SUPPLEMENTARY MATERIAL

A Calibration of the parameters

We only prove the calibration in the setting of linear regression with square loss (i.e., for Theorem 3 only and not for the general Theorem 1). It remains an open question whether the calibration of the parameters can be performed in the general setting of Section 3. We leave this question for future research. Furthermore, for the sake of clarity the adaption to Y (which is only necessary for clipping) is not considered here. However, it can be achieved simultaneously by updating the clipping range based on the past observations Y_s, $s \leqslant t-1$ (see Gerchinovitz, 2013, Section 4.5).

The calibration algorithm (Algorithm 3) works as follows. We define large enough grids of parameters for each doubling session $j \geqslant 0$

$$\begin{aligned}\mathcal{G}_j = \Big\{ (d_0, & \alpha, U, B) \in [1, \dots, d] \times \mathbb{R}_+^3 \quad \text{such that} \\ & d_0 \in \{0\} \cup \big\{ 2^k, k = 0, \dots, \lceil \log_2 d \rceil \big\} \\ & \alpha \in \big\{ 2^k, k = -2j + \lceil \log_2 (B d_0 / Y^2) \rceil, \dots, \\ & \qquad\qquad\qquad\qquad j + \lceil \log_2 d_0 \rceil \big\} \\ & U \in \big\{ 2^k, k = -2j, \dots, 2j + \lceil 2 \log_2 Y \rceil \big\} \\ & B \in \big\{ 2^k, k = -2j, \dots, 2j + \lceil 2 \log_2 Y \rceil \big\} \Big\} . \end{aligned} \tag{7}$$

For each set of parameters $p = (d_0, \alpha, U, B) \in \mathcal{G}_j$, we perform a local version of SAEW to obtain an estimator $\tilde{\theta}_{p,j}$ at time $t = 2^j - 1$. Then, the calibration algorithm uses the online aggregation procedure BOA of Wintenberger (2014) to make predictions from $t = 2^j$ to $2^{j+1} - 1$. Its predictions are based on online combinations of the (clipped) forecasts made by the $\tilde{\theta}_{p,j}$.

Theorem 4. *Let $Y > \max_{t=1,\dots,T} |Y_t|$ almost surely. With probability $1 - \delta$, the excess risk of the estimator $\tilde{f}_T$ produced by Algorithm 3 is of order*

$$\mathcal{O}_T \bigg(\frac{Y^2}{T} \log \Big(\frac{(\log d)(\log T + \log Y)}{\delta} \Big) + \frac{d_0 X^2 \sigma^2}{\alpha^* T} \log \Big(\frac{d \log T}{\delta} \Big) \bigg) ,$$

where $d_0 = \|\theta^*\|_0$ and $\alpha^* > 0$ is the largest value of α satisfying Inequality (SC).

The proof is postponed to Appendix B.7.

```
Algorithm 3: Calibration algorithm
Parameters: $Y > 0$, $\delta > 0$
Initialization: $t_0 = t = 1$ and $\bar{\theta}^{(0)} = 0$
For each $j = 0, 1, \dots$
  • Define the grid $\mathcal{G}_j$ as in (7)
  • For parameters $p = (d_0, \alpha, U, B) \in \mathcal{G}_j$:
      – Define $\delta_j = \delta / (2(j+1)^2)$
      – Run SAEW with parameter $(d_0, \alpha, U, B, \delta_j)$
        for $t = 0, \dots, 2^j - 1$ and get the estimator
        $\tilde{\theta}_{2^j - 1}$, denoted by $\tilde{\theta}_{p,j}$.
      – Define the clipped predictor
            $f_{p,j} : x \mapsto [x^\top \tilde{\theta}_{p,j}]_Y$
        where $[\,\cdot\,]_Y := \max\big\{ -Y, \min\{\,\cdot\,, Y\} \big\}$.
  • For $t = 2^j, \dots, 2^{j+1} - 1$,
      – predict $\widehat{f}_{t-1}(X_t)$ by performing BOA with
        experts $(f_{p,j})_{p \in \mathcal{G}_j}$
      – output the estimator $\tilde{f}_{t-1} = \bar{f}_j$
  • Define the average estimator
            $\bar{f}_{j+1} = 2^{-j} \sum_{t=2^j}^{2^{j+1}-1} \widehat{f}_{t-1}$ .
```

Remark A.1. Similarly to the restricted eigenvalue condition of the Lasso, we believe that the strong convexity condition for α^* might be necessary on subspaces of dimension lower than or equal to d_0 only. However, to do so, SAEW should be used with a subroutine that produces sparse $\widehat{\theta}_{t-1}$. Up to our knowledge, such procedures do not exist for convex optimization in the ℓ_1-ball. As stated previously, sparse procedures such as RDA of Xiao (2010) cannot be used as subroutines since they perform optimization in the ℓ_2-ball and suffer a linear dependence on d. We leave this question for future work.

Remark A.2. For the sake of clarity, the above result is only stated asymptotically. However the bound also holds in finite time up to universal multiplicative constant (as done in the proof). Additional negligible terms of order $\mathcal{O}(1/T^2)$ then appear in the bound. Furthermore, the finite time bound also achieves the best of the two regimes (slow rate vs fast rate) as in Theorem 3.

Remark A.3. Theorem 4 has been proven only for square linear regression. However, it also holds for any strongly-convex loss function, with locally bounded gradients (i.e., with LIST condition, see Wintenberger, 2014).

Remark A.4. To perform the calibration, we left the original framework of Section 2. First, because of the clipping, the estimators $\tilde{f}_{t-1}$ produced by Algorithm 3 are not linear any-more. Second, the meta-algorithm implies that we can observe the gradients of all subroutines SAEW simultaneously. Tuning the parameters in the original setting is left for future work.

B Proofs

B.1 Lemma 5

We first state Lemma 5, a classical result in strong convexity, as it will be useful in the proofs. It relates the ℓ_2-error of an estimator with its excess risk when the risk is strongly convex.

Lemma 5. *If the risk is 2α-strongly convex, then*

$$\|\theta - \theta^*\|_2^2 \leqslant \alpha^{-1} \operatorname{Risk}(\theta)$$

for all $\theta \in \mathbb{R}^d$.

Proof. Let $\theta \in \mathbb{R}^d$, by (SC) applied with $\theta_1 = \theta^*$ and $\theta_2 = \theta$, we get

$$\big\|\theta - \theta^*\big\|_2^2 \leqslant \alpha^{-1}\mathbb{E}\big[\ell_t(\theta) - \ell_t(\theta^*)\big] + \alpha^{-1}\mathbb{E}\big[\nabla\ell_t(\theta^*)\big]^\top(\theta^* - \theta)\,.$$

But, $\mathbb{E}\big[\nabla\ell_t(\theta^*)\big]^\top(\theta^* - \theta) \leqslant 0$. Otherwise, taking into account the convexity of the domain, the direction $\theta - \theta^*$ is a decreasing feasible direction, which contradicts the optimality of θ^*. □

B.2 Proof of Theorem 1

Let (δ_i) be a non-increasing sequence in $(0,1)$ such that $\sum_{i=1}^\infty \delta_i \leqslant \delta$.

Step 1. Proof by induction that the subroutines always perform the optimization in the correct ℓ_1-ball. We prove by induction on $i \geqslant 0$ that with probability at least $1 - \sum_{j=1}^i \delta_j$

$$\big\|\theta^* - [\bar\theta_{t_i-1}]_{d_0}\big\|_1 \leqslant U2^{-i/2}\,. \qquad (\mathcal{H}_i)$$

$\mathcal{H}_0$ is satisfied by assumption since $\|\theta^*\|_1 \leqslant U$ and $[\bar\theta_{t_0-1}]_{d_0} = [\bar\theta_0]_{d_0} = 0$ (see SAEW for the definition of $[\bar\theta_0]$).

Let $i \geqslant 0$ and assume $(\mathcal{H}_i)$. The following Lemma (whose proof is postponed to Appendix B.3) states that the gradients are indeed upper-bounded by B in sup-norm.

Lemma 6. *Let $i \geqslant 0$. Under $(\mathcal{H}_i)$, for all $t \in [t_i, t_{i+1}-1]$, $\|\nabla\ell_t(\widehat\theta_{t-1})\|_\infty \leqslant B$ almost surely.*

Therefore, from the regret bound (5), the subroutine $\mathcal{S}_i$ satisfies for all $t \in [t_i, t_{i+1}-1]$

$$\sum_{s=t_i}^t \ell_s(\widehat\theta_{s-1}) - \ell_s(\theta^*) \leqslant U2^{-i/2}\Bigg(a\sqrt{\sum_{s=t_i}^t \big\|\nabla\ell_s(\widehat\theta_{s-1})\big\|_\infty^2} + bB\Bigg)\,.$$

Bounding the cumulative risk with the regret thanks to Theorem 10 in Appendix C.2, it yields with probability at least $1 - \sum_{j=1}^{i+1}\delta_j$,

$$\sum_{s=t_i}^t \mathbb{E}[\ell_s](\widehat\theta_{s-1}) - \mathbb{E}[\ell_s](\theta^*) \leqslant U2^{-i/2}\operatorname{Err}_t \qquad (8)$$

where $\operatorname{Err}_t := a_i'\sqrt{\sum_{t_i}^t \big\|\nabla\ell_s(\widehat\theta_{s-1})\big\|_\infty^2} + b_i'B$ with

$$a_i' := a + \sqrt{2}\sqrt{\log\Big(1 + \frac{1}{2}\log\Big(\frac{t-t_i+1}{2}\Big)\Big) - \log\delta_{i+1}}\,,$$

and

$$(9)$$

$$b_i' := b + \frac{1}{2} + \log\Big(1 + \frac{1}{2}\log\Big(\frac{t-t_i+1}{2}\Big)\Big) - \log\delta_{i+1}\,. \qquad (10)$$

Thus, recalling that by definition (see SAEW)

$$\bar\theta_t := (t - t_i + 1)^{-1}\sum_{s=t_i}^t \widehat\theta_{s-1}\,,$$

and because the losses are i.i.d., Jensen's inequality yields

$$\begin{aligned}\operatorname{Risk}(\bar\theta_t) &= \mathbb{E}[\ell_{t+1}](\bar\theta_t) - \mathbb{E}[\ell_{t+1}](\theta^*)\\ &\overset{\text{Jensen}}{\leqslant} (t-t_i+1)^{-1}\sum_{s=t_i}^t \mathbb{E}[\ell_s](\widehat\theta_{s-1}) - \mathbb{E}[\ell_s](\theta^*)\\ &\overset{(8)}{\leqslant} \frac{U\operatorname{Err}_t}{2^{i/2}(t-t_i+1)}\,. \qquad (11)\end{aligned}$$

Together with the strong convexity of the risk (Lemma 5), this entails

$$\big\|\bar\theta_t - \theta^*\big\|_2^2 \leqslant \frac{U\operatorname{Err}_t}{\alpha 2^{i/2}(t-t_i+1)}\,. \qquad (12)$$

We thus control the ℓ_2-error of $\bar\theta_t$. However, in order to control the ℓ_1-error without paying a factor d, we need to truncate some coordinates of $\bar\theta_t$. By definition of $[\bar\theta_t]_{d_0}$ (see SAEW), we have

$$[\bar\theta_t]_{d_0} \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d:\|\theta\|_0\leqslant d_0}\big\{\big\|\bar\theta_t - \theta\big\|_2\big\}\,. \qquad (13)$$

Now, (13) together with $\|\theta^*\|_0 \leqslant d_0$ (by assumption) yields

$$\big\|\bar\theta_t - [\bar\theta_t]_{d_0}\big\|_2 \leqslant \big\|\bar\theta_t - \theta^*\big\|_2\,. \qquad (14)$$

Furthermore, because both $\|\theta^*\|_0 \leqslant d_0$ and $\big\|[\bar{\theta}_t]_{d_0}\|_0 \leqslant d_0$, we have

$$\|[\bar{\theta}_t]_{d_0} - \theta^*\|_0 \leqslant 2d_0 \,. \tag{15}$$

Therefore, with probability at least $1 - \sum_{j=0}^{i+1} \delta_j$

$$\begin{aligned} \big\|[\bar{\theta}_t]_{d_0} - \theta^*\big\|_1 &\overset{(15)}{\leqslant} \sqrt{2d_0}\,\big\|[\bar{\theta}_t]_{d_0} - \theta^*\big\|_2 \\ &\leqslant \sqrt{2d_0}\Big(\big\|[\bar{\theta}_t]_{d_0} - \bar{\theta}_t\big\|_2 + \big\|\bar{\theta}_t - \theta^*\big\|_2\Big) \\ &\overset{(14)}{\leqslant} 2\sqrt{2d_0}\big\|\bar{\theta}_t - \theta^*\big\|_2 \\ &\overset{(12)}{\leqslant} 2\sqrt{2d_0\alpha^{-1}U\,\mathrm{Err}_t\,2^{-i/2}(t-t_i+1)^{-1}} \\ &=: \varepsilon_t \,, \end{aligned} \tag{16}$$

where the last equality holds by definition of ε_t (see SAEW). Finally, $(\mathcal{H}_{i+1})$ is fulfilled by definition of t_{i+1} (see SAEW), which satisfies $\varepsilon_{t_{i+1}-1} \leqslant U2^{-(i+1)/2}$. The induction is thus completed.

In the rest of the proof, we consider that $(\mathcal{H}_i)$ are satisfied for all $i \geqslant 0$. This occurs with probability $1 - \sum_{j=1}^{\infty} \delta_j \geqslant 1 - \delta$ as stated by Step 1.

Step 2. Fast rate for the excess risk of $\tilde{\theta}_t$. First, we prove that the excess risk of $\tilde{\theta}_t$ is upper-bounded as

$$\mathrm{Risk}(\tilde{\theta}_t) \leqslant \frac{d_0 B^2}{\alpha}\left(\frac{2^7 a'^2}{t} + \frac{2^{11} b'^2}{t^2}\right) + \frac{2\alpha U^2}{d_0 t^2}\,, \tag{17}$$

for all $t \geqslant 1$, where $a' = a'_{\lfloor 2\log_2 t\rfloor}$ and $b' = b'_{\lfloor 2\log_2 t\rfloor}$.

To do so, we start from the risk inequality (11). From the definition of ε_t (see (16)), we get

$$\mathrm{Risk}(\bar{\theta}_t) \leqslant \frac{\alpha\varepsilon_t^2}{8d_0}\,, \qquad t \geqslant 1\,. \tag{18}$$

Thus, by definition of $\tilde{\theta}_t := \bar{\theta}_{\arg\min_{s\leqslant t}\varepsilon_s}$, we have

$$\mathrm{Risk}(\tilde{\theta}_t) \leqslant \frac{\alpha \min_{s\leqslant t}\varepsilon_s^2}{8d_0} \tag{19}$$

We conclude the proof of (17) with the following lemma proved in Appendix B.4

Lemma 7. *Let $i \geqslant 0$. Let $t_i - 1 \leqslant t \leqslant t_{i+1}$, then*

$$\min_{s\leqslant t}\varepsilon_s \leqslant U\left(\frac{\sqrt{2}\gamma a_i'}{\sqrt{t}} + \frac{2+4\gamma b_i'}{t}\right)\,,$$

where $\gamma := 2^4 d_0 B/(\alpha U)$.

Let $i \geqslant 0$ such that $t_i - 1 \leqslant t \leqslant t_{i+1}$. Lemma 7 together with (19) and $(x+y)^2 \leqslant 2x^2 + 2y^2$ for $x, y \geqslant 0$, yields

$$\mathrm{Risk}(\tilde{\theta}_t) \leqslant \frac{\alpha U^2\gamma^2}{8d_0}\Big(\frac{\sqrt{2}a_i'}{\sqrt{t}} + \frac{2\gamma^{-1} + 4b_i'}{t}\Big)^2 \tag{20}$$

$$\leqslant \frac{\alpha U^2\gamma^2}{d_0}\left(\frac{a_i'^2}{2t} + \frac{2\gamma^{-2} + 8b_i'^2}{t^2}\right)\,. \tag{21}$$

Now, remark that if $i \geqslant 2\log t$, then $\varepsilon_{t_i-1} \leqslant U2^{-i} \leqslant U/t$ and from (19), $\mathrm{Risk}(\tilde{\theta}_t) \leqslant \alpha U^2/(8d_0t^2)$. Together, with (21), we get

$$\mathrm{Risk}(\tilde{\theta}_t) \leqslant \frac{\alpha U^2\gamma^2}{d_0}\left(\frac{a'^2}{2t} + \frac{2\gamma^{-2} + 8b'^2}{t^2}\right)\,,$$

with $a' = a'_{\lfloor 2\log_2 t\rfloor}$ and $b' = b'_{\lfloor 2\log_2 t\rfloor}$. Substituting $\gamma = 2^4 d_0 B/(\alpha U)$ concludes the proof of Inequality (17).

Step 3. Slow rate for the excess risk of $\tilde{\theta}_t$. Now, we prove that

$$\mathrm{Risk}(\tilde{\theta}_t) \leqslant UB\left(\frac{a'}{\sqrt{t/2}} + \frac{4b'}{t}\right) + \frac{\alpha U^2}{8d_0 t},\ t \geqslant 1\,. \tag{22}$$

For small values of t, the slow rate will be satisfied from the initial bound of the subroutine during the first session. At some time $\tau > 0$, the fast rate becomes better than the slow rate. This splitting time is defined as the solution of the equality

$$\frac{\mathrm{Err}_{t_1-1}}{t_1-1} = B\Big(\frac{\sqrt{2}a'}{\sqrt{\tau}} + \frac{2\gamma^{-1}+4b'}{\tau}\Big)\,. \tag{23}$$

Let $t \geqslant 1$. To control $\mathrm{Risk}(\tilde{\theta}_t)$, we distinguish three cases:

- if $t \leqslant t_1 - 1$, then, since by definition of ε_s

  $$\arg\min_{s\leqslant t}\frac{\mathrm{Err}_s}{s} = \arg\min_{s\leqslant t}\varepsilon_s\,,$$

  we get from Inequality (11) that

  $$\begin{aligned}\mathrm{Risk}(\tilde{\theta}_t) &= \mathrm{Risk}(\bar{\theta}_{\arg\min_{s\leqslant t}\varepsilon_s}) \\ &\leqslant U2^{-0/2}\min_{s\leqslant t}\frac{\mathrm{Err}_s}{s} \\ &\leqslant U\frac{\mathrm{Err}_t}{t}\,.\end{aligned}$$

  By definition of Err_t (see (8)) and upper-bounding the gradients by B, we get

  $$\mathrm{Risk}(\tilde{\theta}_t) \leqslant UB\Big(\frac{a_0'}{\sqrt{t}} + \frac{b_0'}{t}\Big)\,.$$

- if $t_1 \leqslant t \leqslant \tau$, then following the same reasoning as above, we have

  $$\mathrm{Risk}(\tilde{\theta}_t) \leqslant U\frac{\mathrm{Err}_{t_1-1}}{t_1-1}\,,$$

which yields by definition of τ (see Equality (23)) and by using $t \leqslant \tau$:

$$\begin{aligned}\mathrm{Risk}(\tilde{\theta}_t) &\leqslant UB\Big(\frac{\sqrt{2}a'}{\sqrt{\tau}} + \frac{2\gamma^{-1}+4b'}{\tau}\Big)\\ &\leqslant UB\Big(\frac{\sqrt{2}a'}{\sqrt{t}} + \frac{2\gamma^{-1}+4b'}{t}\Big).\end{aligned}$$

- if $\tau \leqslant t$, since by definition of t_1 (see SAEW), $\varepsilon_{t_1-1} \leqslant U/2$, then by definition of ε_{t_1-1} (see (16)),

  $$2\sqrt{2d_0\alpha^{-1}U\frac{\mathrm{Err}_{t_1-1}}{t_1-1}} \leqslant \frac{U}{2}\,,$$

  and thus taking the square and rearranging the terms

  $$\frac{d_0}{\alpha} \leqslant \frac{U}{2^5}\Big(\frac{t_1-1}{\mathrm{Err}_{t_1-1}}\Big)\,.$$

  Using the definition of $\gamma = 2^4 d_0 B/(\alpha U)$ and substituting Err_{t_1-1} with Equality (23), this yields

  $$\frac{\alpha U^2\gamma^2}{8d_0} = \frac{2^5 d_0 B^2}{\alpha} \leqslant UB\Big(\frac{\sqrt{2}a'}{\sqrt{\tau}} + \frac{2\gamma^{-1}+4b'}{\tau}\Big)^{-1}.$$

  Finally from Inequality (20), and using $\tau \leqslant t$

  $$\mathrm{Risk}(\tilde{\theta}_t) \leqslant UB\left(\frac{\sqrt{2}a'}{\sqrt{t}} + \frac{2\gamma^{-1}+4b'}{t}\right).$$

Combining the three cases together and substituting $\gamma = 2^4 d_0 B/(\alpha U)$, concludes the proof of Inequality (22).

Step 4. Conclusion of the proof Combining Inequalities (17) and (22), we get the risk inequality stated in the Theorem 1 for $\tilde{\theta}_t$. It only remains to choose $\delta_j = \delta/(j+1)^2$ so that $\sum_{j=1}^{\infty}\delta_j \leqslant \delta$ and to control $a' = a'_{\lfloor 2\log_2 t\rfloor}$ and $b' = b'_{\lfloor 2\log_2 t\rfloor}$. From (9), we can use $\delta_{\lfloor 2\log_2 t\rfloor+1} \geqslant \delta/(1+2\log_2 t)^2$ and $T_i \leqslant t$. Simple calculation yields that $a'-a$ is lower than

$$\begin{aligned}&\sqrt{2\big(\log(1+1/2\log(t/2)) - \log\delta + 2\log(1+2\log_2 t)\big)}\\ &\leqslant \sqrt{6\log(1+3\log t) - 2\log\delta}.\end{aligned}$$

Similarly, for $b'-b$. It is upper-bounded by

$$\begin{aligned}&\frac{1}{2} + \log\big(1+(1/2)\log(t/2)\big) - \log\delta + 2\log(1+2\log_2 t)\\ &\leqslant 1/2 + 3\log(1+3\log t) - \log\delta\,.\end{aligned}$$

This concludes the proof.

B.3 Proof of Lemma 6

Since by assumption $B \geqslant \max_{\theta:\|\theta\|_1\leqslant 2U}\|\nabla\ell_t(\theta)\|_\infty$ a.s. it suffices to show that $\|\widehat{\theta}_{t-1}\|_1 \leqslant 2U$. By definition of the session $\mathcal{S}_i$,

$$\widehat{\theta}_{t-1} \in \mathcal{B}_1([\bar{\theta}_{t_i-1}]_{d_0}, U2^{-i/2})\,.$$

Thus:

- if $i=0$, since $[\bar{\theta}_0]_{d_0} = 0$, $\|\widehat{\theta}_{t-1}\|_1 \leqslant U$.
- if $i=1$, since $\|[\bar{\theta}_{t_1-1}]\|_1 \leqslant U$ as a truncated average of vectors in $\mathcal{B}_1(0,U)$, we have

  $$\begin{aligned}\|\widehat{\theta}_{t-1}\|_1 &\leqslant \|\widehat{\theta}_{t-1} - [\bar{\theta}_{t_1-1}]_{d_0}\|_1 + \|[\bar{\theta}_{t_1-1}]_{d_0}\|_1\\ &\leqslant U/\sqrt{2} + U \leqslant 2U;\end{aligned}$$

- otherwise, $i \geqslant 2$ and $\|\widehat{\theta}_{t-1}\|_1$ is bounded by

  $$\begin{aligned}&\|\widehat{\theta}_{t-1} - [\bar{\theta}_{t_i-1}]_{d_0}\|_1 + \|[\bar{\theta}_{t_i-1}]_{d_0} - \theta^*\|_1 + \|\theta^*\|_1\\ &\overset{(\mathcal{H}_i)}{\leqslant} U2^{-i/2} + U2^{-i/2} + U \leqslant 2U\,.\end{aligned}$$

Putting the tree cases together, $\|\widehat{\theta}_{t-1}\|_1 \leqslant 2U$, which concludes the proof.

B.4 Proof of Lemma 7

It is enough to control $\varepsilon_{t_i-1} \geqslant \min_{s\leqslant t}\varepsilon_s$. To do so, we prove that for every $j \geqslant 0$, $T_j := t_{j+1} - t_j$ cannot be too large, so that at time t, i will be at least of order $\log_2 t$.

Let $j \geqslant 0$. We can assume $t_{j+1} > t_j$, otherwise $T_j = 0$. Thus, from the bound on the gradients (Lemma 6) and from the definition of Err_t (see (8)) for all $t \in [t_j+1, t_{j+1}]$,

$$\mathrm{Err}_{t-1} \leqslant B(a'_j\sqrt{t-t_j} + b'_j)\,, \tag{24}$$

and from the definition of ε_{t-1} (see (16))

$$\varepsilon_{t-1} \leqslant 2\sqrt{2d_0\alpha^{-1}UB\frac{a'_j\sqrt{t-t_j}+b'_j}{2^{j/2}(t-t_j)}}\,.$$

By definition, t_{j+1} is the smallest integer after t_j satisfying $\varepsilon_{t_{j+1}-1} \leqslant U2^{-(j+1)/2}$. Hence, we have $\varepsilon_{t_{j+1}-2} \geqslant U2^{-(j+1)/2}$, which implies

$$\begin{aligned}&2\sqrt{2d_0\alpha^{-1}UB\frac{a'_j\sqrt{T_j-1}+b'_j}{2^{j/2}(T_j-1)}} \geqslant U2^{-(j+1)/2}\\ &\Leftrightarrow\ 2^{j/2}\underbrace{2^4 d_0\alpha^{-1}U^{-1}B}_{:=\gamma}\big(a'_j\sqrt{T_j-1}+b'_j\big) \geqslant T_j - 1\end{aligned}$$

Then, by solving a second order equation in $\sqrt{T_j-1}$ (see for instance Gaillard et al., 2014, Lemma 10), the above inequality entails

$$T_j \leqslant 1 + 2^j\,\gamma^2 a_j'^2 + 2^{j/2}\gamma b'_j\,. \tag{25}$$

Therefore, summing over $j = 0, \dots, i$

$$\begin{aligned} t_{i+1} &= \cancel{t_0} + \sum_{j=0}^{i} T_j \\ &\leqslant \sum_{j=0}^{i} \big(1 + 2^j \gamma^2 a_j'^2 + 2^{j/2}\gamma b_j\big) \\ &\leqslant 2^{1+i} \gamma^2 a_i'^2 + (1+\sqrt{2}) 2^{(i+1)/2} \gamma b_i' + i + 1 \\ &\leqslant 2^{1+i} \gamma^2 a_i'^2 + 2^{(i+1)/2}\sqrt{2}\big(2\gamma b_i' + 1\big)\,, \end{aligned}$$

where the last inequality is because $2^{(i+1)/2} \geqslant \sqrt{2}(i+1)$ for $i \geqslant 0$. Solving the second-order inequality in $2^{(i+1)/2}$ we get

$$2^{-(i+1)/2} \leqslant \frac{\gamma a_i'}{\sqrt{t_{i+1}}} + \sqrt{2}\frac{1 + 2\gamma b_i'}{t_{i+1}}\,.$$

Thus, since $\varepsilon_{t_i - 1} \leqslant U 2^{-i/2}$, we have

$$\varepsilon_{t_i - 1} \leqslant U\gamma \left(\frac{\sqrt{2} a_i'}{\sqrt{t_{i+1}}} + \frac{2\gamma^{-1} + 4 b_i'}{t_{i+1}} \right)\,.$$

The proof of Lemma 7 finally follows using $t \leqslant t_{i+1}$.

B.5 Proof of Theorem 2

With probability $1 - \delta$, all inequalities provided in the proof of Theorem 1 are satisfied. We also consider the notation of the previous proof. Let $t \geqslant 1$.

Step 1. Slow rate We remark that for any $i \geqslant 0$,

$$\begin{aligned} \sum_{s=t_i}^{(t_{i+1}-1)\wedge t} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) &\overset{(8)}{\leqslant} U 2^{-i/2}\, \mathrm{Err}_{(t_{i+1}-1)\wedge t} \\ &\overset{(24)}{\leqslant} UB2^{-i/2}(a_i'\sqrt{t} + b_i') \qquad (26) \end{aligned}$$

where, in the last inequality, we use that $(t_{i+1} - 1) \wedge t \leqslant t$ and $t_i \geqslant 1$. We will use this inequality for $i \leqslant \lfloor 2 \log t \rfloor$. For $i > \lfloor 2 \log t \rfloor$, we use the fact that the gradients are bounded by B, so that by convexity of the risk

$$\begin{aligned} &\sum_{s=t_i}^{(t_{i+1}-1)\wedge t} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) \\ &\quad \leqslant \sum_{s=t_i}^{(t_{i+1}-1)\wedge t} \big\|\mathbb{E}[\nabla \ell_s](\widehat{\theta}_{s-1})\big\|_\infty \|\widehat{\theta}_{s-1} - \theta^*\|_1 \\ &\quad \leqslant UB2^{-i/2} t\,. \qquad (27) \end{aligned}$$

Summing (26) over $i = 0, \dots, \lfloor 2\log_2 t \rfloor$ and (27) over $i = \lceil 2 \log_2 t \rceil, \dots, \infty$, we get

$$\begin{aligned} \mathrm{Risk}_{1:t}(\widehat{\theta}_{0:(t-1)}) &:= \sum_{s=1}^{t} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) \\ &\leqslant UB \sum_{i=0}^{\lfloor 2\log_2 t \rfloor} 2^{-i/2}(a_i'\sqrt{t} + b_i') \\ &\qquad + UBt \sum_{i=\lceil 2\log_2 t \rceil}^{\infty} 2^{-i/2}\,. \qquad (28) \end{aligned}$$

The second sum is controlled as

$$\sum_{i=\lceil 2\log_2 t \rceil}^{\infty} 2^{-i/2} \leqslant t^{-1} \sum_{i=0}^{\infty} 2^{-i/2}\,.$$

Thus, since $\sum_{i=0}^{\infty} 2^{-i/2} = 2 + \sqrt{2} \leqslant 4$, we have

$$\mathrm{Risk}_{1:t}(\widehat{\theta}_{0:(t-1)}) \leqslant 4UB(a'\sqrt{t} + b') + 4UB\,,$$

where we recall that $a' = a'_{\lfloor 2\log_2 t \rfloor}$ and $b' = b'_{\lfloor 2\log_2 t \rfloor}$. This concludes Step 1.

Step 2. Fast rate Let us now prove the fast rate

$$\begin{aligned} \mathrm{Risk}_{1:t}\left(\widehat{\theta}_{0:(t-1)}\right) &\leqslant \frac{2^5 d_0 B^2}{\alpha} a'^2 \log_2 t \\ &\qquad + 4BU(1 + b') + U^2 \frac{\alpha}{8 d_0}\,, \end{aligned}$$

for all $t \geqslant 1$.

First, we remark that similarly to (18), we get for all $i \geqslant 0$ that

$$\begin{aligned} \sum_{s=t_i}^{t_{i+1}-1} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) &\overset{(8)}{\leqslant} U \frac{\mathrm{Err}_{t_{i+1}-1}}{2^{-i/2} T_i} T_i \\ &\overset{(16)}{\leqslant} \frac{\alpha \varepsilon^2_{t_{i+1}-1}}{8 d_0} T_i \\ &\leqslant \frac{\alpha U^2 2^{-i}}{16 d_0} T_i \qquad (29) \end{aligned}$$

where the last inequality is because $\varepsilon_{t_{i+1}-1} \leqslant U 2^{-(i+1)/2}$ by definition of t_{i+1} (see SAEW). We will use this inequality for $i \leqslant \lfloor 2 \log t \rfloor$. Summing (29) over $i = 0, \dots, \lfloor 2\log_2 t \rfloor$ and (27) over $i = \lceil 2\log_2 t \rceil, \dots, \infty$, we get

$$\begin{aligned} \mathrm{Risk}_{1:t}(\widehat{\theta}_{0:(t-1)}) &:= \sum_{s=1}^{t} \mathbb{E}[\ell_s](\widehat{\theta}_{s-1}) - \mathbb{E}[\ell_s](\theta^*) \\ &\leqslant \frac{U^2 \alpha}{2^4 d_0} \sum_{i=0}^{\lfloor 2\log_2 t \rfloor} 2^{-i} T_i \\ &\qquad + UBt \sum_{i=\lceil 2\log_2 t \rceil}^{\infty} 2^{-i/2}\,. \qquad (30) \end{aligned}$$

We upper bound both sums. The second one is controlled as we did for (28). The first one is upper-bounded thanks to (25)

$$\sum_{i=0}^{\lfloor 2\log_2 t \rfloor} 2^{-i} T_i \leqslant \sum_{i=0}^{\lfloor 2\log_2 t \rfloor} \left(\gamma^2 a_i'^2 + 2^{-i/2}\gamma b_i' + 2^{-i}\right)$$

$$\leqslant 2\gamma^2 a'^2 \log_2 t + 4\gamma b' + 2\,.$$

Therefore, substituting the two sums into (30), the cumulative risk $\mathrm{Risk}_{1:t}(\widehat{\theta}_{0:(t-1)})$ is upper-bounded by

$$\frac{U^2\alpha}{2^4 d_0}\Big(2\gamma^2 a'^2 \log_2 t + 4\gamma b' + 2\Big) + 4UB\,,$$

which, by substituting $\gamma = 2^4 d_0 B/(\alpha U)$, is equal to

$$\frac{2^5 d_0 B^2}{\alpha} a'^2 \log_2 t + 4BU(1+b') + \frac{\alpha U^2}{8 d_0}\,.$$

This concludes the proof.

B.6 Proof of Theorem 3

Let first check that we are indeed in the setting of Theorem 1. The risk is strongly convex because for any $\theta_1, \theta_2 \in \mathbb{R}^d$

$$\mathbb{E}[\ell_t(\theta_1) - \ell_t(\theta_2)] = \mathbb{E}\Big[(Y_t - X_t^\top\theta_1)^2 - (Y_t - X_t^\top\theta_2)^2\Big]$$
$$= \mathbb{E}\Big[-2(Y_t - X_t^\top\theta_1)X_t^\top(\theta_1-\theta_2) - \big(X_t^\top(\theta_1-\theta_2)\big)^2\Big]$$
$$= \nabla\mathbb{E}[\ell_t](\theta_1)^\top(\theta_1-\theta_2) - (\theta_1-\theta_2)^\top\mathbb{E}\big[X_tX_t^\top\big](\theta_1-\theta_2)\,.$$

Assumption (SC) is thus satisfied with $\alpha = \lambda_{\min}(\mathbb{E}\big[X_tX_t^\top\big])$. Besides, for all θ such that $\|\theta\|_1 \leqslant 2U$, we have

$$\|\nabla\ell_t(\theta)\|_\infty = \|2(Y_t - X_t^\top\theta)X_t\|_\infty \leqslant 2(Y+2XU)X = B\,.$$

Now, we mimic the proof of Theorem 1. In the rest of the proof, we consider that $(\mathcal{H}_i)$ are satisfied for all $i \geqslant 0$. This occurs with probability $1-\delta$ and all inequalities stated in the proof of Theorem 1 are satisfied.

The proof is based on the following Lemma that we substitute to Inequality (24) from the proof of Theorem 1.

Lemma 8. *For all $t \in [t_i, t_{i+1}-1]$, with probability $1-\delta_{i+1}$,*

$$\mathrm{Err}_{t-1} \leqslant 2\sqrt{2}X\sigma a_i'\sqrt{t-t_i} + Bc_i'\,,$$

where $c_i' := b_i' + a_i'\big(\sqrt{\log\delta_{i+1}^{-1}} + \sqrt{2b} + 2a\big)$. We recall that Err_{t-1} is defined in (8).

Proof of Lemma 8. In the particular case of the square loss, the gradients are given by $\nabla\ell_t(\theta) = 2X_t(X_t^\top\theta - Y_t)$, so that

$$\|\nabla\ell_t(\widehat{\theta}_{t-1})\|_\infty^2 \leqslant 4X^2\ell_t(\widehat{\theta}_{t-1})\,. \tag{31}$$

Following Gerchinovitz, 2013, Corollary 2.2, we get from Inequality (5) that

$$\sum_{t=t_{\mathrm{start}}}^{t_{\mathrm{end}}} \ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta^*) \leqslant 2aUX\sqrt{\sum_{t_{\mathrm{start}}}^{t_{\mathrm{end}}}\ell_t(\widehat{\theta}_{t-1})} + bUB\,.$$

Solving the second-order inequality (see Gaillard et al., 2014, Lemma 10), it yields the improvement for small losses

$$\sqrt{\sum_{t=t_{\mathrm{start}}}^{t_{\mathrm{end}}}\ell_t(\widehat{\theta}_{t-1})} \leqslant \sqrt{\sum_{t=t_{\mathrm{start}}}^{t_{\mathrm{end}}}\ell_t(\theta^*)} + \sqrt{bUB} + 2aUX\,.$$

Thus, from (31),

$$\sqrt{\sum_{s=t_i}^{t-1}\big\|\nabla\ell_s(\widehat{\theta}_{s-1})\big\|_\infty^2} \leqslant 2X\sqrt{\sum_{s=t_i}^{t-1}\ell_s(\theta^*)}$$
$$+ 2X\sqrt{bUB} + 4aUX^2\,.$$

But, with probability $1-\delta_{i+1}$, we have from Theorem 9

$$\sum_{s=t_i}^{t-1}\ell_s(\theta^*) \leqslant (e-1)\sum_{s=t_i}^{t-1}\mathbb{E}[\ell_s(\theta^*)] + (Y+XU)^2\log\delta_{i+1}^{-1}$$
$$\leqslant 2\sigma^2(t-t_i) + (Y+XU)^2\log\delta_{i+1}^{-1}\,,$$

where $\sigma^2 = \mathbb{E}[\ell_t(\theta^*)]$. Plugging into the previous inequality and using $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$ for $x, y > 0$, this yields

$$2^{-1}X^{-1}\sqrt{\sum_{s=t_i}^{t-1}\big\|\nabla\ell_s(\widehat{\theta}_{s-1})\big\|_\infty^2} \tag{32}$$
$$\leqslant \sqrt{2}\sigma\sqrt{t-t_i} + (Y+XU)\sqrt{\log\delta_{i+1}^{-1}} + \sqrt{bUB} + 2aUX$$
$$\leqslant \sqrt{2}\sigma\sqrt{t-t_i} + 2^{-1}BX^{-1}\big(\sqrt{\log\delta_{i+1}^{-1}} + \sqrt{2b} + 2a\big)\,, \tag{33}$$

where the second inequality is because $B/(2X) \geqslant (Y + XU) \geqslant XU$. The proof of Lemma 8 is concluded by using the definition of Err_{t-1} (see (8)). □

The proof of Theorem 3 is then completed following the one of Theorem 1 by using Lemma 8 instead of Inequality (24). Finally, it only suffices to substitute Ba_i' with $2\sqrt{2}X\sigma a_i'$ and b_i' with c_i' in the final results. At the end, b' of Theorem 1 must thus be substituted with

$$c' := b' + a'\big(\sqrt{2\log(1+2\log_2 T) - \log\delta} + \sqrt{2b} + 4a\big)$$
$$\leqslant 1/2 + b + 3\log(1+3\log T) - \log\delta$$
$$+ \big(a + \sqrt{6\log(1+3\log t) - 2\log\delta}\big)$$
$$\big(\sqrt{2\log(1+3\log T) - \log\delta} + \sqrt{2b} + 2a\big)$$
$$\leqslant \frac{1}{2} + b + 3\log(1+3\log T) - \log\delta + 4a^2$$
$$+2b + 6\log(1+3\log T) - 2\log\delta)$$
$$\leqslant 1/2 + 3b + 4a^2 + 9\log(1+3\log T) - 3\log\delta\,.$$
$$\lesssim 1 + b + a^2 + \log\log T - \log\delta$$

However, in contrast to the bound B on the gradients, Lemma 8 only holds with probability $1-\delta_{i+1}$ (instead of almost surely). A union bound over all events states that the final result only holds with probability $1-\delta-\sum_{i=1}^{\infty}\delta_{i+1} = 1-2\delta$. To get a result with probability $1-\delta$, δ must thus be multiplied by 2 in the results.

This gives that, from the risk bound of Theorem 1, with probability $1-\delta$, $\mathrm{Risk}\left(\tilde\theta_t\right)$ is upper-bounded by

$$\min\Bigg\{4U\left(\frac{X\sigma a'}{\sqrt{T}}+\frac{Bc'}{T}\right)+\frac{\alpha U^2}{8d_0T}, \\ \frac{d_0}{\alpha}\left(\frac{2^{10}X^2\sigma^2a'^2}{T}+\frac{2^{11}B^2c'^2}{T^2}\right)+\frac{2\alpha U^2}{d_0T^2}\Bigg\},$$

where $a' = 2a+2\sqrt{6\log(1+3\log T)+2\log(2/\delta)}$ and $c' = 1+3b+4a^2+9\log(1+3\log T)+3\log(2/\delta)$.

The bound of the theorem is then obtained by using that $B = 2X(Y+2XU)$.

B.7 Proof of Theorem 4

For the sake of clarity, we only perform this proof up to universal constants. Let $B^* = 2X(Y+2X\|\theta^*\|_1) \geqslant \max_{\theta\in\mathcal{B}(0,2\|\theta^*\|_1)}\|\nabla\ell_t(\theta)\|_\infty$ almost surely. We also define by α^* the maximal number strong convexity parameter that satisfies (SC).

Let $T \geqslant 1$. Then, by definition (see Alg. 3), $\tilde f_{T-1} = \bar f_j$ for $j = \lfloor \log_2 T\rfloor - 1$.

We aim at controlling the excess risk of the average estimator $\bar f_j = \sum_{t=2^j}^{2^{j+1}-1}\widehat f_t$. To do so, we control the cumulative risk for $t = 2^j,\dots,2^{j+1}-1$

$$\mathrm{Risk}^{(j)} := \sum_{t=2^j}^{2^{j+1}-1}\mathbb{E}_{t-1}\big[(Y_t-\widehat f_t(X_t))^2\big] \\ -\mathbb{E}\big[(Y_t-X_t^\top\theta^*)^2\big],$$

where $\mathbb{E}_{t-1}[\,\cdot\,] = \mathbb{E}\big[\,\cdot\,|(X_1,Y_1),\dots,(X_{t-1},Y_{t-1})\big]$. We will use that

$$\mathrm{Risk}(\bar f_j) \leqslant \mathrm{Risk}^{(j)}\,2^{-j} \lesssim \frac{\mathrm{Risk}^{(j)}}{T}. \tag{34}$$

We first prove that it exists a predictor $f_{p,j}$ with $p\in\mathcal{G}_j$ that has a small excess risk. Then, we will apply Theorem 4.5 of Wintenberger, 2014 to show that BOA almost achieves this performance.

Step 1. Either it exists a predictor $f_{p,j}$ with small excess risk or $\mathrm{Risk}^{(j)}$ is small. Since all predictions $\widehat f_t(X_t)$ lie in $[-Y,Y]$ almost surely,

$$\mathrm{Risk}^{(j)} \leqslant Y^2 2^j \leqslant Y^2T\,. \tag{35}$$

Let d_0 in $\mathcal{G}_j$ (i.e., a power of 2) such that $d_0/2 \leqslant \|\theta^*\|_0 \leqslant d_0$. We show that if the conditions of Theorem 3 cannot be satisfied with any parameter of the grid $\mathcal{G}_j$, the cumulative risk $\mathrm{Risk}^{(j)}$ is small enough. We start with the choice of the parameter U, which should be of order $\|\theta^*\|_1$:

a) If $\|\theta^*\|_1 \leqslant 2^{-2j}$. It exists a predictor in $\mathcal{G}_j$ such that $f_{p,j} = 0$ (consider $d_0 = 0$). In this case,

$$\mathrm{Risk}(f_{p,j}) = \mathbb{E}[(Y_t-0)^2] \leqslant B^*\|\theta^*\|_1 \\ \leqslant B^*2^{-2j} \lesssim B^*T^{-2}\,,$$

where we used that $2^{-j}\lesssim T^{-1}$.

b) If $\|\theta^*\|_1 \geqslant 2^{2j+\lceil 2\log Y\rceil}$, then $2^j \leqslant Y^{-2}\|\theta^*\|_1 2^{-j}$ and from Inequality (35),

$$\mathrm{Risk}^{(j)} \leqslant \|\theta^*\|_1 2^{-j} \lesssim \frac{\|\theta^*\|_1}{T} \lesssim \frac{\|\theta^*\|_0(B^*)^2}{\alpha^*T}\,.$$

Otherwise, we can choose U in $\mathcal{G}_j$ such that $U/2 \leqslant \|\theta^*\|_1 \leqslant U$. Similarly for B:

c) if $B < 2^{-2j}$, then for $f_{p,j} = 0$,

$$\mathrm{Risk}(f_{p,j}) = \mathbb{E}[\ell(Y_t,0)] \leqslant B^*\|\theta^*\|_1 \\ \leqslant \|\theta^*\|_1 2^{-2j} \lesssim \frac{\|\theta^*\|_1}{T^2}\,,$$

d) if $B > 2^{2j+\lceil 2\log Y\rceil}$, then from Inequality (35), $\mathrm{Risk}^{(j)} \leqslant B^*2^{-j} \lesssim B^*T^{-1}$.

Otherwise, we can choose B in $\mathcal{G}_j$ such that $B/2 \leqslant B^* \leqslant B$. Finally, for α:

e) if $\alpha^* < 2^{-2j+\lceil\log_2(B^2d_0/Y^2)\rceil} \leqslant d_0B^22^{-2j}/Y^2$, then $2^j \leqslant d_0B^2/(Y^2\alpha^*2^j)$ and thus

$$\mathrm{Risk}^{(j)} \leqslant Y^22^j \leqslant Y^2\frac{d_0B^2}{Y^2\alpha^*2^j} \lesssim \frac{\|\theta^*\|_0(B^*)^2}{\alpha^*T}\,.$$

Otherwise, we can choose α in $\mathcal{G}_j$ such that $\min\{d_0/T,\alpha^*/2\} \leqslant \alpha \leqslant \alpha^*$.

f) Applying Theorem 3, with high probability the excess risk of the estimator $f_{p,j}$ with the choice (d_0,α,U,B) described above satisfies

$$\mathrm{Risk}(f_{p,j}) \overset{\text{clipping}}{\leqslant} \mathrm{Risk}(\tilde\theta_{p,j}) \\ \lesssim \min\Bigg\{\frac{X^2}{\gamma}\left(\frac{\sigma^2a'^2}{T}+\frac{(Y+X\|\theta^*\|_1)^2c'^2}{T^2}\right)+\frac{\gamma\|\theta^*\|_1^2}{T^2}, \\ \|\theta^*\|_1X\left(\frac{\sigma a'}{\sqrt{T}}+\frac{(Y+X\|\theta^*\|_1)c'}{T}\right)+\frac{\gamma\|\theta^*\|_1^2}{T}\Bigg\},$$

with $\gamma = \max\{d_0/\alpha, 1/T\}$.

Putting everything together, either (for cases b), d), and e))

$$\mathrm{Risk}^{(j)} \lesssim \Big(B^* + \frac{\|\theta^*\|_0 (B^*)^2}{\alpha^*}\Big) T^{-1} \tag{36}$$

or, for cases a), c), and f), there exists $p \in \mathcal{G}_j$ such that with high probability

$$\mathrm{Risk}(f_{p,j}) \lesssim \min\left\{ \frac{1}{\gamma}\left(\frac{X^2\sigma^2 a'^2}{T} + \frac{(B^* c')^2}{T^2}\right) + \frac{\gamma\|\theta^*\|_1^2}{T^2}, \right.$$
$$\left. \|\theta^*\|_1 X \left(\frac{\sigma a'}{\sqrt{T}} + \frac{(Y + X\|\theta^*\|_1)c'}{T}\right) + \frac{\gamma\|\theta^*\|_1^2}{T}\right\} + \frac{B^*}{T^2}\,, \tag{37}$$

Step 2. Bound of the meta-algorithm. Using that the square loss is $4Y$-Lipschitz over the domain $[-Y, Y]$ and 2-strongly convex, we can apply Theorem 4.5 of Wintenberger, 2014 with $C_b = 4Y$, $C_\ell = 2$, and $M = \#\mathcal{G}_j$. We get that with high enough probability

$$\mathrm{Risk}^{(j)} \lesssim T \min_{p\in\mathcal{G}_j} \mathrm{Risk}(f_{p,j})$$
$$+Y^2\big(\log \#\mathcal{G}_j + \log(\log T + \log Y) - \log\delta\big)\,.$$

Substituting

$$\#\mathcal{G}_j \lesssim (j + \log Y)^3 \log d \lesssim (\log T + \log Y)^3 \log d\,,$$

this yields

$$\mathrm{Risk}^{(j)} \lesssim \frac{\min_{p\in\mathcal{G}_j}\mathrm{Risk}(f_{p,j})}{T}$$
$$+Y^2\big(\log\log d + \log(\log T + \log Y) - \log\delta\big)\,.$$

Combining with Inequality (37), we obtain that $\mathrm{Risk}^{(j)}$ is at most of order

$$\mathrm{Risk}^{(j)} \lesssim Y^2\big(\log\log d + \log(\log T + \log Y) - \log\delta\big)$$
$$+ \min\left\{\frac{1}{\gamma}\left(X^2\sigma^2 a'^2 + \frac{(B^*c')^2}{T}\right) + \frac{\gamma\|\theta^*\|_1^2}{T},\right.$$
$$\left.\|\theta^*\|_1 X\left(\sigma a'\sqrt{T} + (Y + X\|\theta^*\|_1)c'\right) + \gamma\|\theta^*\|_1^2\right\} + \frac{B^*}{T}.$$

Finally, using Inequality (34), keeping only the main asymptotic term in $1/T$, and substituting $a' \lesssim \log((d\log T)/\delta)$ concludes the proof.

C Martingale inequalities

In this section, we prove two martingale inequalities that are used in the analysis.

C.1 Poissonian inequality

First, we prove a Poissonian inequality which only works for nonnegative increments.

Theorem 9. *Let $T \geqslant 1$. Let $(X_t)_{t\geqslant 1}$ be a sequence of random variables such that $X_t \in [0, B]$ almost surely, then with probability at least $1-\delta$*

$$\sum_{t=1}^T X_t \leqslant (e-1)\sum_{t=1}^T \mathbb{E}_{t-1}[X_t] + B\log(1/\delta)\,.$$

Proof. Let $Z_t = X_t/B \in [0, 1]$. From Cesa-Bianchi and Lugosi, 2006, Lemma A.3, for all $t \geqslant 1$, and all $s > 0$

$$\mathbb{E}_{t-1}\Big[\exp\big(sZ_t - (e^s - 1)\mathbb{E}_{t-1}[Z_t]\big)\Big] \leqslant 1\,.$$

Thus,

$$\mathbb{E}\Big[\exp\Big(s\sum_{t=1}^T Z_t - (e^s-1)\sum_{t=1}^T \mathbb{E}_{t-1}[Z_t]\Big)\Big]$$
$$= \mathbb{E}\Big[\mathbb{E}_{T-1}\Big[\exp\big(sZ_T - (e^s-1)\mathbb{E}_{T-1}[Z_T]\big)\Big]$$
$$\exp\Big(s\sum_{t=1}^{T-1} Z_t - (e^s-1)\sum_{t=1}^{T-1}\mathbb{E}_{t-1}[Z_t]\Big)\Big]$$
$$\leqslant \mathbb{E}\Big[\exp\Big(s\sum_{t=1}^{T-1} Z_t - (e^s-1)\sum_{t=1}^{T-1}\mathbb{E}_{t-1}[Z_t]\Big)\Big]$$

By induction, we get

$$\mathbb{E}\Big[\exp\Big(s\sum_{t=1}^T Z_t - (e^s-1)\sum_{t=1}^T \mathbb{E}_{t-1}[Z_t]\Big)\Big] \leqslant 1\,.$$

We conclude thanks to Markov's inequality, with probability at least $1-\delta$

$$\sum_{t=1}^T Z_t \leqslant \frac{e^s-1}{s}\sum_{t=1}^T \mathbb{E}_{t-1}[Z_t] + \frac{1}{s}\log(1/\delta)\,.$$

The final result is obtained by substituting $Z_t = X_t/B$ and by choosing $s = 1$. □

C.2 From cumulative regret to cumulative risk

Theorem 10. *Let $x > 0$. Assume $\theta^* \in \mathcal{B}_1(\theta_{\mathrm{center}}, \varepsilon)$. The cumulative risk of any convex optimization procedure in $\mathcal{B}_1(\theta_{\mathrm{center}}, \varepsilon)$ satisfies, with probability $1-\delta$*

$$\mathrm{Risk}_{1:T}(\widehat{\theta}_{0:(T-1)}) - \mathrm{Reg}_{1:T}(\widehat{\theta}_{0:(T-1)})$$
$$\leqslant \varepsilon\sqrt{2\log\Big(\frac{2+\log(T/2)}{2\delta}\Big)\sum_{t=1}^T \big\|\nabla\ell_t(\widehat{\theta}_{t-1})\big\|_\infty^2}$$

$$+\Big(\frac{1}{2} + \log\big(1 + \frac{1}{2}\log(T/2)\big) - \log\delta\Big)\varepsilon B\,,$$

where $B \geqslant \max_{\theta\in\mathcal{B}_1(\theta_{\text{center}},\varepsilon)} \big\|\nabla\ell_t(\widehat{\theta}_{t-1})\big\|_\infty$ *almost surely.*

Proof. This is a consequence of Theorem 4.1 of Wintenberger, 2014. Let $\delta \in (0,1)$ and $(\eta_t)_{t\geqslant 0}$ be a sequence adapted to the filtration $(\mathcal{F}_t = \{\ell_1,\dots,\ell_{t-1}\})_{t\geqslant 0}$. Then, with the notation $\ell_{j,t}^2 \leqslant \varepsilon^2\|\nabla\ell_t(\widehat{\theta}_{t-1})\|_\infty^2$, applying Theorem 4.1 of Wintenberger (2014), we get that with probability $1-\delta$

$$\begin{aligned} R_T &:= \text{Risk}_{1:T}(\widehat{\theta}_{0:(T-1)}) - \text{Reg}_{1:T}(\widehat{\theta}_{0:(T-1)}) \\ &\leqslant \varepsilon^2 \sum_{t=1}^T \eta_{t-1}\|\nabla\ell_t(\widehat{\theta}_{t-1})\|_\infty^2 \\ &\quad + \frac{\log\Big(1 + \mathbb{E}\big[\log(\eta_1/\eta_T)\big]\Big) - \log\delta}{\eta_T}\,, \end{aligned} \tag{38}$$

where $\text{Risk}_{1:T}(\widehat{\theta}_{0:(T-1)}) := \sum_{t=1}^T \mathbb{E}[\ell_t](\widehat{\theta}_{t-1}) - \mathbb{E}[\ell_t](\theta^*)$ and $\text{Reg}_{1:T}(\widehat{\theta}_{0:(T-1)}) := \sum_{t=1}^T \ell_t(\widehat{\theta}_{t-1}) - \ell_t(\theta^*)$.

We obtain the stated inequality from (38), by properly setting the tuning parameters

$$\eta_t := \frac{1}{\varepsilon}\min\left\{\frac{1}{B}, \frac{c\Gamma}{V_{t-1}}\right\},$$

where c will be set by the analysis and

$$\Gamma := \sqrt{\log\big(1 + \log(\sqrt{T}/c)\big) - \log\delta}\,,$$

and

$$V_{t-1} := \sqrt{\sum_{s=1}^{t-1} \big\|\nabla\ell_s(\widehat{\theta}_{s-1})\big\|_\infty^2}\,.$$

Indeed, first we use that that $\eta_1/\eta_T \leqslant \sqrt{T}/c$ so that $\mathbb{E}[\log(\eta_1/\eta_T)] \leqslant \log(\sqrt{T}/c)$. Then, similarly to the proof of Cesa-Bianchi et al., 2007, Theorem 5, we can show that the first term in the right-hand side of (38) is upper-bounded as

$$\sum_{t=1}^T \eta_{t-1}\|\nabla\ell_t(\widehat{\theta}_{t-1})\|_\infty^2 \leqslant \frac{B}{2\varepsilon} + \frac{c\Gamma}{2\varepsilon}V_T.$$

But, by definition of η_T, the second term is also controlled as

$$\frac{\log\Big(1 + \mathbb{E}\big[\log(\eta_1/\eta_T)\big]\Big) - \log\delta}{\eta_T} \leqslant \varepsilon\Gamma\max\left\{B\Gamma, \frac{1}{c}V_T\right\}.$$

Plugging these two last inequalities into (38) leads to

$$R_T \leqslant \frac{B\varepsilon}{2} + \frac{c\Gamma\varepsilon}{2}V_T + \varepsilon\Gamma\max\left\{B\Gamma, \frac{V_T}{c}\right\}.$$

We then need to distinguish two cases

- if $c\Gamma B \leqslant V_T$, then optimizing in $c = \sqrt{2}$

$$R_T \leqslant \frac{B\varepsilon}{2} + \Big(\frac{c}{2} + \frac{1}{c}\Big)\Gamma\varepsilon V_T \leqslant \frac{B\varepsilon}{2} + \sqrt{2}\Gamma\varepsilon V_T$$

- if $c\Gamma B \geqslant V_T$, then

$$R_T \leqslant \frac{B\varepsilon}{2} + \frac{1}{\sqrt{2}}\Gamma\varepsilon V_T + \varepsilon B\Gamma^2\,.$$

Therefore, putting the two cases together

$$R_T \leqslant \frac{B\varepsilon}{2} + \sqrt{2}\Gamma\varepsilon V_T + \varepsilon B\Gamma^2\,.$$

We conclude the proof by substituting Γ and V_T with their definitions. □