Computing Full Conformal Prediction Set with Approximate Homotopy

Eugene Ndiaye
RIKEN Center for Advanced Intelligence Project
eugene.ndiaye@riken.jp

Ichiro Takeuchi
Nagoya Institute of Technology
takeuchi.ichiro@nitech.ac.jp

Abstract

If you are predicting the label y of a new object with $\hat{y}$, how confident are you that $y = \hat{y}$? Conformal prediction methods provide an elegant framework for answering such question by building a $100(1-\alpha)\%$ confidence region without assumptions on the distribution of the data. It is based on a refitting procedure that parses all the possibilities for y to select the most likely ones. Although providing strong coverage guarantees, conformal set is impractical to compute exactly for many regression problems. We propose efficient algorithms to compute conformal prediction set using approximated solution of (convex) regularized empirical risk minimization. Our approaches rely on a new homotopy continuation technique for tracking the solution path with respect to sequential changes of the observations. We also provide a detailed analysis quantifying its complexity.

1 Introduction

In many practical applications of regression models it is beneficial to provide, not only a point-prediction, but also a prediction set that has some desired coverage property. This is especially true when a critical decision is being made based on the prediction, e.g., in medical diagnosis or experimental design. *Conformal prediction* is a general framework for constructing non-asymptotic and distribution-free prediction sets. Since the seminal work of [27, 23], the statistical properties and computational algorithms for conformal prediction have been developed for a variety of machine learning problems such as density estimation, clustering, and regression - see the review of [3].

Let $\mathcal{D}_n = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ be a sequence of features and labels of random variables in $\mathbb{R}^p \times \mathbb{R}$ from a distribution $\mathbb{P}$. Based on observed data $\mathcal{D}_n$ and a new test instance x_{n+1} in $\mathbb{R}^p$, the goal of conformal prediction is to build a $100(1-\alpha)\%$ confidence set that contains the unobserved variable y_{n+1} for α in $(0, 1)$, without any specific assumptions on the distribution $\mathbb{P}$.

The conformal prediction set for y_{n+1} is defined as the set of $z \in \mathbb{R}$ whose *typicalness* is sufficiently large. The typicalness of each z is defined based on the residuals of the regression model, trained with an augmented training set $\mathcal{D}_{n+1}(z) = \mathcal{D}_n \cup (x_{n+1}, z)$. On average, prediction sets constructed within a conformal prediction framework are shown to have a desirable coverage property, as long as the training instances $\{(x_i, y_i)\}_{i=1}^{n+1}$ are exchangeable, and the regression estimator is symmetric with respect to the training instances (even when the model is not correctly specified).

Despite these attractive properties, the computation of conformal prediction sets has been intractable since one needs to fit infinitely many regression models with an augmented training set $\mathcal{D}_{n+1}(z)$, for all possible $z \in \mathbb{R}$. Except for simple regression estimators with quadratic loss (such as least-square regression, ridge regression or lasso estimators) where an explicit and exact solution of the model parameter can be written as a piece of a linear function in the observation vectors, the computation of the full and exact conformal set for the general regression problem is challenging and still open.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Contributions. We propose a general method to compute the full conformal prediction set for a wider class of regression estimators. The main novelties are summarized in the following points:

- We introduce a new homotopy continuation technique, inspired by [9, 18], which can efficiently update an approximate solution with tolerance $\epsilon > 0$, when the data are streamed sequentially. For this, we show that the variation of the optimization error only depends on the loss on the new input data. Thus, exploiting the regularity of the loss, we can provide a range of observations for which an approximate solution is still valid. This allows us to approximately fit infinitely many regression models for all possible z in a pre-selected range $[y_{\min}, y_{\max}]$, using only a finite number of candidate z. For example, when the loss function is smooth, the number of model fittings required for constructing the prediction set is $O(1/\sqrt{\epsilon})$.
- Exploiting the approximation error bounds of the proposed homotopy continuation method, we can construct the prediction set based on the ϵ-solution, which satisfies the same valid coverage properties under the same mild assumptions as the conformal prediction framework. When the approximation tolerance ϵ decreases to 0, the prediction set converges to the *exact* conformal prediction set which would be obtained by fitting an infinitely large number of regression models. Furthermore, if the loss function of the regression estimator is smooth and some other regularity conditions are satisfied, the prediction set constructed by the proposed method is shown to contain the *exact* conformal prediction set.

For reproducibility, our implementation is available in

`https://github.com/EugeneNdiaye/homotopy_conformal_prediction`

Notation. For a non zero integer n, we denote $[n]$ to be the set $\{1, \cdots, n\}$. The dataset of size n is denoted $\mathcal{D}_n = (x_i, y_i)_{i\in[n]}$, the row-wise feature matrix $X = [x_1, \cdots, x_{n+1}]^\top$, and $X_{[n]}$ is its restriction to the n first rows. Given a proper, closed and convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, we denote $\mathrm{dom} f = \{x \in \mathbb{R}^n : f(x) < +\infty\}$. Its Fenchel-Legendre transform is $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined by $f^*(x^*) = \sup_{x\in\mathrm{dom} f} \langle x^*, x\rangle - f(x)$. The smallest integer larger than a real value r is denoted $\lceil r \rceil$. We denote by $Q_{1-\alpha}$, the $(1-\alpha)$-quantile of a real valued sequence $(U_i)_{i\in[n+1]}$, defined as the variable $Q_{1-\alpha} = U_{(\lceil (n+1)(1-\alpha) \rceil)}$, where $U_{(i)}$ are the i-th order statistics. For j in $[n+1]$, the rank of U_j among $U_1, \cdots, U_{n+1}$ is defined as $\mathrm{Rank}(U_j) = \sum_{i=1}^{n+1} \mathbb{1}_{U_i \le U_j}$. The interval $[a-\tau, a+\tau]$ will be denoted $[a \pm \tau]$.

2 Background and Problem Setup

We consider the framework of regularized empirical risk minimization (see for instance [24]) with a convex loss function $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$, a convex regularizer $\Omega : \mathbb{R} \mapsto \mathbb{R}$ and a positive scalar λ:

$$\hat{\beta} \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} P(\beta) := \sum_{i=1}^{n} \ell(y_i, x_i^\top \beta) + \lambda\Omega(\beta) \ . \tag{1}$$

For simplicity, we will assume that for any real values z and z_0, we have $\ell(z_0, z)$ and $\ell(z, z_0)$ are non negative, $\ell(z_0, z_0)$ and $\ell^*(z_0, 0)$ are equal to zero. These assumptions are easy to satisfy and we refer the reader to the appendix for more details.

Examples. A popular example of a loss function found in the literature is `power norm regression`, where $\ell(a, b) = |a - b|^q$. When $q = 2$, this corresponds to classical linear regression. Cases where $q \in [1, 2)$ are common in robust statistics. In particular, $q = 1$ is known as least absolute deviation. The `logcosh` loss $\ell(a, b) = \gamma \log(\cosh(a - b)/\gamma)$ is a differentiable alternative to the ℓ_∞ norm (Chebychev approximation). One can also have the `Linex` loss function [10, 5] which provides an asymmetric loss $\ell(a, b) = \exp(\gamma(a - b)) - \gamma(a - b) - 1$, for $\gamma \neq 0$. Any convex regularization functions Ω *e.g.* Ridge [12] or sparsity inducing norm [2] can be considered.

For a new test instance x_{n+1}, the goal is to construct a prediction set $\hat{\Gamma}^{(\alpha)}(x_{n+1})$ for y_{n+1} such that

$$\mathbb{P}^{n+1}(y_{n+1} \in \hat{\Gamma}^{(\alpha)}(x_{n+1})) \ge 1 - \alpha \ \text{ for } \alpha \in (0, 1) \ . \tag{2}$$

2.1 Conformal Prediction

Conformal prediction [27] is a general framework for constructing confidence sets, with the remarkable properties of being distribution free, having a finite sample coverage guarantee, and being able to be adapted to any estimator under mild assumptions. We recall the arguments in [23, 16].

Let us introduce the extension of the optimization problem (1) with augmented training data $\mathcal{D}_{n+1}(z) := \mathcal{D}_n \cup \{(x_{n+1}, z)\}$ for $z \in \mathbb{R}$:

$$\hat{\beta}(z) \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} P_z(\beta) := \sum_{i=1}^{n} \ell(y_i, x_i^\top \beta) + \ell(z, x_{n+1}^\top \beta) + \lambda \Omega(\beta) \ . \tag{3}$$

Then, for any z in $\mathbb{R}$, we define the conformity measure for $\mathcal{D}_{n+1}(z)$ as

$$\forall i \in [n],\ \hat{R}_i(z) = \psi(y_i, x_i^\top \hat{\beta}(z)) \text{ and } \hat{R}_{n+1}(z) = \psi(z, x_{n+1}^\top \hat{\beta}(z)) \ , \tag{4}$$

where ψ is a real-valued function that is invariant with respect to any permutation of the input data. For example, in a linear regression problem, one can take the absolute value of the residual to be a conformity measure function *i.e.* $\hat{R}_i(z) = |y_i - x_i^\top \hat{\beta}(z)|$.

The main idea for constructing a conformal confidence set is to consider the *typicalness* of a candidate point z measured as

$$\hat{\pi}(z) = \hat{\pi}(\mathcal{D}_{n+1}(z)) := 1 - \frac{1}{n+1} \mathrm{Rank}(\hat{R}_{n+1}(z)) \ . \tag{5}$$

If the sequence $(x_i, y_i)_{i \in [n+1]}$ is exchangeable and identically distributed, then $(\hat{R}_i(y_{n+1}))_{i \in [n+1]}$ is also , by the invariance of $\hat{R}$ *w.r.t.* permutations of the data. Since the rank of one variable among an exchangeable and identically distributed sequence is (sub)-uniformly distributed (see [4]) in $\{1, \cdots, n+1\}$, we have $\mathbb{P}^{n+1}(\hat{\pi}(y_{n+1}) \leq \alpha) \leq \alpha$ for any α in $(0, 1)$. This implies that the function $\hat{\pi}$ takes a small value on atypical data. Classical statistics for hypothesis testing, such as a p-value function, satisfy such a condition under the null hypothesis (see [14, Lemma 3.3.1]). In particular, this implies that the desired coverage guarantee in Equation (2) is verified by the conformal set defined as

$$\hat{\Gamma}^{(\alpha)}(x_{n+1}) := \{z \in \mathbb{R} : \hat{\pi}(z) > \alpha\} \ . \tag{6}$$

The conformal set gathers the real value z such that $\hat{\pi}(z) > \alpha$, if and only if $\hat{R}_{n+1}(z)$ is ranked no higher than $\lceil (n+1)(1-\alpha) \rceil$, among $\hat{R}_i(z)$ for all i in $[n]$. For regression problems where y_{n+1} lies in a subset of $\mathbb{R}$, obtaining the conformal set $\hat{\Gamma}^{(\alpha)}(x_{n+1})$ in Equation (6) is computationally challenging. It requires re-fitting the prediction model $\hat{\beta}(z)$ for infinitely many candidates z in $\mathbb{R}$ in order to compute a conformity measure such as $\hat{R}_i(z) = |y_i - x_i^\top \hat{\beta}(z)|$.

Existing Approaches for Computing a Conformal Prediction Set. In Ridge regression, for any x in $\mathbb{R}^p$, $z \mapsto x^\top \hat{\beta}(z)$ is a linear function of z, implying that $\hat{R}_i(z)$ is piecewise linear. Exploiting this fact, an exact conformal set $\hat{\Gamma}^{(\alpha)}(x_{n+1})$ for Ridge regression was efficiently constructed in [20]. Similarly, using the piecewise linearity in z of the Lasso solution, [15] proposed a piecewise linear homotopy under mild assumptions, when a single input sample point is perturbed. Apart from these cases of quadratic loss with Ridge and Lasso regularization, where an explicit formula of the estimator is available, computing such a set is often infeasible. Also, a known drawback of exact path computation is its exponential complexity in the worst case [8], and numerical instabilities due to multiple inversions of potentially ill-conditioned matrices.

Another approach is to split the dataset into a training set - in which the regression model is fitted, and a calibration set - in which the conformity scores and their ranks are computed. Although this approach avoids the computational bottleneck of the full conformal prediction framework, statistical efficiencies are lost both in the model fitting stage and in the conformity score rank computation stage, due to the effect of a reduced sample size. It also adds another layer of randomness, which may be undesirable for the construction of prediction intervals [15].

A common heuristic approach in the literature is to evaluate the typicalness $\hat{\pi}(z)$ only for an arbitrary finite number of grid points. Although the prediction set constructed by those finite number of $\hat{\pi}(z)$ might roughly mimic the conformal prediction set, the desirable coverage properties are no longer maintained. To overcome this issue, [6] proposed a discretization strategy with a more careful procedure to round the observation vectors, but failed to exactly preserve the $1-\alpha$ coverage guarantee. In the appendix, we discuss in detail critical limitations of such an approach.

Algorithm 1 ϵ-`online_homotopy`

Input: $\mathcal{D}_n = \{(x_1, y_1), \cdots, (x_n, y_n)\}, x_{n+1}, [y_{\min}, y_{\max}], \epsilon_0 < \epsilon$
Initialization: $z_{t_0} = x_{n+1}^\top \beta$ where β is an ϵ_0-solution for the problem (1) using only $\mathcal{D}_n$
repeat
  $z_{t_{k+1}} = z_{t_k} \pm s_\epsilon$ where $s_\epsilon = \sqrt{\frac{2}{\nu}(\epsilon - \epsilon_0)}$ if the loss is ν-smooth
  Get $\beta(z_{t_{k+1}})$ by minimizing $P_{z_{t_{k+1}}}$ up to accuracy $\epsilon_0 < \epsilon$ {warm started with $\beta(z_{t_k})$}
until $[y_{\min}, y_{\max}]$ is covered
Return: $\{z_{t_k}, \beta(z_{t_k})\}_{k \in [T_\epsilon]}$

3 Homotopy Algorithm

In constructing an exact conformal set, we need to be able to compute the entire path of the model parameters $\hat{\beta}(z)$; which is obtained after solving the augmented optimization problem in Equation (3), for any z in $\mathbb{R}$. In fact, two problems arise. First, even for a single z, $\hat{\beta}(z)$ may not be available because, in general, the optimization problem *cannot* be solved exactly [19, Chapter 1]. Second, except for simple regression problems such as Ridge or Lasso, the entire exact path of $\hat{\beta}(z)$ cannot be computed infinitely many times.

Our basic idea to circumvent this difficulty is to rely on approximate solutions at a given precision $\epsilon > 0$. Here, we call an ϵ-solution any vector β such that its objective value satisfies

$$P_z(\beta) - P_z(\hat{\beta}(z)) \leq \epsilon \ . \tag{7}$$

An ϵ-solution can be found efficiently, under mild assumptions on the regularity of the function being optimized. In this section, we show that finite paths of ϵ-solutions can be computed for a wider class of regression problems. Indeed, it is not necessary to re-calculate a new solution for neighboring observations - *i.e.* $\beta(z)$ and $\beta(z_0)$ have the same performance when z is close to z_0. We develop a precise analysis of this idea. Then, we show how this can be used to effectively approximate the conformal prediction set in Equation (6) based on exact solution, while preserving the coverage guarantee.

We recall the dual formulation [22, Chapter 31] of Equation (3):

$$\hat{\theta}(z) \in \underset{\theta \in \mathbb{R}^{n+1}}{\arg\max}\, D_z(\theta) := -\sum_{i=1}^{n} \ell^*(y_i, -\lambda\theta_i) - \ell^*(z, -\lambda\theta_{n+1}) - \lambda\Omega^*(X^\top\theta) \ . \tag{8}$$

For a primal/dual pair of vectors $(\beta(z), \theta(z))$ in $\mathrm{dom}P_z \times \mathrm{dom}D_z$, the duality gap is defined as

$$\mathrm{Gap}_z(\beta(z), \theta(z)) := P_z(\beta(z)) - D_z(\theta(z)) \ .$$

Weak duality ensures that $P_z(\beta(z)) \geq D_z(\theta(z))$, which yields an upper bound for the approximation error of $\beta(z)$ in Equation (7) *i.e.*

$$P_z(\beta(z)) - P_z(\hat{\beta}(z)) \leq \mathrm{Gap}_z(\beta(z), \theta(z)) \ .$$

This will allow us to keep track of the approximation error when the parameters of the objective function change. Given any β such that $\mathrm{Gap}(\beta, \theta) \leq \epsilon$ *i.e.* an ϵ-solution for problem (1), we explore the candidates for y_{n+1} with the parameterization of the real line z_t defined as

$$z_t := z_0 + t, \text{ for } t \in \mathbb{R} \text{ and } z_0 = x_{n+1}^\top \beta \ . \tag{9}$$

This additive parameterization was used in [15] for the case of the Lasso. It provides the nice property that adding (x_{n+1}, z_0) as the $(n+1)$-th observation does not change the objective value of β *i.e.* $P(\beta) = P_{z_0}(\beta)$. Thus, if a vector β is an ϵ-solution for P, it will remain so for P_{z_0}. Interestingly, such a choice is still valid for a sufficiently small t. We show that, depending on the regularity of the loss function, we can precisely derive a range of the parameter t so that β remains a valid ϵ-solution for P_{z_t} when the dataset $\mathcal{D}_n$ is augmented with $\{(x_{n+1}, z_t)\}$.

We define the variation of the duality gap between real values z and z_0 to be

$$\Delta G(x_{n+1}, z, z_0) := \mathrm{Gap}_z(\beta, \theta) - \mathrm{Gap}_{z_0}(\beta, \theta) \ .$$

Lemma 1. *For any* $(\beta, \theta) \in \mathrm{dom} P_w \times \mathrm{dom} D_w$ *for* $w \in \{z_0, z\}$*, we have*

$$\Delta G(x_{n+1}, z, z_0) = [\ell(z, x_{n+1}^\top \beta) - \ell(z_0, x_{n+1}^\top \beta)] + [\ell^*(z, -\lambda\theta_{n+1}) - \ell^*(z_0, -\lambda\theta_{n+1})] \ .$$

Lemma 1 showed that the variation of the duality gap between z and z_0 depends only on the variation of the loss function ℓ, and its conjugate ℓ^*. Thus, it is enough to exploit the regularity (*e.g.* smoothness) of the loss function in order to obtain an upper bound for the variation of the duality gap (and therefore the optimization error).

Construction of Dual Feasible Vector. A generic method for producing a dual-feasible vector is to re-scale the output of the gradient mapping. For a real value z, let $\beta(z)$ be any primal vector and let us denote $Y_z = (y_1, \cdots, y_n, z)$.

Optimality conditions for (3) and (8) implies $\hat{\theta}(z) = -\nabla\ell(Y_z, X\hat{\beta}(z))/\lambda$, which suggests we can make use of [18]

$$\theta(z) := \frac{-\nabla\ell(Y_z, X\beta(z))}{\max\{\lambda_t, \sigma^\circ_{\mathrm{dom}\Omega^*}(X^\top \nabla\ell(Y_z, X\beta(z)))\}} \in \mathrm{dom} D_z \ , \tag{10}$$

where σ is the support function and σ° its polar function. When the regularization is a norm $\Omega(\cdot) = \|\cdot\|$, then $\sigma^\circ_{\mathrm{dom}\Omega^*}$ is the associated dual norm $\|\cdot\|_*$. When Ω is strongly convex, then the dual vector in Equation (10) simplifies to $\theta(z) = -\nabla\ell(Y_z, X\beta(z))/\lambda$.

Using $\theta(z_0)$ in Equation (10) with $z_0 = x_{n+1}^\top \beta$ greatly simplifies the expression for the variation of the duality gap between z_t and z_0 in Lemma 1 to

$$\Delta G(x_{n+1}, z_t, z_0) = \ell(z_t, x_{n+1}^\top \beta) \ .$$

This directly follows from the assumptions $\ell(z_0, z_0) = \ell^*(z_0, 0) = 0$ and by construction of the dual vector $\theta_{n+1} \propto \partial_2 \ell(z_0, x_{n+1}^\top \beta) = \partial_2 \ell(z_0, z_0) = 0$. Whence, assuming that the loss function is ν-smooth (see the appendix for more details and extensions to other regularity assumptions) and using the parameterization in Equation (9), we obtain

$$\Delta G(x_{n+1}, z_t, z_0) \leq \frac{\nu}{2}(z_t - z_0)^2 = \frac{\nu}{2} t^2 \ .$$

Proposition 1. *Assuming that the loss function* ℓ *is* ν*-smooth, the variations of the gap* $\Delta G(x_{n+1}, z_t, z_0)$ *are smaller than* ϵ *for all* t *in* $[-\sqrt{2\epsilon/\nu}, \sqrt{2\epsilon/\nu}]$*. Moreover, assuming that* $\mathrm{Gap}_{z_0}(\beta(z_0), \theta(z_0)) \leq \epsilon_0 < \epsilon$*, we have* $(\beta(z_0), \theta(z_0))$ *being a primal/dual* ϵ*-solution for the optimization problem* (3) *with augmented data* $\mathcal{D}_n \cup \{(x_{n+1}, z_t)\}$ *as long as*

$$|z_t - z_0| \leq \sqrt{\frac{2}{\nu}(\epsilon - \epsilon_0)} =: s_\epsilon \ .$$

Complexity. A given interval $[y_{\min}, y_{\max}]$ can be covered by Algorithm 1 with T_ϵ steps where

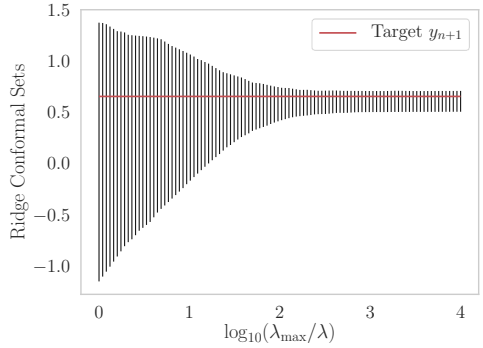
$$T_\epsilon \leq \left\lceil \frac{y_{\max} - y_{\min}}{s_\epsilon} \right\rceil \in O\left(\frac{1}{\sqrt{\epsilon}}\right) \ .$$

We can notice that the step sizes s_ϵ (smooth case) for computing the whole path are independent of the data and the intermediate solutions. Thus, for computational efficiency, the latter can be computed in parallel or by sequentially warm-starting the initialization. Also, since the grid can be constructed by decreasing or increasing the value of z_t, one can observe that the number of solutions calculated along the path can be halved by using only $\beta(z_t)$ as an ϵ-solution on the whole interval $[z_t \pm s_\epsilon]$.

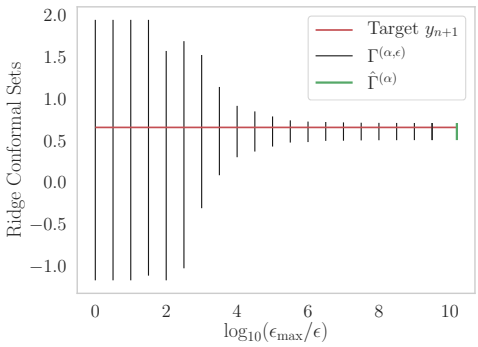
Lower Bound. Using the same reasoning when the loss is μ-strongly convex, we have

$$\Delta G(x_{n+1}, z_t, z_0) \geq \frac{\mu}{2}(z_t - z_0)^2 \ .$$

Hence $\Delta G(x_{n+1}, z_t, z_0) > \epsilon$ as soon as $|z_t - z_0| > \sqrt{\frac{2}{\mu}(\epsilon - \epsilon_0)}$. Thus, in order to guarantee ϵ approximation errors at any candidate z_t, all the step sizes are necessarily of order $\sqrt{\epsilon}$.



(a) Exact conformal prediction set for ridge regression with one hundred regularization parameters ranging from $\lambda_{\max} = \log(p)$ to $\lambda_{\min} = \lambda_{\max}/10^4$, spaced evenly on a log scale.



(b) Evolution of the conformal set of the proposed homotopy method with different optimization errors, spaced evenly on a geometric scale ranging from $\epsilon_{\max} = \|(y_1, \cdots, y_n)\|^2$ to $\epsilon_{\min} = \epsilon_{\max}/10^{10}$.

Figure 1: Illustration of conformal prediction sets at level $\alpha = 0.1$ with exact solutions and approximate solutions for ridge regression. We use a synthetic data set generated using `sklearn` with $X, y = \texttt{make\_regression}(n = 100, p = 50)$. We have chosen the hyperparameter with the smallest confidence set in Figure (a) to generate Figure (b).

Choice of $[y_{\min}, y_{max}]$. We follow the actual practice in the literature [15, Remark 5] and set $y_{\min} = y_{(1)}$ and $y_{\max} = y_{(n)}$. In that case, we have $\mathbb{P}(y_{n+1} \in [y_{\min}, y_{\max}]) \geq 1 - 2/(n+1)$. This implies a loss in the coverage guarantee of $2/(n+1)$, which is negligible when n is sufficiently large.

Related Works on Approximate Homotopy. Recent papers [9, 18] have developed approximation path methods when a function is concavely parameterized. Such techniques cannot be used here since, for any $\beta \in \mathbb{R}^p$, the function $z \mapsto P_z(\beta)$ is not concave. Thus, it does not fit within their problem description.

Using homotopy continuation to update an exact Lasso solution in the online setting was performed by [7, 15]. Allowing an approximate solution allows us to extensively generalize those approaches to a broader class of machine learning tasks, with a variety of regularity assumptions.

4 Practical Computation of a Conformal Prediction Set

We present how to compute a conformal prediction set, based on the approximate homotopy algorithm in Section 3. We show that the set obtained preserves the coverage guarantee, and tends to the exact set when the optimization error ϵ decreases to zero. In the case of a smooth loss function, we present a variant of conformal sets with an approximate solution, which contains the exact conformal set.

4.1 Conformal Sets Directly Based on Approximate Solution

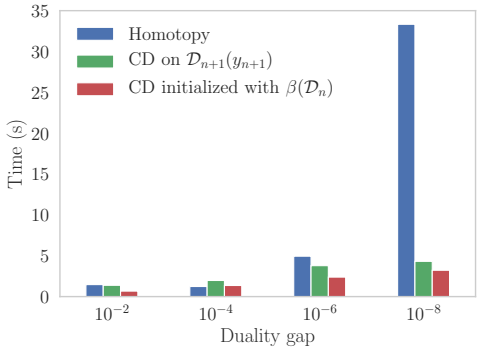
For a real value z, we cannot evaluate $\hat{\pi}(z)$ in Equation (5) in many cases because it depends on the exact solution $\hat{\beta}(z)$, which is unknown. Instead, we only have access to a given ϵ-solution $\beta(z)$ and the corresponding (approximate) conformity measure given as:

$$\forall i \in [n],\ R_i(z) = \psi(y_i, x_i^\top \beta(z)) \text{ and } R_{n+1}(z) = \psi(z, x_{n+1}^\top \beta(z))\ . \tag{11}$$

However, for establishing a coverage guarantee, one can note that *any* estimator that preserves exchangeability can be used. Whence, we define

$$\pi(z, \epsilon) := 1 - \frac{1}{n+1}\mathrm{Rank}(R_{n+1}(z)), \qquad \Gamma^{(\alpha,\epsilon)}(x_{n+1}) := \{z \in \mathbb{R} :\ \pi(z, \epsilon) > \alpha\}\ . \tag{12}$$

Proposition 2. *Given a significance level $\alpha \in (0, 1)$ and an optimization tolerance $\epsilon > 0$, if the observations $(x_i, y_i)_{i \in [n+1]}$ are exchangeable and identically distributed under probability $\mathbb{P}$, then the conformal set $\Gamma^{(\alpha,\epsilon)}(x_{n+1})$ satisfies the coverage guarantee $\mathbb{P}^{n+1}(y_{n+1} \in \Gamma^{(\alpha,\epsilon)}(x_{n+1})) \geq 1 - \alpha$.*



	Coverage	Length	Time (s)
Oracle	0.9	1.685	0.59
Split	0.9	3.111	0.26
1e-2	0.9	1.767	2.17
1e-4	0.9	1.727	8.02
1e-6	0.9	1.724	45.94
1e-8	0.9	1.722	312.56

Table 1: Computing a conformal set for a Lasso regression problem on a climate data set `NCEP/NCAR Reanalysis` [13] with $n = 814$ observations and $p = 73570$ features. On the left, we compare the time needed to compute the full approximation path with our homotopy strategy, single coordinate descent (CD) on the full data $\mathcal{D}_{n+1}(y_{n+1})$, and an update of the solution after initialization with an approximate solution using $\mathcal{D}_n$. On the right, we display the coverage, length and time of different methods averaged over 100 randomly held-out validation data sets.

The conformal prediction set $\Gamma^{(\alpha,\epsilon)}(x_{n+1})$ (with an approximate solution) preserves the $1-\alpha$ coverage guarantee and converges to $\Gamma^{(\alpha,0)}(x_{n+1}) = \hat{\Gamma}^{(\alpha)}(x_{n+1})$ (with an exact solution) when the optimization error decreases to zero. It is also easier to compute in the sense that only a finite number of candidates z need to be evaluated. Indeed, as soon as an approximate solution $\beta(z)$ is allowed, we have shown in Section 3 that a solution update is not necessary for neighboring observation candidates.

We consider the parameterization in Equation (9). It holds that

$$\Gamma^{(\alpha,\epsilon)} = \{z \in \mathbb{R} : \pi(z,\epsilon) > \alpha\} = \{z_t : t \in \mathbb{R}, \pi(z_t,\epsilon) > \alpha\} \ .$$

Using Algorithm 1, we can build a set $\{z_{t_1}, \cdots, z_{t_{T_\epsilon}}\}$ that covers $[y_{\min}, y_{\max}]$ with ϵ-solutions *i.e.* :

$$\forall z \in [y_{\min}, y_{\max}], \exists k \in [T_\epsilon] \text{ such that } \mathrm{Gap}_z(\beta(z_{t_k}), \theta(z_{t_k})) \leq \epsilon \ .$$

Using the classical conformity measure $\hat{R}_i(z) = |y_i - x_i^\top \hat{\beta}(z)|$ and computing a piecewise constant approximation of the solution path $t \mapsto \hat{\beta}(z_t)$ with the set $\{\beta(z_{t_k}) : k \in [T_\epsilon]\}$, we have

$$\Gamma^{(\alpha,\epsilon)} \cap [y_{\min}, y_{\max}] = \bigcup_{k \in [T_\epsilon]} [z_{t_k}, z_{t_{k+1}}] \cap [x_{n+1}^\top \beta(z_{t_k}) \pm Q_{1-\alpha}(z_{t_k})] \ .$$

where $Q_{1-\alpha}(z)$ is the $(1-\alpha)$-quantile of the sequence of approximate residuals $(R_i(z))_{i \in [n+1]}$.

Details and extensions to the more general cases of conformity measures are discussed in the appendix.

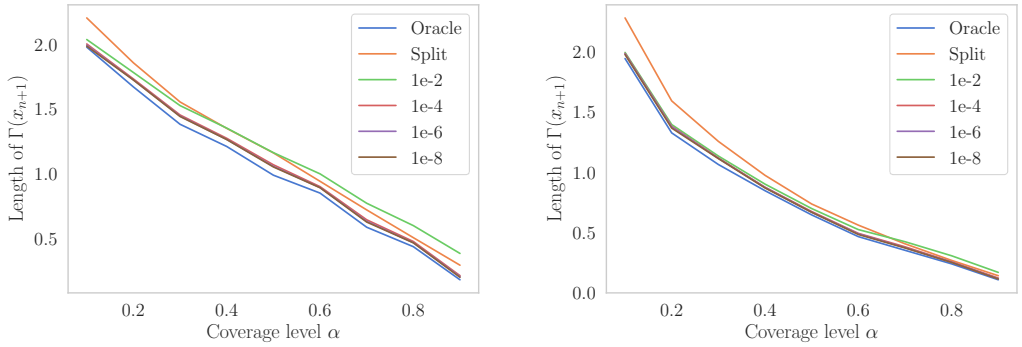
4.2 Wrapping the Exact Conformal Set

Previously, we showed that a full conformal set can be efficiently computed with an approximate solution, and it converges to the conformal set with an exact solution when the optimization error decreases to zero. When the loss function is smooth and, under a gradient-based conformity measure (introduced below), we provide a stronger guarantee that the exact conformal set can be included in a conformal set, using only approximate solutions. For this, we show how the conformity measure can be bounded *w.r.t.* to the optimization error, when the input observation z changes.

Gradient based Conformity Measures. The separability of the loss function implies that the coordinate-wise absolute value of the gradient of the loss function preserves the excheangeability of the data, and then the coverage guarantee. Whence it can be safely used as a conformity measure *i.e.*

$$\hat{R}_:(z) = |\nabla \ell(Y_z, X\hat{\beta}(z))|, \qquad R_:(z) = |\nabla \ell(Y_z, X\beta(z))| \ . \tag{13}$$

Using Equation (13), we show how the function $\hat{\pi}$ can be approximated from above and below, thanks to a fine bound on the dual optimal solution, which is related to the gradient of the loss function.



(a) Linear regression with ℓ_1 regularization on Diabetes dataset ($n = 442, p = 10$).

(b) `Logcosh` regression with ℓ_2^2 regularization on Boston dataset ($n = 506, p = 13$).

Figure 2: Length of the conformal prediction sets at different coverage level $\alpha \in \{0.1, 0.2, \cdots, 0.9\}$. For all α, we display the average over 100 repetitions of randomly held-out validation data sets.

Lemma 2. *If the loss function $\ell(z, \cdot)$ is ν-smooth, for any real value z, we have*

$$\|\theta(z) - \hat{\theta}(z)\|^2 \leq \frac{2\nu}{\lambda^2} \operatorname{Gap}_z(\beta(z), \theta(z)), \quad \forall (\beta(z), \theta(z)) \in \operatorname{dom} P_z \times \operatorname{dom} D_z \ .$$

Using Equation (13) and further assuming that the dual vector $\theta(z)$ constructed in Equation (10) coincides [1] with $-\nabla \ell(Y_z, X\beta(z))/\lambda$ in $\operatorname{dom} D_z$, we have $\hat{R}_:(z) = \|\lambda \hat{\theta}(z)\|$ and $R_:(z) = \|\lambda \theta(z)\|$.

Thus, combining the triangle inequality and Lemma 2 we have

$$\forall i \in [n+1],\ (R_i(z) - \hat{R}_i(z))^2 \leq \|R_:(z) - \hat{R}_:(z)\|^2 = \lambda^2 \|\theta(z) - \hat{\theta}(z)\|^2 \leq 2\nu\epsilon \ ,$$

where the last inequality holds as soon as we can maintain $\operatorname{Gap}_z(\beta(z), \theta(z))$ to be smaller than ϵ, for any z in $\mathbb{R}$. Whence, $\hat{R}_i(z)$ belongs to $[R_i(z) \pm \sqrt{2\nu\epsilon}]$ for any i in $[n+1]$. Noting that

$$\hat{\pi}(z) = 1 - \frac{1}{n+1} \operatorname{Rank}(\hat{R}_{n+1}(z)) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}_{\hat{R}_i(z) \geq \hat{R}_{n+1}(z)} \ ,$$

the function $\hat{\pi}$ can be easily approximated from above and below by the functions $\underline{\pi}(z, \epsilon)$ and $\overline{\pi}(z, \epsilon)$, which do not depend on the exact solution and are defined as:

$$\underline{\pi}(z, \epsilon) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}_{R_i(z) \geq R_{n+1}(z) + 2\sqrt{2\nu\epsilon}}, \quad \overline{\pi}(z, \epsilon) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1}_{R_i(z) \geq R_{n+1}(z) - 2\sqrt{2\nu\epsilon}} \ .$$

Proposition 3. *We assume that the loss function is ν-smooth and that we use a gradient based conformity measure* (13)*. Then, we have $\underline{\pi}(z, \epsilon) \leq \hat{\pi}(z) \leq \overline{\pi}(z, \epsilon)$ and the approximated lower and upper bounds of the exact conformal set are $\underline{\Gamma}^{(\alpha,\epsilon)} \subset \hat{\Gamma}^{(\alpha)} \subset \overline{\Gamma}^{(\alpha,\epsilon)}$ where*

$$\underline{\Gamma}^{(\alpha,\epsilon)} = \{z \in \mathbb{R} : \underline{\pi}(z, \epsilon) > \alpha\}, \qquad \overline{\Gamma}^{(\alpha,\epsilon)} = \{z \in \mathbb{R} : \overline{\pi}(z, \epsilon) > \alpha\} \ .$$

In the baseline case of quadratic loss, such sets can be easily computed as

$$\overline{\Gamma}^{(\alpha,\epsilon)} \cap [y_{\min}, y_{\max}] = \bigcup_{k \in [T_\epsilon]} [z_{t_k}, z_{t_{k+1}}] \cap [x_{n+1}^\top \beta(z_{t_k}) \pm Q^-_{1-\alpha}(t_k)] \ ,$$

$$\underline{\Gamma}^{(\alpha,\epsilon)} \cap [y_{\min}, y_{\max}] = \bigcup_{k \in [T_\epsilon]} [z_{t_k}, z_{t_{k+1}}] \cap [x_{n+1}^\top \beta(z_{t_k}) \pm Q^+_{1-\alpha}(t_k)] \ ,$$

where we have denoted $Q^-_{1-\alpha}(t_k)$ (resp. $Q^+_{1-\alpha}(t_k)$) as the $(1-\alpha)$-quantile of the sequence of shifted approximate residuals $(R_i(z_{t_k}) - 2\sqrt{2\nu\epsilon})_{i \in [n+1]}$ (resp. $(R_i(z_{t_k}) + 2\sqrt{2\nu\epsilon})_{i \in [n+1]}$) corresponding to the approximate solution $\beta(z_{t_k})$ for k in $[T_\epsilon]$.

[1] This holds whenever Ω is strongly convex or its domain is bounded. Also, one can guarantee this condition when $\beta(z)$ is build using any converging iterative algorithm, with sufficient iterations, for solving Equation (3).

	Oracle	Split	1e-2	1e-4	1e-6	1e-8
`Smooth Chebychev Approx.`						
Coverage	0.92	0.95	0.92	0.92	0.92	0.92
Length	1.940	2.271	1.998	1.990	1.987	1.981
Time (s)	0.019	0.016	0.073	0.409	3.742	36.977
`Linex regression`						
Coverage	0.91	0.93	0.91	0.91	0.91	0.91
Length	2.189	2.447	2.231	2.209	2.205	2.199
Time (s)	0.013	0.012	0.050	0.234	2.054	20.712

Table 2: Computing a conformal set for a `logcosh` (resp. `linex`) regression problem regularized with a Ridge penalty on the Boston (resp. Diabetes) dataset with $n = 506$ observations and $p = 13$ features (resp. $n = 442$ and $p = 10$). We display the coverage, length and time of the different methods, averaged over 100 randomly held-out validation data sets.

5 Numerical Experiments

We illustrate the approximation of a full conformal prediction set for both linear and non-linear regression problems, using synthetic and real datasets that are publicly available in `sklearn`. All experiments were conducted with a coverage level of 0.9 ($\alpha = 0.1$) and a regularization parameter selected by cross-validation on a randomly separated training set (for real data, we used 33% of the data).

In the case of Ridge regression, *exact* and *full* conformal prediction sets can be computed without any assumptions [20]. We show in Figure 1, the conformal sets *w.r.t.* different regularization parameters λ, and our proposed method based on an approximated solution for different optimization errors. The results indicate that high precision is not necessary to obtain a conformal set close to the exact one.

For other problem formulations, we define an `Oracle` as the set $[x_{n+1}^\top \hat{\beta}(y_{n+1}) \pm \hat{Q}_{1-\alpha}(y_{n+1})]$ obtained from the estimator trained with machine precision on the oracle data $\mathcal{D}_{n+1}(y_{n+1})$ (the target variable y_{n+1} is not available in practice). For comparison, we display the average over 100 repetitions of randomly held-out validation data sets, the empirical coverage guarantee, the length, and time needed to compute the conformal set with splitting and with our approach.

We illustrated in Table 1 the computational cost of our proposed homotopy for Lasso regression, using vanilla coordinate descent (CD) optimization solvers in `sklearn` [21]. For a large range of duality gap accuracies ϵ, the computational time of our method is roughly the same as a single run of CD on the full data set. However, when ϵ becomes very small ($\approx 10^{-8}$), we lose computational time efficiency due to large complexity T_ϵ. This is visible in regression problems with non-quadratic loss functions Table 2.

The computational times depend only on the data fitting part and the computation of the conformity score functions. Thus, the computational efficiency is independent of the coverage level α. We show in Figure 2, the variations of the length of the conformal prediction set for different coverage level.

Overall, the results indicate that the homotopy method provides valid and near-perfect coverage, regardless of the optimization error ϵ. The lengths of the confidence sets generated by homotopy methods gradually increase as ϵ increases, but all of the sets are consistently smaller than those of splitting approaches. Our experiments showed that high accuracy has only limited benefits.

Acknowledgments

We would like to thank the reviewers for their valuable feedbacks and detailed comments which contributed to improve the quality of this paper. This work was partially supported by MEXT KAKENHI (17H00758, 16H06538), JST CREST (JPMJCR1502), RIKEN Center for Advanced Intelligence Project, and JST support program for starting up innovation-hub on materials research by information integration initiative.

References

[1] D. Azé and J-P. Penot. Uniformly convex and uniformly smooth convex functions. *Annales de la faculté des sciences de Toulouse*, 1995.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 2012.

[3] V. Balasubramanian, S-S. Ho, and V. Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Elsevier, 2014.

[4] J. Bröcker and H. Kantz. The concept of exchangeability in ensemble forecasting. *Nonlinear Processes in Geophysics*, 2011.

[5] Y-C. Chang and W-L. Hung. Linex loss functions with applications to determining the optimum process parameters. *Quality & Quantity*, 2007.

[6] W. Chen, K-J. Chun, and R. F. Barber. Discretized conformal prediction for efficient distribution-free inference. *Stat*, 2018.

[7] P. Garrigues and L. E. Ghaoui. An homotopy algorithm for the lasso with online observations. In *Advances in neural information processing systems*, pages 489–496, 2009.

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

[9] J. Giesen, J. K. Müller, S. Laue, and S. Swiercy. Approximating concavely parameterized optimization problems. *Advances in neural information processing systems*, 2012.

[10] M. Gruber. *Regression estimators: A comparative study*. JHU Press, 2010.

[11] J-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms. II*. Springer-Verlag, 1993.

[12] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970.

[13] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, et al. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, 1996.

[14] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[15] J. Lei. Fast exact conformalization of lasso using piecewise linear homotopy. *Biometrika*, 2019.

[16] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 2018.

[17] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparsity enforcing penalties. *J. Mach. Learn. Res*, 2017.

[18] E. Ndiaye, T. Le, O. Fercoq, J. Salmon, and I. Takeuchi. Safe grid search with optimal complexity. *ICML*, 2019.

[19] Y. Nesterov. *Introductory lectures on convex optimization*. Kluwer Academic Publishers, 2004.

[20] I. Nouretdinov, T. Melluish, and V. Vovk. Ridge regression confidence machine. *ICML*, 2001.

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

[22] R. T. Rockafellar. *Convex analysis*. Princeton University Press, 1997.

[23] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 2008.

[24] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[25] A. Shibagaki, M. Karasuyama, K. Hatano, and I. Takeuchi. Simultaneous safe screening of features and samples in doubly sparse modeling. *ICML*, 2016.

[26] T. Sun and Q. Tran-Dinh. Generalized self-concordant functions: A recipe for newton-type methods. *Mathematical Programming*, 2018.

[27] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, 2005.

[28] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, 2005.

6 Appendix

More examples of Loss Function. Popular instances of loss functions can be found in the literature. For instance, in `power norm regression`, $\ell(a,b) = |a-b|^q$. When $q=2$, it corresponds to classical linear regression and the cases where $q \in [1,2)$ are common in robust statistics. In particular $q=1$ is known as least absolute deviation. One can also have the `log-cosh` loss $\ell(a,b) = \gamma \log(\cosh(a-b)/\gamma)$ as a differentiable alternative for the ℓ_∞ norm (chebychev approximation). One also have the `Linex` loss function [10, 5] which provide an asymmetric loss $\ell(a,b) = \exp(\gamma(a-b)) - \gamma(a-b) - 1$, for $\gamma \neq 0$. Least square fitting with non linear transformation where the relation between the observations and the features are described as $y_i \approx \phi(x_i, \beta)$ where ϕ is derived from physical or biological prior knowledge on the data. For instances, we have the exponential model $\phi(x_i,\beta) = a\exp(bx_i)$. Any convex regularization functions Ω can be considered. Popular examples are sparsity inducing norm [2], Ridge [12], elastic net [28], total variation, ℓ_∞, sorted ℓ_1 norm etc.

6.1 Homotopy with Different Regularity

We recall that from Lemma 1, we have

$$\Delta G(x_{n+1}, z_t, z_0) = [\ell(z, x_{n+1}^\top \beta) - \ell(z_0, x_{n+1}^\top \beta)] + [\ell^*(z_t, -\lambda\theta_{n+1}) - \ell^*(z_0, -\lambda\theta_{n+1})] \ . \quad (14)$$

We also recall the assumptions on the loss function.

Assumption A1. The functions ℓ and Ω are bounded from below. Thus, without loss of generality, we can also assume for any real value z_0 that $\ell^*(z_0, 0) = -\inf_z \ell(z_0, z) = 0$ otherwise one can always replace $\ell(z_0, \cdot)$ by $\ell(z_0, \cdot) - \inf_z \ell(z_0, z)$.

Assumption A2. For any real values z and z_0, we have $\ell(z_0, z), \ell(z, z_0) \geq 0$ and $\ell(z_0, z_0) = 0$.

This assumptions helps to simplify the first order expansion of ℓ at z_0 since $z_0 = \arg\min_z \ell(z, z_0)$ which is equivalent to $\partial_1 \ell(z_0, z_0) = 0$. Similarly, we also have $\partial_2 \ell(z_0, z_0) = 0$

Now we apply the formula Equation (14) to $z_0 = x_{n+1}^\top \beta$ and $z_t = z_0 + t$. Furthermore, using the dual vector in Equation (10), we have by construction $\theta_{n+1} \propto \partial_2 \ell(z_0, x_{n+1}^\top \beta) = \partial_2 \ell(z_0, z_0) = 0$. Then the variation of the gap between z_t and z_0 simplifies to

$$\Delta G(x_{n+1}, z_t, z_0) = \ell(z_t, z_0) \ . \quad (15)$$

Smooth Loss. To simplify the notation, given a real value b, we denote $\ell_{[b]}(a) = \ell(a,b)$ which is assumed to be a ν-smooth function *i.e.*

$$\ell_{[b]}(a) \leq \ell_{[b]}(a_0) + \langle \ell'_{[b]}(a_0), a - a_0 \rangle + \frac{\nu}{2}(a-a_0)^2, \quad \forall a, a_0 \ . \quad (16)$$

By assumption, $\ell_{[b]}(b) = 0$ and $\ell_{[b]}(a) \geq 0$. Thus we have $b = \arg\min_a \ell_{[b]}(a)$ which implies $\ell'_{[b]}(b) = 0$. Then $\ell(a,b) = \ell_{[b]}(a) \leq \frac{\nu}{2}(a-b)^2$; applied to $a = z_t$ and $b = z_0$, it reads:

$$\Delta G(x_{n+1}, z_t, z_0) \leq \frac{\nu}{2}(z_{n+1}(t) - z_0)^2 = \frac{\nu}{2}t^2 \ . \quad (17)$$

Lipschitz Loss. We suppose that the loss function is ν-Lipschitz *i.e.*

$$|\ell_{[b]}(a) - \ell_{[b]}(a_0)| \leq \nu |a - a_0| \ . \quad (18)$$

Applying Equation (18) to $a = z_t$ and $b = a_0 = z_0$ reads:

$$\Delta G(x_{n+1}, z_t, z_0) \leq \nu |z_t - z_0| = \nu |t| \ .$$

Whence the variation of the gap $\Delta G(x_{n+1}, z_t, z_0)$ are smaller than ϵ as soon as $t \in [-\epsilon/\nu, \epsilon/\nu]$. In that case, the complexity of the homotopy for covering the interval $[y_{\min}, y_{\max}]$ is

$$T_\epsilon \leq \left\lceil \frac{y_{\max} - y_{\min}}{\epsilon/\nu} \right\rceil \in O\left(\frac{1}{\epsilon}\right) \ .$$

$\mathcal{V}$-smooth Loss. We suppose that the loss function is uniformly smooth *i.e.*

$$\ell_{[b]}(a) \leq \ell_{[b]}(a_0) + \langle \ell'_{[b]}(a_0), a - a_0 \rangle + \mathcal{V}_{a_0}(a - a_0), \quad \forall a, a_0 \ , \tag{19}$$

where $\mathcal{V}_{a_0}$ is a non negative functions vanishing at zero.

Applying Equation (19) to $a = z_t$ and $b = a_0 = z_0$ reads:

$$\Delta G(x_{n+1}, z_t, z_0) \leq \mathcal{V}_{z_0}(z_t - z_0) = \mathcal{V}_{z_0}(t) \ . \tag{20}$$

The $\mathcal{V}$-smooth regularity contains two important known cases of local and global smoothness with different order:

- **Uniformly Smooth Loss** [1]. In this case, $\mathcal{V}_{a_0}(a - a_0) = \mathcal{V}(\|a - a_0\|)$ does not depends on a_0 and where $\mathcal{V}$ is any non increasing function from $[0, +\infty)$ to $[0, +\infty]$ *e.g.* $\mathcal{V}(t) = \frac{\mu}{d} t^d$. When $d = 2$, we recover the classical smoothness in (16).
  Thus, the variation of the gap $\Delta G(x_{n+1}, z_t, z_0)$ are smaller than ϵ as soon as $t \in [-\mathcal{V}^{-1}(\epsilon), \mathcal{V}^{-1}(\epsilon)]$. This leads to a generalized complexity of the homotopy for covering $[y_{\min}, y_{\max}]$ in T_ϵ steps where

$$T_\epsilon \leq \left\lceil \frac{y_{\max} - y_{\min}}{\mathcal{V}^{-1}(\epsilon)} \right\rceil \in O\left(\frac{1}{\mathcal{V}^{-1}(\epsilon)} \right) \ .$$

- **Generalized Self-Concordant Loss** [26]. 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, [26, Proposition 10] have shown that one could write:

$$\mathcal{V}_{\ell_{[b]}, a_0}(a - a_0) = w_\nu(d_\nu(a_0, a)) \, \|a - a_0\|_{a_0}^2 \ ,$$

  where the last equality holds if $d_\nu(a_0, a) < 1$ for the case $\nu > 2$. Closed-form expressions of $w_\nu(\cdot)$ and $d_\nu(\cdot)$ are given as follow:

$$d_\nu(a_0, a) := \begin{cases} M_{\ell_{[b]}} \, \|a - a_0\|_2 & \text{if } \nu = 2, \\ \left(\frac{\nu}{2} - 1\right) M_{\ell_{[b]}} \, \|a - a_0\|_2^{3 - \nu} \, \|a - a_0\|_{a_0}^{\nu - 2} & \text{if } \nu > 2, \end{cases} \tag{21}$$

  and

$$w_\nu(\tau) := \begin{cases} \frac{e^\tau - \tau - 1}{\tau^2} & \text{if } \nu = 2, \\ \frac{-\tau - \log(1 - \tau)}{\tau^2} & \text{if } \nu = 3, \\ \frac{(1 - \tau)\log(1 - \tau) + \tau}{\tau^2} & \text{if } \nu = 4, \\ \left(\frac{\nu - 2}{4 - \nu}\right) \frac{1}{\tau} \left[\frac{\nu - 2}{2(3 - \nu)\tau} \left((1 - \tau)^{\frac{2(3 - \nu)}{2 - \nu}} - 1 \right) - 1 \right] & \text{otherwise.} \end{cases} \tag{22}$$

  Power loss function $\ell_{[b]}(a, b) = (a - b)^q$ for $q \in (1, 2)$, popular in robust regression, is covered with $M_{\ell_{[b]}} = \frac{2 - q}{(2 - q)\sqrt[(2-q)]{q(q-1)}}, \nu = \frac{2(3-q)}{2-q} \in (4, +\infty)$.
  We refer to [26] for more details and examples.

Note that when local smoothness is used, the step sizes depend on the current candidate z_{t_k} along the path: the generated grid is adaptive and the step sizes can be computed numerically.

6.2 Proofs

Lemma 3 (c.f. Lemma 1)**.** *For any $(\beta, \theta) \in \mathrm{dom} P_z \times \mathrm{dom} D_z$ for $z \in \{z_0, y\}$, we have*

$$\Delta G(x_{n+1}, z, z_0) = [\ell(z, x_{n+1}^\top \beta) - \ell(z_0, x_{n+1}^\top \beta)] + [\ell^*(z, -\lambda \theta_{n+1}) - \ell^*(z_0, -\lambda \theta_{n+1})] \ . \tag{23}$$

Proof. By definition,

$$\begin{aligned} \Delta G(x_{n+1}, z, z_0) &= \mathrm{Gap}_z(\beta, \theta) - \mathrm{Gap}_{z_0}(\beta, \theta) = [P_z(\beta) - D_z(\theta)] - [P_{z_0}(\beta) - D_{z_0}(\theta)] \\ &= [P_z(\beta) - P_{z_0}(\beta)] - [D_z(\theta) - D_{z_0}(\theta)] \ . \end{aligned}$$

The conclusion follows from the fact that the first term is

$$P_z(\beta) - P_{z_0}(\beta) = \ell(z, x_{n+1}^\top \beta) - \ell(z_0, x_{n+1}^\top \beta) \ ,$$

and the second term is

$$D_z(\theta) - D_{z_0}(\theta) = \ell^*(z_0, -\lambda\theta_{n+1}) - \ell^*(z, -\lambda\theta_{n+1}) \ .$$

□

For the initialization, we start with a couple of vector $(\beta, \theta) \in \mathrm{dom}P \times \mathrm{dom}D \subset \mathbb{R}^p \times \mathbb{R}^n$ that we need to extent to $(\beta, \theta^+) \in \mathrm{dom}P_z \times \mathrm{dom}D_z \subset \mathbb{R}^p \times \mathbb{R}^{n+1}$. For better clarity, we restate the previous lemma to this specific case.

Lemma 4. *Let (β, θ) be any primal/dual vector in $\mathrm{dom}P \times \mathrm{dom}D$ and $\theta^+ = (\theta, 0)$ in $\mathbb{R}^{n+1}$. For any real value z, the variation of the duality gap is equal to the loss between z and $x_{n+1}^\top \beta$* i.e.

$$\Delta G(x_{n+1}, z) := \mathrm{Gap}_z(\beta, \theta^+) - G(\beta, \theta) = \ell(z, x_{n+1}^\top \beta) \ .$$

Proof. Let $\delta \in \mathbb{R}$ such that $\theta_\delta^+ = (\theta, \delta)^\top \in \mathbb{R}^{n+1} \in \mathrm{dom}D_z$. We have

$$\begin{aligned}
\Delta G(\delta) &:= \mathrm{Gap}_z(\beta, \theta_\delta^+) - G(\beta, \theta) \\
&= [P_z(\beta) - P(\beta)] - [D_z(\theta_\delta^+) - D(\theta)] \\
&= \ell(z, x_{n+1}^\top \beta) + \ell^*(z, -\lambda\delta) + \lambda \left[\Omega^*(X_{[n]}^\top \theta + \delta x_{n+1}^\top) - \Omega^*(X_{[n]}^\top \theta)\right] \ .
\end{aligned}$$

We choose $\delta = 0$; which is admissible because $0 \in \mathrm{dom}\ell^*(z, \cdot)$ if and only if $\ell(z, \cdot)$ is bounded from below as assumed. The result follows the observation that $\Delta G(0) = \Delta G(x_{n+1}, z)$. □

Proposition 4 (c.f. Proposition 2)**.** *Given a significance level $\alpha \in (0, 1)$ and an optimization tolerance $\epsilon > 0$, if the observations $(x_i, y_i)_{i \in [n+1]}$ are exchangeable and identically distributed under probability $\mathbb{P}$, then the conformal set $\Gamma^{(\alpha,\epsilon)}(x_{n+1})$ satisfies the coverage guarantee*

$$\mathbb{P}^{n+1}(y_{n+1} \in \Gamma^{(\alpha,\epsilon)}(x_{n+1})) \geq 1 - \alpha \ .$$

Proof. The separability of the loss function in P_z implies that

$$\beta((x_i, y_i)_{i \in [n+1]}) = \beta((x_{\sigma(i)}, y_{\sigma(i)})_{i \in [n+1]}) \ ,$$

for any permutation σ of the index set $\{1, \cdots, n+1\}$. Whence the sequence of conformity measure $(R_i(y_{n+1}))_{i \in [n+1]}$ is invariant *w.r.t.* permutation of the data.

The exchangeability of the sequence $\{(x_i, y_i)_{i \in [n+1]}\}$ implies that of $(R_i(y_{n+1}))_{i \in [n+1]}$.

The rests of the proof are based on the fact that the rank of one variable among an *exchangeable and identically distributed* sequence is (sub)-uniformly distributed [4].

Lemma 5. *Let $U_1, \ldots, U_{n+1}$ be exchangeable and identically distributed sequence of real valued random variables. Then for any $\alpha \in (0, 1)$, we have $\mathbb{P}^{n+1}(\mathrm{Rank}(U_{n+1}) \leq (n+1)(1-\alpha)) \geq 1 - \alpha$.*

Using Lemma 5, we deduce that the rank of $R_{n+1}(y_{n+1})$ among $(R_i(y_{n+1}))_{i \in [n+1]}$ is sub-uniformly distributed on the discrete set $\{1, \cdots, n+1\}$. Recalling the definition of *typicalness*

$$\forall z \in \mathbb{R}, \quad \pi(z, \epsilon) = 1 - \frac{1}{n+1} \mathrm{Rank}(R_{n+1}(z)) \ ,$$

We have

$$\mathbb{P}^{n+1}(\pi(y_{n+1}, \epsilon) > \alpha) = \mathbb{P}^{n+1}(\mathrm{Rank}(R_{n+1}(y_{n+1}) < (n+1)(1-\alpha)) \geq 1 - \alpha \ .$$

The proof for conformal set with exact solution corresponds to $\epsilon = 0$. □

Lemma 6 (c.f. Lemma 2)**.** *Assuming that $\ell(y_i, \cdot)$ is ν-smooth, we have*

$$\|\theta(z) - \hat{\theta}(z)\|^2 \leq \frac{2\nu}{\lambda^2} \mathrm{Gap}_z(\beta(z), \theta(z)) \ . \tag{24}$$

Note that such a bound on the dual optimal solution, leveraging duality gap, was used in optimization [17, 25] to bound the Lagrange multipliers for identifying sparse components in lasso type problems.

Proof. Remember that $\forall i \in [n], \ell(y_i, \cdot)$ is ν-smooth. As a consequence, $\forall i \in [n], \ell^*(y_i, \cdot)$ is $1/\nu$-strongly convex [11, Theorem 4.2.2, p. 83] and so the dual function D_λ is λ^2/ν-strongly concave:

$$\forall(\theta_1, \theta_2) \quad D_z(\theta_2) \leq D_z(\theta_1) + \langle \nabla D_z(\theta_1), \theta_2 - \theta_1 \rangle - \frac{\lambda^2}{2\nu} \|\theta_1 - \theta_2\|^2 \ .$$

Specifying the previous inequality for $\theta_1 = \hat{\theta}(z), \theta_2 = \theta(z)$, one has

$$D_z(\theta) \leq D_z(\hat{\theta}(z)) + \langle \nabla D_z(\hat{\theta}(z)), \theta(z) - \hat{\theta}(z) \rangle - \frac{\lambda^2}{2\nu} \|\hat{\theta}(z) - \theta(z)\|^2 \ .$$

By definition, $\hat{\theta}(z)$ maximizes D_z, so, $\langle \nabla D_z(\hat{\theta}(z)), \theta(z) - \hat{\theta}(z) \rangle \leq 0$. This implies

$$D_z(\theta(z)) \leq D_z(\hat{\theta}(z)) - \frac{\lambda^2}{2\nu} \|\hat{\theta}(z) - \theta(z)\|^2.$$

By weak duality, we have $D_z(\hat{\theta}(z)) \leq P_z(\beta(z))$, hence

$$D_z(\theta(z)) \leq P_z(\beta(z)) - \frac{\lambda^2}{2\nu} \|\hat{\theta}(z) - \theta(z)\|^2$$

and the conclusion follows. □

Proposition 5 (c.f. Proposition 3)**.** *We assume that the loss function is ν-smooth and that we use a gradient based conformity measure* (13)*. Then, we have $\underline{\pi}(z, \epsilon) \leq \hat{\pi}(z) \leq \overline{\pi}(z, \epsilon)$ and the approximated lower and upper bounds of the exact conformal set are $\underline{\Gamma}^{(\alpha,\epsilon)} \subset \hat{\Gamma}^{(\alpha)} \subset \overline{\Gamma}^{(\alpha,\epsilon)}$ where*

$$\underline{\Gamma}^{(\alpha,\epsilon)} = \{z \in \mathbb{R} : \underline{\pi}(z, \epsilon) > \alpha\}, \qquad \overline{\Gamma}^{(\alpha,\epsilon)} = \{z \in \mathbb{R} : \overline{\pi}(z, \epsilon) > \alpha\} \ .$$

Proof. We recall that for any i in $[n+1]$, we have $\hat{R}_i(z)$ belongs to $[R_i(z) \pm \sqrt{2\nu\epsilon}]$. Then

$$\begin{aligned} \hat{R}_i(z) \geq \hat{R}_{n+1}(z) &\Longrightarrow R_i(z) + \sqrt{2\nu\epsilon} \geq \hat{R}_i(z) \geq \hat{R}_{n+1}(z) \geq R_{n+1}(z) - \sqrt{2\nu\epsilon} \\ &\Longrightarrow R_i(z) \geq R_{n+1}(z) - 2\sqrt{2\nu\epsilon} \ . \end{aligned}$$

Whence $\hat{\pi}(z) \leq \overline{\pi}(z, \epsilon)$. The inequality $\underline{\pi}(z, \epsilon) \leq \hat{\pi}(z)$ follows from the fact that $R_i(z) - \sqrt{2\nu\epsilon}$ (resp. $R_{n+1}(z) + \sqrt{2\nu\epsilon}$) is a lower bound of $\hat{R}_i(z)$ (resp. upper bound of $\hat{R}_{n+1}(z)$). □

6.3 Details on Practical Computations

For simplicity, let us first restrict to the case of quadratic loss where the conformity measure is defined such as $\hat{R}_i(z) = |y_i - x_i^\top \hat{\beta}(z)|$.

Note that $\pi(z, \epsilon) > \alpha$ if and only if $R_{n+1}(z) \leq Q_{1-\alpha}(z)$ where $Q_{1-\alpha}(z)$ is the $(1-\alpha)$-quantile of the sequence of approximate residual $(R_i(z))_{i \in [n+1]}$. Then the approximate conformal set can be conveniently written as

$$\Gamma^{(\alpha,\epsilon)} = \{z \in \mathbb{R} : R_{n+1}(z) \leq Q_{1-\alpha}(z)\} = \bigcup_{z \in \mathbb{R}} [x_{n+1}^\top \beta(z) \pm Q_{1-\alpha}(z)] \ .$$

Let $\{\beta(z_{t_k}) : k \in [T_\epsilon]\}$ be the set of solutions outputted by the approximation homotopy method, the functions $t \mapsto \beta(z_t)$ ϵ-solution of optimization problem (3) using the data $\mathcal{D}_{n+1}(z_t)$ and $t \mapsto x^\top \beta(z_t)$, are piecewise constant on the intervals (t_k, t_{k+1}). Also, the map $t \mapsto R_{n+1}(z_t)$ (resp. $t \mapsto R_i(z_t)$ for i in $[n]$) is piecewise linear (resp. piecewise constant) on $[t_k, t_{k+1}]$. Thus,we have

$$\begin{aligned} \Gamma^{(\alpha,\epsilon)} \cap [y_{\min}, y_{\max}] &= \{z_t : t \in \mathbb{R}, R_{n+1}(z_t) \leq Q_{1-\alpha}(z_t\} \cap [y_{\min}, y_{\max}] \\ &= \bigcup_{k \in [T_\epsilon]} \{z_t : t \in [t_k, t_{k+1}], R_{n+1}(z_t) \leq Q_{1-\alpha}(z_t)\} \\ &= \bigcup_{k \in [T_\epsilon]} [z_{t_k}, z_{t_{k+1}}] \cap [x_{n+1}^\top \beta(z_{t_k}) \pm Q_{1-\alpha}(z_{t_k})] \ . \end{aligned}$$

where $Q_{1-\alpha}(z)$ is the $(1-\alpha)$-quantile of the sequence of approximate residual $(R_i(z))_{i \in [n+1]}$.

Extensions to Others Nonconformity Measure. We consider a generic conformity measure in Equation (4). We basically follow the same step than the derivation of conformal set for ridge in [27].

For any i in $[n+1]$, we denote the intersection points of the functions $R_i(z_t)$ and $R_{n+1}(z_t)$ restricted on the interval $[t_k, t_{k+1}]$: $t_{k,i}^-$ and $t_{k,i}^+$.

We however assume that there is only two intersection points. The computations for more finitely points are the same. For instance, using the absolute value as conformity measure, we have

$$t_{k,i}^- = (\mu_{t_k}(x_{n+1}) - R_i(z_{t_k}) - z_0), \qquad t_{k,i}^+ = (\mu_{t_k}(x_{n+1}) + R_i(z_{t_k}) - z_0) \ ,$$

where $\mu_{t_k}(x_{n+1}) := x_{n+1}^\top \beta(z_{t_k})$. Now, let us define

$$S_i = \{t \in [t_{\min}, t_{\max}] : R_i(z_t) \geq R_{n+1}(z_t)\} = \bigcup_{k\in[T_\epsilon]} S_i \cap [t_k, t_{k+1}] = \bigcup_{k\in[T_\epsilon]} [t_{k,i}^-, t_{k,i}^+] \ .$$

For any k in $[T_\epsilon]$, we denote the set of solutions $t_{k,1}^-, t_{k,1}^+, \cdots, t_{k,n+1}^-, t_{k,n+1}^+$ in increasing order as $t_k = t_{k,0} < t_{k,1} < \cdots < t_{k,l_k} = t_{k+1}$. Whence for any $t \in [t_k, t_{k+1}]$, it exists a unique index $j = \mathcal{J}(t)$ such that $t \in (t_{k,j}, t_{k,j+1})$ or $t \in \{t_{k,j}, t_{k,j+1}\}$ and for any $t \in [t_k, t_{k+1}]$, we have

$$(n+1)\pi(z_t) = \sum_{i=1}^{n+1} \mathbb{1}_{t\in S_i \cap [t_k, t_{k+1}]} = N_k(\mathcal{J}(t)) + M_k(\mathcal{J}(t))$$

where the functions

$$N_k(j) = \sum_{i=1}^{n+1} \mathbb{1}_{(t_{k,j}, t_{k,j+1}) \subset [t_{k,i}^-, t_{k,i}^+]} \text{ and } M_k(j) = \sum_{i=1}^{n+1} \mathbb{1}_{t_{k,j} \in [t_{k,i}^-, t_{k,i}^+]}$$

Note that $\mathcal{J}^{-1}([t_k, t_{k+1}]) = \{0, 1, \cdots, l_k\}$ and $\mathcal{J}^{-1}(j) = [t_{k,j}, t_{j+1}]$. Finally, we have

$$\Gamma^{(\alpha,\epsilon)} \cap [y_{\min}, y_{\max}] = \bigcup_{k\in[T_\epsilon]} \Gamma^{(\alpha,\epsilon)} \cap [t_k, t_{k+1}] \tag{25}$$

$$= \bigcup_{k\in[T_\epsilon]} \bigcup_{\substack{j\in[0:l_k] \\ N_k(j) > (n+1)\alpha}} (t_{k,j}, t_{k,j+1}) \ \cup \bigcup_{\substack{j\in[0:l_k] \\ M_k(j) > (n+1)\alpha}} \{t_j\} \ . \tag{26}$$

6.4 Alternative Grid based Strategies.

Another line of attempts to find a discretization of the set $\hat{\Gamma}^{(\alpha)}$ consist in roughly approximating the conformal set by restricting $\hat{\Gamma}^{(\alpha)}(x_{n+1})$ to an arbitrary fine grid of candidate $\hat{\mathcal{Y}}$ *i.e.* $\bigcup_{z\in\hat{\mathcal{Y}}}[x_{n+1}^\top \hat{\beta}(z) \pm \hat{Q}_{1-\alpha}(z)]$ [16]. Such approximation did not show any coverage guarantee. To overcome this issue, [6] have proposed a discretization strategy with a more carefully rounding procedure of the observation vectors.

Given an arbitrary finite set $\hat{\mathcal{Y}}$ and any discretization function $\hat{d} : \mathbb{R} \mapsto \hat{\mathcal{Y}}$, define

$$\Gamma^{\alpha,1} = \{z \in \mathbb{R} : \hat{d}(z) \in [x_{n+1}^\top \hat{\beta}(d(z)) \pm \hat{Q}_{1-\alpha}(\hat{d}(z))]\} \ , \tag{27}$$

$$\Gamma^{\alpha,2} = \bigcup_{z\in\hat{\mathcal{Y}}} \hat{d}^{-1}(z) \cap [x_{n+1}^\top \hat{\beta}(z) \pm \hat{Q}_{1-\alpha}(z)] \ . \tag{28}$$

Then [6][Theorem 2] showed that for any exchangeable finite set $\tilde{\mathcal{Y}} = \tilde{\mathcal{Y}}(\mathcal{D}_{n+1})$ and discretization function $\tilde{d} : \mathbb{R} \mapsto \tilde{\mathcal{Y}}$, we have the coverage

$$\mathbb{P}^{n+1}(y_{n+1} \in \Gamma^{\alpha,i}) \geq 1 - \alpha - \mathbb{P}^{n+1}((\hat{\mathcal{Y}}, \hat{d}) \neq (\tilde{\mathcal{Y}}, \tilde{d})) \text{ for } i \in \{1, 2\} \ . \tag{29}$$

A noticeable weakness of this result is that it strongly depends on the relation between the couples $(\hat{\mathcal{Y}}, \hat{d})$ and $(\tilde{\mathcal{Y}}, \tilde{d})$. Equation (29) fails to provide any meaningful informations in many situations *e.g.* the bound is vacuous anytime $|\hat{\mathcal{Y}}| \neq |\tilde{\mathcal{Y}}|$ or two different discretizations are chosen. Thus, the sets $\Gamma^{\alpha,1}, \Gamma^{\alpha,2}$ need a careful choice of the finite grid point $\hat{\mathcal{Y}}$ to be practical. This paper shows how to automatically and efficiently calibrate such set without loss in the coverage guarantee. Our approach provide optimization stopping criterion for each grid point while for arbitrary discretization, one must solve problem (3) at unnecessarily high accuracy. Last but not least, when the loss function is smooth, our approach provide an unprecedented guarantee to contains the full, exact conformal set.

6.5 Additional Experiments

6.5.1 Sparse Nonlinear Regression

We run experiments on the `Friedman1` regression problem available in `sklearn` where the inputs X are independent features uniformly distributed on the interval $[0, 1]$. The output z is nonlinearly generated using only 5 features.

$$y = 10\sin(\pi X_{:,1}X_{:,2}) + 20(X_{:,3} - 0.5)^2 + 10X_{:,4} + 5X_{:,5} + 0.5\mathcal{N}(0, 1) \ . \tag{30}$$

The results are displayed in Table 3

	Oracle	Split	1e-2	1e-4	1e-6	1e-8
`Lasso`						
Coverage	0.91	0.88	0.93	0.89	0.89	0.89
Length	1.50	2.320	2.272	2.011	1.956	1915
Time	0.005	0.003	0.020	0.076	0.397	3.499

Table 3: Computing conformal set for `lasso` regression problem `friedman1` dataset with $n = 506$ observations and $p = 13$ features (resp. $n = 500$ and $p = 50$). We display the coverage, length and time of different methods averaged over 100 randomly left out validation data.

6.5.2 Real Data with Large Number of Observations

In this benchmark, we illustrate the performances of the different conformal prediction strategies when the number of observations is large. We use the California housing dataset available in `sklearn`. Results are reported in Table 4.

	Oracle	Split	1e-2	1e-4	1e-6	1e-8
`Smooth Chebychev Approx.`						
Coverage	0.92	0.92	0.92	0.92	0.92	0.92
Length	0.014	0.014	0.014	0.014	0.014	0.014
Time	0.096	0.065	0.203	0.269	0.942	7.578

Table 4: Computing conformal set for `logcosh` regression problem regularized with Ridge penalty on California housing dataset with $n = 20640$ observations and $p = 8$ features. We display the coverage, length and time of different methods averaged over 100 randomly left out validation data.

In this example, both the Splitting and the proposed homotopy method achieves the same performances than the `Oracle` (which use the target y_{n+1} in the model fitting). Due to the large number of observations n, the efficiency of the Splitting approach is less affected by its inherent reduction of the sample size. We still note that, a rough approximation of the optimal solution is sufficient to get a good conformal prediction set with homotopy.