# An Information-Theoretic View of Generalization via Wasserstein Distance

Hao Wang*, Mario Diaz†, José Cândido S. Santos Filho*‡, and Flavio P. Calmon*
*Harvard University, {hao_wang, candido}@g.harvard.edu, flavio@seas.harvard.edu
†Centro de Investigación en Matemáticas A.C., diaztorres@cimat.mx
‡University of Campinas, candido@decom.fee.unicamp.br

***Abstract*—We capitalize on the Wasserstein distance to obtain two information-theoretic bounds on the generalization error of learning algorithms. First, we specialize the Wasserstein distance into total variation, by using the discrete metric. In this case we derive a generalization bound and, from a strong data-processing inequality, show how to narrow the bound by adding Gaussian noise to the output hypothesis. Second, we consider the Wasserstein distance under a generic metric. In this case we derive a generalization bound by exploiting the geometric nature of the Kantorovich-Rubinstein duality theorem. We illustrate the use of these bounds with examples. Our bounds can handle certain cases in which existing bounds via mutual information fail.**

## I. Introduction

From an information-theoretic viewpoint, a learning algorithm can be thought of as a channel [1], the input being the training dataset and the output being the learned hypothesis (e.g., regression coefficients in linear regression, layer weights in a neural network). In practice, such a channel is typically noisy: a single input dataset may lead to different hypotheses, where the probability of observing a given output hypothesis depends on the learning model to be trained, on the risk function to be minimized, and on the (randomized) minimization algorithm itself. Although the output hypothesis is determined by minimizing the empirical risk function w.r.t. a particular training dataset, it is ultimately expected to *generalize*, i.e., yield a low risk w.r.t. the true (and unknown) underlying data distribution. As a result, an essential performance measure of learning algorithms is given by the difference between the population risk and the empirical risk, known as the *generalization error*.

The generalization error is usually analyzed in terms of upper bounds. A traditional approach relies on characterizing the complexity of the hypothesis class, e.g., via VC-dimension or Rademacher complexity [2]. This leads to algorithm-independent bounds, in the sense that, except for the hypothesis class, any particulars about the inner structure of the learning process at hand are fully ignored. In view of that, the resulting bounds are usually loose, representing a sort of worst-case scenario over all learning algorithms that share the hypothesis class under consideration. For example, the k-Nearest Neighbors algorithm (k-NN) has an infinite VC-dimension, while it still generalizes well [3]. This drawback has motivated the development of algorithm-dependent bounds, such as PAC-Bayesian bounds [4] and algorithm stability bounds [3].

More recently, Russo and Zou [5] introduced an information-based framework to analyze overfitting in the context of adaptive data analysis. In that work, generalization bounds were cast in terms of mutual information by leveraging the intrinsic connection between sub-Gaussian concentration and KL-divergence bounds on transportation cost (cf. [5, Proposition 3.1] and the "transportation lemma" [6, Lemma 4.18]). This framework was further explored by Xu and Raginsky [1]. In [7], Alabdulmohsin used total variation to bound the generalization error when the loss function is bounded. In [8], Raginsky *et al.* proposed measures of algorithmic stability based on total variation, mutual information, and Wasserstein distance, respectively, and used them to upper bound the generalization error. Other information-theoretic metrics have been used to derive alternative upper bounds [9]–[11]. All these bounds follow a similar vein, with the generalization error being related to how much the output hypothesis depends on the input dataset.

In this work, we take a similar approach and derive new upper bounds on the generalization error of learning algorithms. Motivated by the drawback that mutual information — as well as the error bounds based on it — may become infinite, we derive generalization results via the Wasserstein distance. First, we specialize the Wasserstein distance to total variation by considering the discrete metric. In this case we derive an upper bound on the generalization error and, by applying a strong data-processing inequality [12], we show how to tighten the bound by adding Gaussian noise to the learning output. Second, at the expense of requiring the loss function to be Lipschitz continuous with respect to the hypothesis, we allow for the Wasserstein distance to be used along with a generic metric. In this case we derive an upper bound on the generalization error by virtue of the Kantorovich-Rubinstein duality theorem [13]. An interesting feature of this bound is that it can accommodate the Euclidean metric, thereby assessing the dependency between the learned hypothesis and the input dataset in a geometric manner. Another advantage of this bound is its capability to handle deterministic learning mappings, in which case existing bounds based on mutual information would fail. We also provide examples to illustrate the derived bounds.

The previous works more closely related to ours are those by Alabdulmohsin [7] and Raginsky *et al.* [8]. Compared to [7], we provide an upper bound based on total variation

that can accommodate unbounded loss functions, as well as we capitalize on a strong data-processing inequality to produce a more comprehensive explanation about why adding noise to the learning output may reduce overfitting. Moreover, we avoid assumptions on the stability of the learning algorithms made in [8]. After submitting the present work, we became aware of the results by Lopez and Jog [14] where generalization error bounds using Wasserstein distances are also provided. The approach in [14] differs from ours in a couple of aspects. First, we include an analysis of generalization error tailored to total variation (see Theorem 1), i.e., for the Wasserstein distance under the discrete metric. Second, our focus is on exploiting the geometric concentration of the returned hypotheses via a Wasserstein-distance bound (see Example 2), a feature which is not explored in [14].

## II. Problem Statement and Preliminaries

We follow the notation from [1]. Consider an instance space $\mathcal{Z}$, a hypothesis space $\mathcal{W}$, and a loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$. A learning algorithm can be represented as a conditional distribution $P_{W|S}$ that takes as input $n$ i.i.d. data samples $S \triangleq (Z_1, \cdots, Z_n)$, with $Z_i \sim \mu$, and outputs a hypothesis $W \in \mathcal{W}$. The population risk of a given hypothesis $w \in \mathcal{W}$ w.r.t. $\mu$ is defined as

$$L_\mu(w) \triangleq \mathbb{E}\left[\ell(w, Z)\right] = \int_{\mathcal{Z}} \ell(w, z)\mu(dz). \tag{1}$$

Since $\mu$ is unknown, $L_\mu(w)$ cannot be computed directly. We instead compute the empirical risk of $w$ w.r.t. the input dataset $S$ by averaging the loss function over all data samples:

$$L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i). \tag{2}$$

The difference $L_\mu(W) - L_S(W)$ is called generalization error and its expected value is denoted as

$$\mathsf{gen}(\mu, P_{W|S}) \triangleq \mathbb{E}\left[L_\mu(W) - L_S(W)\right], \tag{3}$$

where the expectation is taken over the joint distribution $P_{S,W} = \mu^{\otimes n} \otimes P_{W|S}$.

Our main goal is to derive upper bounds on (3) (see Section III) via Wasserstein distance. The Wasserstein distance between two probability distributions $\mu$ and $\nu$ is defined as

$$\mathbb{W}(\mu, \nu) \triangleq \inf_{\gamma \in \Gamma(\mu,\nu)} \int d(x, x')d\gamma(x, x'). \tag{4}$$

Here, $d(\cdot, \cdot)$ is a metric and $\Gamma(\mu, \nu)$ denotes the set of all couplings of $\mu$ and $\nu$ (i.e., all joint distributions with marginals $\mu$ and $\nu$). The Kantorovich-Rubinstein duality [13] states that

$$\mathbb{W}(\mu, \nu) = \sup_{f:\mathrm{Lip}(f)\le 1} \left\{\mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(V)\right]\right\}, \tag{5}$$

where $U \sim \mu$, $V \sim \nu$, and the supremum is taken over all 1-Lipschitz functions in the metric $d$, i.e., functions $f$ such that, for all $x, x'$,

$$|f(x) - f(x')| \le d(x, x'). \tag{6}$$

The total-variation distance between two probability distributions $\mu$ and $\nu$ is defined as

$$\mathsf{TV}(\mu, \nu) \triangleq \sup_E |\mu(E) - \nu(E)|, \tag{7}$$

where the supremum is taken over all measurable sets $E$. Note that the total-variation distance equals the Wasserstein distance associated to the discrete metric $d(x, x') = \mathbb{I}_{x\neq x'}$, where $\mathbb{I}$ denotes the indicator function. With this notation, the T-information between two random variables $U$ and $V$ is defined as $\mathsf{T}(U; V) \triangleq \mathsf{TV}(P_{U,V}, P_U \otimes P_V)$.

Finally, a random variable $U$ is called $\sigma$-subgaussian if, for all $\lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda(U-\mathbb{E}[U])}\right] \le e^{\lambda^2\sigma^2/2}$. If $U$ is $\sigma$-subgaussian, then [6] for all $t \ge 0$,

$$\max\left\{\Pr(U - \mathbb{E}\left[U\right] > t), \Pr(U - \mathbb{E}\left[U\right] < -t)\right\} \le e^{-\frac{t^2}{2\sigma^2}}.$$

The converse also holds true, up to a constant factor modifying the subgaussianity constant.

## III. Bounds on the Generalization Error

In this section we investigate the generalization error using Wasserstein distance. We start with total variation, the Wasserstein distance associated to the discrete metric, from which we provide a first bound on the generalization error. Then we derive an alternative bound using the Wasserstein distance with a generic metric, though requiring in this case the loss function to be Lipschitz continuous.

### A. Discrete Metric

We start with a technical lemma, which will be used to provide an upper bound on the expected generalization error.

**Lemma 1.** *Let $f : \mathcal{X} \to \mathbb{R}$ be a measurable and non-negative function. If $f(\bar{X})$ is $\sigma$-subgaussian with $\bar{X} \sim Q$, then for any probability measure $P$ and $X \sim P$,*

$$\mathbb{E}\left[f(\bar{X}) - f(X)\right] \le \frac{\delta}{1-\delta}\left[\mathbb{E}\left[f(X)\right] + \sigma\frac{1 + 4\log(1/\delta)}{\sqrt{2\log(1/\delta)}}\right], \tag{8}$$

*where $\delta \triangleq \mathsf{TV}(P, Q)$.*

*Proof.* See Appendix A. □

In Theorem 1 (on the next page), $\mathbb{E}\left[f(X)\right]$ will be instantiated as the expected empirical risk in order to upper bound the generalization error. Alas, the upper bound (8) contains $\mathbb{E}\left[f(X)\right]$ on the RHS of the inequality. Next, we provide an example that shows that this dependence is unavoidable and, in general, the subgaussian constant $\sigma$ and the total variation $\mathsf{TV}(P, Q)$ do not suffice to upper-bound the LHS of (8).

**Example 1.** Let $g : (0, +\infty) \times (0, 1) \to \mathbb{R} \cup \{+\infty\}$ be a function such that, under the assumptions in Lemma 1,

$$\mathbb{E}\left[f(\bar{X}) - f(X)\right] \le g(\sigma, \mathsf{TV}(P, Q)). \tag{9}$$

We will show that $g$ is necessarily trivial, i.e., $g \equiv +\infty$.

Let $t \in (0,1)$, $\sigma > 0$, and $c > 1+\sigma$. We consider the function $f(x) = x\mathbb{I}_{(0,+\infty)}(x)$ and the probability distributions $Q$ and $P$ with associated probability density functions

$$q(x) = \frac{1}{2\sigma}\mathbb{I}_{[c-\sigma,c+\sigma]}(x), \tag{10}$$

$$p(x) = \frac{1}{2\sigma}\mathbb{I}_{[c-\sigma t,c+\sigma t]}(x) + \mathbb{I}_{[0,1-t]}(x). \tag{11}$$

In this setting, it is straightforward to verify that $f(\bar{X})$ is $\sigma$-subgaussian with $\bar{X} \sim Q$, $\mathbb{E}[f(X)] = tc + \frac{1}{2}(1-t)^2$, and $\mathsf{TV}(P,Q) = 1-t$. A direct computation shows that

$$\begin{aligned}\mathbb{E}\left[f(\bar{X}) - f(X)\right] &= (1-t)c - \frac{1}{2}(1-t)^2 \\ &= \frac{\delta}{1-\delta}\left(\mathbb{E}[f(X)] - \frac{\delta}{2}\right).\end{aligned}$$

Thus, for $\sigma > 0$ and $t \in (0,1)$ fixed, $\mathbb{E}\left[f(\bar{X}) - f(X)\right] \to \infty$ as $c \to \infty$. Therefore, $g(\sigma, \mathsf{TV}(P,Q)) = +\infty$, as we wanted to show. Note that, in this example, $\mathrm{D}_{\mathrm{KL}}(P\|Q) = \infty$.

It is possible to provide an upper bound on the generalization error of a learning algorithm by applying Lemma 1 to $P = P_{S,W}$, $Q = P_S \otimes P_W$, and $f(s,w) = \frac{1}{n}\sum_{i=1}^{n}\ell(w,z_i)$. This is attained in the following theorem.

**Theorem 1.** *Suppose that the loss function $\ell(\cdot,\cdot)$ is non-negative and $\ell(w,Z)$ is $\sigma$-subgaussian under $Z \sim \mu$ for all $w \in \mathcal{W}$. Then*

$$\mathrm{gen}(\mu, P_{W|S}) \leq \frac{\delta}{1-\delta}\left(\mathbb{E}[L_S(W)] + \frac{\sigma}{\sqrt{n}}\frac{1+4\log(1/\delta)}{\sqrt{2\log(1/\delta)}}\right),$$

*where $\delta \triangleq \mathsf{T}(S;W)$. In particular,*

$$\mathbb{E}[L_\mu(W)] \leq \frac{1}{1-\delta}\left(\mathbb{E}[L_S(W)] + \frac{\sigma\delta}{\sqrt{n}}\frac{1+4\log(1/\delta)}{\sqrt{2\log(1/\delta)}}\right).$$

*Proof.* See Appendix B. □

In practice, the population risk quantifies how well the output hypothesis performs over new fresh data samples. Observe that the upper bounds in Theorem 1 becomes tighter when (i) the number of samples $n$ increases, (ii) the expected empirical risk $\mathbb{E}[L_S(W)]$ decreases, or (iii) the T-information $\mathsf{T}(S;W)$ decreases. Contributions (i) and (ii) can be accomplished by collecting more samples and training a model that better fits the input dataset, respectively. As for contribution (iii), an important remark is in order, as follows.

Strong data-processing inequalities (see, e.g., [12]) establish that the T-information decreases by adding Gaussian noise to the output hypothesis. More specifically, if $S \to W \to Y$, where $\|W\| \leq A$ and $Y = W + N$, with $N \sim \mathcal{N}(0, \mathbf{I}_d)$, then

$$\mathsf{T}(S;Y) \leq \eta_{\mathsf{TV}}(A)\mathsf{T}(S;W), \tag{12}$$

where $\eta_{\mathsf{TV}}(A) \triangleq 1 - \sqrt{\frac{2}{\pi}}\int_A^\infty e^{-t^2/2}\mathrm{d}t$. In particular, we have $\mathsf{T}(S;Y) < 1$. Combining (12) with the upper bound on the generalization error in Theorem 1, we can obtain a new upper bound on the generalization error that is guaranteed to be finite. This contrasts with the mutual-information approach, in which case the RHS of (12) can remain infinite.

### *B. Generic Metric*

In Section III-A, we provided an upper bound on the generalization error via total variation and connected it with a strong data-processing inequality. In this section we take an approach to generalization bounds using a generic metric for the Wasserstein distance.

To better motivate our transition from total-variation distance to a generic Wasserstein distance, let us discuss some drawbacks of the former. On the one hand, it is known that, on a compact space, convergence in distribution is equivalent to convergence in Wasserstein distance (see, e.g., [15]). On the other hand, convergence in distribution does not necessarily imply convergence in total-variation distance (see [16] for a simple example on this issue). Hence, it might be hard to combine total variation with classical probability results relying on convergence in distribution or related notions.

The above observation naturally motivates an alternative bound on the generalization error by exploiting the geometric nature of the Wasserstein distance on compact spaces. The following theorem, whose proof relies on the Kantorovich-Rubinstein duality, provides an alternative upper bound on the generalization error that takes into account the geometric concentration of the output hypothesis of a learning algorithm.

**Theorem 2.** *Suppose that the empirical risk $L_s(\cdot)$ is $L$-Lipschitz for all $s \in \mathcal{Z}^n$, i.e., $|L_s(w) - L_s(w')| \leq Ld(w,w')$. Then*

$$\left|\mathrm{gen}(\mu, P_{W|S})\right| \leq L \cdot \mathbb{E}\left[\mathbb{W}(P_W, P_{W|S})\right], \tag{13}$$

*where the expectation is taken over $P_S$.*

*Proof.* See Appendix C. □

Theorem 2 does not require the empirical risk function to have the form in (2). Nonetheless, if $\ell(\cdot,z)$ is $L$-Lipschitz for all $z \in \mathcal{Z}$, the empirical risk as defined in (2) is also $L$-Lipschitz. While the results of the previous section are more general as they do not require any Lipschitzianity, many loss functions in machine learning are indeed Lipschitz.

Observe that the Lipschitzianity requirement is equivalent to boundedness under the discrete metric. Hence, Theorem 2 implies that if the empirical risk function is bounded by $L$, i.e., $|L_s(w)| \leq L,\ \forall s \in \mathcal{Z}^n, \forall w \in \mathcal{W}$, then

$$\left|\mathrm{gen}(\mu, P_{W|S})\right| \leq 2L \cdot \mathsf{TV}(S;W). \tag{14}$$

As shown in Theorem 2, the generalization error is upper bounded by the expectation of the Wasserstein distance between the distribution of the output hypothesis and the corresponding distribution conditioned on the input dataset. So, it is natural to consider using $\mathbb{W}(P_W, P_{W|S=s})$ as a regularization term and minimize the empirical risk function along with the Wasserstein distance. Since $P_W$ depends on the (unknown) underlying data distribution $\mu$, we replace it with

the distribution $\delta_0$ (i.e., Dirac delta function). In other words, we consider

$$\begin{aligned}&\underset{P_{W|S=s}}{\operatorname{argmin}} \Big( \mathbb{E}\left[L_s(W)|S=s\right] + \lambda \mathbb{W}(P_{W|S=s}, \delta_0) \Big) \\ &= \underset{P_{W|S=s}}{\operatorname{argmin}} \Big( \mathbb{E}\left[L_s(W) + \lambda \|W\| | S=s\right] \Big). \end{aligned} \tag{15}$$

Note that in (15) the proposed regularization based on the Wasserstein distance echos the spirit of the well-known minimum description length problem [2].

Finally, we provide an example to illustrate the generalization bound given in Theorem 2.

**Example 2.** Let $Z_1, \ldots, Z_n$ be i.i.d. Gaussian random variables $\mathcal{N}(m, \sigma^2)$ and consider the empirical risk $L_S(w) = ||w| - |\bar{Z}||$, where $S = (Z_1, \cdots, Z_n)$ and $\bar{Z} \triangleq \frac{Z_1 + \cdots + Z_n}{n}$. Assume that $w$ is found using gradient descent with step size $\epsilon \ll m$ and initial point uniformly distributed in $(-\epsilon, \epsilon)$.

In order to bound $\mathbb{W}(P_W, P_{W|\bar{Z}=\bar{z}})$ for a given $\bar{z} \in \mathbb{R}$, we introduce

$$\widetilde{V}_{\bar{z}} = B\left(\bar{z} + U\right) \quad \text{and} \quad V = B(m + R),$$

where $\Pr(B = 1) = \Pr(B = -1) = 0.5$, $U$ is uniformly distributed over $(-\epsilon/2, \epsilon/2)$, and $R$ has a density function

$$p_R(r) = \frac{1}{\epsilon} \int_{-\infty}^{\infty} \mathbb{I}_{[r-\epsilon/2, r+\epsilon/2]}(\bar{z}) p_{\bar{Z}-m}(\bar{z}) \mathrm{d}\bar{z}. \tag{16}$$

Due to the triangle inequality,

$$\begin{aligned}&\mathbb{W}(P_W, P_{W|\bar{Z}=\bar{z}}) \\ &\leq \mathbb{W}(P_W, P_V) + \mathbb{W}(P_V, P_{\widetilde{V}_{\bar{z}}}) + \mathbb{W}(P_{\widetilde{V}_{\bar{z}}}, P_{W|\bar{Z}=\bar{z}}). \end{aligned} \tag{17}$$

We first provide an upper bound for $\mathbb{W}(P_V, P_{\widetilde{V}_{\bar{z}}})$. By the definition of Wasserstein distance, we have

$$\begin{aligned}\mathbb{W}(P_V, P_{\widetilde{V}_{\bar{z}}}) &\leq \mathbb{E}\left[\left|\widetilde{V}_{\bar{z}} - V\right|\right] \\ &\leq |\bar{z} - m| + \mathbb{E}\left[|U|\right] + \mathbb{E}\left[|R|\right] \\ &= |\bar{z} - m| + \frac{\epsilon}{4} + \mathbb{E}\left[|R|\right].\end{aligned}$$

Observe that, for all $r \in \mathbb{R}$, $p_R(-r) = p_R(r)$. Hence,

$$\begin{aligned}\mathbb{E}\left[|R|\right] &= 2\int_0^{2\epsilon} r p_R(r) \mathrm{d}r + 2\int_{2\epsilon}^{\infty} r p_R(r) \mathrm{d}r \\ &\leq 2\epsilon + 2\int_{2\epsilon}^{\infty} \int_{-\infty}^{\infty} \frac{r}{\epsilon} \mathbb{I}_{[r-\epsilon/2, r+\epsilon/2]}(\bar{z}) p_{\bar{Z}-m}(\bar{z}) \mathrm{d}\bar{z}\mathrm{d}r,\end{aligned}$$

where we used (16). By Tonelli's theorem,

$$\begin{aligned}\mathbb{E}\left[|R|\right] &\leq 2\epsilon + 2\int_{-\infty}^{\infty} \int_{2\epsilon}^{\infty} \frac{r}{\epsilon} \mathbb{I}_{[\bar{z}-\epsilon/2, \bar{z}+\epsilon/2]}(r) p_{\bar{Z}-m}(\bar{z}) \mathrm{d}r\mathrm{d}\bar{z} \\ &\leq 2\epsilon + 4\int_0^{\infty} \bar{z} p_{\bar{Z}-m}(\bar{z}) \mathrm{d}\bar{z} \\ &= 2\epsilon + 2\sqrt{\frac{2\sigma^2}{\pi n}},\end{aligned}$$

where the last equality follows from the fact that $\mathbb{E}\left[|N - m|\right] = \sqrt{2\sigma^2/\pi}$ whenever $N \sim \mathcal{N}(m, \sigma^2)$. Therefore,

$$\mathbb{W}(P_V, P_{\widetilde{V}_{\bar{z}}}) \leq |\bar{z} - m| + \frac{9\epsilon}{4} + 2\sqrt{\frac{2\sigma^2}{\pi n}}. \tag{18}$$

The randomness of choosing an initial point will be translated into the randomness of the output hypothesis produced by gradient descent and, consequently,

$$p_{W|\bar{Z}=\bar{z}}(w) = \frac{1}{2\epsilon}\left(\mathbb{I}_{|w+\bar{z}| \leq \epsilon/2} + \mathbb{I}_{|w-\bar{z}| \leq \epsilon/2}\right),$$

whenever $|\bar{z}| > \epsilon/2$. In this case, $\mathbb{W}(P_{\widetilde{V}_{\bar{z}}}, P_{W|\bar{Z}=\bar{z}}) = 0$. When $|\bar{z}| \leq \epsilon/2$, it can be verified that $p_{W|\bar{Z}=\bar{z}}$ and $p_{\widetilde{V}_{\bar{z}}}$ are supported within $[-\epsilon, \epsilon]$. Therefore,

$$\mathbb{W}(P_{\widetilde{V}_{\bar{z}}}, P_{W|\bar{Z}=\bar{z}}) \leq \begin{cases} 2\epsilon & |\bar{z}| \leq \epsilon/2, \\ 0 & |\bar{z}| > \epsilon/2. \end{cases} \tag{19}$$

Finally, it can be shown that, for $|w| > \epsilon$,

$$\begin{aligned}p_W(w) &= \int_{-\infty}^{\infty} p_{\bar{Z}}(\bar{z}) p_{W|\bar{Z}=\bar{z}}(w) \mathrm{d}\bar{z} \\ &= \int_{-w-\epsilon/2}^{-w+\epsilon/2} \frac{p_{\bar{Z}}(\bar{z})}{2\epsilon} \mathrm{d}\bar{z} + \int_{w-\epsilon/2}^{w+\epsilon/2} \frac{p_{\bar{Z}}(\bar{z})}{2\epsilon} \mathrm{d}\bar{z} \\ &= p_V(w).\end{aligned}$$

A routine argument then reveals that

$$\mathbb{W}(P_W, P_V) \leq 2\epsilon. \tag{20}$$

By plugging (18), (19), and (20) in (17), we obtain that

$$\mathbb{E}\left[\mathbb{W}(P_W, P_{W|\bar{Z}=\bar{z}})\right] \leq \frac{25}{4}\epsilon + 3\sqrt{\frac{2\sigma^2}{\pi n}},$$

where we used the fact that $\mathbb{E}\left[|\bar{Z} - m|\right] = \sqrt{\frac{2\sigma^2}{\pi n}}$. Since $L_s(\cdot)$ is 1-Lipschitz for all $s$, Theorem 2 implies in this case that

$$\left|\mathrm{gen}(\mu, P_{W|S})\right| \leq \frac{25}{4}\epsilon + 3\sqrt{\frac{2\sigma^2}{\pi n}}. \tag{21}$$

Observe that in this example $\mathbb{E}\left[L_S(W)\right] \approx \epsilon$. Hence, the bound in (21) bears a similar structure as the bound in Theorem 1. Nonetheless, computing the bound in Theorem 1 for this example proves to be much harder than the present derivations. The reason is that two distributions supported over $(-\epsilon, \epsilon)$ may have a large (even maximum!) total variation, while their Wasserstein distance is always not greater than $2\epsilon$.

## IV. Conclusions

By following an information-theoretic perspective, we derived two bounds on the generalization error of learning algorithms. Motivated by inherent drawbacks of existing bounds based on mutual information, we rather employed two different instances of the Wasserstein distance, by considering either the discrete metric or — at the expense of introducing a Lipschitz constraint on the loss function — a generic metric. We hope our results provide a geometric insight on the interplay between the in-sample approximation and out-of-sample generalization.

## REFERENCES

[1] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 2524–2533.

[2] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[3] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, no. Mar, pp. 499–526, 2002.

[4] D. A. McAllester, "Some PAC-Bayesian theorems," *Mach. Learn.*, vol. 37, no. 3, pp. 355–363, 1999.

[5] D. Russo and J. Zou, "How much does your data exploration overfit? Controlling bias via information usage," *arXiv preprint arXiv:1511.05219*, 2015.

[6] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[7] I. M. Alabdulmohsin, "Algorithmic stability and uniform generalization," in *Adv. Neural Inf. Process. Syst.*, 2015, pp. 19–27.

[8] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, "Information-theoretic analysis of stability and bias of learning algorithms," in *IEEE Inf. Theory Workshop*, 2016, pp. 26–30.

[9] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *Proc. 2017 IEEE Int. Symp. on Inf. Theory*, 2017, pp. 1475–1479.

[10] I. Issa and M. Gastpar, "Computable bounds on the exploration bias," in *Proc. 2018 IEEE Int. Symp. on Inf. Theory*, 2018, pp. 576–580.

[11] A. R. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," *arXiv preprint arXiv:1806.03803*, 2018.

[12] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 35–55, 2016.

[13] C. Villani, *Optimal transport: Old and new*. Springer Science & Business Media, 2008, vol. 338.

[14] A. T. Lopez and V. Jog, "Generalization error bounds using wasserstein distances," in *IEEE Inf. Theory Workshop*. IEEE, 2018, pp. 1–5.

[15] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Int. Stat. Rev.*, vol. 70, no. 3, pp. 419–435, 2002.

[16] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Int. Conf. on Mach. Learn.*, 2017, pp. 214–223.


## APPENDIX A
## PROOF OF LEMMA 1

Let $\bar{U} \triangleq f(\bar{X}) - \mathbb{E}\left[f(\bar{X})\right]$ and $U \triangleq f(X) - \mathbb{E}\left[f(\bar{X})\right]$. For any $T \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[f(\bar{X}) - f(X)\right] = \mathbb{E}\left[\bar{U}\right] - \mathbb{E}\left[U\right] \\
&= \mathbb{E}\left[\bar{U}\mathbb{I}_{-\mathbb{E}\left[f(\bar{X})\right]\leq\bar{U}<0}\right] + \mathbb{E}\left[\bar{U}\mathbb{I}_{0\leq\bar{U}\leq T}\right] + \mathbb{E}\left[\bar{U}\mathbb{I}_{T<\bar{U}}\right] \\
&\quad - \mathbb{E}\left[U\mathbb{I}_{-\mathbb{E}\left[f(\bar{X})\right]\leq U<0}\right] - \mathbb{E}\left[U\mathbb{I}_{0\leq U\leq T}\right] - \mathbb{E}\left[U\mathbb{I}_{T<U}\right] \\
&\leq \left|\mathbb{E}\left[\bar{U}\mathbb{I}_{-\mathbb{E}\left[f(\bar{X})\right]\leq\bar{U}<0}\right] - \mathbb{E}\left[U\mathbb{I}_{-\mathbb{E}\left[f(\bar{X})\right]\leq U<0}\right]\right| \\
&\quad + \left|\mathbb{E}\left[\bar{U}\mathbb{I}_{0\leq\bar{U}\leq T}\right] - \mathbb{E}\left[U\mathbb{I}_{0\leq U\leq T}\right]\right| + \mathbb{E}\left[\bar{U}\mathbb{I}_{T<\bar{U}}\right]. \quad (22)
\end{aligned}$$

To bound the first term in (22), we define, for all $x \in \mathcal{X}$,

$$g(x) \triangleq \left(\mathbb{E}\left[f(\bar{X})\right] - f(x)\right)\mathbb{I}_{0\leq f(x)\leq\mathbb{E}\left[f(\bar{X})\right]}.$$

Note that $g(x) \geq 0$ and $\|g\|_\infty \leq \mathbb{E}\left[f(\bar{X})\right]$. It can be proved that for a measurable and non-negative function $h : \mathcal{X} \to \mathbb{R}$,

$$\left|\mathbb{E}\left[h(\bar{X}) - h(X)\right]\right| \leq \|h\|_\infty \mathsf{TV}(P, Q), \quad (23)$$

where $\bar{X} \sim Q$ and $X \sim P$. Hence,

$$\begin{aligned}
&\left|\mathbb{E}\left[\bar{U}\mathbb{I}_{-\mathbb{E}\left[f(\bar{X})\right]\leq\bar{U}<0}\right] - \mathbb{E}\left[U\mathbb{I}_{-\mathbb{E}\left[f(\bar{X})\right]\leq U<0}\right]\right| \\
&= \left|\mathbb{E}\left[g(\bar{X}) - g(X)\right]\right| \leq \mathbb{E}\left[f(\bar{X})\right] \cdot \mathsf{TV}(P, Q). \quad (24)
\end{aligned}$$

Similarly,

$$\left|\mathbb{E}\left[\bar{U}\mathbb{I}_{0\leq\bar{U}\leq T}\right] - \mathbb{E}\left[U\mathbb{I}_{0\leq U\leq T}\right]\right| \leq T \cdot \mathsf{TV}(P, Q). \quad (25)$$

We now bound the last term in (22). Recall that for any non-negative random variable $V$, $\mathbb{E}[V] = \int_0^\infty \Pr(V > v)\mathrm{d}v$. Thus,

$$\mathbb{E}\left[\bar{U}\mathbb{I}_{\bar{U}>T}\right] = T\Pr\left(\bar{U} > T\right) + \int_T^\infty \Pr\left(\bar{U} > t\right)\mathrm{d}t. \quad (26)$$

Since $\bar{U} = f(\bar{X}) - \mathbb{E}\left[f(\bar{X})\right]$ is $\sigma$-subgaussian, we have that $\Pr(\bar{U} > t) \leq e^{-t^2/(2\sigma^2)}$ for all $t \geq 0$. In particular,

$$\int_T^\infty \Pr\left(\bar{U} > t\right)\mathrm{d}t \leq \int_T^\infty e^{-t^2/(2\sigma^2)}\mathrm{d}t \leq \frac{\sigma^2}{T}e^{-T^2/(2\sigma^2)},$$

where the last inequality follows from Mill's inequality. Hence,

$$\mathbb{E}\left[\bar{U}\mathbb{I}_{\bar{U}>T}\right] \leq Te^{-T^2/(2\sigma^2)} + \frac{\sigma^2}{T}e^{-T^2/(2\sigma^2)}. \quad (27)$$

By plugging (24), (25), and (27) in (22), we conclude that $\mathbb{E}\left[f\left(\bar{X}\right) - f(X)\right]$ is upper bounded by

$$\left(\mathbb{E}\left[f(\bar{X})\right] + T\right)\mathsf{TV}(P, Q) + \left(T + \frac{\sigma^2}{T}\right)e^{-T^2/(2\sigma^2)}.$$

By taking $T = \sqrt{-2\sigma^2\log(\delta)}$ with $\delta \triangleq \mathsf{TV}(P, Q)$, we obtain that $\mathbb{E}\left[f\left(\bar{X}\right) - f(X)\right]$ is upper bounded by

$$\mathbb{E}\left[f(\bar{X})\right] \cdot \delta + \sqrt{2}\sigma\delta\left(2\sqrt{-\log(\delta)} + \frac{1}{2\sqrt{-\log(\delta)}}\right).$$

Therefore, we get the desired conclusion.

## APPENDIX B
## PROOF OF THEOREM 1

We set $P = P_{S,W}$, $Q = P_S \otimes P_W$, and $f(s, w) = \frac{1}{n}\sum_{i=1}^n \ell(w, z_i)$ in Lemma 1 and follow a similar procedure as the decoupling estimate proposed in [1]. Specifically, the expected generalization error can be written as

$$\mathrm{gen}(\mu, P_{W|S}) = \mathbb{E}\left[f(\bar{S}, \bar{W})\right] - \mathbb{E}\left[f(S, W)\right], \quad (28)$$

where the joint distributions of $(S, W)$ and $(\bar{S}, \bar{W})$ are $P_{S,W} = \mu^{\otimes n} \otimes P_{W|S}$ and $P_{\bar{S},\bar{W}} = P_S \otimes P_W$, respectively. If $\ell(w, Z)$ is $\sigma$-subgaussian for all $w \in \mathcal{W}$, then $f(S, w)$ is $\sigma/\sqrt{n}$-subgaussian due to the i.i.d. assumption on $Z_i$. Consequently, $f(\bar{S}, \bar{W})$ is $\sigma/\sqrt{n}$-subgaussian since $\bar{S}$ and $\bar{W}$ are independent.

## APPENDIX C
## PROOF OF THEOREM 2

Let $(S, W) \sim P_{S,W}$, $(\bar{S}, \bar{W}) \sim P_S \otimes P_W$. Observe that

$$\begin{aligned}
\left|\mathrm{gen}(\mu, P_{W|S})\right| &= \left|\mathbb{E}\left[L_\mu(W) - L_S(W)\right]\right| \\
&= \left|\mathbb{E}\left[L_{\bar{S}}(\bar{W})\right] - \mathbb{E}\left[L_S(W)\right]\right|.
\end{aligned}$$

By a conditioning argument, we obtain that

$$\begin{aligned}
&\left|\mathrm{gen}(\mu, P_{W|S})\right| \\
&= \left|\int \mathrm{d}P_S(s)\left(\mathbb{E}\left[L_s(\bar{W})\right] - \mathbb{E}\left[L_s(W)|S = s\right]\right)\right| \\
&\leq L \cdot \mathbb{E}\left[\mathbb{W}(P_W, P_{W|S})\right], \quad (29)
\end{aligned}$$

where (29) follows from the Kantorovich-Rubinstein duality.