

## Appendix to “Model Averaging via Penalized Regression for Tracking Concept Drift” published in the Journal of Computational and Graphical Statistics

Kyupil Yeon, Moon Sup Song, Yongdai Kim, Hosik Choi, and Cheolwoo Park

### 1 Proof of Theorem 2

*Proof.* For brevity of notation, we replace  $\hat{f}_j(\mathbf{x}_i)$  with  $x_{ij}$ . Therefore, the objective function can be denoted by

$$S = \sum_{i=1}^n \left( y_i - \sum_{j=1}^m w_j x_{ij} \right)^2 + \lambda \sum_{j=1}^m w_j^2, \quad \sum_{j=1}^m w_j = 1.$$

Differentiating  $S$  with respect to  $w_k$  ( $k \neq m$ ), and equating it to 0, we have

$$\frac{\partial S}{\partial w_k} = 2 \sum_{i=1}^n (y_i - \sum_{j=1}^m w_j x_{ij})(x_{im} - x_{ik}) + 2\lambda(w_k - w_m) = 0,$$

which becomes the following equation with  $w_m = 1 - \sum_{j=1}^{m-1} w_j$ ,

$$\begin{aligned} & \left( \sum_{i=1}^n (x_{ik} - x_{im})^2 + 2\lambda \right) w_k + \sum_{j \neq k}^{m-1} \left( \sum_{i=1}^n (x_{ik} - x_{im})(x_{ij} - x_{im}) + \lambda \right) w_j \\ &= \sum_{i=1}^n (x_{ik} - x_{im})(y_i - x_{im}) + \lambda \end{aligned} \quad (4)$$

for  $k = 1, 2, \dots, m-1$ .

Let us define

$$\begin{aligned} a_k &= \sum_i (x_{ik} - x_{im})^2, \\ b_{kj} &= \sum_i (x_{ik} - x_{im})(x_{ij} - x_{im}), \\ c_k &= \sum_i (x_{ik} - x_{im})(y_i - x_{im}). \end{aligned}$$

Then (4) is denoted as a matrix form,

$$\begin{bmatrix} a_1 + 2\lambda & b_{12} + \lambda & \cdots & b_{1,m-1} + \lambda \\ b_{21} + \lambda & a_2 + 2\lambda & \cdots & b_{2,m-1} + \lambda \\ \vdots & \vdots & \vdots & \vdots \\ b_{m-1,1} + \lambda & b_{m-1,2} + \lambda & \cdots & a_{m-1} + 2\lambda \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{m-1} \end{bmatrix} = \begin{bmatrix} c_1 + \lambda \\ c_2 + \lambda \\ \vdots \\ c_{m-1} + \lambda \end{bmatrix}$$

or  $\mathbf{T}\mathbf{w} = \mathbf{C}$ . Thus, the solution to  $w_1, w_2, \dots, w_{m-1}$  will be obtained by  $\mathbf{w} = \mathbf{T}^{-1}\mathbf{C}$  and  $w_m = 1 - \sum_{j=1}^{m-1} w_j$ .

When  $\lambda \rightarrow \infty$ , only the coefficients of  $\lambda$  are important, that is,

$$\begin{bmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{m-1} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

and the solution is  $w_1 = w_2 = \cdots = w_{m-1} = 1/m$ , and also  $w_m = 1/m$ .  $\square$

## 2 Proof of Theorem 3

*Proof.* It suffices to show that  $\hat{\mathbf{w}}^T \hat{\mathbf{w}}$  is a decreasing function of  $\lambda$ .

The original setting (2) can be re-expressed using the Karush-Kuhn-Tucker (KKT) conditions. For  $\gamma = (\gamma_1, \dots, \gamma_m)$  and  $\alpha \geq 0$ , we minimize

$$\mathcal{L}(\mathbf{w}, \alpha, \gamma) = \sum_{i=1}^n \left( y_i - \sum_{j=1}^m w_j \hat{f}_j(\mathbf{x}_i) \right)^2 + \lambda \sum_{j=1}^m w_j^2 + \alpha \left( \sum_{j=1}^m w_j - 1 \right) - \sum_{j=1}^m \gamma_j w_j,$$

that is,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_k} &= 0, & \sum_{j=1}^m w_j &= 1, \\ \gamma_k w_k &= 0, & \gamma_k &\geq 0, \end{aligned}$$

for  $k = 1, \dots, m$ . Then,

$$\frac{\partial \mathcal{L}}{\partial w_k} = 2 \sum_{i=1}^n \left( y_i - \sum_{j=1}^m w_j \hat{f}_j(\mathbf{x}_i) \right) (-\hat{f}_k(\mathbf{x}_i)) + 2\lambda w_k + \alpha - \gamma_k = 0, \quad (5)$$

for  $k = 1, \dots, m$ . Multiplying  $w_k$  to (5), then summing over  $k$  leads us to

$$-2 \sum_{k=1}^m \left( \sum_{i=1}^n y_i \hat{f}_k(\mathbf{x}_i) \right) w_k + 2 \sum_{i=1}^n \left( \sum_{k=1}^m w_k \hat{f}_k(\mathbf{x}_i) \right)^2 + 2\lambda \sum_{k=1}^m w_k^2 + \alpha \sum_{k=1}^m w_k = 0, \quad (6)$$

since  $\gamma_k w_k = 0$ . The matrix notation of (6) is expressed as

$$-2\mathbf{y}^T \mathbf{X}_1 \mathbf{w} + 2\mathbf{w}^T \mathbf{X}_1^T \mathbf{X}_1 \mathbf{w} + 2\lambda \mathbf{w}^T \mathbf{w} + \alpha \mathbf{1}^T \mathbf{w} = 0. \quad (7)$$

where

$$\mathbf{X}_1 = \begin{pmatrix} \hat{f}_1(\mathbf{x}_1) & \cdots & \hat{f}_m(\mathbf{x}_1) \\ \hat{f}_1(\mathbf{x}_2) & \cdots & \hat{f}_m(\mathbf{x}_2) \\ \vdots & \ddots & \vdots \\ \hat{f}_1(\mathbf{x}_n) & \cdots & \hat{f}_m(\mathbf{x}_n) \end{pmatrix}.$$

The solution to (7) is given by

$$\hat{\mathbf{w}} = (\mathbf{X}_1^T \mathbf{X}_1 + \lambda \mathbf{I})^{-1} \left( \mathbf{X}_1^T \mathbf{y} - \frac{\alpha}{2} \mathbf{1} \right),$$

and therefore,

$$\hat{\mathbf{w}}^T \hat{\mathbf{w}} = \left( \mathbf{y}^T \mathbf{X}_1 - \frac{\alpha}{2} \mathbf{1}^T \right) (\mathbf{X}_1^T \mathbf{X}_1 + \lambda \mathbf{I})^{-2} \left( \mathbf{X}_1^T \mathbf{y} - \frac{\alpha}{2} \mathbf{1} \right),$$

which is a decreasing function of  $\lambda$ .  $\square$

### 3 Moving hyperplane data

We report the results for support vector machines (SVM) and linear discriminant analysis (LDA).

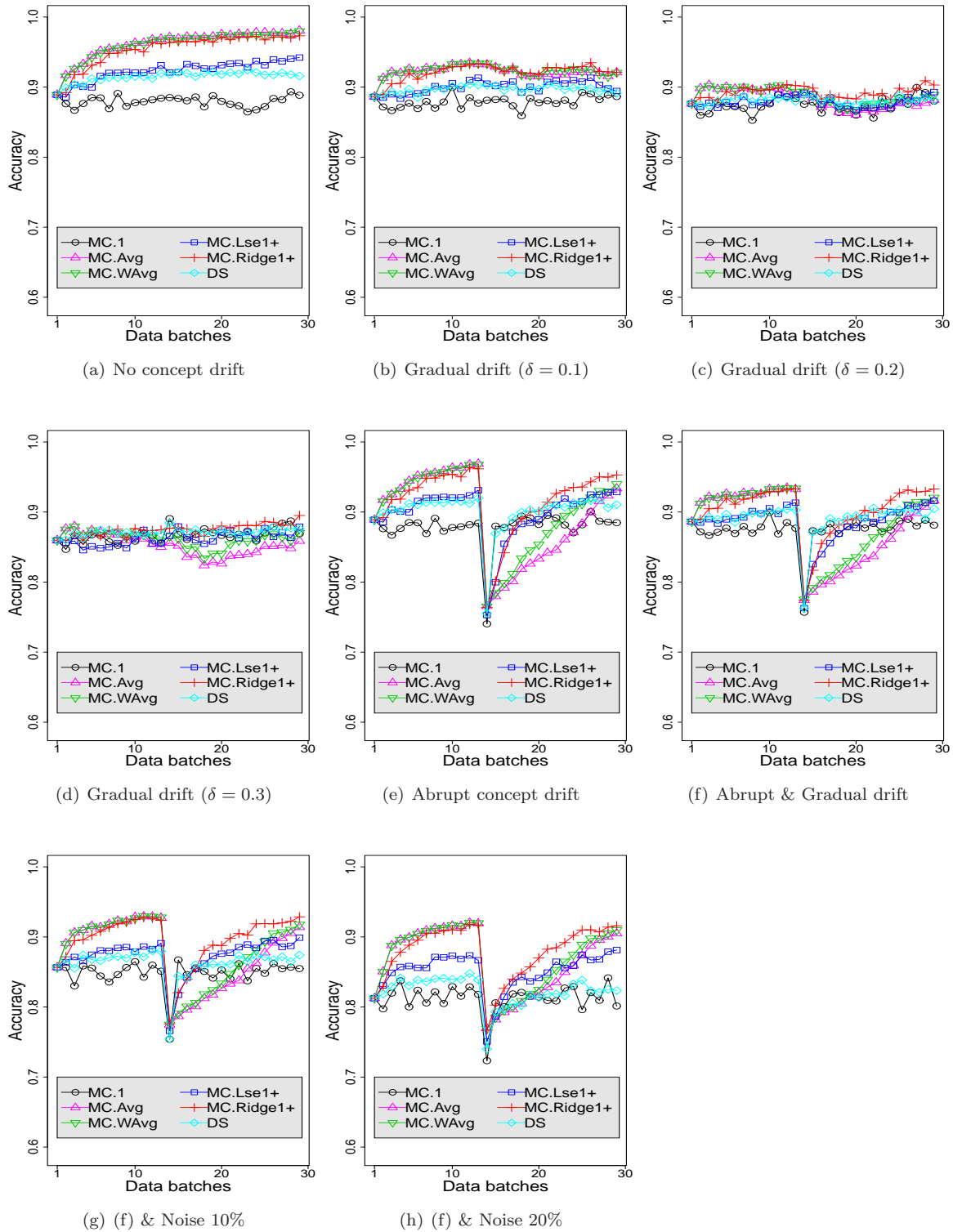


Figure 1: Concept drift tracking in Moving-hyperplane data: SVM

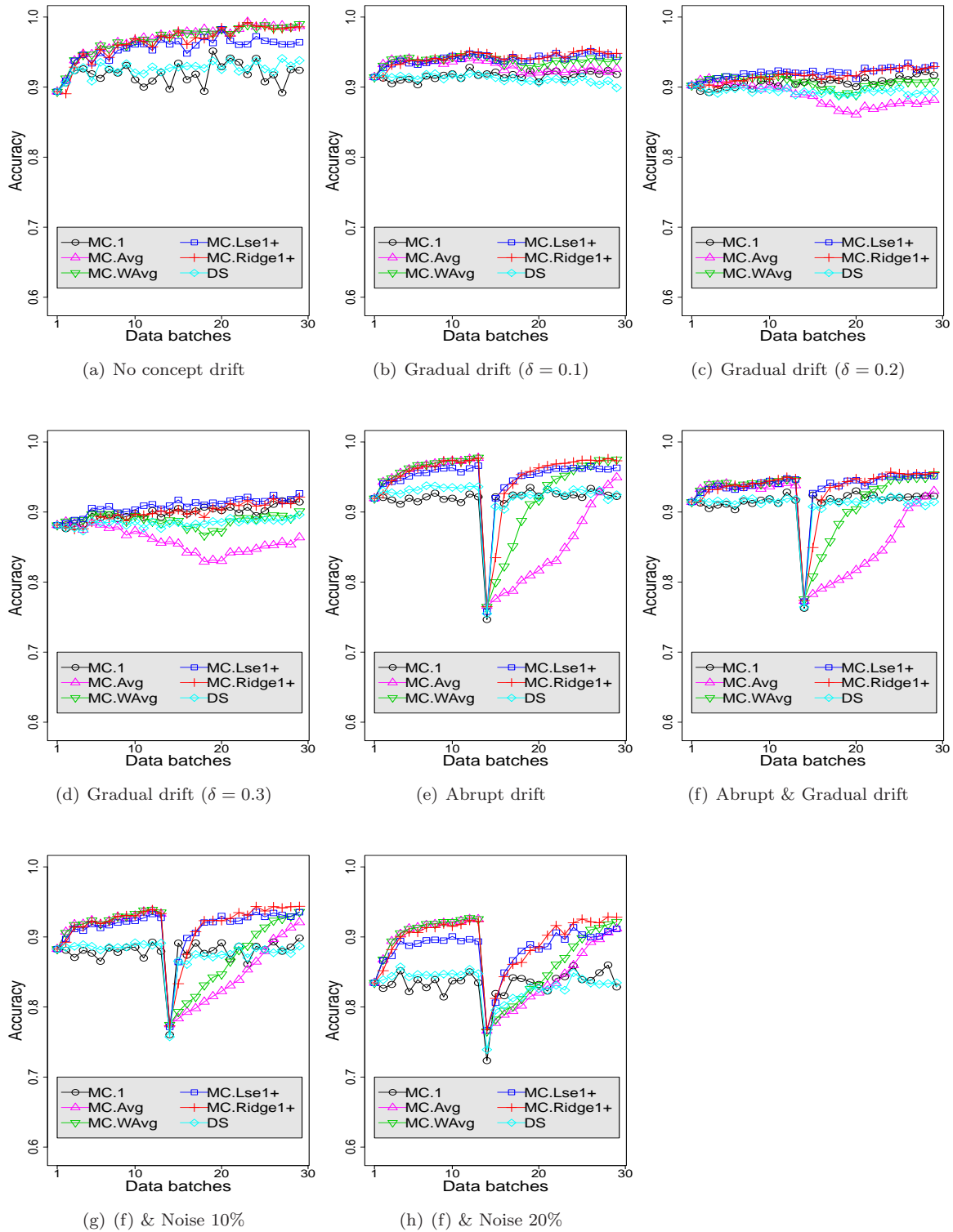


Figure 2: Concept drift tracking in Moving-hyperplane data: LDA