# Faster logistic regression?

Dirk van der Hoeven

April 19, 2023

# Step 1: data collection

I am interested in finding out the probability that you prefer Pierce Brosnan over Sean Connery as James Bond Given Age.

Who is the best actor for James Bond (Sean Connery or Pierce Brosnan) and what is your age?

# Step 2: choose and train our model

Many option to choose from

- boosting
- neural network
- decision tree
- random forest
- **logistic regression**

But how to train our model?

# Today

**How to guarantee that you are predicting as well as the best model in class?**

# Today

**How to guarantee that you are predicting as well as the best model in class?**

And how to use online learning to achieve our goal

# Today

**How to guarantee that you are predicting as well as the best model in class?**

And how to use online learning to achieve our goal

1. statistical learning (batch setting)
2. online learning
3. online to batch conversion in expectation
4. online to batch conversion with high-probability
5. possible approach for fast logistic regression
6. self-concordant functions

# Statistical learning

1 The learner receives i.i.d. data $D = (X_1, Y_1), (X_2, Y_2), \ldots, (X_T, Y_T)$

# Statistical learning

1 The learner receives i.i.d. data $D = (X_1, Y_1), (X_2, Y_2), \ldots, (X_T, Y_T)$

2 The learner chooses $\bar{f} : \mathcal{X} \to \mathcal{Y}$

# Statistical learning

1 The learner receives i.i.d. data $D = (X_1, Y_1), (X_2, Y_2), \ldots, (X_T, Y_T)$

2 The learner chooses $\bar{f} : \mathcal{X} \to \mathcal{Y}$

**Goal**: control the excess risk

$$R = \mathbb{E}_{X,Y}[\ell(\bar{f}(X), Y)] - \min_{f \in \mathcal{F}} \mathbb{E}_{X,Y}[\ell(f(X), Y)]$$

either in expectation or with high-probability over $D$.

# Statistical learning: an interactive example

$D = (1,1), (1,1), (1,1), (1,1), (1,1), (1,1), (1,0)$

$\mathcal{F} = \{0, 1\}$

Q: what is a reasonable predictor $\bar{f}$?

# Statistical learning: an interactive example

$D = (1,1),(1,1),(1,1),(1,1),(1,1),(1,1),(1,0)$

$\mathcal{F} = \{0, 1\}$

Q: what is a reasonable predictor $\bar{f}$?

Intuitive option: ERM (empirical risk minimization)

$$\bar{f} = \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \ell(f(X_t), Y_t)$$

# Logistic regression

Today's objective is to control the logistic loss for linear $\mathcal{F}$, i.e.

$$\ell(x^\top \theta, y) = -\mathbb{1}[y = 1] \log(\sigma(x^\top \theta)) - \mathbb{1}[y = -1] \log(1 - \sigma(x^\top \theta)),$$

where $\sigma(z) = \frac{1}{1+\exp(z)}$

# Logistic regression

Today's objective is to control the logistic loss for linear $\mathcal{F}$, i.e.

$$\ell(x^\top \theta, y) = -\mathbb{1}[y=1]\log(\sigma(x^\top \theta)) - \mathbb{1}[y=-1]\log(1-\sigma(x^\top \theta)),$$

where $\sigma(z) = \frac{1}{1+\exp(z)}$

An alternative formulation:

$$\ell(x^\top \theta, y) = \log(1+\exp(yx^\top \theta))$$

# Logistic regression: **improper** excess risk

The excess risk is defined as:

$$R = \mathbb{E}_{X,Y}\Big[ - \mathbb{1}[Y = 1] \log(\bar{\sigma}(X)) - \mathbb{1}[Y = -1] \log(1 - \bar{\sigma}(X))\Big] \\ - \min_{\theta \in \mathcal{B}(b)} \mathbb{E}_{X,Y} \log(1 + \exp(YX^\top \theta))$$

We assume that $\|X\| \leq r$ almost surely.

**Important** our estimator might be improper: $\bar{\sigma}(x)$ is not necessarily equal to $\sigma(x^\top \hat{\theta})$ for some $\hat{\theta}$.

# Excess risk of ERM

For logistic regression, **in expectation over $D$**, ERM is known to obtain

$$\theta_{ERM} = \operatorname*{arg\,min}_{\theta \in \mathcal{B}(b)} \frac{1}{T} \sum_{t=1}^{T} \log(1 + \exp(Y_t, X_t^\top \theta))$$
$$R = O\left(\min\left\{\frac{br}{\sqrt{T}}, \frac{d \exp(br)}{T}\right\}\right)$$

# Excess risk of ERM

For logistic regression, **in expectation over $D$**, ERM is known to obtain

$$\theta_{ERM} = \operatorname*{arg\,min}_{\theta \in \mathcal{B}(b)} \frac{1}{T} \sum_{t=1}^{T} \log(1 + \exp(Y_t, X_t^\top \theta))$$

$$R = O\left(\min\left\{\frac{br}{\sqrt{T}}, \frac{d\exp(br)}{T}\right\}\right)$$

But is it optimal?

# Excess risk of ERM

For logistic regression, **in expectation over $D$**, ERM is known to obtain

$$\theta_{ERM} = \arg\min_{\theta \in \mathcal{B}(b)} \frac{1}{T} \sum_{t=1}^{T} \log(1 + \exp(Y_t, X_t^\top \theta))$$
$$R = O\left(\min\left\{\frac{br}{\sqrt{T}}, \frac{d \exp(br)}{T}\right\}\right)$$

But is it optimal?

Yes: for **proper learners** it is known that one can not do better than the above rate.

# Excess risk of improper learners **in expectation**

All in big-O

- Foster et al (2018)
  excess risk: $\frac{d \log(rbT/d)}{T}$ runtime $=$ poly$(T, d)$
- Jezequel et al (2020), Agarwal (2021)
  excess risk: $\frac{drb \log(T)}{T}$ runtime $T(d^2 + \log(T))$
- Mourtada and Gaiffas (2021)
  excess risk: $\frac{d+(rb)^2}{T}$ runtime $=$ ERM

# Excess risk of improper learners **w.h.p.**

All in big-O, with probability at least $1 - \delta$

- Vijaykumar (2021) (*slightly stronger guarantee)
  Excess risk: $\frac{d}{T} \log(T)(\log(Trb) + \log(1/\delta))$ runtime = ???
- Puchkin and Zhivotovskiy (2023)
  Excess risk: $\frac{\exp(br)(d+\log(1/\delta))}{T}$ runtime = ERM
- **Van der Hoeven et al (to appear 2023)**
  Excess risk: $\frac{d}{T} \log(Trb) + \frac{\log(T) \log(1/\delta)}{T}$ runtime = poly($d$, $T$)

# How to obtain our bounds?

Online to batch conversion in expectation:

1. run an online learning algorithm sequentially over the data using the original loss function
2. average the predictors of the online learning algorithm
3. use Jensen's inequality

# How to obtain our bounds?

Online to batch conversion in expectation:

1. run an online learning algorithm sequentially over the data using the original loss function
2. average the predictors of the online learning algorithm
3. use Jensen's inequality

Online to batch conversion with high probability:

1. run an online learning algorithm sequentially over the data using a **shifted loss** function
2. average the predictors of the online learning algorithm
3. use Jensen's inequality + a version of **freedman's inequality** for martingales
4. **cancel some quadratic terms**

# The program

1 ~~statistical learning~~

2 online learning

3 online to batch conversion in expectation

4 online to batch conversion with high-probability

5 possible approach for fast logistic regression

6 self-concordant functions

# Online learning

Online learning proceeds in rounds $t = 1, \ldots, T$ In each round $t$

1 the **environment picks** an outcome $y_t \in \mathcal{Y}$ and reveals context $\boldsymbol{x}_t \in \mathcal{X}$

# Online learning

Online learning proceeds in rounds $t = 1, \dots, T$ In each round $t$

1. the **environment picks** an outcome $y_t \in \mathcal{Y}$ and reveals context $\boldsymbol{x}_t \in \mathcal{X}$
2. the learner chooses $f_t$ and issues prediction $f_t(\boldsymbol{x}_t) \in \mathcal{Y}$

# Online learning

Online learning proceeds in rounds $t = 1, \ldots, T$ In each round $t$

1. the **environment picks** an outcome $y_t \in \mathcal{Y}$ and reveals context $\boldsymbol{x}_t \in \mathcal{X}$
2. the learner chooses $f_t$ and issues prediction $f_t(\boldsymbol{x}_t) \in \mathcal{Y}$
3. the learner suffers $\ell(f_t(\boldsymbol{x}_t), y_t)$

# Online learning

Online learning proceeds in rounds $t = 1, \ldots, T$ In each round $t$

1 the **environment picks** an outcome $y_t \in \mathcal{Y}$ and reveals context $\boldsymbol{x}_t \in \mathcal{X}$

2 the learner chooses $f_t$ and issues prediction $f_t(\boldsymbol{x}_t) \in \mathcal{Y}$

3 the learner suffers $\ell(f_t(\boldsymbol{x}_t), y_t)$

4 the environment reveals $y_t$.

# Online learning

Online learning proceeds in rounds $t = 1, \ldots, T$ In each round $t$

1. the **environment picks** an outcome $y_t \in \mathcal{Y}$ and reveals context $\boldsymbol{x}_t \in \mathcal{X}$
2. the learner chooses $f_t$ and issues prediction $f_t(\boldsymbol{x}_t) \in \mathcal{Y}$
3. the learner suffers $\ell(f_t(\boldsymbol{x}_t), y_t)$
4. the environment reveals $y_t$.

I assume that $\ell$ is $\alpha-$exp concave and is known to the learner at the start of the game.

# Goal: control the (expected) regret

$$\mathcal{R}_T = \underbrace{\sum_{t=1}^{T} \ell(f_t(\boldsymbol{x}_t), y_t)}_{\text{how did the learner do?}} - \underbrace{\min_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(\boldsymbol{x}_t), y_t)}_{\text{how well does the best fixed prediction do?}}$$

# Goal: control the (expected) regret

$$\mathcal{R}_T = \underbrace{\sum_{t=1}^{T} \ell(f_t(\boldsymbol{x}_t), y_t)}_{\text{how did the learner do?}} - \underbrace{\min_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(\boldsymbol{x}_t), y_t)}_{\text{how well does the best fixed prediction do?}}$$

Example: for **any sequence** of $(X_1, Y_1), \ldots, (X_T, Y_T)$, with $\alpha$-**exp concave**, the exponential weights algorithm guarantees a $\frac{\log(|\mathcal{F}|)}{\alpha}$ regret bound, which is known to be optimal (up to constants).

# Exp-concavity

A function $g : \mathcal{Z} \to [0, m]$ is called $\alpha$-exp concave if

$$\tilde{g}(z) = \exp(-\alpha g(z))$$

is a concave function.

# Exp-concavity

A function $g : \mathcal{Z} \to [0, m]$ is called $\alpha$-exp concave if

$$\tilde{g}(z) = \exp(-\alpha g(z))$$

is a concave function. **Alternatively** a function is exp-concave if

$$\alpha(g'(z))^2 \leq g''(z)$$

**Examples**: squared loss (linear regression), log loss (density estimation), logistic loss (logistic regression)

# Online learning: an interactive example

In each round $t$

# Online learning: an interactive example

In each round $t$

1. I choose $y_t \in \{0, 1\}$ (there is no context)

# Online learning: an interactive example

In each round $t$

1. **I** choose $y_t \in \{0, 1\}$ (there is no context)
2. **You** choose prediction $\in \{0, 1\}$

# Online learning: an interactive example

In each round $t$

1. **I** choose $y_t \in \{0, 1\}$ (there is no context)
2. **You** choose prediction $\in \{0, 1\}$
3. **You** suffer loss $(\hat{y}_t - y_t)^2$

# Online learning: an interactive example

In each round $t$

1. **I** choose $y_t \in \{0, 1\}$ (there is no context)
2. **You** choose prediction $\in \{0, 1\}$
3. **You** suffer loss $(\hat{y}_t - y_t)^2$
4. **I** reveal $y_t$

# Online learning: an interactive example

**You**: predict $\hat{y}_t \in \{0, 1\}$

**I**: give you the answer

Outcomes:

# Online learning: an interactive example

**You**: predict $\hat{y}_t \in \{0, 1\}$

**I**: give you the answer

Outcomes: $y_1 = 1,$

# Online learning: an interactive example

**You**: predict $\hat{y}_t \in \{0, 1\}$

**I**: give you the answer

Outcomes: $y_1 = 1,\ \ y_2 = 1,$

# Online learning: an interactive example

**You**: predict $\hat{y}_t \in \{0, 1\}$

**I**: give you the answer

Outcomes: $y_1 = 1,\ \ y_2 = 1,\ \ y_3 = 1,$

# Online learning: an interactive example

**You**: predict $\hat{y}_t \in \{0, 1\}$

**I**: give you the answer

Outcomes: $y_1 = 1,\ \ y_2 = 1,\ \ y_3 = 1,\ \ y_4 = 1,$

# Online learning: an interactive example

**You**: predict $\hat{y}_t \in \{0, 1\}$

**I**: give you the answer

Outcomes: $y_1 = 1,\ \ y_2 = 1,\ \ y_3 = 1,\ \ y_4 = 1,\ \ y_5 = 1,$

# Online learning: an interactive example

**You**: predict $\hat{y}_t \in \{0, 1\}$

**I**: give you the answer

Outcomes: $y_1 = 1, \;\; y_2 = 1, \;\; y_3 = 1, \;\; y_4 = 1, \;\; y_5 = 1, \;\; y_6 = 1,$

# Online learning: an interactive example

**You**: predict $\hat{y}_t \in \{0, 1\}$

**I**: give you the answer

Outcomes: $y_1 = 1,\ \ y_2 = 1,\ \ y_3 = 1,\ \ y_4 = 1,\ \ y_5 = 1,\ \ y_6 = 1,\ \ y_7 = 0$

# Why standard tools do not work

Methods relying solely on exp-concavity will not work because the logistic loss is $\exp(-br)$-exp concave:

$$\ell'(z, y) = y\frac{\exp(zy)}{1 + exp(zy)}$$

$$\ell''(z, y) = y^2\frac{exp(zy)}{(1 + exp(zy))^2}$$

Tools based on $\alpha$-exp concavity have bounds that scale with

$$\frac{1}{\alpha}$$

which is why we need improper methods: these give us more leverage.

# Online to batch conversion

We will use:

$$\bar{\sigma} = \frac{1}{T}\sum_{t=1}^{T}\sigma_{\mu,t}$$

where $\sigma_{\mu,t}, \ldots, \sigma_{\mu,t}$ are the functions output by an online learning algorithm.

# Online to batch conversion

We will use:

$$\bar{\sigma} = \frac{1}{T}\sum_{t=1}^{T}\sigma_{\mu,t}$$

where $\sigma_{\mu,t}, \dots, \sigma_{\mu,t}$ are the functions output by an online learning algorithm.

For example, we could use **Online Newton Step** (ONS) to obtain some $\theta_t$ and output

$$\sigma_{\mu,t}(x) = (1-\mu)\sigma(x^\top \theta_t) + \mu\tfrac{1}{2}$$

# Online to batch conversion

We will use:

$$\bar{\sigma} = \frac{1}{T}\sum_{t=1}^{T}\sigma_{\mu,t}$$

where $\sigma_{\mu,t}, \ldots, \sigma_{\mu,t}$ are the functions output by an online learning algorithm.

For example, we could use **Online Newton Step** (ONS) to obtain some $\theta_t$ and output

$$\sigma_{\mu,t}(x) = (1-\mu)\sigma(x^\top \theta_t) + \mu\frac{1}{2}$$

Alternatively, we could use **continuous exponential weights** to obtain a distribution $P_t$ over $\theta$ to obtain

$$\sigma_t(x) = \mathbb{E}_{P_t}[(1-\mu)\sigma(x^\top \theta)] + \mu\frac{1}{2}$$

# Online to batch conversion in Expectation over data

Let $D_{t-1} = (X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1})$
By convexity of $\ell$ we have

$$
\begin{aligned}
&\mathbb{E}_{X,Y,D}[\ell(\bar{\sigma}(X), Y)] \\
&\leq \mathbb{E}_{X,Y,D}[\frac{1}{T}\sum_{t=1}^{T} \ell(\sigma_{\mu,t}(X), Y)] \\
&= \mathbb{E}_D\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{X_t,Y_t}[\ell(\sigma_{\mu,t}(X_t), Y_t)|D_{t-1}]\right] \\
&= \mathbb{E}_D\left[\frac{1}{T}\sum_{t=1}^{T} \ell(\sigma_{\mu,t}(X_t), Y_t)\right] \qquad \text{why it's easy in } \mathbb{E}
\end{aligned}
$$

# Online to batch conversion in expectation over data

Similarly

$$\min_{f \in \mathcal{F}} \mathbb{E}_{X,Y}[\ell(f(X), Y)] \geq \mathbb{E}_D[\min_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \ell(f_t(X_t), Y_t)]$$

and thus

$$\mathbb{E}_{X,Y,D}[\ell(\bar{\sigma}(X), Y)] - \min_{f \in \mathcal{F}} \mathbb{E}_{X,Y,D}[\ell(f(X), Y)] \leq \mathbb{E}_D[\frac{1}{T}\mathcal{R}_T]$$

We have converted the guarantees of an online learning algorithms to the statistical learning (batch) setting

# Intermission: the problem with improper methods

Consider a simple setting where $\mathcal{F} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$.
Audibert (2007) showed that there exists a $\delta \in [0, 1]$ such that with probability at least $\delta$, for the squared loss and OTB of exponential weights must suffer

$$\mathbb{E}_{X,Y}[\ell(\bar{f}(X), Y)] - \min_{f \in \mathcal{F}} \mathbb{E}_{X,Y}[\ell(f(X), Y)] \geq c\sqrt{\frac{\log(e\delta^{-1})}{T}}$$

for some $c > 0$.

# Intermission: the problem with improper methods

Consider a simple setting where $\mathcal{F} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$. Audibert (2007) showed that there exists a $\delta \in [0, 1]$ such that with probability at least $\delta$, for the squared loss and OTB of exponential weights must suffer

$$\mathbb{E}_{X,Y}[\ell(\bar{f}(X), Y)] - \min_{f \in \mathcal{F}} \mathbb{E}_{X,Y}[\ell(f(X), Y)] \geq c\sqrt{\frac{\log(e\delta^{-1})}{T}}$$

for some $c > 0$.

# Intermission: the problem with improper methods

Consider a simple setting where $\mathcal{F} = \{0,1\}$ and $\mathcal{Y} = \{0,1\}$.
Audibert (2007) showed that there exists a $\delta \in [0,1]$ such that with probability at least $\delta$, for the squared loss and OTB of exponential weights must suffer

$$\mathbb{E}_{X,Y}[\ell(\bar{f}(X), Y)] - \min_{f \in \mathcal{F}} \mathbb{E}_{X,Y}[\ell(f(X), Y)] \geq c\sqrt{\frac{\log(e\delta^{-1})}{T}}$$

for some $c > 0$.

The **cause of this behaviour** is at the heart of several issues of **improper learners**: **sometimes better** than the best in class, **sometimes worse** than the best in class, which leads to **high variance**.

Foster et al (2018) did not account for this behavior in their high-probability bound.

# Star aggregation

Audibert (2007) also showed that for the squared loss an algorithm called star aggregation does in fact guarantee that, with probability at least $1 - \delta$

$$\mathbb{E}_{X,Y}[\ell(\bar{f}(X), Y)] - \min_{f \in \mathcal{F}} \mathbb{E}_{X,Y}[\ell(f(X), Y)] \leq C \frac{\log(\delta^{-1}) + \log(|\mathcal{F}|)}{T}$$

for some $C > 0$

# Star aggregation

Audibert (2007) also showed that for the squared loss an algorithm called star aggregation does in fact guarantee that, with probability at least $1 - \delta$

$$\mathbb{E}_{X,Y}[\ell(\bar{f}(X), Y)] - \min_{f \in \mathcal{F}} \mathbb{E}_{X,Y}[\ell(f(X), Y)] \leq C \frac{\log(\delta^{-1}) + \log(|\mathcal{F}|)}{T}$$

for some $C > 0$

**But.....** it is horrifically slow.

The algorithm of Vijaykumar (2021) is based on this idea.

# A look at the problem with OTB

Recall $D_{t-1} = (X_1, Y_1), \ldots, (X_{t-1}, Y_{t-1})$

By convexity of $\ell$ we have

$$\begin{aligned}&\mathbb{E}_{X,Y}[\ell(\bar{\sigma}(X), Y)]\\ &\leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{X_t,Y_t}[\ell(\sigma_{\mu,t}(X_t), Y_t)|D_{t-1}]\end{aligned}$$

How to recover the loss suffered by the learner, i.e. $\sum_{t=1}^{T} \ell(\sigma_{\mu,t}(X_t), Y_t)$? Or perhaps, how to recover the regret?

# Martingales

Denote by $f^\star = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{X,Y}[\ell(f(X), Y)]$

Let

$$\begin{aligned} r_t &= \ell(\sigma_{\mu,t}(X_t), Y_t) - \ell(f^\star(X_t), Y_t) \\ Z_t &= \mathbb{E}_{X_t,Y_t}[r_t | D_{t-1}] - r_t \end{aligned}$$

$Z_1, \ldots, Z_T$ can be recognised as a martingale difference sequence, for which standard concentration results exist.

# Martingales

Suppose that $|Z_t| \leq K$ **(!!)** almost surely

Azuma-Hoeffding: with probability at least $1 - \delta$

$$\sum_{t=1}^{T} Z_t \leq K\sqrt{T \log(1/\delta)}$$

hopeless because of the $\sqrt{T}$...

# Martingales

Suppose that $|Z_t| \leq K$ almost surely

(A version of) Freedman's inequality: for a fixed $\lambda \in (0, 1/K]$, with probability at least $1 - \delta$

$$\sum_{t=1}^{T} Z_t \leq \frac{\log(1/\delta)}{\lambda} + \lambda \sum_{t=1}^{T} \mathbb{E}[Z_t^2 | D_{t-1}]$$

Nicer, but how to deal with the $\sum_{t=1}^{T} \mathbb{E}_{X_t, Y_t}[Z_t^2 | D_{t-1}]$?

# Exp-concavity to the rescue

## Lemma

*Suppose that* $g : \mathcal{Z} \to [0, m]$ *is an* $\alpha$*-exp-concave function. Let* $\gamma = 4 \max \left\{ m, \frac{1}{\alpha} \right\}$*. Then*

$$g\left(\tfrac{1}{2}x + \tfrac{1}{2}z\right) \leq \tfrac{1}{2}g(x) + \tfrac{1}{2}g(z) - \frac{\big(g(x) - g(z)\big)^2}{4\gamma}\ , \qquad \textit{for all } x, z \in \mathcal{Z}\ .$$

# Exp-concavity to the rescue

## Lemma

*Suppose that* $g : \mathcal{Z} \to [0, m]$ *is an* $\alpha$*-exp-concave function. Let* $\gamma = 4 \max\left\{m, \frac{1}{\alpha}\right\}$. *Then*

$$g\left(\tfrac{1}{2}x + \tfrac{1}{2}z\right) \leq \tfrac{1}{2}g(x) + \tfrac{1}{2}g(z) - \frac{\big(g(x) - g(z)\big)^2}{4\gamma}\,, \qquad \textit{for all } x, z \in \mathcal{Z}\ .$$

**"inequalities write papers"**

# How to use the inequality

Recall $\ell(p(x), y) = -y\log(p(x)) - (1-y)\log(1-p(x))$. This is not bounded in a nice manner!
Instead: we predict with $\sigma_{\mu,t}(x) = (1-\mu)\sigma_t(x) + \mu\frac{1}{2}$.
For $\mu \leq \frac{1}{2}$ we can show that

$$\ell(p(x), y) \geq \ell((1-\mu)p(x) + \mu\tfrac{1}{2}, y) - 2\mu$$

We use $\ell_\mu(p(x)) = \ell((1-\mu)p(x) + \mu\frac{1}{2}, y) \in [0, \log(2/\mu)]$. We have that

$$\ell(\sigma_{\mu,t}(x), y) - \ell(p(x), y) \leq \ell_\mu(\sigma_t(x), y) - \ell_\mu(p(x), y) + 2\mu$$

# How to use the inequality

Alternatively, we can set $\mu = 0$ and simply use that

$$\log(1 + exp(z)) \leq \log(2) + |z| \tag{1}$$

# How to use the inequality

Let $f^\star(x) = \sigma(x^\top \theta^\star)$
Recall that $r_t = \ell(\sigma_{\mu,t}(X_t), Y_t) - \ell(f^\star(X_t), Y_t)$

Since $\ell$ is 1-exp concave in its first argument, we can use the inequality to show that

$$r_t \leq 2\ell_\mu(\sigma_t(X_t), Y_t) - 2\ell_\mu(\tfrac{1}{2}f^\star(X_t) + \tfrac{1}{2}\sigma_t(X_t), Y_t) - \frac{(r_t)^2}{2\gamma} + 2\mu$$

# How to use the inequality

Let $f^\star(x) = \sigma(x^\top \theta^\star)$
Recall that $r_t = \ell(\sigma_{\mu,t}(X_t), Y_t) - \ell(f^\star(X_t), Y_t)$

Since $\ell$ is 1-exp concave in its first argument, we can use the inequality to show that

$$r_t \leq 2\ell_\mu(\sigma_t(X_t), Y_t) - 2\ell_\mu(\tfrac{1}{2}f^\star(X_t) + \tfrac{1}{2}\sigma_t(X_t), Y_t) - \frac{(r_t)^2}{2\gamma} + 2\mu$$

We can now use $\sum_{t=1}^T -\frac{(r_t)^2}{2\gamma}$ to **compensate** for $\sum_{t=1}^T \mathbb{E}_{X_t, Y_t}[Z_t^2 | D_{t-1}]$ from **Freedman's inequality**!

# But wait...

How do we deal with

$$\sum_{t=1}^{T} \left(\ell_\mu(\sigma_t(X_t), Y_t) - \ell_\mu(\tfrac{1}{2}f^\star(X_t) + \tfrac{1}{2}\sigma_t(X_t), Y_t)\right) \text{ ?}$$

# But wait...

How do we deal with

$$\sum_{t=1}^{T} \big(\ell_\mu(\sigma_t(X_t), Y_t) - \ell_\mu(\tfrac{1}{2} f^\star(X_t) + \tfrac{1}{2}\sigma_t(X_t), Y_t)\big) \ ?$$

Define $\tilde{\ell}_t(\sigma(\theta^\top X_t)) = \ell(\frac{1}{2}\sigma(\theta^\top X_t) + \frac{1}{2}\sigma_t(X_t), Y_t)$ and bound the **shifted regret**

$$\begin{aligned}
&\sum_{t=1}^{T} \big(\ell_\mu(\sigma_t(X_t), Y_t) - \ell_\mu(\tfrac{1}{2} f^\star(X_t) + \tfrac{1}{2}\sigma_t(X_t), Y_t)\big) \\
&= \sum_{t=1}^{T} \big(\tilde{\ell}_t(\sigma_t(X_t)) - \tilde{\ell}_t(f^\star(X_t))\big)
\end{aligned}$$

# Online learning to the rescue!

Because $-\log$ is 1-exp concave we have that $\tilde{\ell}_t$ is also 1-exp concave. This means that we can run our favorite algorithm (**exponential weights** with a **uniform prior** over the ball) on losses $\tilde{\ell}_t$ to guarantee that

$$\tilde{\mathcal{R}}_T = \sum_{t=1}^{T} \left(\tilde{\ell}_t(\sigma_t(X_t)) - \tilde{\ell}_t(f^\star(X_t))\right) \leq Cd \log(1 + brT/d)$$

For some constant $C > 0$

# Combining it all

## Theorem

*Suppose that the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, m]$ is $\alpha$-exp concave in its first argument. Then the risk of the averaged estimator*

$$\bar{f} = \frac{1}{T} \sum_{t=1}^{T} f_t$$

*satisfies, with probability at least $1 - \delta$ with respect to the random draw of $(X_t, Y_t)_{t=1}^{T}$,*

$$\mathbb{E}_{X,Y}[\ell(\bar{f}(X), Y)] - \min_{f \in \mathcal{F}} \mathbb{E}_{X,Y}[\ell(f(X), Y)] \leq \frac{2\tilde{\mathcal{R}}_T + 2\gamma \log(1/\delta)}{T}$$

*where $\gamma = 4 \max\left\{m, \frac{1}{\alpha}\right\}$.*

# Interpreting it all

The results implies that we online need to run our online learning algorithms on the **shifted losses** $\tilde{\ell}_t$ to obtain **high-probability** bound on the **excess risk**

# Interpreting it all

The results implies that we online need to run our online learning algorithms on the **shifted losses** $\tilde{\ell}_t$ to obtain **high-probability** bound on the **excess risk**

For logistic regression, this implies that we have an algorithm with runtime poly$(dT)$ that obtains a

$$O(\frac{d}{T}\log(Trb) + \frac{\log(T)\log(1/\delta)}{T})$$

excess risk bound with probability at least $\geq 1 - \delta$.

Alternative excess risk bound with $\mu = 0$: $O(\frac{d}{T}\log(Trb) + \frac{rb\log(1/\delta)}{T})$

# Open problems

1. We do not understand logistic regression fully. Is there a fast algorithm with a nice bound w.h.p.?
2. Extensions to self-concordant losses?

# A new hope

Suppose that $|u|, |z| \leq br$

Recent works are able to use the lower bound

$$\log(1 + exp(yu)) \geq \log(1 + exp(yz)) \\ + y\frac{\exp(yz)}{(1 + exp(yz))}(u - z) + \frac{1}{br + 1}\frac{\exp(yz)}{(1 + exp(yz))^2}(u - z)^2$$

To obtain a $O(dbr\log(T)$ regret bound with $\tilde{O}(d^2 T)$ runtime.
**Their magic idea:** let parameter $\theta_t$ in $\sigma_t(x^\top \theta_t)$ depend on feature $x_t$

# Why we need to work

Denote by $\kappa_t = \frac{1}{br+1}\frac{\exp(Y_t X_t^\top \theta_t(X_t))}{(1+exp(Y_t X_t^\top \theta_t(X_t)))^2}$, $\beta_t = Y_t \frac{exp(Y_t X_t^\top \theta_t(X_t))}{1+\exp(Y_t X_t^\top \theta_t(X_t))}$

$$\ell_t(\theta) = \beta_t(X_t^\top \theta - X_t^\top \theta_t(X_t)) + \kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta)^2$$

Suppose that we use $\bar{\sigma}(x) = \frac{1}{T}\sum_{t=1}^T \sigma(x^\top \theta_t(x))$. We have that

$$\begin{aligned}
R &\leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{X,Y}[\ell_t(\theta_t(X_t)) - \ell_t(\theta^\star)|D_{t-1}] \\
&= \frac{1}{T}\left(\sum_{t=1}^T \mathbb{E}_{X,Y}[\beta_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star) - \tfrac{1}{2}\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2|D_{t-1}] \right.\\
&\quad \left. - \tfrac{1}{2}\,\mathbb{E}_{X,Y}[\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2|D_{t-1}]\right)
\end{aligned}$$

# Why we need to work

Applying Friedman to conclude that, with probability at least $1-\delta$

$$\sum_{t=1}^{T} \mathbb{E}_{X,Y}[\beta_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star) - \tfrac{1}{2}\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2 | D_{t-1}]$$

$$\leq \sum_{t=1}^{T} \beta_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star) - \tfrac{1}{2}\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2 + \frac{\log(1/\delta)}{\lambda}$$

$$+ \lambda \sum_{t=1}^{T} \mathbb{E}_{X,Y}\left[\left(\beta_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star) - \tfrac{1}{2}\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2\right)^2 | D_{t-1}\right],$$

for some fixed $\lambda \in [0, \frac{1}{R}]$, where $R$ is such that $2|\ell_t(\theta^\star)| \leq R$

# Why we need to work

Since $\frac{1}{2}\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2 \leq 2(br)^2$ we have that

$$\mathbb{E}_{X,Y}\left[\left(\beta_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star) - \tfrac{1}{2}\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2\right)^2 | D_{t-1}\right]$$
$$\leq \mathbb{E}_{X,Y}\left[2\left(\beta_t^2(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2 + 4\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2\right)^2 | D_{t-1}\right]$$

Thus, setting $\lambda = \frac{\alpha}{(br)^2}$ for some fixed $\alpha \in [0, \frac{1}{2}]$ we find...

# Why we need to work

with probability at least $1 - \delta$

$$R \leq \frac{1}{T}\left( \underbrace{\sum_{t=1}^{T} \beta_t (X_t^\top \theta_t(X_t) - X_t^\top \theta^\star) - \tfrac{1}{2}\kappa_t (X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2}_{\textbf{regret}} \right.$$
$$+ \frac{\alpha}{(br)^2}\,\mathbb{E}_{X,Y}\left[2\Big(\beta_t^2 (X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2 | D_{t-1}\right]$$
$$\left. + (4\alpha - \tfrac{1}{2})\,\mathbb{E}_{X,Y}[\kappa_t (X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2 | D_{t-1}] \right) + \frac{(br)^2 \log(1/\delta)}{\alpha}$$

Now, setting $\alpha < 1/16$....

# Why we need to work

with probability at least $1 - \delta$

$$R \leq \frac{1}{T}\Bigg(\underbrace{\sum_{t=1}^{T} \beta_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star) - \tfrac{1}{2}\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2}_{\textbf{regret = (?) } O(dbr\log(T))}$$

$$+ \frac{\alpha}{(br)^2}\mathbb{E}_{X,Y}\Big[2\Big(\beta_t^2(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2 | D_{t-1}\Big]$$

$$\underbrace{- \frac{1}{4}\mathbb{E}_{X,Y}[\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2 | D_{t-1}]}_{\textbf{use to compensate the positive quadratic?}}\Bigg) + \frac{(br)^2\log(1/\delta)}{\alpha}$$

# Why we need to work

recall

$$\beta_t^2 = \frac{\exp(2Y_t X_t^\top \theta_t(X_t))}{(1 + exp(Y_t X_t^\top \theta_t(X_t)))^2}$$

$$\kappa_t = \frac{1}{br+1} \frac{\exp(Y_t X_t^\top \theta_t(X_t))}{(1 + exp(Y_t X_t^\top \theta_t(X_t)))^2}$$

This suggest us to set $\alpha = O((br)^{-1} \exp(-br))$ to find ...

# Why we need to work

with probability at least $1-\delta$

$$R \leq \frac{1}{T}\Bigg( \underbrace{\sum_{t=1}^{T} \beta_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star) - \tfrac{1}{2}\kappa_t(X_t^\top \theta_t(X_t) - X_t^\top \theta^\star)^2}_{\textbf{regret = (?)}\ O(dbr\log(T))}$$

$$+ \exp(br)br\log(1/\delta))$$

$$= O(\frac{dbr\log(T) + \exp(br)(br)\log(1/\delta)}{T})$$

**There is still hope though**: the algorithm that obtains the $O(dbr\log(T))$ regret bound does not use that $\beta^2 \leq \exp(br)\kappa_t$, even though this is the standard approach in online learning.

# Other approaches?

- boosting improper learners: standard idea relies on Markov's inequality for the excess risk...
- Mourtada and Gaiffas (2021) obtain a $\frac{d+(rb)^2}{T}$ excess risk bound in $\mathbb{E}$: try to get a high-probability version of their algorithm?
- Other approaches?

# Second open problem

If $g$ satisfies

$$2(g''(x))^{3/2} \geq g'''(x)$$

Then $g$ is called self-concordant and

# Second open problem

If $g$ satisfies

$$2(g''(x))^{3/2} \geq g'''(x)$$

Then $g$ is called self-concordant and

$$g(x) - g(u) \leq g'(x)(x-u) - \sqrt{(x-u)^2 g''(x)} + \ln(1 + \sqrt{(x-u)^2 g''(x)})$$
$$g(x) - g(u) \leq g'(u)(x-u) - \sqrt{(x-u)^2 g''(u)} - \ln(1 - \sqrt{(x-u)^2 g''(u)})$$

Is this sufficient to avoid scaling $L$ and/or $b$ for $\alpha$-exp concave $g$ where $|g'(x)| \leq L$ and $\|x\| \leq b$ in high-probability bounds?

# Second open problem: why

The high-probability part of our bound scales with $\frac{1}{\alpha} + m$, where $m$ is such that $\ell(f(x), y) \in [0, m]$. W.p. at least $1 - \delta$

$$
\begin{aligned}
R &= O\left(\frac{1}{T}\Big(R_T + (\frac{1}{\alpha} + m)\log(1/\delta)\Big)\right) \\
&= O\left(\frac{1}{T}\Big(R_T + \underbrace{(\frac{1}{\alpha} + Lb)}_{\color{red}{Lb\ \textbf{can be big}}}\log(1/\delta)\Big)\right)
\end{aligned}
$$

Application in portfolio selection, logistic regression(?), statistical learning (MLE of covariance matrices, s.c. huber loss), and generally quite nice. Also, if the function is also a barrier for a domain **then no need for projections** anymore!

# Second open problem: why

The high-probability part of our bound scales with $\frac{1}{\alpha} + m$, where $m$ is such that $\ell(f(x), y) \in [0, m]$. W.p. at least $1 - \delta$

$$R = O\left(\frac{1}{T}\Big(R_T + (\frac{1}{\alpha} + m)\log(1/\delta)\Big)\right)$$

$$= O\left(\frac{1}{T}\Big(R_T + \underbrace{(\frac{1}{\alpha} + Lb)}_{Lb \textbf{ can be big}}\log(1/\delta)\Big)\right)$$

Application in portfolio selection, logistic regression(?), statistical learning (MLE of covariance matrices, s.c. huber loss), and generally quite nice. Also, if the function is also a barrier for a domain **then no need for projections** anymore!

**Reason for hope:**

- recent results for portfolio selection show that regret bounds do not need to scale with $L$
- Self-concordance is used in optimization for similar gains.

# Credits

**Co-authors:** Nikita Zhivotovskiy (Berkeley statistics) and Nicolò Cesa-Bianchi (University of Milan)

**Where to find our things**

Van der Hoeven, D., Zhivotovskiy, N., & Cesa-Bianchi, N. (2022). A Regret-Variance Trade-Off in Online Learning. NeurIPS 2022.

Van der Hoeven, D., Zhivotovskiy, N., & Cesa-Bianchi, N. (2023). Excess Risk Bounds via Sequential Predictors. To appear 2023. **Ask for a preprint!**