Entropic approximation for multi-dimensional martingale optimal transport

Hadrien De March†
December 31, 2018

Abstract

We study the existing algorithms that solve the multidimensional martingale optimal transport. Then we provide a new algorithm based on entropic regularization and Newton’s method. Then we provide theoretical convergence rate results and we check that this algorithm performs better through numerical experiments. We also give a simple way to deal with the absence of convex ordering among the marginals. Furthermore, we provide a new universal bound on the error linked to entropy.

Key words. Martingale optimal transport, entropic approximation, numerics, Newton.

1 Introduction

The problem of martingale optimal transport was introduced as the dual of the problem of robust (model-free) superhedging of exotic derivatives in financial mathematics, see Beiglböck, Henry-Labordère & Penkner [6] in discrete time, and Galichon, Henry-Labordère & Touzi [29] in continuous-time. This robust superhedging problem was introduced by Hobson [38], and was addressing specific examples of exotic derivatives by means of corresponding solutions of the Skorokhod embedding problem, see [22, 36, 37], and the survey [35].

Given two probability measures $\mu, \nu$ on $\mathbb{R}^d$, with finite first order moment, martingale optimal transport differs from standard optimal transport in that the set of all interpolating probability measures $\mathcal{P}(\mu, \nu)$ on the product space is reduced to the subset $\mathcal{M}(\mu, \nu)$ restricted by the martingale condition. We recall from Strassen [57] that $\mathcal{M}(\mu, \nu) \neq \emptyset$ if and only if $\mu \leq \nu$.

The author gratefully acknowledges the financial support of the ERC 321111 Rofirm, and the Chairs Financial Risks (Risk Foundation, sponsored by Société Générale) and Finance and Sustainable Development (IEF sponsored by EDF and CA).

†CMAP, École Polytechnique, hadrien.de-march@polytechnique.org.
in the convex order, i.e. $\mu(f) \leq \nu(f)$ for all convex functions $f$. Notice that the inequality $\mu(f) \leq \nu(f)$ is a direct consequence of the Jensen inequality, the reverse implication follows from the Hahn-Banach theorem.

This paper focuses on giving numerical aspects of martingale optimal transport for finite marginals. Henry-Labordère [33] used dual linear programming techniques to solve this problem, chosing well the cost functions so that the dual constraints were much easier to check. Alfonsi, Corbetta & Jourdain noticed the difficulty, when going to higher dimension to get a discrete approximation of continuous marginals in convex order, that are still in convex order in higher dimension. So they mainly solve this problem, and then do several optimal transport resolutions with primal linear programming. Guo & Oblój [31] provide convergence results in the one dimensional setting of the discrete problem converges to the continuous problem, and they provide a Bregman projection scheme for solving the martingale optimal transport problem in the one dimensional setting. We also mention Tan & Touzi [58] who used a dynamic programming approach to solve a continuous-time version of martingale optimal transport.

The idea of using Bregman projection comes from classical optimal transport. Christian Leonard [42] was the first to have the idea of introducing an entropic penalization in an optimal transport problem. The entropic penalization makes this problem smooth and strictly convex and gives a Gibbs structure to the optimal probability, which has an explicit formula as a function of the dual optimizer. The unanimous adoption of entropic methods for solving optimal transport problems came from Marco Cuturi [23] who noticed that finding the dual solution of the entropic problem was equivalent to finding two diagonal matrices that made a full matrix bistochastic, therefore allowing to use the celebrated Sinkhorn algorithm.

Historically in classical optimal transport, the practitioners used linear programming algorithm to solve it, such as the Hungarian method [41], the auction algorithm [15], the network simplex [2], we may also mention [30]. However, this method was so costly that only small problems could be treated because of the polynomial cost of linear programming algorithms. Later, Benhamou & Brenier [11] found another way of solving numerically the optimal transport problem by making it a dynamic programming problem with a final penalization on the mismatch of the final marginal of the dynamic process with the target marginal. For particular cases, it was also possible to use the Monge-Ampere equation. In the case of the square distance cost, Brenier [18] proved that the optimal coupling is concentrated on a deterministic map, which was the gradient of a “potential” convex function $u$. When furthermore the marginals have densities with respect to the Lebesgue measure, we may prove that $u$ is a solution of the Monge-Ampere equation $\det D^2 u = \frac{g \circ \nabla u}{f}$, where $f$ is the density of $\mu$ and $g$ is the density of $\nu$. This equation satisfies a maximum principle, allowing to solve it in practice, see [14] and [13]. We also mention a smart strategy by Merigot [46], using semi-discrete transport. Levy [43] introduced a Newton method to solve the semi-discrete problem very fast.

For the entropic resolution, Leonard [42] proved that the value of the entropic penalized optimal transport converged to the one of the unpenalized problem, while the optimal transports converged as well to a solution of the optimal transport. See [19] and [21] for more precise
studies of this convergence in particular cases. It have been observed by [40] that the entropic formulation was particularly useful for numerical resolution, as it allowed to use the celebrated Sinkhorn algorithm [53]. The power of this technique has been rediscovered by [23], and widely adopted by the community, see [55], [50], or [60]. This method has already been adapted to different transport problems, such as Wasserstein barycenters [11] and multi-marginal transport problems [12], gradient flows problems [48], unbalanced transport [20], and one dimensional martingale optimal transport [31].

The remarkable work by Schmitzer [52] gives very practical considerations and tricks on how to actually make the Bregman projection algorithm converge fast and stay stable in practice. Cuturi & Peyre [24] used a quasi-Newton method to solve the smooth entropic optimal transport. Their conclusion seems that the Sinkhorn algorithm is still more effective. However, [16] use an inexact Newton method (i.e. including the use of the second derivative) and manage to beat the performance of the Sinkhorn algorithm. We also mention [4] which introduces a "Greenkhorn algorithm" that outperforms the Sinkhorn algorithm according to their experiment, and similarly [59] introduces an overrelaxed version of the Sinkhorn algorithm that squares the linear convergence coefficient, and accelerates the algorithm.

Our subsequent work differs from Guo & Oblój as we explain how to deal with higher dimension, give a more effective algorithm for martingale optimal transport by inexact Newton method. We also provide a speed of convergence for the Bregman projection algorithm, and explains how to deal with the lack of convex ordering of the marginals. Finally the universal bound that we give for the error linked to the entropy term is much sharper than the previous state-of-the-art. This bound may be extended to classical optimal transport for which it does not seem to be in the literature either.

In this paper we introduce several existing algorithms for solving martingale optimal transport such as linear programming, non-smooth semi-dual optimization, and Bregman projections. We introduce the smooth Newton algorithms, and the Newton semi-implied algorithm. Then we give some theoretical results on the speed of convergence of these algorithms, together with solutions to stabilize them and make them work in practice, like the preconditioning for the Newton method, or how to deal with marginals that are not in convex order. We provide new convergence rates for the entropic approximation of the martingale optimal transport, that are much better than the existing ones. The known result is an error of the order $\varepsilon (\ln(N) - 1)$, where $N$ is the size of the discretized grid, while we prove that we can get a result of order $\varepsilon d^2$, where $d$ is the dimension of the space of the problem (1 or 2 in this paper). These rates rely on very strong hypotheses that may be hard to check in practice. However we see on the numerical example that they are well verified in practice.

The paper is organized as follows. Section 2 gives the problem to solve, Section 3 give the different algorithms that we will compare. In Section 4 we provide practical solutions to some usual problems, Section 5 provides theoretical convergence rates for the algorithms, Section 6 gathers the proofs of the theoretical results, and finally Section 7 contains numerical results.
Notation We fix an integer $d \geq 1$.

In all this paper, $\mathbb{R}^d$ is endowed with the Euclidean structure, the Euclidean norm of $x \in \mathbb{R}^d$ will be denoted $|x|$. Let $A \subset \mathbb{R}^d$ we denote $|A|$ the Lebesgue volume of $A$. The map $\iota_A$ is the map equal to 0 on $A$, and $\infty$ otherwise. If $V$ is a topological affine space and $A \subset V$ is a subset of $V$, $\text{int}A$ is the interior of $A$, $\text{cl}A$ is the closure of $A$, $\text{aff}A$ is the smallest affine subspace of $V$ containing $A$, $\text{conv}A$ is the convex hull of $A$, and $\dim(A) := \dim(\text{aff}A)$. Let $(u_\varepsilon)_{\varepsilon > 0}, (v_\varepsilon)_{\varepsilon > 0} \subset V$. We denote that $u_\varepsilon = o(v_\varepsilon)$ if $\lim_{\varepsilon \to 0} \frac{|u_\varepsilon|}{|v_\varepsilon|} = 0$. We further denote $u_\varepsilon \ll v_\varepsilon$. A classical property of $o(\cdot)$ is that

$$u_\varepsilon = v_\varepsilon + o(v_\varepsilon) \quad \text{if and only if} \quad u_\varepsilon = v_\varepsilon + o(u_\varepsilon). \quad (1.1)$$

Let $x_0 \in \mathbb{R}^d$, and $r > 0$, we denote $\text{zoom}^{x_0}_r : x \mapsto x_0 + rx$, $B_r(x_0)$ is the closed ball centered in $x_0$ with radius $r$, and we only write $B_r$ when the center is 0. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we denote $\|f\|_\infty := \sup_{x \in \mathbb{R}^d} f(x)$ its infinite norm, and $\|f\|_R := \sup_{x \in B_R} f(x)$ its infinite norm when restricted to the ball $B_R$, for $R > 0$. Let $a, b \in \mathbb{R}^d$, we denote $a \otimes b := (a_i b_j)_{1 \leq i, j \leq d}$, the only matrix in $M_d(\mathbb{R})$ such that for all $x \in \mathbb{R}^d$, we have $(a \otimes b) x = (b \cdot x) a$. Let $1 \leq k \leq d + 1$ and $u_1, ..., u_k \in \mathbb{R}^d$, we denote $\text{det}_{\text{aff}}(u_1, ..., u_k) := \left| \det \left( (e_j \cdot (u_i - u_k))_{1 \leq i, j \leq k-1} \right) \right|$, where $(e_j)_{1 \leq j \leq k-1}$ is an orthonormal basis of $\text{Vect}(u_1 - u_k, ..., u_{k-1} - u_k)$.

Let $M \in M_d(\mathbb{R})$, a real matrix of size $d$, we denote $\text{det} M$ the determinant of $M$. We also denote $\text{Com}(M)$ the comatrix of $M$: for $1 \leq i, j \leq d$, $\text{Com}(M)_{i,j} = (-1)^{i+j} \text{det} M_{i,j}$, where $M_{i,j}$ is the matrix of size $d - 1$ obtained by removing the $i^{th}$ line and the $j^{th}$ row of $M$. Recall the useful comatrix formula:

$$\text{Com}(M)^t M = M \text{Com}(M)^t = (\text{det} M) I_d. \quad (1.2)$$

As a consequence, whenever $M$ is invertible, $M^{-1} = \frac{1}{\text{det} M} \text{Com}(M)^t$.

We denote $\Omega := \mathbb{R}^d \times \mathbb{R}^d$ and define the two canonical maps

$$X : (x, y) \in \Omega \mapsto x \in \mathbb{R}^d \quad \text{and} \quad Y : (x, y) \in \Omega \mapsto y \in \mathbb{R}^d.$$ 

For $\varphi, \psi : \mathbb{R}^d \rightarrow \mathbb{R}$, and $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$, we denote

$$\varphi \oplus \psi := \varphi(X) + \psi(Y), \quad \text{and} \quad h \otimes := h(X) \cdot (Y - X),$$

with the convention $\infty - \infty = \infty$.

For a Polish space $\mathcal{X}$, we denote by $\mathcal{P}(\mathcal{X})$ the set of all probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Let $\mathcal{Y}$ be another Polish space, and $\mathcal{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. The corresponding conditional kernel $\mathcal{P}_x$ is defined by:

$$\mathcal{P}(dx, dy) = \mathcal{P} \circ X^{-1}(dx) \mathcal{P}_x(dy).$$

We also use this notation for finite measures. For a measure $m$ on $\mathcal{X}$, we denote $L^1(\mathcal{X}, m) := \{ f \in L^0(\mathcal{X}) : m(|f|) < \infty \}$. We also denote simply $L^1(\mathcal{X}, m) := L^1(\mathbb{R}, m)$. 

4
2 Preliminaries

Throughout this paper, we consider two probability measures $\mu$ and $\nu$ on $\mathbb{R}^d$ with finite first order moment, and $\mu \leq \nu$ in the convex order, i.e. $\nu(f) \geq \mu(f)$ for all integrable convex $f$. We denote by $\mathcal{M}(\mu, \nu)$ the collection of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals $\mathbb{P} \circ X^{-1} = \mu$ and $\mathbb{P} \circ Y^{-1} = \nu$. Notice that $\mathcal{M}(\mu, \nu) \neq \emptyset$ by Strassen [57].

For a derivative contract defined by a non-negative coupling function $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+$, the martingale optimal transport problem is defined by:

$$S_{\mu, \nu}(c) := \sup_{P \in \mathcal{M}(\mu, \nu)} \mathbb{P}[c].$$

The corresponding robust superhedging problem is

$$I_{\mu, \nu}(c) := \inf_{(\varphi, \psi, h) \in \mathcal{D}_{\mu, \nu}(c)} \mu(\varphi) + \nu(\psi),$$

where

$$\mathcal{D}_{\mu, \nu}(c) := \{(\varphi, \psi, h) \in \mathbb{L}^1(\mu) \times \mathbb{L}^1(\nu) \times \mathbb{L}^1(\mu, \mathbb{R}^d) : \varphi \otimes \psi + h \otimes c \geq 0\}.$$ 

The following inequality is immediate:

$$S_{\mu, \nu}(c) \leq I_{\mu, \nu}(c).$$

This inequality is the so-called weak duality. For upper semi-continuous coupling, we get from Beiglböck, Henry-Labordère, and Penckner [6], and Zaev [63] that there is strong duality, i.e. $S_{\mu, \nu}(c) = I_{\mu, \nu}(c)$. For any Borel coupling function bounded from below, Beiglböck, Nutz & Touzi [9] in dimension 1, and De March [26] in higher dimension proved that duality holds for a quasi-sure formulation of dual problem and proved dual attainability thanks to the structure of martingale transports evidenced in [27].

Along all this paper, we assume that $\mu$ and $\nu$ are discrete, i.e. we may find finite $X$ and $Y$ so that $\mu = \sum_{x \in X} \mu_x \delta_x$, and $\nu = \sum_{y \in Y} \nu_y \delta_y$, so that all the coordinates of $\mu$ and $\nu$ are positive. Similarly, duality clearly holds thanks to the finiteness of the support, and the dual problem becomes discretized as well: for $(\varphi, \psi, h) \in \mathcal{D}_{\mu, \nu}(c)$, we can denote $\varphi$, $\psi$, and $h$ as vectors $(\varphi(x))_{x \in X}$, $(\varphi(y))_{y \in Y}$, and $(h_i(x))_{x \in X, 1 \leq i \leq d}$.

To solve the martingale transport problem in practice, it seems necessary to discretize the problem. Guo & Oblój [31] prove that the martingale optimal transport problem with continuous $\mu$, $\nu$, and $c$ is a limit of this kind of discrete problem in dimension one under reasonable assumptions. This paper does not focus on proving the convergence of the discretized problem towards the continuous problem, we focus on how to solve the discretized problem.
3 Algorithms

3.1 Primal and dual simplex algorithm

3.1.1 Primal

The natural strategy to solve this problem will be to use linear programming techniques such as simplex algorithm. One major problem with this approach is that the set $\mathcal{M}(\mu, \nu)$ may be empty, because in practice, the discretization of the marginals may break the convex ordering between them, thus making the set $\mathcal{M}(\mu, \nu)$ empty by Strassen theorem. This problem was relieved by Guo & Oblój [31], and by Alfonsi, Corbetta & Jourdain [3]. In [31], they deal with the problem by replacing the convex ordering constraint by an approximate convex ordering constraint which is more resilient to perturbing the marginals. In [3], they go beyond and gives several algorithms to find measures $\nu'$ (resp. $\mu'$) that are in convex order with $\mu$ (resp. with $\nu$) and satisfy some optimality criteria such as minimality of $\nu - \nu'$ (resp. $\mu - \mu'$) in terms of $p-$Wasserstein distance. We also give in Subsubsection 4.3.2 a technique to avoid this issue.

3.1.2 Dual

One huge weakness of the Primal algorithm is that the size of the problem is $|\mathcal{X}| |\mathcal{Y}|$, which is the size of $\mathcal{X} \times \mathcal{Y}$, the support of the probabilities we consider. When $|\mathcal{X}|$ and $|\mathcal{Y}|$ are big, it becomes a problem for memory storage. We notice that the number of constraints is $(d + 1)|\mathcal{X}| + |\mathcal{Y}|$, which is much smaller, because the dual functions $\varphi$, and $h$ are respectively in $\mathbb{R}^\mathcal{X}$ and in $(\mathbb{R}^d)\setminus \mathcal{X}$, and the dual function $\psi$ lies in $\mathbb{R}^\mathcal{Y}$. This is why in practice it makes sense to solve the Kuhn & Tucker dual problem instead of the primal one. We will see considerations on the speed of convergence in Subsubsection 5.4.1.

3.2 Semi-dual non-smooth convex optimization approach

It is well known from classical transport that solving directly the linear programming problem is too costly (see [47]) consequently, some alternative techniques have been developed like the Benamou-Brenier [11] approach, which inspired Tan & Touzi [58] for the continuous time optimal transport problem. The idea consists in solving a Hamilton-Jacobi-Bellman problem with a penalization on the distance between the final marginal and $\nu$. Then an extension of this idea to our two-steps MOT problem gives the following resolution algorithm, suggested by Guo and Oblój [31]. We denote $\mathcal{M}(\mu) := \{\mathbb{P} \in \mathcal{P}(\Omega) : \mathbb{P} \circ X^{-1} = \mu, \text{ and } \mathbb{P}[Y|X] = X, \mu - \text{a.s.}\}$, and
Let we assume that it may be decomposed as $H$ where $\mu, \nu$.

Then with high probability, on each gradient step, the function $V(\psi)$ is a convex function in the variable $\psi$. Then the problem becomes a simple convex optimization problem. It seems appropriate in these conditions to solve the problem with using a classical gradient descent algorithm. It is proved in [58] that $V(\psi)$ is a convex function in the variable $\psi$. Then the subgradient of $V$ at $\psi$ is given by

$$\partial V(\psi) = \left\{ \sum_{x \in \mathcal{X}} \mu_x \sum_i \lambda_i(x) \delta_{y_i(x)} - \nu : (y(x), \lambda(x)) \in \text{argconc}_{c(x)} - \psi(x), \text{ for all } x \in \mathcal{X} \right\}$$

Notice that this set is a singleton for a.e. $\psi \in L^1(\mathcal{Y})$, as $V$ is a convex function in finite dimensions. Then with high probability, on each gradient step, the function $V$ will be differentiable on this point. In practice there is always uniqueness after the first step.

### 3.3 Entropic algorithms

#### 3.3.1 The entropic problem in optimal transport

In practice, this problem is added some regularity by the addition of an entropic penalization (see Leonard [42], Cuturi [23]). Let $\varepsilon > 0$,

$$S^\varepsilon_{\mu, \nu}(c) := \sup_{P \in \mathcal{P}(\mu, \nu)} \mathbb{P}[c] - \varepsilon \mathbb{H}(\mathbb{P}|m_0),$$

where $\mathbb{H}(\mathbb{P}|m_0) := \int_{\mathcal{X}} \left( \ln \left( \frac{d\mathbb{P}}{dm_0} \right) - 1 \right) \frac{d\mathbb{P}}{dm_0} m_0(d\omega)$. The measure $m_0$ is the "reference measure", we assume that it may be decomposed as $m_0 := m^X \otimes m^Y$ such that $\mu$ is dominated by $m^X \in \mathcal{M}(\mathbb{R}^d)$, and $\nu$ is dominated by $m^Y \in \mathcal{M}(\mathbb{R}^d)$. For this text we chose $m_0 := \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \delta_{(x,y)}$.

By the finiteness of the supports of $\mu$ and $\nu$, we know that $\mathbb{P}$ is absolutely continuous with respect to $m_0$. Denote $p := \frac{d\mathbb{P}}{dm_0}$, and abuse notation writing $p \in \mathcal{P}(\mu, \nu)$. Then $H(\mathbb{P}|m_0) := \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} (\ln p(x, y) - 1) p(x, y)$. This problem can be written with Lagrange multipliers,

$$\sup_{\mathbb{P} \in \mathcal{P}(\mu, \nu)} \mathbb{P}[c] - \varepsilon \mathbb{H}(\mathbb{P}|m_0) = \inf_{(\varphi, \psi) \in \mathcal{L}(\mu) \times \mathcal{L}(\nu)} \sup_{\mathbb{P} \in \mathcal{P}(\mu, \nu)} \mathbb{P}[c - \varphi \oplus \psi] - \varepsilon \mathbb{H}(\mathbb{P}|m_0) + \mu[\varphi] + \nu[\psi].$$
which leads to an explicit Gibbs form for the optimal kernel \( p \). Then as the supports are finite we easily get the shape of the optimizer \( p(x, y) = \exp \left( -\frac{\varphi(x) + \psi(y) - c(x, y)}{\varepsilon} \right) \), and the associated dual problem becomes

\[
I_{\mu, \nu}^\varepsilon(c) := \inf_{(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)} \mu[\varphi] + \nu[\psi] + \varepsilon \sum_{x, y} \exp \left( -\frac{\varphi(x) + \psi(y) - c(x, y)}{\varepsilon} \right).
\]

One important property that we need is the \( \Gamma \)-convergence. We say that

\[
F_{\varepsilon} \Gamma \text{-converges to } F \quad \text{when } \varepsilon \to 0 \quad \text{if for all sequence } \varepsilon_n \to 0,
\]

(i) For all sequences \( x_n \to x \), we have \( F(x) \geq \lim \sup \_n F_{\varepsilon_n}(x_n) \).

(ii) There exists a sequence \( x_n \to x \) such that \( F(x) \leq \lim \inf \_n F_{\varepsilon_n}(x_n) \).

The \( \Gamma \)-convergence implies that \( \min_{F_{\varepsilon}} F \to \min_F F \) when \( n \to \infty \), and that if \( x_n \) is a minimizer of \( F_n \) for all \( n \geq 1 \), and if \( x_n \to x \), then \( x \) is a minimizer of \( F \). Leonard [12] proved this \( \Gamma \)-convergence of the penalized problem to the optimal transport problem.

### 3.3.2 The Bregman iterations algorithm

Coupled with the Sinkhorn algorithm [53] introduced by Marco Cuturi for optimal transport [23], this method allows an exponentially fast approximated resolution. Notice that the operator

\[
V_{\varepsilon}(\varphi, \psi) := \mu[\varphi] + \nu[\psi] + \varepsilon \sum_{x, y} \exp \left( -\frac{\varphi(x) + \psi(y) - c(x, y)}{\varepsilon} \right)
\]

is smooth convex. The Euler-Lagrange equations \( \partial_\varphi V_{\varepsilon} = 0 \) (resp. \( \partial_\psi V_{\varepsilon} = 0 \)) are exactly equivalent to the marginal relations \( \mathbb{P} \circ X^{-1} = \mu \) (resp. \( \mathbb{P} \circ Y^{-1} = \nu \)). It was noticed in [23] that these partial optimizations can be obtained in closed form:

\[
\varphi(x) = \varepsilon \ln \left( \frac{1}{\mu_x} \sum_{y} \exp \left( -\frac{\psi(y) - c(x, y)}{\varepsilon} \right) \right),
\]

and

\[
\psi(y) = \varepsilon \ln \left( \frac{1}{\nu_y} \sum_{x} \exp \left( -\frac{\varphi(x) - c(x, y)}{\varepsilon} \right) \right).
\]

By iterating these partial optimization, we obtain the so-called Sinkhorn algorithm (see [53]) that is equivalent to a block optimization of the smooth function \( V_{\varepsilon} \) which dual is called Bregman projection [17], and converges exponentially fast, see Knight [39].

### 3.3.3 The entropic approach for the one period martingale optimal transport problem

As observed by Guo & Obłój [31] in dimension 1, the Sinkhorn algorithm can be extended to the martingale optimal transport problem. With exactly the same computations, we get

\[
S_{\mu, \nu}^\varepsilon(c) := \sup_{\mathbb{P} \in \mathcal{M}(\mu, \nu)} \mathbb{P}[c] - \varepsilon H(\mathbb{P} | m_0)
\]

\[
= I_{\mu, \nu}^\varepsilon(c) := \inf_{(\varphi, \psi, h) \in L^1(\mu) \times L^1(\nu) \times L^0(\mathbb{R}^d)} \mu[\varphi] + \nu[\psi] + \varepsilon \sum_{x, y} \exp \left( -\frac{\varphi(x) + \psi(y) + h(x, y) - c(x, y)}{\varepsilon} \right).
\]
First notice that the $\Gamma$–convergence still holds in this easy finite case.

**Proposition 3.1.** Let $F_\varepsilon : \mathbb{P} \in \mathcal{M}(\mu, \nu) \mapsto \mathbb{P}[c] - \varepsilon H(\mathbb{P} | m_0)$. For $\varepsilon > 0$, $F_\varepsilon$ is strictly concave upper semi-continuous. Furthermore, $F_\varepsilon$ $\Gamma$–converges to $F_0$ when $\varepsilon \to 0$.

**Proof.** This $\Gamma$–convergence is easy by finiteness as the entropy is bounded by $\ln(|\mathcal{X}||\mathcal{Y}|) - 1$ when $\mathbb{P}$ is a probability measure. \qed

We denote $\Delta := \varphi \oplus \psi + h \otimes c$, the convex function to minimize becomes

$$V_\varepsilon(\varphi, \psi, h) := \mu[\varphi] + \nu[\psi] + \varepsilon \sum_{x,y} \exp \left( -\frac{\Delta(x, y)}{\varepsilon} \right).$$

Then the Sinkhorn algorithm is complemented by another step so as to account for the martingale relation:

$$\varphi(x) = \varepsilon \ln \left( \frac{1}{\mu_x} \sum_y \exp \left( -\frac{\psi(y) + h(x) \cdot (y - x) - c(x, y)}{\varepsilon} \right) \right), \quad (3.3)$$

$$\psi(y) = \varepsilon \ln \left( \frac{1}{\nu_y} \sum_x \exp \left( -\frac{\varphi(x) + h(x) \cdot (y - x) - c(x, y)}{\varepsilon} \right) \right),$$

$$0 = \frac{1}{\mu_x} \sum_y \exp \left( -\frac{\Delta(x, y)}{\varepsilon} \right) (y - x),$$

$$= \frac{-1}{\mu_x} \frac{\partial}{\partial h(x)} \varepsilon \sum_y \exp \left( -\frac{\Delta(x, y)}{\varepsilon} \right).$$

Notice that the martingale step is not closed form and is only implied. However, it may be computed almost as fast as $\varphi$, and $\psi$, thanks to the Newton algorithm applied to each smooth strongly convex function $F_x$ of $d$ variables given, for each $x \in \mathcal{X}$ with its derivatives by

$$F_x(h) := \varepsilon \sum_y \exp \left( -\frac{\Delta(x, y)}{\varepsilon} \right), \quad (3.4)$$

$$\nabla F_x(h) = -\sum_y \exp \left( -\frac{\Delta(x, y)}{\varepsilon} \right) (y - x),$$

$$D^2 F_x(h) = \frac{1}{\varepsilon} \sum_y \exp \left( -\frac{\Delta(x, y)}{\varepsilon} \right) (y - x) \otimes (y - x).$$

Notice that the optimization of $F_{x_1}$ and $F_{x_2}$ are independent for $x_1 \neq x_2$.

### 3.3.4 Truncated Newton method

For these problems, it may make sense to use a Newton method, as the problems are smooth, and the Newton method converges very fast. For very highly dimensional problems (here $(d + 1)|\mathcal{X}| + |\mathcal{Y}|$), the inversion of the hessian is too much costly. Then it is in general preferred to use quasi-Newton. Instead of computing the Newton step $D^2 V_\varepsilon^{-1} \nabla V_\varepsilon$, we use a conjugate
gradient algorithm to find by iterations a vector \( p \in D \) such that \( |D^2V^p - \nabla V| \) is small enough, generally in practice this quantity is chosen to be smaller than \( \min \left( \frac{1}{\sqrt{2}}, \sqrt{|\nabla V|} \right) \).

The conjugate gradient algorithm approximates the solution of the equation \( Ax = b \) by solving it "direction by direction" along the most important direction, until a stopping criterion is reached. The exact algorithm may be found in [62].

### 3.3.5 Implied truncated Newton method

Some instabilities may appear from Newton steps as any term of the form \( \exp(X/\varepsilon) \) can easily explode when \( \varepsilon \) is very small and \( X > 0 \). The dimension may also make the conjugate gradient from the quasi-Newton algorithm slow. A good way to avoid this problem and exploit the near-closed formula for the optimal \( \varphi \) and \( h \) when \( \psi \) is fixed, or optimal \( \psi \) when \( \varphi \) and \( h \) are fixed.

Instead of applying the truncated method to \( V^\varepsilon(\varphi, \psi, h) \), we apply the truncated Newton method to \( V^\varepsilon(\psi) := \min_{\varphi, h} V^\varepsilon(\varphi, \psi, h) \). It is elementary that with these definitions we have

\[
\inf_{\varphi, \psi, h} V^\varepsilon(\varphi, \psi, h) = \inf_{\psi} V^\varepsilon(\psi).
\]

Doing this variable implicitation is easy by the fact that we have a closed formula for \( \varphi \) and a quasi-closed formula for \( h \). It brings the great advantage of having the first marginal and the martingale relationship verified, this fact will be exploited in Subsubsection 4.3.2.

Now we give a general framework that allows to use variables implicitation. The following framework should be used with \( F = V^\varepsilon, x = \psi, \) and \( y = (\varphi, h) \). Proposition 3.2 below provides the appropriate convexity result together with the closed formulas for the two first derivatives of \( V^\varepsilon(X) \) that are necessary to apply the truncated Newton algorithm. Let \( A \) and \( B \) finite dimensional spaces and \( F : A \times B \longrightarrow \mathbb{R} \), we say that \( F \) is \( \alpha \)-convex if

\[
\lambda F(\omega_1) + (1 - \lambda)F(\omega_2) - F(\lambda \omega_1 + (1 - \lambda)\omega_2) \geq \alpha \frac{(1 - \lambda)}{2} |\omega_1 - \omega_2|^2,
\]

for all \( \omega_1, \omega_2 \in A \times B \), and \( 0 \leq \lambda \leq 1 \). The case \( \alpha = 0 \) corresponds to the standard notion of convexity.

**Proposition 3.2.** Let \( F : A \times B \longrightarrow \mathbb{R} \) be a \( \alpha \)-convex function. Then the map \( \tilde{F} : x \longmapsto \inf_{y \in B} F(x, y) \) is \( \alpha \)-convex. Furthermore, if \( \alpha > 0 \) and \( F \) is \( C^2 \), then \( y(x) := \arg\min_y F(x, y) \) is unique and we have

\[
\nabla y(x) = -\partial^2_y F^{-1} \partial^2_{yx} F(x, y(x)),
\]

\[
\nabla \tilde{F}(x) = \partial_x F(x, y(x)),
\]

\[
D^2 \tilde{F}(x) = (\partial^2_x F - \partial^2_{xy} F \partial^2_y F^{-1} \partial^2_{yx} F) (x, y(x)).
\]

The proof of Proposition 3.2 is reported in Subsection 6.1.

**Remark 3.3.** The matrix \( \partial^2_{xy} F \partial^2_y F^{-1} \partial^2_{yx} F \) is symmetric positive definite, therefore the curvature of the function \( \tilde{F} \) is reduced by the implicitation process, making heuristically the minimization easier. This fact is also observed in practice.
This method shall be used for the optimization of $V_\varepsilon$, but also for the optimization of $F_x$ that gives the martingale step, see (3.3). Indeed the value of $\varphi(x)$ does not change the martingale optimality of $F_x$. We provide these important formulas.

**The map $\hat{V}_\varepsilon$ and its derivatives:** Let $\psi \in \mathbb{R}^Y$, we denote $(\hat{\varphi}_\psi, \hat{h}_\psi) := \text{argmin}_{\varphi, h} V_\varepsilon(\varphi, \psi, h)$, that are unique and may be found in quasi-closed form from (3.3). Now we give the formula for $\hat{V}_\varepsilon$ and its derivatives. We directly get from Proposition 3.2 that

$$
\hat{V}_\varepsilon(\psi) = V_\varepsilon \left( \hat{\varphi}_\psi, \hat{\psi}, \hat{h}_\psi \right),
$$

$$
\nabla \hat{V}_\varepsilon(\psi) = \partial \psi \left( \hat{\varphi}_\psi, \psi, \hat{h}_\psi \right),
$$

$$
D^2 \hat{V}_\varepsilon = \left( \partial^2 \hat{V}_\varepsilon \varepsilon \right) \left( \hat{\varphi}_\psi \right) \left( \hat{\psi}, \hat{h}_\psi \right).
$$

The last additive decomposition of $\partial \psi \hat{V}_\varepsilon^{-1}$ stems from the fact that $\partial \psi \hat{V}_\varepsilon \left( \hat{\varphi}_\psi, \psi, \hat{h}_\psi \right)$ is diagonal. Indeed, $V_\varepsilon$ is a sum of functions of $(\varphi(x), h(x))$ for $x \in \mathcal{X}$, and the crossed derivative $\partial \psi(h(x)) \hat{V}_\varepsilon = \sum_y (y - x) \exp \left( -\frac{\Delta (x,y)}{\varepsilon} \right)$ cancels at $\left( \hat{\varphi}_\psi(x), \hat{h}_\psi(x) \right)$ by the martingale property induced by the optimality in $h(x)$. The same holds for $\partial h_i(x) \hat{V}_\varepsilon$ for $i \neq j$. We denote $\Delta_\psi := \hat{\varphi}_\psi + \hat{h}_\psi - c$, and we have

$$
\hat{V}_\varepsilon(\psi) = \mu \left[ \hat{\varphi}_\psi \right] + \nu[\psi] + \varepsilon \sum_{x,y} \exp \left( -\frac{\Delta_\psi(x,y)}{\varepsilon} \right),
$$

$$
\nabla \hat{V}_\varepsilon(\psi) = \left( \nu_y - \sum_x \exp \left( -\frac{\Delta_\psi(x,y)}{\varepsilon} \right) \right)_{y \in Y},
$$

$$
D^2 \hat{V}_\varepsilon = \varepsilon^{-1} \text{diag} \left( \sum_x \exp \left( -\frac{\Delta_\psi(x,y)}{\varepsilon} \right) \right)_{y \in Y},
$$

$$
-\varepsilon^{-1} \sum_x \left( \sum_y \exp \left( -\frac{\Delta_\psi(x,y)}{\varepsilon} \right) \right)^{-1} \times \left( \exp \left( -\frac{\Delta_\psi(x,y_1)}{\varepsilon} \right) \exp \left( -\frac{\Delta_\psi(x,y_2)}{\varepsilon} \right) \right)_{y_1,y_2 \in Y},
$$

$$
-\varepsilon^{-1} \sum_x \sum_{i=1}^d \left( \sum_y (y_i - x_i)^2 \exp \left( -\frac{\Delta_\psi(x,y)}{\varepsilon} \right) \right)^{-1} \times \left( (y_1 - x_1) \exp \left( -\frac{\Delta_\psi(x,y_1)}{\varepsilon} \right) (y_2 - x_1) \exp \left( -\frac{\Delta_\psi(x,y_2)}{\varepsilon} \right) \right)_{y_1,y_2 \in Y}.
$$

Notice that for the conjugate gradient algorithm, we only need to be able to compute the product $(D^2 \hat{V}_\varepsilon) p$ for $p \in \mathbb{R}^Y$. Then if the RAM is not sufficient to store the whole matrix $D^2 \hat{V}_\varepsilon$, it may be convenient to only store $D_\psi := \text{diag} \left( \sum_x \exp \left( -\frac{\Delta_\psi(x,y)}{\varepsilon} \right) \right)$, $D_\varphi^{-1} := \text{diag} \left( \frac{-\Delta_\psi(x,y)}{\varepsilon} \right)$, and $D_{h_i}^{-1} := \text{diag} \left( \sum_y (y_i - x_i)^2 \exp \left( -\frac{\Delta_\psi(x,y)}{\varepsilon} \right) \right)^{-1}$ for all $1 \leq i \leq d$. 


\[ i \leq d. \text{ Then we compute } \left( D^2 \tilde{V}_\varepsilon \right) p \text{ in the following way:} \]

\[
\begin{align*}
p_\psi & := (D_\psi) p \\
p_\varphi & := \left( \sum_y \exp \left( -\frac{\Delta_\varepsilon^\psi(x, y)}{\varepsilon} \right) p_y \right)_{x \in \mathcal{X}}, \\
p_{hi} & := \left( \sum_y (y_i - x_i) \exp \left( -\frac{\Delta_\varepsilon^\psi(x, y)}{\varepsilon} \right) p_y \right)_{x \in \mathcal{X}}, \\
p'_\varphi & := (D_\phi^{-1}) p_\varphi, \\
p'_{hi} & := (D_{hi}^{-1}) p_{hi}, \\
p''_\varphi & := \left( \sum_x \exp \left( -\frac{\Delta_\varepsilon^\psi(x, y)}{\varepsilon} \right) (p''_\phi)_x \right)_{y \in \mathcal{Y}}, \\
p''_{hi} & := \left( \sum_x (y_i - x_i) \exp \left( -\frac{\Delta_\varepsilon^\psi(x, y)}{\varepsilon} \right) (p''_{hi})_x \right)_{y \in \mathcal{Y}}, \\
\end{align*}
\]

\[
\left( D^2 \tilde{V}_\varepsilon \right) p = \varepsilon^{-1} \left( p_\psi - p_\varphi - \sum_{i=1}^d p''_{hi} \right).
\]

**The map \( \tilde{F}_x \) and its derivatives:** In this paragraph we fix \( \psi \in \mathbb{R}^\mathcal{Y} \) and \( \varepsilon > 0 \). Recall the map \( F_x \) from (3.4). This map may be seen as a function of \((\varphi(x), h(x))\). Then we set

\[
\tilde{F}_x(h) := \min_{\varphi(x) \in \mathbb{R}} \mu_x \varphi(x) + \varepsilon \sum_y \exp \left( -\frac{\varphi(x) + h \cdot (y - x) - c(x, y)}{\varepsilon} \right).
\]

The optimizer is given by (3.3), hence by the closed formula

\[
\varphi_h(x) := \arg\min_{\varphi(x) \in \mathbb{R}} \mu_x \varphi(x) + F_x(\varphi(x), h) = \varepsilon \ln \left( \frac{1}{\mu_x} \sum_y \exp \left( -\frac{\varphi(y) + h \cdot (y - x) - c(x, y)}{\varepsilon} \right) \right).
\]

A direct application of Proposition 3.2 gives

\[
\begin{align*}
\tilde{F}_x(h) & = \min_{\varphi(x) \in \mathbb{R}} \mu_x \varphi_h(x) + \varepsilon \sum_y \exp \left( -\frac{\hat{\Delta}_h(x, y)}{\varepsilon} \right), \\
\nabla \tilde{F}_x(h) & = -\sum_y (y - x) \exp \left( -\frac{\hat{\Delta}_h(x, y)}{\varepsilon} \right), \\
D^2 \tilde{F}_x(h) & = \varepsilon^{-1} \left( \sum_y (y - x)^{2\otimes} \exp \left( -\frac{\hat{\Delta}_h(x, y)}{\varepsilon} \right) - \mu_x^{-1} \left( \sum_y (y - x) \exp \left( -\frac{\hat{\Delta}_h(x, y)}{\varepsilon} \right) \right)^{2\otimes} \right),
\end{align*}
\]

where we denote \( \hat{\Delta}_h(x, y) := \varphi_h(x) + \psi(y) + h \cdot (y - x) - c(x, y) \), and \( u^{2\otimes} := u \otimes u \) for \( u \in \mathbb{R}^d \).
4 Solutions to practical problems

4.1 Preventing numerical explosion of the exponentials

As we want to make $\varepsilon$ go to $0$, all the terms like $\exp\left(\frac{\varepsilon}{\delta}\right)$ tend to explode numerically. Here are the different risks that we have to deal with, and how we deal with them.

First the Newton algorithm is very local, and nothing guarantees that after one iteration, the value function will not explode. From our practical experience, the algorithm tends to explode for $\varepsilon \leq 10^{-3}$. Notice that the numerical experiment given by [16] does not go beyond $10^{-3}$, we may imagine that this is because they do not use the variable implicitation technique. Furthermore, we notice from our numerical experimenting that this variable implicitation, additionally to the stabilizing the numerical scheme, makes the convergence of the Newton algorithm much faster. Moreover, impliciting in $\varphi$ and $h$ is much more effective than impliciting in $\psi$, even though we have to do the implicitation in $h$ which is much more costly than the implicitation in $\psi$.

For the computation of the implicitations [3.3], the computation of the formula $\varphi(x) = \varepsilon \ln \left( \frac{1}{\mu} \sum_y \exp \left( -\frac{\psi(y) + h(x)(y-x) - c(x,y)}{\varepsilon} \right) \right)$ should be done as follows to prevent numerical explosion. First we compute $M_x := \max_{y \in Y} \left\{ -\frac{\psi(y) + h(x)(y-x) - c(x,y)}{\varepsilon} \right\}$, and then the computation that we do effectively is

$$\varphi(x) = M_x + \varepsilon \ln \left( \sum_y \exp \left( -\frac{\psi(y) + h(x)(y-x) - c(x,y)}{\varepsilon} - M_x \right) \right) - \varepsilon \ln \mu_x$$  \hspace{1cm} (4.5)

In (4.5), the exponential arguments are always smaller than 1, and one of them is equal to 1, then any explosion makes the exponential be totally negligible when compared to $\exp(0) = 1$, this computation rule makes it very stable. Notice also the separation of $\ln \mu_x$ that allows to treat the case when the value of $\mu_x$ is extremely low (like for example when you discretise a Gaussian measure on a grid) even if in this case, it may be smarter to just remove the value from the grid.

Notice that the variable implicitation should also be used during each partial optimisation in $h(x)$ for $x \in X$, as this Newton algorithm is highly susceptible to explode as well. The implicitation simply consists in minimizing in $\varphi(x)$ the maps $F_x$ from [3.4], and has a closed form.

Another thing to take care of about $h$ is the initial value taken for the next partial optimization of $V_\varepsilon$ in $h$. On a first hand, choosing the last value for $h$ helps diminishing the number of steps for the optimization. Also, when $\varepsilon$ is very small, even with the implicitation, the Newton optimization may get hard if the initial value is too far from the optimum.
4.2 Customization of the Newton method

4.2.1 Preconditioning

The conjugate gradient algorithm used to compute the search direction for the Newton algorithm has a convergence rate given by

\[ |x_k - x^*|_A \leq 2 \left( \frac{\kappa(A)^{-1}}{\sqrt{\kappa(A)^{-1} + 1}} \right)^k |x_0 - x^*|_A, \]

where \( x_k \) is the \( k \)-th iterate, \( x^* \) is the solution of the problem, \( |x|_A := x^t A x \) is the Euclidean norm associated to the scalar product \( A \), and \( \kappa(A) := \|A\| \|A^{-1}\| \) is the conditioning of \( A \). This conditioning is the fundamental parameter for this convergence speed. When \( \varepsilon \) is getting small, the conditioning raises. We also observe on the numerics that is happens when the marginals have a thin tail (e.g. Gaussian distributions). The simplest way of dealing with this conditioning problem consists in applying a "preconditioning" algorithm. We find a matrix \( P \) that is easy to invert (for example a diagonal matrix) and we use the fact that solving \( Ax = b \) is equivalent with solving \( P^t A P x' = P^t b \), where \( x' := P^{-1} x \). We use the most classical and simple preconditioning which consists in taking \( P := \sqrt{\text{diag}(A)^{-1}} \). See [62] for the precise algorithm.

4.2.2 Line search

An important advantage of the Bregman projection algorithm over the primitive Newton algorithm is that \( V_\varepsilon \) is a Lyapunov function as the steps only consist of block minimizations of this function, whereas the Newton step may get very wrong and lose the optimal region if we are not close enough to the minimum. However in practice, some ingredients need to be added to the Newton step. Indeed, once the direction of search is decided by the conjugate gradient algorithm, in practice it is necessary to make a line search algorithm, i.e. to find a point on the line on which the value function \( V_\varepsilon \) is strictly smaller, and so does the directional gradient absolute value \( |\nabla V_\varepsilon \cdot p| \), where \( p \) is the descent direction. This 'descent' condition is called the Wolfe condition. A very good line search algorithm that is commonly used in practice is detailed in [62].

**Remark 4.1.** Notice that if a value is rejected by the line search, it is important to throw away the value of \( h \) given by this wrong point, and to come back to the last value of \( h \) corresponding to a point that was not rejected by the line search.

4.3 Penalization

4.3.1 The penalized problem

The dual solution may not be unique, which may lead to numerical unstabilities. As an example we may add any constant to \( \varphi \) while subtracting it to \( \psi \) without changing the value of \( V \). A straightforward solution is to add a penalization to the minimization problem. I.e. we have the new problem

\[ \min_{\psi \in \mathbb{R}^n} \tilde{V}_\varepsilon(\psi) + \alpha f(\psi) \quad (4.6) \]
where \( f \) is a strictly convex superlinear function, so that there is a unique minimum by the fact that the gradient of \( \tilde{V}_\varepsilon \) is a difference of probabilities, which proves that this convex function is Lipschitz, whence the strict convexity and super-linearity of \( \tilde{V}_\varepsilon(\psi) + \alpha f(\psi) \). In practice we take \( f(\psi) := \frac{1}{2} \sum_{y \in \mathcal{Y}} a_y \psi(y)^2 \), for some \( a \in \mathbb{R}^\mathcal{Y} \), so that \( \nabla f(\psi) = \sum_{y \in \mathcal{Y}} a_y \psi(y) e_y \), where \((e_y)_{y \in \mathcal{Y}}\) is the canonical basis, and \( D f(\psi) = \text{diag}(a) \) have these easy closed expressions. In practice we take \( a = (1), a = \nu, a = \nu^2 \), or \( a = \nu/\psi_0 \), where \( \psi_0 \) is a fixed estimate of \( \psi \) from the last step of \( \varepsilon \)-scaling (see Subsection 4.4).

### 4.3.2 Marginals not in convex order

Problem 4.6 allows to solve the problem of mismatch in the convex ordering thanks to the following theorem that allows for probability measures \( \mu, \nu \) not in convex order to find another probability measure \( \tilde{\nu} \) in convex order with \( \mu \) that satisfies some optimality criterion, for example in terms of distance from \( \nu \).

**Theorem 4.2.** Let \((\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})\) not in convex order. Let \( \nu_\alpha := \mathbb{P}_\alpha \circ Y^{-1} \), where \( \mathbb{P}_\alpha \) is the optimal probability for Problem (4.6), where \( f \) is a super-linear, differentiable, strictly convex, and \( p \)-homogeneous function \( \mathbb{R}^\mathcal{Y} \rightarrow \mathbb{R} \) for some \( p > 1 \). Then \( \nu_\alpha \rightarrow \nu_l \) when \( \alpha \rightarrow 0 \), for some \( \nu_l \succeq c \mu \) satisfying

\[
f^*(\nu_l - \nu) = \min_{\tilde{\nu} \succeq c \mu} f^*(\tilde{\nu} - \nu).
\]

The proof of Theorem 4.2 is reported in Subsection 6.2. Notice that for \( f(\psi) := \frac{1}{2} \sum_{y \in \mathcal{Y}} a_y \psi(y)^2 \), we have \( f^*(\gamma) = \frac{1}{2} \sum_{y \in \mathcal{Y}} a_y \gamma(y)^2 \), whence the idea of taking \( a = \nu^2 \).

### 4.3.3 Conjugate gradient improvement and stabilization

Adding a penalization also allows to accelerate the conjugate gradient algorithm, indeed it reduces the conditioning of the Hessian matrix by killing the small eigenvalues, and therefore accelerates the conjugate gradient algorithm’s convergence. It also stabilizes this algorithm, indeed when \( \varepsilon \) is small we observe that without penalization, the numerical error may cause instabilities by returning a non positive definite Hessian. Adding the positive definite Hessian of the penalization function bypasses this instability.

### 4.4 Epsilon scaling

For all entropic algorithms, we observe that when \( \varepsilon \) is small, the algorithm may be very slow to find the region of optimality. For the Bregman projection, the formula for the speed of convergence in Subsection 5.4.3 suggests to have a strategy of \( \varepsilon \)-scaling: i.e. we solve the problem for \( \varepsilon = 1 \), so that the function to optimize is very smooth. Then solve the problem for \( \varepsilon' < \varepsilon \), with the previous optimum as an initial point. We continue this algorithm until we reach the desired value for \( \varepsilon \). In practice we divide \( \varepsilon \) by 2 at each step.
4.5  Grid size adaptation

It may be a huge loss of time to run the algorithm on full resolution since the beginning of $\varepsilon$-scaling. To prevent this waste of time, Schmitzer [52] suggests to raise the size of the grid at the same time than shrinking $\varepsilon$. In practice we give to each new point of the grid for $\varphi$, $\psi$, and $h$ the value of the closest point in the previous grid. We use heuristic criteria to decide when to double the size of the grid, avoiding for example to double is when $\varepsilon$ is too small as is seriously challenges the stability of the resolution scheme.

4.6  Kernel truncation

While $\varepsilon$ shrinks to 0, we observe that the optimal transport tends to concentrate on graphs, as suggested in [25]. Because of the exponential, the value of the optimal probability far enough to these graphs tends to become completely negligible. For this reason, Schmitzer [52] suggests to truncate the grid in order to do much less calculation. In dimensions higher than 1, the gain in term of number of operation may quickly reach a factor 100 for small $\varepsilon$. In practice we removed the points in the grid when their probability were smaller than $10^{-7}\mu_x$ (resp. $10^{-7}\nu_y$) for all $x \in \mathcal{X}$ (resp. for all $y \in \mathcal{Y}$).

4.7  Computing the concave hull of functions

We were not able to find algorithms that compute the concave hull of a function in the literature, so we provide here the one we used. Let $f : \mathcal{Y} \rightarrow \mathbb{R}$.

In dimension 1 the algorithm is linear in $|\mathcal{Y}|$, we use the McCallum & Avis [45] algorithm to find the points of the convex hull of the upper graph of $f$ in a linear time and then we go through these points until we find the two consecutive points $y_1, y_2 \in \mathcal{Y}$ around the convex hull such that $y_1 < x \leq y_2$. Then $f_{\text{conc}}(x) = \frac{y_2-x}{y_2-y_1} f(y_1) + \frac{x-y_1}{y_2-y_1} f(y_2)$.

In higher dimension we use Algorithm 1 in order to compute the convex hull of a function. We do not know if a better algorithm exists, but this one should be the fastest when the active points of the convex hull are already close to the maximum, this will be useful to compute $(c(x, \cdot) - \psi)_{\text{conc}}(x)$ from Theorem 5.5 below, so as the field "gradient" of the result that allows to find the right $h$. We believe that the complexity of this algorithm is quadradic in the (not so improbable) worst case of a concave function, $O(n \ln(n))$ on average for a "random" function, and linear when the guess of the gradient is good. These assesments are formal and based on the observation of numerics, we do not prove anything about Algorithm 1 not even the fact that it cannot go on infinite loops. We provide it for the reader who would like to reproduce the numerical experiments without having to search for an algorithm by himself.
Algorithm 1 Concave hull of $f$.

1: procedure CONCAVEHULL($f, x, grid, gradientGuess$)
2:  \hspace{1em} \textbf{if} gradientGuess is None \textbf{then}
3:  \hspace{2em} grad $\leftarrow$ vector of zeros with the same size than $x$
4:  \hspace{2em} gridF $\leftarrow$ $f$(grid)
5:  \textbf{else}
6:  \hspace{2em} grad $\leftarrow$ gradientGuess
7:  \hspace{2em} gridF $\leftarrow$ $f$(grid) $-$ grad $\cdot$ grid
8:  \hspace{2em} $y \leftarrow \text{argmax}_{gridF}$
9:  \hspace{2em} support $\leftarrow$ [$y$]
10: \hspace{2em} gridF $\leftarrow$ gridF $-$ gridF[$y$]
11: \hspace{2em} \textbf{while} True \textbf{do}
12:  \hspace{3em} \textbf{if} $x \in \text{aff support}$ \textbf{then}
13:  \hspace{4em} bary $\leftarrow$ barycentric coefficients of $x$ in the basis support
14:  \hspace{4em} \textbf{if} bary are all $> 0$ \textbf{then}
15:  \hspace{5em} value $\leftarrow$ sum(bary $\times$ $f$(support))
16:  \hspace{5em} \textbf{return} \{"value" : value,"support" : support;
17:  \hspace{5em} "barycentric coefficients" : bary,"gradient" : grad\}
18:  \hspace{4em} \textbf{else}
19:  \hspace{5em} i $\leftarrow$ argmin bary
20:  \hspace{5em} remove entry $i$ in support
21:  \hspace{5em} remove entry $i$ in bary
22:  \hspace{4em} \textbf{else}
23:  \hspace{5em} projx $\leftarrow$ orthogonal projection of $x$ on aff support
24:  \hspace{5em} $p = x - \text{projx}$
25:  \hspace{5em} scalar $\leftarrow$ $p \cdot (grid - x)$
26:  \hspace{4em} \textbf{if} scalar are all $\leq 0$ \textbf{then}
27:  \hspace{5em} Fail with error "x not in the convex hull of grid."
28:  \hspace{4em} y $\leftarrow$ argmax{gridF/scalar such that scalar $> 0$}
29:  \hspace{4em} add $y$ to support
30:  \hspace{4em} a $\leftarrow$ $-gridF[y]/\text{scalar}[y]$
31:  \hspace{4em} gridF $\leftarrow$ gridF + $a$ $\times$ scalar
32:  \hspace{4em} grad $\leftarrow$ grad $-$ $a$ $\times$ $p$
5 Convergence rates

5.1 Discretization error

Proposition 5.1. Let \( \mu \leq \nu \) in convex order in \( \mathcal{P}(\mathbb{R}^d) \) having a dual optimizer \((\varphi, \psi, h) \in \mathcal{D}_{\mu,\nu}(c) \) such that \( \varphi \) is \( L_\varphi \)-Lipschitz, and \( \psi \) is \( L_\psi \)-Lipschitz. Then for all \( \mu' \leq \nu' \) in convex order in \( \mathcal{P}(\mathbb{R}^d) \) having a dual optimizer \((\varphi', \psi', h') \in \mathcal{D}_{\mu,\nu}(c) \) such that \( \varphi' \) is \( L_{\varphi'} \)-Lipschitz, and \( \psi' \) is \( L_{\psi'} \)-Lipschitz, we have

\[
|S_{\mu,\nu}(c) - S_{\mu',\nu'}(c)| \leq \max (L_{\varphi}, L_{\varphi'}) W_1(\mu', \mu) + \max (L_{\psi}, L_{\psi'}) W_1(\nu', \nu)
\]

The proof of Proposition 5.1 is reported in Subsection 6.3

Remark 5.2. Guo & Obłój [31] provide a very similar result in Proposition 2.2. It is however very different as they need to introduce an approximately martingale optimal transport problem, and our result makes hypotheses on the Lipschitz property of the dual optimizers, which are unknown, and even their existence in unknown. In dimension 1, thanks to the work by Beiglböck, Lim & Obłój [8], we may prove the existence of these Lipschitz dual, thanks to some regularity assumptions on c. In higher dimension, there are ongoing investigations about the existence of similar results. However, by Example 4.1 in [26], it will be necessary to make assumptions on \( \mu, \nu \) as well, as the smoothness of \( c \) cannot guarantee the existence of a dual optimizer. Proposition 5.1 is entitled to be a proposition of practical use, we may formally assume that the partial dual functions that we get converge to the continuous dual and assume that their Lipschitz constant converges to the Lipschitz constant of the limit.

We refer to Subsection 2.2 in [31] for a study of the discrete \( W_1 \)-approximation of the continuous marginals. In dimensions higher than 3, it is necessary to use a Monte-Carlo type approximation of \( \mu \) and \( \nu \) to avoid the curse of dimensionality linked to a grid type approximation. However, Proposition 5.1 is not well-adapted to estimate the error, as we know from [28] that the Wasserstein distance between a measure and its Monte-Carlo estimate is of order \( n^{-1/2} \). Next proposition deals with this issue. For two sequences \((u_N)_{N \geq 0}\) and \((v_N)_{N \geq 0}\), we denote \( u_N \approx v_N \) when \( N \to \infty \) if \( u_N/v_N \) converges to 1 in probability, when \( N \to \infty \).

Proposition 5.3. Let \( \mu \leq \nu \) in convex order in \( \mathcal{P}(\mathbb{R}^d) \) having a dual optimizer \((\varphi, \psi, h) \in \mathcal{D}_{\mu,\nu}(c) \), and \( \mu_N \) and \( \nu_M \), independent Monte-Carlo estimates of \( \mu \) and \( \nu \) with \( N \) and \( M \) samples. If furthermore \( \mu_N' \leq \nu_M' \) is in convex order in \( \mathcal{P}(\mathbb{R}^d) \) having a dual optimizer \((\varphi_N', \psi_M', h') \in \mathcal{D}_{\mu_N',\nu_M'}(c) \) such that when \( N, M \to \infty \), we have

(i) \( (\mu_N - \mu)[\varphi] \approx (\mu_N - \mu)[\varphi_N] \), \( (\nu_M - \nu)[\psi] \approx (\nu_M - \nu)[\psi_M] \),

(ii) \( (\mu_N - \mu_N)[\varphi] \approx (\mu_N - \mu_N)[\varphi_N] \), \( (\nu_M - \nu_M')[\psi] \approx (\nu_M - \nu_M')[\psi_M] \).

Then we have

\[
|S_{\mu,\nu}(c) - S_{\mu',\nu'}(c)| \leq \alpha \sqrt{\frac{\text{Var}_\mu[\varphi]}{N} + \frac{\text{Var}_\nu[\psi]}{M} + (\mu_N - \mu_N')[\varphi_N] + (\nu_M - \nu_M')[\psi_M]},
\]

with probability converging to \( 1 - 2\int_0^\infty \exp(-x^2/2) \, dx \), when \( N, M \to \infty \).
The proof of Proposition 5.3 is reported in Subsection 6.3

**Remark 5.4.** In Proposition 5.3, we introduce \( \mu_N, \nu_M \) because the Monte-Carlo approximation will not conserve the convex order for \( \mu_X, \nu_N \) in general. Then we obtain \( \mu_N', \nu_M' \) from "convex ordering processes" such as the one suggested in Subsection 4.3.2 or the one suggested in [3]. In both cases, the quantity \( |(\mu_N - \mu_N')[\varphi_N] + (\nu_M - \nu_M')[\psi_M]| \) may be computed exactly numerically.

### 5.2 Entropy error

In this subsection, \( m_\varepsilon \) is a generic finite measure and no assumptions are made on \( \mu_\varepsilon \) and \( \nu_\varepsilon \).

**Theorem 5.5.** Let \( \mu_\varepsilon \preceq \nu_\varepsilon \) in convex order in \( \mathcal{P}(\mathbb{R}^d) \) with dual optimizer \((\varphi_\varepsilon, \psi_\varepsilon, h_\varepsilon) \in \mathcal{D}_{\mu_\varepsilon, \nu_\varepsilon}(c)\) to the \( \varepsilon \)-entropic dual problem with reference measure \( m_\varepsilon \), such that we may find \( \gamma, \eta, \beta > 0 \), sets \( (D_\varepsilon^X, D_\varepsilon^Y)_{\varepsilon>0} \subset B(\mathbb{R}^d) \), and parameters \((\alpha_\varepsilon, A_\varepsilon)_{\varepsilon>0} \subset (0, \infty)\), such that if we denote \( r_\varepsilon := \varepsilon^{\frac{1}{2}-\eta} \), \( m_\varepsilon := \mu \circ O^{-1} \), and \( \Delta_\varepsilon := \varphi \oplus \psi + h^\Theta - c \), for \( \varepsilon > 0 \) small enough we have:

(i) \( \frac{dp_{\mu_\varepsilon}}{dm_{\varepsilon}}(x) \leq \varepsilon^{-\eta} \), \( \mu_\varepsilon - a.e., \mu_\varepsilon[\Omega] \leq \varepsilon^{-\gamma} \), and \( A_\varepsilon \ll \varepsilon^{-\eta} \) for all \( q > 0 \).

(ii) For \( m_\varepsilon - a.e. \), \( (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \), we have \( (m_\varepsilon)_x \left[ B_{A_\varepsilon}(y) \right] \geq \varepsilon^{\gamma} \), and for \( (m_\varepsilon)_x - a.e. \) \( y' \in B_{A_\varepsilon}(y) \), we have \( |\Delta_\varepsilon(x, y) - \Delta_\varepsilon(x, y')| \leq \gamma \varepsilon \ln(e^{-1}) \).

(iii) \( \mu_\varepsilon \left[ (D_\varepsilon^X)^c \right] \ll 1/\ln(e^{-1}) \) and for all \( x \in D_\varepsilon^X \) we may find \( k_\varepsilon \in \mathbb{N}, \lambda_\varepsilon \in (B_{A_\varepsilon})^{k_\varepsilon} \), and \( \lambda_\varepsilon \in [0, 1]^{k_\varepsilon} \) with \( \text{det}(S_\varepsilon^x) \geq A_\varepsilon^{-1} \), \( \min \lambda_\varepsilon \geq A_\varepsilon^{-1} \), and \( \sum_{l=1}^{k_\varepsilon} \lambda_\varepsilon^l \leq x = \text{convex combination} \).

(iv) On \( B_{r_\varepsilon}(S_\varepsilon^x) \), \( \Delta_\varepsilon(x, \cdot) \) is \( C^2 \), \( A_\varepsilon^{-1}I_d \geq \delta_\varepsilon^2 \Delta_\varepsilon(x, \cdot) \leq A_\varepsilon I_d \), and for all \( y, y' \in B_{r_\varepsilon}(S_\varepsilon^x) \), we have \( |\partial_y \Delta_\varepsilon(x, y) - \partial_y \Delta_\varepsilon(x, y')| \leq \varepsilon^{\gamma} \).

(v) For \( x \in D_\varepsilon^X \) and \( y \not\in B_{r_\varepsilon}(S_\varepsilon^x) \), we have that \( \Delta_\varepsilon(x, y) \geq \sqrt{\varepsilon} \text{dist}(y, S_\varepsilon^x) \).

(vi) \( \nu_\varepsilon \left[ (D_\varepsilon^Y)^c \right] \ll 1/\ln(e^{-1}) \) and for all \( y_0 \in D_\varepsilon^Y \), \( R, L \geq 1 \), and \( f : \mathbb{R}^d \to \mathbb{R}_+ \) such that \( f_{\bar{y}_\varepsilon} \) is \( L \)-Lipschitz, we have

\[
\left| \int_{B_R} f(y) \left[ \frac{d(m_\varepsilon)_x \circ \text{zoom}_{y_0}^\varepsilon}{(m_\varepsilon)_y} \right] - \frac{dy}{|B_R|^\varepsilon} \right| \leq [R + L] \varepsilon^{\beta} \int_{B_R} f(y) \frac{dy}{|B_R|^\varepsilon}.
\]

Then if we denote \( \varphi_\varepsilon := e^{-\Delta_\varepsilon} m_\varepsilon \), we have

\[
\lim_{\varepsilon \to 0} \frac{\mu_\varepsilon[\varphi_\varepsilon] + \nu_\varepsilon[\psi_\varepsilon] - \varphi_\varepsilon[c]}{\varepsilon} = \frac{d}{2}, \quad \text{where} \quad \varphi_\varepsilon := (c(X, \cdot) - \psi_\varepsilon)_{\text{conc}}(X).
\]

The proof of Theorem 5.5 is reported to Subsection 6.4.

**Corollary 5.6.** Under the assumptions of 5.5, we have that \( \varphi_\varepsilon[c] \geq S_{\mu, \nu}(c) - \frac{d}{2} \varepsilon + o(\varepsilon) \), when \( \varepsilon \to 0 \).

**Proof.** We fix \( h(x) \in \partial(c(x, \cdot) - \psi_\varepsilon)_{\text{conc}}(x) \) for all \( x \in \mathcal{X}, \) then

\[
((c(X, \cdot) - \psi_\varepsilon)_{\text{conc}}(X), \psi_\varepsilon, h) \in \mathcal{D}_{\mu_\varepsilon, \nu_\varepsilon}(c),
\]

and therefore \( \mu_\varepsilon[(c(X, \cdot) - \psi_\varepsilon)_{\text{conc}}(X)] + \nu_\varepsilon[\psi_\varepsilon] \geq I_{\mu_\varepsilon, \nu_\varepsilon}(c) \geq S_{\mu_\varepsilon, \nu_\varepsilon}(c) \geq \varphi_\varepsilon[c]. \) Theorem 5.5 concludes the proof. \( \square \)
Figure 1 gives numerical examples of the convergence of the duality gap when \( \varepsilon \) converges to 0. In these graphs, the blue curve gives the ratio of the dominator of the duality gap \( \mu_\varepsilon \left[ (c(X, \cdot) - \psi_\varepsilon)_{\text{conc}}(X) \right] + \nu_\varepsilon [\psi_\varepsilon] - \mathbb{P}_\varepsilon [c] \) with respect to \( \varepsilon \). It is meant to be compared to the green flat curve which is its theoretical limit \( \frac{d}{2} \) according to Theorem 5.5. Finally the orange curve provides the ratio of the weaker dominator \( \mu_\varepsilon \left[ \sup (c(X, \cdot) - \psi_\varepsilon) \right] + \nu_\varepsilon [\psi_\varepsilon] - \mathbb{P}_\varepsilon [c] \) of the duality gap with respect to \( \varepsilon \). The interest of this last weaker dominator is that it avoids computing the concave envelop which may be a complicated issue, while having a reasonable comparable performance in practice than the concave hull dominator as showed by the graphs and by Remark 5.13 below.

Figure 1a provides these curves for the one-dimensional cost function \( c := XY^2 \), \( \mu \) uniform on \([-1, 1]\), and \( \nu = |Y|^{1.5} \mu \). The grid size adaptation method is used and the size of the grid goes from 10 when \( \varepsilon = 1 \) to 10000 when \( \varepsilon = 10^{-5} \). Figure 1b provides these curves for the two-dimensional cost function \( c : (x, y) \in \mathbb{R}^2 \times \mathbb{R}^2 \mapsto x_1(y_1^2 + 2y_2^2) + x_2(2y_1^2 + y_2^2) \), \( \mu \) uniform on \([-1, 1]^2\), and \( \nu = (|Y_1|^{1.5} + |Y_2|^{1.5}) \mu \). The grid size adaptation method is used again and the size of the grid goes from 10 \times 10 \) when \( \varepsilon = 1 \) to 160 \times 160 \) when \( \varepsilon = 10^{-4} \).

Remark 5.7. The hypotheses of regularity are impossible to check at the current state of the art. One element that could argue in this direction is the fact that the limit of \( \Delta_\varepsilon \) when \( \varepsilon \to 0 \), that we shall denote \( \Delta_0 \), should satisfy the Monge-Ampère equation \( \det (\partial^2 \Delta_\varepsilon) = f \), similar to classical transport [61] that may provide some regularity, but less than the one needed to satisfy (ii) of Theorem 5.5, see Chapter 5 of [32]. It also justifies, in the case where \( f > 0 \), that \( \partial^2 \Delta_\varepsilon \in \text{GL}_d(\mathbb{R}) \).

Remark 5.8. Assumption (vi) on the local convergence of the reference measures is justified is we take a regular grid for \( \mathcal{Y} \) that becomes fine fast enough. If the grid does not become fine fast enough, then the local decrease of \( \Delta_\varepsilon - \Delta_\varepsilon (x) - \nabla \Delta_\varepsilon (x) \cdot (Y - x) \) is exponential because of the shape of the kernel. We observe on the experiments that in this case the duality gap becomes indeed smaller, however as a downside, the convergence of the scheme becomes much less effective because the marginal error stays high (see Proposition 5.1). See Figure 1b.

Remark 5.9. Assumption (ii) from the theorem seems to hold in one dimension, but it seems to be wrong in two dimensions, as shown in the example of Figure 4. However, we may still find a formula similar to (5.7) below. Therefore, we may reasonably assume that the error is still of order \( \varepsilon \), as confirmed in a the numerical examples by Figure 1b.

Remark 5.10. In the case when \( \mathcal{X} \) is obtained from Monte-Carlo methods, (vi) is verified with a constant that depends on the point, and is probabilistic. the exponent \( \alpha \) may be taken taken equal to \( \frac{1}{d} \), see [28].

Remark 5.11. Despite the difficulty to check the assumptions, this result is inspired and satisfied by observation on the numerics, we have tried with several cost functions, differentiable or not, and the result seems to be always satisfied, probably with the help of its universality. See figure 4.
Remark 5.12. An easier version of Theorem 5.5 could be obtained by the same strategy for classical optimal transport under the twist condition \(c_{x}(x, \cdot)\) injective, replacing \(c(X, \cdot) - \psi_{\varepsilon}\) by \(\sup (c(X, \cdot) - \psi_{\varepsilon})\).

Remark 5.13. By similar reasoning, we may prove that even for martingale optimal transport, replacing \(c_{x}(x, \cdot)\) with \(\psi_{\varepsilon}\), would still give a good result (see Figure 1). We may formally estimate the new limit of the duality gap \(\varepsilon\): for all \(y_{\varepsilon}(x)\) that are not the optimizer (say \(y_{0}(x)\)), the weight added to \(d_{2}\) is \(\lambda\varepsilon_{i_{x}}(x, y_{\varepsilon}(x))\). By using the tools of the proof of Theorem 5.5, we get the formal formula, if we denote \(\lambda_{i}\) for the limit of \(\lambda_{i}(x)\), and \(y_{i}\) for the limit of \(y_{i}(x)\), we have

\[
\mu_{\varepsilon}\left[\sup (c(X, \cdot) - \psi_{\varepsilon})(X)\right] + \nu_{\varepsilon}(\psi_{\varepsilon}) - P_{\varepsilon}[c] \approx \alpha\varepsilon, \quad (5.7)
\]

with \(\alpha = \frac{d}{\varepsilon} + \frac{1}{\varepsilon} \int_{\mathbb{R}^{d}} \sum_{i>0} \lambda_{i} \ln \left(\frac{dm^{\varepsilon}(y_{i})}{dm^{\varepsilon}(y_{0})}\right) \lambda_{i} \det \left(\frac{\Delta_{0}(y_{0})}{\Delta_{0}(y_{i})}\right) \} d\mu\). Then we could reasonably make the assumption the the second term in \(\alpha\) does not explode, and then the limit is still of the order of \(\varepsilon\), as we may see on the numerical experiments of Figure 1. This result also generalises to optimal transport when there are several transport maps.

5.3 Penalization error

Proposition 5.14. Let \((\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})\) in convex order. Let \(\nu_{\alpha} := \mathbb{P}_{\alpha} \circ Y^{-1}\), where \(\mathbb{P}_{\alpha}\) is the optimal probability for the entropic dual implied problem with an additional penalization \(\alpha f\), where \(f\) is a super-linear, strictly convex, and differentiable function \(\mathbb{R}^{\mathcal{Y}} \rightarrow \mathbb{R}\). Then let \(\psi_{0}\) be the only optimizer for the entropic dual implied problem with minimal \(f(\psi)\), we have

\[
\frac{\nu_{\alpha} - \nu}{\alpha} \rightarrow \nabla f(\psi_{0}), \quad \text{when} \quad \alpha \rightarrow 0.
\]

The proof of Proposition 5.14 is reported to Subsection 6.5.
5.4 Convergence rates of the algorithms

5.4.1 Convergence rate for the simplex algorithm

Precise results on the convergence rate of the simplex algorithm is an open problem. Roos [51] gave an example in which the convergence takes an exponential time in the number of parameters. However the simplex algorithm is much more efficient in practice, Smale [54] proved that in average, the number of necessary steps in polynomial in the number of entries, and Spielman & Teng [56] refined this analysis by including the number of constraints in the polynomial.

However, none of these papers provide the real time of convergence of this algorithm. Schmitzer [52] reports that this algorithm is not very useful in practice as it only allows to solve a discretized problem with no more than hundreds of points.

5.4.2 Convergence rate for the semi-dual algorithm

We notice that any subgradient of this function is a difference of probabilities, and then the gradient is bounded. Furthermore the function $V$ is a supremum of a finite number of affine functions, and therefore it does not have a smooth second derivative. In this condition the best theoretical way to optimize this function is by a gradient descent with a step size of order $O(1/\sqrt{n})$ at the $n$–th step, see Ben-Tal & Nemirovski [10]. Then by Theorem 5.3.1 of [10], the rate of convergence is $O(1/\sqrt{n})$ as well, which is quite slow. Furthermore, the time of computation of one step needs to compute one convex hull which has the average complexity $O(|\mathcal{Y}| \ln(|\mathcal{Y}|))$ for each $x \in \mathcal{X}$, and $O(|\mathcal{Y}|)$ in dimension 1, see Subsection 4.7. However, we give in Subsection 4.7 an algorithm that computes the concave hull in a linear time if the relying points of the concave hull do not change too much. Then let us be optimistic and assume that the computation of one concave hull is on average $O(|\mathcal{Y}|)$, then we have that the complexity is $O(|\mathcal{X}||\mathcal{Y}|)$ operations for each step. Although this algorithm is highly parallelizable, its complexity imposes to the grid to be very coarse. Indeed, to get a precision of $10^{-2}$, we need an order of $10^4$ operations. We shall see that the entropic algorithms are much more performing for this low precision.

Notice that even though the best theoretical algorithm is the last gradient descent, Lewis & Overton [44] showed that in most case, quasi-Newton methods converge faster, even when the convex function is non-smooth, however they find a particular case in which quasi-Newton fails at being better. The L-BFGS method is a quasi-Newton method that is adapted to high-dimensions problems. The Hessian (even though it does not exist) is approximated by a low-dimensional estimate, and the classical Newton step method is applied. See [62] for the exact algorithm. We see on simulations that this algorithm is indeed much more efficient.

Even if the quasi-Newton algorithm gives better results, the smooth entropic algorithms are much more effective in practice.
5.4.3 Convergence rate for the Sinkhorn algorithm

In practice, if we want to observe the transport maps from [25] to have a good precision on the estimation of the support of the optimal transport, we need to set $\epsilon = 10^{-4}$. The rate of convergence of the Sinkhorn algorithm is given by $\kappa^{2n}$ after $n$ iterations for some $0 < \kappa < 1$, see [39]. This result is extended by [31] to the one-dimensional martingale Sinkhorn algorithm. However, we have $\kappa := \sqrt{\theta_1}$, and

$$\theta := \max_{(x_1,y_1),(x_2,y_2) \in \Omega} \exp \left( \frac{c(x_1,y_1) + c(x_2,y_2) - c(x_2,y_1) - c(x_1,y_2)}{\epsilon} \right),$$

in the case of classical transport. Then $\theta$ is of the order of $\exp(K(c)/\epsilon)$ for some map $K(c)$ bounded from below. For $\epsilon = 10^{-5}$, this $\theta$ is so big that $\kappa^2$ is so close to 1 that $\kappa^{2n}$, with $n$ the number of iterations will remain approximately equal to 1.

We also see in practice for the martingale Sinkhorn algorithm that the rate of convergence is not exponential for small values of $\epsilon$, see Figure 2 in the numerical experiment part, as the graph is logarithmic in the error, an exponential convergence rate would be characterized by a straight line. However we observe that for the Bregman projection algorithm we do not have a straight line during the first part of the iteration for the one-dimensional case, and it never happens in the two-dimensional case.

In this regime of $\epsilon$ small, another convergence theory looks to have a better fit with this algorithm. The Sinkhorn algorithm may be interpreted as a block coordinates descent for the optimization of the map $V_\epsilon(\varphi, \psi)$. We optimize alternatively in $\varphi$, and in $\psi$. We know from Beck & Tetruashvili [5] that this optimization problem has a speed of convergence given by $\frac{LR(x_0)^2}{n}$, where $R(x_0)$ is a quantity that is of the order of $|x_0 - x^*|$ in practice, where $x^*$ is the closest optimizer of $V_\epsilon$ and $L$ is the Lipschitz constant of the gradient. This speed is more comparable to the convergence observed in practice. More precisely, $L$ is of the order of $1/\epsilon$.

This formula shows that in order to minimize the problem for a very small $\epsilon$, we first need to make $R(x_0)$ small to compensate $L$. This can be done by minimizing the problem for larger $\epsilon$. In practice, we divide $\epsilon$ by 2 until we reach a sufficiently small $\epsilon$. Then we make the grid finer as $\epsilon$ becomes small, and exploit the sparsity in the problem that appears when $\epsilon$ gets small. See Schmitzer [52].

We may apply the same theory for the martingale $V_\epsilon$ and its block optimization in $(\varphi, h)$ and in $\psi$. Let $\mathcal{D}_{X,Y} := \{(\varphi, \psi, h) \in \mathbb{R}^X \times \mathbb{R}^Y \times (\mathbb{R}^d)^Y \approx \mathbb{R}^{(d+1)|X|+|Y|}, and for $x := (\varphi, \psi, h) \in \mathcal{D}_{X,Y}$, let $\Delta(x) := (\varphi \oplus \psi + h \bar{\Omega})_{X \times Y}$.

**Theorem 5.15.** Let $x_0 = (\varphi_0, \psi_0, h_0) \in \mathcal{D}_{X,Y}$ such that $\left( e^{-\varphi_0 \oplus \psi_0 + h_0 \bar{\Omega} \cdot \cdot / \epsilon} \right)_{X \times Y}$ sums to 1, and for $n \geq 0$, let the $n^{th}$ iteration of the martingale Sinkhorn algorithm:

$$x_{n+1/2} := (\varphi_n, \psi_{n+1} := \arg\min_{\psi} V_\epsilon(\psi_n, \cdot, h_n, h_n),$$

$$x_{n+1} := (\varphi_{n+1} := \arg\min_{\varphi} V_\epsilon(\cdot, \psi_{n+1}, \cdot), \psi_{n+1}, h_{n+1} := \arg\min_h V_\epsilon(\cdot, \psi_{n+1}, \cdot)),$$
Furthermore let $P_0 \in \mathcal{M}(\mu, \nu)$ and let $X^*$ be the minimizing affine space of $V_\varepsilon$ and let $V_\varepsilon^*$ be its minimum, then we have

$$V_\varepsilon(x_n) - V_\varepsilon^* \leq \frac{\beta R(x_0)^2 \varepsilon^{-1}}{n},$$

(5.8)

$$V_\varepsilon(x_n) - V_\varepsilon^* \leq \left(1 - \frac{2|\mathcal{X}|}{\beta \lambda_2} \varepsilon^{-D(x_0)} \right)^n (V_\varepsilon(x_0) - V_\varepsilon^*),$$

(5.9)

and $\text{dist}(x_n, X^*) \leq \sqrt{\frac{\lambda \varepsilon^{\frac{D(x_0)}{\varepsilon}}}{|\mathcal{X}|} (V_\varepsilon(x_n) - V_\varepsilon^* \frac{1}{2}}$,}

where $\beta := (2 \ln(2) - 1)^{-1} \approx 2.6$, $\lambda_p := |\mathcal{X}|^{-1} \inf_{x \in D_{\mathcal{X}, \mathcal{Y}}: \Delta(x) \neq 0} \max_{\delta \in \Delta(x)} \sup_{p \in \mathbb{N}} \frac{\|\Delta(x)\|_p^p}{\|\delta\|_p^p}$, for $p = 1, 2$, $D(x_0) := \lambda_1 \max(1, \|\mathcal{Y} - \mathcal{X}\|_\infty) \frac{V_\varepsilon(x_0) - P_0 \varepsilon}{|\mathcal{X}|^{\min}}$, and $\|\cdot\|_\infty := \sup_{x \in \mathcal{X}} \|y - x\|_\infty$, 

$$\text{and we may choose } R(x_0) \text{ among } R(x_0) := \left\{ \begin{array}{ll}
\sup_{x \in \mathcal{X}} \text{dist}(x, X^*) \\
\sqrt{\frac{\lambda \varepsilon^{\frac{D(x_0)}{\varepsilon}}}{|\mathcal{X}|} (V_\varepsilon(x_0) - V_\varepsilon^* \frac{1}{2}}
\end{array} \right.$$

The proof of Theorem 5.15 is reported in Subsection 6.6.

Remark 5.16. By the same arguments, we may prove a similar theorem for the case of optimal transport with $\lambda_1 = \min\left(1, \frac{|\mathcal{Y}|}{|\mathcal{X}|}\right)$, and $\lambda_2 = \frac{1 + \left(\frac{|\mathcal{Y}|}{|\mathcal{X}|}\right)^2}{1 + \left(\frac{|\mathcal{Y}|}{|\mathcal{X}|}\right)}$.

Remark 5.17. The theoretical rate of convergence given by Theorem 5.15 becomes pretty bad when $\varepsilon \to 0$. We observe it in practice when we apply this algorithm with a small $\varepsilon$ and a starting point $x_0 = 0$. This emphasizes the need of using the epsilon scaling trick, of Subsection 4.4.

Remark 5.18. Depending on the experiment, in some cases we observe a linear convergence like [5.8] (see Figure 2a), however in other cases, we observe a convergence speed that looks more like [5.9] (see Figure 2b). However, the convergence rates that we provide here are generic, if we wanted to have convergence rates that look more like the one observed, we would need to look for the asymptotic convergence rates like it was suggested by Peyré in [37] for the case of classical transport.

Remark 5.19. The positive probability $P_0 \in \mathcal{M}(\mu, \nu)$ is necessary. We know from [27] that for some (possibly elementary) $\mu \leq \nu$ in convex order, we may find $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbb{P}[(x_0, y_0)] = 0$ for all $\mathbb{P} \in \mathcal{M}(\mu, \nu)$ even thought $\mu\{x_0\} > 0$ and $\nu\{y_0\} > 0$ (see Example 2.2 in [27]). Therefore, in this situation there is no optimal $x^* \in D_{\mathcal{X}, \mathcal{Y}}$, as this would mean that $\Delta(x^*)_{x_0, y_0} = \infty$. 

24
5.4.4 Convergence rate for the Newton algorithm

When the current point gets close enough from the optimum, the convergence rate of the Newton algorithm is quadratic if the hessian is Lipschitz, i.e. $|x_k - x^*|$ and $|\nabla V(x_k)|$ both converge quadratically to 0, see Theorem 3.5 in [62]. The truncated Newton is a bit slower, but still has a superlinear convergence rate, see Theorem 7.2 in [62].

5.4.5 Convergence rate for the implied Newton algorithm

The important parameter for Newton algorithm is the Lipschitz constant of the Hessian of the objective function. However in the case of variable implicitation, the presence of $B^2 y F^{-1}_1$ in the Hessian, and the addition of the variation of $y$ in the Lipschitz analysis may kill the Lipschitz property of the Hessian of $\tilde{F}$. The following proposition solves this problem.

**Proposition 5.20.** Let $(\tilde{x}_n)_{n \geq 0}$ the Newton iterations applied to $\tilde{F}$ starting from $\tilde{x}_0 := x_0 \in X$.

Now let $(x_n)_{n \geq 0}$ be the sequence defined by recurrence by $x_0 := x_0$, then for all $n \geq 0$, $y_n := y(x_n)$, and let $(x, y)$ be the result of a Newton step from $(x_n, y_n)$, and we set $x_{n+1} := x$.

Then $(\tilde{x}_n)_{n \geq 0} = (x_n)_{n \geq 0}$.

The proof of Proposition 5.20 is reported to Subsection 6.7. This proposition implies that the theoretical convergence of the Newton algorithm on $F$ can be extended to the Newton algorithm applied to $\tilde{F}$, indeed the partial minimization in $y$ only decreases the distance from the current point to the minimum around the minimum. In practice we observe that the convergence for this implied algorithm is much faster and much more stable than the non-implied Newton algorithm.

6 Proofs of the results

6.1 Minimized convex function

**Proof of Proposition 3.2** Let $x_1, x_2 \in A$, $y_1, y_2 \in B$, and $0 \leq \lambda \leq 1$. We have

\[
\lambda F(x_1, y_1) + (1 - \lambda) F(x_2, y_2) \geq F(\lambda(x_1, y_1) + (1 - \lambda)(x_2, y_2)) + \alpha \frac{\lambda(1-\lambda)}{2} |(x_1, y_1) - (x_2 - y_2)|^2 \\
\geq \tilde{F}(\lambda x_1 + (1 - \lambda)x_2) + \alpha \frac{\lambda(1-\lambda)}{2} |x_1 - x_2|^2.
\]

By minimizing over $y_1$ and $y_2$, we get

\[
\lambda \tilde{F}(x_1, y_1) + (1 - \lambda) \tilde{F}(x_2, y_2) \geq \tilde{F}(\lambda x_1 + (1 - \lambda)x_2) + \alpha \frac{\lambda(1-\lambda)}{2} |x_1 - x_2|^2,
\]

which establishes the $\alpha$–convexity of $\tilde{F}$.

Now if we further assume that $\alpha > 0$ and $F$ is $C^2$, $y \mapsto F(x, y)$ is $\alpha$–convex, and therefore strictly convex and super-linear. Hence, there is a unique minimizer $y(x)$. Using
the first order derivative condition of this optimum, we have $\partial_y F(x, y(x)) = 0$. By the fact that $\partial_y^2 F$ is positive definite (bigger than $\alpha I_d$ by $\alpha$-convexity), we may apply the local inversion theorem, which proves that $y(x)$ is $C^1$ in the neighborhood of $x$. We also obtain $\partial_{yx}^2 F(x, y(x)) + \partial_y^2 F(x, y(x)) \nabla y(x) = 0$, which gives the following expression of $\nabla y$:

$$\nabla y(x) = -\partial_y^2 F^{-1} \partial_{yx}^2 F(x, y(x)).$$

Now we may compute the derivatives of $\tilde{F}$. By definition, we have $\tilde{F}(x) = F(x, y(x))$, then just differentiating this expression, we get

$$\nabla \tilde{F}(x) = \partial_x F(x, y(x)) + \partial_y F(x, y(x)) \nabla y(x) = \partial_x F(x, y(x)),$$

where the second equality comes from the fact that $\partial_y F(x, y(x)) = 0$ because $y(x)$ is a minimizer. Finally we get the Hessian by deriving again this expression and injecting the value of $\nabla y(x)$.

\[ \square \]

6.2 Limit marginal

**Proof of Theorem 4.2** Let $\alpha > 0$, we are considering the following minimization problem:

$$\inf_{\psi} \tilde{V}_\varepsilon(\psi) + \alpha f(\psi) = \inf_{\varphi, \psi, h} \mu[\varphi] + \nu[\psi] + \varepsilon \int e^{-\Delta} dm_0 + \alpha f(\psi) = \inf_{\psi} \sup_{\mathbb{P} \in \mathcal{M}(\mu)} \mathbb{P}[c - \psi] - \varepsilon H(\mathbb{P}|m_0) + \nu[\psi] + \alpha f(\psi)$$

$$= \sup_{\mathbb{P} \in \mathcal{M}(\mu)} \inf_{\psi} \mathbb{P}[c] - \varepsilon H(\mathbb{P}|m_0) + (\nu - \mathbb{P} \circ Y^{-1})[\psi] + \alpha f(\psi)$$

$$= \sup_{\mathbb{P} \in \mathcal{M}(\mu)} \mathbb{P}[c] - \varepsilon H(\mathbb{P}|m_0) - (\alpha f)^*(\mathbb{P} \circ Y^{-1} - \nu)$$

$$= \sup_{\mathbb{P} \in \mathcal{M}(\mu)} \mathbb{P}[c] - \varepsilon H(\mathbb{P}|m_0) - \alpha^{-\frac{1}{p-1}} f^*(\mathbb{P} \circ Y^{-1} - \nu),$$

where the first equality comes from a mutualisation of the infima, the second comes from a partial dualisation of the infimum in $\varphi, h$ in a supremum over $\mathbb{P} \in \mathcal{M}(\mu)$, we obtain the third equality by applying the minimax theorem and reordering the terms, the fourth equality the definition of the Fenchel-Legendre transform, and the fifth and final equality is just a consequence of the transformation of a multiplier of a $p$-homogeneous function by the Fenchel-Legendre conjugate. Let $(\alpha_n)_{n \geq 1}$ converging to 0. As $\mathcal{Y}$ is finite, the set $\mathcal{P}(\mathcal{Y})$ is compact. Then we may assume up to extracting a subsequence that $\nu_{\alpha_n}$ converges to some limit $\nu_1$. The first order optimality equation for all $y \in \mathcal{Y}$ gives that $\nu - \nu_{\alpha_n} + \alpha_n \nabla f(\psi_n)$, where $\psi_n$ is the unique optimizer of $\tilde{V}_\varepsilon + \alpha f$. By the $p$-homogeneity of $f$, the gradient $\nabla f$ is $(p - 1)$--homogeneous. Then we have the convergence $\psi_n := \frac{\psi_n}{\alpha_n^{\frac{1}{p-1}}} \rightarrow n \rightarrow \infty \psi_1 := \nabla f^{-1}(\nu_1 - \nu)$. As we have by (6.10)

$$\inf_{\psi} \tilde{V}_\varepsilon(\psi) + \alpha_n f(\psi) = \inf_{\psi} \sup_{\mathbb{P} \in \mathcal{M}(\mu)} \mathbb{P}[c - \psi] - \varepsilon H(\mathbb{P}|m_0) + \nu[\psi] + \alpha_n f(\psi)$$

26
By dividing this equation by $\alpha_n^{-1}$, we have that $\psi_l$ is the minimizer of the strictly convex function $\sup_{\mu \in M(\mu)} L[\psi] + f(\psi)$, it is therefore unique. Then $\psi_l$ is unique as well. By (6.10), $P_{\alpha_n}$ tends to minimize $f^*(P \circ Y^{-1} - \nu)$, by the fact that $\psi_l = \lim_{\alpha \to 0} P_{\alpha} \circ Y^{-1}$, which concludes the proof.

6.3 Discretization error

Proof of Proposition 5.1 We have that $(\varphi, \psi, h)$ is a dual optimizer for $(\mu, \nu)$. Then $\varphi \oplus \psi + h^\circ \geq c$, and if $P^l \in M(\mu', \nu')$, then $P^l[c] \leq \mu'\varphi + \nu'\psi + L_\varphi W_1(\mu', \mu) + L_\psi W_1(\nu', \nu)$. If we take the supremum in $P^l$, we get that

$$S_{\mu', \nu'}(c) \leq \mu'[\varphi] + \nu'[\psi] + L_\varphi W_1(\mu', \mu) + L_\psi W_1(\nu', \nu)$$

As the reasoning may be symmetrical in $(\mu, \nu), (\mu', \nu')$, we get the result.

Proof of Proposition 5.3 Similar to the proof of Proposition 5.1, we have that

$$S_{\mu', \nu'}(c) - S_{\mu, \nu}(c) \leq (\mu' - \mu)[\varphi] + (\nu' - \nu)[\psi],$$

and

$$S_{\mu, \nu}(c) - S_{\mu', \nu'}(c) \leq (\mu - \mu')[\varphi] + (\nu - \nu')[\psi].$$

The first inequality gives

$$S_{\mu', \nu'}(c) - S_{\mu, \nu}(c) \leq (\mu' - \mu)[\varphi] + (\nu' - \nu)[\psi] + (\mu' - \mu)[\varphi] + (\nu' - \nu)[\psi].$$

The two first terms are independent and their sum $(\mu' - \mu)[\varphi] + (\nu' - \nu)[\psi]$ is equivalent in law to $\sqrt{\text{Var}_N[\varphi] + \text{Var}_M[\psi]}N(0, 1)$ when $N, M$ go to infinity. Then doing the same work on the symmetric inequality and using the Assumptions (i) to (iv), we get the result.

6.4 Entropy error

Lemma 6.1. Let $r, a > 0$, $F : \mathbb{R}^d \to \mathbb{R}^d$ and $x_0 \in \mathbb{R}^d$ such that $\|F^{-1}(x_0)\| \leq a^{-1}$, and on $B_r(x_0)$, we have that $F$ is $C^1$, that $\nabla F$ is invertible, and that $\|\nabla F - \nabla F(x_0)\| < a/2$. Then $F$ is a $C^1$-diffeomorphism on $B_r(x_0)$.

Proof. We claim that $F$ is injective on $B_r(x_0)$, we also have that $\nabla F$ is invertible on this set. Then by the global inversion theorem, $F$ is a $C^1$-diffeomorphism on $B_r(x_0)$.
Now we prove the claim that $F$ is injective on $B_r(x_0)$. Let $x, y \in B_r(x_0)$,

\[
F(y) - F(x) = \int_0^1 \nabla F(tx + (1 - t)y)(y - x) dt = \nabla F(x)(y - x) + \int_0^1 \left[ \nabla F(tx + (1 - t)y) - \nabla F(x) \right] (y - x) dt = \nabla F(x) \left( y - x + \nabla F(x)^{-1} \int_0^1 \left[ \nabla F(tx + (1 - t)y) - \nabla F(x) \right] (y - x) dt \right).
\]

Then we assume that $F(y) = F(x)$, and we suppose for contradiction that $x \neq y$. Therefore by the fact that $\nabla F$ is invertible, we have

\[
|y - x| = \left| \nabla F(x)^{-1} \int_0^1 \left[ \nabla F(tx + (1 - t)y) - \nabla F(x) \right] (y - x) dt \right| < \| \nabla F(x)^{-1} \| |2\frac{d}{2}|y - x| \leq |y - x|,
\]

Then we get the contradiction $|y - x| < |y - x|$. The injectivity is proved. 

In order to prove Lemma 6.3, we first need the following technical lemma.

**Lemma 6.2.** Let an integer $d \geq 1$, $k \leq d$, $r > 0$, and $(y_i)_{1 \leq i \leq k + 1}$, $k + 1$ differentiable maps $B_r \rightarrow \mathbb{R}^d$ such that $|\nabla y_i| \leq A$, $|y_i| \leq A$ and $\det_{\text{aff}}(y_1, \ldots, y_k) \geq A^{-1}$. Let $(u_i)_{1 \leq i \leq d}$ an orthonormal basis of $((y_i(0) - y_{k+1}(0))_{1 \leq i \leq k}) ^{\perp}$, $(e_i)_{1 \leq i \leq d}$ be the orthonormal basis obtained from $((y_i - y_{k+1})_{1 \leq i \leq k})^{\perp}$, $p$ the orthogonal projection of 0 on $\text{aff}(y_1, \ldots, y_{k+1})$, and $(\lambda_i)_{1 \leq i \leq k+1}$ the unique coefficients such that $p = \sum_{i=1}^{k+1} \lambda_i y_i$, barycentric combination. Then the maps $e_i$, $\lambda_i$, and $p$ are differentiable on $B_r$ and we may find $C, q > 0$, only depending on $d$ such that if $r \leq C^{-1} A^{-q}$, then we have that $|\nabla e_i| \leq CA^q$, $|\nabla p| \leq CA^q$, $|\nabla \lambda_i| \leq CA^q$, and $\nabla \det_{\text{aff}}(y_1, \ldots, y_{k+1}) \leq CA^q$.

**Proof.** The determinant is a polynomial expression of the coefficients, therefore if these coefficients are bounded by $A$. Then we may find $C_{\text{det}}, q_{\text{det}} > 0$ (only depending on $d$) such that $|\nabla \det| \leq C_{\text{det}} A^{q_{\text{det}}}$.

Let $(v_i)_{1 \leq i \leq d} := ((y_i - y_{k+1})_{1 \leq i \leq k}, (u_i)_{k+1 \leq i \leq d})$. Notice that for all $i$, we have $|\nabla v_i| \leq 2A$, and by the fact that

\[
\det_{\text{aff}}(y_1(0), \ldots, y_{k+1}(0)) := |\det(y_1(0) - y_{k+1}(0), \ldots, y_k(0) - y_{k+1}(0), u_{k+1}, \ldots, u_d)| \geq A^{-1},
\]

we have that $|\det(v_1, \ldots, v_d)| \geq \frac{1}{2} A^{-1}$ on $B_r$ for $r \leq C_{\text{det}}^{-1} A^{-q_{\text{det}}}$, $\frac{1}{2} A^{-1}$.

Recall that $e_1 := \frac{v_1}{|v_1|}$. By the fact that $|v_1| \cdots |v_d| \geq |\det(v_1, \ldots, v_d)|$, we have that $|v_1| \geq A^{-d}$.

Then we may find $C_1, q_1 > 0$ such that $\nabla e_1 \leq C_1 A^{q_1}$. Now for $1 \leq i \leq k$, we have that

\[
\nabla \lambda_i = \sum_{j=1}^{d} (\lambda_i)_{j} \nabla v_j \leq C_1 A^{q_1} \sum_{j=1}^{d} |v_j| \leq C_2 A^{q_1}.
\]

28
Let \( D : \mathbb{N} \to \mathbb{R}^d \) be defined as the Gram-Schmidt basis obtained from \( \{ e_i \} \). Notice that \( e_i := \frac{v_i - \sum_{j<i} (e_j \cdot v_i) e_j}{|v_i - \sum_{j<i} (e_j \cdot v_i) e_j|} \). Then we have that \( |v_i - \sum_{j<i} (e_j \cdot v_i) e_j| \geq A^{-d} \). Therefore, by induction, we may find \( C_i, q_i \) such that \( |\nabla e_i| \leq C_i A^{q_i} \).

Now notice that \( p := y_{k+1} + \sum_{i=1}^k e_i \cdot (0 - y_{k+1}) e_i \). Then we may find \( C_0, q_0 > 0 \) such that \( |\nabla p| \leq C_0 A^{q_0} \).

Finally let \( \lambda := (\lambda_1, ..., \lambda_{k+1}), M' := [e_i \cdot (y_j - y_{k+1})]_{1 \leq i \leq k, 1 \leq j \leq k+1}, M := \begin{bmatrix} 1...1 \end{bmatrix}^T, \) and \( P := \begin{bmatrix} 1 \ \frac{1}{p - y_{k+1}} \end{bmatrix} \). We have that \( M \lambda = P \), and therefore \( \lambda = M^{-1} P \). Recall that \( M^{-1} = \det(M)^{-1} \text{Com}(M)^T \) (see \([1],[2]\)), therefore we may find \( \lambda' / \lambda' \) such that \( |M^{-1}| \leq \lambda' A^q \), and \( |\nabla (M^{-1})| \leq \lambda' A^q \). Then we may find \( C'' \), \( q'' > 0 \) such that \( |\nabla \lambda_i| \leq C'' A^{q''} \) for all \( i \).

Finally, by the fact that
\[ \det(\text{aff}(y_1, ..., y_{k+1}) = |\det(y_1 - y_{k+1}, ..., y_k - y_{k+1}, e_{k+1}, ..., e_d)|, \]
we may find \( C''', q''' \) such that \( \det(\text{aff}(y_1, ..., y_{k+1}) \leq C''' A^{q'''} \).

The lemma is proved for
\[ C := \max(C_0, ..., C_d, C', C'', C''', 2C_{det}), \] and for \( q := \max(q_0, ..., q_d, q', q'', q''', q_{det} + 1) \).

\[ \square \]

**Lemma 6.3.** Let \( A, r, e, \delta, h, H > 0 \) and \( F : \mathbb{R}^d \to \mathbb{R} \) such that we may find \( k \in \mathbb{N} \) and \( S \in (B_1)^k \) such that for all \( y \in S \), we have on \( B_r(y) \) that \( F \) is \( C^2 \), \( A^{-1} I_d \leq D^2 F \leq A I_d \), and \( |D^2 F - D^2 F(y)| \leq e \). Furthermore, \( \det(\text{aff}(S)) \geq A^{-1} \), \( \nabla F(S) = \{0\} \), and \( \sum_{i=1}^k \lambda_i S_i \leq H \), convex combination with \( \min \lambda \geq A^{-1} \). Furthermore, assume that \( F(S) \subset [-h, h] \), and \( F \geq \delta \text{dist}(Y, S) \) on \( B_r(S)^c \). Then we may find \( C, q > 0 \) such that if \( \delta, r \geq CA^q H, e, H \leq C^{-1} A^{-q}, \) and \( h \leq rH \), then we have that \( (0, F_{\text{conv}}(0)) = \sum_{i=1}^k \lambda_i (\tilde{y}_i, F(\tilde{y}_i)), \) convex combination with \( |\tilde{y}_i - S_i| \leq CA^q H \) for all \( i \).

**Proof.**

**Step 1:** For all \( i \), the map \( y \mapsto \nabla F(y) \) is a \( C^1 \)-diffeomorphism on \( B_r(y_i) \) by Lemma 6.1. Then we define the map \( z_i(a) := \nabla F^{-1}(a) \) which is defined on \( B_{rA^{-1}} \). Notice that its gradient is given by \( \nabla z_i(a) := D^2 F^{-1}(a) \). Now we define the map \( \Phi : \mathbb{R}^d \to \mathbb{R}^d \) as follows: \( \Phi_i(a) := a \cdot (z_i(a) - z_{k+1}(a)) - (F(z_i(a)) - F(z_{k+1}(a))) \), for \( 1 \leq i \leq k \), and \( \Phi_k(a) := e_i(a) \cdot p(a) \), for \( k+1 \leq i \leq d \), where \( p(a) \) is the orthogonal projection of \( 0 \) on \( \text{aff}(z_1, ..., z_k) \), and \( (e_{k+1}(a), ..., e_d(a)) \) is the orthonormal basis of \( \text{aff}(z_1(a), ..., z_k(a))^\perp \) defined as the Gram-Schmidt basis obtained.
from \((z_1(a) - z_{k+1}(a), ..., z_k(a) - z_{k+1}(a), u_{k+1}, ..., u_d)\), where \((u_{k+1}, ..., u_d)\) is a fixed basis of \((z_1(0) - z_{k+1}(0), ..., z_k(0) - z_{k+1}(0))\).

Step 2: Now we prove that the convex hull \((F)_{\text{conv}}(0)\) is determined by the equation \(\Phi(a) = 0\) for \(a\) small enough. Let \(|a| \leq rA^{-1}\) such that \(\Phi(a) = 0\). Then \(a(z_i(a) - z_{k+1}(a)) - (F(z_1(a)) - F(z_{k+1}(a))) = 0\), and therefore let \(b := F(z_1(a)) - az_1(a) = ... = F(z_k(a)) - az_k(a)\). Then the map \(y \mapsto ay + b\) is tangent to \(F\) at all \(z_i\). Furthermore, \(p(a)\) is orthogonal to \((\text{aff}(z_1(a), ..., z_k(a)))\), implying that \(p(a) = 0\). Then \(0 \in \text{aff}(z_1(a), ..., z_k(a))\). By Lemma 6.2 we may find \(C_1, q_1 > 0\) (only depending on \(d\)) such that \(|\nabla \det\text{aff}(z_1, ..., z_k)| \leq C_1A^q_1\). Therefore, if \(a \leq \frac{1}{2}A^{-2}C_1^{-1}A^{-q_1}\), we have that \(|\det\text{aff}(z_1, ..., z_k)| \geq \frac{1}{2}A^{-1}\) and we may find \((\lambda_i)_{1 \leq i \leq k+1}\) together with the fact that \(\min \lambda \geq A^{-1}\). Now we prove that \(F \geq aY + b\). This holds on each \(B_r(z_i(a))\) by convexity of \(F\) on these balls, together with the fact that \(aY + b\) is tangent to \(F\). Now out of these balls, \(F \geq \delta \text{dist}(Y, S)\) by assumption. Furthermore, \(|z_i(a) - z_i(0)| \leq A|a|\), and \(|\nabla F(z_i(a))| \leq A^2|a|\), while similar, we have \(|F(z_i(a))| \leq h + \frac{1}{2}A^3|a|^2\). Notice that as it is tangent, \(aY + b = \nabla F(z_i(a))(Y - z_i(a)) + F(z_i(a)) = \nabla F(z_i(a))(Y - S_i) + \nabla F(z_i(a))(S_i - z_i(a)) + F(z_i(a))\), for all \(i\). Then,

\[
|aY + b| \leq |\nabla F(z_i(a))||Y - S_i| + |\nabla F(z_i(a))(S_i - z_i(a)) + F(z_i(a))|
\leq A^2|a||Y - S_i| + \frac{3}{2}A^3|a|^2 + h
\]

Therefore, if \(\delta \geq A^2|a| + (\frac{3}{2}A^3|a|^2 + h) r^{-1}\), then \(F \geq \delta \text{dist}(Y, S) \geq aY + b\). This holds in particular if \(r \geq h/H\), implying that \(A^2|a| + (\frac{3}{2}A^3|a|^2 + h) r^{-1} \leq A^2|a| + \frac{3}{2}A^3|a|^2 r^{-1} + hH/h \leq A^2|a| + \frac{3}{2}A^3|a| + H\). Finally, the following domination is sufficient:

\[
\delta \geq A^2|a| + \frac{3}{2}A^3|a| + H. \quad (6.12)
\]

Step 3: Now we prove that \(\Phi\) may be locally inverted. If \(1 \leq i \leq k\), we have \(\nabla \Phi_i(a) = z_i(a) - z_{k+1}(a)\). If \(k + 1 \leq i \leq d\), we have \(\nabla \Phi_i(a) = \nabla p(\epsilon_i) + \nabla e_i(p(a))\). We may rewrite the previous expression by introducing the locally smooth maps \(\lambda_j(a)\) such that \(p(a) = \sum_{j=1}^k \lambda_j(a)z_j(a)\), convex combination. Then \(\nabla p(a) = \sum_{j=1}^k \lambda_j(a)\nabla z_j(a) + \sum_{j=1}^k \nabla \lambda_j(a)z_j(a)\). Notice that by the relationship \(\sum_{j=1}^k \lambda_j(a) = 1\), we have that \(\sum_{j=1}^k \nabla \lambda_j(a) = 0\), therefore \(\sum_{j=1}^k \nabla \lambda_j(a)(z_j(a) - z_{k+1}(a)) = \sum_{j=1}^k \nabla \lambda_j(a)z_j(a) - \sum_{j=1}^k \nabla \lambda_j(a)(z_j(a) - z_{k+1}(a)) = 0\) as \(z_j - z_{k+1} \perp \epsilon_i\). Finally, we have \(\nabla p(a)e_i(a) = \sum_{j=1}^k \lambda_j(a)(\lambda_j(a)D^2F^{-1}(z_j))\epsilon_i(a)\).

Step 4: Now we provide a bound for \(\nabla e_i(a)p(a)\). We have the control \(|p(a)| \leq |p(0)| + \sup_{B_a} |\nabla p| r \leq H + C_2A^q a\) for \(a \in B_r\) by Lemma 6.2 for some \(C_2, q_2 > 0\). Therefore by Lemma 6.2 we may find \(C_3, q_3 > 0\) so that if \(H \leq C_3^{-1}A^{-q_3}\), we have the control \(|\nabla e_i(a)| \leq C_3A^{q_1}\), whence the inequality

\[
|\nabla e_i(a)p(a)| \leq C_3A^{q_1}(H + C_2A^q a). \quad (6.13)
\]

Step 5: Now we provide a lower bound to \(\text{det} \nabla \Phi\). Notice that \(\nabla \Phi = P_0 + P'\) with \(P_0 := \langle z_i - z_{k+1} : i \leq k, De_i : i \geq k+1 \rangle\), and \(P' := (\nabla e_i P : i \geq k+1)\), where \(D := \sum_{j=1}^k \lambda_j D^2F^{-1}(z_j)\).
Let $M_{\text{basis}} := \text{Mat}(z_i - z_{k+1} : i \leq k, e_i : i \geq k + 1)$, then $P_0M^{-1}_{\text{basis}}$ may be written as a block matrix as follows: $P_0M^{-1}_{\text{basis}} = \begin{bmatrix} I_k & \cdot \\ 0 & D_{\text{basis}} \end{bmatrix}$, with $D_{\text{basis}} := (e_i^TDe_j)_{k+1 \leq i, j \leq d}$. Then $\det(P_0M^{-1}_{\text{basis}}) = \det(D_{\text{basis}}) \geq A^{-(d-k)}$, $\det(M_{\text{basis}}) = \det(z_i - z_{k+1} : i \leq k) \geq A^{-1}$, and therefore $\det P_0 \geq A^{-(d-k+1)}$. Then by Lemma 6.2 as $k$ lines of $P_0$ are dominated by $2A$ and $d - k$ are dominated by $A$, we have $\det \nabla \Phi(a) \geq C_4^{-1}A^{-q_4}$ if $a, H \leq C_4^{-1}A^{-q_4}$, and $a \leq r$ for some $C_4, q_4 > 0$. 

Step 6: Finally $|\Phi| \leq A + A = 2A$. In order to apply Lemma 6.1 we need to control $|\nabla \Phi(a) - \nabla \Phi(a')|$ for $a, a' \in B_r$. For $i \leq k$, $|\nabla \Phi_i(a) - \nabla \Phi_i(a')| = |D^2F(z_i(a))^{-1} - D^2F(z_{k+1}(a'))^{-1}| \leq 2A^2e_i$. For $i \geq k + 1$,

$$|\nabla \Phi_i(a) - \nabla \Phi_i(a')| = |\nabla (p(a) \cdot e_i(a)) - \nabla (p(a') \cdot e_i(a'))| \leq |\nabla p(a)e_i(a) - \nabla p(a')e_i(a')| + |\nabla e_i(a)p(a) - \nabla e_i(a')p(a')| \leq |\nabla p(a)e_i(a) - \nabla p(a')e_i(a')| + 2(H + C_2A^{q_2}(|a| + |a'|))C_1A^{q_1}.$$  

We consider the first term:

$$|\nabla p(a)e_i(a) - \nabla p(a')e_i(a')| \leq \left| \sum_{j=1}^{k} \lambda_j(a') (D^2F^{-1}(z_j(a)) - D^2F^{-1}(z_j(a'))) \right| + \sum_{j=1}^{k} (\lambda_j(a) - \lambda_j(a')) D^2F^{-1}(z_j(a)) | \leq A^2e + C_1A^{q_1}|a - a'|A,$$

and therefore we may find $C_5, q_5 > 0$ such that if $|a|, e \leq C_5^{-1}A^{-q_5}$, then we have that $|\nabla \Phi(a) - \nabla \Phi(a')| \leq \frac{1}{2} |\det \nabla \Phi(0)|^{-1} \|\text{Com}(\nabla \Phi(0))\| = \|\nabla \Phi(0)\|$. Then we may apply Lemma 6.1.

$\Phi$ is a $C^1$-diffeomorphism on $B_r$, we may find $C_6, q_6 > 0$ such that $C_6^{-1}A^{-q_6} \leq |\nabla \Phi| \leq C_6A^{q_6}$. By assumption, we have $|\Phi(0)| \leq dH$. Moreover, $B_{rc_6^{-1}A^{-q_6}}(\Phi(0)) \subset \Phi(B_r)$. Therefore, if $H \leq rC_6^{-1}dA^{-q_6}$, then we may find $a_0 \in B_r$ such that $\Phi(a_0) = 0$. We have

$$|a_0| = |\Phi^{-1}(0) - \Phi^{-1}(\Phi(0))| \leq C_6A^{q_6}|\Phi(0)| \leq C_6dA^{q_6}H.$$  

By Step 2, $z_1(a_0), \ldots, z_k(a_0)$ have the required property. Moreover,

$$|z_i(a_0) - S_i| = |z_i(a_0) - z_i(0)| \leq C_6dA^{q_6+1}H.$$  

Finally, (6.12) is satisfied if $\delta \geq C_7A^{q_7}H$, with $C_7 := C_6 + \frac{5}{2}$, and $q_7 := \max(3, q_6)$. The lemma is proved for $C := \max(3, C_1, \ldots, C_7, C_6d)$ and $q := \max(3, q_1, \ldots, q_7, q_6 + 1).$ 

\[\square\]

**Proof of Theorem 5.5**

Step 1: We claim that we may find $C_1 > 0$ such that for $\varepsilon$ small enough, we have $\Delta_\varepsilon \geq -C_1\varepsilon \ln(\varepsilon^{-1})$, $m_\varepsilon$-a.e. Indeed, by the fact that $(\varphi_\varepsilon, \psi_\varepsilon, h_\varepsilon)$ is an optimum, we have that $e^{-\frac{\Delta_\varepsilon}{\varepsilon}} m_\varepsilon$ is a probability distribution. Therefore, $e^{-\frac{\Delta_\varepsilon}{\varepsilon}}(m_\varepsilon)^X/\frac{dm_\varepsilon}{dm_\varepsilon^X}$ is a probability measure,
Claim that we may find by the fact that

\[ \mu \leq C \epsilon \]

Hence \( \Delta \epsilon (X, Y) \geq -3 \gamma_c \ln(\epsilon^{-1}), \) \( \mu \) a.s. The claim is proved for \( \epsilon \) small enough.

**Step 2:** We claim that we may find \( C_2 > 0 \) such that \( \int_{\Delta \epsilon > C_2 \epsilon \ln(\epsilon^{-1})} e^{-\Delta \epsilon \epsilon} \rho \leq \epsilon \), and \( \int_{\Delta \epsilon > C_2 \epsilon \ln(\epsilon^{-1})} e^{-\Delta \epsilon \epsilon} \rho \leq 1 \). Indeed, let \( C_2 > 0 \). By (i), we have

\[ \int_{\Delta \epsilon > C_2 \epsilon \ln(\epsilon^{-1})} \Delta \epsilon e^{-\Delta \epsilon \epsilon} \rho \leq m_{\epsilon}[\Omega] C_2 \epsilon \ln(\epsilon^{-1}) \epsilon e^{C_2} \epsilon \leq C_2 \epsilon \ln(\epsilon^{-1}) \epsilon C_2 + \gamma. \]

Similar, we have that \( \int_{\Delta \epsilon > C_2 \epsilon \ln(\epsilon^{-1})} e^{-\Delta \epsilon \epsilon} \rho \leq C_2 \epsilon \ln(\epsilon^{-1}) \epsilon C_2 + \gamma. \) Therefore, up to choosing \( C_2 \) large enough, the claim holds.

**Step 3:** Let \( \Delta \epsilon := \Delta \epsilon - (\Delta \epsilon (X, \cdot))_{\text{conv}} (X) - \nabla (\Delta \epsilon (X, \cdot))_{\text{conv}} (X) \cdot (Y - X) \). We claim that \( \int x \notin D^X \Delta \epsilon e^{-\Delta \epsilon \epsilon} \rho \leq \epsilon \). Indeed by (iii) we have that \( \mu \epsilon [D^X] \leq \epsilon \), therefore we have

\[
\int_{r \notin D^X} \Delta \epsilon e^{-\Delta \epsilon \epsilon} \rho \leq \int_{r \notin D^X} C \epsilon \ln(\epsilon^{-1}) e^{-\Delta \epsilon \epsilon} \rho + \int_{\Delta \epsilon > C_2 \epsilon \ln(\epsilon^{-1})} \Delta \epsilon e^{-\Delta \epsilon \epsilon} \rho \\
\leq C \epsilon \ln(\epsilon^{-1}) \mu \epsilon [D^X] + \int_{\Delta \epsilon > C_2 \epsilon \ln(\epsilon^{-1})} \Delta \epsilon e^{-\Delta \epsilon \epsilon} \rho \\
\leq \epsilon.
\]

Finally, by the martingale property of \( e^{-\Delta \epsilon \epsilon} \rho \), we have

\[
\int_{r \notin D^X} \Delta \epsilon e^{-\Delta \epsilon \epsilon} \rho = \int_{r \notin D^X} \left( \Delta \epsilon - (\Delta \epsilon (X, \cdot))_{\text{conv}} (X) \right) v^{-\Delta \epsilon \epsilon} \rho \\
= \int_{r \notin D^X} \left( \Delta \epsilon (X, \cdot)_{\text{conv}} (X) \right) (Y - X) v^{-\Delta \epsilon \epsilon} \rho \\
\leq \int_{r \notin D^X} \Delta \epsilon e^{-\Delta \epsilon \epsilon} \rho + \mu \epsilon [D^X] C_1 \epsilon \ln(\epsilon^{-1}) \\
\leq \epsilon.
\]

**Step 4:** Let \( x \notin D^X \), we denote \( S^x_k = \{ s_1, \ldots, s_k \} \), where \( k := k^x \). We claim that for all \( S' := (s'_1, \ldots, s'_k) \in \mathbb{R}^d \) such that \( s'_i \in B_{\delta}(s_i) \) for all \( i \), and \( \sum \lambda_i s'_i = x' \in B_{\delta}(x) \) we have for \( \epsilon > 0 \) small enough that \( S' \in (B_{2A_{\epsilon}})^k, |\det S' | \geq \frac{1}{2} A_{\epsilon}^{-1}, \) \( \min \lambda_i \geq \frac{1}{2} A_{\epsilon}^{-1} \).

Indeed, \( \sum \lambda_i s'_i + x - x' = x \), with \( s'_i + x - x' \in B_{2\epsilon}(s_i) \). By Lemma 6.2, we may find \( C_3, \epsilon_3 \) such that \( |\lambda'_i - \lambda_i| \leq C_3 A_{\epsilon}^{1} \epsilon_{\delta}, |\det S' - \det S| \leq C_3 A_{\epsilon}^{1} \epsilon_{\delta}, \) and \( |s'_i| \leq A_{\epsilon} + \epsilon_{\delta} \). Now by the fact that \( \epsilon_{\delta} \leq A_{\epsilon}^{1} \), the claim is proved.

**Step 5:** We claim that up to shrinking \( D^X \), we may assume that for \( x \notin D^X \), we have \( \frac{d\mu_\epsilon}{d\mu_\epsilon^X} (x) \geq \epsilon^{\gamma + 1} \). Indeed, \( \mu_\epsilon^X \left[ \frac{d\mu_\epsilon}{d\mu_\epsilon^X} (x) \leq \epsilon^{\gamma + 1} \right] \leq \epsilon^{\gamma + 1} m_\epsilon \mu_\epsilon^X \left[ \mathbb{R}^d \right] \). Therefore,

\[
\mu_\epsilon^X \left[ \frac{d\mu_\epsilon}{d\mu_\epsilon^X} (x) \leq \epsilon^{\gamma + 1} \right] \leq \epsilon^{\gamma + 1} m_\epsilon \mu_\epsilon^X \left[ \mathbb{R}^d \right] = \epsilon^{\gamma + 1} m_\epsilon \mathbb{E} \leq \epsilon \leq 1/\ln(\epsilon^{-1}).
\]

32
Therefore we may shrink $D_e^\delta$ by removing $\{ \frac{dm_e}{dm_\varepsilon} > e^{\gamma + 1} \}$ from it.

Step 6: We claim that for $\varepsilon > 0$ small enough, we may find unique $y_i \in B_{r_\varepsilon}(s_i)$ such that $\nabla \Delta(x, y_i) = 0$ for all $i$, $B_{r_\varepsilon}(y_i) \subset B_{r_\varepsilon}(s_i)$, with $r_\varepsilon := e^{\frac{1}{2} - \eta}$, where $0 < \eta < \frac{\eta}{2}$, and finally $\Delta_\varepsilon(x, y) \geq \frac{1}{2} \sqrt{\varepsilon} \text{dist}(y, (y_1, \ldots, y_k))$ for $y \notin \cup_{i=1}^k B_{r_\varepsilon}(y_i)$. Indeed $\Delta_\varepsilon(x, \cdot)$ is strictly convex on $B_{r_\varepsilon}(s_i)$ for all $i$. Furthermore, let

$$m_i := \frac{1}{\lambda_i} \int_{B_{r_\varepsilon}(s_i)} ye^{-\frac{\Delta_\varepsilon(x, y)}{\varepsilon}} dm_\varepsilon^Y \left( \frac{dm_e}{dm_\varepsilon} \right)^{-1},$$

with $\lambda_i := \int_{B_{r_\varepsilon}(s_i)} e^{-\frac{\Delta_\varepsilon(x, y)}{\varepsilon}} dm_\varepsilon^Y \left( \frac{dm_e}{dm_\varepsilon} \right)^{-1}$. By the martingale property of $e^{-\frac{\Delta_\varepsilon}{\varepsilon}} m_\varepsilon$, we have

$$\sum_{i=1}^k \lambda_i m_i = x - \int_{(\cup_{i=1}^k B_{r_\varepsilon}(s_i))^c} ye^{-\frac{\Delta_\varepsilon(x, y)}{\varepsilon}} dm_\varepsilon^Y \left( \frac{dm_e}{dm_\varepsilon} \right)^{-1},$$

therefore by Step 5,

$$\left| \sum_{i=1}^k \lambda_i m_i - x \right| \leq \int_{(\cup_{i=1}^k B_{r_\varepsilon}(s_i))^c} |y|e^{-\frac{\varepsilon}{2} \text{dist}(y, S_\varepsilon^Y)} m_\varepsilon^Y(dy) e^{-\gamma - 1}$$

$$\leq \int_{(\cup_{i=1}^k B_{r_\varepsilon}(s_i))^c} |y|e^{-\frac{\varepsilon}{2} \text{dist}(y, S_\varepsilon^Y)} m_\varepsilon^Y(dy) e^{-\gamma - 1}$$

$$\leq \int_{(\cup_{i=1}^k B_{r_\varepsilon}(s_i))^c, |y| \geq 2A_e} 2A_e e^{-\frac{\varepsilon}{2} \text{dist}(y, S_\varepsilon^Y)} m_\varepsilon^Y(dy) e^{-\gamma - 1}.$$

Observe that as $y_i \leq A_e$, we have $\text{dist}(y, S_\varepsilon^Y) \geq |y| - A_e$. Furthermore, if $A_e \geq 2$, and $|y| \geq 2A_e$, we have $|y| \leq e|y| - A_e$. Then we have

$$\left| \sum_{i=1}^k \lambda_i m_i - x \right| \leq \int_{(\cup_{i=1}^k B_{r_\varepsilon}(s_i))^c, |y| \geq 2A_e} e^{-\frac{\varepsilon}{2} - 1} m_\varepsilon^Y(dy) e^{-\gamma - 1}$$

$$+ \int_{(\cup_{i=1}^k B_{r_\varepsilon}(s_i))^c, |y| < 2A_e} 2A_e e^{-\frac{\varepsilon}{2} r_\varepsilon} m_\varepsilon^Y(dy) e^{-\gamma - 1}$$

$$\leq e^{-\frac{\varepsilon}{2} - 1} A_e m_\varepsilon^Y [\mathbb{R}^d] e^{-\gamma - 1} + 2A_e e^{-\gamma - 1} m_\varepsilon^Y [\mathbb{R}^d] e^{-\gamma - 1}$$

$$\leq \varepsilon. \quad (6.14)$$

Similarly, we have $\sum_i \lambda_i = 1 + o(\varepsilon)$, with uniform convergence of $o(\varepsilon)$ in $x$. By Step 4, we have that $\lambda_i \geq \frac{1}{2} A_e^{-1}$ for $\varepsilon > 0$ small enough, as $\varepsilon \ll r_\varepsilon$. Therefore, we may find $y \in B_{r_\varepsilon}(s_i)$ such that $\Delta_\varepsilon(x, y) < e^{\frac{1}{2} - \frac{\varepsilon}{2}}$, as otherwise, similar to (6.14), we would have $\lambda_i \ll \varepsilon$. Notice that for $y$ in the boundary of $B_{r_\varepsilon}(s_i)$, by (v), we have $\Delta_\varepsilon(x, y) \geq e^{\frac{1}{2} - \eta} = e^{\frac{1}{2} - \eta} > e^{\frac{1}{2} - \frac{\varepsilon}{2}}$, then as $\Delta_\varepsilon(x, \cdot)$ is strictly convex on $B_{r_\varepsilon}(s_i)$, we may find a unique minimizer $y_i \in B_{r_\varepsilon}(s_i)$. Now let $l := \text{dist}(y_i, \partial B_{r_\varepsilon}(s_i))$, we have that $\Delta_\varepsilon(x, y_i) + \frac{1}{2} A_e l^2 \geq e^{1 - \eta}$. Then by the inequality $\Delta_\varepsilon(x, y_i) \leq e^{\frac{1}{2} - \frac{\varepsilon}{2}}$, we have that $l \geq \sqrt{2A_e^{-1} e^{\frac{1}{2} - \frac{\varepsilon}{2}} \sqrt{1 - e^{\frac{\varepsilon}{2}}}}$. Now, let $0 < \eta < \frac{\eta}{2}$, we have $l \leq e^{\frac{1}{2} - \eta}$ for $\varepsilon > 0$ small enough. Finally, let $y \notin \cup_{i=1}^k B_{r_\varepsilon}(y_i)$, we treat two cases. Case 1: $y \in B_{r_\varepsilon}(y_i) \setminus B_{r_\varepsilon}(y_i)$ for some $i$.
Then by (iv) we have \( \Delta_\varepsilon(x,y) \geq \Delta_\varepsilon(x,y_i) + \frac{1}{2} A_\varepsilon^{-1} r_\varepsilon^2 \geq -C_1 \varepsilon \ln(\varepsilon^{-1}) + \sqrt{\varepsilon} r_\varepsilon \geq \frac{1}{2} \sqrt{\varepsilon} \text{dist}(y,(y_1,\ldots,y_k)) \), for \( \varepsilon > 0 \) small enough.

Case 2: \( y \notin B_{r_\varepsilon}(y_i) \) for all \( i \). Then let \( 1 \leq i \leq k \), we have \( |y - s_i| \leq r_\varepsilon \), and recall that \( |y_i - s_i| \leq r_\varepsilon \). Then \( |y - y_i| \leq |y - s_i| + r_\varepsilon \leq 2|y - y_i| \), and therefore dist \( (y,(y_1,\ldots,y_k)) \leq 2\text{dist}(y,S_\varepsilon^x) \). By (v) we have \( \Delta_\varepsilon(x,y) \geq \sqrt{\varepsilon} \text{dist}(y,S_\varepsilon^x) \geq \frac{1}{2} \sqrt{\varepsilon} \text{dist}(y,(y_1,\ldots,y_k)) \), for \( \varepsilon > 0 \) small enough.

The claim is proved. Now, up to changing \( \eta \) to \( \tilde{\eta} \), the properties (i) to (vi) are still satisfied, and the properties of (iv) and (v) also hold if we replace \( S_\varepsilon^x \) by \( (y_1,\ldots,y_k) \).

Step 7: Let \( D_\varepsilon^{X\rightarrow Y} := \{ x \in D_\varepsilon^X : B_{r_\varepsilon}(y) \setminus D^Y_\varepsilon = \emptyset \}, \) for some \( y \in S_\varepsilon^x \). We claim that we have \( \mu_\varepsilon[D_\varepsilon^{X\rightarrow Y}] \ll \frac{1}{\ln(\varepsilon^{-1})} \). Indeed, for \( x \in D_\varepsilon^X \), and for all \( y \in S_\varepsilon^x \), we have \( \langle \nu_\varepsilon \rangle_x[\varepsilon B_{r_\varepsilon}(y)] \geq A_\varepsilon^{-1} \) by Step 6. Therefore, if for some such \( y \), we have that \( \varepsilon B_{r_\varepsilon}(y) \subset (D^Y_\varepsilon)^c \), then \( A_\varepsilon^{-1} \leq \langle \nu_\varepsilon \rangle_x[\varepsilon B_{r_\varepsilon}(y)] \leq \langle \nu_\varepsilon \rangle_x[(D^Y_\varepsilon)^c] \). Then, if we integrate along \( \mu_\varepsilon \) on \( D_\varepsilon^{X\rightarrow Y} \), together with (vi) we get that

\[
A_\varepsilon^{-1} \mu_\varepsilon[D_\varepsilon^{X\rightarrow Y}] \leq \langle \nu_\varepsilon \rangle(Y \neq D^Y_\varepsilon) = \nu_\varepsilon[(D^Y_\varepsilon)^c] \ll A_\varepsilon^{-1} / \ln(\varepsilon^{-1}).
\]

The claim is proved. Now up to shrinking \( D_\varepsilon^X \), we may assume that \( D_\varepsilon^X \cap D_\varepsilon^{X\rightarrow Y} = \emptyset \).

Step 8: We claim that up to raising \( \gamma \), if \( \| f \|_2 \leq \frac{1}{4} \) and \( R \geq \varepsilon / \sqrt{\varepsilon} \), then (vi) may be applied to \( y_i \) (even if \( y_i \notin D^Y_\varepsilon \)), up to replacing \( (m) \varepsilon_x[B_{R_\varepsilon}(y)] \) by \( (m) \varepsilon_x[B_{2R_\varepsilon}(y)] \). Let \( f \) have such property. Now let \( \tilde{f} \) defined by \( f = \tilde{f} \) on \( B_{R/2} \), and \( \tilde{f}(y) := \left( 1 - \frac{2|y|}{R} \right) \tilde{f}(\frac{R}{2|y|}) \). Let \( L \), \( R \geq 1 \) such that \( f \) is \( L \)-Lipschitz, then \( \tilde{f} \) is \( L \)-Lipschitz. Therefore, we have

\[
\left| \int_{B_{R}} f(y) \left[ (m) \varepsilon_x \circ \text{zoom}_{\varepsilon}^{y_i} \left( (m) \varepsilon_x(B_{2R}(y_i)) \right) - \frac{dy}{|B_{2R}|} \right] \right| \leq [2R + L]^\gamma \int_{B_2} \tilde{f}(y) \frac{dy}{|B_R|} \quad \text{from (vi)}.
\]

As we may find \( y^* \in B_{R/2} \) such that \( f(y^*) = \| f \|_2 \), we have \( f(y) \geq \| f \|_2 (1 - L|y - y^*|) \), and

\[
\int_{B_{R}} f(y) \frac{dy}{|B_{2R}|} \geq |B_1| L^{-d}.
\]

Therefore, as \( |B_{2R}| = 2^d |B_R| \), we have

\[
\left| \int_{B_{R}} f(y) \left[ (m) \varepsilon_x \circ \text{zoom}_{\varepsilon}^{y_i} \left( (m) \varepsilon_x(B_{2R}(y_i)) \right)/2^d \frac{dy}{|B_{2R}|} \right] \right| \leq \left( [2R + L]^\gamma + |B_1|^{-1} L^d \varepsilon^\beta \right) \int_{B_{R}} f(y) \frac{dy}{|B_{2R}|} \leq [R + L]^{2\gamma + 2} \varepsilon^\beta \int_{B_{R}} f(y) \frac{dy}{|B_{2R}|}. \quad (6.15)
\]

From now we replace \( \gamma \) by \( \gamma' := 2\gamma + 2 \).

Step 9: As a preparation for this step, we observe that (i) to (vi) are preserved if we replace \( \eta \) by any \( 0 \leq \eta' \leq \eta \). Then, up to lowering \( \eta \), we may assume without loss of generality that \( \eta < \beta/\gamma \). Therefore

\[
\beta - \eta \gamma > 0. \quad (6.16)
\]
We claim that
\[ \sum_{i=1}^{b_\varepsilon} \int_{B_\varepsilon(y_i)} (\Delta_\varepsilon(x, y) - \Delta_\varepsilon(x, y_i)) \, (P_\varepsilon)_x(dy) = \frac{d}{2} \varepsilon + o(\varepsilon), \]
where the convergence of \( o(\varepsilon) \) is uniform in \( x \). Indeed, consider
\[ \lambda_i := \int_{B_\varepsilon(y_i)} \exp \left( -\frac{\Delta_\varepsilon(x, y)}{\varepsilon} \right) m^\gamma(dy) \frac{d\mu_\varepsilon}{dm^\varepsilon}(x)^{-1} \]
\[ = \frac{d\mu_\varepsilon}{dm^\varepsilon}(x)^{-1} \int_{B_{\varepsilon-\eta}} \exp \left( -\frac{\Delta_\varepsilon(x, \text{zoom} \frac{y_i}{\varepsilon})(y)}{\varepsilon} \right) m^\gamma \circ \text{zoom} \frac{y_i}{\varepsilon}(dy). \] (6.17)

We want to compare \( \lambda_i \) to
\[ \lambda_i' := \frac{d\mu_\varepsilon}{dm^\varepsilon}(x)^{-1} \int_{B_{\varepsilon-\eta}} \exp \left( -\frac{\Delta_\varepsilon(x, \text{zoom} \frac{y_i}{\varepsilon})(y)}{\varepsilon} \right) dy \left( m_\varepsilon \right)_x \left[ B_{2r_\varepsilon}(y_i') \right]^{2d} \frac{d\mu_\varepsilon}{dm^\varepsilon}(x)^{-1}. \]

Notice that the map \( F : y \to \Delta_\varepsilon \frac{\text{zoom} \frac{y_i}{\varepsilon}(y)}{\varepsilon} \) may be differentiated in
\[ \nabla F = \frac{\partial_y \Delta_\varepsilon \frac{\text{zoom} \frac{y_i}{\varepsilon}(y)}{\varepsilon}}{\varepsilon}, \]
which is bounded by \( A_\varepsilon \varepsilon^{-\eta} \), by the fact that \( \nabla F(0) = 0 \) and \( D^2 F \leq A_\varepsilon I_d \) by (iv). Then, we observe that the map \( f : y \to \exp \left( -\frac{\Delta_\varepsilon \frac{\text{zoom} \frac{y_i}{\varepsilon}(y)}{\varepsilon}}{\varepsilon} \right) \) satisfies that \( \| f \|_{\varepsilon^{-\eta} / \varepsilon} \) is \( A_\varepsilon \varepsilon^{-\eta} \)-Lipschitz.

We may apply (6.15) to \( f \), with \( R = r_\varepsilon / \sqrt{\varepsilon} \), as for \( y \in B_{2R} \setminus B_R \), we have
\[ |F(y)| \leq e^{-\frac{\Delta_\varepsilon \frac{\text{zoom} \frac{y_i}{\varepsilon}(y)}{\varepsilon}}{\varepsilon} - A_\varepsilon^{-1} |y|^2} \leq |F|^{2R} e^{-A_\varepsilon \varepsilon^{-2\eta}} \leq \frac{1}{4} \varepsilon^\beta, \]
for \( \varepsilon > 0 \) small enough, and get that
\[ |\lambda_i - \lambda_i'| \leq \left| \int_{B_{\varepsilon-\eta}} f(y) \left[ \left( m_\varepsilon \right)_x \circ \text{zoom} \frac{y_i}{\varepsilon}(dy) \right] \left( m_\varepsilon \right)_x \left[ B_{2r_\varepsilon}(y_i') \right]^{2d} \frac{d\mu_\varepsilon}{dm^\varepsilon}(x)^{-1} \right| \]
\[ \leq \left[ \varepsilon^{-\eta}(A_\varepsilon + 1) \right]^{\beta} \int_{B_{\varepsilon-\eta}} f(y) \left( m_\varepsilon \right)_x \left[ B_{2r_\varepsilon}(y_i') \right]^{2d} \frac{d\mu_\varepsilon}{dm^\varepsilon}(x)^{-1} \]
\[ = (A_\varepsilon + 1)^{\gamma \varepsilon^{-\beta - \eta}} \lambda_i'. \] (6.18)

Similar, we claim that the map \( g \), defined by
\[ g : y \to \left( \Delta_\varepsilon \left( x, \text{zoom} \frac{y_i}{\varepsilon}(y) \right) - \Delta_\varepsilon(x, y_i) \right) \exp \left( -\frac{\Delta_\varepsilon \frac{\text{zoom} \frac{y_i}{\varepsilon}(y)}{\varepsilon}}{\varepsilon} \right) \]
satisfies that \( \| g \|_{\varepsilon^{-\eta} / \varepsilon} \) is \( e^{-1} A_\varepsilon \varepsilon^{-\eta} \)-Lipschitz. Now, we want to compare
\[ D_i := \int_{B_{\varepsilon-\eta}} g(y) d(m_\varepsilon \circ \text{zoom} \frac{y_i}{\varepsilon})(dy) \frac{d\mu_\varepsilon}{dm^\varepsilon}(x)^{-1}, \]
with
\[ D_i' := \int_{B_{\varepsilon-\eta}} g(y) dy \left( m_\varepsilon \right)_x \left[ B_{2r_\varepsilon}(y_i') \right]^{2d} \frac{d\mu_\varepsilon}{dm^\varepsilon}(x)^{-1}. \]
Hence similar, by (6.15), we have

\[ |D_i - D_i'| = \left| \int_{B_{\varepsilon-\eta}} g(y) \frac{d(m_{\varepsilon})}{\sqrt{\varepsilon}} (dy) \right| \leq \left[ \varepsilon^{-\eta}(e^1A\varepsilon + 1) \right]^\gamma \varepsilon^\beta \int_{B_{\varepsilon-\eta}} g(x) \frac{dx}{\sqrt{\varepsilon}} \leq (e^{-1}A\varepsilon + 1)^\gamma \varepsilon^{\beta - \eta} D_i'. \]

Now we compute the value of \( D_i' \). By the fact that \( \lambda_i \) is uniformly smaller than \( \varepsilon \) applied to a symmetrical positive definite matrix denotes the only symmetrical positive definite square root of the matrix \( A \), we get that \( D_i' = \varepsilon K_i \sqrt{\Pi^d} \sqrt{\det \frac{\partial^2 \Delta_\varepsilon(x, y_i) y^2}{\varepsilon}} \).
Finally, using (1.1) and the fact that we can sum up positive $o(\cdot)$, we get
\[
\sum_{i=1}^{k_\delta} \int_{B_{r_\delta}(y_i)} (\Delta_\varepsilon (x, y) - \Delta_\varepsilon (x, y_i)) (\mathbb{P}_e)_x(dy) = \sum_{i=1}^{k} D_i \\
= \sum_{i=1}^{k} D''_i + o \left( \sum_{i=1}^{k} D'_i + D'''_i \right) \\
= \varepsilon K_{i} d \sqrt{\det \delta_{l}^2 \Delta_\varepsilon (x, y_i)} + o \left( \sum_{i=1}^{k} D_i \right) \\
= \varepsilon \frac{d}{2} \sum_{i=1}^{k} \lambda''_i + o \left( \sum_{i=1}^{k} D_i \right) \\
= \varepsilon \frac{d}{2} + o \left( \varepsilon + \sum_{i=1}^{k} D_i \right) \\
= \varepsilon \frac{d}{2} + o(\varepsilon),
\]
where all the $o(\cdot)$ are uniform in $x$, thanks to all the controls established in this step. The claim is proved.

**Step 10:** We claim that for $\varepsilon > 0$ small enough, we have
\[
\text{dist}(x, \text{aff}(y_1, \ldots, y_k)) \leq C_5 A_{q_5} \varepsilon^{\frac{1}{2} + \beta - \eta} + C_6 A_{q_6} \varepsilon^{\frac{1}{2} + \eta} + \varepsilon.
\]
(6.22)

Indeed, let $z_i := \lambda_i^{-1} \int_{B_{r_\varepsilon}(y_i)} ye^{-\frac{\Delta_\varepsilon(x,y)}{\varepsilon}} m_\varepsilon(y) \left( \frac{d\mu_\varepsilon}{dm_\varepsilon} \right)^{-1}$, where recall that
\[
\lambda_i = \int_{B_{r_\varepsilon}(y_i)} e^{-\frac{\Delta_\varepsilon(x,y)}{\varepsilon}} m_\varepsilon(y) \left( \frac{d\mu_\varepsilon}{dm_\varepsilon} \right)^{-1}.
\]
We have that $x = \int_{\mathbb{R}^d} ye^{-\frac{\Delta_\varepsilon(x,y)}{\varepsilon}} m_\varepsilon(y) \left( \frac{d\mu_\varepsilon}{dm_\varepsilon} \right)^{-1}$, by the martingale property of $e^{-\frac{\Delta_\varepsilon}{\varepsilon}} m_\varepsilon$.

Similar to (6.14), we have $\left| \sum_{i=1}^{k} \lambda_i z_i - x \right| \ll \varepsilon$. Similar, we also have $\sum_{i=1}^{k} \lambda_i = 1 + o(\varepsilon)$. Therefore, dist$(x, \text{aff}(z_1, \ldots, z_k)) \ll \varepsilon$. Now let
\[
z_i' := \frac{1}{\lambda_i} \int_{B_{r_\varepsilon}(y_i)} ye^{-\frac{\Delta_\varepsilon(x,y)}{\varepsilon}} m_\varepsilon(y) \left( \frac{d\mu_\varepsilon}{dm_\varepsilon} \right)^{-1} \frac{dy}{|B_{r_\varepsilon}|},
\]
where recall that $\lambda_i' = \int_{B_{r_\varepsilon}(y_i)} e^{-\frac{\Delta_\varepsilon(x,y)}{\varepsilon}} m_\varepsilon(y) \left( \frac{d\mu_\varepsilon}{dm_\varepsilon} \right)^{-1}$. We claim that we may find universal $C_5, q_5 > 0$, such that $|\lambda_i (z_i - y_i) - \lambda_i' (z_i' - y_i)| \leq C_5 A_{q_5} \varepsilon^{\frac{1}{2} + \beta - \eta}$. Indeed, if $u$ is a unit vector, we have
\[
h : y \rightarrow \sqrt{\varepsilon} |y \cdot u| \exp \left( -\frac{\Delta_\varepsilon (x, \text{zoom}_\varepsilon^{-1}(y)) - \Delta_\varepsilon (x, y_i)}{\varepsilon} \right).
\]
We claim that \( \frac{h}{|h|^\alpha} \) is \( \varepsilon^{-\eta}A_\varepsilon \left( 1 + \frac{\sqrt{\gamma}}{2\gamma_0} \right) \) -Lipschitz, where \( \gamma_0 > 0 \) is the unique positive real satisfying \( 2y_0^2 = \varepsilon^{-\eta}y \), furthermore the function is small enough on \( B_{2R(y_i)} \backslash B_R(y_i) \) so that we may apply (6.15):

\[
|u \cdot (\lambda_i(z_i - y_i) - \lambda_i'(z_i' - y_i))| = \left| \int_{B_{\varepsilon-\eta}(y_i)} h(y) \frac{d(m_x)_{x} \circ \text{zoom}_{\varepsilon_1}(dy)}{(m_x)_{x}[B_{2r_i}(y_i)]/2^d} - \frac{dy}{|B_{\varepsilon-\eta}|} \right| K_i |B_{\varepsilon-\eta}|
\]

\[
\leq \left[ \varepsilon^{-\eta}A_\varepsilon \left( 1 + \frac{\sqrt{A_\varepsilon}}{2\gamma_0^3} \right) \right]^\gamma \varepsilon^{\delta} \int_{B_{\varepsilon-\eta}(y_i)} h(y)dyK_i,
\]

\[
\leq \left[ A_\varepsilon \left( 1 + \frac{\sqrt{A_\varepsilon}}{2\gamma_0^3} \right) \right]^\gamma \varepsilon^{1+\gamma |I\varepsilon|} A_\varepsilon K_i,
\]

(6.23)

where \( I := \int_{\mathbb{R}} |y \cdot u|^2 e^{-|y|^2} dy \), as recall that by Step 9, \( K_i = \frac{\chi_n'}{\sqrt{2\pi} \sqrt{\det \varepsilon^2 \Delta_{\varepsilon}(x,y)}} \leq A_{\varepsilon}^{d/2} \). Now we consider again a unit vector \( u \), and

\[
(z_i' - y_i) \cdot u = \frac{\sqrt{\varepsilon}}{\lambda_i'} \int_{B_{\varepsilon-\eta}} ((y \cdot u)_+ + (y \cdot u)_-) e^{-\Delta_{\varepsilon}(x,y_i)} dyK_i
\]

\[
\leq \frac{\sqrt{\varepsilon}}{\lambda_i'} \int_{B_{\varepsilon-\eta}} (y \cdot u)_+ e^{-\varepsilon \Delta_{\varepsilon}(x,y_i)} dyK_i
\]

\[
- \frac{\sqrt{\varepsilon}}{\lambda_i'} \int_{B_{\varepsilon-\eta}} (y \cdot u)_- e^{-\varepsilon \Delta_{\varepsilon}(x,y_i)} dyK_i,
\]

\[
\leq \frac{K_i \sqrt{\varepsilon}}{\lambda_i'} (I/2) \left( \sqrt{\varepsilon^2 \Delta_{\varepsilon}(x,y_i)} - \varepsilon^2 I_d^{-1} u^2 - \sqrt{\varepsilon^2 \Delta_{\varepsilon}(x,y_i)} + \varepsilon^2 I_d^{-1} u^2 \right) \]

\[
\leq C_6 A_{\varepsilon}^{q_6} \varepsilon^{\frac{1}{2} + \gamma},
\]

for some \( C_6, q_6 > 0 \), independent of \( x \) and \( u \), as \( \varepsilon^2 \Delta_{\varepsilon}(x,y_i) \geq A_{\varepsilon}^{-1} I_d \geq \varepsilon^2 I_d \), and \( \lambda_i' \geq \frac{1}{2} A_{\varepsilon}^{-1} \). Then \( |z_i' - y_i| \leq C_6 A_{\varepsilon}^{q_6} \varepsilon^{\frac{1}{2} + \gamma} \).

Finally, with (6.18) and (6.20), up to raising \( C_5, C_6, q_5, q_6 \), we get the estimate \( |z_i - y_i| \leq C_5 A_{\varepsilon}^{q_5} \varepsilon^{\frac{1}{2} + \gamma - \eta} + C_6 A_{\varepsilon}^{q_6} \varepsilon^{\frac{1}{2} + \gamma - \eta} \). We finally get the desired estimate from the fact that \( \text{dist}(x, \text{aff}(y_1, \ldots, y_k)) \).

Step 11: We now claim that

\[
\int_{B_{\varepsilon}(y_i)} \Delta_{\varepsilon}(x, \cdot) d(P_\varepsilon)_x = \int_{B_{\varepsilon}(y_i)} (\Delta_{\varepsilon}(x, \cdot) - \Delta_{\varepsilon}(x, y_i)) d(P_\varepsilon)_x + o(\varepsilon),
\]

(6.24)

where the convergence speed of \( o(\varepsilon) \) is independent of the choice of \( x \) and \( i \). Indeed, by (6.22) and the fact that

\[
|\Delta_{\varepsilon}(x, y_i)| \leq \max(\Delta_1, \Delta_2) \varepsilon \ln(\varepsilon^{-1}) \leq \varepsilon^{1 - \frac{\alpha}{2}} \leq r_\varepsilon H_{\varepsilon},
\]

with \( H_{\varepsilon} := \varepsilon^{\frac{1}{2} + \frac{1}{2} \min(\theta, \beta - \eta)} \) for \( \varepsilon \) small enough. Furthermore, by (6.22), we have that for \( \varepsilon > 0 \) small enough, \( \text{dist}(x, \text{aff}(y_1, \ldots, y_k)) \leq H_{\varepsilon} \). Then, we may apply Lemma 6.3 we may find \( C_7, q_7 > 0 \) such that we have \( \Delta_{\varepsilon}(x, y) = \Delta_{\varepsilon}(x, y_i) - \Delta_{\varepsilon}(x, \bar{y}_i) - \nabla \Delta_{\varepsilon}(x, \bar{y}_i) \cdot (y - \bar{y}_i) \), with \( |y_i - \bar{y}_i| \leq C_7 A_{\varepsilon}^{q_7} H_{\varepsilon} \leq C_7 A_{\varepsilon}^{q_7} \varepsilon^{\frac{1}{2}} \), as by Step 6, we have that for \( \eta \neq \cup_{i=1}^k B_{\varepsilon}(y_i) \), we have \( \Delta_{\varepsilon}(x, y) \geq \frac{1}{2} \sqrt{\varepsilon} \text{dist}(y, (y_1, \ldots, y_k)) \), where \( \frac{1}{2} \sqrt{\varepsilon} \geq C_7 A_{\varepsilon}^{q_7} (\varepsilon) \), and \( r_\varepsilon \geq C_7 A_{\varepsilon}^{q_7} H_{\varepsilon} \) as
well. Notice that therefore, we have $0 \leq \Delta_\varepsilon(x, y_i) - \Delta_\varepsilon(x, y_i) \leq \frac{1}{2}A_\varepsilon C_2^2 A_\varepsilon^2 H_\varepsilon^2 \ll \varepsilon$, and similar, we have $| \nabla \Delta_\varepsilon(x, y_i) \cdot (y_i - y_i) | \leq A_\varepsilon C_2^2 A_\varepsilon^2 H_\varepsilon^2 \ll \varepsilon$. Finally, $S_{B_\varepsilon(y)}(y_i) \Delta_\varepsilon(x, y_i) \cdot (y_i - y_i)(F_\varepsilon)_x = \nabla \Delta_\varepsilon(x, y_i) \cdot S_{B_\varepsilon(y)}(y_i) (y_i - y_i)$. We have from the computations in Step 10 that $| S_{B_\varepsilon(y)}(y_i) (y_i - y_i)(F_\varepsilon)_x | \leq C_3 A_\varepsilon^2 \varepsilon \frac{1}{2}$, and $| \nabla \Delta_\varepsilon(x, y_i) \cdot S_{B_\varepsilon(y)}(y_i) (y_i - y_i) | \leq C_3 C_7 A_\varepsilon^2 + 2^r \varepsilon \frac{1}{2} H_\varepsilon \ll \varepsilon$, whence the result.

Step 12: Now using Step 3 and Step 9, integrating against $\mu_\varepsilon$, with the uniform error estimate \([6.24]\), together with controls that are independent of $x$, similar to \([6.14]\), to deal with $(\cup_{i=1}^n B_\varepsilon(y_i))^c$, we get

$$\int \tilde{\Delta}_\varepsilon d\mu_\varepsilon = \int D_\varepsilon \sum_{i=1}^{k_\varepsilon} \int B_\varepsilon(y_i) (\Delta_\varepsilon(x, y) - \Delta_\varepsilon(x, y_i)) (\mu_\varepsilon)(dy) \varepsilon(x) + o(\varepsilon)
= \frac{d}{\varepsilon^2} + o(\varepsilon).$$

Finally, notice that

$$\tilde{\Delta}_\varepsilon = \Delta_\varepsilon - (\Delta_\varepsilon(X, \cdot))_{\text{conv}}(X) - \nabla(\Delta_\varepsilon(X, \cdot))_{\text{conv}}(X) \cdot (Y - X)
= \psi_\varepsilon - c - (\psi_\varepsilon - c(X, \cdot))_{\text{conv}}(X) - \nabla(\psi_\varepsilon - c(X, \cdot))_{\text{conv}}(X) \cdot (Y - X)
= \tilde{\psi}_\varepsilon + \psi_\varepsilon - c - \nabla(\psi_\varepsilon - c(X, \cdot))_{\text{conv}}(X) \cdot (Y - X).$$

Therefore we have $\mu_\varepsilon[\tilde{\Delta}_\varepsilon] = \mu_\varepsilon[\tilde{\psi}_\varepsilon] + \nu_\varepsilon[\psi_\varepsilon] - \mu_\varepsilon[c]$. Whence the result of the theorem. □

### 6.5 Asymptotic penalization error

**Proof of Proposition 5.14** As $\mu \leq \nu$, $\tilde{V}_\varepsilon$ is convex with a finite global minimum. Then the minimum of $\tilde{V}_\varepsilon + \alpha f$ converges to a minimum of $\tilde{V}_\varepsilon$. More precisely, let $\psi_\alpha$ be the only global minimizer of $\tilde{V}_\varepsilon + \alpha f$, then $\psi_\alpha$ is also the minimizer of the map $\frac{1}{\alpha}(\tilde{V}_\varepsilon - \tilde{V}_\varepsilon(\psi_\alpha)) + f$, which $\Gamma$–converges to $\{\psi_\varepsilon \mid \psi_\varepsilon(\psi_\alpha) \neq \tilde{V}_\varepsilon(\psi_\alpha)\} + f$, whose unique global minimizer is $\psi_0$. Therefore $\psi_\alpha \rightarrow \psi_0$ when $\alpha \rightarrow 0$. Now the first order condition gives that $\frac{\psi_\alpha - \nu}{\alpha} = \nabla f(\psi_\alpha) \rightarrow \nabla f(\psi_0)$, when $\alpha \rightarrow 0$, by convexity and differentiability of $f$, guaranteeing that $\psi \mapsto \nabla f(\psi)$ is continuous. □

### 6.6 Convergence of the martingale Sinkhorn algorithm

**Proof of Theorem 5.15** This result stems from an indirect application of Theorem 5.2 in [5]. By a direct application of this theorem we get that

$$V_\varepsilon(x_k) - V_\varepsilon^* \leq \frac{2 \min(L_1, L_2) R^2(x_0)}{n - 1},$$

and that

$$V_\varepsilon(x_k) - V_\varepsilon^* \leq \left(1 - \frac{\sigma}{\min(L_1, L_2)}\right)^{n-1} (V_\varepsilon(x_0) - V_\varepsilon^*),$$

with $R(x_0) := \sup_{V_\varepsilon(x) \leq V_\varepsilon(x_0)} \text{dist}(x, X^*)$, $L_1$ (resp. $L_2$) is the Lipschitz constant of the $\psi$–gradient (resp. $(\varphi, h)$–gradient) of $V_\varepsilon$, and $\sigma$ is the strong convexity parameter of $V_\varepsilon$. 39
Furthermore, the strong convexity gives that
\[ \text{dist}(x_k, \mathcal{X}^*) \leq \sqrt{\frac{2}{\sigma}(V_\varepsilon(x_k) - V_\varepsilon^*)}. \] (6.27)

However the gradient $\nabla V_\varepsilon$ is locally but not globally Lipschitz, nor $V_\varepsilon$ strongly convex. Therefore we need to refine the theorem by looking carefully at where these constants are used in its proof.

**Step 1:** The constant $L_1$ is used for Lemma 5.1 in [5]. We need for all $k \geq 0$ to have $V_\varepsilon(x_k) - V_\varepsilon(x_{k+1/2}) \geq \frac{1}{2L_1} |\nabla V_\varepsilon(x_k)|^2$. We may start from $x_1$, this way all the $x_k$ are such that $(e^{-\Delta(x_k)h,i})_{i,j \in \mathcal{X}}$ is a probability distribution. Then $|\partial_\psi^2 V_\varepsilon(x_k)| \leq \varepsilon^{-1}$. Let $u \in \mathbb{R}^\mathcal{Y}$, then $|\partial_\psi^2 V_\varepsilon(\varphi_k, \psi_k + u, h_k)| \leq \varepsilon^{-1} e^{\frac{|w|}{\varepsilon}}$.

We want to find $C, L > 0$ such that $V_\varepsilon(x_k) - V_\varepsilon(\varphi_k, \psi_k - C\partial_\psi V_\varepsilon(x_k), h_k) \geq \frac{1}{2L} |\partial_\psi V_\varepsilon(x_k)|$, then $L$ may be used to replace $L_1$ in the final step of the proof of Lemma 5.1 in [5]. Recall that $|\partial_\psi V_\varepsilon(x_k)|_\infty \leq 1$, as it is the difference of two probability vectors. We have

\[
V_\varepsilon(x_k) - V_\varepsilon(\varphi_k, \psi_k - C\partial_\psi V_\varepsilon(x_k), h_k)
= -\partial_\psi V_\varepsilon(x_k) \cdot (-C\partial_\psi V_\varepsilon(x_k))
- C^2 \int_0^1 (1-t)\partial_\psi V_\varepsilon(x_k)^t \partial_\psi^2 V_\varepsilon(\varphi_k, \psi_k - tC\partial_\psi V_\varepsilon(x_k), h_k) \partial_\psi V_\varepsilon(x_k) dt
\leq C|\partial_\psi V_\varepsilon(x_k)|^2 - C^2|\partial_\psi V_\varepsilon(x_k)|^2 \int_0^1 \varepsilon^{-1}(1-t)e^{\frac{C^2}{\varepsilon^2}} dt
= C|\partial_\psi V_\varepsilon(x_k)|^2 - C^2|\partial_\psi V_\varepsilon(x_k)|^2 \varepsilon^{-1} e^{\frac{C^2}{\varepsilon^2}} C^2|\partial_\psi V_\varepsilon(x_k)|^2
= \left( C - \varepsilon \left( e^{\frac{C^2}{\varepsilon^2}} - 1 - \frac{C}{\varepsilon} \right) \right) |\partial_\psi V_\varepsilon(x_k)|^2.
\]

Deriving with respect to $C$ gives the equation $C = \varepsilon \ln(2)$. We get

\[ V_\varepsilon(x_k) - V_\varepsilon(\varphi_k, \psi_k - C\partial_\psi V_\varepsilon(x_k), h_k) \geq \varepsilon (2\ln(2) - 1)|\partial_\psi V_\varepsilon(x_k)|^2. \]

Therefore we may use $L := \varepsilon^{-1}(4\ln(2) - 2)^{-1}$.

**Step 2:** The constant $\sigma$ is used to get the result from (3.21) in [5]. Then we just need the inequality

\[ V_\varepsilon(y) \geq V_\varepsilon(x) + \nabla V_\varepsilon(x) \cdot (y - x) + \frac{\sigma}{2} |y - x|^2, \] (6.28)

to hold for some $y \in \mathcal{X}^*$ and $x = x_k$ for all $k \geq 0$. Now we give a lower bound for $\sigma$. Notice that $V_\varepsilon = \mu[\varphi] + \nu[\psi] + \varepsilon \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \exp \left( \frac{-\varepsilon}{\varepsilon} \right) \circ \Delta_{x,y}$. Then for $x_0, u \in \mathcal{D}_{\mathcal{X}, \mathcal{Y}}$, we have

\[
u^t D^2 V_\varepsilon(x_0) u = \varepsilon^{-1} \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \exp \left( \frac{-\varepsilon}{\varepsilon} \right) \circ \Delta_{x,y}(x_0) \Delta_{x,y}(u)^2 \geq \varepsilon^{-1} \exp \left( \frac{-|\Delta(x)|_\mathcal{X}}{\varepsilon} \right) |\Delta(x)|^2.
\]
Then, by definition of $\lambda_2$, we may find $\tilde{u}$ such that $\Delta(u) = \Delta(\tilde{u})$, and
\begin{equation}
 u' D^2 V_\varepsilon(x_0) u \geq \frac{|X|}{\lambda_2 \varepsilon} \exp \left( - \frac{|\Delta(x)| \varepsilon}{\varepsilon} \right) |\tilde{u}|^2. \tag{6.29}
\end{equation}
Now, we claim that $|\Delta(x)| \varepsilon \leq D(x_0)$. Then let $x^* \in X^*$ and consider (6.29) for $u = x^* - x$. Then we have that $x + \tilde{u} \in X^*$, and therefore, we may take $y = x + \tilde{u}$ for (6.28), and therefore use
\begin{equation}
 \sigma := \frac{|X|}{\lambda_2 \varepsilon} \exp \left( - \frac{D(x_0)}{\varepsilon} \right), \tag{6.30}
\end{equation}
in this equation.

Step 3: Now we prove our claim that $|\Delta(x)| \varepsilon \leq D(x_0)$. Indeed
\begin{equation}
 V_\varepsilon(x_0) \geq V_\varepsilon(x) = \mu[\varphi] + \nu[\psi] + \varepsilon = P_0[\varphi \oplus \psi + h^\oplus] + \varepsilon = P_0[\Delta(x)] + P_0[c] + \varepsilon.
\end{equation}
Therefore we have $P_0[\Delta(x)] \leq V_\varepsilon(x_0) - P_0[c] - \varepsilon$, and finally
\begin{equation}
 (P_0)_{\text{min}} \Delta(x) \leq V_\varepsilon(x_0) - P_0[c] - \varepsilon. \tag{6.31}
\end{equation}
Then $|\Delta(x)| \varepsilon \leq D(x_0)$ stems from the definition of $\lambda_1$.

Step 4: Now we provide the bound on $R(x_0)$. From the proof of Theorem 5.2 in [5], what is needed to make the proof work is $R(x_0) = \sup_{k \geq 0} \text{dist}(x_k, X^*)$, which is smaller than $\sup_{x \in V_\varepsilon(x_0)} \text{dist}(x, X^*)$. Furthermore, from (6.27) together with (6.30), we get that the supremum $\sup_{k \geq 0} \text{dist}(x_k, X^*)$ is also smaller than $\sqrt{\frac{|X|}{\lambda_2 \varepsilon} \exp \left( - \frac{D(x_0)}{\varepsilon} \right)} \left( V_\varepsilon(x_0) - V_\varepsilon^* \right)^{\frac{1}{2}}$. Finally, from (6.31), together with the definition of $\lambda_1$, we may find that $\tilde{x}^* \in X^*$ such that $\Delta(x_k) = \Delta(\tilde{x}_k)$, $\tilde{x}^* \in X^*$, $|\tilde{x}_k|_1 \leq \lambda_1 \frac{V_\varepsilon(x_0) - P_0[c] - \varepsilon}{|X|(|P_0|_{\text{min}} - P_0[c])}$, and $|\tilde{x}^*|_1 \leq \lambda_1 \frac{V_\varepsilon(x_0) - P_0[c] - \varepsilon}{|X|(|P_0|_{\text{min}} - P_0[c])}$. Then $|\tilde{x}_k - \tilde{x}^*| \leq |\tilde{x}_k - \tilde{x}^*|_1 \leq 2 \lambda_1 \frac{V_\varepsilon(x_0) - P_0[c] - \varepsilon}{|X|(|P_0|_{\text{min}} - P_0[c])}$ by the fact that $|\cdot|_1 \leq |\cdot|$. Finally, as $\tilde{x}^* + x_k - \tilde{x}_k \in X^*$, we have $\text{dist}(x_k, X^*) \leq 2 \lambda_1 \frac{V_\varepsilon(x_0) - P_0[c] - \varepsilon}{|X|(|P_0|_{\text{min}} - P_0[c])}$. Therefore the same bound holds for $\sup_{k \geq 0} \text{dist}(x_k, X^*)$.

Step 5: Finally, as we focus on the $L_1$ optimization phase, we may replace $n - 1$ by $n$ in the convergence formula (6.25) and (6.26), see the proof of Theorem 5.2 in [5]. The result is proved.

\section*{6.7 Implied Newton equivalence}

\textbf{Proof of Proposition 5.20} We apply the Newton step in the algorithm to $(x, y(x))$. We are looking for $p$ such that $D^2 F p = \nabla F$. First $\nabla F(x, y(x)) = \partial_x F(x, y(x)) = \nabla \hat{F}(x)$, then if we decompose $p = p_x \oplus p_y \in X \oplus Y$, the equation becomes
\begin{equation}
 \partial_{xx}^2 F p_x + \partial_{xy}^2 F p_y = \nabla \hat{F}(x), \quad \text{and} \quad \partial_{yx}^2 F p_x + \partial_{yy}^2 F p_y = 0.
\end{equation}
The solution to this equation system is given by $p_y = -\partial_{yy}^2 F^{-1} \partial_{yx}^2 F p_x$, and
\begin{equation}
 \left( \partial_{xx}^2 F p_x - \partial_{xx}^2 F \partial_{yx}^2 F^{-1} \partial_{yy}^2 F \partial_{xx}^2 F \right) p_x = \nabla \hat{F}(x) \tag{6.32}
\end{equation}
The conclusion follows from the fact that (6.32) is the step for the Newton algorithm applied to $\hat{F}$. The Newton step on $y$ does not matter, as $y$ will be immediately thrown away and replaced by $y(x)$.
7 Numerical experiment

7.1 An hybrid algorithm

The steps of the Newton algorithm are theoretically very performing if the current point is close enough to the optimum. What is really time-consuming is the computation of the descent direction with the conjugate gradient algorithm. The idea of preferring the Newton method to the Bregman projection method in the case of martingale optimal transport comes from the fact that, unlike in the case of classical transport, projecting on the martingale constraint is more costly than projecting on the marginal constraints, as we use a Newton algorithm instead of a closed formula. From the experiment, we would say that in dimension 1 it takes $7$ times more time, and $20$ times more in dimension 2. The implied Newton algorithm performs this projection only for the Newton step, whereas it is not necessary for the conjugate gradient algorithm.

We Notice that the Bregman projection algorithm is more effective at the beginning, to find the optimal region, and then it converges slower. In contrast, the Newton algorithm is slow at the beginning when it is searching the neighborhood of the optimum, but when its finds this neighborhood, the convergence gets very fast. Then it makes sense to apply an hybrid algorithm that starts with Bregman projections, and concludes with the Newton method. We call this dual-method algorithm the hybrid algorithm. We see on the simulations that it generally out-performs the two other algorithms.

Figure 2 compares the evolution of the gradient error in dimension 1 and 2 of the longest step of the three algorithms in terms of computation time. What we call here the gradient error is the norm 1 of the gradient of the function $\tilde{V_\varepsilon}$ that we are minimizing, and which is also equal to the difference between the target measure $\nu$ and the current measure. In the case of Newton algorithms, the penalization gradient is also included, then we use a coefficient in front of this penalization so that it does not interfere too much with the equation between the current and the target measure. We use the $\varepsilon$-scaling technique. For each value of $\varepsilon$, we iterate the minimization algorithm until the error is smaller than $10^{-2}$. Then at the final iteration we lower the target error to the one we want.

Figure 2a compares the evolution of the gradient error in dimension 1 and 2 of the longest step of the three algorithms in terms of computation time. What we call here the gradient error is the norm 1 of the gradient of the function $\tilde{V_\varepsilon}$ that we are minimizing, and which is also equal to the difference between the target measure $\nu$ and the current measure. In the case of Newton algorithms, the penalization gradient is also included, then we use a coefficient in front of this penalization so that it does not interfere too much with the equation between the current and the target measure. We use the $\varepsilon$-scaling technique. For each value of $\varepsilon$, we iterate the minimization algorithm until the error is smaller than $10^{-2}$. Then at the final iteration we lower the target error to the one we want.

The green line corresponds to the Bregman projections algorithm. The orange line corresponds to the implied truncated Newton algorithm. All the techniques evocated in Section 4 are applied. We use the diagonal of the Hessian to precondition the conjugate gradient algorithm. The coefficient in front of the quadratic penalization, which is normalized by $\nu^2$, is set to $10^{-2}$. Finally the blue line corresponds to the "hybrid algorithm", which consists in doing some Bregman projection steps before switching to the implied truncated Newton algorithm. The moment of switching is chosen by very empirical criteria: we do it after having the initial error divided by 2 or after 100 iteration, or if the initial error is divided by 1.1 if the initial error is smaller than 0.1.

Figure 2a gives the computation times of these three entropic algorithms, for a grid size going from 10 to 2500 while $\varepsilon$ goes from 1 to $10^{-4}$, with the cost function $c := XY^2$, $\mu$ uniform
on $[-1,1]$, and $\nu := \frac{1}{K}|Y|^{1.5}\mu$, where $K := (|Y|^{1.5}\mu)[\mathbb{R}]$. By [34] the optimal coupling that we get is the "left curtain" coupling studied in [7]. We show the curves for the value of $\epsilon$ that takes the largest amount of time, the one for which the time of computation is the most important for $\epsilon = 4.2 \times 10^{-4}$.

We conduct the same experiment on a two dimensional problem. The difference of efficiency between the algorithms should be even bigger, as the computing of the optimal $h$ becomes more costly, as the optimization of a convex function of two variables. Let $d = 2$, $c : (x, y) \in \mathbb{R}^2 \times \mathbb{R}^2 \mapsto x_1(y_1^2 + 2y_2^2) + x_2(2y_1^2 + y_2^2)$, $\mu$ uniform on $[-1,1]^2$, and $\nu = \frac{1}{K}(|Y_1|^{1.5} + |Y_2|^{1.5})\mu$ where $K := ((|Y_1|^{1.5} + |Y_2|^{1.5})\mu)[\mathbb{R}^2]$. We start with a $10 \times 10$ grid and scale it to a $160 \times 160$ one while $\epsilon$ scales from $1$ to $10^{-4}$. Figure 2b gives the computation times of the three entropic algorithms. Once again we show the curves for the value of $\epsilon$ that takes the largest amount of time, the one for which the time of computation is the most important for $\epsilon = 7.4 \times 10^{-3}$.

![Figure 2: Log plot of the size of the gradient VS time for the Bregman projection algorithm, the Newton algorithm, and the Hybrid algorithm.](image)

7.2 Results for some simple cost functions

7.2.1 Examples in one dimension

Figure 3 give the solution for three different costs for $\epsilon = 10^{-5}$ with $\mu := (\mu_1 + \mu_2)/2$ and $\nu := (\nu_1 + \nu_2)/2$ with $\mu_1$ uniform on $[-1,1]$, $\nu_1 = \frac{1}{K}|Y|^{1.5}\mu_1$ with $K := (|Y|^{1.5}\mu_1)[\mathbb{R}]$, $\mu_2$ is the law of $\exp(\mathcal{N}(-\frac{1}{2}\sigma_1^2, \sigma_1^2)) - 1$ with $\sigma_1 = 0.1$, and $\nu_2$ is the law of $\exp(\mathcal{N}(-\frac{1}{2}\sigma_2^2, \sigma_2^2)) - 1$ with $\sigma_2 = 0.2$. The scale indicates the mass in each point of the grid, the mass of the entropic approximation of the optimal coupling is the yellow zone. Notice that in all the cases the optimal coupling is supported on at most two maps. We saw this in all our experiment, we conjecture that for almost all $\mu, \nu$ this is the case.
Figure 3: Optimal coupling for different costs in dimension one.

Figure 3a shows well the left curtain coupling from [7] and [34]. Figure 3b shows the optimal coupling for the distance cost. This coupling has been studied by Hobson & Neuberger [37]. They predict that this coupling is concentrated on two graphs. Finally, Figure 3c shows how we may find solutions for any kind of cost function.

7.2.2 Example in two dimensions

Figure 4: Optimal coupling conditioned to several values of $X$.

In dimension 2, it has been proved in [25] that for the cost function $c : (x, y) \in \mathbb{R}^2 \times \mathbb{R}^2 \mapsto x_1(y_1^2 + 2y_2^2) + x_2(2y_1^2 + y_2^2)$, the kernel of optimal probabilities are concentrated on the intersection of two ellipses with fixed characteristics, except for their position and their scale. Figure 4 is meant to test this theoretical result. We do an entropic approximation with a grid $160 \times 160$, and $\varepsilon = 10^{-4}$. Then we selected 4 points $x_1 := (-0.45, -0.65)$, $x_2 := (0.3, -0.66)$, $x_3 := (0.2, 0.44)$, and $x_4 := (0.13, 0.16)$ and draw the kernels of the approximated optimal transport $\mathbb{P}^*$ conditioned to $X = x_i$ for $i = 1, 2, 3, 4$. We see on this figure that for $i = 1, 2, 3, 4$, $\mathbb{P}^*(\cdot | X = x_i)$ is concentrated on exactly $i$ points, showing that all the numbers between 1 and 4 are reached. It seems that no trivial result can be proved on the number of maps that we may reach.
References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. Network flows. 1988.

[3] Aurélien Alfonsi, Jacopo Corbetta, and Benjamin Jourdain. Sampling of probability measures in the convex order and approximation of martingale optimal transport problems. 2017.

[4] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1961–1971, 2017.

[5] Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4):2037–2060, 2013.

[6] Mathias Beiglböck, Pierre Henry-Labordère, and Friedrich Penkner. Model-independent bounds for option prices: a mass transport approach. *Finance and Stochastics*, 17(3):477–501, 2013.

[7] Mathias Beiglböck and Nicolas Juillet. On a problem of optimal transport under marginal martingale constraints. *The Annals of Probability*, 44(1):42–106, 2016.

[8] Mathias Beiglböck, Tongseok Lim, and Jan Oblój. Dual attainment for the martingale transport problem. *arXiv preprint arXiv:1705.04273*, 2017.

[9] Mathias Beiglböck, Marcel Nutz, and Nizar Touzi. Complete duality for martingale optimal transport on the line. *arXiv preprint arXiv:1507.00671*, 2015.

[10] Ahron Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

[11] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[12] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[13] Jean-David Benamou, Francis Collino, and Jean-Marie Mirebeau. Monotone and consistent discretization of the monge-ampere operator. *Mathematics of computation*, 85(302):2743–2775, 2016.

[14] Jean-David Benamou, Brittany D Froese, and Adam M Oberman. Numerical solution of the optimal transportation problem using the monge–ampère equation. *Journal of Computational Physics*, 260:107–126, 2014.

[15] Dimitri P Bertsekas. The auction algorithm: A distributed relaxation method for the assignment problem. *Annals of operations research*, 14(1):105–123, 1988.
[16] Christoph Brauer, Christian Clason, Dirk Lorenz, and Benedikt Wirth. A sinkhorn-newton method for entropic optimal transport. *arXiv preprint arXiv:1710.06635*, 2017.

[17] Lev 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.

[18] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[19] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

[20] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*, 2016.

[21] Roberto Cominetti and Jaime San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1-3):169–187, 1994.

[22] Alexander MG Cox and Jan Oblój. Robust pricing and hedging of double no-touch options. *Finance and Stochastics*, 15(3):573–605, 2011.

[23] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[24] Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[25] Hadrien De March. Local structure of the optimizer of multi-dimensional martingale optimal transport. *arXiv preprint arXiv:1805.09469*, 2018.

[26] Hadrien De March. Quasi-sure duality for multi-dimensional martingale optimal transport. *arXiv preprint arXiv:1805.01757*, 2018.

[27] Hadrien De March and Nizar Touzi. Irreducible convex paving for decomposition of multi-dimensional martingale transport plans. *arXiv preprint arXiv:1702.08298*, 2017.

[28] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[29] Alfred Galichon, Pierre Henry-Labordere, and Nizar Touzi. A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *The Annals of Applied Probability*, 24(1):312–336, 2014.

[30] Andrew V Goldberg and Robert E Tarjan. Finding minimum-cost circulations by canceling negative cycles. *Journal of the ACM (JACM)*, 36(4):873–886, 1989.

[31] Gaoyue Guo and Jan Obloj. Computational methods for martingale optimal transport problems. *arXiv preprint arXiv:1710.07911*, 2017.

[32] Cristian E Gutiérrez. *The Monge-Ampere equation*, volume 42. Springer, 2001.
[33] Pierre Henry-Labordere. *Model-free Hedging: A Martingale Optimal Transport Viewpoint*. CRC Press, 2017.

[34] Pierre Henry-Labordère and Nizar Touzi. An explicit martingale version of the one-dimensional brentier theorem. *Finance and Stochastics*, 20(3):635–668, 2016.

[35] David Hobson. The skorokhod embedding problem and model-independent bounds for option prices. In *Paris-Princeton Lectures on Mathematical Finance 2010*, pages 267–318. Springer, 2011.

[36] David Hobson and Martin Klimmek. Robust price bounds for the forward starting straddle. *Finance and Stochastics*, 19(1):189–214, 2015.

[37] David Hobson and Anthony Neuberger. Robust bounds for forward start options. *Mathematical Finance*, 22(1):31–56, 2012.

[38] David G Hobson. Robust hedging of the lookback option. *Finance and Stochastics*, 2(4):329–347, 1998.

[39] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

[40] JJ Kosowsky and Alan L Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. *Neural networks*, 7(3):477–490, 1994.

[41] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97, 1955.

[42] Christian Léonard. From the schrödinger problem to the monge–kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[43] Bruno Lévy. A numerical algorithm for l2 semi-discrete optimal transport in 3d. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1693–1715, 2015.

[44] Adrian S Lewis and Michael L Overton. Nonsmooth optimization via quasi-newton methods. *Mathematical Programming*, 141(1-2):135–163, 2013.

[45] Duncan McCallum and David Avis. A linear algorithm for finding the convex hull of a simple polygon. *Information Processing Letters*, 9(5):201–206, 1979.

[46] Quentin Mérigot. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library, 2011.

[47] Adam M Oberman and Yuanlong Ruan. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.

[48] Gabriel Peyré. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[49] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. Technical report, 2017.
[50] Julien Rabin and Nicolas Papadakis. Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 256–269. Springer, 2015.

[51] Cornelis Roos. An exponential example for terlaky’s pivoting rule for the criss-cross simplex method. *Mathematical Programming*, 46(1-3):79–84, 1990.

[52] Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *arXiv preprint arXiv:1610.06519*, 2016.

[53] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[54] Steve Smale. On the average number of steps of the simplex method of linear programming. *Mathematical programming*, 27(3):241–262, 1983.

[55] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

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

[57] Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, pages 423–439, 1965.

[58] Xiaolu Tan, Nizar Touzi, et al. Optimal transportation under controlled stochastic dynamics. *The annals of probability*, 41(5):3201–3240, 2013.

[59] Alexis Thibault, Lénaïc Chizat, Charles Dossal, and Nicolas Papadakis. Overrelaxed sinkhorn-knopp algorithm for regularized optimal transport. *arXiv preprint arXiv:1711.01851*, 2017.

[60] Matthew Thorpe, Serim Park, Soheil Kolouri, Gustavo K Rohde, and Dejan Slepčev. A transportation $l^p$ distance for signal analysis. *Journal of Mathematical Imaging and Vision*, 59(2):187–210, 2017.

[61] Neil S Trudinger and Xu-Jia Wang. On the second boundary value problem for monge-ampere type equations and optimal transportation. *arXiv preprint math/0601086*, 2006.

[62] Stephen Wright and Jorge Nocedal. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

[63] Danila A Zaev. On the monge–kantorovich problem with additional linear constraints. *Mathematical Notes*, 98(5-6):725–741, 2015.