Fast Rates with Unbounded Losses

Peter D. Grünwald PDG@CWI.NL
Centrum Wiskunde & Informatica (CWI) and Leiden University

Nishant A. Mehta MEHTA@CWI.NL
Centrum Wiskunde & Informatica (CWI)

Abstract

We present new excess risk bounds for randomized and deterministic estimators for general unbounded loss functions including log loss and squared loss. Our bounds are expressed in terms of the information complexity and hold under the recently introduced v-*central condition*, allowing for high-probability bounds, and its weakening, the v-*pseudoprobability convexity condition*, allowing for bounds in expectation even under heavy-tailed distributions. The parameter v 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-conditions generalize; favorable v in combination with small information complexity leads to $\tilde{O}(1/n)$ rates. While these fast rate conditions control the lower tail of the excess loss, the upper tail is controlled by a new type of *witness-of-badness condition* which allows us to connect the excess risk to a generalized Rényi divergence, generalizing previous results connecting Hellinger distance to KL divergence.

Keywords: statistical learning theory, fast rates, PAC-Bayes, information geometry

1. Introduction

In statistical learning, a learning agent which we will call Learner is presented with samples from an unknown probability distribution; Learner then plays a distribution over a set of actions in order to minimize their expected loss over future samples. More formally, we consider a probability distribution P over a sample space $\mathcal{Z}$, with independent and identically distributed (i.i.d.) samples $Z_1, \ldots, Z_n \in \mathcal{Z}$ drawn from P.[1] The loss function $\ell : \bar{\mathcal{F}} \times \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$ maps an action $f \in \bar{\mathcal{F}}$ and sample $z \in \mathcal{Z}$ to a loss value, and Learner plays a randomized action over a model $\mathcal{F} \subset \bar{\mathcal{F}}$. Classification, regression and (potentially misspecified) density estimation are special cases of this statistical learning problem, which we represent compactly via the tuple $(P, \ell, \mathcal{F})$. Learner's goal in this problem is to select a function $f \in \mathcal{F}$ that minimizes the excess risk $\mathbf{E}_{Z \sim P}[\ell(f, Z)] - \inf_{f \in \mathcal{F}} \mathbf{E}_{Z \sim P}[\ell(f, Z)]$, a setting equivalent to Vapnik's (1995) *general setting of the learning problem*. For simplicity we assume the existence of $f^* \in \mathcal{F}$ satisfying $\mathbf{E}[\ell_{f^*}] = \inf_{f \in \mathcal{F}} \mathbf{E}[\ell_f]$ as is done in many related works such as Bartlett et al. (2005) and Mendelson (2014a), and so the excess risk is measured with respect to an optimal comparator in $\mathcal{F}$. We also generalize the setting by allowing Learner to play distributions Π over $\mathcal{F}$.

This work establishes new bounds on the performance of such randomized estimators using information-theoretic arguments, including strengthenings of standard PAC-Bayesian excess risk bounds (Catoni, 2003; Audibert, 2004). Our analysis recovers bounded loss results, but, more significantly, also covers unbounded losses under a newly introduced set of assumptions. Specifically,

[1] We generally ignore all measurability issues by implicitly assuming a suitable σ-algebra on $\mathcal{Z}$.

we establish fast rates of decay for the excess risk (i.e. faster than the "slow" rate of $O(1/\sqrt{n})$), under the recently introduced v-central condition and a weakening thereof, the v-pseudoprobability convexity (PPC) condition (van Erven et al., 2015), both generalizations of fast rate conditions that were called *stochastic mixability* by van Erven et al. (2012) and Mehta and Williamson (2014). For bounded losses and $v(x) \asymp x^{1-\beta}$, both v-central and PPC are equivalent to the Bernstein condition with exponent β (Audibert, 2004; Bartlett and Mendelson, 2006) and thus generalize Tsybakov's margin condition (Tsybakov, 2004) to cases where f^* is only optimal within $\mathcal{F}$ and not the Bayes optimal decision. Yet, if $\mathcal{F}$ has unbounded excess loss then they become incomparable to Bernstein (as they are one-sided, imposing restrictions on the lower tail of random variable $\ell_f - \ell_{f^*}$ but not its upper tail, whereas Bernstein is two-sided, restricting both tails). If $\mathcal{F}$ has unbounded excess *risk* then the Bernstein condition cannot hold any more, yet the PPC condition can still hold, under polynomial tail decay assumptions.

We do need some minimal control over $\ell_f - \ell_{f^*}$, however. For this we employ a second, new assumption, the *witness condition* (see Definition 12), which automatically holds for bounded losses and finite classes. There is some similarity between this assumption and the recently introduced small-ball assumption of Mendelson (2014a), as discussed in Section 3.

A by-product of our analysis is Theorem 13 (Section 4) that implies a tight upper bound on the ratio of the KL-divergence between two probability densities to their η*-Hellinger divergence* (Liese and Vajda, 2006), generalizing previous results of Yang and Barron (1998) and Birgé and Massart (1998). We now proceed to Section 2 to formalize the setting; Section 2.1 gives an extended overview of the paper. Section 3 discusses our conditions in detail, Section 4 presents the main results Theorem 14 and Theorem 15 and discusses related work. After page 12 we provide a short proof sketch (Section 5) and suggest future work. Most proofs, technical details concerning infinities, and an extensive list of examples can be found in the appendix.

2. Setting and Goal of the Paper

Let $\ell_f(z) := \ell(f, z)$ denote the loss of action $f \in \bar{\mathcal{F}}$ 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 $\ell_f(x, y) = |y - f(x)|$, 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, $\mathcal{F}$ 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).

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 $\Pi_{|n}$. $\Pi_{|n}$ 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 $\mathcal{F}$. We assume in the sequel that this measure has a density $\pi \mid z^n$ relative to a fixed background measure ρ. Whenever we consider a distribution Π on $\mathcal{F}$ for a problem $(P, \ell, \mathcal{F})$, we denote its outcome, a random variable, as $\underline{f}$.

All random variables are assumed to be functions of $Z, Z_1, \ldots, Z_n$ which are i.i.d. $\sim \overline{P}$. If we write ℓ_f we mean $\ell_f(Z)$. For loss functions that can be unbounded both above and below, we must avoid undefined expectations and problems with interchanging order of expectations; the only practically relevant loss function of this kind that we are aware of is the log loss (see below) combined with densities on uncountable sample spaces. To avoid any such problems, we make two additional, very mild requirements on the learning problems and on the learning algorithms that we consider. These requirements, which automatically hold for all standard loss functions we are aware of except log loss, are given in Appendix D as (28) and (29), where we explain in detail how we deal with these infinity issues. From now on we tacitly only consider learning problems for which they hold and which are nontrivial in the sense that for some $f \in \mathcal{F}$, $\mathbf{E}_{Z \sim P}[\ell_f(Z)] < \infty$.

Given $(P, \ell, \mathcal{F})$, we say that comparator f^* is *optimal* if $\mathbf{E}[\ell_f - \ell_{f^*}] \geq 0$ for all $f \in \mathcal{F}$. We are usually interested in comparing the performance of estimators to an optimal comparator; our main measures of 'easiness' of a learning problem with comparator, the v-central and v-PPC conditions (defined in the next section), both imply that the comparator is optimal. The only exception is Proposition 4 on complexity (not risk) bounds, which holds for general comparators as long as $\mathbf{E}_{Z \sim P}[\ell_{f^*}(Z)] < \infty$.

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 $x \in \mathcal{X}$, $p(\cdot \mid x)$ is a density on $\mathcal{Y}$ relative to a fixed underlying measure μ. Take *log loss*, defined on outcome $z = (x, y)$ as $\ell_f(x, y) = -\log p_f(y \mid x)$. The excess risk, now $\mathbf{E}[\ell_f - \ell_{f^*}] = \mathbf{E}_{Z \sim P}\left[\log \frac{p_{f^*}(Y|X)}{p_f(Y|X)}\right]$, 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 $\mathbf{E}[\ell_f - \ell_{f^*}] = \mathrm{KL}(p \,\|\, p_f) - \mathrm{KL}(p \,\|\, p_{f^*})$. Thus, under log loss our goal is equivalent to learning a distribution minimizing the KL divergence from P over $\{p_f : f \in \mathcal{F}\}$. We take an optimal comparator, with $\inf_{f \in \mathcal{F}} \mathrm{KL}(p \,\|\, p_f) = \mathrm{KL}(p \,\|\, p_{f^*}) = \epsilon \geq 0$, where, if $\epsilon = 0$, 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^* represents the true P. If $\epsilon > 0$, we still have $\inf_{f \in \mathcal{F}} \mathbf{E}[\ell_f - \ell_{f^*}] = 0$ 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, they are most usefully applied to generalized Bayesian, two-part (MAP/MDL) or empirical risk minimization (ERM) estimators, which we now define. Fix a distribution $\Pi_{|0}$ on $\mathcal{F}$ with density π, henceforth called *prior*, and a *learning rate* $\eta > 0$. The η*-generalized Bayesian posterior* based on $\mathcal{F}$ and sample $z_1, \ldots, z_n$ and prior $\Pi_{|0}$ is the distribution $\Pi^B_{|n}$ on $f \in \mathcal{F}$, defined by its density

$$\pi^B_{|n}(f) = \pi^B(f \mid z_1, \ldots, z_n) := \frac{\exp\left(-\eta \sum_{i=1}^n \ell_f(z_i)\right) \cdot \pi(f)}{\int_{\mathcal{F}} \exp\left(-\eta \sum_{i=1}^n \ell_f(z_i)\right) \cdot \pi(f) d\rho(f)}. \tag{1}$$

We will only consider priors $\Pi_{|0}$ satisfying the natural requirement that for all $z \in \mathcal{Z}$, $\Pi_{|0}(f \in \mathcal{F} : \ell_f(z) < \infty) > 0$, so that (1) 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 $\Pi_{|0}$ with mass function π on $\ddot{\mathcal{F}}$ and define the η*-generalized two-part MDL estimator for prior* Π *at sample size* n as $\ddot{f} := \arg\min_{f \in \ddot{\mathcal{F}}} \sum_{i=1}^n \ell_f(Z_i) + \frac{1}{\eta} \cdot (-\log \pi(f))$, 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 by setting the prior Π to be uniform over $\mathcal{F}$. 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 6.

2.1. Overview: Excess Risk Bounds and How We Prove Them

Here is an informal rendering of the type of statement we prove in our main results, Theorem 14 and 15 in Section 4, with all technical terms informally explained further below:

Theorem Template Let $(P, \ell, \mathcal{F})$ be a learning problem with comparator f^*. Let Π be an arbitrary learning algorithm as defined above. Assume that $(P, \ell, \mathcal{F})$ satisfies *(a) the witness condition* and (b) *a v-fast rate condition*, where $v : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is a nonnegative increasing function. Then for any $0 < \eta < \frac{v(\epsilon)}{2}$, there are constants $c_2 > 0$ and $\eta' \asymp \eta$ such that

$$\mathbf{E}_{\underline{f} \sim \Pi_{|n}} \left[\mathbf{E}[\ell_{\underline{f}} - \ell_{f^*}] \right] \trianglelefteq_{n \cdot \eta'} c_2 \left(\mathrm{IC}_{n,\eta}(f^* \parallel \Pi_|) + \epsilon \right). \tag{2}$$

with information complexity IC defined as:

$$\mathrm{IC}_{n,\eta}(f^* \parallel \Pi_|) := \mathbf{E}_{\underline{f} \sim \Pi_{|n}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\ell_{\underline{f}}(Z_i) - \ell_{f^*}(Z_i) \right) \right] + \frac{\mathrm{KL}(\Pi_{|n} \parallel \Pi_{|0})}{\eta \cdot n}. \tag{3}$$

Thus, we bound the expected excess risk we get by randomizing over $\Pi_{|n}$ in terms of a complexity term, which is just a variation of the term arising in standard PAC-Bayesian bounds (Catoni, 2003; McAllester, 2003), involving a KL divergence between 'posterior' and 'prior', which was first presented in this exact form by Zhang (2006b). In case the estimator $\hat{f}$ is deterministic, the left-hand side of (2) simplifies to $\mathbf{E}[\ell_{\hat{f}} - \ell_{f^*}]$. The nonstandard inequality $\trianglelefteq$ refers to inequality both in expectation and with high probability over the sample $Z^n \sim P$, and will be further explained below; if the weak version of our fast-rate condition holds, we only get inequality in expectation. To make further sense of (2), we need to explain the information complexity $\mathrm{IC}_{n,\eta}$ and the need for and meaning of preconditions (a) and (b), which will be done below. We then devote Section 3 to the preconditions and present the precise theorems in Section 4. There we also provide examples in which the theorems yield fast rates when applied with an optimally balanced feasible pair (η, ϵ).

2.2. Exponential Stochastic Inequality $\trianglelefteq$

Both our results and the v-fast rate conditions may be expressed succinctly via the notion of *exponential stochastic inequality*.

Definition 1 (Exponential Stochastic Inequality (ESI)) *Let $\eta > 0$ and let U, U' be random variables on some probability space with probability measure P. We define*

$$U \trianglelefteq_{\eta}^{P} U' \quad \Leftrightarrow \quad \mathbf{E}_{U,U' \sim P} \left[e^{\eta(U - U')} \right] \leq 1. \tag{4}$$

When clear from the context, we omit the distribution P from the notation. In particular, when U, U' are defined relative to a learning problem $(P, \ell, \mathcal{F})$, then $\trianglelefteq_\eta$ stands for $\trianglelefteq_\eta^P$. Also, if a distribution Π on $\mathcal{F}$ is specified, it stands for $\trianglelefteq_\eta^{P \otimes \Pi}$, i.e. P in (4) is the product distribution of P and Π.

An ESI simultaneously captures "with (very) high probability" and "in expectation" results.

Proposition 2 (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 $1 - e^{-K}$, $U \leq U' + K/\eta$.*

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 3 (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: $\mathbf{E}[e^{\frac{\eta}{2}((U-a)+(V-b))}] \leq \frac{1}{2}\mathbf{E}[e^{\eta(U-a)}] + \frac{1}{2}\mathbf{E}[e^{\eta(V-b)}]$. ■

2.3. Information Complexity

The present form of the information complexity is due to Zhang (2006b), with precursors from Rissanen (1989) and Yamanishi (1998). It relates to covering numbers via the following proposition:

Proposition 4 *Consider a learning problem $(P, \ell, \mathcal{F})$ with comparator f^* and let $Z_1, \ldots, Z_n$ satisfy $\sum_{i=1}^n \ell_{f^*}(Z_i) < \infty$. Let $\Pi_|$ be an η-Bayesian posterior. We have for all $\eta > 0$ that*

$$\mathrm{IC}_{n,\eta}(f^* \,\|\, \Pi^B_|) = \inf_{\Pi \in \mathrm{RAND}} \mathrm{IC}_{n,\eta}(f^* \,\|\, \Pi_|) \leq \inf_{\dot{f} \in \mathrm{DET}} \mathrm{IC}_{n,\eta}(f^* \,\|\, \dot{f}) = \mathrm{IC}_{n,\eta}(f^* \,\|\, \ddot{f}), \tag{5}$$

where RAND *is the set of* all *learning algorithms* Π *and* DET *the set of* all *deterministic estimators that can be defined relative to $(P, \ell, \mathcal{F})$. Furthermore, suppose $\mathcal{F} = \bigcup_{j \in \mathbf{N}} \mathcal{F}_j$ is a countable union of sub-models such that for $\delta > 0$, $\ddot{\mathcal{F}}_{j,\delta} \subset \mathcal{F}_j$ is a minimal δ-cover of $\ddot{\mathcal{F}}_j$ in the ℓ_∞-norm (i.e. $\sup_{f \in \mathcal{F}_j} \min_{\dot{f} \in \ddot{\mathcal{F}}_{j,\delta}} \|\ell_f - \ell_{\dot{f}}\|_\infty \leq \delta$). Define $\Gamma := \{2^0, 2^{-1}, \ldots, 2^{-K}\}$ for $K := \lceil \log_2(n) \rceil$. Assume that for all j, $|\ddot{\mathcal{F}}_{j,\delta}| = \mathcal{N}(\mathcal{F}_j, \delta) < \infty$, let $\pi_\mathbb{N}$ be a probability mass function on $\mathbb{N}$ and let $\Pi_{|0}$ be the prior on $\bigcup_{j \in \mathbf{N}, \delta \in \Gamma} \ddot{F}_{j,\delta}$ with mass function π given by, for $f \in \ddot{F}_{j,2^{-k}}$, $\pi(f) = \pi_\mathbb{N}(j)/(K \cdot \mathcal{N}(\mathcal{F}_j, 2^{-k}))$. Then the right-hand side of (5) is further bounded as*

$$\inf_{f' \in \ddot{\mathcal{F}}} \left\{ \frac{1}{n} \sum_{j=1}^n (\ell_{f'}(Z_j) - \ell_{f^*}(Z_j)) + \frac{-\log \pi(f')}{\eta \cdot n} \right\} \leq \inf_{j \in \mathbb{N}, \delta \in \Gamma} \left\{ \delta + \frac{\log \mathcal{N}(\mathcal{F}_j, \delta) + \log \log_2(2n) - \log \pi_\mathbb{N}(j)}{\eta \cdot n} \right\}.$$

In the special case with singleton $\mathcal{F}_j$'s, the right-hand side reduces to $-\log \pi_\mathbb{N}(j)$.

Zhang (2006a) showed (5) in the i.i.d. setting. The second inequality is an immediate consequence of the definitions; we omit a detailed proof. Additionally, it is well known that the information complexity of a generalized Bayesian posterior is equal to the generalized Bayesian marginal likelihood (see e.g. equation (5) of Zhang (2006b)), which, in the special case of $\eta = 1$ and log loss recovers the marginal likelihood of the data relative to f^*.

From Theorem 16 and this result, we see that we have three equivalent characterizations of information complexity for η-Bayesian estimators. First, there is just the basic definition (3) with $\Pi_{|n}$ instantiated to the η-Bayes posterior. Second, there is the characterization as the minimizer of (3) for the given data, over all distributions Π over λ, as given by (5). And third, there is the characterization in terms of the generalized Bayesian marginal likelihood.

2.4. The need for Conditions: from annealed to standard risk

For $\eta > 0$, $f \in \bar{\mathcal{F}}$, we employ a generalization of excess risk that we call *annealed excess risk* (terminology from statistical mechanics; see e.g. Haussler et al. (1996)), or *generalized Rényi divergence* (terminology from information theory, see e.g. van Erven and Harremoës (2014)):

$$\mathbf{E}^{\mathrm{ANN}(\eta)}_{Z\sim P}\left[\ell_f - \ell_{f^*}\right] = -\frac{1}{\eta}\log \mathbf{E}_{Z\sim P}\left[e^{-\eta(\ell_f - \ell_{f^*})}\right], \tag{6}$$

with $\log$ the natural logarithm. As shown by Zhang (2006b), the central result (2) holds *with* $\epsilon = 0$ *and under no further conditions* if we replace the $\Pi_{|n}$-expected risk on the left by its annealed version $\mathbf{E}_{\underline{f}\sim\Pi_{|n}}\mathbf{E}^{\mathrm{ANN}(\eta)}_{Z\sim P}\left[\ell_{\underline{f}} - \ell_{f^*}\right]$. Thus, our strategy in proving our theorems will be to determine conditions under which the η-annealed excess risk is similar enough to the standard risk for (2) to hold. The excess risk does satisfy $\lim_{\eta\downarrow 0}\mathbf{E}^{\mathrm{ANN}(\eta)}_{Z\sim P}\left[\ell_f - \ell_{f^*}\right] = \mathbf{E}_{Z\sim P}\left[\ell_f - \ell_{f^*}\right]$ (see Proposition 21), but for $\eta \downarrow 0$ the information complexity diverges to infinity, so a bound in terms of $\mathrm{IC}_{n,\eta}$ becomes useless. We provide two types of conditions, the *v-fast rate conditions* and the *witness condition*. When both hold, $\mathbf{E}^{\mathrm{ANN}(\eta)}_{Z\sim P}\left[\ell_f - \ell_{f^*}\right]$ and $\mathbf{E}_{Z\sim P}\left[\ell_f - \ell_{f^*}\right]$ can be linked so that (2) holds. The novelty of this paper is that these conditions are far less restrictive than those implicitly employed in earlier works using similar proof strategies such as Zhang (2006a) and Audibert (2004).

$\mathbf{E}^{\mathrm{ANN}(\eta)}_{Z\sim P}\left[\ell_f - \ell_{f^*}\right]$ is decreasing in η and without further condition can become negative and even $-\infty$ for $\eta > 0$. The first *v-fast rate conditions* (van Erven et al., 2015), presented in Section 3, ensure that the annealed excess risk is positive, or at least larger than $-\epsilon$, for all $\eta < v(\epsilon)$ (the slack term in (2) arises because it may still become slightly negative). The parameter v determines how small we must make η to get sufficiently small ϵ to make (2) useful; the faster v grows, the better.

To see that we need a second condition, consider the density estimation Example 1 again. If we assume a correct model, $p = p_{f^*}$, then the central condition holds automatically (van Erven et al., 2015), for all $\eta \le 1$. $\mathbf{E}^{\mathrm{ANN}(\eta)}_{Z\sim P}\left[\ell_f - \ell_{f^*}\right]$ is now equal to the η-Rényi divergence between p and p_f, which is an (often tight) upper bound of the η-Hellinger divergence $\mathrm{H}_\eta(p \,\|\, p_f) := \eta^{-1}(1 - \mathbf{E}^*_{Z\sim p}[(p_f/p)^\eta])$; note that $\mathrm{H}_{1/2}$ coincides with the standard squared Hellinger distance $2 - 2\int\sqrt{p\cdot p_f}d\mu$. To bound the risk in terms of the η-annealed excess risk it is now sufficient to bound the KL divergence in terms of η-Hellinger divergence. Yet, the latter is immediately seen to be bounded for $\eta < 1$, whereas in general we can have $\mathrm{KL}(p \,\|\, p_f) = \infty$. We thus need an extra condition. The simplest such condition is that the likelihood ratio of p to p_f is bounded for all $f \in \mathcal{F}$ (in terms of log loss this means that the loss is bounded). For that case, Birgé and Massart (1998) and Yang and Barron (1998) proved a tight bound on the ratio between KL and standard ($\eta = 1/2$) Hellinger. Theorem 13 in Section 4 represents a vast generalization of their result to arbitrary η, misspecified $\mathcal{F}$, and general loss functions under the witness condition (Section 3), which allows unbounded losses. It is the cornerstone for proving our main results, Theorem 14 and Theorem 15.

3. Getting a Grip on the Conditions

We now turn to our fast rate conditions. The central and PPC conditions below were introduced by van Erven et al. (2015), where they also were studied and compared to other conditions in detail.

Definition 5 (Central Condition) *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 $f^* \in \mathcal{F}$ such that*

$$\ell_{f^*} - \ell_f \trianglelefteq_\eta \epsilon \qquad \textit{for all } f \in \mathcal{F}, \tag{7}$$

If it satisfies the η-central condition up to 0, we say that the strong η-central condition holds.

The condition thus expresses that f^* is not just the risk minimizer in $\mathcal{F}$ but also that it provides an exponential bound on the probability that $\ell_{f^*} - \ell_f$ is large (both of which follow from Proposition 2). This is clearly desirable for learning, because, for any fixed $f \in \mathcal{F}$ that is worse than f^*, the condition makes the probability that f outperforms f^* on the sample exponentially small.

The special case of this condition with $\eta = 1$ 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). Space limitations preclude doing justice to all the implications of the central condition and its equivalences to other conditions. Here we merely note that the strong central condition holds for density estimation with log loss in the well-specified setting and, under a convex model, in the misspecified setting (see Example 2.2 of van Erven et al. (2015)), as well as that, for classification and other bounded loss cases, it can be related to the Bernstein condition (Audibert, 2004; Bartlett and Mendelson, 2006) (as discussed immediately before Definition 9 below).

Although not all learning problems satisfy the strong η-central condition with respect to comparator f^*, it turns out that by adopting a different comparator g we always are guaranteed to have $\ell_g - \ell_f \trianglelefteq_\eta 0$ for all $f \in \mathcal{F}$, and moreover, the performance of f^* can be related to the performance of g under conditions much weaker than the strong η-central condition. This comparator, formally defined below, is an instance of what we call a *generalized reversed information projection* (GRIP), a versatile generalization of the reversed information projection of Barron and Li (1999) (a major inspiriation for this work). The original projection was used in the context of density estimation under $\log$ loss; we extend it to general learning problems:

Definition 6 (GRIP) *Let $(P, \ell, \mathcal{F})$ be a learning problem. Define[2] the set of pseudoprobability 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 $\operatorname{conv}(\mathcal{E})$ is defined as the pseudo-loss ℓ_g satisfying*

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

From the above definition, we see that a GRIP is only a *pseudo-predictor*, meaning that it may fail to correspond to any actual prediction function; however, the corresponding loss for a GRIP *is* well-defined, as shown in Appendix C. It will be convenient to call the quantity appearing in the center expectation above a "mix-loss", defined for a distribution $Q \in \Delta(\mathcal{F})$ as $\ell_Q := -\frac{1}{\eta} \log \mathbf{E}_{\underline{f} \sim Q}\left[e^{-\eta \ell_{\underline{f}}} \right]$. We defer showing that $\ell_g - \ell_f \trianglelefteq_\eta 0$ for all $f \in \mathcal{F}$ until after Definition 10, as the proof is more natural using that definition.

The next definition is a weakening of the strong η-central condition that relates ϵ to η.

Definition 7 (v-Central Condition) *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 $\epsilon > 0$, there exists a function $f^* \in \mathcal{F}$ such that (7) is satisfied with $\eta = v(\epsilon)$.*

[2] This transformation is known as *entropification* (Grünwald, 1999).

When the v-central condition holds but the strong η-central condition does not, we may instead use the GRIP comparator g and the following proposition, relating the random variables ℓ_{f^*} and ℓ_g:

Proposition 8 *Under the v-central condition, we have $\ell_{f^*} - \ell_g \trianglelefteq_{v(\epsilon)} \epsilon$.*

Proof Let $Q \in \Delta(\mathcal{F})$ be arbitrary and let ℓ_Q be the mix-loss with respect to $\eta = v(\epsilon)$. Then

$$\mathbf{E}\left[e^{v(\epsilon)(\ell_{f^*}-\ell_Q)}\right] = \mathbf{E}\left[\mathbf{E}_{\underline{f}\sim Q}\left[e^{v(\epsilon)(\ell_{f^*}-\ell_{\underline{f}})}\right]\right] \leq e^{v(\epsilon)\epsilon}.$$

Now, as we show in Appendix C, there exists a sequence $\{Q_k\}$ such that $\{\ell_{Q_k}\}$ converges to ℓ_g in $L_1(P)$, and so we also have $\mathbf{E}\left[e^{-v(\epsilon)(\ell_{f^*}-\ell_g)}\right] \leq e^{v(\epsilon)\epsilon}$. ■

This proposition in particular implies that the information complexity with respect to comparator g is, with high probability, not much larger than the information complexity with respect to comparator f^*, a crucial property in our transferring risk bounds with respect to the pseudo-predictor g to risk bounds with respect to our actual comparator f^*.

One of the main results of van Erven et al. (2015) (in their Section 5) is that for bounded loss functions, the v-central condition holds for some v with $v(\epsilon) \asymp \epsilon^{1-\beta}$ iff the Bernstein condition below holds for exponent β and some $B > 0$. The Bernstein condition is known to characterize the rates that can be obtained in bounded loss problems for proper learners, and the same thus holds for the central condition.

Definition 9 (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 $f \in \mathcal{F}$, $\mathbf{E}[(\ell_f - \ell_{f^*})^2] \leq C\left(\mathbf{E}[\ell_f - \ell_{f^*}]\right)^{\beta}$.*

The best case of the Bernstein condition is exponent 1, corresponding to a v with $v(0) > 0$, i.e. to the strong central condition.

The v-central condition, for $v(\epsilon) \asymp \epsilon^{1-\beta}$, $\beta \in [0,1]$, requires exponential tails of $\ell_{f^*} - \ell_f$ and thus can fail to hold in the case of regression with squared loss under polynomially decaying tails. A weakening of it, the v-PPC condition, may then still hold. For example, in bounded regression (Example 6 in Appendix E.2), the v-central condition fails to hold but the v-PPC condition does hold for $v(\epsilon) = \sqrt{\epsilon}$ under the assumption that the tails satisfy $\mathbf{E}[|Y|^4] < \infty$. To prepare for the condition, note that (7) really means that $\mathbf{E}_{Z\sim P}\left[\exp(-\eta(\ell_f(Z) - \ell_{f^*}(Z)))\right] \leq \exp(\eta\epsilon)$ and thus is clearly equivalent to:

$$\mathbf{E}_{\underline{f}\sim Q}\,\mathbf{E}_{Z\sim P}\left[e^{-\eta(\ell_{\underline{f}}(Z)-\ell_{f^*}(Z))}\right] \leq e^{\eta\epsilon} \qquad \text{for all } Q \in \Delta(\mathcal{F}). \tag{8}$$

Using the mix-loss notation, the left-hand side can be rewritten as $\mathbf{E}_{Z\sim P}\left[e^{-\eta(\ell_Q(Z)-\ell_{f^*}(Z))}\right]$, so that (8) becomes

$$\ell_{f^*}(Z) - \ell_Q(Z) \trianglelefteq_\eta \epsilon \qquad \text{for all } Q \in \Delta(\mathcal{F}), \tag{9}$$

which is thus equivalent to the η-central condition (7). The PPC-condition is simply the strict weakening of (9) to its in-expectation form:

Definition 10 (Pseudoprobability convexity condition) *Let $\eta > 0$ and $\epsilon \geq 0$. We say that $(P, \ell, \mathcal{F})$ satisfies the η-pseudoprobability convexity condition up to ϵ if there exists some $f^* \in \mathcal{F}$ such that*

$$\mathbf{E}_{Z\sim P}\left[\ell_{f^*}(Z) - \ell_Q(Z)\right] \leq \epsilon \qquad \textit{for all } Q \in \Delta(\mathcal{F}). \tag{10}$$

Taking v as in Definition 7, we say that $(P, \ell, \mathcal{F})$ satisfies the v-PPC condition if, for all $\epsilon > 0$, there exists a function $f^* \in \mathcal{F}$ such that (10) is satisfied with $\eta = v(\epsilon)$.

The learning implications of the v-PPC condition are best viewed through the lens of the GRIP pseudo-predictor g, which satisfies $\mathbf{E}[\ell_g] = \inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}\left[\ell_Q\right]$. The v-PPC condition can be re-expressed as the condition that g has at most ϵ less risk than f^*, while from earlier we recall that g itself satisfies the key property $\ell_g - \ell_f \trianglelefteq_\eta 0$ for all $f \in \mathcal{F}$ (implying that $\ell_f - \ell_g$ has an exponential lower tail). The latter helps us guarantee that the excess risk with respect to g decays at a faster rate, while the former helps us transfer this guarantee to the excess risk with respect to our actual comparator f^*.

Finally, as promised, we show that:

Proposition 11 *For all $f \in \mathcal{F}$, we have $\ell_g - \ell_f \trianglelefteq_\eta 0$.*

Proof Consider the *loss class* $\ell_{\mathcal{F}'} := \{\ell_Q : Q \in \Delta(\mathcal{F})\} \cup \{\ell_g\}$. Observe that the PPC and central conditions can be defined using a loss class and P rather than explicitly requiring a tuple $(P, \ell, \mathcal{F})$. Now, we have $\ell_g \in \ell_{\mathcal{F}'}$, and $(P, \ell_{\mathcal{F}'})$ clearly satisfies the (suitably reparametrized) strong η-PPC condition. Theorem 3.10 of van Erven et al. (2015) then implies that the (suitably reparametrized) strong η-central condition also holds, *a fortiori* implying the result (which need only hold for the Dirac distributions Q_f supported on some $f \in \mathcal{F}$). ■

We now turn to the second condition needed to relate annealed to standard excess risk.

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

$$\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u\}}\right] \geq c\,\mathbf{E}[\ell_f - \ell_{f^*}]. \tag{11}$$

*More generally, for $M > 1$ we say that $(P, \ell, \mathcal{F})$ with comparator f^* satisfies the M-*witness of badness condition *if there exist constants $u > 0$ and $c \in (0, 1]$ such that for all $f \in \mathcal{F}$*

$$\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u(1 \vee (M^{-1}\mathbf{E}[\ell_f - \ell_{f^*}]))\}}\right] \geq c\,\mathbf{E}[\ell_f - \ell_{f^*}]. \tag{12}$$

We see that the witness condition (11) is just the M-witness condition for $M = \infty$, which is suitable for the case where $\mathcal{F}$ has unbounded loss but bounded excess risk (Theorem 14). We need the more complicated condition with $M < \infty$ only to deal with the unbounded excess risk case (Theorem 15). We also see that it trivially holds if $\mathcal{F}$ is finite or ℓ is bounded.

The intuitive reason for imposing this condition is to rule out situations in which learnability simply does not hold. For instance, consider a setting where, with probability $1 - \delta$, we have $\ell_f = 0$ and $\ell_{f^*} = \frac{1}{1-\delta}$, while with probability δ, we have $\ell_f = \frac{2}{\delta}$ and $\ell_{f^*} = 0$. Then $\mathbf{E}[\ell_f - \ell_{f^*}] = 1$, but as δ goes to zero, empirically we will never *witness the badness of f* as it almost surely achieves lower loss than f^*. We now provide an example where the witness condition holds:

Example 2 (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 $\mathbf{E}[Y^2] \leq C$, the function class $\mathcal{F}$ consists of functions f for the predictions $f(X)$ are bounded as $|f(X)| \leq r$ almost surely, and 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 last assumption 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.

As we show in Appendix E.1, in this setup the Bernstein condition holds with exponent 1 and multiplicative constant $8(\sqrt{C}+r)^2$. Moreover, as we also show, a Bernstein condition with exponent 1 and multiplicative constant B *always* implies that the witness condition holds if $u \geq B$ with constant $c = 1 - \frac{B}{u}$. In particular, in the current setting the witness condition holds with $u = 16(\sqrt{C}+r)^2$ and $c = \frac{1}{2}$. □

Intriguingly, on an intuitive level the witness condition bears some similarity to the recent small-ball assumption of Mendelson (2014a). 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 $\Pr\left(|f - h| \geq \kappa \|f - h\|_{L_2}\right) \geq \varepsilon$. Under this assumption, Mendelson (2014a) established bounds on the L_2-parameter estimation error $\|\hat{f} - f^*\|_{L_2}$ 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-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 9 in Appendix E.2); 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.

4. Main Results

We now present Theorem 13, underlying our two main results further below.

Theorem 13 *Let $\bar{\eta} > 0$. Assume that $\mathbf{E}\, e^{-\bar{\eta}(\ell_f - \ell_{f^*})} \leq 1$. Let $u > 0$ and $c \in (0,1]$ be constants for which $\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u\}}\right] \geq c\, \mathbf{E}[\ell_f - \ell_{f^*}]$. Then for all $\eta \in (0, \bar{\eta})$, with $c_1 := \frac{1}{c} \frac{\eta u + 1}{1 - \frac{\eta}{\bar{\eta}}}$,*

$$\mathbf{E}[\ell_f - \ell_{f^*}] \;\leq\; c_1 \cdot \frac{1}{\eta} \mathbf{E}\left[1 - e^{-\eta(\ell_f - \ell_{f^*})}\right] \;\leq\; c_1 \cdot \mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_f - \ell_{f^*}\right]. \tag{13}$$

This result generalizes results of Birgé and Massart (1998, Lemma 5) and Yang and Barron (1998, Lemma 4) that bound the ratio between the standard KL-divergence $\mathrm{KL}(P \,\|\, Q)$ and the (standard) 1/2-Hellinger divergence $\mathrm{H}_\eta(P \,\|\, Q)$ 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 LHS becomes $\mathrm{KL}(P \,\|\, Q)$ and $\frac{1}{\eta} \mathbf{E}[1 - e^{-\eta \ell_f - \ell_{f^*}}] = \frac{1}{\eta}(1 - 1\, \mathbf{E}[(q/p)^\eta])$. Under a bounded density ratio $p/q \leq V$, we can take $u = \log V$ and $c = 1$ (the witness condition is then trivially satisfied), so that $c_1 = \frac{\eta \log V + 1}{1 - \eta}$, which for $\eta = 1/2$ coincides with the Birgé-Massart-Yang-Barron-bound. Wong et al. (1995, Theorem 5) also bound the ratio between the standard KL and Hellinger divergences, even when the density ratio random variable p/q is unbounded, although with exponentially decaying tails; we will compare a version of our result to their bound in future work. Lastly, in the general learning setting but with excess loss bounded by some constant b, we may always take $u = b$ and $c = 1$ so that the witness condition is trivially satisfied.

Example 3 (Example 2 and Theorem 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 2. This result thus gives an analogue of Theorem 13 for bounded regression with Gaussian noise and also allows us to prove versions of our two main results below for this model. Our earlier Example 2 extends Birgé's result, since it shows that the excess risk can be bounded by a constant times the annealed excess risk if the tail of the target Y has bounded second moment, which, in the well-specified setting in particular, specializes to ξ having bounded second moment 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_1 for which (13) holds for all $f \in \mathcal{F}$. From this we infer, by using Theorem 13 in the contrapositive direction, that for such $\mathcal{F}$ the witness condition also does not hold. □

Let $(P, \ell, \mathcal{F})$ be a learning problem with comparator f^*. We now present our first main result, an excess risk bound that holds under the witness condition, which allows unbounded losses but requires bounded excess risk, i.e. $\sup_{f \in \mathcal{F}} \mathbf{E}[\ell_f - \ell_{f^*}] < \infty$. Let Π be a learning algorithm.

Theorem 14 (Excess Risk Bound - Bounded Excess Risk Case) *Assume that $(P, \ell, \mathcal{F})$ satisfies the witness condition (11). Let $c_2 := \frac{1}{c} \frac{2\eta u + 1}{1 - \frac{2\eta}{v(\epsilon)}}$. Then, with information complexity* IC *as in (3), under the v-central condition, for any $\eta < \frac{v(\epsilon)}{2}$:*

$$\mathbf{E}_{\underline{f} \sim \Pi_{|n}} \left[\mathbf{E}[\ell_{\underline{f}} - \ell_{f^*}] \right] \trianglelefteq_{\frac{\eta \cdot n}{2 c_2}} c_2 \left(\mathrm{IC}_{n,\eta}(f^* \parallel \Pi_|) + \epsilon \right). \tag{14}$$

whereas under the v-PPC condition, for any $\eta < \frac{v(\epsilon)}{2}$:

$$\mathbf{E}_{Z_1^n} \left[\mathbf{E}_{\underline{f} \sim \Pi_{|n}} \left[\mathbf{E}[\ell_{\underline{f}} - \ell_{f^*}] \right] \right] \leq c_2 \left(\mathbf{E}_{Z_1^n} \left[\mathrm{IC}_{n,\eta}(f^* \parallel \Pi_|) \right] + \epsilon \right). \tag{15}$$

The factor c_2 explodes if $\eta \uparrow v(\epsilon)/2$. If the v-central or v-PPC condition holds for some v, it clearly also holds for any smaller v, in particular for $v'(\epsilon) := v(\epsilon) \wedge 1$. Applying the theorem with v' (which will not affect the rates obtained), we may thus take $\eta = v'(\epsilon)/4$, so that c_2 is bounded by $\frac{1}{c}(u + 2)$. The ESI in the first result then implies that with probability at least $1 - e^{-K}$ the left-hand side exceeds the right-hand side by at most $\frac{2 c_2 K}{\eta n}$ for constant c_2; it also implies (15). For the case of bounded loss, we can further take u to be $\sup_{f \in \mathcal{F}} \| \ell_f - \ell_{f^*} \|_\infty$ and $c = 1$. Finally, in the special case when strong $\bar{\eta}$-central holds, we can take $\epsilon = 0$ and $v(0) = \bar{\eta}$; then, as explained in the proof, $\trianglelefteq_{\frac{\eta \cdot n}{2 c_2}}$ may be replaced by $\trianglelefteq_{\frac{\eta \cdot n}{c_2}}$, yielding slightly better concentration.

While the theorem holds for arbitrary learning algorithms, good bounds on $\mathrm{IC}_{n,\eta}$ are available mainly for $\Pi_{|n}$ set to a η-generalized Bayes or two-part or ERM estimators. A very simple such bound is given in Proposition 4. Applying the prior $\Pi_{|0}$ given there we see that for both η-generalized Bayes and two-part estimators, (14) implies that with high probability and in expectation, taking for simplicity $\eta = v(\epsilon)/4$,

$$\mathbf{E}_{\underline{f} \sim \Pi_{|n}} \left[\mathbf{E}[\ell_{\underline{f}} - \ell_{f^*}] \right] \leq c_2 \left(\inf_{\substack{j \in \mathbb{N} \\ \delta}} \left\{ \delta + 4 \cdot \frac{\log \mathcal{N}(\mathcal{F}_j, \delta) + \log \log_2(2n) - \log \pi_{\mathbb{N}}(j)}{v(\epsilon) \cdot n} \right\} + \epsilon \right).$$

where the infimum is over all δ of form 2^{-k} that are larger than $1/n$. If f^* is contained in some $\mathcal{F}_j$ with covering number polynomial in δ, we can take $\delta \asymp 1/n$ and then we get a bound of $O(\epsilon + \log(n)/(nv(\epsilon)))$, which holds uniformly for all $\epsilon > 0$ so we can optimize for ϵ.

Example 4 (Bounded Loss, Bernstein Condition) For bounded losses, the Bernstein condition is automatically satisfied with exponent $\beta = 0$, implying that v-central automatically holds with $v(\epsilon) = C\epsilon^{1-\beta} = C\epsilon$ for some $C > 0$ (see Section 3). If the metric entropy $\log \mathcal{N}(\mathcal{F}, \delta)$ is logarithmic in δ, then in optimizing over ϵ we should take $\epsilon \asymp 1/\sqrt{n}$, giving the well-known 'slow rate' bounds for bounded losses, of $\tilde{O}(1/\sqrt{n})$ (suppressing log-factors). If the set of models $\mathcal{F}_j$ under consideration is finite, we get the same result with ERM, which can be applied without knowledge of η; for finite classes we can also get rid of the $\log n$ term. At the other extreme, if Bernstein holds with exponent $\beta = 1$, then v-central holds with $v(\epsilon) = \bar{\eta} > 0$ constant, and we get $\tilde{O}(1/n)$ rates. □

We further remark that, for the case of log loss, Barron and Cover (1991) and Zhang (2006b) provide extensive bounds on information complexity for nonparametric density estimation that lead to the minimax convergence rates in many cases (with k-dimensional parametric $\mathcal{F}$ one gets the standard $(k/2) \log n$ BIC term). If one is content with in-expectation results, one can use (15) and it is sufficient to bound $\mathrm{IC}_{n,\eta}$ *in expectation* rather than almost surely. One then gets sophisticated analogues of Proposition 4 with (smaller) covering numbers defined directly in terms of excess risk rather than the sup-norm. The main contribution of our work in this density-estimation context is that we extend such works to hold under (a) misspecification (the v-fast rate conditions are much weaker than earlier conditions for Bayesian nonparametric density estimation under misspecification such as by Kleijn and van der Vaart (2006)) and (b) in terms of KL divergence, under the witness condition — earlier results are invariably geared towards Hellinger distance.

We now present a result for a learning problem $(P, \ell, \mathcal{F})$ with unbounded excess risk.

Theorem 15 (Excess Risk Bound - Unbounded Excess Risk Case) *Assume that $(P, \ell, \mathcal{F})$ satisfies the M-witness condition (12). Let $c_2 := \frac{1}{c} \frac{2\eta u+1}{1-\frac{2\eta}{v(\epsilon)}}$. Let $\{b_n\}$ be a decreasing sequence. Then under the v-central condition, for any $\eta < \frac{v(\epsilon)}{2}$:*

$$\Pi_{|n}\big(\{f : \mathbf{E}[\ell_f - \ell_{f^*}] \geq b_n\}\big) \;\trianglelefteq_{(b_n \wedge M)\cdot\frac{n\cdot\eta}{4c_2}}\; c_2 \left(\frac{1}{M} + \frac{1}{b_n}\right) \cdot \big(\mathrm{IC}_{n,\eta}(f^* \parallel \Pi_|) + \epsilon\big).$$

whereas under the v-PPC condition, if $\Pi_{|n}$ is a selector *concentrated on $\hat{f}_n$, for any $\eta < \frac{v(\epsilon)}{2}$:*

$$P\left(\mathbf{E}[\ell_{\hat{f}_n} - \ell_{f^*}] \geq b_n\right) \leq c_2 \left(\frac{1}{M} + \frac{1}{b_n}\right) \cdot \left(\mathbf{E}_{Z_1^n}[\mathrm{IC}_{n,\eta}(f^* \parallel \Pi_|)] + \epsilon\right), \tag{16}$$

where P is the distribution of Z^n.

We note that, as can be seen from the first display in the proof in Appendix B, the first statement implies the second, so that, as before, the result under the central condition is stronger than under the PPC condition. This result is harder to interpret than the previous one, so for simplicity we will focus on the case that $\Pi_{|n}$ corresponds to a selector $\hat{f}_n$.

We can now take any sequence η_n and ϵ_n so that $d_n := \mathbf{E}_{Z_1^n}[\mathrm{IC}_{n,\eta_n}(f^* \parallel \Pi_|)] + \epsilon_n$ converges to 0 (we may take pairs η_n and ϵ_n that optimize the bound as this is sample-independent; e.g. in the setting of Example 4, if $v(\epsilon) = \epsilon^{1-\beta}$ then we get a rate of $n^{-1/(2-\beta)}$), and further any nondecreasing sequence a_n such that $b_n := a_n d_n$ is decreasing. Plugging b_n into (16) gives, for all n:

$$P\left(\mathbf{E}[\ell_{\hat{f}_n} - \ell_{f^*}] \geq a_n \cdot (\mathbf{E}[\mathrm{IC}_{n,\eta_n}(f^* \parallel \Pi_|)] + \epsilon_n)\right) \leq c_2 \left(\frac{d_n}{M} + \frac{1}{a_n}\right) \xrightarrow[n\to\infty]{} 0,$$

and thus it implies $\mathbf{E}[\ell_{\hat{f}_n} - \ell_{f^*}] = O_P(\mathbf{E}[\mathrm{IC}_{n,\eta_n}(f^* \parallel \Pi_|) + \epsilon_n])$, so, while in this unbounded risk case we have no 'exponential in-probability' results such as in (14), the risk still converges at rate $\mathrm{IC}_{n,\eta_n} + \epsilon_n$ in the O_P-sense often considered in statistics (see e.g. (Van de Geer, 2000)).

4.1. Related work

Proper vs. Improper There exist learning problems $(P, \ell, \mathcal{F})$ on which no proper learner, that always predicts inside $\mathcal{F}$, can achieve a rate as good as an improper learner, that can select $\hat{f}_n \notin \mathcal{F}$ (Audibert, 2007; van Erven et al., 2015). Here we consider *randomized* proper estimators, to which the same lower bounds apply; hence, they cannot in general compete with improper methods such as exponential weighted forecasters and other aggregation methods. Such methods achieve fast rates under conditions such as stochastic exp-concavity (Juditsky et al., 2008), which are very similar to strong PPC, as explained by van Erven et al. (2015).

Empirical process vs Information-theoretic There also are approaches based on empirical process theory (EPT) like (Bartlett et al., 2005; Bartlett and Mendelson, 2006; Koltchinskii, 2006; Mendelson, 2014a; Liang et al., 2015), and 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 great advantage of EPT approaches is that they often can achieve optimal rates of convergence for 'large' models $\mathcal{F}$ with metric entropy $\log \mathcal{N}(\mathcal{F}, \epsilon)$ that increases polynomially in $1/\epsilon$; prior-based approaches (including ours) may yield suboptimal rates in such cases (see Audibert (2009) for discussion and Audibert and Bousquet (2007) for a first step into combining both approaches). On the other hand, 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, as can be seen from the final display in Proposition 4. 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.

As for unbounded losses and EPT methods, Mendelson (2014a,b) provides bounds on the L_2-estimation error $\|\hat{f} - f^*\|_{L_2}^2$ and Liang et al. (2015) on the related squared loss risk: for other loss functions not much seems to be known (Mendelson (2014b) shows that improved L_2-estimation error rates may be obtained by using other, proxy loss functions during training; but the target remains L_2-estimation). In contrast, our approach allows for general loss functions ℓ_f including density estimation, but we do not specially study proxy training losses. As explained in Section 3, in situations in which L_2-estimation error and excess squared risk coincide, the bounds remain incomparable due to incomparability of the small-ball assumption and the witness condition.

These last three EPT-based works can deal with $(P, \ell, \mathcal{F})$ with unbounded excess (squared loss) risk. This is in contrast to the prior-based methods; 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 15) for general models including countable infinite unions of models as in Proposition 4. Previous works dealing with unbounded 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 the excess risk $\sup_{f \in \mathcal{F}} \mathbf{E}[\ell_f - \ell_{f^*}] = \infty$ is unbounded, as follows trivially from inspecting Definition 9, whereas the v-central and PPC-conditions *can* hold. See for instance Example 8 in Appendix E.2, 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 15 applies (for some moderate M).

In the unbounded-loss-yet-bounded-risk case, the difference between these works and ours opaques, as there are cases where the Bernstein condition holds for some β but the v-PPC condition does not hold for $v(\epsilon) \asymp \epsilon^{1-\beta}$, but also the opposite can happen. In Appendix E.2 we provide examples of both: the first example, Example 6, is a well-specified estimation of means problem with heavy tails. If the second moment is finite, the Bernstein condition holds with exponent $\beta = 1$ and Corollary 6.2 of Audibert (2009) implies a fast rate of $\tilde{O}(1/n)$. On the other hand, the v-central condition fails to hold for any non-trivial v, and only the v-PPC condition holds for $v(\epsilon) = O(\epsilon^{2/s})$, provided that $\mathbf{E}[|Y|^s] < \infty$ for some $s \geq 2$. If $\mathbf{E}[|Y|^2] < \infty$, the witness condition also holds. Theorem 14 then implies a suboptimal rate of $\tilde{O}(n^{-s/(s+2)})$. In the second example, Example 7, 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 14 therefore applies whereas the results of Audibert (2009) and Zhang (2006b) do not. Finally we note that Audibert (2009) proves his bounds for a specialized, ingenious learning algorithm, whereas Zhang's and our bounds hold for general estimators.

5. Proof Sketch of Theorem 14

We develop the results of Theorem 14 through a certain chain of information-theoretic relationships. Establishing the theorem will take an additional level of generality afforded by using a *dynamic comparator* function $\phi : \mathcal{F} \to \bar{\mathcal{F}}$ in place of f^*. We use the notation $\mathrm{KL}(f^* \,\|\, \underline{f}) = \mathbf{E}[\ell_{\underline{f}} - \ell_{f^*}]$, $\mathrm{R}_\eta(\phi(\underline{f}) \,\|\, \underline{f}) = \mathbf{E}^{\mathrm{ANN}(\eta)}[\ell_{\underline{f}} - \ell_{\phi(\underline{f})}]$, and $\mathrm{H}_\eta(\phi(f) \,\|\, f) = \frac{1}{\eta}\,\mathbf{E}\left[1 - e^{-\eta(\ell_{\underline{f}} - \ell_{\phi(\underline{f})})}\right]$. Take some fixed $\epsilon \geq 0$ and define $\bar{\eta} := v(\epsilon)$. All inequalities in the following chain hold in expectation with respect to $\underline{f}$ drawn from the posterior $\Pi_{|n}$:

$$\overbrace{\mathrm{KL}(f^* \,\|\, \underline{f}) \lesssim \mathrlap{\underbrace{\phantom{\mathrm{H}_{2\eta}(g_{\underline{f}} \,\|\, \underline{f}) \leq \mathrm{H}_\eta(g \,\|\, \underline{f})}}_{\text{(b) Lemma 19}}}\mathrm{H}_{2\eta}(g_{\underline{f}} \,\|\, \underline{f})}^{\text{(a) Lemma 17 + Theorem 18}} \leq \overbrace{\mathrm{H}_\eta(g \,\|\, \underline{f}) \leq \mathrlap{\underbrace{\phantom{\mathrm{R}_\eta(g \,\|\, \underline{f}) \trianglelefteq_{n,\bar{\eta}} \mathrm{IC}_{n,\eta}(g \,\|\, \Pi_|)}}_{\text{(d) Theorem 16}}}\mathrm{R}_\eta(g \,\|\, \underline{f})}^{\text{(c) } 1-x \leq -\log x} \trianglelefteq_{n,\bar{\eta}} \overbrace{\mathrm{IC}_{n,\eta}(g \,\|\, \Pi_|) \trianglelefteq_{n,\bar{\eta}} \mathrm{IC}_{n,\eta}(f^* \,\|\, \Pi_|) + \epsilon}^{\text{(e) } \ell_{f^*} \trianglelefteq_{\bar{\eta}} \ell_g + \epsilon}.$$

Here, $\lesssim$ denotes inequality up to a constant. Before sketching the ideas above, we address the new character g_f; it is an instance of a GRIP. The (full) GRIP ℓ_g was defined in Definition 6, and for each $f \in \mathcal{F}$ the *mini-grip* ℓ_{g_f} is the GRIP defined by replacing $\mathcal{F}$ with the dyad $\{f^*, f\}$ in Definition 6.

Now, onwards with grasping the links in the chain. Our analysis begins with the ESI (d); this result is essentially due to Zhang (2006b), though we note that our application of that result using comparator g is non-standard.

(1) Consider first the case when f^* satisfies the strong central condition. Then $g = g_f = f^*$, so that (e) is no longer necessary. Inequality (c) is always true, and, since $g_f = f^*$, it turns out

that rather than showing (a) and (b) we instead directly can show, for fixed f, that $\mathrm{KL}(f^* \,\|\, f) \lesssim \mathrm{H}_\eta(f^* \,\|\, f)$. This inequality constitutes all the real work in the case of the strong central condition and is given by Theorem 18 with static comparator $\phi(f) = f^*$ (Lemma 19 is not necessary); we remark that Theorem 13 was merely a special case of Theorem 18. The proof of Theorem 18 navigates very carefully to control the KL-divergence by using a key lemma, Lemma 23, to handle what happens on the event $\{\ell_f - \ell_{f^*} \leq u\}$, while leveraging the witness condition to handle what happens on the complementary event $\{\ell_f - \ell_{f^*} > u\}$.

(2) When only the v-central condition holds for non-constant v, we do not have $g = f^*$ and can fail to have $g_{\underline{f}} = f^*$. As the strong central condition fails, we cannot apply Theorem 18 with comparator $\phi(f) = f^*$. Fortunately, the excess loss with respect to either $\phi(f) = g$ or $\phi(\underline{f}) = g_{\underline{f}}$ *does* satisfy $\mathbf{E}\left[e^{-\eta(\ell_f - \ell_{\phi(f)})}\right] \leq 1$ (by Proposition 11), and moreover, due to a property related to GRIPs, in either case the v-central condition implies for fixed f that $\ell_{f^*} \trianglelefteq_{v(\epsilon)} \ell_{\phi(f)} + \epsilon$ (by Proposition 8).

However, using solely either g or $g_{\underline{f}}$ is problematic. If we use only comparator g, then Theorem 18 with $\phi(f) = g$ requires a weak form of the witness condition with respect to comparator g; we so far have been unable to prove that the witness condition with comparator f^* implies a weak witness condition with comparator g. On the other hand, if we use only comparator $g_{\underline{f}}$, such a weak witness condition *is* implied by the Witness Protection Lemma (Lemma 17), but now (e) with $g_{\underline{f}}$ in place of g fails to hold because $g_{\underline{f}}$ depends on the sample (unlike the fixed function g). Fortunately, we can circumvent the pitfalls of both comparators while still enjoying their respective advantages. The critical link (b) lets us achieve the best of both worlds by way of Lemma 19, which links a Hellinger divergence with respect to $g_{\underline{f}}$ to a Hellinger divergence with respect to g. Applying Theorem 18 with $\phi(f) = g_f$ for (a) and applying $\ell_{f^*} \trianglelefteq_{v(\epsilon)} \ell_g + \epsilon$ for (e) then yields the desired result.

(3) Under the v-PPC condition, everything works as in the case of the v-central condition except that the ESI's are replaced by inequalities in expectation over the sample $Z_1, \ldots, Z_n$.

6. Future Work

We have seen that while in the bounded loss case the β-Bernstein and $v(\epsilon) \asymp \epsilon^{1-\beta}$-PPC conditions are equivalent, in the unbounded loss case there is a strange discrepancy between them; a main goal for future work is to extend our bounds to cover faster rates under a weaker condition implied by either of Bernstein and PPC. A second goal is simply to establish whether or not the witness condition holds for commonly used classes (such as misspecified regression, common probability models with density estimation, etc.). A third goal is to extend our results to the case of VC-type classes, for which it seems any analysis must proceed via some sort of symmetrization. A key work along this direction is that of Audibert and Bousquet (2007), but the results in the inductive setting there still require some additional work before becoming fully inductive bounds. Another goal is to extend our ideas to the realm of empirical process-type methods, where optimal convergence rates for large models can be obtained. Finally, our bounds become useful mostly for ERM, two-part, and generalized Bayesian estimators. To apply the latter two, Learner must know the optimal η. In previous works (Audibert (2009) and many others) it is suggested to do this using e.g. cross-validation, but recent works such as Grünwald (2011) and Grünwald (2012) present 'safe Bayesian' methods for doing this which provably select the right η under Bernstein conditions and bounded losses. To apply these methods here, they should be generalized to unbounded losses, which, it seems, is feasible.

References

Jean-Yves Audibert. PAC-Bayesian statistical learning theory. *These de doctorat de lUniversité 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 Olivier Bousquet. Combining pac-bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8:863–889, 2007.

Andrew R. Barron and Thomas M. Cover. Minimum complexity density estimation. *Information Theory, IEEE Transactions on*, 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.

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.

Olivier Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, 2003.

Imre Csiszar. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

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.

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.

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

B.J.K. Kleijn and A.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.

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

Jonathan Qiang 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.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *Information Theory, IEEE Transactions on*, 52(10):4394–4412, 2006.

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

Nishant A. Mehta and Robert C. Williamson. From stochastic mixability to fast rates. In *Advances in Neural Information Processing Systems*, pages 1197–1205, 2014.

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

Shahar Mendelson. Learning without concentration for general loss functions. *arXiv preprint arXiv:1410.3192*, 2014b.

Shahar Mendelson. On aggregation for heavy-tailed classes. *arXiv preprint arXiv:1502.07097*, 2015.

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

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

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

Sara Van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

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

Tim van Erven, Peter D. Grünwald, Mark D. Reid, and Robert C. Williamson. Mixability in statistical learning. In *Advaces in Neural Information Processing Systems*, pages 1691–1699, 2012.

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.

Wing Hung Wong, Xiaotong Shen, et al. 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. *Information Theory, IEEE Transactions on*, 44(4):1424–1439, 1998.

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

Appendix A. Proof of Theorem 14 and Supporting Results

This section assembles the machinery for proving Theorem 14 and concludes with the proof thereof. We begin by working towards presenting Theorem 18, which goes most of the way in proving Theorem 14. To this end, we first sculpt a modified version of the witness condition for use only in the proofs (no additional assumptions are made). The proof of Theorem 18 itself navigates carefully around issues with unboundedness by way of this modified witness condition, and it further uses a critical lemma, Lemma 23, to suitably control the bounded part. Finally, we prove Theorem 14. As we will see, the key to making the proof work is to use instances of a generalized reversed information projection in just the right way.

A.1. PAC-Bayesian inequality

The following is a slight restatement of Theorem 2.1 of Zhang (2006b).

Theorem 16 *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:*

$$\mathbf{E}_{\underline{f}\sim\Pi_{|n}}\left[\mathbf{E}^{\mathrm{ANN}(\eta)}_{Z\sim P}\left[\ell_{\underline{f}} - \ell_{\phi(\underline{f})}\right]\right] \trianglelefteq_{\eta\cdot n} \mathrm{IC}_{n,\eta}\left(\phi(\underline{f}) \,\|\, \Pi_{|}\right). \tag{17}$$

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

$$\mathrm{IC}_{n,\eta}\left(\phi(\underline{f}) \,\|\, \Pi_{|}\right) := \mathbf{E}_{\underline{f}\sim\Pi|n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\ell_{\underline{f}}(Z_i) - \ell_{\phi(\underline{f})}(Z_i)\right)\right] + \frac{\mathrm{KL}(\Pi_{|n} \,\|\, \Pi_{|0})}{\eta\cdot n}. \tag{18}$$

By the finiteness considerations of Appendix D, $\mathrm{IC}_{n,\eta}(\phi(\underline{f}) \,\|\, \Pi_{|})$ 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; comparator functions will be critical to our application of Theorem 16. This theorem, in various forms, is folklore (see e.g. Zhang (2006a)). The result 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.

A.2. Proof of Theorem 14

Theorem 18 below represents a pivotal generalization of its special case Theorem 13 from the main text. We now work towards stating and proving this more general result.

For $f \in \mathcal{F}$, we work with the excess loss $\ell_f - \ell_{\phi(f)}$, where $\phi : \mathcal{F} \to L_1(P)$ is a *comparator map* which, for a given f, yields some $\phi(f)$ satisfying $\mathbf{E}[\ell_{\phi(f)}] \leq \mathbf{E}[\ell_f]$.

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 $f \in \mathcal{F}$,*

$$\mathbf{E}\left[\left(\ell_f - \ell_{\phi(f)}\right)\cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} \leq u(1 \vee (M^{-1}\,\mathbf{E}[\ell_f - \ell_{f^*}]))\}}\right] \geq c\,\mathbf{E}[\ell_f - \ell_{\phi(f)}]. \tag{19}$$

If we modify the RHS of (19) *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 $\psi : f \mapsto f^*$, 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, ℓ_{g_f}, an instance of a GRIP (see Definition 6).

The main motivation for our introducing the GRIP is that $(P, \ell, \mathcal{F})$ with comparator ℓ_g satisfies the η-central condition (from Proposition 11). 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 C 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" ℓ_{g_f}, for each f, defined by replacing $\mathcal{F}$ with $\{f^*, f\}$ in Definition 6. The mini-grip with respect to f then has the simple, characterizing property of satisfying

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

Also, as will be used to critical effect in the application of Theorem 18, for each f the learning problem $(P, \{f^*, f\}, \ell)$ with comparator ℓ_{g_f} satisfies the η-central condition.

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 $\phi : f \mapsto g_f$.[3]

Lemma 17 (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* $(P, \ell, \mathcal{F})$ *satisfies the weak witness condition with dynamic comparator* $\phi : f \mapsto g_f$ *with the same constants* (M, u, c)*.*

The above result and the next are proved in Appendix F.

Theorem 18 *Let* $\bar{\eta} > 0$*. For dynamic comparator* ϕ*, assume that* $\mathbf{E}\left[e^{-\bar{\eta}(\ell_f - \ell_{\phi(f)})}\right] \leq 1$*. Let* $u > 0$ *and* $c \in (0,1]$ *be constants for which* $\mathbf{E}\left[(\ell_f - \ell_{\phi(f)}) \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} \leq u\}}\right] \geq c\,\mathbf{E}[\ell_f - \ell_{f^*}]$*. Then for all* $\eta \in (0, \bar{\eta})$

$$\mathbf{E}[\ell_f - \ell_{f^*}] \;\leq\; c_1 \cdot \frac{1}{\eta} \mathbf{E}\left[1 - e^{-\eta(\ell_f - \ell_{\phi(f)})}\right] \;\leq\; c_1 \cdot \mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_f - \ell_{\phi(f)}\right],$$

with $c_1 := \frac{1}{c} \frac{\eta u + 1}{1 - \frac{\eta}{\bar{\eta}}}$.

[3] Technically, g_f need not be well-defined, but we will always use g_f only via ℓ_{g_f}, which *is* well-defined.

Hellinger mini-grip to GRIP Theorem 18 already is enough to obtain the implication of Theorem 14 under the strong central condition. However, we need to do a bit more work to obtain the results under the v-central condition and the PPC and v-PPC conditions.

Lemma 19 *Let g_f be the mini-grip with respect to η and f, and let g be the GRIP with respect to η.*

$$\frac{1}{\eta}\left(1 - \mathbf{E}\left[e^{-\eta(\ell_f - \ell_{g_f})}\right]\right) \le \frac{1}{\eta/2}\left(1 - \mathbf{E}\left[e^{-\frac{\eta}{2}(\ell_f - \ell_g)}\right]\right) \tag{20}$$

Proof Observe that

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

where the first inequality follows from Jensen's and for the second inequality we use the to-be-proved inequality

$$\mathbf{E}\left[e^{-\eta(\ell_{g_f} - \ell_g)}\right] \le 1. \tag{21}$$

We now prove (21). First, recall that $g_f = -\frac{1}{\eta}\log\left((1-\alpha)e^{-\eta\ell_{f^*}} + \alpha e^{-\eta\ell_f}\right)$. Using this representation:

$$\mathbf{E}\left[e^{-\eta(\ell_{g_f} - \ell_g)}\right] = (1-\alpha)\,\mathbf{E}\left[e^{-\eta(\ell_f^* - \ell_g)}\right] + \alpha\,\mathbf{E}\left[\alpha e^{-\eta(\ell_f - \ell_g)}\right] \le 1.$$

■

Next, we chain $1 - x \le -\log x$, Lemma 19, and Theorem 18 to obtain a bound that we will use in the proofs of Theorems 14 and 15.

Corollary 20 *Let $f \in \mathcal{F}$. Let g be the GRIP with respect to $\bar{\eta}$. Let $u > 0$ and $c \in (0,1]$ be constants for which $\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \le u\}}\right] \ge c\,\mathbf{E}[\ell_f - \ell_{f^*}]$. Then for all $\eta \in (0, \frac{\bar{\eta}}{2})$*

$$\mathbf{E}[\ell_f - \ell_{f^*}] \le c_2 \mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_f - \ell_g\right],$$

with $c_2 := \frac{1}{c}\frac{2\eta u + 1}{1 - \frac{2\eta}{\bar{\eta}}}$.

Proof Since $1 - x \le -\log x$, it holds that $\mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_f - \ell_g\right]$ is lower bounded as

$$\mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_f - \ell_g\right] \ge \frac{1}{\eta}\left(1 - \mathbf{E}\left[e^{-\eta(\ell_f - \ell_g)}\right]\right),$$

which, from Lemma 19, is further lower bounded by

$$\frac{1}{2\eta}\left(1 - \mathbf{E}\left[e^{-2\eta(\ell_f - \ell_{g_f})}\right]\right).$$

Finally, Lemma 17 establishes the weak witness condition with respect to comparator g_f (defined using $\bar{\eta}$), and from Proposition 11 this comparator further satisfies $\mathbf{E}\left[e^{-\bar{\eta}(\ell_f - \ell_{g_f})}\right] \leq 1$, so that we may apply Theorem 18 with $\phi(f) = g_f$ to further lower bound the above by $\frac{1}{c_2}\mathbf{E}[\ell_f - \ell_{f^*}]$. ■

Everything is now in place to prove Theorem 14.

Proof (of Theorem 14) Fix some $\epsilon \geq 0$. We define g and g_f (for each $f \in \mathcal{F}$) using $v(\epsilon)$. First, Theorem 16 states for our particular choice of η that

$$\mathbf{E}_{\underline{f}\sim\Pi_{|n}}\left[-\frac{1}{\eta}\log\mathbf{E}\left[e^{-\eta(\ell_{\underline{f}} - \ell_g)}\right]\right] \trianglelefteq_{\eta\cdot n} \mathbf{E}_{\underline{f}\sim\Pi_{|n}}\left[\frac{1}{n}\sum_{j=1}^{n}(\ell_{\underline{f}}(Z_j) - \ell_g(Z_j))\right] + \frac{\mathrm{KL}(\Pi_{|n} \,\|\, \Pi_{|0})}{\eta n}. \tag{22}$$

Next, Corollary 20 implies that, for $c_2 := \frac{1}{c}\frac{2\eta u+1}{1-\frac{2\eta}{v(\epsilon)}}$,

$$\mathbf{E}_{\underline{f}\sim\Pi_{|n}}\left[\mathbf{E}[\ell_{\underline{f}} - \ell_{f^*}]\right] \trianglelefteq_{\frac{\eta\cdot n}{c_2}} c_2 \cdot \left(\mathbf{E}_{\underline{f}\sim\Pi_{|n}}\left[\frac{1}{n}\sum_{j=1}^{n}(\ell_{\underline{f}}(Z_j) - \ell_g(Z_j))\right] + \frac{\mathrm{KL}(\Pi_{|n} \,\|\, \Pi_{|0})}{\eta n}\right). \tag{23}$$

Starting from (23), we now prove the results in the theorem statement. For completeness, we first prove a slight strengthening of part 1 when the strong $\bar{\eta}$-central condition holds. In this case, for all $f \in \mathcal{F}$ we have that $g = f^*$, we have $v(\epsilon) = \bar{\eta}$ for all $\epsilon \geq 0$, we actually may take any $\eta < \bar{\eta}$, we can apply Theorem 18 (with $\phi(f) = f^*$) rather than Corollary 20, and we hence can improve c_2 to the value c_1. Taking $\epsilon = 0$, (14) holds with $\trianglelefteq_{\frac{\eta\cdot n}{2}}$ replaced by $\trianglelefteq_{\eta\cdot n}$.

Part 1 - When the v-central condition holds. First, Proposition 8 implies that $\ell_{f^*} - \ell_g \trianglelefteq_{v(\epsilon)} \epsilon$. This fact, taken together with $Z_1, \ldots, Z_n$ being i.i.d. and $\eta < v(\epsilon)$, implies that

$$\frac{1}{n}\sum_{j=1}^{n}\left(\ell_{f^*}(Z_j) - \ell_g(Z_j)\right) \trianglelefteq_{\eta\cdot n} \epsilon.$$

By the weak-transitivity property of $\trianglelefteq$ (Proposition 3), we can add the above display (after multiplying the LHS and RHS by c_2 and replacing $\trianglelefteq_{\eta\cdot n}$ by $\trianglelefteq_{\frac{\eta\cdot n}{c_2}}$) to (23), yielding (14).

Part 2 - When the v-PPC condition holds. The exponential stochastic inequality in (23), taken together with part (i) of Proposition 2, implies the following in-expectation bound:

$$\mathbf{E}_{Z_1^n}\left[\mathbf{E}_{\underline{f}\sim\Pi_{|n}}\left[\mathbf{E}[\ell_{\underline{f}} - \ell_{f^*}]\right]\right] \;\leq\; c_2\,\mathbf{E}_{Z_1^n}\left[\mathbf{E}_{\underline{f}\sim\Pi_{|n}}\left[\frac{1}{n}\sum_{j=1}^{n}(\ell_{\underline{f}}(Z_j) - \ell_g(Z_j))\right] + \frac{\mathrm{KL}(\Pi_{|n} \,\|\, \Pi_{|0})}{\eta n}\right].$$

Now, from the v-PPC condition combined the convergence argument in the proof of Proposition 11 (i.e. there exists a sequence $\{\ell_{Q_k}\}$ converging to ℓ_g in $L_1(P)$) implies that $\mathbf{E}[\ell_{f^*}] \leq \mathbf{E}[\ell_g] + \epsilon$, implying the result (15). ■

Appendix B. Proof of Theorem 15

Proof (of Theorem 15) We only prove the result in the case of the v-central condition. For the bound under the v-PPC condition, replace all ESI's in the proof below with inequalities in expectation over Z_1^n, and observe that for a selector $\Pi_{|n}$ concentrated on $\hat{f}_n$:

$$\mathbf{E}\left[\Pi_{|n}\left(\{f : \mathbf{E}[\ell_f - \ell_{f^*}] \geq b_n\}\right)\right] = \mathbf{E}\left[\mathbf{1}_{\{\mathbf{E}[\ell_{\hat{f}_n} - \ell_{f^*}] \geq b_n\}}\right] = \Pr\left(\mathbf{E}[\ell_{\hat{f}_n} - \ell_{f^*}] \geq b_n\right).$$

We now prove the first result (under the v-central condition). Recall that c_2 is defined as $c_2 := \frac{1}{c}\frac{2\eta u+1}{1-\frac{2\eta}{\bar{\eta}}}$. Fix some $\epsilon > 0$ and take $\bar{\eta} := v(\epsilon)$ and some $\eta < \frac{\bar{\eta}}{2}$.

Observe that

$$\begin{aligned}
&\Pi_{|n}(\{f : \mathbf{E}[\ell_f - \ell_{f^*}] \geq b_n\}) \\
&\quad \leq \Pi_{|n}\big(\{f : \mathbf{E}[\ell_f - \ell_{f^*}] > M\}\big) && (24) \\
&\quad\quad + \Pi_{|n}\big(\{f : b_n \leq \mathbf{E}[\ell_f - \ell_{f^*}] \leq M\}\big). && (25)
\end{aligned}$$

We bound each part in turn. To proceed, we will shift the analysis to the GRIP ℓ_g with respect to $\mathcal{F}$ and $\bar{\eta}$.

To bound (24), for functions with excess risk larger than M we establish a constant lower bound on the Hellinger divergence with respect to the comparator ℓ_{g_f}. This lower bound is a simple consequence of Corollary 20 with the u from that result taken to be $u\frac{\mathbf{E}[\ell_f - \ell_{f^*}]}{M}$, yielding:

$$\frac{M}{c_2} \leq \mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_f - \ell_g\right]. \tag{26}$$

Now, (24) may be bounded by making use of (26):

$$\begin{aligned}
\Pi_{|n}\big(\{f : \mathbf{E}[\ell_f - \ell_{f^*}] > 1\}\big)\big) &\leq \Pi_{|n}\big(\{f : \mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_f - \ell_g\right] > \frac{M}{c_2}\}\big)\big) \\
&\leq \frac{c_2}{M}\,\mathbf{E}_{\underline{f}\sim\Pi_{|n}}[\mathbf{E}^{\mathrm{ANN}(\eta)}\left[\ell_{\underline{f}} - \ell_g\right]] \\
&\trianglelefteq_{\frac{M}{c_2}\cdot n\cdot\eta}\, \frac{c_2}{M}\mathrm{IC}_{n,\eta}(g \,\|\, \Pi_|),
\end{aligned}$$

where we employed the generalized notion of information complexity from (18).

Under the v-central condition, Proposition 8 implies that

$$\ell_{f^*} - \ell_g \trianglelefteq_{v(\epsilon)} \epsilon, \tag{27}$$

which in turn implies that

$$\frac{c_2}{M}\mathrm{IC}_{n,\eta}(f^* \,\|\, \Pi_|) - \frac{c_2}{M}\mathrm{IC}_{n,\eta}(g \,\|\, \Pi_|) \trianglelefteq_{\frac{M}{c_2}n\cdot\eta}\, \frac{c_2}{M}\epsilon.$$

Thus, we have

$$\Pi_{|n}\big(\{f : \mathbf{E}[\ell_f - \ell_{f^*}] > 1\}\big) \;\trianglelefteq_{\frac{M}{c_2}\cdot\frac{n\cdot\eta}{2}}\; \frac{c_2}{M}\left(\mathrm{IC}_{n,\eta}(f^* \,\|\, \Pi_|) + \epsilon\right).$$

We proceed to bound (25). Observe that if $b_n > M$, then the measure is zero. Thus, this case can also be handled by any nonnegative upper bound developed in the case of $b_n \leq M$; assume we are in this latter case. Now, for any f satisfying $\mathbf{E}[\ell_f - \ell_{f^*}] \leq M$, the threshold in Definition 12 simplifies to u. Thus, for such f, we may apply Corollary 20 (with the same u and c) to yield

$$\mathbf{E}[\ell_f - \ell_{f^*}] \leq c_2 \cdot \mathbf{E}^{\mathrm{ANN}(\eta)} \left[\ell_f - \ell_g\right].$$

Leveraging this inequality, we have

$$\Pi_{|n}\big(\{f : b_n \leq \mathbf{E}[\ell_f - \ell_{f^*}] \leq 1\}\big) \leq \Pi_{|n}\big(\{f : \mathbf{E}^{\mathrm{ANN}(\eta)} \left[\ell_f - \ell_g\right] \geq \frac{b_n}{c_2}\}\big).$$

Now, Proposition 11 implies that $\ell_g - \ell_f \trianglelefteq_\eta 0$, and hence the random variable $\mathbf{E}^{\mathrm{ANN}(\eta)} \left[\ell_f - \ell_g\right]$ is nonnegative. We therefore may apply Markov's inequality, yielding the upper bound

$$\begin{aligned}\Pi_{|n}\big(\{f : \mathbf{E}^{\mathrm{ANN}(\eta)} \left[\ell_f - \ell_g\right] \geq \frac{b_n}{c_2}\}\big) &\leq \frac{c_2}{b_n} \mathbf{E}_{\underline{f} \sim \Pi_{|n}} \left[\mathbf{E}^{\mathrm{ANN}(\eta)} \left[\ell_{\underline{f}} - \ell_g\right]\right] \\ &\trianglelefteq_{\frac{b_n}{c_2} \cdot n \cdot \eta} \frac{c_2}{b_n} \mathrm{IC}_{n,\eta}(g \,\|\, \Pi_|),\end{aligned}$$

where the first inequality holds for all samples Z_1^n and the ESI follows from Theorem 16.

Using (27) as before, we have $\frac{c_2}{b_n} \mathrm{IC}_{n,\eta}(g \,\|\, \Pi_|) \trianglelefteq_{\frac{b_n}{c_2} \cdot n \cdot \eta} \frac{c_2}{b_n} \mathrm{IC}_{n,\eta}(f^* \,\|\, \Pi_|) + \frac{c_2}{b_n} \epsilon$, and so we finally get

$$\Pi_{|n}\big(\{f : b_n \leq \mathbf{E}[\ell_f - \ell_{f^*}] \leq 1\}\big) \trianglelefteq_{\frac{b_n}{c_2} \cdot \frac{n \cdot \eta}{2}} \frac{c_2}{b_n} \left(\mathrm{IC}_{n,\eta}(f^* \,\|\, \Pi_|) + \epsilon\right).$$

The result follows by combining the two ESI's using the weak-transitivity property of $\trianglelefteq$ (Proposition 3, with the η there taken as $(b_n \wedge M) \cdot \frac{n \cdot \eta}{4 c_2}$). ■

Appendix C. 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 ℓ_g of P onto $\mathrm{conv}(\mathcal{E}_{\mathcal{F},\eta})$. Because $\mathcal{F}$ and η are fixed throughout, we adopt the notation $\mathcal{E} := \mathcal{E}_{\mathcal{F},\eta}$ and $\mathcal{C} := \mathrm{conv}(\mathcal{E}_{\mathcal{F},\eta})$.

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

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

Let us 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.

C.1. Proving q^* exists

Throughout, we will need to assume the existence of a certain sequence $\{q_n\}$ in $\mathcal{C}$ such that $\mathrm{KL}(p; q_0 \,\|\, q_n) < \infty$ for all n. This is not problematic, as we now explain. Recall the definition of the mix-loss $\ell_Q = -\frac{1}{\eta} \log \mathbf{E}_{\underline{f}\sim Q}\big[e^{-\eta \ell_{\underline{f}}}\big]$ for any distribution Q over $\mathcal{F}$. Under the η-PPC condition up to ϵ (and hence a fortiori under the η-central condition up to ϵ),[4] it holds that

$$\mathbf{E}\left[\ell_{f^*}(Z)\right] \leq \inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}[\ell_Q] + \epsilon,$$

and hence

$$-\epsilon \leq \inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}[\ell_Q] - \mathbf{E}\left[\ell_{f^*}(Z)\right] \leq 0.$$

Also, from the property of the infimum, for any $\delta > 0$, there exists Q_δ for which

$$\inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}[\ell_Q] \leq \mathbf{E}[\ell_{Q_\delta}] \leq \inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}[\ell_Q] + \delta,$$

and so the same Q_δ satisfies

$$-\epsilon \leq \mathbf{E}[\ell_{Q_\delta}] - \mathbf{E}\left[\ell_{f^*}(Z)\right] \leq \delta.$$

It therefore follows that for any sequence $\{\delta_n\}$ there exists a corresponding sequence $\{Q_n\}$ for which $\mathbf{E}[\ell_{Q_n} - \ell_{f^*}] \to \inf_{Q \in \Delta(\mathcal{F})} \mathbf{E}[\ell_Q - \ell_{f^*}]$ and $\mathbf{E}[\ell_{Q_n} - \ell_{f^*}] \in [-\epsilon, \delta_n]$.

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

Let $\{q_n\}$ be a sequence in $\mathcal{C}$ for which $\mathrm{KL}(p; q_0 \,\|\, q_n) \to \inf_{q \in \mathcal{C}} \mathrm{KL}(p; q_0 \,\|\, q)$. Taking $q_0 = e^{-\ell_{f^*}}$, 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 $\bar{q}_n$.

[4] It might not be necessary to appeal to the PPC or central conditions, but doing so makes proving finiteness much easier. Since all our results rely on these conditions, we are free to exploit them here.

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 $\{\bar{q}_n\}$ 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 $\log(\bar{q}_n)$ 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 $\mathrm{KL}(p; q_0 \,\|\, q^*) = \lim_{n \to \infty} \mathrm{KL}(p; q_0 \,\|\, \bar{q}_n)$ since

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

Appendix D. Details on Infinity, Affinity and Exponential Stochastic Inequality

D.1. Definitions and conventions concerning ∞ and $-\infty$

Since we allow loss functions to take on the value ∞ and to be unbounded both from above and from below, we 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 $f : \mathcal{U} \to \mathbb{R} \cup \{-\infty, \infty\}$. 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.

To ensure well-defined expectations, we need two conditions. First, we restrict ourselves to learning problems with distribution P and loss function ℓ such that for all $f \in \bar{\mathcal{F}}$, we have:

$$\mathbf{E}_{Z \sim P}[(\ell_f(Z))^-] > -\infty. \tag{28}$$

Whenever in the main text we refer to a learning problem, we automatically assume that (28) holds. This is a very mild requirement: it will automatically hold if the loss is bounded from below, which is the case for all loss functions we usually encounter except for the log loss, and even for the log loss it is always the case except if we have continuous sample spaces $\mathcal{Y}$. Example 5 below shows that even then it is a natural requirement.

The second condition needed for well-defined expectations is similar: we also require, relative to given learning problem with model $\mathcal{F}$, for every distribution Π on $\mathcal{F}$, that for all $z \in \mathcal{Z}$,

$$\mathbf{E}_{\underline{f} \sim \Pi}[(\ell_{\underline{f}}(z))^-] > -\infty. \tag{29}$$

Again, whenever in the main text we write 'a distribution on $\mathcal{F}$', we mean a distribution for which (29) holds; whenever we refer to a learning algorithm $\Pi_|$ relative to a given learning problem, we assume that it is designed such that, for all $z_1, \ldots, z_n \in \mathcal{Z}^n$, for all $0 \leq i \leq n$, (29) holds with Π set to $\Pi | z^i$. Again, this condition holds automatically for all the usual loss functions (since they are bounded below) except for log loss with continuous $\mathcal{Y}$. For that case, (29) is a real restriction. For the case of η-Bayesian estimators with $\eta \leq 1$ (our primary interest with log loss) we can conveniently enforce it by imposing a natural condition on the prior Π: $\mathbf{E}_{\underline{f} \sim \Pi}[\exp(-\ell_{\underline{f}}(z))] = \int p_f(z) d\Pi(f) < \infty$. Then the requirement holds for Π and for η-Bayesian estimators as defined underneath (1) it will then automatically hold for $\Pi_{|i}$, for all $0 \leq i \leq n$ as well.

Example 5 (Density Estimation) When the observed data has zero density according to some p_f with $f \in \mathcal{F}$, the log loss becomes infinite; such cases are easily handled by the definitions above. However, if (and only if) the space $\mathcal{Y} = \mathbb{R}^k$ is uncountable, there is another complication for log loss: while for each fixed z and f, $-\log p_f(z) > -\infty$, we may have that $\mathbf{E}\big[\big(\ell_f(Z)\big)^-\big] =$

$\mathbf{E}[(-\log p_f(Z))^-] = -\infty$, where either $f \in \mathcal{F}$ is fixed and the expectation is over $Z \sim P$; or Z is a fixed $z \in \mathcal{Z}$ and the expectation is over $f = \underline{f} \sim \Pi$, Π being the output of some learning algorithm. If we allowed this, viz. the definitions above, the expected loss could become undefined. We will prevent this from happening by requiring (28) and (29). The first is a natural requirement, since for all likelihood-based estimators it will automatically hold if P itself has a density p relative to the given underlying measure μ. For in that case, we can replace each density p_f by density $p'_f := p_f / p$. Since $p(Z) > 0$ almost surely, $-\log p'_f(Z)$ is well-defined and $> -\infty$ and hence is a well-defined extended random variable, and we have (Csiszar, 1975) $\mathbf{E}_{Z \sim P}[-\log p'_f] = \text{KL}(P \| P_f) \geq 0$, KL denoting KL-divergence. Hence, if we worked with densities p'_f rather than p_f then (28) would hold automatically as long as P has a density. Now if we work with a learning algorithm whose output is a function of the likelihood and a prior Π (as it is for, e.g., Bayesian, 2-part MDL, and maximum likelihood estimators), then the output (a distribution over $\mathcal{F}$) given data z^n remains unaffected if we base inference on p_f or p'_f — in the latter case we are merely dividing the likelihood by a constant which is the same for all $f \in \mathcal{F}$, hence the relative likelihood remains unchanged.

To see how the second requirement (29) pans out, consider the Gaussian scale family with $\mathcal{Z} = \mathcal{Y} = \mathbb{R}$ and $\{p_f \mid f \in \mathcal{F}\}$ where $\mathcal{F} = \mathbb{R}^+$ and $p_f(y) \propto \exp(-y^2/f)$, i.e. p_f represents the normal distribution with mean 0 and variance $\sigma^2 := f$. Then under log loss we have $\ell_f(y) = \frac{y^2}{f} + \frac{1}{2}\log(\pi f)$. Now, if the prior places sufficient mass around 0, for instance via a prior $\pi(j^{-1}) \propto j^{-2}$ for $j = 1, 2, \ldots$, and if moreover, P places a point mass at 0, then (29) can fail to hold. Hence, our requirement prohibits this sort of prior here. □

Extended Fubini and well-defined stochastic inequalities Under the assumption that (28) and (29) hold, we can apply Lemma 3.1. of Grünwald and Dawid (2004), essentially an extension of Fubini's theorem, which gives that

$$\mathbf{E}_{Z \sim P}[\mathbf{E}_{\underline{f} \sim \Pi}[\ell_{\underline{f}}(Z)]] = \mathbf{E}_{\underline{f} \sim \Pi}[\mathbf{E}_{Z \sim P}[\ell_{\underline{f}}(Z)]] = \mathbf{E}_{(Z, \underline{f}) \sim P \otimes \Pi}[\ell_{\underline{f}}(Z)],$$

the final expectation being over the product distribution of P and Π. This allows us to exchange expectations relative to P and Π whenever convenient, and we will do so in the main text without explicitly mentioning it.

Moreover, as long as $\tilde{f}$ is a nontrivial comparator, the quantity $\ell_{\tilde{f}}(Z) - \ell_f(Z)$ is a well-defined extended random variable under $Z \sim P$. If $Z = z$ for a fixed z with $\ell_{\tilde{f}}(z) < \infty$, it is also a well-defined extended random variable under all distributions Π on $\mathcal{F}$ satisfying (29). Since throughout the text we only use nontrivial comparators and invariably assume (29), this ensures that $\ell_{\tilde{f}}(Z) - \ell_f(Z)$ and hence also $\exp(\eta(\ell_{\tilde{f}}(Z) - \ell_f(Z)))$ are well-defined extended random variables, so that our exponential-stochastic-inequality statements in the main text are all well-defined.

Finally, using again that $\exp(\eta(\ell_{\tilde{f}}(Z) - \ell_f(Z)))$ is well-defined we can now apply Lemma 3.1. of Grünwald and Dawid (2004) again to give us that, for all $\eta > 0$,

$$\mathbf{E}_{Z \sim P}\left[\mathbf{E}_{\underline{f} \sim \Pi}\left[e^{\eta(\ell_{\tilde{f}} - \ell_{\underline{f}})}\right]\right] = \mathbf{E}_{\underline{f} \sim \Pi}\left[\mathbf{E}_{Z \sim P}\left[e^{\eta(\ell_{\tilde{f}} - \ell_{\underline{f}})}\right]\right] = \mathbf{E}_{(Z, \underline{f}) \sim P \otimes \Pi}\left[e^{\eta(\ell_{\tilde{f}} - \ell_{\underline{f}})}\right] \in [0, \infty],$$

which implies that we can again exchange expectations over $\underline{f}$ and Z if we so desire; we will again do so in the text without explicit mention.

D.2. Affinity

Proposition 21

(a) If $\mathbf{E}[e^{-\eta X}] < \infty$, *we have*

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

(b) $\eta \mapsto \mathbf{E}^{\mathrm{ANN}(\eta)}[X]$ *is non-increasing.*

Before proceeding to the proof, we remark that the condition $\mathbf{E}[e^{-\eta X}]$ will always hold when we invoke part (a) of the above proposition, as the central condition will hold whenever we use (a); in fact even the v-central condition is sufficient.

Proof 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 $\mathrm{sign}(x e^{-\eta x} - \frac{1 - e^{-\eta x}}{\eta}) = -\mathrm{sign}(e^{\eta x} - (\eta x + 1)) \leq 0$.

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

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

Taking a sequence of $\alpha = \alpha_j \in \{\alpha_i\}_i$ going to zero, starting at some $\alpha_1 < 1$, we have that for all j that $x \mapsto \frac{1}{\alpha_j \bar{\eta}}\left(1 - e^{-\alpha_j \bar{\eta} x}\right) - \frac{1}{\bar{\eta}}(1 - e^{-\bar{\eta} x})$ 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

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

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 $\eta' \le \eta$, the concavity of $x \mapsto x^{\eta'/\eta}$ together with Jensen's inequality implies that

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

■

Appendix E. Examples

E.1. Proof for heavy-tailed regression with bounded predictions example

Proof (of claims in Example 2) The excess loss is of the form

$$\ell_f - \ell_{f^*} = (f(X) - f^*(X)) \cdot (-2Y + f(X) + f^*(X)).$$

To see that a Bernstein condition holds, observe that

$$\begin{aligned} (\ell_f - \ell_{f^*})^2 &\le (f(X) - f^*(X))^2 \cdot 4 \max\left\{(f(X) - Y)^2, (f^*(X) - Y)^2\right\} \\ &\le 4(f(X) - f^*(X))^2 \cdot \max\left\{(Y - r)^2, (Y + r)^2\right\}, \end{aligned}$$

and so

$$\begin{aligned} \mathbf{E}\left[(\ell_f - \ell_{f^*})^2\right] &\le 4\,\mathbf{E}\left[\mathbf{E}\left[(f(X) - f^*(X))^2 \cdot \max\left\{(Y - r)^2, (Y + r)^2\right\} \mid X\right]\right] \\ &= 4\,\mathbf{E}\left[(f(X) - f^*(X))^2\, \mathbf{E}\left[\max\left\{(Y - r)^2, (Y + r)^2\right\} \mid X\right]\right] \\ &\le 8(\sqrt{C} + r)^2\, \mathbf{E}\left[(f(X) - f^*(X))^2\right], \end{aligned}$$

where in the last step we summed the terms in the maximum and applied Hölder's inequality.

Next, if it holds that

$$\mathbf{E}\,(f^*(X) - Y)(f(X) - f^*(X))] \ge 0, \tag{30}$$

then it is easy to verify that

$$\mathbf{E}\left[(f(X) - f^*(X))^2\right] \le \mathbf{E}[\ell_f - \ell_{f^*}].$$

The condition (30) holds for all $f \in \mathcal{F}$ under our assumption that the risk minimizer f^* over $\mathcal{F}$ continues to be a minimizer when taking the minimum risk over the convex hull of $\mathcal{F}$. To see this, we observe that if we instead consider the function class $\operatorname{conv}(\mathcal{F})$, then f^* is still a minimizer and (30) holds for all $f \in \operatorname{conv}(\mathcal{F})$ from Mendelson (2015) (see the text around equation (1.3) therein).

Putting the above together, we see a Bernstein condition does indeed hold:

$$\mathbf{E}\left[(\ell_f - \ell_{f^*})^2\right] \le 8(\sqrt{C} + r)^2\, \mathbf{E}[\ell_f - \ell_{f^*}].$$

Next, we show that the Bernstein condition with exponent 1 and multiplicative constant A implies the witness condition. Let $u > 0$ be a to-be-determined constant. Then

$$\begin{aligned}
\mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} > u\}}\right] &\leq \mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \frac{\ell_f - \ell_{f^*}}{u} \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \geq 0\}}\right] \\
&= \frac{1}{u} \mathbf{E}\left[\ell_f - \ell_{f^*}^2 \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \geq 0\}}\right] \\
&\leq \frac{1}{u} \mathbf{E}\left[\ell_f - \ell_{f^*}^2\right] \\
&\leq \frac{A}{u} \mathbf{E}\left[\ell_f - \ell_{f^*}\right].
\end{aligned}$$

It follows that for any choice $u \geq A$, the witness condition holds with constant $c = 1 - \frac{A}{u}$. ■

E.2. Comparative examples

Example 6 (Estimation of means with second moment) Let $\mathcal{Z} = \mathcal{Y}$, let the model be $\mathcal{F} = [\mu_0, \mu_1]$, and take Y to have mean $\mu \in [\mu_0, \mu_1]$. Take squared loss, $\ell_\nu(Y) = (\nu - Y)^2$. Then we are in the well-specified setting and $f^* = \mu$. Assume for some $s \geq 2$ that we have $\mathbf{E}[|Y|^s] < \infty$; in particular, the variance σ^2 is then finite. Then, for any $\nu \in \mathcal{F}$, we have

$$\mathbf{E}[(\nu - Y)^2 - (\mu - Y)^2] = (\nu - \mu)^2$$

and (from tedious algebra)

$$\mathbf{E}[((\nu - Y)^2 - (\mu - Y)^2)^2] = \left((\nu - \mu)^2 + 4\sigma^2\right)(\nu - \mu)^2$$

Thus, the Bernstein condition holds with exponent 1 and constant $B = \left((\mu_1 - \mu_0)^2 + 4\sigma^2\right)^{-1}$. Applying Corollary 6.2 of Audibert (2009) then implies (after discretization) a rate of $\tilde{O}\left(\frac{\log n}{n}\right)$ in expectation using the SeqRand algorithm. The $\tilde{O}$ notation here hides an at most $\log n$ factor due to a (purely theoretical) discretization argument using a uniform ϵ-net.

On the other hand, without subexponential tail decay, the v-central condition fails to hold for any non-trivial v, but, as shown by van Erven et al. (2015, Example 5.10), the v-PPC condition holds for $v(\epsilon) = O(\epsilon^{2/s})$.[5] As we showed in Example 2, the witness condition holds if $\mathbf{E}[|Y|^2] < \infty$ (i.e. $s = 2$). Thus, for $s \geq 2$, Theorem 14 implies a rate of $\tilde{O}(n^{-s/(s+2)})$ in expectation, where, similar to before, the notation hides an at most $(\log n)^{s/(s+2)}$ factor due to discretization.

Notably, the SeqRand algorithm incorporates a second-order loss-difference term that appears to be the key to its superior performance in this example. □

Example 7 (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.

[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.

It follows that $f^* = f_1$, and for every $j > 1$ we have $\mathbf{E}[\ell_{f_j} - \ell_{f^*}] = \frac{3a}{8} + 1$. Thus, the excess risk is bounded for all f_j. The witness condition holds because for all $j > 1$ we have $\Pr(\ell_{f_j} - \ell_{f^*} = 1) = a$ and $\mathbf{E}[(\ell_{f_j} - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_{f_j} - \ell_{f^*} \leq 1\}}] \geq \frac{3a}{8}$. 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 $\mathbf{E}[(\ell_{f_j} - \ell_{f^*})^2] = a + j^2 \to \infty$ 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 14 still applies while e.g. the results of Zhang (2006b) and Audibert (2009) do not (see Section 4.1). □

Example 8 (Bernstein condition does not hold, unbounded excess risk I) 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 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 $\mathbf{E}[\ell_{f_\mu} - \ell_{f^*}] = \frac{1}{2}\nu^2$.

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

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

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

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

which is at most

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

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 9 (Small-ball assumption violated) To properly compare to the small-ball assumption of Mendelson (2014a), we consider regression with squared loss in the well-specified setting, so that the parameter estimation error bounds of Mendelson (2014a) 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:

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

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 $\mathbf{E}[\ell_{f_j} - \ell_{f^*}] = p_j$.

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

$$\begin{aligned}
&\mathbf{E}\left[\left(\ell_{f_j}(Z) - \ell_{f^*}(Z)\right) \cdot \mathbf{1}_{\{\ell_{f_j}(Z)-\ell_{f^*}(Z)>u\}}\right] \\
&= \mathbf{E}\left[\left(f_j^2(X) - 2f_j(X)Y\right) \cdot \mathbf{1}_{\{f_j^2(X)-2f_j(X)Y>u\}}\right] \\
&= p_j\, \mathbf{E}\left[(1-2Y) \cdot \mathbf{1}_{\{1-2Y>u\}}\right] \\
&= p_j \left(\Pr\left(Y < \frac{1-u}{2}\right) - 2\,\mathbf{E}\left[Y \cdot \mathbf{1}_{\{Y<\frac{1-u}{2}\}}\right]\right).
\end{aligned} \tag{31}$$

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 (31) 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 14. □

Appendix F. Proofs

Proof (of Lemma 17 (Witness Protection Lemma)) Let f be arbitrary. For brevity we define $u' := u(1 \vee (M^{-1}\,\mathbf{E}[\ell_f - \ell_{f^*}]))$. Observe that

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

Rewriting, we have

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

which we rearrange as

$$\begin{aligned}
\mathbf{E}\left[(\ell_f - \ell_{g_f}) \cdot \mathbf{1}_{\{\ell_f - \ell_{g_f} \leq u'\}}\right] &\geq \mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u'\}}\right] + \mathbf{E}[\ell_f - \ell_{g_f}] - \mathbf{E}[\ell_f - \ell_{f^*}] \\
&= \mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u'\}}\right] + \mathbf{E}[\ell_{f^*} - \ell_{g_f}] \\
&\geq \mathbf{E}\left[(\ell_f - \ell_{f^*}) \cdot \mathbf{1}_{\{\ell_f - \ell_{f^*} \leq u'\}}\right].
\end{aligned}$$

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

Proof (of Theorem 18) For any $\eta \in [0, \bar{\eta}]$, define:

$$h_{f,\eta} \coloneqq \frac{1}{\eta}\left(1 - e^{-\eta(\ell_f - \ell_{\phi(f)})}\right) \qquad S_{f,\eta} \coloneqq h_{f,\eta} - h_{f,\bar{\eta}} \qquad \mathrm{H}_{\eta,f} \coloneqq \mathbf{E}[h_{f,\eta}].$$

Note that $\mathbf{E}[\ell_f - \ell_{\phi(f)}] = \mathbf{E}[h_{f,0}]$ from part (a) of Proposition 21. Therefore, we may rewrite the excess risk of f as

$$\begin{aligned}\mathbf{E}[\ell_f - \ell_{\phi(f)}] &= \mathbf{E}[h_{f,0} - h_{f,\bar{\eta}} + h_{f,\bar{\eta}}] \\ &= \mathbf{E}[S_{f,0}] + \mathrm{H}_{\bar{\eta},f}.\end{aligned}$$

Splitting up the expectation into two components, we have

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

Now, from Lemma 22 (stated and proved immediately after this proof) and using $\bar{C} \coloneqq C_{\bar{\eta},\eta,u}$ to avoid cluttering notation, we have

$$\begin{aligned}\mathbf{E}[\ell_f - \ell_{\phi(f)}] &\le \bar{C}\,\mathbf{E}[S_{f,\eta} \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} \le u\}}] + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} > u\}}] + \mathrm{H}_{\bar{\eta},f} \\ &\le \bar{C}\,\mathbf{E}[S_{f,\eta}] + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} > u\}}] + \mathrm{H}_{\bar{\eta},f} \\ &= \bar{C}\left(\mathrm{H}_{\eta,f} - \mathrm{H}_{\bar{\eta},f}\right) + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} > u\}}] + \mathrm{H}_{\bar{\eta},f} \\ &= \bar{C}\mathrm{H}_{\eta,f} - (\bar{C} - 1)\mathrm{H}_{\bar{\eta},f} + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} > u\}}].\end{aligned}$$

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

$$\mathbf{E}[\ell_f - \ell_{\phi(f)}] \le \bar{C}\mathrm{H}_{\eta,f} + \mathbf{E}[S_{f,0} \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} > u\}}].$$

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

$$\mathbf{E}[\ell_f - \ell_{\phi(f)}] - \mathbf{E}[(\ell_f - \ell_{\phi(f)}) \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} > u\}}] \le \bar{C}\mathrm{H}_{\eta,f},$$

which can be rewritten as

$$\mathbf{E}[(\ell_f - \ell_{\phi(f)}) \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} \le u\}}] \le \bar{C}\mathrm{H}_{\eta,f},$$

Now, since we assume that

$$c\,\mathbf{E}[\ell_f - \ell_{f^*}] \le \mathbf{E}[(\ell_f - \ell_{\phi(f)}) \cdot \mathbf{1}_{\{\ell_f - \ell_{\phi(f)} \le u\}}],$$

the first inequality is proved:

$$\mathbf{E}[\ell_f - \ell_{f^*}] \le \frac{\bar{C}}{c}\mathrm{H}_{\eta,f}.$$

Finally, $1 - x \le -\log x$ yields the second inequality. ■

Lemma 22 ("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 $C_{\bar{\eta},\eta,u} := \frac{\eta u + 1}{1 - \frac{\eta}{\bar{\eta}}}$.

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 g_η as $g_\eta(r) = \frac{1}{\eta}(1 - r^\eta) - (1 - r)$. From Lemma 23, 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 $C_{\bar{\eta},\eta,u} = \frac{\eta u + 1}{1 - \frac{\eta}{\bar{\eta}}}$. ■

Lemma 23 *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 $r \ge \frac{1}{V}$:*

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

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

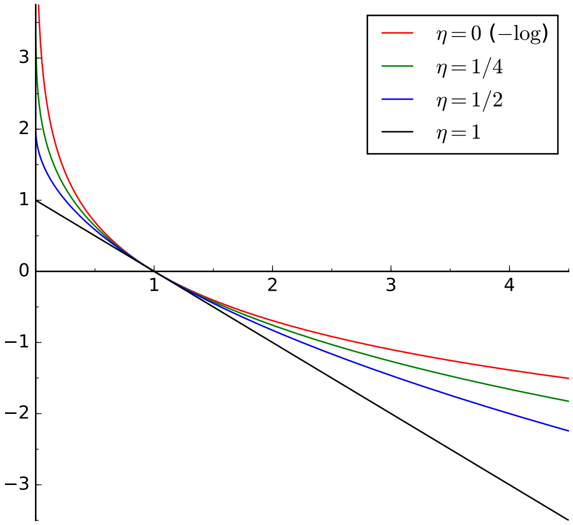

Figure 1: The function $r :\to \eta^{-1}(1 - r^{\eta})$ for various values of r. $g_\eta(r)$ is the difference of the line for η at r and the line for $\eta = 1$ at r, which is always positive.

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 decreases: 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

$$s(r) = \frac{(-r^{\eta'} + r) \cdot g_\eta(r) + (r^\eta - r) \cdot g_{\eta'}(r)}{g_\eta(r)^2}. \tag{32}$$

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 a easier to analyze. (32) shows that for $r > 0, r \neq 1$, $h'(r) = 0$, i.e. h reaches an extremum, iff $s(r) = 0$, i.e. iff the numerator in (32) 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

$$f'(1) = \frac{1}{2} \cdot \frac{1-\eta'}{1-\eta} \left(\eta' - \eta\right) \ , \ h'(1) = \frac{1}{3} \cdot \frac{1-\eta'}{1-\eta} \left(\eta' - \eta\right), \tag{33}$$

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

$$h < f \text{ on } (L, 1) \quad ; \quad h > f \text{ on } (1, R). \tag{34}$$

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 (34), 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 (33), $h'(r)$ 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 $0 \leq \eta' < \eta < 1$, 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 $y \in (a, b)$,

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

Straightforward differentiation and simplification gives that $u'_a(y) = ay^{a-1}(a - 1)(1 - y^{-1})$ 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 (32) 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 $t(r) \leq s(r)$ for all $r \leq 1$. We already know that $h_{\eta',\eta}$ is decreasing, so that $s(r) \leq 0$ for all r, so we have $t(r) \leq s(r) \leq 0$ for all $r \leq 1$. 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 $0 < r \leq 1$. Then $\tilde{h}'(r) = -(\eta/(1-\eta))r^{-1} \leq t(r)r^{-1} \leq s(r)r^{-1} = h'_{0,\eta}(r) \leq 0$ 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 $\tilde{h}(r) \geq h_{0,\eta}(r)$ for $0 < r < 1$. We can thus bound $h_{0,\eta}(1/V)$ by $\tilde{h}(1/V)$, and the result follows. ■