Supplementary Material to Confusion-Based Online Learning and a Passive-Aggressive Scheme

Liva Ralaivola
QARMA, Laboratoire d'Informatique Fondamentale de Marseille
Aix-Marseille University, France
liva.ralaivola@lif.univ-mrs.fr

1 Proof of COPA's update procedure

A bit of notation. Input space is $\mathcal{X} \doteq \mathsf{R}^d$, $W \doteq [w_1 \cdots w_Q]$; $\mathbf{1} \doteq [1 \cdots 1]^\top$ of the appropriate dimension (always clear from context). Therefore, $W\mathbf{1} = \sum_{q=1,\dots,Q} w_q$.

1.1 Primal and Dual Problems

Ultimate goal: we want to solve the following problem, for p (and thus (x, y)) fixed,

$$\min_{W, W\mathbf{1}=0} F(W) \doteq \frac{1}{2}\sum_{q=1}^{Q} \left\|w_q - w_q^t\right\|^2 + \frac{C}{2}\sum_{q\neq p} \left|\langle w_q, x\rangle + \Delta\right|_+^2, \tag{1}$$

where $\Delta > 0$ (it is $\Delta = 1/(Q-1)$ in the main text). Thich can be equivalently written as

$$\min_{W,\boldsymbol{\xi}} G(W,\boldsymbol{\xi}) \doteq \frac{1}{2}\sum_{q=1}^{Q} \left\|w_q - w_q^t\right\|^2 + \frac{C}{2}\sum_{q\neq p} \xi_q^2 \tag{2}$$

$$\text{s.t. } \sum_{q=1}^{Q} w_q = \mathbf{0} \;\wedge\; \xi_q \geq \langle w_q, x\rangle + \Delta,\; q \neq y. \tag{3}$$

Here, 'equivalently' means that a solution W^* of the former optimization problem is also a solution of the latter (and *vice versa*). The optimal slack variables $\boldsymbol{\xi}^*$ are then such that

$$\xi_q^* = \left|\langle w_q^*, x\rangle + \Delta\right|_+,\; q \neq y.$$

To solve this optimization problem, we may introduce the Lagrangian of the previous problem:

$$L(W,\boldsymbol{\xi},\boldsymbol{\alpha}) = G(W,\boldsymbol{\xi}) - \sum_{q\neq y} \alpha_q \left[\xi_q - \langle w_q, x\rangle - \Delta\right] - \boldsymbol{\lambda}^\top W\mathbf{1}, \tag{4}$$

where $\boldsymbol{\lambda} \in \mathsf{R}^d$ and $\alpha_q \geq 0$, for $q \neq y$.

Taking derivatives of L with respect to the primal variables W and $\boldsymbol{\xi}$ and making the gradient be zero (a necessary condition on L for the primal variables to be optimal) gives:

$$\nabla_{w_q} L = w_q - w_q^t + \alpha_q x - \boldsymbol{\lambda} = 0,\; q \neq y \tag{5}$$

$$\nabla_{w_y} L = w_y - w_y^t - \boldsymbol{\lambda} = 0 \tag{6}$$

$$\nabla_{\xi_q} L = \alpha_q - C\xi_q \tag{7}$$

Or, stated otherwise,

$$w_q = w_q^t - \alpha_q x + \boldsymbol{\lambda},\; q = 1,\dots,Q \tag{8}$$

$$\alpha_q = C\xi_q, \tag{9}$$

where we have introduced a Lagrangian multiplier α_y that is *clamped* to 0 —this allows us to lighten the notation by not having to write $q \neq y$ when referring to index q.

Note, otherwise, that the Karush-Kuhn-Tucker optimality conditions give that, for all $q \neq y$

$$\alpha_q \left[\xi_q - \langle w_q, x\rangle - \Delta\right] = 0. \tag{10}$$

Summing the Q equations in (8), using the fact that $\sum_q w_q^t = \mathbf{0}$ and that we require $\sum_q w_q = \mathbf{0}$ for the new vectors that we are computing leads to:

$$\sum_q w_q = \sum_q w_q^t - \sum_q \alpha_q x - Q\boldsymbol{\lambda} \Leftrightarrow \mathbf{0} = \mathbf{0} - \sum_q \alpha_q x + Q\boldsymbol{\lambda} \tag{11}$$

$$\Leftrightarrow \boldsymbol{\lambda} = \frac{s_\alpha}{Q} x, \tag{12}$$

where, we have introduced the notation s_α for the sum of the α_q's:

$$s_\alpha \doteq \sum_{q=1}^{Q} \alpha_q = \boldsymbol{\alpha}^\top \mathbf{1}. \tag{13}$$

Henceforth the necessary condition (8) for W to be optimal rewrites as

$$w_q = w_q^t - \left(\alpha_q - \frac{s_\alpha}{Q}\right) x, \; q = 1, \ldots, Q \tag{14}$$

After some algebra, replacing W and $\boldsymbol{\xi}$ in the Lagrangian (4) thanks to Equations (9) and (14) allow us to get the dual objective $H(\boldsymbol{\alpha})$ of (1):

$$H(\boldsymbol{\alpha}) \doteq -\frac{1}{2}\left(\|x\|^2 + \frac{1}{C}\right)\sum_{q=1}^{Q} \alpha_q^2 + \frac{1}{2}\frac{\|x\|^2}{Q}\left(\sum_{q=1}^{Q} \alpha_q\right)^2 + \sum_{q=1}^{Q} \alpha_q \ell_q^t \tag{15}$$

where, for the sake of readability, the following notation is introduced:

$$\ell_q^t \doteq \langle w_q^t, x\rangle + \Delta \tag{16}$$

$$\kappa \doteq \frac{1}{C} + \|x\|^2. \tag{17}$$

Given the convexity of optimization problems (1) and (2), the solution $\boldsymbol{\alpha}^*$ of the convex optimization problem

$$\max_{\boldsymbol{\alpha}} H(\boldsymbol{\alpha}) \quad \text{s.t. } \alpha_y = 0 \;\wedge\; \alpha_q \geq 0, q \neq y \tag{18}$$

provides a solution W^* of (1) thanks to (14) through

$$w_q^* = w_q^t - \left(\alpha_q^* - \frac{1}{Q}\sum_{q=1}^{Q} \alpha_q^*\right) x, \; q = 1, \ldots, Q. \tag{19}$$

The following lemma shows that the dual objective H given by (15) is strictly concave in $\boldsymbol{\alpha}$: the dual optimization problem (18) therefore admits a unique maximum $\boldsymbol{\alpha}^*$, and it is thus valid to refer to $\boldsymbol{\alpha}^*$ as *the* optimal solution of (18).

Lemma 1. *The dual objective H* (15) *is strictly concave and optimization problem* (18) *admits* a unique *maximizer $\boldsymbol{\alpha}^*$.*

Proof. It is sufficient to show that the Hessian of $-H$ is strictly positive, i.e. that is only has positive eigenvalues. Rewriting things in matrix form, and leaving the linear part of $-H$ aside, this means it is sufficient to show that the application

$$R : \boldsymbol{\alpha} \mapsto R(\boldsymbol{\alpha}) \doteq \boldsymbol{\alpha}^\top \left(\frac{1}{C}\mathbb{I} + \|x\|^2 \left(\mathbb{I} - \frac{1}{Q}\mathbf{1}\mathbf{1}^\top\right)\right) \boldsymbol{\alpha}$$

is strictly convex. Observing that

$$\left(\mathbb{I} - \mathbf{1}\mathbf{1}^\top/Q\right)^2 = \left(\mathbb{I} - \mathbf{1}\mathbf{1}^\top/Q\right)$$

tells you that $(\mathbb{I} - \mathbf{1}\mathbf{1}^\top/Q)$ is a projection operator, and that its only eigenvalues are therefore 0 and 1 (see, e.g. [2]). Hence, since the only eigenvalue of $\mathbb{I}/C$ is obviously $1/C$, the eigenvalues of the Hessian of R are $1/C$ and $1 + 1/C$, and R is stricly convex.

This leads to the fact that $-H$ (adding the linear —convex— term to R) is strictly convex as well. The domain over which $-H$ has to be minimized is made of nonnegative constraints only and is therefore convex: minimizing $-H$ over the domain is therefore a (strict) convex optimization problem and it admits a *unique* solution, $\boldsymbol{\alpha}^*$. □

1.2 Families $(\boldsymbol{\alpha}(\mathcal{I}))_{\mathcal{I}}$ and $(W(\mathcal{I}))_{\mathcal{I}}$

We now show that finding $\boldsymbol{\alpha}^*$ (and therefore W^*) might be done in constant time, without recoursing to any optimization procedure. The idea of the proof is similar to what is encountered when performing projection on mixed-norm balls [], and more closely related to the work of [1].

In order to state the main theorem of this section, it is handy to introduce the family $(\boldsymbol{\alpha}(\mathcal{I}))_{\mathcal{I}\subseteq\mathcal{Y}\backslash\{y\}}$ of vectors defined as follows.

Definition 1 (Family $(\boldsymbol{\alpha}(\mathcal{I}))_{\mathcal{I}\subseteq\mathcal{Y}\backslash\{y\}}$)**.** The family $(\boldsymbol{\alpha}(\mathcal{I}))_{\mathcal{I}\subseteq\mathcal{Y}\backslash\{y\}}$ is such that the components $\alpha_q(\mathcal{I})$ of $\boldsymbol{\alpha}(\mathcal{I})$ verify:

$$\alpha_q(\mathcal{I}) \doteq \begin{cases} \dfrac{1}{\kappa}\left(\ell_q^t + \dfrac{\|x\|^2}{Q} s_\alpha(\mathcal{I})\right) & \text{if } q \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where, $I \doteq |\mathcal{I}|$ being the size of $\mathcal{I}$,

$$s_\alpha(\mathcal{I}) \doteq \frac{Q}{\kappa Q - I\|x\|^2} \sum_{q\in\mathcal{I}} \ell_q^t. \tag{21}$$

Remark 1. A few observations may be issued regarding $\boldsymbol{\alpha}(\mathcal{I})$. First, the denominator appearing in (21) *cannot* be zero: it suffices to recall the definition of κ in (17) and the fact that I is striclty lower than Q. Then, with no additional constraint on $\mathcal{I}$, there is no reason for the $\alpha_q(\mathcal{I}), q \in \mathcal{I}$ not to be negative —as we shall see, we will later on build a set $\mathcal{I}^*$ such that $\alpha_q(\mathcal{I}^*) > 0$ whenever $q \in \mathcal{I}^*$. Finally, for $p, q \in \mathcal{I}$, if $\ell_p^t \geq \ell_q^t$ then $\alpha_p(\mathcal{I}) \geq \alpha_q(\mathcal{I})$ (this directly comes from (20)).

The family $(\boldsymbol{\alpha}(\mathcal{I}))_{\mathcal{I}}$ directly induces a family $(W(\mathcal{I}))_{\mathcal{I}}$ as follows.

Definition 2 (Family $(W(\mathcal{I}))_{\mathcal{I}\subseteq\mathcal{Y}\backslash\{y\}}$)**.** The family $(W(\mathcal{I}) = [w_1(\mathcal{I})\cdots w_Q(\mathcal{I})])_{\mathcal{I}\subseteq\mathcal{Y}\backslash\{y\}}$ is deduced from $(\boldsymbol{\alpha}(\mathcal{I}))_{\mathcal{I}\subseteq\mathcal{Y}\backslash\{y\}}$ as follows:

$$w_q(\mathcal{I}) \doteq w_q^t - \left(\alpha_q(\mathcal{I}) - \frac{1}{Q} s_\alpha(\mathcal{I})\right) x,\ q = 1, \ldots, Q. \tag{22}$$

1.3 Efficient Updates

From now on, we assume we have at hand a permutation $\sigma : \{1, \ldots, Q-1\} \to \mathcal{Y}\backslash\{y\}$ such that

$$\ell_{\sigma(1)}^t \geq \cdots \geq \ell_{\sigma(Q-1)}^t.$$

The main theorem of this section follows.

Theorem 1. *Let I^* be the largest index $I \in \{1, \ldots, Q-1\}$ such that*

$$\ell_{\sigma(I)}^t + \frac{\|x\|^2}{\kappa Q - (I-1)\|x\|^2} \sum_{q=1}^{I-1} \ell_{\sigma(q)}^t > 0. \tag{23}$$

If $\mathcal{I}^*$ is set to $\mathcal{I}^* \doteq \{\sigma(1), \ldots, \sigma(I^*)\}$, then $\boldsymbol{\alpha}^* \doteq \boldsymbol{\alpha}(\mathcal{I}^*)$ is the solution of problem (18)*, and*

$$w_q^* = w_q^t - \left(\alpha_q(\mathcal{I}^*) - \frac{1}{Q}\sum_{q=1}^{I^*} \alpha_q(\mathcal{I}^*)\right) x,\ q = 1, \ldots, Q \tag{24}$$

is the solution of problem (2)*, i.e. it provides us with the update equation to perform learning.*

The proof of this theorem develops upon two ideas, that are established in Lemma 2 and Lemma 3. Lemma 2 establishes the analytic form of $\boldsymbol{\alpha}^*$, by proving that it is an element of the family $(\alpha(\mathcal{I}))_{\mathcal{I}}$ introduced before. The question raised by the latter lemma is therefore that of finding the correct $\mathcal{I}^*$. Lemma 3 explains why the set $\mathcal{I}^*$ given in Theorem 1 is indeed an optimal set of indices.

Lemma 2. *The solution $\boldsymbol{\alpha}^*$ of Problem* (18) *is such that $\boldsymbol{\alpha}^* \in (\boldsymbol{\alpha}(\mathcal{I}))_{\mathcal{I}\subseteq\mathcal{Y}\setminus\{y\}}$, i.e. the components α_q^* of $\boldsymbol{\alpha}^*$ obey* (20) *(see Definition 1).*

Proof. We denote by W^*, w_q^*, $\boldsymbol{\xi}^*$ the primal variable at the optimum of (2).

Suppose that we know the set $\mathcal{I}^*$ of indices such that for $q \in \mathcal{I}^*$, $\alpha_q^* > 0$ and denote $I^* = |\mathcal{I}^*|$ the size of $\mathcal{I}^*$. Given optimality condition (9), we have $\xi_q^* = \alpha_q^*/C$, for $q \in \mathcal{I}^*$. Combining the complementarity condition (10) and the expression of w_q^* given by (14), we get that, for $q \in \mathcal{I}^*$:

$$\begin{aligned}\frac{\alpha_q^*}{C} - \left\langle w_q^t - \left(\alpha_q^* - \frac{1}{Q}s_{\alpha^*}\right)x, x\right\rangle - \Delta = 0 &\Leftrightarrow \frac{\alpha_q^*}{C} - \ell_q^t + \left(\alpha_q^* - \frac{1}{Q}s_{\alpha^*}\right)\|x\|^2 = 0\\ &\Leftrightarrow \kappa\alpha_q^* - \ell_q^t - \frac{\|x\|^2}{Q}s_{\alpha^*} = 0\\ &\Leftrightarrow \alpha_q^* = \frac{1}{\kappa}\left(\ell_q^t + \frac{\|x\|^2}{Q}s_{\alpha^*}\right),\end{aligned}$$

where $s_{\alpha^*} = \sum_{q\in\mathcal{I}^*}\alpha_q^*$. Summing over $q \in \mathcal{I}^*$ gives

$$s_{\alpha^*} = \frac{1}{\kappa}\left(\sum_{q\in\mathcal{I}^*}\ell_q^t + I^*\frac{\|x\|^2}{Q}s_{\alpha^*}\right) \Leftrightarrow s_{\alpha^*} = \frac{Q}{\kappa Q - I^*\|x\|^2}\sum_{q\in\mathcal{I}^*}\ell_q^t.$$

This completes the proof. □

Lemma 3. *If $\mathcal{I}^*$ is chosen as recommended by Theorem 1 then $\boldsymbol{\alpha}(\mathcal{I}^*)$ is the solution of Problem* (18)*.*

Proof. Let I^* be chosen as the largest I fulfilling (23) and $\mathcal{I}^* \doteq \{\sigma(1),\ldots,\sigma(I^*)\}$.

On the one hand,

$$\begin{aligned}\ell_{\sigma(I^*)}^t + \frac{\|x\|^2}{\kappa Q - (I^*-1)\|x\|^2}\sum_{q=1}^{I^*-1}\ell_{\sigma(q)}^t > 0 &\Leftrightarrow \left(\kappa Q - (I^*-1)\|x\|^2\right)\ell_{\sigma(I^*)}^t + \|x\|^2\sum_{q=1}^{I^*-1}\ell_{\sigma(q)}^t > 0\\ &\Leftrightarrow \left(\kappa Q - (I^*-1)\|x\|^2\right)\ell_{\sigma(I^*)}^t + \|x\|^2\sum_{q=1}^{I^*}\ell_{\sigma(q)}^t - \|x\|^2\ell_{\sigma(I^*)}^t > 0\\ &\Leftrightarrow \left(\kappa Q - I^*\|x\|^2\right)\ell_{\sigma(I^*)}^t + \|x\|^2\sum_{q=1}^{I^*}\ell_{\sigma(q)}^t > 0\\ &\Leftrightarrow \ell_{\sigma(I^*)}^t + \frac{\|x\|^2}{Q}\frac{Q}{\kappa Q - I^*\|x\|^2}\sum_{q\in\mathcal{I}^*}\ell_q^t > 0\\ &\Leftrightarrow \ell_{\sigma(I^*)}^t + \frac{\|x\|^2}{Q}s_\alpha(\mathcal{I}^*) > 0 \Leftrightarrow \alpha_{\sigma(I^*)} > 0\end{aligned}$$

where we used that the denominators are strictly positive in the first and next-to-last lines, and that $\sum_{q=1}^{I^*}\ell_{\sigma(q)}^t = \sum_{q\in\mathcal{I}^*}\ell_q^t$, by the definition of $\mathcal{I}^*$. Using $\alpha_p(\mathcal{I}) \geq \alpha_q(\mathcal{I})$ for any $\mathcal{I}$, whenever $\ell_p^t \geq \ell_q^t$ for $p, q \in \mathcal{I}$ (see Remark 1), this first series of equations says that

$$\alpha_q(\mathcal{I}^*) > 0,\ q \in \mathcal{I}^*. \tag{25}$$

On the other hand, we have

$$\ell_{\sigma(J)}^t + \frac{\|x\|^2}{Q}s_\alpha(\mathcal{I}^*) \leq 0,\ \forall J > I^*. \tag{26}$$

Indeed, it suffices to observe that, using the definition of $s_\alpha(\mathcal{I}^*)$ (see (21))

$$\ell^t_{\sigma(I^*+1)} + \frac{\|x\|^2}{Q}s_\alpha(\mathcal{I}^*) > 0 \Leftrightarrow \ell^t_{\sigma(I^*+1)} + \frac{\|x\|^2}{\kappa Q - I^*\|x\|^2}\sum_{q=1}^{I^*}\ell^t_{\sigma(q)} > 0,$$

which is impossible because it would mean that I^*+1 also fulfills equation (23) while being larger than I^*. Hence

$$\ell^t_{\sigma(I^*+1)} + \frac{\|x\|^2}{Q}s_\alpha(\mathcal{I}^*) \leq 0.$$

As for all $J \geq I^*+1$, $\ell^t_{\sigma(J)^*} \leq \ell^t_{\sigma(I^*+1)}$, Equation (26) indeed holds.

We are now ready to prove that $\boldsymbol{\alpha}(\mathcal{I}^*)$ is the optimal solution of (18). To do so, we are simply going to show that the duality gap between the primal and dual objective is zero when considering $W(\mathcal{I}^*)$ and $\boldsymbol{\alpha}(\mathcal{I}^*)$, i.e.

$$F(W(\mathcal{I}^*)) - H(\boldsymbol{\alpha}(\mathcal{I}^*)) = 0.$$

As the primal optimization problem is convex, having a zero duality gap is a necessary and sufficient condition for $\boldsymbol{\alpha}(\mathcal{I}^*)$ (and thus, $W(\mathcal{I}^*)$) to be the solution of (18).

A few calculations give the following:

$$F(W(\mathcal{I}^*)) = \frac{1}{2}\kappa\sum_{q=1}^{I^*}\alpha_q^2(\mathcal{I}^*) - \frac{1}{2}\frac{\|x\|^2}{Q}s_\alpha(\mathcal{I}^*) + \frac{C}{2}\sum_{q=I^*+1}^{Q}\left|\ell^t_{\sigma(q)} + \frac{\|x\|^2}{Q}s_\alpha(\mathcal{I}^*)\right|_+^2$$

$$H(\boldsymbol{\alpha}(\mathcal{I}^*)) = \frac{1}{2}\kappa\sum_{q=1}^{I^*}\alpha_q^2(\mathcal{I}^*) - \frac{1}{2}\frac{\|x\|^2}{Q}s_\alpha(\mathcal{I}^*),$$

and the duality gap is therefore given by

$$F(W(\mathcal{I}^*)) - H(\boldsymbol{\alpha}(\mathcal{I}^*)) = \frac{C}{2}\sum_{q=I^*+1}^{Q}\left|\ell^t_{\sigma(q)} + \frac{\|x\|^2}{Q}s_\alpha(\mathcal{I}^*)\right|_+^2,$$

and, as established in (26), $\ell^t_{\sigma(q)} + \frac{\|x\|^2}{Q}s_\alpha(\mathcal{I}^*) \leq 0$ for $q > I^*$. We thus have the desired result: $F(W(\mathcal{I}^*)) - H(\boldsymbol{\alpha}(\mathcal{I}^*)) = 0$.

All in all, we have constructed a vector of coefficients $\boldsymbol{\alpha}^{\mathcal{I}}$ fulfilling the nonnegativity constraints and realizing a zero-duality gap: $\boldsymbol{\alpha}(\mathcal{I})$ is indeed the solution of Problem (18). Consequently, $W(\mathcal{I}^*)$ is the solution of Problem (2). □

References

[1] S. Matsushima, N. Shimizu, K. Yoshida, T. Ninomiya, and H. Nakagawa. Exact passive-aggressive algorithm for multiclass classification using support class. In *SDM 10*, pages 303–314, 2010.

[2] C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.