Fast Rates for Exp-concave Empirical Risk Minimization

Authored by:

Tomer Koren
Kfir Levy

Abstract

We consider Empirical Risk Minimization (ERM) in the context of stochastic optimization with exp-concave and smooth losses—a general optimization framework that captures several important learning problems including linear and logistic regression, learning SVMs with the squared hinge-loss, portfolio selection and more. In this setting, we establish the first evidence that ERM is able to attain fast generalization rates, and show that the expected loss of the ERM solution in d dimensions converges to the optimal expected loss in a rate of d/n. This rate matches existing lower bounds up to constants and improves by a $log{n}$ factor upon the state-of-the-art, which is only known to be attained by an online-to-batch conversion of computationally expensive online algorithms.

1 Paper Body

Statistical learning and stochastic optimization with exp-concave loss functions captures several fundamental problems in statistical machine learning, which include linear regression, logistic regression, learning support-vector machines (SVMs) with the squared hinge loss, and portfolio selection, amongst others. Exp-concave functions constitute a rich class of convex functions, which is substantially richer than its more familiar subclass of strongly convex functions. Similarly to their strongly-convex counterparts, it is well-known that exp-concave loss functions are amenable to fast generalization rates. Specifically, a standard online-to-batch conversion [6] of either the Online Newton Step algorithm [8] or exponential weighting schemes ? [5, 8] in d dimensions gives rise to convergence rate of d/n, as opposed to the standard 1/ n rate of generic (Lipschitz) stochastic convex optimization. Unfortunately, the latter online methods are highly inefficient computationally-wise; e.g., the runtime complexity of the Online Newton Step algorithm scales as d4 with the dimension of the problem, even in very simple optimization scenarios [13]. An alternative and widely-used learning paradigm is that of Empirical Risk Minimization

(ERM), which is often regarded as the strategy of choice due to its generality and its statistical efficiency. In this scheme, a sample of training instances is drawn from the underlying data distribution, and the minimizer of the sample average (or the regularized sample average) is computed. As opposed to methods based on online-to-batch conversions, the ERM approach enables the use of any optimization procedure of choice and does not restrict one to use a specific online algorithm. Furthermore, the ERM solution often enjoys several distribution-dependent generalization bounds in conjunction, and thus is able to obliviously adapt to the properties of the underlying data distribution. In the context of exp-concave functions, however, nothing is known about the generalization abilities ? of ERM besides the standard 1/ n convergence rate that applies to any convex losses. Surprisingly, it appears that even in the specific and extensively-studied case of linear regression with the squared loss, the state of affairs remains unsettled: this important case was recently addressed by Shamir 1

[19], who proved a ?(d/n) lower bound on the convergence rate of any algorithm, and conjectured that the rate of an ERM approach should match this lower bound. In this paper, we explore the convergence rate of ERM for stochastic exp-concave optimization. We show that when the exp-concave loss functions are also smooth, a slightly-regularized ERM approach yields a convergence rate of O(d/n), which matches the lower bound of Shamir [19] up to constants. In fact, our result shows for ERM a generalization rate tighter than the state-of-the-art obtained by the Online Newton Step algorithm, improving upon the latter by a log n factor. Even in the specific case of linear regression with the squared loss, our result improves by a log(n/d) factor upon the best known fast rates provided by the Vovk-Azoury-Warmuth algorithm [3, 22]. Our results open an avenue for potential improvements to the runtime complexity of exp-concave stochastic optimization, by permitting the use of accelerated methods for large-scale regularized loss minimization. The latter has been the topic of an extensive research effort in recent years, and numerous highly-efficient methods have been developed; see, e.g., Johnson and Zhang [10], ShalevShwartz and Zhang [16, 17] and the references therein. On the technical side, our convergence analysis relies on stability arguments introduced by Bousquet and Elisseeff [4]. We prove that the expected loss of the regularized ERM solution does not change significantly when a single instance, picked uniformly at random from the training sample, is discarded. Then, the technique of Bousquet and Elisseeff [4] allows us to translate this average stability property into a generalization guarantee. We remark that in all previous stability analyses that we are aware of, stability was shown to hold uniformly over all discarded training intances, either with probability one [4, 16] or in expectation [20]; in contrast, in the case of exp-concave functions it is crucial to look at the average stability. In order to bound the average stability of ERM, we make use of a localized notion of strong convexity, defined with respect to a local norm at a certain point in the optimization domain. Roughly speaking, we show that when looking at the right norm, which is determined by the local properties of the empirical risk at the right point, the minimizer of the empirical risk becomes stable. This part of our

analysis is inspired by recent analysis techniques of regularization-based online learning algorithms [1], that use local norms to study the regret performance of online linear optimization algorithms. 1.1

Related Work

The study of exp-concave loss functions was initiated in the online learning community by Kivinen and Warmuth [12], who considered the problem of prediction with expert advice with exp-concave losses. Later, Hazan et al. [8] considered a more general framework that allows for a continuous decision set, and proposed the Online Newton Step (ONS) algorithm that attains a regret bound that grows logarithmically with the number of optimization rounds. Mahdavi et al. [15] considered the ONS algorithm in the statistical setting, and showed how it can be used to establish generalization bounds that hold with high probability, while still keeping the fast 1/n rate. Fast convergence rates in stochastic optimization are known to be achievable under various conditions. Bousquet and Elisseeff [4] and Shalev-Shwartz et al. [18] have shown, via a uniform stability argument, that ERM guarantees a convergence rate of 1/n for strongly convex functions. Sridharan et al. [21] proved a similar result, albeit using the notion of localized Rademacher complexity. For the case of smooth and non-negative losses, Srebro et al. [20] established a 1/n rate in low-noise conditions, i.e., when the expected loss of the best hypothesis is of order 1/n. For further discussion of fast rates in stochastic optimization and learning, see [20] and the references therein.

2

Setup and Main Results

We consider the problem of minimizing a stochastic objective F (w) = E[f (w, Z)] d

(1)

over a closed and convex domain W ? R in d-dimensional Euclidean space. Here, the expectation is taken with respect to a random variable Z distributed according to an unknown distribution over a parameter space Z. Given a budget of n samples z1 , . . . , zn of the random variable Z, we are required to produce an estimate w b ? W whose expected excess loss, defined by 2

E[F (w)] b ? minw?W F (w), is small. (Here, the expectation is with respect the randomization of the training set z1 , . . . , zn used to produce w.) b We make the following assumptions over the loss function f . First, we assume that for any fixed parameter z ? Z, the function f (?, z) is ?-exp-concave over the domain W for some ? ¿ 0, namely, that the function exp (??f (?, z)) is concave over W. We will also assume that f (?, z) is ?-smooth over W with respect to Euclidean norm k ? k2 , which means that its gradient is ?-Lipschitz with respect to the same norm: ? w, w0 ? W ,

k?f (w, z) ? ?f (w0 , z)k2 ? ?kw ? w0 k2 .

(2)

In particular, this property implies that f (?, z) is differentiable. For simplicity, and without loss of generality, we assume ? ? 1. Finally, we assume that f (?, z) is bounded over W, in the sense that —f (w, z) ? f (w0 , z)— ? C for all w, w0 ? W for some C ¿ 0. In this paper, we analyze a regularized Empirical

Risk Minimization (ERM) procedure for optimizing the stochastic objective in Eq. (1), that based on the sample z1 , . . . , zn computes w b = arg min Fb(w)
,

(3)

w?W

where n

1X 1 Fb(w) = f (w, zi) + R(w) . n i=1 n

(4)

The function R : W 7? R serves as a regularizer, which is assumed to be 1-strongly-convex with respect to the Euclidean norm; for instance, one can simply choose R(w) = 12 kwk22 . The strong convexity of R implies in particular that Fb is also strongly convex, which ensures that the optimizer w b is unique. For our bounds, we will assume that —R(w) ? R(w0)— ? B for all w, w0 ? W for some constant B ¿ 0. Our main result, which we now present, establishes a fast 1/n convergence rate for the expected excess loss of the ERM estimate w b given in Eq. (3). Theorem 1. Let f : W ? Z 7? R be a loss function defined over a closed and convex domain W ? Rd , which is ?-exp-concave, ?-smooth and B-bounded with respect to its first argument. Let R : W 7? R be a 1-strongly-convex and B-bounded regularization function. Then, for the regularized ERM estimate w b defined in Eqs. (3) and (4) based on an i.i.d. sample z1 , . . . , zn , the expected excess loss is bounded as

d 24?d 100Cd B + + = O . E[F (w)] b ? min F (w) ? w?W ?n n n n In other words, the theorem states that for ensuring an expected excess loss of at most , a sample of size n = O(d/) suffices. This result improves upon the best known fast convergence rates for exp-concave functions by a O(log n) factor, and matches the lower bound of Shamir [19] for the special case where the loss function is the squared loss. For this particular case, our result affirms the conjecture of Shamir [19] regarding the sample complexity of ERM for the squared loss; see Section 2.1 below for details. It is important to note that Theorem 1 establishes a fast convergence rate with respect to the actual expected loss F itself, and not for a regularized version thereof (and in particular, not with respect to the expectation of ? Fb). Notably, the magnitude of the regularization we use is only O(1/n), as opposed to the O(1/ n) regularization used ? in standard regularized loss minimization methods (that can only give rise to a traditional O(1/ n) rate). 2.1

Results for the Squared Loss

In this section we focus on the important special case where the loss function f is the squared loss, namely, f (w; x, y) = 12 (w ?x ? y)2 where x ? Rd is an instance vector and y ? R is a target value. This case, that was extensively studied in the past, was recently addressed by Shamir [19] who gave lower bounds on the sample complexity of any learning algorithm under mild assumptions. 3

Shamir [19] analyzed learning with the squared loss in a setting where the domain is W = {w ? Rd : kwk2 ? B} for some constant B ¿ 0, and the parameters distribution is supported over {x ? Rd : kxk2 ? 1} ? {y ? R : —y— ? B}. It is not hard to verify that in this setup, for the squared loss we can take ? = 1, ? = 4B 2 and C = 2B 2 . Furthermore, if we choose the standard regularizer R(w) = 21 kwk22 , we have —R(w) ? R(w0)— ? 12 B 2 for all w, w0 ? W. As a consequence, Theorem 1 implies that the expected excess loss of the regularized ERM estimator w b we defined in Eq. (3) is bounded by O(B 2 d/n). On the other hand, standard uniform convergence results for generalized linear functions ? [e.g., 11] show that, under the same conditions, ERM also enjoys an upper bound of O(B 2 / n) over its expected excess risk. Overall, we conclude: Corollary 2. For the squared loss f (w; x, y) = 12 (w ? x ? y)2 over the domain W = {w ? Rd : kwk2 ? B} with Z = {x ? Rd : kxk2 ? 1} ? {y ? R : —y— ? B}, the regularized ERM estimator w b defined in Eqs. (3) and (4) based on an i.i.d. sample of n instances has

2

B d B2 , ? E[F (w)] b ? min F (w) = O min . w?W n n This result slightly improves, by a log(n/d) factor, upon the bound conjectured by Shamir [19] for the ERM estimator, and matches the lower bound proved therein up to constants.1 Previous fast-rate results for ERM that we are aware of either included excess log factors [2] or were proven under additional distributional assumptions [14, 9]; see also the discussion in [19]. We remark that Shamir conjectures this bound for ERM without any regularization. For the specific case of the squared loss, it is indeed possible to obtain the same rates without regularizing; we defer details to the full version of the paper. However, in practice, regularization has several additional benefits: it renders the ERM optimization problem well-posed (i.e., ensures that the underlying matrix that needs to be inverted is well-conditioned), and guarantees it has a unique minimizer.

3

Proof of Theorem 1

Our proof of Theorem 1 proceeds as follows. First, we relate the expected excess risk of the ERM estimator w b to its average leave-one-out stability [4]. Then, we bound this stability in terms of certain local properties of the empirical risk at the point w. b To introduce the average stability notion we study, we first define for each i = 1, . . . , n the following empirical leave-one-out risk: 1 1X f (w, zj) + R(w) (i = 1, . . . , n) . Fbi (w) = n n j6=i

Namely, Fbi is the regularized empirical risk corresponding to the sample obtained by discarding the instance zi . Then, for each i we let w bi = arg minw?W Fbi (w) be the ERM estimator bi . The average leave-one-out stability of w corresponding to F b is then defined as the quantity Pn 1 (f (w b , z) ? f (w, b z)). i i i i=1 n Intuitively, the average leave-one-out stability serves as an unbiased estimator of the amount of change in the expected loss of the ERM estimator when one of the instances z1 , . . . , zn , chosen uniformly at random, is removed from the training sample. We note that looking at the average is crucial for us, and the stronger condition of (expected) uniform stability does not hold for expconcave functions. For further discussion of the various stability notions, refer to Bousquet and Elisseeff [4]. Our main step in proving Theorem 1 involves bounding the average leave-one-out stability of w b defined in Eq. (3), which is the purpose of the next theorem. Theorem 3 (average leave-one-out stability). For any z1 , . . . , zn ? Z and for w b1 , . . . , w bn and w b as defined above, we have n

24?d 100Cd 1X f (w bi , zi) ? f (w, b zi) ? + . n i=1 ?n n 1 We remark that Shamir?s result assumes two different bounds over the magnitude of the predictors w and the target values y, while here we assume both are bounded by the same constant B. We did not attempt to capture this refined dependence on the two different parameters.

4

Before proving this theorem, we first show how it can be used to obtain our main theorem. The proof follows arguments similar to those of Bousquet and Elisseeff [4] and Shalev-Shwartz et al. [18]. Proof of Theorem 1. To obtain the stated result, it is enough to upper bound the expected excess loss of w bn , which is the minimizer of the regularized empirical risk over the i.i.d. sample {z1 , . . . , zn?1 }. To this end, fix an arbitrary w? ? W. We first write 1 R(w?) = E[Fb(w?)] ? E[Fb(w)] b , n which holds true since w b is the minimizer of Fb over W. Hence, F (w?) +

1 E[F (w bn)]? F (w?) ? E[F (w bn) ? Fb(w)] b + R(w?) . n

(5)

Next, notice that the random variables w b1 , . . . , w bn have exactly the same distribution: each is the output of regularized ERM on an i.i.d. sample of n ? 1 examples. Also, notice that w bi , which is the minimizer of the sample obtained by discarding the i?th example, is independent of zi . Thus, we have n

E[F (w bn)] =

n

1X 1X E[F (w bi)] = E[f (w bi , zi)] . n i=1 n i=1

Furthermore, we can write n

1 1X E[f (w, b zi)] + E[R(w)] E[Fb(w)] b = b . n i=1 n Plugging these expressions into Eq. (5) gives a bound over the expected excess loss of w bn in terms of the average stability: n

E[F (w bn)]? F (w?) ?

1X 1 E[f (w bi , zi) ? f (w, b zi)] + E[R(w?) ? R(w)] b . n i=1 n

Using Theorem 3 for bounding average stability term on the right-hand side, and our assumption that supw,w0 ?W —R(w) ? R(w0)— ? B to bound the second term, we obtain the stated bound over the expected excess loss of w bn . The remainder of the section is devoted to the proof of Theorem 3. Before we begin with the proof of the theorem itself, we first present a useful tool for analyzing the stability of minimizers of convex functions, which we later apply to the empirical leave-one-out risks. 3.1

Local Strong Convexity and Stability

Our stability analysis for exp-concave functions is inspired by recent analysis techniques of regularization-based online learning algorithms, that make use of strong convexity with respect to local norms [1]. The crucial strong-convexity property is summarized in the following definition. Definition 4 (Local strong convexity). We say that a function g : K 7? R is locally ?-stronglyconvex over a domain K ? Rd at x with respect to a norm k?k, if ?y?K,

g(y) ? g(x) + ?g(x)?(y ? x) +

? ky ? xk2 . 2

In words, a function is locally strongly-convex at x if it can be lower bounded (globally over its entire domain) by a quadratic tangent to the function at x; the nature of the quadratic term in this lower bound is determined by a choice of a local norm, which is typically adapted to the local properties of the function at the point x. With the above definition, we can now prove the following stability result for optima of convex functions, that underlies our stability analysis for exp-concave functions. 5

Lemma 5. Let g1 , g2 : K 7? R be two convex functions defined over a closed and convex domain K ? Rd , and let x1 ? arg minx?K g1 (x) and x2 ? arg minx?K g2 (x). Assume that g2 is locally ?-strongly-convex at x1 with respect to a norm k?k. Then, for h = g2 ? g1 we have kx2 ? x1 k ?

2 k?h(x1)k? . ?

Furthermore, if h is convex then 2 2 k?h(x1)k? . ?

0 ? h(x1) ? h(x2) ? Proof. The local strong convexity of g2 at x1 implies

?g2 (x1)?(x1 ? x2) ? g2 (x1) ? g2 (x2) +

? kx2 ? x1 k2 . 2

Notice that g2 (x1) ? g2 (x2) ? 0, since x2 is a minimizer of g2 . Also, since x1 is a minimizer of g1 , first-order optimality conditions imply that ?g1 (x1)?(x1 ? x2) ? 0, whence ?g2 (x1)?(x1 ? x2) = ?g1 (x1)?(x1 ? x2) + ?h(x1)?(x1 ? x2) ? ?h(x1)?(x1 ? x2) . Combining the observations yields ? kx2 ? x1 k2 ? ?h(x1)?(x1 ? x2) ? k?h(x1)k? ?kx1 ? x2 k , 2 where we have used H?older?s inequality in the last inequality. This gives the first claim of the lemma. To obtain the second claim, we first observe that g1 (x2) + h(x2) ? g1 (x1) + h(x1) ? g1 (x2) + h(x1) where we used the fact that x2 is the minimizer of g2 = g1 + h for the first inequality, and the fact that x1 is the minimizer of g1 for the second. This establishes the lower bound 0 ? h(x1) ? h(x2). For the upper bound, we use the assumed convexity of h to write h(x1) ? h(x2) ? ?h(x1)?(x1 ? x2) ? k?h(x1)k? ?kx1 ? x2 k ?

2 2 k?h(x1)k? , ?

where the second inequality follows from H?older?s inequality, and the final one from the first claim of the lemma. 3.2

Average Stability Analysis

With Lemma 5 at hand, we now turn to prove Theorem 3. First, a few definitions are needed. For brevity, we henceforth denote fi (?) = f (?, zi) for all P i. We let hi = ?fi (w) b be theP gradient of fi n at the point w b defined in Eq. (3), and let H = ?1 Id + i=1 hi hTi and Hi = ?1 Id + j6=i hj hTj for 1 all i, where ? = 12 min{ 4C , ?}. Finally, we will use k?kM to denote the norm induced by a positive ? definite matrix M , ? i.e., kxkM = xT M x. In this case, the dual norm kxk?M induced by M simply equals kxkM ?1 = xT M ?1 x. In order to obtain an upper bound over the average stability, we first bound each of the individual stability expressions fi (w bi)?fi (w) b in terms of a certain norm of the gradient hi of the corresponding function fi . As the proof below reveals, this norm is the local norm at w b with respect to which the leave-one-out risk Fbi is locally strongly convex. Lemma 6. For all i = 1, . . . , n it holds that fi (w bi) ? fi (w) b ?

2 6? khi k?Hi . ?

Notice that the expression on the right-hand side might be quite large for a particular function fi ; indeed, uniform stability does not hold in our case. However, as we show below, the average of these expressions is small. The proof of Lemma 6 relies on Lemma 5 above and the following property of exp-concave functions, established by Hazan et al. [8]. 6

Lemma 7 (Hazan et al. [8], Lemma 3). Let f : K 7? R be an ?-exp-concave function over a convex 1 domain K ? Rd such that —f (x) ? f (y)— ? C for any x, y ? K. Then for any ? ? 21 min{ 4C , ?} it holds that 2 ? ? x, y ? K , f (y) ? f (x) + ?f (x)?(y ? x) + (6) ?f (x)?(y ? x) . 2 Proof of Lemma 6. We apply Lemma 5 with g1 = Fb and g2 = Fbi (so that h = ? n1 fi). We should first verify that Fbi is indeed (?/n)-strongly-convex at w b with respect to the norm k?kHi . Since each fi is ?-exp-concave, Lemma 7 shows that for all w ? W, 2 ? , (7) fi (w) ? fi (w) b + ?fi (w)?(w b ? w) b + hi ?(w ? w) b 2 with our choice of ? =

1 2

1 min{ 4C , ?}. Also, the strong convexity of the regularizer R implies that

1 R(w) ? R(w) b + ?R(w)?(w b ? w) b + kw ? wk b 22 . 2 Summing Eq. (7) over all j 6= i with Eq. (8) and dividing through by n gives 2 ? X 1 Fbi (w) ? Fbi (w) b + ?Fbi (w)?(w b ? w) b + hi ?(w ? w) b + kw ? wk b 22 2n 2n

(8)

j6=i

? kw ? wk b 2Hi , = Fbi (w) b + ?Fbi (w)?(w b ? w) b + 2n which establishes the strong convexity. Now, applying Lemma 5 gives 2n 2 k?h(w)k b ?Hi = khi k?Hi . ? ? On the other hand, since fi is convex, we have kw bi ? wk b Hi ?

(9)

fi (w bi) ? fi (w) b ? ?fi (w bi)?(w bi ? w) b

= ?fi (w)?(b w bi ? w) b + ?fi (w bi) ? ?fi (w) b ?(w bi ? w) b .

(10)

The first term can be bounded using H?older?s inequality and Eq. (9) as 2 2 khi k?Hi . ? Also, since fi is ?-smooth (with respect to the Euclidean norm), we can bound the second term in Eq. (10) as follows:

?fi (w bi) ? ?fi (w) b ?(w bi ? w) b ? k?fi (w bi) ? ?fi (w)k b 2 ?kw bi ? wk b 2 ? ?kw bi ? wk b 22 , ?fi (w bi)?(w bi ? w) b = hi ?(w bi ? w) b ? khi k?Hi ?kw bi ? wk b Hi ?

and since Hi (1/?)Id , we can further bound using Eq. (9), 2 4 khi k?Hi . ? Combining the bounds (and simplifying using our assumption ? ? 1) gives the lemma. kw bi ? wk b 22 ? ?kw bi ? wk b 2Hi ?

Next, we bound a sum involving the local-norm terms introduced in Lemma 6. Lemma 8. Let I = {i ? [n] : khi k?H ¿ 12 }. Then —I— ? 2d, and we have X 2 khi k?Hi ? 2d . i?I /

Proof. Denote ai = hTi H ?1 hi for all i = 1, . . . , n. First, we claim that ai ¿ 0 for all i, and P ?1 being positive-definite. For the sum of the i ai ? d. The fact that ai ¿ 0 is evident from H ai ?s, we write: n X i=1

ai =

n X i=1

hTi H ?1 hi =

n X

tr(H ?1 hi hTi) ? tr(H ?1 H) = tr(Id) = d ,

i=1

7

(11)

where we have used the linearity of the trace, and the fact that H

Pn

i=1

hi hTi .

Now, our claim that —I— ? 2d is evident: if khi k?H ¿ 12 for more than 2d terms, then the sum P P T ?1 hi must be larger than d, which is a contradiction to Eq. (11). To prove i?I ai = i?I hi H our second claim, we first write Hi = H ? hi hTi and use the Sherman-Morrison identity [e.g., 7] to obtain Hi?1 = (H ? hi hTi)?1 = H ?1 +

H ?1 hi hTi H ?1 1 ? hTi H ?1 hi

for all i ? / I. Note that for i ? / I we have hTi H ?1 hi ¡ 1, so that the identity applies and the inverse on the right-hand side is well defined. We therefore have:

2 hT H ?1 hi 2 a2i = hTi Hi?1 hi = hTi H ?1 hi + i T ?1 khi k?Hi = ai + ? 2ai , 1 ? ai 1 ? hi H hi where the inequality follows from the fact that 1 ? ai ? ai for i ? / I. Summing this inequality over i? / I and recalling that the ai ?s are nonnegative, we obtain X

khi k?Hi

2

? 2

i?I /

X

ai ? 2

n X

ai = 2d ,

i=1

i?I /

which concludes the proof. Theorem 3 is now obtained as an immediate consequence of our lemmas above. Proof of Theorem 3. As a consequence of Lemmas 6 and 8, we have

1X C—I— 2Cd fi (w bi) ? fi (w) b ? ? , n n n i?I

and

2 1X 6? X 12?d fi (w bi) ? fi (w) b ? khi k?Hi . ? n ?n ?n i?I /

Summing the inequalities and using

4

i?I /

1 ?

= 2 max{4C, ?1 } ? 2(4C + ?1) gives the result.

Conclusions and Open Problems

We have proved the first fast convergence rate for a regularized ERM procedure for exp-concave loss functions. Our bounds match the existing lower bounds in the specific case of the squared loss up to constants, and improve by a logarithmic factor upon the best known upper bounds achieved by online methods. Our stability analysis required us to assume smoothness of the loss functions, in addition to their exp-concavity. We note, however, that the Online Newton Step algorithm of Hazan et al. [8] for online exp-concave optimization does not require such an assumption. Even though most of the popular exp-concave loss functions are also smooth, it would be interesting to understand whether smoothness is indeed required for the convergence of the ERM estimator we study in the present paper, or whether it is simply a limitation of our analysis. Another interesting issue left open in our work is how to obtain bounds on the excess risk of ERM that hold with high probability, and not only in expectation. Since the excess risk is non-negative, one can always apply Markov?s inequality to obtain a bound that holds with probability 1 ? ? but scales linearly with 1/?. Also, using standard concentration inequalities p (or success amplification techniques), we may also obtain high probability bounds that scale with log(1/?)/n, losing the fast 1/n rate. We leave the problem of obtaining bounds that depends both linearly on 1/n and logarithmically on 1/? for future work. 8

2 References

[1] J. D. Abernethy, E. Hazan, and A. Rakhlin. Interior-point methods for full-information and bandit online learning. Information Theory, IEEE Transactions on, 58(7):4164?4175, 2012. [2] M. Anthony and P. L. Bartlett. Neural network learning: Theoretical foundations. cambridge university press, 2009. [3] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. Machine Learning, 43(3):211?246, 2001. [4] O. Bousquet and A. Elisseeff. Stability and generalization. The Journal of Machine Learning Research, 2:499?526, 2002. [5] N. Cesa-Bianchi and G. Lugosi. Prediction, learning, and games. Cambridge University Press, 2006. [6] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. IEEE Transactions on Information Theory, 50(9):2050?2057, 2004. [7] G. H. Golub and C. F. Van Loan. Matrix computations, volume 3. JHU Press, 2012. [8] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. Machine Learning, 69(2-3):169?192, 2007. [9] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. Foundations of Computational Mathematics, 14(3):569?600, 2014. [10] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In Advances in Neural Information Processing Systems, pages 315?323, 2013. [11] S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In Advances in neural information processing systems, pages 793?800, 2009. [12] J. Kivinen and M. K. Warmuth. Averaging expert predictions. In Computational Learning Theory, pages 153?167. Springer, 1999. [13] T. Koren. Open problem: Fast stochastic

exp-concave optimization. In Conference on Learning Theory, pages 1073?1075, 2013. [14] G. Lecu?e and S. Mendelson. Performance of empirical risk minimization in linear aggregation. arXiv preprint arXiv:1402.5763, 2014. [15] M. Mahdavi, L. Zhang, and R. Jin. Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In Proceedings of The 28th Conference on Learning Theory, 2015. [16] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. The Journal of Machine Learning Research, 14(1):567?599, 2013. [17] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. Mathematical Programming, pages 1?41, 2014. [18] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. The Journal of Machine Learning Research, 11:2635?2670, 2010. [19] O. Shamir. The sample complexity of learning linear predictors with the squared loss. arXiv preprint arXiv:1406.5143, 2014. [20] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In Advances in neural information processing systems, pages 2199?2207, 2010. [21] K. Sridharan, S. Shalev-Shwartz, and N. Srebro. Fast rates for regularized objectives. In Advances in Neural Information Processing Systems, pages 1545?1552, 2009. [22] V. Vovk. Competitive on-line statistics. International Statistical Review, 69(2):213?248, 2001.

9