Stochastic Online Optimization using Kalman Recursion

Joseph de Vilmarest  JOSEPH.DE_VILMAREST@SORBONNE-UNIVERSITE.FR
Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, CNRS
4 place Jussieu, 75005 Paris, France
and Électricité de France R&D

Olivier Wintenberger  OLIVIER.WINTENBERGER@SORBONNE-UNIVERSITE.FR
Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, CNRS
4 place Jussieu, 75005 Paris, France

Editor: Tong Zhang

Abstract

We study the Extended Kalman Filter in constant dynamics, offering a bayesian perspective of stochastic optimization. For generalized linear models, we obtain high probability bounds on the cumulative excess risk in an unconstrained setting, under the assumption that the algorithm reaches a local phase. In order to avoid any projection step we propose a two-phase analysis. First, for linear and logistic regressions, we prove that the algorithm enters a local phase where the estimate stays in a small region around the optimum. We provide explicit bounds with high probability on this convergence time, slightly modifying the Extended Kalman Filter in the logistic setting. Second, for generalized linear regressions, we provide a martingale analysis of the excess risk in the local phase, improving existing ones in bounded stochastic optimization. The algorithm appears as a parameter-free online procedure that optimally solves some unconstrained optimization problems.

Keywords: extended kalman filter, online learning, stochastic optimization

1. Introduction

The optimization of convex functions is a long-standing problem with many applications. In supervised machine learning it frequently arises in the form of the prediction of an observation $y_t \in \mathbb{R}$ given explanatory variables $X_t \in \mathbb{R}^d$. The aim is to minimize a cost depending on the prediction and the observation. We focus in this article on linear predictors, hence the loss function is of the form $\ell(y_t, \theta^\top X_t)$.

Two important settings have emerged in order to analyse learning algorithms. In the online setting $(X_t, y_t)$ may be set by an adversary. The assumption required is boundedness and the goal is to upper estimate the regret (cumulative excess loss compared to the optimum). In the stochastic setting with independent identically distributed (i.i.d.) $(X_t, y_t)$, we define the risk $L(\theta) = \mathbb{E}[\ell(y, \theta^\top X)]$. We focus on the cumulative excess risk $\sum_{t=1}^n L(\hat{\theta}_t) - L(\theta^*)$ where $\theta^*$ minimizes the risk. We obtain bounds holding with high probability simultaneously for any horizon, that is, we control the whole trajectory of the risk. Furthermore, our bounds on the cumulative risk all lead to similar ones on the excess risk at any step for the averaged version of the algorithm.

Due to its low computational cost the Stochastic Gradient Descent (SGD) of Robbins and Monro (1951) has been widely used, along with its equivalent in the online setting.
the online gradient descent (Zinkevich, 2003) and a simple variant where the iterates are averaged (Ruppert, 1988; Polyak and Juditsky, 1992). More recently Bach and Moulines (2013) provided a sharp bound in expectation on the excess risk for a two-step procedure that has been extended to the average of SGD with a constant step size (Bach, 2014). Second-order methods based on stochastic versions of Newton-Raphson algorithm have been developed in order to converge faster in iterations, although with a bigger computational cost per iteration (Hazan et al. 2007).

In order to obtain a parameter-free second-order algorithm we apply a bayesian perspective, seeing the loss as a negative log-likelihood and approximating the maximum-likelihood estimator at each step. We get a state-space model interpretation of the optimization problem: in a well-specified setting the space equation is $y_t \sim p_{\theta_t}(\cdot \mid X_t) \propto \exp(-\ell(\cdot, \theta_t^T X_t))$ with $\theta_t \in \mathbb{R}^d$ and the state equation defines the dynamics of the state $\theta_t$. The stochastic convex optimization setting corresponds to a degenerate constant state-space model $\theta_t = \theta_{t-1}$ called static. As usual in State-Space models, the optimization is realized with the Kalman recursion (Kalman and Bucy, 1961) for the quadratic loss and the Extended Kalman Filter (EKF) (Fahrmeir, 1992) in a more general case. A correspondence has recently been made by Ollivier (2018) between the static EKF and the online natural gradient (Amari, 1998). This motivates a risk analysis in order to enrich the link between Kalman filtering and the optimization community. We may see the static EKF as the online approximation of bayesian model averaging, and similarly to its analysis derived by Kakade and Ng (2005) our analysis is robust to misspecification, that is we don’t assume the data to be generated by the probabilistic model.

The static EKF is very close to the Online Newton Step (ONS) introduced by Hazan et al. (2007) as both are second-order online algorithms and our results are of the same flavor as those obtained on the ONS (Mahdavi et al., 2015). However the ONS requires the knowledge of the region in which the optimization is realized. It is involved in the choice of the gradient step size and a projection step is done to ensure that the search stays in the chosen region. On the other hand the static EKF yields two advantages at the cost of being less generic.

First, there is no costly projection step and each recursive update runs in $O(d^2)$ operations, where $d$ is the dimension of the features $X_t$. Therefore, our comparison of the static EKF with the ONS provides a lead to the open question of Koren (2013). Indeed, the problem of the ONS pointed out by Koren (2013) is to control the cost of the projection step and the question is whether it is possible to perform better than the ONS in the stochastic exp-concave setting. We don’t answer the open question in the general setting. However, we suggest a general way to get rid of the projection by dividing the analysis between a convergence proof of the algorithm to the optimum and a second phase where the estimate stays in a small region around the optimum where no projection is required.

Second, the algorithm is (nearly) parameter-free. We believe that bayesian statistics is the reasonable approach in order to obtain parameter-free online algorithms in the unconstrained setting. Parameter-free is not exactly correct as there are initialization parameters, which we see as a smoothed version of the hard constraint imposed by bounded algorithm, but they have no impact on the leading terms of our bounds. Static Kalman filter coincides with the ridge forecaster and similarly the static EKF may be seen as the online approximation of a regularized empirical risk minimizer.
1.1 Related Work

Theoretical guarantees for online and stochastic algorithms are multi-criteria and of various natures. The comparison of upper-bounds or computational complexity highly depends on the values of \( d \) the dimension of the explanatory vectors and \( n \) the time horizon, leading to different views on whether the dependence to \( d \) or \( n \) is the most important. The nature of the guarantee obviously depends on the objective pursued.

In the adversarial setting, the learner suffers a loss \( \ell_t(\hat{\theta}_t) \) depending on its estimate \( \hat{\theta}_t \) at each time step \( t \). It is natural to minimize the cumulative loss, or equivalently the regret

\[
\sum_{t=1}^{n} \ell_t(\hat{\theta}_t) - \sum_{t=1}^{n} \ell_t(\theta^*) ,
\]

where \( \theta^* \) reaches the minimum value of the cumulative loss and thus highly depends on \((\ell_t)_{1 \leq t \leq n}\). Under an assumption of bounded gradients, Zinkevich (2003) proved that a first-order online gradient descent yields a regret bound in \( O(\sqrt{n}) \). The Online Newton Step (ONS) is a second-order online gradient descent that has been designed to obtain a regret bound in \( O(\ln n) \) (Hazan et al., 2007) under the assumption that the losses are exp-concave. The improved guarantee comes at a cost of \( O(d^2) \) operations per step instead of \( O(d) \), along with a projection at each step whose cost depends on the data.

In the stochastic setting where the losses \( (\ell_t) \) are assumed i.i.d., the aim is to minimize the risk \( L(\theta) = \mathbb{E}[\ell(\theta)] \). A natural candidate is the Empirical Risk Minimizer (ERM), whose asymptotics are well understood (see for example Murata and Amari (1999)). Assuming the existence of \( \theta^* \) minimizing the risk and that the Hessian matrix \( H^* = \nabla^2 L(\theta^*) \) is positive definite, the ERM \( \hat{\theta}_n \) satisfies

\[
\mathbb{E}[L(\hat{\theta}_n)] - L(\theta^*) = \frac{\text{tr}(G^*H^{*-1})}{2n} + o(1/n) ,
\]

\[
G^* = \mathbb{E}\left[ \nabla^2 \ell(y, \theta^* X) \right] .
\]

Although in the well-specified setting the identity \( \text{tr}(G^*H^{*-1}) = d \) holds, in the misspecified case there is no general estimate for \( \text{tr}(G^*H^{*-1}) \). Recently a non-asymptotic bound \( L(\hat{\theta}_n) - L(\theta^*) = O(\text{tr}(G^*H^{*-1})\ln \delta^{-1}/n) \) holding with probability \( 1 - \delta \) has been shown by Ostrovskii and Bach (2021) on the ERM. However the ERM is defined only implicitly and may have important computational cost, hence recursive algorithms based on gradient descent have been studied under different sets of assumptions to bound \( \text{tr}(G^*H^{*-1}) \).

The most simple is Stochastic Gradient Descent (SGD), where each step is in the opposite direction of the gradient. This algorithm has been widely used with various step sizes. Sharp results have been obtained by Bach (2014) for a constant gradient step size \( C/\sqrt{n} \) with fixed horizon \( n \). Under the assumption that gradients are bounded by \( R \) we have \( \text{tr}(G^*H^{*-1}) \leq R^2/\mu \) where \( \mu \) is the minimal eigenvalue of \( H^* \). The fast rate \( \mathbb{E}[L(\bar{\theta}_n)] - L(\theta^*) = O(R^2/(\mu n)) \) is obtained by Bach (2014) for the averaged estimate \( \bar{\theta}_n \) of SGD. In the same article the author also derives a bound with high probability but with a slower rate: it degrades into \( L(\bar{\theta}_n) - L(\theta^*) = O(\log \delta^{-1}/\sqrt{n}) \) with probability \( 1 - \delta \). Finally, in the quadratic setting a fast rate \( L(\bar{\theta}_n) - L(\theta^*) = O(1/(n\delta^\alpha)) \) is achieved with probability \( 1 - \delta \) for a defined \( \alpha > 0 \) (Bach and Moulines, 2013).

To obtain fast rates with high probability beyond the quadratic setting, it seems necessary to use second-order information as in the ONS (Mahdavi et al., 2015). Under the
Table 1: Summary of existing results along with the static EKF for which we prove the bound for the cumulative excess risk. We focus on the dependence on $n$, and $\delta$ for the bounds holding with probability $1 - \delta$ (w.h.p.). $S(\delta)$ is the cumulative excess risk of the convergence phase.

<table>
<thead>
<tr>
<th></th>
<th>ERM</th>
<th>Averaged SGD</th>
<th>ONS</th>
<th>This article</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regret</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Excess risk in expectation</td>
<td>$O(\frac{1}{n})$</td>
<td>$O(\frac{1}{n})$</td>
<td>$O(\ln n)$</td>
<td></td>
</tr>
<tr>
<td>Excess risk w.h.p.</td>
<td>$O(\frac{\ln \delta^{-1}}{n})$</td>
<td>$O(\frac{\ln \delta^{-1}}{n})$</td>
<td>$O(\ln n + \ln \delta^{-1})$</td>
<td></td>
</tr>
<tr>
<td>Cumulative excess risk w.h.p.</td>
<td>$O(\ln n + \ln \delta^{-1})$</td>
<td>$O(\ln n + \ln \delta^{-1} + S(\delta))$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost per iteration</td>
<td>Implicit</td>
<td>$O(d)$</td>
<td>$O(d^2 + T_{proj})$</td>
<td>$O(d^2)$</td>
</tr>
</tbody>
</table>

assumption that the loss is $\alpha$-exp-concave, $\text{tr}(G^*H^{*-1}) \leq d/\alpha$ and for the averaged version of the ONS the rate $L(\tilde{\theta}_n) - L(\theta^*) = O(d(\ln n + \ln \delta^{-1})/(\alpha n))$ with probability $1 - \delta$ is obtained. From our perspective, the result is stronger than what is claimed by Mahdavi et al. (2015): the bound obtained is

$$\sum_{t=1}^{n} L(\hat{\theta}_n) - L(\theta^*) = O(\ln n + \ln \delta^{-1}),$$

(1)

holding simultaneously for any $n$ with probability $1 - \delta$. Note that although averaging this bound with Jensen’s inequality leads to a sub-optimal bound on the excess risk of the last averaged estimate, it is conversely not possible to obtain Equation (1) from

$$L(\hat{\theta}_n) - L(\theta^*) = O(\ln \delta^{-1}/n),$$

holding with probability $1 - \delta$.

1.2 Contributions

Our first contribution is a local analysis of the static EKF under assumptions defined in Section 2, and provided that consecutive steps stay in a small ball around the optimum $\theta^*$. We derive local bounds on the cumulative risk with high probability from a martingale analysis. Our analysis of Section 3 is similar to the one of Mahdavi et al. (2015) and we slightly refine their constants as a by-product.

We then show that the convergence property crucial in our analysis is reachable. To that end we focus on linear regression and logistic regression as these two well-known problems are challenging in the unconstrained setting. In linear regression, the gradient of the loss is not bounded globally. In logistic regression, the loss is strictly convex, but neither strongly convex nor exp-concave in the unconstrained setting. In Section 4, we develop a global bound in the logistic setting on a slight modification of the algorithm introduced by Bercu et al. (2020). We prove that this modified algorithm converges and stays into a local region around $\theta^*$ after a finite number of steps. Moreover we show that it coincides with the static EKF and thus our local analysis applies. In Section 5, we apply our local analysis to the quadratic setting. We rely on Hsu et al. (2012) to obtain the convergence of the algorithm after exhibiting the correspondence between Kalman filter in constant dynamics and the ridge forecaster, and we therefore obtain similarly a global bound.
Finally, we demonstrate numerically the competitiveness of the static EKF for logistic regression in Section 6.

2. Definitions and Assumptions

We consider loss functions that may be written as the negative log-likelihood of a generalized linear model (McCullagh and Nelder, 1989). Formally, the loss is defined as
\[
\ell(y, \theta^T X) = -\log p_{\theta}(y | X) \quad \text{where} \quad \theta \in \mathbb{R}^d, \ (X, y) \in \mathcal{X} \times \mathcal{Y} \quad \text{for some} \quad \mathcal{X} \subset \mathbb{R}^d \quad \text{and} \quad \mathcal{Y} \subset \mathbb{R} \quad \text{and} \quad p_{\theta} \text{ is of the form}
\]
\[
p_{\theta}(y | X) = h(y) \exp \left( \frac{y \theta^T X - b(\theta^T X)}{a} \right),
\]

where \(a\) is a constant and \(h\) and \(b\) are one-dimensional functions on which a few assumptions are required (Assumption 3). This includes logistic and quadratic regressions, see Sections 4 and 5. We display the static EKF in Algorithm 1 in this setting (see Appendix A for a derivation relying on Durbin and Koopman, 2012). In the quadratic setting, noting that the EKF estimate \(\hat{\theta}_t\) does not depend on the (unknown) variance \(\sigma^2\), we consider the quadratic loss \(\ell(y, \hat{y}) = (y - \hat{y})^2 / 2\) by convention.

\begin{algorithm}
\caption{Static Extended Kalman Filter for Generalized Linear Model}
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 \(P_{t+1} = P_t - P_t X_t X_t^T P_t^{-1} \alpha_t\) with \(\alpha_t = \frac{b'(\hat{\theta}_t^T X_t)}{a}\).
   \(b\) Update \(\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1} (y_t - b'(\hat{\theta}_t^T X_t)) X_t\).
\end{algorithm}

Due to matrix-vector and vector-vector multiplications, Algorithm 1 has a running-time complexity of \(O(d^2)\) at each iteration and thus \(O(nd^2)\) for \(n\) iterations.

Note that although we need the loss function to be derived from a likelihood of the form (2), we do not need the data to be generated under this process. We need two standard hypotheses on the data. The first one is the i.i.d. assumption and bounded random design (all along the paper \(\|\|\) is the Euclidean norm):

**Assumption 1** The observations \((X_t, y_t)\) are i.i.d. copies of the pair \((X, y) \in \mathcal{X} \times \mathcal{Y}\) and there exists \(D_X\) such that \(\|X_t\| \leq D_X\) almost surely (a.s.).

Working under Assumption 1, we define the risk function \(L(\theta) = \mathbb{E} [\ell(y, \theta^T X)]\). Note that in Section 3 we don’t need \(\mathbb{E}[XX^T]\) invertible, but we will make such an assumption in Sections 4 and 5 to prove the convergence of the algorithm in the logistic and quadratic settings, respectively. In order to work on a well-defined optimization problem we assume there exists a minimum:

**Assumption 2** There exists \(\theta^* \in \mathbb{R}^d\) such that \(L(\theta^*) = \inf_{\theta \in \mathbb{R}^d} L(\theta)\).
We treat two different settings requiring different assumptions, summarized in Assumption 3 and 4 respectively. First, motivated by logistic regression we define:

**Assumption 3** There exist $(\kappa_\varepsilon)_{\varepsilon > 0}, (h_\varepsilon)_{\varepsilon > 0}$ and $\rho_\varepsilon \xrightarrow{\varepsilon \to 0} 1$ such that for any $\varepsilon > 0$ and any $\theta, \theta_0 \in \mathbb{R}^d$ satisfying $\max(\|\theta - \theta^*\|, \|\theta_0 - \theta^*\|) \leq \varepsilon$, we have

- $\ell'(y, \theta^\top X)^2 \leq \kappa_\varepsilon \ell''(y, \theta^\top X)$ a.s.
- $\ell''(y, \theta^\top X) \leq h_\varepsilon$ a.s.
- $\ell''(y, \theta^\top X) \geq \rho_\varepsilon \ell''(y, \theta_0^\top X)$ a.s.

Here $\ell'$ and $\ell''$ are the first and second derivatives of $\ell$ with respect to the second variable.

Assumption 3 requires local exp-concavity (around $\theta^*$) along with some regularity on $\ell''$ ($\ell''$ continuous and $\ell''(y, \theta^\top X) \geq \gamma > 0$ a.s. is sufficient). That setting implies $Y$ bounded, because $\ell'$ depends on $y$ whereas $\ell''$ doesn’t. In logistic regression, $Y = \{-1, +1\}$ and Assumption 3 is satisfied for $\kappa_\varepsilon = e^{D_X(\|\theta^*\| + \varepsilon)}$, $h_\varepsilon = \frac{1}{4}$, $\rho_\varepsilon = e^{-\varepsilon D_X}$.

Second, we consider the quadratic loss, corresponding to a Gaussian model. In order to include the well-specified setting and to bound $G^* = \mathbb{E}[(y - \theta^\top X)^2XX^\top]$, we assume $y$ sub-gaussian conditionally to $X$ and not too far away from the model:

**Assumption 4** The distribution of $(X, y) \in X \times Y$ satisfies

- There exists $\sigma^2 > 0$ such that for any $s \in \mathbb{R}$, $\mathbb{E}\left[e^{s(y - \mathbb{E}[y|X])} | X\right] \leq e^{\sigma^2 s^2}$ a.s.,
- There exists $D_{\text{app}} \geq 0$ such that $|\mathbb{E}[y | X] - \theta^\top X| \leq D_{\text{app}}$ a.s.

Both conditions of Assumption 4 hold with $Y = \mathbb{R}$ and $D_{\text{app}} = 0$ for the well-specified Gaussian linear model with random bounded design. The second condition of Assumption 4 is satisfied for $D_{\text{app}} > 0$ in misspecified sub-gaussian linear model with a.s. bounded approximation error.

3. The Algorithm Around the Optimum

In this section, we analyse the cumulative risk under a strong convergence assumption. Precisely we define

$$\tau(\varepsilon) = \min\{k \in \mathbb{N} | \forall t > k, \|\hat{\theta}_t - \theta^*\| \leq \varepsilon\},$$

where $(\hat{\theta}_t)_t$ are the estimates of the static EKF, and with the convention $\min \emptyset = +\infty$. We assume a bound on $\tau(\varepsilon)$ holding with high probability:

**Assumption 5** For any $\delta, \varepsilon > 0$, there exists $T(\varepsilon, \delta) \in \mathbb{N}$ such that

$$\mathbb{P}(\tau(\varepsilon) \leq T(\varepsilon, \delta)) \geq 1 - \delta.$$

Assumption 5 states that with high probability there exists a convergence time after which the algorithm stays trapped in a local region around the optimum. Sections 4 and 5 are devoted to define explicitly such a convergence time for a modified EKF in the logistic setting and for the true EKF in the quadratic setting.
We present our result in the bounded and sub-gaussian settings. The results and their proofs are very similar, but two crucial steps are different. First, Assumption 3 yields a bound on the gradient holding almost surely. We relax the boundedness condition for the quadratic loss with a sub-gaussian hypothesis, requiring a specific analysis. Second, our analysis is based on a second-order expansion. The quadratic loss is equal to its second-order Taylor expansion but we need Assumption 5 along with the third point of Assumption 3 otherwise.

The following theorem is our result in the bounded setting.

**Theorem 1** Starting the static EKF from any $\hat{\theta}_1 \in \mathbb{R}^d$, $P_1 > 0$, if Assumptions 1, 2, 3, 5 are satisfied and if $\rho_\varepsilon > 0.95$, for any $\delta > 0$, it holds for any $n \geq 1$ simultaneously

$$
\sum_{t=\tau(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} L(\hat{\theta}_t) - L(\theta^*) \leq \frac{5}{2} d \kappa_\varepsilon \ln \left(1 + n \frac{\lambda_{\max}(P_1)D_X^2}{\varepsilon^2} \right) + 5 \lambda_{\max} \left(P_{T(\varepsilon, \delta)+1}^{-1}\right) \varepsilon^2 
$$

$$+ 30 \left(2 \kappa_\varepsilon + \frac{\lambda_{\max}(P_1)D_X^2}{\varepsilon} \right) \ln \delta^{-1},$$

with probability at least $1 - 3\delta$.

The constant 0.95 may be chosen arbitrarily close to 0.5 with growing constants in the bound on the cumulative risk. There is a hidden trade-off in $\varepsilon$: on the one hand, the smaller $\varepsilon$ the better our upper-bound, but on the other hand $T(\varepsilon, \delta)$ increases when $\varepsilon$ decreases, and thus our bound applies after a bigger convergence time.

For the quadratic loss, we obtain the following result under the sub-gaussian hypothesis.

**Theorem 2** In the quadratic setting, starting the static EKF from any $\hat{\theta}_1 \in \mathbb{R}^d$, $P_1 > 0$, if Assumptions 1, 2, 4 and 5 are satisfied, for any $\delta > 0$ and any $\varepsilon > 0$, it holds for any $n \geq 1$ simultaneously

$$
\sum_{t=\tau(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} L(\hat{\theta}_t) - L(\theta^*) \leq \frac{15}{2} d \left(8 \sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2 \right) \ln \left(1 + n \frac{\lambda_{\max}(P_1)D_X^2}{\varepsilon^2} \right)
$$

$$+ 5 \lambda_{\max} \left(P_{T(\varepsilon, \delta)+1}^{-1}\right) \varepsilon^2 + 115 \left(\varepsilon^2 \left(4 + \frac{\lambda_{\max}(P_1)D_X^2}{4} \right) + D_{\text{app}}^2 + 2 \varepsilon^2 D_X^2 \right) \ln \delta^{-1},$$

with probability at least $1 - 3\delta$.

We observe a similar trade-off in $\varepsilon$ as in Theorem 1. Up to numerical constants, the tight constant $d(\sigma^2 + D_{\text{app}}^2)$ (see for instance Hsu et al. 2012) is achieved by choosing $\varepsilon$ arbitrarily small, at the cost of a loose control of the $I(\varepsilon, \delta)$ first steps.

Both results follow from a regret analysis close to the one on the ONS (see Section 3.1), and on a control on the martingales stated below:

**Lemma 3** Let $k \geq 0$ and $(\Delta N_t)_{t \geq k}$ be any martingale difference adapted to the filtration $(\mathcal{F}_t)_{t \geq k}$ such that for any $t > k$, $\mathbb{E}[\Delta N_t^2 | \mathcal{F}_{t-1}] < \infty$. For any $\delta, \lambda > 0$, we have the simultaneous property

$$
\sum_{t=k+1}^{k+n} \left(\Delta N_t - \frac{\lambda}{2} (\Delta N_t) + \mathbb{E}[(\Delta N_t) | \mathcal{F}_{t-1}] \right) \leq \frac{\ln \delta^{-1}}{\lambda}, \quad n \geq 1,
$$

with probability at least $1 - \delta$. 

7
Algorithm 2: Recursive updates of the ONS and the static EKF

<table>
<thead>
<tr>
<th>Online Newton Step</th>
<th>Static Extended Kalman Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{t+1}^{-1} = P_t^{-1} + \ell'(y_t, w_t^T X_t)^2 X_t X_t^T$</td>
<td>$P_{t+1}^{-1} = P_t^{-1} + \ell''(y_t, \hat{\theta}_t^T X_t) X_t X_t^T$,</td>
</tr>
<tr>
<td>$\nabla_t = \ell'(y_t, w_t^T X_t) X_t$</td>
<td>$\nabla_t = \ell'(y_t, \hat{\theta}_t^T X_t) X_t$,</td>
</tr>
<tr>
<td>$w_{t+1} = \prod_{K} \left( w_t - \frac{1}{\gamma} P_{t+1} \nabla_t \right)$</td>
<td>$\hat{\theta}_{t+1} = \hat{\theta}<em>t - P</em>{t+1} \nabla_t$,</td>
</tr>
<tr>
<td>$P_{t+1}^{-1}$ is the projection on $K$ for the norm $| |$.</td>
<td></td>
</tr>
</tbody>
</table>

This result proved in Appendix B.1 is a corollary of a martingale inequality from Bercu and Touati (2008) and a stopping time construction (Freedman, 1975).

We detail the key ideas of the proofs in the rest of the Section, and we defer to Appendix B the proof of the intermediate results along with the detailed proof of Theorems 1 and 2. Specifically, we display in Section 3.1 the parallel with the ONS, where we compare with the existing result on the cumulative risk, and a similar analysis yields an adversarial bound on a second-order expansion of the loss. In Section 3.2 we compare the excess risk with its second-order expansion thanks to Assumption 5, and we use a martingale analysis to obtain a bound on the cumulative excess risk.

3.1 Comparison with Online Newton Step and Adversarial Analysis

We display the parallel between the ONS and the static EKF in Algorithm 2 through their recursive updates. We observe that the square of the first derivative of $\ell$ is replaced with the second derivative. Thus $t P_t^{-1}$ in the static EKF is an estimate of the Hessian $H^*$ which is the optimal preconditioning matrix as shown in Corollary 3 of Murata and Amari (1999).

Then the recursion on the parameter ($w_t$ and $\hat{\theta}_t$) has two differences: there is a gradient step size $1/\gamma$ in the ONS absent in the static EKF, and after the gradient step the ONS applies a projection. Lemma 3 yields the following refinement on the bound of Mahdavi et al. (2015) obtained on the cumulative excess risk of the ONS:

**Corollary 4** Assume the search region $K$ has diameter $D$ and the gradients are bounded by $R$. Let $(w_t)_t$ be the ONS estimates starting from $w_1 \in K$, $P_1 = \lambda I$ and using a step-size $\gamma = \frac{1}{2} \min\left(\frac{1}{4RD}, \alpha\right)$ with $\alpha$ the exp-concavity constant of $\ell$ on $K$. Then for any $\delta > 0$, it holds for any $n \geq 1$ simultaneously

$$
\sum_{t=1}^{n} L(w_t) - L(\theta^*) \leq \frac{3}{2\gamma} d \ln \left( 1 + \frac{nR^2}{\lambda d} \right) + \frac{\lambda \gamma}{6} D^2 + \left( \frac{12}{\gamma} + \frac{4\gamma R^2 D^2}{3} \right) \ln \delta^{-1},
$$

with probability at least $1 - 2\delta$.

For the sake of consistency, we display Corollary 4 as a bound on the cumulative excess risk, whereas Theorem 3 of Mahdavi et al. (2015) is a bound on the excess risk of the averaged
ONS. The latter follows directly from an application of Jensen’s inequality. The proof of Corollary 4 consists in replacing Theorem 4 of Mahdavi et al. (2015) with Lemma 3. We obtain similar constants in Theorem 1 and in Corollary 4, as \( \kappa \) is the inverse of the exp-concavity constant \( \alpha \). The use of second-order methods with well-tuned preconditioning is crucial in order to replace the leading constant \( R^2/\mu \) obtained for first-order methods by \( d/\alpha \) (\( \mu \) is the minimum eigenvalue of the hessian \( H^* \)).

Our results on the static EKF are less general than the ones obtained on the ONS as a control of the convergence time \( \tau(\varepsilon) \leq T(\varepsilon, \delta) \) is required with high probability. On the other hand, the results obtained on the ONS require the knowledge of the exp-concavity constant \( \alpha \) whereas the static EKF is parameter-free. That is why we argue that the static EKF provides an optimal way to tune the step size and the preconditioning matrix. Indeed, as \( \varepsilon \) is a parameter of the EKF analysis but not of the algorithm, we can improve the leading constant \( \kappa \) on a local region arbitrarily small around \( \theta^* \), at a cost of a loose control of the \( T(\varepsilon, \delta) \) first steps. In the ONS the choice of a diameter \( D > \|\theta^*\| \) makes the gradient step-size sub-optimal and impacts the leading constant.

Once the parallel between the ONS and the static EKF has been displayed (Algorithm 2), it is natural to adopt an approach similar to the one in Hazan et al. (2007). The cornerstone of our local analysis is the derivation of an adversarial bound on the second-order Taylor expansion of \( \ell \), from the recursive update formulae.

**Lemma 5** For any sequence \( (X_t, y_t)_t \), starting from any \( \hat{\theta}_1 \in \mathbb{R}^d, P_1 > 0 \), it holds for any \( \theta^* \in \mathbb{R}^d \) and \( n \in \mathbb{N} \) that

\[
\sum_{t=1}^{n} \left( (\ell'(y_t, \hat{\theta}_t^\top X_t) X_t) \top (\hat{\theta}_t - \theta^*) - \frac{1}{2} (\hat{\theta}_t - \theta^*) \top \left( \ell''(y_t, \hat{\theta}_t^\top X_t) X_t X_t^\top \right) (\hat{\theta}_t - \theta^*) \right) \\
\leq \frac{1}{2} \sum_{t=1}^{n} X_t^\top P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^\top X_t)^2 + \frac{\|\hat{\theta}_1 - \theta^*\|^2}{\lambda_{\min}(P_1)}.
\]

We cannot compare the excess loss with the second-order Taylor expansion in general, and it is natural to use a step size parameter. In Hazan et al. (2007), the regret analysis of the ONS is based on a very similar bound on

\[
\left( (\ell'(y_t, w_t^\top X_t) X_t) \top (w_t - \theta^*) - \frac{\gamma}{2} (w_t - \theta^*) \top \left( \ell''(y_t, w_t^\top X_t) X_t X_t^\top \right) (w_t - \theta^*) \right),
\]

where \( \gamma \) is a step size parameter. Then the regret bound follows from the exp-concavity property, bounding the excess loss \( \ell(y_t, w_t^\top X_t) - \ell(y_t, \theta^*^\top X_t) \) with the previous quantity for a specific \( \gamma \). The dependence of \( \gamma \) on the exp-concavity constant and the bound on the gradients require that the algorithm stays in a bounded region around the optimum \( \theta^* \), and a projection on this region is used, potentially at each step.

We follow a very different approach, to stay parameter-free, unconstrained and to avoid any additional cost in the leading constant. In the stochastic setting, we observe that we can upper-bound the excess risk with a second-order expansion, up to a multiplicative factor.

### 3.2 From Adversarial to Stochastic: the Cumulative Risk

In order to compare the excess risk with a second-order expansion, we compare the first-order term with the second-order one.
Proposition 6 If Assumptions 1, 2 and 3 are satisfied, for any \( \theta \in \mathbb{R}^d \), it holds
\[
\frac{\partial L}{\partial \theta}^\top_\theta (\theta - \theta^*) \geq \rho_{||\theta-\theta^*||} (\theta - \theta^*)^\top \frac{\partial^2 L}{\partial \theta^2} |_\theta (\theta - \theta^*).
\]

This result leads immediately to the following proposition, using the first-order convexity property of \( L \).

Proposition 7 If Assumptions 1, 2 and 3 are satisfied, for any \( \theta \in \mathbb{R}^d \), \( 0 < c < \rho_{||\theta-\theta^*||} \), it holds
\[
L(\theta) - L(\theta^*) \leq \frac{\rho_{||\theta-\theta^*||}}{\rho_{||\theta-\theta^*||} - c} \left( \frac{\partial L}{\partial \theta}^\top_\theta (\theta - \theta^*) - c(\theta - \theta^*)^\top \frac{\partial^2 L}{\partial \theta^2} |_\theta (\theta - \theta^*) \right).
\]

Lemma 5 motivates the use of \( c > \frac{1}{2} \), thus we need at least \( \rho_{||\theta-\theta^*||} > \frac{1}{2} \). In the quadratic setting, it holds as an equality with \( \rho = 1 \) because the second derivative of the quadratic loss is constant. In the bounded setting we need to control the second derivative in a small range, and we can achieve that only locally, therefore we separate the condition \( \rho_{||\theta-\theta^*||} > \frac{1}{2} \) between the third point of Assumption 3 and Assumption 5.

Then we are left to obtain a bound on the cumulative risk from Lemma 5. In order to compare the derivatives of the risk and the losses, we need to control the martingale difference adapted to the natural filtration (\( F_t \)) and defined as
\[
\Delta M_t = \left( \frac{\partial L}{\partial \theta} |_{\hat{\theta}_t} - \nabla_t \right)^\top (\hat{\theta}_t - \theta^*), \quad \text{where} \ \nabla_t = \ell'(y_t, \hat{\theta}_t^\top X_t) X_t.
\]
We thus apply Lemma 3 to this martingale difference.

Lemma 8 Starting the static EKF from any \( \hat{\theta}_1 \in \mathbb{R}^d, P_1 > 0 \), if Assumptions 1 and 2 are satisfied, for any \( k \geq 0 \) and \( \delta, \lambda > 0 \), it holds simultaneously
\[
\sum_{t=k+1}^{k+n} \left( \Delta M_t - \lambda (\hat{\theta}_t - \theta^*)^\top \left( \nabla_t \nabla_t^\top + \frac{3}{2} \mathbb{E} [\nabla_t \nabla_t^\top | F_{t-1}] \right) (\hat{\theta}_t - \theta^*) \right) \leq \frac{\ln \delta^{-1}}{\lambda}, \quad n \geq 1,
\]
with probability at least \( 1 - \delta \).

The proof of Theorems 1 and 2 deferred to Appendix B builds on the above results. Summing Lemma 5 and 8, we obtain for any \( \delta, \lambda > 0 \) the simultaneous bound
\[
\sum_{t=T(\epsilon,\delta)+1}^{T(\epsilon,\delta)+n} \left( \frac{\partial L}{\partial \theta} |_{\hat{\theta}_t} (\hat{\theta}_t - \theta^*)^\top (\hat{\theta}_t - \theta^*) - (\hat{\theta}_t - \theta^*)^\top \left( \frac{1}{2} \nabla_t^{(2)} + \lambda \nabla_t \nabla_t^\top + \frac{3}{2} \lambda \mathbb{E} [\nabla_t \nabla_t^\top | F_{t-1}] \right) (\hat{\theta}_t - \theta^*) \right)
\]
\[
\leq \frac{1}{2} \sum_{t=T(\epsilon,\delta)+1}^{T(\epsilon,\delta)+n} X_t^\top P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^\top X_t)^2 + \frac{||\hat{\theta}_1 - \theta^*||^2}{\lambda_{\min}(P_{T(\epsilon,\delta)+1})} + \frac{\ln \delta^{-1}}{\lambda}, \quad n \geq 1,
\]
with probability at least \( 1 - \delta \), where we define \( \nabla_t^{(2)} = \ell''(y_t, \hat{\theta}_t^\top X_t) X_t X_t^\top \) for any \( t \). In the last equation, we control (see Appendix B.4 and B.5) the quadratic term in \( \hat{\theta}_t - \theta^* \) on the left hand-side in terms of \( (\hat{\theta}_t - \theta^*)^\top \frac{\partial^2 L}{\partial \theta^2} |_{\hat{\theta}_t} (\hat{\theta}_t - \theta^*) \) in order to lower-bound the left expression proportionally to the cumulative excess risk using Proposition 7 for well chosen \( \lambda \).
Algorithm 3: Truncated Extended Kalman Filter for Logistic Regression

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 $P_{t+1} = P_t - \frac{P_t X_t X_t^T P_t}{1 + e^{\hat{\theta}_t^T X_t}} \alpha_t$, with $\alpha_t = \max \left( \frac{1}{t \beta}, \frac{1}{(1 + e^{\hat{\theta}_t^T X_t})(1 + e^{-\hat{\theta}_t^T X_t})} \right)$.

   (b) Update $\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1} \frac{y_t X_t}{1 + e^{\hat{\theta}_t^T X_t}}$.

4. **Logistic Setting**

Logistic regression is a widely used statistical model in classification. The prediction of a binary random variable $y \in \mathcal{Y} = \{-1, 1\}$ consists in modelling $\mathcal{L}(y \mid X)$ with

$$p_\theta(y \mid X) = \frac{1}{1 + e^{-y \theta^T X}} = \exp \left( \frac{y \theta^T X - (2 \ln(1 + e^{\theta^T X}) - \theta^T X)}{2} \right).$$

In the GLM notations, it yields $a = 2$ and $b(\theta^T X) = 2 \ln(1 + e^{\theta^T X}) - \theta^T X$.

4.1 Results for the Truncated Algorithm

In order to prove the convergence of the algorithm needed in the local phase, we follow a trick introduced by Bercu et al. (2020) consisting in changing slightly the update on $P_t$. Indeed, when the authors tried to prove the asymptotic convergence of the static EKF (which they named stochastic Newton step) using Robbins-Siegmund Theorem, they needed the convergence of $\sum_t \lambda_{\max}(P_t)^2$. This seems very likely to hold as we have intuitively $P_t \propto 1/t$.

However, in order to obtain $\lambda_{\max}(P_t) = O(1/t)$, one needs to lower-bound $\alpha_t$, that is, to upper-bound $|\hat{\theta}_t^T X_t|$, and that is impossible in the global logistic setting. Therefore, the idea is to force a lower-bound on $\alpha_t$ in its definition. We thus define, for some $0 < \beta < 1/2$, $t \geq 1$,

$$\alpha_t = \max \left( \frac{1}{t \beta}, \frac{1}{(1 + e^{\hat{\theta}_t^T X_t})(1 + e^{-\hat{\theta}_t^T X_t})} \right).$$

This modification yields Algorithm 3, where we keep the notations $\hat{\theta}_t, P_t, \tau(\epsilon)$ with some abuse in the rest of this section. We impose a decreasing threshold on $\alpha_t$ ($\beta > 0$) and we prove that the recursion coincides with Algorithm 1 after some steps. The sensitivity of the algorithm to $\beta$ is discussed at the end of Section 4.2. Also, note that the threshold could be $c/t^\beta$, $c > 0$, as in Bercu et al. (2020). We consider $1/t^\beta$ for clarity. We control the convergence time $\tau(\epsilon)$ of this version of the EKF:

**Proposition 9** Starting Algorithm 3 from $\hat{\theta}_1 = 0$ and any $P_1 > 0$, if Assumptions 1 and 2 are satisfied and $\mathbb{E}[XX^T]$ is invertible, for any $\epsilon, \delta > 0$, it holds $\tau(\epsilon) \leq T(\epsilon, \delta)$ along with

$$\forall t > T(\epsilon, \delta), \quad \alpha_t = \frac{1}{(1 + e^{\hat{\theta}_t^T X_t})(1 + e^{-\hat{\theta}_t^T X_t})},$$

for some $0 < \beta < 1/2$. 


with probability at least $1 - \delta$, where $T(\varepsilon, \delta) \in \mathbb{N}$ is defined in Corollary 13.

Besides the convergence of the truncated EKF, the proposition states that the truncated recursions coincide with the static EKF ones after the first $T(\varepsilon, \delta)$ steps. Thus we can apply our analysis of Section 3. We state the global result for $\varepsilon = 1/(20D_X)$:

**Theorem 10** Under the assumptions of Proposition 9, for any $\delta > 0$, it holds for any $n \geq 1$ simultaneously

$$
\sum_{t=1}^{n} L(\hat{\theta}_t) - L(\theta^*) \leq 3de^{D_X\|\theta^*\|} \ln \left(1 + n \frac{\lambda_{\text{max}}(P_1)D_X^2}{4d}\right) + \frac{\lambda_{\text{max}}(P_1^{-1})}{75D_X^2} + 64e^{D_X\|\theta^*\|} \ln \delta^{-1} + T\left(\frac{1}{20D_X}, \delta\right) \left(\frac{1}{300} + D_X\|\hat{\theta}_1 - \theta^*\|\right) + T\left(\frac{1}{20D_X}, \delta\right)^2 \frac{\lambda_{\text{max}}(P_1)D_X^2}{2},
$$

with probability at least $1 - 4\delta$, where $T(1/(20D_X), \delta)$ is defined in Corollary 13.

### 4.2 Explicit Definition of $T(\varepsilon, \delta)$ in Proposition 9

It is proved that $\|\hat{\theta}_n - \theta^*\|^2 = O(\ln n/n)$ almost surely (Bercu et al., 2020, Theorem 4.2). We don’t obtain a non-asymptotic version of this rate of convergence, but the aim of this paragraph is to prove Proposition 9 for an explicit value of $T(\varepsilon, \delta)$ for any $\delta, \varepsilon > 0$.

The objective of the truncation introduced in the algorithm is to improve the control on $P_t$. We state that fact formally with a concentration result relying on Tropp (2012). We define $\Lambda_{\min}$ the smallest eigenvalue of $E[XX^\top]$.

**Proposition 11** Under the assumptions of Proposition 9, for any $\delta > 0$, it holds simultaneously

$$
\forall t > \left(\frac{20D_X^4}{\Lambda_{\min}^2 \ln \left(\frac{625dD_X^8}{\Lambda_{\min}^4 \delta}\right)}\right)^{1/(1-\beta)}, \quad \lambda_{\text{max}}(P_t) \leq \frac{4}{\Lambda_{\min} \lambda_{\min}^{1-\beta}},
$$

with probability at least $1 - \delta$.

This proposition justifies the choice $\beta < 1/2$ in the introduction of the truncated algorithm to satisfy the condition $\sum \lambda_{\text{max}}(P_t)^2 < +\infty$ with high probability. Motivated by Proposition 11, we define, for $C > 0$, the event

$$A_C := \bigcap_{t=1}^{\infty} \left(\lambda_{\text{max}}(P_t) \leq \frac{C}{t^{1-\beta}}\right).$$

To obtain a control on $P_t$ holding for any $t$, we use the relation $\lambda_{\text{max}}(P_t) \leq \lambda_{\text{max}}(P_1)$ holding almost surely. We thus define

$$C_\delta = \max \left(\frac{4}{\Lambda_{\min}^2}, \lambda_{\text{max}}(P_1) \left(\frac{20D_X^4}{\Lambda_{\min}^2 \ln \left(\frac{625dD_X^8}{\Lambda_{\min}^4 \delta}\right)}\right)^{1/(1-\beta)}\right),$$

and we obtain $\mathbb{P}(A_{C_\delta}) \geq 1 - \delta$. We obtain the following theorem under that condition.
Theorem 12 Under the assumptions of Proposition 9, we have for any $\delta, \varepsilon > 0$ and $t \geq 1$
\[
\mathbb{P}(\|\hat{\theta}_t - \theta^*\| > \varepsilon \mid A_{C_3}) \leq (\sqrt{t} + 1) \exp \left( - \frac{\Lambda_{\min}^6 (1 - 2\beta) \varepsilon^4}{2^{16} D_X^2 C_2^2 (1 + e^{D_X (\|\theta^*\| + \varepsilon)})^6} \ln(t)^2 \right) + t \exp \left( - \frac{\Lambda_{\min}^2 (1 - 2\beta) \varepsilon^4}{2^{11} D_X^2 C_2^2 (1 + e^{D_X (\|\theta^*\| + \varepsilon)})^2 (\sqrt{t} - 1)^{1 - 2\beta}} \right).
\]

The beginning of our convergence proof starts similarly as the analysis of Bercu et al. (2020): we obtain a recursive inequality ensuring that $(L(\hat{\theta}_t))$, is decreasing in expectation. However, in order to obtain a non-asymptotic result we cannot apply Robbins-Siegmund Theorem. Instead we use the fact that the variations of the algorithm $\hat{\theta}_t$ are slow thanks to the control on $P_t$. Thus, if the algorithm was far from the optimum, the last estimates were far too which contradicts the decrease in expectation of the risk. Consequently, we look at the last $k \leq t$ such that $\|\hat{\theta}_k - \theta^*\| < \varepsilon/2$, if it exists. We decompose the probability of being outside the local region in two scenarii, yielding the two terms in Theorem 12. If $k < \sqrt{t}$, the recursive decrease in expectation makes it unlikely that the estimate stays far from the optimum for a long period. If $k > \sqrt{t}$, the control on $P_t$ allows a control on the probability that the algorithm moves fast, in $t - k$ steps, away from the optimum.

The following corollary explicitly defines a guarantee for the convergence time.

Corollary 13 Proposition 9 holds with for any $\varepsilon, \delta > 0$
\[
T(\varepsilon, \delta) = \max \left( \left( 2(1 + e^{D_X (\|\theta^*\| + \varepsilon)}) \right)^{1/\beta}, \exp \left( \frac{3 \cdot 2^{15} D_X^2 C_2^2 (1 + e^{D_X (\|\theta^*\| + \varepsilon)})^6}{\Lambda_{\min}^6 (1 - 2\beta)^{3/2} \varepsilon^4} \right), 6\delta^{-1} \right).
\]

This definition of $T(\varepsilon, \delta)$ allows a discussion on the dependence of the bound Theorem 10 to the different parameters. Note that the choice $\varepsilon = 1/(20D_X)$ in Theorem 10 is artificially made for simplifying constants since the bound actually holds for any $\varepsilon > 0$ simultaneously. The truncation has introduced an extra parameter $0 < \beta < 1/2$ that does not impact the leading term in Theorem 10. However, it impacts the first step control in an intricate way. On the one hand, when $\beta$ is close to 0, the algorithm is slow to coincide with the true EKF as $T(\varepsilon, \delta) = e^{O(1)/\beta}$. On the other hand, the larger $\beta$, the larger our control on $\lambda_{\max}(P_t)$ and thus we get $T(\varepsilon, \delta) = e^{O(1)/(1 - 2\beta)^{3/2}}$. Practical considerations show that the truncation is artificial and can even deteriorate the performance of the EKF, see Section 6. Thus Bercu et al. (2020) suggest to choose $\beta = 0.49$.

The dependence on $\delta$ is even more complex. The third constraint on $T(\varepsilon, \delta)$ is $O(\delta^{-1})$ which should not be sharp. To improve this lousy dependence in the bound, one needs a better control of $P_t$. It would follow from a specific analysis of the $O(\ln \delta^{-1})$ first recursions in order to "initialize" the control on $P_t$. However the objective of Corollary 13 was to prove Proposition 9 and not to get an optimal value of $T(\varepsilon, \delta)$. A refinement of our convergence analysis following from a tighter control on $P_t$ of the EKF than the one provided by Tropp (2012) is a very important and challenging open question.
5. Quadratic Setting

We obtain a global result for the quadratic loss where Algorithm 1 becomes the standard Kalman filter (recall that we take $\sigma^2 = 1$, that is $\ell(y, \hat{y}) = (y - \hat{y})^2 / 2$ and $a = 1$, $b' (\hat{\theta}_t^\top X_t) = \hat{\theta}_t^\top X_t, \alpha_t = 1$).

The parallel with the ridge forecaster was evoked by Diderrich (1985), and it is crucial that the static Kalman filter is the ridge regression estimator for a decaying regularization parameter. It highlights that the static EKF may be seen as an approximation of the regularized ERM:

**Proposition 14** In the quadratic setting, for any sequence $(X_t, y_t)$, starting from any $\hat{\theta}_1 \in \mathbb{R}^d$ and $P_1 > 0$, the static EKF satisfies the optimisation problem

$$
\hat{\theta}_t = \arg \min_{\theta \in \mathbb{R}^d} \left( \frac{1}{2} \sum_{s=1}^{t-1} (y_s - \theta^\top X_s)^2 + \frac{1}{2} (\theta - \hat{\theta}_1)^\top P_1^{-1} (\theta - \hat{\theta}_1) \right), \quad t \geq 1.
$$

Note that the static Kalman filter provides automatically a right choice of the ridge regularization parameter. This equivalence yields a logarithmic regret bound for the Kalman filter (Theorem 11.7 of Cesa-Bianchi and Lugosi, 2006). It follows from Lemma 5 as the quadratic loss coincides with its second-order Taylor expansion. The leading term of the bound is $d \ln n \max_t (y_t - \hat{\theta}_t^\top X_t)^2$, thus $y_t - \hat{\theta}_t^\top X_t$ needs to be bounded.

As the static Kalman filter estimator is exactly the ridge forecaster, we can also use the regularized empirical risk minimization properties to control $T(\epsilon, \delta)$. In particular, we apply the ridge analysis of Hsu et al. (2012), and we check Assumption 5:

**Proposition 15** Starting from any $\hat{\theta}_1 \in \mathbb{R}^d$ and $P_1 > 0$, if Assumptions 1, 2 and 4 hold and if $E[XX^\top]$ is invertible then Assumption 5 holds for $T(\epsilon, \delta)$ defined explicitly in Appendix D, Corollary 26.

Up to universal constants, defining $\Lambda_{\min}$ as the smallest eigenvalue of $E[XX^\top]$, we get

$$
T(\epsilon, \delta) \lesssim h \left( \frac{\epsilon^{-1}}{\Lambda_{\min}} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{\sigma^2} + \frac{D_X^2}{\Lambda_{\min}} (1 + D_{\text{app}}^2) \sqrt{\ln \delta^{-1}} + \sigma^2 d \right)
+ \left( \frac{D_X}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_X \|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|^2}{\sqrt{\sigma^2 d}} \ln \delta^{-1} \right) \right),
$$

with $h(x) = x \ln x$. We obtain a much less dramatic dependence in $\epsilon$ than in the logistic setting. However we could not avoid a $\Lambda_{\min}^{-1}$ factor in the definition of $T(\epsilon, \delta)$. It is not surprising since the convergence phase relies deeply on the behavior of $P_t$.

As for the logistic setting, we split the cumulative risk into two sums. The sum of the first terms is roughly bounded by a worst case analysis, and the sum of the last terms is estimated thanks to our local analysis (Theorem 2). However, as the loss and its gradient are not bounded we cannot obtain a similar almost sure upper-bound on the convergence phase. The sub-gaussian assumption provides a high probability bound instead.
Theorem 16 Under the assumptions of Proposition 15, for any $\varepsilon, \delta > 0$, it holds simultaneously
\[
\sum_{t=1}^{n} L(\hat{\theta}_t) - L(\theta^*) \leq \frac{15}{2} d \left( 8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2 \right) \ln \left( 1 + n \frac{\lambda_{\max}(P_1)D_X^2}{d} \right) + 5\lambda_{\max}(P_1^{-1})\varepsilon^2 \\
+ 115 \left( \sigma^2 \left( 4 + \frac{\lambda_{\max}(P_1)D_X^2}{4} \right) + D_{\text{app}}^2 + 2\varepsilon^2 D_X^2 \right) \ln \delta^{-1} \\
+ D_X^2 \left( 5\varepsilon^2 + 2\|\hat{\theta}_1 - \theta^*\|^2 + 3\lambda_{\max}(P_1)D_X\sigma \ln \delta^{-1})^2 \right) T(\varepsilon, \delta) \\
+ \frac{2\lambda_{\max}(P_1)^2D_X^4(3\sigma + D_{\text{app}})^2}{3} T(\varepsilon, \delta)^3, \quad n \geq 1,
\]
with probability at least $1 - 6\delta$.

Note that the dependence of the cumulative excess risk of the convergence phase in terms of $\delta$ is $O(\log(\delta^{-1})^3)$.

6. Experiments

We experiment the static EKF for logistic regression. Precisely, we compare the following sequential algorithms that we all initialize at 0:

- The static EKF and the truncated version (Algorithm 3). We take the default value $P_1 = I_d$ along with the value $\beta = 0.49$ suggested by Bercu et al. (2020). Note that a threshold $10^{-10}/\varepsilon^{0.49}$ as recommended by Bercu et al. (2020) would coincide with the static EKF.

- The ONS and the averaged version. The convex region of search is a ball centered in 0 and of radius $D_\theta = 1.1\|\theta^*\|$, a setting where we have good knowledge of $\theta^*$. We consider two choices of the exp-concavity constant on which the ONS crucially relies to define the gradient step size. First, we use the only available bound $e^{-D_\theta D_X}$. Second, in the settings where the step size is so small that the ONS doesn’t move, we use the exp-concavity constant $\kappa_0$ at $\theta^*$. This yields a bigger step size, though the exp-concavity is not satisfied on the region of search.

- Two Averaged Stochastic Gradient Descent as described by Bach (2014). First we test the choice of the gradient step size $\gamma = 1/(2D_X^2\sqrt{N})$ denoted by ASGD and a second version with $\gamma = \|\theta^*\|/(D_X\sqrt{N})$ denoted by ASGD oracle. Note that these algorithms are with fixed horizon, thus at each step $t$, we have to re-run the whole procedure.

6.1 Synthetic Data

We first consider well-specified data generated by the process of Bercu et al. (2020). The explanatory variables $X = (1, Z^\top)^\top$ are of dimension $d = 11$ where $Z$ is a random vector composed of 10 independent components uniformly generated in $[0, 1]$, thus $D_X = \sqrt{d}$. With this distribution for $X$ we define three synthetic settings that we evaluate:
We evaluate the different algorithms with the mean squared error $\mathbb{E}[\|\hat{\theta}_t - \theta^*\|^2]$ that we approximate by its empirical version on 100 samples. We display the results in Figure 2.

6.2 Real Data Sets

To illustrate better the robustness to misspecification, we run the same procedures on real data sets:

- **Forest cover-type** (Blackard and Dean, 1999): the feature vector is of dimension $d = 54$, and as it is a multi-class task (7 classes) we focus on classifying 2 versus all others. There are $n = 581012$ instances and we randomly split in two halves for training and testing.

- **Adult income** (Kohavi, 1996): the objective is to predict whether a person’s annual income is smaller or bigger than 50K. There are 14 explanatory variables, and we...
obtain \( d = 98 \) once categorical variables are transformed into binary variables. We use the canonical split between training (32561 instances) and testing (16281 instances).

For each data set, we standardize \( X \) such that each feature ranges from 0 to 1. At each step we sample within the training set (with replacement). We evaluate through an empirical version of \( \mathbb{E}[L(\hat{\theta}_n)] - L(\theta^*) \) estimated on 100 samples and where \( L \) is estimated on the test set, see Figure 3.

### 6.3 Summary

Our experiments show the superiority of the EKF for logistic regression compared to the ONS or to averaged SGD in all the settings we tested. We display in Table 2 a few indicators of the data sets. In particular, it is interesting that the static EKF works well even in a setting where the Hessian matrix \( H^* \) is singular.

It appears clear that low exp-concavity constants are responsible of the poor performances of the ONS. One may tune the gradient step size at the cost of losing the exp-concavity property and thus the regret guarantee of (Hazan et al., 2007) or its analogous for the cumulative risk (Mahdavi et al., 2015). Averaging is crucial for the ONS, whereas it is useless for the static EKF. Indeed we chose not to plot the averaged version of the EKF for clarity, but the EKF performs better than its averaged version.

It is important to note that in the first synthetic setting the truncation deteriorates the performance of the EKF, as well as in the adult income data set to a lesser extent, whereas the results are the same in the other settings. Bercu et al. (2020) argue that the truncation is artificially introduced for the convergence property, thus they use the threshold \( 10^{-10}/t^{0.49} \) instead of \( 1/t^{0.49} \) and the truncated version almost coincides with the true EKF. We confirm here that the truncation may be damaging if the threshold is set too high and
Figure 3: Excess test risk for forest cover type (left) and adult income (right). As the ONS doesn’t move when applied with the exp-concavity constant $e^{-D_\theta D_X}$ we use instead the exp-concavity constant at $\theta^\star$: $\kappa_0 \approx 1.4 \cdot 10^{-3}$ for forest cover type and $\kappa_0 \approx 5.5 \cdot 10^{-6}$ for adult income. The EKF and the truncated version almost coincide for both data sets.

<table>
<thead>
<tr>
<th>Setting</th>
<th>$d$</th>
<th>$\lambda_{\text{max}}(H^\star)/\mu$</th>
<th>$\text{tr}(G^\star H^{\star-1})$</th>
<th>$R^2/\mu$</th>
<th>$de^{D_\theta D_X}$</th>
<th>$d\kappa_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic well-specified 1</td>
<td>11</td>
<td>$6.9 \cdot 10^2$</td>
<td>$1.7 \cdot 10^2$</td>
<td>$1.0 \cdot 10^3$</td>
<td>$9.2 \cdot 10^{17}$</td>
<td>$6.4 \cdot 10^{15}$</td>
</tr>
<tr>
<td>Synthetic well-specified 2</td>
<td>11</td>
<td>$1.5 \cdot 10^2$</td>
<td>$7.1 \cdot 10^1$</td>
<td>$2.5 \cdot 10^3$</td>
<td>$5.4 \cdot 10^4$</td>
<td>$3.3 \cdot 10^2$</td>
</tr>
<tr>
<td>Synthetic misspecified 3</td>
<td>11</td>
<td>$1.5 \cdot 10^2$</td>
<td>$7.1 \cdot 10^1$</td>
<td>$2.0 \cdot 10^3$</td>
<td>$3.6 \cdot 10^3$</td>
<td>$7.4 \cdot 10^1$</td>
</tr>
<tr>
<td>Forest cover type</td>
<td>54</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$9.2 \cdot 10^{32}$</td>
<td>$3.8 \cdot 10^{14}$</td>
</tr>
<tr>
<td>Adult income</td>
<td>98</td>
<td>$2.5 \cdot 10^4$</td>
<td>$7.2 \cdot 10^5$</td>
<td>$5.3 \cdot 10^8$</td>
<td>$1.8 \cdot 10^{32}$</td>
<td>$1.8 \cdot 10^{14}$</td>
</tr>
</tbody>
</table>

Table 2: For the different experimental settings we display the dimension $d$ and the condition number of the Hessian at $\theta^\star$ ($\lambda_{\text{max}}(H^\star)$ and $\mu$ are the maximal and minimal eigenvalues of $H^\star$). We present the value of $\text{tr}(G^\star H^{\star-1})$ which is bounded either by $R^2/\mu$, or by $de^{D_\theta D_X}$ because $e^{-D_\theta D_X}$ bounds the exp-concavity constant on the centered ball of radius $D_\theta$. We add to the table $d\kappa_0 \leq de^{D_\theta D_X}$ where $\kappa_0$ is the inverse of the exp-concavity constant of the loss at $\theta^\star$. 

Synthetic well-specified 1

Synthetic well-specified 2

Synthetic misspecified 3

Forest cover type

Adult income
we recommend to use the EKF in practice, or equivalently the truncated version with the low threshold suggested by Bercu et al. (2020).

7. Conclusion

We studied an efficient way to tackle some unconstrained optimization problems, in which we get rid of the projection step of bounded algorithm such as the ONS. We presented a bayesian approach where we transformed the loss into a negative log-likelihood. We used the Kalman recursion to provide a parameter free approximation of the maximum-likelihood estimator. We demonstrated the optimality of the local phase for locally exp-concave losses which can be expressed as GLM log-likelihoods. We proved the finiteness of the convergence phase in logistic and quadratic regressions. We illustrated our theoretical results with numerical experiments for logistic regression. It would be interesting to generalize our results to a larger class of optimization problems.

Finally, this article aimed at strengthening the bridge between Kalman recursion and the optimization community. Therefore we made the i.i.d. assumption, standard in the stochastic optimization literature and we focus on the static EKF. It may lead the way to a risk analysis of the general EKF in non i.i.d. state-space models.
The Appendix follows the structure of the article:

- Appendix A presents the EKF for generalized linear models.
- Appendix B contains the proofs of Section 3. Precisely, Lemma 3 is proved in Section B.1. the intermediate results of Sections 3.1 and 3.2 are proved in Sections B.2 and B.3. then Theorem 1 is proved in Section B.4 and Theorem 2 in Section B.5.
- Appendix C contains the proofs of Section 4. We derive the global bound (Theorem 10) in Section C.1. then we obtain the concentration result on $P_1$ in Section C.2. and finally we prove the convergence of the truncated algorithm in Section C.3.
- Appendix D contains the proofs of Section 5. We prove Theorem 16 in Section D.1 and then in Section D.2 we prove the convergence of the algorithm, and we define an explicit value of $T(\epsilon, \delta)$ satisfying Assumption 5.

Appendix A. Derivation of the Static EKF for Generalized Linear Models

As in Section 10.2 of Durbin and Koopman (2012) we consider the following state-space model:

\[ y_t = Z_t(\theta_t) + \varepsilon_t, \]
\[ \theta_{t+1} = T_t(\theta_t) + \eta_t. \]

where $\varepsilon_t$ and $\eta_t$ are independent with mean zero and variances $h_t(\theta_t), Q_t(\theta_t)$. The state-space version of equation (2) is

\[ p(y_t \mid X_t) = h(y_t) \exp \left( \frac{y_t \theta_t^\top X_t - b(\theta_t^\top X_t)}{a} \right). \]

The preceding equation matches the space equation form with $Z_t(\theta_t) = b'(\theta_t^\top X_t)$ and $h_t(\theta_t) = ab''(\theta_t^\top X_t)$. Thus we can write the EKF as follows (see Equation 10.4 of Durbin and Koopman, 2012): denoting by $T_t$ the derivative of $T_t$,

\[ v_t = y_t - b'(\hat{\theta}_t^\top X_t), \]
\[ F_t = X_t^\top P_t X_t b''(\hat{\theta}_t^\top X_t)^2 + ab''(\hat{\theta}_t^\top X_t), \]
\[ \hat{\theta}_{t|t} = \hat{\theta}_t + P_t X_t b''(\hat{\theta}_t^\top X_t) F_t^{-1} v_t, \]
\[ P_{t|t} = P_t - P_t X_t F_t^{-1} X_t^\top P_t b''(\hat{\theta}_t^\top X_t)^2, \]
\[ \hat{\theta}_{t+1} = T_t(\hat{\theta}_{t|t}), \]
\[ P_{t+1} = T_t P_{t|t} T_t^\top + Q_t(\hat{\theta}_{t|t}). \]

We focus on the static setting where the state equation becomes $\theta_{t+1} = \theta_t$, thus we have $\hat{\theta}_{t+1} = \hat{\theta}_{t|t}$ and $P_{t+1} = P_{t|t}$. We rewrite the update on $P_t$ as follows:

\[ P_{t+1} = P_t - \frac{P_t X_t X_t^\top P_t b''(\hat{\theta}_t^\top X_t)}{X_t^\top P_t X_t b''(\hat{\theta}_t^\top X_t) + a}. \]

Moreover we have $P_{t+1} X_t = P_t X_t F_t^{-1} ab''(\hat{\theta}_t^\top X_t)$ thus we can rewrite the update on $\hat{\theta}_t$ as follows:

\[ \hat{\theta}_{t+1} = \hat{\theta}_t + \frac{P_{t+1} X_t (y_t - b'(\hat{\theta}_t^\top X_t))}{a}. \]

This yields Algorithm 1.
Appendix B. Proofs of Section 3

B.1 Proof of Lemma 3

We prove the following Lemma inspired by the stopping time technique of Freedman (1975) from which we derive Lemma 3. We give a general form useful in several proofs.

Lemma 17 Let \((F_n)\) be a filtration, and we consider a sequence of events \((A_n)\) that is adapted to \((F_n)\). Let \((V_n)\) be a sequence of random variables adapted to \((F_n)\) satisfying \(V_0 = 1, V_n \geq 0\) almost surely for any \(n\), and

\[
\mathbb{E}[V_n \mid F_{n-1}, A_{n-1}] \leq V_{n-1}, \quad n \geq 1.
\]

Then for any \(\delta > 0\), it holds

\[
\mathbb{P}\left(\bigcup_{n=1}^{\infty} V_n > \delta^{-1}\right) \leq \delta + \mathbb{P}\left(\bigcup_{n=0}^{\infty} A_n\right).
\]

An important particular case is when \((V_n)\) is a super-martingale adapted to the filtration \((F_n)\) satisfying \(V_0 = 1\) and \(V_n \geq 0\) almost surely: then we have simultaneously \(V_n \leq \delta^{-1}\) for \(n \geq 1\) with probability larger than \(1 - \delta\).

Proof We define

\[
E_k = \bigcup_{n=1}^{k} (V_n > \delta^{-1} \cup A_{n-1}^-).
\]

As \((E_k)\) is increasing, we have, for any \(k \geq 1\),

\[
\mathbb{P}(E_k) = \sum_{n=1}^{k} \mathbb{P}(E_n \cap E_{n-1}^c)
\]

\[
= \sum_{n=1}^{k} \mathbb{P}(A_{n-1}^- \cap E_{n-1}^c) + \sum_{n=1}^{k} \mathbb{P}(V_n > \delta^{-1} \cap E_{n-1}^c \cap A_{n-1}^-).
\]

First, we have

\[
\sum_{n=1}^{k} \mathbb{P}(A_{n-1}^- \cap E_{n-1}^c) \leq \mathbb{P}\left(\bigcup_{n=0}^{k-1} A_n^-\right).
\]

Second, we apply Markov’s inequality:

\[
\sum_{n=1}^{k} \mathbb{P}(V_n > \delta^{-1} \cap E_{n-1}^c \cap A_{n-1}^-) \leq \sum_{n=1}^{k} \mathbb{E}\left[\frac{V_n}{\delta^{-1} \mathbb{1}_{E_n \cap E_{n-1}^c \cap A_{n-1}^-}}\right]
\]

\[
= \delta \sum_{n=1}^{k} \mathbb{E}\left[V_n \mathbb{1}_{E_{n-1}^c \cap A_{n-1}^-} - \mathbb{1}_{E_n^-}\right]
\]

\[
= \delta \sum_{n=1}^{k} \left(\mathbb{E}\left[V_n \mathbb{1}_{E_{n-1}^c \cap A_{n-1}^-}\right] - \mathbb{E}\left[V_n \mathbb{1}_{E_n^-}\right]\right).
\]
The second line is obtained since $E_n \subset (E_{n-1} \cap A_{n-1})$. According to the tower property and the super-martingale assumption,

$$
\mathbb{E} \left[ V_n 1_{E_{n-1} \cap A_{n-1}} \right] = \mathbb{E} \left[ \mathbb{E} [V_n | F_{n-1}, A_{n-1}] 1_{E_{n-1} \cap A_{n-1}} \right] \\
\leq \mathbb{E} \left[ \mathbb{E} [V_n | F_{n-1}, A_{n-1}] E_{n-1} \right] \\
\leq \mathbb{E} \left[ V_{n-1} 1_{E_{n-1}} \right].
$$

Therefore, a telescopic argument along with $V_0 = 1$ and $V_k 1_{E_k} \geq 0$ yields

$$
\sum_{n=1}^{k} \mathbb{P} \left( V_n > \delta^{-1} \cap E_{n-1} \cap A_{n-1} \right) \leq \delta.
$$

Finally, for any $k \geq 1$, we obtain

$$
\mathbb{P} (E_k) \leq \mathbb{P} \left( \bigcup_{n=0}^{k-1} A_n \right) + \delta
$$

and the desired result follows by letting $k \to \infty$.

**Proof of Lemma 3.** Let $\lambda > 0$. For any $n \geq 1$, we define

$$
V_n = \exp \left( \sum_{t=k+1}^{k+n} \left( \lambda \Delta N_t - \frac{\lambda^2}{2} ((\Delta N_t)^2 + \mathbb{E}[(\Delta N_t)^2 | F_{t-1}]) \right) \right).
$$

Lemma B.1 of Bercu and Touati (2008) states that $(V_n)$ is a super-martingale adapted to the filtration $(F_{k+n})$. Moreover $V_0 = 1$ and for any $n$, it holds $V_n \geq 0$ almost surely. Therefore we can apply Lemma 17.

**B.2 Proof of Lemma 5**

**Proof of Lemma 5.** We start from the update formula $\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1} (y_t - b'(\hat{\theta}_t^T X_t)) X_t$ yielding

$$
(\hat{\theta}_{t+1} - \theta^*)^T P_{t+1}^{-1} (\hat{\theta}_{t+1} - \theta^*) = (\hat{\theta}_t - \theta^*)^T P_{t+1}^{-1} (\hat{\theta}_t - \theta^*) + 2 \frac{(y_t - b'(\hat{\theta}_t^T X_t)) X_t^T}{a} (\hat{\theta}_t - \theta^*)
$$

$$
+ X_t^T P_{t+1} X_t \left( \frac{y_t - b'(\hat{\theta}_t^T X_t)}{a} \right)^2.
$$
With a summation argument, re-arranging terms, we obtain:

\[
\sum_{t=1}^{n} \left( \frac{(b'(\hat{\theta}_t^\top X_t) - y_t)X_t^\top}{a} (\hat{\theta}_t - \theta^*) - \frac{1}{2}(\hat{\theta}_t - \theta^*)^\top (P^{-1}_{t+1} - P^{-1}_t)(\hat{\theta}_t - \theta^*) \right)
\]

\[
= \frac{1}{2} \sum_{t=1}^{n} X_t^\top P_{t+1} X_t \left( \frac{y_t - b'(\hat{\theta}_t^\top X_t)}{a} \right)^2
\]

\[
+ \frac{1}{2} \sum_{t=1}^{n} \left( (\hat{\theta}_t - \theta^*)^\top P^{-1}_{t+1}(\hat{\theta}_t - \theta^*) - (\hat{\theta}_{t+1} - \theta^*)^\top P^{-1}_{t+1}(\hat{\theta}_{t+1} - \theta^*) \right)
\]

We bound the telescopic sum: as \( P^{-1}_{n+1} \succ 0 \), we have

\[
\sum_{t=1}^{n} \left( (\hat{\theta}_t - \theta^*)^\top P^{-1}_{t}(\hat{\theta}_t - \theta^*) - (\hat{\theta}_{t+1} - \theta^*)^\top P^{-1}_{t+1}(\hat{\theta}_{t+1} - \theta^*) \right)
\]

\[
\leq (\hat{\theta}_1 - \theta^*)^\top P^{-1}_1(\hat{\theta}_1 - \theta^*) \leq \frac{||\hat{\theta}_1 - \theta^*||^2}{\lambda_{\text{min}}(P_1)}
\]

The result follows from the identities

\[
\frac{(b'(\hat{\theta}_t^\top X_t) - y_t)X_t}{a} = \ell'(y_t, \hat{\theta}_t^\top X_t)X_t, \quad P^{-1}_{t+1} - P^{-1}_t = \ell''(y_t, \hat{\theta}_t^\top X_t)X_t X_t^\top.
\]

**B.3 Proofs of Section 3.2**

**Proof of Proposition 6.** The first-order condition satisfied by \( \theta^* \) is

\[
\mathbb{E} \left[ \frac{(y - b'(\theta^*^\top X)X)}{a} \right] = 0,
\]

yielding \( \mathbb{E} [yX] = \mathbb{E}[b'(\theta^*^\top X)X] \). Therefore

\[
\mathbb{E} \left[ \frac{(b'(\theta^\top X) - y)X}{a} \right]^\top (\theta - \theta^*) = \frac{1}{a}(\theta - \theta^*)^\top \mathbb{E} \left[ X(b'(\theta^\top X) - b'(\theta^*^\top X)) \right].
\]

Considering the function \( f : \lambda \to (\theta - \theta^*)^\top \mathbb{E} \left[ Xb'(\theta^\top X + \lambda(\theta - \theta^*)^\top X) \right] \), we know there exists \( \lambda \in [0, 1] \) such that \( f'(\lambda) = f(1) - f(0) \). This translates into

\[
\frac{\partial L}{\partial \theta}^\top (\theta - \theta^*) = \frac{1}{a}(\theta - \theta^*)^\top \mathbb{E} \left[ Xb''(\theta^\top X + \lambda(\theta^* - \theta)^\top X) (\theta - \theta^*)^\top X \right].
\]

Then we use Assumption 3:

\[
\frac{b''(\theta^\top X + \lambda(\theta^* - \theta)^\top X)}{b''(\theta^\top X)} = \frac{\ell''(y_t, \theta^\top X + \lambda(\theta^* - \theta)^\top X)}{\ell''(y_t, \theta^\top X)} \geq \rho_{\|\theta - \theta^*\|},
\]

\[
\mathbb{E} \left[ Xb''(\theta^\top X + \lambda(\theta^* - \theta)^\top X) (\theta - \theta^*)^\top X \right] \geq \rho_{\|\theta - \theta^*\|}.
\]
yielding
\[
\frac{\partial L}{\partial \theta}^T_{\theta} (\theta - \theta^*) \geq \rho_{\|\theta - \theta^*\|} (\theta - \theta^*)^T \mathbb{E} \left[ \ell''(y, \theta^T X) XX^T \right] (\theta - \theta^*) \\
= \rho_{\|\theta - \theta^*\|} (\theta - \theta^*)^T \frac{\partial^2 L}{\partial \theta^2} \bigg|_{\theta} (\theta - \theta^*) .
\]

\[\square\]

**Proof of Proposition 7.** We first recall that \( L(\theta) - L(\theta^*) \leq \frac{\partial L}{\partial \theta}^T_{\theta} (\theta - \theta^*) \), then Proposition 6 yields
\[
\frac{\partial L}{\partial \theta}^T_{\theta} (\theta - \theta^*) - c(\theta - \theta^*)^T \frac{\partial^2 L}{\partial \theta^2} \bigg|_{\theta} (\theta - \theta^*) \geq (1 - \frac{c}{\rho_{\|\theta - \theta^*\|}^2}) \frac{\partial L}{\partial \theta}^T_{\theta} (\theta - \theta^*) ,
\]
and the result follows.

\[\square\]

**Proof of Lemma 8.** We first develop \((\Delta M_t)^2\):
\[
(\Delta M_t)^2 = \left( \left[ \mathbb{E} [\nabla_t | F_{t-1}] - \nabla_t \right]^T (\hat{\theta}_t - \theta^*) \right)^2 \\
= (\hat{\theta}_t - \theta^*)^T \left( \mathbb{E} [\nabla_t | F_{t-1}] \mathbb{E} [\nabla_t | F_{t-1}]^T + \nabla_t \nabla_t^T \\
- \nabla_t \mathbb{E} [\nabla_t | F_{t-1}]^T - \mathbb{E} [\nabla_t | F_{t-1}] \nabla_t^T \right) (\hat{\theta}_t - \theta^*) \\
\leq 2(\hat{\theta}_t - \theta^*)^T \left( \mathbb{E} [\nabla_t \nabla_t^T | F_{t-1}] + \nabla_t \nabla_t^T \right) (\hat{\theta}_t - \theta^*) \\
\leq 2(\hat{\theta}_t - \theta^*)^T \left( \mathbb{E} [\nabla_t \nabla_t^T | F_{t-1}] \right) (\hat{\theta}_t - \theta^*) .
\]

The third line holds because if \( U, V \in \mathbb{R}^d \), it holds \(-UV^T - VU^T \preceq UU^T + VV^T\). The last one comes from \( \mathbb{E} \left[ (\nabla_t - \mathbb{E} [\nabla_t | F_{t-1}]) (\nabla_t - \mathbb{E} [\nabla_t | F_{t-1}])^T \mid F_{t-1} \right] \succeq 0\).

Also, we have the relation
\[
\mathbb{E} [(\Delta M_t)^2 | F_{t-1}] \leq (\hat{\theta}_t - \theta^*)^T \mathbb{E} [\nabla_t \nabla_t^T | F_{t-1}] (\hat{\theta}_t - \theta^*) .
\]

It yields
\[
(\Delta M_t)^2 + \mathbb{E} [(\Delta M_t)^2 | F_{t-1}] \leq (\hat{\theta}_t - \theta^*)^T \left( 3\mathbb{E} [\nabla_t \nabla_t^T | F_{t-1}] + 2\nabla_t \nabla_t^T \right) (\hat{\theta}_t - \theta^*) ,
\]
and the result follows from Lemma 3.

\[\square\]

We derive the following Lemma in order to control the right-hand side of Lemma 5, in both settings.

**Lemma 18** Assume the second point of Assumption 3 holds. For any \( k, n \geq 1 \), if \( \|\hat{\theta}_t - \theta^*\|^2 \leq \varepsilon \) for any \( k < t \leq k + n \) then we have
\[
\sum_{t=k+1}^{k+n} \text{Tr} \left( P_{t+1} (P_{t+1}^{-1} - P_t^{-1}) \right) \leq d \ln \left( 1 + n \frac{h_\varepsilon \lambda_{\max}(P_{k+1}) D_X^2}{d} \right) .
\]
Proof. We apply Lemma 11.11 of Cesa-Bianchi and Lugosi (2006):

\[
\sum_{t=k+1}^{k+n} \text{Tr}(P_{t+1}(P_t^{-1} - P_{t+1}^{-1})) = \sum_{t=k+1}^{k+n} \left( 1 - \frac{\det(P_t^{-1})}{\det(P_{t+1}^{-1})} \right)
\]

\[
\leq \sum_{t=k+1}^{k+n} \ln \left( \frac{\det(P_{t+1}^{-1})}{\det(P_t^{-1})} \right)
\]

\[
= \ln \left( \frac{\det(P_{k+n+1}^{-1})}{\det(P_k^{-1})} \right)
\]

\[
\leq \ln \det \left( I + \sum_{t=k+1}^{k+n} \ell''(y_t, \hat{\theta}_t^T X_t)(P_{k+1}^{1/2} X_t)(P_{k+1}^{1/2} X_t)^\top \right)
\]

\[
= \sum_{i=1}^d \ln(1 + \lambda_i),
\]

where \(\lambda_1, ..., \lambda_d\) are the eigenvalues of \(\sum_{t=k+1}^{k+n} \ell''(y_t, \hat{\theta}_t^T X_t)(P_{k+1}^{1/2} X_t)(P_{k+1}^{1/2} X_t)^\top\). Therefore we have

\[
\sum_{t=k+1}^{k+n} \text{Tr}(P_{t+1}(P_t^{-1} - P_{t+1}^{-1})) \leq d \ln \left(1 + \frac{1}{d} \sum_{i=1}^d \lambda_i \right)
\]

\[
\leq d \ln \left(1 + \frac{1}{d} n h_\varepsilon \lambda_{\max}(P_{k+1}) D_X^2 \right).
\]

B.4 Bounded Setting (Assumption 3)

Proof of Theorem 1. Let \(\delta > 0\). On the one hand, we sum Lemma 5 and 8. We obtain, for any \(\lambda > 0\),

\[
\sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} \left( \mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^\top(\hat{\theta}_t - \theta^*) - \frac{1}{2} Q_t 
\]

\[
- \lambda(\hat{\theta}_t - \theta^*)^\top \left( \nabla_t \nabla_t^\top + \frac{3}{2} \mathbb{E}[\nabla_t \nabla_t^\top \mid \mathcal{F}_{t-1}] \right)(\hat{\theta}_t - \theta^*) \right)
\]

\[
\leq \frac{1}{2} \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} X_t^\top P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^T X_t)^2 + \frac{\|\hat{\theta}_t - \theta^*\|^2}{\lambda_{\min}(P_{T(\varepsilon, \delta)+1}) + \frac{\ln \delta^{-1}}{\lambda}, \quad n \geq 1, \quad (3)
\]

with probability at least \(1 - \delta\), where we define \(Q_t = (\hat{\theta}_t - \theta^*)^\top \left( \ell''(y_t, \hat{\theta}_t^T X_t)X_tX_t^\top \right)(\hat{\theta}_t - \theta^*)\) for any \(t\).
On the other hand, thanks to Assumption 3, we can apply Proposition 7 with \( c = 0.75 \) to obtain, for any \( t \geq 1 \),
\[
\|\hat{\theta}_t - \theta^*\| \leq \varepsilon \\
\implies L(\hat{\theta}_t) - L(\theta^*) \leq \frac{\rho_\varepsilon}{\rho_\varepsilon - 0.75} (\frac{\partial L}{\partial \theta}^\top (\hat{\theta}_t - \theta^*) - 0.75(\hat{\theta}_t - \theta^*)^\top \frac{\partial^2 L}{\partial \theta^2} (\hat{\theta}_t - \theta^*)) , \\
\implies L(\hat{\theta}_t) - L(\theta^*) \leq 5\left( E[\nabla_t | F_{t-1}]^\top (\hat{\theta}_t - \theta^*) - 0.75E[Q_t | F_{t-1}] \right) ,
\]
(4)
because \( \rho_\varepsilon > 0.95 \).

In order to bridge the gap between Equations (3) and (4), we need to control the quadratic terms of Equation (3) with \( E[Q_t | F_{t-1}] \). First, for any \( t \), if \( \|\hat{\theta}_t - \theta^*\| \leq \varepsilon \), we have \( Q_t \in [0, h_\varepsilon^2 D_X^2] \), and we apply Lemma A.3 of Cesa-Bianchi and Lugosi (2006) to the random variable \( \frac{1}{h_\varepsilon^2 D_X} Q_t \in [0, 1] \): for any \( s > 0 \),
\[
E \left[ \exp \left( \frac{s}{h_\varepsilon^2 D_X} Q_t - \frac{e^s - 1}{h_\varepsilon^2 D_X} E[Q_t | F_{t-1}] \right) \bigg| F_{t-1}, \|\hat{\theta}_t - \theta^*\| \leq \varepsilon \right] \leq 1 .
\]
We fix \( s = 0.1 \) and we define
\[
V_n = \exp \left( \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} \frac{0.1}{h_\varepsilon^2 D_X} Q_t - \frac{e^{0.1} - 1}{h_\varepsilon^2 D_X} E \left[ \frac{1}{h_\varepsilon^2 D_X} Q_t | F_{t-1} \right] \right) .
\]
The sequence \( (V_n) \) is adapted to \( (F_{T(\varepsilon, \delta)+n}) \), almost surely we have \( V_0 = 1 \) and \( V_n \geq 0 \). Finally,
\[
E \left[ V_n | F_{T(\varepsilon, \delta)+n-1}, \|\hat{\theta}_{T(\varepsilon, \delta)+n} - \theta^*\| \leq \varepsilon \right] \leq V_{n-1} ,
\]
and \( (\|\hat{\theta}_{T(\varepsilon, \delta)+n} - \theta^*\| \leq \varepsilon) \) belongs to \( F_{T(\varepsilon, \delta)+n-1} \). We apply Lemma 17
\[
P \left( \bigcup_{n=1}^{\infty} V_n > \delta^{-1} \right) \cup \left( \bigcup_{n=1}^{\infty} (\|\hat{\theta}_{T(\varepsilon, \delta)+n} - \theta^*\| > \varepsilon) \right) \leq \delta + P \left( \bigcup_{n=1}^{\infty} (\|\hat{\theta}_{T(\varepsilon, \delta)+n} - \theta^*\| > \varepsilon) \right) .
\]
We define \( A_k^\varepsilon = \bigcap_{n=k+1}^{\infty} (\|\hat{\theta}_n - \theta^*\| \leq \varepsilon) \) for any \( k \). The last inequality is equivalent to
\[
P \left( \bigcup_{n=1}^{\infty} \left( \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} Q_t > 10(e^{0.1} - 1) \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} E \left[ Q_t | F_{t-1} \right] + 10h_\varepsilon^3 D_X^2 \ln \delta^{-1} \right) \cap A_k^\varepsilon \right) \leq \delta .
\]
(5)
We then bound the two quadratic terms coming from Lemma 8: using Assumption 3 we have the implications
\[
\|\hat{\theta}_t - \theta^*\| \leq \varepsilon \implies (\hat{\theta}_t - \theta^*)^\top \nabla_t \nabla_t^\top (\hat{\theta}_t - \theta^*) \leq \kappa_\varepsilon Q_t ,
\]
\[
\|\hat{\theta}_t - \theta^*\| \leq \varepsilon \implies (\hat{\theta}_t - \theta^*)^\top E[\nabla_t \nabla_t^\top | F_{t-1}] (\hat{\theta}_t - \theta^*) \leq \kappa_\varepsilon E[Q_t | F_{t-1}] .
\]
Therefore, we get from (5)

\[
\mathbb{P}\left( \bigcup_{n=1}^{\infty} \sum_{t=T(\varepsilon,\delta)+n}^{T(\varepsilon,\delta)+n} \left( \frac{1}{2}Q_t + \lambda (\hat{\theta}_t - \theta^*)^\top \nabla_t \nabla_t^\top (\hat{\theta}_t - \theta^*) \right) + \frac{3}{2} \lambda (\hat{\theta}_t - \theta^*)^\top \mathbb{E}[\nabla_t \nabla_t^\top | F_{t-1}] (\hat{\theta}_t - \theta^*) > 0 \right) > \\
\left( 10(e^{0.1} - 1) \left( \frac{1}{2} + \lambda \kappa \right) + \frac{3}{2} \lambda \kappa \right) \mathbb{E}[Q_t | F_{t-1}] \\
+ 10 \left( \frac{1}{2} + \lambda \kappa \right) h_x \varepsilon^2 D_X^2 \ln \delta^{-1} \right) \cap A_{T(\varepsilon,\delta)}^\varepsilon \leq \delta.
\]

We set \( \lambda = \frac{0.75 - 5(e^{0.1} - 1)}{(10(e^{0.1} - 1) + \frac{3}{2})\kappa} \), so that

\[
10(e^{0.1} - 1) \left( \frac{1}{2} + \lambda \kappa \right) + \frac{3}{2} \lambda \kappa = 0.75,
\]

\[
\frac{1}{2} + \lambda \kappa = \frac{1}{2} + \frac{0.75 - 5(e^{0.1} - 1)}{10(e^{0.1} - 1) + \frac{3}{2}} \approx 0.59 \leq 0.6,
\]

and consequently

\[
\mathbb{P}\left( \bigcup_{n=1}^{\infty} \sum_{t=T(\varepsilon,\delta)+n}^{T(\varepsilon,\delta)+n} \left( \mathbb{E}[\nabla_t | F_{t-1}]^\top (\hat{\theta}_t - \theta^*) - 0.75 \mathbb{E}[Q_t | F_{t-1}] \right) > 6h_x \varepsilon^2 D_X^2 \ln \delta^{-1} \\
+ \sum_{t=T(\varepsilon,\delta)+n}^{T(\varepsilon,\delta)+n} \left( \mathbb{E}[\nabla_t | F_{t-1}]^\top (\hat{\theta}_t - \theta^*) - \frac{1}{2} Q_t \\
- \lambda (\hat{\theta}_t - \theta^*)^\top \left( \nabla_t \nabla_t^\top + \frac{3}{2} \mathbb{E}[\nabla_t \nabla_t^\top | F_{t-1}] (\hat{\theta}_t - \theta^*) \right) \right) \cap A_{T(\varepsilon,\delta)}^\varepsilon \right) \leq \delta.
\]

We plug Equation (4) in the last inequality:

\[
\mathbb{P}\left( \bigcup_{n=1}^{\infty} \sum_{t=T(\varepsilon,\delta)+n}^{T(\varepsilon,\delta)+n} \left( L(\hat{\theta}_t) - L(\theta^*) \right) > 30h_x \varepsilon^2 D_X^2 \ln \delta^{-1} \\
+ 5 \sum_{t=T(\varepsilon,\delta)+n}^{T(\varepsilon,\delta)+n} \left( \mathbb{E}[\nabla_t | F_{t-1}]^\top (\hat{\theta}_t - \theta^*) - \frac{1}{2} Q_t \\
- \lambda (\hat{\theta}_t - \theta^*)^\top \left( \nabla_t \nabla_t^\top + \frac{3}{2} \mathbb{E}[\nabla_t \nabla_t^\top | F_{t-1}] (\hat{\theta}_t - \theta^*) \right) \right) \cap A_{T(\varepsilon,\delta)}^\varepsilon \right) \leq \delta.
\]
Lemma 1.4 of Rigollet and Hutter (2015): for any \( \epsilon > 0 \) and \( \delta > 0 \), there exists a constant \( C > 0 \) such that

\[
\mathbb{P}
\left(\bigcup_{n=1}^{\infty} \mathbb{P}
\left(\sum_{t=T(\epsilon, \delta)+1}^{T(\epsilon, \delta)+n} (L(\hat{\theta}_t) - L(\theta^*)) > \frac{5}{2} \sum_{t=T(\epsilon, \delta)+1}^{T(\epsilon, \delta)+n} X_t^\top P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^\top X_t)^2
\right) + \frac{5\|\hat{\theta}_1 - \theta^*\|^2}{\lambda_{\text{min}}(P_{T(\epsilon, \delta)+1})} + 30(2\kappa_{\epsilon} + h_{\epsilon}\varepsilon D_X^2) \ln \delta^{-1}\right) \cap A^\epsilon_{T(\epsilon, \delta)} \right) \leq 2\delta.
\]

Thanks to Assumption 3, we have

\[
X_t^\top P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^\top X_t)^2 \leq \kappa_{\epsilon} \text{Tr} \left( P_{t+1}(P_{t+1} - P_t^{-1}) \right), \quad t > T(\epsilon, \delta),
\]

therefore we apply Lemma 18 for any \( n \), it holds

\[
\sum_{t=T(\epsilon, \delta)+1}^{T(\epsilon, \delta)+n} X_t^\top P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^\top X_t)^2 \leq d\kappa_{\epsilon} \ln \left( 1 + n \frac{h_{\epsilon}\lambda_{\text{max}}(P_{T(\epsilon, \delta)+1})D_X^2}{d} \right),
\]

As \( P_{T(\epsilon, \delta)+1} \ll P_1 \), we obtain

\[
\mathbb{P}
\left(\bigcup_{n=1}^{\infty} \mathbb{P}
\left(\sum_{t=T(\epsilon, \delta)+1}^{T(\epsilon, \delta)+n} (L(\hat{\theta}_t) - L(\theta^*)) > \frac{5}{2} \sum_{t=T(\epsilon, \delta)+1}^{T(\epsilon, \delta)+n} X_t^\top P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^\top X_t)^2
\right) + \frac{5\|\hat{\theta}_1 - \theta^*\|^2}{\lambda_{\text{min}}(P_{T(\epsilon, \delta)+1})} + 30(2\kappa_{\epsilon} + h_{\epsilon}\varepsilon D_X^2) \ln \delta^{-1}\right) \cap A^\epsilon_{T(\epsilon, \delta)} \right) \leq 2\delta.
\]

To conclude, we use Assumption 5. \( \blacksquare \)

B.5 Quadratic Setting (Assumption 4)

We recall two definitions introduced in the previous subsection:

\[
A^\epsilon_k = \bigcap_{n=k+1}^{\infty} \{\|\hat{\theta}_n - \theta^*\| \leq \epsilon\}, \quad k \geq 1,
\]

\[
Q_t = (\hat{\theta}_t - \theta^*)^\top X_t X_t^\top (\hat{\theta}_t - \theta^*), \quad t \geq 1.
\]

The sub-gaussian hypothesis requires a different treatment of several steps in the proof. In the following proofs, we use a consequence of the first points of Assumption 4. We apply Lemma 1.4 of Rigollet and Hutter (2015): for any \( X \in \mathbb{R}^d \),

\[
\mathbb{E}[(y - \mathbb{E}[y \mid X])^2 \mid X] \leq 2i(2\sigma^2)^i \Gamma(i) = 2(2\sigma^2)^i i!, \quad i \in \mathbb{N}^*.
\]  

(6)

First, we control the quadratic terms in \( \nabla_t = -(y_t - \hat{\theta}_t^\top X_t)X_t \) in the following lemma.
Lemma 19  
1. For any \( k \in \mathbb{N} \) and \( \delta > 0 \), we have
\[
P\left( \bigcup_{n=1}^{\infty} \left( \sum_{t=k+1}^{k+n} (\hat{\theta}_t - \theta^*)^\top \nabla_t \nabla_t^\top (\hat{\theta}_t - \theta^*) \right) > 3 \left( 8\sigma^2 + \epsilon^2 D_{\text{app}}^2 + \epsilon^2 D_X^2 \right) \sum_{t=k+1}^{k+n} Q_t + 12\epsilon^2 D_X^2 \sigma^2 \ln \delta^{-1} \right) \cap A_k \right) \leq \delta.
\]

2. For any \( t \), it holds almost surely
\[
(\hat{\theta}_t - \theta^*)^\top \mathbb{E}[\nabla_t \nabla_t^\top | \mathcal{F}_{t-1}](\hat{\theta}_t - \theta^*) \leq 3 \left( \sigma^2 + D_{\text{app}}^2 + \|\hat{\theta}_t - \theta^*\|^2 D_X^2 \right) \mathbb{E}[Q_t | \mathcal{F}_{t-1}].
\]

Proof

1. We recall that for any \( a, b, c \), we have \((a + b + c)^2 \leq 3(a^2 + b^2 + c^2)\). Thus
\[
(\hat{\theta}_t - \theta^*)^\top \nabla_t \nabla_t^\top (\hat{\theta}_t - \theta^*) = Q_t(y_t - \hat{\theta}_t^\top X_t)^2
\leq 3Q_t \left( (y_t - \mathbb{E}[y_t | X_t])^2 + (\mathbb{E}[y_t | X_t] - \theta^*^\top X_t)^2 + ((\theta^* - \hat{\theta}_t)^\top X_t)^2 \right)
\leq 3Q_t \left( (y_t - \mathbb{E}[y_t | X_t])^2 + D_{\text{app}}^2 + \|\hat{\theta}_t - \theta^*\|^2 D_X^2 \right).
\]  

To obtain the last inequality, we use the second point of Assumption 4 to bound the middle term. Then we use Taylor series for the exponential, and we apply Equation (6). For any \( t \) and any \( \mu \) satisfying \( 0 < \mu \leq \frac{1}{4\epsilon^2 D_X^2 \sigma^2} \), we have
\[
\mathbb{E} \left[ \exp \left( \mu Q_t(y_t - \mathbb{E}[y_t | X_t])^2 \right) \right] = 1 + \sum_{i \geq 1} \frac{\mu^i Q_t^i \mathbb{E}[(y_t - \mathbb{E}[y_t | X_t])^{2i}]}{i!}
\leq 1 + 2 \sum_{i \geq 1} \frac{\mu^i Q_t^i (2\sigma^2)^i}{i!}
\leq 1 + 2 \sum_{i \geq 1} (2\mu Q_t \sigma^2)^i
\leq 1 + 8\mu Q_t \sigma^2,
\]
\[
\leq \exp(8\mu Q_t \sigma^2).
\]

Therefore, for any \( t \),
\[
\mathbb{E} \left[ \exp \left( \frac{1}{4\epsilon^2 D_X^2 \sigma^2} Q_t \left( (y_t - \mathbb{E}[y_t | X_t])^2 - 8\sigma^2 \right) \right) \right] \leq 1.
\]

We define the random variable
\[
V_n \equiv \exp \left( \frac{1}{4\epsilon^2 D_X^2 \sigma^2} \sum_{t=k+1}^{k+n} Q_t \left( (y_t - \mathbb{E}[y_t | X_t])^2 - 8\sigma^2 \right) \right), \quad n \in \mathbb{N}.
\]
\((V_n)_n\) is adapted to the filtration \((\sigma(X_1, y_1, \ldots, X_{k+n}, y_{k+n}, X_{k+n+1})_n\), moreover \(V_0 = 1\) and \(V_n \geq 0\) almost surely, and

\[
\mathbb{E} \left[ V_n \mid X_1, y_1, \ldots, X_{k+n-1}, y_{k+n-1}, X_{k+n}, \|\hat{\theta}_{k+n} - \theta^*\| \leq \varepsilon \right] \leq V_{n-1}.
\]

Therefore we apply Lemma 17 for any \(\delta > 0\),

\[
P \left( \bigcup_{n=1}^{\infty} (V_n > \delta^{-1}) \cap A_{\delta}^k \right) \leq \delta,
\]

which is equivalent to

\[
P \left( \bigcup_{n=1}^{\infty} \left( \sum_{t=k+1}^{k+n} Q_t (y_t - \mathbb{E}[y_t \mid X_t])^2 > 8\sigma^2 \sum_{t=k+1}^{k+n} Q_t + 4\varepsilon^2 D_X^2 \sigma^2 \ln \delta^{-1} \right) \cap A_{\delta}^k \right) \leq \delta.
\]

Substituting in Equation (7), we obtain the desired result.

2. We apply the same decomposition as for Equation (7): for any \(t\),

\[
(\hat{\theta}_t - \theta^*)^\top \mathbb{E}[\nabla_t \nabla_t^\top \mid F_{t-1}](\hat{\theta}_t - \theta^*)
\]

\[
\leq 3(\hat{\theta}_t - \theta^*)^\top \mathbb{E} \left[ X_t X_t^\top (y_t - \mathbb{E}[y_t \mid X_t])^2 + D_{\text{app}}^2 \|\theta^* - \hat{\theta}_t\|^2 D_X^2 \right] (\hat{\theta}_t - \theta^*)
\]

Assumption 4 implies that for any \(X_t\), \(\mathbb{E}[(y_t - \mathbb{E}[y_t \mid X_t])^2 \mid X_t] \leq \sigma^2\). Thus, the tower property yields

\[
(\hat{\theta}_t - \theta^*)^\top \mathbb{E}[\nabla_t \nabla_t^\top \mid F_{t-1}](\hat{\theta}_t - \theta^*)
\]

\[
\leq 3 \left( \sigma^2 + D_{\text{app}}^2 \|\theta^* - \hat{\theta}_t\|^2 D_X^2 \right)(\hat{\theta}_t - \theta^*)^\top \mathbb{E}[X_t X_t^\top \mid F_{t-1}](\hat{\theta}_t - \theta^*).
\]

Second, we bound the right-hand side of Lemma 5, that is the objective of the following lemma.

**Lemma 20** Let \(k \in \mathbb{N}\). For any \(\delta > 0\), we have

\[
P \left( \bigcup_{n=1}^{\infty} \left( \sum_{t=k+1}^{k+n} X_t^\top P_{t+1} X_t (y_t - \hat{\theta}_t^\top X_t)^2 > 12\lambda_{\text{max}}(P_1) D_X^2 \sigma^2 \ln \delta^{-1} \right. \right.
\]

\[
+ 3 \left( 8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2 \right) \left. d \ln \left( 1 + n \frac{\lambda_{\text{max}}(P_{k+1}) D_X^2}{d} \right) \right) \cap A_{\delta}^k \right) \leq \delta.
\]

**Proof** We apply a similar analysis as in the proof of Lemma 19 in order to use the sub-gaussian assumption, and then we apply the telescopic argument as in the bounded setting.
We decompose $y_t - \hat{\theta}^T_t X_t$:

$$X_t^TP_{t+1}X_t(y_t - \hat{\theta}^T_t X_t)^2 \leq 3X_t^TP_{t+1}X_t((y_t - \mathbb{E}[y_t \mid X_t])^2 + (\mathbb{E}[y_t \mid X_t] - \hat{y}'(\theta^*^T X_t))^2 + ((\theta^* - \hat{\theta})^T X_t)^2) \leq 3X_t^TP_{t+1}X_t \left((y_t - \mathbb{E}[y_t \mid X_t])^2 + D_{app}^2 + \|\hat{\theta}_t - \theta^*\|^2D^2_X \right). \tag{8}$$

To control $(y_t - \mathbb{E}[y_t \mid X_t])^2X_t^TP_{t+1}X_t$, we use its positivity along with Equation (6). Precisely, for any $t$ and any $\mu > 0$ satisfying $0 < \mu \leq \frac{1}{4X_t^TP_{t+1}X_t\sigma^2}$, we have

$$\mathbb{E}\left[\exp\left(\mu(y_t - \mathbb{E}[y_t \mid X_t])^2X_t^TP_{t+1}X_t \mid \mathcal{F}_{t-1}, X_t\right)\right] = 1 + \sum_{i \geq 1} \frac{\mu^i(X_t^TP_{t+1}X_t)^i}{i!} \mathbb{E}\left[(y_t - \mathbb{E}[y_t \mid X_t])^{2i} \mid X_t\right]$$

$$\leq 1 + 2 \sum_{i \geq 1} \frac{\mu^i(X_t^TP_{t+1}X_t)^i i!(2\sigma^2)^i}{i!} = 1 + 2 \sum_{i \geq 1} \left(2\mu X_t^TP_{t+1}X_t\sigma^2\right)^i$$

$$\leq 1 + 8\mu X_t^TP_{t+1}X_t\sigma^2, \quad 0 < 2\mu X_t^TP_{t+1}X_t\sigma^2 \leq \frac{1}{2} \leq \exp\left(8\mu X_t^TP_{t+1}X_t\sigma^2\right).$$

We apply the previous bound with a uniform $\mu = \frac{1}{4\lambda_{\max}(P_1)D^2_X\sigma^2}$. As $\lambda_{\max}(P_{t+1}) \leq \lambda_{\max}(P_1)$ for any $t$, we get $\mu \leq \frac{1}{4X_t^TP_{t+1}X_t\sigma^2}$. Thus, we define

$$V_n = \exp\left(\frac{1}{4\lambda_{\max}(P_1)D^2_X\sigma^2} \sum_{t=k+1}^{k+n} ((y_t - \mathbb{E}[y_t \mid X_t])^2 - 8\sigma^2) X_t^TP_{t+1}X_t\right), \quad n \in \mathbb{N}.$$

$(V_n)$ is a super-martingale adapted to the filtration $(\sigma(X_1, y_1, ..., X_{k+n-1}, y_{k+n-1}, X_{k+n}))_n$ satisfying almost surely $V_0 = 1, V_n \geq 0$, thus we apply Lemma 17:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} (V_n > \delta^{-1})\right) \leq \delta,$$

or equivalently

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} \left(\sum_{t=k+1}^{k+n} X_t^TP_{t+1}X_t(y_t - \mathbb{E}[y_t \mid X_t])^2 > 8\sigma^2 \sum_{t=k+1}^{k+n} X_t^TP_{t+1}X_t + 4\lambda_{\max}(P_1)D^2_X\sigma^2 \ln \delta^{-1}\right)\right) \leq \delta.$$
Combining it with Equation (8), we get

\[ \mathbb{P}\left( \bigcup_{n=1}^{\infty} \left( \sum_{t=k+1}^{k+n} X_t^T P_{t+1} X_t (y_t - \hat{\theta}_t^T X_t)^2 > 3 \left( 8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2 \right) \sum_{t=k+1}^{k+n} X_t^T P_{t+1} X_t + 12\lambda_{\text{max}}(P_{t+1}) D_X^2 \sigma^2 \ln \delta^{-1} \right) \cap A_k^\varepsilon \right) \leq \delta. \]

Then we apply Lemma 18: the second point of Assumption 3 holds with \( h_\varepsilon = 1 \), thus

\[ \sum_{t=k+1}^{k+n} \text{Tr} \left( P_{t+1}(P_{t+1}^{-1} - P_{t}^{-1}) \right) \leq d \ln \left( 1 + n \frac{\lambda_{\text{max}}(P_{t+1}) D_X^2}{d} \right), \quad n \geq 1. \]

We conclude with \( X_t^T P_{t+1} X_t = \text{Tr}(P_{t+1}(P_{t+1}^{-1} - P_{t}^{-1})) \).

We sum up our findings and we prove the result for the quadratic loss. The structure of the proof is the same as the one of Theorem 1.

**Proof of Theorem 2.** On the one hand, we sum Lemma 5 and Lemma 8: for any \( \lambda, \delta > 0 \)

\[ \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} \left( \mathbb{E}[\nabla_t \mid F_{t-1}]^\top (\hat{\theta}_t - \theta^*) - \frac{1}{2} Q_t - \lambda(\hat{\theta}_t - \theta^*)^\top \left( \nabla_t \nabla_t^\top + \frac{3}{2} \mathbb{E}[\nabla_t \nabla_t^\top \mid F_{t-1}] \right) (\hat{\theta}_t - \theta^*) \right) \]

\[ \leq \frac{1}{2} \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} X_t^T P_{t+1} X_t (y_t - \hat{\theta}_t^T X_t)^2 + \frac{||\hat{\theta}_{T(\varepsilon, \delta)+1} - \theta^*||^2}{\lambda_{\text{min}}(P_{T(\varepsilon, \delta)+1})} + \frac{\ln \delta^{-1}}{\lambda}, \quad n \geq 1, \quad (9) \]

with probability at least \( 1 - \delta \). On the other hand, we have

\[ \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} (L(\hat{\theta}_t) - L(\theta^*)) \leq \frac{1}{1 - 0.8} \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} \left( \mathbb{E}[\nabla_t \mid F_{t-1}]^\top (\hat{\theta}_t - \theta^*) - 0.8 \mathbb{E}[Q_t \mid F_{t-1}] \right). \quad (10) \]

We aim to relate Equations (9) and (10) as in the proof of Theorem 1. To that end, we apply Lemma 19

\[ \mathbb{P}\left( \bigcup_{n=1}^{\infty} \left( \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} \left( \frac{1}{2} Q_t + \lambda(\hat{\theta}_t - \theta^*)^\top \left( \nabla_t \nabla_t^\top + \frac{3}{2} \mathbb{E}[\nabla_t \nabla_t^\top \mid F_{t-1}] \right) (\hat{\theta}_t - \theta^*) \right) \right) > \left( \frac{1}{2} + 3\lambda(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) \right) \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} Q_t \]

\[ + \frac{9}{2} \lambda \left( \sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2 \right) \sum_{t=T(\varepsilon, \delta)+1}^{T(\varepsilon, \delta)+n} \mathbb{E}[Q_t \mid F_{t-1}] + 12\lambda \varepsilon^2 D_X^2 \sigma^2 \ln \delta^{-1} \right) \]

\[ \cap A_{T(\varepsilon, \delta)}^\varepsilon \right) \leq \delta. \]
As in the proof of Theorem 1 we apply Lemma A.3 of (Cesa-Bianchi and Lugosi, 2006) and Lemma 17: for any \( \delta > 0 \),

\[
\mathbb{P}\left( \bigcup_{n=1}^{\infty} \left( \sum_{t=T(\epsilon, \delta)+1}^{T(\epsilon, \delta)+n} Q_t > 10(e^{0.1} - 1) \sum_{t=T(\epsilon, \delta)+1}^{T(\epsilon, \delta)+n} \mathbb{E}[Q_t | \mathcal{F}_{t-1}] + 10\epsilon^2 D^2_X \ln \delta^{-1} \right) \cap A^T_{T(\epsilon, \delta)} \right) \leq \delta.
\]

We combine the last two inequalities:

\[
\mathbb{P}\left( \bigcup_{n=1}^{\infty} \left( \sum_{t=T(\epsilon, \delta)+1}^{T(\epsilon, \delta)+n} \left( \frac{1}{2} Q_t + \lambda (\hat{\theta}_t - \theta^*)^\top \left( \nabla_t \nabla_t^\top + \frac{3}{2} \mathbb{E}[\nabla_t \nabla_t^\top | \mathcal{F}_{t-1}] \right) (\hat{\theta}_t - \theta^*) \right) > \left( 10(e^{0.1} - 1) \left( \frac{1}{2} + 3\lambda(8\sigma^2 + D^2_{app} + \epsilon^2 D^2_X) \right) + \frac{9}{2} \lambda(\sigma^2 + D^2_{app} + \epsilon^2 D^2_X) \right) \right) \right.
\]

\[
\left. \sum_{t=T(\epsilon, \delta)+1}^{T(\epsilon, \delta)+n} \mathbb{E}[Q_t | \mathcal{F}_{t-1}] \right)
\]

\[
+ \left( 10\epsilon^2 D^2_X \left( \frac{1}{2} + 3\lambda(8\sigma^2 + D^2_{app} + \epsilon^2 D^2_X) \right) + 12\lambda \epsilon^2 D^2_X \sigma^2 \right) \ln \delta^{-1} \right) \cap A^T_{T(\epsilon, \delta)} \right) \leq 2\delta.
\]

We set

\[
\lambda = (0.8 - 5(e^{0.1} - 1)) \left( 30(e^{0.1} - 1)(8\sigma^2 + D^2_{app} + \epsilon^2 D^2_X) + \frac{9}{2}(\sigma^2 + D^2_{app} + \epsilon^2 D^2_X) \right)^{-1}
\]

in order to obtain

\[
10(e^{0.1} - 1) \left( \frac{1}{2} + 3\lambda(8\sigma^2 + D^2_{app} + \epsilon^2 D^2_X) \right) + \frac{9}{2} \lambda(\sigma^2 + D^2_{app} + \epsilon^2 D^2_X) = 0.8 \frac{1}{109\sigma^2 + 28D^2_{app} + 28\epsilon^2 D^2_X} < \lambda < \frac{1}{108\sigma^2 + 27D^2_{app} + 27\epsilon^2 D^2_X},
\]

\[
10\epsilon^2 D^2_X \left( \frac{1}{2} + 3\lambda(8\sigma^2 + D^2_{app} + \epsilon^2 D^2_X) \right) + 12\lambda \epsilon^2 D^2_X \sigma^2 \leq 8\epsilon^2 D^2_X
\]

\[
\frac{1}{\lambda} \leq 28(4\sigma^2 + D^2_{app} + \epsilon^2 D^2_X). \]
Combining Equations (9), (10) and (11), we obtain
\[
\mathbb{P}\left( \bigcup_{n=1}^{\infty} \left( 0.2 \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} (L(\hat{\theta}_t) - L(\theta^*)) \right) > \frac{1}{2} \sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} X_t^\top P_{t+1} X_t (y_t - \hat{\theta}_t^\top X_t)^2 + \frac{\varepsilon^2}{\lambda_{\min}(P_{T(\varepsilon,\delta)+1})} \\
+ 28(4\sigma^2 + D_{\text{approx}}^2 + \varepsilon^2 D_X^2) \ln \delta^{-1} + 8\varepsilon^2 D_X^2 \ln \delta^{-1} \right) \leq 3\delta.
\]
Finally, we apply Lemma 20 with \( P_{T(\varepsilon,\delta)+1} \approx P_1 \) and we use Assumption 5: it holds simultaneously
\[
\sum_{t=T(\varepsilon,\delta)+1}^{T(\varepsilon,\delta)+n} L(\hat{\theta}_t) - L(\theta^*) \leq 5 \left( \frac{3}{2} (8\sigma^2 + D_{\text{approx}}^2 + \varepsilon^2 D_X^2) \ln \left( 1 + n \frac{\lambda_{\max}(P_1) D_X^2}{d} \right) \\
+ \lambda_{\max}(P_{T(\varepsilon,\delta)+1}) \varepsilon^2 + 28(4\sigma^2 + D_{\text{approx}}^2 + \varepsilon^2 D_X^2) \ln \delta^{-1} \\
+ 8\varepsilon^2 D_X^2 \ln \delta^{-1} + 6\lambda_{\max}(P_1) D_X^2 \sigma^2 \ln \delta^{-1} \right), \quad n \geq 1,
\]
with probability at least \( 1 - 5\delta \). To conclude, we write
\[
28(4\sigma^2 + D_{\text{approx}}^2 + \varepsilon^2 D_X^2) + 8\varepsilon^2 D_X^2 + 6\lambda_{\max}(P_1) D_X^2 \sigma^2 \\
\leq 28 \left( \sigma^2 \left( 4 + \frac{\lambda_{\max}(P_1) D_X^2}{4} \right) + D_{\text{approx}}^2 + 2\varepsilon^2 D_X^2 \right).
\]

Appendix C. Proofs of Section 4

C.1 Proof of Theorem 10

Proof of Theorem 10. We check Assumption 3 with \( \kappa_\varepsilon = e^{D_X(\|\theta^*\|+\varepsilon)} \), \( h_\varepsilon = \frac{1}{4} \) and \( \rho_\varepsilon = e^{-\varepsilon D_X} > 0.95 \). We can thus apply Theorem 1 with
\[
\lambda_{\max}(P_{T(\varepsilon,\delta)+1}) \leq \lambda_{\max}(P_1) + \frac{1}{4} \sum_{t=1}^{T(\varepsilon,\delta)} \|X_t\|^2,
\]
\[
\frac{5\kappa_\varepsilon}{2} < 3e^{D_X(\|\theta^*\|)}, \quad 30 \left( 2\kappa_\varepsilon + \frac{\varepsilon^2 D_X^2}{4} \right) < 64e^{D_X(\|\theta^*\|)}, \quad 5\varepsilon^2 D_X^2 \leq 1/75.
\]
We then control the first terms. To that end, we use a rough bound at any time \( t \geq 1 \):
\[
L(\hat{\theta}_t) - L(\theta^*) \leq \mathbb{E} \left[ \frac{y_X}{1 + e^{y\hat{\theta}_t^\top X}} | \hat{\theta}_t \right]^\top (\hat{\theta}_t - \theta^*) \\
\leq D_X \|\hat{\theta}_t - \theta^*\| \\
\leq D_X(\|\hat{\theta}_1 - \theta^*\| + (t - 1)\lambda_{\max}(P_1) D_X),
\]
because for any \( s \geq 1 \), we have \( P_s \preceq P_1 \) and therefore \( \| \hat{\theta}_{s+1} - \hat{\theta}_s \| \leq \lambda_{\max}(P_1)D_X \). Summing from 1 to \( T(\frac{1}{20D_X}, \delta) \) yields the result.

\[ \square \]

### C.2 Concentration of \( P_t \)

We prove a concentration result based on Tropp (2012), which will be used on the inverse of \( P_t \).

**Lemma 21** If Assumption 1 is satisfied, then for any \( 0 \leq \beta < 1 \) and \( t \geq 4^{1/(1-\beta)} \), it holds

\[
\mathbb{P}\left( \lambda_{\min}\left( \sum_{s=1}^{t-1} \frac{X_sX_s^\top}{s^\beta} \right) < \frac{\lambda_{\min}t^{1-\beta}}{4(1-\beta)} \right) \leq d \exp\left( -t^{1-\beta} \frac{\Lambda_{\min}^2}{10D_X^4} \right).
\]

**Proof** We wish to center the matrices \( X_sX_s^\top \) by subtracting their (common) expected value. We use that if \( A \) and \( B \) are symmetric, \( \lambda_{\min}(A - B) \leq \lambda_{\min}(A) - \lambda_{\min}(B) \). Indeed, denoting by \( v \) any eigenvector of \( A \) associated with its smallest eigenvalue,

\[
\lambda_{\min}(A - B) = \min_x \frac{x^\top (A - B)x}{\|x\|^2} \\
\leq \frac{v^\top (A - B)v}{\|v\|^2} \\
= \lambda_{\min}(A) - \frac{v^\top Bv}{\|v\|^2} \\
\leq \lambda_{\min}(A) - \min_x \frac{x^\top Bx}{\|x\|^2} \\
= \lambda_{\min}(A) - \lambda_{\min}(B).
\]

We obtain:

\[
\lambda_{\min}\left( \sum_{s=1}^{t-1} \frac{X_sX_s^\top}{s^\beta} - \sum_{s=1}^{t-1} \mathbb{E}\left[ \frac{X_sX_s^\top}{s^\beta} \right] \right) \leq \lambda_{\min}\left( \sum_{s=1}^{t-1} \frac{X_sX_s^\top}{s^\beta} \right) - \lambda_{\min}\left( \sum_{s=1}^{t-1} \mathbb{E}\left[ \frac{X_sX_s^\top}{s^\beta} \right] \right) \\
= \lambda_{\min}\left( \sum_{s=1}^{t-1} \frac{X_sX_s^\top}{s^\beta} \right) - \Lambda_{\min}\sum_{s=1}^{t-1} \frac{1}{s^\beta} \\
\leq \lambda_{\min}\left( \sum_{s=1}^{t-1} \frac{X_sX_s^\top}{s^\beta} \right) - \Lambda_{\min}\frac{t^{1-\beta} - 1}{1 - \beta}.
\]
Therefore, we obtain

\[
P\left( \lambda_{\min}\left( \frac{t^{-1} \sum_{s=1}^{t^{-1}} X_s X_s^\top}{s^\beta} \right) < \frac{\Lambda_{\min}(t^{1-\beta} - 2)}{2(1 - \beta)} \right)
\]

\[
\leq P\left( \lambda_{\min}\left( \frac{t^{-1} \sum_{s=1}^{t^{-1}} (X_s X_s^\top - \mathbb{E}\left[ \frac{X_s X_s^\top}{s^\beta} \right])}{s^\beta} \right) < \frac{\Lambda_{\min}(t^{1-\beta} - 2)}{2(1 - \beta)} - \Lambda_{\min}\frac{t^{1-\beta} - 1}{1 - \beta} \right)
\]

\[
= P\left( \lambda_{\max}\left( \frac{t^{-1} \sum_{s=1}^{t^{-1}} \left( \mathbb{E}\left[ \frac{X_s X_s^\top}{s^\beta} \right] - \frac{X_s X_s^\top}{s^\beta} \right) \right)}{s^\beta} \right) > \frac{\Lambda_{\min}t^{1-\beta}}{2(1 - \beta)} \right)
\]

We check the assumptions of Theorem 1.4 of Tropp (2012):

- Obviously \( \mathbb{E}\left[ \frac{X_s X_s^\top}{s^\beta} \right] - \frac{X_s X_s^\top}{s^\beta} \) is centered,

- \( \lambda_{\max}\left( \mathbb{E}\left[ \frac{X_s X_s^\top}{s^\beta} \right] - \frac{X_s X_s^\top}{s^\beta} \right) \leq \lambda_{\max}\left( \mathbb{E}\left[ \frac{X_s X_s^\top}{s^\beta} \right] \right) \leq D_X^2 \) almost surely.

As \( 0 \lesssim \mathbb{E}\left[ \left( \mathbb{E}\left[ \frac{X_s X_s^\top}{s^\beta} \right] - \frac{X_s X_s^\top}{s^\beta} \right)^2 \right] \lesssim \mathbb{E}\left[ \left( \frac{X_s X_s^\top}{s^\beta} \right)^2 \right] \lesssim \frac{D_X^4}{s^2} \mathbb{I} < \frac{D_X^4}{s^2} \mathbb{I} \), we get

\[
0 \lesssim \sum_{s=1}^{t^{-1}} \mathbb{E}\left[ \left( \mathbb{E}\left[ \frac{X_s X_s^\top}{s^\beta} \right] - \frac{X_s X_s^\top}{s^\beta} \right)^2 \right] \lesssim \left( \sum_{s=1}^{t^{-1}} \frac{D_X^4}{s^2} \right) \mathbb{I} \lesssim \left( D_X^4 \frac{t^{1-\beta}}{1 - \beta} \right) \mathbb{I}
\]

Therefore we can apply Theorem 1.4 of Tropp (2012):

\[
P\left( \lambda_{\max}\left( \sum_{s=1}^{t^{-1}} \left( \mathbb{E}\left[ \frac{X_s X_s^\top}{s^\beta} \right] - \frac{X_s X_s^\top}{s^\beta} \right) \right) > \frac{\Lambda_{\min}t^{1-\beta}}{2(1 - \beta)} \right)
\]

\[
\leq d \exp\left( -\frac{\Lambda_{\min}^2 t^{2(1-\beta)} / (8(1 - \beta)^2)}{D_X^4 t^{1-\beta} / (1 - \beta) + D_X^2 \Lambda_{\min} t^{1-\beta} / (6(1 - \beta))} \right)
\]

\[
= d \exp\left( -t^{1-\beta} \frac{\Lambda_{\min}^2}{8D_X^4} \frac{1 / (1 - \beta)^2}{\Lambda_{\min} / (6D_X^2 (1 - \beta))} \right)
\]

\[
= d \exp\left( -t^{1-\beta} \frac{\Lambda_{\min}^2}{8D_X^4} \left( 1 - \beta + \frac{\Lambda_{\min}(1 - \beta)}{6D_X^2} \right)^{-1} \right).
\]

Using \( \Lambda_{\min} / D_X^2 \leq 1 \) and \( \beta \geq 0 \), we obtain \( 8(1 - \beta + \frac{\Lambda_{\min}(1 - \beta)}{6D_X^2}) \leq 8(1 + 1/6) = 28/3 \leq 10 \), therefore

\[
P\left( \lambda_{\min}\left( \sum_{s=1}^{t^{-1}} \frac{X_s X_s^\top}{s^\beta} \right) < \frac{\Lambda_{\min}(t^{1-\beta} - 2)}{2(1 - \beta)} \right) \leq d \exp\left( -t^{1-\beta} \frac{\Lambda_{\min}^2}{10D_X^4} \right).
\]

The result follows from \( \frac{1}{2} t^{1-\beta} - 2 > 0 \) for \( t \geq 4^{1/(1-\beta)} \).

We can now do a union bound to obtain Proposition 11.
Proof of Proposition 11. We first move our problem to the setting of Lemma 21:

$$\lambda_{\text{max}}(P_t) = \lambda_{\text{min}} \left( P^{-1}_1 + \sum_{s=1}^{t-1} X_s X_s^\top \alpha_s \right)^{-1}$$

$$\leq \lambda_{\text{min}} \left( P^{-1}_1 + \sum_{s=1}^{t-1} \frac{X_s X_s^\top}{s^\beta} \right)^{-1},$$

because \( \alpha_s \geq 1/s^\beta \). Therefore, for \( t \geq 8 \geq 4^{1/(1-\beta)} \),

$$P \left( \lambda_{\text{max}}(P_t) > \frac{4}{\Lambda_{\text{min}} t^{1-\beta}} \right) \leq P \left( \lambda_{\text{min}} \left( P^{-1}_1 + \sum_{s=1}^{t-1} \frac{X_s X_s^\top}{s^\beta} \right)^{-1} > \frac{4}{\Lambda_{\text{min}} t^{1-\beta}} \right)$$

$$= P \left( \lambda_{\text{min}} \left( P^{-1}_1 + \sum_{s=1}^{t-1} \frac{X_s X_s^\top}{s^\beta} \right) < \frac{\Lambda_{\text{min}} t^{1-\beta}}{4} \right)$$

$$\leq P \left( \lambda_{\text{min}} \left( \sum_{s=1}^{t-1} \frac{X_s X_s^\top}{s^\beta} \right) < \frac{\Lambda_{\text{min}} t^{1-\beta}}{4} \right)$$

$$\leq d \exp \left( -t^{1-\beta} \frac{\Lambda_{\text{min}}^2}{10D_X^4} \right),$$

where we applied Lemma 21 to obtain the last line. We take a union bound to obtain, for any \( k \geq 7 \),

$$P \left( \exists t > k, \lambda_{\text{max}}(P_t) > \frac{4}{\Lambda_{\text{min}} t^{1-\beta}} \right) \leq \sum_{t>k} d \exp \left( -t^{1-\beta} \frac{\Lambda_{\text{min}}^2}{10D_X^4} \right)$$

$$\leq d \sum_{t>k} \exp \left( -t^{1-\beta} \frac{\Lambda_{\text{min}}^2}{10D_X^4} \right)$$

$$= d \sum_{m \geq 1} \exp \left( -m \frac{\Lambda_{\text{min}}^2}{10D_X^4} \right) \sum_{t>k} 1_{[t^{1-\beta}=m]}$$

We bound \( \sum_{t>k} 1_{[t]=m} \): for any \( m \)

$$[t^{1-\beta}] = m \implies m^{1/(1-\beta)} \leq t < (m + 1)^{1/(1-\beta)} ,$$

then using \( e^x \leq 1 + 2x \) for any \( 0 \leq x \leq 1 \), we have

$$(m + 1)^{1/(1-\beta)} = m^{1/(1-\beta)}(1 + 1/m)^{1/(1-\beta)}$$

$$= m^{1/(1-\beta)} \exp(\ln(1 + 1/m)/(1 - \beta))$$

$$\leq m^{1/(1-\beta)} \exp(1/(m(1 - \beta)))$$

$$\leq m^{1/(1-\beta)}(1 + 2/(m(1 - \beta))) ,$$

as long as \( m \geq 2 \geq 1/(1 - \beta) \). Therefore

$$(m + 1)^{1/(1-\beta)} - m^{1/(1-\beta)} + 1 \leq 2m^{1/(1-\beta)-1}/(1 - \beta) + 1 \leq 4m + 1 \leq 4(m + 1) ,$$

37
and that is true for \( m = 1 \) too. Hence

\[
\mathbb{P} \left( \exists t > k, \lambda_{\text{max}}(P_t) > \frac{4 \Lambda_{\text{min}}}{10D_X^4} \right) \leq 4d \sum_{m \geq \lceil k^{1-\beta} \rceil} (m + 1) \exp \left( -\frac{m \Lambda_{\text{min}}^2}{10D_X^4} \right)
\]

where the second line is obtained deriving both sides of

\[
\sum_{m \geq \lceil k^{1-\beta} \rceil} \exp \left( -\frac{m \Lambda_{\text{min}}^2}{10D_X^4} \right)
\]

respect to \( r \). Also, as \( 1 - e^{-x} \geq xe^{-x} \) for any \( x \in \mathbb{R} \), we get

\[
\mathbb{P} \left( \exists t > k, \lambda_{\text{max}}(P_t) > \frac{4 \Lambda_{\text{min}}}{10D_X^4} \right) 
\]

Furthermore, as \( xe^{-x} \leq e^{-1} \) for any \( x \geq 0 \), we get for any \( k \geq 7 \):

\[
\left( k^{1-\beta} + \frac{10D_X^4}{\Lambda_{\text{min}}^2} \exp \left( \frac{\Lambda_{\text{min}}^2}{10D_X^4} \right) \right) \exp \left( -k^{1-\beta} \frac{\Lambda_{\text{min}}^2}{20D_X^4} \right) 
\]

Combining the last two inequalities, we obtain

\[
\mathbb{P} \left( \exists t > k, \lambda_{\text{max}}(P_t) > \frac{4 \Lambda_{\text{min}}}{10D_X^4} \right) 
\]

and the result follows. The last line comes from \( \Lambda_{\text{min}} \leq D_X^2 \) and consequently

\[
800e^{-1} \exp \left( 2 \frac{\Lambda_{\text{min}}^2}{10D_X^4} + \frac{1}{2} \exp \left( \frac{\Lambda_{\text{min}}^2}{10D_X^4} \right) \right) \leq 800e^{-1+0.2+0.5e^{0.1}} \approx 624.7 \leq 625.
\]

The condition \( k \geq 7 \) is not necessary because

\[
\left( \frac{20D_X^4}{\Lambda_{\text{min}}^2} \ln \left( \frac{625dD_X^8}{\Lambda_{\text{min}}^4 \delta} \right) \right)^{1/(1-\beta)} \geq 20 \ln(625\delta^{-1}),
\]
and either $\delta \geq 1$ and the result is trivial, either $\delta < 1$ and $20 \ln(625\delta^{-1}) \geq 128$.

C.3 Convergence of the Truncated Algorithm

In order to prove Theorem 12, we state and prove an intermediate lemma.

**Lemma 22** Let $\theta \in \mathbb{R}^d$.

1. For any $\eta > 0$, we have
   
   $$L(\theta) - L(\theta^*) > \eta \implies \left\| \frac{\partial L}{\partial \theta} \right\|_{\theta} \geq D_\eta$$

   where $D_\eta = \frac{\Lambda_{\min} \sqrt{\eta}}{\sqrt{2DX} (1 + e^{DX(\|\theta^*\| + \sqrt{\eta D_X^2})})}$.

2. For any $\varepsilon > 0$, we have
   
   $$\|\theta - \theta^*\| > \varepsilon \implies L(\theta) - L(\theta^*) > \frac{\Lambda_{\min}}{4(1 + e^{DX(\|\theta^*\| + \varepsilon)})} \varepsilon^2.$$

**Proof** Both points derive from a second-order identity, turned in an upper-bound in the one case and in a lower-bound in the other. Using $\frac{\partial L}{\partial \theta}(\theta^*) = 0$, there exists $0 \leq \lambda \leq 1$ such that

$$L(\theta) = L(\theta^*) + \frac{1}{2} (\theta - \theta^*)^\top \mathbb{E} \left[ \frac{1}{(1 + e^{(\lambda \theta + (1-\lambda)\theta^*)^\top X})(1 + e^{-(\lambda \theta + (1-\lambda)\theta^*)^\top X})} XX^\top \right] (\theta - \theta^*).$$

1. We first have

$$L(\theta) - L(\theta^*) \leq \frac{D_X^2}{8} \|\theta - \theta^*\|^2.$$ 

Assume $L(\theta) - L(\theta^*) > \eta$. Then $\|\theta - \theta^*\| \geq \sqrt{8\eta / D_X^2}$. Also, using the Taylor expansion of $\theta^*$ around some $\theta_0 \in \mathbb{R}^d$, we get

$$L(\theta^*) \geq L(\theta_0) + \frac{\partial L}{\partial \theta}|_{\theta_0} (\theta^* - \theta_0) + \frac{1}{4(1 + e^{DX(\|\theta^*\| + \|\theta_0 - \theta^*\|)})} (\theta_0 - \theta^*)^\top \mathbb{E} [XX^\top] (\theta_0 - \theta^*),$$

and that yields

$$\left\| \frac{\partial L}{\partial \theta} \right\|_{\theta_0} (\theta_0 - \theta^*) \geq L(\theta_0) - L(\theta^*) + \frac{\Lambda_{\min}}{4(1 + e^{DX(\|\theta^*\| + \|\theta_0 - \theta^*\|)})} \|\theta_0 - \theta^*\|^2.$$ 

Therefore, as $L(\theta_0) - L(\theta^*) \geq 0$,

$$\left\| \frac{\partial L}{\partial \theta} \right\|_{\theta_0} \geq \frac{\Lambda_{\min}}{4(1 + e^{DX(\|\theta^*\| + \|\theta_0 - \theta^*\|)})} \|\theta_0 - \theta_{\text{true}}\|. $$
Finally, as $L$ is convex of minimum $\theta^*$,

$$\left\| \frac{\partial L}{\partial \theta} \right\| \geq \min_{\|\theta_0 - \theta^*\| = \sqrt{8\eta/D_X}} \left\| \frac{\partial L}{\partial \theta} \right\| \theta_0 \geq \frac{\Lambda_{\min}}{4(1 + e^{D_X(\|\theta^*\| + \sqrt{8\eta/D_X})})} \sqrt{8\eta/D_X}
\geq \frac{\Lambda_{\min}}{\sqrt{2D_X(1 + e^{D_X(\|\theta^*\| + \sqrt{8\eta/D_X})})}} \sqrt{\eta}.$$  

2. On the other hand we have

$$L(\theta) \geq L(\theta^*) + \frac{\Lambda_{\min}}{4(1 + e^{D_X(\|\theta^*\| + \|\theta - \theta^*\|)})} \|\theta - \theta^*\|^2.$$  

Thus, as $L$ is convex of minimum $\theta^*$, if $\|\theta - \theta^*\| > \varepsilon$ it holds

$$L(\theta) - L(\theta^*) > \min_{\|\theta_0 - \theta^*\| = \varepsilon} L(\theta_0) - L(\theta^*) \geq \frac{\Lambda_{\min}}{4(1 + e^{D_X(\|\theta^*\| + \varepsilon)})} \varepsilon^2.$$  

\[\square\]

**Proof of Theorem 12.** We prove the convergence of $(L(\hat{\theta}_t))_t$ to $L(\theta^*)$ and then the convergence of $(\hat{\theta}_t)_t$ to $\theta^*$ follows. The convergence of $(L(\hat{\theta}_t))_t$ comes from the first point of Lemma 22. The link between the two convergences is stated in the second point.

To study the evolution of $L(\hat{\theta}_t)$ we first apply a second-order Taylor expansion: for any $t \geq 1$ there exists $0 \leq \alpha_t \leq 1$ such that

$$L(\hat{\theta}_{t+1}) = L(\hat{\theta}_t) + \left. \frac{\partial L}{\partial \theta} \right|_{\hat{\theta}_t}^\top (\hat{\theta}_{t+1} - \hat{\theta}_t) + \frac{1}{2} (\hat{\theta}_{t+1} - \hat{\theta}_t)^\top \left. \frac{\partial^2 L}{\partial \theta^2} \right|_{\hat{\theta}_t + \alpha_t(\hat{\theta}_{t+1} - \hat{\theta}_t)} (\hat{\theta}_{t+1} - \hat{\theta}_t). \quad (12)$$

We have $\frac{\partial^2 L}{\partial \theta^2} \preceq \frac{1}{4} E[XX^\top]$, therefore, using the update formula on $\hat{\theta}$, the second-order term is bounded with

$$\left. \frac{\partial^2 L}{\partial \theta^2} \right|_{\hat{\theta}_t + \alpha_t(\hat{\theta}_{t+1} - \hat{\theta}_t)} (\hat{\theta}_{t+1} - \hat{\theta}_t) \leq \frac{1}{(1 + e^{y_t^\top \hat{\theta}_t} x_t)^2} X_t^\top P_{t+1} X_t \frac{E[XX^\top]}{4} P_{t+1} X_t
\leq \frac{1}{4} D_X^2 \lambda_{\max}(P_{t+1})^2 \leq \frac{1}{4} D_X^4 \lambda_{\max}(P_t)^2.$$  

The first-order term is controlled using the definition of the algorithm:

$$\hat{\theta}_{t+1} - \hat{\theta}_t = \left(P_t - \frac{P_t X_t X_t^\top P_t}{1 + X_t^\top P_t X_t \alpha_t} \alpha_t \right) \frac{y_t X_t}{1 + e^{y_t^\top \hat{\theta}_t} x_t},$$  

and as $\alpha_t \leq 1$,

$$\left\| -\alpha_t \frac{P_t X_t X_t^\top P_t}{1 + X_t^\top P_t X_t \alpha_t} \frac{y_t X_t}{1 + e^{y_t^\top \hat{\theta}_t} x_t} \right\| \leq D_X^2 \lambda_{\max}(P_t)^2.$$
Also, \( \| \frac{\partial L}{\partial \theta} \| \leq D_X \). Substituting our findings in Equation (12), we obtain
\[
L(\hat{\theta}_{t+1}) \leq L(\hat{\theta}_t) + \frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_t} P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t} X_t} + 2D_X^2 \lambda_{\max}(P_t)^2.
\] (13)

We define
\[
M_t = \frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_t} P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t} X_t} - \mathbb{E} \left[ \frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_t} P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t} X_t} \bigg| X_1, y_1, \ldots, X_{t-1}, y_{t-1} \right] \]
\[
= \frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_t} P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t} X_t} + \frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_t} P_t \frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_t}.
\]

Hence we have
\[
\frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_t} P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t} X_t} \leq M_t - \lambda_{\min}(P_t) \left\| \frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_t} \right\|^2 \leq M_t - \frac{1}{tD_X^2} \left\| \frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_t} \right\|^2,
\]
because \( P_s \succ \frac{1}{sD_X^2} \). Combining it with Equation (13) and summing consecutive terms, we obtain, for any \( k < t \),
\[
L(\hat{\theta}_t) - L(\hat{\theta}_k) \leq \sum_{s=k}^{t-1} \left( M_s - \frac{1}{sD_X^2} \left\| \frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_s}\right\|^2 + 2D_X^4 \lambda_{\max}(P_s)^2 \right) \]. (14)

We recall that there exists \( C_\delta \) such that \( \mathbb{P}(A_{C_\delta}) \geq 1 - \delta \) where
\[
A_{C_\delta} := \bigcap_{t=1}^{\infty} \left( \lambda_{\max}(P_t) \leq \frac{C_\delta}{t^{1-\beta}} \right).
\]

On the previous inequality, we see that the right-hand side is the sum of a martingale and a term which is negative for \( s \) large enough, under the event \( A_{C_\delta} \).

We are then interested in \( \mathbb{P}((L(\hat{\theta}_t) - L(\theta^*) > \eta) \mid A_{C_\delta}) \) for some \( \eta \) > 0. For \( 0 \leq k \leq t \), we define \( B_{k,t} \) the event \( (\forall k < s < t, L(\hat{\theta}_s) - L(\theta^*) > \eta/2) \). Then we use the law of total probability:
\[
\mathbb{P}(L(\hat{\theta}_t) - L(\theta^*) > \eta \mid A_{C_\delta}) \leq \mathbb{P} \left( (L(\hat{\theta}_t) - L(\theta^*) > \eta) \cap B_{0,t} \mid A_{C_\delta} \right)
\]
\[
+ \sum_{k=1}^{t-1} \mathbb{P} \left( (L(\hat{\theta}_k) - L(\theta^*) > \eta) \cap (L(\hat{\theta}_k) - L(\theta^*) \leq \frac{\eta}{2}) \cap B_{k,t} \mid A_{C_\delta} \right) \] (16)
\[
\leq \mathbb{P} \left( (L(\hat{\theta}_t) - L(\theta^*) > \eta) \cap B_{0,t} \mid A_{C_\delta} \right)
\]
\[
+ \sum_{k=1}^{t-1} \mathbb{P} \left( (L(\hat{\theta}_k) - L(\hat{\theta}_k) > \frac{\eta}{2}) \cap B_{k,t} \mid A_{C_\delta} \right).
\]

Lemma 22 yields
\[
L(\hat{\theta}_s) - L(\theta^*) > \frac{\eta}{2} \implies \left\| \frac{\partial L}{\partial \theta} \bigg|_{\hat{\theta}_s} \right\| \geq D_\eta.
\]
We combine the last equation, along with Equation (14) and the definition of $A_{C_s}$ to get, for any $1 \leq k < t$,

$$
\mathbb{P}
\left(
(L(\hat{\theta}_t) - L(\hat{\theta}_k) > \eta/2) \cap B_{k,t} \mid A_{C_s}
\right)
\leq
\mathbb{P}
\left(
\sum_{s=k}^{t-1} M_s > f(k,t) \cap B_{k,t} \mid A_{C_s}
\right)
\leq
\mathbb{P}
\left(
\sum_{s=k}^{t-1} M_s > f(k,t) \mid A_{C_s}
\right),
$$

where $f(k,t) = \frac{\eta}{2} + 2D_X^2 \sum_{s=k}^{t-1} \frac{1}{s} - 2D_X^4 C_\delta^2 \sum_{s=k}^{t-1} \frac{1}{s^{(1-\beta)}},$ for any $1 \leq k < t$.

Similarly, we get

$$
\mathbb{P}
\left(
(L(\hat{\theta}_t) - L(\theta^*) > \eta) \cap B_{0,t} \mid A_C
\right)
\leq
\mathbb{P}
\left(
\sum_{s=1}^{t-1} M_s > f_0(t) \mid A_C
\right),
$$

with $f_0(t) = \eta - (L(\hat{\theta}_1) - L(\theta^*)) + 2D_X^2 \sum_{s=1}^{t-1} \frac{1}{s} - 2D_X^4 C_\delta^2 \sum_{s=1}^{t-1} \frac{1}{s^{(1-\beta)}},$ for any $t \geq 1$.

We have $\mathbb{E}[M_s \mid X_1, y_1, ..., X_{s-1}, y_{s-1}] = 0$, and almost surely $|M_s| \leq 2D_X^2 \lambda_{\text{max}}(P_s)$. We can therefore apply Azuma-Hoeffding inequality: for $t, k$ such that $f(k,t) > 0$,

$$
\mathbb{P}
\left(
\sum_{s=k}^{t-1} M_s > f(k,t) \mid A_{C_s}
\right)
\leq
\exp\left(-f(k,t)^2 \frac{(1-2\beta)\max\left(1/2, (k-1)^{1-2\beta}\right)}{8D_X^4 C_\delta^2}\right),
$$

because $\sum_{s=k}^{\infty} \frac{1}{s^{2(1-\beta)}} \leq \frac{1}{(1-2\beta)\max(1/2, (k-1)^{1-2\beta})}.$ Similarly, for $t$ such that $f_0(t) > 0$,

$$
\mathbb{P}
\left(
\sum_{s=1}^{t-1} M_s > f_0(t) \mid A_{C_s}
\right)
\leq
\exp\left(-f_0(t)^2 \frac{1}{16D_X^4 C_\delta^2}\right).
$$

We need to control $f(k,t), f_0(t)$. We see that for $t$ large enough, when $k$ is small compared to $t$, $f(k,t)$ is driven by $\frac{D_X^2}{4} \ln(t)$ and when $k \approx t$, $f(k,t)$ is driven by $\eta/2$. The following Lemma formally states these approximations as lower-bounds. We prove it right after the end of this proof.

**Lemma 23** For $t \geq \max\left(\frac{16D_X^5 C_\delta^2}{e^{D_X^2(1-2\beta)}}, \left(1 + \frac{8D_X^4 C_\delta^2}{\eta(1-2\beta)}\right)^{1-2\beta}\right)$, it holds

$$
f(k,t) \geq \frac{D_X^2}{4} \ln(t), \quad 1 \leq k < \sqrt{t},
$$

$$
f(k,t) \geq \frac{\eta}{4}, \quad \sqrt{t} \leq k < t.
$$

Similarly, for $t \geq e^{\frac{D_X^2}{8\eta}} \left(L(\hat{\theta}_1) - L(\theta^*) + \frac{4D_X^4 C_\delta^2}{(1-2\beta)}\right)$, we have

$$
f_0(t) \geq \frac{D_X^2}{2} \ln(t).
$$
Then, defining \( C_1 = \frac{D_\delta^2(1-2\beta)}{256D_XC_\delta^3} \) and \( C_2 = \frac{\eta^2(1-2\beta)}{128D_XC_\delta^2} \), we finally get for \( t \) large enough:

\[
\begin{align*}
\mathbb{P}\left( (L(\hat{\theta}_k) - L(\theta^*)) > \eta \cap B_{0,t} \mid A_{C_\delta} \right) &\leq \exp\left( -4C_1 \ln(t)^2 \right), \\
\mathbb{P}\left( (L(\hat{\theta}_k) - L(\theta^*)) > \eta \cap (L(\hat{\theta}_k) - L(\theta^*)) \leq \frac{\eta}{2} \cap B_{k,t} \mid A_{C_\delta} \right) &\leq \begin{cases} 
\exp\left( -C_1 \ln(t)^2 \right) & \text{if } 1 \leq k < \sqrt{t}, \\
\exp\left( -C_2(k-1)^{1-2\beta} \right) & \text{if } \sqrt{t} \leq k < t.
\end{cases}
\end{align*}
\]

Substituting in Equation (16) yields:

\[
\mathbb{P}(L(\hat{\theta}_t) - L(\theta^*) > \eta \mid A_C) \leq \exp\left( -4C_1 \ln(t)^2 \right) + \sum_{k=1}^{\lfloor \sqrt{t} \rfloor - 1} \exp\left( -C_1 \ln(t)^2 \right) + \sum_{k=\lfloor \sqrt{t} \rfloor}^{t-1} \exp\left( -C_2(k-1)^{1-2\beta} \right) \leq (\sqrt{t} + 1) \exp\left( -C_1 \ln(t)^2 \right) + t \exp\left( -C_2(\sqrt{t} - 1)^{1-2\beta} \right).
\]

Finally, Point 2 of Lemma 22 allows to obtain the result: defining \( \eta = \frac{\lambda_{\min} \epsilon^2}{4(1+e^{D_X(\|\theta^*\|+\epsilon)})} \), we obtain

\[
\mathbb{P}(\|\hat{\theta}_t - \theta^*\| > \epsilon \mid A_{C_\delta}) \leq \mathbb{P}(L(\hat{\theta}_t) - L(\theta^*) > \eta \mid A_{C_\delta}) \leq (\sqrt{t} + 1) \exp\left( -C_1 \ln(t)^2 \right) + t \exp\left( -C_2(\sqrt{t} - 1)^{1-2\beta} \right).
\]

In order to obtain the constants involved in the Theorem, we write

\[
D_{\eta} = \frac{\lambda_{\min} \sqrt{3(1+e^{D_X(\|\theta^*\|+\epsilon)})}}{2D_X(1 + \exp\left( D_X(\|\theta^*\| + \sqrt{\frac{\lambda_{\min} \epsilon^2}{D_X(1+e^{D_X(\|\theta^*\|+\epsilon)})}}) \right))} \geq \left( \frac{\lambda_{\min}}{1 + e^{D_X(\|\theta^*\|+\epsilon)}} \right)^{3/2} \frac{\epsilon}{4D_X},
\]

\[
C_1 \geq \frac{\lambda_{\min}^2(1-2\beta)\epsilon^4}{216D_X^2C_\delta^3(1 + e^{D_X(\|\theta^*\|+\epsilon)})^6},
\]

\[
C_2 \geq \frac{\lambda_{\min}^2(1-2\beta)\epsilon^4}{211D_X^2C_\delta^2(1 + e^{D_X(\|\theta^*\|+\epsilon)})^2},
\]

and the conditions of Lemma 23 become

\[
t \geq \exp\left( \frac{2^8D_X^4C_\delta^2(1 + e^{D_X(\|\theta^*\|+\epsilon)})^3}{\lambda_{\min}^3(1-2\beta)\epsilon^2} \right),
\]

\[
t \geq \left( 1 + \frac{32D_X^4C_\delta^2(1 + e^{D_X(\|\theta^*\|+\epsilon)})^{\frac{1}{1-2\beta}}}{(1-2\beta)\lambda_{\min} \epsilon^2} \right)^2,
\]

\[
t \geq \exp\left( \frac{32D_X^4(1 + e^{D_X(\|\theta^*\|+\epsilon)})^3}{\lambda_{\min}^3 \epsilon^2} \left( L(\hat{\theta}_1) - L(\theta^*) + \frac{4D_X^4C_\delta^2}{1 - 2\beta} \right) \right).
\]
The third line is obtained with the inequality $\ln(1 + x) \leq \sqrt{x}$ for any $x > 0$. Obviously, as $0 < 1 - 2\beta < 1$, the first threshold on $t$ is bounded by:

$$
\exp\left(\frac{2^8 D_X^8 C_\delta^2 (1 + e^{D_X(\|\theta^*\| + \epsilon)})^3}{\Lambda_{\min}^3 (1 - 2\beta)^{3/2} \epsilon^2}\right) \leq \exp\left(\frac{2^8 D_X^8 C_\delta^2 (1 + e^{D_X(\|\theta^*\| + \epsilon)})^3}{\Lambda_{\min}^3 (1 - 2\beta)^{3/2} \epsilon^2}\right).
$$

To handle the third one, we use $D_X^2 C_\delta \geq \frac{4D_X^2}{\Lambda_{\min}} \geq 4$ and as $\hat{\theta}_1 = 0$ we obtain $L(\hat{\theta}_1) - L(\theta^*) \leq \ln 2 \leq \frac{4D_X^4 C_\delta^2}{1 - 2\beta}$, hence

$$
\exp\left(\frac{32 D_X^4 (1 + e^{D_X(\|\theta^*\| + \epsilon)})^3}{\Lambda_{\min}^3 \epsilon^2} \left(L(\hat{\theta}_1) - L(\theta^*) + \frac{4D_X^4 C_\delta^2}{1 - 2\beta}\right)\right)
\leq \exp\left(\frac{2^8 D_X^8 C_\delta^2 (1 + e^{D_X(\|\theta^*\| + \epsilon)})^3}{\Lambda_{\min}^3 (1 - 2\beta)^{3/2} \epsilon^2}\right).
$$

Proof of Lemma 23. We recall that for any $k \geq 1$,

$$
\sum_{s=k}^{t-1} \frac{1}{s} \geq \ln t - \ln k, \quad \sum_{s=k}^{t-1} \frac{1}{s^{(1 - \beta)}} \leq \frac{1}{1 - 2\beta} \max(1/2, (k - 1)^{1 - 2\beta}).
$$

Therefore:

$$
f(k, t) \geq \frac{\eta}{2} + \frac{D_\eta^2}{D_X^2} (\ln t - \ln k) - \frac{2D_X^4 C_\delta^2}{1 - 2\beta} \frac{1}{\max(1/2, (k - 1)^{1 - 2\beta})},
$$

$$
f_0(t) \geq \eta - (L(\hat{\theta}_1) - L(\theta^*)) + \frac{D_\eta^2}{D_X^2} \ln t - \frac{4D_X^4 C_\delta^2}{1 - 2\beta}.
$$
• For any $1 \leq k < \sqrt{t}$, it holds $\ln k \leq \frac{1}{2} \ln t$, and we have 
\[ f(k, t) \geq \frac{D^2_{\eta}}{2D^2_X} \ln(t) - \frac{4D^4_X C^2_{\delta}}{1 - 2\beta}. \]
Taking $t \geq e^{\frac{16D^2_X C^2_{\delta}}{I}}$ yields $f(k, t) \geq \frac{D^2_{\eta}}{4D^2_X} \ln(t)$.

• For $t \geq 2$ and any $k \geq \sqrt{t}$, we have 
\[ f(k, t) \geq \frac{\eta}{2} - \frac{2D^4_X C^2_{\delta}}{(1 - 2\beta)(k - 1)^{1 - 2\beta}} \geq \frac{\eta}{2} - \frac{2D^4_X C^2_{\delta}}{(1 - 2\beta)(\sqrt{t} - 1)^{1 - 2\beta}}. \]

Then if $t \geq \left(1 + \left(\frac{8D^4_X C^2_{\delta}}{\eta(1 - 2\beta)}\right)^\frac{1}{2\beta}\right)^2$, we get $f(k, t) \geq \frac{\eta}{4}$.

• Last point comes from $f_0(t) \geq \frac{D^2_{\eta}}{D^2_X} \ln t - (L(\hat{\theta}_1) - L(\theta^*)) - \frac{45D^4_X C^2_{\delta}}{1 - 2\beta}$.

\[\square\]

**Proof of Corollary 13.** We apply Theorem 12: for any $t \geq \exp\left(\frac{2^8 D^8 X e^{2^4 D^2_X (1 + e^D_X (||\theta^*|| + \epsilon))^3}}{A^3_{\min} (1 - 2\beta)^{3/2} \epsilon^2}\right)$,

\[\mathbb{P} (||\hat{\theta}_t - \theta^*|| > \varepsilon \mid A_{C^{\sqrt{t}/2}}) \leq (\sqrt{t} + 1) \exp\left(-C_1 \ln(t)^2\right) + t \exp\left(-C_2 (\sqrt{t} - 1)^{1 - 2\beta}\right),\]

where
\[C_1 = \frac{A^6_{\min} (1 - 2\beta) \varepsilon^4}{2^{10} D^{12} X C^{2^2} (1 + e^D_X (||\theta^*|| + \epsilon))^6}, \quad C_2 = \frac{A^2_{\min} (1 - 2\beta) \varepsilon^4}{2^{11} D^2_X C^{2^2} (1 + e^D_X (||\theta^*|| + \epsilon))^2}.\]

We use a union bound: for any $T \geq \exp\left(\frac{2^8 D^8 X e^{2^4 D^2_X (1 + e^D_X (||\theta^*|| + \epsilon))^3}}{A^3_{\min} (1 - 2\beta)^{3/2} \epsilon^2}\right)$,

\[\mathbb{P} \left(\bigcup_{t = T + 1}^{\infty} (||\hat{\theta}_t - \theta^*|| > \varepsilon \mid A_{C^{\sqrt{t}/2}})\right) \leq \sum_{t > T} (\sqrt{t} + 1) \exp\left(-C_1 \ln(t)^2\right) + t \exp\left(-C_2 (\sqrt{t} - 1)^{1 - 2\beta}\right),\]

• If $T \geq e^{\frac{3}{\varepsilon^2 C_1}}$, we have
\[\sum_{t > T} (\sqrt{t} + 1) \exp\left(-C_1 \ln(t)^2\right) \leq \sum_{t > T} (\sqrt{t} + 1) \frac{1}{t^{3/2}} \leq 2/T.\]
• For \( t \geq 4, 1 - 1/\sqrt{t} \geq 1/2 \), then for \( t \geq \left( \frac{12}{C_2(1 - 2\beta)} \right)^{4/(1-2\beta)} \),

\[
t^3 \exp \left( -C_2(\sqrt{t} - 1)^{1-2\beta} \right) \leq \exp \left( 3 \ln(t) - \frac{C_2 t(1-2\beta)}{2} \right) \leq \exp \left( \frac{12}{1 - 2\beta} \ln \left( \frac{12}{C_2(1 - 2\beta)} \right) - \frac{6}{1 - 2\beta} \left( \frac{12}{C_2(1 - 2\beta)} \right) \right) \leq 1,
\]

because for any \( x > 0 \), we have \( \ln x \leq x/2 \).

Thus for \( T \geq \left( \frac{12}{C_2(1 - 2\beta)} \right)^{4/(1-2\beta)} \)

\[
\sum_{t > T} t \exp \left( -C_2(\sqrt{t} - 1)^{1-2\beta} \right) \leq 1/T.
\]

Finally, for \( T \) satisfying the previous conditions as well as \( T \geq 6\delta^{-1} \), we obtain

\[
P \left( \bigcup_{t = T+1}^{\infty} (\|\hat{\theta}_t - \theta^*\| > \varepsilon) \bigg| A_{C_{\delta/2}} \right) \leq 3/T \leq \delta/2.
\]

We now compare the constants involved. As long as \( \varepsilon D_X \leq 1 \), we have

\[
\exp \left( \frac{2^8 D_X^8 C_{\delta/2}^2 (1 + e^{D_X(\|\theta^*\|+\epsilon)} \right)^3}{\Lambda_{\min}(1 - 2\beta)^{3/2} \varepsilon^2} \right) \leq \exp \left( \frac{3 \cdot 2^{15} D_X^{12} C_{\delta/2}^2 (1 + e^{D_X(\|\theta^*\|+\epsilon)} \right)^6}{\Lambda_{\min}(1 - 2\beta)^{3/2} \varepsilon^4} \right).
\]

Furthermore, as \( 1 - 2\beta \leq 1 \), we have

\[
\exp \left( \frac{3}{2C_1} \right) = \exp \left( \frac{3 \cdot 2^{15} D_X^{12} C_{\delta/2}^2 (1 + e^{D_X(\|\theta^*\|+\epsilon)} \right)^6}{\Lambda_{\min}(1 - 2\beta)^{3/2} \varepsilon^4} \right) \leq \exp \left( \frac{3 \cdot 2^{15} D_X^{12} C_{\delta/2}^2 (1 + e^{D_X(\|\theta^*\|+\epsilon)} \right)^6}{\Lambda_{\min}(1 - 2\beta)^{3/2} \varepsilon^4} \right).
\]
Finally,

\[
\left( \frac{12}{C_2(1-2\beta)} \right)^{4/(1-2\beta)} = \exp \left( \frac{4}{1-2\beta} \ln \frac{12}{C_2(1-2\beta)} \right)
\]

\[
= \exp \left( \frac{4}{1-2\beta} \ln \frac{12 \cdot 2^{11} D_X^4 C_2^2 \theta^* (1 + e^{D_X(\|\theta^*\| + \varepsilon)}^2)}{\Lambda_{\min}^2 (1-2\beta)^2 \varepsilon^4} \right)
\]

\[
= \exp \left( \frac{8}{1-2\beta} \ln \frac{12 \cdot 2^{11} D_X^4 C_2^2 \theta^* (1 + e^{D_X(\|\theta^*\| + \varepsilon)}^2)}{\Lambda_{\min}^2 (1-2\beta)^2 \varepsilon^4} \right)
\]

\[
\leq \exp \left( \frac{8}{1-2\beta} \sqrt{\frac{3 \cdot 2^{13} D_X^4 C_2^2 \theta^* (1 + e^{D_X(\|\theta^*\| + \varepsilon)}^2)}{\Lambda_{\min}^2 (1-2\beta)^3 \varepsilon^4}} \right)
\]

\[
= \exp \left( \frac{\sqrt{62^9 D_X^4 C_2^2 \theta^* (1 + e^{D_X(\|\theta^*\| + \varepsilon)}^2)}^{\frac{1}{2}}}{\Lambda_{\min} (1-2\beta)^{3/2} \varepsilon^2} \right)
\]

\[
\leq \exp \left( \frac{3 \cdot 2^{15} D_X^4 C_2^2 \theta^* (1 + e^{D_X(\|\theta^*\| + \varepsilon)}^6)}{\Lambda_{\min}^6 (1-2\beta)^{3/2} \varepsilon^4} \right).
\]

\]

Appendix D. Proofs of Section 5

Proof of Proposition 14. The first order condition of the optimum yields

\[
\arg \min_{\theta \in \mathbb{R}^d} \sum_{s=1}^{t-1} (y_s - \hat{\theta}^T X_s)^2 + \frac{1}{2} (\theta - \hat{\theta}_1)^T P_1^{-1} (\theta - \hat{\theta}_1) = \hat{\theta}_1 + P_t \sum_{s=1}^{t-1} (y_s - \hat{\theta}_1^T X_s) X_s.
\]

Therefore we prove recursively that \( \hat{\theta}_t - \hat{\theta}_1 = P_t \sum_{s=1}^{t-1} (y_s - \hat{\theta}_1^T X_s) X_s \). It is clearly true at \( t = 1 \). Assuming it is true for some \( t \geq 1 \), we use the update formula

\[
\hat{\theta}_{t+1} - \hat{\theta}_1 = (I - P_{t+1} X_t X_t^T) (\hat{\theta}_t - \hat{\theta}_1) + P_{t+1} y_{t+1} X_t - P_{t+1} X_t X_t^T \hat{\theta}_1
\]

\[
= (I - P_{t+1} X_t X_t^T) P_t \sum_{s=1}^{t-1} (y_s - \hat{\theta}_1^T X_s) X_s + P_{t+1} (y_t - \hat{\theta}_1^T X_t) X_t.
\]

We conclude with the following identity:

\[
(I - P_{t+1} X_t X_t^T) P_t = P_t - P_t X_t X_t^T P_t + \frac{P_t X_t X_t^T P_t X_t X_t^T P_t}{X_t^T P_t X_t + 1} = P_t - \frac{P_t X_t X_t^T P_t}{X_t^T P_t X_t + 1} = P_{t+1}.
\]

\]

47
D.1 Proof of Theorem 16

We first prove a result controlling the first estimates of the algorithm.

**Lemma 24** Provided that Assumptions 1, 2 and 4 are satisfied, starting from any \( \hat{\theta}_1 \in \mathbb{R}^d \) and \( P_1 \succ 0 \), for any \( \delta > 0 \), it holds simultaneously

\[
\| \hat{\theta}_t - \theta^* \| \leq \| \hat{\theta}_1 - \theta^* \| + \lambda_{\max}(P_1)D_X ((3\sigma + D_{\text{approx}})(t - 1) + 3\sigma \ln \delta^{-1}) , \quad t \geq 1,
\]

with probability at least \( 1 - \delta \).

**Proof** From Proposition 14, we obtain, for any \( t \geq 1 \),

\[
\hat{\theta}_t - \theta^* = P_t \sum_{s=1}^{t-1} (y_s - \hat{\theta}_1^\top X_s)X_s - P_t \left( P_1^{-1} + \sum_{s=1}^{t-1} X_sX_s^\top \right) (\theta^* - \hat{\theta}_1).
\]

Consequently,

\[
\hat{\theta}_t - \theta^* = P_t \sum_{s=1}^{t-1} (y_s - \theta^* X_s)X_s + P_t P_1^{-1} (\hat{\theta}_1 - \theta^*),
\]

and using \( P_t P_1^{-1} \ll I \), we obtain

\[
\| \hat{\theta}_t - \theta^* \| \leq \| \hat{\theta}_1 - \theta^* \| + \lambda_{\max}(P_1)D_X \sum_{s=1}^{t-1} |y_s - \theta^* X_s|.
\]

We apply Lemma 1.4 of Rigollet and Hüttter (2015) in the second line of the following: for any \( \mu \) such that \( 0 < \mu < \frac{1}{2\sqrt{2}\sigma} \),

\[
\mathbb{E} [\exp(\mu |y_t - \mathbb{E}[y_t | X_t]|)] = 1 + \sum_{i \geq 1} \mu^i \mathbb{E} [\mathbb{E}[|y_t - \mathbb{E}[y_t | X_t]|^i]]
\]

\[
\leq 1 + \sum_{k \geq 1} \mu^k (2\sigma^2)^{k/2} i \Gamma(i/2)
\]

\[
\leq 1 + \sum_{i \geq 1} \left( \sqrt{2} \mu \sigma \right)^i , \quad \text{because } \Gamma(i/2) \leq \Gamma(i) = (i - 1)!
\]

\[
\leq 1 + 2\sqrt{2} \mu \sigma , \quad \text{because } 0 < \sqrt{2} \mu \sigma \leq \frac{1}{2}
\]

\[
\leq \exp \left( 2\sqrt{2} \mu \sigma \right).
\]

Therefore \( \left( \exp \left( \frac{1}{2\sqrt{2} \sigma} \sum_{s=1}^{t} (|y_s - \mathbb{E}[y_s | X_s]| - 2\sqrt{2}\sigma) \right) \right)_t \) is a super-martingale to which we can apply Lemma 17. We obtain, for any \( \delta > 0 \),

\[
\sum_{s=1}^{t-1} |y_t - \mathbb{E}[y_t | X_t]| \leq 2\sqrt{2}(t - 1)\sigma + 2\sqrt{2}\sigma \ln \delta^{-1} , \quad t \geq 1,
\]

\[
48
\]
with probability at least $1 - \delta$. The result follows from Equation (17) and $2\sqrt{2} \leq 3$. ■

**Proof of Theorem 16.** We first apply Theorem 2: with probability at least $1 - 5\delta$, it holds simultaneously for all $n \geq T(\varepsilon, \delta)$

$$\sum_{t=T(\varepsilon, \delta)+1}^{n} L(\hat{\theta}_t) - L(\theta^*) \leq \frac{15}{2} d (8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) \ln \left(1 + (n - T(\varepsilon, \delta)) \frac{\lambda_{\max}(P_1) D_X^2}{d}\right)$$

$$+ 5\lambda_{\max} \left(P_{T(\varepsilon, \delta)+1}^{-1}\right) \varepsilon^2$$

$$+ 115 \left(\sigma^2 (4 + \frac{\lambda_{\max}(P_1) D_X^2}{4}) + D_{\text{app}}^2 + 2\varepsilon^2 D_X^2\right) \ln \delta^{-1}.$$

Moreover, $\lambda_{\max} \left(P_{T(\varepsilon, \delta)+1}^{-1}\right) \leq \lambda_{\max}(P_1^{-1}) + T(\varepsilon, \delta) D_X^2$.

Then we derive a bound on the first $T(\varepsilon, \delta)$ terms. For any $t \geq 1$, we have $L(\hat{\theta}_t) - L(\theta^*) \leq D_X^2 \|\hat{\theta}_t - \theta^*\|^2$, thus, using $(a + b)^2 \leq 2(a^2 + b^2)$ and applying Lemma 24 we obtain the simultaneous property

$$L(\hat{\theta}_t) - L(\theta^*) \leq 2D_X^2 \|\hat{\theta}_t - \theta^*\|^2 + 3\lambda_{\max}(P_1) D_X \sigma \ln \delta^{-1})^2$$

$$+ 2\lambda_{\max}(P_1)^2 D_X^4 (3 \sigma + D_{\text{app}})^2 (t - 1)^2, \quad t \geq 1,$n

with probability at least $1 - \delta$. A summation argument yields, for any $\delta > 0$,

$$\sum_{t=1}^{T(\varepsilon, \delta)} L(\hat{\theta}_t) - L(\theta^*) \leq 2D_X^2 \|\hat{\theta}_t - \theta^*\|^2 + 3\lambda_{\max}(P_1) D_X \sigma \ln \delta^{-1})^2 T(\varepsilon, \delta)$$

$$+ \lambda_{\max}(P_1)^2 D_X^4 (3 \sigma + D_{\text{app}})^2 (T(\varepsilon, \delta) - 1) T(\varepsilon, \delta) (2T(\varepsilon, \delta) - 1)/3,$n

with probability at least $1 - \delta$. ■

D.2 Definition of $T(\varepsilon, \delta)$

We now focus on the definition of $T(\varepsilon, \delta)$. We first transcript the result of Hsu et al. (2012) to our notations in the following lemma.

**Lemma 25** Provided that Assumptions 1, 2 and 4 are satisfied, starting from any $\hat{\theta}_1 \in \mathbb{R}^d$ and $P_1 = p_1 I$, $p_1 > 0$, we have, for any $0 < \delta < e^{-2}\delta$ and $t \geq 6\frac{D_X^2}{\sigma_{\min}} (\ln d + \ln \delta^{-1})$,

$$\|\hat{\theta}_{t+1} - \theta^*\|_2^2 \leq \frac{3}{t} \left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1} + \frac{D_X^2}{\sigma_{\min}} D_{\text{app}}^2 \frac{4(1 + \sqrt{8\ln \delta^{-1}})}{0.07^2} + \frac{3\sigma^2 (d/0.035 + \ln \delta^{-1})}{0.07}\right)$$

$$+ \frac{12}{0.07^2 t^2} \left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1} \frac{D_X^2}{\sigma_{\min}} (1 + \sqrt{8\ln \delta^{-1}})\right.$$ 

$$+ \left(\frac{D_X}{\sqrt{\sigma_{\min}}} (D_{\text{app}} + D_X \|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2p_1}}\right)^2 (\ln \delta^{-1})^2\left),

with probability at least $1 - 4\delta$. 49
We obtain, for any $0 < \delta < e^{−6}$ and $t \geq 6\frac{D}{\sqrt{\lambda min}}(\ln(d) + \ln \delta^{-1})$, it holds that $\|\hat{\beta}_{t+1,\lambda} - \beta\|_2^2 \leq 3(\|\beta_\lambda - \beta\|_2^2 + \varepsilon_{bs} + \varepsilon_{vr})$ with probability $1 - 4\delta$, with

$$
\varepsilon_{bs} \leq \frac{4}{0.07^2}\left(\frac{D^2}{\lambda_{min}^2} \mathbb{E}[(\mathbb{E}[y | X] - \beta^\top X)^2] + \left(1 + \frac{D^2}{\lambda_{min}^2}\right)\|\beta_\lambda - \beta\|_2^2\right) + \frac{\left(\frac{D^2}{\lambda_{min}^2} (D_{app} + D_X\|\beta - \hat{\beta}_1\|) + \|\beta_\lambda - \beta\|_2^2\right)^2}{t^2} - \left(\ln \delta^{-1}\right)^2\right),
$$

$$
\varepsilon_{vr} \leq \frac{2\sigma^2}{0.07^2 t} + 2\sigma^2 \frac{\ln \delta^{-1}}{0.07^3 / 2 t} + \frac{2\sigma^2 \ln \delta^{-1}}{0.07 t}.
$$

Moreover $\mathbb{E}[(\mathbb{E}[y | X] - \beta^\top X)^2] \leq D_{app}^2$ and $\lambda_{min} \leq D_X^2$; hence, using $\|\beta_\lambda - \beta\|_2 \leq \lambda\|\beta - \hat{\beta}_1\|$ we transfer the result in our KF notations, that is, $\theta_t = \hat{\beta}_{t,T^{-1}/2(t-1)}, \beta_1 = \hat{\theta}_1, \beta = \theta^*$. We obtain, for any $0 < \delta < e^{−6}$ and $t \geq 6\frac{D}{\sqrt{\lambda min}}(\ln(d) + \ln \delta^{-1})$,

$$
\varepsilon_{bs} \leq \frac{4}{0.07^2}\left(\frac{D^2}{\lambda_{min}^2} D_{app}^2 + \frac{D^2}{\lambda_{min}^2} \|\hat{\theta}_1 - \theta^*\|^2}{p_1 t} (1 + \sqrt{8\ln \delta^{-1}}) + \frac{\left(\frac{D^2}{\lambda_{min}^2} (D_{app} + D_X\|\theta^*\|) + \|\hat{\theta}_1 - \theta^*\|^2\right)^2}{t^2} - \left(\ln \delta^{-1}\right)^2\right),
$$

$$
\varepsilon_{vr} \leq \frac{\sigma^2 d(1 + \delta_f)}{0.07^2 t} + 2\sigma^2 \frac{\sqrt{d(1 + \delta_f)}}{0.07^3 / 2 t} \frac{\ln \delta^{-1}}{0.07 t} + \frac{2\sigma^2 \ln \delta^{-1}}{0.07 t},
$$

$$
\|\hat{\theta}_{t+1} - \theta^*\|_2^2 \leq 3\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{2 p_1 t} + \varepsilon_{bs} + \varepsilon_{vr}\right),
$$
Stochastic Online Optimization using Kalman Recursion

with probability at least $1 - 4\delta$. For $t \geq \frac{D_t^2}{\Lambda_{\min}} \ln \delta^{-1}$, as $\ln \delta^{-1} \geq 1$, we get

$$
\delta_t \leq \frac{1}{\sqrt{6} \ln \delta^{-1}} (1 + \sqrt{8 \ln \delta^{-1}}) + \frac{1}{6} \sqrt{\frac{1}{d} + 1} \leq 1 + \frac{\sqrt{8}}{\sqrt{6}} + \frac{2\sqrt{3}}{9} \approx 1.9 \leq 2.
$$

Thus, as $\sqrt{ab} \leq \frac{a+b}{2}$ for any $a, b > 0$, we have

$$
\epsilon_{vt} \leq \frac{\sigma^2}{0.07t} \left( \frac{3d}{0.07} + 2\sqrt{3d \ln \delta^{-1} \frac{0.07}{0.07} + 2\ln \delta^{-1}} \right)
$$

$$
\leq \frac{\sigma^2}{0.07t} \left( \frac{6d}{0.07} + 3\ln \delta^{-1} \right)
$$

$$
\leq 3\sigma^2(d/0.035 + \ln \delta^{-1})
.$$ 

It yields the result.

Lemma 25 allows the definition of an explicit value for $T(\epsilon, \delta)$, as displayed in the following Corollary.

**Corollary 26** Assumption 5 is satisfied for $T(\epsilon, \delta) = \max(T_1(\delta), T_2(\epsilon, \delta), T_3(\epsilon, \delta))$ where we define

$$
T_1(\delta) = \max \left( \frac{12 \frac{D_t^2}{\Lambda_{\min}} (\ln d + \ln \delta^{-1}) + \frac{48 \frac{D_t^2}{\Lambda_{\min}}}{\Lambda_{\min}} \ln \frac{24 \frac{D_t^2}{\Lambda_{\min}}}{\Lambda_{\min}}}{0.07} \right),
$$

$$
T_2(\epsilon, \delta) = \frac{24\epsilon^{-1}}{\Lambda_{\min}} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1} + \frac{\frac{D_t^2}{\Lambda_{\min}}}{\Lambda_{\min}} \frac{D_t^2}{\Lambda_{\min}} \frac{2(1 + \sqrt{8 \ln \delta^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035 + \ln \delta^{-1})}{0.07} \right)
$$

$$
\ln \frac{12\epsilon^{-1}}{\Lambda_{\min}} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1} + \frac{\frac{D_t^2}{\Lambda_{\min}}}{\Lambda_{\min}} \frac{D_t^2}{\Lambda_{\min}} \frac{2(1 + \sqrt{8 \ln \delta^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035 + \ln \delta^{-1})}{0.07} \right),
$$

$$
T_3(\epsilon, \delta) = \sqrt{\frac{96\epsilon^{-1}}{0.07^2\Lambda_{\min}}} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1} \frac{\frac{D_t^2}{\Lambda_{\min}}}{\Lambda_{\min}} (1 + \sqrt{8 \ln \delta^{-1}})
$$

$$
+ \left( \frac{D_t}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_t \|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2p_1}} \right)^2 (\ln \delta^{-1}) \right)^{1/2}
$$

$$
\ln \frac{96\epsilon^{-1}}{0.07^2\Lambda_{\min}} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1} (1 + \frac{\frac{D_t^2}{\Lambda_{\min}}}{\Lambda_{\min}}) (1 + \sqrt{8 \ln \delta^{-1}})
$$

$$
+ \left( \frac{D_t}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_t \|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2p_1}} \right)^2 (\ln \delta^{-1}) \right)^{1/2}.
$$

We recall that for any $\eta \leq 1$, we have $\frac{\ln t}{t} \leq \eta$ for $t \geq 2\eta^{-1} \ln(\eta^{-1})$, and we use it in the following proof.

**Proof of Corollary 26.** We define $\delta_t = \delta/t^2$ for any $t \geq 1$. In order to apply Lemma 25 with a union bound, we need $t \geq \frac{6 \frac{D_t^2}{\Lambda_{\min}} (\ln d + \ln \delta^{-1})}{\Lambda_{\min}}$. If $t \geq \frac{12 \frac{D_t^2}{\Lambda_{\min}} (\ln d + \ln \delta^{-1})}{\Lambda_{\min}}$ and
Lemma 25. We get the simultaneous property for all $t \geq \frac{4 D^2_{\hat{X}}}{\Lambda_{\min}} \ln \frac{2 D^2_{\hat{X}}}{\Lambda_{\min}}$, we obtain
\[
 t \geq \frac{t}{2} + \frac{\sqrt{t}}{2} \sqrt{t} \\
 \geq 6 \frac{D^2_{\hat{X}}}{\Lambda_{\min}} \ln d + \ln \delta^{-1} + \frac{12 D^2_{\hat{X}}}{\Lambda_{\min}} \ln t, \quad \text{as } \ln t \leq \sqrt{t} \\
 = 6 \frac{D^2_{\hat{X}}}{\Lambda_{\min}} (\ln d + \ln \delta^{-1}).
\]

Therefore, we define $T_1(\delta) = \max \left( 12 \frac{D^2_{\hat{X}}}{\Lambda_{\min}} (\ln d + \ln \delta^{-1}), \frac{4 D^2_{\hat{X}}}{\Lambda_{\min}} \ln \frac{24 D^2_{\hat{X}}}{\Lambda_{\min}} \right)$, and we apply Lemma 25. We get the simultaneous property
\[
\|\hat{\theta}_{t+1} - \theta^*\|^2 \leq \frac{3}{t} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{2 p_1} + \frac{D^2_{\hat{X}}}{\Lambda_{\min}} D^2_{\text{app}} \frac{4(1 + \sqrt{8 \ln \delta^{-1}})}{0.07^2} + \frac{3 \sigma^2 (d/0.035 + \ln \delta^{-1})}{0.07} \right) \\
+ \frac{12}{0.07^2 t^2} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1} \frac{D^2_{\hat{X}}}{\Lambda_{\min}} (1 + \sqrt{8 \ln \delta^{-1}}) \\
+ \left( \frac{D_X}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_X \|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2 p_1}} \right)^2 (\ln \delta^{-1})^2 \right)
\]
for all $t \geq T_1(\delta)$, with probability at least $1 - 4 \delta \sum_{t \geq T_1(\delta)} t^{-2} \geq 1 - \delta$ because $T_1(\delta) > 4$.

Thus, as $\ln t \geq 1$ for $t \geq T_1(\delta)$ and $\|\hat{\theta}_{t+1} - \theta^*\|^2 \geq \Lambda_{\min} \|\hat{\theta}_{t+1} - \theta^*\|^2$, we obtain
\[
\|\hat{\theta}_{t+1} - \theta^*\| \leq \frac{6 \ln t}{\Lambda_{\min} t} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{2 p_1} + \frac{D^2_{\hat{X}}}{\Lambda_{\min}} D^2_{\text{app}} \frac{4(1 + \sqrt{8 \ln \delta^{-1}})}{0.07^2} + \frac{3 \sigma^2 (d/0.035 + \ln \delta^{-1})}{0.07} \right) \\
+ \frac{48 (\ln t)^2}{0.07^2 \Lambda_{\min} t^2} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1} \frac{D^2_{\hat{X}}}{\Lambda_{\min}} (1 + \sqrt{8 \ln \delta^{-1}}) \\
+ \left( \frac{D_X}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_X \|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2 p_1}} \right)^2 (\ln \delta^{-1})^2 \right)
\]
for all $t \geq T_1(\delta)$, with probability at least $1 - \delta$. Finally, both terms of the last inequality are bounded by $\varepsilon/2$.

From Corollary 26, we obtain the asymptotic rate by comparing $T_2(\delta)$ and $T_3(\delta)$. We write $T_2(\delta) = 2 A_2(\delta) \ln A_2(\delta), T_3(\delta) = 2 A_3(\delta) \ln A_3(\delta)$ with
\[
A_2(\delta) \lesssim \frac{\varepsilon^{-1}}{\Lambda_{\min}} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1} + \frac{D^2_{\hat{X}}}{\Lambda_{\min}} D^2_{\text{app}} \sqrt{\ln \delta^{-1}} + \sigma^2 (d + \ln \delta^{-1}) \right) \\
A_3(\delta) \lesssim \left[ \frac{\varepsilon^{-1}}{\Lambda_{\min}} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1} \frac{D^2_{\hat{X}}}{\Lambda_{\min}} \sqrt{\ln \delta^{-1}} + \left( \frac{D_X (D_{\text{app}} + D_X \|\theta^*\|)}{\sqrt{\Lambda_{\min}}} + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{p_1}} \right)^2 (\ln \delta^{-1})^2 \right) \right].
\]
where the symbol \( \lesssim \) means less than up to universal constants. As \( \sqrt{a+b} \lesssim \sqrt{a} + \sqrt{b} \) and \( \sqrt{ab} \lesssim a + b \), we obtain

\[
A_3(\delta) \lesssim \sqrt{\frac{\varepsilon^{-1}}{\Lambda_{\min}}} \left( \sqrt{\frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1}} \frac{D_X^2}{\Lambda_{\min}} \sqrt{\ln \delta^{-1}} + \left( \frac{D_X}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_X \|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{p_1}} \right) \ln \delta^{-1} \right)
\]

\[
\lesssim \sqrt{\frac{\varepsilon^{-1}}{\Lambda_{\min}}} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1} + \frac{D_X^2}{\Lambda_{\min}} \sqrt{\ln \delta^{-1}} + \left( \frac{D_X}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_X \|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{p_1}} \right) \ln \delta^{-1} \right).
\]

Thus, as long as \( \frac{\varepsilon^{-1}}{\Lambda_{\min}} \leq 1 \), we get

\[
A_2(\delta), A_3(\delta) \lesssim \frac{\varepsilon^{-1}}{\Lambda_{\min}} \left( \frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1} + \frac{D_X^2}{\Lambda_{\min}} (1 + D_{\text{app}}^2) \sqrt{\ln \delta^{-1}} + \sigma^2 d \right.
\]

\[
+ \left( \frac{D_X}{\sqrt{\Lambda_{\min}}} (D_{\text{app}} + D_X \|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{p_1}} + \sigma^2 \right) \ln \delta^{-1} \bigg) .
\]

References


