From Generalisation Error to Transportation-cost Inequalities and Back

Amedeo Roberto Esposito, Michael Gastpar
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland
{amedeo.esposito, michael.gastpar}@epfl.ch

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834354

***Abstract*—In this work, we connect the problem of bounding the expected generalisation error with transportation-cost inequalities. Exposing the underlying pattern behind both approaches we are able to generalise them and go beyond Kullback-Leibler Divergences/Mutual Information and sub-Gaussian measures. In particular, we are able to provide a result showing the equivalence between two families of inequalities: one involving functionals and one involving measures. This result generalises the one proposed by Bobkov and Götze that connects transportation-cost inequalities with concentration of measure. Moreover, it allows us to recover all standard generalisation error bounds involving mutual information and to introduce new, more general bounds, that involve arbitrary divergence measures.**

***Index Terms*—Wasserstein Distance, Kullback-Leibler Divergence, Information Measures, Duality, Young's Inequality, Transportation-Cost Inequalities, Generalisation Error**

I. Introduction

A recent and interesting line of research has explored the problem of bounding the generalization error of learning algorithms via information measures [1]–[9]. The starting observation is that one can interpret a learning algorithm as a (potentially) randomised mapping that takes as input a data-set and provides as output a hypothesis (e.g., a classifier). The purpose is to retrieve a classifier that has good performance both on the training set and on an *independent* test set. Intuitively, if the outcome of the learning algorithm **depends** too much on its input then its performance on new data-points will be poor. In this case, the algorithm is said to *overfit* to the training set. One way of measuring said dependence is through information measures. To this day, virtually every information measure has been connected to the problem: Mutual Information [1], [2], [8], [9], Total Variation [3], Rényi's α-Divergences, Sibson's $\alpha-$Mutual Information, $f-$divergences and $f-$Mutual Information [7], etc. The connection has been drawn with respect to expected generalisation error [1], [3], [4], [10] and with respect to the probability of having a large generalisation error [7].

There exists a similarity between the bound connecting expected generalisation error to mutual information and transportation-cost inequalities. Transportation-cost inequalities link Wasserstein distances and Kullback-Leibler divergences to the concentration of measure phenomenon. In this work we will connect the dots and expose the pattern connecting these objects. Moreover, unearthing the underlying mechanism we will generalise both type of results (generalisation-error bounds and transportation-cost inequalities) and go beyond Kullback-Leibler Divergences and sub-Gaussian tails.

II. Background and definitions

Following the De Finetti's notation, instead of denoting the expectation of a function f with respect to a probability measure μ with $\mathbb{E}_\mu[f]$ we will denote it with $\mu(f)$. Spaces will be denoted with calligraphic letters $\mathcal{X},\mathcal{Y}$, random variables with capital letters X,Y and measures with Greek lower-case letters μ,ν. Given a function $f:\mathcal{X}\to\mathcal{Y}$ we will denote with $f^\star$ its Legendre-Fenchel dual, defined as follows.

$$f^\star(x^\star)=\sup_{x\in\mathcal{X}}\langle x,x^\star\rangle-f(x), \tag{1}$$

where $\langle x,x^\star\rangle$ denotes the natural pairing between a space $\mathcal{X}$ and its (topological) dual $\mathcal{X}^\star$, the space of (continuous) linear functionals defined over $\mathcal{X}$ *i.e.*, $\langle x,x^\star\rangle=x^\star(x)$. A different but related notion of duality is Young's duality. Given a function $f:\mathbb{R}\to\mathbb{R}$ we denote with $f_Y^\star$ its Young's complementary function, defined as follows:

$$f_Y^\star(x^\star)=\sup_{x>0}\langle x,x^\star\rangle-f(x),\ \ x^\star>0. \tag{2}$$

Notice that the supremum in (2) is over a different set with respect to (1). We will always specify which notion of duality we are using throughout the paper. Given two measures μ,ν such that ν is absolutely continuous with respect to μ (denoted with $\nu\ll\mu$) and a convex functional φ we denote with $D_\varphi(\nu\|\mu)=\mu\left(\varphi\left(\frac{d\nu}{d\mu}\right)\right)$ the φ-divergence, where $\frac{d\nu}{d\mu}$ is the Radon-Nikodym derivative. With $\varphi(x)=x\log x$ one recovers the KL-divergence $D(\nu\|\mu)$, with $\varphi(x)=|x-1|$ one recovers the Total Variation distance $TV(\nu,\mu)$ and with $\varphi(x)=\frac{|x|^\alpha}{\alpha}$, $\alpha>0$ one recovers the Hellinger integral $H_\alpha(\nu\|\mu)$. Given a locally compact Hausdorff space $\mathcal{X}$, $C_c(\mathcal{X})$ denotes the space of continuous functions $f:\mathcal{X}\to\mathbb{R}$ defined on a compact support.

A. Optimal Transport Theory

Denote with $\mathcal{M}(\mathcal{X})$ the set of Radon measures over $\mathcal{X}$ and with $\mathcal{P}(\mathcal{X})$ the set of all probability measures over $\mathcal{X}$. Let $\mu,\nu\in\mathcal{P}(\mathcal{X})$, consider the set $\Pi(\mu,\nu)$ of all the joint probability measures $\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{X})$ with marginals equal to μ and ν, *i.e.*, such that $\pi(\cdot\times\mathcal{X})=\mu(\cdot)$ and $\pi(\mathcal{X}\times\cdot)=\nu(\cdot)$. The problem advanced by Kantorovich was the following: given μ and ν and a Borel function $d:\mathcal{X}\times\mathcal{X}\to[0,+\infty]$, can

we find a joint measure π that minimises $\pi(d)$? π represents a **transport plan** between μ, ν. Under mild assumptions on d, optimal transport plans are guaranteed to exists for general spaces $\mathcal{X}$ [11, Theorem 4.1]. If d itself is a metric over $\mathcal{X}$, then $\inf_\pi \pi(d)$ represents a distance over the space of probability measures. More precisely, given a metric d, let us denote with $\mathcal{P}_p(\mathcal{X})$ the set of probability measures μ on $\mathcal{X}$ such that $\mu(d(X, x_0)^p)^{1/p} < +\infty$ for some $x_0 \in \mathcal{X}$.

Definition 1 ([11, Def. 6.1])**.** Let $(\mathcal{X}, d)$ be a Polish space and $p \in [1, +\infty)$. Let $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, the $p-$Wasserstein distance between μ and ν is defined as $W_p(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)}(\pi(d^p))^{1/p}$.

Wasserstein distances satisfy interesting properties [12, Lemma 3.4.1]. Moreover, when connected to KL-divergences, through what are known in the literature as "Transportation-cost Inequalities", they have interesting implications in the concentration of measure phenomenon.

Definition 2 (Transportation-Cost Inequality)**.** Let $(\mathcal{X}, d)$ be a Polish space and μ a probability measure on $\mathcal{X}$, we say that μ satisfies an L^p-transportation-cost inequality with constant c (or $T_p(c)$ in short) if for every $\nu \ll \mu$

$$W_p(\mu, \nu) \leq \sqrt{2cD(\nu\|\mu)}. \tag{3}$$

When $p = 1$, for instance, these inequalities are equivalent to concentration in the following sense:

Theorem $\diamond$ ([13, Thm 3.1])**.** *Let $\mu \in \mathcal{P}_1(\mathcal{X})$ be a Borel probability measure. There exists a c such that for every $\lambda \in \mathbb{R}$*

$$\log \mu(\exp(\lambda f)) \leq \left(\frac{c\lambda^2}{2}\right) \tag{4}$$

for every $1-$Lipschitz μ-integrable function $f : \mathcal{X} \to \mathbb{R}$ such that $\mu(f) = 0$, if and only if μ satisfies a $T_1(c)$ inequality, i.e., for every $\nu \ll \mu$

$$W_1(\mu, \nu) \leq \sqrt{2cD(\nu\|\mu)}. \tag{5}$$

Example 1. Let $\mathcal{X}$ be a discrete space and $d(x, y) = \mathbb{1}_{x \neq y}$. We have that $W_1(\mu, \nu) = TV(\mu, \nu)$, *i.e.*, the Total Variation distance between μ, ν [12, Prop. 3.4.1]. In this case Equation (5) is known under the name of Pinsker's inequality and holds for every $\mu \in \mathcal{P}_1(\mathcal{X})$ with $c = 1/4$.

B. Learning Theory

For reasons of space we skip a thorough description of the learning framework. The setting we consider is the one described in [1]. In particular, in this work, we focus on the expected generalisation error which can be defined as follows:

Definition 3. Let $\mathcal{P}_Z$ be a probability measure over the instance space $\mathcal{Z}$. Let $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be a loss function and $\mathcal{H}$ denotes the space of classifiers. The error (or risk) of a prediction rule $h \in \mathcal{H}$ with respect to $\mathcal{P}_\mathcal{Z}$ is defined as $L_{\mathcal{P}_\mathcal{Z}}(h) = \mathcal{P}_Z(\ell(h, Z))$ (the expected value under $\mathcal{P}_Z$ of $\ell(h, Z)$ for a given h). Given a sample $S = (z_1, \ldots, z_n)$, the empirical error of h with respect to $S = Z^n$ is defined as $L_S(h) = \frac{1}{n}\sum_{i=1}^n \ell(h, z_i)$. Moreover, given a learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$, denoting with $\mathcal{P}_{HS}$ the joint measure induced by $\mathcal{A}$ on $\mathcal{H} \times \mathcal{Z}^n$, its generalization error with respect to $S = Z^n$ is defined as follows:

$$\text{gen-err}_\mathcal{P}(\mathcal{A}, S) = \mathcal{P}_{HS}(L_S(\mathcal{A}(S)) - L_\mathcal{P}(\mathcal{A}(S))). \tag{6}$$

Remember that using the De Finetti notation (and given that $\mathcal{P}_Z$, $\mathcal{P}_{SH}$ are probability measures) then, given a function f, $\mathcal{P}_Z(f)$ and $\mathcal{P}_{HS}(f)$ denote the expectation of f, respectively under $\mathcal{P}_Z$ and $\mathcal{P}_{HS}$.

III. Main Result

Let us start with the main result and then elaborate on it to show how it is connected to both generalisation error and transportation-cost inequalities.

Theorem 1. *Let $f \in C_c(\mathcal{X})$ and $\phi : \mathbb{R} \to \mathbb{R}$ be a strictly convex function such that its Legendre-Fenchel dual $\phi^\star$ admits a generalised inverse $\phi^{\star -1}(t) = \inf\{s : \phi(s) \geq t\}$. Let $\psi^\star : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ be the Legendre-Fenchel dual of $\psi : C_c(\mathcal{X}) \to \mathbb{R}$. If*

$$\psi(\lambda f) \leq \phi(\lambda) \text{ for every } \lambda > 0 \tag{7}$$

then, for every measure ν such that $\nu(f) < +\infty$ and $\phi'^{-1}\left(\phi^{\star -1}(\psi^\star(\nu))\right) > 0$,

$$\nu(f) \leq \phi^{\star -1}\left(\psi^\star(\nu)\right). \tag{8}$$

Proof. By definition of Legendre-Fenchel transform we know that for a given function f and any given measure ν:

$$\psi^\star(\nu) = \sup_{f \in C_c(\mathcal{X})} \langle f, \nu\rangle - \psi(f) \tag{9}$$

$$\geq \lambda\langle f, \nu\rangle - \psi(\lambda f) = \lambda\nu(f) - \psi(\lambda f). \tag{10}$$

Hence, we can say that, given $f \in C_c(\mathcal{X})$, $\nu \in \mathcal{M}(\mathcal{X})$ and $\lambda > 0$

$$\nu(f) \leq \frac{\psi(\lambda f) + \psi^\star(\nu)}{\lambda} \leq \frac{\phi(\lambda) + \psi^\star(\nu)}{\lambda}, \tag{11}$$

where in (11) we used Equation (7). Denoting with $c = \psi^\star(\nu)$ and choosing $\lambda = \phi'^{-1}\left(\phi^{\star -1}(c)\right)$ gives us that

$$\nu(f) \leq \phi^{\star -1}(c) = \phi^{\star -1}(\psi^\star(\nu)). \tag{12}$$

Indeed, let us denote, for simplicity, $\phi^{\star -1}(c) = t$, then replacing λ with $\phi'^{-1}\left(\phi^{\star -1}(c)\right)$ in (11) one has

$$\frac{c + \phi(\phi'^{-1}(t))}{\phi'^{-1}(t)} = \frac{c + t\phi'^{-1}(t) - \phi^\star(t)}{\phi'^{-1}(t)} \tag{13}$$

$$= t + \frac{c - \phi^\star(t)}{\phi'^{-1}(t)} \tag{14}$$

$$= \phi^{\star -1}(c) + \frac{c - \phi^\star(\phi^{\star -1}(c))}{\phi'^{-1}(t)} \tag{15}$$

$$\leq \phi^{\star -1}(\psi^\star(\nu)), \tag{16}$$

where (13) follows from [14, Eq. (1.11)]. □

Remark 1. In most settings of interest the assumption that $\phi'^{-1}\left(\phi^{\star -1}(\psi^\star(\nu))\right) > 0$ is easily satisfied. If one considers

$\psi^\star(\nu)$ to be a φ-Divergence, *i.e.,* $\psi^\star(\nu) = D_\varphi(\nu\|\mu)$, with μ a probability measure fixed before-hand, then $\psi^\star(\nu) \geq 0$ for every $\nu \in \mathcal{P}(\mathcal{X})$. Typical functions ϕ will be of the form $\phi(x) = x^\alpha/\alpha$ with $\alpha > 1$. This implies that $\phi'^{-1}\left(\phi^{\star -1}(\psi^\star(\nu))\right) = (\beta\psi^\star(\nu))^{\frac{1}{\alpha}}$ with $\beta = \alpha/(\alpha-1) > 0$, which is clearly positive for $\psi^\star(\nu) > 0$.

The idea behind Theorem 1 is the following: suppose one wishes to bound the expected value of a function f (*e.g.*, a random variable X) with respect to ν (*i.e.*, in our current notation, $\nu(f)$ or $\nu(X)$) using a function of a divergence between ν and a measure μ fixed before-hand (*e.g.*, $D(\nu\|\mu)$). In order to provide such a result, leveraging Theorem 1 it is necessary to assume (or prove) a bound on the dual of the divergence (*e.g.*, $\log(\mu(\exp(\cdot)))$ in the case of KL, cf. Equation 7). The shape of the bound on the expected value will then depend on the bound one can provide on the dual of the divergence (cf. Equation 8). For instance, in Theorem $\diamond$, assuming a bound like the one in Equation 4 leads to a bound on the expected value $\nu(f)$ that behaves like $\sqrt{kD(\nu\|\mu)}$. This setting will be exemplified further in the following Corollary, where a a well-known result that involves the KL-Divergence (and that has already appeared in [1], [10]) will be recovered:

Corollary 1. *Let X be a random variable over the probability space $(\mathcal{X}, \mathcal{F}, \mu)$ and assume X to be a zero-mean σ^2-sub-Gaussian random variable* [1] *with respect to μ. We have that for every measure ν such that $\nu \ll \mu$ and such that $\nu(X) < +\infty$*

$$\nu(X) \leq \sqrt{2\sigma^2 D(\nu\|\mu)}. \tag{17}$$

Proof. The proof follows from Theorem 1, selecting $\psi(f) = \log(\mu(\exp(f)))$ and thus $\psi^\star(\nu) = D(\nu\|\mu)$ with $\nu \in \mathcal{P}(\mathcal{X})$. The assumption that X is sub-Gaussian under μ implies that $\psi(\lambda X) \leq \frac{\lambda^2\sigma^2}{2} = \phi(\lambda)$ for every $\lambda \in \mathbb{R}$, where ϕ is a strongly convex function. To complete the argument, we observe that $\phi^\star(\lambda^\star) = \frac{\lambda^{\star 2}}{2\sigma^2} \implies \phi^{\star -1}(t) = \sqrt{2\sigma^2 t}$, which establishes the claimed bound. □

IV. Generalisaton Error Bounds

Various results can be provided using Theorem 1 and the well-known Corollary 1. However, to offer something even more concrete to the reader, let us set up a very specific framework and restrict our assumptions further. Given a convex functional φ, we will consider the following family of functionals over measures $\psi^\star_\mu(\cdot) = D_\varphi(\cdot\|\mu)$, with μ a measure fixed beforehand. Consequently, ψ_μ denotes the Legendre-Fenchel dual of $D_\varphi(\cdot\|\mu)$. Similarly to before, we will assume that given $f \in C_c(\mathcal{X})$, $\psi_\mu(\lambda f) \leq \phi(\lambda)$ for some convex function ϕ and for every $\lambda > 0$.

Let us set ourselves in a classical learning setting. Let $S = (Z_1, \ldots, Z_n)$ and $H = \mathcal{A}(S)$ be two random variables respectively over the spaces $(\mathcal{Z}^n, \mathcal{F}_{\mathcal{Z}^n}, \mathcal{P}_S = \mathcal{P}_Z^{\otimes n})$, $(\mathcal{H}, \mathcal{F}_\mathcal{H}, \mathcal{P}_H)$. Let $\mathcal{P}_S\mathcal{P}_H$ denote the joint measure induced by the product of the marginals of S and H over $(\mathcal{Z}^n \times \mathcal{H}, \mathcal{F})$. Let S be the input of a learning algorithm $\mathcal{A}$ and $H = \mathcal{A}(S)$ the corresponding output. Let $\mathcal{P}_{SH}$ be the joint measure induced by $\mathcal{A}$ and assume that $\mathcal{P}_{SH} \ll \mathcal{P}_S\mathcal{P}_H$.

In order to match the framework that is typically utilised in the Learning Theory literature, we will work with $\psi^\star_{\mathcal{P}_S\mathcal{P}_H}(\cdot) = D_\varphi(\cdot\|\mathcal{P}_S\mathcal{P}_H)$ and, following the structure of Theorem 1, one has to assume something about $\psi_{\mathcal{P}_S\mathcal{P}_H}$ in order to provide a bound that involves $\psi^\star_{\mathcal{P}_S\mathcal{P}_H}$. However, our assumptions (again, matching the literature) will involve $\psi_{\mathcal{P}_S}$ or $\psi_{\mathcal{P}_Z}$. The reason why we can do this is the following. Given the choice of $\psi^\star_\mu = D_\varphi(\cdot\|\mu)$, we typically know the shape that ψ_μ will have. If $\psi^\star_\mu$ is the KL divergence then $\psi_\mu(f) = \log\mu(\exp(f))$ while if $\psi^\star_\mu$ is a $\varphi-$Divergence then $\psi_\mu(f) = \mu(\varphi^\star(f))$ [14]. This naturally implies that if we consider product measures, one has the following characterisation for the dual $\psi_{\mu\times\xi}(f) = \xi(\mu(\varphi^\star(f))) = \xi(\psi_\mu(f))$ for φ-Divergences and $\exp(\psi_{\mu\times\xi}(f)) = \xi(\mu(\exp(f))) = \xi(\exp(\psi_\mu(f)))$ for KL.

Consequently, since we will consider the product measure $\mathcal{P}_S\mathcal{P}_H$ we have that, given the structure of the dual, an upper-bound on $\psi_{\mathcal{P}_S}$ for every $h \in \mathcal{H}$ naturally implies an upper-bound on $\psi_{\mathcal{P}_S\mathcal{P}_H}$. Another important consideration is that, an assumption of the form $\psi_{\mathcal{P}_Z}(\lambda(\ell(h,Z) - \mathcal{P}_Z(\ell(h,Z)))) \leq \phi(\lambda)$ for every h typically implies an assumption of the form $\psi_{\mathcal{P}_S}(\frac{\lambda}{n}(\sum_i(\ell(h,Z_i) - \mathcal{P}_S(\ell(h,Z_i))))) \leq \phi(\lambda)/n$, with S a sequence of n iid samples distributed according to $\mathcal{P}_Z^{\otimes n}$, something that we will informally define as the "n-sum property" of $\psi_{\mathcal{P}_Z}, \psi_{\mathcal{P}_S}$ and ϕ. Under this framework, we can state the following result.

Corollary 2. *Let $\psi^\star_{\mathcal{P}_S\mathcal{P}_H}(\cdot) = D_\varphi(\cdot\|\mathcal{P}_S\mathcal{P}_H)$ for some convex functional φ. Assume that $\psi_{\mathcal{P}_S\mathcal{P}_H}(\lambda(L_S(H) - \mathcal{P}_H(L_\mathcal{P}(H)))) \leq \phi(\lambda)/n$ for every $h \in \mathcal{H}$ and $n, \lambda > 0$. One has that*

$$\textit{gen-err}_\mathcal{P}(\mathcal{A}, S) \leq \frac{\phi^{\star -1}(nD_\varphi(\mathcal{P}_{SH}\|\mathcal{P}_S\mathcal{P}_H))}{n}. \tag{18}$$

Proof. Let us denote with $\phi_n(\lambda) = \frac{\phi(\lambda)}{n}$ one has that $\phi^\star_n(\lambda^\star) = \frac{1}{n}\phi^\star(\lambda n)$ and consequently $\phi^{\star -1}_n(t) = \frac{1}{n}\phi^{\star -1}(nt)$. The statement then follows from Theorem 1. □

From this we can now easily recover all the bounds on the expected generalisation-error that involve KL and that are present in the literature, for instance [1, Thm. 1]:

Corollary 3. *Let $\varphi(x) = x\log x$ in Corollary 2, hence $D_\varphi(\cdot\|\mathcal{P}_S\mathcal{P}_H)$ is the KL-Divergence. Assume that the loss function $\ell(h, Z_i)$ is σ^2-sub-Gaussian under $\mathcal{P}_Z$ for every h. We have that*

$$\textit{gen-err}_\mathcal{P}(\mathcal{A}, S) \leq \sqrt{\frac{2\sigma^2 I(S;\mathcal{A}(S))}{n}}. \tag{19}$$

Proof. Since $\ell(h, Z_i)$ is σ^2-sub-Gaussian under $\mathcal{P}_Z$, then $\psi_{\mathcal{P}_S}\left(\frac{1}{n}\sum_i \lambda(\ell(h,Z_i) - \mathcal{P}_Z(\ell(h,Z)))\right) \leq \frac{\phi(\lambda)}{n}$. This implies that $\psi_{\mathcal{P}_Z}, \psi_{\mathcal{P}_S}$ and ϕ satisfy the "n-sum property". The argument then follows directly from Corollary 2 along with the fact that $\frac{1}{n}\phi^{\star -1}(nt) = \sqrt{\frac{2\sigma^2 t}{n}}$. □

[1] Given a zero-mean random variable X we say that it is σ^2-sub-Gaussian if the following holds true for every $\lambda \in \mathbb{R} : \log(\mu(e^{\lambda X})) \leq \frac{\lambda^2\sigma^2}{2}$.

To conclude the section and emphasise the generality of this approach, let us choose a different divergence and derive a result which, to the best of our knowledge, has not appeared before in the literature.

Corollary 4. *Let $\varphi(x) = \frac{x^2}{2}$ in Corollary 2, hence $\psi^\star_{\mathcal{P}_S\mathcal{P}_H}(\cdot) = D_\varphi(\cdot\|\mathcal{P}_S\mathcal{P}_H) = (\chi^2(\cdot\|\mathcal{P}_S\mathcal{P}_H)+1)/2$. Assume that given the loss function $\ell(h,Z_i)$ there exists a constant $K>0$ such that for every h $\mathcal{P}_Z((\lambda(\ell(h,Z_i)-\mathcal{P}_Z(\ell(h,Z)))^2) \le K\lambda^2$. One has that*

$$\textit{gen-err}_{\mathcal{P}}(\mathcal{A},S) \le \sqrt{\frac{2K(\chi^2(\mathcal{P}_{SH}\|\mathcal{P}_S\mathcal{P}_H)+1)}{n}}. \tag{20}$$

Proof. Given that $\psi^\star_{\mathcal{P}_S\mathcal{P}_H}(\nu) = \frac{1}{2}(\chi^2(\nu\|\mathcal{P}_S\mathcal{P}_H)+1)$ one has that $\psi_{\mathcal{P}_S\mathcal{P}_H}(f) = \frac{1}{2}\mathcal{P}_S\mathcal{P}_H(f^2)$. Since the Z_i are assumed to be iid random variables, the $\ell(h,Z_i)$ are also iid. By assumption, one has that $\mathcal{P}_S((\lambda\frac{1}{n}\sum_i(\ell(h,Z_i)-\mathcal{P}_Z(\ell(h,Z)))^2) \le \frac{\lambda^2}{n}K$. Thus, $\psi_{\mathcal{P}_Z},\psi_{\mathcal{P}_S}$ and ϕ satisfy the "n-sum property". The argument then follows from Corollary 2 and by noticing that $\phi^\star(\lambda^\star) = \frac{\lambda^{\star 2}}{2K}$ which in turn implies that $\phi^{\star -1}(t) = \sqrt{2Kt}$. □

Remark 2. Assuming the σ^2-sub-Gaussianity of $\ell(h,Z_i)-\mathcal{P}_Z(\ell(h,Z))$ naturally implies a bound on $\mathcal{P}_Z((\lambda(\ell(h,Z_i)-\mathcal{P}_Z(\ell(h,Z)))^\kappa)$ for every $\kappa \ge 1$. For instance, σ^2-sub-Gaussianity implies that in Corollary 4, $\mathcal{P}_Z((\lambda(\ell(h,Z_i)-\mathcal{P}_Z(\ell(h,Z)))^2) \le K\lambda^2$ holds with $K = 2\sigma^2e^{2/e}$. However, assuming that ℓ has bounded variance (for every h or on expectation wrt $\mathcal{P}_H$) is much less restrictive than assuming it is sub-Gaussian. As an example of this, suppose that the loss ℓ is σ^2-sub-Gaussian. One has that ℓ^2 is sub-Exponential with parameters $(4\sqrt{2}\sigma^2, 4\sigma^2)$. This means that ℓ^2 has a finite log-moment generating function only for $\lambda < 1/(4\sigma^2)$. Considering the framework given by Theorem 1 and Corollary 3, in order to solve the infimum in (11) (which, in turn, provides Eq (19)), one needs to select $\lambda^\star = \sqrt{\frac{2I(S;H)}{\sigma^2}}$. Consequently, if $I(S;H) \ge \frac{1}{2^5\sigma^2}$ the bound is not valid as, for that choice of λ, the log-moment generating function of ℓ with respect to $\mathcal{P}_S\mathcal{P}_H$ is actually unbounded (and, thus, cannot be bounded by $\phi(\lambda)$). On the other hand, a sub-Exponential random variable is such that all of its moments are bounded. Hence, an approach as the one suggested by Corollary 4 can be successful.

A. Recovering other known results

Other known results can be recovered, for instance noticing that gen-err$_{\mathcal{P}}(\mathcal{A},S)$ can be re-written as $\frac{1}{n}\sum_{i=1}^n(\mathcal{P}_{HZ}(\ell(H,Z_i)) - \mathcal{P}_H\mathcal{P}_Z(\ell(H,Z)))$ and using Corollary 1 on each term inside the summation recovers exactly the main result in [8, Thm. 2].

Alternatively, the observation that one can consider a super-sample $\tilde{S}$ of length $2n$ and take two random subsets of length n S,S' independently, allows us to rewrite the generalisation error as follows gen-err$_{\mathcal{P}}(\mathcal{A},\tilde{S}) = \mathcal{P}_{\tilde{S}}(\mathcal{P}_{SH|\tilde{S}}(L_{\tilde{S}}(H) - L_{\mathcal{P}}(H)))$. One can then use Theorem 1 for every given choice of $\tilde{S}$ on $\mathcal{P}_{SH|\tilde{S}=\tilde{s}}(L_{\tilde{S}}(H) - L_{\mathcal{P}}(H))$ and exactly recover [9, Thm 5.1, Corollary 5.2, 5.3]. It represents a bound involving $D(\mathcal{P}_{SH|\tilde{S}=\tilde{s}}\|\mathcal{P}_{S|\tilde{S}=\tilde{s}}\mathcal{P}_{H|\tilde{S}=\tilde{s}})$ after assuming/showing that $\psi_{\mathcal{P}_{S|\tilde{S}=\tilde{s}}\mathcal{P}_{H|\tilde{S}=\tilde{s}}}(\lambda(L_{\tilde{s}}(H) - L_{\mathcal{P}}(H))) \le \frac{\lambda^2}{2n}c(\tilde{s})$ with $c(\tilde{s})$ a constant that depends on the realisation of $\tilde{S}$. Clearly the possible combinations are numerous, however the pattern remains the same: a bound on the dual of the targeted divergence (or functional of measure) implies a bound on a difference of expectations. Indeed, still following the approach undertaken in [8] but with the general spirit that characterises this work, one can show the following: gen-err$_{\mathcal{P}}(\mathcal{A},S) \le \frac{1}{n}\sum_{i=1}^n\sqrt{2K(\chi^2(\mathcal{P}_{Z_iH}\|\mathcal{P}_{Z_i}\mathcal{P}_H)+1)}$. Otherwise, following [9] one can recover bounds involving $\mathcal{P}_{\tilde{S}}\left(\chi^2\left(\mathcal{P}_{SH|\tilde{S}}\|\mathcal{P}_{S|\tilde{S}}\mathcal{P}_{H|\tilde{S}}\right)\right)$.

V. Transportation-Cost inequalities

A recurring theme in the field that lies at the intersection between Information Theory and Optimal Transport is showing that Transportation-Cost Inequalities are equivalent to some form of concentration of measure. One such example is Theorem ⋄ in Section II-A. An in-depth review of this connection can be found in [12, Section 3.4]. This type of results are generally made of two parts: the "if part" (if μ satisfies concentration for every function in some family then the $T_p(c)$ holds) and the "only if" part (if μ satisfies a $T_p(c)$ inequality then one has concentration for every function in some family). So far we have essentially established the "if part" for general functionals of measures. Let us now establish the "only if" part and then discuss how it relates to the classical framework.

Theorem 2. *Let $\psi : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ be a functional and let $f \in \mathcal{M}(\mathcal{X})^\star$. If there exist a Young function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ and its complementary function $\phi^\star$which admits a generalised inverse $\phi^{\star -1}$ such that for every ν with $\nu(f) < +\infty$ one has*

$$\nu(f) \le \phi_Y^{\star\,-1}(\psi(\nu)), \tag{21}$$

then

$$\psi^\star(\lambda f) \le \phi(\lambda) \text{ for every } \lambda > 0, \tag{22}$$

where $\psi^\star : \mathcal{M}(\mathcal{X})^\star \to \mathbb{R}$ is the Legendre-Fenchel dual of ψ.

Proof. Given that ϕ is a Young function, one has that by [15, Lemma 2.4]

$$\nu(f) \le \phi_Y^{\star\,-1}(\psi(\nu)) = \inf_{\lambda>0}\frac{\phi(\lambda)+\psi(\nu)}{\lambda}. \tag{23}$$

This means that for every $\lambda > 0$ and every ν-integrable function f,

$$\phi(\lambda) \ge \lambda\nu(f) - \psi(\nu). \tag{24}$$

Taking the supremum with respect to $\nu \in \mathcal{M}(\mathcal{X})$ one recovers the statement: $\phi(\lambda) \ge \psi^\star(\lambda f)$. □

In order to see how Theorem 1 and Theorem 2 represent, respectively, the "if" and "only if" part connecting $T_p(c)-$like inequalities to concentration let us recover Theorem ⋄.
In particular, select:

- $\psi^\star(\nu) = D(\nu\|\mu)$ for a given μ (consequently, one has that $\psi(\lambda f) = \log \mu(\exp(\lambda f))$, cf. [16, Lemma 6.2.13]);
- $\phi(\lambda) = \frac{c\lambda^2}{2}$ (which implies ${\phi_Y^\star}^{-1}(\kappa) = \sqrt{2c\kappa}$);

Theorem 1 is what allows us to reach (5) starting from (4). Keeping the same ϕ but inverting the roles of ψ and $\psi^\star$ in Theorem 2 is what allows us to reach (4) starting from (5). Some extra technical steps are necessary in order to bring in Wasserstein Distances (which will be considered in the proof just below). Given the generality of the results we can, as an example, consider a setting similar to Theorem $\diamond$ but involving a different divergence:

Theorem 3. *Let $\mu \in \mathcal{P}_1(\mathcal{X})$ and $\beta > 1$. There exists a c such that for every λ and every 1-Lipschitz function f*

$$\mu\left(|\lambda f|^\beta\right) \leq (c\lambda)^\beta, \tag{25}$$

if and only if, for every $\nu \ll \mu$

$$W_1(\mu,\nu) \leq \left(\alpha c^\alpha H_\alpha(\nu\|\mu)\right)^{\frac{1}{\alpha}}, \tag{26}$$

where $\alpha = \frac{\beta}{\beta-1}$. Setting $\beta = 2$ we recover the following:

$$W_1(\mu,\nu) \leq \sqrt{c^2 2(\chi^2(\nu\|\mu)+1)}. \tag{27}$$

Proof. Let μ be a probability measure and let $\psi^\star(\nu) = H_\alpha(\nu\|\mu)$ with $\alpha > 1$, one has that $\psi(f) = \frac{\mu(|f|^\beta)}{\beta}$ (cf. [14, Theorem 3.3]). Moreover, let $\phi(\lambda) = \frac{|c\lambda|^\beta}{\beta}$, one has that $\phi^\star(\lambda^\star) = \frac{|\lambda^\star/c|^\alpha}{\alpha}$ with $\alpha = \frac{\beta}{\beta-1}$. Consequently, for positive κ, ${\phi^\star}^{-1}(\kappa) = (\alpha c^\alpha \kappa)^{\frac{1}{\alpha}}$. Let f be a function such that $\|f\|_{Lip} \leq 1$ and $\mu(f) = 0$. Assume that for every $\nu \ll \mu$, $W_1(\mu,\nu) \leq (\alpha c^\alpha H_\alpha(\nu\|\mu))^{\frac{1}{\alpha}}$. By the Kantorovich-Rubenstein dual representation of W_1 (cf. $W_1(\mu,\nu) = \sup_{f:\|f\|_{Lip}\leq 1} |\mu(f) - \nu(f)|$ [11, Thm. 5.10]) one has that for every function f such that $\|f\|_{Lip} \leq 1$ and $\mu(f) = 0$ one can rewrite Equation 26 as follows:

$$\nu(f) \leq (\alpha c^\alpha H_\alpha(\nu\|\mu))^{\frac{1}{\alpha}} = {\phi^\star}^{-1}(H_\alpha(\nu\|\mu)). \tag{28}$$

Consequently, by Theorem 2, one has that for every such f

$$\psi^\star(\lambda f) = \mu\left(\frac{|\lambda f|^\beta}{\beta}\right) \leq \frac{(c\lambda)^\beta}{\beta} = \phi(\lambda). \tag{29}$$

Similarly, assuming (29) leads to (28) for every such f via Theorem 1. Repeating the same argument with -f one reaches the following statement:

$$|\mu(f) - \nu(f)| \leq (\alpha c^\alpha H_\alpha(\nu\|\mu))^{\frac{1}{\alpha}}. \tag{30}$$

The assumption that $\mu(f) = 0$, can be dropped replacing f with $f - \mu(f)$. Taking then the supremum over all the $1-$Lipschitz functions f in (30) one reaches Equation (26). □

Drawing inspiration from Theorem 3 one can consider almost any $\varphi-$Divergence. Some restrictions on the possible choices of φ arise naturally in order to have access to both variational representations (cf. [14, Theorems 3.3, 3.4]).

Remark 3. In the classical $T_1(c)$-setting one assumes that $\log(\mu(\exp(\lambda f))) \leq \frac{\lambda^2 c^2}{2}$ for every f that is $1-$Lipschitz and consequently recovers (5). Consider Theorem 3 instead: in Equation (25) we are only asking for the β-th moment to be bounded. The same argument as in Remark 2 holds: even though it is well known that in general $\chi^2(\nu\|\mu) \geq D(\nu\|\mu)$ (leading thus, to a worse bound on W_1), in order to get (27) one only needs to bound the second moment of f with respect to μ and such a bound can exist for functions with unbounded log-moment generating function, which are excluded from a classical $T_1(c)$ setting (cf. Remark 2).

Theorem 3 highlights the following approach: starting from the variational representation of Wasserestein distances W_p one understands the restriction on the family of functions that needs to be considered (cf. [11, Thm. 5.10])). Let us denote such a family with $\mathcal{F}_p$. The next step is then to fix a measure μ and a functional $\psi_\mu^\star$ (in Theorem 3, the choice of $\psi_\mu^\star$ has fallen on the Hellinger integral). The upper-bound that one can provide on $\psi(\lambda f)$ for $f \in \mathcal{F}_p$, characterised by $\phi(\lambda)$ (cf. Equations (4), (7) and (25)), will determine the shape of the $T_p(c)$-like inequality (through ${\phi_Y^\star}^{-1}$, cf. Equations (5), (8) and (26)), here denoted $\phi_p^\psi(c)$-inequalities for convenience. Vice versa, bounding a Wasserstein distance W_p through a divergence $\psi_\mu^\star(\nu)$ via a $\phi_p^\psi(c)$-inequality (cf. Equations (5), (21) and (26)) implies a bound on the dual ψ_μ (cf. Equations (4), (22) and (25)) and that can imply concentration according to ψ_μ, ϕ and $\mathcal{F}_p$.

VI. Conclusions

In this work we linked transportation-cost inequalities with generalisation error bounds. The thread connecting the two approaches is Legendre-Fenchel/Young duality. As a result, we managed to generalise both approaches to various divergences and to random variables that are not necessarily sub-Gaussian. The approach undertaken in this work grants two extra degrees of freedom: $\psi^\star$ and ϕ. In particular, the functional $\psi^\star$ does not necessarily have to be the KL-Divergence and the function ϕ can be any convex function that allows us to upper-bound $\psi(\lambda f)$. Considering KL, this is tantamount to showing concentration properties of μ. Assuming a Gaussian-like behaviour of the log-moment generating function (hence, choosing $\phi(\lambda) = c\lambda^2/2$) leads to the familiar $\sqrt{2cD(\nu\|\mu)}$. On the other hand, trying to show a classical $T_p(c)$-like inequality for some measure μ while fixing the shape of the inequality to be approximately $W_p(\mu,\nu) \leq \sqrt{cD(\nu\|\mu)}$ (which means, essentially, fixing ϕ) can be impossible. This depends on the concentration properties of μ or the family of functions that we have to consider $\mathcal{F}_p$. One can thus relax the inequality by either assuming a behaviour of the cumulant generating function that is different from Gaussian-like (same $\psi^\star$ but different ϕ) or by picking a completely different divergence (changing $\psi^\star$ and/or ϕ).

Acknowledgment

The authors would like to thank Professor Ugo Vaccaro for insightful comments and suggestions on an early draft of this work. The work in this paper was supported in part by the Swiss National Science Foundation under Grant 200364.

REFERENCES

[1] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems*, 2017, p. 2521–2530.

[2] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, "Learners that use little information," ser. Proceedings of Machine Learning Research, vol. 83. PMLR, 07–09 Apr 2018, pp. 25–55.

[3] H. Wang, M. Diaz, J. C. S. S. Filho, and F. P. Calmon, "An information-theoretic view of generalization via Wasserstein distance," *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 577–581, 2019.

[4] A. T. Lopez and V. S. Jog, "Generalization error bounds using wasserstein distances," *2018 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2018.

[5] A. Pensia, V. Jog, and P.-L. Loh, "Generalization error bounds for noisy, iterative algorithms," *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 546–550, 2018.

[6] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, pp. 1–1, 2020.

[7] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rényi-, f-divergences and Maximal Leakage," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 4986–5004, 2021.

[8] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 587–591.

[9] T. Steinke and L. Zakynthinou, "Reasoning About Generalization via Conditional Mutual Information," in *Proceedings of Thirty Third Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, J. Abernethy and S. Agarwal, Eds., vol. 125. PMLR, 09–12 Jul 2020, pp. 3437–3452. [Online]. Available: https://proceedings.mlr.press/v125/steinke20a.html

[10] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 51. PMLR, 09–11 May 2016, pp. 1232–1240.

[11] C. Villani, *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

[12] M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Information Theory, Communications, and Coding: Second Edition*. Now Foundations and Trends, 2014.

[13] S. Bobkov and F. Götze, "Exponential integrability and transportation cost related to logarithmic sobolev inequalities," *Journal of Functional Analysis*, vol. 163, no. 1, pp. 1 – 28, 1999.

[14] M. Broniatowski and A. Keziou, "Minimization of divergences on sets of signed measures," *Studia Scientiarum Mathematicarum Hungarica*, vol. 43, 03 2010.

[15] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[16] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, ser. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2009. [Online]. Available: https://books.google.ch/books?id=d3nnjwEACAAJ