



I N S E A





2

8

Cross-validation

Machine learning classic: zero-to-hero

Idée: Effectuer plusieurs déjeuners et dîners à l'extérieur de test

2. Choisir une liste de valeurs de **C**, par ex: [0.01, 0.05, 0.1, 1, 10]

Pour chaque C :

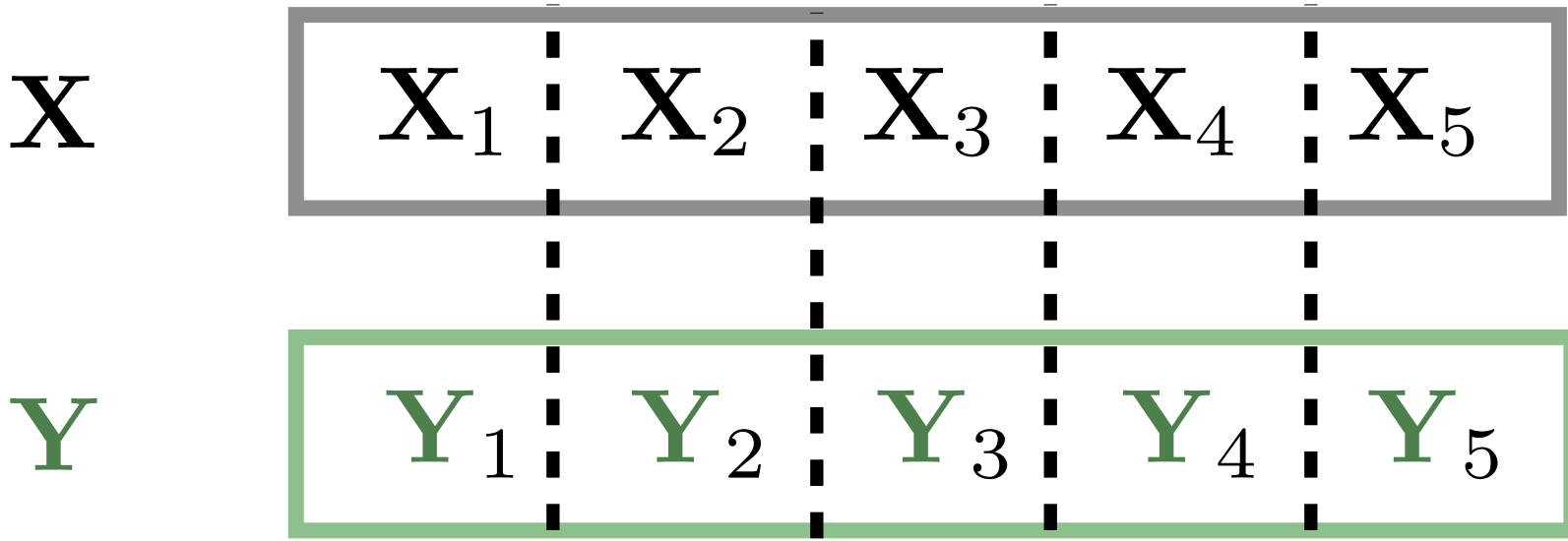
1. Optimiser sur $\mathbf{X}_{\text{train}} \mathbf{Y}_{\text{train}}$
$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \text{loss}(f_{\theta}(\mathbf{x}_i), y_i) + \frac{1}{C} \text{pénalité}(\theta)$$
2. Évaluer l'erreur de prédiction sur $\mathbf{X}_{\text{test}} \mathbf{Y}_{\text{test}}$

4. Pour chaque C , calculer l'erreur de prédiction moyenne

5. Choisir le **C** avec l'erreur de prédiction moyenne la plus petite

3. Pour chaque k in $[1, 2, 3, 4, 5]$, créer un d'écoupage train/test

1. Couper le dataset en 5 parties (folds)



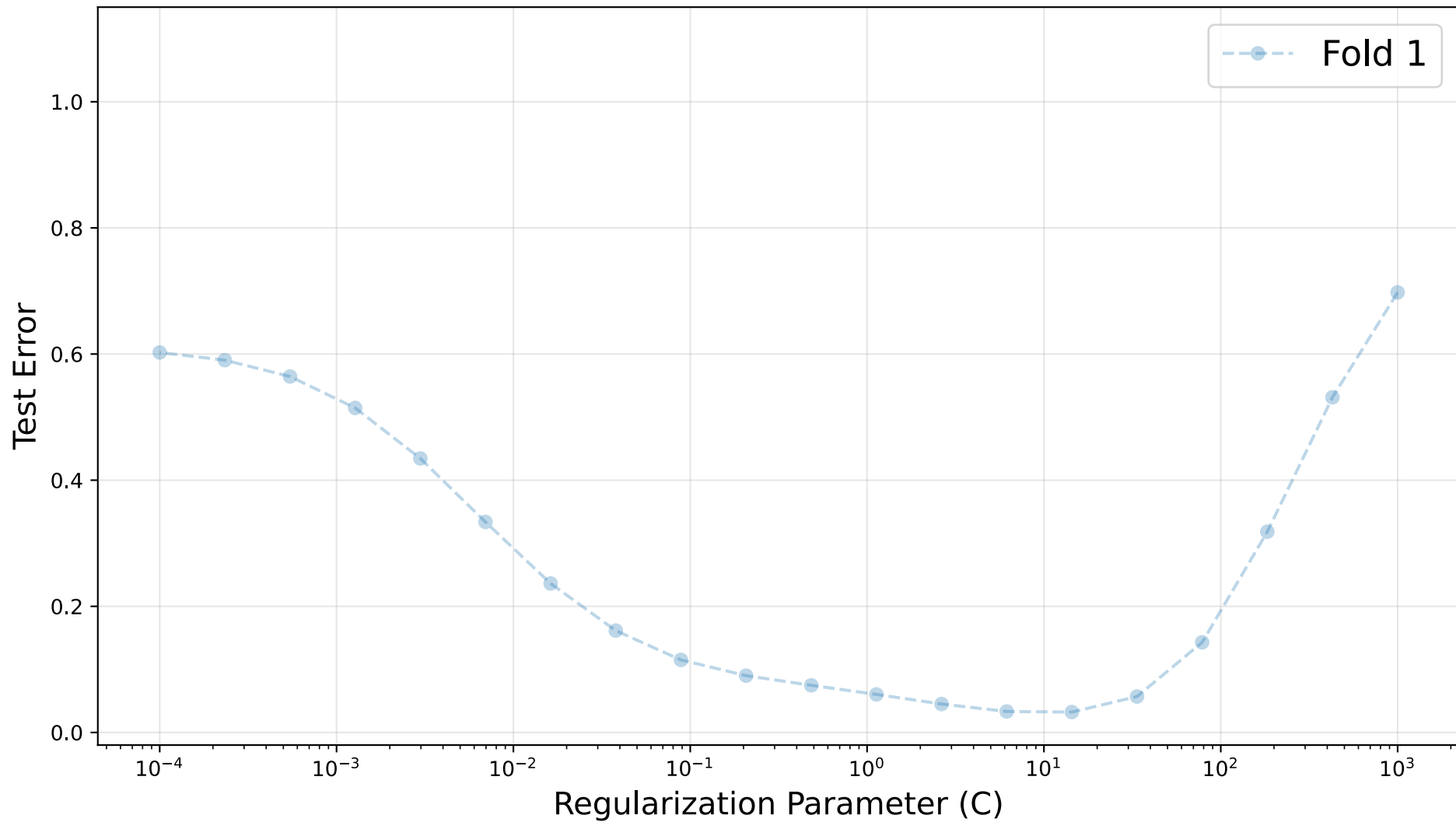
The first part of the paper discusses the importance of understanding the cultural context of the research. It highlights the need for researchers to be sensitive to the values and beliefs of the communities they are studying. This is particularly important in the field of education, where cultural differences can significantly impact learning outcomes. The paper then moves on to discuss the challenges of conducting research in culturally diverse settings. It notes that researchers often face difficulties in establishing rapport with participants and in interpreting their responses. To address these challenges, the paper suggests several strategies, including the use of local informants and the development of culturally appropriate research instruments. The final part of the paper discusses the importance of ethical considerations in cross-cultural research. It emphasizes the need for researchers to obtain informed consent from participants and to ensure that their research does not cause harm to the communities they are studying.

C'est l'erreur de validation croisée

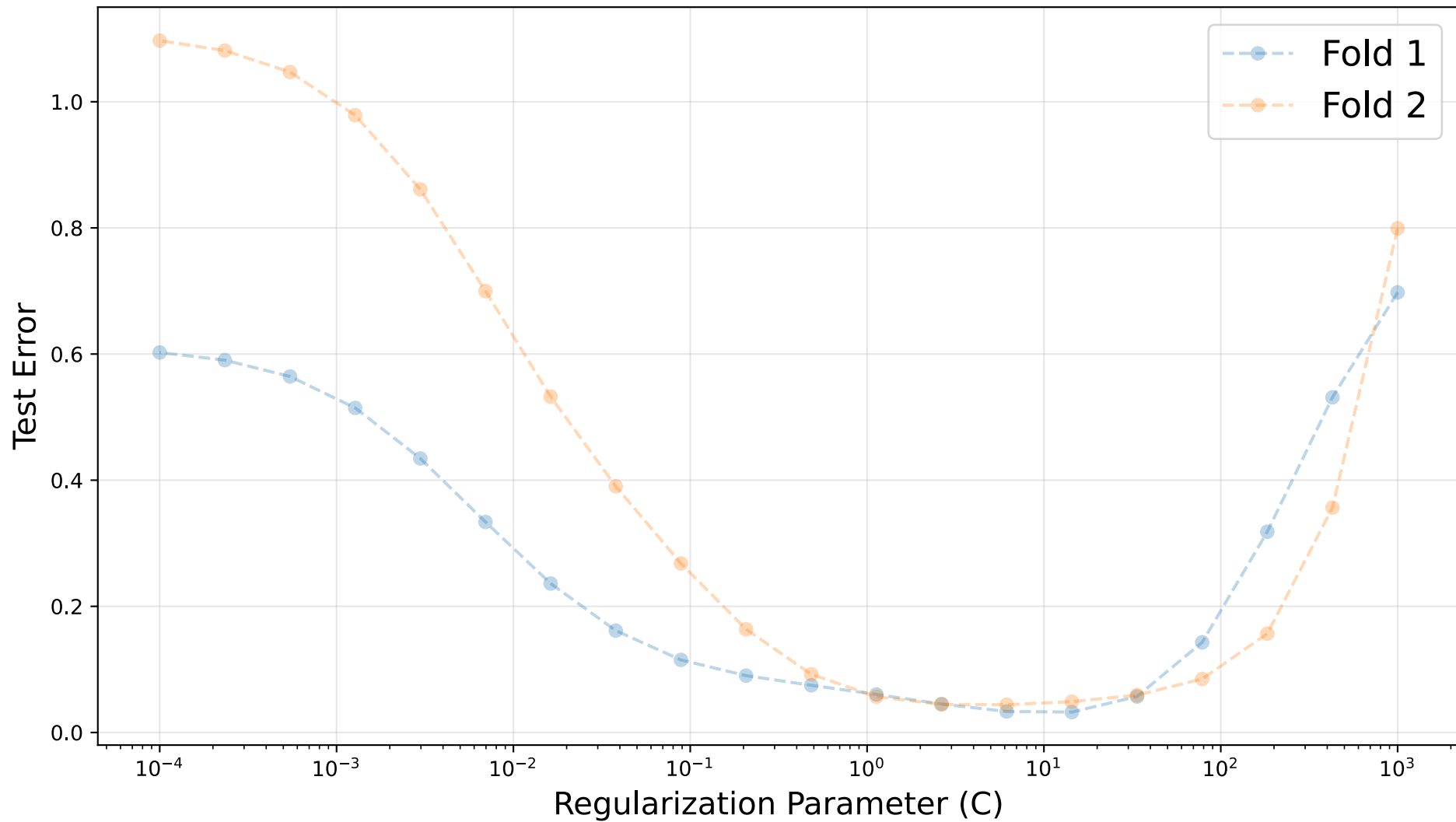
5-Fold cross validation

$$\mathbf{X}_{\text{test}}, \mathbf{Y}_{\text{test}} = \mathbf{X}_k, \mathbf{Y}_k \quad \mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}} = [\mathbf{X}_{\text{sans}} \mathbf{X}_k], \dots, [\mathbf{Y}_{\text{sans}} \mathbf{Y}_k]$$

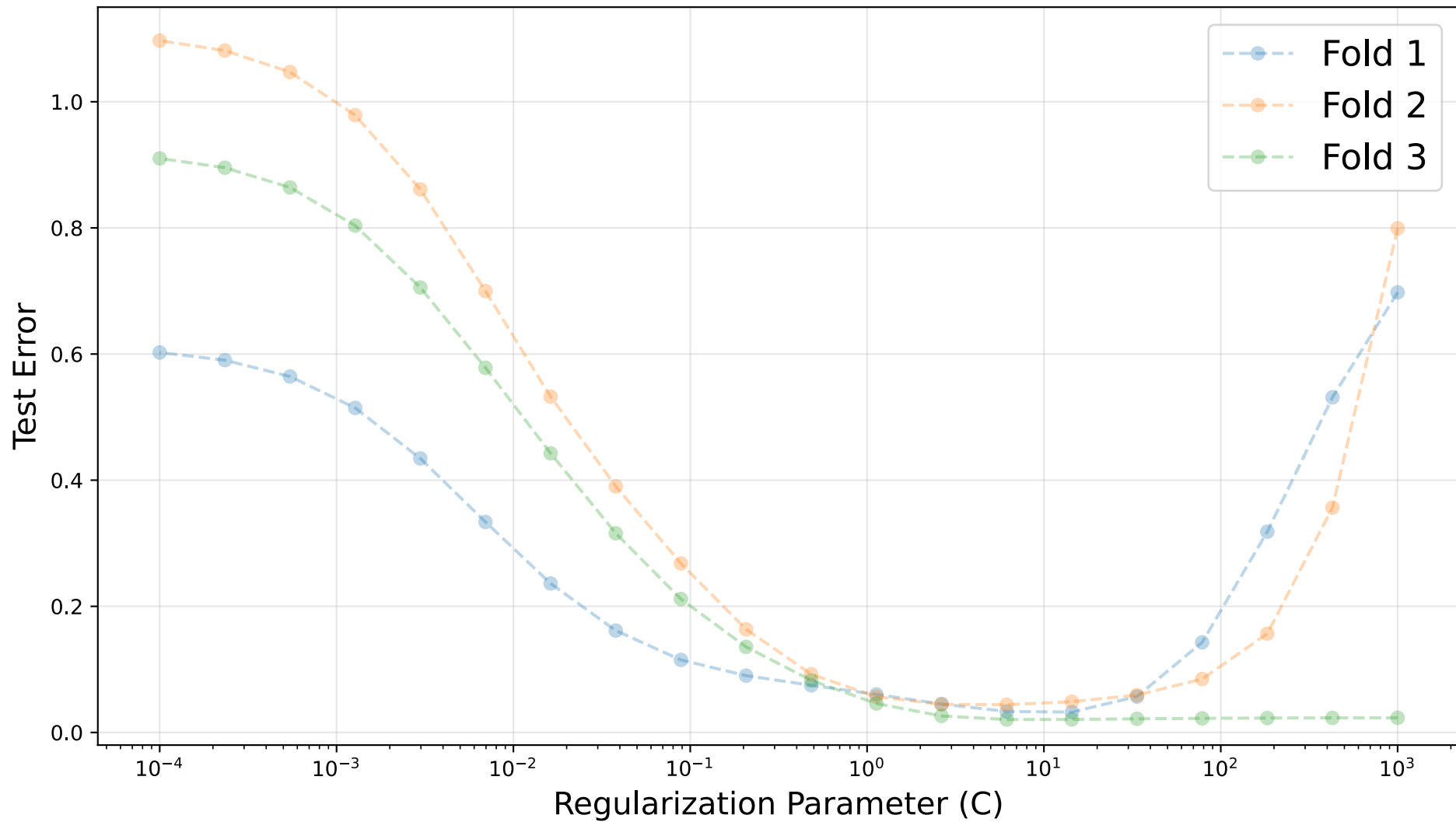
5-Fold Cross-Validation Error



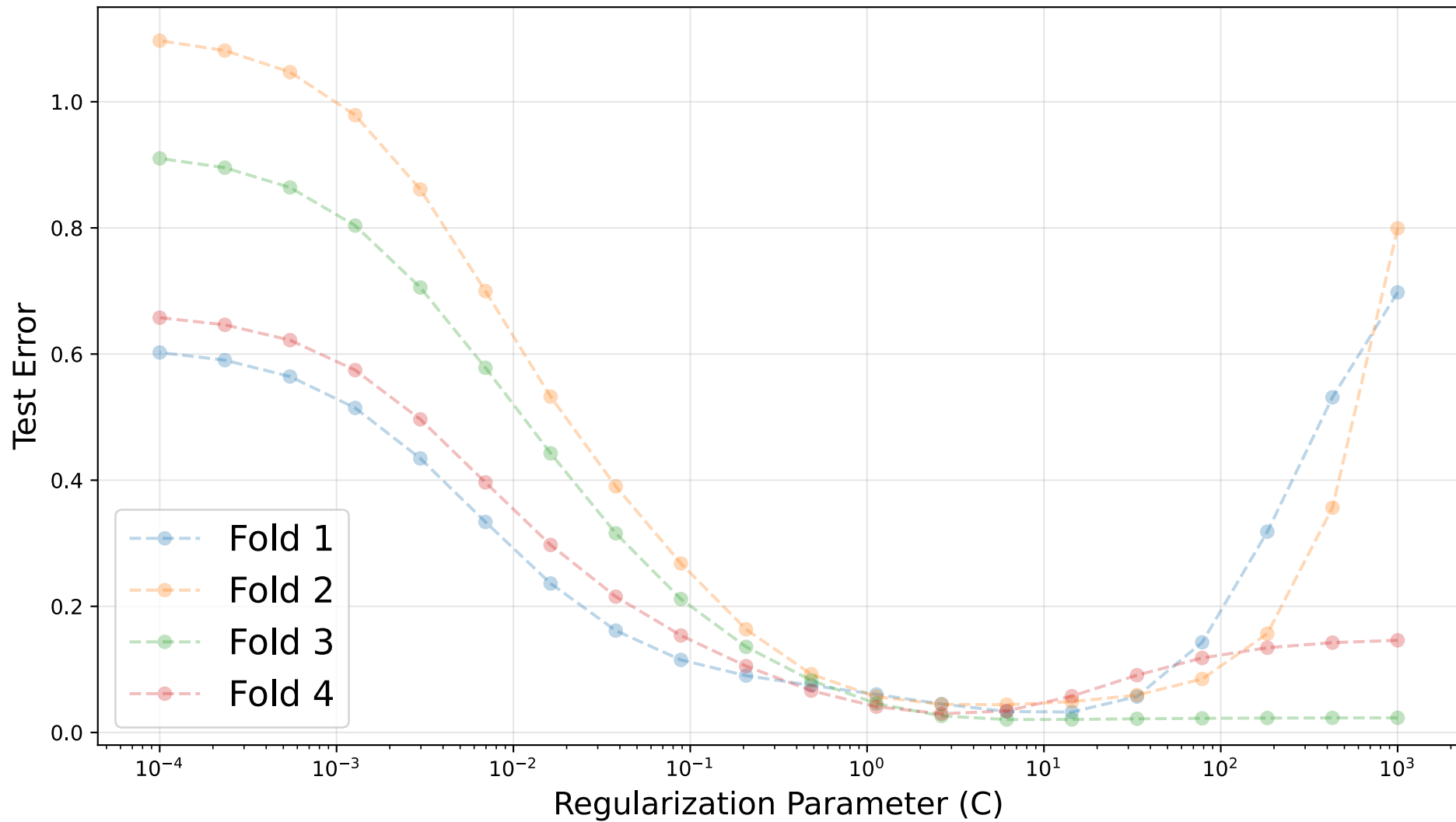
5-Fold Cross-Validation Error



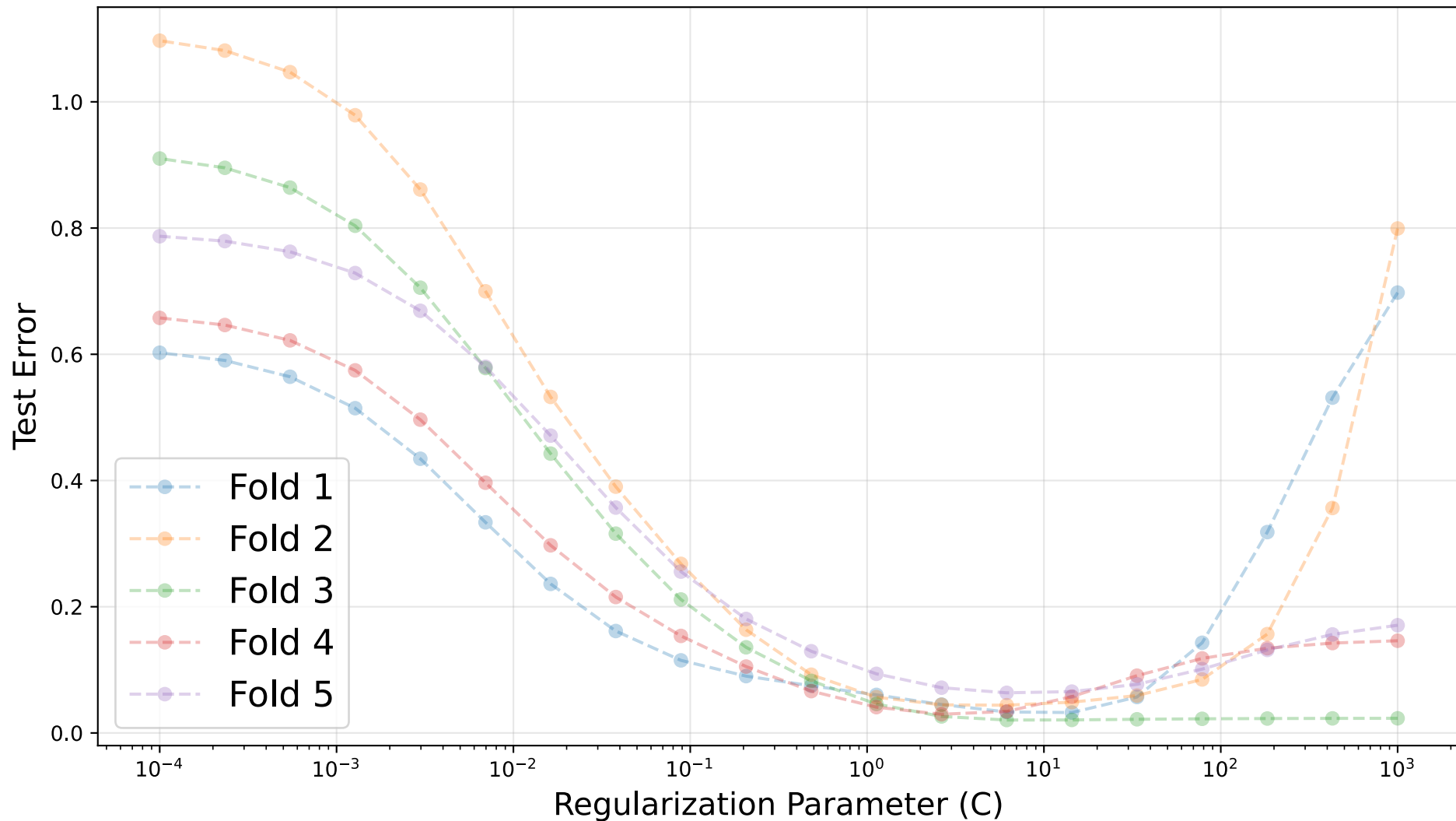
5-Fold Cross-Validation Error



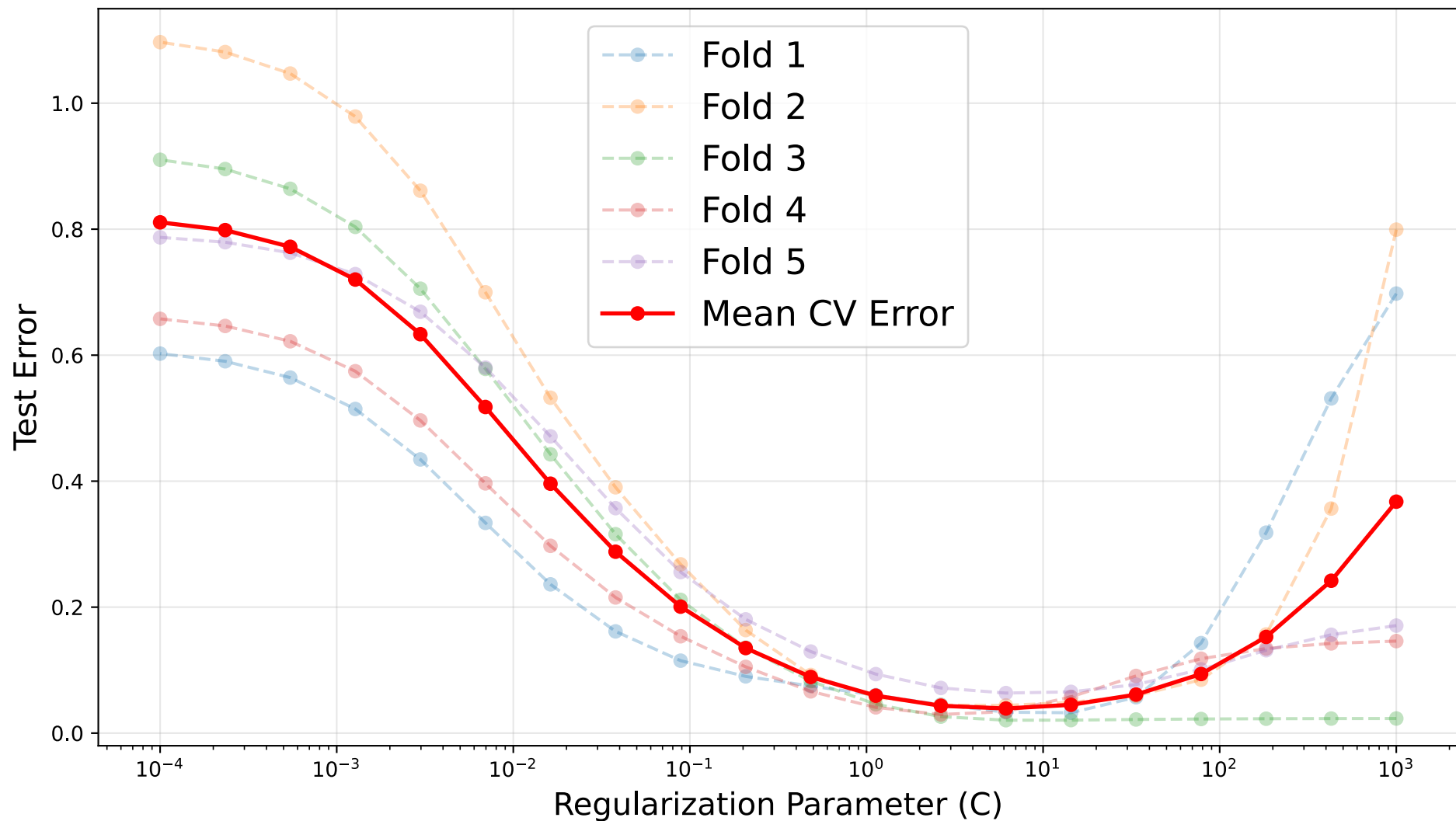
5-Fold Cross-Validation Error



5-Fold Cross-Validation Error



5-Fold Cross-Validation Error

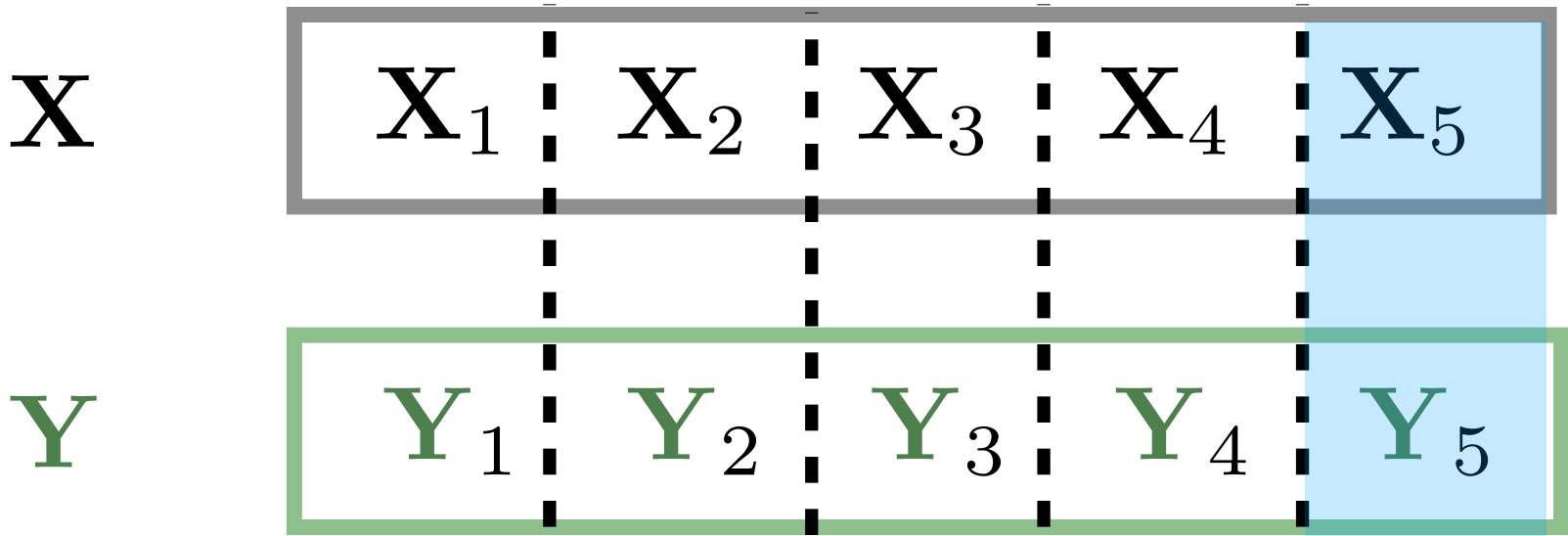




C optimal

Idée: Effectuer plusieurs découpages et moyenner l'erreur de test

1. Couper le dataset en 5 parties (folds)



2. Choisir une liste de valeurs de **C**, par ex: [0.01, 0.05, 0.1, 1., 10]

3. Pour chaque **k** in [1, 2, 3, 4, 5], créer un découpage train/test

$\mathbf{X}_{\text{test}}, \mathbf{Y}_{\text{test}} = \mathbf{X}_k, \mathbf{Y}_k$

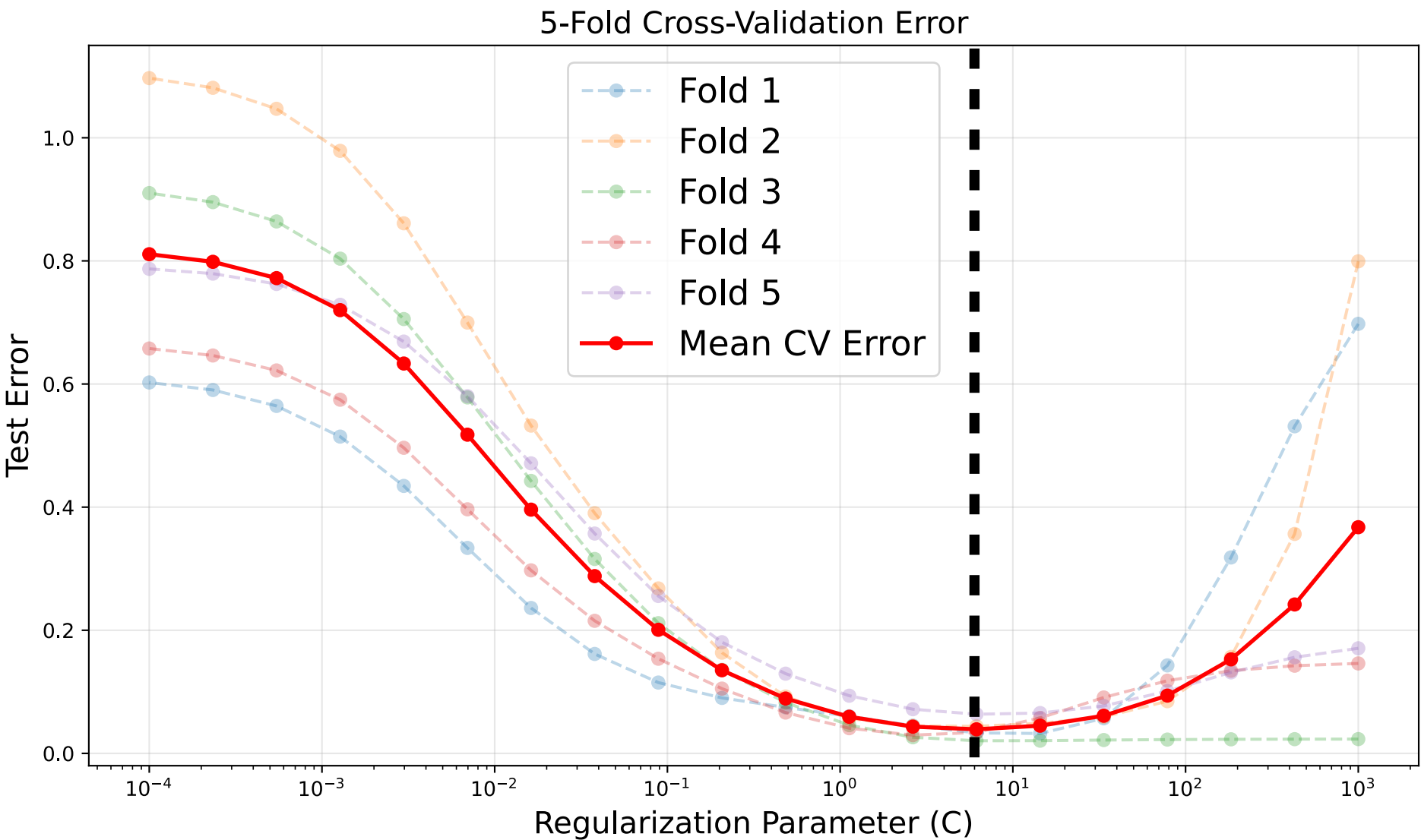
$\mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}} = [\mathbf{X} \text{ sans } \mathbf{X}_k], \dots, [\mathbf{Y} \text{ sans } \mathbf{Y}_k]$

Pour chaque **C**:

1. Optimiser sur $\mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}}$ $\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \text{loss}(f_{\theta}(\mathbf{x}_i), y_i) + \frac{1}{C} \text{pénalité}(\theta)$
2. Évaluer l'erreur de prédiction sur $\mathbf{X}_{\text{test}}, \mathbf{Y}_{\text{test}}$

4. Pour chaque **C**, calculer l'erreur de prédiction moyenne

5. Choisir le **C** avec l'erreur de prédiction moyenne la plus petite



C optimal

C'est l'erreur de validation croisée

5-Fold cross validation



Et si les données ressemblent à ceci ?

