Journal of Machine Learning Research 21 (2020) 1-80 Submitted 7/18; Revised 11/19; Published 3/20

Fast Rates for General Unbounded Loss Functions: From ERM to Generalized Bayes

Peter D. Grünwald PDG@CWI.NL
Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
Leiden University, Mathematical Institute, Leiden, The Netherlands

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

Editor: Mehryar Mohri

Abstract

We present new excess risk bounds for general unbounded loss functions including log loss and squared loss, where the distribution of the losses may be heavy-tailed. The bounds hold for general estimators, but they are optimized when applied to η-generalized Bayesian, MDL, and empirical risk minimization estimators. In the case of log loss, the bounds imply convergence rates for generalized Bayesian inference under misspecification in terms of a generalization of the Hellinger metric as long as the learning rate η is set correctly. For general loss functions, our bounds rely on two separate conditions: the v-GRIP (generalized reversed information projection) conditions, which control the lower tail of the excess loss; and the newly introduced witness condition, which controls the upper tail. The parameter v in the v-GRIP conditions determines the achievable rate and is akin to the exponent in the Tsybakov margin condition and the Bernstein condition for bounded losses, which the v-GRIP conditions generalize; favorable v in combination with small model complexity leads to $\tilde{O}(1/n)$ rates. The witness condition allows us to connect the excess risk to an "annealed" version thereof, by which we generalize several previous results connecting Hellinger and Rényi divergence to KL divergence.

Keywords: Statistical Learning Theory, Fast Rates, PAC-Bayes, Misspecification, Generalized Bayes

1. Introduction

Much of statistical learning theory has operated under the restrictive assumption that the loss suffered for any prediction falls into some finite interval, which to say that the losses are bounded. In addition, much of this theory for deterministic estimators and even more so for randomized estimators only yields "slow" convergence rates of the risk of the predictor to the minimum risk achievable via the model in use; these are the best rates possible in the face of a worst case distribution. Faster rates of convergence are often possible under various, practically-applicable conditions on the learning problem, and showing such improvements is important as they can translate to drastic reductions on the number of examples needed to achieve a fixed level of error. We provide a novel theory of excess risk bounds for deterministic and randomized estimators in settings with general unbounded loss functions which may have heavy-tailed distributions — important applications include regression in situations with heavy-tailed noise and density estimation with log loss without assuming boundedness of likelihood ratios. These bounds have implications for two different areas: in statistical learning, they establish that with unbounded losses, under weak conditions, one can obtain estimators with *fast* convergence rates of their risk — such conditions previously were only well understood in the bounded case (earlier work on generalization bounds for unbounded loss functions such as (Meir and Zhang, 2003; Cortes et al., 2019) typically needs much stronger conditions to obtain fast rates). In density estimation under misspecification, the new bounds imply convergence rates for η-generalized Bayesian posteriors, in which the likelihood is raised to a power η not necessarily equal to 1, under surprisingly weak conditions. Finally, the bounds highlight the close similarity between PAC-Bayesian and η-generalized Bayesian learning methods under misspecification; these methods usually are studied within different communities. We now consider these applications in turn:

1. Statistical Learning In Statistical Learning Theory (Vapnik, 1995) the goal is to learn an action or predictor $\hat{f}$ from some set of actions, or *model*, $\mathcal{F}$ based on i.i.d. data $Z^n \equiv Z_1, Z_2, \ldots, Z_n \sim P$, where P is an unknown probability distribution over a sample space $\mathcal{Z}$. One hopes to learn an $\hat{f}$ with small risk, i.e., expected loss $\mathbf{E}[\ell_{\hat{f}}(Z)]$, for some given loss function ℓ. Here, $\mathbf{E}$ denotes expectation under P, and $\hat{f} \equiv \hat{f}(Z^n)$ is a function from $\mathcal{Z}^n$ to $\mathcal{F}$ that represents a learning algorithm; a prototypical example is empirical risk minimization (ERM). Thus, as is common, with some abuse of notation a learning algorithm is really a function, i.e., we do not insist it to be computable; and, in statistical contexts, we sometimes refer to learning algorithms as *estimators*, simply because this is common usage. A *learning problem* can thus be summarized as a tuple $(P, \ell, \mathcal{F})$. Well-known special cases include classification (with ℓ the 0-1 loss or some convex surrogate thereof) and regression (with ℓ the squared loss). As is customary (see e.g. (Bartlett et al., 2005) and (Mendelson, 2014)), in most of our results we assume existence of an optimal $f^* \in \mathcal{F}$ achieving $\mathbf{E}[\ell_{f^*}(Z)] = \inf_{f \in \mathcal{F}} \mathbf{E}[\ell_f(Z)]$, and we define the excess loss of f as $L_f = \ell_f - \ell_{f^*}$.

When the losses are almost surely bounded under P, there exists a well-established theory that gives optimal convergence rates of the excess risk $\mathbf{E}[L_{\hat{f}}]$ of estimator $\hat{f}$ in terms of sample size n. Broadly speaking, in the bounded case the optimal rate is usually of order

$$O\left(\left(\frac{\text{COMP}_n}{n}\right)^{\gamma}\right), \tag{1}$$

where COMP_n is a measure of model complexity such as the Vapnik-Chervonenkis (VC) dimension or the log-cardinality of an optimally chosen ϵ-net over $\mathcal{F}$, among others. For the models usually studied in statistics, such complexity measures are sublinear in n, and for "simple" models (often called parametric models, like those of finite VC dimension in classification) are finite or logarithmic in n. The exponent γ, which is in the range $[1/2, 1]$ in practically all cases of interest, reflects the *easiness* of a learning problem by depending on both geometric and statistical properties of $(P, \ell, \mathcal{F})$. This exponent is equal to 1/2 in the worst case but can be larger, allowing for faster rates, if the loss ℓ has sufficient curvature, e.g., if it is *exponentially concave (exp-concave)* or *mixable* (Cesa-Bianchi and Lugosi, 2006), or if $(P, \ell, \mathcal{F})$ satisfies "easiness" conditions such as the *Tsybakov margin condition* (Tsybakov, 2004), a *Bernstein condition* (Audibert, 2004; Bartlett and Mendelson, 2006), or *(stochastic) exp-concavity* (Juditsky et al., 2008). Because these conditions and the others on which this paper centers can allow for learning at faster rates, when any of the conditions hold a learning problem is intuitively easier. We thus call all such conditions *easiness conditions* throughout this work. In this literature, one often calls (1) with $\gamma = 1/2$ the *slow rate* and (1) with $\gamma = 1$ the *fast rate*. We note, however, that the terminology "fast rate" is somewhat imprecise, as there are special cases for which rates even faster than n^{-1} are possible (Audibert and Tsybakov, 2007). A more precise term may be "optimistic rate" (see (Mendelson, 2017a) for a lucid discussion), as this is the rate obtainable in the optimistic situation where an easiness condition holds. We opt for "fast" primarily for historical reasons.

Van Erven et al. (2015) showed that, in the case when the excess losses are bounded[1], all the "easiness" conditions above are subsumed by what they term the v-central condition, where v is a function that effectively modulates γ. While Van Erven et al. (2015) do show connections between such conditions for unbounded excess losses as well, they left open the question of whether the conditions still imply fast rates in that case. Thus, the first main target of the present paper is to extend this "fast rate theory" to the unbounded and heavy-tailed excess loss case. A main consequence of our bounds is that under *v-GRIP* conditions ("GRIP" stands for *generalized reversed information projection*), which consist of the v-central condition and a weakening thereof, and an additional *witness* condition, the obtainable rates remain the same as in the bounded case.

2. Density Estimation under Misspecification Letting $\mathcal{F}$ index a set of probability densities $\{p_f : f \in \mathcal{F}\}$ and setting the loss ℓ to the log loss, $\ell_f(z) = -\log p_f(z)$, we find that the statistical learning problem becomes equivalent to density estimation, the excess risk becomes equal to the *generalized Kullback-Leibler (KL) divergence*

$$D(f^* \parallel \hat{f}) = \mathbf{E}_{Z\sim P}[\log(p_{f^*}(Z)/p_{\hat{f}}(Z))],$$

and ERM becomes maximum likelihood estimation. We call a model $\mathcal{F}$ *well-specified* if it is correct, i.e., if p_{f^*} is the density of the true distribution P; in that case $D(f^* \parallel \hat{f})$ becomes the standard KL divergence. In this setting, our results thus automatically become convergence bounds of estimators $\hat{f}$ to the KL-optimal density within $\mathcal{F}$, where the convergence itself is in terms of KL divergence rather than more usual, weaker metrics such as Hellinger distance. Here, our results vastly generalize earlier results on KL bounds which typically rely on strong conditions such as boundedness of likelihood ratios or exponential tail conditions (Birgé and Massart, 1998; Yang and Barron, 1998; Wong and Shen, 1995; Sason and Verdú, 2016); in this work, the much weaker witness condition suffices.

We also provide bounds that are more similar to the standard Hellinger-type bounds and that hold without the witness condition, having a generalization of squared Hellinger distance (suitable for misspecification) rather than KL divergence on the left. Our bounds also allow for estimators that output a distribution Π on $\mathcal{F}$ rather than a single $\hat{f}$ and are particularly well-suited for η-generalized Bayesian posteriors, in which the likelihood in the prior-posterior update is raised to a power η; standard Bayes corresponds to $\eta = 1$. We thus can compare our rates to classical results on Bayesian rates of convergence in the well-specified case, such as in the influential paper (Ghosal, Ghosh, and van der Vaart, 2000) (GGV from now on). In this case, we generally obtain rates comparable to those of GGV,

[1] Van Erven et al. (2015) actually assume that the losses are bounded, but inspection of the results therein reveals that all that is needed is in fact bounded *excess* losses.

but under weaker conditions, as long as we take η (arbitrarily close to but) smaller than 1, a fact already noted for η-generalized Bayes by Zhang (2006a); Martin et al. (2017); Walker and Hjort (2002). In contrast to earlier work, however, our results remain valid in the misspecified case, although η has to be adjusted there to get convergence at all; moreover, the rates obtained are with respect to a new "misspecification metric" and hence are not always comparable to those obtained in the well-specified case. The optimal η depends on the "best" parameter v for which a v-GRIP condition holds. Grünwald and Van Ommen (2017) give a simple example which shows that taking $\eta = 1$ (standard Bayes) in regression under misspecification can lead to results that are dramatically worse than taking the right η, thus showing that our results do have practical implications.

3. η-generalized Bayes and PAC-Bayes The η-generalized Bayesian posterior can be further generalized: for general loss functions ℓ, we can define "posteriors" Π_n^B with densities given by

$$\frac{d\Pi_n^B}{d\Pi_0}(f) \equiv \pi_n^B(f) \equiv \pi^B(f \mid z_1, \ldots, z_n) := \frac{\exp\left(-\eta \sum_{i=1}^n \ell_f(z_i)\right)}{\int_{\mathcal{F}} \exp\left(-\eta \sum_{i=1}^n \ell_h(z_i)\right) \cdot d\Pi_0(h)}, \tag{2}$$

for some "prior" distribution Π_0 on $\mathcal{F}$. This idea goes back at least to Vovk (1990) and is central in the PAC-Bayesian approach to statistical learning (McAllester, 2003). Recently, it has also been embraced within the Bayesian community (Bissiri et al., 2016; Miller and Dunson, 2018). Nevertheless, the communities studying frequentist convergence of Bayesian methods under misspecification and PAC-Bayesian analysis are still largely separate; yet, the present paper shows that the approaches can be analyzed using the very same machinery and that it is fruitful to do so. To wit, *all* our results are based on an existing lemma due to T. Zhang (2006b; 2006a) which provides convergence bounds in terms of an "annealed" pseudo-excess risk for general estimators; these bounds are optimized if one plugs in η-generalized Bayesian estimators of the general form above. Zhang's bound is itself based on earlier works in the information theory literature (in particular, the Minimum Description Length (MDL) literature) (Barron and Cover, 1991; Li, 1999)) and the PAC-Bayesian literature (Catoni, 2003; Audibert, 2004). Of course, the technique also has some disadvantages, to which we return in the Discussion (Section 7).

1.1. Overview and Main Insights of the Paper

Section 2 formalizes the setting; Section 7 discusses additional related work and potential future work and provides discussion. The paper ends with appendices containing all long proofs, technical details concerning infinities, and some additional examples. The main results are in Sections 3–6:

Section 3: Zhang's Bound; Information Complexity In Section 3, for which we do not claim any novelty, we present Lemma 5; this lemma is T. Zhang's (2006b; 2006a) result that bounds a pseudo-excess risk of estimator $\hat{f} : \mathcal{Z}^n \to \mathcal{F}$ in terms of the *information complexity* $\mathrm{IC}_{n,\eta}$. A very simplified form of this lemma is

$$\mathbf{E}_{Z\sim P}^{\mathrm{ANN}(\eta)}\left[L_{\hat{f}}\right] \trianglelefteq_{\eta\cdot n} \mathrm{IC}_{n,\eta}, \tag{3}$$

where the pseudo-excess risk $\mathbf{E}^{\text{ANN}(\eta)}_{Z\sim P}$ is formally defined in (11) and $\trianglelefteq$ indicates *exponential stochastic inequality* (ESI), a useful notational tool which we define. ESI implies both inequality in expectation and with high probability over the sample Z^n that determines $\hat{f} \equiv \hat{f}(Z^n)$; the subscript $\eta{\cdot}n$ is only relevant for the in-probability version (see Proposition 3) and can be ignored for now. The actual bound (14) we provide in Lemma 5 generalizes (3), also allowing for estimators that output a distribution such as generalized Bayesian posteriors as given by (2). $\text{IC}_{n,\eta}$ is a notion of model complexity which, apart from n and η, also depends (for now suppressed in the notation) on the data Z^n, the choice of estimator $\hat{f}$ or Π_n, and on a distribution Π_0 on $\mathcal{F}$ which we may think of as "something like" a prior: while the bound holds for any fixed Π_0, the estimator that *minimizes* $\text{IC}_{n,\eta}$ for given prior Π_0 and data Z^n is the corresponding η-generalized Bayesian posterior Π^B_n given by (2).

For this choice of estimator, one can often design priors such that, with high probability and in expectation, $\text{IC}_{n,\eta}$ for the η-generalized Bayesian estimator can be upper bounded as

$$\text{IC}_{n,\eta} = \tilde{O}\left(\frac{\text{COMP}_n}{\eta n}\right), \tag{4}$$

for functions COMP_n that rely on the model $\mathcal{F}$'s complexity as indicated above (the $\tilde{O}$-notation suppresses logarithmic factors). In Section 3 we show that in the application to well-specified density estimation, priors can always be chosen such that the classical posterior contraction rates of GGV are (essentially) recovered for any fixed $\eta > 0$, in the sense that (3) would imply the same rates if the left-hand side were replaced by a squared Hellinger distance. For example, for standard finite and parametric statistical models, we obtain for Bayesian estimators that $\text{COMP}_n = \tilde{O}(1)$; for the nonparametric statistical models considered by GGV, we obtain $\text{COMP}_n = \tilde{O}(n^\alpha)$ for an α such that (4) becomes the minimax optimal rate. Similar bounds on $\text{IC}_{n,\eta}$ with general loss functions are given in Section 6. Henceforth, we use the term *parametric* to refer to $\mathcal{F}$ for which generalized Bayes estimators give $\text{COMP}_n = O(\log n) = \tilde{O}(1)$.

We would thus get good convergence bounds if the left-hand side of (3) were the actual excess risk, but instead it is an "annealed" version thereof, always smaller than the actual excess risk and sometimes even negative. All of our own results can be viewed as establishing conditions under which the annealed excess risk can either be related to the actual excess risk or otherwise to a (generalized Hellinger) metric measuring "distance" between f^* and f in some manner; this is done by modifying η. Both the information complexity and its upper bound (4) can only increase as we decrease η (Proposition 6); yet, for small enough η, annealed convergence implies convergence in the sense in which we are interested (either excess risk or generalized Hellinger distance) up to some constant factor (Sections 4 and 5) and sometimes with an additional slack term (Sections 5 and 6). Thus, the optimal η is given by a tradeoff between information complexity and these additional factors and terms.

Sections 4–6 each contain (a) a condition enabling a link between annealed excess risk and the divergence of interest in that section; (b) a new theoretical concept underlying the condition, (c) convergence result(s) relating information complexity to an actual metric or excess risk, and (d) example(s) that illustrate it.

Section 4: The Strong Central Condition and a New Metric; First Convergence Result The *strong central condition* (Van Erven et al., 2015) expresses that the lower tail of the excess loss $L_f := \ell_f - \ell_{f^*}$ is exponential, i.e., $P(\ell_{f^*} - \ell_f > A)$ is exponentially small in A. It has a parameter $\bar{\eta} > 0$ that determines the precise bound that can be obtained. While this may sound like a very strong condition, due to the nature of the log loss it automatically holds for density estimation with $\bar{\eta} = 1$ if the model is well-specified or convex. We show (Theorem 10) that the $\bar{\eta}$-strong central condition is sufficient for convergence in a new "misspecification" metric $d_{\bar{\eta}}$ (Definition 8) that generalizes the Hellinger distance: there exist estimators such that for every $0 < \eta < \bar{\eta}$,

$$d^2_{\bar{\eta}}(f^*, \hat{f}) \ \trianglelefteq_{\eta \cdot n} C_\eta \cdot \mathrm{IC}_{n,\eta},$$

where C_η is a constant that tends to ∞ as $\eta \uparrow \bar{\eta}$ and is bounded by 1 if $\eta \leq \bar{\eta}/2$. For misspecified models, $\bar{\eta}$ can in principle be either smaller or larger than 1. This metric is mainly of interest in the density estimation application of our work, and we thus compare our results to those of GGV for well-specified density estimation and illustrate them for the case of misspecified generalized linear models (GLMs). Plugging in any fixed $\eta < \bar{\eta}$ in (4) and comparing to (1), we see that under the strong central condition, we can always achieve the fast rate, i.e., (1) with $\gamma = 1$.

Section 5: The Witness Condition and a First Excess Risk Convergence Result Here we consider when, under the strong central condition, we can get bounds on the actual excess risk (or, in density estimation, on the generalized KL divergence). We provide a new concept, the *empirical witness of badness condition*, or *witness condition* for short, which provides control over the upper tail of the excess loss $L_f = \ell_f - \ell_{f^*}$ (whereas the central condition concerns the lower tail). Essentially, the witness condition says that whenever $f \in \mathcal{F}$ is worse than f^* in expectation, the probability that we witness this in our training example should not be negligibly small. We thus rule out the case that f has extremely large loss with extremely small probability. This condition turns out to be quite weak — it can still hold if, for example, the excess loss $\ell_f - \ell_{f^*}$ is heavy-tailed (it suffices for the conditional second moment of the target to be uniformly bounded almost surely; see Example 7). Thus we establish our first excess risk convergence result, Theorem 14, which, in its simplest form, says that if both the central condition holds with parameter $\bar{\eta}$ and the witness condition holds, then for all $0 < \eta < \bar{\eta}$,

$$\mathbf{E}[L_{\hat{f}}] \ \trianglelefteq_{\eta \cdot n / a_\eta} \ a_\eta \cdot \mathrm{IC}_{n,\eta}, \tag{5}$$

where a_η is a constant that again tends to ∞ as $\eta \uparrow \bar{\eta}$. Once again, by combining (5) and (4), we see that under a witness and $\bar{\eta}$-central condition, we can achieve the fast rate by taking $\gamma = 1$ in (1).

The witness condition vastly generalizes earlier conditions such as boundedness of likelihood ratios in density estimation (Birgé and Massart, 1998; Yang and Barron, 1998) and the exponential tail condition of Wong and Shen (1995). Moreover, (5) (Theorem 14) is based on Lemma 13, which generalizes earlier results relating KL divergence to Hellinger and Rényi-type divergences such as those of Yang and Barron (1999), Haussler and Opper (1997), Birgé and Massart (1998), Wong and Shen (1995), and Sason and Verdú (2016). We also discuss the similarity between the witness condition and the recently introduced *small-ball assumption* of Mendelson (2014).

Section 6: Weaker Fast Rate Conditions; the GRIP The $\bar{\eta}$-central condition of Section 4 can be generalized to the v-central condition, where $v : \mathbb{R}^+ \to \mathbb{R}^+$ is a nondecreasing function; nonconstant $v(x)$ gives weaker conditions that still allow for fast rates. Van Erven et al. (2015) showed that for the bounded excess loss case, most existing easiness conditions can be shown to be equivalent to either a v-central condition or to what they call a v-PPC *(pseudo-probability-convexity)* condition. In one of their central results, they show these two seemingly different conditions to be equivalent to one another, and also, if v is of the form $v(x) \asymp x^{1-\beta}$, (essentially) equivalent to a (B, β)-*Bernstein condition* (Audibert, 2004; Bartlett and Mendelson, 2006). In this section we show that for unbounded excess losses, the v-central and v-PPC conditions become quite different from each other (and also from the Bernstein condition): the v-PPC condition allows for heavy- (polynomial) tailed loss distributions, whereas the v-central condition does not.

We first present Theorem 22, an excess risk bound under the v-central condition that is a relatively straightforward consequence of Theorem 14, our risk bound under the $\bar{\eta}$-central condition. We then move to Theorem 29, a similar excess risk bound under the v-PPC condition. This theorem involves the *GRIP*, the novel, fundamental concept of this section (Definition 23). GRIP stands for *generalized reversed information projection* and generalizes the concept of reversed information projection introduced by Li (1999). The GRIP $m^{\eta}_{\mathcal{F}}$ is an η-dependent pseudo-predictor (it might achieve smaller risk than any f for which ℓ_f is defined). We show that, for each η, if f^* is replaced by the GRIP $m^{\eta}_{\mathcal{F}}$, then the convergence result (5) above holds. We can interpret the v-PPC condition as controlling the excess risk of f^* over the GRIP $m^{\eta}_{\mathcal{F}}$ as a function of η: the smaller η, the smaller this excess risk. This determines, for each sample size, an optimal η at which the bound (5) and the excess risk of f^* relative to $m^{\eta}_{\mathcal{F}}$ balance. Theorem 22 establishes that whenever the witness condition holds and a v-central condition holds, we have, for every $\epsilon > 0$, for $\eta < v(\epsilon)$,

$$\mathbf{E}[L_{\hat{f}}] \;\trianglelefteq_{\eta \cdot n / a'_{\eta}}\; a'_{\eta} \cdot \mathrm{IC}_{n,\eta} + \epsilon; \tag{6}$$

where again a'_{η} is a constant. Theorem 29 shows that if a v-PPC condition holds, the same result holds whenever $\eta < v(\epsilon)/2$, but now only in expectation, for yet another a'_{η}. Thus, the optimal rate now depends on v; in particular, if $v(\epsilon) \propto \epsilon^{1-\beta}$, then we can optimize over ϵ using upper bound (4) and find that, as long as $\textsc{comp}_n$ is logarithmic in n (as in parametric settings), by setting η at sample size n equal to $\eta \asymp n^{-(1-\beta)/(2-\beta)}$ we obtain the rate

$$\mathbf{E}[L_{\hat{f}}] = \tilde{O}\left(n^{-\frac{1}{2-\beta}}\right) \tag{7}$$

which interpolates between the fast rate ((1) with $\gamma = 1$) and the slow rate ($\gamma = 1/2$), where $\gamma = 1/(2-\beta)$ depends on β. Such calculations are well-known for the bounded loss case, and our results establish that the same story continues to hold for the unbounded excess loss case, as long as a witness condition holds — even for heavy-tailed losses. While Theorems 22 and 29 are applicable to the unbounded-loss-yet-bounded risk case (for which $\sup_{f \in \mathcal{F}} \mathbf{E}[\ell_f] < \infty$), Theorem 31 extends this result to the unbounded risk case, requiring a slight generalization of the witness condition. Examples 11 and 12 illustrate our results by considering regression with heavy-tailed losses, the latter example further linking the aforementioned small-ball assumption to our generalized witness condition.

The Picture that Emerges Our results point to three separate factors that determine achievable convergence rates for generalized Bayesian, two-part MDL, and empirical risk minimization (ERM) estimators, which often, but not always (see below) coincide with minimax rates:

1. The *information complexity* $\mathrm{IC}_{n,\eta}$, which determines the "richness" of the model. It is data- and algorithm- dependent, but we can often bound it with high probability or even independently of the underlying P. In addition, to see what rates can be achieved, we can plug in the (η-generalized Bayesian) learning algorithms that minimize it.

2. The *interaction between P, ℓ, and $\mathcal{F}$* that determines, for each $f \in \mathcal{F}$, the distribution of the *lower tail* of the excess loss L_f. This interaction is sometimes called the *easiness* of the problem (Koolen et al., 2016); it determines the optimal η at which a bound on η-information complexity implies a bound on the generalized Hellinger-type metric. This is captured by our v-GRIP conditions, which generalize several existing easiness conditions.

3. The *interaction between P, ℓ, and $\mathcal{F}$* that determines the distribution of the *upper tail* of the excess loss. This interaction plays *no* role for bounded excess losses and *no* role for density estimation if one only cares about convergence in the weak misspecification metric. Yet for unbounded excess losses with the excess risk target (or density estimation with KL-type target), this interaction becomes crucial to take into account and is done so via the witness condition.

In the Discussion (Section 7), Figure 1 summarizes how the various conditions hang together and are in some special cases (e.g. squared loss) implied by existing, better-known easiness conditions imposed in other works.

What We Do *Not* Cover We stress at the outset that we do not cover everything there is to know about the type of convergence bounds we prove. First of all, our bounds are most useful for ERM, η-generalized Bayesian, and MDL estimators, for a specific η that depends on the learning problem $(P, \ell, \mathcal{F})$ and often also on n. Thus to apply generalized Bayes/MDL in practice, η needs to be determined in some data-driven way; we discuss various ways to do this in Section 7. Note though that our bounds can be directly used for ERM, which can be implemented without knowledge of η.

We also leave untouched the fact that for parametric models, Zhang's bounds lead to an unnecessary $\log n$-factor in the convergence rates. Zhang (2006b; 2006a), following Catoni (2003), addresses this issue by a relatively straightforward "localized" modification of his bound; since it distracts from our main points (the witness and GRIP conditions, which lead to polynomial gains in rate), we will simply ignore all logarithmic factors in this paper.

Third, the new convergence rates for η-generalized Bayesian, MDL, and ERM estimators that we establish are in some cases, but not always, minimax optimal. We do explicitly discuss for each example below whether the obtained rates are optimal and discuss exceptions, unknowns, and potential remedies in Section 7.

Finally, we only discuss *proper* and *randomized proper* learning algorithms and estimators here. This means that our estimators either output an $\hat{f} \in \mathcal{F}$ or, if they output a distribution $\Pi \mid Z^n$, it is always a distribution on $\mathcal{F}$, and the quality of this distribution is evaluated by the expected loss incurred if one draws an f randomly from $\Pi \mid Z^n$. The terminology "proper" is from learning theory (Lee et al., 1996); in statistics such estimators are sometimes called "in-model" (Grünwald, 2007). In learning theory, one often considers more general "improper" set-ups in which one can play an element of (say) $\operatorname{conv}(\mathcal{F})$, the convex hull of $\mathcal{F}$, which sometimes improves the obtainable rates. We briefly return to this issue in Example 11 and Section 7.

Notation	Description	Page
General notation		
Z^n	i.i.d. sample; $Z^n = (Z_1, Z_2, \ldots, Z_n) \sim P^n$	2
P	Probability distribution over $\mathcal{Z}$	2
$\hat{f}$	Deterministic estimator or learning algorithm; $\hat{f} \equiv \hat{f}(Z^n)$	2
$(P, \ell, \mathcal{F})$	Learning problem for distribution P, loss function ℓ, and model $\mathcal{F}$	2
ℓ_f	Loss of hypothesis f; $\ell_f(z) \equiv \ell(f, z)$ and $\ell_f \equiv \ell_f(Z)$	11
f^*	Risk minimizer within $\mathcal{F}$	2
L_f	Excess loss (w.r.t. f^*) of f; $L_f(z) \equiv \ell_f(z) - \ell_{f^*}(z)$ and $L_f \equiv L_f(Z)$	2
Π_n^B (and π_n^B)	η-generalized Bayesian posterior (and its density relative to Π_0)	4
$\trianglelefteq_\eta$	Exponential stochastic inequality (E.S.I.)	14
$\Pi_{\vert}$	Randomized estimator or learning algorithm; $\Pi_{\vert} : \bigcup_{n=0}^{\infty} \mathcal{Z}^n \to \Delta(\mathcal{F})$	11
Π_n	Output of algorithm $\Pi_{\vert}$ based on sample Z^n; $\Pi_n \equiv \Pi \mid Z^n$	11
Π_0	Prior; $\Pi_0 \equiv \Pi \mid \{\}$	13
μ	Common dominating measure for $\{p_f\}_{f \in \mathcal{F}}$ in the case of log loss	12
$\ddot{f}_{\text{2-P}}$	η-generalized two-part MDL estimator for prior Π_0 at sample size n	13
$(\hat{f}, \Pi_0)$	Deterministic estimator $\hat{f}$ viewed as randomized estimator	13
$\mathbf{E}^{\mathrm{HE}(\eta)}[U]$	Hellinger-transformed expectation; $\mathbf{E}^{\mathrm{HE}(\eta)}[U] = \frac{1}{\eta}\left(1 - \mathbf{E}\left[e^{-\eta U}\right]\right)$	14
$\mathbf{E}^{\mathrm{ANN}(\eta)}[U]$	Annealed expectation; $\mathbf{E}^{\mathrm{ANN}(\eta)}[U] = -\frac{1}{\eta} \log \mathbf{E}\left[e^{-\eta U}\right]$	14
$\mathrm{IC}_{n,\eta}(\Pi_{\vert})$	Information complexity	15
$p_{f,\eta}$	Entropified loss; $p_{f,\eta}(z) = p(z) \frac{\exp(-\eta L_f(z))}{\mathbf{E}[\exp(-\eta L_f(Z))]}$	19
$d_{\bar{\eta}}(\cdot, \cdot)$	Misspecification metric	20
$\mathcal{N}(\mathcal{A}, \Vert \cdot \Vert, \epsilon)$	ε-covering number of $(\mathcal{A}, \Vert \cdot \Vert)$	32
Divergences		
$\mathrm{KL}(\cdot \Vert \cdot)$	Standard Kullback-Leibler divergence	13
$\mathrm{H}_{1/2}(\cdot \Vert \cdot)$	Standard (squared) Hellinger distance	20
$\mathrm{H}_\eta(\cdot \Vert \cdot)$	η-generalized Hellinger divergence	20
$D_\alpha(p \Vert q)$	Rényi divergence of order α; $D_\alpha(p \Vert q) = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} d\mu$	47
Pseudo-predictors		
$\bar{\mathcal{F}}$	enlarged action space $\bar{\mathcal{F}} \supseteq \mathcal{F}$ that also contains pseudo-predictors	30
f_ϵ^*	pseudo-predictor, defined via its loss by $\ell_{f_\epsilon^*}(z) = \ell_{f^*}(z) - \epsilon$, $\forall z \in \mathcal{Z}$	31
$\mathcal{E}_{\mathcal{F},\eta}$	set of pseudoprobability densities; $\mathcal{E}_{\mathcal{F},\eta} = \left\{e^{-\eta \ell_f} : f \in \mathcal{F}\right\}$	33
ξ_Q	mixture of pseudoprobability densities; $\xi_Q = \mathbf{E}_{\underline{f} \sim Q}[e^{-\eta \ell_{\underline{f}}}]$	33
$m_{\mathcal{F}}^\eta$ or ℓ_{g_η}	GRIP; $\mathbf{E}[m_{\mathcal{F}}^\eta] = \inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}\left[-\frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim Q}\left[e^{-\eta \ell_{\underline{f}}}\right]\right]$	33
m_Q^η	mix loss for $Q \in \Delta(\mathcal{F})$; $m_Q^\eta = -\frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim Q}\left[e^{-\eta \ell_{\underline{f}}}\right]$	33
m_A^η	generalized GRIP w.r.t. $A \subseteq \bar{\mathcal{F}}$; $\mathbf{E}\left[m_A^\eta\right] = \inf\limits_{Q \in \Delta(A \cup \{f^*\})} \mathbf{E}\left[m_Q^\eta\right]$	33
m_f^η	mini-grip w.r.t. f; $\mathbf{E}[m_f^\eta] = \inf\limits_{\alpha \in [0,1]} \mathbf{E}\left[-\dfrac{1}{\eta} \log\left((1-\alpha) e^{-\eta \ell_{f^*}} + \alpha e^{-\eta \ell_f}\right)\right]$	58
$g_{\mathcal{F}}^\eta$ and g_f^η	pseudo-actions for GRIP losses $m_{\mathcal{F}}^\eta$ and m_f^η respectively	59

Notation	Description	Page
Conditions		
(β, B)-Bernstein	$\mathbf{E}[L_f^2] \leq B\left(\mathbf{E}[L_f]\right)^{\beta}$ for all $f \in \mathcal{F}$	28
strong $\bar{\eta}$-central	$\exists f^* \in \mathcal{F}$ s.t. $\ell_{f^*} - \ell_f \trianglelefteq_{\bar{\eta}} 0$ for all $f \in \mathcal{F}$	19
η-central up to ε	$\exists f^* \in \mathcal{F}$ s.t. $\ell_{f^*} - \ell_f \trianglelefteq_{\eta} \epsilon$ for all $f \in \mathcal{F}$	31
v-central	for all $\varepsilon \geq 0$, $\exists f^* \in \mathcal{F}$ s.t. $\ell_{f^*} - \ell_f \trianglelefteq_{v(\varepsilon)} \epsilon$ for all $f \in \mathcal{F}$	31
η-PPC up to ε	$\exists f^* \in \mathcal{F}$ s.t. $\mathbf{E}_{Z \sim P}\left[\ell_{f^*} - m_{\mathcal{F}}^{\eta}\right] \leq \epsilon$	34
v-PPC	for all $\varepsilon \geq 0$, $\exists f^* \in \mathcal{F}$ s.t. $\mathbf{E}_{Z \sim P}\left[\ell_{f^*} - m_{\mathcal{F}}^{v(\varepsilon)}\right] \leq \epsilon$	34
(u, c)-witness	$\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u\}}\right] \geq c\,\mathbf{E}[\ell_f - \ell_{f^*}]$ for all $f \in \mathcal{F}$	24
(τ, c)-witness	generalized version of (u, c)-witness condition (see Definition 12)	25
witness w.r.t. ϕ	(u, c)-witness condition with dynamic comparator (see Assump. 1)	58
weak witness w.r.t. ϕ	weakened version of the previous condition (see Assumption 1)	58
unif. exp. upper tail	U_f (for $f \in \mathcal{F}$) has condition if $\exists \kappa \in (0, \infty)$ s.t. $\sup_{f \in \mathcal{F}} \mathbf{E}\left[e^{\kappa U_f}\right] < \infty$	27
small-ball assumption	$\exists \kappa > 0$ and $\epsilon \in (0,1)$ s.t. $\forall f, h \in \mathcal{F}$, $\Pr\left(\lvert f - h \rvert \geq \kappa \lVert f - h \rVert_{L_2(P)}\right) \geq \varepsilon$	29
convex luckiness	(for squared loss); $\arg\min_{f \in \mathcal{F}} \mathbf{E}[\ell_f] = \arg\min_{f \in \mathrm{conv}(\mathcal{F})} \mathbf{E}[\ell_f]$	28

2. Setting, Technical Preliminaries, Global Assumptions

We now formally introduce the problem setting, cover some preliminaries, and state the assumptions used throughout this work. A glossary appearing on this page and the last one describes all frequently used symbols and conditions.

Let $\ell_f(z) := \ell(f, z) \in \mathbb{R} \cup \{\infty\}$ denote the loss of action $f \in \mathcal{F}$ under outcome $z \in \mathcal{Z}$. In the classical statistical learning problems of classification and regression with i.i.d. samples, we have $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Classification (0-1 loss) is recovered by taking $\mathcal{Y} = \{0, 1\}$ and $\ell_f(x, y) = \lvert y - f(x) \rvert$, and we obtain regression with squared loss by taking $\mathcal{Y} = \mathbb{R}$ and $\ell_f(x, y) = (y - f(x))^2$. In either case, the class $\mathcal{F}$ is some subset of the set of all functions $f : \mathcal{X} \to \mathcal{Y}$, such as the set of decision trees of depth at most 5 for classification. Our setting also includes conditional density estimation (see Example 1). Unless we explicitly state otherwise, whenever we introduce a random variable we assume it is a function of $Z, Z_1, \ldots, Z_n$ which are i.i.d. $\sim P$. If we write ℓ_f we mean $\ell_f(Z)$.

While in frequentist statistics one mostly considers learning algorithms (often called "estimators") that always output a single $f \in \mathcal{F}$, we also will consider algorithms that output *distributions* on $\mathcal{F}$. Such distributions can, but need not, be Bayesian or generalized Bayesian posteriors as described below. Formally, a learning algorithm based on a set of predictors $\mathcal{F}$ is a function $\Pi_{|} : \bigcup_{n=0}^{\infty} \mathcal{Z}^n \to \Delta(\mathcal{F})$, where Δ is the set of distributions on $\mathcal{F}$. The output of algorithm $\Pi_{|}$ based on sample Z^n is written as $\Pi \mid Z^n$ and abbreviated to Π_n. Π_n is a function of Z^n and hence a random variable under P. For fixed given z^n, $\Pi \mid z^n$ is a measure on $\mathcal{F}$. Importantly, our learning algorithms are always defined such that they can also output a distribution Π_0 based on an empty data sequence; we may think of this as a "prior" guess of f. We explain below how to recast standard estimators such as ERM, for which Π_0 is undefined, in this framework. Whenever we consider a distribution Π on $\mathcal{F}$ for a problem $(P, \ell, \mathcal{F})$, we denote its outcome, a random variable, as $\underline{f}$. Whenever we compare the performance of a learning algorithm $\Pi_{|}$ to a fixed $\tilde{f} \in \mathcal{F}$, we call $\tilde{f}$ a *comparator*. $\tilde{f}$ is called *optimal* or *risk-minimizing* if $\mathbf{E}[\ell_f(Z) - \ell_{\tilde{f}}(Z)] \geq 0$ for all $f \in \mathcal{F}$; under the assumptions below, this expectation is always well-defined. We usually (but not in Section 6 and the proofs) take as our comparator $\tilde{f} = f^*$, where f^* is a risk minimizer. Whenever this cannot cause confusion, we write $L_f = \ell_f - \ell_{f^*}$ for the *excess loss* relative to f^*.

Assumptions on Learning Algorithms $\Pi_|$ Whenever in the sequel we mention a learning algorithm $\Pi_|$, we make the following (very mild) assumptions: (1) for all n, $z^n \in \mathcal{Z}^n$, Π_n has a density $\pi_n \equiv \pi \mid z^n$ relative to the prior distribution Π_0; (2) Π_0 satisfies the natural requirement that for all $z \in \mathcal{Z}$, $\Pi_0(f \in \mathcal{F} : \ell_f(z) < \infty) > 0$.

Assumptions on and Conventions for Learning Problems $(P, \ell, \mathcal{F})$ All of our mathematical results concern learning problems $(P, \ell, \mathcal{F})$ for which we invariably make the following assumptions:

1. Unless the loss function ℓ is log-loss or conditional log-loss (see the example below), it is is uniformly bounded from below in the sense that $\inf_{f \in \mathcal{F}} \inf_{z \in \mathcal{Z}} \ell_f(Z) > -\infty$.

2. For (conditional) log-loss, we assume for all $f \in \mathcal{F}$ that p_f is a probability density relative to some fixed common dominating measure μ, so that P_f, the distribution with density p_f, is absolutely continuous with respect to μ; we also assume that P itself is absolutely continuous with respect to μ. Moreover, we additionally assume that

$$\mathrm{KL}(P \,\|\, P_{f^*}) < \infty \tag{8}$$

   and, with $H(P)$ the differential entropy of P relative to μ,

$$H(P) > -\infty. \tag{9}$$

3. The learning problem is nontrivial in the sense that for some $f \in \mathcal{F}$, $\mathbf{E}_{Z \sim P}[\ell_f(Z)] < \infty$ (we require this irrespective of whether ℓ is log-loss).

4. There exists an optimal $f \in \mathcal{F}$. We fix any one among these (our results hold no matter which we take) and denote it by f^*.

Some of our results continue to hold without the final assumption; we shall in all cases say so explicitly. Since we invariably want to impose these assumptions, from now on learning problems $(P, \ell, \mathcal{F})$ are *defined* to be such that they satisfy these four assumptions, and we will not explicitly mention them any more. The assumptions, and all other issues concerning unboundedness and infinities, are discussed in detail in Appendix H. The requirement that the loss is bounded from below ensures that there are no issues involving undefined expectations or problems with interchanging order of expectations, as we show in Appendix H.1. It holds for just about all loss functions encountered in the literature, except for log-loss defined on continuous outcome spaces, where the log-loss can be unbounded both from above and below; in Appendix H.2 we motivate the requirements we impose on log-loss and show that, while very mild, they are still sufficient to make all expectations well-defined.

Example 1 (Conditional Density Estimation) Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and let $\{p_f \mid f \in \mathcal{F}\}$ be a statistical model of conditional densities for $Y \mid X$, i.e., for each $x \in \mathcal{X}$, $p_f(\cdot \mid x)$ is a probability density on $\mathcal{Y}$ relative to a fixed underlying measure μ. Take (conditional) *log loss*, defined on outcome $z = (x, y)$ as $\ell_f(x, y) = -\log p_f(y \mid x)$. The excess risk, now $\mathbf{E}[L_f] = \mathbf{E}_{Z \sim P}\Big[\log \frac{p_{f^*}(Y|X)}{p_f(Y|X)}\Big]$, is formally equivalent to the *generalized KL divergence*, as already defined in the original paper by Kullback and Leibler (1951) that also introduced what is now the "standard" KL divergence. Assuming that P has a density p relative to the underlying measure, and denoting standard KL divergence by KL, we have $\mathrm{KL}(p \,\|\, p_f) = \mathbf{E}_{Z \sim P}\left[\log \frac{p(Y|X)}{p_f(Y|X)}\right]$, so that $\mathbf{E}[L_f] = \mathrm{KL}(p \,\|\, p_f) - \mathrm{KL}(p \,\|\, p_{f^*})$. Thus, minimizing the excess risk under log loss is equivalent to learning a distribution minimizing the KL divergence from P over $\{p_f : f \in \mathcal{F}\}$. We have $\inf_{f \in \mathcal{F}} \mathrm{KL}(p \,\|\, p_f) = \mathrm{KL}(p \,\|\, p_{f^*}) = \epsilon \geq 0$. If $\epsilon = 0$, we must have $p_{f^*} = p$, so we deal with a standard *well-specified* density estimation problem, i.e., the model $\{p_f \mid f \in \mathcal{F}\}$ is "correct" and $f^* \in \mathcal{F}$ represents the true P. If $\epsilon > 0$, we still have $\inf_{f \in \mathcal{F}} \mathbf{E}[L_f] = 0$ and may view our problem as learning an f that is closest to f^* in generalized KL divergence. □

Generalized (PAC-) Bayesian, Two-Part, and ERM Estimators Although our main results hold for general estimators, Proposition 6 below indicates that they are especially suited for generalized Bayesian, two-part MDL, or ERM estimators, since these minimize the bounds provided by our theorems under various constraints. To define these estimators, fix a distribution Π_0 on $\mathcal{F}$, henceforth called *prior*, and a *learning rate* $\eta > 0$. The *η-generalized Bayesian posterior* based on prior Π_0, $\mathcal{F}$ and sample $z_1, \ldots, z_n$ is the distribution Π_n^B on $f \in \mathcal{F}$, defined by (2). By our requirement that for all $z \in \mathcal{Z}$, $\Pi_0(f \in \mathcal{F} : \ell_f(z) < \infty) > 0$, (2) is guaranteed to be well-defined.

Now, given a learning problem as defined above, fix a countable subset $\ddot{\mathcal{F}}$ of $\mathcal{F}$, a distribution Π_0 concentrated on $\ddot{\mathcal{F}}$ and define the *η-generalized two-part MDL estimator for prior* Π_0 *at sample size* n as

$$\ddot{f}_{\text{2-P}} := \arg\min_{f \in \ddot{\mathcal{F}}} \sum_{i=1}^{n} \ell_f(Z_i) + \frac{1}{\eta} \cdot (-\log \Pi_0(\{f\})), \tag{10}$$

where, if the minimum is achieved by more than one $f \in \ddot{\mathcal{F}}$, we take the smallest in the countable list, and if the minimum is not achieved, we take the smallest f in the list that is within $1/n$ of the minimum. Note that the η-two part estimator is *deterministic*: it concentrates on a single function. ERM is recovered for finite $\mathcal{F}$ by setting the prior Π_0 to be uniform over $\ddot{\mathcal{F}}$. We may view the η-two part estimator as a learning algorithm $\Pi_|$ in our sense by defining Π_0 to be the prior on $\ddot{\mathcal{F}}$ as above and, for each n, Π_n as the distribution that puts all of its mass at $\ddot{f}_{\text{2-P}}$ at sample size n. While we could denote this estimator as $\Pi_{|\text{2-P}}$, it will be convenient to write $(\ddot{f}_{\text{2-P}}, \Pi_0)$ so as to also specify the prior. In the same way, general priors Π_0 combined with general deterministic estimators $\hat{f}$ defined for samples of length ≥ 1 may be viewed as learning algorithms $\Pi_|$ which we will denote as $(\hat{f}, \Pi_0)$.

Finally, we formally define the ERM estimator as the $f \in \mathcal{F}$ that minimizes $\sum_{j=1}^{n} \ell_f(Z_j)$; whenever we refer to ERM we will make sure that at least one such f exists; ties can then be broken in any way desired. It is important to note that ERM can be applied without knowledge of η; however, for general two-part and Bayesian estimators we need to know η — we return to this issue in Section 7.

3. Annealed Risk, ESI, and Complexity

In this section we present Lemma 5, a PAC-Bayesian style bound that underlies all our results to follow. Remarkably, it holds without any regularity conditions. However, on the left hand side it has an "annealed" version of the risk rather than the actual risk. In Sections 4, 5, and 6 we give conditions under which the annealed risk can be replaced by either a Hellinger-type distance or the standard risk, which is what we are really interested in. Lemma 5 relates the annealed risk to an information complexity via *exponential stochastic inequality* (ESI). We now introduce the technical notions of annealed expectation and ESI. We then present Lemma 5 and discuss its right-hand side, the information complexity. We do not claim any novelty for the technical results in this section — the lemma below can be found in (Zhang, 2006b,a), for example. Still, we need to treat these results in some detail to prepare the new results in subsequent sections.

3.1. Main Concepts: Annealed and Hellinger Risk, ESI

For $\eta > 0$ and general random variables U, we define, respectively, the *Hellinger-transformed expectation* and the *annealed expectation* (terminology from statistical mechanics; see e.g. (Haussler et al., 1996)), also known as *Rényi-transformed expectation* (terminology from information theory, see e.g. (Van Erven and Harremoës, 2014)) as

$$\mathbf{E}^{\mathrm{HE}(\eta)}\left[U\right] := \frac{1}{\eta}\left(1 - \mathbf{E}\left[e^{-\eta U}\right]\right) \quad ; \quad \mathbf{E}^{\mathrm{ANN}(\eta)}\left[U\right] := -\frac{1}{\eta}\log \mathbf{E}\left[e^{-\eta U}\right], \tag{11}$$

with log the natural logarithm. We will frequently use that for $\eta > 0$,

$$\mathbf{E}^{\mathrm{HE}(\eta)}\left[U\right] \leq \mathbf{E}^{\mathrm{ANN}(\eta)}\left[U\right] \leq \mathbf{E}[U] \tag{12}$$

where the first inequality follows from $-\log x \geq 1 - x$ and the second from Jensen. We also note that if, for example, U is bounded, then the inequalities become equalities in the limit:

Proposition 1 *If $\mathbf{E}[e^{-\eta X}] < \infty$, we have $\lim_{\eta \downarrow 0} \mathbf{E}^{\mathrm{HE}(\eta)}[X] = \mathbf{E}[X]$ and we also have that $\eta \mapsto \mathbf{E}^{\mathrm{ANN}(\eta)}[X]$ is non-increasing.*

All our results below may be expressed succinctly via the notion of *exponential stochastic inequality.*

Definition 2 (Exponential Stochastic Inequality (ESI)) *Let $\eta > 0$ and let U, U' be random variables on some probability space with probability measure P. We define*

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

In all our applications of this notation, P is the distribution appearing in a given learning problem $(P, \ell, \mathcal{F})$ that will be clear from the context; hence, we omit it in the ESI notation. An ESI simultaneously captures "with (very) high probability" and "in expectation" results.

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

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

The following proposition will be extremely convenient for our proofs:

Proposition 4 (Weak Transitivity) *Let (U, V) be a pair of random variables with joint distribution P. For all $\eta > 0$ and $a, b \in \mathbb{R}$, if $U \trianglelefteq_\eta a$ and $V \trianglelefteq_\eta b$, then $U + V \trianglelefteq_{\eta/2} a + b$.*

Proof From Jensen's inequality: $\mathbf{E}[e^{\frac{\eta}{2}((U-a)+(V-b))}] \leq \frac{1}{2}\mathbf{E}[e^{\eta(U-a)}] + \frac{1}{2}\mathbf{E}[e^{\eta(V-b)}]$. ■

3.2. PAC-Bayesian Style Inequality

All our results are based on the following lemma due to Zhang (2006b):

Lemma 5 *Let $(P, \ell, \mathcal{F})$ represent a learning problem with L_f the excess loss relative to an optimal f^*. Let $\Pi_|$ be a learning algorithm (defining a "prior" Π_0) for this learning problem that outputs distributions on $\mathcal{F}$. For all $\eta > 0$, $n \in \mathbb{N}$, we have:*

$$\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{E}^{\mathrm{ANN}(\eta)}_{Z\sim P}\left[L_{\underline{f}}\right]\right] \trianglelefteq_{\eta \cdot n} \mathrm{IC}_{n,\eta}\left(\Pi_|\right). \tag{14}$$

where $\mathrm{IC}_{n,\eta}$ is the information complexity, defined as:

$$\mathrm{IC}_{n,\eta}(\Pi_|) := \mathbf{E}_{\underline{f} \sim \Pi_n}\left[\frac{1}{n}\sum_{i=1}^{n} L_{\underline{f}}(Z_i)\right] + \frac{\mathrm{KL}(\Pi_n \,\|\, \Pi_0)}{\eta \cdot n}. \tag{15}$$

By the finiteness considerations of Appendix H, $\mathrm{IC}_{n,\eta}(\Pi_|)$ is always well-defined but may in some cases be equal to $-\infty$ or ∞. We prove a generalized form of this result, which does not require existence of an optimal f^*, in Appendix A.1 The proof is essentially taken from the proof of Theorem 2.1 of Zhang (2006b) and is presented only for completeness.

This result is similar to various results that have been called *PAC-Bayesian inequalities*, although this name is sometimes reserved for a different type of inequality involving an empirical (observable) quantity on the right that does not involve f^* (McAllester, 2003). Lemma 5 generalizes earlier in-expectation results by Barron and Li (1999) for deterministic estimators rather than (randomized) learning algorithms; these in-expectation results further refine in-probability results of Barron and Cover (1991), arguably the starting point of this research.

To explain the potential usefulness of Lemma 5, let us weaken (14) to an in-expectation statement via Proposition 3, so that it reduces to:

$$\mathbf{E}_{Z^n \sim P}\left[\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{E}^{\mathrm{ANN}(\eta)}[L_{\underline{f}}]\right]\right] \leq \mathbf{E}_{Z^n \sim P}\left[\mathrm{IC}_{n,\eta}\left(\Pi_|\right)\right]. \tag{16}$$

If the annealed expectation were a standard expectation, the left-hand side would be an expected excess risk. Then we would have a great theorem: by (16), the lemma bounds the expected excess risk of estimator $\Pi_|$ by a complexity term, which, as we will see below, generalizes a large number of previous complexity terms (and allows us to get the same rates), both for well-specified density estimation and for general loss functions. The non-standard inequality $\trianglelefteq$ implies that we get such bounds not only in expectation but also in probability. The only problem is that the left-hand side in Lemma 5 is not the standard risk but the annealed risk, which is always smaller and can even be negative. It turns out however that — as already suggested, but not proved by Proposition 1 — by making η small enough, the left-hand side can in many cases be related to the standard excess risk or another divergence-like measure after all. The conditions which allow this are the subject of Sections 4–6; but first, in the remainder of the this section we study the complexity term in detail.

3.3. Information Complexity

The present form of the information complexity is due to Zhang (2006b), with precursors from Rissanen (1989); Barron and Cover (1991); Yamanishi (1998). For generalized Bayesian, two-part MDL and standard ERM, a first further bound is given via the following proposition, the first part of which is also from Zhang (2006b); we note that this result can be extended to the generalized definition of $\mathrm{IC}_{n,\eta}$ given in Section A.1; the extended result does not rely on the existence of f^*.

Proposition 6 *Consider a learning problem $(P, \ell, \mathcal{F})$ and let $Z^n \equiv Z_1, \ldots, Z_n$ be any sample with $\sum_{i=1}^n \ell_{f^*}(Z_i) < \infty$ (this will hold a.s. if $Z^n \sim P$). Let Π_0 be a distribution on $\mathcal{F}$. and let $\Pi^B_{|}$ be the corresponding η-generalized Bayesian posterior, with, for each n, π^B_n given by* (2)*. We have for all $\eta > 0$ that $\mathrm{IC}_{n,\eta}(\Pi^B_{|})$ is non-increasing in η, and that*

$$n \cdot \mathrm{IC}_{n,\eta}(\Pi^B_{|}) = n \cdot \inf_{\Pi_| \in \mathrm{RAND}} \mathrm{IC}_{n,\eta}(\Pi_|) = -\frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim \Pi_0} \exp\left(-\eta \sum_{i=1}^n L_{\underline{f}}(Z_i) \right) \tag{17}$$

$$\leq \inf_A \left\{ -\frac{1}{\eta} \log \Pi_0(A) + n \cdot \mathrm{IC}_{n,\eta}(\Pi^B_{|} \mid f \in A) \right\} \tag{18}$$

$$\leq \inf_A \left\{ -\frac{1}{\eta} \log \Pi_0(A) + \mathbf{E}_{\underline{f} \sim \Pi_0 | A} \left[\sum_{i=1}^n L_{\underline{f}}(Z_i) \right] \right\}, \tag{19}$$

where RAND *is the set of* all *learning algorithms $\Pi'_|$ that can be defined relative to $(P, \ell, \mathcal{F})$ with $\Pi'_0 = \Pi_0$ and the second infimum is over all measurable subsets $A \subseteq \mathcal{F}$. In the special case that Π_0 has countable support $\ddot{\mathcal{F}}$ so that the η-two part estimator (10) is defined, we further have*

$$n \cdot \mathrm{IC}_{n,\eta}(\Pi^B_{|}) \leq n \cdot \inf_{\dot{f} \in \mathrm{DET}} \mathrm{IC}_{n,\eta}((\dot{f}, \Pi_0)) \tag{20}$$

$$= n \cdot \mathrm{IC}_{n,\eta}(f^* \,\|(\ddot{f}_{2\text{-P}}, \Pi_0)) \leq \inf_{f \in \ddot{\mathcal{F}}} \left\{ -\frac{1}{\eta} \log \Pi_0(\{f\}) + \sum_{i=1}^n L_f(Z_i) \right\},$$

where DET *is the set of* all *deterministic estimators with range $\ddot{\mathcal{F}}$.*

From Lemma 5 and this result, we see that we have three equivalent characterizations of information complexity for η-generalized Bayesian estimators. First, there is just the basic definition (15) with Π_n instantiated to the η-generalized Bayesian posterior. Second, there is the characterization as the minimizer of (15) for the given data, over all distributions Π_n on $\mathcal{F}$. And third, there is the characterization in terms of a generalized Bayesian marginal likelihood: (19) shows that for $\eta = 1$ and ℓ the log loss, the information complexity $\mathrm{IC}_{n,\eta}(\Pi^{\mathrm{B}}_{|})$ is the log Bayes marginal likelihood of the data relative to f^*, divided by n. If furthermore $\mathcal{F}$ is a sufficiently regular k-dimensional parametric probability model equipped with a prior Π_0 with full support on $\mathcal{F}$, and the model is correct, i.e., $Z_1, Z_2, \ldots$ are sampled i.i.d. from a distribution with density in $\mathcal{F}$, then, as is well-known, the information complexity will almost surely coincide, up to $O(1/n)$, with the BIC penalty: $n \cdot \mathrm{IC}_{n,\eta}(\Pi^{\mathrm{B}}_{|}) = (k/2)\log n + O(1)$; see Grünwald (2007) for precise results.

3.3.1. Bounds on Information Complexity for η-Generalized Bayes

Ghosal et al. (2000) (GGV from now on) presented several theorems implying concentration of the (standard) Bayesian posterior around the true distribution in the well-specified i.i.d. case; their results were employed in many subsequent papers such as, for example, (Ghosal and Van Der Vaart, 2007; Ghosal et al., 2008; Bickel and Kleijn, 2012). We compare our results to theirs in Example 2 in Section 4. One of the conditions they impose is the existence of a sequence $(\epsilon_n)_{n \geq 1}$ such that $n\epsilon_n^2 \to \infty$, and, for some constant $C > 0$, for all n, a certain ϵ_n^2-ball around the true distribution has prior mass at least $\exp(-nC\epsilon_n^2)$. Generalizing from log loss to arbitrary loss functions, their condition reads

$$\Pi_0\left(f : \mathbf{E}[L_f] \leq \epsilon_n^2 \ ; \ \mathbf{E}\left(L_f\right)^2 \leq \epsilon_n^2\right) \geq e^{-nC\epsilon_n^2}. \tag{21}$$

They then show that, under this and further conditions, the posterior concentrates with Hellinger rate ϵ_n (see Example 2 of Section 4 for the precise meaning). Now note that (21) implies the weaker

$$\Pi_0\left(f : \mathbf{E}[L_f] \leq \epsilon_n^2\right) \geq e^{-nC\epsilon_n^2}, \tag{22}$$

which in turn implies, via (19), for any $0 < \eta \leq 1$, the following bound on IC for the η-generalized Bayesian estimator:

$$\mathbf{E}_{Z^n \sim P}\left[\mathrm{IC}_{n,\eta}\left(\Pi_{|}\right)\right] \leq \epsilon_n^2 \cdot (1 + (C/\eta)), \tag{23}$$

To see this, note that (19) and (22) imply

$$\begin{aligned} &\mathrm{IC}_{n,\eta}(\Pi^{\mathrm{B}}_{|}) \\ &\leq -\frac{1}{n}\sum_{i=1}^{n} \ell_{f^*}(Z_i) - \frac{1}{n\eta}\log \Pi_0\{f : \mathbf{E}[L_f] \leq \epsilon_n^2\} + \frac{1}{n}\mathbf{E}_{\underline{f} \sim \Pi_0 | \{f : \mathbf{E}[L_f] \leq \epsilon_n^2\}}\left[\sum_{i=1}^{n}\left(\ell_{\underline{f}}(Z_i)\right)\right] \\ &\leq C\frac{\epsilon_n^2}{\eta} + \frac{1}{n}\mathbf{E}_{\underline{f} \sim \Pi_0 | \{f : \mathbf{E}[L_f] \leq \epsilon_n^2\}}\left[\sum_{i=1}^{n}\left(L_{\underline{f}}(Z_i)\right)\right]. \end{aligned} \tag{24}$$

This implies (23).

All the examples of nonparametric families provided by GGV (including priors on sieves, log-spline models and Dirichlet processes) rely on showing that condition (21) above holds for specific priors, and hence in all these cases we get bounds on the expected-information complexity which, by (16) allows us to establish comparable rates in expectation for the

η-generalized Bayesian estimator in the well-specified case, for any η such that the left-hand side can be linked to an actual distance measure — see Example 2 in Section 4.

We also would like to bound the excess risk in probability in terms of the expected information complexity. For this, we can proceed in either of two ways: we either start with an expectation bound such as (16) and then use Markov's inequality (since the excess risk of any estimator is a.s. nonnegative) to go back from expectation to in-probability. However, under GGV's condition (21) (the weaker (22) is not sufficient here), we can also use the in-probability version of Lemma 5 directly. In combination with Lemma 8.1 of GGV (which straightforwardly extends to our setting with general loss and η) this implies that for all $\delta > 0$:

$$P\left(\mathrm{IC}_{n,\eta}\left(\Pi_{|}\right) \geq \left(1 + \delta^{-1/2}\right)\epsilon_n^2\right) \leq \frac{\delta}{n\epsilon_n^2}. \tag{25}$$

It follows that under (21), since $n\epsilon_n^2 \to \infty$, ϵ_n^2 is, up to constant factors depending on δ, an upper bound both on $\mathbf{E}\left[\mathrm{IC}_{n,\eta}\left(\Pi_{|}\right)\right]$, and, for every δ, with probability at least $1-\delta$, on $\mathrm{IC}_{n,\eta}\left(\Pi_{|}\right)$ — see the discussion below Theorem 31 in Section 6.

Finally, there often exist nontrivial worst-case (sup norm) or almost-sure bounds on the information complexity; such bounds — mostly developed for parametric models but also, e.g., for Gaussian processes (Seeger et al., 2008) have historically mostly been established within the MDL literature; see (Grünwald, 2007) for an extensive overview. While we will not go into such bounds in detail here, below we provide a very simple such bound for countably infinite classes, which shows the ease by which IC allows for model aggregation.

Suppose that we have a countably infinite collection of classes $\mathcal{F}_1, \mathcal{F}_2, \ldots$ and a corresponding set of priors $\Pi_0^{(1)}, \Pi_0^{(2)}, \ldots$. Let us select a new prior $q : \mathbb{N} \to \mathbb{R}^+$ over the collection $\mathcal{F} := \bigcup_{j\in\mathbb{N}} \mathcal{F}_j$. Then we may define a new prior $\Pi_0 = \sum_{j\in\mathbb{N}} q(j)\Pi_0^{(j)}$ over $\mathcal{F}$. We will assume that the risk minimizer in the full class, f^*, is equal to $f^*_{j^*}$ for some $j^* \in \mathbb{N}$. By Proposition 6, Eq. (18), we must now have, for all data $Z_1, \ldots, Z_n$, that

$$n \cdot \mathrm{IC}_{n,\eta}\left(\Pi_{|}\right) \leq -\frac{1}{\eta}\log q(j^*) + n \cdot \mathrm{IC}_{n,\eta}(\Pi \mid f \in \mathcal{F}_{j^*}), \tag{26}$$

where $\Pi \mid f \in \mathcal{F}_{j^*}$ is the η-generalized Bayesian estimator based on the prior $\Pi_0^{(j^*)}$ within $\mathcal{F}_{j^*}$.

If we now further assume that, for each j, the GGV-type condition (22) is satisfied (with prior $\Pi_0^{(j)}$ and with f^* replaced by f^*_j, the risk minimizer over $\mathcal{F}_j$), then taking expectations in (26) implies that (22) holds for Π_0, with $f^* = f^*_{j^*}$, with the RHS scaled by a factor $q(j^*)$. A simple adaptation of (23) then gives

$$\mathbf{E}_{Z^n\sim P}\left[\mathrm{IC}_{n,\eta}\left(\Pi_{|}\right)\right] \leq \epsilon_n^2 \cdot (1 + (C/\eta)) + \frac{-\log q(j^*)}{n\eta}. \tag{27}$$

Thus, the overhead in information complexity for combining the classes is simply $\frac{-\log q(j^*)}{n\eta}$. Moreover, in the case of a finite collection of M classes, we may take q uniform and the overhead becomes $\frac{\log M}{n\eta}$.

4. The Strong Central Condition

As we explained below Lemma 5, our strategy in proving our theorems will be to determine conditions under which the η-annealed excess risk is similar enough to either the standard risk or a meaningful weakening thereof for Lemma 5 to be useful. In this section we present the simplest such condition, which is still quite strong — it requires an exponentially small upper tail of the distribution of $\ell_{f^*} - \ell_f$. This *strong central* condition has a parameter $\bar{\eta} > 0$, and whenever we want to make this explicit we refer to it as "the $\bar{\eta}$-central condition". *Intuitively*, its usefulness for learning is obvious: it ensures that the probability that a "bad" f outperforms f^* by more than L is exponentially small in L. *Technically*, its use is that it ensures that the annealed risk is positive for all $\eta < \bar{\eta}$. This allows us to turn Lemma 5 into a useful result by replacing its left-hand side by a metric which (for log loss) generalizes the squared Hellinger distance.

4.1. Definitions and Main Results

We now turn to the strong central condition, which, along with its weakened versions discussed in Section 6 was introduced by Van Erven et al. (2015).

Definition 7 (Central Condition) *Let $\bar{\eta} > 0$. We say that $(P, \ell, \mathcal{F})$ satisfies the* strong $\bar{\eta}$-central condition *if there exists some $\tilde{f} \in \mathcal{F}$ such that*

$$\mathbf{E}\left[e^{-\bar{\eta}(\ell_f - \ell_{\tilde{f}})}\right] \leq 1, \text{ i.e., } \ell_{\tilde{f}} - \ell_f \trianglelefteq_{\bar{\eta}} 0 \qquad \text{for all } f \in \mathcal{F}. \tag{28}$$

Jensen's inequality implies that if a $\tilde{f}$ exists satisfying (28), it must be optimal; hence we can take $\tilde{f} = f^*$. The special case of this condition with $\bar{\eta} = 1$ under log loss has appeared previously, often implicitly, in works studying rates of convergence in density estimation (Barron and Cover, 1991; Li, 1999; Zhang, 2006a; Kleijn and van der Vaart, 2006; Grünwald, 2011). For details about the myriad of implications of the central condition and its equivalences to other conditions we refer to Van Erven et al. (2015). Here we merely highlight the most important facts. First, trivially, the strong central condition automatically holds for density estimation with log loss in the well-specified setting since then p_{f^*} is the density of P (see Example 1), as we then have

$$\mathbf{E}_{Z\sim P}\left[e^{-\bar{\eta}(\ell_f - \ell_{f^*})}\right] = \mathbf{E}_{Z\sim P}\left[\frac{p_f(Z)}{p_{f^*}(Z)}\right] = 1 \tag{29}$$

Second, less trivially, it also automatically holds under a convex model in the misspecified setting (see Li (1999) and Example 2.2 of Van Erven et al. (2015)). Third, for classification and other bounded excess loss cases, it can be related to the *Massart condition*, a special case of the Bernstein condition (Audibert, 2004; Bartlett and Mendelson, 2006) (as discussed immediately before Definition 17 in Section 5).

We now introduce a new metric which is derived from the Hellinger metric, introduced below (as is common) in terms of its square.

Definition 8 (Misspecification Metric) *For a given learning problem $(P, \ell, \mathcal{F})$, associate each $f \in \mathcal{F}$ and $\eta > 0$ with a probability density*

$$p_{f,\eta}(z) := p(z)\frac{\exp(-\eta L_f(z))}{\mathbf{E}[\exp(-\eta L_f(Z))]}, \tag{30}$$

where p is the density of P. Now define $d_{\bar{\eta}}(f, f')$ as the Hellinger distance between $p_{f,\bar{\eta}}$ and $p_{f',\bar{\eta}}$:

$$\begin{aligned} d^2_{\bar{\eta}}(f, f') &:= \frac{2}{\bar{\eta}}\left(1 - \int \sqrt{p_{f,\bar{\eta}}(z) p_{f',\bar{\eta}}(z)}\, d\mu(z)\right) \\ &= \mathbf{E}^{\mathrm{HE}(\bar{\eta}/2)}\left[L_f - \mathbf{E}^{\mathrm{ANN}(\bar{\eta})}\left[L_f\right] + L_{f'} - \mathbf{E}^{\mathrm{ANN}(\bar{\eta})}\left[L_{f'}\right]\right]. \end{aligned} \tag{31}$$

The following result is obvious:

Proposition 9 *If ℓ is log loss and $\mathcal{F}$ is well-specified relative to P we can take $\bar{\eta} = 1$ and then for every $f \in \mathcal{F}$, $d^2_{\bar{\eta}}(f^*, f)$ coincides with the standard squared Hellinger distance $\mathrm{H}_{1/2}(P_{f^*} \| P_f)$ defined by $\mathrm{H}_{1/2}(P_f \| P_{f'}) := 2\left(1 - \int \sqrt{p_f(z) p_{f'}(z)}\, d\mu(z)\right)$.*

Since $d_{\bar{\eta}}$ is always interpretable as a Hellinger distance, it is clearly a metric. This is different from an existing, more well-known generalization of the Hellinger distance for the well-specified case (Sason and Verdú, 2016), $\mathrm{H}_\eta(P \,\|\, Q) := \eta^{-1}\left(1 - \mathbf{E}_{Z\sim P}\left(q(z)/p(z)\right)^\eta\right)$ which does not define a metric except for $\eta = 1/2$ (and then coincides with d_1). The $d_{\bar{\eta}}$ metric is of interest in the misspecified density estimation setting — with density estimation, we may not necessarily be interested in log loss prediction and a metric weaker than excess risk (i.e. generalized KL divergence) may be sufficient for our purposes. With other loss functions, the main interest will usually be learning an $\hat{f}$ with small prediction error. Then the metric above, while still well-defined, may not be appropriate, and one is interested in the excess risk bounds of the next section instead.

Theorem 10 *Suppose that the $\bar{\eta}$-strong central condition holds. Then for any $0 < \eta < \bar{\eta}$, the metric $d_{\bar{\eta}}$ satisfies*

$$\mathbf{E}_{\underline{f}\sim\Pi_n}\left[d^2_{\bar{\eta}}(f^*, \underline{f})\right] \trianglelefteq_{\eta\cdot n} C_\eta \cdot \mathrm{IC}_{n,\eta}\left(\Pi_|\right),$$

with $C_\eta = \eta/(\bar{\eta} - \eta)$. In particular, $C_\eta < \infty$ for $0 < \eta < \bar{\eta}$, and $C_\eta = 1$ for $\eta = \bar{\eta}/2$.

Example 2 (Comparison to Results by GGV) Following (Zhang, 2006a) we illustrate the considerable leverage provided in the well-specified density estimation case by allowing η-generalized Bayesian estimators for $\eta < 1$. GGV show that for the standard Bayesian estimator, under condition (21) (which only refers to *local* properties of the prior in neighborhoods of the true density p_{f^*}), in combination with a rather stringent *global* entropy condition, the following holds: there exists a constant C' such that $\Pi_n\left(f \in \mathcal{F} : d^2_1(f^*, f) > C'\epsilon^2_n\right) \to 0$ in P-probability, i.e., for every $B > 0$,

$$P\left(\Pi_n\left(f \in \mathcal{F} : d^2_1(f^*, f) > C'\epsilon^2_n\right) > B\right) \to 0.$$

Now, suppose the model is correct so that the $\bar{\eta}$-central condition holds for $\bar{\eta} = 1$. Then we get from Theorem 10 that for any $\eta < \bar{\eta}$, using *only* condition (22), the following holds: for any $\gamma_1, \gamma_2, \ldots$ such that $\gamma_n/\epsilon_n \to \infty$, the generalized Bayesian estimator satisfies $\Pi_n\left(f \in \mathcal{F} : d^2_1(f^*, f) > C'\gamma^2_n\right) \to 0$ in P-probability, i.e., for every $B > 0$,

$$P\left(\Pi_n\left(f \in \mathcal{F} : d^2_1(f^*, f) > C'\gamma^2_n\right) > B\right) \to 0, \tag{32}$$

as immediately follows from applying Markov's inequality twice as done below. Thus, by taking $\eta < 1$ we need neither the stronger condition (21) nor the much stronger GGV global entropy condition; for this we pay only a slight price since our bound is not in terms of ϵ_n^2 but is instead in terms of γ_n^2, which we have to take slightly larger (a factor $\log \log n$ is of course sufficient). Under well-specification, we thus obtain the same rates as GGV for all the statistical models they consider, up to a $\log \log n$ factor; as GGV show, these rates are usually minimax optimal. Interestingly, other works on Bayesian and MDL nonparametric consistency for the well-specified case also consider $\eta < 1$ (Barron and Cover, 1991; Zhang, 2006a; Walker and Hjort, 2002; Martin et al., 2017) or invoke an alternative stringent condition to deal with $\eta = 1$ ((Zhang, 2006a, Section 5.2), Barron et al. (1999)); see Zhang (2006a) for a very detailed discussion. While it may be argued that one should be able to deal with standard Bayes ($\eta = 1$), in this paper we also aim to deal with misspecification where we need to take $\eta < 1$ (and cannot take it arbitrarily close to 1) even for simple problems (Grünwald and Van Ommen, 2017), and then there is no special reason to handle $\eta = 1$ via additional conditions.

To show (32), note that, if the $\bar{\eta}$-central condition holds, then for general $A, B > 0$, we have

$$
\begin{aligned}
&P\left(\Pi_n(f \in \mathcal{F} : d_{\bar{\eta}}^2(f^*, f) > A) > B\right) \leq B^{-1}\, \mathbf{E}_{Z^n}\left[\Pi_n(f \in \mathcal{F} : d_{\bar{\eta}}^2(f^*, f) > A)\right] \\
&\quad \leq (AB)^{-1}\, \mathbf{E}_{Z^n}\, \mathbf{E}_{\underline{f} \sim \Pi_n}\left[d_{\bar{\eta}}^2(f^*, \underline{f})\right] \leq (AB)^{-1}\, \mathbf{E}_{Z^n}\left[\mathrm{IC}_{n,\bar{\eta}/2}\left(\Pi_|\right)\right],
\end{aligned}
$$

where we applied Markov's inequality twice, and the final inequality is from Theorem 10. Plugging in $A = C'\gamma_n^2$ and $\epsilon_n^2 \geq \mathbf{E}\left[\mathrm{IC}_{n,\bar{\eta}/2}\left(\Pi_|\right)\right]$ (using (23)), this can be further bounded as $B^{-1}\epsilon_n^2/\gamma_n^2 \to 0$. □

4.2. Applying Theorem 10 in Misspecified Density Estimation

From the above it is clear that Theorem 10 has plenty of applications whenever the model under consideration is correct. We now consider applications of Theorem 10 to misspecified models of probability densities $\mathcal{F}$ with generalized Bayesian estimators $\Pi_|^B$. For this we must establish (a) that the central condition holds for $\mathcal{F}$, and (b) suitable bounds on the information complexity relative to $\Pi_|^B$. As to (a), we know that the $\bar{\eta}$-central condition holds for $\bar{\eta} = 1$ whenever the set of distributions $\{p_f : f \in \mathcal{F}\}$ is correct or convex; as shown elsewhere and illustrated in Example 3 below, it also holds for 1-dimensional (nonconvex) exponential families and high-dimensional generalized linear models (GLMs) under potentially severe misspecification of the noise, as long as the regression function is well-specified and P has exponentially small tails. As to (b), we may consider priors such that in the well-specified case, the GGV condition holds for some sequence $\epsilon_1^2, \epsilon_2^2, \ldots$ as in Example 2. As explained in the example, the GGV condition then automatically holds for GLMs under misspecification as well, so that the same bounds on information complexity can be given as in the well-specified case. It appears that this is a special property of GLMs though — for general $\mathcal{F}$, we only have the following proposition which shows that, if the GGV condition holds for some specific prior in the well-specified case with some bounds $\epsilon_1, \epsilon_2, \ldots$, then, as long as p_{f^*} dominates p, it must still hold in the misspecified case for the same prior for a strictly larger sequence $\epsilon_1', \epsilon_2', \ldots$, leading to a potential deterioration of the bound given by Theorem 10.

Proposition 11 *Consider a learning problem $(P, \ell, \mathcal{F})$ where $\mathcal{F}$ indexes a set of probability distributions $\{P_f : f \in \mathcal{F}\}$ with densities p_f, and suppose that* $\sup_{z \in \mathcal{Z}} \frac{dP(z)}{dP_{f^*}(z)} = C < \infty$. *Then for all $f \in \mathcal{F}$,*

$$\mathbf{E}_{Z \sim P}[L_f] \leq C \cdot \left(\mathbf{E}_{Z \sim P_{f^*}}[L_f] + \sqrt{2\, \mathbf{E}_{Z \sim P_{f^*}}[L_f]} \right). \tag{33}$$

Proof Observe that

$$\mathbf{E}_{Z \sim P}[L_f] \leq \mathbf{E}_{Z \sim P}[0 \vee L_f] \leq C\, \mathbf{E}_{Z \sim P_{f^*}}[0 \vee L_f] \leq C \cdot \left(D(f^* \| f) + \sqrt{2 D(f^* \| f)} \right),$$

where $\mathbf{E}_{Z \sim P_{f^*}}[L_f] = D(f^* \| f)$ is the KL divergence between f^* and f and the last inequality is from Yang and Barron (1998) (see the remark under their Lemma 3); for completeness we provide a proof in the appendix. ■

As a trivial consequence, whenever the weakened GGV condition (22) holds for all P_f with $f \in \mathcal{F}$ for a sequence $\epsilon_1, \epsilon_2, \ldots$, it will still hold for a sequence $\epsilon'_1, \epsilon'_2, \ldots$ with $\epsilon'_j \asymp \sqrt{\epsilon_j}$. It follows from (23) that we now automatically have a bound of order ϵ'_n / n on the misspecified expected information complexity. Theorem 10 now establishes that whenever the GGV condition holds in the well-specified case, under the further (weak) condition that $\sup_{z \in \mathcal{Z}} dP(z)/dP_{f^*}(z) = C < \infty$, we automatically get a form of consistency for η-generalized Bayes, for $\eta < \bar{\eta}$. The question whether we get the same rates of convergence is obfuscated in two ways: first, the misspecification metric is in general incomparable to the Hellinger metric; second, even in cases in which the misspecification metric dominates the standard Hellinger, for nonparametric $\mathcal{F}$ with $\mathbf{E}[\mathrm{IC}_{n,\eta}] \asymp n^{-\gamma}$, the conversion $\epsilon'_j \asymp \sqrt{\epsilon_j}$ worsens the rates obtained by Theorem 10 to $n^{-\gamma/2}$. To deal with the first problem, one could establish a condition under which the misspecification metric dominates standard Hellinger; but this is tricky and will be left for future work. The second problem is still of interest in the next section, in which the misspecification metric is replaced by the excess risk, which has the same meaning irrespective of whether $\mathcal{F}$ is well-specified. As indicated below, for generalized linear models we can get rid of the square root in (33), but whether this can be done more generally also remains an important open problem for future work. An alternative, also to be considered for future work, is to refrain from using the priors constructed for the well-specified case altogether and instead directly design priors for the misspecified case, with hopefully better bounds on information complexity.

Example 3 (Exponential Families and Generalized Linear Models) Consider a learning problem $(P, \ell, \mathcal{F})$ in the conditional density estimation setting of Example 1, so that ℓ is the conditional log-loss; $Z = (X, Y)$ with X taking values in $\mathcal{X} \subset \mathbb{R}^k$; and $\{p_f : f \in \mathcal{F}\}$ for some $\mathcal{F} \subset \mathbb{R}^k$ represents a k-dimensional generalized linear model (GLM), given in its standard parameterization (so that $\langle x, f \rangle$ is the linear predictor fed into the link function) (McCullagh and Nelder, 1989). Heide et al. (2019, Theorem 2) show[2] that, under three further conditions on $(P, \ell, \mathcal{F})$, the central condition holds for some $\bar{\eta} > 0$, even under misspecification. In essence, the conditions require (1) that Y has exponential tails, in

[2] In previous arXiv versions of this paper, we gave these results in full detail, adding 7 pages to its length. Following referee's comments and consultation with the associate editor, we moved them to the paper (Heide et al., 2019), where they are further illustrated by means of actual experiments with misspecified GLMs.

the sense that $\sup_{x\in\mathcal{X}} \mathbf{E}[\exp(\eta|Y|) \mid X = x] < \infty$ for some $\eta > 0$ (a requirement that is automatically satisfied for, e.g., logistic regression, for which $\mathcal{Y}$ is finite); (b) that $\mathcal{F}$ is restricted to a compact (though possibly very high dimensional) set, and (c), that the misspecification is of a certain type: the noise may be misspecified in arbitrary ways, but the GLM should contain the distribution with the correct generalized regression function. That is, there should be an $f \in \mathcal{F}$ indexing distribution P_f with the correct conditional mean, so that $\mathbf{E}_{P_f}[Y \mid X] = \mathbf{E}_P[Y \mid X]$. This f will then in fact be equal to the risk-optimal f^*. By taking $\mathcal{X}$ to be a singleton, a GLM becomes a 1-dimensional natural exponential family, and the result thus also applies to such families. For this simplified case, Heide et al. (2019) show that the smallest $\bar{\eta}$ for which the $\bar{\eta}$-central condition holds is upper bounded by, and in some cases not much smaller than, the ratio of variances $\mathbf{E}_{P_{f^*}}[(Y - \mathbf{E}_{P_{f^*}}[Y])^2]/\mathbf{E}_P[(Y - \mathbf{E}_P[Y])^2]$.

Heide et al. (2019, Proposition 2) shows that, if $\mathcal{F}$ represents a GLM, then under the same three conditions, we have $\mathbf{E}_{Z\sim P}[L_f] = \mathbf{E}_{Z\sim P_{f^*}}[L_f]$, so that there is no need to resort to Proposition 11. This implies that for any prior satisfying the GGV condition in the well-specified case, the same prior can be used in the misspecified case and, using Theorem 10, we can prove the same risk bounds, up to a constant factor, as in the well-specified case for generalized Bayes with any fixed $\eta < \bar{\eta}$. In particular, k-dimensional GLMs being sufficently general parametric models, we can use any continuous prior on $\mathcal{F}$ that is bounded away from 0 and obtain that, for any fixed η, $n \cdot \mathrm{IC}_{n,\eta}(\Pi_{|}^{\mathrm{B}}) \le (k/2\eta)\log n + O(1)$, cf. the remark after Proposition 6. Theorem 10 then gives a bound of $\tilde{O}(k/n)$, which is within a log factor of the minimax optimal parametric rate $O(k/n)$ for squared Hellinger distance in the well-specified case. □

Example 4 (Comparison to Bhattacharya et al. (2019)) After our submission of the present paper, we became aware of (Bhattacharya et al., 2019). The analysis and results of that paper (first submitted to arXiv in 2016, around the same time as the present paper) overlap with our Theorem 10, and some of their examples have implications for our work as well. Bhattacharya et al. (2019) focus exclusively on generalized Bayesian estimators $\Pi_{|}^{B}$. Their Theorem 3.6 is a variation of Zhang's Lemma 5, extended to handle non-i.i.d. P. Their α-Rényi divergence is just our η-annealed excess risk, with $\eta = 1 - \alpha$. For $\mathcal{F}$ satisfying the $\bar{\eta}$-central condition, they provide Theorem 3.1, which has some similarity to Theorem 10: their result extends ours in that it allows non-i.i.d. P; it rephrases ours so that the result is directly stated in terms of GGV-style conditions on Π_0 rather than on bounds on $\mathrm{IC}_{n,\eta}(\Pi_{|}^{B})$, similar to our (32); and it stays closer to Lemma 5 in that it keeps the annealed excess risk on the left (a nonsymmetric divergence) where Theorem 10 has a (symmetric) metric. In their Lemma 2.1. they re-prove the result of Li (1999) and Van Erven et al. (2015) that 1-strong central holds for convex probability models. Also, they provide (Section 5.1) an interesting novel example in which the strong 1-central condition holds: Gaussian regression, with probability densities $p_f(y \mid x) \propto \exp(-(y - f(x))^2/2\sigma^2)$ with fixed variance σ^2, where the true noise is Gaussian and the set of regression functions $\mathcal{F}$ is convex (but the corresponding density functions $\{p_f : f \in \mathcal{F}\}$ are not, so Li's result does not apply). The model is misspecified in that $\mathcal{F}$ does not contain the true regression function; in contrast, in Example 3 above we considered the reverse case in which the noise is misspecified yet the regression function is not. They show that in their setting, bounds on the annealed excess risk imply bounds on the $L_2(P)$-parameter estimation error that we consider in Example 9. They do not consider the non-annealed excess risk bounds and weaker forms of the central condition that we will turn to in the following sections. □

5. The Witness Condition

We have seen via Theorem 10 that under the $\bar{\eta}$-central condition, Lemma 5 provides a bound on a weak Hellinger-type metric. For problems different from density estimation, i.e., loss functions different from log loss, we often mainly are interested in a bound on the excess risk. To get such bounds, we need a second condition on top of the $\bar{\eta}$-central condition. To see why, consider again the density estimation example (Example 1). If we assume a correct model, $p = p_{f^*}$, then from (29) the $\bar{\eta}$-central condition holds automatically for all $\bar{\eta} \leq 1$, and so Theorem 10 gives a bound on the Hellinger distance. Yet, while the Hellinger distance is bounded, in general we can have $\mathrm{KL}(p \,\|\, p_f) = \infty$. If, for example, $\mathcal{F}$ is the set of densities for the Bernoulli model, P is Bernoulli(1/2), and we use ERM for log loss (so that $\hat{f}$ is the maximum likelihood estimator for the Bernoulli model), we observe with positive probability only 0's. In this case, we will infer $\hat{f}$ with $p_{\hat{f}}(Y = 0) = 1$, and thus with positive probability the excess risk between $\hat{f}$ and f^* is ∞ even though the expected Hellinger distance is of order $O(1/n)$. We thus need an extra condition.

For log loss, the simplest such condition is that the likelihood ratio of p_{f^*} to p_f is uniformly bounded for all $f \in \mathcal{F}$. For that case, Birgé and Massart (1998) proved a tight bound on the ratio between the standard KL divergence and the standard ($\eta = 1/2$) Hellinger distance. Lemma 13 below represents a generalization of their result to arbitrary η, misspecified $\mathcal{F}$, and general loss functions under the *witness condition* which we introduce below, and which is a significant weakening of the bounded likelihood ratio condition. It is the cornerstone for proving our subsequent results: Theorems 14, 22, 29, and 31. Whereas the strong central condition imposes exponential decay of the lower tail of the excess loss $\ell_f - \ell_{f^*}$, the witness condition imposes a much weaker type of control on the upper tail of $\ell_f - \ell_{f^*}$.

Below, we show that the witness condition generalizes not only conditions of Birgé and Massart (1998) but also of Sason and Verdú (2016) and Wong and Shen (1995) (Example 6). We also show that it holds in a variety of settings, e.g., with exponential families with suitably restricted parameter spaces in the well-specified setting and when the log likelihood has exponentially small tails (Example 5), but also with bounded regression under heavy-tailed distributions (Example 7). Moreover, although the conditions are not equivalent, there is an intriguing similarity to the recent *small-ball assumption* of Mendelson (2014) (Example 9).

5.1. Definition and Main Result

Definition 12 (Empirical Witness of Badness) *We say $(P, \ell, \mathcal{F})$ satisfies the (u, c)-*empirical witness of badness condition *(or* witness condition*) for constants $u > 0$ and $c \in (0, 1]$ if for all $f \in \mathcal{F}$*

$$\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u\}}\right] \geq c\, \mathbf{E}[\ell_f - \ell_{f^*}]. \tag{34}$$

More generally, for a function $\tau : \mathbb{R}^+ \to [1, \infty)$ *and constant* $c \in (0,1)$ *we say* $(P, \ell, \mathcal{F})$ *satisfies the* (τ, c)-witness condition *if for all* $f \in \mathcal{F}$, $\mathbf{E}[\ell_f - \ell_{f^*}] < \infty$ *and*

$$\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \le \tau(\mathbf{E}[\ell_f - \ell_{f^*}])\}}\right] \ge c\, \mathbf{E}[\ell_f - \ell_{f^*}]. \tag{35}$$

The (u, c)-witness condition (34) is just the (τ, c)-witness condition for the constant function τ identically equal to u. In our results we frequently use the fact that, by adding $\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} > u\}}\right]$ to both sides of (34) and rearranging, the (u, c)-witness condition holds if and only if for $c' = 1 - c$ (and hence $c' \in (0,1)$),

$$\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} > u\}}\right] \le c'\, \mathbf{E}[\ell_f - \ell_{f^*}], \tag{36}$$

and similarly for the τ-version.

The intuitive reason for imposing this condition is to rule out situations in which learnability simply cannot hold. For instance, consider a setting with $\mathcal{F} = \{f^*, f_1, f_2, \ldots\}$ where $\ell_{f^*} = 1$ with probability 1 and, for each $j \ge 1$, ℓ_{f_j} is equal to 0 with probability $1 - \frac{1}{j}$ and equal to $2j$ with probability $\frac{1}{j}$. Then for all j, $\mathbf{E}[\ell_{f_j} - \ell_{f^*}] = 1$, but as $j \to \infty$, empirically we will never *witness the badness of* f_j as it almost surely achieves lower loss than f^*. On the other hand, if the excess loss is upper bounded by some constant b, we may always take $u = b$ and $c = 1$ so that a witness condition is trivially satisfied. Below we provide several nontrivial examples besides bounded excess losses and finite $\mathcal{F}$ in which the witness condition holds.

The following result shows how the witness condition, combined with the strong central condition, leads to fast-rate excess risk bounds:

Lemma 13 *Let* $\bar{\eta} > 0$. *Assume that the* $\bar{\eta}$*-strong central condition (28) holds and let, for arbitrary* $0 < \eta < \bar{\eta}$, $c_u := \frac{1}{c} \frac{\eta u + 1}{1 - \frac{\eta}{\bar{\eta}}}$. *Suppose further that the* (u, c)*-witness condition holds for* $u > 0$ *and* $c \in (0, 1]$. *Then for all* $f \in \mathcal{F}$, *all* $\eta \in (0, \bar{\eta})$:

$$\mathbf{E}[L_f] \;\le\; c_u \cdot \mathbf{E}^{\mathrm{HE}(\eta)}\left[L_f\right] \;\le\; c_u \cdot \mathbf{E}^{\mathrm{ANN}(\eta)}\left[L_f\right]. \tag{37}$$

More generally, suppose that the $\bar{\eta}$*-central condition and the* (τ, c)*-witness condition hold for* $c \in (0, 1]$ *and a* non-increasing *function* τ. *Then for all* $\lambda > 0$, *all* $f \in \mathcal{F}$,

$$\mathbf{E}[L_f] \;\le\; \lambda \vee \left(c_{\tau(\lambda)} \cdot \mathbf{E}^{\mathrm{HE}(\eta)}\left[L_f\right]\right) \;\le\; \lambda \vee \left(c_{\tau(\lambda)} \cdot \mathbf{E}^{\mathrm{ANN}(\eta)}\left[L_f\right]\right). \tag{38}$$

Note that for large u, c_u is approximately linear in u/c.

The following theorem is now an almost immediate corollary of Lemma 5 and Lemma 13:

Theorem 14 *Consider a learning problem* $(P, \ell, \mathcal{F})$ *and a learning algorithm* $\Pi_|$. *Suppose that the* $\bar{\eta}$*-strong central condition holds. If the* (u, c)*-witness condition holds, then for any* $\eta \in (0, \bar{\eta})$,

$$\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{E}[L_f]\right] \;\trianglelefteq_{\frac{\eta \cdot n}{c_u}}\; c_u \cdot \mathrm{IC}_{n,\eta}\left(\Pi_|\right),$$

with c_u *as in Lemma 13. If instead the* (τ, c)*-witness condition holds for some non-increasing function* τ *as above, then for any* $\lambda > 0$

$$\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{E}[L_f]\right] \;\trianglelefteq_{\frac{\eta \cdot n}{c_{\tau(\lambda)}}}\; \lambda + c_{\tau(\lambda)} \cdot \mathrm{IC}_{n,\eta}\left(\Pi_|\right). \tag{39}$$

Proof The first and second inequalities are from chaining Lemma 5 with Lemma 13 ((37) and (38) respectively). The first inequality is immediate using that for general random variables U, V, we have $U \trianglelefteq_a V \Leftrightarrow cU \trianglelefteq_{a/c} cV$. For the second inequality, we first upper bound the max on the RHS of (38) by the sum of the terms. ■

This theorem is applicable if the (τ, c)-witness condition holds for a non-increasing τ. If the risk $\sup_{f \in \mathcal{F}} \mathbf{E}[L_f]$ is unbounded, we can only expect the witness condition to hold for τ such that for large x, $\tau(x)$ is increasing; such τ are considered in Section 6.3. Non-increasing τ are often appropriate for scenarios with bounded risk (even though the loss may be unbounded and even heavy-tailed); we encounter one instance thereof in the exponential family example below. There, $\lim_{x \downarrow 0} \tau(x) = \infty$, but the increase as $x \downarrow 0$ is so slow that the optimal λ at sample size n is of order $O(1/n)$ and $c_{\tau(\delta)} = O(\log n)$, leading only to an additional log factor in the bound compared to the case where the (u, c)-witness condition holds for constant u.

Some Existing Bounds Generalized by Lemma 13 Lemma 13 generalizes a result of Birgé and Massart (1998, Lemma 5) (also stated and proved in Yang and Barron (1998, Lemma 4)) that bounds the ratio between the standard KL divergence $\mathrm{KL}(P \,\|\, Q)$ and the (standard) 1/2-squared Hellinger distance $\mathrm{H}_{1/2}(P \| Q)$ for distributions P and Q. To see this, take density estimation under log loss in the well-specified setting with $\eta < \bar{\eta} = 1$, so that $f^* = p$ and $f = q$; then the left-hand side becomes $\mathrm{KL}(P \,\|\, Q)$ and the right-hand side $\frac{1}{\eta} \mathbf{E}[1 - e^{-\eta L_f}] = \frac{1}{\eta}(1 - \mathbf{E}[(q/p)^\eta]) = \mathrm{H}_\eta(P \,\|\, Q)$ (this notation was introduced below Proposition 9). Under a bounded density ratio $p/q \leq V$, we can take $u = \log V$ and $c = 1$ (the (u, c)-witness condition is then trivially satisfied), so that $c_u = \frac{\eta \log V + 1}{1 - \eta}$, which for $\eta = 1/2$ coincides with the Birgé-Massart bound. The case of general $\eta \in (0, 1)$ first was handled by Haussler and Opper (1997) (see Lemma 4 therein), but their bound stops short of providing an explicit upper bound for the ratio.

Sason and Verdú (2016) independently obtained an upper bound (see Theorem 9 therein) on the ratio of the standard KL divergence $\mathrm{KL}(P \,\|\, Q)$ to the η-generalized Hellinger divergence in the case of bounded density ratio $\operatorname{ess\,sup} \frac{dP}{dQ}$, for general η. Theorem 13 generalizes Theorem 9 of Sason and Verdú (2016) by allowing for misspecification in the case of density estimation with log loss, allowing for general losses, and, critically for our applications, allowing for unbounded density ratios under a witness condition. We note that in the case of bounded density ratio $\frac{dP}{dQ}$ and the regime $\eta \in (0, 1)$ (corresponding to $\alpha = 1 - \eta \in (0, 1)$ in Theorem 9 of Sason and Verdú (2016)), their bound and the unsimplified form of our bound (see $C_{0 \leftarrow \eta}(V)$ in Lemma 36 in Appendix C) are identical, as they should be since both bounds are tight. The additional, slightly looser simplified bound that we provide greatly helps to simplify the treatment for unbounded excess losses under the witness condition. We stress though that Sason and Verdú (2016) treat general F-divergences under well-specification, including a wide array of divergences beyond η-generalized Hellinger for $\eta \in (0, 1)$, so in that respect, their bounds are far more general. In the next section we establish that Lemma 13 also generalizes a bound by Wong and Shen (1995).

5.2. Example Situations in which the Witness Condition Holds

We now present some examples of common learning problems in which the (τ, c)-witness condition holds for a suitable τ. We first consider a case where the distribution of the excess loss has exponentially decaying tails in both directions. The (u, c)-witness condition (34) does not always hold for such excess losses, but we now show that the τ-witness condition is *always* guaranteed to hold in such cases for a non-increasing function τ, which leads to a bound on excess risk that is only a log factor worse than the direct bound on the annealed risk of Lemma 5.

Definition 15 *Suppose that for given* $(P, \ell, \mathcal{F})$ *and a collection of random variables* $\{U_f : f \in \mathcal{F}\}$*, there is a* $0 < \kappa < \infty$ *such that* $\sup_{f \in \mathcal{F}} \mathbf{E}\left[e^{\kappa U_f}\right] < \infty$*. Then we say that* U_f has a uniformly exponential upper tail.

The name reflects that U_f has uniformly exponential upper tails if and only if there are constants $c_1, c_2 > 0$ such that for all $u > 0$, $f \in \mathcal{F}$, $P(U_f \geq u) \leq c_1 e^{-c_2 u}$, as is easily shown (we omit the details).

Lemma 16 *Define* $M_\kappa := \sup_{f \in \mathcal{F}} \mathbf{E}\left[e^{\kappa L_f}\right]$ *and assume that* L_f *has a uniformly exponential upper tail, so that* $M_\kappa < \infty$*. Then, for the map* $\tau : x \mapsto 1 \vee \kappa^{-1} \log \frac{2M_\kappa}{\kappa x} = O(1 \vee \log(1/x))$*, the* (τ, c)*-witness condition holds with* $c = 1/2$*.*

Now let $\bar{\eta} > 0$. Assume both the $\bar{\eta}$-strong central condition, i.e., $\mathbf{E}\left[e^{-\bar{\eta} L_f}\right] \leq 1$, and that L_f has a uniformly exponential upper tail. As an immediate consequence of the lemma above, Theorem 14 now gives that for any learning algorithm $\Pi_|$ for any $\eta \in (0, \bar{\eta})$, (using $\lambda = 1/n$), there is $C_\eta < \infty$ such that

$$\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{E}[L_{\underline{f}}]\right] \trianglelefteq_{\frac{\eta \cdot n}{C_\eta \log n}} \frac{1}{n} + C_\eta \cdot (\log n) \mathrm{IC}_{n,\eta}\left(\Pi_|\right), \tag{40}$$

so we obtain an excess risk bound that is only a log factor worse than the bound that can be obtained for the generalized Hellinger metric in Theorem 14.

Example 5 (Generalized Linear Models and Witness) Consider again Example 3, about GLMs. Heide et al. (2019, Appendix B) show that, under the three assumptions that we informally listed in Example 3, the conditions of Lemma 16 are satisfied. We can thus use (40) to give us that, up to log-factors, for misspecified GLMs satisfying the three conditions mentioned in Example 3 and generalized Bayesian estimators based on priors that are continuous and bounded away from 0 on $\mathcal{F}$, we can prove a rate of order $\tilde{O}(d/n)$, which, up to log factors, is equal to the minimax parametric rate. □

As a second consequence of Lemma 16, this time combined with (38) from Lemma 13 with $\lambda = \mathbf{E}^{\mathrm{HE}(\eta)}\left[L_f\right]$, we find that under the conditions of Lemma 16, there is $C_\eta < \infty$ such that

$$\mathbf{E}[L_{\underline{f}}] \leq \max\left\{\mathbf{E}^{\mathrm{HE}(\eta)}\left[L_f\right], C_\eta \cdot \mathbf{E}^{\mathrm{HE}(\eta)}\left[L_f\right] \cdot \log \frac{1}{\mathbf{E}^{\mathrm{HE}(\eta)}\left[L_f\right]}\right\}. \tag{41}$$

The above result generalizes a bound due to Wong and Shen (1995), as we now show.

Example 6 The bound (41) generalizes a bound of Wong and Shen (1995). Their result, the first part of their Theorem 5, allows one to bound KL divergence in terms of Hellinger distance, i.e., it holds in the special case of well-specified density estimation under log loss with the choice $\bar{\eta} = 1$, $\eta = 1/2$. Formally, consider probability model $\{P_f \mid f \in \mathcal{F}\}$ where each P_f has density p_f, and assume the model is well-specified in that $Z \sim P = P_{f^*}$ with $f^* \in \mathcal{F}$. Wong and Shen (1995) consider the condition that for some $0 < \kappa < 1$, it holds that $M'_\kappa := \sup_{f \in \mathcal{F}} \int_{(p_f/p_{f^*}) \geq e^{1/\kappa}} p_{f^*} (p_{f^*}/p_f)^\kappa < \infty$. They show that, under this condition, the following holds for all $f \in \mathcal{F}$ in the regime $\mathrm{H}_{1/2}(P_{f^*} \parallel P_f) = \mathbf{E}^{\mathrm{HE}(\eta)}\left[L_f\right] \leq \frac{1}{2}\left(1 - e^{-1}\right)^2$:

$$\mathbf{E}[L_f] \leq \left(6 + \frac{2 \log 2}{(1 - e^{-1})^2} + \frac{4}{\kappa} \max\left\{2, \log \frac{M'_\kappa}{\mathbf{E}^{\mathrm{HE}(\eta)}\left[L_f\right]}\right\}\right) \mathbf{E}^{\mathrm{HE}(\eta)}\left[L_f\right], \tag{42}$$

where $\ell_f = -\log p_f$ is log loss. Now, note that for this loss function and in the case $\bar{\eta} = 1$ (where their result applies too), M_κ in Lemma 16 and M'_κ in (42) satisfy $M'_\kappa \leq M_\kappa \leq M'_\kappa + e$. Comparing (42) to (41), we see that up to values of the constants, our result generalizes Wong and Shen's. □

We just showed that a τ-witness condition always holds under exponential tails of the loss. The following example shows that even if the loss random variables ℓ_f have fat (polynomial) tails, the witness condition often holds, even for constant τ. Before providing the example, we first recall the Bernstein condition (Audibert, 2004; Bartlett and Mendelson, 2006) and a useful proposition that will be leveraged in the example.

Definition 17 (Bernstein Condition) *For some $B > 0$ and $\beta \in (0, 1]$, we say $(P, \ell, \mathcal{F})$ satisfies the (β, B)-Bernstein condition if, for all $f \in \mathcal{F}$, $\mathbf{E}[L_f^2] \leq B\left(\mathbf{E}[L_f]\right)^\beta$.*

The best case of the Bernstein condition is when the exponent β is equal to 1. In past works, the Bernstein condition has mostly been used to characterize fast rates in the bounded excess loss regime, where the (u, c)-witness condition holds automatically. In that regime, the Bernstein condition for $\beta = 1$ and the central condition become equivalent (i.e. for each (β, C) pair there is some $\bar{\eta}$ and vice versa, where the relationship depends only on the upper bound on the loss; see Theorem 5.4 of Van Erven et al. (2015)). The following proposition shows that with unbounded excess losses, the Bernstein condition can also be related to the witness condition:

Proposition 18 (Bernstein implies Witness) *If $(P, \ell, \mathcal{F})$ satisfies the (β, B)-Bernstein condition, then, for any $u > B$, $(P, \ell, \mathcal{F})$ satisfies the (τ, c)-witness condition with $\tau(x) = u \cdot (1/x)^{1-\beta}$ and $c = 1 - \frac{B}{u}$. In particular, if $\beta = 1$ then $(P, \ell, \mathcal{F})$ satisfies the (u, c)-witness condition with constant u.*

The special case of this result for $\beta = 1$ will be put to use in Example 11 in Section 6.

Example 7 (Heavy-tailed Regression with Bounded Predictions) Consider a regression problem with squared loss, so that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Further assume that the risk minimizer f^* over $\mathcal{F}$ continues to be a minimizer when taking the minimum risk over the convex hull of $\mathcal{F}$. We call this assumption *convex luckiness* for squared loss. It is implied, for example, when $\mathcal{F}$ is convex or when the model is well-specified in the sense that $Y = f^*(X) + \xi$ for ξ a zero-mean random variable that is independent of X. Thus, when $\mathcal{F}$ is convex, we can enforce it; if we are not willing to work with a convex $\mathcal{F}$ (for example, because this would blow up the COMP_n in (4)), then we are "lucky" if it holds — since it allows, in general, for better rates (see Section 7 for additional discussion).

Now assume further that $\mathbf{E}[Y^2 \mid X] \le C$ a.s. and the function class $\mathcal{F}$ consists of functions f for which the predictions $f(X)$ are bounded as $|f(X)| \le r$ almost surely. Proposition 19 shows that in this setup, the Bernstein condition holds with exponent 1 and multiplicative constant $8(\sqrt{C} + r)^2$. Proposition 18 then implies that the (u, c)-witness condition holds with $u = 16(\sqrt{C} + r)^2$ and $c = \frac{1}{2}$. □

Proposition 19 *Under the assumptions of the example above, the* $(1, 8(\sqrt{C}+r)^2)$*-Bernstein condition holds.*

We note that Theorem 14 cannot be used with squared loss when Y is heavy-tailed as then the strong central condition cannot hold. Thus, while Example 7 might imply in this case that a (u, c)-witness condition holds, we do not yet have the machinery to put this fact to use. However, in Example 11, we show that weaker easiness conditions can still hold and fast rates can still be obtained.

Example 8 (Example 7 and Lemma 13 in Light of Birgé (2004)) Proposition 1 of Birgé (2004) shows that, in the case of well-specified bounded regression with Gaussian noise ξ, the excess risk is bounded by the 1/2-annealed excess risk times a constant proportional to r^2, where r is the bound on $|f(X)|$ as in Example 7. This result thus gives an analogue of Lemma 13 for bounded regression with Gaussian noise and also allows us to apply one of our main results, Theorem 29 below (excess risk bounds with heavy-tailed losses), for this model. Our earlier Example 7 extends Birgé's result, since it shows that the excess risk can be bounded by a constant times the annealed excess risk if the target Y has an almost surely uniformly bounded conditional second moment, which, in the well-specified setting in particular, specializes to $\xi \mid X$ almost surely having (uniformly) bounded second moment (and thus potentially having quite heavy tails) rather than Gaussian tails. On the other hand, (Birgé, 2004, Section 2.2) also gives a negative result for sets $\mathcal{F}$ that are not bounded (i.e. $\sup_{x \in \mathcal{X}, f \in \mathcal{F}} |f(x)| = \infty$): even in the "nice" case of Gaussian regression, there exist such sets for which the ratio between excess risk and annealed excess risk can be arbitrarily large, i.e., there exists no finite constant c_u for which (37) holds for all $f \in \mathcal{F}$. From this we infer, by using Lemma 13 in the contrapositive direction, that for such $\mathcal{F}$ the witness condition also does not hold. □

Example 9 (Witness vs. the Small-ball Assumption) Intriguingly, the witness condition intuitively bears some similarity to the small-ball assumption of Mendelson (2014). This assumption states that there exist constants $\kappa > 0$ and $\epsilon \in (0, 1)$ such that, for all $f, h \in \mathcal{F}$, we have

$$\Pr\left(|f - h| \ge \kappa \|f - h\|_{L_2(P)}\right) \ge \varepsilon. \tag{43}$$

Under this assumption, Mendelson (2014) established bounds on the $L_2(P)$-parameter estimation error $\|\hat{f} - f^*\|_{L_2(P)}$ in function learning. For the special case that $h = f^*$, one can read the small-ball assumption as saying that "no f behaving very similarly to f^* with high probability is very different from f^* only with very small probability so that it is still quite different on average." The witness condition reads as "there should be no f that is no worse than f^* with high probability and yet with very small probability is much worse than f^*, so that on average it is still substantially worse". Despite this similarity, the details are quite different. In order to compare the approaches, we may consider regression with squared loss in the well-specified setting as in the example above. Then the $L_2(P)$-estimation error becomes equivalent to the excess risk, so both Mendelson's and our results below bound the same quantity. But in that setting one can easily construct an example where the witness and strong central conditions hold (so Theorem 14 applies) yet the small-ball assumption does not (Example 16 in Appendix I); but it is also straightforward to construct examples of the opposite by noting that small-ball assumption does not refer to Y whereas the witness condition does. In Section 6.3 we will see that, nevertheless, the small-ball assumption can be related to the τ-witness condition for a particular τ that is needed in the unbounded risk scenario (Theorem 31). □

6. Bounds under Weaker Easiness Conditions

In many learning problems, there is no $\eta > 0$ such that the strong η-central condition is satisfied. Yet, it turns out that in many cases of interest there still exist weaker conditions under which fast convergence rates are possible. We consider two types of conditions. Both are best understood by generalizing the notion of excess risk: whereas hitherto, this was invariably defined as the risk (expected loss of some learner $\Pi_|$) relative to the comparator f^* that was optimal within $\mathcal{F}$, we will now also allow more general comparators that lie outside $\mathcal{F}$. In particular we will consider as comparator a *pseudo-predictor* g with risk $\mathbf{E}[\ell_g] = \mathbf{E}[\ell_{f^*}] - \epsilon$ for some small $\epsilon > 0$. Being better than f^*, g does not correspond to an action that can be actually played, but one can often find a g such that, with f^* replaced by g, the η-central condition does hold for some $\eta > 0$ while, simultaneously, ϵ is so small that an excess risk bound relative to g implies also a good excess risk bound relative to the original comparator f^*. We will soon introduce a function v that modulates how large one can take η for a desired ϵ (the larger η, the better the bounds that ensue).

In order to work with comparators that are pseudo-predictors, we now introduce $\bar{\mathcal{F}}$, an enlarged action space that is a superset of $\mathcal{F}$ and that also contains the pseudo-predictors we use in the remainder of this work. These pseudo-predictors always will be deterministic and typically will be constant-shifted versions of ℓ_f (for some $f \in \mathcal{F}$) or versions of a GRIP (introduced in Definition 23). Although a given pseudo-predictor $f \in \bar{\mathcal{F}}$ can fail to be well-defined as a playable action, the loss ℓ_f of any pseudo-action we employ will always be well-defined. We thus extend our loss notation $\ell_f(z)$ to all $f \in \bar{\mathcal{F}}$.

We first consider the *v-central condition*, a strict weakening of the strong central condition which applies if the excess loss is bounded or has exponential tails; here the comparator can be taken to be a trivial modification of f^*. We next consider the *v-PPC condition*, a strict weakening of the v-central condition, which applies if the losses have polynomial tails. It is based on using a new type of comparator, the *generalized reversed information projection* (*GRIP*), which generalizes a concept from Barron and Li (1999). In Section 6.1 we present the v-central condition and a corresponding excess risk bound for bounded excess risks. Section 6.2 presents the v-PPC condition, the GRIP, and the corresponding excess risk bound for bounded excess risks. Finally, Section 6.3 shows risk bounds under the v-PPC and v-central conditions for unbounded excess risks.

6.1. The v-Central Condition

Definition 20 (v-Central Condition (Van Erven et al., 2015)) *Let $\eta > 0$ and $\epsilon \geq 0$. We say that $(P, \ell, \mathcal{F})$ satisfies the η-central condition up to ϵ if there exists some $\tilde{f} \in \mathcal{F}$ such that*

$$\ell_{\tilde{f}} - \ell_f \trianglelefteq_\eta \epsilon \qquad \textit{for all } f \in \mathcal{F}. \tag{44}$$

Let $v : [0, \infty) \to [0, \infty)$ be a bounded, non-decreasing function satisfying $v(\epsilon) > 0$ for all $\epsilon > 0$. We say that $(P, \ell, \mathcal{F})$ satisfies the v-central condition if, for all $\epsilon \geq 0$, there exists a function $\tilde{f} \in \mathcal{F}$ such that (44) *is satisfied with $\eta = v(\epsilon)$.*

The special case with constant $v(\epsilon) \equiv \bar{\eta}$ reduces to the earlier strong $\bar{\eta}$-central condition (and then $\tilde{f}$ must be optimal so we can take $\tilde{f} = f^*$); for nonconstant v, the condition is weaker in that it allows a little slack ϵ, and to make ϵ small, we need to take η small. For each $\epsilon \geq 0$, we now define f^*_ϵ in terms of its loss by $\forall z \in \mathcal{Z} : \ell_{f^*_\epsilon}(z) := \ell_{f^*}(z) - \epsilon$. This f^*_ϵ plays the role of alternative comparator referred to above. We can now apply Lemma 5 with f^*_ϵ instead of f^* to get a bound on the annealed excess risk:

$$\mathbf{E}_{\underline{f} \sim \Pi_n} \left[\mathbf{E}^{\text{ANN}(\eta)} \left[\ell_f - \ell_{f^*_\epsilon} \right] \right] \quad \trianglelefteq_{\eta \cdot n} \quad \text{IC}_{n,\eta}(\Pi_|) + \epsilon. \tag{45}$$

Analogous to the story in Section 5.1, we want to turn this bound into an actual excess risk bound. This is done by the following lemma, which is a straightforward consequence from the first part of Lemma 13 and only differs from it in that it has ℓ_{f^*} on the right-hand side replaced by $\ell_{f^*_\epsilon}$ and a slightly larger constant factor.

Lemma 21 *Let $(P, \ell, \mathcal{F})$ be a learning problem that satisfies the v-central condition for some v. Let $f \in \mathcal{F}$. Suppose that (34) holds for some $u > 0$ and $c \in (0, 1]$, i.e., $(P, \ell, \{f, f^*\})$ satisfies the (u, c)-witness condition. Fix $\epsilon \geq 0$ and let $\bar{\eta} = v(\epsilon)$. As in Lemma 13, let $c_u = \frac{1}{c} \frac{\eta u + 1}{1 - \frac{\eta}{\bar{\eta}}}$. Then for all $\eta \in (0, \bar{\eta})$,*

$$\mathbf{E}[L_f] \;\leq\; c_{u+\epsilon} \cdot \mathbf{E}^{\text{ANN}(\eta)} \left[\ell_f - \ell_{f^*_\epsilon} \right]. \tag{46}$$

In particular, if $(P, \ell, \mathcal{F})$ satisfies the (u, c)-witness condition then (46) holds for all $f \in \mathcal{F}$.

The key to the proof is that, if $(P, \ell, \mathcal{F})$ satisfies the v-central condition, then we have that

$$(P, \ell, \mathcal{F} \cup \{f^*_\epsilon\}) \text{ satisfies the } \eta\text{-central condition with } \eta = v(\epsilon). \tag{47}$$

We now show how Lemma 21 straightforwardly implies a strict strengthening of Theorem 14, one which holds under the v-central condition rather than just the $\bar{\eta}$-central condition: since (46) holds for all $f \in \mathcal{F}$, it also holds in expectation over f, under any arbitrary distribution Π over f. We can thus take expectations over Π_n on both sides of (46) and chain the resulting inequality with ESI (45). Using that for general random variables U, V and $c > 0$, $U \trianglelefteq_a V \Leftrightarrow cU \trianglelefteq_{u/c} cV$, this gives:

Theorem 22 (v-Central Excess Risk Bound - Bounded Excess Risk Case) *Let $\Pi_|$ be an arbitrary learning algorithm based on $\mathcal{F}$. Assume that $(P, \ell, \mathcal{F})$ satisfies the (u,c)-witness condition (34) and let c_u be defined as in Lemma 21. Then under the v-central condition, for any $\epsilon \geq 0$, any $0 < \eta < v(\epsilon)$:*

$$\mathbf{E}_{\underline{f} \sim \Pi_n} \Big[\mathbf{E}[L_{\underline{f}}] \Big] \quad \trianglelefteq_{\frac{\eta \cdot n}{c_{u+\epsilon}}} \quad c_{u+\epsilon} \cdot \big(\mathrm{IC}_{n,\eta}(\Pi_|) + \epsilon \big) . \tag{48}$$

Analogously to the second part of Lemma 13 and Theorem 14, one can give versions of this result for the τ-witness condition as well, but for simplicity we will not do so. This theorem allows unbounded losses but is only useful when the excess risk is bounded, i.e., $\sup_{f \in \mathcal{F}} \mathbf{E}[L_f] < \infty$, because for unbounded risk, the required (u,c)-witness condition is excessively strong; see Section 6.3.

The factor $c_{u+\epsilon}$ explodes if $\eta \uparrow v(\epsilon)$. If the v-central condition holds for some v, it clearly also holds for any smaller v, in particular for $\underline{v}(\epsilon) \coloneqq v(\epsilon) \wedge 1$. Applying the theorem with $\underline{v}$ (which will not affect the rates obtained), we may thus take $\eta = \underline{v}(\epsilon)/2$, so that $c_{u+\epsilon}$ is bounded by $\frac{1}{c}(u + \epsilon + 2)$. The ESI in (48) then implies that with probability at least $1 - e^{-K}$ the left-hand side exceeds the right-hand side by at most $\frac{(u+\epsilon+2)K}{c \eta n}$. For the case of bounded excess loss, we can further take u to be $\sup_{f \in \mathcal{F}} \|L_f\|_\infty$ and $c = 1$. Finally, in the special case when the strong $\bar{\eta}$-central condition holds, we can take $\epsilon = 0$ and $v(0) = \bar{\eta}$ and Theorem 22 specializes to Theorem 14.

In Section 6.2 below we introduce the v-PPC condition. One of the main results of Van Erven et al. (2015) (in their Section 5) is that, for bounded excess losses, the v-central condition holds for some v with $v(\epsilon) \asymp \epsilon^{1-\beta}$ if and only if the v-PPC condition hold for some v with $v(\epsilon) \asymp \epsilon^{1-\beta}$ if and only if the Bernstein condition holds for exponent β and some $B > 0$; the three conditions are thus equivalent up to constant factors in the bounded excess loss case. The best case of the Bernstein condition of $\beta = 1$ corresponds to a v with $v(0) > 0$, i.e., to the strong central condition. The Bernstein condition is known to characterize the rates that can be obtained in bounded excess loss problems for proper learners, and the same thus holds for the v-central and v-PPC conditions. It is also implied by the well-known *Tsybakov margin condition* as long as $\mathcal{F}$ contains the Bayes optimal classifier (see (Lecué, 2011) and (Van Erven et al., 2015) for discussion).

We now illustrate Theorem 22 for the case of ERM over certain parametric classes when the v-central condition holds for v of the form $v(\epsilon) \asymp \epsilon^{1-\beta}$, so that a Bernstein condition holds with exponent β. We will see that for bounded losses our result recovers, up to log factors, rates that are known to be minimax optimal. We first need some notation. For a pseudo-metric space $(\mathcal{A}, \| \cdot \|)$ and any $\epsilon > 0$, let $\mathcal{N}(\mathcal{A}, \| \cdot \|, \epsilon)$ be the ϵ-covering number of $(\mathcal{A}, \epsilon)$, defined as the minimum number of radius-ϵ balls whose union contains $\mathcal{A}$.

Example 10 (Lipschitz (and Bounded) Loss) Suppose that *(i)* for each $z \in \mathcal{Z}$, the loss ℓ is G-Lipschitz as a function of $f \in \mathcal{F}$; *(ii)* $\mathcal{F}$ has bounded metric entropy in some pseudometric $\|\cdot\|$; and *(iii)* the loss is uniformly bounded over $\mathcal{F}$ (so that a witness condition holds). Let $\mathcal{F}_\epsilon$ be an optimal ϵ-net with respect to $\| \cdot \|$. Take a uniform prior over $\mathcal{F}$, and (purely for the analysis) consider the randomized predictor $\Pi_|$ that predicts by drawing an f uniformly from a radius-ϵ ball around $\hat{f}$, the ERM predictor. If the v-central condition holds, it follows that the information complexity of $\Pi_|$ is bounded as $G\epsilon + \frac{\log \mathcal{N}(\mathcal{F}, \|\cdot\|, \epsilon)}{v(\varepsilon) n}$. To see this, for any $A \subset \mathcal{F}$ let A^ϵ be the ϵ-extension of A, defined as $\left\{f \in \mathcal{F} \colon \inf_{f' \in A} \|f - f'\| \le \epsilon\right\}$. Then observe that

$$e^{\mathrm{KL}(\Pi_n \,\|\, \Pi_0)} = \frac{\mathrm{vol}(\mathcal{F})}{\mathrm{vol}(\{\hat{f}\}^\epsilon)} \le \frac{\mathrm{vol}(\bigcup_{f \in \mathcal{F}_\epsilon} \{f\}^\epsilon)}{\mathrm{vol}(\{\hat{f}\}^\epsilon)} \le \frac{\sum_{f \in \mathcal{F}_\epsilon} \mathrm{vol}(\{f\}^\epsilon))}{\mathrm{vol}(\{\hat{f}\}^\epsilon)} = \mathcal{N}(\mathcal{F}, \|\cdot\|, \epsilon).$$

Moreover, it is easy to see that the risk of standard ERM (rather than its randomized version) over the entire class $\mathcal{F}$ is at most the risk of Π_n plus an additional $G\epsilon$. Hence, if v satisfies $v(\epsilon) = C\epsilon^{1-\beta}$ for some $\beta \in [0,1]$ and if the metric entropy is logarithmic in ϵ, then by tuning ϵ and η as in (7) we see from (48) that ERM obtains a rate of $\tilde{O}(n^{-1/(2-\beta)})$ (suppressing log-factors) with high probability — which is the minimax optimal rate in this setting (Van Erven et al., 2015). Note that the Bernstein condition is automatically satisfied for $\beta = 0$, yielding the slow rate of $\tilde{O}(1/\sqrt{n})$, and the other extreme of $\beta = 1$ yields the fast rate of $\tilde{O}(1/n)$. □

6.2. The v-PPC Condition and the GRIP

Trivially, if the v-central condition holds for some function v, then there exists $\epsilon > 0$ such that, with $c = e^{\epsilon v(\epsilon)}$, for all $f \in \mathcal{F}$, $\mathbf{E}[e^{-v(\epsilon) L_f}] \le c$, so that $-L_f$ must have a uniformly exponential upper tail as in Definition 15. Thus, if $-L_f$ has a polynomial upper tail, the v-central condition cannot hold. The v-PPC condition is a further weakening of the v-central condition which can still hold in the latter case. We achieve this by replacing the comparator f^*_ϵ by a more sophisticated pseudo-predictor $m^\eta_{\mathcal{F}}$, the *generalized reversed information projection* (GRIP). The original projection (Li, 1999) was used in the context of density estimation under log loss. We now extend it to general learning problems:

Definition 23 (GRIP) *Let $(P, \ell, \mathcal{F})$ be a learning problem. Define*[3] *the set of pseudo-probability densities $\mathcal{E}_{\mathcal{F},\eta} := \left\{e^{-\eta \ell_f} : f \in \mathcal{F}\right\}$. For $Q \in \Delta(\mathcal{F})$, define $\xi_Q := \mathbf{E}_{\underline{f} \sim Q}[e^{-\eta \ell_{\underline{f}}}]$. The generalized reversed information projection of P onto* $\mathrm{conv}(\mathcal{E})$ *is defined as the pseudo-loss ℓ_{g_η} satisfying*

$$\mathbf{E}[\ell_{g_\eta}] = \inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}\left[-\frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim Q}\left[e^{-\eta \ell_{\underline{f}}}\right]\right] = \inf_{\xi_Q \in \mathrm{conv}(\mathcal{E})} \mathbf{E}\left[-\frac{1}{\eta} \log \xi_Q\right].$$

Following terminology from the individual-sequence prediction literature, we call the quantity appearing in the center expectation above a "mix loss" (De Rooij et al., 2014) defined for a distribution $Q \in \Delta(\mathcal{F})$ as $m^\eta_Q := -\frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim Q}\left[e^{-\eta \ell_{\underline{f}}}\right]$. The notion of mix loss can be extended from distributions to sets by defining, for any $A \subseteq \bar{\mathcal{F}}$, the object m^η_A as the pseudo-loss satisfying $\mathbf{E}\left[m^\eta_A\right] = \inf_{Q \in \Delta(A \cup \{f^*\})} \mathbf{E}\left[m^\eta_Q\right]$.[4] *We thus have that $\ell_{g_\eta} = m^\eta_{\mathcal{F}}$, and we use the latter notation from here on out.*

[3] This transformation is known as *entropification* (Grünwald, 1999). For $\eta = 1$ and log-loss, pseudo-probability densities are just standard probability densities, while for general η and ℓ, the analogy to probability densites is still useful, hence the name; in particular, ξ_Q shares some properties of mixture distributions (Van Erven et al., 2015).

[4] The reason for automatically taking the union of A with f^* is to lessen the notation for the mini-grip, introduced in Appendix E.2.1.

Even though the GRIP is only a pseudo-predictor, meaning that it may fail to correspond to any actual prediction function, the corresponding loss for a GRIP *is* well-defined, as shown in Appendix G. The main use of the GRIP lies in the fact that the probability that its loss exceeds that of any $f \in \mathcal{F}$ is exponentially small:

Proposition 24 *For all $f \in \mathcal{F}$, for every $\eta > 0$, we have $m^{\eta}_{\mathcal{F}} - \ell_f \trianglelefteq_{\eta} 0$.*

The proposition implies that $m^{\eta}_{\mathcal{F}} \trianglelefteq_{\eta} \ell_{f^*}$ and hence $\mathbf{E}[m^{\eta}_{\mathcal{F}}] \le \mathbf{E}[\ell_{f^*}]$ and, for any $\eta > 0$, $\mathcal{F} \cup \{m^{\eta}_{\mathcal{F}}\}$ satisfies the η-central condition, with $m^{\eta}_{\mathcal{F}}$ in the role of f^*. We can now define the v-PPC condition:

Definition 25 (Pseudoprobability Convexity (PPC) Condition) *Let $\eta > 0$ and $\varepsilon \ge 0$. We say that $(P, \ell, \mathcal{F})$ satisfies the η-PPC condition up to ε if there exists some $\tilde{f} \in \mathcal{F}$ such that*

$$\mathbf{E}_{Z \sim P}\left[\ell_{\tilde{f}}\right] - \inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}\left[-\frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim Q}\left[e^{-\eta \ell_{\underline{f}}}\right]\right] \le \epsilon, \quad \text{i.e.,} \quad \mathbf{E}_{Z \sim P}\left[\ell_{\tilde{f}} - m^{\eta}_{\mathcal{F}}\right] \le \epsilon. \tag{49}$$

Let $v : [0, \infty) \to [0, \infty)$ be a bounded, non-decreasing function satisfying $v(\epsilon) > 0$ for all $\epsilon > 0$. We say that $(P, \ell, \mathcal{F})$ satisfies the v-PPC condition if, for all $\epsilon \ge 0$, there exists a function $\tilde{f} \in \mathcal{F}$ such that (49) *is satisfied with $\eta = v(\epsilon)$.*

In both the v-central and v-PPC conditions, we look at pairs (η, ϵ) such that there exists a comparator g which has risk no better than $\mathbf{E}[\ell_{f^*}] - \epsilon$, and for which $(P, \ell, \mathcal{F} \cup \{g\})$ satisfies the η-central condition. We achieve this for any (η, ϵ) with $0 < \eta \le v(\epsilon)$, where for the v-central condition, the comparator was $g = f^*_{\epsilon}$ (see (47)) and for the v-PPC condition, it is $g = m^{\eta}_{\mathcal{F}}$.

The name "PPC" stems from the fact that the condition expresses a pseudo-convexity property of the set of pseudoprobability densities mentioned in Definition 23; see Van Erven et al. (2015) for a graphical illustration and for the proof that the v-central condition implies the v-PPC condition for the same v. We already mentioned that Van Erven et al. (2015) (in their Section 5) proved the reverse implication, hence equivalence of the v-central and v-PPC conditions, up to constant factors, for bounded excess losses. To give some initial intuition for the unbounded case, we note that the v-PPC condition is satisfied for $v(\epsilon) = C \cdot \epsilon$ for a suitable constant C whenever the witness condition holds. While this was known for bounded excess losses (where linear v corresponds to the weakest Bernstein condition, which automatically holds), by Proposition 26 below it turns out to hold even if the excess losses are heavy-tailed (so the v-central condition can never hold) and the risk can be unbounded, as long as the second moment of the risk of f^* is finite. This will imply, for example, (Theorem 31 below and discussion) that the "slow" $\tilde{O}\left(1/\sqrt{n}\right)$ excess risk rate for parametric models can be obtained in-probability by η_n-generalized Bayes (with the optimal η_n depending on the sample size as $\eta_n \asymp 1/\sqrt{n}$) under hardly any conditions.

Proposition 26 *Let $(P, \ell, \mathcal{F})$ be such that for all $f \in \mathcal{F}$, all $z \in \mathcal{Z}$, $\ell_f(z) \ge 0$ and such that for some fixed $u > 0$, for all $f \in \mathcal{F}$ with $\mathbf{E}[L_f] > 0$,*

$$\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \le u\}}\right] \ge 0. \tag{50}$$

(in particular this is implied by the (u,c)-witness condition (34)). Then for all $\eta \leq 1/\mathbf{E}[\ell_{f^*}]$,

$$\mathbf{E}_{Z\sim P}\left[\ell_{f^*} - m_{\mathcal{F}}^{\eta}\right] \leq \eta \cdot e \cdot \left(u^2 + \frac{3}{2}\mathbf{E}[\ell_{f^*}^2]\right).$$

As a consequence of this result, if we have $\mathbf{E}_{Z\sim P}\left[\ell_{f^*}^2\right] < \infty$, then the v-PPC condition holds with $v(\epsilon) = (C\epsilon) \wedge (1/\mathbf{E}[\ell_{f^*}])$, where $C = e^{-1}\cdot(u^2 + \frac{3}{2}\mathbf{E}[\ell_{f^*}^2])^{-1}$.

The proof of this proposition is based on the following fact, interesting in its own right and also used in the proof of later results:

Proposition 27 *For given learning problem $(P,\ell,\mathcal{F})$, let ℓ' be such that (a) for all $f \in \mathcal{F}$, all $z \in \mathcal{Z}$, $\ell'_f(z) \leq \ell_f(z)$, and (b), $\ell'_{f^*}(z) = \ell_{f^*}(z)$. If the "smaller-loss" learning problem $(P,\ell',\mathcal{F})$ satisfies the v-PPC condition for some function v, then so does $(P,\ell,\mathcal{F})$.*

We now work towards a first risk bound under the v-PPC condition, using the GRIP. The development is entirely analogous to that leading up to Theorem 22, our risk bound under the v-central condition. We start with the following result, which essentially only differs from Lemma 13 and the corresponding lemma for the v-central condition and f_ϵ^*-comparator, Lemma 21, in that it has ℓ_{f^*} (as in Lemma 13) and $\ell_{f_\epsilon^*}$ (as in Lemma 21) on the right-hand side replaced by the GRIP loss $m_{\mathcal{F}}^{\bar{\eta}}$ and requires $\eta < \bar{\eta}/2$. The proof is much more involved though since the comparators on the left and the right are not connected in a straightforward manner.

Lemma 28 *Let $(P,\ell,\mathcal{F})$ be a learning problem and let $f \in \mathcal{F}$. Let $\bar{\eta} > 0$. Suppose that (34) holds for some $u > 0$ and $c \in (0,1]$, i.e., $(P,\ell,\{f,f^*\})$ satisfies the (u,c)-witness condition. Let $c'_u := \frac{1}{c}\frac{\eta\cdot u+1}{1-\frac{2\eta}{\bar{\eta}}}$. Then for all $\eta \in (0,\bar{\eta}/2)$,*

$$\mathbf{E}[L_f] \;\leq\; c'_{2u}\cdot \mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_f - m_{\mathcal{F}}^{\bar{\eta}}\right]. \tag{51}$$

In particular, if $(P,\ell,\mathcal{F})$ satisfies the (u,c)-witness condition then (51) holds for all $f \in \mathcal{F}$.

Based on this lemma it is now easy to prove analogues of Theorem 14. Below we first present our second main result, an excess risk bound that holds under the basic witness condition. The result allows unbounded and heavy-tailed losses but is only useful when the excess risk is bounded; see Section 6.3.

Theorem 29 (Excess Risk Bound - Bounded Excess Risk Case) *Let $\Pi_|$ be an arbitrary learning algorithm based on $\mathcal{F}$. Assume that $(P,\ell,\mathcal{F})$ satisfies the (u,c)-witness condition (34). Let c'_u be as in Lemma 28. Then under the v-PPC condition, for any $\eta < \frac{v(\epsilon)}{2}$,*

$$\mathbf{E}_{Z_1^n}\left[\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\mathbf{E}[L_{\underline{f}}]\right]\right] \leq c'_{2u}\left(\mathbf{E}_{Z_1^n}\left[\mathrm{IC}_{n,\eta}(\Pi_|)\right] + \epsilon\right). \tag{52}$$

The result is entirely analogous to Theorem 22 (and the remarks made there apply here as well), with two differences: first, v is replaced by $v/2$, which will worsen the obtainable bounds by a factor of 2 and hence will not affect the rates. Second, the ESI in (48) is replaced by an expectation. Thus, we have an exponential in-probability bound (holding with probability $1-\delta$ up to an $O(\log(1/\delta))$-term) under the v-central condition but not under the v-PPC condition. That such a deviation bound does not hold under the v-PPC condition is inevitable since all of our bounds are valid for ERM estimators, which, under heavy-tailed loss distributions, are known to behave poorly in probability (Catoni, 2012, Proposition 6.2). There exist specialized M-estimators for mean estimation problems (Catoni, 2012) or more generally (for regression problems) that achieve better high-probability bounds by employing a variation of the median-of-means idea (Nemirovskii and Yudin, 1983; Hsu and Sabato, 2016; Lugosi and Mendelson, 2019).

To illustrate Theorem 29, we now provide an example where the v-central condition cannot hold because the excess risk has polynomially decaying tails; yet, the v-PPC condition may still hold for v that allow for faster rates than the "slow" $\tilde{O}(1/\sqrt{n})$.

Example 11 (Heavy-tailed Regression with Bounded Predictions, Continued) We continue with the setting of Example 7. In addition to assuming that $\mathbf{E}[Y^2 \mid X] \le C$ a.s. for a constant C, we also assume that $\mathbf{E}[|Y|^s] < \infty$ for some $s \ge 2$; note that the first assumption already implies the second for $s = 2$. We further assume that $\mathcal{F}$ has bounded metric entropy in sup-norm, with covering numbers $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon)$ growing polynomially in ϵ. Without subexponential tail decay, the v-central condition fails to hold for any nontrivial v; however, as shown by Van Erven et al. (2015, Example 5.10) (based on a result of Juditsky et al. (2008)), if $\mathbf{E}[|Y|^s] < \infty$ for some $s \ge 2$, then the v-PPC condition holds for $v(\epsilon) = O(\epsilon^{2/s})$.[5] Moreover, as we showed in Example 7, the witness condition holds if $\mathbf{E}[Y^2 \mid X] < \infty$ a.s.; there, we also established that the Bernstein condition holds with $\beta = 1$.

Now, take a uniform prior over $\mathcal{F}$, and take the randomized predictor $\Pi_|$ as in Example 10 which randomizes over an ϵ-ball around the ERM predictor $\hat{f}$. Then, for $s \ge 2$, Theorem 29 implies that the expected excess risk of Π_n is at most

$$\mathbf{E}_{Z_1^n}\left[\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\frac{1}{n}\sum_{j=1}^n L_{\underline{f}}(Z_j)\right]\right] + \frac{\log\mathcal{N}(\mathcal{F}, \|\cdot\|, \epsilon)}{v(\epsilon)n} + \epsilon.$$

The first term can be bounded as

$$\begin{aligned}
&\mathbf{E}_{Z_1^n}\left[\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\frac{1}{n}\sum_{j=1}^n\Big(L_{\hat{f}}(Z_j) + \ell_{\underline{f}}(Z_j) - \ell_{\hat{f}}(Z_j)\Big)\right]\right]\\
&\le \mathbf{E}_{Z_1^n}\left[\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\frac{1}{n}\sum_{j=1}^n\Big(\ell_{\underline{f}}(Z_j) - \ell_{\hat{f}}(Z_j)\Big)\right]\right]\\
&= \mathbf{E}_{Z_1^n}\left[\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\frac{1}{n}\sum_{j=1}^n\big(\underline{f}^2(X_j) - \hat{f}^2(X_j) + 2Y_j(\hat{f}(X_j) - \underline{f}(X_j))\big)\right]\right]\\
&\le \mathbf{E}_{Z_1^n}\left[2\epsilon\left(\|\mathcal{F}\|_\infty + \left(\frac{1}{n}\sum_{j=1}^n Y_j^2\right)^{1/2}\right)\right],
\end{aligned}$$

[5] What is actually shown there is that a property called v-stochastic exp-concavity holds, but, the results of that paper imply then that v-stochastic mixability holds which in turn implies that the v-PPC condition holds.

which is at most $2\epsilon\left(\|\mathcal{F}\|_\infty + \|Y\|_{L_2(P)}\right) = O(\epsilon)$, and it is simple to verify that the ERM predictor $\hat{f}$ satisfies the same bound. Tuning ϵ in $O\left(\epsilon + \frac{\log \mathcal{N}(\mathcal{F},\|\cdot\|,\epsilon)}{\epsilon^{2/s} n}\right)$ yields a rate of $\tilde{O}(n^{-s/(s+2)})$ in expectation, where the notation hides log factors. □

Two remarks are in order about the rate obtained in the above example.

First, Juditsky et al. (2008) previously obtained this rate for finite classes $\mathcal{F}$ without the assumption that $\mathbf{E}[Y^2 \mid X]$ is almost surely uniformly bounded; their result is achieved by an online-to-batch conversion of a sequential algorithm which, after the conversion, plays actions in the convex hull of $\mathcal{F}$. It is unclear if we truly need the assumption on the conditional second moment of Y or if the need for this assumption is just an artifact of our analysis. In the regime where our stronger assumption holds, in the case of convex luckiness (see Example 7) the rates obtained in the present paper match those of Juditsky et al. (2008). However, if convex luckiness does not hold, then the results of Juditsky et al. (2008) still enjoy the rate of $\tilde{O}(n^{-s/(s+2)})$ whereas we cannot guarantee this rate. This is not surprising: without convex luckiness, "improper learners" that play in the convex hull of $\mathcal{F}$ are inherently more powerful than (randomized) proper learners.

Second, even when convex luckiness does hold, the rate obtained in Example 11 above is not optimal. The reason is that in the setting of this example, a Bernstein condition with $\beta = 1$ does hold, as was established earlier in Example 7. Thus, via Corollary 6.2 of Audibert (2009) it is possible to obtain the better rate of $\tilde{O}(1/n)$ in expectation using Audibert's SeqRand algorithm. Notably, the SeqRand algorithm for statistical learning involves using a sequential learning algorithm which incorporates a second-order loss-difference term. For new predictions, SeqRand employs an online-to-batch conversion based on drawing functions uniformly at random from the set of previously played functions. It is thus a randomized proper learning algorithm. There are now two possibilities. The first is that there exist $\mathcal{F}$ satisfying the condition of Example 7 for which ERM and η-generalized Bayes simply do not achieve the rate of $\tilde{O}(1/n)$; in that case either SeqRand's second-order nature or its online-to-batch step may be needed to get the fast rate. The other possibility is that ERM and η-generalized Bayes do generally attain the fast rate under the Bernstein condition and a.s. bounded $\mathbf{E}[Y^2 \mid X]$-condition, in which case Theorem 29 is suboptimal for this situation — we return to this issue in the Discussion (Section 7). In any case, SeqRand is computationally intractable for most infinite classes, and we are not aware of any polynomial-time learning algorithms that match the rate of SeqRand.

6.3. Bounds for Unbounded Excess Risk

We now present a result for a learning problem $(P, \ell, \mathcal{F})$ with unbounded excess risk. Once again, the result follows (now with some work) from Lemma 28, but now we need to be careful because the (u,c)-witness condition with fixed u and c cannot be expected to hold: it would become an exceedingly strong condition for $\mathbf{E}[L_f] \to \infty$. We will thus require the τ-witness condition for a particular, easier τ, namely $\tau(x) = u(1 \vee x)$ for some $u \geq 1$, so that for large x, $\tau(x) \asymp x$. We first show, in Proposition 30 below, that at least for the squared loss this condition can be expected to hold in a variety of situations. The price to pay for using this τ is that we only get in-probability results — we show those in Theorem 31 (we do not know whether in-expectation results hold as well). Note that one could obtain better constants in that theorem if one employed $\tau(x) = a \vee (bx)$ for the best possible a and b, but for simplicity we did not do this.

Proposition 30 (Bernstein plus small-ball implies unbounded witness) *Consider the setting of Example 7, i.e., regression with ℓ the squared loss and convex luckiness. We still assume convex luckiness and make the weaker assumption $\mathbf{E}[Y^2] < \infty$, but now we do not* not *assume that the risk is bounded; i.e., we can have $\sup_{f \in \mathcal{F}} \mathbf{E}[\ell_f] = \infty$. Fix some $b > 0$ and suppose that there exists constants $\kappa > 0, \epsilon \in (0,1)$ such that*

1. *for all $f \in \mathcal{F}$ with $\mathbf{E}[L_f] > b$, Mendelson's (2014) small-ball assumption (43) holds with constants ϵ, κ for f, f^* (i.e. with f^* in the role of h),*
2. *For all $c_0 > b$, all $f \in \mathcal{F}$ with $\mathbf{E}[L_f] \le c_0$, there is a B such that the $(1, B)$-Bernstein condition holds, i.e., $\mathbf{E}[L_f^2] \le B\, \mathbf{E}[L_f]$.*

Then $(P, \ell, \mathcal{F})$ satisfies the (τ, c)-witness condition, with $\tau(x) = u(1 \vee x)$ for some $u \ge 1$ and with $c \in (0,1]$ which depends only on κ, ϵ, b, and $\mathbf{E}[\ell_{f^*}]$.

Example 12 (Heavy-tailed Regression, Continued) Mendelson provides several examples of convex $\mathcal{F}$ for which the small-ball assumption holds; the proposition above shows that for all these examples, the τ-witness condition holds as well as soon as, for f with small excess risk, the Bernstein condition holds. For example, under the following "meta"-condition the small-ball assumption holds (see (Mendelson, 2014, Lemma 4.1)) and, as we show in Appendix C.3, the Bernstein condition holds as well for $\mathcal{F}_{c_0} := \{f \in \mathcal{F} : \mathbf{E}[L_f] < c_0\}$, for all $c_0 \ge b$, as long as we assume convex luckiness (see Example 7).

$$\mathbf{E}[\ell_{f^*}^2] < \infty \quad \text{and} \quad \text{for some } A > 0, \text{ for all } f \in \mathcal{F}_{c_0},$$
$$\mathbf{E}[(f(X) - f^*(X))^4]^{1/2} \le A \cdot \mathbf{E}[(f(X) - f^*(X))^2].$$

We stress however that our theorem below does not recover Mendelson's rates for $L_2(P)$-estimation error (Section 7), which rely on further highly sophisticated analysis of the squared loss situation; our goal here is merely to show that our τ-witness condition for the unbounded risk case is not a very strong one. □

Theorem 31 (Excess Risk Bound - Unbounded Excess Risk Case) *Assume that $(P, \ell, \mathcal{F})$ satisfies the (τ, c)-witness condition (35) with $\tau : x \mapsto u(1 \vee x)$ for some $u \ge 1$ and constant c. Let $\epsilon_1, \epsilon_2, \ldots$ and $\eta_1, \eta_2, \ldots$ be sequences such that*

$$\epsilon_n \to 0, \qquad n\eta_n \to \infty.$$

Let $c_u := \frac{u}{c} \frac{\eta_n + 1}{1 - \frac{\eta_n}{v(\epsilon_n)}}$ and $c'_u := \frac{u}{c} \frac{\eta_n + 1}{1 - \frac{2\eta_n}{v(\epsilon_n)}}$. Suppose that $\mathrm{IC}_{n,\eta} := \mathrm{IC}_{n,\eta}(\Pi_|)$ is nontrivial in the sense that $\mathbf{E}[\mathrm{IC}_{n,\eta_n}] \to 0$.

1. *Let $\Pi_| \equiv (\hat{f}, \Pi_0)$ represent a deterministic estimator. Suppose that, for given function v, the v-PPC condition holds and that for all n, $0 < \eta_n < v(\epsilon_n)/2$. Then for all n larger than some n_0, the right-hand side of the following equation is bounded by 1, and for all such n, for all $\delta > 0$, with probability at least $1 - \delta$,*

$$\mathbf{E}[L_{\hat{f}}] \le \left(c'_{2u} \cdot \frac{1}{\delta} \right) \cdot \text{BOUND}, \quad \textit{with} \ \ \text{BOUND} = \left(\mathbf{E}\left[\mathrm{IC}_{n,\eta_n} \right] + \epsilon_n \right). \tag{53}$$

Now suppose that, more strongly, the v-central condition holds as well. Let $\overline{\mathrm{IC}}_{n,\eta}$ be any upper bound on $\mathrm{IC}_{n,\eta}(f^* \| \Pi_|)$ that is nontrivial in that $\mathbf{E}[\overline{\mathrm{IC}}_{n,\eta_n}] \to 0$. Let $C_{n,\delta}$ be a function of $\delta \in (0,1)$ such that for all $\delta \in (0,1)$, $C_{n,\delta} > 2\log(2/\delta)$ and

$$P\left(\overline{\mathrm{IC}}_{n,\eta_n} \geq C_{n,\delta} \cdot \mathbf{E}\left[\overline{\mathrm{IC}}_{n,\eta_n}\right]\right) \leq \delta. \tag{54}$$

Then for all n larger than some n_0, the right-hand side of the following equation is bounded by 1, and for all such n, for all $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\mathbf{E}[L_{\hat{f}}] \leq \left(c'_{u+\epsilon_n} \cdot C_{n,\delta}\right) \cdot \text{BOUND}, \quad \textit{with} \;\; \text{BOUND} = \left(\mathbf{E}\left[\overline{\mathrm{IC}}_{n,\eta_n}\right] + \epsilon_n + \frac{2}{n\eta_n}\right). \tag{55}$$

2. *Now let $\Pi_|$ be a general, potentially nondeterministic estimator, suppose that the v-PPC condition holds and let $\overline{\mathrm{IC}}_{n,\eta_n}$ be any bound on $\mathrm{IC}(\Pi_|)$ that is slightly larger than IC_{n,η_n}, i.e., there exist a sequence $a_1, a_2, \ldots \to \infty$ such that, for all n, all z^n, $\overline{\mathrm{IC}}_{n,\eta_n} \geq a_n \mathrm{IC}_{n,\eta_n}$. Then*

$$\Pi_n\left(\left\{f \in \mathcal{F} : \mathbf{E}[L_f] > c'_{2u} \cdot \text{BOUND}\right\}\right) \to 0 \;\; \textit{in } P\textit{-probability}, \tag{56}$$

with BOUND $= \mathbf{E}\left[\overline{\mathrm{IC}}_{n,\eta_n}\right] + \epsilon_n$.

When $\Pi_|$ represents a deterministic estimator $\hat{f}$ such as an η-two part MDL estimator, the result is just a standard convergence-in-probability result. For learning algorithms that output a distribution such as generalized Bayes, the result seems fairly weak as nothing is said about the rate at which the deviation probability goes to 0. Note, however, that the same holds for most standard results about posterior convergence in Bayesian statistics; for example, the results of GGV (see Example 2) are stated in exactly the same manner.

Note that the factor for the PPC-results increases quickly with δ; depending on how strong a bound (54) can be given, the v-central results can thus become substantially stronger asymptotically. This is the case even though their bound has an additional $1/(n\eta_n)$ term. Indeed, this extra term is of the right order, comparable to the upper bound on IC_{n,η_n} given by (4). Therefore, for $v(x) \asymp x^{1-\beta}$, optimization of ϵ_n and η_n can be done in the same way as for the bounded risk case, leading to a rate of $\tilde{O}(n^{-1/(2-\beta)})$ as in (7). To give an example in which the bound for the v-central condition gets a better dependence on δ than v-PPC consider generalized Bayesian posteriors under the GGV condition (21) discussed in Section 3.3; in that case, we get the bound (25) which implies (54) for a $C_{n,\delta} = o(\delta^{-1/2})$ (rather than the $O(\delta^{-1})$ in the PPC-result) and with ϵ_n, as defined there used as an upper bound on $\mathrm{IC}_{n,\eta}$. Still, in this example $C_{n,\delta}$ is polynomial in δ whereas Theorem 22 had only a logarithmic dependence on δ. As mentioned earlier, this stronger dependence on δ is unavoidable as the results under the v-PPC condition apply to methods like ERM, which have poor deviation properties.

To derive further corollaries from this theorem, we mention the following extension of Proposition 26:

Proposition 32 (When (τ, c)-witness implies v-PPC) *Suppose that the (τ, c)-witness condition holds for given learning problem $(P, \ell, \mathcal{F})$ with $\tau : x \mapsto u(1 \vee x)$ for some $u \geq 1$ and constant $c \in (0,1]$ as in Theorem 31. Further suppose that that $\ell_f(z) \geq 0$ for all $f \in \mathcal{F}$ and all $z \in \mathcal{Z}$. Then the v-PPC condition holds with $v(\epsilon) = (C\epsilon) \wedge (1/\mathbf{E}[\ell_{f^*}])$, where $C = e^{-1} \cdot (u^2 \left(1 \vee \left(\mathbf{E}[\ell_{f^*}]/c\right)^2\right) + \frac{3}{2}\mathbf{E}[\ell^2_{f^*}])^{-1}$.*

The above proposition implies that if the τ-witness condition holds with τ as in Theorem 31 above, then the results (53) and (56) automatically hold with choice $2\eta_n < (C\epsilon_n) \wedge (1/\mathbf{E}[\ell_{f^*}])$, which for large n is equivalent to $\eta_n < C\epsilon_n/2$. For parametric $\mathcal{F}$ we can take $\epsilon_n \asymp 1/\sqrt{n}$, so that the v-PPC condition is satisfied with $\eta_n \asymp 1/\sqrt{n}$. Thus, under quite weak conditions (for all f, z, $\ell_f(z) \geq 0$, $\mathbf{E}[\ell^2_{f^*}] < \infty$, and the τ-witness condition holds as above), but with unbounded, heavy tailed losses and without explicitly imposing any GRIP conditions, we get in all three cases of Theorem 31, by choosing $\eta_n \asymp 1/\sqrt{n}$, that $\text{BOUND} = \tilde{O}\left(1/\sqrt{n}\right)$. Consequently, even under very weak assumptions, we still get convergence for generalized η_n-Bayesian estimators, albeit at the "slow" rate.

7. Discussion & Open Questions

In this paper we presented several theorems that gave convergence rates for general estimators, including pseudo-Bayesian and ERM estimators, under general "easiness conditions". We end by putting these conditions in context and discussing some of the limitations of our approach, thereby pointing to avenues for future work.

Easiness Conditions We proved our convergence rates under the *GRIP* conditions (the v-central and v-PPC conditions) and the τ-*witness* condition, and we provided some relations to other conditions such as convex luckiness for squared loss (defined in Example 7), Bernstein conditions (Definition 17), and uniformly exponential tails (Definition 15). As promised in the beginning of this paper, our conditions and results complement those of Van Erven et al. (2015) which are mostly for the bounded case. The most important conditions of that paper that did not show up here are (a) the extension of *convex luckiness* beyond the squared loss (it is formally defined for general losses by Van Erven et al. (2015) under the name "Assumption B") and (b) the v-*stochastic mixability* condition (see Definition 5.9 of Van Erven et al. (2015)). We will restrict discussion of the v-stochastic mixability condition to the case where the decision set $\mathcal{F}_d$ from Van Erven et al. (2015) is equal to $\text{conv}(\mathcal{F})$. In the present paper, where the set $\mathcal{P}$ from Van Erven et al. (2015) is always equal to the singleton $\{P\}$, it is easy to see that v-stochastic mixability is equivalent to the v-PPC condition but with the minimizer f^* over $\mathcal{F}$ replaced by the minimizer f^*_{conv} over $\text{conv}(\mathcal{F})$. Van Erven et al. (2015) show that for bounded excess losses, v-stochastic mixability characterizes obtainable rates for *improper* learners that are allowed to play in the convex hull of $\mathcal{F}$. v-stochastic mixability is in turn implied by the easiness conditions of Juditsky et al. (2008), (for constant v) by conditions on the loss function such as mixability and exp-concavity (Cesa-Bianchi and Lugosi, 2006), and by strong convexity. For clarity we give an overview of the relevant implications between our conditions and those of Van Erven et al. (2015) in Figure 1.

Misspecification We showed that our methods are particularly well-suited for proving a form of consistency for (generalized Bayesian) density estimation under misspecification; under only the $\bar{\eta}$-central condition, a weak condition on the support of p_{f^*}, and using a prior such that the weakened GGV condition (22) holds, we can show that for any $\eta < \bar{\eta}$, the η-generalized Bayesian posterior is consistent in the sense of our misspecification metric (see Proposition 11 and discussion below it). As stated there, an interesting open question is under which conditions the metric entropy for the misspecified case is of the same order as

excess loss is...	condition type	loss function	result
bounded	GRIP	general	v-PPC $\Leftrightarrow$ v-central (vE) $x^{1-\beta}$-PPC $\Leftrightarrow$ x^{β}-Bernstein (vE)
	witness	general	(u, c)-witness always holds (trivial)
unbounded	GRIP	general	convex luckiness + v-stochastic mixability $\Rightarrow$ v-PPC (vE)
		general	v-central $\Rightarrow$ v-PPC (vE)
		general	v-central $\Rightarrow$ L_f has uniformly exponential lower tail (vE)
		log loss	convex luckiness $\Rightarrow$ 1-central (vE)
		squared loss	convex luckiness + bounded predictions + $Y \mid X$ has a.s. uniformly bounded 2^{nd} moment $\Rightarrow$ $(1, B)$-Bernstein (GM, Example 7)
unbounded	witness	general	(β, B)-Bernstein $\Rightarrow$ (τ, c)-witness, $\tau(x) \asymp x^{\beta-1}$ (GM, Proposition 18)
		general	L_f has uniformly exponential upper tail $\Rightarrow$ (τ, c)-witness, $\tau(x) \asymp 1 \vee \log(1/x)$ (GM, Lemma 16)
		log loss, correct model	Wong-Shen $\Leftrightarrow$ L_f has uniformly exponential tails (GM, Example 6)

Figure 1: GM stands for "established in the present paper", vE refers to Van Erven et al. (2015). All implications hold up to constant factors. Note that boundedness always refers to *excess loss*. For example, for Lipschitz losses on a bounded domain, the losses themselves may have heavy tails but the excess loss will be bounded.

the metric entropy for the well-specified case, as then the misspecification metric dominates the standard Hellinger metric.

Proper vs. Improper There exist learning problems $(P, \ell, \mathcal{F})$ on which no proper learner — one which always predicts inside $\mathcal{F}$ — can achieve a rate as good as that of an improper learner, which can select $\hat{f}_n \notin \mathcal{F}$ (Audibert, 2007; Van Erven et al., 2015). In this paper we considered *randomized* proper estimators, to which the same lower bounds apply; hence, they cannot in general compete with improper methods such as exponentially weighted average forecasters and other aggregation methods. Such methods achieve fast rates under conditions such as stochastic exp-concavity (Juditsky et al., 2008), which imply the "stochastic mixability" condition that, as explained by Van Erven et al. (2015), is sufficient for fast rates for aggregation methods. To get rates comparable to those of improper learners, we invariably need to make a "convex luckiness" assumption under which, as again shown by Van Erven et al. (2015), v-stochastic mixability implies the v-PPC condition (see also Figure 1); the latter allows for fast rates for randomized proper learners. An interesting question for future work is whether our proof techniques can be extended to incorporate, and get the right rates for, improper methods such as the empirical star estimator (Audibert, 2007) and Q-aggregation (Lecué and Rigollet, 2014). Since the original analysis of these methods bears some similarity to our techniques, this might very well be possible.

While superior rates for improper learners are inevitable, it is more worrying that the rate we showed for ERM in heavy-tailed bounded regression is worse than the rate for the SeqRand algorithm, which is also randomized proper (see Example 11 and text below it). We do not know whether the rate we obtain is the actual worst-case rate that ERM achieves under our conditions, or whether ERM achieves the same rate as SeqRand, or something in between. In the latter two cases, it would mean that our bounds are suboptimal. Sorting this out is a major goal for future work.

Empirical Process vs Information-theoretic Broadly speaking, one can distinguish approaches to proving excess risk bounds into two main groups: on the one hand are approaches based on empirical process theory (EPT) such as (Bartlett et al., 2005; Bartlett and Mendelson, 2006; Koltchinskii, 2006; Mendelson, 2014; Liang et al., 2015; Dinh et al., 2016) and most work involving VC dimension in classification. On the other hand are information-theoretic approaches based on prior measures, change-of-measure arguments, and KL penalties such as PAC-Bayesian and MDL approaches (Barron and Cover, 1991; Li, 1999; Catoni, 2003; Audibert, 2004; Grünwald, 2007; Audibert, 2009). A significant advantage of EPT approaches is that they often can achieve optimal rates of convergence for "large" models $\mathcal{F}$ with metric entropy $\log \mathcal{N}(\mathcal{F}, \|\cdot\|, \epsilon)$ that increases polynomially in $1/\epsilon$, where $\|\cdot\|$ is the $L_1(P)$ or $L_2(P)$-metric. Prior-based approaches (including the one in this paper) may yield suboptimal rates in such cases (see Audibert (2009) for discussion). A closely related advantage of EPT approaches is that they can handle empirical covers of $\mathcal{F}$, thus allowing one to prove bounds for VC classes, among others.

An advantage of prior-based approaches is that they inherently penalize, so that whenever one has a countably infinite union of classes $\mathcal{F} = \bigcup_{j \in \mathbb{N}} \mathcal{F}_j$, the approaches automatically adapt to the rate that can be obtained as if the best $\mathcal{F}_j$ containing f^* were known in advance; this adaptation was illustrated at various places in this paper (see final display in Proposition 6, equation (27)). This happens even if for every n, there is a j and $f \in \mathcal{F}_j$

with empirical error 0; in such a case unpenalized methods as often used in EPT methods would overfit. In the paper (Grünwald and Mehta, 2019), a companion paper to the present one, we show for bounded excess losses that the two approaches may be combined. In fact one can provide a single excess risk bound in which the information complexity is replaced by a strictly smaller quantity and instead of a prior one uses a more general "luckiness function" (Grünwald, 2007) that is better suited for dealing with penalized estimators. For some choices of luckiness function, one gets a slight strengthening of the excess risk bounds given in this paper; for other choices, one gets bounds in terms of Rademacher complexity, $L_2(P)$ and empirical $L_2(P_n)$ covering numbers. Thus, the best of both worlds is achievable, but for the time being only for bounded excess losses.

Another major goal for future work is thus to provide such a combined EPT-information theoretic bound for unbounded excess losses that allows for heavy-tailed excess loss. Within the EPT literature, some work has been done: Mendelson (2014, 2017b) provides bounds on the $L_2(P)$-estimation error $\|\hat{f} - f^*\|^2_{L_2(P)}$ and Liang et al. (2015) on the related squared loss risk. For other loss functions not much seems to be known: Mendelson (2017b) shows that improved $L_2(P)$-estimation error rates may be obtained by using other, proxy loss functions during training; however, the target remains $L_2(P)$-estimation. In contrast, our approach allows for general loss functions ℓ including density estimation, but we do not specially study proxy training losses.

The last three EPT-based works can deal with $(P, \ell, \mathcal{F})$ with unbounded excess (squared loss) risk. This is in contrast to earlier papers in the information-theoretic/PAC-Bayes tradition; as far as we know, our work is the first one that allows one to prove excess risk convergence rates in the unbounded risk case (Theorem 31) for general models including countable infinite unions of models as in Proposition 6. Previous works dealing with unbounded excess loss all rely on a Bernstein condition — we are aware of (Zhang, 2006a), requiring $\beta = 1$; (Audibert, 2004), for the transductive setting rather than our inductive setting; and, the most general, (Audibert, 2009). However, for convex or linear losses, a Bernstein condition can *never* hold if $\sup_{f \in \mathcal{F}} \mathbf{E}[L_f]$ is unbounded, as follows trivially from inspecting Definition 17, whereas the v-central and PPC-conditions *can* hold. See for instance Example 15 in Appendix I, where $\mathcal{F}$ is just the densities of the normal location family without any bounds on the mean: here the Bernstein condition must fail, yet the strong central condition and the witness condition both hold and thus Theorem 31 applies (for some moderate M).

In the unbounded-excess-loss-yet-bounded-risk case, the difference between these works and ours opaques: there may well be cases (though we have not produced one) where the Bernstein condition holds for some β but the v-PPC condition does not hold for $v(\epsilon) \asymp \epsilon^{1-\beta}$; the opposite certainly can happen (note however that in the bounded excess loss case these two conditions are equivalent; see Figure 1). Indeed, Example 14 in Appendix I exhibits an $\mathcal{F}$ for which the excess risk is bounded but its second moment is not, whence the Bernstein condition fails to hold for *any* positive exponent, while both the strong central condition and the witness condition hold. Theorem 29 therefore applies whereas the results of Audibert (2009) and Zhang (2006b) do not. Finally we note that Audibert (2009) proves his bounds for his ingenious SeqRand learning algorithm, whereas Zhang's and our bounds hold for general estimators.

Yet another major goal for current work is thus to disentangle the role of the PPC condition and the Bernstein condition for unbounded excess losses; ideally we would extend our bounds to cover faster rates under a weaker condition implied by either of the Bernstein or PPC conditions.

Additional Future Work: Learning η A general issue with generalized Bayesian and MDL methods, but one that is avoided by ERM, is the fact that they depend on the learning rate parameter η. While this is often pragmatically resolved by cross-validation (see e.g. Audibert (2009) and many others), Grünwald (2011, 2012) give a method for learning η that provably finds the "right" η (i.e. optimal for the best Bernstein condition that holds for the given learning problem) for bounded excess loss functions and likelihood ratios; experiments (Grünwald and Van Ommen, 2017) indicate that this "safe Bayesian" method works excellently in the unbounded case as well. While it seems that the proof technique to handle learning η carries over to the present unbounded setting, actually proving that the SafeBayes method still works remains a task for future work.

Acknowledgments

We would like to thank Bob Williamson (who saw the use of bringing structure in the various existing 'easiness conditions'), Tim van Erven (who brought our attention to the fact that GGV's entropy condition is not needed whenever $\eta < 1$, as earlier noted by Tong Zhang) and Bas Kleijn (who brought Wong and Shen (1995) to our attention), Andrew Barron (for various discussions), and Alice Kirichenko (for pointing out numerous small mistakes). An anonymous referee made some highly useful suggestions. This research was supported by the Netherlands Organization for Scientific Research (NWO) VICI Project Nr. 639.073.04.

Appendix A. Proofs for Section 3

Proof (of Proposition 1) First, we prove (a), i.e.,

$$\lim_{\eta \downarrow 0} -\frac{1}{\eta} \log \mathbf{E}[e^{-\eta X}] = \lim_{\eta \downarrow 0} \frac{1}{\eta} \left(1 - \mathbf{E}[e^{-\eta X}]\right) = \mathbf{E}[X].$$

Define $y_\eta := \mathbf{E}[e^{-\eta X}]$; we will use the fact that $\lim_{\eta \downarrow 0} \mathbf{E}[e^{-\eta X}] = 1$ (from Fatou's Lemma, using the nonnegativity of $e^{-\eta x}$).

Now, from Lemma 2 of Van Erven and Harremoës (2014), for $y \geq \frac{1}{2}$ we have $(y - 1)\left(1 + \frac{1-y}{2}\right) \leq \log y \leq y - 1$. Hence,

$$\lim_{\eta \downarrow 0} -\frac{1}{\eta} \log \mathbf{E}[e^{\eta X}] = \lim_{\eta \downarrow 0} -\frac{1}{\eta} \log y_\eta = \lim_{\eta \downarrow 0} -\frac{1}{\eta} (y_\eta - 1) = \lim_{\eta \downarrow 0} \frac{1}{\eta} \mathbf{E}[1 - e^{-\eta X}],$$

which completes the proof of the first equality.

Now, for all x the function $\eta \to \frac{1}{\eta}(1 - e^{-\eta x})$ is non-increasing, as may be verified since $\operatorname{sign}(x e^{-\eta x} - \frac{1 - e^{-\eta x}}{\eta}) = -\operatorname{sign}(e^{\eta x} - (\eta x + 1)) \leq 0$.

Next, we rewrite the following Hellinger-divergence-like quantity:

$$\mathbf{E}\left[\frac{1}{\alpha\bar{\eta}}\left(1-e^{-\alpha\bar{\eta}X}\right)\right]=\mathbf{E}\left[\frac{1}{\alpha\bar{\eta}}\left(1-e^{-\alpha\bar{\eta}X}\right)-\frac{1}{\bar{\eta}}(1-e^{-\bar{\eta}X})\right]+\frac{1}{\bar{\eta}}\mathbf{E}\left[1-e^{-\bar{\eta}X}\right].$$

Now take any decreasing sequence $\alpha=\alpha_j\in(\alpha_i)_{i\geq 1}$ going to zero with $\alpha_1<1$. We have for all j that $x\mapsto\frac{1}{\alpha_j\bar{\eta}}(1-e^{-\alpha_j\bar{\eta}x})-\frac{1}{\bar{\eta}}(1-e^{-\bar{\eta}x})$ is a positive function, and the corresponding sequence with respect to j is non-decreasing. Hence, the monotone convergence theorem applies and we may interchange the limit and expectation, yielding

$$\begin{aligned}
&\lim_{\alpha\downarrow 0}\mathbf{E}\left[\frac{1}{\alpha\bar{\eta}}\left(1-e^{-\alpha\bar{\eta}X}\right)-\frac{1}{\bar{\eta}}(1-e^{-\bar{\eta}X})\right]+\frac{1}{\bar{\eta}}\mathbf{E}\left[1-e^{-\bar{\eta}X}\right]\\
&=\mathbf{E}\left[\lim_{\alpha\downarrow 0}\frac{1}{\alpha\bar{\eta}}\left(1-e^{-\alpha\bar{\eta}X}\right)-\frac{1}{\bar{\eta}}(1-e^{-\bar{\eta}X})\right]+\frac{1}{\bar{\eta}}\mathbf{E}\left[1-e^{-\bar{\eta}X}\right]\\
&=\mathbf{E}\left[\lim_{\eta\downarrow 0}\frac{1-e^{-\eta X}}{\eta}\right]=\mathbf{E}\left[\frac{\lim_{\eta\downarrow 0}Xe^{-\eta X}}{1}\right]=\mathbf{E}[X],
\end{aligned}$$

where the penultimate equality follows from L'Hôpital's rule. This concludes the proof of the second part of (a). Next, we show (b). Observe that for any $\eta'\leq\eta$, the concavity of $x\mapsto x^{\eta'/\eta}$ together with Jensen's inequality implies that

$$-\frac{1}{\eta'}\log\mathbf{E}\left[e^{-\eta' X}\right]=-\frac{1}{\eta'}\log\mathbf{E}\left[\left(e^{-\eta X}\right)^{\eta'/\eta}\right]\geq-\frac{1}{\eta'}\log\left(\mathbf{E}\left[e^{-\eta X}\right]\right)^{\eta'/\eta}=-\frac{1}{\eta}\log\mathbf{E}\left[e^{-\eta X}\right].$$

■

A.1. Proof of Lemma 33, Extending Lemma 5

We begin with an extension of Lemma 5. This more general result will be used in the proof of Theorem 29. It generalizes Lemma 5 in that it allows general comparators $\phi(f)$, which depend on the f being compared, instead of just the risk-minimizing f^* (and it continues to hold even if $\mathcal{F}$ does not contain an optimal f^*). Formally, let $(P,\ell,\mathcal{F})$ be a learning problem. For $f\in\mathcal{F}$, we work with the excess loss $\ell_f-\ell_{\phi(f)}$, where $\phi:\mathcal{F}\to\bar{\mathcal{F}}$ is a *comparator map*[6] which, in the special case of Lemma 5, is simply the trivial function mapping each $f\in\mathcal{F}$ to f^*.

Lemma 33 *Let $(P,\ell,\mathcal{F})$ represent a learning problem. Let $\Pi_|$ be a learning algorithm for this learning problem that outputs distributions on $\mathcal{F}$. Let $\phi:\mathcal{F}\to\bar{\mathcal{F}}$ be any deterministic function mapping the predictor $\underline{f}\sim\Pi_n$ to a set of nontrivial comparators. Then for all $\eta>0$, we have:*

$$\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\mathbf{E}^{\mathrm{ANN}(\eta)}_{Z\sim P}\left[\ell_{\underline{f}}-\ell_{\phi(\underline{f})}\right]\right]\trianglelefteq_{\eta\cdot n}\mathrm{IC}_{n,\eta}\left(\phi(\underline{f})\parallel\Pi_|\right).\tag{57}$$

where IC_η is the (generalized) information complexity, defined as

$$\mathrm{IC}_{n,\eta}\left(\phi(\underline{f})\parallel\Pi_|\right):=\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\ell_{\underline{f}}(Z_i)-\ell_{\phi(\underline{f})}(Z_i)\right)\right]+\frac{\mathrm{KL}(\Pi_n\parallel\Pi_0)}{\eta\cdot n}.\tag{58}$$

[6] The set $\bar{\mathcal{F}}$ is defined at the beginning of Section 6.

By the finiteness considerations of Appendix H, $\mathrm{IC}_{n,\eta}(\phi(\underline{f}) \,\|\, \Pi_{|})$ is always well-defined but may in some cases be equal to $-\infty$ or ∞. The explicit use above of a comparator function ϕ differs from Zhang's statement, in which the ability to use such a mapping was left quite implicit; however, inspection of the proof of Theorem 2.1 of Zhang (2006b) reveals that our version above with comparator functions is also true. Comparator functions will be critical to our application of Lemma 33. For completeness, we provide a proof of this generalized result.

Proof (of Lemma 33) For any measurable function $\psi : \mathcal{F} \times \mathcal{Z}^n \to \mathbb{R}$ it holds that

$$\mathbf{E}_{f\sim\Pi_n}[\psi(f,Z^n)] - \mathrm{KL}(\Pi_n \,\|\, \Pi_0) \le \log \mathbf{E}_{f\sim\Pi_0}\left[e^{\psi(f,Z^n)}\right]. \tag{59}$$

This result, a variation of the "Donsker-Varadhan variational bound" follows from convex duality; see Zhang (2006b) for an explicit proof.

Define the function $R_n\colon \mathcal{F} \times \mathcal{Z}^n \to \mathbb{R}$ as $R_n(f,z^n) = \sum_{j=1}^n \left(\ell_f(z_j) - \ell_{\phi(f)}(z_j)\right)$. Then (59) with the choice $\psi(f,Z^n) = -\eta R_n(f,Z^n) - \log \mathbf{E}_{\bar{Z}^n\sim P^n}\left[e^{-\eta R_n(f,\bar{Z}^n)}\right]$ yields

$$\mathbf{E}_{f\sim\Pi_n}\left[-\eta R_n(f,Z^n) - \log \mathbf{E}_{\bar{Z}^n}\left[e^{-\eta R_n(f,\bar{Z}^n)}\right]\right] - \mathrm{KL}(\Pi_n \,\|\, \Pi_0) \le \log \mathbf{E}_{f\sim\Pi_0}\left[\frac{e^{-\eta R_n(f,Z^n)}}{\mathbf{E}_{\bar{Z}^n}\left[e^{-\eta R_n(f,\bar{Z}^n)}\right]}\right],$$

which, after exponentiating and taking the expectation with respect to $Z^n \sim P^n$, gives

$$\begin{aligned}
&\mathbf{E}_{Z_n}\left[\exp\left(\mathbf{E}_{f\sim\Pi_n}\left[-\eta R_n(f,Z^n) - \log \mathbf{E}_{\bar{Z}^n}\left[e^{-\eta R_n(f,\bar{Z}^n)}\right]\right] - \mathrm{KL}(\Pi_n \,\|\, \Pi_0)\right)\right] \\
&\le \mathbf{E}_{Z^n}\left[\mathbf{E}_{f\sim\Pi_0}\left[\frac{e^{-\eta R_n(f,Z^n)}}{\mathbf{E}_{\bar{Z}^n}\left[e^{-\eta R_n(f,\bar{Z}^n)}\right]}\right]\right].
\end{aligned}$$

From the Tonelli-Fubini theorem (see e.g. (Dudley, 2002, p. 137)), we can exchange the two outermost expectations on the RHS, and so the RHS is at most 1. Using ESI notation, we then have

$$\mathbf{E}_{f\sim\Pi_n}\left[-\log \mathbf{E}_{\bar{Z}^n}\left[e^{-\eta R_n(f,\bar{Z}^n)}\right]\right] \trianglelefteq_1 \mathbf{E}_{f\sim\Pi_n}\left[\eta R_n(f,Z^n)\right] + \mathrm{KL}(\Pi_n \,\|\, \Pi_0).$$

Using that the $\bar{Z}_1,\ldots,\bar{Z}_n$ are drawn i.i.d. from P and dividing by $\eta \cdot n$ then yields

$$\mathbf{E}_{f\sim\Pi_n}\left[-\frac{1}{\eta}\log \mathbf{E}_Z\left[e^{-\eta(\ell_f(Z)-\ell_{\phi(f)}(Z))}\right]\right] \trianglelefteq_{\eta\cdot n} \mathbf{E}_{f\sim\Pi_n}\left[\frac{1}{n}\sum_{j=1}^n\left(\ell_f(Z_j) - \ell_{\phi(f)}(Z_j)\right)\right] + \frac{1}{\eta}\mathrm{KL}(\Pi_n \,\|\, \Pi_0).$$

■

Proof (of Proposition 6) Zhang (2006a) showed the first inequality in (17) and (20). The equality of the first and third terms and the inequality in (17) are "folklore" in the individual sequence-prediction and MDL communities. For completness we provide a proof.

The two equalities in (17) are easy to see after rewriting the center term as

$$n \cdot \inf_{\Pi_{|}\in\mathrm{RAND}} \mathrm{IC}_{n,\eta}(\Pi_{|}) = -\frac{1}{\eta} \sup_{\Pi\in\Delta(\mathcal{F})}\left\{-\sum_{j=1}^n L_f(Z_j) - \mathrm{KL}(\Pi \,\|\, \Pi_0)\right\}.$$

Now, from Legendre duality, we have for some map $\varphi : \mathcal{X} \to \mathbb{R}$ that

$$\sup_{\nu \in \Delta(\mathcal{X})} \left\{ \mathbf{E}_{X \sim \nu}[\varphi(X)] - \mathrm{KL}(\nu \,\|\, \mu) \right\} = \log \mathbf{E}_{X \sim \mu}\left[e^{\varphi(X)} \right],$$

and the supremum is achieved by taking $\nu(dx) = \frac{e^{\varphi}(dx)}{\mathbf{E}_{X \sim \mu}\left[e^{\varphi(X)}\right]}$. This proves the equalities in (17).

To see (18) and (19), observe that for any $A \subset \mathcal{F}$, we have

$$\begin{aligned}
-\frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim \Pi_0}\left[e^{-\sum_{j=1}^n L_{\underline{f}}(Z_j)} \right] &= -\frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim \Pi_0}\left[\left(\mathbf{1}_{\{\underline{f} \in A\}} + \mathbf{1}_{\{\underline{f} \notin A\}} \right) e^{-\sum_{j=1}^n L_{\underline{f}}(Z_j)} \right] \\
&\leq -\frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim \Pi_0}\left[\mathbf{1}_{\{\underline{f} \in A\}} \cdot e^{-\sum_{j=1}^n L_{\underline{f}}(Z_j)} \right] \\
&= -\frac{1}{\eta} \log \Pi_0(A) - \frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim \Pi_0 | A}\left[e^{-\sum_{j=1}^n L_{\underline{f}}(Z_j)} \right] \\
&\leq -\frac{1}{\eta} \log \Pi_0(A) + \mathbf{E}_{\underline{f} \sim \Pi_0 | A}\left[\sum_{j=1}^n L_{\underline{f}}(Z_j) \right],
\end{aligned}$$

where the last line follows from Jensen's inequality. Together with the second equality in the already-established (17), the third line implies (18); the last line implies (19).

For (20), the first inequality is obvious since the infimum over DET is at least the infimum over RAND. The equality is immediate from the definition of the two-part MDL estimator. The second inequality follows as a special case of the inequality in (17). ■

Appendix B. Proofs for Section 4

Proof (of Theorem 10) The Rényi divergence (Van Erven and Harremoës, 2014) of order α is defined as $D_\alpha(p\|q) = \frac{1}{\alpha-1} \log \int p^\alpha q^{1-\alpha} d\mu$, so that, for $0 < \alpha < 1$, with $\eta = (1-\alpha)\bar{\eta}$,

$$\begin{aligned}
D_\alpha(p_{f^*,\bar{\eta}} \| p_{f,\bar{\eta}}) &= \frac{1}{\alpha - 1} \log \int p(z) \frac{e^{-\alpha \bar{\eta} L_{f^*}} \cdot e^{-(1-\alpha)\bar{\eta} L_f}}{(\mathbf{E}[e^{-\bar{\eta} L_{f^*}(Z)}])^\alpha (\mathbf{E}[e^{-\bar{\eta} L_f(Z)}])^{1-\alpha}} d\mu \\
&= \frac{1}{\alpha - 1} \log \int p(z) \frac{e^{-(1-\alpha)\bar{\eta} L_f}}{(\mathbf{E}[e^{-\bar{\eta} L_f(Z)}])^{1-\alpha}} d\mu \\
&= -\frac{\bar{\eta}}{\eta} \left(\log \mathbf{E}[e^{-\eta L_f}] - \frac{\eta}{\bar{\eta}} \log \mathbf{E}[e^{-\bar{\eta} L_f(Z)}] \right) \\
&= \bar{\eta} \mathbf{E}^{\mathrm{ANN}(\eta)}[L_f] + \log \mathbf{E}[e^{-\bar{\eta} L_f(Z)}] \\
&\leq \bar{\eta} \mathbf{E}^{\mathrm{ANN}(\eta)}[L_f],
\end{aligned}$$

where we used the $\bar{\eta}$-central condition. Van Erven and Harremoës (2014) show that the squared Hellinger distance between two densities p and q is always bounded by their Rényi divergence of order 1/2 and also that the latter is bounded by the Rényi divergence of order $0 < \alpha < 1/2$ via $D_{1/2}(p\|q) \leq \frac{1-\alpha}{\alpha} D_\alpha(p\|q)$, so that we get

$$d_{\bar{\eta}}^2(f, f') \leq \frac{1}{\bar{\eta}} \cdot \frac{1-\alpha}{\alpha} \cdot \bar{\eta} \mathbf{E}^{\mathrm{ANN}(\eta)}[L_f] = \frac{\eta}{\bar{\eta} - \eta} \mathbf{E}^{\mathrm{ANN}(\eta)}[L_f].$$

The result is now immediate from Lemma 5. ■

Proof (of Proposition 11, Cont.) We use the familiar rewrite of the KL divergence $\mathbf{E}_{Z\sim P_{f^*}}[L_f] = D(f^*\|f)$ as $\mathbf{E}_{Z\sim P_{f^*}}[L_f] = \mathbf{E}[L_f + S]$, with $S = (p_f(Z)/p_{f^*}(Z)) - 1$, where as is well-known, $L_f + S$ is nonnegative on $\mathcal{Z}$. Using this in the second inequality below gives:

$$\begin{aligned}\mathbf{E}_{Z\sim P_{f^*}}[L_f \vee 0] &= \mathbf{E}_{Z\sim P_{f^*}}[\mathbf{1}_{\{L_f\geq 0\}}(L_f + S)] - \mathbf{E}_{Z\sim P_{f^*}}[\mathbf{1}_{\{L_f\geq 0\}}\, S] \leq \mathbf{E}_{Z\sim P_{f^*}}[L_f] + \mathbf{E}_{Z\sim P_{f^*}}[|S|] \\ &= \mathbf{E}_{Z\sim P_{f^*}}[L_f] + \int p_{f^*}\left|\frac{p_f - p_{f^*}}{p_{f^*}}\right| d\mu(z) \leq D(f^*\|f) + \int \left|p_f - p_{f^*}\right| d\mu,\end{aligned}$$

and the result follows by Pinsker's inequality. ■

Appendix C. Proofs for Section 5 and Example 12

C.1. Proof of Lemma 34, Extending Lemma 13

Below we state and prove Lemma 34 which generalizes Lemma 13 in the main text in that it allows general comparators $\phi(f)$, as introduced above Lemma 33. This extension is pivotal for our results in Section 6 involving the GRIP.

Lemma 34 *Let $\bar\eta > 0$. Let ϕ be any comparator map ϕ such that for any given f, $\phi(f)$ satisfies $\mathbf{E}[\ell_{\phi(f)}] \leq \mathbf{E}[\ell_f]$. Assume that the strong $\bar\eta$-central condition is satisfied with respect to comparator ϕ for some fixed $f \in \mathcal{F}$, i.e.,*

$$\ell_f - \ell_{\phi(f)} \trianglelefteq_{\bar\eta} 0. \tag{60}$$

Furthermore assume that the (u, c)-witness condition holds for this f, relative to $\phi(f)$, for some constants $u > 0$ and $c \in (0, 1]$, i.e.,

$$c\,\mathbf{E}[L_f] \leq \mathbf{E}[(\ell_f - \ell_{\phi(f)}) \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} \leq u\}}]. \tag{61}$$

Then for all $\eta \in (0, \bar\eta)$

$$\mathbf{E}[L_f] \;\leq\; c_u \cdot \mathbf{E}^{\mathrm{HE}(\eta)}\left[\ell_f - \ell_{\phi(f)}\right] \;\leq\; c_u \cdot \mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_f - \ell_{\phi(f)}\right], \tag{62}$$

with $c_u := \frac{1}{c}\frac{\eta u+1}{1-\frac{\eta}{\bar\eta}}$. Moreover, suppose that the (τ, c)-witness condition holds for a non-increasing τ and c as in Definition 12, for all $f \in \mathcal{F}$, relative to comparator $\phi(\cdot)$, i.e., $\mathbf{E}\big[(\ell_f - \ell_{\phi(f)}) \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} \leq \tau(\mathbf{E}[\ell_f - \ell_{\phi(f)}])\}}\big] \geq c\,\mathbf{E}[L_f]$. For all $f \in \mathcal{F}$, all $\eta \in (0, \bar\eta)$, all $\epsilon > 0$, we have:

$$\mathbf{E}[L_f] \;\leq\; \epsilon \vee c_{\tau(\epsilon)} \cdot \mathbf{E}^{\mathrm{HE}(\eta)}\left[\ell_f - \ell_{\phi(f)}\right] \;\leq\; \epsilon \vee c_{\tau(\epsilon)} \cdot \mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_f - \ell_{\phi(f)}\right]. \tag{63}$$

Proof
Proof of (62). Define $L'_f := \ell_f - \ell_{\phi(f)}$. For any $\eta \in [0, \bar\eta]$, define:

$$h_{f,\eta} := \frac{1}{\eta}\left(1 - e^{-\eta L'_f}\right) \qquad S_{f,\eta} := h_{f,\eta} - h_{f,\bar\eta} \qquad H_{f,\eta} := \mathbf{E}^{\mathrm{HE}(\eta)}[L'_f] = \mathbf{E}[h_{f,\eta}].$$

It is easy to verify that the map $\eta \mapsto h_{f,\eta}$ is non-increasing, and hence $S_{f,\eta}$ is a positive random variable for any $\eta \in [0, \bar\eta]$. It also is easy to verify that $\lim_{\eta\downarrow 0} h_{f,\eta} = L'_f$. We thus can define $h_{f,0} = L'_f$ and $S_{f,0} = L'_f - h_{f,\bar{\eta}}$ and hence can rewrite the excess risk of f (with respect to $\phi(f)$) as

$$\mathbf{E}[L'_f] = \mathbf{E}[h_{f,0} - h_{f,\bar{\eta}} + h_{f,\bar{\eta}}] = \mathbf{E}[S_{f,0}] + H_{f,\bar{\eta}}.$$

Splitting up the expectation into two components, we have

$$\mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{L'_f \le u\}}] + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{L'_f > u\}}] + H_{f,\bar{\eta}}.$$

Now, from Lemma 35 (stated and proved immediately after this proof), the positivity of $S_{f,\eta}$, and using $\bar{C} \coloneqq C_{\bar{\eta},\eta,u}$ to avoid cluttering notation, we have

$$\begin{aligned}\mathbf{E}[L'_f] \le \bar{C}\,\mathbf{E}[S_{f,\eta} \cdot \mathbf{1}_{\{L'_f \le u\}}] + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{L'_f > u\}}] + H_{f,\bar{\eta}} \le \bar{C}\,\mathbf{E}[S_{f,\eta}] + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{L'_f > u\}}] + H_{f,\bar{\eta}} \\ = \bar{C}\left(H_{f,\eta} - H_{f,\bar{\eta}}\right) + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{L'_f > u\}}] + H_{f,\bar{\eta}} = \bar{C} H_{f,\eta} - (\bar{C} - 1) H_{f,\bar{\eta}} + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{L'_f > u\}}].\end{aligned}$$

We observe that $H_{f,\bar{\eta}} \ge 0$ since $H_{f,\bar{\eta}} = \frac{1}{\bar{\eta}} \mathbf{E}\left[1 - e^{-\bar{\eta} L'_f}\right] \ge 0$, where the inequality is implied by the strong $\bar{\eta}$-central condition (i.e. $\mathbf{E}\left[e^{-\bar{\eta} L'_f}\right] \le 1$). Therefore, since it always holds that $\bar{C} \ge 1$ we have

$$\mathbf{E}[L'_f] \le \bar{C} H_{f,\eta} + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{L'_f > u\}}]. \tag{64}$$

Next, we claim that $\mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{L'_f > u\}}] \le \mathbf{E}[L'_f \cdot \mathbf{1}_{\{L'_f > u\}}]$. To see this, observe that $S_{f,0} = L'_f + \frac{1}{\bar{\eta}}\left(e^{-\bar{\eta} L'_f} - 1\right)$, and that the second term is negative on the event $L'_f > u$. We thus have

$$\mathbf{E}[L'_f] - \mathbf{E}[L'_f \cdot \mathbf{1}_{\{L'_f > u\}}] \le \bar{C} H_{f,\eta},$$

which can be rewritten as

$$\mathbf{E}[L'_f \cdot \mathbf{1}_{\{L'_f \le u\}}] \le \bar{C} H_{f,\eta}, \tag{65}$$

Now, since we assume (61), the first inequality in (62) is proved, and the second then follows from (12):

$$\mathbf{E}[L_f] \le \frac{\bar{C}}{c} H_{f,\eta}.$$

Proof of (63). Fix arbitrary $f \in \mathcal{F}$. We know that for this particular f, either $\mathbf{E}[L_f] \le \epsilon$ in which case there is nothing to prove, or $\mathbf{E}[L_f] > \epsilon$. Then for this f, the (u, c)-witness condition holds with $u = \tau(\mathbf{E}[L_f]) \le \tau(\epsilon)$. But then the result follows as above. ■

Lemma 35 ("Bounded Part" Lemma) *For $u, \bar{\eta} > 0$ and $\eta \in [0, \bar{\eta})$, we have*

$$\mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} \le u\}}] \le C_{\bar{\eta},\eta,u}\,\mathbf{E}[S_{f,\eta} \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} \le u\}}],$$

where $C_{\bar{\eta},\eta,u} \coloneqq \frac{\eta u + 1}{1 - \frac{\eta}{\bar{\eta}}}$.

Proof It is sufficient to show that on the set $\{\ell_f - \ell_{\phi(f)} \le u\}$, it holds that $S_{f,0} \le C S_{f,\eta}$ for some constant C. This may be rewritten as wanting to show, for $\eta_0 \to 0$:

$$\frac{1}{\eta_0}(1 - e^{-\eta_0(\ell_f - \ell_{\phi(f)})}) - \frac{1}{\bar{\eta}}(1 - e^{-\bar{\eta}(\ell_f - \ell_{\phi(f)})}) \le C\left(\frac{1}{\eta}(1 - e^{-\eta(\ell_f - \ell_{\phi(f)})}) - \frac{1}{\bar{\eta}}(1 - e^{-\bar{\eta}(\ell_f - \ell_{\phi(f)})})\right).$$

Letting $r = e^{-\bar{\eta}(\ell_f - \ell_{\phi(f)})}$, this is equivalent to showing that

$$\frac{1}{\bar{\eta}}\left(\frac{1}{\eta_0/\bar{\eta}}(1 - r^{\eta_0/\bar{\eta}}) - (1 - r)\right) \le \frac{C}{\bar{\eta}}\left(\frac{1}{\eta/\bar{\eta}}(1 - r^{\eta/\bar{\eta}}) - (1 - r)\right).$$

Now, for any $\eta \ge 0$, define[7] the function g_η as $g_\eta(r) = \frac{1}{\eta}(1 - r^\eta) - (1 - r)$. From Lemma 36, for any $\eta' \ge 0$, if $r \ge \frac{1}{V}$ for some $V > 1$ then $g_0(r) \le \frac{1}{1-\eta'}(\eta' \log V + 1) g_{\eta'}(r)$.

Applying this inequality, taking $\eta_0 \to 0$ and $\eta' := \frac{\eta}{\bar{\eta}}$, and observing that on the set $\{\ell_f - \ell_{\phi(f)} \le u\}$ we may take $V = e^{\bar{\eta} u} > 1$, we see that whenever $\ell_f - \ell_{\phi(f)} \le u$,

$$\left(\frac{1}{\eta_0}(1 - r^{\eta_0}) - (1 - r)\right) \le \frac{1}{1 - \eta'}(\eta' \bar{\eta} u + 1)\left(\frac{1}{\eta'}(1 - r^{\eta'}) - (1 - r)\right).$$

Thus, $S_{f,0} \le C_{\bar{\eta},\eta,u} S_{f,\eta}$ indeed holds for $C_{\bar{\eta},\eta,u} = \frac{\eta u + 1}{1 - \frac{\eta}{\bar{\eta}}}$. ■

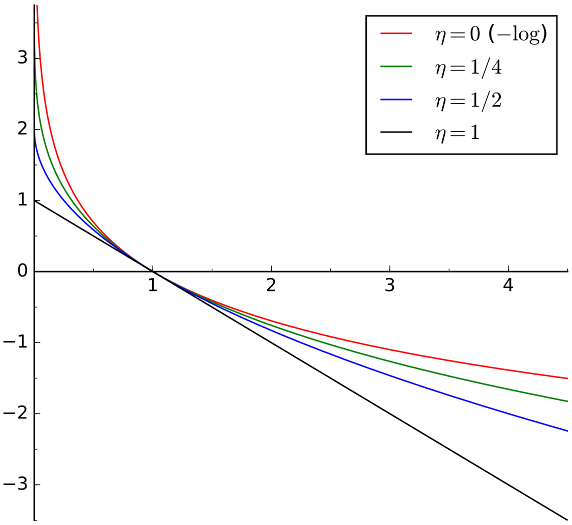

Figure 2: The function $r :\to \eta^{-1}(1 - r^\eta)$ for various values of r. $g_\eta(r)$ is the difference of the line for η at r and the line for $\eta = 1$ at r, which is always positive.

Lemma 36 *Let $0 \le \eta' < \eta < 1$ and $1 < V < \infty$. Define $g_\eta(r) := \eta^{-1}(1 - r^\eta) - (1 - r)$, a positive function. Then for $\eta' > 0$ and $r \ge \frac{1}{V}$:*

$$g_{\eta'}(r) \le C_{\eta' \leftarrow \eta}(V) g_\eta(r),$$

[7] Note that the g_η used here is *not* a GRIP.

where $C_{\eta'\leftarrow\eta}(V) \le ((\eta')^{-1}-1)/(\eta^{-1}-1)$, *and*

$$\lim_{\eta'\downarrow 0} g_{\eta'}(r) \le C_{0\leftarrow\eta}(V) g_\eta(r),$$

where $C_{0\leftarrow\eta}(V) = \frac{\log V-(1-V^{-1})}{\frac{1}{\eta}(1-V^{-\eta})-(1-V^{-1})} \le \frac{\eta}{1-\eta}\log V + \frac{1}{1-\eta}$.

Proof Let $0 \le \eta' < \eta$. We will prove that, for all $r \ge \frac{1}{V}$, we have $g_{\eta'}(r) \le C \cdot g_\eta(r)$ for some constant C. Hence it suffices to bound

$$h_{\eta',\eta}(r) := \frac{g_{\eta'}(r)}{g_\eta(r)} = \frac{(\eta')^{-1}(1-r^{\eta'})-(1-r)}{\eta^{-1}(1-r^\eta)-(1-r)}.$$

We can extend the definition of this function to $\eta' = 0$ and $r = 1$ so that it becomes well-defined for all $r > 0$, $0 \le \eta' < \eta < 1$: $(0)^{-1}(1-r^0)$ is defined as $\lim_{\eta'\downarrow 0}(\eta')^{-1}(1-r^{\eta'}) = -\log r$. $h_{\eta',\eta}(1)$ is set to $\lim_{r\uparrow 1} h_{\eta',\eta}(r) = \lim_{r\downarrow 1} h_{\eta',\eta}(r)$ which is calculated using L'Hôpital's rule twice, together with the fact that for $0 \le \eta \le 1$ (note $\eta = 0$ is allowed), $g'_\eta(r) = -r^{\eta-1}+1$, $g''_\eta(r) = (1-\eta)r^{\eta-2}$. Then, because $g_\eta(1) = g_0(1) = g'_\eta(1) = g'_0(1) = 0$, we get:

$$h_{\eta',\eta}(1) := \lim_{r\downarrow 1} g_{\eta'}(r)/g_\eta(r) = \lim_{r\downarrow 1} g'_{\eta'}(r)/g'_\eta(r) = \lim_{r\downarrow 1} g''_{\eta'}(r)/g''_\eta(r) = \frac{1-\eta'}{1-\eta}.$$

We have $\lim_{r\to\infty} h_{\eta',\eta}(r) = 1$, and we show below that $h_{\eta',\eta}(r)$ is strictly decreasing in r for each $0 \le \eta' < \eta < 1$, so the maximum value is achieved for the minimum $r = 1/V$. We have $h_{\eta',\eta}(1/V) \le h_{\eta',\eta}(0) = (\eta'^{-1}-1)/(\eta^{-1}-1)$ and $h_{0,\eta}(1/V) = (\log V - (1-V^{-1}))/(\eta^{-1}(1-V^{-\eta})-(1-V^{-1}))$. The result follows by defining $C_{\eta'\leftarrow\eta}(V) = h_{\eta',\eta}(1/V)$. It only remains to show that $h_{\eta',\eta}(r)$ is decreasing in r and that the upper bound on $C_{0\leftarrow\eta}(V)$ stated in the lemma holds.

Proof that h is decreasing: The derivative of $h \equiv h_{\eta',\eta}$ for fixed $0 \le \eta' < \eta < 1$ is given by $h'_{\eta',\eta}(r) = r^{-1} \cdot s(r)$, where

$$s(r) = \frac{(-r^{\eta'}+r)\cdot g_\eta(r) + (r^\eta - r)\cdot g_{\eta'}(r)}{g_\eta(r)^2}. \tag{66}$$

Although we tried hard, we found neither a direct argument that $h' \le 0$ or that $h'' > 0$ (which would also imply the result in a straightforward manner). We resolve the issue by relating h to a function f which is easier to analyze. (66) shows that for $r > 0, r \ne 1$, $h'(r) = 0$, i.e., h reaches an extremum, iff $s(r) = 0$, i.e., iff the numerator in (66) is 0, i.e., iff $\frac{g_{\eta'}(r)}{g_\eta(r)} = \frac{r^{\eta'}-r}{r^\eta - r}$, i.e., iff

$$h(r) = f(r), \quad \text{where } f(r) := \frac{r^{\eta'-1}-1}{r^{\eta-1}-1}.$$

We can extend f to its discontinuity point $r = 1$ by using L'Hôpital's rule similar to its use above, and then we find that $f(1) = h(1)$; similarly, we find that the discontinuities of $f'(r)$ and $h'(r)$ at $r = 1$ are also removable, again by aggressively using L'Hôpital, which gives

$$f'(1) = \frac{1}{2}\cdot\frac{1-\eta'}{1-\eta}\left(\eta'-\eta\right)\ ,\ h'(1) = \frac{1}{3}\cdot\frac{1-\eta'}{1-\eta}\left(\eta'-\eta\right), \tag{67}$$

and we note that both derivatives are < 0 and also that there is $L < 1, R > 1$ such that

$$h < f \text{ on } (L, 1) \quad ; \quad h > f \text{ on } (1, R). \tag{68}$$

Below we show that f is strictly decreasing on $(0, \infty)$. But then h cannot have an extremum on $(0, 1)$; for if it had, there would be a point $0 < r_0 < 1$ with $h'(r_0) = 0$ and therefore $h(r_0) = f(r_0)$, so that, since $f'(r_0) < 0$, h lies under f in an open interval to the left of r_0 and above f to the right of r_0. But by (68), this means that there is another point r_1 with $r_0 < r_1 < 1$ at which h and f intersect such that h lies *above* f directly to the left of r_1. But we already showed that at any intersection, in particular at r_1, $h'(r_1) = 0$. Since $f'(r_1) < 0$, this implies that h must lie *below* f directly to the left of r_1, and we have reached a contradiction. It follows that h has no extrema on $(0, 1)$; entirely analogously, one shows that h cannot have any extrema on $(1, \infty)$. By (67), $h'(r)$ is negative in an open interval containing 1, so it follows that h is decreasing on $(0, \infty)$.

It thus only remains to be shown that f is strictly decreasing on $(0, \infty)$. To this end we consider a monotonic variable transformation, setting $y = r^{\eta - 1}$ so that $r^{\eta' - 1} = y^{(1-\eta')/(1-\eta)}$ and, for $a > 1$, define $f_a(y) = (y^a - 1)/(y - 1)$. Note that with $a = (1 - \eta')/(1 - \eta)$, $f_a(r^{\eta - 1}) = f(r)$. Since $0 < \eta < 1$, y is strictly decreasing in r, so it is sufficient to prove that, for all a corresponding to some choice of $0 \le \eta' < \eta < 1$, i.e., for all $a > 1$, f_a is strictly increasing on $y > 0$. Differentiation with respect to y gives that f_a is strictly increasing on interval (a, b) if, for all $y \in (a, b)$,

$$u_a(y) \equiv a y^a - y^a + 1 - a y^{a-1} > 0.$$

Straightforward differentiation and simplification gives that $u'_a(y) = a y^{a-1}(a - 1)(1 - y^{-1})$ which is strictly negative for all $y < 1$ and strictly positive for $y > 1$. Since trivially, $u_a(1) = 0$, it follows that $u_a(y) > 0$ on $(0, 1)$ and $u_a(y) > 0$ on $(1, \infty)$, so that f_a is strictly increasing on $(0, 1)$ and on $(1, \infty)$. But then f_a must also be strictly increasing at $r = 1$, so f_a is strictly increasing on $(0, \infty)$, which is what we had to prove.

Proof of upper bound on $C_{0 \leftarrow \eta}(V)$: The right term in $s(r)$ as given by (66) is positive for $r < 1$, and $g_{\eta'}(x) > g_\eta(x)$, so setting $t(r)$ to $s(r)$, but with $g_{\eta'}(r)$ in the right term in the numerator replaced by $g_\eta(r)$, i.e.,

$$t(r) := \frac{(-r^{\eta'} + r) \cdot g_\eta(r) + (r^\eta - r) \cdot g_\eta(r)}{g_\eta(r)^2} = \frac{-r^{\eta'} + r^\eta}{g_\eta(r)},$$

we have $t(r) \le s(r)$ for all $r \le 1$. We already know that $h_{\eta', \eta}$ is decreasing, so that $s(r) \le 0$ for all r, so we have $t(r) \le s(r) \le 0$ for all $r \le 1$. In particular, this holds for the case $\eta' = 0$, for which $t(r)$ simplifies to $t(r) = (-1 + r^\eta)/g_\eta(r) = -(1 - r^\eta)/(\eta^{-1}(1 - r^\eta) - (1 - r))$. A simple calculation shows that (a) $\lim_{r \downarrow 0} t(r) = -1/(\eta^{-1} - 1) = -\eta/(1 - \eta)$ and (b) $t(r)$ is increasing on $0 < r < 1$ for all $0 < \eta < 1$.

Now define $\tilde{h}$ by setting $\tilde{h}(r) = (1/(1 - \eta)) \cdot (1 - \eta \log r)$ for $0 < r \le 1$. Then $\tilde{h}'(r) = -(\eta/(1 - \eta))r^{-1} \le t(r) r^{-1} \le s(r) r^{-1} = h'_{0,\eta}(r) \le 0$ by all the above together. Since $\tilde{h}(1) = h_{0,\eta}(1)$, and for $r < 1$, $h_{0,\eta}$ is decreasing but $\tilde{h}$ is decreasing even faster, we must have $\tilde{h}(r) \ge h_{0,\eta}(r)$ for $0 < r < 1$. We can thus bound $h_{0,\eta}(1/V)$ by $\tilde{h}(1/V)$, and the result follows. ■

C.2. Proof of Lemma 16

Proof Markov's inequality implies that for all $f \in \mathcal{F}$, $\Pr(e^{\delta L_f} > u) < \frac{M_\delta}{u}$ for any $u \geq 0$. Therefore, for some map $\tau : \mathbb{R}_+ \to \mathbb{R}_+$ to be set later:

$$\begin{aligned} &\mathbf{E}\left[L_f \cdot \mathbf{1}_{\{L_f > \tau(\mathbf{E}[L_f])\}}\right] = \int_0^\infty \Pr(L_f \cdot \mathbf{1}_{\{L_f > \tau(\mathbf{E}[L_f])\}} > t) dt = \\ &\int_{\tau(\mathbf{E}[L_f])}^\infty \Pr(L_f > t) dt = \int_{\tau(\mathbf{E}[L_f])}^\infty \Pr(e^{\delta L_f} > e^{\delta t}) dt \leq \int_{\tau(\mathbf{E}[L_f])}^\infty M_\delta e^{-\delta t} dt \quad = \frac{M_\delta}{\delta} e^{-\delta \tau(\mathbf{E}[L_f])}. \end{aligned} \tag{69}$$

Taking $\tau : x \mapsto 1 \vee \frac{\log \frac{2M_\delta}{\delta x}}{\delta}$, the last line above is bounded by $\frac{1}{2}\mathbf{E}[L_f]$, and so the (τ, c)-witness condition holds with $c = 1/2$. ■

C.3. Proofs Related to Heavy-tailed Regression

We start with some general facts. For squared loss, the excess loss can be written as (abbreviating $f(X)$ and $f^*(X)$ to f and f^*, resp.),

$$L_f = (f(X) - f^*(X)) \cdot (-2Y + f(X) + f^*(X)) \tag{70}$$

$$= (f - f^*) \cdot ((f - f^*) + 2(f^* - Y)) \tag{71}$$

$$= (f - f^*)^2 + 2(f^* - Y)(f - f^*). \tag{72}$$

Now, recall that in both Examples 7 and 12, we assumed that the risk minimizer f^* over $\mathcal{F}$ continues to be a minimizer when taking the minimum risk over the convex hull of $\mathcal{F}$. This implies that for all $f \in \mathcal{F}$,

$$\mathbf{E}\,(f^*(X) - Y)(f(X) - f^*(X))] \geq 0, \tag{73}$$

To see this, we observe that if we instead consider the function class $\mathrm{conv}(\mathcal{F})$, then f^* is still a minimizer and (73) holds for all $f \in \mathrm{conv}(\mathcal{F})$ from Mendelson (2017a) (see the text around equation (1.3) therein).

But now (73) with (72) implies that, under our assumptions,

$$\mathbf{E}\left[(f(X) - f^*(X))^2\right] \leq \mathbf{E}[L_f]. \tag{74}$$

Proof (of Proposition 18) Let $u > 0$ be a to-be-determined constant. Then

$$\begin{aligned} &\mathbf{E}\left[L_f \cdot \mathbf{1}_{\{L_f > \tau(\mathbf{E}[L_f])\}}\right] \leq \mathbf{E}\left[L_f \cdot \frac{L_f}{\tau(\mathbf{E}[L_f])} \cdot \mathbf{1}_{\{L_f \geq 0\}}\right] = \\ &\frac{1}{\tau(\mathbf{E}[L_f])} \mathbf{E}\left[L_f^2 \cdot \mathbf{1}_{\{L_f \geq 0\}}\right] \leq \frac{1}{\tau(\mathbf{E}[L_f])} \mathbf{E}\left[L_f^2\right] \leq \frac{B}{u} \frac{\left(\mathbf{E}\left[L_f\right]\right)^\beta}{(\mathbf{E}[L_f])^{\beta - 1}} = \frac{B}{u} \mathbf{E}[L_f], \end{aligned}$$

and the result follows. ■

Proof (of Proposition 19) To see that a Bernstein condition holds if $\mathbf{E}[Y^2 \mid X] \le C$ a.s. and $|f(X)| \le r$ almost surely, observe that from (70),

$$L_f^2 \le 2(f(X) - f^*(X))^2 \left(4Y^2 + (f(X) - f^*(X))^2\right),$$

and hence

$$\begin{aligned}\mathbf{E}\left[L_f^2\right] &\le 8\left(\mathbf{E}\left[(f(X) - f^*(X))^2\,\mathbf{E}[Y^2 \mid X]\right] + r^2\,\mathbf{E}\left[(f(X) - f^*(X))^2\right]\right) \\ &\le 8(C + r^2)\,\mathbf{E}\left[(f(X) - f^*(X))^2\right].\end{aligned}$$

Invoking (74), we see that a Bernstein condition does indeed hold:

$$\mathbf{E}\left[L_f^2\right] \le 8(C + r^2)\,\mathbf{E}[L_f].$$

■

Proof (of Claim in Example 12) From (71), Cauchy-Schwarz, and our assumption,

$$\mathbf{E}[L_f^2] \le \sqrt{\mathbf{E}[(f(X) - f^*(X))^4]} \cdot \sqrt{C} \le A\,\mathbf{E}[(f(X) - f^*(X))^2] \cdot \sqrt{C} \le A\,\mathbf{E}[L_f] \cdot \sqrt{C}, \quad (75)$$

where the final inequality follows from (74) and

$$\begin{aligned}C &= \mathbf{E}[((f - f^*) + 2(Y - f^*))^4] \le \mathbf{E}[(2(f - f^*)^2 + 8(Y - f^*)^2)^2] \\ &\le \mathbf{E}[8(f - f^*)^4 + 32(Y - f^*)^4] \le 8A^2\,\mathbf{E}[(f - f^*)^2]^2 + 32\,\mathbf{E}[\ell_{f^*}^2] \\ &\le 8A^2\,\mathbf{E}[L_f]^2 + 32\,\mathbf{E}[\ell_{f^*}^2] \le 8A^2 c_0^2 + 32\,\mathbf{E}[\ell_{f^*}^2],\end{aligned}$$

where the third and fifth inequality follow from our assumptions and the fourth follows from (74). This quantity is bounded, so (75) implies the Bernstein condition. ■

Appendix D. Proofs for Section 6.1

D.1. Proof of Lemma 21

We first prove (47) from the main text: suppose that $(P, \ell, \mathcal{F})$ satisfies the v-central condition. We then have for all $f \in \mathcal{F}$,

$$\mathbf{E}\left[e^{v(\epsilon)\cdot(\ell_{f_\epsilon^*} - \ell_f)}\right] = \mathbf{E}\left[e^{v(\epsilon)\cdot(\ell_{f^*} - \ell_f)}\right] \cdot e^{-v(\epsilon)\epsilon} \le 1,$$

where the inequality follows because $(P, \ell, \mathcal{F})$ satisfies the v-central condition. Now suppose further that $(P, \ell, \{f\} \cup \{f^*\})$ satisfies the (u, c)-witness condition. This gives:

$$\begin{aligned}c\,\mathbf{E}[L_f] &\le \mathbf{E}[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \le u\}}] = \mathbf{E}[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f_\epsilon^*} \le u + \epsilon\}}] \\ &= \mathbf{E}[(\ell_f - (\ell_{f_\epsilon^*} + \epsilon)) \cdot \mathbf{1}_{\{\ell_f - \ell_{f_\epsilon^*} \le u + \epsilon\}}] \le \mathbf{E}[(\ell_f - \ell_{f_\epsilon^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f_\epsilon^*} \le u + \epsilon\}}],\end{aligned}$$

whence the $(u{+}\epsilon, c)$ witness condition holds for $(P, \ell, \{f, f_\epsilon^*\})$. By this fact and (47) (proven above), we can apply Lemma 34 (our extension of Lemma 13 from the main text), with $\phi(f)$ set to f_ϵ^* (i.e. $\phi(f)$ does not depend on f). The result, (46), follows.

Appendix E. Proofs for Section 6.2

E.1. Proof of Propositions 24–27

Proof (of Proposition 24) Consider the learning problem $(P, \tilde{\ell}, \tilde{\mathcal{F}})$ with

$$\tilde{\mathcal{F}} := \left\{ m_Q^\eta : Q \in \Delta(\mathcal{F}) \right\} \cup \{ m_{\mathcal{F}}^\eta \}$$

and $\tilde{\ell}_{\tilde{f}} := \tilde{f}$ for $\tilde{f} \in \tilde{\mathcal{F}}$.

We will show that the strong η-PPC condition (Van Erven et al., 2015) holds for this problem with $m_{\mathcal{F}}^\eta$ taking the role of the optimal action. That is,

$$\mathbf{E}\left[m_{\mathcal{F}}^\eta \right] \leq \inf_{\tilde{Q} \in \Delta(\tilde{\mathcal{F}})} \mathbf{E}\left[-\frac{1}{\eta} \log \mathbf{E}_{\tilde{f} \sim \tilde{Q}} \left[e^{-\eta \tilde{\ell}_{\tilde{f}}} \right] \right]. \tag{76}$$

In one of their main results, (Van Erven et al., 2015, Theorem 3.10 and Corollary 3.11), again extending an argument of Li (1999), show that the strong η-PPC condition implies the strong η-central condition for any tuple $(P, \ell, \tilde{\mathcal{F}})$ under the sole assumption that $\tilde{\mathcal{F}}$ contains a risk minimizer, i.e., there exists $f' \in \tilde{\mathcal{F}}$ with $\min_{f \in \tilde{\mathcal{F}}} \mathbf{E}[\ell_f] = \mathbf{E}[\ell_{f'}]$. But we construct $\tilde{\mathcal{F}}$ so that this holds, since it contains $m_{\mathcal{F}}^\eta$. Thus, if (76) indeed holds (as we will soon show), then $(P, \tilde{\ell}, \tilde{\mathcal{F}})$ also satisfies the the strong η-central condition. But this implies that, for all $\tilde{f} \in \tilde{\mathcal{F}}$,

$$\mathbf{E}\left[e^{-\eta(\tilde{\ell}_{\tilde{f}} - m_{\mathcal{F}}^\eta)} \right] \leq 1.$$

The statement above holds in particular for any $\tilde{f} = m_Q^\eta$, which includes the special case of the Dirac mix losses of the form $m_{\delta_f}^\eta = \ell_f$ for any $f \in \mathcal{F}$, and hence we have, for all $f \in \mathcal{F}$,

$$\mathbf{E}\left[e^{-\eta(\ell_f - m_{\mathcal{F}}^\eta)} \right] \leq 1 \qquad \text{for all } f \in \mathcal{F},$$

which is what we wanted.

Let us now prove inequality (76). We start with the RHS of (76) and, via a sequence of lower bounds, will arrive at the LHS. First, observe that the RHS can be rewritten as

$$\begin{aligned} &\inf_{\alpha \in [0,1]} \inf_{\tilde{Q} \in \Delta(\Delta(\mathcal{F}))} \mathbf{E}\left[-\frac{1}{\eta} \log \left(\alpha e^{-\eta m_{\mathcal{F}}^\eta} + (1-\alpha) \mathbf{E}_{Q \sim \tilde{Q}} \left[e^{-\eta m_Q^\eta} \right] \right) \right] \\ &= \inf_{\alpha \in [0,1]} \inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}\left[-\frac{1}{\eta} \log \left(\alpha e^{-\eta m_{\mathcal{F}}^\eta} + (1-\alpha) m_Q^\eta \right) \right]. \end{aligned}$$

Next, for each α and Q, we introduce a function $\Gamma_{\alpha,Q} \colon \mathbb{R} \to \mathbb{R}$, defined as

$$\Gamma_{\alpha,Q}(x) = -\frac{1}{\eta} \log \left(\alpha e^{-\eta x} + (1-\alpha) m_Q^\eta \right),$$

so that the last line in the above display may be rewritten as

$$\inf_{\alpha \in [0,1]} \inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}\left[\Gamma_{\alpha,Q}(m_{\mathcal{F}}^\eta) \right].$$

Now, as we show in Appendix G, there exists a sequence $(Q_n)_{n\geq 1}$ such that $m^{\eta}_{Q_n}$ converges to $m^{\eta}_{\mathcal{F}}$ in $L_1(P)$. For any $n \geq 1$, we have

$$\mathbf{E}\left[\Gamma_{\alpha,Q}(m^{\eta}_{\mathcal{F}})\right] = \mathbf{E}\left[\Gamma_{\alpha,Q}(m^{\eta}_{Q_n})\right] + \mathbf{E}\left[\Gamma_{\alpha,Q}(m^{\eta}_{\mathcal{F}}) - \Gamma_{\alpha,Q}(m^{\eta}_{Q_n})\right] \tag{77}$$

Note that $\Gamma_{\alpha,Q}$ is 1-Lipschitz, since (for any choice of α and Q),

$$\frac{d\Gamma_{\alpha,Q}}{dx}\Gamma_{\alpha,Q}(x) = -\frac{1}{\eta}\frac{-\eta\alpha e^{-\eta x}}{\alpha e^{-\eta x} + (1-\alpha)e^{-\eta m^{\eta}_{Q}}} = \frac{\alpha e^{-\eta x}}{\alpha e^{-\eta x} + (1-\alpha)e^{-\eta m^{\eta}_{Q}}} \in [0,1].$$

Consequently, it holds that (77) is lower bounded by

$$\mathbf{E}\left[\Gamma_{\alpha,Q}(m^{\eta}_{Q_n})\right] - \mathbf{E}\left[\left|\Gamma_{\alpha,Q}(m^{\eta}_{\mathcal{F}}) - \Gamma_{\alpha,Q}(m^{\eta}_{Q_n}\right|\right] \geq \mathbf{E}\left[\Gamma_{\alpha,Q}(m^{\eta}_{Q_n})\right] - \mathbf{E}\left[\left|m^{\eta}_{\mathcal{F}} - m^{\eta}_{Q_n}\right|\right].$$

Next, since $m^{\eta}_{Q_n}$ converges to $m^{\eta}_{\mathcal{F}}$ in $L_1(P)$, taking the limit as $n \to \infty$, the RHS of the last line above converges to $\mathbf{E}\left[\Gamma_{\alpha,Q}(m^{\eta}_{Q_n})\right]$. Thus, we have shown that

$$\mathbf{E}\left[\Gamma_{\alpha,Q}(m^{\eta}_{\mathcal{F}})\right] \geq \lim_{n\to\infty} \mathbf{E}\left[\Gamma_{\alpha,Q}(m^{\eta}_{Q_n})\right],$$

and so:

$$\begin{aligned}
&\inf_{\alpha\in[0,1]}\inf_{Q\in\Delta(\mathcal{F})} \mathbf{E}\left[-\frac{1}{\eta}\log\left(\alpha e^{-\eta m^{\eta}_{\mathcal{F}}} + (1-\alpha)e^{-\eta m^{\eta}_{Q}}\right)\right] \\
&\geq \inf_{\alpha\in[0,1]}\inf_{Q\in\Delta(\mathcal{F})}\lim_{n\to\infty} \mathbf{E}\left[-\frac{1}{\eta}\log\left(\alpha e^{-\eta m^{\eta}_{Q_n}} + (1-\alpha)e^{-\eta m^{\eta}_{Q}}\right)\right] \\
&= \inf_{\alpha\in[0,1]}\inf_{Q\in\Delta(\mathcal{F})}\lim_{n\to\infty} \mathbf{E}\left[m^{\eta}_{\alpha Q_n + (1-\alpha)Q}\right] \\
&\geq \inf_{\alpha\in[0,1]}\inf_{Q\in\Delta(\mathcal{F})}\lim_{n\to\infty} \mathbf{E}\left[m^{\eta}_{\mathcal{F}}\right] \\
&= \mathbf{E}\left[m^{\eta}_{\mathcal{F}}\right],
\end{aligned}$$

where we used that the quantity inside $\lim_{n\to\infty}$ is equal to $\mathbf{E}[m^{\eta}_{Q'}]$ for some $Q' \in \Delta(\mathcal{F})$, and hence by definition not smaller than $\mathbf{E}[m^{\eta}_{\mathcal{F}}]$. Thus, inequality (76) indeed holds. ■

Proof (of Proposition 26) Fix $\eta > 0$ and let u be as in (50). For each $f \in \mathcal{F}$, let f' be defined by $\ell_{f'} = \ell_f$ if $\ell_f \leq \ell_{f^*} + u$ and $\ell_{f'} = \ell_{f^*}$ otherwise and let $\mathcal{F}'$ be the resulting model. Then $m^{\eta}_{\mathcal{F}'}$ is the GRIP relative to η and the class $\mathcal{F}'$; from Appendix G this GRIP is guaranteed to exist. By definition, for every $\delta > 0$ there is a distribution Q' on $\mathcal{F}'$ such that $\mathbf{E}_{Z\sim P}[m^{\eta}_{Q'} - m^{\eta}_{\mathcal{F}'}] \leq \delta$. Define f° such that it has constant loss, i.e., for all $z \in \mathcal{Z}$, $\ell_{f^\circ}(z) := \mathbf{E}[\ell_{f^*}]$. By using $-\log x \geq 1 - x$ and we have for each $z \in \mathcal{Z}$, for some $\eta' \in (0, \eta)$:

$$\begin{aligned}
m^{\eta}_{Q'} - \ell_{f^\circ} = -\frac{1}{\eta}\log \mathbf{E}_{f'\sim Q'}\, e^{-\eta(\ell_{f'} - \ell_{f^\circ})} &\geq \frac{1}{\eta}\left(1 - \mathbf{E}_{f'\sim Q'}\, e^{-\eta(\ell_{f'} - \ell_{f^\circ})}\right) \\
&= \mathbf{E}_{f'\sim Q'}\left[\ell_{f'} - \ell_{f^\circ}\right] - \frac{1}{2}\eta\, \mathbf{E}(\ell_{f'} - \ell_{f^\circ})^2 e^{-\eta'(\ell_{f'} - \ell_{f^\circ})} \\
&\geq \mathbf{E}_{f'\sim Q'}\left[\ell_{f'} - \ell_{f^\circ}\right] - \frac{1}{2}e^{\eta \ell_{f^\circ}} \cdot \eta\, \mathbf{E}_{f'\sim Q'}(\ell_{f'} - \ell_{f^\circ})^2.
\end{aligned}$$

Now use that

$$\begin{aligned}
\mathbf{E}_{f' \sim Q'}\left[(\ell_{f'} - \ell_{f^\circ})^2\right] &= \mathbf{E}_{f' \sim Q'}\left[\left((\ell_{f'} - \ell_{f^*}) + (\ell_{f^*} - \ell_{f^\circ})\right)^2\right] \\
&\leq 2\left(\mathbf{E}_{f' \sim Q'}\left[(\ell_{f'} - \ell_{f^*})^2\right] + (\ell_{f^*} - \ell_{f^\circ})^2\right) \\
&\leq 2\left(\mathbf{E}_{f' \sim Q'}\left[\mathbf{1}_{\{\ell_{f'} > \ell_{f^*}\}}(\ell_{f'} - \ell_{f^*})^2 + \mathbf{1}_{\{\ell_{f'} \leq \ell_{f^*}\}}(\ell_{f'} - \ell_{f^*})^2\right] + (\ell_{f^*} - \ell_{f^\circ})^2\right) \\
&\leq 2u^2 + 2\ell_{f^*}^2 + (\ell_{f^*} - \ell_{f^\circ})^2.
\end{aligned}$$

Combining this with the previous inequality and taking the expectation with respect to Z yields

$$\begin{aligned}
\mathbf{E}_{Z \sim P}\left[m^\eta_{\mathcal{F}'} - \ell_{f^*}\right] &= \mathbf{E}_{Z \sim P}\left[m^\eta_{Q'} - \ell_{f^\circ}\right] - \delta \\
&\geq \mathbf{E}_{Z \sim P}\,\mathbf{E}_{f' \sim Q'}\left[\ell_{f'} - \ell_{f^*}\right] - \frac{1}{2}\eta e^{\eta \ell_{f^\circ}} \cdot \left(2u^2 + \mathbf{E}_{Z \sim P}\left[2\ell^2_{f^*} + (\ell_{f^*} - \ell_{f^\circ})^2\right]\right) - \delta \\
&\geq \mathbf{E}_{Z \sim P}\,\mathbf{E}_{f' \sim Q'}\left[(\ell_{f'} - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_{f'} - \ell_{f^*} \leq u\}}\right] - \frac{1}{2}\eta e^{\eta\,\mathbf{E}[\ell_{f^*}]} \cdot \left(2u^2 + 3\,\mathbf{E}[\ell^2_{f^*}]\right) - \delta \\
&= \mathbf{E}_{f \sim Q}\,\mathbf{E}_{Z \sim P}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u\}}\right] - \frac{1}{2}\eta e^{\eta\,\mathbf{E}[\ell_{f^*}]} \cdot \left(2u^2 + 3\,\mathbf{E}[\ell^2_{f^*}]\right) - \delta \\
&\geq -\frac{1}{2}\eta e^{\eta\,\mathbf{E}[\ell_{f^*}]} \cdot \left(2u^2 + 3\,\mathbf{E}[\ell^2_{f^*}]\right) - \delta,
\end{aligned}$$

where $Q \in \Delta(\mathcal{F})$ is the distribution defined by taking $dQ(f) = dQ'(f')$ (where we make use of the bijection between $\mathcal{F}$ and $\mathcal{F}'$ from the definition of $\ell_{f'}$ in terms of f, for all $f' \in \mathcal{F}$), and the final inequality invokes (50). We now take $\eta \leq 1/\mathbf{E}[\ell_{f^*}]$, yielding

$$\mathbf{E}_{Z \sim P}\left[\ell_{f^*} - m^\eta_{\mathcal{F}'}\right] \leq \eta \cdot e \cdot \left(u^2 + \frac{3}{2}\,\mathbf{E}[\ell^2_{f^*}]\right) + \delta. \tag{78}$$

The result now follows from Proposition 27, using that the reasoning above holds for every $\delta > 0$. ■

Proof (of Proposition 27) Define the set $\mathcal{F}'$ such that for each $f \in \mathcal{F}$, there is an $f' \in \mathcal{F}$ with $\ell'_f = \ell_{f'}$ and vice versa. Note that we must have:

$$\mathbf{E}_{Z \sim P}\left[m^\eta_{\mathcal{F}'}\right] \leq \mathbf{E}_{Z \sim P}\left[m^\eta_{\mathcal{F}}\right]. \tag{79}$$

To see this, assume for contradiction that there exists some $\varepsilon > 0$ such that $\mathbf{E}_{Z \sim P}\left[m^\eta_{\mathcal{F}}\right] \leq \mathbf{E}_{Z \sim P}\left[m^\eta_{\mathcal{F}'}\right] - \varepsilon$. Let $(Q_j)_{j \geq 1}$ be a sequence for which $\mathbf{E}_{Z \sim P}[m^\eta_{Q_j}] \leq \mathbf{E}_{Z \sim P}[m^\eta_{\mathcal{F}}] + \frac{\varepsilon}{2}$. We will make use of the fact that, for each $Q' \in \Delta(\mathcal{F}')$, $m^\eta_{Q'} \leq m^\eta_Q$ since for each f' the corresponding f has, on all z, either the same or larger loss. This setup then implies the following contradiction:

$$\mathbf{E}_{Z \sim P}[m^\eta_{\mathcal{F}'}] \leq \mathbf{E}_{Z \sim P}[m^\eta_{Q'_j}] \leq \mathbf{E}_{Z \sim P}[m^\eta_{Q_j}] \leq m^\eta_{\mathcal{F}} + \frac{\varepsilon}{2} \leq m^\eta_{\mathcal{F}'} - \frac{\varepsilon}{2}.$$

Now, since by assumption $\ell_{f^*} \equiv \ell_{(f^*)'}$, (79) implies that

$$\mathbf{E}_{Z \sim P}\left[\ell_{f^*} - m^\eta_{\mathcal{F}}\right] \leq \mathbf{E}_{Z \sim P}\left[\ell_{f^*} - m^\eta_{\mathcal{F}'}\right]$$

which implies the statement of the proposition. ■

E.2. Proof of Lemma 28

The proof of Lemma 28 is based on relating the loss $m^{\bar{\eta}}_{\mathcal{F}}$ of the GRIP comparator appearing in that lemma to the loss of a related "dynamic" comparator $m^{\bar{\eta}}_{f}$ (which we will call "mini-GRIP") that *varies* with f. This requires us to first re-define the witness condition for such dynamic comparators, relate this dynamic witness condition to the standard witness condition, and relate the GRIP loss to the mini-GRIP loss; this is all achieved in the following subsection.

E.2.1. Witness Protection and Mini-grip

Assumption 1 (Advanced Empirical Witness of Badness) *Let $M \geq 1$ be a parameter of the assumption. We say that $(P, \ell, \mathcal{F})$ satisfies the* empirical witness of badness condition *(abbreviated as* witness condition*) with respect to* dynamic *comparator ϕ if there exist constants $u > 0$ and $c \in (0,1]$ such that for all $f \in \mathcal{F}$,*

$$\mathbf{E}\left[(\ell_f - \ell_{\phi(f)}) \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} \leq u(1 \vee (M^{-1}\mathbf{E}[L_f]))\}}\right] \geq c\,\mathbf{E}[\ell_f - \ell_{\phi(f)}]. \tag{80}$$

If we modify the RHS of (80) *so that the term $\mathbf{E}[\ell_f - \ell_{\phi(f)}]$ is replaced by the potentially smaller $\mathbf{E}[\ell_f - \ell_{f^*}]$, then we call the condition the* weak empirical witness of badness condition *(abbreviated as* weak witness condition*).*

In practice, we will assume only that the witness condition holds for the *static* comparator $\psi : f \mapsto f^*$ (so named because the comparator does not vary with f), as can already be handled through the simpler witness condition of Definition 12. However, because the central condition may not necessarily be satisfied with comparator f^*, it is beneficial if a witness condition holds for a suitably-related comparator for which the central condition *does* hold. The ideal candidate for this comparator turns out to be an f-dependent pseudo-loss, m^{η}_{f}, an instance of a GRIP (see Definition 23).

The main motivation for our introducing the GRIP is that $(P, \ell, \mathcal{F})$ with comparator $m^{\eta}_{\mathcal{F}}$ satisfies the η-central condition (from Proposition 24). The GRIP arises as a generalization of the reversed information projection of Li (1999), which is the special case of the above with $\eta = 1$, log loss, and $\mathcal{F}$ a class of probability distributions. In this case, the GRIP, now a reversed information projection, is the (limiting) distribution P^* which minimizes the KL divergence $\mathrm{KL}(P \,\|\, P^*)$ over the convex hull of $\mathcal{P}$; note that P^* is not necessarily in $\mathrm{conv}(\mathcal{P})$. Li (1999, Theorem 4.3) proved the existence of the reversed information projection; for completeness, in Appendix G we present a lightly modified proof of the existence of the GRIP.

As mentioned above, in our technical results exploiting both the central and witness conditions, we will need not only the "full" GRIP but also a "mini-grip" m^{η}_{f}, for each f, defined by replacing $\mathcal{F}$ with $\{f^*, f\}$ in Definition 23. The mini-grip with respect to f then has the simple, characterizing property of satisfying

$$\mathbf{E}[m^{\eta}_{f}] = \inf_{\alpha \in [0,1]} \mathbf{E}\left[-\frac{1}{\eta} \log\left((1-\alpha) e^{-\eta \ell_{f^*}} + \alpha e^{-\eta \ell_f}\right)\right].$$

Also, as will be used to critical effect in the application of Lemma 34, for each f the learning problem $(P, \{f^*, f\}, \ell)$ with comparator m^{η}_{f} satisfies the η-central condition.

Although up until now it has sufficed to refer to GRIPs only via their loss, for convenience of notation we now let $g_{\mathcal{F}}^{\eta}$ denote the pseudo-action obtaining the GRIP loss $m_{\mathcal{F}}^{\eta}$, and we let g_f^{η} denote the pseudo-action obtaining the mini-GRIP loss m_f^{η}. It should be emphasized that neither $g_{\mathcal{F}}^{\eta}$ nor g_f^{η} need be well-defined; this is of no consequence, however, as we will use both only via their losses $m_{\mathcal{F}}^{\eta}$ and m_f^{η}, which *are* well-defined.

We now show that if the witness condition holds with respect to the static comparator $\psi : f \mapsto f^*$, then the weak witness condition holds with respect to the comparator $\phi : f \mapsto g_f^{\eta}$.

Lemma 37 (Witness Protection Lemma) *Assume that $(P, \ell, \mathcal{F})$ satisfies the witness condition with static comparator $\psi : f \mapsto f^*$ and constants (M, u, c). Then, for any $\eta > 0$, $(P, \ell, \mathcal{F})$ satisfies the weak witness condition with dynamic comparator $\phi : f \mapsto g_f^{\eta}$ with the same constants (M, u, c).*

Proof (of Lemma 37 (Witness Protection Lemma)) Let f be arbitrary. For brevity we define $u' \coloneqq u(1 \vee (M^{-1}\,\mathbf{E}[L_f]))$. Observe that

$$\mathbf{E}\left[(\ell_f - m_f^{\eta}) \cdot \mathbf{1}_{\{\ell_f - m_f^{\eta} > u'\}}\right] \le \mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} > u'\}}\right].$$

Rewriting, we have

$$\mathbf{E}[\ell_f - m_f^{\eta}] - \mathbf{E}\left[(\ell_f - m_f^{\eta}) \cdot \mathbf{1}_{\{\ell_f - m_f^{\eta} \le u'\}}\right] \le \mathbf{E}[L_f] - \mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \le u'\}}\right],$$

which we rearrange as

$$\begin{aligned}
\mathbf{E}\left[(\ell_f - m_f^{\eta}) \cdot \mathbf{1}_{\{\ell_f - m_f^{\eta} \le u'\}}\right] &\ge \mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \le u'\}}\right] + \mathbf{E}[\ell_f - m_f^{\eta}] - \mathbf{E}[L_f] \\
&= \mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \le u'\}}\right] + \mathbf{E}[\ell_{f^*} - m_f^{\eta}] \\
&\ge \mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \le u'\}}\right].
\end{aligned}$$

From the assumed witness condition with static comparator $\psi : f \mapsto f^*$, the RHS is lower bounded by $c\,\mathbf{E}[L_f]$, and so we have established the weak witness condition with dynamic comparator ϕ and the same constants (M, u, c). ■

From Hellinger mini-grip to GRIP

Lemma 38 *For any $\eta > 0$ and $f \in \mathcal{F}$,*

$$\mathbf{E}^{\mathrm{HE}(\eta)}\left[\ell_f - m_f^{\eta}\right] \le \mathbf{E}^{\mathrm{HE}(\eta/2)}\left[\ell_f - m_{\mathcal{F}}^{\eta}\right]. \tag{81}$$

Proof Observe that

$$\begin{aligned}
\frac{1}{\eta/2}\left(1 - \mathbf{E}\left[e^{-\frac{\eta}{2}(\ell_f - m_{\mathcal{F}}^{\eta})}\right]\right) &= \frac{1}{\eta/2}\left(1 - \mathbf{E}\left[e^{-\frac{\eta}{2}(\ell_f - m_f^{\eta} + m_f^{\eta} - m_{\mathcal{F}}^{\eta})}\right]\right) \\
&\ge \frac{1}{\eta/2}\left(1 - \frac{1}{2}\mathbf{E}\left[e^{-\eta(\ell_f - m_f^{\eta})}\right] - \frac{1}{2}\mathbf{E}\left[e^{-\eta(m_f^{\eta} - m_{\mathcal{F}}^{\eta})}\right]\right) \\
&\ge \frac{1}{\eta/2}\left(\frac{1}{2} - \frac{1}{2}\mathbf{E}\left[e^{-\eta(\ell_f - m_f^{\eta})}\right]\right) = \frac{1}{\eta}\left(1 - \mathbf{E}\left[e^{-\eta(\ell_f - m_f^{\eta})}\right]\right),
\end{aligned}$$

where the first inequality follows from Jensen's and for the second inequality we use that, as we will now show, $\mathbf{E}\left[e^{-\eta(m_f^\eta - m_{\mathcal{F}}^\eta)}\right] \leq 1$. To show that this is indeed the case, recall that $m_f^\eta = -\frac{1}{\eta}\log\left((1-\alpha)e^{-\eta\ell_{f^*}} + \alpha e^{-\eta\ell_f}\right)$. Using this representation we find:

$$\mathbf{E}\left[e^{-\eta(m_f^\eta - m_{\mathcal{F}}^\eta)}\right] = (1-\alpha)\,\mathbf{E}\left[e^{-\eta(\ell_f^* - m_{\mathcal{F}}^\eta)}\right] + \alpha\,\mathbf{E}\left[\alpha e^{-\eta(\ell_f - m_{\mathcal{F}}^\eta)}\right] \leq 1.$$

■

Next, we chain $1 - x \leq -\log x$, Lemma 38, and Lemma 34 to obtain a bound that we will use in the proofs of Theorems 29 and 31.

E.3. Actual Proof of Lemma 28

Let $f \in \mathcal{F}$. Let $u > 0$ and $c \in (0,1]$ be constants for which $\mathbf{E}\left[L_f \cdot \mathbf{1}_{\{L_f \leq u\}}\right] \geq c\,\mathbf{E}[L_f]$, i.e., the (u,c)-witness condition holds. Below we show that for all $\eta \in (0, \frac{\bar{\eta}}{2})$

$$\mathbf{E}[L_f] \leq c'_{2u}\mathbf{E}^{\text{ANN}(\eta)}\left[\ell_f - m_{\mathcal{F}}^{\bar{\eta}}\right], \tag{82}$$

with $c'_{2u} = \frac{1}{c}\frac{2\eta u + 1}{1 - \frac{2\eta}{\bar{\eta}}}$.

Proof of (82). We have from (12) and Lemma 38 that

$$\mathbf{E}^{\text{ANN}(\eta)}\left[\ell_f - m_{\mathcal{F}}^{\bar{\eta}}\right] \geq \mathbf{E}^{\text{HE}(\eta)}\left[\ell_f - m_{\mathcal{F}}^{\bar{\eta}}\right].$$

Now Lemma 37 establishes the weak witness condition with respect to comparator $g_f^{\bar{\eta}}$, and from Proposition 24 this comparator further satisfies $\mathbf{E}\left[e^{-\bar{\eta}(\ell_f - m_f^{\bar{\eta}})}\right] \leq 1$, so that we may apply Lemma 34 with $\phi(f) = g_f^{\bar{\eta}}$ to further lower bound the above by $\frac{1}{c'_{2u}}\mathbf{E}[L_f]$.

E.4. Proof of Theorem 29

Theorem 29 now follows easily from Lemma 28: fix some $\epsilon \geq 0$. First, Lemma 33 (our extension of Lemma 5 from the main text) states for our particular choice of η that

$$\mathbf{E}_{\underline{f}\sim\Pi_n}\left[-\frac{1}{\eta}\log\mathbf{E}\left[e^{-\eta(\ell_{\underline{f}} - m_{\mathcal{F}}^{v(\epsilon)})}\right]\right] \trianglelefteq_{\eta\cdot n} \mathbf{E}_{\underline{f}\sim\Pi_n}\left[\frac{1}{n}\sum_{j=1}^{n}(\ell_{\underline{f}}(Z_j) - m_{\mathcal{F}}^{v(\epsilon)}(Z_j))\right] + \frac{\mathrm{KL}(\Pi_n \parallel \Pi_0)}{\eta n}. \tag{83}$$

Weakening this to an in-expectation statement via part (i) of Proposition 3, and combining the in-expectation version with Lemma 28, (51) implies that, for $c'_{2u} = \frac{1}{c}\frac{2\eta u+1}{1-\frac{2\eta}{v(\epsilon)}}$,

$$\mathbf{E}_{Z_1^n}\left[\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\mathbf{E}[L_{\underline{f}}]\right]\right] \;\leq\; c'_{2u}\,\mathbf{E}_{Z_1^n}\left[\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\frac{1}{n}\sum_{j=1}^{n}(\ell_{\underline{f}}(Z_j) - m_{\mathcal{F}}^{v(\epsilon)}(Z_j))\right] + \frac{\mathrm{KL}(\Pi_n \parallel \Pi_0)}{\eta n}\right]. \tag{84}$$

Now, the v-PPC condition implies that $\mathbf{E}[\ell_{f^*}] \leq \mathbf{E}[m_{\mathcal{F}}^{v(\epsilon)}] + \epsilon$, implying the result (52).

Appendix F. Proofs for Section 6.3

F.1. Proof of Proposition 30

We first state another proposition that is of independent interest, relating generalized "small-ball" assumptions to weakenings thereof which resemble the witness condition.

Definition 39 *We say that a collection of nonnegative random variables $\{S_a : a \in \mathcal{A}\}$ satisfies the* generalized small-ball condition *if there exist constants C_1, C_2 with for all $a \in \mathcal{A}$, $P(S_a \geq C_1 \mathbf{E}[S_a]) \geq C_2$ (Mendelson's (2014) small-ball assumption in Example 9 and 12 is the case with $\mathcal{A} = \mathcal{F} \times \mathcal{F}$, $S_{f,g} := (f(X) - g(X))^2$, $C_1 = \kappa^2, C_2 = \epsilon$). We say that $\{S_a : a \in \mathcal{A}\}$ satisfies the* generalized weakened small-ball condition *if there exist constants C'_1, C'_2 with for all $a \in \mathcal{A}$, $\mathbf{E}[\mathbf{1}_{\{S_a < C'_1 \mathbf{E}[S_a]\}} \cdot S_a] \geq C'_2 \mathbf{E}[S_a]$.*

The term "weakened" comes from the following proposition:

Proposition 40 *Suppose that the generalized small-ball condition holds with constants C_1 and C_2. Then the generalized weakened small-ball condition holds with constants $C'_1 = 2/C_2$ and $C'_2 = (C_1 C_2)/2$.*

Proof From Markov's inequality, we have for all $a \in \mathcal{A}$, $P(S_a < (2/C_2) \mathbf{E}[S_a]) \geq 1 - C_2/2$. In combination with the small-ball assumption, this implies

$$P\left(C_1 \mathbf{E}[S_a] \leq S_a < \frac{2}{C_2} \mathbf{E}[S_a]\right) \geq \frac{C_2}{2},$$

and so, since $S_a \geq 0$,

$$\mathbf{E}\left[\mathbf{1}_{\{S_a < (2/C_2)\mathbf{E}[S_a]\}} \cdot S_a\right] \geq \mathbf{E}\left[\mathbf{1}_{\{C_1 \mathbf{E}[S_a] \leq S_a < (2/C_2)\mathbf{E}[S_a]\}} \cdot S_a\right] \geq \frac{C_2}{2} \cdot C_1 \cdot \mathbf{E}[S_a],$$

and the result follows. ∎

Proof (of Proposition 30) Take some $c_0 > b$, with a precise value to be established later. First consider the set $\{f \in \mathcal{F} : \mathbf{E}[L_f] > c_0\}$. Define the random variable $S_f := (f(X) - f^*(X))^2$ and $T_f := 2(f^*(X) - Y)(f - f^*)$. From (72) we see that $L_f = S_f + T_f$. Hence for every $c > 0$,

$$\begin{aligned}
&\mathbf{E}[L_f \cdot \mathbf{1}_{\{L_f \geq c \mathbf{E}[L_f]\}}] \\
&\leq \mathbf{E}[S_f \cdot \mathbf{1}_{\{S_f \geq T_f\}} \cdot \mathbf{1}_{\{S_f + T_f \geq c \mathbf{E}[L_f]\}}] + \mathbf{E}[S_f \cdot \mathbf{1}_{\{S_f < T_f\}} \cdot \mathbf{1}_{\{S_f + T_f \geq c \mathbf{E}[L_f]\}}] + \mathbf{E}[|T_f|] \\
&\leq \mathbf{E}[S_f \cdot \mathbf{1}_{\{S_f \geq T_f\}} \cdot \mathbf{1}_{\{2S_f \geq c \mathbf{E}[L_f]\}}] + \mathbf{E}[T_f \cdot \mathbf{1}_{\{S_f < T_f\}} \cdot \mathbf{1}_{\{S_f + T_f \geq c \mathbf{E}[L_f]\}}] + \mathbf{E}[|T_f|] \\
&\leq \mathbf{E}[S_f \cdot \mathbf{1}_{\{S_f \geq T_f\}} \cdot \mathbf{1}_{\{S_f \geq (c/2) \mathbf{E}[S_f]\}}] + 2\mathbf{E}[|T_f|], \qquad (85)
\end{aligned}$$

where the last inequality follows since $\mathbf{E}[S_f] \leq \mathbf{E}[L_f]$, owing to (73).

We now bound both terms further. By Cauchy-Schwarz, the second term satisfies

$$\begin{aligned}
2\mathbf{E}[|T_f|] &= 4\mathbf{E}[|Y - f^*||f - f^*|] \\
&\leq 4\sqrt{\mathbf{E}[(Y - f^*)^2] \cdot \mathbf{E}[S_f^2]} \leq 4\sqrt{\frac{\mathbf{E}[\ell_{f^*}]}{\mathbf{E}[L_f]}} \cdot \mathbf{E}[L_f] < 4\sqrt{\frac{\mathbf{E}[\ell_{f^*}]}{c_0}} \cdot \mathbf{E}[L_f].
\end{aligned}$$

Plugging in $c' := (c/2) = 2/\epsilon$, the first term can be rewritten, by Proposition 40 and our assumption that the small-ball assumption holds, as

$$\mathbf{E}[S_f] - \mathbf{E}[S_f \cdot \mathbf{1}_{\{S_f < (c/2)\,\mathbf{E}[S_f]\}}] \leq \mathbf{E}[S_f] - \frac{\kappa^2 \epsilon}{2}\,\mathbf{E}[S_f] = (1 - \frac{\kappa^2 \epsilon}{2})\,\mathbf{E}[S_f] \leq (1 - \frac{\kappa^2 \epsilon}{2})\,\mathbf{E}[L_f],$$

so that with (85) we get

$$\mathbf{E}[L_f \cdot \mathbf{1}_{\{L_f \geq c'\,\mathbf{E}[L_f]\}}] \leq C'\,\mathbf{E}[L_f],$$

for $C' = \left(\left(1 - \frac{\kappa^2\epsilon}{2}\right) + 4\sqrt{\frac{\mathbf{E}[\ell_{f^*}]}{c_0}}\right)$. We now pick c_0 large enough such that $C' < 1$. It then follows by the characterization (36) of the witness condition that the set $\{f \in \mathcal{F} : \mathbf{E}[L_f] \geq c_0\}$ satisfies the (τ, c)-witness condition with $\tau(x) = c'x$ for $c' = 2/\epsilon$ and constant $c = 1 - C'$.

For the set $\{f \in \mathcal{F} : \mathbf{E}[L_f] < c_0\}$, note that we have already shown (Example 7) that the Bernstein condition implies the basic witness condition. This implies that there exists $u > 0$ such that $\{f \in \mathcal{F} : \mathbf{E}[L_f] \leq c_0\}$ satisfies the (u, c)-witness condition for $c = \frac{1}{2}$.

Putting the two statements for both subsets of $\mathcal{F}$ together, it follows that $\mathcal{F}$ satisfies the (τ, c)-witness condition with any τ such that $\tau(x) \geq u \vee \frac{2x}{\epsilon}$ for all x and with $c = (1 - C') \wedge \frac{1}{2}$; the result follows. ∎

F.2. Proof of Theorem 31

We will need the following lemma, whose proof is a straightforward extension of the proofs of Theorem 22 and Theorem 29:

Lemma 41 *With τ as in the statement of Theorem 31, we get for any $\epsilon \geq 0$, any $0 < \eta < \frac{v(\epsilon)}{2}$:*

$$\textit{under v-central:} \quad \mathbf{E}_{\underline{f} \sim \Pi_n}\left[\xi(\mathbf{E}[L_{\underline{f}}])\right] \ \trianglelefteq_{\frac{\eta \cdot n}{2 c_{u+\epsilon}}} \ c_{u+\epsilon}\left(\mathrm{IC}_{n,\eta}(\Pi_|) + \epsilon\right) \tag{86}$$

$$\textit{under v-PPC:} \quad \mathbf{E}_{Z_1^n}\left[\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\xi(\mathbf{E}[L_{\underline{f}}])\right]\right] \leq c'_{2u}\left(\mathbf{E}_{Z_1^n}\left[\mathrm{IC}_{n,\eta}(\Pi_|)\right] + \epsilon\right), \tag{87}$$

where $c_u := \frac{u}{c}\frac{\eta+1}{1 - \frac{\eta}{v(\epsilon)}}$ and $c'_{2u} := \frac{u}{c}\frac{2\eta+1}{1-\frac{2\eta}{v(\epsilon)}}$ and $\xi(\mathbf{E}[L_f]) = 1 \wedge \mathbf{E}[L_f]$.

Proof (86) follows by following essentially the same steps as in the proof of Theorem 22, but splitting the expectation in two parts:

$$\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\xi(\mathbf{E}[L_{\underline{f}}])\right] = \mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{1}_{\{\mathbf{E}[L_{\underline{f}}] < 1\}} \cdot \mathbf{E}[L_{\underline{f}}]\right] + \mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{1}_{\{\mathbf{E}[L_{\underline{f}}] \geq 1\}} \cdot 1\right]. \tag{88}$$

Fix some $\epsilon \geq 0$. The first term on the right of (88) can be bounded as follows, using Lemma 21 and the fact that a (u, c)-witness condition is assumed for f with $\mathbf{E}[L_f] < 1$ in combination with (83) and the fact that for $c > 0$ and general random variables U, V, we have $U \trianglelefteq_a V \Leftrightarrow cU \trianglelefteq_{a/c} cV$:

$$\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{1}_{\{\mathbf{E}[L_{\underline{f}}] < 1\}} \cdot \mathbf{E}[L_{\underline{f}}]\right] \trianglelefteq_{\eta n / c_{u+\epsilon}} \ c_{u+\epsilon} \cdot \left(\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\frac{1}{n}\sum_{j=1}^{n}(\ell_{\underline{f}}(Z_j) - \ell_{f^*_\epsilon}(Z_j))\right] + \frac{\mathrm{KL}(\Pi_n \,\|\, \Pi_0)}{\eta n}\right).$$

The second term on the right of (88) can similarly be bounded, using that $\tau(\mathbf{E}[L_f]) = u\,\mathbf{E}[L_f]$ for all f with $\mathbf{E}[L_f] \geq 1$:

$$\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\mathbf{1}_{\{\mathbf{E}[L_{\underline{f}}]\geq 1\}} \cdot \frac{\mathbf{E}[L_{\underline{f}}]}{\mathbf{E}[L_{\underline{f}}]}\right] \trianglelefteq_{\eta n/B}$$

$$B \cdot \left(\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\frac{1}{n}\sum_{j=1}^{n}(\ell_{\underline{f}}(Z_j) - \ell_{f^*_\epsilon}(Z_j))\right] + \frac{\mathrm{KL}(\Pi_n \,\|\, \Pi_0)}{\eta n}\right),$$

where $B = \sup_{f:\mathbf{E}[L_f]\geq 1} c_{u\,\mathbf{E}[L_f]+\epsilon}/\,\mathbf{E}[L_f]$. The result (86) now follows by adding the two terms using Proposition 3 and bounding B by using that $c_{u\cdot a+\epsilon}/a \leq c_{u+\epsilon}$ for $a \geq 1$.

(87) follows in similar fashion, by repeating the proof of Theorem 29, but again splitting the expectation of $\xi(L_f)$ in two parts, just like above; we omit the details. ∎

Proof (of Theorem 31) We start by establishing the key inequality (90) below both under the v-central and the v-PPC condition, but with different values for r_n in (90). For this, we invoke Lemma 41. This gives that the v-PPC condition implies, via (87) and Markov's inequality, that for all $\delta \geq 0$, with probability at least $1-\delta$,

$$\mathbf{E}_{\underline{f}\sim\Pi_n}\left[\xi(\mathbf{E}[L_{\underline{f}}])\right] \leq r_n, \tag{89}$$

where $r_n = \frac{c'_{2u}}{\delta} \cdot (\mathbf{E}\left[\mathrm{IC}_{n,\eta_n}\right] + \epsilon_n)$.

On the other hand, under the v-central condition, (86) holds and via Proposition 3 we can turn it into a high probability bound. Combining this bound with (54) via a standard union bound argument gives that, for all $\delta > 0$, with probability at least $1-\delta$, (89) holds, with ξ as before but now with $r_n = c_{u+\epsilon_n} C_{n,\delta}\left(\mathbf{E}\left[\overline{\mathrm{IC}}_{n,\eta_n}\right] + \epsilon_n + \frac{2}{n\eta_n}\right)$. Rewriting (89) gives that, with probability at least $1-\delta$,

$$\Pi_n\left(\left\{\underline{f} : \mathbf{E}[L_{\underline{f}}] \geq 1\right\}\right) + \mathbf{E}_{\underline{f}\sim\Pi_n}\left[\mathbf{1}_{\{\mathbf{E}[L_{\underline{f}}]<1\}} \cdot \mathbf{E}[L_{\underline{f}}])\right] \leq r_n. \tag{90}$$

Part 1, Deterministic Estimators. For deterministic $\Pi_| \equiv (\hat{f}, \Pi_0)$,
(90) simplifies to $\mathbf{1}_{\{\mathbf{E}[L_{\hat{f}}]\geq 1\}} + \mathbf{1}_{\{\mathbf{E}[L_{\hat{f}}]<1\}} \cdot \mathbf{E}[L_{\hat{f}}] \leq r_n$, which further implies that with probability at least $1-\delta$, simultaneously,

$$\mathbf{1}_{\{\mathbf{E}[L_{\hat{f}}]\geq 1\}} \leq r_n \text{ and } \mathbf{1}_{\{\mathbf{E}[L_{\hat{f}}]<1\}} \cdot \mathbf{E}[L_{\hat{f}}] \leq r_n, \tag{91}$$

and both the result for the v-PPC condition (53) and the v-central condition (55) follow by noting that we may assume n large enough so that $r_n < 1$, so that (91) is logically equivalent to

$$\mathbf{E}[L_{\hat{f}}] < 1 \text{ and } \mathbf{1}_{\{\mathbf{E}[L_{\hat{f}}]<1\}} \cdot \mathbf{E}[L_{\hat{f}}] \leq r_n,$$

which in turn is equivalent to $\mathbf{E}[L_{\hat{f}}] \leq r_n$, and thus the results are implied.

Part 2, General Learning Algorithms. Here we assume the v-PPC condition, so we can use (90) with r_n as in the v-PPC case.

By Markov's inequality, for any sequence $b_1, b_2, \ldots$ of positive numbers tending to ∞,

$$\begin{aligned} \Pi_n\left(\left\{f \in \mathcal{F} : 1 > \mathbf{E}[L_f] > b_n r_n\right\}\right) &= \Pi_n\left(\mathbf{1}_{\{\mathbf{E}[L_{\underline{f}}]<1\}} \cdot \mathbf{E}[L_{\underline{f}}] > b_n r_n\right) \\ &\leq \frac{\mathbf{E}_{\underline{f} \sim \Pi_n}\left[\mathbf{1}_{\{\mathbf{E}[L_{\underline{f}}]<1\}} \cdot \mathbf{E}[L_{\underline{f}}]\right]}{b_n r_n}. \end{aligned}$$

Combining this with (90) (dropping the leftmost term in that inequality) gives that with probability at least $1 - \delta$,

$$\Pi_n\left(\left\{f \in \mathcal{F} : 1 > \mathbf{E}[L_f] > b_n r_n\right\}\right) \leq \frac{1}{b_n}.$$

Combining this again with (90), now dropping the second term in the inequality and using a standard union bound, gives that with probability at least $1 - 2\delta$,

$$\Pi_n\left(\left\{f \in \mathcal{F} : \mathbf{E}[L_f] > b_n r_n\right\}\right) \leq \frac{1}{b_n} + r_n,$$

which, plugging in the definition of r_n and $\overline{\mathrm{IC}}_{n,\eta}$ on the left, can be rewritten as, for each n, each δ, with a_n as in the theorem statement:

$$\text{With probability} \geq 1 - 2\delta\text{: } \Pi_n\left(\left\{f \in \mathcal{F} : \mathbf{E}[L_f] > \frac{b_n}{a_n} \cdot \frac{c'_{2u}}{\delta} \cdot \left(\mathbf{E}[\overline{\mathrm{IC}}_{n,\eta} + \epsilon_n]\right)\right\}\right) \leq \frac{1}{b_n} + r_n. \quad (92)$$

Now choose $\delta = 1/\sqrt{a_n} \to 0$ as a function of n, and choose $b_n = \sqrt{a_n} \to \infty$. Then (92) implies the result. ∎

F.3. Proof of Proposition 32

Proof Let c, u and τ be as in the statement of the proposition. For each $f \in \mathcal{F}$, we will define modified predictors f', defined in terms of their losses $\ell_{f'}$ so that for all such f', we have

$$\mathbf{E}[\left(\ell_{f'} - \ell_{f^*}\right) \cdot \mathbf{1}_{\{\ell_{f'} - \ell_{f^*} \leq u'\}}] \geq 0, \text{ for } u' = u \cdot \left(\frac{\mathbf{E}[\ell_{f^*}]}{c} \vee 1\right), \quad (93)$$

which allows us to apply Proposition 26 to the set of f'; we will also ensure that for all $z \in \mathcal{Z}$,

$$\ell_{f'}(z) \leq \ell_f(z) \text{ and } \ell_{(f^*)'}(z) = \ell_{f^*}(z), \quad (94)$$

which will allow us to apply Proposition 27 so that results for f' transfer to the original f. Once we have shown (93) and (94), the result follows.

Case 1: $\mathbf{E}[L_f] \leq (\mathbf{E}[\ell_{f^*}]/c) \vee 1$. For all f with $\mathbf{E}[L_f] \leq (\mathbf{E}[\ell_{f^*}]/c) \vee 1$ (including f^*), we simply set $f' = f$. Then (94) holds trivially. To see that (93) holds, note that the assumed τ-witness condition holds for $\tau(\mathbf{E}[L_f]) = u(1 \vee \mathbf{E}[L_f]) \leq u(1 \vee (\mathbf{E}[\ell_{f^*}]/c \vee 1))$, which is no larger than the u' mentioned in (93), which then immediately follows by the assumed witness condition.

Case 2: $\mathbf{E}[L_f] > (\mathbf{E}[\ell_{f^*}]/c) \vee 1$. For these f, we define

$$\ell_{f'}(z) = \begin{cases} \ell_f(z) & \text{if } \ell_f(z) \leq \ell_{f^*}(z) \\ \frac{\ell_f(z) - \ell_{f^*}(z)}{c'} + \ell_{f^*}(z) & \text{if } \ell_f(z) > \ell_{f^*}(z), \end{cases}$$

with $c' := \mathbf{E}[L_f]/(\mathbf{E}[\ell_{f^*}/c] \vee 1)$, which by construction must satisfy $c' > 1$. This implies after rearranging terms that (94) holds. It thus remains to prove (93). To see that it holds, first note that $\ell_{f'} > \ell_{f^*} \Leftrightarrow \ell_f > \ell_{f^*}$ and that $\ell_f \geq 0$ on all z. Using these facts we find that:

$$\begin{aligned}
&\mathbf{E}\big[\big(\ell_{f'} - \ell_{f^*}\big) \cdot \mathbf{1}_{\{\ell_{f'} - \ell_{f^*} \leq u'\}}\big] \\
&\geq -\mathbf{E}\big[\mathbf{1}_{\{\ell_{f'} - \ell_{f^*} \leq 0\}}\, \ell_{f^*}\big] + \mathbf{E}\big[\big(\ell_{f'} - \ell_{f^*}\big) \cdot \mathbf{1}_{\{\ell_{f'} - \ell_{f^*} > 0\}} \cdot \mathbf{1}_{\{\ell_{f'} - \ell_{f^*} \leq u'\}}\big] \\
&\geq -\mathbf{E}[\ell_{f^*}] + \mathbf{E}\big[\big(\ell_{f'} - \ell_{f^*}\big) \cdot \mathbf{1}_{\{\ell_f > \ell_{f^*}\}} \cdot \mathbf{1}_{\{\ell_{f'} - \ell_{f^*} \leq u'\}}\big] \\
&= -\mathbf{E}[\ell_{f^*}] + \mathbf{E}\left[\left(\frac{\ell_f - \ell_{f^*}}{c'}\right) \cdot \mathbf{1}_{\{\ell_f > \ell_{f^*}\}} \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u'c'\}}\right] \\
&= -\mathbf{E}[\ell_{f^*}] + \frac{1}{c'}\, \mathbf{E}\big[\big(\ell_f - \ell_{f^*}\big) \cdot \mathbf{1}_{\{\ell_f > \ell_{f^*}\}} \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u'\, \mathbf{E}[L_f]/((\mathbf{E}[\ell_{f^*}]/c) \vee 1)\}}\big] \\
&= -\mathbf{E}[\ell_{f^*}] + \frac{1}{c'}\, \mathbf{E}\big[\big(\ell_f - \ell_{f^*}\big) \cdot \mathbf{1}_{\{\ell_f > \ell_{f^*}\}} \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u\, \mathbf{E}[L_f]\}}\big]. \\
&= -\mathbf{E}[\ell_{f^*}] + \frac{1}{c'}\, \mathbf{E}\big[\big(\ell_f - \ell_{f^*}\big) \cdot \mathbf{1}_{\{\ell_f > \ell_{f^*}\}} \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u(\mathbf{E}[L_f] \vee 1)\}}\big]. && (95) \\
&\geq -\mathbf{E}[\ell_{f^*}] + \frac{1}{c'}\, \mathbf{E}\big[\big(\ell_f - \ell_{f^*}\big) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u(\mathbf{E}[L_f] \vee 1)\}}\big] \\
&\geq -\mathbf{E}[\ell_{f^*}] + \frac{1}{c'} \cdot c\, \mathbf{E}[L_f] \geq -\mathbf{E}[\ell_{f^*}] + \mathbf{E}[\ell_{f^*}] = 0, && (96)
\end{aligned}$$

where (95) follows because all f's we consider here have $\mathbf{E}[L_f] > 1$ and (96) follows by our assumption of the τ-witness condition.

■

Appendix G. The Existence of the Generalized Reversed Information Projection

Recall that $\mathcal{E}_{\mathcal{F},\eta}$ is the the entropification-induced set $\left\{e^{-\eta \ell_f} : f \in \mathcal{F}\right\}$. In this section, we prove the existence of the generalized reversed information projection $m_{\mathcal{F}}^{\eta}$ of P onto $\operatorname{conv}(\mathcal{E}_{\mathcal{F},\eta})$. Because $\mathcal{F}$ and η are fixed throughout, we adopt the notation $\mathcal{E} := \mathcal{E}_{\mathcal{F},\eta}$ and $\mathcal{C} := \operatorname{conv}(\mathcal{E}_{\mathcal{F},\eta})$.

Formally, we will show that there exists q^* (not necessarily in $\mathcal{C}$) satisfying

$$\mathbf{E}[-\log q^*(Z)] = \inf_{q \in \mathcal{C}} \mathbf{E}[-\log q(Z)].$$

One might think that there is an easy proof by simply taking q^* to lie in the closure of $\mathcal{C}$ under some appropriate topology, but it is not evident what topology to take. For example, even in the simple case with $\eta = 1$ and ℓ_f is the log-loss so that $\mathcal{E}$ and $\mathcal{C}$ are sets of probability densities, it may happen that q^* is a sub-density (integrating to less than 1) (Li, 1999) so that it would not lie in the closure of any standard topology which we may impose on $\mathcal{C}$. We thus follow a different approach. We first rewrite the above in the language of information geometry. To provide easier comparison to Li (1999) we use the following modified KL notation here for a generalized KL divergence, which in particular makes the underlying distribution P explicit:

$$\mathrm{KL}(p; q_0 \,\|\, q) := \mathbf{E}_{Z \sim P}\left[\log \frac{q_0(Z)}{q(Z)}\right],$$

where q_0 and q are nonnegative but neither need be a normalized probability density. Then the existence question above is equivalent to the existence of q^* such that

$$\mathrm{KL}(p; q_0 \,\|\, q^*) = \inf_{q \in \mathcal{C}} \mathrm{KL}(p; q_0 \,\|\, q);$$

here, the only restriction on q_0 is that $\mathbf{E}_{Z \sim P}[\log q_0]$ be finite.

Now, Li (1999) already showed the above in the case of density estimation with log loss, $\eta = 1$, and $q_0 = p$; in that setting, we have $e^{-\eta \ell_f} = f$, and so mixtures of elements of $\mathcal{E}$ correspond to mixtures of probability distributions in $\mathcal{F}$. Hence, our setting is more general, yet Li's argument (with minor adaptations) still works. To be sure, we go through his argument step-by-step and show that it all still works in our setting.

In the remainder of this section, we treat two cases simultaneously unless a separate treatment is indicated: the case when the loss is uniformly bounded from below (as in Appendix H.1) and the case of log loss (with the loss not uniformly bounded from below, as in Appendix H.2). In the former case, we always take $q_0 = e^{-\eta \ell_{f^*}}$. In the latter case, we always take $q_0 = p$.

G.1. Proving q^* Exists

Throughout, we will need to assume the existence of a certain sequence $(q_n)_{n \geq 1}$ in $\mathcal{C}$, satisfying $\mathrm{KL}(p; q_0 \,\|\, q_n) \to \inf_{q \in \mathcal{C}} \mathrm{KL}(p; q_0 \,\|\, q)$, for which $\mathrm{KL}(p; q_0 \,\|\, q_n)$ is finite for all n. This is not problematic, as we now explain. We treat separately the case of losses uniformly bounded from below and the case of log loss without a uniform lower bound on the loss.

Losses uniformly bounded from below. First, observe that for any $q_n \in \mathcal{C}$,

$$\mathrm{KL}(p; q_0 \,\|\, q_n) \geq -\|\ell_-\|_\infty - \mathbf{E}[\ell_{f^*}] > -\infty.$$

To see this, observe that $q_n = \mathbf{E}_{f \sim R_n}[e^{-\eta \ell_f}]$ for some distribution $R_n \in \Delta(\mathcal{F})$; then assumption (98) gives the first inequality. The second inequality holds because we only deal with non-trivial learning problems, and so f^* obtains risk less than $+\infty$. Next, since the particular choice $q_n = e^{-\eta \ell_{f^*}}$ yields $\mathrm{KL}(p; q_0 \,\|\, q_n) = 0$, we may always restrict to sequences for which we have $\mathrm{KL}(p; q_0 \,\|\, q_n) < \infty$ for all n. Hence, we indeed can take the sequence satisfying the finiteness requirement.

Log loss. First, we show for any q_n that $\mathrm{KL}(p; q_0 \,\|\, q_n)$ is well-defined; its well-definedness is not immediately clear since each q_n need not be a probablity density. For convenience, we introduce the notation that, for any n, the distribution R_n satisfies $q_n = \mathbf{E}_{f \sim R_n}[e^{-\eta \ell_f}]$. Therefore, $-\log q_n = m^{\eta}_{R_n}$.

Now, defining the pseudo-loss $\ell_p(Z) = -\log p(Z)$ corresponding to playing the pseudo-action p, our present goal is to show that $\mathbf{E}\left[m^{\eta}_{R_n} - \ell_P\right]$ is well-defined for each j. To this end, we make the following claim:

$$\mathbf{E}_{Z\sim P}\left[\left(m^{\eta}_{R_n}(Z) - \ell_p(Z)\right)^{-}\right] > -\frac{1}{\eta}\log 2. \tag{97}$$

To see the claim, define for $f \in \mathcal{F}$ the excess loss $\ell_{f,p}(Z) \coloneqq \ell_f(Z) - \ell_p(Z)$ and observe that (we simplify by writing R instead of R_n)

$$\begin{aligned}
&\mathbf{E}_{Z\sim P}\left[\left(m^{\eta}_{R} - \ell_p\right)^{-}\right] \\
&= \mathbf{E}_{Z\sim P}\left[-\frac{1}{\eta}\log \mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]\cdot \mathbf{1}_{\left\{\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]>e\right\}}\right] \\
&= \frac{1}{\eta}\mathbf{E}_{Z\sim P}\left[-\log\left(\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]\cdot \mathbf{1}_{\left\{\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]>e\right\}} + \mathbf{1}_{\left\{\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]\le e\right\}}\right)\right] \\
&\ge -\frac{1}{\eta}\log \mathbf{E}_{Z\sim P}\left[\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]\cdot \mathbf{1}_{\left\{\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]>e\right\}} + \mathbf{1}_{\left\{\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]\le e\right\}}\right] \\
&\ge -\frac{1}{\eta}\log \mathbf{E}_{Z\sim P}\left[\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]\cdot \mathbf{1}_{\left\{\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]>e\right\}} + 1\right],
\end{aligned}$$

where Jensen's inequality was applied for the first inequality. It remains to show that

$$\mathbf{E}_{Z\sim P}\left[\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]\cdot \mathbf{1}_{\left\{\mathbf{E}_{f\sim R}\left[e^{-\eta\ell_{f,p}(Z)}\right]>e\right\}}\right] < \infty.$$

Rewriting the LHS, we have

$$\begin{aligned}
\mathbf{E}_{Z\sim P}\left[\mathbf{E}_{f\sim R}\left[\left(\frac{p_f}{p}\right)^{\eta}\right]\cdot \mathbf{1}_{\left\{\left(\frac{p_f}{p}\right)^{\eta}>e\right\}}\right] &\le \mathbf{E}_{Z\sim P}\left[\mathbf{E}_{f\sim R}\left[\left(\frac{p_f}{p}\right)^{\eta}\right]\right] \\
&\le \left(\mathbf{E}_{Z\sim P}\left[\mathbf{E}_{f\sim R}\left[\frac{p_f}{p}\right]\right]\right)^{\eta} \\
&= 1,
\end{aligned}$$

where the inequality follows from $\eta \le 1$, the concavity of the map $x \mapsto x^{\eta}$, and Jensen's inequality. The claim thus follows.

Now that we have shown that $\mathrm{KL}(p; q_0 \,\|\, q_n)$ is well-defined for all n, we also conclude from assumption (8) that we may always take a sequence such that $\mathrm{KL}(p; q_0 \,\|\, q_n) < \infty$ for all n. Moreover, from (97), this can be strengthened to $\mathrm{KL}(p; q_0 \,\|\, q_n) \in [-\eta^{-1}\log 2, \infty)$, and so this quantity is finite as desired.

In the remainder of this section, the two cases of loss assumptions are treated simulataneously (recall that q_0 is defined differently for each).

Step 1: Existence of minimizer $\bar{q}_n$ in convex hull of finite sequence

Let $(q_n)_{n\ge 1}$ be a sequence in $\mathcal{C}$ for which $\mathrm{KL}(p; q_0 \,\|\, q_n) \to \inf_{q\in\mathcal{C}} \mathrm{KL}(p; q_0 \,\|\, q)$. From the argument above we may restrict the sequence to one for which $\mathrm{KL}(p; q_0 \,\|\, q_n)$ is finite for all n. Take $\mathcal{C}_n$ to be $\mathrm{conv}(\{q_1, \ldots, q_n\})$.

We introduce the representation $D(t) : \Delta^{n-1} \to \mathbb{R}_+$, where $D(t) = \mathrm{KL}(p; q_0 \,\|\, q_t)$ with $q_t = \sum_{j=1}^n t_j q_j$.

The first claim is that $t \mapsto D(t)$ is a continuous function. Li's Lemma 4.2 proves continuity of D when $q_0 = p$, $\mathrm{KL}(p \,\|\, q_i) < \infty$ for $i \in [n]$ and each q_i is a probability distribution. However, inspection of the proof reveals that the result still holds for general q_0 and when both q_0 and q_i are only pseudoprobability densities, as long as we still have $\mathrm{KL}(p; q_0 \,\|\, q_i) < \infty$ for $i \in [n]$. But we already have established the latter requirement, and so D is indeed continuous. Since D also has compact domain, it follows that D is globally minimized by an element in $\mathcal{C}_n$. Call this element $\bar{q}_n$.

Step 2: Beneficial properties of minimizer $\bar{q}_n$

We claim for all $q \in \mathcal{C}_n$ that $\int p \frac{q}{\bar{q}_n} \leq 1$. This follows from a suitably adapted version of Li's Lemma 4.1. First, we observe that even though Li's Lemma 4.1 is for the case of the KL divergence $\mathrm{KL}(p \,\|\, q) = \int p \log \frac{p}{q}$, changing the $\log p$ term to $\log q_0$ has no effect on the proof. Therefore, this result also works for $\mathrm{KL}(p; q_0 \,\|\, q)$. Next, the proof works without modification even when its q^* and q are only pseudoprobability densities. To apply Li's Lemma 4.1, *mutatis mutandis*, we instantiate its $\mathcal{C}$ as $\mathcal{C}_n$, its p as p, its q as q, and its q^* as $\bar{q}_n$.

Step 3: $(\log \bar{q}_n)_n$ is Cauchy sequence in $L_1(P)$

We can find a sequence $(\bar{q}_n)_{n \geq 1}$ such that $\{\mathrm{KL}(p; q_0 \,\|\, \bar{q}_n)\}$ both is non-increasing and converges to $\inf_{q \in \mathcal{C}} \mathrm{KL}(p \,\|\, q)$.

Next, let $n \leq m$ throughout the rest of this step and observe that

$$\mathrm{KL}(p; q_0 \,\|\, \bar{q}_n) - \mathrm{KL}(p; q_0 \,\|\, \bar{q}_m) = \int p \log \frac{p}{\frac{p\bar{q}_n}{\bar{q}_m} / c_{m,n}} + \log \frac{1}{c_{m,n}}$$

with $c_{m,n} := \int \frac{p\bar{q}_n}{\bar{q}_m}$.

Now, due to the normalization by $c_{m,n}$ the first term on the RHS is a KL divergence and hence nonnegative. Also, since $c_{m,n} \leq 1$, the second term also is nonnegative.

Next, observe that $\mathrm{KL}(p; q_0 \,\|\, \bar{q}_n) - \mathrm{KL}(p; q_0 \,\|\, \bar{q}_m) \to 0$ as $n, m \to \infty$, and so we have

$$\int p \log \frac{p}{\frac{p\bar{q}_n}{\bar{q}_m} / c_{m,n}} = \mathrm{KL}\left(p \,\|\, \frac{p\bar{q}_n}{\bar{q}_m} / c_{m,n}\right) \to 0$$

as well as

$$\log \frac{1}{c_{m,n}} \to 0 \quad \Rightarrow \quad c_{m,n} \to 1.$$

Next, we apply the following inequality due to Barron/Pinsker, holding for any probability distributions p_1 and p_2:

$$\int p_1 |\log(p_1) - \log(p_2)| \leq \mathrm{KL}(p_1 \,\|\, p_2) + \sqrt{2\mathrm{KL}(p_1 \,\|\, p_2)}.$$

This yields

$$\int p \left| \log \frac{p}{\frac{p\bar{q}_n}{\bar{q}_m} / c_{m,n}} \right| \to 0.$$

Since $c_{m,n} \to 1$, it therefore follows that

$$\int p | \log(\bar{q}_n) - \log(\bar{q}_m) | \to 0.$$

Therefore $(\log(\bar{q}_n))_{n \geq 1}$ is a Cauchy sequence in $L_1(P)$, and from the completeness of this space, $\log(\bar{q}_n)$ converges to some $\log(q^*) \in L_1(P)$.

Finally, we observe that $\mathrm{KL}(p; q_0 \parallel q^*) = \lim_{n \to \infty} \mathrm{KL}(p; q_0 \parallel \bar{q}_n)$ since

$$\begin{aligned} \mathrm{KL}(p; q_0 \parallel q^*) - \lim_{n \to \infty} \mathrm{KL}(p; q_0 \parallel \bar{q}_n) &= \lim_{n \to \infty} \int p(\log \bar{q}_n - \log q^*) \\ &\leq \lim_{n \to \infty} \int p | \log \bar{q}_n - \log q^* | \\ &= 0. \end{aligned}$$

Appendix H. Definitions and Conventions Concerning ∞ and $-\infty$

For general losses we allow the loss to take on the value ∞, and for density estimation under log loss we allow the loss to take on the value ∞ and to be unbounded from below; see Appendix H.2 for a full description of our assumptions in this latter setting. We thus need to take care to avoid ambiguous expressions such as $\infty - \infty$; here we follow the approach of Grünwald and Dawid (2004). We generally permit operations on the extended real line $[-\infty, \infty]$, with definitions and exceptions as in (Rockafellar, 1970, Section 4). For a given distribution P on some space $\mathcal{U}$ with associated σ-algebra, we define the *extended random variable* U as any measurable function $U : \mathcal{U} \to \mathbb{R} \cup \{-\infty, \infty\}$. We say that U is *well-defined* if either $P(U = \infty) = 0$ or $P(U = -\infty) = 0$. Now let U be a well-defined extended random variable. For any function $f : [-\infty, \infty] \to [-\infty, \infty]$, we say that $f(U)$ is well-defined if either $P(f(U) = \infty) = 0$ or $P(f(U) = -\infty) = 0$ and we abbreviate the expectation $\mathbf{E}_{U \sim P}[f(U)]$ to $\mathbf{E}[f]$, hence we think of f as an extended random variable itself. If f is bounded from below and above $\mathbf{E}[f]$ is defined in the usual manner. Otherwise we interpret $\mathbf{E}[f]$ as $\mathbf{E}[f^+] + \mathbf{E}[f^-]$ where $f^+(u) := \max\{f(u), 0\}$ and $f^-(u) := \min\{f(u), 0\}$, allowing either $\mathbf{E}[f^+] = \infty$ or $\mathbf{E}[f^-] = -\infty$, but not both. In the first case, we say that $\mathbf{E}[f]$ is well-defined; in the latter case, $\mathbf{E}[f]$ is undefined. In the remainder of this section we introduce conditions under which all extended random variables and all expectations occurring in the main text are always well-defined.

The quantities which we need to show to be well-defined, both in the case of general losses and log loss, are (i) the risk for deterministic estimators; (ii) the risk for randomized estimators; (iii) the excess risk for either deterministic or randomized estimators; and (iv) certain ESIs and posterior expectations of annealed expectations. The GRIP is handled separately in Appendix G.

H.1. When the Loss is Uniformly Bounded from Below (General Losses)

Here, we show that the relevant expressions are well-defined when the loss is uniformly bounded from below.

Risk for deterministic/randomized estimators and relevant comparators

We first show that the risk of any deterministic estimator is well-defined. Our assumption that the loss is uniformly bounded from below is equivalent to the existence of a finite constant $\|\ell_-\|_\infty$ for which

$$\inf_{f \in \mathcal{F}} \inf_{z \in \mathcal{Z}} \ell_f(Z) \geq -\|\ell_-\|_\infty. \tag{98}$$

We thus have for any $f \in \mathcal{F}$ that $\mathbf{E}_{Z \sim P}[(\ell_f(Z))^-] > -\infty$, and so the risk $\mathbf{E}_{Z \sim P}[\ell_f(Z)]$ is well-defined. Moreover, since $\inf_{f \in \mathcal{F}} \mathbf{E}[\ell_f(Z)] > -\infty$, we also have that for any distribution Π on $\mathcal{F}$ that $\mathbf{E}_{f \sim \Pi}\left[\mathbf{E}_{Z \sim P}[\ell_f(Z)]\right]$ is well-defined.

For all comparators $\tilde{f}$ used in this paper, assumption (98) also implies that

$$\inf_{z \in \mathcal{Z}} \ell_{\tilde{f}}(Z) > -\infty.$$

To see this, observe that the only comparators we use from the set $\bar{\mathcal{F}} \setminus \mathcal{F}$ are GRIPs (which for a given $z \in \mathcal{Z}$ cannot obtain loss lower than $\inf_{f \in \mathcal{F}} \ell_f(z)$) and versions of the loss of a GRIP or some $f \in \mathcal{F}$ that are shifted by a finite constant. Thus, the risk is well-defined for all comparators used in this paper.

Excess risk for randomized estimators

Next, the excess risk of any randomized estimator relative to a non-trivial comparator also is well-defined, since, by definition of a non-trivial comparator $\tilde{f}$ and the uniformly-bounded-below assumption, we have $-\infty < \mathbf{E}_{Z \sim P}\left[\ell_{\tilde{f}}(Z)\right] < \infty$.

ESI / Posterior-expectation of annealed expectations

Finally, we verify that all ESIs and annealed expectations of excess losses also are well-defined. The relevant quantities are (for all non-trivial comparators $\tilde{f}$)

$$\mathbf{E}_{Z \sim P}\left[e^{\eta\left(\ell_{\tilde{f}}(Z) - \ell_f(Z)\right)}\right] \quad \text{for all } f \in \mathcal{F} \tag{99}$$

and

$$\mathbf{E}_{f \sim Q}\left[-\frac{1}{\eta} \log \mathbf{E}_{Z \sim P}\left[e^{\eta\left(\ell_{\tilde{f}}(Z) - \ell_f(Z)\right)}\right]\right] \quad \text{for all } Q \in \Delta(\mathcal{F}). \tag{100}$$

A potential issue with the ESI (99) being well-defined is that we can have both $\ell_{\tilde{f}}(z) = +\infty$ and $\ell_f(z) = +\infty$ for all z in some set $A \subset \mathcal{Z}$ of P-measure zero. To show that the expectation is well-defined, we define for $j = 1, 2, \ldots$ the random variable

$$g_j(Z) = \exp\left(\eta\left(\left(j \wedge \ell_{\tilde{f}}(Z)\right) - \ell_f(Z)\right)\right).$$

Now, for each $j = 1, 2, \ldots$, the expectation $\mathbf{E}[g_j(Z)]$ is well-defined. Moreover, letting A be precisely the subset of $\mathcal{Z}$ for which $\ell_{\tilde{f}}(z) = +\infty$, it holds that $\{g_j\}$ converges to $\exp\left(\eta(\ell_{\tilde{f}} - \ell_f)\right)$ pointwise on $\mathcal{Z} \setminus A$. Hence, from Levi's monotone convergence theorem, $\mathbf{E}_{Z\sim P}\left[e^{\eta\left(\ell_{\tilde{f}}(Z) - \ell_f(Z)\right)}\right]$ is well-defined.

Next, we show that annealed expectations of the form (100) also are well-defined. From Hölder's inequality,

$$\begin{aligned}\mathbf{E}\left[e^{\eta(\ell_{\tilde{f}}(Z) - \ell_f(Z))}\right] &= \mathbf{E}\left[e^{\eta\ell_{\tilde{f}}(Z)} e^{-\eta\ell_f(Z)}\right] \\ &\le e^{\|\ell_-\|_\infty}\, \mathbf{E}\left[e^{\eta\ell_{\tilde{f}}(Z)}\right] \\ &< \infty,\end{aligned}$$

where the final inequality follows because $\ell_{\tilde{f}}(Z) < \infty$ with probability 1. Therefore, the negative logarithm of the above is lower bounded by a finite negative constant that is independent of $f \in \mathcal{F}$. It follows that (100) is well-defined.

H.2. Log Loss

In the common case of log loss with uncountable sample spaces, the loss is not always uniformly bounded from below; see Example 13 below for a concrete illustration. To allow for this case while avoiding issues with infinities we need to make the alternative assumptions of Section 2, which we now discuss. Recall that we assumed for all $f \in \mathcal{F}$ that p_f is absolutely continuous with respect to a common dominating measure μ, and that furthermore we have (8) and (9). To motivate these assumptions, observe that $H(P)$ is the Bayes risk with respect to all possible probability measures, whereas $\mathrm{KL}(P \,\|\, P_{f^*})$ is the approximation error due to playing the optimal in-model predictor f^* rather than P. Now, (8) is a reasonable requirement, as it simply means that the approximation error is finite; this is discussed further in Example 13. Now, if we have $H(P) = -\infty$, then in light of (8), we would also have to have $\mathbf{E}_{Z\sim P}\left[\ell_{f^*}(Z)\right] = -\infty$, which would imply that for any $f \in \mathcal{F}$ with $\mathbf{E}[\ell_f] \neq \mathbf{E}[\ell_{f^*}]$, the excess risk is infinite; this would make learning meaningless. We thus assume (9).[8]

Risk for deterministic estimators

Because for log loss we do not assume that losses are bounded from below, we need to ensure that the risk is well-defined.

We do this in two steps. First, we show that $\mathrm{KL}(P \,\|\, Q)$ is well-defined for any probability distribution Q with density q (with respect to μ). We do this by showing that $\mathbf{E}\left[\left(\log \frac{p}{q}\right)^-\right] >$

[8] A referee asked the natural question why we do not simply impose the more standard condition that $P \ll P_f$ for all $f \in \mathcal{F}$, thus avoiding use of differential entropy. But this is not sufficient, as explained below (101).

$-\infty$:

$$
\begin{aligned}
\mathbf{E}[\mathbf{1}_{\{q/p>1\}}(-\log q+\log p)] &= \mathbf{E}[-\log(\mathbf{1}_{\{q/p>1\}}\cdot(q/p)+\mathbf{1}_{\{q/p\le 1\}}\cdot 1)] \\
&\ge -\log \mathbf{E}[\mathbf{1}_{\{q/p>1\}}\cdot(q/p)+\mathbf{1}_{\{q/p\le 1\}}\cdot 1] \\
&\ge -\log 2,
\end{aligned}
$$

where the application of Jensen's inequality for the first inequality is legitimate because the expectation is of a nonpositive quantity. The above holds in particular for q set to any p_f (for $f \in \mathcal{F}$). Next, we use the decomposition

$$
\mathbf{E}[\ell_f] = \mathbf{E}[-\log p_f + \log p - \log p] = \mathrm{KL}(P \,\|\, Q) + H(P). \tag{101}
$$

Since the KL divergence term is nonnegative and $H(P) < -\infty$ (recall assumption (9)), the above is well-defined.

We note that it is not sufficient to replace (9) by the standard requirement that $P \ll P_f$ for all $f \in \mathcal{F}$, for then (101) may become undefined. To see this, note that, for two probability measures P and R, we may have $\mathrm{KL}(P \,\|\, R) = \infty$ even if $P \ll R$ (take, for example, P a distribution on $\mathbb{N}$ with mass function $p(i) \propto i^{-1-\alpha}$ for $0 < \alpha \le 1$ and R with mass function $r(i) = 2^{-i}$). Since $H(P)$ defined relative to base measure R is equal to $-\mathrm{KL}(P \,\|\, R)$ we may in general also have $H(P) = -\infty$ even if P has a density relative to R. Thus, without the requirement (9) we could have $\mathrm{KL}(P \,\|\, Q) + H(P) = \infty - \infty$ which is undefined.

Risk for randomized estimators

The above argument can be trivially modified (adding an outer expectation over $f \sim \Pi$ everywhere) to show that the risk of any randomized estimator Π is also well-defined.

Excess risk with respect to randomized estimators

Finally, because we only consider situations in this paper for which the GRIP obtains risk less than positive infinity, the excess risk of any Π with respect to the GRIP is well-defined; the same is true for the excess risk with respect to the comparator f^*, since we only consider situations where the risk of f^* is close to the risk of the GRIP.

ESI / Posterior-expectation of annealed expectations

Finally, we verify that all ESIs and annealed expectations of excess losses also are well-defined. The relevant quantities are (for all non-trivial comparators $\tilde{f}$)

$$
\mathbf{E}_{Z\sim P}\left[e^{\eta\left(\ell_{\tilde{f}}(Z)-\ell_f(Z)\right)}\right] \quad \text{for all } f \in \mathcal{F} \tag{102}
$$

and, taking the comparator to be the GRIP $m^{\eta}_{\mathcal{F}}$ as this is all that we require for annealed expectations in this paper,

$$
\mathbf{E}_{f\sim Q}\left[-\frac{1}{\eta}\log \mathbf{E}_{Z\sim P}\left[e^{\eta\left(m^{\eta}_{\mathcal{F}}(Z)-\ell_f(Z)\right)}\right]\right] \quad \text{for all } Q \in \Delta(\mathcal{F}). \tag{103}
$$

A potential issue with the ESI (102) being well-defined is that we can have $\ell_{\tilde{f}}(z) = \ell_f(z) = +\infty$ or $\ell_{\tilde{f}}(z) = \ell_f(z) = -\infty$ for all z in some set $A \subset \mathcal{Z}$ of P-measure zero. To show that the expectation is well-defined, we define for $j = 1, 2, \ldots$ the random variable

$$g_j(Z) = \exp\left(\eta\left(\left[j \wedge \ell_{\tilde{f}}(Z)\right] - \left[(-j) \vee \ell_f(Z)\right]\right)\right).$$

Now, for each $j = 1, 2, \ldots$, the expectation $\mathbf{E}[g_j(Z)]$ is well-defined. Moreover, letting A be precisely the subset of $\mathcal{Z}$ for which either $\ell_{\tilde{f}}(z) = +\infty$ or $\ell_f(z) = -\infty$, it holds that $\{g_j\}$ converges to $\exp\left(\eta(\ell_{\tilde{f}} - \ell_f)\right)$ pointwise on $\mathcal{Z} \setminus A$. Hence, from Beppo Levi's monotone convergence theorem, $\mathbf{E}_{Z\sim P}\left[e^{\eta\left(\ell_{\tilde{f}}(Z) - \ell_f(Z)\right)}\right]$ is well-defined.

Finally, we verify that (103) is well-defined. Indeed, it is well-defined as a trivial consequence of $\mathbf{E}_{Z\sim P}\left[e^{\eta\left(m^{\eta}_{\mathcal{F}}(Z) - \ell_f(Z)\right)}\right] \le 1$ which holds by virtue of the comparator being the GRIP.

Example 13 (Density Estimation) Consider the Gaussian scale family with $\mathcal{Z} = \mathbb{R}$ and $\{p_f \mid f \in \mathcal{F}\}$ where $\mathcal{F} = \mathbb{R}^+$ and $p_f(y) \propto \exp(-y^2/2f)$, i.e., p_f is the density, relative to standard Lebesgue measure, of the normal distribution with mean 0 and variance $\sigma^2 := f$. Then under log loss we have $\ell_f(y) = \frac{y^2}{f} + \frac{1}{2}\log(\pi(f))$. Obviously, we do not want to rule out a model as standard like this, yet the loss is unbounded from below, which illustrates the need for treating log-loss separately from other loss functions. The requirements (8) and (9) above do allow for this model, as long as the underlying distribution P (a) has a density relative to Lebesgue measure (otherwise (9) does not hold); (b) is not too-heavy tailed (it needs to have a second moment, otherwise (8) does not hold), and (c) is not excessively peaked at 0 (for example, the probability distribution P on $(0, 1/\exp(1))$ with density $p(x) = 1/(x \cdot \log^2 x)$ has $H(P) = -\infty$, but distribution P' with density $p'(x) = 3/(x \cdot \log^4 x)$ has finite $H(P')$. If one restricts the model to contain only $f \ge \sigma_0^2$ for some $\sigma_0^2 > 0$, then the log loss is bounded from below, and the requirements (8) and (9) do not need to be imposed; in that situation, one could allow for an underlying distribution P with a point mass at some outcome, so that P does not have a density relative to Lebesgue measure and $D(P \| P_{f^*}) = \infty$, yet all our concepts remain well-defined. □

Appendix I. Comparative examples

Example 14 (Bernstein condition does not hold, bounded excess risk) Consider regression with squared loss, so that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Select P such that X and Y are independent. Let X follow the law P such that $P(X = 0) = P(X = 1) = \frac{a}{2}$, for $a := 2 - \frac{\pi^2}{6} \in (0, 1)$, and, for $j = 2, 3, \ldots$, $P(X = j) = \frac{1}{j^2}$. Let $Y = 0$ surely. Take as $\mathcal{F}$ the countable class $\{f_1, f_2, \ldots\}$ such that $f_1(1) = 0.5$ and f_1 is identically 0 for all other values of $x \in \mathcal{X}$; for each $j = 2, 3, \ldots$, the function f_j is defined as $f_j(0) = 1$, $f_j(j) = j$, and f_j takes the value 0 otherwise.

It follows that $f^* = f_1$, and for every $j > 1$ we have $\mathbf{E}[L_{f_j}] = \frac{3a}{8} + 1$. Thus, the excess risk is bounded for all f_j. The witness condition holds because for all $j > 1$ we have $\Pr(L_{f_j} = 1) = a$ and $\mathbf{E}[L_{f_j} \cdot \mathbf{1}_{\{L_{f_j} \le 1\}}] \ge \frac{3a}{8}$. Also, it is easy to verify that the strong central condition holds with $\eta = 2$. On the other hand, the Bernstein condition fails to hold in this example because $\mathbf{E}[L_{f_j}^2] = a + j^2 \to \infty$ as $j \to \infty$, while the excess risk is finite. In fact, even the variance of the excess risk is unbounded as $j \to \infty$, precluding the use of a weaker variance-based Bernstein condition as in equation (5.3) of Koltchinskii (2006). Therefore, Theorem 22 still applies while, e.g., the results of Zhang (2006b) and Audibert (2009) do not (see Section 7). □

Example 15 (Bernstein condition does not hold, unbounded excess risk) The setup of this example was presented in Example 5.7 of Van Erven et al. (2015) and is reproduced here for convenience. For f_μ the univariate normal density with mean μ and variance 1, let $\mathcal{P}$ be the normal location family and let $\mathcal{F} = \{f_\mu : \mu \in \mathbb{R}\}$ be the set of densities of the distributions in $\mathcal{P}$. Then, since the model is well-specified, for any $P \in \mathcal{P}$ with density f_ν we have $f^* = f_\nu$. As shown in Van Erven et al. (2015), the Bernstein condition does not hold in this example, although we note that the weaker, variance-based Bernstein condition of (Koltchinskii, 2006, equation (5.3)) does hold. However, we are not aware of any analyses that make use of the variance-based Bernstein condition in the unbounded excess losses regime.

Since the model is well-specified, the strong central condition holds with $\eta = 1$. Next, we show that the witness condition holds with $M = 2$, $u = 4$, and $c = 1 - \sqrt{\frac{2}{\pi}}$. From location-invariance, we assume $\nu > \mu = 0$ without loss of generality.

First, observe that the excess risk is equal to $\mathbf{E}[L_{f_\mu}] = \frac{1}{2}\nu^2$.

As $M = 2 < \infty$, the witness condition has two cases: the case of excess risk at least 2 and the case of excess risk below 2. We begin with the first case, in which $\nu \geq 1$. Then the contribution to the excess risk from the upper tail is

$$\begin{aligned}\mathbf{E}\left[L_{f_\mu} \cdot \mathbf{1}_{\{L_{f_\mu} > u\,\mathbf{E}[L_{f_\mu}]\}}\right] &= \mathbf{E}\left[\left(-\frac{\nu^2}{2} + X\nu\right) \cdot \mathbf{1}_{\{-\frac{\nu^2}{2} + X\nu > u\frac{\nu^2}{2}\}}\right] \\ &= \mathbf{E}\left[\left(-\frac{\nu^2}{2} + X\nu\right) \cdot \mathbf{1}_{\{X > \frac{u\nu}{2} + \frac{\nu}{2}\}}\right] \leq \nu\,\mathbf{E}\left[X \cdot \mathbf{1}_{\{X > \frac{u\nu}{2}\}}\right],\end{aligned}$$

which is at most

$$\begin{aligned}\nu\,\mathbf{E}\left[X \cdot \mathbf{1}_{\{X - \nu > (\frac{u}{2} - 1)\nu\}}\right] &= \nu \int_0^\infty \Pr(X \cdot \mathbf{1}_{\{X - \nu > (\frac{u}{2} - 1)\nu\}} > t)dt \\ &\leq \nu \frac{1}{\sqrt{2\pi}} \frac{e^{-(\frac{u}{2}-1)^2\nu^2/2}}{(\frac{u}{2} - 1)\nu} = \frac{1}{\sqrt{2\pi}} \frac{e^{-(\frac{u}{2}-1)^2\nu^2/2}}{(\frac{u}{2} - 1)}.\end{aligned}$$

Since $u = 4$, the above is at most $\frac{1}{\sqrt{2\pi}}$ and so, in this regime, the witness condition indeed is satisfied with $c = 1 - \sqrt{2/\pi}$.

Consider now the case of $\nu < 1$. In this case, the threshold simplifies to the constant u and the upper tail's contribution to the excess risk is

$$\begin{aligned}\mathbf{E}\left[L_{f_\mu} \cdot \mathbf{1}_{\{L_{f_\mu} > u\}}\right] &= \mathbf{E}\left[\left(-\frac{\nu^2}{2} + X\nu\right) \cdot \mathbf{1}_{\{-\frac{\nu^2}{2} + X\nu > u\}}\right] \\ &= \mathbf{E}\left[\left(-\frac{\nu^2}{2} + X\nu\right) \cdot \mathbf{1}_{\{X > \frac{u}{\nu} + \frac{\nu}{2}\}}\right] \leq \nu\,\mathbf{E}\left[X \cdot \mathbf{1}_{\{X > \frac{u}{\nu}\}}\right],\end{aligned}$$

which is at most

$$\nu \, \mathbf{E}\left[X \cdot \mathbf{1}_{\{X-\nu > \frac{u}{\nu}-\nu\}}\right] = \nu \int_0^\infty \Pr(X \cdot \mathbf{1}_{\{X-\nu > \frac{u}{\nu}-\nu\}} > t) dt$$

$$\leq \nu \frac{1}{\sqrt{2\pi}} \frac{e^{-(\frac{u}{\nu}-\nu)^2/2}}{\frac{u}{\nu}-\nu} = \nu^2 \frac{1}{\sqrt{2\pi}} \frac{e^{-(\frac{u}{\nu}-\nu)^2/2}}{u-\nu^2}.$$

Since $u = 4$ and $\nu < 1$, the above is at most $\frac{\nu^2}{\sqrt{18\pi}}$, and so the value of c from before still works and the witness condition holds in this regime as well. □

Example 16 (Small-ball Assumption Violated) To properly compare to the small-ball assumption of Mendelson (2014), we consider regression with squared loss in the well-specified setting, so that the parameter estimation error bounds of Mendelson (2014) directly transfer to excess loss bounds for squared loss. Take X and Y be independent. The distribution of X is defined as, for $j = 1, 2, \ldots$, $P(X = j) = p_j := \frac{1}{a} \cdot \frac{1}{j^2}$ for $a = \frac{\pi^2}{6}$. Let the distribution of Y be zero-mean Gaussian with unit variance. For the class $\mathcal{F}$, we take the following countable class of indicator functions: for each $j = 0, 1, 2, \ldots$, define $f_j(i) = \mathbf{1}_{\{i=j\}}$, for any positive integer i. Since $f_0(x) = \mathbf{E}[Y \mid X = x] = 0$ for all $x \in \{1, 2, \ldots\}$, we have $f^* = f_0$.

The small-ball assumption fails in this setting, since, for any constant $\kappa > 0$ and for all $j = 1, 2, \ldots$:

$$\Pr\left(|f_j - f^*| > \kappa \|f_j - f^*\|_{L_2(P)}\right) \leq \Pr\left(|f_j - f^*| > 0\right) = p_j = \frac{1}{aj^2} \to 0 \text{ as } j \to \infty.$$

On the other hand, the strong central condition holds with $\eta = \frac{1}{2}$, since, for all $j = 1, 2, \ldots$ and all x:

$$\mathbf{E}\left[e^{-\eta L_{f_j}}\right] = \mathbf{E}\left[\frac{e^{-\eta(f_j(x)-Y)^2}}{e^{-\eta Y^2}}\right] = \int \frac{\frac{1}{\sqrt{2\pi\eta^{-1}}} e^{-\eta(f_j(x)-Y)^2}}{\frac{1}{\sqrt{2\pi\eta^{-1}}} e^{-\eta Y^2}} p(Y) dy$$

which is equal to 1 for $\eta = \frac{1}{2}$, since $Y \sim \mathcal{N}(0, 1)$.

It remains to check the witness condition. Observe that, for each j, we have $\mathbf{E}[L_{f_j}] = p_j$.

Next, we study how much of the excess risk comes from the upper tail, above some threshold u:

$$\begin{aligned}\mathbf{E}\left[L_{f_j} \cdot \mathbf{1}_{\{L_{f_j} > u\}}\right] &= \mathbf{E}\left[\left(f_j^2(X) - 2f_j(X)Y\right) \cdot \mathbf{1}_{\{f_j^2(X) - 2f_j(X)Y > u\}}\right] \\ &= p_j \, \mathbf{E}\left[(1 - 2Y) \cdot \mathbf{1}_{\{1-2Y > u\}}\right] \\ &= p_j \left(\Pr\left(Y < \frac{1-u}{2}\right) - 2\, \mathbf{E}\left[Y \cdot \mathbf{1}_{\{Y < \frac{1-u}{2}\}}\right]\right). \qquad (104)\end{aligned}$$

Now, let $K := \frac{u-1}{2}$. It is easy to show that

$$\Pr(Y > K) \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-K^2/2}}{K}.$$

In addition, for $u \geq 3$ (and hence $K \geq 1$), we have

$$\begin{aligned}\mathbf{E}\left[Y \cdot \mathbf{1}_{\{Y>K\}}\right] &= \int_0^\infty \Pr(Y \cdot \mathbf{1}_{\{Y>K\}} > t)dt = \int_K^\infty \Pr(Y > t)dt \\ &\leq \int_K^\infty \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t} dt \leq \int_K^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-K^2/2}}{K} dt.\end{aligned}$$

Thus, taking $u = 3$, we see that (104) is at most $p_j\sqrt{\frac{2}{\pi}}e^{-1/2} \leq \frac{p_j}{2}$, the witness condition therefore holds, and so we may apply the first part of Theorem 22. □

References

Jean-Yves Audibert. PAC-Bayesian statistical learning theory. *Thèse de doctorat de l'Université Paris*, 6:29, 2004.

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. In *NIPS*, 2007.

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

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Andrew Barron, Mark J. Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561, 1999.

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

Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

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

Anirban Bhattacharya, Debdeep Pati, Yun Yang, et al. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.

Peter J. Bickel and Bas J.K. Kleijn. The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40(1):206–237, 2012.

Lucien Birgé. Model selection for Gaussian regression with random design. *Bernoulli*, 10 (6):1039–1051, 2004.

Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Olivier Catoni. A PAC-Bayesian approach to adaptive classification. Technical report, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, 2003.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

Nicòlo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning and Games.* Cambridge University Press, Cambridge, UK, 2006.

Corinna Cortes, Spencer Greenberg, and Mehryar Mohri. Relative deviation learning bounds and generalization with unbounded loss functions. *Ann. Math. Artif. Intell.*, 85(1):45–70, 2019.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15: 1281–1316, 2014.

Vu C. Dinh, Lam S. Ho, Binh Nguyen, and Duy Nguyen. Fast learning rates with heavy-tailed losses. In *Advances in Neural Information Processing Systems 29*, pages 505–513. Curran Associates, Inc., 2016.

Richard M. Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002.

Subhashis Ghosal and Aad W. Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

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.

Subhashis Ghosal, Jüri Lember, and Aad W. van der Vaart. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008.

Peter D. Grünwald. Viewing all models as "probabilistic". In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 171–182. ACM, 1999.

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

Peter D. Grünwald. Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *COLT*, pages 397–420, 2011.

Peter D. Grünwald. The safe Bayesian: learning the learning rate via the mixability gap. In *Proceedings 23rd International Conference on Algorithmic Learning Theory (ALT '12)*. Springer, 2012.

Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004.

Peter D. Grünwald and Nishant A. Mehta. A tight excess risk bound via a unified PAC-Bayesian-Rademacher-Shtarkov-MDL complexity. In *Proceedings 30th Conference on Algorithmic Learning Theory (ALT '19)*, 2019.

Peter D. Grünwald and Thijs Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 2017.

David Haussler and Manfred Opper. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

David Haussler, Michael Kearns, H. Sebastian Seung, and Naftali Tishby. Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25(2-3):195–236, 1996.

R. De Heide, A. Kirichenko, P. Grünwald, and N. Mehta. Safe-Bayesian generalized linear regression. *arXiv preprint arXiv:1910.....*, 2019.

Daniel J. Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

Anatoli Juditsky, Philippe Rigollet, and Alexandre B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.

Bas J.K. Kleijn and Aad W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877, 2006.

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

Wouter M. Koolen, Peter Grünwald, and Tim van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems*, pages 4457–4465, 2016.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Guillaume Lecué. *Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis.* Habilitation à diriger des recherches, Université Paris-Est, 2011.

Guillaume Lecué and Philippe Rigollet. Optimal learning with Q-aggregation. *The Annals of Statistics*, 42(1):211–224, 2014.

Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6): 2118–2132, 1996.

Qiang (Jonathan) Li. *Estimation of mixture models.* PhD thesis, Yale University, 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.

Gábor Lugosi and Shahar Mendelson. Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli*, 25(3):2075–2106, 2019.

Ryan Martin, Raymond Mess, and Stephen G. Walker. Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847, 2017.

David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

Peter McCullagh and John Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, Boca Raton, second edition, 1989.

R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

Shahar Mendelson. Learning without concentration. In *Proceedings of The 27th Conference on Learning Theory*, pages 25–39, 2014.

Shahar Mendelson. On aggregation for heavy-tailed classes. *Probability Theory and Related Fields*, 168(3-4):641–674, 2017a.

Shahar Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, Jun 2017b.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, pages 1–13, 2018.

Arkadii Nemirovskii and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

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

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Matthias W. Seeger, Sham M. Kakade, and Dean P. Foster. Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, 2008.

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

Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

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.

Vladimir N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag New York, Inc., 1995.

Vladimir Vovk. Aggregating strategies. In *Proceedings of the third annual workshop on Computational learning theory*, pages 371–383. Morgan Kaufmann Publishers Inc., 1990.

Stephen Walker and Nils Lid Hjort. On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):811–821, 2002.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362, 1995.

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

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Yuhong Yang and Andrew R Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116, 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. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006b.