Logistic Regression: The Importance of Being Improper

Dylan Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan

djfoster@cs.cornell.edu, satyenkale@google.com, haipengl@usc.edu, mohri@cs.nyu.edu, sridharan@cs.cornell.edu

Online Logistic Regression

For $t = 1, \ldots, T$

- receive example $x_t \in \mathbb{R}^d$ (wlog $\|x_t\|_\star \le 1$)
- predict a "score" $z_t \in \mathbb{R}$
- observe true label $y_t \in \{-1, +1\}$
- suffer logistic loss: $\ell(z_t, y_t) = \ln(1 + e^{-y_t z_t})$ 

Goal: minimize regret

$$\mathrm{Reg} = \sum_{t=1}^T \ell(z_t, y_t) - \min_{w \,:\, \|w\| \le B} \sum_{t=1}^T \ell(\langle w, x_t \rangle, y_t)$$
$$= \text{loss of learner} - \text{loss of best linear predictor}$$

Optimal regret in terms of d, T, B?

Results

Rates for Logistic Regression:

- Gradient Descent (Zinkevich'03): $\mathrm{Reg} = \mathcal{O}\left(B \cdot \sqrt{T}\right)$
- Online Newton Step (HAK'07): $\mathrm{Reg} = \mathcal{O}\left(\exp(B) \cdot d \ln T\right)$
- Open problem (MS'12): Is $\mathrm{Reg} = \mathcal{O}\left(\mathrm{poly}(B) \cdot d \ln T\right)$ possible?
- Negative answer (HKL'14): No, $\exp(B)$ is optimal for $\log T$.

Our work:

$$\mathrm{Reg} = \mathcal{O}(d \ln(BT))$$

i.e. *doubly-exponential improvement*.

- Via improper learning — (HKL'14) applies only to **proper** learner, i.e., $z_t = \langle w_t, x_t \rangle$ for some weight vector w_t.
- Matching lower bound.

Applications:

- bandit multiclass, resolving open problem of AR'09.
- online boosting, resolving open problem of BKL'15, JGT'17.

Extensions:

- statistical learning setting with high probability
- general non-linear function classes (non-constructively)
- (weighted) multiclass logistic regression

Multiclass Logistic Regression

For $t = 1, \ldots, T$

- receive example $x_t \in \mathbb{R}^d$
- predict a score vector $z_t \in \mathbb{R}^K$
- sample a label from softmax function $\sigma(z_t) = \left(\cdots, \frac{e^{z_{t,y}}}{\sum_{k \in [K]} e^{z_{t,k}}}, \cdots\right)$
- observe true label $y_t \in [K]$
- suffer logistic loss $\ell(z_t, y_t) = \ln(1 + \sum_{k \ne y_t} e^{z_{t,k} - z_{t,y_t}}) = -\ln \sigma(z_t)_{y_t}$

$$\mathrm{Reg} = \sum_{t=1}^T \ell(z_t, y_t) - \min_{\substack{W \in \mathbb{R}^{K \times d} \\ \|W_k\| \le B}} \sum_{t=1}^T \ell(W x_t, y_t)$$

Online Newton Step: $\mathcal{O}(e^B d K \ln T)$

Our work: $\mathcal{O}(dK \ln(BT))$

Lower bound

Theorem Any online logistic regression algorithm must have

$$\mathbf{Regret} = \Omega\left(d \log\left(\frac{B}{\sqrt{d} \log(T)}\right)\right).$$

Via reduction to Perceptron-type mistake bound setting.

Algorithmic Approach

Mixability of logistic loss

Vovk'95: $\ell : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ is η-mixable if:
$\forall$ distribution π over $\mathcal{Z}$, $\exists z_\pi \in \mathcal{Z}$ s.t. $\forall y \in \mathcal{Y}$ we have

$$\mathbb{E}_{z \sim \pi}[\exp(-\eta \ell(z, y))] \le \exp(-\eta \ell(z_\pi, y))$$

η-exp-concavity implies η-mixability: take $z_\pi = \mathbb{E}_{z \sim \pi}[z]$

Main observation: while logistic loss is e^{-B}-exp-concave, it's *1-mixable!*

$$\begin{aligned}\mathbb{E}_{z \sim \pi}[\exp(-\ell(z, y))] &= \mathbb{E}_{z \sim \pi}[\sigma(z)_y] \\ &= \exp(\ln \mathbb{E}_{z \sim \pi}[\sigma(z)]_y) \\ &= \exp(-\ell(\sigma^\dagger(\mathbb{E}_{z \sim \pi}[\sigma(z)]), y))\end{aligned}$$

Simple, but requires improper perspective

Vovk's Aggregating Algorithm

Predict $z_t = \sigma^\dagger(\mathbb{E}_{W \sim P_t}[\sigma(W x_t)])$ (which is improper) where

$$P_t(W) \propto \exp\left(-\sum_{\tau < t} \ell(W x_\tau, y_\tau)\right) \quad \forall W \in \mathbb{R}^{K \times d} : \|W_k\| \le B$$

Implementation

- requires log-concave sampler, e.g. hit-and-run (LV'06), projected Langevin Monte Carlo (BEL'15), Dikin walk (NR'17), ...

  (running time is a large polynomial)

Application 1: Bandit Multiclass

For $t = 1, \ldots, T$

- receive example $x_t \in \mathbb{R}^d$.
- predict a label $\hat{y}_t \in [K]$.
- observe "**right or wrong**", i.e., $\mathbf{1}\{\widehat{y}_t \ne y_t\}$.

Goal: mistake bound relative to surrogate loss:

$$\mathbb{E}\left[\sum_{t=1}^T \mathbf{1}\{\widehat{y}_t \ne y_t\}\right] \le \inf_{W : \|W_k\| \le B} \sum_{t=1}^T \ell(W x_t, y_t) + \mathbf{Reg}.$$

Abernethy and Rakhlin, COLT 2009 open problem:

Efficient algorithm with $O(\sqrt{T})$ regret for logistic loss?

Approach

1) Run improper logistic regression and sample $\hat{y}_t \sim \boldsymbol{\sigma}(z_t)$.
2) Loss not always observed! But can construct unbiased estimator:

$$\hat{\ell}(z, y_t) = -\frac{\ln \sigma(z_t)_{y_t}}{\sigma(z_t)_{y_t}} \mathbf{1}\{\widehat{y}_t = y_t\}$$

Leads to worse mixability parameter.

Final result : $\mathcal{O}(K\sqrt{dT \ln(BT)})$

Previous result (HK'11): $\mathcal{O}(dK^3 BT^{2/3})$

Extension 1: Statistical Learning

Setup:

Receive samples $(x_1, y_1), \ldots, (x_n, y_n)$ i.i.d. from unknown distribution $\mathcal{D}$.

Goal: Output predictor $\widehat{g} : \mathcal{X} \to \mathbb{R}^K$ such that with prob $\ge 1 - \delta$,

$$\mathbb{E}_{\mathcal{D}}\, \ell(\widehat{g}(x), y) - \inf_{W : \|W_k\| \le B} \mathbb{E}_{\mathcal{D}}\, \ell(W x, y) \le \mathbf{ExcessRisk}(n, d, K, B, \delta).$$

Improper learning: $\widehat{g}$ is possibly non-linear.

Theorem: For every δ, there exists predictor $\widehat{g}$ such that with probability at least $1 - \delta$,

$$\mathbb{E}_{\mathcal{D}}\, \ell(\widehat{g}(x), y) - \inf_{W : \|W_k\| \le B} \mathbb{E}_{\mathcal{D}}\, \ell(W x, y) \le \widetilde{O}\left(\frac{dK \log(Bn) \log\left(\frac{1}{\delta}\right)}{n}\right).$$

Approach: Adapt the "boosting the confidence" strategy from Mehta'17.

Application 2: Online Boosting

For $t = 1, \ldots, T$

- receive example x_t
- predict a label $\widehat{y}_t \in [K]$
- observe true label y_t

Goal: minimize #mistakes with the help of N *weak learners* who predicts $\widehat{y}_t^1, \ldots, \widehat{y}_t^N$.

Weak Learning Condition (JGT'17)

Let a weak learner i make predictions $\hat{y}_t^i \in [K]$.

There exists edge γ and sample complexity S such that for all outcomes $(y_t)_{t \le T}$ and cost matrices $(C_t)_{t \le T}$, we have

$$\sum_{t=1}^T C_t(y_t, \hat{y}_t^i) \le \sum_{t=1}^T \mathbb{E}_{k \sim u_{\gamma, y_t}}[C_t(y_t, k)] + S.$$

Randomized baseline: $u_{\gamma,y} \in \Delta_K$ has $u_{\gamma,y}(k) = (1-\gamma)/K$ for $k \ne y$ and $u_{\gamma,y}(k) = (1-\gamma)/K + \gamma$ for $k = y$.

Cost matrices: Live in
$\mathcal{C} = \left\{C \in \mathbb{R}_+^{K \times K} \mid \forall y \in [K], C(y,y) = 0 \text{ and } \|C(y, \cdot)\|_1 \le 1\right\}$.

Prior work: BKL'15: Gave optimal algorithm and separate adaptive algorithm with sub-optimal sample complexity. JGT'17: Same, but for multiclass.

Our result:

Adaptive online boosting with optimal sample complexity.

Sample complexity $\mathcal{O}(\frac{1}{\epsilon^2 \gamma^4}) \Rightarrow \mathcal{O}(\frac{1}{\epsilon \gamma^2})$ where ϵ is error rate (average # mistakes) and γ is edge.
Via improper logistic + BKL'15 strategy.

Also gives optimal+adaptive algorithm for multiclass and ranking variants (JGT'17, JT'18).

Extension 2: General Function Classes

General function class setup:

General instance space $\mathcal{X}$, benchmark class $\mathcal{F} \subseteq (\mathcal{X} \to \mathbb{R}^K)$.

$$\mathrm{Reg} = \sum_{t=1}^T \ell(z_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^T \ell(f(x_t), y_t)$$

Minimax regret: $\mathcal{V}_n(\mathcal{F})$.

Definition: Tree

For any set $\mathcal{Z}$, a $\mathcal{Z}$-valued K-ary tree of depth n is a sequence $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ of n mappings with $\mathbf{z}_t : [K]^{t-1} \to \mathcal{Z}$.

Definition: Cover

A set V of $\mathbb{R}^K$-valued K-ary trees is an α-cover (w.r.t. the L_p norm) of $\mathcal{F}$ on an $\mathcal{X}$-valued K-ary tree $\mathbf{x}$ of depth n with loss ℓ if

$$\forall f \in \mathcal{F},\ y \in [K]^n,\ \exists \boldsymbol{v} \in V \text{ s.t.}$$
$$\left(\frac{1}{n} \sum_{t=1}^n \max_{y'_t \in [K]} |\ell(f(\mathbf{x}_t(y)), y'_t) - \ell(\boldsymbol{v}_t(y), y'_t)|^p\right)^{1/p} \le \alpha.$$

The L_p covering number of $\mathcal{F}$ on tree $\mathbf{x}$ is defined as

$$\mathcal{N}_p(\alpha, \ell \circ \mathcal{F}, \mathbf{x}) := \min\{|V| : V \text{ is an } \alpha\text{-cover of } \mathcal{F} \text{ on } \mathbf{x}\}.$$

Further, define $\mathcal{N}_p(\alpha, \ell \circ \mathcal{F}) = \sup_{\mathbf{x}} \mathcal{N}_p(\alpha, \ell \circ \mathcal{F}, \mathbf{x})$.

Theorem:
Any function class $\mathcal{F}$ that is uniformly bounded over $\mathcal{X}$ enjoys the minimax value bound:

$$\mathcal{V}_n(\mathcal{F}) \le \inf_{\gamma > 0} \Bigg\{ \log \mathcal{N}_2(\gamma, \ell \circ \mathcal{F}) + \inf_{\gamma \ge \alpha > 0} \left\{ 8\alpha n + 24\sqrt{n} \int_\alpha^\gamma \sqrt{\log(\mathcal{N}_2(\delta, \ell \circ \mathcal{F}) \cdot n)}\, d\delta \right\} \Bigg\}.$$

Matches minimax square loss rate (RS'14), with no explicit dependence on scale B.

Examples

- s-sparse B-bounded linear classes:

$$\mathcal{V}_n(\mathcal{F}) \le s \log(dnB).$$

- All 1-Lipschitz B-bounded functions over $[0,1]^d$, $d \ge 2$:

$$\mathcal{V}_n(\mathcal{F}) \le (\log(Bn))^{1/d} n^{1 - 1/d}.$$