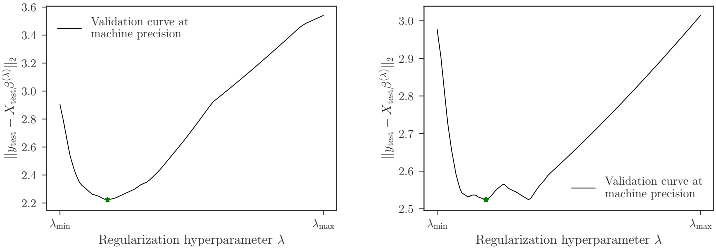
Safe Grid Search with Optimal Complexity

E. Ndiaye
Riken AIP

Joint work with: T. Le, O. Fercoq, J. Salmon, I. Takeuchi

Hyperparameter Tuning

- Learning Task: $\hat{\beta}^{(\boldsymbol{\lambda})} \in \arg\min_{\beta \in \mathbb{R}^p} f(X_{\text{train}}\beta) + \boldsymbol{\lambda}\Omega(\beta)$
- Evaluation: $E_v(\hat{\beta}^{(\boldsymbol{\lambda})}) = \mathcal{L}(y_{\text{test}},\ X_{\text{test}}\hat{\beta}^{(\boldsymbol{\lambda})})$



How to approximate the best hyperparameter?

Hyperparameter Tuning

The optimal hyperparameter is given by

$$\underset{\lambda \in [\lambda_{\min}, \lambda_{\max}]}{\arg\min} \; E_v(\hat{\beta}^{(\lambda)}) = \mathcal{L}(y_{\text{test}}, X_{\text{test}} \hat{\beta}^{(\lambda)})$$
$$\text{s.t. } \hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \, f(X_{\text{train}} \beta) + \lambda \Omega(\beta)$$

Issues:

- The objective $\lambda \mapsto E_v(\hat{\beta}^{(\lambda)})$ is **non-smooth** and **non-convex**
- Often, It is **unpractical** to evaluate $E_v(\hat{\boldsymbol{\beta}}^{(\boldsymbol{\lambda})})$

Tracking the curve of solutions

$$\hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min}\, f(X\beta) + \lambda\Omega(\beta)$$

Exact Path: For $(f, \Omega) =$ (Piecewise Quadratic, Piecewise Linear) the function $\lambda \longmapsto \hat{\beta}^{(\lambda)}$ is piecewise linear (`Lars`[1] algorithm).

[1](Efron *et al.* , 2004)
[2](Mairal and Yu, 2012)
[3](Bousquet and Bottou, 2008)

Tracking the curve of solutions

$$\hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} f(X\beta) + \lambda \Omega(\beta)$$

Exact Path: For $(f, \Omega) =$ (Piecewise Quadratic, Piecewise Linear) the function $\lambda \longmapsto \hat{\beta}^{(\lambda)}$ is piecewise linear (`Lars`[1] algorithm).

Drawbacks:

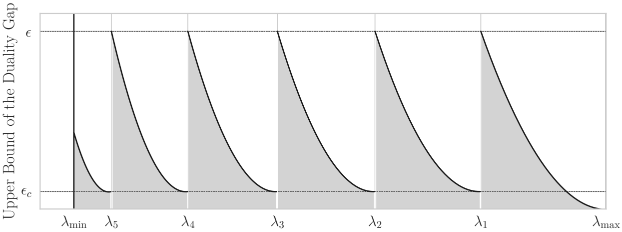
- Exponential [2] complexity for Lasso $O((3^p + 1)/2)$
- Numerical instabilities
- Hard to generalize to others (loss, regularization)
- Cannot benefited of early stopping rule [3].

[1] (Efron *et al.* , 2004)
[2] (Mairal and Yu, 2012)
[3] (Bousquet and Bottou, 2008)

Approximation of the solution path [4]

Training Task: $$\hat{\beta}^{(\lambda)} \in \arg\min_{\beta \in \mathbb{R}^p} f(X\beta) + \lambda\Omega(\beta) =: P_\lambda(\beta)$$

Suboptimal gap: $$P_\lambda(\beta^{(\lambda_t)}) - P_\lambda(\hat{\beta}^{(\lambda)}) \leq Q_{t,\mathcal{V}_{f^*}}\left(1 - \frac{\lambda}{\lambda_t}\right) \quad .$$



$Q_{t,\mathcal{V}_{f^*}}(\rho) :=$ optimization error at $\lambda_t +$ approximation error(λ, λ_t) ,

[4](Giesen et al. 2012)

Bound the validation Gap

$$\left|E_v(\hat{\beta}^{(\lambda)}) - E_v(\beta^{(\lambda_t)})\right| \leq \max_{\beta \in \mathcal{B}_\lambda} \mathcal{L}(X'\beta, X'\beta^{(\lambda_t)}) \ ,$$

$$\mathcal{B}_\lambda = \mathrm{Ball}\left(\beta^{(\lambda_t)}, \textbf{Suboptimal gap on the training}\ \right) \ni \hat{\beta}^{(\lambda)}$$

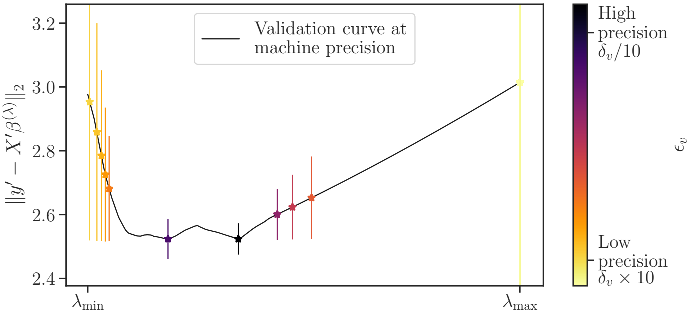
- $\longrightarrow$ Approximate the validation path !

Bound the validation Gap

$$\left|E_v(\hat\beta^{(\lambda)}) - E_v(\beta^{(\lambda_t)})\right| \leq \max_{\beta \in \mathcal{B}_\lambda} \mathcal{L}(X'\beta, X'\beta^{(\lambda_t)}) \enspace ,$$

$$\mathcal{B}_\lambda = \mathrm{Ball}\left(\beta^{(\lambda_t)}, \textbf{Suboptimal gap on the training}\right) \ni \hat\beta^{(\lambda)}$$

- $\longrightarrow$ Approximate the validation path !



$$\min_{\lambda_t \in \Lambda_{\mathrm{val}(\epsilon_v)}} E_v(\beta^{(\lambda_t)}) - \min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} E_v(\hat{\beta}^{(\lambda)}) \leq \epsilon_v \ .$$

Code: `https://github.com/EugeneNdiaye/safe_grid_search`

Let's talk during the poster session ;-)