A. Bound for Online Gradient Descent with Per-Dimension Learning Rates

We remind the update of OGD with per-dimension learning rates:

$$w_{t+1,i} = w_{t,i} - \eta_i \nabla_{t,i}, \qquad i = 1, \ldots, d,$$

with $\boldsymbol{w}_1 = \mathbf{0}$. For any $u_i \in \mathbb{R}$, we have:

$$(u_i - w_{t+1,i})^2 - (u_i - w_{t,i})^2 = (u_i - w_{t,i} + \eta_i \nabla_{t,i})^2 - (u_i - w_{t,i})^2 = 2\eta_i \nabla_{t,i}(u_i - w_{t,i}) + \eta_i^2 \nabla_{t,i}^2.$$

Summing over trials $t = 1, \ldots, T$ and rearranging:

$$2\eta_i \sum_{t=1}^{T} \nabla_{t,i}(w_{t,i} - u_i) = u_i^2 - (u_i - w_{T+1,i})^2 + \eta_i^2 \sum_{t=1}^{T} \nabla_{t,i}^2.$$

Dividing by $2\eta_i$, upper bounding and summing over $i = 1, \ldots, d$:

$$\sum_{t=1}^{T} \nabla_t^\top (\boldsymbol{w}_t - \boldsymbol{u}) \le \sum_{i=1}^{d} \left(\frac{u_i^2}{2\eta_i} + \frac{\eta_i}{2} \sum_{t=1}^{T} \nabla_{t,i}^2 \right).$$

Finally, using (3) shows that the right-hand side of the above upper bounds the regret.

B. Scale Invariance of Algorithm 1 and Algorithm 2

Let $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^T$ be a data sequence and define a transformed sequence $\{(\boldsymbol{A}\boldsymbol{x}_t, y_t)\}_{t=1}^T$, where $\boldsymbol{A} = \operatorname{diag}(a_1, \ldots, a_d)$ with $a_1, \ldots, a_d > 0$. We will show that the sequence of predictions $\widehat{y}_1, \ldots, \widehat{y}_T$ generated by the algorithms on the original and the transformed data sequences are the same. This can easily be done inductively: assuming $\widehat{y}_1, \ldots, \widehat{y}_t$ are the same on both sequences, this implies $g_1, \ldots, g_t$ are also the same (as $g_t = \partial_{\widehat{y}_t} \ell(y_t, \widehat{y}_t)$, while y_t are the same in both sequences). Given that, a closer inspection of the algorithms lets us determine the behavior of all maintained statistics under the feature transformation $x_{t,i} \mapsto a_i x_{t,i}$.

For both algorithms we have:

$$M_{t,i} = \max_{j \le t} |x_{j,i}| \mapsto a_i M_{t,i}, \quad S_{t,i}^2 = \sum_{j \le t} (g_j x_{j,i})^2 \mapsto a_i^2 S_{t,i}^2, \quad G_{t,i} = -\sum_{j \le t} g_j x_{j,i} \mapsto a_i G_{t,i},$$

This means that for Algorithm 1:

$$\beta_{t,i} \mapsto \beta_{t,i}, \quad \theta_{t,i} \mapsto \theta_{t,i}, \quad w_{t,i} \mapsto a_i^{-1} w_{t,i},$$

so that $x_{t,i} w_{t,i} \mapsto x_{t,i} w_{t,i}$ and thus $\widehat{y}_t = \boldsymbol{x}_t^\top \boldsymbol{w}_t$ is invariant under the scale transformation.

Similarly, for Algorithm 2 we have:

$$\eta_{t,i} \mapsto \eta_{t,i}, \quad \theta_{i,i} \mapsto \theta_{t,i}, \quad w_{t,i} \mapsto a_i^{-1} w_{t,i},$$

and the scale invariance follows.

C. Proof of Theorem 3.1

Before proving the theorem, we need two auxiliary results:

Lemma C.1. *Let $f(x) = \alpha \left(e^{|x|/\gamma} - |x|/\gamma - 1 \right)$ with $\alpha, \gamma > 0$. Its Fenchel conjugate is given by:*

$$\begin{aligned} f^*(u) &\stackrel{def}{=} \sup_x \{ux - f(x)\} \\ &= (|u|\gamma + \alpha) \ln(1 + |u|\gamma/\alpha) - |u|\gamma \\ &\le |u|\gamma \ln(1 + |u|\gamma/\alpha). \end{aligned} \tag{6}$$

Proof. Note that since $f(x)$ is symmetric in x,

$$\begin{aligned}\sup_x\{ux - f(x)\} &= \sup_{x\geq 0}\{|u|x - f(x)\} \\ &= \sup_{x\geq 0}\Big\{\underbrace{|u|x - \alpha\left(e^{x/\gamma} - x/\gamma - 1\right)}_{g(x)}\Big\}. \end{aligned} \tag{7}$$

Setting the derivative of $g(x)$ to zero gives its unconstrained maximizer $x^* = \gamma\ln(1 + |u|\gamma/\alpha)$, and since $x^* \geq 0$, it is also the maximizer of $g(x)$ under constraint $x \geq 0$. Thus:

$$f^*(u) = g(x^*) = (|u|\gamma + \alpha)\gamma\ln(1 + |u|\gamma/\alpha) - |u|\gamma.$$

The inequality in the lemma follows from an elementary inequality $\ln(1 + x) \leq x$ applied to $\alpha\ln(1 + |u|\gamma/\alpha)$. □

Lemma C.2. *For any $v \in \mathbb{R}$ and any $q \in [-1, 1]$:*

$$\frac{q\,\mathrm{sgn}(v)}{2}\left(e^{\frac{|v|}{2}} - 1\right) + e^{\frac{|v-q|}{2\sqrt{1+q^2}}} - \frac{|v-q|}{2\sqrt{1+q^2}} \leq e^{\frac{|v|}{2}} - \frac{|v|}{2} + q^2.$$

Proof. It suffices to prove the lemma for $v \geq 0$. Indeed, the inequality holds for some $v \geq 0$ and $q \in [-1, 1]$ if and only if it holds for $-v$ and $-q$. Denote:

$$\widetilde{v} = \frac{|v-q|}{\sqrt{1+q^2}}.$$

In this notation and with the assumption $v \geq 0$, the inequality translates to:

$$e^{\frac{\widetilde{v}}{2}} - \frac{\widetilde{v}}{2} \leq e^{\frac{v}{2}}\left(1 - \frac{q}{2}\right) - \frac{v-q}{2} + q^2 \tag{8}$$

We will split the proof into three sub-cases: (i) $q \geq v$, (ii) $q \leq v \leq 3$, and (iii) $v \geq 3$. Since $q \leq 1$, these cases cover all allowed values of v and q.

Case (i): $q \geq v$. We have $\widetilde{v} = \frac{q-v}{\sqrt{1+q^2}} \leq q - v$. Since the function $e^x - x$ is increasing in x for $x \in (1, \infty)$, it holds:

$$e^{\frac{\widetilde{v}}{2}} - \frac{\widetilde{v}}{2} \leq e^{\frac{q-v}{2}} - \frac{q-v}{2} = e^{\frac{v}{2}}e^{\frac{q-2v}{2}} - \frac{q-v}{2}.$$

From $q \leq 1$ and $v \geq 0$ it follows $\frac{q-2v}{2} \leq \frac{1-2v}{2} \leq \frac{1}{2}$. Since function $f(x) = \frac{e^x - x - 1}{x^2}$ is nondecreasing in x (see, e.g., (Cesa-Bianchi & Lugosi, 2006), Section A.1.2), we have:

$$e^x - x - 1 \leq x^2\frac{e^{1/2} - 1/2 - 1}{1/4} \leq 0.6x^2 \qquad \text{for } x \leq \frac{1}{2}. \tag{9}$$

Thus, we bound $e^{\frac{q-2v}{2}}$ by $1 + \frac{q-2v}{2} + 0.15(q-2v)^2$ and get:

$$\begin{aligned} e^{\frac{\widetilde{v}}{2}} - \frac{\widetilde{v}}{2} &\leq e^{\frac{v}{2}}\left(1 + \frac{q-2v}{2}\right) - \frac{q-v}{2} + 0.15e^{\frac{v}{2}}(q-2v)^2 \\ &= e^{\frac{v}{2}}\left(1 - \frac{q}{2}\right) - \frac{v-q}{2} + (e^{\frac{v}{2}} - 1)(q-v) + 0.15e^{\frac{v}{2}}(q-2v)^2 \\ &\leq e^{\frac{v}{2}}\left(1 - \frac{q}{2}\right) - \frac{v-q}{2} + v(q-v) + \frac{1}{4}(q-2v)^2, \end{aligned}$$

where the last inequality follows from the fact that $v \leq 1$ (as $q \geq v$ and $q \leq 1$), which by (9) implies $e^{\frac{v}{2}} \leq 1 + \frac{v}{2} + 0.6\frac{v^2}{4} = 1 + 0.5v + 0.15v^2 \leq 1 + v$, and furthermore $0.15e^{\frac{v}{2}} \leq 0.15e^{\frac{1}{2}} \leq \frac{1}{4}$. But $v(q-v) + \frac{1}{4}(q-2v)^2 = \frac{1}{4}q^2 \leq q^2$, which proves (8) for $q \geq v$.

Case (ii): $q \le v \le 3$. We have $\widetilde{v} = \frac{v-q}{\sqrt{1+q^2}} \le v - q$, and by the monotonicity of function $e^x - x$ for $x \in (1, \infty)$:

$$e^{\frac{\widetilde{v}}{2}} - \frac{\widetilde{v}}{2} \le e^{\frac{v-q}{2}} - \frac{v-q}{2} = e^{\frac{v}{2}} e^{-\frac{q}{2}} - \frac{v-q}{2}.$$

Using (9) and $q \ge -1$, we bound $e^{-q/2} \le 1 - \frac{q}{2} + 0.15q^2$ to get:

$$e^{\frac{\widetilde{v}}{2}} - \frac{\widetilde{v}}{2} \le e^{\frac{v}{2}} \left(1 - \frac{q}{2}\right) - \frac{v-q}{2} + 0.15 e^{\frac{v}{2}} q^2.$$

Using $0.15 e^{\frac{v}{2}} \le 0.15 e^{\frac{3}{2}} \le 0.68 \le 1$ proves (8) for $q \le v \le 3$.

Case (iii): $v > 3$. We lower-bound the right-hand side of (8):

$$e^{\frac{v}{2}} \left(1 - \frac{q}{2}\right) - \frac{v-q}{2} + q^2 \ge e^{\frac{v}{2}} \left(1 - \frac{q}{2}\right) - \frac{v - q - \frac{q^2}{2}}{2} \ge e^{\frac{1}{2}(v - q - \frac{q^2}{2})} - \frac{v - q - \frac{q^2}{2}}{2},$$

where the first inequality is simply from $q^2 \ge \frac{q^2}{4}$, while the second follows from $1 - x \ge e^{-x-x^2}$ for $x \le \frac{1}{2}$ (see, .e.g., (Cesa-Bianchi & Lugosi, 2006), Lemma 2.4). Now, using the monotonicity of function $e^x - x$,

$$e^{\frac{1}{2}(v - q - \frac{q^2}{2})} - \frac{v - q - \frac{q^2}{2}}{2} \ge e^{\frac{\widetilde{v}}{2}} - \frac{\widetilde{v}}{2} \qquad \Longleftrightarrow \qquad v - q - \frac{q^2}{2} \ge \widetilde{v},$$

thus it suffices to show the latter to finish the proof. We have:

$$v - q - \frac{q^2}{2} - \widetilde{v} = (v-q)\left(1 - \frac{1}{\sqrt{1+q^2}}\right) - \frac{q^2}{2} \ge (3-q)\left(1 - \frac{1}{\sqrt{1+q^2}}\right) - \frac{q^2}{2}.$$

Using elementary inequality $\sqrt{1+x} \le 1 + \frac{x}{2}$, we have: $\frac{1}{\sqrt{1+q^2}} = \frac{\sqrt{1+q^2}}{1+q^2} \le \frac{1+q^2/2}{1+q^2}$, and thus:

$$\begin{aligned} v - q - \frac{q^2}{2} - \widetilde{v} &\ge (3-q)\left(1 - \frac{1+q^2/2}{1+q^2}\right) - \frac{q^2}{2} = (3-q)\frac{q^2/2}{1+q^2} - \frac{q^2}{2} \\ &= \frac{q^2}{2}\left(\frac{3-q}{1+q^2} - 1\right) \ge \frac{q^2}{2}\left(\frac{3-1}{1+1} - 1\right) = 0. \end{aligned}$$

This shows that $v - q - \frac{q^2}{2} \ge \widetilde{v}$ and thus proves (9) for $v > 3$. □

Before we state the next result, we summarize the notation which will be used in what follows. For any $i = 1, \ldots, d$ and any $t = 1, \ldots, T$, let:

$$M_{t,i} = \max_{j \le t} |x_{j,i}|, \qquad G_{t,i} = -\sum_{j \le t} g_j x_{j,i}, \qquad S_{t,i}^2 = \sum_{j \le t} (g_j x_{j,i})^2,$$

be, respectively, the maximum input value, the negative cumulative gradient, and the sum of squared gradients at i-th coordinate up to (and including) trial t, and we also denote $M_{0,i} = G_{0,i} = S_{0,i}^2 = 0$. Moreover, define:

$$\beta_{t,i} = \begin{cases} \min\left\{\beta_{t-1,i}, \epsilon \frac{S_{t-1,i}^2 + M_{t,i}^2}{x_{t,i}^2 t}\right\} & \text{when } x_{t,i} \ne 0, \\ \beta_{t-1,i} & \text{when } x_{t,i} = 0, \end{cases}$$

with $\beta_{1,i} = \epsilon$. The weight vector at trial t is given by:

$$w_{t,i} = \frac{\beta_{t,i} \mathrm{sgn}(G_{t-1,i})}{2\sqrt{S_{t-1,i}^2 + M_{t,i}^2}} \left(e^{\frac{|G_{t-1,i}|}{2\sqrt{S_{t-1,i}^2 + M_{t,i}^2}}} - 1 \right), \tag{10}$$

as long as $M_{t,i} > 0$; if $M_{t,i} = 0$ (which means that $x_{j,i} = 0$ for all $j \le t$), we set $w_{t,i} = 0$, but any other value of $w_{t,i}$ would lead to the same loss. Finally, define $\hat{S}_{t,i}^2 = S_{t,i}^2 + M_{t,i}^2$.

Lemma C.3. *Define:*

$$\psi_{t,i}(x) = \begin{cases} \beta_{t,i}\left(e^{|x|/(2\hat{S}_{t,i})} - \frac{|x|}{2\hat{S}_{t,i}} - 1\right) & \text{for } \hat{S}_{t,i} \neq 0, \\ 0 & \text{for } \hat{S}_{t,i} = 0. \end{cases}$$

For any $i = 1, \ldots, d$ and any $t = 1, \ldots, T$ we have:

$$w_{t,i} g_t x_{t,i} \;\leq\; \psi_{t-1,i}(G_{t-1,i}) - \psi_{t,i}(G_{t,i}) + \frac{\epsilon}{t}.$$

Proof. Fix $i \in \{1, \ldots, d\}$, and let τ_i be the first trial t such that $x_{t,i} \neq 0$. This means that $\hat{S}_{t,i} = x_{t,i} = 0$ for all $t < \tau_i$, and the inequality is trivially satisfied for any $t < \tau_i$, as the left-hand side is zero, while the right-hand side is ϵ/t. Thus, assume $t \geq \tau_i$.

Fix t and define $v = \frac{G_{t-1,i}}{\sqrt{S^2_{t-1,i}+M^2_{t,i}}}$ and $q = \frac{g_t x_{t,i}}{\sqrt{S^2_{t-1,i}+M^2_{t,i}}}$. As $|q| \leq \frac{|g_t x_{t,i}|}{M_{t,i}} \leq \frac{|x_{t,i}|}{\max_{j \leq t}|x_{j,i}|} \leq 1$, we can apply Lemma C.2 to such v and q, which, after subtracting 1 and multiplying by $\beta_{t,i}$ on both sides, gives:

$$\begin{aligned} \beta_{t,i}\frac{q\,\mathrm{sgn}(v)}{2}\left(e^{\frac{|v|}{2}} - 1\right) + \beta_{t,i}\left(e^{\frac{|v-q|}{2\sqrt{1+q^2}}} - \frac{|v-q|}{2\sqrt{1+q^2}} - 1\right) \\ \leq \beta_{t,i}\left(e^{\frac{|v|}{2}} - \frac{|v|}{2} - 1\right) + \beta_{t,i}q^2. \end{aligned} \tag{11}$$

Using the definition of the weight vector (10) we identify the first term on the left-hand side of (11):

$$\beta_{t,i}\frac{q\,\mathrm{sgn}(v)}{2}\left(e^{|v|/2} - 1\right) = w_{t,i} g_t x_{t,i}.$$

Next, since:

$$\frac{G_{t,i}}{\hat{S}_{t,i}} = \frac{G_{t,i}}{\sqrt{S^2_{t,i}+M^2_{t,i}}} = \frac{G_{t-1,i} - g_t x_{t,i}}{\sqrt{S^2_{t-1,i}+M^2_{t,i}+(g_t x_{t,i})^2}} = \frac{v-q}{\sqrt{1+q^2}},$$

the second term on the left-hand side of (11) is equal to $\psi_{t,i}(G_{t,i})$. Thus, (11) can be rewritten as:

$$w_{t,i} g_t x_{t,i} + \psi_{t,i}(G_{t,i}) \leq \beta_{t,i}\left(e^{\frac{|v|}{2}} - \frac{|v|}{2} - 1\right) + \beta_{t,i}q^2,$$

and to finish the proof, it suffices to show that the two terms on the right-hand side are upper bounded, respectively, by $\psi_{t-1,i}(G_{t-1,i})$ and $\frac{\epsilon}{t}$.

To bound $\beta_{t,i}q^2$ note that if $x_{t,i} = 0$ then $\beta_{t,i}q^2 = 0$, whereas if $x_{t,i} \neq 0$ then by the definition of $\beta_{t,i}$:

$$\beta_{t,i}q^2 = \beta_{t,i}\frac{(g_t x_{t,i})^2}{S^2_{t-1,i}+M^2_{t,i}} \leq \epsilon\frac{S^2_{t-1,i}+M^2_{t,i}}{x^2_{t,i}t}\frac{(g_t x_{t,i})^2}{S^2_{t-1,i}+M^2_{t,i}} \leq \frac{\epsilon g_t^2}{t} \leq \frac{\epsilon}{t}.$$

To bound $\beta_{t,i}(e^{|v|/2} - |v|/2 - 1)$ by $\psi_{\tau_i-1,i}(G_{t-1,i})$ note that both are zero if $t = \tau_i$ (because $G_{\tau_i-1,i} = 0$ and $v = 0$). On the other hand, for $t > \tau_i$ we have:

$$|v| = \frac{|G_{t-1,i}|}{\sqrt{S^2_{t-1,i}+M^2_{t,i}}} \leq \frac{|G_{t-1,i}|}{\sqrt{S^2_{t-1,i}+M^2_{t-1,i}}} = \frac{|G_{t-1,i}|}{\hat{S}_{t-1,i}},$$

and by the monotonicity of $f(x) = e^x - x - 1$:

$$\beta_{t,i}(e^{|v|/2} - |v|/2 - 1) \leq \beta_{t,i}\left(e^{\frac{|G_{t-1,i}|}{2\hat{S}_{t-1,i}}} - \frac{|G_{t-1,i}|}{2\hat{S}_{t-1,i}} - 1\right) \leq \psi_{t-1,i}(G_{t-1,i}),$$

where in the last inequality we used $\beta_{t,i} \leq \beta_{t-1,i}$ (which follows from the definition) and the fact that $e^x - x - 1 \geq 0$ for all x. □

We are now ready to prove Theorem 3.1, which we restate here for convenience:

Theorem. *For any $\boldsymbol{u} \in \mathbb{R}$ the regret of* ScInOL_1 *is upper-bounded by:*

$$R_T(\boldsymbol{u}) \leq \sum_{i=1}^{d} \left(2|u_i|\hat{S}_{T,i} \ln(1 + 2|u_i|\hat{S}_{T,i}\epsilon^{-1}T) + \epsilon(1 + \ln T)\right) = \sum_{i=1}^{d} \tilde{O}(|u_i|\hat{S}_{T,i}),$$

where $\hat{S}_{T,i} = \sqrt{S_{T,i}^2 + M_{T,i}^2}$ and $\tilde{O}(\cdot)$ hides the constants and logarithmic factors.

Proof. Applying Lemma (C.3) for a fixed $i \in \{1, \ldots, d\}$ and all $t = 1, \ldots, T$, and summing over trials gives:

$$\sum_{t=1}^{T} w_{t,i} g_t x_{t,i} \leq -\psi_{T,i}(G_{T,i}) + \sum_{t=1}^{T} \frac{\epsilon}{t} \leq -\psi_{T,i}(G_{T,i}) + \epsilon\,(1 + \ln T)\,,$$

where we used $\psi_{0,i}(G_{0,i}) = 0$. By (3),

$$\begin{aligned}
R_T(\boldsymbol{u}) &\leq \sum_{t=1}^{T} g_t \boldsymbol{x}_t^\top (\boldsymbol{w}_t - \boldsymbol{u}) = \sum_{i=1}^{d} \left(\sum_{t=1}^{T} g_t x_{t,i} w_{t,i} + G_{T,i} u_i\right) \\
&\leq \sum_{i=1}^{d} \left(G_{T,i} u_i - \psi_{T,i}(G_{T,i})\right) + d\epsilon\,(1 + \ln(T)) \\
&\leq \sum_{i=1}^{d} \sup_x \left\{x u_i - \psi_{T,i}(x)\right\} + d\epsilon\,(1 + \ln(T)) \\
&\leq \sum_{i=1}^{d} 2|u_i|\hat{S}_{T,i} \ln\left(1 + 2|u_i|\hat{S}_{T,i}/\beta_{T,i}\right) + d\epsilon\,(1 + \ln(T))\,,
\end{aligned}$$

where in the last inequality we used Lemma C.1 for each i with $\alpha = \beta_{T,i}$ and $\gamma = 2\hat{S}_{T,i}$. To finish the proof, it suffices to show that $\beta_{T,i} \geq \frac{\epsilon}{T}$, which we do by induction on t. For $t = 1$, we have by the definition $\beta_{t,i} = \epsilon$. Now, assume $\beta_{t-1,i} \geq \frac{\epsilon}{t-1}$, and we will show $\beta_{t,i} \geq \frac{\epsilon}{t}$. If $x_{t,i} = 0$, $\beta_{t,i} = \beta_{t-1,i} \geq \frac{\epsilon}{t-1} > \frac{\epsilon}{t}$; on the other hand, if $x_{t,i} \neq 0$, from the definition of $\beta_{t,i}$:

$$\beta_{t,i} = \min\left\{\beta_{t-1,i}, \epsilon\frac{S_{t-1,i}^2 + M_{t,i}^2}{x_{t,i}^2 t}\right\} \geq \min\left\{\frac{\epsilon}{t-1}, \epsilon\frac{x_{t,i}^2}{x_{t,i}^2 t}\right\} = \frac{\epsilon}{t},$$

where we used $S_{t-1,i}^2 + M_{t,i}^2 \geq M_{t,i}^2 = \max_{j \leq t} x_{j,i}^2 \geq x_{t,i}^2$. □

D. Proof of Theorem 3.2

Similarly as in the previous section, we proceed the proof of the theorem with several auxiliary results. Define:

$$h(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } |x| \leq 1, \\ |x| - \frac{1}{2} & \text{for } |x| > 1 \end{cases} \tag{12}$$

(see Figure 3). Note that $h(x) = h(|x|)$, and $h(|x|)$ is monotonic in $|x|$. Moreover, for all $x \in \mathbb{R}$:

$$|x| - \frac{1}{2} \leq h(x) \leq \frac{1}{2}x^2. \tag{13}$$

The lower bound in (13) is clearly satisfied for $|x| < 1$, while for $|x| \leq 1$ we have $h(x) - (|x| - \frac{1}{2}) = \frac{1}{2}(|x| - 1)^2 \geq 0$. On the other hand, the upper bound in (13) is clearly satisfied for $|x| \leq 1$, while for $|x| > 1$ we have $h(x) - \frac{1}{2}x^2 = -\frac{1}{2}(|x| - 1)^2 \leq 0$.

Lemma D.1. *Let $f(x) = \alpha e^{|x|/\gamma}$ with $\alpha, \gamma > 0$. Its Fenchel conjugate $f^*(u) = \sup_x\{ux - f(x)\}$ satisfies $f^*(u) \leq |u|\gamma(\ln(|u|\gamma/\alpha) - 1)$ for all u.*

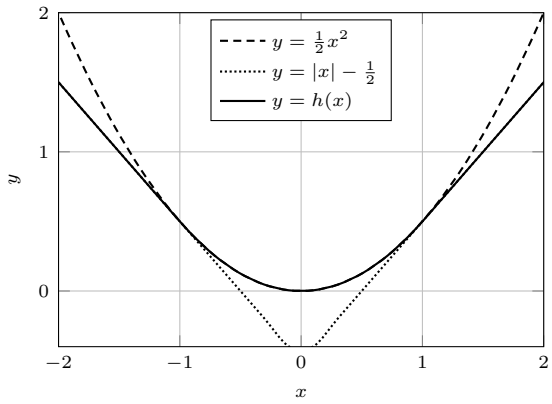

Figure 3. Function $h(x)$

Proof. Since $f(x)$ is symmetric in x, $\sup_x\{ux - f(x)\} = \sup_{x\geq 0}\{|u|x - f(x)\} = \sup_{x\geq 0} g(x)$, where $g(x) = |u|x - \alpha e^{x/\gamma}$. Setting the derivative of $g(x)$ to zero gives its unconstrained maximizer $x^* = \gamma \ln(|u|\gamma/\alpha)$, for which $g(x^*) = |u|\gamma(\ln(|u|\gamma/\alpha) - 1)$. The proof is finished by noticing that $\sup_{x\geq 0} g(x) \leq \sup_{x\in\mathbb{R}} g(x) = g(x^*)$. □

Lemma D.2. *For any $v \in \mathbb{R}$ and any $q \in [-1, 1]$:*

$$\exp\left\{\frac{1}{2}h\left(\frac{v-q}{1+q^2}\right) - \frac{1}{2}h(v) - \frac{1}{2}q^2\right\} \leq 1 - \frac{1}{2}q\,\mathrm{sgn}(v)\min\{|v|, 1\}$$

Proof. It suffices to prove the lemma for $v \geq 0$. Indeed, the inequality holds for some $v \geq 0$ and $q \in [-1, 1]$ if and only if it holds for $-v$ and $-q$. Denote:

$$\widetilde{v} = \frac{|v-q|}{\sqrt{1+q^2}}.$$

In this notation and with the assumption $v \geq 0$, the inequality translates to:

$$e^{\frac{1}{2}(h(\widetilde{v}) - h(v) - q^2)} \leq 1 - \frac{1}{2}q\min\{v, 1\}. \tag{14}$$

To prove (14), it suffices to show that:

$$h(\widetilde{v}) - h(v) - q^2 \leq -q\min\{v, 1\} - \frac{1}{2}(q\min\{v, 1\})^2, \tag{15}$$

because (15) together with $q \leq 1$ and inequality $e^{-x-x^2} \leq 1 - x$ for $x \leq \frac{1}{2}$ (see, e.g., (Cesa-Bianchi & Lugosi, 2006), Section A.1.2) implies (14).

We will split the proof of (15) into three sub-cases: (i) $v \leq 1$, (ii) $v \geq 1$ and $\widetilde{v} \geq 1$, (iii) $v \geq 1$ and $\widetilde{v} < 1$.

Case (i): $v \leq 1$. From the definition, $h(v) = \frac{1}{2}v^2$ and by (13) we upper bound $h(\widetilde{v}) \leq \frac{1}{2}\widetilde{v}^2$. Using $\widetilde{v} \leq |v - q|$ we have:

$$h(\widetilde{v}) - h(v) - q^2 \leq \frac{1}{2}\widetilde{v}^2 - \frac{1}{2}v^2 - q^2 \leq \frac{1}{2}(v-q)^2 - \frac{1}{2}v^2 - q^2 = -vq - \frac{1}{2}q^2 \leq -vq - \frac{1}{2}v^2q^2,$$

and since $\min\{v, 1\} = v$, this implies (15).

Case (ii): $v \geq 1$ and $\widetilde{v} \geq 1$. As $q \leq 1 \leq v$, we have $|v - q| = v - q$, and by the definition, $h(v) = v - \frac{1}{2}$, $h(\widetilde{v}) = \widetilde{v} - \frac{1}{2}$. Therefore:

$$h(\widetilde{v}) - h(v) - q^2 = \widetilde{v} - v - q^2 \leq v - q - v - q^2 \leq -q - q^2/2,$$

where in the first inequality we used $\widetilde{v} \leq |v - q| = v - q$. As $\min\{v, 1\} = 1$, this implies (15).

Case (iii): $v \geq 1$ and $\widetilde{v} < 1$. We have:

$$\widetilde{v} < 1 \iff \frac{(v-q)^2}{1+q^2} \leq 1 \iff v^2 - 2vq - 1 \leq 0 \iff v \leq q + \sqrt{1+q^2},$$

where the last equivalence follows from solving a quadratic inequality with respect to $v \geq 1$ for fixed q. We now note that function:

$$g(v) = h(\widetilde{v}) - h(v) - q^2 = \frac{1}{2}\widetilde{v}^2 - \left(v - \frac{1}{2}\right) - q^2 = \frac{(v-q)^2}{2(1+q^2)} - v - q^2 + \frac{1}{2}$$

is convex in v and hence it is maximized at the boundaries $\{1, q + \sqrt{1+q^2}\}$ of the allowed range of v. When $v = 1$, we have:

$$g(v) = \frac{(1-q)^2}{2(1+q^2)} - 1 - q^2 + \frac{1}{2} \leq \frac{1}{2}(1-q)^2 - q^2 - \frac{1}{2} = -q - \frac{1}{2}q^2,$$

whereas if $v = q + \sqrt{1+q^2}$, we have

$$g(v) = \frac{1}{2} - \left(q + \sqrt{1+q^2}\right) - q^2 + \frac{1}{2} \leq -q - q^2 \leq -q - \frac{1}{2}q^2,$$

so that $g(v) \leq -q - \frac{1}{2}q^2$ in the entire range of allowed values of v. As $\min\{v, 1\} = 1$, this implies (15). □

Before stating further results, we summarize the notation: for $i = 1, \ldots, d$ and $t = 1, \ldots, T$,

$$M_{t,i} = \max_{j \leq t} |x_{j,i}|, \quad G_{t,i} = -\sum_{j \leq t} g_j x_{j,i}, \quad S_{t,i}^2 = \sum_{j \leq t} (g_j x_{j,i})^2, \quad \eta_{t,i} = \epsilon - \sum_{j \leq t} g_t x_{t,i} w_{t,i},$$

with the convention $M_{0,i} = G_{0,i} = S_{0,i}^2 = 0$ and $\eta_{0,i} = \epsilon$. As before, we also use $\hat{S}_{t,i}^2 = S_{t,i}^2 + M_{t,i}^2$. The weight vector at trial t is given by:

$$w_{t,i} = \frac{\operatorname{sgn}(G_{t-1,i}) \min\left\{\frac{|G_{t-1,i}|}{\sqrt{S_{t-1,i}^2 + M_{t,i}^2}}, 1\right\}}{2\sqrt{S_{t-1,i}^2 + M_{t,i}^2}} \eta_{t-1,i} \tag{16}$$

as long as $M_{t,i} > 0$; if $M_{t,i} = 0$, we set $w_{t,i} = 0$.

Lemma D.3. *Define:*

$$\psi_{t,i}(x) = \begin{cases} e^{\frac{1}{2}h\left(\frac{x}{\hat{S}_{t,i}}\right)} & \text{for } \hat{S}_{t,i} \neq 0, \\ 1 & \text{for } \hat{S}_{t,i} = 0, \end{cases}$$

with $h(\cdot)$ defined in (12). For any $i = 1, \ldots, d$, let τ_i be the first trial in which $x_{t,i} \neq 0$. We have for any $= 1, \ldots, d$ and any $t = \tau_i, \ldots, T$:

$$\frac{\eta_{t,i}}{\eta_{t-1,i}} \geq \frac{\psi_{t,i}(G_{t,i})}{\psi_{t-1,i}(G_{t-1,i})} e^{-\delta_{t,i}},$$

where $\delta_{t,i} = \frac{(g_t x_{t,i})^2}{2(S_{t-1,i}^2 + M_{t,i}^2)}$

Proof. Fix i and $t \geq \tau_i$, and define $v = \frac{G_{t-1,i}}{\sqrt{S_{t-1,i}^2 + M_{t,i}^2}}$ and $q = \frac{g_t x_{t,i}}{\sqrt{S_{t-1,i}^2 + M_{t,i}^2}}$. As $|q| \leq \frac{|g_t x_{t,i}|}{M_{t,i}} \leq 1$, we can apply Lemma D.2 to such v and q, which gives:

$$e^{\frac{1}{2}h\left(\frac{v-q}{1+q^2}\right) - \frac{1}{2}h(v) - \frac{1}{2}q^2} \leq 1 - \frac{1}{2}q \operatorname{sgn}(v) \min\{|v|, 1\} \tag{17}$$

Using the definition of weight vector (16), we identify the right-hand side of (17) with $1 - \frac{g_t x_{t,i} w_{t,i}}{\eta_{t-1,i}} = \frac{\eta_{t,i}}{\eta_{t-1,i}}$. Since $\frac{1}{2}q^2 = \delta_{t,i}$ and $\frac{G_{t,i}}{\hat{S}_{t,i}} = \frac{v-q}{\sqrt{1+q^2}}$ (see the proof of Lemma C.3), we also identify the left-hand side of (17) with $\psi_{t,i}(G_{t,i}) e^{-\frac{1}{2}h(v)} e^{-\delta_{t,i}}$. Hence, (17) can be rewritten as:

$$\frac{\eta_{t,i}}{\eta_{t-1,i}} \geq \frac{\psi_{t,i}(G_{t,i})}{e^{\frac{1}{2}h(v)}} e^{-\delta_{t,i}},$$

and thus to prove the lemma, it suffices to show:

$$e^{\frac{1}{2}h(v)} \leq \psi_{t-1,i}(G_{t-1,i}). \tag{18}$$

When $t = \tau_i$, we have $v = 0$ as well as $G_{t-1,i} = 0$, and (18) holds as its both sides are equal to 1. For $t > \tau_i$, (18) reduces to $h(v) \leq h(G_{t-1,i}/\hat{S}_{t-1,i})$, which holds because:

$$|v| = \frac{|G_{t-1,i}|}{\sqrt{S_{t-1,i}^2 + M_{t,i}^2}} \leq \frac{|G_{t-1,i}|}{\sqrt{S_{t-1,i}^2 + M_{t-1,i}^2}} = \frac{|G_{t-1,i}|}{\hat{S}_{t-1,i}},$$

and $h(x) = h(|x|)$ is monotonic in $|x|$. □

We are now ready to prove Theorem 3.2, which we restate here for convenience:

Theorem. *For any $\boldsymbol{u} \in \mathbb{R}$ the regret of* ScInOL$_2$ *is upper-bounded by:*

$$R_T(\boldsymbol{u}) \leq d\epsilon + \sum_{i=1}^{d} 2|u_i|\hat{S}_{T,i}\left(\ln(3|u_i|\hat{S}_{T,i}^3\epsilon^{-1}/x_{\tau_i,i}^2) - 1\right),$$

where $\hat{S}_{T,i} = \sqrt{S_{T,i}^2 + M_{T,i}^2}$ and $\tau_i = \min\{t\colon |x_{t,i}| \neq 0\}$.

Proof. Fixing $i \in \{1, \ldots, d\}$, applying Lemma (C.3) for $t = \tau_i, \ldots, T$, and multiplying over trials gives:

$$\frac{\eta_{T,i}}{\eta_{\tau_i-1,i}} \geq \frac{\psi_{T,i}(G_{T,i})}{\psi_{\tau_i-1,i}(G_{\tau_i-1,i})}e^{-\Delta_{T,i}},$$

where we denoted $\Delta_{T,i} = \sum_{t=\tau_i}^{T} \delta_{t,i}$. From the definition of τ_i, we have $\eta_{\tau_i-1,i} = \epsilon$ and $\psi_{\tau_i-1,i} \equiv 1$. Using $\eta_{T,i} = \epsilon - \sum_{t\leq T} g_t x_{t,i} w_{t,i}$ we get:

$$\sum_{t=1}^{T} g_t x_{t,i} w_{t,i} \leq \epsilon - \epsilon\psi_{T,i}(G_{T,i})e^{-\Delta_{T,i}} \leq \epsilon - \epsilon e^{-\Delta_{T,i}+|G_{T,i}|/(2\hat{S}_{T,i})-\frac{1}{4}},$$

where we used (12) to bound $h(x) \geq |x| - \frac{1}{2}$. By (3),

$$\begin{aligned}
R_T(\boldsymbol{u}) &\leq \sum_{t=1}^{T} g_t \boldsymbol{x}_t^\top(\boldsymbol{w}_t - \boldsymbol{u}) = \sum_{i=1}^{d}\left(\sum_{t=1}^{T} g_t x_{t,i} w_{t,i} + G_{T,i}u_i\right) \\
&\leq d\epsilon + \sum_{i=1}^{d}\left(G_{T,i}u_i - \epsilon e^{-\Delta_{T,i}-\frac{1}{4}}e^{|G_{T,i}|/(2\hat{S}_{T,i})}\right) \\
&\leq d\epsilon + \sum_{i=1}^{d}\sup_x\left\{xu_i - \epsilon e^{-\Delta_{T,i}-\frac{1}{4}}e^{|x|/(2\hat{S}_{T,i})}\right\} \\
&\leq d\epsilon + \sum_{i=1}^{d} 2|u_i|\hat{S}_{T,i}\left(\ln\left(2\epsilon^{-1}|u_i|\hat{S}_{T,i}e^{\frac{1}{4}+\Delta_{T,i}}\right) - 1\right),
\end{aligned}$$

where in the last inequality we used Lemma D.1 for each i with $\alpha = \epsilon e^{-\Delta_{T,i}-\frac{1}{4}}$ and $\gamma = 2\hat{S}_{T,i}$. We will now show that

$$\Delta_{T,i} \leq \ln\left(\frac{\hat{S}_{T,i}^2}{x_{\tau_i,i}^2}\right), \tag{19}$$

which, together with $2e^{1/4} \leq 3$ will finish the proof. To prove (19), we use $M_{t,i}^2 \geq x_{t,i}^2 \geq (g_t x_{t,i})^2 = S_{t,i}^2 - S_{t-1,i}^2$ to get:

$$\delta_{t,i} = \frac{(g_t x_{t,i})^2}{2(S_{t-1,i}^2 + M_{t,i}^2)} \leq \frac{(g_t x_{t,i})^2}{S_{t-1,i}^2 + 2M_{t,i}^2} \leq \frac{(g_t x_{t,i})^2}{S_{t,i}^2 + M_{t,i}^2} = \frac{(M_{t,i}^2 + S_{t,i}^2) - (M_{t,i}^2 + S_{t-1,i}^2)}{S_{t,i}^2 + M_{t,i}^2}.$$

Using $\frac{a-b}{a} \leq \ln \frac{a}{b}$ for any $a \geq b > 0$ (which follows from the concavity of the logarithm):

$$\delta_{t,i} \leq \ln \frac{M_{t,i}^2 + S_{t,i}^2}{M_{t,i}^2 + S_{t-1,i}^2} \leq \ln \frac{M_{t+1,i}^2 + S_{t,i}^2}{M_{t,i}^2 + S_{t-1,i}^2},$$

where for $t = T$, we define $M_{T+1,i} = M_{T,i}$. Summing the above over trials $t = \tau_i, \ldots, T$:

$$\Delta_{T,i} = \sum_{t=\tau_i}^{T} \delta_{t,i} \leq \ln \frac{M_{T+1,i}^2 + S_{T,i}^2}{M_{\tau_i,i}^2 + S_{\tau_i-1,i}^2} = \ln \frac{M_{T,i}^2 + S_{T,i}^2}{x_{\tau_i,i}^2} = \ln \frac{\hat{S}_{T,i}^2}{x_{\tau_i,i}^2},$$

which was to be shown. □

E. Datasets

MNIST dataset is available at Yann Lecun's page. All other datasets are availableat the UCI repository. Scale is computed as a ratio of highest to lowest positive L_2 norms of features.

Name	features	records	classes	scale
Bank	53	41188	2	6.05E+05
Census	381	299285	2	1.81E+06
Covertype	54	581012	7	1.31E+06
Madelon	500	2600	2	1.09E+00
MNIST	728	70000	10	5.83E+03
Shuttle	9	58000	7	7.46E+00

Table 2. Short summary of datasets

F. Experiment: Classification Accuracy Plots

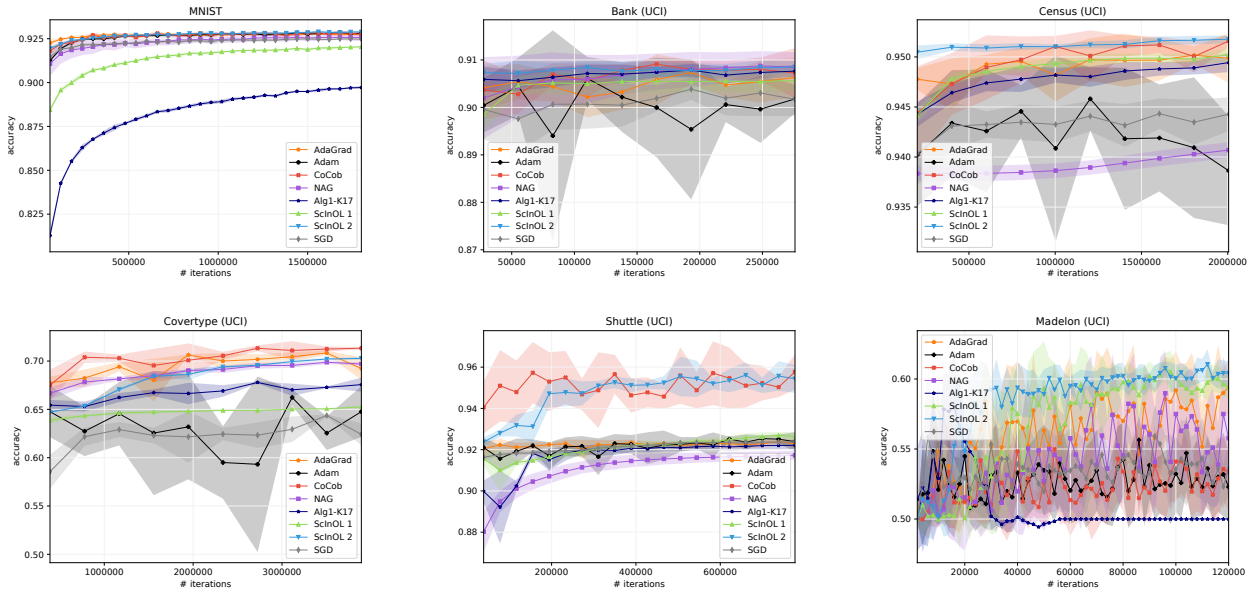

Figure 4. Accuracy results for linear classification experiments.

G. Multivariate Predictions

For simplicity, in the paper we focus on loss functions defined for real-valued predictions $\widehat{y} \in \mathbb{R}$. Sometimes, however, it is natural to consider a setup of multivariate predictions $\widehat{\boldsymbol{y}} \in \mathbb{R}^K$. For instance, the multinomial logistic loss (cross-entropy loss) is defined for $y \in \{1, \ldots, K\}$ as:

$$\ell(y, \widehat{\boldsymbol{y}}) = -\sum_{k=1}^{K} \mathbf{1}[y = k] \ln \sigma_k(\widehat{\boldsymbol{y}}) = -\widehat{y}_y + \ln\left(\sum_{k=1}^{K} e^{\widehat{y}_k}\right),$$

where $\sigma_k(\widehat{\boldsymbol{y}}) = \frac{e^{\widehat{y}_k}}{\sum_{j=1}^{K} e^{\widehat{y}_j}}$ is the soft-max transform.

We assume the multivariate losses $\ell_t(\widehat{\boldsymbol{y}}) = \ell(y_t, \widehat{\boldsymbol{y}})$ are convex and L-Lipschitz in the sense that the max-norm of subgradient $\nabla \ell_t(\widehat{\boldsymbol{y}})$ for any $\widehat{\boldsymbol{y}}$ is bounded, $\|\nabla \ell_t(\widehat{\boldsymbol{y}})\|_\infty \le L$ (which is satisfied with $L = 1$ by the multinomial logistic loss). We consider the class of comparators which are parameterized by $\boldsymbol{U} \in \mathbb{R}^{d \times K}$, a $d \times K$ parameter matrix, and the regret of the algorithms against $\boldsymbol{U}$ for a sequence of data $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^{T}$ is defined as:

$$R_T(\boldsymbol{U}) = \sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{y}}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{U}^\top \boldsymbol{x}_t).$$

Consider an algorithm which at trial t predicts with a weight matrix $\boldsymbol{W}_t \in \mathbb{R}^{d \times K}$, $\widehat{\boldsymbol{y}}_t = \boldsymbol{W}_t^\top \boldsymbol{x}_t$. Using the convexity of the loss, for any $\widehat{\boldsymbol{y}}, \widehat{\boldsymbol{y}}'$ and any t we have $\ell_t(\widehat{\boldsymbol{y}}') \ge \ell_t(\widehat{\boldsymbol{y}}) + \nabla \ell_t(\widehat{\boldsymbol{y}})^\top (\widehat{\boldsymbol{y}}' - \widehat{\boldsymbol{y}})$. Denoting $\nabla \ell_t(\widehat{\boldsymbol{y}}_t)$ by $\boldsymbol{g}_t = (g_{t,1}, \ldots, g_{t,K})$ with $g_{t,k} \in [-L, L]$ for all $k = 1, \ldots, K$, and using the bound above with $\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{y}}_t = \boldsymbol{W}_t^\top \boldsymbol{x}_t$ and $\widehat{\boldsymbol{y}}' = \boldsymbol{U}^\top \boldsymbol{x}_t$ we have:

$$R_T(\boldsymbol{U}) = \sum_{i=1}^{d} \sum_{k=1}^{K} \left(\sum_{t=1}^{T} g_{t,k} x_{t,i} (W_{t;i,k} - U_{i,k}) \right).$$

The regret decouples into a sum over individual coordinates and dimensions of the prediction vector, and the extension of our algorithms is now straightforward (see Algorithm (3) and (4) below). Also, the analysis can be carried out in full analogy to the univariate loss case resulting in the following bounds (for $L = 1$):

Theorem G.1. *For any $\boldsymbol{U} \in \mathbb{R}^{d \times K}$ the regret of* ScInOL_1 *is upper-bounded by:*

$$R_T(\boldsymbol{U}) \le \sum_{i=1}^{d} \sum_{k=1}^{K} \Big(2|U_{i,k}| \hat{S}_{T;i,k} \ln(1 + 2|U_{i,k}| \hat{S}_{T;i,k} \epsilon^{-1} T) + \epsilon(1 + \ln T) \Big)$$

where $\hat{S}_{T;i,k} = \sqrt{S_{T;i,k}^2 + M_{T;i}^2}$.

Theorem G.2. *For any $\boldsymbol{U} \in \mathbb{R}^{d \times K}$ the regret of* ScInOL_2 *is upper-bounded by:*

$$R_T(\boldsymbol{U}) \le dK\epsilon + \sum_{i=1}^{d} \sum_{k=1}^{K} 2|U_{i,k}| \hat{S}_{T;i,k} \left(\ln(3|U_{i,k}| \hat{S}_{T;i,k}^3 \epsilon^{-1} / x_{\tau_i,i}^2) - 1 \right),$$

where $\hat{S}_{T;i,k} = \sqrt{S_{T;i,k}^2 + M_{T;i}^2}$ and $\tau_i = \min\{t \colon |x_{t,i}| \ne 0\}$.

Algorithm 3: ScInOL$_1(\epsilon)$ for multivariate losses

Initialization : $S^2_{0;i,k}, G_{0;i,k}, M_{0;i} \leftarrow 0, \beta_{0;i,k} \leftarrow \epsilon \ (i = 1, \dots, d;\ k = 1, \dots, K)$
for $t = 1, \dots, T$ **do**
  Receive $\boldsymbol{x}_t \in \mathbb{R}^d$
  for $i = 1, \dots, d$ **do**
    $M_{t;i} \leftarrow \max\{M_{t-1;i}, |x_{t,i}|\}$
    for $k = 1, \dots, K$ **do**
      $\beta_{t;i,k} \leftarrow \min\{\beta_{t-1;i,k}, \epsilon(S^2_{t-1;i,k} + M^2_{t;i})/(x^2_{t,i}t)\}$
      $W_{t;i,k} = \frac{\beta_{t;i,k}\mathrm{sgn}(\theta_{t;i,k})}{2\sqrt{S^2_{t-1;i,k}+M^2_{t;i}}}\Big(e^{|\theta_{t;i,k}|/2} - 1\Big)$, where $\theta_{t;i,k} = \frac{G_{t-1;i,k}}{\sqrt{S^2_{t-1;i,k}+M^2_{t;i}}}$
  Predict with $\widehat{\boldsymbol{y}}_t = \boldsymbol{W}_t^\top \boldsymbol{x}_t$, receive loss $\ell_t(\widehat{\boldsymbol{y}}_t)$ and compute $\boldsymbol{g}_t = \nabla_{\widehat{\boldsymbol{y}}_t}\ell_t(\widehat{\boldsymbol{y}}_t)$
  for $i = 1, \dots, d$ **do**
    for $k = 1, \dots, K$ **do**
      $G_{t;i,k} \leftarrow G_{t-1;i,k} - g_{t,k}x_{t,i}$
      $S^2_{t;i,k} \leftarrow S^2_{t-1;i,k} + (g_{t,k}x_{t,i})^2$

Algorithm 4: ScInOL$_2(\epsilon)$ for multivariate losses

Initialization : $S^2_{0;i,k}, G_{0;i,k}, M_{0;i} \leftarrow 0, \eta_{0;i,k} \leftarrow \epsilon \ (i = 1, \dots, d;\ k = 1, \dots, K)$
for $t = 1, \dots, T$ **do**
  Receive $\boldsymbol{x}_t \in \mathbb{R}^d$
  for $i = 1, \dots, d$ **do**
    $M_{t;i} \leftarrow \max\{M_{t-1;i}, |x_{t,i}|\}$
    for $k = 1, \dots, K$ **do**
      $W_{t;i,k} = \frac{\mathrm{sgn}(\theta_{t;i,k})\min\{|\theta_{t;i,k}|,1\}}{2\sqrt{S^2_{t-1;i,k}+M^2_{t;i}}}\eta_{t-1;i,k}$, where $\theta_{t;i,k} = \frac{G_{t-1;i,k}}{\sqrt{S^2_{t-1;i,k}+M^2_{t;i}}}$
  Predict with $\widehat{\boldsymbol{y}}_t = \boldsymbol{W}_t^\top \boldsymbol{x}_t$, receive loss $\ell_t(\widehat{\boldsymbol{y}}_t)$ and compute $\boldsymbol{g}_t = \nabla_{\widehat{\boldsymbol{y}}_t}\ell_t(\widehat{\boldsymbol{y}}_t)$
  for $i = 1, \dots, d$ **do**
    for $k = 1, \dots, K$ **do**
      $G_{t;i,k} \leftarrow G_{t-1;i,k} - g_{t,k}x_{t,i}$
      $S^2_{t;i,k} \leftarrow S^2_{t-1;i,k} + (g_{t,k}x_{t,i})^2$
      $\eta_{t;i,k} \leftarrow \eta_{t-1;i,k} - g_{t,k}x_{t,i}w_{t,i,k}$