From exp-concavity to variance control: $O(1/n)$ rates and online-to-batch conversion with high probability

Nishant A. Mehta
Centrum Wiskunde & Informatica (CWI)
mehta@cwi.nl

Abstract

We present an algorithm for the statistical learning setting with a bounded exp-concave loss in $d$ dimensions that obtains excess risk $O(d/n)$ with high probability: the dependence on the confidence parameter $\delta$ is polylogarithmic in $1/\delta$. The core technique is to boost the confidence of recent in-expectation $O(d/n)$ excess risk bounds for empirical risk minimization (ERM), without sacrificing the rate, by leveraging a Bernstein condition which holds due to exp-concavity. This Bernstein condition implies that the variance of excess loss random variables are controlled in terms of their excess risk. Using this variance control, we further show that a regret bound for any online learner in this setting translates to a high probability excess risk bound for the corresponding online-to-batch conversion of the online learner. We also show that with probability $1 - \delta$ the standard ERM method obtains excess risk $O(d\log(n) + \log(1/\delta))/n)$.

1 Introduction

In the statistical learning problem, a learning agent observes a samples of $n$ points $Z_1, \ldots, Z_n$ drawn i.i.d. from an unknown distribution $P$ over an outcome space $Z$. The agent then seeks an action $f$ in an action space $F$ that minimizes their expected loss, or risk, $E_{Z \sim P}[\ell(f, Z)]$, where $\ell$ is a loss function $\ell: F \times Z \rightarrow \mathbb{R}$. Several recent works have studied this problem in the situation where the loss is exp-concave and bounded, $F$ and $Z$ are subsets of $\mathbb{R}^d$, and $F$ is convex. Mahdavi et al. (2015) were the first to show that there exists a learner for which, with probability at least $1 - \delta$, the excess risk decays at the rate $d\log(n) + \log(1/\delta))/n)$. Through novel algorithmic stability arguments applied to empirical risk minimization (ERM), Koren and Levy (2015) and Gonen and Shalev-Shwartz (2016) succeeded in discarding the $\log n$ factor to obtain a rate of $d/n$, but their bounds only hold in expectation. All three works highlighted the open problem of obtaining a high probability excess risk bound with the rate $d\log(1/\delta)/n$. Whether this is possible is far from a trivial question in light of a result of Audibert (2008): when learning over a finite class with bounded exp-concave losses satisfying some additional regularity assumptions, a progressive mixture rule (a Cesàro mean of pseudo-Bayesian estimators) with appropriate learning rate obtains expected excess risk $O(1/n)$ but, for any learning rate, these rules suffer from severe deviations of order $\sqrt{\log(1/\delta)/n}$.

We resolve the high probability question: we present a learning algorithm with an excess risk bound (Corollary 1) which has rate $(d\log(1/\delta) + (\log(1/\delta))^2)/n$ with probability $1 - \delta$. In fact, ERM already obtains $O((d\log(n) + \log(1/\delta))/n)$ excess risk, which apparently was not known although implicit in the literature. To vanquish the $\log n$ factor with the small polylog$(1/\delta)$ price it suffices to run a two-phase ERM method based on a confidence-boosting device. The key to our analysis is connecting exp-concavity to the central condition of Van Erven et al. (2015), which in turn implies a Bernstein condition. We then exploit the variance control of the excess loss random variables afforded by the Bernstein condition to boost the confidence trick of Schapire (1990).

In the next section we formally define the setting and describe the previous $O(d/n)$ in-expectation bounds. We present the results for standard ERM and our confidence-boosted ERM method in Sections 3 and 4 respectively. In Section 5, we discuss a brief history of the work in this area. Finally, Section
extend the results of Kakade and Tewari (2009) to exp-concave losses, showing that under a bounded loss assumption a regret bound for any online exp-concave learner transfers to a high probability excess risk bound via an online-to-batch conversion. This extension comes at no additional technical price: it is a consequence of the variance control implied by exp-concavity, control leveraged by Freedman’s inequality for martingales to obtain a fast rate with high probability. This result continues the line of work of Cesa-Bianchi et al. (2001) and Kakade and Tewari (2009) and accordingly is about the generalization ability of online exp-concave learning algorithms.

2 Exp-concave statistical learning and in-expectation bounds

We now describe the setting more formally. In this work \( F \) is always assumed to be convex. We say a function \( A: F \to \mathbb{R} \) has diameter \( C \) if \( \sup_{f_1, f_2 \in F} |A(f_1) - A(f_2)| \leq C \). Assume for each \( z \in Z \) that the loss map \( \ell(\cdot, z): f \mapsto \ell(f, z) \) is \( \eta \)-exp-concave, i.e. \( f \mapsto e^{-\eta \ell(f, z)} \) is concave over \( F \). We further assume, for each outcome \( z \), that the loss \( \ell(\cdot, z) \) has diameter \( B \). We adopt the notation \( \ell_f(z) := \ell(f, z) \).

Given a sample of \( n \) points drawn i.i.d. from an unknown distribution \( P \) over \( Z \), our objective is to select a hypothesis \( f \in F \) that minimizes the excess risk \( E_{Z \sim P} [\ell_f(Z)] - \inf_{f \in F} E_{Z \sim P} [\ell_f(Z)] \). We assume there exists \( f^* \in F \) satisfying \( E[\ell_f(Z)] = \inf_{f \in F} E_{Z \sim P} [\ell_f(Z)] \); this assumption also was made by Gonen and Shalev-Shwartz (2016) and Kakade and Tewari (2009).

Let \( A_F \) be an algorithm, defined for a function class \( F \) as a mapping \( A_F: \bigcup_{n \geq 0} \mathbb{Z}^n \to F \); we drop the subscript \( F \) when it is clear from the context. Our starting point will be an algorithm \( A \) which, when provided with a sample \( Z \) of \( n \) i.i.d. points, satisfies an expected risk bound of the form

\[
E_{Z \sim P^n} [E_{Z \sim P} [\ell_{A(f, Z)}(Z) - \ell_{f^*}(Z)]] \leq \psi(n). \tag{1}
\]

Koren and Levy (2015) and Gonen and Shalev-Shwartz (2016) both established in-expectation bounds of the form (1) that obtain a rate of \( O(d/n) \) in the case when \( F \subset \mathbb{R}^d \), each in a slightly different setting. Koren and Levy (2015) assume, for each outcome \( z \in Z \), that the loss \( \ell(\cdot, z) \) has diameter \( B \) and is \( \beta \)-smooth, i.e. for all \( f, f' \in F \), the gradient is \( \beta \)-Lipschitz:

\[
\|\nabla_f \ell(f, z) - \nabla_f \ell(f', z)\|_2 \leq \beta \|f - f'\|_2.
\]

They also use a 1-strongly convex regularizer \( \Gamma: F \to \mathbb{R} \) with diameter \( R \). Under these assumptions, they show that ERM run with the weighted regularizer \( \frac{1}{n} \Gamma \) has expected excess risk at most

\[
\psi(n) = \frac{1}{n} \left( \frac{24\beta d}{n} + 100Bd + R \right).
\]

It is not known if the smoothness assumption is necessary to eliminate the \( \log n \) factor.

Gonen and Shalev-Shwartz (2016) work in a slightly different setting that captures all known exp-concave losses; whether they capture the entire class of exp-concave losses is not clear. They assume that the loss is of the form \( \ell_f(z) = \phi_y(f(x)) \), for \( F \subset \mathbb{R}^d \). They further assume, for each \( z = (x, y) \), that the mapping \( f \mapsto \phi_y(f(x)) \) is \( \alpha \)-strongly convex and \( L \)-Lipschitz, but they do not assume smoothness. They show that standard, unregularized ERM has expected excess risk at most

\[
\psi(n) = \frac{2L^2d}{\alpha n} = \frac{2d}{\eta n},
\]

where \( \eta = \alpha/L^2 \); the purpose of the rightmost expression is that the loss is \( \eta \)-exp-concave. Although this bound ostensibly is independent of the loss’s diameter \( B \), the dependence may be masked by \( \eta \): for logistic loss, \( \eta = e^{-B}/4 \), while squared loss admits the more favorable \( \eta = 1/(4B)^2 \).

\[1\]This assumption is not explicit from Koren and Levy (2015), but their other assumptions might imply it. In any case, provided that the guarantees of Koren and Levy (2015) and Gonen and Shalev-Shwartz (2016) hold, our analysis in Section 4 can be adapted to work if the infimal risk is not achieved, i.e. if \( f^* \in F \) does not exist.
3 A high probability bound for ERM

Before showing how to get a high probability \(O(d/n)\) excess risk bound, as a warm-up we first show that ERM itself obtains excess risk \(O(d \log(n)/n)\) with high probability; here and elsewhere, if \(\delta\) is omitted the dependence is \(\text{polylog}(1/\delta)\). That ERM satisfies such a bound was largely implicit in the literature, and so we present this result explicitly here. The closest such known result, Theorem 1 of Mahdavi and Jin (2014), is not directly applicable as it relies on an additional assumption (see Assumption (I) therein). Our assumptions subtly differ from elsewhere in this work. We assume that \(\mathcal{F} \subset \mathbb{R}^d\) satisfies \(\sup_{f', f \in \mathcal{F}} \|f - f'\|_2 \leq R\) and that, for each outcome \(z \in \mathcal{Z}\), the loss \(\ell(\cdot, z)\) is \(L\)-Lipschitz and \(|\ell_f(z) - \ell_{f'}(z)| \leq B\). The first two assumptions already imply the last for \(B = LR\). All these assumptions were made by Mahdavi and Jin (2014) and Koren and Levy (2015), sometimes implicitly, and while Gonen and Shalev-Shwartz (2016) only make the Lipschitz assumption, for all known \(\eta\)-exp-concave losses the constant \(\eta\) depends on \(B\) (which itself typically will depend on \(R\)).

The first, critical observation is that exp-concavity implies good concentration properties of the excess loss random variable. This is easiest to see by way of the \(\eta\)-central condition, which the excess loss satisfies. This concept, studied by Van Erven et al. (2015) and first introduced by Van Erven et al. (2012) as “stochastic mixability”, is defined as follows.

**Definition 1 (Central condition)** We say that \((P, \ell, \mathcal{F})\) satisfies the \(\eta\)-central condition for some \(\eta > 0\) if there exists a comparator \(f^* \in \mathcal{F}\) such that, for all \(f \in \mathcal{F}\),

\[
E_{Z \sim P} \left[ e^{-\eta(\ell_f(Z) - \ell_{f^*}(Z))} \right] \leq 1.
\]

Jensen’s inequality implies that if this condition holds, the corresponding \(f^*\) must be a risk minimizer.

Next, we show that in our setting \((P, \ell, \mathcal{F})\) satisfies the \(\eta\)-central condition.

**Lemma 1** Let \(\mathcal{F}\) be convex. Take \(\ell\) to be a loss function \(\ell: \mathcal{F} \times \mathcal{Z} \to \mathbb{R}\), and assume that, for each \(z \in \mathcal{Z}\), the map \(\ell(\cdot, z): f \mapsto \ell(f, z)\) is \(\eta\)-exp-concave. Then, for all distributions \(P\) over \(\mathcal{Z}\), if there exists an \(f^* \in \mathcal{F}\) that minimizes the risk under \(P\), then \((P, \ell, \mathcal{F})\) satisfies the \(\eta\)-central condition.

With the central condition in our grip, Theorem 7 of Mehta and Williamson (2014) directly implies an \(O(d \log(n)/n)\) bound for ERM; however, a considerably simpler version of that result yields much smaller constants. The version below, proved in the appendix for completeness, only requires an \((\varepsilon/L)\)-net of \(\mathcal{F}\) in the \(\ell_2\) norm, which induces an \(\varepsilon\)-net of \(\{\ell_f: f \in \mathcal{F}\}\) in the sup norm.

**Theorem 1** Let \(\mathcal{F} \subset \mathbb{R}^d\) be a convex set satisfying \(\sup_{f, f' \in \mathcal{F}} \|f - f'\|_2 \leq R\). Suppose, for all \(z \in \mathcal{Z}\), that the loss \(\ell(\cdot, z)\) is \(\eta\)-exp-concave and \(L\)-Lipschitz. Let \(\sup_{z \in \mathcal{Z}, f \in \mathcal{F}} |\ell_f(z) - \ell_{f^*}(z)| \leq B\). Then if \(n \geq 5\), with probability at least \(1 - \delta\), ERM learns a hypothesis \(\hat{f}\) with excess risk bounded as

\[
E_{Z \sim P}[\ell_{\hat{f}}(Z) - \ell_{f^*}(Z)] \leq \frac{1}{n} \left( 8 \left( B \vee \frac{1}{\eta} \right) \left( d \log(16LRn) + \log \frac{1}{\delta} \right) + 1 \right). \tag{2}
\]

4 Boosting the confidence for high probability bounds

The two existing excess risk bounds mentioned in Section 2 decay at the rate \(1/n\). A naïve application of Markov’s inequality unsatisfyingly yields excess risk bounds of order \(\psi(n)/\delta\) that hold with probability \(1 - \delta\). In this section, we present and analyze our meta-algorithm, CONFIDENCEBOOST, which boosts these in-expectation bounds to hold with probability at least \(1 - \delta\) at the price of a factor polylogarithmic in \(1/\delta\). This method is essentially the “boosting the confidence” trick of Schapire (1990); the novelty lies in a refined analysis that exploits a Bernstein-type condition to improve the rate in the final high probability bound from the typical \(O(1/\sqrt{n})\) to the desired \(O(1/n)\).

Our analysis of CONFIDENCEBOOST actually applies more generally than the exp-concave learning setting, requiring only that \(A\) satisfy an in-expectation bound of the form (1), the loss \(\ell(\cdot, z)\) have bounded diameter for each \(z \in \mathcal{Z}\), and the problem \((P, \ell, \mathcal{F})\) satisfy a \((C, q)\)-Bernstein condition.

---

See also the exposition by Kearns and Vazirani (1994, Chapter 4.2).
Definition 2 (Bernstein condition) We say that \((\mathcal{P}, \ell, \mathcal{F})\) satisfies the \((C, q)\)-Bernstein condition for some \(C > 0\) and \(q \in (0, 1]\) if there exists a comparator \(f^* \in \mathcal{F}\) such that, for all \(f \in \mathcal{F}\),

\[
E_{\mathcal{Z} \sim \mathcal{P}} \left[ \left( \ell_f(Z) - \ell_{f^*}(Z) \right)^2 \right] \leq C \ E_{\mathcal{Z} \sim \mathcal{P}} \left[ \ell_f(Z) - \ell_{f^*}(Z) \right]^q .
\]

Before getting to CONFIDENCEBoost, we first show that in the exp-concave learning setting the Bernstein condition holds with best possible exponent, \(q = 1\), so that this setting is indeed a special case of the more general setting we analyze. Recall from Lemma 1 that the \(\eta\)-central condition holds for \((\mathcal{P}, \ell, \mathcal{F})\). The next lemma shows that the \(\eta\)-central condition, together with the boundedness of the loss, implies that a Bernstein condition holds.

Lemma 2 (Central to Bernstein) Let \(X\) be a random variable taking values in \([-B, B]\). Assume that \(E[e^{-\eta X}] \leq 1\). Then \(E[X^2] \leq 4/(1 + \eta) E[X]\).

Boosting the boosting the confidence trick. First, consider running \(\mathcal{A}\) on a sample \(Z_1, \ldots, Z_n\) of i.i.d. points. The excess risk random variable \(E[Z(\mathcal{A}(Z_1)), Z] - \ell_{f^*}(Z)]\) is nonnegative, and so Markov’s inequality implies that the expected excess risk bounded by \(\psi(n)\) imply that

\[
\Pr \{ E[Z(\mathcal{A}(Z_1)), Z] - \ell_{f^*}(Z)] \geq c \cdot \psi(n) \} \leq \frac{1}{c}.
\]

Now, let \(Z_1, \ldots, Z_K\) be independent samples, each of size \(n\). Running \(\mathcal{A}\) on each sample yields \(\hat{f}_1 := \mathcal{A}(Z_1), \ldots, \hat{f}_K := \mathcal{A}(Z_K)\). Applying Markov’s inequality as above, combined with independence, implies that with probability at least \(1 - e^{-K}\) there exists \(j \in [K]\) such that \(E_{Z \sim P} [\ell_{\hat{f}_j}(Z) - \ell_{f^*}(Z)] \leq c \cdot \psi(n)\). Let us call this good event GOOD.

Our quest is now to show that on event GOOD, we can identify any of the hypotheses \(\hat{f}_1, \ldots, \hat{f}_K\) approximately satisfying \(E[Z, \mathcal{A}(Z)], Z) - \ell_{f^*}(Z)] \leq c \cdot \psi(n)\), where by “approximately” we mean up to some slack that weakens the order of our resulting excess risk bound by a multiplicative factor of at most \(K\). As we will see, it suffices to run ERM over this finite subclass using a fresh sample. The proposed meta-algorithm is presented in Algorithm 1.

Algorithm 1: CONFIDENCEBoost

| Input: \(Z_1, \ldots, Z_{K+1} \overset{iid}{\sim} P^n\); learner \(\mathcal{A}_F\) |
| for \(j = 1 \to K\) do \(\hat{f}_j = \mathcal{A}_F(Z_j)\) return |
| ERM \(F_K(Z_{K+1})\), with \(F_K = \{\hat{f}_1, \ldots, \hat{f}_K\}\) |

Analysis. From here on out, we treat the initial sample of size \(Kn\) as fixed and de-hat the \(\hat{K}\) estimators above, referring to them as \(f_1, \ldots, f_K\). It is without loss of generality that we further assume that they are sorted in order of increasing risk (breaking ties arbitrarily). Our goal now is to show that running ERM on the finite class \(F_K := \{f_1, \ldots, f_K\}\) yields low excess risk with respect to comparator \(f_1\). A typical analysis of the boosting the confidence trick would apply Hoeffding’s inequality to select a risk minimizer optimal to resolution \(1/\sqrt{n}\), but this is not good enough here. As a further boost to the trick, this time with respect to its resolution, we will establish that a Bernstein condition holds over a particular subclass of \(F_K\) with high probability, which will in turn imply that ERM obtains \(O(1/n^{1/(2-q)})\) excess risk over \(F_K\).

We first establish an approximate Bernstein condition for \((\mathcal{P}, \ell, F_K)\). Observe that for all \(f_j \in F_K\),

\[
\|\ell_{f_j} - \ell_{f_1}\|_{L_2(P)} \leq \|\ell_{f_j} - \ell_{f^*}\|_{L_2(P)} + \|\ell_{f_1} - \ell_{f^*}\|_{L_2(P)}.
\]

and so from the \((C, q)\)-Bernstein condition

\[
\|\ell_{f_j} - \ell_{f_1}\|_{L_2(P)}^2 \leq C \left( E[|\ell_{f_j} - \ell_{f^*}|^2] + E[|\ell_{f_1} - \ell_{f^*}|^2] \right) + 2 \left( E[|\ell_{f_j} - \ell_{f^*}|^q] \cdot E[|\ell_{f_1} - \ell_{f^*}|^q] \right) \leq C \left( 3 E[|\ell_{f_j} - \ell_{f^*}|^q] + E[|\ell_{f_1} - \ell_{f^*}|^q] \right) \leq C \left( 3 E[|\ell_{f_j} - \ell_{f_1}|^q] + 4 E[|\ell_{f_1} - \ell_{f^*}|^q] \right),
\]

where the last step follows because the map \(x \mapsto x^q\) is concave and hence subadditive.

We call this bound an approximate Bernstein condition because, on event GOOD, for all \(f_j \in F_K\):

\[
\|\ell_{f_j} - \ell_{f_1}\|_{L_2(P)}^2 \leq C \left( 3 E[|\ell_{f_j} - \ell_{f_1}|^q] + 4(c \cdot \psi(n))^q \right).
\]
Define the class $\mathcal{F}_K := \{f_1\} \cup \{f_j \in \mathcal{F}_K : \mathbb{E}[\ell_{f_j} - \ell_{f_1}] \geq 4^{1/q_e} \cdot \psi(n)\}$. Then with probability $\Pr(\text{GOOD}) \geq 1 - e^{-K}$, the problem $(P, \ell, \mathcal{F}_K)$ satisfies the $(4C, q)$-Bernstein condition.

We now analyze the outcome of running ERM on $\{f_1, \ldots, f_k\}$ using a fresh sample of $n$ points. The next lemma shows that ERM performs favorably under a Bernstein condition, a well-known result.

**Lemma 3** Let $\mathcal{G}$ be a finite class of functions $\{f_1, \ldots, f_K\}$ and assume without loss of generality that $f_1$ is a risk minimizer. Let $\mathcal{G}' \subset \mathcal{G}$ be a subclass for which, for all $f \in \mathcal{G}'$:

$$\mathbb{E}[(\ell_f - \ell_{f_1})^2] \leq C \mathbb{E}[\ell_f - \ell_{f_1}]^q,$$

and $\ell_f - \ell_{f_1} \leq B$ almost surely. Then, with probability at least $1 - \delta$, ERM run on $\mathcal{G}$ will not select any function $f$ in $\mathcal{G}'$ whose excess risk satisfies

$$\mathbb{E}[\ell_f - \ell_{f_1}] \geq \left( \frac{2(C + \frac{B^2 - q}{3}) \log \left( \frac{\mathcal{G}' - 1}{\delta} \right)}{n} \right)^{1/(2-q)}.$$

Applying Lemma 3 with $\mathcal{G} = \mathcal{F}_K$ and $\mathcal{G}' = \mathcal{F}'_K$, we see that with probability at least $1 - \delta$ over the fresh sample, ERM selects a function $f_j$ falling in one of two cases:

- $\mathbb{E}_{Z \sim p}[\ell_{f_j}(Z) - \ell_{f_1}(Z)] \leq 4^{1/q_e} \cdot \psi(n)$;
- $\mathbb{E}_{Z \sim p}[\ell_{f_j}(Z) - \ell_{f_1}(Z)] \leq \left( \frac{2(C + \frac{B^2 - q}{3}) \log \frac{\mathcal{G}' - 1}{\delta}}{n} \right)^{1/(2-q)}$ (using $|\mathcal{F}_K'| - 1 \leq K$).

Now, consider running CONFIDENCEBoost with $K = \lceil \log(2/\delta) \rceil$ on a sample of $n$ points, so that $n_e = n/\lceil K + 1 \rceil$; for simplicity, assume that $K + 1$ divides $n$. Then taking the failure probability for the ERM phase to be $\delta/2$, CONFIDENCEBoost admits the following guarantee.

**Theorem 2** Let $(P, \ell, \mathcal{F})$ satisfy the $(C, q)$-Bernstein condition, and assume for all $z \in \mathcal{Z}$ that the loss $\ell(\cdot, z)$ has diameter $B$. Impose any necessary assumptions such that algorithm $A$ obtains a bound of the form (1). Then, with probability at least $1 - \delta$, CONFIDENCEBoost run with $K = \log(2/\delta)$ and $n_e = n/K$ learns a hypothesis $\hat{f}$ with excess risk $\mathbb{E}_{Z \sim p}[\ell_{\hat{f}}(Z) - \ell_{f_1}(Z)]$ at most

$$e \cdot \psi\left( \frac{n}{\log \frac{2}{\delta}} \right) + \max \left\{ 4^{1/q_e} \cdot \psi\left( \frac{n}{\log \frac{2}{\delta}} \right), \left( \frac{4(C + \frac{B^2 - q}{3}) \log \frac{\mathcal{G}' - 1}{\delta}}{n} \right)^{1/(2-q)} \right\}.$$

The next result for the exp-concave learning setting of interest is immediate.

**Corollary 1** Applying Theorem 2 with $A_F$ the algorithm of Koren and Levy (2015) and their assumptions, the excess risk bound in Theorem 2 specializes to

$$O\left( \left( \frac{\beta d/n + B d + R}{n} \right) \log \frac{1}{\delta} + (1/\eta + B + 1) \left( \log \frac{1}{\delta} \right)^2 \right).$$

Similarly taking $A_F$ the algorithm of Gonen and Shalev-Shwartz (2016) and their assumptions yields

$$O\left( \left( \frac{d/\eta}{n} \right) \log \frac{1}{\delta} + (1/\eta + B + 1) \left( \log \frac{1}{\delta} \right)^2 \right).$$

**Remarks.** As we saw from Lemmas 1 and 2, in the exp-concave setting a Bernstein condition holds for the class $\mathcal{F}$. A natural inquiry is if one could use this Bernstein condition to show directly a high probability fast rate of $O(d/n)$ for ERM. Indeed, under strong convexity, Sridharan et al. (2009) show that a similar bound for ERM is possible; however, they used strong convexity to bound a localized complexity. It is unclear if exp-concavity can be used to bound a localized complexity, and the Bernstein condition alone seems insufficient; such a bound may be possible via ideas from the local norm analysis of Koren and Levy (2015). While we think controlling a localized complexity from exp-concavity is a very interesting and worthwhile direction, we leave this to future work.
the Online Newton Step (ONS) algorithm. This algorithm has bypassing the Bernstein condition entirely.

In the case of infinite classes, such as when $\mathcal{F} \subset \mathbb{R}^d$ as we consider here, Hazan et al. (2007) designed the Online Newton Step (ONS) algorithm. This algorithm has $O(d \log T)$ regret over $T$ rounds, which, after online-to-batch conversion yields $O(d \log(n)/n)$ excess risk in expectation. Until recently, it was unclear whether one could obtain a similar high probability result. However, Mahdavi et al. (2015) showed that an online-to-batch conversion of ONS does indeed enjoy excess risk bounded by $O(d \log(n)/n)$ with high probability. While this resolved the statistical complexity of learning up to log $n$ factors, ONS (though efficient) can have a high computational complexity of $O(d^3)$ even in simple cases like learning over the unit $\ell_2$ ball, and in general its complexity may be as high as $O(d^2)$ per projection step (Hazan et al., 2007; Koren, 2013).

If one hopes to eliminate the log $n$ factor, the additional hardness of the online setting makes it unlikely that one can proceed via an online-to-batch conversion approach. Moreover, computational considerations suggest circumventing ONS anyways. In this vein, as we discussed in Section 2 both Koren and Levy (2015) and Gonen and Shalev-Shwartz (2016) recently established in-expectation excess risk bounds for a lightly penalized ERM algorithm and ERM itself respectively, without resorting to an online-to-batch conversion. Notably, both works developed arguments based on algorithmic stability, thereby circumventing the typical reliance on chaining-based arguments to discard log $n$ factors. In analogy to the empirical star algorithm of Audibert (2008), which for convex $\mathcal{F}$ reduces to ERM itself, we conjecture that ERM also enjoys excess risk bounded by $O((d + \log(1/\delta))/n)$ with high probability; note that the conjectured effect of log $(1/\delta)$ is additive rather than multiplicative. Table 1 summarizes what is known and our new results.

| Algorithm                | In-expectation | With probability $1 - \delta$ |
|--------------------------|----------------|-------------------------------|
| Progressive mixture rule | $O(\log |\mathcal{F}|/n)$ | $\Omega(\sqrt{\log(1/\delta)}/n)$ |
| Empirical star           | $O(\log |\mathcal{F}|/n)$ | $O((\log |\mathcal{F}| + \log(1/\delta))/n)$ |
| Online Newton Step       | $O(d \log n/n)$ | $O(d \log n + \log(1/\delta))/n)$ |
| EWOO                     | $O((d \log n + \log(1/\delta))/n)$ | $O((d \log n + \log(1/\delta))/n)$ |
| ERM                      | $O(d/n)$      | $O((d \log(1/\delta) + (\log(1/\delta))^2)/n)$ |
| Boosted ERM              | —             | $O((d \log(1/\delta) + (\log(1/\delta))^2)/n)$ |

Table 1: Excess risk bounds, with new results in bold. Boosted ERM applies CONFIDENCEBOOST to ERM. “ERM” refers to either penalized ERM (Koren and Levy, 2015) or ERM (Gonen and Shalev-Shwartz, 2016). For simplicity we only show dependence with respect to $d$, $n$, and $\delta$.

5 Previous work: On the log $n$ factor

Learning under exp-concave losses with finite classes dates back to the seminal work of Vovk (1990) and the game of prediction with expert advice, with the first explicit treatment for exp-concave losses due to Kivinen and Warmuth (1999). Vovk (1990) showed that if a game is $\eta$-mixable (which is implied by $\eta$-exp-concavity), one can guarantee that the worst-case individual sequence regret against the best of $K$ experts is at most $\frac{\log K}{\eta}$. An online-to-batch conversion then implies an in-expectation excess risk bound of the same order in the stochastic i.i.d. setting.

Quite relatedly, Audibert (2008) showed that when learning over a finite class with certain exp-concave losses, no progressive mixture rule can obtain a high probability excess risk bound of order better than $\sqrt{\log(1/\delta)/n}$. Note that Audibert’s result actually holds for progressive indirect mixture rules, which correspond to Vovk’s Aggregating Algorithm (when using the same learning rate and substitution function in both algorithms). Audibert (2008) overcame this deviations shortcoming of progressive mixture rules by devising the empirical star algorithm, which first runs ERM on $\mathcal{F}$, obtaining $f_{\text{ERM}}$, and then runs ERM a second time on the star convex hull of $\mathcal{F}$ with respect to $f_{\text{ERM}}$. This algorithm does achieve $O(\log |\mathcal{F}|/n)$ with high probability; the rate was only proved for squared loss with targets $y$ in $[-1, 1]$, but it was claimed that the result can be extended to general, bounded losses $y \mapsto \ell(y, \hat{y})$ satisfying smoothness and strong convexity as a function of predictions $\hat{y}$ (not hypotheses $f$). After such an extension, the empirical star algorithm could replace ERM in the second phase of CONFIDENCEBOOST, bypassing the Bernstein condition entirely.

In the case of infinite classes, such as when $\mathcal{F} \subset \mathbb{R}^d$ as we consider here, Hazan et al. (2007) designed the Online Newton Step (ONS) algorithm. This algorithm has $O(d \log T)$ regret over $T$ rounds, which, after online-to-batch conversion yields $O(d \log(n)/n)$ excess risk in expectation. Until recently, it was unclear whether one could obtain a similar high probability result. However, Mahdavi et al. (2015) showed that an online-to-batch conversion of ONS does indeed enjoy excess risk bounded by $O(d \log(n)/n)$ with high probability. While this resolved the statistical complexity of learning up to log $n$ factors, ONS (though efficient) can have a high computational complexity of $O(d^3)$ even in simple cases like learning over the unit $\ell_2$ ball, and in general its complexity may be as high as $O(d^2)$ per projection step (Hazan et al., 2007; Koren, 2013).

If one hopes to eliminate the log $n$ factor, the additional hardness of the online setting makes it unlikely that one can proceed via an online-to-batch conversion approach. Moreover, computational considerations suggest circumventing ONS anyways. In this vein, as we discussed in Section 2 both Koren and Levy (2015) and Gonen and Shalev-Shwartz (2016) recently established in-expectation excess risk bounds for a lightly penalized ERM algorithm and ERM itself respectively, without resorting to an online-to-batch conversion. Notably, both works developed arguments based on algorithmic stability, thereby circumventing the typical reliance on chaining-based arguments to discard log $n$ factors. In analogy to the empirical star algorithm of Audibert (2008), which for convex $\mathcal{F}$ reduces to ERM itself, we conjecture that ERM also enjoys excess risk bounded by $O((d + \log(1/\delta))/n)$ with high probability; note that the conjectured effect of log $(1/\delta)$ is additive rather than multiplicative. Table 1 summarizes what is known and our new results.
6 High probability online-to-batch-conversion with log $n$ factors

In Section 4, we saw how to get high probability bounds that avoided $\log n$ factors. The present section’s purpose is to show that if one is willing to accept the additional $\log n$ factor in a high probability bound, then it is sufficient to use an online-to-batch conversion of an online exp-concave learner whose worst-case cumulative regret (over $T$ rounds) is logarithmic in $T$. Using such a conversion, it is easy to get an excess risk bound with the additional $\log n$ factor that holds in expectation. The key difficulty is making such a bound hold with high probability. This result provides an alternative to the high probability $O(\log n/n)$ result for ERN in Section 3.

Before we begin, we note that Mahdavi et al. (2015) previously considered an online-to-batch conversion of ONS and established the first explicit high probability $O(\log n/n)$ excess risk bound in the exp-concave statistical learning setting. Their analysis is quite elegant but seems to be intimately coupled to ONS; it consequently is unclear if their analysis can be used to grasp excess risk bounds by online-to-batch conversions of other online exp-concave learners. This leads us to our next point and a new path: it is possible to transfer regret bounds to high probability excess risk bounds via online-to-batch conversion for general online exp-concave learners. Our analysis builds almost entirely off of the analysis of Kakade and Tewari (2009) in the strongly convex setting.

On the generalization ability of online exp-concave learning algorithms. We first consider a different, related setting: online convex optimization (OCO) under a $B$-bounded, $\nu$-strongly convex loss that is $L$-Lipschitz with respect to the action. An OCO game unfolds over $T$ rounds. An adversary first selects a sequence of $T$ convex loss functions $c_1, \ldots, c_T$. In round $t$, the online learner plays $f_t \in F$, the environment subsequently reveals cost function $c_t$, and the learner suffers loss $c_t(f_t)$. Note that the adversary is oblivious, and so the learner does not necessarily need to randomize. Because we are interested in analyzing the statistical learning setting, we constrain the adversary to play a sequence of $T$ points $z_1, \ldots, z_T \in Z$, inducing cost functions $\ell(\cdot, z_1), \ldots, \ell(\cdot, z_T)$.

Consider an online learner that sequentially plays actions $f_1, \ldots, f_T \in F$ in response to $z_1, \ldots, z_T$, so that $f_t$ depends on $(z_1, \ldots, z_{t-1})$. The (cumulative) regret is defined as

$$\sum_{t=1}^{T} \ell(f_t(z_t)) - \inf_{f \in F} \sum_{t=1}^{T} \ell(f(z_t)).$$

When the losses are bounded, strongly convex, and Lipschitz, Kakade and Tewari (2009) showed that if there is an online algorithm with regret $R_T$ on an i.i.d. sequence $Z_1, \ldots, Z_T \sim P$, online-to-batch conversion by simple averaging of the iterates $f_T := \frac{1}{T} \sum_{t=1}^{T} f_t$ admits the following guarantee.

**Theorem 3 (Corollary 5 of Kakade and Tewari (2009))** For all $z \in Z$, assume that $\ell(\cdot, z)$ is bounded by $B$, $\nu$-strongly convex, and $L$-Lipschitz. Then with probability at least $1 - 4 \log(T)\delta$ the action $\hat{f}_T$ satisfies excess risk bound

$$\mathbb{E}_{Z \sim P}[\ell_{\hat{f}_T}(Z) - \ell_{f^*}(Z)] \leq \frac{R_T}{T} + 4 \sqrt{\frac{L^2 \log \frac{1}{\delta}}{\nu} \frac{\sqrt{R_T}}{T} + \max \left\{ \frac{16L^2}{\nu}, 6B \right\} \frac{\log \frac{1}{\delta}}{T}}.$$

Under various assumptions, there are OCO algorithms that obtain worst-case regret (under all sequences $z_1, \ldots, z_T$) $R_T = O(\log T)$. For instance, Online Gradient Descent (Hazan et al., 2007) admits the regret bound $R_T \leq \frac{G^2}{2T}(1 + \log T)$, where $G$ is an upper bound on the gradient.

What if we relax strong convexity to exp-concavity? As we will see, it is possible to extend the analysis of Kakade and Tewari (2009) to $\eta$-exp-concave losses. Of course, such a regret-to-excess-risk bound conversion is useful only if we have online algorithms and regret bounds to start with. Indeed, at least two such algorithms and bounds exist, due to Hazan et al. (2007):

- **ONS**, with $R_T \leq 5 \left( \frac{1}{\eta} + GD \right) d \log T$, where $G$ is a bound on the gradient and $D$ is a bound on the diameter of the action space.
that Var is equal to easiness the actual regret can be involve a log

Lemma 4 (Conditional variance control) Define the Martingale difference sequence

\[ \xi_t := E_Z [\ell_{f_t}(Z) - \ell_{f^*}(Z)] - (\ell_{f_t}(Z_t) - \ell_{f^*}(Z_t)). \]

Then

\[ \text{Var} [\xi_t \mid Z_1, \ldots, Z_{t-1}] \leq 4 \left( \frac{1}{\eta} + B \right) E_Z [\ell_{f_t}(Z) - \ell_{f^*}(Z)]. \]

**Proof** Observe that \( \text{Var} [\xi_t \mid Z_1, \ldots, Z_{t-1}] = \text{Var} [\ell_{f_t}(Z_t) - \ell_{f^*}(Z_t) \mid Z_1, \ldots, Z_{t-1}] \). Treating the sequence \( Z_1, \ldots, Z_{t-1} \) as fixed and also treating \( f_t \) as a fixed parameter, the above conditional variance is equal to \( \text{Var} [\ell_f(Z) - \ell_{f^*}(Z)] \), where the randomness lies entirely in \( Z \sim P \). Then, Lemma 2 implies that \( \text{Var} [\ell_f(Z) - \ell_{f^*}(Z)] \leq 4 \left( \frac{1}{\eta} + B \right) E [\ell_f(Z) - \ell_{f^*}(Z)] \).

The next corollary is from a retrace of the proof of Theorem 2 of Kakade and Tewari (2009).

**Corollary 2** For all \( z \in \mathcal{Z} \), let \( \ell(\cdot, z) \) be bounded by \( B \) and \( \eta \)-exp-concave with respect to the action \( f \in \mathcal{F} \). Then with probability at least \( 1 - \delta \), for any \( T \geq 3 \), the excess risk of \( \bar{f}_T \) is at most

\[ \frac{R_T}{T} + 4 \sqrt{\left( \frac{1}{\eta} + B \right) \log \frac{4 \log T}{\delta} \cdot \sqrt{\frac{R_T}{T}} + 16 \left( \frac{1}{\eta} + B \right) \log \frac{4 \log T}{\delta}}. \]

In particular, an online-to-batch conversion of EWOO yields excess risk of order

\[ \frac{d \log T}{\eta T} + \frac{\sqrt{d \log T}}{T} \left( \sqrt{\frac{\log \log T}{\eta}} B + \frac{1}{\eta} + \frac{B}{\eta} \right) \log \frac{1}{\delta} + \frac{(\log \log T)B + 8 \log \frac{1}{\delta}}{T}. \]

By proceeding similarly one can get a guarantee for ONS, under the additional assumptions that \( \mathcal{F} \) has bounded diameter and that, for all \( z \in \mathcal{Z} \), the gradient \( \nabla_f \ell(f, z) \) has bounded norm.

**Obtaining \( o(\log T) \) excess risk.** Although the known worst-case regret bounds in this online setting involve a \( \log T \) factor, when the environment is stochastic and the distribution satisfies some notion of easiness the actual regret can be \( o(\log T) \). In such situations the excess risk similarly can be \( o(\log T) \) because our excess risk bounds depend not on worst-case regret bounds but rather the actual regret. Let us briefly explore one scenario where such improvement is possible. Suppose that the loss is also \( \beta \)-smooth; then, in situations when the cumulative loss of \( f^* \) is small, a more favorable regret bound is available using the analysis of Orabona et al. (2012, Theorem 1) for ONS: they show a regret bound of order \( \log (1 + \sum_{t=1}^T \ell_{f^*}(Z_t)) \). As a simple example, consider the case when the problem is realizable, in the sense that \( \ell_{f^*}(Z) = 0 \) almost surely. Then the regret bound becomes constant and the rate with respect to \( T \) for the excess risk in Corollary 2 becomes \( \frac{\log \log T}{T} \).
7 Discussion and Open Problems

We presented the first high probability $O(d/n)$ excess risk bound for exp-concave statistical learning. The key to proving this bound was the connection between exp-concavity and the central condition, a connection which suggests that exp-concavity implies a low noise condition. Here, low noise can be interpreted either in terms of the central condition, by the exponential decay of the negative tail of the excess loss random variables, or in terms of the Bernstein condition, by the variance of the excess loss of a hypothesis $f$ being controlled by its excess risk. All our results were based on this low noise interpretation of exp-concavity. In contrast, the previous in-expectation $O(d/n)$ results of Koren and Levy (2015) and Gonen and Shalev-Shwartz (2016) were based on the geometric/convexity-interpretation of exp-concavity, which we further boosted to high probability results using the low noise interpretation. It would be interesting to develop a high probability $O(d/n)$ result that proceeds purely from a low noise interpretation or purely from a geometric/convexity one.

Many results flowing from algorithmic stability often only yield in-expectation bounds, with high probability bounds stemming either from (i) a posthoc confidence boosting procedure — typically involving Hoeffding’s inequality, which “slows down” fast rate results; or (ii) quite strong stability notions — e.g. uniform stability allows one to apply McDiarmid’s inequality to a single run of the algorithm (Bousquet and Elisseeff, 2002). Is it a limitation of algorithmic stability techniques that high probability $O(d/n)$ fast rates seem to be out of reach without a posthoc confidence boosting procedure, or are we simply missing the right perspective? One reason to avoid a confidence boosting procedure is that the resulting bounds suffer from a multiplicative $\log(1/\delta)$ factor (on top of the standard additive $\log(1/\delta)$ factor in bounds like Theorem 1). As we mentioned earlier, we conjecture that the basic ERM method obtains a high probability $O(d/n)$ rate, and a potential path to show this rate would be to control a localized complexity as done by Sridharan et al. (2009) but using a more-involved argument based on exp-concavity rather than strong convexity.

An additional question of interest is if there exist exp-concave losses that are both non-smooth and that fail to fit into the framework of Gonen and Shalev-Shwartz (2016). If the answer is no, then the present work fully resolves the high probability analysis for bounded exp-concave losses (modulo the multiplicative effect of the $\log(1/\delta)$ term). If the answer is yes, and if there does exist an algorithm that can guarantee $O(1/n)$ expected excess risk for all (potentially non-smooth) bounded exp-concave losses, then we can boost the confidence of such a result as well.
References

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. In Advances in Neural Information Processing Systems, pages 41–48, 2008.

Olivier Bousquet and André Elisseeff. Stability and generalization. Journal of Machine Learning Research, 2:499–526, 2002.

Bernd Carl and Irntraud Stephani. Entropy, compactness, and the approximation of operators, volume 98. Cambridge University Press, 1990.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. In Advances in Neural Information Processing Systems, pages 359–366, 2001.

David A. Freedman. On tail probabilities for martingales. The Annals of Probability, 3(1):100–118, 1975.

Alon Gonen and Shai Shalev-Shwartz. Tightening the sample complexity of empirical risk minimization via preconditioned stability. arXiv preprint arXiv:1601.04011, 2016.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. Machine Learning, 69(2-3):169–192, 2007.

Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In Advances in Neural Information Processing Systems, pages 801–808, 2009.

Michael J. Kearns and Umesh Vazirani. An introduction to computational learning theory. MIT press, 1994.

Jyrki Kivinen and Manfred K Warmuth. Averaging expert predictions. In Computational Learning Theory, pages 153–167. Springer, 1999.

Tomer Koren. Open problem: Fast stochastic exp-concave optimization. In Conference on Learning Theory, pages 1073–1075, 2013.

Tomer Koren and Kfir Levy. Fast rates for exp-concave empirical risk minimization. In Advances in Neural Information Processing Systems, pages 1477–1485, 2015.

Mehrdad Mahdavi and Rong Jin. Excess risk bounds for exponentially concave losses. arXiv preprint arXiv:1401.4566, 2014.

Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In Proceedings of The 28th Conference on Learning Theory, pages 1305–1320, 2015.

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

Francesco Orabona, Nicolo Cesa-Bianchi, and Claudio Gentile. Beyond logarithmic bounds in online learning. In International Conference on Artificial Intelligence and Statistics, pages 823–831, 2012.

Robert E. Schapire. The strength of weak learnability. Machine learning, 5(2):197–227, 1990.

Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In Advances in Neural Information Processing Systems, pages 1545–1552, 2009.

Tim van Erven, Peter Grünwald, Mark D. Reid, and Robert C. Williamson. Mixability in statistical learning. In Advances 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.
Mathukumalli Vidyasagar. *Learning and Generalization with Applications to Neural Networks*. Springer, 2002.

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

### A Proofs

**Proof of Lemma 1** The exp-concavity of \( f \mapsto \ell(f, z) \) for each \( z \in Z \) implies that, for all \( z \in Z \) and all distributions \( Q \) over \( \mathcal{F} \):

\[
E_{f \sim Q} \left[ e^{-\eta \ell(f, z)} \right] \leq e^{-\eta \ell(f, z)} \iff \ell(E_{f \sim Q}[f], z) \leq -\frac{1}{\eta} \log E_{f \sim Q} \left[ e^{-\eta \ell(f, z)} \right].
\]

It therefore holds that for all distributions \( P \) over \( Z \), for all distributions \( Q \) over \( \mathcal{F} \), there exists (from convexity of \( \mathcal{F} \)) \( f^* = E_{f \sim Q}[f] \in \mathcal{F} \) satisfying

\[
E_{Z \sim P}[\ell(f^*, Z)] \leq E_{Z \sim P} \left[ -\frac{1}{\eta} \log E_{f \sim Q} \left[ e^{-\eta \ell(f, Z)} \right] \right].
\]

This condition is equivalent to stochastic mixability as well as the pseudoprobability convexity (PPC) condition, both defined by Van Erven et al. (2015). To be precise, for stochastic mixability, in Definition 4.1 of Van Erven et al. (2015), take their \( \mathcal{F}_d \) and \( \mathcal{F} \) both equal to our \( \mathcal{F} \), their \( \mathcal{P} \) equal to \( \{P\} \), and \( \psi(f) = f^* \); then strong stochastic mixability holds. Likewise, for the PPC condition, in Definition 3.2 of Van Erven et al. (2015) take the same settings but instead \( \phi(f) = f^* \); then the strong PPC condition holds. Now, Theorem 3.10 of Van Erven et al. (2015) states that the PPC condition implies the (strong) central condition.

**Proof of Theorem 1** First, from Lemma 1, the convexity of \( \mathcal{F} \) together with \( \eta \)-exp-concavity implies that \((P, \ell, \mathcal{F})\) satisfies the \( \eta \)-central condition.

The remainder of the proof is a drastic simplification of the proof of Theorem 7 of Mehta and Williamson (2014). Technically, Theorem 7 of that works applies directly, but one can get substantially smaller constants by avoiding much of the technical machinery needed there to handle VC-type classes (e.g. symmetrization, chaining, Talagrand’s inequality).

Denote by \( \mathcal{L}_f := \ell_f - \ell_{f^*} \) the excess loss with respect to comparator \( f^* \). Our goal is to show that, with high probability, ERM does not select any function \( f \in \mathcal{F} \) whose excess risk \( E[\mathcal{L}_f] \) is larger than \( \frac{a}{n} \) for some constant \( a \). Clearly, with probability 1 ERM will never select any function for which both \( \mathcal{L}_f \geq 0 \) almost surely and with some positive probability \( \mathcal{L}_f > 0 \); we call these functions the empirically inadmissible functions. For any \( \gamma_n > 0 \), let \( \mathcal{F}_{\geq \gamma_n} \) be the subclass formed by starting with \( \mathcal{F} \), retaining only functions whose excess risk is at least \( \gamma_n \), and further removing the empirically inadmissible functions.

Our goal now may be expressed equivalently as showing that, with high probability, ERM does not select any function \( f \in \mathcal{F}_{\geq \gamma_n} \) where \( \gamma_n = \frac{a}{n} \) and \( a > 1 \) is some constant to be determined later. Let \( \mathcal{F}_{\geq \gamma_n, \varepsilon} \) be an optimal proper \((\varepsilon/L)\)-cover for \( \mathcal{F}_{\geq \gamma_n} \) in the \( \ell_2 \) norm. From the Lipschitz property of the loss it follows that this cover induces an \( \varepsilon \)-cover in sup norm over the loss-composed function class \( \{\ell_f : f \in \mathcal{F}_{\geq \gamma_n}\} \). Observe that an \( \varepsilon \)-cover of \( \mathcal{F}_{\geq \gamma_n} \) in the \( \ell_2 \) norm has cardinality at most \((4R/\varepsilon)^d\) (Carl and Stephani, 1990, equation 1.1.10), and the cardinality of an optimal proper \( \varepsilon \)-cover is at most the cardinality of an optimal \((\varepsilon/2)\)-cover. (Vidyasagar, 2002, Lemma 2.1). It hence follows that \( \|\mathcal{F}_{\geq \gamma_n, \varepsilon}\| \leq (\frac{4R}{2\varepsilon})^d \).

Let us consider some fixed \( f \in \mathcal{F}_{\geq \gamma_n, \varepsilon} \). Since we removed the empirical inadmissible functions, there exists some \( \eta_f \geq \eta \) for which \( E[e^{-\eta_f \mathcal{L}_f}] = 1 \). Theorem 3 and Lemma 4, both from Mehta and Williamson (2014), imply that

\[
\log E_{Z \sim P} \left[ e^{-(\eta_f/2)\mathcal{L}_f} \right] \leq -\frac{0.18\eta_f a}{(B\eta_f \vee 1)n}.
\]
Applying Theorem 1 of Mehta and Williamson (2014) with $t = \frac{a}{2n}$ and the $\eta$ in that theorem set to $\eta_f/2$ yields:

$$\Pr \left( \frac{1}{n} \sum_{j=1}^{n} L_f(Z_j) \leq \frac{a}{2n} \right) \leq \exp \left( -0.18 \frac{\eta_f}{B \eta_f \vee 1} a + \frac{a \eta_f}{4n} \right).$$

Taking a union bound over $F_{\geq \gamma_n, \varepsilon}$ and using $\eta \leq \eta_f$ for all $f \in F_{\geq \gamma_n, \varepsilon}$, we have that

$$\Pr \left( \exists f \in F_{\geq \gamma_n, \varepsilon} : \frac{1}{n} \sum_{j=1}^{n} L_f(Z_j) \leq \frac{a}{2n} \right) \leq \left( 8 LR \varepsilon \right)^d \exp \left( -0.18 \frac{\eta_f}{B \eta_f \vee 1} a + \frac{a \eta}{4n} \right).$$

Setting $\varepsilon = \frac{1}{2n}$ and taking $n \geq 5$, from inversion it follows that with probability at least $1 - \delta$, for all $f \in F_{\geq \gamma_n, \varepsilon}$, we have

$$\frac{1}{n} \sum_{j=1}^{n} L_f(Z_j) \leq \frac{a}{2n},$$

where

$$a = 8 \left( B \vee \frac{1}{\eta} \right) \left( d \log(16 LR n) + \log \frac{1}{\delta} \right).$$

Now, since $\sup_{f \in F_{\geq \gamma_n}} \min_{f_r \in F_{\geq \gamma_n, \varepsilon}} \| f - f_r \|_\infty \leq \frac{1}{2n}$, and increasing $a$ by 1 to guarantee that $a > 1$, with probability at least $1 - \delta$, for all $f \in F_{\geq \gamma_n}$, we have $\frac{1}{n} \sum_{j=1}^{n} L_f(Z_j) > 0$. □

**Proof (of Lemma 2)** The main tool we use is part 2 of Theorem 5.4 of Van Erven et al. (2015). First, as per the proof of Lemma 1, note that the central condition as defined in the present work is equivalent to the strong PPC condition of Van Erven et al. (2015). We actually can improve that result due to our easier setting because we may take their function $v$ to be the constant function identically equal to $\eta$. Consequently, in equation (70) of Van Erven et al. (2015), we may take $\varepsilon = 0$, improving their constant $c_2$ by a factor of 3; moreover, their result actually holds for the second moment, not just the variance, yielding:

$$E[X^2] \leq \frac{2}{\eta \kappa(-2\eta B)} E[X], \quad (6)$$

where $\kappa(x) = \frac{e^x - x - 1}{x^2}$.

We now study the function

$$x \mapsto \frac{1}{\kappa(-x)} = \frac{x^2}{e^{-x} + x - 1}.$$

We claim that for all $x \geq 0$:

$$\frac{x^2}{e^{-x} + x - 1} \leq 2 + x.$$  

L'Hôpital’s rule implies that the inequality holds for $x = 0$, and so it remains to consider the case of $x > 0$.

First, observe that the denominator is nonnegative, and so we may rewrite this inequality as

$$x^2 \leq (2 + x)(e^{-x} + x - 1),$$

which simplifies to

$$0 \leq 2e^{-x} + x + xe^{-x} - 2 \quad \iff \quad 2(1 - e^{-x}) \leq x(1 + e^{-x}).$$

Therefore, we just need to show that, for all $x > 0$,

$$\frac{2}{x} \leq \frac{1 + e^{-x}}{1 - e^{-x}} = \frac{e^{x/2} + e^{-x/2}}{e^{x/2} - e^{-x/2}} = \text{coth}(x/2),$$

where $\text{coth}(x/2)$ is the hyperbolic cotangent function.
which is equivalent to showing that for all \( x > 0 \),

\[
\tanh(x) \leq x.
\]

But this indeed holds, since

\[
\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{2(x + \frac{x^3}{3!} + \frac{x^5}{5!} + \ldots)}{2(1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots)} = x \cdot \frac{1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots}{1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots} \leq x.
\]

The desired inequality is now established.

Returning to (6), we have

\[
E[X^2] \leq \frac{2}{\eta} (2 + 2\eta B) E[X] \leq 4 \left( \frac{1}{\eta} + B \right) E[X].
\]

**Proof (of Lemma 3)** The following simple version of Bernstein’s inequality will suffice for our analysis. Let \( X_1, \ldots, X_n \) be independent random variables satisfying \( X_j \geq B \) almost surely. Then

\[
\Pr \left( \frac{1}{n} \sum_{j=1}^{n} X_j - E[X] \geq t \right) \leq \exp \left( -\frac{nt^2}{2 \left( E \left[ \frac{1}{n} \sum_{j=1}^{n} X_j^2 \right] + \frac{Bt}{3} \right)} \right).
\]

Denote by \( \mathcal{L}_f := \ell_f - \ell_{f_1} \) the excess loss with respect to comparator \( f_1 \). Fix some \( f \in \mathcal{G}' \setminus \{f_1\} \), take \( X = -\mathcal{L}_f \), and set \( t = E[\mathcal{L}_f] \), yielding:

\[
\Pr \left( \frac{1}{n} \sum_{j=1}^{n} \mathcal{L}_f(Z_j) \leq 0 \right) \leq \exp \left( -\frac{n E[\mathcal{L}_f]^2}{2 \left( E \left[ \mathcal{L}_f^2 \right] + \frac{1}{2} B E[\mathcal{L}_f] \right)} \right)
\]

\[
\leq \exp \left( -\frac{n E[\mathcal{L}_f]^2}{2 \left( C E[\mathcal{L}_f]^q + \frac{1}{2} B E[\mathcal{L}_f] \right)} \right)
\]

\[
= \exp \left( -\frac{n E[\mathcal{L}_f]^{2-q}}{2 \left( C + \frac{1}{2} B E[\mathcal{L}_f]^{1-q} \right)} \right)
\]

\[
\leq \exp \left( -\frac{n E[\mathcal{L}_f]^{2-q}}{2 \left( C + \frac{1}{2} B^{2-q} \right)} \right).
\]

Therefore, if

\[
E[\mathcal{L}_f] \geq \left( \frac{2 \left( C + \frac{B^{2-q}}{3} \right) \log \frac{|\mathcal{G}'|}{\delta}}{n} \right)^{1/(2-q)}
\]

then it holds with probability at least \( 1 - \frac{8 \delta}{|\mathcal{G}'| - 1} \) that \( \frac{1}{n} \sum_{j=1}^{n} \mathcal{L}_f(Z_j) > 0 \). The result follows by taking a union bound over the subclass of \( \mathcal{G}' \setminus \{f_1\} \) for which (7) holds.  

\[\blacksquare\]