Fast rates for prediction with limited expert advice

El Mehdi Saad [1] & Gilles Blanchard [2]

[1] *LMO, université Paris-Saclay, 91400 Orsay.*
E-mail: el-mehdi.saad@universite-paris-saclay.fr
[2] *LMO, université Paris-Saclay, 91400 Orsay.*
E-mail: gilles.blanchard@universite-paris-saclay.fr

Résumé. Nous investiguons le problème de la minimisation de l'erreur de generalisation par rapport au meilleur expert dans une famille finie, sous la contrainte d'un accès limité à l'information. Nous présentons de nouveaux algorithmes permettant d'atteindre un excès de risque de l'ordre de $\mathcal{O}(1/T)$ (où T représente le nombre de données) avec grande probabilité, sous l'hypothèse que la fonction de perte est Lipschitz et fortement convexe.

Mots-clés. Apprentissage statistique, Apprentissage en-ligne, Apprentissage avec budget, Prédiction avec avis d'experts.

Abstract. We investigate the problem of minimizing the excess generalization error with respect to the best expert prediction in a finite family in the stochastic setting, under limited access to information. Assuming that the loss function is Lipschitz and strongly convex, we design novel algorithms achieving fast rates ($\mathcal{O}(1/T)$) with high probability.

Keywords. Statistical Learning, Online Learning, Budgeted Learning, Prediction with expert advice.

1 Introduction and main contributions

We consider a generic prediction problem in a stochastic setting: a target random variable Y taking values in $\mathcal{Y}$ is to be predicted by a user-determined forecast F, also modelled as a random variable, taking values in a closed convex subset $\mathcal{X}$ of $\mathbb{R}^d$. The mismatch between the two is measured via a loss function $l(F, Y)$. The quality of the agent's output is measured by its generalization risk

$$R(F) := \mathbb{E}\big[l(F, Y)\big].$$

To assist us in this task, the forecast or "advice" of a number of "experts" $(F_1, \ldots, F_K)$ (also modeled as random variables) can be requested. The agent's objective is to achieve a risk as close as possible to the risk of the best expert $R^* = \min_{i \in [\![K]\!]} R(F_i)$. We measure the performance of the user's forecast via its excess risk with respect to that best expert.

We investigate several scenarios for prediction with limited access to expert advice. Furthermore, our emphasis is on obtaining *fast convergence rates* guarantees on the excess risk (i.e. $O(1/T)$ or $O(1/C)$, where C is the total query budget). These are possible under a strong convexity assumption of the loss, specified below.

Assumption 1.1. *$\forall y \in \mathcal{Y}$: $x \in \mathcal{X} \subseteq \mathbb{R}^d \mapsto l(x, y)$ is L-Lipschitz and ρ-strongly convex.*

This assumption implies in particular that the loss range is bounded by $B := 8\frac{\rho^2}{L}$.
Our contributions are the following.

- As a preliminary, we revisit (Section 2) the *full information setting*, with no limitations on queries. We contribute a new algorithm that is both simpler than existing ones. Furthermore, for forecast we only need to consult 2 experts.
- We then investigate (Section 3) the *budgeted setting* where we have a total query budget constraint C for the training phase; we give precise efficiency guarantees on the number of training expert queries needed to achieve a given precision for forecast.
- Finally, we investigate the *m-query* setting where the agent is limited to $m \geq 2$ queries per training round. We give some lower bounds (Section 4) were we show that fast rates cannot be achieved if the agent is only allowed to consult one single expert per training round.

2 Full-information setting

We assume the output of all experts are observed for T rounds (in other words, T i.i.d. training examples), which is the full information or "batch" setting. We want to output a final prediction rule with prediction risk controlled with high probability over the training phase.

We start with putting forward an apparently new rule. The underlying principle will guide us to construct a budget efficient expert selection rule in the sequel.

Define $\hat{R}(F_i) := T^{-1}\sum_{t=1}^{T} l(F_{i,t}, Y_t)$ the empirical loss of expert i, and let $\hat{d}_{ij} := (T^{-1}\sum_{t=1}^{T}(F_{i,t} - F_{j,t})^2)^{\frac{1}{2}}$ the empirical L_2 distance between experts i and j over T rounds. Finally let $\alpha = \alpha(\delta) := (\log(4K\delta^{-1})/T)^{\frac{1}{2}}$, where $\delta \in (0, 1)$ is a fixed confidence parameter. Define

$$\Delta_{ij} := \hat{R}(F_j) - \hat{R}(F_i) - 6\alpha \max\left\{L\hat{d}_{ij}, B\alpha\right\}.$$

Consider the following set of experts:

$$S := \left\{j \in [\![K]\!] : \sup_{j \in [\![K]\!]} \Delta_{ij} \leq 0\right\}.$$

Our new full information setting rule is the following:

$$\text{choose } \bar{k} \in S \text{ arbitrarily } ; \quad \text{pick } \bar{j} \in \operatorname{Arg}\operatorname{Max}_{j \in S} \hat{d}_{\bar{k}j}; \quad \text{predict } \widehat{F} := \frac{1}{2}(F_{\bar{k}} + F_{\bar{j}}). \tag{1}$$

The theorem below establishes fast convergence rate for the excess risk of this rule:

Theorem 2.1. *If Assumption 1.1 holds and $\delta \in (0,1)$ is fixed, then for the prediction rule $\widehat{F}$ defined by (1), it holds with probability $1-3\delta$ over the training phase (c is an absolute constant):*

$$R(\widehat{F}) \leq R^* + cB\frac{\log(4K\delta^{-1})}{T}.$$

3 Global-budget setting

In this section, we consider the global budget setting. More precisely, given an a-priori defined budget C, at each round the decision-maker selects an arbitrary subset of experts and asks for their predictions. The player then pays a unit for each observed expert's advice. The game finishes when the budget is exhausted, at which point the player outputs a convex combination of experts for prediction.

We convert the batch rule defined in the full information setting to an "online" rule by performing the test $\Delta_{ji} > 0$ for each pair (i,j) after each allocation. This extension allows us to derive instance dependent bounds, which cover the rates obtained in the batch setting in the worst case.

Let $\mathcal{S}^* := \operatorname{Arg}\operatorname{Min}_{i \in [\![K]\!]} R(F_i)$ denote the set of optimal experts. For $i,j \in [\![K]\!]$, we denote by $d_{ij} := (\mathbb{E}[(F_i - F_j)^2])^{1/2}$ the L_2 distance between the experts F_i and F_j. For $i \in [\![K]\!]$, we introduce the following quantity:

$$\Lambda_i := \min_{i^* \in \mathcal{S}^*} \max\left\{ \frac{L^2 d_{ii^*}^2}{|R(F_i) - R(F_{i^*})|^2}; \frac{B}{R(F_i) - R(F_{i^*})} \right\}.$$

Define the following set of experts: $\mathcal{S}_\epsilon := \left\{ i \in [\![K]\!] : \Lambda_i > \frac{1}{\epsilon} \right\}$, and let $\mathcal{S}_\epsilon^c$ be its complementary.

Theorem 3.1. *(Instance dependent bound) Suppose Assumption 1.1 holds. Let $C \geq K$ denote the global budget on queries and denote $\hat{g}$ the output of the procedure described above. For any $\epsilon \geq 0$, if:*

$$C > 578 C_\epsilon \log\bigl(K\delta^{-1} C_\epsilon\bigr),$$

where

$$C_\epsilon := \sum_{i \in \mathcal{S}_\epsilon^c} \Lambda_i + |\mathcal{S}_\epsilon| \ \min\left\{ \frac{1}{\epsilon}; \Lambda^* \right\},$$

where $\Lambda^* := \max_{i:\Lambda_i < +\infty} \Lambda_i$, then, with probability at least $1-\delta$: $R(\hat{g}) \leq R^* + cB\epsilon$, where c is an absolute constant.

4 Local-budget setting

In this section, we suppose that the decision-maker is constrained to see only $m \geq 2$ experts' advice per round. We suppose that the horizon is unknown; when the game is halted, the player outputs a convex combination of the experts.

To circumvent the limitation of observing only m experts per round, in each round, we sample m experts out of S in a uniform way, where S is the set of non-eliminated experts. Then the tests $\Delta_{ji} \leq 0$ and $\Delta_{ij} \leq 0$ are performed for all sampled experts. If i or j fail the test, it is eliminated from S.

Finally, when the algorithm is halted, we apply the rule (1) to the set S. The following theorem gives theoretical guarantees on the excess risk of the output $\hat{g}$:

Theorem 4.1. *Suppose Assumption 1.1 holds. Let $T \geq 1$, if $m \geq 2$, then with probability at least $1 - \delta$:*

$$R(\hat{g}) \leq \min_{i \in [\![K]\!]} R_i + cB \, \frac{(K/m)^2 \log(2TK\delta^{-1})}{T},$$

where c is an absolute constant.

Finally, we show that if the learner is restricted to see only one expert feedback per round ($m = 1$), it is impossible to do better than an excess risk $\mathcal{O}\big(1/\sqrt{T}\big)$ in deviation.

Lemma 4.2. *Consider the squared loss function. For $K = 2$, and $m = 1$, for any $T > 0$, for any convex combination of the experts $\hat{g}$ output after T training rounds, there exists a probability distribution for experts $\{F_1, F_2\}$ and target variable Y such that with probability at least* 0.1*,*

$$\hat{R}_T(\hat{g}) - R^* \geq \frac{1}{2\sqrt{T}}.$$

References

[1] J.-Y. Audibert. Progressive mixture rules are deviation suboptimal / Supplemental "Proof of the optimality of the empirical star algorithm". In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, 2008.

[2] G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probability theory and related fields*, 145(3-4):591–613, 2009.

[3] W. S. Lee, P. Bartlett, and R. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.