JMLR: Workshop and Conference Proceedings vol 35:1–13, 2014

# Logistic Regression: Tight Bounds for Stochastic and Online Optimization

**Elad Hazan** EHAZAN@IE.TECHNION.AC.IL
**Tomer Koren** TOMERK@TECHNION.AC.IL
**Kfir Y. Levy** KFIRYL@TX.TECHNION.AC.IL
*Technion—Israel Institute of Technology, Haifa 32000, Israel*

## Abstract

The logistic loss function is often advocated in machine learning and statistics as a smooth and strictly convex surrogate for the 0-1 loss. In this paper we investigate the question of whether these smoothness and convexity properties make the logistic loss preferable to other widely considered options such as the hinge loss. We show that in contrast to known asymptotic bounds, as long as the number of prediction/optimization iterations is sub exponential, the logistic loss provides no improvement over a generic non-smooth loss function such as the hinge loss. In particular we show that the convergence rate of stochastic logistic optimization is bounded from below by a polynomial in the diameter of the decision set and the number of prediction iterations, and provide a matching tight upper bound. This resolves the COLT open problem of McMahan and Streeter (2012).

**Keywords:** Logistic regression, Stochastic optimization, Online learning, Lower bounds.

## 1. Introduction

In many applications, such as estimation of click-through-rate in web advertising, and predicting whether a patient has a certain disease, the logistic loss is often the loss of choice. It appeals as a convex surrogate of the 0-1 loss, and as a tool that not only yields categorical prediction but also able to estimate the underlying probabilities of the categories. Moreover, Friedman et al. (2000) and Collins et al. (2002) have shown that logistic regression is strongly connected to boosting.

A long standing debate in the machine learning community has been the optimal choice of surrogate loss function for binary prediction problems (see Langford (2009), Bulatov (2007)). Amongst the arguments in support of the logistic loss are its smoothness and strict-convexity properties, which unlike other loss functions (such as the hinge loss), permit the use of more efficient optimization methods. In particular, the logistic loss is exp-concave, and thus second-order methods are applicable and give rise to theoretically superior convergence and/or regret bounds.

More technically, under standard assumptions on the training data, the logistic loss is 1-Lipschitz and $e^{-D}$-exp-concave over the set of linear $n$-dimensional classifiers whose $L_2$-norm is at most $D$. Thus, the Online Newton Step algorithm (Hazan et al., 2007) can be applied to the logistic regression problem and gives a convergence rate of $\widetilde{O}(e^D n/T)$ over $T$ iterations. On the other hand, first order methods can be used to attain a rate of $O(D/\sqrt{T})$, which is attainable in general for any Lipschitz convex loss function. The exponential dependence on $D$ of the first bound suggests that second order methods might present poor performance in practical logistic regression problems, even when compared to the slow $1/\sqrt{T}$ rate of first-order methods. The gap between the two rates raises the question: ***is a fast convergence rate of the form $\widetilde{O}(poly(D)/T)$ achievable for logistic regression?***

This question has received much attention lately. Bach (2013), relying on a property called "generalized self-concordance", gave an algorithm with convergence rate of $O(D^4/\mu^* T)$, where $\mu^*$ is the smallest eigenvalue of the Hessian at the optimal point. This translates to a $O(\text{poly}(D)/T)$ rate whenever the expected loss function is "locally strongly convex" at the optimum. More recently, Bach and Moulines (2013) extended this result and presented an elegant algorithm that attains a rate of the form $O(\rho^3 D^4 n/T)$, without assuming strong convexity (neither global or local) — but rather depending on a certain data-dependent constant $\rho$.

In this paper, we resolve the above question and give tight characterization of the achievable convergence rates for logistic regression. We show that as long as the target accuracy $\epsilon$ is not exponentially small in $D$, a rate of the form $\widetilde{O}(\text{poly}(D)/T)$ is not attainable. Specifically, we prove a lower bound of $\Omega(\sqrt{D/T})$ on the convergence rate, that can also be achieved (up to a $\sqrt{D}$ factor) by stochastic gradient descent algorithms. In particular, this shows that in the worst case, the magnitude of data-dependent parameters used in previous works are exponentially large in the diameter $D$. The latter lower bound only applies for multi-dimensional regression (i.e., when $n \geq 2$); surprisingly, in one-dimensional logistic regression we find a rate of $\Theta(T^{-2/3})$ to be tight. As far as we know, this is the first natural setting demonstrating such a phase transition in the optimal convergence rates, with respect to the dimensionality of the problem.

| Setting | Previous | This Paper | |
|---|---|---|---|
| | | $n = 1$ | $n \geq 2$ |
| Stochastic | $O\left(\frac{D}{\sqrt{T}}\right)$ [Zinkevich]<br>$O\left(\frac{e^D \log T}{T}\right)$ [Hazan et al.] | $O\left(\frac{D^3}{T^{2/3}}\right)$ [Cor. 10]<br>$\Omega\left(\frac{D^{2/3}}{T^{2/3}}\right)$ [Thm. 2] | $\Omega\left(\sqrt{\frac{D}{T}}\right)$ [Thm. 4] |
| Online | $O(D\sqrt{T})$ [Zinkevich]<br>$O(e^D \log T)$ [Hazan et al.] | $O(D^3 T^{1/3})$ [Thm. 9]<br>$\Omega(D^{2/3} T^{1/3})$ [Cor. 7] | $\Omega(\sqrt{DT})$ [Cor. 8] |

Table 1: Convergence rates and regret bounds for the logistic loss, in the regime $T = O(e^D)$.

We also consider the closely-related online optimization setting, where on each round $t = 1, 2, \ldots, T$ an adversary chooses a certain logistic function and our goal is to minimize the $T$-round regret, with respect to the best fixed decision chosen with the benefit of hindsight. In this setting, McMahan and Streeter (2012) investigated the one-dimensional case and showed that if the adversary is restricted to pick binary (i.e. $\pm 1$) labels, a simple follow-the-leader algorithm attains a regret bound of $O(\sqrt{D} + \log T)$. This discovery led them to conjecture that bounds of the form $O(\text{poly}(D) \log T)$ should be achievable in the general multi-dimensional case with continuous labels set.

Our results extend to the online optimization setup and resolve the COLT 2012 open problem of McMahan and Streeter (2012) on the negative side. Namely, we show that as long as the number of rounds $T$ is not exponentially large in $D$, an upper bound of $O(\text{poly}(D) \log T)$ cannot be attained in general. We obtain lower bounds on the regret of $\Omega(\sqrt{D\,T})$ in the multi-dimensional case and $\Omega(D^{2/3}T^{1/3})$ in the one-dimensional case, when allowing the adversary to use a continuous label set. We are not aware of any other natural problem that exhibits such a dichotomy between the minimax regret rates in the one-dimensional and multi-dimensional cases.

It is interesting to note that our bounds apply to a finite interval of time, namely when $T = O(e^D)$, which is arguably the regime of interest for reasonable values of $D$. This is the reason our lower bounds do not contradict the logarithmic known regret bounds.

We prove the tightness of our one-dimensional lower bounds, in both the stochastic and online settings, by devising an online optimization algorithm specialized for one-dimensional online logistic regression that attains a regret of $O(D^3\,T^{1/3})$. This algorithm maintains approximations of the observed logistic loss functions, and use these approximate losses to form the next prediction by a follow-the-regularized-leader (FTRL) procedure. As opposed to previous works that utilize approximate losses based on *local* structure (Zinkevich, 2003; Hazan et al., 2007), we find it necessary to employ approximations that rely on the *global* structure of the logistic loss.

The paper is organized as follows. In Section 2 we describe the settings we consider and give the necessary background. We present our lowers bounds in Section 3, and in Section 4 we prove our upper bound for one dimensional logistic regression. We conclude in Section 5.

## 2. Setting and Background

In this section we formalize the settings of stochastic logistic regression and online logistic regression and give the necessary background on both problems.

### 2.1. Stochastic Logistic Regression

In the problem of stochastic logistic regression, there is an unknown distribution $\mathcal{D}$ over instances $x \in \mathbb{R}^n$. For simplicity, we assume that $\|x\| \leq 1$. The goal of an optimization algorithm is to minimize the expected loss of a linear predictor $w \in \mathbb{R}^n$,

$$L(w) \;=\; \mathbf{E}_{x\sim\mathcal{D}}[\,\ell(w,x)\,]\,, \tag{1}$$

where $\ell$ is the logistic loss function[1],

$$\ell(w,x) \;=\; \log\big(1+\exp(x\cdot w)\big)$$

that expresses the negative log-likelihood of the instance $x$ under the logit model. While we may try to optimize $L(w)$ over the entire Euclidean space, for generalization purposes we usually restrict the optimization domain to some bounded set. In this paper, we focus on optimizing the expected loss over the set $\mathcal{W} = \{w \in \mathbb{R}^n : \|w\| \leq D\}$, the Euclidean ball of radius $D$. We define the *excess loss* of a linear predictor $w \in \mathcal{W}$ as the difference $L(w) - \min_{w^*\in\mathcal{W}} L(w^*)$ between the expected loss of $w$ and the expected loss of the best predictor in the class $\mathcal{W}$.

An algorithm for the stochastic optimization problem, given a sample budget $T$ as a parameter, may use a sample $x_1, \ldots, x_T$ of $T$ instances sampled independently from the distribution $\mathcal{D}$, and

[1] The logistic loss is commonly defined as $\ell(w;x,y) = \log\big(1+\exp(-yx\cdot w)\big)$ for instances $(x,y) \in \mathbb{R}^n \times [-1,1]$. For ease of notation and without loss of generality, we ignore the variable $y$ in the instance $(x,y)$ by absorbing it into $x$.

produce an approximate solution $\overline{w}_T$. The *rate of convergence* of the algorithm is then defined as the expected excess loss of the predictor $\overline{w}_T$, given by

$$\mathbf{E}[L(\overline{w}_T)] \; - \; \min_{w^* \in \mathcal{W}} L(w^*) \; ,$$

where the expectation is taken with respect to both the random choice of the training set and the internal randomization of the algorithm (which is allowed to be randomized).

### 2.2. Online Logistic Regression

Another optimization framework we consider is that of online logistic optimization, which we formalize as the following game between a player and an adversary. On each round $t = 1, 2, \ldots, T$ of the game, the adversary first picks an instance $x_t \in \mathbb{R}^n$, the player then chooses a linear predictor $w_t \in \mathcal{W} = \{w \in \mathbb{R}^n \, : \, \|w\| \leq D\}$, observes $x_t$ and incurs loss

$$\ell(w_t, x_t) \; = \; \log\big(1 + \exp(x_t \cdot w_t)\big) \; .$$

For simplicity we again assume that $\|x_t\| \leq 1$ for all $t$. The goal of the player is to minimize his regret with respect to a fixed prediction from the set $\mathcal{W}$, which is defined as

$$\mathrm{Regret}_T \; = \; \sum_{t=1}^{T} \ell(w_t, x_t) \; - \; \min_{w^* \in \mathcal{W}} \sum_{t=1}^{T} \ell(w^*, x_t) \; .$$

### 2.3. Information-Theoretic Tools

As a part of our lower bound proofs, we utilize two impossibility theorems that assert the minimal number of samples needed in order to distinguish between two distributions. We prove the following lower bound on the performance of any algorithm for this task.

**Theorem 1** *Assume a coin with bias either $p$ or $p+\epsilon$, where $p \in (0, \frac{1}{2}]$, is given. Any algorithm that correctly identifies the coin's bias with probability at least $3/4$, needs no less than $p/16\epsilon^2$ tosses.*

The theorem applies to both deterministic and randomized algorithms; in case of random algorithms the probability is with respect to both the underlying distribution of the samples, and the randomization of the algorithm. The proof of Theorem 1 is given, for completeness, in Hazan et al. (2014).

## 3. Lower Bounds for Logistic Regression

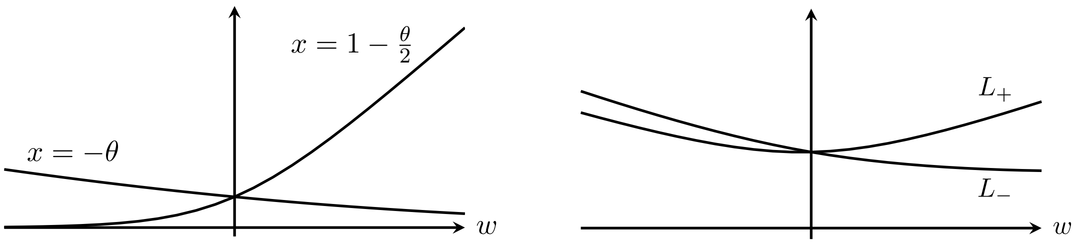
In this section we derive lower bounds for the convergence rate of stochastic logistic regression. For clarity, we lower bound the number of observations $T$ required in order to attain excess loss of at most $\epsilon$, which we directly translate to a bound for the convergence rate. The stochastic optimization lower bounds are then used to obtain corresponding bounds for the online setting.

In Section 3.1 we prove a lower bound for the one dimensional case, in Section 3.2 we prove another lower bound for the multidimensional case, and in Section 3.3 we present our lower bounds for the online setting.

### 3.1. One-dimensional Lower Bound for Stochastic Optimization

We now show that any algorithm for one-dimensional stochastic optimization with logistic loss, must observe at least $\Omega(D/\epsilon^{1.5})$ instances before it provides an instance with $\epsilon$ expected excess loss. This directly translates to a convergence rate of $\Omega(D^{2/3}/T^{2/3})$. Formally, the main theorem of this section is the following.

**Theorem 2** *Consider the one dimensional stochastic logistic regression setting with a fixed sample budget $T = O(e^D)$. For any algorithm $\mathcal{A}$ there exists a distribution $\mathcal{D}$ for which the expected excess loss of $\mathcal{A}$'s output is at least $\Omega(D^{2/3}/T^{2/3})$.*



($a$) Logistic loss functions corresponding to instances in the set $\{1 - \frac{\theta}{2}, -\theta\}$.

($b$) Expected loss functions induced by the distributions $\mathcal{D}_+, \mathcal{D}_-$.

Figure 1: Loss functions used in the one-dimensional construction, and the induced expected loss functions.

The proof of Theorem 2 is given at the end of this section; here we give an informal proof sketch. Consider distributions $\mathcal{D}$ over the two-element set $\{1 - \frac{\theta}{2}, -\theta\}$. For $w \in [D/2, D]$ and $\theta \ll 1$, the losses of these instances are approximately linear/quadratic with opposed slopes (see Fig. 1($a$)). Consequently, we can build a distribution with an expected loss which is quadratic in $w$; upon perturbing the latter distribution by $\pm\epsilon$ we get two distributions $\mathcal{D}_+, \mathcal{D}_-$ with expected losses $L_+, L_-$ that are approximately linear in $w$ with slopes $\pm\epsilon$ (see Fig. 1($b$)). An algorithm that attains a low expected excess loss on both these distributions can be used to distinguish between them, we then utilize an information theoretic impossibility theorem to bound the number of observations needed in order to distinguish between two distributions.

In Fig. 2 we present two distributions, which we denote by $\mathcal{D}_+$ and $\mathcal{D}_-$. We denote by $L_+, L_-$ the expected logistic loss of a predictor $w \in \mathcal{W}$ with respect to $\mathcal{D}_+, \mathcal{D}_-$, i.e.,

$$\begin{aligned} L_\chi(w) &= \mathbf{E}_{\mathcal{D}_\chi}[\ell(w,x)] \\ &= \left(\frac{\theta}{2} + \chi\frac{\epsilon}{D}\right)\ell\left(w, 1 - \frac{\theta}{2}\right) + \left(1 - \frac{\theta}{2} - \chi\frac{\epsilon}{D}\right)\ell\left(w, -\theta\right) \,, \qquad \chi \in \{-1, 1\} \,. \end{aligned}$$

The following lemma states that it is impossible attain a low expected excess loss on both $\mathcal{D}_+$ and $\mathcal{D}_-$ simultaneously. Here we only give a sketch of the proof; the complete proof can be found in Hazan et al. (2014).

**Input**: $\epsilon > 0$, $D \geq 1$, $\chi \in \{-1, 1\}$

1. Set $\theta = \sqrt{\epsilon}/D$
2. Define a distribution $\mathcal{D}_\chi$ over instances $x \in \{1 - \frac{\theta}{2}, -\theta\}$ as follows:

$$x = \begin{cases} 1 - \frac{\theta}{2} & \text{w.p. } \frac{\theta}{2} + \chi\frac{\epsilon}{D} \\ -\theta & \text{w.p. } 1 - \left(\frac{\theta}{2} + \chi\frac{\epsilon}{D}\right) \end{cases}$$

**Output**: For $t = 1, 2, \ldots$, sample $x_t \sim \mathcal{D}_\chi$

Figure 2: Two distributions: $\mathcal{D}_\chi$, $\chi \in \{-1, 1\}$; any algorithm that attains an $\epsilon$ expected excess logistic loss on both of them requires $\Omega(D/\epsilon^{1.5})$ observations.

**Lemma 3** *Given $D \geq 1$ and $\Omega(e^{-D}) \leq \epsilon \leq 1/25$, consider the distributions $\mathcal{D}_+, \mathcal{D}_-$ defined in Fig. 2. Then the following holds:*

$$\begin{aligned} L_+(w) - \min_{w^* \in \mathcal{W}} L_+(w^*) &\geq \epsilon/20\,, \qquad \forall\, w \in [\tfrac{3}{4}D, D]\,, \\ L_-(w) - \min_{w^* \in \mathcal{W}} L_-(w^*) &\geq \epsilon/20\,, \qquad \forall\, w \in [-D, \tfrac{3}{4}D]\,. \end{aligned}$$

**Proof (sketch)** First we show that for $w \in [\frac{1}{2}D, D]$, the losses of the instances $1 - \frac{\theta}{2}, -\theta$ are approximately linear/quadratic, i.e.,

$$\begin{aligned} \left|\ell(w, 1 - \tfrac{\theta}{2}) - (1 - \tfrac{\theta}{2})w\right| &\leq \frac{\epsilon}{40}\,, \qquad \forall\, w \in [\tfrac{1}{2}D, D]\,, \\ \left|\ell(w, -\theta) - \left(\log 2 - \tfrac{\theta}{2}w + \tfrac{1}{8}(\theta w)^2\right)\right| &\leq \frac{\epsilon}{40}\,, \qquad \forall\, w \in [\tfrac{1}{2}D, D]\,. \end{aligned}$$

Using the above approximations and $\theta = \sqrt{\epsilon}/D$, we show that $L_+(w) \approx \epsilon w/D + \epsilon w^2/8D^2$ and $L_-(w) \approx -\epsilon w/D + \epsilon w^2/8D^2$ for $w \in [\frac{1}{2}D, D]$, where "$\approx$" denotes equality up to an additive term of $\epsilon/40$. Thus,

$$\begin{aligned} L_+(w) - \min_{w^* \in \mathcal{W}} L_+(w^*) &\geq L_+(w) - L_+(D/2) \geq \epsilon/20\,, \qquad \forall\, w \in [\tfrac{3}{4}D, D]\,, \\ L_-(w) - \min_{w^* \in \mathcal{W}} L_-(w^*) &\geq L_-(w) - L_-(D) \geq \epsilon/20\,, \qquad \forall\, w \in [\tfrac{1}{2}D, \tfrac{3}{4}D]\,. \end{aligned}$$

Showing that $L_-$ is monotonically decreasing in $[-D, \frac{1}{2}D]$, extends the latter inequality to $[-D, \frac{3}{4}D]$. ■

We are now ready to prove Theorem 2.

**Proof of Theorem 2** Consider some algorithm $\mathcal{A}$; we will show that if $\mathcal{A}$ observes $T$ samples from a distribution $\mathcal{D}$ which is either $\mathcal{D}_+$ or $\mathcal{D}_-$, then the expected excess loss $\tilde{\epsilon}$ that $\mathcal{A}$ can guarantee is lower bounded by $\Omega(D^{2/3}T^{-2/3})$.

The excess loss is non negative; therefore, if $\mathcal{A}$ guarantees an expected excess loss smaller than $\tilde{\epsilon} := \epsilon/80$, then by Markov's inequality it achieves an excess loss smaller than $\epsilon/20$, w.p. $\geq 3/4$. Denoting by $\overline{w}_T$ the predictor that $\mathcal{A}$ outputs after $T$ samples, then according to Lemma 3, attaining an excess loss smaller than $\epsilon/20$ on the distribution $\mathcal{D}_+$ (respectively $\mathcal{D}_-$) implies $\overline{w}_T \leq \frac{3}{4}D$ (respectively $\overline{w}_T > \frac{3}{4}D$).

Since $\mathcal{A}$ achieves an excess loss smaller than $\epsilon/20$ w.p. $\geq 3/4$ for any distribution $\mathcal{D}$ we can use its output to identify the right distribution w.p. $\geq 3/4$. This can be done as follows:

$$\begin{aligned}&\text{If } \ \overline{w}_T \leq \tfrac{3}{4}D, \ \ \text{Return: "}\mathcal{D}_+\text{" ;}\\ &\text{If } \ \overline{w}_T > \tfrac{3}{4}D, \ \ \text{Return: "}\mathcal{D}_-\text{" .}\end{aligned}$$

According to Theorem 1 distinguishing between these two distributions ("coins") w.p. $\geq 3/4$ requires that the number of observations $T$ to be lower bounded as follows:

$$T \;\geq\; \frac{\theta/2 - \epsilon/D}{16(2\epsilon/D)^2} \;\geq\; \frac{1}{256}\frac{D}{\epsilon^{1.5}}\;,$$

We used $\theta/2 - \epsilon/D$ as a lower bound on the bias of $\mathcal{D}_-$; since $\theta = \sqrt{\epsilon}/D$ and $\epsilon \leq 1/25$ it follows that $\theta/2 - \epsilon/D \geq \sqrt{\epsilon}/4D$. We also used $2\epsilon/D$ as the bias between the "coins" $\mathcal{D}_+$, $\mathcal{D}_-$. Using the above inequality together with $\tilde{\epsilon} = \epsilon/80$ yields a lower bound of $\frac{1}{4000}D^{2/3}T^{-2/3}$ on the expected excess loss. ■

### 3.2. Multidimensional Lower Bound for Stochastic Optimization

We now construct two distribution over instance vectors from the unit ball of $\mathbb{R}^2$, and prove that any algorithm that attains an expected excess loss at most $\epsilon$ on both distributions requires $\Omega(D/\epsilon^2)$ samples in the worst case. This directly translates to a convergence rate of $\Omega(\sqrt{D/T})$. For $n > 2$ dimensions, we can embed the same construction in the unit ball of $\mathbb{R}^n$, thus our bound holds in any dimension greater than one. The main theorem of this section is the following.

**Theorem 4** *Consider the multidimensional stochastic logistic regression setting with $D \geq 2$ and a fixed sample budget $T = O(e^D)$. For any algorithm $\mathcal{A}$ there exists a distribution $\mathcal{D}$ such that the expected excess loss of $\mathcal{A}$'s output is at least $\Omega(\sqrt{D/T})$.*

Theorem 4 is proved at the end of this section. We bring here an informal description of the proof:

Consider distributions that choose instances among the set $\{x_0, x_l, x_r\}$ depicted in Fig. 3. The shaded areas in Fig. 3 depict regions in the domain $\mathcal{W}$ where either $\ell(\cdot, x_l)$ or $\ell(\cdot, x_r)$ is approximately linear. The dark area represents the region in which both loss functions are approximately linear. By setting the probability of $x_0$ much larger than the others we can construct a distribution over the instances $\{x_0, x_l, x_r\}$ such that the minima of the induced expected loss function lies in the black area. Perturbing this distribution by $\pm\epsilon$ over the odds of choosing $x_l, x_r$ we attain two distributions $\mathcal{D}_+, \mathcal{D}_-$ whose induced expected losses $L_+, L_-$ are almost linear over in the dark area, with opposed $\pm\epsilon$ slopes. An algorithm that attains a low expected excess loss on both distributions can be used to distinguish between them. This allows us to use information theoretic arguments to lower bound the number of samples needed for the optimization algorithm. In Fig. 4 we present the

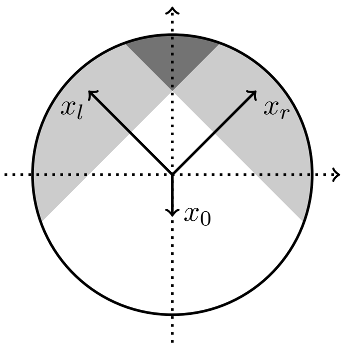


Figure 3: Instances used in multidimensional lower bound.

**Input**: $\epsilon > 0$, $D \geq 2$, $\chi \in \{-1, 1\}$

1. Set $p \in [0, 1]$ such that $\frac{p}{1-p} = \frac{D}{\sqrt{2}} \frac{1+e^{0.9}}{1+e^{-0.9D/\sqrt{2}}}$ and define:

$$x_0 = \tfrac{1}{D}(0, -1)^\top \,, \quad x_l = \tfrac{1}{\sqrt{2}}(-1, 1)^\top \,, \quad x_r = \tfrac{1}{\sqrt{2}}(1, 1)^\top$$

2. Define a distribution $\mathcal{D}_\chi$, that choose instances $x \in \{x_0, x_l, x_r\}$ as follows:

$$x = \begin{cases} x_0 & \text{w.p.} \;\; p \\ x_l & \text{w.p.} \;\; \frac{1+\chi\epsilon}{2} \cdot (1-p) \\ x_r & \text{w.p.} \;\; \frac{1-\chi\epsilon}{2} \cdot (1-p) \end{cases}$$

**Output**: For $t = 1, 2, \ldots$, sample $x_t \sim \mathcal{D}_\chi$

Figure 4: Two distributions: $\mathcal{D}_\chi$, $\chi \in \{-1, 1\}$; any algorithm that attains an $\epsilon$ excess logistic loss on both of them requires $\Omega(D/\epsilon^2)$ observations.

distributions $\mathcal{D}_+, \mathcal{D}_-$. We denote by $L_+$ and $L_-$ the expected loss functions induced by $\mathcal{D}_+$ and $\mathcal{D}_-$ respectively, that are given by

$$L_\chi(w) \;=\; p \cdot \ell(w, x_0) + \frac{1+\chi\epsilon}{2}(1-p) \cdot \ell(w, x_l) + \frac{1-\chi\epsilon}{2}(1-p) \cdot \ell(w, x_r), \qquad \chi \in \{-1, 1\}$$

In the following lemma we state that it is impossible attain a low expected excess loss on both $\mathcal{D}_+$ and $\mathcal{D}_-$ simultaneously.

**Lemma 5** *Given $D \geq 2$ and $\Omega(e^{-D}) \leq \epsilon \leq 1/10D$, consider $\mathcal{D}_+, \mathcal{D}_-$ as defined in Fig. 4. Then the following holds:*

$$\begin{aligned} L_+(w) - \min_{w^* \in \mathcal{W}} L_+(w^*) \;&\geq\; \epsilon/20 \,, \qquad \forall\, w : w[1] \leq 0 \,, \quad \textit{and} \\ L_-(w) - \min_{w^* \in \mathcal{W}} L_-(w^*) \;&\geq\; \epsilon/20 \,, \qquad \forall\, w : w[1] \geq 0 \,. \end{aligned}$$

Here we only give a sketch of the proof; for the complete proof, refer to Hazan et al. (2014).

**Proof (sketch)** Let $L_0$ be the unperturbed ($\epsilon = 0$) version of $L_+, L_-$, i.e.,

$$L_0(w) \;=\; p\ell(w, x_0) + \frac{1-p}{2}\ell(w, x_l) + \frac{1-p}{2}\ell(w, x_r)\;.$$

Note that $L_0$ is constructed such that its minima is attained at $w_0 = (0, 0.9D)$, which belongs to the shaded area in Fig. 3. Thus, in the neighborhood of this minima both $\ell(w, x_l), \ell(w, x_r)$ are approximately linear. Using linear approximations of $\ell(w, x_l), \ell(w, x_r)$ around $w_0$, we show that the value of $L_+$ at $w_a = (0.3D, 0.9D)$ is smaller by $\epsilon/20$ than the minimal value of $L_0$, hence

$$\min_{w^* \in \mathcal{W}} L_+(w^*) \;\le\; L_+(w_a) \;\le\; L_0(w_0) - \epsilon/20\;. \tag{2}$$

Moreover, $L_+$ is shown to be the sum of $L_0$ and a function which is positive whenever $w[1] \le 0$, thus

$$L_+(w) \;\ge\; L_0(w)\;, \quad \forall\, w \;:\; w[1] \le 0\;. \tag{3}$$

Combining Eqs. (2) and (3) we get

$$L_+(w) - \min_{w^* \in \mathcal{W}} L_+(w^*) \;\ge\; L_0(w) - \big(L_0(w_0) - \epsilon/20\big) \;\ge\; \epsilon/20\;, \qquad \forall\, w \;:\; w[1] \le 0\;,$$

where the last inequality follows from $w_0$ being the minimizer of $L_0(w)$. A similar argument shows that for predictors $w$ such that $w[1] \ge 0$, it holds that $L_-(w) - \min_{w^* \in \mathcal{W}} L_-(w^*) \ge \epsilon/20$. ■

For the proof of Theorem 4 we require a lemma that lower-bounds the minimal number of samples needed in order to distinguish between the distributions $\mathcal{D}_+, \mathcal{D}_-$ defined in Fig. 4. To this end, we use the following modified version of Theorem 1.

**Lemma 6** *Let $p \in (0, 1/2]$. Consider a distribution supported on three atoms with probabilities $\{q_0, (1-q_0)(p+\chi\epsilon), (1-q_0)(1-p-\chi\epsilon)\}$, with $\chi$ being either $0$ or $1$. Any algorithm that identifies the distribution correctly with probability at least $3/4$, needs no less than $p/16(1-q_0)\epsilon^2$ samples.*

Lemma 6 can be proved similarly to Theorem 1 (see Hazan et al. (2014)). We are now ready to prove Theorem 4.

**Proof of Theorem 4** Consider some algorithm $\mathcal{A}$; we will show that if $\mathcal{A}$ observes $T$ samples from a distribution $\mathcal{D}$ which is either $\mathcal{D}_+$ or $\mathcal{D}_-$, then the expected excess loss $\tilde{\epsilon}$ that $\mathcal{A}$ can guarantee is lower bounded by $\Omega(\sqrt{D/T})$.

The excess loss is non negative; therefore if $\mathcal{A}$ guarantees an expected excess loss smaller than $\tilde{\epsilon} = \epsilon/80$, then by Markov's inequality it achieves an excess loss smaller than $\epsilon/20$, w.p. $\ge 3/4$. Denoting by $\overline{w}_T$ the predictor that $\mathcal{A}$ outputs after $T$ samples, then according to Lemma 5, attaining an excess loss smaller than $\epsilon/20$ on distribution $\mathcal{D}_+$(respectively $\mathcal{D}_-$) implies $\overline{w}_T[1] > 0$ (respectively $\overline{w}_T[1] < 0$).

Since $\mathcal{A}$ achieves an excess loss smaller than $\epsilon/20$ w.p. $\ge 3/4$ for any $\mathcal{D}$ among $\mathcal{D}_+, \mathcal{D}_-$ we can use its output to identify the right distribution w.p. $\ge 3/4$. This can be done as follows:

$$\text{if } \;\overline{w}_T[1] \ge 0, \;\; \text{return ``}\mathcal{D}_+\text{''}\;\;;$$
$$\text{if } \;\overline{w}_T[1] < 0, \;\; \text{return ``}\mathcal{D}_-\text{''}\;\;.$$

According to Lemma 6, distinguishing between these two distributions w.p.$\geq 3/4$ requires that the number of observations $T$ to be upper bounded as follows:

$$T \;\geq\; \frac{0.5(1-\epsilon)}{16(1-p)(2\epsilon)^2} \;\geq\; \frac{D}{256}\frac{1}{\epsilon^2}\;,$$

We used $0.5(1-\epsilon)$ as a lower bound on the bias of distribution $\mathcal{D}_-$ conditioned that the instance $x_0$ was not chosen; since $\epsilon \leq 1/10D$, $D \geq 2$ it follows that $0.5(1-\epsilon) \geq 0.25$. We also used $2\epsilon$ as the bias between the distributions $\mathcal{D}_+$ and $\mathcal{D}_-$ conditioned that the label $x_0$ was not chosen. Finally we used $1-p \leq 1/D$. The above inequality together with $\tilde{\epsilon} = \epsilon/80$ yields a lower bound of $\frac{1}{1300}\sqrt{D/T}$ on the expected excess loss. ■

### 3.3. Lower Bounds for Online Optimization

In Section 3 we proved two lower bounds for the convergence rate of stochastic logistic regression. Standard online-to-batch conversion (Cesa-Bianchi et al., 2004) shows that any online algorithm attaining a regret of $R(T)$ can be used to attain a convergence rate of $R(T)/T$ for stochastic optimization. Hence, the lower bounds stated in Theorems 2 and 4 imply the following:

**Corollary 7** *Consider the one dimensional online logistic regression setting with $T = O(e^D)$. For any algorithm $\mathcal{A}$ there exists a sequence of loss functions such that $\mathcal{A}$ suffers a regret of at least $\Omega(D^{2/3}T^{1/3})$.*

**Corollary 8** *Consider the multidimensional online logistic regression setting with $T = O(e^D)$, $D \geq 2$. For any algorithm $\mathcal{A}$ there exists a sequence of loss functions such that $\mathcal{A}$ suffers a regret of at least $\Omega(\sqrt{DT})$.*

## 4. Upper Bound for One-dimensional Regression

In this section we consider online logistic regression in one dimension; here an adversary chooses instances $x_t \in [-1,1]$, then a learner chooses predictors $w_t \in \mathcal{W} = \{w \in \mathbb{R} : |w| \leq D\}$, and suffers a logistic loss $\ell(w_t, x_t) = \log(1+e^{x_t w_t})$. We provide an upper bound of $O(T^{1/3})$ for logistic online regression in one dimension, thus showing that the lower bound found in Theorem 2 is tight. Formally, we prove:

**Theorem 9** *Consider the one dimensional online regression with logistic loss. Then a player that chooses predictors $w_t \in \mathcal{W}$ according to Algorithm 1 with $\eta = T^{-1/3}$ and $D \geq 2$, achieves the following guarantee:*

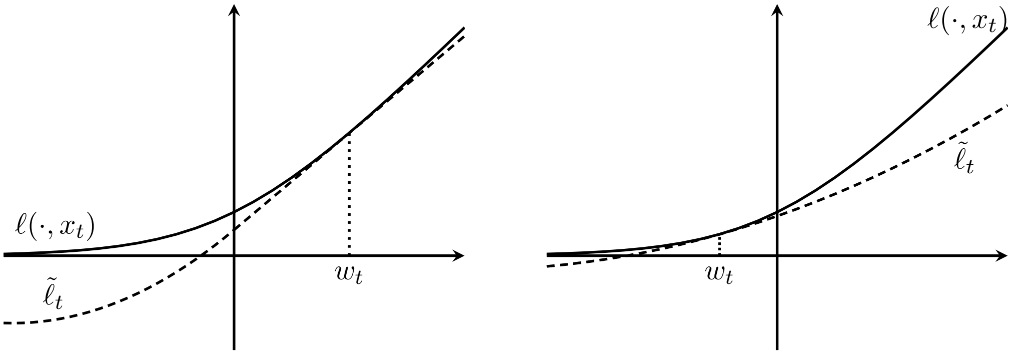
$$\textit{Regret}_T \;=\; \sum_{t=1}^{T}\log(1+e^{x_t w_t}) - \min_{w\in\mathcal{W}}\sum_{t=1}^{T}\log(1+e^{x_t w}) \;=\; O(D^3\,T^{1/3})\,.$$

Using standard online-to-batch conversion techniques Cesa-Bianchi et al. (2004), we can translate the upper bound given in the above lemma to an upper bound for stochastic optimization.

**Corollary 10** *Consider the one dimensional stochastic logistic regression setting with $D \geq 2$ and a budget of $T$ samples. Then for any distribution $\mathcal{D}$ over instances, an algorithm that chooses predictors $w_1, \dots, w_t \in \mathcal{W}$ according to Algorithm 1 with $\eta = T^{-1/3}$ and outputs $\overline{w}_T = \frac{1}{T}\sum_{\tau=1}^{T} w_\tau$, achieves the following guarantee:*

$$\mathbf{E}[L(\overline{w}_T)] \;-\; \min_{w^* \in [-D,D]} L(w^*) \;=\; O(D^3/T^{2/3})\,.$$

Following Zinkevich (2003) and Hazan et al. (2007), we approximate the losses received by the adversary, and use the approximate losses in a follow-the-regularized-leader (FTRL) procedure in order to choose the predictors.



(*a*) Mixed linear/quadratic approximation

(*b*) Quadratic approximation

Figure 5: Approximate losses used by Algorithm 1.

First, note the following lemma due to Zinkevich (2003) (proof is found in Hazan et al. (2007)).

**Lemma 11** *Let $\ell_1, \dots, \ell_T$ be an arbitrary sequence of loss functions, and let $w_1, \dots, w_T \in \mathcal{K}$. Let, $\tilde{\ell}_1, \dots, \tilde{\ell}_T$ be a sequence of loss function that satisfy $\tilde{\ell}_t(w_t) = \ell_t(w_t)$, and $\tilde{\ell}_t(w) \leq \ell_t(w)$ for all $w \in \mathcal{K}$. Then*

$$\sum_{t=1}^{T} \ell_t(w_t) - \min_{w \in \mathcal{K}} \sum_{t=1}^{T} \ell_t(w) \;\leq\; \sum_{t=1}^{T} \tilde{\ell}_t(w_t) - \min_{w \in \mathcal{K}} \sum_{t=1}^{T} \tilde{\ell}_t(w)\,.$$

Thus, the regret on the original losses is bounded by the regret of the approximate losses. For the logistic losses, $\ell(w, x_t) = \log(1 + e^{x_t w})$, we define approximate losses $\tilde{\ell}_t$ that satisfy the conditions of the last lemma. Depending on $x_t, w_t$, we divide into 3 cases:

$$\tilde{\ell}_t(w) = \begin{cases} a_0 + y_t w + \frac{\beta}{2} y_t^2 w^2 \mathbb{1}_{w \leq 0} & \text{if } w_t \geq 0 \text{ and } x_t \geq \frac{1}{D}\,; \\ a_0 + y_t w + \frac{\beta}{2} y_t^2 w^2 \mathbb{1}_{w \geq 0} & \text{if } w_t \leq 0 \text{ and } x_t \leq -\frac{1}{D}\,; \\ a_0 + y_t w + \frac{\beta}{2} y_t^2 (w - w_t)^2 & \text{if } |x_t| \leq \frac{1}{D} \text{ or } x_t w_t \leq 0\,, \end{cases} \tag{4}$$

where,

$$y_t = \left.\frac{\partial \ell(w, x_t)}{\partial w}\right|_{w_t} = g_t x_t\,, \quad g_t = \frac{e^{x_t w_t}}{1 + e^{x_t w_t}}\,, \quad \beta = 1/8D\,, \quad a_0 = \log(1 + e^{x_t w_t}) - g_t x_t w_t\,.$$

Thus, if $|x_t| \leq 1/D$ or $x_t w_t \leq 0$, then we use a quadratic approximation, else we use a loss that changes from linear to quadratic on $w = 0$. Note that if the approximation loss $\tilde{\ell}_t$ is partially linear, then the magnitude of its slope $|y_t|$ is greater than $1/2D$.

The approximations are depicted in Fig. 5. In Fig. 5($a$) the approximate loss changes from linear to quadratic in $w = 0$ , where in Fig. 5($b$) the approximate loss is quadratic everywhere. The following technical lemma states that the losses $\{\tilde{\ell}_t\}$ satisfy the conditions of Lemma 11.

**Lemma 12** *Assume that $D \geq 2$. Let $\ell(\cdot, x_1), \dots, \ell(\cdot, x_T)$ be a sequence of logistic loss functions and let $w_1, \dots, w_T \in \mathcal{W}$. The approximate losses $\tilde{\ell}_1, \dots, \tilde{\ell}_T$ defined above satisfy $\tilde{\ell}_t(w_t) = \ell(w_t, x_t)$ and $\tilde{\ell}_t(w) \leq \ell(w, x_t)$ for all $w \in \mathcal{W}$.*

Lemma 12 is proved in Hazan et al. (2014). We are now ready to describe our algorithm that obtains a regret of $O(D^3 T^{1/3})$ for one-dimensional online regression, given in Algorithm 1.

**Algorithm 1** FTRL for logistic losses

```
  Input: Learning rate $\eta > 0$, diameter $D$
  let $R(w) = \frac{1}{16D} w^2$
  for $t = 1, 2 \dots T$ do
    set $w_t = \arg\min_{w \in [-D,D]} \left\{ \sum_{\tau=1}^{t-1} \tilde{\ell}_\tau(w) + \frac{1}{\eta} R(w) \right\}$
    observe $x_t \in [-1, 1]$ and suffer loss $\ell(w_t, x_t) = \log(1 + e^{x_t w_t})$
    compute $\tilde{\ell}_t$ according to Eq. (4)
  end for
```

We conclude with a proof sketch of Theorem 9; for the complete proof, see Hazan et al. (2014).

**Proof of Theorem 9 (sketch)** First we show that the regret of Algorithm 1 is upper bounded by the sum of differences $\sum_{t=1}^{T} \tilde{\ell}'_t(w_t)(w_t - w_{t+1})$, and then divide the analysis into two cases. In the first case we show that the accumulated regret in rounds where $\tilde{\ell}_t$ is quadratic around $w_t$ is upper bounded by $O(D \log T)$. The second case analyses rounds in which $\tilde{\ell}_t$ is linear around $w_t$; due to the regularization, in the first such $T^{2/3}$ rounds our regret is bounded by $O(T^{1/3})$ and if the number of such rounds is greater than $T^{2/3}$ we show that the quadratic part of the accumulated losses is large enough so the above sum of differences is smaller than $O(D^3 T^{1/3})$. Since the approximations $\tilde{\ell}_t$ may change from linear to quadratic in $w = 0$, our analysis splits into two cases: the case where consecutive predictors $w_t, w_{t+1}$ have the same sign, and the case where they have opposite signs. ■

## 5. Summary and Open Questions

We have given tight bounds for stochastic and online logistic regression that preclude the existence of fast rates for logistic regression without exponential factors. As a consequence, we have also resolved the COLT 2012 open problem of McMahan and Streeter (2012). Our lower bounds can be extended to the multidimensional setting in which the instances are normalized and the labels are binary.

Our results suggest that second-order methods might present poor performance in practical logistic regression problems. Indeed, in the derivation of our lower bounds we have constructed a distribution over instances such that the induced expected loss function is approximately linear around its optimum.

An interesting feature of our results is that our regret/convergence bounds apply to *a finite range of $T$*, and are different than the known asymptotic bounds. Arguably, the range of $T$ for which our results apply is the important one in practice (sub-exponential in the size of the hypothesis class). Are there other natural settings in which regret bounds for bounded number of iterations differ from the asymptotic bound?

## Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 336078 – ERC-SUBLRN.

## References

Francis Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *arXiv preprint arXiv:1303.6149*, 2013.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26*, pages 773–781. 2013.

Yaroslav Bulatov. Log loss or hinge loss? http://yaroslavvb.blogspot.co.il/2007/06/log-loss-or-hinge-loss.html, June 2007.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Michael Collins, Robert E Schapire, and Yoram Singer. Logistic regression, adaboost and bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Elad Hazan, Tomer Koren, and Kfir Y. Levy. Logistic regression: Tight bounds for stochastic and online optimization. *arXiv preprint arXiv:1405.3843*, 2014.

John Langford. Optimal proxy loss for classification. http://hunch.net/?p=547, April 2009.

Brendan H McMahan and Matthew J Streeter. Open problem: Better bounds for online logistic regression. *Journal of Machine Learning Research-Proceedings Track*, 23:44–1, 2012.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.