JMLR: Workshop and Conference Proceedings 19 (2011) 813–816 24th Annual Conference on Learning Theory

Bounds on Individual Risk for Log-loss Predictors

Peter D. Grünwald PDG@CWI.NL and **Wojciech Kotłowski** KOTLOWSK@CWI.NL
Centrum Wiskunde & Informatica
Amsterdam, the Netherlands

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

In sequential prediction with log-loss as well as density estimation with risk measured by KL divergence, one is often interested in the *expected instantaneous loss*, or, equivalently, the *individual risk* at a given fixed sample size n. For Bayesian prediction and estimation methods, it is often easy to obtain bounds on the *cumulative risk.* Such results are based on bounding the individual sequence regret, a technique that is very well known in the COLT community. Motivated by the easiness of proofs for the cumulative risk, our open problem is to use the results on cumulative risk to prove corresponding individual-risk bounds.

Background We consider sequential prediction (online learning) with log-loss (Cesa-Bianchi and Lugosi, 2006). In each iteration $n = 1, 2, \ldots$, after observing a sequence of past outcomes $x^n = x_1, x_2, \ldots, x_n \in \mathcal{X}^n$, a prediction strategy assigns a probability distribution on $\mathcal{X}$, denoted $\hat{P}(\cdot \mid x^n)$. Then, a next outcome x_{n+1} is revealed and the strategy incurs the *log loss* $-\log \hat{P}(x_{n+1} \mid x^n)$. The goal of the prediction strategy is to be not much worse than the best in a reference set of distributions (also called "experts"), which we call the *model* $\mathcal{M}$.

In online learning, the performance of a prediction strategy is usually measured by the *regret*, which is the difference between the accumulated loss of the prediction strategy and the best distribution in the model. The goal is then to minimize the regret in the worst case over all possible data sequences. This problem is relatively well-explored as it has been investigated in such fields as statistics, information theory, finance and machine learning. For example, it is known that: (1) when the model is finite (contains a finite number of distributions, say N), it is possible to obtain a constant bound $\log N$ on the regret, (2) when the model is infinite, but parametric (e.g. exponential families), a bound of the form $\frac{k}{2} \log n + O(1)$ is usually possible, where k is the number of parameters (Grünwald, 2007).

In statistics, more focus is traditionally put on the *instantaneous* rather than cumulative losses of the prediction strategy: one wants the loss when predicting x_n to be small for fixed n, and to go to 0 at a fast rate as n increases. Since it is not possible to meaningfully bound instantaneous loss for adversarial data, one assumes that the data are sampled form a distribution P^*. Then, it is reasonable to define the *individual risk* or *instantaneous redundancy* in the n-th iteration as the difference between the expected loss of the prediction strategy and the expected loss of the best (w.r.t. P^*) distribution in the model:

$$RISK_n(\hat{P}, P^*) = \mathbb{E}_{P^*}[-\log \hat{P}(X_{n+1}|X^n)] - \inf_{P \in \mathcal{M}} \{\mathbb{E}_{P^*}[-\log P(X_{n+1} \mid X^n)]\}$$

(note that we use capitals to denote both distributions and their densities/mass functions). Although our questions can be phrased more generally, for simplicity we will assume that data are i.i.d. and that $P^* \in \mathcal{M}$. Then the infimum in the above is attained by P^* and the expression simplifies to:

$$RISK_n(\hat{P}, P^*) = \mathbb{E}_{X^{n+1} \sim P^*}[-\log \hat{P}(X_{n+1}|X^n)] - \mathbb{E}_{X \sim P^*}[-\log P^*(X)]$$

The aforementioned results about regret immediately imply results about *cumulative* risk. For example, for k- parameter exponential families, Bayesian, ML (maximum likelihood prediction), NML (normalized maximum likelihood ("Shtarkov")) and several other well-known strategies achieve cumulative risk $\sum_{i=1}^n RISK_i(\hat{P}, P^*) = (k/2)\log n + O(1)$. In general, especially for Bayesian strategies, it is easy to obtain bounds on the cumulative risk. For example, if $\mathcal{M}$ is countable and $\hat{P}$ is the Bayesian predictive distribution based on prior W such that $W(P^*) > 0$, then one has that

$$\sum_{i=1}^n RISK_i(\hat{P}, P^*) \leq -\log W(P^*).$$

The proof is completely straightforward using an individual-sequence regret argument (Grünwald, 2007, Chapter 6). The question we ask is what can be said about the individual risk of a prediction strategy $\hat{P}$, given the performance in terms of such cumulative risk. In particular, let $\hat{P}$ be defined as a Bayesian predictive distribution. We ask:

Our Questions

1. When the model is a k-parameter exponential family, is a bound of the form $\frac{k}{2n} + O(1/n^2)$ possible?

2. When the model is countably infinite, is it possible to obtain a bound on the individual risk of the form $\frac{-\log W(P^*)}{n^\gamma}$ for some $\gamma > 0$ (preferrably $\gamma \geq 1$)?

Known results.

1. Cesaro As noted by e.g. Barron (1998); Yang (2000) (see also Catoni (1997)) it is possible to establish a relationship between the cumulative risk of a prediction strategy and the individual risk of a modified strategy using the notion of *Cesaro averaging.* Let $\hat{P}$ be any prediction strategy and define: $\hat{P}_{\text{Cesaro}}(x_{n+1}|x^n) = \frac{1}{n}\sum_{i=1}^n \hat{P}(x_{i+1}|x^i)$. It turns out that $RISK_n(\hat{P}_{\text{Cesaro}}, P^*) \leq \frac{1}{n}\sum_{i=1}^n RISK_i(\hat{P}, P^*)$. Unfortunately, this statement does not say anything about the individual risk of the original strategy $\hat{P}$, only about its Cesaro average $\hat{P}_{\text{Cesaro}}$. In practice, the Cesaro average will often perform worse than the original $\hat{P}$: *Cesaro averaging is good to prove things, not to improve things.* Moreover, in question 1 above the Cesaro-strategy, when applied to Bayesian strategies, gives a rate bounded by $n^{-1}(k/2)\log n$, which is suboptimal by a factor of $\log n$: it is known that, e.g. with the ML estimator, an individual risk of $O(1/n)$ is achieveable.

2. Follow the Leader So far, the only case for which individual risk results are relatively well-studied seems to be the maximum likelihood (also known as "follow the leader") strategy for exponential families. Grünwald and de Rooij (2005) proved that for one-parameter exponential family, when the data are generated i.i.d. by a distribution P^*, possibly outside $\mathcal{M}$, then the individual risk of the maximum likelihood decreases as:

$$RISK_n(\hat{P}, P^*) = \frac{1}{2n} \mathrm{var}(P^*) \cdot I(\bar{P}) + O(1/n^2), \tag{1}$$

where $\mathrm{var}(P^*)$ is a variance of P^*, I is a Fisher information, while $\bar{P}$ is the element in $\mathcal{M}$ closest to P^* in terms of KL-divergence $D(P^* \| P)$. In particular, if $P^* \in \mathcal{M}$, then $P^* = \bar{P}$ and $\mathrm{var}(P^*) = I^{-1}(P^*)$, so that the bound takes the form $\frac{1}{2n} + O(1/n^2)$, which is the optimal rate for a one-dimensional exponential family. Forster and Warmuth (2002) considered maximum likelihood for k-dimensional exponential families and managed to prove the bound of the form:

$$RISK_n(\hat{P}, P^*) \leq \frac{1}{2(n-1)} \mathrm{tr}\{\mathrm{cov}(P^*)\} \cdot \sup_{P \in \mathcal{M}} \|I(P)\|, \tag{2}$$

where $\mathrm{cov}(P^*)$ is the covariance matrix for P^*. The bounds (1) and (2) are very similar, but essentially incomparable. The latter is a true bound, which holds for all n and any exponential family, but the constant in front of $O(\frac{1}{n})$ is not optimal. The former is an asymptotic expansion of the individual risk with the optimal constant in front of $O(\frac{1}{n})$. Both results concern only a particular prediction strategy (ML), which is known to be suboptimal when $P^* \notin \mathcal{M}$, and cannot be easily extended to say anything about any asymptotically optimal strategy, such as Bayes.

3. 2-part MDL If $\hat{P}(X_{n+1} \mid X^n)$ is taken to be the 2-part MDL estimator achieving $\min_{P \in \mathcal{M}} - \log W(P) - \log P(X^n)$, then one can use a result due to Barron, Cover, Li and Zhang to get a bound on the squared Hellinger distance between $\hat{P}$ and P^*. If all distributions in $\mathcal{M}$ have uniformly bounded density ratios, i.e. $\sup_{x \in \mathcal{X}, P, Q \in \mathcal{M}} P(X)/Q(X) < \infty$, then this translates into a bound on the instantaneous risk. With the original bound (see (Grünwald, 2007, Chapter 15) for a simple statement and proof), one gets a bound $O(-\log W(P^*)(\log n)/n)$ on the individual risk for the two-part MDL prediction strategy. This can be refined (Zhang, 2006) to get $O(-\log W(P^*)/n)$. Strangely, if $\hat{P}$ is set to be a Bayesian predictive distribution (which usually works better in practice), then nothing is known about the individual risk.

4. Decreasing risk!? Let $a_1, a_2, \ldots$ be any sequence of number such that $\sum_{i=1}^n a_n \leq C \log n$. It can be easily shown (Grünwald, 2007) that such a sequence does not necessarily converge to 0. Bounding $\sum_{i=1}^n a_n \leq C$ does imply that a_n converges to 0, but it can converge at arbitrarily slow rate. However, if we additionally assume that the sequence a_n is non-increasing, we immediately get optimal-rate bounds $a_n \leq \frac{C}{n}$ in the first question. Thus, one strategy to address our questions for a given model $\mathcal{M}$ would be to first show that individual risks of $\hat{P}$ are monotonically *decreasing*. It is known that e.g. if $\mathcal{M}$ is the Gaussian location family, then the risk of the ML predictions is strictly decreasing; on the other hand in some cases the risk of the Bayesian strategy can slightly increase at some n. Consider e.g. the Bernoulli model with a uniform prior, and assume the data is a sequence of independent fair coin flips, i.e. they are i.i.d. Bernoulli 1/2. In that case the risk at sample size 1 is 0, because the Bayesian predictive distribution based on the uniform prior and no data is $P(X_1 = 1) = 1/2$. At sample size 2, the Bayesian predictive distribution is $P(X_2 = 1 \mid X_1 = x)$ which is either 2/3 (if $x = 1$) or 1/3 (if $x = 0$). In both cases, the risk increases Barron (1998); Grünwald (2007). So increasing risk is possible. Still, no examples are known of substantially increasing risk at large n. Thus, maybe one might prove that some tight enough upper bound on the risk is still decreasing...

References

A.R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In A.P. Dawid J.M. Bernardo, J.O. Berger and A.F.M. Smith, editors, *Bayesian Statistics*, volume 6, pages 27–52. Oxford University Press, Oxford, 1998.

O. Catoni. A mixture approach to universal model selection. preprint LMENS 97-30, 1997. Available from `http://www.dma.ens.fr/edition/preprints/Index.97.html`.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games.* Cambridge University Press, 2006.

Jürgen Forster and Manfred K. Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 64(1):76–102, 2002.

P. D. Grünwald. *The Minimum Description Length Principle.* MIT Press, Cambridge, MA, 2007.

P. D. Grünwald and S. de Rooij. Asymptotic log-loss of prequential maximum likelihood codes. In *Conference on Learning Theory (COLT 2005)*, pages 652–667, 2005.

Y. Yang. Mixing strategies for density estimation. 28(1):75–87, 2000.

Tong Zhang. From ϵ-entropy to KL entropy: analysis of minimum information complexity density estimation. 34(5):2180–2210, 2006.