Learning by Mirror Averaging

A. Judisky, P. Rigollet and A. B. Tsybakov
The Annals of Statistics

July 5, 2018

Abstract

- Construct a new estimator, called aggregate
- The aggregate consist of a simple recursive procedure which solves an auxiliary stochastic linear programming problem
- The aggregate satisfies sharp oracle inequalities under some general assumptions
- The results are applied to several problems - regression, classification, density estimation

Terminologies in Model selection

- Given a collection of M estimators, construct a new estimator which is nearly as good as the best among them w.r.t. a given risk criterion

- $(\mathcal{Z}, \mathfrak{F})$: measurable space (data)
- Θ is a M-dim simplex (set of a weight for each estimator)

$$\Theta = \left\{ \theta \in \mathbb{R}^M : \sum_{j=1}^{M} \theta^{(j)} = 1, \theta^{(j)} \geq 0, j = 1, \cdots, M \right\}$$

- Z : r.v with values in $\mathcal{Z}$, P : distribution of Z, E : corresponding expectation
- Observed n i.i.d. r.v.s : $Z_1, \cdots, Z_n \sim P$
  P_n, E_n : joint distribution and the corresponding expactation of $Z_1, \cdots, Z_n$

Model selection

- $Q : \mathcal{Z} \times \Theta \to \mathbb{R}$: measurable function (loss)
- Average risk function corresponding to Q :

$$A(\theta) = EQ(Z, \theta)$$

- **Aim** : "mimic the oracle" $\min_j A(e_j)$ (e_j : j-th coordinate unit vector in $\mathbb{R}^M$)
  Construct an estimator $\tilde{\theta}_n$ m'sble w.r.t. $Z_1, \cdots, Z_n$ s.t.

$$\begin{aligned} E_n A(\tilde{\theta}_n) &\le \min_{1 \le j \le M} A(e_j) + \Delta_{n,M} \\ &= \min_{\theta \in \{e_1, \cdots, e_M\}} A(\theta) + \Delta_{n,M} \end{aligned} \tag{1}$$

- $\Delta_{n,M}$: remaindeer term, should be as small as possible

Example

- $H = (h_1, \cdots, h_M)^\top$: a vector of preliminary estimators constructed from a training sample
  - supposed to be frozen in this thesis (each h_j is a fixed function)

- $Q(z, \theta) = l(z, \theta^\top H)$ for some loss function l

Goal

- Under some assumption, the smallest possible value of $\Delta_{n,M}$ in a minimax sense has the form

$$\Delta_{n,M} = \frac{C \log M}{n} \tag{2}$$

with some constant $C > 0$

- Constuct $\tilde{\theta}_n$ satisfies (1), (2)
  - Selector cannot achieve (2) - "Model Mixture" rather than Model selection
  - construct $\tilde{\theta}_n$ with Mirror algorithm
  - $\tilde{\theta}_n$ satisfies (1), (2)
  - Some examples

Suboptimality of selectors

- $Q(z,\theta) = \frac{1}{2}\theta^\top\theta - z^\top\theta, z \in \mathbb{R}^M$: squared loss
- P^k : M dimension gaussian distribution w/ mean $\frac{\sigma}{2}\sqrt{\frac{\log M}{n}}e_k$ and the covariance matrix $\sigma^2 I$

Proposition 2.1. Let Q be the squared loss function. Assume that we observe i.i.d. random vectors $Z_1, \cdots, Z_n$ with the same distribution as Z. Denote by E_n^k the expectation w.r.t. the sample $Z_1, \cdots, Z_n$ when Z has distribution P_k. Then there exists an absolute constant $c > 0$ s.t.

$$\inf_{T_n} \sup_{k=1,\cdots,M} \left\{ E_n^k\left[A_k(T_n)\right] - \min_{1\le j\le M} A_k(e_j) \right\} \ge c\sigma\sqrt{\frac{\log M}{n}} \tag{3}$$

where the infimum is taken over all the selectors T_n

Mirror algorithm

- The name *mirror averaging* reflect the fact that the algorithm does a s.g.d. in the dual space with further "mirroring" to the primal space and averaging.
- If A is convex, then we can bound it from above by a linear function:

$$A(\theta) \leq \sum_{j=1}^{M} \theta^{(j)} A(e_j) \triangleq \tilde{A}(\theta)$$

where $\tilde{A}(\theta) = E\tilde{Q}(Z, \theta)$ with

$$\tilde{Q}(Z, \theta) \triangleq \theta^{\top} u(Z), u(Z) \triangleq (Q(Z, e_1), \cdots, Q(Z, e_M))^{\top}$$

Then $\tilde{A}(e_j) = A(e_j)$ for all j and $\min_{\theta \in \Theta} \tilde{A}(\theta) = \min_{1 \leq j \leq M} A(e_j)$ since Θ is a simplex.

Mirror algorithm

- For $\beta > 0$, define the function $W_\beta : \mathbb{R}^M \to \mathbb{R}$ by

$$W_\beta(z) \triangleq \beta \log \left(\frac{1}{M} \sum_{j=1}^{M} e^{-z^{(j)}/\beta} \right) \tag{4}$$

- The gradient of W_β is given by

$$\nabla W_\beta(z) = \left[-\frac{e^{-z^{(j)}/\beta}}{\sum_{k=1}^{M} e^{-z^{(k)}/\beta}} \right]_{j=1}^{M} \in \Theta$$

- Let $u_i = u(Z_i)$. Then $u_i = \nabla_\theta \tilde{Q}(Z_i, \theta)$.

Mirror algorithm

- Fix the initial calues $\theta_0 \in \Theta$ and $\zeta_0 \in \mathbb{R}^M$
- For $i = 1, \cdots, n-1$, do the recursive update

$$\begin{aligned} \zeta_i &= \zeta_{i-1} + u_i \\ \theta_i &= -\nabla W_\beta(\zeta_i) \end{aligned} \tag{5}$$

- Output at iteration n the average

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_{i-1} \tag{6}$$

- The "mirroring" function ∇W_β maps the variables $\zeta_i \in (\mathbb{R}^M, l_\infty)$ to (Θ, l_1)
- W_β in (4) is not the only possible choice.

Main results

Let Q_1 be the function on $\mathcal{Z} \times \Theta \times \Theta$ defined by $Q_1(z, \theta, \theta') = Q(z, \theta) - Q(z, \theta')$ for all $z \in \mathcal{Z}$ and all $\theta, \theta' \in \Theta$.

Theorem 4.1. Assume that Q_1 can be decomposed into the sum of two functinos $Q_1 = Q_2 + Q_3$ such that:

- The mapping $\theta \to -Q_2(z, \theta, \theta')/\beta$ is exponentially concave on the simplex Θ, for all $z \in \mathcal{Z}, \theta' \in \Theta$, and $Q_2(z, \theta, \theta) = 0$ for all $z \in \mathcal{Z}, \theta \in \Theta$
- There exists a function R on $\mathcal{Z}$ integrable with respect to P and such that $Q_3(z, \theta, \theta) \leq R(z)$, for all $z \in \mathcal{Z}, \theta, \theta' \in \Theta$.

Then the aggregate $\hat{\theta}_n$ satisfies, for any $M \geq 2, n \geq 1$, the following oracle inequality:

$$E_{n-1} A(\hat{\theta}_n) \leq \min_{1 \leq j \leq M} A(e_j) + \frac{\beta log M}{n} + E[R(Z)].$$

Main results

Theorem 4.2. Assume that for some $\beta > 0$ there exists a Borel function $\Psi_\beta : \Theta \times \Theta \to \mathbb{R}_+$ such that the mapping $\theta \to \Psi_\beta(\theta, \theta')$ is concave on the simplex Θ for any fixed $\theta \in \Theta$, $\Psi_\beta(\theta, \theta) = 1$ and $E \exp(Q_1(Z, \theta, \theta)/\beta) \leq \Psi_\beta(\theta, \theta')$ for all $\theta, \theta' \in \Theta$. Then the aggregate $\hat{\theta}_n$ satisfies, for any $M \geq 2, n \geq 1$, the following oracle inequality:

$$E_{n-1} A(\hat{\theta}_n) \leq \min_{1 \leq j \leq M} A(e_j) + \frac{\beta \log M}{n} + E[R(Z)].$$

regression model

Corrollary 5.1. Consider the regression model $Y = f(X) + \xi$ where $X \in \mathcal{X}, Y \in \mathbb{R}, f : \mathcal{X} \to \mathbb{R}$ and $\xi = Y - f(X)$ is a real-valued random variable satisfying $E(\xi|X) = 0$. Assume also that $E(Y^2) < \infty$ and $||f_j||_\infty \leq L, j = 1, \cdots, M$, for some finite constant $L > 0$. Then for any positive constants $B \geq (4L+2)^{-2}, LB < b < 1/4$ and any $\beta \geq (b/B)^2$, the aggregate estimator $\tilde{f}_n(x) = \hat{\theta}_n^\top H(x), x \in \mathcal{X}$, where $\hat{\theta}_n$ is obtained by the mirror averaging algorithm, satisfies

$$E_{n-1}||\hat{f}_n - f||^2_{2,P_X} \leq \min_{1 \leq j \leq M} ||f_j - f||^2_{2,P_X} + \frac{\beta \log M}{n} + E[R_\beta(Y)] \qquad (7)$$

where

$$R_\beta(Y) = 4L|y|\mathbb{I}_{\{|y| \geq B\beta\}} + \frac{4L^2 y^2}{B\beta}\mathbb{I}_{\{b\sqrt{\beta} < |y| < B\beta\}}$$

- If $|Y| \leq L_0$ and $\beta > 16L_0^2, R_\beta(Y) \equiv 0$

regression model

Corrollary 5.6. Consider the regression model $Y = f(X) + \xi$ where $X \in \mathcal{X}, Y \in \mathbb{R}, f : \mathcal{X} \to \mathbb{R}$ and, conditionally on X, the random variable $\xi = Y - f(X)$ is Gaussian with zero mena and variance bounded by σ^2. Assume that $||f - f_j||_\infty \leq \tilde{L}$, for some finite constant $\tilde{L} > 0$. Then for any $\beta \geq 2\sigma^2 + 2\tilde{L}^2$, the aggregate estimator $\tilde{f}_n(x) = \hat{\theta}_n^\top H(x), x \in \mathcal{X}$, where $\hat{\theta}_n$ is obtained by the mirror averaging algorithm, satisfies

$$E_{n-1}||\hat{f}_n - f||^2_{2,P_X} \leq \min_{1 \leq j \leq M} ||f_j - f||^2_{2,P_X} \tag{8}$$

where

$$R_\beta(Y) = 4L|y|\mathbb{I}_{\{|y| \geq B\beta\}} + \frac{4L^2y^2}{B\beta}\mathbb{I}_{\{b\sqrt{\beta} < |y| < B\beta\}}$$