Supplementary material

A Proof of Lemma 1

First, to get rid of the absolute value in Eq. (3), we can expand the constraint $|\boldsymbol{w} \cdot \boldsymbol{x}_t - y_t| \leq \epsilon$ into two as follows:

$$\begin{aligned} \boldsymbol{w}_{t+1} = \underset{\boldsymbol{w} \in \mathbf{R}_+^m}{\text{argmin}} \quad & \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2 \\ \text{subject to} \quad & \boldsymbol{w} \cdot \boldsymbol{x}_t - y_t \leq \epsilon \\ & -\boldsymbol{w} \cdot \boldsymbol{x}_t + y_t \leq \epsilon. \end{aligned} \tag{16}$$

The Lagrangian of this optimization problem is:

$$\mathcal{L}(\boldsymbol{w}, \theta, \kappa, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2 + \theta(\boldsymbol{w} \cdot \boldsymbol{x}_t - y_t - \epsilon) + \kappa(-\boldsymbol{w} \cdot \boldsymbol{x}_t + y_t - \epsilon) - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{w},$$

where $\theta \in \mathbf{R}_+$, $\kappa \in \mathbf{R}_+$ and $\boldsymbol{\mu} \in \mathbf{R}_+^m$ are Lagrange multipliers. Differentiating the Lagrangian with respect to w_j and solving for zero gives the optimality condition:

$$w_j^* = w_{t,j} - (\theta^* - \kappa^*)x_{t,j} + \mu_j^* \ \forall j.$$

Additionally, the KKT complementary slackness conditions require that:

$$w_j^* \mu_j^* = 0 \ \forall j \tag{17}$$

$$\theta^*(\boldsymbol{w}^* \cdot \boldsymbol{x}_t - y_t - \epsilon) = 0 \tag{18}$$

$$\kappa^*(-\boldsymbol{w}^* \cdot \boldsymbol{x}_t + y_t - \epsilon) = 0. \tag{19}$$

If $w_j^* > 0$, Eq. (17) implies that $\mu_j^* = 0$. Therefore, $w_j^* = w_{t,j} + (\theta^* - \kappa^*)x_{t,j}$. To update the model, at least one weight must be changed. Therefore, $\theta^* \neq \kappa^*$. Using Eq. (18) and (19), this implies that either θ^* or κ^* is non-zero, but not both.

If $w_j^* = 0$, we must have $w_{t,j} - (\theta^* - \kappa^*)x_{t,j} + \mu_j^* = 0$. Next, we assume that $x_{t,j} > 0$ and $w_{t,j} > 0$ as otherwise the optimal solution is clearly $w_j^* = w_{t,j}$. Using these assumptions together with $\mu_j^* \geq 0$, we obtain that $\theta^* > \kappa^*$. Assume for a moment that $\kappa^* > 0$. Using the fact that the right-hand side of Eq. (19) must be zero, we obtain that $\boldsymbol{w}^* \cdot \boldsymbol{x}_t = y_t - \epsilon$. Injecting that in the right-hand side of Eq. (18), we obtain -2ϵ. This implies that $\theta^* = 0$, which contradicts $\theta^* > \kappa^*$. Therefore, $\kappa^* = 0$. Thus, a non-negativity constraint (i.e., $\mu_j^* > 0$) can only be effective if $\theta^* > 0$.

If $\theta^* > 0$, $\boldsymbol{w}^*$ needs to satisfy $\boldsymbol{w}^* \cdot \boldsymbol{x}_t - y_t \leq \epsilon$, which implies that $\boldsymbol{w}_t \cdot \boldsymbol{x}_t > y_t + \epsilon \geq y_t$. If $\kappa^* > 0$, $\boldsymbol{w}^*$ needs to satisfy $-\boldsymbol{w}^* \cdot \boldsymbol{x}_t + y_t \leq \epsilon$, which implies that $\boldsymbol{w}_t \cdot \boldsymbol{x}_t < y_t - \epsilon \leq y_t$.

Note that Eq. (5) and Eq. (6) contain *equality* constraints. This is because minimizing the loss $\frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2$ can obviously be achieved by satisfying the inequality constraints in Eq. (16) at their boundary. Lemma 1 is illustrated in Figure 3.

B Proof of Lemma 3

When $\hat{y}_t > y_t$, we need to solve:

$$\begin{aligned} \boldsymbol{w}_{t+1}, \xi^* = \underset{\boldsymbol{w} \in \mathbf{R}_+^m, \xi \in \mathbf{R}_+}{\text{argmin}} \quad & \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2 + C\xi \\ \text{subject to} \quad & \boldsymbol{w} \cdot \boldsymbol{x}_t = y_t + \epsilon + \xi \end{aligned} \tag{20}$$

After rearranging terms, the Lagrangian of the optimization problem in Eq. (20) is:

$$\mathcal{L}(\boldsymbol{w}, \theta, \lambda, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2 + \theta(\boldsymbol{w} \cdot \boldsymbol{x}_t - y_t - \epsilon) + \xi(C - \theta - \lambda) - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{w}, \tag{21}$$

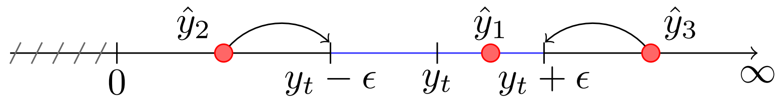

Figure 3: Illustration of Lemma 1. Given the true target y_t, we consider three points $\hat{y}_1$, $\hat{y}_2$ and $\hat{y}_3$, corresponding to three possible scenarios. In the first scenario, $\hat{y}_1$ is between $y_t - \epsilon$ and $y_t + \epsilon$ and satisfies $|\hat{y}_1 - y_t| \leq \epsilon$. The model thus does not need to be updated. In the second scenario, $\hat{y}_2 < y_t - \epsilon \leq y_t$. The smallest model change can be achieved by respecting the constraint at the boundary. This can only be done by making some model coefficients bigger and thus the non-negativity can be ignored. In the third scenario, $\hat{y}_3 > y_t + \epsilon \geq y_t$. Again, the smallest model change can be achieved by respecting the constraint at the boundary. However, this time, some model coefficients must be made smaller and thus the non-negativity must be respected.

where $\theta \in \mathbf{R}_+$, $\lambda \in \mathbf{R}_+$ and $\boldsymbol{\mu} \in \mathbf{R}^m_+$. In the following, we denote the primal optimal point by $(\boldsymbol{w}^*, \xi^*)$ and the dual optimal point by $(\theta^*, \lambda^*, \boldsymbol{\mu}^*)$. We denote the root of f_t by θ_f. Differentiating Eq. (21) with respect to ξ and setting the result to zero gives $C - \theta^* - \lambda^* = 0$. Using $\lambda^* \geq 0$, we obtain $\theta^* \leq C$. Assume that $y_t + \epsilon > 0$ (when $y_t + \epsilon = 0$, it can easily be shown that $\theta^* = \min(C, \theta_u)$). We know that f_t is strictly negative in $(\theta_f, +\infty)$ and strictly positive in $(-\infty, \theta_f)$. Thus, if $f_t(C) < 0$, then $\theta_f < C$. We can thus readily choose $\theta^* = \theta_f$. On the other hand, if $f_t(C) > 0$, then $\theta_f > C$. Using our previous result $\theta^* \leq C$, we thus get $\theta^* < \theta_f$, and therefore $f_t(\theta^*) > 0$. We know from the equality constraint in Eq. (20) that $\boldsymbol{w}^* \cdot \boldsymbol{x}_t = \max(\boldsymbol{w}_t - \theta^* \boldsymbol{x}_t, 0) \cdot \boldsymbol{x}_t = y_t + \epsilon + \xi^* \Leftrightarrow f_t(\theta^*) = \xi^*$. Therefore, $\xi^* > 0$. Following the KKT complementary slackness condition $\xi^* \lambda^* = 0$, we obtain $\lambda^* = 0$. Finally, using $C - \theta^* - \lambda^* = 0$, we obtain $\theta^* = C$. When $f_t(C) = 0$, clearly $\theta^* = C$. In summary, if $f_t(C) \geq 0$ then $\theta^* = C$, otherwise $\theta^* = \theta_f$.

C Proof of Theorem 1

To prove regret bounds for our algorithms, we adopt a primal-dual view of online learning [Shalev-Shwartz and Singer, 2007]. In this view, we define an optimization problem and we cast online learning as the task of incrementally increasing the dual objective function. The amount by which the dual increases serves as the notion of progress. In this paper, we consider the following optimization problem

$$\underset{\boldsymbol{w} \in \mathbf{R}^m_+}{\text{minimize}}\ \mathcal{P}(\boldsymbol{w}) = C \sum_{i=1}^{l} \ell^\epsilon(\boldsymbol{w}; (\boldsymbol{x}_i, y_i)) + \mathcal{R}(\boldsymbol{w}),$$

where ℓ^ϵ is defined in Eq. (4) and $\mathcal{R}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2$. The dual of this problem is

$$\underset{\substack{\boldsymbol{\kappa} \in [0,C]^l \\ \boldsymbol{\theta} \in [0,C]^l}}{\text{maximize}}\ \mathcal{D}(\boldsymbol{\kappa}, \boldsymbol{\theta}) = \sum_{i=1}^{l} (\kappa_i - \theta_i) y_i - \epsilon \sum_{i=1}^{l} (\kappa_i + \theta_i) - \mathcal{R}^* \left(\sum_{i=1}^{l} (\kappa_i - \theta_i) \boldsymbol{x}_i \right),$$

where $\mathcal{R}^*$ is the Fenchel conjugate of the function $\mathcal{R}$, as defined as follows

$$\begin{aligned} \mathcal{R}^*(\boldsymbol{u}) &= \sup_{\boldsymbol{w} \in \mathbf{R}^m_+} \boldsymbol{w} \cdot \boldsymbol{u} - \mathcal{R}(\boldsymbol{w}) \\ &= \frac{1}{2} \| \max(\boldsymbol{u}, 0) \|^2. \end{aligned}$$

Let κ_i^t and θ_i^t be the dual variables on round t. The primal-dual relationship is given by

$$\begin{aligned} \boldsymbol{w}_t &= \underset{\boldsymbol{w} \in \boldsymbol{R}^m_+}{\text{argmax}}\ \boldsymbol{w} \cdot \boldsymbol{u}_t - \mathcal{R}(\boldsymbol{w}) \\ &= \nabla \mathcal{R}^*(\boldsymbol{u}_t) \\ &= \max(\boldsymbol{u}_t, 0), \end{aligned} \tag{22}$$

where $\nabla \mathcal{R}^*$ is the gradient of $\mathcal{R}^*$ and $\boldsymbol{u}_t = \sum_{i=1}^{l} (\kappa_i^t - \theta_i^t) \boldsymbol{x}_i$. $\nabla \mathcal{R}^*(\cdot)$ is sometimes called link function. Let $\mathcal{P}^*$ be the primal optimal objective value. From the weak duality theorem, we immediately obtain the lower bound

$\forall t\ \mathcal{D}(\boldsymbol{\kappa}^t, \boldsymbol{\theta}^t) \le \mathcal{P}^*$, that is

$$\forall t\ \forall \boldsymbol{w} \in \mathbf{R}_+^m \quad \mathcal{D}(\boldsymbol{\kappa}^t, \boldsymbol{\theta}^t) \le C \sum_{i=1}^{l} \ell^\epsilon(\boldsymbol{w}; (\boldsymbol{x}_i, y_i)) + \mathcal{R}(\boldsymbol{w}). \tag{23}$$

Let $\Delta_t = \mathcal{D}(\boldsymbol{\kappa}^{t+1}, \boldsymbol{\theta}^{t+1}) - \mathcal{D}(\boldsymbol{\kappa}^t, \boldsymbol{\theta}^t)$. Assuming that $\boldsymbol{\kappa}^1 = \mathbf{0}$ and $\boldsymbol{\theta}^1 = \mathbf{0}$, we have $\mathcal{D}(\boldsymbol{\kappa}^1, \boldsymbol{\theta}^1) = 0$. Thus, we obtain

$$\mathcal{D}(\boldsymbol{\kappa}^{l+1}, \boldsymbol{\theta}^{l+1}) = \sum_{t=1}^{l} \Delta_t, \tag{24}$$

since $\sum_{t=1}^{l} \Delta_t$ is a telescoping sum. Next, we first prove a lower-bound on Δ_t for a simplified version of NN-PA-I, which sets either κ_i^t or θ_i^t to $\min(\frac{\ell_t^\epsilon}{\|\boldsymbol{x}\|^2}, C)$ if $i = t$. Otherwise, it keeps all other dual variables fixed. Let $s_t = \text{sign}(y_t - \boldsymbol{w}_t \cdot \boldsymbol{x}_t)$ and $\alpha_t = \min(\frac{\ell_t^\epsilon}{\|\boldsymbol{x}\|^2}, C)$. Assume t is a round on which an update occurred. Then, the dual increase is

$$\Delta_t = s_t \alpha_t y_t - \epsilon \alpha_t - \mathcal{R}^*(\boldsymbol{u}_t + s_t \alpha_t \boldsymbol{x}_t) + \mathcal{R}^*(\boldsymbol{u}_t). \tag{25}$$

Next, we make use of the quadratic bound lemma [Mangasarian, 2002], which states that if f has a Lipschitz continuous gradient ∇f with constant K, then

$$f(\boldsymbol{v}) - f(\boldsymbol{s}) - \nabla f(\boldsymbol{s}) \cdot (\boldsymbol{v} - \boldsymbol{s}) \le \frac{K}{2} \|\boldsymbol{v} - \boldsymbol{s}\|^2.$$

It is easy to verify that $\mathcal{R}^*$ has Lipschitz continuous gradient $\nabla \mathcal{R}^*$ with constant 1. By using $\boldsymbol{v} = \boldsymbol{u}_t + s_t \alpha_t \boldsymbol{x}_t$ and $\boldsymbol{s} = \boldsymbol{u}_t$, we obtain

$$\mathcal{R}^*(\boldsymbol{u}_t + s_t \alpha_t \boldsymbol{x}_t) \le \mathcal{R}^*(\boldsymbol{u}_t) + \nabla \mathcal{R}^*(\boldsymbol{u}_t) \cdot (s_t \alpha_t \boldsymbol{x}_t) + \frac{1}{2} \alpha_t^2 \|\boldsymbol{x}_t\|^2.$$

Plugging the result in Eq. (25), we obtain

$$\begin{aligned}
\Delta_t &= s_t \alpha_t y_t - \epsilon \alpha_t - \mathcal{R}^*(\boldsymbol{u}_t + s_t \alpha_t \boldsymbol{x}_t) + \mathcal{R}^*(\boldsymbol{u}_t) \\
&\ge \alpha_t (s_t y_t - \epsilon - s_t \nabla \mathcal{R}^*(\boldsymbol{u}_t) \cdot \boldsymbol{x}_t) - \frac{1}{2} \alpha_t^2 \|\boldsymbol{x}_t\|^2 \\
&= \alpha_t (s_t y_t - \epsilon - s_t \boldsymbol{w}_t \cdot \boldsymbol{x}_t) - \frac{1}{2} \alpha_t^2 \|\boldsymbol{x}_t\|^2 \\
&= \alpha_t (|\boldsymbol{w}_t \cdot \boldsymbol{x}_t - y_t| - \epsilon) - \frac{1}{2} \alpha_t^2 \|\boldsymbol{x}_t\|^2 \\
&= \alpha_t \ell_t^\epsilon - \frac{1}{2} \alpha_t^2 \|\boldsymbol{x}_t\|^2,
\end{aligned}$$

where in the third line we used Eq. (22). Following Shalev-Shwartz and Singer [2007, Section 5], we define a "mitigating" function μ_ρ

$$\begin{aligned}
\mu_\rho(\ell) &= \frac{1}{C} \left(\min(\frac{\ell}{\rho}, C) \ell - \frac{1}{2} \min(\frac{\ell}{\rho}, C)^2 \rho \right) \\
&= \frac{1}{C\rho} \min(\ell, C\rho) \left(\ell - \frac{1}{2} \min(\ell, C\rho) \right).
\end{aligned}$$

If we choose $\rho = \max_i \|\boldsymbol{x}_i\|^2$, we obtain

$$\Delta_t \ge C \mu_\rho \left(\ell^\epsilon(\boldsymbol{w}_t; (\boldsymbol{x}_t, y_t)) \right). \tag{26}$$

The rest of the proof follows almost exactly Shalev-Shwartz and Singer [2007, Section 5] and is given for completeness. Combining Eq. (23), Eq. (24) and Eq. (26), we get

$$\sum_{t=1}^{l} \mu_\rho \left(\ell^\epsilon(\boldsymbol{w}_t; (\boldsymbol{x}_t, y_t)) \right) \le \sum_{t=1}^{l} \ell^\epsilon(\boldsymbol{w}; (\boldsymbol{x}_t, y_t)) + \frac{\mathcal{R}(\boldsymbol{w})}{C}.$$

By dividing both sides by l and using the fact that μ_ρ is convex, we obtain

$$\mu_\rho\left(\frac{1}{l}\sum_{t=1}^{l}\ell^\epsilon(\boldsymbol{w}_t;(\boldsymbol{x}_t,y_t))\right) \leq \frac{1}{l}\sum_{t=1}^{l}\mu_\rho(\ell^\epsilon(\boldsymbol{w}_t;(\boldsymbol{x}_t,y_t))) \leq \frac{1}{l}\sum_{t=1}^{l}\ell^\epsilon(\boldsymbol{w};(\boldsymbol{x}_t,y_t)) + \frac{\mathcal{R}(\boldsymbol{w})}{lC}$$

Since $\mu_\rho^{-1}(\ell)$ is defined and monotonically increasing when $\ell \geq 0$, we can apply it on both sides. Moreover, it can be verified that $\mu_\rho^{-1}(\ell) \leq \ell + \frac{C\rho}{2}$. After rearranging terms, we thus obtain

$$\forall \boldsymbol{w} \in \mathbf{R}_+^m \quad \sum_{t=1}^{l}\ell^\epsilon(\boldsymbol{w}_t;(\boldsymbol{x}_t,y_t)) - \sum_{t=1}^{l}\ell^\epsilon(\boldsymbol{w};(\boldsymbol{x}_t,y_t)) \leq \frac{\mathcal{R}(\boldsymbol{w})}{C} + \frac{lC\rho}{2}.$$

This concludes the proof for the approximate version of NN-PA-I. For the regular version, we note that the dual increase must be by definition at least as large as for the approximate version. Therefore, its regret bound must be at least as good.

D Solving problem (6) exactly by pivot algorithm

When $\boldsymbol{x}_t$ is a vector of all ones and $y_t + \epsilon = 1$, the optimization problem defined in Eq. (6) reduces to the well-known problem of Euclidean projection onto the standard simplex. This problem was well-studied in the literature; for example, Duchi et al. [2008] adapt the randomized pivot algorithm for median finding [Cormen et al., 2001] to this problem. We now derive a similar pivot algorithm for solving the optimization problem (6). Our algorithm includes Duchi et al.'s algorithm as a special case and has expected $O(m)$ complexity.

First, we extend lemmas 2 and 3 of Shalev-Shwartz and Singer [2006] to the more general problem of Eq. (6). The two lemmas below show that we can easily solve (6) if we sort the elements of $\boldsymbol{w}_t = [w_{t,1},\dots,w_{t,m}]$ and $\boldsymbol{x}_t = [x_{t,1},\dots,x_{t,m}]$.

Lemma 4 *Let $\boldsymbol{w}^* = [w_1^*,\dots,w_m^*]$ be the optimal solution to the minimization problem in Eq.* (6)*. Let s and j be two indices such that $\frac{w_{t,s}}{x_{t,s}} > \frac{w_{t,j}}{x_{t,j}}$. If $w_s^* = 0$, then w_j^* must be zero as well.*

Proof of Lemma 4 is most similar to the proof of Lemma 2 of Shalev-Shwartz and Singer [2006], and is thus omitted. Let I be the sequence of indices such that $\frac{w_{t,I_1}}{x_{t,I_1}} \geq \frac{w_{t,I_2}}{x_{t,I_2}} \geq \dots \geq \frac{w_{t,I_m}}{x_{t,I_m}}$. Let $\bar{\boldsymbol{w}}^t$ and $\bar{\boldsymbol{x}}^t$ be the vectors $\boldsymbol{w}_t$ and $\boldsymbol{x}_t$ sorted by I. Lemma 4 implies that there exists a positive number ρ^* such that:

$$\boldsymbol{w}^* \cdot \boldsymbol{x}_t = \sum_{j=1}^{m}\max(w_{t,j} - \theta^* x_{t,j}, 0)x_{t,j} = \sum_{j=1}^{\rho^*}(\bar{w}_{t,j} - \theta^*\bar{x}_{t,j})\bar{x}_{t,j} = z_t,$$

where $z_t = y_t + \epsilon$. From the above, we easily obtain $\theta^* = \theta(\rho^*)$, where

$$\theta(\rho) = \frac{(\sum_{j=1}^{\rho}\bar{w}_{t,j}\bar{x}_{t,j}) - z_t}{\sum_{j=1}^{\rho}(\bar{x}_{t,j})^2}. \quad (27)$$

Therefore, the problem of finding θ^* reduces to the problem of finding ρ^*. The next lemma offers a simple solution.

Lemma 5 *Let $\boldsymbol{w}^*$ be the optimal solution to the minimization problem in Eq.* (6)*. Let $\bar{\boldsymbol{w}}_t$ and $\bar{\boldsymbol{x}}_t$ be the vectors $\boldsymbol{w}_t$ and $\boldsymbol{x}_t$ sorted by I. Then the number of strictly positive elements in $\boldsymbol{w}^*$ is*

$$\rho^* = \max_j\{j : \bar{w}_{t,j} - \theta(j)\bar{x}_{t,j} > 0\}.$$

Algorithm 3 Randomized pivot algorithm for solving Eq. (6)

Input: $\boldsymbol{w}_t$, $\boldsymbol{x}_t$, y_t, ϵ
Initialize $U \leftarrow \{j : x_{t,j} > 0\}$, $s \leftarrow 0$, $\sigma \leftarrow 0$
Set $z_t = y_t + \epsilon$
while $U \neq \emptyset$ **do**
  Pick $k \in U$ at random
  Partition U:
    $G \leftarrow \{j \in U : \frac{w_{t,j}}{x_{t,j}} \geq \frac{w_{t,k}}{x_{t,k}}\}$
    $L \leftarrow \{j \in U : \frac{w_{t,j}}{x_{t,j}} < \frac{w_{t,k}}{x_{t,k}}\}$
  Compute:
    $\Delta s = \sum_{j \in G} w_{t,j} x_{t,j}$
    $\Delta \sigma = \sum_{j \in G} (x_{t,j})^2$
    $\theta = \frac{s + \Delta s - z_t}{\sigma + \Delta \sigma}$
  if $w_{t,k} - \theta x_{t,k} > 0$ **then**
    Update $s \leftarrow s + \Delta s$, $\sigma \leftarrow \sigma + \Delta \sigma$
    Set $U \leftarrow L$
  else
    Set $U \leftarrow G \setminus \{k\}$
  end if
end while
Output: θ

Lemma 5 shows that finding ρ^* (and therefore, θ^*) can easily be done provided that $\bar{\boldsymbol{w}}^t$ and $\bar{\boldsymbol{x}}^t$ are available. However, this requires obtaining I (i.e., sorting $\frac{w_{t,j}}{x_{t,j}}$ $\forall x_{t,j} > 0$), which unfortunately takes $O(m \log m)$ time complexity.

Building upon Lemma 4 and Lemma 5, we can extend Duchi et al.'s efficient pivot algorithm. The procedure is outlined in Algorithm 3. The algorithm identifies ρ^* and the pivot value $\bar{w}_{t,\rho^*}$ without sorting $\boldsymbol{w}_t$ and $\boldsymbol{x}_t$ thanks to a divide and conquer procedure which at each iteration eliminates elements shown to be strictly smaller than $\bar{w}_{t,\rho^*}$. While doing so, it also accumulates the sums s (numerator) and σ (denominator) needed to compute θ^* from Eq. (27). The algorithm has *expected* linear time complexity with respect to m.