Safe Grid Search with Optimal Complexity

Eugene Ndiaye, Tam Le, Olivier Fercoq, Joseph Salmon, Ichiro Takeuchi

Safe Grid Search with Optimal Complexity

Eugene Ndiaye
Télécom ParisTech

Tam Le
RIKEN AIP

Olivier Fercoq
Télécom ParisTech

Joseph Salmon
Université de Montpellier

Ichiro Takeuchi
Nagoya Institute of Technology

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 $d > 0$ 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 by previous works). We leverage our technique to provide refined bounds on the validation error as well as a practical algorithm for hyperparameter tuning. The later 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 it 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 non negative hyperparameter λ. The choice of λ is crucial since it directly influences the generalization performance of the estimator, *i.e.,* its score on unseen data sets. One of the most popular method in such a context is cross-validation, see Arlot and Celisse (2010) for a detailed review. For simplicity, we investigate here the holdout version that consists in splitting the data in two parts: on the first part (*training set*) the method is trained for a pre-defined collection of candidates $\Lambda_T := \{\lambda_0, \dots, \lambda_{T-1}\}$, and on the second part (*validation set*), the best parameter is selected.

For a piecewise quadratic loss f and a piecewise linear regularization Ω (*e.g.,* the Lasso estimator), (Osborne et al., 2000; Rosset and 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 and 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. Even in favorable cases, a complexity linear in p can be expensive to compute for large p.
- 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 crucial issue whenever the regularization parameter decreases.
- 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 apply 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 and Bousquet, 2008), it is not necessary to optimize below the statistical error to

Email: eugene.ndiaye@telecom-paristech.fr; tam.le@riken.jp; olivier.fercoq@telecom-paristech.fr; joseph.salmon@umontpellier.fr; takeuchi.ichiro@nitech.ac.jp

reach good generalization property. Exact regularization path algorithms require maintaining optimality conditions when the hyperparameter varies, which is time consuming.

To overcome these issues, an approximation of the solution path up to accuracy $\epsilon > 0$ was proposed and optimal complexity was proven to be $O(1/\epsilon)$ by Giesen et al. (2010) in a fairly general setting. A noticeable contribution was proposed by Mairal and Yu (2012), that come up with an algorithm whose complexity is $O(1/\sqrt{\epsilon})$ for the Lasso case. The later result was then extended by Giesen et al. (2012) to objective function that has a quadratic lower bound while providing a lower and upper bound of order $O(1/\sqrt{\epsilon})$. Unfortunately, these assumptions fail to hold for a large class of 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 later 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 an upper and lower bound on 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 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 optimality guarantees. We discuss 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 $d > 0$ (see (Bauschke and 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 and Tran-Dinh, 2017). Finally, 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.

Notation. Given a proper, closed and convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, we denote $\operatorname{dom} f = \{x \in \mathbb{R}^n : f(x) < +\infty\}$. If f is a twice continuously differentiable function with positive definite Hessian $\nabla^2 f(x)$ at any $x \in \operatorname{dom} f$, 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 $f^*(x^*) = \sup_{x \in \operatorname{dom} f} \langle x^*, x\rangle - f(x)$. 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 $\{1, \dots, T\}$ for any non zero integer T. The vector of observations is $y \in \mathbb{R}^n$ and the design matrix $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}$ has n observations row-wise, and p features (column-wise).

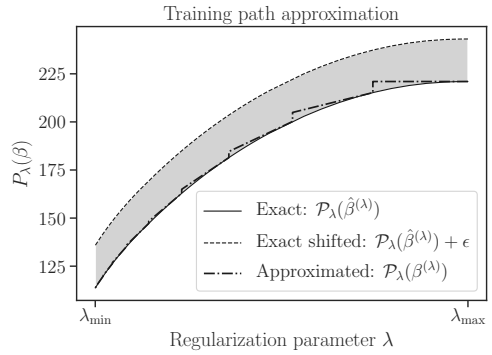


Figure 1: Illustration of the approximation path for the Lasso at accuracy $\epsilon = \|y\|_2^2/20$. We choose $\lambda_{\max} = \|X^\top y\|_\infty$ and $\lambda_{\min} = \lambda_{\max}/50$. The shaded gray region shows the interval where any ϵ-path must lie. The exact path is computed with the `LassoLars`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

2 Problem setup

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

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

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 and Kennard, 1970) $\Omega(\cdot) = \|\cdot\|_2^2/2$ 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

	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 $\mathrm{Nh}(x) = x\log(x) + (1-x)\log(1-x)$

ϵ for readability) is referred to as an ϵ-solution for λ if its (primal) objective value is optimal at precision ϵ:

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

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

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

We call *path complexity* T_ϵ for Problem (1) 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):

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

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

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

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

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

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.,* when f is strongly convex, we consider a 1-dimensional quadratic). This bound allows to control the optimization error as λ varies while preserving an optimal complexity on the approximation path.

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

[2] the c stands for computational in ϵ_c

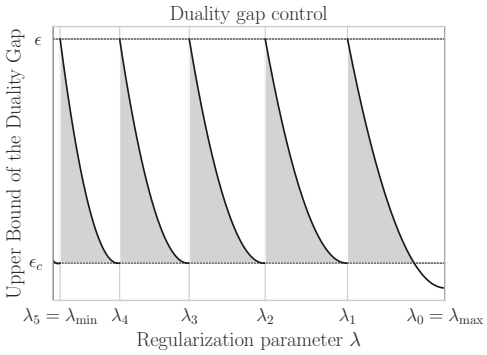


Figure 2: Illustration of the construction of an ϵ-path for the Lasso on synthetic dataset generated with `sklearn` as $X, y =$ `make_regression`$(n = 30, p = 150)$ at accuracy $\epsilon = \|y\|_2^2/40$ and $\epsilon_c = \epsilon/10$. We choose $\lambda_{\max} = \|X^\top y\|_\infty$ and $\lambda_{\min} = \lambda_{\max}/10$ leading to a path complexity $T_\epsilon = 6$. For Lasso the bound is piecewise quadratic.

3 Bounds and approximation path

Definition 2. Given a differentiable function f and $x \in \mathrm{dom}\, f$, 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 $z \in \mathrm{dom}\, f$

$$\mathcal{U}_{f,x}(z - x) \leq f(z) - f(x) - \langle \nabla f(x), z - x \rangle \ , \tag{6}$$
$$\mathcal{V}_{f,x}(z - x) \geq f(z) - f(x) - \langle \nabla f(x), z - x \rangle \ . \tag{7}$$

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:

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

where $\mathcal{U}(\cdot)$ and $\mathcal{V}(\cdot)$ are increasing from $[0, +\infty)$ to $[0, +\infty]$ vanishing at 0; see Azé and 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 and Nesterov, 2014) or (Bauschke and 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 $\forall x \in \mathrm{dom}\, f$ 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 and Tran-Dinh, 2017, Proposition

10) shows 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 for $w_\nu(\cdot)$ and $d_\nu(\cdot)$ are recalled in Appendix for logistic and quadratic losses.

Approximating the duality gap path. Assume we have constructed primal/dual feasible vectors for a finite grid of parameters $\Lambda_T = \{\lambda_0, \dots, \lambda_{T-1}\}$, *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

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

In the previous display, the $\mathcal{G}_t$ 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) hand side of Inequality (9) holds true

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

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 parameter λ when using primal/dual vector $(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ for λ_t close to λ. When the function satisfies both inequalities, the tightness of the bound can be related to the conditioning $\mathcal{U}_{f^*}/\mathcal{V}_{f^*}$ of the dual loss f^*. Equality holds for the least-squares case ($\mathcal{U}_{f^*} \equiv \mathcal{V}_{f^*} \equiv \|\cdot\|_2^2/2$), which certifies the tightness of the bounds.

From Lemma 1, we have $\mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \le \epsilon$ as soon as $Q_{t,\mathcal{V}_{f^*}}(\rho) \le \epsilon$ 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.

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

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

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) for $\lambda = \lambda_t$ to accuracy $\epsilon_c < \epsilon$
  Compute $\rho_t^\ell(\epsilon) = \max\{\rho \text{ s.t. } Q_{t,\mathcal{V}_{f^*}}(\rho) \le \epsilon\}$
  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: $\{\beta^{(\lambda_t)} \ : \ \lambda_t \in \Lambda\}$

Conversely, given a grid[4] of T parameters $\Lambda_T = \{\lambda_0, \dots, \lambda_{T-1}\}$, 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 $\lambda \mapsto \mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$:

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

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 $\epsilon_{\Lambda_T} \le \max_{t\in[T]} Q_{t,\mathcal{V}_{f^*}}(1 - \lambda_t^\star/\lambda_t)$ where for all $t \in \{0, \dots, T-1\}$, $\lambda_t^\star$ is the largest $\lambda \in [\lambda_{t+1}, \lambda_t]$ such that $Q_{t,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_t) \ge Q_{t+1,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_{t+1})$.*

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

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

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

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

This choice guarantees that the duality gap $\mathcal{G}_t$ and Δ_t converge to 0 when $\beta^{(\lambda_t)}$ converges to a solution $\hat\beta^{(\lambda_t)}$.

Finding ρ. Following Prop. 1, one needs to solve 1-dimensional equations like $Q_{t,\mathcal{V}_{f^*}}(\rho) = \epsilon$ 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):

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

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:

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

where $\delta_t := \epsilon - \mathcal{G}_t$ and $\tilde{\delta}_t := \Delta_t - \mathcal{G}_t$.

3.1 Discretization strategies

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

For regularized supervised learning methods, it is customary to start from a large regularizer[5] $\lambda_0 = \lambda_{\max}$ and then to perform iteratively 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. Generally, models are computing by increasing complexity which allows important speed-ups due to *warm start* (Friedman et al., 2007), provided that the parameters λ's are close to each other. Knowing λ_t we will provide recursive strategy constructing λ_{t+1}.

Adaptive unilateral. The strategy we call *Unilateral* consist 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. $\mathcal{G}_t \leq \epsilon_c < \epsilon$ 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 on a Lasso example. It stands as a generic algorithm for computing an approximation path for loss functions satisfying the loose regularity assumption 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 from $\lambda_{\max}$ to $\lambda_{\min}$. 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 $\lambda_{t'} \leq \lambda_t^\ell$ 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.,* $\lambda_{t'} \leq \lambda_t$. 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 $\|\nabla f(X\beta^{(\lambda_{t'})})\|_* \leq \widetilde{R}_t$, 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 $\Delta_{t'} \leq \widetilde{\Delta}_t$, where $\widetilde{\Delta}_t := \widetilde{R}_t \times \mathcal{U}_f^{-1}(\epsilon_c)$ as well as $\mathcal{G}_\lambda(\beta^{(\lambda_{t'})}, \theta^{(\lambda_{t'})}) \leq Q_{t',\mathcal{V}_{f^*}}(\rho) \leq \widetilde{Q}_{t,\mathcal{V}_{f^*}}(\rho)$, 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 $\widetilde{Q}_{t,\mathcal{V}_{f^*}}(\rho) \leq \epsilon$ 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. $\mathcal{G}_t \leq \epsilon_c < \epsilon$ for all t. Then the set $\{\beta^{(\lambda_t)} \ : \ \lambda_t \in \Lambda^{(b)}(\epsilon)\}$ is an ϵ-path for Problem (1).

Uniform unilateral and bilateral. Given the initial information from the initialization $(\beta^{(\lambda_0)}, \theta^{(\lambda_0)})$, we can build a (crude) uniform grid that guarantees an ϵ-approximation before solving any optimization problem. Indeed, by applying Lemma 3 at $t = 0$, we have $\mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leq \widetilde{Q}_{0,\mathcal{V}_{f^*}}(\rho)$. We can define $\widetilde{\rho}_0^\ell(\epsilon)$ (resp. $\widetilde{\rho}_0^r(\epsilon)$) as the largest non-negative ρ s.t. $\widetilde{Q}_{0,\mathcal{V}_{f^*}}(\rho) \leq \epsilon$ (resp. $\widetilde{Q}_{0,\mathcal{V}_{f^*}}(-\rho) \leq \epsilon$) and also

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

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*

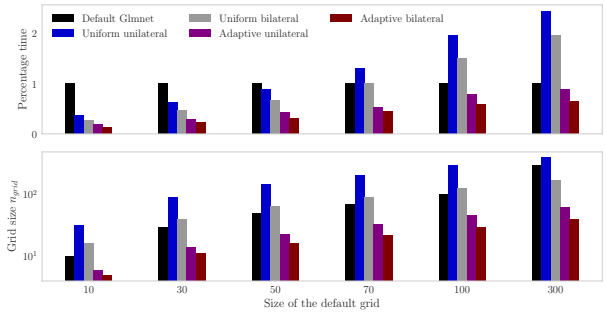
$$\lambda_0 = \lambda_{\max}, \quad \lambda_{t+1} = \lambda_t \times (1 - \rho_0(\epsilon)) \ ,$$

and $\forall t \in [T], (\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ s.t. $\mathcal{G}_t \leq \epsilon_c < \epsilon$. 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*

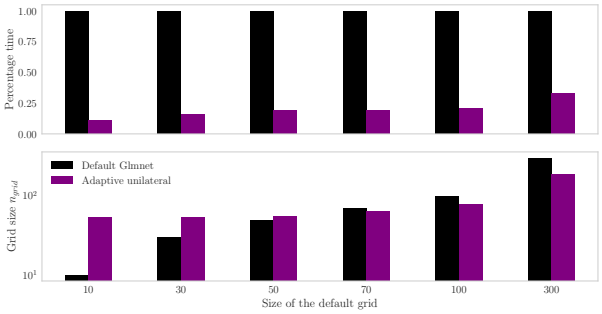
$$T_\epsilon = \left\lfloor \frac{\log(\lambda_{\max}/\lambda_{\min})}{\log(1 - \rho_0(\epsilon))} \right\rfloor \ .$$

Remark 2. For cases such as the Lasso, an explicit value of $\beta^{(\lambda_{\max})}$ can be obtained, say $\beta^{(\lambda_0)} = 0$, for $\lambda_0 = \left\|X^\top y\right\|_\infty$. Since the uniform grid depends only on this initially known value $\beta^{(\lambda_0)}$, it can be computed prior any optimization solver is launched. Hence, this case is well suited for naive parallel computation over the grid of parameters.

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



(a) ℓ_1 least-squares regression on climate data set NCEP/NCAR Reanalysis whith $n = 814$ 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 at the same error than the default grid.

3.2 Limitations of previous framework

Previous algorithms and analysis for computing an ϵ-path have been initially developed with a complexity of $O(1/\epsilon)$ (Clarkson, 2010; Giesen et al., 2010) in a large class of problem. Nevertheless, data fitting functions arising in machine earning have often nicer regularities that must be exploited. This is all the more striking in the Lasso example where a better complexity in $O(1/\sqrt{\epsilon})$ was obtained by Mairal and Yu (2012); Giesen et al. (2012).

The relation between the complexity of the path and the regularity of the objective function remains unclear and previous methods do not apply to many machine learning problem. For instance, for the logistic regression, the dual loss f^* is not uniformly smooth: so in order to apply the previous theory, one needs to restrict the solution on a (potentially badly pre-selected) compact set.

Let us consider the one dimensional toy example where $\beta \in \mathbb{R}$, $X = \mathrm{Id}$ and $y = -1$, $f(X\beta) = \log(1 + \exp(\beta))$. 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 of the dual can be reasonably estimated as $\nu_{f^*} = (1+\exp(\log(2)/\lambda))^2$ at each step. This leads to an unreasonable algorithm with tiny step sizes in Corollary 1, since for $\lambda_{\min} = \lambda_{\max}/10$ we already have $\nu_{f^*} \approx \exp(100)$. Also, note that the dual function is not polynomial; hence, the algorithm proposed by Giesen et al. (2012) can not be applied for the logistic loss.

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.

3.3 Complexity and regularity

Lower bound on the path complexity. For our method, the lower bound on the duality gap quantifies how close the proposed step in Proposition 1 is from the best possible step one can achieve for smooth loss functions. Indeed, at the 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 λ between $\lambda_{\min}$ and λ_t such that $\mathcal{U}_{f^*}(-\zeta_t \times \rho) > \epsilon$. Hence, *any* algorithm for computing an ϵ-path with the duality gap 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 (details on the constants appearing in the following result are provided in Appendix).

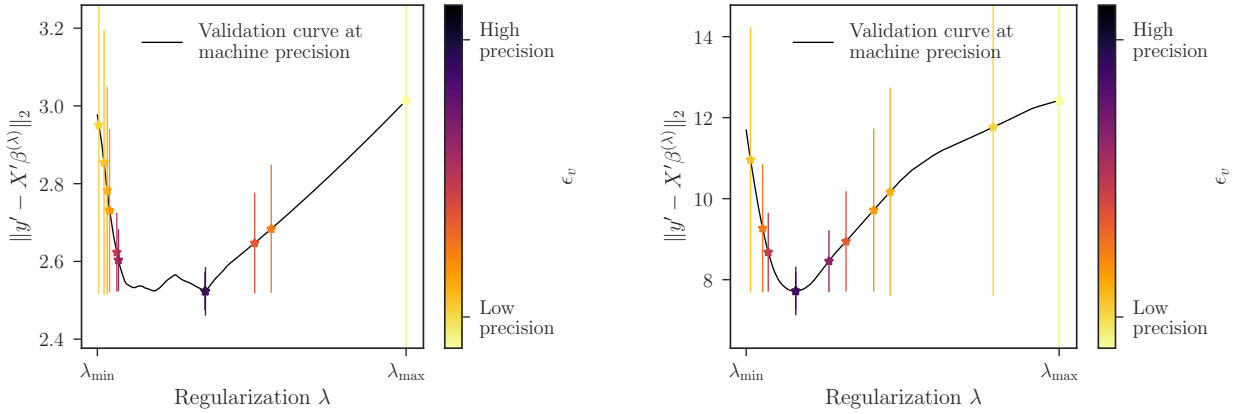
Proposition 6 (Approximation path: complexity)**.** *For $\epsilon_c < \epsilon$, there exists $C_f(\epsilon_c) > 0$ such that*

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

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

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

Moreover, $C_f(\epsilon_c)$ tends to a finite constant C_f when ϵ_c goes to 0.



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

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

Figure 4: Safe selection of the optimal hyperparameter for Elastic Net regression when the targeted accuracy ϵ_v on the validation set (30% of the observations) is refined. The range of parameters investigated is generated by Algorithm 2 (with bilateral path) between $\lambda_{\max} = \|X^\top y\|_\infty$ and $\lambda_{\min} = \lambda_{\max}/100$. Note that for loose precision, suboptimal parameter are identified, but better and better candidates are found as the accuracy ϵ_v decreases.

4 Validation path

To achieve good generalization performance, estimators defined as solution of Problem 1 require careful adjustment of the regularization parameter λ, to balance between data-fitting and regularization. A standard approach in machine learning to tune such a parameter is to select it by comparing the validation errors on a finite grid (potentially using K-fold cross-validation). Unfortunately, it is often difficult to determine a priori the number of points in the grid or how they should be distributed to achieve low validation error, and practitioners often rely on rules of thumb for that task.

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

$$E_v(\beta) = \mathcal{L}(y', X'\beta) \ . \tag{12}$$

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

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

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. However, 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 depends 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. The following conditions, on the validation loss and on the inner optimization objective, are assumed through the section:

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

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

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

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

between λ and λ_t as

$$\Delta E_v(\lambda_t, \lambda) := \left| E_v(\hat{\beta}^{(\lambda)}) - E_v(\beta^{(\lambda_t)}) \right| \ . \tag{13}$$

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 Assumption A2, any primal solution $\hat{\beta}^{(\lambda)}$ belongs to the Euclidean ball with center $\beta^{(\lambda_t)}$ and radius

$$r_{t,\mu}(\lambda) = \sqrt{\frac{2}{\mu} \mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})} \ . \tag{14}$$

Such a ball relying on a duality gap evaluation has been recently proved useful to speed-up sparse optimization solvers. Their good performance comes from their ability to iteratively identify the sparsity structure of the optimal solutions, and 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 (14) 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*

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

where $\left(\frac{x'^{\top}_{(\lfloor n\epsilon_v \rfloor + 1)} \beta^{(\lambda_t)}}{\left\| x'_{(\lfloor n\epsilon_v \rfloor + 1)} \right\|} \right)^2$ *is the* $(\lfloor n\epsilon_v \rfloor + 1)$*-th smallest value of* $\left(\frac{x_i'^{\top} \beta^{(\lambda_t)}}{\|x'_i\|} \right)^2$ *for* $i \in [n']$*. 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*

$$\lambda \in \lambda_t \times \left[1 - \rho_t^\ell(\epsilon_{v,\mu}),\ 1 + \rho_t^r(\epsilon_{v,\mu}) \right],$$

where $\rho_t^\ell(\epsilon_{v,\mu})$*,* $\rho_t^r(\epsilon_{v,\mu})$ *are defined in Proposition 1.*

Remark 3. Considering the current regularization parameter λ_t, we have

$$\Delta E_v(\lambda_t, \lambda_t) = \left| E_v(\hat{\beta}^{(\lambda_t)}) - E_v(\beta^{(\lambda_t)}) \right| \leq \epsilon_v \ ,$$

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
$\Lambda(\epsilon_{v,\mu}) =$ `training_path` $(f, \Omega, \epsilon_{v,\mu}, [\lambda_{\min}, \lambda_{\max}])$
Return: $\Lambda(\epsilon_{v,\mu})$

as soon as $\mathcal{G}_t \leq \epsilon_{v,\mu}$, which gives us a stopping criterion for solving the optimization problem on the training set (X, y) relative to the desired accuracy ϵ_v on the validation set (X', y'). This point of view has the appealing property of relieving the practitioner from selecting the stopping criterion ϵ_c while optimizing on the training set.

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

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

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

Proposition 8. *Under Assumptions A1 and A2,* $\{\beta^{(\lambda_t)}$ *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 $[\lambda_{\min}, \lambda_{\max}]$.

We first consider the default grid in `sklearn` (Pedregosa et al., 2011) and `glmnet` (Friedman et al., 2010) which is defined as $\lambda_t = \lambda_{\max} \times 10^{-\delta t/(T-1)}$ (here $\delta = 3$). 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 a smaller approximation error. Our experiments were conducted on the following datasets: `leukemia`, available in `sklearn` and the climate data `NCEP/NCAR Reanalysis` (Kalnay et al., 1996). Results are reported in Figure 3 for both classification and regression problem. Our approach leads to a better approximation of the regularization path w.r.t. the default grid, as larger steps are obtain at each λ_t. This often results in a significant gain in computating time.

The convergence of our method is illustrated for Elasatic Net in Figure 4 on synthetic databases generated with `sklearn` as a random regression problem `make_regression` and `make_sparse_uncorrelated`

(with low number of informative features) presented in Celeux et al. (2012). For a decreasing sequence of validation errors, we represent the hyperparameter selected by our algorithm and its corresponding safe interval. Note that, even if the validation curve is often non smooth and non convex, the output of the safe grid search converges to the global minimum as guaranteed in Proposition 8.

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. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

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

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

Bauschke, H. H. and Combettes, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830.

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

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

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

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

Sun, T. and Tran-Dinh, Q. (2017). Generalized self-concordant functions: A recipe for newton-type methods. *arXiv preprint arXiv:1703.04599*.

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

6 Appendix

6.1 Generalized self-concordant functions

Proposition 9 (Sun and Tran-Dinh (2017), Proposition 10)**.** *If (M_f, ν)-generalized self concordant, then*

$$w_\nu(-d_\nu(x,y)) \|y-x\|_x^2 \leq f(y) - f(x) - \langle \nabla f(x), y-x\rangle \leq w_\nu(d_\nu(x,y)) \|y-x\|_x^2 \tag{15}$$

where the right-hand side inequality holds if $d_\nu(x,y) < 1$ for the case $\nu > 2$ and where

$$d_\nu(x,y) := \begin{cases} M_f \|y-x\|_2 & \text{if } \nu = 2, \\ \left(\frac{\nu}{2} - 1\right) M_f \|y-x\|_2^{3-\nu} \|y-x\|_x^{\nu-2} & \text{if } \nu > 2, \end{cases} \tag{16}$$

and

$$w_\nu(\tau) := \begin{cases} \frac{e^\tau - \tau - 1}{\tau^2} & \text{if } \nu = 2, \\ \frac{-\tau - \log(1-\tau)}{\tau^2} & \text{if } \nu = 3, \\ \frac{(1-\tau)\log(1-\tau) + \tau}{\tau^2} & \text{if } \nu = 4, \\ \left(\frac{\nu-2}{4-\nu}\right) \frac{1}{\tau} \left[\frac{\nu-2}{2(3-\nu)\tau} \left((1-\tau)^{\frac{2(3-\nu)}{2-\nu}} - 1\right) - 1\right] & \text{otherwise.} \end{cases} \tag{17}$$

Remark 4. The dual of the logistic loss is Generalized Self-Concordant with $M_{f^*} = 1, \nu = 4$.

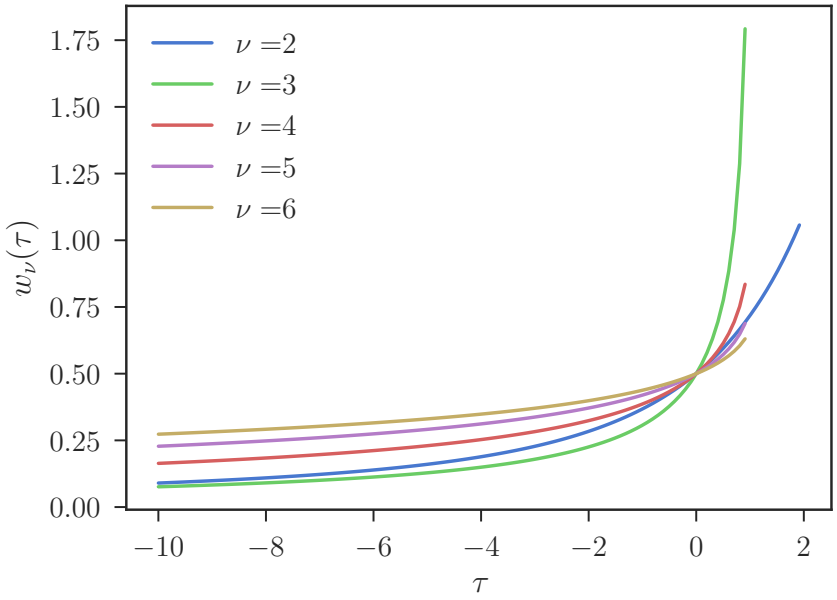


Figure 5: Illustration of the functions in self concordant bounds Equation (17)

6.2 Proof of the bounds for the warm start error and approximation path error

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), then, for $\rho = 1 - \lambda/\lambda_t$, the right (resp. left) hand side of Inequality* (18) *holds true*

$$Q_{t,\mathcal{U}_{f^*}}(\rho) \leq \mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leq Q_{t,\mathcal{V}_{f^*}}(\rho). \tag{18}$$

Proof. We recall that $\mathcal{G}_t := \mathcal{G}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)})$ and we denote for simplicity

$$\mathcal{G}_\lambda^{\lambda_t} := \mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \quad \text{and} \quad \Gamma_t := \Omega(\beta^{(\lambda_t)}) + \Omega^*(X^\top \theta^{(\lambda_t)}) \ .$$

By definition for any $(\beta, \theta) \in \operatorname{dom} P_\lambda \times \operatorname{dom} D_\lambda$ we have $\mathcal{G}_\lambda(\beta, \theta) = f(X\beta) + f^*(-\lambda\theta) + \lambda(\Omega(\beta) + \Omega^*(X^\top\theta))$, so the following holds

$$\frac{1}{\lambda_t}[\mathcal{G}_t - f(X\beta^{(\lambda_t)}) - f^*(-\lambda_t\theta^{(\lambda_t)})] = \Gamma_t \ . \tag{19}$$

Hence using Equality (19) in the definition of $\mathcal{G}_\lambda^{\lambda_t}$, we have:

$$\begin{aligned}\mathcal{G}_\lambda^{\lambda_t} &= f(X\beta^{(\lambda_t)}) + f^*(-\lambda\theta^{(\lambda_t)}) + \lambda\Gamma_t \\ &\overset{(19)}{=} \frac{\lambda}{\lambda_t}\,\mathcal{G}_t + \left(1 - \frac{\lambda}{\lambda_t}\right)[f(X\beta^{(\lambda_t)}) + f^*(-\lambda_t\theta^{(\lambda_t)})] + f^*(-\lambda\theta^{(\lambda_t)}) - f^*(-\lambda_t\theta^{(\lambda_t)}).\end{aligned}$$

Let us write the proof for the upper bound (the proof for the lower bound is similar). We apply the smoothness property and the Fenchel-Young Inequality (22) to the function $f^*(\cdot)$ with $z = -\lambda\theta^{(\lambda_t)}$ and $x = \zeta_t := -\lambda_t\theta^{(\lambda_t)}$ to obtain

$$\mathcal{G}_\lambda^{\lambda_t} \leq \frac{\lambda}{\lambda_t}\,\mathcal{G}_t + \left(1 - \frac{\lambda}{\lambda_t}\right)\Delta_t + \mathcal{V}_{f^*,\zeta_t}\left((\lambda_t - \lambda)\theta^{(\lambda_t)}\right) \ ,$$

where we have used the equality case in the Fenchel-Young Inequality (20) to get:

$$\Delta_t = f(X\beta^{(\lambda_t)}) + f^*(\zeta_t) + \langle \nabla f^*(\zeta_t), -\zeta_t\rangle = f(X\beta^{(\lambda_t)}) - f(\nabla f^*(\zeta_t)) \ .$$

We conclude by noticing that $\frac{\lambda}{\lambda_t}\,\mathcal{G}_t + \left(1 - \frac{\lambda}{\lambda_t}\right)\Delta_t = \mathcal{G}_t + \left(1 - \frac{\lambda}{\lambda_t}\right)(\Delta_t - \mathcal{G}_t)$. □

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

Proof. From the upper bound $\mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leq Q_{t,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_t)$ for all λ and λ_t, and since one can partition the parameter set as $[\lambda_{\min}, \lambda_{\max}] = \cup_{t\in[0:T-1]}[\lambda_{t+1}, \lambda_t]$, we have

$$\begin{aligned}\epsilon_{\Lambda_t} &\leq \max_{t\in[0:T-1]} \sup_{\lambda\in[\lambda_{t+1},\lambda_t]} \min_{\lambda_t\in\Lambda_T} Q_{t,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_t) \\ &\leq \max_{t\in[0:T-1]} \sup_{\lambda\in[\lambda_{t+1},\lambda_t]} \min_{t'\in\{t+1,t\}} Q_{t',\mathcal{V}_{f^*}}(1 - \lambda/\lambda_{t'}) \ .\end{aligned}$$

where the last inequality holds since $\{\lambda_{t+1}, \lambda_t\}$ is a subset of Λ_T. Let us define

$$\forall\lambda \in [\lambda_{t+1}, \lambda_t], \quad \psi_t(\lambda) := \min\{Q_{t+1,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_{t+1}), Q_{t,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_t)\} \ .$$

The quantity $Q_{t+1,\mathcal{V}_{f^*}}(1-\lambda/\lambda_{t+1})$ (resp. $Q_{t,\mathcal{V}_{f^*}}(1-\lambda/\lambda_t)$) is monotonically increasing *w.r.t.* λ (resp. decreasing), so $\sup_{\lambda\in[\lambda_{t+1},\lambda_t]}\psi_t(\lambda)$ is reached for the largest λ satisfying

$$Q_{t,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_t) \geq Q_{t+1,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_{t+1}) \ .$$

□

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 $\tilde{\delta}_t := \Delta_t - \mathcal{G}_t$.

Proof. If f^* is $\frac{\nu}{2}\|\cdot\|^2$-smooth (which is equivalent to f is $\frac{1}{2\nu}\|\cdot\|^2$-strongly convex), we have from Lemma 1

$$\mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leq Q_{t,\mathcal{V}_{f^*}}(\rho) = \mathcal{G}_t + \rho(\Delta_t - \mathcal{G}_t) + \frac{\nu\rho^2}{2}\left\|\zeta_t\right\|^2 \ .$$

Hence we conclude by solving in ρ the inequality $Q_{t,\mathcal{V}_{f^*}}(\rho) \leq \epsilon$. □

6.3 Useful convexity inequalities

Lemma 5 (Fenchel-Young inequalities)**.** Let f be a continuously differentiable function. For all x, x^*, we have

$$f(x) + f^*(x^*) \geq \langle x^*, x \rangle, \tag{20}$$

with equalities if and only if $x^* = \nabla f(x)$ (or equivalently $x = \nabla f^*(x^*)$). Moreover, if f is $\mathcal{U}_{f,x}$-convex (resp. $\mathcal{V}_{f,x}$-smooth) Inequality (21) (resp. Inequality (22)) holds true:

$$f(x) + f^*(x^*) \geq \langle x^*, x \rangle + \mathcal{U}_{f,x}(x - \nabla f^*(x^*)), \tag{21}$$
$$f(x) + f^*(x^*) \leq \langle x^*, x \rangle + \mathcal{V}_{f,x}(x - \nabla f^*(x^*)). \tag{22}$$

Proof. We have from the $\mathcal{U}_{f,x}$-convexity and the equality $f(z) + f^*(\nabla f(z)) = \langle \nabla f(z), z \rangle$

$$-f^*(\nabla f(z)) + \langle \nabla f(z), x \rangle + \mathcal{U}_{f,x}(x - z) = f(z) + \langle \nabla f(z), x - z \rangle + \mathcal{U}_{f,x}(x - z) \leq f(x).$$

We conclude by applying the inequality at $z = \nabla f^*(x^*)$ and remark that $\nabla f(z) = x^*$. The same proof holds for the upper bound (22). □

Applying Fenchel-Young Inequalities (21) and (22) give the following bounds.

Lemma 6. We assume that $-\lambda\theta \in \mathrm{Dom}(f^*)$ and $X^\top\theta \in \mathrm{Dom}(\Omega^*)$. Then, the Inequality (23) (resp. (23)) provided that f is $\mathcal{U}_f$-convex (resp. $\mathcal{V}_f$-smooth).

$$\lambda\widetilde{\Omega}(\beta, \theta) + \mathcal{U}_f(X\beta - \nabla f^*(-\lambda\theta)) \leq \mathcal{G}_\lambda(\beta, \theta) \tag{23}$$
$$\lambda\widetilde{\Omega}(\beta, \theta) + \mathcal{V}_f(X\beta - \nabla f^*(-\lambda\theta)) \geq \mathcal{G}_\lambda(\beta, \theta) \tag{24}$$

where $\widetilde{\Omega}(\beta, \theta) = \Omega(\beta) + \Omega^*(X^\top\theta) + \langle \beta, -X^\top\theta \rangle$.

Proof. We apply the Fenchel-Young Inequality (21) to obtain

$$\begin{aligned}
\mathcal{G}_\lambda(\beta, \theta) =& f(X\beta) + f^*(-\lambda\theta) + \lambda(\Omega(\beta) + \Omega^*(X^\top\theta)) \\
\geq& \langle X\beta, -\lambda\theta \rangle + \mathcal{U}_f(X\beta - \nabla f^*(-\lambda\theta)) + \lambda(\Omega(\beta) + \Omega^*(X^\top\theta)) \\
=& \mathcal{U}_f(X\beta - \nabla f^*(-\lambda\theta)) + \lambda\left(\Omega(\beta) + \Omega^*(X^\top\theta) + \langle \beta, -X^\top\theta \rangle\right).
\end{aligned}$$

The same technique applies for the upper bound with the Fenchel-Young Inequality (22). □

Remark 5. From the Fenchel-Young Inequality (20), we have $\Omega(\beta) + \Omega^*(X^\top\theta) \geq \langle \beta, X^\top\theta \rangle$, so the lower bound is always non negative.

Variation of the loss function along the path

Lemma 7. Let $\beta^{(\lambda_t)}$ (resp. $\beta^{(\lambda_{t'})}$) be an ϵ-solution at parameter λ_t (resp. $\lambda_{t'}$), then we have

$$\left(1 - \frac{\lambda_{t'}}{\lambda_t}\right)\left(f(X\beta^{(\lambda_{t'})}) - f(X\beta^{(\lambda_t)})\right) \leq \mathcal{G}_{t'} + \frac{\lambda_{t'}}{\lambda_t}\mathcal{G}_t\,.$$

where $\mathcal{G}_s := \mathcal{G}_{\lambda_s}(\beta^{(\lambda_s)}, \theta^{(\lambda_s)})$ for $s \in \{t, t'\}$. Moreover, the mapping $\lambda \mapsto f(X\hat{\beta}^{(\lambda)})$ is non-increasing.

Proof. Since $\hat{\beta}^{(\lambda)}$ is optimal at parameter λ, we have:

$$f(X\beta^{(\lambda)}) + \lambda\Omega(\beta^{(\lambda)}) - \epsilon \leq f(X\hat{\beta}^{(\lambda)}) + \lambda\Omega(\hat{\beta}^{(\lambda)}) \leq f(X\beta^{(\lambda_t)}) + \lambda\Omega(\beta^{(\lambda_t)})\ .$$

Moreover,

$$\begin{aligned}
f(X\beta^{(\lambda_t)}) + \lambda\Omega(\beta^{(\lambda_t)}) &= \frac{\lambda}{\lambda_t}\left(f(X\beta^{(\lambda_t)}) + \lambda_t\Omega(\beta^{(\lambda_t)})\right) + \left(1 - \frac{\lambda}{\lambda_t}\right)f(X\beta^{(\lambda_t)}) \\
&\leq \frac{\lambda}{\lambda_t}\left(f(X\hat{\beta}^{(\lambda_t)}) + \lambda_t\Omega(\hat{\beta}^{(\lambda_t)}) + \epsilon\right) + \left(1 - \frac{\lambda}{\lambda_t}\right)f(X\beta^{(\lambda_t)}) \\
&\leq \frac{\lambda}{\lambda_t}\left(f(X\beta^{(\lambda)}) + \lambda_t\Omega(\beta^{(\lambda)}) + \epsilon\right) + \left(1 - \frac{\lambda}{\lambda_t}\right)f(X\beta^{(\lambda_t)})\ .
\end{aligned}$$

The last inequality comes from the optimality of $\hat{\beta}^{(\lambda_t)}$ at parameter λ_t. Hence,

$$f(X\beta^{(\lambda)}) + \lambda\Omega(\beta^{(\lambda)}) - \epsilon \leq \frac{\lambda}{\lambda_t}\left(f(X\beta^{(\lambda)}) + \lambda_t\Omega(\beta^{(\lambda)}) + \epsilon\right) + \left(1 - \frac{\lambda}{\lambda_t}\right) f(X\beta^{(\lambda_t)}) \ .$$

At optimality, $\epsilon = 0$ and we can deduce that $\left(1 - \frac{\lambda}{\lambda_t}\right) f(X\hat{\beta}^{(\lambda)}) \leq \left(1 - \frac{\lambda}{\lambda_t}\right) f(X\hat{\beta}^{(\lambda_t)})$, hence the second result. □

Bounding the gradient along the path

We can furthermore bound the norm of the gradient of the loss when the parameter λ varies.

Lemma 8. For $x \in \mathrm{Dom}(f)$, if f is $\mathcal{V}_{f,x}$-smooth, then writing $\mathcal{V}_{f,x}^* = (\mathcal{V}_{f,x})^*$ for the Fenchel-Legendre transform, one has

$$\mathcal{V}_{f,x}^*(-\nabla f(x)) \leq f(x) - \inf_z f(z) \ .$$

Proof. From the smoothness of f, we have

$$\inf_z f(z) \leq \inf_z \left(f(x) + \langle \nabla f(x), z - x\rangle + \mathcal{V}_{f,x}(z - x)\right) = f(x) - (\mathcal{V}_{f,x})^*(-\nabla f(x)) \ .$$

□

A direct application of Lemma 8 and Lemma 7 yields:

Lemma 9. Assume that f is uniformly smooth and let $\beta^{(\lambda_{t'})}$ (resp. $\beta^{(\lambda_t)}$) be an ϵ-solution at parameter $\lambda_{t'}$ (resp. λ_t). Then for $\delta_\epsilon(\lambda_{t'}, \lambda_t) := \frac{\lambda_t + \lambda_{t'}}{\lambda_t - \lambda_{t'}}\ \epsilon$, we have

$$\mathcal{V}_f^*(-\nabla f(X\beta^{(\lambda_{t'})})) \leq f(X\beta^{(\lambda_t)}) + \delta_\epsilon(\lambda_{t'}, \lambda_t) \ .$$

At optimality $\epsilon = 0$ and so $\delta_\epsilon(\lambda_{t'}, \lambda_t) = 0$ and we have

$$\mathcal{V}_f^*(-\nabla f(X\hat{\beta}^{(\lambda_{t'})})) \leq f(X\hat{\beta}^{(\lambda_t)}).$$

Lemma 10. Assuming f is uniformly smooth and $\rho_t^\ell(\epsilon) = 1 - \frac{\lambda_{t'}}{\lambda_t}$, we have $\|\nabla f(X\beta^{(\lambda_{t'})})\| \leq \widetilde{R}_t$. If in addition, f is uniformly convex, we have $\Delta_{t'} \leq \widetilde{\Delta}_t$.

Proof. If f is uniformly smooth, from Lemma 9, we have:

$$\begin{aligned}
\mathcal{V}_f^*(-\nabla f(X\beta^{(\lambda_{t'})})) &\leq f(X\beta^{(\lambda_t)}) + \delta_\epsilon(\lambda_{t'}, \lambda_t) \\
\|\nabla f(X\beta^{(\lambda_{t'})})\|_* &\leq \mathcal{V}_f^{*-1}\left(f(X\beta^{(\lambda_t)}) + \frac{2\epsilon}{\rho_t^\ell(\epsilon)}\right) =: \widetilde{R}_t
\end{aligned}$$

where the first line follows from Lemma 9 and the second follows from the fact that for $\mathcal{V}_f = \mathcal{V} \circ \|\cdot\|$, we have $\mathcal{V}_f^* := (\mathcal{V}_f)^* = \mathcal{V}^* \circ \|\cdot\|_*$ and since $\lambda_{t'} \leq \lambda_t$, then $\delta_\epsilon(\lambda_{t'}, \lambda_t) \leq 2\epsilon\lambda_t/(\lambda_t - \lambda_{t'}) = 2\epsilon/\rho_t^\ell(\epsilon)$.

Since f is convex, we have

$$\begin{aligned}
\Delta_t &:= f(X\beta^{(\lambda_t)}) - f(\nabla f^*(-\lambda_t\theta^{(\lambda_t)})) \leq -\langle \nabla f(X\beta^{(\lambda_t)}), \nabla f^*(-\lambda_t\theta^{(\lambda_t)}) - X\beta^{(\lambda_t)}\rangle \\
&\leq \|\nabla f(X\beta^{(\lambda_t)})\|_* \|\nabla f^*(-\lambda_t\theta^{(\lambda_t)}) - X\beta^{(\lambda_t)}\| \\
&\leq \|\nabla f(X\beta^{(\lambda_t)})\|_* \times \mathcal{U}_f^{-1}(\mathcal{G}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}))
\end{aligned}$$

where the two last inequalities comes from Holder inequality and Lemma 6. □

6.4 Proof of the complexity bound

We denote T_ϵ the cardinality of the grid returned by Algorithm 1 and let $(\rho_t)_{t\in[0:T_\epsilon-1]}$ be the set of step size needed to cover the interval $[\lambda_{\min}, \lambda_{\max}]$. Using $\rho_t = 1 - \frac{\lambda_{t+1}}{\lambda_t}$, we have

$$\log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) = \log\left(\prod_{t=0}^{T_\epsilon-1} \frac{\lambda_t}{\lambda_{t+1}}\right) = \sum_{t=0}^{T_\epsilon-1} \log\left(\frac{1}{1-\rho_t}\right) \ .$$

Hence, denoting $\rho_{\min}(\epsilon) = \min_{t\in[0:T_\epsilon-1]} \rho_t$, we have

$$T_\epsilon \times \rho_{\min}(\epsilon) \leq \log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) \ . \tag{25}$$

Moreover, to simplify our analysis we will suppose that at each step λ_t, we have solved the optimization problem with two measures of accuracy $\mathcal{G}_t \leq \epsilon_c$ and $\Delta_t \leq \epsilon_c$ for $\epsilon_c < \epsilon$. Also, we assume that we explore the parameter range in decreasing order. Then we recall from Lemma 1 that $\mathcal{G}_\lambda(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leq Q_{t,\mathcal{V}_{f^*}}(\rho)$ which is smaller than ϵ as soon as $\mathcal{V}_{f^*,\zeta_t}(-\zeta_t\cdot\rho) \leq \epsilon - \epsilon_c$. Since $\rho_{\min}(\epsilon) = \min_{t\in[0:T_\epsilon-1]} \rho_t = \min_{t\in[0:T_\epsilon-1]} \sup\{\rho \ : \ Q_{t,\mathcal{V}_{f^*}}(\rho) \leq \epsilon\}$, then

$$\rho_{\min}(\epsilon) \geq \min_{t\in[0:T_\epsilon-1]} \sup\{\rho \ : \ \mathcal{V}_{f^*,\zeta_t}(-\zeta_t\cdot\rho) \leq \epsilon - \epsilon_c\} \ . \tag{26}$$

Hence the complexity of the path is bounded as follows.

Proposition 6 (Complexity of the approximation path)**.** *For $\epsilon_c < \epsilon$, there exists $C_f(\epsilon_c) > 0$ such that*

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

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

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

Moreover, $C_f(\epsilon_c)$ tends to a finite constant C_f when ϵ_c goes to 0.

Proof. In the uniformly convex case, $\mathcal{V}_{f^*,\zeta_t}(-\zeta_t\cdot\rho) = \mathcal{V}_{f^*}(\rho\left\|\zeta_t\right\|_*)$, hence we can deduce from Equation (25) and (26) that

$$T_\epsilon \leq \frac{1}{\rho_{\min}(\epsilon)} \times \log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) \leq \log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) \times \frac{\max_{t\in[0:N_\epsilon-1]} \left\|\zeta_t\right\|_*}{\mathcal{V}_{f^*}^{-1}(\epsilon - \epsilon_c)} \ ,$$

so we just need to uniformly bound $\|\zeta_s\|$. By construction of the dual point Lemma 2, we have:

$$\left\|\zeta_t\right\|_* = \frac{\lambda_t}{\max(\lambda_t, \sigma^\circ_{\operatorname{dom}\Omega^*}(X^\top\nabla f(X\beta^{(\lambda_t)}))} \left\|\nabla f(X\beta^{\lambda_t})\right\|_* \leq \left\|\nabla f(X\beta^{\lambda_t})\right\|_* \leq \widetilde{R}_0 \ , \tag{27}$$

where the last inequality comes from Lemma 10.

For the Generalized Self-Concordant case, we first recall that the functions $w_\nu(\cdot)$ in Equation (17) are increasing and $w_\nu(0) = 1/2$. Then there exists a positive constant a_ν such that $w_\nu(\tau) \leq 1$ for $\tau \in [0, a_\nu]$ (in fact $a_\nu = 1$ for the logistic regression). Thus, provided $\rho d_\nu(\zeta_t) \leq a_\nu$, we can derive the bound $\mathcal{V}_{f^*}(-\zeta_t \times \rho) \leq \rho^2 \left\|\zeta_t\right\|^2_{\zeta_t}$.

Like in the uniformly convex case, in order to get the complexity of the ϵ-path, we also need a uniform bound on $\left\|\zeta_t\right\|_{\zeta_t}$. By taking (6) on f^* with $x = \zeta_t$ and $z = 0$, we obtain

$$\begin{aligned} w_\nu(-d_\nu(\zeta_t))\left\|\zeta_t\right\|^2_{\zeta_t} = \mathcal{U}_{f^*,\zeta_t}(-\zeta_t) &\leq f^*(0) - f^*(\zeta_t) - \langle\nabla f^*(\zeta_t), -\zeta_t\rangle = f(\nabla f^*(\zeta_t)) = f(X\beta^{(\lambda_t)}) - \Delta_t \\ &\leq f(X\beta^{(\lambda_t)}) + \epsilon_c \leq f(X\beta^{(\lambda_0)}) + \frac{2\epsilon_c}{\rho_0^\ell(\epsilon)} + \epsilon_c =: \bar{R}_0 \end{aligned}$$

where we used the inequality case of Fenchel-Young Inequality and the fact that $f^*(0) = -\inf f = 0$. This shows that $\psi(\zeta_t) := \mathcal{U}_{f^*,\zeta_t}(-\zeta_t) \leq \bar{R}_0$. Since the function ψ is continuous, then its level set is closed *i.e.,* $\{z \in \mathbb{R}^n : \psi(z) \leq \bar{R}_0\}$ is closed. Recalling Equation (27), we have $\|\zeta_t\|_* \leq \widetilde{R}_0$. Then we have

$$d_\nu(\zeta_t) \leq sup_{z\in\mathcal{H}_f} d_\nu(z) =: B_f \text{ where } \mathcal{H}_f := \{z \in \mathbb{R}^n : \psi(z) \leq \bar{R}_0\} \cap \{z : \|z\|_* \leq \widetilde{R}_0\} \text{ is a compact set.}$$

Since the function $w_\nu(\cdot)$ is increasing, we have $w_\nu(-B_f)\|\zeta_t\|^2_{\zeta_t} \leq w_\nu(-d_\nu(\zeta_t))\|\zeta_t\|^2_{\zeta_t} \leq \bar{R}_0$ which implies that $\|\zeta_t\|^2_{\zeta_t} \leq \frac{\bar{R}_0}{w_\nu(-B_f)}$. Thus, provided $\rho d_\nu(\zeta_t) \leq \bar{a}_\nu$, we can derive the bound $\mathcal{V}_{f^*}(\zeta_t \times \rho) \leq \rho^2 \|\zeta_t\|^2_{\zeta_t}$. Whence $\rho_{\min}(\epsilon) \geq \min_t \frac{\sqrt{\epsilon-\epsilon_0}}{\|\zeta_t\|_{\zeta_t}}$. Hence the complexity is bounded as $T_\epsilon \leq \log\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) \frac{\sqrt{\bar{R}_0/w_\nu(-B_f)}}{\sqrt{\epsilon-\epsilon_c}}$. □

6.5 Proof of the validation error bounds

Proposition 7 (Grid for a prescribed validation error)**.** *Suppose that we have solved problem* (1) *for a parameter* λ_t *up to accuracy* $\mathcal{G}_{\lambda_t}(\beta^{(\lambda_t)}, \theta^{(\lambda_t)}) \leq \xi(\epsilon_v, \mu, X')$*, then we have* $\Delta E_v(\lambda_t, \lambda) \leq \epsilon_v$ *for all*

$$\lambda \in \lambda_t \times \left[1 - \rho^\ell_t(\xi(\epsilon_v, \mu, X')),\, 1 + \rho^r_t(\xi(\epsilon_v, \mu, X'))\right]$$

where $\rho^\ell_t(\epsilon)$ *and* $\rho^r_t(\epsilon)$ *for* $\epsilon > 0$ *are defined in Proposition 1.*

Proof. We distinguish the two cases of interest: classification and regression.

- Case where the loss function is a norm:
  we have
  $$\max_{\beta\in\mathcal{B}(\beta^{(\lambda_t)},r)} \mathcal{L}(X'\beta, X'\beta^{(\lambda_t)}) = \max_{\beta\in\mathcal{B}(\beta^{(\lambda_t)},r)} \|X'(\beta - \beta^{(\lambda_t)})\| \leq r_{\lambda,\mu}\|X'\|$$
  where $r_{\lambda,\mu}$ is the duality gap safe radius defined in Equation (14). Hence by using the bounds on the duality gap in Lemma 1, we can ensure $\Delta E_v(\lambda_t, \lambda) \leq \epsilon_v$ for all $\rho = 1 - \lambda/\lambda_t$ such that $Q_{t,\mathcal{V}_{f^*}}(\rho) \leq \frac{\mu\epsilon_v^2}{2\|X'\|^2}$.
- Case where the loss function is the indicator function:
  using the inequality $-2ab \leq (a-b)^2 - b^2$ for $a = {x'}_i^\top\beta$ and $b = {x'}_i^\top\beta^{(\lambda_t)}$ and $|{x'}_i^\top(\beta - \beta^{(\lambda_t)})| \leq r\|{x'}_i\|$ for all $\beta \in \mathcal{B}(\beta^{(\lambda_t)}, r)$ we have:
  $$-2({x'}_i^\top\beta)({x'}_i^\top\beta^{(\lambda_t)}) \leq (r\|{x'}_i\|)^2 - ({x'}_i^\top\beta^{(\lambda_t)})^2.$$
  Hence we obtain the following upper bound
  $$\max_{\beta\in\mathcal{B}(\beta^{(\lambda_t)},r)} \mathcal{L}(X'\beta, X'\beta^{(\lambda_t)}) = \max_{\beta\in\mathcal{B}(\beta^{(\lambda_t)},r)} \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{({x'}_i^\top\beta^{(\lambda_t)})({x'}_i^\top\beta)<0} \leq \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{|{x'}_i^\top\beta^{(\lambda_t)}|\leq r\|{x'}_i\|}.$$
  By using the bound on the duality gap, we can ensure $\Delta E_v(\lambda_0, \lambda) \leq \epsilon_v$ for all λ such that:
  $$\#\left\{i \in [n] : \xi_i := \frac{\mu}{2}\left(\frac{{x'}_i^\top\beta^{(\lambda_t)}}{\|x'_i\|}\right)^2 \leq Q_{t,\mathcal{V}_{f^*}}(1 - \lambda/\lambda_t)\right\} \leq \lfloor n\epsilon_v \rfloor.$$
  By denoting $\left(\xi_{(i)}\right)_{i\in[n]}$ the (increasing) ordered sequence, we need the inequality to be true for at most the $\lfloor n\epsilon_v \rfloor$ first values *i.e.,* we choose λ such that:
  $$Q_{t,\mathcal{V}_{f^*}}\left(1 - \frac{\lambda}{\lambda_t}\right) < \frac{\mu}{2}\left(\frac{{x'}^\top_{(\lfloor n\epsilon_v\rfloor+1)}\beta^{(\lambda_t)}}{\left\|{x'}_{(\lfloor n\epsilon_v\rfloor+1)}\right\|}\right)^2 .$$

□