - 1)(1 - f(1 - \frac{1}{N_\alpha})) + \frac{1}{4N_\alpha^3} \|f'\|_\infty \|f''\|_\infty \frac{1}{1 - f(1 - \frac{1}{N_\alpha})}$$

At this point there seems to be a trade off between $(1 - f(1 - \frac{1}{N_\alpha}))$ and its inverse. We should therefore make a reasonable assumption on the decay of f' to 0 in 1: we want to suppose that

$$1 - f(1 - \frac{1}{N_\alpha}) \sim \left(\frac{1}{N_\alpha} \right)^p \quad (7.6)$$

for some $p > 0$ such that the terms $1 - f(1 - \frac{1}{N_\alpha})$ and $(N_\alpha^3(1 - f(1 - \frac{1}{N_\alpha})))^{-1}$ are of the same order, that is $p = 3 - p$ i.e. $p = 3/2$. This could be fine and lead to $\text{tv}_i^h(u) \leq 1 + c_2 h^{\frac{3}{2}\alpha}$ so $E^h(u) \leq \bar{E} + c_1 h^{1-\alpha} + c_2 h^{\frac{3}{2}\alpha}$ which we would optimize into $\alpha = 2/5$ and have finally $E^h(u) \leq \bar{E} + ch^{3/5}$.

However, the assumption $1 - f(1 - \frac{1}{N_\alpha}) \sim (N_\alpha)^{-3/2}$ contradicts the fact that f'' is bounded near 1 which we use in the upper bound for d_n . For a bounded f'' one can only chose $p \geq 2$ in (7.6) and get a N_α^{-1} term from $(N_\alpha^3(1 - f(1 - \frac{1}{N_\alpha})))^{-1}$. Finally this general analysis seems to only lead to a rate in \sqrt{h} as before.

7.3 Result for a particular function

If we compute exactly the d_n term for an easy to deal with function f , the derivative terms in f' and f'' are multiplied and oddly give a better estimate. We chose to take a polynomial f given by:

$$f(t) = \begin{cases} 0 & \text{if } t \leq -1 \\ \frac{1}{2}(1+t)^k & \text{if } -1 \leq t \leq 0 \\ 1 - \frac{1}{2}(1-t)^k & \text{if } 0 \leq t \leq 1 \\ 1 & \text{if } t \geq 1 \end{cases} \quad (7.7)$$

for some integer $k \geq 1$. As $1 - f(1 - \frac{1}{N_\alpha}) = \frac{1}{N_\alpha^k}$ we have from equation (7.5):

$$\text{tv}_i^h(u) \leq 1 + (\sqrt{2} - 1) \frac{1}{N_\alpha^k} + \sum_1^{N_\alpha-1} d_n \quad (7.8)$$

and then we can explicitly the sum of the d_n for small values of k . As we have seen in the beginning, for $k = 1$ one finds $\sum d_n$ to be of order N_α^{-1} and this leads to the \sqrt{h} rate. For $k = 2$ one can obtain

$$\sum_1^{N_\alpha-1} d_n \leq c \frac{\log N_\alpha}{N_\alpha^2}$$

for some constant $c > 0$, and this would lead to a rate in h^θ for any $\theta < 2/3$. We get even better when taking $k = 3$ (and numerical results seem to show that taking $k > 3$ does not lead to better estimates):

Fact II.2. *For the choice of f given by (7.7) with $k = 3$, one has*

$$\sum_1^{N_\alpha-1} d_n \leq \frac{3}{N_\alpha^2}$$

Proof. Let us denote

$$\begin{aligned}\Delta_+ &:= f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n}{N_\alpha}\right) = \frac{1}{2} \left(\left(1 - \frac{n}{N_\alpha}\right)^3 - \left(1 - \frac{n+1}{N_\alpha}\right)^3 \right) \\ &= \frac{1}{2} \left(3\left(1 - \frac{n}{N_\alpha}\right)^2 \frac{1}{N_\alpha} - 3\left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha^2} + \frac{1}{N_\alpha^3} \right) \\ &= \frac{3}{2N_\alpha} \left(\left(1 - \frac{n}{N_\alpha}\right)^2 - \left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha} + \frac{1}{3N_\alpha^2} \right)\end{aligned}$$

Similarly (that is, taking $n \leftarrow n - 1$),

$$\Delta_- := f\left(\frac{n}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right) = \frac{3}{2N_\alpha} \left(\left(1 - \frac{n}{N_\alpha}\right)^2 + \left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha} + \frac{1}{3N_\alpha^2} \right)$$

so that

$$\begin{aligned}\Delta_+ + \Delta_- &= f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right) = \frac{3}{N_\alpha} \left(\frac{1}{3N_\alpha^2} + \left(1 - \frac{n}{N_\alpha}\right)^2 \right) \\ \Delta_- - \Delta_+ &= 2f\left(\frac{n}{N_\alpha}\right) - f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right) = \frac{3}{N_\alpha^2} \left(1 - \frac{n}{N_\alpha}\right)\end{aligned}$$

and

$$\begin{aligned}\Delta_+ \times \Delta_- &= \left(f\left(\frac{n+1}{N_\alpha}\right) - f\left(\frac{n}{N_\alpha}\right) \right) \left(f\left(\frac{n}{N_\alpha}\right) - f\left(\frac{n-1}{N_\alpha}\right) \right) \\ &= \frac{9}{4N_\alpha^2} \left(\left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{N_\alpha^3} - \left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha} \right) \\ &\quad \times \left(\left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{N_\alpha^3} + \left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha} \right) \\ &= \frac{9}{4N_\alpha^2} \left(\left(\left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{3N_\alpha^2} \right)^2 - \left(\left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha} \right)^2 \right)\end{aligned}$$

We can now estimate d_n thanks to expression (7.4):

$$\begin{aligned}d_n &= \frac{1}{8} \times \frac{\left(\frac{3}{N_\alpha} \left(\frac{1}{3N_\alpha^2} + \left(1 - \frac{n}{N_\alpha}\right)^2 \right) \right) \times \left(\frac{3}{N_\alpha^2} \left(1 - \frac{n}{N_\alpha}\right) \right)^2}{\frac{9}{4N_\alpha^2} \left(\left(\left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{3N_\alpha^2} \right)^2 - \left(\left(1 - \frac{n}{N_\alpha}\right) \frac{1}{N_\alpha} \right)^2 \right)} \\ &= \frac{3}{2N_\alpha^3} \left(1 - \frac{n}{N_\alpha}\right)^2 \times \frac{\frac{1}{3N_\alpha^2} + \left(1 - \frac{n}{N_\alpha}\right)^2}{\left(\left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{3N_\alpha^2} \right)^2 - \left(1 - \frac{n}{N_\alpha}\right)^2 \frac{1}{N_\alpha^2}}\end{aligned}$$

Then as $n \leq N_\alpha - 1$ we can use

$$\left(\left(1 - \frac{n}{N_\alpha}\right)^2 + \frac{1}{3N_\alpha^2} \right)^2 - \left(1 - \frac{n}{N_\alpha}\right)^2 \frac{1}{N_\alpha^2} \geq \left(1 - \frac{n}{N_\alpha}\right)^4 - \frac{1}{3N_\alpha^2} \left(1 - \frac{n}{N_\alpha}\right)^2 > 0$$

and make the variable change $n \leftarrow N - n$ to get:

$$\sum_{n=1}^{N_\alpha-1} d_n \leq \frac{3}{2N_\alpha^3} \sum_{n=1}^{N_\alpha-1} n^2 \frac{\frac{1}{3N_\alpha^2} + n^2}{n^4 - \frac{1}{3N_\alpha^2} n^2} \leq \frac{3}{2N_\alpha^3} \sum_{n=1}^{N_\alpha-1} 2 \leq \frac{3}{N_\alpha^2}$$

because $\frac{1}{3N_\alpha^2} + n^2 \leq 2(n^2 - \frac{1}{3N_\alpha^2})$. \square

This concludes our proof of the upper bound inequality for the main Theorem II.1. Indeed, when combining the total variation estimate (7.5) with the L^2 estimate (7.2), we finally are able to state the following: for any $c_1 > \frac{1}{\sqrt{2}\lambda} \int_0^1 (f-1)^2 = \frac{1}{28\sqrt{2}\lambda}$ and $c_2 > 3$, there exists $\underline{h} > 0$ such that

$$\forall h \leq \underline{h}, E^h(u) \leq \bar{E} + c_1 h^{1-\alpha} + c_2 h^{2\alpha} \quad (7.9)$$

Taking $\alpha = 1/3$ then proves our result (7.1). More precisely, given c_1, c_2 and $h > 0$, the best α in (7.9) must satisfy $-c_1 h^{1-\alpha} \log h + 2c_2 h^{2\alpha} \log h = 0$ which leads to $\alpha = \frac{1}{3} - \frac{\log(2c_2/c_1)}{3 \log h}$ and gives the upper bound $E^h(u) \leq \bar{E} + ch^{2/3}$ for the constant $c = (2^{1/3} + 2^{-2/3}) c_1^{2/3} c_2^{1/3}$. We note that the value of c varies in $\lambda^{-2/3}$.

7.4 Modifications for Neumann boundary conditions

In this section we adjust the admissible variable u from the previous section to explain why the upper bound result (7.1) remains valid for Neumann boundary conditions. In the following, c denotes a constant depending only on λ that can change from line to line.

Remember from section 6.2 that with the Neumann boundary conditions, the limit continuous value of the energy is changed to $\bar{E} = \bar{E}_N = 1 - \sqrt{2}\lambda$ when $\lambda \leq \lambda^* = \frac{\sqrt{2}}{4}$. Because of the form of this continuous solution, it is natural to consider, for u the cubic transition in the Dirichlet setting of the previous section, the variable v given by

$$\forall n \in \llbracket -N, N \rrbracket, v_n = \frac{1}{2} + \mu(u_n - \frac{1}{2})$$

Here $\mu \in (0, 1)$ is a shrinking parameter that we adjust so that $v_N = 1 - a_{opt} = 1 - \sqrt{2}\lambda$: as $u_N = 1$ this corresponds to taking $\mu = 1 - 2\sqrt{2}\lambda$.

We write $v_n = f_\mu(\frac{n}{N_\alpha})$ for the function $f_\mu = \frac{1}{2} + \mu(f - \frac{1}{2})$ which is such that $f_\mu(x) = \frac{1+\mu}{2} = 1 - \sqrt{2}\lambda$ for $x \geq 1$. This leads to splitting the fidelity term into:

$$\frac{h}{2} \|v - g\|_2^2 = h \sum_{n=1}^{N_\alpha} (v_n - 1)^2 + h \sum_{n=N_\alpha+1}^N (v_n - 1)^2$$

Then on one hand when $N_\alpha \rightarrow \infty$,

$$h \sum_{n=1}^{N_\alpha} (v_n - 1)^2 \sim h^{1-\alpha} \int_0^1 (f_\mu - 1)^2 \text{ so } h \sum_{n=1}^{N_\alpha} (v_n - 1)^2 \leq ch^{1-\alpha}$$

and on the other hand

$$h \sum_{n=N_\alpha+1}^N (v_n - 1)^2 = h(N - N_\alpha) \times 2\lambda^2 \leq 2\lambda^2$$

For the total variation term, we have

$$\begin{aligned} \mathbf{tv}_i^h(v) &= \mu \mathbf{tv}_i^h(u) = (1 - 2\sqrt{2}\lambda) \mathbf{tv}_i^h(u) \\ &\leq (1 - 2\sqrt{2}\lambda)(1 + ch^{2\alpha}) \\ &\leq 1 - 2\sqrt{2}\lambda + ch^{2\alpha} \end{aligned}$$

so finally

$$\begin{aligned} E^h(v) &= \frac{h}{2\sqrt{2}\lambda} \|v - g\|_2^2 + \mathbf{tv}_i^h(v) \\ &\leq \sqrt{2}\lambda + ch^{1-\alpha} + 1 - 2\sqrt{2}\lambda + ch^{2\alpha} \\ &\leq \bar{E} + ch^{2/3} \end{aligned}$$

when taking $\alpha = 1/3$, and we indeed have the same estimate for Neumann than for Dirichlet boundary conditions.

CHAPTER 8

LOWER BOUND ESTIMATE

In this chapter we now prove the lower bound estimate of the point 2 of Theorem II.1, that is: $\exists \underline{h}, c > 0$ such that

$$\forall h \leq \underline{h}, ch^{2/3} \leq \overline{E}^h - \overline{E}$$

Symmetrically to what we did before, we will obtain this bound by proposing an admissible solution, but for the dual problem. However, the proof we present will be less direct than in the previous chapter as we will first have to study the continuous problem corresponding to the limit value of the rescaled energy $h^{-2/3}(\overline{E}^h - \overline{E})$.

8.1 Dual problem

Writing that

$$\sqrt{(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2} = \max_{p_n^2 + q_n^2 \leq 1} q_n(u_{n+1} - u_n) + p_n(u_n - u_{n-1})$$

we obtain the following dual problem of (6.1):

$$\begin{aligned} & \max_{\substack{p_n^2 + q_n^2 \leq 1 \\ -N \leq n \leq N}} \min_{u \in \mathbb{R}^{2N+1}} \frac{1}{\sqrt{2}} \left\{ \sum_{n=-N}^N \frac{h}{2\lambda} (u_n - g_n)^2 + q_n(u_{n+1} - u_n) + p_n(u_n - u_{n-1}) \right\} \\ = & \max_{\substack{p_n^2 + q_n^2 \leq 1 \\ -N \leq n \leq N}} \min_{u \in \mathbb{R}^{2N+1}} \frac{1}{\sqrt{2}} \left\{ \sum_{n=-N}^N \frac{h}{2\lambda} (u_n - g_n)^2 + \sum_{n=-N+1}^{N-1} u_n (q_{n-1} - q_n + p_n - p_{n+1}) \right. \\ & \left. + u_N (q_{N-1} - q_N + p_N) + u_{-N} (-q_{-N} + p_{-N} - p_{-N+1}) \right. \\ & \left. + u_{N+1} q_N - u_{-N-1} p_{-N} \right\} \end{aligned}$$

From this point on, we focus exclusively on Dirichlet boundary conditions, that is $u_N = u_{N+1} = 1$; $u_{-N} = u_{-N-1} = 0$. See section 8.5 for Neumann boundary conditions.

For $|n| < N$, we find that $u_n = g_n - \frac{\lambda}{h}(q_{n-1} - q_n + p_n - p_{n+1})$, and the value of the dual problem is consequently (after simplification using the value of g_n):

$$\max_{\substack{p_n^2 + q_n^2 \leq 1 \\ -N \leq n \leq N}} \frac{1}{\sqrt{2}} \left\{ \frac{1}{2}(q_{-1} + q_0 + p_0 + p_1) - \frac{\lambda}{2h} \sum_{n=-N+1}^{N-1} (q_{n-1} - q_n + p_n - p_{n+1})^2 \right\}$$

Now we make two more simplifications before turning to an evaluation of the convergence rate of this quantity. First, one easily checks that the objective is concave and invariant by the change $(q_n, p_n) \rightarrow (p_{-n}, q_{-n})$: as a consequence, one can find a solution satisfying $q_n = p_{-n}$ for all n .

Second, duality indicates that at the optimum one should have for all n the relation $\sqrt{(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2} = q_n(u_{n+1} - u_n) + p_n(u_n - u_{n-1})$. Taking $n = 0$ in this equality gives, thanks to the symmetry of u (Proposition II.2), that $\sqrt{2}|u_1 - u_0| = (q_0 + p_0)(u_1 - u_0)$ so that $q_0 = p_0 = \frac{\sqrt{2}}{2}$ because $u_1 > u_0 = \frac{1}{2}$ (Proposition II.3). Simplifying the term $(q_{n-1} - q_n + p_n - p_{n+1})^2$ which is invariant by $n \rightarrow -n$ and vanishes at $n = 0$, we finally get

$$\bar{E}^h = \max \frac{1}{2} + \frac{1}{\sqrt{2}} p_1 - \frac{\lambda}{\sqrt{2}h} \sum_{n=1}^{N-1} (p_{-n+1} - p_{-n} + p_n - p_{n+1})^2 \quad (8.1)$$

with the constraint that $p_n^2 + p_{-n}^2 \leq 1$ for all $n \in \llbracket 1, N \rrbracket$ and that $p_0 = \frac{\sqrt{2}}{2}$.

8.2 Change of variables

We are interested in the evaluation of the convergence rate of the value of the problem (8.1) towards its continuous limit $\bar{E} = \bar{E}_D = 1$. To begin with, we notice that taking $p_n \equiv \sqrt{2}/2$ gives $\bar{E}^h \geq \bar{E}$. Consequently, we expect the optimal value of p to be close to $\sqrt{2}/2$ for N large. Together with the symmetry regarding $n \rightarrow -n$ of the objective, this leads us to proposing the following change of variables: for $n \in \llbracket 0, N \rrbracket$

$$s_n = \frac{1}{\sqrt{2}}(p_n + p_{-n}) - 1; \quad r_n = \frac{1}{\sqrt{2}}(p_n - p_{-n})$$

for which we calculate

$$p_{-n+1} - p_{-n} + p_n - p_{n+1} = \frac{1}{\sqrt{2}}(s_{n-1} - s_{n+1} + 2r_n - r_{n-1} - r_{n+1})$$

$$p_n^2 + p_{-n}^2 \leq 1 \iff s_n^2 + 2s_n + r_n^2 \leq 0$$

and it gives rise to

$$\overline{E}^h - \overline{E} = \max_{\substack{(s_n, r_n)_{0 \leq n \leq N} \\ s_0 = r_0 = 0 \\ s_n^2 + 2s_n + r_n^2 \leq 0}} \frac{1}{2}(s_1 + r_1) - \frac{\lambda}{2\sqrt{2}h} \sum_{n=1}^{N-1} (s_{n-1} - s_{n+1} + 2r_n - r_{n-1} - r_{n+1})^2$$

We would like to show that $\overline{E}^h - \overline{E} \geq cN^{-\alpha}$ for some exponent $0 < \alpha < 1$. If we introduce $\tau = 1/N^\beta$ for some $\beta \in (0, \alpha)$ and $\sigma_n = N^\alpha s_n$, $\rho_n = N^{\alpha-\beta} r_n$, then we can force the appearance of first and second discrete derivatives for σ and ρ as

$$\begin{aligned} (\overline{E}^h - \overline{E})N^\alpha &= \max_{\substack{(\sigma_n, \rho_n) \\ 0 \leq n \leq N}} \frac{1}{2}(\sigma_1 + \frac{\rho_1}{\tau}) \\ &\quad - \frac{\lambda}{2\sqrt{2}} N^{1-\alpha-\beta} \tau \sum_{n=1}^{N-1} \left(\frac{\sigma_{n-1} - \sigma_{n+1}}{\tau} + \frac{2\rho_n - \rho_{n-1} - \rho_{n+1}}{\tau^2} \right)^2 \end{aligned}$$

along with the constraints $\sigma_0 = \rho_0 = 0$ and $N^{-\alpha}\sigma_n^2 + 2\sigma_n + N^{2\beta-\alpha}\rho_n^2 \leq 0$.

If $1 - \alpha - \beta = 0$, we find that as $N \rightarrow \infty$, the limiting energy in the variational problem should be of the form,

$$\max \frac{1}{2}\rho'(0) - \frac{\lambda}{2\sqrt{2}} \int_0^\infty |2\sigma' + \rho''|^2$$

for functions $\sigma, \rho : [0, \infty) \rightarrow \mathbb{R}$ with $\sigma(0) = \rho(0) = 0$. The constraint, on the other hand, becomes

$$\begin{cases} \rho^2 = 0 & \text{if } 2\beta - \alpha > 0 \\ 2\sigma + \rho^2 \leq 0 & \text{if } 2\beta - \alpha = 0 \Leftrightarrow \beta = 1/3, \alpha = 2/3 \\ 2\sigma \leq 0 & \text{if } 2\beta - \alpha < 0. \end{cases}$$

In the first case, which is when $\alpha < 2/3$, this continuous limit problem has value zero so we may expect that the discrete renormalized energy goes to zero, and as a consequence that $\overline{E}^h - \overline{E} = o(N^{-\alpha})$ as $N \rightarrow \infty$. In the third case, the continuous problem has value $+\infty$ and we expect that $N^\alpha(\overline{E}^h - \overline{E}) \rightarrow +\infty$ for $\alpha > 2/3$. We would like to show that in the second case, that is $\alpha = 2/3$, the limiting problem has a positive value c so that $\overline{E}^h - \overline{E} \geq cN^{-2/3}$ for sufficiently large N . Consequently we deal with the problem

$$\max_{(\sigma, \rho) \in S} \frac{1}{2}\rho'(0) - \frac{\lambda}{2\sqrt{2}} \int_0^\infty (2\sigma' + \rho'')^2 =: D(\sigma, \rho) \quad (8.2)$$

where S is the set of couples of functions $\sigma, \rho : [0, \infty) \rightarrow \mathbb{R}$ such that: $\sigma(0) = \rho(0) = 0$, $2\sigma + \rho^2 \leq 0$, ρ admits a right derivative at 0 and the distributional derivative $2\sigma' + \rho''$ is in $L^2(0, \infty)$.

Our strategy is now the following: in section 8.3 we prove that problem (8.2) has a positive value and investigate the form of the solution (σ, ρ) . Then in section 8.4 we explain how to discretize it in order to get the positivity, for h small enough, of our discrete problem:

$$(\bar{E}^h - \bar{E})h^{-2/3} = \max_{\substack{(\sigma_n, \rho_n)_{0 \leq n \leq N} \\ \sigma_0 = \rho_0 = 0 \\ N^{-2/3} \sigma_n^2 + 2\sigma_n + \rho_n^2 \leq 0}} D^h(\sigma, \rho) \quad (8.3)$$

$$\text{with } D^h(\sigma, \rho) := \frac{1}{2} \left(\sigma_1 + \frac{\rho_1}{\tau} \right) - \frac{\lambda}{2\sqrt{2}} \tau \sum_{n=1}^{N-1} \left(\frac{\sigma_{n-1} - \sigma_{n+1}}{\tau} + \frac{2\rho_n - \rho_{n-1} - \rho_{n+1}}{\tau^2} \right)^2$$

8.3 Study of the limit problem

First, the change of variables $\hat{\sigma}(t) = \lambda^{-2/3} \sigma(t\lambda^{-1/3})$, $\hat{\rho}(t) = \lambda^{-1/3} \rho(t\lambda^{-1/3})$ shows that (adding the parameter λ to the arguments of D)

$$\max D(\sigma, \rho, \lambda) = \lambda^{-2/3} \max D(\sigma, \rho, 1)$$

or even, for any $\lambda_0 > 0$, that

$$\max D(\sigma, \rho, \lambda) = \left(\frac{\lambda}{\lambda_0} \right)^{-2/3} \max D(\sigma, \rho, \lambda_0)$$

Consequently, we can restrict our study of the problem to a single value of λ . In all the following, we chose to take $\lambda = \frac{1}{\sqrt{2}}$ so that problem (8.2) writes:

$$\frac{1}{2} \max_{(\sigma, \rho) \in S} \rho'(0) - \frac{1}{2} \int_0^\infty (2\sigma' + \rho'')^2$$

8.3.1 A dual of the dual

To understand the solution of problem (8.2), we derive a dual of it writing

$$-\frac{1}{2} \int_0^\infty (2\sigma' + \rho'')^2 = \inf_{\psi} \int_0^\infty (2\sigma' + \rho'')\psi + \frac{1}{2} \int_0^\infty \psi^2 \quad (8.4)$$

where the infimum lies on $\psi \in \mathcal{C}_c^\infty([0, \infty))$, the set of the restrictions to $[0, \infty)$ of smooth functions with compact support in \mathbb{R} . Note that if σ, ρ are regular enough one has at the optimum $\psi = -(2\sigma' + \rho'')$. Integrating by parts and using the fact that $\sigma(0) = \rho(0) = 0$ for any $(\sigma, \rho) \in S$, we obtain the dual problem

$$\frac{1}{2} \inf_{\psi} \frac{1}{2} \int_0^\infty \psi^2 + \sup_{(\sigma, \rho) \in S} (1 - \psi(0))\rho'(0) + \int_0^\infty (\rho\psi'' - 2\sigma\psi')$$

First, taking for ρ a bounded smooth function with $|\rho'(0)|$ as large as we want, we see that one must have $\psi(0) = 1$. Second, we relax the constraint $(\sigma, \rho) \in S$ in the

remaining integral into just $2\sigma + \rho^2 \leq 0$ (we will show below that strong duality with problem (8.2) actually occurs) to get:

$$\frac{1}{2} \inf_{\psi(0)=1} \frac{1}{2} \int_0^\infty |\psi|^2 + \int_0^\infty H(\psi', \psi'')$$

where the function H is defined for $x, y \in \mathbb{R}$ by

$$H(x, y) = \sup_{2\sigma + \rho^2 \leq 0} -2\sigma x + \rho y = \begin{cases} +\infty & \text{if } x > 0 \text{ or } x = 0, y \neq 0, \\ 0 & \text{if } (x, y) = (0, 0), \\ \frac{y^2}{4|x|} & \text{if } x < 0. \end{cases}$$

(taking $\rho = 0$, $\sigma = -N$ for the case $x > 0$ or $x = 0, y \neq 0$; and $\rho = -y/2x$, $\sigma = -\rho^2/2$ for the case $x < 0$). Observe that necessarily $\psi' \leq 0$. Denoting $\phi = \sqrt{-\psi'}$ gives $\phi' = -\psi''/(2\sqrt{-\psi'})$ so that $H(\psi', \psi'') = |\phi'|^2$. Then, one has $\psi(x) = 1 - \int_0^x \phi(t)^2 dt$. In particular as ψ^2 is integrable, one must have $\int_0^\infty \phi(t)^2 dt = 1$ and $\psi(x) = \int_x^\infty \phi(t)^2 dt$. Hence the dual problem can be rewritten (extending the search of ϕ to $H^1(0, \infty)$ by density)

$$\frac{1}{2} \inf_{(\phi, \psi) \in S'} \left\{ \frac{1}{2} \int_0^\infty |\psi|^2 + \int_0^\infty |\phi'|^2 \right\} \quad (8.5)$$

where $S' = \{(\phi, \psi) \text{ s.t. } \phi \in H^1(0, \infty), \|\phi\|_{L^2}^2 = 1 \text{ and } \psi(x) = \int_x^\infty \phi(t)^2 dt\}$.

It turns out this problem has a positive value:

Proposition II.4. *Problem (8.5) admits a minimizer $(\phi, \psi) \in S'$.*

Proof. Consider a minimizing sequence (ϕ_n, ψ_n) : as ϕ_n is bounded in $H^1(0, \infty)$, up to a subsequence it converges to some ϕ , moreover the convergence is strong in $L^2(0, T)$ for any $T > 0$, and $\int_0^\infty \phi^2 \leq 1$. We also assume that ψ_n converges, weakly in $L^2(0, +\infty)$, to some ψ . In addition, $\psi_n(x) = 1 - \int_0^x \phi_n^2 \rightarrow 1 - \int_0^x \phi^2 =: \tilde{\psi}(x)$ for any $x \geq 0$, and one even has $|\psi_n(x) - \tilde{\psi}(x)| = |\int_0^x (\phi_n - \phi)(\phi_n + \phi)| \leq 2\|\phi_n - \phi\|_{L^2(0, x)}$ hence the convergence is locally uniform. Consequently, it must be that $\tilde{\psi} = \psi$. As $\int_0^\infty |\psi|^2 < +\infty$, we deduce that ψ (which is nonincreasing) goes to 0 at infinity, hence $\int_0^\infty \phi^2 = 1$. It follows that (ϕ, ψ) is a minimizer of (8.5). \square

To recover the positive value of problem (8.2), we now need to show that strong duality holds. To do that we first prove some properties of the minimizer (ϕ, ψ) .

Proposition II.5. *The minimizer $(\phi, \psi) \in S'$ of problem (8.5) satisfies:*

1. $\psi, \phi \in C^\infty([0, \infty)) \cap L^2(0, \infty)$.
2. $\phi'(0) = 0$ and $\phi'' = k\phi$ where k is the primitive of ψ given by $k(t) = \int_0^t \psi - A$ with $A = \|\phi'\|_{L^2}^2 + \|\psi\|_{L^2}^2$.
3. $\phi \geq 0$, $\phi(0) > 0$, and ϕ is nonincreasing and tends to zero at infinity.

Proof. A first remark is that $\psi' = -\phi^2 \in L^1(0, \infty)$ and $\psi'' = -2\phi\phi' \in L^1(0, \infty)$, hence ψ is at least C^1 . Moreover, if (ϕ, ψ) is a minimizer, so is $(|\phi|, \psi)$. The solution of (8.5) being unique, one has $\phi \geq 0$.

From this solution (ϕ, ψ) , let us form for $\varepsilon \in \mathbb{R}$ and for a test function η

$$\phi_\varepsilon = \frac{\phi + \varepsilon\eta}{\|\phi + \varepsilon\eta\|_{L^2}}; \quad \psi_\varepsilon(x) = \int_x^\infty \phi_\varepsilon^2$$

Then $(\phi_\varepsilon, \psi_\varepsilon)$ are admissible in the dual of the dual problem and one computes:

$$\begin{aligned} \phi_\varepsilon^2 &= \phi^2 + 2\varepsilon\eta\phi - 2\varepsilon\phi^2 \int_0^\infty \phi\eta + O(\varepsilon^2) \\ \psi_\varepsilon^2(x) &= \psi^2(x) + 4\varepsilon\psi(x) \int_x^\infty \phi\eta - 4\varepsilon\psi^2(x) \int_0^\infty \phi\eta + O(\varepsilon^2) \\ \phi'_\varepsilon &= \phi' + \varepsilon\eta' - \varepsilon\phi' \int_0^\infty \phi\eta + O(\varepsilon^2) \end{aligned}$$

so that, after noting that $\int_0^\infty \psi(x) \int_x^\infty \phi\eta dx = \int_0^\infty \phi\eta\nu$ with $\nu(t) = \int_0^t \psi$, one has

$$\begin{aligned} \int_0^\infty |\phi'_\varepsilon|^2 &= \int_0^\infty |\phi'|^2 - 2\varepsilon \int_0^\infty |\phi'|^2 \int_0^\infty \phi\eta + 2\varepsilon \int_0^\infty \phi'\eta' + O(\varepsilon^2) \\ \int_0^\infty |\psi_\varepsilon|^2 &= \int_0^\infty |\psi|^2 - 4\varepsilon \int_0^\infty |\psi|^2 \int_0^\infty \phi\eta + 4\varepsilon \int_0^\infty \phi\eta\nu + O(\varepsilon^2) \end{aligned}$$

Now the optimality of (ψ, ϕ) in problem (8.5) leads to

$$\int_0^\infty \phi\eta\nu - \int_0^\infty |\psi|^2 \int_0^\infty \phi\eta + \int_0^\infty \phi'\eta' - \int_0^\infty |\phi'|^2 \int_0^\infty \phi\eta = 0 \quad (8.6)$$

First, as this relation holds for any $\eta \in C_c^\infty(0, \infty)$, we have $\phi'' = k\phi$ (with $k = \nu - A$ where $A = \|\psi\|_{L^2}^2 + \|\phi'\|_{L^2}^2$) in the weak sense. However this relation induces the regularity of ϕ and ψ which are finally C^∞ . In addition, re-evaluating the relation (8.6) with now $\eta \in C_c^\infty([0, \infty))$, we also deduce that $\phi'(0) = 0$.

To finish with, one must have $\phi(0) > 0$ as otherwise ϕ would be zero everywhere as solution of $\phi'' = k\phi$, $\phi'(0) = \phi(0) = 0$. And for its monotonicity, note that $\phi'' = k\phi$ has the sign of k which is nonincreasing since $k' = \psi \geq 0$. Hence ϕ'' is first nonpositive (starting at $\phi''(0) = -A\phi(0) \leq 0$) then possibly nonnegative. As a consequence, ϕ' is first nonincreasing, and hence nonpositive since $\phi'(0) = 0$, then can become nondecreasing. But even in that case, ϕ' has to remain nonpositive otherwise one has $\phi'(t) \geq c > 0$ for t large enough so $\phi(t) \geq ct + c'$ which contradicts the fact that ϕ^2 is integrable. This concludes the proof. \square

In the following, functions ϕ, ψ, k and constant A denote the solution described in Proposition II.5. We show that strong duality holds between problems (8.2) and (8.5) by building an admissible couple $(\sigma, \rho) \in S$ satisfying $\psi = -(2\sigma' + \rho'')$ and using identity (8.4). We divide our study in two cases: either $\phi > 0$ on \mathbb{R}^+ (the ‘‘positive’’ case), or $\phi > 0$ on $[0, a)$ and $\phi = 0$ on $[a, +\infty[$ for some $a > 0$ (the ‘‘compact support’’ case). Note that numerical experiments seem to show we actually are in the ‘‘compact support’’ case, see Figure 8.1.

8.3.2 Strong duality result

In the “positive” case, recalling how the dual problem was obtained, one defines the following functions: $\rho = -\phi'/\phi$, $\sigma = -\rho^2/2$ and then checks that $2\sigma + \rho^2 \leq 0$, $\sigma(0) = \rho(0) = 0$, $\rho'(0) = A$ and $2\sigma' + \rho'' = -\psi$ so that strong duality holds as:

$$\frac{1}{2} \left\{ \rho'(0) - \frac{1}{2} \int_0^\infty |2\sigma' + \rho''|^2 \right\} = \frac{1}{2} \left\{ \int_0^\infty \phi'^2 + \frac{1}{2} \int_0^\infty \psi^2 \right\}$$

In the “compact support” case, one still defines $\rho = -\phi'/\phi$ and $\sigma = -\rho^2/2$ on $[0, a)$. Then one has to decide what to do on $[a, +\infty)$. First, for $t < a$:

$$\rho(t) = -\frac{\phi'(t)}{\phi(t)} = \frac{1}{\phi(t)} \int_t^a \phi''(s) ds = \int_t^a \frac{\phi(s)}{\phi(t)} k(s) ds$$

As ϕ is nonincreasing, $\frac{\phi(s)}{\phi(t)} \leq 1$ in the above integral and we deduce

$$|\rho(t)| \leq \int_t^a k(s) ds \rightarrow 0 \text{ when } t \rightarrow a$$

and also $\sigma(t) = -\rho(t)^2/2 \rightarrow 0$ when $t \rightarrow a$. The first guess would then consist in extending σ and ρ by continuity and one could set $\sigma = \rho = 0$ on $[a, +\infty)$.

This would actually lead to a discontinuous ρ' . Indeed on one hand ρ is right differentiable in a with right derivative $\rho'(a^+) = 0$. On the other hand, $\rho'(t) = \rho^2(t) - k(t)$ for $t \in (0, a)$ with $\rho(t) \rightarrow 0$ and $k(t) \rightarrow k(a)$ when $t \rightarrow a$, hence¹ ρ is also left differentiable in a but with left derivative $\rho'(a^-) = -k(a)$. Finally ρ' is discontinuous at a (and C^∞ elsewhere), so ρ'' has a dirac mass at a . Whereas $\sigma = -\rho^2/2$ on $(0, a)$ as well as on $[a, +\infty)$ is continuous and has derivative $\sigma' = -\rho'\rho$ also continuous at a as $\rho(a) = 0$. As a consequence $2\sigma' + \rho'' \notin L^2$ and $(\sigma, \rho) \notin S$.

This is why one should not take $\sigma = 0$ but rather $\sigma = -k(a)/2$ on $(a, +\infty)$ and still $\rho = 0$. With this setting, $2\sigma + \rho'$ is continuous at a and the two dirac masses compensate each other so that $2\sigma' + \rho'' \in L^2$. One just needs to check that $2\sigma + \rho^2 \leq 0$,

¹This is an easy result we recall for clarity:

Lemma II.3. *Suppose $f : (a, b) \rightarrow \mathbb{R}$ is a C^1 function such that $f'(t) \rightarrow \ell \in \mathbb{R}$ when $t \rightarrow b$. Then f admits a limit in b , is differentiable in b , and $f'(b) = \ell$.*

Proof. For any $\eta > \varepsilon > 0$, we have according to the mean value theorem that $|f(b - \varepsilon) - f(b - \eta)| = (\eta - \varepsilon)|f'(b - \nu)|$ for some $\nu \in (\varepsilon, \eta)$. As f' converges in b it is bounded, hence $(f(b - \varepsilon))_{\varepsilon \rightarrow 0}$ is Cauchy and converges. We further denote $f(b)$ its limit.

For the differentiability of f in b : fix $\varepsilon > 0$ and take $\delta > 0$ such that $|f'(t) - \ell| \leq \varepsilon$ for any $t \in (b - \delta, b)$. For any $\eta < \delta$ we can write:

$$\left| \frac{f(b) - f(b - \eta)}{\eta} - \ell \right| \leq \frac{1}{\eta} \int_{b-\eta}^b |f'(t) - \ell| dt \leq \frac{1}{\eta} \times \eta\varepsilon = \varepsilon$$

□

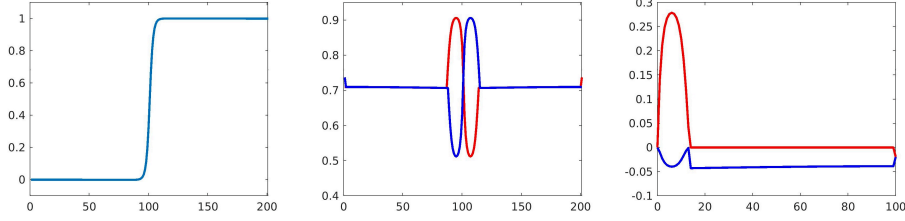


Figure 8.1 – Primal solution u (left), dual solutions p and q (center), corresponding σ (blue) and ρ (red) (right) in the Dirichlet setting with $N = 100$.

that is that $k(a) \geq 0$. This comes again from the fact that $\phi'' = k\phi$: if $k(a) < 0$ then, as $\phi > 0$ on $[0, a)$ and k is nondecreasing, one would obtain that ϕ' is (strictly) decreasing on $[0, a)$. Starting with $\phi'(0) = 0$ we obtain that $\phi'(a) < 0$, which contradicts the regularity of ϕ as $\phi = 0$ on $[a, +\infty)$. Finally, $(\sigma, \rho) \in S$ and, just as before, $2\sigma' + \rho'' = -\psi$ so strong duality holds.

8.4 Return to the discrete problem

Recall we denoted

$$D(\sigma, \rho) = \frac{1}{2} \left\{ \rho'(0) - \frac{1}{2} \int_0^\infty |2\sigma' + \rho''|^2 \right\}$$

$$D^h(\boldsymbol{\sigma}, \boldsymbol{\rho}) = \frac{1}{2} \left\{ \sigma_1 + \frac{1}{\tau} \rho_1 - \frac{\tau}{2} \sum_{n=1}^{N-1} \left(\frac{\sigma_{n-1} - \sigma_{n+1}}{\tau} + \frac{2\rho_n - \rho_{n-1} - \rho_{n+1}}{\tau^2} \right)^2 \right\}$$

the objectives of the continuous and discrete problems respectively (where again $\tau = N^{-1/3}$). The constraints on $\sigma, \rho : \mathbb{R}^+ \rightarrow \mathbb{R}$ and $\boldsymbol{\sigma}, \boldsymbol{\rho} \in \mathbb{R}^{N+1}$ are

$$\sigma(0) = \rho(0) = 0 \text{ and } 2\sigma + \rho^2 \leq 0 \text{ on } \mathbb{R}^+$$

$$\boldsymbol{\sigma}_0 = \boldsymbol{\rho}_0 = 0 \text{ and } \forall n \in \llbracket 1, N \rrbracket, N^{-2/3} \boldsymbol{\sigma}_n^2 + 2\boldsymbol{\sigma}_n + \boldsymbol{\rho}_n^2 \leq 0$$

Given an admissible (σ, ρ) of the continuous problem with $D(\sigma, \rho) > 0$ we chose the following discretization: set $\boldsymbol{\sigma}_0 = 0$ and $\forall n \geq 1, \boldsymbol{\sigma}_n = \sigma(\tau n) - \tau$ as well as $\boldsymbol{\rho}_n = \rho(\tau n)$ for all $n \in \llbracket 0, N \rrbracket$. Then – provided σ is bounded – $(\boldsymbol{\sigma}, \boldsymbol{\rho})$ is indeed admissible in the discrete problem as $\boldsymbol{\sigma}_0 = \boldsymbol{\rho}_0 = 0$ and

$$N^{-2/3} \boldsymbol{\sigma}_n^2 + 2\boldsymbol{\sigma}_n + \boldsymbol{\rho}_n^2 = N^{-2/3} (\sigma(\tau n) - \tau)^2 - 2\tau + 2\sigma(\tau n) + \rho(\tau n)^2$$

$$\leq N^{-2/3} (\sigma(\tau n) - \tau)^2 - 2N^{-1/3}$$

with this quantity being nonpositive as soon as $|\sigma(\tau n) - N^{-1/3}| \leq \sqrt{2}N^{1/6}$ which is true for N sufficiently large when σ is bounded.

Therefore we just need to check that with this discretization $D^h(\sigma, \rho)$ converges to $D(\sigma, \rho)$ when $N \rightarrow \infty$ as expected in the first place. First note that $\sigma_1 = \sigma(\tau) - \tau \rightarrow \sigma(0) = 0$ (as long as σ is continuous) and that $\frac{1}{\tau}\rho_1 = \frac{\rho(\tau) - \rho(0)}{\tau} \rightarrow \rho'(0)$. As a result we focus next on the convergence of the Riemann sum towards the desired integral.

Second, we can in fact take σ to be $\sigma_n = \sigma(\tau n)$. Indeed this only affects the first term of the sum adding:

$$-\frac{\tau}{2} \left| \frac{\sigma(2\tau) - \tau}{\tau} + \frac{\rho(2\tau) - 2\rho(\tau) + \rho(0)}{\tau^2} \right|^2 + \frac{\tau}{2} \left| \frac{\sigma(2\tau) - \sigma(0)}{\tau} + \frac{\rho(2\tau) - 2\rho(\tau) + \rho(0)}{\tau^2} \right|^2$$

with $\frac{\rho(2\tau) - 2\rho(\tau) + \rho(0)}{\tau^2} \rightarrow \rho''(0)$, $\frac{\sigma(2\tau) - \sigma(0)}{\tau} \rightarrow 2\sigma'(0)$ and $\frac{\sigma(2\tau) - \tau}{\tau} \rightarrow 2\sigma'(0) - 1$, so that this quantity tends to zero when $\tau \rightarrow 0$.

To ensure the convergence of the sum, we will need additional regularity on σ and ρ . In the compact support case, we find a new couple (σ, ρ) , more regular and still satisfying $D(\sigma, \rho) > 0$ whereas in the positive case we stick with the (σ, ρ) defined above but show they decrease exponentially fast.

8.4.1 Compact support case

Recall that in this case we have $\sigma, \rho : \mathbb{R}^+ \rightarrow \mathbb{R}$ satisfying $D(\sigma, \rho) > 0$ with $\rho = \sigma' = 0$ on $(a, +\infty)$ and σ, ρ of class \mathcal{C}^∞ on $[0, \infty) \setminus \{a\}$. We extend σ and ρ to \mathbb{R}^- by 0 and regularize them into \mathcal{C}^∞ functions on $[0, \infty)$ while keeping their admissibility in problem (8.2) as well as the compactness of their support and the value of $\rho'(0)$.

To this end, we first regularize by convolution with a function $\eta \in \mathcal{C}_c^\infty(\mathbb{R})$ such that $\eta \geq 0$, $\int_0^\infty \eta = 1$, and $\eta(x) = 0$ for any $x \notin (0, 1)$: we obtain functions $\rho_\varepsilon = \int_{\mathbb{R}} \rho(\cdot + \varepsilon t) \eta(t) dt$ and $\sigma_\varepsilon = \int_{\mathbb{R}} \sigma(\cdot + \varepsilon t) \eta(t) dt$ which are \mathcal{C}^∞ on $[0, \infty)$ and satisfy $\rho_\varepsilon = \sigma'_\varepsilon = 0$ on (a, ∞) as well as $2\sigma_\varepsilon + \rho_\varepsilon^2 \leq 0$ since this constraint is convex, that is $C = \{(s, r) \in \mathbb{R}^2 : 2s + r^2 \leq 0\}$ is a convex set.

However, we lost the values of $\rho(0), \sigma(0)$ and more importantly of $\rho'(0)$ which appears in problem (8.2). To this end, take $\nu \in \mathcal{C}^\infty$ a plateau function such that $\nu = 1$ on $(-\infty, \frac{a}{3})$ and $\nu = 0$ on $(\frac{2a}{3}, +\infty)$, and set $\hat{\sigma}_\varepsilon = \nu\sigma + (1 - \nu)\sigma_\varepsilon$, $\hat{\rho}_\varepsilon = \nu\rho + (1 - \nu)\rho_\varepsilon$. As σ and ρ are \mathcal{C}^∞ on $[0, +\infty)$ except in a which is avoided, $\hat{\sigma}_\varepsilon$ and $\hat{\rho}_\varepsilon$ are \mathcal{C}^∞ on $[0, +\infty)$, and as $\hat{\rho}_\varepsilon = \rho$, $\hat{\sigma}_\varepsilon = \sigma$ near 0 we keep $\hat{\sigma}_\varepsilon(0) = \hat{\rho}_\varepsilon(0) = 0$ and $\hat{\rho}'_\varepsilon(0) = \rho'(0)$. Furthermore, the constraint $2\hat{\sigma}_\varepsilon + \hat{\rho}_\varepsilon^2 \leq 0$ is still fulfilled by convex combination. Finally one checks that:

$$2\hat{\sigma}'_\varepsilon + \hat{\rho}''_\varepsilon = 2\sigma'_\varepsilon + \rho''_\varepsilon + \{2(\sigma' - \sigma'_\varepsilon) + (\rho'' - \rho''_\varepsilon)\}\nu + \{(\sigma - \sigma_\varepsilon) + 2(\rho - \rho_\varepsilon)\}\nu' + \{\rho - \rho_\varepsilon\}\nu''$$

so that when ε goes to 0:

- $2\sigma'_\varepsilon + \rho''_\varepsilon$ converges to $2\sigma' + \rho''$ in $L^2(0, \infty)$.
- σ', ρ'' are continuous on $[0, \frac{2a}{3}]$ hence $2(\sigma' - \sigma'_\varepsilon) + (\rho'' - \rho''_\varepsilon)$ converges to 0 uniformly on $[0, \frac{2a}{3}]$. As $\nu = 0$ on $(\frac{2a}{3}, +\infty)$ this implies that $\{2(\sigma' - \sigma'_\varepsilon) + (\rho'' - \rho''_\varepsilon)\}\nu$ converges to 0 in $L^2(0, \infty)$.

- $\nu' = \nu'' = 0$ outside of $[\frac{a}{3}, \frac{2a}{3}]$ where σ, σ' and ρ' are continuous hence $\{(\sigma - \sigma_\varepsilon) + 2(\rho' - \rho'_\varepsilon)\}\nu' + \{\rho - \rho_\varepsilon\}\nu''$ converges to 0 uniformly hence in $L^2(0, \infty)$.

To conclude, $D(\hat{\sigma}_\varepsilon, \hat{\rho}_\varepsilon) \rightarrow D(\sigma, \rho)$. This shows that one can find (σ, ρ) admissible in the continuous problem such that $D(\sigma, \rho) > 0$ and σ, ρ are C^∞ on $[0, +\infty)$, with ρ and σ' having compact supports. In particular all the functions $\sigma, \sigma', \sigma'', \rho, \rho', \rho''$ and ρ''' can be uniformly bounded by some constant $M > 0$.

Then to estimate convergence of $D^h(\sigma, \rho)$ towards $D(\sigma, \rho)$ we can truncate the Riemann sum at $n = \lfloor \frac{a}{\tau} \rfloor$ where the supports of σ' and ρ are included in $[0, a]$. Doing so it is easy to show that

$$\tau \sum_{n=1}^{N-1} \left| \frac{\sigma_{n+1} - \sigma_{n-1}}{\tau} + \frac{\rho_{n+1} - 2\rho_n + \rho_{n-1}}{\tau^2} \right|^2 = \tau \sum_{n=1}^{\lfloor \frac{a}{\tau} \rfloor} |2\sigma'(\tau n) + \rho''(\tau n)|^2 + O(\tau)$$

And we conclude saying that as $(2\sigma' + \rho'')^2$ is Riemann integrable one has

$$\tau \sum_{n=1}^{\lfloor \frac{a}{\tau} \rfloor} |2\sigma'(\tau n) + \rho''(\tau n)|^2 \rightarrow \int_0^a (2\sigma' + \rho'')^2 = \int_0^\infty (2\sigma' + \rho'')^2$$

hence the desired convergence.

8.4.2 Positive case

Recall that in this case we have $\sigma, \rho : \mathbb{R}^+ \rightarrow \mathbb{R}$ satisfying $D(\sigma, \rho) > 0$ with $\sigma = -\rho^2/2$ and $\rho = -\phi'/\phi$ for some $\phi > 0$ which is C^∞ on \mathbb{R}^+ . We also had that $\phi' \leq 0$ and $\phi'' = k\phi$ for some function k such that $k' \geq 0$. Therefore ρ satisfies on \mathbb{R}^+ :

$$\rho' = -\frac{\phi''}{\phi} + \frac{\phi'^2}{\phi^2} = \rho^2 - k$$

This relation allows us to show that the derivatives of ρ tends to 0 exponentially fast, which will compensate the non compactness of their support. It is important to note that the key argument in the following proofs is that this relation holds on the whole \mathbb{R}^+ : in the case of compact support it only holds on $[0, a)$ and one cannot obtain the same conclusions (especially, in the compact support case, we cannot have $\rho'(t) \geq 0$ for all $t \geq 0$ as shown below). Our analysis begins with the two following lemmas that derive from easy manipulations and antidifferentiation.

Lemma II.4. *Let $\rho, k : \mathbb{R}^+ \rightarrow \mathbb{R}$ be C^1 functions such that for all $t \geq 0$, $\rho'(t) = \rho^2(t) - k(t)$, $\rho(t) \geq 0$ and $k'(t) \geq 0$. Then for all $t \geq 0$, $\rho'(t) \geq 0$.*

Proof. Suppose $\rho'(t) = -r < 0$ for some $t \geq 0$, then thanks to the hypotheses $\rho(t), k'(t) \geq 0$ we obtain that $\rho''(t) \leq 0$. From this reasoning it is easy to prove that ρ' will remain nonpositive on (t, ∞) . But then again, as $\rho, k' \geq 0$, this implies that $\rho' = \rho^2 - k$ is nonincreasing. Consequently, $\rho'(s) \leq -r$ for any $s \geq t$ and we obtain $\rho(s) \leq \rho(t) - r(s - t)$ which cannot stand with the hypothesis that $\rho \geq 0$. \square

Lemma II.5. *Let $t_1 \in \mathbb{R}$ and let $\rho : [t_1, +\infty[\rightarrow \mathbb{R}^+$ be a C^1 function. There is no $L \in \mathbb{R}$ such that $\forall t \geq t_1$*

$$\rho^2(t) - L \neq 0 \text{ and } \frac{\rho'(t)}{\rho^2(t) - L} \geq 1$$

Proof. Depending, on the sign of L , one integrates $\frac{\rho'}{\rho^2 - L}$ into: $-\frac{1}{\rho}$ if $L = 0$, $\log\left(\frac{\rho - \sqrt{L}}{\rho + \sqrt{L}}\right)$ if $L > 0$, or $\arctan\left(\frac{\rho}{\sqrt{-L}}\right)$ if $L < 0$. In either cases, taking the limit at infinity leads to a contradiction. \square

Thanks to the first lemma, ρ is nonnegative and nondecreasing (and not zero everywhere), so $\rho(t) \rightarrow R \in (0, +\infty]$ when $t \rightarrow \infty$. In particular there exist $c > 0$ and $t_0 > 0$ such that $\forall t \geq t_0$, $-\frac{\phi'(t)}{\phi(t)} = \rho(t) \geq c > 0$ which leads to $\phi(t) \leq \phi(t_0) \exp(-c(t - t_0))$. As a consequence, we deduce that the function k is bounded. To do so, remember that $k(t) = \int_0^t \psi(u) du - A$ with $\psi(u) = \int_u^\infty \phi^2(s) ds$ hence $k(t) = \int_0^t \int_u^\infty \phi^2(s) ds du - A = \int_0^\infty \min(s, t) \phi^2(s) ds - A$. Finally the exponential bound on ϕ shows that k is bounded, and since it is increasing it converges to some $L \in \mathbb{R}$. In addition the convergence is exponential since :

$$L - k(t) = \int_t^\infty (s - t) \phi^2(s) ds \leq M \exp(-2ct) \text{ for some } M > 0$$

Next we must have $R < +\infty$ and more precisely $R^2 \leq L$. Indeed, otherwise we would have a $t_1 > 0$ such that $\forall t \geq t_1$, $\rho'(t) = \rho^2(t) - k(t) \geq \rho^2(t) - L > 0$ hence $\frac{\rho'(t)}{\rho(t)^2 - L} \geq 1$ which is impossible according to the second lemma. To finish with, as $\rho' = \rho^2 - k$ stays nonnegative and converges to $R^2 - L$, $R^2 = L$ and finally:

$$\forall t \geq 0, \rho'(t) = \rho^2(t) - L + L - k(t) \leq L - k(t) \leq M \exp(-2ct)$$

As a consequence, $\sigma', \rho'', \sigma''$ and ρ''' decrease exponentially to zero. Indeed:

- $\sigma' = -\rho' \rho$ with ρ bounded.
- $\rho'' = 2\sigma' \sigma - \psi$ with $\sigma = -\rho^2/2$ bounded and ψ decreasing exponentially to zero (as $\psi(t) = \int_t^\infty \phi^2$ with ϕ decreasing exponentially).
- $\sigma'' = -\rho'^2 - \rho'' \rho$.
- $\rho''' = 2\rho'' \rho + 2\rho'^2 + \phi^2$.

Then we get the following estimate for our discretization: write for $n \in \llbracket 1, N - 1 \rrbracket$

$$\frac{\sigma_{n+1} - \sigma_{n-1}}{\tau} = 2\sigma'(\tau n + \eta_n) \text{ and } \frac{\rho_{n+1} - 2\rho_n + \rho_{n-1}}{\tau^2} = \rho''(\tau n + \tilde{\eta}_n)$$

for some $\eta_n, \tilde{\eta}_n \in (-\tau, \tau)$, so that we have:

$$\left| \left| \frac{\sigma_{n+1} - \sigma_{n-1}}{\tau} + \frac{\rho_{n+1} - 2\rho_n + \rho_{n-1}}{\tau^2} \right|^2 - |2\sigma'(\tau n) + \rho''(\tau n)|^2 \right| = \Delta_n^- \times \Delta_n^+$$

with

$$\begin{aligned}\Delta_n^- &:= |2\sigma'(\tau n + \eta_n) - 2\sigma'(\tau n) + \rho''(\tau n + \tilde{\eta}_n) - \rho''(\tau n)| \\ &\leq 2\tau \times (2\|\sigma''\|_{\infty,(\tau n-\tau,\tau n+\tau)} + \|\rho'''\|_{\infty,(\tau n-\tau,\tau n+\tau)}) \\ &\leq \tau M \exp(-c(\tau n - \tau)) \\ \Delta_n^+ &:= |2\sigma'(\tau n + \eta_n) + 2\sigma'(\tau n) + \rho''(\tau n + \tilde{\eta}_n) + \rho''(\tau n)| \\ &\leq 4\|\sigma'\|_{\infty,(\tau n-\tau,\tau n+\tau)} + 2\|\rho''\|_{\infty,(\tau n-\tau,\tau n+\tau)} \\ &\leq M \exp(-c(\tau n - \tau))\end{aligned}$$

for some constants $M, c > 0$ and finally one can write (for other constants $M, c > 0$):

$$\begin{aligned}&\left| \tau \sum_{n=1}^{N-1} \left| \frac{\sigma_{n+1} - \sigma_{n-1}}{\tau} + \frac{\rho_{n+1} - 2\rho_n + \rho_{n-1}}{\tau^2} \right|^2 - \tau \sum_{n=1}^{N-1} |2\sigma'(\tau n) - \rho''(\tau n)|^2 \right| \\ &\leq \tau^2 \sum_{n=1}^{N-1} M \exp(-c(\tau n - \tau)) \\ &\leq M\tau^2 \sum_{n=0}^{\infty} \exp(-c\tau)^n = M \frac{\tau^2}{1 - \exp(-c\tau)} \sim M \frac{\tau^2}{c\tau} \rightarrow 0 \text{ as } N \rightarrow \infty\end{aligned}$$

To conclude, that is to obtain $D^h(\sigma, \rho) \rightarrow D(\sigma, \rho)$, we state that

$$\tau \sum_{n=1}^{N-1} (2\sigma'(\tau n) + \rho''(\tau n))^2 \rightarrow \int_0^\infty (2\sigma' + \rho'')^2 \text{ as } N \rightarrow \infty$$

This comes from taking $f = (2\sigma' + \rho'')^2 = \psi^2$ – which is indeed nonincreasing as $\psi' = -\phi^2 \leq 0$ and $\psi \geq 0$ – in the following result:

Proposition II.6. *Let $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ be a continuous and nonincreasing function such that $\int_0^\infty f$ converges. Let $a > b > 0$ and $c_1, c_2, c_3 \in \mathbb{R}$ constants. Then*

$$S_N := \frac{1}{N^b} \sum_{l=[c_1]}^{[c_2 N^a + c_3]} f\left(\frac{l}{N^b}\right) \rightarrow \int_0^\infty f \text{ when } N \rightarrow \infty$$

Proof. To simplify, assume $c_1 = c_3 = 0, c_2 = 1$ and a, b are integers. Let $\varepsilon > 0$ and let $M \in \mathbb{N}^*$ such that $\forall x, y \geq M - 1, \left| \int_x^y f \right| \leq \varepsilon$. As f tends to zero at infinity it is uniformly continuous so one can find $\delta > 0$ such that $|x - y| \leq \delta \Rightarrow |f(x) - f(y)| \leq \frac{\varepsilon}{M}$. For N such that $N^a > MN^b$ and $N^b > \frac{1}{\delta}$, write $S_N = S_N^1 + S_N^2$ with

$$S_N^1 = \frac{1}{N^b} \sum_0^{MN^b-1} f\left(\frac{l}{N^b}\right) \text{ and } S_N^2 = \frac{1}{N^b} \sum_{MN^b}^{N^a} f\left(\frac{l}{N^b}\right)$$

Then on one hand, using that $|\frac{l}{N^b} - t| < \delta$ we have:

$$\left| S_N^1 - \int_0^M f \right| = \left| \sum_0^{MN^b-1} \int_{\frac{l}{N^b}}^{\frac{l+1}{N^b}} f\left(\frac{l}{N^b}\right) - f(t) dt \right| \leq MN^b \times \frac{1}{N^b} \times \frac{\varepsilon}{M} = \varepsilon$$

and on the other hand, writing $S_N^2 = \sum_{MN^b}^{N^a} \int_{\frac{1}{N^b}}^{\frac{1+1}{N^b}} f(\frac{1}{N^b}) = \sum_{MN^{b-1}}^{N^a-1} \int_{\frac{1}{N^b}}^{\frac{1+1}{N^b}} f(\frac{1+1}{N^b})$ and using that f is nonincreasing we get:

$$\int_{M-\frac{1}{N^b}}^{N^{a-b}} f \geq S_N^2 \geq \int_M^{N^{a-b}+\frac{1}{N^b}} f$$

with these two integrals bounded by ε in absolute value by choice of M and N . Finally,

$$\left| S_N - \int_0^\infty f \right| \leq \left| S_N^1 - \int_0^M f \right| + |S_N^2| + \left| \int_M^\infty f \right| \leq 3\varepsilon$$

which completes the proof. \square

8.5 Modifications for Neumann boundary conditions

Dealing with Neumann boundary conditions takes us back to the 1D problem (6.1), where we now take $u_{N+1} = u_N$ and $u_{-N-1} = u_{-N}$. We also suppose $\lambda < \lambda^*$ so that $\bar{u} \neq \frac{1}{2}$. Thanks to Proposition II.3, we can suppose $p_0 = q_0 = \sqrt{2}/2$ in the dual problem (8.1), and one checks that it is changed into

$$\bar{E}^h = \max_{\substack{p_n^2 + p_{-n}^2 \leq 1 \\ -N \leq n \leq N \\ p_0 = \sqrt{2}/2}} \frac{1}{2} + \frac{1}{\sqrt{2}} p_1 - \frac{\lambda}{h\sqrt{2}} \sum_{n=1}^{N-1} (p_{-n+1} - p_{-n} + p_n - p_{n+1})^2 - \frac{\lambda}{h\sqrt{2}} (p_{-N+1} + p_N)^2$$

Remember from section 6.2 that the limit value when $h = \frac{1}{N} \rightarrow 0$ is $\bar{E} = \bar{E}_N = 1 - \sqrt{2}\lambda$. This value is (almost) achieved when taking $p_n = \sqrt{2}/2 - |n|/\sqrt{2}N$ as it gives $\bar{E}^h \geq 1 - \sqrt{2}\lambda + \frac{3\lambda - \sqrt{2}}{2\sqrt{2}}h$ (but $3\lambda - \sqrt{2} < 0$). Let us denote

$$F(p, \lambda) = \frac{1}{2} + \frac{1}{\sqrt{2}} p_1 - \frac{\lambda}{\sqrt{2}h} \sum_{n=1}^{N-1} (p_{-n+1} - p_{-n} + p_n - p_{n+1})^2$$

$$\tilde{F}(\tilde{p}, \lambda) = \frac{1}{2} + \frac{1}{\sqrt{2}} \tilde{p}_1 - \frac{\lambda}{\sqrt{2}h} \sum_{n=1}^{N-1} (\tilde{p}_{-n+1} - \tilde{p}_{-n} + \tilde{p}_n - \tilde{p}_{n+1})^2 - \frac{\lambda}{\sqrt{2}h} (\tilde{p}_{-N+1} + \tilde{p}_N)^2$$

Note that the constraint on p in Dirichlet and Neumann problems is the same: $p_0 = \sqrt{2}/2$ and $p_n^2 + p_{-n}^2 \leq 1$. Now suppose p is the Dirichlet variable constructed in the previous sections, and form $\tilde{p}_n = p_n - \frac{|n|}{\sqrt{2}N}$. We want to compare $\tilde{F}(\tilde{p}, \lambda) - \bar{E}_N$ to $F(p, \lambda) - \bar{E}_D$. As $\bar{E}_N = 1 - \lambda\sqrt{2} = \bar{E}_D - \lambda\sqrt{2}$, we split $\lambda\sqrt{2}$ into $N \times \frac{\lambda}{\sqrt{2}h} \times \frac{2}{N^2}$ and allocate each $\frac{2}{N^2}$ to a term involving p^2 in the expression of \tilde{E} . This will make appear:

$$\begin{aligned} (\tilde{p}_{-n+1} - \tilde{p}_{-n} + \tilde{p}_n - \tilde{p}_{n+1})^2 - \frac{2}{N^2} &= (p_{-n+1} - p_{-n} + p_n - p_{n+1} + \frac{\sqrt{2}}{N})^2 - \frac{2}{N^2} \\ &= x_n^2 + \frac{2\sqrt{2}}{N} x_n \end{aligned}$$

where we denoted $x_n = p_{-n+1} - p_{-n} + p_n - p_{n+1}$. When summing, we will recover the term in x_n^2 appearing in $E(p, \lambda)$, along with

$$\sum_{n=1}^{N-1} x_n = p_1 - p_N + p_0 - p_{-N+1} = \left(p_1 - \frac{\sqrt{2}}{2}\right) - (p_N + p_{N-1} - \sqrt{2})$$

Besides, one has

$$(\tilde{p}_{-N+1} + \tilde{p}_N)^2 - \frac{2}{N^2} = (p_{-N+1} + p_N - \sqrt{2})^2 + \frac{\sqrt{2}}{N}(p_{-N+1} + p_N - \sqrt{2}) - \frac{3}{N^2}$$

Then we obtain:

$$\begin{aligned} \tilde{F}(\tilde{p}, \lambda) - \bar{E}_N &= \frac{1}{2} + \frac{1}{\sqrt{2}}p_1 - \frac{1}{2N} - 1 - \frac{\lambda}{\sqrt{2}h} \sum_{n=1}^{N-1} x_n^2 \\ &\quad - \frac{\lambda}{\sqrt{2}h} \times \frac{2\sqrt{2}}{N} \left(p_1 - \frac{\sqrt{2}}{2}\right) - \frac{\lambda}{\sqrt{2}h} (p_{-N+1} + p_N - \sqrt{2})^2 \\ &\quad - \frac{\lambda}{\sqrt{2}h} \times \frac{\sqrt{2}}{N} (p_{-N+1} + p_N - \sqrt{2}) + \frac{\lambda}{\sqrt{2}h} \frac{3}{N^2} \\ &= F(p, \lambda) - \bar{E}_D - 2\lambda \left(p_1 - \frac{\sqrt{2}}{2}\right) + R \end{aligned} \quad (8.7)$$

where $R = \lambda(p_N + p_{-N+1} - \sqrt{2}) - \frac{\lambda}{\sqrt{2}h} (p_N + p_{-N+1} - \sqrt{2})^2 + \frac{3\sqrt{2}\lambda-1}{2N}$.

At this point, remember p was obtained from continuous functions σ and ρ through

$$\begin{cases} p_n = \frac{1}{\sqrt{2}}(\sigma_n + 1 + \rho_n); & p_{-n} = \frac{1}{\sqrt{2}}(\sigma_n + 1 - \rho_n) \\ \text{with } \sigma_n = N^{-2/3}(\sigma(\tau n) - \tau); & \rho_n = N^{-1/3}\rho(\tau n) \end{cases}$$

As ρ and σ are bounded, one sees that p_n converges to $\frac{\sqrt{2}}{2}$ uniformly as N goes to infinity (that is $\max_{-N \leq n \leq N} |p_n - \frac{\sqrt{2}}{2}| \rightarrow 0$ as $N \rightarrow \infty$). This first shows that \tilde{p} is admissible in the dual problem (meaning that $\tilde{p}_n^2 + \tilde{p}_{-n}^2 \leq 1$): indeed p is itself admissible and $p_n \geq \tilde{p}_n \geq -1 \geq -p_n$ for N sufficiently large. Second, remember that, at infinity, σ converges to $-k(a) < 0$ or to 0, and ρ converges to 0. Writing

$$p_{-N+1} + p_N - \sqrt{2} = \frac{1}{\sqrt{2}}(\sigma_N + \sigma_{N-1} + \rho_N - \rho_{N-1})$$

one sees that $N^{2/3}R \rightarrow 0$ when $N \rightarrow \infty$. Then we apply a last trick to include $2\lambda(p_1 - \frac{\sqrt{2}}{2})$ from (8.7) into our energies: we remark that

$$F(p, \lambda) - \bar{E}_D - 2\lambda \left(p_1 - \frac{\sqrt{2}}{2}\right) = (1 - 2\sqrt{2}\lambda) \left(F\left(p, \frac{\lambda}{1 - 2\sqrt{2}\lambda}\right) - \bar{E}_D\right)$$

This finally shows that

$$N^{2/3} \left(\tilde{F}(\tilde{p}, \lambda) - \bar{E}_N\right) = N^{2/3} \left((1 - 2\sqrt{2}\lambda) \left(F\left(p, \frac{\lambda}{1 - 2\sqrt{2}\lambda}\right) - \bar{E}_D\right)\right) + N^{2/3}R$$

converges to a positive value when N tends to infinity; hence the $O(h^{2/3})$ rate is also true in the Neumann setting.

Remark II.1. Although it gives the correct estimate, note that the variable \tilde{p} obtained in the way we described from the optimal p of the Dirichlet setting is probably not optimal in the Neumann setting, see the numerical experiments 8.2. The ideal situation would be to understand the continuous limit of the Neumann problem as we did in the Dirichlet setting. But doing the natural changes of variables $\sigma_n = \frac{1}{\sqrt{2}}(p_n + p_{-n}) - 1 + \frac{n}{2N}$; $\rho_n = \frac{1}{\sqrt{2}}(p_n - p_{-n})$ followed by $\tilde{\sigma}_n = N^{2/3}\sigma_n$; $\tilde{\rho}_n = N^{1/3}\rho_n$ leads to expressing $N^{2/3}(E^h(u^h) - E(\bar{u}))$ as the supremum of $(\frac{1}{\sqrt{2}} \times)$

$$\begin{aligned} & (\lambda - \lambda^2)\left(\tilde{\sigma}_1 + \frac{\tilde{\rho}_1}{\tau} - \tau/2\right) + \lambda^2\left(\tilde{\sigma}_{N-1} + \tilde{\sigma}_N + \frac{\tilde{\rho}_N - \tilde{\rho}_{N-1}}{\tau}\right) \\ & - \frac{\lambda^2}{2}\tau \sum_1^{N-1} \left(\frac{\tilde{\sigma}_{n-1} - \tilde{\sigma}_{n+1}}{\tau} + \frac{-\tilde{\rho}_{n-1} + 2\tilde{\rho}_n - \tilde{\rho}_{n+1}}{\tau^2} \right)^2 \\ & - \frac{\lambda^2}{2}\tau \left(\frac{\tilde{\sigma}_{N-1} - \tilde{\sigma}_N}{\sqrt{\tau}} + \frac{\tilde{\rho}_N - \tilde{\rho}_{N-1}}{\tau\sqrt{\tau}} + \frac{\sqrt{\tau}}{2} \right)^2 \end{aligned}$$

with still $\tau = N^{-1/3}$ and under the constraint

$$\tau^2\left(\tilde{\sigma}_n - \frac{1}{2}n\tau\right)^2 + 2\left(\tilde{\sigma}_n - \frac{1}{2}n\tau\right) + \tilde{\rho}_n^2 \leq 0$$

When $N \rightarrow \infty$, this constraint becomes $2\sigma(t) + \rho^2(t) \leq t$ (recall that $n\tau$ corresponds to t); however the energy at stake remains unclear to us.

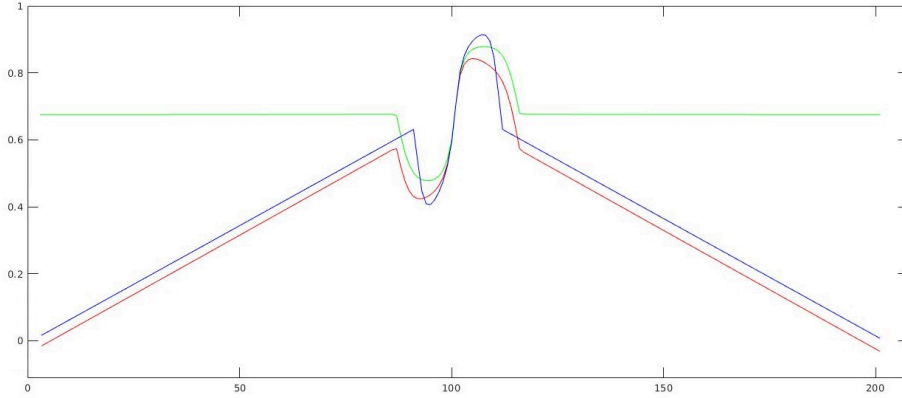


Figure 8.2 – The variable \tilde{p} (in red) obtained from the optimal value p of the Dirichlet setting (in green) does *not* match the optimal value of the Neumann problem (in blue)

**THE ROF MODEL WITH RAVIART-THOMAS
TOTAL VARIATION**

In this chapter, we present and analyze another choice of discrete total variation that is based on the space of Raviart-Thomas fields. These fields have already been used in the context of discretization of the ROF problem, for instance in [DJS07, DJS12, HHS⁺18]. We give a convergence analysis and obtain a $O(h)$ error bound on the energy under standard hypothesis.

9.1 Definitions

The idea behind the definition of the isotropic total variation is of course to catch the L^1 norm of the gradient of u based on a discretization of the expression $\text{TV}(u) = \int_{\Omega} |\nabla u|$. To do so, one chooses a finite differences operator D , defined on the mesh $\Omega = \cup C_{i,j}$ introduced in section 5.2, and designed to approximate ∇ . However, the non isotropy of the grid itself prevents D from being isotropic, as it has to involve a notion of neighbor on this two-directional grid. On the contrary, the dual definition of TV offers the possibility to discretize a field rather than an operator. In the formulas

$$\begin{aligned} \text{TV}_N(u) &= \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi, \phi \in \mathcal{C}_c^1(\Omega, \mathbb{R}^2) \text{ s.t. } \|\phi\|_{\infty} \leq 1 \right\} \\ \text{TV}_D(u) &= \sup \left\{ - \int_{\Omega} u \operatorname{div} \phi + \int_{\partial\Omega} \langle \phi, \vec{n} \rangle, \phi \in \mathcal{C}^1(\Omega, \mathbb{R}^2) \text{ s.t. } \|\phi\|_{\infty} \leq 1 \right\} \end{aligned} \quad (9.1)$$

we will keep the exact operator div but replace the spaces $\mathcal{C}_c^1(\Omega, \mathbb{R}^2)$ and $\mathcal{C}^1(\Omega, \mathbb{R}^2)$ of (compactly supported) \mathcal{C}^1 fields from Ω to \mathbb{R}^2 by a space of discrete fields $RT0$ favouring no direction in the sense that it contains any constant field $\phi \equiv \nu \in \mathbb{R}^2$. This

will indeed bring more isotropy. Indeed, let us define for either continuous or discrete function and boundary term u, b :

$$\mathrm{TV}_{RT,D}^h(u) = \sup \left\{ -\int_{\Omega} u \operatorname{div} \phi + \int_{\partial\Omega} b \langle \phi | \vec{n} \rangle, \phi \in RT0 \text{ s.t. } \|\phi\|_{\infty} \leq 1 \right\} \quad (9.2)$$

Then provided one can approximate $RT0$ fields by C^1 fields (which will be the case, see Lemma II.6), one will have for any u , $\mathrm{TV}_{RT,D}^h(u) \leq \mathrm{TV}(u)$. And for $u = g_{\nu}$, the fact that $\nu \in RT0$ finally gives $\mathrm{TV}_{RT,D}^h(g_{\nu}) = \mathrm{TV}(g_{\nu}) = \int_{\partial\Omega} g_{\nu} \langle \nu | \vec{n} \rangle$. As a consequence the same reasoning than in the continuous case section 5.1 shows that for $b = g_{\nu}$ one has $\bar{u} = g_{\nu}$ for any ν in the following ROF model (mixing a discrete TV term to a continuous L^2 term):

$$\bar{u} = \arg \min_{\substack{u \in BV \cap L^2(\Omega) \\ u|_{\partial\Omega} = g_{\nu}|_{\partial\Omega}}} \frac{1}{2\lambda} \|u - g_{\nu}\|_{L^2}^2 + \mathrm{TV}_{RT,D}^h(u)$$

The most simple space one can think of to play the role of $RT0$ is the space of piecewise affine functions $\phi : \Omega \rightarrow \mathbb{R}^2$. Denoting $p_{i,j} = (x_{i,j}, y_{i,j})$ the center of the square $C_{i,j}$, these fields are given by

$$\forall (x, y) \in C_{i,j}, \phi(x, y) = \begin{pmatrix} a_{i,j}^1(x - x_{i,j}) + b_{i,j}^1(y - y_{i,j}) + c_{i,j}^1 \\ a_{i,j}^2(x - x_{i,j}) + b_{i,j}^2(y - y_{i,j}) + c_{i,j}^2 \end{pmatrix}$$

for any real numbers $a_{i,j}^k, b_{i,j}^k, c_{i,j}^k, k \in \{1, 2\}$. In fact, it is interesting to lower a bit the dimension of our space. First, as only $\operatorname{div} \phi$ is involved one can take $b_{i,j}^1 = a_{i,j}^2 = 0$. Second, to ensure the identity $-\int_{\Omega} u \operatorname{div} \phi = \int_{\Omega} \langle \nabla u | \phi \rangle$ holds for regular u , one requires that the boundary terms $\int_{\partial C_{i,j}} u \langle \phi | \nu \rangle$ cancel each other. Imposing this cancellation for any u and any ν gives, for the horizontal cancellation between cells (i, j) and $(i+1, j)$, the equations:

$$\forall i, j \in \llbracket 1, N-1 \rrbracket, a_{i,j}^1 \frac{h}{2} + c_{i,j}^1 = a_{i+1,j}^1 \frac{-h}{2} + c_{i+1,j}^1 =: f_{i+\frac{1}{2},j}$$

We get similar equations for the vertical cancellation. In addition, in the Neumann setting we also require the cancellation of this flux at the boundary of the domain: $f_{N+\frac{1}{2},j} = a_{N,j}^1 \frac{h}{2} + c_{N,j}^1 = 0$, $f_{\frac{1}{2},j} = a_{1,j}^1 \frac{-h}{2} + c_{1,j}^1 = 0$. Finally, all these equations on a, b, c lead to the definition of the lowest order Raviart-Thomas fields as presented in [RT77]. These fields are defined via their fluxes through the edges of the squares. We will denote $f_{i+\frac{1}{2},j}$ (resp. $f_{i,j+\frac{1}{2}}$) the flux through the edge between the squares $C_{i,j}$ and $C_{i+1,j}$ (resp. $C_{i,j}$ and $C_{i,j+1}$), and remember that $(x_{i,j}, y_{i,j})$ denotes the center of the square $C_{i,j}$. Then the Raviart-Thomas fields are the elements of

$$RT0 = \left\{ \phi : \Omega \rightarrow \mathbb{R}^2 \text{ s.t. } \exists (f_{i+\frac{1}{2},j}, f_{i,j+\frac{1}{2}})_{i,j} \text{ s.t. } \forall i, j \in \llbracket 1, N \rrbracket, \right. \\ \left. \phi(x, y) = \begin{pmatrix} \frac{f_{i+\frac{1}{2},j} + f_{i-\frac{1}{2},j}}{2} + (f_{i+\frac{1}{2},j} - f_{i-\frac{1}{2},j}) \frac{x - x_{i,j}}{h} \\ \frac{f_{i,j+\frac{1}{2}} + f_{i,j-\frac{1}{2}}}{2} + (f_{i,j+\frac{1}{2}} - f_{i,j-\frac{1}{2}}) \frac{y - y_{i,j}}{h} \end{pmatrix} \text{ in } C_{i,j} \right\} \quad (9.3)$$

In the sequel, we will write $\phi = \phi_f \in RT0$ to precise that f denotes the fluxes of the Raviart-Thomas field ϕ according to (9.3). In the Neumann setting, we add the condition that these fields vanish on the boundary of Ω , and we denote:

$$RT0_0 = \left\{ \phi_f \in RT0 \text{ s.t. } \forall i, j \in \llbracket 1, N \rrbracket, f_{\frac{1}{2},j} = f_{N+\frac{1}{2},j} = f_{i,\frac{1}{2}} = f_{i,N+\frac{1}{2}} = 0 \right\}$$

Finally, the definitions of the discrete Raviart-Thomas total variation that we will use are the following. In the Neumann setting, we define for any $u^h \in P0$:

$$\text{TV}_{RT,N}^h(u^h) = \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi, \phi = \phi_f \in RT0_0 \text{ s.t. } \|\phi\|_{\infty} \leq 1 \right\} \quad (9.4)$$

while in the Dirichlet setting, and for any discrete boundary term b^h , we define:

$$\text{TV}_{RT,D}^h(u^h) = \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi + \int_{\partial\Omega} b^h \langle \phi | \vec{n} \rangle, \phi = \phi_f \in RT0 \text{ s.t. } \|\phi\|_{\infty} \leq 1 \right\} \quad (9.5)$$

We stress the fact that no discontinuity jump appears in the calculus of $\operatorname{div} \phi_f$ so that, for instance in the Neumann setting, for $\phi_f \in RT0_0$:

$$\begin{aligned} - \int_{\Omega} u^h \operatorname{div} \phi_f &= - \sum_{i,j} h^2 u_{i,j}^h \frac{1}{h} (f_{i+\frac{1}{2},j} - f_{i-\frac{1}{2},j} + f_{i,j+\frac{1}{2}} - f_{i,j-\frac{1}{2}}) \\ &= h \sum_{i,j} f_{i+\frac{1}{2},j} (u_{i+1,j}^h - u_{i,j}^h) + h \sum_{i,j} f_{i,j+\frac{1}{2}} (u_{i,j+1}^h - u_{i,j}^h) \\ &= h \sum_{i,j} \left\langle \begin{pmatrix} f_{i+\frac{1}{2},j} \\ f_{i,j+\frac{1}{2}} \end{pmatrix} \middle| \begin{pmatrix} (u^h)_{i+1,j} - (u^h)_{i,j} \\ (u^h)_{i,j+1} - (u^h)_{i,j} \end{pmatrix} \right\rangle = h \langle f | D u^h \rangle \end{aligned}$$

for the classical finite difference operator D used previously in section 5.2. In particular, as noted by Lee and co-authors in [LPP19], the isotropic total variation can be recovered in the context of Raviart-Thomas fields total variation as

$$\text{TV}_i^h(u^h) = \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi, \phi = \phi_f \in RT0_0 \text{ s.t. } \forall i, j, \left| \begin{pmatrix} f_{i+\frac{1}{2},j} \\ f_{i,j+\frac{1}{2}} \end{pmatrix} \right| \leq 1 \right\}$$

It has been noticed by Condat [Con17] that the fact that the variables $u_{i,j}^h$ and $f_{i+\frac{1}{2},j}, f_{i,j+\frac{1}{2}}$ lie on different grids is an obstruction to the isotropy of TV_i^h . Following Hintermüller and co-authors [HRH14], he proposes to fix this offset by interpolating the two grids at stake through constraints more complicated than $|(f_{i+\frac{1}{2},j}, f_{i,j+\frac{1}{2}})| \leq 1$ and defines:

$$\text{TV}_{\text{Condat}}^h(u_h) = \sup \left\{ - \int_{\Omega} u_h \operatorname{div} \phi_f, \phi_f \in RT0_0 \text{ s.t. } \forall i, j \in \llbracket 1, N \rrbracket, \right. \\ \left. \max(|(L_{\bullet} \phi_f)_{i,j}|, |(L_{\leftrightarrow} \phi_f)_{i,j}|, |(L_{\dagger} \phi_f)_{i,j}|) \leq 1 \right\}$$

where

$$\begin{cases} (L_{\bullet}\phi_f)_{i,j} = \frac{1}{2} \begin{pmatrix} f_{i+\frac{1}{2},j} + f_{i-\frac{1}{2},j} \\ f_{i,j+\frac{1}{2}} + f_{i,j-\frac{1}{2}} \end{pmatrix} \\ (L_{\leftrightarrow}\phi_f)_{i,j} = \begin{pmatrix} \frac{1}{4}(f_{i+\frac{1}{2},j} + f_{i-\frac{1}{2},j} + f_{i+\frac{1}{2},j+1} + f_{i-\frac{1}{2},j+1}) \\ f_{i,j+\frac{1}{2}} \end{pmatrix} \\ (L_{\uparrow}\phi_f)_{i,j} = \begin{pmatrix} f_{i+\frac{1}{2},j} \\ \frac{1}{4}(f_{i,j+\frac{1}{2}} + f_{i,j-\frac{1}{2}} + f_{i+1,j+\frac{1}{2}} + f_{i+1,j-\frac{1}{2}}) \end{pmatrix} \end{cases}$$

With the point of view of Raviart-Thomas fields, one can check that these operators correspond to taking the values in the center of the cells – denoted $p_{i,j} = (x_{i,j}, y_{i,j})$ – and two averages of the values in the middle of the edges:

$$\begin{cases} (L_{\bullet}\phi_f)_{i,j} = \phi_f(x_{i,j}, y_{i,j}) \\ (L_{\leftrightarrow}\phi_f)_{i,j} = \frac{1}{2} \left(\phi_f(x_{i,j} + \frac{h}{2}^-, y_{i,j}) + \phi_f(x_{i,j} + \frac{h}{2}^+, y_{i+1,j}) \right) \\ (L_{\uparrow}\phi_f)_{i,j} = \frac{1}{2} \left(\phi_f(x_{i,j}, y_{i,j} + \frac{h}{2}^-) + \phi_f(x_{i,j+1}, y_{i,j} + \frac{h}{2}^+) \right) \end{cases}$$

where the exponent $+$ or $-$ indicates that we take the right or left limit of ϕ_f at the point at stake; and where we note that the coordinates of the centers of the cells satisfy $x_{i,j+1} = x_{i,j}$ and $y_{i+1,j} = y_{i,j}$. In comparison, the isotropic total variation imposed $|L_1(\phi_f)_{i,j}| \leq 1$ where

$$L_1(\phi_f)_{i,j} = \begin{pmatrix} f_{i+\frac{1}{2},j} \\ f_{i,j+\frac{1}{2}} \end{pmatrix} = \phi_f(x_{i,j} + \frac{h}{2}^-, y_{i,j} + \frac{h}{2}^-)$$

is the value of the field at the upright corner of the cell.

In TV_{RT}^h however, the constraint on ϕ_f is the same as on ϕ on the continuous TV, namely that $|\phi_f(x)| \leq 1$ for all $x \in \Omega$. Note that since the two components of ϕ_f are piecewise affine, the constraint of being less than 1 everywhere on Ω reduces to being less than 1 in the corners of the mesh, that is

$$\begin{aligned} \text{TV}_{RT,N}^h(u_h) &= \sup \left\{ - \int_{\Omega} u_h \operatorname{div} \phi_f, \phi_f \in RT0_0 \text{ s.t.} \right. \\ &\quad \left. \forall i, j \in \llbracket 1, N \rrbracket, \max_{1 \leq k \leq 4} |(L_k \phi_f)_{i,j}| \leq 1 \right\} \\ \text{TV}_{RT,D}^h(u_h) &= \sup \left\{ - \int_{\Omega} u_h \operatorname{div} \phi_f + \int_{\partial\Omega} b^h \langle \phi_f | \vec{n} \rangle, \phi_f \in RT0 \text{ s.t.} \right. \\ &\quad \left. \forall i, j \in \llbracket 1, N \rrbracket, \max_{1 \leq k \leq 4} |(L_k \phi_f)_{i,j}| \leq 1 \right\} \end{aligned}$$

$$\text{with } \begin{cases} (L_1\phi_f)_{i,j} = \phi_f(x_{i,j} + \frac{h}{2}^-, y_{i,j} + \frac{h}{2}^-) \\ (L_2\phi_f)_{i,j} = \phi_f(x_{i,j} - \frac{h}{2}^-, y_{i,j} + \frac{h}{2}^-) \\ (L_3\phi_f)_{i,j} = \phi_f(x_{i,j} - \frac{h}{2}^-, y_{i,j} - \frac{h}{2}^-) \\ (L_4\phi_f)_{i,j} = \phi_f(x_{i,j} + \frac{h}{2}^-, y_{i,j} - \frac{h}{2}^-) \end{cases} \quad \begin{array}{|c|c|} \bullet & \bullet \\ \hline 2 & 1 \\ \hline (x_{i,j}, y_{i,j}) & \\ \hline \times & \\ \hline 3 & 4 \\ \hline \bullet & \bullet \end{array}$$

Finally, the Raviart-Thomas formulation offers a unified framework to deal with these three total variations, see chapter 10.

9.2 Error estimate

In [CP20] Chambolle and Pock have studied a total variation based on Crouzeix-Raviart finite elements on a triangular mesh. This total variation can be computed by approximating the dual fields with Raviart-Thomas fields under a norm constraint in the center of each triangle. Given a source term $g \in L^\infty$, and under a regularity assumption on the dual field, they show there exists a constant c (depending on g and the value of the continuous ROF problem) such that $|\bar{E} - \bar{E}^h| \leq ch$ where we recall that \bar{E} and \bar{E}^h are respectively the optimal values of the continuous and discrete problems:

$$\bar{u} = \arg \min_{u \in BV(\Omega)} \frac{1}{2\lambda} \|u - g\|_{L^2}^2 + \text{TV}(u) =: E(u) \quad (9.6)$$

$$\bar{u}^h = \arg \min_{u^h \in P_0} \frac{1}{2\lambda} \|u^h - g^h\|_{L^2}^2 + \text{TV}_{RT}^h(u^h) =: E^h(u^h) \quad (9.7)$$

with appropriate variants for Dirichlet and Neumann boundary conditions (recall that when no subscript N or D is specified, the proposed results are valid for both settings). Thanks to the strong convexity of the energy these estimates are also controlling the squared L^2 error between \bar{u} and \bar{u}^h . We also refer the reader to [Bar20] for a similar result (Proposition 4.2.) and for many extensions. This study easily transposes to our context and this section is devoted to the proof of the following theorem.

Theorem II.2. *Suppose that $g \in L^\infty(\Omega)$ and that the dual problem of the continuous ROF model (9.6) has a Lipschitz solution. Then there exists a constant $c > 0$ depending only on λ and \bar{E} such that*

$$\forall h > 0, \|\bar{u}^h - \bar{u}\|_{L^2(\Omega)} \leq c\sqrt{h}$$

If in addition $g \in BV(\Omega)$, then there exists a constant $c' > 0$ depending only on λ and \bar{E} such that

$$\forall h > 0, |\bar{E}^h - \bar{E}| \leq c'h$$

The proof of this error estimate is two-fold: a first estimate comes from the primal problems, a second one from the dual problems.

9.2.1 Primal estimate

The first part of the proof relies on the conformal aspect of our discrete total variation TV_{RT}^h (9.4), (9.5) with respect to the continuous total variation TV (9.1). Following a similar strategy than in [BNS15], it follows from the following TV-diminishing lemma (recall that Π_{P0} is the projection on $P0$ defined in section 5.2):

Lemma II.6. *For any function $u \in BV \cap L^2(\Omega)$ satisfying $u|_{\partial\Omega} = b$ if one works in the Dirichlet setting with boundary condition b , one has $\text{TV}_{RT}^h(\Pi_{P0}u) \leq \text{TV}(u)$.*

Proof. The main argument is that if $\phi \in RT0$, then $\text{div } \phi$ is piecewise constant so that $u^h = \Pi_{P0}(u)$ satisfies

$$\int_{\Omega} u^h \text{div } \phi = \int_{\Omega} u \text{div } \phi$$

From that, $\text{TV}_{RT}^h(u^h)$ appears in both Dirichlet and Neumann setting as a supremum over a smaller set of admissible fields ϕ (and hence is lower than $\text{TV}(u)$) modulo a density result about Raviart-Thomas fields that we detail below.

In the Neumann setting, for $\phi \in RT0_0$ such that $|\phi| \leq 1$, one wants to find a sequence $(\phi_n) \in (\mathcal{C}_c^1(\Omega, \mathbb{R}^2))^{\mathbb{N}}$ such that $|\phi_n| \leq 1$ and $\text{div } \phi_n$ converges to $\text{div } \phi$ in $L^2(\Omega)$ (actually, a weak convergence would be sufficient). As on $\partial\Omega$, $\langle \phi, \vec{n} \rangle = 0$, we naturally extend ϕ by zero to $\mathbb{R}^2 \setminus \Omega$. Then one would like to regularize ϕ by convolution. However, this would not lead to functions with compact support in Ω . To fix this, we introduce a small offset and rather deal with $\psi = \phi \circ \Delta$ where $\Delta : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ realizes a shrinkage centered at $p = (\frac{1}{2}, \frac{1}{2})$ (the center of Ω) through

$$\forall x \in \mathbb{R}^2, \Delta(p + x) = p + (1 + \delta)x$$

where $\delta > 0$ is designed to tend to zero. Doing so, one checks that if $\|x\|_{\infty} \geq \frac{1}{2(1+\delta)}$ then $\Delta(x) \notin \Omega$, hence $\psi(x) = 0$. We then regularize and set

$$\forall x \in \mathbb{R}^2, \phi_n(x) = \int_{\mathbb{R}^2} \psi(x - y) n^2 \rho(ny) dy$$

where ρ is a smooth function such that $\int_{\mathbb{R}^2} \rho = 1$ and $\rho(x) = 0$ for $\|x\|_{\infty} \geq 1$. We find that $\phi_n(x) = 0$ for $\|x\|_{\infty} \geq \frac{1}{n} + \frac{1}{2(1+\delta)}$, so that for any $\delta > 0$, $\phi_n \in \mathcal{C}_c^{\infty}(\Omega, \mathbb{R}^2)$ for n large enough. It is also clear that if $|\phi| \leq 1$ then $|\phi_n| \leq 1$. Now let us prove the desired convergence: first as $\text{div } \psi \in L^2(\Omega)$, $(\text{div } \phi_n)$ converges to $\text{div } \psi$ in $L^2(\Omega)$. Then one has to compare $\text{div } \psi$ to $\text{div } \phi$. As

$$\forall x \in \mathbb{R}^2, \text{div } (\psi)(x) = (1 + \delta)(\text{div } \phi)(\Delta(x))$$

we get, denoting M a bound for the piecewise constant field $\text{div } \phi$:

$$\|\text{div } \psi - \text{div } \phi\|_{L^2(\Omega)} \leq \delta M + \|(\text{div } \phi) \circ \Delta - \text{div } \phi\|_{L^2(\Omega)}$$

To finish with, as $\text{div } \phi$ is piecewise constant, for most of the $x \in \Omega$ we will have $(\text{div } \phi)(\Delta(x)) - \text{div } \phi(x) = 0$. The $x \in \Omega$ for which this is not true are those such that

x and $\Delta(x)$ belong to different cells $C_{i,j}$. This situation can only happen if x is close, precisely at distance at most δ , to a border of a cell. The volume of the set of these particular x is easily bounded by $4N\delta$ (very broadly, this set is included in $2N$ stripes centered on the edges of the mesh and of width 2δ). In addition, for any of these x , we can bound $(\operatorname{div} \phi)(\Delta(x)) - \operatorname{div} \phi(x)$ by $2M$. Therefore, we get

$$\|(\operatorname{div} \phi) \circ \Delta - \operatorname{div} \phi\|_{L^2(\Omega)} \leq 4M\sqrt{N\delta}$$

which finally shows that $\operatorname{div} \psi$ converges to $\operatorname{div} \phi$ in $L^2(\Omega)$ when $\delta \rightarrow 0$, and concludes the proof.

In the Dirichlet setting, one also has to ensure that the quantity $\int_{\partial\Omega} b\langle\phi_n|\vec{n}\rangle$ converges to $\int_{\partial\Omega} b\langle\phi|\vec{n}\rangle$. We use the exact same offset and regularization as before except we extend ϕ by

$$\forall x = (x_1, x_2) \in \mathbb{R}^2 \setminus \Omega, \phi(x) = \begin{cases} \phi(0, x_2) & \text{if } x_1 < 0 \text{ and } x_2 \in (0, 1) \\ \phi(1, x_2) & \text{if } x_1 > 1 \text{ and } x_2 \in (0, 1) \\ \phi(x_1, 0) & \text{if } x_2 < 0 \text{ and } x_1 \in (0, 1) \\ \phi(x_1, 1) & \text{if } x_2 > 1 \text{ and } x_1 \in (0, 1) \\ (0, 0) & \text{otherwise} \end{cases}$$

We divide the integral over $\partial\Omega$ into eight integrals over the half sides of the square, and focus here on the bottom left segment $S = \{0\} \times [0, \frac{1}{2}]$ (the proof is of course the same for the other segments). Thanks to the shift given by Δ , the function ϕ_n is actually constant to ϕ on a significant part of S . Indeed, for j such that $jh \in [0, \frac{1}{2}]$, we find that if $x = (0, x_2)$ is such that $\frac{2jh+\delta}{2(1+\delta)} < x_2 < \frac{2(j+1)h+\delta}{2(1+\delta)}$, then $\Delta(x) \in (-\infty, 0) \times (jh, (j+1)h)$ and consequently $\psi(x) = \phi(0, (j+\frac{1}{2})h)$. As a result, after convolution we obtain that if $\frac{2jh+\delta}{2(1+\delta)} + \frac{1}{n} < x_2 < \frac{2(j+1)h+\delta}{2(1+\delta)} - \frac{1}{n}$ then $\phi_n(x) = \phi(0, (j+\frac{1}{2})h) = \phi(x)$. Finally for n such that for any j , $\frac{2(j+1)h+\delta}{2(1+\delta)} - \frac{1}{n} > (j+1)h$ we can write

$$\begin{aligned} \left| \int_S \langle \phi_n - \phi | \vec{n} \rangle \right| &= \left| \sum_j \int_{jh}^{(j+1)h} \langle \phi_n(0, x_2) - \phi(0, x_2) | \vec{n}(0, x_2) \rangle dx_2 \right| \\ &\leq \sum_j \int_{jh}^{\frac{2jh+\delta}{2(1+\delta)} + \frac{1}{n}} \|\phi_n - \phi\|_\infty \\ &\leq \sum_j \left(\frac{2jh+\delta}{2(1+\delta)} + \frac{1}{n} - jh \right) \times 2 \\ &\leq N \left(\frac{\delta}{2(1+\delta)} + \frac{1}{n} \right) \end{aligned}$$

hence the desired convergence when $n \rightarrow \infty$ and $\delta \rightarrow 0$. \square

This lemma leads the first estimate:

Proposition II.7. *The solutions \bar{u}, \bar{u}^h of (9.6), (9.7) satisfy*

$$\frac{1}{2\lambda} \|\bar{u}^h - \Pi_{P_0} \bar{u}\|_{L^2}^2 \leq \bar{E} - \bar{E}^h - \frac{1}{2\lambda} (\|\bar{u} - g\|_{L^2}^2 - \|\Pi_{P_0}(\bar{u} - g)\|_{L^2}^2)$$

Proof. First, use the strong convexity of \bar{E}^h and write

$$\frac{1}{2\lambda} \|\bar{u}^h - \Pi_{P_0} \bar{u}\|_{L^2}^2 \leq E^h(\Pi_{P_0} \bar{u}) - \bar{E}^h$$

second, thanks to the previous lemma we have

$$E^h(\Pi_{P_0} \bar{u}) \leq \frac{1}{2\lambda} \|\Pi_{P_0} \bar{u} - g^h\|_{L^2}^2 + \text{TV}(\bar{u}) = E(\bar{u}) + \frac{1}{2\lambda} (\|\Pi_{P_0} \bar{u} - g^h\|_{L^2}^2 - \|\bar{u} - g\|_{L^2}^2)$$

and the result follows. \square

9.2.2 Dual estimate

The second estimate relies on the evaluation of the dual problems of (9.6) and (9.7). In the continuous setting, switching the min operator from (9.6) with the supremum defining the total variation leads to the following dual problems:

$$\bar{\phi}_N \in \arg \max_{\substack{\phi \in \mathcal{H}^0 \text{ s.t.} \\ \|\phi\|_\infty \leq 1}} - \int_{\Omega} g \operatorname{div} \phi - \frac{\lambda}{2} \|\operatorname{div} \phi\|_{L^2}^2 =: D_N(\phi) \quad (9.8)$$

$$\bar{\phi}_D \in \arg \max_{\substack{\phi \in \mathcal{H} \text{ s.t.} \\ \|\phi\|_\infty \leq 1}} - \int_{\Omega} g \operatorname{div} \phi - \frac{\lambda}{2} \|\operatorname{div} \phi\|_{L^2}^2 + \int_{\partial\Omega} b \langle \phi | \bar{n} \rangle =: D_D(\phi) \quad (9.9)$$

where $\mathcal{H} = \{\phi \in L^\infty(\Omega) \text{ s.t. } \operatorname{div} \phi \in L^2(\Omega)\}$ and \mathcal{H}^0 is the subset of \mathcal{H} made of fields vanishing at the boundary in the weak sense $\mathcal{H}^0 = \{\phi \in \mathcal{H} \text{ s.t. } \forall u \in H^1(\Omega), \int_{\Omega} \langle \nabla u | \phi \rangle = - \int_{\Omega} u \operatorname{div} \phi\}$. As usual, because of how these problems are obtained one has $D(\phi) \leq E(u)$ for any admissible couple (ϕ, u) . We find in [CCC⁺10] a proof that strong duality holds between these dual and primal problems. It relies on the Euler-Lagrange equation of the ROF problem that states, for instance in the Neumann setting, that \bar{u} is a minimizer of (9.6) if and only if there exists $\bar{\phi} \in \mathcal{H}$ such that $\bar{u} - g = \lambda \operatorname{div} \bar{\phi}$, $\|\bar{\phi}\|_\infty \leq 1$ and $-\int_{\Omega} \bar{u} \operatorname{div} \bar{\phi} = \text{TV}(\bar{u})$. Choosing $\phi = \bar{\phi}$ in the above inequality shows strong duality between primal and dual problems. Finally, this result also holds in the Dirichlet setting and one has: $\bar{D} := D(\bar{\phi}) = \bar{E}$ through the relation $\bar{u} = g + \lambda \operatorname{div} \bar{\phi}$.

The same relations arise in the discrete case, which is completely similar, and the discrete dual problems are:

$$\bar{\phi}_N^h \in \arg \max_{\substack{\phi^h \in RT_0 \\ \|\phi^h\|_\infty \leq 1}} - \int_{\Omega} g^h \operatorname{div} \phi^h - \frac{\lambda}{2} \|\operatorname{div} \phi^h\|_2^2 =: D_N^h(\phi^h)$$

$$\bar{\phi}_D^h \in \arg \max_{\substack{\phi^h \in RT_0 \\ \|\phi^h\|_\infty \leq 1}} - \int_{\Omega} g^h \operatorname{div} \phi^h - \frac{\lambda}{2} \|\operatorname{div} \phi^h\|_2^2 + \int_{\partial\Omega} b^h \langle \phi^h | \bar{n} \rangle =: D_D^h(\phi^h)$$

As previously, one has to be able to get a discrete field from a continuous one through a projection operator similar to Π_{P0} . This will be achieved by the operator $\Pi_{RT0} : \mathcal{H} \rightarrow RT0$ which takes $z = (z_1, z_2) : \Omega \rightarrow \mathbb{R}^2$ to $\phi = \phi_f \in RT0$ given by the fluxes through the edges of the mesh f that are defined as

$$f_{i+1/2,j} = \frac{1}{h} \int_{E_{i+1/2,j}} z_2 ; f_{i,j+1/2} = \frac{1}{h} \int_{E_{i,j+1/2}} z_1$$

where $E_{i+1/2,j} = \partial C_{i,j} \cap \partial C_{i+1,j}$ and $E_{i,j+1/2} = \partial C_{i,j} \cap \partial C_{i,j+1}$. This projection operator enjoys two properties that derive from simple integration formulas.

Lemma II.7. $\forall \phi \in \mathcal{H}, \operatorname{div}(\Pi_{RT0}(\phi)) = \Pi_{P0}(\operatorname{div} \phi)$.

Proof. Using density of smooth functions, we suppose $\phi = (\phi_1, \phi_2)$ is \mathcal{C}^1 , and focus on a square C of the mesh, centered at (x_C, y_C) and with edges labeled e_1 (respectively e'_1, e_2, e'_2) for the left (respectively right, bottom and top) side of C . According to the definition of the projection operator Π_{RT0} , we have that

$$\operatorname{div} \Pi_{RT0}(\phi) = \frac{1}{h^2} \left(\int_{e'_1} \phi_1 - \int_{e_1} \phi_1 + \int_{e'_2} \phi_2 - \int_{e_2} \phi_2 \right)$$

Besides, recall that the side of the square has length h so

$$\begin{aligned} \int_{e'_1} \phi_1 - \int_{e_1} \phi_1 &= \int_{y_C - \frac{h}{2}}^{y_C + \frac{h}{2}} \phi_1(x_C + \frac{h}{2}, y) - \phi_1(x_C - \frac{h}{2}, y) dy \\ &= \int_{y_C - \frac{h}{2}}^{y_C + \frac{h}{2}} \int_{x_C - \frac{h}{2}}^{x_C + \frac{h}{2}} \frac{\partial \phi_1}{\partial x}(x, y) dx dy = \int_C \frac{\partial \phi_1}{\partial x} \end{aligned}$$

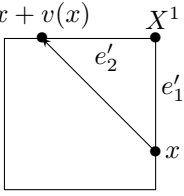
Similarly, $\int_{e'_2} \phi_2 - \int_{e_2} \phi_2 = \int_C \frac{\partial \phi_2}{\partial y}$ and the result follows. \square

Lemma II.8. *If $\phi : \Omega \rightarrow \mathbb{R}^2$ is L -Lipschitz and if $\|\phi\|_\infty \leq 1$ then its projection $\phi^h = \Pi_{RT0}(\phi)$ satisfies $\|\phi^h\|_\infty \leq 1 + \frac{\sqrt{2}}{2} Lh$*

Proof. Denote $\phi = (\phi_1, \phi_2)$ and consider a square C of the mesh. To prove that $|\phi^h| \leq 1 + \frac{\sqrt{2}}{2} Lh$ in C one only has to prove it at the four corners of C as this field is affine. We only focus on the top right corner, denoted X^1 and linked to the right and top edges, denoted e'_1 and e'_2 . According to the definition of the projection operator Π_{RT0} , the value of the field ϕ^h at X^1 is

$$\phi^h(X^1) = \frac{1}{h} \begin{pmatrix} \int_{e'_1} \phi_1 \\ \int_{e'_2} \phi_2 \end{pmatrix} \text{ so } |\phi^h(X^1)|^2 = \left(\frac{1}{h} \int_{e'_1} \phi_1 \right)^2 + \left(\frac{1}{h} \int_{e'_2} \phi_2 \right)^2$$

Then using Jensen's inequality and denoting $Id + v$ the rotation taking e'_1 to e'_2 (see the drawing below) we have:

$$\begin{aligned}
 |\phi^h(X^1)|^2 &\leq \frac{1}{h} \left(\int_{e'_1} \phi_1^2 + \int_{e'_2} \phi_2^2 \right) \\
 &= \frac{1}{h} \left(\int_{e'_1} \phi_1^2(x) + \phi_2^2(x+v(x)) \, dx \right) \\
 &= \frac{1}{h} \left(\int_{e'_1} \phi_1^2 + \phi_2^2 \right) + \frac{1}{h} \left(\int_{e'_1} \phi_2^2(x+v(x)) - \phi_2^2(x) \, dx \right) \\
 &\leq 1 + \frac{1}{h} \int_{e'_1} |\phi_2(x+v(x)) - \phi_2(x)| |\phi_2(x+v(x)) + \phi_2(x)| \, dx \\
 &\leq 1 + \frac{2L}{h} \int_{e'_1} |v(x)| \, dx
 \end{aligned}$$


where we used that ϕ_2 is L -Lipschitz and $|\phi_2| \leq 1$. To finish with, we compute:

$$\int_{e'_1} |v(x)| \, dx = \int_0^h t\sqrt{2} \, dt = \frac{h^2}{2} \sqrt{2}$$

so that $|\phi^h(X^1)| \leq \sqrt{1 + \sqrt{2}Lh} \leq 1 + \frac{\sqrt{2}}{2}Lh$. \square

In our analysis, we will consequently need a Lipschitz hypothesis to hold on the optimal dual field $\bar{\phi}$. As noticed by [CP20], this hypothesis is reasonable in the sense that it is known to hold when g is the characteristic of a disk and $\Omega = \mathbb{R}^2$, as well as in the case $g = g_\nu$ (where one can even take $L = 0$ as $\bar{\phi} = \nu$ is a solution). It seems plausible that this hypothesis is satisfied as long as $g \in L^\infty(\Omega)$ (when working in a bounded convex domain Ω of \mathbb{R}^2), however we are not aware that such a result is known for the time being.

We now apply these two lemmas to get an admissible solution in the discrete dual problem from a (by hypothesis) Lipschitz solution of the continuous dual problem. We get the second estimate:

Proposition II.8. *Suppose the dual continuous problem (9.8) (9.9) admits a L -Lipschitz solution, then one has:*

$$\begin{aligned}
 \bar{D}_N &\leq \left(1 + \frac{\sqrt{2}}{2}Lh\right) \bar{D}_N^h + \frac{1}{2\lambda} \|g - g^h\|_{L^2(\Omega)}^2 \\
 \bar{D}_D &\leq \left(1 + \frac{\sqrt{2}}{2}Lh\right) \bar{D}_D^h + \frac{1}{2\lambda} \|g - g^h\|_{L^2(\Omega)}^2 + \|b - b^h\|_{L^1(\partial\Omega)}
 \end{aligned}$$

Proof. We apply Lemma II.8 to this L -Lipschitz continuous solution $\bar{\phi}$ and we denote $\tilde{\phi} = \Pi_{RT0}(\bar{\phi})$ and $l = \frac{\sqrt{2}}{2}L$. Then $\phi^h = \frac{1}{1+lh}\tilde{\phi}$ is admissible in the dual discrete problem and we write, for Neumann setting:

$$\begin{aligned} D_N^h(\tilde{\phi}) &= -\langle g^h | \operatorname{div} \tilde{\phi} \rangle - \frac{\lambda}{2} \|\operatorname{div} \tilde{\phi}\|_{L^2}^2 \\ &= -\frac{\lambda}{2} (1+lh)^2 \|\operatorname{div} \phi^h\|_{L^2}^2 - (1+lh) \langle g^h | \operatorname{div} \phi^h \rangle \\ &\leq -\frac{\lambda}{2} (1+lh) \|\operatorname{div} \phi^h\|_{L^2}^2 - (1+lh) \langle g^h | \operatorname{div} \phi^h \rangle = (1+lh) D_N^h(\phi^h) \\ &\leq (1+lh) \bar{D}_N^h \end{aligned}$$

Besides, thanks to Lemma II.7 and Jensen's inequality:

$$\begin{aligned} D_N^h(\tilde{\phi}) &= -\frac{\lambda}{2} \|\operatorname{div} \tilde{\phi}\|_{L^2}^2 - \langle g^h | \operatorname{div} \tilde{\phi} \rangle = \frac{1}{2\lambda} \|g^h\|_{L^2}^2 - \frac{1}{2\lambda} \|g^h + \lambda \operatorname{div} \tilde{\phi}\|_{L^2}^2 \\ &= \frac{1}{2\lambda} \|g^h\|_{L^2}^2 - \frac{1}{2\lambda} \|\Pi_{P0}(g + \lambda \operatorname{div} \bar{\phi})\|_{L^2}^2 \\ &\geq \frac{1}{2\lambda} \|g^h\|_{L^2}^2 - \frac{1}{2\lambda} \|g + \lambda \operatorname{div} \bar{\phi}\|_{L^2}^2 = D_N(\bar{\phi}) - \frac{1}{2\lambda} (\|g\|_{L^2}^2 - \|g^h\|_{L^2}^2) \end{aligned}$$

The desired estimation follows (just noticing that $\|g - g^h\|_{L^2}^2 = \|g\|_{L^2}^2 - \|g^h\|_{L^2}^2$ because $g^h = \Pi_{P0}(g)$). In the Dirichlet setting, the same proof applies to get the estimate $D_D^h(\tilde{\phi}) \leq (1+lh) \bar{D}_D^h$; and noticing that $\int_{\partial\Omega} b^h \langle \tilde{\phi} | \vec{n} \rangle = \int_{\partial\Omega} b^h \langle \bar{\phi} | \vec{n} \rangle$, one finally obtains

$$D_D^h(\tilde{\phi}) \geq \bar{D}_D + \frac{1}{2\lambda} (\|g^h\|_{L^2}^2 - \|g\|_{L^2}^2) + \int_{\partial\Omega} (b^h - b) \langle \bar{\phi} | \vec{n} \rangle$$

which leads to the result. \square

9.2.3 Combination of the two estimates

Now combining Propositions II.7 and II.8, we deduce, for instance in the Neumann setting, that there exists a constant $c > 0$, depending on the optimal energy \bar{E} such that, thanks to Jensen's inequality:

$$\begin{aligned} \frac{1}{2\lambda} \|\bar{u}^h - \Pi_{P0}\bar{u}\|_{L^2}^2 &\leq ch + \frac{1}{2\lambda} (\|g - g^h\|_{L^2}^2 - (\|\bar{u} - g\|_{L^2}^2 + \|\Pi_{P0}(\bar{u} - g)\|_{L^2}^2)) \\ &= ch + \frac{1}{2\lambda} \left(\|\Pi_{P0}\bar{u}\|_{L^2}^2 - \|\bar{u}\|_{L^2}^2 + 2 \int_{\Omega} \bar{u}g - 2 \int_{\Omega} (\Pi_{P0}\bar{u})g^h \right) \\ &\leq ch + \frac{1}{\lambda} \int_{\Omega} g(\bar{u} - \Pi_{P0}\bar{u}) \\ &\leq ch + \frac{1}{\lambda} \|g\|_{\infty} \|\bar{u} - \Pi_{P0}\bar{u}\|_{L^1(\Omega)} \\ &\leq c'h \end{aligned}$$

where in the final inequality we used the first part of following lemma:

Lemma II.9. For any $f \in BV(\Omega)$ and $b \in BV(\partial\Omega)$, one has:

$$\|f - f^h\|_{L^1(\Omega)} \leq \frac{4}{3}\pi\sqrt{2}h\text{TV}(f) \quad \text{and} \quad \|b - b^h\|_{L^1(\partial\Omega)} \leq h\text{TV}(b)$$

Proof. We give here the proof for f , the arguments are the same for b . As there exists a sequence $(f_n) \in (C^\infty(\Omega))^{\mathbb{N}}$ such that $f_n \rightarrow f$ in $L^1(\Omega)$ (so that $f_n^h \rightarrow f^h$ in $L^1(\Omega)$ as well) and $\text{TV}(f_n) \rightarrow \text{TV}(f)$, one can suppose $f \in C^\infty(\Omega)$. In addition, as f is continuous, $\text{TV}_\Omega(f) = \sum_{i,j} \text{TV}_{C_{i,j}}(f)$ so we only need to prove the estimate on a square $C = C_{i,j}$ for some (i,j) . We have:

$$\begin{aligned} \|f - f^h\|_{L^1(C)} &= \int_{x \in C} \left| f(x) - \frac{1}{|C|} \int_{y \in C} f(y) dy \right| dx \\ &= \frac{1}{|C|} \int_{x \in C} \left| \int_{y \in C} f(x) - f(y) dy \right| dx \\ &\leq \frac{1}{|C|} \int_{x,y \in C} |f(x) - f(y)| dy dx \end{aligned}$$

We then write $y = x + a$ and note that as C is a square of size h , one has $|a| \leq \sqrt{2}h$ so that denoting B the disk of radius $\sqrt{2}h$ we can estimate:

$$\begin{aligned} \|f - f^h\|_{L^1(C)} &\leq \frac{1}{|C|} \int_{x \in C} \int_{\substack{a \in B \\ x+a \in C}} |f(x) - f(x+a)| da dx \\ &\leq \frac{1}{|C|} \int_{a \in B} \int_{\substack{x \in C \\ x+a \in C}} |f(x) - f(x+a)| dx da \\ &\leq \frac{1}{|C|} \int_{a \in B} |a| \text{TV}_C(f) da = \frac{4}{3}\pi\sqrt{2}h \text{TV}_C(f) \end{aligned}$$

where we used the following lemma for the last inequality:

Lemma II.10. Let $D \subset \mathbb{R}^2$ be a convex domain and $f \in C^1(D) \cap BV(D)$, then for any $a \in \mathbb{R}^2$,

$$\int_{\substack{x \in D \\ x+a \in D}} |f(x+a) - f(x)| dx \leq |a| \text{TV}_D(f)$$

Proof. Using the convexity of D to integrate along the line $[x, x+a]$ we get:

$$\begin{aligned} \int_{\substack{x \in D \\ x+a \in D}} |f(x+a) - f(x)| dx &\leq \int_{\substack{x \in D \\ x+a \in D}} \int_0^1 |\nabla f(x+sa)| |a| dx ds \\ &\leq |a| \int_0^1 \int_{\substack{x \in D \\ x+sa \in D}} |\nabla f(x+sa)| dx ds \\ &\leq |a| \int_0^1 \int_{y \in D} |\nabla f(y)| dy ds = |a| \text{TV}_D(f) \end{aligned}$$

which concludes the proof. \square

This lemma also applies to get the same estimate in the Dirichlet setting using that $b \in BV(\partial\Omega)$. We finally showed that, under the hypotheses $g \in L^\infty(\Omega)$ and $\bar{\phi}$ is Lipschitz, there exists a constant $c > 0$ such that $\|\bar{u}^h - \Pi_{P_0}\bar{u}\|_{L^2} \leq c\sqrt{h}$. However, to estimate the convergence of the energies $|\bar{E} - \bar{E}^h|$ it seems mandatory to control the term $\|g - g^h\|_{L^2}^2$ through Lemma II.9 and ask that $g \in BV(\Omega)$. In this situation, we finally get the announced convergence rate: for some $c > 0$ depending on g and on the continuous energy \bar{E} , provided $\bar{\phi}$ is Lipschitz,

$$\forall h > 0, |\bar{E} - \bar{E}^h| \leq ch$$

As the error $\|g - g^h\|_{L^2}^2$ made on the discretization of $g \in BV(\Omega)$ is precisely of order $O(h)$, this estimate on the energy is in a way “optimal”. Note finally that the same rates would be obtained with a weaker TV-diminishing lemma only demanding: $\text{TV}^h(\Pi_{P_0}(u)) \leq (1 + ch)\text{TV}(u)$ which could be true for other discrete total variations.

Remark II.2. One could also chose to consider the discrete problem where g^h is replaced by g in the L^2 term, that is to minimize $\widetilde{E}^h(u^h) = \frac{1}{2\lambda}\|u^h - g\|_{L^2}^2 + \text{TV}_{RT}^h(u^h)$. Actually this leads to the same optimizer \bar{u}^h as $\widetilde{E}^h = E^h + \frac{1}{2\lambda}\|g - g^h\|_{L^2}^2$. However, denoting $\widetilde{\bar{E}}^h$ the optimal value of this energy, Proposition II.8 then writes $\bar{E} - \widetilde{\bar{E}}^h \leq ch$. Meanwhile, after using the already mentioned calculation

$$\|\bar{u} - g\|_{L^2}^2 - \|\Pi_{P_0}(\bar{u} - g)\|_{L^2}^2 = \|\bar{u} - \Pi_{P_0}\bar{u}\|_{L^2}^2 + \|g - g^h\|_{L^2}^2 - 2 \int_{\Omega} g(\bar{u} - \Pi_{P_0}\bar{u})$$

one sees that Proposition II.7 implies $\bar{E} - \widetilde{\bar{E}}^h \geq -\frac{1}{\lambda} \int_{\Omega} g(\bar{u} - \Pi_{P_0}\bar{u}) \geq -ch$ so that finally, even when $g \in L^\infty(\Omega)$ is such that $g \notin BV(\Omega)$ (but still under the assumption that there exists a Lipschitz dual field $\bar{\phi}$), one has

$$|\bar{E} - \widetilde{\bar{E}}^h| \leq ch$$

CHAPTER 10

IMPLEMENTATION AND RESULTS

In this chapter we briefly present the way we implemented the different total variations through a united Raviart-Thomas framework, and the way we computed solutions of the ROF problem through a primal-dual algorithm. Finally, we give numerical results illustrating the $O(h^{2/3})$ estimate obtained for the isotropic total variation and the behavior of the Raviart-Thomas total variation on simple problems.

10.1 A united framework

As we have seen, the Raviart-Thomas fields offer a united framework to deal with different total variations. Indeed, TV_i^h , TV_{RT}^h as well as the total variation proposed in [Con17, HRH14] $\text{TV}_{\text{Condat}}^h$ can all be expressed in the form:

$$\begin{aligned}\text{TV}_N^L(u^h) &= \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi, \phi \in RT0_0 \text{ s.t. } \| |L\phi| \|_{\infty} \leq 1 \right\} \\ \text{TV}_D^L(u^h) &= \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi + \int_{\partial\Omega} b^h \langle \phi | \vec{n} \rangle, \phi \in RT0 \text{ s.t. } \| |L\phi| \|_{\infty} \leq 1 \right\}\end{aligned}$$

where $L : RT0 \rightarrow (\mathbb{R}^2)^{\mathcal{I}}$ is some linear operator and \mathcal{I} a finite set of indices. This operator gives the constraints that the dual field must satisfy, namely that for all $k \in \mathcal{I}$, $\forall i, j, |(L_k \phi)_{i,j}| \leq 1$.

In the case of the isotropic total variation, the set \mathcal{I} only has one element, and, using notations of section 9.1, one has $L = L_1$. For the Raviart-Thomas total variation, there are $|\mathcal{I}| = 4$ types of constraints and $L = (L_1, L_2, L_3, L_4)$. Finally, for Condat total variation, $|\mathcal{I}| = 3$ and $L = (L_{\bullet}, L_{\leftrightarrow}, L_{\leftrightarrow})$.

All these L operators are easily computed through the following relations, expressing their values on $\phi = \phi_f \in RT0$: for all $i, j \in \llbracket 0, N \rrbracket$,

$$\begin{aligned} (L_1(\phi_f))_{i,j} &= \begin{pmatrix} f_{i+\frac{1}{2},j} \\ f_{i,j+\frac{1}{2}} \end{pmatrix} & (L_2(\phi_f))_{i,j} &= \begin{pmatrix} f_{i-\frac{1}{2},j} \\ f_{i,j+\frac{1}{2}} \end{pmatrix} \\ (L_3(\phi_f))_{i,j} &= \begin{pmatrix} f_{i-\frac{1}{2},j} \\ f_{i,j-\frac{1}{2}} \end{pmatrix} & (L_4(\phi_f))_{i,j} &= \begin{pmatrix} f_{i+\frac{1}{2},j} \\ f_{i,j-\frac{1}{2}} \end{pmatrix} \\ (L_\bullet \phi_f)_{i,j} &= \frac{1}{2} \begin{pmatrix} f_{i+\frac{1}{2},j} + f_{i-\frac{1}{2},j} \\ f_{i,j+\frac{1}{2}} + f_{i,j-\frac{1}{2}} \end{pmatrix} \\ (L_{\leftrightarrow} \phi_f)_{i,j} &= \begin{pmatrix} \frac{1}{4}(f_{i+\frac{1}{2},j} + f_{i-\frac{1}{2},j} + f_{i+\frac{1}{2},j+1} + f_{i-\frac{1}{2},j+1}) \\ f_{i,j+\frac{1}{2}} \end{pmatrix} \\ (L_{\updownarrow} \phi_f)_{i,j} &= \begin{pmatrix} f_{i+\frac{1}{2},j} \\ \frac{1}{4}(f_{i,j+\frac{1}{2}} + f_{i,j-\frac{1}{2}} + f_{i+1,j+\frac{1}{2}} + f_{i+1,j-\frac{1}{2}}) \end{pmatrix} \end{aligned}$$

with $f_{k,l} = 0$ for couples (k, l) such that this quantity is not defined.

Note that the four variants of the isotropic total variation obtained through the four combinations of directions selected to discretize the ∇ operator (and that we denoted $\text{TV}_{i,\oplus,\ominus}^h$ for $\oplus, \ominus \in \{+, -\}$ in the introduction) correspond to enforcing the constraints $\| |L_k(\phi_f)| \|_\infty \leq 1$ for $1 \leq k \leq 4$ separately. On the contrary, the Raviart-Thomas total variation enforces the four of them simultaneously.

10.2 Resolution by a primal-dual algorithm

To solve it numerically, we write the (dual) ROF problem in the following way, for instance for Neumann boundary conditions:

$$\begin{aligned} & \min_{u^h \in P0} \frac{1}{2\lambda} \|u^h - g^h\|_{L^2}^2 + \sup \left\{ - \int_{\Omega} u^h \operatorname{div} \phi_f, \phi_f \in RT0_0 \text{ s.t. } \| |L\phi_f| \|_\infty \leq 1 \right\} \\ &= \sup_{\phi_f \in RT0_0} \min_{u^h \in P0} \frac{1}{2\lambda} \|u^h - g^h\|_{L^2}^2 - \int_{\Omega} u^h \operatorname{div} \phi_f - F(L\phi_f) \\ &= - \min_{\phi_f \in RT0_0} G(\phi_f) + F(L\phi_f) \end{aligned}$$

where $G(\phi_f) = \frac{\lambda}{2} \|\operatorname{div} \phi_f\|_{L^2}^2 + \int_{\Omega} g^h \operatorname{div} \phi_f$ and $F : (\mathbb{R}^2)^{\mathcal{I}} \rightarrow \mathbb{R}$ is given by $F(z) = 0$ if $\| |z| \|_\infty \leq 1$, $+\infty$ otherwise. Note that the optimal primal solution \bar{u}^h will be obtained from the optimal ϕ_f through $\bar{u}^h = g^h + \lambda \operatorname{div} \phi_f$.

This allows one to use one of the primal-dual algorithms presented in [CP11] for which one needs to calculate the following proximal operators (we denote F^* the convex conjugate of F and use the Moreau identity to calculate its prox, see [BC11]):

$$(Id + \tau \partial G)^{-1}(\phi_f) = \left(\frac{1}{\tau} Id + \lambda DD^* \right)^{-1} \left(\frac{1}{\tau} \phi_f + Dg^h \right)$$

$$(Id + \sigma \partial F^*)^{-1}(z) = \left\{ \begin{array}{l} 0 \text{ if } |z_i| \leq \sigma \\ z_i \left(1 - \frac{\sigma}{|z_i|}\right) \text{ otherwise} \end{array} \right\}_{i \in \mathcal{I}}$$

where $D = -\text{div}^*$ is the opposite of the dual operator of the divergence on the $RT0$ fields, which corresponds to a finite difference approximation of the gradient.

Finally, we use the simplest version of the primal-dual algorithms of [CP11] and obtain the following:

Algorithm II.1. From $\phi_f^0 \in RT0_0$, $z^0 \in (\mathbb{R}^2)^{\mathcal{I}}$, and $\sigma, \tau > 0$ such that $\sigma\tau \|L\|^2 \leq 1$, set $\bar{\phi}_f^0 = \phi_f^0$ and do for $n = 0, 1, \dots$

$$z^{n+1} = (Id + \sigma \partial F^*)^{-1}(z^n + \sigma L^* \bar{\phi}_f^n)$$

$$\phi_f^{n+1} = \left(\frac{1}{\tau} Id + \lambda DD^* \right)^{-1} \left(\frac{1}{\tau} \phi_f^n - L^* z^{n+1} + Dg^h \right)$$

$$\bar{\phi}_f^{n+1} = 2\phi_f^{n+1} - \phi_f^n$$

One checks that in the Dirichlet setting, the function G is replaced by $G(\phi_f) = \frac{\lambda}{2} \|\text{div} \phi_f\|_{L^2}^2 + \int_{\Omega} g^h \text{div} \phi_f - \int_{\partial\Omega} b^h \langle \phi_f | \vec{n} \rangle$ and that the same algorithm applies just replacing Dg^h with the appropriate correction to take into account the boundary term (namely in Neumann $Dg^h = 0$ on the boundary edges, while in Dirichlet Dg^h has value Dg_b^h such that $\int_{\partial\Omega} \langle \phi_f | Dg_b^h \rangle = \int_{\partial\Omega} b^h \langle \phi_f | \vec{n} \rangle$).



Figure 10.1 – Denoising experiment using the ROF model with TV_{RT}^h : on the left, the noisy image g^h ; on the right, the denoised image \bar{u}^h

10.3 Numerical results

First, computation of the 1D problem (6.1) reveals that the $O(h^{2/3})$ rate is almost observed in practice. In Figure 10.2 we plotted the value of the energy \bar{E}^h in the Dirichlet setting for N ranging in $[100, 5000]$ with a stepsize of 100. The corresponding log – log graph exhibits a numerical convergence rate of h^θ with $\theta = 0.6240$.

We present in Figure 10.3 the results for the denoising of a line, that is $g = g_\nu$ in the Dirichlet setting for different orientations ν and for the three total variations we considered: isotropic, Raviart-Thomas and Condat. We also give the results for the denoising of the circle we first showed in Figure 5.2 for the isotropic total variation.

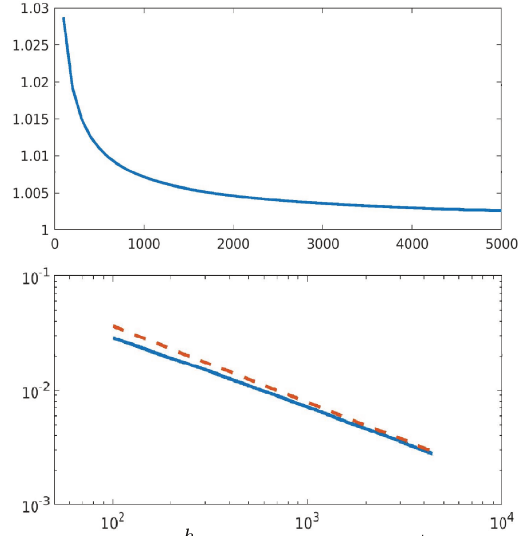


Figure 10.2 – \bar{E}^h (plain blue) and $h^{2/3}$ (dashed red) as functions of $N = \frac{1}{h}$ (cartesian scale (top) and log – log scale (bottom))

We see that the Raviart-Thomas total variation performs as well as the Condat total variation of [Con17, HRH14]. However, it is important to notice that this good behavior relies heavily on the presence of the L^2 term $\|u - g\|_{L^2}^2$ in the problem we considered. Indeed, when tackling the inpainting problem, that is the completion of a missing image (here, a plain discontinuity) from its boundary datum:

$$\begin{aligned} \arg \min_{\substack{u^h \in P0 \\ \text{s.t. } u^h|_B = g^h|_B}} \text{TV}_D^h(u^h) \end{aligned} \quad (10.1)$$

where B denotes the $4N - 4$ border pixels of the image, the Raviart-Thomas total variation does *worse* than the isotropic total variation, while the Condat total variation still produces sharp discontinuities, see Figure 10.4.

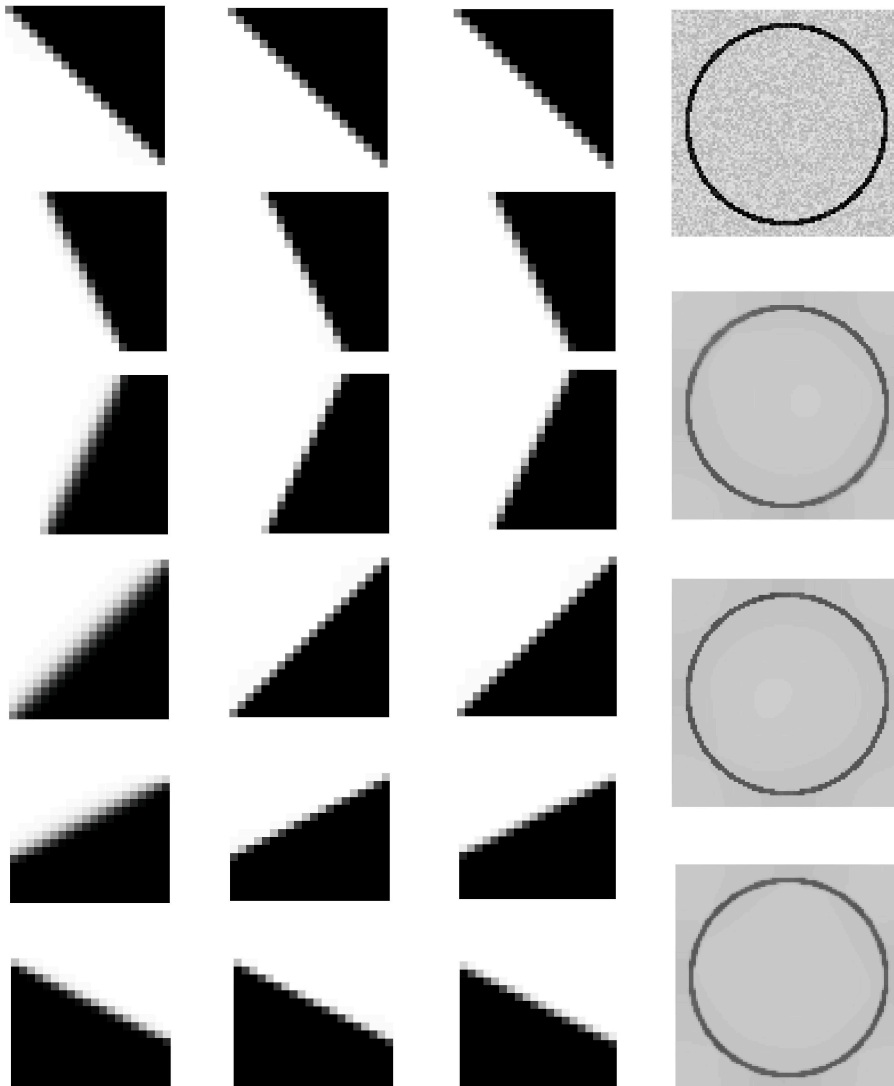


Figure 10.3 – Denoising lines and circle with TV_i^h (left column and second circle), TV_{RT}^h (middle column and third circle) and TV_{Condat}^h (right column and fourth circle)

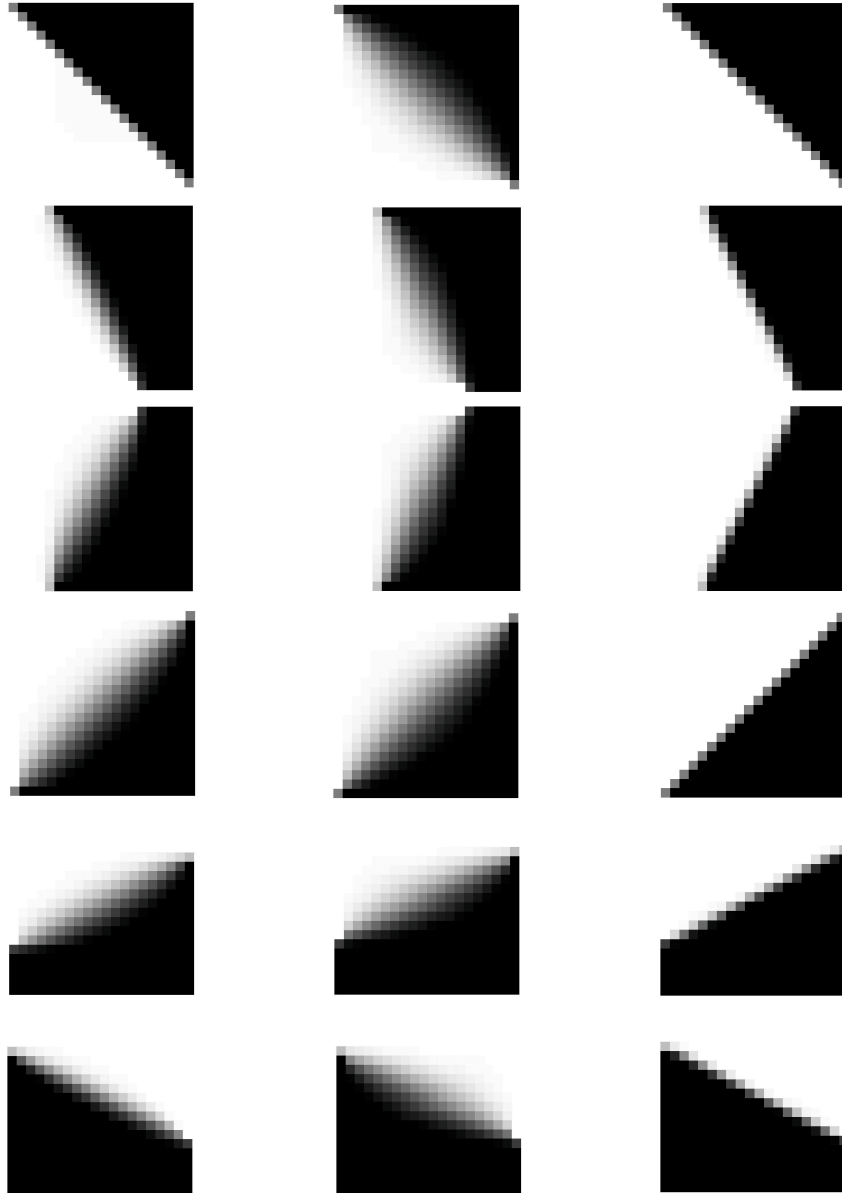


Figure 10.4 – Inpainting lines with TV_i^h (left), TV_{RT}^h (middle) and TV_{Condat}^h (right)

CONCLUSION OF PART II

In the second part of this thesis, we studied two discretizations of the ROF denoising problem, that is the minimization of the sum of a quadratic term and a total variation term. We first proved a $O(h^{2/3})$ error estimate when using the standard “isotropic” discretization of the total variation for discontinuities in the “bad” direction $\frac{1}{\sqrt{2}}(-1, 1)$. Our proof relies on the analysis of a translational invariant problem which reduces to a 1D problem with a non standard discretization of the 1D total variation of the form $\sum_n \sqrt{(u_{n+1} - u_n)^2 + (u_n - u_{n-1})^2}$. We only stated our result in terms of energies of the problems, leaving the convergence rate of the minimizer to later studies. For the same problem, we find that the error is essentially zero in the 90° flipped “good” direction $\frac{1}{\sqrt{2}}(1, 1)$. Additional investigations on the behavior of the problem in intermediate directions could be led; for instance the direction $\frac{1}{\sqrt{5}}(2, 1)$ could be reduced to a 1D denoising problem with a total variation of the form $\sum_n \sqrt{(u_{n+2} - u_n)^2 + (u_n - u_{n-1})^2}$.

In a second part, we performed the study of error rates for the same problem, but with a discretization of the total variation based on Raviart-Thomas fields. We found an optimal $O(h)$ error estimate under standard hypotheses. We numerically showed that this total variation behaves well on the denoising problem but we also observed that it is quite poor at “inpainting” tasks, that is the completion of missing features, such as discontinuities. It would remain to establish consistency and error estimates for total variations doing well on both problems such as a discretization recently analyzed by Condat.

Finally, as we presented briefly in the introduction of Part II, the two parts of this thesis can be united in the context of optimization problems involving both a total variation term and a Wasserstein distance term. Among these problems are the Wasserstein flow of the total variation or of the perimeter, and the computation of total variation-regularized barycenters of indicator functions. We hope that this can serve as a starting point for future works on these subjects.

BIBLIOGRAPHY

- [AC11] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.
- [Ach93] E. Achilles. Implications of convergence rates in Sinkhorn balancing. *Linear Algebra Appl.*, 187:109–112, 1993.
- [AFP00] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.
- [AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.
- [ANWR17] J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems 30*, pages 1964–1974. Curran Associates, Inc., 2017.
- [AZJ⁺18] A. Arnab, S. Zheng, S. Jayasumana, B. Romera-Paredes, M. Larsson, A. Kirillov, B. Savchynskyy, C. Rother, F. Kahl, and P. H. S. Torr. Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction. *IEEE Signal Processing Magazine*, 35(1):37–52, 2018.
- [Bar20] S. Bartels. Nonconforming discretizations of convex minimization problems and precise relations to mixed methods. <https://arxiv.org/abs/2002.02359>, 2020.
- [BC89] D. Bertsekas and D. Castanon. The auction algorithm for the transportation problem. *Annals of Operations Research*, 20:67–96, 1989.

- [BC11] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011.
- [BCC⁺15] J-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.*, 37(2):A1111–A1138, 2015.
- [BCP19] J. Bigot, E. Cazelles, and N. Papadakis. Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration. *Inf. Inference*, 8(4):719–755, 2019.
- [BDM09] R. Burkard, M. Dell’Amico, and S. Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, USA, 2009.
- [BE88] D. P. Bertsekas and J. Eckstein. Dual coordinate step methods for linear network flow problems. *Mathematical Programming*, 42(1):203–243, 1988.
- [BHH⁺87] J. C. Bezdek, R. J. Hathaway, R. E. Howard, C. A. Wilson, and M. P. Windham. Local convergence analysis of a grouped variable version of coordinate descent. *J. Optim. Theory Appl.*, 54(3):471–477, 1987.
- [Bir57] G. Birkhoff. Extensions of Jentzsch’s theorem. *Trans. Amer. Math. Soc.*, 85:219–227, 1957.
- [BL00] H. H. Bauschke and A. S. Lewis. Dykstra’s algorithm with Bregman projections: a convergence proof. *Optimization*, 48(4):409–427, 2000.
- [BL18] J. Brossard and C. Leuridan. Iterated proportional fitting algorithm and infinite products of stochastic matrices. In *Séminaire de Probabilités XLIX*, volume 2215 of *Lecture Notes in Mathematics*, pages 75–117. Springer, 2018.
- [BNS15] S. Bartels, R. H. Nochetto, and A. J. Salgado. A total variation diminishing interpolation operator and applications. *Math. Comp.*, 84(296):2569–2587, 2015.
- [BP79] A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*. Computer science and applied mathematics. Academic Press, 1979.
- [BR97] R. B. Bapat and T. E. S. Raghavan. *Nonnegative Matrices and Applications*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1997.
- [Bra02] A. Braides. Γ -convergence for beginners, volume 22 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2002.

- [Bre67] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967.
- [Bru68] R. A. Brualdi. Convex sets of non-negative matrices. *Canadian J. Math.*, 20:144–157, 1968.
- [Car04] J. E. Carroll. Birkhoff’s contraction coefficient. *Linear Algebra and its Applications*, 389:227 – 234, 2004.
- [CC20] C. Caillaud and A. Chambolle. Error estimates for finite differences approximations of the total variation. <https://hal.archives-ouvertes.fr/hal-02539136/>, 2020.
- [CCC⁺10] A. Chambolle, V. Caselles, S. Cremers, M. Novaga, and T. Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.
- [Cha05] A. Chambolle. Total variation minimization and a class of binary mrf models. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 136–152, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
- [CK18] D. Chakrabarty and S. Khanna. Better and simpler error analysis of the Sinkhorn-Knopp algorithm for matrix scaling. In *1st Symposium on Simplicity in Algorithms*, volume 61 of *OASiCs OpenAccess Ser. Inform.*, pages Art. No. 4, 11. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2018.
- [CL97] A. Chambolle and P-L. Lions. Image recovery via total variation minimization and related problems. *Numer. Math.*, 76(2):167–188, 1997.
- [CL19] A. Chambolle and T. Laux. Mullins-Sekerka as the Wasserstein flow of the perimeter. <https://arxiv.org/abs/1910.02508>, 2019.
- [Con17] L. Condat. Discrete total variation: new definition and minimization. *SIAM J. Imaging Sci.*, 10(3):1258–1290, 2017.
- [CP11] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, 2011.
- [CP19] G. Carlier and C. Poon. On the total variation Wasserstein gradient flow and the TV-JKO scheme. *ESAIM Control Optim. Calc. Var.*, 25:Paper No. 42, 21, 2019.
- [CP20] A. Chambolle and T. Pock. Crouzeix–Raviart approximation of the total variation on simplicial meshes. *Journal of Mathematical Imaging and Vision*, 2020.

- [CSM94] R. Cominetti and J. San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Math. Programming*, 67(2, Ser. A):169–187, 1994.
- [CT93] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.*, 3(3):538–543, 1993.
- [CTV17] A. Chambolle, P. Tan, and S. Vaiter. Accelerated alternating descent methods for Dykstra-like problems. *Journal of Mathematical Imaging and Vision*, 2017.
- [Cut13] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems*, 26, 2013.
- [CZ92] Y. Censor and S. A. Zenios. Proximal minimization algorithm with D -functions. *J. Optim. Theory Appl.*, 73(3):451–464, 1992.
- [dB] P-T. de Boer. Translation of [Kru37]. <https://wwwhome.ewi.utwente.nl/~ptdeboer/misc/kruithof-1937-translation.html>.
- [DGK18] P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn’s algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.
- [DJS07] P. Destuynder, M. Jaoua, and H. Sellami. A dual algorithm for denoising and preserving edges in image processing. *J. Inverse Ill-Posed Probl.*, 15(2):149–165, 2007.
- [DJS12] P. Destuynder, M. Jaoua, and H. Sellami. An error estimate in image processing. *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées*, 15:61–81, 2012.
- [FL89] J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra Appl.*, 114/115:717–735, 1989.
- [Gur03] L. Gurvits. Classical deterministic complexity of edmonds’ problem and quantum entanglement. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, STOC ’03, page 10–19, New York, NY, USA, 2003. Association for Computing Machinery.
- [Hes06] T. Heskes. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *J. Artificial Intelligence Res.*, 26:153–190, 2006.

- [HHS⁺18] M. Herrmann, R. Herzog, S. Schmidt, J. Vidal-Núñez, and G. Wachsmuth. Discrete total variation with finite elements and applications to imaging. *Journal of Mathematical Imaging and Vision*, 61:411–431, 2018.
- [HRH14] M. Hintermüller, C. N. Rautenberg, and J. Hahn. Functional-analytic and numerical issues in splitting methods for total variation-based image reconstruction. *Inverse Problems*, 30(5):055014, 34, 2014.
- [HRS88] D. Hershkowitz, U. G. Rothblum, and H. Schneider. Classifications of nonnegative matrices using diagonal equivalence. *SIAM Journal on Matrix Analysis and Applications*, 9(4):455–460, 1988.
- [Ide16] M. Idel. A review of matrix scaling and Sinkhorn’s normal form for matrices and positive maps. <https://arxiv.org/abs/1609.06349>, 2016.
- [Kan42] L. Kantorovitch. On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 37:199–201, 1942.
- [Kar37] S. Karczmarz. Angenaherte auflosung von systemen linearer gleichungen. *Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat.*, pages 355–357, 1937.
- [KLRS08] B. Kalantari, I. Lari, F. Ricca, and B. Simeone. On the complexity of general matrix scaling and entropy minimization via the RAS algorithm. *Mathematical Programming*, 112:371–401, 04 2008.
- [KMDL19] Y. Kushinsky, H. Maron, N. Dym, and Y. Lipman. Sinkhorn algorithm for lifted assignment problems. *SIAM J. Imaging Sci.*, 12(2):716–735, 2019.
- [Kni08] P. A. Knight. The Sinkhorn-Knopp algorithm: convergence and applications. *SIAM J. Matrix Anal. Appl.*, 30(1):261–275, 2008.
- [Kru37] J. Kruithof. Telefoonverkeersrekening. *De Ingenieur*, 1937.
- [KSS⁺20] P. Knöbelreiter, C. Sormann, A. Shekhovtsov, F. Fraundorfer, and T. Pock. Belief propagation reloaded: Learning bp-layers for labeling problems. <https://arxiv.org/abs/2003.06258>, 2020.
- [Kuh55] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Res. Logist. Quart.*, 2:83–97, 1955.
- [LG12] O. Lezoray and L. Grady. *Image Processing and Analysis with Graphs: Theory and Practice*. Digital imaging and computer vision series. CRC Press, 2012.

- [LLW09] M.-J. Lai, B. Lucier, and J. Wang. *The Convergence of a Central-Difference Discretization of Rudin-Osher-Fatemi Model for Image Denoising*, pages 514–526. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
- [LPP19] C.-O. Lee, E.-H. Park, and J. Park. A finite element approach for the dual Rudin-Osher-Fatemi model and its nonoverlapping domain decomposition methods. *SIAM J. Sci. Comput.*, 41(2):B205–B228, 2019.
- [McC97] R. J. McCann. A convexity principle for interacting gases. *Adv. Math.*, 128(1):153–179, 1997.
- [Men68] M. V. Menon. Matrix links, an extremization problem, and the reduction of a non-negative matrix to one with prescribed row and column sums. *Canadian J. Math.*, 20:225–232, 1968.
- [Min88] H. Minc. *Nonnegative matrices*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., New York, 1988. A Wiley-Interscience Publication.
- [MJGF09] O. Meshi, A. Jaimovich, A. Globerson, and N. Friedman. Convexifying the Bethe free energy. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09*, page 402–410, Arlington, Virginia, USA, 2009. AUAI Press.
- [Mon81] G. Monge. Mémoire sur la théorie des déblais et des remblais. <https://gallica.bnf.fr/ark:/12148/bpt6k35800/f796>, 1781.
- [MS69] M. V. Menon and H. Schneider. The spectrum of a nonlinear operator associated with a matrix. *Linear Algebra Appl.*, 2:321–334, 1969.
- [OS88] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *J. Comput. Phys.*, 79(1):12–49, 1988.
- [Ott98] F. Otto. Dynamics of labyrinthine pattern formation in magnetic fluids: a mean-field theory. *Arch. Rational Mech. Anal.*, 141(1):63–103, 1998.
- [PA02] P. Pakzad and V. Anantharam. Belief propagation and statistical physics. In *Princeton University*, 2002.
- [PC19] G. Peyré and M. Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Foundations and Trends in Machine Learning Series. Now Publishers, 2019.
- [Pea88] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. The Morgan Kaufmann Series in Representation and Reasoning. Morgan Kaufmann, San Mateo, CA, 1988.

- [ROF92] L. Rudin, S. J. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992. [also in *Experimental Mathematics: Computational Issues in Nonlinear Science* (Proc. Los Alamos Conf. 1991)].
- [RT77] P. A. Raviart and J. M. Thomas. A mixed finite element method for 2nd order elliptic problems. In *Mathematical Aspects of Finite Element Methods*, pages 292–315, Berlin, Heidelberg, 1977. Springer Berlin Heidelberg.
- [Rus95] L. Ruschendorf. Convergence of the iterative proportional fitting procedure. *Ann. Statist.*, 23(4):1160–1174, 1995.
- [Sch19] B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM J. Sci. Comput.*, 41(3):A1443–A1481, 2019.
- [SGG11] M. Sharify, S. Gaubert, and L. Grigori. Solution of the optimal assignment problem by diagonal scaling algorithms. <https://arxiv.org/abs/1104.3830>, 2011.
- [Sin64] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.
- [Sin67] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.
- [Sin72] R. Sinkhorn. Continuous dependence on A in the D_1AD_2 theorems. *Proc. Amer. Math. Soc.*, 32:395–398, 1972.
- [SK67] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21:343–348, 1967.
- [Sou91] G. W. Soules. The rate of convergence of Sinkhorn balancing. In *Proceedings of the First Conference of the International Linear Algebra Society (Provo, UT, 1989)*, volume 150, pages 3–40, 1991.
- [TCDP17] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis. Overrelaxed Sinkhorn-Knopp Algorithm for Regularized Optimal Transport. In *NIPS’17 Workshop on Optimal Transport & Machine Learning*, Long Beach, United States, 2017.
- [VKV16] K. Vanjigounder, Narayanankutty K.A., and S. Veni. Performance comparison of total variation based image regularization algorithms. *International Journal on Advanced Science, Engineering and Information Technology*, 6, 2016.
- [Wal17] J. D. Walsh. *The boundary method and general auction for optimal mass transportation and Wasserstein distance computation*. PhD thesis, Georgia Institute of Technology, 2017.

- [Wika] Wikipedia contributors. Iterative proportional fitting — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Iterative_proportional_fitting.
- [Wikb] Wikipedia contributors. Sinkhorn's theorem — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Sinkhorn's_theorem.
- [WJ08] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- [WL11] J. Wang and B. J. Lucier. Error bounds for finite-difference methods for Rudin-Osher-Fatemi image smoothing. *SIAM J. Numer. Anal.*, 49(2):845–868, 2011.
- [XWWZ18] Y. Xie, X. Wang, R. Wang, and H. Zha. A fast proximal point method for computing Wasserstein distance. <https://arxiv.org/abs/1802.04307>, 2018.
- [YFW05] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51:2282 – 2312, 2005.

Titre : Estimations asymptotiques pour des algorithmes de traitement d'images et de données : une étude de l'algorithme de Sinkhorn et une analyse numérique de la minimisation de la variation totale

Mots clés : Optimisation convexe ; Analyse numérique ; Transport Optimal ; Traitement d'images

Résumé : Cette thèse traite de problèmes discrets d'optimisation convexe et s'intéresse à des estimations de leurs taux de convergence. Elle s'organise en deux parties indépendantes. Dans la première partie, nous étudions le taux de convergence de l'algorithme de Sinkhorn et de certaines de ses variantes. Cet algorithme apparaît dans le cadre du Transport Optimal (TO) par l'intermédiaire d'une régularisation entropique. Ses itérations, comme celles de ses variantes, s'écrivent sous la forme de produits composante par composante de matrices et de vecteurs positifs. Pour les étudier, nous proposons une nouvelle approche basée sur des inégalités de convexité simples et menant au taux de convergence linéaire observé en pratique. Nous étendons ce résultat à un certain type de variantes de l'algorithme que nous appelons algorithmes de Sinkhorn équilibrés de dimension 1. Nous présentons ensuite des techniques numériques traitant le cas de la convergence vers zéro du paramètre de régularisation des problèmes de TO. Enfin, nous menons l'analyse complète du taux de convergence en dimension 2. Dans la deuxième partie, nous donnons des estimations d'erreur pour

deux discrétisations de la variation totale (TV) dans le modèle de Rudin, Osher et Fatemi (ROF). Ce problème de débruitage d'image, qui revient à calculer l'opérateur proximal de la variation totale, bénéficie de propriétés d'isotropie assurant la conservation de discontinuités nettes dans les images débruitées, et ce dans toutes les directions. En discrétisant le problème sur un maillage carré de taille h et en utilisant une variation totale discrète standard dite TV isotrope, cette propriété est perdue. Nous démontrons que dans une direction particulière l'erreur sur l'énergie est d'ordre $h^{2/3}$, ce qui est relativement élevé face aux attentes pour de meilleures discrétisations. Notre preuve repose sur l'analyse d'un problème équivalent en dimension 1 et de la TV perturbée qui y intervient. La deuxième variation totale discrète que nous considérons copie la définition de la variation totale continue en remplaçant les champs duaux habituels par des champs discrets dits de Raviart-Thomas. Nous retrouvons ainsi le caractère isotrope du modèle ROF discret. Pour conclure, nous prouvons, pour cette variation totale et sous certaines hypothèses, une estimation d'erreur en $O(h)$.

Title : Asymptotical estimates for some algorithms for data and image processing: a study of the Sinkhorn algorithm and a numerical analysis of total variation minimization

Keywords : Convex optimization; Numerical analysis; Optimal Transportation; Image processing

Abstract : This thesis deals with discrete optimization problems and investigates estimates of their convergence rates. It is divided into two independent parts. The first part addresses the convergence rate of the Sinkhorn algorithm and of some of its variants. This algorithm appears in the context of Optimal Transportation (OT) through entropic regularization. Its iterations, and the ones of the Sinkhorn-like variants, are written as componentwise products of nonnegative vectors and matrices. We propose a new approach to analyze them, based on simple convex inequalities and leading to the linear convergence rate that is observed in practice. We extend this result to a particular type of variants of the algorithm that we call 1D balanced Sinkhorn-like algorithms. In addition, we present some numerical techniques dealing with the convergence towards zero of the regularizing parameter of the OT problems. Lastly, we conduct the complete analysis of the convergence rate in dimension 2. In the second part, we establish error estimates

for two discretizations of the total variation (TV) in the Rudin-Osher-Fatemi (ROF) model. This image denoising problem, that is solved by computing the proximal operator of the total variation, enjoys isotropy properties ensuring the preservation of sharp discontinuities in the denoised images in every direction. When the problem is discretized into a square mesh of size h and one uses a standard discrete total variation – the so-called isotropic TV – this property is lost. We show that in a particular direction the error in the energy is of order $h^{2/3}$ which is relatively large with respect to what one can expect with better discretizations. Our proof relies on the analysis of an equivalent 1D denoising problem and of the perturbed TV it involves. The second discrete total variation we consider mimics the definition of the continuous total variation replacing the usual dual fields by discrete Raviart-Thomas fields. Doing so, we recover an isotropic behavior of the discrete ROF model. Finally, we prove a $O(h)$ error estimate for this variant under standard hypotheses.