Chemins de régularisation pour la régression ν-SVR

Gilles Gasso, Karina Zapien, Stéphane Canu

LITIS, EA 4108
Avenue de l'Université, BP 76801 Saint-Etienne du Rouvray, France
{gilles.gasso, karina.zapien, stephane.canu}@insa-rouen.fr

Résumé – L'article décrit le calcul des chemins de régularistion de l'algorithme ν-SVR. Dans la formulation classique de cet algorithme, l'utilisateur fournit deux hyper-paramètres : ν qui détermine la largeur du tube du coût ϵ-insensible optimisé par le SVR et le paramètre de régularisation λ qui règle le compromis entre la régularité de la fonction de régression et l'erreur. L'article présente une façon efficace d'explorer l'ensemble des solutions losrque les hyper-paramètres varient.

Abstract – The paper describes the computation of the full paths of the well-known ν-SVR. In the classical method, the user provides two parameters: the regularization parameter λ and ν which settles the width of the tube of the ϵ-insensitive cost optimized by SVR. The paper proposes an efficient way to get all the solutions when ν and λ vary.

1 Introduction

L'approche SVR (*Support vector regression*) est une technique maintenant bien connue pour traiter les problèmes de régression [1]. Elle découle directement des principes des algorithmes de type machine à vecteur support. Dans cette approche, on minimise un coût de type L_1 dit ϵ-insensible (défini par $\max(0, |y - f(x)| - \epsilon)$) avec une pénalisation $\|f\|^2$. Le compromis entre l'erreur et la pénalité sur la régularité de f est assuré par un paramètre de régularisation λ qui est à choisir. En plus de λ, l'utilisateur doit fournir la valeur ϵ de la largeur du tube.

En général, pour une application donnée, il est difficile de spécifier la bonne valeur de ϵ. Pour contourner ce problème, l'approche ν-SVR a été introduite et permet la détermination automatique de ϵ [1]. De plus $0 \leq \nu \leq 1$ a une interprétation intuitive car elle définit la borne inférieure de la proportion du nombre de points supports (ce qui introduit la parcimonie de la fonction de régression) et la borne supérieure de la proportion de points pouvant être en dehors du tube. Malgré l'interprétation qu'on peut associer à la valeur de ν, son choix automatique par l'utilisateur pour une application donnée reste problématique.

Plusieurs travaux ont été dédiés au choix des deux hyper-paramètres. Certains reposent sur une recherche en grille (*grid search*) dans l'espace des hyper-paramètres couplée avec l'exploitation de mesures de performances comme le critère de validation croisée ou des critères sur les bornes [2] pour aider au choix du bon modèle. D'autres méthodes font appel à l'optimisation non-linéaire de critère de validation croisée par rapport aux hyper-paramètres [3].

Tout récemment de nouvelles approches ont été étudiées et sont basées sur le calcul du chemin de régularisation [4, 5] c'est-à-dire le calcul d'une façon rapide de toutes les solutions optimales lorsque le paramètre de régularisation varie. En partant d'une solution initiale, les paramètres de la solution suivante sont simplement obtenus en résolvant un système linéaire. Compte tenu de l'efficacité de ces algorithmes, nous proposons ici de les adapter au choix des hyper-paramètres du ν-SVR. Ceci nous conduit à proposer deux chemins de régularisation (λ-chemin et ν-chemin) pour explorer l'espace des hyper-paramètres.

La suite de l'article décrit le calcul des chemins de régularisation et leurs tests sur quelques applications.

2 Formulation du ν-SVR

On dispose d'un ensemble de N données d'apprentissage $\{(x_i, y_i) \in \mathcal{X} \times \mathbb{R}\}$. La méthode de régression ν-SVR est basée sur l'optimisation du coût ϵ-insensible $L(y, f(x)) = \max(0, |y - f(x)| - \epsilon)$ représenté sur la figure 1.

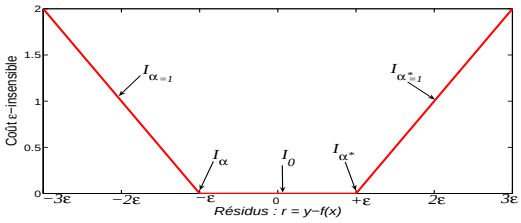

FIG. 1 – Illustration du coût ϵ-insensible

La formulation primale du problème s'écrit :

$$\begin{cases} \min_{f,\epsilon,\xi,\xi^*} \quad \frac{\lambda}{2}\|f\|^2_{\mathcal{H}} + \nu\epsilon + \sum_{i=1}^N (\xi_i + \xi_i^*) \\ \text{s.c.} \quad -\epsilon - \xi_i \leq y_i - f(x_i) \leq \epsilon + \xi_i^*, \quad \forall i = 1, \ldots, N \\ \quad \xi_i \geq 0, \ \ \xi_i^* \geq 0, \quad \forall i = 1, \ldots, N \quad \text{et } \epsilon \geq 0 \end{cases}$$

où λ est le paramètre de régularisation et $f(x)$, la fonction de régression supposée appartenir à un espace de Hilbet à noyau reproduisant $\mathcal{H}$.

Remarquons que dans cette formulation, le paramètre ν varie entre 0 et N au lieu de l'intervalle $[0, 1]$. Par conséquent, ν détermine la borne supérieure du nombre de points pouvant se trouver à l'extérieur du tube et une borne inférieure sur le nombre de points supports de la fonction de régression. Cette dernière donnée par l'expression $f(x) = \frac{1}{\lambda}\sum_{i=1}^N (\alpha_i^* - \alpha_i)k(x_i, x) + b$ est la solution du problème primal. Dans cette équation, $k(.,.)$ est la fonction noyau et les α_i et α_i^* représentent les multiplicateurs de

Lagrange qui sont les solutions du problème dual (avec K la matrice de Gram) :

$$\begin{cases} \min\limits_{\alpha^*,\alpha\in\mathbb{R}^N} \frac{1}{2\lambda}(\alpha^*-\alpha)^\top K(\alpha^*-\alpha) - (\alpha^*-\alpha)^\top \mathbf{y} \\ \text{s.c.} \quad (\alpha^*-\alpha)^\top \mathbf{1}_N = 0 \quad \text{et} \quad (\alpha^*+\alpha)^\top \mathbf{1}_N \leq \nu \\ \qquad 0 \leq \alpha_i,\ \alpha_i^* \leq 1, \quad \forall i = 1,\ldots,N \end{cases} \tag{1}$$

3 Les chemins de régularisation

Supposons λ et ν fixés. Les conditions KKT permettent de calculer b et ϵ [1]. La qualité du modèle obtenu dépend des valeurs choisies. Plutôt que de faire une recherche exhaustive dans l'espace de ces hyper-paramètres, nous proposons dans cet article d'analyser de façon efficace et rapide l'évolution de $f(x)$ en fonction de λ et ν. Pour ce faire, en fixant ν, on peut calculer l'ensemble des solutions en fonction de λ : c'est le λ-chemin. Inversément, pour λ fixé, il est possible d'examiner la qualité de $f(x)$ en fonction de ν, ce qui donne le ν-chemin. En exploitant l'idée originelle de Gunter et Zhu [5], il est aisé de montrer que sur ces chemins les paramètres du modèle varient linéairement par morceaux. La problématique est alors de déterminer les zones et les directions de variation linéaire.

Considérons les ensembles de points suivants (correspondant aux différentes zones du critère sur la figure 1) :

$\mathcal{I}_{\alpha=1}$:	$r_i < -\epsilon$,	$\alpha_i = 1$,	$\alpha_i^* = 0$
$\mathcal{I}_{\alpha^*=1}$:	$r_i > \epsilon$,	$\alpha_i = 0$,	$\alpha_i^* = 1$
$\mathcal{I}_0$:	$\lvert r_i\rvert < \epsilon$,	$\alpha_i = 0$,	$\alpha_i^* = 0$
$\mathcal{I}_\alpha$:	$r_i = -\epsilon$,	$0 \leq \alpha_i \leq 1$,	$\alpha_i^* = 0$
$\mathcal{I}_{\alpha^*}$:	$r_i = \epsilon$,	$\alpha_i = 0$,	$0 \leq \alpha_i^* \leq 1$

avec $r_i = y_i - f(x_i)$ le résidu. Les ensembles $\mathcal{I}_{\alpha=1}$ et $\mathcal{I}_{\alpha^*=1}$ contiennent respectivement les points dont les résidus se situent respectivement dans les parties gauche et droite du critère. Les points de $\mathcal{I}_\alpha$ et $\mathcal{I}_{\alpha^*}$ appartiennent aux coudes gauche et droite du critère. Enfin $\mathcal{I}_0$ contient les points se trouvant dans le tube, c'est-à-dire de coût nul selon le critère. Précisons que pour une fonction $f(x)$ donnée correspondant à la solution du problème (1), chaque donnée d'apprentissage est classée dans un et un seul de ces ensembles. Pour des valeurs de λ et ν fixées, les ensembles sont aussi fixés et ainsi donc $f(x)$. L'intérêt des chemins de régularisation est de déterminer de façon efficiente les mouvements des points entre ces ensembles en fonction des hyper-paramètres, ce qui donne l'évolution de $f(x)$.

3.1 Calcul du λ-chemin

On suppose ν constant. La fonction f peut s'écrire :

$$f(x) = \frac{1}{\lambda}\left(\sum_{i=1}^{N}(\alpha_i^* - \alpha_i)k(x_i,x) + \beta_0\right)$$

avec $\beta_0 = \lambda b$. Soit $f^t(x)$, la solution correspondant à λ^t. Les ensembles associés sont notés $\mathcal{I}_{\alpha=1}^t$, $\mathcal{I}_{\alpha^*=1}^t$, $\mathcal{I}_0^t$, $\mathcal{I}_{\alpha^*}^t$ et $\mathcal{I}_\alpha^t$. On montre plus loin que les paramètres de $f^t(x)$ varient linéairement en fonction de λ tant que les ensembles demeurent inchangés. Le point clé de l'algorithme est donc la détermination de la valeur particulière de λ pour laquelle un ou des points changent d'ensemble. Ce changement arrive si l'un de ces mouvements potentiels se produit : de $\mathcal{I}_\alpha^t$ à $\mathcal{I}_{\alpha=1}^t$ ($Evt1$), de $\mathcal{I}_{\alpha^*}^t$ à $\mathcal{I}_{\alpha^*=1}^t$ ($Evt2$), de $\mathcal{I}_\alpha^t$ ou $\mathcal{I}_{\alpha^*}^t$ à $\mathcal{I}_0^t$ ($Evt3$), de $\mathcal{I}_{\alpha=1}^t$ à $\mathcal{I}_\alpha^t$ ($Evt4$), de $\mathcal{I}_{\alpha^*=1}^t$ à $\mathcal{I}_{\alpha^*}^t$ ($Evt5$) et finalement de $\mathcal{I}_0^t$ à $\mathcal{I}_\alpha^t$ ou de $\mathcal{I}_0^t$ à $\mathcal{I}_{\alpha^*}^t$ ($Evt6$). Pour déterminer sous quelles conditions ces mouvements interviennent, nous exprimons la fonction de régression sous la forme $\lambda f(x) = \lambda f(x) - \lambda^t f^t(x) + \lambda^t f^t(x)$ pour $\lambda^{t+1} < \lambda < \lambda^t$. On a alors

$$\lambda f(x) = \sum_{i\in\mathcal{I}_\alpha^t\cup\mathcal{I}_{\alpha^*}^t} (\delta\alpha_i^* - \delta\alpha_i)\, k(x_i,x) + \delta\beta_0 + \lambda^t f^t(x) \tag{2}$$

avec $\delta\alpha_i = \alpha_i - \alpha_i^t$, $\delta\alpha_i^* = \alpha_i^* - \alpha_i^{*t}$ et $\delta\beta_0 = \beta_0 - \beta_0^t$. Dans l'équation (2), la somme est prise seulement sur $\mathcal{I}_\alpha^t$ et $\mathcal{I}_{\alpha^*}^t$ car les paramètres de Lagrange associés aux points des autres ensembles sont fixes et valent 0 ou 1. L'équation (2) montre qu'en faisant varier λ, la fonction de régression correspondante f s'obtient à partir de f^t et des variations des paramètres associés aux points dans $\mathcal{I}_\alpha^t$ et $\mathcal{I}_{\alpha^*}^t$ (les points sur le tube) et du biais b.

Pour calculer ces variations, nous exploitons les conditions liées aux points se trouvant sur le tube. En effet, pour $j \in \mathcal{I}_\alpha^t$, on a : $y_j - f^t(x_j) = -\epsilon^t$. Si pour $f(x)$, ce point reste dans $\mathcal{I}_\alpha^t$, on aurait $y_j - f(x_j) = -\epsilon$. Par conséquent, en utilisant (2), la relation suivante est vérifiée :

$$\lambda(y_j+\epsilon) = \sum_{i\in\mathcal{I}_{\alpha^*}^t\cup\mathcal{I}_\alpha^t} (\delta\alpha_i^* - \delta\alpha_i)k(x_i,x_j) + \delta\beta_0 + \lambda^t(y_j+\epsilon^t)$$

Posons $d = \lambda\epsilon$ et $\delta d = \lambda\epsilon - \lambda^t\epsilon^t$. En réarrangeant cette dernière équation nous obtenons $\forall j \in \mathcal{I}_\alpha^t$:

$$(\lambda-\lambda^t)y_j = \sum_{i\in\mathcal{I}_{\alpha^*}^t\cup\mathcal{I}_\alpha^t} (\delta\alpha_i^* - \delta\alpha_i)k(x_i,x_j) + \delta\beta_0 - \delta d \tag{3}$$

De façon similaire, on établit que $\forall j \in \mathcal{I}_{\alpha^*}^t$, :

$$(\lambda-\lambda^t)y_j = \sum_{i\in\mathcal{I}_{\alpha^*}^t\cup\mathcal{I}_\alpha^t} (\delta\alpha_i^* - \delta\alpha_i)k(x_i,x_j) + \delta\beta_0 + \delta d \tag{4}$$

En considérant les deux dernières séries d'équations, nous avons à ce stade $|\mathcal{I}_\alpha| + |\mathcal{I}_{\alpha^*}|$ équations pour $|\mathcal{I}_\alpha^t| + |\mathcal{I}_{\alpha^*}^t| + 2$ inconnues ($\delta\alpha_i, i \in \mathcal{I}_\alpha^t$, $\delta\alpha_i^*, i \in \mathcal{I}_{\alpha^*}^t$, $\delta\beta_0$ et δd qui fournit une indication sur la variation de la largeur du tube). Le système est donc sous-déterminé. Pour rendre la solution unique, exploitons les contraintes du problème dual (équation 1). De la contrainte d'équilibrage $(\alpha^*-\alpha)^\top\mathbf{1} = 0$, on déduit :

$$\sum_{i\in\mathcal{I}_{\alpha^*}^t\cup\mathcal{I}_\alpha^t} (\delta\alpha_i^* - \delta\alpha_i) = 0 \tag{5}$$

La deuxième contrainte $(\alpha^*+\alpha)^\top\mathbf{1} \leq \nu$ fournit l'équation manquante. En effet, si la largeur du tube n'est pas nulle, les conditions KKT impliquent que cette condition inégalité devient une contrainte égalité. Par conséquent, pour peu qu'on s'assure dans l'algorithme de la non-nullité de ϵ, on peut alors écrire (noter ici que ν est fixe) :

$$\sum_{i\in\mathcal{I}_{\alpha^*}^t\cup\mathcal{I}_\alpha^t} (\delta\alpha_i^* + \delta\alpha_i) = 0 \tag{6}$$

En regroupant les équations (3-6), on aboutit alors à ce système linéaire :

$$A\boldsymbol{\delta\theta} = (\lambda-\lambda^t)\mathbf{z} \text{ où } \boldsymbol{\delta\theta} = \begin{bmatrix} \delta\boldsymbol{\alpha} & \delta\boldsymbol{\alpha}^* & \delta\beta_0 & \delta d \end{bmatrix}^\top$$

$$A=\begin{bmatrix} -K(\mathcal{I}_\alpha^t,\mathcal{I}_\alpha^t) & K(\mathcal{I}_\alpha^t,\mathcal{I}_{\alpha^*}^t) & \mathbf{1} & -\mathbf{1} \\ -K(\mathcal{I}_\alpha^t,\mathcal{I}_{\alpha^*}^t)^\top & K(\mathcal{I}_{\alpha^*}^t,\mathcal{I}_{\alpha^*}^t) & 1 & 1 \\ -\mathbf{1}^\top & \mathbf{1}^\top & 0 & 0 \\ \mathbf{1}^\top & \mathbf{1}^\top & 0 & 0 \end{bmatrix}, \mathbf{z}=\begin{bmatrix} \mathbf{y}_{\mathcal{I}_\alpha^t} \\ \mathbf{y}_{\mathcal{I}_{\alpha^*}^t} \\ 0 \\ 0 \end{bmatrix}$$

Soit $\eta = A^{-1}\mathbf{z}$. On constate alors que les paramètres du modèle de régression sont linéaires par rapport à λ :

$$\alpha^{t+1} = \alpha^t + (\lambda - \lambda^t)\eta_\alpha,\ \alpha^{*t+1} = \alpha^{*t} + (\lambda - \lambda^t)\eta_{\alpha^*} \quad (7)$$

$$\beta_0^{t+1} = \beta_0^t + (\lambda - \lambda^t)\eta_{\beta_0} \quad (8)$$

$$d^{t+1} = d^t + (\lambda - \lambda^t)\eta_d \text{ avec } d = \lambda\epsilon,\ d^t = \lambda^t\epsilon^t \quad (9)$$

Cette variation linéaire est valable tant que les ensembles $\mathcal{I}_\alpha^t$, $\mathcal{I}_{\alpha^*}^t$, $\mathcal{I}_{\alpha=1}^t$, $\mathcal{I}_{\alpha^*=1}^t$ et $\mathcal{I}_0^t$ sont inchangés. Examinons comment déterminer les valeurs de λ qui conduisent à un mouvement d'un point d'un ensemble vers un autre.

3.1.1 Points dans $\mathcal{I}_\alpha^t$ ou $\mathcal{I}_{\alpha^*}^t$ et détection des mouvements $Evt1$, $Evt2$ et $Evt3$

L'occurence de ces évènements intervient respectivement quand les paramètres α atteignent leur borne 1, les paramètres α^* atteignent la borne 1 et les paramètres α et α^* s'annulent 0. En utilisant l'équation (7), on déduit la valeur de λ correspondant à ces évènements :

$$\lambda_{Evt1}^{t+1} = \frac{1-\alpha_i^t}{\eta_{\alpha_i}} + \lambda^t,\ i \in \mathcal{I}_\alpha^t\ ;\ \lambda_{Evt2}^{t+1} = \frac{1-\alpha_i^{*t}}{\eta_{\alpha_i^*}} + \lambda^t,\ i \in \mathcal{I}_{\alpha^*}^t$$

$$\lambda_{Evt3}^{t+1} = \left\{ \frac{-\alpha_i^t}{\eta_{\alpha_i}} + \lambda^t, \quad i \in \mathcal{I}_\alpha^t \right\} \cup \left\{ \frac{-\alpha_i^{*t}}{\eta_{\alpha_i^*}} + \lambda^t, \quad i \in \mathcal{I}_{\alpha^*}^t \right\}$$

3.1.2 Points dans $\mathcal{I}_{\alpha=1}^t$, $\mathcal{I}_{\alpha^*=1}^t$, $\mathcal{I}_0^t$ et détection de $Evt4$, $Evt5$ et $Evt6$

Ces mouvements se produisent respectivement pour $r_i = -\epsilon^{t+1}$ pour des points $i \in \mathcal{I}_{\alpha=1}^t$, $r_i = \epsilon^{t+1}$ pour des points $i \in \mathcal{I}_{\alpha^*=1}^t$ et pour $|r_i| = \epsilon^{t+1}$, avec $i \in \mathcal{I}_0^t$. Rappelons que $r_i = y_i - f(x_i)$ est le résidu.

En injectant les équations (7 - 8) dans (2) et après quelques calculs mathématiques, on obtient :

$$f(x) = \frac{\lambda^t}{\lambda}\left[f^t(x) - h^t(x)\right] + h^t(x)$$

avec $h^t(x) = \sum_{i \in \mathcal{I}_{\alpha^*}^t \cup \mathcal{I}_\alpha^t} (\delta\alpha_i^* - \delta\alpha_i)k(x_i, x_j) + \delta\beta_0$. En se basant sur (9) et la dernière équation, on établit que les valeurs de λ associées aux évènements 4 à 6 sont :

$$\lambda_{Evt4}^{t+1} = \frac{\lambda^t\left(f^t(x_i) - h^t(x_i) - \epsilon^t + \eta_d\right)}{y_i - h^t(x_i) + \eta_d}$$

$$\lambda_{Evt5}^{t+1} = \frac{\lambda^t\left(f^t(x_i) - h^t(x_i) + \epsilon^t - \eta_d\right)}{y_i - h^t(x_i) - \eta_d}$$

$$\lambda_{Evt6}^{t+1} = \left\{\lambda_{Evt4}^{t+1}, \quad i \in \mathcal{I}_0^t\right\} \cup \left\{\lambda_{Evt5}^{t+1}, \quad i \in \mathcal{I}_0^t\right\}$$

3.1.3 Algorithme du λ-chemin

A l'étape $t+1$ de l'algorithme, la valeur de λ^{t+1} à retenir est la plus grande des valeurs obtenues aux sections 3.1.1 et 3.1.2 et immédiatement inférieures à λ^t. Nous initialisons notre algorithme en résolvant le problème dual avec une valeur de λ_0 élevée. On fait ensuite tourner l'algorithme (détecter le bon évènement et trouver la bonne valeur de λ, mettre à jour les paramètres et les ensembles) jusqu'à ce que $\mathcal{I}_{\alpha^*}$ ou $\mathcal{I}_\alpha$ devienne vide ou λ très petit.

Précisons qu'au cours de l'algorithme, pour des raisons de continuité, un point en dehors du tube ne passe dans le tube qu'après avoir transité sur le tube et vice-versa. Par exemple, un point dans $\mathcal{I}_{\alpha^*=1}$ ne peut passer dans $\mathcal{I}_0$ que s'il a transité par $\mathcal{I}_{\alpha^*}$. C'est cette continuité qui assure que l'algorithme fournit l'ensemble des solutions $f(x)$ quand λ varie.

3.2 Calcul du ν-chemin

On suppose ici que le paramètre λ est fixé et on examine l'influence de ν sur la fonction de régression. La démarche est très similaire à celle du λ-chemin. Soit $f^t(x)$ la solution correspondant à ν^t. Considérons également $\nu^t < \nu < \nu^{t+1}$ tel que les ensembles de points de l'étape t demeurent inchangés. λ étant constant, de la relation (2), on écrit :

$$\lambda\left(f(x) - f^t(x)\right) = \sum_{i \in \mathcal{I}_{\alpha^*}^t \cup \mathcal{I}_\alpha^t} (\delta\alpha_i^* - \delta\alpha_i)k(x_i, x)) + \delta\beta_0.$$

Comme dans le cas du λ-chemin, les points appartenant aux ensembles $\mathcal{I}_\alpha$ et $\mathcal{I}_{\alpha^*}$ vérifient nécessairement les conditions respectives $y_i - f(x_i) = -\epsilon$ et $y_i - f(x_i) = \epsilon$ et on en déduit par conséquent :

$$\sum_{i \in \mathcal{I}_{\alpha^*}^t \cup \mathcal{I}_\alpha^t} (\delta\alpha_i^* - \delta\alpha_i)k(x_i, x_j) + \delta\beta_0 - \delta d = 0 \quad \forall j \in \mathcal{I}_\alpha^t \quad (10)$$

$$\sum_{i \in \mathcal{I}_{\alpha^*}^t \cup \mathcal{I}_\alpha^t} (\delta\alpha_i^* - \delta\alpha_i)k(x_i, x_j) + \delta\beta_0 + \delta d = 0 \quad \forall j \in \mathcal{I}_{\alpha^*}^t \quad (11)$$

avec cette fois-ci $\delta d = \lambda(\epsilon - \epsilon^t)$ puisqu'on considère ici le paramètre λ fixe. Pour calculer la variation des paramètres, nous exploitons aussi les contraintes du problème dual : ainsi l'équation (5) reste valable alors que la contrainte inégalité $(\alpha^* + \alpha)^\top\mathbf{1} \leq \nu$ est transformée en une contrainte égalité $\sum_{i \in \mathcal{I}_{\alpha^*}^t \cup \mathcal{I}_\alpha^t} (\delta\alpha_i^* + \delta\alpha_i) = \nu - \nu^t$ car ici ν varie.

En regroupant toutes ces équations, on établit un système linéaire $A\boldsymbol{\delta\theta} = (\nu - \nu^t)\mathbf{w}$ avec $\mathbf{w} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & 0 & 1 \end{bmatrix}^\top$ Les paramètres sont donc linéaires par rapport à ν comme dans les équations (7) à (9). Les valeurs de ν correspondant aux différents évènements sont calculées en appliquant les mecanismes exposées dans les sections 3.1.1 et 3.1.2. Précisons que les évènements $Evt4$, $Evt5$ and $Evt6$ sont déterminés en utilisant la relation $f(x) = f^t(x) + \frac{\nu - \nu^t}{\lambda}h^t(x)$ obtenue en remplaçant dans l'expression de $f(x)$ les équations de mise à jour des paramètres par rapport à ν.

Le ν-chemin est donc similaire au λ-chemin. Son initialisation est relativement facile puisqu'on peut fixer ν à l'une de ses valeurs seuil. Choisir $\nu \approx 0$ signifie qu'aucun point d'apprentissage n'est autorisé à être dans le tube : le tube est large et la solution obtenue est parcimonieuse car la plupart des points appartiennent à $\mathcal{I}_\alpha$ ou $\mathcal{I}_{\alpha^*}$ (sur les frontières du tube) ou à $\mathcal{I}_0$ (dans le tube). L'autre alternative est $\nu \approx m$: le tube a une largeur très fine et la solution initiale est nettement moins parcimonieuse.

Hormis l'initialisation de ces algorithmes, une autre question intéressante est comment passer du ν-chemin au λ-chemin ? Comme à chaque étape des algorithmes, les paramètres et les ensembles des points sont connus, le passage d'un chemin à l'autre se fait aisément pour peu que les ensembles $\mathcal{I}_\alpha$ ou $\mathcal{I}_{\alpha^*}$ ne soient pas vides. Remarquons que les paramètres λ et ν peuvent coître ou décroître sur les chemins. Il suffira d'adapter le choix du bon évènement.

4 Application des algorithmes

Les chemins de régularisation sont testés dans un premier temps sur un problème jouet consistant à approximer la fonction non-linéaire $y = sin(exp(3 * x))$. Les observations x_i sont échantillonnées selon une loi uniforme sur l'intervalle [0, 1]. L'ensemble d'apprentissage comporte 150 points. Un noyau gaussien de largeur de bande 0.1 a été utilisé. A chaque étape, les performances en généralisation de la fonction de régression sont évaluées en calculant l'erreur LOO (leave-one-out) $LOO = \frac{1}{m}\sum_{i=1,i\neq k}^{m} |y_i - f_k(x_i)|$ avec $f_k(x)$, la fonction obtenue en excluant le point k de l'ensemble d'apprentissage. Les résultats issus de l'application du λ-chemin sont reportés sur la figure 2. Précisons que l'évolution de l'erreur LOO est similaire à celle d'un critère de validation croisée non présenté ici pour des raisons de manque place. On constate que quand λ diminue, l'erreur LOO décroit rapidement ce qui permet l'arrêt prématuré de l'algorithme. Le même constat peut être fait en regardant l'évolution de la largeur du tube. Pour les faibles valeurs de ν, comme la solution initiale est très parcimonieuse, le calcul du LOO est rapide. L'arrêt prématuré de l'algorithme fournit donc une solution parcimonieuse et on n'a pas besoin d'explorer la totalité du chemin.

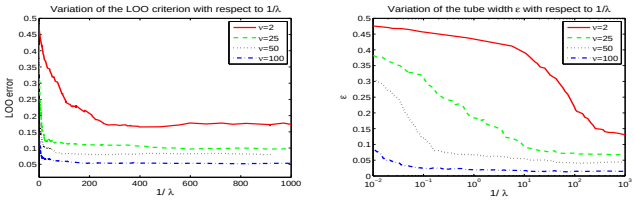

FIG. 2 – Illustration du λ-chemin. L'axe des abscisses du 2e graphique est en échelle logarithmique

L'illustration du ν-chemin pour différentes valeurs de λ est portée sur la figure 3. Pour des valeurs croissantes de ν, le tube rétrécit et l'erreur LOO décroit. Ces constats sont cohérents avec ceux du λ-chemin.

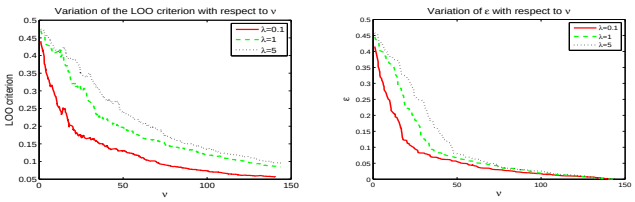

FIG. 3 – Illustration du ν-chemin.

Nous avons ensuite évalué les performances des algorithmes en termes de temps de calcul. Ainsi pour le problème jouet, nous obtenons pour le λ-chemin les résultats du tableau 1 (données moyennées sur 10 tests et pour différentes valeurs de ν fixées a priori). L'apprentissage du

TAB. 1 – Temps de calcul du λ-chemin et d'une procédure d'apprentissage de ν-SVR avec *warm restart*. $N = 1500$.

	$\nu = 0.01N$	$\nu = 0.5N$	$\nu = 0.75N$
λ-chemin	1.70 ± 0.076	1.95 ± 0.03	2 ± 0.031
WR	4.30 ± 0.053	21.8 ± 0.15	21.15 ± 0.12

ν-SVR basé sur la procédure du « redémarrage à chaud » (*warm restart*) consiste à utiliser les paramètres de la fonction de régression courante comme initialisation dans la résolution du problème dual. On constate que le calcul du chemin de régularisation est plus performant (gain de temps de l'ordre de 11) et donne toutes les solutions à partir desquelles l'utilisateur choisira son meilleur modèle. Par contre, la procédure de démarrage à chaud doit être vue ici comme une méthode de *grid search* sur des valeurs de λ déterminant les points de cassure de la variation linéaire par morceaux sur le chemin et ne donne pas toutes les solutions. Des performances de rapidité du ν-chemin sont également observées(non reportées ici).

Pour étudier l'influence des paramètres du noyau, le λ-chemin a été testé sur des données de Boston Housing (dépôt UCI). Des noyaux gaussiens avec différentes largeurs de bande σ sont utilisés. Les résultats moyennés sur 10 tests sont portés dans le tableau 2. On remarque que les gains de temps vont de 4 à 9. Cette différence s'explique par le fait que σ influence le nombre de mouvements des points d'un ensemble de points à l'autre ce qui peut accéler ou ralentir l'algorithme. Néanmoins le calcul du chemin est plus efficient qu'un *grid search*.

TAB. 2 – Données Boston Housing. Comparaison des temps de calcul. $N = 406$ points $x_i \in \mathbb{R}^{13}$.

$\sigma = 1$			
	$\nu = 0.01N$	$\nu = 0.5N$	$\nu = 0.75N$
λ-chemin	0.95 ± 0.32	1.95 ± 0.35	2.06 ± 1.31
WR	8.6 ± 1.96	13.08 ± 5.17	13.77 ± 5.15
$\sigma = 0.1$			
	$\nu = 0.01$	$\nu = 0.5$	$\nu = 0.75$
λ-chemin	12.31 ± 0.34	12.29 ± 0.44	12.27 ± 0.38
WR	51.44 ± 0.78	51.63 ± 1.24	51.32 ± 0.95

5 Conclusion

Cet article a présenté le calcul efficace de l'ensemble des solutions $f(x)$ d'un problème de régression en utlisant l'approche ν-SVR. Ces chemins donnent automatiquement l'évolution de $f(x)$ en fonction des hyper-paramètres. Les extensions de ce travail sont de deux sortes : en premier lieu, l'exploitation intelligente de ces chemins pour balayer l'espace des hyper-paramètres et en deuxième lieu l'évaluation de l'efficacité de ces algorithmes pour des larges bases de données.

Références

[1] B. Schölkopf and A. Smola. *Leaning with Kernels*. MIT Press, 2001.

[2] M.-W. Chang and C.-J. Lin. Leave-one-out bounds for support vector regression model selection. *Neural Computation*, 17 :1188–1222, 2005.

[3] K. Kobayashi, D. Kitakoshi, and R. Nakano. Yet faster method to optimize svr hyperparameters based on minimizing cross-validation error. In *Proc. of the IJCNN05*, 2005.

[4] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *JMLR*, 5 :1391–1415, 2004.

[5] Lacey Gunter and Ji Zhu. Computing the solution path for the regularized support vector regression. In *NIPS*, 2005.