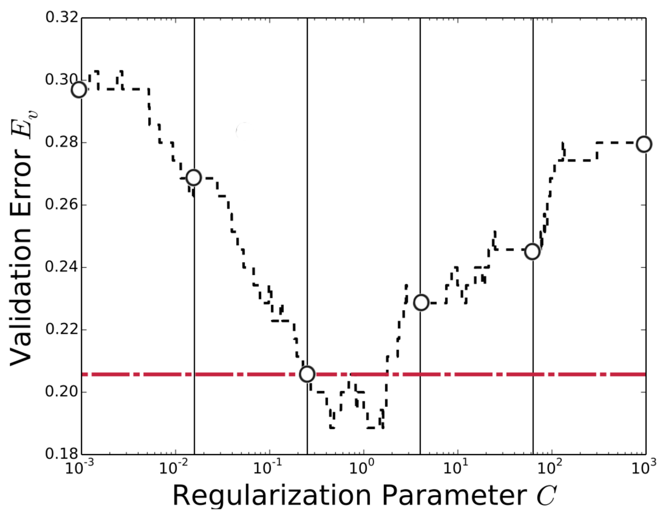
Approximately optimal selection of regularization parameter

for L2 regularized convex loss minimization problems for supervised classifications

Atsushi Shibagaki, Yoshiki Suzuki, Masayuki Karasuyama, Ichiro Takeuchi (Nagoya Institute of Technology)

Abstract and problem formulation

Background : Grid search for model selection



- Can you find out about the difference in quality between exact optimal and a selected parameter ?
- If you use our algorithm, you can find out it in the sense that the validation error

Target problems: L2 regularized loss minimization problems (e.g. SVM)

- Training instances and labels : $\{(x_i, y_i) \in \mathbb{R}^d \times \{-1,1\}\}_{i\in[n]}$
- Validation instances and labels : $\{(x'_i, y'_i) \in \mathbb{R}^d \times \{-1,1\}\}_{i\in[n']}$

$$w^*_C := \arg\min_{w\in\mathbb{R}^d} \frac{1}{2}\|w\|^2 + C\sum_{i\in[n]} \ell(y_i, w^\top x_i) \quad (1)$$

Goal : Finding a theoretical approximation guarantee in the sense that the validation error for the regularization parameter is at most greater by $\varepsilon \in [0,1]$ than the smallest possible the validation error (2)

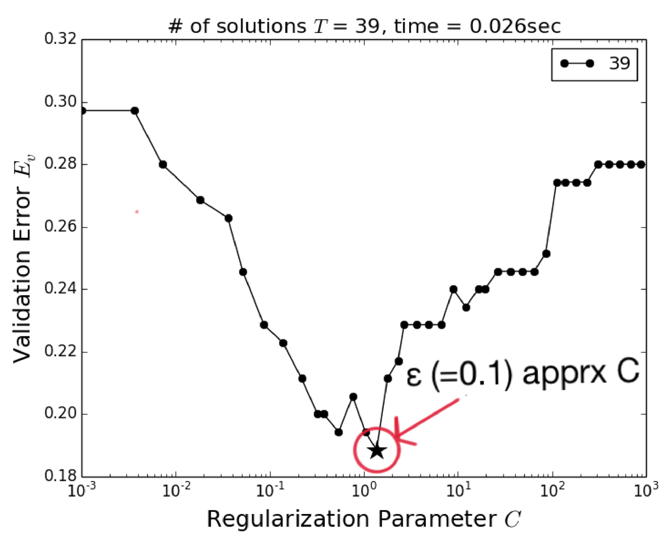
- Validation error : $E_v(w) := \frac{1}{n'}\sum_{i\in[n']} I(y'_i w^\top x'_i < 0)$
- ε-approximate regularization parameters ($\varepsilon \in [0,1]$) :

$$C(\varepsilon) := \left\{C \in [C_l, C_u] \,\middle|\, E_v(w^*_C) - (\text{the lowest } E_v \text{ in } [C_l, C_u]) \le \varepsilon\right\} \quad (2)$$

Contribution : The algorithm for finding an ε-approximate regularization parameter (input: ε, output: an ε-approximate regularization parameter)

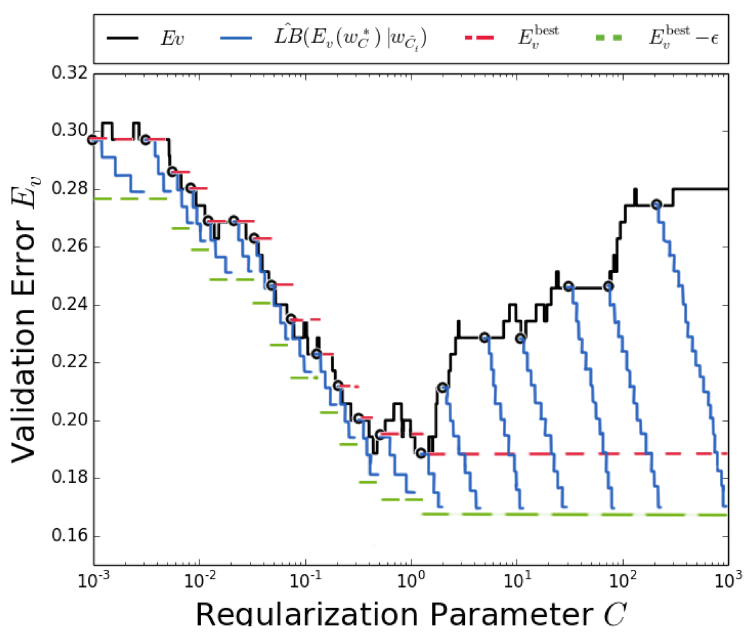
Approach : Computing the validation error lower bound as a function of the regularization parameter in the entire interval

Example : An illustration of the proposed algorithm (dataset : `ionosphere`)



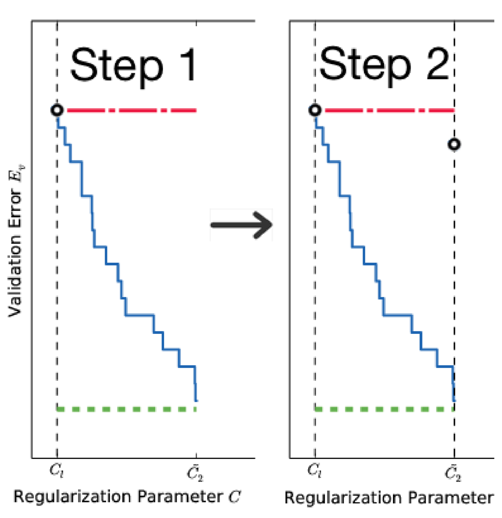
- The algorithm automatically selected 39 regularization parameter values in $[10^{-3}, 10^3]$
- And an upper bound of the validation error for each of the 39 regularization parameter values is obtained by solving an optimization problem (1)
- Among those 39 values, the one with the smallest validation error upper bound (indicated as ★ at $C = 1.368$) is guaranteed to be $\varepsilon(=0.1)$ *approximate regularization parameter* (2)

Details of approach : the behavior of the algorithm and the illustration



Our algorithm

- is built on novel technique for computing validation error lower bounds
- needs some solutions of (1) for computing validation error lower bound in the entire interval



Step0 Compute a solution $\hat{w}_{C_l}$ at the interval of the left end C_l

Step1 Compute validation error lower bound by using a solution we obtained at a previous step, and update the current best validation error upper bound

Step2 Find $\tilde{C}_{t+1}$ so that the best regularization parameter obtained is an ε-approximate regularization parameter in the interval $[C_l, \tilde{C}_{t+1}]$ and compute $\hat{w}_{\tilde{C}_{t+1}}$ at $C = \tilde{C}_{t+1}$

Step3 Continue Step1,2 until reach the interval of the right end C_u

Extensions : Cross-validation setup

- Proposed algorithm can be straightforwardly adapted to a cross-validation (CV) setup.

Theory : Validation error lower bounds $LB(E_v(w^*_C))$

$$LB(E_v(w^*_C)) = \frac{\text{\# validation instances that can be guaranteed to be mis-classified by using } w^*_C}{\text{\# validation instances}}$$

1. Compute lower and upper bound of inner product $w^{*\top}_C x'_i$
   - Construct existing ranges of optimal solutions (hypersphere) by using $\hat{w}_{\tilde{C}}$

1.1 Optimal condition : $(\underbrace{w^*_C + C\sum \xi_i(w^*_C)}_{\text{a subgradient of (1)}: g(w^*_C)})^\top (w^*_C - \hat{w}_{\tilde{C}}) \le 0$

$(\xi_i(w^*_C)$ is a subgradient of the loss function $\ell_i(w) := \ell(y_i, w^\top x_i))$

1.2 Definition of subgradient : $\ell_i(w^*_C) \ge \ell_i(\hat{w}_{\tilde{C}}) + \xi_i(\hat{w}_{\tilde{C}})^\top (w^*_C - \hat{w}_{\tilde{C}})$
$\ell_i(\hat{w}_{\tilde{C}}) \ge \ell_i(w^*_C) + \xi_i(w^*_C)^\top (\hat{w}_{\tilde{C}} - w^*_C)$

$$\Rightarrow \Big\|w^*_C - \underbrace{\frac{1}{2}\Big(\hat{w}_{\tilde{C}} - \frac{C}{\tilde{C}}(g(\hat{w}_{\tilde{C}}) - \hat{w}_{\tilde{C}})\Big)}_{\text{center}}\Big\|^2 \le \Big(\underbrace{\frac{1}{2}\Big\|\hat{w}_{\tilde{C}} + \frac{C}{\tilde{C}}(g(\hat{w}_{\tilde{C}}) - \hat{w}_{\tilde{C}})\Big\|}_{\text{radius}}\Big)^2 \quad (3)$$

- Solve following optimization problems :

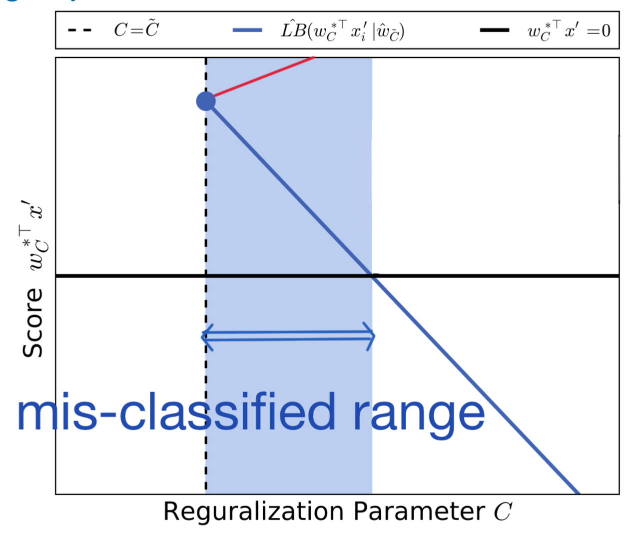
Lower bound : $w^{*\top}_C x'_i \ge \hat{LB}(w^{*\top}_C x'_i \mid \hat{w}_{\tilde{C}}) := \min_w w^\top x'_i$ s.t. (3)

Upper bound : $w^{*\top}_C x'_i \le \hat{UB}(w^{*\top}_C x'_i \mid \hat{w}_{\tilde{C}}) := \max_w w^\top x'_i$ s.t. (3)

   - have explicit solutions
   - are change linearly with a regularized parameter

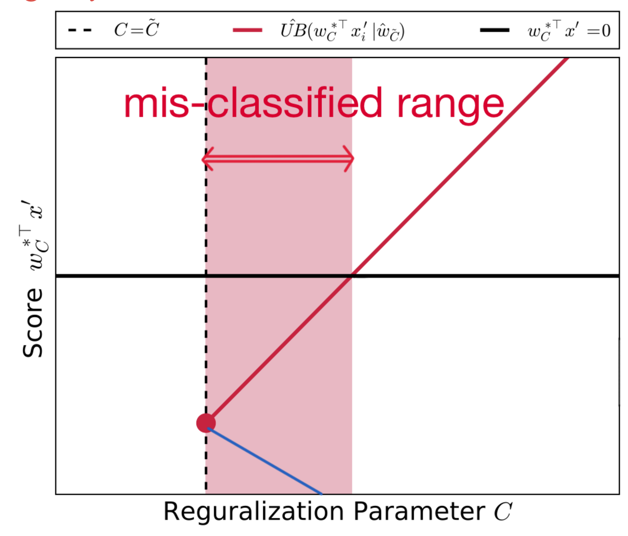
2. Identify the interval of C within which the validation instance is guaranteed to be mis-classified
   - similar to following two images (according to the sign of labels y'_i)

$\hat{LB}(w^{*\top}_C x'_i|\hat{w}_{\tilde{C}})$:monotonic decrease ($C > \tilde{C}$)



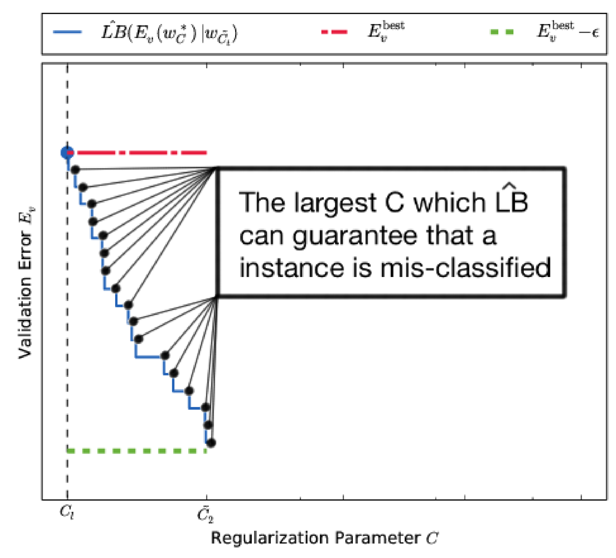
mis-classified case 1 ($y'_i = -1$) :
$0 < LB(w^{*\top}_C x'_i|\hat{w}_{\tilde{C}}) \Rightarrow$
the interval between $\tilde{C}$ and C that $\hat{LB}(w^{*\top}_C x'_i|\hat{w}_{\tilde{C}})$ becomes 0

$\hat{UB}(w^{*\top}_C x'_i|\hat{w}_{\tilde{C}})$:monotonic increace ($C > \tilde{C}$)



mis-classified case2 ($y'_i = +1$) :
$UB(w^{*\top}_C x'_i|\hat{w}_{\tilde{C}}) < 0 \Rightarrow$
the interval between $\tilde{C}$ and C that $\hat{UB}(w^{*\top}_C x'_i|\hat{w}_{\tilde{C}})$ becomes 0

3. Compute validation error lower bound $\hat{LB}(E_v(w^*_C)|\hat{w}_{\tilde{C}})$

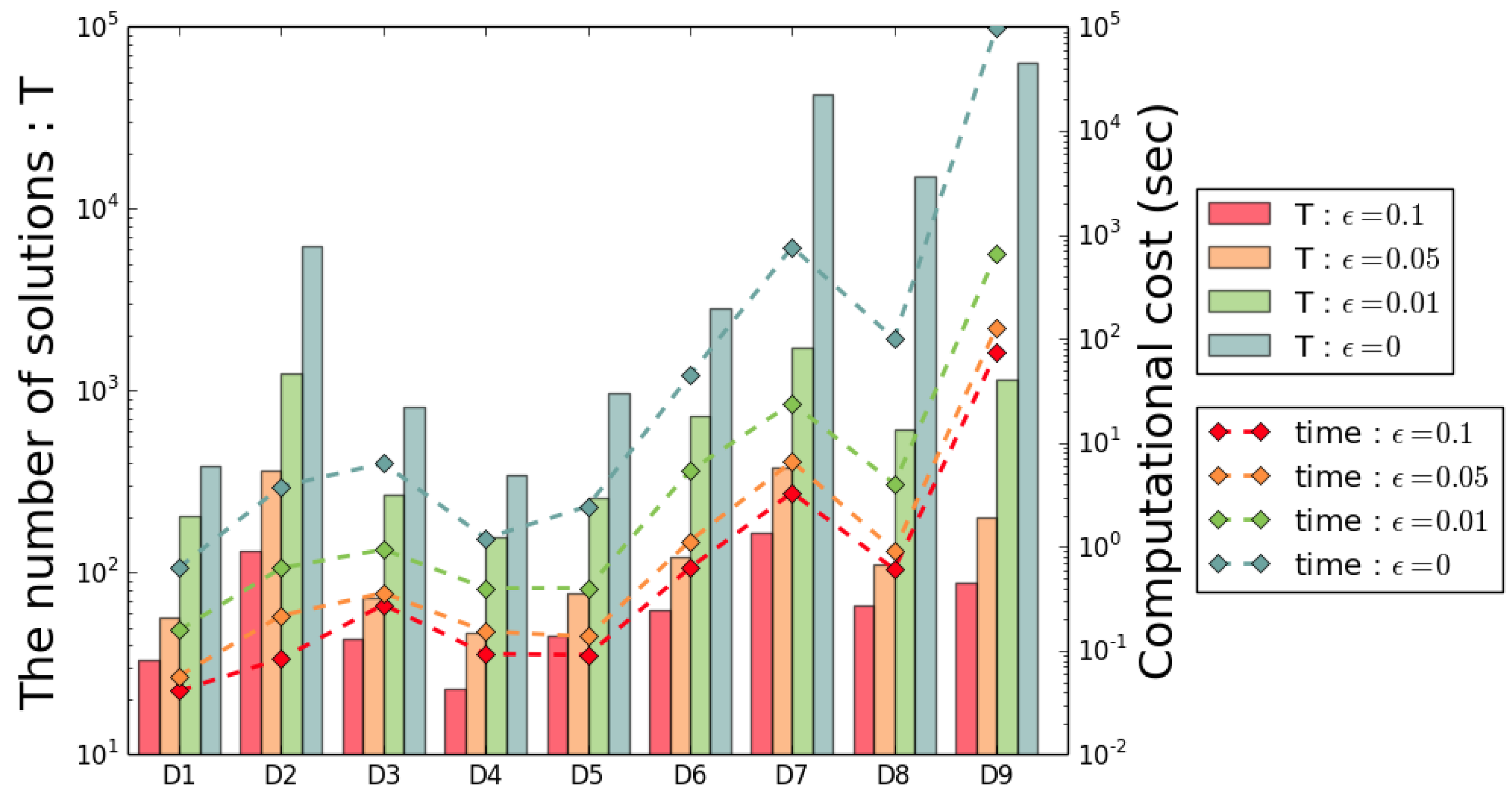


- When $\hat{LB}(w^{*\top}_C x'_i|\hat{w}_{\tilde{C}})$ cannot guarantee that a validation instance x'_i is mis-classified, the validation error lower bound decrease $1/n'$
- Therefore, the validation error lower bound is staircase function

Experiments : finding an ε-approximate regularization parameter

- Under 10-fold cross validation setup
- The entire interval of regularization parameter : $[10^{-3}, 10^3]$
- Loss function $\ell_i(w)$: smooth hinge-loss
- Input : $\varepsilon = \{0.1,\ 0.05,\ 0.01,\ 0\}$

	dataset name	sample size	input dimension		dataset name	sample size	input dimension
D1	`heart`	270	13	D6	`german.numer`	1000	24
D2	`liver-disorders`	345	6	D7	`svmguide3`	1284	21
D3	`ionosphere`	351	34	D8	`svmguide1`	7089	4
D4	`australian`	690	14	D9	`a1a`	32561	123
D5	`diabetes`	768	8				



Shibagaki et al. Regularization Path of Cross-Validation Error Lower Bounds. arXiv:1502.02344.