A randomized online learning algorithm for better variance control

Jean-Yves Audibert

ParisTech - Ecole des Ponts
CERTIS

Conference on Learning Theory, 2006

Outline

1. Motivation
   - The learning task
   - The progressive mixture rule
   - A striking sequential prediction result in least square regression
2. Contributions
   - The variance function
   - The algorithm and its risk bound
   - Application to general loss function
   - Application to least square loss

Outline

A standard learning framework...

- **Training data Z_1^n:** $Z_i = (X_i, Y_i) \quad i = 1, \dots, n \quad$ i.i.d. $\sim \mathbb{P}$
- **Prediction function:** $g : \mathcal{X} \to \mathcal{Y}$
- **Loss:** $L(Z, g)$
- **Risk:** $R(g) = \mathbb{E}_{\mathbb{P}(dZ)} L(Z, g)$
- **Model:**
  - $\mathcal{P} =$ the set of proba on $\mathcal{Z}$ in which we assume that $\mathbb{P}$ is
  - $\mathcal{G} =$ a set of prediction functions
- **Best prediction function in $\mathcal{G}$:** $\tilde{g} = \text{argmin}_{\mathcal{G}} R$

The $(L, \mathcal{P}, \mathcal{G})$-learning task:

Predict as well as $\tilde{g}$. More formally: find a mapping $Z_1^n \mapsto \hat{g}$ such that for any $\mathbb{P} \in \mathcal{P}$, we have

$$\mathbb{E}_{Z_1^n} R(\hat{g}) \leq R(\tilde{g}) + \textit{small term}$$

A standard learning framework...

- **Training data Z_1^n:** $Z_i = (X_i, Y_i) \quad i = 1, \ldots, n \quad$ i.i.d. $\sim \mathbb{P}$
- **Prediction function:** $g : \mathcal{X} \to \mathcal{Y}$
- **Loss:** $L(Z, g)$
- **Risk:** $R(g) = \mathbb{E}_{\mathbb{P}(dZ)} L(Z, g)$
- **Model:**
  - $\mathcal{P}$ = the set of proba on $\mathcal{Z}$ in which we assume that $\mathbb{P}$ is
  - $\mathcal{G}$ = a set of prediction functions
- **Best prediction function in $\mathcal{G}$:** $\tilde{g} = \mathrm{argmin}_{\mathcal{G}} R$

The $(L, \mathcal{P}, \mathcal{G})$-learning task:

Predict as well as $\tilde{g}$. More formally: find a mapping $Z_1^n \mapsto \hat{g}$ such that for any $\mathbb{P} \in \mathcal{P}$, we have

$$\mathbb{E}_{Z_1^n} R(\hat{g}) \leq R(\tilde{g}) + \textit{small term}$$

A standard learning framework...

- **Training data Z_1^n:** $Z_i = (X_i, Y_i) \quad i = 1, \ldots, n \quad$ i.i.d. $\sim \mathbb{P}$
- **Prediction function:** $g : \mathcal{X} \to \mathcal{Y}$
- **Loss:** $L(Z, g)$
- **Risk:** $R(g) = \mathbb{E}_{\mathbb{P}(dZ)} L(Z, g)$
- **Model:**
  - $\mathcal{P} =$ the set of proba on $\mathcal{Z}$ in which we assume that $\mathbb{P}$ is
  - $\mathcal{G} =$ a set of prediction functions
- **Best prediction function in $\mathcal{G}$:** $\tilde{g} = \text{argmin}_{\mathcal{G}} R$

The $(L, \mathcal{P}, \mathcal{G})$-learning task:

Predict as well as $\tilde{g}$. More formally: find a mapping $Z_1^n \mapsto \hat{g}$ such that for any $\mathbb{P} \in \mathcal{P}$, we have

$$\mathbb{E}_{Z_1^n} R(\hat{g}) \leq R(\tilde{g}) + \; C(\log|\mathcal{G}|)/n \quad \text{for } L(Z, g) = [Y - g(X)]^2$$

...however unusual properties

- To be “optimal”, we need to choose $\hat{g}$ outside the model $\mathcal{G}$.
- For least square loss (i.e. $L(Z, g) = [Y - g(X)]^2$), the only known optimal algorithm is the progressive mixture rule (see next slides)
- The proof is not based on bounds on the supremum of empirical processes

The progressive mixture rule
Notation

- **Cumulative loss of g up to time i:** $\Sigma_i(g) = \sum_{j=1}^{i} L(Z_j, g)$
- **Prior distribution on $\mathcal{G}$:** π
- **Gibbs distribution:** for any $h : \mathcal{G} \to \mathbb{R}$,

$$\pi_{-h}(dg) = \frac{e^{-h(g)}}{\mathbb{E}_{g' \sim \pi} e^{-h(g')}} \cdot \pi(dg) \propto e^{-h(g)} \cdot \pi(dg)$$

Key idea:

π_{-h} concentrates on the prediction functions for which h is minimum.

- **Typical example of Gibbs distribution:** $\pi_{-\lambda\Sigma_i}$ with $\lambda > 0$

The progressive mixture rule
Definition and property

Definition :

Let $\lambda > 0$. Predict according to $\hat{g} = \frac{1}{n+1} \sum_{i=0}^{n} \mathbb{E}_{\pi_{-\lambda\Sigma_i}(dg)} g$.

Property [Catoni (1999), Juditsky, Rigollet & Tsybakov (2005)]:

For the least square loss, under the assumptions

- the output has exponential moments
  (i.e. $\exists \alpha, M > 0 \quad \forall x \in \mathcal{X} \; E[e^{\alpha|Y|} | X = x] \leq M$)
- the functions of the model are uniformly bounded
  $\exists B > 0 \; \forall g \in \mathcal{G}, \|g\|_\infty \leq B$
- λ small enough, i.e. $\lambda \leq C(\alpha, M, B)$

$$\mathbb{E} R(\hat{g}) \leq R(\tilde{g}) + \frac{\log |\mathcal{G}|}{\lambda(n+1)}.$$

Sequential prediction framework

- $\mathcal{G}$ = set of prediction functions (or static experts)
- No probabilistic assumption on the data
- **Context:** At time i, you know $Z_1, \dots, Z_{i-1}$ and you have to give a prediction function $\hat{h}_i$, which will be only used to predict the output associated with X_i.
- **Target:** Predict as well as the best function in terms of cumulative loss:

$$\sum_{i=1}^n L(Z_i, \hat{h}_i) \leq \min_{g \in \mathcal{G}} \Sigma_n(g) + \textit{small term}$$

Sequential prediction in least square setting

Key idea [Vovk (1990), Haussler, Kivinen & Warmuth (1998)]:

Assume that $\mathcal{Y} = [-B; B]$ (i.e. bounded outputs). Let $\lambda = \frac{1}{2B^2}$. For any $i \in \{1, \dots, n\}$, let $\hat{h}_i$ be a prediction function such that

$$\forall z \in \mathcal{Z} \qquad L(z, \hat{h}_i) \leq -\frac{1}{\lambda} \log \mathbb{E}_{\pi_{-\lambda\Sigma_{i-1}}(dg)} e^{-\lambda L(z,g)}.$$

- $\hat{h}_i$ exists even if it has no simple explicit formula!

Theorem [Haussler, Kivinen & Warmuth (1998)]:

The cumulative loss on Z_1^n of the strategy in which the prediction at time i is done according to $\hat{h}_i$ is bounded with

$$\min_{g \in \mathcal{G}} \Sigma_n(g) + 2B^2 \log |\mathcal{G}|.$$

Theorem [Haussler, Kivinen & Warmuth (1998)]:

The strategy in which the prediction at time i is done according to $\hat{h}_i$ satisfies $\sum_{i=1}^{n+1} L(Z_i, \hat{h}_{i-1}) \leq \inf_{g \in \mathcal{G}} \Sigma_{n+1}(g) + 2B^2 \log|\mathcal{G}|$.

$\Downarrow$

Result

The algorithm predicting according to $\hat{g} = \frac{1}{n+1} \sum_{i=0}^{n} \hat{h}_i$ satisfies

$$\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + 2B^2 \frac{\log|\mathcal{G}|}{n+1}$$

- To be compared with
  $\mathbb{E}R(\text{progressive mixture rule}) \leq R(\tilde{g}) + C(\alpha, M, B) \frac{\log|\mathcal{G}|}{n+1}$,
- Worst case analysis leads to
  - optimal convergence rate for our learning task
  - even better constants when the output is bounded!

Theorem [Haussler, Kivinen & Warmuth (1998)]:

The strategy in which the prediction at time i is done according to $\hat{h}_i$ satisfies $\quad \sum_{i=1}^{n+1} L(Z_i, \hat{h}_{i-1}) \leq \inf_{g \in \mathcal{G}} \Sigma_{n+1}(g) + 2B^2 \log |\mathcal{G}|$.

$$\Downarrow$$

Result

The algorithm predicting according to $\hat{g} = \frac{1}{n+1} \sum_{i=0}^{n} \hat{h}_i$ satisfies

$$\mathbb{E} R(\hat{g}) \leq R(\tilde{g}) + 2B^2 \frac{\log |\mathcal{G}|}{n+1}$$

- To be compared with
  $$\mathbb{E} R(\text{progressive mixture rule}) \leq R(\tilde{g}) + C(\alpha, M, B) \frac{\log |\mathcal{G}|}{n+1},$$
- Worst case analysis leads to
  - optimal convergence rate for our learning task
  - even better constants when the output is bounded!

The new concept: the variance function

Variance function associated with the $(L, \mathcal{P}, \mathcal{G})$-learning task

Let $\bar{\mathcal{G}}$ be the set of all prediction functions (not only those in $\mathcal{G}$).
For any $\lambda > 0$, let $v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ be such that

$$\forall\, \rho \text{ proba on } \mathcal{G} \quad \exists\, \hat{\pi}(\rho) \text{ proba on } \bar{\mathcal{G}} \quad \forall \mathbb{P} \in \mathcal{P}$$
$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda\left[L(Z,g') - L(Z,g) - v_\lambda(Z,g,g')\right]} \leq 0.$$

The new concept: the variance function

Variance function associated with the $(L, \mathcal{P}, \mathcal{G})$-learning task

Let $\bar{\mathcal{G}}$ be the set of all prediction functions (not only those in $\mathcal{G}$).
For any $\lambda > 0$, let $v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ be such that

$$\forall \rho \text{ proba on } \mathcal{G} \quad \exists \hat{\pi}(\rho) \text{ proba on } \bar{\mathcal{G}} \quad \forall \mathbb{P} \in \mathcal{P}$$
$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda \left[L(Z,g') - L(Z,g) - v_\lambda(Z,g,g')\right]} \leq 0.$$

To be compared with

$$\log \mathbb{E}_{\mathbb{P}(dZ)} e^{\lambda[\mathbb{E}_{\mathbb{P}(dZ)} L(Z,g) - L(Z,g) - \phi(\lambda) \mathrm{Var}_{\mathbb{P}(dZ)} L(Z,g)]} \leq 0.$$

The new concept: the variance function

Variance function associated with the $(L, \mathcal{P}, \mathcal{G})$-learning task

Let $\bar{\mathcal{G}}$ be the set of all prediction functions (not only those in $\mathcal{G}$).
For any $\lambda > 0$, let $v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ be such that

$$\forall\, \rho \text{ proba on } \mathcal{G} \quad \exists\, \hat{\pi}(\rho) \text{ proba on } \bar{\mathcal{G}} \quad \forall \mathbb{P} \in \mathcal{P}$$
$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda\left[L(Z,g') - L(Z,g) - v_\lambda(Z,g,g')\right]} \leq 0.$$

Probabilistic version of Vovk, Haussler, Kivinen and Warmuth's condition:

$$\forall\, z \in \mathcal{Z} \qquad L(z, \hat{h}_i) \leq -\frac{1}{\lambda} \log \mathbb{E}_{\pi_{-\lambda\Sigma_i}(dg)} e^{-\lambda L(z,g)}.$$

$$\forall\, z \in \mathcal{Z} \qquad L(z, h_\rho) \leq -\frac{1}{\lambda} \log \mathbb{E}_{\rho(dg)} e^{-\lambda L(z,g)}$$

$$\forall \mathbb{P} \qquad \mathbb{E}_{\delta_{h_\rho}(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda[L(Z,g') - L(Z,g)]} \leq 0.$$

$$\Rightarrow v_\lambda \equiv 0 \quad \text{and} \quad \hat{\pi}(\rho) = \delta_{h_\rho}$$

The new concept: the variance function

Variance function associated with the $(L, \mathcal{P}, \mathcal{G})$-learning task

Let $\bar{\mathcal{G}}$ be the set of all prediction functions (not only those in $\mathcal{G}$).
For any $\lambda > 0$, let $v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ be such that

$$\forall\, \rho \text{ proba on } \mathcal{G} \quad \exists\, \hat{\pi}(\rho) \text{ proba on } \bar{\mathcal{G}} \quad \forall \mathbb{P} \in \mathcal{P}$$
$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda \left[L(Z,g') - L(Z,g) - v_\lambda(Z,g,g') \right]} \leq 0.$$

Probabilistic version of Vovk, Haussler, Kivinen and Warmuth's condition:

$$\forall\, z \in \mathcal{Z} \qquad L(z, \hat{h}_i) \leq -\frac{1}{\lambda} \log \mathbb{E}_{\pi_{-\lambda \Sigma_i}(dg)} e^{-\lambda L(z,g)}.$$
$$\forall\, z \in \mathcal{Z} \qquad L(z, h_\rho) \leq -\frac{1}{\lambda} \log \mathbb{E}_{\rho(dg)} e^{-\lambda L(z,g)}$$
$$\forall \mathbb{P} \qquad \mathbb{E}_{\delta_{h_\rho}(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda [L(Z,g') - L(Z,g)]} \leq 0.$$
$$\Rightarrow v_\lambda \equiv 0 \quad \text{and} \quad \hat{\pi}(\rho) = \delta_{h_\rho}$$

The new concept: the variance function

Variance function associated with the $(L, \mathcal{P}, \mathcal{G})$-learning task

Let $\bar{\mathcal{G}}$ be the set of all prediction functions (not only those in $\mathcal{G}$).
For any $\lambda > 0$, let $v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ be such that

$$\forall\, \rho \text{ proba on } \mathcal{G} \quad \exists\, \hat{\pi}(\rho) \text{ proba on } \bar{\mathcal{G}} \quad \forall \mathbb{P} \in \mathcal{P}$$
$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda \left[L(Z,g') - L(Z,g) - v_\lambda(Z,g,g') \right]} \leq 0.$$

Probabilistic version of Vovk, Haussler, Kivinen and Warmuth's condition:

$$\forall z \in \mathcal{Z} \qquad L(z, \hat{h}_i) \leq -\frac{1}{\lambda} \log \mathbb{E}_{\pi_{-\lambda\Sigma_i}(dg)} e^{-\lambda L(z,g)}.$$
$$\forall z \in \mathcal{Z} \qquad L(z, h_\rho) \leq -\frac{1}{\lambda} \log \mathbb{E}_{\rho(dg)} e^{-\lambda L(z,g)}$$
$$\forall \mathbb{P} \qquad \mathbb{E}_{\delta_{h_\rho}(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda [L(Z,g') - L(Z,g)]} \leq 0.$$
$$\Rightarrow v_\lambda \equiv 0 \quad \text{and} \quad \hat{\pi}(\rho) = \delta_{h_\rho}$$

The new concept: the variance function

Variance function associated with the $(L, \mathcal{P}, \mathcal{G})$-learning task

Let $\bar{\mathcal{G}}$ be the set of all prediction functions (not only those in $\mathcal{G}$).
For any $\lambda > 0$, let $v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ be such that

$$\forall\, \rho \text{ proba on } \mathcal{G} \quad \exists\, \hat{\pi}(\rho) \text{ proba on } \bar{\mathcal{G}} \quad \forall \mathbb{P} \in \mathcal{P}$$
$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda \left[L(Z,g') - L(Z,g) - v_\lambda(Z,g,g') \right]} \leq 0.$$

Probabilistic version of Vovk, Haussler, Kivinen and Warmuth's condition:

$$\forall\, z \in \mathcal{Z} \qquad L(z, \hat{h}_i) \leq -\tfrac{1}{\lambda} \log \mathbb{E}_{\pi_{-\lambda \Sigma_i}(dg)} e^{-\lambda L(z,g)}.$$
$$\forall\, z \in \mathcal{Z} \qquad L(z, h_\rho) \leq -\tfrac{1}{\lambda} \log \mathbb{E}_{\rho(dg)} e^{-\lambda L(z,g)}$$
$$\forall \mathbb{P} \qquad \mathbb{E}_{\delta_{h_\rho}(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda [L(Z,g') - L(Z,g)]} \leq 0.$$
$$\Rightarrow v_\lambda \equiv 0 \quad \text{and} \quad \hat{\pi}(\rho) = \delta_{h_\rho}$$

The new concept: the variance function

Variance function associated with the $(L, \mathcal{P}, \mathcal{G})$-learning task

Let $\bar{\mathcal{G}}$ be the set of all prediction functions (not only those in $\mathcal{G}$).
For any $\lambda > 0$, let $v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ be such that

$$\forall\, \rho \text{ proba on } \mathcal{G} \quad \exists\, \hat{\pi}(\rho) \text{ proba on } \bar{\mathcal{G}} \quad \forall \mathbb{P} \in \mathcal{P}$$
$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda \left[L(Z,g') - L(Z,g) - v_\lambda(Z,g,g') \right]} \leq 0.$$

Whatever L, $\mathcal{P}$ and $\mathcal{G}$ are, we can take

$$v_\lambda(z,g,g') = \frac{\lambda}{2} \left[L(z,g) - L(z,g') \right]^2 \qquad \text{and} \qquad \hat{\pi}(\rho) = \rho.$$

The algorithm based on the variance function

Generic Algorithm:

1. Let $\lambda > 0$. Let $S_0(g) = 0$ for any $g \in \mathcal{G}$.
   Define $\hat{\rho}_0 \triangleq \hat{\pi}(\pi)$ in the sense of the variance function definition.
   Draw a function $\hat{g}_0$ according to this distribution.
2. For any $i \in \{1, \dots, n\}$, iteratively define

   $$S_i(g) \triangleq S_{i-1}(g) + L(Z_i, g) + v_\lambda(Z_i, g, \hat{g}_{i-1}) \qquad \text{for any } g \in \mathcal{G}.$$

   and

   $$\hat{\rho}_i \triangleq \hat{\pi}(\pi_{-\lambda S_i})$$

   and draw a function $\hat{g}_i$ according to the distribution $\hat{\rho}_i$.
3. Predict with a function drawn according to the uniform distribution on $\{\hat{g}_0, \dots, \hat{g}_n\}$.

Its generalization error bound

Main theorem

Let π be uniform on $\mathcal{G}$ finite.
Let $\Delta_\lambda(g, g') \triangleq \mathbb{E}_{\mathbb{P}(dZ)} v_\lambda(Z, g, g')$ for $g \in G$ and $g' \in \bar{\mathcal{G}}$.
The expected risk of the generic algorithm satisfies

$$\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + \mathbb{E}\Delta_\lambda(\tilde{g}, \hat{g}) + \frac{\log|\mathcal{G}|}{\lambda(n+1)},$$

where $\mathbb{E}$ denotes the expectation w.r.t. the training data distribution and the randomizing distributions.

Symmetrization trick on prediction functions:

Let $z \in \mathcal{Z}$ and $\alpha(g', g) \triangleq \lambda[L(z, g') - L(z, g)]$. We have

$$\mathbb{E}_{\rho(dg')}\mathbb{E}_{\rho(dg)} e^{\alpha(g',g) - \frac{\alpha^2(g',g)}{2}} \leq 1$$

- Whatever L, $\mathcal{P}$ and $\mathcal{G}$ are, we can take

$$v_\lambda(z, g, g') = \frac{\lambda}{2}\big[L(z, g) - L(z, g')\big]^2 \qquad \text{and} \qquad \hat{\pi}(\rho) = \rho.$$

Corollary of the main theorem

Let $V(g, g') = \mathbb{E}_{\mathbb{P}(dZ)}\big\{[L(Z, g) - L(Z, g')]^2\big\}$. Our generic algorithm applied with $v_\lambda(Z, g, g') = \lambda[L(Z, g) - L(Z, g')]^2/2$ and $\hat{\pi}(\rho) = \rho$ satisfies

$$\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + \frac{\lambda}{2}\mathbb{E}V(\tilde{g}, \hat{g}) + \frac{\log|\mathcal{G}|}{\lambda(n+1)}$$

Symmetrization trick on prediction functions:

Let $z \in \mathcal{Z}$ and $\alpha(g', g) \triangleq \lambda[L(z, g') - L(z, g)]$. We have

$$\mathbb{E}_{\rho(dg')}\mathbb{E}_{\rho(dg)}\, e^{\alpha(g',g) - \frac{\alpha^2(g',g)}{2}} \leq 1$$

- Whatever L, $\mathcal{P}$ and $\mathcal{G}$ are, we can take

$$v_\lambda(z, g, g') = \frac{\lambda}{2}\big[L(z, g) - L(z, g')\big]^2 \qquad \text{and} \qquad \hat{\pi}(\rho) = \rho.$$

Corollary of the main theorem

Let $V(g, g') = \mathbb{E}_{\mathbb{P}(dZ)}\big\{[L(Z, g) - L(Z, g')]^2\big\}$. Our generic algorithm applied with $v_\lambda(Z, g, g') = \lambda[L(Z, g) - L(Z, g')]^2/2$ and $\hat{\pi}(\rho) = \rho$ satisfies

$$\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + \tfrac{\lambda}{2}\mathbb{E}V(\tilde{g}, \hat{g}) + \frac{\log|\mathcal{G}|}{\lambda(n+1)}$$

Making the bound more explicit

$$\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + \frac{\lambda}{2}\mathbb{E}V(\tilde{g},\hat{g}) + \frac{\log|\mathcal{G}|}{\lambda(n+1)}$$

Generalized Mammen and Tsybakov's assumption

There exist $0 \leq \gamma \leq 1$ and a prediction function g^* (not necessarily in $\mathcal{G}$) such that $V(g,g^*) \leq c[R(g) - R(g^*)]^\gamma$ for any $g \in \mathcal{G}$

$$\Downarrow$$

- When $\gamma = 1$,

$$\mathbb{E}R(\hat{g}) - R(g^*) \leq \frac{1+c\lambda}{1-c\lambda}\left[R(\tilde{g}) - R(g^*)\right] + \frac{\log|\mathcal{G}|}{(1-c\lambda)\lambda(n+1)}$$

  In particular, for $\lambda = 1/2c$, when g^* belongs to $\mathcal{G}$, we get $\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + \frac{4c\log|\mathcal{G}|}{n+1}$.

- When $\gamma < 1$, for any $0 < \beta < 1$ and for $\tilde{R} \triangleq R(\tilde{g}) - R(g^*)$,

$$\mathbb{E}R(\hat{g}) - R(g^*) \leq \left\{\frac{1}{\beta}\left([\tilde{R} + c\lambda\tilde{R}^\gamma] + \frac{\log|\mathcal{G}|}{\lambda(n+1)}\right)\right\} \vee \left(\frac{c\lambda}{1-\beta}\right)^{\frac{1}{1-\gamma}}.$$

Making the bound more explicit

$$\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + \frac{\lambda}{2}\mathbb{E}V(\tilde{g},\hat{g}) + \frac{\log|\mathcal{G}|}{\lambda(n+1)}$$

Generalized Mammen and Tsybakov's assumption

There exist $0 \leq \gamma \leq 1$ and a prediction function g^* (not necessarily in $\mathcal{G}$) such that $V(g,g^*) \leq c[R(g) - R(g^*)]^\gamma$ for any $g \in \mathcal{G}$

$$\Downarrow$$

- When $\gamma = 1$,

$$\mathbb{E}R(\hat{g}) - R(g^*) \leq \frac{1+c\lambda}{1-c\lambda}\left[R(\tilde{g}) - R(g^*)\right] + \frac{\log|\mathcal{G}|}{(1-c\lambda)\lambda(n+1)}$$

  In particular, for $\lambda = 1/2c$, when g^* belongs to $\mathcal{G}$, we get $\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + \frac{4c\log|\mathcal{G}|}{n+1}$.

- When $\gamma < 1$, for any $0 < \beta < 1$ and for $\tilde{R} \triangleq R(\tilde{g}) - R(g^*)$,

$$\mathbb{E}R(\hat{g}) - R(g^*) \leq \left\{\frac{1}{\beta}\left([\tilde{R} + c\lambda\tilde{R}^\gamma] + \frac{\log|\mathcal{G}|}{\lambda(n+1)}\right)\right\} \vee \left(\frac{c\lambda}{1-\beta}\right)^{\frac{1}{1-\gamma}}.$$

Comparaison with standard-style risk bounds

Recall $V(g,g') = \mathbb{E}_{\mathbb{P}(dZ)}\{[L(Z,g) - L(Z,g')]^2\}$.

- Symmetrization on the prediction functions space leads to $\hat{g}$ such that $\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + \frac{\lambda}{2}\mathbb{E}V(\tilde{g},\hat{g}) + \frac{\log|\mathcal{G}|}{\lambda(n+1)}$
- Vapnik-Cervonenkis' symmetrization (i.e. use of a second sample) leads to $\hat{g}_{\mathrm{ERM}}$ such that

$$\mathbb{E}R(\hat{g}_{\mathrm{ERM}}) \leq R(\tilde{g}) + \lambda\mathbb{E}V(\tilde{g},\hat{g}_{\mathrm{ERM}}) + \frac{\log(e|\mathcal{G}|)}{\lambda n} + \lambda\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}[L(Z_i,\tilde{g}) - L(Z_i,\hat{g}_{\mathrm{ERM}})]^2.$$

- Straightforward approach without symmetrizing but requiring

$$\sup_{g\in\mathcal{G},g'\in\mathcal{G}} |L(Z,g') - L(Z,g)| \leq A$$

leads to $\hat{g}_{\mathrm{ERM}}$ such that

$$\mathbb{E}R(\hat{g}_{\mathrm{ERM}}) \leq R(\tilde{g}) + \lambda\varphi(\lambda A)\mathbb{E}V(\tilde{g},\hat{g}_{\mathrm{ERM}}) + \frac{\log(e|\mathcal{G}|)}{\lambda n},$$

where $\varphi(t) \triangleq \frac{e^t-1-t}{t^2}$ and $\varphi(0) = \frac{1}{2}$ by continuity.

Comparaison with standard-style risk bounds

Recall $V(g,g') = \mathbb{E}_{\mathbb{P}(dZ)}\{[L(Z,g) - L(Z,g')]^2\}$.

- Symmetrization on the prediction functions space leads to $\hat{g}$ such that $\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + \frac{\lambda}{2}\mathbb{E}V(\tilde{g},\hat{g}) + \frac{\log|\mathcal{G}|}{\lambda(n+1)}$
- Vapnik-Cervonenkis' symmetrization (i.e. use of a second sample) leads to $\hat{g}_{\text{ERM}}$ such that

$$\mathbb{E}R(\hat{g}_{\text{ERM}}) \leq R(\tilde{g}) + \lambda\mathbb{E}V(\tilde{g},\hat{g}_{\text{ERM}}) + \frac{\log(e|\mathcal{G}|)}{\lambda n} + \lambda\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}[L(Z_i,\tilde{g}) - L(Z_i,\hat{g}_{\text{ERM}})]^2.$$

- Straightforward approach without symmetrizing but requiring

$$\sup_{g\in\mathcal{G},g'\in\mathcal{G}} |L(Z,g') - L(Z,g)| \leq A$$

leads to $\hat{g}_{\text{ERM}}$ such that

$$\mathbb{E}R(\hat{g}_{\text{ERM}}) \leq R(\tilde{g}) + \lambda\varphi(\lambda A)\mathbb{E}V(\tilde{g},\hat{g}_{\text{ERM}}) + \frac{\log(e|\mathcal{G}|)}{\lambda n},$$

where $\varphi(t) \triangleq \frac{e^t-1-t}{t^2}$ and $\varphi(0) = \frac{1}{2}$ by continuity.

Comparaison with standard-style risk bounds

Recall $V(g, g') = \mathbb{E}_{\mathbb{P}(dZ)}\{[L(Z, g) - L(Z, g')]^2\}$.

- Symmetrization on the prediction functions space leads to $\hat{g}$ such that $\mathbb{E}R(\hat{g}) \leq R(\tilde{g}) + \frac{\lambda}{2}\mathbb{E}V(\tilde{g}, \hat{g}) + \frac{\log|\mathcal{G}|}{\lambda(n+1)}$
- Vapnik-Cervonenkis' symmetrization (i.e. use of a second sample) leads to $\hat{g}_{\text{ERM}}$ such that

$$\mathbb{E}R(\hat{g}_{\text{ERM}}) \leq R(\tilde{g}) + \lambda\mathbb{E}V(\tilde{g}, \hat{g}_{\text{ERM}}) + \frac{\log(e|\mathcal{G}|)}{\lambda n} + \lambda\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}[L(Z_i, \tilde{g}) - L(Z_i, \hat{g}_{\text{ERM}})]^2.$$

- Straightforward approach without symmetrizing but requiring

$$\sup_{g\in\mathcal{G}, g'\in\mathcal{G}} |L(Z, g') - L(Z, g)| \leq A$$

leads to $\hat{g}_{\text{ERM}}$ such that

$$\mathbb{E}R(\hat{g}_{\text{ERM}}) \leq R(\tilde{g}) + \lambda\varphi(\lambda A)\mathbb{E}V(\tilde{g}, \hat{g}_{\text{ERM}}) + \frac{\log(e|\mathcal{G}|)}{\lambda n},$$

where $\varphi(t) \triangleq \frac{e^t - 1 - t}{t^2}$ and $\varphi(0) = \frac{1}{2}$ by continuity.

Application to least square loss

Study of the influence of the tail distribution

Framework:

- $L(Z,g) = [Y - g(X)]^2$
- $\exists B > 0 \quad \forall g \in \mathcal{G} \quad \|g\|_\infty \leq B$
- Predict as well as the best function in $\mathcal{G}$

Three cases:

- Bounded output : $|Y| \leq B$ a.s.
- Output with finite exponential moments :
  $$\exists \alpha, M > 0 \quad \forall x \in \mathcal{X} \quad E[e^{\alpha|Y|}|X = x] \leq M$$
- Output with finite moments :
  $$\mathbb{E}|Y|^s \leq A \qquad \text{for some } s \geq 2 \text{ and } A > 0$$

Bounded output : $|Y| \leq B$ a.s.

The variance function (recall):

$v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ is s.t. $\forall \rho$ proba on $\mathcal{G}, \exists \hat{\pi}(\rho)$ proba on $\bar{\mathcal{G}}, \forall \mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda \left[L(Z,g') - L(Z,g) - v_\lambda(Z,g,g') \right]} \leq 0.$$

Theorem

One can choose $v_{1/(2B^2)} \equiv 0$. The corresponding generic algorithm satisfies

$$R(\hat{g}) \leq R(\tilde{g}) + 2B^2 \frac{\log |\mathcal{G}|}{n+1}$$

$v_{1/(2B^2)}$ can be associated with $\hat{\pi}(\rho) = \delta_{h_\rho}$, where $h_\rho \in \bar{\mathcal{G}}$ is taken s.t.

$$\forall (x,y) \in \mathcal{Z} \quad [y - h_\rho(x)]^2 \leq -2B^2 \log \mathbb{E}_{\rho(dg)} e^{-[y-g(x)]^2/(2B^2)}.$$

Bounded output : $|Y| \leq B$ a.s.

The variance function (recall):

$v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ is s.t. $\forall \rho$ proba on $\mathcal{G}, \exists \hat{\pi}(\rho)$ proba on $\bar{\mathcal{G}}, \forall \mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda \left[L(Z,g') - L(Z,g) - v_\lambda(Z,g,g')\right]} \leq 0.$$

Theorem

One can choose $v_{1/(2B^2)} \equiv 0$. The corresponding generic algorithm satisfies

$$R(\hat{g}) \leq R(\tilde{g}) + 2B^2 \frac{\log |\mathcal{G}|}{n+1}$$

$v_{1/(2B^2)}$ can be associated with $\hat{\pi}(\rho) = \delta_{h_\rho}$, where $h_\rho \in \bar{\mathcal{G}}$ is taken s.t.

$$\forall (x,y) \in \mathcal{Z} \quad [y - h_\rho(x)]^2 \leq -2B^2 \log \mathbb{E}_{\rho(dg)} e^{-[y-g(x)]^2/(2B^2)}.$$

Output with finite exponential moments: $\exists \alpha, M > 0 \quad \forall x \in \mathcal{X} \quad E[e^{\alpha|Y|}|X = x] \leq M$

The variance function (recall):

$v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ is s.t. $\forall\, \rho$ proba on $\mathcal{G}, \exists\, \hat{\pi}(\rho)$ proba on $\bar{\mathcal{G}}, \forall \mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}_{\hat{\pi}(\rho)(dg')}\mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda\left[L(Z,g')-L(Z,g)-v_\lambda(Z,g,g')\right]} \leq 0.$$

Theorem

For an appropriate $\lambda = C(\alpha, M, B)$, we can choose $v_\lambda \equiv 0$.
The corresponding generic algorithm satisfies

$$R(\hat{g}) \leq R(\tilde{g}) + \frac{1}{\lambda}\frac{\log|\mathcal{G}|}{n+1}$$

v_λ can be associated with $\hat{\pi}(\rho) = \delta_{\mathbb{E}_{\rho(dg)}g}$.

Output with finite exponential moments: $\exists \alpha, M > 0 \quad \forall x \in \mathcal{X} \quad E[e^{\alpha |Y|} | X = x] \leq M$

The variance function (recall):

$v_\lambda : \mathcal{Z} \times \mathcal{G} \times \bar{\mathcal{G}} \to \mathbb{R}$ is s.t. $\forall \rho$ proba on $\mathcal{G}, \exists \hat{\pi}(\rho)$ proba on $\bar{\mathcal{G}}, \forall \mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}_{\hat{\pi}(\rho)(dg')} \mathbb{E}_{\mathbb{P}(dZ)} \log \mathbb{E}_{\rho(dg)} e^{\lambda \left[L(Z,g') - L(Z,g) - v_\lambda(Z,g,g') \right]} \leq 0.$$

Theorem

For an appropriate $\lambda = C(\alpha, M, B)$, we can choose $v_\lambda \equiv 0$. The corresponding generic algorithm satisfies

$$R(\hat{g}) \leq R(\tilde{g}) + \frac{1}{\lambda} \frac{\log |\mathcal{G}|}{n+1}$$

v_λ can be associated with $\hat{\pi}(\rho) = \delta_{\mathbb{E}_{\rho(dg)} g}$.

Output with finite moments: $\mathbb{E}|Y|^s \leq A$ for some $s \geq 2$ and $A > 0$

Theorem

Let $N = \frac{n+1}{\log|\mathcal{G}|}$. For $\lambda = \frac{C}{B^2} N^{-\frac{2}{s+2}}$, we can choose

$$v_\lambda(z, g, g') = C\Big[B|y|\mathbf{1}_{|y| \geq CBN^{\frac{2}{s+2}}} + N^{-\frac{2}{s+2}} y^2 \mathbf{1}_{CBN^{\frac{1}{s+2}} \leq |y| < CBN^{\frac{2}{s+2}}}\Big]$$

The corresponding generic algorithm satisfies

$$R(\hat{g}) \leq R(\tilde{g}) + CB^2 N^{-\frac{s}{s+2}}.$$

v_λ can be associated with $\hat{\pi}(\rho) = \delta_{\mathbb{E}_{\rho(dg)} g}$.

Output with finite moments: $\mathbb{E}|Y|^s \le A$ for some $s \ge 2$ and $A > 0$

Theorem

Let $N = \frac{n+1}{\log|\mathcal{G}|}$. For $\lambda = \frac{C}{B^2} N^{-\frac{2}{s+2}}$, we can choose

$$v_\lambda(z, g, g') = C\left[B|y|\mathbf{1}_{|y| \ge CBN^{\frac{2}{s+2}}} + N^{-\frac{2}{s+2}} y^2 \mathbf{1}_{CBN^{\frac{1}{s+2}} \le |y| < CBN^{\frac{2}{s+2}}}\right]$$

The corresponding generic algorithm satisfies

$$R(\hat{g}) \le R(\tilde{g}) + CB^2 N^{-\frac{s}{s+2}}.$$

v_λ can be associated with $\hat{\pi}(\rho) = \delta_{\mathbb{E}_{\rho(dg)} g}$.

Conclusion

- Define the concept of variance function
- Obtain a randomized algorithm that
  - allows to recover recent model selection type results from Juditsky, Rigollet and Tsybakov (2005)
  - benefits from worst-case analysis type arguments
- Propose a new symmetrization trick on the prediction function space that improves
  - a standard-style statistical bound
  - bounds in heavy noise setting

More details in ...

- D. Haussler, J. Kivinen and M. K. Warmuth,
  Sequential prediction of individual sequences under general loss functions,
  IEEE Trans. on Information Theory, 44(5):1906–1925, 1998.
- J. Kivinen and M. K. Warmuth,
  Averaging Expert Predictions,
  Lecture Notes in Computer Science, 1572:153–167, 1999.
- A. Juditsky, P. Rigollet and A. B. Tsybakov,
  Learning by mirror averaging,
  Technical report available from ArXiv website, 2005.
- J.-Y. Audibert,
  Model selection type aggregation with better variance control,
  Technical report available from my webpage, 2006.