Méthode de Plans Sécants Régularisée pour l'Optimisation Non convexe : Annexe à l'article dans JMLR

Trinh Minh Tri Do, Thierry Artières

Regularized Bundle Methods for convex and non convex risks: Additional material to the JMLR paper

Trinh-Minh-Tri Do[†,‡] Tri.Do@idiap.ch
[†]*Idiap Research Institute*
Rue Marconi 19
1920 Martigny, Switzerland

Thierry Artières[‡] Thierry.Artieres@lip6.fr
[‡]*LIP6 - Université Pierre et Marie Curie (UPMC)*
4 Place Jussieu 75005 Paris, France

Editor: LIP6 - UPMC internal report - November 2012

1. Abstract

This report is an additional material to our article in the Journal of Machine Learning Research (JMLR). Both documents deal with an algorithm that we designed, named NRBM (Non convex Regularized Bundle Methods), to deal efficiently with regularized non convex risks as often encountered in the machine learning field. The JMLR article provides more details on NRBM and report empirical evaluation on many real machine learning problems while this report provides additional theoretical results related to the convergence analysis of algorithm NRBM.

Main part of this work was done while the first author, Trinh Minh Tri DO, was with UPMC.

2. Introduction

The algorithm we describe here is designed to deal with the following general unconstrained optimization problem

$$\begin{array}{ll} \min_{\mathbf{w}} & f(\mathbf{w}) \\ with & f(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + R(\mathbf{w}) \end{array} \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^D$ are the model parameters and $R(\mathbf{w})$ (the main objective) is a data-fitting measurement to be minimized which we consider to be not necessarily smooth everywhere nor convex.

This report provides theoretical elements on the convergence of our algorithm named-NRBM. As will be seen proofs hold for general risks, including non convex and non smooth risks under two assumptions, the first one being very standard (Lipschitz continuous risk) while the second is moot. Hence, although our analysis does not clearly prove that NRBM algorithm converges towards a stationary solution for non convex risks it still provides in our opinion valuable results that may be of interest.

We first recall the notations and the algorithm as they are described in the paper on NRBM, then we detail our analysis on convergence.

3. Non-Convex Regularized Bundle Method (NRBM)

3.1 Notations

Cutting Planes. A cutting plane (CP) $c_{\mathbf{w}'}$ is an approximation of f which is accurate for $\mathbf{w}$ lying in the vicinity of $\mathbf{w}'$ where the CP is defined, i.e. where the gradient is computed. It is defined as :

$$\begin{array}{ll} c_{\mathbf{w}'}(\mathbf{w}) & = \langle \mathbf{a}_{\mathbf{w}'}, \mathbf{w} \rangle + b_{\mathbf{w}'} \\ with & \mathbf{a}_{\mathbf{w}'} \in \partial f(\mathbf{w}') \\ & b_{\mathbf{w}'} = f(\mathbf{w}') - \langle \mathbf{a}_{\mathbf{w}'}, \mathbf{w}' \rangle \end{array} \tag{2}$$

Raw and modified cutting plane. We distinguish between a raw linear cutting plane of the risk $c_{\mathbf{w}_j}$ (with $c_{\mathbf{w}_j}(\mathbf{w}) = \langle \mathbf{a}_{\mathbf{w}_j}, \mathbf{w} \rangle + b_{\mathbf{w}_j}$) that is built at a particular iteration j of the algorithm and the eventually modified versions of this cutting plane that might be used in posterior iterations. Indeed a cutting plane may be modified multiple times for solving conflicts as in standard NBM method. At iteration t we note c_j^t (with $c_j^t(\mathbf{w}) = \langle \mathbf{a}_j, \mathbf{w} \rangle + b_j^t$) the cutting plane which is derived from $c_{\mathbf{w}_j}$, the raw CP originally built at iteration j. Unlike NBM, the normal vector $\mathbf{a}_j$ in our algorithm might be different than the subgradient $\mathbf{a}_{\mathbf{w}_j}$ computed at $\mathbf{w}_j$, due to our particular solving conflict method. However, once defined at iteration j, the normal vector $\mathbf{a}_j$ keeps fixed over iterations. On the contrary, the offset might be modified for solving conflicts occurring after iteration j, and we use a superscript t indicating the iteration number for the cutting plane's offset b_j^t.

Current and best solutions. We note $\mathbf{w}_t$ the current solution found at iteration t. It is the one that minimizes the approximated problem at the previous iteration. Also, we note $\mathbf{w}_t^*$ the best solution up to iteration t, i.e. the solution $\mathbf{w}_j (j \leq t)$ with minimum value $f(\mathbf{w}_j)$.

Approximated problem. Like in Bundle Methods (CRBM, NBM, etc), we build an accurate approximation of f by using an ensemble of cutting planes of R built at different coordinates. Let $\mathcal{C}$ be a working set of (active) cutting planes, the approximation of f is:

$$f(\mathbf{w}) \approx g_t(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \max_{c \in \mathcal{C}} c(\mathbf{w})$$

The minimization of the approximation function is called the approximated problem. Consider a working set C of active cutting planes, it is defined as:

$$\tilde{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \max_{c \in \mathcal{C}} c(\mathbf{w}) \tag{3}$$

We note $\tilde{\mathbf{w}}_t$ the minimizer of the curent approximation function $g_t(\mathbf{w})$ and v_t is the minimum of the curent approximation function, i.e. $v_t = g_t(\tilde{\mathbf{w}}_t)$.

Bundle. The bundle $\mathbb{B}_t$ denotes the state of the algorithm at iteration t. It consists in a bundle of cutting planes which were built at previous solutions, c_j^t for $j = 1..t$. Similarly to non-convex bundle methods, we define a locality measure which is associated to any active cutting plane. It is related to the locality between the point where the cutting plane was built and the best current observed solution. We note s_j^t the locality measure between cutting plane c_j^t and the best observed solution up to iteration t, $\mathbf{w}_t^*$. The full bundle information is:

$$\mathbb{B}_t = \{c_1^t, ..., c_t^t, \tilde{c}_{t-1}^t, s_1^t, ..., s_t^t, \tilde{s}_{t-1}^t\} \tag{4}$$

where $\tilde{c}_{t-1}^t$ is an aggregated cutting plane and $\tilde{s}_{t-1}^t$ is its locality measure to the best observed solution $\mathbf{w}_t^*$.

Aggregation cutting plane. The aggregation cutting plane at iteration t is noted $\tilde{c}_t^t$. This aggregation cutting plane is not directly built from a solution, it may be viewed as a convex combination of the cutting planes in the bundle, designed to accumulate information from past cutting planes (allowing for limited memory vesrions where one maintains only a fixed number of CP in the bundle). Our aggregation technique differs from standard NBM in that we define a locality measure, $\tilde{s}_t^t$, for the aggregated CP which is estimated as a similar convex combination (with the same weights) of the locality measures associated to the cutting planes in the bundle.

Locality measure. At iteration t, we define the locality measure between CP c_j^t built at $\mathbf{w}_j$ and $\mathbf{w}_t^*$ as:

$$s_j^t = s(\mathbf{w}_j, \mathbf{w}_t^*) = \frac{\lambda}{2}\left(\|\mathbf{w}_j - \mathbf{w}_j^*\|^2 + \sum_{k=j+1}^{t} \|\mathbf{w}_k^* - \mathbf{w}_{k-1}^*\|^2\right)$$

which yields a natural recursive formulate:

$$s_j^t = s_j^{t-1} + \frac{\lambda}{2}\|\mathbf{w}_t^* - \mathbf{w}_{t-1}^*\|^2, \forall j < t \tag{5}$$

3.2 Algorithm

The main algorithm is described in Algorithm 1. It takes as input an initial solution and values for hyper parameters λ and ϵ. It produces as output a solution of the optimization problem, $\mathbf{w}^*$. It calls Algorithm 2, which itself calls Algorithm 3. We focus here on a full variant of the NRBM algorithm where the bundle at iteration t includes all the cutting planes built at previous iterations. Yet since all our discussion and theoretical elements require only the aggregation cutting plane and the cutting planes built at iteration t to belong to the bundle. Hence all the discussion hereafter holds for a limited memory variant of NRBM where the bundle includes only a subset of the cutting planes.

Note that there are two conditions that should hold when modifying the offset of a cutting plane when a conflict is solved. The two conditions are given below and provide an upper bound U and a lower bound L for b_t^t, these conditions are used in Algorithm 3.

$$\begin{array}{ll} b_t^t & \leq R(\mathbf{w}_t^*) - \langle \mathbf{a}_{\mathbf{w}_t}, \mathbf{w}_t^* \rangle - s_t^t = U \\ b_t^t & \geq f(\mathbf{w}_t^*) - \frac{\lambda}{2}\|\mathbf{w}_t\|^2 - \langle \mathbf{a}_{\mathbf{w}_t}, \mathbf{w}_t \rangle = L \end{array} \tag{6}$$

If $L \leq U$ any value in (L, U) works (in our implementation we set $b_t^t = L$).

Algorithm 1 NRBM

Input: $\mathbf{w}_1$, λ, ϵ
Output: $\mathbf{w}^*$
Initialization:
  Compute cutting plane $c_{\mathbf{w}_1}$ of R
  $[c_1^1, s_1^1] = [\tilde{c}_1^1, \tilde{s}_1^1] = [c_{\mathbf{w}_1}, 0]$
  $\tilde{\mathbf{w}}_1 = -\mathbf{a}_1/\lambda$
  $B_1 = \{c_1^1, s_1^1, \tilde{c}_1^1, \tilde{s}_1^1\}$
for $t = 2$ to ∞ **do**
  $\mathbf{w}_t \leftarrow \tilde{\mathbf{w}}_{t-1}$
  Compute cutting plane $c_{\mathbf{w}_t}$ of R
  $\mathbf{w}_t^* = \text{argmin}_{\mathbf{w} \in \{\mathbf{w}_1, \ldots \mathbf{w}_t\}} f(\mathbf{w}_t)$
  $B_t = \text{UpdateBundle}(B_{t-1}, \mathbf{w}_{t-1}^*, \mathbf{w}_t^*, c_{\mathbf{w}_t}, \mathbf{w}_t)$
  $(\tilde{\mathbf{w}}_t, v_t, \tilde{c}_t^t, \tilde{s}_t^t) = \text{MinimizeApproximationProblem}(B_t, \lambda)$
  $gap_t = f(\mathbf{w}_t^*) - v_t$
  if $gap_t < \epsilon$ **then return** $\mathbf{w}_t^*$
end for

Algorithm 2 UpdateBundle

Input: $B_{t-1} = \{c_1^{t-1}, ..., c_{t-1}^{t-1}, \tilde{c}_{t-1}^{t-1}, s_1^{t-1}, ...s_{t-1}^{t-1}, \tilde{s}_{t-1}^{t-1}\}, \mathbf{w}_{t-1}^*, \mathbf{w}_t^*, \mathbf{w}_t, c_{\mathbf{w}_t}$
Output: $B_t = \{c_1^t, ..., c_t^t, \tilde{c}_{t-1}^t, s_1^t, ...s_t^t, \tilde{s}_{t-1}^t\}$
if $\mathbf{w}_t^* \neq \mathbf{w}_{t-1}^*$ **then** Descent Step
  for $j = 1..t-1$
    $s_j^t = s_j^{t-1} + \frac{\lambda}{2}\|\mathbf{w}_t^* - \mathbf{w}_{t-1}^*\|^2$
    $b_j^t = \min[b_j^{t-1}, R(\mathbf{w}_t^*) - \langle \mathbf{a}_j, \mathbf{w}_t^* \rangle - s_j^t]$
  end
  $\tilde{s}_{t-1}^t = \tilde{s}_{t-1} + \frac{\lambda}{2}\|\mathbf{w}_t^* - \mathbf{w}_{t-1}^*\|^2$
  $\tilde{b}_{t-1}^t = \min[\tilde{b}_{t-1}^{t-1}, R(\mathbf{w}_t^*) - \langle \tilde{\mathbf{a}}_{t-1}, \mathbf{w}_t^* \rangle - \tilde{s}_{t-1}^t]$
  $\tilde{c}_{t-1}^t(\mathbf{w}) := \langle \tilde{\mathbf{a}}_{t-1}, \mathbf{w} \rangle + \tilde{b}_{t-1}^t$
  $[c_t^t, s_t^t] = [c_{\mathbf{w}_t}, 0]$
else Null Step
  $[c_1^t, ..., c_{t-1}^t, \tilde{c}_{t-1}^t, s_1^t, ...s_{t-1}^t, \tilde{s}_{t-1}^t] = [c_1^{t-1}, ..., c_{t-1}^{t-1}, \tilde{c}_{t-1}^{t-1}, s_1^{t-1}, ...s_{t-1}^{t-1}, \tilde{s}_{t-1}^{t-1}]$
  if $(b_{\mathbf{w}_t} \leq R(\mathbf{w}_t^*) - \langle \mathbf{a}_{\mathbf{w}_t}, \mathbf{w}_t^* \rangle - s(\mathbf{w}_t, \mathbf{w}_t^*))$ **then**
    $[c_t^t, s_t^t] = \text{SolveConflictNullStep}(\mathbf{w}_t^*, \mathbf{w}_t, c_{\mathbf{w}_t})$
  else $[c_t^t, s_t^t] = [c_{\mathbf{w}_t}, \frac{\lambda}{2}\|\mathbf{w}_t - \mathbf{w}_t^*\|^2]$
end

4. Convergence analysis for NRBM

In this section, we present the analysis of convergence rate of NRBM and the result on the convergence toward a stationary solution. Our analysis of convergence rate of NRBM is based on the fact the approximation gap decreases towards zero, and the algorithm requires a limited number of iteration to reach a gap lower than a given (positive) precision ϵ. Concerning the nature of the solution found by NRBM, we show that if the algorithm

Algorithm 3 SolveConflictNullStep

Input: $\mathbf{w}_t^*, \mathbf{w}_t, c_{\mathbf{w}_t}$ with parameters $(\mathbf{a}_{\mathbf{w}_t}, b_{\mathbf{w}_t})$
Output: c_t^t with parameters $(\mathbf{a}_t, b_t^t)$ and s_t^t
$s_t^t = \frac{\lambda}{2}\|\mathbf{w}_t^* - \mathbf{w}_t\|^2$
Compute L, U according to (6)
if $L \leq U$ **then** $[\mathbf{a}_t, b_t] = [\mathbf{a}_{\mathbf{w}_t}, L]$ **else**
    $\mathbf{a}_t = -\lambda \mathbf{w}_t^*$ NullStep2 case
    $b_t^t = f(\mathbf{w}_t^*) - \frac{\lambda}{2}\|\mathbf{w}_t\|^2 - \langle \mathbf{a}_t, \mathbf{w}_t \rangle$

reaches a null approximation gap after a finite number of iterations then a stationary solution is found. Furthermore, if the algorithms does not reach a null gap after a finite number of iterations then it generates cluster points which are stationary solutions.

These results are gained under two assumptions that we presented and discussed in section 4. In Section 4.2, we present some useful results concerning the gap, the locality measure of the aggregation cutting plane, and the minimum of the approximation problem. These preliminary results are then used to derive our main results concerning convergence rate analysis convergence to a stationary solution (Section 4.3 and 4.4).

4.1 Assumptions

The necessary assumptions for proving our main results are the following:

- H1 : The empirical risk is Lipschitz continuous with a constant G.
- H2 : The number of iterations where a conflict is solved by modifying the normal vector $\mathbf{a}_t$ (NullStep2 case in Algorithm 1) is finite.

H1 is a rather standard assumption. It was used for instance in (Smola et al., 2008; Shalev-Shwartz et al., 2007; Joachims, 2006), for proving convergence results. It is in particular a reasonable assumption in case of smooth almost everywhere risks such as those one gets using hinge loss and maximum margin criterion (SVM, structured output prediction, etc).

H2 is less intuitive. Recall that there is a NullStep2 at iteration t if and only if the raw cutting plane built at current solution $\mathbf{w}_t$ is not compatible with the best observed solution $\mathbf{w}_t^*$. Hence, since the current solution and the best observed solution get closer as the iteration number increases we may hope that NullStep2 do not arise after a finite number of iterations. Furthermore, it is very likely that if the algorithm gets close enough to a stationary solution $\mathbf{w}^*$ lying within a smooth area then it should converge towards this stationary solution without conflicts anymore, as it would do in case of a convex and smooth objective. This is particularly expected for our algorithm (compared to standard non convex bundle methods) since it focuses on maintaining a good approximation function around the best current solution. Another important point is that we did not observe any case of infinite number of conflicts in our experiments (on both academic optimization problems and machine learning problems) where NullStep2 mainly occurred in a few early iterations.

At the end these claims are still not proved so that the convergence of NRBM to a stationary solution is not fully proved here, but we believe that our convergence analysis establishes some important elements towards a fast and fully proved bundle method for minimizing non-convex regularized function.

4.2 Preliminary results

The three preliminary results below establish, first a link between the gap, $gap_t = f(\mathbf{w}_t^*) - v_t$, and the locality measure $\tilde{s}_t^t$ of the aggregation cutting plane, second a lower bound on the minimum of the approximation function, $v_t = \min_{\mathbf{w}} g_t(\mathbf{w})$, and finally the bounds on the solution generated by NRBM.

Lemma 4.1 *Following inequality always holds:*

$$\tilde{s}_t^t + \frac{\lambda}{2}\|\tilde{\mathbf{w}}_t - \mathbf{w}_t^*\|^2 \leq gap_t \tag{7}$$

As a consequence:

$$\tilde{s}_t^{t+1} \leq gap_t \tag{8}$$

Proof

First, we know that:

$$R(\mathbf{w}_t^*) - \tilde{c}_t^t(\mathbf{w}_t^*) \geq \tilde{s}_t^t \tag{9}$$

Then:

$$\begin{array}{lll} & \tilde{s}_t^t + \langle \tilde{\mathbf{a}}_t, \mathbf{w}_t^* \rangle + \tilde{b}_t^t & \leq R(\mathbf{w}_t^*) \\ \iff & \tilde{s}_t^t + \langle \tilde{\mathbf{a}}_t, \mathbf{w}_t^* \rangle + \tilde{b}_t^t + \frac{\lambda}{2}\|\mathbf{w}_t^*\|^2 - v_t & \leq R(\mathbf{w}_t^*) + \frac{\lambda}{2}\|\mathbf{w}_t^*\|^2 - v_t \\ \iff & \tilde{s}_t^t + \langle \tilde{\mathbf{a}}_t, \mathbf{w}_t^* \rangle + \tilde{b}_t^t + \frac{\lambda}{2}\|\mathbf{w}_t^*\|^2 - v_t & \leq f(\mathbf{w}_t^*) - v_t = gap_t \end{array} \tag{10}$$

Next, using that:

$$\tilde{\mathbf{w}}_t = -\frac{\tilde{\mathbf{a}}_t}{\lambda} \tag{11}$$

and:

$$\begin{array}{lll} g_t(\tilde{\mathbf{w}}_t) = v_t & = -\frac{1}{2\tilde{\lambda}}\|\boldsymbol{\alpha}_t A_t\|^2 + \boldsymbol{\alpha}_t B_t & = -\frac{\lambda}{2}\|\frac{\tilde{\mathbf{a}}_t}{\lambda}\|^2 + \tilde{b}_t^t \\ & = \frac{\lambda}{2}\|\frac{\tilde{\mathbf{a}}_t}{\lambda}\|^2 - \lambda\|\frac{\tilde{\mathbf{a}}_t}{\lambda}\|^2 + \tilde{b}_t^t & = \frac{\lambda}{2}\|\tilde{\mathbf{w}}_t\|^2 - \langle \tilde{\mathbf{a}}_t, \frac{\tilde{\mathbf{a}}_t}{\lambda} \rangle + \tilde{b}_t^t \\ & = \frac{\lambda}{2}\|\tilde{\mathbf{w}}_t\|^2 + \langle \tilde{\mathbf{a}}_t, \tilde{\mathbf{w}}_t \rangle + \tilde{b}_t^t & \end{array} \tag{12}$$

we get:

$$\begin{array}{ll} & \langle \tilde{\mathbf{a}}_t, \mathbf{w}_t^* \rangle + \tilde{b}_t^t + \frac{\lambda}{2}\|\mathbf{w}_t^*\|^2 - v_t \\ = & \langle \tilde{\mathbf{a}}_t, \mathbf{w}_t^* \rangle + \tilde{b}_t^t + \frac{\lambda}{2}\|\mathbf{w}_t^*\|^2 - \frac{\lambda}{2}\|\tilde{\mathbf{w}}_t\|^2 - \langle \tilde{\mathbf{a}}_t, \tilde{\mathbf{w}}_t \rangle - \tilde{b}_t^t \\ = & -\lambda\langle \tilde{\mathbf{w}}_t, \mathbf{w}_t^* \rangle + \frac{\lambda}{2}\|\mathbf{w}_t^*\|^2 - \frac{\lambda}{2}\|\tilde{\mathbf{w}}_t\|^2 + \lambda\|\tilde{\mathbf{w}}_t\|^2 \\ = & \frac{\lambda}{2}\|\mathbf{w}_t^* - \tilde{\mathbf{w}}_t\|^2 \end{array} \tag{13}$$

so that Eq. (7) is satisfied.

Furthermore, since $\mathbf{w}_{t+1}^*$ is either $\tilde{\mathbf{w}}_t$ or $\mathbf{w}_t^*$ (depending on the iteration $t+1$ being a descend step or a null step), $\|\mathbf{w}_{t+1}^* - \mathbf{w}_t^*\|^2 \leq \|\tilde{\mathbf{w}}_t - \mathbf{w}_t^*\|^2$. Then:

$$\tilde{s}_t^{t+1} = \tilde{s}_t^t + \frac{\lambda}{2}\|\mathbf{w}_{t+1}^* - \mathbf{w}_t^*\|^2 \leq \tilde{s}_t^t + \frac{\lambda}{2}\|\tilde{\mathbf{w}}_t - \mathbf{w}_t^*\|^2 \leq gap_t \tag{14}$$

■

Lemma 4.2 *The minimum value of the approximation function at iteration t, $g_t(\mathbf{w})$, is lower bounded by:*

$$v_t \geq \max_{\alpha_t \in [0,1]} -\frac{q}{2}(\alpha_t)^2 + l\alpha_t + m \tag{15}$$

where

$$\begin{array}{ll} q & = \frac{1}{\lambda}\|\mathbf{a}_t - \tilde{\mathbf{a}}_{t-1}\|^2 \\ l & = c_t^t(\mathbf{w}_t) - \tilde{c}_{t-1}^t(\mathbf{w}_t) \\ m & = \frac{\lambda}{2}\|\mathbf{w}_t\|^2 + \tilde{c}_{t-1}^t(\mathbf{w}_t). \end{array}$$

Furthermore

$$l \geq \tilde{s}_{t-1}^t > 0 \tag{16}$$

$$m \geq f(\mathbf{w}_t^*) - gap_{t-1} \tag{17}$$

$$l \geq f(\mathbf{w}_t^*) - m. \tag{18}$$

Proof

Since the aggregated cutting plane and the new added cutting plane at iteration t are a subset of the active cutting planes at iteration t, we can define a simple lower bound on v_t, noted v_t^{low}, through:

$$v_t \geq v_t^{low} = \min_{\mathbf{w}} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \max\left[\langle \tilde{\mathbf{a}}_{t-1}, \mathbf{w}\rangle + \tilde{b}_{t-1}^t, \langle \mathbf{a}_t, \mathbf{w}\rangle + b_t^t\right] \tag{19}$$

Note that v_t^{low} may be characterized as the maximum of the above objective in its dual form:

$$\begin{array}{lll} v_t^{low} = \max_{\tilde{\alpha}_{t-1}, \alpha_t} & -\frac{\lambda}{2}\|\frac{\tilde{\alpha}_{t-1}\tilde{\mathbf{a}}_{t-1} + \alpha_t \mathbf{a}_t}{\lambda}\|^2 + \tilde{\alpha}_{t-1}\tilde{b}_{t-1}^t + \alpha_t b_t^t \\ s.t & 0 \leq \tilde{\alpha}_{t-1}, \alpha_t \leq 1 \\ & \tilde{\alpha}_{t-1} + \alpha_t = 1 \end{array} \tag{20}$$

where $\tilde{\alpha}_{t-1}, \alpha_t \in \mathbb{R}$ are Lagrange multipliers. The above quadratic program of the two variables may be easily rewritten so that we get:

$$\begin{array}{lll} v_t \geq v_t^{low} & = \max\limits_{\alpha_t \in [0,1]} & -\frac{1}{2\lambda}\|\tilde{\mathbf{a}}_{t-1} + \alpha_t(\mathbf{a}_t - \tilde{\mathbf{a}}_{t-1})\|^2 + \alpha_t(b_t^t - \tilde{b}_{t-1}^t) + \tilde{b}_{t-1}^t \\ & = \max\limits_{\alpha_t \in [0,1]} & -\frac{1}{2\lambda}\|\mathbf{a}_t - \tilde{\mathbf{a}}_{t-1}\|^2(\alpha_t)^2 + (\frac{\|\tilde{\mathbf{a}}_{t-1}\|^2}{\lambda} - \frac{\langle \mathbf{a}_t, \tilde{\mathbf{a}}_{t-1}\rangle}{\lambda} + b_t^t - \tilde{b}_{t-1}^t)\alpha_t - \frac{\|\tilde{\mathbf{a}}_{t-1}\|^2}{2\lambda} + \tilde{b}_{t-1}^t \end{array} \tag{21}$$

which has the same shape as in Eq. (15), with $q = \frac{1}{\lambda}\|\mathbf{a}_t - \tilde{\mathbf{a}}_{t-1}\|^2$, $l = (\frac{\|\tilde{\mathbf{a}}_{t-1}\|^2}{\lambda} - \frac{\langle \mathbf{a}_t, \tilde{\mathbf{a}}_{t-1}\rangle}{\lambda} + b_t^t - \tilde{b}_{t-1}^t)$ and $m = -\frac{\|\tilde{\mathbf{a}}_{t-1}\|^2}{2\lambda} + \tilde{b}_{t-1}^t$. Identifying the linear term in Eq. (15), while noticing that $\mathbf{w}_t = \tilde{\mathbf{w}}_{t-1} = -\frac{\tilde{\mathbf{a}}_{t-1}}{\lambda}$, we get the right expression for l:

$$\begin{array}{ll} l & = \frac{\|\tilde{\mathbf{a}}_{t-1}\|^2}{\lambda} - \frac{\langle \mathbf{a}_t, \tilde{\mathbf{a}}_{t-1}\rangle}{\lambda} + b_t^t - \tilde{b}_{t-1}^t \\ & = \langle \mathbf{a}_t, \mathbf{w}_t\rangle + b_t^t - \langle \tilde{\mathbf{a}}_{t-1}, \mathbf{w}_t\rangle - \tilde{b}_{t-1}^t \\ & = c_t^t(\mathbf{w}_t) - \tilde{c}_{t-1}^t(\mathbf{w}_t) \end{array} \tag{22}$$

To show that $l > 0$ we use the positive linearization error constraint in first equation of Eq. (6) for $\tilde{c}_{t-1}^t$ and $\mathbf{w}_t^*$:

$$R(\mathbf{w}_t^*) - \tilde{s}_{t-1}^t \geq \tilde{c}_{t-1}^t(\mathbf{w}_t^*) \tag{23}$$

Considering the second constraint in Eq. (6):

$$\frac{\lambda}{2}\|\mathbf{w}_t\|^2 + c_t^t(\mathbf{w}_t) \geq f(\mathbf{w}_t^*) \tag{24}$$

Summing the two above inequalities we finally get:

$$c_t^t(\mathbf{w}_t) - \tilde{s}_{t-1}^t \geq \tilde{c}_{t-1}^t(\mathbf{w}_t^*) \tag{25}$$

then $l > 0$ since the locality measure is positive.

To show the result in Eq. (17), we rewrite the constant term m based on Eq. (15):

$$\begin{aligned} m &= -\frac{\|\tilde{\mathbf{a}}_{t-1}\|^2}{2\lambda} + \tilde{b}_{t-1}^t \\ &= \frac{\|\tilde{\mathbf{a}}_{t-1}\|^2}{2\lambda} - \frac{\|\tilde{\mathbf{a}}_{t-1}\|^2}{\lambda} + \tilde{b}_{t-1}^t \\ &= \frac{\lambda}{2}\|\mathbf{w}_t\|^2 + \langle \tilde{\mathbf{a}}_{t-1}, \mathbf{w}_t \rangle + \tilde{b}_{t-1}^t \\ &= \frac{\lambda}{2}\|\mathbf{w}_t\|^2 + \tilde{c}_{t-1}^t(\mathbf{w}_t). \end{aligned} \tag{26}$$

Using the definition of $\tilde{b}_{t-1}^t$, it may further be rewritten as:

$$\begin{aligned} m &= \frac{\lambda}{2}\|\mathbf{w}_t\|^2 + \langle \tilde{\mathbf{a}}_{t-1}, \mathbf{w}_t \rangle + \tilde{b}_{t-1}^t \\ &= \frac{\lambda}{2}\|\tilde{\mathbf{w}}_{t-1}\|^2 + \langle \tilde{\mathbf{a}}_{t-1}, \tilde{\mathbf{w}}_{t-1} \rangle + \tilde{b}_{t-1}^{t-1} + \min[0, R(\mathbf{w}_t^*) - \langle \tilde{\mathbf{a}}_{t-1}, \mathbf{w}_t^* \rangle - \tilde{b}_{t-1}^{t-1} - \tilde{s}_{t-1}^t] \\ &= v_{t-1} + \min[0, R(\mathbf{w}_t^*) - \tilde{c}_{t-1}^{t-1}(\mathbf{w}_t^*) - \tilde{s}_{t-1}^t] \end{aligned} \tag{27}$$

In the case there is no conflict between $\tilde{c}_{t-1}^{t-1}$ and $\mathbf{w}_t^*$, i.e. $R(\mathbf{w}_t^*) - \tilde{c}_{t-1}^{t-1}(\mathbf{w}_t^*) \geq \tilde{s}_{t-1}^t$, then $m = v_{t-1}$ and:

$$f(\mathbf{w}_t^*) - m = f(\mathbf{w}_t^*) - v_{t-1} \leq f(\mathbf{w}_{t-1}^*) - v_{t-1} = gap_{t-1} \tag{28}$$

On the contrary, if there is conflict between $\tilde{c}_{t-1}^{t-1}$ and $\mathbf{w}_t^*$, i.e. $R(\mathbf{w}_t^*) - \tilde{c}_{t-1}^{t-1}(\mathbf{w}_t^*) < \tilde{s}_{t-1}^t$, using Eq. (8):

$$m = v_{t-1} + R(\mathbf{w}_t^*) - \tilde{c}_{t-1}^{t-1}(\mathbf{w}_t^*) - \tilde{s}_{t-1}^t \geq v_{t-1} + R(\mathbf{w}_t^*) - \tilde{c}_{t-1}^{t-1}(\mathbf{w}_t^*) - gap_{t-1} \tag{29}$$

In this case, $\mathbf{w}_t^* \neq \mathbf{w}_{t-1}^*$ because it is not possible to exist a conflict between $\tilde{c}_{t-1}^{t-1}$ and $\mathbf{w}_{t-1}^*$ (by construction of aggregated CP, see Eq. (9)), which implies it is a descent step. Hence $\mathbf{w}_t^* = \tilde{\mathbf{w}}_{t-1}$, and we may write:

$$\begin{aligned} & m \geq v_{t-1} + R(\mathbf{w}_t^*) - \tilde{c}_{t-1}^{t-1}(\tilde{\mathbf{w}}_{t-1}) - gap_{t-1} \\ \iff \quad & m \geq \frac{\lambda}{2}\|\tilde{\mathbf{w}}_{t-1}\|^2 + \tilde{c}_{t-1}^{t-1}(\tilde{\mathbf{w}}_{t-1}) + R(\mathbf{w}_t^*) - \tilde{c}_{t-1}^{t-1}(\tilde{\mathbf{w}}_{t-1}) - gap_{t-1} \end{aligned} \tag{30}$$

since v_{t-1} is the minimum value of $\frac{\lambda}{2}\|\mathbf{w}\|^2 + \tilde{c}_{t-1}^{t-1}(\mathbf{w})$ which is minimized at $\tilde{\mathbf{w}}_{t-1}$. Then:

$$\begin{aligned} & m \geq \frac{\lambda}{2}\|\tilde{\mathbf{w}}_{t-1}\|^2 + R(\mathbf{w}_t^*) - gap_{t-1} \\ \iff \quad & m \geq f(\mathbf{w}_t^*) - gap_{t-1} \end{aligned} \tag{31}$$

where we used $\mathbf{w}_t^* = \tilde{\mathbf{w}}_{t-1}$ again.

Finally, we show the result in Eq (18). Using Eq (22) and Eq (26), we have

$$\begin{array}{lll} & l & \geq f(\mathbf{w}_t^*) - m \\ \iff & l + m & \geq f(\mathbf{w}_t^*) \\ \iff & c_t^t(\mathbf{w}_t) - \tilde{c}_{t-1}^t(\mathbf{w}_t) + \frac{\lambda}{2}\|\mathbf{w}_t\|^2 + \tilde{c}_{t-1}^t(\mathbf{w}_t) & \geq f(\mathbf{w}_t^*) \\ \iff & \frac{\lambda}{2}\|\mathbf{w}_t\|^2 + c_t^t(\mathbf{w}_t) & \geq f(\mathbf{w}_t^*) \end{array} \tag{32}$$

which resumes to our second condition when solving conflict (Cf Eq. (6)). ■

Lemma 4.3 *Let $\bar{G} = \max(\|\lambda \mathbf{w}_1\|, G)$ where G is the Lipschitz constant of R, we have:*

$$\forall t, \|\mathbf{w}_t\| \leq \frac{\bar{G}}{\lambda}. \tag{33}$$

Furthermore:

$$\begin{array}{ll} \forall t, \|\mathbf{a}_t\| & \leq \bar{G} \\ \forall t, \|\tilde{\mathbf{a}}_t\| & \leq \bar{G}. \end{array} \tag{34}$$

Proof

First, $\mathbf{w}_1$ trivially satisfies the inequality. Next assume Eq. (33) is satisfied for $\mathbf{w}_1$ ... $\mathbf{w}_k$. We show now it is also true for $\mathbf{w}_{k+1}$. Actually, $\mathbf{w}_{k+1}$ coincides with $\tilde{\mathbf{w}}_k$, the minimizer of $g_k(\mathbf{w})$, so that:

$$\mathbf{w}_{k+1} = \tilde{\mathbf{w}}_k \quad = \frac{1}{\lambda}\left(\sum_{j=1..k} \alpha_j \mathbf{a}_j + \alpha_{k+1}\tilde{\mathbf{a}}_{k-1}\right) \tag{35}$$

where $\sum_{j=1..k+1} \alpha_j = 1$ and $\alpha_j \geq 0 \ \forall j$. Using Eq. (11), we have $\tilde{\mathbf{a}}_{k-1} = -\lambda \tilde{\mathbf{w}}_{k-1} = -\lambda \mathbf{w}_k$, so that:

$$\mathbf{w}_{k+1} \quad = \sum_{j=1..k} \alpha_j \frac{\mathbf{a}_j}{\lambda} + \alpha_{k+1}(-\mathbf{w}_k) \tag{36}$$

Furthermore, $\|\mathbf{a}_j\| \leq \bar{G}$ for $j < k+1$ since:

a) if $\mathbf{a}_j$ is a raw subgradient of R then $\|\mathbf{a}_j\| \leq G \leq \bar{G}$

b) otherwise, $\mathbf{a}_j$ is a modified normal vector at iteration j: $\mathbf{a}_j = \lambda \mathbf{w}_j^*$ (NullStep2 case) where $\mathbf{w}_j^* \in \{\mathbf{w}_i\}_{i=1..j}$, so that $\|\mathbf{a}_j\| \leq \bar{G}$ by induction hypothesis.

At the end, $\mathbf{w}_{k+1}$ is a convex combination of vectors upper bounded by $\frac{\bar{G}}{\lambda}$, then it is also upper bounded by $\frac{\bar{G}}{\lambda}$.

Finally the results in Eq. (34) are straightforward from the above proof.

■

4.3 Convergence Rate

We provide here an upper bound on the convergence rate of our algorithm. The analysis is based on a lower bound of the decrease of the gap (Lemma 4.4). Using this lower bound, Theorem 4.1 proves that Algorithm 1 converges to a solution with accuracy ϵ with a rate $O(1/\lambda\epsilon)$. The only hypothesis required in this section is Hypothesis $H1$.

Lemma 4.4 *Approximation gap produced by Algorithm 1 satisfies:*

$$gap_{t-1} - gap_t \geq \min(\frac{gap_{t-1}}{2}, \frac{(gap_{t-1})^2\lambda}{8\bar{G}^2}) \tag{37}$$

Proof Starting from the lower bound in Lemma 4.2, we can apply a result from (Teo et al., 2007) which states that the minimum of $\frac{1}{2}qx^2 - lx$ with $l, q > 0$ and $x \in [0,1]$ is upper bounded by $-\frac{l}{2}\min(1, l/q)$. Then, using the same notation for q, l and m as in Lemma 4.2:

$$\begin{array}{lll} & \min_{\alpha_t^t \in [0,1]} \frac{q}{2}(\alpha_t^t)^2 - l\alpha_t^t & \leq -\frac{l}{2}\min(1, l/q) \\ \iff & -\max_{\alpha_t^t \in [0,1]} -\frac{q}{2}(\alpha_t^t)^2 + l\alpha_t^t & \leq \frac{l}{2}\min(1, l/q) \\ \iff & -\max_{\alpha_t^t \in [0,1]} -\frac{q}{2}(\alpha_t^t)^2 + l\alpha_t^t + m & \leq m + \frac{l}{2}\min(1, l/q) \\ \Rightarrow & -v_t & \leq m + \frac{l}{2} min(1, l/q) \\ \Rightarrow & f(\mathbf{w}_t^*) - v_t & \leq f(\mathbf{w}_t^*) - m - \frac{l}{2}\min(1, l/q) \end{array} \tag{38}$$

From Eq. (18), we have $l \geq f(\mathbf{w}_t^*) - m$ so that:

$$f(\mathbf{w}_t^*) - v_t \leq f(\mathbf{w}_t^*) - m - \frac{f(\mathbf{w}_t^*) - m}{2}\min(1, \frac{f(\mathbf{w}_t^*) - m}{q}) \tag{39}$$

Combining $f(\mathbf{w}_t^*) - m \leq gap_{t-1}$ (Lemma 4.2) and the monotonicity of $h(x) = x - \frac{x}{2}\min(1, x/q)$ into Eq. (39) leads to:

$$gap_t \leq gap_{t-1} - \frac{gap_{t-1}}{2}(1, \frac{gap_{t-1}}{q}) \tag{40}$$

Finally subsituting the value of q and using hypothesis Lemma 4.3, $q = \frac{1}{\lambda}\|\mathbf{a}_t - \tilde{\mathbf{a}}_{t-1}\|^2 \leq 4\bar{G}^2/\lambda$, and we get the claim. ■

Theorem 4.1 *Algorithm 1 reaches a gap below ϵ with a number of iterations $O(1/\lambda\epsilon)$.*

Proof Let consider the two quantities occurring in Eq. (37), $gap_{t-1}/2$ and $\lambda gap_{t-1}^2/8\bar{G}^2$.

We first show that the situation where $gap_{t-1}/2 > \lambda gap_{t-1}^2/8\bar{G}^2$ (i.e. $gap_{t-1} > 4\bar{G}^2/\lambda$) may only happen a finite number of iterations, T_0.

Actually if $gap_{t-1} > 4\bar{G}^2/\lambda$ Lemma 4.4 shows that $gap_t \leq gap_{t-1}/2$ and the gap is at least divided by two every iteration. Then $gap_{t-1} > 4\bar{G}^2/\lambda$ may arise for at most $T_0 = log_2(\lambda gap_1/4\bar{G}^2) + 1$ where $gap_1 = \frac{\lambda}{2}\|\mathbf{w}_1 + \mathbf{a}_1/\lambda\|^2$ (it may be obtained analytically since the approximation function in the first iteration is quadratic).

Hence after at most T_0 iterations the gap decreases according to $gap_t - gap_{t-1} \leq -gap_{t-1}^2/8\bar{G}^2 \leq 0$. To go further we introduce a function $u(t)$ which is an upper bound of gap_t (Teo et al., 2007). Solving differential equation $u'(t) = -\frac{\lambda}{8\bar{G}^2}u^2(t)$ with boundary condition $u(T_0) = 4\bar{G}^2/\lambda$ gives $u(t) = \frac{8\bar{G}^2}{\lambda(t+2-T_0)} \geq gap_t / \forall t \geq T_0$. Next solving $u(t) \leq \epsilon \iff t \geq 8\bar{G}^2/\lambda\epsilon + T_0 - 2$, so that a solution is reached with accuracy ϵ within $\left[T_0 + 8\bar{G}^2/\lambda\epsilon - 2\right]$ iterations. ■

4.4 Stationary Solution

We show now that under assumptions listed in Section 4.1, Algorithm 1 either converges towards a stationary solution, or it generates cluster points that are stationary solutions. The proof requires both assumptions $H1$ and $H2$.

We prove first a preliminary result.

Lemma 4.5 *The normal vector, $\tilde{\mathbf{a}}_t$, and the locality measure, $\tilde{s}_t^t$, of the aggregation cutting plane $\tilde{c}_t$ is a convex combination of $\{(\mathbf{a}_j, s_j^t)_{j=1..t}\}$, i.e. there are numbers $\boldsymbol{\beta}^t \in [0,1]^t$ such that:*

$$(\tilde{\mathbf{a}}_t, \tilde{s}_t^t) = \sum_{j=1..t} \beta_j^t (\mathbf{a}_j, s_j^t) \text{ and } \sum_{j=1..t} \beta_j^t = 1 \tag{41}$$

where $\mathbf{a}_j$ is the normal vector of cutting plane added at iteration i, which has eventually been modified CP.

Proof

We prove the lemma by induction.

The lemma is true for first iteration since $[\tilde{\mathbf{a}}_1, \tilde{s}_1^1] = [\mathbf{a}_1, s_1^1] = [\mathbf{a}_{\mathbf{w}_1}, 0]$. Now assume that the lemma is true for t first iterations, we show that the lemma is also true at iteration $t+1$. By definition:

- $(\tilde{\mathbf{a}}_{t+1}, \tilde{s}_{t+1}^{t+1})$ is a convex combination of $\{(\mathbf{a}_1, s_1^{t+1}), ...(\mathbf{a}_{t+1}, s_{t+1}^{t+1}), (\tilde{\mathbf{a}}_t, \tilde{s}_t^{t+1})\}$ (*).

Next since lemma is assumed to be true for iteration t, $(\tilde{\mathbf{a}}_t, \tilde{s}_t^t)$ is a convex combination of $\{(\mathbf{a}_1, s_1^t), ..., (\mathbf{a}_t, s_t^t)\}$. And, by definition of locality measure in Algorithm 1:

$$\begin{array}{ll} s_j^{t+1} & = s_j^t + \frac{\lambda}{2}\|\mathbf{w}_{t+1}^* - \mathbf{w}_t^*\|^2 \ \forall j \leq t \\ \tilde{s}_t^{t+1} & = \tilde{s}_t^t + \frac{\lambda}{2}\|\mathbf{w}_{t+1}^* - \mathbf{w}_t^*\|^2 \end{array} \tag{42}$$

Then:

- $(\tilde{\mathbf{a}}_t, \tilde{s}_t^{t+1})$ is a convex combination of $\{(\mathbf{a}_1, s_1^{t+1}), ..., (\mathbf{a}_t, s_t^{t+1})\}$ (**).

Now, combining (*) and (**) we get that $(\tilde{\mathbf{a}}_{t+1}, \tilde{s}_{t+1}^{t+1})$ is also a convex combination of $\{(\mathbf{a}_1, s_1^{t+1}), ..., (\mathbf{a}_t, s_t^{t+1}), (\mathbf{a}_{t+1}, s_{t+1}^{t+1})\}$. ■

Theorem 4.2 *If $gap_t = 0$ at iteration t of Algorithm 1, then $\mathbf{w}_t^* = \tilde{\mathbf{w}}_t$ and $\mathbf{w}_t^*$ is a stationary point of the objective function f, i.e. $0 \in \partial f(\mathbf{w}_t^*)$.*

Proof

From Lemma 4.5 $(\tilde{\mathbf{a}}_t, \tilde{s}_t^t)$ is a convex combination of $\{(\mathbf{a}_j, s_j^t)\}_{i=1..t}$ with coefficients β_j^t.

Let $LA = \left\{ i | \beta_j^t > 0 \right\}$ be the set of "active" cutting plane indexes with non-null β coefficients, which is not empty since $\sum \beta_j^t = 1$. Lemma 4.1 implies that if $gap_t = 0$ then $\mathbf{w}_t^* = \tilde{\mathbf{w}}_t$ and $\tilde{s}_t^t = 0$. Since $\tilde{s}_t^t$ is a convex combination (with same β coefficients as above) of s_j^t then: $\tilde{s}_t^t = 0 \Rightarrow s_j^t = 0 \ \forall i \in LA$. Futhermore, $\mathbf{s}_j^t = 0 \Rightarrow \mathbf{w}_j \equiv \mathbf{w}_t^*$ so that $\mathbf{w}_j$ is the best observed solution up to iteration t, and it is also the best observed solution up to iteration

i, meaning that iteration i was a descent step, which implies that $\mathbf{a}_j \equiv \mathbf{a}_{\mathbf{w}_j}$ is a subgradient of R at $\mathbf{w}_j = \mathbf{w}_t^*$.

At the end, since $\tilde{\mathbf{a}}_t$ is a convex combination (considering only $\beta_j^t > 0$) of subgradients of R at $\mathbf{w}_t^*$, it is a subgradient of R at $\mathbf{w}_t^*$. Then $(\lambda \mathbf{w}_t^* + \tilde{\mathbf{a}}_t)$ is a subgradient of $f(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + R(\mathbf{w})$ at $\mathbf{w}_t^*$. Recalling that $\mathbf{w}_t^* = \tilde{\mathbf{w}}_t$ and that $\tilde{\mathbf{w}}_t = -\frac{\tilde{\mathbf{a}}_t}{\lambda}$ (Cf. (Eq. 11)), we get that $(\lambda \tilde{\mathbf{w}}_t + \tilde{\mathbf{a}}_t) = 0 \in \partial f(\mathbf{w}_t^*)$ so that $\mathbf{w}_t^*$ is a stationary solution of f. ■

Theorem 4.3 *If Algorithm 1 does not reach a stationary solution in a limited number of iterations, the two infinite sequences* $(\mathbf{w}_t)$ *and* $(\mathbf{w}_t^*)$ *produced by NRBM Algorithm 1 have cluster points*[1]*.*

Proof In the case where Algorithm 1 does not reach a null gap within a finite number of iteration, the two sequences $(\mathbf{w}_t)$ and $(\mathbf{w}_t^*)$ are infinite. Note that the sequence $(\mathbf{w}_t)$ is bounded (Lemma 4.3), and the sequence $(\mathbf{w}_t^*)$ is also bounded as it picks elements from the sequence $(\mathbf{w}_t)$.

Since the two sequences $(\mathbf{w}_t)$ and $(\mathbf{w}_t^*)$ are bounded, they have cluster points since the Bolzano-Weierstrass theorem (see (Moore, 2008)) states that any bounded sequence in R^D has a convergent subsequence. ■

The following theorem show that any cluster point of the sequence $(\mathbf{w}_t^*)$ are stationary solution of the objective function $f(\mathbf{w})$.

Theorem 4.4 *Let* $\mathbf{w}^*$ *be a cluster point of the sequence* $(\mathbf{w}_t^*)$*. Then under assumptions H1 and H2,* $\mathbf{w}^*$ *is a stationary solution of* $f(\mathbf{w})$*.*

Proof

We first give an outline of the proof. We prove the theorem by showing that the subdifferential of the objective function at the cluster point contains a null subgradient, i.e. $\mathbf{0} \in \partial f(\mathbf{w}^*)$. This is equivalent to show there exist a subgradient $\mathbf{a}^*$ of $R(\mathbf{w})$ at $\mathbf{w}^*$ that statisfies $\lambda \mathbf{w}^* + \mathbf{a}^* = \mathbf{0}$. However, since the subdifferential of R at $\mathbf{w}^*$ cannot be computed explicitly, we rely on the fact that if a sequence of subgradients of R converges, its limit is also a subgradient of R (Luksan and Vlcek, 2000). Actually the subgradient $\mathbf{a}^*$ that we look for is actually a convex combination of limits of subgradient sequences. Since each of the subgradient sequence converges toward a subgradient, $\mathbf{a}^*$ converges towards a convex combination of subgradients so that it is itself a subgradient of R.

The proof is organized as follows. First, we propose a candidate $\mathbf{a}^*$ satisfying the equality $\lambda \mathbf{w}^* + \mathbf{a}^* = \mathbf{0}$ (Eq. 46). Then we show that this candidate is actually a subgradient of $R(\mathbf{w})$ at $\mathbf{w}^*$ so that the claim is correct.

From Lemma 4.5, $(\tilde{\mathbf{a}}_t, \tilde{s}_t^t)$ is a convex combination of points in the set $\{(\mathbf{a}_j, s_j^t)\}_{i=1..t}$. Then according to Caratheodory's Theorem (Luksan and Vlcek, 2000), there are at most $D+2$ indexes $\{u_{t,j} \in [1,t] | j = 1..D+2\}$ and $D+2$ weights $\alpha_{u_{t,j}} \in [0,1]$ (where D denotes the dimension of $\mathbf{a}_j$'s) such that:

1. Let $\{x_n\}$ be a sequence of real vectors, then x is a cluster point of $\{x_n\}$ if for every $\epsilon > 0$, there are infinitely many points x_n such that $\|x - x_n\| < \epsilon$.

$$(\tilde{\mathbf{a}}_t, \tilde{s}_t^t) = \sum_{j=1..D+2} \alpha_{u_{t,j}} (\mathbf{a}_{u_{t,j}}, s_{u_{t,j}}^t) \tag{43}$$

Let $\mathbf{w}_{u_{t,j}}$ be the point where the cutting plane with normal vector $\mathbf{a}_{u_{t,j}}$ was built and consider the concatenation of vectors $\mathbf{q}_t = (\mathbf{w}_{u_{t,1}}, \mathbf{w}_{u_{t,2}}, \mathbf{w}_{u_{t,D+2}}, \mathbf{a}_{u_{t,1}}, \mathbf{a}_{u_{t,2}}, \mathbf{a}_{u_{t,D+2}}, \alpha_{u_{t,1}}, \alpha_{u_{t,2}}, \alpha_{u_{t,D+2}})$. A similar reasoning as in Theorem 4.3 applies to the sequence of vectors $\mathbf{q}_t$ so that the sequence $(\mathbf{q}_t)$ has cluster points. Let $\mathbb{K}$ be an infinite set of indices such that $\mathbf{w}_t^* \overset{\mathbb{K}}{\to} \mathbf{w}^*$, then there exists an infinite set $\mathbb{K}1 \subset \mathbb{K}$ such that $\forall j = 1..(D+2)$ $(\mathbf{a}_{u_{t,j}}, \mathbf{w}_{u_{t,j}}, \alpha_{u_{t,j}}) \overset{\mathbb{K}1}{\to} (\mathbf{a}_j^*, \mathbf{w}_j^*, \alpha_j^*)$. Then $(\tilde{\mathbf{a}}_t, \tilde{s}_t^t) \overset{\mathbb{K}1}{\to} (\tilde{\mathbf{a}}^*, \tilde{s}^*)$ with $(\tilde{\mathbf{a}}^*, \tilde{s}^*) = \sum_{j=1..D+2} \alpha_j^* (\mathbf{a}_j^*, s_j^*)$.

We show now that $\forall j, \alpha_j^* > 0 \Rightarrow \mathbf{w}_j^* \equiv \mathbf{w}^*$. First the the measure definition obeys:

$$\tilde{s}_t^t = \sum_{j=1..t} \alpha_j s_j^t + \alpha_{t+1} \tilde{s}_{t-1}^t \tag{44}$$

Then from Lemma A.1, Eq. (43):

$$gap_t \geq \tilde{s}_t^t = \sum_{j=1..D+2} \alpha_{u_{t,j}} s_{u_{t,j}}^t \geq \sum_{j=1..D+2} \alpha_{u_{t,j}} \|\mathbf{w}_{u_{t,j}} - \mathbf{w}_t^*\|^2 \tag{45}$$

Then, since $gap_t \overset{t\to\infty}{\to} 0$ we get that $\forall j \ s.t. \ \alpha_j^* > 0 \ : \ \|\mathbf{w}_{u_{t,j}} - \mathbf{w}_t^*\|^2 \overset{t\to\infty}{\to} 0$. Finally using that $\mathbf{w}_{u_{t,j}} \overset{\mathbb{K}1}{\to} \mathbf{w}_j^*$ and $\mathbf{w}_t^* \overset{\mathbb{K}1}{\to} \mathbf{w}^*$, we obtain that $\forall j, \ \alpha_j^* > 0 \Rightarrow \mathbf{w}_j^* \equiv \mathbf{w}^*$.

Next recalling that $\tilde{\mathbf{w}}_t = -\frac{\tilde{\mathbf{a}}_t}{\lambda}$, $\tilde{\mathbf{w}}_t \overset{\mathbb{K}}{\to} \mathbf{w}^*$ and $\tilde{\mathbf{a}}_t \overset{\mathbb{K}}{\to} \tilde{\mathbf{a}}^*$:

$$\begin{array}{lll} & \mathbf{w}^* & = -\frac{\tilde{\mathbf{a}}^*}{\lambda} \\ \iff & \mathbf{w}^* & = -\sum_{j=1}^{D+2} \frac{\alpha_j^* \mathbf{a}_j^*}{\lambda} \\ \iff & \lambda \mathbf{w}^* + \sum_{j=1}^{D+2} \alpha_j^* \mathbf{a}_j^* & = \mathbf{0} \end{array} \tag{46}$$

In the following we show that $\sum_{j=1}^{D+2} \alpha_j^* \mathbf{a}_j^*$ is a subgradient of R at $\mathbf{w}^*$. Based on this result $\lambda \mathbf{w}^* + \sum_{j=1}^{D+2} \alpha_j^* \mathbf{a}_j^*$ is a subgradient of f at $\mathbf{w}^*$, and since this subgradient is null $\mathbf{w}^*$ is a stationary solution of f.

As a first possibility every $\mathbf{a}_j^*$ (with non-null α_j^*) may be a subgradient of R at $\mathbf{w}^*$ (recall that we showed that $\mathbf{w}_j^* \equiv \mathbf{w}^*$). This is the case if all $(\mathbf{w}_j^t)_{t \in K1}$ are subgradients of R since in this case $\mathbf{a}_{u_{t,j}} \overset{\mathbb{K}1}{\to} \mathbf{a}_j^*$ and R being Lipschitz continuous implies that $\mathbf{a}_j^*$ is a subgradient of R. Furthermore a convex combination of subgradients is a subgradient, hence in this case $\sum_{j=1}^{D+2} \alpha_j^* \mathbf{a}_j^*$ is a subgradient of R.

Unfortunately, the NullStep2 case in our Algorithm 3) makes that a particular $\mathbf{a}_{u_{t,j}}$ may not be a subgradient of R at $\mathbf{w}_{u_{t,j}}$.

In this case we have to rely on Hypothesis H2. Let T_L be the index of the last NullStep2 iteration and let k denote the number of a NullStep2 iteration (hence $k \leq T_L$). Since there are only descent steps after iteration T_L, and since a locality measure cannot decrease: $\forall t > T_L$, $s_k^t > s_k^k > 0$ (cf. Line 5 in Algorithm 2). This has to be contrasted to the fact that $\forall j, \alpha_j^* > 0 \Rightarrow \mathbf{s}_{u_{t,j}}^t \overset{\mathbb{K}1}{\to} 0$. Hence there exists an infinite subset $\mathbb{K}2 \subset \mathbb{K}1$ (with indices larger

than T_L) such that : $\forall t \in \mathbb{K}2$ $\mathbf{s}^t_{u_{t,j}} < s^k_k$. As a conclusion, iteration number $s^t_{u_{t,j}}$ cannot correspond to a NullStep2 iteration, and $\mathbf{a}_{u_{t,j}}$ is actually a subgradient of R at $\mathbf{w}^{t,j}$.

Finally, since $(\mathbf{a}_{u_{t,j}}, \mathbf{w}_{u_{t,j}}, \alpha_{u_{t,j}}) \overset{\mathbb{K}2}{\to} (\mathbf{a}^*_j, \mathbf{w}^*_j, \alpha^*_j)$, and assuming R is Lipschitz continuous, $\mathbf{a}^*_j$ (for j such that $\alpha^*_j > 0$) are subgradient of R at $\mathbf{w}^*$ (Luksan and Vlcek, 2000). Then $\mathbf{w}^*$ is a stationary solution of f.

■

4.5 Extension to Line Search Variant

All previous results hold for the *full line search strategy* except that we have to show that the norm of the normal vectors a_i of all cutting planes is bounded by a constant $\overline{G}$. Since Lemma A.3 shows this is true for all CP built at all curent solutions $\mathbf{w}_t$, it is sufficient here to show this is true for the solutions found by linesearch.

Actually using Hypothesis H1, it is easy to show that:

$$S(\mathbf{w}_1) = \{\mathbf{w} \mid f(\mathbf{w}) \leq f(\mathbf{w}_1)\} \subset Ball(\mathbf{w}_1, 2\|\mathbf{w}_1\| + \frac{2G}{\lambda})$$

Let note $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_1$. Then:

$$\begin{array}{lll} & \frac{\lambda}{2}\|\mathbf{w}\|^2 + R(\mathbf{w}) & \leq \frac{\lambda}{2}\|\mathbf{w}_1\|^2 + R(\mathbf{w}_1) \\ \Leftrightarrow & \frac{\lambda}{2}\|\mathbf{w}_1 + \Delta\mathbf{w}\|^2 + R(\mathbf{w}_1) - (R(\mathbf{w}_1) - R(\mathbf{w})) & \leq \frac{\lambda}{2}\|\mathbf{w}_1\|^2 + R(\mathbf{w}_1) \\ \Rightarrow & \lambda\langle\Delta\mathbf{w}, \mathbf{w}_1\rangle + \frac{\lambda}{2}\|\Delta\mathbf{w}\|^2 - G\|\Delta\mathbf{w}\| & \leq 0 \\ \Rightarrow & \frac{\lambda}{2}\|\Delta\mathbf{w}\|(\|\Delta\mathbf{w}\| - 2\|\mathbf{w}_1\| - \frac{2G}{\lambda}) & \leq 0 \end{array} \quad (47)$$

And we finally get that a solution $\mathbf{w}$ of a line search necessarily lies in a ball centered at $\mathbf{w}_1$ with radius $2\|\mathbf{w}_1\| + \frac{2G}{\lambda})$. Hence the norm of the normal vectors of all CP in the *full linesearch strategy* is bounded by a constant and the proofs in the appendix hold for this variant.

Finally, the things are more complicated for the *greedy line search strategy* and the proofs do not hold anymore in their actual shape. Yet, such a strategy is less expensive than the *full* one and it is efficient in practice. All results of the line search variant in the experiment section have been gained using this implementation.

References

T. Joachims. Training linear SVMs in linear time. In *ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD)*, pages 217 – 226, 2006.

Ladislav Luksan and Jan Vlcek. Introduction to nonsmooth analysis. theory and algorithms. Technical report, Universita degli Studi di Bergamo, 2000.

Gregory H. Moore. The emergence of open sets, closed sets, and limit points in analysis and topology. *Hist. Math.*, 35(3):220–241, 2008.

Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated subgradient solver for svm. In *ICML '07*, pages 807–814, New York, USA, 2007. ACM Press. ISBN 978-1-59593-793-3. doi: http://doi.acm.org/10.1145/1273496.1273598.

Alex Smola, S V N Vishwanathan, and Quoc Le. Bundle methods for machine learning. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *NIPS 20*, pages 1377–1384, Cambridge, MA, 2008. MIT Press.

Choon Hui Teo, Quoc Viet Le, Alex Smola, and S. V.N. Vishwanathan. A scalable modular convex solver for regularized risk minimization. In *KDD '07*, pages 727–736, New York, USA, 2007. ACM. ISBN 978-1-59593-609-7. doi: http://doi.acm.org/10.1145/1281192.1281270.