Competitive online generalized linear regression under square loss

Fedor Zhdanov

Computer Learning Research Centre
Royal Holloway, University of London

Notation

- We predict events $\omega_t \in \Omega = [0, 1]$.
- Our predictions are $\gamma_t \in \Gamma = [0, 1]$.
- The quality of predictions is measured by the square loss function $(\gamma_t - \omega_t)^2$.
- Define class of experts Θ such that predictions of experts $\xi^\theta(x) \in \mathbb{R}$ are defined for every $x \in \mathbf{X} \subseteq \mathbb{R}^n$.

Protocol

$L_t^\theta := 0$ for all $\theta \in \Theta$
$L_t := 0$ for the algorithm
FOR $t = 1, 2, \ldots$
    Input vector $x_t \in \mathbf{X}$
    Algorithm predicts $\gamma_t \in [0, 1]$
    Reality announces the actual outcome $\omega_t \in [0, 1]$

    Experts $L_t^\theta := L_{t-1}^\theta + (\xi^\theta(x_t) - \omega_t)^2$, $\theta \in \Theta$
    Learner $L_t := L_{t-1} + (\gamma_t - \omega_t)$
END FOR
Goal: provide for any (for the best) $\theta \in \Theta$

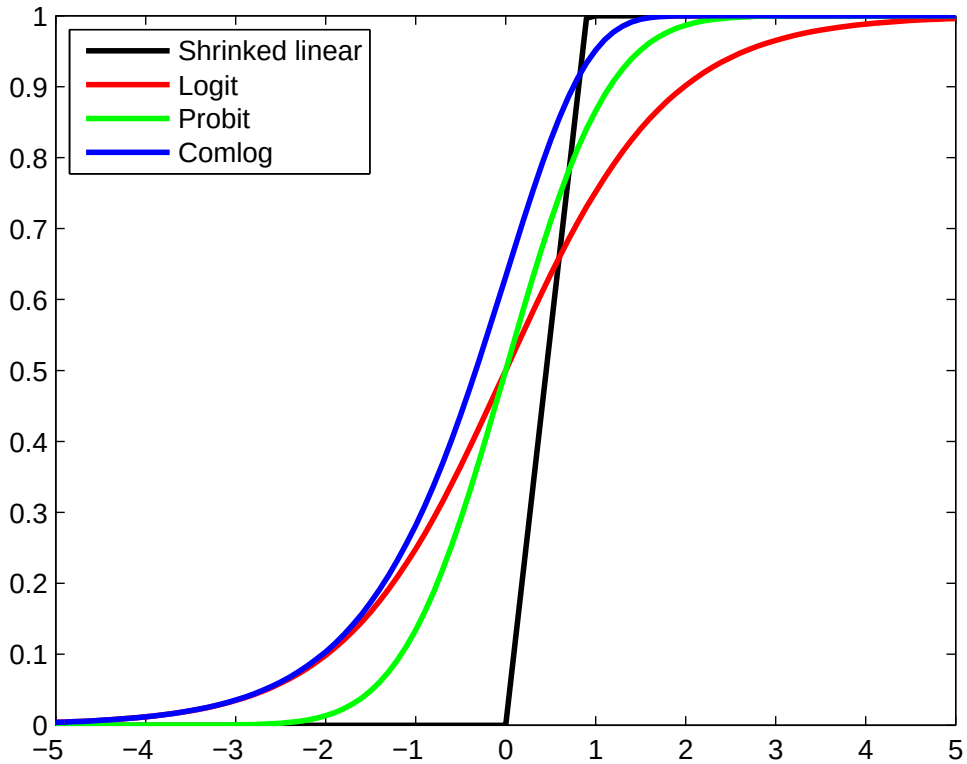
$$L_T \leq L_T^\theta + R_T$$

at any step T with small regret term $R_T = o(T)$.

Generalized linear models

Classes of experts with $\xi^\theta(x) = \sigma(\theta' x)$ for $\theta \in \Theta = \mathbb{R}^n$:

- Linear $\sigma(z) = z$
- Logistic $\sigma(z) = \frac{1}{1+e^{-z}}$
- Probit $\sigma(z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-v^2/2} dv$
- Complementary log-log $\sigma(z) = 1 - \exp(-\exp(z))$



Square loss! $\sum_{t=1}^{T} (\sigma(\theta' x_t) - \omega_t)^2 + a\|\theta\|^2$ is not necessarily convex in θ.

Aggregating Algorithm

Initialize experts' weights $P_0(d\theta)$.

Find suitable learning rate parameter: $\eta \in (0, 2]$.

FOR $t = 1, 2, \ldots$:

Read experts' predictions ξ_t^θ.

Calculate average for all ω:

$$g_t(\omega) = -\frac{1}{\eta} \ln \int_\Theta e^{-\eta(\xi_t^\theta - \omega)^2} P_{t-1}^*(d\theta).$$

Predict $\gamma_t = S(g_t)$.

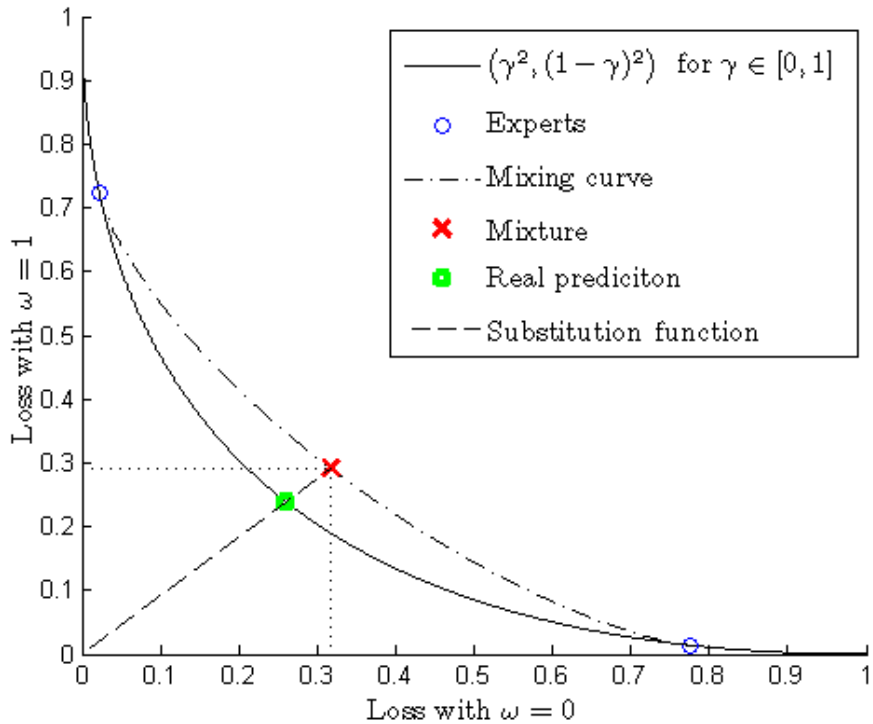
Read the actual outcome ω_t.

Update and normalize weights:

$$P_t(d\theta) = e^{-\eta(\xi_t^\theta - \omega_t)^2} P_{t-1}(d\theta)$$

$$P_t^*(d\theta) = P_t(d\theta) / \int_\Theta P_t(d\theta)$$

END FOR.



Loss bounds

Require:

$$b(u,z) := \left(\frac{d\sigma(z)}{dz}\right)^2 + (\sigma(z) - u)\frac{d^2\sigma(z)}{dz^2}$$

is uniformly bounded: $b := \sup_{u\in[0,1],z\in\mathbb{R}} |b(u,z)| < \infty$. We have $b = 1$ for linear models, $b \le \frac{5}{64}$ for logistic models, $b \le \frac{25}{128}$ for probit models, $b \le \frac{17}{64}$ for comlog models.

Theorem

Let $a > 0$. There exists a prediction algorithm such that for any T, every sequence $(x_1, y_1, \ldots, x_T, y_T)$ such that $\|x_t\|_\infty \le X$, and every $\theta \in \mathbb{R}^n$, the cumulative loss L_T satisfies

$$L_T \le L_T^\theta + a\|\theta\|^2 + \frac{n}{4}\ln\left(1 + \frac{bX^2T}{a}\right).$$

Algorithm

Predicts $\gamma_t = \frac{1}{2} - G_t(1) + G_t(0)$ for

$$G_t(\omega) = -\frac{1}{2} \ln \int_\Theta e^{-2\left((\sigma(\theta' x_T)-\omega)^2 + \sum_{t=1}^{T-1}(\sigma(\theta' x_t)-\omega_t)^2 + a\|\theta\|^2\right)} d\theta.$$

The integral is calculated using Markov Chain Monte Carlo: Metropolis sampling.
To sample from $\mathcal{P}$ with density $f_{\mathcal{P}}(\theta)$ over Θ:

- Take initial $\theta^0 \in \Theta$ and Gaussian proposal distribution $N(0, \sigma^2)$.
- $\theta^i = \theta^{i-1} + \zeta^i, \;\; \zeta^i \sim N(0, \sigma^2)$
- θ^i is accepted with the probability $\min\left(1, \frac{f_{\mathcal{P}}(\theta^i)}{f_{\mathcal{P}}(\theta^{i-1})}\right)$. Thus move closer to the maximum of $f_{\mathcal{P}}$.

Experiments

- Daily maximum of 1 hour (ozone1) ozone concentration and of the average over 8 consecutive hours (ozone8) in Texas, USA, 1998-2004.
- Two ways of prediction: online (retrain every day) or incremental batch (train on all previous years, predict for the next year).
- Best parameters are found on the first year.

Table: Average MSEs of different algorithms over the last 6 years of ozone1/ozone8.

Algorithm	MSE online	MSE batch
AAGLM	**8.2159**/15.3288	**8.2163**/**15.7552**
SVM rbf	9.7607/**14.8806**	9.4497/15.9679
SVM linear	8.8624/16.1532	8.8500/16.5264
Comlog	14.4578/24.2686	13.8096/27.9474
Zeros	9.6667/21.3333	9.6667/21.3333

Thank you

Thank you

Proof of competitiveness

For $\eta \in (0, 2]$ for any t there exists function S such that for $\gamma_t = S(g_t)$

$$(\gamma_t - \omega)^2 \leq g_t(\omega)$$

for all $\omega \in \Omega$.

Lemma (Vovk, 2001, Lemma 1)

For any learning rate $\eta > 0$, initial distribution P_0, and T,

$$\sum_{t=1}^{T} g_t(\omega_t) = -\frac{1}{\eta} \ln \int_{\Theta} e^{-\eta L_T^{\theta}} P_0(d\theta),$$

where L_T^{θ} is the cumulative loss of the expert θ at the step T.

Then

$$L_T(AA) \leq -\frac{1}{\eta} \ln \int_{\Theta} e^{-\eta L_T^{\theta}} P_0(d\theta) \leq L_T^{\theta} + R_T.$$

Large classes of experts

Recall:

$$L_T \le -\frac{1}{\eta} \ln \int_\Theta e^{-\eta L_T^\theta} P_0(d\theta).$$

Thus

$$L_T \le L_T^{\theta_0} + a\|\theta\|_0^2 - \frac{1}{\eta} \ln \int_\Theta e^{-\eta \theta' (aI + b\sum_{t=1}^T x_t x_t')\theta} d\theta.$$