Proceedings of Machine Learning Research vol 98:1–33, 2019 30th International Conference on Algorithmic Learning Theory

A Tight Excess Risk Bound via a Unified PAC-Bayesian–Rademacher–Shtarkov–MDL Complexity

Peter D. Grünwald PDG@CWI.NL
CWI and Leiden University, The Netherlands

Nishant A. Mehta NMEHTA@UVIC.CA
University of Victoria, Canada

Editors: Aurélien Garivier and Satyen Kale

Abstract

We present a novel notion of complexity that interpolates between and generalizes some classic complexity notions in learning theory: for empirical risk minimization (ERM) with arbitrary bounded loss, it is upper bounded in terms of data-independent Rademacher complexity; for generalized Bayesian estimators, it is upper bounded by the data-dependent information (KL) complexity. For ERM, the new complexity reduces to normalized maximum likelihood complexity, i.e., a minimax log-loss individual sequence regret. Our first main result bounds excess risk in terms of the new complexity. Our second main result links the new complexity to $L_2(P)$ entropy via Rademacher complexity, generalizing earlier results of Opper, Haussler, Lugosi, and Cesa-Bianchi who covered the log-loss case with L_∞ entropy. Together, these results recover optimal bounds for VC-type and large (polynomial entropy) classes, replacing local Rademacher complexities by a simpler analysis which almost completely separates the two aspects that determine the achievable rates: 'easiness' (Bernstein) conditions and model complexity.

Keywords: NML, MDL, Rademacher complexity, PAC-Bayes, minimax excess risk

1. Introduction

We simultaneously address three questions of learning theory: *(A)* We precisely relate Rademacher complexities for arbitrary bounded losses and the minimax cumulative log-loss regret, also known as the *Shtarkov integral* and *normalized maximum likelihood (NML) complexity*. *(B)* We bound this minimax regret in terms of L_2 entropy; past results were based on L_∞ entropy. *(C)* We introduce a new type of complexity that enables a unification of data-dependent PAC-Bayesian and empirical-process-type excess risk bounds into a single bound that often is minimax optimal.

These results are part of the tree of bounds in Figure 1. The $\leftarrow$ arrow stands for 'bounded in terms of'; the precise bounds are given in the respective results in the paper. Most formulas are also given in the glossary on page 18. Red arrows indicate new results. We start with a sample space $\mathcal{Z}$, a family of predictors $\mathcal{F}$ for an arbitrary loss function $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$, and data $Z^n = Z_1, Z_2, \ldots, Z_n$ i.i.d. $\sim P$. Examples of ℓ include log-loss, squared loss, and $0/1$-loss. In its simplest form, the novel complexity $\mathrm{COMP}_\eta(\mathcal{F})$ depends on η, $\mathcal{F}$, and (suppressed in the notation) P, n, and ℓ; the parameter η may for now be thought of as the learning rate used by an exponentially weighted aggregation forecaster. By ③ = ⑤, Section 2.1, $\mathrm{COMP}_\eta(\mathcal{F})$ is equal to the *minimax cumulative individual sequence regret* for sequential prediction with log-loss relative to a family $\mathcal{Q}_{\mathcal{F},\eta}$ of probability measures defined in terms of $\mathcal{F}$ and η (and, suppressed in notation, P and ℓ). This minimax cumulative regret is also known as

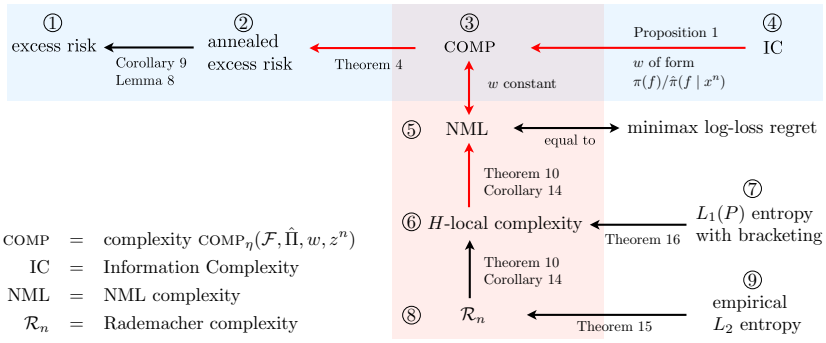

Figure 1: The tree of bounds we provide; red arrows indicate new results.

the log-*Shtarkov integral* or *Normalized Maximum Likelihood (NML) complexity* and has been much studied in the minimum description length (MDL) literature (Rissanen, 1996; Grünwald, 2007). In the sequel, when η is clear from the context we simply use the notations $\mathrm{COMP}(\mathcal{F})$ and $\mathcal{Q}_{\mathcal{F}}$.

Problem A: NML and Rademacher Theorem 10 and Corollary 14 (⑤ ← ⑥ ← ⑧) establish a precise and tight link between NML and Rademacher complexity via a new complexity we introduce and dub *H-local complexity*. Both Rademacher and NML complexities are used as penalties in model selection (although with very different motivations), but while their close conceptual similarity has been noted by several authors (e.g. Grünwald (2007); Zhu et al. (2009); Roos (2016)), so far any formal link has been lacking. The proof of Theorem 10 relies on a novel use of Talagrand's inequality, outlined in Section 4.1 (see Lemma 12). As we now explain, the result also allows us to prove new, concrete regret bounds for log-loss.

Problem B: Bounding NML complexity with L_2 entropies If $\mathcal{F}$ is a class of polynomial empirical L_2 entropy, the Rademacher complexity can be bounded (Theorem 15, ⑧ ← ⑨, (Koltchinskii, 2011)) in terms of empirical L_2 entropy; if $\mathcal{F}$ admits polynomial $L_1(P)$ entropy with bracketing, then H-local complexity is bounded (Theorem 16, ⑥ ← ⑦, (Massart and Nédélec, 2006)) in terms of $L_1(P)$ entropy with bracketing. In conjunction with our new result Theorem 10, ⑤ ← ⑥, this becomes of significant interest for log-loss individual sequence prediction, as we now explain. For any class $\mathcal{F}$ of static experts (probability distributions) for log-loss prediction, we can arrange things such that $\mathcal{Q}_{\mathcal{F},1} = \mathcal{F}$. Theorem 10 then implies a bound on the minimax regret of $\mathcal{F}$ in terms of $L_1(P)$ entropy and empirical L_2 entropy, where P can be any member of the class $\mathcal{Q}_{\mathcal{F}}$, significantly improving previous bounds on minimax log-loss regret that have the same functional form as ours but which relied on L_∞ entropy instead (Opper and Haussler, 1999; Cesa-Bianchi and Lugosi, 2001). These bounds, and also more recent bounds on minimax log-loss regret by Rakhlin and Sridharan (2015), become void whenever the L_∞ entropy is unbounded, whereas our bounds are still meaningful.[1] In Section 4.3, we present one such concrete example where $\mathcal{F}$ is the class of monotone densities and the loss is bounded. In Appendix A, we compare our results more closely to the three aforementioned papers.

Towards Problem C To further explain Figure 1, consider an estimator $\hat{\Pi}$ which on each sample $Z^n = Z_1, \ldots, Z_n$ outputs a distribution $\hat{\Pi} \mid Z^n$ on $\mathcal{F}$; deterministic estimators $\hat{f}$ such as empirical

1. To be clear, we *do not* handle the case of unbounded losses. The L_∞ entropy of a loss-composed class can be unbounded even when the loss is bounded.

risk minimization (ERM) are represented via $\delta_{\hat{f}}$, the Dirac measure on $\hat{f}$. In its general form, our novel complexity $\text{COMP}_\eta(\mathcal{F}, \hat{\Pi}, w, Z^n)$ depends not just on $\mathcal{F}$ but also on the choice of estimator $\hat{\Pi}$,[2] the data $Z^n \in \mathcal{Z}^n$ itself, and a *luckiness function* $w : \mathcal{Z}^n \times \mathcal{F} \to \mathbb{R}_0^+$. The function w generalizes both the idea of an estimation penalty and the 'prior' in PAC-Bayesian bounds, and it can be chosen freely; different choices lead to different complexities and excess risk bounds. If w is taken constant, the complexity becomes data-independent and can be written as $\text{COMP}_\eta(\mathcal{F}, \hat{\Pi})$; the special case $\text{COMP}_\eta(\mathcal{F})$ considered before is the supremum of $\text{COMP}_\eta(\mathcal{F}, \hat{\Pi})$ over all $\mathcal{F}$-valued estimators.

We now turn to the first line in Figure 1. ② ← ③, Theorem 4, bounds the *annealed excess risk* of any estimator $\hat{\Pi}$ in terms of its empirical risk on the training data Z^n plus the complexity in its general form $\text{COMP}_\eta(\mathcal{F}, \hat{\Pi}, w, Z^n)$, for any w chosen a priori. Annealed excess risk is a proxy of actual excess risk, the expected loss difference between predicting with $\hat{\Pi} \mid Z^n$ and predicting with the actual risk minimizer f^* in $\mathcal{F}$. The annealed version is defined via an η-annealed expectation of the form $-\eta^{-1} \log \mathsf{E}[e^{-\eta U}]$ for a random variable U. The bound ① ← ② (Corollary 9) bounds the actual excess risk in terms of the annealed excess risk, so that we get a true excess risk bound for $\hat{\Pi}$. In particular, for w of the form $\pi(f)/\hat{\pi}(f \mid z^n)$, where π is the density of a 'prior' distribution Π on $\mathcal{F}$, the complexity becomes, by ③ ← ④ (Proposition 1) (strictly) upper bounded by the *information complexity* of Zhang (2006a,b), involving a Kullback-Leibler (KL) divergence term $\text{KL}(\hat{\Pi}|Z^n \,\|\, \Pi)$. Information complexity generalizes earlier complexities and associated bounds from information theory such as (extended) *stochastic complexity* (Rissanen, 1989; Yamanishi, 1998), *resolvability* (Barron and Cover, 1991), and PAC-Bayesian excess risk bounds (Audibert, 2004; Catoni, 2007). Together, ① ← ② ← ③ ← ④ recover and strengthen Zhang's bounds.

Problem C: Unifying data-dependent and empirical process-type risk bounds As lamented by Audibert (2004, 2009), despite their considerable appeal, standard PAC-Bayes/KL excess risk bounds do not yield the right rates for large classes, i.e., with polynomial $L_2(P)$ entropy. On the other hand, standard Rademacher complexity generalization and excess risk bound analyses do not easily extend to penalized estimators or generalized Bayesian estimators based on updating a prior distribution; also, handling log-loss appears difficult with Rademacher complexity. Yet ① ← ② ← ③ shows that there exists a single bound capturing all these applications: by varying the function w one can get both (a strict strengthening of) the KL bounds and a Rademacher complexity-type excess risk bound. Thus, the chain of bounds ① ← … ← ⑦/⑨ recovers rates for ERM that either are minimax optimal (for classification) or the best known rates for ERM (for other losses) for VC-type and polynomial entropy classes; the rates depend in the right way on the 'easiness' of the problem as modulated by Tsybakov's (2004) margin condition and Bernstein conditions (Bartlett et al., 2005).

Summary and contents *Technically*, our most important results are Theorem 10 and Corollary 14, leading to nontrivial bounds for minimax log-loss regret in situations in which, to the best of our knowledge, it previously was unknown how to obtain such bounds. *Conceptually* (in the sense of 'giving insight'), the most important result is Theorem 4, which gives a sharp (in a sense we will explain) bound on the annealed risk in terms of the new complexity, and allows us to relate cumulative regret (an individual-sequence notion) to excess risk (a probabilistic notation). Together, Theorems 10 and 4 lead to the full chain of implications which unifies the PAC-Bayesian bounds — which are suitable for log-loss, can incorporate priors, but are suboptimal for large classes — with

2. A complexity being estimator-dependent has precedent from the notion of information complexity (Zhang, 2006b).

the Rademacher style bounds — which do not easily incorporate priors or log-loss but are minimax optimal for large classes.

In Section 2.1, we introduce the simple data-independent version of our complexity, $\text{COMP}(\mathcal{F}, \hat{f})$, which is really the NML complexity. In Section 2.2 we extend our notion of complexity to the generalized data-dependent form $\text{COMP}(\mathcal{F}, \hat{\Pi}, w, z^n)$. Section 3 contains our main conceptual result, Theorem 4. In Section 4, we derive our main technical result, Theorem 10 and its Corollary 14, a bound on $\text{COMP}(\mathcal{F}, \hat{f})$ in terms of Rademacher complexity; we also present a concrete application of this result, Theorem 20, which provides the best known rates for ERM under Bernstein conditions for bounded loss functions in a number of situations.

2. The Novel Complexity Notion

Preliminaries In the statistical learning problem (Vapnik, 1998), a labeled sample $Z^n = Z_1, \ldots, Z_n$ is drawn independently from probability distribution P over $\mathcal{Z} = (\mathcal{X} \times \mathcal{Y})$, where $Z_j = (X_j, Y_j)$ for $j \in [n]$. We are given a model $\mathcal{F}$ and a loss function $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$, with the loss of predictor f on z denoted as $\ell_f(z)$. Loss functions such as 0-1 loss and log-loss (for joint densities on $z = (x, y)$) can be expressed this way: for 0-1 loss $\mathcal{F}$ consists of functions $f : \mathcal{X} \to \mathcal{Y}$ with $\ell_f(x, y) = |y - f(x)|$, and for log-loss $\mathcal{F}$ is a set of probability densities on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ relative to some underlying measure ν and $\ell_f(x, y) = -\log f(x, y)$, with $\log$ the natural logarithm. An *estimator* or *learner* $\hat{\Pi}$ maps from $\mathcal{Z}^n$ to distributions over $\mathcal{F}$. We write $\hat{\Pi} \mid z^n$ to denote the distribution chosen for data z^n. When $\hat{\Pi}$ is supported entirely on a single function $\hat{f} \in \mathcal{F}$, we write the estimator as $\hat{f}$ and the f chosen for given data z^n as $\hat{f}_{|z^n}$. An example of such a *deterministic estimator* is ERM. An example of a *randomized estimator* is $\hat{\Pi} \mid z^n$, the generalized η-Bayesian posterior (Zhang, 2006b). We use the term estimator for both deterministic and randomized $\hat{\Pi}$.

We aim to learn distributions $\hat{\Pi}$ that obtain low expected *risk* $\mathsf{E}_{f \sim \hat{\Pi}}[\mathsf{E}_{Z \sim P}[\ell_f(Z)]]$, where the risk of a predictor f is $\mathsf{E}_{Z \sim P}[\ell_f(Z)]$. The quality of $\hat{\Pi}$ on data z^n is naturally measured via the *excess risk* $\mathsf{E}_{f \sim \hat{\Pi}|z^n}[\mathsf{E}_{Z \sim P}[\ell_f(Z) - \ell_{f^*}(Z)]]$, where f^* is a minimizer of the risk over $\mathcal{F}$; like many other authors (e.g. Bartlett et al. (2005)) we assume that f^* exists. We use the notation $R_f(z) = \ell_f(z) - \ell_{f^*}(z)$, extended to samples $z^n = (z_1, \ldots, z_n) \in \mathcal{Z}^n$ as $R_f(z^n) = \sum_{i=1}^n \left(\ell_f(z_i) - \ell_{f^*}(z_i)\right)$.

2.1. The Novel Complexity Measure, Simple Case

To prepare for the definition of our complexity measure COMP, we first need to associate each $f \in \mathcal{F}$ with a probability distribution Q_f. We may assume without loss of generality that the underlying distribution P on Z has a density p with respect to some base measure ν. Now for each $f \in \mathcal{F}$, we define Q_f to be the distribution over $\mathcal{Z}$ with density (with respect to ν)

$$q_f(z) := \frac{p(z) \cdot e^{-\eta R_f(z)}}{\mathsf{E}_{Z \sim P}\left[e^{-\eta R_f(Z)}\right]}. \tag{1}$$

We extend the definition to n outcomes via the product density $q_f(z^n) := \prod_{i=1}^n q_f(z_i)$. In this way the model $\mathcal{F}$ is itself mapped to a set $\mathcal{Q}_\mathcal{F} = \{q_f : f \in \mathcal{F}\}$ of probability densities, the mapping depending on the loss function ℓ of interest, but also (suppressed in notation) on η, f^*, and on the 'true' P; this is an instance of the 'entropification procedure' suggested by Grünwald (1999).

We now define our new complexity measure. For simplicity we first present the data-independent version in the special case of deterministic estimators $\hat{f}$; this case suffices to make the connection to

minimax regret and Rademacher complexities in Section 4. Define the *Shtarkov integral* as

$$\mathrm{S}(\mathcal{F}; \hat{f}) := \mathsf{E}_{Z^n \sim P}\left[\frac{e^{-\eta R_{\hat{f}_{|Z^n}}(Z^n)}}{C(\hat{f}_{|Z^n})}\right] = \int_{\mathcal{Z}^n} q_{\hat{f}_{|z^n}}(z^n) d\nu(z^n) \overset{(\star)}{=} \int_{\mathcal{Z}^n} p_{\hat{f}_{|z^n}}(z^n) d\nu(z^n), \tag{2}$$

where, for any $f \in \mathcal{F}$, $C(f) := \mathsf{E}_{Z^n \sim P}[e^{-\eta R_f(Z^n)}]$ is the normalization constant. When $\mathrm{S}(\mathcal{F}, \hat{f})$ is finite, as is the case with bounded loss, the corresponding *complexity of model $\mathcal{F}$ equipped with $\hat{f}$* is

$$\textsc{comp}(\mathcal{F}, \hat{f}) := \eta^{-1} \log \mathrm{S}(\mathcal{F}, \hat{f}). \tag{3}$$

COMP, S, q_f, and normalizer C all depend on η, but this is suppressed in notation unless needed for clarity. The $(\star)$ equality in (2) holds in the very special case that the original loss function is log-loss, $\eta = 1$, and $\mathcal{F}$ contains the density p of P ('the model is correct'). In that case $f^* = p$, $C(f) = 1$ for all $f \in \mathcal{F}$, $\mathcal{Q}_{\mathcal{F}}$ is equal to $\mathcal{F}$, and $R_f(z) = -\log f(z) + \log p(z)$; thus, (1) reduces to $q_f(z) = f(z)$, and the $(\star)$ equality follows. We also define the *maximal complexity* $\textsc{comp}(\mathcal{F})$ as

$$\mathrm{S}(\mathcal{F}) := \int_{\mathcal{Z}^n} \sup_{f \in \mathcal{F}} q_f(z^n) d\nu(z^n) \quad ; \quad \textsc{comp}(\mathcal{F}) := \eta^{-1} \log \mathrm{S}(\mathcal{F}) = \sup_{\hat{f}} \textsc{comp}(\mathcal{F}, \hat{f}); \tag{4}$$

the final equality is trivial from the definition, the sup ranging over all deterministic estimators on $\mathcal{F}$.

Let $\mathcal{K}$ be a finite set and let $\{\mathcal{F}_k : k \in \mathcal{K}\}$ be a partition of $\mathcal{F}$. Then, as shown by (e.g.) Opper and Haussler (1999) (a proof is in Appendix E.1 for convenience), for every deterministic estimator,

$$\textsc{comp}(\mathcal{F}, \hat{f}) \leq \eta^{-1} \log |\mathcal{K}| + \max_{k \in \mathcal{K}} \textsc{comp}(\mathcal{F}_k). \tag{5}$$

Using (5), we can link COMP to Rademacher complexity, as shown in Section 4.1. Below, we first link COMP to log-loss prediction, extend it to encompass data-dependent and PAC-Bayesian complexities, and present our excess risk bound for the general complexities.

Minimax cumulative log-loss interpretation of COMP For any estimator $\hat{f}$, we can define a density r on $\mathcal{Z}^n$ relative to ν by setting $r(z^n) := \frac{q_{\hat{f}}(z^n)}{\mathrm{S}(\mathcal{F}, \hat{f})}$, which evidently integrates to 1 and hence is a probability density (different choices of estimator $\hat{f}$ lead to different r; this is suppressed in the notation). We can use density r to sequentially predict $Z_1, Z_2, \ldots, Z_n$ by predicting Z_i with the corresponding conditional density $r(Z_i \mid Z^{i-1})$. The cumulative log-loss thus obtained is given by

$$\textstyle\sum_{i=1}^n -\log r(Z_i \mid Z^{i-1}) = -\log r(Z^n),$$

Because of the correspondence, via Kraft's inequality, of log-loss prediction and data compression, we can also think of this quantity as a codelength. Similarly, $\min_{f \in \mathcal{F}} -\log q_f(Z^n)$ is the minimum cumulative loss one could have obtained *with hindsight*, i.e., if one had sequentially predicted the Z_i by the q_f that turned out to minimize $-\log q_f$ on Z^n. Assuming this minimum is well-defined, it is achieved by $\hat{f}^{\mathrm{ML}}$, the maximum likelihood estimator relative to $\mathcal{Q}_{\mathcal{F}}$, for which evidently also $\textsc{comp}(\mathcal{F}) = \textsc{comp}(\mathcal{F}, \hat{f}^{\mathrm{ML}})$. Thus, we get that for all $z^n \in \mathcal{Z}^n$,

$$\begin{aligned} \eta \cdot \textsc{comp}(\mathcal{F}, \hat{f}) = \log \mathrm{S}(\mathcal{F}, \hat{f}) &= -\log r(z^n) - \left(-\log q_{\hat{f}}(z^n)\right) \\ &\overset{\text{if } \hat{f} = \hat{f}^{\mathrm{ML}}}{=} \eta \cdot \textsc{comp}(\mathcal{F}) = -\log r(z^n) - \min_{f \in \mathcal{F}} \left(-\log q_f(z^n)\right), \end{aligned} \tag{6}$$

the first equation holding for general $\hat{f}$ and the second for $\hat{f}^{\mathrm{ML}}$. The final expression is just the (cumulative log-loss) *regret* of r on data z^n, which, by (6), is constant on z^n. As first noted by Shtarkov (1987), this implies that (6) is also the *minimax individual sequence regret* relative to the model $\mathcal{Q}_{\mathcal{F}}$ when sequentially predicting outcomes $Z_1, \ldots, Z_n$ with the log-loss; the corresponding optimal sequential prediction strategy r is usually called the normalized maximum likelihood (NML) or Shtarkov density; see (Rissanen, 1996; Grünwald, 2007) for details.

Allowing data-dependency We now generalize the complexity definition above to arbitrary deterministic $\hat{f}$ so that it becomes data-dependent. The central concept we need is that of a *luckiness function* $w : \mathcal{Z}^n \to \mathbb{R}_0^+$; every combination of estimator and luckiness function will, up to scaling, define a unique version of complexity; and every such complexity induces a different data-dependent bound on excess risk. We call w a 'luckiness function' since it will improve our excess risk bounds if we are 'lucky' in the sense that P is such that $w(Z^n)$ will be large with high probability.

The *generalized Shtarkov integral* (Grünwald, 2007) for estimator $\hat{f}$ relative to luckiness function w is defined as

$$\mathrm{S}(\mathcal{F}, \hat{f}, w) := \mathsf{E}_{Z^n \sim P}\left[\frac{e^{-\eta R_{\hat{f}_{|Z^n}}(Z^n)}}{C(\hat{f}_{|Z^n})} \cdot w(Z^n)\right] = \int_{\mathcal{Z}^n} q_{\hat{f}_{|z^n}}(z^n) w(z^n) d\nu(z^n), \tag{7}$$

and, whenever $\mathrm{S}(\mathcal{F}, \hat{f}, w) < \infty$, we define the corresponding data-dependent complexity as

$$\textsc{comp}(\mathcal{F}, \hat{f}, w, z^n) := \tfrac{1}{\eta}\left(-\log w(z^n) + \log \mathrm{S}(\mathcal{F}, \hat{f}, w)\right). \tag{8}$$

Both expressions reduce to (2) and (3) if we take w constant over $\mathcal{Z}^n$. The cumulative log-loss interpretation of COMP that held for constant w can be extended to nonuniform w (for an explanation and an example of a nonuniform w, see Appendix C.1).

2.2. The Novel Complexity Measure, General Case

Here we further generalize the complexity definition so that it can output distributions $\hat{\Pi} \mid Z^n$ on $\mathcal{F}$. For this we need to extend the domain of the luckiness function to encompass $\mathcal{F}$, i.e., we now take arbitrary functions of the form $w : \mathcal{Z}^n \times \mathcal{F} \to \mathbb{R}_0^+$.

The generalized Shtarkov integral for estimator $\hat{\Pi}$ relative to luckiness function w is defined as

$$\mathrm{S}(\mathcal{F}, \hat{\Pi}, w) := \mathsf{E}_{Z^n \sim P}\left[\exp\Big(-\mathsf{E}_{\underline{f} \sim \hat{\Pi} | Z^n}\left[\eta R_{\underline{f}}(Z^n) + \log C(\underline{f}) - \log w(Z^n, \underline{f})\right]\Big)\right], \tag{9}$$

and the generalized (data-dependent) model complexity corresponding to (9) is now defined as

$$\textsc{comp}(\mathcal{F}, \hat{\Pi}, w, z^n) := \tfrac{1}{\eta} \cdot \Big(\mathsf{E}_{\underline{f} \sim \hat{\Pi} | z^n}\left[-\log w(z^n, \underline{f})\right] + \log \mathrm{S}(\mathcal{F}, \hat{\Pi}, w)\Big). \tag{10}$$

(9) and (10) are readily seen to generalize (7) and (8) respectively: if, for a given deterministic estimator $\hat{f}$, we take $\hat{\Pi}(\cdot \mid Z^n)$ to be $\delta_{\hat{f}}$ (the Dirac measure on $\hat{f}_{|Z^n}$) and we take a function $w(z^n, f) \equiv w(z^n)$ that does not depend on f, then the expressions above simplify trivially to (7) and (8) respectively; thus $\textsc{comp}(\mathcal{F}, \delta_{\hat{f}}, w, z^n) = \textsc{comp}(\mathcal{F}, \hat{f}, w, z^n)$. Finally, we define

$$\textsc{comp}^{\mathrm{FULL}}(\mathcal{F}, \hat{\Pi}, w, z^n) := \textsc{comp}(\mathcal{F}, \hat{\Pi}, w, z^n) + \mathsf{E}_{\underline{f} \sim \hat{\Pi} | z^n}\left[R_{\underline{f}}(z^n)\right], \tag{11}$$

the sum of the complexity and, for fixed data, the expected excess loss a random draw from $\hat{\Pi}$ achieves on that data. Admittedly, in this fully general form the novel complexity is no longer an easily interpretable quantity; instead, its power derives from the fact that in one special case it simplifies to the readily-interpretable version for deterministic estimators (8) and hence can be related to Rademacher complexity, while in another special case it specializes to the readily-interpretable information complexity. Thus, it allows us to unify two notions of complexity traditionally viewed as distinct, and associated with excess risk bounds based on different proof techniques, to a single notion that allows us to prove excess risk bounds using a single theorem/proof technique (Theorem 4). We already showed how (11) relates to the simpler form (8); we now explain the connection to information complexity.

COMP generalizes information complexities To explain how PAC-Bayesian type complexities arise as a special case of COMP, we consider luckiness measures w that are defined in terms of probability distributions Π on $\mathcal{F}$ that do not depend on the data; we call these 'priors'. For notational convenience it is useful to assume (without loss of generality) that Π has a density π relative to some underlying measure ρ on $\mathcal{F}$ and that, for all $z^n \in \mathcal{Z}^n$, $\hat{\Pi} \mid z^n$ also has a density $\hat{\pi} \mid z^n$ relative to ρ.

Proposition 1 *Consider arbitrary Π and $\hat{\Pi}$ as above with densities π and $\hat{\pi} \mid z^n$ relative to some ρ. Set $w(z^n, f) := \pi(f)/\hat{\pi}(f \mid z^n)$. Then we have $S(\mathcal{F}, \hat{\Pi}, w) \leq 1$. Consequently,*

$$\text{COMP}^{\text{FULL}}(\mathcal{F}, \hat{\Pi}, w, z^n) \leq \mathsf{E}_{\underline{f} \sim \hat{\Pi} | z^n}\left[R_{\underline{f}}(z^n)\right] + \eta^{-1} \cdot \text{KL}(\ (\hat{\Pi} \mid z^n) \,\|\, \Pi\ ), \tag{12}$$

where $\text{KL}(\ (\hat{\Pi} \mid z^n) \,\|\, \Pi\ ) = \mathsf{E}_{\underline{f} \sim \hat{\Pi} | z^n}\left[\log\left(\hat{\pi}(f \mid z^n)/\pi(f)\right)\right]$ *is KL-divergence.*

Thus, $\text{COMP}^{\text{FULL}}$ is upper bounded by *information complexity* defined relative to prior Π (Zhang, 2006a,b), which is just the RHS of (12) divided by n. The notion of information complexity is also used to bound excess risk in the PAC-Bayesian approach of Catoni (2007) and Audibert (2004). As noted by Zhang (2006b), the right-hand side of (12) is minimized if $\hat{\Pi}$ is taken to be a Gibbs estimator, and in that case it evaluates to the *extended stochastic complexity* (Yamanishi, 1998) $-\eta^{-1} \log \mathsf{E}_{\underline{f} \sim \Pi}[\exp(-\eta R_{\underline{f}}(z^n))]$, which for $\eta = 1$ and ℓ the log-loss, coincides with the standard log Bayesian marginal likelihood (see Grünwald and Mehta (2016, Proposition 6) for details). The cumulative log-loss interpretation of COMP extends to this case as well (see Appendix C.2).

3. First Main Result: Bounding Excess Risk in terms of New Complexity

From now on we restrict to the bounded loss setting. Specifically, we assume that

$$\sup_{f,g \in \mathcal{F}} \operatorname{ess\,sup} |\ell_f(Z) - \ell_g(Z)| \leq \frac{1}{2}, \tag{A1}$$

as this always can be accomplished by an appropriate scaling of a bounded loss function. Before presenting our first main result, we introduce a variant of an ordinary expectation as well as some notation. For $\eta > 0$ and general random variables U, we define the *annealed expectation* (see Grünwald and Mehta (2016) for the origin of this terminology) as $\mathsf{E}^{\text{ANN},\eta}[U] = -\frac{1}{\eta} \log \mathsf{E}\left[e^{-\eta U}\right]$.

Below we will first bound the annealed excess risk rather than the standard excess risk and then continue to bound the latter in terms of the former. Our first main result below may be expressed succinctly via the notion of *exponential stochastic inequality* (ESI).

Definition 2 (ESI) *Let $\eta > 0$ and let U, U' be random variables with joint distribution P. We define*

$$U \trianglelefteq_\eta U' \quad \Leftrightarrow \quad \mathsf{E}_{U,U' \sim P}\left[e^{\eta(U-U')}\right] \leq 1, \tag{13}$$

and we write $U \trianglelefteq^*_\eta U'$ iff the right-hand side of (13) holds with equality.

Clearly $U \trianglelefteq^*_\eta U' \Rightarrow U \trianglelefteq_\eta U'$. An ESI captures both high probability and in-expectation results:

Proposition 3 (ESI Implications) *For all $\eta > 0$, if $U \trianglelefteq_\eta U'$ then, (i), $\mathsf{E}[U] \leq \mathsf{E}[U']$; and, (ii), for all $K > 0$, with P-probability at least $1 - e^{-K}$, $U \leq U' + K/\eta$.*

We now present our first main result, a new bound that interpolates between Zhang's bound (see below) and standard empirical process theory bounds for handling large classes and that is sharp in the sense that it really is an equality of exponential moments.

Theorem 4 *For every randomized estimator $\hat{\Pi}$ and every luckiness function $w : \mathcal{Z}^N \times \mathcal{F} \to \mathbb{R}_0^+$,*

$$\mathsf{E}_{\underline{f} \sim \hat{\Pi}_{|Z^n}}\left[\mathsf{E}^{\mathrm{ANN},\eta}_{\bar{Z}\sim P}\left[R_{\underline{f}}(\bar{Z})\right]\right] \trianglelefteq^*_{n\eta} \frac{1}{n} \cdot \mathrm{COMP}^{\mathrm{FULL}}(\mathcal{F}, \hat{\Pi}, w, Z^n). \tag{14}$$

The proof is a sequence of straightforward rewritings, where the key observation is that for every $f \in \mathcal{F}$, the annealed risk $\mathsf{E}^{\mathrm{ANN},\eta}_{\bar{Z}\sim P}\left[R_f(\bar{Z})\right]$ is related to the normalization factor appearing in the definition (1) of the probability density q_f and its n-fold product $C(f)$ appearing in (2) via the following equality, as follows immediately from the definitions:

$$\mathsf{E}^{\mathrm{ANN},\eta}_{\bar{Z}\sim P}\left[R_f(\bar{Z})\right] = \tfrac{1}{n} \cdot \tfrac{-\log C(f)}{\eta}. \tag{15}$$

By taking w as in Proposition 1, via (12), this theorem strictly generalizes Theorem 3.1 of Zhang (2006b), the left-hand side of Zhang's inequality being equal to the annealed excess risk and the right-hand side to the information complexity, i.e., the right-hand side of (12). However, by taking different w, we get different bounds which are not covered by Zhang's results and which, as we will see, can be used to recover minimax excess risk bounds for certain large classes of polynomial entropy. We mention in passing that whereas Zhang's bound is purely data-dependent and hence one can select an estimator $\hat{\Pi}$ that optimizes the bound, $\mathrm{COMP}^{\mathrm{FULL}}(\mathcal{F}, \hat{\Pi}, w, Z^n)$ additionally incorporates the distribution dependent quantity $\mathrm{S}(\mathcal{F}, \hat{\Pi}, w)$ and hence cannot be optimized with respect to w.

The above ESI's have annealed expectations on their left-hand sides and thus still fall short of providing excess risk bounds. This gap can be resolved under the *v-central condition*.

Definition 5 *Let $v\colon [0,\infty) \to [0,\infty)$ be a bounded, non-decreasing function. We say that $(P, \ell, \mathcal{F})$ satisfies the v-central condition if, for all $\gamma > 0$, $\mathsf{E}[e^{-v(\gamma)R_f(Z)}] \leq e^{v(\gamma)\cdot\gamma}$.*

In the special case of constant $v \equiv \eta \in (0,\infty)$, we say that the η-central condition holds.

If the loss is η-exp-concave and $\mathcal{F}$ is convex, the η-central condition holds (Van Erven et al., 2015, p. 1798)). For bounded losses the v-central condition is equivalent to the *Bernstein condition*.

Definition 6 *Let $\beta \in [0,1]$. We say that $(P, \ell, \mathcal{F})$ satisfies the β-Bernstein condition if, for a constant $B < \infty$, it holds that $\mathsf{E}\left[R_f(Z)^2\right] \leq B\,\mathsf{E}\left[R_f(Z)\right]^\beta$ for all $f \in \mathcal{F}$.*

We only recall one direction of the equivalence of the v-central condition to the Bernstein condition here; the full equivalence is due to Van Erven et al. (2015, Theorem 5.4).

Lemma 7 *Assume for all $f \in \mathcal{F}$ that $R_f(Z) \in [-1/2, 1/2]$ a.s. If the β-Bernstein condition holds for a $\beta \in [0,1]$ and a constant B, then the v-central condition holds for $v(\gamma) = \min\{\gamma^{1-\beta}/B, 1\}$.*

Note that for such bounded loss functions, the weakest Bernstein condition with $\beta = 0$ holds automatically, as does the v-central condition with $v(\gamma) \propto \gamma$.

The following lemma is a translation of Lemma 2 of Grünwald (2012) which addresses the aforementioned gap between the annealed and actual expectations.

Lemma 8 *Suppose that the v-central condition holds for some v as in Definition 5. If $R_f(Z) \in [-1/2, 1/2]$ a.s., then for all $\gamma > 0$, for all $\eta \leq \frac{v(\gamma)}{2}$, and for a constant $C_\eta := 2 + 2\eta$,*

$$\mathsf{E}_{Z\sim P}\left[R_f(Z)\right] \leq C_\eta \cdot \mathsf{E}^{\mathrm{ANN},\eta}_{Z\sim P}\left[R_f(Z)\right] + (C_\eta - 1)\eta^{-1} v(\gamma) \cdot \gamma,$$

A version of the above result also holds for general bounded losses.

The next two excess risk bounds in terms of COMP are nearly immediate.

Corollary 9 *Take the setup of Lemma 8. For any randomized estimator $\hat{\Pi}$ and luckiness function w,*

$$\mathsf{E}_{\underline{f}\sim\hat{\Pi}_{|Z^n}}\left[\mathsf{E}_{Z\sim P}\left[R_{\underline{f}}(Z)\right]\right] \trianglelefteq_{v(\gamma)\cdot n/6} \ \tfrac{3}{n} \cdot \mathrm{COMP}^{\mathrm{FULL}}_{v(\gamma)/2}(\mathcal{F}, \hat{\Pi}, w, Z^n) + 4\gamma. \tag{16}$$

If $\hat{f}$ is ERM, then
$$\mathsf{E}_{Z\sim P}\left[R_{\hat{f}}(Z)\right] \trianglelefteq_{v(\gamma)\cdot n/6} \ \tfrac{3}{n} \cdot \mathrm{COMP}_{v(\gamma)/2}(\mathcal{F}, \hat{f}) + 4\gamma. \tag{17}$$

To help interpret Corollary 9, we give two special cases of (16). In both cases, we will suppose (as in Proposition 1) that $w(z^n, f) := \pi(f)/\hat{\pi}(f \mid z^n)$, where π is the density of a fixed probability measure on $\mathcal{F}$ independent of the sample, so that COMP is bounded by information complexity. First, if the η-central condition holds, then, setting $\eta' = \eta/2$ and using (12), it further follows that

$$\mathsf{E}_{\underline{f}\sim\hat{\Pi}_{|Z^n}}\left[\mathsf{E}_{Z\sim P}\left[R_{\underline{f}}(Z)\right]\right] \trianglelefteq_{\frac{n\eta}{6}} \ \tfrac{3}{n}\Big(\mathsf{E}_{\underline{f}\sim\hat{\Pi}|Z^n}[R_{\underline{f}}(Z^n)] + \tfrac{2}{\eta} \cdot \mathrm{KL}(\ (\hat{\Pi} \mid Z^n) \,\|\, \Pi\ )\Big).$$

In the second case, we take $\hat{\Pi}$ to be any posterior whose $\hat{\Pi}$-expected empirical risk is at most the empirical risk of f^* (e.g. $\hat{\Pi}$ could be the Dirac measure on ERM), and for simplicity we further assume that a β-Bernstein condition holds for some $B \geq 2$ (if it holds for a smaller B it also holds for $B = 2$). Thus, from the bounded loss assumption the v-central condition holds for $v(\gamma) = \frac{\gamma^{1-\beta}}{B}$ (provided that we only consider $\gamma \leq B^{1/(1-\beta)}$), and tuning γ yields $\gamma = A_1 \cdot n^{-1/(2-\beta)}\,\mathrm{KL}(\ (\hat{\Pi} \mid Z^n) \,\|\, \Pi\ )^{1/(2-\beta)}$ for a constant A_1 depending only on β and B, so that

$$\mathsf{E}_{\underline{f}\sim\hat{\Pi}_{|Z^n}}\left[\mathsf{E}_{Z\sim P}\left[R_{\underline{f}}(Z)\right]\right] \trianglelefteq_{n\cdot a_n} \ A_1 \cdot \left(\tfrac{1}{n} \cdot \mathrm{KL}(\ (\hat{\Pi} \mid Z^n) \,\|\, \Pi\ )\right)^{1/(2-\beta)}, \tag{18}$$

where $a_n = A_2(\mathrm{KL}(\ (\hat{\Pi} \mid Z^n) \,\|\, \Pi\ )/n)^{(1-\beta)/(2-\beta)}$ for a constant A_2 depending only on β and B. Lastly, in both cases when the class is finite and the prior Π is uniform, the KL-divergence term reduces to $\log|\mathcal{F}|$. We thus retrieve the familiar $O(n^{-1/2})$ rate in the worst case ($\beta = 0$, for which the Bernstein condition holds vacuously for bounded losses) and $O(n^{-1})$ in the best case, $\beta = 1$.

4. Bounds on Maximal Complexity COMP($\mathcal{F}$) and Excess Risk Bounds they Imply

We now leverage and extend ideas of Opper and Haussler (1999) and results from empirical process theory to get explicit bounds on COMP for two important types of large classes: classes whose empirical entropy grows polynomially and sets of classifiers with polynomial entropy with bracketing. Along the way, we form vital connections to expected suprema of certain empirical processes, including Rademacher complexity. Finally, we present explicit excess risk bounds which are simple consequences of our bounds on COMP. Analogues of all these results, including excess risk bounds with optimal rates, also hold for classes of VC-type; for space, we leave these results to Appendix D.

Preliminaries To properly capture losses like log-loss and supervised losses like 0-1 loss and squared loss, we introduce two parameterizations of the loss function: the *supervised loss* parameterization $\ell_f(z) = \ell(y, f(x))$ for $z = (x, y)$; and the *direct* parameterization $\ell_f(z) = f(z)$. We use the direct parameterization for density estimation with log-loss, where we define $f(z) = \ell_f(z) = -\log p_f(z)$ (we return to conditional density estimation with p_f of the form $p_f(y|x)$ in Appendix A). Thus, each $f \in \mathcal{F}$ has domain $\mathcal{Z}$, and the equivalence $\mathcal{F} = \{\ell_f : f \in \mathcal{F}\}$ holds. For supervised losses, however, each $f \in \mathcal{F}$ has domain $\mathcal{X}$ while each loss-composed function ℓ_f has domain $\mathcal{Z}$.

Unlike previous sections, in this section we require an additional assumption in the case of the supervised loss parameterization: we assume that, for each outcome $(x, y) = z \in \mathcal{Z}$, the loss $\ell_f(z) = \ell(y, f(x))$ is L-Lipschitz in its second argument, i.e. for all $f, g \in \mathcal{F}$,

$$\left|\ell(y, f(x)) - \ell(y, g(x))\right| \leq L \left|f(x) - g(x)\right|. \tag{A2}$$

In the case of classification with 0-1 loss, $\mathcal{F}$ is the set of classifiers taking values in $\{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$, and so (A2) will hold with $L = 1$ (and is in fact an equality). For convenience in the analysis, in the case of the direct parameterization we may always take $L = 1$.

We review some of the standard notions of complexity before presenting our bounds; for more details, see, e.g., Van der Vaart and Wellner (1996) . Let $\mathcal{H}$ be a class of functions mapping from some space $\mathcal{S}$ to $\mathbb{R}$; we typically will take $\mathcal{S}$ equal to either $\mathcal{X}$ or $\mathcal{Z}$. For a pseudonorm $\|\cdot\|$, the *ε-covering number* $\mathcal{N}(\mathcal{H}, \|\cdot\|, \varepsilon)$ is the minimum number of radius-ε balls in the pseudonorm $\|\cdot\|$ whose union contains $\mathcal{H}$. We will work with the $L_2(Q)$ (or $L_1(Q)$) pseudonorms for some probability measure Q. A case that will occur frequently is when $Q = P_n$ is the empirical measure $\frac{1}{n}\sum_{j=1}^n \delta_{S_j}$ based on a sample $S_1, \ldots, S_n$; here, δ_s (for $s \in \mathcal{S}$) is a Dirac measure, and the sample will always be clear from the context.

For two functions $h^{(l)}$ and $h^{(u)}$, the *bracket* $[h^{(l)}, h^{(u)}]$ is the set of all functions f that satisfy $h^{(l)} \leq f \leq h^{(u)}$. An *$\varepsilon$-bracket* (in some pseudonorm $\|\cdot\|$) is a bracket $[h^{(l)}, h^{(u)}]$ satisfying $\|h^{(l)} - h^{(u)}\| \leq \varepsilon$. The *$\varepsilon$-bracketing number* $\mathcal{N}_{[]}(\mathcal{H}, \|\cdot\|, \varepsilon)$ is the minimum number of ε-brackets that cover $\mathcal{H}$; the logarithm of the ε-bracketing number is called the *ε-entropy with bracketing*.

Let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher random variables (distributed uniformly on $\{-1, 1\}$). The *empirical Rademacher complexity* of $\mathcal{H}$ and the *Rademacher complexity* of $\mathcal{H}$ respectively are

$$\mathcal{R}_n(\mathcal{H} \mid S^n) := \mathsf{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\sup_{h\in\mathcal{H}}\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n \epsilon_i h(S_i)\right|\right]; \qquad \mathcal{R}_n(\mathcal{H}) := \mathsf{E}\left[\sup_{h\in\mathcal{H}}\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n \epsilon_i h(S_i)\right|\right],$$

where the first expectation is conditional on $S^n = (S_1, \ldots, S_n)$.

4.1. H-local Complexity and Rademacher Complexity Bounds on the NML Complexity

We first show that the simple form of the complexity $\text{COMP}(\mathcal{F}, \hat{f}) \leq \text{COMP}(\mathcal{F})$ can be directly upper bounded in terms of two other complexity notions, the H-local complexity (defined below) and Rademacher complexity, up to a constant depending on $\sup_{f,g\in\mathcal{F}} \|f-g\|_{L_2(P)}$, the $L_2(P)$ diameter of $\mathcal{F}$.

Theorem 10 (Main Technical Result) *Fix $\varepsilon > 0$ and let $\mathcal{F}$ have diameter ε in the $L_2(P)$ pseudo-metric. Define $\sigma := e\,L\,\varepsilon$, fix arbitrary $f_0 \in \mathcal{F}$, and define the loss class $\mathcal{G} := \{\ell_{f_0} - \ell_f : f \in \mathcal{F}\}$.*

$$\textit{Define} \qquad T_n := \sup_{f\in\mathcal{F}} \left\{ \sum_{j=1}^n \left(\ell_{f_0}(Z_j) - \ell_f(Z_j)\right) - \mathsf{E}_{Z^n\sim Q_{f_0}} \left[\sum_{j=1}^n \left(\ell_{f_0}(Z_j) - \ell_f(Z_j)\right)\right] \right\}.$$

$$\textit{Then} \qquad \text{COMP}_\eta(\mathcal{F}) \;\leq\; 3\,\mathsf{E}_{Z^n\sim Q_{f_0}}\left[T_n\right] + n\eta\sigma^2 \;\leq\; 6n\,\mathsf{E}_{Z^n\sim Q_{f_0}}\left[\mathcal{R}_n(\mathcal{G} \mid Z^n)\right] + n\eta\sigma^2. \tag{19}$$

We call the quantity $\mathsf{E}_{Z^n\sim Q_{f_0}}[T_n]$ an *entropified local complexity*, or *H-local complexity* for short. The "local" nomenclature stems from how *(a)* in the empirical process inside the supremum for T_n; the loss is localized around ℓ_{f_0}, and *(b)* we apply this H-local complexity only for subclasses of small diameter. "Entropified" refers to the sample being distributed according to Q_{f_0}, itself defined via entropification. The attentive reader may have noticed that in the above theorem, the expectation in the Rademacher complexity is relative to the distribution Q_{f_0} for arbitrary $f_0 \in \mathcal{F}$, rather than the distribution P generating the data. The appearance of Q_{f_0} appears to dampen the utility of $\mathcal{F}$ having small $L_2(P)$ diameter. This apparent mismatch will be of no concern due to a technical lemma (Lemma 27 in Appendix E.3), which relates the $L_2(Q_{f_0})$ and $L_2(P)$ pseudometrics.

The proof of Theorem 10 is in three steps; the first part of (19) is a consequence of Lemmas 11 and 12 below. Step 1 is a simple generalization of an argument of Opper and Haussler (1999):

Lemma 11 *Take arbitrary $\mathcal{F}$ and fix arbitrary $f_0 \in \mathcal{F}$. Then $\text{COMP}_\eta(\mathcal{F}) \leq \frac{1}{\eta}\log \mathsf{E}_{Z^n\sim Q_{f_0}}\left[e^{\eta T_n}\right]$.*

Step 2 is to bound $\mathsf{E}[e^{\eta T_n}]$. The next lemma (proved via Talagrand's inequality) does this.

Lemma 12 ('Reverse Jensen' for T_n) *Let $\mathcal{F}$, σ, and $\mathcal{G}$ be as in Theorem 10. Then*

$$\mathsf{E}_{Z^n\sim Q_{f_0}}[e^{\eta T_n}] \leq \exp\left(3\eta\,\mathsf{E}_{Z^n\sim Q_{f_0}}\left[T_n\right] + n\eta^2\sigma^2\right). \tag{20}$$

Opper and Haussler (1999) obtained a result similar to (20) but under the considerably stronger assumption that the original class has finite sup-norm entropy and, consequently, that the class $\mathcal{G}_k$ has sup-norm radius at most $O(\varepsilon)$. The first inequality of (19) now follows. Step 3 proves the second inequality via standard results from empirical process theory:

Lemma 13 *Take the setting of Theorem 10. Then $\mathsf{E}_{Z^n\sim Q_{f_0}}\left[T_n\right] \leq 2n\,\mathsf{E}_{Z^n\sim Q_{f_0}}\left[\mathcal{R}_n(\mathcal{G} \mid Z^n)\right]$.*

To use Theorem 10 for $\mathcal{F}$ with large $L_2(P)$ diameter, we first decompose $\text{COMP}_\eta(\mathcal{F})$ in terms of the $L_2(P)$ covering numbers at a small, optimally-tuned resolution ε plus the maximal complexity among all Voronoi cells induced by the cover, as in (5). We then use known bounds on H-local complexity and Rademacher complexity to sharply bound $\text{COMP}(\mathcal{F})$ in terms of covering numbers. To this end, let $\mathcal{F}$ be arbitrary and let $\{f_1, \ldots, f_{N_\varepsilon}\}$ be an $(\varepsilon/2)$-cover for $\mathcal{F}$ in the $L_2(P)$ pseudo-metric, with $N_\varepsilon := \mathcal{N}(\mathcal{F}, L_2(P), \varepsilon/2)$, and let $\mathcal{F}_{\varepsilon,1}, \ldots, \mathcal{F}_{\varepsilon,N_\varepsilon}$ be the corresponding partition of $\mathcal{F}$ into Voronoi cells according to the $L_2(P)$ pseudometric. That is, for $k \in [N_\varepsilon]$, Voronoi cell $\mathcal{F}_{\varepsilon,k}$ is defined (breaking ties arbitrarily) as $\left\{f \in \mathcal{F} : k = \arg\min_{i\in[N_\varepsilon]} \|f - f_i\|_{L_2(P)}\right\}$. Clearly, each cell $\mathcal{F}_{\varepsilon,k}$ has $L_2(P)$ diameter at most ε. For each k, fix an arbitrary $f_k \in \mathcal{F}_{\varepsilon,k}$ and let $T_n^{(k)}$ be defined as T_n above with f_k in the role of f_0. Via (5), the following corollary of Theorem 10 is now immediate.

Corollary 14 *Let $\sigma := e\,L\,\varepsilon$ and, for $k \in [N_\varepsilon]$, define the loss class $\mathcal{G}_k := \{\ell_{f_k} - \ell_f : f \in \mathcal{F}_{\varepsilon,k}\}$.*

Then
$$\textsc{comp}_\eta(\mathcal{F}) \le \eta^{-1}\log N_\varepsilon + \max_{k\in[N_\varepsilon]}\left\{3\,\mathsf{E}_{Z^n\sim Q_{f_k}}\left[T_n^{(k)}\right] + \eta n\sigma^2\right\} \tag{21}$$
$$\le \eta^{-1}\log N_\varepsilon + \max_{k\in[N_\varepsilon]}\left\{6n\,\mathsf{E}_{Z^n\sim Q_{f_k}}\left[\mathcal{R}_n(\mathcal{G}_k \mid Z^n)\right] + \eta n\sigma^2\right\}. \tag{22}$$

4.2. From H-local Complexity and Rademacher Complexity to Excess Risk Bounds

We now show concrete implications of our link between COMP, $\mathsf{E}[T_n]$, and $\mathcal{R}_n$ for classes with polynomial empirical entropy and sets of classifiers of polynomial $L_2(P)$ entropy with bracketing. Let $\mathcal{H}$ be a class of functions over a space $\mathcal{S}$. Class $\mathcal{H}$ is said to have *polynomial empirical entropy* if, for some $A \in (0,\infty), \rho \in (0,1)$, for all $\varepsilon > 0$, the empirical entropy of $\mathcal{H}$ satisfies

$$\sup_{s_1,\ldots,s_n\in\mathcal{S}} \log \mathcal{N}(\mathcal{H}, L_2(P_n), \varepsilon) \le \left(A/\varepsilon\right)^{2\rho}. \tag{23}$$

We say class $\mathcal{H}$ has *polynomial $L_1(P)$ entropy with bracketing* if, for some $A \in (0,\infty), \rho \in (0,1)$, for all $\varepsilon > 0$, the $L_1(P)$ entropy with bracketing of $\mathcal{H}$ satisfies

$$\log \mathcal{N}_{[]}(\mathcal{H}, L_1(P), \varepsilon) \le \left(A^2/\varepsilon\right)^{\rho}. \tag{24}$$

To obtain explicit bounds from Corollary 14, we require suitable upper bounds on either the Rademacher complexity $\mathsf{E}_{Z^n\sim Q_{f_k}}\left[\mathcal{R}_n(\mathcal{G}_k \mid Z^n)\right]$ or directly on the H-local complexity $\mathsf{E}_{Q_{f_k}}\left[T_n^{(k)}\right]$ itself for the above two types of classes. It is simple to obtain such bounds using Dudley's entropy integral, a product of the chaining technique of empirical process theory. However, the trick here is to somehow leverage that $\mathcal{G}_k$ has small $L_2(P)$ diameter. Koltchinskii (2011) (see equation (3.19)) obtained a bound which improves with reductions in the $L_2(P)$ diameter; we restate a simplified version here. In the sequel, $\lesssim$ means inequality up to multiplication by a universal constant.

Theorem 15 *Let $\mathcal{H}$ be a class of functions over $\mathcal{Z}$ with: polynomial empirical entropy as in (23) with exponent ρ; $\sup_{h\in\mathcal{H}} \mathsf{E}_{Z\sim Q}[h(Z)^2] \le \sigma^2$; and $U := \sup_{h\in\mathcal{H}} \|h\|_\infty$. If $Q \in \Delta(\mathcal{Z})$, then*

$$\mathsf{E}_{Z^n\sim Q}\left[\mathcal{R}_n(\mathcal{H} \mid Z^n)\right] \lesssim \max\left\{A^\rho \sigma^{1-\rho} n^{-1/2}, A^{2\rho/(\rho+1)} U^{(1-\rho)/(1+\rho)} n^{-1/(1+\rho)}\right\}. \tag{25}$$

For classes of polynomial entropy with bracketing, we appeal to upper bounds on $\mathsf{E}_{Q_{f_k}}\left[T_n^{(k)}\right]$. If the class $\mathcal{G}_k$ has small $L_1(Q_{f_k})$ diameter and, moreover, if it also has polynomial $L_1(Q_{f_k})$ entropy with bracketing, then Lemma A.4 of Massart and Nédélec (2006) provides precisely such a bound. Below, we present a straightforward consequence thereof.

Theorem 16 *Let $\mathcal{H}$ be a class of functions over $\mathcal{Z}$ with: polynomial entropy with bracketing as in (24) with exponent ρ; $\sup_{h\in\mathcal{H}} \mathsf{E}_{Z\sim Q}[|h(Z)|] \le \sigma^2$; and $\sup_{h\in\mathcal{H}} \|h\|_\infty \le 1$. If $Q \in \Delta(\mathcal{Z})$, then*

$$\mathsf{E}_{Z^n\sim Q}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{j=1}^n h(Z_j) - \mathsf{E}[h(Z)]\right\}\right] \lesssim \max\left\{A^\rho \sigma^{1-\rho} n^{-1/2}, A^{2\rho/(\rho+1)} n^{-1/(1+\rho)}\right\}. \tag{26}$$

The following theorem builds on Corollary 14 and *nearly* follows by plugging in (25) into (22) and tuning ε in terms of n and η (which gives the non-bracketing case of (27)), and plugging in (26) into (21) and then tuning (which gives the bracketing case of (27)). The remaining work is to resolve a minor discrepancy between $L_2(P)$ pseudonorms and $L_2(Q_{f_k})$ pseudonorms (or the L_1 versions thereof). This theorem will allow us to show optimal rates under Bernstein conditions.

Theorem 17 *If $\mathcal{F}$ has polynomial empirical entropy as in (23) or is a set of classifiers of polynomial entropy with bracketing as in (24) with exponent ρ, then, for all $\eta \in (0,1]$,*

$$n^{-1}\mathrm{COMP}_\eta(\mathcal{F}) \lesssim (AL)^{\frac{2\rho}{1+\rho}} \cdot n^{-\frac{1}{1+\rho}} \cdot \eta^{-\frac{1-\rho}{1+\rho}}. \tag{27}$$

We now prepare for our results on the rates of ERM on a class $\mathcal{F}$. In the following corollary, note that in both cases, the occurrence of the Bernstein exponent β (or κ^{-1}) is consistent with its occurrence in the simple finite $\mathcal{F}$ setting of (18).

Corollary 18 *Assume that a β-Bernstein condition holds for $\mathcal{F}$ as in Definition 6 for some β and B, and impose assumption (A1). Define $\kappa := \beta^{-1}$. Let $\hat{f}$ be ERM over $\mathcal{F}$. Then (27) further implies, taking $\gamma = n^{-\frac{\kappa}{2\kappa-1+\rho}}$ and assuming that $n > (2B)^{-\frac{2\kappa-1+\rho}{\kappa-1}}$, that for all $z^n \in \mathcal{Z}^n$,*

$$n^{-1}\mathrm{COMP}_{v(\gamma)}(\mathcal{F}, \hat{f}) + \gamma \lesssim \left((AL)^{\frac{2\rho}{\rho+1}} + 1\right) \cdot B^{\frac{1-\rho}{1+\rho}} n^{-\frac{\kappa}{2\kappa-1+\rho}}. \tag{28}$$

We used the notation $\kappa = \beta^{-1}$ here to make the results more easily comparable to Tsybakov (2004) and Audibert (2004); however, the result still holds for the case $\beta = 0$ if we simply replace κ^{-1} by β in all exponents above; e.g. $\kappa/(2\kappa - 1 + \rho)$ in (28) becomes $1/(2 - \kappa^{-1} + \kappa^{-1}\rho) = 1/(2 - \beta + \beta\rho)$.

4.3. Applications

We now sketch two applications. The premise for the first is that the results of Opper and Haussler (1999) hinge on the log-loss-composed class having finite sup-norm entropy. Yet, our bounds on COMP (and hence the cumulative minimax regret under log-loss) instead only require this class to have finite empirical L_2 entropy. Our first application illustrates how different these metric entropies can be via the example of individual sequence prediction under log-loss with a class of monotone probability densities. The second application shows how our results recover optimal rates under Bernstein conditions for large classes.

Monotone densities and bounded log-loss regret Let $c_{\min}$ and $c_{\max}$ be positive constants satisfying $0 < c_{\min} < 1$ and $2 \leq c_{\max} < \infty$. Consider the class $\mathcal{P}$ of monotone probability densities on $[0,1]$ for which $c_{\min} \leq p(x) \leq c_{\max}$ for all $p \in \mathcal{P}$ and all $x \in [0,1]$; we require the densities to be uniformly lower bounded by a constant due to our bounded loss assumption. This class is well-studied in nonparametric statistics (see e.g. Ghosal et al. (2000); Giné and Nickl (2016)).

As we explain in Appendix B, the loss-composed class $\ell \circ \mathcal{P} := \{-\log p : p \in \mathcal{P}\}$ has polynomial empirical L_2 entropy that grows as $O(\varepsilon^{-1})$. The bound (23) thus holds with $\rho = \frac{1}{2}$. Taking $\eta = 1$, applying Theorem 17, and leveraging the connection between COMP and the minimax individual sequence regret under log-loss (from (6)), we immediately have:

Theorem 19 *For the class $\mathcal{P}$, the minimax individual sequence regret under log-loss is $O(n^{1/3})$.*

Using the fact that the log-loss satisfies the 1-central condition and that (as is easily shown) $\mathrm{COMP}_\eta(\mathcal{F})$ is increasing in η, we can also apply our excess risk bounds with $\eta = 1/2$ and $\hat{f}$ set to ERM (i.e., maximum likelihood) to this problem. In combination with Theorem 19 this yields an excess log-risk (i.e. KL divergence) rate of $O(n^{1/3}/n) = O(n^{-2/3})$, implying a maximum likelihood Hellinger convergence rate of $O(\sqrt{n^{-2/3}}) = O(n^{-1/3})$ which is known to be minimax optimal (Ghosal et al., 2000). On the other hand, we show in Appendix B that the sup-norm entropy of $\mathcal{P}$ cannot be finite for small enough ε; this in fact holds even for a 1-dimensional parametric subclass! Consequently, the results of Opper and Haussler (1999) and Rakhlin and Sridharan (2015) cannot yield non-trivial regret bounds here.

Recovering bounds under Bernstein conditions for large classes We now show how Lemma 8 can recover optimal rates under the Tsybakov margin condition and the best known rates for ERM under Bernstein-type conditions. While other techniques can achieve the same rates for ERM, we feel that our approach embodies a simpler analysis; it also leads to new results for other estimators, as shown in the long version.

Theorem 20 *Assume that the β-Bernstein condition holds for $\mathcal{F}$ as in Corollary 18 and define $\kappa := \beta^{-1}$. Let $\hat{f}$ be ERM over $\mathcal{F}$. Suppose that $\mathcal{F}$ has polynomial empirical entropy as in (23) or is a set of classifiers of polynomial entropy with bracketing as in (24) with exponent ρ. Then there is a C_2 such that for all n large enough so that $n > (2B)^{-\frac{2\kappa-1+\rho}{\kappa-1}}$, we have, with $\psi_2(n) = \frac{1}{6} \cdot n^{\frac{\kappa+\rho}{2\kappa-1+\rho}}$,*

$$\mathsf{E}_{Z\sim P}\left[R_{\hat{f}}(Z)\right] \trianglelefteq_{\psi_2(n)} C_2\left[\left((AL)^{\frac{2\rho}{\rho+1}}+1\right)\cdot B^{\frac{1-\rho}{1+\rho}}\cdot n^{-\frac{\kappa}{2\kappa-1+\rho}}\right], \tag{29}$$

In the long version, we extend this result to general deterministic estimators. Theorem 20 combined with part (ii) of Proposition 3 implies that, with probability at least $1-\delta$, ERM obtains the rate $n^{-\frac{\kappa}{2\kappa-1+\rho}} + n^{-\frac{\kappa+\rho}{2\kappa-1+\rho}} \cdot \log\frac{1}{\delta}$. For sets of classifiers of polynomial entropy with bracketing, the rate $n^{-\frac{\kappa}{(2\kappa-1+\rho)}}$ is known to be optimal, matching results of Tsybakov (2004, Theorem 1), Audibert (2004) (see the discussion after Theorem 3.3), and Koltchinskii (2006, p. 36). Outside of classification, for classes of polynomial empirical entropy the rate we obtain is to our knowledge the best known for ERM. In particular, if the nonparametric class is convex and the loss is exp-concave, then $\kappa = \beta = 1$, and our rates for ERM are minimax optimal (Rakhlin et al., 2017, Theorem 7). Yet, there are cases where $\beta < 1$ in which an aggregation scheme can obtain a rate as if $\beta = 1$; one such example is in the case of squared loss with a non-convex class (Rakhlin et al., 2017; Liang et al., 2015).

Additional insights and results Theorem 20 offers a distribution-dependent bound that recovers minimax optimal rates for ERM. We can extend the result to obtain the optimal rates in the (easier) case of VC-type classes as well, as we show in Appendix D. In the long version of the paper, we further extend these results to general deterministic estimators. Our bounds are arguably simpler than those based on local Rademacher complexity; further discussion is in Appendix A ("Discussion"). There, we also discuss why our minimax cumulative log-loss regret results do not easily transfer to conditional density estimation, and we indicate how our Corollary 9 leads to a general 1-to-1 correspondence between excess risk bounds, luckiness functions w, and lossless codes for data compression: as just two special cases, the right-hand side of PAC-Bayesian bounds with w set as in Proposition 1 can be interpreted as the log-loss/codelength regret of a Bayesian coding strategy; and the data-independent NML bound on excess risk is obtained with $w = 1$.

Acknowledgments

We would like to thank Dylan Foster for useful discussions.

References

Jean-Yves Audibert. *PAC-Bayesian statistical learning theory*. PhD thesis, Université Paris VI, 2004.

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

Andrew R. Barron and Thomas M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, 1991.

Peter Bartlett, Peter Grünwald, Peter Harremoës, Fares Hedayati, and Wojciech Kotlowski. Horizon-independent optimal prediction with log-loss in exponential families. In *Conference on Learning Theory*, pages 639–661, 2013.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

Olivier Catoni. *PAC-Bayesian Supervised Classification*. Lecture Notes-Monograph Series. IMS, 2007.

Nicolò Cesa-Bianchi and Gábor Lugosi. Worst-case bounds for the logarithmic loss of predictors. *Machine Learning*, 43(3):247–264, 2001.

Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-dimensional Statistical Models*, volume 40. Cambridge University Press, 2016.

Peter Grünwald. The safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer, 2012.

Peter D. Grünwald. Viewing all models as "probabilistic". In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, pages 171–182, 1999.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

Peter D. Grünwald and Nishant A. Mehta. Fast rates with unbounded losses. *arXiv preprint* `1605.00252`, 2016.

David Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

Vladimir Koltchinskii and Dmitry Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, 1999.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset Rademacher complexity. In *Proceedings of The 27th Conference on Learning Theory (COLT 2015)*, pages 1260–1285, 2015.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34 (5):2326–2366, 2006.

Manfred Opper and David Haussler. Worst case prediction over sequences under log loss. In *The Mathematics of Information Coding, Extraction and Distribution*, pages 81–90. Springer, 1999.

Alexander Rakhlin and Karthik Sridharan. Sequential probability assignment with binary alphabets and large classes of experts. *arXiv preprint arXiv:1501.07340*, 2015.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Hackensack, NJ, 1989.

Jorma Rissanen. Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory*, 42(1): 40–47, 1996.

Teemu Roos. Informal remark, 2016. Remarks made during the discussion of J. Shawe-Taylor's invited talk at the WITMSE 2016 conference.

Yu. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, 1987.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

AW van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 1996.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015. Special issue in Memory of Alexey Chervonenkis.

Vladimir N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

Mathukumalli Vidyasagar. *Learning and Generalization with Applications to Neural Networks*. Springer, 2002.

Kenji Yamanishi. A decision-theoretic extension of stochastic complexity and its applications to learning. *Information Theory, IEEE Transactions on*, 44(4):1424–1439, 1998.

Tong Zhang. From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006a.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *Information Theory, IEEE Transactions on*, 52(4):1307–1321, 2006b.

Xiaojin Zhu, Bryan R Gibson, and Timothy T Rogers. Human Rademacher complexity. In *Advances in Neural Information Processing Systems*, pages 2322–2330, 2009.

Glossary

Notation	Description	Page
Losses		
$R_f(z)$	Excess loss, $\ell_f(z) - \ell_{f^*}(z)$	4
$R_f(z^n)$	Cumulative excess loss, $\ell_f(z^n) - \ell_{f^*}(z^n)$	4
$q_f(z)$	Entropified version of f, $\frac{p(z)\cdot e^{-\eta R_f(z)}}{\mathsf{E}_{Z\sim P}\left[e^{-\eta R_f(Z)}\right]}$	4
Notation		
$U \trianglelefteq_\eta U'$	Exponential stochastic inequality, $\mathsf{E}_{U,U'\sim P}\left[e^{\eta(U-U')}\right] \le 1$	7
$\mathsf{E}^{\mathrm{ANN},\eta}[U]$	Annealed expectation, $-\frac{1}{\eta}\log \mathsf{E}\left[e^{-\eta U}\right]$	7
Complexities		
$C(f)$	Normalization constant for Shtarkov integral, $\mathsf{E}_{Z^n\sim P}\left[e^{-\eta R_f(Z^n)}\right]$	5
$\mathrm{S}(\mathcal{F};\hat{f})$	Shtarkov integral (deterministic version), $\mathsf{E}_{Z^n\sim P}\left[\frac{e^{-\eta R_{\hat{f}_{\mid Z^n}}(Z^n)}}{C(\hat{f}_{\mid Z^n})}\right]$	5
$\mathrm{S}(\mathcal{F})$	Maximal Shtarkov integral (deterministic version), $\int_{\mathcal{Z}^n}\sup_{f\in\mathcal{F}} q_f(z^n)d\nu(z^n)$	5
$\mathrm{S}(\mathcal{F};\hat{f},w)$	Generalized Shtarkov integral (deterministic version) $\mathsf{E}_{Z^n\sim P}\left[\frac{e^{-\eta R_{\hat{f}_{\mid Z^n}}(Z^n)}}{C(\hat{f}_{\mid Z^n})}\cdot w(Z^n)\right] = \int_{\mathcal{Z}^n} q_{\hat{f}_{\mid z^n}}(z^n)w(z^n)d\nu(z^n)$	6
$\mathrm{S}(\mathcal{F},\hat{\Pi},w)$	Generalized Shtarkov integral, $\mathsf{E}_{Z^n\sim P}\left[\exp\left(-\mathsf{E}_{\underline{f}\sim\hat{\Pi}\mid Z^n}\left[\eta R_{\underline{f}}(Z^n) + \log C(\underline{f}) - \log w(Z^n,\underline{f})\right]\right)\right]$	6
$\mathrm{COMP}(\mathcal{F},\hat{f})$	Complexity, $\eta^{-1}\log \mathrm{S}(\mathcal{F},\hat{f})$	5
$\mathrm{COMP}(\mathcal{F},\hat{f},w,z^n)$	Generalized complexity (deterministic version), $\frac{1}{\eta}\left(-\log w(z^n) + \log \mathrm{S}(\mathcal{F},\hat{f},w)\right)$	6
$\mathrm{COMP}(\mathcal{F})$	Maximal complexity, $\eta^{-1}\log \mathrm{S}(\mathcal{F}) = \sup_{\hat{f}} \mathrm{COMP}(\mathcal{F},\hat{f})$	5
$\mathrm{COMP}(\mathcal{F},\hat{\Pi},w,z^n)$	Generalized complexity, $\frac{1}{\eta}\cdot\left(\mathsf{E}_{\underline{f}\sim\hat{\Pi}\mid z^n}\left[-\log w(z^n,\underline{f})\right] + \log \mathrm{S}(\mathcal{F},\hat{\Pi},w)\right)$	6
$\mathrm{COMP}^{\mathrm{FULL}}(\mathcal{F},\hat{\Pi},w,z^n)$	Full generalized complexity, $\mathrm{COMP}(\mathcal{F},\hat{\Pi},w,z^n) + \mathsf{E}_{\underline{f}\sim\hat{\Pi}\mid z^n}[R_{\underline{f}}(z^n)]$	6
T_n	$\sup_{f\in\mathcal{F}}\left\{\sum_{j=1}^n\left(\ell_{f_0}(Z_j)-\ell_f(Z_j)\right) - \mathsf{E}_{Z^n\sim Q_{f_0}}\left[\sum_{j=1}^n\left(\ell_{f_0}(Z_j)-\ell_f(Z_j)\right)\right]\right\}$	11
$\mathsf{E}_{Z^n\sim Q_{f_0}}[T_n]$	H-local complexity	11
$\mathcal{N}(\mathcal{H},\|\cdot\|,\varepsilon)$	ε-covering number for $\mathcal{H}$ in the norm $\|\cdot\|$	10
$\mathcal{N}_{[]}(\mathcal{H},\|\cdot\|,\varepsilon)$	ε-bracketing number for $\mathcal{H}$ in the norm$\|\cdot\|$	10
$\mathcal{R}_n(\mathcal{H}\mid S^n)$	Empirical Rademacher complexity, $\mathsf{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\sup_{h\in\mathcal{H}}\left\lvert\frac{1}{n}\sum_{i=1}^n\epsilon_i h(S_i)\right\rvert\right]$	10
$\mathcal{R}_n(\mathcal{H})$	Rademacher complexity, $\mathsf{E}\left[\sup_{h\in\mathcal{H}}\left\lvert\frac{1}{n}\sum_{i=1}^n\epsilon_i h(S_i)\right\rvert\right]$	10

Appendix A. Further Discussion and Comparison to Existing Work

Implications and insights for log-loss individual sequence regret Our strategy for controlling $\mathrm{COMP}(\mathcal{F})$ owes much to an ingenious argument of Opper and Haussler (1999). They analyzed the minimax regret in the individual sequence prediction setting with log-loss, where the class of comparators is the set of static experts (i.e. experts that predict according to the same distribution in each round). Cesa-Bianchi and Lugosi (2001) obtain bounds in the more general setting where the comparator class consists of *arbitrary* experts that can predict conditionally on the past. For a further considerable extension within the realm of log-loss, see Rakhlin and Sridharan (2015) who allow prediction based not just on the past but also on side information, and who bound the minimax cumulative regret in terms of sequential complexities and sequential covering numbers. These three papers all end up with bounds in terms of variations of an L_∞ metric entropy. While the precise variation differs from paper to paper, due to the different generality of the setting, in all cases they would lead to vacuous bounds for classes $\mathcal{F}$ of static experts admitting finite $L_2(P)$ but no finite L_∞ covers, such as the set of nonincreasing densities we consider in Appendix B.

The present paper does handle such cases; we note, however, that unlike the non-i.i.d. setting of (Cesa-Bianchi and Lugosi, 2001; Rakhlin and Sridharan, 2015), the present paper is restricted to the unconditional (i.e. without side information) i.i.d./static experts setting. In some of their examples (see e.g. their Section 6), Rakhlin and Sridharan (2015) do consider settings with static experts (like ours), but (unlike us) predictions can make use of side information: thus, the observations are of the form $z_i = (x_i, y_i)$, the regret of a predictor $\tilde{p}$ is

$$\sum_{i=1}^{n} -\log \tilde{p}(y_i \mid x_i, z^{i-1}) - \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} (-\log p_f(y_i \mid x_i)),$$

and we want to find the $\tilde{p}$ achieving minimax regret. One might conjecture that our techniques are extendable to this conditional setting, based on the fact that any conditional density estimation problem can be turned into an unconditional one by fixing a distribution $p'(x)$ on $\mathcal{X}$ and defining $p'_f(x, y) := p_f(y|x)p(x)$. For Bayesian prediction strategies, the $-\log p(x_i)$ terms indeed cancel in the regret and the conditional and unconditional settings are equivalent; but this does not turn out to be the case for the minimax optimal prediction strategy, and it is at this point unclear whether our nonsequential $L_2(P)$-covering numbers can replace the sequential L_∞ covers of Rakhlin and Sridharan (2015) for the static, conditional setting.

Insights for the relation between excess risk bounds and data compression For general (possibly randomized) estimators, Corollary 9 with the choice $w(z^n, f) := \pi(f)/\hat{\pi}(f \mid z^n)$ for prior π provides an excess risk bound in terms of information complexity which itself can be expressed in terms of a (generalization of) the cumulative log-loss of a Bayesian sequential prediction strategy (Zhang, 2006a; Grünwald, 2012) defined relative to the constructed probability model $\mathcal{Q}_{\mathcal{F}}$. By the correspondence between codelengths and cumulative log-loss, we may say that we bound an *excess risk in terms of a codelength*. Recalling Figure 1, the equality ③ = ⑤ shows that we also get a useful excess risk bound in terms of the codelengths of the minimax (NML) code. In Appendix C, we extend this idea by showing, much more generally, that *every* luckiness function (up to scaling) uniquely defines an idealized code and vice versa; and moreover, every such luckiness function/code uniquely defines an excess risk bound.

Insights for excess risk bounds Theorem 20 offers a distribution-dependent bound that recovers minimax optimal rates for ERM. We can extend the result to obtain the optimal rates in the (easier) case of VC-type classes as well, as we show in Appendix D. In the long version of the paper, we further extend these results to general deterministic estimators. Theorem 20 and its extension to VC classes, Theorem 26, thus offer minimax optimal distribution-dependent excess risk bounds whose derivation we view as simpler than similar bounds based on local Rademacher complexities. In particular, our strategy completely avoids complicated (at least in the view of the authors) fixed point equations that have been used to obtain good excess risk bounds in other works (such as Koltchinskii and Panchenko (1999); Bartlett et al. (2005); Koltchinskii (2006)). We note, however, that the bounds in the present paper lack the kind of data-dependence exhibited by previous works leveraging local Rademacher complexities. Indeed, the bound in Theorem 20 is an exact oracle inequality which is distribution-dependent and, consequently, is not computable by a practitioner who does not know the β for which a Bernstein condition holds. In contrast, bounds obtained via local Rademacher complexities can be computed without distributional knowledge and have been shown to behave like the correct (but unknown to the practitioner) distribution-dependent bounds asymptotically (see Theorem 4.2 of Bartlett et al. (2005)).

Yet, the present work gives rise to results which allow a different kind of data-dependence: a PAC-Bayesian improvement for situations when the posterior distribution is close to a prior distribution; it also shows that, as long one restricts attention to oracle inequalities, there exists a single unifying proof technique and ensuing bound that can recover optimal rates for large and small classes, and (bounded versions of) log-loss and other losses alike.

Appendix B. A nontrivial regret bound for a class of monotone probability densities

B.1. Proof of Theorem 19

For convenience, we restate the definition of the class of monotone probability densities $\mathcal{P}$. Let $c_{\min}$ and $c_{\max}$ be positive constants satisfying $0 < c_{\min} < 1$ and $2 \leq c_{\max} < \infty$. Consider the class $\mathcal{P}$ of monotone probability densities on $[0,1]$ for which $c_{\min} \leq p(x) \leq c_{\max}$ for all $p \in \mathcal{P}$ and all $x \in [0,1]$.

We first show that, for all $\varepsilon > 0$, the $L_2(P)$ ε-entropy of the loss-composed class $\ell \circ \mathcal{P} := \{-\log p : p \in \mathcal{P}\}$ is finite. We will make use of Proposition 3.5.17 of Giné and Nickl (2016) (see also Theorem 2.7.5 of Van der Vaart and Wellner (1996)), restated here for convenience.

Theorem 21 *Let $\mathcal{F}$ be the class of monotone functions on $\mathbb{R}$ for which, for constants $-\infty < a < b < \infty$, we have for all $f \in \mathcal{F}$ and all $x \in \mathbb{R}$ that $a \leq f(x) \leq b$. We have for some positive constant K depending only on a and b, uniformly for all Borel probability measures P on $\mathbb{R}$,*

$$\log \mathcal{N}_{[]}(\mathcal{F}, L_2(P), \varepsilon) \leq \frac{K}{\varepsilon} \qquad \text{for } 0 < \varepsilon \leq b - a.$$

Taking $a = c_{\min}$, $b = c_{\max}$, and restricting to the domain to $[0,1]$, the same result holds for $\mathcal{F} = \mathcal{P}$. Since the result holds uniformly for all (Borel) probability measures on $\mathbb{R}$, it also holds for the empirical L_2 bracketing numbers.

Moreover, from Van der Vaart and Wellner (1996) (see the top display on page 84), for any probability measure P (including the empirical measures), the covering numbers are upper bounded

by the bracketing numbers:

$$\mathcal{N}(\mathcal{P}, L_2(P), \varepsilon) \le \mathcal{N}_{[]}(\mathcal{P}, L_2(P), \varepsilon).$$

Finally, since $r \mapsto \log r$ is $(1/c_{\min})$-Lipschitz on the domain $r \ge c_{\min}$, it holds that the loss-composed class $\ell \circ \mathcal{P}$ satisfies

$$\mathcal{N}(\ell \circ \mathcal{P}, L_2(P), \varepsilon) \le \mathcal{N}(\mathcal{P}, L_2(P), c_{\min} \cdot \varepsilon).$$

Consequently, we have, for some constant A depending only on $c_{\min}$ and $c_{\max}$,

$$\sup_{z_1,\ldots,z_n \in \mathcal{Z}} \mathcal{N}(\ell \circ \mathcal{P}, L_2(P_n), \varepsilon) \le \frac{A}{\varepsilon}.$$

Next, we observe that $\ell \circ \mathcal{P}$ satisfies the polynomial empirical entropy bound (23) with $\rho = \frac{1}{2}$, and so applying Theorem 17 with $\eta = 1$, we have

$$\mathrm{COMP}_1(\ell \circ \mathcal{P}) = O(n^{1/3}).$$

But, from (6), the $\mathrm{COMP}_1(\ell \circ \mathcal{P})$ is precisely equal to the minimax individual sequence regret relative to the class $\mathcal{P}$ under log-loss.

B.2. The sup-norm entropy of $\ell \circ \mathcal{P}$ fails to be finite

We now show that sup-norm ε-covering numbers of $\ell \circ \mathcal{P}$ are unbounded for sufficiently small ε. We actually show this to be true even when $\mathcal{P}$ is replaced by the following, significantly smaller parametric subclass:

$$\mathcal{P}_\Theta := \left\{ p_\theta : 0 \le \theta \le \frac{1 - c_{\min}}{c_{\max}} \right\},$$

where density p_θ is defined as

$$p_\theta(x) = \begin{cases} c_{\max} & \text{if } x \in [0, \theta], \\ \frac{1 - c_{\max} \cdot \theta}{1 - \theta} & \text{if } x \in (\theta, 1]. \end{cases}$$

We first show that the class of densities itself cannot admit a finite ε-cover in sup-norm (for small enough ε). We then show why an ε-cover in sup-norm for $\ell \circ \mathcal{P}_\Theta$ is a $(K\varepsilon)$-cover in sup-norm for $\mathcal{P}_\Theta$, thereby prohibiting the existence of the former for small enough ε.

We begin by taking any $\varepsilon \in (0, \frac{1}{2}]$, finite N, and $\theta_1 < \theta_2 < \cdots < \theta_N$. We will show that the set $\{p_{\theta_1}, p_{\theta_2}, \ldots, p_{\theta_N}\}$ cannot be a proper ε-cover (in sup-norm) for our parametric class.

To this end, let $\theta' = \frac{\theta_1 + \theta_2}{2}$ index a probability density in the parametric class. For any $x \in (\theta_1, \theta')$, it holds that $p_{\theta'}(x) = c_{\max} \ge 2$ whereas $p_{\theta_1}(x) = \frac{1 - c_{\max} \cdot \theta_1}{1 - \theta_1} \le 1$, and so $|p_{\theta_1}(x) - p_{\theta'}(x)| \ge 1$. Hence, $p_{\theta'}$ cannot be covered by an ε-ball centered at p_{θ_1}. Similarly, for any $x \in (\theta', \theta_2)$, we have that $|p_{\theta_2}(x) - p_{\theta'}(x)| \ge 1$ (and from monotonicity the same is true if replacing p_{θ_2} by p_{θ_j} for any $j > 2$). Consequently, the ε-covering property fails to hold.

Finally, suppose that there exists a finite proper ε-cover for $\ell \circ \mathcal{P}_\Theta$, say, $\mathcal{P}_{\Theta_\varepsilon} := \{p_\theta : \theta \in \Theta_\varepsilon\}$ for some finite set $\Theta_\varepsilon \subset \Theta$. Then, for any $\theta \in \Theta$, there exists $\theta' \in \Theta_\varepsilon$ such that, for all $x \in [0, 1]$, we have

$$|\log p_\theta(x) - \log p_{\theta'}(x)| \le \varepsilon.$$

But $r \mapsto e^r$ is $e^{\log c_{\max}}$-Lipschitz for $r \leq \log c_{\max}$, and so

$$\begin{aligned} |p_\theta(x) - p_{\theta'}(x)| &= |\exp(\log p_\theta(x)) - \exp(\log p_{\theta'}(x))| \\ &\leq c_{\max} |\log p_\theta(x) - \log p_{\theta'}(x)| \\ &\leq c_{\max} \cdot \varepsilon, \end{aligned}$$

a contradiction of the non-existence of a finite $(c_{\max} \cdot \varepsilon)$-cover for $\ell \circ \mathcal{P}_\Theta$ for small enough ε.

Appendix C. Cumulative log-loss interpretation

C.1. For deterministic estimators and general w

Fix an arbitrary estimator $\hat{f}$. Then for any luckiness function w with $\mathrm{S}(\mathcal{F}, \hat{f}, w) < \infty$, we can define the probability density

$$r_w(z^n) := \frac{q_{\hat{f}|z^n}(z^n) \cdot w(z^n)}{\mathrm{S}(\mathcal{F}, \hat{f}, w)}, \tag{30}$$

with $r(z^n) := \frac{q_{\hat{f}}(z^n)}{\mathrm{S}(\mathcal{F}, \hat{f})}$ (introduced in the last paragraph on page 5) being the special case with $w \equiv 1$. Just as with r_1, for general such w, r_w can be thought of as a sequential prediction strategy, and $\eta \cdot \textsc{comp}(\mathcal{F}, \hat{f}, w, z^n) - \log q_{\hat{f}_{|z^n}}(z^n) = -\log r_w(z^n)$ is the cumulative log-loss achieved by r_w. Different (up to scaling) w generate different log-loss prediction strategies (codes) and corresponding complexities. Conversely, for every probability density r' relative to ν on Z^n, we can set a luckiness measure $w(z^n)$ proportional to $r'(z^n)/q_{\hat{f}|z^n}(z^n)$; with the appropriately scaled choice of w, r_w will coincide r'; we thus have a 1-to-1-correspondence between luckiness functions w with $\mathrm{S}(\mathcal{F}, \hat{f}, w) < \infty$, codes and complexities.

Grünwald (2007) and Bartlett et al. (2013) consider various nonuniform w. To give but one example, if $\mathcal{Z} = \mathbb{R}^k$, $\mathcal{F} \subseteq \mathbb{R}^k$, and (for $f \in \mathcal{F}$) $q_f(z) = p_f(z) \propto \exp(-\|z - f\|_2^2/2)$ denotes a Gaussian with mean f, then a natural choice is to take w of Gaussian form $w(z^n) = \exp(-\|\hat{f}_{|z^n}\|_2^2)$ for a given estimator $\hat{f}$; the corresponding sequential prediction strategy r_w gives smaller cumulative log-loss to data with $\hat{f}_{|z^n}$ close to 0.

C.2. For general estimators and general w

Just as for deterministic estimators, we note that every randomized estimator $\hat{\Pi}$ and luckiness function w defines a probability density/prediction strategy on Z^n by setting

$$r_w(z^n) := \frac{\exp\left(\mathsf{E}_{\underline{f} \sim \hat{\Pi}|z^n}\left[\log q_{\underline{f}}(z^n) \cdot w(z^n, \underline{f})\right]\right)}{S(\mathcal{F}, \hat{\Pi}, w)},$$

and just as before, COMP can be interpreted in terms of the 'code' r_w.

Appendix D. Full version of Section 4 (includes results for VC-type classes)

For space, we limited the results appearing in the main text to large classes of polynomial entropy or polynomial entropy with bracketing. In this section, we restate Section 4.2 and the second part of Section 4.3 with the bounds for VC-type classes included.

D.1. From H-local complexity and Rademacher complexity to excess risk bounds

We now show concrete implications of our link between COMP, $\mathsf{E}[T_n]$, and $\mathcal{R}_n$ for three types of classes: classes of VC-type, classes with polynomial empirical entropy, and sets of classifiers of polynomial $L_2(P)$ entropy with bracketing. Let $\mathcal{H}$ be a class of functions over a space $\mathcal{S}$. The class $\mathcal{H}$ is said to be of *VC-type* if, for some $A \in (0,\infty)$ and $V > 0$, for all $\varepsilon > 0$, the empirical covering numbers of $\mathcal{H}$ satisfy

$$\sup_{s_1,\ldots,s_n \in \mathcal{S}} \mathcal{N}(\mathcal{H}, L_2(P_n), \varepsilon) \leq \left(A/\varepsilon\right)^V . \tag{31}$$

Class $\mathcal{H}$ is said to have *polynomial empirical entropy* if, for some $A \in (0,\infty), \rho \in (0,1)$, for all $\varepsilon > 0$, the empirical entropy of $\mathcal{H}$ satisfies

$$\sup_{s_1,\ldots,s_n \in \mathcal{S}} \log \mathcal{N}(\mathcal{H}, L_2(P_n), \varepsilon) \leq \left(A/\varepsilon\right)^{2\rho} . \tag{32}$$

We say class $\mathcal{H}$ has *polynomial $L_1(P)$ entropy with bracketing* if, for some $A \in (0,\infty), \rho \in (0,1)$, for all $\varepsilon > 0$, the $L_1(P)$ entropy with bracketing of $\mathcal{H}$ satisfies

$$\log \mathcal{N}_{[]}(\mathcal{H}, L_1(P), \varepsilon) \leq \left(A^2/\varepsilon\right)^{\rho}. \tag{33}$$

To obtain explicit bounds from Corollary 14, we require suitable upper bounds on either the Rademacher complexity $\mathsf{E}_{Z^n \sim Q_{f_k}}\left[\mathcal{R}_n(\mathcal{G}_k \mid Z^n)\right]$ or directly on the H-local complexity $\mathsf{E}_{Q_{f_k}}\left[T_n^{(k)}\right]$ itself for the above three types of classes. It is simple to obtain such bounds using Dudley's entropy integral, a product of the chaining technique of empirical process theory. However, the trick here is to somehow leverage that $\mathcal{G}_k$ has small $L_2(P)$ diameter. Koltchinskii (2011) (see equations (3.17) and (3.19)) obtained a bound which improve with reductions in the $L_2(P)$ diameter; we restate simplified versions of these bounds here. In the sequel, $\lesssim$ means inequality up to multiplication by a universal constant.

Theorem 22 *Let $\mathcal{H}$ be a class of functions over $\mathcal{Z}$, and let $Q \in \Delta(\mathcal{Z})$. Let $\sup_{h\in\mathcal{H}} \mathsf{E}_{Z\sim Q}[h(Z)^2] \leq \sigma^2$ and $U := \sup_{h\in\mathcal{H}} \|h\|_\infty$. Assume that $\mathcal{H}$ is of VC-type as in (31) with exponent V. Then, for $\sigma^2 \geq \frac{c}{n}$ (for some constant c)*

$$\mathsf{E}_{Z^n\sim Q}\left[\mathcal{R}_n(\mathcal{H} \mid Z^n)\right] \lesssim \max\left\{\sqrt{\frac{V}{n}}\sigma\sqrt{\log\frac{A}{\sigma}}, \frac{VU}{n}\log\frac{A}{\sigma}\right\}. \tag{34}$$

If instead $\mathcal{H}$ is of polynomial empirical entropy as in (32) with exponent ρ, then

$$\mathsf{E}_{Z^n\sim Q}\left[\mathcal{R}_n(\mathcal{H} \mid Z^n)\right] \lesssim \max\left\{\frac{A^\rho}{\sqrt{n}}\sigma^{1-\rho}, \frac{A^{2\rho/(\rho+1)}U^{(1-\rho)/(1+\rho)}}{n^{1/(1+\rho)}}\right\}. \tag{35}$$

For classes of polynomial entropy with bracketing, we appeal to upper bounds on $\mathsf{E}_{Q_{f_k}}\left[T_n^{(k)}\right]$. If the class $\mathcal{G}_k$ has small $L_1(Q_{f_k})$ diameter and, moreover, if it also has polynomial $L_1(Q_{f_k})$ entropy with bracketing, then Lemma A.4 of Massart and Nédélec (2006) provides precisely such a bound. Below, we present a straightforward consequence thereof.

Theorem 23 *Let $\mathcal{H}$ be a class of functions over $\mathcal{Z}$ with: polynomial entropy with bracketing as in (33) with exponent ρ; $\sup_{h\in\mathcal{H}} \mathsf{E}_{Z\sim Q}[|h(Z)|] \le \sigma^2$; and $\sup_{h\in\mathcal{H}} \|h\|_\infty \le 1$. If $Q \in \Delta(\mathcal{Z})$, then*

$$\mathsf{E}_{Z^n\sim Q}\left[\sup_{h\in\mathcal{H}}\left\{\frac{1}{n}\sum_{j=1}^{n} h(Z_j) - \mathsf{E}[h(Z)]\right\}\right] \lesssim \max\left\{A^{\rho}\sigma^{1-\rho}n^{-1/2}, A^{2\rho/(\rho+1)}n^{-1/(1+\rho)}\right\}. \quad (36)$$

The following theorem builds on Corollary 14 and *nearly* follows by plugging in either (34) or (35) into (22) and tuning ε in terms of n and η (which gives the VC case, (37)) , and plugging in (36) into (21) and then tuning (which gives the polynomial entropy case, (38)). The remaining work is to resolve a minor discrepancy between $L_2(P)$ pseudonorms and $L_2(Q_{f_k})$ pseudonorms (or the L_1 versions thereof). This theorem will allow us to show optimal rates under Bernstein conditions.

Theorem 24 *If $\mathcal{F}$ is of VC-type as in (31) with exponent V, then for all $\eta \in (0,1]$,*

$$\frac{\mathrm{COMP}_\eta(\mathcal{F})}{n} \lesssim V\log\frac{ALn}{V} \cdot n^{-1}\cdot\eta^{-1}. \quad (37)$$

If $\mathcal{F}$ has polynomial empirical entropy as in (32) or is a set of classifiers of polynomial entropy with bracketing as in (33) with exponent ρ, then, for all $0<\eta<1$,

$$\frac{\mathrm{COMP}_\eta(\mathcal{F})}{n} \lesssim (AL)^{\frac{2\rho}{1+\rho}}\cdot n^{-\frac{1}{1+\rho}}\cdot\eta^{-\frac{1-\rho}{1+\rho}}. \quad (38)$$

The proof of Theorem 24 can be found in Appendix E.4. We now prepare for our results on the rates of ERM on a class $\mathcal{F}$. In the following corollary, note that in both cases, the occurrence of the Bernstein exponent β (or κ^{-1}) is consistent with its occurrence in the simple finite $\mathcal{F}$ setting of (18).

Corollary 25 *Assume that a β-Bernstein condition holds for $\mathcal{F}$ as in Definition 6 for some β and B, and impose assumption (A1). Define $\kappa := \beta^{-1}$. Let $\hat{f}$ be ERM over $\mathcal{F}$. Then, (37) further implies, taking $\gamma = \left(B\left(\frac{V\log\frac{ALn}{V}}{n}\right)\right)^{\kappa/(2\kappa-1)}$ and taking n large enough so that $\frac{V}{n}\log\frac{ALn}{V} \le B^{1/(1-\beta)}$, that for all $z^n \in \mathcal{Z}^n$,*

$$\frac{\mathrm{COMP}_{v(\gamma)}(\mathcal{F},\hat{f})}{n} + \gamma \lesssim \left(B\left(V\log\frac{ALn}{V}\right)\right)^{\frac{\kappa}{2\kappa-1}}\cdot n^{-\frac{\kappa}{2\kappa-1}}. \quad (39)$$

Analogously, under such a Bernstein condition, (38) further implies, taking $\gamma = n^{-\frac{\kappa}{2\kappa-1+\rho}}$ and assuming that $n > (2B)^{-\frac{2\kappa-1+\rho}{\kappa-1}}$, that for all $z^n \in \mathcal{Z}^n$,

$$\frac{\mathrm{COMP}_{v(\gamma)}(\mathcal{F},\hat{f})}{n} + \gamma \lesssim \left((AL)^{\frac{2\rho}{\rho+1}} + 1\right)\cdot B^{\frac{1-\rho}{1+\rho}}n^{-\frac{\kappa}{2\kappa-1+\rho}}. \quad (40)$$

We used the notation $\kappa = \beta^{-1}$ here to make the results more easily comparable to Tsybakov (2004) and Audibert (2004); however, the result still holds for the case $\beta = 0$ if we simply replace κ^{-1} by β in all exponents above; e.g. $\kappa/(2\kappa-1)$ in (39) becomes $1/(2-\kappa^{-1}) = 1/(2-\beta)$.

D.2. Recovering Bounds under Bernstein Conditions for Large Classes

We now show how Lemma 8 can recover the optimal rates under the Tsybakov margin condition and the best known rates for ERM under Bernstein-type conditions. While other techniques can achieve the same rates for ERM, we feel that our approach embodies a simpler analysis for the polynomial entropy case; it also leads to new results for other estimators, as shown in the long version.

Theorem 26 *Assume that the β-Bernstein condition holds for $\mathcal{F}$ as in Corollary 25 and define $\kappa := \beta^{-1}$. Let $\hat{f}$ be ERM on $\mathcal{F}$. First suppose $\mathcal{F}$ is of VC-type as in (31) with exponent V. Then there is a universal constant C_1 such that for all n large enough so that $\frac{V}{n}\log\frac{ALn}{V} + \tau \leq B^{1/(1-\beta)}$, we have*

$$\mathsf{E}_{Z\sim P}\left[R_{\hat{f}}(Z)\right] \preceq_{\psi_1(n)} C_1 \left(B\left(V\log\frac{ALn}{V}\right)\right)^{\frac{\kappa}{2\kappa-1}} \cdot n^{-\frac{\kappa}{2\kappa-1}}, \tag{41}$$

where $\psi_1(n) = \frac{1}{6B}\cdot\left(B\left(V\log\frac{ALn}{V}\right)\right)^{(\kappa-1)/(2\kappa-1)} n^{\kappa/(2\kappa-1)} \asymp (\log n)^{(\kappa-1)/(2\kappa-1)}\cdot n^{\kappa/(2\kappa-1)}$. Analogously, suppose that $\mathcal{F}$ has polynomial empirical entropy as in (32) or is a set of classifiers of polynomial entropy with bracketing as in (33) with exponent ρ. Then there is a C_2 such that for all n large enough so that $n > (2B)^{-\frac{2\kappa-1+\rho}{\kappa-1}}$, we have

$$\mathsf{E}_{Z\sim P}\left[R_{\hat{f}}(Z)\right] \preceq_{\psi_2(n)} C_2\left[\left((AL)^{\frac{2\rho}{\rho+1}}+1\right)\cdot B^{\frac{1-\rho}{1+\rho}}\cdot n^{-\frac{\kappa}{2\kappa-1+\rho}}\right], \tag{42}$$

where $\psi_2(n) = \frac{1}{6}\cdot n^{\frac{\kappa+\rho}{2\kappa-1+\rho}}$.

In the long version, we extend this result to general deterministic estimators. The ESI (42) of Theorem 26 combined with part (ii) of Proposition 3 implies that, with probability at least $1-\delta$, ERM obtains the rate $n^{-\frac{\kappa}{2\kappa-1+\rho}} + n^{-\frac{\kappa+\rho}{2\kappa-1+\rho}}\cdot\log\frac{1}{\delta}$. For sets of classifiers of polynomial entropy with bracketing, the rate $n^{-\frac{\kappa}{(2\kappa-1+\rho)}}$ is known to be optimal, matching results of Tsybakov (2004, Theorem 1), Audibert (2004) (see the discussion after Theorem 3.3), and Koltchinskii (2006, p. 36). Outside of classification, for classes of polynomial empirical entropy the rate we obtain is to our knowledge the best known for ERM. In particular, if the nonparametric class is convex and the loss is exp-concave, then $\kappa = \beta = 1$, and our rates for ERM are minimax optimal (Rakhlin et al., 2017, Theorem 7). Yet, there are cases where $\beta < 1$ in which an aggregation scheme can obtain a rate as if $\beta = 1$; one such example is in the case of squared loss with a non-convex class (Rakhlin et al., 2017; Liang et al., 2015).

Appendix E. Proofs omitted from the main text

E.1. Proofs for Section 2

Proof of (5) The proof of this result is simple enough to state in just a few lines:

$$\begin{aligned}\textsc{comp}(\mathcal{F}, \hat{f}) &\leq \frac{1}{\eta} \log \int_{\mathcal{Z}^n} \max_{k \in \mathcal{K}} \sup_{f \in \mathcal{F}_k} q_f(z^n) d\nu(z^n) \\ &\leq \frac{1}{\eta} \log \int_{\mathcal{Z}^n} \sum_{k \in \mathcal{K}} \sup_{f \in \mathcal{F}_k} q_f(z^n) d\nu(z^n) \\ &\leq \frac{1}{\eta} \log |\mathcal{K}| + \frac{1}{\eta} \max_{k \in \mathcal{K}} \log \int_{\mathcal{Z}^n} \sup_{f \in \mathcal{F}_k} q_f(z^n) d\nu(z^n).\end{aligned}$$

■

Proof of Proposition 1 The first result follows since, by Jensen's inequality applied to (9), we have, using the definition of w and Fubini's theorem, that $\mathrm{S}(\mathcal{F}, \hat{\Pi}, w)$ is at most

$$\mathsf{E}_{Z^n \sim P} \mathsf{E}_{\underline{f} \sim \hat{\Pi}|z^n} \left[\frac{e^{-\eta R_{\underline{f}}(z^n)}}{C(\underline{f})} \cdot w(z^n, \underline{f}) \right] = \mathsf{E}_{Z^n \sim P} \mathsf{E}_{\underline{f} \sim \Pi} \left[\frac{e^{-\eta R_{\underline{f}}(z^n)}}{C(\underline{f})} \right] = \mathsf{E}_{\underline{f} \sim \Pi} \left[\int_{z^n} q_{\underline{f}}(z^n) d\nu(z^n) \right],$$

which equals 1. For (12), plug in our choice of w into the definition of COMP. ■

E.2. Proofs for Section 3

Proof of Proposition 3 Jensen's inequality yields (i). Apply Markov's inequality to $e^{-\eta(U-U')}$ for (ii). ■

Proof of Theorem 4 Let us abbreviate $\textsc{ann}(f) = n \, \mathsf{E}^{\textsc{ann},\eta}_{\bar{Z} \sim P} \left[R_f(\bar{Z}) \right]$. By the definition of ESI (13) we see that the statement in the theorem is equivalent to

$$\mathsf{E}_{Z^n \sim P} \left[\exp \left(\eta \cdot \left(\mathsf{E}_{\underline{f} \sim \hat{\Pi}_{|Z^n}} \left[\textsc{ann}(\underline{f}) \right] - \textsc{comp}^{\textsc{full}}(\mathcal{F}, \hat{\Pi}, w, Z^n) \right) \right) \right] = 1. \tag{43}$$

Plugging in the definition of $\textsc{comp}^{\textsc{full}}$ and then COMP, the left-hand side can be rewritten as

$$\mathsf{E} \left[\exp \left(\eta \cdot \left(\mathsf{E}_{\underline{f} \sim \hat{\Pi}_{|Z^n}} \left[\textsc{ann}(\underline{f}) - R_{\underline{f}}(Z^n) \right] - \frac{1}{\eta} \cdot \left(\mathsf{E}_{\underline{f} \sim \hat{\Pi}|Z^n} \left[-\log w(\underline{f}, Z^n) \right] + \log \mathrm{S}(\mathcal{F}, \hat{\Pi}, w) \right) \right) \right) \right]$$

$$= \frac{\mathsf{E} \left[\exp \left(\eta \cdot \left(\mathsf{E}_{\underline{f} \sim \hat{\Pi}_{|Z^n}} \left[\textsc{ann}(\underline{f}) - R_{\underline{f}}(Z^n) \right] - \frac{1}{\eta} \cdot \mathsf{E}_{\underline{f} \sim \hat{\Pi}|Z^n} \left[-\log w(\underline{f}, Z^n) \right] \right) \right) \right]}{\mathsf{E}_{Z^n \sim P} \left[\exp \left(- \mathsf{E}_{\underline{f} \sim \hat{\Pi}|z^n} \left[\eta R_{\underline{f}}(z^n) + \log C(\underline{f}) - \log w(z^n, \underline{f}) \right] \right) \right]},$$

where the denominator is just the definition of S. It is thus sufficient to prove that this expression is equal to 1. But this is immediate from the definition of $C(f)$ and $\textsc{ann}(\cdot)$. ■

Proof of Lemma 7 For clarity, let $\bar{a}$ and $\bar{b}$ refer to the constants a and b from part 1(a) of Theorem 5.4 of Van Erven et al. (2015). Apply that result with $\bar{b} = \frac{1}{2\bar{a}}$, $\bar{a} = 1/2$, and the u function there set to $x \mapsto Bx^\beta$. Note that although the statement of Theorem 5.4 actually imposes the stronger condition that the loss ℓ be $[0, 1/2]$-valued, the proof thereof only requires that $R_f \in [-1/2, 1/2]$ a.s. for all $f \in \mathcal{F}$. ■

Proof of Corollary 9 For (16), start with Lemma 8 with $\eta = v(\gamma)/2$ for the desired $\gamma > 0$, and then apply Theorem 4 to (stochastically) upper bound the annealed excess risk term. Since $v(\gamma) \leq 1$ by assumption, we have $C_{v(\gamma)/2} \leq 3$. For (17), start with (16), and take $w \equiv 1$ and $\hat{\Pi}(\cdot \mid \mathcal{Z}^n)$ equal to the Dirac measure on $\hat{f}_{|Z^n}$. From these settings and the optimality of ERM for the empirical risk, $\mathrm{COMP}^{\mathrm{FULL}}(\mathcal{F}, \hat{\Pi}, w, Z^n)$ reduces to the simpler form $\mathrm{COMP}(\mathcal{F}, \hat{f})$. ■

E.3. Proof of Theorem 10

We first prove the results that imply the first inequality of (19) and then prove the result that implies the second inequality.

E.3.1. PROOF OF FIRST INEQUALITY OF (19)

The first inequality of (19) from Theorem 10 is a consequence of Lemmas 11 and 12, which we prove in turn.

Proof of Lemma 11

$$\begin{aligned} e^{\eta \cdot \mathrm{COMP}_\eta(\mathcal{F})} = \mathrm{S}(\mathcal{F}) &= \mathsf{E}_{Z^n \sim Q_{f_0}} \left[\sup_{f \in \mathcal{F}} \frac{q_f(Z^n)}{q_{f_0}(Z^n)} \right] \\ &= \mathsf{E}_{Z^n \sim Q_{f_0}} \left[\exp\left(\sup_{f \in \mathcal{F}} \log \frac{q_f(Z^n)}{q_{f_0}(Z^n)} \right) \right] \\ &\leq \mathsf{E}_{Z^n \sim Q_{f_0}} \left[\exp\left(\sup_{f \in \mathcal{F}} \left\{ \log \frac{q_f(Z^n)}{q_{f_0}(Z^n)} - \mathsf{E}_{Z^n \sim Q_{f_0}} \left[\log \frac{q_f(Z^n)}{q_{f_0}(Z^n)} \right] \right\} \right) \right], \end{aligned}$$

where the inequality follows because the second term inside the supremum is a negative KL-divergence. Now, using the definition of Q_f and Q_{f_0}, the above is equal to

$$\mathsf{E}_{Z^n \sim Q_{f_0}} \left[\exp\left(\eta \underbrace{\sup_{f \in \mathcal{F}} \left\{ \sum_{j=1}^n \left(\ell_{f_0}(Z_j) - \ell_f(Z_j) \right) - \mathsf{E}_{Z^n \sim Q_{f_0}} \left[\sum_{j=1}^n \left(\ell_{f_0}(Z_j) - \ell_f(Z_j) \right) \right] \right\}}_{T_n} \right) \right].$$

■

It remains to prove Lemma 12.

Proof of Lemma 12 First, from our assumption on the loss and $\eta \leq 1$ together imply that

$$\sup_{f,g\in\mathcal{F}} \operatorname{ess\,sup} \left\{ \eta \left(\ell_f(Z) - \ell_g(Z) - \mathsf{E}[\ell_f(Z) - \ell_g(Z)] \right) \right\} \leq 1.$$

Our goal now is to be able to apply Talagrand's inequality. To this end, observe that

$$\sup_{f,g\in\mathcal{F}} \mathsf{Var}\left[\eta \left(\ell_f(Z) - \ell_g(Z) - \mathsf{E}[\ell_f(Z) - \ell_g(Z)] \right) \right] \leq \eta^2 \sup_{f,g\in\mathcal{F}} \|(\ell_f - \ell_g)\|^2_{L_2(Q_{f_0})}.$$

Now, *if* $\mathcal{F}$ had a small $L_2(Q_{f_0})$ diameter, then the Lipschitzness of the loss would imply that the above term is also small. However, by assumption, the class $\mathcal{F}$ is only known to have small $L_2(P)$ diameter (of at most ε). Lemma 27 (stated after this proof) effectively bridges the gap between these two pseudonorms, showing that

$$\sup_{f,g\in\mathcal{F}} \|\ell_f - \ell_g\|_{L_2(Q_{f_0})} \leq e\, L \sup_{f,g\in\mathcal{F}} \|f - g\|_{L_2(P)}, \tag{44}$$

which is then at most $e\, L\, \varepsilon = \sigma$.

Bousquet's version of Talagrand's inequality (see Theorem 2.3 of Bousquet (2002) or, for a more direct presentation, Theorem 12.5 of Boucheron et al. (2013)) now yields

$$\mathsf{E}_{Q_{f_0}}[e^{\lambda \eta T^{(k)}_{n,\eta}}] \leq \exp\left(\mathsf{E}_{Q_{f_0}}[\eta T^{(k)}_n] + (e^\lambda - (\lambda+1))(n\eta^2\sigma^2 + 2\,\mathsf{E}_{Q_{f_0}}[\eta T^{(k)}_n]) \right).$$

Inequality (20) now follows by taking $\lambda = 1$. ■

The following lemma was used to control the complexity of the class $\mathcal{F}$.

Lemma 27 *For the supervised loss parameterization,*

$$\|\ell_f - \ell_g\|_{L_2(Q_{f_0})} \leq e \cdot L \|f - g\|_{L_2(P)}. \tag{45}$$

For the direct parameterization,

$$\|\ell_f - \ell_g\|_{L_2(Q_{f_0})} \leq e \|f - g\|_{L_2(P)}. \tag{46}$$

Proof of Lemma 27 We first prove (45), the supervised loss parameterization result. The Lipschitz assumption on the loss implies that

$$\mathsf{E}_{(X,Y)\sim Q_{f_0}} \left[\left(\ell_f(X,Y) - \ell_g(X,Y) \right)^2 \right] \leq L^2\, \mathsf{E}_{X\sim Q_{f_0}} \left[\left(f(X) - g(X) \right)^2 \right].$$

Next, observe that for $\Delta(x) = \frac{q_{f_0}(x)}{p(x)}$

$$\mathsf{E}_{X\sim Q_{f_0}} \left[\left(f(X) - g(X) \right)^2 \right] = \mathsf{E}_{X\sim P} \left[\Delta(x) \left(f(X) - g(X) \right)^2 \right].$$

Since the inside of the expectation is nonnegative, it remains to upper bound $\Delta(x)$. By definition,

$$\Delta(x) = \frac{p(x) \int p(y \mid x) e^{-\eta R_f(x,y)} dy}{p(x)\, \mathsf{E}_{(\bar{X},\bar{Y})\sim P} \left[e^{-\eta R_f(\bar{X},\bar{Y})} \right]} = \frac{\mathsf{E}_{Y\sim P|X=x} \left[e^{-\eta R_f(x,Y)} \right]}{\mathsf{E}_{(\bar{X},\bar{Y})\sim P} \left[e^{-\eta R_f(\bar{X},\bar{Y})} \right]} \leq e^\eta \leq e,$$

since $\eta \leq 1$ and the excess loss random variable takes values in $[-1/2, 1/2]$.

We now prove the direct parameterization result (46). Observe that for $\Delta(z) = \frac{q_{f_0}(z)}{p(z)}$

$$\mathsf{E}_{Z\sim Q_{f_0}}\left[\left(\ell_f(Z) - \ell_g(Z)\right)^2\right] = \mathsf{E}_{Z\sim P}\left[\Delta(Z)\left(f(Z) - g(Z)\right)^2\right],$$

where we use the fact that $\ell_f = f$ for all $f \in \mathcal{F}$ in the direct parameterization. As above, it remains to upper bound $\Delta(z)$. By definition,

$$\Delta(z) = \frac{p(z)e^{-\eta R_f(z)}}{p(z)\,\mathsf{E}_{\bar{Z}\sim P}\left[e^{-\eta R_f(\bar{Z})}\right]} \leq e^{\eta} \leq e.$$

■

E.3.2. PROOF OF SECOND INEQUALITY OF (19)

The second inequality of (19) is a consequence of the first part of (19) and a standard empirical process theory result, Lemma 13. For completeness, we provide a proof of this result below.

Proof of Lemma 13 Recall that $\mathcal{G} = \{\ell_{f_0} - \ell_f : f \in \mathcal{F}\}$, and let $\epsilon_1, \ldots \epsilon_n$ be independent Rademacher random variables. In the below, both Z^n and $\bar{Z}^n$ are drawn from Q_{f_0}.

The following sequence of inequalities is a standard use of symmetrization from empirical process theory:

$$\begin{aligned}
&\mathsf{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{j=1}^n\left(\ell_{f_0}(Z_j) - \ell_f(Z_j)\right) - \mathsf{E}\left[\sum_{j=1}^n\left(\ell_{f_0}(\bar{Z}_j) - \ell_f(\bar{Z}_j)\right)\right]\right\}\right] \\
&= \mathsf{E}\left[\sup_{g\in\mathcal{G}}\left\{\sum_{j=1}^n g(Z_j) - \mathsf{E}\left[\sum_{j=1}^n g(\bar{Z}_j)\right]\right\}\right] \\
&\leq \mathsf{E}\left[\sup_{g\in\mathcal{G}}\sum_{j=1}^n\left(g(Z_j) - g(\bar{Z}_j)\right)\right] \\
&= \mathsf{E}\left[\sup_{g\in\mathcal{G}}\sum_{j=1}^n\epsilon_j\left(g(Z_j) - g(\bar{Z}_j)\right)\right] \\
&\leq 2\,\mathsf{E}\left[\sup_{g\in\mathcal{G}}\sum_{j=1}^n\epsilon_j g(Z_j)\right] \\
&\leq 2\,\mathsf{E}\left[\sup_{g\in\mathcal{G}}\left|\sum_{j=1}^n\epsilon_j g(Z_j)\right|\right].
\end{aligned}$$

■

E.4. Proof of Theorem 24 (subsumes proof of Theorem 17)

Proof of Theorem 24 Taking the results of Corollary 14 and dividing by n gives the two inequalities

$$\frac{\mathrm{COMP}_\eta(\mathcal{F})}{n} \le \frac{\log \mathcal{N}(\mathcal{F}, L_2(P), \varepsilon/2)}{n\eta} + \frac{3}{n} \max_{k \in [N_\varepsilon]} \mathsf{E}_{Z^n \sim Q_{f_k}} \left[T_n^{(k)} \right] + \eta\sigma^2 \tag{47}$$

and

$$\frac{\mathrm{COMP}_\eta(\mathcal{F})}{n} \le \frac{\log \mathcal{N}(\mathcal{F}, L_2(P), \varepsilon/2)}{n\eta} + 6 \max_{k \in [N_\varepsilon]} \mathsf{E}_{Z^n \sim Q_{f_k}} \left[\mathcal{R}_n(\mathcal{G}_k \mid Z^n) \right] + \eta\sigma^2, \tag{48}$$

where we remind the reader that $N_\varepsilon = \mathcal{N}(\mathcal{F}, L_2(P), \varepsilon/2)$.

In the below applications of Theorems 22 and 23, we make use of the following two observations. First, from Lemma 27 (which we previously applied to yield (44)), it follows that the $L_2(Q_{f_k})$ diameter of $\mathcal{G}_k$ is at most σ. Second, for any distribution $Q \in \Delta(\mathcal{Z})$, for all $u > 0$,

$$\mathcal{N}(\mathcal{G}_k, L_2(Q), u) = \mathcal{N}(\{\ell_f : f \in \mathcal{F}_{\varepsilon,k}\}, L_2(Q), u) \le \mathcal{N}(\mathcal{F}_{\varepsilon,k}, L_2(Q), u/L) \tag{49}$$

and (in the case of sets of classifiers)

$$\begin{aligned} N_{[]}(\mathcal{G}_k, L_2(Q_{f_k}), u) = \mathcal{N}_{[]}(\{\ell_f : f \in \mathcal{F}_{\varepsilon,k}\}, L_2(Q_{f_k}), u) &= \mathcal{N}_{[]}(\mathcal{F}_{\varepsilon,k}, L_2(Q_{f_k}), u) \\ &\le \mathcal{N}_{[]}(\mathcal{F}_{\varepsilon,k}, L_2(P), u/e); \end{aligned} \tag{50}$$

in both (49) and (50), the first equality holds because $\mathcal{G}_k$ is a shifted version of $\{\ell_f : f \in \mathcal{F}\}$. In the case of the supervised loss parameterization, the inequality in (49) holds from the Lipschitzness of the loss, and, in the case of the direct parameterization, the inequality is actually equality (recall that $L = 1$ in this case). The second equality of (50) holds because we only consider sets of classifiers with 0-1 loss. Lastly, the inequality in (50) is due to the 1-Lipschitzness of 0-1 loss for sets of classifiers and Lemma 27. From (49), if $\mathcal{F}$ is a VC-type class (and hence so is $\mathcal{F}_{\varepsilon,k}$), then $\mathcal{G}_k$ also is a VC-type class. Analogously, if $\mathcal{F}$ has polynomial empirical entropy, the same property extends to $\mathcal{G}_k$. From (50), if $\mathcal{F}$ is a class whose $L_2(P)$ entropy with bracketing is polynomial (and hence so is $\mathcal{F}_{\varepsilon,k}$), then $\mathcal{G}_k$ is a class whose $L_2(Q_{f_k})$ entropy with bracketing is polynomial with the same exponent.

VC-type classes. First, Theorem 28 (stated after this proof) implies that, for all $u > 0$,

$$\mathcal{N}(\mathcal{F}, L_2(P), u) \le \left(\frac{2A}{u}\right)^V.$$

Starting from (48), inequality (34) from Theorem 22 combined with (49) then implies that (coarsely using $\eta \le 1$)

$$\begin{aligned} \frac{\mathrm{COMP}_\eta(\mathcal{F})}{n} &\lesssim \frac{V \log \frac{4A}{\varepsilon}}{n\eta} + \max\left\{ \sqrt{\frac{V}{n}} \sigma \sqrt{\log \frac{AL}{\sigma}}, \frac{VU}{n} \log \frac{AL}{\sigma} \right\} + \eta\sigma^2 \\ &\lesssim \frac{V \log \frac{4A}{\varepsilon}}{n\eta} + \max\left\{ \sqrt{\frac{V}{n}} L\,\varepsilon \sqrt{\log \frac{A}{\varepsilon}}, \frac{VU}{n} \log \frac{A}{\varepsilon} \right\} + (L\,\varepsilon)^2. \end{aligned}$$

Finally, setting $\varepsilon = \frac{4}{L}\sqrt{\frac{V}{n}}$ yields (up to a universal multiplicative constant) the bound

$$\frac{V \log \frac{A\,L\,n}{V}}{n\eta} + \max\left\{ \frac{V}{n} \sqrt{\log \frac{ALn}{V}}, \frac{V}{n} \log \frac{ALn}{V} \right\} + \frac{V}{n} \lesssim \frac{V \log \frac{A\,L\,n}{V}}{n\eta},$$

where we used the assumption that $\eta \le 1$. This proves (37).

Classes of polynomial empirical entropy or polynomial entropy with bracketing. The first order of business is to control $N_\varepsilon = \log \mathcal{N}(\mathcal{F}, L_2(P), \varepsilon/2)$. In the case of classes of polynomial empirical entropy, we again invoke Theorem 28 to conclude that, for all $u > 0$,

$$\log \mathcal{N}(\mathcal{F}, L_2(P), u) \leq \left(\frac{2A}{u}\right)^{2\rho}.$$

In the case of sets of classifiers of polynomial entropy with bracketing, the $L_2(P)$ entropy can be controlled by the relationship

$$\log \mathcal{N}(\mathcal{F}, L_2(P), u) \leq \log \mathcal{N}_{[]}(\mathcal{F}, L_2(P), u) = \log \mathcal{N}_{[]}(\mathcal{F}, L_1(P), u^2) \leq \left(\frac{A}{u}\right)^{2\rho}.$$

Next, for *(i)* classes of polynomial empirical entropy, we start from (48) and apply inequality (35) from Theorem 22 combined with (49); or *(ii)* for classes of polynomial entropy with bracketing, we start from (47) and apply[3] Theorem 23 combined with (50); both cases imply that, for $0 < \eta \leq 1$, using $\rho < 1$,

$$\begin{aligned}\frac{\mathrm{COMP}_\eta(\mathcal{F})}{n} &\lesssim \frac{1}{n\eta}\left(\frac{2A}{\varepsilon}\right)^{2\rho} + \max\left\{\frac{(AL)^\rho}{\sqrt{n}}\sigma^{1-\rho}, \frac{(AL)^{2\rho/(\rho+1)}U^{(1-\rho)/(1+\rho)}}{n^{1/(1+\rho)}}\right\} + \eta\sigma^2 \\ &\lesssim \frac{1}{n\eta}\left(\frac{A}{\varepsilon}\right)^{2\rho} + \frac{A^\rho L}{\sqrt{n}}\varepsilon^{1-\rho} + \eta^{\frac{\rho-1}{\rho+1}} \cdot \frac{(AL)^{2\rho/(\rho+1)}}{n^{1/(1+\rho)}} + \eta \cdot (L\,\varepsilon)^2. \end{aligned} \tag{51}$$

(the enlargement of the third term will not affect the rates, as will now become clear). We now set $\epsilon := C_0 n^{-\frac{1}{2(1+\rho)}} \cdot \eta^{-\frac{1}{1+\rho}}$ for a constant $C_0 > 0$ to be determined later (this choice for ϵ was obtained by minimizing the sum of the first and second terms in the last line of (51) by setting the derivative to 0). With this choice, we get, as a very simple yet tedious calculation shows:

$$\begin{aligned} n^{-1}\eta^{-1}\epsilon^{-2\rho} &= C_0^{-2\rho} \cdot n^{-\frac{1}{1+\rho}} \cdot \eta^{\frac{\rho-1}{\rho+1}} \\ n^{-1/2}\epsilon^{1-\rho} &= C_0^{1-\rho} \cdot n^{-\frac{1}{1+\rho}} \cdot \eta^{\frac{\rho-1}{\rho+1}} \\ \eta\epsilon^2 &= C_0^2 \cdot n^{-\frac{1}{1+\rho}} \cdot \eta^{\frac{\rho-1}{\rho+1}} \end{aligned}$$

so that (51) becomes

$$\frac{\mathrm{COMP}_\eta(\mathcal{F})}{n} \lesssim C_{A,C_0,L} \cdot n^{-\frac{1}{1+\rho}} \cdot \eta^{\frac{\rho-1}{\rho+1}} \tag{52}$$

where

$$C_{A,C_0,L} = \left(\frac{2A}{C_0}\right)^{2\rho} + A^\rho L^\rho (C_0\, e)^{1-\rho} + (AL)^{2\rho/(\rho+1)} + (e\, L\, C_0)^2. \tag{53}$$

Plugging in $C_0 = A^{\rho/(\rho+1)} L^{-1/(\rho+1)}$, the four terms become of the same order:

$$\begin{aligned} C_{A,C_0,L} &\lesssim \left(L^{1/(\rho+1)} A^{1-\frac{\rho}{\rho+1}}\right)^{2\rho} + L^{1-\frac{1-\rho}{1+\rho}} A^{\rho+\frac{\rho(1-\rho)}{\rho+1}} + (AL)^{2\rho/(\rho+1)} + (L^{1-\frac{1}{1+\rho}} A^{\frac{\rho}{\rho+1}})^2 \\ &\lesssim (AL)^{2\rho/(\rho+1)}, \end{aligned}$$

3. Note that in classification, for any Q, the $L_1(Q)$ diameter is equal to the square of the $L_2(Q)$ diameter.

and (38) follows.

■

The above proof made use of the universal $L_2(P)$ metric entropy being essentially equivalent to the universal $L_2(P_n)$ metric entropy. This result extends an analogous result of Haussler (1995) for VC classes (see Corollary 1 therein).

Theorem 28 (Extended Haussler)
Let $\mathcal{F}$ be a class of functions over a space $\mathcal{S}$. Suppose that, for all $\varepsilon > 0$ and all $n \in \mathbb{N}$, there is some function $\psi\colon \mathbb{R}_+ \to \mathbb{N}$ such that

$$\sup_{s_1,\ldots,s_n \in \mathcal{S}} \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon) \leq \psi(\varepsilon).$$

Then, for any probability measure $P \in \Delta(\mathcal{S})$ and any $\varepsilon > 0$,

$$\mathcal{N}(\mathcal{F}, L_2(P), \varepsilon) \leq \psi(\varepsilon/2).$$

The proof is essentially due to Haussler with little change to the argument for the more general result.

Proof of Theorem 28 Let d be some pseudometric on $\mathcal{F}$. We say that $U \subset \mathcal{F}$ is ε separated if, for all $f, g \in U$, it holds that $d(f, g) > \varepsilon$. Let the ε-packing number $\mathcal{M}(\mathcal{F}, d, \varepsilon)$ be the maximal size of an ε-separated set in $\mathcal{F}$.

The packing numbers and covering numbers satisfy the following relationship (Vidyasagar, 2002, Lemma 2.2)

$$\mathcal{M}(\mathcal{F}, d, \varepsilon) \leq \mathcal{N}(\mathcal{F}, d, \varepsilon/2).$$

Thus, it is sufficient to bound $\mathcal{M}(\mathcal{F}, L_2(P), \varepsilon)$.

Suppose that $\mathcal{M}(\mathcal{F}, L_2(P), \varepsilon) > \mathcal{M}(\mathcal{F}, L_2(P_n), \varepsilon)$, and take U to be some ε-separated subset of $\mathcal{F}$ in the $L_2(P)$ pseudometric of cardinality $|U| > \mathcal{M}(\mathcal{F}, L_2(P_n), \varepsilon)$.

Next, draw $s_1, \ldots, s_n$ i.i.d. from P. Since U is finite, by taking n large enough we can ensure that the event $A_{f,g}$, defined as,

$$\|f - g\|_{L_2(P_n)} = \left(\frac{1}{n} \sum_{j=1}^{n} (f(s_j) - g(s_j)) \right)^{1/2} < \varepsilon,$$

occurs with probability at most $\frac{1}{|U|^2}$. Since $\binom{|U|}{2} < |U|^2$, it follows that the probability that no event $A_{f,g}$ occurs among all $f, g \in U$ is positive. Hence, there exists a set of points $s_1, \ldots, s_n$ for which U is an ε-packing in the $L_2(P_n)$ pseudometric. But then it must be the case that $\mathcal{M}(\mathcal{F}, L_2(P_n), \varepsilon) \geq |U|$, contradicting our assumption that $|U| > \mathcal{M}(\mathcal{F}, L_2(P_n), \varepsilon)$. ■

E.5. Proof of Corollary 25 (subsumes proof of Corollary 18)

Proof of Corollary 25 To see (39), we begin by upper bounding $\frac{\text{COMP}_{v(\gamma)}(\mathcal{F})}{n}$ using (37) with $\eta = v(\gamma) = \min\left\{\frac{\gamma^{1-\beta}}{B}, 1\right\}$ (from Lemma 7). Tentatively suppose that $B\gamma^{-(1-\beta)} \geq 1$; then $v(\gamma)^{-1} \lesssim B\gamma^{-(1-\beta)}$, and hence

$$\frac{\text{COMP}_{v(\gamma)}(\mathcal{F}, \hat{f})}{n} + \gamma \lesssim \frac{B}{n}\left(V \log \frac{ALn}{V} + \tau\right)\gamma^{-(1-\beta)} + \gamma.$$

Tuning γ such that it is equal to the first term on the RHS above yields (39); it is simple to verify that the supposition $B\gamma^{-(1-\beta)} \geq 1$ is ensured by the constraint on n stated in the corollary.

We now prove (40). Let $\gamma_n := n^{-\frac{\kappa}{2\kappa-1+\rho}}$ be the value of γ used at sample size n. We get from the definition of the Bernstein condition and Lemma 7 that

$$\eta_n := v(\gamma_n) = B^{-1}\left(n^{-\frac{\kappa}{2\kappa-1+\rho}}\right)^{(\kappa-1)/\kappa} = B^{-1}n^{-\frac{\kappa-1}{2\kappa-1+\rho}}$$

for all n for which the RHS above is at most 1. This will hold whenever $n \geq (1/B)^{\frac{2\kappa-1+\rho}{\kappa-1}}$. For such n, we will also have $\eta_n \leq 1$ and thus can apply (38), plugging in $\eta = \eta_n = v(\gamma_n)$. The result follows by simple algebra for all n larger than the given bound. ■

E.6. Proof of Theorem 26 (subsumes proof of Theorem 20)

Proof of Theorem 26 For both results, observe from Corollary 9 that, for all $\gamma > 0$,

$$\mathsf{E}_{Z\sim P}\left[R_{\hat{f}}(Z)\right] \trianglelefteq_{v(\gamma)\cdot n/6} \frac{3\left(\text{COMP}_{v(\gamma)/2}(\mathcal{F}, \hat{f})\right)}{n} + 4\gamma. \tag{54}$$

The result now follows by plugging in (39) and (40). ■