Electronic Journal of Probability

Electron. J. Probab. **20** (2015), no. 114, 1–27.
ISSN: 1083-6489 DOI: 10.1214/EJP.v20-3558

Weak transport inequalities and applications to exponential and oracle inequalities

Olivier Wintenberger*

Abstract

We study the dimension-free inequalities, see Talagrand [49], for non-product measures extending Marton's [39] weak transport from the Hamming distance to other metrics. The Euclidian norm is proved to be appropriate for dealing with non-product measures associated with classical time series. Our approach to address dependence, based on coupling of trajectories, weakens previous contractive arguments used in [20] and [41]. Following Bobkov-Götze's [10] approach, we derive sub-Gaussianity and a convex Poincaré inequality for non-product measures that are not uniformly mixing, extending the Samson's [48] results. Such dimension-free inequalities are useful for applications in statistics. Expressing the concentration properties of the ordinary least squares estimator as a weak transport problem, we obtain new oracle inequalities with fast rates of convergence for classical time series models.

Keywords: transport inequalities; concentration of measures; time series; oracle inequalities.
AMS MSC 2010: 60E15; 28A35.
Submitted to EJP on May 28, 2014, final version accepted on July 2, 2015.

1 Introduction

In his remarkable paper [49], Talagrand proved that convex distances have dimension-free concentration properties. Marton [37]'s seminal work showed that transport inequalities efficiently yield such dimension-free concentration inequalities. Using a duality argument, Bobkov and Götze [10] further proved that transport inequalities are equivalent to some concentration inequalities. Our references on the subject are the monograph of Villani [51], the survey of Gozlan and Léonard [25] and the textbook of Boucheron *et al.* [16] for applications in statistic. In dependent settings, Samson [48] showed that Marton's weak transport extends to the uniformly mixing setting [29]. Considering Marton's weak transport for metrics other than the Hamming one, this article develops transport and exponential inequalities, and fast rates of convergence in statistical applications for classical time series models. First, we motivate the choice of Marton's weak transport approach.

*Sorbonne Universités, UPMC Univ Paris 06 LSTA, Case 158, 4 place Jussieu, 75005 Paris, FRANCE &Department of Mathematical Sciences University of Copenhagen, DENMARK.
E-mail: olivier.wintenberger@upmc.fr http://wintenberger.fr

In the case of product measures with common margin (iid case), the modified log-Sobolev approach developed in [42] provides useful and optimal dimension-free concentration inequalities of Bernstein-type. However, such inequalities are not valid in their optimal form in many situations for non product measures. The reason is the following: in the bounded iid case, Bernstein's inequality yields Gaussian behavior for deviations, depending on the essential supremum. In various bounded Markovian cases, there exists a unique regeneration scheme of iid cycles with random length. Then the variance terms in the Bernstein-type inequalities are perturbed by the concentration properties of the random length that cannot be bounded; see [8]. The perturbations yield an additional logaritheoremic term which cannot be removed; see [1]. This fact is a drawback for statistical applications, where one needs to recover a variance term similar to the iid case. To bypass this problem, various authors assume contractive properties on the kernel of the Markov chains; see Marton [38] under geometric ergodicity and Lezaud [35] under a spectral gap condition. For symmetric Markov processes, the spectral gap condition is more general than uniform ergodicity and it is also necessary for Bernstein's inequality; see [28]. More general contractive conditions are used by Marton [41] and Djellout *et al.* [20]. They extend the dimension-free transport inequality $T_{2,d_2}(C)$ described below. Consider a lower semi-continuous metric d on a Polish space E. Write $\mathcal{K}(Q|P) = Q[\log(dQ/dP)]$, where $P[h]$ denotes $\int hdP$ for any probability measure P and measurable function h. A measure P on E^n satisfies $T_{2,d}(C)$, $C > 0$, if, for any probability measure Q on E^n,

$$\inf_{\pi} \pi\Big[\sum_{k=1}^{n} d(X_k, Y_k)^2\Big] \leq 2C\mathcal{K}(Q|P).$$

Here π is any coupling scheme of $(X_k, Y_k)_{1\leq k\leq n}$ with margins (P, Q). This inequality is dimension-free if the "variance term" C does not depend on n. Moreover, if this "variance term" is sufficiently close to the marginal variance then a Bernstein-type inequality is recovered thanks to Bobkov-Götze's [10] duality argument. Then fast convergence rates in statistical applications can be achieved, see for instance [30].

Many classical time series models do not satisfy the contractive conditions of [41, 20]. Therefore, we inted to obtain fast rates of convergence in statistical applications based on weaker assumptions. Samson [48] already extended Marton's weak transport approach to non-contractive time series. He worked under a weak dependence condition closely related to the uniform mixing condition [29]. Samson's results yield fast convergence rates of order n^{-1} in statistical applications for uniformly mixing sequences; see [2]. His approach relies on the maximal coupling properties and cannot be extended in a direct way to more general dependent settings because the maximal coupling exists only for Hamming's distance; see [19]. Recalling Marton's [39] orignal approach and denoted Hamming's distance $\mathbf{1}$, the main result in [48] is expressed as a weak transport inequality $\tilde{T}_{2,\mathbf{1}}(C)$: for any probability measure Q,

$$\inf_{\pi} \sup_{\alpha} \frac{\sum_{k=1}^{n} \pi[\alpha_k(Y)\,\mathbf{1}_{X_k\neq Y_k}]}{(\sum_{k=1}^{n} Q[\alpha_k(Y)^2])^{1/2}} \leq \sqrt{2C\mathcal{K}(Q|P)}.$$

Here α is any non-negative measurable function and C does not depend on n. We extend these weak transport inequalities to more general dependent settings by considering metrics different from Hamming's. We say that a probability measure P on E^n satisfies the weak transport inequality $\tilde{T}_{p,d}(C)$ for $C > 0$ and $1 \leq p \leq 2$, if, for any probability measure Q on E^n,

$$\inf_{\pi} \sup_{\alpha} \frac{\sum_{k=1}^{n} \pi[\alpha_k(Y)d(X_k, Y_k)]}{(\sum_{k=1}^{n} Q[\alpha_k(Y)^q])^{1/q}} \leq \sqrt{2C\mathcal{K}(Q|P)}, \tag{1.1}$$

with $1/p+1/q=1$ and the convention $+\infty/+\infty=0/0=0$. By duality of the spaces ^{p} and ℓ^p with ^{q} and ℓ^q, respectively, (1.1) is equivalent to

$$\inf_\pi Q\Big[\sum_{k=1}^n \pi[d(X_k,Y_k)\mid Y]^p\Big]^{1/p}\le\sqrt{2C\mathcal{K}(Q|P)}.$$

By Jensen's inequality, $\tilde{T}_{p,d}(C)$ appears as a weak version of the classical transport inequality $T_{p,d}(C)$ of [20]:

$$\inf_\pi\Big(\sum_{k=1}^n \pi[d^p(X_k,Y_k)]\Big)^{1/p}\le\sqrt{2C\mathcal{K}(Q|P)}.$$

Contrary to the classical $T_{p,d}(C)$ transport inequalities, the weak transport inequalities extend nicely to non-product non-contractive measures P on E^n, $n\ge1$. Using a new coupling scheme for trajectories, our main result Theorem 3.15 states that there exists $C'>0$ such that

$$\sup_\alpha\inf_\pi\frac{\sum_{j=1}^n\pi[\alpha_j(Y)d(X_j,Y_j)]}{(\sum_{j=1}^n Q[\alpha_j(Y)^q])^{1/q}}\le\sqrt{2n^{2/p-1}C'\mathcal{K}(Q|P)}. \tag{1.2}$$

Sion's minimax theorem actually shows that (1.2) is equivalent to $\tilde{T}_{p,d}(n^{2/p-1}C')$. The corresponding concentration properties are dimension-free only in the case $p=2$. We introduce the notion of $\Gamma_{d,d'}(p)$-weak dependence in Definition 3.14 to assert the existence of a coupling between the trajectories $(X_{i+1},\dots,X_n)$, given the same past X_k, $k<i$, controlling possible deviations in the present X_i through an auxiliary metric d' satisfying $d\le Md'$, $M>0$.

When the metrics $d=d'=\mathbf{1}$ are Hamming's, $\Gamma(2)$-weak dependence is related with the weak dependence condition used by Samson [48], and we slightly improve his results. However, to deal with classical time series models, it is preferable to choose d as the Euclidian norm; see Section 4. For $p=1$, $\tilde{T}_{1,d}(C)=T_{1,d}(C)$ by definition and $\Gamma(1)$-weak dependence coincides with the setting of [46] when $d'=\mathbf{1}$ and the one of [20] when $d'=d$. Thus we recover Hoeffding's inequalities of [46, 20]. They are not dimension-free because $n^{2/p-1}=n$ for $p=1$. In the case $p=2$, we prove the first dimension-free concentration result for ARMA processes under the minimal dependence assumption that the stationary distribution exists. This result provides a positive answer to an important question raised in [20], Remark 3.6. Our approach considerably improves upon the existing methods based on contractive arguments [20, 41]. For instance, consider the Markov chain (X_t,ξ_t) formed by an ARMA(1,1) process $X_t=\phi X_{t-1}+\xi_t+\theta\xi_{t-1}$. Then the contraction condition is $\phi^2+\theta^2<1$ whereas the trajectory coupling scheme exists when $|\phi|<1$.

Weakening transport inequalities does not deteriorate the concentrations property useful for statistical applications. We prove that $\tilde{T}_{2,d}(C)$ yields the convex distance dimension-free estimate due to Talagrand:

$$P(A)P(d_c(X,A)>t)\le\exp\Big(-\frac{t^2}{4C}\Big),\qquad t>0,\ \text{for any measurable set } A.$$

Here $d_c(x,A)$ is the convex distance of Talagrand [49], when d is the Hamming distance, and the Euclidian distance to the convex hull of A as in Maurey [36], when d is the Euclidian norm. Following Bobkov and Ledoux [9], we obtain this result by analyzing the Bobkov-Götze [10] dual form of the weak transport inequality: if P satisfies $\tilde{T}_{2,d}(C)$ on E

then for any function f such that $f(x) - f(y) \le \alpha(x)d(x,y)$, $x, y \in E$, for some function α we have

$$P\Big[\exp\Big(\lambda(f - P[f]) - \frac{C\lambda^2}{2}\alpha^2\Big)\Big] \le 1, \qquad \lambda > 0. \tag{1.3}$$

When d is the Hamming distance, inequality (1.3) yields the classical Bernstein inequality with suboptimal constants, see Ledoux [33] in the independent setting and Samson [48] in the uniform mixing setting. When the function f is convex, the condition above is automatically satisfied for $\alpha = \partial f$ (the sub-gradients) and d the Euclidian norm. The inequality (1.3) coincides with generalizations of Tsirel'son's inequality discovered in [50] for Gaussian measures; see [12]. Using the dual form (1.3), we also prove that $\tilde{T}_{2,d}(C)$ implies sub-Gaussianity and the convex Poincaré inequality [11]. Then, the weak transport approach provides dimension-free concentration properties of ARMA processes under minimal assumptions. It is sufficient for extending fast rates of convergence in statistical applications from the classical iid setting to the $\Gamma(2)$-weak dependence one.

As the transport inequalities yield concentration of measures via relative entropy, we use the statistical PAC-Bayesian paradigm that describes the accuracy of estimators in terms of relative entropy; see [43]. We introduce the conditional weak transport approach that provides sharp oracle 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. If R denotes the risk of prediction, an oracle inequality holds if $R(\hat{\theta}) \le (1+\eta)R(\overline{\theta}) + \Delta_n \eta^{-1_{\eta\neq 0}}$ where $\eta \ge 0$, $\overline{\theta}$ is the oracle defined as $R(\overline{\theta}) \le R(\theta)$ for all θ and Δ_n is the rate of convergence. If $\eta = 0$ then the oracle inequality is said to be exact and otherwise it is non-exact; see [32]. The dimension-free concentration properties yield fast rates of convergence $\Delta_n \propto n^{-1}$. Moreover, in the $\Gamma(2)$-weak dependent case for $d = \mathbf{1}$, the "variance term" $C > 0$ in $\tilde{T}_{2,\mathbf{1}}(C)$ coincides with the marginal variance. The marginal variance is crucial to obtain exact oracle inequalities with fast rates of convergence under the Bernstein condition of [7]. Thus, we obtain new exact oracle inequalities with fast convergence rates for the OLS $\hat{\theta}$ in that specific case. However, in the more general $\Gamma(2)$-weak dependent cases when $d = N$, the Bernstein condition cannot hold as the "variance term" $C > 0$ does not coincide with the marginal variance. However, Tsirel'son's inequality still holds and we achieve new non-exact oracle inequalities in this case. The non-asymptotic efficiency of the OLS is proved for the first time for many models, including classical ARMA models.

The paper is organized as follows. In Section 2 we develop the preliminaries to be used in the proof of our main result, a weak transport inequality for non-product measures stated in Section 3. In Section 3 We also study the dual form of the weak transport inequalities, the Tsirel'son inequality and the connection with Talagrand's inequalities. Section 4 is devoted to some examples of $\Gamma(p)$-weak dependent processes. Finally, new oracle inequalities with fast rates of convergence are given in Section 5.

2 Weak transport costs, glueing Lemma and Markov couplings

2.1 Weak transport costs on E

Let $M(E)$ denote the set of probability measures on the Polish space E, $M^+(E)$ 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 convention that $Q[\alpha^q]^{1/q} = \operatorname{ess\,sup} \alpha(Y)$ when $q = \infty$ and $+\infty/+\infty = 0/0 = 0$. For fixed $\alpha \in M^+(F)$, let us denote

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

Notice 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 $P = Q$. We assumed that $\alpha \in M^+(E)$ and d are lower semi-continuous such that the optimal transport in the definition of the weak transport cost exists; see [25]. 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 glueing lemma:

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

Proof. In view of the classical glueing Lemma (see for example [51]) we can choose $\pi_{x,y,z} = \pi_{x|y}\pi_{z|y}\pi_y$. The margins correspond to $\pi_{x|y}\pi_y = \pi_{x,y}$ and $\pi_{z|y}\pi_y = \pi_{y,z}$ and the conditional independence as $\pi_{x,z|y} = \pi_{x,y,z}/\pi_y$ follows by definition. □

The conditional independence in the glueing Lemma 2.1 is the main ingredient for proving 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

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

We 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

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

Now choose $\pi^*_{x,y}$ satisfying

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

Using Jensen's inequality we have $Q[\tilde{\alpha}^q] = Q[\pi^*_{z|y}[\alpha(Z)]^q] \le R[\alpha^q]$. Let us denote $\pi^* = \pi^*_{x,y,z}$ obtained by the glueing Lemma 2.4 applied to $\pi^*_{x,y}$ 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

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

Taking the supremum over α in the last inequality, the desired result follows from the definition of $\tilde{W}_{p,d}(Q,R)$. □

2.2 Markov coupling schemes

In this section, we consider with no loss of generality Markov coupling schemes only on the product space E^n with $n = 2$. The case $n \geq 2$ follows by an induction argument.

Definition 2.3. *For P, $Q \in M(E^2)$, the set of Markov couplings $\tilde{M}(P,Q)$ is defined as the product $\pi = \pi_1\pi_{2|1}$ with π_1 a coupling of (P_1,Q_1) and $\pi_{2|1}$ a coupling of ($P_{2|1}$,$Q_{2|1}$).*

The terminology "Markov coupling" was introduced by Rüschendorf in [47], even if the coupling is not the distribution of a Markov process. Similar couplings are used by Marton in [39]. The property of conditional independence in the glueing Lemma 2.1 is well compatible with Markov couplings:

Lemma 2.4. *For any Markov couplings $\pi_{x,y} \in \tilde{M}(P,Q)$ and $\pi_{y,z} \in \tilde{M}(P,Q)$ with $P,\ Q,\ R \in \tilde{M}(E^2)$ there exists a distribution $\pi_{x,y,z}$ with margins $\pi_{x,y}$, $\pi_{y,z}$ 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

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

We check that $\pi_{x,y,z}$ has the correct margins. First, in view of Lemma 2.1 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. By the definition of Markov coupling schemes, 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

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

and the same reasoning applies to the second margin.

From the glueing Lemma 2.1 we already know that X_1 and Z_1 are independent conditional on Y_1, i.e. that $\pi_{x_1,z_1|y_1} = \pi_{x_1|y_1}\pi_{z_1|y_1}$. We show independence conditional on Y_1 and Y_2 as well. 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 $\pi_{y_2|y_1} = \pi_{y_2|x_1,y_1,z_1}$ due to the relation (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 $\pi_{x_1,z_1|y_1,y_2} = \pi_{x_1|y_1,y_2}\pi_{z_1|y_1,y_2}$ since

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

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 when 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

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

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

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 $n \geq 2$. For P, $Q \in M(E^n)$ we define

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

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

$$\tilde{W}_{\alpha,d}(P,Q) = \inf_{\pi \in \tilde{M}(P,Q)} \sum_{k=1}^n \pi[\alpha_k(Y) d(X_k, Y_k)] \tag{2.7}$$

for any fixed $\alpha = (\alpha_k)_{1\leq k\leq n} \in M^+(E^n)^n$. Considering Markov couplings, we use the conditional independence in the glueing Lemma 2.4 to assert that the weak transport cost on E^n also satisfies a useful inequality stronger than the triangular one:

Lemma 2.5. *For any $P, Q, R \in M(E^n)$, for any $\alpha \in M^+(E^n)^n$ there exists $\tilde{\alpha} \in M^+(E^n)^n$ satisfying $Q[\tilde{\alpha}_k^q(Y)] \leq R[\alpha_k^q(Z)]$ for $1 \leq k \leq n$ 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 2.6. As a consequence of Lemma 2.5, 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 using the relation $Q[\tilde{\alpha}_k^q(Y)] \leq R[\alpha_k^q(Z)]$ and taking the supremum over α in (2.8).

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

$$\begin{aligned} \pi^*_{y,z}\Big[\sum_{k=1}^n \alpha_j(Z) d(X_k, Z_k)\Big] &= \tilde{W}_{\alpha,d}(Q,R), \\ \pi^*_{x,y}\Big[\sum_{k=1}^n \pi^*_{z|y}[\alpha_k(Z)] d(X_k, Y_k)\Big] &= \tilde{W}_{\pi^*_{z|y}[\alpha(Z)],d}(P,Q). \end{aligned}$$

Write $\pi^* = \pi^*_{x,y,z}$ obtained by glueing $\pi^*_{x,y}$ and $\pi^*_{y,z}$; see Lemma 2.4. Then

$$\begin{aligned} \pi^*_{x,z}\Big[\sum_{k=1}^n \alpha_k(Z) d(X_k, Z_k)\Big] &\leq \pi^*_{x,y,z}\Big[\sum_{k=1}^n \alpha_k(Z) d(X_k, Y_k)\Big] + \pi^*\Big[\sum_{k=1}^n \alpha_k(Z) d(Y_k, Z_k)\Big] \\ &\leq \pi^*_{x,y}\Big[\sum_{k=1}^n \pi^*_{z|y}[\alpha_k(Z)] d(X_k, Y_k)\Big] + \pi^*_{y,z}\Big[\sum_{k=1}^n \alpha_k(Z) d(Y_k, Z_k)\Big] \\ &\leq \tilde{W}_{\pi^*_{z|y}[\alpha(Z)],d}(P,Q) + \tilde{W}_{\alpha,d}(Q,R). \end{aligned} \tag{2.10}$$

Inequality (2.8) follows from (2.10) if we write $\tilde{\alpha}_k(y) = \pi^*_{z|y}[\alpha_k(Z)]$ and notice that the relation $Q[\tilde{\alpha}_k^q(Y)] \leq R[\alpha_k^q(Z)]$ holds by an application of Jensen's inequality. □

3 Weak transport inequalities

3.1 Weak transport inequalities and Bobkov-Götze dual forms

We say that the probability measure P on E^n, $n \geq 1$, satisfies the weak transport inequality $\tilde{T}_{p,d}(C)$, $C > 0$, if for any distribution Q on E^n we have

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

We say that P satisfies the inverted weak transport inequality $\tilde{T}^{(i)}_{p,d}(C)$, $C>0$ if

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

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 $\tilde{T}^{(i)}_{p',d}(C)$ with $p' \ge p$. We have $\tilde{T}_{1,d}(C) = \tilde{T}^{(i)}_{1,d}(C) = T_{1,d}(C)$ for the classical transport inequality $T_{1,d}(C)$ defined by the relation

$$\inf_{\pi \in \tilde{M}(P,Q)} \sum_{k=1}^n \pi[d(X_k,Y_k)] \le \sqrt{2C\mathcal{K}(Q|P)}.$$

Following [10], we investigate the dual form of weak transport. Denote

$$f_{\alpha,d}(y) = \inf_{x\in E^n} \Big\{ \sum_{k=1}^n \alpha_k(y) d(x_k,y_k) + f(x) \Big\}$$

and $\mathcal{C}_b$ the set of continuous bounded functions taking value in $\mathbb{R}$. We have the following Bobkov-Götze dual forms of the weak transport inequalities:

Theorem 3.1. *The weak transport inequalities $\tilde{T}_{p,d}(C)$ and $\tilde{T}^{(i)}_{p,d}(C)$ are equivalent to, respectively,*

$$\sup_{\lambda>0} \sup_{\alpha\in M^+(E^n)^n} \sup_{f\in\mathcal{C}_b} P\Big[\exp\Big(\lambda(f_{\alpha,d} - P[f]) - C\lambda^2 \Big(\frac{\sum_{k=1}^n \alpha_k^q - 1}{q} + \frac{1}{2} \Big)\Big)\Big] \le 1, \tag{3.3}$$

$$\sup_{\lambda>0} \sup_{\alpha\in M^+(E^n)^n} \sup_{f\in\mathcal{C}_b} P\Big[\exp\Big(\lambda(f_{\alpha,d} - P[f]) - C\lambda^2 \Big(\frac{\sum_{k=1}^n P[\alpha_k^q] - 1}{q} + \frac{1}{2} \Big)\Big)\Big] \le 1, \tag{3.4}$$

with the convention that $\alpha_k^q/q \to 0$ as $q\to\infty$.

Remark 3.2. In the case $p=1$ and $q=\infty$ we recognize the dual form of the transport inequality $T_{1,d}(C)$, i.e. Hoeffding's inequality [10]:

$$\sup_{\lambda>0} \sup_{f\in \mathrm{Lip}_1(d)} P\Big[\exp\Big(\lambda(f - P[f]) - \frac{C\lambda^2}{2} \Big)\Big] \le 1.$$

Here Lip $_1$ is the set of 1-Lipschitz functions f with respect to d on E^n satisfying

$$|f(x) - f(y)| \le \sum_{k=1}^n d(x_k,y_k).$$

Proof. We focus on the case $1<p\le 2$ as Remark 3.2 already deals with the case $p=1$. As proofs are similar, we prove the first dual form only. We first show that $\tilde{T}_{p,d}(C)$ implies the dual form (3.3). For any $\alpha\in M^+(E^n)^n$, Kantorovich duality provides the identity

$$\tilde{W}_{\alpha,d}(P,Q) = \inf_\pi \pi\Big[\sum_{k=1}^n \alpha_k(Y) d(X_k,Y_k) \Big] = \sup_{f\in\mathcal{C}_b} Q[f_{\alpha,d}] - P[f].$$

Then a measure P satisfies $\tilde{T}_{p,d}(C)$ if for any probability measure $Q\in M(E^n)$

$$\sup_{f\in\mathcal{C}_b} Q[f_{\alpha,d}] - P[f] \le \Big(\sum_{k=1}^n Q[\alpha_k^q] \Big)^{1/q} \sqrt{2C\mathcal{K}(Q|P)}.$$

From the variational identity $ab = \inf_{\lambda>0} \lambda a^q/q + b^p/(\lambda^{p-1}p)$, we derive that for any $\lambda>0$

$$Q[(f_{\alpha,d} - P[f])] \le \frac{\lambda C}{q} \sum_{k=1}^n Q[\alpha_k^q] + \frac{\mathcal{K}(Q|P)^{p/2} 2^{p/2} C^{1-p/2}}{\lambda^{p-1} p}.$$

Hence

$$\frac{p}{2}Q\Big[\Big(\frac{p}{C}\Big)^{1-p/2}\lambda^{p-1}\Big(f_{\alpha,d}-P[f]-\frac{\lambda C}{q}\sum_{k=1}^n\alpha_k^q\Big)\Big]^{2/p}\le\mathcal{K}(Q|P).$$

From Young's inequality

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

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

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

For $x=Q[\lambda(f_{\alpha,d}-P[f]-\lambda C/q\sum_{k=1}^n\alpha_k^q)]$ we obtain

$$Q\Big[\lambda\Big(f_{\alpha,d}-P[f]-\frac{\lambda C}{q}\sum_{k=1}^n\alpha_k^q\Big)\Big]-\Big(\frac{1}{p}-\frac{1}{2}\Big)C\lambda^2\le\mathcal{K}(Q|P).$$

Denote

$$\Psi:=\lambda(f_{\alpha,d}-P[f])-C\lambda^2\Big(\frac{\sum_{k=1}^n\alpha_k^q-1}{q}+\frac{1}{2}\Big),$$

and choose Q as $dQ/dP=e^\Psi/P[e^\Psi]$. Then $Q[\Psi]\le\mathcal{K}(Q|P)$ implies that $P[e^\Psi]\le 1$. This is the desired result.

Conversely, assume that the dual form $P[e^\Psi]\le 1$ holds. Then $Q[\Psi]\le\mathcal{K}(Q|P)$ for any measure Q from the variational form of the entropy and we obtain

$$Q[f_{\alpha,d}]-P[f]\le\mathcal{K}(Q|P)/\lambda+\Big(\frac{\sum_{k=1}^n Q[\alpha_k^q]-1}{q}+\frac{1}{2}\Big)C\lambda=\mathcal{K}(Q|P)/\lambda+C\lambda/2,\qquad\lambda>0,$$

where the identity only holds for normalized $\alpha\in M^+(E^n)^n$ such that $\sum_{k=1}^n Q[\alpha_k^q]=1$. Then, minimizing with respect to $\lambda>0$, we obtain the desired result for such normalized α. The weak transport inequality $\tilde{T}_{p,d}(C)$ follows as one can always renormalized $\alpha\in M^+(E^n)^n$ by considering $\alpha/(\sum_{k=1}^n Q[\alpha_k^q])^{1/q}$. □

Now we turn to the case $p=2$. Because $\lambda f=\lambda f_{\lambda\alpha,d}$ we get the following expressions of the dual forms (3.3) and (3.4):

$$\sup_{\alpha\in M^+(E^n)^n}\sup_{f\in\mathcal{C}_b}P\Big[\exp\Big(f_{\alpha,d}-P[f]-\frac{C}{2}\sum_{k=1}^n\alpha_k^2\Big)\Big]\le 1,$$

$$\sup_{\alpha\in M^+(E^n)^n}\sup_{f\in\mathcal{C}_b}P\Big[\exp\Big(f_{\alpha,d}-P[f]-\frac{C}{2}\sum_{k=1}^n P[\alpha_k^2]\Big)\Big]\le 1.$$

In the case $p=2$, we are not able to identify the map $f\to f_{\alpha,d}$ but we still have

Corollary 3.3. *If P satisfies $\tilde{T}_{2,d}(C)$ or $\tilde{T}_{2,d}^{(i)}(C)$ then for any f and α_k satisfying*

$$f(y)-f(x)\le\sum_{k=1}^n\alpha_k(y)d(x_k,y_k),\qquad x,y\in E^n,1\le k\le n,\tag{3.5}$$

we have, respectively,

$$P\Big[\exp\Big(f-P[f]-\frac{C}{2}\sum_{k=1}^n\alpha_k^2\Big)\Big]\le 1,\qquad P\Big[\exp\Big(P[f]-f-\frac{C}{2}\sum_{k=1}^n P[\alpha_k^2]\Big)\Big]\le 1.$$

These two inequalities are particularly useful when the second order terms depending on $\sum_{k=1}^n\alpha_k^2$ can be suitably bounded; below see two classical settings.

3.2 The specific case $d = \mathbf{1}$ of the Hamming distance

The weak transport inequalities when $d(x,y) = \mathbf{1}_{x\neq y}$ for short $d = \mathbf{1}$, was introduced by Marton in [39] who obtained the universality of the weak transports when $n = 1$:

Proposition 3.4. *Any measure P on E satisfies $\tilde{T}_{2,\mathbf{1}}(C)$ and $\tilde{T}^{(i)}_{2,\mathbf{1}}(C)$ with $C = 1$.*

Extended to E^n, $n > 1$, the weak transport inequalities $\tilde{T}_{2,\mathbf{1}}(C)$ and $\tilde{T}^{(i)}_{2,\mathbf{1}}(C)$ imply the concentration of measures; see [39]. Following [16], we introduce the notion of self-boundedness: a function f is self-bounding if there exist functions f_k on E^{n-1}, $1 \le k \le n$, satisfying

$$0 \le f(x) - f_k(x_1,\dots,x_{k-1},x_{k+1},\dots,x_n) \le 1,$$
$$\sum_{i=1}^n (f(x) - f_k(x_1,\dots,x_{k-1},x_{k+1},\dots,x_n))^2 \le f(x).$$

Then f satisfies (3.5) with functions $\alpha_k = f - f_k$ such that $\sum_{k=1}^n \alpha_k^2 \le f$. More generally, we will refer to the self-bounding property for the Hamming distance when f satisfies (3.5) with $\sum_{k=1}^n \alpha_k^2 \le f$. For self-bounding functions, an application of Corollary 3.3 provides that if P satisfies $\tilde{T}_{2,\mathbf{1}}(C)$ or $\tilde{T}^{(i)}_{2,\mathbf{1}}(C)$ then

$$P[\exp((\lambda - C\lambda^2/2)f)] \le \exp(\lambda P[f]), \qquad P[\exp((\lambda - C\lambda^2/2)P[f] - \lambda f)] \le 1, \qquad \lambda > 0.$$

An important example of self-bounding functions is the convex distance d_T introduced by Talagrand [49]. Let $A \subset E^n$ be a measurable set. Denoting

$$d_T(x,A) = \sup_{\|c\|\le 1} \inf_{y\in A} \sum_{k=1}^n c_k(x) 1_{x_k\neq y_k} = \inf_{y\in A} \sum_{k=1}^n c_k^*(x) 1_{x_k\neq y_k},$$

where c^* are the weights that achieve the supremum in d_T, we have

$$\begin{aligned} d_T(x,A) - d_T(y,A) &\le \inf_{x'\in A} \sum_{k=1}^n c_k^*(x) 1_{x_k\neq x'_k} - \inf_{y'\in A} \sum_{k=1}^n c_k^*(y) 1_{y_k\neq y'_k} \\ &\le \sum_{k=1}^n c_k^*(x) 1_{x_k\neq y_k}. \end{aligned}$$

Then, by the convex inequality $x^2 - y^2 \le 2x(x-y)$, we obtain

$$d_T(x,A)^2 - d_T(y,A)^2 \le \sum_{k=1}^n 2d_T(x,A) c_k^*(x) 1_{x_k\neq y_k}.$$

Thus $f(x) = d_T(x,A)^2$ obeys (3.5) with $\alpha_k(x)$ satisfying $\sum_{k=1}^n \alpha_k^2(x) \le 4d_T(x,A)^2$ and $d_T^2(x,A)/4$ is self-bounding. An application of Corollary 3.3 provides Talagrand's inequality [49]; see Section 7.5 of [16]:

Proposition 3.5. *If the law P of $X = (X_1,\dots,X_n)$ satisfies $\tilde{T}_{2,\mathbf{1}}(C)$ and $\tilde{T}^{(i)}_{2,\mathbf{1}}(C)$ then for all measurable set A we have*

$$P(A)P(d_T(X,A) > t) \le \exp\Big(-\frac{t^2}{4C}\Big), \qquad t > 0.$$

Remark 3.6. If the X_ks are independent, Theorem 3.15 yields that P satisfies $\tilde{T}_{2,\mathbf{1}}(1)$ or $\tilde{T}^{(i)}_{2,\mathbf{1}}(1)$ and the constant $C = 1$ is optimal; see [49].

3.3 The specific case $d = N$ of the Euclidian metric

Next we consider the case of $E = \mathbb{R}^n$ equipped with the Euclidian norm $\|\cdot\|$. We write $d = N$. Applying Corollary 3.3, we obtain an exponential inequality for separately convex functions similar to [33]. Recall that a function is separately convex if it is convex in each coordinate.

Corollary 3.7. *If P satisfies $\tilde{T}_{2,d}(C)$ or $\tilde{T}^{(i)}_{2,d}(C)$ then, respectively,*

$$P[\exp(f - P[f] - C\|\nabla f\|^2/2] \le 1, \text{ for any separately convex function } f, \tag{3.6}$$

$$P[\exp(f - P[f] - CP[\|\nabla f\|^2]/2] \le 1, \text{ for any separately concave function } f. \tag{3.7}$$

Proof. Any separately convex function f satisfies (3.5) for $\alpha_k = \partial f_k$, the partial sub derivative of f. Then we identify $\sum_{k=1}^n \alpha_k^2$ with $\|\nabla f\|^2$, where $\nabla f \in \mathbb{R}^n$ is the vector of the partial sub derivatives. □

Remark 3.8. Inequality (3.6) is called the Tsirel'son inequality who discovered it for independent Gaussian random variables with the optimal constant $C = 1$. Corollary 6.1 in [12] states that it holds for any measures satisfying the log-Sobolev inequality. In particular, $\tilde{T}_{2,1_2}(C)$ holds for log-concave measures $dP/dx = e^{-V}$ with C-strongly convex function V.

Thanks to Corollary (3.7), one can relate the weak transport inequalities to more classical notions of concentration. Recall that a measure P on $E = \mathbb{R}^n$ is sub-Gaussian if there exists $c > 0$ such that

$$P[\exp(\lambda\|X\|^2)] < \infty \quad \text{for} \quad 0 < \lambda < c.$$

This property is equivalent to $T_{1,N}(C)$ for some $C > 0$, see [20, 13], and it is a common assumption in statistics. We say that P satisfies the *convex* Poincaré inequality if for any separately convex function g

$$P[(g - P[g])^2] \le CP[\|\nabla g\|^2].$$

Remark 3.9. The convex Poincaré inequality on $E = \mathbb{R}$ has been studied in [11]. It is satisfied for X standard normal and $X \in [0, 1]$ with $C = 1$. It also holds with the same constant $C = 1$ for the corresponding product measure on $\mathbb{R}^n$, $n > 1$.

Notice that the convex Poincaré inequality is equivalent to a *concave* Poincaré inequality.

Theorem 3.10. *The weak transport inequality $T_{2,N}$ or $T^{(i)}_{2,N}$ implies sub-Gaussianity and the convex Poincaré inequality.*

Remark 3.11. In a personal communication, N. Gozlan and P.-M. Samson showed me that the converse is not true using the counter-example given on p.15 in [27].

Proof. The arguments developed in this proof are classical; see [34]. We focus on the case of $T_{2,N}$ when $n = 1$, because the proof for $n > 1$ and $T^{(i)}_{2,N}$ follows the same reasoning. Assume that P satisfies $T_{2,d}(C)$ or $T^{(i)}_{2,d}(C)$ and apply (3.6) to $g(x) = \lambda x$: $P[\exp(\lambda(X - P[X]))] \le \exp(C\lambda^2/2)$, $\lambda > 0$. Then P must be sub-Gaussian. Now, applying (3.6) or (3.7) to tg as $t \to 0$ we obtain the convex Poincaré inequality in both cases. □

Tsirel'son's inequality (3.6) quantifies the concentration of "self-bounding" functions with respect to the Euclidian norm, i.e. convex functions f such that $\|\nabla f\|^2 \le f$. Let A be a measurable set of $\mathbb{R}^n$ and B its convex hull, then $d_N^2(x, B)/4$ with $d_N(x, B) = \inf_{y \in B}\|x - y\|$ is a self-bounding function. Following the same reasoning as in Section 7.5 of [16], we obtain the Euclidian version of Talagrand's concentration inequality of [36].

Proposition 3.12. *If the law P of $X=(X_1,\dots,X_n)$ satisfies $\tilde T_{2,N}(C)$ and $\tilde T^{(i)}_{2,N}(C)$ then*

$$P(A)P(d_N(X,A)>t)\le \exp\Big(-\frac{t^2}{4C}\Big),\qquad t>0.$$

Remark 3.13. Via the convex property (τ)he result is proved in [36] for independent X_js on $[0,1]$ and standard normals with the optimal constant $C=1$.

3.4 Coupling trajectories

In order to obtain the concentration inequalities of Sections 3.1-3.3 for non-product measures on E^n, $n\ge 2$, we prove that the weak transport inequalities hold under the new notion of $\Gamma_{d,d'}(p)$-weak dependence. This notion asserts the existence of a coupling scheme between trajectories $(X_{i+1},\dots,X_n)$ given the same past and controlling possible deviations in the present value X_i. To be precise, we add to any law P on E^n an artificial initial state $X_0=Y_0=x_0=y_0$ for a fixed point $y_0\in E$. Denote $x^{(i)}=(x_i,\dots,x_0)$ for $i\ge 0$ and write $P_{|x^{(i)}}$ for the conditional laws of $(X_{i+1},\dots,X_n)$ given $(X_i,\dots,X_0)=x^{(i)}=(x_i,\dots,x_0)$. Let d and d' be two lower semi-continuous metrics on E such that $d\le Md'$ for some $M>0$. We will work under the following weak dependence assumption:

Definition 3.14. *For any $1\le p\le 2$, the probability measure P is $\Gamma_{d,d'}(p)$-weakly dependent if for any $1\le i\le n$, any $(x^{(i)},y_i)\in E^{i+2}$ there exist coefficients $\gamma_{k,i}(p)\ge 0$ and a coupling π_i of $(P_{|x^{(i)}},P_{y_i,x^{(i-1)}})$ satisfying*

$$P_{x^{(i)}}[\pi_{|i}[d(X_k,Y_k)\mid X]^p]^{1/p}\le \gamma_{k,i}(p)d'(x_i,y_i),\qquad i<k\le n. \tag{3.8}$$

Note that the role of x_i and y_i can be interchanged in (3.8). By the symmetry of x_i and y_i, (3.8) also holds when replacing $P_{x^{(i)}}[\pi_{|i}[d(X_k,Y_k)\mid X]^p]^{1/p}$ by $P_{y_i,x^{(i-1)}}[\pi_{|i}[d(X_k,Y_k)\mid Y]^p]^{1/p}$. We introduce the $n\times n$ matrix

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

We equip $\mathbb{R}^n$ with the ℓ^r-norm, $1\le r<\infty$, and the set of the $n\times n$ matrices A with the subordinated norms

$$\|A\|_{2,r}=\max_{x\neq 0}\frac{\|Ax\|_r}{\|x\|_2},\qquad 1\le r<\infty.$$

We are now ready to formulate the main result of the paper.

Theorem 3.15. *For any $1\le p\le 2$, if P is $\Gamma_{d,d'}(p)$-weakly dependent and $P_{x_j|x^{(j-1)}}$ satisfies $\tilde T_{p,d}(C)$ or $\tilde T^{(i)}_{p,d}(C)$ for all $1\le j\le n$ then P satisfies $\tilde T_{p,d}(C\|\Gamma(p)\|^2_{2,p})$ or $\tilde T^{(i)}_{p,d}(C\|\Gamma(p)\|^2_{2,p})$ respectively.*

Remark 3.16. For $p=1$, one has the explicit form $\|\Gamma(p)\|^2_{2,1}=\sum_{k=1}^n(M+\sum_{j>k}\gamma_{j,k}(1))^2$. For $1\le p\le 2$ we have $n^{2/p-1}M^2\le\|\Gamma(p)\|^2_{2,p}\le n^{2/p-1}\|\Gamma(p)\|^2_{2,2}$. The weak transport inequalities can be dimension-free only when $p=2$. When $\gamma_{i,j}(p)=\gamma_{k,\ell}(p)$ for $j-i=k-\ell$ (for instance when the process (X_t) is stationary), from the estimate $\|\Gamma(p)\|_{2,2}\le M+\sum_{k=1}^{n-1}\gamma_{k,0}(p)$ we obtain $\|\Gamma(p)\|^2_{2,2}\le n^{2/p-1}(M+\sum_{k=1}^{n-1}\gamma_{k,0}(p))^2$. Theorem 3.15 yields weak transport inequalities $\tilde T_{p,d}(n^{2/p-1}C')$ or $\tilde T^{(i)}_{p,d}(n^{2/p-1}C')$ if $\sum_{k\ge 1}\gamma_{k,0}(p)<\infty$.

Remark 3.17. Under $\Gamma_{d,\mathbf{1}}(p)$-weak dependence, the results of Theorem 3.15 are easily extendable to non-product measures on $E_1 \times \cdots \times E_n$, where E_i are subspaces of E with finite diameters Δ_k, $1 \le k \le n$. It suffices to replace M by Δ_i on the diagonal of the matrix $\Gamma(p)$.

Proof. We only detail the proof of the weak transport inequality $\tilde T_{p,d}$ as the proof of the inverse weak transport inequality $\tilde T^{(i)}_{p,d}$ follows the same arguments.Let us fix a lower semi-continuous $\alpha \in M^+(E^n)^n$ such that $Q[\alpha_k^q] < \infty$ for all $1 \le k \le n$. The proof is based on the notion of $\Gamma_{d,d'}(p)$-weak dependence and on a recursive argument due to Lemma 3.19 stated below.

We will use the following Markov coupling $\tilde\pi$, defined as $\tilde\pi = \tilde\pi_{n|n-1} \cdots \tilde\pi_{2|1}\tilde\pi_{1|0} \in \tilde M(E^n)$. Here the $\tilde\pi_{i|i-1} = \tilde\pi_{x_i,y_i|x^{(i-1)},y^{(i-1)}}$s achieve the minimum in the auxiliary transport problems

$$\tilde\pi_{i|i-1}\Big[\sum_{k=i}^n Q[\alpha_k^q|Y_i, y^{(i-1)}]^{1/q}\gamma_{k,i}(p)d'(X_i,Y_i)\Big] = \tilde W_{Q[\alpha_k^q|\cdot,y^{(i-1)}]^{1/q}\gamma_{k,i}(p),d'}(P_{x_i|x^{(i-1)}}, Q_{y_i|y^{(i-1)}})$$

for any $x^{(i-1)}, y^{(i-1)}$ in E^{i-1}, $1 \le i \le n$. Such coupling exist thanks to the lower semi-continuity of the function $Q[\alpha_k^q|\cdot, y^{(i-1)}]^{1/q} \in M^+(E)$, for any $y^{(i-1)}$ in E^{i-1}. Moreover, due to the definition of the weak transport metric $\tilde W_{p,d'}(P_{x_i|x^{(i-1)}}, Q_{y_i|y^{(i-1)}})$, the coupling $\tilde\pi$ satisfies the following important relation, for any $x^{(i-1)}, y^{(i-1)}$ in E^{i-1}, $1 \le i \le n$,

$$\begin{aligned}
&\tilde\pi_{i|i-1}\Big[\sum_{k=i}^n Q[\alpha_k^q|Y_i, y^{(i-1)}]^{1/q}\gamma_{k,i}(p)d'(X_i,Y_i)\Big]\\
&\qquad\le Q\Big[\Big(\sum_{k=i}^n Q[\alpha_k^q|Y_i, y^{(i-1)}]\gamma_{k,i}(p)\Big)^q \mid y^{(i-1)}\Big]^{1/q}\tilde W_{p,d'}(P_{x_i|x^{(i-1)}}, Q_{y_i|y^{(i-1)}})\\
&\qquad\le \sum_{k=i}^n Q[\alpha_k^q|y^{(i-1)}]^{1/q}\gamma_{k,i}(p)\tilde W_{p,d'}(P_{x_i|x^{(i-1)}}, Q_{y_i|y^{(i-1)}}),
\end{aligned} \tag{3.9}$$

where the last inequality follows from the triangle inequality.

The crucial Lemma 3.19 below heavily relies on the necessary and sufficient conditions for the existence of Markov couplings due to Rüschendorf [47]. Let us recall the result in full generality. For any cost function $\sigma : E^n \times E^n \mapsto \mathbb{R}_+$, $n \ge 2$, the section of σ in $(x_1, y_1) \in E^2$ is defined as

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

Theorem 3.18 (Theorem 3 in [47])**.** *We have the equivalence between*

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

Let $\alpha_k^{(i)}$ denote the section of α_k in $y^{(i)}$ such that $\alpha_k^{(i)}(y_{i+1},\dots,y_n) = \alpha_k(y)$ and write $\alpha^{(i)} = (\alpha_k^{(i)})_{k>i}$. A simple corollary of this theorem is the following result, which will be used in our recursive argument:

Lemma 3.19. *For any* $1 \le i \le n$, $x^{(i-1)}$, $y^{(i-1)}$ *in* E^{i-1}, *let* P, $Q \in M(E^{n-i+1})$ *be decomposed as* $P_{|x^{(i-1)}} = P_iP_{|X_i,x^{(i-1)}}$ *and* $Q_{|y^{(i-1)}} = Q_iQ_{|Y_i,y^{(i-1)}}$ *for some* P_i, $Q_i \in M(E)$. *Then for any* $\alpha \in M^+(E^n)^n$ *and any coupling* $\pi_i \in \tilde M(P_i,Q_i)$ *we have*

$$\tilde W_{\alpha^{(i-1)},d}(P_{|x^{(i-1)}}, Q_{|y^{(i-1)}}) \le \pi_i[Q[\alpha_i|Y_i,y^{(i-1)}]d(X_i,Y_i) + \tilde W_{\alpha^{(i)},d}(P_{|X_i,y^{(i-1)}}, Q_{|Y_i,y^{(i-1)}})].$$

Proof. We assume that for almost all $x^{(i)}$, $y^{(i)} \in E$ we have $\tilde{W}_{\alpha^{(i)},d}(P_{|x^{(i)}}, Q_{|y^{(i)}}) < \infty$. Then, by lower semi-continuity, there exists $\pi^*_{|x^{(i)},y^{(i)}}$ such that

$$\pi^*_{|x^{(i)},y^{(i)}}\Big[\sum_{k=i+1}^{n} \alpha_k^{(i)}(Y_{i+1},\dots,Y_n)d(X_k,Y_k)\Big] = \tilde{W}_{\alpha^{(i)},d}(P_{|x^{(i)}}, Q_{|y^{(i)}}).$$

Note that for any $x^{(i)}$, $y^{(i)} \in E^i$ and any coupling $\pi_{|x^{(i)},y^{(i)}}$ of $(P_{|x^{(i)}}, Q_{|y^{(i)}})$ we have

$$\pi_{|x^{(i)},y^{(i)}}[\alpha_i^{(i)}(y_{i+1},\dots,y_n)d(x_i,y_i)] = \pi_{|x^{(i)},y^{(i)}}[\alpha_i^{(i)}(y_{i+1},\dots,y_n)]d(x_i,y_i) = Q[\alpha_i|y^{(i)}]d(x_i,y_i),$$

as the margins are fixed by definition of the coupling. Consider the cost function

$$\sigma_i((x_i,y_i),(x_k,y_k)_{i+1\le k\le n}) = \sum_{k=i}^{n} \alpha_k^{(i-1)}(y_i,\dots,y_n)d(x_k,y_k).$$

Then the section of σ_i in (x_i,y_i) is $\sum_{k=i}^n \alpha_k^{(i)}(y_{i+1},\dots,y_n)d(x_k,y_k)$. We can apply Theorem 3.18 and the solution of the optimization in 2.(a) of Theorem 3.18 is given by $\pi^*_{|x^{(i)},y^{(i)}}$ as this optimization does not depend on $\pi_{|x^{(i)},y^{(i)}}[\alpha_i^{(i)}(y_{i+1},\dots,y_n)d(x_i,y_i)]$ that is constraint to be equal to $Q[\alpha_i|y^{(i)}]d(x_i,y_i)$. The desired result follows optimizing 2.(b). □

We are now ready to describe the recursive argument that we will use. Applying Lemmas 3.19 and 2.5 for any $1 \le i \le n$, we obtain:

$$\begin{aligned}\tilde{W}_{\alpha^{(i-1)},d}(P_{|y^{(i-1)}}, Q_{|y^{(i-1)}}) &\le \tilde{\pi}_{i|i-1}[Q[\alpha_i|Y_i,y^{(i-1)}]d(X_i,Y_i) + \tilde{W}_{\alpha^{(i)},d}(P_{|X_i,y^{(i-1)}}, Q_{|Y_1,y^{(i-1)}})]\\ &\le \tilde{\pi}_{i|i-1}[Q[\alpha_i|Y_i,y^{(i-1)}]d(X_i,Y_i) + \tilde{W}_{\alpha^{(i)},d}(P_{|Y_i,y^{(i-1)}}, Q_{|Y_i,y^{(i-1)}})\\ &\quad + \tilde{W}_{\tilde{\alpha}^{(i)},d}(P_{|X_i,y^{(i-1)}}, P_{|Y_i,y^{(i-1)}})],\end{aligned} \tag{3.10}$$

where $\tilde{\alpha}^{(i)}$ satisfies $P_{|Y_i,y^{(i-1)}}[(\tilde{\alpha}_k^{(i)})^q] = P[\tilde{\alpha}_k^q \mid Y_i, y^{(i-1)}] \le Q[\alpha_k^q \mid Y_i, y^{(i-1)}]$, $k > i$. To bound the last term, we will use the definition of $\Gamma_{d,d'}(p)$-weak dependence. The $\Gamma_{d,d'}(p)$-weak dependence condition ensures the existence of a coupling $\pi_{|i}$ of $(P_{|x^{(i)}}, P_{y_i,x^{(i-1)}})$ satisfying (3.8). Using the duality in ^{p} and ℓ^p, it implies that

$$\sum_{k=i+1}^{n} \pi_{|i}[\alpha_k d(X_k,Y_k)] \le \sum_{k=i+1}^{n} P[\alpha_k^q|Y_i,y^{(i-1)}]^{1/q}\gamma_{k,i}(p)d'(x_i,y_i).$$

Then, for any $x^{(i)}$, $y^{(i)} \in E$ we obtain

$$\begin{aligned}\tilde{W}_{\alpha^{(i)},d}(P_{|x_i,y^{(i-1)}}, P_{|y^{(i)}}) &\le \pi_{|i}\Big[\sum_{k=i+1}^{n} \alpha_k d(X_k,Y_k)\Big]\\ &\le \sum_{k=i+1}^{n} P[\alpha_k^q|Y_i,y^{(i-1)}]^{1/q}\gamma_{k,i}(p)d'(x_i,y_i)\\ &\le \sum_{k=i+1}^{n} Q[\alpha_k^q \mid Y_i,y^{(i-1)}]^{1/q}\gamma_{k,i}(p)d'(x_i,y_i).\end{aligned} \tag{3.11}$$

Denoting $\gamma_{i,i}(p) = M$, the relations $d(X_i,Y_i) \le \gamma_{i,i}(p)d'(X_i,Y_i)$ hold by assumption for all $1 \le i \le n$. Collecting the inequalities (3.10) and (3.11), we obtain

$$\begin{aligned}\tilde{W}_{\alpha^{(i-1)},d}(P_{|y^{(i-1)}}, Q_{|y^{(i-1)}}) \le \tilde{\pi}_{i|i-1}\Big[&\sum_{k=i}^{n} Q[\alpha_k^q|Y_i,y^{(i-1)}]^{1/q}\gamma_{k,i}d'(X_i,Y_i)\\ &+ \tilde{W}_{\alpha^{(i)},d}(P_{|Y_i,y^{(i-1)}}, Q_{|Y_i,y^{(i-1)}})\Big].\end{aligned}$$

For the specific Markov coupling considered here, relation (3.9) holds. As its second margin is $Q_{y_i|y^{(i-1)}}$, we also have

$$\tilde{W}_{\alpha^{(i-1)},d}(P_{|y^{(i-1)}},Q_{|y^{(i-1)}}) \le \sum_{k=i}^{n} Q[\alpha_k^q|y^{(i-1)}]\gamma_{k,i}(p)\tilde{W}_{p,d'}(P_{x_i|y^{(i-1)}},Q_{y_i|y^{(i-1)}}) + Q[\tilde{W}_{\alpha^{(i)},d}(P_{|Y_i,y^{(i-1)}},Q_{|Y_i,y^{(i-1)}})].$$

We can now apply a recursive argument on $Q[\tilde{W}_{\alpha^{(i)},d}(P_{|Y^{(i)}},Q_{|Y^{(i)}})]$. Starting from the relation $\tilde{W}_{\alpha,d}(P,Q) = Q[\tilde{W}_{\alpha^{(0)},d}(P_{|y^{(0)}},Q_{|y^{(0)}})]$, we obtain

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

the second inequality following from the assumption $P_{x_j|y^{(j-1)}} \in \tilde{T}_{p,d}(C)$. Let $\mathbf{Q}$ be the vector $(Q[\alpha_k^q|Y^{(k-1)}]^{1/q})'_{1\le k\le n}$ and $\mathbf{W}$ the vector $((2C\mathcal{K}(P_{x_k|Y^{(k-1)}}|Q_{y_k|Y^{(k-1)}}))^{1/2})'_{1\le k\le n}$. With $<;>$ denoting the scalar product on $\mathbb{R}^n$, we obtain

$$\tilde{W}_{\alpha,d}(P,Q) \le Q[<\mathbf{Q};\Gamma(p)\mathbf{W}>] \le Q[\|\mathbf{Q}\|_q\|\Gamma(p)\mathbf{W}\|_p] \le Q[\|\mathbf{Q}\|_q^q]^{1/q}Q[\|\Gamma(p)\mathbf{W}\|_p^p]^{1/p},$$

where we used the Hölder's inequality twice. By definition of the matrix norm $\|\cdot\|_{2,p}$, we also have

$$\tilde{W}_{\alpha,d}(P,Q) \le Q[\|\mathbf{Q}\|_q^q]^{1/q}\|\Gamma(p)\|_{2,p}Q[\|\mathbf{W}\|_2^p]^{1/p}.$$

By definition, we identify the three terms in the upper bound

$$\begin{aligned}
Q[\|\mathbf{Q}\|_q^q] &= \sum_{k=1}^{n} Q[\alpha_k^q],\\
Q[\|\mathbf{W}\|_2^p] &= Q\Big[\Big(\sum_{k=1}^{n} 2C\mathcal{K}(Q_{y_k|Y^{(k-1)}}|P_{x_k|Y^{(k-1)}})]\Big)^{p/2}\Big],\\
\mathcal{K}(Q|P) &= \sum_{k=1}^{n} Q[2C\mathcal{K}(Q_{y_k|Y^{(k-1)}}|P_{x_k|Y^{(k-1)}})].
\end{aligned}$$

As $p/2 \le 1$, an application of Jensen's inequality yields

$$Q[\|\mathbf{W}\|_2^p] \le \Big(\sum_{k=1}^{n} Q[2C\mathcal{K}(Q_{y_k|Y^{(k-1)}}|P_{x_k|Y^{(k-1)}})]\Big)^{p/2} = (2C\mathcal{K}(Q|P))^{p/2},$$

where the last identity is the tensorization property of the entropy; see Lemma 1 in [48] for instance. Finally, we obtain

$$\frac{\sum_{k=1}^{n}\tilde{\pi}[\alpha_k(Y)d(X_k,Y_k)]}{(\sum_{k=1}^{n} Q[\alpha_k^q])^{1/q}} \le \sqrt{2C\|\Gamma(p)\|_{2,p}^2\mathcal{K}(Q|P)}.$$

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

4 Examples of $\Gamma_{d,d'}(p)$-weakly dependent processes

4.1 $\Gamma_{d,d'}(1)$-weakly dependent examples

When $p=1$, the dual form of $\tilde{T}_{1,d}(C\|\Gamma(1)\|_{2,1}^2) = \tilde{T}_{1,d}^i(C\|\Gamma(1)\|_{2,1}^2) = T_{1,d}(C\|\Gamma(1)\|_{2,1}^2)$ is the Hoeffding inequality which is not dimension-free; see Remark 3.16. We then recover concentration results that have been proved using the bounded difference approach of [44]. Applying Kantorovich-Rubinstein's inequality, we obtain an explicit upper-bound:

$$\sum_{k=i+1}^{n} \gamma_{k,i}(1) \le \sup_{f\in \mathrm{Lip}_1(d)} \sup_{x^{(i)},y_i} \frac{P[f(X_{i+1},\dots,X_n)|x^{(i)}] - P[f(X_{i+1},\dots,X_n)|y_i,x^{(i-1)}]|}{d'(x_i,y_i)}.$$

In the bounded case $d' = \mathbf{1}$, the $\Gamma_{d,\mathbf{1}}(1)$-weak dependence condition coincides with the one introduced by Rio [46] for more general spaces $E_1 \times \cdots \times E_n$, see Remark 3.17. As the conditional probabilities $P_{x_j|x^{(j-1)}}$ automatically satisfy Pinsker's inequality $\tilde{T}_{\mathbf{1},\mathbf{1}}(1/4)$, Theorem 3.15 recovers Hoeffding's inequality [46]. The context of $\Gamma_{N,\mathbf{1}}(1)$-weak dependence is extensive and we refer the reader to Section 7 of [18] for a detailed study and many examples including causal functions of stationary sequences, iterated random functions, Markov kernels and expanding maps.

When $d = d'$ the $\Gamma_{d,d}(1)$-weak dependence condition is implied by condition $(C_1)'$ of [20]: for any $f \in \mathrm{Lip}_1(d)$,

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

From Remark 3.16 we have $\|\Gamma(1)\|_{2,1} \le n(1+S)$ and thus Theorem 3.15 recovers the Hoeffding inequality of [20]. Examples of $\Gamma_{d,d}(1)$-weakly dependent time series are given in [20]. In particular, ARMA processes with sub-Gaussian innovations satisfy the conditions of Theorem 3.15 for $p = 1$, $d = d' = N$. Thus they satisfy Hoeffding's inequality.

4.2 $\Gamma_{\mathbf{1},\mathbf{1}}(p)$-weakly dependent examples

In the case $d = \mathbf{1}$, the best coupling scheme $\pi_{|i}$ is provided by the maximal coupling of [24] for any $1 \le i \le n$. We then have

$$Q[\pi_{|i}[d(X_k,Y_k) \mid Y]^p] = \int \Big(1 - \frac{dP_{|x^{(i)}}}{dP_{|y_i,x^{(i-1)}}}\Big)_+^p dP_{y_i,x^{(i-1)}}, \qquad i < k \le n,$$

and

$$\gamma_{k,i}(p)^p = \sup_{x_i\ne y_i} \int \Big(1 - \frac{dP_{|x^{(i)}}}{dP_{|y_i,x^{(i-1)}}}\Big)_+^p dP_{y_i,x^{(i-1)}}.$$

We recover the condition in [38] for contractive Markov chains and $p = 2$. We deduce from that expression the estimates

$$\sup_{x_i\ne y_i} \|P_{|x^{(i)}} - P_{|y_i,x^{(i-1)}}\|_{TV} \le \gamma_{k,i}(p) \le \sup_{x_i\ne y_i} \|P_{|x^{(i)}} - P_{|y_i,x^{(i-1)}}\|_{TV}^{1/p}$$

where $\|P - Q\|_{TV} = \sup_A |P(A) - Q(A)|$ for any distributions P and Q. The upper bound coincides with the coefficients introduced in [48] for $p = 2$. That the weak-dependence conditions here slightly improve those of [48] as discussed in [40]. In the stationary case, $\gamma_{k,i}(p)^p \le 2\phi_{k-i}$ where ϕ_k are the uniform mixing coefficients introduced in [29].

We extend the transport inequality of [48] for $p = 2$: as any $P_{x_j|x^{(j-1)}}$ satisfies $\tilde{T}_{2,\mathbf{1}}(1)$, Theorem 3.15 yields

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

In the stationary case, as $\|\tilde{\Gamma}(p)\|_{2,2}^2 \le 1 + \sum_{k=1}^n \gamma_{k,0}(2)$ we obtain

$$\inf_{\pi\in\tilde{M}} \Big(\sum_{i=1}^{n} Q[\pi[X_i \ne Y_i \mid Y_i]^p\Big)^{1/p} \le \Big(1 + \sum_{k=1}^{n} \tilde{\gamma}_{k,0}(2)\Big)\sqrt{2\mathcal{K}(Q|P)}.$$

This result can be extended to non-stationary sequences; see [31] for examples. When E is a real vector space, the choice of the Hamming distance is not natural and the resulting weak dependence conditions are often too restrictive.

4.3 $\Gamma_{N,N}(2)$-weakly dependent examples

In what follows, we show that the choice $d = N$ of the Euclidian metric is natural for many time series with state space $E = \mathbb{R}^K$, $K \geq 1$. We focus on two generic examples: the Stochastic Recurrent equations, treated in [20] for $p = 1$ only, and chains with infinite memory [21]. As we are not aware of an explicit expression of the $\gamma_{k,i}(2)$ coefficients when $d = N$, we use the natural coupling provided by the dynamics of the models to estimate them.

4.3.1 Stochastic Recurrent equations (SREs)

Define the SRE on E as (also called Iterated Random Functions in [22] and Random Dynamical Systems in [20])

$$X_0(x) := x \in E, \qquad X_{k+1}(x) = \psi_{k+1}(X_k(x)), \qquad k \geq 0, \tag{4.1}$$

where (ψ_k) is a sequence of iid random maps. By P we denote the distribution of the whole process $(\psi_k)_{k\geq 1}$. Assume in the next proposition that d and d' are any semi-lower continuous metrics satisfying $d \leq Md'$ for some $M > 0$.

Proposition 4.1. *For $1 \leq p \leq 2$, assume that the distribution of $\psi_1(x)$ satisfies $\tilde{T}_{p,d'}(C)$ or $\tilde{T}^{(i)}_{p,d'}(C)$ for any $x \in E$ and that there exists some $S > 0$ satisfying*

$$\sum_{k=1}^{\infty} P[d^p(X_k(x), X_k(y))]^{1/p} \leq Sd'(x,y), \qquad x, y \in E. \tag{4.2}$$

Then the distribution P^n_x of $(X_k(x))_{1\leq k\leq n}$ satisfies $\tilde{T}_{p,d}(C(M+S)^2 n^{2/p-1})$ or $\tilde{T}^{(i)}_{p,d}(C(M+S)^2 n^{2/p-1})$, for any $x \in E$, respectively.

Proof. The result is proved by an application of Theorem 3.15. The condition of $\Gamma_{d,d'}(p)$-weak dependence is satisfied because the joint law of $(X_k(x), X_k(y))_{t\geq 1}$ is a natural coupling scheme $\pi_{|0}$ of the law of $(X_t)_{t\geq 1}$ given that $(X_0, X_{-1}, X_{-2} \ldots) = (x, x_{-1}, x_{-2}, \ldots)$ and $(X_0, X_{-1}, X_{-2} \ldots) = (y, x_{-1}, x_{-2}, \ldots)$. Similarly, we obtain natural coupling schemes $\pi_{|i}$ for any $i \geq 0$ and the coefficients $\gamma_{k,i}$ satisfy the relation $\sum_{k>i} \gamma_{k,i}(p) \leq S$. Using similar arguments as in Remark 3.16, we obtain $\|\Gamma_{d,d'}(2)\|_p \leq M + S$. The result is proved because, by the Markov property, $P_{x_j|x^{(j-1)}}$ is the law of $\psi_j(x_{j-1})$ and satisfies $\tilde{T}_{p,d'}(C)$ or $\tilde{T}^{(i)}_{p,d'}(C)$ by assumption. □

Condition (4.2) becomes very tractable when $d = d' = N$ is the Euclidian norm of $E = \mathbb{R}^K$ as it is equivalent to the following Lyapunov condition in ^{p}:

Corollary 4.2. *Assume that the Lyapunov exponent satisfies*

$$\lambda_{max}({}^p) := \lim_{k\to\infty} \Big(\sup_{x\neq y} \frac{P[\|X_k(x) - X_k(y)\|^2]}{\|x-y\|^2} \Big)^{1/k}$$

is less than one. The distribution of $\psi_1(x)$ satisfies $\tilde{T}_{p,d'}(C)$ or $\tilde{T}^{(i)}_{p,d'}(C)$ for any $x \in E$ then there exists some $S > 0$ such that the distribution P^n_x of $(X_k(x))_{1\leq k\leq n}$ satisfies $\tilde{T}_{p,d}(C(M+S)^2 n^{2/p-1})$ or $\tilde{T}^{(i)}_{p,d}(C(M+S)^2 n^{2/p-1})$, for any $x \in E$, respectively.

This corollary answers an important question raised in Remark 3.6 of [20]: the dimension-free concentration properties (in a weak form) hold for SRE under Lyapunov's condition in ^{2}.

We consider in detail two classical SREs, ARMA models and general affine processes. The two first examples cannot be treated with the same generality by using the contractive conditions of [20, 41].

Example 4.3 (ARMA models)**.** Consider the ARMA model

$$X_0(x) = x, \qquad X_{k+1}(x) = AX_k(x) + \xi_{k+1}, \qquad k \geq 1,$$

in $E = \mathbb{R}^K$ where $A \in \mathcal{M}_{K,K}$ (the space of $K \times K$ matrices) and (ξ_k) is a sequence of iid 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 $\Gamma_{N,N}(p)$-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 existence of a causal solution (X_k) of the ARMA model. The conditions of Proposition 4.1 are satisfied if the law of ξ_1 satisfies $\tilde{T}_{p,N}(C)$ or $\tilde{T}^{(i)}_{p,N}(C)$. Thus it is in particular true for $p = 2$ for bounded or Gaussian innovations ξ_k. Note that the notion of $\Gamma_{N,N}(2)$-weak dependence is more general than the mixing ones. For instance, the solution of $X_{t+1} = 2^{-1}(X_t + \xi_{t+1})$ with $\xi_1 \sim \mathcal{B}(1/2)$ is $\Gamma_{N,N}(2)$-weakly dependent but not mixing; see [4].

Example 4.4 (General affine processes)**.** Consider now the specific SRE

$$X_0(x) = x, \quad X_{t+1}(x) = f(X_k(x)) + M(X_k(x))\xi_{k+1}, \qquad k \geq 1,$$

where $E = \mathbb{R}^K$, $\xi_t \in \mathbb{R}^{K'}$, $K' \geq 1$, $f : \mathbb{R}^K \mapsto \mathbb{R}^K$, $M : \mathbb{R}^K \mapsto \mathcal{M}_{K,K'}$ (the space of $K \times K'$ matrices) and the innovations ξ_k are iid random vectors of $\mathbb{R}^{K'}$. Fix $p = 2$ and assume that:

- (a) $P_\xi \in \tilde{T}_{2,N}(C)$ or $\tilde{T}^{(i)}_{2,N}(C)$ on $\mathbb{R}^{K'}$;
- (b) $\|M(x)\| \leq C'$, $\forall x \in \mathbb{R}^K$, $C' > 0$, $\|\cdot\|$ denoting also the operator norm on $\mathcal{M}_{K,K'}$ associated with the Euclidian norms of $\mathbb{R}^K$ and $\mathbb{R}^{K'}$;
- (c) the Lyapunov condition $\lambda_{max}(^2) < 1$ is satisfied.

Using Lemma 2.1 in [20] we obtain that the conditions (a) and (b) implies that $P_{x_i|x_{i-1}}$ satisfies $\tilde{T}_{2,N}(CK^2)$ or $\tilde{T}^{(i)}_{2,N}(CK^2)$. Moreover, condition (4.2) is satisfied for some $S > 0$ and thus P^n_x satisfies $\tilde{T}_{2,N}(CK^2(1+S)^2)$ or $\tilde{T}^{(i)}_{2,N}(CK^2(1+S)^2)$ for any $x \in E$.

4.3.2 Chains with Infinite Memory

Here we assume that $d = d' = d$ is any semi lower-continuous distance (not necessarily the Euclidian norm). Consider chains with infinite memory defined in [21] for any function $F :\ E^{\mathbb{N}} \times \mathcal{X} \mapsto E$ by the relation:

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

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

$$X_k(y_0, x^{(-1)}) = F(X_{k-1}(y, x^{(-1)}), X_{k-2}(y_0, x^{(-1)}), \ldots, Y_1, y_0, x_{-1}, x_{-2}, \ldots; \xi_k), \qquad \forall k \geq 1,$$

where the innovations $(\xi_k)_{k\geq 1}$ are the same as in (4.3). Then the natural coupling scheme $\pi_{|0}$ is the distribution of $(X_k(x), X_k(y_0, x^{(-1)}))_{k\geq 1}$. By P denote the law of the innovations process (ξ_k) and P_x^n the law of $(X_k(x))_{1\leq k\leq n}$ on E^n.

Proposition 4.5. *Assume there exists a sequence of non-negative numbers* (a_i) *such that* $\sum_{i\geq 1} a_i = a < 1$, $\sum_{i\geq 1} i\log(i)a_i < \infty$ *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* $F(x; \xi_1)$, $x \in E^{\mathbb{N}}$, *satisfies* $\tilde{T}_{p,d}(C)$ *or* $\tilde{T}^{(i)}_{p,d}(C)$ *then there exists* $C' > 0$ *such that* P_x^n *satisfies* $\tilde{T}_{p,d}(C' n^{2/p-1})$ *or* $\tilde{T}^{(i)}_{p,d}(C' n^{2/p-1})$, *respectively.*

Proof. We estimate the coefficients $\gamma_{k,0}(p)$, $k \geq 0$, only, as the same reasoning holds for $\gamma_{k+i,i}(p)$, $i \geq 1$. By (4.4) we can choose $\gamma_{1,0}(p) = a_1$ as

$$P[d^p(X_1(x), X_1(y_0, x^{(-1)}))]^{1/p} = P[d(F(x_0, x^{(-1)}; \xi_1), F(y_0, x^{(-1)}; \xi_1))^p]^{1/p} \leq a_1 d(x_0, y_0).$$

Applying (4.4) recursively, we obtain the existence of the coefficients $\gamma_{k,0}(p)$, satisfying the relation

$$\gamma_{k,0}(p) \leq \sum_{j=1}^{k} a_j \gamma_{k-j,0}(p), \qquad k \geq 1, \qquad \gamma_{0,0}(p) = 1 \text{ by convention.}$$

Arguments similar to the proof of Theorem 3.1 in [21] yield

$$\gamma_{k,0}(p) \leq \gamma_{1,0}(p) \inf_{1\leq p\leq t} \Big\{ a^{t/p} + \sum_{j\geq p} a_j \Big\}.$$

The desired result follows by choosing $p = cr/\log(r)$ such that $\sum_{t\geq 1} \gamma_{k,0}(p) < \infty$. ☐

Example 4.6 (AR(∞) models)**.** As an example of chains with infinite memory in $E = \mathbb{R}$, considerthe stationary solution to the autoregressive equation

$$X_t = \sum_{i\geq 1} a_i X_{t-i} + \xi_t, \qquad t \in \mathbb{Z},$$

where the real numbers a_i are such that $\sum_{i\geq 1} |a_i| < 1$ and $\sum_{i\geq 1} i\log(i)|a_i| < \infty$. Then, if ξ_1 satisfies $\tilde{T}_{2,N}(C)$ or $\tilde{T}^{(i)}_{2,N}(C)$, the distribution of $(X_1, \dots, X_n)$ satisfies $\tilde{T}_{2,N}(C')$ or $T^{(i)}_{2,N}(C')$, $C' > 0$ for any $n \geq 1$ given the past $(X_0, X_{-1}, \dots)$ a.s..

Example 4.7 (General affine processes with infinite memory)**.** Consider the process on $E = \mathbb{R}^K$ defined as the stationary solution of the equation

$$X_k = f(X_{k-1}, X_{k-2}, X_{k-3}, \dots) + M(X_{k-1}, X_{k-2}, X_{k-3}, \dots)\xi_k, \qquad \forall k \geq 1$$

where f and M are Lipschitz continuous functions with respect to the Euclidian norms in $\mathbb{R}^K$ and $\mathcal{M}(K, K')$ respectively. These general affine models include many classical econometric models and are estimated efficiently by the quasi maximum likelihood estimator [6]. Let Ψ denote either f or M and $(\alpha_i(\Psi))_{i\geq 1}$ be the Lipschitz coefficients

$$\|\Psi(x) - \Psi(y)\| \leq \sum_{i\geq 1} \alpha_i(\Psi)\|x_i - y_i\|, \qquad \forall x,\ y \in E^{\mathbb{N}}.$$

If the condition (a) of Example 4.4 is satisfied and $\|M(x)\| \leq C'$, $\forall x \in E^{\mathbb{N}}$, $C' > 0$, $\sum_{i\geq 1} \alpha_i(f) + P_\xi[\xi^2]^{1/2}\alpha_i(M) < 1$ and $\sum_{i\geq 1} i\log(i)(\alpha_i(f) + P_\xi[\xi^2]^{1/2}\alpha_i(M)) < \infty$ then the distribution of $(X_1, \dots, X_n)$ satisfies $\tilde{T}_{2,N}(C'')$ or $T^{(i)}_{2,N}(C')$, $C'' > 0$, for any $n \geq 1$.

5 Applications to oracle inequalities with fast convergence rates

In this section, we use the weak transport approach to obtain new oracle inequalities with fast rate of convergence in the $\Gamma(2)$-weakly dependent setting with $d = N$ and $d = \mathbf{1}$. Instead of using the classical approach based on exponential concentration inequalities as those given in Propositions 3.5 and 3.12, we prefer to use a more direct approach based on the new concept of conditional weak transport inequalities introduced in Section 5.2.

5.1 The statistical setting

We focus on oracle inequalities for the ordinary least squares estimator. We consider the case of linear regressiobyn, where $X = (Y, Z) = (Y, Z^{(1)}, \dots, Z^{(d)})$ on $\mathbb{R}^{K+1}$, $K > 0$. 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 and $\theta \in \mathbb{R}^K$ is a parameter. In our context, the observations are not necessarily independent nor identically distributed. We denote by P their distribution and we assume that P satisfies $T_{2,d}(C)$ and $T_{2,d}^{(i)}(C)$ for some $C > 0$ independent of $n \ge 1$, with $d = N$ or $\mathbf{1}$. In view of Theorem 3.15 and since $p = 2$, one can get such dimension-free concentration inequalities under $\Gamma(p)$-weak dependence. But there is a tradeoff due to the two possible choices of d: if $d = N$ then the setting is restricted to sub-Gaussian marginals but the dependence in the observations includes most of the classical time series. On the contrary, if $d = \mathbf{1}$ there is no assumption on the margins but the observations are restricted to be uniformly mixing.

The risk of prediction is defined as

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

The aim is to approximate the predictive performance of the oracle, the value $\overline{\theta} \in \mathbb{R}^K$ such that $R(\overline{\theta}) \le R(\theta)$, $\theta \in \mathbb{R}^K$. We consider the Ordinary Least Square (OLS) estimator $\hat{\theta}$ of $\overline{\theta}$ defined as $r(\hat{\theta}) \le r(\theta)$ for all $\theta \in \mathbb{R}^d$. We 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 $n \times K$ 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 matrix. Assume that G is a definite positive matrix and denote $\rho = \max(1, \rho_{sp}(G^{-1}))$ when $d = N$ or $\rho = \max(1, \max_{i,j}|G_{i,j}^{-1}|)$ when $d = N$. Then the change of variables $(\mathcal{Z}, \theta) \to (\mathcal{Z}G^{-1/2}, G^{1/2}\theta)$ is a $\sqrt{\rho}$-Lipschitz function for both choices of metric. Thus $\mathcal{Z}G^{-1/2}$ satisfies $\tilde{T}_{2,N}(\rho C)$ and $\tilde{T}_{2,N}^{(i)}(\rho C)$ by an application of Lemma 2.1 of [20]. To simplify the presentation, we thus consider $G = I_K$, the identity matrix on R^K, $\mathcal{Z} \in \tilde{T}_{2,N}(\rho C)$ and $\tilde{T}_{2,N}^{(i)}(\rho C)$. With this notation, $P[\|\mathcal{Z}\|_n^2] = K$ and $\|\hat{\theta} - \overline{\theta}\|^2 = R(\hat{\theta}) - R(\overline{\theta})$.

5.2 Conditional weak transport inequalities

We recall the classical approach based on the empirical process concentration to motivate our new approach. Following [42], oracle inequalities will follow from the concentration properties of $\overline{r}(\hat{\theta})$. However, as the distribution of $\hat{\theta}$ is difficult to deal with, one studies the concentration of the supremum of the empirical process

$$f(X_1, \dots, X_n) = \sup_{\theta\in\Theta}\{\overline{r}(\theta) - \overline{R}(\theta)\}.$$

If f is a self-bounding function (for $d = \mathbf{1}$) one can use the weak transport to extend Bernstein inequalities for f to dependent settings; see [48]. To obtain oracle inequalities, one studies the expected value of the supremum of the empirical process. In an independent context, a classical solution consists of using a chain argument and a metric entropy approach; see Chapter 13 of [16]. In dependent settings, it is not an easy task because the metric entropy depends on the mixing properties; see [45].

We use a different approach that does not deal with the supremum of the empirical process. Following the PAC-Bayesian approach [43, 17], the idea is to consider probability measures ρ_θ centered at θ. Denote any measure on $E^n \times \Theta$ by $Q\nu$ where $Q \in M^+(E^n)$ is the distribution of $(Y_1, \dots, Y_n)$ and ν is defined conditional on the Y_ks. The concentration properties of $\overline{r}(\hat\theta)$ will follow from the weak transport properties of the measure $P\rho_{\hat\theta}$ on $E^n \times \Theta$. We still denote by $d = N$ or $d = \mathbf{1}$ the metric chosen on that product space. Notice that $\rho_{\hat\theta}$ is a probability measure defined conditional on the observations X_k. Thus, the properties of the measure $P\rho_{\hat\theta}$ are not easily handle in a direct way. The PAC-Bayesian approach consists of introducing a prior measure $\rho_{\overline{\theta}}$ that does not depend on the X_ks. The conditional weak transport approach then extends the transport from P for the metric $d_\theta(x, y) = d((x, \theta), (y, \theta))$ to $\rho_{\overline{\theta}} P$ for d. Then we transport $P\rho_{\hat\theta}$ to $Q\rho_{\hat\theta}$, for any Q.

Proposition 5.1. *If P satisfies $\tilde T_{p,d_\theta}(C_\theta)$ or $\tilde T^{(i)}_{p,d_\theta}(C_\theta)$, $\theta \in \Theta$, then for any $\mu \in M^+(\Theta)$, $\mu \otimes P$ satisfies $\tilde T_{p,d}$ or $\tilde T^{(i)}_{p,d}$ with constant $\mu[C_\theta]$ for $p = 1$ and $\sup_\Theta C_\theta$ for $1 < p \le 2$.*

Remark 5.2. The result does not depend on the transport properties of μ. In the case $d = \mathbf{1}$ or $d = N$ on E^n and on $E^n \times \Theta$, if P satisfies $\tilde T_{p,d}(C)$ or $\tilde T^{(i)}_{p,d}(C)$ then it satisfies also $\tilde T_{p,d_\theta}(C)$ or $\tilde T^{(i)}_{p,d_\theta}(C)$ because d_θ is a 1-Lipschitz function for any $\theta \in \Theta$.

Proof. By the proof of Theorem 3.1, $\tilde T_{p,d}$ is equivalent to the linear inequality

$$Q\nu[\lambda(f_{\alpha,d} - \lambda C_\theta \alpha^q/q)] \le (1/p - 1/2) C_\theta \lambda^2 + \mathcal{K}(Q\nu|\mu \otimes P) + \mu \otimes P[\lambda f].$$

Denote Q_θ the conditional probability measure such that $\mu Q_\theta = Q\nu$. By virtue of $\tilde T_{p,d_\theta}(C_\theta)$ we have

$$Q_\theta[\lambda(f_{\alpha,d} - \lambda \sup_\Theta C_\theta \alpha^q/q)] \le Q_\theta[\lambda(f_{\alpha,d} - \lambda C_\theta \alpha^q/q)] \le (1/p - 1/2) C_\theta \lambda^2 + \mathcal{K}(Q_\theta|P) + P[\lambda f].$$

We obtain the desired result by linearity, integrating with respect to μ and observing that $\mathcal{K}(Q\nu|\mu \otimes P) = \mathcal{K}(\mu Q_\theta|\mu \otimes P) = \mu[\mathcal{K}(Q_\theta|P)]$ for $1 < p < 2$. For $p = 1$, we notice that $\lambda C_\theta \alpha^q/q = 0$ by convention and the desired result follows in a similar way. □

Using this approach, we will obtain oracle inequalities noticing that $f = \overline{r} - \overline{R}$ or $f = r - R$ has nice "self-bounding" properties for $d_\theta = \mathbf{1}$ or $d = N$, respectively.

5.3 A non-exact oracle inequality for $\Gamma_{N,N}(2)$-weakly dependent sequences

Our first result is a bound on the excess of risk of the OLS estimator for $\Gamma_{N,N}(2)$-weakly dependent observations X_k, $1 \le k \le n$. We first give an oracle inequality that follows from the conditional weak transport described above:

Theorem 5.3. *Assume that $X = (X_1, \dots, X_n)$ satisfies $T_{2,N}(\rho C)$ and $T^{(i)}_{2,N}(\rho C)$ for $\rho \ge 1$ and $C > 0$. For any measure Q we have*

$$Q[\overline{R}(\hat\theta)] \le Q[\|\mathcal{Z}\|_n^2]/\beta + 4\sqrt{\rho C Q[L] n^{-1}(\mathcal{K}(Q|P) + \beta Q[\overline{R}(\hat\theta)]/2)}, \qquad \beta > 0, \tag{5.1}$$

where

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

Proof. We first study the self-bounding properties of $f = \overline{r}$. Using the inequality $x^2 - y^2 \le 2x(x-y) \le 2\|x\|\|x-y\|$ for any $x, y \in \mathbb{R}$ we obtain

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

We apply the conditional weak transport approach, suppressing $d = N$ in the notation. By definition of $\tilde{W}_2$ in (2.7) and using the Cauchy-Schwarz inequality, we have for any Q_θ defined given θ

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

As P satisfies $\tilde{T}_{2,N}(\rho C)$ and $\tilde{T}^{(i)}_{2,N}(\rho C)$, again applying the Cauchy-Schwarz inequality, we obtain

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

Let ρ_θ denote $\mathcal{N}_d(\theta, \beta^{-1}I_K)$ for any $\beta > 0$. Integration with respect to $\rho_{\overline{\theta}}$ yields

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

Choosing Q_θ such as $\rho_{\overline{\theta}}Q_\theta = Q\rho_{\hat{\theta}}$, we have $\rho_{\overline{\theta}}[\mathcal{K}(Q_\theta|P)] = \mathcal{K}(Q|P) + Q[\mathcal{K}(\rho_{\hat{\theta}}|\rho_{\overline{\theta}})]$. From the entropy of the Gaussian distributions, we have $\mathcal{K}(\rho_{\hat{\theta}}|\rho_{\overline{\theta}})] = \beta/2\|\hat{\theta} - \overline{\theta}\| = \beta/2(R(\hat{\theta}) - R(\overline{\theta}))$. Collecting these identities, we obtain

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

By Jensen's inequality $Q\rho_{\hat{\theta}}[R(\theta)] \ge Q[R(\hat{\theta})]$. Standard computations on the Gaussian distribution yield $Q\rho_{\hat{\theta}}[r(\theta)] \le r(\hat{\theta}) + Q[\|\mathcal{Z}\|_n^2]/\beta \le r(\overline{\theta}) + Q[\|\mathcal{Z}\|_n^2]/\beta$. Collecting these bounds, we obtain

$$Q[R(\hat{\theta}) - R(\overline{\theta}) - \|\mathcal{Z}\|_n^2]/\beta] \le$$
$$4\sqrt{\rho C n^{-1}(\mathcal{K}(Q|P) + \beta/2Q[R(\hat{\theta}) - R(\overline{\theta})])\rho_{\overline{\theta}}[(1+\|\theta)\|^2)R(\theta)] + (1+\|\overline{\theta}\|^2)Q[r(\overline{\theta})])}. \quad (5.2)$$

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

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

Then we can decompose the last term of the sum as follows

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

where $\mathcal{Y} = (Y_1, \ldots, Y_n)$. Standard computations on the Gaussian distribution yield

$$\rho_{\overline{\theta}}[R(\theta)] = R(\overline{\theta}) + K/\beta$$
$$\rho_{\overline{\theta}}[\|\theta\|^2] = \|\overline{\theta}\|^2 + K/\beta$$
$$\rho_{\overline{\theta}}[\|\theta\|^2(\theta - \overline{\theta})] = 2\overline{\theta}/\beta$$
$$\rho_{\overline{\theta}}[\|\theta\|^2\|\theta - \overline{\theta}\|^2] = (\|\overline{\theta}\|^2 + K/\beta)(K-1)/\beta + \|\overline{\theta}\|^2/\beta + 3K/\beta.$$

The desired result follows by plugging these identities into (5.2) and noticing that $4P[\mathcal{Y}\mathcal{Z}]\overline{\theta} \le 2nR(\overline{\theta})$. ☐

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

$$Q\mu[\overline{R}(\theta)] \leq Q\mu[\overline{r}(\theta)] + 4\sqrt{\rho C n^{-1}\mathcal{K}(Q\mu|P\nu)(\nu[(1+\|\theta\|^2)R(\theta)] + (1+\|\overline{\theta}\|^2)Q[r(\overline{\theta})]}. \quad (5.3)$$

We discuss the choice $\mu = \rho_{\hat{\theta}}$ and $\nu = \rho_{\overline{\theta}}$ made above. The goal is to obtain an upper bound on $Q[R(\hat{\theta})]$ with $Q\mu[\overline{r}(\theta)]$ as small as possible. As soon as μ is centered at $\hat{\theta}$, Jensen's inequality yields $Q\mu[R(\theta)] \geq Q[R(\hat{\theta})]$. If μ is concentrated sufficiently close to $\hat{\theta}$ then $Q\mu[r(\theta) - r(\overline{\theta})]$ is small as $r(\hat{\theta}) - r(\overline{\theta}) < 0$. One cannot choose μ as the Dirac measure at $\hat{\theta}$ in view of 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_k forces us to choose a measure supported by the whole space $\mathbb{R}^K$. The measure μ could be chosen in order to optimize the upper bound explicitly. Due to the presence of $\mathcal{K}(\mu,\nu)$ in the upper bound, it provides Gibbs estimators that are nice alternatives to the OLS; see Chapter 4 of the textbook of Catoni [17] 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 $Q\mu[r(\theta) - r(\overline{\theta})]$ by $Q[\|\mathcal{Z}\|_n^2]/\beta$. This term can easily be estimated by $P[\|\mathcal{Z}\|_n^2]/\beta = K/\beta$ plus a concentration term implying the entropy $\mathcal{K}(Q|P)$. Thus we obtain a non-exact oracle inequality

Corollary 5.4. *For any $\eta > 0$ such that $(K+2)/n < \eta < 1$ we have with probability $1-\varepsilon$, $0 < \varepsilon < 1$,*

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

where $B_1 = 2(3 + 2\|\overline{\theta}\|^2 + \eta/n)$, $B_2 = 2(5 + \|\overline{\theta}\|^2)$ and $B_3 = 2(K(K-1) + K/n)$.

Remark 5.5. This result extends some non-exact oracle inequalities of [32] from the iid to a dependent context. Such non-exact oracles inequalities are preliminary results as they are useful only if $R(\overline{\theta})$ is small. They are usually complemented by a model selection procedure and yield optimal non-exact oracle inequalities for the model selection procedure; see [32].

Proof. Applying Young's inequality to (5.1), we obtain

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

Notice that, by definition of L, we have

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

By similar arguments as in the proof of Theorem 5.3 we obtain

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

Since $P[\|\mathcal{Z}\|_n^2] = K$, we also have

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

Collecting these bounds and using the Cauchy-Schwarz inequality, we get

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

Using again Young's inequality, by definition of B_1, B_2 and B_3 we obtain

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

Choose Q as the probability P restricted to the complement of A, the event corresponding to the desired oracle inequality. On this complement, we have

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

Combining the last two inequalities, we obtain $-\log(\varepsilon)\le\mathcal{K}(Q|P)$. By definition of Q, the relative entropy is explicitly given by $\mathcal{K}(Q|P)=-\log(1-P(A))$. The desired result follows. □

5.4 Exact oracle inequalities for $\Gamma_{\mathbf{1},\mathbf{1}}(2)$-weakly-dependent sequences

Now we provide an equivalent form of the conditional weak transport inequality (5.3) when $d=d'=\mathbf{1}$. Then any function f has the following "self-bounding" property $f(x)-f(y)\le|f(x)|1_{x\ne y}+|f(y)|1_{x\ne y}$. Following the lines of the proof of (5.3) with $f=\overline{r}$, we obtain

$$Q\mu[\overline{R}]\le Q\mu[\overline{r}]+2\sqrt{2\rho C\mathcal{K}(Q\mu|P\nu)(P\nu[\overline{r}^2]+Q\mu[\overline{r}^2])}. \tag{5.4}$$

We again choose $\mu=\rho_{\hat\theta}$ and $\nu=\rho_{\overline{\theta}}$ such that we can use Lemma 1.2 in the supplementary material of [5]: for any $\theta\in\mathbb{R}^K$ we have

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

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 in full generality. We will work under the so-called Bernstein condition that bounds the variance of the excess of risk by its expectation [7]. This condition links the set of parameters $\Theta\subseteq\mathbb{R}^K$ and the support of P: there exists some finite $B>0$ satisfying

$$B=\sup_{\theta\in\Theta}\frac{\sum_{i=1}^n\|Z_i\theta\|_\infty}{\sum_{i=1}^n P[Z_i\theta]^2}. \tag{5.5}$$

In the iid case this Bernstein condition was used in [5]. Under (5.5) and as $P[\|\mathcal{Z}\|_n^2]=K$, we also have $\|\mathcal{Z}\|_n^2\le BK$ and $\|\mathcal{Z}\|_n^4\le(BK)^2$. Moreover, using computations given in the supplementary material of [5], we have

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

Collecting these bounds, we get the following conditional weak transport result for $d=\mathbf{1}$ that improves Theorem 5.3:

Theorem 5.6. *If condition* (5.5) *holds then*

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

In the above bound the terms involving $r(\overline{\theta})$ are nuisance terms without additional condition 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 upper bound provided by Bernstein's inequality under conditions as [5]. Indeed, this conditions estimate the variance term by $\overline{R}(\hat\theta)$; that is why they were called after Bernstein in [7]. The fact that the excess risk also appears in the estimate is a major advantage when considering the Hamming distance (the Bernstein's inequality) compared with the Euclidian distance (the Tsirel'son's inequality); there the term $Q[R(\hat\theta)]\ge Q[\overline{R}(\hat\theta)]$ appeared because $\overline{r}$ is "self-bounding" when $d=\mathbf{1}$ but only r is "self-bounding" when $d=N$. As $Q[\overline{R}(\hat\theta)]$ is the quantity of interest, we obtain

Corollary 5.7. *If condition* (5.5) *holds then we have with probability* $1-\varepsilon$, $0<\varepsilon<1$,

$$R(\hat\theta)\le R(\bar\theta)+160\frac{B^2+4BM}{n}\times$$
$$\times\Big(BK+8\rho C(\log(\varepsilon^{-1})-\log P(r(\bar\theta)>M))+\frac{K(R(\bar\theta)+M)}{10B+40M}+\frac{8(BK)^2}{n}\Big),\qquad M>0.$$

Remark 5.8. Under (5.5), the exact oracle inequality above holds for any $\Gamma_{\mathbf{1},\mathbf{1}}(2)$-weakly dependent sequence without assumptions on the marginals (because any probability measure satisfies $\tilde T_{2,\mathbf{1}}(1)$). These oracle inequalities are new, even in the iid case. We refer the reader to [5] for estimates of the term $\log P(r(\bar\theta)>M)$ in the iid case under finite moments of order 4 only.

Proof. We write $A=\{r(\bar\theta)\le M\}$ and denote the restriction of P to A by P_A, i.e. $P_A(B)=P(B\cap A)$ for any measurable set B on E^n. We do not know wether P_A satisfies weak transport inequalities. However, a similar reasoning as for deriving (5.4) yields

$$Q[\overline R(\hat\theta)]\le BK/\beta+\rho_{\hat\theta}\Big[\sqrt{(4BK\overline R(\bar\theta)/\beta+(4BK/\beta)^2)n^{-1}\tilde W_2(Q_\theta,P_A)}$$
$$+\sqrt{((10B^2+40BM)Q[\overline R(\hat\theta)]+4BKM/\beta+(4BK/\beta)^2)n^{-1}\tilde W_2(P_A,Q_\theta)}\Big].$$

We use the triangle inequality of the weak transport cost (2.9):

$$\tilde W_2(P_A,Q_\theta)\le\tilde W_2(P_A,P)+\tilde W_2(P,Q_\theta),$$
$$\tilde W_2(Q_\theta,P_A)\le\tilde W_2(Q_\theta,P)+\tilde W_2(P,P_A).$$

Because P satisfies $\tilde T_2(\rho C)$ and $\tilde T_2^{(i)}(\rho C)$, both RHS terms are estimated by

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

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

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

Using Young's inequality with $\lambda=\beta=n(40B^2+160BM)^{-1}$, we get

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

We conclude choosing Q as P restricted to the complement of the event corresponding to the desired oracle inequality. □

Acknowledgments

I thank the GT EVOL's team, especially F. Bolley, I. Gentil, C. Léonard 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 [39]. Finally, I would like to deeply thank F. Merlevède, T. Mikosch and an anonymous referee for helpful comments.

References

[1] Adamczak, R. (2008) A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13**, 1000–1034. MR-2424985

[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. MR-2948906

[3] Alquier, P. and Wintenberger, O. (2012) Model selection for weakly dependent time series forecasting, *Bernoulli* **18 (3)**, 883–913. MR-2948906

[4] Andrews, D.W.K. (1984) Nonstrong mixing autoregressive processes. *J. Appl. Probab.* **21**, 930–934. MR-0766830

[5] Audibert, J.-Y. and Catoni, O. (2011) Robust linear least squares regression *Ann. Statist.* **39 (5)**, 2766–2794. MR-2906886

[6] Bardet, J.-M. and Wintenberger, O. (2009) Asymptotic normality of the Quasi Maximum Likelihood Estimator for multidimensional causal processes. *Ann. Statist.* **37**, 2730–2759. MR-2541445

[7] Bartlett, P. L. and Mendelson, S. (2006) Empirical minimization. *Probab. Theory Related Fields* **135**, 311–334. MR-2240689

[8] Bertail, P. and Clémencon, S. (2010) Sharp bounds for the tails of functionals of Markov chains. *Theory Probab. Appl.* **54**, 505–515. MR-2766354

[9] Bobkov, S. and Ledoux, M. (1997) Poincaré's inequalities and Talagrand's concentration phenomenon for the exponential distribution. *Probab. Theory and Related Fields*, **107**, 383-400. MR-1440138

[10] Bobkov, S. G. and Götze, F. (1999) Exponential integrability and transportation cost related to logaritheoremic Sobolev inequalities. *J. Funct. Anal.* **163**, 1–28. MR-1682772

[11] Bobkov, S. G. and Götze, F. (1999) Discrete isoperimetric and Poincaré-type inequalities. *Probab. Theory Related Fields* **114**, 245–277. MR-1701522

[12] Bobkov, S. G., Gentil, I., Ledoux, M. (2001) Hypercontractivity of Hamilton-Jacobi equations. *J. Math. Pures Appl.* **80**, 669–696. MR-1846020

[13] Bolley, F. and Villani, C. (2005) Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Ann. Fac. Sci. Toulouse Math.* **14** 331–352. MR-2172583

[14] Boucheron, S., Lugosi, G. and Massart, P. (2009) On concentration of self-bounding functions. *Electron. J. Probab.* **14**, 1884–1899. MR-2540852

[15] Boucheron, S., Lugosi, G. and Massart, P. (2000) A sharp concentration inequality with applications. *Random Struct. Algor.* **16**, 277–292. MR-1749290

[16] Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press. MR-3185193

[17] Catoni, O. (2004) *Statistical Learning Theory and Stochastic Optimization*, Lecture Notes in Mathematics. Springer, Berlin. MR-2163920

[18] Dedecker, J. and Prieur, C. (2005) New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields*, **132**, 203–236. MR-2199291

[19] 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. MR-2283252

[20] Djellout, H., Guillin, A. and Wu, L. (2004) Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Ann. Probab.* **32**, 2702–2732. MR-2078555

[21] Doukhan, P. and Wintenberger, O. (2008) Weakly dependent chains with infinite memory. *Stochastic Process. Appl.* **118**, 1997–2013. MR-2462284

[22] Duflo, M. (1997) Random iterative models. *Applications of Mathematics*, **34**. Springer-Verlag, Berlin. MR-1485774

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

[24] Goldstein, S. (1979) Maximal coupling. *Z. Wahrsch. Verw. Gebiete* **46**, 193–204. MR-0516740

[25] Gozlan, N. and Léonard, C. (2010) Transport inequalities. A survey. *Markov Processes and Related Fields*, **16**, 635–736. MR-2895086

[26] Gozlan, N., Roberto, C., Samson, P.-M. et Tetali, P. (2012) Displacement convexity of entropy and related inequalities on graphs. Preprint. MR-3256809

[27] Gozlan, N. (2012) Transport-Entropy inequalities on the line. *Electron. J. Probab*, **17**, 1–18. MR-2946156

[28] Guillin, A., Léonard, C., Wu, L. and Yao, N. Transportation-information inequalities for Markov processes. *Probab. Theory Related Fields* **144**, 669–695. MR-2496446

[29] Ibragimov, I. A. (1962) Some limit theorems for stationary processes. *Theory Probab. Appl.* **7**, 349–382. MR-0148125

[30] Joulin, A. and Ollivier, Y. (2010) Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.* **38**, 2418–2442. MR-2683634

[31] Kontorovich, A. and Ramanan, K. (2008) Concentration Inequalities for Dependent Random Variables via the Martingale Method. *Ann. Probab.* **36**, 2126–2158. MR-2478678

[32] Lecué, G. and Mendelson, S. (2012) General nonexact oracle inequalities for classes with a subexponential envelope. *Ann. Statist.* **40 (2)**, 832–860. MR-2933668

[33] Ledoux, M. (1996) On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.* **1**, 63–87. MR-1399224

[34] Ledoux, M. (2005) *The Concentration of Measure Phenomenon* **89**, AMS Bookstore.

[35] Lezaud, P. (1998) Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.* **8**, 849–867. MR-1627795

[36] Maurey, B. (1991) Some deviation inequalities. *GAFA*, **1**, 188–197. MR-1097258

[37] Marton K. (1986) A simple proof of the blowing-up lemma. *IEEE Trans. Inform. Theory* **32**, 445–446. MR-0838213

[38] Marton K. (1996) Bounding $\overline{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.* **24**, 857–866. MR-1404531

[39] Marton, K. (1996) A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.* **6** , 556–571. MR-1392329

[40] Marton, K. (2003) Measure concentration and strong mixing. *Studia Scientiarum Mathematicarum Hungarica* **32 (3B)**, **40**, 95–113. MR-2002993

[41] Marton, K. (2004) Measure concentration for Euclidean distance in the case of dependent random variables. *Ann. Probab.* **32 (3B)**, 2526–2544. MR-2078549

[42] Massart, P. (2007), *Concentration Inequalities and Model Selection*. Springer, Berlin. MR-2319879

[43] McAllester, D. A. (1999) PAC-Bayesian model averaging. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 164–170, ACM, New York. MR-1811612

[44] McDiarmid C. (1989) On the method of bounded differences, in: Surveys of Combinatorics, *Siemons J. (Ed.)*, **Lect. Notes Series 141**, London Math. Soc. MR-1036755

[45] Rio, E. (1998) Processus empiriques absolument réguliers et entropie universelle. *Prob. Th. Rel. Fields,* **111**, 585–608. MR-1641838

[46] Rio, E. (2000) Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.* **330**, 905–908. MR-1771956

[47] Rüschendorf, L. (1985) The Wasserstein distance and approximation theorems, *Z.Wahrsch. Verw. Gebiete* **70**, 117–129. MR-0795791

[48] Samson, P.-M. (2000) Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.* **28**, 416–461. MR-1756011

[49] Talagrand, M. (1995) Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.*, **81**, 73–205. MR-1361756

[50] 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. MR-0816291

[51] Villani, C. (2009) *Optimal Transport, Old and New.* Springer-Verlag, Berlin, 2009. MR-2459454

[52] Wintenberger, O. (2010) Deviation inequalities for sums of weakly dependent time series. *Electron. Commun. Probab.* **15**, 489–503. MR-2733373

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)

Economical model of EJP-ECP

- Low cost, based on free software (OJS[1])
- Non profit, sponsored by IMS[2], BS[3], PKP[4]
- Purely electronic and secure (LOCKSS[5])

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]OJS: Open Journal Systems http://pkp.sfu.ca/ojs/
[2]IMS: Institute of Mathematical Statistics http://www.imstat.org/
[3]BS: Bernoulli Society http://www.bernoulli-society.org/
[4]PK: Public Knowledge Project http://pkp.sfu.ca/
[5]LOCKSS: Lots of Copies Keep Stuff Safe http://www.lockss.org/
[6]IMS Open Access Fund: http://www.imstat.org/publications/open.htm