Safe Grid Search with Optimal Complexity

Eugene Ndiaye [1] **Tam Le** [1] **Olivier Fercoq** [2] **Joseph Salmon** [3] **Ichiro Takeuchi** [4]

Abstract

Popular machine learning estimators involve regularization parameters that can be challenging to tune, and standard strategies rely on grid search for this task. In this paper, we revisit the techniques of approximating the regularization path up to predefined tolerance ϵ in a unified framework and show that its complexity is $O(1/\sqrt[d]{\epsilon})$ for uniformly convex loss of order $d \geq 2$ and $O(1/\sqrt{\epsilon})$ for Generalized Self-Concordant functions. This framework encompasses least-squares but also logistic regression, a case that as far as we know was not handled as precisely in previous works. We leverage our technique to provide refined bounds on the validation error as well as a practical algorithm for hyperparameter tuning. The latter has global convergence guarantee when targeting a prescribed accuracy on the validation set. Last but not least, our approach helps relieving the practitioner from the (often neglected) task of selecting a stopping criterion when optimizing over the training set: our method automatically calibrates this criterion based on the targeted accuracy on the validation set.

1. Introduction

Various machine learning problems are formulated as minimization of an empirical loss function f plus a regularization function Ω whose calibration is controlled by a hyperparameter λ. The choice of λ is crucial in practice since it directly influences the generalization performance of the estimator, *i.e.,* its score on unseen data sets. The most popular method in such a context is cross-validation (or some variant, see (Arlot & Celisse, 2010) for a detailed review). For simplicity, we investigate here the simplest case, the holdout version. It consists in splitting the data in two parts: on the first part (*training set*) the method is trained for a predefined collection of candidates $\Lambda_T := \{\lambda_0, \ldots, \lambda_{T-1}\}$, and on the second part (*validation set*), the best parameter is selected among the T candidates.

For a piecewise quadratic loss f and a piecewise linear regularization Ω (*e.g.,* the Lasso estimator), Osborne et al. (2000); Rosset & Zhu (2007) have shown that the set of solutions follows a piecewise linear curve *w.r.t.* to the parameter λ. There are several algorithms that can generate the full path by maintaining optimality conditions when the regularization parameter varies. This is what `LARS` is performing for Lasso (Efron et al., 2004), but similar approaches exist for SVM (Hastie et al., 2004) or generalized linear models (GLM) (Park & Hastie, 2007). Unfortunately, these methods have some drawbacks that can be critical in many situations:

- their worst case complexity, *i.e.,* the number of linear segments, is exponential in the dimension p of the problem (Gärtner et al., 2012) leading to unpractical algorithms. Recently, Li & Singer (2018) have shown that for some specific design matrix with n observations, a polynomial complexity of $O(n \times p^6)$ can be obtained. Note that even in a more favorable cases of linear complexity in p, the exact path can be expensive to compute when the dimension p is large.

- they suffer from numerical instabilities due to multiple and expensive inversion of ill-conditioned matrix. As a result, these algorithms may fail before exploring the entire path, a common issue for small regularization parameter.

- they lack flexibility when it comes at incorporating different statistical learning tasks because they usually rely on specific algebra to handle the structure of the regularization and loss functions. As far as we know, they can be applied only to a limited number of cases and we are not aware of a general framework that bypasses these issues.

- they cannot benefit of early stopping. Following Bottou & Bousquet (2008), it is not necessary to optimize below the statistical error for suitable generalization. Exact regularization path algorithms need to maintain optimality conditions as the hyperparameter varies, which is time consuming.

To overcome these issues, an ϵ-approximation of the solution path was proposed and optimal complexity was proven to be $O(1/\epsilon)$ by (Giesen et al., 2010) in a fairly general

[1] Riken AIP [2] LTCI, Télécom ParisTech, Université Paris-Saclay [3] IMAG, Univ Montpellier, CNRS, Montpellier, France [4] Nagoya Institute of Technology. Correspondence to: E. Ndiaye <eugene.ndiaye@riken.jp>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

	Lasso	Logistic regr.
$f_i(z)$	$(y_i - z)^2/2$	$\log(1 + \mathrm{e}^z) - y_i z$
$f_i^*(u)$	$((u - y_i)^2 - y_i^2)/2$	$\mathrm{Nh}(u + y_i)$
$\mathcal{V}_{f^*,x}(u)$	$\|u\|_2^2/2$	$w_4(\|u\|_x^2/\|u\|_2)\|u\|_u^2$

Table 1. $w_4(\tau) = \frac{(1-\tau)\log(1-\tau)+\tau}{\tau^2}$ and $\mathrm{Nh}(x) = x\log(x) + (1-x)\log(1-x)$

setting. Then, Mairal & Yu (2012) provided an interesting algorithm whose complexity is $O(1/\sqrt{\epsilon})$ for the Lasso case. The latter result was then extended by (Giesen et al., 2012) to objective function with quadratic lower bound while providing a lower and upper bound of order $O(1/\sqrt{\epsilon})$. Unfortunately, these assumptions fail to hold for many problems, including logistic regression or Huber loss.

Following such ideas, (Shibagaki et al., 2015) have proposed, for classification problems, to approximate the regularization path on the hold-out cross-validation error. Indeed, the latter is a more natural criterion to monitor when one aims at selecting a hyperparameter guaranteed to achieve the best validation error. The main idea is to construct upper and lower bounds of the validation error as simple functions of the regularization parameter. Hence by sequentially varying the parameters, one can estimate a range of parameter for which the validation error gap (*i.e.,* the difference with the validation error achieved by the best parameter) is smaller than an accuracy $\epsilon_v > 0$.

Contributions. We revisit the approximation of the solution and validation path in a unified framework under general regularity assumptions commonly met in machine learning. We encompass both classification and regression problems and provide a complexity analysis along with explicit optimality guarantees. We highlight the relationship between the regularity of the loss function and the complexity of the approximation path. We prove that its complexity is $O(1/\sqrt[d]{\epsilon})$ for uniformly convex loss of order $d \geq 2$ (see Bauschke & Combettes (2011, Definition 10.5)) and $O(1/\sqrt{\epsilon})$ for the logistic loss thanks to a refined measure of its curvature throughout its Generalized Self-Concordant properties (Sun & Tran-Dinh, 2017). As far as we know, the previously known approximation path algorithms cannot handle these cases. We provide an algorithm with global convergence property for selecting a hyperparameter with a validation error ϵ_v-close to the optimal hyperparameter from a given grid. We bring a natural stopping criterion when optimizing over the training set making this criterion automatically calibrated.

Our implementation is available at `https://github.com/EugeneNdiaye/safe_grid_search`.

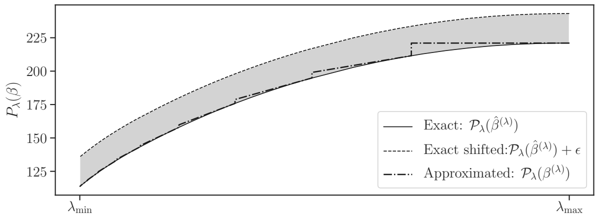


Figure 1. Illustration of the approximation path for the Lasso at accuracy $\epsilon = \|y\|_2^2/20$. We choose $\lambda_{\max} = \|X^\top y\|_\infty$ and $\lambda_{\min} = \lambda_{\max}/50$. The shaded gray region shows the interval where any ϵ-path must lie. The exact path is computed with the `LassoLars` on `diabetes` data from `sklearn`.

Notation. Given a proper, closed and convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, we denote $\mathrm{dom}\, f = \{x \in \mathbb{R}^n : f(x) < +\infty\}$. If f is a twice continuously differentiable function with positive definite Hessian $\nabla^2 f(x)$ at any $x \in \mathrm{dom}\, f$, we denote $\|z\|_x = \sqrt{\langle \nabla^2 f(x) z, z\rangle}$. The Fenchel-Legendre transform of f is the function $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined by $f^*(x^*) = \sup_{x \in \mathrm{dom}\, f} \langle x^*, x\rangle - f(x)$. The support function of a nonempty set C is defined as $\sigma_C(x) = \sup_{c \in C} \langle c, x\rangle$. If C is closed, convex and contains 0, we define its polar as $\sigma_C^\circ(x^*) = \sup_{\sigma_C(x) \leq 1} \langle x^*, x\rangle$. We denote by $[T]$ the set $\{1, \ldots, T\}$ for any non zero integer T. The vector of observations is $y \in \mathbb{R}^n$ and the design matrix $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times p}$ has n observations row-wise, and p features (column-wise).

2. Problem setup

Let us consider the class of regularized learning methods expressed as convex optimization problems, such as (regularized) GLM (McCullagh & Nelder, 1989):

$$\hat{\beta}^{(\lambda)} \in \underset{\beta \in \mathbb{R}^p}{\arg\min}\, \underbrace{f(X\beta) + \lambda\Omega(\beta)}_{P_\lambda(\beta)} \quad \text{(Primal)}. \qquad (1)$$

We highlight two important cases: the regularized least-squares and logistic regression where the loss functions are written as an empirical risk $f(X\beta) = \sum_{i \in [n]} f_i(x_i^\top \beta)$ with the f_i's given in Table 1. The penalty term is often used to incorporate prior knowledges by enforcing a certain regularity on the solutions. For instance, choosing a Ridge penalty (Hoerl & Kennard, 1970) $\Omega(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ improves the stability of the resolution of inverse problems while $\Omega(\cdot) = \|\cdot\|_1$ imposes sparsity at the feature level, a motivation that led to the Lasso estimator (Tibshirani, 1996); see also (Bach et al., 2012) for extensions to other structured penalties.

In practice, obtaining $\hat{\beta}^{(\lambda)}$, an exact solution to Problem (1) is unpractical and one aims achieving a prescribed precision $\epsilon > 0$. More precisely, a (primal) vector $\beta^{(\lambda)} := \beta^{(\lambda,\epsilon)}$ (we will drop the dependency in ϵ for readability) is referred to as an ϵ-solution for λ if its (primal) objective value is optimal at precision ϵ:

$$P_\lambda(\beta^{(\lambda)}) - P_\lambda(\hat{\beta}^{(\lambda)}) \leq \epsilon \ . \qquad (2)$$

We recall and illustrate the notion of approximation path in Figure 1 as described by Giesen et al. (2012).

Definition 1 (ϵ-path). A set $\mathcal{P}_\epsilon \subset \mathbb{R}^p$ is called an ϵ-path for a parameter range $[\lambda_{\min}, \lambda_{\max}]$ if

$$\forall \lambda \in [\lambda_{\min}, \lambda_{\max}], \exists \text{ an } \epsilon\text{-solution } \beta^{(\lambda)} \in \mathcal{P}_\epsilon \ . \quad (3)$$

We call *path complexity* T_ϵ the cardinality of the ϵ-path.

To achieve the targeted ϵ-precision in (2) over a whole path and construct an ϵ-path [1], we rely on *duality gap* evaluations. For that, we compute ϵ_c-solutions[2] (for an accuracy $\epsilon_c < \epsilon$) over a finite grid, and then we control the gap variations w.r.t. λ to achieve the prescribed ϵ-precision over the whole range $[\lambda_{\min}, \lambda_{\max}]$; see Algorithm 1. We now recall the Fenchel duality (Rockafellar, 1997, Chapter 31):

$$\hat{\theta}^{(\lambda)} \in \underset{\theta \in \mathbb{R}^n}{\arg\max} \underbrace{-f^*(-\lambda\theta) - \lambda\Omega^*(X^\top\theta)}_{D_\lambda(\theta)} \quad \text{(Dual)}. \quad (4)$$

For a primal/dual pair $(\beta, \theta) \in \operatorname{dom} P_\lambda \times \operatorname{dom} D_\lambda$, the duality gap is the difference between primal and dual objectives:

$$\mathcal{G}_\lambda(\beta, \theta) = f(X\beta) + f^*(-\lambda\theta) + \lambda(\Omega(\beta) + \Omega^*(X^\top\theta)) \ .$$

Weak duality yields $D_\lambda(\theta) \leq P_\lambda(\beta)$ and

$$P_\lambda(\beta) - P_\lambda(\hat{\beta}^{(\lambda)}) \leq \mathcal{G}_\lambda(\beta, \theta) \ , \quad (5)$$

explaining the interest of the duality gap as an optimality certificate. Using (5), we can safely construct an approximation path for Problem (1) : if $\beta^{(\lambda)}$ is an ϵ-solution for λ, it is guaranteed to remain one for all parameters λ' such that $\mathcal{G}_{\lambda'}(\beta^{(\lambda)}, \theta^{(\lambda)}) \leq \epsilon$. Since the function $\lambda' \mapsto \mathcal{G}_{\lambda'}(\beta^{(\lambda)}, \theta^{(\lambda)})$ does not exhibit a simple dependence in λ, we rely on an upper bound on the gap encoding the structural regularity of the loss function (*e.g.,* 1-dimensional quadratics for strongly convex functions). This bound controls the optimization error as λ varies while preserving optimal complexity on the approximation path.

3. Bounds and approximation path

We introduce the tools to design an approximation path.

3.1. Preliminary results and technical tools

Definition 2. Given a differentiable function f and $x \in \operatorname{dom} f$, let $\mathcal{U}_{f,x}(\cdot)$ and $\mathcal{V}_{f,x}(\cdot)$ be non negative functions that vanish at 0. We say that f is $\mathcal{U}_{f,x}$-convex (resp. $\mathcal{V}_{f,x}$-smooth) at x when Inequality (6) (resp. (7)) is satisfied for any $z \in \operatorname{dom} f$

$$\mathcal{U}_{f,x}(z - x) \leq f(z) - f(x) - \langle \nabla f(x), z - x \rangle \ , \quad (6)$$

$$\mathcal{V}_{f,x}(z - x) \geq f(z) - f(x) - \langle \nabla f(x), z - x \rangle \ . \quad (7)$$

This extends μ-strong convexity and ν-smoothness (Nesterov, 2004) and encompasses smooth uniformly convex losses and generalized self-concordant ones.

Smooth uniformly convex case: In this case, we have

$$\mathcal{U}_{f,x}(z - x) = \mathcal{U}(\|z - x\|),$$
$$\mathcal{V}_{f,x}(z - x) = \mathcal{V}(\|z - x\|),$$

where $\mathcal{U}(\cdot)$ and $\mathcal{V}(\cdot)$ are increasing from $[0, +\infty)$ to $[0, +\infty]$ vanishing at 0; see Azé & Penot (1995). Examples of such functions are $\mathcal{U}(t) = \frac{\mu}{d}t^d$ and $\mathcal{V}(t) = \frac{\nu}{d}t^d$ where d, μ and ν are positive constants. The case $d = 2$ corresponds to strong convexity and smoothness; in general they are called *uniformly convex of order* d, see (Juditski & Nesterov, 2014) or (Bauschke & Combettes, 2011, Ch. 10.2 and 18.5) for details.

Generalized self-concordant case: a $\mathcal{C}^3$ convex function f is (M_f, ν)-generalized self-concordant of order $\nu \geq 2$ and $M_f \geq 0$ if $\forall x \in \operatorname{dom} f$ and $\forall u, v \in \mathbb{R}^n$:

$$\left|\langle \nabla^3 f(x)[v]u, u \rangle\right| \leq M_f \|u\|_x^2 \|v\|_x^{\nu-2} \|v\|_2^{3-\nu} .$$

In this case, Sun & Tran-Dinh (2017, Proposition 10) have shown that one could write:

$$\mathcal{U}_{f,x}(y - x) = w_\nu(-d_\nu(x, y)) \|y - x\|_x^2 \ ,$$
$$\mathcal{V}_{f,x}(y - x) = w_\nu(d_\nu(x, y)) \|y - x\|_x^2 \ ,$$

where the last equality holds if $d_\nu(x, y) < 1$ for the case $\nu > 2$. Closed-form expressions of $w_\nu(\cdot)$ and $d_\nu(\cdot)$ are recalled in Appendix for logistic, quadratic and power losses.

Approximating the duality gap path. Assume we have constructed primal/dual feasible vectors for a finite grid of parameters $\Lambda_T = \{\lambda_0, \ldots, \lambda_{T-1}\}$, *i.e.,* we have at our disposal $(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ for all $\lambda_t \in \Lambda_T$. Let us denote $\mathcal{G}_t = \mathcal{G}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$, and for $\zeta_t = -\lambda_t\theta^{(\lambda_t)}$, $\Delta_t = f(X\beta^{(\lambda_t)}) - f(\nabla f^*(\zeta_t))$. For any function $\phi : \mathbb{R}^n \to [0, +\infty]$ that vanishes at 0, $\rho \in \mathbb{R}$, we define

$$Q_{t,\phi}(\rho) = \mathcal{G}_t + \rho \cdot (\Delta_t - \mathcal{G}_t) + \phi(-\rho \cdot \zeta_t) \ . \quad (8)$$

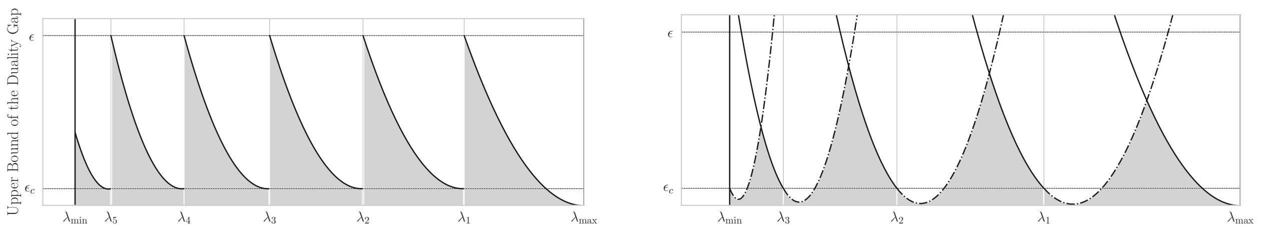
The terms $\mathcal{G}_t$ and Δ_t represent a measure of the optimization error at λ_t. The notation introduced in (8) will be convenient to write concisely upper and lower bounds on the duality gap. This is the goal of the next lemma which leverages regularity of the loss function f, as introduced in Definition 2. This provides control on how the duality gap deviates when one evaluates it for another (close) parameter λ.

Lemma 1. We assume that $-\lambda\theta^{(\lambda_t)} \in \operatorname{dom} f^*$ and $X^\top\theta^{(\lambda_t)} \in \operatorname{dom} \Omega^*$. If f^* is $\mathcal{V}_{f^*}$-smooth (resp. $\mathcal{U}_{f^*}$-convex)[3], then for $\rho = 1 - \lambda/\lambda_t$, the right (resp. left)

[1] note that such a path depends on exact solutions $\hat{\beta}^{(\lambda)}$'s

[2] the c stands for computational in ϵ_c

[3] we drop x in $\mathcal{U}_{f,x}$ and write $\mathcal{U}_f$ if no ambiguity holds.



(a) Uniform unilateral approximation path (as in Proposition 3)

(b) Bilateral approximation path (as in Proposition 4)

Figure 2. Illustration of the construction of ϵ-paths for the Lasso on synthetic dataset generated with `sklearn` as $X, y =$ `make_regression`$(n = 30, p = 150)$ at accuracy $\epsilon = \|y\|_2^2/40$ and $\epsilon_c = \epsilon/10$. We choose $\lambda_{\max} = \|X^\top y\|_\infty$ and $\lambda_{\min} = \lambda_{\max}/20$. For Lasso the bounds are piece-wise quadratic. The shaded gray regions correspond to the regions where the true value of the duality gap lies. We obtain a path complexity of $T_\epsilon = 6$ (resp. $T_\epsilon = 4$) for the unilateral (resp. bilateral) path over $[\lambda_{\min}, \lambda_{\max}]$.

hand side of Inequality (9) holds true

$$Q_{t,\mathcal{U}_{f^*}}(\rho) \leq \mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leq Q_{t,\mathcal{V}_{f^*}}(\rho) \ . \tag{9}$$

Proof. Proof for this result and for other propositions and theorems are deferred to the Appendix. □

The function ϕ, chosen as $\mathcal{V}_{f^*}$ (resp. $\mathcal{U}_{f^*}$) for the upper (resp. lower) bound, essentially captures the regularity needed to approximate the duality gap at λ when using primal/dual vector $(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ for λ_t close to λ. When the function satisfies both inequalities, tightness of the bounds can be related to the conditioning $\mathcal{U}_{f^*}/\mathcal{V}_{f^*}$ of the dual loss f^*. Equality holds for $\mathcal{U}_{f^*} \equiv \mathcal{V}_{f^*} \equiv \frac{1}{2}\|\cdot\|_2^2$ (least-squares), showing the tightness of the bounds.

From Lemma 1, we have $\mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leq \epsilon$ as soon as $Q_{t,\mathcal{V}_{f^*}}(\rho) \leq \epsilon$ where $\rho = 1 - \lambda/\lambda_t$ varies with λ. Hence, we obtain the following proposition that allows to track the regularization path for an arbitrary precision on the duality gap. It proceeds by choosing the largest $\rho = \rho_t$ such that the upper bound in Equation (9) remains below ϵ and leads to Algorithm 1 for computing an ϵ-path.

Proposition 1 (Grid for a prescribed precision)**.** *Given $(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ such that $\mathcal{G}_t \leq \epsilon_c < \epsilon$, for all $\lambda \in \lambda_t \times \left[1 - \rho_t^\ell(\epsilon), 1 + \rho_t^r(\epsilon)\right]$, we have $\mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leq \epsilon$ where $\rho_t^\ell(\epsilon)$ (resp. $\rho_t^r(\epsilon)$) is the largest non-negative ρ s.t. $Q_{t,\mathcal{V}_{f^*}}(\rho) \leq \epsilon$ (resp. $Q_{t,\mathcal{V}_{f^*}}(-\rho) \leq \epsilon$).*

Conversely, given a grid[4] of T parameters $\Lambda_T = \{\lambda_0, \ldots, \lambda_{T-1}\}$, we define ϵ_{Λ_T}, the error of the approximation path on $[\lambda_{\min}, \lambda_{\max}]$ by using a piecewise constant approximation of the map $\lambda \mapsto \mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$:

$$\epsilon_{\Lambda_T} := \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \min_{\lambda_t \in \Lambda_T} \mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \ . \tag{10}$$

This error is however difficult to evaluate in practice so we rely on a tight upper bound based on Lemma 1 that often leads to closed-form expressions.

Proposition 2 (Precision for a given grid)**.** *Given a grid of parameters Λ_T, the set $\{\beta^{(\lambda)} : \lambda \in \Lambda_T\}$ is an ϵ_{Λ_T}-path with $\epsilon_{\Lambda_T} \leq \max_{t \in [T]} Q_{t,\mathcal{V}_{f^*}}(1 - \lambda_t^\star/\lambda_t)$ where for all $t \in \{0, \ldots, T-1\}$, $\lambda_t^\star$ is the largest $\lambda \in [\lambda_{t+1}, \lambda_t]$ such that $Q_{t,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_t) \geq Q_{t+1,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_{t+1})$.*

Construction of dual feasible vector. We rely on gradient rescaling to produce a dual feasible vector:

Lemma 2. For any $\beta^{(\lambda_t)} \in \mathbb{R}^p$, the vector

$$\theta^{(\lambda_t)} = \frac{-\nabla f(X\beta^{(\lambda_t)})}{\max(\lambda_t, \sigma^\circ_{\mathrm{dom}\,\Omega^*}(X^\top \nabla f(X\beta^{(\lambda_t)}))} \ ,$$

is feasible: $-\lambda\theta^{(\lambda_t)} \in \mathrm{dom}\, f^*$, $X^\top\theta^{(\lambda_t)} \in \mathrm{dom}\,\Omega^*$.

Remark 1. When the regularization is a norm, $\Omega(\cdot) = \|\cdot\|$ then $\sigma^\circ_{\mathrm{dom}\,\Omega^*}$ is the associated dual norm $\|\cdot\|_*$.

The dual $\theta^{(\lambda_t)}$ in Lemma 2 implies that $\mathcal{G}_t$ and Δ_t converge to 0 when $\beta^{(\lambda_t)}$ converges to $\hat{\beta}^{(\lambda_t)}$ (Ndiaye et al., 2017).

Finding ρ. Following Proposition 1, a 1-dimensional equation $Q_{t,\mathcal{V}_{f^*}}(\rho) = \epsilon$ needs to be solved to obtain an ϵ-path. This can be done efficiently at high precision by numerical solvers if no explicit solution is available.

As a corollary from Lemma 1 and Proposition 2, we recover the analysis by Giesen et al. (2012):

Corollary 1. If the function f^* is $\frac{\nu}{2}\|\cdot\|^2$-smooth, the left (ρ_t^ℓ) and right (ρ_t^r) step sizes defined in Proposition 1 have closed-form expressions:

$$\rho_t^\ell = \frac{\sqrt{2\nu\delta_t \left\|\zeta_t\right\|^2 + \tilde{\delta}_t^2} - \tilde{\delta}_t}{\nu \left\|\zeta_t\right\|^2}, \rho_t^r = \frac{\sqrt{2\nu\delta_t \left\|\zeta_t\right\|^2 + \tilde{\delta}_t^2} + \tilde{\delta}_t}{\nu \left\|\zeta_t\right\|^2},$$

where $\delta_t := \epsilon - \mathcal{G}_t$ and $\tilde{\delta}_t := \Delta_t - \mathcal{G}_t$. This is simplified to $\delta_t = \epsilon - \epsilon_c$ and $\tilde{\delta}_t = 0$ when $\max(\mathcal{G}_t, \Delta_t) \leq \epsilon_c$.

[4] we assume a decreasing order $\lambda_{t+1} < \lambda_t$, reflecting common practices for GLM, *e.g.,* for the Lasso.

Algorithm 1 `training_path`

```
Input: $f, \Omega, \epsilon, \epsilon_c, [\lambda_{\min}, \lambda_{\max}]$
Initialization: $t = 0$, $\lambda_0 = \lambda_{\max}$, $\Lambda = \{\lambda_{\max}\}$
repeat
  Get $\beta^{(\lambda_t)}$ solving (1) to accuracy $\mathcal{G}_t \le \epsilon_c < \epsilon$
  Compute the step size $\rho_t^\ell(\epsilon)$ following Proposition 3, 4, 5.
  Set $\lambda_{t+1} = \max(\lambda_t \times (1 - \rho_t^\ell), \lambda_{\min})$
  $\Lambda \leftarrow \Lambda \cup \{\lambda_{t+1}\}$ and $t \leftarrow t + 1$
until $\lambda_t \le \lambda_{\min}$
Return: $\{\beta^{(\lambda_t)} \ : \ \lambda_t \in \Lambda\}$
```

3.2. Discretization strategies

We now establish new strategies for the exploration of the hyperparameter space in the search for an ϵ-path.

For regularized learning methods, it is customary to start from a large regularizer[5] $\lambda_0 = \lambda_{\max}$ and then to perform the computation of $\hat{\beta}^{(\lambda_{t+1})}$ after the one of $\hat{\beta}^{(\lambda_t)}$, until the smallest parameter of interest $\lambda_{\min}$ is reached. Models are generally computed by increasing complexity, allowing important speed-ups due to *warm start* (Friedman et al., 2007) when the λ's are close to each other. Knowing λ_t, we provide a recursive strategy to construct λ_{t+1}.

Adaptive unilateral. The strategy we call *unilateral* consists in computing the new parameter as $\lambda_{t+1} = \lambda_t \times (1 - \rho_t^\ell(\epsilon))$ as in Proposition 1.

Proposition 3 (Unilateral approximation path)**.** *Assume that f^* is $\mathcal{V}_{f^*}$-smooth. We construct the grid of parameters $\Lambda^{(u)}(\epsilon) = \{\lambda_0, \dots, \lambda_{T_\epsilon - 1}\}$ by*

$$\lambda_0 = \lambda_{\max}, \quad \lambda_{t+1} = \lambda_t \times (1 - \rho_t^\ell(\epsilon)) \ ,$$

and $(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ s.t. $\mathcal{G}_t \le \epsilon_c < \epsilon$ for all t. Then, the set $\{\beta^{(\lambda_t)} \ : \ \lambda_t \in \Lambda^{(u)}(\epsilon)\}$ is an ϵ-path for Problem (1).

This strategy is illustrated in Figure 2(a) on a Lasso case, and stands as a generic one to compute an approximation path for loss functions satisfying assumptions in Definition 2.

Adaptive bilateral. For uniformly convex functions, we can make a larger step by combining the information given by the left and right step sizes. Indeed, let us assume that we explore the parameter range $[\lambda_{\min}, \lambda_{\max}]$. Starting from a parameter λ_t, we define the next step, given by Proposition 1, $\lambda_t^\ell := \lambda_t(1 - \rho_t^\ell)$. Then it exists $\lambda_{t'} \le \lambda_t^\ell$ such that $\lambda_{t'}^r := \lambda_{t'}(1 + \rho_{t'}^r) = \lambda_t^\ell$. Thus a larger step can be done by using $\lambda_{t'} = \lambda_t \times (1 - \rho_t^\ell)/(1 + \rho_{t'}^r)$. However $\rho_{t'}^r$ depends on the (approximated) solution $\beta^{(\lambda_{t'})}$ that we do not know before optimizing the problem for parameter $\lambda_{t'}$ when computing sequentially the grid points in decreasing order *i.e.,* $\lambda_{t'} \le \lambda_t$. We overcome this issue in Lemma 3 by (upper) bounding all the constants in $Q_{t',\mathcal{V}_{f^*}}(\rho)$ that depend on the solution $\beta^{(\lambda_{t'})}$, by constants involving only information given by $\beta^{(\lambda_t)}$.

Lemma 3. Assuming f uniformly smooth yields $\|\nabla f(X\beta^{(\lambda_{t'})})\|_* \le \widetilde{R}_t$, where $\widetilde{R}_t := \mathcal{V}_f^{*-1}\big(f(X\beta^{(\lambda_t)}) + \frac{2\epsilon_c}{\rho_t^\ell(\epsilon)}\big)$. If additionally f is uniformly convex, this yields $\Delta_{t'} \le \widetilde{\Delta}_t$, where $\widetilde{\Delta}_t := \widetilde{R}_t \times \mathcal{U}_f^{-1}(\epsilon_c)$ as well as $\mathcal{G}_\lambda(\beta^{(\lambda_{t'})}, \theta^{(\lambda_{t'})}) \le Q_{t',\mathcal{V}_{f^*}}(\rho) \le \widetilde{Q}_{t,\mathcal{V}_{f^*}}(\rho)$, where

$$\widetilde{Q}_{t,\mathcal{V}_{f^*}}(\rho) = \epsilon_c + \rho \cdot (\widetilde{\Delta}_t - \epsilon_c) + \mathcal{V}_{f^*}\left(|\rho| \cdot \widetilde{R}_t\right) \ .$$

Let us now define $\rho_t^{(b)}(\epsilon) = \dfrac{\rho_t^\ell(\epsilon) + \tilde{\rho}_t^r(\epsilon)}{1 + \tilde{\rho}_t^r(\epsilon)}$, where $\rho_t^\ell(\epsilon)$ is defined in Proposition 1 and $\tilde{\rho}_t^r(\epsilon)$ is the largest non negative ρ such that $\widetilde{Q}_{t,\mathcal{V}_{f^*}}(\rho) \le \epsilon$ in Lemma 3.

Proposition 4 (Bilateral Approximation Path)**.** *Assume that f is uniformly convex and smooth. We construct the grid $\Lambda^{(b)}(\epsilon) = \{\lambda_0, \dots, \lambda_{T_\epsilon - 1}\}$ by*

$$\lambda_0 = \lambda_{\max}, \quad \lambda_{t+1} = \lambda_t \times (1 - \rho_t^{(b)}(\epsilon)) \ ,$$

and $(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ s.t. $\mathcal{G}_t \le \epsilon_c < \epsilon$ for all t. Then the set $\{\beta^{(\lambda_t)} \ : \ \lambda_t \in \Lambda^{(b)}(\epsilon)\}$ is an ϵ-path for Problem (1).

This strategy is illustrated in Figure 2(b) on a Lasso example.

Uniform unilateral and bilateral. In some cases, it may be advantageous to have access to a predefined grid before launching a hyperparameter selection procedure such as `hyperband` (Li et al., 2017) or for parallel computations. Given the initial information from the initialization $(\beta^{(\lambda_0)}, \theta^{(\lambda_0)})$, we can build a uniform grid that guarantees an ϵ-approximation before solving any optimization problem. Indeed, by applying Lemma 3 at $t = 0$, we have $\mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \le \widetilde{Q}_{0,\mathcal{V}_{f^*}}(\rho)$. We can define $\widetilde{\rho}_0^\ell(\epsilon)$ (resp. $\widetilde{\rho}_0^r(\epsilon)$) as the largest non-negative ρ s.t. $\widetilde{Q}_{0,\mathcal{V}_{f^*}}(\rho) \le \epsilon$ (resp. $\widetilde{Q}_{0,\mathcal{V}_{f^*}}(-\rho) \le \epsilon$) and also

$$\rho_0(\epsilon) = \begin{cases} \widetilde{\rho}_0^\ell(\epsilon) & \text{for \textit{unilateral} path,} \\ \frac{\widetilde{\rho}_0^\ell(\epsilon) + \widetilde{\rho}_0^r(\epsilon)}{1 + \widetilde{\rho}_0^r(\epsilon)} & \text{for \textit{bilateral} path.} \end{cases} \tag{11}$$

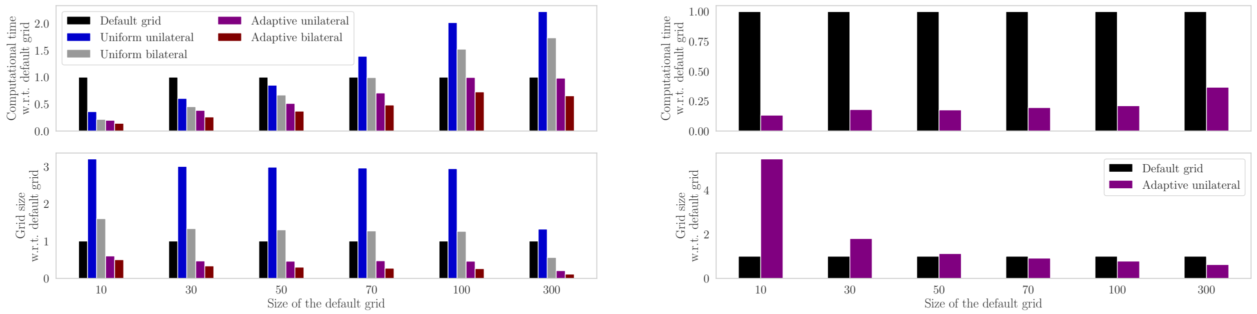
Proposition 5 (Uniform approximation path)**.** *Assume that f is uniformly convex and smooth, and define the grid $\Lambda^{(0)}(\epsilon) = \{\lambda_0, \dots, \lambda_{T_\epsilon - 1}\}$ by*

$$\lambda_0 = \lambda_{\max}, \quad \lambda_{t+1} = \lambda_t \times (1 - \rho_0(\epsilon)) \ ,$$

and $\forall t \in [T], (\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ s.t. $\mathcal{G}_t \le \epsilon_c < \epsilon$. Then the set $\{\beta^{(\lambda_t)} \ : \ \lambda_t \in \Lambda^{(0)}(\epsilon)\}$ is an ϵ-path for Problem (1) *with at most T_ϵ grid points where*

$$T_\epsilon = \left\lfloor \frac{\log(\lambda_{\min}/\lambda_{\max})}{\log(1 - \rho_0(\epsilon))} \right\rfloor \ . \tag{12}$$

[5] for the Lasso one often chooses $\lambda_0 = \lambda_{\max} := \|X^\top y\|_\infty$



(a) ℓ_1 least-squares regression on climate data set `NCEP/NCAR Reanalysis`; $n = 814$ observations and $p = 73577$ features.

(b) ℓ_1 logistic regression on `leukemia` data set with $n = 72$ observations and $p = 7129$ features.

Figure 3. Computation of the approximation path to reach the same error than the default grid ($\epsilon = 10^{-4} \left\|y\right\|^2$ for the least-squares case and $\epsilon = 10^{-4} \min(n_1, n_2)/n$ where n_i is the number of observations in the class $i \in \{0, 1\}$, for the logistic case). We have used the same (vanilla) coordinate descent optimization solver with warm start between parameters for all grids. Note that a smaller grid do not imply faster computation, as the interplay with the warm-start can be intricate in our sequential approach.

3.3. Limitations of previous framework

Previous algorithms for computing ϵ-paths have been initially developed with a complexity of $O(1/\epsilon)$ (Clarkson, 2010; Giesen et al., 2010) in a large class of problems. Yet, losses arising in machine learning have often nicer regularities that can be exploited. This is all the more striking in the Lasso case where a better complexity in $O(1/\sqrt{\epsilon})$ was obtained by Mairal & Yu (2012); Giesen et al. (2012).

The relation between path complexity and regularity of the objective function remains unclear and previous methods do not apply to all popular learning problems. For instance the dual loss f^* of the logistic regression is not uniformly smooth. So to apply the previous theory, one needs to optimize on a (potentially badly pre-selected) compact set.

Let us consider the one dimensional toy example where $p = 1$, $\beta \in \mathbb{R}$, $X = \mathrm{Id}_p$, $y = -1$ and the loss function $f(X\beta) = \log(1 + \exp(\beta))$. We have, $\nabla^2 f(\beta) = \exp(\beta)/(1 + \exp(\beta))^2$. Then for Problem (1), since $P_\lambda(\hat{\beta}^{(\lambda)}) \leq P_\lambda(0)$, we have $|\hat{\beta}^{(\lambda)}| \in [0, \log(2)/\lambda]$ and a smoothness constant $\nu_{f^*} \approx \exp(\lambda)$ for the dual can be estimated at each step. This leads to an unreasonable algorithm with tiny step sizes in Corollary 1. Also, the algorithm proposed by Giesen et al. (2012) can not be applied for the logistic loss since the dual function is not polynomial.

Our proposed algorithm does not suffer from such limitations and we introduce a finer analysis that takes into account the regularity of the loss functions.

3.4. Complexity and regularity

Lower bound on path complexity. For our method, the lower bound on the duality gap quantifies how close the from Proposition 1 is from the best possible one can achieve for smooth loss functions. Indeed, at optimal solution, we have $\mathcal{G}_t = \Delta_t = 0$. Thus the largest possible step — starting at λ_t and moving in decreasing order — is given by the smallest $\lambda \in [\lambda_{\min}, \lambda_t]$ such that $\mathcal{U}_{f^*}(-\hat{\zeta}_t \times \rho) > \epsilon$ where $\hat{\zeta}_t = -\lambda_t \hat{\theta}^{(\lambda_t)}$. Hence, *any* algorithm for computing an ϵ-path for $\mathcal{U}_{f^*}$-uniformly convex dual loss, have necessarily a complexity of order at least $O(1/\mathcal{U}_{f^*}^{-1}(\epsilon))$.

Upper bounds. We remind that we write T_ϵ for the complexity of our proposed approximation path *i.e.,* the cardinality of the grid returned by Algorithm 1. In the following proposition, we propose a bound on the complexity *w.r.t.* the regularity of the loss function. Discussions on the constants and assumptions are provided in the Appendix.

Proposition 6 (Approximation path: complexity)**.** *Assuming that* $\max(\mathcal{G}_t, \Delta_t) \leq \epsilon_c < \epsilon$ *at each step* t*, there exists an explicit constant* $C_f(\epsilon_c) > 0$ *such that*

$$T_\epsilon \leq \log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) \times \frac{C_f(\epsilon_c)}{\mathcal{W}_{f^*}(\epsilon - \epsilon_c)} \quad , \tag{13}$$

where for all $t > 0$*, the function* $\mathcal{W}_{f^*}$ *is defined by*

$$\mathcal{W}_{f^*}(\cdot) = \begin{cases} \mathcal{V}_{f^*}^{-1}(\cdot), & \textit{if } f \textit{ is uniformly convex and smooth} \\ \sqrt{\cdot}, & \textit{if } f \textit{ is Generalized Self-Concordant} \\ & \quad \textit{and uniformly-smooth.} \end{cases}$$

Moreover, $C_f(\epsilon_c)$ *is an uniform upper bound of* $\left\|\zeta_t\right\|_*$ *along the path, that tends to a constant* C_f *when* ϵ_c *goes to* 0.

Proposition 6 applied to the special case when f is ν-smooth and μ-strongly convex reads $\log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)\sqrt{\frac{\nu}{\mu}\frac{f(X\beta^{(\lambda_0)})}{\epsilon - \epsilon_c}}$ for the complexity for any data X, y. This is not explicitly dependent on the dimension n and p and are more scalable.

Algorithm 2 ϵ_v-path for Validation Set

```
    **Input:** $f, \Omega, \epsilon_v, [\lambda_{\min}, \lambda_{\max}]$
    Compute $\epsilon_{v,\mu}$ as in Proposition 7
    $\Lambda(\epsilon_{v,\mu}) = \texttt{training\_path}(f, \Omega, \epsilon_{v,\mu}, [\lambda_{\min}, \lambda_{\max}])$
    **Return:** $\Lambda(\epsilon_{v,\mu})$
```

4. Validation path

To achieve good generalization performance, estimators defined as solutions of Problem 1 require a careful adjustment of λ to balance data-fitting and regularization. A standard approach to calibrate such a parameter is to select it by comparing the validation errors on a finite grid (say with K-fold cross-validation). Unfortunately, it is often difficult to determine a priori the grid limits, the number of λ's (number of points in the grid) or how they should be distributed to achieve low validation error.

Considering the validation data (X', y') with n' observations and loss[6] $\mathcal{L}$, we define the validation error for $\beta \in \mathbb{R}^p$:

$$E_v(\beta) = \mathcal{L}(y', X'\beta) \ . \tag{14}$$

For selecting a hyperparameter, we leverage our approximation path to solve the bi-level problem

$$\begin{aligned} \mathop{\arg\min}_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \quad & E_v(\hat{\beta}^{(\lambda)}) = \mathcal{L}(y', X'\hat{\beta}^{(\lambda)}) \\ & \text{s.t. } \hat{\beta}^{(\lambda)} \in \mathop{\arg\min}_{\beta \in \mathbb{R}^p} f(X\beta) + \lambda\Omega(\beta) \ . \end{aligned}$$

Recent works have addressed this problem by using gradient-based algorithms, see for instance Pedregosa (2016); Franceschi et al. (2018) who have shown promising results in computational time and scalability w.r.t. multiple hyperparameters. Yet, they require assumptions such as smoothness of the validation function E_v and non-singular Hessian of the inner optimization problem at optimal values which are difficult to check in practice since they depend on the optimal solutions $\hat{\beta}^{(\lambda)}$. Moreover, they can only guarantee convergence to stationary point.

In this section, we generalize the approach of (Shibagaki et al., 2015) and show that with a safe and simple exploration of the parameter space, our algorithm has a global convergence property. For that, we assume the following conditions on the validation loss and on the inner optimization objective throughout the section:

A1 $|\mathcal{L}(a, b) - \mathcal{L}(a, c)| \leq \mathcal{L}(b, c)$ for any $a, b, c \in \mathbb{R}^n$.

A2 The function $\beta \mapsto P_\lambda(\beta)$ is μ-strongly convex.

The assumption on the loss function is verified for norms (regression) and indicator functions (classification). Indeed, if $\mathcal{L}(a, b) = \|a - b\|$, **A1** corresponds to the triangle inequality. For the $\ell_{0/1}$-loss $\mathcal{L}(a, b) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{a_i b_i < 0}$, since for any real s, u and v, $|\mathbf{1}_{us<0} - \mathbf{1}_{uv<0}| \leq \mathbf{1}_{sv<0}$, one has

$$\left| \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{a_i b_i < 0} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{a_i c_i < 0} \right| \leq \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{b_i c_i < 0} \ .$$

Definition 3. Given a primal solution $\hat{\beta}^{(\lambda)}$ for parameter λ and a primal point $\beta^{(\lambda_t)}$ returned by an algorithm, we define the gap on the validation error between λ and λ_t as

$$\Delta E_v(\lambda_t, \lambda) := \left| E_v(\hat{\beta}^{(\lambda)}) - E_v(\beta^{(\lambda_t)}) \right| \ . \tag{15}$$

Suppose we have fixed a tolerance ϵ_v on the gap on validation error *i.e.*, $\Delta E_v(\lambda_t, \lambda) \leq \epsilon_v$. Based on Assumption **A1**, if there is a region $\mathcal{R}_\lambda$ that contains the optimal solution $\hat{\beta}^{(\lambda)}$ at parameter λ, then we have

$$\begin{aligned} \Delta E_v(\lambda_t, \lambda) &\leq \mathcal{L}(X'\hat{\beta}^{(\lambda)}, X'\beta^{(\lambda_t)}) \\ &\leq \max_{\beta \in \mathcal{R}_\lambda} \mathcal{L}(X'\beta, X'\beta^{(\lambda_t)}) \ . \end{aligned}$$

A simple strategy consists in choosing $\mathcal{R}_\lambda$ as a ball.

Lemma 4 (Gap safe region Ndiaye et al. (2017))**.** Under **A2**, any primal solution $\hat{\beta}^{(\lambda)}$ belongs to the Euclidean ball with center $\beta^{(\lambda_t)}$ and radius

$$r_{t,\mu}(\lambda) = \sqrt{\frac{2}{\mu}\,\mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})} \ . \tag{16}$$

Such a safe ball leveraging duality gap has been proved useful to speed-up sparse optimization solvers. The improve performance relies on the ability to identify the sparsity structure of the optimal solutions; approaches of this type are referred to as *safe screening rules* as they provide *safe* certificates for such structures (El Ghaoui et al., 2012; Fercoq et al., 2015; Shibagaki et al., 2016; Ndiaye et al., 2017).

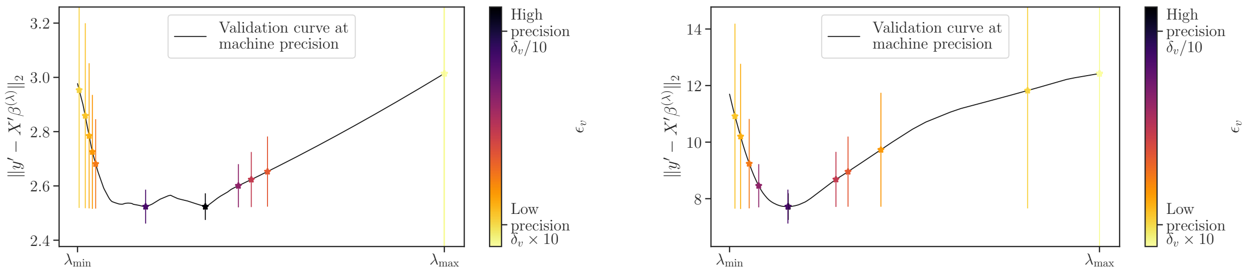
Since the radius in Equation (16) depends explicitly on the duality gap, we can sequentially track a range of parameters for which the gap on the validation error remains below a prescribed tolerance by controlling the optimization error.

Proposition 7 (Grid for prescribed validation error)**.** *Under Assumptions **A1** and **A2**, let us define for* $i \in [n']$ *an index in the test set,* $\xi_i = \left(\frac{x_i'^\top \beta^{(\lambda_t)}}{\|x_i'\|}\right)^2$ *and*

$$\epsilon_{v,\mu} = \begin{cases} \frac{\mu}{2} \times \left(\frac{\epsilon_v}{\|X'\|}\right)^2, & \textit{(regression)} \\ \frac{\mu}{2} \times \xi_{(\lfloor n\epsilon_v \rfloor + 1)}, & \textit{(classification)} \end{cases} \tag{17}$$

where $\xi_{(\lfloor n\epsilon_v \rfloor + 1)}$ *is the* $(\lfloor n\epsilon_v \rfloor + 1)$*-th smallest value of* ξ_i*'s. Given* $(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ *such that* $\mathcal{G}_t \leq \epsilon_{v,\mu}$*, we have* $\Delta E_v(\lambda_t, \lambda) \leq \epsilon_v$ *for all parameter* λ *in the interval* $\lambda_t \times \left[1 - \rho_t^\ell(\epsilon_{v,\mu}),\, 1 + \rho_t^r(\epsilon_{v,\mu})\right]$ *, where* $\rho_t^\ell(\epsilon_{v,\mu})$*,* $\rho_t^r(\epsilon_{v,\mu})$ *are defined in Proposition 1.*

[6] the data-fitting terms might differ from training to testing; for instance for logistic regression the $\ell_{0/1}$-loss is used for validation but the logistic function is optimized at training.



(a) Synthetic data set generated using the `sklearn` command $X, y =$ `make_sparse_uncorrelated`$(n = 30, p = 50)$.

(b) Synthetic data set generated using the `sklearn` command $X, y =$ `make_regression`$(n = 500, p = 5000)$.

Figure 4. Safe selection of the optimal hyperparameter for Enet on the validation set (30% of the observations). The targeted accuracy ϵ_v is refined from $\delta_v \times 10$ to $\delta_v/10$ with $\delta_v = \max_{\lambda_t \in \Lambda} E_v(\beta^{(\lambda_t)}) - \min_{\lambda_t \in \Lambda} E_v(\beta^{(\lambda_t)})$ and Λ is the default grid between $\lambda_{\max} = \|X^\top y\|_\infty$ and $\lambda_{\min} = \lambda_{\max}/100$ of size $T = 200$. The stars represent the worst case solution amount the one generated by Algorithm 2 (with bilateral path). For loose precision suboptimal parameters are identified, but better ones are found as the accuracy ϵ_v decreases.

Remark 2 (Stopping criterion for training)**.** For the current parameter λ_t, $\Delta E_v(\lambda_t, \lambda_t) \leq \epsilon_v$ as soon as $\mathcal{G}_t \leq \epsilon_{v,\mu}$, which gives us a stopping criterion for optimizing on the training part (X, y) relative to the desired accuracy ϵ_v on the validation data (X', y'). This has the appealing property of relieving the practitioner from selecting the stopping criterion ϵ_c when optimizing on the training set.

Algorithm 2 outputs a discrete set of parameters $\Lambda(\epsilon_{v,\mu})$ such that $\{\beta^{(\lambda_t)}$ for $\lambda_t \in \Lambda(\epsilon_{v,\mu})\}$ is an ϵ_v-path for the validation error. Thus, for any $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, there exists $\lambda_t \in \Lambda(\epsilon_{v,\mu})$ such that

$$E_v(\beta^{(\lambda_t)}) - \epsilon_v \leq E_v(\hat{\beta}^{(\lambda)}) \ . \tag{18}$$

The following proposition is obtained by taking the minimum on both sides of the inequality.

Proposition 8. *Under Assumptions **A1** and **A2**, the set* $\{\beta^{(\lambda_t)}$ *for* $\lambda_t \in \Lambda(\epsilon_{v,\mu})\}$ *is an* ϵ_v*-path for the error and*

$$\min_{\lambda_t \in \Lambda(\epsilon_{v,\mu})} E_v(\beta^{(\lambda_t)}) - \min_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} E_v(\hat{\beta}^{(\lambda)}) \leq \epsilon_v \ .$$

5. Numerical experiments

We illustrate our method on ℓ_1-regularized least squares and logistic regression by comparing the computational times and number of grid points needed to compute an ϵ-path for a given range $[\lambda_{\min}, \lambda_{\max}]$ for several strategies.

The ”Default grid” is the one used by default in the packages `glmnet` (Friedman et al., 2010) and `sklearn` (Pedregosa et al., 2011). It is defined as $\lambda_t = \lambda_{\max} \times 10^{-\delta t/(T-1)}$ (here $\delta = 3$). The proposed grids are the adaptive unilateral/bilateral and uniform unilateral/bilateral grids that are defined in Propositions 3, 4 and 5.

Thanks to Proposition 2, we measure the approximation path error ϵ of the default grid of size T and report the times and numbers of grid points T_ϵ needed to achieve such a precision. Our experiments were conducted on the `leukemia` dataset, available in `sklearn` and the climate dataset `NCEP/NCAR Reanalysis` (Kalnay et al., 1996). The optimization algorithms are the same for all the grid, hence we compare only the grid construction impact. Results are reported in Figure 3 for classification and regression problem. Our approach leads to better guarantees for approximating the regularization path w.r.t. the default grid and often significant gain in computing time.

Figure 4 illustrates convergence for Elastic Net (Enet) (Zou & Hastie, 2005), on synthetic data generated by `sklearn` as random regression problems `make_regression` and `make_sparse_uncorrelated` (Celeux et al., 2012). For a decreasing levels of validation error, we represent the λ selected by our algorithm and its corresponding safe interval. Even when the validation curve is non smooth and non convex, the output of the safe grid search converges to the global minimum as stated in Proposition 8.

6. Conclusion

We have shown how to efficiently construct one dimensional grids of regularization parameters for convex risk minimization, and to get an automatic calibration, optimal in term of hold-out test error. Future research could examine how to adapt our framework to address multi-dimensional parameter grids. This case is all the more interesting that it naturally arises when addressing non-convex problems, *e.g.,* MCP or SCAD, with re-weighted ℓ_1-minimization. Approximation of a full path then requires to optimize up to precision ϵ_c at each step, even for non promising hyperparameter, which is time consuming. Combining our approach with safe elimination procedures could provide faster hyperparameter selection algorithms.

Acknowledgements

This work was supported by the Chair Machine Learning for Big Data at Télécom ParisTech. TL acknowledges the support of JSPS KAKENHI Grant number $17K12745$.

References

Arlot, S. and Celisse, A. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

Azé, D. and Penot, J.-P. Uniformly convex and uniformly smooth convex functions. In *Annales de la faculté des sciences de Toulouse*, pp. 705–730. Université Paul Sabatier, 1995.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Convex optimization with sparsity-inducing norms. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

Bauschke, H. H. and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, 2011. ISBN 978-1-4419-9466-0.

Bottou, L. and Bousquet, O. The tradeoffs of large scale learning. In *NIPS*, pp. 161–168, 2008.

Celeux, G., Anbari, M. E., Marin, J.-M., and Robert, C. P. Regularization in regression: comparing bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7(2):477–502, 2012.

Clarkson, K. L. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):63:1–63:30, 2010.

Efron, B., Hastie, T., Johnstone, I. M., and Tibshirani, R. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

El Ghaoui, L., Viallon, V., and Rabbani, T. Safe feature elimination in sparse supervised learning. *J. Pacific Optim.*, 8 (4):667–698, 2012.

Fercoq, O., Gramfort, A., and Salmon, J. Mind the duality gap: safer rules for the lasso. In *ICML*, pp. 333–342, 2015.

Franceschi, L., Frasconi, P., Salzo, S., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, pp. 1563–1572, 2018.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2): 302–332, 2007.

Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1, 2010.

Gärtner, B., Jaggi, M., and Maria, C. An exponential lower bound on the complexity of regularization paths. *Journal of Computational Geometry*, 2012.

Giesen, J., Jaggi, M., and Laue, S. Approximating parameterized convex optimization problems. In *European Symposium on Algorithms*, pp. 524–535, 2010.

Giesen, J., Müller, J. K., Laue, S., and Swiercy, S. Approximating concavely parameterized optimization problems. In *NIPS*, pp. 2105–2113, 2012.

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. The entire regularization path for the support vector machine. *J. Mach. Learn. Res.*, 5:1391–1415, 2004.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Juditski, A. and Nesterov, Y. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792*, 2014.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77(3): 437–471, 1996.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

Li, Y. and Singer, Y. The well tempered lasso. *ICML*, 2018.

Mairal, J. and Yu, B. Complexity analysis of the lasso regularization path. In *ICML*, pp. 353–360, 2012.

McCullagh, P. and Nelder, J. A. *Generalized Linear Models*. Chapman & Hall, 1989.

Ndiaye, E., Fercoq, O., Gramfort, A., and Salmon, J. Gap safe screening rules for sparsity enforcing penalties. *J. Mach. Learn. Res.*, 18(128):1–33, 2017.

Nesterov, Y. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

Osborne, M. R., Presnell, B., and Turlach, B. A. A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.*, 20(3):389–403, 2000.

Park, M. Y. and Hastie, T. L1-regularization path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B*, 69(4):659–677, 2007.

Pedregosa, F. Hyperparameter optimization with approximate gradient. In *ICML*, pp. 737–746, 2016.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

Rockafellar, R. T. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

Rosset, S. and Zhu, J. Piecewise linear regularized solution paths. *Ann. Statist.*, 35(3):1012–1030, 2007.

Shibagaki, A., Suzuki, Y., Karasuyama, M., and Takeuchi, I. Regularization path of cross-validation error lower bounds. In *NIPS*, pp. 1666–1674, 2015.

Shibagaki, A., Karasuyama, M., Hatano, K., and Takeuchi, I. Simultaneous safe screening of features and samples in doubly sparse modeling. In *ICML*, pp. 1577–1586, 2016.

Sun, T. and Tran-Dinh, Q. Generalized self-concordant functions: A recipe for newton-type methods. *Mathematical Programming*, 2017.

Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, 67(2): 301–320, 2005.