HAL open science

Logarithmic Regret for parameter-free Online Logistic Regression

Joseph de Vilmarest, Olivier Wintenberger

To cite this version:

Joseph de Vilmarest, Olivier Wintenberger. Logarithmic Regret for parameter-free Online Logistic Regression. 2019. hal-02046843

HAL Id: hal-02046843

https://hal.science/hal-02046843v1

Preprint submitted on 22 Feb 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LOGARITHMIC REGRET FOR PARAMETER-FREE ONLINE LOGISTIC REGRESSION

A PREPRINT

Joseph de Vilmarest
josephdevilmarest@gmail.com
LPSM, Sorbonne Université
4 Place Jussieu, 75005 Paris, France

Olivier Wintenberger
olivier.wintenberger@upmc.fr
LPSM, Sorbonne Université
4 Place Jussieu, 75005 Paris, France

February 22, 2019

ABSTRACT

We consider online optimization procedures in the context of logistic regression, focusing on the Extended Kalman Filter (EKF). We introduce a second-order algorithm close to the EKF, named Semi-Online Step (SOS), for which we prove a $\mathcal{O}(\log(n))$ regret in the adversarial setting, paving the way to similar results for the EKF. This regret bound on SOS is the first for such parameter-free algorithm in the adversarial logistic regression. We prove for the EKF in constant dynamics a $\mathcal{O}(\log(n))$ regret in expectation and in the well-specified logistic regression model.

***K*eywords** Kalman filter, Logistic Regression, Online Optimization

1 Introduction

In the convex online optimization literature (Hazan, 2016), a crucial issue is the tuning of parameters. Our aim is to develop parameter-free algorithms in the context of logistic regression. One observes $y_t \in \{-1, 1\}$ recursively through time $t = 1, 2, \ldots$. At each instance $t-1$, the objective is to construct a prediction of the next value y_t. In hand, we have explanatory variables X_t in $\mathbb{R}^d$ along with the past pairs $(X_s, y_s)_{s<t}$. We reduce the prediction to a d-dimensional optimization problem thanks to the logistic loss function

$$\ell_t(y_t, \hat{\theta}_t) = \log(1 + \exp(-y_t \hat{\theta}_t^T X_t)), \qquad t = 1, 2, \ldots,$$

where $\hat{\theta}_t \in \mathbb{R}^d$, $t = 1, 2, \ldots$, are provided by a recursive algorithm. The aim of online convex optimization is to provide regret bounds on the cumulative losses $\sum_{t=1}^n \ell_t(y_t, \hat{\theta}_t)$ for algorithms whose recursive update step is of constant complexity.

The logistic loss is exp-concave, property that guarantees the existence of online procedures achieving $O(\log(n))$ regret in the adversarial setting. The seminal paper of Hazan et al. (2007) proposed two such algorithms, Online Newton Step and Follow The Approximate Leader, that achieve this rate of convergence. Both methods require the knowledge of some constants unknown in practice, namely the constant of exp-concavity and an upper-bound on the gradients of the losses. They also require a projection step on a convex set of finite diameter. We consider these methods as localized ones because they use the strongly convex paraboloid local approximation of any exp-concave functions stated in Lemma 3 in Hazan et al. (2007).

On the contrary, some recent papers (Bach and Moulines, 2013; Gadat and Panloup, 2017; Godichon-Baggioni, 2018) propose global algorithms in the stochastic setting. Bach and Moulines (2013) provide sharp regret bounds for a two-step procedure where the crucial step is the averaging of a Stochastic Gradient Descent (SGD) with constant learning rate that has to be tuned. In Gadat and Panloup (2017); Godichon-Baggioni (2018) the authors propose non-asymptotic regret bounds with large constants of the averaging of a SGD with more robust learning rates that does not need to be tuned. Our results have the same flavor on a very popular online algorithm, the Extended Kalman Filter (EKF), whose non asymptotic properties have not yet been studied.

For linear regression, Kalman filters as originally described in Kalman and Bucy (1961) present a Bayesian perspective. The idea is to estimate the conditional expectation of the future state and its variance, given a prior on the initial state and past observations that follow a dynamic model. Kalman recursion is exactly the ridge regression estimator, see Diderrich (1985), so Kalman filter achieves a $\mathcal{O}(\log(n))$ regret for quadratic losses in adversarial setting. Note that the global strong convexity of the loss is crucial in the analysis of the regret in Cesa-Bianchi and Lugosi (2006).

The Extended Kalman Filter (EKF) of Fahrmeir (1992) yields an online parameter-free algorithm for logistic regression. More generally, EKF works in any misspecified Generalized Linear Model as defined in Rigollet (2012). Recently, the equivalence between Kalman filtering under constant dynamics and Online Natural Gradient has been noticed by Ollivier (2018). It is our belief that Kalman filtering offers an optimal way to choose the step-size in an online gradient descent algorithm. Up to our knowledge, regret bounds have been derived for the batch Maximum Likelihood Estimator only, also called Follow The Leader in the online learning literature. The complexity of this batch algorithm is prohibitive, see the discussion in Hazan et al. (2007). In our paper, we view the EKF as an approximation of FTL in order to derive a $\mathcal{O}(\log(n))$ regret bound.

As an intermediate step, we prove a $\mathcal{O}(\log(n))$ regret in the logistic regression problem for a second-order algorithm between FTL and EKF. We name it the Semi-Online Step (SOS) algorithm as it requires t computations at each step, i.e. its complexity is quadratic in the number of iterations. Despite its inefficiency, SOS analysis is interesting as the non-asymptotic guarantee is valid in any adversarial setting. One can also interpret the extra t computations per iteration compared to the EKF as the cost of the estimation of the local strong convexity constant of the paraboloid approximation.

The EKF is the natural online approximation of the SOS. It is efficient (constant time per iteration) and we prove a $\mathcal{O}(\log(n))$ regret, in expectation and in the well-specified logistic regression setting only. The analysis of the regret splits in two steps. When the algorithm is close to the optimum, its regret is logarithmic with high probability. This logarithmic rate is due to the nice martingale properties of the gradients of the losses. The conditional expectation of the gradient is proportional to its quadratic variation. The logarithmic regret bound follows from the local paraboloid approximation of Hazan et al. (2007). The other phase, when the algorithm explores the optimization space, is much more problematic to analyze because the local paraboloid approximation does not apply uniformly. To circumvent this issue, we appeal at more robust potential arguments as in Gadat and Panloup (2017). We got a logarithmic control on the number of iterations spent in the first phase in expectation only. It is an open question whether this number of iterations can be controlled with high probability.

The paper is organized as follows. In Section 2, we introduce the SOS algorithm and we give its $\mathcal{O}(\log(n))$ regret in Theorem 1 followed by its proof. In Theorem 6 of Section 3, we present our result in expectation for the EKF. We present the main steps of the proof of Theorem 6 in Section 4. Finally we discuss the results and future work in Section 5.

2 Semi-Online Step algorithm

In Section 2.1, we introduce the SOS algorithm as a semi-online approximation of the batch FTL algorithm

$$\theta_t^* \in \arg\min_{\theta} \sum_{s=1}^{t-1} l_s(y_s, \theta)\,. \tag{1}$$

We see in Section 2.2 that SOS is also very close to the EKF but with higher complexity. Then we prove a bound on the regret of SOS in Section 2.3.

2.1 Construction of the SOS algorithm

The Semi-Online Step is described in Algorithm 1. We derive it from the Taylor approximation

$$\frac{\partial}{\partial\theta}\left[\sum_{s=1}^{t} l_s(y_s,\theta)\right] \approx \frac{\partial}{\partial\theta}\left[\sum_{s=1}^{t} l_s(y_s,\theta)\right]\Big|_{\theta=\theta_t^*} + \frac{\partial^2}{\partial\theta^2}\left[\sum_{s=1}^{t} l_s(y_s,\theta)\right]\Big|_{\theta=\theta_t^*}(\theta-\theta_t^*)\,,$$

which transforms the first order condition of the optimization problem (1) realized by θ_{t+1}^* into

$$\frac{\partial}{\partial\theta}\left[\sum_{s=1}^{t} l_s(y_s,\theta)\right]\Big|_{\theta=\theta_t^*} + \frac{\partial^2}{\partial\theta^2}\left[\sum_{s=1}^{t} l_s(y_s,\theta)\right]\Big|_{\theta=\theta_t^*}(\theta_{t+1}^*-\theta_t^*) \approx 0\,.$$

Algorithm 1: Semi-Online Steps

1. *Initialization*: $\tilde{P}_1$ is any positive definite matrix, $\tilde{\theta}_1$ is any initial parameter in $\mathbb{R}^d$.
2. *Iteration*: at each time step $t = 1, 2, \ldots$
   (a) Compute the matrix $\tilde{P}_{t+1} = \left(\tilde{P}_1^{-1} + \sum_{s=1}^{t} \frac{X_s X_s^T}{(1+e^{\tilde{\theta}_t^T X_s})(1+e^{-\tilde{\theta}_t^T X_s})}\right)^{-1}$:
   Starting from $\tilde{P}_{t+1}^{(0)} = \tilde{P}_1$, we compute
   $\tilde{P}_{t+1}^{(u)} = \left(\tilde{P}_1^{-1} + \sum_{s=1}^{u} \frac{X_s X_s^T}{(1+e^{\tilde{\theta}_t^T X_s})(1+e^{-\tilde{\theta}_t^T X_s})}\right)^{-1}$ thanks to the recursion
   $$\tilde{P}_{t+1}^{(u)} = \tilde{P}_{t+1}^{(u-1)} - \frac{\tilde{P}_{t+1}^{(u-1)} X_u X_u^T \tilde{P}_{t+1}^{(u-1)}}{1 + X_u^T \tilde{P}_{t+1}^{(u-1)} X_u \tilde{p}_t^{(u)}(1-\tilde{p}_t^{(u)})} \tilde{p}_t^{(u)}(1-\tilde{p}_t^{(u)}),$$
   with $\tilde{p}_t^{(u)} = 1/(1+e^{-\tilde{\theta}_t^T X_u})$ for any $u = 1, \ldots, t$, so that $\tilde{P}_{t+1} = \tilde{P}_{t+1}^{(t)}$.
   (b) Update
   $$\tilde{\theta}_{t+1} = \tilde{\theta}_t + \tilde{P}_{t+1} \frac{y_t X_t}{1 + e^{y_t \tilde{\theta}_t^T X_t}}.$$

Using the definition of θ_t^* we have

$$\frac{\partial}{\partial \theta}\left[\sum_{s=1}^{t-1} l_s(y_s, \theta_t^*)\right] = 0.$$

Combining this identity and the definition of the derivatives of the logistic loss we obtain

$$\frac{\partial}{\partial \theta}\left[\sum_{s=1}^{t} l_s(y_s, \theta)\right]\Big|_{\theta=\theta_t^*} = \frac{\partial}{\partial \theta} l_t(y_t, \theta)\Big|_{\theta=\theta_t^*} = \frac{-y_t X_t}{1 + e^{y_t \theta_t^{*T} X_t}},$$

$$\frac{\partial^2}{\partial \theta^2}\left[\sum_{s=1}^{t} l_s(y_s, \theta)\right]\Big|_{\theta=\theta_t^*} = \sum_{s=1}^{t} \frac{X_s X_s^T}{(1+e^{\theta_t^{*T} X_s})(1+e^{-\theta_t^{*T} X_s})}.$$

Therefore θ_{t+1}^* satisfies approximately

$$\left(\sum_{s=1}^{t} \frac{X_s X_s^T}{(1+e^{\theta_t^{*T} X_s})(1+e^{-\theta_t^{*T} X_s})}\right)(\theta_{t+1}^* - \theta_t^*) \approx \frac{y_t X_t}{1 + e^{y_t \theta_t^{*T} X_t}}.$$

If the Hessian matrix were invertible, we would obtain

$$\theta_{t+1}^* \approx \theta_t^* + \left(\sum_{s=1}^{t} \frac{X_s X_s^T}{(1+e^{\theta_t^{*T} X_s})(1+e^{-\theta_t^{*T} X_s})}\right)^{-1} \frac{y_t X_t}{1 + e^{y_t \theta_t^{*T} X_t}}.$$

This relation approximately satisfied by the optima sequence (θ_t^*) motivates the introduction of the SOS algorithm as defined in Algorithm 1. The computation of $\tilde{P}_{t+1}$ relies on the Sherman-Morrison formula: if $A \in \mathbb{R}^{d \times d}$ and $u, v \in \mathbb{R}^d$,

$$\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}. \tag{2}$$

We introduce the regularization matrix $\tilde{P}_1$ which guarantees the positive definiteness of $\tilde{P}_t$ in Algorithm 1. A good choice is for instance $\tilde{P}_1^{-1} = \frac{I}{p_1}$, $p_1 > 0$. SOS then corresponds to the approximation

$$\tilde{\theta}_t \approx \arg\min_{\theta}\left(\sum_{s=1}^{t-1} l_s(y_s, \theta) + \frac{1}{2p_1}\|\theta\|^2\right), \qquad t = 1, 2, \ldots.$$

Algorithm 2: Extended Kalman Filter

1. *Initialization*: P_1 is any positive definite matrix, $\hat{\theta}_1$ is any initial parameter in $\mathbb{R}^d$.
2. *Iteration*: at each time step $t = 1, 2, \ldots$
   (a) Update
   $$P_{t+1} = P_t - \frac{P_t X_t X_t^T P_t}{1 + X_t^T P_t X_t \hat{p}_t(1-\hat{p}_t)} \hat{p}_t(1-\hat{p}_t)\,,$$
   with $\hat{p}_t = 1/(1+e^{-\hat{\theta}_t^T X_t})$.
   (b) Update
   $$\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1} \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}}\,.$$

2.2 Comparison with EKF

The Extended Kalman Filter was introduced by Fahrmeir (1992) for any Dynamic Generalized Linear Model. For constant dynamics, the EKF is shown to be equivalent to the Online Natural Gradient algorithm in Ollivier (2018), yielding the recursion

$$P_{t+1}^{-1} = P_t^{-1} + \frac{X_t X_t^T}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})}\,,$$
$$\hat{\theta}_{t+1} = \hat{\theta}_t - P_{t+1} \frac{\partial}{\partial \theta} l_t(y_t, \theta)\Big|_{\theta=\hat{\theta}_t}\,.$$

This EKF recursion departs from SOS in the update of the matrix P_t which satisfies

$$P_{t+1} = \left(P_1^{-1} + \sum_{s=1}^{t} \frac{X_s X_s^T}{(1+e^{\hat{\theta}_s^T X_s})(1+e^{-\hat{\theta}_s^T X_s})} \right)^{-1}, \qquad t = 1, 2, \ldots.$$

In EKF, we add a rank-one matrix to get P_{t+1}^{-1} from P_t^{-1} in order to update the matrix efficiently. On the contrary, the matrix $\tilde{P}_t$ in SOS is recomputed at each step because the Hessian has to be computed at the current estimate $\tilde{\theta}_t$. Despite the similarity between P_t and $\tilde{P}_t$ we were not able to control their differences. Our analysis of SOS and EKF are distinct and the obtained regret bounds are different in nature.

Thanks to the Sherman-Morrison formula (2), we describe the EKF in Algorithm 2 avoiding any inversion of matrices. The spatial complexity of the two algorithms is $\mathcal{O}(d^2)$ due to the storage of the matrices P_{t+1} and $\tilde{P}_{t+1}$. In term of running time, at each step of the SOS algorithm we have to compute recursively $\tilde{P}_{t+1}^{(u)}$ for $u = 1, \ldots, t$ and then $\tilde{\theta}_{t+1}$. Each recursion on $\tilde{P}_{t+1}^{(u)}$ in u requires the computation of a rank-one matrix (product vector-vector) and its addition to the sum, its complexity is $\mathcal{O}(d^2)$. Thus, the complexity of step t in SOS is $\mathcal{O}(td^2)$. As a comparison, the EKF updates P_t online and therefore requires only $\mathcal{O}(d^2)$ operations at each step.

2.3 The regret bound for SOS and its proof

In what follows, we denote

$$D_X = \max_{1 \le t \le n} \|X_t\|,\ D_\theta = \max_{1 \le t \le n} \|\tilde{\theta}_t\|,\ D = \max_{1 \le t \le n} |\tilde{\theta}_t^T X_t|.$$

SOS offers the advantage to be easier to analyse than EKF. We prove a $\mathcal{O}(\log(n))$ regret bound on SOS in Theorem 1. Note that the leading constant is the inverse square of the exp-concavity constant times $d^{3/2} D_X (D_\theta + \|\theta\|)$. The localized algorithms of Hazan et al. (2007) satisfy finer regret bounds with the inverse of the exp-concavity constant times d as the leading constant. We believe that Theorem 1 could be improved to get a constant proportional to the inverse of the exp-concavity constant instead of the square inverse, see the end of the proof of Lemma 2 where we use a very loose bound bringing a $(1+e^D)/2$. Up to our knowledge, SOS is the first parameter-free algorithm that achieves a $\mathcal{O}(\log(n))$ regret bound in the adversarial logistic regression setting.

Theorem 1. *Starting from $\tilde{P}_1 = p_1 I$ and $\tilde{\theta}_1 \in \mathbb{R}^d$, for any $(X_t, y_t)_{1\le t\le n}$ and $\theta \in \mathbb{R}^d$, the SOS algorithm achieves the regret bound*

$$\sum_{t=1}^{n}\Big(l_t(y_t,\tilde{\theta}_t) - l_t(y_t,\theta)\Big) \le \left(\frac{\sqrt{d}D_X(D_\theta + \|\theta\|)\left(1+e^D\right)}{4} + 1\right)\frac{1+e^D}{2} d\log(1+(n-1)p_1 D_X^2) + \frac{\|\tilde{\theta}_1\|^2 + \|\theta\|^2}{2p_1} + D_X(D_\theta + \|\theta\|)\,, \qquad n \ge 1.$$

Proof. We first apply a telescopic sum argument

$$\begin{aligned}
\sum_{t=1}^{n}\Big(l_t(y_t,\tilde{\theta}_t) - l_t(y_t,\theta)\Big) &= \sum_{t=1}^{n}\left(\sum_{s=1}^{t} l_s(y_s,\tilde{\theta}_t) - \sum_{s=1}^{t-1} l_s(y_s,\tilde{\theta}_t) - l_t(y_t,\theta)\right)\\
&= \sum_{t=1}^{n-1}\left(\sum_{s=1}^{t} l_s(y_s,\tilde{\theta}_t) - \sum_{s=1}^{t} l_s(y_s,\tilde{\theta}_{t+1})\right) + \sum_{s=1}^{n}\Big(l_s(y_s,\tilde{\theta}_n) - l_s(y_s,\theta)\Big)\\
&= \sum_{t=1}^{n-1}\left(\sum_{s=1}^{t} l_s(y_s,\tilde{\theta}_t) + \frac{1}{2}\tilde{\theta}_t^T\tilde{P}_1^{-1}\tilde{\theta}_t - \sum_{s=1}^{t} l_s(y_s,\tilde{\theta}_{t+1}) - \frac{1}{2}\tilde{\theta}_{t+1}^T\tilde{P}_1^{-1}\tilde{\theta}_{t+1}\right)\\
&\quad + \sum_{s=1}^{n} l_s(y_s,\tilde{\theta}_n) + \frac{1}{2}\tilde{\theta}_n^T\tilde{P}_1^{-1}\tilde{\theta}_n - \sum_{s=1}^{n} l_s(y_s,\theta) - \frac{1}{2}\theta^T\tilde{P}_1^{-1}\theta\\
&\quad + \frac{1}{2}\theta^T\tilde{P}_1^{-1}\theta - \frac{1}{2}\tilde{\theta}_1^T\tilde{P}_1^{-1}\tilde{\theta}_1\,.
\end{aligned}$$

Then, defining $S_t(\theta) = \frac{\partial}{\partial\theta}\left[\sum_{s=1}^{t-1} l_s(y_s,\theta) + \frac{1}{2}\theta^T\tilde{P}_1^{-1}\theta\right]$, we use the convexity of $S_t + l_t$ to obtain linear bounds:

$$\begin{aligned}
\sum_{t=1}^{n}\Big(l_t(y_t,\tilde{\theta}_t) - l_t(y_t,\theta)\Big) \le &\sum_{t=1}^{n-1}\left(S_t(\tilde{\theta}_t) + \frac{\partial l_t(y_t,\theta)}{\partial\theta}\Big|_{\tilde{\theta}_t}\right)^T(\tilde{\theta}_t - \tilde{\theta}_{t+1})\\
&+ \left(S_n(\tilde{\theta}_n) + \frac{\partial l_n(y_n,\theta)}{\partial\theta}\Big|_{\tilde{\theta}_n}\right)^T(\tilde{\theta}_n - \theta)\\
&+ \frac{1}{2}\theta^T\tilde{P}_1^{-1}\theta - \frac{1}{2}\tilde{\theta}_1^T\tilde{P}_1^{-1}\tilde{\theta}_1\,.
\end{aligned}$$

We apply another telescopic argument in order to get

$$\sum_{t=1}^{n-1} S_t(\tilde{\theta}_t)^T(\tilde{\theta}_t - \tilde{\theta}_{t+1}) = \sum_{t=1}^{n-1}\Big(S_{t+1}(\tilde{\theta}_{t+1}) - S_t(\tilde{\theta}_t)\Big)^T\tilde{\theta}_{t+1} + S_1(\tilde{\theta}_1)^T\tilde{\theta}_1 - S_n(\tilde{\theta}_n)^T\tilde{\theta}_n\,.$$

As $S_1(\tilde{\theta}_1) = \tilde{P}_1^{-1}\tilde{\theta}_1$, we sum up our findings to achieve the regret bound

$$\begin{aligned}
\sum_{t=1}^{n}\Big(l_t(y_t,\tilde{\theta}_t) - l_t(y_t,\theta)\Big) \le &\sum_{t=1}^{n-1}\Big(S_{t+1}(\tilde{\theta}_{t+1}) - S_t(\tilde{\theta}_t)\Big)^T\tilde{\theta}_{t+1} - S_n(\tilde{\theta}_n)^T\theta + \frac{1}{2}\tilde{\theta}_1^T\tilde{P}_1^{-1}\tilde{\theta}_1 + \frac{1}{2}\theta^T\tilde{P}_1^{-1}\theta\\
&+ \sum_{t=1}^{n-1}\left(\frac{\partial l_t(y_t,\theta)}{\partial\theta}\Big|_{\tilde{\theta}_t}\right)^T(\tilde{\theta}_t - \tilde{\theta}_{t+1}) + \left(\frac{\partial l_n(y_n,\theta)}{\partial\theta}\Big|_{\tilde{\theta}_n}\right)^T(\tilde{\theta}_n - \theta)\,. \qquad (3)
\end{aligned}$$

Next we use the following Lemma proved in Appendix A.

Lemma 2. *For any $t = 1, 2, \ldots$, we have*

$$\Big\|S_{t+1}(\tilde{\theta}_{t+1}) - S_t(\tilde{\theta}_t)\Big\| \le \frac{\sqrt{d}D_X\left(1+e^D\right)}{4}\frac{X_t^T\tilde{P}_{t+1}X_t}{(1+e^{y_t\tilde{\theta}_t^T X_t})^2}\,.$$

Applying Lemma 2 on the norm of the first term in the previous regret bound (3), we get

$$\left\|\sum_{t=1}^{n-1}\left(S_{t+1}(\tilde{\theta}_{t+1})-S_t(\tilde{\theta}_t)\right)^T\tilde{\theta}_{t+1}\right\| \leq \sum_{t=1}^{n-1}\left\|S_{t+1}(\tilde{\theta}_{t+1})-S_t(\tilde{\theta}_t)\right\|\|\tilde{\theta}_{t+1}\|$$
$$\leq \frac{\sqrt{d}D_X D_\theta\left(1+e^D\right)}{4}\sum_{t=1}^{n-1}\frac{X_t^T\tilde{P}_{t+1}X_t}{(1+e^{y_t\hat{\theta}_t^T X_t})^2}\,.$$

Similarly, we estimate the second term of the regret bound (3) as

$$\left\|S_n(\tilde{\theta}_n)^T\theta\right\| \leq \sum_{t=1}^{n-1}\left\|S_{t+1}(\tilde{\theta}_{t+1})-S_t(\tilde{\theta}_t)\right\|\|\theta\| \leq \frac{\sqrt{d}D_X\|\theta\|\left(1+e^D\right)}{4}\sum_{t=1}^{n-1}\frac{X_t^T\tilde{P}_{t+1}X_t}{(1+e^{y_t\tilde{\theta}_t^T X_t})^2}\,.$$

Finally, we easily control the last two terms of (3) as we identify

$$\sum_{t=1}^{n-1}\left(\frac{\partial l_t(y_t,\theta)}{\partial\theta}\Big|_{\tilde{\theta}_t}\right)^T(\tilde{\theta}_t-\tilde{\theta}_{t+1}) = \sum_{t=1}^{n-1}\frac{X_t^T\tilde{P}_{t+1}X_t}{(1+e^{y_t\tilde{\theta}_t^T X_t})^2}\,,$$

and we use the upper-bound

$$\left\|\left(\frac{\partial l_n(y_n,\theta)}{\partial\theta}\Big|_{\tilde{\theta}_n}\right)^T(\tilde{\theta}_n-\theta)\right\| \leq \left\|\frac{\partial l_n(y_n,\theta)}{\partial\theta}\Big|_{\tilde{\theta}_n}\right\|(\|\tilde{\theta}_n\|+\|\theta\|) \leq D_X(D_\theta+\|\theta\|)\,.$$

Therefore,

$$\sum_{t=1}^{n}\left(l_t(y_t,\tilde{\theta}_t)-l_t(y_t,\theta)\right) \leq \left(\frac{\sqrt{d}D_X(D_\theta+\|\theta\|)\left(1+e^D\right)}{4}+1\right)\sum_{t=1}^{n-1}\frac{X_t^T\tilde{P}_{t+1}X_t}{(1+e^{y_t\tilde{\theta}_t^T X_t})^2}$$
$$+\frac{1}{2}\tilde{\theta}_1^T\tilde{P}_1^{-1}\tilde{\theta}_1+\frac{1}{2}\theta^T\tilde{P}_1^{-1}\theta+D_X(D_\theta+\|\theta\|)\,.$$

In order to conclude, we follow ideas from Cesa-Bianchi and Lugosi (2006) (in particular Lemma 11.11) to prove in Appendix A the following proposition which yields the result of Theorem 1.

Proposition 3. *For any sequence $(X_t,y_t)_{1\leq t\leq n}$ we have*

$$\sum_{t=1}^{n-1}\frac{X_t^T\tilde{P}_{t+1}X_t}{(1+e^{y_t\hat{\theta}_t^T X_t})^2} \leq \frac{1+e^D}{2}d\log(1+(n-1)p_1D_X^2)\,.$$

□

3 Extended Kalman Filter

We were not able to bound the regret of the EKF algorithm in the adversarial setting as we did not control the difference between the matrices $\tilde{P}_t$ and P_t. Thus, our EKF regret analysis holds in a restrictive well-specified stochastic setting.

3.1 Discussion on the assumptions

We assume that the stochastic sequence (X_t,y_t) follows the logistic regression model: there exists $\theta_{\text{true}}\in\mathbb{R}^d$ such that

$$p(y_t|X_t,\theta_{\text{true}}) = \frac{1}{1+e^{-y_t\theta_{\text{true}}^T X_t}}, \qquad t=1,2,\ldots. \tag{4}$$

We do not make any assumption on the dependence of the stochastic process (X_t) so far. We consider the regret in term of the expected loss conditionally on X_t: for any random variable Z, we note $\mathbb{E}_t[Z]$ the conditional expectation $\mathbb{E}[Z\mid X_1,y_1,...,X_{t-1},y_{t-1},X_t]$ (we know the past pairs $(X_s,y_s)_{s<t}$ along with X_t the explanatory variables at time t). We first observe that for any t, $\theta\to\mathbb{E}_t[l_t(y_t,\theta)]$ is a convex function minimized in θ_{true}. Even if $(\mathbb{E}_t[l_t(y_t,\theta)])$ is a stochastic sequence, we apply a convexity argument on the expected losses in order to bound the regret by a linear regret

$$\sum_{t=1}^{n}\left(\mathbb{E}_t[l_t(y_t,\hat{\theta}_t)]-\mathbb{E}_t[l_t(y_t,\theta_{\text{true}})]\right) \leq \sum_{t=1}^{n}\mathbb{E}_t\left[\frac{y_tX_t^T}{1+e^{y_t\hat{\theta}_t^T X_t}}\right](\theta_{\text{true}}-\hat{\theta}_t)\,.$$

All the regret bounds on EKF provided hereafter actually come from identical bounds on the linear regret. We identify the expected gradients in the linear regret as $\mathbb{E}_t[y_t X_t^T/(1+e^{y_t\theta^T X_t})] = \mathbb{E}_{y\sim p(y|X_t,\theta_{\text{true}})}[yX_t^T/(1+e^{y\theta^T X_t})]$. We observe a key property satisfied by the logistic gradients, proved in Appendix B.

Proposition 4. *For any $\theta, X \in \mathbb{R}^d$, there exists $c > 0$ satisfying $e^{-|(\theta_{true}-\theta)^T X|} < c < e^{|(\theta_{true}-\theta)^T X|}$ and*

$$\mathbb{E}_{y\sim p(y|X,\theta_{true})}\left[\frac{yX^T(\theta_{true}-\theta)}{1+e^{y\theta^T X}}\right] = c\frac{(\theta_{true}-\theta)^T XX^T(\theta_{true}-\theta)}{(1+e^{\theta^T X})(1+e^{-\theta^T X})}\,.$$

Such Bernstein's type conditions yield fast rates of convergence. However, the constant c in Proposition 4 is relative to the error $|(\theta_{\text{true}}-\theta)^T X|$ and the fast rate holds only locally: If there exists some τ so that $|(\theta_{\text{true}}-\hat{\theta}_t)^T X_t| \le 1/2$ for any $t > \tau$ then an application of Corollary 7 and Theorem 8 (with $\varepsilon = 0.5$ and $\alpha = 0.05$ so that $1/2+\alpha = c < e^{-\varepsilon}$) yields the following regret bound

$$\sum_{t=\tau+1}^{n+\tau}\mathbb{E}_t\left[\frac{y_t X_t^T}{1+e^{y_t\hat{\theta}_t^T X_t}}\right](\theta_{\text{true}}-\hat{\theta}_t) \le 30\Big(20(1+e^D)\log(\frac{1}{\delta}) + \frac{1+e^D}{4}d\log(1+np_1D_X^2) + \frac{1}{2p_1}\|\theta_{\text{true}}\|^2\Big)\,,$$

with probability at least $1-\delta$, $\delta > 0$.

In order to get the global regret bound, we need two extra assumptions on the law of X_t:

Assumption 1. *There exists $m_1 > 0$ such that for any $t = 1, 2, \ldots$,*

$$\frac{m_1 I}{t} \prec \mathbb{E}\left[P_{t+1}X_tX_t^T \mid X_1, y_1, \ldots, X_{t-1}, y_{t-1}\right]\,.$$

Assumption 2. *There exists $M_2 > 0$ such that for any $t = 1, 2, \ldots$,*

$$\mathbb{E}\left[X_t^T P_{t+1}^2 X_t\right] \le \frac{M_2}{t^2}\,.$$

One checks these assumptions under the invertibility of the matrix $\mathbb{E}[XX^T]$ for bounded iid (X_t):

Proposition 5. *In the iid case, if $\lambda_{\min} = \lambda_{\min}(\mathbb{E}[XX^T]) > 0$ and if $\|X\| \le D_X$ a.s. then we have*

$$\frac{\lambda_{\min}}{t(1+D_X^2)^2} \le \lambda_{\min}\left(\mathbb{E}\left[P_{t+1}X_tX_t^T \mid X_1, y_1, \ldots, X_{t-1}, y_{t-1}\right]\right)\,,$$

$$\lambda_{\max}\left(\mathbb{E}\left[P_{t+1}^2\right]\right) \le \frac{16(1+e^D)^2}{\lambda_{\min}^2 t^2}\left(1+\frac{1}{t}\frac{2de^{-3}\left(3D_X^4+D_X^2\lambda_{\min}/2\right)^3}{(\lambda_{\min}^2/8)^2}\right)\,.$$

The results of Proposition 5 imply Assumption 1 and Assumption 2. Proposition 5 is proved in Appendix B.

3.2 Regret bound in expectation for the EKF

In what follows we assume that

$$D_X \ge \max_{1\le t\le n}\|X_t\|,\ D_\theta \ge \max\left(\max_{1\le t\le n}\|\hat{\theta}_t\|, \|\theta_{\text{true}}\|\right) \text{ and } D \ge \max_{1\le t\le n}|\hat{\theta}_t^T X_t| \quad a.s.$$

It is important to note that these constants are not used in the EKF Algorithm 2, making it parameter-free.

Theorem 6. *Assume that (X_t, y_t) satisfies the logistic regression (4) for any $\theta_{\text{true}} \in \mathbb{R}^d$. If the EKF starts with $P_1 = p_1 I$, $p_1 > 0$, $\hat{\theta}_1 = 0$ and if the assumptions 1 and 2 are satisfied, we have*

$$\sum_{t=1}^{n}(\mathbb{E}[l(y,\hat{\theta}_t)] - \mathbb{E}[l(y,\theta_{true})]) \le 30\Big(20(1+e^D) + \frac{1+e^D}{4}d\log(1+np_1D_X^2) + \frac{1}{2p_1}\|\theta_{true}\|^2\Big)$$
$$+ (62D_XD_\theta + 60D_X^2D_\theta^2 + 15D_X^2)\Big(1+\frac{4^{k+1}D_X^{2k}b_k}{ka-1}\log(n)\Big), \qquad n \ge 1\,.$$

Here $a = e^{-D}m_1/(1+e^D) < 1$, $b_k = \frac{5M_2}{p_1^2D_X^2}\left(4D_\theta^2 + 2p_1D_\theta D_X + p_1^2D_X^2\right)^k$ and k is any positive integer satisfying $1 < ka < 2$.

Note that m_1 as defined in Assumption 1 may be chosen such that $a < 1$.

4 The sketch of the proof of Theorem 6

One has to distinguish between the localized steps where $|(\theta_{\text{true}} - \hat{\theta}_t)^T X_t| < \varepsilon$ and the others. To this end, we define $T_\varepsilon = \{1 \leq t \leq n \mid |(\hat{\theta}_t - \theta_{\text{true}})^T X_t| \leq \varepsilon\}$. In Section 4.1 we exhibit a bound on the sum of the localized terms $t \in T_\varepsilon$, and in Section 4.2 we upper-bound the expected value of the non-localized terms $t \notin T_\varepsilon$. Finally Section 4.3 merges these results to prove Theorem 6. The proofs of the intermediate results used in Sections 4.1, 4.2 and 4.3 are deferred to Appendix B.

4.1 Bounding the localized steps with high probability

In the well-specified logistic regression, Proposition 4 provides an upper-bound on the linearized regret

$$\mathbb{E}_t\left[\frac{y_t X_t^T}{1+e^{y_t \hat{\theta}_t^T X_t}}\right](\theta_{\text{true}} - \hat{\theta}_t) = \mathbb{E}_{y\sim p(y|X_t,\theta_{\text{true}})}\left[\frac{y X_t^T}{1+e^{y \hat{\theta}_t^T X_t}}\right](\theta_{\text{true}} - \theta_t)\,.$$

The following corollary of Proposition 4 provides a simple estimate for localized steps:

Corollary 7. *For any step $t = 1, 2, \ldots$, if we have $|(\theta_{true} - \hat{\theta}_t)^T X_t| < \varepsilon$ and $c < e^{-\varepsilon}$ then it holds*

$$\mathbb{E}_t\left[\frac{y_t X_t^T}{1+e^{y_t \hat{\theta}_t^T X_t}}\right](\theta_{true} - \hat{\theta}_t) < \frac{e^\varepsilon}{e^{-\varepsilon} - c}\left(\mathbb{E}_t\left[\frac{y_t X_t^T}{1+e^{y_t \hat{\theta}_t^T X_t}}\right](\theta_{true} - \hat{\theta}_t) - c\frac{(\theta_{true} - \hat{\theta}_t)^T X_t X_t^T (\theta_{true} - \hat{\theta}_t)}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})}\right).$$

The upper-bound is controlled thanks to the negative quadratic term which is responsible of the fast rate of convergence. It can be seen as a local strong convexity term. We derive the following regret bound with high probability, parameterizing $c = 1/2 + \alpha$ for some $\alpha > 0$:

Theorem 8. *For any $\varepsilon, \alpha, \delta > 0$ and starting the EKF from $P_1 = p_1 I$, $p_1 > 0$ and $\hat{\theta}_1 = 0$, we have*

$$\begin{aligned}\sum_{t\in T_\varepsilon}\Big(\mathbb{E}_t\left[\frac{y_t X_t^T}{1+e^{y_t \hat{\theta}_t^T X_t}}\right](\theta_{true} - \hat{\theta}_t) - \big(\frac{1}{2}+\alpha\big)\frac{(\theta_{true} - \hat{\theta}_t)^T X_t X_t^T (\theta_{true} - \hat{\theta}_t)}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})}\Big)\\ \leq \frac{1+e^D}{\alpha}\log(\delta^{-1}) + \frac{1+e^D}{4} d\log(1+np_1 D_X^2) + \frac{1}{2p_1}\|\theta_{true}\|^2\\ - \frac{1}{2}\sum_{t\notin T_\varepsilon}\Big((\hat{\theta}_t - \theta_{true})^T P_t^{-1}(\hat{\theta}_t - \theta_{true}) - (\hat{\theta}_{t+1} - \theta_{true})^T P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta_{true})\Big),\end{aligned}$$

with probability at least $1 - \delta$.

Proof. We begin with a lemma whose proof is very much inspired by the proof of the regret bound of the ONS algorithm in Hazan (2016). We note that the constant $1/2$ in front of the quadratic term is responsible to the fast rate of convergence, as there exists a gap with respect to the constant $c \approx 1$ in Proposition 4.

Lemma 9. *For any $\varepsilon > 0$ and any sequence (X_t, y_t), starting the EKF from $P_1 = p_1 I$, $p_1 > 0$ and $\hat{\theta}_1 = 0$, we have*

$$\begin{aligned}\sum_{t\in T_\varepsilon}\Big(\frac{y_t X_t^T}{1+e^{y_t \hat{\theta}_t^T X_t}}(\theta_{true} - \hat{\theta}_t) - \frac{1}{2}(\theta_{true} - \hat{\theta}_t)^T (P_{t+1}^{-1} - P_t^{-1})(\theta_{true} - \hat{\theta}_t)\Big)\\ \leq \frac{1+e^D}{4} d\log(1+np_1 D_X^2) + \frac{1}{2p_1}\|\theta_{true}\|^2\\ - \frac{1}{2}\sum_{t\notin T_\varepsilon}\Big((\hat{\theta}_t - \theta_{true})^T P_t^{-1}(\hat{\theta}_t - \theta_{true}) - (\hat{\theta}_{t+1} - \theta_{true})^T P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta_{true})\Big).\end{aligned}$$

Then we prove the following lemma which is a corollary of a martingale inequality from Bercu and Touati (2008):

Lemma 10. *We define* $\Delta M_t = \mathbb{E}_t\left[\frac{y_t X_t^T}{1+e^{y_t\hat{\theta}_t^T X_t}}\right](\theta_{true}-\hat{\theta}_t) - \frac{y_t X_t^T}{1+e^{y_t\hat{\theta}_t^T X_t}}(\theta_{true}-\hat{\theta}_t)$. *Then for any* $\varepsilon > 0$ *and* $\alpha > 0$, *it holds*

$$\sum_{t\in T_\varepsilon}\left(\Delta M_t - \alpha\frac{(\theta_{true}-\hat{\theta}_t)^T X_t X_t^T(\theta_{true}-\hat{\theta}_t)}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})}\right) \leq \frac{1+e^D}{\alpha}\log(\delta^{-1}),$$

with probability at least $1-\delta$.

Adding the inequalities of Lemma 9 and Lemma 10 yield the result. □

From Theorem 8 it is easy to bound the linearized regret of the localized steps in expectation:

Corollary 11. *For any* $\varepsilon, \alpha > 0$, *we have*

$$\begin{aligned}\sum_{t\in T_\varepsilon}\Big(\mathbb{E}\left[\frac{y_t X_t^T}{1+e^{y_t\hat{\theta}_t^T X_t}}(\theta_{true}-\hat{\theta}_t)\right] - (\frac{1}{2}+\alpha)\frac{(\theta_{true}-\hat{\theta}_t)^T X_t X_t^T(\theta_{true}-\hat{\theta}_t)}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})}\Big)\\ \leq \frac{1+e^D}{\alpha} + \frac{1+e^D}{4}d\log(1+np_1D_X^2) + \frac{1}{2p_1}\|\theta_{true}\|^2\\ - \frac{1}{2}\sum_{t\notin T_\varepsilon}\Big((\hat{\theta}_t-\theta_{true})^T P_t^{-1}(\hat{\theta}_t-\theta_{true}) - (\hat{\theta}_{t+1}-\theta_{true})^T P_{t+1}^{-1}(\hat{\theta}_{t+1}-\theta_{true})\Big).\end{aligned}$$

4.2 Bounding the expected number of unlocalized steps

It is essential in our proof to lower-bound the cardinal of T_ε because the remaining terms in the regret of Theorem 6 are essentially controlled by it:

$$\begin{aligned}\sum_{t\notin T_\varepsilon}\Big\|\mathbb{E}_t\Big[\frac{y_t X_t^T}{1+e^{y_t\hat{\theta}_t^T X_t}}(\theta_{\text{true}}-\hat{\theta}_t)\Big]\\ - \frac{e^\varepsilon}{e^{-\varepsilon}-(\frac{1}{2}+\alpha)}\frac{1}{2}\Big((\hat{\theta}_t-\theta_{\text{true}})^T P_t^{-1}(\hat{\theta}_t-\theta_{\text{true}}) - (\hat{\theta}_{t+1}-\theta_{\text{true}})^T P_{t+1}^{-1}(\hat{\theta}_{t+1}-\theta_{\text{true}})\Big)\Big\|\\ \leq \left(2D_XD_\theta + \frac{e^\varepsilon}{e^{-\varepsilon}-(\frac{1}{2}+\alpha)}(2D_XD_\theta + 2D_X^2D_\theta^2 + \frac{1}{2}D_X^2)\right)(n - Card(T_\varepsilon)).\end{aligned} \quad (5)$$

We bound the number of non-localized steps $n - Card(T_\varepsilon)$ in expectation only, yielding the regret bound in Theorem 6 in expectation.

Theorem 12. *We define* $a = e^{-D}m_1/(1+e^D)$. *For any positive integer* k *such that* $1 < ka < 2$, *provided that Assumptions 1 and 2 are satisfied, we have*

$$\mathbb{E}\left[n - Card(T_\varepsilon)\right] \leq 1 + \frac{4D_X^{2k}b_k}{\varepsilon^{2k}(ka-1)}\log(n).$$

with $b_k = \frac{5M_2}{p_1^2D_X^2}\left(4D_\theta^2 + 2p_1D_\theta D_X + p_1^2D_X^2\right)^k$.

Proof. We first find a recursive bound on $\mathbb{E}\left[\|\hat{\theta}_t - \theta_{\text{true}}\|^{2k}\right]$ for any positive integer k:

Lemma 13. *Provided that Assumptions 1 and 2 are satisfied, it holds*

$$\mathbb{E}\left[\|\hat{\theta}_{t+1}-\theta_{true}\|^{2k}\right] \leq \mathbb{E}\left[\|\hat{\theta}_t-\theta_{true}\|^{2k}\right]\left(1-\frac{ka}{t}\right) + \frac{b_k}{t^2}, \qquad t \geq 1.$$

It is easy to derive from this lemma the following corollary:

Corollary 14. *If Assumptions 1 and 2 are satisfied and* $1 < ka < 2$ *then we have*

$$\mathbb{E}\left[\|\hat{\theta}_t-\theta_{true}\|^{2k}\right] \leq \frac{4b_k}{t(ka-1)}, \qquad t \geq 2.$$

Then, using first the Cauchy-Schwarz inequality $|(\hat{\theta}_t - \theta_{\text{true}})^T X_t| \leq \|\hat{\theta}_t - \theta_{\text{true}}\| \|X_t\|$ and second the Markov inequality, we get

$$\begin{aligned}
\mathbb{P}\left(|(\hat{\theta}_t - \theta_{\text{true}})^T X_t| > \varepsilon\right) &\leq \mathbb{P}\left(\|\hat{\theta}_t - \theta_{\text{true}}\| > \frac{\varepsilon}{\|X_t\|}\right) \\
&\leq \frac{\mathbb{E}\left[\|\hat{\theta}_t - \theta_{\text{true}}\|^{2k}\right]}{\varepsilon^{2k}/\|X_t\|^{2k}} \\
&\leq \frac{4\|X_t\|^{2k} b_k}{\varepsilon^{2k}(ka-1)} \frac{1}{t}\,.
\end{aligned}$$

which proves the Theorem by a summation argument for $2 \leq t \leq n$ together with the trivial bound $\mathbb{P}\left(|(\hat{\theta}_1 - \theta_{\text{true}})^T X_1| > \varepsilon\right) \leq 1$. □

4.3 Proof of Theorem 6

Summing Corollary 11 and Equation 5 along with Theorem 12 yields the regret bound in expectation

$$\begin{aligned}
&\sum_{t=1}^{n} \left(\mathbb{E}[l(y, \hat{\theta}_t)] - \mathbb{E}[l(y, \theta_{\text{true}})]\right) \\
&\leq \frac{e^{\varepsilon}}{e^{-\varepsilon} - (\frac{1}{2} + \alpha)} \Big(\frac{1 + e^{D}}{\alpha} + \frac{1 + e^{D}}{4} d \log(1 + n p_1 D_X^2) + \frac{1}{2p_1} \|\theta_{\text{true}}\|^2\Big) \\
&\quad + \Big(2 D_X D_\theta + \frac{e^{\varepsilon}}{e^{-\varepsilon} - (\frac{1}{2} + \alpha)} (2 D_X D_\theta + 2 D_X^2 D_\theta^2 + \frac{1}{2} D_X^2)\Big)\Big(1 + \Big(\frac{D_X}{\varepsilon}\Big)^{2k} \frac{4 b_k}{ka - 1} \log(n)\Big)\,,
\end{aligned}$$

for any $0 < \alpha < \frac{1}{2}$ and $0 < \varepsilon < -\log(1/2 + \alpha)$. Choosing $\varepsilon = 0.5$ and $\alpha = 0.05$, we get $\frac{e^{\varepsilon}}{e^{-\varepsilon} - (\frac{1}{2} + \alpha)} \approx 29.2 < 30$ and we obtain the result of Theorem 6.

5 Conclusion and future work

We have designed an algorithm called SOS that is a second-order algorithm very close to the EKF. We obtain a compromise between time complexity and regret guarantee. Indeed, its complexity lies between FTL, which is computationally greedy, and the EKF, which is the least expensive second-order algorithm. Moreover, we prove that SOS achieves the optimal $\mathcal{O}(\log(n))$ regret in the adversarial setting. The dependence of the constant of Theorem 1 on the exp-concavity constant might be reduced.

An interesting challenge is to adjust this regret bound for the EKF. In this paper we obtained weaker guarantees for the EKF, a $\mathcal{O}(\log(n))$ regret bound with prohibitive constants, in expectation and in the well-specified logistic regression. It would be interesting to obtain the result in the misspecified setting. Another improvement would be to find conditions for the convergence of $\sum_t \mathbb{P}(|(\hat{\theta}_t - \theta_{\text{true}})^T X_t| > \varepsilon)$, transforming our regret in expectation to a regret with high probability.

Also, as though we focused on logistic regression, it seems to us that our approach might be applied to any Generalized Linear Model.

References

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Advances in neural information processing systems*, pages 773–781, 2013.

B. Bercu and A. Touati. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18(5):1848–1869, 2008.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge university press, 2006.

G. T. Diderrich. The Kalman filter from the perspective of Goldberger–Theil estimators. *The American Statistician*, 39(3):193–198, 1985.

L. Fahrmeir. Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association*, 87(418):501–509, 1992.

S. Gadat and F. Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. arXiv preprint arXiv:1709.03342, 2017.

A. Godichon-Baggioni. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. arxiv preprint arxiv:1609.05479, 2018.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

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

R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of basic engineering*, 83 (1):95–108, 1961.

Y. Ollivier. Online natural gradient as a Kalman filter. *Electronic Journal of Statistics*, 12(2):2930–2961, 2018.

P. Rigollet. Kullback–Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2): 639–665, 2012.

J. A. Tropp. User-friendly tail bounds for matrix martingales. Technical report, CALIFORNIA INST OF TECH PASADENA, 2011.

A Details for the proof of Theorem 1

Proof of Lemma 2. Let $S_t^{(i)}(\theta) = \frac{\partial}{\partial \theta_i}\left[\sum_{s=1}^{t-1} l_s(y_s,\theta) + \frac{1}{2}\theta^T \tilde{P}_1^{-1}\theta\right]$ be the ith coordinate of $S_t(\theta)$. We prove that

$$\left|S_{t+1}^{(i)}(\tilde{\theta}_{t+1}) - S_t^{(i)}(\tilde{\theta}_t)\right| \leq \frac{D_X\left(1+e^D\right)}{4}\frac{X_t^T\tilde{P}_{t+1}X_t}{(1+e^{y_t\tilde{\theta}_t^T X_t})^2}. \tag{6}$$

Applying a Taylor expansion, there exists $0 < \alpha_t^{(i)} < 1$ satisfying

$$S_{t+1}^{(i)}(\tilde{\theta}_{t+1}) - S_{t+1}^{(i)}(\tilde{\theta}_t) - \left(\frac{\partial S_{t+1}^{(i)}(\theta)}{\partial\theta}\Big|_{\tilde{\theta}_t}\right)^T(\tilde{\theta}_{t+1}-\tilde{\theta}_t) = \frac{1}{2}(\tilde{\theta}_{t+1}-\tilde{\theta}_t)^T\left(\frac{\partial^2 S_{t+1}^{(i)}(\theta)}{\partial\theta^2}\Big|_{\tilde{\theta}_t+\alpha_t^{(i)}(\tilde{\theta}_{t+1}-\tilde{\theta}_t)}\right)(\tilde{\theta}_{t+1}-\tilde{\theta}_t).$$

Thanks to the update of $\tilde{\theta}_{t+1}$, we have the relation

$$S_{t+1}^{(i)}(\tilde{\theta}_t) - S_t^{(i)}(\tilde{\theta}_t) + \left(\frac{\partial S_{t+1}^{(i)}(\theta)}{\partial\theta}\Big|_{\tilde{\theta}_t}\right)^T(\tilde{\theta}_{t+1}-\tilde{\theta}_t) = 0.$$

Therefore, summing last two identities, we get

$$S_{t+1}^{(i)}(\tilde{\theta}_{t+1}) - S_t^{(i)}(\tilde{\theta}_t) = \frac{1}{2}(\tilde{\theta}_{t+1}-\tilde{\theta}_t)^T\left(\frac{\partial^2 S_{t+1}^{(i)}(\theta)}{\partial\theta^2}\Big|_{\tilde{\theta}_t+\alpha_t^{(i)}(\tilde{\theta}_{t+1}-\tilde{\theta}_t)}\right)(\tilde{\theta}_{t+1}-\tilde{\theta}_t).$$

From the definition of the logistic loss, we identify

$$\frac{\partial^2 S_{t+1}^{(i)}(\theta)}{\partial\theta^2} = -\sum_{s=1}^{t}\frac{X_s^{(i)}X_sX_s^T(e^{\theta^T X_s}-e^{-\theta^T X_s})}{(1+e^{\theta^T X_s})^2(1+e^{-\theta^T X_s})^2}, \qquad \theta\in\mathbb{R}^d.$$

Using the bound $\frac{|e^{\theta^T X_s}-e^{-\theta^T X_s}|}{(1+e^{\theta^T X_s})(1+e^{-\theta^T X_s})} < 1$, we have

$$\left|\frac{1}{2}(\tilde{\theta}_{t+1}-\tilde{\theta}_t)^T\left(\frac{\partial^2 S_{t+1}^{(i)}(\theta)}{\partial\theta^2}\Big|_{\tilde{\theta}_t+\alpha(\tilde{\theta}_{t+1}-\tilde{\theta}_t)}\right)(\tilde{\theta}_{t+1}-\tilde{\theta}_t)\right| \leq \frac{D_X}{2}(\tilde{\theta}_{t+1}-\tilde{\theta}_t)^T$$
$$\left(\sum_{s=1}^{t}\frac{X_sX_s^T}{(1+e^{(\tilde{\theta}_t+\alpha^{(i)}(\tilde{\theta}_{t+1}-\tilde{\theta}_t))^T X_s})(1+e^{-(\tilde{\theta}_t+\alpha^{(i)}(\tilde{\theta}_{t+1}-\tilde{\theta}_t))^T X_s})}\right)(\tilde{\theta}_{t+1}-\tilde{\theta}_t).$$

Noticing that

$$\frac{1}{(1+e^{(\tilde{\theta}_t+\alpha^{(i)}(\tilde{\theta}_{t+1}-\tilde{\theta}_t))^T X_s})(1+e^{-(\tilde{\theta}_t+\alpha^{(i)}(\tilde{\theta}_{t+1}-\tilde{\theta}_t))^T X_s})} < \frac{1+e^D}{2}\frac{1}{(1+e^{\tilde{\theta}_t^T X_s})(1+e^{-\tilde{\theta}_t^T X_s})},$$

we obtain

$$\left(\sum_{s=1}^{t}\frac{X_s X_s^T}{(1+e^{(\tilde{\theta}_t+\alpha^{(i)}(\tilde{\theta}_{t+1}-\tilde{\theta}_t))^T X_s})(1+e^{-(\tilde{\theta}_t+\alpha^{(i)}(\tilde{\theta}_{t+1}-\tilde{\theta}_t))^T X_s})}\right) \prec \frac{1+e^D}{2}\tilde{P}_{t+1}^{-1}.$$

Combining our findings with the updates $\tilde{\theta}_{t+1}-\tilde{\theta}_t = \tilde{P}_{t+1}\frac{y_t X_t}{1+e^{y_t\tilde{\theta}_t^T X_t}}$ yields Eqn. (6). The desired result follows easily as

$$\left\|S_{t+1}(\tilde{\theta}_{t+1}) - S_t(\tilde{\theta}_t)\right\| \leq \sqrt{d}\max_{1\leq i\leq d}\left|S_{t+1}^{(i)}(\tilde{\theta}_{t+1}) - S_t^{(i)}(\tilde{\theta}_t)\right|.$$

□

Proof of Proposition 3. For any $1 \leq s \leq t \leq n$ we have $\frac{1}{(1+e^{\tilde{\theta}_t^T X_s})(1+e^{-\tilde{\theta}_t^T X_s})} > \frac{1}{2(1+e^D)}$ and

$$\frac{X_s X_s^T}{(1+e^{\tilde{\theta}_t^T X_s})(1+e^{-\tilde{\theta}_t^T X_s})} \succ \frac{X_s X_s^T}{2(1+e^D)}.$$

Summing this inequality along with $1 > \frac{1}{2(1+e^D)}$ and $\tilde{P}_1^{-1} \succ 0$ yields

$$\tilde{P}_{t+1}^{-1} = \tilde{P}_1^{-1} + \sum_{s=1}^{t}\frac{X_s X_s^T}{(1+e^{\tilde{\theta}_t^T X_s})(1+e^{-\tilde{\theta}_t^T X_s})} \succ \frac{1}{2(1+e^D)}\left(\tilde{P}_1^{-1} + \sum_{s=1}^{t} X_s X_s^T\right).$$

Using that if A and B are positive definite matrices, $A \succ B \implies A^{-1} \prec B^{-1}$, we get

$$\tilde{P}_{t+1} \prec 2(1+e^D)\left(\tilde{P}_1^{-1} + \sum_{s=1}^{t} X_s X_s^T\right)^{-1}.$$

We then apply Lemma 11.11 of Cesa-Bianchi and Lugosi (2006) to get

$$\begin{aligned} X_t^T\left(\tilde{P}_1^{-1} + \sum_{s=1}^{t} X_s X_s^T\right)^{-1} X_t &= 1 - \frac{\det\left(\tilde{P}_1^{-1} + \sum_{s=1}^{t-1} X_s X_s^T\right)}{\det\left(\tilde{P}_1^{-1} + \sum_{s=1}^{t} X_s X_s^T\right)} \\ &\leq \log\left(\frac{\det\left(\tilde{P}_1^{-1} + \sum_{s=1}^{t} X_s X_s^T\right)}{\det\left(\tilde{P}_1^{-1} + \sum_{s=1}^{t-1} X_s X_s^T\right)}\right), \end{aligned}$$

thanks to the inequality $1 - x \leq \log(1/x)$ for any $x > 0$. As $\frac{1}{(1+e^{\tilde{\theta}_t^T X_t})(1+e^{-\tilde{\theta}_t^T X_t})} \leq \frac{1}{4}$, we get

$$\frac{X_t^T \tilde{P}_{t+1} X_t}{(1+e^{\tilde{\theta}_t^T X_t})(1+e^{-\tilde{\theta}_t^T X_t})} \leq \frac{1+e^D}{2}\log\left(\frac{\det\left(\tilde{P}_1^{-1} + \sum_{s=1}^{t} X_s X_s^T\right)}{\det\left(\tilde{P}_1^{-1} + \sum_{s=1}^{t-1} X_s X_s^T\right)}\right).$$

Summing from 1 to $n-1$, we obtain

$$\begin{aligned} \sum_{t=1}^{n-1}\frac{X_t^T \tilde{P}_{t+1} X_t}{(1+e^{\tilde{\theta}_t^T X_t})(1+e^{-\tilde{\theta}_t^T X_t})} &\leq \frac{1+e^D}{2}\log\left(\frac{\det\left(\tilde{P}_1^{-1} + \sum_{s=1}^{n-1} X_s X_s^T\right)}{\det(\tilde{P}_1^{-1})}\right) \\ &= \frac{1+e^D}{2}\log\left(\det\left(I + p_1\sum_{s=1}^{n-1} X_s X_s^T\right)\right), \end{aligned}$$

because $\tilde{P}_1 = p_1 I$. The biggest eigenvalue of $\left(I + p_1 \sum_{s=1}^{n-1} X_s X_s^T\right)$ is bounded by $1 + p_1 \sum_{s=1}^{n-1} X_s^T X_s$. We get

$$\det\left(I + p_1 \sum_{s=1}^{n-1} X_s X_s^T\right) \leq \left(1 + p_1 \sum_{s=1}^{n-1} X_s^T X_s\right)^d \leq \left(1 + p_1(n-1)D_X^2\right)^d.$$

and that concludes the proof. □

B Details for the proof of Theorem 6

Proof of Proposition 4. We develop the expectation

$$\mathbb{E}_{y\sim p(y|X,\theta_{\text{true}})}\left[\frac{yX^T(\theta_{\text{true}} - \theta)}{1 + e^{y\theta^T X}}\right] = \frac{(\theta_{\text{true}} - \theta)^T X}{(1 + e^{\theta^T X})(1 + e^{-\theta^T X})}\left[\frac{1 + e^{-\theta^T X}}{1 + e^{-\theta_{\text{true}}^T X}} - \frac{1 + e^{\theta^T X}}{1 + e^{\theta_{\text{true}}^T X}}\right].$$

Therefore, we bound $\left[\frac{1+e^{-\theta^T X}}{1+e^{-\theta_{\text{true}}^T X}} - \frac{1+e^{\theta^T X}}{1+e^{\theta_{\text{true}}^T X}}\right]$ in terms of $X^T(\theta_{\text{true}} - \theta)$. We first rearrange it:

$$\begin{aligned}\frac{1 + e^{-\theta^T X}}{1 + e^{-\theta_{\text{true}}^T X}} - \frac{1 + e^{\theta^T X}}{1 + e^{\theta_{\text{true}}^T X}} &= \left(1 + \frac{e^{-\theta^T X} - e^{-\theta_{\text{true}}^T X}}{1 + e^{-\theta_{\text{true}}^T X}}\right) - \left(1 + \frac{e^{\theta^T X} - e^{\theta_{\text{true}}^T X}}{1 + e^{\theta_{\text{true}}^T X}}\right) \\ &= \frac{e^{(\theta_{\text{true}} - \theta)^T X} - 1}{1 + e^{\theta_{\text{true}}^T X}} - \frac{e^{-(\theta_{\text{true}} - \theta)^T X} - 1}{1 + e^{-\theta_{\text{true}}^T X}}.\end{aligned}$$

There exists $|\alpha| \leq |(\theta_{\text{true}} - \theta)^T X|$ such that

$$\frac{e^{(\theta_{\text{true}} - \theta)^T X} - 1}{1 + e^{\theta_{\text{true}}^T X}} - \frac{e^{-(\theta_{\text{true}} - \theta)^T X} - 1}{1 + e^{-\theta_{\text{true}}^T X}} = \left[\frac{e^{\alpha}}{1 + e^{\theta_{\text{true}}^T X}} + \frac{e^{-\alpha}}{1 + e^{-\theta_{\text{true}}^T X}}\right] X^T(\theta_{\text{true}} - \theta).$$

As the function $x \to \frac{a}{1+x} + \frac{b}{1+1/x}$ is monotonic with limits a and b in 0 and $+\infty$ respectively, we get

$$e^{-(\theta_{\text{true}} - \theta)^T X} < \frac{e^{\alpha}}{1 + e^{\theta_{\text{true}}^T X}} + \frac{e^{-\alpha}}{1 + e^{-\theta_{\text{true}}^T X}} < e^{(\theta_{\text{true}} - \theta)^T X}.$$

□

Proof of Proposition 5. We define $S_{t+1} = \left(I + \sum_{s=1}^{t} X_s X_s^T\right)^{-1}$. As $S_t \prec P_t \prec 2(1 + e^D)S_t$, we give the desired results first on S_t and the desired results on P_t follow easily.

We first give a lower bound on $\mathbb{E}[S_{t+1} X_t X_t^T \mid X_1, y_1, \ldots, X_{t-1}, y_{t-1}]$. Using the relation $S_{t+1} = S_t - \frac{S_t X_t X_t^T S_t}{1 + X_t^T S_t X_t}$, we write

$$S_{t+1} X_t X_t^T = S_t X_t X_t^T - \frac{S_t X_t X_t^T S_t}{1 + X_t^T S_t X_t} X_t X_t^T = \frac{1}{1 + X_t^T S_t X_t} S_t X_t X_t^T.$$

Noting that $S_t \succcurlyeq I$, $\|X_t\| \leq D_X$, and $S_t X_t X_t^T \succcurlyeq 0$ as a rank-one matrix with eigenvalue $X_t^T S_t X_t > 0$, we get $S_{t+1} X_t X_t^T \succcurlyeq \frac{S_t X_t X_t^T}{1 + D_X^2}$. It implies

$$\begin{aligned}\mathbb{E}[S_{t+1} X_t X_t^T \mid X_1, y_1, \ldots, X_{t-1}, y_{t-1}] &\succcurlyeq \frac{1}{1 + D_X^2} \mathbb{E}[S_t X_t X_t^T \mid X_1, y_1, \ldots, X_{t-1}, y_{t-1}] \\ &= \frac{1}{1 + D_X^2} S_t \mathbb{E}[X_t X_t^T \mid X_1, y_1, \ldots, X_{t-1}, y_{t-1}].\end{aligned}$$

The independence hypothesis yields $\mathbb{E}[X_t X_t^T \mid X_1, y_1, \ldots, X_{t-1}, y_{t-1}] = \mathbb{E}[XX^T]$. Also, from $\lambda_{\max}(S_t^{-1}) \leq 1 + (t-1)D_X^2$ we obtain $\lambda_{\min}(S_t) \geq \frac{1}{1+(t-1)D_X^2} \geq \frac{1}{t(1+D_X^2)}$. Therefore

$$\lambda_{\min}(\mathbb{E}[S_{t+1} X_t X_t^T \mid X_1, y_1, \ldots, X_{t-1}, y_{t-1}]) \geq \frac{1}{1 + D_X^2} \lambda_{\min}(S_t) \lambda_{\min}(\mathbb{E}[XX^T]) \geq \frac{\lambda_{\min}}{t(1 + D_X^2)^2}.$$

In order to get an upper bound on $\lambda_{\max}(\mathbb{E}[P_{t+1}^2])$, we first bound $\mathbb{P}\left(\lambda_{\max}(S_t) > \frac{2}{t\lambda_{\min}}\right)$. Then we estimate $\lambda_{\max}(\mathbb{E}[P_{t+1}^2]) \leq \mathbb{E}[\lambda_{\max}(P_{t+1}^2)] \leq 4(1+e^D)^2\mathbb{E}[\lambda_{\max}(S_t)^2]$ with

$$\begin{aligned}\mathbb{E}[\lambda_{\max}(S_t)^2] &= \mathbb{E}[\lambda_{\max}(S_t)^2 \mathbb{1}_{\lambda_{\max}(S_t) > \frac{2}{t\lambda_{\min}}}] + \mathbb{E}[\lambda_{\max}(S_t)^2 \mathbb{1}_{\lambda_{\max}(S_t) \leq \frac{2}{t\lambda_{\min}}}] \\ &\leq \mathbb{P}\left(\lambda_{\max}(S_t) > \frac{2}{t\lambda_{\min}}\right) + \frac{4}{t^2\lambda_{\min}^2},\end{aligned}$$

because $\lambda_{\max}(S_t) \leq 1$ and $\mathbb{P}\left(\lambda_{\max}(S_t) \leq \frac{2}{t\lambda_{\min}}\right) \leq 1$.

We control the deviations of $\lambda_{\max}(S_t)$ first by centering as

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}(S_t) > \frac{2}{t\lambda_{\min}}\right) &= \mathbb{P}\left(\lambda_{\min}(S_t^{-1}) < \frac{\lambda_{\min}}{2}t\right) \\ &= \mathbb{P}\left(\lambda_{\min}\Big(\sum_{s=1}^t X_s X_s^T\Big) < \frac{\lambda_{\min}}{2}t - 1\right) \\ &\leq \mathbb{P}\left(\lambda_{\min}\Big(\sum_{s=1}^t X_s X_s^T\Big) < \frac{\lambda_{\min}}{2}t\right) \\ &= \mathbb{P}\left(\lambda_{\min}\Big(\sum_{s=1}^t X_s X_s^T\Big) - t\lambda_{\min}(\mathbb{E}[XX^T]) < -\frac{\lambda_{\min}}{2}t\right).\end{aligned}$$

Then we want to show that $X_s X_s^T$ and $\mathbb{E}[XX^T]$ are commuting in order to rewrite the centered smallest eigenvalue as the smallest eigenvalue of a centered matrix and apply the Bernstein inequality of Tropp (2011) on it. We note that $X_s X_s^T \mathbb{E}[XX^T]$ is a rank-one matrix with eigenvalue $X_s^T \mathbb{E}[XX^T] X_s > 0$ if $X_s \neq 0$. Therefore $X_s X_s^T \mathbb{E}[XX^T] \succcurlyeq 0$ is symmetric and $X_s X_s^T \mathbb{E}[XX^T] = (X_s X_s^T \mathbb{E}[XX^T])^T = \mathbb{E}[XX^T] X_s X_s^T$. Similarly, we get that $\sum_1^t X_s X_s^T$ and $\mathbb{E}[XX^T]$ are commuting, thus they are simultaneously diagonalizable. Using their joint diagonalization, we infer that

$$\lambda_{\min}\left(\sum_{s=1}^t (X_s X_s^T - \mathbb{E}[XX^T])\right) \leq \lambda_{\min}\left(\sum_{s=1}^t X_s X_s^T\right) - \lambda_{\min}\left(\sum_{s=1}^t \mathbb{E}[XX^T]\right).$$

Combining those results, we obtain that

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}(S_t) > \frac{2}{t\lambda_{\min}}\right) &\leq \mathbb{P}\left(\lambda_{\min}\left(\sum_{s=1}^t (X_s X_s^T - \mathbb{E}[XX^T])\right) < -\frac{\lambda_{\min}}{2}t\right) \\ &= \mathbb{P}\left(\lambda_{\max}\left(\sum_{s=1}^t (\mathbb{E}[XX^T] - X_s X_s^T)\right) > \frac{\lambda_{\min}}{2}t\right).\end{aligned}$$

We apply Theorem 1.3 of Tropp (2011) which is a Bernstein inequality on the largest eigenvalue of sums of independent centered matrices. We check the conditions:

- $\mathbb{E}\left[\mathbb{E}[XX^T] - X_s X_s^T\right] = 0$,
- $\lambda_{\max}(\mathbb{E}[XX^T] - X_s X_s^T) \leq D_X^2$ a.s.,
- from $\|X_s\| \leq D_X^2$ as $0 \preccurlyeq \mathbb{E}\left[(\mathbb{E}[XX^T] - X_s X_s^T)^2\right] \preccurlyeq \mathbb{E}\left[(X_s X_s^T)^2\right]$, the largest singular value of $\mathbb{E}\left[(\mathbb{E}[XX^T] - X_s X_s^T)^2\right]$ is upper-bounded by D_X^4.

Therefore we obtain

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{s=1}^t (\mathbb{E}[XX^T] - X_s X_s^T)\right) > u\right) \leq d\exp\left(-\frac{u^2/2}{tD_X^4 + D_X^2 u/3}\right), \qquad u > 0.$$

Applying it with $u = \frac{\lambda_{\min}}{2}t$, we get

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}(S_t) > \frac{2}{t\lambda_{\min}}\right) &\leq d\exp\left(-\frac{(\lambda_{\min}/2)^2 t^2/2}{tD_X^4 + D_X^2(\lambda_{\min}/2)t/3}\right)\\ &= d\exp\left(-t\frac{\lambda_{\min}^2/8}{D_X^4 + D_X^2\lambda_{\min}/6}\right)\\ &\leq \frac{1}{t^3}\frac{27de^{-3}\left(D_X^4 + D_X^2\lambda_{\min}/6\right)^3}{(\lambda_{\min}^2/8)^3},\end{aligned}$$

because $\max_{x\in\mathbb{R}}(e^{-ax}x^3) = \frac{27e^{-3}}{a^3}$ for any $a > 0$. Therefore the desired result is obtained as

$$\mathbb{E}[\lambda_{\max}(S_t)^2] \leq \frac{4}{\lambda_{\min}^2 t^2}\left(1 + \frac{1}{t}\frac{2de^{-3}\left(3D_X^4 + D_X^2\lambda_{\min}/2\right)^3}{(\lambda_{\min}^2/8)^2}\right).$$

□

Proof of Corollary 7. Denoting $E_t = \mathbb{E}_t\left[\frac{y_t X_t^T}{1+e^{y_t\hat{\theta}_t^T X_t}}\right](\theta_{\text{true}} - \hat{\theta}_t)$ and $Q_t = \frac{(\theta_{\text{true}}-\hat{\theta}_t)^T X_t X_t^T(\theta_{\text{true}}-\hat{\theta}_t)}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})}$, we have the sandwich relationship $e^{-\varepsilon}Q_t < E_t < e^{\varepsilon}Q_t$ according to Proposition 4. Therefore we obtain

$$E_t - cQ_t > (e^{-\varepsilon} - c)Q_t > \frac{e^{-\varepsilon} - c}{e^{\varepsilon}}E_t,$$

and the Corollary follows. □

Proof of Lemma 9. We start from the Kalman recursion

$$\begin{aligned}P_{t+1}^{-1} &= P_t^{-1} + \frac{1}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})}X_t X_t^T,\\ \hat{\theta}_{t+1} &= \hat{\theta}_t + P_{t+1}\frac{y_t X_t}{1+e^{y_t\hat{\theta}_t^T X_t}}.\end{aligned} \tag{7}$$

Multiplying Equation (7) by P_{t+1}^{-1} and (7) and subtracting θ_{true}, we obtain

$$(\hat{\theta}_{t+1} - \theta_{\text{true}})^T P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta_{\text{true}}) = (\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1}^{-1}(\hat{\theta}_t - \theta_{\text{true}}) + \frac{X_t^T P_{t+1} X_t}{(1+e^{y_t\hat{\theta}_t^T X_t})^2} + 2\frac{y_t X_t^T}{1+e^{y_t\hat{\theta}_t^T X_t}}(\hat{\theta}_t - \theta_{\text{true}}),$$

yielding the following equality:

$$\begin{aligned}&\sum_{t\in T_\varepsilon}\left((\frac{y_t X_t}{1+e^{y_t\hat{\theta}_t^T X_t}})^T(\theta_{\text{true}} - \hat{\theta}_t) - \frac{1}{2}(\hat{\theta}_t - \theta_{\text{true}})^T(P_{t+1}^{-1} - P_t^{-1})(\hat{\theta}_t - \theta_{\text{true}})\right)\\ &\quad= \frac{1}{2}\sum_{t\in T_\varepsilon}\frac{X_t^T P_{t+1} X_t}{(1+e^{y_t\hat{\theta}_t^T X_t})^2} + \frac{1}{2}\sum_{t\in T_\varepsilon}\left((\hat{\theta}_t - \theta_{\text{true}})^T P_t^{-1}(\hat{\theta}_t - \theta_{\text{true}}) - (\hat{\theta}_{t+1} - \theta_{\text{true}})^T P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta_{\text{true}})\right)\\ &\quad= \frac{1}{2}\sum_{t=1}^{n}\frac{X_t^T P_{t+1} X_t}{(1+e^{y_t\hat{\theta}_t^T X_t})^2} + \frac{1}{2}\sum_{t=1}^{n}\left((\hat{\theta}_t - \theta_{\text{true}})^T P_t^{-1}(\hat{\theta}_t - \theta_{\text{true}}) - (\hat{\theta}_{t+1} - \theta_{\text{true}})^T P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta_{\text{true}})\right)\\ &\qquad- \frac{1}{2}\sum_{t\notin T_\varepsilon}\frac{X_t^T P_{t+1} X_t}{(1+e^{y_t\hat{\theta}_t^T X_t})^2} - \frac{1}{2}\sum_{t\notin T_\varepsilon}\left((\hat{\theta}_t - \theta_{\text{true}})^T P_t^{-1}(\hat{\theta}_t - \theta_{\text{true}}) - (\hat{\theta}_{t+1} - \theta_{\text{true}})^T P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta_{\text{true}})\right).\end{aligned}$$

As $\sum_{t\notin T_\varepsilon}\frac{X_t^T P_{t+1} X_t}{(1+e^{y_t\hat{\theta}_t^T X_t})^2} \geq 0$, we obtain

$$\begin{aligned}&\sum_{t\in T_\varepsilon}\left((\frac{y_t X_t}{1+e^{y_t\hat{\theta}_t^T X_t}})^T(\theta_{\text{true}} - \hat{\theta}_t) - \frac{1}{2}(\theta_{\text{true}} - \hat{\theta}_t)^T(P_{t+1}^{-1} - P_t^{-1})(\theta_{\text{true}} - \hat{\theta}_t)\right)\\ &\quad\leq \frac{1}{2}\sum_{t=1}^{n}\frac{X_t^T P_{t+1} X_t}{(1+e^{y_t\hat{\theta}_t^T X_t})^2} + \frac{1}{2}\sum_{t=1}^{n}\left((\hat{\theta}_t - \theta_{\text{true}})^T P_t^{-1}(\hat{\theta}_t - \theta_{\text{true}}) - (\hat{\theta}_{t+1} - \theta_{\text{true}})^T P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta_{\text{true}})\right)\\ &\qquad- \frac{1}{2}\sum_{t\notin T_\varepsilon}\left((\hat{\theta}_t - \theta_{\text{true}})^T P_t^{-1}(\hat{\theta}_t - \theta_{\text{true}}) - (\hat{\theta}_{t+1} - \theta_{\text{true}})^T P_{t+1}^{-1}(\hat{\theta}_{t+1} - \theta_{\text{true}})\right).\end{aligned}$$

Using similar arguments than in the proof of Proposition 3, we obtain

$$\sum_{t=1}^{n} \frac{X_t^T P_{t+1} X_t}{(1+e^{y_t \hat{\theta}_t^T X_t})^2} \leq \frac{(1+e^D)}{2} d \log\left(1 + n p_1 D_X^2\right).$$

The telescopic sum yields the desired result

$$\sum_{t=1}^{n} \left((\hat{\theta}_t - \theta_{\text{true}})^T P_t^{-1} (\hat{\theta}_t - \theta_{\text{true}}) - (\hat{\theta}_{t+1} - \theta_{\text{true}})^T P_{t+1}^{-1} (\hat{\theta}_{t+1} - \theta_{\text{true}}) \right) \leq \frac{1}{p_1} \|\theta_{\text{true}}\|^2.$$

□

Proof of Lemma 10. We apply Lemma B.1 of Bercu and Touati (2008)) on the martingale difference $\Delta M_t \mathbb{1}_{|X_t^T(\hat{\theta}_t - \theta_{\text{true}})| < \varepsilon}$ (as $X_t^T(\hat{\theta}_t - \theta_{\text{true}})$ is adapted to the filtration $\sigma(X_1, y_1, \ldots, X_{t-1}, y_{t-1}, X_t)$) in order to obtain

$$\mathbb{E}\left[\exp\left(\sum_{t \in T_\varepsilon} \left(\lambda \Delta M_t - \frac{\lambda^2}{2}((\Delta M_t)^2 + \mathbb{E}_t[(\Delta M_t)^2])\right)\right)\right] \leq 1, \qquad \lambda > 0.$$

We will prove that

$$(\Delta M_t)^2 + \mathbb{E}_t[(\Delta M_t)^2]) \leq 2(1+e^D) \frac{(\theta_{\text{true}} - \hat{\theta}_t)^T X_t X_t^T (\theta_{\text{true}} - \hat{\theta}_t)}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})} \tag{8}$$

in order to achieve

$$\mathbb{E}\left[\exp\left(\sum_{t \in T_\varepsilon} \left(\lambda \Delta M_t - \lambda^2 (1+e^D) \frac{(\theta_{\text{true}} - \hat{\theta}_t)^T X_t X_t^T (\theta_{\text{true}} - \hat{\theta}_t)}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})}\right)\right)\right]$$
$$\leq \mathbb{E}\left[\exp\left(\sum_{t \in T_\varepsilon} \left(\lambda \Delta M_t - \frac{\lambda^2}{2}((\Delta M_t)^2 + \mathbb{E}_t[(\Delta M_t)^2])\right)\right)\right] \leq 1. \tag{9}$$

We obtain the inequality (8) by first developing the quadratic term $\mathbb{E}_t[(\Delta M_t)^2])$ as

$$\mathbb{E}_t[(\Delta M_t)^2]) = (\theta_{\text{true}} - \hat{\theta}_t)^T \frac{X_t X_t^T}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})} (\theta_{\text{true}} - \hat{\theta}_t) \mathbb{E}_t\left[\frac{1+e^{-y_t \hat{\theta}_t^T X_t}}{1+e^{y_t \hat{\theta}_t^T X_t}}\right],$$

$$\mathbb{E}_t\left[\frac{1+e^{-y_t \hat{\theta}_t^T X_t}}{1+e^{y_t \hat{\theta}_t^T X_t}}\right] = \frac{1+e^{-\hat{\theta}_t^T X_t}}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\theta_{\text{true}}^T X_t})} + \frac{1+e^{\hat{\theta}_t^T X_t}}{(1+e^{-\hat{\theta}_t^T X_t})(1+e^{\theta_{\text{true}}^T X_t})}$$
$$= \frac{a}{1+x} + \frac{a^{-1}}{1+x^{-1}},$$

with $a = \frac{1+e^{-\hat{\theta}_t^T X_t}}{1+e^{\hat{\theta}_t^T X_t}}$ and $x = e^{-\theta_{\text{true}}^T X_t}$. As the function $x \to \frac{a}{1+x} + \frac{a^{-1}}{1+1/x}$ is monotonic with limits a and a^{-1} in 0 and $+\infty$, we get $\mathbb{E}_t\left[\frac{1+e^{-y_t \hat{\theta}_t^T X_t}}{1+e^{y_t \hat{\theta}_t^T X_t}}\right] \leq \frac{1+e^D}{1+e^{-D}} < 1 + e^D$. To conclude to the inequality (8) we write

$$(\Delta M_t)^2 = \frac{(\theta_{\text{true}} - \hat{\theta}_t)^T X_t X_t^T (\theta_{\text{true}} - \hat{\theta}_t)}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})} \frac{1+e^{y_t \hat{\theta}_t^T X_t}}{1+e^{-y_t \hat{\theta}_t^T X_t}},$$

and we notice that $\frac{1+e^{y_t \hat{\theta}_t^T X_t}}{1+e^{-y_t \hat{\theta}_t^T X_t}} \leq \frac{1+e^D}{1+e^{-D}} < 1 + e^D$. Therefore the inequality (8) is proved.
Using (9) and the Chernoff's bound, we get for any $\alpha, \gamma > 0$ and $\lambda = \alpha/(1+e^D)$,

$$\mathbb{P}\left(\sum_{t \in T_\varepsilon} \left(\Delta M_t - \alpha \frac{(\theta_{\text{true}} - \hat{\theta}_t)^T X_t X_t^T (\theta_{\text{true}} - \hat{\theta}_t)}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})}\right) > \gamma\right)$$
$$\leq \exp(-\lambda\gamma) \mathbb{E}\left[\exp\left(\sum_{t \in T_\varepsilon} \left(\lambda \Delta M_t - \lambda^2 (1+e^D) \frac{(\theta_{\text{true}} - \hat{\theta}_t)^T X_t X_t^T (\theta_{\text{true}} - \hat{\theta}_t)}{(1+e^{\hat{\theta}_t^T X_t})(1+e^{-\hat{\theta}_t^T X_t})}\right)\right)\right]$$
$$\leq e^{-\frac{\alpha\gamma}{1+e^D}}.$$

Setting $\gamma = \frac{1+e^D}{\alpha} \log(\delta^{-1})$ for any $\delta > 0$ yields the result. □

Proof of Lemma 13. We use power functions of $\|\hat{\theta}_{t+1} - \theta_{\text{true}}\|^2$ identified as

$$(\hat{\theta}_{t+1} - \theta_{\text{true}})^T(\hat{\theta}_{t+1} - \theta_{\text{true}}) = (\hat{\theta}_t - \theta_{\text{true}})^T(\hat{\theta}_t - \theta_{\text{true}}) + 2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} + \frac{X_t^T P_{t+1}^2 X_t}{(1 + e^{y_t \hat{\theta}_t^T X_t})^2}.$$

Developing the power function of order k, we obtain

$$\|\hat{\theta}_{t+1} - \theta_{\text{true}}\|^{2k} = \left(\|\hat{\theta}_t - \theta_{\text{true}}\|^2 + 2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right)^k$$
$$+ \frac{X_t^T P_{t+1}^2 X_t}{(1 + e^{y_t \hat{\theta}_t^T X_t})^2} \sum_{i=1}^{k} \binom{k}{i} \left(\|\hat{\theta}_t - \theta_{\text{true}}\|^2 + 2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right)^{k-i} \left(\frac{X_t^T P_{t+1}^2 X_t}{(1 + e^{y_t \hat{\theta}_t^T X_t})^2} \right)^{i-1}.$$

By definition we note that $P_{t+1} \preccurlyeq P_1 = p_1 I$ so that

$$\left| \|\hat{\theta}_t - \theta_{\text{true}}\|^2 + 2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right| \leq \|\hat{\theta}_t - \theta_{\text{true}}\|^2 + 2p_1 \|\hat{\theta}_t - \theta_{\text{true}}\| \|X_t\| \leq 4D_\theta^2 + 2p_1 D_\theta D_X,$$

and

$$\frac{X_t^T P_{t+1}^2 X_t}{(1 + e^{y_t \hat{\theta}_t^T X_t})^2} \leq p_1^2 D_X^2.$$

Assumption 2 gives the rate

$$\mathbb{E}\left[\frac{X_t^T P_{t+1}^2 X_t}{(1 + e^{y_t \hat{\theta}_t^T X_t})^2} \right] \leq \frac{M_2}{4t^2}.$$

Summing those bounds in the binomial expansion of order k, we get

$$\mathbb{E}\left[\frac{X_t^T P_{t+1}^2 X_t}{(1 + e^{y_t \hat{\theta}_t^T X_t})^2} \sum_{i=1}^{k} \binom{k}{i} \left(\|\hat{\theta}_t - \theta_{\text{true}}\|^2 + 2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right)^{k-i} \left(\frac{X_t^T P_{t+1}^2 X_t}{(1 + e^{y_t \hat{\theta}_t^T X_t})^2} \right)^{i-1} \right]$$
$$\leq \frac{M_2}{4t^2} \frac{1}{p_1^2 D_X^2} \sum_{i=1}^{k} \binom{k}{i} \left(4D_\theta^2 + 2p_1 D_\theta D_X\right)^{k-i} \left(p_1^2 D_X^2\right)^i \leq \frac{b_{k,1}}{t^2},$$

with $b_{k,1} = \frac{M_2}{4p_1^2 D_X^2} \left(4D_\theta^2 + 2p_1 D_\theta D_X + p_1^2 D_X^2\right)^k$. Similarly, we use again the binomial expansion of order k in order to obtain

$$\left(\|\hat{\theta}_t - \theta_{\text{true}}\|^2 + 2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right)^k = \|\hat{\theta}_t - \theta_{\text{true}}\|^{2k} + 2k \|\hat{\theta}_t - \theta_{\text{true}}\|^{2(k-1)} \frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}}$$
$$+ \left(2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right)^2 \sum_{i=2}^{k} \binom{k}{i} \|\hat{\theta}_t - \theta_{\text{true}}\|^{2(k-i)} \left(2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right)^{i-2}.$$

We use again the elementary bound

$$\left| 2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right| \leq 2p_1 D_\theta D_X,$$

and the following estimate which holds under Assumption 2

$$\mathbb{E}\left[\left(2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right)^2 \right] \leq (4D_\theta)^2 \mathbb{E}\left[X_t^T P_{t+1}^2 X_t \right] \leq (4D_\theta)^2 \frac{M_2}{t^2}.$$

Summing the terms in the binomial expansion of order k we get

$$\mathbb{E}\left[\left(2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right)^2 \sum_{i=2}^{k} \binom{k}{i} \|\hat{\theta}_t - \theta_{\text{true}}\|^{2(k-i)} \left(2\frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right)^{i-2} \right] \leq \frac{b_{2,k}}{t^2},$$

with

$$b_{2,k} = (4D_\theta)^2 M_2 \frac{1}{(2p_1 D_\theta D_X)^2}(4D_\theta^2 + 2p_1 D_\theta D_X)^k = \frac{4M_2}{p_1^2 D_X^2}(4D_\theta^2 + 2p_1 D_\theta D_X)^k.$$

Hence we have

$$\mathbb{E}\left[\|\hat{\theta}_{t+1} - \theta_{\text{true}}\|^{2k}\right] \leq \mathbb{E}\left[\|\hat{\theta}_t - \theta_{\text{true}}\|^{2k}\right] + 2k\mathbb{E}\left[\|\hat{\theta}_t - \theta_{\text{true}}\|^{2(k-1)} \frac{(\hat{\theta}_t - \theta_{\text{true}})^T P_{t+1} y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}}\right] + \frac{b_k}{t^2},$$

with $b_k \geq b_{1,k} + b_{2,k}$. We then apply Proposition 4 deriving

$$\mathbb{E}_t\left[\frac{y_t}{1 + e^{y_t \hat{\theta}_t^T X_t}}\right] = -\frac{X_t^T(\hat{\theta}_t - \theta_{\text{true}})}{(1 + e^{\hat{\theta}_t^T X_t})(1 + e^{-\hat{\theta}_t^T X_t})} c_t\,,$$

with $e^{-D} < c_t < e^D$ and the tower property in order to obtain

$$\mathbb{E}\left[\|\hat{\theta}_{t+1} - \theta_{\text{true}}\|^{2k}\right] \leq \mathbb{E}\left[\|\hat{\theta}_t - \theta_{\text{true}}\|^{2k}\right] - \frac{kc_t}{1 + e^D}\mathbb{E}\left[\|\hat{\theta}_t - \theta_{\text{true}}\|^{2(k-1)}(\hat{\theta}_t - \theta_{\text{true}})^T \left(P_{t+1} X_t X_t^T\right)(\hat{\theta}_t - \theta_{\text{true}})\right] + \frac{b_k}{t^2}\,.$$

Then Assumption 1 applied thanks to the tower property yields

$$\mathbb{E}\left[\|\hat{\theta}_{t+1} - \theta_{\text{true}}\|^{2k}\right] \leq \mathbb{E}\left[\|\hat{\theta}_t - \theta_{\text{true}}\|^{2k}\right]\left(1 - \frac{e^{-D} k m_1}{t(1 + e^D)}\right) + \frac{b_k}{t^2}.$$

□

Proof of Corollary 14. Defining $l_t = \mathbb{E}\left[\|\hat{\theta}_t - \theta_{\text{true}}\|^{2k}\right]$ and according to Lemma 13 we have the inequality

$$l_{t+1} \leq l_t\Big(1 - \frac{ka}{t}\Big) + \frac{b_k}{t^2}, \qquad t \geq 1\,.$$

By a recursive argument it yields to the estimate for $t \geq 2$

$$l_t \leq \sum_{\tau=1}^{t-1} \frac{b_k}{\tau^2} \prod_{s=\tau+1}^{t-1}\Big(1 - \frac{ka}{s}\Big) + l_1 \prod_{s=1}^{t-1}\Big(1 - \frac{ka}{s}\Big) \leq \sum_{\tau=1}^{t-1} \frac{b_k}{\tau^2} \prod_{s=\tau+1}^{t-1}\Big(1 - \frac{ka}{s}\Big),$$

because $l_1 > 0$, $1 - ka < 0$ and for $s > 1$, $1 - \frac{ka}{s} > 0$. Moreover, taking the logarithm of the products, we estimate

$$\sum_{s=\tau+1}^{t-1} \log\Big(1 - \frac{ka}{s}\Big) \leq -ka \sum_{s=\tau+1}^{t-1} \frac{1}{s} \leq -ka \int_{\tau+1}^{t} \frac{du}{u} = ka\left(\log(\tau+1) - \log(t)\right)\,.$$

It provides the bound

$$\prod_{s=\tau+1}^{t-1}(1 - \frac{ka}{s}) \leq \frac{(\tau+1)^{ka}}{t^{ka}},$$

yielding the estimate

$$l_t \leq \frac{b_k}{t^{ka}} \sum_{\tau=1}^{t-1}(\tau+1)^{ka-2}\left(\frac{\tau+1}{\tau}\right)^2 \leq \frac{4b_k}{t^{ka}} \sum_{\tau=1}^{t-1}(\tau+1)^{ka-2}.$$

As $-1 < ka - 2 < 0$, we infer that

$$\sum_{\tau=1}^{t-1}(\tau+1)^{ka-2} \leq \int_1^t u^{ka-2} du = \frac{1}{ka-1}(t^{ka-1} - 1),$$

so that

$$l_t \leq \frac{4b_k}{ka-1}(\frac{1}{t} - \frac{1}{t^{ka}}),$$

and Corollary 14 follows for $t \geq 2$. □