From Stochastic Mixability to Fast Rates

Authored by:

Robert C. Williamson
Nishant A. Mehta

Abstract

Empirical risk minimization (ERM) is a fundamental learning rule for statistical learning problems where the data is generated according to some unknown distribution $mathsf{P}$ and returns a hypothesis f chosen from a fixed class $mathcal{F}$ with small loss ell. In the parametric setting, depending upon $(ell, mathcal{F},mathsf{P})$ ERM can have slow $(1/sqrt{n})$ or fast $(1/n)$ rates of convergence of the excess risk as a function of the sample size n. There exist several results that give sufficient conditions for fast rates in terms of joint properties of ell, $mathcal{F}$, and $mathsf{P}$, such as the margin condition and the Bernstein condition. In the non-statistical prediction with expert advice setting, there is an analogous slow and fast rate phenomenon, and it is entirely characterized in terms of the mixability of the loss ell (there being no role there for $mathcal{F}$ or $mathsf{P}$). The notion of stochastic mixability builds a bridge between these two models of learning, reducing to classical mixability in a special case. The present paper presents a direct proof of fast rates for ERM in terms of stochastic mixability of $(ell,mathcal{F}, mathsf{P})$, and in so doing provides new insight into the fast-rates phenomenon. The proof exploits an old result of Kemperman on the solution to the general moment problem. We also show a partial converse that suggests a characterization of fast rates for ERM in terms of stochastic mixability is possible.

1 Paper Body

Empirical risk minimization (ERM) is a fundamental learning rule for statistical learning problems where the data is generated according to some unknown distribution P and returns a hypothesis f chosen from a fixed class F with small loss ? ‘. In the parametric setting, depending upon (‘, F, P) ERM can have slow (1/ n) or fast (1/n) rates of convergence of the excess risk as a function of the sample size n. There exist several results that give sufficient conditions for fast rates in terms of joint properties of ‘, F, and P, such as the margin condition and the Bernstein condition. In the non-statistical prediction with expert advice setting, there is an analogous slow and fast rate phenomenon, and it is

entirely characterized in terms of the mixability of the loss ‘ (there being no role there for F or P). The notion of stochastic mixability builds a bridge between these two models of learning, reducing to classical mixability in a special case. The present paper presents a direct proof of fast rates for ERM in terms of stochastic mixability of (‘, F, P), and in so doing provides new insight into the fast-rates phenomenon. The proof exploits an old result of Kemperman on the solution to the general moment problem. We also show a partial converse that suggests a characterization of fast rates for ERM in terms of stochastic mixability is possible.

1

Introduction

Recent years have unveiled central contact points between the areas of statistical and online learning. These include Abernethy et al.?s [1] unified Bregman-divergence based analysis of online convex optimization and statistical learning, the online-to-batch conversion of the exponentially weighted average forecaster (a special case of the aggregating algorithm for mixable losses) which yields the progressive mixture rule as can be seen e.g. from the work of Audibert [2], and most recently Van Erven et al.?s [21] injection of the concept of mixability into the statistical learning space in the form of stochastic mixability. It is this last connection that will be our departure point for this work. Mixability is a fundamental property of a loss that characterizes when constant regret is possible in the online learning game of prediction with expert advice [23]. Stochastic mixability is a natural adaptation of mixability to the statistical learning setting; in fact, in the special case where the function class consists of all possible functions from the input space to the prediction space, stochastic mixability is equivalent to mixability [21]. Just as Vovk and coworkers (see e.g. [24, 8]) have developed a rich convex geometric understanding of mixability, stochastic mixability can be understood as a sort of effective convexity. In this work, we study the O(1/n)-fast rate phenomenon in statistical learning from the perspective of stochastic mixability. Our motivation is that stochastic mixability might characterize fast rates in statistical learning. As a first step, Theorem 5 herein establishes via a rather direct argument that stochastic mixability implies an exact oracle inequality (i.e. with leading constant 1) with a fast rate for finite function classes, and Theorem 7 extends this result to VC-type classes. This result can be understood as a new chapter in an evolving narrative that started with Lee et al.?s [13] seminal paper 1

showing fast rates for agnostic learning with squared loss over convex function classes, and that was continued by Mendelson [18] who showed that fast rates are possible for p-losses (y, y?) 7? —y ? y?—p over effectively convex function classes by passing through a Bernstein condition (defined in (12)). We also show that when stochastic mixability does not hold in a certain sense (described in Section 5), then the risk minimizer is not unique in a bad way. This is precisely the situation at the heart of the works of Mendelson [18] and Mendelson and Williamson [19], which show that having non-unique minimizers is symptomatic of bad geometry of the learning problem. In such situations, there are certain targets (i.e. output conditional distributions) close to the original target under which empirical risk minimization learns (ERM) at a slow rate, where the guilty target depends on the sample size and the target sequence approaches the original target asymptotically. Even the best known upper bounds have constants that blow up in the case of non-unique minimizers. Thus, whereas stochastic mixability implies fast rates, a sort of converse is also true, where learning is hard in a ?neighborhood? of statistical learning problems for which stochastic mixability does not hold. In addition, since a stochastically mixable problem?s function class looks convex from the perspective of risk minimization, and since when stochastic mixability fails the function class looks non-convex from the same perspective (it has multiple well-separated minimizers), stochastic mixability characterizes the effective convexity of the learning problem from the perspective of risk minimization. Much of the recent work in obtaining faster learning rates in agnostic learning has taken place in settings where a Bernstein condition holds, including results based on local Rademacher complexities [3, 10]. The Bernstein condition appears to have first been used by Bartlett and Mendelson [4] in their analysis of ERM; this condition is subtly different from the margin condition of Mammen and Tsybakov [15, 20], which has been used to obtain fast rates for classification. Lecu?e [12] pinpoints that the difference between the two conditions is that the margin condition applies to the excess loss relative to the best predictor (not necessarily in the model class) whereas the Bernstein condition applies to the excess loss relative to the best predictor in the model class. Our approach in this work is complementary to the approaches of previous works, coming from a different assumption that forms a bridge to the online learning setting. Yet this assumption is related; the Bernstein condition implies stochastic mixability under a bounded losses assumption [21]. Further understanding the connection between the Bernstein condition and stochastic mixability is an ongoing effort. ? Contributions. The core contribution of this work is to show a new path to the O(1/n)-fast rate in statistical learning. We are not aware of previous results that show fast rates from the stochastic mixability assumption. Secondly, we establish intermediate learning rates that interpolate between the fast and slow rate under a weaker notion of stochastic mixability. Finally, we show that in a certain sense stochastic mixability characterizes the effective convexity of the statistical problem. In the next section we formally define the statistical problem, review stochastic mixability, and explain our high-level approach toward getting fast rates. This approach involves directly appealing to the Cram?er-Chernoff method, from which nearly all known concentration inequalities arose in one way or another. In Section 3, we frame the problem of computing a particular moment of a certain excess loss random variable as a general moment problem. We sufficiently bound the optimal value of the moment, which allows for a direct application of the Cram?er-Chernoff method. These results easily imply a fast rates bound for finite classes that can be extended to parametric (VC-type) classes, as shown in Section 4. We describe in Section 5 how stochastic mixability characterizes a certain notion of convexity of the statistical learning problem. In Section 6, we extend the fast rates results to classes that obey a notion we call weak stochastic mixability. Finally, Section 7 concludes this work with connections to related topics in statistical learning theory and a discussion of open problems.

2

Stochastic mixability, Cram?er-Chernoff, and ERM

Let (‘, F, P) be a statistical learning problem with ‘ : Y ? R ? R+ a non-negative loss, F ? RX a compact function class, and P a probability measure over X ? Y for input space X and output/target space Y. Let Z be a random variable defined as Z = (X, Y) ? P. We assume for all f ? F, ‘(Y, f (X)) ? V almost surely (a.s.) for some constant V . A probability measure P operates on functions and loss-composed functions as:

P ‘(?, f) = E(X,Y)?P ‘ Y, f (X) .

P f = E(X,Y)?P f (X) 2

Similarly, an empirical measure Pn associated with an n-sample z, comprising n iid samples (x1 , y1), . . . , (xn , yn), operates on functions and loss-composed functions as: n n

1X 1X Pn f = f (xj) Pn ‘(?, f) = ‘ yj , f (xj) . n j=1 n j=1 Let f ? be any function for which P ‘(?, f ?) = inf f ?F P ‘(?, f). For each f ? F define the excess risk random variable Zf := ‘ Y, f (X) ? ‘ Y, f ? (X) . We frequently work with the following two subclasses. For any ? ¿ 0, define the subclasses F? := {f ? F : P Zf ? ?} F? := {f ? F : P Zf ? ?} . 2.1

Stochastic mixability

For ? ¿ 0, we say that (‘, F, P) is ?-stochastically mixable if for all f ? F log E exp(??Zf) ? 0. (1) If ?-stochastic mixability holds for some ? ¿ 0, then we say that (‘, F, P) is stochastically mixable. Throughout this paper it is assumed that the stochastic mixability condition holds, and we take ? ? to be the largest ? such that ?-stochastic mixability holds. Condition (1) has a rich history, beginning from the foundational thesis of Li [14] who studied the special case of ? ? = 1 in density estimation with log loss from the perspective of information geometry. The connections that Li showed between this condition and convexity were strengthened by Gr?unwald [6, 7] and Van Erven et al. [21]. 2.2

Cram?er-Chernoff

The high-level strategy taken here is to show that with high probability ERM will not select a fixed hypothesis function f with excess risk above na for some constant a ¿ 0. For each hypothesis, this guarantee will flow from the Cram?er-Chernoff method [5] by controlling the cumulant generating function (CGF) of ?Zf in a particular way to yield exponential concentration. This control will be possible because the ? ? -stochastic mixability condition implies that the CGF of ?Zf takes the value 0 at some ? ? ? ? , a fact later exploited by our key tool Theorem 3. Let Z be a real-valued random variable. Applying Markov?s inequality to an exponentially transformed random variable yields that, for any ? ? 0 and t ? R Pr(Z ? t) ? exp(??t + log E exp(?Z)); (2) the inequality is non-trivial only if t ¿ E Z and ? ¿ 0. 2.3

Analysis of ERM

We consider the ERM estimator f?z := arg minf ?F Pn ‘(?, f). That is, given an n-sample z, ERM selects any f?z ? F minimizing the empirical risk Pn ‘(?, f). We say ERM is ?-good when f?z ? F? . In order to show that ERM is ?-good it is sufficient to show that for all f ? F F? we have P Zf ¿ 0.

The goal is to show that with high probability ERM is ?-good, and we will do this by showing that with high probability uniformly for all f ? F F? we have Pn Zf ¿ t for some slack t ¿ 0 that will come in handy later. For a real-valued random variable X, recall that the cumulant generating function of X is ? 7? ?X (?) := log E e?X ; we allow ?X (?) to be infinite for some ? ¿ 0. Theorem 1 (Cram?er-Chernoff Control on ERM). Let a ¿ 0 and select f such that E Zf ¿ 0. Let t ¡ E Zf . If there exists ? ¿ 0 such that ??Zf (?) ? ? na , then n o Pr Pn '(?, f) ? Pn '(?, f ?) + t ? exp(?a + ?t). Pn Proof. Let Zf,1 , . . . , Zf,n be iid copies of Zf , and define the sum Sf,n := j=1 ?Zf,j . Since 1 (?t) ¿ E n Sf,n , then from (2) we have X

n 1 1 Pr Zf,j ? t = Pr Sf,n ? ?t ? exp (?t + log E exp(?Sf,n)) n j=1 n n = exp(?t) E exp(??Zf) . 3

Making the replacement ??Zf (?) = log E exp(??Zf) yields

1 log Pr Sf,n ? ?t ? ?t + n??Zf (?). n By assumption, ??Zf (?) ? ? na , and so Pr{Pn Zf ? t} ? exp(?a + ?t) as desired. This theorem will be applied by showing that for an excess loss random variable Zf taking values in [?1, 1], if for some ? ¿ 0 we have E exp(??Zf) = 1 and if E Zf = na for some constant a (that can and must depend on n), then ??Zf (?/2) ? ? c?a n where c ¿ 0 is a universal constant. This is the nature of the next section. We then extend this result to random variables taking values in [?V, V].

3

Semi-infinite linear programming and the general moment problem

The key subproblem now is to find, for each excess loss random variable Zf with mean na and ??Zf (?) = 0 (for some ? ? ? ?), a pair of constants ?0 ¿ 0 and c ¿ 0 for which ??Zf (?0) ? ? ca n. Theorem 1 would then imply that ERM will prefer f ? over this particular f with high probability for ca large enough. This subproblem is in fact an instance of the general moment problem, a problem on which Kemperman [9] has conducted a very nice geometric study. We now describe this problem. The general moment problem. Let P(A) be the space of probability measures over a measurable space A = (A, S). For real-valued measurable functions h and (gj)j?[m] on a measurable space A = (A, S), the general moment problem is inf EX?? h(X) ??P(A) (3) subject to EX?? gj (X) = yj , j ? {1, . . . , m}. Let the vector-valued map g : A ? Rm be defined in terms of coordinate functions as (g(x))j = gj (x), and let the vector y ? Rm be equal to (y1 , . . . , ym). Let D? ? Rm+1 be the set

m X ? ? m+1 D := d = (d0 , d1 , . . . , dm) ? R : h(x) ? d0 + dj gj (x) for all x ? A .

(4)

j=1

Theorem 3 of [9] states that if y ? int conv g(A), the optimal value of problem (3) equals

m X ? ? dj yj : d = (d0 , d1 , . . . , dm) ? D . sup d0 +

(5)

j=1

Our instantiation. We choose A = [?1, 1], set m = 2 and define h, (gj)j?{1,2} , and y ? R2 as: a h(x) = ?e(?/2)x , g1 (x) = x, g2 (x) = e?x , y1 = ? , y2 =

1, n for any ? ¿ 0, a ¿ 0, and n ? N. This yields the following instantiation of problem (3): inf ??P([?1,1])

subject to

EX?? ?e(?/2)X a n = 1.

(6a)

EX?? X = ?

(6b)

EX?? e?X

(6c)

Note that equation (5) from the general moment problem now instantiates to n o a sup d0 ? d1 + d2 : d? = (d0 , d1 , d2) ? D? , n with D? equal to the set n o d? = (d0 , d1 , d2) ? R3 : ?e(?/2)x ? d0 + d1 x + d2 e?x for all x ? [?1, 1] .

(7)

(8)

Applying Theorem 3 of [9] requires the condition y ? int conv g([?1, 1]). We first characterize when y ? conv g([?1, 1]) holds and handle the int conv g([?1, 1]) version after Theorem 3. 4

Lemma 2 (Feasible Moments). The point y = ? na , 1 ? conv g([?1, 1]) if and only if cosh(?) ? 1 a e? + e?? ? 2 = ? . n e? ? e?? sinh(?)

(9)

Proof. Let W denote the convex hull of g([?1, 1]). We need to see if ? na , 1 ? W . Note that W is the convex set formed by starting with the graph of x 7? e?x on the domain [?1, 1], including the line segment connecting this curve?s endpoints (?1, e??) to (1, e?x), and including all of the points below this line segment but above the aforementioned graph. That is, W is precisely the set

e? + e?? e? ? e?? W := (x, y) ? R2 : e?x ? y ? + x, ?x ? [?1, 1] . 2 2 It remains to check that 1 is sandwiched between the lower and upper bounds at x = ? na . Clearly the lower bound holds. Simple algebra shows that the upper bound is equivalent to condition (9). Note that if (9) does not hold, then the semi-infinite linear program (6) is infeasible; infeasibility in turn implies that such an excess loss random variable cannot exist. Thus, we need not worry about whether (9) holds; it holds for any excess loss random variable satisfying constraints (6b) and (6c). The following theorem is a key technical result for using stochastic mixability to control the CGF. The proof is long and can be found in Appendix A. Theorem 3 (Stochastic Mixability Concentration). Let f be an element of F with Zf taking values in [?1, 1], n ? N, E Zf = na for some a ¿ 0, and ??Zf (?) = 0 for some ? ¿ 0. If e? + e?? ? 2 a ¡ , n e? ? e?? E e(?/2)(?Zf) ? 1 ?

then

(10)

0.18(? ? 1)a . n

Note that since log(1 ? x) ? ?x when x ¡ 1, we have ??Zf (?/2) ? ? 0.18(?n? 1)a . In order to apply Theorem 3, we need (10) to hold, but only (9) is guaranteed to hold. The corner case is if (9) holds with equality. However, observe that one can always approximate the random variable X by a perturbed version

X 0 which has nearly identical mean a0 ? a and a nearly identical 0 0 ? 0 ? ? for which EX 0 ??0 e? X = 1, and yet the inequality in (9) is strict. Later, in the proof of Theorem 5, for any random variable that required perturbation to satisfy the interior condition (10), we implicitly apply the analysis to the perturbed version, show that ERM would not pick the (slightly different) function corresponding to the perturbed version, and use the closeness of the two functions to show that ERM also would not pick the original function. We now present a necessary extension for the case of losses with range [0, V], proved in Appendix A. Lemma 4 (Bounded Losses). Let g1 (x) = x and y2 = 1 be common settings for the following two problems. The instantiation of problem (3) with A = [?V, V], h(x) = ?e(?/2)x , g2 (x) = e?x , and y1 = ? na has the same optimal value as the instantiation of problem (3) with A = [?1, 1], h(x) = ?e(V ?/2)x , g2 (x) = e(V ?)x , and y1 = ? a/V n .

4

Fast rates

We now show how the above results can be used to obtain an exact oracle inequality with a fast rate. We first present a result for finite classes and then present a result for VC-type classes (classes with logarithmic universal metric entropy). ? Theorem 5 (Finite Classes Exact Oracle Inequality). Let (‘, F, P) be ? -stochastically mixable, where —F— = N , ‘ is a nonnegative loss, and supf ?F ‘ Y, f (X) ? V a.s. for a constant V . Then for all n ? 1, with probability at least 1 ? ? n o

6 max V, ?1? log 1? + log N ? P ‘(?, f?z) ? P ‘(?, f) + . n 5

(?)

Proof. Let ?n = na for a constant a to be fixed later. For each ? ¿ 0, let F?n ? F?n correspond to those functions in F?n for which ? is the largest constant such that E exp(??Zf) = 1. Let hyper F? ? F?n correspond to functions f in F?n for which lim??? E exp(??Zf) ¡ 1. Clearly, n S (?) hyper F?n = ??[? ? ,?) F?n ? F?n . The excess loss random variables corresponding to elements hyper f ? F? are ?hyper-concentrated? in the sense that they are infinitely stochastically mixable. n However, Lemma 10 in Appendix B shows that for each hyper-concentrated Zf , there exists another excess loss random variable Zf0 with mean arbitrarily close to that of Zf , with E exp(??Zf0) = 1 for some arbitrarily large but finite ?, and with Zf0 ? Zf with probability 1. The last property implies that the empirical risk of Zf0 is no greater than that of Zf ; hence for each hyper-concentrated Zf it is sufficient (from the perspective of ERM) to study a corresponding Zf0 . From now on, we implicitly S (?) make this replacement in F?n itself, so that we now have F?n = ??[?? ,?) F?n . (?)

Consider an arbitrary a ¿ 0. For some fixed ? ? [? ? , ?) for which —F?n — ¿ 0, consider (?)

the subclass F?n . Individually for each such function, we will apply Theorem 1 as follows. From Lemma 4, we have ??Zf (?/2) = ?? V1 Zf (V ?/2). From Theorem 3, the latter is at most 1)(a/V) ? 0.18(V ? ? = ? (V0.18?a n ? ? 1)n . Hence, Theorem 1 with t = 0 and the ? from the Theorem taken to be ?/2 implies that the probability of the event Pn ‘(?, f) ? Pn ‘(?, f ?) is at most

exp ?0.18 V ??? 1 a . Applying the union bound over all of F?n , we conclude that

0.18a Pr {?f ? F?n : Pn ‘(?, f) ? Pn ‘(?, f ?)} ? N exp ?? ? . V ?? ? 1

Since ERM selects hypotheses on their empirical risk, from inversion it holds that with probability at 6 max{V, ?1? }(log ?1 +log N) . least 1 ? ? ERM will not select any hypothesis with excess risk at least n Before presenting the result for VC-type classes, we require some definitions. For a pseudometric space (G, d), for any ? ¿ 0, let N (?, G, d) be the ?-covering number of (G, d); that is, N (?, G, d) is the minimal number of balls of radius ? needed to cover G. We will further constrain the cover (the set of centers of the balls) to be a subset of G (i.e. to be proper), thus ensuring that the stochastic mixability assumption transfers to any (proper) cover of F. Note that the ?proper? requirement at most doubles the constant K below, as shown by Vidyasagar [22, Lemma 2.1]. We now state a localization-based result that allows us to extend the result for finite classes to VCtype classes. Although the localization result can be obtained by combining standard techniques,1 we could not find this particular result in the literature. Below, an ?-net F? of a set F is a subset of F such that F is contained in the union of the balls of radius ? with centers in F? . Theorem 6. Let F be a separable function class whose functions have range bounded in [0, V] and for which, for a constant K ? 1, for each u ? (0, K] the L2 (P) covering numbers are bounded as C K N (u, F, L2 (P)) ? . (11) u Suppose F? is a minimal ?-net for F in the L2 (P) norm, with ? = n1 . Denote by ? : F ? F? an L2 (P)-metric projection from F to F? . Then, provided that ? ? 21 , with probability at most ? can there exist f ? F such that s !

V 1 e Pn f ¡ Pn (?(f)) ? 1080C log(2Kn) + 90 log C log(2Kn) + log . n ? ? The proof is presented in Appendix C. We now present the fast rates result for VC-type classes. The proof (in Appendix C) uses Theorem 6 and the proof of the Theorem 5. Below, we denote the loss-composed version of a function class F as ‘ ? F := {‘(?, f) : f ? F}. 1

See e.g. the techniques of Massart and N?ed?elec [16] and equation (3.17) of Koltchinskii [11].

6

Theorem 7 (VC-Type Classes Exact Oracle Inequality). Let (‘, F, P) be ? ? -stochastically mixable with ‘ ? F separable, where, for a constant K ? 1, for each ? ? (0, K] we have C

N (‘ ? F, L2 (P), ?) ? K , and supf ?F ‘ Y, f (X) ? V a.s. for a constant V ? 1. Then ? for all n ? 5 and ? ? 12 , with probability at least 1 ? ? o n ?

? C log(Kn) + log 2? , 8 max V, ?1? 1 ? q

P ‘(?, fz) ? P ‘(?, f) + max

? 2V 1080C log(2Kn) + 90 log 2 C log(2Kn) + log n ?

? ?

?

5

2e ?

?

+

1 . n

Characterizing convexity from the perspective of risk minimization

In the following, when we say (‘, F, P) has a unique minimizer we mean that any two minimizers f1? , f2? of P ‘(?, f) over F satisfy ‘ Y, f1? (X) = ‘ Y, f2? (X) a.s. We say the excess loss class {‘(?, f) ? ‘(?, f ?) : f ? F} satisfies a (?, B)-Bernstein condition with respect to P for some B ¿ 0 and 0 ¡ ? ? 1 if, for all f ? F: 2 ? P ‘(?, f) ? ‘(?, f ?) ? B P ‘(?, f) ? ‘(?, f ?) . (12) It already is known that the stochastic mixability condition guarantees that there is a unique minimizer [21]; this is a simple consequence of Jensen?s inequality. This leaves open the question: if stochastic mixability does not hold, are there necessarily non-unique minimizers? We show that in a certain sense this is indeed the case, in bad way: the set of minimizers will be a disconnected set.

For any ? ¿ 0, define G? as the class G? := {f ? } ? f ? F : kf ? f ? kL1 (P) ? ? , where in case there are multiple minimizers in F we arbitrarily select one of them as f ? . Since we assume that F is compact and G? {f ? } is equal to F minus an open set homeomorphic to the unit L1 (P) ball, G? {f ? } is also compact. Theorem 8 (Non-Unique Minimizers). Suppose there exists some ? ¿ 0 such that G? is not stochastically mixable. Then there are minimizers f1? , f2? ? F of P ‘(?, f) over F such that it is

? ? not the case that ‘ Y, f1 (X) = ‘ Y, f2 (X) a.s. Proof. Select ? ¿ 0 as in the theorem and some fixed ? ¿ 0. Since G? is not ?-stochastically mixable, there exists f? ? G? such that ??Zf? (?) ¿ 0. Note that there exists ? 0 ? (0, ?) with ??Zf? (? 0) = 0; if not, lim??0

??Zf (?)???Zf ?

?

?

(0)

¿ 0 ? ?0?Zf? (0) ¿ 0, so ?0?Zf? (0) = E(?Zf?)

implies that E Zf? ¡ 0, a contradiction! From Lemma 2, E Zf? ? 0

cosh(? 0)?1 sinh(? 0) ;

for ? 0 ? 0 the RHS

0

0

)?1 2 0 1 has upper bound ?2 since the derivative of ?2 ? cosh(? sinh(? 0) is the nonnegative function 2 tanh (? /2) 0

0)?1 and ?2 ? cosh(? —?0 =0 = 0. Thus, E Zf? ? 0 as ? ? 0. As G? {f ? } is compact, we can take sinh(? 0)

a positive decreasing sequence (?j)j approaching 0, corresponding to a sequence (f?j)j ? G? {f ? } with limit point g ? ? G? {f ? } for which E Zg? = 0, and so there is a risk minimizer in G? {f ? }. The implications of having non-unique risk minimizers. In the case of non-unique risk minimizers, Mendelson [17] showed that for p-losses (y, y?) 7? —y ? y?—p with p ? [2, ?) there is an n-indexed sequence of probability measures (P(n))n approaching the true probability measure as n ? ? such that, for each n, ERM learns at a slow rate under sample size n when the true distribution is P(n) . This behavior is a consequence of the statistical learning problem?s poor geometry: there are multiple minimizers and the set of minimizers is not even connected. Furthermore, in this case, the best known fast rate upper bounds (see [18] and [19]) have a multiplicative constant that approaches ? as the target probability measure approaches a probability measure for which there are non-unique minimizers. The reason for the poor upper bounds in this case is that the constant B in the Bernstein condition explodes, and the upper bounds rely upon the Bernstein condition.

6

Weak stochastic mixability

For some ? ? [0, 1], we say (‘, F, P) is (?, ?0)-weakly stochastically mixable if, for every ? ¿ 0, for all f ? {f ? } ? F? , the inequality log E exp(??? Zf) ? 0 holds with ?? := ?0 ?1?? . This concept was introduced by Van Erven et al. [21] without a name. 7

Suppose that some fixed function has excess risk a = ?. Then, roughly, with high probability ERM does not make a mistake provided that a?a = n1 , i.e. when ? ? ?0 ?1?? = n1 and hence when ? = (?0 n)?1/(2??) . Modifying the proof of the finite classes result (Theorem 5) to consider all functions in the subclass F?n for ?n = (?0 n)?1/(2??) yields the following corollary of Theorem 5. Corollary 9. Let (‘, F, P) be (?, ?0)-weakly stochastically mixable for some ? ? [0, 1], where —F— = N , ‘ is a nonnegative loss, and supf ?F ‘ Y, f (X) ? V a.s. for a constant V . Then for any n ? ?10 V (1??)/(2??) , with probability at least 1 ? ?

6 log 1? + log N ? ? P ‘(?, fz) ? P ‘(?, f) + . (?0 n)1/(2??) It is simple to show a similar result for VC-type classes; the ?-net can still be taken at the resolution 1 ?1/(2??) . n , but we need only apply the analysis to the subclass of F with excess risk at least (?0 n)

7

Discussion

We have shown that stochastic mixability implies fast rates for VC-type classes, using a direct argument based on the Cram?er-Chernoff method and sufficient control of the optimal value of a certain instance of the general moment problem. The approach is amenable to localization in that the analysis separately controls the probability of large deviations for individual elements of F. An important open problem is to extend the results presented here for VC-type classes to results for nonparametric classes with polynomial metric entropy, and moreover, to achieve rates similar to those obtained for these classes under the Bernstein condition. There are still some unanswered questions with regards to the connection between the Bernstein condition and stochastic mixability. Van Erven et al. [21] showed that for bounded losses the Bernstein condition implies stochastic mixability. Therefore, when starting from a Bernstein condition, Theorem 5 offers a different path to fast rates. An open problem is to settle the question of whether the Bernstein condition and stochastic mixability are equivalent. Previous results [21] suggest that the stochastic mixability does imply a Bernstein condition, but the proof was non-constructive, and it relied upon a bounded losses assumption. It is well known (and easy to see) that both stochastic mixability and the Bernstein condition hold only if there is a unique minimizer. Theorem 8 shows in a certain sense that if stochastic mixability does not hold, then there cannot be a unique minimizer. Is the same true when the Bernstein condition fails to hold? Regardless of whether stochastic mixability is equivalent to the Bernstein condition, the direct argument presented here and the connection to classical mixability, which does characterize constant regret in the simpler non-stochastic setting, motivates further study of stochastic mixability. Finally, it would be of great interest to discard the bounded losses assumption. Ignoring the dependence of the metric entropy on the maximum possible loss, the upper bound on the loss V enters the final bound through the difficulty of controlling the minimum value of u? (?1) when ? is large (see the proof of Theorem 3). From extensive experiments with a grid-approximation linear program, we have observed that the worst (CGF-wise) random variables for fixed negative mean and fixed optimal stochastic mixability constant are those which place very little probability mass at ?V and most of the probability mass at a small positive number that scales with the mean. These random variables correspond to functions that with low probability beat f ? by a large (loss) margin but with high probability have slightly higher loss than f ? . It would be useful to understand if this exotic behavior is a real concern and, if not, find a simple, mild condition on the moments that rules it out. Acknowledgments RCW thanks Tim van Erven for the initial discussions around the Cram?er-Chernoff method during his visit to Canberra in 2013 and for his gracious permission to proceed with the present paper without him as an author, and both authors thank him for the further enormously helpful spotting of a serious error in our original proof for fast rates for VC-type classes. This work was supported by the Australian Research Council (NAM and RCW) and NICTA (RCW). NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence program. 8

2 References

[1] Jacob Abernethy, Alekh Agarwal, Peter L. Bartlett, and Alexander Rakhlin. A stochastic view of optimal regret through minimax duality. In Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009), 2009. [2] Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. The Annals of Statistics, 37(4):1591?1646, 2009. [3] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. The Annals of Statistics, 33(4):1497?1537, 2005. [4] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. Probability Theory and Related Fields, 135(3):311?334, 2006. [5] St?ephane Boucheron, G?abor Lugosi, and Pascal Massart. Concentration inequalities: A nonasymptotic theory of independence. Oxford University Press, 2013. [6] Peter Gr?unwald. Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In Proceedings of the 24th International Conference on Learning Theory (COLT 2011), pages 397?419, 2011. [7] Peter Gr?unwald. The safe Bayesian. In Proceedings of the 23rd International Conference on Algorithmic Learning

Theory (ALT 2012), pages 169?183. Springer, 2012. [8] Yuri Kalnishkan and Michael V. Vyugin. The weak aggregating algorithm and weak mixability. In Proceedings of the 18th Annual Conference on Learning Theory (COLT 2005), pages 188?203. Springer, 2005. [9] Johannes H.B. Kemperman. The general moment problem, a geometric approach. The Annals of Mathematical Statistics, 39(1):93?122, 1968. [10] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. The Annals of Statistics, 34(6):2593?2656, 2006. [11] Vladimir Koltchinskii. Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole dEt?e de Probabilit?es de Saint-Flour XXXVIII-2008, volume 2033. Springer, 2011. [12] Guillaume Lecu?e. Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis. Habilitation a‘ diriger des recherches, Universit?e ParisEst, 2011. [13] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. IEEE Transactions on Information Theory, 44(5):1974?1980, 1998. [14] Jonathan Qiang Li. Estimation of mixture models. PhD thesis, Yale University, 1999. [15] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. The Annals of Statistics, 27(6):1808?1829, 1999. ? [16] Pascal Massart and Elodie N?ed?elec. Risk bounds for statistical learning. The Annals of Statistics, 34(5):2326?2366, 2006. [17] Shahar Mendelson. Lower bounds for the empirical minimization algorithm. IEEE Transactions on Information Theory, 54(8):3797?3803, 2008. [18] Shahar Mendelson. Obtaining fast error rates in nonconvex situations. Journal of Complexity, 24(3):380? 397, 2008. [19] Shahar Mendelson and Robert C. Williamson. Agnostic learning nonconvex function classes. In Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT 2002), pages 1?13. Springer, 2002. [20] Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. The Annals of Statistics, 32(1):135?166, 2004. [21] Tim Van Erven, Peter D. Gr?unwald, Mark D. Reid, and Robert C. Williamson. Mixability in statistical learning. In Advances in Neural Information Processing Systems 25 (NIPS 2012), pages 1700?1708, 2012. [22] Mathukumalli Vidyasagar. Learning and Generalization with Applications to Neural Networks. Springer, 2002. [23] Volodya Vovk. A game of prediction with expert advice. Journal of Computer and System Sciences, 56(2):153?173, 1998. [24] Volodya Vovk. Competitive on-line statistics. International Statistical Review, 69(2):213?248, 2001.

9