Weak transport inequalities and applications to exponential inequalities and oracle inequalities

Olivier Wintenberger

WEAK TRANSPORT INEQUALITIES AND APPLICATIONS TO EXPONENTIAL AND ORACLE INEQUALITIES

OLIVIER WINTENBERGER

Abstract. We extend the weak transport as defined by Marton in [32] to other metrics than the Hamming distance. We obtain new weak transport inequalities for non products measures extending the results of Samson in [39]. 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 [40] for convex functions that are Lipschitz continuous. Expressing the concentration properties of the ordinary least square estimator as a conditional weak 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 5
2.1. Weak transport costs on E 5
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 $\Gamma_{d_1,d_2}(p)$-weakly dependent processes 12
4.1. $\Gamma_{d_1,d_2}(1)$-weakly dependent examples 12
4.2. $\tilde{\Gamma}(p)$-weakly dependent examples 13
4.3. $\Gamma_{\|\cdot\|,d_2}(2)$-weakly dependent exemples 14
5. New exponential inequalities 17
5.1. General exponential inequalities 17
5.2. Extensions of classical concentration inequalities to dependent cases 18
5.3. The specific case of the Hamming distance 19
6. Applications to oracle inequalities with fast convergence rates 20
6.1. The statistical setting 20
6.2. Nonexact oracle inequality for $\Gamma_{\|\cdot\|,d_2}(2)$-weakly dependent sequences 21
6.3. Exact oracle inequalities for $\tilde{\Gamma}(2)$-weakly-dependent sequences 24
References 25

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 [30], transport inequalities are known to efficiently yield dimension free concentration inequalities. Using a duality argument, Bobkov and Gotze [9] even proved that transport inequalities are equivalent to some concentration inequalities. Our references on the subject are the monograph of Villani [42] and the survey of Gozlan and Leonard [21]. Transport inequalities appear as a nice alternative to the classical modified log-Sobolev approach of Massart [34] 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 develops new kinds of transport inequalities, new exponential inequalities and new oracle inequalities with fast rates of convergence.

In the case of product measures with common margin (iid case), the classical modified log-Sobolev approach developed by Massart in [34] 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 Markovian cases, their 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 Clémencon [8]. The variance terms in the Bernstein's type inequalities are 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]. It is a drawback for statistical applications for whom the variance term of the iid case is essential. To bypass this problem, many authors assume contractions conditions on the kernel of Markov chains, see Marton [31] under geometric ergodicity and Lezaud [29] under a spectral gap condition. For symmetric Markov process, the spectral gap condition is more general than uniform ergodicity and is also necessary for Bernstein's inequality, see Guillin *et al.* [23].

Many classical models in time series analysis do not satisfy such conditions. Fortunately, the classical Bernstein's inequality also holds for non Markovian processes under $\tilde{\Gamma}$-weakly dependent conditions, closely related with the uniform mixing condition, see Samson [39]. 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 [15]. For other metrics, non optimal couplings are used by Marton [33] and Djellout *et al.* [16] to extend classical dimension free transport inequalities $T_{2,d}(C)$ in a dependent context for metrics d different than the Hamming one. 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 [25]. Otherwise,a tradeoff must be done between the estimates of the variance terms and the accuracy of the coupling schemes, see Wintenberger [43] for details. The fast rates of convergence in oracle inequalities are not achieved because the variance terms of the Bernstein's types inequalities do not have the same order than in the iid case. On the contrary, the Hoeffding inequality is easily extended to very general dependent case using the bounded difference method, see McDiarmid [36], Rio [37] and Djellout *et al.* [16]. Unfortunately, the Hoeffding inequality, equivalent

to the $T_1(C)$ transport inequality, is not dimension free. Thus, this probabilistic inequality yields oracle inequalities with slow rates of convergence of order $n^{-1/2}$, see Alquier and Wintenberger [3].

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

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

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 $\mathcal{K}(Q|P)$ is the relative entropy $Q[\log(dP/dQ)]$ (also called the Kullback-Leibler divergence). As the roles of P and Q are not the same, we also introduce $\tilde{T}^{(i)}_{p,d}(C)$ where P and Q are interchanged in the left hand side term. An application of Sion's minimax theorem shows that the weak transport inequalities are extensions of the transport inequalities introduced by Marton [32]

$$\inf_{\pi} \sup_{\alpha} \frac{\pi[\alpha(Y) 1_{X \ne Y}]}{(Q[\alpha(Y)^q])^{1/q}} \le \sqrt{2C\mathcal{K}(Q|P)}.$$

These inequalities are weakened forms of the classical $T_{p,d}(C)$ transport inequality

$$W_{p,d}(P,Q) := \inf_{\pi} \pi[d^p(X,Y)]^{1/p} \le \sqrt{2C\mathcal{K}(Q|P)}.$$

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

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

The main assumptions hold on the conditional laws $P_{|x^{(i)}}$ of $(X_{i+1}, \dots, X_n)$ given that $(X_i, \dots, X_0) = x^{(i)} = (x_i, \dots, x_0)$. Fix a lower semi-continuous auxiliary metric d' satisfying $d' \ge Md$. We assume the existence of a coupling scheme $\pi_{|i}$ of the E^{n-i}-supported measures $(P_{|x^{(i)}}, P_{|y_i, x^{(i-1)}})$ such that

$$\pi_{|i}(d^p(X_k, Y_k))^{1/p} \le \gamma_{k,i}(p)\, d'(x_i, y_i), \qquad \forall\, i < k \le n.$$

The existence of such coefficients $\gamma_{k,i}(p)$ for any $1 \le i < k \le n$ is called the $\Gamma(p)$-weakly dependent condition. When $d = d'$ is the Hamming distance, the $\Gamma(2)$-weak dependance coincides with the weak dependence already studied by Samson [39] and we recover his results. We keep the notation of [39] and denote $\tilde{\Gamma}(p)$ the weak dependence when d is the Hamming distance. However, to deal with more general and classical time series, we prefer to choose d as the euclidian norm, see Section 4. When $p = 1$ and $\tilde{T}_{1,d}(C) = \tilde{T}^{(i)}_{1,d}(C) = T_{1,d}(C)$ by definition, the $\Gamma(1)$-weak dependence coincides with the one of [37] when d' is the Hamming distance and the one of [16] when $d' = d$. Thus we recover the Hoeffding's inequalities of [37, 16]. They are not dimension free because $n^{2/p-1} = n$ as $p = 1$ and we prefer to focus our study on the case $p = 2$ and d the euclidian norm which is new.

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. If P satisfies $\tilde{T}_{2,d}(C)$ 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 $f(x)-f(y)\le\sum_{j=1}^n L_j(x)d(x_j,y_j)$ for any $x,y\in(\mathbb{R}^d)^n$ we have

$$P\Big[\exp\Big(\lambda(f-P[f])-\frac{C\lambda^2}{2}\sum_{j=1}^n L_j^2\Big)\Big]\le 1,\qquad \lambda>0. \tag{1.2}$$

When d is the Hamming distance, inequality (1.2) yields to the Bernstein inequality, see Ledoux [28] in the independent setting and Samson [39] in the $\tilde{\Gamma}(2)$-weakly dependent setting. When the function f is a convex function, the condition above is automatically satisfied with $L_j=\partial_j f$ (the sub-gradients) and d the euclidian norm. The inequality (1.2) coincides with generalizations of the Tsirel'son inequality given in [41] (also implied by the T_2 transport inequality, see Bobkov *et al.* [10]). For convex functions that are also Lipschitz continuous on $[0,1]^n$, it extends for $\Gamma(2)$-weakly dependent measures the classical exponential inequality for products measures due to Talagrand [40].

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 [35]. The oracle inequalities are thus expressed as conditional weak transport inequalities. 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 $R(\hat{\theta})\le(1+\eta)R(\overline{\theta})+\Delta_n\eta^{-1_{\eta\neq 0}}$ where $\eta\ge 0$, $\overline{\theta}$ is the oracle defined as $R(\hat{\theta})\le R(\theta)$ for all θ and Δ_n is the rate of convergence. Oracle inequalities are standard non asymptotic criteria for the efficiency of statistical estimators, see Massart [34]. If $\eta=0$ then the oracle inequality is said to be exact and otherwise it is non exact, see Lecué and Mendelson [27] for a discussion. The dimension free concentration properties obtained from the weak transport inequalities with $p=2$ 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. When d is the Hamming distance, we recover in the conditional weak transport inequalities the variance terms of the iid case. These variance terms play a crucial role through Bernstein's condition introduced by Bartlett and Mendelson [7] to obtain oracle inequalities with fast rates of convergence. Thus, when d is 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 follows: 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 measures 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 $1 \le p \le 2$ and $1/p + 1/q = 1$. Let us define the weak transport cost as

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

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

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

Note that $\tilde{W}$ is not symmetric and that $\tilde{W}_{p,d}(P,Q) = \tilde{W}_{p,d}(Q,P) = \tilde{W}_{\alpha,d}(P,Q) = \tilde{W}_{\alpha,d}(Q,P) = 0$ if $P = Q$. Note that $\alpha \in M^+$ and d are assumed to be lower semi-continuous such as the optimal transport in the weak transport cost definition exists, see for example [21]. 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 conditional on Y, i.e. $\pi_{x,z|y} = \pi_{x|y}\pi_{z|y}$.*

Proof. From the classical gluing Lemma, se for example the Villani's textbook [42], 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. □

The conditional independence in the gluing Lemma 2.1 is the main ingredient to prove the triangular inequality on $\tilde{W}_{p,d}$:

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

$$\tilde{W}_{p,d}(P,R) \le \tilde{W}_{p,d}(P,Q) + \tilde{W}_{p,d}(Q,R) \tag{2.3}$$

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

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

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

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

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

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

Note that $Q[\tilde{\alpha}^q] = Q[\pi^*_{z|y}[\alpha(Z)|Y]^q] \le R[\alpha^q]$ 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 $\pi^*[\alpha(Z)d(X,Y)] \le R[\alpha^q]\tilde{W}_{p,d}(P,Q)$. We obtain

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

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

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 Rüschendorf in [38]. Similar couplings are used by Marton in [32]. 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 conditional 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

$$\pi_{x_1,y_1,z_1}=\pi_{x_1|y_1}\pi_{z_1|y_1}\pi_{y_1}, \tag{2.4}$$

and

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

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 conditional 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 conditional 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 conditional 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 conditional 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 idenity (2.5). Thus the result is proved:

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

□

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

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

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

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

for any fixed $\alpha = (\alpha_j)_{1 \leq j \leq n} \in M^+(E^n)$. Considering Markov couplings, we 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, $\tilde{W}_{\alpha,d}$ 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 $Q[\tilde{\alpha}_j(Y)]^q \leq R[\alpha_j^q(Z)]$ fo any $1 \leq j \leq n$ and*

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

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

$$\tilde{W}_{p,d}(P,R) \leq \tilde{W}_{p,d}(P,Q) + \tilde{W}_{p,d}(Q,R) \tag{2.9}$$

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

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

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

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

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. □

3. Weak transport inequalities

3.1. **Weak transport inequalities.** Let us say that the probability measure P on E^n, $n \geq 1$, satisfies the weak transport inequality $\tilde{T}_{p,d}(C)$ when for all distribution Q on E^n we have

$$\tilde{W}_{p,d}(P,Q) \leq \sqrt{2C\mathcal{K}(Q|P)}. \tag{3.1}$$

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

$$\tilde{W}_{p,d}(Q,P) \leq \sqrt{2C\mathcal{K}(Q|P)}. \tag{3.2}$$

By an application of Jensen's inequality, P satisfies $\tilde{T}_{p,d}(C)$ and $\tilde{T}^{(i)}_{p,d}(C)$ as soon as $\tilde{T}_{p',d}(C)$ and $\tilde{T}^{(i)}_{p',d}(C)$ reciprocally with $p' \geq p$. Moreover $\tilde{T}_{1,d}(C) = \tilde{T}^{(i)}_{1,d}(C) = T_{1,d}(C)$ where $T_{p,d}(C)$ is the classical transport inequality defined by the relation

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

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

Theorem 3.1.

1. *Any $P \in M(E)$ satisfies $\tilde{T}_{2,d}(1)$ and $\tilde{T}^{(i)}_{2,d}(1)$ when d is the Hamming distance $d(x,y) = 1_{x \neq y}$.*
2. *Any $P \in M(E)$ satisfies $\tilde{T}_{2,d}(D^2)$ and $\tilde{T}^{(i)}_{2,d}(D^2)$ 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. Alternative proofs are given in [32, 39]. The constant 1 is optimal, see the discussion in Section 5.3. For more involved examples of measures on discrete state spaces satisfying the weak transport $\tilde{T}_{2,d}$ we refer the reader to Gozlan *et al.* [22].

Remark 3.2. By definition every P satisfying the classical transport inequality $T_{2,d}(C)$ such that gaussian or log-concave measure satisfies also $\tilde{T}_{2,d}(C)$ and $\tilde{T}^{(i)}_{2,d}(C)$. However, any distribution having a support with finite diameter satisfies $\tilde{T}_{2,d}(C)$ but not necessarily $\tilde{T}_{2,d}(C)$. For any metric d the weak transport inequalities $\tilde{T}_{2,d}(C)$ and $\tilde{T}^{(i)}_{2,d}(C)$ have dual forms given below in (3.3) and (3.4). These expression 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 $\tilde{W}_{\alpha,d}$ for $\alpha \in M^+(E)$ fixed we have

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

where $f_{\alpha,d}(y) = \inf_x\{\alpha(y)d(x,y) + f(x)\}$. Then a measure P satisfies $\tilde{T}_{2,d}(C)$ if for any $\alpha \in M^+(E)$ and any probability measure Q

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

Thus P satisfies $\tilde{T}_{2,d}(C)$ if for any measure Q it holds

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

By the variational form of the entropy we obtain

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

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

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

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

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

$\tilde{T}_{2,d}(1)$ follows by taking $X = \lambda f$. To prove that $\tilde{T}^{(i)}_{2,d}(1)$ holds we start from its dual form. For equivalent reasons than the preceding dual form (3.3), our weak transport inequality $\tilde{T}^{(i)}_{2,d}(C)$ holds for any $C > 0$ iff

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

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, $n \geq 2$.** We use a new coupling technique based on the following $\Gamma_{d_1,d_2}(p)$-weak dependence condition of any law 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 $x^{(i)} = (x_i, \ldots, x_0)$ for $i \geq 0$ and $P_{|x^{(i)}}$ the conditional laws of $(X_{i+1}, \ldots, X_n)$ given that $(X_i, \ldots, X_0) = x^{(i)} = (x_i, \ldots, x_0)$. Let d_1 and d_2 be two lower semi-continuous distances on E such that $d_1 \leq M d_2$ for some $M > 0$. Let us work under the following weak dependence assumption:

Definition 3.1. For any $1 \leq p \leq 2$, the probability measure P is $\Gamma_{d_1,d_2}(p)$-weakly dependent if for any $1 \leq i \leq n$, any $(x^{(i)}, y_i) \in E^{i+2}$ there exists a coupling scheme $\pi_{|i}$ of $(P_{|x^{(i)}}, P_{|x^{(i-1)},y_i})$ and coefficients $\gamma_{k,i}(p) \geq 0$ such that

$$W_{p,d_1}(P_{x_k|x^{(i)}}, P_{x_k|x^{(i-1)},y_i}) \leq \gamma_{k,i}(p)\, d_2(x_i, y_i), \qquad \forall i < k \leq n. \tag{3.5}$$

Let us denote

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

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. *For any $1 \leq p \leq 2$, if P is $\Gamma_{d_1,d_2}(p)$-weakly dependent and $P_{x_j|x^{(j-1)}}$ satisfies $\tilde{T}_{p,d_1}(C)$ or $\tilde{T}^{(i)}_{p,d_1}(C)$ for all $1 \leq j \leq n$ then P satisfies $\tilde{T}_{p,d_1}(C\|\Gamma(p)\|_p^2 n^{2/p-1})$ or $\tilde{T}^{(i)}_{p,d_1}(C\|\Gamma(p)\|_p^2 n^{2/p-1})$ 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 $\|A\|_p \le \|A\|_1^{1/p}\|A\|_\infty^{1-1/p}$ and the fact that $\|\Gamma\|_1 = \|\Gamma\|_\infty = M + \sum_{i=1}^n \gamma_{i,0}(p)$, then $\|A\|_p \le M + \sum_{i=1}^n \gamma_{i,0}(p)$.

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 $1 \le j \le n$. As preliminaries, we recall the following result of existence of the optimal Markov coupling due to from Rüschendorf, [38] and a simple and useful consequence 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 [38])**.** *We have the equivalence between (1) and (2)*

(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) $\inf_{\pi_1} \pi_1[h] = \pi_1^*[h] < \infty$.

Denote $\alpha_j^{(i)}$ denotes the section of α_j in $y^{(i)}$ as $\alpha_j^{(i)}(y_{i+1}, \ldots, y_n) = \alpha_j(y)$ and $\alpha^{(i)} = (\alpha_j^{(i)})_{j>i}$. 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*

$$\tilde{W}_{\alpha,d_1}(P,Q) \le \pi_1[Q_{|Y_1}[\alpha_1|Y_1]d_1(X_1,Y_1) + \tilde{W}_{\alpha^{(1)},d_1}(P_{|X_1}, Q_{|Y_1})]. \tag{3.6}$$

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

$$\pi^*_{|x_1,y_1}\Big[\sum_{j=2}^n \alpha_j^{(1)} d_1(X_j, Y_j)\Big] = \tilde{W}_{\alpha^{(1)},d_1}(P_{|x_1}, Q_{|y_1})].$$

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

$$\pi^*_{|x_1,y_1}[\alpha_1^{(1)} d_1(x_1,y_1)] = \pi^*_{|x_1,y_1}[\alpha_1^{(1)}|x_1,y_1]d_1(x_1,y_1) = Q_{|y_1}[\alpha_1|y_1]d_1(x_1,y_1)$$

by definition of Markov couplings. □

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

$$\begin{aligned}\tilde{\pi}_{j|j-1}\Big[\sum_{k=j}^n Q_{|Y_j,y^{(j-1)}}[\alpha_k^q|Y_j, y^{(j-1)}]^{1/q}\gamma_{k,j}(p)d_2(X_j,Y_j)\Big] \\ = \Big(\sum_{k=j}^n Q_{|y^{(j-1)}}[\alpha_k^q|y^{(j-1)}]\gamma_{k,j}(p)^q\Big)^{1/q}\tilde{W}_{p,d_2}(P_{x_j|x^{(j-1)}}, Q_{y_j|y^{(j-1)}})\end{aligned} \tag{3.7}$$

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

We are now ready to prove the result iterating several times the same reasoning. Let us detail the case $j = 1$ when considering probabilities conditional on y_0. Applying (3.6) and (2.8) we have

$$\tilde{W}_{\alpha,d_1}(P,Q) \le \tilde{\pi}_{1|y^{(0)}}[Q_{|Y_1,y^{(0)}}[\alpha_1|Y_1,y^{(0)}]d_1(X_1,Y_1) + \tilde{W}_{\alpha^{(1)},d_1}(P_{|X_1,y^{(0)}}, Q_{|Y_1,y^{(0)}})]$$

$$\leq \tilde{\pi}_{1|y^{(0)}}[Q_{|Y_1,y^{(0)}}[\alpha_1|Y_1,y^{(0)}]d_1(X_1,Y_1)+\tilde{W}_{\alpha^{(1)},d_1}(P_{|Y_1,y^{(0)}},Q_{|Y_1,y^{(0)}}) \\ +\tilde{W}_{\tilde{\alpha}^{(1)},d_1}(P_{|X_1,y^{(0)}},P_{|Y_1,y^{(0)}})]. \tag{3.8}$$

To bound the last term, we use the definition of the $\Gamma_{d_1,d_2}(p)$-weak dependence:

Lemma 3.5. *For any $\alpha_k \in M^+(E)$ for all $j<k\leq n$ and any $\Gamma_{d_1,d_2}(p)$-weakly dependent probability measure P we have*

$$\tilde{W}_{\alpha^{(j)},d_1}(P_{|x_j,y^{(j-1)}},P_{|y^{(j)}}) \leq \sum_{k=j+1}^{n} Q_{|y^{(j)}}[\alpha_k^q|y^{(j)}]^{1/q}\gamma_{k,j}(p)d_2(x_j,y_j) \tag{3.9}$$

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

$$\begin{aligned}
\tilde{W}_{\alpha^{(j)},d_1}(P_{|x_j,y^{(j-1)}},P_{|y^{(j)}}) &= \inf_{\pi_{|j}}\pi_{|j}\Big[\sum_{k=j+1}^{n}\alpha_k d_1(X_k,Y_k)\Big] \\
&\leq \inf_{\pi_{|j}}\sum_{k=j+1}^{n}\pi_{|j}[\alpha_k^q|y^{(j)}]^{1/q}\pi_{|j}[d_1^p(X_k,Y_k)]^{1/p} \\
&\leq \inf_{\pi_{|j}}\sum_{k=j+1}^{n}Q_{|y^{(j)}}[\alpha_k^q|y^{(j)}]^{1/q}\pi_{|j}[d_1^p(X_k,Y_k)]^{1/p} \\
&\leq \sum_{k=j+1}^{n}Q_{|y^{(j)}}[\alpha_k^q|y^{(j)}]^{1/q}\gamma_{k,j}(p)d_2(x_j,y_j)
\end{aligned}$$

because the $\Gamma(p)$-weak dependence condition ensures the existence of a coupling scheme satisfying

$$\pi_{|j}[d_1^p(X_k,Y_k)]^{1/p}\leq\gamma_{k,j}(p)d_2(x_j,y_j) \qquad \forall\, j<k\leq n.$$

□

Denoting $\gamma_{i,i}=M$ for all $1\leq i\leq n$, note that by assumption we have the relation$d_1(X_i,Y_i)\leq\gamma_{i,i}d_2(X_i,Y_i)$. Collecting the bounds (3.8) and (3.9) we obtain

$$\tilde{W}_{\alpha,d_1}(P_{|y^{(0)}},Q_{|y^{(0)}}) \leq \tilde{\pi}_{1|y^{(0)}}\Big[\sum_{k=1}^{n}Q_{|Y^{(1)},y_0}[\alpha_k^q|Y^{(1)},y_0]^{1/q}\gamma_{k,1}d_2(X_1,Y_1) \\ +\tilde{W}_{\alpha^{(1)},d_1}(P_{|Y_1,y^{(0)}},Q_{|Y_1,y^{(0)}})].$$

Let us do the same reasoning than above for any $1\leq j\leq n$ conditional on $y^{(j)}$ on $\tilde{W}_{\alpha^{(j-1)},d_1}(P_{|y^{(j)}},Q_{|y^{(j)}})$. For any $1\leq j\leq n$, we obtain:

$$\tilde{W}_{\alpha^{(j-1)},d_1}(P_{|y^{(j-1)}},Q_{|y^{(j-1)}}) \leq \tilde{\pi}_{j|y^{(j-1)}}\Big[\sum_{k=j}^{n}Q[\alpha_k^q|Y_j,y^{(j-1)}]^{1/q}\gamma_{k,j}(p)d_2(X_j,Y_j) \\ +\tilde{W}_{\alpha^{(j)},d_1}(P_{|Y_j,y^{(j-1)}},Q_{|Y_j,y^{(j-1)}})].$$

For the specific Markov coupling considered here, the identity (3.7) holds and

$$\begin{aligned}
&\tilde{W}_{\alpha^{(j-1)},d_1}(P_{|y^{(j-1)}},Q_{|y^{(j-1)}}) \\
&\quad\leq\Big(\sum_{k=j}^{n}Q_{|y^{(j-1)}}[\alpha_k^q|y^{(j-1)}]\gamma_{k,j}(p)^q\Big)^{1/q}\tilde{W}_{p,d_2}(P_{x_j|y^{(j-1)}},Q_{y_j|y^{(j-1)}}) \\
&\qquad+\tilde{\pi}_{j|y^{(j-1)}}[\tilde{W}_{\alpha^{(j)},d_1}(P_{|Y_j,y^{(j-1)}},Q_{|Y_j,y^{(j-1)}})] \\
&\quad\leq\sum_{k=j}^{n}Q_{|y^{(j-1)}}[\alpha_k^q|y^{(j-1)}]^{1/q}\gamma_{k,j}(p)\tilde{W}_{p,d_2}(P_{x_j|y^{(j-1)}},Q_{y_j|y^{(j-1)}})
\end{aligned}$$

$$+\tilde{\pi}_{j|y^{(j-1)}}[\tilde{W}_{\alpha^{(j)},d_2}(P_{|Y_j,y^{(j-1)}},Q_{|Y_j,y^{(j-1)}})]$$

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

$$\begin{aligned}\tilde{W}_{\alpha,d}(P,Q) &\leq Q\Big[\sum_{j=1}^n\sum_{k=j}^n Q[\alpha_k^q|Y^{(j-1)}]^{1/q}\gamma_{k,j}(p)\tilde{W}_{p,d_2}(P_{x_j|Y^{(j-1)}},Q_{y_j|Y^{(j-1)}})\Big]\\ &\leq \sum_{j=1}^n\sum_{k=j}^n Q[\alpha_k^q]^{1/q}\gamma_{k,j}(p)Q[\tilde{W}_{p,d_2}(P_{x_j|Y^{(j-1)}},Q_{y_j|Y^{(j-1)}})^p]^{1/p}\\ &\leq \sum_{j=1}^n\sum_{k=j}^n Q[\alpha_k^q]^{1/q}\gamma_{k,j}(p)Q[2C\mathcal{K}(Q_{y_j|Y^{(j-1)}}|P_{x_j|Y^{(j-1)}})^{p/2}]^{1/p}\end{aligned}$$

the second inequality following from Hölder's and Jensen's inequalities and the last one from the assumption $P_{x_j|y^{(j-1)}} \in \tilde{T}_{p,d}(C)$. 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 $<;>$ denoting the scalar product, we obtain

$$\tilde{W}_{\alpha,d}(P,Q) \leq < \mathbf{Q};\Gamma(p)\mathbf{W} > \leq \|\mathbf{Q}\|_q\|\Gamma(p)\|_p\|\mathbf{W}\|_p.$$

Note that we have the identities

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

As $p/2 \leq 1$, successive applications of Jensen's and Holder's inequalities yield

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

Finally, we obtain

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

The desired result follows by taking the supremum over all $\alpha \in M^+(E^n)$. □

4. Examples of $\Gamma_{d_1,d_2}(p)$-weakly dependent processes

4.1. **$\Gamma_{d_1,d_2}(1)$-weakly dependent examples.** When $p = 1$, the dual form of $\tilde{T}_{1,d}(C) = \tilde{T}^i_{1,d}(C) = T_{1,d}(C)$ is the Hoeffding inequality. Moreover we apply the Kantorovitch-Rubinstein inequality to obtain an explicit expression of the $\gamma_{k,i}(p)$ coefficients:

$$\sum_{k=i+1}^n \gamma_{k,i}(1)$$

$$= \sup_{f\in \mathrm{Lip}_1(i)} \sup_{x^{(i)},y_i} \frac{P[f(X_{i+1},\dots,X_n)|x^{(i)}] - P[f(X_{i+1},\dots,X_n)|y_i,x^{(i-1)}]|}{d_2(x_i,y_i)}$$

where $\mathrm{Lip}_1(i)$ is the set of Lipschitz functions f satisfying

$$|f(x)-f(y)| \le \sum_{j=i+1}^{k} d_1(x_j,y_j), \quad \forall x=(x_{i+1},\dots,x_n),\ y=(y_{i+1},\dots,y_n)\in E^{n-i}.$$

In the bounded case $d_1 \le M$ and $d_2(x,y)=1_{x\neq y}$ is the Hamming distance, the Γ_{d_1,d_2}-weak dependence condition coincides with the one introduces in Rio [37]. As the conditional probabilities $P_{x_j|x^{(j-1)}}$ automatically satisfy Pinsker's inequality $\tilde{T}_{1,d_2}(1/4)$, Theorem 3.2 recovers the Hoeffding inequality of [37].

The context of $\Gamma_{d_1,d_2}(1)$-weak dependence is very general for d_1 the distance associated with the euclidian norm and d_2 the Hamming distance. We refer the reader to Section 7 of [14] for a detailed study of many examples in this case, including causal functions of stationary sequences, iterated random functions, Markov kernels and expanding maps.

When $d_1=d_2=d$ for any metric d, the Γ_{d_1,d_2}-weak dependence condition coincides with the condition $(C_1)'$ of [16]: for any 1-Lipschitz function f it holds

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

From Remark 3.3 we have $\|\Gamma(1)\|_1 \le 1+S$ and thus Theorem 3.2 recovers the Hoeffding inequality of [16]. Examples of $\Gamma_{d,d}(1)$-weakly dependent time series are given in [16]. In particular, ARMA processes with sub-gaussian or log-concave innovations satisfy the conditions of Theorem 3.2 for $p=1$. Thus they satisfy an Hoeffding's type inequality (which is not dimension free).

4.2. **$\tilde{\Gamma}(p)$-weakly dependent examples.** When $d_1=d_2$ is the Hamming distance $1_{x\neq y}$, let us denote $\Gamma_{d_1,d_2}=\tilde{\Gamma}$. For this choice of metrics the optimal coupling scheme is the maximal coupling given in [20]. Then we have

$$\tilde{\gamma}_{k,i}(p) = \sup_{x^{(i)},y_i} \|P_{|x^{(i)}} - P_{|y_i,x^{(i-1)}}\|_{TV}^{1/p}$$

where $\|P-Q\|_{TV}=\sup_A |P(A)-Q(A)|$ for any distributions P et Q. The $\tilde{\Gamma}(p)$ weakly dependent condition coincides with the ones used by Samson [39] for $p=2$ and by Rio [37] and Kontorovitch and Ramanan [26] for $p=1$.

For Markov chains, the $\tilde{\Gamma}(p)$ weakly dependent condition is equivalent to the uniform ergodicity condition. In the stationary case, $\tilde{\gamma}_{k,i}^p \le 2\phi_{k-i}$ where ϕ is the uniform mixing coefficient introduced by Ibragimov [24]. For $p=2$, we recover the transport inequality obtained by Samson [39] as $\tilde{T}_{1,1_{x\neq y}}(1)$ holds for any distribution and by an application of Theorem 3.2 we obtain

$$\inf_{\pi\in\tilde{M}} \Big(\sum_{i=1}^n Q[\pi[X_i\neq Y_i \mid Y_i]^2\Big)^{1/2} \le \|\tilde{\Gamma}(2)\|_2 \sqrt{2\mathcal{K}(Q|P)}.$$

Note that we use here the minimax theorem of Sion and the Proposition 1 of [32] to obtain the identities:

$$\tilde{W}_{2,1_{x\neq y}}(P,Q) = \inf_{\pi\in\tilde{M}} \sup_{\alpha_j>0} \frac{\sum_{j=1}^n \pi[\alpha_j(Y)1_{X_j\neq Y_j}]}{(\sum_{j=1}^n Q[\alpha_j(Y)^2])^{1/2}} = \inf_{\pi\in\tilde{M}} \Big(\sum_{i=1}^n Q[\pi[X_i\neq Y_i \mid Y_i]^2\Big)^{1/2}.$$

Note also that this weak transport inequality yields dimension free concentration properties in term of the convex distance as it is done in Talagrand in [40] or in

Marton in [31].

In the stationary case, any ϕ-mixing processes are $\tilde{\Gamma}$-weakly dependent with $\|\tilde{\Gamma}(p)\|_p \le 1 + \sum_{i=1}^n (2\phi_i)^{1/p}$ for any $1 \le p \le 2$, see [39]. But the $\tilde{\Gamma}(p)$-weakly dependence is also satisfied for non stationary sequences, see [26]. However, when E is a real vector space, the choice of the Hamming distance is not natural and the resulting weakly dependent conditions are often too restrictive.

4.3. $\Gamma_{\|\cdot\|,d_2}(2)$**-weakly dependent exemples.** In what follows, we focus on the choice $d_1 = \|\cdot\|$ the euclidian norm that is natural in many examples in $E = \mathbb{R}^k$. We also fix $p = 2$ in order to obtain dimension free transport inequalities when applying Theorem 3.2. We focus on two generic examples: the stochastic recurrent equations already treated in [16] when $p = 1$ and the chains with infinite memory introduced in [17]. In each example, we use the natural coupling provided by the structure of the example to estimate the coefficients $\gamma_{k,i}$. Note that an explicit expression of these coefficients is not required.

Example 4.1 (Stochastic Recurrent Equations (SREs))**.** Consider the SRE (also called Iterated Random Functions in [18] and Random Dynamical Systems in [16])

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

where (ψ_t) is a sequence of i.i.d. random maps. Let us denote here P the probability of the whole process $(\psi_t)_{t\ge1}$. Assume in the next proposition in full generality that d_1 and d_2 are any semi-lower continuous metrics satisfying $d_1 \le M d_2$ for some $M > 0$.

Proposition 4.1. *For any $1 \le p \le 2$, if the distribution of $\psi_1(x)$ belongs to $\tilde{T}_{p,d_2}(C)$ or $\tilde{T}^{(i)}_{p,d_2}(C)$ for any $x \in E$ and if there exists some $S > 0$ satisfying*

$$\sum_{t=1}^{\infty} P[d_1^p(X_t(x), X_t(x'))]^{1/p} \le S d_2(x, x') \qquad \forall x, x' \in E. \tag{4.2}$$

then for any $x \in E$ we have that the law P_x^n of $(X_t(x))_{1\le t\le n}$ on E^n satisfies $\tilde{T}_{p,d_1}(C(M+S)^2 n^{2/p-1})$ or $\tilde{T}^{(i)}_{p,d_1}(C(M+S)^2 n^{2/p-1})$ respectively.

Proof. The result is proved by an application of Theorem 3.2. The condition of Γ_{d_1,d_2}-weak dependence is satisfied because the joint law of $(X_t(x), X_t(x))_{t\ge1}$ is a natural coupling scheme $\pi_{|0}$ of the law of $(X_t)_{t\ge1}$ given that $(X_0, X_{-1}, X_{-2}\ldots) = (x, x_{-1}, x_{-2}, \ldots)$ and $(X_0, X_{-1}, X_{-2}\ldots) = (x', x_{-1}, x_{-2}, \ldots)$. We obtain similarly natural coupling schemes $\pi_{|i}$ for any $i \ge 0$ and the coefficients $\gamma_{k,i}$ satisfy the relation $\sum_{k>i} \gamma_{k,i}(p) \le S$. The fact that the relation $\sum_{k<n} \gamma_{n,k}(p) \le S$ also holds for any $n \ge 2$ follows from the exchangeability of $(\psi_1, \ldots, \psi_n)$. Using similar arguments than in Remark 3.3, we obtain that $\|\Gamma(p)\|_p \le M + S$. The result is proved as, by the Markov property, $P_{x_j|x^{(j-1)}}$ is the law of $\psi_j(x_{j-1})$ that satisfies $\tilde{T}_{p,d_2}(C)$ or $\tilde{T}^{(i)}_{p,d_2}(C)$ by assumption. □

Let us detail two classical SREs, the ARMA models and the general affine processes when $d_1 = d_2$ equals the euclidian norm $\|\cdot\|$.

Example 4.2 (ARMA models)**.** Consider the ARMA model

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

in $E = \mathbb{R}^k$ where $A \in \mathcal{M}_{k,k}$ (the space of $k \times k$ matrices) and (ξ_t) is a sequence of i.i.d. random vectors in $\mathbb{R}^k$ called the innovations. This model is a particular case

of the general model above with $\psi_t(x) = Ax + \xi_t$. The $\Gamma_{\|\cdot\|,\|\cdot\|}(p)$-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). The conditions of Proposition 4.1 are satisfied if the law of ξ_1 satisfies $\tilde{T}_{p,\|\cdot\|}(C)$ or $\tilde{T}^{(i)}_{p,\|\cdot\|}(C)$. It is in particular the case with $p = 2$ for bounded, gaussian or log-concave innovations ξ_t.

For instance, the stationary solution of $X_{t+1} = 1/2X_t + \xi_{t+1}$ with $\xi_1 \sim \mathcal{B}(1/2)$ is such that the distributions of $(X_1, \dots, X_n)$ satisfy the dimension free inequalities $\tilde{T}_{2,\|\cdot\|}(C)$ and $\tilde{T}^{(i)}_{2,\|\cdot\|}(C)$ for any $n \geq 1$. However, this process is not $\tilde{\Gamma}(2)$-weakly dependent, see [4].

Example 4.3 (General affine processes)**.** Consider now the specific SRE

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

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

(a) $P_\xi \in \tilde{T}_{2,\|\cdot\|}(C)$ or $\tilde{T}^{(i)}_{2,\|\cdot\|}(C)$ on $\mathbb{R}^{k'}$;
(b) $\|M(x)\| \leq K$, $\forall x \in \mathbb{R}^k$, $K > 0$, $\|\cdot\|$ denoting also the operator norm on $\mathcal{M}_{k,k'}$ associated with the euclidian norms on $\mathbb{R}^k$ and $\mathbb{R}^{k'}$;
(c) the Lyapunov exponent in L^2 satisfies

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

Using a version of Lemma 2.1 in [16] we obtain that conditions (1) and (2) implies that $P_{x_i|x_{i-1}} \in \tilde{T}_{2,\|\cdot\|}(CK^2)$ or $\tilde{T}^{(i)}_{2,\|\cdot\|}(CK^2)$. Moreover condition (4.2) is satisfied for some $S > 0$ and thus P^n_x satisfies $\tilde{T}_{2,\|\cdot\|}(CK^2(1+S)^2)$ or $\tilde{T}^{(i)}_{2,\|\cdot\|}(CK^2(1+S)^2)$ for any $x \in E$. We answer positively to a question raised in Remark 3.6 in [16].

Example 4.4 (Chains with Infinite Memory)**.** Here assume that $d_1 = d_2 = d$ is any semi lower-continuous distance. Consider chains with infinite memory introduced by Doukhan and Wintenberger [17] for any function $F : E^{\mathbb{N}} \times \mathcal{X} \mapsto E$ by the relation:

$$X_t(x) = F(X_{t-1}, X_{t-2}, \dots, X_1, x_0, x_{-1}, x_{-2}, \dots; \xi_t), \qquad \forall t \geq 1, \tag{4.3}$$

for any sequence $x = (x_{-t})_{t\geq 0} \in E^{\mathbb{N}}$ and any iid innovations ξ_t on some measurable space $\mathcal{X}$. This model does not satisfy the Markov property. However, it still exists a natural coupling scheme of the law of $(X_t)_{t\geq 1}$ given that $(X_0, X_{-1}, X_{-2} \dots) = x$ and $(X_0, X_{-1}, X_{-2} \dots) = (y_0, x_{-1}, x_{-2}, \dots)$. Indeed, define recursively the trajectory $(Y_t)_{t\geq 1}$ by the relation

$$Y_t = F(Y_{t-1}, Y_{t-2}, \dots, Y_1, y_0, x_{-1}, x_{-2}, \dots; \xi_t), \qquad \forall t \geq 1,$$

where the innovations $(\xi_t)_{t\geq 1}$ are the same than in (4.3). Then the natural coupling scheme $\pi_{|0}$ is the distribution of $(X_t(x), Y_t)_{t\geq 1}$. Denote P the law of the innovations process (ξ_t), we have

Proposition 4.2. *Assume there exists a sequence of non negative numbers (a_i) such that $\sum_{i\geq 1} a_i = a < 1$ and*

$$P[d(F(x_1, x_2, \dots; \xi), F(y_1, y_2, \dots; \xi))^p]^{1/p} \leq \sum_{i\geq 1} a_i d(x_i, y_i), \tag{4.4}$$

for any $x=(x_1,x_2,\dots)$ and $y=(y_1,y_2,\dots)$ in $E^{\mathbb{N}}$. If the distribution of $F(y;\xi_1)$ satisfies $\tilde{T}_{p,d}(C)$ or $\tilde{T}^{(i)}_{p,d}(C)$ then for any $y\in E^{\mathbb{N}}$ we have that P_x^n satisfies $\tilde{T}_{p,d}(C^2(1-a)^{-2}n^{2/p-1})$ or $\tilde{T}^{(i)}_{p,d}(C^2(1-a)^{-2}n^{2/p-1})$ respectively, P_x^n being the law of $(X_t(x))_{1\le t\le n}$ on E^n.

Proof. Let us compute a bound for the coefficients $P[d^p(X_t(x),Y_t)]^{1/p}$ for all $t\ge 0$ that are estimates of the coefficients $\gamma_{t,0}$. Fix $w_0=1$ and define recursively

$$w_t \le \sum_{j=1}^{t} a_j w_{t-j}, \qquad j\ge 1.$$

Note that coupling schemes $\pi_{|i}$ can be constructed similarly than for the case $i=0$. By construction and by a recursive argument using (4.4) we have the relation $\gamma_{k,i}\le w_t$ for any $k-i\le t$. Denote $A(s)=\sum_{t\ge1}a_t s^t$ and $W(s)=\sum_{t\ge1}w_t s^t$ for any $|s|\ge 1$. From the relation $W(s)=1+A(s)W(s)$ we deduce that $W(1)=\sum_{t\ge1}w_t=(1-A(1))^{-1}=(1-a)^{-1}$. The desired result follows by an application of similar arguments to the ones at the end of the proof of Proposition 4.1. □

Remark that, under $a<1$ and $\mathbb{E}[d^p(F(y,\xi_1),z)]<\infty$ for any $y\in E^{\mathbb{N}}$ and $z\in E$, the existence of a stationary solution with moment of order p satisfying

$$X_t(x)=F(X_{t-1},X_{t-2},X_{t-3}\dots;\xi_t), \qquad \forall t\in\mathbb{Z}$$

is proved in [17]. If moreover the law of $F(x;\xi_1)$ satisfies $T_{p,d}(C)$ then P_x^n satisfies $T_{p,d}(C^2(1-a)^{-2}n^{2/p-1})$ by an application of Theorem 2.5 in [16].

Example 4.5 (AR(∞) models)**.** As an example of chains with infinite memory in $E=\mathbb{R}$, consider (X_t) the stationary solution to the autoregressive equation

$$X_t=\sum_{i\ge1}a_iX_{t-i}+\xi_t, \qquad t\in\mathbb{Z},$$

where the real numbers a_i are such that $a:=\sum_{i\ge1}|a_i|<1$. Then if ξ_1 satisfies $\tilde{T}_{2,|\cdot|}(C)$ or $\tilde{T}^{(i)}_{p,|\cdot|}(C)$ then the distribution of $(X_1,\dots,X_n)$ satisfies $\tilde{T}_{2,|\cdot|}(C^2(1-a)^{-2})$ or $T^{(i)}_{p,d_1}(C^2(1-a)^{-2})$ for any $n\ge1$. The same result holds when replacing the weak transport $\tilde{T}$ with the classical transport T.

Example 4.6 (General affine processes with infinite memory)**.** Consider the process on $E=\mathbb{R}^k$ defined as the solution of

$$X_t(x)=f(X_{t-1},X_{t-2},X_{t-3},\dots)+M(X_{t-1},X_{t-2},X_{t-3},\dots)\xi_t, \qquad \forall t\ge1$$

where f and M are Lipschitz continuous functions with value in $\mathbb{R}^k$ and $\mathcal{M}(k,k')$ respectively. These general affine models includes classical econometric models and is estimated in a parametric setting by the quasi maximum likelihood estimator in [6]. Denote for $\Psi=f$ and $\Psi=M$ the Lipschitz coefficients

$$\|\Psi(x)-\Psi(y)\|\le\sum_{i\ge1}\alpha_i(\Psi)\|x_i-y_i\|, \qquad \forall x,\ y\in E^{\mathbb{N}}.$$

If the condition (a) of Example 4.3 is satisfied and $\|M(x)\|\le K$, $\forall x\in E^{\mathbb{N}}$, $K>0$ and $a:=\sum_{i\ge1}\alpha_i(f)+P_\xi[\xi^2]^{1/2}\alpha_i(M)<1$ then the distribution of $(X_1,\dots,X_n)$ satisfies $\tilde{T}_{2,\|\cdot\|}(C^2(1-a)^{-2})$ or $T^{(i)}_{2,\|\cdot\|}(C^2(1-a)^{-2})$ for any $n\ge1$. That $P_\xi[\xi^2]$ is finite follows from the finiteness of the exponential moments of P_ξ implied by the weak transport inequalities, see the next section.

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 $L_j : E^n \mapsto \mathbb{R}_+$, $1 \le j \le n$ satisfying

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

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

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

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

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

$$P\Big[\exp\Big(\lambda(f - P[f]) - \frac{C\lambda^2}{2}\Big(\Big(2 - \frac{2}{p}\Big)\sum_{j=1}^{n} L_j^{\frac{p}{p-1}} + \Big(\frac{2}{p} - 1\Big)\Big)\Big)\Big] \le 1. \tag{5.3}$$

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

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

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

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

and by definition of $\tilde{W}$ we obtain

$$Q[f] - P[f] \le Q\Big[\sum_{j=1}^{n} L_j^q\Big]^{1/q} \tilde{W}_{p,d}(P, Q).$$

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

$$Q[(f - P[f])] \le Q\Big[\sum_{j=1}^{n} L_j^q\Big]^{1/q} \sqrt{2C\mathcal{K}(Q|P)}. \tag{5.5}$$

From the variational identity

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

we get for all $\lambda > 0$:

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

We can rewrite it as

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

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

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

Then the desired result follows from the variational formula of the entropy. □

5.2. **Extensions of classical concentration inequalities to dependent cases.** Let us recall the classical inequality obtained by Talagrand in the iid case [40]: let f be a separately convex Lipschitz function on $[0,1]^n$ then

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

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 uniformly ergodic Markov chains in Marton [31] and to $\tilde{\Gamma}(2)$-weakly dependent processes in Samson [39] for convex Lipschitz functions f. The extension to the more general $\Gamma_{\|\cdot\|,1_{x\neq y}}(2)$-weakly dependent context follows from Theorem 5.1 as P satisfies $\tilde{T}_{2,\|\cdot\|}(\|\Gamma(2)\|_2^2)$ and $\tilde{T}^{(i)}_{2,\|\cdot\|}(\|\Gamma(2)\|_2^2)$ choosing $d_1 = \|\cdot\|$ and d_2 the Hamming distance ($M = 1$). Note that with no loss of generality we can assume that f is smooth enough (see Samson [39] 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 $\Gamma_{\|\cdot\|,1_{x\neq y}}(2)$-weakly dependent measure P on E^n, for any convex L-Lipschitz function f we have*

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

This type of inequalities has a lot of applications in probability theory, see [40].

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.* [11].

Corollary 5.3. *Assume that* $|g(x)-g(y)| \le Ld(x,y)\ \forall x,y \in E$ *with* $L>0$. *If* P *satisfies* $\tilde{T}_{2,d}(C)$ *or* $\tilde{T}^{(i)}_{2,d}(C)$ *then for every* $t \ge 0$ *we have respectively*

$$P(f \ge P[f]+t) \le \exp\Big(-\frac{t^2}{8CL^2(P[f]+t)}\Big),$$
$$P(f \le P[f]-t) \le \exp\Big(-\frac{t^2}{8CL^2P[f]}\Big).$$

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

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

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

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

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

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

The desired inequality follows by the Chernoff argument. □

5.3. **The specific case of the Hamming distance.** We fix $d_1(x,y)=d_2(x,y)=1_{x\neq y}$ as in Samson [39]. Thus the result of this section holds for any $\tilde{\Gamma}(2)$-weakly dependent sequence (with bounded margins). Using exactly the same arguments than above, an extension of the classical exponential inequality (5.6) also holds in this more restrictive case, see [39]:

$$P(|f-P[f]| \ge t) \le 2\exp\Big(-\frac{t^2}{2\|\tilde{\Gamma}(2)\|_2^2 L^2}\Big) \qquad \forall t \ge 0.$$

The case of the Hamming distance is specific because for any non negative function f we can replace the convexity argument $x^2-y^2 \le 2x(x-y)$ by the simple inequality $f(x)-f(y) \le f(x)1_{x\neq y}$. 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 $\sum_{j=1}^n L_j^2 \le Mf$ and $-f$ satisfies (5.2) with $\sum_{j=1}^n {L_j^{(i)}}^2 \le Mf$. Applying Theorem 5.1 we recover the results of Theorem 2 of [39]:

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

$$P(f \ge P[f]+t) \le \exp\Big(-\frac{t^2}{2M\|\tilde{\Gamma}(2)\|_2^2(P[f]+t)}\Big),$$
$$P(f \le P[f]-t) \le \exp\Big(-\frac{t^2}{2M\|\tilde{\Gamma}(2)\|_2^2 P[f]}\Big).$$

The constant 1 in $\tilde{T}_{2,1_{x\neq y}}(1)$ or $\tilde{T}^{(i)}_{2,1_{x\neq y}}(1)$ is optimal in Theorem as discussed in Boucheron *et al.* [12] 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 $f(x)-f(y) \le f(x)1_{x\neq y}$, it is also possible to extend classical Bernstein's inequality in the $\tilde{\Gamma}(2)$-weakly dependent case:

Theorem 5.5 ([39] (page 460, line 7))**.** *Let g be a measurable function $\mathbb{R} \to [-M, M]$ and let*

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

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

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

This inequality is applied in [2] to obtain exact oracle inequality with fast rates in the $\tilde{\Gamma}(2)$-weakly dependent context under a so called Bernstein's condition that relates the variance term (similar than in the iid case) and the statistical risk.

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 new exact oracle inequalities in the $\tilde{\Gamma}(2)$-weakly dependent setting. Instead of using the inequalities given in the last Section we prefer to use a more direct approach using conditional weak transport inequalities.

6.1. **The statistical setting.** We focus on oracle inequalities for the the ordinary least square estimator. Let us consider the case of the linear regression where $X = (Y, Z) = (Y, Z^{(1)}, \dots, Z^{(d)})$ and $E = \mathbb{R}^{d+1}$ is 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 $(X_i)_{1\le i\le n} = (Y_i, Z_i)_{1\le i\le n}$ are the observations. In our context, these observations are not necessarily independent nor identically distributed and we denote by P their distribution. The risk of prediction is denoted

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

The aim is to estimate the value $\overline{\theta} \in \mathbb{R}^d$ such that $R(\overline{\theta}) \le R(\theta)$, $\forall \theta \in \mathbb{R}^d$. We consider the ordinary least square estimator $\hat{\theta}$ of $\overline{\theta}$ such that $r(\hat{\theta}) \le r(\theta)$ for all $\theta \in \mathbb{R}^d$. Let us denote the excess of risk $\overline{R}(\theta) = R(\theta) - R(\overline{\theta}) \ge 0$, $\overline{r}$ its empirical counterpart, $\mathcal{Z} = (Z_i)_{1\le i\le n}$ 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 $T_{2,d}(C)$ and $T_{2,d}^{(i)}(C)$ for some $C > 0$ on E^n with d issued from the euclidian norm or the Hamming distance. In view of Theorem 3.2 and for applications perspective in time series we are interested in $\Gamma_{\|\cdot\|,d_2}(2)$ 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 $var\phi(Y_1, \dots, Y_{i-1})$. Here the function φ is known and one can think of the projection on the last coordinates (case of linear autoregression), functions on Fourier basis or wavelets, etc. 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 $\ell \ge 1$ fixed, we have $\gamma(2)_{k,0} \le \gamma(2)_{\lceil k/\ell\rceil,0}$ and in the non-linear autoregressive case, $\tilde{\gamma}(2)_{k,0} \le \|\varphi\|_\infty \tilde{\gamma}(2)_{\lceil k/\ell\rceil,0}$. 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 $\Gamma_{\|\cdot\|,d_2}(2)$-weakly dependent sequences.** Our first result is a bound on the excess of risk for $\Gamma_{\|\cdot\|,d_2}(2)$-weak dependence where $d_2 = \|\cdot\|$ or $d_2(x,y) = 1_{x\neq y}$ (more general in bounded cases). Let us first give a new inequality called a conditional weak transport inequality because one needs to work conditionally on θ, see the discussion after the proof:

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

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

where

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

Proof. Considering the change $(\mathcal{Z},\theta) \to (\mathcal{Z}G^{-1/2}, G^{1/2}\theta)$, 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 $\tilde{T}_{2,\|\cdot\|}(\rho C)$ and $\tilde{T}^{(i)}_{2,\|\cdot\|}(\rho C)$ using similar arguments than in Lemma 2.1 in [16]. In all the sequel, we thus consider $G = I_d$, $\mathcal{Z} \in \tilde{T}_{2,\|\cdot\|}(\rho C)$ and $\tilde{T}^{(i)}_{2,\|\cdot\|}(\rho C)$. With this notation, $P[\|\mathcal{Z}\|_n^2] = d$ and $\|\hat{\theta} - \overline{\theta}\|^2 = R(\hat{\theta}) - R(\overline{\theta})$. We adopt the so called PAC-bayesian approach considering that $\hat{\theta} = \rho_{\hat{\theta}}[\theta]$ where $\rho_{\hat{\theta}} = \mathcal{N}_d(\overline{\theta}, \beta^{-1}I_d)$ 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 $\rho_{\overline{\theta}}$ called a priori because it does not depend on the observations (X_i). Let us fix some measure Q and denote Q_θ the conditional probability measure such that $\rho_{\overline{\theta}}Q_\theta = Q\rho_{\hat{\theta}}$ (it exists as the support of ρ_θ does not depend on θ).

Let us first study similar properties than in (5.1) of the function $f = \overline{r}$. 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

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

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

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

As P satisfies $\tilde{T}_{2,\|\cdot\|}(\rho C)$ and $\tilde{T}^{(i)}_{2,\|\cdot\|}(\rho C)$ and using the Cauchy-Schwartz inequality we obtain

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

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

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

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

$Q\rho_{\hat{\theta}}[R(\theta) - R(\overline{\theta}) - r(\theta) + r(\overline{\theta})] \leq$

$$4\sqrt{\rho C n^{-1}(\mathcal{K}(Q|P)+\beta/2Q[R(\hat\theta)-R(\overline\theta)])\rho_{\overline\theta}[(1+\|\theta)\|^2)R(\theta)]+(1+\|\overline\theta\|^2)Q[r(\overline\theta)]}.$$

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

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

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

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

Let us decompose the last term:

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

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

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

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

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)

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

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

Let us discuss the choices $\mu=\rho_{\hat\theta}$ and $\nu=\rho_{\overline\theta}$ made above. 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 $Q\mu[r(\theta)-r(\overline\theta)]$ is small as $r(\hat\theta)-r(\overline\theta)<0$. Choosing μ as the Dirac mass in $\hat\theta$ is excluded by the condition of 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 $Q\mu[r(\theta)-r(\overline\theta)]$ can be seen as an alternative to the classical VC-dimension, see Mc Allester [35]. The measure μ should be chosen in order to bound this term (and the entropy $\mathcal{K}(\mu|\nu)$). It leads to Gibbs estimators that are nice alternatives to classical estimators, see Chapter 4 of the textbook of Catoni [13] in the iid case, [3, 2] in weakly dependent settings. Here we choose the gaussian measures $\mu=\rho_{\hat\theta}$ and $\nu=\rho_{\overline\theta}$ as in Audibert and Catoni [5] for simplicity because $\mathcal{K}(\nu|\mu)=\beta/2\|\hat\theta-\overline\theta\|^2$. This choice leads to estimate the term $Q\mu[r(\theta)-r(\overline\theta)]$ by $Q[\|\mathcal{Z}\|_n^2]/\beta$. This term can easily be estimated by the sum of d/β and a concentration term implying the entropy $\mathcal{K}(Q|P)$. Thus we obtain nonexact oracle inequalities:

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

$$R(\hat{\theta}) \leq (1 + B_1\eta)R(\overline{\theta}) + \frac{B_2 d + 16\rho C \log(\varepsilon^{-1})}{n\eta} + \frac{B_3}{(n\eta)^2}$$

where

$$\begin{aligned} B_1 &= 2(3 + 2\|\overline{\theta}\|^2 + \eta/n), \\ B_2 &= 2(5 + \|\overline{\theta}\|^2), \\ B_3 &= 2(d(d-1) + d/n). \end{aligned}$$

This result extends nonexact oracle inequalities as developed by Lecué and Mendelson [27] to a dependent context but for the OLS only. Similar results for regularized estimators in the iid case are given in [27]. Nonexact oracle inequalities are weak form of exact oracle inequalities valid in cases where the Bernstein assumption of Bartlett and Mendelson [7] does not hold. Here, the nonexact oracle inequality holds without any constraint on $\theta \in \mathbb{R}^d$ whereas the assumption (6.4) is crucial for the exact oracle inequality given in Corollary 6.4.

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

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

Notice that by definition of K we have

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

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

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

As $P[\|Z\|^2] = d$ we obtain similarly

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

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

$$\begin{aligned} Q[\|Z\|^2/\beta + \lambda/n r(\overline{\theta})] \leq\, & d/\beta + \lambda/n R(\overline{\theta}) \\ & + 4\sqrt{\rho C n^{-1}(d/\beta^2 + (\lambda/n)^2 R(\overline{\theta}))\mathcal{K}(Q|P)}. \end{aligned}$$

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

$$Q[\overline{R}(\hat{\theta}) - B_1\eta R(\overline{\theta}) - B_2 d/(n\eta) - B_3/(n\eta)^2] \leq \frac{16\rho C\mathcal{K}(Q|P)}{n\eta}.$$

Choose Q as the probability P restricted to the complementary of the event corresponding to the desired oracle inequality that we denote A. Then

$$\frac{16\rho C \log(\varepsilon^{-1})}{n\eta} \leq Q[\overline{R}(\hat{\theta}) - B_1\eta R(\overline{\theta}) - B_2 d/(n\eta) - B_3/(n\eta)^2]$$

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

6.3. **Exact oracle inequalities for $\tilde{\Gamma}(2)$-weakly-dependent sequences.** Let us now give an equivalent of (6.2) when $d_1(x,y) = d_2(x,y) = 1_{x\neq y}$. 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 $f = \overline{r}$ we obtain easily

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

For the specific choice $\mu = \rho_{\hat{\theta}}$ and $\nu = \rho_{\overline{\theta}}$ we use computations given in Lemma 1.2 in the supplementary material of [5]: for any $\theta \in \mathbb{R}^d$ we have

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

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 probability measures 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

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

Similar assumption has been used in the iid case by Audibert and Catoni in [5]. 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 [5] we obtain easily that

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

It leads to the following equivalent of Theorem 6.1

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

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

In the above estimate the terms involving $r(\overline{\theta})$ are nuisance terms without additional on $\overline{\theta}$. However, if this term is bounded then the last term of the inequality is proportional to the excess risk $Q[\overline{R}(\hat{\theta})]$. Similarly, in the classical approach [7], the excess risk also appears in the concentration under the Bernstein condition that controls this variance term by $\overline{R}(\hat{\theta})$. It is the major advantage considering the Hamming distance compared with the Euclidian distance where instead of $Q[\overline{R}(\hat{\theta})]$ the term $Q[R(\hat{\theta})]$ appeared. As $Q[\overline{R}(\hat{\theta})]$ is the quantity of interest, we obtain here exact oracle inequalities

Corollary 6.4. *If condition* (6.4) *holds, for any* $0 < \varepsilon < 1$ *and any* $M > 0$ *we have with probability* $1-\varepsilon$*:*

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

The condition (6.4) on the parameters θ is crucial for obtaining such exact oracle inequalities; it is equivalent to the Bernstein condition of Bartlett and Mendelson [7] in our statistical setting when the empirical risk is bounded, see [5]. Except the condition (6.4), the exact oracle inequality holds for any $\tilde{\Gamma}(2)$-weakly dependent sequences without assumptions on the margins (because any probability measure satisfies $\tilde{T}_{2,1_{x\neq y}}(1)$). These oracle inequalities are knew, even in the iid case. We

refer the reader to Audibert and Catoni [5] for estimates of the term $\log P(r(\overline{\theta}) > M)$ in the iid case under finite moments assumption on P of order 4 only.

Proof. Let us denote $A = \{r(\overline{\theta}) \leq M\}$ 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 weak transport inequalities. However, a similar reasoning than to obtain (6.3) yields

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

We now use 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 by

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

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

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

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

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

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. □

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 and G. Lecué for helpful discussions on the PAC-bayesian approach and (non)exact oracle inequalities respectively. I am also grateful to N. Gozlan for indicating me the minimax identity of the weak transport studied here and the one of [32]. Finally, I would like to deeply thank F. Merlevède for pointing out errors in a previous version.

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., Li, X. and Wintenberger, O. (2012) Prediction of time series by statistical learning: general losses and fast rates, preprint available on arxiv.org.

[3] Alquier, P. and Wintenberger, O. (2012) Model selection for weakly dependent time series forecasting, *Bernoulli* **18 (3)**, 883–913.

[4] Andrews, D.W.K. (1984) Nonstrong mixing autoregressive processes. *J. Appl. Probab.* **21**, 930–934.

[5] Audibert, J.-Y. and Catoni, O. (2011) Robust linear least squares regression *Ann. Statist.* **39 (5)**, 2766–2794.

[6] Bardet, J.-M. et Wintenberger, O. (2009) Asymptotic normality of the Quasi Maximum Likelihood Estimator for multidimensional causal processes. *Ann. Statist.* **37**, 2730–2759.

[7] Bartlett, P. L. and Mendelson, S. (2006) Empirical minimization. *Probab. Theory Related Fields* **135 (3)**, 311–334.

[8] Bertail, P. and Clémencon, S. (2010) Sharp bounds for the tails of functionals of Markov chains. *Theory Probab. Appl.* **54 (3)**, 505–515,

[9] Bobkov, S. G. and Götze, F. (1999) Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal.* **163 (1)**, 1–28.

[10] Bobkov, S. G., Gentil, I., Ledoux, M. (2001) Hypercontractivity of Hamilton-Jacobi equations. *J. Math. Pures Appl.* **80 (7)**, 669–696.

[11] Boucheron, S., Lugosi, G. and Massart, P. (2009) On concentration of self-bounding functions. *Electron. J. Probab.* **14 (64)**, 1884–1899.

[12] Boucheron, S., Lugosi, G. and Massart, P. (2000). A sharp concentration inequality with applications. *Random Structures and Algorithms* **16 (3)**, 277–292.

[13] Catoni, O. (2004) Statistical Learning Theory and Stochastic Optimization, *Lecture Notes in Mathematics*, Springer, Berlin.

[14] Dedecker, J. and Prieur, C. (2005) New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields*, **132(2)**, 203–236.

[15] 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.

[16] 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.

[17] Doukhan, P. and Wintenberger, O. (2008) Weakly dependent chains with infinite memory. *Stochastic Process. Appl.* **118 (11)**, 1997–2013.

[18] Duflo, M. (1997) Random iterative models. *Applications of Mathematics*, **34**. Springer-Verlag, Berlin.

[19] Gao, F., Guillin, A., and Wu, L. (2012) Bernstein types concentration inequalities for symmetric Markov processes. To appear in *SIAM Theor. Probab. Appl.*

[20] Goldstein, S. (1979) Maximal coupling. *Z. Wahrsch. Verw. Gebiete* **46**, 193–204.

[21] Gozlan, N. and Léonard, C. (2010) Transport inequalities. A survey. *Markov Processes and Related Fields*, **16**, 635–736.

[22] Gozlan, N., Roberto, C., Samson, P.-M. et Tetali, P. (2012) Displacement convexity of entropy and related inequalities on graphs. Preprint.

[23] Guillin, A., Léonard, C., Wu, L. and Yao, N. Transportation-information inequalities for Markov processes. *Probab. Theory Related Fields* **144 (3-4)**, 669–695.

[24] Ibragimov, I. A. (1962) Some limit theorems for stationary processes. *Theory Probab. Appl.* **7**, 349–382.

[25] Joulin, A. and Ollivier, Y. (2010) Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.* **38 (6)**, 2418–2442.

[26] Kontorovich, A. and Ramanan, K. (2008) Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Ann. Probab.* **36(6)**, 2126–2158.

[27] Lecué, G. and Mendelson, S. (2012) General nonexact oracle inequalities for classes with a subexponential envelope. *Ann. Statist.* **40 (2)**, 832–860.

[28] Ledoux, M. (1996) On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.* **1**, 63–87

[29] Lezaud, P. (1998) Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.* **8 (3)**, 849–867.

[30] Marton K. (1986) A simple proof of the blowing-up lemma. *IEEE Trans. Inform. Theory* **32(3)**, 445–446.

[31] Marton K. (1996) Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.* **24 (2)**, 857–866.

[32] Marton, K. (1996) A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.* **6 (3)**, 556–571.

[33] Marton, K. (2004) Measure concentration for Euclidean distance in the case of dependent random variables. *Ann. Probab.* **32 (3B)**, 2526–2544.

[34] Massart, P. (2007), *Concentration inequalities and model selection*. Springer, Berlin.

[35] McAllester, D. A. (1999) PAC-Bayesian model averaging. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 164–170, ACM, New York.

[36] McDiarmid C. (1989) On the method of bounded differences, in: Surveys of Combinatorics, *Siemons J. (Ed.)*, **Lect. Notes Series 141**, London Math. Soc.

[37] 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

[38] Rüschendorf, L. (1985) The Wasserstein Distance and Approximation Theorems, *Z.Wahrsch. Verw. Gebiete* **70**, 117–129.

[39] Samson, P.-M. (2000) Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.* **28 (1)**, 416–461.
[40] Talagrand, M. (1995) Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l?I.H.E.S.*, **81**, 73–205.
[41] 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.
[42] Villani, C. (2009) *Optimal transport, old and new.* Springer-Verlag, Berlin, 2009.
[43] 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 Peri, 92245 Malakoff Cedex, FRANCE

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