Open Problem: Better Bounds for Online Logistic Regression

H. Brendan McMahan MCMAHAN@GOOGLE.COM
Google Inc., Seattle, WA

Matthew Streeter MSTREETER@GOOGLE.COM
Google Inc., Pittsburgh, PA

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

Known algorithms applied to online logistic regression on a feasible set of L_2 diameter D achieve regret bounds like $\mathcal{O}(e^D \log T)$ in one dimension, but we show a bound of $\mathcal{O}(\sqrt{D} + \log T)$ is possible in a binary 1-dimensional problem. Thus, we pose the following question: Is it possible to achieve a regret bound for online logistic regression that is $\mathcal{O}(\text{poly}(D)\log(T))$? Even if this is not possible in general, it would be interesting to have a bound that reduces to our bound in the one-dimensional case.

Keywords: online convex optimization, online learning, regret bounds

1. Introduction and Problem Statement

Online logistic regression is an important problem, with applications like click-through-rate prediction for web advertising and estimating the probability that an email message is spam. We formalize the problem as follows: on each round t the adversary selects an example $(x_t, y_t) \in \mathbb{R}^n \times \{-1, 1\}$, the algorithm chooses model coefficients $w_t \in \mathbb{R}^n$, and then incurs loss

$$\ell(w_t; x_t, y_t) = \log(1 + \exp(-y_t w_t \cdot x_t)), \tag{1}$$

the negative log-likelihood of the example under a logistic model. For simplicity we assume $\|x_t\|_2 \le 1$ so that any gradient $\|\nabla\ell(w_t)\|_2 \le 1$. While conceptually any $w \in \mathbb{R}^n$ could be used as model parameters, for regret bounds we consider competing with a feasible set $\mathcal{W} = \{w \mid \|w\|_2 \le D/2\}$, the L_2 ball of diameter D centered at the origin.

Existing algorithms for online convex optimization can immediately be applied. First-order algorithms like online gradient descent (Zinkevich, 2003) achieve bounds like $\mathcal{O}(D\sqrt{T})$. On a bounded feasible set logistic loss (Eq. (1)) is exp-concave, and so we can use second-order algorithms like Follow-The-Approximate-Leader (FTAL), which has a general bound of $\mathcal{O}((\frac{1}{\alpha} + GD)n \log T)$ (Hazan et al., 2007) when the loss functions are α-exp-concave on the feasible set; we have $\alpha = e^{-D/2}$ for the logistic loss (see Appendix A), which leads to a bound of $\mathcal{O}((\exp(D) + D)n \log T)$ in the general case, or $\mathcal{O}(\exp(D) \log T)$ in the one-dimensional case. The exponential dependence on the diameter of the feasible set can make this bound worse than the $\mathcal{O}(D\sqrt{T})$ bounds for practical problems where the post-hoc optimal probability can be close to zero or one.

We suggest that better bounds may be possible. In the next section, we show that a simple Follow-The-Regularized-Leader (FTRL) algorithm can achieve a much better result, namely

$\mathcal{O}(\sqrt{D} + \log T)$, for one-dimensional problems where the adversary is further constrained[1] to pick $x_t \in \{-1, 0, +1\}$. A single mis-prediction can cost about $D/2$, and so the additive dependence on the diameter of the feasible set is less than the cost of one mistake. The open question is whether such a bound is achievable for problems of arbitrary finite dimension n. Even the general one-dimensional case, where $x_t \in [-1, 1]$, is not obvious.

2. Analysis in One Dimension

We analyze an FTRL algorithm. We can ignore any rounds when $x_t = 0$, and then since only the sign of $y_t x_t$ matters, we assume $x_t = 1$ and the adversary picks $y_t \in \{-1, 1\}$. The cumulative loss function on P positive examples and N negative examples is

$$c(w; N, P) = P \log(1 + \exp(-w)) + N \log(1 + \exp(w)).$$

Let N_t denote the number of negative examples seen through the t'th round, with P_t the corresponding number of positive examples. We play FTRL, with

$$w_{t+1} = \arg\min_w c(w; N_t + \lambda, P_t + \lambda),$$

for a constant $\lambda > 0$. This is just FTRL with a regularization function $r(w) = c(w; \lambda, \lambda)$. Using the FTRL lemma (e.g., McMahan and Streeter (2010, Lemma 1)), we have

$$\text{Regret} \le r(w^*) + \sum_{t=1}^{T} f_t(w_t) - f_t(w_{t+1})$$

where $f_t(w) = \ell(w; x_t, y_t)$.

It is easy to verify that $r(w) \le \lambda(|w| + 2 \log 2)$. It remains to bound $f_t(w_t) - f_t(w_{t+1})$. Fix a round t. For compactness, we write $N = N_{t-1}$ and $P = P_{t-1}$. Suppose that $y_t = -1$, so $N_t = N + 1$ and $P_t = P$ (the case when $y_{t+1} = +1$ is analogous). Since f_t is convex, by definition $f_t(w) \ge f_t(w_t) + g_t(w - w_t)$ where $g_t = \triangledown f_t(w_t)$. Taking $w = w_{t+1}$ and re-arranging, we have

$$f_t(w_t) - f_t(w_{t+1}) \le g_t(w_t - w_{t+1}) \le |g_t||w_t - w_{t+1}|.$$

It is easy to verify that $|g_t| \le 1$, and also that

$$w_t = \log\left(\frac{P + \lambda}{N + \lambda}\right).$$

Since $y_t = -1$, $w_{t+1} < w_t$, and so

$$\begin{aligned} |w_t - w_{t+1}| &= \log\left(\frac{P + \lambda}{N + \lambda}\right) - \log\left(\frac{P + \lambda}{N + 1 + \lambda}\right) \\ &= \log(N + 1 + \lambda) - \log(N + \lambda) \\ &= \log\left(1 + \frac{1}{N + \lambda}\right) \le \frac{1}{N + \lambda}. \end{aligned}$$

[1] Constraining the adversary in this way is reasonable in many applications. For example, re-scaling each x_t so $\|x_t\|_2 = 1$ is a common pre-processing step, and many problems also are naturally featurized by $x_{t,i} \in \{0, 1\}$, where $x_{t,i} = 1$ indicates some property i is present on the t'th example.

Thus, if we let $T^- = \{t \mid y_t = -1\}$, we have

$$\sum_{t \in T^-} f_t(w_t) - f_t(w_{t+1}) \leq \sum_{N=0}^{N_T} \frac{1}{N+\lambda} \leq \frac{1}{\lambda} + \sum_{N=1}^{N_T} \frac{1}{N} \leq \frac{1}{\lambda} + \log(N_T) + 1.$$

Applying a similar argument to rounds with positive labels and summing over the rounds with positive and negative labels independently gives

$$\text{Regret} \leq \lambda(|w^*| + 2\log 2) + \log(P_T) + \log(N_T) + \frac{2}{\lambda} + 2.$$

Note $\log(P_T) + \log(N_T) \leq 2\log T$. We wish to compete with w^* where $|w^*| \leq D/2$, so we can choose $\lambda = \frac{1}{\sqrt{D/2}}$ which gives

$$\text{Regret} \leq \mathcal{O}(\sqrt{D} + \log T).$$

References

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69, December 2007.

H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *COLT*, 2010.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.

Appendix A. The Exp-Concavity of the Logistic Loss

Theorem 1 *The logistic loss function* $\ell(w_t; x_t, y_t) = \log(1 + \exp(-y_t w_t \cdot x_t))$, *from Eq.* (1)*, is* α*-exp-concave with* $\alpha = \exp(-D/2)$ *over set* $\mathcal{W} = \{w \mid \|w\|_2 \leq D/2\}$ *when* $\|x_t\|_2 \leq 1$ *and* $y_t \in \{-1, 1\}$.

Proof Recall that a function ℓ is α-exp-concave if $\nabla^2 \exp(-\alpha\ell(w)) \preceq 0$. When $\ell(w) = g(w \cdot x)$ for $x \in \mathbb{R}^n$, we have $\nabla^2 \exp(-\alpha\ell(w)) = \nabla^2 f''(z) x x^\top$, where $f(z) = \exp(-\alpha g(z))$. For the logistic loss, we have $g(z) = \log(1 + \exp(z))$ (without loss of generality, we consider a negative example), and so $f(z) = (1 + \exp(z))^{-\alpha}$. Then,

$$f''(z) = \alpha e^z (1 + e^z)^{-\alpha-2}(\alpha e^z - 1).$$

We need the largest α such that $f''(z) \leq 0$, given a fixed z. We can see by inspection that $\alpha = 0$ is a zero. Since $e^z(1 + e^z)^{-\alpha-2} > 0$, from the term $(\alpha e^z - 1)$ we conclude $\alpha = e^{-z}$ is the largest value of α where $f''(z) \leq 0$. Note that $z = w_t \cdot x_t$, and so $|z| \leq D/2$ since $\|x_t\|_2 \leq 1$, and so taking the worst case over $w_t \in \mathcal{W}$ and x_t with $\|x_t\|_2 \leq 1$, we have $\alpha = \exp(-D/2)$. ■