Fast rates with high probability in exp-concave statistical learning

Nishant A. Mehta
Centrum Wiskunde & Informatica
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 $O(d \log(1/\delta)/n)$ with probability $1-\delta$. 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. We also 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. Lastly, we present high probability bounds for the exp-concave model selection aggregation problem that are quantile-adaptive in a certain sense. One bound obtains a nearly optimal rate without requiring the loss to be Lipschitz continuous, and another requires Lipschitz continuity but obtains the optimal rate.

1 Introduction

In the statistical learning problem, a learning agent observes a sample of n points $Z_1, \ldots, Z_n$ drawn i.i.d. from an unknown distribution P over an outcome space $\mathcal{Z}$. The agent then seeks an action f in an action space $\mathcal{F}$ that minimizes their expected loss, or risk, $\mathsf{E}_{Z \sim P}[\ell(f, Z)]$, where ℓ is a loss function $\ell \colon \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$. Several recent works have studied this problem in the situation where the loss is exp-concave and bounded, $\mathcal{F}$ and $\mathcal{Z}$ are subsets of $\mathbb{R}^d$, and $\mathcal{F}$ 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 $d(\log n + \log(1/\delta))/n$. Via new algorithmic stability arguments applied to empirical risk minimization (ERM), Koren and Levy (2015) and Gonen and Shalev-Shwartz (2016) discarded 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 η-exp-concave losses, the progressive mixture rule (a Cesàro mean of pseudo-Bayesian estimators) with learning rate η obtains expected excess risk $O(1/n)$ but, for *any* learning rate, these rules suffer from severe deviations of order $\sqrt{\log(1/\delta)/n}$.

This work resolves the high probability question: we present a learning algorithm with an excess risk bound (Corollary 1) which has rate $d \log(1/\delta)/n$ with probability at least $1 - \delta$. ERM also obtains $O((d \log(n) + \log(1/\delta))/n)$ excess risk, a fact that apparently was not widely known although it follows from results in the literature. To vanquish the $\log n$ factor with the small $\log(1/\delta)$ 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 boosting the confidence trick of Schapire (1990).

In the next section, we discuss a brief history of the work in this area. In Section 3, 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 4 and 5 respectively. Section 6 extends 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

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

the variance control implied by exp-concavity, and this control can be 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. One powerful consequence of this result is a new guarantee for model selection aggregation: we present a method (Section 7) for the model selection aggregation problem over finite classes with exp-concave losses that obtains a rate of $O((\log |\mathcal{F}| + \log n)/n)$ with high probability, with no dependence on the Lipschitz continuity of the loss function. All previous bounds of which we are aware depend on the Lipschitz continuity of the problem. Moreover, the bound is a quantile-like bound in that it improves with the prior measure on a subclass of nearly optimal hypotheses.

2 A history of exp-concave learning

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 η-mixable (which is implied by η-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.

Audibert (2008) showed that when learning over a finite class with exp-concave losses, no progressive mixture rule can obtain a high probability excess risk bound of order better than $\sqrt{\log(1/\delta)/n}$. ERM fares even worse, with a lower bound of $\sqrt{\log |\mathcal{F}|/n}$ *in expectation.* (Juditsky et al., 2008). Audibert (2008) overcame the deviations shortcoming of progressive mixture rules via his *empirical star* algorithm, which first runs ERM on $\mathcal{F}$, obtaining $\hat{f}_{\text{ERM}}$, and then runs ERM a second time on the star convex hull of $\mathcal{F}$ with respect to $\hat{f}_{\text{ERM}}$. This algorithm achieves $O(\log |\mathcal{F}|/n)$ with high probability; the rate was only proved for squared loss with targets Y and predictions $\hat{y}$ in $[-1, 1]$, but it was claimed that the result can be extended to general, bounded losses $\hat{y} \mapsto \ell(y, \hat{y})$ satisfying smoothness and strong convexity as a function of predictions $\hat{y}$. Under similar assumptions, Lecué and Rigollet (2014) proved that a method, Q-aggregation, also obtains this rate but can further take into account a prior distribution.

For convex classes, such as $\mathcal{F} \subset \mathbb{R}^d$ as we consider here, Hazan et al. (2007) designed the Online Newton Step (ONS) and Exponentially Weighted Online Optimization (EWOO) algorithms. Both have $O(d \log n)$ regret over n rounds, which, after online-to-batch conversion yields $O(d \log(n)/n)$ excess risk in expectation. Mahdavi et al. (2015) showed that an online-to-batch conversion of ONS enjoys 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 cost of $O(d^3)$ even in simple cases like learning over the unit ℓ_2 ball, and in general its complexity may be as high as $O(d^4)$ per projection step (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 discuss in the next section 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. Table 1 summarizes what is known and our new results.

3 Rate-optimal in-expectation bounds

We now describe the setting more formally. In this work $\mathcal{F}$ is always assumed to be convex, except in Section 7, which studies the model selection aggregation problem for countable classes. We say a function $A \colon \mathcal{F} \to \mathbb{R}$ has diameter C if $\sup_{f_1, f_2 \in \mathcal{F}} |A(f_1) - A(f_2)| \leq C$. Assume for each $z \in \mathcal{Z}$ that the loss map $\ell(\cdot, z) \colon f \mapsto \ell(f, z)$ is η-exp-concave, i.e. $f \mapsto e^{-\eta \ell(f,z)}$ is concave over $\mathcal{F}$. 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 $\mathcal{Z}$, our objective is to select a hypothesis $f \in \mathcal{F}$ that minimizes the excess risk $\mathsf{E}_{Z \sim P}[\ell_f(Z)] - \inf_{f \in \mathcal{F}} \mathsf{E}_{Z \sim P}[\ell_f(Z)]$. We assume that there exists $f^* \in \mathcal{F}$ satisfying $\mathsf{E}[\ell_{f^*}(Z)] = \inf_{f \in \mathcal{F}} \mathsf{E}_{Z \sim P}[\ell_f(Z)]$; this assumption also was made by Gonen and Shalev-Shwartz (2016) and Kakade and Tewari (2009).[1]

Let $\mathcal{A}_{\mathcal{F}}$ be an algorithm, defined for a function class $\mathcal{F}$ as a mapping $\mathcal{A}_{\mathcal{F}} \colon \bigcup_{n \geq 0} \mathcal{Z}^n \to \mathcal{F}$; we drop the subscript $\mathcal{F}$ when it is clear from the context. Our start-

[1] This assumption is not explicit from Koren and Levy (2015), but their other assumptions might imply it. Regardless, if their results and those of Gonen and Shalev-Shwartz (2016) hold, our analysis in Section 5 can be adapted to work if the infimal risk is not achieved.

	Convex $\mathcal{F}$		Finite $\mathcal{F}$	
Algorithm	Expectation	Probability $1-\delta$	Expectation	Probability $1-\delta$
Progressive mixture	—	—	$\log\lvert\mathcal{F}\rvert/n$	$\Omega(\sqrt{\log(1/\delta)/n})$
Empirical star / Q-agg.	—	—	$\log\lvert\mathcal{F}\rvert/n$	$(\log\lvert\mathcal{F}\rvert+\log(1/\delta))/n$
Online Newton Step	$d\log n/n$	$d(\log n+\log(1/\delta))/n$	—	—
EWOO	$d\log n/n$	$\boldsymbol{(d\log n+\log(1/\delta))/n}$	—	—
ERM	d/n	$\boldsymbol{(d\log n+\log(1/\delta))/n}$	$\Omega(\sqrt{\log\lvert\mathcal{F}\rvert/n})$	—
Boosted ERM	—	$\boldsymbol{d\log(1/\delta)/n}$	—	—

Table 1: Excess risk bounds (new results in bold). Upper bounds are big-O. Boosted ERM applies ConfidenceBoost to ERM. "ERM" is either penalized ERM (Koren and Levy, 2015) or ERM (Gonen and Shalev-Shwartz, 2016). For simplicity we only show dependence in d, n, and δ and restrict Q-aggregation to uniform prior.

ing point will be an algorithm $\mathcal{A}$ which, when provided with a sample $\mathbf{Z}$ of n i.i.d. points, satisfies an expected risk bound of the form

$$\mathsf{E}_{\mathbf{Z}\sim P^n}\left[\mathsf{E}_{Z\sim P}\left[\ell_{\mathcal{A}(\mathbf{Z})}(Z)-\ell_{f^*}(Z)\right]\right]\leq\psi(n). \qquad (1)$$

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 $\mathcal{F}\subset\mathbb{R}^d$, each in a slightly different setting. Koren and Levy (2015) assume, for each outcome $z\in\mathcal{Z}$, that the loss $\ell(\cdot,z)$ has diameter B and is β-smooth for some $\beta\geq 1$, i.e. for all $f,f'\in\mathcal{F}$, the gradient is β-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 $\Gamma\colon\mathcal{F}\to\mathbb{R}$ 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

$$\psi(n)=\tfrac{1}{n}\left(\tfrac{24\beta d}{\eta}+100Bd+R\right).$$

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. They assume that the loss is of the form $\ell_f(z)=\phi_y(\langle f,x\rangle)$, for $\mathcal{F}\subset\mathbb{R}^d$. They further assume, for each $z=(x,y)$, that the mapping $\hat{y}\mapsto\phi_y(\hat{y})$ is α-strongly convex and L-Lipschitz, but they do not assume smoothness. They show that standard, unregularized ERM has expected excess risk at most

$$\psi(n)=\tfrac{2L^2d}{\alpha n}=\tfrac{2d}{\eta n},$$

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

4 A high probability bound for ERM

As a warm-up to proving a high probability $O(d/n)$ excess risk bound, we first show that ERM itself obtains excess risk $O(d\log(n)/n)$ with high probability; here and elsewhere, if δ is omitted the dependence is $\log(1/\delta)$. That ERM satisfies such a bound was largely implicit in the literature, and so we make this result explicit. The closest such result, Theorem 1 of Mahdavi and Jin (2014), does not apply as it relies on an additional assumption (see their Assumption (I)). 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 η-exp-concave losses the constant η 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 η-central condition, which the excess loss satisfies. This concept, 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 η-*central condition *for some $\eta>0$ if there exists a comparator $f^*\in\mathcal{F}$ such that, for all $f\in\mathcal{F}$, $\mathsf{E}_{Z\sim P}\left[e^{-\eta(\ell_f(Z)-\ell_{f^*}(Z))}\right]\leq 1$.*

Jensen's inequality implies that if this condition holds, the corresponding f^* must be a risk minimizer. It is known (Van Erven et al., 2015, Section 4.2.2) that in our setting $(P,\ell,\mathcal{F})$ satisfies the η-central condition.

Lemma 1 *Let $\mathcal{F}$ be convex. Take ℓ to be a loss function $\ell\colon\mathcal{F}\times\mathcal{Z}\to\mathbb{R}$, and assume that, for each $z\in\mathcal{Z}$, the map $\ell(\cdot,z)\colon f\mapsto\ell(f,z)$ is η-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 η-central condition.*

Under the central condition, Theorem 7 of Mehta and Williamson (2014) directly implies an $O(d \log(n)/n)$ bound for ERM; however, a far simpler version of that result yields much smaller constants. The proof of the version below, in the appendix for completeness, only makes use of an (ε/L)-net of $\mathcal{F}$ in the ℓ_2 norm, which induces an ε-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 η-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*

$$\begin{aligned} &\mathsf{E}_{Z \sim P}[\ell_{\hat{f}}(Z) - \ell_{f^*}(Z)] && (2) \\ &\leq \tfrac{1}{n} \left(8 \left(B \vee \tfrac{1}{\eta} \right) \left(d \log(16LRn) + \log \tfrac{1}{\delta} \right) + 1 \right). \end{aligned}$$

5 Boosting the confidence for high probability bounds

The two existing excess risk bounds mentioned in Section 3 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 $\log(1/\delta)$ factor. 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 $\mathcal{A}$ 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.*

Definition 2 (Bernstein condition) *We say that $(P, \ell, \mathcal{F})$ 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 $f \in \mathcal{F}$,*

$$\mathsf{E}_{Z \sim P}\Big[\left(\ell_f(Z) - \ell_{f^*}(Z) \right)^2 \Big] \leq C \, \mathsf{E}_{Z \sim P} \left[\ell_f(Z) - \ell_{f^*}(Z) \right]^q.$$

Algorithm 1: ConfidenceBoost

```
Input: $\mathbf{Z}_1, \ldots, \mathbf{Z}_K \overset{iid}{\sim} P^{n_{\mathrm{I}}}$, $\mathbf{Z}_{\mathrm{II}} \sim P^{n_{\mathrm{II}}}$, learner $\mathcal{A}_{\mathcal{F}}$
for $j = 1 \to K$ do $\hat{f}_j = \mathcal{A}_{\mathcal{F}}(\mathbf{Z}_j)$
return $\mathrm{ERM}_{\mathcal{F}_K}(\mathbf{Z}_{\mathrm{II}})$, with $\mathcal{F}_K = \{\hat{f}_1, \ldots, \hat{f}_K\}$
```

Before getting to ConfidenceBoost, we first show that the exp-concave learning setting satisfies the Bernstein condition with the best exponent, $q = 1$, and so is a special case of the more general setting we analyze. Recall from Lemma 1 that the η-central condition holds for $(P, \ell, \mathcal{F})$. The next lemma, which adapts a result of Van Erven et al. (2015), shows that the η-central condition, together with 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 $\mathsf{E}[e^{-\eta X}] \leq 1$. Then $\mathsf{E}[X^2] \leq 4 \left(1/\eta + B \right) \mathsf{E}[X]$.*

Boosting the "boosting the confidence" trick. First, consider running $\mathcal{A}$ on a sample $\mathbf{Z}_1$ of n i.i.d. points. The excess risk random variable $\mathsf{E}_Z[\ell_{\mathcal{A}(\mathbf{Z}_1)}(Z) - \ell_{f^*}(Z)]$ is nonnegative, and so Markov's inequality and the expected excess risk being bounded by $\psi(n)$ imply that

$$\Pr\left(\mathsf{E}_Z[\ell_{\mathcal{A}(\mathbf{Z}_1)}(Z) - \ell_{f^*}(Z)] \geq e \cdot \psi(n) \right) \leq \tfrac{1}{e}.$$

Now, let $\mathbf{Z}_1, \ldots, \mathbf{Z}_K$ be independent samples, each of size n. Running $\mathcal{A}$ on each sample yields $\hat{f}_1 := \mathcal{A}(\mathbf{Z}_1), \ldots, \hat{f}_K := \mathcal{A}(\mathbf{Z}_K)$. 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 $\mathsf{E}_{Z \sim P}\big[\ell_{\hat{f}_j}(Z) - \ell_{f^*}(Z)\big] \leq e \cdot \psi(n)$. 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 $\mathsf{E}_{Z \sim P}\big[\ell_{\hat{f}_j}(Z) - \ell_{f^*}(Z)\big] \leq e \cdot \psi(n)$, 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.

Analysis. From here on out, we treat the initial sample of size Kn as fixed and unhat the K estimators above, referring to them as $f_1, \ldots, f_K$. Without loss of generality, 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 $\mathcal{F}_K := \{f_1, \ldots, f_K\}$ 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 $\mathcal{F}_K$ with high probability, which will in turn imply that ERM obtains $O(1/n^{1/(2-q)})$ excess risk over $\mathcal{F}_K$.

We first establish an *approximate* Bernstein condition for $(P, \ell, \mathcal{F}_K)$. Since $\|\ell_{f_j} - \ell_{f_1}\|_{L_2(P)} \leq \|\ell_{f_j} - \ell_{f^*}\|_{L_2(P)} + \|\ell_{f_1} - \ell_{f^*}\|_{L_2(P)}$ for all $f_j \in \mathcal{F}_K$, from the (C, q)-Bernstein condition, $\|\ell_{f_j} - \ell_{f_1}\|^2_{L_2(P)}$ is at most

$$\begin{aligned}
&\leq C \left(\mathsf{E}[\ell_{f_j} - \ell_{f^*}]^q + \mathsf{E}[\ell_{f_1} - \ell_{f^*}]^q \right) \\
&\quad + 2C \left(\mathsf{E}[\ell_{f_j} - \ell_{f^*}] \cdot \mathsf{E}[\ell_{f_1} - \ell_{f^*}] \right)^{q/2} \\
&\leq C \big(3 \, \mathsf{E}[\ell_{f_j} - \ell_{f^*}]^q + \mathsf{E}[\ell_{f_1} - \ell_{f^*}]^q \big) \\
&= C \big(3 \left(\mathsf{E}[\ell_{f_j} - \ell_{f_1}] + \mathsf{E}[\ell_{f_1} - \ell_{f^*}] \right)^q + \mathsf{E}[\ell_{f_1} - \ell_{f^*}]^q \big) \\
&\leq C \big(3 \, \mathsf{E}[\ell_{f_j} - \ell_{f_1}]^q + 4 \, \mathsf{E}[\ell_{f_1} - \ell_{f^*}]^q \big);
\end{aligned}$$

the last step uses the subadditivity of the concave map $x \mapsto x^q$. We call this bound an approximate Bernstein condition because, on event GOOD, for all $f_j \in \mathcal{F}_K$:

$$\|\ell_{f_j} - \ell_{f_1}\|^2_{L_2(P)} \leq C \left(3 \, \mathsf{E}[\ell_{f_j} - \ell_{f_1}]^q + 4 (e \cdot \psi(n))^q \right).$$

Define the class $\mathcal{F}'_K$ as the set $\{f_1\} \cup \left\{ f_j \in \mathcal{F}_K : \mathsf{E}[\ell_{f_j} - \ell_{f_1}] \geq 4^{1/q} e \cdot \psi(n) \right\}$. 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 $\{g_1, \ldots, g_K\}$ and assume without loss of generality that g_1 is a risk minimizer over $\mathcal{G}$. Let $\mathcal{G}' \subset \mathcal{G}$ be a subclass for which, for all $f \in \mathcal{G}'$:*

$$\mathsf{E}[(\ell_f - \ell_{g_1})^2] \leq C \, \mathsf{E}[\ell_f - \ell_{g_1}]^q,$$

and $\ell_f - \ell_{g_1} \leq B$ 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*

$$\mathsf{E}[\ell_f - \ell_{g_1}] \geq \left(\frac{2\left(C + \frac{B^{2-q}}{3}\right) \log \frac{|\mathcal{G}'| - 1}{\delta}}{n} \right)^{1/(2-q)} .$$

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

- $\mathsf{E}_{Z \sim P}[\ell_{f_j}(Z) - \ell_{f_1}(Z)] \leq 4^{1/q} e \cdot \psi(n)$;
- $\mathsf{E}_{Z \sim P}[\ell_{f_j}(Z) - \ell_{f_1}(Z)] \leq \left(\frac{2\left(C + \frac{B^{2-q}}{3}\right) \log \frac{K}{\delta}}{n} \right)^{1/(2-q)}$

We now run ConfidenceBoost with $K = \lceil \log(2/\delta) \rceil$ on a sample of n points, with $n_{\mathrm{I}} = \frac{n}{2K}$ and $n_{\mathrm{II}} = \frac{n}{2}$; for simplicity, we assume that $2K$ divides n. 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 $\mathcal{A}$ obtains a bound of the form (1). Then, with probability at least $1 - \delta$,* ConfidenceBoost *run with $K = \lceil \log(2/\delta) \rceil$, $n_{\mathrm{I}} = n/(2K)$, and $n_{\mathrm{II}} = n/2$ learns a hypothesis $\hat{f}$ with excess risk $\mathsf{E}_{Z \sim P}[\ell_{\hat{f}}(Z) - \ell_{f^*}(Z)]$ at most*

$$\begin{aligned}
& e \cdot \psi \left(\frac{n}{2 \lceil \log \frac{2}{\delta} \rceil} \right) \qquad (3) \\
& + \max \left\{ \begin{array}{c} 4^{1/q} e \cdot \psi \left(\frac{n}{2 \log \lceil \frac{2}{\delta} \rceil} \right), \\ \left(\frac{4\left(C + \frac{B^{2-q}}{3}\right)\left(\log \frac{1}{\delta} + \log \lceil \log \frac{2}{\delta} \rceil\right)}{n} \right)^{1/(2-q)} \end{array} \right\}.
\end{aligned}$$

The next result for exp-concave learning is immediate.

Corollary 1 *Applying Theorem 2 with $\mathcal{A}_\mathcal{F}$ the algorithm of Koren and Levy (2015) and their assumptions (with $\beta \geq 1$), the bound in Theorem 2 specializes to*

$$O\left(\frac{\log \frac{1}{\delta}}{n} \left(\frac{d\beta}{\eta} + dB + R \right) \right). \qquad (4)$$

Similarly taking $\mathcal{A}_\mathcal{F}$ the algorithm of Gonen and Shalev-Shwartz (2016) and their assumptions yields

$$O\left(\frac{\log \frac{1}{\delta}}{n} \left(\frac{d}{\eta} + B \right) \right). \qquad (5)$$

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 (which is strictly stronger than exp-concavity), 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).

6 Online-to-batch-conversion

This section shows 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 n rounds) is logarithmic in n. 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 ERM in Section 4.

Mahdavi et al. (2015) considered an online-to-batch conversion of ONS and established the first high probability $O(\log n/n)$ excess risk bound in the exp-concave statistical learning setting. Their analysis is elegant but seems intimately coupled to ONS; it is thus unclear if their analysis can yield excess risk bounds for online-to-batch conversions of other online exp-concave learners. This leads 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 strongly on the analysis of Kakade and Tewari (2009) in the strongly convex setting.

We first consider a different, related setting: online convex optimization (OCO) under a B-bounded, ν-strongly convex loss that is L-Lipschitz with respect to the action. An OCO game unfolds over n rounds. An adversary first selects a sequence of n convex loss functions $c_1, \ldots, c_n$. In round t, the online learner plays $f_t \in \mathcal{F}$, the environment subsequently reveals cost function c_t, and the learner suffers loss $c_t(f_t)$. Because we are interested in analyzing the statistical learning setting, we constrain the adversary to play a sequence of n points $z_1, \ldots, z_n \in \mathcal{Z}$, inducing cost functions $\ell(\cdot, z_1), \ldots, \ell(\cdot, z_n)$.

Consider an online learner that sequentially plays actions $f_1, \ldots, f_n \in \mathcal{F}$ in response to $z_1, \ldots, z_n$, so that f_t depends on $(z_1, \ldots, z_{t-1})$. The (cumulative) regret is defined as $\sum_{t=1}^n \ell_{f_t}(z_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \ell_f(z_t)$. When the losses are bounded, strongly convex, and Lipschitz, Kakade and Tewari (2009) showed that if an online algorithm has regret $\mathcal{R}_n$ on an i.i.d. sequence $Z_1, \ldots, Z_n \sim P$, online-to-batch conversion by simple averaging $\bar{f}_n := \frac{1}{n} \sum_{t=1}^n f_t$ has the following guarantee.

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

$$\begin{aligned} &\mathsf{E}_{Z \sim P}[\ell_{\bar{f}_n}(Z) - \ell_{f^*}(Z)] \\ &\leq \frac{\mathcal{R}_n}{n} + 4\sqrt{\frac{L^2 \log \frac{1}{\delta}}{\nu}} \frac{\sqrt{\mathcal{R}_n}}{n} + \max\left\{\frac{16L^2}{\nu}, 6B\right\} \frac{\log \frac{1}{\delta}}{n}. \end{aligned}$$

Under various assumptions, there are OCO algorithms that obtain worst-case regret (under all sequences $z_1, \ldots, z_n$) $\mathcal{R}_n = O(\log n)$; e.g., Online Gradient Descent (Hazan et al., 2007) satisfies $\mathcal{R}_n \leq \frac{G^2}{2\nu}(1 + \log n)$, 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 η-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 $\mathcal{R}_n \leq 5\left(\frac{1}{\eta} + GD\right) d \log n$, where G is a bound on the gradient and D is a bound on the diameter of the action space.
- Exponentially Weighted Online Optimization (EWOO), with $\mathcal{R}_n \leq \frac{1}{\eta} d\,(1 + \log(n+1))$. The better regret bound comes at the price of not being computationally efficient.

We now show how to extend the analysis of Kakade and Tewari (2009) to exp-concave losses. While similar results can be obtained from the work of Mahdavi et al. (2015) for the specific case of ONS, our analysis is agnostic of the base algorithm. As a consequence, our analysis applies to EWOO, which, although highly impractical, offers a better regret bound. The key insight is that exp-concavity implies a variance inequality similar to Lemma 1 of Kakade and Tewari (2009), a pivotal result of that work that unlocks Freedman's inequality for martingales (Freedman, 1975). Let Z_1^t denote the sequence $Z_1, \ldots, Z_t$.

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

$$\xi_t := \mathsf{E}_Z\left[\ell_{f_t}(Z) - \ell_{f^*}(Z)\right] - \left(\ell_{f_t}(Z_t) - \ell_{f^*}(Z_t)\right).$$

Then

$$\mathrm{Var}\left[\xi_t \mid Z_1^{t-1}\right] \leq 4\left(\frac{1}{\eta} + B\right) \mathsf{E}_Z\left[\ell_{f_t}(Z) - \ell_{f^*}(Z)\right].$$

Proof Observe that $\mathrm{Var}\left[\xi_t \mid Z_1^{t-1}\right] = \mathrm{Var}\big[\ell_{f_t}(Z_t) - \ell_{f^*}(Z_t) \mid Z_1^{t-1}\big]$. Treating the sequence Z_1^{t-1} as fixed and also treating f_t as a fixed parameter f, the above conditional variance equals $\mathrm{Var}\big[\ell_f(Z) - \ell_{f^*}(Z)\big]$, where only $Z \sim P$ is random. Then, Lemma 2 implies that $\mathrm{Var}\big[\ell_f(Z) - \ell_{f^*}(Z)\big] \leq 4\left(\frac{1}{\eta} + B\right) \mathsf{E}\left[\ell_f(Z) - \ell_{f^*}(Z)\right]$. ■

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 η-exp-concave with respect to the action $f \in \mathcal{F}$. Then with probability at least $1 - \delta$, for any $n \geq 3$, the excess risk of $\bar{f}_n$ is at most*

$$\begin{aligned} &\frac{\mathcal{R}_n}{n} + 4\sqrt{\left(\frac{1}{\eta} + B\right) \log \frac{4 \log n}{\delta}} \cdot \frac{\sqrt{\mathcal{R}_n}}{n} \\ &\quad + 16\left(\frac{1}{\eta} + B\right) \frac{\log \frac{4 \log n}{\delta}}{n}. \end{aligned}$$

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

$$\begin{aligned} &\frac{d \log n}{\eta n} + \frac{(\log \log n) B + B \log \frac{1}{\delta}}{n} \\ &\quad + \frac{\sqrt{d \log n}}{n} \left(\sqrt{\frac{(\log \log n) B}{\eta}} + \sqrt{\left(\frac{1}{\eta^2} + \frac{B}{\eta}\right) \log \frac{1}{\delta}}\right). \end{aligned}$$

Proceeding similarly yields a bound for ONS under the additional assumptions that $\mathcal{F}$ has bounded diameter and that $\nabla_f \ell(f, z)$ has bounded norm for all $z \in \mathcal{Z}$.

Obtaining $o(\log(n)/n)$ excess risk. The worst-case regret bounds in this online setting have a $\log n$ factor, but when the environment is stochastic and the distribution satisfies some notion of easiness the actual regret can be $o(\log n)$. In such situations the excess risk similarly can be $o(\log(n)/n)$ because our excess risk bounds depend not on worst-case regret bounds but rather the actual regret. We explore one scenario where such improvement is possible. Suppose that the loss is also β-smooth; then, when the cumulative loss of f^* is small, the analysis of Orabona et al. (2012, Theorem 1) for ONS yields an improved regret bound of order $\log\bigl(1 + \sum_{t=1}^n \ell_{f^*}(Z_t)\bigr)$. 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 is constant and the rate with respect to n for the excess risk in Corollary 2 is $\frac{\log \log n}{n}$.

7 Model selection aggregation

In the model selection aggregation problem for exp-concave losses, we are given a countable class $\mathcal{F}$ of functions from an input space $\mathcal{X}$ to an output space $\mathcal{Y}$ and a loss $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$; for each $y \in \mathcal{Y}$, the mapping $\hat{y} \mapsto \ell(y, \hat{y})$ is η-exp-concave. The loss is a supervised loss, as in supervised classification and regression, unlike the more general loss functions used in the rest of the paper. The random points $Z \sim P$ now decompose into an input-output pair $Z = (X, Y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We often use the notation $\ell_f(Z) := \ell(Y, f(X))$. The goal is the same as in the stochastic exp-concave optimization problem, but now $\mathcal{F}$ fails to be convex.

After Audibert (2008) showed that the progressive mixture rule cannot obtain fast rates with high probability, several works developed methods that departed from progressive mixture rules and gravitated instead toward ERM-style rules, starting with the empirical star algorithm of Audibert (2008) and a subsequent method of Lecué and Mendelson (2009) which runs ERM over the convex hull of a data-dependent subclass. Lecué and Rigollet (2014) extended these results to take into account a prior on the class using their Q-aggregation procedure. All the methods require Lipschitz continuity of the loss[2] and are for finite classes. In this section, we present an algorithm that carefully composes exponential weights-type algorithms and still obtains a fast rate with high probability for the model selection aggregation problem. One incarnation can do so with the fast rate of $O(\log |\mathcal{F}|/n)$ for finite $|\mathcal{F}|$, by relying on Boosted ERM. Another, "pure" version based on exponential weights-type procedures alone can get a rate of $O(\log |\mathcal{F}|/n + \log n/n)$ with no dependence on the Lipschitz continuity of the loss. To our knowledge, this is the first fast rate high probability bound for model selection aggregation that lacks dependence on the Lipschitz constant of the loss. Both results hold more generally, allowing for countable classes, taking into account a prior distribution π over $\mathcal{F}$, and providing a quantile-like improvement when there is a low quantile with close to optimal risk.

As $\mathcal{F}$ is countable and hence not convex, algorithms for stochastic exp-concave optimization do not apply. Our approach is to apply stochastic exp-concave optimization to the convex hull of a certain small cardinality, data-dependent subset of $\mathcal{F}$. The first phase obtains this subset via the progressive mixture rule. We offer two variants for the second phase: PM-EWOO (Algorithm 2) and PM-CB (Algorithm 3). In the algorithms, $\mathcal{A}^{\mathsf{pm}}$ and $\mathcal{A}^{\mathsf{ew}}$ are online-to-batch conversions of the progressive mixture rule and EWOO respectively, $\mathcal{A}^{\mathsf{cb}}$ is ConfidenceBoost, and $\mathcal{A}^{\mathsf{erm}}$ is ERM.

Our interest in PM-EWOO is two-fold: *(i)* it is a "purely" exponential weights type method in that it is based only on the progressive mixture rule and EWOO; *(ii)* it does not require any Lipschitz assumption on the loss function, unlike all previous work.

Theorem 4 *Let $\mathcal{F}$ be a countable and π a prior distribution over $\mathcal{F}$. Assume that for each y the loss $\ell \colon \hat{y} \mapsto \ell(y, \hat{y})$ is η-exp-concave. Further assume that $\sup_{f, f' \in \mathcal{F}} |\ell(y, f(x)) - \ell(y, f'(x))| \leq B$ for all (x, y) in the support of P. Then with probability at least $1 - \delta$,* PM-EWOO *run with $K = \lceil \log(2/\delta) \rceil$, $n_{\mathrm{I}} = n/(2K)$ and $n_{\mathrm{II}} = n/2$ learns a hypothesis $\hat{f}$ satisfying*

$$\begin{aligned} &\mathsf{E}_{Z \sim P}\left[\ell_{\hat{f}}(Z) - \ell_{f^*}(Z)\right] \\ &\leq e \cdot \textsc{BayesRed}_\eta\left(\frac{n}{2\lceil \log \frac{2}{\delta} \rceil}, \pi\right) + \theta_{\mathrm{EW}}(\delta, n), \end{aligned}$$

with $\theta_{\mathrm{EW}}(\delta, n) = O\left(\frac{\sqrt{B}\left(\log \frac{1}{\delta} + \sqrt{\log \frac{1}{\delta}} \log n\right)}{\eta n} + \frac{B \log \frac{\log n}{\delta})}{n}\right)$.

Here, $\textsc{BayesRed}_\eta\,(n, \pi)$ is the *η-generalized Expected Bayesian Redundancy* (Grünwald, 2012), defined as

$$\inf_{\rho \in \Delta(\mathcal{F})} \left\{\mathsf{E}_Z\left[\mathsf{E}_{f \sim \rho}\left[\ell_f(Z)\right] - \ell_{f^*}(Z)\right] + \frac{D(\rho \,\|\, \pi)}{\eta(n+1)}\right\},$$

for $D(\cdot \,\|\, \cdot)$ the KL-divergence. The bound can be rewritten as a quantile-like bound; for all $\rho \in \Delta(\mathcal{F})$:

$$\begin{aligned} &\mathsf{E}_{Z \sim P}\left[\ell_{\hat{f}}(Z) - \mathsf{E}_{f \sim \rho}\left[\ell_f(Z)\right]\right] \\ &\leq (e-1)\textsc{gap}(\rho, f^*) + \frac{2e\left\lceil \log \frac{2}{\delta} \right\rceil D(\rho \,\|\, \pi)}{\eta n} + \theta_{\mathrm{EW}}(\delta, n), \end{aligned}$$

[2] Audibert (2008) analyzed bounded squared loss, with a suggestion for analyzing exp-concave losses; from the techniques used, Lipschitz continuity likely would be required.

Algorithm 2: PM-EWOO

```
Input: $\mathbf{Z}_1, \ldots, \mathbf{Z}_K \overset{iid}{\sim} P^{n_{\mathrm{I}}}$, $\mathbf{Z}_{\mathrm{II}} \sim P^{n_{\mathrm{II}}}$
for $j = 1 \to K$ do $\hat{f}_j = \mathcal{A}^{\mathsf{pm}}_{\mathcal{F}}(\mathbf{Z}_j)$
return $\mathcal{A}^{\mathsf{ew}}_{\mathcal{F}_K}(\mathbf{Z}_{\mathrm{II}})$, with $\mathcal{F}_K = \mathrm{conv}(\{\hat{f}_1, \ldots, \hat{f}_K\})$
```

Algorithm 3: PM-CB

```
Input: $\mathbf{Z}_1, \ldots, \mathbf{Z}_{2K} \overset{iid}{\sim} P^{n_{\mathrm{I}}}$, $\mathbf{Z}_{\mathrm{II}} \sim P^{n_{\mathrm{II}}}$
for $j = 1 \to K$ do $\hat{f}_j = \mathcal{A}^{\mathsf{pm}}_{\mathcal{F}}(\mathbf{Z}_j)$
return $\mathcal{A}^{\mathsf{cb}}(\mathbf{Z}_{K+1}, \ldots, \mathbf{Z}_{2K}, \mathbf{Z}_{\mathrm{II}}, \mathcal{A}^{\mathsf{erm}}_{\mathrm{conv}(\{\hat{f}_1, \ldots, \hat{f}_K\})})$
```

where $\textsc{gap}(\rho, f^*) := \mathsf{E}_Z\left[\mathsf{E}_{f \sim \rho}\left[\ell_f(Z)\right] - \ell_{f^*}(Z)\right]$. This bound enjoys a quantile-like improvement when $\textsc{gap}(\rho, f^*)$ is small. For instance, if there is a set $\mathcal{F}'$ of large prior measure which has excess risk close to f^*, then Theorem 4 pays $\log(1/\pi(\mathcal{F}'))$ for the complexity; in contrast, Theorem A of Lecué and Rigollet (2014) pays a higher complexity price of $\log(1/\pi(f^*))$.

Lastly, we provide a simpler bound by specializing to the case of ρ concentrated entirely on f^*. Then

$$\mathsf{E}_{Z \sim P}\left[\ell_{\hat{f}}(Z) - \ell_{f^*}(Z)\right] \leq \frac{2e\left\lceil \log \frac{2}{\delta} \right\rceil \log \frac{1}{\pi(f^*)}}{\eta n} + \theta_{\mathrm{EW}}(\delta, n).$$

Theorem 4 does not require Lipschitz continuity of the loss, but the rate is suboptimal due to the $\log n$ factor. The next result obtains the correct rate by using ConfidenceBoost for the second stage of the procedure.

Theorem 5 *Take the assumptions of Theorem 4, but instead assume that for each y the loss $\ell \colon \hat{y} \mapsto \ell(y, \hat{y})$ is α-strongly convex and L-Lipschitz (so (α/L^2)-exp-concavity holds). Then with probability at least $1 - \delta$,* PM-CB *run with $K = \lceil \log(3/\delta) \rceil$, $n_{\mathrm{I}} = n/(4K)$ and $n_{\mathrm{II}} = n/2$ learns a hypothesis $\hat{f}$ satisfying*

$$\begin{aligned} &\mathsf{E}_{Z \sim P}\left[\ell_{\hat{f}}(Z) - \ell_{f^*}(Z)\right] \\ &\leq e \cdot \textsc{BayesRed}_\eta\left(\frac{n}{4\left\lceil \log \frac{3}{\delta} \right\rceil}, \pi\right) + \theta_{\mathrm{CB}}(\delta, n), \end{aligned}$$

with $\theta_{\mathrm{CB}}(\delta, n) = O\left(\frac{\left(\log \frac{1}{\delta}\right)^2}{\eta n} + \frac{B \log \frac{1}{\delta}}{n}\right)$.

The proofs of Theorems 4 and 5 are nearly identical. We sketch a proof here, as it uses a novel reduction of the second phase to a low-dimensional stochastic exp-concave optimization problem. For simplicity, we restrict to the case of finite $\mathcal{F}$, uniform prior π, and competing with f^*. We start with an initial procedure that drastically reduces the set of candidates to a set of $O(\log(1/\delta))$. To this end, note that an online-to-batch conversion of the progressive mixture rule run on n samples obtains expected excess risk at most $\log|\mathcal{F}|/(\eta(n+1))$. Hence, K independent runs yield a hypothesis with the same bound inflated by a factor e with probability at least $1 - e^{-K}$. By taking the convex hull of this set of K predictors and reparameterizing the problem, we get a stochastic η-exp-concave optimization problem over the K-dimensional simplex; the best predictor in the convex hull clearly is no worse than the best one in $\mathcal{F}$. Thus, our analyses of EWOO and ConfidenceBoost apply and the results follow.

8 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, the link between exp-concavity and the central condition, suggests that exp-concavity implies a *low noise* condition. Here, low noise can be viewed through the central condition, by the exponential decay of the lower tail of the excess loss random variables, or the Bernstein condition, by the variance of the excess loss of a hypothesis being controlled by its excess risk. The previous in-expectation $O(d/n)$ results of Koren and Levy (2015) and Gonen and Shalev-Shwartz (2016) used the geometric interpretation of exp-concavity, which we boosted to high probability results using the low noise interpretation. It would be interesting to get a high probability $O(d/n)$ result that proceeds purely from a low noise interpretation or purely from a geometric one.

We also developed high probability quantile-like risk bounds for model selection aggregation, one with an optimal rate and another with a slightly suboptimal rate but no dependence on the Lipschitz continuity of the loss. However, our bound form is not yet a full quantile-type bound; it degrades when the $\textsc{gap}$ term is large, while the bound of Lecué and Rigollet (2014) does not have this problem. Yet, our bound provides an improvement when there is a neighborhood around f^* with large prior mass, which the bound of Lecué and Rigollet cannot do. It is an open problem to get a bound with the best of both worlds.

Acknowledgements

This research was supported by the Netherlands Organization for Scientific Research (NWO) VICI Project Nr. 639.073.04. Thanks to Tim van Erven and Bob Williamson for highlighting the connection between exp-concavity and the central condition back in 2014, to Cristóbal Guzmán, Wouter Koolen, and Peter Grünwald for interesting discussions on this work, and to Prateek Jain for first mentioning Kakade and Tewari (2009) in the context of online-to-batch conversions under strong convexity, which ultimately led to the discovery that the results extend to the case of exp-concavity.

References

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

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

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

Nicolò 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.

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

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

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

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.

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.

Guillaume Lecué and Shahar Mendelson. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3-4):591–613, 2009.

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

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.