A stochastic Gauss-Newton algorithm for regularized semi-discrete optimal transport

Bernard Bercu[1], Jérémie Bigot[1], Sébastien Gadat[2], Emilia Siviero[3]

[1]Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), Université de Bordeaux
[2]Toulouse School of Economics, Université Toulouse 1 Capitole
[3] LTCI, Télécom Paris, Institut Polytechnique de Paris

March 3, 2022

Abstract

We introduce a new second order stochastic algorithm to estimate the entropically regularized optimal transport cost between two probability measures. The source measure can be arbitrary chosen, either absolutely continuous or discrete, while the target measure is assumed to be discrete. To solve the semi-dual formulation of such a regularized and semi-discrete optimal transportation problem, we propose to consider a stochastic Gauss-Newton algorithm that uses a sequence of data sampled from the source measure. This algorithm is shown to be adaptive to the geometry of the underlying convex optimization problem with no important hyperparameter to be accurately tuned. We establish the almost sure convergence and the asymptotic normality of various estimators of interest that are constructed from this stochastic Gauss-Newton algorithm. We also analyze their non-asymptotic rates of convergence for the expected quadratic risk in the absence of strong convexity of the underlying objective function. The results of numerical experiments from simulated data are also reported to illustrate the finite sample properties of this Gauss-Newton algorithm for stochastic regularized optimal transport, and to show its advantages over the use of the stochastic gradient descent, stochastic Newton and ADAM algorithms.

Keywords: Stochastic optimization; Stochastic Gauss-Newton algorithm; Optimal transport; Entropic regularization; Convergence of random variables.

AMS classifications: Primary 62G05; secondary 62G20.

1 Introduction

1.1 Computational optimal transport for data science

The use of optimal transport (OT) and Wasserstein distances for data science has recently gained an increasing interest in various research fields such as machine learning [1, 20, 21, 23, 25, 42, 44], statistics [5, 7, 8, 13, 30, 35, 40, 46, 49] and image processing or computer vision [4, 12, 19, 26, 39, 45]. Solving a problem of OT between two probability measures μ and ν is known to be computationally challenging, and entropic regularization [15, 16] has emerged as an efficient tool to approximate and smooth the variational Wasserstein problems arising in computational optimal transport for data science. A detailed presentation of the recent

research field of computational optimal transport is given in [38], while recent reviews on the application of optimal transport to statistics can be found in [9, 36].

Recently, approaches [5, 23] based on first order stochastic algorithms have gained popularity to solve (possibly regularized) OT problems using data sampled from μ. These approaches are based on the semi-dual formulation [16] of regularized OT problems that can be rewritten as a *non-strongly convex* stochastic optimization problem. In this paper, for the purpose of obtaining stochastic algorithms for OT with faster convergence in practice, we introduce a second order stochastic algorithm to solve regularized semi-discrete OT between an arbitrary probability measure μ, typically absolutely continuous, and a *known* discrete measure ν with finite support of size J. More precisely, we focus on the estimation of an entropically regularized optimal transport cost $W_\varepsilon(\mu, \nu)$ between such measures (where $\varepsilon > 0$ is an entropic regularization parameter) using a class of stochastic quasi-Newton algorithms that we refer to as Gauss-Newton algorithms and which use the knowledge of a sequence (X_n) of independent random vectors sampled from μ.

Applications of semi-discrete optimal transport can be found in computational geometry and computer graphics [33, 34], as well as in the problem of optimal allocation of resources from online observations [5]. For an introduction to semi-discrete optimal transport problems and related references, we also refer to [38, Chapter 5]. In a deterministic setting where the full knowledge of μ is used and in the unregularized case, the convergence of a Newton algorithm for semi-discrete optimal transport has been studied in depth in [29]. An extension of the formulation of semi-discrete OT to include an entropic regularization is proposed in [16]. The main advantage of incorporating such a regularization term in classical OT is to obtain a dual formulation leading to a smooth convex minimization problem allowing the implementation of simple and numerically more stable algorithms as shown in [16]. The use of regularized semi-discrete OT has then found applications in image processing using generative adversarial networks [24, 43]. In these works, samples from the generative model are typically drawn from an absolutely continuous source measure in order to fit a discrete target distribution.

1.2 Main contributions and structure of the paper

As discussed above, we introduce a stochastic Gauss-Newton (SGN) algorithm for regularized semi-discrete OT for the purpose of estimating $W_\varepsilon(\mu, \nu)$, and the main goal of this paper is to study the statistical properties of such an approach. This algorithm is shown to be adaptive to the geometry of the underlying convex optimization problem with no important hyperparameter to be accurately tuned. Then, the main contributions of our work are to derive the almost sure rates of convergence, the asymptotic normality and the non-asymptotic rates of convergence (in expectation) of various estimators of interest that are constructed using the SGN algorithm to be described below. Although the underlying stochastic optimization problem is not strongly convex, fast rates of convergence can be obtained by combining the so-called notion of *generalized self-concordance* introduced in [2] that has been shown to hold for regularized OT in [5], and the Kurdyka-Łojasiewicz inequality as studied in [22]. We also report the results from various numerical experiments on simulated data to illustrate the finite sample properties of this algorithm, and to compare its performances with those of the stochastic gradient descent (SGD), stochastic Newton (SN) and ADAM [28] algorithms for stochastic regularized OT.

The paper is then organized as follows. The definitions of regularized semi-discrete OT

and the stochastic algorithms used for solving this problem are given in Section 2. The main results of the paper are stated in Section 3, while the important properties related to the regularized OT are given in Section 5. In Section 4, we describe a fast implementation of the SGN algorithm for regularized semi-discrete OT, and we assess the numerical ability of SGN to solve OT problems. In particular, we report numerical experiments on simulated data to compare the performances of various stochastic algorithms for regularized semi-discrete OT. The statistical properties of the SGN algorithm are established in an extended Section 6 that gathers the proof of our main results. Finally, two technical appendices A and B contain the proofs of auxiliary results.

2 A stochastic Gauss-Newton algorithm for regularized semi-discrete OT

In this section, we introduce the notion of regularized semi-discrete OT and the stochastic algorithm that we propose to solve this problem.

2.1 Notation, definitions and main assumptions on the OT problem

Let $\mathcal{X}$ and $\mathcal{Y}$ be two metric spaces. Denote by $\mathcal{M}_+^1(\mathcal{X})$ and $\mathcal{M}_+^1(\mathcal{Y})$ the sets of probability measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\mathbf{1}_J$ be the column vector of $\mathbb{R}^J$ with all coordinates equal to one, and denote by $\langle\ ,\ \rangle$ and $\|\ \|$ the standard inner product and norm in $\mathbb{R}^J$. We also use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the largest and smallest eigenvalues of a symmetric matrix A, whose spectrum will be denoted by $\mathrm{Sp}(A)$ and Moore-Penrose inverse by A^-. By a slight abuse of notation, we sometimes denote by $\lambda_{\min}(A)$ the smallest non-zero eigenvalue of a positive semi-definite matrix A. Finally, $\|A\|_2$ and $\|A\|_F$ stand for the operator and Frobenius norms of A, respectively. For $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, let $\Pi(\mu,\nu)$ be the set of probability measures on $\mathcal{X}\times\mathcal{Y}$ with marginals μ and ν. As formulated in [23], the problem of entropically regularized optimal transport between $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ is defined as follows.

Definition 2.1. *For any $(\mu,\nu) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})$, the Kantorovich formulation of the regularized optimal transport between μ and ν is the following convex minimization problem*

$$W_\varepsilon(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) + \varepsilon KL(\pi|\mu\otimes\nu), \tag{2.1}$$

where $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ is a lower semi-continuous function referred to as the cost function of moving mass from location x to y, $\varepsilon \geqslant 0$ is a regularization parameter, and KL stands for the Kullback-Leibler divergence between π and a positive measure ξ on $\mathcal{X}\times\mathcal{Y}$, up to the additive term $\int_{\mathcal{X}\times\mathcal{Y}} d\xi(x,y)$, namely

$$KL(\pi|\xi) = \int_{\mathcal{X}\times\mathcal{Y}} \Big(\log\Big(\frac{d\pi}{d\xi}(x,y)\Big) - 1\Big) d\pi(x,y).$$

For $\varepsilon = 0$, the quantity $W_0(\mu,\nu)$ is the *standard OT cost*, while for $\varepsilon > 0$, we refer to $W_\varepsilon(\mu,\nu)$ as the *regularized OT cost* between the two probability measures μ and ν. In

this framework, we shall consider cost functions that are lower semi-continuous and that satisfy the following standard assumption (see e.g. [48, Part I-4]), for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$0 \leqslant c(x, y) \leqslant c_{\mathcal{X}}(x) + c_{\mathcal{Y}}(y), \tag{2.2}$$

where $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are real-valued functions such that $\int_{\mathcal{X}} c_{\mathcal{X}}(x) d\mu(x) < +\infty$ and $\int_{\mathcal{Y}} c_{\mathcal{Y}}(y) d\nu(y) < +\infty$. Under condition (2.2), $W_\varepsilon(\mu, \nu)$ is finite regardless any value of the regularization parameter $\varepsilon \geqslant 0$. Moreover, note that $W_\varepsilon(\mu, \nu)$ can be negative for $\varepsilon > 0$, and that we always have the lower bound $W_\varepsilon(\mu, \nu) \geqslant -\varepsilon$. In this paper, we concentrate on *the regularized case where* $\varepsilon > 0$, and on the *semi-discrete setting* where $\mu \in \mathcal{M}_+^1(\mathcal{X})$ is an arbitrary probability measure (e.g. either discrete or absolutely continuous with respect to the Lesbesgue measure), and ν is a discrete measure with finite support $\mathcal{Y} = \{y_1, \dots, y_J\}$ that can be written as

$$\nu = \sum_{j=1}^{J} \nu_j \delta_{y_j}.$$

Here, δ stands for the standard Dirac measure, the locations $\{y_1, \dots, y_J\}$ as well as the positive weights $\{\nu_1, \dots, \nu_J\}$ are assumed to be known and summing up to one. We shall also use the notation

$$\min(\nu) = \min_{1 \leqslant j \leqslant J} \{\nu_j\} \qquad \text{and} \qquad \max(\nu) = \max_{1 \leqslant j \leqslant J} \{\nu_j\}.$$

We shall also sometimes refer to the *discrete setting* when μ is also a discrete measure. Let us now define the semi-dual formulation of the minimization problem (2.1) as introduced in [23]. In the semi-discrete setting and for $\varepsilon > 0$, using the semi-dual formulation of the minimization problem (2.1), it follows that $W_\varepsilon(\mu, \nu)$ can be expressed as the following convex optimization problem

$$W_\varepsilon(\mu, \nu) = - \inf_{v \in \mathbb{R}^J} H_\varepsilon(v) \tag{2.3}$$

with

$$H_\varepsilon(v) = \mathbb{E}[h_\varepsilon(X, v)] = \int_{\mathcal{X}} h_\varepsilon(x, v) d\mu(x), \tag{2.4}$$

where X stands for a random variable drawn from the unknown distribution μ, and for any $(x, v) \in \mathcal{X} \times \mathbb{R}^J$,

$$h_\varepsilon(x, v) = \varepsilon + \varepsilon \log\Big(\sum_{j=1}^{J} \exp\Big(\frac{v_j - c(x, y_j)}{\varepsilon}\Big)\nu_j\Big) - \sum_{j=1}^{J} v_j \nu_j. \tag{2.5}$$

Throughout the paper, we shall assume that, for any $\varepsilon > 0$, there exists $v^* \in \mathbb{R}^J$ that minimizes the function H_ε, leading to

$$W_\varepsilon(\mu, \nu) = -H_\varepsilon(v^*).$$

The above equality is the key result allowing to formulate (2.3) as a convex stochastic minimization problem, and to consider the issue of estimating $W_\varepsilon(\mu, \nu)$ in the setting of

stochastic optimization. For a discussion on sufficient conditions implying the existence of such a minimizer v^*, we refer to [5, Section 2]. As discussed in Section 5, the function H_ε possesses a one-dimensional subspace of global minimizers, defined by $\{v^* + t\boldsymbol{v}_J,\ t \in \mathbb{R}\}$ where

$$\boldsymbol{v}_J = \frac{1}{\sqrt{J}}\mathbf{1}_J.$$

Therefore, we will constrain our algorithm to live in $\langle \boldsymbol{v}_J \rangle^\perp$, which denotes the orthogonal complement of the one-dimensional subspace $\langle \boldsymbol{v}_J \rangle$ of $\mathbb{R}^J$ spanned by $\boldsymbol{v}_J$. In that setting, the OT problem (2.3) becomes identifiable.

2.2 Pre-conditionned stochastic algorithms

In the context of regularized OT, we first introduce a general class of stochastic pre-conditionned algorithms that are also referred to as quasi-Newton algorithms in the literature. Starting from Section 2.1, our approach is inspired by the recent works [5, 23], which use the property that

$$W_\varepsilon(\mu, \nu) = -H_\varepsilon(v^*) = -\min_{v \in \mathbb{R}^J} \mathbb{E}[h_\varepsilon(X, v)]$$

where $h_\varepsilon(x, v)$ is the smooth function defined by (2.5) that is simple to compute. For a sequence (X_n) of independent and identically distributed random variables sampled from the distribution μ, the class of pre-conditionned stochastic algorithms is defined as the following family of recursive stochastic algorithms to estimate the minimizer $v^* \in \langle \boldsymbol{v}_J \rangle^\perp$ of H_ε. These algorithms can be written as

$$\widehat{V}_{n+1} = P_J\left(\widehat{V}_n - n^\alpha S_n^{-1} \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)\right) \tag{2.6}$$

for some constant $0 \leqslant \alpha < 1/2$, where $\nabla_v h_\varepsilon$ stands for the gradient of h_ε with respect to v, $\widehat{V}_0$ is a random vector belonging to $\langle \boldsymbol{v}_J \rangle^\perp$, and S_n is a symmetric and *positive definite* $J \times J$ random matrix which is measurable with respect to the σ-algebra $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. In addition, P_J is the orthogonal projection matrix onto $\langle \boldsymbol{v}_J \rangle^\perp$,

$$P_J = I_J - \boldsymbol{v}_J \boldsymbol{v}_J^T.$$

This stochastic algorithm allows us to estimate $W_\varepsilon(\mu, \nu)$ by the recursive estimator

$$\widehat{W}_n = -\frac{1}{n}\sum_{k=1}^{n} h_\varepsilon(X_k, \widehat{V}_{k-1}). \tag{2.7}$$

The special case where $S_n = s^{-1} n I_J$ with some constant $s > 0$, corresponds to the well-known stochastic gradient descent (SGD) algorithm that has been introduced in the context of stochastic OT in [23], and recently investigated in [5]. Following some recent contributions [6, 14] in stochastic optimization using Newton-type stochastic algorithms, another potential choice is $S_n = \mathbb{S}_n$ where $\mathbb{S}_n$ is the natural Newton recursion defined as

$$\mathbb{S}_n = I_J + \sum_{k=1}^{n} \nabla_v^2 h_\varepsilon(X_k, \widehat{V}_{k-1}) = \mathbb{S}_{n-1} + \nabla_v^2 h_\varepsilon(X_n, \widehat{V}_{n-1}) \tag{2.8}$$

where $\nabla_v^2 h_\varepsilon$ stands for the Hessian matrix of h_ε with respect to v, We refer to the choice (2.8) for $\mathbb{S}_n$ as the stochastic Newton (SN) algorithm. Unfortunately, from a computational point of view, a major limitation of this SN algorithm is the need to compute the inverse of $\mathbb{S}_n$ at each iteration n in equation (2.6). As $\mathbb{S}_n$ is given by the recursive equation (2.8), it is tempting to use the Sherman-Morrison-Woodbury (SMW) formula [27], that is recalled in Lemma A.1, in order to compute $\mathbb{S}_n^{-1}$ from the knowledge of $\mathbb{S}_{n-1}^{-1}$ in a recursive manner. However, as detailed in Section A.2 of Appendix A, the Hessian matrix $\nabla_v^2 h_\varepsilon(X_n, \widehat{V}_{n-1})$ does not have a sufficiently low-rank structure that would lead to a fast recursive approach to compute $\mathbb{S}_n^{-1}$. Therefore, for the SN algorithm, the computational cost to evaluate $\mathbb{S}_n^{-1}$ appears to be of order $\mathcal{O}(J^3)$ which only leads to a feasible algorithm for very small values of J. This important computational limitation then drew our investigation towards the SGN algorithm instead of the SN approach.

2.3 The stochastic Gauss-Newton algorithm

Historically, the Gauss-Newton adaptation of the Newton algorithm consists in replacing the Hessian matrix $\nabla_v^2 h_\varepsilon(X_n, \widehat{V}_{n-1})$ by a tensor product of the gradient $\nabla_v h_\varepsilon(X_n, \widehat{V}_{n-1})$. In our framework, it leads to another pre-conditionned stochastic algorithm. We introduce S_n recursively as

$$\begin{aligned} S_n &= I_J + \sum_{k=1}^{n} \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1}) \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1})^T + \gamma \left(1 + \left\lfloor \frac{k}{J} \right\rfloor\right)^{-\beta} Z_k Z_k^T \qquad (2.9) \\ &= S_{n-1} + \nabla_v h_\varepsilon(X_n, \widehat{V}_{n-1}) \nabla_v h_\varepsilon(X_n, \widehat{V}_{n-1})^T + \gamma \left(1 + \left\lfloor \frac{n}{J} \right\rfloor\right)^{-\beta} Z_n Z_n^T, \end{aligned}$$

for some constants $\gamma > 0$ and $0 < \beta < 1/2$, and where $(Z_1, \ldots, Z_n)$ is a *deterministic* sequence of vectors defined, for all $1 \leqslant k \leqslant n$, by

$$Z_k = \sqrt{\nu_{\ell_k}} e_{\ell_k}$$

with $\ell_k = 1 + (k-1) \pmod J$, where $(e_1, \ldots, e_J)$ stands for the canonical basis of $\mathbb{R}^J$. We shall refer to the choice (2.9) for S_n as the regularized stochastic Gauss-Newton (SGN) algorithm and from now on, the notation S_n refers to this definition. We also use the convention that $S_0 = I_J$.

3 Main results on the SGN algorithm

Throughout this section, we investigate the statistical properties of the recursive sequence $(\widehat{V}_n)$ defined by (2.6) with $0 \leqslant \alpha < 1/2$, where (S_n) is the sequence of random matrices defined by (2.9) with $0 < \beta < 1/2$ that yields the SGN algorithm. The initial value $\widehat{V}_0$ is assumed to be a square integrable random vector that belongs to $\langle \boldsymbol{v}_J \rangle^\perp$. Then, thanks to the projection step in equation (2.6), it follows that for all $n \geqslant 1$, $\widehat{V}_n$ also belongs to $\langle \boldsymbol{v}_J \rangle^\perp$. To derive the convergence properties of the SGN algorithm, we first need to introduce the matrix-valued function $G_\varepsilon(v)$ defined as

$$G_\varepsilon(v) = \mathbb{E}\left[\nabla_v h_\varepsilon(X, v) \nabla_v h_\varepsilon(X, v)^T\right] \qquad (3.1)$$

that will be shown to be a key quantity to analyze the SGN algorithm. In particular, we shall derive our results under the following assumption on the smallest eigenvalue of

$G_\varepsilon(v^*)$ associated to eigenvectors belonging to $\langle \boldsymbol{v}_J \rangle^\perp$.

Invertibility assumption. The matrix $G_\varepsilon(v^*)$ satisfies

$$\min_{v \in \langle \boldsymbol{v}_J \rangle^\perp} \left\{ \frac{v^T G_\varepsilon(v^*) v}{\|v\|^2} \right\} > 0.$$

In all the sequel, we suppose that this invertibility assumption is satisfied. We denote by $G_\varepsilon^-(v^*)$ the Moore-Penrose inverse of $G_\varepsilon(v^*)$ and by $G_\varepsilon^{-1/2}(v^*)$ its square-root. We now discuss, in what follows, the next keystone inequality.

Proposition 3.1. *Assume that the regularization parameter $\varepsilon > 0$ satisfies*

$$\varepsilon \leqslant \frac{\min(\nu)}{\max(\nu) - \min(\nu)}. \tag{3.2}$$

Then, in the sense of partial ordering between positive semi-definite matrices, we have

$$G_\varepsilon(v^*) \leqslant \nabla^2 H_\varepsilon(v^*). \tag{3.3}$$

Inequality (3.3) is an important property of the SGN algorithm to prove its adaptivity to the geometry of the stochastic optimization problem (2.3). Of course, one can observe that no hyperparameter depending on the Hessian of H_ε needs to be tuned to run this algorithm provided that condition (3.2) holds. One can also remark that there is no restriction on the regularization parameter ε when ν is the uniform distribution, that is when $\nu_j = 1/J$, for all $1 \leqslant j \leqslant J$, implying that $\max(\nu) = \min(\nu)$. Throughout the paper, we suppose that condition (3.2) holds true. Below, we denote by $\lambda_{\min}^{\langle \boldsymbol{v}_J \rangle^\perp}(A)$ the smallest non-zero eigenvalue of a positive semi-definite matrix A, when the associated eigenvectors belong to $\langle \boldsymbol{v}_J \rangle^\perp$, the orthogonal complement of $\boldsymbol{v}_J$.

It immediately follows from inequality (3.3) that

$$1 \leqslant \lambda_{\min}^{\langle \boldsymbol{v}_J \rangle^\perp}(G_\varepsilon^{-1/2}(v^*) \nabla^2 H_\varepsilon(v^*) G_\varepsilon^{-1/2}(v^*)). \tag{3.4}$$

Inequality (3.4) will be a key property in this paper to derive the rates of convergence of the estimators obtained from the SGN algorithm. Note that the (pseudo) inverse of the Hessian matrix $\nabla^2 H_\varepsilon(v^*)$ somehow represents an ideal deterministic pre-conditioning matrix, whose use would lead to the second order Newton algorithm: this ideal pre-conditioned algorithm is *non-adaptive* since it requires the use of $\nabla^2 H_\varepsilon(v^*)$, which is unknown in practice.

Indeed, in our SGN algorithm, adaptivity is tightly related to the limiting recursion induced by Equation (2.6). If we admit (temporarily) the almost sure convergence of the SGN algorithm towards v^* and of $n^{-1} S_n$ towards $G_\epsilon(v^*)$, the recursion induced by (2.6) looks very similar to a discretization of a dynamical system with a step size $n^{-(1-\alpha)}$ and with a limiting linearized drift of the form $-G_\epsilon(v^*)^{-1} \nabla^2 H_\epsilon(v^*)(v - v^*)$. For further details on this point, we refer to the so-called ODE method (see *e.g.* [3]). The keystone property induced by Proposition 3.1 is that thanks to Equation (3.3) and Equation (3.4), the linearized drift of the limiting deterministic dynamical system has its eigenvalues that are lower bounded by 1, regardless of the value of the Hessian matrix $\nabla^2 H_\epsilon(v^*)$. This translates an adaptation of the algorithm to the curvature of H_ϵ near the target point v^*. Therefore, the matrix $G_\epsilon(v^*)$ that is learnt on-line, and automatically adapts to the eigenspaces associated to the smallest eigenvalues of $\nabla^2 H_\epsilon(v^*)$. Therefore, one may interpret (3.4) as the adaptivity of the SGN algorithm to the geometry of the semi-dual formulation of regularized OT.

3.1 Almost sure convergence

The almost sure convergence of the sequences $(\widehat{V}_n)$, $(\widehat{W}_n)$ and $(\overline{S}_n)$ are as follows where $\overline{S}_n = \frac{1}{n} S_n$.

Theorem 3.1. *Assume that $\alpha \in [0, 1/2[$ and $\alpha + \beta < 1/2$. Then, we have*

$$\lim_{n \to +\infty} \widehat{V}_n = v^* \qquad a.s. \tag{3.5}$$

and

$$\lim_{n \to +\infty} \overline{S}_n = G_\varepsilon(v^*) \qquad a.s. \tag{3.6}$$

The following result is an immediate corollary of Theorem 3.1, thanks to the continuity of the function h_ε.

Corollary 3.1. *Assume that $\alpha \in [0, 1/2[$ and $\alpha + \beta < 1/2$. Suppose that the cost function c satisfies, for any $1 \leqslant j \leqslant J$,*

$$\int_{\mathcal{X}} c^2(x, y_j) d\mu(x) < +\infty. \tag{3.7}$$

Then, we have

$$\lim_{n \to +\infty} \widehat{W}_n = W_\varepsilon(\mu, \nu) \qquad a.s.$$

We now derive results on the almost sure rates of convergence of the sequences $(\widehat{V}_n)$ and $(\overline{S}_n)$ that are the keystone in the proof of the asymptotic normality of the estimator $\widehat{V}_n$ studied in Section 3.2. We emphasize that we restrict our study to the case $\alpha = 0$, which yields the fastest rates of convergence and that corresponds to the meaningful situation from the numerical point of view.

Theorem 3.2. *Assume that $\alpha = 0$. Then, we have the almost sure rate of convergence*

$$\left\| \widehat{V}_n - v^* \right\|^2 = \mathcal{O}\Big(\frac{\log n}{n} \Big) \qquad a.s. \tag{3.8}$$

In addition, we also have

$$\left\| \overline{S}_n - G_\varepsilon(v^*) \right\|_F = \mathcal{O}\Big(\frac{1}{n^\beta} \Big) \qquad a.s. \tag{3.9}$$

and

$$\left\| \overline{S}_n^{-1} - G_\varepsilon^-(v^*) \right\|_F = \mathcal{O}\Big(\frac{1}{n^\beta} \Big) \qquad a.s. \tag{3.10}$$

3.2 Asymptotic normality

The asymptotic normality of our estimates depends on the magnitude of the smallest eigenvalue (associated to eigenvectors belonging to $\langle \boldsymbol{v}_J \rangle^\perp$) of the matrix

$$\Gamma_\varepsilon(v^*) = G_\varepsilon^{-1/2}(v^*) \nabla^2 H_\varepsilon(v^*) G_\varepsilon^{-1/2}(v^*). \tag{3.11}$$

Thanks to the key inequality (3.4), we have that the smallest eigenvalue of $\Gamma_\varepsilon(v^*)$ is always greater than 1, in the sense that

$$\min_{v \in \langle \boldsymbol{v}_J \rangle^\perp} \frac{v^T \Gamma_\varepsilon(v^*) v}{\|v\|^2} \geqslant 1.$$

One can observe that we also restrict our study to the case $\alpha = 0$ which yields the usual $\sqrt{n}$ rate of convergence for the central limit theorem that is stated below.

Theorem 3.3. *Assume that $\alpha = 0$. Then, we have the asymptotic normality*

$$\sqrt{n}\big(\widehat{V}_n - v^*\big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, G_\varepsilon^{-1/2}(v^*)\big(2\Gamma_\varepsilon(v^*) - P_J\big)^{-} G_\varepsilon^{-1/2}(v^*)\Big). \tag{3.12}$$

In addition, suppose that the cost function c satisfies, for any $1 \leqslant j \leqslant J$,

$$\int_{\mathcal{X}} c^4(x, y_j) d\mu(x) < +\infty. \tag{3.13}$$

Then, we also have

$$\sqrt{n}\big(\widehat{W}_n - W_\varepsilon(\mu, \nu)\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \sigma_\varepsilon^2\big) \tag{3.14}$$

where the asymptotic variance $\sigma_\varepsilon^2 = \mathbb{E}[h_\varepsilon^2(X, v^*)] - W_\varepsilon^2(\mu, \nu)$.

In order to discuss the above result on the asymptotic normality of $\widehat{V}_n$, we denote by

$$\Sigma_\varepsilon(v^*) = G_\varepsilon^{-1/2}(v^*)\big(2\Gamma_\varepsilon(v^*) - P_J\big)^{-} G_\varepsilon^{-1/2}(v^*)$$

the asymptotic covariance matrix in (3.12). One can check that $\Sigma_\varepsilon(v^*)$ satisfies the Lyapunov equation

$$\Big(\frac{1}{2} P_J - A\nabla^2 H_\varepsilon(v^*)\Big)\Sigma_\varepsilon(v^*) + \Sigma_\varepsilon(v^*)\Big(\frac{1}{2} P_J - A\nabla^2 H_\varepsilon(v^*)\Big)^T = -A G_\varepsilon(v^*) A \tag{3.15}$$

with $A = G_\varepsilon^{-}(v^*)$. Moreover, one can observe that

$$G_\varepsilon(v^*) = \lim_{n \to +\infty} \mathbb{E}[\varepsilon_{n+1} \varepsilon_{n+1}^T | \mathcal{F}_n] \qquad \text{a.s.}$$

is the asymptotic covariance matrix of the martingale increment $\varepsilon_{n+1} = \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) - \nabla H_\varepsilon(\widehat{V}_n)$. Hence, to better interpret the asymptotic covariance matrix $\Sigma_\varepsilon(v^*)$, let us consider the following sub-class of pre-conditionned stochastic algorithms

$$\widetilde{V}_{n+1} = P_J\Big(\widetilde{V}_n - \frac{1}{n} A \nabla_v h_\varepsilon(X_{n+1}, \widetilde{V}_n)\Big), \tag{3.16}$$

where A is a deterministic positive semi-definite matrix satisfying the stability condition

$$A\nabla^2 H_\varepsilon(v^*) \geqslant \frac{1}{2} P_J. \tag{3.17}$$

Then, adapting well-known results on stochastic optimisation (see e.g. [18, 37]), one may prove that

$$\sqrt{n}\big(\widetilde{V}_n - v^*\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \Sigma(A)\big)$$

where $\Sigma(A)$ is the solution of the Lyapunov equation (3.15) with $\Sigma(A)$ instead of $\Sigma_\varepsilon(v^*)$. Hence, the asymptotic normality of the SGN algorithm coincides with the one of the pre-conditionned stochastic algorithm (3.16) for the choice $A = G_\varepsilon^-(v^*)$. Hence, the main advantage of the SGN algorithm is to be fully data-driven as $G_\varepsilon^-(v^*)$ is obviously unknown. Among the deterministic pre-conditionning matrices satisfying condition (3.17), it is also known (see e.g. [18, 37]) that the best choice is to take $A = \nabla^2 H_\varepsilon^-(v^*)$ that corresponds to an ideal Newton algorithm and which yields the optimal asymptotic covariance matrix

$$\Sigma^* = \Sigma\left(\nabla^2 H_\varepsilon(v^*)\right) = \nabla^2 H_\varepsilon(v^*)^- G_\varepsilon(v^*) \nabla^2 H_\varepsilon(v^*)^- \leqslant \Sigma_\varepsilon(v^*).$$

Therefore, the SGN algorithm does not yield an estimator $\widehat{V}_n$ having an asymptotically optimal covariance matrix. Note that, as shown in [5, Theorem 3.4], using an average version of the standard SGD algorithm, that is with $S_n = s^{-1} n I_J$ and $0 < \alpha < 1/2$, allows to obtain an estimator having an asymptotic distribution with optimal covariance matrix Σ^*. However, in numerical experiments, it appears that the choice of s for the averaged SGD algorithm is crucial but difficult to tune. The results from [5] suggests to take $s = \varepsilon/(2\min(\nu))$ which follows from the property that

$$\lambda_{\min}\left(\nabla^2 H_\varepsilon(v^*)\right) \geqslant \frac{\min(\nu)}{\varepsilon},$$

that is discussed in Section 5. Hence, the choice $s = \varepsilon/(2\min(\nu))$ ensures that the pre-conditioning matrix $A = s^{-1} I_J$ satisfies the stability condition (3.17). However, as shown by the numerical experiments carried out in Section 4, it appears that the SGN algorithm automatically adapts to the geometry of the optimisation problem with better results than the SGD algorithm. Finally, we remark from the asymptotic normality (3.14) and [5, Theorem 3.5] that the asymptotic variance of the recursive estimator $\widehat{W}_n$ is the same when $\widehat{V}_n$ is either computed using the SGN or the SGD algorithm.

3.3 Non-asymptotic rates of convergence

The last contribution of our paper is to derive non-asymptotic upper bounds on the expected risk of various estimators arising from the use of the SGN algorithm when (S_n) is the sequence of positive definite matrices defined by (2.9). In particular, we derive the rate of convergence of the expected quadratic risks

$$\mathbb{E}\big[\big\|\widehat{V}_n - v^*\big\|^2\big] \qquad \text{and} \qquad \mathbb{E}\big[\big\|\overline{S}_n - G_\varepsilon(v^*)\big\|_F^2\big].$$

We also analyze the rate of convergence of the expected excess risk $W_\varepsilon(\mu,\nu) - \mathbb{E}[\widehat{W}_n]$ of the recursive estimator $\widehat{W}_n$ defined by (2.7) used to approximate the regularized OT cost $W_\varepsilon(\mu,\nu)$.

Theorem 3.4. *Assume that $\alpha \in]0, 1/2[$ and that $4\beta < 1 - 2\alpha$. Then, there exists a positive constant c_ε such that for any $n \geqslant 1$,*

$$\mathbb{E}\big[\big\|\widehat{V}_n - v^*\big\|^2\big] \leqslant \frac{c_\varepsilon}{n^{1-\alpha}} \qquad \textit{and} \qquad \mathbb{E}\big[\big\|\overline{S}_n - G_\varepsilon(v^*)\big\|_F^2\big] \leqslant \frac{c_\varepsilon}{n^{2\beta}}. \tag{3.18}$$

Moreover, we also have

$$\big|\mathbb{E}[\widehat{W}_n] - W_\varepsilon(\mu,\nu)\big| \leqslant \frac{c_\varepsilon}{n^{1-\alpha}}, \tag{3.19}$$

and if the cost function c satisfies $\int_{\mathcal{X}} c^2(x, y_j) d\mu(x) < +\infty$, for any $1 \leqslant j \leqslant J$, then

$$\mathbb{E}\big[\big|\widehat{W}_n - W_\varepsilon(\mu, \nu)\big|\big] \leqslant \frac{c_\varepsilon}{\sqrt{n}}. \tag{3.20}$$

Note that the value of the constant c_ε appearing in Theorem 3.4 may also depend on α and β, but we remove this dependency in the notation to simplify the presentation. One can observe that choosing $\alpha > 0$ allows the algorithm to be fully adaptative in the sense that no important hyperparameter needs to be tune to obtain non-asymptotic rates of convergence. The case $\alpha = 0$ could also be considered but this will require to introduce a multiplicative positive constant c in the definition of the SGN algorithm by replacing $n^\alpha S_n^{-1}$ in equation (2.6) by cS_n^{-1}. Then, provided that c is sufficiently large, one may obtain faster rate of convergence for the expected quadratic risk of the order $\mathcal{O}(\log(n)/n)$. However, in our numerical experiments, we have found that introducing such a large multiplicative constant c makes the convergence of the SGN algorithm too slow. Therefore, results on non-asymptotic convergence rates in the case $\alpha = 0$ are not reported here.

4 Implementation of the SGN algorithm and numerical experiments

In this section, we first discuss computational considerations on the implementation of the SGN algorithm, and we also make several remarks to justify its use. Then, we report the results of numerical experiments.

4.1 A fast recursive approach to compute S_n^{-1}.

In this paragraph, we discuss on the computational benefits of using the Gauss-Newton method as an alternative to the Newton algorithm. A key point to define the SGN algorithm consists in replacing in equation (2.8) that defines the SN algorithm, the positive definite Hessian matrices $\nabla_v^2 h_\varepsilon(X_k, \widehat{V}_{k-1})$ by the tensor product $\nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1}) \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1})^T$ of the gradient of h_ε at $(X_k, \widehat{V}_{k-1})$. A second important ingredient in the definition (2.9) of the SGN algorithm is the additive regularization terms $\gamma \left(1 + \left\lfloor \frac{k}{J} \right\rfloor\right)^{-\beta} Z_k Z_k^T$ whose role is discussed in the sub-section below.

Note that $Z_k Z_k^T = \nu_{\ell_k} e_{\ell_k} e_{\ell_k}^T$ (with $\ell_k = (k-1) \pmod J + 1$) is a diagonal matrix, such that all its diagonal elements are equal to zero, except the ℓ_k-th one which is equal to ν_{ℓ_k}. In this manner, the difference $S_n - S_{n-1} = \phi_n \phi_n^T + \gamma \left(1 + \left\lfloor \frac{n}{J} \right\rfloor\right)^{-\beta} Z_n Z_n^T$ is thus the sum of two rank one matrices, where $\phi_n = \nabla_v h_\varepsilon(X_n, \widehat{V}_{n-1})$. Therefore, one may easily obtain S_n^{-1} from the knowledge of S_{n-1}^{-1} as follows. Introducing the intermediate matrix $S_{n-\frac{1}{2}} = S_{n-1} + \gamma \left(1 + \left\lfloor \frac{n}{J} \right\rfloor\right)^{-\beta} Z_n Z_n^T$, we observe that $S_n = S_{n-\frac{1}{2}} + \phi_n \phi_n^T$. Consequently, by applying the SMW formula (A.3), we first notice that

$$\begin{aligned} S_{n-\frac{1}{2}}^{-1} &= (S_{n-1} + \gamma \left(1 + \left\lfloor \frac{n}{J} \right\rfloor\right)^{-\beta} Z_n Z_n^T)^{-1} \\ &= S_{n-1}^{-1} - (Z_n^T S_{n-1}^{-1} Z_n + \gamma^{-1} \left(1 + \left\lfloor \frac{n}{J} \right\rfloor\right)^{\beta})^{-1} S_{n-1}^{-1} Z_n Z_n^T S_{n-1}^{-1} \end{aligned}$$

Using that $Z_n = \sqrt{\nu_{\ell_n}} e_{\ell_n}$, we furthermore have that

$$S_{n-\frac{1}{2}}^{-1} = S_{n-1}^{-1} - \nu_{\ell_n} \frac{(S_{n-1}^{-1})_{\cdot,\ell_n} (S_{n-1}^{-1})_{\ell_n,\cdot}^T}{\nu_{\ell_n} (S_{n-1}^{-1})_{\ell_n,\ell_n} + \gamma^{-1} \left(1 + \left\lfloor \frac{n}{J} \right\rfloor\right)^{\beta}}. \tag{4.1}$$

Secondly, applying again the SMW formula (A.3), we obtain that

$$S_n^{-1} = S_{n-\frac{1}{2}}^{-1} - \frac{S_{n-\frac{1}{2}}^{-1} \phi_n \phi_n^T S_{n-\frac{1}{2}}^{-1}}{\phi_n^T S_{n-\frac{1}{2}}^{-1} \phi_n + 1}. \tag{4.2}$$

Hence, the recursive formulas (4.1) and (4.2) allow, at each iteration n, a much more faster computation of S_n^{-1} from the knowledge of S_{n-1}^{-1}, which is a key advantage of the SGN algorithm over the use of the SN algorithm. Indeed, the cost of computing S_n^{-1} using the above recursive formulas is that of matrix vector multiplication which is of order $\mathcal{O}(J^2)$.

4.2 The role of regularization.

Let us denote by $R_n = \sum_{k=1}^{n} \gamma \left(1 + \left\lfloor \frac{k}{J} \right\rfloor\right)^{-\beta} Z_k Z_k^T$ the sum of the deterministic regularization terms in (2.9) implying that S_n can be decomposed as

$$S_n = I_J + \sum_{k=1}^{n} \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1}) \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1})^T + R_n. \tag{4.3}$$

If $n = pJ$ for some integer $p \geqslant 1$, the regularization by the matrices $\gamma \left(1 + \left\lfloor \frac{k}{J} \right\rfloor\right)^{-\beta} Z_k Z_k^T$ in (2.9) sum up to a simple expression given by

$$R_n = \left(\sum_{m=1}^{p} m^{-\beta} \right) \gamma \, \mathrm{diag}(\nu).$$

The following two important comments can be made to clarify the role of this regularization effect:

- adding the supplementary matrix R_n in (4.3) implies that S_n is invertible as soon as $n \geq J$ with a known lower bound on its smallest eigenvalue. Indeed, thanks to the condition $0 < \beta < 1/2$ and to the property that

$$\left(\frac{1}{p} \sum_{m=1}^{p} m^{-\beta} \right) \sim \frac{1}{1-\beta} p^{-\beta},$$

the additive term R_n allows to regularize the smallest eigenvalue of S_n. This is important for the evolution of the stochastic algorithm: this regularization allows to show that $\widehat{V}_n$ converges almost surely to v^*. More precisely, while $\hat{V}_{n+1} - \hat{V}_n$ is essentially modified in the direction $-n^\alpha S_n^{-1} \nabla_v h_\epsilon(X_{n+1}, \hat{V}_n)$, it is well known that too large step sizes are prohibited to obtain a good behavior of stochastic algorithms. Therefore, taking a sufficiently small β guarantees a suitable upper bound of the increments of the SGN, that in turn limits the effect of the noise at each iteration of the algorithm.

- the growth of R_n is sublinear for large values of n whereas

$$\sum_{k=1}^{n} \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1}) \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1})^T$$

grows linearly with n so that this last term will become dominant in the decomposition of S_n, inducing a "learning" of the curvature of the landscape function H_ϵ. Recalling that $\overline{S}_n = \frac{1}{n} S_n$, it will be shown in Section 6 that

$$\lim_{n \to +\infty} \overline{S}_n = G_\varepsilon(v^*) \quad \text{a.s.}$$

with

$$G_\varepsilon(v^*) = \mathbb{E}\left[\nabla_v h_\varepsilon(X, v^*) \nabla_v h_\varepsilon(X, v^*)^T\right] = \mathrm{diag}(\nu) - \nu\nu^T - \varepsilon \nabla^2 H_\varepsilon(v^*),$$

where the last equality above follows the proof of Proposition 3.1. Hence, when $n \longrightarrow +\infty$, the regularization disappears as long as $\beta > 0$. Note that this would not be the case if β was chosen to be equal to 0.

To sum up, taking $\beta \in (0, 1/2)$ will be a crucial assumption to derive the almost sure convergence rates that are stated in Theorem 3.2.

4.3 Numerical experiments

In this section, we report numerical results on the performances of stochastic algorithms for regularized optimal transport when the source measure μ is either discrete or absolutely continuous. We shall compare the SGD, ADAM, SGN and SN algorithms. For the SGD algorithm, following the results in [5], we took $\alpha = 1/2$ and $S_n = s^{-1} n I_J$ with $s = \varepsilon/(2\min(\nu))$. The ADAM algorithm has been implemented following the parametrization made in the seminal paper [28] except the value of the stepsize (as defined in [28, Algorithm 1]) that is set to 0.005 instead of 0.001, which improves the performances of ADAM in our numerical experiments. For the SGN algorithm, we set $\alpha = 0$ and we have taken $\gamma = 10^{-3}$ (a small value) and $\beta = 0.49$. For the results reported in this paper, we have found that the performances of the SGN algorithm are not very sensitive to the value of $\beta \in (0, 1/2)$. Finally, for the SN algorithm, we chose $\alpha = 0$ and $S_n = \mathbb{S}_n$ as defined by (2.8).

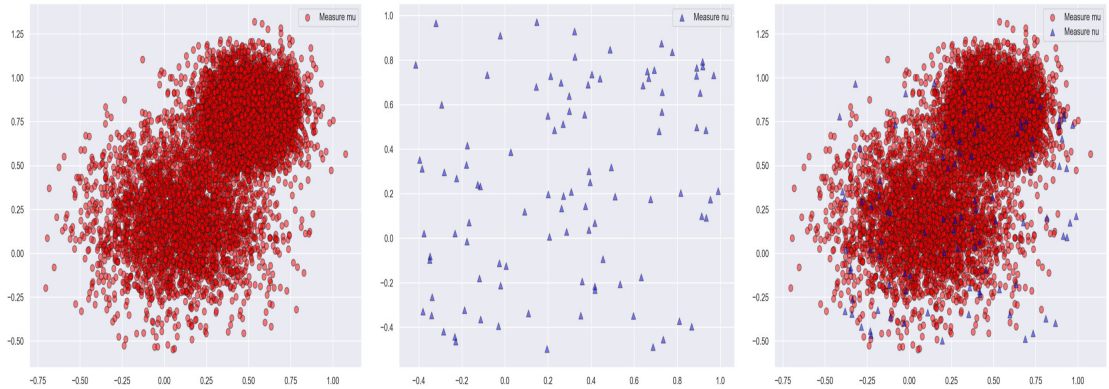
In the discrete setting, we shall also compare the performances of these stochastic algorithms to those of the Sinkhorn algorithm [15], which is a deterministic iterative procedure that uses the full knowledge of the measures μ and ν at each iteration. Let us recall that, for the SGD and ADAM algorithms, the computational cost of one iteration from n to $n+1$ is of order $\mathcal{O}(J)$, while it is of order $\mathcal{O}(J^2)$ for the SGN algorithm and $\mathcal{O}(J^3)$ for the SN algorithm. Each iteration of the Sinkhorn algorithm is of order $\mathcal{O}(IJ)$, where I denotes the size of the support of μ in the discrete setting.

In these numerical experiments, we investigate the numerical behavior of the recursive estimators $\widehat{W}_n$ and $\widehat{V}_n$. The performances of the various stochastic algorithms used to compute these estimators are compared in terms of the expected excess risks $\mathbb{E}\big[\big|\widehat{W}_n - W_\varepsilon(\mu,\nu)\big|\big]$ and $\mathbb{E}\big[\big\|\widehat{V}_n - v^*\big\|^2\big]$. For the SGN algorithm, we also analyze the convergence of the estimator $\overline{S}_n$ to the matrix $G_\varepsilon(v^*)$. The expected value involved in these expected risks is approximated using 100 Monte-Carlo replications. When the measure μ is discrete,

we use the Sinkhorn algorithm [15] to preliminary compute $W_\varepsilon(\mu,\nu)$ and v^*. When μ is absolutely continuous, the regularized OT cost $W_\varepsilon(\mu,\nu)$ is preliminary approximated by running the SN algorithm with a very large value of iterations (e.g. $n = 10^6$). To the best of our knowledge, apart from stochastic approaches as in [23], there is no other method to evaluate $W_\varepsilon(\mu,\nu)$ in the semi-discrete setting. Note that we shall compare the evolution of these excess risks as a function of the computational time (observed on the computer) of each algorithm. Moreover, the estimators $\widehat{W}_n$ and $\widehat{V}_n$ obviously depends on the regularization parameter ε. However, for the sake of simplicity, we have chosen to denote them as $\widehat{W}_n$ and $\widehat{V}_n$, although we carry out numerical experiments for different values of ε. Finally, we also analyze the asymptotic distributions of $\widehat{W}_n$ and $\widehat{V}_n$ to illustrate the results on asymptotic normality given in Section 3.2.

4.3.1 Discrete setting in dimension $d = 2$

In this section, the cost function is chosen as the squared Euclidean distance that is $c(x,y) = \|x-y\|^2$. We focus our attention when both $\mu = \sum_{i=1}^I \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^J \nu_j \delta_{y_j}$ are uniform discrete measures supported on $\mathbb{R}^2$, that is $\mu_i = \frac{1}{I}$ and $\nu_j = \frac{1}{J}$. The points $(x_i)_{1\leqslant i\leqslant I}$ (resp. $(y_j)_{1\leqslant j\leqslant J}$) are drawn randomly (once for all) from a Gaussian mixture with two components (resp. from the uniform distribution on $[0,1]^2$). An example of two such measures is displayed in Figure 1 for $I = 10^4$ and $J = 100$. The number of iterations of the four stochastic algorithms is fixed to $n = 10^5$ except for the experiments on the asymptotic distribution of $\widehat{W}_n$ and $\widehat{V}_n$, where n is let being larger. Finally, the Sinkhorn algorithm is let running until convergence is reached to provide a reference value for $W_\varepsilon(\mu,\nu)$ and v^* considered as the ground truth.



(a) (b) (c)

Figure 1: (a) Discrete measure μ supported on $I = 10^4$ points drawn from a mixture of two Gaussian distributions, and (b) discrete measure ν supported on $J = 100$ points randomly drawn from the uniform distribution on $[0,1]^2$. (c) Superposition of μ and ν.

Convergence of the excess risks. We first report results for $I = 10^4$ (size of the support of μ) and $J \in \{100, 400\}$ (size of the support of ν), and two small values of the regularization

parameter $\varepsilon \in \{0.01, 0.005\}$. For different combinations of these hyperparameters, we display from Figure 2 to Figure 5 the value of the expected excess risks (in logarithmic scale) $\mathbb{E}\big[\big|\widehat{W}_n - W_\varepsilon(\mu,\nu)\big|\big]$ and $\mathbb{E}\big[\big\|\widehat{V}_n - v^*\big\|^2\big]$ as functions of the averaged (along the 100 Monte Carlo replications) computational time of each iteration of the stochastic algorithms. We also draw the evolution of the metrics (in logarithmic scale) $\big|W_k - W_\varepsilon(\mu,\nu)\big|$ and $\big\|V_k - v^*\big\|^2$ as functions of the computational time of the iterations of the Sinkhorn algorithm, where $W_k \in \mathbb{R}$ and $V_k \in \mathbb{R}^J$ are the output of the Sinkhorn algorithm at its k-th iteration.

In Figure 2 to Figure 5, the various curves are displayed as functions of the computational time until the convergence of the Sinkhorn algorithm is reached, that is until $k = k_{\max}$ (the maximum number of Sinkhorn iterations). Note that for $k \approx k_{\max}$ then $\big|W_k - W_\varepsilon(\mu,\nu)\big| \approx 0$ and $\big\|V_k - v^*\big\|^2 \approx 0$. Hence, for such large values of k, these metrics have necessarily smaller values than those that are used to evaluate the stochastic algorithms. In the discussion that follows, we thus consider that the stochastic algorithms have reached convergence when the values of either $\mathbb{E}\big[\big|\widehat{W}_n - W_\varepsilon(\mu,\nu)\big|\big]$ or $\mathbb{E}\big[\big\|\widehat{V}_n - v^*\big\|^2\big]$ stabilize, although these metrics may be larger than the metrics used to evaluate the Sinkhorn algorithm for large values of the computational time. This is due to the randomness of the stochastic algorithms and their resulting positive variance (even for large values of n). Then, the following comments can be made from the output of these numerical experiments.

- For $I = 10^4$, $J = 100$ and $\varepsilon = 0.01$, the four stochastic algorithms reach convergence faster than the Sinkhorn algorithm. The convergence is much faster for the metric $\mathbb{E}\big[\big\|\widehat{V}_n - v^*\big\|^2\big]$ than for the metric $\mathbb{E}\big[\big|\widehat{W}_n - W_\varepsilon(\mu,\nu)\big|\big]$.

- For $I = 10^4$, $J = 100$ and $\varepsilon = 0.005$, SGD fails to converge either for the estimator $\widehat{W}_n$ or the estimator $\widehat{V}_n$. For this smallest value of ϵ, SGN and SN have similar performances for the metric $\mathbb{E}\big[\big\|\widehat{V}_n - v^*\big\|^2\big]$. The SN algorithm is slightly better than SGN for the metric $\mathbb{E}\big[\big|\widehat{W}_n - W_\varepsilon(\mu,\nu)\big|\big]$. We also observe that SN and SGN converge much faster than Sinkhorn, and that they have better performances than ADAM for the two metrics.

- For $I = 10^4$, $J = 400$ and $\varepsilon \in \{0.01, 0.005\}$, it can be seen that the SGD algorithm does not converge. For the metric $\mathbb{E}\big[\big\|\widehat{V}_n - v^*\big\|^2\big]$, the convergence of the SN and SGN algorithms is much faster than Sinkhorn, and these two algorithms outperform ADAM.

Therefore, these numerical experiments suggest that the SGN algorithm has interesting benefits over the SGD, ADAM and Sinkhorn algorithms for moderate values of J and for small values of the regularization parameter ε. In these settings, SGN seems to be particularly relevant for the estimation of v^*, and it reaches performances similar to those of SN for the metric $\mathbb{E}\big[\big\|\widehat{V}_n - v^*\big\|^2\big]$. The SGN algorithm may also converge much faster than the Sinkhorn algorithm for either the estimation of $W_\varepsilon(\mu,\nu)$ or v^* as the size I of the support of μ is large.

Asymptotic distribution of the stochastic algorithms. Now, we illustrate the results from Section 3.2 on the asymptotic distributions of $\widehat{W}_n$ and $\widehat{V}_n$. To this end, we consider the setting $I = 10^3$ and $J = 50$. Then, we display in Figure 6 for $\varepsilon = 0.1$ and $n = 2 \times 10^5$

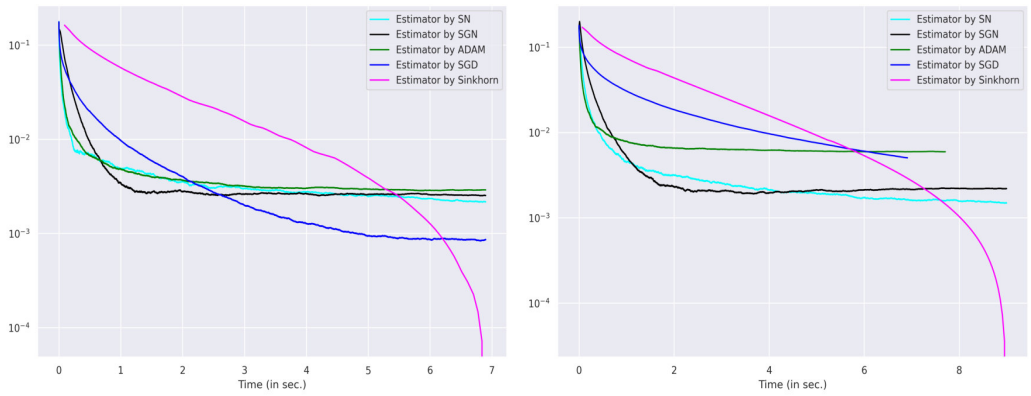
iterations (resp. Figure 7 for $\varepsilon = 0.01$ and $n = 4 \times 10^5$ iterations) the histograms of 200 independent realizations of

$$\widetilde{W}_n = \frac{\sqrt{n}\left(\widehat{W}_n - W_\varepsilon(\mu,\nu)\right)}{\widehat{\sigma}_n}$$

using each of the four stochastic algorithms, where

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n} h_\varepsilon^2(X_k, \widehat{V}_{k-1}) - \widehat{W}_n^2,$$

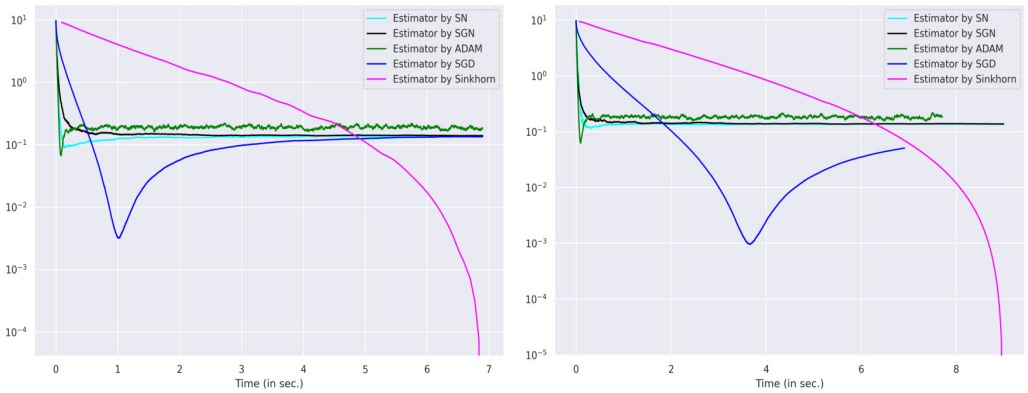
is a recursive estimator of the asymptotic variance of $\widehat{W}_n$ that has been introduced in [5]. For all the algorithms, it can be seen in Figure 6 and Figure 7 that $\widehat{W}_n$ is normally distributed. For the SGD and the SN algorithms, the histograms of $\widetilde{W}_n$ are very close to the standard Gaussian distribution, while the SGN is seen to be slightly biased. The bias is much more important for the ADAM algorithm.



(a) $\varepsilon = 0.01$ (b) $\varepsilon = 0.005$

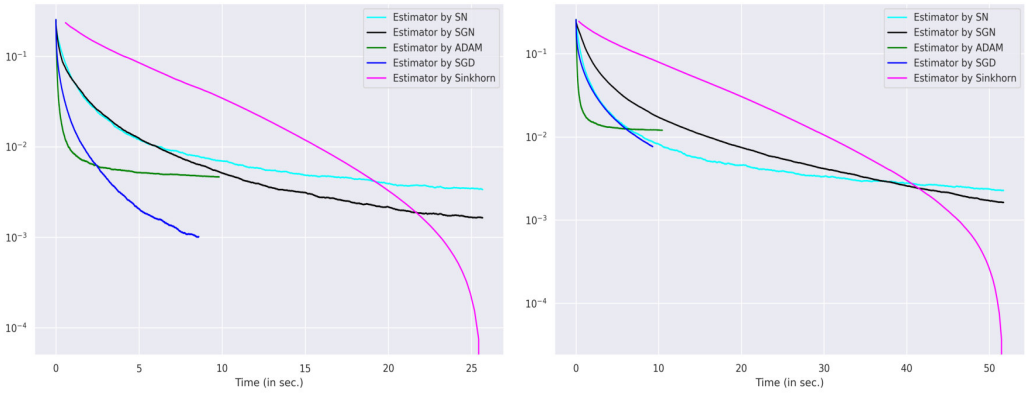
Figure 2: Discrete setting with $I = 10^4$ and $J = 100$ with $n = 10^5$ iterations. Expected excess risk (in logarithmic scale) $\log(\mathbb{E}\big[\big|\widehat{W}_n - W_\varepsilon(\mu,\nu)\big|\big])$ (resp. metric $\log(\big|W_k - W_\varepsilon(\mu,\nu)\big|)$) as a function of the averaged computational cost of the iterations of the four stochastic algorithms (resp. the Sinkhorn algorithm) for different values of the regularization parameter ε.

In Figure 8 and Figure 9, we also display the histograms of 200 independent realizations of $\widetilde{V}_n = n\big\|\widehat{V}_n - v^*\big\|^2$ for the four stochastic algorithms. The distribution of $\widetilde{V}_n$ has the shape of a χ^2-distribution but the "number of degrees of freedom" is highly varying from one algorithm to the other. It can be seen from Figure 8 and Figure 9, that $\widetilde{V}_n$ reaches its smallest variance for the SN algorithm, and that the second smallest variance is obtained with the SGN algorithm. The SGD and the ADAM algorithms finally have a much larger variance.



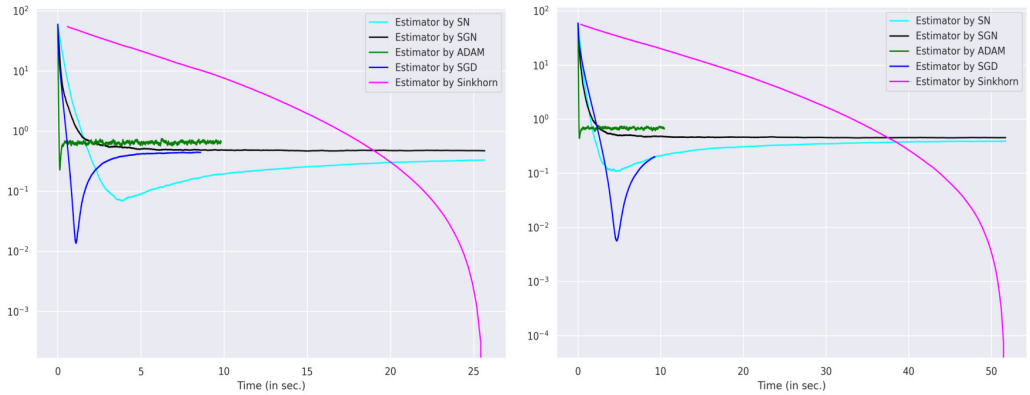
(a) $\varepsilon = 0.01$ (b) $\varepsilon = 0.005$

Figure 3: Discrete setting with $I = 10^4$ and $J = 100$ with $n = 10^5$ iterations. Expected excess risk (in logarithmic scale) $\log(\mathbb{E}[\|\widehat{V}_n - v^*\|^2])$ (resp. metric $\log(\|V_k - v^*\|^2)$) as a function of the averaged computational cost of the iterations of the four stochastic algorithms (resp. the Sinkhorn algorithm) for different values of the regularization parameter ε.



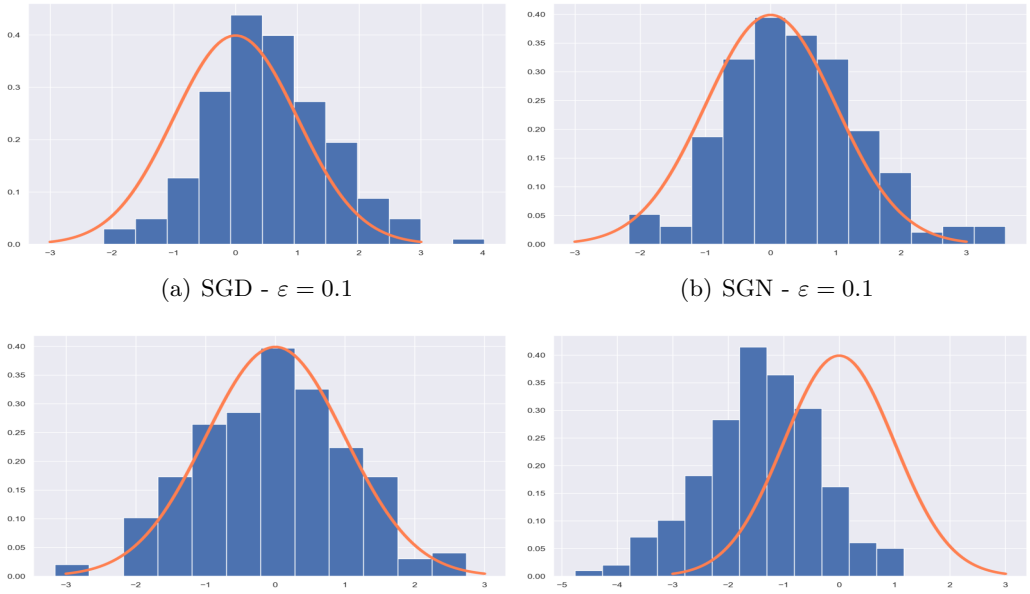
(a) $\varepsilon = 0.01$ (b) $\varepsilon = 0.005$

Figure 4: Discrete setting with $I = 10^4$ and $J = 400$ with $n = 10^5$ iterations. Expected excess risk (in logarithmic scale) $\log(\mathbb{E}[|\widehat{W}_n - W_\varepsilon(\mu,\nu)|])$ (resp. metric $\log(|W_k - W_\varepsilon(\mu,\nu)|)$) as a function of the averaged computational cost of the iterations of the four stochastic algorithms (resp. the Sinkhorn algorithm) for different values of the regularization parameter ε.



(a) $\varepsilon = 0.01$ (b) $\varepsilon = 0.005$

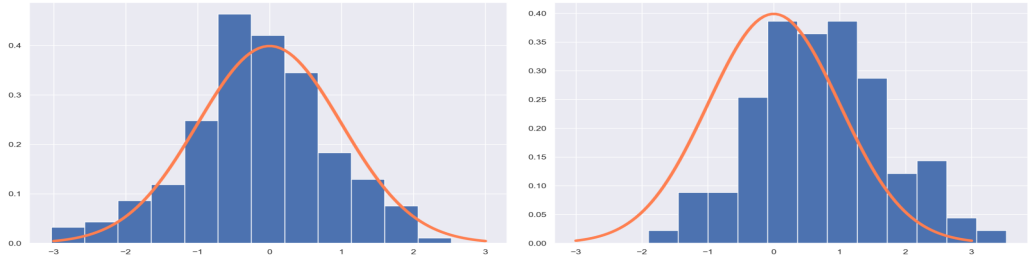
Figure 5: Discrete setting with $I = 10^4$ and $J = 400$ with $n = 10^5$ iterations. Expected excess risk (in logarithmic scale) $\log(\mathbb{E}[\|\widehat{V}_n - v^*\|^2])$ (resp. metric $\log(\|V_k - v^*\|^2)$) as a function of the averaged computational cost of the iterations of the four stochastic algorithms (resp. the Sinkhorn algorithm) for different values of the regularization parameter ε.



(a) SGD - $\varepsilon = 0.1$ (b) SGN - $\varepsilon = 0.1$

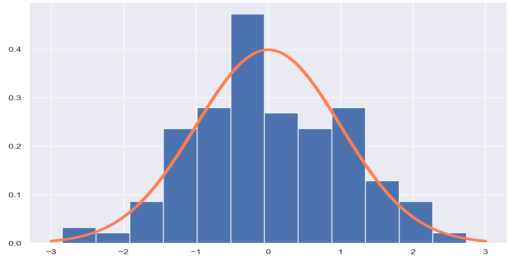
(c) SN - $\varepsilon = 0.1$ (d) ADAM - $\varepsilon = 0.1$

Figure 6: Discrete setting with $I = 10^3$, $J = 50$ and $\varepsilon = 0.1$. Histogram of 200 independent realizations of $\frac{\sqrt{n}(\widehat{W}_n - W_\varepsilon(\mu,\nu))}{\widehat{\sigma}_n}$ with $n = 2 \times 10^5$ using each of the four stochastic algorithms. The orange curve is the density of the standard Gaussian distribution.

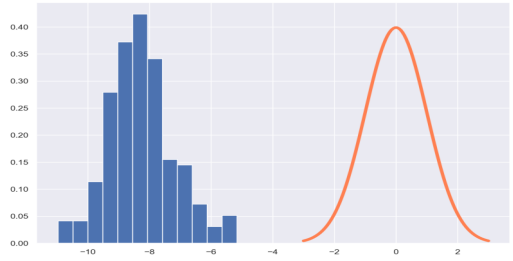


(a) SGD - $\varepsilon = 0.01$

(b) SGN - $\varepsilon = 0.01$



(c) SN - $\varepsilon = 0.01$



(d) ADAM - $\varepsilon = 0.01$

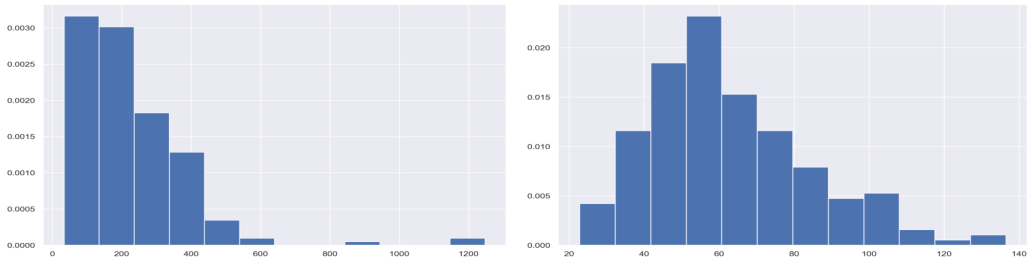
Figure 7: Discrete setting with $I = 10^3$, $J = 50$ and $\varepsilon = 0.01$. Histogram of 200 independent realizations of $\frac{\sqrt{n}\left(\widehat{W}_n - W_\varepsilon(\mu,\nu)\right)}{\widehat{\sigma}_n}$ with $n = 4 \times 10^5$ iterations using each of the four stochastic algorithms. The orange curve is the density of the standard Gaussian distribution.

Therefore, these numerical experiments clearly show that using the SGN algorithm has interesting benefits as it outperforms SGD and ADAM for the estimation of v^* since it yields an estimator $\widetilde{V}_n$ with a smaller variance.

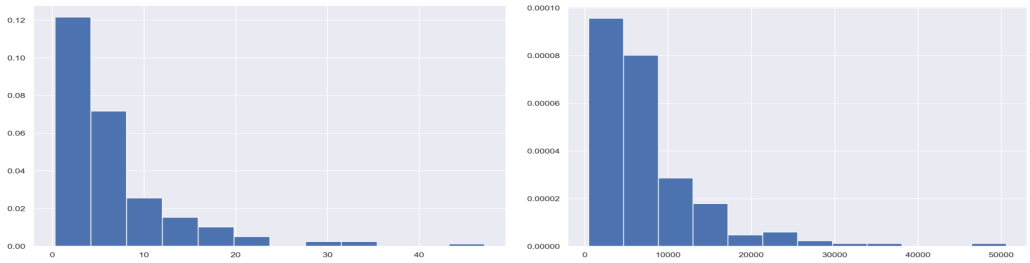
Convergence of the pre-conditionning matrix for the SN algorithm. Finally, we report in Figure 10 numerical results (with $I = 10^4$ and $J \in \{100, 200\}$) on the convergence of $\overline{S}_n$ to $G_\varepsilon(v^*)$ as a function the computational time of the SGN algorithm (using $n = 10^5$ iterations) for different values of the regularization parameter ε. We observe that the convergence becomes slower as ε decreases.

4.3.2 Semi-discrete setting

In this section, the cost function is chosen as the following normalized squared Euclidean distance $c(x, y) = \frac{1}{d}\|x - y\|^2$. We now consider the framework where μ is a mixture of three Gaussian densities in dimension d. In these numerical experiments, the size $J = 100$ of the support of ν is held fixed, and it is chosen as the uniform discrete probability measure supported on J points drawn uniformly on the hypercube $[0, 1]^d$. The value of the dimension d is let growing, and we analyze its influence on the performances of the stochastic algorithms with either $n = 5 \times 10^5$ or $n = 10^6$ iterations. We also study the performances of the Sinkhorn algorithm from a full-batch sample, that is using the

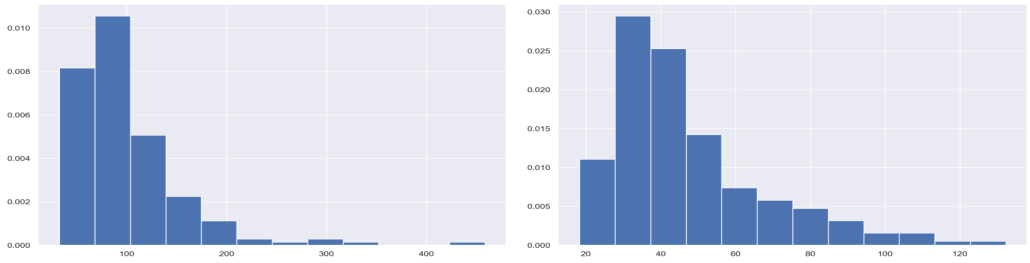


(a) SGD - $\varepsilon = 0.1$ (b) SGN - $\varepsilon = 0.1$

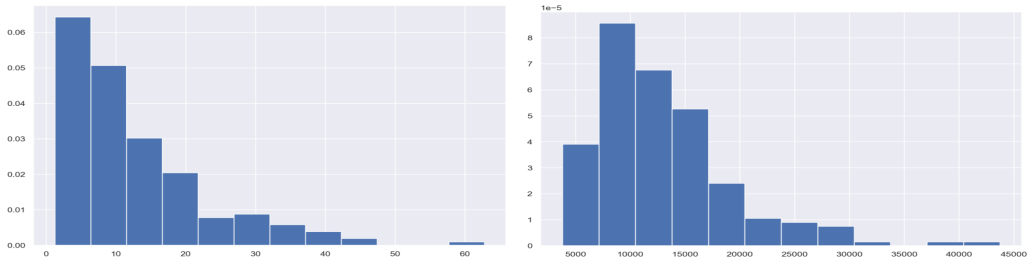


(c) SN - $\varepsilon = 0.1$ (d) ADAM - $\varepsilon = 0.1$

Figure 8: Discrete setting with $I = 10^3$, $J = 50$ and $\varepsilon = 0.1$. Histogram of 200 independent realizations of $n\left\|\widehat{V}_n - v^*\right\|^2$ with $n = 2 \times 10^5$ iterations using each of the four stochastic algorithms.

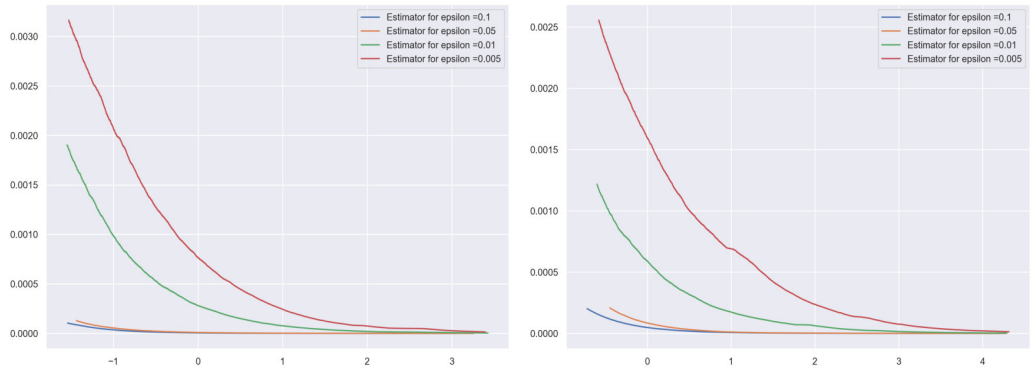


(a) SGD - $\varepsilon = 0.01$ (b) SGN - $\varepsilon = 0.01$



(c) SN - $\varepsilon = 0.01$ (d) ADAM - $\varepsilon = 0.01$

Figure 9: Discrete setting with $I = 10^3$, $J = 50$ and $\varepsilon = 0.01$. Histogram of 200 independent realizations of $n\left\|\widehat{V}_n - v^*\right\|^2$ with $n = 4 \times 10^5$ iterations using each of the four stochastic algorithms.



(a) $I = 10^4$ and $J = 100$ (b) $I = 10^4$ and $J = 200$

Figure 10: Discrete setting and convergence of $\log\left(\left\|\overline{S}_n - G_\varepsilon(v^*)\right\|_F^2\right)$ as a function the computational time of the SGN algorithm (using $n = 10^5$ iterations) for different values of the regularization parameter ε.

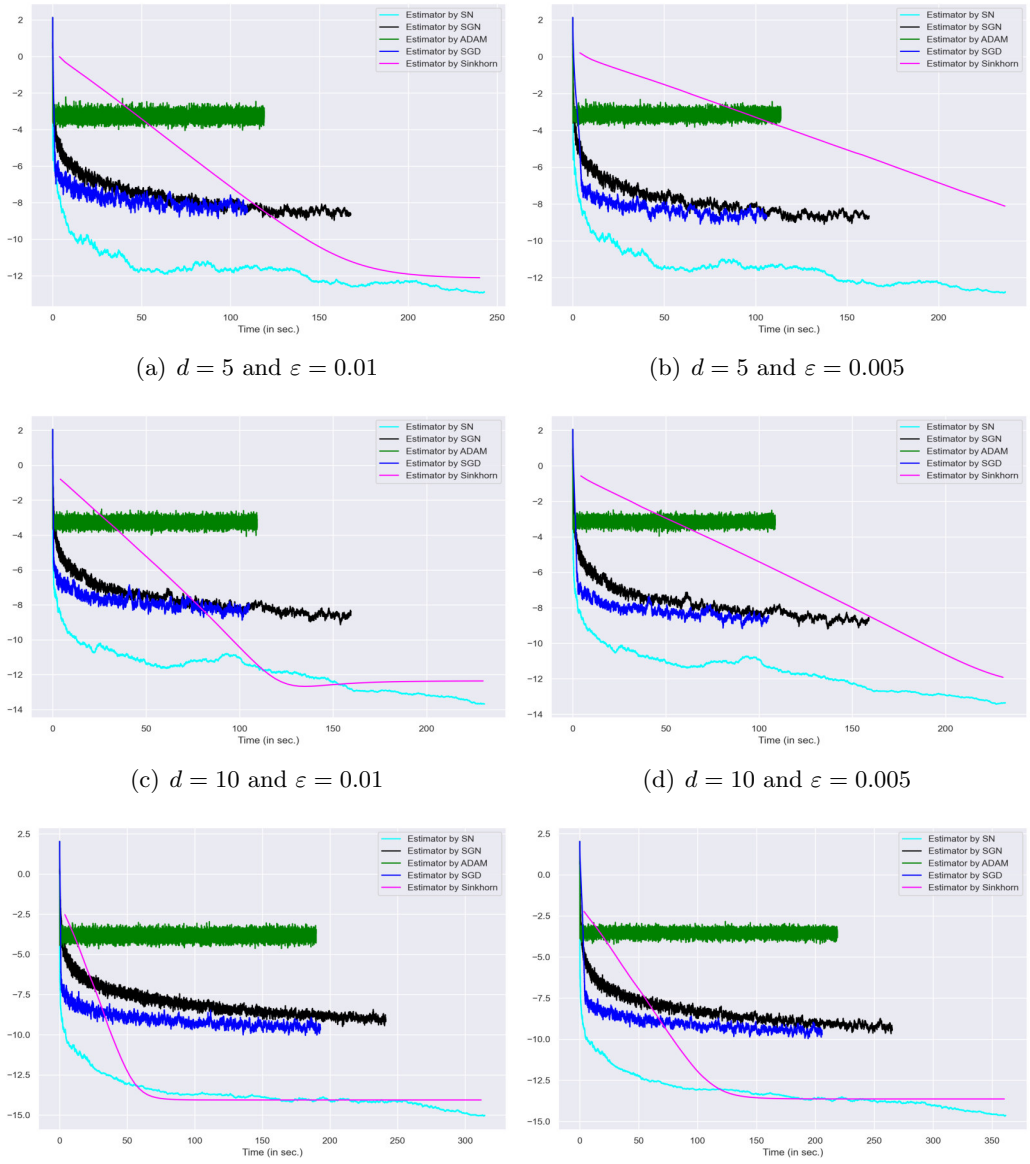
empirical measure

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad \text{where} \quad X_1, \ldots, X_n \sim_{iid} \mu,$$

to compute a solution of the regularized OT problem $W_\varepsilon(\hat{\mu}_n, \nu)$ as an approximation of $W_\varepsilon(\mu, \nu)$. At each iteration, the cost of this Sinkhorn algorithm is thus $\mathcal{O}(nJ)$ as n is the size of the support of $\hat{\mu}_n$. We have chosen to display results for only one simulation as averaging over Monte-Carlo replications does not change our main conclusions from these numerical experiments. All stochastic algorithms (resp. Sinkhorn algorithm) are compared for the metric $\left\|\widehat{V}_n - v^*\right\|^2$ (resp. $\left\|V_k - v^*\right\|^2$), where the vector v^* is preliminary approximated by running the SN algorithm with a large value of iterations $n_{\max} = 10^6$.

In Figure 11 (for $n = 5 \times 10^5$ iterations) and Figure 12 (for $n = n_{\max} = 10^6$ iterations) , we display these metrics (in logarithmic scale) as functions of the computational time of the stochastic and the Sinkhorn algorithms for different values of dimension $d \in \{5, 10, 50\}$ and $\varepsilon \in \{0.01, 0.005\}$. First, for either $n = 5 \times 10^5$ or $n = 10^6$, it can be observed that the SN algorithm has always the best performances, while ADAM has the worst ones. Moreover, the SGN algorithm has slightly better performances than SGD. In Figure 12, the very fast decay of the error after 450 seconds for SN is due to the fact that the ground truth value v^* has been preliminary computed with the SN algorithm with $n = n_{\max} = 10^6$ iterations. For either $n = 5 \times 10^5$ or $n = 10^6$, we also remark that the SGN outperforms Sinkhorn for $\varepsilon = 0.005$ and $d \in \{5, 10\}$ in the sense that when the SGN stops the value reached by $\left\|\widehat{V}_n - v^*\right\|^2$ is smaller than $\left\|V_k - v^*\right\|^2$ obtained with Sinkhorn. In larger dimension $d = 50$, the Sinkhorn algorithm appears to have better performances than SGN and SGD, but we recall that Sinkhorn uses the full sample at each iteration.

Therefore, these numerical experiments show that SGN has interesting benefits over Sinkhorn in small dimension d when combined with small values of the regularization parameter ε and large values of n. Moreover, we observe that the best results are always



(a) $d = 5$ and $\varepsilon = 0.01$

(b) $d = 5$ and $\varepsilon = 0.005$

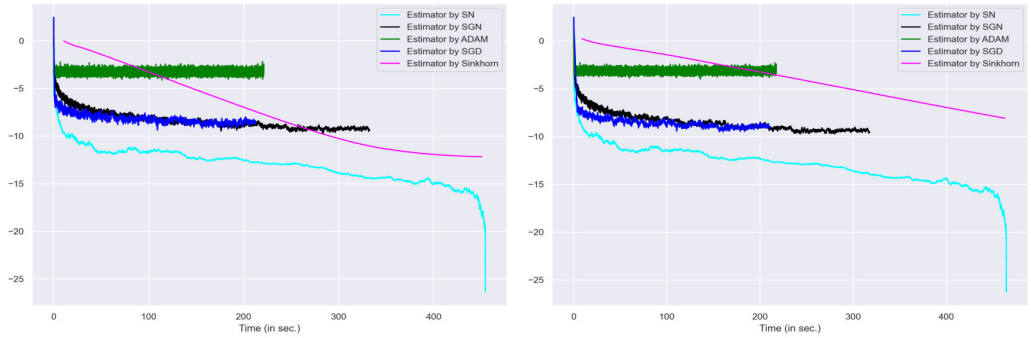
(c) $d = 10$ and $\varepsilon = 0.01$

(d) $d = 10$ and $\varepsilon = 0.005$

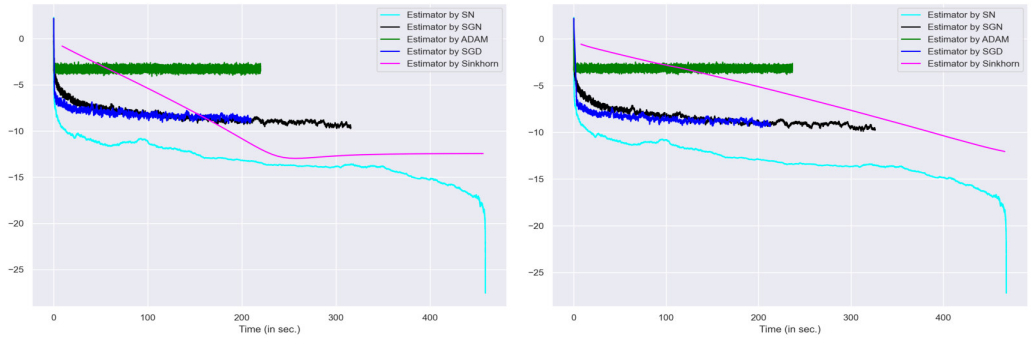
(e) $d = 50$ and $\varepsilon = 0.01$

(f) $d = 50$ and $\varepsilon = 0.005$

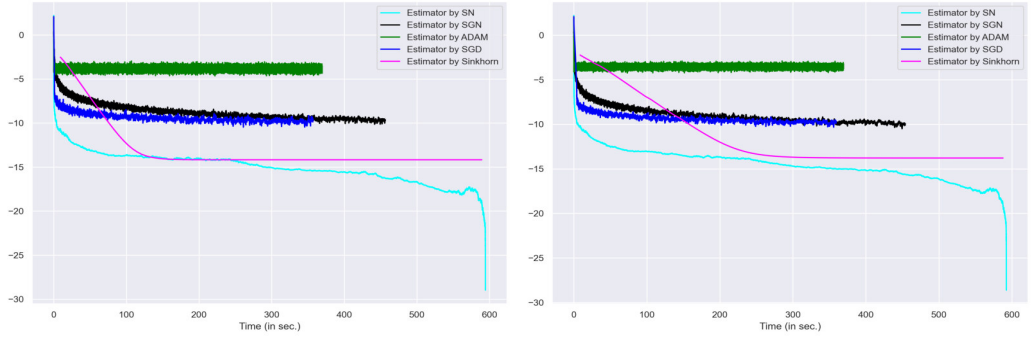
Figure 11: Semi-discrete setting where μ is a mixture of three Gaussian densities, $J = 100$ and $n = 5 \times 10^5$ iterations. Excess risk (in logarithmic scale) $\log(\left\|\widehat{V}_n - v^*\right\|^2)$ (resp. metric $\log(\left\|V_k - v^*\right\|^2)$) as a function of the computational cost of the iterations of the four stochastic algorithms (resp. the Sinkhorn algorithm) for different values of the dimension d and the regularization parameter ε.



(a) $d = 5$ and $\varepsilon = 0.01$ (b) $d = 5$ and $\varepsilon = 0.005$



(c) $d = 10$ and $\varepsilon = 0.01$ (d) $d = 10$ and $\varepsilon = 0.005$



(e) $d = 50$ and $\varepsilon = 0.01$ (f) $d = 50$ and $\varepsilon = 0.005$

Figure 12: Semi-discrete setting where μ is a mixture of three Gaussian densities, $J = 100$ and $n = n_{\max} = 10^6$ iterations. Excess risk (in logarithmic scale) $\log(\|\widehat{V}_n - v^*\|^2)$ (resp. metric $\log(\|V_k - v^*\|^2)$) as a function of the computational cost of the iterations of the four stochastic algorithms (resp. the Sinkhorn algorithm) for different values of the dimension d and the regularization parameter ε.

obtained with the SN algorithm.

5 Properties of the objective function H_ε

The purpose of this section is to discuss various keystone properties of the objective function H_ε that are needed to establish our main results.

5.1 Gradient properties

Let us first remark that, for any $x \in \mathcal{X}$, the function $v \mapsto h_\varepsilon(x, v)$, defined by (2.5), is twice differentiable. For a fixed $x \in \mathcal{X}$, the gradient vector and Hessian matrix of the function h_ε, with respect to its second argument, are given by

$$\nabla_v h_\varepsilon(x, v) = \pi(x, v) - \nu, \tag{5.1}$$

and

$$\nabla_v^2 h_\varepsilon(x, v) = \frac{1}{\varepsilon}\Big(\mathrm{diag}(\pi(x, v)) - \pi(x, v)\pi(x, v)^T\Big), \tag{5.2}$$

where the j^{th} component of the vector $\pi(x, v) \in \mathbb{R}^J$ is such that

$$\pi_j(x, v) = \Big(\sum_{k=1}^{J} \nu_k \exp\Big(\frac{v_k - c(x, y_k)}{\varepsilon}\Big)\Big)^{-1} \nu_j \exp\Big(\frac{v_j - c(x, y_j)}{\varepsilon}\Big).$$

Consequently, the gradient vector and the Hessian matrix of the function H_ε, defined by (2.3), are as follows

$$\nabla H_\varepsilon(v) = \mathbb{E}[\nabla_v h_\varepsilon(X, v)] = \mathbb{E}[\pi(X, v)] - \nu, \tag{5.3}$$

and

$$\nabla^2 H_\varepsilon(v) = \mathbb{E}[\nabla_v^2 h_\varepsilon(X, v)] = \frac{1}{\varepsilon}\mathbb{E}\big[\mathrm{diag}(\pi(X, v)) - \pi(X, v)\pi(X, v)^T\big]. \tag{5.4}$$

Note that the minimizer v^* satisfies $\nabla H_\varepsilon(v^*) = 0$, leading to

$$\mathbb{E}[\pi(X, v^*)] = \nu,$$

which allows us to simplify the expression for the Hessian of H_ε at v^*,

$$\nabla^2 H_\varepsilon(v^*) = \frac{1}{\varepsilon}\left(\mathrm{diag}(\nu) - \mathbb{E}\big[\pi(X, v^*)\pi(X, v^*)^T\big]\right). \tag{5.5}$$

We now discuss some properties of the above gradient vectors and Hessian matrices that will be of interest to study the SGN algorithm.

5.2 Convexity of H_ε and related properties

First of all, the baseline remark is that $\nabla_v^2 H_\varepsilon(v)$ is a positive semi-definite matrix for any $v \in \mathbb{R}^J$, which entails the convexity of H_ε.

Minimizers and rank of the Hessian. It is clear from (5.4) that for any $v \in \mathbb{R}^J$, the smallest eigenvalue of the Hessian matrix $\nabla_v^2 H_\varepsilon(v)$ associated to the eigenvector $\boldsymbol{v}_J$ is equal to zero. Therefore, as indicated in the end of Section 2.1, for any $t \in \mathbb{R}$, the vector $v^* + t\boldsymbol{v}_J$ is also a minimizer of (2.4). Nevertheless, it is well-known [15] that the minimizer v^* of (2.4) is unique up to a scalar translation of its coordinates. We shall thus denote by v^* the minimizer of (2.3) satisfying $\langle v^*, \boldsymbol{v}_J \rangle = 0$. It means that v^* belongs to $\langle \boldsymbol{v}_J \rangle^\perp$, and that the function H_ε admits a unique minimizer over the $J-1$ dimensional subspace $\langle \boldsymbol{v}_J \rangle^\perp$. However, as already shown in [23] and further discussed in [5, Section 3.3], the objective function H_ε is *not strongly convex*, even by restricting the maximization problem (2.3) to the subspace $\langle \boldsymbol{v}_J \rangle^\perp$ since it may be shown that $v \mapsto H_\varepsilon(v)$ may have some vanishing curvature, leading to a flat landscape, *i.e.* to eigenvalues of the Hessian matrix that are arbitrarily close to 0 for large values of $\|v\|$ in $\langle \boldsymbol{v}_J \rangle^\perp$.

Moreover, for any $(x, v) \in \mathcal{X} \times \mathbb{R}^J$, it follows from [5, Lemma A.1] that the matrices $\nabla_v^2 h_\varepsilon(x, v)$ and $\nabla_v^2 H_\varepsilon(v)$ are of rank $J-1$, and therefore, all their eigenvectors associated to non-zero eigenvalues belong to $\langle \boldsymbol{v}_J \rangle^\perp$. Finally, one also has that $\nabla_v h_\varepsilon(x, v) \in \langle \boldsymbol{v}_J \rangle^\perp$ for any $(x, v) \in \mathcal{X} \times \mathbb{R}^J$.

Useful upper and lower bounds. We conclude this section by stating a few inequalities that we repeatedly use in the proofs of our main results. Since ν and $\pi(x, v)$ are vectors with positive entries that sum up to one, it follows from (5.1) that for any $(x, v) \in \mathcal{X} \times \mathbb{R}^J$,

$$\|\nabla_v h_\varepsilon(x, v)\| \leqslant \|\nu\| + \|\pi(x, v)\| \leqslant 2, \tag{5.6}$$

and that the gradient of H_ε is always bounded for any $v \in \mathbb{R}^J$,

$$\|\nabla H_\varepsilon(v)\| \leqslant 2. \tag{5.7}$$

Moreover, thanks to the property that

$$\lambda_{\max}\Big(\frac{1}{\varepsilon}\big(\mathrm{diag}(\pi(x, v)) - \pi(x, v)\pi(x, v)^T\big)\Big) \leqslant \frac{1}{\varepsilon}\mathrm{Tr}\big(\mathrm{diag}(\pi(x, v))\big) = \frac{1}{\varepsilon},$$

we obtain that for any $(x, v) \in \mathcal{X} \times \mathbb{R}^J$,

$$\lambda_{\max}\left(\nabla_v^2 h_\varepsilon(x, v)\right) \leqslant \frac{1}{\varepsilon} \qquad \text{and} \qquad \lambda_{\max}\left(\nabla_v^2 H_\varepsilon(v)\right) \leqslant \frac{1}{\varepsilon}. \tag{5.8}$$

Finally, by [5, Lemma A.1], the second smallest eigenvalue of $\nabla^2 H_\varepsilon(v^*)$ is positive, and one has that

$$\lambda_{\min}\left(\nabla^2 H_\varepsilon(v^*)\right) = \min_{v \in \langle \boldsymbol{v}_J \rangle^\perp} \left\{ \frac{v^T \nabla^2 H_\varepsilon(v^*) v}{\|v\|^2} \right\} \geqslant \frac{1}{\varepsilon} \min(\nu). \tag{5.9}$$

5.3 Generalized self-concordance for regularized semi-discrete OT

Let us now introduce the so-called notion of generalized self-concordance proposed in Bach [2] for the purpose of obtaining fast rates of convergence for stochastic algorithms with non-strongly convex objective functions. Generalized self-concordance has been shown to hold for regularized semi-discrete OT in [5], and we discuss below its implications of some

key properties for the analysis of the SGN algorithm studied in this paper. To this end, for any $v \in \langle \boldsymbol{v}_J \rangle^\perp$ and for all t in the interval $[0,1]$, we denote $v_t = v^* + t(v - v^*)$, and we define the function φ, for all $t \in [0,1]$, as

$$\varphi(t) = H_\varepsilon(v_t).$$

The second-order Taylor expansion of φ with integral remainder is given by

$$\varphi(1) = \varphi(0) + \varphi'(0) - \int_0^1 (t-1)\varphi''(t)\, dt. \tag{5.10}$$

Using that $\varphi(1) = H_\varepsilon(v)$, $\varphi(0) = H_\varepsilon(v^*)$ and $\varphi'(0) = \langle v - v^*, \nabla H_\varepsilon(v^*) \rangle = 0$, it has been first remarked in [5] that inequality (5.8) implies that

$$H_\varepsilon(v) - H_\varepsilon(v^*) \leqslant \frac{1}{2\varepsilon} \|v - v^*\|^2 \tag{5.11}$$

Moreover, it is shown in the proof of [5, Lemma A.2] that the following inequality holds

$$\left|\varphi'''(t)\right| \leqslant \frac{\sqrt{2}}{\varepsilon} \varphi''(t) \|v - v^*\|. \tag{5.12}$$

It means that the function φ satisfies the so-called generalized self-concordance property with constant $s_\varepsilon = \sqrt{2}/\varepsilon$ as defined in Appendix B of [2]. As a consequence of inequality (5.12) and thanks to the arguments in the proof of [5, Lemma A.2], the error of linearizing the gradient $\nabla H_\varepsilon(v) \approx \nabla^2 H_\varepsilon(v^*)(v - v^*)$ is controlled as follows,

$$\|\nabla H_\varepsilon(v) - \nabla^2 H_\varepsilon(v^*)(v - v^*)\| \leqslant 2 s_\varepsilon \|v - v^*\|^2. \tag{5.13}$$

Moreover, generalized self-concordance also implies the following result (which is a consequence of the arguments in the proof of Lemma A.2 in [5]) that may be interpreted as a local strong convexity property of the function H_ε in the neighborhood of v^*.

Lemma 5.1. *For any $v \in \langle \boldsymbol{v}_J \rangle^\perp$, we have*

$$\langle \nabla H_\varepsilon(v), v - v^* \rangle \geqslant \frac{1 - \exp(-\delta(v))}{\delta(v)} (v - v^*)^T \nabla^2 H_\varepsilon(v^*)(v - v^*), \tag{5.14}$$

where $\delta(v) = s_\varepsilon \|v - v^*\|$.

Finally, if we now consider the matrix-valued function $G_\varepsilon(v)$ introduced in equation (3.1), we have the following result which can be interpreted as a local Lipschitz property of $G_\varepsilon(v)$ around $v = v^*$. The proof of this lemma is postponed to Appendix A.

Lemma 5.2. *For any $v \in \langle \boldsymbol{v}_J \rangle^\perp$, we have that*

$$-\frac{4}{\varepsilon} \|v - v^*\| I_J \leqslant G_\varepsilon(v) - G_\varepsilon(v^*) \leqslant \frac{4}{\varepsilon} \|v - v^*\| I_J, \tag{5.15}$$

in the sense of partial ordering between positive semi-definite matrices.

6 Proofs of the main results

This section contains the proofs of our main results that are stated in Section 3. Our results are based on previous important contributions on self-concordance functions (see *e.g.* [2, 5]), regularization of second order algorithms [6] and on the Kurdyka-Łojasiewicz inequality adapted to stochastic algorithms [22]. More specifically, almost sure convergence and almost sure convergence rates crucially depend on the adaptive property (3.4), which induces a contraction rate of the sequence $G_\varepsilon^{1/2}(v^*)\big(\widehat{V}_n - v^*\big)$ (see Equation (6.11) below). The non-asymptotic study is then based on both the self-concordance property, stated in Lemma 5.1 and on the KL inequality stated in Proposition 6.1. The combination of these two properties is an essential novelty brought by our work in order to build a key Lyapunov function in Equation (6.45). We emphasize that to obtain the results stated below, we have derived quantitative computations that are specific to the regularized OT problem. In particular, if the use of the KL inequality is borrowed from [22], the exact values of the constant m_ϵ and M used in Proposition 6.1 crucially depend on the self-concordance property of the regularized OT problem.

6.1 Keystone property

We start with the proof of inequality (3.4) that states the adaptivity of the SGN algorithm to the local geometry of H_ε.

Proof of Proposition 3.1. First of all, one can remark that $G_\varepsilon(v^*)$ is a positive semi-definite matrix whose smallest eigenvalue is equal to zero and associated to the eigenvector $\boldsymbol{v}_J$. Thus, all the eigenvectors of $G_\varepsilon(v^*)$ associated to non-zero eigenvalues belong to $\langle \boldsymbol{v}_J \rangle^\perp$. We already saw from (5.5) that

$$\nabla^2 H_\varepsilon(v^*) = \frac{1}{\varepsilon}\left(\mathrm{diag}(\nu) - \mathbb{E}\big[\pi(X, v^*)\pi(X, v^*)^T\big]\right).$$

Consequently, it follows from (3.1) and (5.1) that

$$G_\varepsilon(v^*) = \mathbb{E}\big[\pi(X, v^*)\pi(X, v^*)^T\big] - \nu\nu^T = \mathrm{diag}(\nu) - \nu\nu^T - \varepsilon\nabla^2 H_\varepsilon(v^*),$$

which implies that

$$G_\varepsilon(v^*) = \nabla^2 H_\varepsilon(v^*) + \Sigma_\varepsilon^*$$

where

$$\Sigma_\varepsilon^* = \mathrm{diag}(\nu) - \nu\nu^T - (1+\varepsilon)\nabla^2 H_\varepsilon(v^*).$$

On the one hand, it is easy to see that $\boldsymbol{v}_J^T \Sigma_\varepsilon^* \boldsymbol{v}_J = 0$. On the other hand, we deduce from inequality (5.9) that for all $v \in \langle \boldsymbol{v}_J \rangle^\perp$,

$$v^T \Sigma_\varepsilon^* v \leqslant \max(\nu) - \left(\frac{1+\varepsilon}{\varepsilon}\right)\min(\nu).$$

Finally, condition (3.2) on the regularization parameter ε leads to $G_\varepsilon(v^*) \leqslant \nabla^2 H_\varepsilon(v^*)$, which completes the proof of Proposition 3.1. □

6.2 Proofs of the most sure convergence results

Proof of Theorem 3.1. In what follows, we borrow some arguments from the proof of [6, Theorem 4.1] to establish the almost sure convergence of the regularized versions of the SGN algorithm as an application of the Robbins-Siegmund Theorem [41].

• We already saw that for all $n \geqslant 0$, $\widehat{V}_n$ belongs to $\langle \boldsymbol{v}_J \rangle^{\perp}$. We clearly have from (5.1) that for all $n \geqslant 0$, $\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)$ also belong to $\langle \boldsymbol{v}_J \rangle^{\perp}$. Hence, we have from (2.6) that for all $n \geqslant 0$,

$$\widehat{V}_{n+1} = \widehat{V}_n - n^\alpha P_J S_n^{-1} P_J \big(\nabla H_\varepsilon(\widehat{V}_n) + \varepsilon_{n+1} \big), \tag{6.1}$$

where the martingale increment ε_{n+1} is given by

$$\varepsilon_{n+1} = \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) - \mathbb{E}[\nabla(h_\varepsilon(X_{n+1}, \widehat{V}_n))|\mathcal{F}_n] = \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) - \nabla H_\varepsilon(\widehat{V}_n)$$

with $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Moreover, it follows from the Taylor-Lagrange formula that

$$H_\varepsilon(\widehat{V}_{n+1}) = H_\varepsilon(\widehat{V}_n) + \nabla H_\varepsilon(\widehat{V}_n)^T (\widehat{V}_{n+1} - \widehat{V}_n) + \frac{1}{2} (\widehat{V}_{n+1} - \widehat{V}_n)^T \nabla^2 H_\varepsilon(\xi_{n+1}) (\widehat{V}_{n+1} - \widehat{V}_n), \tag{6.2}$$

where $\xi_{n+1} = \widehat{V}_n + t(\widehat{V}_{n+1} - \widehat{V}_n)$ with $t \in]0, 1[$. Consequently, we deduce from (6.1) and (6.2) that for all $n \geqslant 0$,

$$\begin{aligned} H_\varepsilon(\widehat{V}_{n+1}) &= H_\varepsilon(\widehat{V}_n) - n^\alpha \big\langle \nabla H_\varepsilon(\widehat{V}_n), P_J S_n^{-1} P_J \big(\nabla H_\varepsilon(\widehat{V}_n) + \varepsilon_{n+1} \big) \big\rangle \\ &\quad + \frac{n^{2\alpha}}{2} \big(P_J S_n^{-1} P_J \big(\nabla H_\varepsilon(\widehat{V}_n) + \varepsilon_{n+1} \big) \big)^T \nabla^2 H_\varepsilon(\xi_{n+1}) \big(P_J S_n^{-1} P_J \big(\nabla H_\varepsilon(\widehat{V}_n) + \varepsilon_{n+1} \big) \big). \end{aligned}$$

Taking the conditional expectation with respect to $\mathcal{F}_n$ on both sides of the previous equality, we obtain that for all $n \geqslant 0$,

$$\begin{aligned} \mathbb{E}[H_\varepsilon(\widehat{V}_{n+1})|\mathcal{F}_n] &= H_\varepsilon(\widehat{V}_n) - n^\alpha \nabla H_\varepsilon(\widehat{V}_n)^T P_J S_n^{-1} P_J \nabla H_\varepsilon(\widehat{V}_n) \\ &\quad + \frac{n^{2\alpha}}{2} \mathbb{E}\Big[\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)^T P_J S_n^{-1} P_J \nabla^2 H_\varepsilon(\xi_{n+1}) P_J S_n^{-1} P_J \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) \big| \mathcal{F}_n \Big] \\ &= H_\varepsilon(\widehat{V}_n) - n^\alpha \nabla H_\varepsilon(\widehat{V}_n)^T P_J S_n^{-1} P_J \nabla H_\varepsilon(\widehat{V}_n) \\ &\quad + \frac{n^{2\alpha}}{2} \mathbb{E}\Big[\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)^T S_n^{-1} \nabla^2 H_\varepsilon(\xi_{n+1}) S_n^{-1} \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) \big| \mathcal{F}_n \Big] \end{aligned} \tag{6.3}$$

using the elementary fact that $P_J \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) = \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)$ as well as $\nabla^2 H_\varepsilon(\xi_{n+1}) v_J = 0$ which implies that $P_J \nabla^2 H_\varepsilon(\xi_{n+1}) P_J = \nabla^2 H_\varepsilon(\xi_{n+1})$. On the one hand, we have from inequality (5.6) that $\|\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)\| \leqslant 2$. On the other hand, it follows from inequality (5.8) that $\lambda_{\max}\big(\nabla^2 H_\varepsilon(\xi_{n+1})\big) \leqslant 1/\varepsilon$. Therefore, we deduce from (6.3) that for all $n \geqslant 0$,

$$\mathbb{E}[H_\varepsilon(\widehat{V}_{n+1})|\mathcal{F}_n] \leqslant H_\varepsilon(\widehat{V}_n) + A_n - B_n \qquad \text{a.s.}$$

where the two positive random variables A_n and B_n are given by

$$A_n = \frac{2n^{2\alpha}}{\varepsilon \big(\lambda_{\min}\big(S_n\big)\big)^2}$$

and $B_n = n^\alpha \nabla H_\varepsilon(\widehat{V}_n)^T P_J S_n^{-1} P_J \nabla H_\varepsilon(\widehat{V}_n)$. Our purpose is now to show that

$$\sum_{n=1}^{\infty} A_n < +\infty \qquad \text{a.s.}$$

We already saw from (4.3) that for all $n \geqslant 1$,

$$S_n = I_J + \sum_{k=1}^{n} \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1}) \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1})^T + R_n$$

where

$$R_n = \sum_{k=1}^{n} \gamma \Big(1 + \lfloor \frac{k}{J} \rfloor\Big)^{-\beta} Z_k Z_k^T.$$

We clearly have $\lambda_{\min}(S_n) \geqslant \lambda_{\min}(R_n)$. Let p_n be the largest integer such that $p_n J \leqslant n$. One can remark that

$$R_n = \Big(\sum_{m=1}^{p_n} m^{-\beta}\Big) \gamma \, \mathrm{diag}(\nu) + \sum_{k=p_n+1}^{n} \gamma \Big(1 + \lfloor \frac{k}{J} \rfloor\Big)^{-\beta} Z_k Z_k^T, \tag{6.4}$$

which implies that

$$\lambda_{\min}(S_n) \geqslant \gamma \min(\nu) \Big(\sum_{m=1}^{p_n} m^{-\beta}\Big).$$

However, for any $0 < \beta < 1/2$ and for all $p_n \geqslant 4$,

$$\sum_{m=1}^{p_n} \frac{1}{m^\beta} \geqslant \frac{p_n^{1-\beta}}{2(1-\beta)}.$$

Consequently, using that $p_n \geqslant nJ^{-1} - 1$, we obtain that

$$\sum_{n=1}^{\infty} A_n \leqslant \frac{8(1-\beta)^2}{\varepsilon(\gamma \min(\nu))^2} \sum_{n=1}^{\infty} \frac{n^{2\alpha}}{p_n^{2(1-\beta)}} \leqslant \frac{16(1-\beta)^2 J^{2(1-\beta)}}{\varepsilon(\gamma \min(\nu))^2} \sum_{n=1}^{\infty} \frac{1}{n^{2(1-\alpha-\beta)}} < +\infty \qquad \text{a.s.}$$

since the assumption $0 < \alpha + \beta < 1/2$ implies that $2(1-\alpha-\beta) > 1$. Therefore, we can apply the Robbins-Siegmund Theorem [41] to conclude that the sequence $(H_\varepsilon(\widehat{V}_n))$ converges almost surely to a finite random variable and that the series

$$\sum_{n=1}^{\infty} B_n < +\infty \qquad \text{a.s.}$$

leading to

$$\sum_{n=1}^{\infty} n^\alpha \frac{\|\nabla H_\varepsilon(\widehat{V}_n)\|^2}{\lambda_{\max}(S_n)} < +\infty \qquad \text{a.s.} \tag{6.5}$$

One can very from inequality (5.6) and (4.3) that for all $n \geqslant 1$,

$$\lambda_{\max}(S_n) \leqslant 1 + 4n + \gamma \max(\nu) \sum_{k=1}^{n} \Big(1 + \lfloor \frac{k}{J} \rfloor\Big)^{-\beta} \leqslant 1 + (4 + \gamma \max(\nu))n.$$

Since $\alpha \geqslant 0$, it implies that

$$\sum_{n=1}^{\infty} \frac{n^{\alpha}}{\lambda_{\max}(S_n)} = +\infty \qquad \text{a.s.} \tag{6.6}$$

The rest of the proof proceeds from standard arguments combining (6.5) and (6.6). Let

$$H_{\varepsilon,\infty} = \lim_{n \to +\infty} H_{\varepsilon}(\widehat{V}_n) \qquad \text{a.s.}$$

and assume by contradiction that $H_{\varepsilon,\infty} > H_{\varepsilon}(v^*)$ where

$$H_{\varepsilon}(v^*) = \min_{v \in \langle \boldsymbol{v}_J \rangle^{\perp}} H_{\varepsilon}(v).$$

Since H_{ε} is a convex function with a unique minimizer v^* on $\langle \boldsymbol{v}_J \rangle^{\perp}$, we necessarily have

$$\lim_{\|v\| \to +\infty} H_{\varepsilon}(v) = +\infty.$$

It means that $(\widehat{V}_n)$ is almost surely bounded since $H_{\varepsilon,\infty}$ is finite. Therefore, we can find a compact set K such that $v^* = \arg\min_{v \in \langle \boldsymbol{v}_J \rangle^{\perp}} H_{\varepsilon}(v) \notin K$ and $\widehat{V}_n \in K$ for all n large enough. Using the continuity of $\|\nabla H_{\varepsilon}\|$ and the compactness of K, we conclude that $\|\nabla H_{\varepsilon}\|$ attains its lower bound, which is strictly positive on K. It ensures the existence of a constant $c > 0$, such that, for all n large enough,

$$\|\nabla H_{\varepsilon}(\widehat{V}_n)\| \geq c > 0.$$

The above lower bound associated with (6.5) and (6.6) yields a contradiction. Hence, we can conclude that

$$\lim_{n \to +\infty} \|\nabla H_{\varepsilon}(\widehat{V}_n)\| = 0 \qquad \text{a.s.}$$

It clearly implies that equation (3.5) holds true since $(\widehat{V}_n)$ is a bounded sequence with a unique adherence point v^*.

• It now remains to investigate the almost sure convergence of the matrix $\overline{S}_n$. We observe from equation (4.3) that S_n can be splitted into two terms,

$$S_n = M_n + \Sigma_n \tag{6.7}$$

with

$$M_n = \sum_{k=1}^{n} \Phi_k \Phi_k^T - G_{\varepsilon}(\widehat{V}_{k-1}) \qquad \text{and} \qquad \Sigma_n = I_J + \sum_{k=1}^{n} G_{\varepsilon}(\widehat{V}_{k-1}) + R_n,$$

where the vector Φ_k stands for $\Phi_k = \nabla_v h_{\varepsilon}(X_k, \widehat{V}_{k-1})$. Using the assumption $0 < \beta < 1/2$, we have from (6.4) that

$$\lim_{n \to +\infty} \frac{1}{n} \big(I_J + R_n \big) = 0.$$

Moreover, it follows from (5.3) that G_{ε} is a continuous function from $\mathbb{R}^J$ to $\mathbb{R}^{J \times J}$. Consequently, we deduce from convergence (3.5) together with the Cesaro mean convergence theorem that

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{k=1}^{n} G_{\varepsilon}(\widehat{V}_{k-1}) = G_{\varepsilon}(v^*) \qquad \text{a.s.}$$

which implies that

$$\lim_{n\to+\infty} \frac{1}{n}\Sigma_n = G_\varepsilon(v^*) \qquad \text{a.s.} \tag{6.8}$$

Hereafter, we focus our attention on the first term M_n in the right-hand side of (6.7). For any $u \in \mathbb{R}^J$, let

$$M_n(u) = u^T M_n u = \sum_{k=1}^{n} \xi_k(u)$$

where, for all $n \geqslant 1$, $\xi_n(u) = u^T\big(\Phi_n\Phi_n^T - G_\varepsilon(\widehat{V}_{n-1})\big)u$. It follows from (3.1) that for all $n \geqslant 1$, $\mathbb{E}[\Phi_{n+1}\Phi_{n+1}^T|\mathcal{F}_n] = G_\varepsilon(\widehat{V}_n)$. Hence, for all $n \geqslant 1$, $\mathbb{E}[\xi_{n+1}(u)|\mathcal{F}_n] = 0$. Furthermore, we obtain from (5.6) and (3.1) that for all $n \geqslant 1$, $\mathbb{E}[\xi_{n+1}^2(u)|\mathcal{F}_n] \leqslant 16||u||^2$. Consequently, $(M_n(u))$ is a locally square-integrable martingale with predictable quadratic variation satisfying

$$\langle M(u)\rangle_n = \sum_{k=1}^{n} \mathbb{E}[\xi_k^2(u)|\mathcal{F}_{k-1}] \leqslant 16n||u||^4.$$

We deduce from the strong law of large numbers for martingales given (e.g. by Theorem 1.3.24 in [18]) that

$$\lim_{n\to+\infty} \frac{1}{n}M_n(u) = 0 \qquad \text{a.s.}$$

which may be translated immediately into the matricial form

$$\lim_{n\to+\infty} \frac{1}{n}M_n = 0 \qquad \text{a.s.} \tag{6.9}$$

Finally, the convergence (3.6) follows from the decomposition (6.7) together with (6.8) and (6.9), which completes the proof of Theorem 3.1. □

It is straightforward to obtain the almost sure convergence of $\widehat{W}_n$ as follows.

Proof of Corollary 3.1. By Theorem 3.1 one has that $\widehat{V}_n$ converges a.s. to v^* under the assumption that $\alpha+\beta < 1/2$. Then, the almost sure convergence of $\widehat{W}_n$ to $W_\varepsilon(\mu,\nu)$ follows from assumption (3.7) and the arguments in the proof of [5, Theorem 3.5]. □

6.3 Proofs of the almost sure rates of convergence

We now establish the almost sure rates of convergence rates for the SGN algorithm. In contrast with the previous results, we emphasize that the regularization parameter ε must now be small enough, in the sense of condition (3.2). This entails the key inequality (3.4) deduced from Proposition 3.1.

Proof of Theorem 3.2. To alleviate the notation, we denote by $\|A\|$ either the operator norm $\|A\|_2$ or the Frobenius norm $\|A\|_F$ all along the proof. Since these two norms are equivalent and verify that $\|.\|_2^2 \leqslant \|.\|_F^2 \leqslant J\|.\|_2^2$, the upper bounds derived below might hold up to multiplicative constant depending on J, which will not affect the results that are purely asymptotic.

• Our starting point when $\alpha = 0$ is equation (2.6) written with $S_n = n\overline{S}_n$. We recall that the martingale increment is $\varepsilon_{n+1} = \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) - \nabla H_\varepsilon(\widehat{V}_n)$, so that for all $n \geqslant 0$,

$$\begin{aligned}\widehat{V}_{n+1} - v^* &= \widehat{V}_n - v^* - P_J S_n^{-1} P_J\big(\nabla H_\varepsilon(\widehat{V}_n) + \varepsilon_{n+1}\big) \\ &= \widehat{V}_n - v^* - \frac{1}{n}\big(P_J\big(\overline{S}_n^{-1} - G_\varepsilon^-(v^*)\big)P_J\big)\big(\nabla H_\varepsilon(\widehat{V}_n) + \varepsilon_{n+1}\big) \\ &\quad - \frac{1}{n}G_\varepsilon^-(v^*)\big(\nabla H_\varepsilon(\widehat{V}_n) + \varepsilon_{n+1}\big)\end{aligned} \tag{6.10}$$

where we decomposed $S_n^{-1} = n^{-1}\overline{S}_n^{\,-1} = n^{-1}G_\varepsilon^-(v^*) + n^{-1}(\overline{S}_n^{-1} - G_\varepsilon^-(v^*))$ and $P_J G_\varepsilon(v^*) = G_\varepsilon(v^*)$ which implies that $P_J G_\varepsilon^-(v^*) P_J = G_\varepsilon^-(v^*)$. The rest of the proof then consists in a linearization of $\nabla H_\varepsilon(\widehat{V}_n)$ around v^*. For that purpose, denote

$$D_n = \big(P_J\big(\overline{S}_n^{-1} - G_\varepsilon^-(v^*)\big)P_J\big) \qquad \text{and} \qquad \delta_n = \nabla H_\varepsilon(\widehat{V}_n) - \nabla^2 H_\varepsilon(v^*)(\widehat{V}_n - v^*).$$

We obtain from (6.10) that for all $n \geqslant 0$,

$$\begin{aligned}\widehat{V}_{n+1} - v^* &= \Big(P_J - \frac{1}{n}G_\varepsilon^-(v^*)\nabla^2 H_\varepsilon(v^*)\Big)\big(\widehat{V}_n - v^*\big) - \frac{1}{n}P_J\overline{S}_n^{-1}P_J\varepsilon_{n+1} \\ &\quad - \frac{1}{n}P_J\overline{S}_n^{-1}P_J\delta_n - \frac{1}{n}D_n\nabla^2 H_\varepsilon(v^*)\big(\widehat{V}_n - v^*\big).\end{aligned}$$

Hence, by setting $\widehat{U}_n = G_\varepsilon^{1/2}(v^*)\big(\widehat{V}_n - v^*\big)$, we obtain that for all $n \geqslant 0$,

$$\widehat{U}_{n+1} = \Big(P_J - \frac{1}{n}\Gamma_\varepsilon(v^*)\Big)\widehat{U}_n - \frac{1}{n}A_n\varepsilon_{n+1} - \frac{1}{n}T_n \tag{6.11}$$

where $\Gamma_\varepsilon(v^*) = G_\varepsilon^{-1/2}(v^*)\nabla^2 H_\varepsilon(v^*)G_\varepsilon^{-1/2}(v^*)$ and $T_n = A_n\delta_n + B_n\big(\widehat{V}_n - v^*\big)$ with

$$A_n = G_\varepsilon^{1/2}(v^*)P_J\overline{S}_n^{-1}P_J, \tag{6.12}$$

$$B_n = G_\varepsilon^{1/2}(v^*)D_n\nabla^2 H_\varepsilon(v^*). \tag{6.13}$$

Thanks to inequality (3.4), we have that $\lambda_{\min}^{\langle \boldsymbol{v}_J\rangle^\perp}(\Gamma_\varepsilon(v^*)) \geqslant 1$. For all $0 \leqslant k \leqslant n$, let

$$P_k^n = \prod_{i=k+1}^{n}\Big(P_J - \frac{1}{i}\Gamma_\varepsilon(v^*)\Big) \tag{6.14}$$

with the usual convention that $P_n^n = P_J$. We deduce from (6.11) that for all $n \geqslant 0$,

$$\widehat{U}_{n+1} = P_0^n\widehat{U}_1 - \sum_{k=1}^{n}\frac{1}{k}P_k^n A_k\varepsilon_{k+1} - \sum_{k=1}^{n}\frac{1}{k}P_k^n T_k. \tag{6.15}$$

The first term of (6.15) is easy to handle. If ρ stands for to the minimal eigenvalue of $\Gamma_\varepsilon(v^*)$ when restricted to act on the subspace $\langle \boldsymbol{v}_J\rangle^\perp$, a simple diagonalization of the matrix $\Gamma_\varepsilon(v^*)$ leads, for all $0 \leqslant k \leqslant n$, to

$$\|P_k^n\| \leqslant \kappa\left(\frac{k}{n}\right)^\rho \tag{6.16}$$

where $\kappa > 0$. Concerning the middle term of (6.15), let (M_n) be the multidimensional martingale defined by $M_1 = 0$ and, for all $n \geqslant 1$,

$$M_{n+1} = \sum_{k=1}^{n} A_k \varepsilon_{k+1}.$$

We infer from (5.3) and (3.1) that $\|\varepsilon_{n+1}\| \leqslant 4$, $\mathbb{E}[\varepsilon_{n+1}|\mathcal{F}_n] = 0$ and

$$\mathbb{E}[\varepsilon_{n+1}\varepsilon_{n+1}^T|\mathcal{F}_n] = G_\varepsilon(\widehat{V}_n) - \nabla H_\varepsilon(\widehat{V}_n)\nabla H_\varepsilon(\widehat{V}_n)^T.$$

Moreover, it follows from (3.5)

$$\lim_{n\to+\infty} \nabla H_\varepsilon(\widehat{V}_n) = \nabla H_\varepsilon(v^*) = 0 \qquad \text{and} \qquad \lim_{n\to+\infty} G_\varepsilon(\widehat{V}_n) = G_\varepsilon(v^*) \qquad \text{a.s.}$$

which ensures via (3.6) and (6.12) that

$$\lim_{n\to+\infty} A_n \mathbb{E}[\varepsilon_{n+1}\varepsilon_{n+1}^T|\mathcal{F}_n]A_n^T = P_J \qquad \text{a.s.}$$

Consequently, we have from the Cesaro mean convergence theorem that the predictable quadratic variation of the multidimensional martingale (M_n) satisfies

$$\lim_{n\to+\infty} \frac{1}{n}\langle M\rangle_n = \lim_{n\to+\infty} \frac{1}{n}\sum_{k=2}^{n} A_{k-1}\mathbb{E}[\varepsilon_k\varepsilon_k^T|\mathcal{F}_{k-1}]A_{k-1} = P_J \qquad \text{a.s.}$$

Hence, we deduce from the strong law of large numbers for multidimensional martingales given by Theorem 4.3.16 in [18] that

$$\|M_n\|^2 = \mathcal{O}(n\log n) \qquad \text{a.s.}$$

Therefore, there exists a finite positive random variable C such that for all $n \geqslant 1$,

$$\|M_{n+1}\| \leqslant C\sqrt{n\log n} \qquad \text{a.s.} \tag{6.17}$$

Hereafter, denote by Q_{n+1} the middle term of (6.15). We obtain from a simple Abel transform that

$$\begin{aligned} Q_{n+1} &= \sum_{k=1}^{n} \frac{1}{k}P_k^n(M_{k+1} - M_k) = \frac{1}{n}M_{n+1} + \sum_{k=1}^{n-1}\frac{1}{k}P_k^n M_{k+1} - \sum_{k=2}^{n}\frac{1}{k}P_k^n M_k \\ &= \frac{1}{n}M_{n+1} + \sum_{k=1}^{n-1}\Big(\frac{1}{k}P_k^n - \frac{1}{k+1}P_{k+1}^n\Big)M_{k+1} \\ &= \frac{1}{n}M_{n+1} + \sum_{k=1}^{n-1}\frac{1}{k(k+1)}\big(P_J - \Gamma_\varepsilon(v^*)\big)P_{k+1}^n M_{k+1} \end{aligned} \tag{6.18}$$

It follows from (6.16), (6.17), (6.18) that for all $n \geqslant 1$,

$$\begin{aligned} \|Q_{n+1}\| &\leqslant C\Big(\frac{\sqrt{n\log n}}{n} + \frac{\kappa}{n^\rho}\sum_{k=1}^{n-1}\frac{(k+1)^\rho}{k(k+1)}\sqrt{k\log k}\Big) \qquad \text{a.s.} \\ &\leqslant C\Big(\Big(\frac{\log n}{n}\Big)^{1/2} + \frac{\kappa\sqrt{\log n}}{n^\rho}\sum_{k=1}^{n}\frac{1}{k^a}\Big) \qquad \text{a.s.} \end{aligned}$$

where $a = 3/2 - \rho < 1$. Consequently, we deduce that for all $n \geqslant 1$,

$$\|Q_{n+1}\| \leqslant C\Big(\Big(\frac{\log n}{n}\Big)^{1/2} + \frac{\kappa n^{1-a}\sqrt{\log n}}{(1-a)n^{\rho}}\Big) \leqslant D\Big(\frac{\log n}{n}\Big)^{1/2} \qquad \text{a.s.} \tag{6.19}$$

where

$$D = \frac{C(1-a+\kappa)}{1-a}.$$

The last term of (6.15) is much more difficult to handle. Denote for all $n \geqslant 1$,

$$\Delta_n = \sum_{k=1}^{n} \frac{1}{k} P_k^n T_k \tag{6.20}$$

We recall that $T_n = A_n \delta_n + B_n\big(\widehat{V}_n - v^*\big)$ where A_n and B_n are given by (6.12) and (6.13). We already saw from (5.13) that

$$\|\delta_n\| \leqslant \frac{2\sqrt{2}}{\varepsilon}\|\widehat{V}_n - v^*\|^2$$

which implies that

$$\|T_n\| \leqslant \frac{2\sqrt{2}}{\varepsilon}\|A_n\|\,\|\widehat{V}_n - v^*\|^2 + \|B_n\|\,\|\widehat{V}_n - v^*\|. \tag{6.21}$$

Moreover, it follows from (3.5) and (3.6) that

$$\lim_{n\to+\infty} \|A_n\| = \|G_{\varepsilon}^{-1/2}(v^*)\| \qquad \text{and} \qquad \lim_{n\to+\infty} \|B_n\| = 0 \qquad \text{a.s.}$$

Consequently, we obtain from (3.5) and (6.21) that it exists a positive constant $b = (4\kappa)^{-1}$ where κ is introduced in (6.16), such that for n large enough,

$$\|T_n\| \leqslant b\|\widehat{V}_n - v^*\| \qquad \text{a.s.} \tag{6.22}$$

Define for all $n \geqslant 1$,

$$L_n = \frac{1}{n}\sum_{k=1}^{n} \|\widehat{V}_k - v^*\|. \tag{6.23}$$

We deduce from (6.20) together with (6.16) and (6.22) that for all $n \geqslant 1$,

$$\|\Delta_n\| \leqslant \frac{\kappa}{n^{\rho}}\sum_{k=1}^{n}\frac{k^{\rho}}{k}\|T_k\| \leqslant \frac{E}{n^{\rho}} + \frac{\kappa b}{n^{\rho}}\sum_{k=1}^{n}\frac{k^{\rho}}{k}\|\widehat{V}_k - v^*\| \leqslant \frac{E}{n^{\rho}} + \kappa b L_n \qquad \text{a.s.} \tag{6.24}$$

where E is a finite positive random variable. Putting together the three contributions (6.16), (6.19) and (6.24), we obtain from (6.15) that for all $n \geqslant 1$,

$$\|\widehat{U}_{n+1}\| \leqslant \frac{\tau\|\widehat{U}_1\| + E}{n^{\rho}} + D\Big(\frac{\log n}{n}\Big)^{1/2} + \kappa b L_n \qquad \text{a.s.}$$

which implies that a finite positive random variable F exists and a constant $0 < c < 1/2$ such that for all $n \geqslant 1$,

$$\|\widehat{V}_{n+1} - v^*\| \leqslant F\Big(\frac{\log n}{n}\Big)^{1/2} + cL_n \qquad \text{a.s.} \tag{6.25}$$

Herafter, we have from (6.23) and (6.25) that or all $n \geqslant 1$,

$$\begin{aligned} L_{n+1} &= \Big(1 - \frac{1}{n+1}\Big) L_n + \frac{1}{n+1} \|\widehat{V}_{n+1} - v^*\|, \\ &\leqslant \Big(1 - \frac{1}{n+1}\Big) L_n + \frac{1}{n+1} \Big(F \Big(\frac{\log n}{n}\Big)^{1/2} + c L_n \Big) \qquad \text{a.s} \\ &\leqslant \Big(1 - \frac{d}{n+1}\Big) L_n + \frac{F}{(n+1)} \Big(\frac{\log n}{n}\Big)^{1/2} \qquad \text{a.s} \end{aligned}$$

where $d = 1 - c$. A straightforward induction yields that for all $n \geqslant 1$,

$$L_n \leqslant \prod_{k=2}^{n} \Big(1 - \frac{d}{k}\Big) L_1 + \sum_{k=2}^{n} \prod_{i=k+1}^{n} \Big(1 - \frac{d}{i}\Big) \frac{F}{(k+1)} \Big(\frac{\log k}{k}\Big)^{1/2} \qquad \text{a.s} \tag{6.26}$$

However, it is well-known that

$$\prod_{k=2}^{n} \Big(1 - \frac{d}{k}\Big) \leqslant \Big(\frac{2}{n+1}\Big)^d \qquad \text{and} \qquad \prod_{i=k+1}^{n} \Big(1 - \frac{d}{i}\Big) \leqslant \Big(\frac{k+1}{n+1}\Big)^d.$$

Hence, we obtain from (6.26) that for all $n \geqslant 1$,

$$L_n \leqslant \Big(\frac{2}{n+1}\Big)^d L_1 + F \Big(\frac{1}{n+1}\Big)^d \sum_{k=2}^{n} \frac{(k+1)^d}{(k+1)} \Big(\frac{\log k}{k}\Big)^{1/2} \qquad \text{a.s}$$

Since $1/2 < d < 1$, it implies that

$$L_n \leqslant \frac{2^d L_1}{n^d} + \frac{F(\log n)^{1/2}}{n^d} \sum_{k=1}^{n} \frac{1}{k^{3/2-d}} \leqslant \frac{2^d L_1}{n^d} + \frac{2F}{2d-1} \Big(\frac{\log n}{n}\Big)^{1/2} \qquad \text{a.s}$$

leading to

$$L_n = \mathcal{O}\Big(\Big(\frac{\log n}{n}\Big)^{1/2}\Big) \qquad \text{a.s} \tag{6.27}$$

Finally, it follows from (6.25) and (6.27) that

$$\|\widehat{V}_n - v^*\|^2 = \mathcal{O}\Big(\frac{\log n}{n}\Big) \qquad \text{a.s.}$$

which completes the proof of (3.8).

• We now focus our attention on (3.9). We have from (6.7) that

$$\overline{S}_n - G_\varepsilon(v^*) = \frac{1}{n} M_n + \frac{1}{n}\big(I_J + R_n\big) + \frac{1}{n} \sum_{k=1}^{n} \big(G_\varepsilon(\widehat{V}_{k-1}) - G_\varepsilon(v^*)\big). \tag{6.28}$$

On the one hand, let $M_n(u) = u^T M_n u$ where $u \in \mathbb{R}^J$. We already saw that $(M_n(u))$ is a locally square-integrable martingale with increments bounded by $8\|u\|^2$. Moreover, its predictable quadratic variation satisfies $\langle M(u) \rangle_n \leqslant 16 n \|u\|^4$. Therefore, we obtain from the third part of Theorem 1.3.24 in [18] that

$$|M_n(u)|^2 = \mathcal{O}(n \log n) \qquad \text{a.s.}$$

which implies that

$$\frac{1}{n}\|M_n\| = \mathcal{O}\Big(\Big(\frac{\log n}{n}\Big)^{1/2}\Big) \qquad \text{a.s.} \tag{6.29}$$

On the other hand, we already saw from Lemma 5.2 that

$$\frac{1}{n}\sum_{k=1}^{n}\|G_\varepsilon(\widehat{V}_{k-1}) - G_\varepsilon(v^*)\| \leqslant \frac{4L_n}{\varepsilon}$$

which ensures via (6.27) that

$$\frac{1}{n}\sum_{k=1}^{n}\|G_\varepsilon(\widehat{V}_{k-1}) - G_\varepsilon(v^*)\| = \mathcal{O}(\Big(\Big(\frac{\log n}{n}\Big)^{1/2}\Big) \qquad \text{a.s.} \tag{6.30}$$

Furthermore, we also have

$$\frac{1}{n}\|R_n\| \leqslant \Big(\frac{\gamma\max(\nu)}{1-\beta}\Big)\frac{1}{n^\beta} \tag{6.31}$$

where $\beta < 1/2$. Consequently, we deduce the almost sure rate of convergence (3.9) for $\overline{S}_n$ from the conjunction of (6.28), (6.29), (6.30) and (6.31). Finally, we obtain the almost sure rate of convergence (3.9) for $\overline{S}_n^{-1}$ from the identity

$$\overline{S}_n^{-1} - G_\varepsilon^-(v^*) = \overline{S}_n^{-1}\big(G_\varepsilon(v^*) - \overline{S}_n\big)G_\varepsilon^-(v^*),$$

which completes the proof of Theorem 3.2. □

6.4 Proofs of the asymptotic normality results

Proof of Theorem 3.3. We now prove the asymptotic normality for the SGN algorithm.
• We recall from (6.15) that for all $n \geqslant 1$,

$$\sqrt{n}\big(\widehat{V}_{n+1} - v^*\big) = -\sqrt{n}G_\varepsilon^{-1/2}(v^*)Q_{n+1} + R_n \tag{6.32}$$

where $R_n = \sqrt{n}G_\varepsilon^{-1/2}(v^*)\big(P_0^n\widehat{U}_1 - \Delta_n\big)$ with $\widehat{U}_1 = G_\varepsilon^{1/2}(v^*)\big(\widehat{V}_1 - v^*\big)$,

$$Q_{n+1} = \sum_{k=1}^{n}\frac{1}{k}P_k^n A_k\varepsilon_{k+1},$$

$$\Delta_n = \sum_{k=1}^{n}\frac{1}{k}P_k^n T_k,$$

On the one hand, we claim that the remainder R_n vanishes almost surely,

$$\lim_{n\to\infty} R_n = 0 \qquad \text{a.s.} \tag{6.33}$$

As a matter of fact, we obviously have from (6.16) that

$$\lim_{n\to\infty}\sqrt{n}P_0^n\widehat{U}_1 = 0 \qquad \text{a.s.}$$

Moreover, we deduce from the proof of Theorem 3.2 together with (6.21) that

$$\|T_n\| = \mathcal{O}\Big(\frac{\log n}{n}\Big) + \mathcal{O}\Big(\frac{1}{n^\beta}\Big(\frac{\log n}{n}\Big)^{1/2}\Big) = \mathcal{O}\Big(\frac{1}{n^\beta}\Big(\frac{\log n}{n}\Big)^{1/2}\Big) \qquad \text{a.s.} \tag{6.34}$$

since $0 < \beta < 1/2$ and

$$\|B_n\| = \mathcal{O}\Big(\frac{1}{n^\beta}\Big) \qquad \text{a.s.}$$

where B_n is defined by (6.13). Therefore, we obtain from (6.34) that there exists a finite positive random variable C such that for all $n \geqslant 1$

$$\|T_n\| \leqslant \frac{C}{n^\beta}\Big(\frac{\log n}{n}\Big)^{1/2} \qquad \text{a.s.} \tag{6.35}$$

Consequently, it follows from (6.16) and (6.35) that for all $n \geqslant 1$,

$$\|\Delta_n\| \leqslant \frac{C\kappa}{n^\rho}\sum_{k=1}^{n}\frac{k^\rho}{k^{1+\beta}}\Big(\frac{\log k}{k}\Big)^{1/2} \leqslant \frac{C\kappa\sqrt{\log n}}{n^\rho}\sum_{k=1}^{n}\frac{1}{k^a} \qquad \text{a.s.}$$

where $a = 3/2 + \beta - \rho$, leading to

$$\|\Delta_n\| \leqslant \frac{D}{n^\beta}\Big(\frac{\log n}{n}\Big)^{1/2} \qquad \text{a.s.} \tag{6.36}$$

with $D = \kappa C/(1-a)$. Hence, as $\beta > 0$, we infer from (6.36) that

$$\lim_{n\to\infty}\sqrt{n}\Delta_n = 0 \qquad \text{a.s}$$

which clearly implies that (6.33) is satisfied. On the other hand, Q_{n+1} is a sum of weighted martingale differences. We deduce from the first part of Proposition B.2 in [50] that

$$\sqrt{n}\,Q_{n+1} \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \Sigma\big) \tag{6.37}$$

where the asymptotic covariance matrix Σ is given by the integral form

$$\begin{aligned}
\Sigma &= \int_0^\infty \Big(\exp\Big(-\Big(\Gamma_\varepsilon(v^*) - \frac{1}{2}P_J\Big)s\Big)\Big)^2\,ds \\
&= \int_0^\infty \Big(\exp\Big(-2s\Big(\Gamma_\varepsilon(v^*) - \frac{1}{2}P_J\Big)\Big)\,ds \\
&= \frac{1}{2}\int_0^\infty \Big(\exp\Big(-t\big(\Gamma_\varepsilon(v^*) - \frac{1}{2}P_J\big)\Big)\,dt \\
&= \frac{1}{2}\Big[-\Big(\Gamma_\varepsilon(v^*) - \frac{1}{2}P_J\Big)^{-}\exp\Big(-t\Big(\Gamma_\varepsilon(v^*) - \frac{1}{2}P_J\Big)\Big]_0^\infty \\
&= \Big(2\Gamma_\varepsilon(v^*) - P_J\Big)^{-}.
\end{aligned}$$

Therefore, we obtain from (6.37) that

$$\sqrt{n}\,G_\varepsilon^{-1/2}(v^*)Q_{n+1} \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, G_\varepsilon^{-1/2}(v^*)\big(2\Gamma_\varepsilon(v^*) - P_J\big)^{-}G_\varepsilon^{-1/2}(v^*)\Big). \tag{6.38}$$

Finally, it follows from (6.32) together with (6.33) and (6.38) that

$$\sqrt{n}\big(\widehat{V}_n - v^*\big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, G_\varepsilon^{-1/2}(v^*)\big(2\Gamma_\varepsilon(v^*) - P_J\big)^{-} G_\varepsilon^{-1/2}(v^*)\Big),$$

which is exactly what we wanted to prove.

• It only remains to prove the asymptotic normality (3.14). We already saw from inequality (5.11) that for all $v \in \mathbb{R}^J$,

$$H_\varepsilon(v) - H_\varepsilon(v^*) \leqslant \frac{1}{2\varepsilon}\|v - v^*\|^2.$$

Moreover, we have from (2.7) that $\widehat{W}_n - W_\varepsilon(\mu,\nu)$ can be splitted into two terms,

$$\begin{aligned}\sqrt{n}\big(\widehat{W}_n - W_\varepsilon(\mu,\nu)\big) &= \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\big(H_\varepsilon(v^*) - h_\varepsilon(X_k, \widehat{V}_{k-1})\big),\\ &= \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\big(H_\varepsilon(\widehat{V}_{k-1}) - h_\varepsilon(X_k, \widehat{V}_{k-1})\big) - \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\big(H_\varepsilon(\widehat{V}_{k-1}) - H_\varepsilon(v^*)\big).\end{aligned} \tag{6.39}$$

The second term in equation (6.39) goes to 0 a.s. thanks to the almost sure rate of convergence (3.8),

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\big(H_\varepsilon(\widehat{V}_{k-1}) - H_\varepsilon(v^*)\big) \leqslant \frac{1}{2\varepsilon\sqrt{n}}\sum_{k=1}^{n}\|\widehat{V}_{k-1} - v^*\|^2 = \mathcal{O}\Big(\frac{\log^2(n)}{\sqrt{n}}\Big) \qquad \text{a.s.}$$

Finally, the first term in equation (6.39) is dealt using argument from the proof of Theorem 3.5 in [5], allowing to prove that it satisfies the asymptotic normality (3.14). This completes the proof of Theorem 3.3. □

6.5 Proofs of the non-asymptotic rates of convergence

We first detail some keystone results related to the use of the Kurdyka-Łojasiewicz functional inequality that is at the heart of the proof of Theorem 3.4.

6.5.1 Kurdyka-Łojasiewicz inequality.

The analysis that we carry out is essentially based on the so-called Kurdyka-Łojasiewicz functional inequality. We refer to the initial works [31,32] and to [10,11] for the use of such inequality in deterministic optimization. To this end, let $\widetilde{H}_\varepsilon$ be the positive and convex function defined, for all $v \in \langle \boldsymbol{v}_J\rangle^\perp$, by

$$\widetilde{H}_\varepsilon(u) = H_\epsilon(G_*^{-1/2}u) - H_\epsilon(G_*^{-1/2}u^*),$$

where $G_*^{-1/2}$ stands for the square root of the Moore-Penrose inverse of $G_* = G_\varepsilon(v^*)$, and $u^* = G_*^{1/2}v^*$. Since $G_*^{-1/2}$ is symmetric, we notice that $\nabla\widetilde{H}_\varepsilon(u) = G_*^{-1/2}\nabla H_\varepsilon(G_*^{-1/2}u)$ and

$\nabla^2 \widetilde{H}_\varepsilon(u) = G_*^{-1/2} \nabla^2 H_\varepsilon(G_*^{-1/2} u) G_*^{-1/2}$. Hence, we obtain from the upper bounds (5.7) and (5.8) that for all $u \in \langle \boldsymbol{v}_J \rangle^\perp$,

$$\|\nabla \widetilde{H}_\varepsilon(u)\| \leqslant 2\lambda_{\max}(G_*^{-1/2}), \tag{6.40}$$

and

$$\lambda_{\max}\big(\nabla^2 \widetilde{H}_\varepsilon(u)\big) \leqslant \frac{1}{\varepsilon} \lambda_{\max}(G_*^-). \tag{6.41}$$

Consequently, it follows from the Taylor-Lagrange formula that for all $u \in \langle \boldsymbol{v}_J \rangle^\perp$,

$$\widetilde{H}_\varepsilon(u) \leqslant \frac{1}{2\varepsilon} \lambda_{\max}(G_*^-) \|u - u^*\|^2. \tag{6.42}$$

First of all, we verify that the function $\widetilde{H}_\varepsilon$ satisfies a Kurdyka-Łojasiewciz inequality as stated in the next proposition. We refer the reader to [22] and the references therein for further details on this topic.

Proposition 6.1. *There exist two positive constants $m_\varepsilon < M$ such that, for all $u \in \langle \boldsymbol{v}_J \rangle^\perp$ with $u \neq u^*$,*

$$0 < m_\varepsilon \leqslant \|\nabla \widetilde{H}_\varepsilon(u)\|^2 + \frac{\|\nabla \widetilde{H}_\varepsilon(u)\|^2}{\widetilde{H}_\varepsilon(u)} \leqslant M < +\infty. \tag{6.43}$$

Moreover, the constant m_ε can be chosen as

$$m_\varepsilon = \varepsilon \lambda_{\min}(G_\varepsilon(v^*)) \min\Big(1, \frac{\varepsilon}{4}\Big). \tag{6.44}$$

The proof of this key inequality is postponned to Appendix B. From equation (6.44), we clearly observe that the magnitude of the constant m_ε depends on ε. Nevertheless, in the analysis carried out in this paper, the regularization parameter ε is held fixed, and we will not be interested in deriving sharp upper bounds depending on ε for the mean square error of $\widehat{V}_n$. We believe that a careful analysis of the role of ε on the convergence of the SGN algorithm is a difficult issue that is left open for future investigation.

6.5.2 Choice of a Lyapunov function

Hereafter, a key step in our analysis is based on the Lyapunov function Φ defined, for all $u \in \langle \boldsymbol{v}_J \rangle^\perp$, by

$$\Phi(u) = \widetilde{H}_\varepsilon(u) \exp(\widetilde{H}_\varepsilon(u)). \tag{6.45}$$

On the one hand, it follows from the elementary inequality $\exp(x) \leqslant 1 + x \exp(x)$ that for all $u \neq u^*$,

$$\frac{\Phi(u)}{\widetilde{H}_\varepsilon(u)} \leqslant 1 + \Phi(u). \tag{6.46}$$

We shall repeatedly use inequality (6.46) in all the sequel. On the other hand, we can easily compute for all $u \neq u^*$,

$$\nabla \Phi(u) = \Big(1 + \frac{1}{\widetilde{H}_\varepsilon(u)}\Big) \Phi(u) \nabla \widetilde{H}_\varepsilon(u),$$

which implies that

$$\langle \nabla\Phi(u), \nabla\widetilde{H}_\varepsilon(u)\rangle = \Big(1 + \frac{1}{\widetilde{H}_\varepsilon(u)}\Big)\Phi(u)\|\nabla\widetilde{H}_\varepsilon(u)\|^2.$$

Consequently, we deduce from Proposition 6.1 that $m_\varepsilon\Phi(u) \leqslant \langle \nabla\Phi(u), \nabla\widetilde{H}_\varepsilon(u)\rangle \leq M\Phi(u)$. Moreover, if Σ denotes a positive semi-definite matrix, by an application of Proposition 6.1, we also obtain the following *key* lower bound

$$\Big\langle \nabla\Phi(u), \Sigma\nabla\widetilde{H}_\varepsilon(u)\Big\rangle \geq m_\varepsilon\lambda_{\min}(\Sigma)\Phi(u), \tag{6.47}$$

that will be useful to control the non-asymptotic rate of convergence of the SGN algorithm. Finally, as remarked in [22], a straightforward computation leads, for all $u \neq u^*$, to

$$\nabla^2\Phi(u) = \Phi(u)\Big(\Big(1 + \frac{2}{\widetilde{H}_\varepsilon(u)}\Big)\nabla\widetilde{H}_\varepsilon(u)\nabla\widetilde{H}_\varepsilon(u)^T + \Big(1 + \frac{1}{\widetilde{H}_\varepsilon(u)}\Big)\nabla^2\widetilde{H}_\varepsilon(u)\Big).$$

Hence, using the fact that $\lambda_{\max}\big(\nabla\widetilde{H}_\varepsilon(u)\nabla\widetilde{H}_\varepsilon(u)^T\big) = \|\nabla\widetilde{H}_\varepsilon(u)\|^2$, we obtain that for all $u \neq u^*$,

$$\lambda_{\max}(\nabla^2\Phi(u)) \leqslant \Phi(u)\Big(\Big(1 + \frac{2}{\widetilde{H}_\varepsilon(u)}\Big)\|\nabla\widetilde{H}_\varepsilon(u)\|^2 + \Big(1 + \frac{1}{\widetilde{H}_\varepsilon(u)}\Big)\lambda_{\max}\big(\nabla^2\widetilde{H}_\varepsilon(u)\big)\Big).$$

Therefore, using inequality (6.46) together with the upper bounds (6.40) and (6.41), we obtain that for all $u \in \langle \boldsymbol{v}_J\rangle^\perp$ with $u \neq u^*$,

$$\lambda_{\max}(\nabla^2\Phi(u)) \leqslant \delta_\epsilon\lambda_{\max}(G_*^-)(1 + \Phi(u)) \tag{6.48}$$

where $\delta_\varepsilon = 2(6 + \varepsilon^{-1})$. Inequality (6.48) will also be crucial to derive the non-asymptotic rate of convergence of the SGN algorithm. Finally, thanks to the following result, we will be able to relate the study of the Lyapunov function Φ to the quadratic risk of $\widehat{V}_n$ and S_n.

Proposition 6.2. *There exists a positive constant d_ε such that for all $u \in \langle \boldsymbol{v}_J\rangle^\perp$,*

$$\|u - u^*\|^2 \leqslant d_\varepsilon\Phi(u). \tag{6.49}$$

Proof. First, one can verify that for all $u \in \langle \boldsymbol{v}_J\rangle^\perp$ in a neighborhood of u^* with $u \neq u^*$, the function $\|u - u^*\|^{-2}\widetilde{H}_\varepsilon(u)$ is lower bounded. Moreover, using that $\widetilde{H}_\varepsilon$ is a convex function on $\langle \boldsymbol{v}_J\rangle^\perp$ that attains its minimal value at u^* with a non-degenerate minimum, we also have

$$\liminf_{\|u\|\to+\infty} \frac{\widetilde{H}_\varepsilon(u)}{\|u\|} > 0.$$

It implies that for any positive t,

$$\lim_{\|u\|\to+\infty} \frac{\exp\big(t\widetilde{H}_\varepsilon(u)\big)}{\|u - u^*\|^2} = +\infty.$$

Since $\widetilde{H}_\varepsilon(u) \geqslant 0$, one always has $\Phi(u) \geqslant \widetilde{H}_\varepsilon(u)$. Consequently, for all $u \neq u^*$,

$$\begin{aligned}\frac{\Phi(u)}{\|u - u^*\|^2} &= \frac{\Phi(u)}{\|u - u^*\|^2}\mathbb{1}_{\|u-u^*\|\leqslant 1} + \frac{\Phi(u)}{\|u - u^*\|^2}\mathbb{1}_{\|u-u^*\|\geqslant 1},\\ &\geqslant \frac{\widetilde{H}_\varepsilon(u)}{\|u - u^*\|^2}\mathbb{1}_{\|u-u^*\|\leqslant 1} + \min_{\|u-u^*\|\geq 1} \frac{\widetilde{H}_\varepsilon(u)\exp\big(\widetilde{H}_\varepsilon(u)\big)}{\|u - u^*\|^2} \geqslant \frac{1}{d_\varepsilon}\end{aligned}$$

for some positive constant d_ε, where we used the local behavior around u^* of $\widetilde{H}_\varepsilon$ to derive a lower bound for the first term and the asymptotic behavior of $\widetilde{H}_\varepsilon$ for the second one with $t = 1$, which is exactly what we wanted to prove. □

6.5.3 A recursive inequality and proof of Theorem 3.4

We first describe the one-step evolution of the sequence $(\Phi(G_*^{1/2}\widehat{V}_n))$ where $(\widehat{V}_n)$ is the recursive sequence defined by (2.6) corresponding to the SGN algorithm. From this analysis, we shall also deduce the rate of convergence of the expected quadratic risk associated with $\widehat{V}_n$ and $\overline{S}_n$. Denote $\widetilde{U}_n = G_*^{1/2}\widehat{V}_n$.

Proposition 6.3. *Assume that $\alpha \in [0, 1/2[$ and that $\alpha + \beta < 1/2$. Then, there exist an integer $n_0 \geqslant J$ and a positive constant $c_\varepsilon > 0$ such that, for all $n \geqslant n_0$,*

$$\mathbb{E}\big[\Phi(\widetilde{U}_{n+1})|\mathcal{F}_n\big] \leqslant \Big(1 - \frac{m_\varepsilon n^\alpha \lambda_{\min}(G_*)\lambda_{\min}(S_n^{-1})}{2}\Big)\Phi(\widetilde{U}_n) + c_\varepsilon n^{2\alpha}\lambda_{\max}^2(S_n^{-1}). \quad (6.50)$$

The proof of Proposition 6.3 is postponed to Appendix B. We are now in position to establish the non-asymptotic rates of convergence for the SGN algorithm.

Proof of Theorem 3.4. In the proof, we use the notation c_ε to denote a positive constant (depending on ε and possibly on α and β) that is independent from n and whose value may change from line to line. We only consider the situation where $\alpha \in]0, 1/2[$ and our analysis is based on the function Φ. We will establish that for all $n \geqslant 1$,

$$\mathbb{E}[\Phi(\widetilde{U}_n)] \leqslant \frac{c_\varepsilon}{n^{1-\alpha}}. \quad (6.51)$$

Step 1: Preliminary rate. Our starting point is inequality (6.50) that we combine with (B.6) and (B.7) to obtain that there exists an integer n_0 such that, for all $n \geq n_0$,

$$\mathbb{E}\big[\Phi(\widetilde{U}_{n+1})|\mathcal{F}_n\big] \leqslant (1 - c_1(n)n^{-1+\alpha})\Phi(\widetilde{U}_n) + c_2(n)n^{2(\alpha+\beta-1)} \quad (6.52)$$

where $n_J = n/J$,

$$c_1(n) = \frac{m_\varepsilon \lambda_{\min}(G_*)}{8 + 4\gamma \max(\nu) + 2n^{-1}},$$
$$c_2(n) = \frac{c_\varepsilon}{J^{2(\beta-1)}\big(\gamma \min(\nu)(1 - n_J^{-1})^{1-\beta} + (1 - 2\gamma \min(\nu))n_J^{-1}\big)^2}.$$

By taking the expectation on both sides of (6.52), we obtain that for all $n \geqslant n_0$,

$$\mathbb{E}[\Phi(\widetilde{U}_{n+1})] \leqslant \big(1 - c_1(n)n^{-1+\alpha}\big)\,\mathbb{E}[\Phi(\widetilde{U}_n)] + c_2(n)n^{-2(1-\alpha-\beta)}.$$

Hereafter, it is not hard to see that for n large enough, there exist two positive constants c_1 and c_2, depending on ε, such that $c_1(n) \geqslant c_1$ and $c_2(n) \leqslant c_2$. Hence, there exists an integer n_0 such that, for all $n \geqslant n_0$,

$$\mathbb{E}[\Phi(\widetilde{U}_{n+1})] \leqslant \big(1 - c_1 n^{-1+\alpha}\big)\,\mathbb{E}[\Phi(\widetilde{U}_n)] + c_2 n^{-2(1-\alpha-\beta)}.$$

Therefore, it follows from the proof of Lemma A.3 in [5] that there exist a positive constant c_ε and an integer n_0 such that, for all $n \geqslant n_0$,

$$\mathbb{E}[\Phi(\widetilde{U}_n)] \leqslant \frac{c_\varepsilon}{n^{1-2(\alpha+\beta)}}.$$

We emphasize that at this stage, we do not obtain the announced result that necessitates further work with a plug-in strategy. The rest of the proof details this additional step.

Step 2: Plug-in. Let us now explain how one may improve the above result from $n^{-(1-2(\alpha+\beta))}$ to $n^{-(1-2\alpha)}$. First of all, thanks to Proposition 6.2, we obtain via Step 1 that

$$\mathbb{E}[\|\widetilde{U}_n - u^*\|^2] \leqslant \frac{c_\varepsilon}{n^{1-2(\alpha+\beta)}}. \tag{6.53}$$

Next, we shall consider the study of the convergence rate of $\overline{S}_n$ to improve the pessimistic bounds (B.6) and (B.7). If p_n denotes the largest integer such that $p_n J \leqslant n$, we already saw from (4.3) and (6.4) that for all $n \geqslant 1$,

$$\overline{S}_n = \overline{\Sigma}_n + \overline{R}_n$$

where

$$\overline{\Sigma}_n = \frac{1}{n}\sum_{k=1}^{n} \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1}) \nabla_v h_\varepsilon(X_k, \widehat{V}_{k-1})^T,$$

$$\overline{R}_n = \frac{1}{n} I_J + \frac{1}{n}\Big(\sum_{m=1}^{p_n} m^{-\beta}\Big)\gamma\,\mathrm{diag}(\nu) + \frac{1}{n}\sum_{k=p_n+1}^{n} \gamma\Big(1 + \lfloor \frac{k}{J} \rfloor\Big)^{-\beta} Z_k Z_k^T.$$

On the one hand, it is not hard to see that it exists a positive constant c_3 such that

$$\|\overline{R}_n\|_F^2 \leqslant \frac{c_3}{n^{2\beta}}. \tag{6.54}$$

On the other hand, starting from the fact that

$$\overline{\Sigma}_{n+1} = \overline{\Sigma}_n + \frac{1}{n+1}\Big(\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)^T - \overline{\Sigma}_n\Big),$$

we obtain that

$$\begin{aligned}\|\overline{\Sigma}_{n+1} - G_*\|_F^2 = \|\overline{\Sigma}_n - G_*\|_F^2 &+ \frac{1}{(n+1)^2}\|\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)^T - \overline{\Sigma}_n\|_F^2 \\ &+ \frac{2}{n+1}\langle \tilde{S}_n - G_*, \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)^T - \overline{\Sigma}_n\rangle_F.\end{aligned}$$

Therefore, by taking the conditional expectation on both sides of the above equality, we obtain that

$$\begin{aligned}\mathbb{E}\Big[\|\overline{\Sigma}_{n+1} - G_*\|_F^2 | \mathcal{F}_n\Big] &= \|\overline{\Sigma}_n - G_*\|_F^2 + \frac{2}{n+1}\langle \overline{\Sigma}_n - G_*, G_\varepsilon(\widehat{V}_n) - \overline{\Sigma}_n\rangle_F \\ &\quad + \frac{1}{(n+1)^2}\mathbb{E}\Big[\|\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)^T - \overline{\Sigma}_n\|_F^2 | \mathcal{F}_n\Big], \\ &= \|\overline{\Sigma}_n - G_*\|_F^2 - \frac{2}{n+1}\|\overline{\Sigma}_n - G_*\|_F^2 + \frac{2}{n+1}\langle \overline{\Sigma}_n - G_*, G_\varepsilon(\widehat{V}_n) - G_*\rangle_F \\ &\quad + \frac{1}{(n+1)^2}\mathbb{E}\Big[\|\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)^T - \overline{\Sigma}_n\|_F^2 | \mathcal{F}_n\Big], \\ &\leqslant \|\overline{\Sigma}_n - G_*\|_F^2\Big(1 - \frac{1}{n+1}\Big) + \frac{1}{n+1}\|G_\varepsilon(\widehat{V}_n) - G_*\|_F^2 \\ &\quad + \frac{1}{(n+1)^2}\mathbb{E}\Big[\|\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)^T - \overline{\Sigma}_n\|_F^2 | \mathcal{F}_n\Big] \end{aligned} \tag{6.55}$$

where the last line follows from Cauchy-Schwarz and Young inequalities. Moreover, we deduce from inequality (5.6) that

$$\mathbb{E}\Big[\|\nabla_v h_\varepsilon(X_{n+1},\widehat{V}_n)\nabla_v h_\varepsilon(X_{n+1},\widehat{V}_n)^T-\overline{\Sigma}_n\|_F^2|\mathcal{F}_n\Big]$$

$$\leqslant 2\mathbb{E}\Big[\|\nabla_v h_\varepsilon(X_{n+1},\widehat{V}_n)\nabla_v h_\varepsilon(X_{n+1},\widehat{V}_n)^T\|_F^2|\mathcal{F}_n\Big]+2\|\overline{\Sigma}_n\|_F^2\leqslant 2\times 4^2+2\times 4=40.$$

Furthermore, we obtain from (5.15) that

$$\|G_\varepsilon(\widehat{V}_n)-G_*\|_F^2\leqslant\frac{16J}{\varepsilon^2}\|\widehat{V}_n-v_*\|^2.$$

Using the previous bounds in (6.55) leads to the recursive inequality

$$\mathbb{E}\Big[\|\overline{\Sigma}_{n+1}-G_*\|_F^2\Big]\leqslant\Big(1-\frac{1}{n+1}\Big)\mathbb{E}\big[\|\overline{\Sigma}_n-G_*\|_F^2\big]+\frac{40}{(n+1)^2}+\frac{16J}{(n+1)\varepsilon^2}\mathbb{E}\big[\|\widehat{V}_n-v^*\|^2\big].$$

Since $\widetilde{U}_n-u^*=G_*^{1/2}\big(\widehat{V}_n-v^*\big)$, inequality (6.53) ensures that

$$\mathbb{E}\Big[\|\overline{\Sigma}_{n+1}-G_*\|_F^2\Big]\leqslant\Big(1-\frac{1}{n+1}\Big)\mathbb{E}\big[\|\overline{\Sigma}_n-G_*\|_F^2\big]+\frac{40}{(n+1)^2}+\frac{16Jc_\varepsilon\lambda_{max}(G_*^-)}{n^{2(1-\alpha-\beta)}\varepsilon^2}.$$

By applying Lemma A.3 in [5], we obtain that for n large enough,

$$\mathbb{E}\big[\|\overline{\Sigma}_n-G_*\|_F^2\big]\leqslant\frac{c_\varepsilon}{n^{(1-2(\alpha+\beta))}}.$$

From this last inequality and (6.54), we conclude that for n large enough,

$$\mathbb{E}\big[\|\overline{S}_n-G_*\|_F^2\big]\leqslant\frac{c_\varepsilon}{n^{(1-2(\alpha+\beta))}}+\frac{c_3}{n^{2\beta}}. \tag{6.56}$$

Since $4\beta<1-2\alpha$, it follows that $1-2(\alpha+\beta)>2\beta$, which means that $n^{-(1-2(\alpha+\beta))}$ decays faster than $n^{-2\beta}$ and the second inequality in (3.18) holds true.

• From now on, we focus our attention on the first inequality in (3.18). It follows from (6.50) and the previous calculation that for n large enough,

$$\mathbb{E}[\Phi(\widetilde{U}_{n+1})]\leqslant(1-c_1(n)n^{-1+\alpha})\mathbb{E}[\Phi(\widetilde{U}_n)]+c_\varepsilon n^{2\alpha-2}\mathbb{E}\big[\lambda_{\max}^2(\overline{S}_n^{-1})\big], \tag{6.57}$$

using that $\overline{S}_n^{-1}=nS_n^{-1}$. The identity $\overline{S}_n^{-1}-G_*^-=G_*^-(G_*-\overline{S}_n)\overline{S}_n^{-1}$ implies that

$$\lambda_{\max}(\overline{S}_n^{-1})\leqslant\|\overline{S}_n^{-1}-G_*^-\|_2+\|G_*^-\|_2\leqslant\lambda_{\max}(G_*^-)\lambda_{\max}(\overline{S}_n^{-1})\|\overline{S}_n-G_*\|_2+\|G_*^-\|_2.$$

It follows from inequality (B.6) that for n large enough, $\lambda_{\max}(\overline{S}_n^{-1})=n\lambda_{\max}(S_n^{-1})\leqslant c_4n^\beta$ where c_4 is a positive constant. It ensures that for n large enough,

$$\lambda_{\max}^2(\overline{S}_n^{-1})\leqslant 2c_4^2\lambda_{\max}^2(G_*^-)n^{2\beta}\|\overline{S}_n-G_*\|_2^2+2\|G_*^-\|_2^2.$$

Since $\|\overline{S}_n-G_*\|_2^2\leqslant\|\overline{S}_n-G_*\|_F^2$, we obtain from inequality (6.56) that for n large enough,

$$\mathbb{E}\Big[\lambda_{\max}^2(\overline{S}_n^{-1})\Big]\leqslant c_\varepsilon\Big(1+\frac{n^{4\beta}}{n^{1-2\alpha}}\Big)+2\|G_*^-\|_2^2.$$

Consequently, we deduce from the condition $4\beta < 1 - 2\alpha$ that there exists a positive constant C_ε such that for all $n \geqslant 1$,

$$\mathbb{E}\Big[\lambda_{\max}^2(\overline{S}_n^{-1})\Big] \leqslant C_\varepsilon. \tag{6.58}$$

It follows from (6.57) and (6.58) that there exist two postive constants c_1 and c_ε such that for n large enough,

$$\mathbb{E}[\Phi(\widetilde{U}_{n+1})] \leqslant (1 - c_1 n^{-1+\alpha})\mathbb{E}[\Phi(\widetilde{U}_n)] + c_\varepsilon n^{2\alpha-2}.$$

Hereafter, using once again Lemma A.3 in [5], we obtain that for all $n \geqslant 1$,

$$\mathbb{E}[\Phi(\widetilde{U}_n)] \leqslant \frac{c_\varepsilon}{n^{1-\alpha}}$$

which is exactly the announced inequality (6.51). Finally, as $\widehat{V}_n - v_* = G_*^{-1/2}\big(\widetilde{U}_n - u^*\big)$, the first inequality in (3.18) clearly follows from (6.51) together with Proposition 6.2.

• It only remains to prove the inequalities (3.19) and (3.20). We already saw the decomposition

$$\begin{aligned}\widehat{W}_n - W_\varepsilon(\mu,\nu) &= \frac{1}{n}\sum_{k=1}^{n}\big(H_\varepsilon(v^*) - h_\varepsilon(X_k,\widehat{V}_{k-1})\big),\\ &= \frac{1}{n}\sum_{k=1}^{n}\xi_k - \frac{1}{n}\sum_{k=1}^{n}\big(H_\varepsilon(\widehat{V}_{k-1}) - H_\varepsilon(v^*)\big),\end{aligned} \tag{6.59}$$

where the martingale increment $\xi_k = -h_\varepsilon(X_k,\widehat{V}_{k-1}) + H_\varepsilon(\widehat{V}_{k-1})$. As $\mathbb{E}[\xi_{k+1}\,|\mathcal{F}_k] = 0$, it follows from (6.59) together with inequality (5.11) and the first inequality in (3.18) that for all $n \geqslant 1$,

$$\Big|\mathbb{E}\big[\widehat{W}_n\big] - W_\varepsilon(\mu,\nu)\Big| = \Big|\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[H_\varepsilon(\widehat{V}_{k-1}) - H_\varepsilon(v^*)\big]\Big| \leqslant \frac{c_\varepsilon}{n}\sum_{k=1}^{n}\frac{1}{k^{1-\alpha}} \leqslant \frac{c_{\varepsilon,\alpha}}{n^{1-\alpha}},$$

which proves inequality (3.19).

Regarding the $\mathbb{L}^1$ risk $\mathbb{E}\big[\big|\widehat{W}_n - W_\varepsilon(\mu,\nu)\big|\big]$, we still use the decomposition (6.59). The triangle inequality and inequality (5.11) imply that

$$\begin{aligned}\mathbb{E}\big[\big|\widehat{W}_n - W_\varepsilon(\mu,\nu)\big|\big] &\leqslant \frac{1}{n}\mathbb{E}\Big[\Big|\sum_{k=1}^{n}\xi_k\Big|\Big] + \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\big[\big|H_\varepsilon(\widehat{V}_{k-1}) - H_\varepsilon(v^*)\big|\big],\\ &\leqslant \frac{1}{n}\Big(\mathbb{E}\Big(\sum_{k=1}^{n}\xi_k\Big)^2\Big)^{1/2} + \frac{1}{2\varepsilon n}\sum_{k=1}^{n}\mathbb{E}[\|\widehat{V}_{k-1} - v^*\|^2],\end{aligned}$$

where the last line comes from the Cauchy-Schwarz inequality. Let us now prove that $\sup_{k\geq 1}\mathbb{E}[\xi_k^2] < +\infty$. To this end, we observe that

$$\mathbb{E}[\xi_k^2] = \mathbb{E}\left(\mathbb{E}\left[\Big(h_\varepsilon(X_k,\widehat{V}_{k-1}) - H_\varepsilon(\widehat{V}_{k-1})\Big)^2\,|\mathcal{F}_k\right]\right) \leqslant \mathbb{E}[h_\varepsilon^2(X_k,\widehat{V}_{k-1})],$$

thanks to the property that $\mathbb{E}\left[h_\varepsilon(X_k, \widehat{V}_{k-1})\,|\mathcal{F}_k\right] = H_\varepsilon(\widehat{V}_{k-1})$. Using that $\|\partial_v h_\varepsilon(x,v)\| = \|\pi(x,v) - \nu\| \leqslant 2$, it follows by integration that $h_\varepsilon(x,v) \leqslant h_\varepsilon(x,v^*) + 2\|v - v^*\|$. We then deduce that

$$\mathbb{E}[\xi_k^2] \leqslant 2\mathbb{E}[h_\varepsilon^2(X, v^*)] + 4\mathbb{E}[\|\widehat{V}_{k-1} - v^*\|^2]$$

Using the first inequality in (3.18), it follows that $(\mathbb{E}[\|\widehat{V}_{k-1} - v^*\|^2])_{k\geq 1}$ is a bounded sequence. Moreover, arguing as in the proof of [5, Theorem 3.5], the condition $\int_{\mathcal{X}} c^2(x, y_j) d\mu(x) < +\infty$, for any $1 \leqslant j \leqslant J$, implies that $\mathbb{E}[h_\varepsilon^2(X, v^*)]$ is finite. Therefore, we conclude that $\sup_{k\geq 1} \mathbb{E}[\xi_k^2] < +\infty$. Hence, using once again the first inequality in (3.18) together with a conditional expectation argument, we obtain that

$$\mathbb{E}\big[\big|\widehat{W}_n - W_\varepsilon(\mu,\nu)\big|\big] \leqslant \frac{c_\varepsilon}{n}\sqrt{n} + \frac{c_\varepsilon}{2\varepsilon n}\sum_{k=1}^{n}\frac{1}{k^{1-\alpha}} \leqslant \frac{c_\varepsilon}{\sqrt{n}},$$

which proves inequality (3.20). This achieves the proof of Theorem 3.4. □

A Appendix - Proofs of auxiliary results

This appendix contains the proofs of some auxiliary results of the paper.

A.1 Proof of Lemma 5.2

Since $\nabla_v h_\varepsilon(X, v) = \pi(X, v) - \nu$, we first remark that

$$\begin{aligned} G_\varepsilon(v) &= L(v) + \nu\nu^T - \nu\mathbb{E}\big[\pi(X,v)\big]^T - \mathbb{E}\big[\pi(X,v)\big]\nu^T, \\ &= L(v) - \nu\nu^T - \nu\nabla H_\varepsilon(v)^T - \nabla H_\varepsilon(v)\nu^T, \end{aligned} \tag{A.1}$$

where $L(v) = \mathbb{E}\big[\pi(X,v)\pi(X,v)^T\big]$. For $u \in \langle \boldsymbol{v}_J\rangle^\perp$ and $t \in [0,1]$, we define the real-valued function $\phi_u(t) = u^T G_\varepsilon(v_t) u$ with $v_t = v^* + t(v - v^*)$. It follows from the decomposition (A.1) that

$$\phi_u(t) = u^T L(v_t) u - \langle u, \nu\rangle^2 - 2\langle u, \nu\rangle\langle \Phi(t), u\rangle$$

where $\Phi(t) = \nabla H_\varepsilon(v_t)$ is a vector-valued function satisfying

$$\Phi'(t) = \nabla^2 H_\varepsilon(v_t)(v - v^*) \qquad \text{and} \qquad \Phi''(t) = \nabla^3 H_\varepsilon(v_t)[v - v^*, v - v^*],$$

where $\nabla^3 H_\varepsilon$ stands for the third-order tensor derivative of H_ε. Now, since $\Phi(1) = \nabla H_\varepsilon(v)$ and $\Phi(0) = \nabla H_\varepsilon(v^*) = 0$, and using the property that $\phi_u(1) - \phi_u(0) = \int_0^1 \phi_u'(t)dt$, we obtain that

$$u^T(L(v) - L(v^*))u - 2\langle u, \nu\rangle\langle \nabla H_\varepsilon(v), u\rangle = \int_0^1 \phi_u'(t)dt. \tag{A.2}$$

We clearly have

$$\phi_u'(t) = \frac{\partial}{\partial t} u^T L(v_t) u - 2\langle u, \nu\rangle\langle \Phi'(t), u\rangle$$

where

$$\frac{\partial}{\partial t} u^T L(v_t) u = \langle \nabla_v \psi_u(v_t), v - v^*\rangle \qquad \text{with} \qquad \psi_u(v) = u^T L(v) u.$$

In addition, we also have $\langle \Phi'(t), u\rangle = u^T \nabla^2 H_\varepsilon(v_t)(v - v^*)$. Since

$$u^T L(v) u = \mathbb{E}\big[\langle u, \pi(X, v)\rangle^2\big]$$

and

$$\nabla_v \pi(x, v) = \frac{1}{\varepsilon}\Big(\mathrm{diag}(\pi(x, v)) - \pi(x, v)\pi(x, v)^T\Big),$$

one obtains that

$$\nabla_v \psi_u(v_t) = \frac{2}{\varepsilon}\mathbb{E}\big[\langle u, \pi(X, v_t)\rangle \left(\mathrm{diag}(\pi(X, v_t))u - \langle u, \pi(X, v_t)\rangle \pi(X, v_t)\right)\big].$$

Consequently,

$$\begin{aligned}\phi_u'(t) = \frac{2}{\varepsilon}\mathbb{E}\big[\langle u, \pi(X, v_t)\rangle \left(u^T \,\mathrm{diag}(\pi(X, v_t))(v - v^*) - \langle u, \pi(X, v_t)\rangle\langle \pi(X, v_t), v - v^*\rangle\right)\big] \\ - 2\langle u, \nu\rangle u^T \nabla^2 H_\varepsilon(v_t)(v - v^*).\end{aligned}$$

Then, we deduce from equality (5.4) that

$$\phi_u'(t) = \frac{2}{\varepsilon}\mathbb{E}\big[\langle u, \pi(X, v_t) - \nu\rangle u^T A_\varepsilon(X, v_t)(v - v^*)\big],$$

with $A_\varepsilon(x, v) = \mathrm{diag}(\pi(x, v)) - \pi(x, v)\pi(x, v)^T$. It follows from Cauchy-Schwarz inequality and the upper bound (5.6) that

$$|\langle u, \pi(X, v_t) - \nu\rangle| \leqslant 2\|u\|,$$

which together with the fact that $\lambda_{\max}(A_\varepsilon(x, v)) \leqslant 1$ yields

$$|\phi_u'(t)| \leqslant \frac{4}{\varepsilon}\|v - v^*\|\|u\|^2.$$

Hence, inserting the above upper bound in (A.2), we obtain that

$$\big|u^T(L(v) - L(v^*))u - 2\langle u, \nu\rangle\langle \nabla H_\varepsilon(v), u\rangle\big| \leqslant \frac{4}{\varepsilon}\|v - v^*\|\|u\|^2.$$

Therefore, in the sense of partial ordering between positive semi-definite matrices, we have shown that

$$-\frac{4}{\varepsilon}\|v - v^*\| I_J \leqslant L(v) - L(v^*) - \nabla H_\varepsilon(v)\nu^T - \nu \nabla H_\varepsilon(v)^T \leqslant \frac{4}{\varepsilon}\|v - v^*\| I_J.$$

Inequality (5.15) thus follows from the decomposition (A.1) since

$$G_\varepsilon(v) - G_\varepsilon(v^*) = L(v) - L(v^*) - \nu \nabla H_\varepsilon(v)^T - \nabla H_\varepsilon(v)\nu^T,$$

which completes the proof of Lemma 5.2.

A.2 A recursive formula to compute the inverse of S_n for the stochastic Newton algorithm

In this section, we discuss the construction of a recursive formula to compute, from the knowledge of S_{n-1}^{-1}, the inverse of the matrix S_n defined by the recursive equation (2.8) that corresponds to the use of the stochastic Newton (SN) algorithm. To this end, let us first recall the following matrix inversion lemma classically referred to as the Sherman-Morrison-Woodbury (SMW) formula [27], also known as Woodbury's formula or Riccati's matrix identity.

Lemma A.1 (SMW formula)**.** *Suppose that A and C are invertible matrices of size $d \times d$ and $q \times q$ respectively. Let U and V be $d \times q$ and $q \times d$ matrices. Then, $A + UCV$ is invertible iff $C^{-1} + VA^{-1}U$ is invertible. In that case, we have*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}. \tag{A.3}$$

A repeated use of the SMW formula allows to prove the following result.

Proposition A.1. *Let $P_J = I_J - \boldsymbol{v}_J \boldsymbol{v}_J^T$ be the projection matrix onto $\langle \boldsymbol{v}_J \rangle^\perp$. Suppose that*

$$\mathbb{S}_n = I_J + \sum_{k=1}^{n} \nabla_v^2 h_\varepsilon(X_k, \widehat{V}_{k-1}) = \mathbb{H}_n + \boldsymbol{v}_J \boldsymbol{v}_J^T,$$

where

$$\mathbb{H}_n = P_J + \sum_{k=1}^{n} \nabla_v^2 h_\varepsilon(X_k, \widehat{V}_{k-1}),$$

with $\mathbb{H}_0 = P_J$. Then, for all $n \geqslant 1$, one has $\mathbb{S}_n^{-1} = \mathbb{H}_n^- + \boldsymbol{v}_J \boldsymbol{v}_J^T$ with $\mathbb{H}_n^-$ that satisfies the recursive formula

$$\mathbb{H}_n^- = P_J \Big(\mathbb{H}_{n-1} + \frac{1}{\varepsilon} \operatorname{diag}\Big(\pi_n\Big)\Big)^{-1} P_J, \tag{A.4}$$

$$= \mathbb{H}_{n-1}^- - \mathbb{H}_{n-1}^- \left(\mathbb{H}_{n-1}^- + \varepsilon \operatorname{diag}\left(\pi_n^{-1}\right)\right)^{-1} \mathbb{H}_{n-1}^-, \tag{A.5}$$

where π_n^{-1} stands for the vector whose entries are the inverse of those of $\pi_n = \pi(X_n, \widehat{V}_{n-1})$.

Proof. For $n \geqslant 1$, we define

$$\tilde{S}_n = I_J + \sum_{k=1}^{n} \Big(\nabla_v^2 h_\varepsilon(X_k, \widehat{V}_{k-1}) + \boldsymbol{v}_J \boldsymbol{v}_J^T \Big) = \tilde{S}_{n-1} + \Sigma_n = \mathbb{H}_n + (n+1) \boldsymbol{v}_J \boldsymbol{v}_J^T, \tag{A.6}$$

where $\Sigma_n = \nabla_v^2 h_\varepsilon(X_n, \widehat{V}_{n-1}) + \boldsymbol{v}_J \boldsymbol{v}_J^T$ and $\tilde{S}_0 = I_J = P_J + \boldsymbol{v}_J \boldsymbol{v}_J^T$. As discussed in Section 5, the eigenvectors of the matrix $\nabla_v^2 h_\varepsilon(X_k, \widehat{V}_{k-1})$ associated to non-zero eigenvalues belong to $\langle \boldsymbol{v}_J \rangle^\perp$ for any $k \geqslant 1$, which implies that $\mathbb{H}_n$ is a matrix of rank $J-1$ and that $\boldsymbol{v}_J$ is its only eigenvector associated to the eigenvalue $\lambda_1 = 0$. Hence, for all $n \geqslant 0$, $\mathbb{H}_n$ is also a matrix such that all its eigenvectors associated to non-zero eigenvalues belong to $\langle \boldsymbol{v}_J \rangle^\perp$. Therefore, for any $n \geqslant 0$, the inverse of the matrix $\tilde{S}_n$ (which is of full rank J) satisfies the identity

$$\tilde{S}_n^{-1} = \mathbb{H}_n^- + \frac{1}{n+1} \boldsymbol{v}_J \boldsymbol{v}_J^T. \tag{A.7}$$

Moreover, given that $\mathbb{S}_n = \mathbb{H}_n + \boldsymbol{v}_J \boldsymbol{v}_J^T$ one has that

$$\mathbb{S}_n^{-1} = \mathbb{H}_n^- + \boldsymbol{v}_J \boldsymbol{v}_J^T = \tilde{S}_n^{-1} + \frac{n}{n+1} \boldsymbol{v}_J \boldsymbol{v}_J^T. \tag{A.8}$$

The computation of $\tilde{S}_n^{-1}$ can be done recursively as follows. By applying the SMW formula (A.3) with $U = V = I_J$ we obtain that

$$\tilde{S}_n^{-1} = \tilde{S}_{n-1}^{-1} - \tilde{S}_{n-1}^{-1} \left(\Sigma_n^{-1} + \tilde{S}_{n-1}^{-1} \right)^{-1} \tilde{S}_{n-1}^{-1}. \tag{A.9}$$

Now, introducing the notation $\pi_n = \pi(X_n, \widehat{V}_{n-1})$, we remark that

$$\nabla_v^2 h_\varepsilon(X_n, \widehat{V}_{n-1}) = \frac{1}{\varepsilon} \left(\operatorname{diag}(\pi_n) - \pi_n \pi_n^T \right)$$

is proportional to a multinomial matrix (up to a minus sign and the multiplicative factor ε^{-1}). Consequently, from the pseudo inverse formula of multinomial matrices [47], we obtain that the inverse of the matrix Σ_n is given by

$$\Sigma_n^{-1} = \varepsilon P_J \operatorname{diag}(\pi_n^{-1}) P_J + \boldsymbol{v}_J \boldsymbol{v}_J^T, \tag{A.10}$$

where π_n^{-1} stands for the vector whose entries are the inverse of those of π_n. Now, introducing the notation $T_{n-1} = \tilde{S}_{n-1}^{-1} + \boldsymbol{v}_J \boldsymbol{v}_J^T$ and $Q_{n-1} = \Sigma_n^{-1} + \tilde{S}_{n-1}^{-1}$, we deduce from equation (A.10) that

$$Q_{n-1} = T_{n-1} + \varepsilon P_J \operatorname{diag}(\pi_n^{-1}) P_J.$$

Consequently, by the SMW formula (A.3), it follows that

$$Q_{n-1}^{-1} = T_{n-1}^{-1} - T_{n-1}^{-1} P_J \left(P_J T_{n-1}^{-1} P_J + \frac{1}{\varepsilon} \operatorname{diag}(\pi_n) \right)^{-1} P_J T_{n-1}^{-1}.$$

Then, applying once again the SMW formula and equality (A.6), one has that

$$\begin{aligned} T_{n-1}^{-1} &= \tilde{S}_{n-1} - \frac{1}{\boldsymbol{v}_J^T \tilde{S}_{n-1} \boldsymbol{v}_J + 1} \tilde{S}_{n-1} \boldsymbol{v}_J \boldsymbol{v}_J^T \tilde{S}_{n-1} = \tilde{S}_{n-1} - \frac{n^2}{n+1} \boldsymbol{v}_J \boldsymbol{v}_J^T, \\ &= \mathbb{H}_{n-1} + \frac{n}{n+1} \boldsymbol{v}_J \boldsymbol{v}_J^T. \end{aligned}$$

Hence, by the fact that $\mathbb{H}_n$ maps the subspace $\langle \boldsymbol{v}_J \rangle^\perp$ onto itself and since P_J is the projection matrix onto $\langle \boldsymbol{v}_J \rangle^\perp$, we thus obtain that

$$Q_{n-1}^{-1} = T_{n-1}^{-1} - \mathbb{H}_{n-1} \left(\mathbb{H}_{n-1} + \frac{1}{\varepsilon} \operatorname{diag}(\pi_n) \right)^{-1} \mathbb{H}_{n-1}.$$

Consequently, we have shown that

$$Q_{n-1}^{-1} = \left(\Sigma_n^{-1} + \tilde{S}_{n-1}^{-1} \right)^{-1} = \mathbb{H}_{n-1} + \frac{n}{n+1} \boldsymbol{v}_J \boldsymbol{v}_J^T - \mathbb{H}_{n-1} \left(\mathbb{H}_{n-1} + \frac{1}{\varepsilon} \operatorname{diag}(\pi_n) \right)^{-1} \mathbb{H}_{n-1}.$$

Therefore, combining the above equality with (A.7), one obtains that

$$\begin{aligned} \tilde{S}_{n-1}^{-1} \left(\Sigma_n^{-1} + \tilde{S}_{n-1}^{-1} \right)^{-1} \tilde{S}_{n-1}^{-1} = \mathbb{H}_{n-1}^- &+ \frac{1}{n(n+1)} \boldsymbol{v}_J \boldsymbol{v}_J^T \\ &- \mathbb{H}_{n-1}^- \mathbb{H}_{n-1} \left(\mathbb{H}_{n-1} + \frac{1}{\varepsilon} \operatorname{diag}(\pi_n) \right)^{-1} \mathbb{H}_{n-1} \mathbb{H}_{n-1}^-. \end{aligned}$$

Inserting the above equality into (A.9) and using again (A.7), one infers that

$$\begin{aligned}\tilde{S}_n^{-1} &= \tilde{S}_{n-1}^{-1} - \mathbb{H}_{n-1}^- - \frac{1}{n(n+1)}\boldsymbol{v}_J\boldsymbol{v}_J^T + \mathbb{H}_{n-1}^-\mathbb{H}_{n-1}\Big(\mathbb{H}_{n-1} + \frac{1}{\varepsilon}\operatorname{diag}(\pi_n)\Big)^{-1}\mathbb{H}_{n-1}\mathbb{H}_{n-1}^-,\\ &= \frac{1}{n+1}\boldsymbol{v}_J\boldsymbol{v}_J^T + \mathbb{H}_{n-1}^-\mathbb{H}_{n-1}\Big(\mathbb{H}_{n-1} + \frac{1}{\varepsilon}\operatorname{diag}(\pi_n)\Big)^{-1}\mathbb{H}_{n-1}\mathbb{H}_{n-1}^-,\\ &= \frac{1}{n+1}\boldsymbol{v}_J\boldsymbol{v}_J^T + P_J\Big(\mathbb{H}_{n-1} + \frac{1}{\varepsilon}\operatorname{diag}(\pi_n)\Big)^{-1}P_J,\end{aligned}$$

by noticing that $\mathbb{H}_{n-1}^-\mathbb{H}_{n-1} = \mathbb{H}_{n-1}\mathbb{H}_{n-1}^- = P_J$. Herafter, we immediately deduce (A.4) from the above identity together with (A.7). Finally, (A.5) follows from an application of a generalization of the SMW formula [17] to the setting of the Moore-Penrose inverse to handle the situation where the matrix A in equation (A.3) is not invertible, which completes the proof of Proposition A.1. □

B Proofs of auxiliary results related to the KL inequality

Proof of Proposition 6.1. The proof consists in a study of $\widetilde{H}_\varepsilon(u)$ when a vector $u \in \langle \boldsymbol{v}_J\rangle^\perp$ is either near u^* or such that $\|u\| \longrightarrow +\infty$. First, we observe that $u \mapsto \|\nabla\widetilde{H}_\varepsilon(u)\|^2 + \frac{\|\nabla\widetilde{H}_\varepsilon(u)\|^2}{\widetilde{H}_\varepsilon(u)}$ is a continuous function except at u^*. Then, since $\widetilde{H}_\varepsilon(u^*) = 0$, a local approximation of $\widetilde{H}_\varepsilon$ using a Taylor expansion shows that, for all $h \in \langle \boldsymbol{v}_J\rangle^\perp$,

$$\widetilde{H}_\varepsilon(u^* + h) = \frac{1}{2}h^T\nabla^2\widetilde{H}_\varepsilon(u^*)h + o(\|h\|^2),$$

and

$$\nabla\widetilde{H}_\varepsilon(u^* + h) = \nabla^2\widetilde{H}_\varepsilon(u^*)h + o(\|h\|).$$

with $\nabla^2\widetilde{H}_\varepsilon(u^*) = G_*^{-1/2}\nabla^2 H_\varepsilon(v^*)G_*^{-1/2}$. Since the matrices $G_*^{-1/2}$ and $\nabla^2 H_\varepsilon(v^*)$ are of rank $J-1$ with all eigenvectors corresponding to positive eigenvalues that belong to $\langle \boldsymbol{v}_J\rangle^\perp$, the Courant-Fischer minmax Theorem yields:

$$\begin{aligned}0 < \lambda_{\min}(\nabla^2\widetilde{H}_\varepsilon(u^*)) &= \liminf_{u\longrightarrow u^*, u\in\langle \boldsymbol{v}_J\rangle^\perp} \frac{\|\nabla\widetilde{H}_\varepsilon(u)\|^2}{\widetilde{H}_\varepsilon(u)}\\ &\leqslant \limsup_{u\longrightarrow u^*, u\in\langle \boldsymbol{v}_J\rangle^\perp} \frac{\|\nabla\widetilde{H}_\varepsilon(u)\|^2}{\widetilde{H}_\varepsilon(u)} = \lambda_{\max}(\nabla^2\widetilde{H}_\varepsilon(u^*)),\end{aligned} \tag{B.1}$$

where $\lambda_{\min}(\nabla^2\tilde{H}_\varepsilon(u^*))$ denotes the second smallest eigenvalue of the matrix $\nabla^2\tilde{H}_\varepsilon(u^*)$, and the notation $u \longrightarrow u^*$ corresponds to the convergence of $u \in \langle \boldsymbol{v}_J\rangle^\perp$. Inequality (B.1) thus implies that the continuous function $u \mapsto \|\nabla\widetilde{H}_\varepsilon(u)\|^2 + \frac{\|\nabla\widetilde{H}_\varepsilon(u)\|^2}{\widetilde{H}_\varepsilon(u)}$ is upper and lower bounded by positive constants in a neighborhood of u^*. Finally, we observe that $\|\nabla\widetilde{H}_\varepsilon\|$ is bounded thanks to inequality (5.7), and it can be checked that the function $\widetilde{H}_\varepsilon$ is coercive (over the finite dimensional vector space $\langle \boldsymbol{v}_J\rangle^\perp$) since it has a unique minimizer at u^*. These facts together with the boundedness of $u \mapsto \|\nabla\widetilde{H}_\varepsilon(u)\|^2 + \frac{\|\nabla\widetilde{H}_\varepsilon(u)\|^2}{\widetilde{H}_\varepsilon(u)}$ in a neighborhood of u^* of implies that inequality (6.43) holds.

Now, let us show that the constant m appearing in inequality (6.43) can be made more explicit thanks to Lemma 5.1. Indeed, since $\nabla \widetilde{H}_\varepsilon(u) = G_*^{-1/2} \nabla H_\varepsilon(G_*^{-1/2} u)$, we immediately obtain from inequality (5.14) that

$$\begin{aligned} \langle \nabla \widetilde{H}_\varepsilon(u), u - u^* \rangle &\geqslant \frac{1 - \exp(-\delta(u))}{\delta(u)} (u - u^*)^T G_*^{-1/2} \nabla^2 H_\varepsilon(v^*) G_*^{-1/2} (u - u^*), \\ &\geqslant \frac{1 - \exp(-\delta(u))}{\delta(u)} \|u - u^*\|^2, \end{aligned} \tag{B.2}$$

where $\delta(u) = \frac{\sqrt{2}}{\varepsilon} \lambda_{\max}(G_*^{-1/2}) \|u - u^*\|$ and the second inequality above follows from the fact that

$$\lambda_{\min}^{\langle \boldsymbol{v}_J \rangle^\perp} \left(G_*^{-1/2} \nabla^2 H_\varepsilon(v^*) G_*^{-1/2} \right) \geqslant 1,$$

by inequality (3.4). Note that inequality (B.2) corresponds to a local strong convex property of the function $\widetilde{H}_\varepsilon$ in the neighborhood of u^*. Then, by the Cauchy-Schwarz inequality, one has that

$$\langle \nabla \widetilde{H}_\varepsilon(u), u - u^* \rangle \leqslant \|\nabla \widetilde{H}_\varepsilon(u)\| \|u - u^*\|.$$

Thus, we obtain by inequality (B.2) that, for any $u \in \langle \boldsymbol{v}_J \rangle^\perp$,

$$\|\nabla \widetilde{H}_\varepsilon(u)\| \geqslant \frac{1 - \exp(-\delta(u))}{\delta(u)} \|u - u^*\|. \tag{B.3}$$

We then consider two cases. If $\|u - u^*\| \leqslant \frac{\varepsilon}{\lambda_{\max}(G_*^{-1/2})}$, then, using that the function $\delta \mapsto \frac{1}{\delta}(1 - \exp(-\delta))$ is decreasing, it follows from (6.42) and (B.3) that,

$$\|\nabla \widetilde{H}_\varepsilon(u)\|^2 + \frac{\|\nabla \widetilde{H}_\varepsilon(u)\|^2}{\widetilde{H}_\varepsilon(u)} \geqslant \frac{\|\nabla \widetilde{H}_\varepsilon(u)\|^2}{\widetilde{H}_\varepsilon(u)} \geqslant \frac{2\varepsilon}{\lambda_{\max}(G_*^-)} \left(1 - \exp\left(-\sqrt{2}\right)\right)^2. \tag{B.4}$$

To the contrary, if $\|u - u^*\| \geqslant \frac{\varepsilon}{\lambda_{\max}(G_*^{-1/2})}$, then

$$\|\nabla \widetilde{H}_\varepsilon(u)\|^2 + \frac{\|\nabla \widetilde{H}_\varepsilon(u)\|^2}{\widetilde{H}_\varepsilon(u)} \geqslant \|\nabla \widetilde{H}_\varepsilon(u)\|^2 \geqslant \frac{\varepsilon^2}{2\lambda_{\max}(G_*^-)} \left(1 - \exp\left(-\sqrt{2}\right)\right)^2. \tag{B.5}$$

using the inequality $1 - \exp\left(-\delta(u)\right) \geqslant 1 - \exp\left(-\sqrt{2}\right)$ that holds for $\delta(u) \geqslant \sqrt{2}$. Consequently, since $\left(1 - \exp\left(-\sqrt{2}\right)\right)^2 > 1/2$ and combining inequalities (B.4) and (B.5), it follows that the constant m appearing in inequality (6.43) can be chosen as $m = m_\varepsilon$ with m_ε defined by (6.44). This concludes the proof of Proposition 6.1. □

We then show the proof of the one-step evolution of the SGN algorithm.

Proof of Proposition 6.3 . First, by using the arguments from the proof of (i) of Theorem 3.1, we remark that the matrix S_n defined by (2.9) satisfies:

$$\lambda_{\min}(S_n) \geqslant 1 + \gamma \min(\nu) \left(\sum_{m=1}^{p_n} m^{-\beta} \right) \quad \text{and} \quad \lambda_{\max}(S_n) \leqslant 1 + 4n + \gamma \max(\nu) \sum_{m=1}^{p_n+1} m^{-\beta},$$

where $p_n \geqslant 1$ denotes the largest integer such that $p_n J \leqslant n$. Consequently, using the fact that $\frac{1}{1-\beta}\big(p_n^{1-\beta}-1\big) \leqslant \sum_{m=1}^{p_n} m^{-\beta} \leqslant \frac{1}{1-\beta}p_n^{1-\beta}$ the above inequalities imply that:

$$\begin{aligned}\lambda_{\max}(S_n^{-1}) &\leqslant \frac{1-\beta}{1-\beta+\gamma\min(\nu)\Big((n_J-1)^{1-\beta}-1\Big)} \\ &\leqslant \frac{1}{1-2\gamma\min(\nu)+\gamma\min(\nu)(n_J-1)^{1-\beta}},\end{aligned} \tag{B.6}$$

where $n_J = n/J$, and

$$\lambda_{\min}(S_n^{-1}) \geqslant \frac{1}{1-\beta+4(1-\beta)n+\gamma\max(\nu)\,(n/J)^{1-\beta}} \geqslant \frac{1}{1+(4+2\gamma\max(\nu))n}. \tag{B.7}$$

Step 1: Taylor expansion. First, we introduce the notation $\widetilde{U}_n = G_*^{1/2}\widehat{V}_n$, and in the proof we repeatedly the property that the eigenvectors of G_* associated to non-zero eigenvalues belong to $\langle \boldsymbol{v}_J\rangle^{\perp}$. Using equation (2.6) and the fact that $P_J G_* = G_*$, a second order Taylor expansion yields

$$\begin{aligned}\Phi(\widetilde{U}_{n+1}) &= \Phi\Big(\widetilde{U}_n - n^{\alpha}G_*^{1/2}S_n^{-1}\nabla_v h_\varepsilon(X_{n+1},\widehat{V}_n)\Big) \\ &= \Phi(\widetilde{U}_n) - n^{\alpha}\Big\langle \nabla\Phi(\widetilde{U}_n), G_*^{1/2}S_n^{-1}\nabla_v h_\varepsilon(X_{n+1},\widehat{V}_n)\Big\rangle \\ &+ \frac{n^{2\alpha}}{2}\nabla^2\Phi(\xi_{n+1})\Big(G_*^{1/2}S_n^{-1}\nabla_v h_\varepsilon(X_{n+1},\widehat{V}_n)\Big)^{\otimes 2},\end{aligned} \tag{B.8}$$

where ξ_{n+1} is such that $\xi_{n+1} = \widetilde{U}_n + t_{n+1}(\widetilde{U}_{n+1}-\widetilde{U}_n)$ with $t_{n+1}\in(0,1)$. Now, applying inequalities (5.6) and (6.48), the second order term in equation (B.8) can be bounded as follows:

$$\begin{aligned}\|\nabla^2\Phi(\xi_{n+1})\Big(G_*^{1/2}S_n^{-1}\nabla_v h_\varepsilon(X_{n+1},\widehat{V}_n)\Big)^{\otimes 2}\| &\leqslant \delta_\varepsilon\lambda_{\max}(G_*^-)(1+\Phi(\xi_{n+1}))\times \\ &\qquad \Big\|G_*^{1/2}S_n^{-1}\ \nabla_v h_\varepsilon(X_{n+1},\widehat{V}_n)\Big\|^2 \\ &\leqslant 4\delta_\varepsilon\lambda_{\max}(G_*^-)\lambda_{\max}^2(G_*^{1/2}S_n^{-1})(1+\Phi(\xi_{n+1})),\end{aligned}$$

which yields the inequality:

$$\begin{aligned}\Phi(\widetilde{U}_{n+1}) &\leqslant \Phi(\widetilde{U}_n) - n^{\alpha}\Big\langle \nabla\Phi(\widetilde{U}_n), G_*^{1/2}S_n^{-1}\nabla_v h_\varepsilon(X_{n+1},\widehat{V}_n)\Big\rangle \\ &\quad +2\delta_\varepsilon\frac{\lambda_{\max}(G_*)}{\lambda_{\min}(G_*)}n^{2\alpha}\lambda_{\max}^2(S_n^{-1})(1+\Phi(\xi_{n+1})).\end{aligned} \tag{B.9}$$

Step 2: An auxiliary inequality. We now establish a technical bound to relate $\Phi(\xi_{n+1})$ and $\Phi(\widetilde{U}_n)$. To this end, by a first order Taylor expansion of the function $s\mapsto \widetilde{H}_\varepsilon(\widetilde{U}_n + s(\widetilde{U}_{n+1}-\widetilde{U}_n))$ where $\widetilde{U}_n^s = \widetilde{U}_n + s(\widetilde{U}_{n+1}-\widetilde{U}_n)$, one has that:

$$\widetilde{H}_\varepsilon(\widetilde{U}_{n+1}) = \widetilde{H}_\varepsilon(\widetilde{U}_n) + \int_0^1 \nabla\widetilde{H}_\varepsilon(\widetilde{U}_n^s)(\widetilde{U}_{n+1}-\widetilde{U}_n)\,ds,$$

and thus, combining the Cauchy-Schwarz inequality with the upper bounds (5.6) and (6.40), we obtain that:

$$\begin{aligned}\widetilde{H}_\varepsilon(\xi_n) &\leqslant \widetilde{H}_\varepsilon(\widetilde{U}_n) + \sup_{t\in[0,1]} \|\nabla \widetilde{H}_\varepsilon(\widetilde{U}_n^T)\| \|\widetilde{U}_{n+1} - \widetilde{U}_n\| \\ &\leqslant \widetilde{H}_\varepsilon(\widetilde{U}_n) + \sup_{t\in[0,1]} \|\nabla \widetilde{H}_\varepsilon(\widetilde{U}_n^T)\| \|n^\alpha G_*^{1/2} S_n^{-1} \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)\| \\ &\leqslant \widetilde{H}_\varepsilon(\widetilde{U}_n) + 4\lambda_{\max}(G_*^{-1/2})\lambda_{\max}(G_*^{1/2}) n^\alpha \lambda_{\max}(S_n^{-1}). \end{aligned} \tag{B.10}$$

Note that, under the condition $\alpha + \beta < 1/2$, it follows from inequality (B.6) that $n^\alpha \lambda_{\max}(S_n^{-1}) \leqslant c_0$ for some constant $c_0 \geqslant 1$ for all $n \geqslant J$. Hence, inserting the upper bound (B.10) into the definition of Φ and using inequality (6.46), we obtain that

$$\begin{aligned}\Phi(\xi_{n+1}) &\leqslant \left(\widetilde{H}_\varepsilon(\widetilde{U}_n) + 4\frac{\lambda_{\max}(G_*^{1/2})}{\lambda_{\min}(G_*^{1/2})} n^\alpha \lambda_{\max}(S_n^{-1})\right) \times \\ &\quad \exp\left(\widetilde{H}_\varepsilon(\widetilde{U}_n) + 4\frac{\lambda_{\max}(G_*^{1/2})}{\lambda_{\min}(G_*^{1/2})} n^\alpha \lambda_{\max}(S_n^{-1})\right) \\ &\leqslant \exp\left(4\frac{\lambda_{\max}(G_*^{1/2})}{\lambda_{\min}(G_*^{1/2})} n^\alpha \lambda_{\max}(S_n^{-1})\right) \times \\ &\quad \left(\widetilde{H}_\varepsilon(\widetilde{U}_n) + 4\frac{\lambda_{\max}(G_*^{1/2})}{\lambda_{\min}(G_*^{1/2})} n^\alpha \lambda_{\max}(S_n^{-1})\right) \exp\left(\widetilde{H}_\varepsilon(\widetilde{U}_n)\right) \\ &\leqslant \exp\left(4c_0\frac{\lambda_{\max}(G_*^{1/2})}{\lambda_{\min}(G_*^{1/2})}\right) \left(\Phi(\widetilde{U}_n) + 4c_0\frac{\lambda_{\max}(G_*^{1/2})}{\lambda_{\min}(G_*^{1/2})}(1+\Phi(\widetilde{U}_n))\right) \\ &\leqslant \tilde{c}_0\left(1+\Phi(\widetilde{U}_n)\right), \end{aligned} \tag{B.11}$$

where $\tilde{c}_0 = \max\left(1, 4c_0\frac{\lambda_{\max}(G_*^{1/2})}{\lambda_{\min}(G_*^{1/2})}\right)\exp\left(4c_0\frac{\lambda_{\max}(G_*^{1/2})}{\lambda_{\min}(G_*^{1/2})}\right)$.

Step 3: Derivation of a recursive inequality. Inserting inequality (B.11) into (B.9), and taking the conditional expectation with respect to $\mathcal{F}_n$, we obtain that:

$$\begin{aligned}\mathbb{E}\left[\Phi(\widetilde{U}_{n+1})\,|\mathcal{F}_n\right] &\leqslant \Phi(\widetilde{U}_n) - n^\alpha \left\langle \nabla\Phi(\widetilde{U}_n), G_*^{1/2} S_n^{-1} G_*^{1/2} \nabla \widetilde{H}_\varepsilon(\widetilde{U}_n)\right\rangle \\ &\quad + 2\delta_\varepsilon \frac{\lambda_{\max}(G_*)}{\lambda_{\min}(G_*)} n^{2\alpha}\lambda_{\max}^2(S_n^{-1})\left(1 + \tilde{c}_0\ \left(1+\Phi(\widetilde{U}_n)\right)\right), \end{aligned}$$

where we used the property that $\nabla H_\varepsilon(\widehat{V}_n) = G_*^{1/2}\nabla\widetilde{H}_\varepsilon(\widetilde{U}_n)$.

Consequently, using inequality (6.47) and introducing $c_\varepsilon = 2\tilde{c}_0\delta_\varepsilon\frac{\lambda_{\max}(G_*)}{\lambda_{\min}(G_*)}$, we have:

$$\begin{aligned}\mathbb{E}\left[\Phi(\widetilde{U}_{n+1})\,|\mathcal{F}_n\right] &\leqslant \left(1 - m_\varepsilon n^\alpha \lambda_{\min}(G_*)\lambda_{\min}(S_n^{-1}) + c_\varepsilon n^{2\alpha}\lambda_{\max}^2(S_n^{-1})\right)\Phi(\widetilde{U}_n) \\ &\quad + c_\varepsilon n^{2\alpha}\lambda_{\max}^2(S_n^{-1}) \end{aligned} \tag{B.12}$$

Now, thanks to inequalities (B.6) and (B.7), and the condition $0 < \alpha + \beta < 1/2$, it follows that there exists an integer n_0 such that, for all $n \geqslant n_0$,

$$m_\varepsilon n^\alpha \lambda_{\min}(G_*)\lambda_{\min}(S_n^{-1}) \geq 2c_\varepsilon n^{2\alpha}\lambda_{\max}^2(S_n^{-1}). \tag{B.13}$$

Hence, by combining inequalities (B.12) and (B.13) we obtain inequality (6.50) which concludes the proof of i). □

Acknowledgments

The authors gratefully acknowledge financial support from the Agence Nationale de la Recherche (MaSDOL grant ANR-19-CE23-0017). J. Bigot and S. Gadat are members of the Institut Universitaire de France (IUF), and part of this work has been carried out with financial support from the IUF.

References

[1] Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In Advances in Neural Information Processing Systems 30. 2017, pp. 1964–1974.

[2] Bach, F. R. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. Journal of Machine Learning Research 15, 1 (2014), 595–627.

[3] Benaïm, M., and Hirsch, M. Asymptotic pseudotrajectories and chain recurrent flows, with applications. Journal of Dynamics and Differential Equations 8, 1 (1996), 141–176.

[4] Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative Bregman projections for regularized transportation problems. SIAM Journal on Scientific Computing 37, 2 (2015), A1111–A1138.

[5] Bercu, B., and Bigot, J. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. Annals of Statistics 49, 2 (2021), 968–987.

[6] Bercu, B., Godichon, A., and Portier, B. An efficient stochastic newton algorithm for parameter estimation in logistic regressions. SIAM Journal on Control and Optimization 58, 1 (2020), 348–367.

[7] Bigot, J., Cazelles, E., and Papadakis, N. Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration. Information and Inference 8, 4 (2019), 719–755.

[8] Bigot, J., Gouet, R., Klein, T., López, A., et al. Geodesic PCA in the Wasserstein space by convex PCA. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 53, 1 (2017), 1–26.

[9] Bigot, Jérémie. Statistical data analysis in the wasserstein space. ESAIM: ProcS 68 (2020), 1–19.

[10] Bolte, J., Daniilidis, A., and Lewis, A. The ł ojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. 17, 4 (2006), 1205–1223.

[11] Bolte, J., Nguyen, P., Peypouquet, J., and Suter, B. W. From error bounds to the complexity of first-order descent methods for convex functions. Math. Program. (A), 165 (2017), 471–507.

[12] Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and radon Wasserstein barycenters of measures. Journal of Mathematical Imaging and Vision 51, 1 (2015), 22–45.

[13] Cazelles, E., Seguy, V., Bigot, J., Cuturi, M., and Papadakis, N. Log-PCA versus Geodesic PCA of histograms in the Wasserstein space. SIAM Journal on Scientific Computing 40, 2 (2018), B429–B456.

[14] Cénac, P., Godichon-Baggioni, A., and Portier, B. An efficient averaged stochastic gauss-newton algorithm for estimating parameters of non linear regressions models. Preprint - arXiv2006.12920, 2020.

[15] Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2292–2300.

[16] Cuturi, M., and Peyré, G. Semidual regularized optimal transport. SIAM Review 60, 4 (2018), 941–965.

[17] Deng, C. Y. A generalization of the Sherman-Morrison-Woodbury formula. Applied Mathematics Letters 24, 9 (2011), 1561 – 1564.

[18] Duflo, M. Random iterative models, vol. 34 of Applications of Mathematics, New York. Springer-Verlag, Berlin, 1997.

[19] Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. Regularized discrete optimal transport. SIAM Journal on Imaging Sciences 7, 3 (2014), 1853–1882.

[20] Flamary, R., Cuturi, M., Courty, N., and Rakotomamonjy, A. Wasserstein discriminant analysis. Machine Learning 107, 12 (Dec 2018), 1923–1945.

[21] Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., and Poggio, T. Learning with a wasserstein loss. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (Cambridge, MA, USA, 2015), NIPS'15, MIT Press, pp. 2053–2061.

[22] Gadat, S., and Panloup, F. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. Preprint - arXiv:1709.03342, May 2022.

[23] Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In Advances in Neural Information Processing Systems 29. 2016, pp. 3440–3448.

[24] Genevay, A., Peyre, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018), A. Storkey and F. Perez-Cruz, Eds., vol. 84 of Proceedings of Machine Learning Research, PMLR, pp. 1608–1617.

[25] Gordaliza, P., Del Barrio, E., Gamboa, F., and Loubes, J.-M. Obtaining fairness using optimal transport theory. In Proceedings of the 36th International Conference on Machine Learning (2019), pp. 2357–2365.

[26] Gramfort, A., Peyré, G., and Cuturi, M. Fast optimal transport averaging of neuroimaging data. In International Conference on Information Processing in Medical Imaging (2015), Springer, pp. 261–272.

[27] Hager, W. W. Updating the inverse of a matrix. SIAM Review 31, 2 (1989), 221–239.

[28] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), Y. Bengio and Y. LeCun, Eds.

[29] Kitagawa, J., Mérigot, Q., and B., T. Convergence of a newton algorithm for semi-discrete optimal transport. Journal of the European Math Society 21, 9 (2019), 2603–2651.

[30] Klatt, M., Tameling, C., and Munk, A. Empirical regularized optimal transport: Statistical theory and applications. SIAM J. Math. Data Sci. 2, 2 (2020), 419–443.

[31] Kurdyka, K. On gradients of functions definable in o-minimal structures. Ann. Inst. Fourier (Grenoble) 48, 3 (1998), 769–783.

[32] Lojasiewicz, S. Une propriété topologique des sous-ensembles analytiques réels. Editions du centre National de la Recherche Scientifique, Paris, Les Équations aux Dérivées Partielles (1963), 87–89.

[33] Mérigot, Q. A multiscale approach to optimal transport. Computer Graphics Forum 30, 5 (2011), 1583–1592.

[34] Mérigot, Q., Meyron, J., and Thibert, B. An algorithm for optimal transport between a simplex soup and a point cloud. SIAM Journal on Imaging Sciences 11, 2 (2018), 1363–1389.

[35] Panaretos, V. M., and Zemel, Y. Amplitude and phase variation of point processes. Annals of Statistics 44, 2 (2016), 771–812.

[36] Panaretos, V. M., and Zemel, Y. Statistical aspects of wasserstein distances. Annual Reviews of Statistics and its Applications 6 (2018), 405–431.

[37] Pelletier, M. Asymptotic almost sure efficiency of averaged stochastic algorithms. SIAM J. Control and Optimization 39 (08 2000), 49–72.

[38] Peyré, G., and Cuturi, M. Computational optimal transport. Foundations and Trends in Machine Learning 11, 5-6 (2019), 355–607.

[39] Rabin, J., and Papadakis, N. Convex color image segmentation with optimal transport distances. In International Conference on Scale Space and Variational Methods in Computer Vision (2015), Springer, pp. 256–269.

[40] Rigollet, P., and Weed, J. Entropic optimal transport is maximum-likelihood deconvolution. Comptes Rendus Mathematique 356, 11 (2018), 1228 – 1235.

[41] Robbins, H., and Siegmund, D. A convergence theorem for non negative almost supermartingales and some applications. In Optimizing methods in statistics. Elsevier, 1971, pp. 233–257.

[42] Rolet, A., Cuturi, M., and Peyré, G. Fast dictionary learning with a smoothed Wasserstein loss. In Proc. International Conference on Artificial Intelligence and Statistics (AISTATS) (2016).

[43] Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. On the convergence and robustness of training gans with regularized optimal transport. In Advances in Neural Information Processing Systems (2018), S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc.

[44] Seguy, V., and Cuturi, M. Principal geodesic analysis for probability measures under the optimal transport metric. In Advances in Neural Information Processing Systems 28. 2015, pp. 3294–3302.

[45] Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. ACM Trans. Graph. 34, 4 (2015), 66:1–66:11.

[46] Sommerfeld, M., and Munk, A. Inference for empirical Wasserstein distances on finite spaces. Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2016).

[47] Steerneman, A., and van Perlo -ten Kleij, F. Properties of the matrix a-xy. Linear Algebra and Its Applications 410 (2005), 70–86.

[48] Villani, C. Optimal transport: old and new, vol. 338. Springer Science & Business Media, 2008.

[49] Zemel, Y., and Panaretos, V. M. Fréchet means and procrustes analysis in wasserstein space. Bernoulli 25, 2 (05 2019), 932–976.

[50] Zhang, L.-X. Central limit theorems of a recursive stochastic algorithm with applications to adaptive designs. Ann. Appl. Probab. 26, 6 (2016), 3630–3658.