



HAL
open science

Weak transport inequalities and applications to exponential inequalities and oracle inequalities

Olivier Wintenberger

► **To cite this version:**

Olivier Wintenberger. Weak transport inequalities and applications to exponential inequalities and oracle inequalities. 2012. hal-00719729v1

HAL Id: hal-00719729

<https://hal.science/hal-00719729v1>

Preprint submitted on 20 Jul 2012 (v1), last revised 5 Mar 2014 (v3)

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

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

WEAK TRANSPORT INEQUALITIES AND APPLICATIONS TO EXPONENTIAL INEQUALITIES AND ORACLE INEQUALITIES

OLIVIER WINTENBERGER

ABSTRACT. We introduce weak transport costs that are weakened forms of the transport costs defined by Marton in [26]. We obtain new weak transport inequalities for non products measures similar than those obtained by Samson in [32] but valid also for other metrics than the Hamming distance. Many examples are provided to show that the euclidian norm is an appropriate metric for many classical time series. The dual form of the weak transport inequalities yield new exponential inequalities and extensions to the dependent case of the classical result of Talagrand [33] for convex functions that are Lipschitz continuous. Expressing the concentration properties of the ordinary least square estimator as a conditional mass transport problem, we derive from the weak transport inequalities new oracle inequalities with fast rates of convergence.

CONTENTS

1. Introduction	2
2. Weak transport costs, gluing lemma and Markov couplings	4
2.1. Weak transport costs on E	4
2.2. Markov couplings	6
2.3. Weak transport costs on E^n , $n \geq 2$	7
3. Weak transport inequalities	8
3.1. Weak transport inequalities	8
3.2. Weak transport inequalities on E	8
3.3. Weak transport inequalities on E^n , $n \geq 2$	9
4. Examples of $\gamma(p)$ -weakly dependent processes	13
4.1. Linear models	14
4.2. Non-linear models	14
5. New exponential inequalities	15
5.1. General exponential inequalities	15
5.2. Extensions of classical concentration inequalities to dependent cases	17
5.3. The specific case of the Hamming distance	18
6. Applications to oracle inequalities with fast convergence rates	19
6.1. The statistical setting	19
6.2. Nonexact oracle inequality for $\gamma(2)$ -weakly-dependent sequences	19
6.3. Exact oracle inequality for $\tilde{\gamma}(2)$ -weakly-dependent sequences	22
References	24

2010 *Mathematics Subject Classification*. Primary 60E15; Secondary 28A35, 62J05, 62M10, 62M20.

Key words and phrases. transport inequalities, concentration of measures, weakly dependent time series, oracle inequalities, ordinary least square estimator, time series prediction.

1. INTRODUCTION

Since the seminal work of Marton [24], transport inequalities efficiently yield dimension free concentration inequalities. Using a duality argument, Bobkov and Gotze [6] even proved that transport inequalities are equivalent to some concentration inequalities. Our references on the subject are the monograph of Villani [36] and the survey of Gozlan and Leonard [16]. Transport inequalities appear as a nice alternative to the classical modified log-Sobolev approach of Massart [28] for obtaining dimension free concentration inequalities useful in mathematical statistics. More specifically, dimension free concentration inequalities are used to get oracle inequalities with fast rates of convergence. This article develop new kinds of transport inequalities, new exponential inequalities and new oracle inequalities with fast rates of convergence.

In the case of product measures, the classical modified log-Sobolev approach developed by Massart in [28] leads to optimal dimension free concentration inequalities of Bernstein's type. However, for non product measures, such inequalities do not hold in their optimal form in many situations. The reason is the following: in the bounded iid case, Bernstein's inequality yields gaussian behavior for deviations less than a bound depending on the essential supremum. In many bounded dependent cases, there exists a unique regeneration scheme of iid cycles with random length. The Bernstein inequality yields gaussian behavior for small deviations less than a bound depending on the essential supremum and also on the concentration properties of the random length, see Bertail and Cléménçon [5]. It is a drawback for statistical applications where the variance term, which is essential, is perturbed by the concentration properties of the random length. It leads, to an additional term, at least logarithmic, which cannot be removed, see Adamcsak [1]. To bypass this problem, many authors assumed contraction conditions on the conditional measure, see Marton [25] for the total variation metric, Lezaud [23] under a spectral gap condition for the kernel of a Markov chain. For symmetric Markov process, this second condition is more general and it is also necessary for Bernstein's inequality, see Guillin *et al.* [17].

Many classical models in time series analysis do not satisfy such conditions. Fortunately, the classical Bernstein's inequality also holds for non contraction conditions but under $\tilde{\gamma}$ -weakly dependent conditions, closely related with uniform mixing conditions, see Samson [32]. This result yields fast convergence rates of order n^{-1} in oracle inequalities (comparable to those in the iid case) in a dependent setting, see [2]. However, this approach relies on the maximal coupling properties of the Hamming distance and cannot be extended to other metrics, see [11]. For other metrics, non optimal couplings are used by Marton [27] and Djellout *et al.* [12] to extend classical dimension free transport inequalities $T_2(C)$ in a dependent context. If the "constant" C in the transport inequality is sufficiently close to the variance term then Bernstein's inequality is recovered and fast convergence rates are achieved, see Joulin and Ollivier [19]. Otherwise, the statistical convergence rates are lower than n^{-1} because a tradeoff must be done between the estimate of the variance and the accuracy of coupling schemes that are not dimension free, see Wintenberger [37] for details. The fast rates of convergence in mathematical statistics are not achieved in general dependent contexts due to the variance term appearing in the classical dimension free inequalities of Bernstein's types. On the contrary, the Hoeffding's inequality that do not have a variance term is easily extended to very general dependent case, see Rio [30] and Djellout *et al.* [12]. Unfortunately, the Hoeffding inequality, equivalent to the $T_1(C)$ transport inequality, is not dimension

free. Thus, this probabilistic inequality yields low rates of convergence of order $n^{-1/2}$, see Alquier and Wintenberger [3].

In this paper we develop new probabilistic tools to obtain dimension free exponential inequalities and thus fast convergence rates in oracle inequalities. Let (E, d) be a Polish space. With the notation $P[h] = \int h dP$ for any probability measure P and any measurable function h , we say that P satisfies the new transport inequality $\tilde{T}_p(C)$ for any $C > 0$ and $1 \leq p \leq q$ if for any measure Q

$$\sup_{\alpha} \inf_{\pi} \frac{\pi[\alpha(Y)d(X, Y)]}{(Q[\alpha(Y)^q])^{1/q}} \leq \sqrt{2C\mathcal{K}(P|Q)}$$

with $1/p + 1/q = 1$ and the convention $+\infty / +\infty = 0/0 = 0$. Here α is any non-negative measurable function, π is any coupling scheme of (X, Y) with margins (P, Q) and $\mathcal{K}(P|Q)$ is the relative entropy $Q[\log(dP/dQ)]$ (also called the Kullback-Leibler divergence). As the role of P and Q is not the same, we also introduce $\tilde{T}_p^{(i)}(C)$ where P and Q are interchanged in the left hand side term. These inequalities are weakened versions of the transport inequalities introduced by Marton [26] (for d being the Hamming distance)

$$\inf_{\pi \in \tilde{\mathcal{M}}} \sup_{\alpha > 0} \frac{\pi[\alpha(Y)d(X, Y)]}{(Q[\alpha(Y)^q])^{1/q}} \leq \sqrt{2C\mathcal{K}(Q|P)}.$$

These inequalities are already weakened forms of the classical $T_p(C)$ transport inequality

$$\inf_{\pi \in \mathcal{M}} \pi[d^p(X, Y)]^{1/p} \leq \sqrt{2C\mathcal{K}(Q|P)}.$$

Contrary to the classical $T_p(C)$ transport inequalities, any compactly supported measure P satisfies the weak $\tilde{T}_p(C)$ transport inequalities for any $1 \leq p \leq 2$. Moreover, the weak transport inequalities extend nicely to non-products non-contractive measures P on E^n , $n \geq 1$. Using a new Markov coupling scheme, our main result in Theorem 3.2 states that there exists $C' > 0$ such that

$$(1.1) \quad \sup_{\alpha} \inf_{\pi} \frac{\sum_{j=1}^n \pi[\alpha_j(Y)d(X_j, Y_j)]}{(\sum_{j=1}^n Q[\alpha_j(Y)^q])^{1/q}} \leq \sqrt{2n^{2/p-1}C'\mathcal{K}(P|Q)}$$

when conditional measures $P_{x_i|x^{(i)}}$, $x^{(i)} = (x_i, \dots, x_0)$ satisfy the weak transport inequalities $\tilde{T}_p(C)$ and under a new $\gamma(p)$ -weak dependent condition:

$$W_p(P_{x_k|x^{(i)}}, P_{x_k|x^{(i-1)}, y_i}) \leq \gamma_{k,i}(p) d(x_i, y_i), \quad 0 \leq i < k \leq n.$$

When d is the Hamming distance, the $\gamma(2)$ -weak dependance coincides with the context of weakly dependence already studied by Samson [32] and we obtain similar results. We keep the notation and denote $\tilde{\gamma}(p)$ the weak dependence coefficients when d is the Hamming distance. However, to tackle much more general and classical time series contexts, we prefer to choose d as the euclidian norm, see Section 4. Then, when $p = 1$ and $\tilde{T}_1(C) = \tilde{T}_1^{(i)}(C) = T_1(C)$ by definition, the $\gamma(1)$ -weak dependence is linked with the weak dependence notion introduced by Rio in [30] as discussed in Djellout *et al.* [12]. Thus we recover the Hoeffding's inequality of [30] which is not dimension free because $n^{2/p-1} = n$ as $p = 1$.

The dual forms of the weak transport inequalities yield new exponential inequalities. Except in the specific case of the Hamming distance, the deviations are not estimated in terms of the variance and contrary to the Hoeffding's inequality, it is a dimension free inequality when $p = 2$. If P satisfies $\tilde{T}_2(C)$ on E^n then for any

function f of the observations (X_1, \dots, X_n) such that there exist functions $L_j(x)$ satisfying $f(x) - f(y) \leq \sum_{j=1}^n L_j(x)d(x_j, y_j)$ for any $x, y \in (\mathbb{R}^d)^n$ we have

$$(1.2) \quad P \left[\exp \left(\lambda(f - P[f]) - \frac{C\lambda^2}{2} \sum_{j=1}^n L_j f^2 \right) \right] \leq \exp(\lambda P[f]), \quad \lambda > 0.$$

When d is the Hamming distance, inequality (1.2) yields to the classical Bernstein's inequality, see Ledoux [22] in the independent setting and Samson [32] in the uniform mixing setting. When the function f is a convex function, it satisfies the above condition with $L_j = \partial_j$ its sub-gradient and the inequality (1.2) coincides with generalizations of the Tsirel'son inequality of [34] (also implied by the T_2 transport inequality, see Bobkov *et al.* [7]). For convex functions that are also Lipschitz continuous the inequality stated above leads to new extensions of the classical exponential inequality due to Talagrand [33] for products measure.

As $n^{2/p-1} = 1$ for $p = 2$, combining inequalities (1.1) and (1.2) we obtain new dimension free exponential inequalities for many dependent classical time series that are $\gamma(2)$ -weakly dependent. As the transport inequalities yield concentration of measures via relative entropy, we couple it with the statistical PAC-bayesian paradigm that describes the accuracy of estimators in term of relative entropy too, see McAllester [29]. The oracle inequalities can thus be expressed as a conditional mass transport problem. We apply this new approach to the Ordinary Least Square (OLS) estimator $\hat{\theta}$ in the linear regression context (other interesting statistical issues will be investigated in the future). Denoting by R the risk of prediction, an oracle inequality states with high probability that $R(\hat{\theta}) \leq (1 + \eta)R(\bar{\theta}) + R_n\eta^{-1}$ where $\eta \geq 0$, $\bar{\theta}$ is the oracle defined as $R(\hat{\theta}) \leq R(\theta)$ for all θ and R_n is the rate of convergence. Oracle inequalities are standard non asymptotic criteria for the efficiency of statistical estimators, see Massart [28]. If $\eta = 0$ then the oracle inequality is said to be exact and otherwise it is non exact, see Lecué and Mendelson [21] for a discussion. The dimension free concentration properties yield to fast rates of convergence $R_n \propto n^{-1}$. For $\gamma(2)$ -weakly dependent time series, we obtain new nonexact oracle inequalities for the OLS $\hat{\theta}$ when the conditional measures satisfies the weak transport inequalities. These assumptions are satisfied for many models such as classical ARMA models with bounded, gaussian or log-concave innovations. In the specific case when d is the Hamming distance, we recover in the conditional mass transport problem the classical variance term as it was the case in exponential inequalities. This variance term plays a crucial role through the so called necessary margins condition introduced by Tsybakov [35]. Thus, fixing d as the Hamming distance, we obtain new exact oracle inequalities with fast convergence rates for the OLS $\hat{\theta}$ in the $\tilde{\gamma}(2)$ -weakly dependent case.

The paper is organized as follow: in Section 2 are developed the properties of the weak transport costs used in the proof of our main result, a weak transport inequalities for non product measures stated in Section 3. Section 4 is devoted to some examples. The dual form of the weak transport inequalities yields new exponential inequalities presented in Section 5. Finally, new oracle inequalities with fast rates of convergence are given in Section 6.

2. WEAK TRANSPORT COSTS, GLUING LEMMA AND MARKOV COUPLINGS

2.1. Weak transport costs on E . Let $M(F)$ denotes the set of probability measure on some space F , $M^+(F)$ the set of lower semi-continuous non negative measurable functions and $\tilde{M}(P, Q)$ the set of coupling measures $\pi_{x,y}$, i.e. $\pi_{x,y} \in M(E^2)$

with margins $\pi_x = P$ and $\pi_y = Q$. Let (p, q) be real numbers satisfying $1 \leq p \leq 2$ and $1/p + 1/q = 1$. Let us define the weak transport cost as

$$(2.1) \quad \tilde{W}_p(P, Q) = \sup_{\alpha \in M^+(E)} \inf_{\pi \in \tilde{M}(P, Q)} \frac{\pi[\alpha(Y)d(X, Y)]}{Q[\alpha^q]^{1/q}}$$

with the classical conventions $Q[\alpha^q]^{1/q} = \text{ess sup } \alpha(Y)$ when $q = \infty$ and $+\infty/+\infty = 0/0 = 0$. For fixed $\alpha \in M^+(F)$, let us denote

$$(2.2) \quad \tilde{W}_\alpha(P, Q) = \inf_{\pi \in \tilde{M}(P, Q)} \pi[\alpha(Y)d(X, Y)].$$

Notice that \tilde{W} is not symmetric and that $\tilde{W}(P, Q) = \tilde{W}(Q, P) = \tilde{W}_\alpha(P, Q) = \tilde{W}_\alpha(Q, P) = 0$ if $P = Q$. Notice that $\alpha \in M^+$ is assumed lower semi-continuous such that the optimal transport in the weak transport costs exist, see for example [16]. Now let us show that the weak transport cost satisfies the triangular inequality. It is a simple consequence of the second assertion of the following version of the gluing Lemma:

Lemma 2.1. *For any coupling $\pi_{x,y} \in \tilde{M}(P, Q)$ and $\pi_{y,z} \in \tilde{M}(Q, R)$ respectively there exists a distribution $\pi_{x,y,z}$ with corresponding margins and such that X and Z are independent conditionally on Y , i.e. $\pi_{x,z|y} = \pi_{x|y}\pi_{z|y}$.*

Proof. From the classical gluing Lemma, see for example the Villani's textbook [36], we can choose $\pi_{x,y,z}$ such that $\pi_{x,y,z} = \pi_{x|y}\pi_{z|y}\pi_y$ as the margins corresponds: $\pi_{x|y}\pi_y = \pi_{x,y}$ and $\pi_{z|y}\pi_y = \pi_{y,z}$. The conditional independence follows from the specific form of $\pi_{x,y,z}$ as $\pi_{x,z|y} = \pi_{x,y,z}/\pi_y$ by definition. \square

The conditional independence in the gluing Lemma 2.1 is the main ingredient to prove the triangular inequality on \tilde{W}_p :

Lemma 2.2. *For any P, Q, R we have*

$$(2.3) \quad \tilde{W}_p(P, R) \leq \tilde{W}_p(P, Q) + \tilde{W}_p(Q, R)$$

Proof. Let us fix $\alpha \in M^+(E)$ such that $R[\alpha^q] < \infty$. We have

$$\pi_{x,z}[\alpha(Z)d(X, Z)] \leq \pi[\alpha(Z)d(X, Y)] + \pi_{y,z}[\alpha(Z)d(Y, Z)].$$

Let us choose $\pi_{y,z}^*$ satisfying

$$\pi_{y,z}^*[\alpha(Z)d(Y, Z)] = \inf_{\pi \in \tilde{M}(Q, R)} \pi[\alpha(Z)d(Y, Z)] \leq R[\alpha^q]^{1/q} \tilde{W}_p(Q, R).$$

By conditional independence in Lemma 2.1, we also have

$$\pi[\alpha(Z)d(X, Y)] = \pi_{x,y}[\pi_{z|y}^*[\alpha(Z)|Y]d(X, Y)] =: \pi_{x,y}[\tilde{\alpha}(Y)d(X, Y)].$$

Let us choose $\pi_{x,y}^*$ satisfying

$$\pi_{x,y}^*[\tilde{\alpha}(Y)d(X, Y)] = \inf_{\pi \in \tilde{M}(P, Q)} \pi[\tilde{\alpha}(Y)d(X, Y)] \leq Q[\tilde{\alpha}^q]^{1/q} \tilde{W}_p(P, Q).$$

Notice that $Q[\tilde{\alpha}^q] = Q[\pi_{z|y}^*[\alpha(Z)|Y]^q] \leq R[\alpha^q]$ using Jensen's inequality. Let us denote $\pi^* = \pi_{x,y,z}^*$ obtained by the gluing Lemma 2.3 of $\pi_{x,y}^*$ and $\pi_{y,z}^*$. Collecting all these bounds we have $\pi^*[\alpha(Z)d(X, Y)] \leq R[\alpha^q] \tilde{W}_p(P, Q)$. We obtain

$$\frac{\pi_{x,z}^*[\alpha(Z)d(X, Z)]}{R[\alpha^q]^{1/q}} \leq (\tilde{W}_p(P, Q) + \tilde{W}_p(Q, R)).$$

and taking the supremum on α the desired result follows from the definition of $\tilde{W}_p(Q, R)$. \square

2.2. Markov couplings. In this section, we only consider Markov couplings on the product space E^n with $n = 2$, the cases $n \geq 2$ following by simple induction reasoning.

Definition 2.1. Let $P, Q \in M(E^2)$, the set of Markov couplings $\tilde{M}(P, Q)$ are defined as the products $\pi = \pi_1 \pi_{2|1}$ with π_1 a coupling of P_1 and Q_1 and $\pi_{2|1}$ a coupling of $P_{2|1}$ and $Q_{2|1}$.

The terminology of Markov couplings was introduced by Rüschemdorf in [31]. Similar couplings have been used by Marton in [26]. The property of conditional independence in the gluing Lemma 2.1 is nicely compatible with Markov couplings:

Lemma 2.3. *For any Markov couplings $\pi_{x,y} \in \tilde{M}(P, Q)$ and $\pi_{y,z} \in \tilde{M}(P, Q)$ with $P, Q, R \in \tilde{M}(E^2)$ it exists a distribution $\pi_{x,y,z}$ with corresponding margins and such that $X = (X_1, X_2)$ and $Z = (Z_1, Z_2)$ are independent conditionally on $Y = (Y_1, Y_2)$.*

Proof. By assumption $\pi_{x,y} = \pi_{x_1,y_1} \pi_{x_2,y_2|x_1,y_1}$ and $\pi_{y,z} = \pi_{y_1,z_1} \pi_{y_2,z_2|y_1,z_1}$. Let us define $\pi_{x,y,z}$ as $\pi_{x_1,y_1,z_1} \pi_{x_2,y_2,z_2|x_1,y_1,z_1}$ by the relation

$$(2.4) \quad \pi_{x_1,y_1,z_1} = \pi_{x_1|y_1} \pi_{z_1|y_1} \pi_{y_1},$$

and

$$(2.5) \quad \pi_{x_2,y_2,z_2|x_1,y_1,z_1} = \pi_{x_2|x_1,y_1,y_2} \pi_{z_2|y_1,z_1,y_2} \pi_{y_2|y_1}.$$

Let us check that $\pi_{x,y,z}$ has the correct margins. First, from the classical gluing lemma we know that π_{x_1,y_1,z_1} has the correct margins. It remains to prove that $\pi_{x_2,y_2,z_2|x_1,y_1,z_1}$ has the correct margins. Notice that from the definition of Markov couplings, we have $\pi_{y_2|y_1} = \pi_{y_2|x_1,y_1} = \pi_{y_2|y_1,z_1}$. Thus the first margin of $\pi_{x_2,y_2,z_2|x_1,y_1,z_1}$ is equal to

$$\pi_{x_2|x_1,y_1,y_2} \pi_{y_2|y_1} = \pi_{x_2|x_1,y_1,y_2} \pi_{y_2|x_1,y_1} = \pi_{x_2,y_2|x_1,y_1}.$$

The same reasoning show that the second margin is also the correct one.

We proved above that by construction X_1 and Z_1 are independent conditionally on Y_1 , i.e. that $\pi_{x_1,z_1|y_1} = \pi_{x_1|y_1} \pi_{z_1|y_1}$. Let us show that it is also the case conditionally on Y_1 and Y_2 . We have

$$\pi_{x_1,z_1|y_1,y_2} = \frac{\pi_{x_1,z_1,y_1,y_2}}{\pi_{y_1,y_2}} = \frac{\pi_{y_2|y_1} \pi_{x_1,z_1,y_1}}{\pi_{y_2|y_1} \pi_{y_1}} = \pi_{x_1,z_1|y_1}$$

the third identity following from the identity $\pi_{y_2|y_1} = \pi_{y_2|x_1,y_1,z_1}$ by the identity (2.5). Thus, using that X_1 and Z_1 are independent conditionally on Y_1 we obtain the identity $\pi_{x_1,z_1|y_1,y_2} = \pi_{x_1|y_1} \pi_{z_1|y_1}$. We conclude that $\pi_{x_1,z_1|y_1,y_2} = \pi_{x_1|y_1,y_2} \pi_{z_1|y_1,y_2}$ as

$$\pi_{x_1|y_1} = \frac{\pi_{y_2|y_1} \pi_{x_1,y_1}}{\pi_{y_2|y_1} \pi_{y_1}} = \frac{\pi_{x_1,y_1,y_2}}{\pi_{y_1,y_2}} = \pi_{x_1|y_1,y_2}$$

the third identity following from the identity $\pi_{y_2|y_1} = \pi_{y_2|x_1,y_1}$ by definition of Markov couplings (the same is true replacing x_1 by z_1).

It remains to prove that X_2 is independent of Z_2 conditionally on (X_1, Z_1) and (Y_1, Y_2) . Indeed, we have by construction

$$\pi_{x_2,z_2|x_1,y_1,z_1,y_2} = \frac{\pi_{x_2,y_2,z_2|x_1,y_1,z_1}}{\pi_{y_2|x_1,y_1,z_1}} = \frac{\pi_{x_2,y_2,z_2|x_1,y_1,z_1}}{\pi_{y_2|y_1}} = \pi_{x_2|x_1,y_1,y_2} \pi_{z_2|y_1,z_1,y_2},$$

the last identity following from the identity (2.5). Thus the result is proved:

$$\pi_{x,z|y} = \pi_{x_2,z_2|x_1,y_1,z_1,y_2} \pi_{x_1,z_1|y_1,y_2} = \pi_{x_2|x_1,y_1,y_2} \cdot \pi_{z_2|y_1,z_1,y_2}$$

□

2.3. Weak transport costs on E^n , $n \geq 2$. We extend the definition of \tilde{W} on the product space E^n for $n \geq 2$. Let $P, Q \in M(E^n)$ we define

$$(2.6) \quad \tilde{W}_p(P, Q) = \sup_{\alpha \in M^+(E^n)} \inf_{\pi \in \tilde{M}(P, Q)} \frac{\sum_{j=1}^n \pi[\alpha_j(Y)d(X_j, Y_j)]}{(\sum_{j=1}^n Q[\alpha_j(Y)^q])^{1/q}}$$

with the convention $(\sum_{j=1}^n Q[\alpha_j(Y)^q])^{1/q} = \max_{1 \leq j \leq n} \text{ess sup } \alpha_j$ if $q = \infty$ and

$$(2.7) \quad \tilde{W}_\alpha(P, Q) = \inf_{\pi \in \tilde{M}(P, Q)} \sum_{j=1}^n \pi[\alpha_j(Y)d(X_j, Y_j)]$$

for any fixed $\alpha = (\alpha_j)_{1 \leq j \leq n} \in M^+(E^n)$. Considering Markov couplings, we can use the conditional independence in the gluing Lemma 2.3 to assert that the weak transport cost on E^n also satisfies the triangular inequality. More useful, \tilde{W}_α satisfies an inequality similar than the triangular one:

Lemma 2.4. *For any $P, Q, R \in M(E^n)$, for any $\alpha \in M^+(E^n)$ there exists $\tilde{\alpha} \in M^+(E^n)$ satisfying $Q[\tilde{\alpha}_j(Y)]^q \leq R[\alpha_j^q(Z)]$ for any $1 \leq j \leq n$ and*

$$(2.8) \quad \tilde{W}_\alpha(P, R) \leq \tilde{W}_{\tilde{\alpha}}(P, Q) + \tilde{W}_\alpha(Q, R)$$

Remark 2.1. As a consequence of the Lemma 2.4, we obtain the triangular inequality for \tilde{W}

$$(2.9) \quad \tilde{W}_p(P, R) \leq \tilde{W}_p(P, Q) + \tilde{W}_p(Q, R)$$

by taking the supremum on α on both sides of (2.8) and using the relation $Q[\tilde{\alpha}_j(Y)]^q \leq R[\alpha_j^q(Z)]$.

Proof. Let us fix $\alpha \in M^+(E^n)$ such that $R[\alpha_j^q] < \infty$ for all $1 \leq j \leq n$. Define recursively the couplings $\pi_{y,z}^*$ and $\pi_{x,y}^* \in \tilde{M}(E^2)$ such that

$$\begin{aligned} \pi_{y,z}^* \left[\sum_{j=1}^n \alpha_j(Z)d(X_j, Z_j) \right] &= \tilde{W}_\alpha(Q, R), \\ \pi_{x,y}^* \left[\sum_{j=1}^n \pi_{z|y}^*[\alpha_j(Z)|Y]d(X_j, Y_j) \right] &= \tilde{W}_{\pi_{z|y}^*[\alpha(Z)|Y]}(P, Q). \end{aligned}$$

where we use Jensen's inequality. Let us denote $\pi^* = \pi_{x,y,z}^*$ obtained by the gluing Lemma 2.3 of $\pi_{x,y}^*$ and $\pi_{y,z}^*$. Then

$$\begin{aligned} \pi_{x,z}^* \left[\sum_{j=1}^n \alpha_j d(X_j, Z_j) \right] &\leq \pi_{y,z}^* \left[\sum_{j=1}^n \alpha_j(Z)d(X_j, Y_j) \right] + \pi^* \left[\sum_{j=1}^n \alpha_j(Z)d(Y_j, Z_j) \right] \\ &\leq \pi_{x,y}^* \left[\sum_{j=1}^n \pi_{z,y}^*[\alpha_j(Z)|Y]d(X_j, Y_j) \right] \\ &\quad + \pi_{y,z}^* \left[\sum_{j=1}^n \alpha_j(Z)d(Y_j, Z_j) \right] \\ (2.10) \quad &\leq \tilde{W}_{\pi_{z|y}^*[\alpha(Z)|Y]}(P, Q) + \tilde{W}_\alpha(Q, R). \end{aligned}$$

The inequality (2.8) follows from (2.10) taking $\tilde{\alpha}_j = \pi_{y,z}^*[\alpha_j(Z)|Y = \cdot]$ and noticing that the relation $Q[\tilde{\alpha}_j^2(Y)] \leq R[\alpha_j^2(Z)]$ holds by an application of Jensen's inequality. \square

3. WEAK TRANSPORT INEQUALITIES

3.1. Weak transport inequalities. Let us denote $\mathcal{K}(P|Q)$ the relative entropy (or the Kullback-Leibler divergence) defined as $\mathcal{K}(P|Q) = Q[\log(dQ/dP)]$ when $Q \ll P$, $\mathcal{K}(P|Q) = \infty$ otherwise. Let us say that the probability measure P on E^n satisfies the weak transport inequality $\tilde{T}_p(C)$ when for all distribution Q on E^n we have

$$(3.1) \quad \tilde{W}_p(P, Q) \leq \sqrt{2C\mathcal{K}(P|Q)}.$$

Let us say that P satisfies the inverted weak transport inequality $\tilde{T}_p^{(i)}(C)$ when

$$(3.2) \quad \tilde{W}_p(Q, P) \leq \sqrt{2C\mathcal{K}(P|Q)}.$$

Notice that by definition and by Jensen's inequality P satisfies $\tilde{T}_p(C)$ and $\tilde{T}_p^{(i)}(C)$ as soon as $\tilde{T}_{p'}(C)$ and $\tilde{T}_{p'}^{(i)}(C)$ reciprocally with $p' \geq p$. Moreover $\tilde{T}_1(C) = \tilde{T}_1^{(i)}(C) = T_1(C)$ where $T_p(C)$ is the classical transport inequality defined for any $1 \leq p \leq 2$ as

$$\inf_{\pi \in \tilde{M}(P, Q)} \pi[d^p(X, Y)]^{1/p} \leq \sqrt{2C\mathcal{K}(P|Q)}.$$

3.2. Weak transport inequalities on E . Let us consider in this section P a probability measure on E (case $n = 1$). Let us show the following

Theorem 3.1.

- (1) Any $P \in M(E)$ satisfies $\tilde{T}_2(1)$ and $\tilde{T}_2^{(i)}(1)$ when d is the Hamming distance $d(x, y) = 1_{x \neq y}$.
- (2) Any $P \in M(E)$ satisfies $\tilde{T}_2(D^2)$ and $\tilde{T}_2^{(i)}(D^2)$ for any metric d such that $\sup_{(x, y) \in E^2} d(x, y) =: D < \infty$.

Remark 3.1. Below is the proof of point (1) for the sake of completeness. However, it is a direct consequence of Theorem 2 in Marton [26], with an alternative proof in Samson [32] as their transport costs are stronger than ours. The constant 1 is still optimal for our weaker transport inequality, see the discussion in Section 5.3.

Remark 3.2. By definition every P satisfying the classical transport inequality $T_2(C)$ such that gaussian or log-concave measure satisfies also $\tilde{T}_2(C)$ and $\tilde{T}_2^{(i)}(C)$. However, any distribution having a support with finite diameter satisfies $\tilde{T}_2(C)$ by point (2) of the above Theorem but not necessarily $\tilde{T}_2^{(i)}(C)$. For any metric d the weak transport inequalities $\tilde{T}_2(C)$ and $\tilde{T}_2^{(i)}(C)$ have dual forms given below in (3.3) and (3.4). These expressions are particularly explicit when d is the Hamming distance.

Proof. Let \mathcal{C}_b denotes the set of all continuous bounded functions. From the dual form of \tilde{W}_α for $\alpha \in M^+(E)$ fixed we have

$$\tilde{W}_\alpha(P, Q) = \inf_{\pi} \pi[\alpha(Y)d(X, Y)] = \sup_{f \in \mathcal{C}_b} Q[f_\alpha] - P[f]$$

where $f_\alpha(y) = \inf_x \{\alpha(y)d(x, y) + f(x)\}$. Then a measure P satisfies $\tilde{T}_2(C)$ if for any $\alpha \in M^+(E)$ and any probability measure Q

$$\sup_{f \in \mathcal{C}_b} Q[f_\alpha] - P[f] \leq \sqrt{2CQ[\alpha^2]\mathcal{K}(Q|P)} = \inf_{\lambda > 0} \lambda CQ[\alpha^2]/2 + \frac{\mathcal{K}(Q|P)}{\lambda}.$$

Thus P satisfies $\tilde{T}_2(C)$ if for any measure Q it holds

$$\sup_{\lambda > 0} \sup_{\alpha > 0} \sup_{f \in \mathcal{C}_b} Q[\lambda(f_\alpha - P[f]) - (\lambda\alpha)^2 C/2] - \mathcal{K}(Q|P) \leq 0.$$

By the variational form of the entropy we obtain

$$(3.3) \quad \sup_{\lambda > 0} \sup_{\alpha > 0} \sup_{f \in \mathcal{C}_b} P[\exp(\lambda(f_\alpha - P[f]) - (\lambda\alpha)^2 C/2)] \leq 1.$$

In the specific case $d(x, y) = 1_{x \neq y}$ we have the explicit expression $f_\alpha(y) = (\alpha(y) + \inf f) \wedge f(y)$. As the difference $f_\alpha - f$ is unchanged when adding a constant on f , we can take $\inf f = 0$ with no loss of generality and

$$\sup_{\alpha > 0} P[\exp(\lambda(f_\alpha - P[f]) - (\lambda\alpha)^2 C/2)] = P[\exp(\lambda(f - P[f]) - \lambda^2 f^2 C/2)].$$

But for any $X > 0$ we have $X - X^2/2 \leq \log(1 + X)$ and thus

$$P[\exp(X - X^2/2)] \leq 1 + P[X] \leq \exp(P[X]).$$

$\tilde{T}_2(1)$ follows by taking $X = \lambda f$. To prove that $\tilde{T}_2^{(i)}(1)$ holds we start from its dual form. For equivalent reasons than the preceding dual form (3.3), our weak transport inequality $\tilde{T}_2^{(i)}(C)$ holds for any $C > 0$ iff

$$(3.4) \quad \sup_{\lambda > 0} \sup_{\alpha > 0} \sup_{f \in \mathcal{C}_b} P[\exp(\lambda(f_\alpha - P[f]) - P[(\lambda\alpha)^2]C/2)] \leq 1.$$

Noticing that we can restrict to $\alpha(x) \leq \sup f - f(x)$, taking $\sup f = 0$ and $C = 1$ we obtain the sufficient condition

$$\sup_{f < 0} P[\exp(\lambda(f - P[f]) - P[(\lambda f)^2]/2)] \leq 1.$$

For any non positive r.v. X we have $\exp(X) \leq 1 + X + X^2/2$ and the desired result follows.

Point (2) is proved noticing that $d(x, y) \leq D1_{x \neq y}$. □

3.3. Weak transport inequalities on E^n , $n \geq 2$. Let us present a new coupling technique based on the following so called $\tilde{\gamma}(p)$ -weakly dependent properties of any measure P on E^n . Add artificially time 0 and put $X_0 = Y_0 = x_0 = y_0$ for a fixed point $y_0 \in E$. Denote $x^{(i)} = (x_i, \dots, x_0)$ for $i \geq 0$. Recall the classical Wassertein distance

$$W_p(P, Q) = \inf_{\pi \in \tilde{M}(P, Q)} \pi[d^p(X, Y)]^{1/p}.$$

Let us work under the following weak dependence assumption:

Definition 3.1. For any $1 \leq 2 \leq p$, any measure d , the probability measure P is $\gamma(p)$ -weakly dependent if for any $0 \leq i < k \leq n$ there exists the coefficient $\gamma_{k,i}(p) \geq 0$ such that

$$(3.5) \quad W_p(P_{x_k|x^{(i)}}, P_{x_k|x^{(i-1)}, y_i}) \leq \gamma_{k,i}(p)d(x_i, y_i) \quad \forall x^{(i)} \in E^{i+1}, (x_k, y_k) \in E^2.$$

Let us denote

$$\Gamma(p) = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \gamma_{2,1}(p) & 1 & 0 & \dots & 0 \\ \gamma_{3,1}(p) & \gamma_{3,2}(p) & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1 & 0 \\ \gamma_{n,1}(p) & \gamma_{n,2}(p) & \dots & \gamma_{n,n-1}(p) & 1 \end{pmatrix}.$$

The matrix $\Gamma(p)$ has n rows and n columns. We equip \mathbb{R}^n with the ℓ^p norm and the set of the matrix of size $n \times n$ with the subordinated norm, both denoted $\|\cdot\|_p$ for any $1 \leq p \leq \infty$.

Theorem 3.2. *If P is $\gamma(p)$ -weakly dependent and $P_{x_j|x^{(j-1)}}$ satisfies $\tilde{T}_p(C)$ or $\tilde{T}_p^{(i)}(C)$ for all $1 \leq j \leq n$ then $\tilde{T}_p(C\|\Gamma(p)\|_p^2 n^{2/p-1})$ or $\tilde{T}_p^{(i)}(C\|\Gamma(p)\|_p^2 n^{2/p-1})$ holds respectively.*

Remark 3.3. When the process (X_t) is stationary, we have $\gamma_{i,j}(p) = \gamma_{k,\ell}(p)$ for $j - i = k - \ell$. From the basic inequality $\|A\|_p \leq \|A\|_1^{1/p} \|A\|_\infty^{1-1/p}$ and the fact that $\|\Gamma\|_1 = \|\Gamma\|_\infty = 1 + \sum_{i=1}^n \gamma_{i,0}(p)$, then $\|A\|_p \leq 1 + \sum_{i=1}^n \gamma_{i,0}(p)$.

Remark 3.4. In the case $p = 1$ we recover the Hoeffding's inequality of Djellout *et al.* [12] from the dual form of $\tilde{T}_1(C) = \tilde{T}_1^i(C) = T_1(C)$. Recall the assumption $(C_1)'$ of [12]: for any 1-Lipschitz function f it holds

$$|P[f(X_{k+1}, \dots, X_n)|x^{(k)}] - P[f(X_{k+1}, \dots, X_n)|y_k, x^{(k-1)}]| \leq Sd(x_k, y_k).$$

For any $a = (a_{k+1}, \dots, a_n)'$, as $f = \sum_{j=k+1}^n a_j f_j$ is a 1-Lipschitz function whenever $\|a\|_\infty \leq 1$ and the f_j are 1-Lipschitz functions, we obtain $a' \mathbf{W}_i \leq S$ with $\mathbf{W}_i = d(x_i, y_i)^{-1} (W_p(P_{x_k|x^{(i)}}, P_{x_k|x^{(i-1)}, y_i}))'_{1 \leq i < k \leq n}$. Denoting \mathbf{W} the $n \times n$ matrix of the \mathbf{W}_k completed with 0 we obtain $\|a' \mathbf{W}\|_\infty \leq S$ for all $\|a\|_\infty \leq 1$. By the definition of the matrix norm and by duality it is equivalent to $\|\mathbf{W}'\|_1 \leq S$. Finally, one can always choose Γ such that it coincides with the supremum of \mathbf{W}' overall (x_k, y_k) such that $d(x_k, y_k) \neq 0$ and thus $(C_1)'$ is equivalent to $\|\Gamma\|_1 \leq S$.

Remark 3.5. In the case d is the Hamming distance $1_{x \neq y}$ then by the Kantorovitch-Rubinstein duality, for any $1 \leq p \leq 2$ and any $x^{(i)} \in E^{i+1}$ and $y_i \in E$:

$$\begin{aligned} & \inf_{\pi \in \tilde{M}(P_{x_k|x^{(i)}}, P_{x_k|x^{(i-1)}, y_i})} \pi[d^p(X, Y)] \\ & \leq \sup_{A \in \mathcal{B}_{x^{(i)}, y_i}} \sup_{x^{(i)}, y_i} |\mathbb{P}(X_k \in A | X^{(i)} = x^{(i)}) - \mathbb{P}(X_k \in A | X_i = y_i, X^{(i-1)} = x^{(i-1)})|. \end{aligned}$$

Here \mathcal{B} is the Borel σ -algebra and the supremum in $x^{(i)}, y_i$ is taken almost everywhere. Following the notation of Samson [32] for $p = 2$ and Kontorovitch and Ramanan [20] for $p = 1$, let us define for any $1 \leq p \leq 2$ the $\tilde{\gamma}(p)$ -weakly dependent coefficients as

$$(3.6) \quad \begin{aligned} & \tilde{\gamma}_{k,i}(p) = \\ & \sup_{A \in \mathcal{B}_{x^{(i)}, y_i}} \sup_{x^{(i)}, y_i} |\mathbb{P}(X_k \in A | X^{(i)} = x^{(i)}) - \mathbb{P}(X_k \in A | X_i = y_i, X^{(i-1)} = x^{(i-1)})|^{1/p}. \end{aligned}$$

The probability measure P is said to be $\tilde{\gamma}(p)$ -weakly dependent when its coefficients are finite. For Markov chains, condition (3.6) is equivalent to the uniform ergodicity. By definition, $\tilde{\gamma}_{k,i}^p \leq 2\phi_{k-i}$ where ϕ is the uniform mixing coefficient introduced by Ibragimov [18]. For $p = 2$, we obtain a weakened form of the transport inequality obtained by Samson [32] as

$$\tilde{W}_2(P, Q) \leq \inf_{\pi \in \tilde{M}} \sup_{\alpha_j > 0} \frac{\sum_{j=1}^n \pi[\alpha_j(Y) d(X_j, Y_j)]}{(\sum_{j=1}^n Q[\alpha_j(Y)^2])^{1/2}} = \inf_{\pi \in \tilde{M}} \left(\sum_{i=1}^n Q[\pi[X_i \neq Y_i | Y_i]^2] \right)^{1/2}.$$

However, the dual form of our weak transport inequality yields the same exponential inequalities than those obtained in [32]. Notice however that our notion of transport seems too weak to yield concentration properties in term of the convex distance as it is done in Talagrand in [33] or in Marton in [25].

Proof. The proofs of the two assertions are similar as the weak dependence condition (3.5) is symmetric in x_i and y_i . Thus the proof of the second assertion is omitted.

Let us fix $\alpha \in M^+(E^n)$ such that $Q[\alpha_j^q] < \infty$ for all $1 \leq j \leq n$. As preliminaries, we recall the following result of existence of the optimal Markov coupling due to from Rüschemdorf, [31] and a simple and useful inequality of this result stated in Lemma 3.4.

Let $\sigma : E^n \times E^n \mapsto \mathbb{R}_+$ and the section of σ in $(x_1, y_1) \in E^2$ as

$$\sigma_{x_1, y_1}(x_2, y_2) = \sigma((x_1, x_2), (y_1, y_2)).$$

Theorem 3.3 (Theorem 3 in [31]). *We have the equivalence between (1) and (2) which asserts (a) and (b)*

- (1) $\inf_{\pi \in \tilde{M}} \pi[\sigma] = \pi^*[\sigma]$ with $\pi^* \in \tilde{M}$,
- (2) (a) $h(x, y) := \inf_{\pi_{2|1}} \pi[\sigma_{x,y}] = \pi_{2|1}^*[\sigma_{x,y}|(x, y)]$ is finite π_1 -a.s. and
 (b) $\inf_{\pi_1} \pi_1[h] = \pi_1^*[h] < \infty$.

A simple corollary of this Theorem is the following result:

Lemma 3.4. *Let $P, Q \in M(E^n)$ be decomposed as $P = P_1 P_{|X_1}$ and $Q = Q_1 Q_{|Y_1}$ for $P_1, Q_1 \in M(E)$ and $P_{|x_1}, Q_{|y_1} \in M(E^{n-1})$. Then for any $\alpha \in M^+(E^n)$ and any coupling $\pi_1 \in \tilde{M}(P_1, Q_1)$ we have*

$$(3.7) \quad \tilde{W}_\alpha(P, Q) \leq \pi_1[Q_{|Y_1}[\alpha_1|Y_1]d(X_1, Y_1) + \tilde{W}_{\alpha^{(1)}}(P_{|X_1}, Q_{|Y_1})].$$

Proof. Let us assume that for almost all $x_1, y_1 \in E$ we have $\tilde{W}_{\alpha^{(1)}}(P_{|x_1}, Q_{|y_1}) < \infty$. Then, by lower semi-continuity, it exists $\pi_{|x_1, y_1}^*$ such that:

$$\pi_{|x_1, y_1}^* \left[\sum_{j=2}^n \alpha_j^{(1)} d(X_j, Y_j) \right] = \tilde{W}_{\alpha^{(1)}}(P_{|x_1}, Q_{|y_1}).$$

Thus the desired result follows from Theorem 3.3 remarking that for any $x_1, y_1 \in E$ we have

$$\pi_{|x_1, y_1}^*[\alpha_1^{(1)} d(x_1, y_1)] = \pi_{|x_1, y_1}^*[\alpha_1^{(1)}|x_1, y_1]d(x_1, y_1) = Q_{|y_1}[\alpha_1|y_1]d(x_1, y_1)$$

by definition of Markov couplings. \square

Let us consider now the following coupling scheme denoted $\tilde{\pi}$ defined recursively as $\tilde{\pi} = \tilde{\pi}_{n|n-1} \cdots \tilde{\pi}_{2|1} \tilde{\pi}_{1|0} \in \tilde{M}(E^n)$ where $\tilde{\pi}_{j|j-1} = \tilde{\pi}_{x_j, y_j|x^{(j-1)}, y^{(j-1)}}$ is determined such that

$$(3.8) \quad \tilde{\pi}_{j|j-1} \left[\sum_{k=j}^n Q_{|Y_j, y^{(j-1)}}[\alpha_k^q|Y_j, y^{(j-1)}]^{1/q} \gamma_{k,j} d(X_j, Y_j) \right] \\ = \left(\sum_{k=j}^n Q_{|y^{(j-1)}}[\alpha_k^q|y^{(j-1)}] \gamma_{k,j}^q \right)^{1/q} \tilde{W}_p(P_{x_j|x^{(j-1)}}, Q_{y_j|y^{(j-1)}})$$

for all $x^{(i-1)}, y^{(i-1)}$ in E^{i-1} .

We are now ready to prove the result iterating several time the same reasoning. Let us detail the case $j = 1$ when considering probabilities conditionally on y_0 . Applying (3.7) and (2.8) we have

$$(3.9) \quad \begin{aligned} \tilde{W}_\alpha(P, Q) &\leq \tilde{\pi}_{1|y^{(0)}}[Q_{|Y_1, y^{(0)}}[\alpha_1|Y_1, y^{(0)}]d(X_1, Y_1) + \tilde{W}_{\alpha^{(1)}}(P_{|X_1, y^{(0)}}, Q_{|Y_1, y^{(0)}})] \\ &\leq \tilde{\pi}_{1|y^{(0)}}[Q_{|Y_1, y^{(0)}}[\alpha_1|Y_1, y^{(0)}]d(X_1, Y_1) + \tilde{W}_{\alpha^{(1)}}(P_{|Y_1, y^{(0)}}, Q_{|Y_1, y^{(0)}}) \\ &\quad + \tilde{W}_{\tilde{\alpha}^{(1)}}(P_{|X_1, y^{(0)}}, P_{|Y_1, y^{(0)}})]. \end{aligned}$$

To bound the last term, we use the definition of the $\gamma(p)$ -weak dependence:

Lemma 3.5. *For any $\alpha_k \in M^+(E)$ for all $j < k \leq n$ and any $\gamma(p)$ -weakly dependent probability measure P we have*

$$(3.10) \quad \tilde{W}_{\alpha^{(j)}}(P_{|x_j, y^{(j-1)}}, P_{|y^{(j)}}) \leq \sum_{k=j+1}^n Q_{|y^{(j)}}[\alpha_k^q | y^{(j)}]^{1/q} \gamma_{k,j} d(x_j, y_j)$$

Proof. Assume that $Q[\alpha_k^q] < \infty$ for $j < k \leq n$. Then, applying the Holder inequality and the definition of Markov couplings, we have

$$\begin{aligned} \tilde{W}_{\alpha^{(j)}}(P_{|x_j, y^{(j-1)}}, P_{|y^{(j)}}) &= \inf_{\pi_{|j}} \pi_{|j} \left[\sum_{k=j+1}^n \alpha_k d(X_k, Y_k) \right] \\ &\leq \inf_{\pi_{|j}} \sum_{k=j+1}^n \pi_{|j} [\alpha_k^q | y^{(j)}]^{1/q} \pi_{|j} [d^p(X_k, Y_k)]^{1/p} \\ &\leq \sum_{k=j+1}^n Q_{|y^{(j)}}[\alpha_k^q | y^{(j)}]^{1/q} \inf_{\pi_{|j}} \pi_{|j} [d^p(X_k, Y_k)]^{1/p} \\ &\leq \sum_{k=j+1}^n Q_{|y^{(j)}}[\alpha_k^q | y^{(j)}]^{1/q} W_p(P_{|x_j, y^{(j-1)}}, P_{|y^{(j)}}) \end{aligned}$$

and the result follows by definition of the $\gamma(p)$ -weak dependence coefficients. \square

Collecting the bounds (3.9) and (3.10) we obtain

$$\begin{aligned} \tilde{W}_{\alpha}(P_{|y^{(0)}}, Q_{y^{(0)}}) &\leq \tilde{\pi}_{1|y^{(0)}} \left[\sum_{k=1}^n Q_{|Y^{(1)}, y_0}[\alpha_k^q | Y^{(1)}, y_0]^{1/q} \gamma_{k,1} d(X_1, Y_1) \right. \\ &\quad \left. + \tilde{W}_{\alpha^{(1)}}(P_{|Y_1, y^{(0)}}, Q_{|Y_1, y^{(0)}}) \right]. \end{aligned}$$

Let us do the same reasoning than above for any $1 \leq j \leq n$ conditionally on $y^{(j)}$ on $\tilde{W}_{\alpha^{(j-1)}}(P_{|y^{(j)}}, Q_{y^{(j)}})$ where $\alpha_j^{(i)}$ denotes the section of α_j in $y^{(i)}$ as $\alpha_j^{(i)}(y_{i+1}, \dots, y_n) = \alpha_j(y)$ and $\alpha^{(i)} = (\alpha_j^{(i)})_{j>i}$. For any $1 \leq j \leq n$, we obtain:

$$\begin{aligned} \tilde{W}_{\alpha^{(j-1)}}(P_{|y^{(j-1)}}, Q_{y^{(j-1)}}) &\leq \tilde{\pi}_{j|y^{(j-1)}} \left[\sum_{k=j}^n Q[\alpha_k^q | Y_j, y^{(j-1)}]^{1/q} \gamma_{k,j} d(X_j, Y_j) \right. \\ &\quad \left. + \tilde{W}_{\alpha^{(j)}}(P_{|Y_j, y^{(j-1)}}, Q_{|Y_j, y^{(j-1)}}) \right]. \end{aligned}$$

For the specific Markov coupling we consider, the identity (3.8) holds and

$$\begin{aligned} \tilde{W}_{\alpha^{(j-1)}}(P_{|y^{(j-1)}}, Q_{y^{(j-1)}}) &\leq \left(\sum_{k=j}^n Q_{|y^{(j-1)}}[\alpha_k^q | y^{(j-1)}] \gamma_{k,j}^q \right)^{1/q} \tilde{W}(P_{x_j|y^{(j-1)}}, Q_{y_j|y^{(j-1)}}) \\ &\quad + \tilde{\pi}_{j|y^{(j-1)}} [\tilde{W}_{\alpha^{(j)}}(P_{|Y_j, y^{(j-1)}}, Q_{|Y_j, y^{(j-1)}})] \\ &\leq \sum_{k=j}^n Q_{|y^{(j-1)}}[\alpha_k^q | y^{(j-1)}]^{1/q} \gamma_{k,j} \tilde{W}(P_{x_j|y^{(j-1)}}, Q_{y_j|y^{(j-1)}}) \\ &\quad + \tilde{\pi}_{j|y^{(j-1)}} [\tilde{W}_{\alpha^{(j)}}(P_{|Y_j, y^{(j-1)}}, Q_{|Y_j, y^{(j-1)}})] \end{aligned}$$

where the last inequality follows from the concavity of $x \rightarrow x^{1/q}$ and Jensen's inequality. Applying an inductive argument, we obtain

$$\tilde{W}_{\alpha}(P, Q) \leq Q \left[\sum_{j=1}^n \sum_{k=j}^n Q[\alpha_k^q | Y^{(j-1)}]^{1/q} \gamma_{k,j} \tilde{W}(P_{x_j|Y^{(j-1)}}, Q_{y_j|Y^{(j-1)}}) \right]$$

$$\begin{aligned}
&\leq \sum_{j=1}^n \sum_{k=j}^n Q[\alpha_k^q]^{1/q} \gamma_{k,j} Q[\tilde{W}(P_{x_j|Y^{(j-1)}}, Q_{y_j|Y^{(j-1)}})^p]^{1/p} \\
&\leq \sum_{j=1}^n \sum_{k=j}^n Q[\alpha_k^q]^{1/q} \gamma_{k,j} Q[2C\mathcal{K}(Q_{y_j|Y^{(j-1)}}|P_{x_j|Y^{(j-1)}})^{p/2}]^{1/p}
\end{aligned}$$

the second inequality follows from Hölder's and Jensen's inequalities and the last one from the assumption $P_{x_j|y^{(j-1)}} \in \tilde{T}_p(C)$. Let us denote \mathbf{Q} the row vector $(Q[\alpha_k^q]^{1/q})_{1 \leq k \leq n}$ and \mathbf{W} the column vector $(Q[2C\mathcal{K}(P_{x_j|Y^{(j-1)}}|Q_{y_j|Y^{(j-1)}})^{p/2}]^{1/p})'_{1 \leq j \leq n}$. With $\langle \cdot; \cdot \rangle$ denoting the scalar product, we obtain

$$\tilde{W}_\alpha(P, Q) \leq \langle \mathbf{Q}; \Gamma \mathbf{W} \rangle \leq \|\mathbf{Q}\|_q \|\Gamma\|_p \|\mathbf{W}\|_p.$$

Notice that we have the identities

$$\begin{aligned}
\|\mathbf{Q}\|_q &= \left(\sum_{j=1}^n Q[\alpha_j^q] \right)^{1/q}, \\
\|\mathbf{W}\|_p &= \left(\sum_{j=1}^n Q[2C\mathcal{K}(Q_{y_j|Y^{(j-1)}}|P_{x_j|Y^{(j-1)}})^{p/2}] \right)^{1/p}, \\
\mathcal{K}(P|Q) &= \sum_{j=1}^n Q[2C\mathcal{K}(Q_{y_j|Y^{(j-1)}}|P_{x_j|Y^{(j-1)}})].
\end{aligned}$$

Indeed, noticing that $p/2 \leq 1$, successive applications of Jensen's inequality and Hölder's inequality yield

$$\begin{aligned}
\|\mathbf{W}\|_p &\leq \left(\sum_{j=1}^n Q[2C\mathcal{K}(Q_{y_j|Y^{(j-1)}}|P_{x_j|Y^{(j-1)}})^{p/2}] \right)^{1/p} \\
&\leq n^{1/p-1/2} \left(\sum_{j=1}^n Q[2C\mathcal{K}(Q_{y_j|Y^{(j-1)}}|P_{x_j|Y^{(j-1)}})] \right)^{1/2} \\
&\leq \sqrt{n^{2/p-1} 2C\mathcal{K}(Q|P)}.
\end{aligned}$$

Finally, we obtain

$$\frac{\sum_{j=1}^n \tilde{\pi}[\alpha_j(Y)d(X_j, Y_j)]}{\left(\sum_{j=1}^n Q[\alpha_j^q]\right)^{1/q}} \leq \sqrt{2C\|\Gamma\|_p^2 n^{2/p-1} \mathcal{K}(Q|P)}.$$

The desired result follows by definition of the weak transport cost by taking the supremum over all $\alpha \in M^+(E^n)$. \square

4. EXAMPLES OF $\gamma(p)$ -WEAKLY DEPENDENT PROCESSES

We have already noticed that when d is chosen as the Hamming distance then the $\tilde{\gamma}(p)$ -weakly dependence is, for example, satisfied for ϕ -mixing processes with $\|\Gamma(p)\|_p \leq 1 + \sum_{i=1}^n (2\phi_i)^{1/p}$ for any $1 \leq p \leq 2$, see [32]. But the $\tilde{\gamma}(p)$ -weakly dependence is also satisfied for non stationary sequences, see [20].

For E being a real vector space, the choice of the Hamming distance is not natural and the resulting weakly dependent conditions are often too restrictive. In what follows, we focus on the more natural choice $d = \|\cdot\|$ the euclidian norm. We will extensively use the fact that probability measures satisfying weak transport inequalities admit finite moments of any order for $d(x_0, X)$, $\forall x_0 \in E$ (it also implies exponential moments, see the next Section).

4.1. Linear models.

Example 4.1 (AR(∞) models). Consider (X_t) the stationary solution of the autoregressive equation

$$X_t = \sum_{i \geq 1} a_i X_{t-i} + \xi_t$$

where the real numbers a_i are such that $1 - \sum_{i \geq 1} a_i z^i$ does not have unit root outside the unit circle. Then the weak dependence condition (3.5) is satisfied with $\gamma_{k,j}(p) = |a_{k-j}|$ for any $1 \leq p \leq 2$ and any $0 \leq j < k \leq n$.

Example 4.2 (MA(∞) models). Let consider $X_t = \sum_{i=1}^{\infty} a_i \xi_{t-i}$ with real numbers a_i such that

$$\sum_{i=1}^{\infty} |a_i| < \infty.$$

Then the model is well defined and if it is invertible, i.e. $\sum_{i=1}^{\infty} a_i z^i$ has no root outside the unit circle, then it admits an AR(∞) representation and the weak dependence condition (3.5) holds with $\gamma_{k,j}(p) \leq |\sum_{j \leq k} \sum_{i_1 + \dots + i_j = k} \prod_{\ell=1}^j a_{i_\ell}|$ for any $1 \leq p \leq 2$ and any $0 \leq j < k \leq n$.

Example 4.3 (ARMA models). Let us consider the ARMA model

$$X_0(x) = x, \quad X_{t+1}(x) = AX_t(x) + \xi_{t+1}$$

in $E = \mathbb{R}^d$ where $A \in \mathcal{M}_{d,d}$ (the space of $d \times d$ matrices) and (Z_t) is a sequence of i.i.d. random vectors in \mathbb{R}^d called the innovations. This model is a particular case of the general model above with $\psi_t(x) = Ax + Z_t$. The $\gamma(p)$ -weak dependence condition is equivalent to

$$\rho_{sp}(A) := \max\{|\lambda|; \lambda \text{ is an eigenvalue in } \mathbb{C} \text{ of } A\} < 1,$$

which is the necessary and sufficient condition for the ergodicity of this linear ARMA model (X_t) .

4.2. Non-linear models.

Example 4.4 (Stochastic Recurrent Equation (SRE)). Consider the SRE (also called Iterated Random Functions)

$$(4.1) \quad X_0(x) := x \in E, \quad X_{t+1}(x) = \psi_{t+1}(X_t(x)), \quad t \geq 0,$$

where (ψ_t) is a sequence of i.i.d. random maps. Let us denote also P the probability of the whole process $(X_t)_{t \geq 0}$. For any $1 \leq p \leq 2$, if the distribution of $\psi_1(x)$ belongs to $\tilde{T}_p(C)$ or $\tilde{T}_p^{(i)}(C)$ for any $x \in E$ and that there exists some $S > 0$ satisfying

$$(4.2) \quad \sum_{t=1}^{\infty} P[d(X_t(x), X_t(x'))^p]^{1/p} \leq Sd(x, x') \quad \forall x, x' \in E.$$

then $P \in \tilde{T}_p(C(1+S)^2 n^{2/p-1})$ or $\tilde{T}_p^{(i)}(C(1+S)^2 n^{2/p-1})$.

Example 4.5 (General affine processes). Consider now the specific SRE

$$X_0(x) = x, X_{t+1}(x) = f(X_t(x)) + M(X_t(x))\xi_{t+1},$$

where $E = \mathbb{R}^d$, $\xi_t \in \mathbb{R}^{d'}$, $f : \mathbb{R}^d \mapsto \mathbb{R}^d$, $M : \mathbb{R}^d \mapsto \mathcal{M}_{d,d'}$ (the space of $d \times d'$ matrices) and the noise (ξ_t) is a sequence of i.i.d. random vectors of $\mathbb{R}^{d'}$ such that its distribution P_ξ is centered. Fix $p = 2$ and assume that:

- (1) $P_\xi \in \tilde{T}_2(C)$ or $\tilde{T}_2^{(i)}(C)$ on $\mathbb{R}^{d'}$ w.r.t. the Euclidean metric;
- (2) there exists $K > 0$ such that $\|M(x)\|_2 \leq K, \forall x \in \mathbb{R}^d$;

(3) the Lyapunov exponent in L^2 satisfies

$$\lambda_{max}(L^2) := \lim_{t \rightarrow \infty} \left(\sup_{x \neq y} \frac{P[|X_t(x) - X_t(y)|^2]}{|x - y|^2} \right)^{1/t} < 1.$$

Using a version of Lemma 2.1 in [12] we obtain that conditions (1) and (2) implies that $P_{x_i|x_{i-1}} \in \tilde{T}_2(CK^2)$ or $\tilde{T}_2^{(i)}(CK^2)$. Moreover condition (4.2) is satisfied with $S = (1 - \lambda_{max}(L^2))^{-1/2}$ and thus $P \in \tilde{T}_2(CK^2(1 + (1 - \lambda_{max}(L^2))^{-1/2})^2)$ or $\tilde{T}_2^{(i)}(CK^2(1 + (1 - \lambda_{max}(L^2))^{-1/2})^2)$. We answer positively to a question raised in Remark 3.6 in [12]. However notice that the condition $\lambda_{max}(L^2)$ can be difficult to check on specific models. One possible sufficient condition is the Lipschitz mixing condition of Duflo [14] asserting the existence of $K > 0$ and $0 < r < 1$ such that

$$P[|X_t(x) - X_t(y)|^2] \leq Kr^t|x - y|^2, \forall x, y \in E.$$

Another possibility is given in the next example.

Example 4.6 (Iterated Random Lipschitz Maps). Consider the general SRE (4.1) and assume that the random maps ψ_t are Lipschitz-continuous. Denote the Lipschitz coefficient of any function f by

$$\Lambda(f) := \sup_{x \neq y} \frac{d(f(x), f(y))}{d(x, y)}.$$

Let P_ψ be the distribution of the sequence of iid random maps (ψ_t) . The top Lyapunov exponent Λ^* is defined as $\lim_{t \rightarrow \infty} t^{-1} \log(\psi_0 \circ \psi_{-1} \circ \dots \circ \psi_{-t+1})$. Its existence in $\mathbb{R} \cup -\infty$ is due to the subadditive ergodic theorem. The condition $\Lambda^* < 0$ is sufficient for the existence of the stationary of the SRE. It implies that $\lambda_{max}(L^2)$ when $\psi_1(x)$ belongs to $\tilde{T}_p(C)$ or $\tilde{T}_p^{(i)}(C)$ for any $x \in E$.

Example 4.7 (Chains with Infinite Memory). Let us consider now the case of Chains with Infinite Memory introduced by Doukhan and Wintenberger [13]:

$$X_t = F(X_{t-1}, X_{t-2}, \dots; \xi_t), \quad \forall t \in \mathbb{Z}.$$

This model does not exhibit any Markov property. Assume there exists a sequence of non negative numbers (a_i) such that

$$P_\xi[d(F(x_1, x_2, \dots; \xi), F(y_1, y_2, \dots; \xi))]^{1/p} \leq \sum_{i \geq 1} a_i d(x_i, y_i).$$

If $\sum_{i \geq 1} a_i < 1$ and $F(x_1, x_2, \dots; \xi)$ is in $\tilde{T}_2(C)$ or $\tilde{T}_2^{(i)}(C)$ the stationary measure exists and (3.5) holds with $\gamma_{k,j}(p) \leq a_{k-j}$ for any $0 \leq j < k \leq n$.

5. NEW EXPONENTIAL INEQUALITIES

5.1. General exponential inequalities. Let $X = (X_1, \dots, X_n)$ be distributed as P and consider the function $f : E^n \mapsto \mathbb{R}$ such that there exist auxiliary functions $L_j : E^n \mapsto \mathbb{R}_+$, $1 \leq j \leq n$ satisfying

$$(5.1) \quad f(y) - f(x) \leq \sum_{j=1}^n L_j(y) d(x_j, y_j) \quad \forall x, y \in E^n.$$

Let us consider the function $g : E^n \mapsto \mathbb{R}$ such that there exist auxiliary functions $L_j^{(i)} : E^n \mapsto \mathbb{R}_+$, $1 \leq j \leq n$ such that

$$(5.2) \quad g(y) - g(x) \leq \sum_{j=1}^n L_j^{(i)}(x) d(x_j, y_j) \quad \forall x, y \in E^n.$$

The dual form of the weak transport inequalities implies the following new exponential inequality:

Theorem 5.1. *If P satisfies $\tilde{T}_p(C)$ and f satisfies (5.1) then for all $\lambda > 0$ we have*

$$(5.3) \quad P \left[\exp \left(\lambda(f - P[f]) - \frac{C\lambda^2}{2} \left(\left(2 - \frac{2}{p}\right) \sum_{j=1}^n L_j^{\frac{p}{p-1}} + \left(\frac{2}{p} - 1\right) \right) \right) \right] \leq 1.$$

If P satisfies $\tilde{T}_p^{(i)}(C)$ and g satisfies (5.2) then for all $\lambda > 0$ we have

$$(5.4) \quad P \left[\exp \left(\lambda(g - P[g]) - \frac{C\lambda^2}{2} \left(\left(2 - \frac{2}{p}\right) \sum_{j=1}^n P \left[L_j^{(i)\frac{p}{p-1}} \right] + \left(\frac{2}{p} - 1\right) \right) \right) \right] \leq 1.$$

Remark 5.1. Consider the case of a $\gamma(2)$ -weakly dependent sequence supported by $[0, 1]^n$. Then P satisfies $\tilde{T}(\|\Gamma(2)\|_2^2)$ and $\tilde{T}_2^{(i)}(\|\Gamma(2)\|_2^2)$ by Theorem 3.2 and the above results apply with $C = \|\Gamma(2)\|$.

Proof. The proofs of (5.3) and (5.4) are similar. We only detail the first one. Integrating (5.1) in (x, y) by π with marginals P and Q we get

$$Q[f] - P[f] \leq \pi \left[\sum_{j=1}^n L_j(Y) d(X_j, Y_j) \right]$$

and by definition of \tilde{W} we obtain

$$Q[f] - P[f] \leq Q \left[\sum_{j=1}^n L_j^q \right]^{1/q} \tilde{W}_p(P, Q).$$

Using that $P \in \tilde{T}_2(C)$ we obtain

$$(5.5) \quad Q[(f - P[f])] \leq Q \left[\sum_{j=1}^n L_j^q \right]^{1/q} \sqrt{2CK(P|Q)}.$$

From the variational identity

$$ab = \inf_{\lambda > 0} \lambda a^q / q + b^p / (\lambda^{p-1} p)$$

we get for all $\lambda > 0$:

$$Q[(f - P[f])] \leq \lambda C / q Q \left[\sum_{j=1}^n L_j^q \right] + \mathcal{K}(P|Q)^{p/2} 2^{p/2} C^{1-p/2} / (\lambda^{p-1} p).$$

We can rewrite it as

$$(p/2)Q \left[(p/C)^{1-p/2} \lambda^{p-1} (f - P[f]) - \lambda C / q \sum_{j=1}^n L_j^q \right]^{2/p} \leq \mathcal{K}(P|Q).$$

From the Young inequality

$$(p/2)x^{2/p} \geq yx - (1 - p/2)y^{2/(2-p)}$$

applied with $y = (C\lambda^2/p)^{2/p-1}$ we obtain

$$(p/2)((p/C)^{1-p/2} \lambda^{p-2})^{2/p} x^{2/p} \geq x - (1 - p/2)C\lambda^2/p$$

For $x = Q[\lambda(f - P[f]) - \lambda C / q \sum_{j=1}^n L_j^q]$ we obtain

$$Q \left[\lambda(f - P[f]) - \lambda C / q \sum_{j=1}^n L_j^q \right] - \mathcal{K}(P|Q) \leq (1/p - 1/2)C\lambda^2.$$

Then the desired result follows from the variational formula of the entropy. \square

5.2. Extensions of classical concentration inequalities to dependent cases. A first corollary of Theorem 5.1 is an extension of the classical following inequality: let f be a separately convex Lipschitz function on $[0, 1]^n$ then

$$(5.6) \quad P(|f - P[f]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right)$$

where L satisfies $|f(x) - f(y)| \leq L\|x - y\|$ for any $x, y \in [0, 1]^n$ equipped with the Euclidian norm. This result was extended to contracting Markov chains in Marton [25] and to $\tilde{\gamma}(2)$ -weakly dependent processes in Samson [32] for convex Lipschitz functions f . The extension to the more general $\gamma(2)$ -weakly dependent context is a straightforward Corollary of Theorem 5.1. From Remark 5.1 we know that P satisfies $\tilde{T}(\|\Gamma(2)\|_2^2)$ and $\tilde{T}_2^{(i)}(\|\Gamma(2)\|_2^2)$. Remark that with no loss of generality we can assume that f is smooth enough (see Samson [32] for a detailed proof of this well known fact). Then for any $x, y \in [0, 1]^n$, by convexity we have

$$f(x) - f(y) \leq \sum_{j=1}^n \partial_j f(x)(x_j - y_j) \leq \sum_{j=1}^n |\partial_j f(x)| |x_j - y_j|.$$

Thus f satisfies condition (5.1) with $L_j = \partial_j f$. From the Lipschitz assumption on f we assert that $\sum_{j=1}^n L_j^2(x) = \|\nabla f\|^2 \leq L$ where ∇f denotes the usual gradient of f . An application of Theorem 5.1 yields that

$$P[\exp(\lambda(f - P[f]))] \leq \exp(\|\Gamma(2)\|_2^2 L^2 \lambda^2 / 2).$$

From similar arguments $-f$ satisfies (5.2) with $\sum_{j=1}^n L_j^{(i)2}(x) \leq L$ and the same estimate holds on the Laplace transform of $-f + P[f]$. Applying the classical Chernoff arguments yields

Corollary 5.2. *For any $\gamma(2)$ -weakly dependent sequences on $[0, 1]^n$, for any convex L -Lipschitz function f it holds*

$$P(|f - P[f]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\|\Gamma(2)\|_2^2 L^2}\right).$$

This type of inequalities have a lot of applications, see [33].

From a statistical perspective, it is also interesting to investigate the properties of the empirical process. As a corollary of Theorem 5.1 we also obtain a Poissonian inequality for the empirical process $f(x) = \sup_{\mathcal{G}} \sum_{i=1}^n g^2(X_i)$ for square of real valued Lipschitz functions. Similar results are obtained in Section 3 of Boucheron *et al.* [8].

Corollary 5.3. *Assume that there exists $(\ell_i)_{1 \leq i \leq n}$ such that for any $g \in \mathcal{G}$ we have*

$$|g(x) - g(y)| \leq \sum_{i=1}^n \ell_i d(x_i, y_i), \quad \forall x, y \in E^n$$

with $\sum_{i=1}^n \ell_i^2 = L^2 < \infty$. If P satisfies $\tilde{T}_2(C)$ and $\tilde{T}_2^{(i)}(C)$ then for every $t \geq 0$ we have

$$\begin{aligned} P(f \geq P[f] + t) &\leq \exp\left(-\frac{t^2}{8CL^2(P[f] + t)}\right), \\ P(f \leq P[f] - t) &\leq \exp\left(-\frac{t^2}{8CL^2P[f]}\right). \end{aligned}$$

Proof. From the convex inequality $x^2 - y^2 \leq 2x(x - y)$ we easily check that f satisfies (5.1) with $\sum_{j=1}^n L_j^2 \leq 4L^2 f$ and $-f$ satisfies (5.2) with $\sum_{j=1}^n L_j^{(i)2} \leq 4L^2 f$. As P satisfies $\tilde{T}_2(C)$ an application of (5.3) yields that for all $\lambda > 0$ we have

$$P[\exp(f(\lambda - 4CL^2\lambda^2) - \lambda P[f])] \leq 1.$$

An application of the Chernoff argument yields that for every $0 \leq \lambda \leq (4CL^2)^{-1}$

$$P(f \geq P[f] + t) \leq \exp(-t\lambda(1 - 4CL^2\lambda) + 4CL^2\lambda^2 P[f]).$$

Optimizing in λ we obtain

$$\lambda = \frac{t}{8CL^2(t + P[f])}$$

and the first inequality of the Corollary follows. For the second inequality, we apply inequality (5.4) to obtain, for any $\lambda > 0$, that

$$P[\exp(\lambda(f - P[f]))] \leq \exp(4CL^2\lambda^2 P[f]).$$

The desired inequality follows by the Chernoff argument. \square

5.3. The specific case of the Hamming distance. We fix $d(x, y) = 1_{x \neq y}$ as in Samson [32]. Thus the result of this section holds for any $\tilde{\gamma}(2)$ -weakly dependent sequence (with no restriction on the margins). Using exactly the same arguments than above, an extension of the classical exponential inequality (5.6) also holds in this case

$$P(|f - P[f]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\|\tilde{\Gamma}(2)\|_2^2 L^2}\right)$$

where $\tilde{\Gamma}(2)$ is the matrix corresponding to the coefficients $\tilde{\gamma}(2)$

The case of the Hamming distance is specific because for any non negative function f we can replace the convexity argument $x^2 - y^2 \leq 2x(x - y)$ by the simple inequality $f(x) - f(y) \leq f(x)1_{x \neq y}$. Let us consider the empirical process $f(x) = |\sup_{\mathcal{G}} \sum_{i=1}^n g(X_i)|$ for some set of non negative real functions \mathcal{G} bounding by M . Then f satisfies (5.1) with $\sum_{j=1}^n L_j^2 \leq Mf$ and $-f$ satisfies (5.2) with $\sum_{j=1}^n L_j^{(i)2} \leq Mf$. Applying Theorem 5.1 we recover the results of Theorem 2 of [32]:

Theorem 5.4. *If $0 \leq g \leq M$ for all $g \in \mathcal{G}$ then for every $t \geq 0$*

$$\begin{aligned} P(f \geq P[f] + t) &\leq \exp\left(-\frac{t^2}{2M\|\tilde{\Gamma}(2)\|_2^2(P[f] + t)}\right), \\ P(f \leq P[f] - t) &\leq \exp\left(-\frac{t^2}{2M\|\tilde{\Gamma}(2)\|_2^2 P[f]}\right). \end{aligned}$$

The constant 1 in $\tilde{T}_2(1)$ or $\tilde{T}_2^{(i)}(1)$ is optimal in Theorem as discussed in Boucheron *et al.* [9] for the iid case. We refer the reader to this article for nice statistical applications of this result in the iid case.

Due to the simple inequality $f(x) - f(y) \leq f(x)1_{x \neq y}$, it is also possible to extend classical Bernstein's inequality in the $\tilde{\gamma}(2)$ -weakly dependent case:

Theorem 5.5 ([32] (page 460, line7)). *Let g be a measurable function $\mathbb{R} \rightarrow [-M, M]$ and let*

$$f = \sum_{i=1}^n g(X_i).$$

Then for all $0 \leq \lambda \leq 1/(M\|\tilde{\Gamma}(2)\|_2^2)$ we have

$$P[\exp(\lambda(f - P[f]))] \leq \exp\left(8\|\tilde{\Gamma}(2)\|_2^2 \sum_{i=1}^n P[(g(X_i) - P[g(X_i)])^2] \lambda^2\right).$$

This inequality has been applied to obtain exact oracle inequality with fast rates in the $\tilde{\gamma}(2)$ -weakly dependent context in [2].

6. APPLICATIONS TO ORACLE INEQUALITIES WITH FAST CONVERGENCE RATES

In this section we use the weak transport inequality to obtain new nonexact oracle inequalities in the $\gamma(2)$ -weakly dependent setting and oracle inequalities in the $\tilde{\gamma}(2)$ -weakly dependent setting. Instead of using extensions of classical inequalities given in the last Section we prefer to use a more direct approach using the PAC-bayesian paradigm. It allows us to consider the mathematical statistic problem of asserting oracle inequalities as a problem of conditional mass transport.

6.1. The statistical setting. We focus on the oracle inequalities of the ordinary least square estimator. Let us consider the case of linear regression where $E = \mathbb{R}^{d+1}$, $X = (Y, Z) = (Y, Z^{(1)}, \dots, Z^{(d)})$ equipped with the euclidian norm $\|\cdot\|$. The empirical risk is denoted

$$r(\theta) = \frac{1}{n} \sum_{i=1}^n (Y_i - Z_i \theta)^2$$

where $(X_i)_{1 \leq i \leq n} = (Y_i, Z_i)_{1 \leq i \leq n}$ are the observations. In our context, these observations are not necessarily independent and we denote by P their distribution. The risk of prediction is denoted

$$R(\theta) = P[r(\theta)] \quad \forall \theta \in \mathbb{R}^d.$$

The aim is to estimate the value $\bar{\theta} \in \mathbb{R}^d$ such that $R(\bar{\theta}) \leq R(\theta)$, $\forall \theta \in \mathbb{R}^d$. We consider the ordinary least square estimator $\hat{\theta}$ of $\bar{\theta}$ such that $r(\hat{\theta}) \leq r(\theta)$ for all $\theta \in \mathbb{R}^d$. Let us denote the excess of risk $\bar{R}(\theta) = R(\theta) - R(\bar{\theta}) \geq 0$, \bar{r} its empirical counterpart, $\mathcal{Z} = (Z_i)_{1 \leq i \leq n}$ the $n \times d$ matrix of the design, $\|\mathcal{Z}\|_n^2 = n^{-1} \sum_{i=1}^n \|Z_i\|^2$ and $G = P[\mathcal{Z}^T \mathcal{Z}]$ its corresponding Gram's matrix. Assume that G is a definite positive matrix and denote $\rho = \max(1, \rho_{sp}(G^{-1}))$. All the results of this sections are given for probability measures P satisfying $T_2(C)$ and $T_2^{(i)}(C)$ for some $C > 0$ on E^n . In view of Theorem 3.2 and for applications perspective in time series we are interested on $\gamma(2)$ or $\tilde{\gamma}(2)$ weakly dependent observations. The case of possibly non linear autoregression is of special interest. There the vector Z_i is a function of the past values $\phi(Y_1, \dots, Y_{i-1})$. Here ϕ is known, one can think of the projection on the last coordinates (case of linear autoregression), functions on Fourier basis or wavelets, etc. The regularity of the function ϕ impact the concentrations properties. The constant C in the weak transport inequality has to be estimated in each specific statistical case. For example, in the linear autoregressive case of order $\ell \geq 1$ fixed, we have $\gamma(2)_{k,0} \leq \gamma(2)_{\lceil k/\ell \rceil, 0}$ and in the non-linear autoregressive case, $\tilde{\gamma}(2)_{k,0} \leq \|\varphi\|_\infty \tilde{\gamma}(2)_{\lceil k/\ell \rceil, 0}$. Finally notice that $\tilde{\gamma}(2)$ coefficients are nicely estimated for any bounded measurable functions φ whereas it is not the case of $\gamma(2)$ coefficients that require more regularity on φ .

6.2. Nonexact oracle inequality for $\gamma(2)$ -weakly-dependent sequences. Our first result is a bound on the excess of risk.

Theorem 6.1. *For any measure Q and any $\beta > 0$ we have*

$$(6.1) \quad Q[\bar{R}(\hat{\theta})] \leq Q[\|\mathcal{Z}\|_n^2]/\beta + 4\sqrt{\rho C Q[K] n^{-1} (\mathcal{K}(P|Q) + \beta Q[\bar{R}(\hat{\theta})]/2)}$$

where

$$K := 4\frac{d}{\beta} + \left(1 + \|\bar{\theta}\|^2 + \frac{d+2}{\beta}\right)R(\bar{\theta}) + \left(\|\bar{\theta}\|^2 + \frac{d}{\beta}\right)\frac{d-1}{\beta} + (1 + \|\bar{\theta}\|^2)r(\bar{\theta}).$$

Proof. Considering the change $(\mathcal{Z}, \theta) \rightarrow (\mathcal{Z}G^{-1/2}, G^{1/2}\theta)$, we assume that the Gram matrix G is the identity matrix. This change of variable is ρ -Lipschitz function. Thus $\mathcal{Z}G^{-1/2}$ satisfies $\tilde{T}_2(\rho C)$ and $\tilde{T}_2^{(i)}(\rho C)$ using similar arguments than in Lemma 2.1 in [12]. Thus in the sequel $G = I_d$, $\mathcal{Z} \in \tilde{T}_2(\rho C)$ and $\tilde{T}_2^{(i)}(\rho C)$. With this notation, $P[\|\mathcal{Z}\|_n^2] = d$ where and $\|\hat{\theta} - \bar{\theta}\|^2 = R(\hat{\theta}) - R(\bar{\theta})$. We adopt the so called PAC-bayesian approach considering that $\hat{\theta} = \rho_{\hat{\theta}}[\theta]$ where $\rho_{\hat{\theta}} = \mathcal{N}_d(\bar{\theta}, \beta^{-1}I_d)$ for any $\beta > 0$. This probability measure is measurable with respect to the observations (X_i) . Thus, the properties of the measure $P\rho_{\hat{\theta}}$ are not simple to handle directly. The PAC-bayesian approach consist in introducing artificially the measure $\rho_{\bar{\theta}}$ called a priori because it does not depend on the observations (X_i) . Let us fix some measure Q and denote Q_{θ} the probability measure such that $\rho_{\bar{\theta}}Q_{\theta} = Q\rho_{\hat{\theta}}$.

Let us first study similar properties than in (5.1) of the function $f = \bar{r}$. With some abuse the euclidian norm on any vector space will also be denoted $\|\cdot\|$. Using the inequality $x^2 - y^2 \leq 2x(x - y) \leq 2|x||x - y|$ for any $x, y \in \mathbb{R}$ we obtain

$$\begin{aligned} f(x) - f(x') &\leq \frac{1}{n} \sum_{i=1}^n ((y_i - z_i\theta)^2 - (y'_i - z'_i\theta)^2 + (y'_i - z'_i\bar{\theta})^2 - (y_i - z_i\bar{\theta})^2) \\ &\leq \frac{2}{n} \sum_{i=1}^n (|y_i - z_i\theta| \|(1, \theta)\| \|x_i - x'_i\| + |y'_i - z'_i\bar{\theta}| \|(1, \bar{\theta})\| \|x_i - x'_i\|). \end{aligned}$$

Then by definition of \tilde{W}_2 and using Cauchy-Schwartz inequality we obtain conditionally on θ that

$$P[f] - Q_{\theta}[f] \leq 2\|(1, \theta)\| \sqrt{n^{-1}R(\theta)\tilde{W}_2(Q_{\theta}, P)} + 2\|(1, \bar{\theta})\| \sqrt{n^{-1}Q_{\theta}[r(\bar{\theta})]\tilde{W}_2(P, Q_{\theta})}$$

As P satisfies $\tilde{T}_2(\rho C)$ and $\tilde{T}_2^{(i)}(\rho C)$ and using the Cauchy-Schwartz inequality we obtain

$$Q_{\theta}[P[f] - f] \leq 4\sqrt{\rho C n^{-1} \mathcal{K}(P|Q_{\theta})((1 + \|\theta\|^2)R(\theta) + (1 + \|\bar{\theta}\|^2)Q_{\theta}[r(\bar{\theta})])}.$$

The positivity of the integrand with respect to $\rho_{\bar{\theta}}$ yields

$$\begin{aligned} \rho_{\bar{\theta}}Q_{\theta}[P[f] - f] &\leq 4\rho_{\bar{\theta}} \left[\sqrt{\rho C n^{-1} \mathcal{K}(P|Q_{\theta})((1 + \|\theta\|^2)R(\theta) + (1 + \|\bar{\theta}\|^2)Q_{\theta}[r(\bar{\theta})])} \right] \\ &\leq 4\sqrt{\rho C n^{-1} \rho_{\bar{\theta}}[\mathcal{K}(P|Q_{\theta})] (\rho_{\bar{\theta}}[(1 + \|\theta\|^2)R(\theta) + (1 + \|\bar{\theta}\|^2)Q[r(\bar{\theta})])}. \end{aligned}$$

Notice that by definition $\rho_{\bar{\theta}}Q_{\theta} = Q\rho_{\hat{\theta}}$ such that we have $\rho_{\bar{\theta}}[\mathcal{K}(P|Q_{\theta})] = \mathcal{K}(P|Q) + Q[\mathcal{K}(\rho_{\hat{\theta}}|\rho_{\bar{\theta}})]$. Moreover $\mathcal{K}(\rho_{\hat{\theta}}|\rho_{\bar{\theta}}) \leq \beta/2(R(\hat{\theta}) - R(\bar{\theta}))$ so that we obtain

$$\begin{aligned} Q\rho_{\hat{\theta}}[R(\theta) - R(\bar{\theta}) - r(\theta) + r(\bar{\theta})] &\leq \\ 4\sqrt{\rho C n^{-1} (\mathcal{K}(P|Q) + \beta/2Q[R(\hat{\theta}) - R(\bar{\theta})]) \rho_{\bar{\theta}}[(1 + \|\theta\|^2)R(\theta) + (1 + \|\bar{\theta}\|^2)Q[r(\bar{\theta})]} &\leq \end{aligned}$$

Now, by Jensen's inequality $Q\rho_{\hat{\theta}}[R(\theta)] \geq Q[R(\hat{\theta})]$ and computations gives that $Q\rho_{\hat{\theta}}[r(\theta)] \leq r(\hat{\theta}) + Q[\|\mathcal{Z}\|_n^2]/\beta \leq r(\bar{\theta}) + Q[\|\mathcal{Z}\|_n^2]/\beta$. Collecting those bounds, we obtain

$$\begin{aligned} Q[R(\hat{\theta}) - R(\bar{\theta}) - \|\mathcal{Z}\|_n^2/\beta] &\leq \\ 4\sqrt{\rho C n^{-1} (\mathcal{K}(P|Q) + \beta/2Q[R(\hat{\theta}) - R(\bar{\theta})]) \rho_{\bar{\theta}}[(1 + \|\theta\|^2)R(\theta) + (1 + \|\bar{\theta}\|^2)Q[r(\bar{\theta})]} &\leq \end{aligned}$$

To end the proof, let us compute $\rho_{\bar{\theta}}[(1 + \|\theta\|^2)R(\theta)]$ using the following identity

$$\rho_{\bar{\theta}}[(1 + \|\theta\|^2)R(\theta)] = \rho_{\bar{\theta}}[R(\theta)] + \rho_{\bar{\theta}}[\|\theta\|^2]R(\bar{\theta}) + \rho_{\bar{\theta}}[\|\theta\|^2 R(\theta) - R(\bar{\theta})].$$

Let us decompose the last term:

$$\rho_{\bar{\theta}}[\|\theta\|^2 R(\theta) - R(\bar{\theta})] = \rho_{\bar{\theta}}[\|\theta\|^2 \|\hat{\theta}\bar{\theta}\|] + 2n^{-1}P[\mathcal{Y}\mathcal{Z}]\rho_{\bar{\theta}}[\|\theta\|^2(\theta - \bar{\theta})]$$

where $\mathcal{Y} = (Y_1, \dots, Y_n)$. Simple computations on gaussian random variables give

$$\begin{aligned} \rho_{\bar{\theta}}[R(\theta)] &= R(\bar{\theta}) + d/\beta \\ \rho_{\bar{\theta}}[\|\theta\|^2] &= \|\bar{\theta}\|^2 + d/\beta \\ \rho_{\bar{\theta}}[\|\theta\|^2(\theta - \bar{\theta})] &= 2\bar{\theta}/\beta \\ \rho_{\bar{\theta}}[\|\theta\|^2\|\theta - \bar{\theta}\|^2] &= (\|\bar{\theta}\|^2 + d/\beta)(d-1)/\beta + \|\bar{\theta}\|^2/\beta + 3d/\beta. \end{aligned}$$

The desired result follows collecting all these bounds and noticing that $4P[\mathcal{Y}\mathcal{Z}]\bar{\theta} \leq 2nR(\bar{\theta})$. \square

In the proof above, we obtain the more general result: for any probability measures μ and ν such that there exists Q_θ satisfying $Q\mu = \nu Q_\theta$ we have:

(6.2)

$$Q\mu[\bar{R}] \leq Q\mu[\bar{r}(\theta)] + 4\sqrt{\rho C n^{-1} \mathcal{K}(P\nu|Q\mu)(\nu[(1 + \|\theta\|^2)R(\theta)] + (1 + \|\bar{\theta}\|^2)Q[r(\bar{\theta})])}.$$

This bound is obtained by integrating with respect to ν the conditional mass transport of $Q_\theta \circ r(\theta)^{-1}$ to $P \circ \bar{r}(\theta)^{-1}$. The weak transport inequalities satisfied by P and the convex properties of the function $(x_1, \dots, x_n) \rightarrow \bar{r}(\theta)$ are used to obtain a bound conditionally on θ .

Let us discuss the choices $\mu = \rho_{\hat{\theta}}$ and $\nu = \rho_{\bar{\theta}}$ made above. Notice that μ and ν have the same support from the assumption $Q\mu = \nu Q_\theta$. As soon as μ is centered in $\hat{\theta}$, Jensen's inequality yields $Q\mu[R(\theta)] \geq Q[R(\hat{\theta})]$. Next, if μ is sufficiently concentrated around $\hat{\theta}$ then $Q\mu[r(\theta) - r(\bar{\theta})]$ is small as $r(\hat{\theta}) - r(\bar{\theta}) < 0$. Choosing μ as the Dirac mass in $\hat{\theta}$ is excluded as the existence of some measure Q_θ satisfying $\nu Q_\theta = Q\mu$. The fact that the support of μ cannot depend on the observations (X_i) constrain us to choose measures supported on the whole space \mathbb{R}^d in absence of a priori information on $\hat{\theta}$. The term $Q\mu[r(\theta) - r(\bar{\theta})]$ can be seen as an alternative to the classical VC-dimension, see Mc Allester [29]. The measure μ should be chosen in order to bound this term (and the entropy $\mathcal{K}(\nu|\mu)$). It leads to Gibbs estimators that are nice alternatives to classical estimators, see Chapter 4 of the textbook of Catoni [10] in the iid case, Alquier and Wintenberger [3, 2] in weakly dependent settings. Here we choose the gaussian measures $\mu = \rho_{\hat{\theta}}$ and $\nu = \rho_{\bar{\theta}}$ as in Audibert and Catoni [4] for simplicity because we have an explicit computation $\mathcal{K}(\nu|\mu) = \beta/2\|\hat{\theta} - \bar{\theta}\|^2$. This choice leads to estimate the term $Q\mu[r(\theta) - r(\bar{\theta})]$ by $Q[\|\mathcal{Z}\|_n^2]/\beta$. This term can easily be estimated with d/β and a concentration term implying the entropy $\mathcal{K}(P|Q)$ in order to obtain a nonexact oracle inequality:

Corollary 6.2. *For any $0 < \varepsilon < 1$ and any $(d+2)/n < \eta < 1$ we have with probability $1 - \varepsilon$:*

$$R(\hat{\theta}) \leq (1 + B_1\eta)R(\bar{\theta}) + \frac{B_2d + 16\rho C \log(\varepsilon^{-1})}{n\eta} + \frac{B_3}{(n\eta)^2}$$

where

$$B_1 = 2(3 + 2\|\bar{\theta}\|^2 + \eta/n),$$

$$B_2 = 2(5 + \|\bar{\theta}\|^2),$$

$$B_3 = 2(d(d-1) + d/n).$$

Remark 6.1. This result extends nonexact oracle inequalities as developed by Lecué and Mendelson [21] to a dependent context but for the OLS only (and not also regularized estimators).

Proof. As for any $a, b > 0$ we have $2\sqrt{ab} \leq a\lambda + b/\lambda$ for any $\lambda > 0$ then from (6.1) we obtain

$$Q[\bar{R}(\hat{\theta}) - \|Z\|^2/\beta - K\lambda/n - \beta\bar{R}(\hat{\theta})/(2\lambda)] - \frac{4\rho C\mathcal{K}(P|Q)}{\lambda} \leq 0.$$

Notice that by definition of K we have

$$Q[K] = 4\frac{d}{\beta} + \left(1 + \|\bar{\theta}\|^2 + \frac{d+2}{\beta}\right)R(\bar{\theta}) + \left(\|\bar{\theta}\|^2 + \frac{d}{\beta}\right)\frac{d-1}{\beta} + (1 + \|\bar{\theta}\|^2)Q[r(\bar{\theta})]$$

by similar arguments than in the proof of Theorem 6.1 we have

$$Q[r(\bar{\theta})] - R(\bar{\theta}) \leq 2\sqrt{2\rho CR(\bar{\theta})n^{-1}\mathcal{K}(P|Q)}.$$

Similarly, as $P[\|Z\|^2] = d$ we obtain

$$Q[\|Z\|^2] - d \leq 2\sqrt{2\rho Cdn^{-1}\mathcal{K}(P|Q)}.$$

Collecting those bounds and using Cauchy-Schwartz inequality we obtain

$$\begin{aligned} Q[\|Z\|^2/\beta + \lambda/nr(\bar{\theta})] &\leq d/\beta + \lambda/nR(\bar{\theta}) \\ &\quad + 4\sqrt{\rho Cn^{-1}(d/\beta^2 + (\lambda/n)^2R(\bar{\theta}))\mathcal{K}(P|Q)}. \end{aligned}$$

Using again that $2\sqrt{ab} \leq a\lambda + b/\lambda$, choosing $\beta = \lambda = n\eta$ and by definition of B_1 , B_2 and B_3 we have

$$Q[\bar{R}(\hat{\theta}) - B_1\eta R(\bar{\theta}) - B_2d/(n\eta) - B_3/(n\eta)^2] \leq \frac{16\rho C\mathcal{K}(P|Q)}{n\eta}.$$

Choose Q as the probability P restricted to the complementary of the event corresponding to the desired oracle inequality and denoted A . Then

$$\frac{16\rho C \log(\varepsilon^{-1})}{n\eta} \leq Q[\bar{R}(\hat{\theta}) - B_1\eta R(\bar{\theta}) - B_2d/(n\eta) - B_3/(n\eta)^2]$$

Combining these two inequality we assert that for this specific Q we have $-\log(\varepsilon) \leq \mathcal{K}(P|Q)$. The relative entropy can be computed explicitly $\mathcal{K}(P|Q) = -\log(1-P(A))$ and thus the desired result follows. \square

6.3. Exact oracle inequality for $\tilde{\gamma}(2)$ -weakly-dependent sequences. Let us now give an equivalent of (6.2) when we equipped E with the Hamming distance $d(x, y) = 1_{x \neq y}$. Instead of using the convexity of $x \mapsto x^2$ as above, we use that $f(x) - f(y) \leq |f(x)|1_{x \neq y} + |f(y)|1_{x \neq y}$ for any f . Following the lines of the proof above with $f = \bar{r}$ we obtain easily

$$(6.3) \quad Q\mu[\bar{R}] \leq Q\mu[\bar{r}] + 2\sqrt{2\rho C\mathcal{K}(P\nu|Q\mu)(P\nu[\bar{r}^2] + Q\mu[\bar{r}^2])}.$$

For the specific choice $\mu = \rho_{\hat{\theta}}$ and $\nu = \rho_{\bar{\theta}}$ we use computations given in Lemma 1.2 in the supplementary material of [4] stating that for any $\theta \in \mathbb{R}^d$

$$\rho_{\theta}[\bar{r}^2] \leq 5\bar{r}(\theta)^2 + \frac{4\|\mathcal{Z}\|_n^2}{n\beta}r(\theta) + \frac{4\|\mathcal{Z}\|_n^4}{n\beta^2}$$

where $\|\mathcal{Z}\|_n^4 = n^{-1} \sum_{i=1}^n \|Z_i\|^4$. The quantities $Q[\|\mathcal{Z}\|_n^2 r(\hat{\theta})]$ and $Q[\|\mathcal{Z}\|_n^4]$ can be difficult to estimate for desired choices of Q . Let us work under the following

assumption on the set of parameters $\Theta \subseteq \mathbb{R}^d$ containing the support of P and the unit disc: there exists some finite constant $B > 0$ such that

$$(6.4) \quad B = \sup_{\theta \in \Theta} \frac{\sum_{i=1}^n \|Z_i \theta\|_\infty}{\sum_{i=1}^n P[Z_i \theta]^2}.$$

Similar assumption has been used in the iid case by Audibert and Catoni in [4]. Under (6.4) and the fact that we assume $P[\|\mathcal{Z}\|^2] = d$ with no loss of generality (see discussion in the proof above) we have $\|\mathcal{Z}\|_n^2 \leq Bd$ and $\|\mathcal{Z}\|_n^2 \leq (Bd)^2$. Moreover, using computations given in the supplementary material of [4] we obtain easily that

$$\bar{r}(\theta)^2 \leq n^{-1}(2B^2 + 8Br(\bar{\theta}))\bar{R}(\theta).$$

It leads to the following equivalent of Theorem 6.1

Theorem 6.3. *If condition (6.4) holds, we have*

$$Q[\bar{R}(\hat{\theta})] \leq \frac{Bd}{\beta} + 2\sqrt{2\rho Cn^{-1}(\mathcal{K}(P|Q) + \beta Q[\bar{R}(\hat{\theta})]/2)} \times \\ \sqrt{Q[(10B^2 + 40Br(\bar{\theta}))\bar{R}(\hat{\theta})] + 4Bd(R(\bar{\theta}) + Q[r(\bar{\theta})])/\beta + 8(Bd/\beta)^2}.$$

In the above estimate the terms involving $r(\bar{\theta})$ are nuisance terms because there is no control on $\bar{\theta}$. If this term is bounded then the main term multiplying the entropy is proportional to the excess risk $Q[\bar{R}(\hat{\theta})]$. It is the major advantage considering the Hamming distance compared with the Euclidian distance where instead $Q[R(\hat{\theta})]$ appeared. In the classical approach as developed by Massart in [28], the excess risk also appears via the variance term in Bernstein's inequality under the margin assumption of Tsybakov [35] that estimates this variance term by $\bar{R}(\hat{\theta})$. As $Q[\bar{R}(\hat{\theta})]$ is the quantity of interest, we can obtain exact oracle inequality the following corollary

Corollary 6.4. *For any $0 < \varepsilon < 1$ and any $M > 0$ we have with probability $1 - \varepsilon$:*

$$R(\hat{\theta}) \leq R(\bar{\theta}) + 160 \frac{B^2 + 4BM}{n} \times \\ \times \left(Bd + 8\rho C(\log(\varepsilon^{-1}) - \log P(r(\bar{\theta}) > M)) + \frac{d(R(\bar{\theta}) + M)}{10B + 40M} + \frac{8(Bd)^2}{n} \right).$$

Remark 6.2. As already noticed by Audibert and Catoni in [4] in the iid case, the exact oracle inequality holds for $\tilde{\gamma}(2)$ -weakly dependent sequences without any assumptions on the margins P except (6.4) (because any probability measures have supports diameter bounded by 1 for the Hamming distance d). We refer the reader to [4] for a nice way to bound the term $\log P(r(\bar{\theta}) > M)$ in the iid case under finite moments assumption on P of order 4 only.

Proof. Let us denote $A = \{r(\bar{\theta}) \leq M\}$ and P_A the restriction of P on A defined as $P_A(B) = P(B \cap A)$ for any measurable set B on E^n . We do not know wether P_A satisfies any weak transport inequality. However, a similar reasoning than for obtaining (6.3) yields

$$Q[\bar{R}(\hat{\theta})] \leq Bd/\beta + \rho_{\hat{\theta}} \left[\sqrt{(4Bd\bar{R}(\bar{\theta})/\beta + (4Bd/\beta)^2)n^{-1}\tilde{W}_2(Q_\theta, P_A)} \right. \\ \left. + \sqrt{((10B^2 + 40BM)Q[\bar{R}(\hat{\theta})] + 4BdM/\beta + (4Bd/\beta)^2)n^{-1}\tilde{W}_2(P_A, Q_\theta)} \right].$$

Now let us the triangular inequality of the weak transport cost (2.9):

$$\tilde{W}_2(P_A, Q_\theta) \leq \tilde{W}_2(P_A, P) + \tilde{W}_2(P, Q_\theta), \\ \tilde{W}_2(Q_\theta, P_A) \leq \tilde{W}_2(Q_\theta, P) + \tilde{W}_2(P, P_A).$$

Because P satisfies $\tilde{T}_2(\rho C)$ and $\tilde{T}_2^{(i)}(\rho C)$, both RHS terms are estimated with

$$\sqrt{2\rho CH(P_A|P)} + \sqrt{2\rho CK(P|Q_\theta)} \leq 4\sqrt{\rho C(\mathcal{K}(P|Q_\theta) - \log P(A))}$$

Collecting all these bounds and using the Cauchy-Schwartz inequality we obtain

$$Q[\bar{R}(\hat{\theta})] \leq Bd/\beta + 4 \left[\sqrt{2\rho Cn^{-1}(\mathcal{K}(P|Q) + \beta Q[\bar{R}(\hat{\theta})])/2 - \log P(A)} \times \right. \\ \left. \sqrt{((10B^2 + 40BM)Q[\bar{R}(\hat{\theta})] + 4Bd(R(\bar{\theta}) + M)/\beta + 8(Bd/\beta)^2)}. \right.$$

Using several times the inequality $2\sqrt{ab} \leq a\lambda + b/\lambda$ with $\lambda = \beta = n(40B^2 + 160BM)^{-1}$ yields

$$Q[\bar{R}]/4 \leq 40 \frac{B^2 + 4BM}{n} \left(Bd + 8\rho C(\mathcal{K}(P|Q) - \log P(A)) + \frac{d(R(\bar{\theta}) + M)}{10B + 40M} + \frac{8(Bd)^2}{n} \right)$$

We conclude as in the proof of Corollary 6.2 choosing Q as P restricted to the complementary of the event corresponding to the desired oracle inequality. \square

Acknowledgments. I thank the GT EVOL's team, especially F. Bolley, I. Gentil, C. Leonard and P.-M. Samson, for presenting and explaining patiently to me the transport inequalities subject. I also thank P. Alquier for helpful discussions on the PAC-bayesian approach.

REFERENCES

- [1] ADAMCZAK, R. (2008) A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** (34), 1000–1034.
- [2] ALQUIER, P. AND WINTENBERGER (2012) Fast rates in learning with dependent observations. preprint available on arxiv.org.
- [3] ALQUIER, P. AND WINTENBERGER (2012) Model selection for weakly dependent time series forecasting, *Bernoulli* **18** (3), 883–913.
- [4] AUDIBERT, J.-Y. AND CATONI, O. (2011) Robust linear least squares regression *Ann. Statist.* **39** (5), 2766–2794.
- [5] BERTAIL, P. AND CLÉMENCON, S. (2010) Sharp bounds for the tails of functionals of Markov chains. *Theory Probab. Appl.* **54** (3), 505–515,
- [6] BOBKOV, S. G. AND GÖTZE, F. (1999) Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal.* **163** (1), 1–28.
- [7] BOBKOV, S. G., GENTIL, I., LEDOUX, M. (2001) Hypercontractivity of Hamilton-Jacobi equations. *J. Math. Pures Appl.* **80** (7), 669–696.
- [8] BOUCHERON, S., LUGOSI, G. AND MASSART, P. (2009) On concentration of self-bounding functions. *Electron. J. Probab.* **14** (64), 1884–1899.
- [9] BOUCHERON, S., LUGOSI, G. AND MASSART, P. (2000). A sharp concentration inequality with applications. *Random Structures and Algorithms* **16** (3), 277–292.
- [10] CATONI, O. (2004) Statistical Learning Theory and Stochastic Optimization, *Lecture Notes in Mathematics*, Springer, Berlin.
- [11] DEDECKER, J. PRIEUR, C. AND RAYNAUD DE FITTE, P. (2006) Parametrized Kantorovich-Rubinstein theorem and application to the coupling of random variables. *Lecture Notes in Statist.*, **187**, Springer, New York.
- [12] DJELLOUT, H., GUILLIN, A. AND WU, L. (2004) Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Ann. Probab.* **32** (3B), 2702–2732.
- [13] DOUKHAN, P. AND WINTENBERGER, O. (2008) Weakly dependent chains with infinite memory. *Stochastic Process. Appl.* **118** (11), 1997–2013.
- [14] DUFLO, M. (1997) Random iterative models. *Applications of Mathematics*, **34**. Springer-Verlag, Berlin.
- [15] GAO, F., GUILLIN, A., AND WU, L. (2012) Bernstein types concentration inequalities for symmetric Markov processes. To appear in *SIAM Theor. Probab. Appl.*
- [16] GOZLAN, N. AND LÉONARD, C. (2010) Transport inequalities. A survey. *Markov Processes and Related Fields*, **16**, 635–736.
- [17] GUILLIN, A., LÉONARD, C., WU, L. AND YAO, N. Transportation-information inequalities for Markov processes. *Probab. Theory Related Fields* **144** (3-4), 669–695.

- [18] IBRAGIMOV, I. A. (1962) Some limit theorems for stationary processes. *Theory Probab. Appl.* **7**, 349–382.
- [19] JOULIN, A. AND OLLIVIER, Y. (2010) Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.* **38 (6)**, 2418–2442.
- [20] KONTOROVICH, A. AND RAMANAN, K. (2008) Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Ann. Probab.* **36(6)**, 2126–2158.
- [21] LECUÉ, G. AND MENDELSON, S. (2012) General nonexact oracle inequalities for classes with a subexponential envelope. *Ann. Statist.* **40 (2)**, 832–860.
- [22] LEDOUX, M. (1996) On Talagrand’s deviation inequalities for product measures. *ESAIM Probab. Statist.* **1**, 63–87
- [23] LEZAUD, P. (1998) Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.* **8 (3)**, 849–867.
- [24] MARTON K. (1986) A simple proof of the blowing-up lemma. *IEEE Trans. Inform. Theory* **32(3)**, 445–446.
- [25] MARTON K. (1996) Bounding \bar{d} -distance by informational divergence: a method to prove measure concentration. *Ann. Probab.* **24 (2)**, 857–866.
- [26] MARTON, K. (1996) A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.* **6 (3)**, 556–571.
- [27] MARTON, K. (2004) Measure concentration for Euclidean distance in the case of dependent random variables. *Ann. Probab.* **32 (3B)**, 2526–2544.
- [28] MASSART, P. (2007), *Concentration inequalities and model selection*. Springer, Berlin.
- [29] MCALLESTER, D. A. (1999) PAC-Bayesian model averaging. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 164–170, ACM, New York.
- [30] RIO, E. (2000) Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. (French) *C. R. Acad. Sci. Paris Sér. I Math.* **330 (10)**, 905–908
- [31] RÜSCHENDORF, L. (1985) The Wasserstein Distance and Approximation Theorems, *Z. Wahrsch. Verw. Gebiete* **70**, 117–129.
- [32] SAMSON, P.-M. (2000) Concentration of measure inequalities for Markov chains and Φ -mixing processes. *Ann. Probab.* **28 (1)**, 416–461.
- [33] TALAGRAND, M. (1995) Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l’I.H.E.S.*, **81**, 73–205.
- [34] TSIREL’SON, B. S. (1985) A geometric approach to a maximum likelihood estimation for infinite-dimensional Gaussian location, II, *Theory Probab. Appl.* **30**, 820–828.
- [35] TSYBAKOV, A. B. (2004) Optimal Aggregation of Classifiers in Statistical Learning. *Ann. Statist.* **32 (1)**, 135–166.
- [36] VILLANI, C. (2009) *Optimal transport, old and new*. Springer-Verlag, Berlin, 2009.
- [37] WINTENBERGER, O. (2010) Deviation inequalities for sums of weakly dependent time series. *Electron. Commun. Probab.* **15**, 489?503.

CEREMADE, PLACE DU MARÉCHAL DE LATTRE DE TASSIGNY, 75775 PARIS CEDEX 16, FRANCE AND CREST-LFA, 15 BOULEVARD GABRIEL PÉRI, 92245 MALAKOFF CEDEX, FRANCE

E-mail address: wintenberger@ceremade.dauphine.fr