Online Prediction – Bayes versus Experts

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch http://www.idsia.ch/∼marcus

19–21 July 2004

Abstract

We derive a very general regret bound in the framework of prediction with expert advice, which challenges the best known regret bound for Bayesian sequence prediction. Both bounds of the form $\sqrt{\text{Loss}\times\text{complexity}}$ hold for any bounded loss-function, any prediction and observation spaces, arbitrary expert/environment classes and weights, and unknown sequence length.

Sequential/online predictions. In sequential or online prediction, for $t=1,2,3,...$, a predictor p makes a prediction $y_t^p\in\mathcal{Y}$ based on past observations $x_1,...,x_{t-1}$; thereafter $x_t\in\mathcal{X}$ is observed and p suffers loss $\ell(x_t,y_t^p)$. The goal is to design predictors with small total loss $L_n^p:=\sum_{t=1}^n\ell(x_t,y_t^p)$. Applications are abundant, e.g. weather or stock market forecasting.

Bayesian Sequence Prediction. In the Bayesian approach to sequence prediction, the definition of the Bayes-optimal mixture predictor is straight-forward. The Bayesian framework assumes that the sequence $x_1...x_n$ is sampled from some distribution μ, i.e. the probability of $x_{<t}:=x_1...x_{t-1}$ is $\mu(x_{<t})$ and the probability of the next symbol being x_t, given $x_{<t}$, is $\mu(x_t|x_{<t})$. The μ-expected loss (given $x_{<t}$) when some predictor Λ predicts the t^{th} symbol and the total μ-expected loss in the first n predictions are

$$\bar l_t^\Lambda(x_{<t}) \;:=\; \sum_{x_t}\mu(x_t|x_{<t})\ell(x_t,y_t^\Lambda), \qquad \bar L_n^\Lambda \;:=\; \sum_{t=1}^n\sum_{x_{<t}}\mu(x_{<t})\cdot\bar l_t^\Lambda(x_{<t}).$$

The goal is to minimize the μ-expected loss. More generally, we define the Λ_ρ sequence prediction scheme

$$y_t^{\Lambda_\rho} := \arg\min_{y_t\in\mathcal{Y}}\sum_{x_t}\rho(x_t|x_{<t})\ell(x_t,y_t),$$

which minimizes the ρ-expected loss. If μ is known, Λ_μ is obviously the best prediction scheme in the sense of achieving minimal expected loss ($\bar{l}_t^{\Lambda_\mu} \le \bar{l}_t^\Lambda$ for *all* Λ). Typically μ is unknown, but known to belong to a class of distributions $\mathcal{M}$. For countable $\mathcal{M}$ the Bayesian solution is to consider the mixture distribution $\xi(x) := \sum_{\nu\in\mathcal{M}} \exp(-k^\nu)\nu(x)$ with $\sum_{\nu\in\mathcal{M}}\exp(-k^\nu)=1$, where $\exp(-k^\nu)$ may be interpreted as the prior belief in ν. For finite $\mathcal{M}$, the uniform choice $k^\nu = \ln|\mathcal{E}|\ \forall\nu\in\mathcal{M}$ is common. Under certain conditions, the loss $\bar{L}_n^{\Lambda_\xi}$ is bounded by the loss $\bar{L}_n^\Lambda$ of *any* other predictor Λ (and hence by the loss of the best predictor in hindsight Λ_μ) in the following way:

$$\bar{L}_n^{\Lambda_\xi} \;\le\; \bar{L}_n^\Lambda + 2\sqrt{\bar{L}_n^\Lambda \cdot k^\mu} + 2\cdot k^\mu \quad \forall\mu\in\mathcal{M} \quad \forall\Lambda \tag{1}$$

Note that $\bar{L}_n^\Lambda$ depends on μ. For countable $\mathcal{M}$ and $\mathcal{X}$, finite $\mathcal{Y}$, any k^μ, and any bounded loss function $\ell : \mathcal{X}\times\mathcal{Y}\to[0,1]$, bound (1) has been proven in [Hut03].

Prediction with Expert Advice (PEA). Contrary to the straight-forward definition of Bayes-optimal predictors, designing well-performing PEA-master algorithms is an art. In the PEA framework one considers a countable class of predictors $\mathcal{E}=\{e_1,e_2,...\}$, called experts. Typically no assumptions are made on (the process generating) the observation sequence $x_1...x_n$. The price for this generality is that there are no absolute performance assertions, but there are strong relative guarantees: Consider the expert $\varepsilon := \arg\min_{e\in\mathcal{E}} L_n^e$, which performs best on sequence $x_1...x_n$. Prediction scheme ε is infeasible, since L_n^ε depends on $x_1...x_n$, not known in advance. But we can ask how close we can come to L_n^ε with a *master algorithm* M which dynamically chooses among or combines the experts $e\in\mathcal{E}$ at time t based only on the known past performance L_{t-1}^e. The naive idea of selecting the expert e which worked best in the past (i.e. $y_t^M = \arg\min_{e\in\mathcal{E}} L_{t-1}^e$) can fail due to oscillations, but refinements selecting expert e with high/low *probability* w_t^e if L_{t-1}^e is small/large work. For infinite classes of experts it is also necessary to add a penalty k^e to the loss of each expert e with $\sum_{e\in\mathcal{E}}\exp(-k^e)=1$. For finite $\mathcal{E}$, the uniform choice $k^e=\ln|\mathcal{E}|\ \forall e$ is common. The "Weighted Majority" (WM) algorithm predicts y_t^e with probability $w_t^e \propto \exp(-\eta_t L_{t-1}^e - k^e)$ with suitable learning parameter $\eta_t \searrow 0$ [LW89, Vov90, CB97, ACBG02, YEYS04]. The recently revived "Follow the Perturbed Leader" (FPL) algorithm selects expert e of minimal $\eta_t L_{t-1}^e + k^e + Q_t^e$ for prediction, where Q_t^e is a random perturbation, i.e. $w_t^e = P[\eta_t L_{t-1}^e + k^e + Q_t^e \le \eta_t L_{t-1}^{e'} + k^{e'} + Q_t^{e'} \forall e']$ [Han57, KV03, HP04]. We are interested in the expected loss $\underline{L}_n^M := E[L_n^M]$ of M relative to $\underline{L}_n^\varepsilon := E[L_n^\varepsilon]$ of the best expert in hindsight. If the set $\mathcal{Y}$ is convex, the master algorithm may, instead of a randomized prediction, make the deterministic prediction $y_t^m := \sum_{e\in\mathcal{E}} w_t^e y_t^e \in \mathcal{Y}$. For convex (in y) loss-functions $\ell(x,y)$ an expected bound on L_n^M implies a for-sure bound on L_n^m, since $L_n^m \le \underline{L}_n^M$. There are many *static* bounds ($\eta_t = const.$) if n or L_n is known in advance. We only review *adaptive* bounds which do not require such extra knowledge. Under certain conditions, the following bound can be proven:

$$L_n^m \le \underline{L}_n^M \le L_n^e + a\cdot\sqrt{L_n^e \cdot k^e} + b\cdot k^e \quad \forall e\in\mathcal{E} \quad \forall x_1...x_n, \tag{2}$$

where a and b are small positive constants. For finite $\mathcal{E}$, $k^e = \ln|\mathcal{E}|$, $\mathcal{X} = \mathcal{Y} = [0,1]$, and $\ell(x,y) = |x-y|$, the bound (2) on L_n^m for WM-type masters has been proven in [CB97] with $a = 2.8$ and $b = 4$ via a doubling trick, and in [ACBG02, YEYS04] for smooth $\eta_t \to 0$ with better constants. We have shown that all four assumptions can be relaxed for FPL-type masters: A bound (2) on $\underline{L}_n^M$ (and hence L_n^m for convex $\mathcal{Y}$ and ℓ) for any $\mathcal{X}$ and $\mathcal{Y}$ and any bounded loss function $\ell : \mathcal{X} \times \mathcal{Y} \to [0,1]$ has been derived. For finite $\mathcal{E}$ and $k^e = \ln|\mathcal{E}|$, the constants are $a = 2\sqrt{2}$ and $b = 8$. A hierarchy of experts allowed to generalize this result to infinite $\mathcal{E}$ and arbitrary k^e with constant a arbitrarily close to $2\sqrt{2}$. The interested reader can find the derivation in the Technical Report [HP04].

PEA versus Bayes. The formal similarity and duality between Bayes bound (1) and PEA bound (2) is striking. Whereas randomized PEA M performs well in any environment, but only relative to a given set of experts $\mathcal{E}$, deterministic Bayes Λ_ξ competes with any other predictor Λ, but only in a given set of probabilistic environments $\mathcal{M}$. M depends on the set of experts $\mathcal{E}$, Λ_ξ depends on the set of environments $\mathcal{M}$. $\underline{E}$xpectations in PEA-bounds are over the randomized Master algorithm, while $\bar{E}$xpectations in Bayes-bounds are over environmental sequences. Apart from these formal relations, there is a real connection between both bounds. The class of Bayes-predictors $\{\Lambda_\nu : \nu \in \mathcal{M}\}$ may be regarded as a class of experts $\mathcal{E}$. The corresponding master algorithm M then satisfies bound (2), i.e. $\underline{L}_n^M \leq L_n^{\Lambda_\nu} + a\sqrt{L_n^{\Lambda_\nu} k^\nu} + bk^\nu$. Setting $\nu = \mu$, taking the μ-expectation, using Jensen's inequality and $E[L_n^{\Lambda_\mu}] \equiv \bar{L}_n^{\Lambda_\mu} \leq \bar{L}_n^\Lambda\ \forall\Lambda$, we get:

$$\bar{\underline{L}}_n^M \;\equiv\; E[\bar{L}_n^M] \;\leq\; \bar{L}_n^\Lambda + a\cdot\sqrt{L_n^\Lambda \cdot k^\mu} + b\cdot k^\mu \quad \forall\mu \in \mathcal{M} \quad \forall\Lambda \tag{3}$$

So ignoring the conditions under which the bounds can be applied and the magnitude of the constants a and b, in the Bayesian framework instead of using the Bayes-optimal predictor Λ_ξ, one may use the PEA master algorithm M with same/similar performance guarantees.

Discussion. Our bound (3) represents a real challenge to Bayesian sequence prediction. Ignoring the constants a and b, the PEA master M has the same performance bound as the Bayes predictor Λ_ξ (2)$\Rightarrow$(3)$\widehat{=}$(1). Additionally, PEA has worst-case guarantees, which Bayes lacks. So it seems that PEA is superior to Bayes. The following issues are of interest to corroborate or to attenuate this statement. First, we only compared *bounds* on PEA and Bayes. It would be interesting to know something about the *actual* (practical or theoretical) relative performance of M and Λ_ξ. For instance the regrets are much better (finite) for smooth loss functions. Second, consider general $\mathcal{X}$, $\mathcal{Y}$, $\ell \in [0,1]$, and finite $\mathcal{E}$ with $k^e = \ln|\mathcal{E}|$. What is the optimal (minimal possible) constant a in bound (2)? In the static case $a = \sqrt{2}$ is optimal [Vov95] and achieved by the Hedge algorithm [FS97]. Moving from static to dynamic η_t typically costs an extra factor $\sqrt{2}$. Also, $a = 2$ in the Bayes bound (1). So we conjecture that there exists a PEA-type master (possibly Hedge) with $a = 2$, and this is the best achievable. Can $a = 2\sqrt{2}$ of FPL be improved? A necessary or at least helpful subproblem is to first generalize the existing bounds for WM-type masters to general $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{E}$, and $\ell \in [0,1]$, similarly to FPL. The Hedge algorithm is promising, since such static bounds already exist. Finally, can the PEA bound (2) be generalized to infinite $\mathcal{E}$ and general k^e in a clean way without the hierarchy trick used in [HP04]? Again, looking at the Bayes bound which works without a hierarchy trick, suggests a positive answer. Is it necessary to use an expert dependent η_t^e? Weaker bounds with $\sqrt{L_n}$ in (2) and (1) replaced by $\sqrt{n}$ are typically easier to prove [KV03], and hence the above questions may be approached by first answering them for $\sqrt{n}$.

References.

[ACBG02] P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

[CB97] N. Cesa-Bianchi et al. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[FS97] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[Han57] J. Hannan. Approximation to Bayes risk in repeated plays. In *Contributions to the Theory of Games 3*, pages 97–139. Princeton University Press, 1957.

[HP04] M. Hutter and J. Poland. Prediction with expert advice by following the perturbed leader for general weights. In *Proc. 15th International Conf. on Algorithmic Learning Theory (ALT-2004)*, Berlin, 2004. Springer.

[Hut03] M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003.

[KV03] A. Kalai and S. Vempala. Efficient algorithms for online decision. In *Proc. 16th Annual Conf. on Learning Theory (COLT-2003)*, Lecture Notes in Artificial Intelligence, pages 506–521, Berlin, 2003. Springer.

[LW89] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. In *30th Annual Symposium on Foundations of Computer Science*, pages 256–261, Research Triangle Park, North Carolina, 1989. IEEE.

[Vov90] V. G. Vovk. Aggregating strategies. In *Proc. 3rd Annual Workshop on Computational Learning Theory*, pages 371–383, Rochester, New York, 1990. ACM Press.

[Vov95] V. G. Vovk. A game of prediction with expert advice. In *Proc. 8th Annual Conf. on Computational Learning Theory*, pages 51–60. ACM Press, New York, NY, 1995.

[YEYS04] R. Yaroshinsky, R. El-Yaniv, and S. Seiden. How to better use expert advice. *Machine Learning*, 2004. to appear.