Mixability in Statistical Learning

W75

Tim van Erven, Peter D. Grünwald, Mark D. Reid, Robert C. Williamson

Summary

To measure the quality of predictions we need a framework. Two standard ones, which seem quite different, are **statistical learning** and **sequential prediction**. Some relations between these two frameworks are known, but the theory to characterize fast rates of convergence is completely distinct. We bridge this gap by introducing the unifying concept of **stochastic mixability**, which jointly takes into account the loss function, the hypothesis class and the underlying distribution.

Statistical Learning

learning $(X_1,Y_1),\ldots,(X_n,Y_n) \overset{\text{iid}}{\sim} P^*$ rate

$$d(\hat f, f^*) = \mathbf{E}_{(X,Y)\sim P^*}[\ell(Y,\hat f(X)) - \ell(Y,f^*(X))] = O(n^{-?})$$

$$V(f,f^*) = \mathbf{E}_{(X,Y)\sim P^*}\left(\ell(Y,f(X)) - \ell(Y,f^*(X))\right)^2$$

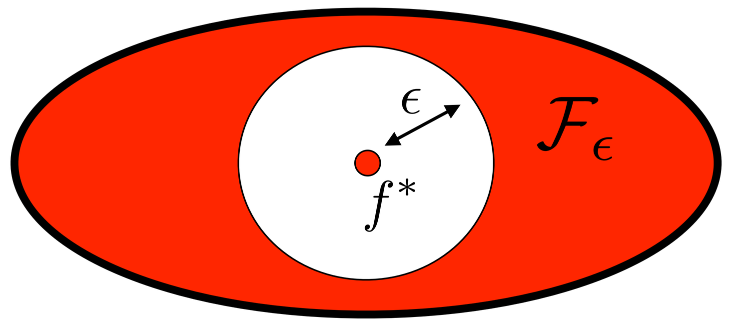
Fast rates of convergence $O(n^{-\kappa/(2\kappa-1)})$ are possible if the **margin condition**

$$c_0 V(f,f^*)^\kappa \leq d(f,f^*) \text{ for all } f \in \mathcal{F}$$

is satisfied with parameters $\kappa \geq 1, c_0 > 0$. [e.g. Tsybakov, 2004] (Smaller κ is better.)

Stochastic Mixability = Margin Condition

Thm [$\kappa = 1$]: Suppose the loss ℓ is bounded. Then $(\ell,\mathcal{F},P^*)$ is η-stochastically mixable if and only if there exists $c_0 > 0$ such that the margin condition is satisfied with $\kappa = 1$.



$$\mathcal{F}_\epsilon = \{f^*\} \cup \{f \in \mathcal{F} \mid d(f,f^*) \geq \epsilon\}$$

Thm [all $\kappa \geq 1$]: Suppose the loss ℓ is bounded. Then the margin condition is satisfied if and only if there exists a constant $C > 0$ such that, for all $\epsilon > 0$, $(\ell,\mathcal{F}_\epsilon,P^*)$ is η-stochastically mixable for $\eta = C\epsilon^{(\kappa-1)/\kappa}$.

Stochastic Mixability

Stochastic mixability provides the link between **fast rates** in both

- statistical learning (via the margin condition),
- sequential prediction (via exp-concavity).

Setting: Predict Y from X

Data: $(X_1,Y_1),\ldots,(X_n,Y_n)$
Hypothesis class (model): $\mathcal{F} = \{f\colon \mathcal{X} \to \mathcal{A}\}$
Loss: $\ell\colon \mathcal{Y}\times\mathcal{A} \to [0,\infty]$

Classification

$\mathcal{Y} = \{0,1\}, \quad \mathcal{A} = \{0,1\}$

$$\ell(y,a) = \begin{cases} 0 & \text{if } y = a \\ 1 & \text{if } y \neq a \end{cases}$$

Density Estimation

Forget about $\mathcal{X}$:

$\mathcal{F} \subset \mathcal{A} = \{$probability densities on $\mathcal{Y}\}$

$\ell(y,p) = -\log p(y)$

Definition

Let $\eta > 0$. Then $(\ell,\mathcal{F},P^*)$ is **η-stochastically mixable** if

$$\mathbf{E}_{(X,Y)\sim P^*}\left[\frac{e^{-\eta\ell(Y,f(X))}}{e^{-\eta\ell(Y,f^*(X))}}\right] \leq 1 \qquad \text{for all } f \in \mathcal{F}$$

where $f^* = \arg\min_{f\in\mathcal{F}} \mathbf{E}_{(X,Y)\sim P^*}[\ell(Y,f(X))]$ is the best f in the model.

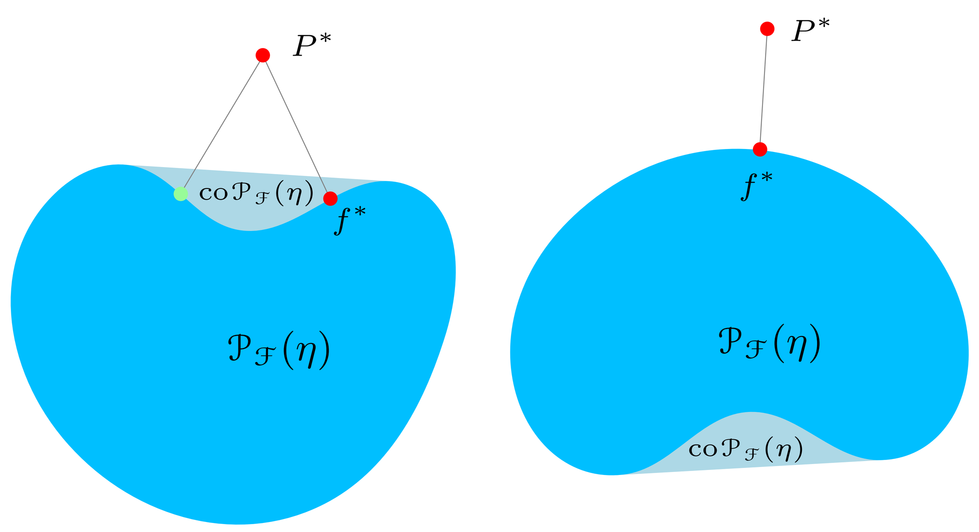
Remark: Special case already known in (Bayesian) density estimation under misspecification [Li, 1999, Kleijn and vdVaart, 2006]

Geometric Interpretation of the Margin Condition

Margin Condition = Stochastic mixability = convexity of the set

$$\mathcal{P}_\mathcal{F}(\eta) = \{e^{-\eta\ell(Y,f(X))} \mid f \in \mathcal{F}\}$$

relative to P^*.



Not stochastically mixable — Stochastically mixable

Example: in density estimation

$$\mathcal{P}_\mathcal{F}(1) = \mathcal{F} = \{\text{probability densities}\}$$

Sequential Prediction

For rounds $t = 1,\ldots,n$:

1. K experts predict $\hat f_t^1,\ldots,\hat f_t^K$
2. Predict (x_t,y_t) by choosing $\hat f_t$
3. Observe (x_t,y_t)

$$\text{Regret} = \frac{1}{n}\sum_{t=1}^n \ell(y_t,\hat f_t(x_t)) - \min_{k\in\{1,\ldots,K\}} \frac{1}{n}\sum_{t=1}^n \ell(y_t,\hat f_t^k(x_t))$$

Best possible worst-case regret is $O(1/n)$ if and only if the loss is **mixable ≈ exp-concave**. [Vovk, 1995]

A loss is **η-mixable** if for any distribution π on $\mathcal{A}$ there exists a prediction $a_\pi \in \mathcal{A}$ such that

$$\mathbf{E}_{A\sim\pi}\left[\frac{e^{-\eta\ell(y,A)}}{e^{-\eta\ell(y,a_\pi)}}\right] \leq 1 \qquad \text{for all } y.$$

Stochastic Mixability = Mixability (under conditions)

$$\mathcal{F}_{\text{full}} = \{\text{all functions from } \mathcal{X} \text{ to } \mathcal{A}\}$$

Thm: Suppose the loss ℓ is *proper* and $\mathcal{X}$ is discrete. Then ℓ is η-mixable if and only if $(\ell,\mathcal{F}_{\text{full}},P^*)$ is η-stochastically mixable for all P^*.

- Proper losses are e.g. 0/1-loss, log-loss, squared loss
- Theorem generalizes to other losses that satisfy two technical conditions

References

A. B. Tsybakov. *Optimal aggregation of classifiers in statistical learning.* The Annals of Statistics, 32(1):135–166, 2004.
V. Vovk. *A game of prediction with expert advice.* In Proceedings of the Eighth Annual Conference on Computational Learning Theory, pages 51–60. ACM, 1995.
Y. Kalnishkan and M. V. Vyugin. *The weak aggregating algorithm and weak mixability.* Journal of Computer and System Sciences, 74:1228–1244, 2008.
J. Li, *Estimation of Mixture Models* (PhD thesis), Yale University, 1999.
B.J.K. Kleijn, A.W. van der Vaart, *Misspecification in Infinite-Dimensional Bayesian Statistics*, The Annals of Statistics, 2006.