Stochastic Online Optimization using Kalman Recursion

Joseph de Vilmarest, Olivier Wintenberger

STOCHASTIC ONLINE OPTIMIZATION USING KALMAN RECURSION

Joseph de Vilmarest [1,2]
joseph.de_vilmarest@upmc.fr

Olivier Wintenberger [1]
olivier.wintenberger@upmc.fr

June 23, 2020

ABSTRACT

We study the Extended Kalman Filter in constant dynamics, offering a bayesian perspective of stochastic optimization. We obtain high probability bounds on the cumulative excess risk in an unconstrained setting. 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. 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 EKF appears as a parameter-free online algorithm with $O(d^2)$ cost per iteration that optimally solves some unconstrained optimization problems.

***K*eywords** 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 $\ell(y_t, \theta^T X_t)$.

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 bound the regret (cumulative excess loss compared to the optimum). In the stochastic setting (X_t, y_t) is i.i.d. thus allowing to define the risk $L(\theta) = \mathbb{E}[\ell(y, \theta^T X)]$. The goal is to bound the excess risk. In this article we focus on the cumulative excess risk and we obtain non-asymptotic bounds holding with high probability. Our bounds hold simultaneously for any horizon, that is, we control the whole trajectory with high probability. Furthermore, our bounds on the cumulative risk all lead to a similar bound on the excess risk at any step for the averaged version of the algorithm.

Due to its low computational cost the Stochastic Gradient Descent 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 Stochastic Gradient Descent (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 $y_t \sim p_{\theta_t}(\cdot \mid X_t) \propto$

[1] Sorbonne Université, CNRS, LPSM, F-75005 Paris, France
[2] EDF R&D, Palaiseau, France

$\exp(-\ell(\cdot, \theta_t^T X_t))$ with $\theta_t \in \mathbb{R}^d$ and the state equation defines the dynamics of the state θ_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 Filter (Kalman and Bucy, 1961) for the quadratic loss and the Extended Kalman Filter (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 the analysis of BMA 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 (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 EKF has no gradient step size parameter nor projection step and thus does not need additional information on the optimal localization, yielding 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. 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. Kalman Filter in constant dynamics is exactly ridge regression with a varying regularization parameter (see Section 3.2), and similarly the static EKF may be seen as the online approximation of a regularized version of the well-studied Empirical Risk Minimizer (see for instance Ostrovskii and Bach (2018)).

1.1 Contributions

Our central contribution is a local analysis of the EKF under assumptions defined in Section 2, and provided that consecutive steps stay in a small ball around the optimum θ^*. We derive local bounds on the cumulative risk with high probability from a martingale analysis. Our analysis is similar to the one of Mahdavi et al. (2015) who obtained comparable results for the ONS, and we slightly refine their constants with an intermediate result (see Theorem 3). That is the aim of Section 3.

We then 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. However, in order to use our local result we first obtain the convergence of the algorithm to θ^*, and for that matter we need a good control of P_t. We therefore modify slightly the algorithm in the fashion of Bercu et al. (2020). This modification is limited in time and thus our local analysis still applies. In Section 5, we apply our analysis to the quadratic setting. We rely on Hsu et al. (2012) to obtain the convergence after exhibiting the correspondence between Kalman Filter in constant dynamics and Ridge Regression, and we therefore obtain similarly a global bound using our local analysis.

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 $\ell(y, \theta^T X) = -\log p_\theta(y \mid X)$ where $\theta \in \mathbb{R}^d$, $(X, y) \in \mathcal{X} \times \mathcal{Y}$ for some $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ and p_θ is of the form

$$p_\theta(y \mid X) = h(y) \exp\left(\frac{y\,\theta^T X - b(\theta^T X)}{a}\right), \tag{1}$$

where a is a constant and h and b are one-dimensional functions on which a few assumptions are required (Assumption 3). This includes linear and logistic regression, see Sections 4 and 5. We display the static EKF in Algorithm 1 in this setting.

Algorithm 1: 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 $t = 1, 2, \dots$
   (a) Update $P_{t+1} = P_t - \frac{P_t X_t X_t^T P_t}{1+X_t^T P_t X_t \alpha_t}\alpha_t$ with $\alpha_t = \frac{b''(\hat{\theta}_t^T X_t)}{a}$.
   (b) Update $\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1}\frac{(y_t - b'(\hat{\theta}_t^T X_t))X_t}{a}$.

Due to only matrix-vector and vector-vector multiplication, 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 (1), 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:

Assumption 1. *The observations $(X_t, y_t)_t$ are i.i.d. copies of the pair $(X, y) \in \mathcal{X} \times \mathcal{Y}$, $\mathbb{E}[XX^T]$ is positive definite and the diameter (for the Euclidian distance) of $\mathcal{X}$ is bounded by D_X.*

Working under Assumption 1, we define the risk function $L(\theta) = \mathbb{E}\left[\ell(y, \theta^T X)\right]$ and $\Lambda_{\min}$ the smallest eigenvalue of $\mathbb{E}[XX^T]$. 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 exists $(\kappa_\varepsilon)_{\varepsilon>0}, (h_\varepsilon)_{\varepsilon>0}$ and $\rho_\varepsilon \xrightarrow[\varepsilon\to 0]{} 1$ such that for any $\varepsilon > 0$ and any $\theta, \theta' \in \mathbb{R}^d$ satisfying $max(\|\theta - \theta^*\|, \|\theta' - \theta^*\|) \leq \varepsilon$, we have*

- $\ell'(y, \theta^T X)^2 \leq \kappa_\varepsilon \ell''(y, \theta^T X)$ *a.s.*
- $\ell''(y, \theta^T X) \leq h_\varepsilon$ *a.s.*
- $\ell''(y, \theta^T X) \geq \rho_\varepsilon \ell''(y, \theta'^T X)$ *a.s.*

Assumption 3 requires local exp-concavity (around θ^*) along with some regularity on ℓ'' (ℓ'' continuous and $\ell''(y, \theta^{*T} X) \geq \mu > 0$ a.s. is sufficient). That setting implies $\mathcal{Y}$ bounded, because ℓ' depends on y whereas ℓ'' doesn't. In logistic regression, $\mathcal{Y} = \{-1, +1\}$ 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, and in order to include the well-specified model, we assume y sub-gaussian conditionally to X, and not too far away from the model:

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

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

Both conditions of Assumption 4 hold with $\mathcal{Y} = \mathbb{R}$ and $D_{\text{app}} = 0$ for the well-specified sub-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:

Assumption 5. *For any $\delta, \varepsilon > 0$, there exists $\tau(\varepsilon, \delta) \in \mathbb{N}$ such that it holds for any $t > \tau(\varepsilon, \delta)$ simultaneously*

$$\|\hat{\theta}_t - \theta^*\| \leq \varepsilon \,,$$

with probability at least $1 - \delta$.

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 logistic and linear regression.

3.1 Main results

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 with larger bounds. Second, our analysis is based on a second-order expansion. The quadratic loss satisfies an identity with 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. 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 ε: on the one hand, the smaller ε the better our upper-bound, but on the other hand $\tau(\varepsilon, \delta)$ increases when ε decreases, and thus our bound applies after a bigger convergence time.

Theorem 1. *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*

$$\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} L(\hat{\theta}_t) - L(\theta^*) \leq \frac{5}{2} d \kappa_\varepsilon \ln\left(1 + n \frac{h_\varepsilon \lambda_{\max}(P_1) D_X^2}{d}\right) + 5\lambda_{\max}\left(P_{\tau(\varepsilon,\delta)+1}^{-1}\right)\varepsilon^2 + 30\left(2\kappa_\varepsilon + h_\varepsilon \varepsilon^2 D_X^2\right)\ln \delta^{-1} \,,$$

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

For the quadratic loss, we obtain the following theorem under the sub-gaussian hypothesis. We observe a similar trade-off in ε.

Theorem 2. *In the quadratic setting, 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*

$$\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} L(\hat{\theta}_t) - L(\theta^*) \leq \frac{15}{2} d \left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) \ln\left(1 + n \frac{\lambda_{\max}(P_1) D_X^2}{d}\right) + 5\lambda_{\max}\left(P_{\tau(\varepsilon,\delta)+1}^{-1}\right)\varepsilon^2 + 115\left(\sigma^2(4 + \frac{\lambda_{\max}(P_1) D_X^2}{4}) + D_{\text{app}}^2 + 2\varepsilon^2 D_X^2\right)\ln \delta^{-1} \,,$$

with probability at least $1 - 5\delta$.

We display the parallel between the ONS and the static EKF in Algorithm 2 through their recursive updates. Our analysis is similar to the one of Mahdavi et al. (2015) and an intermediate result yields the following refinement on their bound on the risk of the averaged ONS:

Theorem 3. *Let $(w_t)_t$ be the ONS estimates starting from $P_1 = \lambda I$ and using a step-size $\gamma = \frac{1}{2}\min(\frac{1}{4GD}, \alpha)$ with α the exp-concavity constant. Assume the gradients are bounded by G and the optimization set $\mathcal{K}$ has diameter D. Then for any $\delta > 0$, it holds for any $n \geq 1$ simultaneously*

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

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

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

Online Newton Step

- $P_{t+1}^{-1} = P_t^{-1} + \ell'(y_t, \hat{\theta}_t^T X_t)^2 X_t X_t^T$.
- $w_{t+1} = \prod_{\mathcal{K}}^{P_{t+1}^{-1}} \Big(w_t - \frac{1}{\gamma} P_{t+1} \nabla_t \Big)$.

Static Extended Kalman Filter

- $P_{t+1}^{-1} = P_t^{-1} + \ell''(y_t, \hat{\theta}_t^T X_t) X_t X_t^T$.
- $\hat{\theta}_{t+1} = \hat{\theta}_t - P_{t+1} \nabla_t$.

where $\nabla_t = \ell'(y_t, \hat{\theta}_t^T X_t) X_t$ and $\prod_{\mathcal{K}}^{P_{t+1}^{-1}}$ is the projection on $\mathcal{K}$ for the norm $\|.\|_{P_{t+1}^{-1}}$.

For consistency with the previous results we display Theorem 3 as a bound on the cumulative risk, whereas Theorem 3 of Mahdavi et al. (2015) is a bound on the risk of the averaged ONS. The latter follows directly from Theorem 3 by an application of Jensen's inequality. The proof of Theorem 3 consists in replacing Theorem 4 of Mahdavi et al. (2015) with the following lemma:

Lemma 4. *Let $k \geq 0$ and $(\Delta N_t)_{t>k}$ be any martingale difference adapted to the filtration $(\mathcal{F}_t)_{t \geq k}$ such that for any $t > k$, $\mathbb{E}[\Delta N_t^2 \mid \mathcal{F}_{t-1}] < \infty$. For any $\delta, \lambda > 0$, we have the simultaneous property*

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

with probability at least $1 - \delta$.

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

The comparison of Theorem 3 with Theorem 1 is difficult because we don't control in general $\tau(\varepsilon, \delta)$. We obtain similar constants, as κ_ε is the inverse of the exp-concavity constant α. However the static EKF is parameter-free whereas α is an input of the ONS through the setting of the step-size γ. That is why we argue that the static EKF provides an optimal way to choose the step size, as does averaged SGD (Bach, 2014). Indeed, as ε is a parameter of the EKF analysis, we can improve the leading constant κ_ε on local region arbitrarily small around θ^*, at a cost for the $\tau(\varepsilon, \delta)$ first terms, whereas in the ONS the choice of a diameter $D > \|\theta^*\|$ makes the gradient step-size sub-optimal and impact the leading constant. Similarly to the ONS analysis, the use of second-order methods learns adaptively the pre-conditioning matrix which is crucial in order to improve the leading constant $D_X^2 / \Lambda_{\min}$ obtained for first-order methods to d.

A similar comparison is possible between the result of Theorem 2 and tight risk bounds obtained for the ordinary least-squares estimator and the ridge regression estimator (Hsu et al., 2012). Up to numerical constants, the tight constant $d(\sigma^2 + D_{\text{app}}^2)$ is achieved by choosing ε arbitrarily small, at a cost for the $\tau(\varepsilon, \delta)$ first terms.

We detail the key ideas of the proofs through intermediate results in Sections 3.2 and 3.3. We defer to Appendix A the proof of these intermediate results along with the detailed proof of Theorems 1 and 2.

3.2 Comparison with Online Newton Step and Ridge Regression: a regret analysis

To begin our analysis, we formalize the strong links between the static EKF, the ONS and the Ridge Regression forecaster. For the quadratic loss, the EKF becomes the Kalman Filter by plugging in Algorithm 1 the identities $a = 1, b'(\hat{\theta}_t^T X_t) = \hat{\theta}_t^T X_t, \alpha_t = 1$.

The parallel with the Ridge Regression 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 empirical risk minimization problem.

Proposition 5. *In the quadratic setting, for any sequence (X_t, y_t) starting from any $\hat{\theta}_1 \in \mathbb{R}^d$ and $P_1 \succ 0$, the EKF satisfies the optimisation problem*

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

Notice that the static Kalman Filter provides automatically a right choice of the Ridge regularization parameter. This proposition is useful in the convergence phase of the quadratic setting.

In order to get a bound that holds sequentially for any $t \geq 1$, we adopt an approach similar as the one in Hazan et al. (2007) on the ONS (Algorithm 2). The cornerstone of our local analysis is the derivation of a bound on the second-order Taylor expansion of ℓ, from the recursive update formulae.

Lemma 6. *For any sequence $(X_t, y_t)_t$, starting from $P_1 \succ 0$ and $\hat{\theta}_1 \in \mathbb{R}^d$, it holds for any $\theta^* \in \mathbb{R}^d$ and $n \in \mathbb{N}$ that*

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

In the quadratic setting there is equality between the quadratic function and its second-order Taylor expansion and a logarithmic regret bound is derived (Cesa-Bianchi and Lugosi (2006), Theorem 11.7). However the factor before the logarithm is not easily bounded, unless we assume $(y_t - \hat{\theta}_t^T X_t)^2$ bounded.

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

$$\left(\ell'(y_t, w_t^T X_t) X_t \right)^T (\hat{\theta}_t - \theta^*) - \frac{\gamma}{2} (w_t - \theta^*)^T \left(\ell'(y_t, w_t^T X_t)^2 X_t X_t^T \right) (w_t - \theta^*) ,$$

where γ is a step size depending on the exp-concavity constant, a bound on the gradients and the diameter of the search region $\mathcal{K}$. Then the regret bound follows from the exp-concavity property, bounding the excess loss $\ell(y_t, w_t^T X_t) - \ell(y_t, \theta^{*T} X_t)$ with the previous quantity.

We follow a very different approach, to stay parameter-free 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.

3.3 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 7. *If Assumptions 1, 2 and 3 are satisfied, for any $\theta \in \mathbb{R}^d$, it holds*

$$\frac{\partial L}{\partial \theta}\Big|_\theta^T (\theta - \theta^*) \geq \rho_{\|\theta - \theta^*\|} (\theta - \theta^*)^T \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta (\theta - \theta^*) .$$

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

Proposition 8. *If Assumptions 1, 2 and 3 are satisfied, for any $\theta \in \mathbb{R}^d$, $0 < c < \rho_{\|\theta - \theta^*\|}$, it holds*

$$L(\theta) - L(\theta^*) \leq \frac{\rho_{\|\theta - \theta^*\|}}{\rho_{\|\theta - \theta^*\|} - c} \left(\frac{\partial L}{\partial \theta}\Big|_\theta^T (\theta - \theta^*) - c (\theta - \theta^*)^T \frac{\partial^2 L}{\partial \theta^2}\Big|_\theta (\theta - \theta^*) \right) .$$

Lemma 6 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. The natural condition becomes the third condition of Assumption 3.

Then we are left to obtain a bound on the cumulative risk from Lemma 6. In order to compare the derivatives of the risk and the losses, we need to control the martingale difference adapted to the natural filtration $(\mathcal{F}_t)$ and defined by

$$\Delta M_t = \left(\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t} - \nabla_t \right)^T (\hat{\theta}_t - \theta^*), \qquad \text{where } \nabla_t = \ell'(y_t, \hat{\theta}_t^T X_t) X_t . \tag{2}$$

We thus apply Lemma 4 to the martingale difference defined in Equation 2.

Lemma 9. *If Assumptions 1 and 2 are satisfied, for any $k \geq 0$ and $\delta, \lambda > 0$, it holds*

$$\sum_{t=k+1}^{k+n} \left(\Delta M_t - \lambda (\hat{\theta}_t - \theta^*)^T \left(\nabla_t \nabla_t^T + \frac{3}{2} \mathbb{E} \left[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1} \right] \right) (\hat{\theta}_t - \theta^*) \right) \leq \frac{\ln \delta^{-1}}{\lambda}, \qquad n \geq 1 ,$$

with probability at least $1 - \delta$.

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 $t = 1, 2, \dots$
   (a) Update $P_{t+1} = P_t - \frac{P_t X_t X_t^T P_t}{1+X_t^T P_t X_t \alpha_t}\alpha_t$, 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 $\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1}\frac{y_t X_t}{1+e^{y_t \hat{\theta}_t^T X_t}}$.

Summing Lemma 6 and 9, the rest of the proof consists in the following two steps:

- We derive poissonian bounds to control the quadratic terms in $\hat{\theta}_t - \theta^*$ in terms of the one of the second-order bound of Proposition 8.
- We upper-bound $\sum_t X_t^T P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^T X_t)^2$ relying on techniques similar to the ridge analysis of the proof of Theorem 11.7 of Cesa-Bianchi and Lugosi (2006).

4 Logistic setting

Logistic regression is a widely used statistical model in order to predict a binary random variable $y \in \mathcal{Y} = \{-1, 1\}$. It consists in estimating $\mathcal{L}(y \mid X)$ 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 The truncated algorithm

For checking Assumption 5, we follow a trick consisting in changing slightly the update on P_t (Bercu et al., 2020). 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 $\lambda_{\max}(P_t) = \mathcal{O}(1/t)$, one needs to lower-bound α_t, that is, lower-bound b'', and that is impossible in the global logistic setting. Therefore, the idea is to force a lower-bound on α_t in its definition. We thus define, for some $0 < \beta < 1/2$,

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

This modification yields Algorithm 3, where we keep the notations $\hat{\theta}_t, P_t$ with some abuse. We impose a decreasing threshold on α_t ($\beta > 0$) so that the recursion coincides with Algorithm 1 after some steps. Then we apply our analysis of Section 3 after slightly changing Assumption 5:

Assumption 6. *For any $\delta, \varepsilon > 0$, there exists $\tau(\varepsilon, \delta) \in \mathbb{N}$ such that it holds for any $t > \tau(\varepsilon, \delta)$*

$$\|\hat{\theta}_t - \theta^*\| \leq \varepsilon \text{ and } \alpha_t = \frac{1}{(1 + e^{\hat{\theta}_t^T X_t})(1 + e^{-\hat{\theta}_t^T X_t})}$$

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

The sensitivity of the algorithm to β is discussed at the end of Section 4.2. Also, note that the threshold could be c/t^β, $c > 0$, as in Bercu et al. (2020), it would not change the proofs nor the local result below.

We first state the result with $\tau(\varepsilon, \delta)$ in our upper-bound, for the choice $\varepsilon = 1/(20 D_X)$. We define its value in the next paragraph, and we discuss its dependence to parameters.

Theorem 10. *If Assumptions 1, 2 and 6 are satisfied, for any $\delta > 0$ it holds for any $n \geq 1$ simultaneously*

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

with probability at least $1 - 4\delta$.

4.2 Definition of $\tau(\varepsilon, \delta)$ in Assumption 6

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 check Assumption 6 with an explicit value of $\tau(\varepsilon, \delta)$ 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 based on Tropp (2012).

Proposition 11. *If Assumption 1 is satisfied, for any $\delta > 0$, it holds simultaneously*

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

with probability at least $1 - \delta$.

The limit $\beta < 1/2$ thus corresponds to the condition $\sum_t \lambda_{\max}(P_t)^2 < +\infty$ with high probability. Motivated by Proposition 11, we define, for $C > 0$, the event

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

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

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

and we obtain $\mathbb{P}(A_{C_\delta}) \geq 1 - \delta$. We obtain the following theorem under that condition.

Theorem 12. *Provided that Assumptions 1 and 2 are satisfied, if $\hat{\theta}_1 = 0$ we have for any $\varepsilon > 0$ and $t \geq \exp\left(\frac{2^8 D_X^8 C_\delta^2 (1+e^{D_X(\|\theta^*\|+\varepsilon)})^3}{\Lambda_{\min}^3(1-2\beta)^{3/2}\varepsilon^2}\right)$,*

$$\mathbb{P}(\|\hat{\theta}_t - \theta^*\| > \varepsilon \mid A_{C_\delta}) \leq (\sqrt{t}+1)\exp\left(-\frac{\Lambda_{\min}^6(1-2\beta)\varepsilon^4}{2^{16}D_X^{12}C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})^6}\ln(t)^2\right) + t\exp\left(-\frac{\Lambda_{\min}^2(1-2\beta)\varepsilon^4}{2^{11}D_X^4C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})^2}(\sqrt{t}-1)^{1-2\beta}\right).$$

The beginning of our convergence proof starts similarly as the analysis of Bercu et al. (2020): we obtain a recursive inequality ensuring that $(L(\hat{\theta}_t))_t$ 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 provided by 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. *Provided that Assumptions 1 and 2 are satisfied, if $\hat{\theta}_1 = 0$ we check Assumption 6 for any $\varepsilon > 0$, $\delta > 0$ and*

$$\tau(\varepsilon, \delta) = \max\left(\left(2(1 + e^{D_X(\|\theta^*\|+\varepsilon)})\right)^{1/\beta}, \exp\left(\frac{3 \cdot 2^{15} D_X^{12} C_{\delta/2}^2 (1 + e^{D_X(\|\theta^*\|+\varepsilon)})^6}{\Lambda_{\min}^6 (1-2\beta)^{3/2} \varepsilon^4}\right), 6\delta^{-1}\right).$$

This definition of $\tau(\varepsilon, \delta)$ allows a discussion on the dependence of the bound Theorem 10 to the different parameters:

- The truncation has introduced an extraparameter β, on which $\tau(\varepsilon, \delta)$ strongly depends with a trade-off. On the one hand, when β is close to 0, the algorithm is slow to coincide with the true Extended Kalman Filter, for which our fast rate holds. Precisely, we have $\tau(\varepsilon, \delta) = e^{O(1)/\beta}$. On the other hand, the truncation was introduced to control P_t. The larger β, the larger our control on $\lambda_{\max}(P_t)$ and thus we get $\tau(\varepsilon, \delta) = e^{O(1)/(1-2\beta)^{3/2}}$.
- As Corollary 13 holds for any $\varepsilon > 0$, the compromise realized with $\varepsilon = 1/(20 D_X)$, made for simplifying constants, is totally arbitrary. The dependence of the convergence time is of the order $\tau(\varepsilon, \delta) = e^{O(1)/\varepsilon^4}$. However the $\log n$ term of the bound has a $e^{D_X \varepsilon}$ factor. Thus the best compromise should be an $\varepsilon > 0$ decreasing with n.
- The dependence to δ is complex. The third constraint on $\tau(\varepsilon, \delta)$ is $O(\delta^{-1})$ which should not be sharp.

To improve this lousy dependence of 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 check Assumption 6 and not to get an optimal value of $\tau(\varepsilon, \delta)$. Moreover practical considerations show that the truncation is artificial and can even deteriorate the performence of the EKF, see Section 6. Thus Bercu et al. (2020) suggest a threshold as low as possible ($10^{-10}/t^{0.49}$) so that the truncation makes no difference in numerical experiments. A tight probability bound on $\lambda_{\max}(P_t)$ of the EKF is a very important and challenging open question.

5 Quadratic setting

We state our result for the quadratic loss where Algorithm 1 becomes the standard Kalman Filter. We first state our result with an upper-bound depending on $\tau(\varepsilon, \delta)$, then we define $\tau(\varepsilon, \delta)$ satisfying Assumption 5.

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 14. *Provided that Assumptions 1, 2, 4 and 5 are satisfied, for any $\varepsilon, \delta > 0$, it holds simultaneously*

$$\begin{aligned}\sum_{t=1}^{n} L(\hat{\theta}_t) - L(\theta^*) \leq{}& \frac{15}{2} d \left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) \ln\left(1 + n\frac{\lambda_{\max}(P_1) D_X^2}{d}\right) + 5\lambda_{\max}(P_1^{-1})\varepsilon^2 \\ &+ 115\left(\sigma^2(4 + \frac{\lambda_{\max}(P_1) D_X^2}{4}) + D_{\text{app}}^2 + 2\varepsilon^2 D_X^2\right) \ln \delta^{-1} \\ &+ D_X^2\left(5\varepsilon^2 + 2(\|\hat{\theta}_1 - \theta^*\|^2 + 3\lambda_{\max}(P_1) D_X \sigma \ln \delta^{-1})^2\right) \tau(\varepsilon, \delta) \\ &+ \frac{2\lambda_{\max}(P_1)^2 D_X^4 (3\sigma + D_{\text{app}})^2}{3} \tau(\varepsilon, \delta)^3, \qquad n \geq 1\,,\end{aligned}$$

with probability at least $1 - 6\delta$.

As Kalman Filter estimator is exactly the Ridge estimator for a varying regularization parameter, we can use the regularized empirical risk minimization properties to control $\tau(\varepsilon, \delta)$. In particular, we apply the ridge analysis provided by Hsu et al. (2012), and we check Assumption 5 by providing a non-asymptotic definition of $\tau(\varepsilon, \delta)$ in Appendix C, Corollary 24. Up to universal constants, we get

$$\begin{aligned}\tau(\varepsilon, \delta) \lesssim h\Bigg(\frac{\varepsilon^{-1}}{\Lambda_{\min}}\Bigg(&\frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1} + \frac{D_X^2}{\Lambda_{\min}}(1 + D_{\text{app}}^2)\sqrt{\ln \delta^{-1}} + \sigma^2 d \\ &+ \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{p_1}} + \sigma^2\right) \ln \delta^{-1}\Bigg)\Bigg),\end{aligned}$$

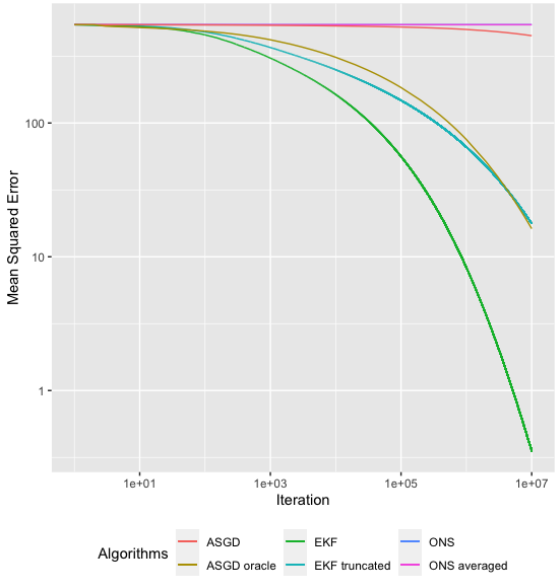

Figure 1: Mean squared error in log-log scale for $\theta^* = (-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T$. The ONS is applied with our optimistic exp-concavity constant $1.5 \cdot 10^{-14}$ instead of $e^{-D\sqrt{d}} \approx 1.2 \cdot 10^{-37}$. We observe that the algorithm still almost doesn't move.

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

6 Experiments

We experiment the static EKF for logistic regression. We first consider well-specified data generated by the same process as Bercu et al. (2020). Then we slightly change the simulation in order to obtain a misspecified setting.

The explanatory variables $X = (1, Z^T)^T$ are of dimension $d = 11$ where Z is a random vector composed of 10 independent components uniformly generated in $[0, 1]$. This yields $D_X = \sqrt{d}$. Then we define $\theta^* = (-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T$, and at each iteration t, the variable $y_t \in \{-1, 1\}$ is a Bernoulli variable of parameter $(1 + e^{-\theta^{*T} X_t})^{-1}$.

We compare the following sequential algorithms that we all initialize at $\hat{\theta}_1 = 0$:

- The 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 $10^{-10}/t^{0.49}$ as recommended by Bercu et al. (2020) would always coincide with the EKF.
- The ONS and the averaged version. The convex region of search is a ball centered in 0 and of radius $D = 1.1\|\theta^*\|$, a setting where we have good knowledge of θ^*. We implement two choices of the exp-concavity constant on which the ONS crucially relies. First, we use the optimal bound $e^{-D\sqrt{d}}$. Second, we use the minimum of the exp-concavity constants of 1000 points of the sphere. This yields an optimistic constant and a bigger step size, though we do not prove that the exp-concavity is satisfied.
- Two Average Stochastic Gradient Descent as described by Bach (2014). First we test the choice of the gradient step size $\gamma = 1/(2d\sqrt{N})$ denoted by ASGD and a second version with $\gamma = \|\theta^*\|/(\sqrt{dN})$ 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.

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 1 for $\theta^* = (-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T$. As this choice of θ^* yields a distribution of the Bernoulli parameter that is almost degenerated on the values 0 with small

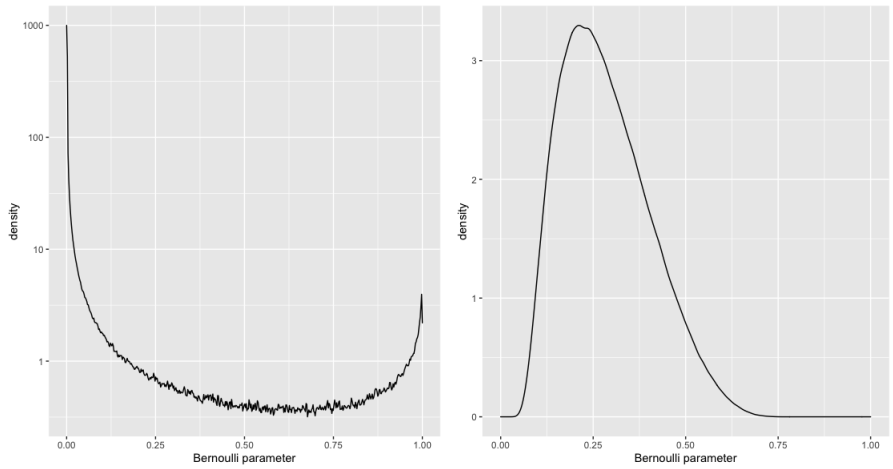

Figure 2: Density of $(1+e^{-\theta^{*T}X})^{-1}$ for $\theta^* = (-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T$ (left, the ordinate is in log scale) and $\theta^* = \frac{1}{10}(-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T$ (right) with 10^6 samples.

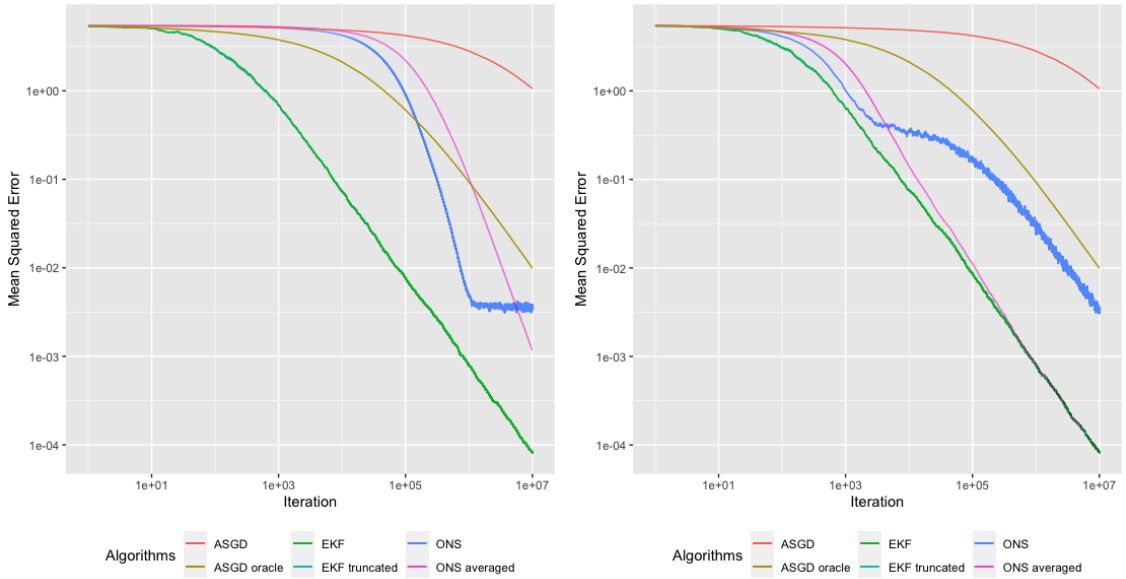

Figure 3: Mean squared error in log-log scale for $\theta^* = \frac{1}{10}(-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T$. The ONS is applied with the exp-concavity constant $e^{-D\sqrt{d}} \approx 2.0 \cdot 10^{-4}$ (left) and with our optimist exp-concavity constant $2.5 \cdot 10^{-2}$ (right).

mass at 1 (cf Figure 2), we run the same experiments for $\theta^* = \frac{1}{10}(-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T$. We display the results in Figure 3 for the second value of θ^*.

Finally, in order to demonstrate the robustness of the EKF we test the algorithms in a misspecified setting switching randomly between two well-specified logistic processes. We define $\theta_1 = \frac{1}{10}(-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T$ and θ_2 where we have only changed the first coefficient from $-9/10$ to $15/10$. Then y is a Bernoulli random variable whose parameter is either $(1+e^{-\theta_1^T X_t})^{-1}$ or $(1+e^{-\theta_2^T X_t})^{-1}$ uniformly at random. We present the results Figure 4.

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.

It appears clear that low exp-concavity constants is 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 in order to obtain a low mean

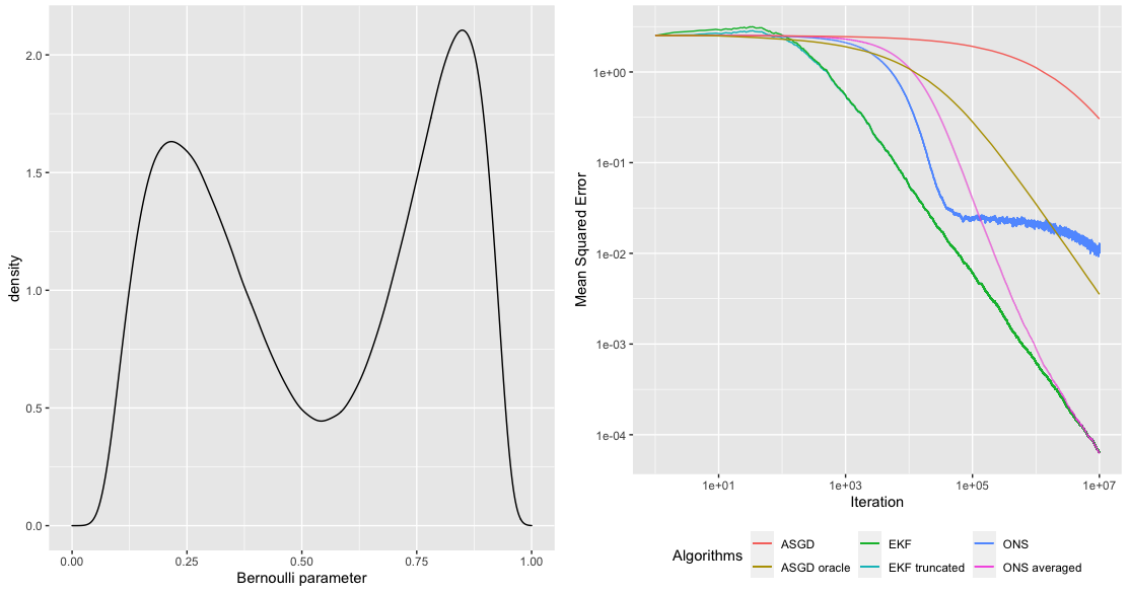

Figure 4: Misspecified setting. Density (left) of the Bernoulli parameter with two modes at $\mathbb{E}[(1+e^{-\theta_1^T X_t})^{-1}] \approx 0.28$ and $\mathbb{E}[(1+e^{-\theta_2^T X_t})^{-1}] \approx 0.79$. Mean squared error (right) where the ONS is applied with the exp-concavity constant $e^{-D\sqrt{d}} \approx 3.0 \cdot 10^{-3}$ and θ^* is estimated with 10^9 iterations of the static EKF.

squared error 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 setting the truncation deteriorates the performance of the EKF. Bercu et al. (2020) argue that the truncation is artificially introduced for the convergence property, they use the threshold $10^{-10}/t^{0.49}$ instead of $1/t^{0.49}$ and thus 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 we recommend to use the EKF in practice, or equivalently the truncated version with the threshold suggested by Bercu et al. (2020). The results are similar for both versions with a smaller θ^*, because our estimates $\hat{\theta}_t$ are smaller too so that the updates of the two versions coincide faster.

7 Conclusion

We have studied an efficient way to tackle some 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 and we used the EKF to approximate the maximum-likelihood estimator. We demonstrate its robustness to misspecification on locally exp-concave losses which can be expressed as GLM log-likelihoods, and 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 Filtering and the optimization community therefore we made the i.i.d. assumption standard in the stochastic optimization literature. It may lead the way to a risk analysis of the EKF in a non i.i.d. setting, where it might be necessary to assume that the data follows a well-specified state-space model.

References

S.-I. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

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

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

B. Bercu, A. Godichon, and B. Portier. An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1):348–367, 2020.

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

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

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

D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

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

D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1, 2012.

S. M. Kakade and A. Y. Ng. Online bounds for bayesian algorithms. In *Advances in neural information processing systems*, pages 641–648, 2005.

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

T. Koren. Open problem: Fast stochastic exp-concave optimization. In *Conference on Learning Theory*, pages 1073–1075, 2013.

M. Mahdavi, L. Zhang, and R. Jin. Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Conference on Learning Theory*, pages 1305–1320, 2015.

P. McCullagh and J. A. Nelder. *Generalized linear models*. London Chapman and Hall, 2nd ed edition, 1989.

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

D. Ostrovskii and F. Bach. Finite-sample analysis of m-estimators using self-concordance. *arXiv preprint arXiv:1810.06838*, 2018.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

P. Rigollet and J.-C. Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

D. Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4): 389–434, 2012.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

Organization of the Appendix

The Appendix follows the structure of the article:

- Appendix A contains the proofs of Section 3. Precisely, Lemma 4 is proved in Section A.1, the intermediate results of Sections 3.2 and 3.3 are proved in Sections A.2 and A.3, then Theorem 1 is proved in Section A.4 and Theorem 2 in Section A.5.
- Appendix B contains the proofs of Section 4. We derive the global bound (Theorem 10) in Section B.1, then we obtain the concentration result on P_t in Section B.2, and finally we prove the convergence of the truncated algorithm in Section B.3.
- Appendix C contains the proofs of Section 5. We prove Theorem 14 in Section C.1 and then in Section C.2 we prove the convergence of the algorithm, and we define an explicit value of $\tau(\varepsilon, \delta)$ satisfying Assumption 5.

A Proofs of Section 3

A.1 Proof of Lemma 4

We prove the following Lemma inspired by the stopping time technique of Freedman (1975) from which we derive Lemma 4. We give a general form useful in several proofs.

Lemma 15. *Let $(\mathcal{F}_n)$ be a filtration, and we consider a sequence of events (A_n) that is adapted to $(\mathcal{F}_n)$. Let (V_n) be a sequence of random variables adapted to $(\mathcal{F}_n)$ satisfying $V_0 = 1$, $V_n \geq 0$ almost surely for any n, and*

$$\mathbb{E}[V_n \mid \mathcal{F}_{n-1}, A_{n-1}] \leq V_{n-1}, \qquad n \geq 1.$$

Then for any $\delta > 0$, it holds

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

An important particular case is when (V_n) is a super-martingale adapted to the filtration $(\mathcal{F}_n)$ 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

$$E_k = \bigcup_{n=1}^{k} \left(V_n > \delta^{-1} \cup \overline{A_{n-1}}\right) .$$

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

$$\begin{aligned}
\mathbb{P}(E_k) &= \sum_{n=1}^{k} \mathbb{P}\left(E_n \cap \overline{E_{n-1}}\right) \\
&= \sum_{n=1}^{k} \mathbb{P}\left(\overline{A_{n-1}} \cap \overline{E_{n-1}}\right) + \sum_{n=1}^{k} \mathbb{P}\left(V_n > \delta^{-1} \cap \overline{E_{n-1}} \cap A_{n-1}\right) .
\end{aligned}$$

First, we have

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

Second, we apply the Chernoff bound:

$$\begin{aligned}
\sum_{n=1}^{k} \mathbb{P}\left(V_n > \delta^{-1} \cap \overline{E_{n-1}} \cap A_{n-1}\right) &= \sum_{n=1}^{k} \mathbb{E}\left[\frac{V_n}{\delta^{-1}} \mathbb{1}_{E_n \cap \overline{E_{n-1}} \cap A_{n-1}}\right] \\
&\leq \delta \sum_{n=1}^{k} \mathbb{E}\left[V_n (\mathbb{1}_{\overline{E_{n-1}} \cap A_{n-1}} - \mathbb{1}_{\overline{E_n}})\right] \\
&= \delta \sum_{n=1}^{k} \left(\mathbb{E}\left[V_n \mathbb{1}_{\overline{E_{n-1}} \cap A_{n-1}}\right] - \mathbb{E}\left[V_n \mathbb{1}_{\overline{E_n}}\right]\right) .
\end{aligned}$$

The second line is obtained since $\overline{E_n} \subset \left(\overline{E_{n-1}} \cap A_{n-1}\right)$. According to the tower property and the super-martingale assumption,

$$\mathbb{E}\left[V_n \mathbb{1}_{\overline{E_{n-1}} \cap A_{n-1}}\right] = \mathbb{E}\left[\mathbb{E}[V_n \mid \mathcal{F}_{n-1}, A_{n-1}] \mathbb{1}_{\overline{E_{n-1}}}\right] \leq \mathbb{E}\left[V_{n-1} \mathbb{1}_{\overline{E_{n-1}}}\right] .$$

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

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

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

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

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

Proof. of Lemma 4. Let $\lambda > 0$. For any $n \geq 1$, we define

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

Lemma B.1 of Bercu and Touati (2008) states that (V_n) is a super-martingale adapted to the filtration $(\mathcal{F}_{k+n})$. Moreover $V_0 = 1$ and for any n, it holds $V_n \geq 0$ almost surely. Therefore we can apply Lemma 15. □

A.2 Proofs of Sections 3.2

Proof. of Proposition 5. The first order condition of the optimum yields

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

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

$$\begin{aligned}\hat{\theta}_{t+1} - \hat{\theta}_1 &= (I - P_{t+1}X_tX_t^T)(\hat{\theta}_t - \hat{\theta}_1) + P_{t+1}y_tX_t - P_{t+1}X_tX_t^T\hat{\theta}_1\\ &= (I - P_{t+1}X_tX_t^T)P_t\sum_{s=1}^{t-1}(y_s - \hat{\theta}_1^T X_s)X_s + P_{t+1}(y_t - \hat{\theta}_1^T X_t)X_t\,.\end{aligned}$$

We conclude with the following identity:

$$(I - P_{t+1}X_tX_t^T)P_t = P_t - P_tX_tX_t^TP_t + \frac{P_tX_tX_t^TP_tX_tX_t^TP_t}{X_t^TP_tX_t+1} = P_t - \frac{P_tX_tX_t^TP_t}{X_t^TP_tX_t+1} = P_{t+1}\,.$$

□

Proof. of Lemma 6. We start from the update formula $\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t+1}\frac{(y_t - b'(\hat{\theta}_t^TX_t))X_t}{a}$ yielding

$$\begin{aligned}(\hat{\theta}_{t+1}-\theta^*)^TP_{t+1}^{-1}(\hat{\theta}_{t+1}-\theta^*) = (\hat{\theta}_t-\theta^*)^TP_{t+1}^{-1}(\hat{\theta}_t-\theta^*) + 2\frac{(y_t - b'(\hat{\theta}_t^TX_t))X_t^T}{a}(\hat{\theta}_t-\theta^*)\\ + X_t^TP_{t+1}X_t\left(\frac{y_t - b'(\hat{\theta}_t^TX_t)}{a}\right)^2.\end{aligned}$$

With a summation argument, re-arranging terms, we obtain:

$$\begin{aligned}&\sum_{t=1}^{n}\left(\frac{(b'(\hat{\theta}_t^TX_t) - y_t)X_t^T}{a}(\hat{\theta}_t-\theta^*) - \frac{1}{2}(\hat{\theta}_t-\theta^*)^T(P_{t+1}^{-1} - P_t^{-1})(\hat{\theta}_t-\theta^*)\right)\\ &=\frac{1}{2}\sum_{t=1}^{n}X_t^TP_{t+1}X_t\left(\frac{y_t - b'(\hat{\theta}_t^TX_t)}{a}\right)^2\\ &\quad+\frac{1}{2}\sum_{t=1}^{n}\left((\hat{\theta}_t-\theta^*)^TP_t^{-1}(\hat{\theta}_t-\theta^*) - (\hat{\theta}_{t+1}-\theta^*)^TP_{t+1}^{-1}(\hat{\theta}_{t+1}-\theta^*)\right).\end{aligned}$$

We bound the telescopic sum: as $P_{n+1}^{-1} \succcurlyeq 0$, we have

$$\begin{aligned}&\sum_{t=\tau+1}^{\tau+n}\left((\hat{\theta}_t-\theta^*)^TP_t^{-1}(\hat{\theta}_t-\theta^*) - (\hat{\theta}_{t+1}-\theta^*)^TP_{t+1}^{-1}(\hat{\theta}_{t+1}-\theta^*)\right)\\ &\qquad\leq (\hat{\theta}_1-\theta^*)^TP_1^{-1}(\hat{\theta}_1-\theta^*) \leq \frac{\|\hat{\theta}_1-\theta^*\|^2}{\lambda_{\min}(P_1)}\,.\end{aligned}$$

The result follows from the identities

$$\frac{(b'(\hat{\theta}_t^TX_t) - y_t)X_t}{a} = \ell'(y_t, \hat{\theta}_t^TX_t)X_t\,,\qquad P_{t+1}^{-1} - P_t^{-1} = \ell''(y_t, \hat{\theta}_t^TX_t)X_tX_t^T\,.$$

□

A.3 Proofs of Section 3.3

Proof. of Proposition 7. We recall that $\mathbb{E}_{y\sim p_{\theta^*}(y|X)}[y] = b'(\theta^{*T}X)$, therefore

$$\mathbb{E}_{y\sim p_{\theta^*}(y|X)}\left[\frac{(b'(\theta^T X)-y)(\theta-\theta^*)^T X}{a}\right] = \frac{(\theta-\theta^*)^T X}{a}\left(b'(\theta^T X) - b'(\theta^{*T}X)\right) .$$

Thus, there exists $\lambda \in [0,1]$ such that

$$\mathbb{E}_{y\sim p_{\theta^*}(y|X)}\left[\frac{(b'(\theta^T X)-y)(\theta-\theta^*)^T X}{a}\right] = \frac{(\theta-\theta^*)^T X}{a} b''\left(\theta^T X + \lambda(\theta^*-\theta)^T X\right)(\theta-\theta^*)^T X .$$

Then we use Assumption 3:

$$\frac{b''\left(\theta^T X + \lambda(\theta^*-\theta)^T X\right)}{b''\left(\theta^T X\right)} = \frac{\ell''\left(y_t, \theta^T X + \lambda(\theta^*-\theta)^T X\right)}{\ell''\left(y_t, \theta^T X\right)} \geq \rho_{\|\theta-\theta^*\|} ,$$

yielding

$$\mathbb{E}_{y\sim p_{\theta^*}(y|X)}\left[\ell'(y,\theta^T X)X\right]^T(\theta-\theta^*) \geq \rho_{\|\theta-\theta^*\|}(\theta-\theta^*)^T\left(\ell''(y,\theta^T X)XX^T\right)(\theta-\theta^*) . \tag{3}$$

The first-order condition satisfied by θ^* is

$$\mathbb{E}\left[-\frac{(y-b'(\theta^{*T}X))X}{a}\right] = 0 ,$$

which is re-written

$$\mathbb{E}\left[yX\right] = \mathbb{E}[b'(\theta^{*T}X)X] = \mathbb{E}[\mathbb{E}_{y\sim p_{\theta^*}(y|X)}[y]X] .$$

Plugging it into Equation 3, we obtain

$$\mathbb{E}[\ell'(y,\theta^T X)X]^T(\theta-\theta^*) \geq \rho_{\|\theta-\theta^*\|}(\theta-\theta^*)^T\mathbb{E}[\ell''(y,\theta^T X)XX^T](\theta-\theta^*) .$$

□

Proof. of Proposition 8. We first recall that $L(\theta) - L(\theta^*) \leq \frac{\partial L}{\partial\theta}\Big|_\theta^T(\theta-\theta^*)$, then Proposition 7 yields

$$\frac{\partial L}{\partial\theta}\Big|_\theta^T(\theta-\theta^*) - c(\theta-\theta^*)^T\frac{\partial^2 L}{\partial\theta^2}\Big|_\theta(\theta-\theta^*) \geq (1-\frac{c}{\rho_{\|\theta-\theta^*\|}})\frac{\partial L}{\partial\theta}\Big|_\theta^T(\theta-\theta^*) ,$$

and the result follows. □

Proof. of Lemma 9. We first develop $(\Delta M_t)^2$:

$$\begin{aligned}
(\Delta M_t)^2 &= \Big(\left(\mathbb{E}\left[\nabla_t \mid \mathcal{F}_{t-1}\right] - \nabla_t\right)^T(\hat{\theta}_t-\theta^*)\Big)^2 \\
&= (\hat{\theta}_t-\theta^*)^T\Big(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^T + \nabla_t\nabla_t^T \\
&\quad - \nabla_t\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^T - \mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]\nabla_t^T\Big)(\hat{\theta}_t-\theta^*) \\
&\leq 2(\hat{\theta}_t-\theta^*)^T\Big(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^T + \nabla_t\nabla_t^T\Big)(\hat{\theta}_t-\theta^*) \\
&\leq 2(\hat{\theta}_t-\theta^*)^T\Big(\mathbb{E}[\nabla_t\nabla_t^T \mid \mathcal{F}_{t-1}] + \nabla_t\nabla_t^T\Big)(\hat{\theta}_t-\theta^*) .
\end{aligned}$$

The third line holds because if $U, V \in \mathbb{R}^d$, it holds $-UV^T - VU^T \preccurlyeq UU^T + VV^T$. The last one comes from $\mathbb{E}\Big[(\nabla_t - \mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}])(\nabla_t - \mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}])^T \mid \mathcal{F}_{t-1}\Big] \succcurlyeq 0$.

Also, we have the relation

$$\mathbb{E}[(\Delta M_t)^2 \mid \mathcal{F}_{t-1}] \leq (\hat{\theta}_t-\theta^*)^T\mathbb{E}[\nabla_t\nabla_t^T \mid \mathcal{F}_{t-1}](\hat{\theta}_t-\theta^*) .$$

It yields

$$(\Delta M_t)^2 + \mathbb{E}[(\Delta M_t)^2 \mid \mathcal{F}_{t-1}] \leq (\hat{\theta}_t-\theta^*)^T\left(3\mathbb{E}[\nabla_t\nabla_t^T \mid \mathcal{F}_{t-1}] + 2\nabla_t\nabla_t^T\right)(\hat{\theta}_t-\theta^*) ,$$

and the result follows from Lemma 4. □

We derive the following Lemma in order to control the right-hand side of Lemma 6, in both settings.

Lemma 16. *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*

$$\sum_{t=k+1}^{k+n} \mathrm{Tr}\left(P_{t+1}(P_{t+1}^{-1} - P_t^{-1})\right) \leq d \ln\left(1 + n\frac{h_\varepsilon \lambda_{\max}(P_{k+1}) D_X^2}{d}\right).$$

Proof. We apply Lemma 11.11 of Cesa-Bianchi and Lugosi (2006):

$$\begin{aligned}
\sum_{t=k+1}^{k+n} \mathrm{Tr}\left(P_{t+1}(P_{t+1}^{-1} - P_t^{-1})\right) &= \sum_{t=k+1}^{k+n} \left(1 - \frac{\det(P_t^{-1})}{\det(P_{t+1}^{-1})}\right) \\
&\leq \sum_{t=k+1}^{k+n} \ln\left(\frac{\det(P_{t+1}^{-1})}{\det(P_t^{-1})}\right) \\
&= \ln\left(\frac{\det(P_{k+n+1}^{-1})}{\det(P_{k+1}^{-1})}\right) \\
&\leq \ln\det\left(I + \sum_{t=k+1}^{k+n} \ell''(y_t, \hat{\theta}_t^T X_t)(P_{k+1}^{1/2} X_t)(P_{k+1}^{1/2} X_t)^T\right) \\
&= \sum_{i=1}^{d} \ln(1+\lambda_i),
\end{aligned}$$

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

$$\begin{aligned}
\sum_{t=k+1}^{k+n} \mathrm{Tr}\left(P_{t+1}(P_{t+1}^{-1} - P_t^{-1})\right) &\leq d \ln\left(1 + \frac{1}{d}\sum_{i=1}^{d} \lambda_i\right) \\
&\leq d \ln\left(1 + \frac{1}{d} n h_\varepsilon \lambda_{\max}(P_{k+1}) D_X^2\right).
\end{aligned}$$

□

A.4 Bounded setting (Assumption 3)

Proof. of Theorem 1. Let $\delta > 0$. On the one hand, we sum Lemma 6 and 9. We obtain, for any $\lambda > 0$,

$$\begin{aligned}
&\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \left(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^T(\hat{\theta}_t - \theta^*) - \frac{1}{2}Q_t - \lambda(\hat{\theta}_t - \theta^*)^T\left(\nabla_t \nabla_t^T + \frac{3}{2}\mathbb{E}\left[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1}\right]\right)(\hat{\theta}_t - \theta^*)\right) \\
&\quad \leq \frac{1}{2}\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} X_t^T P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^T X_t)^2 + \frac{\|\hat{\theta}_1 - \theta^*\|^2}{\lambda_{\min}(P_{\tau(\varepsilon,\delta)+1})} + \frac{\ln \delta^{-1}}{\lambda}, \qquad n \geq 1, \qquad (4)
\end{aligned}$$

with probability at least $1-\delta$, where we define $Q_t = (\hat{\theta}_t - \theta^*)^T\left(\ell''(y_t, \hat{\theta}_t^T X_t) X_t X_t^T\right)(\hat{\theta}_t - \theta^*)$ for any t.

On the other hand, thanks to Assumption 3, we can apply Proposition 8 with $c = 0.75$ to obtain, for any $t \geq 1$,

$$\begin{aligned}
\|\hat{\theta}_t - \theta^*\| \leq \varepsilon \implies L(\hat{\theta}_t) - L(\theta^*) &\leq \frac{\rho_\varepsilon}{\rho_\varepsilon - 0.75}\Big(\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^T(\hat{\theta}_t - \theta^*) - 0.75(\hat{\theta}_t - \theta^*)^T \frac{\partial^2 L}{\partial \theta^2}\Big|_{\hat{\theta}_t}(\hat{\theta}_t - \theta^*)\Big), \\
\implies L(\hat{\theta}_t) - L(\theta^*) &\leq 5\Big(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^T(\hat{\theta}_t - \theta^*) - 0.75\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]\Big), \qquad (5)
\end{aligned}$$

because $\rho_\varepsilon > 0.95$.

In order to bridge the gap between Equations (4) and (5), we need to control the quadratic terms of Equation (4) with $\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]$. First, for any t, if $\|\hat{\theta}_t - \theta^*\| \leq \varepsilon$, we have $Q_t \in [0, h_\varepsilon \varepsilon^2 D_X^2]$, and we apply Lemma A.3 of Cesa-Bianchi and Lugosi (2006) to the random variable $\frac{1}{h_\varepsilon \varepsilon^2 D_X^2} Q_t \in [0,1]$: for any $s > 0$,

$$\mathbb{E}\left[\exp\left(\frac{s}{h_\varepsilon \varepsilon^2 D_X^2} Q_t - \frac{e^s - 1}{h_\varepsilon \varepsilon^2 D_X^2} \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right]\right) \mid \mathcal{F}_{t-1}, \|\hat{\theta}_t - \theta^*\| \leq \varepsilon\right] \leq 1\,.$$

We fix $s = 0.1$ and we define

$$V_n = \exp\left(\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \left(\frac{0.1}{h_\varepsilon \varepsilon^2 D_X^2} Q_t - (e^{0.1} - 1)\mathbb{E}\left[\frac{1}{h_\varepsilon \varepsilon^2 D_X^2} Q_t \mid \mathcal{F}_{t-1}\right]\right)\right)\,.$$

The sequence (V_n) is adapted to $(\mathcal{F}_{\tau(\varepsilon,\delta)+n})$, almost surely we have $V_0 = 1$ and $V_n \geq 0$. Finally,

$$\mathbb{E}[V_n \mid \mathcal{F}_{\tau(\varepsilon,\delta)+n-1}, \|\hat{\theta}_{\tau(\varepsilon,\delta)+n} - \theta^*\| \leq \varepsilon| \leq V_{n-1}\,,$$

and $(\|\hat{\theta}_{\tau(\varepsilon,\delta)+n} - \theta^*\| \leq \varepsilon)$ belongs to $\mathcal{F}_{\tau(\varepsilon,\delta)+n-1}$. We apply Lemma 15:

$$\mathbb{P}\left(\left(\bigcup_{n=1}^{\infty} V_n > \delta^{-1}\right) \cup \left(\bigcup_{n=1}^{\infty} (\|\hat{\theta}_{\tau(\varepsilon,\delta)+n} - \theta^*\| > \varepsilon)\right)\right) \leq \delta + \mathbb{P}\left(\bigcup_{n=1}^{\infty} (\|\hat{\theta}_{\tau(\varepsilon,\delta)+n} - \theta^*\| > \varepsilon)\right)\,.$$

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

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty}\left(\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} Q_t > 10(e^{0.1}-1)\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right] + 10 h_\varepsilon \varepsilon^2 D_X^2 \ln \delta^{-1}\right) \cap A^\varepsilon_{\tau(\varepsilon,\delta)}\right) \leq \delta\,. \quad (6)$$

We then bound the two quadratic terms coming from Lemma 9: using Assumption 3 we have the implications

$$\begin{aligned}
&\|\hat{\theta}_t - \theta^*\| \leq \varepsilon \implies (\hat{\theta}_t - \theta^*)^T \nabla_t \nabla_t^T (\hat{\theta}_t - \theta^*) \leq \kappa_\varepsilon Q_t\,,\\
&\|\hat{\theta}_t - \theta^*\| \leq \varepsilon \implies (\hat{\theta}_t - \theta^*)^T \mathbb{E}\left[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1}\right](\hat{\theta}_t - \theta^*) \leq \kappa_\varepsilon \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right]\,.
\end{aligned}$$

Therefore, we get from (6)

$$\begin{aligned}
\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n}\left(\frac{1}{2}Q_t + \lambda(\hat{\theta}_t - \theta^*)^T \nabla_t \nabla_t^T (\hat{\theta}_t - \theta^*) + \frac{3}{2}\lambda(\hat{\theta}_t - \theta^*)^T \mathbb{E}\left[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1}\right](\hat{\theta}_t - \theta^*)\right) > \\
\left(10(e^{0.1}-1)(\frac{1}{2} + \lambda\kappa_\varepsilon) + \frac{3}{2}\lambda\kappa_\varepsilon\right)\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right] + 10(\frac{1}{2} + \lambda\kappa_\varepsilon) h_\varepsilon \varepsilon^2 D_X^2 \ln \delta^{-1}\Bigg) \cap A^\varepsilon_{\tau(\varepsilon,\delta)}\Bigg) \\
\leq \delta\,.
\end{aligned}$$

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

$$\begin{aligned}
&10(e^{0.1}-1)(\frac{1}{2} + \lambda\kappa_\varepsilon) + \frac{3}{2}\lambda\kappa_\varepsilon = 0.75\,,\\
&\frac{1}{2} + \lambda\kappa_\varepsilon = \frac{1}{2} + \frac{0.75 - 5(e^{0.1}-1)}{10(e^{0.1}-1) + \frac{3}{2}} \approx 0.59 \leq 0.6\,,
\end{aligned}$$

and consequently

$$\begin{aligned}
\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(&\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n}\left(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^T(\hat{\theta}_t - \theta^*) - 0.75\mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]\right) > 6 h_\varepsilon \varepsilon^2 D_X^2 \ln \delta^{-1}\\
&+ \sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n}\left(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^T(\hat{\theta}_t - \theta^*) - \frac{1}{2}Q_t - \lambda(\hat{\theta}_t - \theta^*)^T\left(\nabla_t \nabla_t^T + \frac{3}{2}\mathbb{E}\left[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1}\right]\right)(\hat{\theta}_t - \theta^*)\right)\Bigg)\\
&\cap A^\varepsilon_{\tau(\varepsilon,\delta)}\Bigg) \leq \delta\,.
\end{aligned}$$

We plug Equation (5) in the last inequality:

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty} \Bigg(\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} (L(\hat{\theta}_t) - L(\theta^*)) > 30 h_\varepsilon \varepsilon^2 D_X^2 \ln \delta^{-1} \\ + 5 \sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \Bigg(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^T (\hat{\theta}_t - \theta^*) - \frac{1}{2} Q_t \\ - \lambda (\hat{\theta}_t - \theta^*)^T \left(\nabla_t \nabla_t^T + \frac{3}{2} \mathbb{E}\left[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1} \right] \right) (\hat{\theta}_t - \theta^*) \Bigg) \Bigg) \cap A^\varepsilon_{\tau(\varepsilon,\delta)} \Bigg) \leq \delta \,.$$

We then use Equation (4) with $\frac{1}{\lambda} = \frac{(10(e^{0.1}-1)+\frac{3}{2})\kappa_\varepsilon}{0.75-5(e^{0.1}-1)} \approx 11.4\kappa_\varepsilon \leq 12\kappa_\varepsilon$. It yields

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

Thanks to Assumption 3, we have

$$X_t^T P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^T X_t)^2 \leq \kappa_\varepsilon \operatorname{Tr}\left(P_{t+1}(P_{t+1}^{-1} - P_t^{-1}) \right), \qquad t > \tau(\varepsilon,\delta) \,,$$

therefore we apply Lemma 16: for any n, it holds

$$\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} X_t^T P_{t+1} X_t \ell'(y_t, \hat{\theta}_t^T X_t)^2 \leq d\kappa_\varepsilon \ln\left(1 + n \frac{h_\varepsilon \lambda_{\max}(P_{\tau(\varepsilon,\delta)+1}) D_X^2}{d} \right) .$$

As $P_{\tau(\varepsilon,\delta)+1} \preccurlyeq P_1$, we obtain

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty} \Bigg(\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} (L(\hat{\theta}_t) - L(\theta^*)) > \frac{5}{2} d\kappa_\varepsilon \ln\left(1 + n \frac{h_\varepsilon \lambda_{\max}(P_{\tau(\varepsilon,\delta)+1}) D_X^2}{d} \right) \\ + \frac{5\|\hat{\theta}_1 - \theta^*\|^2}{\lambda_{\min}(P_{\tau(\varepsilon,\delta)+1})} + 30(2\kappa_\varepsilon + h_\varepsilon \varepsilon^2 D_X^2) \ln \delta^{-1} \Bigg) \cap A^\varepsilon_{\tau(\varepsilon,\delta)} \Bigg) \leq 2\delta \,.$$

To conclude, we use Assumption 5. □

A.5 Quadratic setting (Assumption 4)

We recall two definitions introduced in the previous subsection:

$$A_k^\varepsilon = \bigcap_{n=k+1}^{\infty} (\|\hat{\theta}_n - \theta^*\| \leq \varepsilon), \qquad k \geq 1 \,,$$
$$Q_t = (\hat{\theta}_t - \theta^*)^T X_t X_t^T (\hat{\theta}_t - \theta^*), \qquad t \geq 1 \,.$$

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 Hütter (2015): for any $X \in \mathbb{R}^d$,

$$\mathbb{E}[(y - \mathbb{E}[y \mid X])^{2i} \mid X] \leq 2i(2\sigma^2)^i \Gamma(i) = 2(2\sigma^2)^i i!, \qquad i \in \mathbb{N}^* \,. \tag{7}$$

First, we control the quadratic terms in $\nabla_t = -(y_t - \hat{\theta}_t^T X_t) X_t$ in the following lemma.

Lemma 17. *1. For any $k \in \mathbb{N}$ and $\delta > 0$, we have*

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty} \Bigg(\sum_{t=k+1}^{k+n} (\hat{\theta}_t - \theta^*)^T \nabla_t \nabla_t^T (\hat{\theta}_t - \theta^*) \\ > 3\left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) \sum_{t=k+1}^{k+n} Q_t + 12\varepsilon^2 D_X^2 \sigma^2 \ln \delta^{-1} \Bigg) \cap A_k^\varepsilon \Bigg) \leq \delta \,.$$

2. *For any t, it holds almost surely*

$$(\hat{\theta}_t - \theta^*)^T \mathbb{E}[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1}](\hat{\theta}_t - \theta^*) \leq 3\left(\sigma^2 + D_{\text{app}}^2 + \|\hat{\theta}_t - \theta^*\|^2 D_X^2\right) \mathbb{E}[Q_t \mid \mathcal{F}_{t-1}]\,.$$

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

$$\begin{aligned}(\hat{\theta}_t - \theta^*)^T \nabla_t \nabla_t^T (\hat{\theta}_t - \theta^*) &= Q_t (y_t - \hat{\theta}_t^T X_t)^2 \\ &\leq 3Q_t\Big((y_t - \mathbb{E}[y_t \mid X_t])^2 + (\mathbb{E}[y_t \mid X_t] - \theta^{*T} X_t)^2 + ((\theta^* - \hat{\theta}_t)^T X_t)^2\Big) \\ &\leq 3Q_t\left((y_t - \mathbb{E}[y_t \mid X_t])^2 + D_{\text{app}}^2 + \|\hat{\theta}_t - \theta^*\|^2 D_X^2\right)\,. \end{aligned} \tag{8}$$

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 (7). For any t and any μ satisfying $0 < \mu \leq \frac{1}{4Q_t\sigma^2}$, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\mu Q_t (y_t - \mathbb{E}[y_t \mid X_t])^2\right) \mid \mathcal{F}_{t-1}, X_t\right] &= 1 + \sum_{i\geq 1} \frac{\mu^i Q_t^i \mathbb{E}[(y_t - \mathbb{E}\left[y_t \mid X_t\right])^{2i} \mid X_t]}{i!} \\ &\leq 1 + 2\sum_{i\geq 1} \frac{\mu^i Q_t^i i! (2\sigma^2)^i}{i!} \\ &\leq 1 + 2\sum_{i\geq 1} \left(2\mu Q_t \sigma^2\right)^i \\ &\leq 1 + 8\mu Q_t \sigma^2, \qquad 2\mu Q_t\sigma^2 \leq \frac{1}{2} \\ &\leq \exp\left(8\mu Q_t \sigma^2\right)\,.\end{aligned}$$

Therefore, for any t,

$$\mathbb{E}\left[\exp\left(\frac{1}{4\varepsilon^2 D_X^2 \sigma^2} Q_t\left((y_t - \mathbb{E}[y_t \mid X_t])^2 - 8\sigma^2\right)\right) \mid \mathcal{F}_{t-1}, X_t, \|\hat{\theta}_t - \theta^*\| \leq \varepsilon\right] \leq 1\,.$$

We define the random variable

$$V_n = \exp\left(\frac{1}{4\varepsilon^2 D_X^2 \sigma^2} \sum_{t=k+1}^{k+n} Q_t\left((y_t - \mathbb{E}[y_t \mid X_t])^2 - 8\sigma^2\right)\right), \qquad n \in \mathbb{N}\,.$$

$(V_n)_n$ is adapted to the filtration $(\sigma(X_1, y_1, ..., X_{k+n}, y_{k+n}, X_{k+n+1})_n$, moreover $V_0 = 1$ and $V_n \geq 0$ almost surely, and

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

Therefore we apply Lemma 15: for any $\delta > 0$,

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

which is equivalent to

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

Substituting in Equation (8), we obtain the desired result.

2. We apply the same decomposition as for Equation 8: for any t,

$$\begin{aligned}&(\hat{\theta}_t - \theta^*)^T \mathbb{E}[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1}](\hat{\theta}_t - \theta^*) \\ &\quad\leq 3(\hat{\theta}_t - \theta^*)^T \mathbb{E}\left[X_t X_t^T \left((y_t - \mathbb{E}[y_t \mid X_t])^2 + D_{\text{app}}^2 + \|\theta^* - \hat{\theta}_t\|^2 D_X^2\right) \mid \mathcal{F}_{t-1}\right](\hat{\theta}_t - \theta^*)\,.\end{aligned}$$

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

$$(\hat{\theta}_t - \theta^*)^T \mathbb{E}[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1}](\hat{\theta}_t - \theta^*) \\ \leq 3\left(\sigma^2 + D_{\text{app}}^2 + \|\hat{\theta}_t - \theta^*\|^2 D_X^2\right)(\hat{\theta}_t - \theta^*)^T \mathbb{E}[X_t X_t^T \mid \mathcal{F}_{t-1}](\hat{\theta}_t - \theta^*)\,.$$

□

Second, we bound the right-hand side of Lemma 6, that is the objective of the following lemma.

Lemma 18. *Let $k \in \mathbb{N}$. For any $\delta > 0$, we have*

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=k+1}^{k+n} X_t^T P_{t+1} X_t (y_t - \hat{\theta}_t^T X_t)^2 > 3\left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) d \ln\left(1 + n\frac{\lambda_{\max}(P_{k+1}) D_X^2}{d}\right) \\ + 12\lambda_{\max}(P_1) D_X^2 \sigma^2 \ln \delta^{-1}\Bigg) \cap A_k^{\varepsilon}\Bigg) \leq \delta\,.$$

Proof. We apply a similar analysis as in the proof of Lemma 17 in order to use the sub-gaussian assumption, and then we apply the telescopic argument as in the bounded setting. We decompose $y_t - \hat{\theta}_t^T X_t$:

$$X_t^T P_{t+1} X_t (y_t - \hat{\theta}_t^T X_t)^2 \leq 3 X_t^T P_{t+1} X_t \Big((y_t - \mathbb{E}[y_t \mid X_t])^2 + (\mathbb{E}[y_t \mid X_t] - b'(\theta^{*T} X_t))^2 + ((\theta^* - \hat{\theta}_t)^T X_t)^2\Big) \\ \leq 3 X_t^T P_{t+1} X_t \left((y_t - \mathbb{E}[y_t \mid X_t])^2 + D_{\text{app}}^2 + \|\hat{\theta}_t - \theta^*\|^2 D_X^2\right)\,. \tag{9}$$

To control $(y_t - \mathbb{E}[y_t \mid X_t])^2 X_t^T P_{t+1} X_t$, we use its positivity along with Equation (7). Precisely, for any t and any $\mu > 0$ satisfying $0 < \mu \leq \frac{1}{4 X_t^T P_{t+1} X_t \sigma^2}$, we have

$$\mathbb{E}\left[\exp\left(\mu (y_t - \mathbb{E}[y_t \mid X_t])^2 X_t^T P_{t+1} X_t\right) \mid \mathcal{F}_{t-1}, X_t\right] = 1 + \sum_{i \geq 1} \frac{\mu^i (X_t^T P_{t+1} X_t)^i \mathbb{E}\left[(y_t - \mathbb{E}[y_t \mid X_t])^{2i} \mid X_t\right]}{i!} \\ \leq 1 + 2\sum_{i \geq 1} \frac{\mu^i (X_t^T P_{t+1} X_t)^i i! (2\sigma^2)^i}{i!} \\ = 1 + 2\sum_{i \geq 1} \left(2\mu X_t^T P_{t+1} X_t \sigma^2\right)^i \\ \leq 1 + 8\mu X_t^T P_{t+1} X_t \sigma^2, \qquad 0 < 2\mu X_t^T P_{t+1} X_t \sigma^2 \leq \frac{1}{2} \\ \leq \exp\left(8\mu X_t^T P_{t+1} X_t \sigma^2\right)\,.$$

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

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

(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 15:

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

or equivalently

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(\sum_{t=k+1}^{k+n} X_t^T P_{t+1} X_t (y_t - \mathbb{E}[y_t \mid X_t])^2 > 8\sigma^2 \sum_{t=k+1}^{k+n} X_t^T P_{t+1} X_t + 4\lambda_{\max}(P_1) D_X^2 \sigma^2 \ln \delta^{-1}\Bigg)\Bigg) \leq \delta\,.$$

Combining it with Equation (9), we get

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty} \Bigg(\sum_{t=k+1}^{k+n} X_t^T P_{t+1} X_t (y_t - \hat{\theta}_t^T X_t)^2 > 3\left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) \sum_{t=k+1}^{k+n} X_t^T P_{t+1} X_t \\ + 12\lambda_{\max}(P_1) D_X^2 \sigma^2 \ln \delta^{-1} \Bigg) \cap A_k^{\varepsilon} \Bigg) \leq \delta \,.$$

Then we apply Lemma 16: the second point of Assumption 3 holds with $h_\varepsilon = 1$, thus

$$\sum_{t=k+1}^{k+n} \operatorname{Tr}\left(P_{t+1}(P_{t+1}^{-1} - P_t^{-1})\right) \leq d \ln\left(1 + n\frac{\lambda_{\max}(P_{k+1}) D_X^2}{d}\right), \qquad n \geq 1.$$

We conclude with $X_t^T P_{t+1} X_t = \operatorname{Tr}(P_{t+1}(P_{t+1}^{-1} - P_t^{-1}))$. □

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 6 and Lemma 9: for any $\lambda, \delta > 0$,

$$\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \left(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^T (\hat{\theta}_t - \theta^*) - \frac{1}{2} Q_t - \lambda (\hat{\theta}_t - \theta^*)^T \left(\nabla_t \nabla_t^T + \frac{3}{2} \mathbb{E}\left[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1}\right] \right) (\hat{\theta}_t - \theta^*) \right) \\ \leq \frac{1}{2} \sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} X_t^T P_{t+1} X_t (y_t - \hat{\theta}_t^T X_t)^2 + \frac{\|\hat{\theta}_{\tau(\varepsilon,\delta)+1} - \theta^*\|^2}{\lambda_{\min}(P_{\tau(\varepsilon,\delta)+1})} + \frac{\ln \delta^{-1}}{\lambda}, \qquad n \geq 1\,, \quad (10)$$

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

$$\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} (L(\hat{\theta}_t) - L(\theta^*)) \leq \frac{1}{1 - 0.8} \sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \left(\mathbb{E}[\nabla_t \mid \mathcal{F}_{t-1}]^T (\hat{\theta}_t - \theta^*) - 0.8 \mathbb{E}[Q_t \mid \mathcal{F}_{t-1}] \right). \quad (11)$$

We aim to relate Equations (10) and (11) as in the proof of Theorem 1. To that end, we apply Lemma 17:

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty} \Bigg(\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \left(\frac{1}{2} Q_t + \lambda (\hat{\theta}_t - \theta^*)^T \left(\nabla_t \nabla_t^T + \frac{3}{2} \mathbb{E}\left[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1}\right] \right) (\hat{\theta}_t - \theta^*) \right) \\ > \left(\frac{1}{2} + 3\lambda(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) \right) \sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} Q_t \\ + \frac{9}{2}\lambda \left(\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2 \right) \sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right] + 12\lambda\varepsilon^2 D_X^2 \sigma^2 \ln \delta^{-1} \Bigg) \cap A_{\tau(\varepsilon,\delta)}^{\varepsilon} \Bigg) \leq \delta \,.$$

As in the proof of Theorem 1 we apply Lemma A.3 of (Cesa-Bianchi and Lugosi, 2006) and Lemma 15: for any $\delta > 0$,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} \left(\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} Q_t > 10(e^{0.1} - 1) \sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right] + 10\varepsilon^2 D_X^2 \ln \delta^{-1} \right) \cap A_{\tau(\varepsilon,\delta)}^{\varepsilon} \right) \leq \delta \,.$$

We combine the last two inequalities:

$$\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty} \Bigg(\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \left(\frac{1}{2} Q_t + \lambda (\hat{\theta}_t - \theta^*)^T \left(\nabla_t \nabla_t^T + \frac{3}{2} \mathbb{E}\left[\nabla_t \nabla_t^T \mid \mathcal{F}_{t-1}\right] \right) (\hat{\theta}_t - \theta^*) \right) \\ > \left(10(e^{0.1} - 1)\left(\frac{1}{2} + 3\lambda(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) \right) + \frac{9}{2}\lambda(\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) \right) \sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right] \\ + \left(10\varepsilon^2 D_X^2 \left(\frac{1}{2} + 3\lambda(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2) \right) + 12\lambda\varepsilon^2 D_X^2 \sigma^2 \right) \ln \delta^{-1} \Bigg) \cap A_{\tau(\varepsilon,\delta)}^{\varepsilon} \Bigg) \leq 2\delta \,. \quad (12)$$

We set

$$\lambda = \big(0.8 - 5(e^{0.1}-1)\big)\left(30(e^{0.1}-1)(8\sigma^2 + D_{\rm app}^2 + \varepsilon^2 D_X^2) + \frac{9}{2}(\sigma^2 + D_{\rm app}^2 + \varepsilon^2 D_X^2)\right)^{-1}$$

in order to obtain

$$\begin{aligned}
&10(e^{0.1}-1)\Big(\frac{1}{2} + 3\lambda(8\sigma^2 + D_{\rm app}^2 + \varepsilon^2 D_X^2)\Big) + \frac{9}{2}\lambda(\sigma^2 + D_{\rm app}^2 + \varepsilon^2 D_X^2) = 0.8\,,\\
&\frac{1}{109\sigma^2 + 28D_{\rm app}^2 + 28\varepsilon^2 D_X^2} < \lambda < \frac{1}{108\sigma^2 + 27D_{\rm app}^2 + 27\varepsilon^2 D_X^2}\,,\\
&10\varepsilon^2 D_X^2\Big(\frac{1}{2} + 3\lambda(8\sigma^2 + D_{\rm app}^2 + \varepsilon^2 D_X^2)\Big) + 12\lambda D_X^2\varepsilon^2\sigma^2 \le 8\varepsilon^2 D_X^2\\
&\frac{1}{\lambda} \le 28(4\sigma^2 + D_{\rm app}^2 + \varepsilon^2 D_X^2)\,.
\end{aligned}$$

Combining Equations (10), (11) and (12), we obtain

$$\begin{aligned}
\mathbb{P}\Bigg(\bigcup_{n=1}^{\infty}\Bigg(0.2\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n}(L(\hat\theta_t) - L(\theta^*)) > \frac{1}{2}\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} X_t^T P_{t+1} X_t (y_t - \hat\theta_t^T X_t)^2 + \frac{\varepsilon^2}{\lambda_{\min}(P_{\tau(\varepsilon,\delta)+1})}\\
+ 28(4\sigma^2 + D_{\rm approx}^2 + \varepsilon^2 D_X^2)\ln\delta^{-1} + 8\varepsilon^2 D_X^2\ln\delta^{-1}\Bigg) \cap A^{\varepsilon}_{\tau(\varepsilon,\delta)}\Bigg) \le 3\delta\,.
\end{aligned}$$

Finally, we apply Lemma 18 with $P_{\tau(\varepsilon,\delta)+1} \preccurlyeq P_1$ and we use Assumption 5: it holds simultaneously

$$\begin{aligned}
\sum_{t=\tau(\varepsilon,\delta)+1}^{\tau(\varepsilon,\delta)+n} L(\hat\theta_t) - L(\theta^*) \le 5\Bigg(&\frac{3}{2}\left(8\sigma^2 + D_{\rm app}^2 + \varepsilon^2 D_X^2\right) d\ln\left(1 + n\frac{\lambda_{\max}(P_1)D_X^2}{d}\right) + \lambda_{\max}\left(P^{-1}_{\tau(\varepsilon,\delta)+1}\right)\varepsilon^2\\
&+ 28(4\sigma^2 + D_{\rm approx}^2 + \varepsilon^2 D_X^2)\ln\delta^{-1} + 8\varepsilon^2 D_X^2\ln\delta^{-1}\\
&+ 6\lambda_{\max}(P_1)D_X^2\sigma^2\ln\delta^{-1}\Bigg), \qquad n \ge 1\,,
\end{aligned}$$

with probability at least $1 - 5\delta$. To conclude, we write

$$28(4\sigma^2 + D_{\rm approx}^2 + \varepsilon^2 D_X^2) + 8\varepsilon^2 D_X^2 + 6\lambda_{\max}(P_1)D_X^2\sigma^2 \le 28\left(\sigma^2(4 + \frac{\lambda_{\max}(P_1)D_X^2}{4}) + D_{\rm app}^2 + 2\varepsilon^2 D_X^2\right)\,.$$

□

B Proofs of Section 4

B.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

$$\lambda_{\max}(P^{-1}_{\tau(\varepsilon,\delta)+1}) \le \lambda_{\max}(P_1^{-1}) + \frac{1}{4}\sum_{t=1}^{\tau(\varepsilon,\delta)}\|X_t\|^2\,,$$

$$\frac{5\kappa_\varepsilon}{2} < 3e^{D_X\|\theta^*\|}, \qquad 30\Big(2\kappa_\varepsilon + \frac{\varepsilon^2 D_X^2}{4}\Big) < 64e^{D_X\|\theta^*\|}, \qquad 5\varepsilon^2 D_X^2 \le 1/75\,.$$

We then control the first terms. To that end, we use a rough bound at any time $t \ge 1$:

$$\begin{aligned}
L(\hat\theta_t) - L(\theta^*) &\le \mathbb{E}\left[\frac{yX}{1 + e^{y\hat\theta_t^T X}} \mid \hat\theta_t\right]^T(\hat\theta_t - \theta^*)\\
&\le D_X\|\hat\theta_t - \theta^*\|\\
&\le D_X(\|\hat\theta_1 - \theta^*\| + (t-1)\lambda_{\max}(P_1)D_X)\,,
\end{aligned}$$

because for any $s \ge 1$, we have $P_s \preccurlyeq P_1$ and therefore $\|\hat\theta_{s+1} - \hat\theta_s\| \le \lambda_{\max}(P_1)D_X$. Summing from 1 to $\tau(\varepsilon,\delta) \le \tau(\frac{1}{20D_X},\delta)$ yields the result. □

B.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 19. *If Assumption 1 is satisfied, then for any* $0 \leq \beta < 1$ *and* $t \geq 4^{1/(1-\beta)}$*, it holds*

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

Proof. We wish to center the matrices $X_s X_s^T$ 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,

$$\begin{aligned}
\lambda_{\min}(A-B) &= \min_x \frac{x^T(A-B)x}{\|x\|^2} \\
&\leq \frac{v^T(A-B)v}{\|v\|^2} \\
&= \lambda_{\min}(A) - \frac{v^T B v}{\|v\|^2} \\
&\leq \lambda_{\min}(A) - \min_x \frac{x^T B x}{\|x\|^2} \\
&= \lambda_{\min}(A) - \lambda_{\min}(B)\,.
\end{aligned}$$

We obtain:

$$\begin{aligned}
\lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_s X_s^T}{s^\beta} - \sum_{s=1}^{t-1}\mathbb{E}\left[\frac{X_s X_s^T}{s^\beta}\right]\right) &\leq \lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_s X_s^T}{s^\beta}\right) - \lambda_{\min}\left(\sum_{s=1}^{t-1}\mathbb{E}\left[\frac{X_s X_s^T}{s^\beta}\right]\right) \\
&= \lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_s X_s^T}{s^\beta}\right) - \Lambda_{\min}\sum_{s=1}^{t-1}\frac{1}{s^\beta} \\
&\leq \lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_s X_s^T}{s^\beta}\right) - \Lambda_{\min}\frac{t^{1-\beta}-1}{1-\beta}\,.
\end{aligned}$$

Therefore, we obtain

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

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

- Obviously $\mathbb{E}\left[\frac{X_s X_s^T}{s^\beta}\right] - \frac{X_s X_s^T}{s^\beta}$ is centered,
- $\lambda_{\max}\left(\mathbb{E}\left[\frac{X_s X_s^T}{s^\beta}\right] - \frac{X_s X_s^T}{s^\beta}\right) \leq \lambda_{\max}\left(\mathbb{E}\left[\frac{X_s X_s^T}{s^\beta}\right]\right) \leq D_X^2$ almost surely.

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

$$0 \preccurlyeq \sum_{s=1}^{t-1}\mathbb{E}\left[\left(\mathbb{E}\left[\frac{X_s X_s^T}{s^\beta}\right] - \frac{X_s X_s^T}{s^\beta}\right)^2\right] \preccurlyeq \left(\sum_{s=1}^{t-1}\frac{D_X^4}{s^\beta}\right) I \preccurlyeq \left(D_X^4\frac{t^{1-\beta}}{1-\beta}\right) I\,.$$

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

$$\begin{aligned}
&\mathbb{P}\left(\lambda_{\max}\left(\sum_{s=1}^{t-1}\left(\mathbb{E}\left[\frac{X_sX_s^T}{s^\beta}\right]-\frac{X_sX_s^T}{s^\beta}\right)\right)>\frac{\Lambda_{\min}t^{1-\beta}}{2(1-\beta)}\right)\\
&\quad\le d\exp\left(-\frac{\Lambda_{\min}^2t^{2(1-\beta)}/(8(1-\beta)^2)}{D_X^4t^{1-\beta}/(1-\beta)+D_X^2\Lambda_{\min}t^{1-\beta}/(6(1-\beta))}\right)\\
&\quad=d\exp\left(-t^{1-\beta}\frac{\Lambda_{\min}^2}{8D_X^4}\frac{1/(1-\beta)^2}{1/(1-\beta)+\Lambda_{\min}/(6D_X^2(1-\beta))}\right)\\
&\quad=d\exp\left(-t^{1-\beta}\frac{\Lambda_{\min}^2}{8D_X^4}\left(1-\beta+\frac{\Lambda_{\min}(1-\beta)}{6D_X^2}\right)^{-1}\right).
\end{aligned}$$

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

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

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

We can now do a union bound to obtain Proposition 11.

Proof. of Proposition 11. We reduce our problem to the deviations of a sum of centered independent random matrices:

$$\begin{aligned}
\lambda_{\max}(P_t)&=\lambda_{\min}\left(P_1^{-1}+\sum_{s=1}^{t-1}X_sX_s^T\alpha_s\right)^{-1}\\
&\le\lambda_{\min}\left(P_1^{-1}+\sum_{s=1}^{t-1}\frac{X_sX_s^T}{s^\beta}\right)^{-1},
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\max}(P_t)>\frac{4}{\Lambda_{\min}t^{1-\beta}}\right)&\le\mathbb{P}\left(\lambda_{\min}\left(P_1^{-1}+\sum_{s=1}^{t-1}\frac{X_sX_s^T}{s^\beta}\right)^{-1}>\frac{4}{\Lambda_{\min}t^{1-\beta}}\right)\\
&=\mathbb{P}\left(\lambda_{\min}\left(P_1^{-1}+\sum_{s=1}^{t-1}\frac{X_sX_s^T}{s^\beta}\right)<\frac{\Lambda_{\min}t^{1-\beta}}{4}\right)\\
&\le\mathbb{P}\left(\lambda_{\min}\left(\sum_{s=1}^{t-1}\frac{X_sX_s^T}{s^\beta}\right)<\frac{\Lambda_{\min}t^{1-\beta}}{4}\right)\\
&\le d\exp\left(-t^{1-\beta}\frac{\Lambda_{\min}^2}{10D_X^4}\right),
\end{aligned}$$

where we applied Lemma 19 to obtain the last line. We take a union bound to obtain, for any $k\ge 7$,

$$\begin{aligned}
\mathbb{P}\left(\exists t>k,\lambda_{\max}(P_t)>\frac{4}{\Lambda_{\min}t^{1-\beta}}\right)&\le\sum_{t>k}d\exp\left(-t^{1-\beta}\frac{\Lambda_{\min}^2}{10D_X^4}\right)\\
&\le d\sum_{t>k}\exp\left(-\lfloor t^{1-\beta}\rfloor\frac{\Lambda_{\min}^2}{10D_X^4}\right)\\
&=d\sum_{m\ge 1}\exp\left(-m\frac{\Lambda_{\min}^2}{10D_X^4}\right)\sum_{t>k}\mathbb{1}_{\lfloor t^{1-\beta}\rfloor=m}
\end{aligned}$$

We bound $\sum_{t>k}\mathbb{1}_{\lfloor t\rfloor=m}$: for any m

$$\lfloor t^{1-\beta}\rfloor=m\implies m^{1/(1-\beta)}\le t<(m+1)^{1/(1-\beta)},$$

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

$$\begin{aligned}(m+1)^{1/(1-\beta)} &= m^{1/(1-\beta)}(1+1/m)^{1/(1-\beta)} \\ &= m^{1/(1-\beta)} \exp(\ln(1+1/m)/(1-\beta)) \\ &\leq m^{1/(1-\beta)} \exp(1/(m(1-\beta))) \\ &\leq m^{1/(1-\beta)}(1+2/(m(1-\beta)))\,,\end{aligned}$$

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

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

and that is true for $m = 1$ too. Hence

$$\begin{aligned}\mathbb{P}\left(\exists t > k, \lambda_{\max}(P_t) > \frac{4}{\Lambda_{\min} t^{1-\beta}}\right) &\leq 4d \sum_{m \geq \lfloor k^{1-\beta} \rfloor} (m+1) \exp\left(-m \frac{\Lambda_{\min}^2}{10 D_X^4}\right) \\ &= 4d \frac{\exp\left(-\frac{\Lambda_{\min}^2}{10D_X^4}\right)^{\lfloor k^{1-\beta} \rfloor}}{1 - \exp\left(-\frac{\Lambda_{\min}^2}{10D_X^4}\right)} (\lfloor k^{1-\beta} \rfloor + 1 + \frac{\exp\left(-\frac{\Lambda_{\min}^2}{10D_X^4}\right)}{1 - \exp\left(-\frac{\Lambda_{\min}^2}{10D_X^4}\right)}) \\ &\leq 4d \frac{\exp\left(\frac{\Lambda_{\min}^2}{10D_X^4}\right)}{1 - \exp\left(-\frac{\Lambda_{\min}^2}{10D_X^4}\right)} (k^{1-\beta} + \frac{1}{1 - \exp\left(-\frac{\Lambda_{\min}^2}{10D_X^4}\right)}) \exp\left(-\frac{\Lambda_{\min}^2}{10D_X^4}\right)^{k^{1-\beta}}\,,\end{aligned}$$

where the second line is obtained deriving both sides of $\sum_{m \geq \lfloor k^{1-\beta} \rfloor} r^{m+1} = \frac{r^{\lfloor k^{1-\beta} \rfloor + 1}}{1-r}$ with respect to r. Also, as $1 - e^{-x} \geq x e^{-x}$ for any $x \in \mathbb{R}$, we get

$$\begin{aligned}&\mathbb{P}\left(\exists t > k, \lambda_{\max}(P_t) > \frac{4}{\Lambda_{\min} t^{1-\beta}}\right) \\ &\quad \leq 4d \frac{10 D_X^4}{\Lambda_{\min}^2} \exp\left(2 \frac{\Lambda_{\min}^2}{10 D_X^4}\right) (k^{1-\beta} + \frac{10 D_X^4}{\Lambda_{\min}^2} \exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right)) \exp\left(-\frac{\Lambda_{\min}^2}{10 D_X^4}\right)^{k^{1-\beta}}\,.\end{aligned}$$

Also, as $x e^{-x} \leq e^{-1}$ for any $x \geq 0$, we get for any $k \geq 7$:

$$\begin{aligned}\left(k^{1-\beta} + \frac{10 D_X^4}{\Lambda_{\min}^2} \exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right)\right) \exp\left(-k^{1-\beta} \frac{\Lambda_{\min}^2}{20 D_X^4}\right) &\leq \frac{20 D_X^4 e^{-1}}{\Lambda_{\min}^2} \exp\left(\frac{10 D_X^4}{\Lambda_{\min}^2} \exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right) \frac{\Lambda_{\min}^2}{20 D_X^4}\right) \\ &= \frac{20 D_X^4 e^{-1}}{\Lambda_{\min}^2} \exp\left(\frac{1}{2} \exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right)\right)\,.\end{aligned}$$

Combining the last two inequalities, we obtain

$$\begin{aligned}\mathbb{P}\left(\exists t > k, \lambda_{\max}(P_t) > \frac{4}{\Lambda_{\min} t^{1-\beta}}\right) &\leq d \frac{800 D_X^8 e^{-1}}{\Lambda_{\min}^4} \exp\left(2 \frac{\Lambda_{\min}^2}{10 D_X^4} + \frac{1}{2} \exp\left(\frac{\Lambda_{\min}^2}{10 D_X^4}\right)\right) \exp\left(-k^{1-\beta} \frac{\Lambda_{\min}^2}{20 D_X^4}\right) \\ &\leq d \frac{625 D_X^8}{\Lambda_{\min}^4} \exp\left(-k^{1-\beta} \frac{\Lambda_{\min}^2}{20 D_X^4}\right)\,,\end{aligned}$$

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

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

The condition $k \geq 7$ is not necessary because

$$\left(\frac{20 D_X^4}{\Lambda_{\min}^2} \ln\left(\frac{625 d D_X^8}{\Lambda_{\min}^4 \delta}\right)\right)^{1/(1-\beta)} \geq 20 \ln(625 \delta^{-1})\,,$$

and either $\delta \geq 1$ and the result is trivial, either $\delta < 1$ and $20 \ln(625 \delta^{-1}) \geq 128$. □

B.3 Convergence of the truncated algorithm

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

Lemma 20. *Let $\theta \in \mathbb{R}^d$.*

1. *For any $\eta > 0$, we have*
$$L(\theta) - L(\theta^*) > \eta \implies \left\| \frac{\partial L}{\partial \theta}\Big|_\theta \right\| \geq D_\eta$$
for $D_\eta = \frac{\Lambda_{\min}\sqrt{\eta}}{\sqrt{2}D_X(1+e^{D_X(\|\theta^*\|+\sqrt{8\eta/D_X^2})})}$.
2. *For any $\varepsilon > 0$, we have*
$$\|\theta - \theta^*\| > \varepsilon \implies L(\theta) - L(\theta^*) > \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^*\|+\varepsilon)})}\varepsilon^2 .$$

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

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

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 θ^* around some $\theta_0 \in \mathbb{R}^d$, we get
$$L(\theta^*) \geq L(\theta_0) + \frac{\partial L}{\partial \theta}\Big|_{\theta_0}^T (\theta^* - \theta_0) + \frac{1}{4(1+e^{D_X(\|\theta^*\|+\|\theta_0-\theta^*\|)})}(\theta_0 - \theta^*)^T \mathbb{E}\left[XX^T\right](\theta_0 - \theta^*) ,$$
and that yields
$$\frac{\partial L}{\partial \theta}\Big|_{\theta_0}^T (\theta_0 - \theta^*) \geq L(\theta_0) - L(\theta^*) + \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^*\|+\|\theta_0-\theta^*\|)})}\|\theta_0 - \theta^*\|^2 .$$
Therefore, as $L(\theta_0) - L(\theta^*) \geq 0$,
$$\left\| \frac{\partial L}{\partial \theta}\Big|_{\theta_0} \right\| \geq \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^*\|+\|\theta_0-\theta^*\|)})}\|\theta_0 - \theta_{\text{true}}\| .$$
Finally, as L is convex of minimum θ^*,
$$\begin{aligned}\left\| \frac{\partial L}{\partial \theta}\Big|_{\theta} \right\| &\geq \min_{\|\theta_0-\theta^*\|=\sqrt{8\eta/D_X^2}} \left\| \frac{\partial L}{\partial \theta}\Big|_{\theta_0} \right\| \\ &\geq \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^*\|+\sqrt{8\eta/D_X^2})})}\sqrt{8\eta/D_X^2} \\ &\geq \frac{\Lambda_{\min}}{\sqrt{2}D_X(1+e^{D_X(\|\theta^*\|+\sqrt{8\eta/D_X^2})})}\sqrt{\eta} .\end{aligned}$$
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 θ^*, if $\|\theta - \theta^*\| > \varepsilon$ it holds
$$L(\theta) - L(\theta^*) > \min_{\|\theta_0-\theta^*\|=\varepsilon} L(\theta_0) - L(\theta^*) \geq \frac{\Lambda_{\min}}{4(1+e^{D_X(\|\theta^*\|+\varepsilon)})}\varepsilon^2 .$$

□

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 θ^* follows. The convergence of $(L(\hat{\theta}_t))_t$ comes from the first point of Lemma 20. 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

$$L(\hat{\theta}_{t+1}) = L(\hat{\theta}_t) + \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^T (\hat{\theta}_{t+1} - \hat{\theta}_t) + \frac{1}{2}(\hat{\theta}_{t+1} - \hat{\theta}_t)^T \frac{\partial^2 L}{\partial \theta^2}\Big|_{\hat{\theta}_t + \alpha_t(\hat{\theta}_{t+1} - \hat{\theta}_t)} (\hat{\theta}_{t+1} - \hat{\theta}_t)\,. \tag{13}$$

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

$$\begin{aligned}(\hat{\theta}_{t+1} - \hat{\theta}_t)^T \frac{\partial^2 L}{\partial \theta^2}\Big|_{\hat{\theta}_t + \alpha_t(\hat{\theta}_{t+1} - \hat{\theta}_t)} (\hat{\theta}_{t+1} - \hat{\theta}_t) &\leq \frac{1}{(1 + e^{y_t \hat{\theta}_t^T X_t})^2} X_t^T P_{t+1}^T \frac{\mathbb{E}[XX^T]}{4} P_{t+1} X_t \\ &\leq \frac{1}{4} D_X^4 \lambda_{\max}(P_{t+1})^2 \leq \frac{1}{4} D_X^4 \lambda_{\max}(P_t)^2\,.\end{aligned}$$

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

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

and as $\alpha_t \leq 1$,

$$\left\| -\alpha_t \frac{P_t X_t X_t^T P_t}{1 + X_t^T P_t X_t \alpha_t} \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \right\| \leq D_X^3 \lambda_{\max}(P_t)^2\,.$$

Also, $\left\|\frac{\partial L}{\partial \theta}\right\| \leq D_X$. Substituting our findings in Equation (13), we obtain

$$L(\hat{\theta}_{t+1}) \leq L(\hat{\theta}_t) + \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^T P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} + 2 D_X^4 \lambda_{\max}(P_t)^2\,. \tag{14}$$

We define

$$\begin{aligned}M_t &= \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^T P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} - \mathbb{E}\left[\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^T P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \mid X_1, y_1, ..., X_{t-1}, y_{t-1}\right] \\ &= \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^T P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} + \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^T P_t \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}\,.\end{aligned}$$

Hence we have

$$\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}^T P_t \frac{y_t X_t}{1 + e^{y_t \hat{\theta}_t^T X_t}} \leq M_t - \lambda_{\min}(P_t) \left\|\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}\right\|^2 \leq M_t - \frac{1}{t D_X^2} \left\|\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_t}\right\|^2\,,$$

because $P_s \succcurlyeq \frac{I}{s D_X^2}$. Combining it with Equation (14) and summing consecutive terms, we obtain, for any $k < t$,

$$L(\hat{\theta}_t) - L(\hat{\theta}_k) \leq \sum_{s=k}^{t-1} \left(M_s - \frac{1}{s D_X^2} \left\|\frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_s}\right\|^2 + 2 D_X^4 \lambda_{\max}(P_s)^2 \right)\,. \tag{15}$$

We recall that there exists C_δ 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 left-hand side is the sum of a martingale and a term which is negative for s large enough, under the event A_{C_δ}.

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 $B_{k,t}$ be the event $(\forall k < s < t, L(\hat{\theta}_s) - L(\theta^*) > \eta/2)$. Then we use the law of total probability:

$$\begin{aligned}\mathbb{P}(L(\hat{\theta}_t) - L(\theta^*) > \eta \mid A_{C_\delta}) &= \mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^*) > \eta) \cap B_{0,t} \mid A_{C_\delta}\right) \\ &\quad + \sum_{k=1}^{t-1} \mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^*) > \eta) \cap \left(L(\hat{\theta}_k) - L(\theta^*) \leq \frac{\eta}{2}\right) \cap B_{k,t} \mid A_{C_\delta}\right) \qquad (16) \\ &\leq \mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^*) > \eta) \cap B_{0,t} \mid A_{C_\delta}\right) \\ &\quad + \sum_{k=1}^{t-1} \mathbb{P}\left(\left(L(\hat{\theta}_t) - L(\hat{\theta}_k) > \frac{\eta}{2}\right) \cap B_{k,t} \mid A_{C_\delta}\right) .\end{aligned}$$

Lemma 20 yields

$$L(\hat{\theta}_s) - L(\theta^*) > \frac{\eta}{2} \implies \left\| \frac{\partial L}{\partial \theta}\Big|_{\hat{\theta}_s} \right\| \geq D_\eta \, .$$

We combine the last equation, along with Equation (15) and the definition of A_{C_δ} to get, for any $1 \leq k < t$,

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

where $f(k,t) = \frac{\eta}{2} + \frac{D_\eta^2}{D_X^2} \sum_{s=k}^{t-1} \frac{1}{s} - 2D_X^4 C_\delta^2 \sum_{s=k}^{t-1} \frac{1}{s^{2(1-\beta)}}$ 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 $f_0(t) = \eta - (L(\hat{\theta}_1) - L(\theta^*)) + \frac{D_\eta^2}{D_X^2} \sum_{s=1}^{t-1} \frac{1}{s} - 2D_X^4 C_\delta^2 \sum_{s=1}^{t-1} \frac{1}{s^{2(1-\beta)}}$ 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 $|M_s| \leq 2D_X^2 \lambda_{\max}(P_s)$. We can therefore apply Azuma-Hoeffding inequality: for t, k such that $f(k,t) > 0$,

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

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 $f_0(t) > 0$,

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

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 $\frac{D_\eta^2}{D_X^2}\ln(t)$ 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 21. *For* $t \geq \max\left(e^{\frac{16 D_X^6 C_\delta^2}{D_\eta^2 (1-2\beta)}}, \left(1 + \left(\frac{8 D_X^4 C_\delta^2}{\eta(1-2\beta)}\right)^{\frac{1}{1-2\beta}}\right)^2\right)$*, it holds*

$$\begin{aligned} f(k,t) &\geq \frac{D_\eta^2}{4 D_X^2} \ln(t), && 1 \leq k < \sqrt{t}, \\ f(k,t) &\geq \frac{\eta}{4}, && \sqrt{t} \leq k < t \, . \end{aligned}$$

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

$$f_0(t) \geq \frac{D_\eta^2}{2D_X^2}\ln(t)\,.$$

Then, defining $C_1 = \frac{D_\eta^4(1-2\beta)}{256D_X^8C_\delta^2}$ and $C_2 = \frac{\eta^2(1-2\beta)}{128D_X^4C_\delta^2}$, we finally get for t large enough:

$$\begin{aligned}
\mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^*) > \eta) \cap B_{0,t} \mid A_{C_\delta}\right) &\leq \exp\left(-4C_1\ln(t)^2\right), \\
\mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^*) > \eta) \cap (L(\hat{\theta}_k) - L(\theta^*) \leq \frac{\eta}{2}) \cap B_{k,t} \mid A_{C_\delta}\right) &\leq \exp\left(-C_1\ln(t)^2\right), \qquad 1 \leq k < \sqrt{t} \\
\mathbb{P}\left((L(\hat{\theta}_t) - L(\theta^*) > \eta) \cap (L(\hat{\theta}_k) - L(\theta^*) \leq \frac{\eta}{2}) \cap B_{k,t} \mid A_{C_\delta}\right) &\leq \exp\left(-C_2(k-1)^{1-2\beta}\right), \qquad \sqrt{t} \leq k < t
\end{aligned}$$

Substituting in Equation (16) yields:

$$\begin{aligned}
\mathbb{P}(L(\hat{\theta}_t) - L(\theta^*) > \eta \mid A_C) &\leq \exp\left(-4C_1\ln(t)^2\right) + \sum_{k=1}^{\lceil\sqrt{t}\rceil-1}\exp\left(-C_1\ln(t)^2\right) + \sum_{k=\lceil\sqrt{t}\rceil}^{t-1}\exp\left(-C_2(k-1)^{1-2\beta}\right) \\
&\leq (\sqrt{t}+1)\exp\left(-C_1\ln(t)^2\right) + t\exp\left(-C_2(\sqrt{t}-1)^{1-2\beta}\right).
\end{aligned}$$

Finally, Point 2 of Lemma 20 allows to obtain the result: defining $\eta = \frac{\Lambda_{\min}\varepsilon^2}{4(1+e^{D_X(\|\theta^*\|+\varepsilon)})}$, we obtain

$$\begin{aligned}
\mathbb{P}(\|\hat{\theta}_t - \theta^*\| > \varepsilon \mid A_{C_\delta}) &\leq \mathbb{P}(L(\hat{\theta}_t) - L(\theta^*) > \eta \mid A_{C_\delta}) \\
&\leq (\sqrt{t}+1)\exp\left(-C_1\ln(t)^2\right) + t\exp\left(-C_2(\sqrt{t}-1)^{1-2\beta}\right).
\end{aligned}$$

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

$$\begin{aligned}
D_\eta &= \frac{\Lambda_{\min}\sqrt{\frac{\Lambda_{\min}\varepsilon^2}{4(1+e^{D_X(\|\theta^*\|+\varepsilon)})}}}{2D_X(1+\exp\left(D_X(\|\theta^*\| + \sqrt{\frac{\Lambda_{\min}\varepsilon^2}{D_X^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})}}\right))} \geq \left(\frac{\Lambda_{\min}}{1+e^{D_X(\|\theta^*\|+\varepsilon)}}\right)^{3/2}\frac{\varepsilon}{4D_X}, \\
C_1 &\geq \frac{\Lambda_{\min}^6(1-2\beta)\varepsilon^4}{2^{16}D_X^{12}C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})^6}, \\
C_2 &\geq \frac{\Lambda_{\min}^2(1-2\beta)\varepsilon^4}{2^{11}D_X^4C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})^2},
\end{aligned}$$

and the conditions of Lemma 21 become

$$\begin{aligned}
t &\geq \exp\left(\frac{2^8D_X^8C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})^3}{\Lambda_{\min}^3(1-2\beta)\varepsilon^2}\right), \\
t &\geq \left(1+\left(\frac{32D_X^4C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})}{(1-2\beta)\Lambda_{\min}\varepsilon^2}\right)^{\frac{1}{1-2\beta}}\right)^2, \\
t &\geq \exp\left(\frac{32D_X^4(1+e^{D_X(\|\theta^*\|+\varepsilon)})^3}{\Lambda_{\min}^3\varepsilon^2}\left(L(\hat{\theta}_1) - L(\theta^*) + \frac{4D_X^4C_\delta^2}{1-2\beta}\right)\right).
\end{aligned}$$

We would like to obtain a single condition on t, thus we write

$$\begin{aligned}\left(1+\left(\frac{32D_X^4C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})}{(1-2\beta)\Lambda_{\min}\varepsilon^2}\right)^{\frac{1}{1-2\beta}}\right)^2 &= \exp\left(2\ln\left(1+\left(\frac{32D_X^4C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})}{(1-2\beta)\Lambda_{\min}\varepsilon^2}\right)^{\frac{1}{1-2\beta}}\right)\right)\\ &\leq \exp\left(\frac{2}{1-2\beta}\ln\left(1+\frac{32D_X^4C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})}{(1-2\beta)\Lambda_{\min}\varepsilon^2}\right)\right)\\ &\leq \exp\left(\frac{2}{1-2\beta}\sqrt{\frac{32D_X^4C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})}{(1-2\beta)\Lambda_{\min}\varepsilon^2}}\right)\\ &\leq \exp\left(\frac{2^8D_X^8C_\delta^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})^3}{\Lambda_{\min}^3(1-2\beta)^{3/2}\varepsilon^2}\right),\end{aligned}$$

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:

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

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

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

□

Proof. of Lemma 21. We recall that for any $k\geq 1$,

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

Therefore:

$$\begin{aligned}f(k,t)&\geq\frac{\eta}{2}+\frac{D_\eta^2}{D_X^2}(\ln t-\ln k)-\frac{2D_X^4C_\delta^2}{1-2\beta}\frac{1}{\max(1/2,(k-1)^{1-2\beta})}\,,\\ f_0(t)&\geq\eta-(L(\hat{\theta}_1)-L(\theta^*)+\frac{D_\eta^2}{D_X^2}\ln t-\frac{4D_X^4C_\delta^2}{1-2\beta}\,.\end{aligned}$$

- For any $1\leq k<\sqrt{t}$, $\ln k\leq\frac{1}{2}\ln t$, and we have

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

  and taking $t\geq e^{\frac{16D_X^6C_\delta^2}{D_\eta^2(1-2\beta)}}$ yields $f(k,t)\geq\frac{D_\eta^2}{4D_X^2}\ln(t)$.

- For $t\geq 2$ and any $k\geq\sqrt{t}$, we have

  $$f(k,t)\geq\frac{\eta}{2}-\frac{2D_X^4C_\delta^2}{(1-2\beta)(k-1)^{1-2\beta}}\geq\frac{\eta}{2}-\frac{2D_X^4C_\delta^2}{(1-2\beta)(\sqrt{t}-1)^{1-2\beta}}\,.$$

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

- Last point comes from $f_0(t)\geq\frac{D_\eta^2}{D_X^2}\ln t-(L(\hat{\theta}_1)-L(\theta^*)-\frac{4D_X^4C_\delta^2}{1-2\beta}$.

□

Proof. of Corollary 13. We apply Theorem 12: for any $t \geq \exp\left(\frac{2^8 D_X^8 C_{\delta/2}^2 (1+e^{D_X(\|\theta^*\|+\varepsilon)})^3}{\Lambda_{\min}^3 (1-2\beta)^{3/2}\varepsilon^2}\right)$,

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

where

$$C_1 = \frac{\Lambda_{\min}^6 (1-2\beta)\varepsilon^4}{2^{16} D_X^{12} C_{\delta/2}^2 (1+e^{D_X(\|\theta^*\|+\varepsilon)})^6}, \qquad C_2 = \frac{\Lambda_{\min}^2 (1-2\beta)\varepsilon^4}{2^{11} D_X^4 C_{\delta/2}^2 (1+e^{D_X(\|\theta^*\|+\varepsilon)})^2}.$$

We use a union bound: for any $\tau \geq \exp\left(\frac{2^8 D_X^8 C_{\delta/2}^2 (1+e^{D_X(\|\theta^*\|+\varepsilon)})^3}{\Lambda_{\min}^3 (1-2\beta)^{3/2}\varepsilon^2}\right)$,

$$\mathbb{P}\left(\bigcup_{t=\tau+1}^{\infty} (\|\hat{\theta}_t - \theta^*\| > \varepsilon) \mid A_{C_{\delta/2}}\right) \leq \sum_{t>\tau} (\sqrt{t}+1)\exp\left(-C_1 \ln(t)^2\right) + \sum_{t>\tau} t\exp\left(-C_2(\sqrt{t}-1)^{1-2\beta}\right).$$

- If $\tau \geq e^{\frac{3}{2C_1}}$, we have

$$\sum_{t>\tau} (\sqrt{t}+1)\exp\left(-C_1 \ln(t)^2\right) \leq \sum_{t>\tau} (\sqrt{t}+1)\frac{1}{t^{5/2}} \leq 2/\tau,$$

- For $t \geq 4$, $1 - 1/\sqrt{t} \geq 1/2$, then for $t \geq \left(\frac{12}{C_2(1-2\beta)}\right)^{4/(1-2\beta)}$,

$$\begin{aligned} t^3 \exp\left(-C_2(\sqrt{t}-1)^{1-2\beta}\right) &\leq \exp\left(3\ln(t) - \frac{C_2}{2} t^{(1-2\beta)/2}\right) \\ &\leq \exp\left(\frac{12}{1-2\beta}\ln\left(\frac{12}{C_2(1-2\beta)}\right) - \frac{6}{1-2\beta}\left(\frac{12}{C_2(1-2\beta)}\right)\right) \\ &\leq 1, \end{aligned}$$

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

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

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

Finally, for τ big enough, we obtain

$$\mathbb{P}\left(\bigcup_{t=\tau+1}^{\infty} (\|\hat{\theta}_t - \theta^*\| > \varepsilon) \mid A_{C_{\delta/2}}\right) \leq 3/\tau \leq \delta/2,$$

if $\tau \geq 6\delta^{-1}$. We now compare the constants involved. As long as $\varepsilon D_X \leq 1$, we have

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

Furthermore, as $1 - 2\beta \leq 1$, we have

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

Finally,

$$\begin{aligned}
\left(\frac{12}{C_2(1-2\beta)}\right)^{4/(1-2\beta)} &= \exp\left(\frac{4}{1-2\beta}\ln\frac{12}{C_2(1-2\beta)}\right) \\
&= \exp\left(\frac{4}{1-2\beta}\ln\frac{12\cdot 2^{11}D_X^4 C_{\delta/2}^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})^2}{\Lambda_{\min}^2(1-2\beta)^2\varepsilon^4}\right) \\
&= \exp\left(\frac{8}{1-2\beta}\ln\frac{12\cdot 2^{11}D_X^4 C_{\delta/2}^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})^2}{\Lambda_{\min}^2(1-2\beta)\varepsilon^4}\right) \\
&\leq \exp\left(\frac{8}{1-2\beta}\sqrt{\frac{3\cdot 2^{13}D_X^4 C_{\delta/2}^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})^2}{\Lambda_{\min}^2(1-2\beta)\varepsilon^4}}\right) \\
&= \exp\left(\frac{\sqrt{6}2^9 D_X^2 C_{\delta/2}(1+e^{D_X(\|\theta^*\|+\varepsilon)})}{\Lambda_{\min}(1-2\beta)^{3/2}\varepsilon^2}\right) \\
&\leq \exp\left(\frac{3\cdot 2^{15}D_X^{12}C_{\delta/2}^2(1+e^{D_X(\|\theta^*\|+\varepsilon)})^6}{\Lambda_{\min}^6(1-2\beta)^{3/2}\varepsilon^4}\right).
\end{aligned}$$

□

C Proofs of Section 5

C.1 Proof of Theorem 14

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

Lemma 22. *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*

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

with probability at least $1-\delta$.

Proof. From Proposition 5, we obtain, for any $t \geq 1$, $\hat{\theta}_t - \hat{\theta}_1 = P_t\sum_{s=1}^{t-1}(y_s - \hat{\theta}_1^T X_s)X_s$. Consequently,

$$\begin{aligned}
\hat{\theta}_t - \theta^* &= P_t\sum_{s=1}^{t-1}(y_s - \hat{\theta}_1^T X_s)X_s - P_t\left(P_1^{-1} + \sum_{s=1}^{t-1}X_s X_s^T\right)(\theta^* - \hat{\theta}_1) \\
&= P_t\sum_{s=1}^{t-1}(y_s - \theta^{*T}X_s)X_s + P_t P_1^{-1}(\hat{\theta}_1 - \theta^*),
\end{aligned}$$

and using $P_t P_1^{-1} \preccurlyeq I$, we obtain

$$\begin{aligned}
\|\hat{\theta}_t - \theta^*\| &\leq \|\hat{\theta}_1 - \theta^*\| + \lambda_{\max}(P_t)D_X\sum_{s=1}^{t-1}|y_s - \theta^{*T}X_s| \\
&\leq \|\hat{\theta}_1 - \theta^*\| + \lambda_{\max}(P_1)D_X\sum_{s=1}^{t-1}\left(|y_s - \mathbb{E}[y_s \mid X_s]| + D_{\text{app}}\right). \qquad (17)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left[\exp(\mu|y_t - \mathbb{E}[y_t \mid X_t]|)\right] &= 1 + \sum_{i\geq 1} \frac{\mu^i \mathbb{E}[|y_t - \mathbb{E}[y_t \mid X_t]|^i]}{i!} \\
&\leq 1 + \sum_{k\geq 1} \frac{\mu^i (2\sigma^2)^{i/2} i\Gamma(i/2)}{i!} \\
&\leq 1 + \sum_{i\geq 1} \left(\sqrt{2}\mu\sigma\right)^i, \qquad \text{because } \Gamma(i/2) \leq \Gamma(i) = (i-1)! \\
&\leq 1 + 2\sqrt{2}\mu\sigma, \qquad \text{because } 0 < \sqrt{2}\mu\sigma \leq \frac{1}{2} \\
&\leq \exp\left(2\sqrt{2}\mu\sigma\right).
\end{aligned}$$

Thus we can apply Lemma 15 to the super-martingale $\left(\exp\left(\frac{1}{2\sqrt{2}\sigma}\sum_{s=1}^{t}(|y_s - \mathbb{E}[y_s \mid X_s]| - 2\sqrt{2}\sigma)\right)\right)_t$ in order to obtain, for any $\delta > 0$,

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

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

Proof. of Theorem 14. We first apply Theorem 2: with probability at least $1 - 5\delta$, it holds simultaneously

$$\begin{aligned}
\sum_{t=\tau(\varepsilon,\delta)+1}^{n} L(\hat{\theta}_t) - L(\theta^*) \leq{}& \frac{15}{2} d\left(8\sigma^2 + D_{\text{app}}^2 + \varepsilon^2 D_X^2\right) \ln\left(1 + (n - \tau(\varepsilon,\delta))\frac{\lambda_{\max}(P_1)D_X^2}{d}\right) \\
&+ 5\lambda_{\max}\left(P_{\tau(\varepsilon,\delta)+1}^{-1}\right)\varepsilon^2 \\
&+ 115\left(\sigma^2(4 + \frac{\lambda_{\max}(P_1)D_X^2}{4}) + D_{\text{app}}^2 + 2\varepsilon^2 D_X^2\right)\ln\delta^{-1}, \qquad n \geq \tau(\varepsilon,\delta).
\end{aligned}$$

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

Then we derive a bound on the first $\tau(\varepsilon,\delta)$ 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 22 we obtain the simultaneous property

$$\begin{aligned}
L(\hat{\theta}_t) - L(\theta^*) \leq{}& 2D_X^2(\|\hat{\theta}_1 - \theta^*\| + 3\lambda_{\max}(P_1)D_X\sigma\ln\delta^{-1})^2 \\
&+ 2\lambda_{\max}(P_1)^2 D_X^4(3\sigma + D_{\text{app}})^2(t-1)^2, \qquad t \geq 1,
\end{aligned}$$

with probability at least $1 - \delta$.

Thus, a summation argument yields, for any $\delta > 0$,

$$\begin{aligned}
\sum_{t=1}^{\tau(\varepsilon,\delta)} L(\hat{\theta}_t) - L(\theta^*) \leq{}& 2D_X^2(\|\hat{\theta}_1 - \theta^*\| + 3\lambda_{\max}(P_1)D_X\sigma\ln\delta^{-1})^2\tau(\varepsilon,\delta) \\
&+ \lambda_{\max}(P_1)^2 D_X^4(3\sigma + D_{\text{app}})^2 \frac{(\tau(\varepsilon,\delta)-1)\tau(\varepsilon,\delta)(2\tau(\varepsilon,\delta)-1)}{3},
\end{aligned}$$

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

C.2 Definition of $\tau(\varepsilon,\delta)$

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

Lemma 23. *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 $0 < \delta < e^{-2.6}$ and $t \geq 6\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln \delta^{-1})$,*

$$\begin{aligned}\|\hat{\theta}_{t+1} - \theta^*\|_\Sigma^2 \leq & \frac{3}{t}\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1} + \frac{D_X^2}{\Lambda_{\min}} D_{\text{app}}^2 \frac{4(1+\sqrt{8\ln\delta^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035 + \ln\delta^{-1})}{0.07}\right) \\ & + \frac{12}{0.07^2 t^2}\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1}\frac{D_X^2}{\Lambda_{\min}}(1+\sqrt{8\ln\delta^{-1}})\right. \\ & \left. + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2p_1}}\right)^2 (\ln\delta^{-1})^2\right),\end{aligned}$$

with probability at least $1 - 4\delta$.

Proof. We first observe that

$$\arg\min_{w\in\mathbb{R}^d} \frac{1}{t}\sum_{s=1}^t (y_s - w^T X_s)^2 + \lambda\|w - \hat{\beta}_1\|^2 = \arg\min_{w\in\mathbb{R}^d} \frac{1}{t}\sum_{s=1}^t (y_s - \hat{\beta}_1^T X_s - w^T X_s)^2 + \lambda\|w\|^2,$$

therefore we apply ridge analysis of Hsu et al. (2012) to $(X_s, y_s - \hat{\beta}_1^T X_s)$. We note that $(y_s - \hat{\beta}_1^T X_s)$ has the same variance proxy and the same approximation error, it only amounts to translate the optimal w, that is denoted by β.

For any $\lambda > 0$, we observe that $d_{2,\lambda} \leq d_{1,\lambda} \leq d$, $\rho_\lambda \leq \frac{D_X}{\sqrt{d_{1,\lambda}\Lambda_{\min}}}$ and $b_\lambda \leq \rho_\lambda(D_{\text{app}} + D_X\|\beta - \hat{\beta}_1\|)$. Therefore we can apply Theorem 16 of Hsu et al. (2012): for $0 < \delta < e^{-2.6}$ and $t \geq 6\frac{D_X}{\sqrt{\Lambda_{\min}}}(\ln(d) + \ln\delta^{-1})$, the following holds with probability $1 - 4\delta$: $\|\hat{\beta}_{t+1,\lambda} - \beta\|_\Sigma^2 = 3(\|\beta_\lambda - \beta\|_\Sigma^2 + \varepsilon_{\text{bs}} + \varepsilon_{\text{vr}})$, with

$$\begin{aligned}\varepsilon_{\text{bs}} &\leq \frac{4}{0.07^2}\Big(\frac{\frac{D_X^2}{\Lambda_{\min}}\mathbb{E}[(\mathbb{E}[y \mid X] - \beta^T X)^2] + (1 + \frac{D_X^2}{\Lambda_{\min}})\|\beta_\lambda - \beta\|_\Sigma^2}{t}(1+\sqrt{8\ln\delta^{-1}}) \\ &\qquad + \frac{(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\beta - \hat{\beta}_1\|) + \|\beta_\lambda - \beta\|_\Sigma)^2}{t^2}(\ln\delta^{-1})^2\Big), \\ \delta_f &\leq \frac{1}{\sqrt{t}}\frac{D_X}{\sqrt{\Lambda_{\min}}}(1+\sqrt{8\ln\delta^{-1}}) + \frac{1}{t}\frac{4\sqrt{\frac{D_X^4}{\Lambda_{\min}^2 d} + 1}}{3}\ln\delta^{-1}, \\ \varepsilon_{\text{vr}} &\leq \frac{\sigma^2 d(1+\delta_f)}{0.07^2 t} + \frac{2\sigma^2\sqrt{d(1+\delta_f)\ln\delta^{-1}}}{0.07^{3/2}t} + \frac{2\sigma^2\ln\delta^{-1}}{0.07t}.\end{aligned}$$

Moreover $\mathbb{E}[(\mathbb{E}[y \mid X] - \beta^T X)^2] \leq D_{\text{app}}^2$ and $\Lambda_{\min} \leq D_X^2$, hence, using $\|\beta_\lambda - \beta\|_\Sigma \leq \lambda\|\beta - \hat{\beta}_1\|$ we transfer the result in our KF notations, that is, $\hat{\theta}_t = \hat{\beta}_{t, p_1^{-1}/2(t-1)}, \hat{\beta}_1 = \hat{\theta}_1, \beta = \theta^*$. We obtain, for any $0 < \delta < e^{-2.6}$ and $t \geq 6\frac{D_X}{\sqrt{\Lambda_{\min}}}(\ln(d) + \ln\delta^{-1})$,

$$\begin{aligned}\varepsilon_{\text{bs}} &\leq \frac{4}{0.07^2}\Big(\frac{\frac{D_X^2}{\Lambda_{\min}}D_{\text{app}}^2 + \frac{D_X^2}{\Lambda_{\min}}\frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1 t}}{t}(1+\sqrt{8\ln\delta^{-1}}) \\ &\qquad + \frac{(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2p_1 t}})^2}{t^2}(\ln\delta^{-1})^2\Big), \\ \delta_f &\leq \frac{1}{\sqrt{t}}\frac{D_X}{\sqrt{\Lambda_{\min}}}(1+\sqrt{8\ln\delta^{-1}}) + \frac{1}{t}\frac{4\sqrt{\frac{D_X^4}{\Lambda_{\min}^2 d} + 1}}{3}\ln\delta^{-1}, \\ \varepsilon_{\text{vr}} &\leq \frac{\sigma^2 d(1+\delta_f)}{0.07^2 t} + \frac{2\sigma^2\sqrt{d(1+\delta_f)\ln\delta^{-1}}}{0.07^{3/2}t} + \frac{2\sigma^2\ln\delta^{-1}}{0.07t}, \\ \|\hat{\theta}_{t+1} - \theta^*\|_\Sigma^2 &\leq 3\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1 t} + \varepsilon_{\text{bs}} + \varepsilon_{\text{vr}}\right),\end{aligned}$$

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

$$\delta_f \leq \frac{1}{\sqrt{6 \ln \delta^{-1}}}(1+\sqrt{8 \ln \delta^{-1}}) + \frac{1}{6}\frac{4}{3}\sqrt{\frac{1}{d}+1} \leq \frac{1+\sqrt{8}}{\sqrt{6}} + \frac{2\sqrt{2}}{9} \approx 1.9 \leq 2\,.$$

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

$$\begin{aligned}
\varepsilon_{\mathrm{vr}} &\leq \frac{\sigma^2}{0.07t}\left(\frac{3d}{0.07} + 2\sqrt{\frac{3d\ln\delta^{-1}}{0.07}} + 2\ln\delta^{-1}\right) \\
&\leq \frac{\sigma^2}{0.07t}\left(\frac{6d}{0.07} + 3\ln\delta^{-1}\right) \\
&\leq \frac{3\sigma^2(d/0.035 + \ln\delta^{-1})}{0.07t}\,.
\end{aligned}$$

It yields the result. □

Lemma 23 allows the definition of an explicit value for $\tau(\varepsilon, \delta)$, as displayed in the following Corollary.

Corollary 24. *Assumption 5 is satisfied for $\tau(\varepsilon,\delta) = \max(\tau_1(\delta), \tau_2(\varepsilon,\delta), \tau_3(\varepsilon,\delta))$ where we define*

$$\begin{aligned}
\tau_1(\delta) &= \max\left(12\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln\delta^{-1}), \frac{48D_X^2}{\Lambda_{\min}}\ln\frac{24D_X^2}{\Lambda_{\min}}\right), \\
\tau_2(\varepsilon,\delta) &= \frac{24\varepsilon^{-1}}{\Lambda_{\min}}\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1} + \frac{D_X^2}{\Lambda_{\min}}D_{\mathrm{app}}^2\frac{4(1+\sqrt{8\ln\delta^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035+\ln\delta^{-1})}{0.07}\right) \\
&\quad \ln\frac{12\varepsilon^{-1}}{\Lambda_{\min}}\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1} + \frac{D_X^2}{\Lambda_{\min}}D_{\mathrm{app}}^2\frac{4(1+\sqrt{8\ln\delta^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035+\ln\delta^{-1})}{0.07}\right), \\
\tau_3(\varepsilon,\delta) &= \sqrt{\frac{96\varepsilon^{-1}}{0.07^2\Lambda_{\min}}}\Bigg(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1}\frac{D_X^2}{\Lambda_{\min}}(1+\sqrt{8\ln\delta^{-1}}) \\
&\qquad + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\mathrm{app}} + D_X\|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2p_1}}\right)^2(\ln\delta^{-1})^2\Bigg)^{1/2} \\
&\quad \ln\frac{96\varepsilon^{-1}}{0.07^2\Lambda_{\min}}\Bigg(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1}(1+\frac{D_X^2}{\Lambda_{\min}})(1+\sqrt{8\ln\delta^{-1}}) \\
&\qquad + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\mathrm{app}} + D_X\|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2p_1}}\right)^2(\ln\delta^{-1})^2\Bigg)\,.
\end{aligned}$$

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 24. We define $\delta_t = \delta/t^2$ for any $t \geq 1$. In order to apply Lemma 23 with a union bound, we need $t \geq 6\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln\delta_t^{-1})$. If $t \geq 12\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln\delta^{-1})$ and $t \geq \frac{48D_X^2}{\Lambda_{\min}}\ln\frac{24D_X^2}{\Lambda_{\min}}$, we obtain

$$\begin{aligned}
t &\geq \frac{t}{2} + \frac{\sqrt{t}}{2}\sqrt{t} \\
&\geq 6\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln\delta^{-1}) + \frac{12D_X^2}{\Lambda_{\min}}\ln t, \qquad \text{as } \ln t \leq \sqrt{t} \\
&= 6\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln\delta_t^{-1})\,.
\end{aligned}$$

Therefore, we define $\tau_1(\delta) = \max\left(12\frac{D_X^2}{\Lambda_{\min}}(\ln d + \ln \delta^{-1}), \frac{48D_X^2}{\Lambda_{\min}} \ln \frac{24D_X^2}{\Lambda_{\min}}\right)$, and we apply Lemma 23. We get the simultaneous property

$$\begin{aligned}\|\hat{\theta}_{t+1} - \theta^*\|_\Sigma^2 \leq & \frac{3}{t}\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1} + \frac{D_X^2}{\Lambda_{\min}} D_{\text{app}}^2 \frac{4(1+\sqrt{8\ln \delta_t^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035 + \ln \delta_t^{-1})}{0.07}\right) \\ & + \frac{12}{0.07^2 t^2}\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1}\frac{D_X^2}{\Lambda_{\min}}(1+\sqrt{8\ln \delta_t^{-1}})\right. \\ & \qquad \left. + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2p_1}}\right)^2 (\ln \delta_t^{-1})^2\right), \qquad t \geq \tau_1(\delta),\end{aligned}$$

with probability at least $1 - 4\delta \sum_{t \geq \tau_1(\delta)} t^{-2} \geq 1 - \delta$ because $\tau_1(\delta) > 4$.

Thus, as $\ln t \geq 1$ for $t \geq \tau_1(\delta)$ and $\|\hat{\theta}_{t+1} - \theta^*\|_\Sigma^2 \geq \Lambda_{\min}\|\hat{\theta}_{t+1} - \theta^*\|^2$, we obtain

$$\begin{aligned}\|\hat{\theta}_{t+1} - \theta^*\| \leq & \frac{6\ln t}{\Lambda_{\min} t}\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{2p_1} + \frac{D_X^2}{\Lambda_{\min}} D_{\text{app}}^2 \frac{4(1+\sqrt{8\ln \delta^{-1}})}{0.07^2} + \frac{3\sigma^2(d/0.035 + \ln \delta^{-1})}{0.07}\right) \\ & + \frac{48(\ln t)^2}{0.07^2 \Lambda_{\min} t^2}\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1}\frac{D_X^2}{\Lambda_{\min}}(1+\sqrt{8\ln \delta^{-1}})\right. \\ & \qquad \left. + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{2p_1}}\right)^2 (\ln \delta^{-1})^2\right), \qquad t \geq \tau_1(\delta),\end{aligned}$$

with probability at least $1 - \delta$. Finally, both terms of the last inequality are bounded by $\varepsilon/2$. □

From Corollary 24, we obtain the asymptotic rate by comparing $\tau_2(\delta)$ and $\tau_3(\delta)$. We write $\tau_2(\delta) = 2A_2(\delta)\ln A_2(\delta), \tau_3(\delta) = 2A_3(\delta)\ln A_3(\delta)$ with

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

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

$$\begin{aligned}A_3(\delta) &\lesssim \sqrt{\frac{\varepsilon^{-1}}{\Lambda_{\min}}}\left(\sqrt{\frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1}\frac{D_X^2}{\Lambda_{\min}}\sqrt{\ln \delta^{-1}}} + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{p_1}}\right)\ln \delta^{-1}\right) \\ &\lesssim \sqrt{\frac{\varepsilon^{-1}}{\Lambda_{\min}}}\left(\frac{\|\hat{\theta}_1 - \theta^*\|^2}{p_1} + \frac{D_X^2}{\Lambda_{\min}}\sqrt{\ln \delta^{-1}} + \left(\frac{D_X}{\sqrt{\Lambda_{\min}}}(D_{\text{app}} + D_X\|\theta^*\|) + \frac{\|\hat{\theta}_1 - \theta^*\|}{\sqrt{p_1}}\right)\ln \delta^{-1}\right).\end{aligned}$$

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

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