Chained Generalisation Bounds

Eugenio Clerico CLERICO@STATS.OX.AC.UK
Amitis Shidani SHIDANI@STATS.OX.AC.UK
George Deligiannidis DELIGIAN@STATS.OX.AC.UK
Arnaud Doucet DOUCET@STATS.OX.AC.UK
Department of Statistics, University of Oxford, UK.

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

This work discusses how to derive upper bounds for the expected generalisation error of supervised learning algorithms by means of the chaining technique. By developing a general theoretical framework, we establish a duality between generalisation bounds based on the regularity of the loss function, and their chained counterparts, which can be obtained by lifting the regularity assumption from the loss onto its gradient. This allows us to re-derive the chaining mutual information bound from the literature, and to obtain novel chained information-theoretic generalisation bounds, based on the Wasserstein distance and other probability metrics. We show on some toy examples that the chained generalisation bound can be significantly tighter than its standard counterpart, particularly when the distribution of the hypotheses selected by the algorithm is very concentrated.

Keywords: Generalisation bounds; Chaining; Information-theoretic bounds; Mutual information; Wasserstein distance; PAC-Bayes.

1. Introduction

In the supervised setting, a learning algorithm is a procedure that takes a training dataset as input and returns a hypothesis (*e.g.*, regression coefficients, weights of a neural network, etc.). Ideally, the learned hypothesis should perform well on both the input dataset and new data, which were not used for the training. There is hence interest in providing generalisation bounds, namely upper bounds on the algorithm's gap in performance for seen and unseen instances.

The first generalisation bounds were based on characterisations of the hypothesis space's complexity, such as the VC dimension or the Rademacher complexity (Bousquet et al., 2004; Vapnik, 2000; Shalev-Shwartz and Ben-David, 2014). However, due to their algorithm-independent nature, these bounds must hold even for the worst algorithm on the given hypothesis space. Consequently, they are often inadequate for modern over-parameterised neural networks, with the complexity measure usually scaling exponentially with the architecture's depth (Anthony and Bartlett, 2002; Zhang et al., 2021; Belkin et al., 2018).

To address this issue, recent approaches aim at providing algorithm-dependent generalisation bounds. The underlying intuition is that if the output hypothesis is less dependent on the input dataset, it would be less prone to overfitting, and so generalises better. Among the results building on this idea, there are bounds based on uniform stability (Bousquet and Elisseeff, 2002) and differential privacy (Dwork and Roth, 2014), PAC-Bayesian bounds (Guedj, 2019; McAllester, 1998, 1999), and information-theoretic bounds.

In this paper, we shall mainly focus on the information-theoretic framework, where the learning algorithm is seen as a noisy channel connecting the input dataset and the chosen hypothesis.

Russo and Zou (2019) and Xu and Raginsky (2017) were the first to introduce this approach. They upper-bounded the expected generalisation error via the Mutual Information (MI) between the input sample and the learnt hypothesis. This bound is simple and can be applied to a broad class of learning algorithms. However, a major drawback is that it becomes infinite if the choice of the hypothesis is deterministic in the input. Motivated by this problem, several strategies have been proposed.

Bu et al. (2019) gave an individual-sample MI bound, while Steinke and Zakynthinou (2020) introduced a conditional version of the MI, which is always finite. Rodríguez-Gálvez et al. (2020), Haghifam et al. (2020), and Hellström and Durisi (2020) extended and merged these results. Alternatively, different measures of algorithmic stability can replace the MI: Lopez and Jog (2018), Wang et al. (2019), and Rodríguez-Gálvez et al. (2021) proposed bounds based on the Wasserstein distance, while others focused on total variation, f-divergences, and lautum information (Wang et al., 2019; Rodríguez-Gálvez et al., 2021; Esposito et al., 2021; Palomar and Verdú, 2008).

Adopting a different perspective, Asadi et al. (2018) observed that several information-theoretic bounds fail to exploit the dependencies between hypotheses. They hence proposed to combine the original MI bound with the chaining method, a powerful tool from high dimensional probability originally aimed at upper-bounding the expected supremum of random processes. First introduced by Kolmogorov (see van Handel (2016)), the chaining technique has been successfully extended and developed (Dudley, 1967; Talagrand, 2005, 2014). In their Chaining Mutual Information (CMI) bound, Asadi et al. (2018) take finer and finer discretisations of the hypothesis space and rewrite the generalisation error as a telescopic sum, whose terms can be controlled by exploiting the dependencies between the hypotheses. Subsequently, Asadi and Abbe (2020) adapted the CMI technique to the architecture of deep neural nets, while Zhou et al. (2022) introduced bounds based on a stochastic version of chaining. However, it is worth mentioning that previous works had already applied the chaining method to algorithm-dependent bounds. For instance, Audibert and Bousquet (2004) combined the generic chaining from Talagrand (2005) with the PAC-Bayesian approach.

As a final comment, it must be noted that the generalisation bounds from the information-theoretic literature are hard to evaluate in practice, involving expectations with respect to the unknown sample distribution. Nevertheless, they provide useful intuition on the mechanism of the learning process and, as a result, they represent a very active research area. Moreover, recent works have built on them to derive computable analytical bounds for specific algorithms, such as Langevin dynamics, stochastic gradient Langevin dynamics, and stochastic gradient descent (Bu et al., 2019; Negrea et al., 2019; Haghifam et al., 2020; Rodríguez-Gálvez et al., 2020; Neu et al., 2021).

1.1. Our contributions

The CMI bound is an interesting multi-scale reformulation of the original MI result by Russo and Zou (2019). However, in the information-theoretic literature on generalisation bounds, the chaining method has been coupled only with the MI (Asadi et al., 2018; Asadi and Abbe, 2020; Zhou et al., 2022). Two questions then naturally arise. *Is it possible to derive chained versions of other kinds of generalisation bounds? Can these chained bounds be tighter than their original counterparts?*

In the present work, we establish a duality that reads as follows. *Each bound, based on (a certain notion of) regularity of the loss function, corresponds to a chained bound that can be obtained by lifting the regularity condition from the loss to its gradient.* To make sense of this, we first introduce a general framework, standardising the main step in the proof of several information-theoretic bounds from the literature. We then discuss how to extend this framework leveraging the chaining technique, and we provide a simple method to derive novel chained generalisation bounds. We show indeed that in our framework each unchained bound corresponds to a chained one (see Theorems 2 and 4), in a way reflecting the connection between the MI and CMI results.

The framework introduced in this work encompasses several information-theoretic *backward-channel*[1] bounds, and allows us to derive their chained counterparts. However, due to space limitations, many explicit results are deferred to Appendix G (see Table 1) and in the main text we focus on four bounds to concretely illustrate how our framework works: the MI bound from Russo and Zou (2019) and the CMI bound from Asadi et al. (2018) serve as a motivation for our general result, while as an application of our framework we derive a novel Wasserstein bound (see Proposition 15), which is the chained counterpart of a bound from Lopez and Jog (2018).

Moreover, we discuss some possible extensions of our work. On the one hand, our information-theoretic framework can be restated with weaker regularity assumptions on both the loss and the hypothesis space. On the other hand, we present an additional bound that does not fit our theoretical framework but can still be derived using essentially the same technical machinery. It is a chained PAC-Bayesian generalisation result, which has the interesting features of being finite even for deterministic algorithms and not requiring the loss to be bounded by a small constant.

As a final remark, there is no generic answer on whether the chained bounds are tighter than their unchained counterparts. However, the chaining technique turns out to be particularly effective when the hypotheses' distribution is very concentrated. In fact, many of the standard bounds do not exploit this feature, the most pathological case being the MI bound, which can even be infinite. In contrast, the chained bounds can be significantly tighter, intrinsically leveraging the dependencies between different hypotheses. We illustrate this phenomenon through some simple toy examples.

2. Preliminaries

Let the input space $(\mathcal{X}, d_\mathcal{X})$ be a separable complete metric space, and $\Sigma_\mathcal{X}$ the corresponding Borel σ-algebra. We define $\mathcal{S} = \mathcal{X}^m$ and consider a metric $d_\mathcal{S}$ inducing the product σ-algebra $\Sigma_\mathcal{S} = \Sigma_\mathcal{X}^{\otimes m}$. We denote the training dataset as $s = \{x_1, \dots, x_m\} \in \mathcal{S}$. Let $\mathbb{P}_X$ be a probability measure on $\mathcal{X}$ and X a random variable with law $\mathbb{P}_X$. $S = \{X_1, \dots, X_m\} \in \mathcal{S}$ denotes the random training sample, with law $\mathbb{P}_S$. We will always assume that the marginal $\mathbb{P}_{X_i} = \mathbb{P}_X$, for each index i. This is of course the case if the X_i are i.i.d. ($\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$). We will suppose that the hypothesis space $\mathcal{W}$ is a closed subset of $\mathbb{R}^d$, endowed with its Borel σ-algebra $\Sigma_\mathcal{W}$. A learning algorithm consists in a Markov kernel that maps each $s \in \mathcal{S}$ to a probability measure $\mathbb{P}_{W|S=s}$ on $\mathcal{W}$. In turn, this defines a joint probability $\mathbb{P}_{W,S}$ on $\mathcal{W} \times \mathcal{S}$. We denote as $\mathbb{P}_W$ and $\mathbb{P}_S$ the marginal distributions of $\mathbb{P}_{W,S}$, and we let $s \mapsto \mathbb{P}_{W|S=s}$ and $w \mapsto \mathbb{P}_{S|W=w}$ be regular conditional probabilities[2].

In the supervised framework, the goal is to approximate a map $x \mapsto f_\star(x)$ by making use of the information contained in the training sample s (the value of $f^\star(x_i)$ is known for each $x_i \in s$). Each hypothesis w represents a parameterised mapping $x \mapsto f_w(x)$, and the training process consists in tuning w, so as to approximate $f^\star$. The loss $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, allows to assess how far each $f_w(x)$ is

[1] In the information-theoretic literature, the *forward-channel* connects the sample to the hypothesis, while the *backward-channel* goes the other way. Chaining on the hypotheses combines naturally with the *backward-channel*.

[2] The existence of these is ensured by the fact that $\mathcal{S}$ and $\mathcal{W}$ are Polish spaces, cf. Theorem 10.2.2 in Dudley (2002).

from $f^\star(x)$. We will always assume that $\ell(w,\cdot) \in L^1(\mathbb{P}_X)$. Define the empirical and the true loss

$$\mathscr{L}_s(w) = \frac{1}{m}\sum_{i=1}^m \ell(w,x_i)\,; \qquad \mathscr{L}_{\mathcal{X}}(w) = \mathbb{E}_{\mathbb{P}_X}[\ell(w,X)]\,.$$

We call generalisation error the difference $g_s(w) = \mathscr{L}_{\mathcal{X}}(w) - \mathscr{L}_s(w)$. In this work, we are essentially interested in upper-bounding its expected value $\mathcal{G} = \mathbb{E}_{\mathbb{P}_{W,S}}[g_S(W)]$.

The equality $\mathbb{E}_{\mathbb{P}_{W,S}}[\mathscr{L}_{\mathcal{X}}(W)] = \mathbb{E}_{\mathbb{P}_{W\otimes S}}[\mathscr{L}_S(W)]$, where $\mathbb{P}_{W\otimes S} = \mathbb{P}_W \otimes \mathbb{P}_S$, follows from $\mathscr{L}_{\mathcal{X}}(w) = \mathbb{E}_{\mathbb{P}_S}[\mathscr{L}_S(w)]$ and is the starting point of several information-theoretic bounds. Indeed,

$$\mathcal{G} = \mathbb{E}_{\mathbb{P}_{W\otimes S}}[\mathscr{L}_S(W)] - \mathbb{E}_{\mathbb{P}_{W,S}}[\mathscr{L}_S(W)]$$

can be upper-bounded in terms of how "far apart" $\mathbb{P}_{W,S}$ and $\mathbb{P}_{W\otimes S}$ are.

2.1. Further notation and conventions

The following notation will be used throughout the rest of the paper. $(\mathcal{Z},\Sigma_{\mathcal{Z}})$ denotes a generic separable complete metric space, endowed with the Borel σ-algebra induced by its metric $d_{\mathcal{Z}}$. We endow $\mathscr{P}$, the space of all the probability measures on $(\mathcal{Z},\Sigma_{\mathcal{Z}})$, with the topology of the weak convergence, and we denote the corresponding Borel σ-algebra as $\Sigma_{\mathscr{P}}$. For two coupled random variables Z, Z' on $\mathcal{Z}$, we write $\mathbb{P}_{Z\otimes Z'}$ for the independent coupling $\mathbb{P}_Z \otimes \mathbb{P}_{Z'}$. For $v, v' \in \mathbb{R}^q$ (for a generic $q \in \mathbb{N}$) we write $\|v\|$ and $v \cdot v'$ for the Euclidean norm and the standard dot product in $\mathbb{R}^q$ respectively. For a random vector $V \in \mathbb{R}^q$, we write that $V \in L^1$ if $\mathbb{E}_{\mathbb{P}_V}[\|V\|] < +\infty$. When we need to specify the integrability of V with respect to a particular law μ, we explicitly write $V \in L^1(\mu)$, that is $\mathbb{E}_\mu[\|V\|] < +\infty$. Finally, ξ denotes an arbitrary non-negative real number.

3. General framework

3.1. Bounds based on the regularity of the loss

Both the standard MI and Wasserstein bounds from Russo and Zou (2019) and Lopez and Jog (2018) (see Propositions 10 and 11 in Section 4 for the explicit statements) build on some regularity condition on the dependence of ℓ in x, holding uniformly on $\mathcal{W}$. As this is a common assumption for various *backward-channel* bounds in the literature, we will now introduce a unified abstract framework, which allows us to re-derive several information-theoretic bounds, such as many of those based on MI, Wasserstein distances, and other probability metrics. Due to the limited space, in the main text we only give a few concrete applications of our framework (see Section 4). A wide range of additional explicit examples, listed in Table 1, can be found in Appendix G.

Definition 1 ($\mathfrak{D}$-regularity) *Let $\mathfrak{D}$ be a measurable*[3] *map $\mathscr{P} \times \mathscr{P} \to [0,+\infty]$. Fix $\mu \in \mathscr{P}$ and $\xi \geq 0$. We say that $f : \mathcal{Z} \to \mathbb{R}$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$, with respect to μ, if $f \in L^1(\mu)$ and, for every $\nu \in \mathscr{P}$ such that* $\mathrm{Supp}(\nu) \subseteq \mathrm{Supp}(\mu)$ *and $f \in L^1(\nu)$,*

$$|\mathbb{E}_\mu[f(Z)] - \mathbb{E}_\nu[f(Z)]| \leq \xi\, \mathfrak{D}(\mu,\nu)\,.$$

We can extend the definition to functions taking values in $\mathbb{R}^q$, for $q > 1$. We say that $F : \mathcal{Z} \to \mathbb{R}^q$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ (wrt μ) if $z \mapsto v \cdot F(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi\|v\|)$ (wrt μ), for all $v \in \mathbb{R}^q$.

[3] The measurability wrt $\Sigma_{\mathscr{P}}$ is a technical assumption that is required in order to ensure that expressions, such as $\int_{\mathcal{W}} \mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W=w})\mathrm{d}\mathbb{P}_W(w)$ in Theorem 2, make sense. The reader can be assured that it holds whenever $\mathfrak{D}$ is a measurable function of an f-divergence, or the Wasserstein distance. We refer to Appendix F for more details.

The concept of $\mathfrak{D}$-regularity is intrinsically connected to the choice of the measure $\mu \in \mathscr{P}$, in the sense that f might be $\mathcal{R}_{\mathfrak{D}}(\xi)$ regular with respect to μ, but not with respect to some other $\mu' \in \mathscr{P}$. For two simple concrete examples of $\mathfrak{D}$-regularity, we refer to Lemma 9, in Section 4.

Now, let $\mathcal{Z} = \mathcal{S}$ and recall that $\mathcal{W}$ is a closed subset of $\mathbb{R}^d$, with Borel σ-algebra $\Sigma_{\mathcal{W}}$. On the product space $(\mathcal{W} \times \mathcal{S}, \Sigma_{\mathcal{W}} \otimes \Sigma_{\mathcal{S}})$, we consider a probability measure $\mathbb{P}_{W,S}$, with marginals $\mathbb{P}_W$ and $\mathbb{P}_S$. Recall that since $\mathcal{S}$ is a Polish space, $w \mapsto \mathbb{P}_{S|W=w}$ defines a regular conditional probability (cf. Theorem 10.2.2 in Dudley (2002)). The next result, which follows easily from the definition of regularity, is a powerful tool to derive generalisation bounds.

Theorem 2 *Assume that $s \mapsto \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt $\mathbb{P}_S$, $\forall w \in \mathcal{W}$. Then we have*

$$|\mathcal{G}| = |\mathbb{E}_{\mathbb{P}_{W \otimes S}}[\mathscr{L}_S(W)] - \mathbb{E}_{\mathbb{P}_{W,S}}[\mathscr{L}_S(W)]| \leq \xi \, \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W})] \, ,$$

where $\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W})] = \int_{\mathcal{W}} \mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W=w}) \, \mathrm{d}\mathbb{P}_W(w)$.[4]

By specialising the concept of $\mathfrak{D}$-regularity, we can leverage the framework introduced so far and obtain generalisation bounds based on various probability divergences (cf. Table 1). Moreover, individual-sample bounds such as those from Bu et al. (2019) can fit in our framework, as well as bounds based on the random sub-sampling from a super-sample, in the same spirit of the conditional MI bound from Steinke and Zakynthinou (2020). We refer the reader to Appendix G for a more detailed discussion of these results.

3.2. Bounds based on the regularity of the loss's gradient

The bounds based on the chaining technique, such as the CMI bound from Asadi et al. (2018) (see Proposition 12 in Section 4), do not fit naturally in the framework presented so far. We are thus motivated to find an alternative setting that naturally gives rise to chained bounds, thus establishing new generalisation results.

As a starting point, let us notice that the main idea behind the CMI bound is to lift the regularity assumption from $x \mapsto \ell(w,x)$ onto $x \mapsto (\ell(w,x) - \ell(w',x))$. A natural guess is that this approach could provide chained bounds also in our general framework, and this is indeed the case (cf. Theorem 22 in Appendix B.1). However, if ℓ is regular enough we can focus on the gradient $\nabla_w \ell(w,x)$ instead. Since this leads to more intuitive and compact statements, we chose to consider this case in the main text.

Assumptions ♣

- *The set $\mathcal{W} \subset \mathbb{R}^d$ is convex, compact, and with non-empty interior.*
- *The function $w \mapsto \ell(w,x)$ is of class C^1 on $\mathcal{W}$, $\mathbb{P}_X$-a.s.*
- *We have $\sup_{(w,x) \in \mathcal{W} \times \mathcal{X}} |\ell(w,x)| < +\infty$ and $\sup_{(w,x) \in \mathcal{W} \times \mathcal{X}} \|\nabla_w \ell(w,x)\| < +\infty$.*

Let us stress once more that the above assumptions are not necessary in order to obtain the duality chained-unchained bounds. In Appendix B.1 we discuss a more general setting: $\mathcal{W}$ can be non-convex and with empty interior, ℓ continuous on $\mathcal{W}$ ($\mathbb{P}_X$-a.s.) and only bounded in expectation.

The chained bounds involve a sequence of finer and finer discretisations of the hypotheses' space, which can be formalised as follows.

[4] Note that $w \mapsto \mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W=w})$ is measurable, since both $w \mapsto \mathbb{P}_{S|W=w}$ and $(\mu,\nu) \mapsto \mathfrak{D}(\mu,\nu)$ are Borel measurable (see Appendix F).

Definition 3 (Nets and refining sequences of nets) *Given $\varepsilon > 0$, we define an ε-projection on $\mathcal{W}$ as a measurable mapping $\pi : \mathcal{W} \to \mathcal{W}$ such that $\pi(\mathcal{W})$ has finitely many elements and, for all $w \in \mathcal{W}$, $\|\pi(w) - w\| \leq \varepsilon$. The image $\pi(\mathcal{W})$ is called an ε-net on $\mathcal{W}$.*
Consider a positive, vanishing, decreasing sequence $\{\varepsilon_k\}_{n\in\mathbb{N}}$, and assume that there is a $w_0 \in \mathcal{W}$ such that $\|w - w_0\| \leq \varepsilon_0$ for each $w \in \mathcal{W}$. We call $\{\pi_k(\mathcal{W})\}_{n\in\mathbb{N}}$ an $\{\varepsilon_k\}$-refining sequence of nets if $\pi_0(\mathcal{W}) = \{w_0\}$ and, for all $k \geq 1$, we have that π_k is a ε_k-projection and $\pi_{k-1} \circ \pi_k = \pi_{k-1}$.

To simplify the notation, for all $w \in \mathcal{W}$ we let $w_k = \pi_k(w)$, and similarly $W_k = \pi_k(W)$ and $\mathcal{W}_k = \pi_k(\mathcal{W})$. Note that for all k, $w_{k'}$ is determined by w_k whenever $k' \leq k$, as $w_{k'} = \pi_{k'}(w_k)$. Moreover, for all $k \geq 1$, $\|w_k - w_{k-1}\| = \|w_k - \pi_{k-1}(w_k)\| \leq \varepsilon_{k-1}$.

The next theorem is the main result of this work. Together with Theorem 2, it establishes the duality between chained and unchained generalisation bounds, which can essentially be obtained by lifting the regularity from the loss onto its gradient.

Theorem 4 *Assume $\clubsuit$ and that $s \mapsto \nabla_w \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt $\mathbb{P}_S$, $\forall w \in \mathcal{W}$. Then, for any $\{\varepsilon_k\}$-refining sequence of nets on $\mathcal{W}$,*

$$|\mathcal{G}| = |\mathbb{E}_{\mathbb{P}_{W\otimes S}}[\mathscr{L}_S(W)] - \mathbb{E}_{\mathbb{P}_{W,S}}[\mathscr{L}_S(W)]| \leq \xi \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})]\,,$$

where $\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] = \int_{\mathcal{W}} \mathfrak{D}\big(\mathbb{P}_S, \mathbb{P}_{S|W\in\pi_k^{-1}(w)}\big)\,\mathrm{d}\mathbb{P}_W(w)$.

Proof's sketch Here is a sketch of the proof; see Appendix A.3 for the details. Following the standard chaining argument, we control $\mathscr{L}_s(w)$ by the telescopic sum $\sum_{k\geq 1}(\mathscr{L}_s(w_k) - \mathscr{L}_s(w_{k-1}))$. The upper bound will then follow from the fact that the $\mathcal{R}_{\mathfrak{D}}(\xi)$ regularity of $s \mapsto \nabla_w \mathscr{L}_s(w)$ implies the $\mathcal{R}_{\mathfrak{D}}(\varepsilon_{k-1}\xi)$ regularity of $w \mapsto (\mathscr{L}_s(w_k) - \mathscr{L}_s(w_{k-1}))$. ■

Both Theorem 2 and 4 are stated under uniform regularity conditions, in the sense that the value of the regularity's parameter ξ has to be the same for all $w \in \mathcal{W}$. However, we can still achieve generalisation bounds under less strict assumptions. In Appendix B.2 we discuss the case of a measurable map $w \mapsto \xi_w$, such that, for some $p \in [1, +\infty]$, ξ_W is bounded in $L^p(\mathbb{P}_W)$ (or $L^p(\mathbb{P}_{W_k})$, $\forall k \in \mathbb{N}$). Note that choosing $p = +\infty$ brings back the uniform condition.

In a similar spirit, one might try to relax the definition of ε-net, by mimicking the stochastic chaining idea from Zhou et al. (2022). We defer this approach to future work.

4. A few concrete examples: MI and Wasserstein bounds

In the current section we give a few concrete applications of the abstract framework that we have presented so far. We recover some simple generalisation bounds from the literature and establish a novel chained bound, based on the Wasserstein distance.

First, we need to state a few standard definitions.

Definition 5 (Subgaussianity) *A real random variable $Z \in L^1$ is ξ-SubGaussian (ξ-SG) if*

$$\log \mathbb{E}_{\mathbb{P}_Z}[e^{\lambda Z}] \leq \lambda \mathbb{E}_{\mathbb{P}_Z}[Z] + \tfrac{\xi^2\lambda^2}{2}\,, \qquad \forall \lambda > 0\,.$$

A random vector $V \in \mathbb{R}^q$ is ξ-SG if, for all $v \in \mathbb{R}^q$, $V \cdot v$ is $(\|v\|\xi)$-SG. Finally, a stochastic process $\{F_w\}_{w\in\mathcal{W}}$ is ξ-SG if, for every pair $(w, w') \in \mathcal{W}^2$, $F_w - F_{w'}$ is a $(\|w - w'\|\xi)$-SG random variable.

Note that any bounded random variable $Z \in [a,b]$ is $\frac{b-a}{2}$-SG. Moreover, a normally distributed random variable $Z \sim \mathcal{N}(0,\xi)$ is ξ-SG.

Definition 6 (Lipschitzianity) *A function $f : \mathcal{Z} \to \mathbb{R}^q$ is ξ-Lipschitz on $\mathcal{Z}$ if, for all $z, z' \in \mathcal{Z}$,*

$$\|f(z) - f(z')\| \leq \xi d_{\mathcal{Z}}(z,z')\,.$$

Definition 7 (Kullback–Leibler divergence and mutual information) *Let μ and ν be two probability measures on $\mathcal{Z}$. We define the Kullback–Leibler divergence*

$$\mathrm{KL}(\nu\|\mu) = \begin{cases} \mathbb{E}_\nu[\log \mathrm{d}\nu/\mathrm{d}\mu] & \textit{if } \nu \ll \mu; \\ +\infty & \textit{otherwise.} \end{cases}$$

For two coupled random variables Z, Z', the Mutual Information (MI) is defined as

$$I(Z;Z') = \mathrm{KL}(\mathbb{P}_{Z,Z'}\|\mathbb{P}_{Z\otimes Z'}).$$

The KL divergence is non-negative, with $\mathrm{KL}(\nu\|\mu) = 0$ if, and only if, $\mu = \nu$. Similarly the MI is always non-negative, and null if, and only if, $Z \perp\!\!\!\perp Z'$.

Definition 8 (Wasserstein distance) *Given two distributions μ and ν on $\mathcal{Z}$ and fixed $p \geq 1$, their p-Wasserstein distance $\mathfrak{W}_p$ is defined as*

$$\mathfrak{W}_p(\mu,\nu) = \inf_{\pi \in \Pi[\mu,\nu]} \mathbb{E}_{(Z,Z')\sim\pi}[d_{\mathcal{Z}}(Z,Z')^p]^{1/p}\,,$$

where $\Pi[\mu,\nu]$ is the set of all probability measures, on $(\mathcal{Z}^2, \Sigma_{\mathcal{Z}} \otimes \Sigma_{\mathcal{Z}})$, with marginals μ and ν.

It can be shown that for $p > p'$ we have $\mathfrak{W}_p(\mu,\nu) \geq \mathfrak{W}_{p'}(\mu,\nu)$, so that in particular $\mathfrak{W}_1$ is the weakest. For this reason, henceforth we will focus on $\mathfrak{W}_1$, which we will simply denote $\mathfrak{W}$.

Using the concepts that we have just introduced, we can give two simple and concrete examples of $\mathfrak{D}$-regularity.

Lemma 9 *Let $\mathfrak{D}_1 : (\mu,\nu) \mapsto \sqrt{2\mathrm{KL}(\nu\|\mu)}$ and $\mathfrak{D}_2 : (\mu,\nu) \mapsto \mathfrak{W}(\mu,\nu)$. Consider a measurable map $f : \mathcal{Z} \to \mathbb{R}^q$ (with $q \geq 1$). If $f(Z)$ is ξ-SG for $Z \sim \mu \in \mathscr{P}$, then f has regularity $\mathcal{R}_{\mathfrak{D}_1}(\xi)$ wrt μ. If f is ξ-Lipschitz on $\mathcal{Z}$, then f has regularity $\mathcal{R}_{\mathfrak{D}_2}(\xi)$, wrt any $\mu \in \mathscr{P}$ such that $f \in L^1(\mu)$.*

4.1. Standard MI and Wasserstein bounds

We state two simple generalisation bounds that were previously mentioned in the introduction. The proofs that we give leverage the abstract framework of Section 3.1. The first result (Russo and Zou, 2019; Xu and Raginsky, 2017) is an upperbound on $\mathcal{G}$ based on the mutual information between W and S.

Proposition 10 (Standard MI bound) *Let $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$. If $\ell(w,X)$ is ξ-SG, $\forall w \in \mathcal{W}$, then*

$$|\mathcal{G}| \leq \xi\sqrt{\frac{2I(W;S)}{m}}\,.$$

Proof First, since $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$, $\mathscr{L}_S(w)$ is the average of m independent ξ-SG random variables, so it is $(\xi/\sqrt{m})$-SG. In particular, with $\mathfrak{D} : (\mu, \nu) \mapsto \sqrt{2\mathrm{KL}(\nu\|\mu)}$, Lemma 9 shows that $s \mapsto \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi/\sqrt{m})$, $\forall w \in \mathcal{W}$. We conclude by Theorem 2 and Jensen's inequality. ■

The next bound is from Lopez and Jog (2018) and is close in spirit to the previous one, as again it tries to measure how much information about S is enclosed in W. However, now the MI is replaced by an expected Wasserstein distance. In order to get an explicit dependence on $1/\sqrt{m}$, we assume that the metric $d_{\mathcal{S}}$ on $\mathcal{S}$ is related to the one on $\mathcal{X}$ via

$$d_{\mathcal{S}}(s, s') = \left(\sum_{i=1}^{m} d_{\mathcal{X}}(x_i, x_i')^2\right)^{1/2}, \tag{1}$$

where $s = \{x_1, \dots, x_m\}$ and $s' = \{x_1', \dots, x_m'\}$. Note that we do not need $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$.

Proposition 11 (Standard Wasserstein bound) *Suppose that $d_{\mathcal{X}}$ and $d_{\mathcal{S}}$ are related by (1). If, $\forall w \in \mathcal{W}$, $x \mapsto \ell(w, x)$ is ξ-Lipschitz on $\mathcal{X}$, then*

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S|W})] .$$

Proof First notice that

$$d_{\mathcal{S}}(s, s') = \left(\sum_{i=1}^{m} d_{\mathcal{X}}(x_i, x_i')^2\right)^{1/2} \geq \frac{1}{\sqrt{m}} \sum_{i=1}^{m} d_{\mathcal{X}}(x_i, x_i') ,$$

where we used the Cauchy-Schwartz inequality. Consequently, $s \mapsto \mathscr{L}_s(w)$ is $(\xi/\sqrt{m})$-Lipschitz $\forall w \in \mathcal{W}$. In particular, if we let $\mathfrak{D} : (\mu, \nu) \mapsto \mathfrak{W}(\mu, \nu)$, then $s \mapsto \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi/\sqrt{m})$ by Lemma 9. We conclude by Theorem 2. ■

4.2. Chained MI and Wasserstein bounds

As we mentioned in the introduction, one of the main issues with the standard MI bound is that it can easily be vacuous, as it is the case when the learning algorithm defines a deterministic map $\mathcal{S} \to \mathcal{W}$. To address this issue, Asadi et al. (2018) proposed to build on the chaining technique and established the bound below. The setting here is quite different from the one of the standard MI bound, as the process's subgaussianity takes into account the dependencies between different hypotheses. Letting $\{\varepsilon_k\}_{k\in\mathbb{N}}$ be a vanishing decreasing positive sequence, we consider an $\{\varepsilon_k\}$-refining sequence of nets $\{\mathcal{W}_k\}_{k\in\mathbb{N}} = \{\pi_k(\mathcal{W})\}_{k\in\mathbb{N}}$ and recall that $W_k = \pi_k(W)$.

Proposition 12 (CMI bound) *Let $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$ and $\mathcal{W}$ be a compact set, with an $\{\varepsilon_k\}$-refining sequence of nets defined on it. Suppose that $w \mapsto \ell(w, x)$ is continuous, for $\mathbb{P}_X$-almost every x,*[5] *and that $\{\ell(w, X)\}_{w\in\mathcal{W}}$ is a ξ-SG stochastic process. Then we have*

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} \sqrt{2I(W_k; S)} .$$

[5] Note that in Asadi et al. (2018) the result is stated under a weaker assumption of separability of the process. To avoid introducing further definitions and technicalities in the proofs, we decided to focus on the case of a.s. continuity.

We provide a proof of Proposition 12 within the extended general framework of Appendix B.1, while here we establish a similar result, under the more restrictive assumptions ♣.

Leveraging the machinery developed in Section 3.2, we can expect that lifting the subgaussianity from ℓ to $\nabla_w \ell$ we can find a chained version of the MI bound in Proposition 10. Perhaps unsurprisingly, we simply re-obtain the CMI bound of Proposition 12.

Proposition 13 *Let $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$ and assume ♣. If $\nabla_w \ell(w, X)$ is ξ-SG, $\forall w \in \mathcal{W}$, we have that for any $\{\varepsilon_k\}$-refining sequence of nets on $\mathcal{W}$*

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} \sqrt{2I(W_k; S)}\,.$$

Proof As in the proof of Proposition 10, we have that $\nabla_w \mathscr{L}_S(w)$ is $(\xi/\sqrt{m})$-SG, $\forall w \in \mathcal{W}$. In particular, by Lemma 9 we have that $s \mapsto \nabla_w \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi/\sqrt{m})$, $\forall w \in \mathcal{W}$, where $\mathfrak{D} : (\mu, \nu) \mapsto \sqrt{2\mathrm{KL}(\nu \| \mu)}$. Hence, we conclude by Theorem 4 and Jensen's inequality. ■

The next lemma shows that, under the assumptions ♣, Propositions 12 and 13 are equivalent.

Lemma 14 *Under the assumptions ♣, the stochastic process $(\ell(w, X))_{w \in \mathcal{W}}$ is ξ-SG if, and only if, $\nabla_w \ell(w, X)$ is a ξ-SG vector for all $w \in \mathcal{W}$.*

Once again, the main point of the abstract framework presented so far is to underline a duality: to each bound based on the $\mathfrak{D}$-regularity of the loss corresponds a chained bound based on the $\mathfrak{D}$-regularity of its gradient. We can hence apply this idea to the standard Wasserstein bound of Proposition 11 and obtain its chained counterpart, which is a novel result.

Proposition 15 (Chained Wasserstein bound) *Let $d_{\mathcal{X}}$ and $d_{\mathcal{S}}$ be related by (1). Under the assumptions ♣, suppose that $x \mapsto \nabla_w \ell(w, x)$ is ξ-Lipschitz on $\mathcal{X}$, $\forall w \in \mathcal{W}$. Then, for any $\{\varepsilon_k\}$-refining sequence of nets on $\mathcal{W}$,*

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S|W_k})]\,.$$

Proof Let $\mathfrak{D} : (\mu, \nu) \mapsto \mathfrak{W}(\mu, \nu)$. Proceeding as in the proof of Proposition 11, we have that $\nabla_w \mathscr{L}_S(w)$ is $(\xi/\sqrt{m})$-Lipschitz, $\forall w \in \mathcal{W}$. In particular, by Lemma 9, $s \mapsto \nabla_w \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi/\sqrt{m})$, $\forall w \in \mathcal{W}$. Hence, we conclude by Theorem 4. ■

We conclude by recalling once more that, in our framework, any bound based on the regularity of ℓ gives rise to a chained bound. We refer to Table 1 in the appendix for several explicit examples.

5. A chained PAC-Bayesian bound

The framework introduced in Section 3 focuses on the *backward-channel* information-theoretic setting. However, the chaining ideas behind Theorem 4 can fit in a broader context. As an example, we discuss here a PAC-Bayesian result. Although Audibert and Bousquet (2004) have already combined the PAC-Bayesian approach with the chaining technique, their use of an auxiliary sample and of the average distance between nets makes their bounds conceptually different from ours.

The PAC-Bayesian bounds are algorithmic-dependent upper bounds on the expected generalisation error $\mathbb{E}_{\mathbb{P}_{W|S}}[g_S(W)]$ of stochastic classifiers (McAllester, 1998), holding with high probability on the random draw of the training sample S (see Guedj (2019) and Alquier (2021) for recent introductory overviews). They share the same underlying idea with the information-theoretic bounds: the less $\mathbb{P}_{W|S}$ is dependent on S, the better the algorithm generalises. However, in the PAC-Bayesian setting we compare $\mathbb{P}_{W|S}$ not with the marginal $\mathbb{P}_W$, but rather with a fixed probability measure $\mathbb{P}^*_W$, which can be chosen arbitrarily but without making use of the training sample S.

We state here a very simple classical PAC-Bayesian result from Catoni (2009).

Proposition 16 *Assume that ℓ is bounded in $[-\xi, \xi]$. Let $\mathbb{P}^*_W$ be a fixed probability measure on $\mathcal{W}$, chosen independently of S. Fix $\delta \in (0,1)$ and $\lambda > 0$. Then, with probability $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$ larger than $1-\delta$ on the draw of S, we have*

$$\mathbb{E}_{\mathbb{P}_{W|S}}[g_S(W)] \leq \frac{\xi}{\sqrt{2m}}\left(\lambda + \frac{\mathrm{KL}(\mathbb{P}_{W|S}\|\mathbb{P}^*_W) + \log\frac{1}{\delta}}{\lambda}\right).$$

A chained version of the above can be obtained by lifting the boundedness hypothesis from ℓ to $\nabla_w \ell$. This is quite peculiar, as most PAC-Bayesian bounds hold for bounded loss functions $\ell \subseteq [-\xi, \xi]$.

Proposition 17 *Under the assumptions $\clubsuit$, consider a $\{\varepsilon_k\}$-refining sequence of nets on $\mathcal{W}$ and assume that $\nabla_w \ell$ is bounded in $[-\xi, \xi]$. Let $\mathbb{P}^*_W$ be a fixed probability measure on $\mathcal{W}$, chosen independently of S. Fix two sequences $\{\delta_k\}_{k\in\mathbb{N}}$ and $\{\lambda_k\}_{k\in\mathbb{N}}$, such that $\delta_k \in (0,1)$ and $\lambda_k > 0$ for all k. Assume that $\sum_{k\in\mathbb{N}} \delta_k = \delta \in (0,1)$. Then, with probability $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$ larger than $1-\delta$ on the draw of S, we have*

$$\mathbb{E}_{\mathbb{P}_{W|S}}[g_S(W)] \leq \frac{\xi}{\sqrt{2m}}\left(2\sqrt{\log\frac{1}{\delta_0}} + \sum_{k=1}^{\infty}\varepsilon_{k-1}\left(\lambda_k + \frac{\mathrm{KL}(\mathbb{P}_{W_k|S}\|\mathbb{P}^*_{W_k}) + \log\frac{1}{\delta_k}}{\lambda_k}\right)\right).$$

The PAC-Bayesian bound in Proposition 16 is infinite for a deterministic algorithm (that is when $\mathbb{P}_{W|S=s}$ is a Dirac delta for all $s \in \mathcal{S}$). Remarkably, for suitable coefficients λ_k, δ_k, and ε_k, the chained bound of Proposition 17 is always finite, since all the terms $\mathrm{KL}(\mathbb{P}_{W_k|S}\|\mathbb{P}^*_{W_k})$ are bounded by $\log|\mathcal{W}_k|$. However, the best choice of the parameters λ and λ_k is delicate, as it cannot depend on S (and hence on the KL term). We refer to Appendix C for further discussion on this last point.

6. Comparison of chained and unchained bounds

Having established the duality, we are left with the Hamletic question: *chained or unchained, what is the best?* First, we notice that the requirements for the chained bounds are somewhat stronger.

Lemma 18 *Under the assumptions $\clubsuit$, let ε_0 and w_0 be such that $\|w - w_0\| \leq \varepsilon_0$, $\forall w \in \mathcal{W}$. Assume that $s \mapsto \nabla_w \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt $\mathbb{P}_S$, $\forall w \in \mathcal{W}$, and define $\hat{\mathscr{L}}_s(w) = \mathscr{L}_s(w) - \mathscr{L}_s(w_0)$ and $\hat{\mathcal{G}} = \mathbb{E}_{W\otimes S}[\hat{\mathscr{L}}_S(W)] - \mathbb{E}_{W,S}[\hat{\mathscr{L}}_S(W)]$. Then, $\hat{\mathcal{G}} = \mathcal{G}$, and $s \mapsto \hat{\mathscr{L}}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\varepsilon_0\xi)$, wrt $\mathbb{P}_S$ and $\forall w \in \mathcal{W}$.*

Hence, whenever we derive a chained bound $|\mathcal{G}| \leq \xi \sum_{k=1}^{\infty} \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})]$ in our framework, we can always state an unchained counterpart in the form $|\mathcal{G}| \leq \varepsilon_0\xi\, \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W})]$. Nevertheless, the next result shows that conditioning on W_k instead of W can often be helpful.

Lemma 19 *Assume that $\mu \mapsto \mathfrak{D}(\mathbb{P}_S, \mu)$ is convex. For any $\{\varepsilon_k\}$-refining sequence of nets on $\mathcal{W}$, the sequence $\{\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})]\}_{k\in\mathbb{N}}$ is non-decreasing and, $\forall k \in \mathbb{N}$, we have*

$$\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] \le \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W})] \,.$$

$\mathrm{KL}(\nu\|\mu)$ is convex in both ν and μ (Erven and Harremoës, 2014), and the same holds for $\mathfrak{W}(\mu,\nu)$ (Villani, 2009). Thus, $I(W_k; S) \le I(W; S)$[6] and $\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] \le \mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S|W})]$.

Lemma 19 alone is not enough to ensure that the chained bound is tighter than its unchained counterpart. However, if $\mathbb{P}_W$ is very concentrated on a tiny region of $\mathcal{W}$, so that S is almost independent of W_k up to a small scale (*i.e.*, large k), then one can expect the chained result to be the tightest. We will clarify this intuition by means of two simple toy examples. Since Asadi et al. (2018) have already shown that the CMI bound can be much tighter than the MI one, here the focus is on the Wasserstein bounds.

6.1. Comparison of the chained and unchained Wasserstein bounds

In the following we denote by $\mathcal{B}_\ell$ the standard Wasserstein bound (Proposition 11) and by $\mathcal{B}_{\nabla\ell}$ its chained counterpart (Proposition 15). For simplicity, we mainly focus on the case $m = 1$, so that we can write $s = x$ and $\mathcal{G} = \mathbb{E}_{\mathbb{P}_{W\otimes X}}[\ell(W,X)] - \mathbb{E}_{\mathbb{P}_{W,X}}[\ell(W,X)]$.

Example 1 Let $\mathcal{W} = \mathcal{X} = [-1,1]$, $\ell(w,x) = \frac{1}{2}(w-x)^2$, and $\varepsilon_k = 2^{-k}$, for $k \in \mathbb{N}$. We can find mappings π_k that define an $\{\varepsilon_k\}$-refining sequence of nets, with $\mathcal{W}_k = \{2^{1-k}j \,:\, j \in [-2^{k-1} : 2^{k-1}]\}$, where $[a:b] = [a,b]\cap\mathbb{Z}$. Fix $k^\star \in \mathbb{N}$ and define $\theta = 2^{-k^\star}$. Let X be uniformly distributed on $(-\theta,\theta)$. We choose an algorithm that, given x, selects the w minimising $\ell(w,x)$. This means that $\mathbb{P}_{W|X=x} = \delta_x$, where δ_x is the Dirac measure centred on x. Note that $\nabla_w\ell$ is 1-Lipschitz and ℓ is 2-Lipschitz (on $\mathcal{X}$, uniformly on $\mathcal{W}$). However, thanks to Lemma 18 we know that we can consider the loss $\tilde{\ell}(w,x) = \ell(w,x) - \frac{x^2}{2}$, which leads to the same generalisation and is 1-Lipschitz.

In this simple example, we can compute exactly everything we need (see Appendix E.1):

$$|\mathcal{G}| = \frac{1}{3}\,\theta^2 \simeq 0.33\,\theta^2\,; \qquad \frac{1}{2}\,\mathcal{B}_\ell = \mathcal{B}_{\tilde{\ell}} = \frac{2}{3}\,\theta \simeq 0.67\,\theta\,; \qquad \mathcal{B}_{\nabla\ell} = \frac{247}{105}\,\theta^2 \simeq 2.35\,\theta^2\,.$$

Note that, as θ decreases, $\mathbb{P}_W$ becomes more and more concentrated, since W lies with probability 1 in $(-\theta,\theta)$. In particular, X and W_k are independent for $k \le k^\star = -\log_2\theta$, and so the first $k^\star$ terms in the chaining sum are null. For this reason, $\mathcal{B}_{\nabla\ell}$ captures the right behaviour $O(\theta^2)$ of $\mathcal{G}$ for $\theta \to 0$, which is not the case for $\mathcal{B}_\ell$ and $\mathcal{B}_{\tilde{\ell}}$.

Quite interestingly, it is possible to explicitly evaluate the CMI bound ($\mathcal{B}_{\mathrm{CMI}}$) as well. We find $\mathcal{B}_{\mathrm{CMI}} \simeq 3.50\,\theta$, meaning that in this example the chained MI bound fails to capture the right behaviour of $\mathcal{G}$ as $\theta \to 0$. We refer to Section 7 for a few comments about this. On the other hand, the unchained MI bound is infinite, since W is a deterministic function of X.

Finally, if we consider a larger random sample $S = \{X_1, \dots, X_m\}$, with $m > 1$, we still have that the ratio $\mathcal{B}_{\nabla\mathscr{L}}/\mathcal{B}_{\mathscr{L}}$ (between the chained and unchained Wasserstein bounds) vanishes as $O(\theta)$ for $\theta \to 0$. Again, this is a consequence of the fact that S and W_k are independent for $k \le k^\star$, since W is the empirical average $\sum_i X_i/m$ and lies in $(-\theta,\theta)$ with probability 1.

[6] This can also be seen as a trivial consequence of the data-processing inequality.

Example 2 This toy model is inspired by Example 1 in Asadi et al. (2018). Let $\mathcal{W} = \{w \in \mathbb{R}^2 : \|w\| = 1\}$[7] and $\mathcal{X} = \mathbb{R}^2$. Fix $a > 0$ and let $X \sim \mathcal{N}((a, 0), \mathrm{Id})$, a normal distribution centered in $(a, 0)$, with the identity matrix as covariance. The algorithm aims at finding the direction of the mean of X (that is $(1, 0)$), by choosing the w that minimises the loss $\ell(w, x) = -w \cdot x$. Let $w_0 = (1, 0)$ and $\varepsilon_0 = 4$. For $k \geq 1$, let $\mathcal{W}_k = \{w = (\cos\frac{2\pi j}{2^k}, \sin\frac{2\pi j}{2^k}) : j \in [-2^{k-1} : 2^{k-1} - 1]\}$ and $\varepsilon_k = 4/2^k$. We can then easily find projections π_k that make $\{\mathcal{W}_k\}_{k\in\mathbb{N}}$ an $\{\varepsilon_k\}$-sequence of refining nets. Both ℓ and $\nabla_w \ell$ are 1-Lipschitz in $\mathcal{X}$, $\forall w \in \mathcal{W}$. Although it is hard to find the analytic expressions for $\mathcal{B}_\ell$ and $\mathcal{B}_{\nabla\ell}$, we can study their asymptotic behaviour for $a \to \infty$. In this limit, $\mathbb{P}_W$ becomes highly concentrated around $(0, 1)$, as it tends towards a Dirac delta. So, for a large enough we expect the chained bound to be the tightest. Indeed, we find

$$|\mathcal{G}| = \Theta(1/a)\,; \qquad \mathcal{B}_\ell = \Theta(1)\,; \qquad \mathcal{B}_{\nabla\ell} = O((\log a - \log\log a)/a)\,.$$ [8]

Up to logarithmic factors, $\mathcal{B}_{\nabla\ell}$ can capture the correct behaviour of $|\mathcal{G}|$ as $a \to \infty$.

As a final remark, note that in this example the loss ℓ is not Lipschitz on $\mathcal{W}$, uniformly on $\mathcal{X}$, and so the *forward-channel* Wasserstein bound from Wang et al. (2019) does not apply.[9]

6.2. High concentration is not always enough

In both the previous examples, the chained bound was much tighter than its unchained counterpart when $\mathbb{P}_W$ was highly concentrated in a small neighbourhood U of a single point $w_\star$. In particular, if 2ε is the diameter of U, we can expect that just knowing that $W \in U$ is not informative up to a length-scale of order ε. However, this can easily fail when W concentrates around two far apart points (say w_1 and w_2). Indeed, if for small k we already have that $\pi_k(w_1) \neq \pi_k(w_2)$, knowing that the chosen hypothesis is next to w_1 might bring a lot of information about S. On the other hand, one can still imagine situations in which there are multiple points around which W concentrates, yet which one is the nearest to the chosen hypothesis is not informative about the sample.

In Appendix E.1.1, we discuss a high-dimensional version of Example 1, where W does not concentrate around a single point, but in a thin neighbourhood of a one-dimensional line. We show that when θ (the parameter describing the size of the support of W) has the right scaling with the dimension d of $\mathcal{W}$, the ratio $\mathcal{B}_{\nabla\ell}/\mathcal{B}_\ell$ vanishes as $d \to \infty$.

7. Comparison between MI and Wasserstein bounds

We conclude this paper with a few comments on the relation between the MI-based (Propositions 10 and 13) and the Wasserstein-based bounds (Propositions 11 and 15). The problem comes down to comparing the KL divergence with the 1-Wasserstein distance, a task closely related to transportation-cost inequalities (see Raginsky and Sason (2013) for a pedagogical overview). Let μ be a probability measure on the Polish space $(\mathcal{Z}, \Sigma_\mathcal{Z})$. For $\eta > 0$, μ is said to satisfy a L^1 transport-cost inequality with constant η (in short $\mu \in T_1(\eta)$) if, for any $\nu \ll \mu$, $\mathfrak{W}(\mu, \nu) \leq \sqrt{2\eta\,\mathrm{KL}(\nu\|\mu)}$. Hence, whenever $\mathbb{P}_S \in T_1(1)$ we are assured that each one of the two Wasserstein-based bounds is

[7] This is not a convex set. However, one can either suitably extend ℓ to the unit disk in $\mathbb{R}^2$, or easily check that the hypotheses of the extended framework of Theorem 22 in Appendix B.1 are verified (see Section E.2).

[8] Here $f = O(g)$ stands for $\lim_{a\to\infty} |f(a)/g(a)| < \infty$, while by $f = \Theta(g)$ we mean that $f = O(g)$ and $g = O(f)$.

[9] Of course, $\ell(w, x) = -w \cdot x/\|x\|$ would bring the same algorithm and is 1-Lipschitz in w. However this is just due to the radial symmetry. Changing the problem slightly and considering for instance $\ell(w, x) = -w \cdot \psi(x)$, for a general 1-Lipschitz map $\psi : \mathcal{X} \to \mathcal{X}$, would not allow to easily find an equivalent loss that is 1-Lipschitz in w.

tighter than the corresponding MI-based one. For instance, this is the case when $\mathbb{P}_S$ is a multivariate normal whose covariance matrix is the identity (Talagrand, 1996), as in Example 2. However, there is a price to pay: whenever the L^1 transport-inequality holds, then Lipschitzianity is stronger than subgaussianity. More precisely, Bobkov and Götze (1999) showed that $\mu \in T_1(1)$ if, and only if, for every ξ-Lipschitz function $f : \mathcal{Z} \to \mathbb{R}$, $f(Z)$ is ξ-subgaussian for $Z \sim \mu$.

It is worth noticing that, if the size of the support of X is particularly small, the Wasserstein bounds can be much tighter than the MI ones. This is for instance the case in Example 1, where the length-scale of the support of X is given by θ. There, the chained Wasserstein bound goes as θ^2. A factor θ is brought by the chaining technique, which allows us to neglect the contributions of the larger length-scales, whilst the other factor θ is due to the use of the Wasserstein distance, which intrinsically takes into account the considered length-scale. In contrast, since the MI is scale-invariant, the CMI bound has only a linear dependence in θ coming from the chaining method.

7.1. Scaling with the sample size

It is worth mentioning the different roles that the factor $1/\sqrt{m}$ plays in the MI and the Wasserstein bounds. In the MI bound this scaling is linked to concentration properties, since it comes from the fact that the average of m independent ξ-SG random variables is $(\xi/\sqrt{m})$-SG. The requirement that S is made of independent draws is hence essential in this case. On the other hand, in the Wasserstein bound the factor $1/\sqrt{m}$ has a merely geometric origin and follows from the relation (1) between the metrics $d_{\mathcal{X}}$ and $d_{\mathcal{S}}$. In particular, an alternative choice of $d_{\mathcal{S}}$ might yield a different factor in front of the bound, but also change the scaling with m of the Wasserstein distance. A priori, it is not easy to say which $d_{\mathcal{S}}$ would bring the tightest bound. Once more, let us stress that the Wasserstein bound does not require that $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$. Indeed, $\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S|W})$ will take into account the dependencies between the training inputs, and we can expect it to scale poorly with m if the different X_i in S are strongly correlated. However, even in the case of independent X_i, it is hard to say in general what is the exact dependence with m, for both $I(W;S)$ and $\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S|W})$.

As a final remark about the case $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$, just by looking at $\mathbb{P}_X$ it is sometimes possible to establish that both the standard and chained Wasserstein bounds are tighter than their MI counterparts, no matter the size of the training dataset and the choice of the algorithm. To this purpose, we can again exploit some classical results on the transport-cost inequalities (Raginsky and Sason, 2013; Gozlan and Léonard, 2010). For a probability measure μ, we say that $\mu \in T_2(1)$ if, for any $\nu \ll \mu$, $\mathfrak{W}_2(\mu, \nu) \leq \sqrt{2\mathrm{KL}(\nu\|\mu)}$. It is known that $\mu \in T_2(1)$ implies that $\mu^{\otimes m} \in T_2(1)$, $\forall m \geq 1$. In particular, if $\mathbb{P}_X \in T_2(1)$, then we are ensured that $\mathbb{P}_S = \mathbb{P}_X^{\otimes m} \in T_2(1)$. Since $\mathfrak{W} = \mathfrak{W}_1 \leq \mathfrak{W}_2$, $\mathbb{P}_X \in T_2(1)$ actually implies $\mathbb{P}_S \in T_1(1)$, which (as we discussed the beginning of this section) means that each Wasserstein-based bound is tighter than the corresponding MI-based one.

8. Conclusion

We introduced a general framework allowing us to derive new generalisation results leveraging on the chaining technique. By doing so, under suitable regularity conditions we established a duality between chained and unchained generalisation bounds. Although the chained bounds usually come at the price of stricter assumptions, sometimes they better capture the loss function's behaviour, especially in cases where the hypothesis distribution is highly concentrated. We hence believe that combining the chaining method with other information-theoretic techniques is a promising direction in order to tighten the bounds on the generalisation error.

In this work we have mainly focused on the *backward-channel* information-theoretic perspective, as we believe that it combines naturally with the chaining on the hypotheses' space. However, the chained PAC-Bayesian result that we presented is an example of a *forward-channel* bound, as it considers the distribution of the hypotheses, conditioned on the sample. A future direction of study could be to extend our general framework to include *forward-channel* bounds. We believe this should not present major technical difficulties and might bring new interesting results.

Although information-theoretic bounds are usually hard to evaluate in practice, recent works have derived computable analytic bounds for specific algorithms, such as Langevin dynamics or stochastic gradient descent, by upper-bounding information-theoretic generalisation results. We believe that combining these ideas with the chaining technique is a venue worth exploring.

Acknowledgments

Eugenio Clerico is partly supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the grant EP/R513295/1 (DTP scheme). Arnaud Doucet is partly supported by the EPSRC grant EP/R034710/1. He also acknowledges the support of the UK Defence Science and Technology Laboratory (DSTL) and EPSRC under grant EP/R013616/1. This is part of the collaboration between US DOD, UK MOD, and UK EPSRC, under the Multidisciplinary University Research Initiative.

References

P. Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv:2110.11216*, 2021.

G. Aminian, L. Toni, and M. R. D. Rodrigues. Information-theoretic bounds on the moments of the generalization error of learning algorithms. *arXiv:2102.02016*, 2021.

M. Anthony and P. L. Bartlett. *Neural Network Learning - Theoretical Foundations*. Cambridge University Press, 2002.

A. R. Asadi and E. Abbe. Chaining meets chain rule: Multilevel entropic regularization and training of neural nets. *JMLR*, 21, 2020.

A. R. Asadi, E. Abbe, and S. Verdú. Chaining mutual information and tightening generalization bounds. *NeurIPS*, 2018.

J-Y. Audibert and O. Bousquet. PAC-Bayesian generic chaining. *NeurIPS*, 2004.

M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. *ICML*, 2018.

S. G. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1), 1999.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2, 2002.

O. Bousquet, S. Boucheron, and G. Lugosi. *Introduction to Statistical Learning Theory*. Springer, 2004.

Y. Bu, S. Zou, and V. V. Veeravalli. Tightening mutual information based bounds on generalization error. *ISIT*, 2019.

O. Catoni. A PAC-Bayesian approach to adaptive classification. *Preprint LPMA*, 840, 2009.

E. A. Cooper and H. Farid. A toolbox for the radial and angular marginalization of bivariate normal distributions. *arXiv:2005.09696*, 2020.

M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2), 1983.

R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3), 1967.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9, 2014.

T. Erven and P. Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7), 2014.

A. R. Esposito, M. Gastpar, and I. Issa. Generalization error bounds via Rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(3), 2021.

N. Gozlan and C. Léonard. Transport inequalities. A survey. *Markov Processes and Related Fields*, 16, 2010.

B. Guedj. A primer on PAC-Bayesian learning. *Proceedings of the Second Congress of the French Mathematical Society*, 2019.

A. Guntuboyina, S. Saha, and G. Schiebinger. Sharp inequalities for f -divergences. *IEEE Transactions on Information Theory*, 60(1), 2014.

M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *NeurIPS*, 2020.

F. Hellström and G. Durisi. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3), 2020.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301), 1963.

A. S. Kechris. *Classical Descriptive Set Theory*. Springer-Verlag, 1995.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, New York, NY, USA, second edition, 1998.

M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211, 2018.

A. T. Lopez and V. Jog. Generalization error bounds using Wasserstein distances. *IEEE Information Theory Workshop (ITW)*, 2018.

G. Marsaglia, B. Narasimhan, and A. Zaman. The distance between random points in rectangles. *Communications in Statistics - Theory and Methods*, 19, 1990.

D. A. McAllester. Some PAC-Bayesian theorems. *COLT*, 1998.

D. A. McAllester. PAC-Bayesian model averaging. *COLT*, 1999.

J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. *NeurIPS*, 2019.

G. Neu, G. K. Dziugaite, M. Haghifam, and D. M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. *COLT*, 2021.

D. P. Palomar and S. Verdú. Lautum information. *IEEE Transactions on Information Theory*, 54 (3), 2008.

M. Raginsky and I. Sason. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10 (1-2), 2013.

B. Rodríguez-Gálvez, G. Bassi, R. Thobaben, and M. Skoglund. On random subset generalization error bounds and the stochastic gradient Langevin dynamics algorithm. *IEEE Information Theory Workshop (ITW)*, 2020.

B. Rodríguez-Gálvez, G. Bassi, R. Thobaben, and M. Skoglund. Tighter expected generalization error bounds via Wasserstein distance. *arXiv:2101.09315*, 2021.

D. Russo and J. Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1), 2019.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014.

T. Steinke and L. Zakynthinou. Reasoning about generalization via conditional mutual information. *COLT*, 2020.

M. Talagrand. Transportation cost for Gaussian and other product measures. *Geometric and Functional Analysis*, 6(3), 1996.

M. Talagrand. *The Generic Chaining: Upper and Lower Bounds of Stochastic Processes*. Springer-Verlag, 2005.

M. Talagrand. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*. Springer-Verlag, 2014.

R. van Handel. *Probability in High Dimensions*, 2016. URL https://web.math.princeton.edu/~rvan/APC550.pdf. [Online; accessed on February 2022].

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

C. Villani. *Optimal Transport – Old and New*. Springer, 2009.

H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon. An information-theoretic view of generalization via Wasserstein distance. *ISIT*, 2019.

A. D. Wyner. A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, 38(1), 1978.

A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *NeurIPS*, 2017.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3), 2021.

R. Zhou, C. Tian, and T. Liu. Stochastic chaining and strengthened information-theoretic generalization bounds. *arXiv:2201.12192*, 2022.

Appendix A. Omitted proofs of Sections 3 and 4

Here $(\mathcal{Z}, d_{\mathcal{Z}})$ is a separable complete metric space, with Borel σ-algebra $\Sigma_{\mathcal{Z}}$ induced by the metric. $\mathcal{W} \times \mathcal{Z}$ is endowed with the product σ-algebra $\Sigma_{\mathcal{W}} \otimes \Sigma_{\mathcal{Z}}$. $\mathscr{P}$ denotes the space of probability measures on $\mathcal{Z}$ and is endowed with the σ-algebra induced by the topology of weak convergence.

A.1. Proof of Lemma 9

Lemma 9 *Let $\mathfrak{D}_1 : (\mu, \nu) \mapsto \sqrt{2\mathrm{KL}(\nu\|\mu)}$ and $\mathfrak{D}_2 : (\mu, \nu) \mapsto \mathfrak{W}(\mu, \nu)$. Consider a measurable map $f : \mathcal{Z} \to \mathbb{R}^q$ (with $q \geq 1$). If $f(Z)$ is ξ-SG for $Z \sim \mu \in \mathscr{P}$, then f has regularity $\mathcal{R}_{\mathfrak{D}_1}(\xi)$ wrt μ. If f is ξ-Lipschitz on $\mathcal{Z}$, then f has regularity $\mathcal{R}_{\mathfrak{D}_2}(\xi)$, wrt any $\mu \in \mathscr{P}$ such that $f \in L^1(\mu)$.*

Proof First, notice that Lemmas 28 and 29 ensure that both $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are measurable, as required by Definition 1.

Assume that $f(Z)$ is ξ-SG for $Z \sim \mu$. Then, by definition $f \in L^1(\mu)$. Fix ν such that $f \in L^1(\nu)$ and $\mathrm{Supp}(\nu) \subseteq \mathrm{Supp}(\mu)$. If $q = 1$, the Donsker-Varadhan representation of KL (Donsker and Varadhan, 1983) and subgaussianity yield

$$\mathrm{KL}(\nu\|\mu) \geq \sup_{\lambda \in \mathbb{R}} \lambda(\mathbb{E}_\nu[f(Z)] - \mathbb{E}_\mu[f(Z)]) - \lambda^2\xi^2/2 = \frac{(\mathbb{E}_\mu[f(Z)] - \mathbb{E}_\nu[f(Z)])^2}{2\xi^2},$$

from which the $\mathfrak{D}_1$-regularity of f follows immediately. The case of a generic $q > 1$ is trivial, since $v \cdot f(Z)$ is $(\xi\|v\|)$-SG by Definition 5, for all $v \in \mathbb{R}^q$.

Now, let $f \in L^1(\mu)$ be ξ-Lipschitz. If $q = 1$, let π be any coupling with marginals μ and ν. We have that

$$|\mathbb{E}_\mu[f(Z)] - \mathbb{E}_\nu[f(Z)]| = |\mathbb{E}_{(Z,Z')\sim\pi}[f(Z) - f(Z')]| \leq \xi\, \mathbb{E}_{(Z,Z')\sim\pi}[d(Z, Z')]\,.$$

The $\mathfrak{D}_2$-regularity can be established by taking the inf among all the couplings π with marginals μ and ν. The case $q > 1$ follows from the fact that $z \mapsto v \cdot f(z)$ is $(\xi\|v\|)$-Lipschitz. ■

A.2. Proof of Theorem 2

Theorem 2 is equivalent to the following result.

Theorem 20 *Consider a measurable map $F : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$, such that $z \mapsto F(w, z)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt $\mathbb{P}_Z$ and for all $w \in \mathcal{W}$. Then we have*

$$|\mathbb{E}_{\mathbb{P}_{W\otimes Z}}[F(W, Z)] - \mathbb{E}_{\mathbb{P}_{W,Z}}[F(W, Z)]| \leq \xi\, \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_Z, \mathbb{P}_{Z|W})]\,.$$

Proof First, note that $\mathrm{Supp}(\mathbb{P}_{Z|W=w}) \subseteq \mathrm{Supp}(\mathbb{P}_Z)$ by Lemma 30 and $\mathbb{E}_{\mathbb{P}_{Z|W=w}}[|F(w, Z)|] < \infty$, $\mathbb{P}_W$-a.s, since $\mathbb{E}_{\mathbb{P}_{W,Z}}[|F(W, Z)|] < +\infty$. In particular, for $\mathbb{P}_W$-almost every $w \in \mathcal{W}$ we have that

$$|\mathbb{E}_{\mathbb{P}_Z}[F(w, Z)] - \mathbb{E}_{\mathbb{P}_{Z|W=w}}[F(w, Z)]| \leq \xi\, \mathfrak{D}(\mathbb{P}_Z, \mathbb{P}_{Z|W=w})\,.$$

Then the conclusion follows by taking the expectation wrt $\mathbb{P}_W$ and using Jensen's inequality. ■

A.3. Proof of Theorem 4

Theorem 4 follows from the next result, which is a direct corollary of Theorem 22 and Lemma 23, proved in Appendix B.1.

Theorem 21 *Let $\mathcal{W}$ be a compact convex subset of $\mathbb{R}^d$ with non-empty interior. Consider a measurable map $F : \mathcal{W}\times\mathcal{Z} \to \mathbb{R}$, such that $w \mapsto F(w,z)$ is C^1, $\mathbb{P}_Z$-a.s. Assume that $\sup_{(w,z)\in\mathcal{W}\times\mathcal{X}} |F(w,z)| < +\infty$ and $\sup_{(w,z)\in\mathcal{W}\times\mathcal{X}} |\nabla_w F(w,z)| < +\infty$. If $z \mapsto \nabla_w F(w,z)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt $\mathbb{P}_Z$, $\forall w \in \mathcal{W}$, then we have that for any $\{\varepsilon_k\}$-refining sequence of nets on $\mathcal{W}$*

$$|\mathbb{E}_{\mathbb{P}_{W\otimes Z}}[F(W,Z)] - \mathbb{E}_{\mathbb{P}_{W,Z}}[F(W,Z)]| \leq \xi \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_Z, \mathbb{P}_{Z|W_k})]\,.$$

Proof By Lemma 23, the regularity of $\nabla_w F$ implies that the map $z \mapsto (F(w,z) - F(w',z))$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi\|w - w'\|)$, wrt $\mathbb{P}_Z$ and for all $w, w' \in \mathcal{W}$. We conclude by Theorem 22. ■

A.4. Proof of Lemma 14

Lemma 14 *Under the assumptions ♣, the stochastic process $(\ell(w,X))_{w\in\mathcal{W}}$ is ξ-SG if, and only if, $\nabla_w \ell(w,X)$ is a ξ-SG vector for all $w \in \mathcal{W}$.*

Proof First, notice that, without loss of generality, we can consider the case of a one-dimensional $\mathcal{W} \subseteq \mathbb{R}$. Indeed, if $\mathcal{W}$ is higher dimensional, for any two given points w and w', we can always restrict to a line connecting them, making the problem 1D. Moreover, letting $\bar{\ell}(w,x) = \ell(w,x) - \mathbb{E}_{\mathbb{P}_X}[\ell(w,X)]$ we have that the assumptions in ♣ ensure that $\nabla_w \bar{\ell} = \nabla_w \ell - \mathbb{E}_{\mathbb{P}_X}[\nabla_w \ell]$. So, we just need to show that the lemma holds for $\bar{\ell}$.

Now, let $\bar{\ell}$ be a ξ-SG process, so that for $\varepsilon \neq 0$ and $\lambda \in \mathbb{R}$

$$\mathbb{E}_{\mathbb{P}_X}[e^{\lambda(\bar{\ell}(w+\varepsilon,X)-\bar{\ell}(w,X))/\varepsilon}] \leq e^{\frac{\lambda^2}{2\varepsilon^2}\xi^2\varepsilon^2} = e^{\frac{\lambda^2\xi^2}{2}}\,.$$

In particular, by Fatou's lemma we have

$$\mathbb{E}_{\mathbb{P}_X}[e^{\lambda\partial_w\bar{\ell}(w,X)}] = \mathbb{E}_{\mathbb{P}_X}\left[\lim_{\varepsilon\to 0} e^{\lambda\frac{\bar{\ell}(w+\varepsilon,X)-\bar{\ell}(w,X)}{\varepsilon}}\right] \leq \liminf_{\varepsilon\to 0} \mathbb{E}_{\mathbb{P}_X}\left[e^{\lambda\frac{\bar{\ell}(w+\varepsilon,X)-\bar{\ell}(w,X)}{\varepsilon}}\right] \leq e^{\frac{\lambda^2\xi^2}{2}}\,.$$

For the reverse implication, assume that $\partial_w\bar{\ell}(w,X)$ is ξ-SG for all $w \in \mathcal{W}$. Fix $w, w' \in \mathcal{W}$. By the assumptions ♣ we have that, $\mathbb{P}_X$-a.s.

$$\bar{\ell}(w',x) - \bar{\ell}(w,x) = \int_w^{w'} \partial_w\bar{\ell}(u,x)\mathrm{d}u\,.$$

Fix a positive integer N and let $u_j = w + j(w'-w)/N$. We have

$$\mathbb{E}_{\mathbb{P}_X}\left[e^{\lambda\frac{w'-w}{N}\sum_{j=1}^N \partial_w\bar{\ell}(u_j,X)}\right] = \mathbb{E}_{\mathbb{P}_X}\left[\prod_{j=1}^N e^{\lambda\frac{w'-w}{N}\partial_w\bar{\ell}(u_j,X)}\right]$$
$$\leq \prod_{j=1}^N \mathbb{E}_{\mathbb{P}_X}[e^{\lambda(w'-w)\partial_w\bar{\ell}(u_j,X)}]^{1/N} \leq e^{(\lambda^2\xi^2(w-w')^2)/2}\,.$$

Now let $Y_N(x) = \frac{w'-w}{N}\sum_{j=1}^N \partial_w \bar{\ell}(u_j, x)$. Since $w \mapsto \ell(w,x)$ is C^1 ($\mathbb{P}_X$-a.s.) by ♣, we have $\mathbb{P}_X$-a.s. that

$$\lim_{N\to\infty} Y_N(x) = \int_w^{w'} \partial_w \bar{\ell}(u,x)\mathrm{d}u = \ell(w',x) - \ell(w,x)\,.$$

We conclude that

$$\lim_{N\to\infty} \mathbb{E}_{\mathbb{P}_X}\left[e^{\lambda \frac{w'-w}{N}\sum_{j=1}^N \partial_w \bar{\ell}(u_j,x)}\right] = \mathbb{E}_{\mathbb{P}_X}\left[e^{\lambda(\ell(w',x)-\ell(w,x))}\right]\,,$$

since by ♣ $\partial_w \bar{\ell}$ is bounded. ■

Appendix B. Extended general framework

B.1. Weakening the assumptions for the chained bounds

The framework that we presented in the main text required the assumptions ♣ for ℓ (or F in the setting of Theorem 21) for the chained bound. Actually a result equivalent to Theorem 4 can be obtained with weaker assumptions, namely just requiring almost sure continuity and boundedness in expectation for ℓ, and dropping the convexity hypothesis for $\mathcal{W}$.

Theorem 22 *Let $\mathcal{W}$ be a compact subset of $\mathbb{R}^d$ and $\{\mathcal{W}_k\}$ a $\{\varepsilon_k\}$-refining sequence of nets on $\mathcal{W}$. Consider a measurable map $F : \mathcal{W}\times\mathcal{Z} \to \mathbb{R}$, such that $w \mapsto F(w,z)$ is continuous on $\mathcal{W}$, $\mathbb{P}_Z$-a.s., and $\mathbb{E}_{\mathbb{P}_Z}[\sup_{w\in\mathcal{W}} |F(w,Z)|] < +\infty$. Moreover, assume that the function $z \mapsto F(w,z) - F(w',z)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi\|w-w'\|)$ wrt $\mathbb{P}_Z$, for every $(w,w') \in \mathcal{W}^2$. Then, we have*

$$|\mathbb{E}_{\mathbb{P}_{W,Z}}[F(W,Z)] - \mathbb{E}_{\mathbb{P}_{W\otimes Z}}[F(W,Z)]| \leq \xi \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_Z, \mathbb{P}_{Z|W_k})]\,,$$

where $\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_Z, \mathbb{P}_{Z|W_k})] = \int_{\mathcal{W}} \mathfrak{D}(\mathbb{P}_Z, \mathbb{P}_{Z|W\in\pi_k^{-1}(w)})\mathrm{d}\mathbb{P}_W(w)$.

Proof First notice that $w \mapsto F(w,z)$ is uniformly continuous on $\mathcal{W}$, $\mathbb{P}_Z$-a.s., since $\mathcal{W}$ is compact. It follows that $z \mapsto \sup_{w\in\mathcal{W}} |F(w,z) - F(w_k,z)| \to 0$, $\mathbb{P}_Z$-a.s., and so, using the fact that this map is dominated by $z \mapsto 2\sup_{w\in\mathcal{W}} |F(w,z)|$, which is in $L^1(\mathbb{P}_Z)$ by hypothesis, we get that

$$\mathbb{E}_{\mathbb{P}_Z}\left[\sup_{w\in\mathcal{W}} |F(w,Z) - F(w_k,Z)|\right] \to 0\,,$$

as $k \to +\infty$, by dominated convergence. In particular, $\mathbb{E}_{\mathbb{P}_{W,Z}}[|F(W,Z) - F(W_k,Z)|] \to 0$ and $\mathbb{E}_{\mathbb{P}_{W\otimes Z}}[|F(W,Z) - F(W_k,Z)|] \to 0$. Moreover, recalling that $\mathcal{W}_0 = \{w_0\}$ we see that $\mathbb{E}_{\mathbb{P}_{W\otimes Z}}[F(W_0,Z)] - \mathbb{E}_{\mathbb{P}_{W,Z}}[F(W_0,Z)] = 0$. It follows that

$$\begin{aligned}
&\big|\mathbb{E}_{\mathbb{P}_{W\otimes Z}}[F(W,Z)] - \mathbb{E}_{\mathbb{P}_{W,Z}}[F(W,Z)]\big| \\
&\qquad\leq \sum_{k=1}^{\infty} \big|\mathbb{E}_{\mathbb{P}_{W\otimes Z}}[F(W_k,Z) - F(W_{k-1},Z)] - \mathbb{E}_{\mathbb{P}_{W,Z}}[F(W_k,Z) - F(W_{k-1},Z)]\big| \\
&\qquad= \sum_{k=1}^{\infty} \Big|\mathbb{E}_{\mathbb{P}_{W_k\otimes Z}}[F(W_k,Z) - F(W_{k-1},Z)] - \mathbb{E}_{\mathbb{P}_{W_k,Z}}[F(W_k,Z) - F(W_{k-1},Z)]\Big|\,.
\end{aligned} \tag{2}$$

Now, notice that $\mathrm{Supp}(\mathbb{P}_{Z|W_k=w_k}) \subseteq \mathrm{Supp}(\mathbb{P}_Z)$ $\mathbb{P}_W$-a.s. by Lemma 30. Moreover, by the fact that $\mathbb{E}_{\mathbb{P}_Z}[\sup_{w\in\mathcal{W}} |F(w,Z)|] < +\infty$ we have $\mathbb{E}_{\mathbb{P}_{Z|W_k=w_k}}[\sup_{w\in\mathcal{W}} |F(w,Z)|] < +\infty$, and so in particular $\mathbb{E}_{\mathbb{P}_{Z|W_k=w_k}}[|F(w_k,Z) - F(w_{k-1},Z)|] < +\infty$, for $\mathbb{P}_{W_k}$-almost every w_k. Thus, using the regularity of F we find that

$$\begin{aligned}\big|\mathbb{E}_{\mathbb{P}_Z}[F(w_k,Z) - F(w_{k-1},Z)] - \mathbb{E}_{\mathbb{P}_{Z|W_k=w_k}}[F(w_k,Z) - F(w_{k-1},Z)]\big| \\ \leq \xi\|w_k - w_{k-1}\|\mathfrak{D}(\mathbb{P}_Z, \mathbb{P}_{Z|W_k=w_k})\,,\end{aligned} \tag{3}$$

for $\mathbb{P}_{W_k}$-almost every w_k. We can hence conclude by taking the expectation wrt $\mathbb{P}_W$ and using Jensen's inequality. ■

It is easy to see that the current framework includes the one in the main text.

Lemma 23 *Let $\mathcal{W} \subseteq \mathbb{R}^d$ be a convex set. Consider a measurable map $F : \mathcal{W}\times\mathcal{Z} \to \mathbb{R}$ with the following properties: $w \mapsto F(w,z)$ is C^1 $\mathbb{P}_Z$-a.s., $\sup_{(w,z)\in\mathcal{W}\times\mathcal{Z}} |F(w,z)| < +\infty$, and $\sup_{(w,z)\in\mathcal{W}\times\mathcal{Z}} \|\nabla_w F(w,z)\| < +\infty$. If $z \mapsto \nabla_w F(w,z)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt $\mathbb{P}_Z$, $\forall w \in \mathcal{W}$, then $z \mapsto (F(w,z) - F(w',z))$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi\|w-w'\|)$ (wrt $\mathbb{P}_Z$ and $\forall w, w' \in \mathcal{W}$).*

Proof Fix a probability $\hat{\mathbb{P}}_Z$ on $\mathcal{Z}$ such that $\mathrm{Supp}(\hat{\mathbb{P}}_Z) \subseteq \mathrm{Supp}(\mathbb{P}_Z)$. Now, notice that since $\mathcal{W}$ is convex, and F is C^1, for $\mathbb{P}_Z$-almost every z we have

$$F(w,z) - F(w',z) = \int_0^1 \nabla_w F(w_t,z)\cdot(w-w')\,\mathrm{d}t\,,$$

where $w_t = w' + t(w-w')$. Since F is uniformly bounded, we can use Fubini-Tonelli's theorem and Jensen's inequality to write

$$\begin{aligned}|\mathbb{E}_{\mathbb{P}_Z}[F(w,Z)] - &\mathbb{E}_{\hat{\mathbb{P}}_Z}[F(w',Z)]| \\ &\leq \int_0^1 \Big|\mathbb{E}_{\mathbb{P}_Z}[\nabla_w F(w_t,Z)\cdot(w-w')] - \mathbb{E}_{\hat{\mathbb{P}}_Z}[\nabla_w F(w_t,Z)\cdot(w-w')]\Big|\,\mathrm{d}t\,.\end{aligned}$$

Using the fact that $z \mapsto F(w_t,z)$ is in both $L^1(\mathbb{P}_Z)$ and $L^1(\hat{\mathbb{P}}_Z)$ since F is bounded, we conclude by the regularity of $\nabla_w F$. ■

All the bounds of this paper can be restated in this more general framework. We will only give a direct proof of Proposition 12.

Proposition 12 *Let $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$ and $\mathcal{W}$ be a compact set, with an $\{\varepsilon_k\}$-refining sequence of nets defined on it. Suppose that $w \mapsto \ell(w,x)$ is continuous, for $\mathbb{P}_X$-almost every x,*[10] *and that $\{\ell(w,X)\}_{w\in\mathcal{W}}$ is a ξ-SG stochastic process. Then we have*

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}}\sum_{k=1}^{\infty} \varepsilon_{k-1}\sqrt{2I(W_k;S)}\,.$$

[10] Note that in Asadi et al. (2018) the result is stated under a weaker assumption of separability of the process. To avoid introducing further definitions and technicalities in the proofs, we decided to focus on the case of a.s. continuity.

Proof By standard arguments, $\{\mathscr{L}_s(w)\}_{s\in\mathcal{S}}$ is a $(\xi/\sqrt{m})$-SG process. Hence, $\mathscr{L}_S(w)-\mathscr{L}_S(w')$ is $(\xi/\sqrt{m})$-SG for every $w, w'\in\mathcal{W}$. By Lemma 9, $s\mapsto\mathscr{L}_s(w)-\mathscr{L}_s(w')$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi/\sqrt{m})$ wrt $\mathbb{P}_S$ ($\forall w\in\mathcal{W}$), with $\mathfrak{D}:(\mu,\nu)\mapsto\sqrt{2\mathrm{KL}(\nu\|\mu)}$. Finally, let $g(w,s)=\mathscr{L}_{\mathcal{X}}(w)-\mathscr{L}_s(w)$. Clearly g has the same regularity of $\mathscr{L}$. It is not hard to show that $\mathbb{E}_{\mathbb{P}_S}[\sup_{w\in\mathcal{W}}|g(w,S)|]<+\infty$ (this is a straight consequence of Remark 8.1.5 in Vershynin (2018)). We conclude by Theorem 22 and Jensen's inequality. ■

B.2. Bounds for non-uniform $\mathfrak{D}$-regularity

As mentioned at the end of Section 3, the results given so far are stated under uniform regularity assumptions. The next two results show that this is not strictly necessary, and that slightly different bounds can be obtained relaxing these assumptions.

Theorem 24 *Consider a non-negative measurable function $w\mapsto\xi_w$ such that $\|\xi_W\|_{L^p(\mathbb{P}_W)}=\xi$, for some $p\in[1,+\infty]$. Assume that a measurable map $F:\mathcal{W}\times\mathcal{Z}\to\mathbb{R}$ is such that $z\mapsto F(w,z)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi_w)$ wrt $\mathbb{P}_Z$ and for every $w\in\mathcal{W}$. Then we have*

$$|\mathbb{E}_{\mathbb{P}_{W\otimes Z}}[F(W,Z)]-\mathbb{E}_{\mathbb{P}_{W,Z}}[F(W,Z)]|\leq\xi\,\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_Z,\mathbb{P}_{Z|W})^r]^{1/r}\,,$$

where r is such that $1/p+1/r=1$ (with the convention $1/\infty=0$).

Proof The proof is essentially the same as for Theorem 20, the only difference being that now we have

$$|\mathbb{E}_{\mathbb{P}_Z}[F(w,Z)]-\mathbb{E}_{\mathbb{P}_{Z|W=w}}[F(w,Z)]|\leq\xi_w\,\mathfrak{D}(\mathbb{P}_Z,\mathbb{P}_{Z|W=w})\,,$$

whose expectation under $\mathbb{P}_W$ can be upperbounded via Hölder's inequality ■

Theorem 25 *Let $\mathcal{W}$ be a compact subset of $\mathbb{R}^d$ and $\{\pi_k(\mathcal{W})\}$ a $\{\varepsilon_k\}$-refining sequence of nets on $\mathcal{W}$. Consider a measurable map $F:\mathcal{W}\times\mathcal{Z}\to\mathbb{R}$, such that $w\mapsto F(w,z)$ is continuous on $\mathcal{W}$, $\mathbb{P}_Z$-a.s., and $\mathbb{E}_{\mathbb{P}_Z}[\sup_{w\in\mathcal{W}}|F(w,Z)|]<+\infty$. Fix $\xi\geq0$ and consider a measurable map $w\mapsto\xi_w\geq0$ such that $\|\xi_{W_k}\|_{L^p(\mathbb{P}_W)}\leq\xi$, for all $k\in\mathbb{N}$ and for some $p\in[1,+\infty]$. Assume that for every $(w,w')\in\mathcal{W}^2$ the function $z\mapsto F(w,z)-F(w',z)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi_w\|w-w'\|)$, wrt $\mathbb{P}_Z$. Then, we have*

$$|\mathbb{E}_{\mathbb{P}_{W,Z}}[F(W,Z)]-\mathbb{E}_{\mathbb{P}_{W\otimes Z}}[F(W,Z)]|\leq\xi\sum_{k=1}^{\infty}\varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_Z,\mathbb{P}_{Z|W_k})^r]^{1/r}\,,$$

where r is such that $1/p+1/r=1$ (with the convention $1/\infty=0$).

Proof The proof is essentially analogous to the one of Theorem 22, but instead of (3) now we have

$$\begin{aligned}\big|\mathbb{E}_{\mathbb{P}_{Z|W_k=w_k}}[F(w_k,Z)-F(w_{k-1},Z)]-\mathbb{E}_{\mathbb{P}_Z}[F(w_k,Z)-F(w_{k-1},Z)]\big|&\\ \leq\xi_{w_k}\varepsilon_{k-1}\mathfrak{D}(\mathbb{P}_Z,\mathbb{P}_{Z|W_k=w_k})\,.&\end{aligned}$$

The conclusion follows easily by Hölder's inequality. ■

Appendix C. PAC-Bayesian bounds

The next result (Catoni, 2009) is a classical PAC-Bayesian bound. For the sake of completeness we give here a standard proof.

Proposition 16 *Assume that ℓ is bounded in $[-\xi, \xi]$. Let $\mathbb{P}^*_W$ be a fixed probability measure on $\mathcal{W}$, chosen independently of S. Fix $\delta \in (0,1)$ and $\lambda > 0$. Then, with probability $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$ larger than $1-\delta$ on the draw of S, we have*

$$\mathbb{E}_{\mathbb{P}_{W|S}}[g_S(W)] \leq \frac{\xi}{\sqrt{2m}}\left(\lambda + \frac{\mathrm{KL}(\mathbb{P}_{W|S}\|\mathbb{P}^*_W) + \log\frac{1}{\delta}}{\lambda}\right).$$

Proof We define $\mathbb{P}^*_{W\otimes S} = \mathbb{P}^*_W \otimes \mathbb{P}_S$. Fix $\lambda > 0$. Using the Donsker-Varadhan representation of the KL divergence (Donsker and Varadhan, 1983), we have that for all $s \in \mathcal{S}$

$$\mathbb{E}_{\mathbb{P}_{W|S=s}}[g_s(W)] \leq \frac{\xi}{\sqrt{2m}\lambda}\left(\mathrm{KL}(\mathbb{P}_{W|S=s}\|\mathbb{P}^*_W) + \log \mathbb{E}_{\mathbb{P}^*_W}[e^{\sqrt{2m}\lambda g_s(W)/\xi}]\right).$$

By Markov's inequality, we have that

$$\mathbb{P}_S\left(\mathbb{E}_{\mathbb{P}^*_W}[e^{\sqrt{2m}\lambda g_S(W)/\xi}] \leq \frac{1}{\delta}\,\mathbb{E}_{\mathbb{P}^*_{W\otimes S}}[e^{\sqrt{2m}\lambda g_S(W)/\xi}]\right) \geq 1-\delta\,.$$

Now, for all $w \in \mathcal{W}$ we have that $\ell(w, X) \subset [-\xi, \xi]$ is ξ-SG. In particular $g_S(w)$ is $(\xi/\sqrt{m})$-SG, as $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$. Since $\mathbb{E}_{\mathbb{P}_S}[g_S(w)] = 0$ we have

$$\log \mathbb{E}_{\mathbb{P}^*_{W\otimes S}}[e^{\sqrt{2m}\lambda g_S(W)/\xi}] \leq \lambda^2\,,$$

from which we conclude. ■

Note that although Proposition 16 is valid for all $\lambda > 0$, we cannot optimise the final bound wrt λ. Indeed, we have that such a choice of λ would depend on $\mathrm{KL}(\mathbb{P}_{W|S}, \mathbb{P}^*_W)$ and hence on the particular sample used. A possible strategy to overcome this issue consists in selecting a few possible values $\lambda_1, \dots, \lambda_t$ for λ, before drawing the sample S. Then, by mean of a union bound, one can say that with probability $\mathbb{P}_S$ higher than $1 - t\delta$ the generalisation is bounded by the best PAC-Bayesian bound among the t ones obtained.

Proposition 17 *Under the assumptions ♣, consider a $\{\varepsilon_k\}$-refining sequence of nets on $\mathcal{W}$ and assume that $\nabla_w \ell$ is bounded in $[-\xi, \xi]$. Let $\mathbb{P}^*_W$ be a fixed probability measure on $\mathcal{W}$, chosen independently of S. Fix two sequences $\{\delta_k\}_{k\in\mathbb{N}}$ and $\{\lambda_k\}_{k\in\mathbb{N}}$, such that $\delta_k \in (0,1)$ and $\lambda_k > 0$ for all k. Assume that $\sum_{k\in\mathbb{N}} \delta_k = \delta \in (0,1)$. Then, with probability $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$ larger than $1-\delta$ on the draw of S, we have*

$$\mathbb{E}_{\mathbb{P}_{W|S}}[g_S(W)] \leq \frac{\xi}{\sqrt{2m}}\left(2\sqrt{\log\frac{1}{\delta_0}} + \sum_{k=1}^{\infty} \varepsilon_{k-1}\left(\lambda_k + \frac{\mathrm{KL}(\mathbb{P}_{W_k|S}\|\mathbb{P}^*_{W_k}) + \log\frac{1}{\delta_k}}{\lambda_k}\right)\right).$$

Proof By the assumptions in ♣, $w \mapsto \mathscr{L}_s(w)$ is uniformly continuous on $\mathcal{W}$, $\mathbb{P}_S$-a.s. In particular, $\sup_{w\in\mathcal{W}} |\mathscr{L}_s(w) - \mathscr{L}_s(w_k)| \to 0$ as $k \to \infty$, $\mathbb{P}_S$-a.s. As a consequence $\sup_{w\in\mathcal{W}} |\mathbb{E}_{\mathbb{P}_S}[\mathscr{L}_S(w)] -$

$\mathbb{E}_{\mathbb{P}_S}[\mathscr{L}_S(w_k)]| \to 0$ ($\mathbb{P}_S$-a.s.) since the loss is uniformly bounded. It follows that, for $\mathbb{P}_S$-almost every s,

$$\lim_{k\to\infty} \mathbb{E}_{\mathbb{P}_{W|S=s}}[|g_s(W) - g_s(W_k)|] = 0\,.$$

Hence, recalling that $W_0 = w_0$, we have that, $\mathbb{P}_S$-a.s.,

$$\mathbb{E}_{\mathbb{P}_{W|S=s}}[g_s(W)] = g_s(w_0) + \sum_{k=1}^{\infty} \mathbb{E}_{\mathbb{P}_{W_k|S=s}}[g_s(W_k) - g_s(W_{k-1})]\,.$$

On the one hand, by Hoeffding's inequality (Hoeffding, 1963), the first term in the RHS can be upper-bounded with high probability, as

$$\mathbb{P}_S\left(g_S(w_0) > \xi\sqrt{\tfrac{2}{m}\log\tfrac{1}{\delta_0}}\right) \leq \delta_0\,.$$

On the other hand, proceeding as in the proof of Proposition 16, for each term in the telescopic sum we can write, for $\mathbb{P}_S$-almost every s,

$$\begin{aligned}&\mathbb{E}_{\mathbb{P}_{W_k|S=s}}[g_s(W_k) - g_s(W_{k-1})]\\ &\qquad\leq \frac{\varepsilon_{k-1}\xi}{\sqrt{2m}\lambda_k}\left(\mathrm{KL}(\mathbb{P}_{W_k|S=s}\|\mathbb{P}^*_{W_k}) + \log \mathbb{E}_{\mathbb{P}^*_{W_k}}[e^{\sqrt{2m}\lambda_k(g_s(W_k)-g_s(W_{k-1}))/(\varepsilon_{k-1}\xi)}]\right)\,.\end{aligned}$$

Now, $\nabla_w \ell \subset [-\xi, \xi]$, and hence $\nabla_w \ell(w, X)$ is ξ-SG, for all $w \in \mathcal{W}$. By Lemma 14, we have that $\{\ell(w, X)\}_{w\in\mathcal{W}}$ is a ξ-SG process. In particular, $\{g_S(w)\}_{w\in\mathcal{W}}$ is a centred $(\xi/\sqrt{m})$-SG process, as $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$. We have thus obtained that

$$\log \mathbb{E}_{\mathbb{P}^*_{W_k\otimes S}}[e^{\sqrt{2m}\lambda_k(g_S(W_k)-g_S(W_{k-1}))/(\varepsilon_{k-1}\xi)}] \leq \lambda_k^2\,.$$

By Markov's inequality we have that

$$\mathbb{P}_S\left(\mathbb{E}_{\mathbb{P}_{W_k|S}}[g_s(W_k) - g_s(W_{k-1})] > \frac{\varepsilon_{k-1}\xi}{\sqrt{2m}}\left(\lambda_k + \frac{\mathrm{KL}(\mathbb{P}_{W_k|S}\|\mathbb{P}^*_{W_k}) + \log\frac{1}{\delta_k}}{\lambda_k}\right)\right) \leq \delta_k\,.$$

We conclude by a union bound. ■

As for the standard PAC-Bayesian result, here as well we cannot directly optimise the parameters λ_k. Clearly one can again proceed by fixing few possible values for each parameter and then use a union argument to select the best bound. However, in this case this might become particularly hard, due to the large number of parameters. A possible way to address this problem consists in doing some optimisation that does not rely on the value of $\mathrm{KL}(\mathbb{P}_{W_k|S}, \mathbb{P}^*_{W_k})$, to reduce the number of parameters. For instance, we can proceed in the following way. One might suppose that $\mathrm{KL}(\mathbb{P}_{W_k|S}, \mathbb{P}^*_{W_k})$ increases linearly with k. Note that this is for instance the case if the algorithm is deterministic and $\mathbb{P}^*_{W_k}$ is uniform. So, let us say that we believe that $\mathrm{KL}(\mathbb{P}_{W_k|S}, \mathbb{P}^*_{W_k}) = \alpha k$, for some $\alpha > 0$. Then we are allowed to optimise all the λ_k in the chained PAC-Bayesian bound where $\mathrm{KL}(\mathbb{P}_{W_k|S}, \mathbb{P}^*_{W_k})$ is replaced by αk. This leads to

$$\mathbb{E}_{\mathbb{P}_{W|S}}[g_S(W)] \leq \frac{\xi}{\sqrt{2m}}\left(2\sqrt{\log\frac{1}{\delta_0}} + \sum_{k=1}^{\infty}\varepsilon_{k-1}\frac{\mathrm{KL}(\mathbb{P}_{W_k|S}\|\mathbb{P}^*_{W_k}) + \log\frac{1}{\delta_k}}{\sqrt{\alpha k + \log\frac{1}{\delta_k}}}\right)\,,$$

which is a valid bound, holding with probability higher than $1-\delta$, for all $\alpha > 0$. Now we have essentially replaced the λ_k with a single parameter α. Again we cannot optimise directly wrt α, but we can proceed as for the unchained bound, finding a good α by means of a union bound.

As a final remark note that one might want to optimise in terms of δ_k as well. This should be possible, but the constraint $\sum_k \delta_k = \delta$ and the non-convexity of the problem can make the minimisation quite hard in practice. Yet, one can probably resort to numerical methods.

Appendix D. Omitted proofs of Section 6

Lemma 18 *Under the assumptions ♣, let ε_0 and w_0 be such that $\|w-w_0\| \leq \varepsilon_0$, $\forall w \in \mathcal{W}$. Assume that $s \mapsto \nabla_w \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt $\mathbb{P}_S$, $\forall w \in \mathcal{W}$, and define $\hat{\mathscr{L}}_s(w) = \mathscr{L}_s(w) - \mathscr{L}_s(w_0)$ and $\hat{\mathcal{G}} = \mathbb{E}_{W\otimes S}[\hat{\mathscr{L}}_S(W)] - \mathbb{E}_{W,S}[\hat{\mathscr{L}}_S(W)]$. Then, $\hat{\mathcal{G}} = \mathcal{G}$, and $s \mapsto \hat{\mathscr{L}}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\varepsilon_0\xi)$, wrt $\mathbb{P}_S$ and $\forall w \in \mathcal{W}$.*

Proof The fact that $\mathbb{E}_{\mathbb{P}_{W,S}}[\mathscr{L}_S(w_0)] = \mathbb{E}_{\mathbb{P}_{W\otimes S}}[\mathscr{L}_S(w_0)]$ implies that $\hat{\mathcal{G}} = \mathcal{G}$. The regularity of $s \mapsto \hat{\mathscr{L}}_s(w)$ is a direct consequence of Lemma 23. ■

Lemma 19 *Assume that $\nu \mapsto \mathfrak{D}(\mathbb{P}_S, \nu)$ is convex. For any $\{\varepsilon_k\}$-refining sequence of nets on $\mathcal{W}$, the sequence $\{\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})]\}_{k\in\mathbb{N}}$ is non-decreasing and, $\forall k \in \mathbb{N}$, we have*

$$\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] \leq \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W})] \,.$$

Proof Fix $k \geq 0$ and $w_k \in \mathcal{W}_k$ such that $\mathbb{P}_W(W_k = w_k) > 0$. For any measurable set U on $\mathcal{S}$, we have

$$\mathbb{P}_{S|W_k=w_k}(U) = \int_{\mathcal{W}} \mathbb{P}_{S|W=w}(U) \mathrm{d}\mathbb{P}_{W|W_k=w_k}(w) \,,$$

where $\mathrm{d}\mathbb{P}_{W|W_k=w_k}(w) = \frac{\mathrm{d}\mathbb{P}_W(w)}{\mathbb{P}_W(W_k=w_k)}$ if $w \in \pi_k^{-1}(w_k)$, and 0 otherwise. Hence, we can write

$$\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k=w_k}) = \mathfrak{D}\left(\mathbb{P}_S, \int_{\mathcal{W}} \mathbb{P}_{S|W=w}(\cdot) \mathrm{d}\mathbb{P}_{W|W_k=w_k}(w)\right) \,.$$

Since $\nu \mapsto \mathfrak{D}(\mathbb{P}_S, \nu)$ is a convex function, we can use Jensen's inequality to obtain

$$\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k=w_k}) \leq \int_{\mathcal{W}} \mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W=w}) \mathrm{d}\mathbb{P}_{W|W_k=w_k}(w) \,.$$

By taking the expectation wrt $\mathbb{P}_{W_k}$ we conclude that

$$\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] \leq \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W})] \,.$$

Now, for any $k' > k$, the same proof can be used to show that

$$\mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] \leq \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_S, \mathbb{P}_{S|W_{k'}})] \,,$$

by simply replacing $\mathcal{W}$ with $\mathcal{W}_{k'}$ and $\mathbb{P}_W$ with $\mathbb{P}_{W_{k'}}$. ■

Appendix E. Toy Models

E.1. Example 1

Let $\mathcal{W} = \mathcal{X} = [-1, 1]$, $\ell(w, x) = \frac{1}{2}(w - x)^2$, and $\varepsilon_k = 2^{-k}$, for $k \in \mathbb{N}$. We can find mappings π_k that define a $\{\varepsilon_k\}$-refining sequence of nets, with $\mathcal{W}_k = \{2^{1-k}j \ : \ j \in [-2^{k-1} : 2^{k-1}]\}$, where $[a : b] = [a, b] \cap \mathbb{Z}$. Fix $k^\star \in \mathbb{N}$ and define $\theta = 2^{-k^\star}$. Let X be uniformly distributed on $(-\theta, \theta)$, that is $X \sim U_{(-\theta,\theta)}$. We choose an algorithm that, given x, selects the w minimising $\ell(w, x)$. This means that $\mathbb{P}_{W|X=x} = \delta_x$, where δ_x is the Dirac measure on x. Note that $\nabla_w \ell$ is 1-Lipschitz and ℓ is 2-Lipschitz (on $\mathcal{X}$, uniformly on $\mathcal{W}$). However, thanks to Lemma 18 we know that we can consider the loss $\tilde{\ell}(w, x) = \ell(w, x) - \frac{x^2}{2}$, which does not affect the algorithm, leads to the same generalisation, and is 1-Lipschitz. The marginal distribution of W is $W \sim U_{(-\theta,\theta)}$. Moreover, we have $\mathbb{E}_{\mathbb{P}_{W,X}}[\ell(W, X)] = 0$ and $\mathbb{E}_{\mathbb{P}_X}[\ell(w, X)] = \frac{1}{2}\left(w^2 + \frac{\theta^2}{3}\right)$. So,

$$\mathcal{G} = \mathbb{E}_{\mathbb{P}_{W\otimes X}}[\ell(W, X)] - \mathbb{E}_{\mathbb{P}_{W,X}}[\ell(W, X)] = \mathbb{E}_{\mathbb{P}_W}\left[\frac{1}{2}\left(W^2 + \frac{\theta^2}{3}\right)\right] = \frac{\theta^2}{3}\,.$$

Recall that we denote as $\mathcal{B}_\ell$ the bound in Proposition 11 and as $\mathcal{B}_{\nabla\ell}$ the chained bound from Proposition 15. We denote as $\mathcal{B}_{\tilde{\ell}}$ the unchained bound obtained using $\tilde{\ell}$ instead of ℓ. Clearly we have $\mathcal{B}_{\tilde{\ell}} = \mathcal{B}_\ell/2$. We will now evaluate $\mathcal{B}_{\tilde{\ell}}$ and $\mathcal{B}_{\nabla\ell}$. As a starting point, note that the 1-Wasserstein distance between two uniforms measures, on the intervals (A, B) and $(a, b) \subseteq (A, B)$, is given by

$$\mathfrak{W}(U_{(A,B)}, U_{(a,b)}) = \frac{(A - a)^2 + (B - b)^2}{2((B - A) - (b - a))}\,.$$

Note that choosing $a = b \in [A, B]$ in the RHS above gives the 1-Wasserstein distance between a uniform distribution and a Dirac measure. Now, let $a = b = w$, $A = -\theta$ and $B = \theta$. We find that

$$\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W=w}) = \frac{\theta}{2}\left(1 + \frac{w^2}{\theta^2}\right)\,.$$

It follows that

$$\mathcal{B}_{\tilde{\ell}} = \mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W=w})] = \frac{2}{3}\,\theta\,.$$

Comparing $\mathcal{G}$ and $\mathcal{B}_{\tilde{\ell}}$, we realize that the standard Wasserstein bound becomes loose for small θ.

Now, fix $k \in \mathbb{N}$. If $k \leq k^\star$, then $\pi_k(w) = w_0 = 0$ with probability 1. In particular, we have that $W_k \perp\!\!\!\perp X$ and hence $\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k}) = 0$. We will hence focus on the case $k > k^\star$. Let $k = k^\star + k'$. Now, notice that π_k defines $2^{k'} + 1$ intervals in $(-\theta, \theta)$. We will denote them as I_j, where $j \in [-2^{k'-1} : 2^{k'-1}]$. We have $I_{-2^{k'-1}} = (-\frac{1}{2^{k^\star}}, -\frac{2^{k'}-1}{2^k})$ and $I_{2^{k'-1}} = (\frac{2^{k'}-1}{2^k}, \frac{1}{2^{k^\star}})$, while, for $j \in [-2^{k'-1} + 1 : 2^{k'-1} - 1]$, $I_j = (\frac{2j-1}{2^k}, \frac{2j+1}{2^k})$. Note that the two outer intervals will have probability $\mathbb{P}_W(W \in I_{-2^{k'-1}}) = \mathbb{P}_W(W \in I_{-2^{k'-1}}) = 2^{-(k'+1)}$, while for the inner intervals, we have $\mathbb{P}_W(W \in I_j) = 2^{-k'}$.

Now, for $j \in [-2^{k'-1} + 1 : 2^{k'-1} - 1]$ we define $a_j = \frac{2j-1}{2^k}$ and $b_j = \frac{2j+1}{2^k}$. Note that for all these inner intervals we have $b_j - a_j = 2^{1-k}$, $(b_j - a_j)/\theta = 2^{1-k'}$, $a_j/\theta = (2j - 1)/2^{k'}$, and $b_j/\theta = (2j + 1)/2^{k'}$. So, the contribution brought by the inner intervals to $\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k})]$

is given by

$$E_1 = \frac{\theta}{2^{k'}} \sum_{j=-(2^{k'-1}-1)}^{2^{k'-1}-1} \frac{\left(1+\frac{2j-1}{2^{k'}}\right)^2 + \left(-1+\frac{2j+1}{2^{k'}}\right)^2}{4(1-2^{-k'})}$$
$$= \frac{\theta}{6(1-2^{-k'})}(4 - 12 \times 2^{-k'} + 11 \times 2^{-2k'} - 3 \times 2^{-3k'})\,.$$

On the other hand, the contribution of the two outer intervals ($j = \pm 2^{k'-1}$) is given by

$$E_2 = 2 \times \frac{\theta}{2^{k'+1}} \frac{1}{2} \frac{\left(2 - \frac{1}{2^{k'}}\right)^2}{\left(2 - \frac{1}{2^{k'}}\right)} = \frac{\theta}{2^{k'+1}}\left(2 - \frac{1}{2^{k'}}\right) = \theta\left(2^{-k'} - \frac{1}{2} \times 2^{-2k'}\right)\,.$$

We conclude that, for $k' \geq 1$ and $k = k^\star + k'$, we have

$$\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k})] = E_1 + E_2$$
$$= \frac{\theta}{6(1-2^{-k'})}(4 - 12 \times 2^{-k'} + 11 \times 2^{-2k'} - 3 \times 2^{-3k'}) + \theta\left(2^{-k'} - \frac{1}{2} \times 2^{-2k'}\right)\,.$$

We can finally compute $\mathcal{B}_{\nabla \ell}$, as we have

$$\mathcal{B}_{\nabla \ell} = \sum_{k=1}^{\infty} \frac{1}{2^{k-1}} \mathbb{E}_{W_k}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k})]$$
$$= \frac{1}{2^{k^\star}} \sum_{k'=1}^{\infty} \frac{1}{2^{k'-1}} \mathbb{E}_{W_{k^\star+k'}}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_{k^\star+k'}})] = \frac{247}{105}\theta^2 \simeq 2.35\,\theta^2\,.$$

Now, it is interesting to compare these results with the CMI bound. For this purpose, we need to compute $I(W_k; X)$ for a fixed $k \in \mathbb{N}$. Similar to the chained Wasserstein bound, for $k \leq k^\star$ we have that $I(W_k; X) = 0$ as $W_k \perp\!\!\!\perp X$. Therefore, we focus on $k = k^\star + k'$ where $k' \geq 1$. First, notice that the KL divergence between two uniform measures, on the intervals (A, B) and $(a, b) \subseteq (A, B)$, is given by

$$\mathrm{KL}(U_{(a,b)} \| U_{(A,B)}) = \log \frac{B-A}{b-a}\,.$$

As a consequence, we have that for the inner intervals I_j (with $j \in [-2^{k'-1}+1 : 2^{k'-1}-1]$)

$$\mathrm{KL}(\mathbb{P}_{X|W_k \in I_j} \| \mathbb{P}_X) = \log(2^{k-k^\star}) = k' \log 2\,,$$

while for the two outer intervals we have

$$\mathrm{KL}(\mathbb{P}_{X|W_k \in I_{-2^{k'-1}}} \| \mathbb{P}_X) = \mathrm{KL}(\mathbb{P}_{X|W_k \in I_{-2^{k'-1}}} \| \mathbb{P}_X) = \log(2^{k+1-k^\star}) = (k'+1)\log 2\,.$$

Taking the expectation wrt $\mathbb{P}_W$ we obtain

$$I(W_k; X) = \mathbb{E}_{\mathbb{P}_W}[\mathrm{KL}(\mathbb{P}_{X|W_k} \| \mathbb{P}_X)] = \sum_{j=-2^{k'-1}}^{2^{k'-1}} \mathbb{P}_W(W_k \in I_j) \mathrm{KL}(\mathbb{P}_{X|W_k \in I_j} \| \mathbb{P}_X)$$
$$= 2 \times 2^{-(k'+1)}(k'+1)\log 2 + (1 - 2 \times 2^{-(k'+1)})k' \log 2 = (k' + 2^{-k'})\log 2\,.$$

Therefore, the CMI bound is given by

$$\begin{aligned}\mathcal{B}_{\mathrm{CMI}} &= \sum_{k=1}^{\infty} \frac{1}{2^{k-1}} \sqrt{2I(W_k; X)} \\ &= \frac{1}{2^{k^\star}} \sum_{k'=1}^{\infty} \frac{1}{2^{k'-1}} \sqrt{2(k' + 2^{-k'}) \log 2} \simeq 3.50\, \theta .\end{aligned}$$

For $\theta \to 0$ (*i.e.*, $k^\star \to \infty$) $\mathcal{B}_{\nabla\ell}$ is much tighter than $\mathcal{B}_\ell$ and $\mathcal{B}_{\mathrm{CMI}}$, as it captures the asymptotic behaviour of $\mathcal{G} = \theta^2/3$.

Finally, let us consider the case of a random sample $S = \{X_1, \dots, X_m\}$, for $m > 1$. We denote as $\mathcal{B}_{\nabla\mathscr{L}}$ the chained Wasserstein bound, and as $\mathcal{B}_{\mathscr{L}}$ the unchained one. Minimising $\mathscr{L}_S$ leads to

$$W = \frac{1}{m} \sum_{i=1}^{m} X_i \,.$$

Since each X_i lies in $(-\theta, \theta)$ with probability 1, in particular we have that

$$\mathbb{P}_W(W \in (-\theta, \theta)) = 1\,.$$

So, for $k \le k^\star$, W_k is deterministic and hence $S \perp\!\!\!\perp W_k$. We get

$$\mathcal{B}_{\nabla\mathscr{L}} = \frac{1}{\sqrt{m}} \sum_{k=1}^{\infty} \frac{1}{2^{k-1}} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] = \frac{1}{2^{k^\star}\sqrt{m}} \sum_{k'=1}^{\infty} \frac{1}{2^{k'-1}} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] \le 2\theta \mathcal{B}_{\mathscr{L}}\,,$$

where we used Lemma 19 and the fact that $\theta = 2^{-k^\star}$. We have thus seen that even for large samples we still have that for $\theta \to 0$

$$\frac{\mathcal{B}_{\nabla\mathscr{L}}}{\mathcal{B}_{\mathscr{L}}} = O(\theta)\,.$$

E.1.1. HIGHER DIMENSIONAL VARIANT FOR A GENERIC LOSS

We discuss now a higher dimensional version of the above toy model. Fix a positive integer integer $d \ge 1$. Let $\mathcal{W} = \mathcal{X} = [-1, 1]^d$. Fix an integer $k^\star \ge 1$ and define $\theta = 2^{-k^\star}$. We will assume that the choice of $k^\star$ scales with d so that $\theta = \Theta(d^{-\alpha})$ for some $\alpha > 0$. Let X be uniformly distributed on $R_d = (\theta, \theta)^{d-1} \times (-1, 1)$. For $k \in \mathbb{N}$ we let $\varepsilon_k = 2^{-k}\sqrt{d}$ (the rescaling $\sqrt{d}$ is necessary as now $\mathcal{W}$ has diameter $2\sqrt{d}$) and we consider a $\{\varepsilon_k\}$-refining sequence of nets $\mathcal{W}_k = \tilde{\mathcal{W}}_k^{\otimes d}$, where $\tilde{\mathcal{W}}_k = \{2^{1-k} j \; : \; j \in [-2^{k-1} : 2^{k-1}]\}$. We consider a generic loss function ℓ satisfying the assumptions in ♣, and such that $\nabla_w \ell$ is 1-Lipschitz in $\mathcal{X}$, uniformly in $\mathcal{W}$. From Lemma 18 we know that we can find a loss $\tilde{\ell}$ which is $\sqrt{d}$-Lipschitz (as $\varepsilon_0 = \sqrt{d}$), and in general we cannot assume the Lipschitz constant to be smaller. As in the 1D example, we assume that we have an algorithm that given x, selects $w = x$. This means that $\mathbb{P}_{W|X=x} = \delta_x$, where δ_x is the Dirac measure on x, and so the marginal distribution of W is U_{R_d}.

As we are interested in evaluating the Wasserstein bounds, we will need to compute quantities like $\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W=w})$ and $\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k=w_k})$. This can be a pretty hard task if we use the standard 2-norm on $\mathbb{R}^d$ as the distance on $\mathcal{X}$. To give an idea of the challenge, note that already in dimension $d = 2$ computing the expected distance between two uniform distributions on rectangles is far from being trivial (Marsaglia et al., 1990). For this reason, everything is much easier to compute if we endow $\mathcal{X}$ with the distance given by the 1-norm on $\mathbb{R}^d$, that is

$$\hat{d}_{\mathcal{X}}(x,x') = \sum_{i=1}^{d} |x_i - x'_i| \,,$$

where x_i and x'_i are the components of x and x'. We will denote the Wasserstein distances computed in this way as $\hat{\mathfrak{W}}$, and the bounds based on this distance as $\hat{\mathcal{B}}$. Note, however, that we always have that $\mathfrak{W} \leq \hat{\mathfrak{W}}$, where $\mathfrak{W}$ is the Wasserstein distance with cost

$$d_{\mathcal{X}}(x,x') = \|x - x'\| \,,$$

as $d_{\mathcal{X}}(x,x') \leq \hat{d}_{\mathcal{X}}(x,x')$ for all x, x'. Moreover, when x and x' are in R_d, we have that $\hat{d}_{\mathcal{X}}(x,x') - d_{\mathcal{X}}(x,x') = O(\theta\sqrt{d-1})$. For this reason, since $\theta = \Theta(d^{-\alpha})$, we obtain that $\hat{\mathcal{B}}_\ell - \mathcal{B}_\ell = O(d^{1-\alpha})$ and $\hat{\mathcal{B}}_{\nabla\ell} - \mathcal{B}_{\nabla\ell} = O(d^{1-\alpha})$.

Now, for the Wasserstein distatence between $\mathbb{P}_X$ and $\mathbb{P}_{X|W=w}$, thanks to the fact that we are using $\hat{d}_{\mathcal{X}}$, we have

$$\hat{\mathfrak{W}}(\mathbb{P}_X, \mathbb{P}_{X|W=w}) = \sum_{i=1}^{d} \mathfrak{W}_{\mathrm{1D}}(\mathbb{P}_{X_i}, \mathbb{P}_{X_i|W=w}) \,,$$

where $\mathfrak{W}_{\mathrm{1D}}$ is the Wasserstein distance wrt the 1D distance $d_{\mathcal{X}_i}(x_i, x'_i) = |x_i - x'_i|$. Taking the expectation wrt $\mathbb{P}_W$ we find

$$\hat{\mathcal{B}}_{\tilde{\ell}} = \sqrt{d}\,\mathbb{E}_{\mathbb{P}_W}[\hat{\mathfrak{W}}(\mathbb{P}_X, \mathbb{P}_{X|W})] = \frac{2\sqrt{d}}{3}(1 - (d-1)\theta) = \Theta(d^{1/2} + d^{3/2-\alpha}) \,.$$

Since $\hat{\mathcal{B}}_{\tilde{\ell}} - \mathcal{B}_{\tilde{\ell}} = O(d^{1-\alpha})$, if follows that

$$\mathcal{B}_{\tilde{\ell}} = \Theta(d^{1/2} + d^{3/2-\alpha}) \,.$$

We are now left with the task of estimating $\mathcal{B}_{\nabla\ell}$. Fix w_k such that $\mathbb{P}_W(W_k = w_k) > 0$. Now, we have that $\mathbb{P}_{X|W_k=w_k}$ is the uniform distribution on the rectangle $\pi_k^{-1}(\mathcal{W})$. Up to sets of measure 0, we can find d intervals (a_i, b_i) such that

$$\pi_k^{-1}(\mathcal{W}) = (a_1, b_1) \times \cdots \times (a_d, b_d) \,.$$

We can choose a transport plan that is composed of d steps. First we squeeze all the probability mass from $\mathcal{X}$ to $(a_1, b_1) \times (-1,1)^{d-1}$. Then we squeeze the second component, and so on. In this way we find that

$$\hat{\mathfrak{W}}(\mathbb{P}_X, \mathbb{P}_{X|W_k=w_k}) \leq \sum_{i=1}^{d} \mathfrak{W}_{\mathrm{1D}}(\mathbb{P}_{X_i}, \mathbb{P}_{X_i|W_k=w_k}) \,.$$

On the other hand, we have that

$$\begin{aligned} \hat{\mathfrak{W}}(\mathbb{P}_X,\mathbb{P}_{X|W_k=w_k}) &= \inf_{\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_{X|W_k=w_k})} \mathbb{E}_{(X,X')\sim\pi}[\hat{d}_{\mathcal{X}}(X,X')] \\ &= \inf_{\pi \in \Pi(\mathbb{P}_X, \mathbb{P}_{X|W_k=w_k})} \sum_{i=1}^{d} \mathbb{E}_{(X,X')\sim\pi}[|X_i - X'_i|] \geq \sum_{i=1}^{d} \mathfrak{W}_{\mathrm{1D}}(\mathbb{P}_{X_i}, \mathbb{P}_{X_i|W_k=w_k}) \,. \end{aligned}$$

We conclude that

$$\hat{\mathfrak{W}}(\mathbb{P}_X, \mathbb{P}_{X|W_k=w_k}) = \sum_{i=1}^{d} \mathfrak{W}_{\mathrm{1D}}(\mathbb{P}_{X_i}, \mathbb{P}_{X_i|W_k=w_k})\,.$$

We are now back at evaluating Wasserstein distances between uniform distributions on invervals. Proceeding as in the 1D version of the toy example we find

$$\hat{\mathcal{B}}_{\nabla\ell} = \sum_{k=1}^{\infty} \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_W}[\hat{\mathfrak{W}}(\mathbb{P}_S, \mathbb{P}_{S|W_k})] = \frac{247\sqrt{d}}{105}\,(1+(d-1)\theta^2) = \Theta(d^{1/2} + d^{3/2-2\alpha})\,.$$

Again, since $\hat{\mathcal{B}}_{\nabla\ell} - \mathcal{B}_{\nabla\ell} = O(d^{1-\alpha})$ we have that if $\alpha \geq 1/2$

$$\mathcal{B}_{\nabla\ell} = \Theta(d^{1/2})\,.$$

In general, as α might be in $(0, 1/2)$, we can say that (since $\mathcal{B}_{\nabla\ell} \leq \hat{\mathcal{B}}_{\nabla\ell}$)

$$\mathcal{B}_{\nabla\ell} = O(d^{1/2} + d^{3/2-2\alpha})\,.$$

Now, we want to compare the two bounds. We have

$$\frac{\mathcal{B}_{\nabla\ell}}{\mathcal{B}_{\tilde{\ell}}} = O\left(\frac{1+d^{1-2\alpha}}{1+d^{1-\alpha}}\right)\,.$$

If $\alpha \in (0,1)$, we have that this ratio vanishes for $d \to \infty$, meaning that the chained bounds becomes much tighter than its unchained counterpart. On the other hand, for $\alpha > 1$ the ratio is of order 1.

E.2. Example 2

Let $\mathcal{W} = \{w \in \mathbb{R}^2 : \|w\| = 1\}$ and $\mathcal{X} = \mathbb{R}^2$. Fix $a > 0$ and let $X \sim \mathcal{N}(\mathbf{A}, \mathrm{Id})$, a multivariate normal distribution centered in $\mathbf{A} = (a, 0)$, with covariance matrix given by the identity. Let the loss be $\ell(w, x) = -w \cdot x$. As in Example 1, the algorithm selects the w minimising the loss. In practice, we are trying to find the direction of the mean of X, which is $(1, 0)$. Let $\varepsilon_k = 4/2^k$ (for $k \in \mathbb{N}$), $w_0 = (1, 0)$, and $\mathcal{W}_k = \{w = (\cos\frac{2\pi j}{2^k}, \sin\frac{2\pi j}{2^k}\phi) : j \in [-2^{k-1} : 2^{k-1} - 1]\}$ for $k \geq 1$. We can easily define projections π_k that make $\{\mathcal{W}_k\}_{k\in\mathbb{N}}$ a $\{\varepsilon_k\}$-sequence of refining nets. With no difficulty one can verify that ℓ is 1-Lipschitz in $\mathcal{X}$, $\forall w \in \mathcal{W}$. Since $\mathcal{W}$ is not convex, we want to use Theorem 22 to give our chaining bound. It is easy to verify that ℓ satisfies the $\mathfrak{D}$ regularity with $\mathfrak{D} = \mathfrak{W}$, as

$$|(\ell(w,x) - \ell(w,x')) - (\ell(w',x) - \ell(w',x'))| \leq \|x - x'\|\|w - w'\|\,.$$

Since the values of $\mathcal{G}$, $\mathcal{B}_\ell$, and $\mathcal{B}_{\nabla\ell}$ depend on a, we will explicitly write them as functions of a. We will start by finding the exact expression of $|\mathcal{G}(a)|$.

Denote as $\mathfrak{a}$ the Cartesian axis on which $\mathbf{A}$ lies. For $v \in \mathbb{R}^2$, denote as $\alpha(v)$ be the angle between v and $\mathfrak{a}$. Since the learnt w is parallel to x, we have that, with probability 1, $\alpha(X) = \alpha(W)$. Thus, the distribution of $\alpha(W)$ is the distribution of the angle of an isotropic Gaussian centred in $\mathbf{A}$, whose density is given by (Cooper and Farid, 2020)

$$\rho_a(\alpha) = \frac{\phi(a)}{\sqrt{2\pi}}\left(1 + \frac{a\cos\alpha\,\Phi(a\cos\alpha)}{\phi(a\cos\alpha)}\right)\,,$$

where $\phi(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$ and $\Phi(t) = \frac{1}{2}(1 + \mathrm{erf}(t/\sqrt{2}))$.

Now, we can actually give an explicit form for $|\mathcal{G}(a)|$. Indeed, we have

$$|\mathcal{G}(a)| = a\int_{-\pi}^{\pi]}(1-\cos\alpha)\rho_a(\alpha)\mathrm{d}\alpha = a - \frac{\phi(a)}{\sqrt{2\pi}}\int_{-\pi}^{\pi}(a\cos\alpha)^2\frac{\Phi(a\cos\alpha)}{\phi(a\cos\alpha)}\mathrm{d}\alpha\,.$$

Performing a change of variable we get

$$\begin{aligned}\int_{-\pi}^{\pi}(a\cos\alpha)^2\frac{\Phi(a\cos\alpha)}{\phi(a\cos\alpha)}\mathrm{d}\alpha &= 2\int_{-a}^{a}\frac{u^2}{\sqrt{a^2-u^2}}\frac{\Phi(u)}{\phi(u)}\mathrm{d}u\\ &= \frac{a^2e^{a^2/4}\pi^{3/2}}{\sqrt{2}}\left(I_0(a^2/a) + I_1(a^2/4)\right)\,,\end{aligned}$$

where $I_n(t)$ denotes the modified Bessel function of the first kind. So, we have

$$|\mathcal{G}(a)| = a\left(1 - \frac{1}{2}\frac{I_0(a^2/a) + I_1(a^2/4)}{\sqrt{\frac{2}{\pi}\frac{e^{a^2/4}}{a}}}\right)\,.$$

We can now use the asymptotic expansions

$$\begin{aligned}I_0(a^2/4) &= \sqrt{\frac{2}{\pi}}\frac{e^{a^2/4}}{a}\left(1 + \frac{1}{2a^2} + O(a^{-4})\right)\,;\\ I_1(a^2/4) &= \sqrt{\frac{2}{\pi}}\frac{e^{a^2/4}}{a}\left(1 - \frac{3}{2a^2} + O(a^{-4})\right)\,,\end{aligned}$$

to get that

$$|\mathcal{G}(a)| = \frac{1}{2a} + O(a^{-3})\,.$$

Now, we want to show that, as $a \to \infty$, $\mathcal{B}_\ell$ is of order 1. We start by computing a lower bound. For each w, let us consider a new set of Cartesian axes ($\mathfrak{u}(w)$ and $\mathfrak{v}(w)$), such that the angle between $\mathfrak{v}(w)$ and $\mathfrak{a}$ is $\alpha(w)$, and $\mathfrak{u}(w)$ is the normal axis which contains the point $\mathbf{A}$. We choose the orientation of the axes so that in this reference framework we have $\mathbf{A} = (a\sin\alpha(w), 0)$. Since X, conditioned on $\mathcal{W} = w$, has support contained in the axis $\mathfrak{v}(w)$, the Wasserstein distance $\mathcal{W}(\mathbb{P}_X, \mathbb{P}_{X|W=w})$ is lower-bounded by the transport cost of moving every point in $\mathbb{R}^2$ to the closest point on $\mathfrak{v}(w)$. We thus have

$$\begin{aligned}\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W=w}) &\geq \frac{1}{2\pi}\int_{\mathbb{R}^2}|u|e^{-\frac{(u-a\sin\alpha(w))^2+v^2}{2}}\mathrm{d}u\mathrm{d}v\\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}|u|e^{-\frac{(u-a\sin\alpha(w))^2}{2}}\mathrm{d}u \geq a|\sin\alpha(w)|\,.\end{aligned}$$

We can now explicitly compute a lower bound for $\mathcal{B}_\ell(a)$ by taking the expectation wrt $\mathbb{P}_W$. We get

$$\mathcal{B}_\ell(a) \geq \int_{-\pi}^{\pi}a|\sin\theta|\rho_a(\theta)\mathrm{d}\theta = \sqrt{\frac{2}{\pi}}\,\mathrm{erf}\,\frac{a}{\sqrt{2}}\,.$$

In particular, we have established that

$$\liminf_{a\to\infty} \mathcal{B}_\ell(a) \geq \sqrt{\frac{2}{\pi}}\,.$$

We can now look for an upper bound on $\mathcal{B}_\ell(a)$. Fixed w, we can consider the following transport plan from $\mathbb{P}_X$ to $\mathbb{P}_{X|W=w}$. First, we transport all the probability mass on $\mathfrak{v}(w)$, then we arrange the mass on $\mathfrak{v}(w)$ so as to reach the correct density. For the first step, notice that we are simply projecting $\mathbb{P}_X$ on $\mathfrak{v}(w)$. It is not hard to realise that in this way the linear density obtained on $\mathfrak{v}(w)$ is a centred standard normal distribution. The transport cost for this step is given by

$$\frac{1}{2\pi}\int_{\mathbb{R}^2} |u| e^{-\frac{(u-a\sin\alpha(w))^2+v^2}{2}} \mathrm{d}u\mathrm{d}v \leq 1 + a|\sin\alpha(w)|\,.$$

Now let $V \sim \mathcal{N}(0,1)$. The actual distribution of $\mathbb{P}_{X|W=w}$ on $\mathfrak{v}(w)$ is actually given by V, conditioned on $V \geq -a\cos\alpha(w)$, as $-a\cos\alpha(w)$ is the coordinate on $\mathfrak{v}(w)$ of the origin of the standard $\mathbb{R}^2$ Cartesian framework and so $\mathbb{P}_{X|W=w}$ has support $\{v \in \mathfrak{v}(w) : v \geq -a\cos\alpha(w)\}$. We can easily evaluate

$$\mathfrak{W}(\mathbb{P}_V, \mathbb{P}_{V|V\geq -a\cos\alpha(w)}) = \frac{\phi(a\cos\alpha(w))}{\Phi(a\cos\alpha(w))}\,.$$

So we have found that

$$\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W=w}) \leq 1 + a|\sin\alpha(w)| + \frac{\phi(a\cos\alpha(w))}{\Phi(a\cos\alpha(w))}\,.$$

Averaging on w we get that

$$\mathcal{B}_\ell(a) = \mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W})] \leq 1 + \sqrt{\frac{2}{\pi}}\operatorname{erf}\frac{a}{\sqrt{2}} + \frac{e^{-a^2}}{\Phi(-a)}\,,$$

and so

$$\limsup_{a\to\infty} \mathcal{B}_\ell(a) \leq 1 + \sqrt{\frac{2}{\pi}}\,.$$

In particular, we have found that $\mathcal{B}_\ell(a) = \Theta(1)$, for $a \to \infty$.

We are now left with the task of evaluating $\mathcal{B}_{\nabla\ell}(a)$. Recall that, for each $k \geq 1$, we have $\mathcal{W}_k = \{w = (\cos\frac{2\pi j}{2^k}, \sin\frac{2\pi j}{2^k}) : j \in [-2^{k-1} : 2^{k-1}-1]\}$ and $w_0 = (1,0)$. Denote as $\mathcal{U}_k$ the partition on $\mathcal{W}$ induced by π_k, that is

$$\mathcal{U}_k = \{U = \pi_k^{-1}(w) :\, w \in \mathcal{W}_k\}\,.$$

We can certainly suppose that each $U \in \mathcal{U}_k$ is the circular arc enclosed by two adjacent elements of $\mathcal{W}_k$. Now, let $\bar{\mathcal{U}}_k = \{U \in \mathcal{U}_k :\, (1,0) \neq U\}$ and define $\theta_k = \pi/2^k$. Then, we have that, up to points of null measure, $\{W_k = (1,0)\} = \{|\alpha(W)| \leq \theta_k\}$. As a consequence

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k})] &= \textstyle\sum_{U\in\mathcal{U}_k} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W\in U})\mathbf{1}_U(W)] \\ &= \mathbb{P}_W(|\alpha(W)| \leq \theta_k)\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k=w_0}) + \textstyle\sum_{U\in\bar{\mathcal{U}}_k} \mathbb{P}_W(W\in U)\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W\in U})\,,\end{aligned} \tag{4}$$

where $\mathbf{1}_U$ is the indicator function of the event U. We need now to upper-bound the terms of this sum.

Let us define $Z = X - \mathbf{A}$. Clearly $Z \sim \mathcal{N}(0, \mathrm{Id})$. Let ρ be the density of Z, a centered standard multivariate normal, and $\tilde{\rho}$ be the density of Z conditioned on $|\alpha(W)| \leq \theta_k$. We have that

$$\tilde{\rho}(z) = \begin{cases} 0 & \text{if } |\alpha| > \theta; \\ \rho(z)/\mathbb{P}_W(|\alpha(W)| \leq \theta_k) & \text{otherwise.} \end{cases}$$

Let $\zeta = \|Z\|$ and note that $\zeta \sim \chi_2$, the Rayleigh distribution. We notice that

$$\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k=w_0}) = \mathfrak{W}(\mathbb{P}_Z, \mathbb{P}_{Z||\alpha(W)|\leq\theta_k})\,.$$

We can upper-bound this quantity by the transport cost of moving the mass $\mathbb{P}_W(|\alpha(W)| > \theta_k)$ away from $\{|\alpha(W)| > \theta_k\}$, bringing it all on $\mathbf{A}$, and finally redistributing it in the slice $\{|\alpha(W)| \leq \theta_k\}$, proportionally to $\tilde{\rho}$. We hence have

$$\mathfrak{W}(\mathbb{P}_Z, \mathbb{P}_{Z||\alpha(W)|\leq\theta_k}) \leq \mathbb{P}_W(|\alpha(W)| > \theta_k)(\mathfrak{W}(\mathbb{P}_Z, \delta_{\mathbf{A}}) + \mathfrak{W}(\delta_{\mathbf{A}}, \mathbb{P}_{Z||\alpha(W)|\leq\theta_k}))\,.$$

We can evaluate

$$\mathfrak{W}(\mathbb{P}_Z, \delta_{\mathbf{A}}) = \int_{\mathbb{R}^2} \|z\|\rho(z)\mathrm{d}z = \mathbb{E}_\zeta[\zeta] = \sqrt{\frac{\pi}{2}}\,.$$

On the other hand,

$$\begin{aligned} \mathfrak{W}(\delta_{\mathbf{A}}, \mathbb{P}_{Z||\alpha(W)|\leq\theta_k}) &= \int_{\mathbb{R}^2} \|z\|\tilde{\rho}(z)\mathrm{d}z \\ &\leq \frac{1}{\mathbb{P}_W(|\alpha(W)| \leq \theta_k)} \int_{\mathbb{R}^2} \|z\|\rho(z)\mathrm{d}z = \frac{\sqrt{\pi/2}}{\mathbb{P}_W(|\alpha(W)| \leq \theta_k)}\,. \end{aligned}$$

Now notice that

$$\mathbb{P}_W(|\alpha(W)| \leq \theta_k) \geq \mathbb{P}_\zeta(\zeta \leq a\sin\theta_k) = F_\zeta(a\sin\theta_k)\,,$$

where $F_\zeta : u \mapsto 1 - e^{u^2/2}$ is the cdf of ζ. As a consequence we eventually find

$$\begin{aligned} &\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k=w_0}) \\ &\quad \leq \mathbb{P}_W(|\alpha(W)| > \theta_k)\left(1 + \frac{1}{\mathbb{P}_W(|\alpha(W)| \leq \theta_k)}\right)\sqrt{\frac{\pi}{2}} \leq \left(1 + \frac{1}{F_\zeta(a\sin\theta_k)}\right)\sqrt{\frac{\pi}{2}}\,. \end{aligned}$$

Now, for $U \in \bar{\mathcal{U}}_k$, we have that $\{W \in U\} \subseteq \{|\alpha(W)| \geq \theta_k\}$. We can upper-bound $\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W\in U}) = \mathfrak{W}(\mathbb{P}_Z, \mathbb{P}_{Z|W\in U})$ as

$$\mathfrak{W}(\mathbb{P}_Z, \mathbb{P}_{Z|W\in U}) \leq \mathfrak{W}(\mathbb{P}_Z, \delta_{\mathbf{A}}) + \mathfrak{W}(\delta_{\mathbf{A}}, \mathbb{P}_{Z|W\in U})\,.$$

We have already computed $\mathfrak{W}(\mathbb{P}_Z, \delta_{\mathbf{A}}) = \sqrt{\pi/2}$. For the other term we have

$$\begin{aligned} \mathfrak{W}(\delta_{\mathbf{A}}, \mathbb{P}_{Z|W\in U}) &= \frac{1}{\mathbb{P}_W(W \in U)} \int_{(z+\mathbf{A})/\|z+\mathbf{A}\|\in U} \|z\|\rho(z)\mathrm{d}z \\ &\leq \frac{1}{\mathbb{P}_W(W \in U)} \int_{\|z\|>a\sin\theta_k} \|z\|\rho(z)\mathrm{d}z = \frac{1 - F_\zeta(a\sin\theta_k)}{\mathbb{P}(W \in U)}\,. \end{aligned}$$

We have thus obtained that

$$\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W\in U}) \leq \sqrt{\frac{\pi}{2}} + \frac{1 - F_\zeta(a\sin\theta_k)}{\mathbb{P}(W \in U)}\,.$$

Going back to (4), we can now write

$$\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k})] \leq (1 - F_\zeta(a\sin\theta_k))\left(\left(2 + \frac{1}{F_\zeta(a\sin\theta_k)}\right)\sqrt{\frac{\pi}{2}} + 2^k - 1\right)\,, \tag{5}$$

where we used that $\overline{\mathcal{U}}_k$ has $2^k - 1$ elements and that $\sum_{u\in\bar{\mathcal{U}}_k}\mathbb{P}_W(W \in U) \leq (1 - F_\zeta(a\sin\theta_k))$. Now, by plugging into (5) the explicit expressions of F_ζ and θ_k we obtain

$$\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k})] \leq e^{-\frac{1}{2}a^2\sin^2(\pi/2^k)}\left(2^k - 1 + \left(2 + \frac{1}{1 - e^{-\frac{1}{2}a^2\sin^2(\pi/2^k)}}\right)\sqrt{\frac{\pi}{2}}\right) = \mathcal{B}_k(a)\,.$$

Fix $k^\star > 1$. By Lemma 19, we have that for all $k \leq k^\star$, $\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k})] \leq \mathcal{B}_{k^\star}(a)$, and for $k > k^\star$ we have $\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X|W_k})] \leq \mathcal{B}_\ell(a)$. So we have that

$$\mathcal{B}_{\nabla\ell}(a) \leq \sum_{k=1}^{k^\star}\varepsilon_{k-1}\mathcal{B}_{k^\star}(a) + \sum_{k=k^\star}^{\infty}\varepsilon_k\mathcal{B}_\ell(a) \leq 8\,\mathcal{B}_{k^\star}(a) + 4 \times 2^{-k^\star}\mathcal{B}_\ell(a)\,.$$

Now the idea is that we want to choose $k^\star = k^\star_a$ as a function of a, in a way that makes the bound vanish for $a \to +\infty$. Note that if

$$a \geq \frac{2\log 2\sqrt{k^\star_a}}{\sin(\pi/2^{k^\star_a})}\,, \tag{6}$$

then

$$\mathcal{B}_{k^\star_a}(a) \leq 2^{-k^\star_a} + 2^{-2k^\star_a}\left(2 + \frac{1}{1 - 2^{-2k^\star_a}}\right)\sqrt{\frac{\pi}{2}}\,.$$

Notice we can choose $a \mapsto k^\star_a$ such that (6) holds and $a = O(2^{-k^\star_a}\sqrt{k^\star_a})$, for $a \to +\infty$, which implies

$$2^{-k^\star_a} = O\left(\frac{\log a - \log\log a}{a}\right)\,.$$

This proves the asymptotic behaviour for large a

$$\mathcal{B}_{\nabla\ell}(a) = O\left(\frac{\log a - \log\log a}{a}\right)\,.$$

In particular, up to logarithmic factors, the chained bound can capture the correct behaviour of $\mathcal{G}(a)$.

Appendix F. Technicalities

Lemma 26 *The mapping $w \mapsto \mathbb{P}_{Z|W=w}$ is measurable.*

Proof Recall that $\Sigma_{\mathscr{P}}$ is the σ-algebra on $\mathscr{P}$ induced by the weak topology. $\Sigma_{\mathscr{P}}$ is generated by the maps $\phi_U : \mathscr{P} \to [0,1]$, given by $\mu \mapsto \phi_U(\mu) = \mu(U)$, for U ranging in $\Sigma_{\mathcal{Z}}$ (cf. Theorem 17.24 in Kechris (1995)). By definition of regular conditional probability, for every $U \in \Sigma_{\mathcal{Z}}$ the map $w \mapsto \mathbb{P}_{Z|W=w}(U)$ is measurable. Hence $w \mapsto \mathbb{P}_{Z|W=w}$ is a measurable map $\mathcal{W} \to \mathscr{P}$ wrt $\Sigma_{\mathscr{P}}$. ■

Definition 27 *Let $f : (0,+\infty) \to \mathbb{R}$ be a convex lower semi-continuous map such that $f(1) = 0$ and $\lim_{x\to+\infty} f(x)/x = +\infty$. For $\mu, \nu \in \mathscr{P}$ we define the f-divergence*

$$D_f(\nu\|\mu) = \begin{cases} \mathbb{E}_\mu[f(\frac{\mathrm{d}\nu}{\mathrm{d}\mu})] & \textit{if } \nu \ll \mu; \\ +\infty & \textit{otherwise.} \end{cases}$$

Examples of f divergences are the KL divergence ($f : u \mapsto u \log u$) and the p-power divergence ($f : u \mapsto u^p - 1$).

Lemma 28 *$\mathfrak{D} : \mathscr{P} \times \mathscr{P} \to [0,+\infty]$, defined by $\mathfrak{D}(\mu,\nu) = D_f(\nu\|\mu)$, is measurable.*

Proof The measurability follows from the fact D_f is weakly lower semi-continuous (see Corollary 2.9 and Remark 2.1 in Liero et al. (2018)). ■

Lemma 29 *$\mathfrak{D} : \mathscr{P} \times \mathscr{P} \to [0,+\infty]$, defined by $\mathfrak{D}(\mu,\nu) = \mathfrak{W}(\mu,\nu)$, is measurable.*

Proof The measurability follows from the weak lower semi-continuity of $\mathfrak{W}$ (see Villani (2009), Remark 6.12). ■

Lemma 30 $\mathrm{Supp}(\mathbb{P}_{Z|W=w}) \subseteq \mathrm{Supp}(\mathbb{P}_Z)$, *$\mathbb{P}_W$-a.s.*

Proof We start by recalling that given a measure $\mu \in \mathscr{P}$, $\mathrm{Supp}(\mu)$ is the smallest closed subset K of $\mathcal{Z}$ such that $\mu(K) = 1$. Let $U \subseteq \mathcal{W}$ be defined as

$$U = \{w \in \mathcal{W} : \mathbb{P}_{Z|W=w}(\mathrm{Supp}(\mathbb{P}_Z)) < 1\}\,.$$

First, we notice that U is measurable. Indeed, $\mathrm{Supp}(\mathbb{P}_Z)$ is closed, and hence measurable, so $w \mapsto \mathbb{P}_{Z|W=w}(\mathrm{Supp}(\mathbb{P}_Z))$ is a measurable map, by definition of regular conditional probability. Now note that

$$\begin{aligned} 1 = \mathbb{P}_Z(\mathrm{Supp}(\mathbb{P}_Z)) &= \int_{\mathcal{W}} \mathbb{P}_{Z|W=w}(\mathrm{Supp}(\mathbb{P}_Z))\,\mathrm{d}\mathbb{P}_W(w) \\ &\leq 1 - \mathbb{P}_W(U) + \int_U \mathbb{P}_{Z|W=w}(\mathrm{Supp}(\mathbb{P}_Z))\,\mathrm{d}\mathbb{P}_W(w)\,. \end{aligned}$$

As a consequence, we must have that $\int_U \mathbb{P}_{Z|W=w}(\mathrm{Supp}(\mathbb{P}_Z))\,\mathrm{d}\mathbb{P}_W(w) \geq \mathbb{P}_W(U)$. However, by definition $\mathbb{P}_{Z|W=w}(\mathrm{Supp}(\mathbb{P}_Z)) < 1$ for $w \in U$, and so we necessarily have $\mathbb{P}_W(U) = 0$. We conclude by noticing that $\mathrm{Supp}(\mathbb{P}_{Z|W=w}) \supset \mathrm{Supp}(\mathbb{P}_Z)$ if, and only if, $w \in U$. ■

Appendix G. Explicit bounds

In this section we present several bounds that can be derived via the framework of Section 3. To our knowledge, all the chaining bounds that we present here are new, the only exception being the one in Proposition 43, which was recently established in Zhou et al. (2022). However, most of the unchained counterparts were already derived in the literature. The reader can find the bibliographic references in Table 1. Henceforth, all the chained bounds that we state are valid for any $\{\varepsilon_k\}$-sequence of refining nets on $\mathcal{W}$.

G.1. A few examples of $\mathfrak{D}$-regularity

Definition 31 (Power divergence) *Let $p > 1$. Given two probabilities μ and ν on $\mathcal{Z}$, we define the p-power divergence*

$$D^{(p)}(\nu\|\mu) = \begin{cases} \mathbb{E}_\mu\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)^p\right] - 1 & \text{if } \nu \ll \mu; \\ +\infty & \text{otherwise.} \end{cases}$$

For $p = 2$, we denote $D^{(2)}(\nu\|\mu)$ as $\chi^2(\nu\|\mu)$.

Lemma 32 *Fix $p > 1$ and let $r = p/(p-1)$. Let $\mathfrak{D} : (\mu,\nu) \mapsto (D^{(p)}(\nu\|\mu)+1)^{1/p}$ and $f : \mathcal{Z} \to \mathbb{R}^q$ be measurable. Assume that $f \in L^1(\mu)$ and write $f_\mu = \mathbb{E}_\mu[f(Z)]$. If $\mathbb{E}_\mu[\|f(Z) - f_\mu\|^r]^{1/r} \leq \xi$, then f has regularity $\mathcal{R}_\mathfrak{D}(\xi)$ wrt μ.*

Proof Notice that $\mathfrak{D}$ is measurable by Lemma 28. First, we consider the case $q = 1$. Fix $\nu \in \mathscr{P}$ such that $\mathrm{Supp}(\nu) \subseteq \mathrm{Supp}(\mu)$ and $f \in L^1(\nu)$. If ν is not abslutely continuous wrt μ, than the claim is trivially true, so assume $\nu \ll \mu$. Define $f_\mu = \mathbb{E}_\mu[f(Z)]$. We have

$$\begin{aligned}|\mathbb{E}_\mu[f(Z)] - \mathbb{E}_\nu[f(Z)]| \leq \mathbb{E}_\nu[|f(Z) - f_\mu|] = \int_\mathcal{Z} |f(z) - f_\mu| \tfrac{\mathrm{d}\nu}{\mathrm{d}\mu}(z)\mathrm{d}\mu(z) \\ \leq \mathbb{E}_\mu[|f(Z) - f_\mu|^r]^{1/r}(D^{(p)}(\nu\|\mu) + 1)^{1/p} \,,\end{aligned}$$

by Hölder's inequality.

The case $q > 1$ follows form the one-dimesional case, since $\mathbb{E}_\mu[|(f(Z) - f_\mu)\cdot v|^r]^{1/r} \leq \|v\|\mathbb{E}_\mu[\|(f(Z) - f_\mu)\|^r]^{1/r}$ for all $v \in \mathbb{R}^q$. ■

Corollary 33 *Fix $p > 1$ and let $r = p/(p-1)$. Let $\mathfrak{D} : (\mu,\nu) \mapsto (D^{(p)}(\nu\|\mu) + 1)^{1/p}$ and $f : \mathcal{Z} \to \mathbb{R}^q$ be measurable. Assume that $f(Z)$ is ξ-SG for $Z \sim \mu$. Then f has regularity $\mathcal{R}_\mathfrak{D}(e^{1/e}\sqrt{r}\,\xi)$ wrt μ.*

Proof Simply use that $\mathbb{E}_\mu[\|f(Z) - \mathbb{E}_{Z'\sim\mu}[f(Z')]\|^r]^{1/r} \leq e^{1/e}\sqrt{r}\,\xi$ if $f(Z)$ is ξ-SG to conclude by Lemma 32. ■

Lemma 34 *Let $\mathfrak{D} : (\mu,\nu) \mapsto \sqrt{\chi^2(\nu\|\mu)}$. Let $f : \mathcal{Z} \to \mathbb{R}^q$ be measurable. Assume that $\|\mathbb{C}_\mu[f(Z)]\| \leq \xi^2$, where $\mathbb{C}_\mu[f(Z)]$ is the covariance matrix of $f(Z)$ for $Z \sim \mu$. Then, f has regularity $\mathcal{R}_\mathfrak{D}(\xi)$.*

Proof For $q = 1$, the claim is a direct consequence of the HCR bound (Lehmann and Casella, 1998). The case $q > 1$ follows easily. ■

Definition 35 (Total variation) *The total variation of two probability measures $\mu, \nu \in \mathscr{P}$ is defined as*

$$\mathrm{TV}(\mu, \nu) = \sup_{U \in \Sigma_{\mathcal{Z}}} |\mu(U) - \nu(U)| \,.$$

Lemma 36 *Let $\mathfrak{D} : (\mu, \nu) \mapsto 2\mathrm{TV}(\mu, \nu)$. Let $f : \mathcal{Z} \to \mathbb{R}^q$ be a measurable map, bounded in $[-\xi, \xi]$. Then f has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt any $\mu \in \mathscr{P}$.*

Proof First, we need to show that $\nu \mapsto \mathrm{TV}(\mu, \nu)$ is measurable. We have that for all $U \in \Sigma_{\mathcal{Z}}$, the map $\nu \mapsto |\mu(U) - \nu(U)|$ is continuous in the weak topology. In particular, taking the supremum wrt U we get a weakly lower semicontinuous map, which implies the measurability. Now, notice that asking $f \subseteq [-\xi, \xi]$ is equivalent to ask for f to be 2ξ-Lipschitz wrt the discrete metric on $\mathcal{Z}$. We can then proceed as in Lemma 9 using the fact that the total variation coincides with the 1-Wasserstein distance when the transport cost is the discrete metric (Villani, 2009). ■

Corollary 37 *Let $\mathfrak{D} : (\mu, \nu) \mapsto \sqrt{2\mathrm{KL}(\mu \| \nu)}$. Let $f : \mathcal{Z} \to \mathbb{R}^q$ be a measurable map, bounded in $[-\xi, \xi]$. Then f has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$.*

Proof The measurability of $\mathfrak{D}$ is a obvious consequence of Lemma 28. Then, the claim follows directly from Lemma 36 by Pinsker's inequality; see e.g. van Handel (2016). ■

G.2. Some simple bounds based on the $\mathfrak{D}$-regularity

Definition 38 (Power information) *Consider two coupled random variables Z, Z' on $(\mathcal{Z}, \Sigma_{\mathcal{Z}})$. For $p > 1$ we define their p-power information (Guntuboyina et al., 2014) as*

$$I^{(p)}(Z; Z') = D^{(p)}(\mathbb{P}_{Z,Z'} \| \mathbb{P}_{Z \otimes Z'}) \,.$$

Proposition 39 *Fix $p > 1$, let $r = p/(p-1)$ and suppose that $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$. On the one hand, if $\ell(w, X)$ is ξ-SG for $X \sim \mathbb{P}_X$, for all $w \in \mathcal{W}$, then*

$$|\mathcal{G}| \leq \frac{e^{1/e} \sqrt{r}\, \xi}{\sqrt{m}} \left(I^{(p)}(S; W) + 1\right)^{1/p} .$$

On the other hand, under the assumptions $\clubsuit$ if $\nabla_w(\ell, X)$ is ξ-SG for $X \sim \mathbb{P}_X$, for all $w \in \mathcal{W}$, then

$$|\mathcal{G}| \leq \frac{e^{1/e} \sqrt{r}\, \xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} \left(I^{(p)}(S; W_k) + 1\right)^{1/p} .$$

Proof First notice that the ξ-subgaussianity of ℓ (respectively $\nabla_w \ell$) implies that of $\mathscr{L}$ (respectively $\nabla_w \mathscr{L}$) is $(\xi/\sqrt{m})$-SG. Then, the first claim follows by Corollary 33, Theorem 2, and Jensen's inequality, while the second one by Corollary 33, Theorem 4, and Jensen's inequality. ■

Proposition 40 *Suppose that $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$. On the one hand, if $\mathbb{V}_{\mathbb{P}_X}[\ell(w, X)] \leq \xi^2$, for all $w \in \mathcal{W}$, then*

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \mathbb{E}_{\mathbb{P}_W}\left[\sqrt{\chi^2(\mathbb{P}_{S|W} \| \mathbb{P}_S)}\right] .$$

On the other hand, under the assumptions $\clubsuit$ if $\|\mathbb{C}_{\mathbb{P}_X}[\nabla_w \ell(w, X)]\| \leq \xi^2$, for all $w \in \mathcal{W}$, then

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_W}\left[\sqrt{\chi^2(\mathbb{P}_{S|W_k} \| \mathbb{P}_S)}\right] .$$

Proof The claims follow combining Lemma 34 with Theorems 2 and 4. Note that the variance of $\mathscr{L}$ is re-scaled by a factor $1/\sqrt{m}$ wrt the one of ℓ, as $\mathbb{P}_S = \mathbb{P}_X^{\otimes m}$. The same is true for the covariance of $\nabla_w \mathscr{L}$. ■

G.3. Individual-sample bounds

Recall that $S = \{X_1, \ldots, X_m\}$. In this section we will consider a probability measure $\mathbb{P}_S$ on $(\mathcal{S}, \Sigma_{\mathcal{S}})$ such that the marginals $\mathbb{P}_{X_i} = \mathbb{P}_X$ for all $i \in [1 : m]$, but we do not require that the draws are independent. Note moreover that W might depend in a different way on each X_i, so that we can have that $\mathbb{P}_{W,X_i} \neq \mathbb{P}_{W,X_j}$, for $i \neq j$. Now, we specialise Theorems 2 and 4 to obtain individual-sample bounds, such as those from Bu et al. (2019).

Proposition 41 *Assume that $x \mapsto \ell(w, x)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt $\mathbb{P}_X$, $\forall w \in \mathcal{W}$. Then we have*

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_X, \mathbb{P}_{X_i|W})] .$$

Proof Just write

$$\mathcal{G} = \frac{1}{m} \sum_{i=1}^{m} (\mathbb{E}_{\mathbb{P}_{W \otimes X}}[\ell(W, X)] - \mathbb{E}_{\mathbb{P}_{W,X_i}}[\ell(W, X_i)]) .$$

and then conclude by applying Theorem 20 to bound each term of the sum. ■

Proposition 42 *Assume $\clubsuit$ and suppose that $x \mapsto \ell(w, x)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt $\mathbb{P}_X$, $\forall w \in \mathcal{W}$. Then we have*

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_W}[\mathfrak{D}(\mathbb{P}_X, \mathbb{P}_{X_i|W_k})] .$$

Proof Just write

$$\mathcal{G} = \frac{1}{m} \sum_{i=1}^{m} (\mathbb{E}_{\mathbb{P}_{W \otimes X}}[\ell(W, X)] - \mathbb{E}_{\mathbb{P}_{W,X_i}}[\ell(W, X_i)]) .$$

and then conclude by applying Theorem 21 to bound each term of the sum. ■

We can now state several individual-sample generalisation bounds. For the sake of brevity, we will omit the proofs, as they are all direct applications of Propositions 41 and 42, and of the previously established results of $\mathfrak{D}$-regularity.

Proposition 43 *On the one hand, if $\ell(w, X)$ is ξ-SG uniformly on $\mathcal{W}$, then*

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \sqrt{2I(W; X_i)}\,.$$

On the other hand, if $\nabla_w \ell(w, X)$ is ξ-SG uniformly on $\mathcal{W}$, then

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} \sqrt{2I(W_k; X_i)}\,.$$

Proposition 44 *On the one hand, if $x \mapsto \ell(w, x)$ is ξ-Lipschitz uniformly on $\mathcal{W}$, then*

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} E_{\mathbb{P}_W} [\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X_i|W})]\,.$$

On the other hand, assume ♣. if $x \mapsto \nabla_w \ell(w, x)$ is ξ-Lipschitz uniformly on $\mathcal{W}$, then

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} E_{\mathbb{P}_W} [\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X_i|W_k})]\,.$$

Proposition 45 *Fix $p > 1$ and let $r = p/(p-1)$. Write $\bar{\ell}(w)$ for $\mathbb{E}_{\mathbb{P}_X}[\ell(w, X)]$ and $\overline{\nabla_w \ell}(w)$ for $\mathbb{E}_{\mathbb{P}_X}[\nabla_w \ell(w, X)]$. On the one hand, if, for all $w \in \mathcal{W}$, $\mathbb{E}_{\mathbb{P}_X}[|\ell(w, X) - \bar{\ell}(w)|^r] \leq \xi^r$, then*

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} (I^{(p)}(W; X_i) + 1)^{1/p}\,.$$

On the other hand, assume ♣. If $\mathbb{E}_{\mathbb{P}_X}[\|\nabla_w \ell(w, X) - \overline{\nabla_w \ell}(w)\|^r] \leq \xi^r$, for all $w \in \mathcal{W}$, then

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} (I^{(p)}(W_k; X_i) + 1)^{1/p}\,.$$

Proposition 46 *On the one hand, if, for all $w \in \mathcal{W}$, $\mathbb{V}_{\mathbb{P}_X}[\ell(w, X)] \leq \xi^2$, then*

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathbb{P}_W} \left[\sqrt{\chi^2(\mathbb{P}_{X_i|W} \| \mathbb{P}_X)} \right]\,.$$

On the other hand, assume ♣. If $\|\mathbb{C}_{\mathbb{P}_X}[\nabla_w \ell(w, X)]\| \leq \xi^2$, for all $w \in \mathcal{W}$, then

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_W} \left[\sqrt{\chi^2(\mathbb{P}_{X_i|W_k} \| \mathbb{P}_X)} \right]\,.$$

Proposition 47 *On the one hand, if $|\ell(w, x)| \leq \xi$ for all $w \in \mathcal{W}$ and all $x \in \mathcal{X}$, then*

$$|\mathcal{G}| \leq \frac{2\xi}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathbb{P}_W} \left[\mathrm{TV}(\mathbb{P}_X, \mathbb{P}_{X_i|W}) \right]\,.$$

On the other hand, assume ♣. If $\|\nabla_w \ell(w, x)\| \leq \xi$ for all $w \in \mathcal{W}$ and all $x \in \mathcal{X}$, then

$$|\mathcal{G}| \leq \frac{2\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_W} \left[\mathrm{TV}(\mathbb{P}_X, \mathbb{P}_{X_i|W_k}) \right]\,.$$

Definition 48 (Lautum information) *Consider two coupled random variables Z, Z' on $(\mathcal{Z}, \Sigma_\mathcal{Z})$. We define their lautum information (Palomar and Verdú, 2008) as*

$$\mathrm{L}(Z; Z') = \mathrm{KL}(\mathbb{P}_{Z\otimes Z'} \| \mathbb{P}_{Z,Z'})\,.$$

Proposition 49 *On the one hand, if $|\ell(w,x)| \leq \xi$ for all $w \in \mathcal{W}$ and all $x \in \mathcal{X}$, then*

$$|\mathcal{G}| \leq \frac{\xi}{m}\sum_{i=1}^{m}\sqrt{2\mathrm{L}(W;X_i)}\,.$$

On the other hand, assume ♣. If $\|\nabla_w \ell(w,x)\| \leq \xi$ for all $w \in \mathcal{W}$ and all $x \in \mathcal{X}$, then

$$|\mathcal{G}| \leq \frac{\xi}{m}\sum_{i=1}^{m}\sum_{k=1}^{\infty}\varepsilon_{k-1}\sqrt{2\mathrm{L}(W_k;X_i)}\,.$$

G.4. Bounds based on random sub-sampling from a super-sample

We can derive in our framework bounds in the same spirit of the conditional MI bound from Steinke and Zakynthinou (2020).

Let $s^\star = (x^\star_1, \dots, x^\star_m)$ denote a $(2m)$-sample, made of m pairs $x^\star_i = (x_{i,0}, x_{i,1})$. The training sample is in the form $s = (x_1, \dots, x_m)$. The choice of s, given $s^\star$ is determined by a variable $u \in \{0,1\}^n$, in the sense that $x_i = x^\star_{i,u_i}$, where u_i determine which one of the two components of $x^\star_i$ is chosen as x_i. In practice we can write $s = s^\star_u$, with $u \in \{0,1\}^n$. We let $\bar{u} = 1 - u$ (the difference being component-wise), and $\bar{s} = s^\star_{\bar{u}}$. We denote as $S^\star$ the random super-sample and we assume that each $X^\star_i \in S^\star$ has marginal distribution $\mathbb{P}_{X^\star} = \mathbb{P}_X^{\otimes 2}$. Morover, we let $\mathbb{P}_{\bar{U}} = \mathbb{P}_U \sim \mathrm{Bernoulli}(\frac{1}{2})^{\otimes m}$, and we assume that $U \perp\!\!\!\perp S^\star$. Note that this implies that if the super-sample is made of independent pairs ($\mathbb{P}_{S^\star} = \mathbb{P}_{X^\star}^{\otimes m}$) then all the $X_i \in S$ are independent.

Proposition 50 *Let $\mathbb{P}_{S^\star} = \mathbb{P}_{X^\star}^{\otimes m}$. Assume that $s \mapsto \mathscr{L}_s(w)$ has regularity $\mathcal{R}_\mathfrak{D}(\xi)$, wrt $\mathbb{P}_{S|S^\star=s^\star}$, for $\mathbb{P}_{S^\star}$-almost every $s^\star$ and $\forall w \in \mathcal{W}$. Then, we have that*

$$|\mathcal{G}| \leq \xi\,\mathbb{E}_{\mathbb{P}_{W,S^\star}}[\mathfrak{D}(\mathbb{P}_{S|W,S^\star}, \mathbb{P}_{S|S^\star}) + \mathfrak{D}(\mathbb{P}_{\bar{S}|W,S^\star}, \mathbb{P}_{\bar{S}|S^\star})]\,.$$

Proof Let $\hat{g}(w, s^\star, u) = \mathscr{L}_{s^\star_{\bar{u}}}(w) - \mathscr{L}_{s^\star_u}(w)$. Now, recalling that $S = S^\star_U$ and $\bar{S} = S^\star_{\bar{U}}$, we have that $\mathbb{P}_{S|S^\star}$ is the law of $S^\star_U$ and $\mathbb{P}_{\bar{S}|S^\star}$ is the law of $S^\star_{\bar{U}}$, both under $\mathbb{P}_U$ and given $S^\star$. Since $\mathbb{P}_{S^\star} = \mathbb{P}_{X^\star}^{\otimes m} = \mathbb{P}_X^{\otimes 2m}$, then $S \perp\!\!\!\perp \bar{S}$. In particular $\bar{S} \perp\!\!\!\perp W$, and hence $\mathbb{P}_{\bar{S}|W} = \mathbb{P}_{\bar{S}} = \mathbb{P}_S$, so that

$$\mathbb{E}_{\mathbb{P}_{W,S^\star,U}}[\mathscr{L}_{S^\star_{\bar{U}}}(W)] = \mathbb{E}_{\mathbb{P}_{W,\bar{S}}}[\mathscr{L}_{\bar{S}}(W)] = \mathbb{E}_{\mathbb{P}_{W\otimes S}}[\mathscr{L}_S(W)]\,.$$

It follows that $\mathcal{G} = \mathbb{E}_{\mathbb{P}_{W,S^\star,U}}[\hat{g}(W,S^\star,U)]$. Moreover, it is shown in Rodríguez-Gálvez et al. (2021) (cf. proof of Theorem 3 therein) that

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_{W,S^\star,U}}[\hat{g}(W,S^\star,U)] = \mathbb{E}_{\mathbb{P}_{S^\star}}\big[\mathbb{E}_{\mathbb{P}_{W\otimes U|S^\star}}[\mathscr{L}_{S^\star_U}(W)] - \mathbb{E}_{\mathbb{P}_{W,U|S^\star}}[\mathscr{L}_{S^\star_U}(W)]\big]\\ - \mathbb{E}_{\mathbb{P}_{S^\star}}[\mathbb{E}_{\mathbb{P}_{W,U|S^\star}}[\mathscr{L}_{S^\star_{\bar{U}}}(W)] - \mathbb{E}_{\mathbb{P}_{W\otimes U|S^\star}}[\mathscr{L}_{S^\star_{\bar{U}}}(W)]]\,.\end{aligned}$$

We hence have

$$\begin{aligned}|\mathcal{G}| \leq \mathbb{E}_{\mathbb{P}_{S^\star}}\Big[\big|\mathbb{E}_{\mathbb{P}_{W\otimes U|S^\star}}[\mathscr{L}_{S^\star_U}(W)] - \mathbb{E}_{\mathbb{P}_{W,U|S^\star}}[\mathscr{L}_{S^\star_U}(W)]\big|\\ + \big|\mathbb{E}_{\mathbb{P}_{W\otimes U|S^\star}}[\mathscr{L}_{S^\star_{\bar{U}}}(W)] - \mathbb{E}_{\mathbb{P}_{W,U|S^\star}}[\mathscr{L}_{S^\star_{\bar{U}}}(W)]\big|\Big]\,,\end{aligned}$$

which can be rewritten as

$$|\mathcal{G}| \leq \mathbb{E}_{\mathbb{P}_{W,S^\star}}\left[|\mathbb{E}_{\mathbb{P}_{S|S^\star}}[\mathscr{L}_S(W)] - \mathbb{E}_{\mathbb{P}_{S|W,S^\star}}[\mathscr{L}_S(W)]| + |\mathbb{E}_{\mathbb{P}_{\bar{S}|S^\star}}[\mathscr{L}_{\bar{S}}(W)] - \mathbb{E}_{\mathbb{P}_{\bar{S}|W,S^\star}}[\mathscr{L}_{\bar{S}}(W)]| \right] . \tag{7}$$

Now, notice that, since $\mathbb{P}_U = \mathbb{P}_{\bar{U}}$, we have $\mathbb{P}_{S|S^\star} = \mathbb{P}_{\bar{S}|S^\star}$. In particular, $s \mapsto \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$ wrt $\mathbb{P}_{\bar{S}|S^\star = s^\star}$ as well ($\forall w \in \mathcal{W}$ and $\mathbb{P}_{S^\star}$-a.s.). From (7) and Theorem 2, we have that

$$|\mathcal{G}| \leq \xi \, \mathbb{E}_{\mathbb{P}_{W,S^\star}}[\mathfrak{D}(\mathbb{P}_{S|W,S^\star}, \mathbb{P}_{S|S^\star}) + \mathfrak{D}(\mathbb{P}_{\bar{S}|W,S^\star}, \mathbb{P}_{\bar{S}|S^\star})] ,$$

as requested. ■

Proposition 51 *Let $\mathbb{P}_{S^\star} = \mathbb{P}_{X^\star}^{\otimes m}$. Assume ♣ and suppose that $s \mapsto \nabla_w \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$, wrt $\mathbb{P}_{S|S^\star = s^\star}$, for $\mathbb{P}_{S^\star}$-almost every $s^\star$ and $\forall w \in \mathcal{W}$. Then, we have that*

$$|\mathcal{G}| \leq \xi \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_{W,S^\star}}[\mathfrak{D}(\mathbb{P}_{S|W_k,S^\star}, \mathbb{P}_{S|S^\star}) + \mathfrak{D}(\mathbb{P}_{\bar{S}|W_k,S^\star}, \mathbb{P}_{\bar{S}|S^\star})] .$$

Proof We proceed just as in the proof on Proposition 50 until the last step, where we use Theorem 4, instead of Theorem 2, to conclude. ■

We give now some explicit example of bounds that can be obtained via the above two propositions.

Definition 52 (Conditional mutual information, power information, and lautum information) *Let (Z, Z', W) be a random variable on $(\mathcal{Z} \times \mathcal{Z} \times \mathcal{W}, \Sigma_{\mathcal{Z}} \otimes \Sigma_{\mathcal{Z}} \otimes \Sigma_W)$. We define the conditional MI (Wyner, 1978) as*

$$I(Z; Z'|W) = \mathbb{E}_{\mathbb{P}_W}[\mathrm{KL}(\mathbb{P}_{Z,Z'|W} \| \mathbb{P}_{Z \otimes Z'|W})] .$$

For $p > 1$, we define the conditional p-power information as

$$I^{(p)}(Z; Z'|W) = \mathbb{E}_{\mathbb{P}_W}[D^{(p)}(\mathbb{P}_{Z,Z'|W} \| \mathbb{P}_{Z \otimes Z'|W})] .$$

Finally, we define the conditional Lautum information (Palomar and Verdú, 2008) as

$$\mathrm{L}(Z; Z|W) = \mathbb{E}_{\mathbb{P}_W}[\mathrm{KL}(\mathbb{P}_{Z \otimes Z'|W} \| \mathbb{P}_{Z,Z'|W})] .$$

Proposition 53 *Let $\mathbb{P}_{S^\star} = \mathbb{P}_{X^\star}^{\otimes m}$. On the one hand, assume that $|\ell(w, x)| \leq \xi$, for all $w \in \mathcal{W}$ and all $x \in X$. Then, we have that*

$$|\mathcal{G}| \leq 2\xi \sqrt{\frac{2I(W; S|S^\star)}{m}} .$$

On the other hand, assume ♣ and suppose that $\|\nabla_w \ell(w, x)\| \leq \xi$, for all $w \in \mathcal{W}$ and all $x \in \mathcal{X}$. Then we have

$$|\mathcal{G}| \leq 2\xi \sum_{k=1}^{\infty} \varepsilon_{k-1} \sqrt{\frac{2I(W_k; S|S^\star)}{m}} .$$

Proof Assume that $|\ell| \leq \xi$. Note that $\ell(w, X)$ is ξ-SG, for all $w \in \mathcal{W}$, for $X \sim \mathbb{P}_{X|X^\star = x^\star}$ (for all $x^\star$). As the elements of S are independent (even when conditioning on $S^\star$ since $U \perp\!\!\!\perp S^\star$), we have that, $\forall w \in \mathcal{W}$ and $\forall s^\star \in \mathcal{S}^2$, $\mathscr{L}_S(w)$ is $(\xi/\sqrt{m})$-SG for $S \sim \mathbb{P}_{S|S^\star = s^\star}$. We can then conclude by Lemma 9 and Proposition 50, using the fact that $I(W; S|S^\star) = I(W; \bar{S}|S^\star)$, as $\bar{s}$ is fully determined by s (given $s^\star$). The proof for the chained bound is analogous. ■

The proofs for the next propositions are essentially analogous of the one of Proposition 53 and hence are omitted.

Proposition 54 *Let $\mathbb{P}_{S^\star} = \mathbb{P}_{X^\star}^{\otimes m}$ and assume that $d_\mathcal{X}$ and $d_\mathcal{S}$ are related by (1). On the one hand, suppose that $x \mapsto \ell(w, x)$ is ξ-Lipschitz, for all $w \in \mathcal{W}$. Then, we have that*

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \mathbb{E}_{\mathbb{P}_{W,S^\star}} [\mathfrak{W}(\mathbb{P}_{S|S^\star}, \mathbb{P}_{S|W,S^\star}) + \mathfrak{W}(\mathbb{P}_{\bar{S}|S^\star}, \mathbb{P}_{\bar{S}|W,S^\star})] \,.$$

On the other hand, assume $\clubsuit$ and suppose that $x \mapsto \nabla_w \ell(w, x)\| \leq \xi$, for all $w \in \mathcal{W}$ and all $x \in \mathcal{X}$. Then we have

$$|\mathcal{G}| \leq \frac{\xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_{W,S^\star}} [\mathfrak{W}(\mathbb{P}_{S|S^\star}, \mathbb{P}_{S|W_k,S^\star}) + \mathfrak{W}(\mathbb{P}_{\bar{S}|S^\star}, \mathbb{P}_{\bar{S}|W_k,S^\star})] \,.$$

Proposition 55 *Fix $p > 1$, let $r = p/(p-1)$ and suppose that $\mathbb{P}_{S^\star} = \mathbb{P}_{X^\star}^{\otimes m}$. On the one hand, assume that $|\ell(w, x)| \leq \xi$, for all $w \in \mathcal{W}$ and all $x \in X$. Then, we have that*

$$|\mathcal{G}| \leq \frac{2e^{1/e}\sqrt{r}\,\xi}{\sqrt{m}} (I^{(p)}(W; S|S^\star) + 1)^{1/p} \,.$$

On the other hand, assume $\clubsuit$ and suppose that $\|\nabla_w \ell(w, x)\| \leq \xi$, for all $w \in \mathcal{W}$ and all $x \in \mathcal{X}$. Then we have

$$|\mathcal{G}| \leq \frac{2e^{1/e}\sqrt{r}\,\xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} (I^{(p)}(W_k; S|S^\star) + 1)^{1/p} \,.$$

Proposition 56 *Suppose that $\mathbb{P}_{S^\star} = \mathbb{P}_{X^\star}^{\otimes m}$. On the one hand, if $|\ell(w, x)| \leq \xi$, for all $w \in \mathcal{W}$ and all $x \in X$, then*

$$|\mathcal{G}| \leq \frac{2\xi}{\sqrt{m}} \mathbb{E}_{\mathbb{P}_W, S^\star} \left[\sqrt{\chi^2(\mathbb{P}_{S|W,S^\star} \| \mathbb{P}_{S|S^\star})} + \sqrt{\chi^2(\mathbb{P}_{\bar{S}|W,S^\star} \| \mathbb{P}_{\bar{S}|S^\star})} \right] \,.$$

On the other hand, under the assumptions $\clubsuit$ if $\|\nabla_w \ell(w, x)\| \leq \xi$, for all $w \in \mathcal{W}$ and all $x \in \mathcal{X}$, then

$$|\mathcal{G}| \leq \frac{2\xi}{\sqrt{m}} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_{W,S^\star}} \left[\sqrt{\chi^2(\mathbb{P}_{S|W_k,S^\star} \| \mathbb{P}_{S|S^\star})} + \sqrt{\chi^2(\mathbb{P}_{\bar{S}|W_k,S^\star} \| \mathbb{P}_{\bar{S}|S^\star})} \right] \,.$$

One issue with this random sub-sampling approach is that in order to controll $\mathscr{L}_s$ wrt $\mathbb{P}_{S|S^\star = s^\star}$, almost uniformly in $s^\star$, one needs essentially to control the random binary variables $\ell(w, X^\star)$ under $\mathbb{P}_{X|X^\star = (x_0^\star, x_1^\star)}$ (that is $X^\star = x_0^\star$ with probability $1/2$, and $x_1^\star$ with probability $1/2$). This can be easily done in the case of the Wasserstein distance, as the Lipschitzianity guarantees $\mathfrak{W}$-regularity wrt any measure. However for the subgaussianity things are more complicated, and one essentially needs to ask that ℓ is bounded.

It is however possible to restate Proposition 50 (and Proposition 51) without asking that the same regularity holds $\mathbb{P}_{S^\star}$-a.s. The proof of both results follow closely the ones of Propositions 50 and 51, the only difference being a final application of Hölder's inequality.

Proposition 57 *Let $\mathbb{P}_{S^\star} = \mathbb{P}_{X^\star}^{\otimes m}$. Let $p \in [1,+\infty]$ and $r = p/(p-1)$ (with the convention that $1/0 = +\infty$). Assume that $s \mapsto \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi_{s^\star})$, wrt $\mathbb{P}_{S|S^\star=s^\star}$, for $\mathbb{P}_{S^\star}$-almost every $s^\star$ and $\forall w \in \mathcal{W}$, where $\|\xi_{S^\star}\|_{L^p(\mathbb{P}_{S^\star})} = \xi$. Then, we have that*

$$|\mathcal{G}| \leq \xi\, \mathbb{E}_{\mathbb{P}_{W,S^\star}}[|\mathfrak{D}(\mathbb{P}_{S|W,S^\star}, \mathbb{P}_{S|S^\star}) + \mathfrak{D}(\mathbb{P}_{\bar{S}|W,S^\star}, \mathbb{P}_{\bar{S}|S^\star})|^r]^{1/r} \,.$$

Proposition 58 *Let $\mathbb{P}_{S^\star} = \mathbb{P}_{X^\star}^{\otimes m}$. Let $p \in [1,+\infty]$ and $r = p/(p-1)$ (with the convention that $1/0 = +\infty$). Assume ♣ and suppose that $s \mapsto \nabla_w \mathscr{L}_s(w)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi_{s^\star})$, wrt $\mathbb{P}_{S|S^\star=s^\star}$, for $\mathbb{P}_{S^\star}$-almost every $s^\star$ and $\forall w \in \mathcal{W}$, where $\|\xi_{S^\star}\|_{L^p(\mathbb{P}_{S^\star})} = \xi$. Then, we have that*

$$|\mathcal{G}| \leq \xi \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_{W,S^\star}}[|\mathfrak{D}(\mathbb{P}_{S|W_k,S^\star}, \mathbb{P}_{S|S^\star}) + \mathfrak{D}(\mathbb{P}_{\bar{S}|W_k,S^\star}, \mathbb{P}_{\bar{S}|S^\star})|^r]^{1/r} \,.$$

G.5. Individual-sample bounds based on random sub-sampling

We can merge together the ideas of the last two sections.

Proposition 59 *Assume that $x \mapsto \ell(w,x)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$, wrt $\mathbb{P}_{X|X^\star=x^\star}$, for $\mathbb{P}_{X^\star}$-almost every $x^\star$ and $\forall w \in \mathcal{W}$. Then, we have that*

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathbb{P}_{W,X_i^\star}}[\mathfrak{D}(\mathbb{P}_{X_i|W,X_i^\star}, \mathbb{P}_{X_i|X_i^\star}) + \mathfrak{D}(\mathbb{P}_{\bar{X}_i|W,X_i^\star}, \mathbb{P}_{\bar{X}_i|X_i^\star})] \,.$$

Proof Note that $\mathbb{P}_{X|X^\star=x^\star} = \mathbb{P}_{X_i|X_i^\star=x^\star}$. Proceeding as in the proof of Proposition 50, we can show that, for $i \in [1:m]$,

$$\begin{aligned} |\mathbb{E}_{\mathbb{P}_{W\otimes X_i}}[\ell(W,X_i)] - \mathbb{E}_{\mathbb{P}_{W,X_i}}[\ell(W,X_i)]| \\ \leq \mathbb{E}_{\mathbb{P}_{W,X_i^\star}}[\mathfrak{D}(\mathbb{P}_{X_i|W,X_i^\star}, \mathbb{P}_{X_i|X_i^\star}) + \mathfrak{D}(\mathbb{P}_{\bar{X}_i|W,X_i^\star}, \mathbb{P}_{\bar{X}_i|X_i^\star})] \,. \end{aligned}$$

We can immediately conclude by writing $\mathcal{G}$ as in the proof of Proposition 41. ∎

Proposition 60 *Assume ♣ and suppose that $x \mapsto \nabla_w \ell(w,x)$ has regularity $\mathcal{R}_{\mathfrak{D}}(\xi)$, wrt $\mathbb{P}_{X|X^\star=x^\star}$, for $\mathbb{P}_{X^\star}$-almost every $x^\star$ and $\forall w \in \mathcal{W}$. Then, we have that*

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_{X_i^\star,W}}[\mathfrak{D}(\mathbb{P}_{X_i|W_k,X_i^\star}, \mathbb{P}_{X_i|X_i^\star}) + \mathfrak{D}(\mathbb{P}_{\bar{X}_i|W_k,X_i^\star}, \mathbb{P}_{\bar{X}_i|X_i^\star})] \,.$$

Proof We proceed as for proving Proposition 59, but following the proof Proposition 51 instead of 50. ∎

Clearly one can generalise the two results above by using the same observations as in Propositions 57 and 58.

We can now restate all the individual-sample bounds from Section G.3 in the random sub-sampling framework.

Proposition 61 *On the one hand, if $|\ell(w,x)| \leq \xi$, uniformly on $\mathcal{W}$ and $\mathcal{X}$, then*

$$|\mathcal{G}| \leq \frac{2\xi}{m} \sum_{i=1}^{m} \sqrt{2I(W;X_i|X_i^\star)}\,.$$

On the other hand, if $|\nabla_w \ell(w,x)| \leq \xi$, uniformly on $\mathcal{W}$ and $\mathcal{X}$, then

$$|\mathcal{G}| \leq \frac{2\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} \sqrt{2I(W_k;X_i|X_i^\star)}\,.$$

Proposition 62 *On the one hand, if $x \mapsto \ell(w,x)$ is ξ-Lipschitz uniformly on $\mathcal{W}$, then*

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathbb{P}_{W,X_i^\star}}[\mathfrak{W}(\mathbb{P}_{X_i|X_i^\star}, \mathbb{P}_{X_i|W,X_i^\star}) + \mathfrak{W}(\mathbb{P}_{\bar{X}_i|X_i^\star}, \mathbb{P}_{\bar{X}_i|W,X_i^\star})]\,.$$

On the other hand, assume ♣. if $x \mapsto \nabla_w \ell(w,x)$ is ξ-Lipschitz uniformly on $\mathcal{W}$, then

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_{W,X_i^\star}}[\mathfrak{W}(\mathbb{P}_{X_i|X_i^\star}, \mathbb{P}_{X_i|W_k,X_i^\star}) + \mathfrak{W}(\mathbb{P}_{\bar{X}_i|X_i^\star}, \mathbb{P}_{\bar{X}_i|W_k,X_i^\star})]\,.$$

Proposition 63 *Fix $p > 1$ and let $r = p/(p-1)$. On the one hand, if $|\ell(w,x)| \leq \xi$, uniformly on $\mathcal{W}$ and $\mathcal{X}$, then*

$$|\mathcal{G}| \leq \frac{2\xi}{m} \sum_{i=1}^{m} (I^{(p)}(W;X_i|X_i^\star) + 1)^{1/p}\,.$$

On the other hand, assume ♣. If $|\nabla_w \ell(w,x)| \leq \xi$, uniformly on $\mathcal{W}$ and $\mathcal{X}$, then

$$|\mathcal{G}| \leq \frac{2\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} (I^{(p)}(W_k;X_i|X_i^\star) + 1)^{1/p}\,.$$

Proposition 64 *On the one hand, if $|\ell(w,x)| \leq \xi$, uniformly on $\mathcal{W}$ and $\mathcal{X}$, then*

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathbb{P}_{W,X_i^\star}} \left[\sqrt{\chi^2(\mathbb{P}_{X_i|W,X_i^\star} \| \mathbb{P}_{X_i|X_i^\star})} + \sqrt{\chi^2(\mathbb{P}_{\bar{X}_i|W,X_i^\star} \| \mathbb{P}_{\bar{X}_i|X_i^\star})}\right]\,.$$

On the other hand, assume ♣. If $|\nabla_w \ell(w,x)| \leq \xi$, uniformly on $\mathcal{W}$ and $\mathcal{X}$, then

$$|\mathcal{G}| \leq \frac{\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} \mathbb{E}_{\mathbb{P}_{W,X_i^\star}} \left[\sqrt{\chi^2(\mathbb{P}_{X_i|W_k,X_i^\star} \| \mathbb{P}_{X_i|X_i^\star})} + \sqrt{\chi^2(\mathbb{P}_{\bar{X}_i|W_k,X_i^\star} \| \mathbb{P}_{\bar{X}_i|X_i^\star})}\right]\,.$$

Proposition 65 *On the one hand, if $|\ell(w,x)| \leq \xi$, uniformly on $\mathcal{W}$ and $\mathcal{X}$, then*

$$|\mathcal{G}| \leq \frac{2\xi}{m} \sum_{i=1}^{m} \mathbb{E}_{\mathbb{P}_{W,X_i^\star}} \left[\mathrm{TV}(\mathbb{P}_{X_i|X_i^\star}, \mathbb{P}_{X_i|W,X_i^\star}) + \mathrm{TV}(\mathbb{P}_{\bar{X}_i|X_i^\star}, \mathbb{P}_{\bar{X}_i|W,X_i^\star})\right]\,.$$

On the other hand, assume ♣. If $|\nabla_w \ell(w,x)| \leq \xi$, uniformly on $\mathcal{W}$ and $\mathcal{X}$, then

$$|\mathcal{G}| \leq \frac{2\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} \left[\mathrm{TV}(\mathbb{P}_{X_i|X_i^\star}, \mathbb{P}_{X_i|W_k,X_i^\star}) + \mathrm{TV}(\mathbb{P}_{\bar{X}_i|X_i^\star}, \mathbb{P}_{\bar{X}_i|W_k,X_i^\star})\right]\,.$$

Proposition 66 *On the one hand, if $|\ell(w, x)| \leq \xi$, uniformly on $\mathcal{W}$ and $\mathcal{X}$, then*

$$|\mathcal{G}| \leq \frac{2\xi}{m} \sum_{i=1}^{m} \sqrt{2\mathrm{L}(W; X_i | X_i^\star)}\,.$$

On the other hand, assume $\clubsuit$. If $|\nabla_w \ell(w, x)| \leq \xi$, uniformly on $\mathcal{W}$ and $\mathcal{X}$, then

$$|\mathcal{G}| \leq \frac{2\xi}{m} \sum_{i=1}^{m} \sum_{k=1}^{\infty} \varepsilon_{k-1} \sqrt{2\mathrm{L}(W_k; X_i | X_i^\star)}\,.$$

G.6. Summary table

Several explicit bounds that can be derived within our general framework of Section 3 are reported in Table 1. The first column states the regularity condition required on the loss. However, we refer to the corresponding propositions for the detailed assumptions of each bound. All bounds are stated for $\xi = 1$. The last columns give the literature references for each bound, to the best of our knowledge. However, this bibliography should be taken as a mere guideline, as there might possibly be missing references. Those bounds that we could not find in the literature are marked as “New”.

Table 1: Some bounds that can be derived with the framework from Section 3

Assumption ($\forall w \in \mathcal{W}$)	Bound	Prop	Ref
$\ell(w, X)$ 1-SG	$\sqrt{2I(W;S)/m}$	10	Russo and Zou (2019)
$\nabla_w \ell(w, X)$ 1-SG	$\sum_k \varepsilon_{k-1}\sqrt{2I(W_k;S)/m}$	13	Asadi et al. (2018)
$\ell(w, \cdot)$ 1-Lipschitz	$\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S\mid W})]/\sqrt{m}$	11	Lopez and Jog (2018)
$\nabla_w \ell(w, \cdot)$ 1-Lipschitz	$\sum_k \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_S, \mathbb{P}_{S\mid W_k})]/\sqrt{m}$	15	New
$\ell(w, X)$ 1-SG	$e^{1/e}\sqrt{p}(I^{(p)}(W;S)+1)^{1/p}/\sqrt{m(p-1)}$	39	Aminian et al. (2021)
$\nabla_w \ell(w, X)$ 1-SG	$e^{1/e}\sqrt{p}\sum_k \varepsilon_{k-1}(I^{(p)}(W_k;S)+1)^{1/p}/\sqrt{m(p-1)}$	39	New
$\mathbb{V}_{\mathbb{P}_X}[\ell(w, X)] \leq 1$	$\mathbb{E}_{\mathbb{P}_W}[\chi^2(\mathbb{P}_{S\mid W}\Vert\mathbb{P}_S)^{1/2}]/\sqrt{m}$	40	Rodríguez-Gálvez et al. (2021)
$\Vert\mathbb{C}_{\mathbb{P}_X}[\nabla_w \ell(w, X)]\Vert \leq 1$	$\sum_k \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_W}[\chi^2(\mathbb{P}_{S\mid W}\Vert\mathbb{P}_S)^{1/2}]/\sqrt{m}$	40	New
$\ell(w, X)$ 1-SG	$\sum_i \sqrt{2I(W;X_i)}/m$	43	Bu et al. (2019)
$\nabla_w \ell(w, X)$ 1-SG	$\sum_i \sum_k \varepsilon_{k-1}\sqrt{2I(W_k;X_i)}/m$	43	Zhou et al. (2022)
$\ell(w, \cdot)$ 1-Lipschitz	$\sum_i \mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X_i\mid W})]/m$	44	Rodríguez-Gálvez et al. (2021)
$\nabla_w \ell(w, \cdot)$ 1-Lipschitz	$\sum_i \sum_k \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_W}[\mathfrak{W}(\mathbb{P}_X, \mathbb{P}_{X_i\mid W_k})]/m$	44	New
$\mathbb{E}_{\mathbb{P}_X}[\vert\ell(w, X) - \bar{\ell}(w)\vert^{p/(p-1)}] \leq 1$	$\sum_i (I^{(p)}(W;X_i)+1)^{1/p}/m$	45	New
$\mathbb{E}_{\mathbb{P}_X}[\Vert\nabla_w \ell(w, X_i) - \overline{\nabla_w \ell}(w)\Vert^{p/(p-1)}] \leq 1$	$\sum_i \sum_k \varepsilon_{k-1}(I^{(p)}(W_k;X_i)+1)^{1/p}/m$	45	New
$\mathbb{V}_{\mathbb{P}_X}[\ell(w, X)] \leq 1$	$\sum_i \mathbb{E}_{\mathbb{P}_W}[\chi^2(\mathbb{P}_{X_i\mid W}\Vert\mathbb{P}_X)^{1/2}]/m$	46	New
$\Vert\mathbb{C}_{\mathbb{P}_X}[\nabla_w \ell(w, X)]\Vert \leq 1$	$\sum_i \sum_k \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_W}[\chi^2(\mathbb{P}_{X_i\mid W_k}\Vert\mathbb{P}_X)^{1/2}]/m$	46	New
$\vert\ell\vert \leq 1$	$\sum_i \mathbb{E}_{\mathbb{P}_W}[\mathrm{TV}(\mathbb{P}_X, \mathbb{P}_{X_i\mid W})]/m$	47	Rodríguez-Gálvez et al. (2021)
$\Vert\nabla_w \ell\Vert \leq 1$	$\sum_i \sum_k \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_W}[\mathrm{TV}(\mathbb{P}_X, \mathbb{P}_{X_i\mid W_k})]/m$	47	New
$\vert\ell\vert \leq 1$	$\sum_i \sqrt{2\mathrm{L}(W;X_i)}/m$	49	Rodríguez-Gálvez et al. (2021)
$\Vert\nabla_w \ell\Vert \leq 1$	$\sum_i \sum_k \varepsilon_{k-1}\sqrt{2\mathrm{L}(W_k;X_i)}/m$	49	New
$\vert\ell\vert \leq 1$	$2\sqrt{2I(W;S\mid S^\star)/m}$	53	Steinke and Zakynthinou (2020)
$\Vert\nabla_w \ell\Vert \leq 1$	$2\sum_k \varepsilon_{k-1}\sqrt{2I(W_k;S\mid S^\star)/m}$	53	New
$\ell(w, \cdot)$ 1-Lipschitz	$\mathbb{E}_{\mathbb{P}_{W,S^\star}}[\mathfrak{W}(\mathbb{P}_{S\mid S^\star}, \mathbb{P}_{S\mid W,S^\star}) + \ldots$[11]$]/\sqrt{m}$	54	Rodríguez-Gálvez et al. (2021)
$\nabla_w \ell(w, \cdot)$ 1-Lipschitz	$\sum_k \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_{W,S^\star}}[\mathfrak{W}(\mathbb{P}_{S\mid S^\star}, \mathbb{P}_{S\mid W_k,S^\star}) + \ldots]/\sqrt{m}$	54	New
$\vert\ell\vert \leq 1$	$2e^{1/e}\sqrt{p}(I^{(p)}(W;S\mid S^\star)+1)^{1/p}/\sqrt{m(p-1)}$	55	New
$\Vert\nabla_w \ell\Vert \leq 1$	$2e^{1/e}\sqrt{p}\sum_k \varepsilon_{k-1}(I^{(p)}(W_k;S\mid S^\star)+1)^{1/p}/\sqrt{m(p-1)}$	55	New
$\vert\ell\vert \leq 1$	$2\mathbb{E}_{\mathbb{P}_{W,S^\star}}[\chi^2(\mathbb{P}_{S\mid W,S^\star}\Vert\mathbb{P}_{S\mid S^\star})^{1/2} + \ldots]/\sqrt{m}$	56	New
$\Vert\nabla_w \ell\Vert \leq 1$	$2\sum_k \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_{W,S^\star}}[\chi^2(\mathbb{P}_{S\mid W,S^\star}\Vert\mathbb{P}_{S\mid S^\star})^{1/2} + \ldots]/\sqrt{m}$	56	New
$\vert\ell\vert \leq 1$	$2\sum_i \sqrt{2I(W;X_i\mid X_i^\star)}/m$	61	Haghifam et al. (2020)
$\Vert\nabla_w \ell\Vert \leq 1$	$2\sum_i \sum_k \varepsilon_{k-1}\sqrt{2I(W_k;X_i\mid X_i^\star)}/m$	61	New
$\ell(w, \cdot)$ 1-Lipschitz	$\sum_i \mathbb{E}_{\mathbb{P}_{W,X_i^\star}}[\mathfrak{W}(\mathbb{P}_{X_i\mid X_i^\star}, \mathbb{P}_{X_i\mid W,X_i^\star}) + \ldots]/m$	62	Rodríguez-Gálvez et al. (2021)
$\nabla_w \ell(w, \cdot)$ 1-Lipschitz	$\sum_i \sum_k \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_{W,X_i^\star}}[\mathfrak{W}(\mathbb{P}_{X_i\mid X_i^\star}, \mathbb{P}_{X_i\mid W_k,X_i^\star}) + \ldots]/m$	62	New
$\vert\ell\vert \leq 1$	$2\sum_i (I^{(p)}(W;X_i\mid X_i^\star)+1)^{1/p}/m$	63	New
$\Vert\nabla_w \ell\Vert \leq 1$	$2\sum_i \sum_k \varepsilon_{k-1}(I^{(p)}(W_k;X_i\mid X_i^\star)+1)^{1/p}/m$	63	New
$\vert\ell\vert \leq 1$	$\sum_i \mathbb{E}_{\mathbb{P}_{W,X_i^\star}}[\chi^2(\mathbb{P}_{X_i\mid W,X_i^\star}\Vert\mathbb{P}_{X_i\mid X_i^\star})^{1/2} + \ldots]/m$	64	New
$\Vert\nabla_w \ell\Vert \leq 1$	$\sum_i \sum_k \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_{W,X_i^\star}}[\chi^2(\mathbb{P}_{X_i\mid W_k,X_i^\star}\Vert\mathbb{P}_{X_i\mid X_i^\star})^{1/2} + \ldots]/m$	64	New
$\vert\ell\vert \leq 1$	$2\sum_i \mathbb{E}_{\mathbb{P}_{W,X_i^\star}}[\mathrm{TV}(\mathbb{P}_{X_i\mid X_i^\star}, \mathbb{P}_{X_i\mid W,X_i^\star}) + \ldots]/m$	65	Rodríguez-Gálvez et al. (2021)
$\Vert\nabla_w \ell\Vert \leq 1$	$2\sum_i \sum_k \varepsilon_{k-1}\mathbb{E}_{\mathbb{P}_{W,X_i^\star}}[\mathrm{TV}(\mathbb{P}_{X_i\mid X_i^\star}, \mathbb{P}_{X_i\mid W_k,X_i^\star}) + \ldots]/m$	65	New
$\vert\ell\vert \leq 1$	$2\sum_i \sqrt{2\mathrm{L}(W;X_i\mid X_i^\star)}/m$	66	New
$\Vert\nabla_w \ell\Vert \leq 1$	$2\sum_i \sum_k \varepsilon_{k-1}\sqrt{2\mathrm{L}(W_k;X_i\mid X_i^\star)}/m$	66	New

[11] Here and in the following, “$\ldots$” should be read as: “Take the same expression on the left and replace $\mathbb{P}_{S|W,S^\star}$ with $\mathbb{P}_{\bar{S}|W,S^\star}$ (or $\mathbb{P}_{X_i|W,X_i^\star}$ with $\mathbb{P}_{\bar{X}_i|W,X_i^\star}$).”.