Geometry-dependent matching pursuit: a transition phase for convergence of linear regression and LASSO

Céline Moucer[†]*, Francis Bach[†], and Adrien Taylor[†]

Abstract. Greedy first-order methods, such as coordinate descent with Gauss-Southwell rule or matching pursuit, have become popular in optimization due to their natural sparse solutions and their refined convergence guarantees. In this work, we propose a principled approach to generating (regularized) matching pursuit algorithms adapted to the geometry of the problem at hand, as well as their convergence guarantees. Building on these results, we derive approximate convergence guarantees and describe a transition phenomenon from underparametrized to overparametrized models.

Key words. Optimization, first-order methods, matching pursuit, linear regression, LASSO.

MSC codes. 90C25, 68Y25, 60B20

1. Introduction.

Many problems in machine learning and data science take the form of an ℓ_1-regularized minimization problem:

$$G_\star = \min_{\alpha \in \mathbb{R}^d} f(P\alpha) + \lambda \|\alpha\|_1 = F(\alpha) + H(\alpha), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth strongly convex function, $P \in \mathbb{R}^{n \times d}$ and n, d respectively denote the number of samples and d the dimension of the problem. Typically, in the vanilla least-squares regression problem, $H(\alpha) = 0$ and $F(\alpha) = f(P\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2$, where P corresponds to the input data, $y \in \mathbb{R}^n$ to the labels, d to the number of features (or parameters) and n the number of observations. If in addition $H(\alpha) = \lambda\|\alpha\|_1$, Problem (1.1) is exactly the LASSO problem [57], that belongs to more general variational problems appearing in Fenchel duality theory [6, Section 15.3]. Problem (1.1) is often compared to its constrained counterpart,

$$\min_{\alpha \in \mathbb{R}^d} F(\alpha), \text{ such that } \|\alpha\|_1 \leqslant R, \tag{1.2}$$

where λ may be seen as the Lagrange multiplier associated to the constraint $\|\alpha\|_1 \leqslant R$ with $R > 0$. Problems (1.1) and (1.2) arise when looking for sparsity patterns, such as in signal processing where we aim for models depending on a small number of variables, or for trace-norm regularized problems, when looking for low-rank solutions [17]. In particular, Problem (1.1) is a popular way to induce sparsity on the solution for a well-chosen range of λ. Thus, ℓ_1-penalization (or constraint) is strongly connected to sparsity and often presented as a convex substitute for ℓ_0-penalization problems [62, Introduction], that correspond to feature selection.

First-order methods have become popular to solve optimization Problems (1.1) and (1.2), due to their low cost per iteration and to the limited accuracy requirements in machine learning [12, Section 7 and 8]. Within these methods, different algorithms may be used depending

*Ecole Nationale des Ponts et Chaussées, Marne-la-Vallée, France.

[†]DI ENS, École normale supérieure, Université PSL, CNRS, INRIA, 75005 Paris, France (celine.moucer@inria.fr), (adrien.taylor@inria.fr), (francis.bach@inria.fr).

on the properties of functions F and H. For instance, a first-order method may call for the gradient, or for the proximal operator [45] as in the proximal gradient method, or the linear minimization oracle as in the Frank-Wolfe algorithm [28] for the constrained version (1.2). These methods benefit from convergence guarantees, that are mostly helpful to practitioners when they can be numerically computed *a priori*.

In the context of sparsity, traditional first-order methods, such as the proximal gradient, do not always lead to sparse solutions [27]. Boosting strategies (also known as matching pursuit) have been developed to ensure sparse representations of approximate solutions [34, 58]. At each iteration, a possibly new atom (also referred to as a weak-learner in the boosting literature, or a coordinate in the context of coordinate descent) is greedily selected as a best candidate among a set of atoms, and combined to past iterates. While boosting benefits from strong statistical properties [58], from an optimization perspective, their convergence analyses often rely on extra statistical assumptions [63]. More recently, randomized and greedy coordinate descent methods have gained interest due to their low-cost per iteration in high dimension [39] and to their implicit induced sparsity [7, 18].

Correspondences have been highlighted between first-order methods and boosting strategies for non-regularized minimization problems ($\lambda = 0$), leading to convergence guarantees independent of traditional statistical assumptions. For example, coordinate descent has been interpreted as matching pursuit [33], as well as Frank-Wolfe algorithms [28, 32] for constrained Problems (1.2), by formulating them as minimizers of well-chosen quadratic upper approximations. These analyses strongly rely on a well-chosen geometry, characterized by a gauge function [22]. To our knowledge, this comparison was only drawn for non-regularized problems for which $\lambda = 0$ [33] and for constrained problems [54].

Problem (1.1) in $\mathbb{R}^d$ can be naturally formulated as an optimization problem in $\mathbb{R}^n$, letting the gauge geometry appear,

$$\begin{aligned} G_\star &= \min_{\alpha \in \mathbb{R}^d, x \in \mathbb{R}^n} G(\alpha) = f(x) + \lambda \|\alpha\|_1, \text{ such that } x = P\alpha, \\ &= \min_{x \in \mathbb{R}^n} f(x) + \lambda \inf_{\alpha \in \mathbb{R}^d,\ x = P\alpha} \|\alpha\|_1, \\ &= \min_{x \in \mathbb{R}^n} f(x) + \lambda \gamma_{\mathcal{P}}(x), \end{aligned} \tag{1.3}$$

where the gauge function is defined by $\gamma_{\mathcal{P}}(x) = \inf_{\alpha \in \mathbb{R}^d,\ x = P\alpha} \|\alpha\|_1$ with $\mathcal{P} = \operatorname{conv}(\{\pm P_{:,i}, i = 1, \dots, d\})$ the centrally symmetric convex hull of the columns of P. Gauge functions may be seen as generalized versions of the ℓ_1-norm, providing a sparse representation $\alpha \in \mathbb{R}^d$ of a vector $x \in \mathbb{R}^n$ with respect to a set of atoms. Under some assumptions on P, the gauge function may be a norm, as we will see in Section 2. Let us for example take $\mathcal{P} = \operatorname{conv}(\pm e_i)$, then $\gamma_{\mathcal{P}}(x) = \|x\|_1$.

Due to the connection between optimizing in $\mathbb{R}^d$ and in $\mathbb{R}^n$, it is possible to derive algorithms adapted to one or the other geometry, and to formulate geometry-adapted convergence guarantees. For the ℓ_1-geometry, Nutini et al. [43, Section 4] analyzed greedy coordinate descent by considering strong convexity with respect to the ℓ_1-norm, and formulated the strong convexity parameter as an optimization problem [43, Appendix 4.1]. More generally, D'Aspremont et al. [16, Section 2] extended smoothness and strong convexity with respect to the gauge $\gamma_{\mathcal{P}}$, which led to formulations of the smoothness parameter as an optimization

problem in the work of Sun and Bach [54, Section 2.5]. These optimization problems are often hard to solve (yet, they have closed-form reformulation in some cases).

The main idea of this work is to propose a principled view on gradient boosting methods, that are obtained by minimizing a smoothness upper bound with respect to the ℓ_1-norm. This methodology leads to a new boosting strategy for regularized problems, that benefits from (sub)linear convergence properties. Unlike former methods, such as orthogonal matching pursuit under restricted isometry property (RIP) by [62], convergence analysis is performed without statistical assumptions on the data. Convergence guarantees let appear parameters characterizing the class of functions and the geometries of optimization problems (1.1) in $\mathbb{R}^d$ and (1.3) in $\mathbb{R}^n$, but remain mostly intractable. To this end, we compute *a priori* refined estimates of convergence rates for boosting methods applied to a particular least-squares problem. We develop two approaches for computing on the one hand exact estimates using SDP relaxations [23], and on the other hand random approximations using random matrix theory. As a result, we observe a phase transition in the convergence rate of gradient descent (resp. coordinate descent), depending on (n, d). Surprisingly, we conclude that for a fixed number of samples n, adding features (dimension d) improves their convergence, which may be compared to the double descent phenomenon [8] for the generalization error. Building on these results, we experimentally highlight a transition phase for the proximal gradient and regularized matching pursuit on a LASSO problem, depending on the value for λ. Finally, we define an 'ultimate method' converging linearly both in the underparametrized ($n \gg d$) and in the overparametrized ($n \ll d$) regime, that is nonetheless not a boosting method (it may indeed add more than one atom per iteration).

1.1. Prior works. Boosting algorithms. Boosting strategies, also known as matching pursuit in signal processing, have been initiated in the context of sparse recovery [34], and extended to the fitting of weak-learners with 'gradient boosting' techniques such as Adaboost by Freund et al. [21]. Matching pursuit (MP) algorithms produce sparse combinations of atoms by picking a direction from a set of atoms using information on the gradient. Boosting algorithms are suited to both constrained models, with for example orthogonal matching pursuit [51, 58, 62] or greedy algorithms [56], as well as to unconstrained (penalized) optimization problems, with for example the vanilla boosting strategy of Zhang et al. [63], that minimizes a well-chosen quadratic upper-bound. Recently, Locatello et al. [32] have unified the framework for matching pursuit and Frank-Wolfe algorithms [20] leading to non-statistical convergence guarantees for matching pursuit.

Coordinate descent. Coordinate descent has gained interest due to the increasing access to large amounts of data, and thereby to the use of large-scale optimization models. Tseng [59] opened the path to convergence guarantees for proximal coordinate descent on composite minimization problems [60]. Nesterov [39] derived the first coordinate gradient descent method with global guarantees for convex objectives, paving the way to families of randomized coordinate updates [49], and greedy updates [7]. Yet, these analyses often lead to dimension-dependent convergence guarantees. Nutini et al. [43] provided the first convergence guarantee of greedy coordinate descent (or coordinate descent with Gauss-Southwell rule) without dependence in the dimension, formulating the update as the minimization of a smoothness upper bound with respect to the ℓ_1-norm. More precisely, they showed a significantly better performance of

Sinc de ça ? Tseng avant ?
Check ref. dans papier Yuri ?

Aussi : si g est matrice entrainement

greedy coordinate descent compared to randomized coordinate descent. However, the analysis did not extend well to proximal coordinate descent, letting a dependence in the dimension appear in the convergence bound. This led to refined techniques such as the greedy update of Karimireddy [30], with dimension-independent convergence guarantees. Finally, these methods often present the benefit of an induced sparsity, that can be linked to the ℓ_1-norm. Locatello et al. [32] interpreted steepest coordinate descent as a matching pursuit algorithm, where the atoms corresponds to the unitary directions. More precisely, steepest coordinate descent may be seen as the minimization of a smoothness upper bound with respect to the ℓ_1-norm. Considering gauge functions, coordinate descent can be extended to producing solutions sparse with respect to atoms, as Sun and Bach [54] did with the generalized conditional gradient method [3].

Approximate convergence guarantees. Sparse optimization often reveals a gap between theoretical convergence guarantees and observed behaviors. The LASSO has been widely studied for statistical recovery. From an optimization point of view, most of the analyses depend on the statistical recovery efficiency. For constrained optimization problems, Zhang [62] proposed a forward-backward greedy algorithm for which he derived convergence guarantees under restricted isometry properties (RIP). Similarly, Agarwal et al. [1] analyzed the proximal gradient and the projected gradient under restricted strong convexity and smoothness, that comes directly from restricted eigenvalue conditions [48], that appear for example for random Gaussian matrices. A recent focus on average-case analysis of optimization methods under random matrices was initiated by Pedregosa et Scieur [46], coming from the convergence analysis of the simplex method [11, 53]. On the contrary, other works improved global convergence guarantees considering well-chosen geometries. For separable quadratics, Nutini et al. [43, Section 4.1] have computed explicitly the strong convexity parameter in the ℓ_1-geometry. Generalizing unitary atoms from the ℓ_1-geometry to atoms, Sun et Bach [54, Section 2.5] formulated smoothness and strong convexity with respect to gauge functions as optimization problems. However in most cases, since these parameters are hard to compute, both strong convexity and smoothness parameters remains formulated in the ℓ_2-norm. This often leads to additional terms in convergence guarantees, coming from the norm equivalence [43, Appendix 4] or from the geometry such as the pyramidal width [31] or the directional width [32] in Frank-Wolfe techniques, or to the Hoffman constant [26] for linear mappings with strongly convex functions [37, 29, 25].

1.2. Assumptions. Convex optimization framework. In this work, functions f into consideration are convex, differentiable and admit at least one global minimizer $x_\star \in \mathbb{R}^n$. Functions $F(\cdot) = f(P\cdot) : \mathbb{R}^d \to \mathbb{R}$ benefit from the same properties. We restrict ourselves to the analysis of first-order methods, made of linear combinations of past iterates and gradients.

In this paper, functions f may be smooth with respect to a generic norm $\|\cdot\|_{\mathbb{R}^n}$, if they verify for all $x, y \in \mathbb{R}^n$,

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L^f}{2}\|y - x\|^2_{\mathbb{R}^n}. \tag{1.4}$$

Functions $F(\cdot) = f(P\cdot)$ are therefore smooth with respect for any norm $\|\cdot\|_{\mathbb{R}^d}$ with $L^F \leqslant L^f L^P$, where L is defined such that for all $\alpha, \beta \in \mathbb{R}^d$, $\|P(\alpha - \beta)\|^2_{\mathbb{R}^n} \leqslant L^P \|\alpha - \beta\|^2_{\mathbb{R}^d}$, that is

$L^{\mathcal{P}} = \sup_{\|\beta\|_{\mathbb{R}^d} \leqslant 1} \|P\beta\|^2_{\mathbb{R}^n}$. For least-squares, functions F are exactly smooth with $L^F = L^f L^{\mathcal{P}}$. In addition, functions f are strongly convex with respect to a norm $\|\cdot\|_{\mathbb{R}^n}$, if for all $x, y \in \mathbb{R}^n$,

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu^f}{2} \|y - x\|^2_{\mathbb{R}^n}. \tag{1.5}$$

Functions $F(\cdot) = f(P\cdot)$ do not always inherit strong convexity. For example, for least-squares, functions F are not strongly convex as soon as the number of samples n is lower than the dimension d. The 'natural' strong convexity parameter of functions F is given by $\mu^F = \mu^f \mu^{\mathcal{P}}$, with $\mu^{\mathcal{P}} = \inf_{\|\beta\|_{\mathbb{R}^d} \geqslant 1} \|P\beta\|^2_{\mathbb{R}^n}$ and may indeed be zero. As we will see in Section 2.4, F however inherits the Łojasiewicz property with parameter $\mu^{L,F} > 0$, such that for all $\beta \in \mathbb{R}^d$,

$$\frac{1}{2} \|\nabla F(\beta)\|^2_{\mathbb{R}^d,\star} \geqslant \mu^{L,F}(F(\beta) - F_\star). \tag{1.6}$$

Random matrices. A part of this work is devoted to approximating the strong convexity and smoothness parameters of f and F. We consider on the one hand relaxed formulations for strong convexity and smoothness parameters with respect to the data (i.e., geometry) P. On the other hand, we propose random estimates of these parameters, relying on random matrix theory. Random matrices often appears in statistical assumptions, such as with restricted isometry property [14] or the restricted eigenvalue condition [48]. In the machine learning literature, random matrices appear in average-case analysis for quadratics [46] with the Marchenko-Pastur distribution [35], or when studying the double descent phenomena for the generalization error [8, 36, 4] for Gaussian data. Most of the time, these analyses let two regimes appear, depending (among others) on the number of samples n and the dimension d. Throughout this work, we thus consider two regimes depending on the linear mapping structure $P \in \mathbb{R}^{n \times d}$: the *underparametrized* (respectively *overparametrized*) regime, characterized by matrices $P \in \mathbb{R}^{n \times d}$ for which $n \geqslant d$ (resp. $d \geqslant n$) and $P^\top P$ (resp. $PP^\top$) is invertible. Note that the invertibility of $PP^\top$ (resp. $P^\top P$) in the overparametrized (resp. underparametrized) regime can be obtained by adding sufficiently random noise. More assumptions on P and $P^\top P$ will be made across this study.

2. A phase transition for linear regression. We begin with the study of a linear regression problem, where problem (2.1) is a special case of the optimization Problem (1.1) with $\lambda = 0$,

$$\min_{\alpha \in \mathbb{R}^d} \frac{1}{2n} \|P\alpha - y\|_2^2 = F(\alpha) = f(P\alpha), \tag{2.1}$$

where $P \in \mathbb{R}^{n \times d}$, and n, d respectively denotes the number of samples and the dimension.

In this section, we focus on describing the convergence regimes of gradient descent in the ℓ_2-geometry and coordinate descent with the Gauss-Southwell (GS) rule [29, 43] in the ℓ_1-geometry. More precisely, we interpret gradient descent and coordinate descent as the minimizers of smoothness upper bound with respect to well-chosen norms, which leads to linear convergence both in the underparametrized and overparametrized regime. These linear guarantees let smoothness and strong convexity parameters appear, that we formulate as optimization problems in the geometry into account. To ease the computation of these parameters,

we derive numerical estimates. A first technique developed in this work is based on an SDP relaxation, and leads to deterministic estimates. A second technique, inspired from statistical assumptions and average-case analysis, leads to random estimates using random matrices. These random estimates let a phase transition appear between the underparametrized and overparametrized regimes, that we illustrate in particular in a random feature experiment. Finally, we interpret coordinate descent as a matching pursuit algorithm depending on the geometry P.

First, let us formulate smoothness and strong convexity parameters with respect to a generic norm for the least-squares minimization (2.1) as optimization problems. In this context, f is $\frac{1}{n}$-smooth $\frac{1}{n}$-strongly convex with respect to the norm $\|\cdot\|_2$. Thus, F is L^F-smooth with respect to an arbitrary norm $\|\cdot\|$ in $\mathbb{R}^d$, with $L^F = \frac{1}{n}\sup_{\|\beta\|^2\leqslant 1}\|P\beta\|_2^2$. In addition, the function is (possibly) μ^F-strongly convex, with a parameter μ^F explicited in Lemma 2.1 and possibly equal to 0 (especially when dimension $d < n$).

Lemma 2.1. *Let $F = \frac{1}{n}\|P\alpha - y\|_2^2$, where $P \in \mathbb{R}^{n\times d}$. Then, F is μ^F-strongly convex with respect to a norm $\|\cdot\|$ with,*

$$\mu^F = \frac{1}{n}\inf_{\|\beta\|^2\geqslant 1}\|P\beta\|_2^2 \qquad \text{and} \qquad \frac{1}{\mu^F} = n\sup_{\|P\beta\|^2\leqslant 1}\|\beta\|_2^2.$$

Proof. Let us recall the definition for strong convexity (1.5), for all $\alpha, \nu \in \mathbb{R}^d$, $F(\alpha) \geqslant F(\nu) + \langle \nabla F(\nu), \alpha - \nu\rangle + \frac{\mu^F}{2}\|\alpha - \nu\|^2$. Since F is a quadratic, the left-hand side of the inequality can be rephrased into, for all $\beta \in R^d$, $\|P\beta\|_2^2 \geqslant \mu^F\|\beta\|^2$, from which both formulations follow. ∎

In Lemma 2.1, we formulate μ^F, the strong convexity parameter for F, as a nonconvex minimization problem, with a convex objective and concave constraints. Such a problem is usually costly to solve. The function F also verifies the Łojasiewicz inequality (1.6) with $\mu^{L,F}$. Again $\mu^{L,F}$ is formulated as an optimization problem.

Lemma 2.2. *Let $F = \frac{1}{n}\|P\alpha - y\|_2^2$, where $P \in \mathbb{R}^{n\times d}$. Then, F verifies the Łojasiewicz inequality (1.6) with respect to a (dual) norm $\|\cdot\|_\star$, with*

$$\mu^{L,F} = \frac{1}{n}\inf_{\|P\beta\|_2^2\geqslant 1}\|P^\top P\beta\|_\star^2 \qquad \text{and} \qquad \frac{1}{\mu^{L,F}} = n\sup_{\|P^\top P\beta\|_\star^2\leqslant 1}\|P\beta\|_2^2.$$

Proof. The proof follows from the Łojasiewicz inequality. Given that $y = P\alpha_\star$, where $\alpha_\star$ is an optimal point for F, we have for all $\alpha \in \mathbb{R}^d$: $\|\nabla F(\alpha)\|_\star^2 = \frac{1}{n}\|P^\top P(\alpha - \alpha_\star)\|_\star$ and $F(\alpha) - F_\star = \frac{1}{2n}\|P(\alpha - \alpha_\star)\|_2^2$. Then, for all $\beta \in \mathbb{R}^d$, $\|P^\top P(\alpha - \alpha_\star)\|_\star \geqslant \mu^{L,F}\|P\beta\|_2^2$. ∎

Again in Lemma 2.2, $\mu^{L,F}$ is formulated as a (nonconvex) minimization problem. The two quantities μ^F and $\mu^{L,F}$ are compared in Lemma 2.3, with equality in the underparametized regime in which $P^\top P$ is invertible.

Lemma 2.3. *Let $F = \frac{1}{2n}\|P\alpha - y\|_2^2$. Then, we have that $\mu^{L,F} \geqslant \mu^F$ for μ^F (resp. $\mu^{L,F}$) defined in Lemma 2.1 (resp. Lemma 2.2). If $P^\top P$ is invertible, $\mu^{L,F} = \mu^F$.*

Proof. Let us consider the squared-root formulations of μ^F and $\mu^{L,F}$ given in Lemma 2.1 and Lemma 2.2.

$$\frac{1}{\sqrt{n\mu^F}} = \sup_{\|P\beta\|_2\leqslant 1} \|\beta\| = \sup_{\|z\|_\star\leqslant 1, \|P\beta\|_2\leqslant 1} \langle \beta, z\rangle,$$

$$\frac{1}{\sqrt{n\mu^{L,F}}} = \sup_{\|P^\top P\nu\|_\star\leqslant 1} \|P\nu\|_2 = \sup_{\|P^\top P\nu\|_\star\leqslant 1, \|P\beta\|_2\leqslant 1} \langle P\beta, P\nu\rangle = \sup_{\|P^\top P\nu\|_\star\leqslant 1, \|P\beta\|_2\leqslant 1} \langle \beta, P^\top P\nu\rangle.$$

Since $\mathrm{Im}(P^\top P)\subset\mathbb{R}^d$, we have $\frac{1}{\sqrt{n\mu^F}} \geqslant \frac{1}{\sqrt{n\mu^{L,F}}}$, and therefore $\mu^{F,L}\geqslant\mu^F$. In the special case where $P^\top P$ is invertible, $\mathrm{Im}(P^\top P)=\mathbb{R}^d$, and $\mu^{F,PL}=\mu^F$. ∎

In the next sections, we see the role of these parameters in the convergence guarantees of gradient descent and steepest coordinate descent, both in the underparametrized and overparametrized regime. To ease the computation of μ^F and $\mu^{L,F}$, we propose exact estimates of these parameters, as well as concentration results based on a simple random model for P.

2.1. Gradient descent in the ℓ_2-geometry. We are interested in the (approximate) convergence of gradient descent in the underparametrized and the overparametrized regimes. Assume $\mathbb{R}^d$ is equipped with the ℓ_2-norm. The function F is convex, L_2^F-smooth with respect to the norm ℓ_2, with $L_2^F=\frac{1}{n}\lambda_{\max}(P^\top P)$. A common interpretation of gradient descent with fixed step size $\gamma=\frac{1}{L_2^F}$ comes from the minimization of the quadratic smoothness upper bound on F:

$$\alpha_1 = \alpha_0 - \frac{1}{L_2^F}\nabla F(\alpha_0) = \alpha_0 - \frac{1}{L_2^F}P^\top(P\alpha_0 - y). \tag{2.2}$$

In the underparametrized regime, the function F is μ_2^F-strongly convex with respect to the ℓ_2-norm, with $\mu_2^F=\lambda_{\min}(\frac{P^\top P}{n})>0$. As a result, gradient descent (2.2) converges linearly. However, in the overparametrized regime in which $d\geqslant n$, $\mu_2^F=0$, F is not strongly convex. Yet, gradient descent still converges linearly [10], since quadratics benefit from the Łojasiewicz inequality, with $\mu_2^{L,F}=\frac{1}{n}\lambda_{\min}(PP^\top)>0$.

Proposition 2.4. *Let F be convex, L_2^F-smooth with respect to the norm $\|\cdot\|_2$, be μ_2^F-strongly convex and verify a Łojasiewicz inequality with parameter $\mu_2^{L,F}$, with $0\leqslant\mu_2^F\leqslant\mu_2^{L,F}\leqslant L_2^F$. Let $(\alpha_k)_{k\in\mathbb{N}}$ be generated by gradient descent in (2.2) starting from $\alpha_0\in\mathbb{R}^d$. The sequence verifies:*

$$F(\alpha_k) - F_\star \leqslant \left(1 - \frac{\max(\mu_2^F, \mu_2^{L,F})}{L_2^F}\right)^k (F(\alpha_0) - F_\star),$$

where $\mu_2^F=\lambda_{\min}(PP^\top/n)$, $\mu_2^{L,F}=\lambda_{\max}(P^\top P/n)$ and $L_2^F=\lambda_{\max}(P^\top P/n)$

Proof. See appendix A. ∎

The convergence speeds obtained in Proposition 2.4 depends on the maximal and minimal eigenvalues of $P^\top P$ and $PP^\top$. In the case where P is generated randomly, we can derive estimates of these extremal eigenvalues, avoiding a full computation of the extremal eigenvalues,

2.2. Gauss-Southwell coordinate descent in the ℓ_1-geometry. Similar to gradient descent, we study convergence guarantees of coordinate descent based on the Gauss-Southwell (GS) rule. The GS-rule can be obtained from the smoothness upper bound with respect to the ℓ_1-norm, as shown by Nutini et al. [43, Section 4]. For all $\alpha_0, \alpha \in \mathbb{R}^d$,

$$F(\alpha) \leqslant F(\alpha_0) + \langle \nabla F(\alpha_0), \alpha - \alpha_0 \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_0\|_1^2. \tag{2.3}$$

From this inequality, we compute $L_1^F = \frac{1}{n} \max_{\alpha \in \mathbb{R}^d, \|\alpha\|_1 = 1} \|Pz\|_2^2 = \frac{1}{n} \max_{i=1,\dots,d} \|P_{:,i}\|_2^2$ (the maximization problem attains its optimum on the extremal point of the simplex). Gauss-Southwell coordinate descent follows by minimizing over $\alpha \in \mathbb{R}^d$, for a fixed $\alpha_0 \in \mathbb{R}^d$,

$$\begin{aligned} i_0 &= \operatorname*{arg\,max}_{k=1,\dots,d} |\nabla_{i_k} F(\alpha_0)|, \\ \alpha_1 &= \alpha_0 - \frac{1}{L_1^F} \nabla_{i_0} F(\alpha_0) e_{i_0}. \end{aligned} \tag{2.4}$$

As for gradient descent, its convergence speed depends on the parametrization regime. Depending on the (n, d), F may be μ_1^F-strongly convex, or verify the Łojasiewiczinequality with parameter $\mu_1^{L,F}$. Both μ_1^F and $\mu_1^{L,F}$ can be formulated as optimization problems. It follows from the strong convexity characterization given in Lemma 2.1 with the norm $\|\cdot\|_1$, that $\mu_1^F = \frac{1}{n} \inf_{\|z\|_1^2 \geqslant 1} \|Pz\|_2^2$, and from Lemma 2.2 with norm $\|\cdot\|$ for the Łojasiewicz inequality $\mu_1^{L,F} = \inf_{\|P\beta\|_2^2 \geqslant 1} \|P^\top P\beta\|_\infty^2$.

In the regimes under consideration, Proposition 2.7 states that coordinate descent converges linearly, as already proven by Karimi et al. [29, Theorem 1].

Proposition 2.7. *[29, Theorem 1] Let F be convex, L_1^F-smooth with respect to the norm $\|\cdot\|_1$, be μ_1^F-strongly convex and be $\mu_1^{L,F}$-PL with $0 \leqslant \mu_1^F \leqslant \mu_1^{L,F} \leqslant L_1^F$. Let (α_k) be generated by coordinate gradient descent (2.4) starting from $\alpha_0 \in \mathbb{R}^d$. The sequence verifies:*

$$F(\alpha_k) - F_\star \leqslant \left(1 - \frac{\max(\mu_1^F, \mu_1^{L,F})}{L_1^F}\right)^k (F(\alpha_0) - F_\star)$$

Proof. See Appendix A. ∎

The convergence guarantee provided in Proposition 2.7 is however not easily computable. Although $L_1^F = \frac{1}{n} \max_{i=1,\dots,d} \|P_{:,i}\|_2^2$ has a closed-form solution, $\mu_1^F = \frac{1}{n} \inf_{\|z\|_1^2 \geqslant 1} \|Pz\|_2^2$ and $\mu_1^{L,F} = \inf_{\|P\beta\|_2^2 \geqslant 1} \|P^\top P\beta\|_\infty^2$ are formulated as nonconvex minimization problems. We construct estimates to these quantities, so that they may be computed a priori.

SDP relaxations. Building on the formulation of μ_1^F and $\mu_1^{L,F}$ as optimization problems, we rephrase them into relaxed SDPs.

Proposition 2.8. *Gauss-Southwell coordinate descent's linear convergence rate has the following approximations*

- *in the underparametrized regime,* $\frac{1}{\tilde{n}\tilde{\mu}_1^F} = \sup_{X \succcurlyeq 0} \operatorname{Tr}((P^\top P)^{-1} X)$, s.t. $\operatorname{diag}(X) \leqslant 1$,

$$1 - \frac{\pi}{2} \frac{\tilde{\mu}_1^F}{L_1^F} \leqslant 1 - \frac{\mu_1^F}{L_1^F} \leqslant 1 - \frac{\tilde{\mu}_1^F}{L_1^F},$$

- *in the overparametrized regime,* $\frac{1}{n\tilde{\mu}_1^{L,F}} = \sup_{X \succcurlyeq 0} \operatorname{Tr}(P^\top P X) \text{ s.t. } \|P^\top P X P^\top P\|_\infty \leqslant 1$,

$$1 - \frac{\mu_1^{L,F}}{L_1^F} \leqslant 1 - \frac{\tilde{\mu}_1^{L,F}}{L_1^F},$$

where $L_1^F = \frac{1}{n}\max_{i=1,\dots,d} \|P_{:,i}\|_2^2$. *In addition, we still have that* $\tilde{\mu}_1^F \leqslant \tilde{\mu}_1^{PL}$.

Proof. See Appendix B.1. ∎

In Proposition 2.8, we find out SDP relaxations that yield an exact approximation for μ_1^F, and an exact lower bound for $\mu_1^{L,F}$. Yet, the larger n, d, the longer the computation of these SDPs.

Random estimates. We now assume that P is randomly generated, as in the ℓ_2-geometry. Under subgaussian assumptions, we derive in Proposition 2.9 random estimates of μ_1^F, $\mu_1^{L,F}$ and L_1^F. More precisely, we prove that L_1^F concentrates around the variance σ^2, μ_1^F concentrates around $\frac{\sigma^2}{d}$ and $\mu_1^{L,F}$ concentrates around $\frac{\sigma^2}{n}$ with subgaussian tails.

Proposition 2.9. *Let* $P \in \mathbb{R}^{n\times d}$, *with* $P_i \in \mathbb{R}^d$ *i.i.d. subgaussian such that* $\mathbb{E}[P_{i,j}] = 0$, $\mathbb{E}[P_{i,j}] = \sigma^2$. *There exists* $C, C_1, C_2, C_3, C_4 > 0$ *absolute constants such that,*

- *For all* $u \geqslant 2K^2\sqrt{\frac{C_1\log(d)}{n}}$,

$$\left(1 + C_2K^2\frac{1}{\sqrt{n}} - t\right)^2 \leqslant \frac{L_1^F}{\sigma^2} \leqslant \left(1 + 2K^2\sqrt{\frac{C_1\log(d)}{n}} + t\right)^2,$$

holds with probability $1 - e^{-\frac{C}{\sigma^2K^4}\min(u_1(t),u_2(t))}$ *where* $u_1(t) = \log(d)\sigma^2(t + \frac{C_2K^2}{\sqrt{n}})^2$ *and* $u_2(t) = d\sigma^2(t - 2K^2\sqrt{\frac{C_1\log(d)}{n}})^2$.

- *For all* $t \geqslant 0$, *it holds with probability* $1 - 2\exp(-t^2)$,

$$\left(1 - C_3K^2\left(\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}\right)\right)^2 \leqslant \mu_1^F\frac{d}{\sigma^2} \leqslant \left(1 + C_3K^2\left(\sqrt{\frac{1}{n}} + \frac{t}{\sqrt{dn}}\right)\right)^2.$$

- *For all* $t \geqslant 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}}$, *it holds with probability* $1 - 2\exp(-\min(t^2, u_2(t))$,

$$\left(1 - C_4K^2\left(\sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}}\right)\right)^2 \leqslant \mu_1^{L,F}\frac{n}{\sigma^2} \leqslant \left(1 + 2K^2\sqrt{\frac{C_1\log(d)}{n}} + t\right)^2.$$

The constant $K > 0$ *characterizes subgaussian vectors of* P *(and defined in Appendix B.2).*

Proof. See Appendix B.2. ∎

Compared with Proposition 2.4, Proposition 2.9 provides concentration inequalities for L_1^F, μ_1^F and $\mu_1^{L,F}$ depending on dimension d, the variance σ^2, the number of samples n and absolute constants. In the overparametrized regime (resp. underparametrized), we conclude with limiting concentration of the convergence rate for large dimensions (resp. large number of samples).

2.2. Gauss-Southwell coordinate descent in the ℓ_1-geometry. Similar to gradient descent, we study convergence guarantees of coordinate descent based on the Gauss-Southwell (GS) rule. The GS-rule can be obtained from the smoothness upper bound with respect to the ℓ_1-norm, as shown by Nutini et al. [43, Section 4]. For all $\alpha_0, \alpha \in \mathbb{R}^d$,

$$F(\alpha) \leqslant F(\alpha_0) + \langle \nabla F(\alpha_0), \alpha - \alpha_0 \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_0\|_1^2. \tag{2.3}$$

From this inequality, we compute $L_1^F = \frac{1}{n} \max_{\alpha \in \mathbb{R}^d, \|\alpha\|_1 = 1} \|Pz\|_2^2 = \frac{1}{n} \max_{i=1,\dots,d} \|P_{:,i}\|_2^2$ (the maximization problem attains its optimum on the extremal point of the simplex). Gauss-Southwell coordinate descent follows by minimizing over $\alpha \in \mathbb{R}^d$, for a fixed $\alpha_0 \in \mathbb{R}^d$,

$$\begin{aligned} i_0 &= \arg\max_{k=1,\dots,d} |\nabla_{i_k} F(\alpha_0)|, \\ \alpha_1 &= \alpha_0 - \frac{1}{L_1^F} \nabla_{i_0} F(\alpha_0) e_{i_0}. \end{aligned} \tag{2.4}$$

As for gradient descent, its convergence speed depends on the parametrization regime. Depending on the (n, d), F may be μ_1^F-strongly convex, or verify the Łojasiewiczinequality with parameter $\mu_1^{L,F}$. Both μ_1^F and $\mu_1^{L,F}$ can be formulated as optimization problems. It follows from the strong convexity characterization given in Lemma 2.1 with the norm $\|\cdot\|_1$, that $\mu_1^F = \frac{1}{n} \inf_{\|z\|_1^2 \geqslant 1} \|Pz\|_2^2$, and from Lemma 2.2 with norm $\|\cdot\|$ for the Łojasiewicz inequality $\mu_1^{L,F} = \inf_{\|P\beta\|_2^2 \geqslant 1} \|P^\top P\beta\|_\infty^2$.

In the regimes under consideration, Proposition 2.7 states that coordinate descent converges linearly, as already proven by Karimi et al. [29, Theorem 1].

Proposition 2.7. *[29, Theorem 1] Let F be convex, L_1^F-smooth with respect to the norm $\|\cdot\|_1$, be μ_1^F-strongly convex and be $\mu_1^{L,F}$-PL with $0 \leqslant \mu_1^F \leqslant \mu_1^{L,F} \leqslant L_1^F$. Let (α_k) be generated by coordinate gradient descent (2.4) starting from $\alpha_0 \in \mathbb{R}^d$. The sequence verifies:*

$$F(\alpha_k) - F_\star \leqslant \left(1 - \frac{\max(\mu_1^F, \mu_1^{L,F})}{L_1^F}\right)^k (F(\alpha_0) - F_\star)$$

Proof. See Appendix A. ∎

The convergence guarantee provided in Proposition 2.7 is however not easily computable. Although $L_1^F = \frac{1}{n} \max_{i=1,\dots,d} \|P_{:,i}\|_2^2$ has a closed-form solution, $\mu_1^F = \frac{1}{n} \inf_{\|z\|_1^2 \geqslant 1} \|Pz\|_2^2$ and $\mu_1^{L,F} = \inf_{\|P\beta\|_2^2 \geqslant 1} \|P^\top P\beta\|_\infty^2$ are formulated as nonconvex minimization problems. We construct estimates to these quantities, so that they may be computed a priori.

SDP relaxations. Building on the formulation of μ_1^F and $\mu_1^{L,F}$ as optimization problems, we rephrase them into relaxed SDPs.

Proposition 2.8. *Gauss-Southwell coordinate descent's linear convergence rate has the following approximations*

- *in the underparametrized regime,* $\frac{1}{n\tilde{\mu}_1^F} = \sup_{X \succcurlyeq 0} \operatorname{Tr}((P^\top P)^{-1} X)$, s.t. $\operatorname{diag}(X) \leqslant 1$,

$$1 - \frac{\pi}{2} \frac{\tilde{\mu}_1^F}{L_1^F} \leqslant 1 - \frac{\mu_1^F}{L_1^F} \leqslant 1 - \frac{\tilde{\mu}_1^F}{L_1^F},$$

- *in the overparametrized regime,* $\frac{1}{n\tilde{\mu}_1^{L,F}} = \sup_{X \succcurlyeq 0} \operatorname{Tr}(P^\top P X) \text{ s.t. } \|P^\top P X P^\top P\|_\infty \leqslant 1$,

$$1 - \frac{\mu_1^{L,F}}{L_1^F} \leqslant 1 - \frac{\tilde{\mu}_1^{L,F}}{L_1^F},$$

where $L_1^F = \frac{1}{n}\max_{i=1,\dots,d} \|P_{:,i}\|_2^2$. *In addition, we still have that* $\tilde{\mu}_1^F \leqslant \tilde{\mu}_1^{PL}$.

Proof. See Appendix B.1. ∎

In Proposition 2.8, we find out SDP relaxations that yield an exact approximation for μ_1^F, and an exact lower bound for $\mu_1^{L,F}$. Yet, the larger n, d, the longer the computation of these SDPs.

Random estimates. We now assume that P is randomly generated, as in the ℓ_2-geometry. Under subgaussian assumptions, we derive in Proposition 2.9 random estimates of μ_1^F, $\mu_1^{L,F}$ and L_1^F. More precisely, we prove that L_1^F concentrates around the variance σ^2, μ_1^F concentrates around $\frac{\sigma^2}{d}$ and $\mu_1^{L,F}$ concentrates around $\frac{\sigma^2}{n}$ with subgaussian tails.

Proposition 2.9. *Let* $P \in \mathbb{R}^{n\times d}$, *with* $P_i \in \mathbb{R}^d$ *i.i.d. subgaussian such that* $\mathbb{E}[P_{i,j}] = 0$, $\mathbb{E}[P_{i,j}] = \sigma^2$. *There exists* $C, C_1, C_2, C_3, C_4 > 0$ *absolute constants such that,*

- *For all* $u \geqslant 2K^2\sqrt{\frac{C_1\log(d)}{n}}$,

$$\left(1 + C_2K^2\frac{1}{\sqrt{n}} - t\right)^2 \leqslant \frac{L_1^F}{\sigma^2} \leqslant \left(1 + 2K^2\sqrt{\frac{C_1\log(d)}{n}} + t\right)^2,$$

holds with probability $1 - e^{-\frac{C}{\sigma^2K^4}\min(u_1(t),u_2(t))}$ *where* $u_1(t) = \log(d)\sigma^2(t + \frac{C_2K^2}{\sqrt{n}})^2$ *and* $u_2(t) = d\sigma^2(t - 2K^2\sqrt{\frac{C_1\log(d)}{n}})^2$.

- *For all* $t \geqslant 0$, *it holds with probability* $1 - 2\exp(-t^2)$,

$$\left(1 - C_3K^2\left(\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}\right)\right)^2 \leqslant \mu_1^F\frac{d}{\sigma^2} \leqslant \left(1 + C_3K^2\left(\sqrt{\frac{1}{n}} + \frac{t}{\sqrt{dn}}\right)\right)^2.$$

- *For all* $t \geqslant 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}}$, *it holds with probability* $1 - 2\exp(-\min(t^2, u_2(t))$,

$$\left(1 - C_4K^2\left(\sqrt{\frac{n}{d}} + \frac{t}{\sqrt{d}}\right)\right)^2 \leqslant \mu_1^{L,F}\frac{n}{\sigma^2} \leqslant \left(1 + 2K^2\sqrt{\frac{C_1\log(d)}{n}} + t\right)^2.$$

The constant $K > 0$ *characterizes subgaussian vectors of* P *(and defined in Appendix B.2).*

Proof. See Appendix B.2. ∎

Compared with Proposition 2.4, Proposition 2.9 provides concentration inequalities for L_1^F, μ_1^F and $\mu_1^{L,F}$ depending on dimension d, the variance σ^2, the number of samples n and absolute constants. In the overparametrized regime (resp. underparametrized), we conclude with limiting concentration of the convergence rate for large dimensions (resp. large number of samples).

Corollary 2.10. *Under the assumption of Proposition 2.9, if in addition $n, d \to \infty$, coordinate descent with GS rule (2.4) converges with a limiting rate,*

- *in the underparametrized regime, when $n \to \infty$, the convergence guarantee $1 - \frac{\mu_1^F}{L_1^F}$ concentrates in $1 - \frac{1}{d} + O(\frac{1}{\sqrt{n}})$ with subgaussian tails,*
- *in the overparametrized regime, when $d \to \infty$ and $\frac{\log(d)}{n} \to 0$, the convergence guarantee $1 - \frac{1}{\mu_1^{L,F}}$ concentrates in $1 - \frac{1}{n} + O(\frac{1}{\sqrt{d}}) + O(\sqrt{\frac{\log(d)}{n}})$ with subgaussian tails.*

Proof. See the proof in Appendix B.4. ∎

For large overparametrized models (resp. underparametrized), the convergence guarantee of coordinate descent with GS rule concentrates to $(1 - \frac{1}{n})$ (resp. $(1 - \frac{1}{d})$), that is independent of the dimension d (resp. of the number of samples). Note that the condition $\log(d) \ll n$ is indeed reasonable, since e^n grows quickly (when $n = 50$, $e^n \approx 5 \times 10^{21}$). A numerical comparison for the expected and exact lower bounds for μ_1^F and $\mu_1^{L,F}$ is provided in SM2. Unlike the approximate convergence guarantees for gradient descent in the underparametrized regime (resp. overparametrized) detailed in Corollary 2.6, coordinate descent with GS-rule does not improve when adding samples (resp. features).

As for gradient descent in the ℓ_2-geometry, we have formulated coordinate descent with GS rule as the minimization of the smoothness upper bound with respect to the ℓ_1-norm, leading its linear convergence in both the underparametrized and overparametrized regime. For a linear regression problem, nor the strong convexity parameter neither the Łojasiewicz in the ℓ_1-geometry benefit from a closed-form formulation (but it did in the ℓ_2-geometry). In a first approach, we approximate these quantities by SDPs, that may take longer computation in large models (either in the number of samples or the dimension). Instead of that, we consider randomly generated matrices P to approximate these parameters. Under subgaussian data, it appears μ_1^F, $\mu_1^{L,F}$ and L_1^F concentrate to there expectation with subgaussian tails. In the next section, we perform numerical experiments that let a transition phenomenon appear.

2.3. A phase transition phenomenon: experimental results. We compare the approximated convergence guarantees to numerical experimental convergence for gradient descent from Corollary 2.6 and of coordinate descent from Propositions B.1 2.9. More precisely, we verify the expected phase transition in (n, d): in the overparametrized (respectively underparametrized) regime, the larger the dimension (resp. the number of samples), the better the convergence. To this end, we perform experiments on several datasets: several least-squares problems obtained with synthetic quadratics, quadratics from the Leukemia dataset and a random features experiment, that we describe below.

Synthetic quadratics. We consider several least-squares problems (2.1), where the number of samples $n = 50$ is fixed, and the number of dimensions varies so that both the overparametrized and the underparametrized regimes are explored. In this model, the feature matrix P into account is generated such that $P_{:,i} \sim \mathcal{N}(0, I_d)$ are i.i.d, $\alpha_\star \in \{-1, 1\}^d$ has a sparsity equal to $s = 8 < d$, and $y = P\alpha_\star + \epsilon$, where $\epsilon_i \sim \mathcal{N}(0, \sigma)$.

Leukemia dataet. We consider the standard Leukemia dataet [24], where $n = 72$ and $d = 7129$. Again, we consider submatrices, so that the dimensions vary from both the underparametrized regime to the overparametrized. For each model, P has zero mean and the

features have a variance equal to one.

Random features. Let us consider the example of random features for a fixed prediction model. We consider the regression model $\hat{a} = \arg\min_{a\in\mathbb{R}^m} \frac{1}{2n}\|y - f(P,a,\theta)\|_2^2$, where the family of models is given by $\mathcal{F}(\theta) = \{f(P,a,\theta) = \sum_{i=1}^m a_i\sigma(\langle\theta_{:,i}, P_{:,j}\rangle) = \phi_P(\theta)^\top a, a \in \mathbb{R}^N\}$, where $\theta \in \mathbb{R}^{m\times d} \sim \mathcal{N}(0,\nu^2)$, and $\sigma(\cdot) = \max(0,\cdot)$. In this experiment, we increase the number of features m (from 10 to 1000) while the initial data taken from the leukemia dataet is such that $n = 72$, $d = 200$, and $\theta_i \sim \mathcal{N}(0, I_d)$. Compared to the experiments on synthetic quadratics and on the leukemia dataet, the model does not vary in random features: all models converge to the same optimal solution y.

In Figure 1, we plot the iteration number $k(\epsilon)$ at which a certain accuracy ϵ is reached for the three models describes above, both for gradient descent and coordinate descent with GS-rule. We consider precisions $\epsilon = \{10^0, \dots, 10^{10}\}$ and we refer to these curves by ϵ-curves. In Figure 1, both steepest coordinate descent and gradient descent converge faster for the

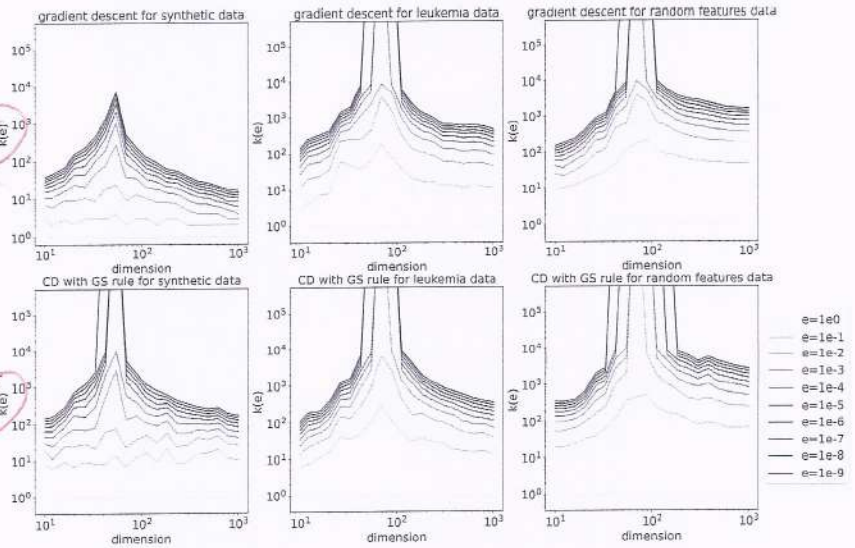

Figure 1: ϵ-curve for gradient descent (top) and coordinate descent with the GS-rule (bottom), for the three models: synthetic quadratics (on the left) with $n = 50$, the leukemia dataset (in the middle) with $n = 72$, a random feature model (on the right) with $n = 72$.

three models when $n \gg d$ (resp. $d \gg n$) in the underparametrized (resp. overparametrized) regime. For $n \approx d$, convergence slows down and tends to be sublinear, as expected from the theory for smooth convex functions. In other words, we observe a transition phenomenon for dimensions $d \approx n$. For the random feature models, a double descent phenomena was empirically highlighted by Belkin et al. [8], and formalized by Mei and Montanari [36]. For a fixed prediction model, as the number of features increase, the excess risk follows is U-shaped for un-

derparametrized optimization models and goes down for overparametrized models. As for the excess rick, we observe a transition at $m \approx n$ as well as a better precision for overparametrized models. Contrary to the generalization error, underparametrized models ($d \ll n$) performs well even when $\frac{d}{n} \to 0$ and are not U-shaped. We thus rather speak of phase transition for gradient and coordinate descent.

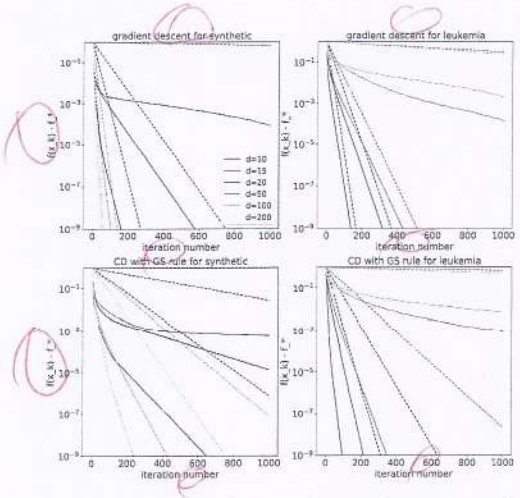

Figure 2: Convergence in function value for gradient descent and coordinate descent with GS rule, on synthetic quadratics ($n = 20$) and on the leukemia dataet ($n = 72$), for several value of dimensions. Dashed line: comparison to the exact upper bound for synthetic quadratics, and to the approximated upper bound for the leukemia dataet.

In Figure 2, we compare the exact and approximated upper bound to the convergence guarantee in function value for gradient descent and coordinate descent with the GS rule. For gradient descent, the theoretical approximation guarantee from Corollary 2.6 matches the real convergence behavior of gradient descent. For steepest coordinate descent, we compare its convergence in function values to the exact upper bound obtained from the SDP relaxation in Proposition 2.8 for 'small' values of d and n, and to its random estimate otherwise. In both cases, they do not match exactly. However, we numerically recover that convergence is improved as the dimension increases.

2.4. Coordinate descent is an instance of matching pursuit. Coordinate descent, as well as gradient descent, converges linearly both in the underparametrized and overparametrized regime, as proven in Proposition 2.7, due respectively to strong convexity and the Łojasiewicz property. Given a certain structure on f, we prove that $F(\cdot) = f(P\cdot)$ inherits some convex properties from f and that coordinate descent can be interpreted as a matching pursuit algorithm in either $\mathbb{R}^d$ or $\mathbb{R}^n$. We now go back to the more general regime,

$$\min_{\alpha \in \mathbb{R}^d} F(\alpha) = f(P\alpha),$$

where f is L^f-smooth, μ^f-strongly convex and F is L_1^F-smooth with respect to the ℓ_1-norm.

In the underparametrized regime, F is μ_1^F-strongly convex (since $P^\top P$ is invertible). The connection between coordinate descent with GS-rule (2.4) and matching pursuit was highlighted by Locatello et al [33]. Considering the set of unitary direction $\mathcal{A} = \operatorname{conv}(\{\pm e_i, i = 1, \dots, d\})$, that is the unit ball L_1, coordinate descent may be rewritten as a matching pursuit:

$$e_{i_0} \in -\operatorname{LMO}_{\mathcal{A}}(\nabla F(\alpha_0)),$$
$$\alpha_1 = \alpha_0 - \frac{1}{L_1^F} \nabla_{i_0} F(\alpha_0) e_{i_0},$$

where $\operatorname{LMO}_{\mathcal{A}}(\nabla F(\alpha_0)) = \inf_{z \in \mathbb{R}^n} \nabla F(\alpha_0)^\top z$. Steepest coordinate descent converges linearly from Proposition 2.7, with the same convergence guarantee as in the context of matching pursuit [33, Theorem 5].

In the overparametrized regime, F does not inherit strong convexity. Yet, for least-squares, F but does inherit some structure from f (see Lemma 2.2). We prove that coordinate descent can be interpreted as a matching pursuit algorithm in $\mathbb{R}^n$.

Recall the gauge function, for $x \in \mathbb{R}^n$, $\gamma_{\mathcal{P}}(x) = \inf_{\alpha \in \mathbb{R}^d, x = P\alpha} \|\alpha\|_1$. Lemma 2.11 ensures $\gamma_{\mathcal{P}}(\cdot)$ is a norm in the overparametrized regime.

Lemma 2.11. *Let $\alpha \in \mathbb{R}^d \to P\alpha \in \mathbb{R}^n$ be a surjection in $\mathbb{R}^d$, and $\mathcal{P} = \operatorname{conv}(P)$ be centrally symmetric. The function $\gamma_{\mathcal{P}}(\cdot)$ is a norm, and its dual norm is $\gamma_{\mathcal{P}}^\star(\cdot) = \sup_{s \in \mathcal{P}} \langle s, \cdot \rangle = \|P^\top \cdot\|_\infty$.*

Proof. See appendix C. ∎

Let f be convex, $L_{\mathcal{P}}^f$-smooth and $\mu_{\mathcal{P}}^f$-strongly convex with respect to the norm $\gamma_{\mathcal{P}}(\cdot)$. We compute in Lemma 2.12 its smoothness and strong convexity parameters with respect to a gauge $\gamma_{\mathcal{P}}(\cdot)$ (that is a thus a norm).

Lemma 2.12. *Let $L_{\mathcal{P}}^f$ (resp. L_2^f) be the smoothness parameter of f with respect to $\gamma_{\mathcal{P}}(\cdot)$ (resp. ℓ_2), and $\mu_{\mathcal{P}}^f$ (resp. μ_2^f) be the strong convexity of f with respect to $\gamma_{\mathcal{P}}$ (resp. ℓ_2). Then, $L_{\mathcal{P}}^f = L_2^f \sup_{j=1,\dots,d} \|P_j\|_2^2$ and $\mu_{\mathcal{P}}^f = \mu_2^f \inf_{z \in \mathbb{R}^n} \|P^\top z\|_\infty^2$, such that $\|z\|_2^2 = 1$.*

Proof. $L_{\mathcal{P}}^f$ is the tightest constant such that for all $x \in \mathbb{R}^n$, $L_2^f \|x\|_2^2 \leqslant L_{\mathcal{P}}^f \gamma_{\mathcal{P}}(x)^2$ and we have that $L_{\mathcal{P}}^f = L_2^f \sup_{u,\beta,u=P\beta,\|\beta\|_1 \leqslant 1} \|u\|_2^2 = L_2^f \sup_{\beta,\|\beta\|_1 \leqslant 1} \|P\beta\|_2^2 = L_2^f \sup_{j=1,\dots,d} \|P_j\|_2^2$. The same reasoning leads to the result for $\mu_{\mathcal{P}}^f$. ∎

In Lemma 2.12, the smoothness (resp. strong convexity) parameter is formulated as optimization problems, multiplied by the smoothness (resp. strong convexity) in the ℓ_2-norm, which are closely related to the parameters for least-squares from Lemma 2.1 and 2.2. In the context of least-squares where $f(x) = \frac{1}{2n}\|x - y\|_2^2$ for $x \in \mathbb{R}^n$, we indeed have that $L_2^f = \mu_2^f = \frac{1}{n}$. For the ℓ_1-norm, that $L_{\mathcal{P}}^f = L_1^F$ as defined in (2.3), and $\mu_{\mathcal{P}}^f = \mu_1^{L,F}$ as soon as $PP^\top$ is invertible (which is true here). Multiplying (2.4) by P, noticing that $\min_{e,\|e\|_1=1} \langle \nabla F(\alpha), e \rangle = \min_{p \in \mathcal{P}} \langle \nabla f(x), p \rangle$ and using Lemma 2.12, coordinate descent with the GS-rule on F can be formulated as match-

ing pursuit on f,

$$z_0 \in \mathrm{LMO}_{\mathcal{P}}(\nabla f(x_0)),$$
$$x_1 = x_0 - \frac{1}{L^f_{\mathcal{P}}}\langle \nabla f(x_0), z_0\rangle z_0. \tag{2.5}$$

Let x_k be generated by matching pursuit (2.5), starting from $x_0 \in \mathbb{R}^n$ for $L^f_{\mathcal{P}}$-smooth and $\mu^f_{\mathcal{P}}$-strongly convex functions, then, Locatello et al [33, Theorem 5] proved linear convergence of the sequence with

$$f(x_k) - f_\star \leqslant \left(1 - \frac{\mu^f_{\mathcal{P}}}{L^f_{\mathcal{P}}}\right)(f(x_0) - f_\star). \tag{2.6}$$

By construction, since $x_k = P\alpha_k$, we have that $F(\alpha_k) - F_\star \leqslant (1 - \frac{\mu^f_{\mathcal{P}}}{L^f_{\mathcal{P}}})(F(\alpha_0) - F_\star) = (1 - \frac{\mu_1^{L,F}}{L_1^F})(F(\alpha_0) - F_\star)$. The same result could have been derived from Proposition 2.7 and the observation that strongly convex functions composed with a linear mapping verify a Łojasiewicz-inequality.

Lemma 2.13. *Let f be $\mu^f_{\mathcal{P}}$-strongly convex with respect to the norm $\gamma_{\mathcal{P}}(\cdot)$. Then, F verifies a Łojasiewicz inequality with parameters $\mu^f_{\mathcal{P}}$, that is for all $\alpha \in \mathbb{R}^d$,*

$$\frac{1}{2}\|\nabla F(\alpha)\|_\infty \geqslant \mu^f_{\mathcal{P}}(F(\alpha) - F_\star).$$

Proof. Let $x \in \mathbb{R}^n$, $f_\star \geqslant f(x) - \sup_z \langle -\nabla f(x), y - x\rangle - \frac{\mu^f_{\mathcal{P}}}{2}\gamma^2_{\mathcal{P}}(y - x) \geqslant f(x) - (\frac{\mu^f_{\mathcal{P}}}{2}\gamma^2_{\mathcal{P}}(\cdot))^\star(-\nabla f(x)) \geqslant f(x) - \frac{1}{2\mu^f_{\mathcal{P}}}\|P^\top \nabla f(x)\|^2_\infty$. Since $F(\cdot) = f(P\cdot)$, the inequality is obtained by taking $x = P\alpha$ and since $\nabla F(\alpha) = P^\top \nabla f(P\alpha)$. ∎

Lemma 2.13 corresponds exactly to the result of Karimi et al. [29, Appendix B], that let a Hoffman constant appear (that is in their context equal to the smallest non-zero eigenvalue of P), as defined in [37, Section 3 and 4.1] by $\theta(P) = \max_{z, \|P^\top z\|_\infty = 1} \|z\|_2^2$. They indeed proved that F verifies the Łojasiewicz inequality, for all $\alpha \in \mathbb{R}^d$, $\frac{1}{2}\|\nabla F(\alpha)\|_2^2 \geqslant \theta(P)\mu^F(F(\alpha_k) - F_\star)$.

Depending on the parametrization regime, we have proven that coordinate descent may be formulated as a (possibly) rebased matching pursuit method. In the underparametrized regime on the one hand, since F inherits all convex properties from f, the atoms are defined by the Euclidean basis and the matching pursuit is formulated in $\mathbb{R}^d$. On the other hand in the overparametrized regime, the introduction of a well-chosen gauge function $\gamma_{\mathcal{P}}$ allows to formulated coordinate descent as a matching pursuit algorithm in $\mathbb{R}^n$, and to perform a convergence analysis using the strong convexity assumption on f. Again, smoothness and strong convexity can be formulated as optimization problems depending on the gauge. The gauge let also appear how F inherits some structure from f. In the next sections, we will see how to generalize this framework for analyzing penalized linear models.

3. Phase transition for penalized linear models. We now consider the penalized linear model,

$$\min_{\alpha\in\mathbb{R}^d} f(P\alpha) + \lambda\|\alpha\|_1 = F(\alpha) + H(\alpha) = G(\alpha), \tag{3.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is L_2^f-smooth, μ_2^f-strongly convex, (and thus, $F : \mathbb{R}^d \to \mathbb{R}$ is L_2^F smooth), $P \in \mathbb{R}^{n\times d}$, $\lambda > 0$ and where $H(\alpha) = \lambda\|\alpha\|_1$ is closed convex and proper. In this section, we derive a new matching pursuit algorithm for a ℓ_1-regularized model, that we compare to proximal coordinate descent with GS rule and to the proximal gradient descent. Building on the result of Section 2, we derive convergence guarantees depending on the properties of f and P, and notice a strong connection to the proximal coordinate descent with GS rule. Yet, in the overparametrized regime, neither the proximal gradient nor the regularized matching pursuit benefits from linear convergence. Instead of that, we describe experimentally the role of λ in the LASSO, as a continuous mapping between low-rank solutions and full-rank solution to the least-squares.

Proximal gradient descent. Proximal gradient descent, a.k.a. forward-backward (see e.g. [15]) was developed for such 'composite' convex optimization problem. Given a starting point $\alpha_0 \in \mathbb{R}^d$, each iterate is obtained by minimizing a smooth quadratic upper bound on F,

$$G(\alpha) \leqslant F(\alpha_k) + \langle \nabla F(\alpha_k), \alpha - \alpha_k\rangle + \frac{L_2^F}{2}\|\alpha_k - \alpha\|_2^2 + \lambda\|\alpha\|_1. \tag{3.2}$$

Minimizing the right side of the inequality leads to the proximal gradient method as follows:

$$\alpha_{k+1} = \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^d} \langle \nabla F(\alpha_k), \alpha - \alpha_k\rangle + \frac{L_2^F}{2}\|\alpha - \alpha_k\|_2^2 + \lambda\|\alpha\|_1.$$

The proximal gradient method converges sublinearly if F is not strongly convex, and linearly if F is in addition μ_2^F-strongly convex, such as in the underparametrized regime. Then, the sequence α_k starting from $\alpha_0 \in \mathbb{R}^d$ verifies ([55, Theorem 2.1]),

$$G(\alpha_k) - G_\star \leqslant \left(1 - \frac{\mu_2^F}{L_2^F}\right)^{2k} (G(\alpha_0) - G_\star).$$

Coordinate descent. In practice, (randomized) coordinate gradient descent is widely used to avoid computing the full gradient (that costs $O(d)$), and is particularly suited to sparse regression problems. Nutini et al. [43] analyzed coordinate descent with the Gauss-Southwell selection rule, that tends to perform better than randomized coordinate descent.

$$\alpha_{k+1} = \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^d} \nabla_{i_k} F(\alpha_k)(\alpha^{(i_k)} - \alpha_k^{(i_k)}) + \frac{L_2^F}{2}(\alpha^{(i_k)} - \alpha_k^{(i_k)})^2 + \lambda|\alpha^{(i_k)}|, \tag{3.3}$$

where $i_k = \arg\min_l \min_{t\in\mathbb{R}} \nabla_l F(P\alpha_k)(t - \alpha^{(l)}) + \frac{L_2^F}{2}(t - \alpha^{(l)})^2 + \lambda|t|$ corresponds to the GS rule. Nutini et al. [43, Appendix 8] proved that coordinate descent with the Gauss-Southwell rule makes at least as much progress as randomized coordinate descent,

$$G(\alpha_{k+1}) - G_\star \leqslant \left(1 - \frac{\mu_2^F}{dL_2^F}\right)(G(\alpha_k) - G_\star).$$

A refinement, that let a sublinear dependence in the parameter μ_1^F appear, is mentioned in [43, Appendix 8]. Coordinate descent with GS-rule is closely related to matching pursuit as for nonpenalized models, as detailed in Supplementary Materials SM3, where we formulate this method as a 'nearly' matching pursuit algorithm.

In the following, we derive a matching pursuit procedure for ℓ_1-regularized problems (3.1), that we compare to classical boosting algorithm and coordinate descent with GS-rule. After that, we compute convergence guarantees for smooth (possibly strong) convex functions. Finally, we propose an interpretation of the convergence regimes depending on the penalty λ.

3.1. Regularized matching pursuit. We propose a new regularized matching pursuit algorithm based on the ℓ_1-geometry. The main idea is to replace the ℓ_2-norm in the minimization Problem (3.2) leading to the proximal gradient by a ℓ_1-norm. Let F be convex, L_1^F-smooth, as for coordinate descent with GS rule in the linear regression problem from Section 2. We define the penalized matching pursuit method starting from $\alpha_0 \in \mathbb{R}^d$ as the sequence minimizing smoothness with respect to the ℓ_1-norm at each iteration:

$$\alpha_{k+1} = \arg\min_{\alpha \in \mathbb{R}^d} \langle P^\top \nabla f(P\alpha), \alpha - \alpha_k \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_k\|_1^2 + \lambda \|\alpha\|_1. \tag{3.4}$$

Whereas the optimization steps in proximal gradient descent (3.2) and proximal coordinate descent with the GS rule (3.3) can be decomposed per coordinate, the function $\alpha \to \|\alpha\|_1^2$, that is not decomposable. Based on the same upper bound (3.4), Song et al. [52, Algorithm 1] generalized greedy coordinate descent with the "SOft ThresOlding PrOjection" (SOTOPO) algorithm using a reweighted least-squares formulation (SM1.1). However, their methods is neither a coordinate-based method, nor a boosting method. We propose instead a regularized matching pursuit algorithm that draws a clean connection to boosting and proximal coordinate descent.

In the following, we formulate this optimization step (3.4) as a matching pursuit algorithm, that only calls for a linear minimization oracle. Using a variational trick detailed in SM1.1 to approach $\|\beta\|_1^2$, let us begin with formulating the problem (3.4) starting from $\alpha_k \in \mathbb{R}^d$ as a decomposable optimization problem,

$$\begin{aligned} V_\star &= \min_{\beta \in \mathbb{R}^d} \langle \nabla F(\alpha_k), \beta \rangle + \frac{L_1^F}{2} \|\beta\|_1^2 + \lambda \|\beta + \alpha_k\|_1, \\ &= \min_{\beta \in \mathbb{R}^d} \max_{z \geqslant 0} \langle \nabla F(\alpha_k), \beta \rangle - \frac{z^2}{2L_1^F} + z\|\beta\|_1 + \lambda \|\beta + \alpha_k\|_1, \\ &= \max_{z \geqslant 0} -\frac{z^2}{2L_1^F} + \min_{\beta \in \mathbb{R}^d} \sum_{i=1}^d \{\nabla_i F(\alpha_k)\beta^{(i)} + z|\beta^{(i)}| + \lambda |\beta^{(i)} + \alpha_k^{(i)}|\}. \end{aligned}$$

At the optimum, $z = L_1^F \|\beta\|_1$. The problem is now decomposable for each coordinate $\beta^{(i)}$, and can be reduced to an optimization problem in $z \geqslant 0$ in Lemma 3.1.

Lemma 3.1. *The optimization step* (3.4) *can be reformulated as*

$$V_\star = \max_{z_{\min} \leqslant z} \left\{ h(z) \triangleq -\frac{z^2}{2L_1^F} + \sum_{i \in I} \min\left(\lambda |\alpha_k^{(i)}|, -\nabla_i F(\alpha_k)\alpha_k^{(i)} + z|\alpha_k^{(i)}|\right), \right\} \tag{3.5}$$

where $z_{\min} = (\max_i |\nabla_i F(\alpha_k)| - \lambda)_+$ *and where* $I = \{i, \alpha_k^{(i)} \neq 0\}$ *is the set of active atoms.*

Proof. The function $\phi_i(z, \beta^{(i)}) = \nabla_i F(\alpha_k)\beta^{(i)} + z|\beta^{(i)}| + \lambda|\beta^{(i)} + \alpha_k^{(i)}|$ is lower bounded if $z \geqslant |\nabla_i F(\alpha_k)| - \lambda$ for all $i \in I$. Then, $\phi_i(z, \cdot)$ attains its minima in $\beta^{(i)} = 0$ with $\phi_i(z, 0) = \lambda|\alpha_k^{(i)}|$ or in $\beta^{(i)} = -\alpha_k^{(i)}$ with $\phi_i(z, -\alpha_k^{(i)}) = -\nabla_i F(\alpha_k)\alpha_k^{(i)} + z|\alpha_k^{(i)}|$ or in every possible value for $\beta^{(i)}$ if $z = \pm\nabla_i F(\alpha_k) - \lambda$ with $\phi_i(z, 0) = \lambda|\alpha_k^{(i)}|$. ∎

Lemma 3.1 leads to a convex constrained optimization problem in $\mathbb{R}^+$, whose objective is piecewise quadratic with slope coefficients changing at $z_i = \lambda + \nabla_i F(\alpha_k)\alpha_k^{(i)}/|\alpha_k^{(i)}|$. Its constraints in $z_{\min}$ includes the LMO. By construction, the LMO given by $z_{\min}$ may correspond to several atoms β_j such that $j \in \arg\min_i |\nabla_i F(\alpha_k)|$. Since we aim at solving Problem (3.4) by constructing a solution as sparse as possible, we introduce Assumption 3.2.

Assumption 3.2. We consider only one atom corresponding to the LMO, that is $i_{\min} \in \arg\max_i |\nabla_i F(\alpha_k)|$, such that $z_{\min} = (\max_i |\nabla_i F(\alpha_k)| - \lambda)_+$.

In the following Lemmas, we compute explicitly the minimum of h. Under Assumption 3.2, we first deal in Lemma 3.3 with the situation in which the objective is a pure quadratic, that is for all $i \in I$, $z_i \geqslant z_{\min}$.

Lemma 3.3. *Let* $(z_\star, \beta_\star)$ *be a solution of* (3.5)*, assume* $\{i, z_i \geqslant z_{\min}\} = \emptyset$ *and verify Assumption 3.2. Then* $z_\star = z_{\min}$, $\beta_\star^{(i_{\min})} = -\operatorname{sign}(\nabla_{i_{\min}} F(\alpha_k))\frac{z_{\min}}{L_1^F}$ *and* $\beta_\star^{(i)} = 0$ *for* $i \neq i_{\min}$.

Proof. The objective is quadratic and attains its minimum at $z_{\min} = L_1^F|\beta^{(i_{\min})}|$. ∎

In the context of Lemma 3.3 and Assumption 3.2, only the atom given by the LMO in $z_{\min} = (\max_i |\nabla_i F(\alpha_k)| - \lambda)_+$ can be added to the set of active atoms. Now, we assume the objective is piecewise quadratic, that is $\mathcal{S} = \{i, z_i \geqslant z_{\min}\} \neq \emptyset$.

Lemma 3.4. *Let* $z_\star, \beta_\star$ *be a solution of* (3.5) *and assume* $\mathcal{S} = \{i, z_i \geqslant z_{\min}\} \neq \emptyset$ *and verify Assumption 3.2. There are four possible solutions to the Problem* (3.5)*,*

- *If* $h'(z_{\min}) \leqslant 0$*, then* $z_\star = z_{\min}$.
  In addition, $\beta_\star^{(i)} = \begin{cases} -\alpha_k^{(i)} & \text{if } z_i \geqslant z_\star, \\ 0 & \text{if } z_i \leqslant z_\star, \\ -\operatorname{sign}(\nabla_{i_{\min}} F(\alpha_k))\frac{z_{\min} - \sum_{i\in\mathcal{S}} |\alpha_k^{(i)}|}{L_1^F} & \text{if } i = i_{\min}. \end{cases}$
- *If there exists* $k \in \mathcal{S}$ *such that* $h'(z_k^+) \geqslant 0$ *and* $h'(z_{k+1}^-) \leqslant 0$*, then* $z_\star \in]z_k, z_{k+1}[$. *In addition,* $\beta_\star^{(i)} = \begin{cases} -\alpha_k^{(i)} & \text{if } z_i \geqslant z_\star, \\ 0 & \text{if } z_i \leqslant z_\star. \end{cases}$
- *If there exists* $k \in \mathcal{S}$ *such that* $h'(z_k^-) \geqslant 0$ *and* $h'(z_k^+) \leqslant 0$ *then* $z_\star = z_k$. *In addition,* $\beta_\star^{(i)} = \begin{cases} -\alpha_k^{(i)} & \text{if } z_i > z_\star, \\ 0 & \text{if } z_i < z_\star, \\ -\operatorname{sign}(\alpha_k^{(i)})\left(\frac{z_k}{L_1^F} - \sum_{i, z_i > z_k} |\alpha_k^{(i)}|\right) & \text{if } i = k. \end{cases}$
- *If* $h'(z_{|I|}) > 0$*, then for all* $i \in I$, $z_\star > z_i$ *and* $\beta_\star^{(i)} = \begin{cases} -\operatorname{sign}(\alpha_k^{(i)})\frac{z_i}{L_1^F} & \text{if } i = |I|, \\ 0 & \text{otherwise.} \end{cases}$

Proof. The function h is strictly concave and quadratic by part on $[z_i, z_{i+1}]$. The solution to the optimization Problem (3.5) is thus obtained by studying the sign of $h'(\cdot)$ at z_i^- and z_i^+. By construction of the solution given in the proof of Lemma 3.1, for all i such that $z_\star > z_i$ (resp. $z_\star < z_i$), then $\beta^{(i)} = 0$ (resp. $\beta^{(i)} = -\alpha_k^{(i)}$). Finally, we have $z_\star = L_1^F \|\beta_\star\|_1$ which gives the solution for $z_\star = z_{\min}$ or $z_\star = z_k$. ∎

Lemmas 3.3 and 3.4 provides a closed form solution by calling only for the linear minimization oracle $\min_i |\nabla_i F(\alpha_k)|$, and performing $O(|I|)$ operations on the active atoms. From that, we deduce Algorithm 3.1.

Algorithm 3.1 Regularized matching pursuit

$\alpha \in \mathbf{R}^d$, $N \in \mathbf{N}$
for $k \in [0, \dots, N]$ **do**
  $z_{\min} = (\max_i |\nabla_i F(\alpha_k)| - \lambda)_+$ and $i_{\min} = \arg\max_i |\nabla_i F(\alpha_k)|$
  For $\alpha_k^{(i)} \neq 0$, compute $z_i = \lambda + \frac{\alpha_k^{(i)}}{|\alpha_k^{(i)}|} \nabla_i F(\alpha_k)$ such that $z_{i+1} \geqslant z_i$
  if $\{i, z_i \geqslant z_{\min}\} = \emptyset$ **then**
    $\beta_{i_{\min}} = -\text{sign}(\nabla_{i_{\min}} F(\alpha_k)) \frac{z_{\min}}{L_1^F}$
  else
    Compute $u = \text{argmin}_i \{z_i \geqslant z_{\min}\}$ and for $i \in [u, v]$, compute $h'(z_i)$
    if $h'(z_{\min}) \leqslant 0$ or $h'(z_u) \leqslant 0$ **then**
      For $i \in [u, v]$, $\beta^{(i)} = -\alpha_k^{(i)}$
      If $h'(z_{\min}) \leqslant 0$, **then** $\beta_{i_{\min}} = -\text{sign}(\nabla_{i_{\min}} F(\alpha_k))(\frac{z_{\min}}{L_1^F} - \sum_{i=u}^v |\alpha_k^{(i)}|)$
    else
      $n = \arg\max\{i, i \in [u, v-1],\ h'(z_i^+), h'(z_{i+1}^-) \geqslant 0\}$
      if $n = v - 1$ **then**
        $\beta_v = -\text{sign}(\alpha_v) \frac{1}{L_1^F} (\lambda + \frac{\alpha_v}{|\alpha_v|} \nabla_v F(\alpha_k))$
      else
        For $i \in [n+1, v]$, $\beta^{(i)} = -\alpha_k^{(i)}$
        If $h'(z_n^+) \leqslant 0$, **then** $\beta^{(n)} = -\text{sign}(\alpha_k^{(n)})(\frac{1}{L_1^F} \left(\lambda + \frac{\alpha_k^{(n)}}{|\alpha_k^{(n)}|} \nabla_n F(\alpha_k)\right) - \sum_{i=n+1}^v |\alpha_k^{(i)}|)$
      end if
    end if
  end if
  $\alpha_{k+1} = \alpha_k + \beta$
end for

In short, Algorithm 3.1 performs three possible actions: either one new atom is added (at most) by calling the LMO $\arg\max_i |\nabla_i F(\alpha_k)| = \arg\max_{p \in \mathcal{P}} p^\top \nabla f(P\alpha_k)$ while some active atoms may be set to zero, or one active atom may be optimized while some active atoms may be set to zero, or some active atoms are set to zero (but none is added nor optimized). To sum it up, at each iteration, it constructs the next iterate using only past active atoms plus possibly a new one generated by the LMO. Therefore, Algorithm 3.1 belongs to the family of boosting algorithms. We refer to it as the regularized matching pursuit.

Compared to the boosting approach of Zhang et al. [64] for metric-norm regularization or to the generalized conditional gradient [3, 54], active atoms are not modified uniformly since only some of them may be reduced to zero. The SOTOPO method of Song [52] minimizes the same upper bound with respect to the ℓ_1-norm (3.4). Yet, it is resolved with a different variational formulation, that does not let a linear minimization oracle appear. Compared to proximal coordinate descent with GS-rule (3.3) applied with L_1^F (instead of L_2^F), the regularized matching pursuit happens to often follow exactly the same path when starting from zero (but not when starting from a nonzero point), as observed in Figure 3. This suggests a connection between regularized matching pursuit and proximal coordinate descent, as proven by Locatello et al. [33] for gradient descent and steepest coordinate descent.

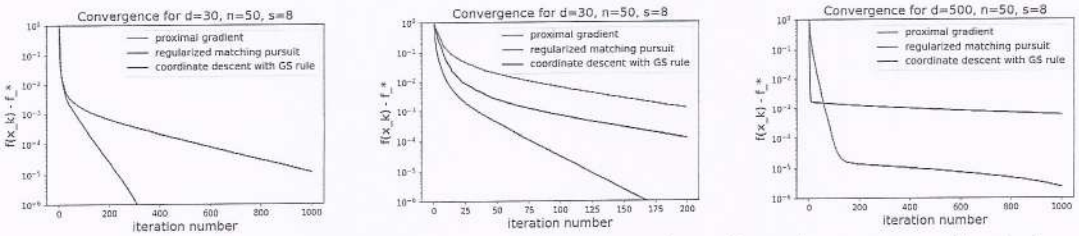

Figure 3: Convergence in function value of the proximal gradient descent, coordinate descent with Gauss-Soutwhell rule and with $L = L_1^F$ (instead of $L = L_2^F$) and of the regularized matching pursuit, for $n = 50$, $s = 8$, $\lambda = 0.001$, $\sigma = 0.5$ and for $d = 30$ starting from zero on the left (underparametrized regime), from a non zero point in the middle (underparametrized regime) and for $d = 500$ on the right (overparametrized regime). RMP and coordinate descent with GS-rule matche exactly in these examples.

In Figure 3, the regularized matching pursuit seems to converge linearly in the underparametrized regime, and sublinearly in the overparametrized regime. We will compute its theoretical convergence guarantees in the next section.

3.2. Convergence guarantee. We now establish convergence guarantees of the regularized matching pursuit, both in for strongly convex and non-strongly convex functions F. We consider a more general composite minimization problem,

$$\min_{\alpha \in \mathbb{R}^d} G(\alpha) \triangleq F(\alpha) + H(\alpha), \tag{3.6}$$

where H is closed, convex, proper, and where F is L_1^F-smooth and (possibly) μ_1^F-strongly convex with respect to the ℓ_1-norm. If in addition, F is a linear mapping, and $H(\cdot) = \|\cdot\|_1$, this is exactly the original optimization Problem (1.1). We evaluate the convergence guarantee a generalized version of the regularized matching pursuit (that is not always a boosting method),

$$\alpha_{k+1} = \arg\min_{\alpha \in \mathbb{R}^d} \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_k\|_1^2 + H(\alpha). \tag{3.7}$$

As we will see, our proofs are closely related to those for randomized coordinate descent [49, Theorem 5, 7].

3.2.1. **Strongly convex functions F.** Let us assume that F is L_1^F-smooth and μ_1^F-strongly convex, typically in the underparametrized regime. Similarly to coordinate gradient descent with GS rule which converges linearly in this context [43], regularized matching pursuit is formulated as the minimization of the smoothness upper bound with respect to the ℓ_1-norm. Therefore, it benefits from linear convergence guarantees, detailed below.

Proposition 3.5. *[41, Appendix 8] If F be convex, L_1^F-smooth with respect to the ℓ_1-norm, and μ_1^F-strongly convex with respect to the ℓ_1-norm. Then, the sequence (α_k) generated by (3.7) verifies,*

$$G(\alpha_{k+1}) - G_\star \leqslant \left(1 - \frac{\mu_1^F}{L_1^F}\right)(G(\alpha_k) - G_\star).$$

Proof. The proof consists in an optimization step over all trajectories, and not on a randomized step (see [49]), as it was done by [41]. See SM4.2. ∎

The regularized matching pursuit is a special case of method (3.7), and verifies the convergence guarantee of Proposition 3.5. As a conclusion, it beats traditional boosting techniques converging sublinearly, such as coordinate descent with GS rule (with $1 - \frac{\mu_2^F}{dL_2^F} \leqslant 1 - \frac{\mu_1^F}{L_1^F}$), or the generalized conditional gradient that is also appropriate to a gauge geometry. In addition, its linear guarantee only depends on the strong convexity and smoothness parameters of F with respect to the ℓ_1-norm. In the special case of the LASSO, the estimates established in Proposition 2.8 and 2.9 still apply. In the overparametrized regime however, Figure 3 suggests that the method does not converge linearly (since it is stuck at a precision around 10^{-5}).

3.2.2. **Smooth convex functions F.** Let now F be L_1^F-smooth, convex, but not strongly convex (which is verified in the overparametrized regime). Usually, guarantees for splitting methods, such as proximal gradient, states a sublinear convergence guarantee. Similarly in Proposition 3.6, we prove sublinear convergence of the (generalized) regularized matching pursuit. To our knowledge, there is no such theorem for sublinear convergence of SOTOPO [52] or of coordinate descent with the Gauss-Southwell rule, which is very close to the generalized regularized matching pursuit.

Proposition 3.6. *Let (α_k) be generated by the generalized regularized matching pursuit (3.7), starting from $\alpha_0 \in \mathbb{R}^d$*

$$G(\alpha_k) - G_\star \leqslant \frac{2L_1^F \mathcal{R}_{\alpha_0}^2}{k+1},$$

where $\mathcal{R}_{\alpha_0}^2 = \max_{\alpha \in \mathbb{R}^d} \max_{\alpha_\star \in \mathbb{R}^d} \{\|\alpha - \alpha_\star\|_1^2, \text{ s.t. } G(\alpha) \leqslant G(\alpha_0)\}$.

Proof. This technique is inspired from the proof for sublinear convergence for randomized proximal coordinate descent established by Richtarik and Takac [49, Theorem 5]. Let $\alpha_{k+1} \in$

$\mathbb{R}^d$ be a minimizer of the smooth upper bound:

$$
\begin{aligned}
G(\alpha_{k+1}) &\leqslant \inf_{\alpha\in\mathbb{R}^d} F(\alpha_k) + \langle \nabla F(\alpha_k), \alpha - \alpha_k\rangle + \frac{L_1^F}{2}\|\alpha-\alpha_k\|_1^2 + H(\alpha), \\
&\leqslant \inf_{\alpha\in\mathbb{R}^d} F(\alpha) + H(\alpha) + \frac{L_1^F}{2}\|\alpha-\alpha_k\|_1^2 \left(= G(\alpha) + \frac{L_1^F}{2}\|\alpha-\alpha_k\|_1^2\right) \text{ (F convex)}, \\
&\leqslant \inf_{t\in[0,1]} G(t\alpha_\star + (1-t)\alpha_k) + \frac{L_1^F t^2}{2}\|\alpha_k-\alpha_\star\|_1^2, \\
&\leqslant \inf_{t\in[0,1]} G(\alpha_k) - t(G(\alpha_k)-G_\star) + \frac{L_1^F t^2}{2}\|\alpha_k-\alpha_\star\|_1^2 \text{ (convexity of } H, F), \\
G(\alpha_{k+1}) - G_\star &\leqslant \inf_{t\in[0,1]} (1-t)(G(\alpha_k)-G_\star) + \frac{L_1^F t^2}{2}\|\alpha_k-\alpha_\star\|_1^2.
\end{aligned}
$$

The solution of this minimization problem is given by $t_\star = \min(1, \frac{G(\alpha_k)-G_\star}{L_1^F\|\alpha_k-\alpha_\star\|_1^2})$. We conclude the minimization bound, depending on the sign of $G(\alpha_k) - G_\star - L_1^F\|\alpha_k-\alpha_\star\|_1^2$:

$$
G(\alpha_{k+1}) - G_\star \leqslant \max\left(1 - \frac{G(\alpha_k)-G_\star}{2L_1\|\alpha_k-\alpha_\star\|_1^2}, \frac{1}{2}\right)(G(\alpha_k)-G_\star).
$$

As a first conclusion, notice that $G(\alpha_k) - G_\star$ is nonincreasing. Recall now that $\mathcal{R}_{\alpha_0}^2 = \max_{\alpha\in\mathbb{R}^d}\max_{\alpha_\star\in\mathbb{R}^d}\{\|\alpha-\alpha_\star\|_1^2, \text{ s.t. } G(\alpha)\leqslant G(\alpha_0)\}$. Then, using the notation $\delta_k = G(\alpha_k)-G_\star$, an upper bound for δ_{k+1} is given by $\delta_{k+1} \leqslant \max\left(1 - \frac{\delta_k}{2L_1^F\mathcal{R}_{\alpha_0}^2}, \frac{1}{2}\right)\delta_k$. Assume now that $\delta_0 \leqslant L_1^F\mathcal{R}_{\alpha_0}^2$ and notice that $\delta_k \leqslant L_1^F\mathcal{R}_{\alpha_0}^2$ since δ_k is nonincreasing. If not, notice that the inequality satisfied at the next iteration $\delta_1 \leqslant \frac{1}{2}\mathcal{R}_{\alpha_0}^2$. Then, we have for $\omega = \frac{1}{2L_1^F\mathcal{R}_{\alpha_0}^2}$, $\delta_{k+1} \leqslant (1-\delta_k\omega)\delta_k$. Following the same argument than in the proof for sublinear convergence of steepest coordinate descent, detailed in SM4.1, it leads to the convergence guarantee $G(\alpha_k) - G_\star \leqslant \frac{2L_1^F\mathcal{R}_{\alpha_0}^2}{k+1}$.

In Proposition 3.6, we have obtained a sublinear convergence guarantee for the regularized matching pursuit for non-strongly convex functions. To our knowledge, this is the first sublinear guarantee for a boosting algorithm under classical assumptions from convex optimization. This method does not benefit from linear convergence guarantee. Yet, as we will see in numerical experiments in the next section, in the case of ℓ_1-regularized model, the regularized matching pursuit converges linearly in certain regimes.

3.3. A phase transition depending on λ: experimental results. The regularized matching pursuit algorithm benefits from convergence guarantees similar to those for the proximal gradient: these methods converge linearly under strong convexity assumptions (underparametrized regime) but sublinearly for non-strongly convex problems (overparametrized regime). In the context of sparsity though, the proximal gradient descent benefits from linear convergence under additional assumptions on the problem classes such as restricted eigenvalue properties [48], and for a well-chosen parameter λ [1, Theorem 2]. In this section, our experiments reveals a transition phenomenon driven by λ on a LASSO problem

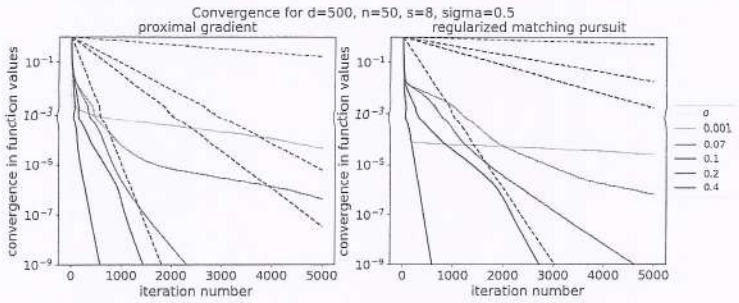

Figure 4: Convergence in function values for the proximal gradient on the left and the regularized matching pursuit on the right for $n = 50$, $d = 500$ and a sparsity $s = 8$ and for several penalty λ. Convergence is compared in dashed lines to local convergence guarantee, taken on the support S on the last iterates and the SDP relaxation from Proposition 2.8.

$F(\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2 + \lambda\|x\|_1$, where P are synthetic Gaussian data in the overparametrized setting as in Section 2.3.

The convergence behavior of proximal gradient descent follows a transition phase, that can be divided into three phases: first, the method converges linearly according to the nonregularized trajectory, then it converges sublinearly, and it converges linearly once the support is identified. Iutzeler and Malick [27, Theorem 1] always identifies the structure of the solution (described by manifolds, or sparsity patterns) after a certain number of steps. For strongly convex functions, Nutini et al. [42] bounded the 'active-set'-complexity of the proximal gradient method. The regularized matching pursuit follows the same behavior. It appears that the sparsity of the solution, and the sparsity identification highly depends on the value of λ: the larger λ, the sparser the solution and the quicker the identification. Thanks to this observation, we derive a posteriori guarantee in Figure 4, based on the sparsity of the solution to the optimization problem. In Figure 4, local strong convexity parameters are given by the estimated of Corollary 2.6 and Proposition 2.8. We recover that large parameter λ both induces a stricter sparsity on the solution and a better convergence.

We describe this transition phase numerically in Figure 5 by plotting the ϵ-curve (see Section 2.3) as a function of λ. For $\lambda = 0$, both methods converge linearly (as expected in the overparametrized regime for gradient descent and coordinate descent with the GS-rule). For 'large' value of λ for which the support is quickly identified, the convergence is linear too. In the middle phase however, convergence depends on the effective dimension of the trajectory, $d_{eff} \approx n$ by construction (and thus, on the effective strong convexity of f long the trajectory). The ϵ-curves can be seen as equivalent in the optimization perspective with the regularization path usually drawn in the context of statistical recovery.

The regularized matching pursuit Algorithm 3.1 formulation allows some intuition regarding the interplay between λ and the sparsity of the solution. Let $\mathcal{A} = \{i, \alpha_k^{(i)} > 0\}$ be the set of active atoms. Algorithm 3.1 may reduce an active atom $i \in \mathcal{A}$ to zero if $\|\nabla F(\alpha)\|_\infty - \frac{\alpha_k^{(i)}}{|\alpha_k^{(i)}|}\nabla_i F(\alpha) \leqslant 2\lambda$. The larger λ, the more active directions may be canceled out.

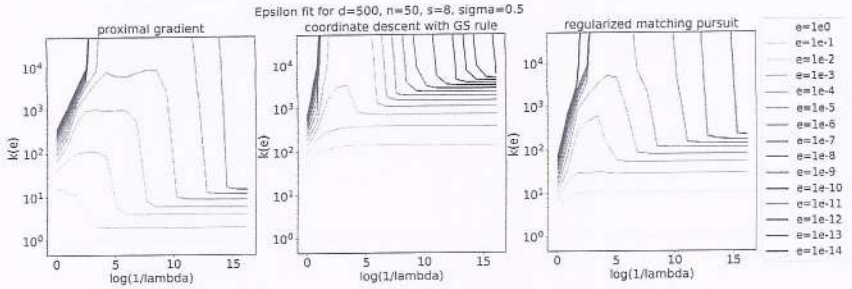

Figure 5: ϵ-curve of the proximal gradient, coordinate descent with the GS rule and regularized matching pursuit for a LASSO problem with $d = 500$, $n = 50$, a sparsity level $s = 8$, $\sigma = 0.5$, after $k = 10000$ iterations for several values of λ.

The smaller λ ($\lambda \ll \|\nabla F(\alpha)\|_\infty$), the closer is regularized matching pursuit to coordinate descent with GS rule (on the right in Figure 5): indeed, only the linear minimization oracle may be added to the set of atoms without modifying other active atoms ($z_i \lesssim z_{\min}$). For $\lambda \approx 0$, the regularized matching pursuit thus converges linearly up to a certain iteration number, which appears with the parallel level lines in Figure 5.

Based on the minimization of a smoothness upper bound with respect to the ℓ_1-norm, we have developed a regularized matching pursuit algorithm, that benefits from linear convergence in the underparametrized regime (where F is strongly convex), and sublinear convergence in the overparametrized regime (where F is not strongly convex). Thanks to the ϵ-curve, we numerically described the role of λ on the convergence of the method (and on the sparsity). In the following section, we propose to develop a method suited to the gauge geometry in the overparametrized regime.

3.4. An ultimate method adapted to the geometry of regularized models. The regularized matching pursuit 3.1 was derived from the ℓ_1-geometry. In Section 2.4, for non-regularized models, coordinate descent with GS-rule was interpreted as a matching pursuit algorithm in the ℓ_1-geometry in the underparametrized regime, but in the $\gamma_{\mathcal{P}}$-geometry in the overparametrized regime. We will see that the regularized matching pursuit as developed above does not benefit from this formulation in the overparametrized regime. Instead, we will propose an 'ultimate method' for the gauge geometry, that benefits from linear convergence in the overparametrized regime but lacks a simple formulation.

Recall the equivalent regularized minimization problems (1.1) and (1.3),

$$\min_{\alpha \in \mathbb{R}^d} f(P\alpha) + \lambda \|\alpha\|_1 = \min_{x \in \mathbb{R}^n} f(x) + \lambda \gamma_{\mathcal{P}}(x),$$

where $\gamma_{\mathcal{P}}$ is a gauge function as defined in Section 2.4, and f is $L^f_{\gamma_{\mathcal{P}}}$-smooth and $\mu^f_{\gamma_{\mathcal{P}}}$-strongly convex with respect to the gauge. The problem in $\mathbb{R}^d$ is reformulated as $\mathbb{R}^n$, of lower dimension.

Remark 3.7. This reformulation only requires $\gamma_{\mathcal{P}}$ to be a gauge function, but not specifically to be a norm. Reversely, minimizing a function penalized by a gauge function (or a

semi-norm) can be reformulated as minimizing a linear mapping penalized by and ℓ_1-norm.

As for the regularized matching pursuit, we formulate an optimization method as the minimization of the smoothness upper bound with respect to the gauge function, starting from $x_0 \in \mathbb{R}^n$.

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \langle \nabla f(x_k), x - x_k \rangle + \frac{L^f_{\gamma_\mathcal{P}}}{2} \gamma_\mathcal{P}(x - x_k)^2 + \lambda \gamma_\mathcal{P}(x). \tag{3.8}$$

We refer to this method as the ultimate method for the gauge $\gamma_\mathcal{P}$, that is adapted to the geometry of the regularized problem 1.1. Let us reformulate the minimization Problem (3.8) on $\mathbb{R}^n$ into a minimization problem in $\mathbb{R}^d$. Let $x_k = P\alpha_k$ with $\alpha_k \in \mathbb{R}^d$, then

$$\begin{aligned}
&\min_{x \in \mathbb{R}^n} \langle \nabla f(x_k), x - x_k \rangle + \frac{L^f_{\gamma_\mathcal{P}}}{2} \gamma_\mathcal{P}(x - x_k)^2 + \lambda \gamma_\mathcal{P}(x), \\
&= \min_{\alpha, \nu \in \mathbb{R}^d} \langle \nabla f(P\alpha_k), P(\alpha - \alpha_k) \rangle + \frac{L^f_{\gamma_\mathcal{P}}}{2} \|\alpha - \alpha_k\|_1^2 + \lambda \|\nu\|_1, \text{ s.t. } x = P\alpha = P\nu, \\
&= \min_{\alpha, \nu \in \mathbb{R}^d} \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L^f_{\gamma_\mathcal{P}}}{2} \|\alpha - \alpha_k\|_1^2 + \lambda \|\nu\|_1, \text{ s.t. } P\alpha = P\nu.
\end{aligned}$$

When $P\alpha = P\nu$ implies $\alpha = \nu$, such as in the underparametrized regime where $P^\top P$ is invertible, the ultimate method for the gauge is equivalent with the regularized matching pursuit (3.1). However, in the overparametrized regime, $P\alpha = P\nu$ does not imply $\alpha = \nu$ in general. We believe this method does not belong to boosting algorithms due to the evaluation of the gauge function in x and in $x - x_k$ in (3.8). In addition, this minimization problem admits neither a simple closed-form solution in general nor a solution based on the KKT conditions (as we did for regularized matching pursuit). While not directly computable in general, the minimization step (3.8) converges linearly to the optimum, as proven below in Proposition 3.8.

Proposition 3.8. *Let f be L^f-smooth and μ^f-strongly convex with respect to the norm $\gamma_\mathcal{P}(\cdot)$. The ultimate matching pursuit (3.8) (x_k) converges linearly with*

$$f(x_k) - f_\star \leqslant \left(1 - \frac{\mu^f_{\gamma_\mathcal{P}}}{L^f_{\gamma_\mathcal{P}}}\right)^k (f(x_0) - f_\star).$$

Proof. The proof follows exactly the proof for Theorem 3.5, replacing the function F by f and the norm $\|\cdot\|_1$ by $\gamma_\mathcal{P}(\cdot)$. ∎

In Proposition 3.8, we prove linear convergence for the ultimate matching pursuit algorithm. We recover the convergence guarantee of coordinate descent with GS rule in the non regularized model (2.6).

In Figure 6, we solve the optimization step for the ultimate method for the gauge with the solver MOSEK [2] on a LASSO problem. It converges linearly in the overparametrized regime, while the other method are first stuck in a sublinear phase. Compared to the proximal gradient, proximal coordinate descent with GS rule, the regularized matching pursuit and the ultimate method starts with sparse solution, and differs after a small number of iteration

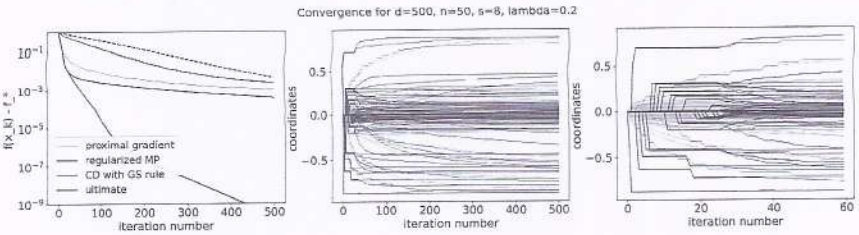

Figure 6: Convergence in function value and coordinates as a function of the iteration number for the proximal gradient descent, the proximal coordinate descent with GS rule, the regularized matching pursuit and the ultimate method on a LASSO problem, with $d = 500$, $n = 50$, $s = 8$, $\lambda = 0.2$. The approximate guarantee in provided in dashed lines.

(about 30 here). In the special case of the LASSO, it is possible to approximate its convergence guarantee as for the linear regression problem. Noticing that $L^f_{\gamma\mathcal{P}} = L^F_1$ and $\mu^f_{\gamma\mathcal{P}} = \mu^F_1$, the estimate of the convergence guarantee of coordinate descent with GS rule from Proposition 2.8 apply here. In the Supplementary Material ??, we propose an inner loop strategy to avoid the use of an optimization solver, together with the convergence analysis of the outer loop given the precision of the inner loop.

Conclusion and future works. In this paper, we developed a principled view for generating optimization algorithms from the minimization of a smoothness upper bound with respect to a well-chosen norm. For non-regularized models, this procedure leads to coordinate descent with GS-rule, that can be interpreted as a matching pursuit algorithm both in the ℓ_1-geometry for underparametrized models, and in the $\gamma\mathcal{P}$-geometry for overparametrized models. Building on these results, we derive a new regularized matching pursuit algorithm based on the minimization of smoothness with respect to the ℓ_1-norm (whose counterpart is proximal gradient descent in the ℓ_2-geometry). While being strongly connected to proximal coordinate descent with GS-rule, the regularized matching pursuit cannot be interpreted as a matching pursuit algorithm in the gauge geometry for overparametrized models, and thus, does not converge linearly in this regime. We finally formulate the ultimate method adapted to overparametrized geometries. Yet, this method lacks a closed-form formulation, that we approximate thanks to an inner-loop strategy.

From this approach, we obtain refined convergence guarantees for (resp. regularized) matching pursuit (resp. coordinate descent with GS rule), that are adapted to the geometry of the problem under consideration. For linear regression and the LASSO, we derive approximate convergence guarantees using SDP relaxations and random matrix theory. As a byproduct, convergence guarantees of both gradient descent and steepest coordinate descent applied to least-squares follow a transition phase from the underparametrized to the overparametrized regime. For ℓ_1-regularized models, a similar transition phase for λ appears, and allows to interpret it as a measure of the sparsity of the solution.

Building on these results, it could be of interest to extend this principled approach to accelerated matching pursuit algorithm (and thus, to accelerated coordinate descent algo-

rithms). Some accelerated techniques have already been developed relying on randomly selected coordinates, such as those of Nesterov et al. [40] for nonregularized minimization and Fercoq and Richtarik [19] or Locatello et al. [33, Section 3] for composite minimization problems, but. Another interesting line of research could be to understand the connection between the observed phase transition for optimization methods and the double descent phenomenon observed for the generalization error in machine learning.

Acknowledgments. This work was funded by MTE and the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grant SEQUOIA 724063).

REFERENCES

[1] A. Agarwal, S. Negahban, and M. J. Wainwright, *Fast global convergence rates of gradient methods for high-dimensional statistical recovery*, in Advances in Neural Information Processing Systems, vol. 23, 2010.

[2] M. ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 10.0.*, 2022, http://docs.mosek.com/9.0/toolbox/index.html.

[3] F. Bach, *Duality between subgradient and conditional gradient methods*, SIAM Journal on Optimization, 25 (2015), pp. 115–129.

[4] F. Bach, *High-dimensional analysis of double descent for linear regression with random projections.* Technical report, arXiv:2303.01372, 2023.

[5] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*, Springer Series in Statistics, (2010).

[6] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2017.

[7] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM Journal on Optimization, 23 (2013), pp. 2037–2060.

[8] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Reconciling modern machine learning and the bias-variance trade-off*, Proceedings of the National Academy of Sciences, 32 (2019), pp. 15849–15854.

[9] R. Berthier, F. Bach, and P. Gaillard, *Accelerated gossip in networks of given dimension using Jacobi polynomial iterations*, SIAM Journal on Mathematics of Data Science, 2 (2020), pp. 24–47.

[10] J. Bolte, A. Daniilinis, O. Ley, and L. Mazet, *Characterization of Łojasiewicz inequalities: subgradient flows, talweg, convexity*, Transactions of the American Mathematical Society, 362 (2010), pp. 3319–3363.

[11] K. Borgwardt, *The average number of pivot steps required by the simplex-method is polynomial*, Zeitschrift für Operations Research, 26 (1987), pp. 157–177.

[12] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Review, 60 (2018), pp. 223–311.

[13] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities. A nonasymptotic theory of independence*, Oxford University Press, 2013.

[14] E. Candes and T. Tao, *Decoding by linear programming*, IEEE Transactions on Information Theory, 51 (2005), pp. 4203–4215.

[15] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Modeling & Simulation, 4 (2005), pp. 1168–1200.

[16] A. d'Aspremont, C. Guzmán, and M. Jaggi, *Optimal affine-invariant smooth minimization algorithms*, SIAM Journal on Optimization, 28 (2018), pp. 2384–2405.

[17] M. Dudik, Z. Harchaoui, and J. Malick, *Lifted coordinate descent for learning with trace-norm regularization*, in Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, 2012, pp. 327–336.

[18] H. Fang, Z. Fan, Y. Sun, and M. P. Friedlander, *Greed meets sparsity: Understanding and improving greedy coordinate descent for sparse optimization*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, 2020.
[19] O. Fercoq and P. Richtárik, *Accelerated, parallel, and proximal coordinate descent*, SIAM Journal on Optimization, 25 (2015), pp. 1997–2023.
[20] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly, 3 (1956), pp. 95–110.
[21] Y. Freund and R. E. Schapire, *A short introduction to boosting*, Japanese Society For Artifical Intelligence, 14 (1999), pp. 771–780.
[22] M. P. Friedlander, I. Macêdo, and T. K. Pong, *Gauge optimization and duality*, SIAM Journal on Optimization, 24 (2014), pp. 1999–2022.
[23] M. X. Goemans and D. P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, Journal of the ACM, 42 (1995), pp. 1115–1145.
[24] T. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring*, Science, 286 (1999), pp. 531–537.
[25] C. Guille-Escuret, B. Goujaud, M. Girotti, and I. Mitliagkas, *A study of condition numbers for first-order optimization*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, 2021, pp. 1261–1269.
[26] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, Journal Research of the National Bureau of Standards, 49 (1957), pp. 263–264.
[27] F. Iutzeler and J. Malick, *Nonsmoothness in machine learning: specific structure, proximal identification, and applications*, Set-Valued and Variational Analysis, 28 (2020), pp. 661–678.
[28] M. Jaggi, *Revisiting Frank-Wolfe: Projection-free sparse convex optimization*, in Proceedings of the International Conference on Machine Learning, no. 1, 2013, pp. 427–435.
[29] H. Karimi, J. Nutini, and M. Schmidt, *Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition*, in Machine Learning and Knowledge Discovery in Databases, 2016, pp. 795–811.
[30] S. P. Karimireddy, A. Koloskova, S. U. Stich, and M. Jaggi, *Efficient greedy coordinate descent for composite problems*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, 2019.
[31] S. Lacoste-Julien and M. Jaggi, *On the global linear convergence of frank-wolfe optimization variants*, in Advances in Neural Information Processing Systems, 2015.
[32] F. Locatello, R. Khanna, M. Tschannen, and M. Jaggi, *A unified optimization view on generalized matching pursuit and Frank-Wolfe*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, 2017, pp. 860–868.
[33] F. Locatello, A. Raj, S. P. Karimireddy, G. Raetsch, B. Schölkopf, S. Stich, and M. Jaggi, *On matching pursuit and coordinate descent*, in Proceedings of the International Conference on Machine Learning, 2018, pp. 3198–3207.
[34] S. G. Mallat and Z. Zhang, *Matching pursuits with time-frequency dictionaries*, IEEE Transactions on Signal Processing, 41 (1993), p. 3397–3415.
[35] V. A. Marchenko and L. A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Matematicheskii Sbornik, 1 (1967), pp. 457 – 483.
[36] S. Mei and A. Montanari, *The generalization error of random features regression: Precise asymptotics and the double descent curve*, Communications on Pure and Applied Mathematics, 75 (2022), pp. 667–766.
[37] I. Necoara, Y. Nesterov, and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Mathematical Programming, 175 (2019), p. 69–107.
[38] Y. Nesterov, *A method for solving the convex programming problem with convergence rate o(1/k2)*, Proceedings of the USSR Academy of Sciences, 269 (1983), pp. 543–547.
[39] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization, 22 (2012), pp. 341–362.
[40] Y. Nesterov and S. U. Stich, *Efficiency of the accelerated coordinate descent method on structured*

optimization problems, SIAM Journal on Optimization, 27 (2017), pp. 110–123.
[41] J. Nutini, *Greed is good: greedy optimization methods for large-scale structured problems*, PhD thesis, University of British Columbia, 2018.
[42] J. Nutini, M. Schmidt, and W. Hare, *"active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern?*, Optimization Letters, 13 (2018), pp. 645–655.
[43] J. Nutini, M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke, *Coordinate descent converges faster with the Gauss-Southwell rule than random selection*, in Proceedings of the International Conference on Machine Learning, 2018, pp. 1632–1641.
[44] C. Paquette, B. van Merriënboer, E. Courtney, and F. Pegregosa, *Halting time is predictable for large models: A universality property and average-case analysis*, Foundations of Computational Mathematics, 23 (2023), p. 597–673.
[45] N. Parikh and S. P. Boyd, *Proximal algorithms*, Foundations and Trends Optimization, 1 (2013), pp. 127–239.
[46] F. Pedregosa and D. Scieur, *Acceleration through spectral density estimation*, in Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 7553–7562.
[47] F. Pedregosa and D. Scieur, *Acceleration through spectral density estimation*, in Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 7553–7562.
[48] G. Raskutti, M. J. Wainwright, and B. Yu, *Restricted eigenvalue properties for correlated gaussian designs*, Journal of Machine Learning Research, 11 (2010), pp. 2241–2259.
[49] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Mathematical Programming, 144 (2014).
[50] D. Scieur and F. Pedregosa, *Universal asymptotic optimality of Polyak momentum*, in Proceedings of the 37th International Conference on Machine Learning, 2020, pp. 8565–8572.
[51] S. A. B. Sheng Chen and W. Luo, *Orthogonal least squares methods and their application to non-linear system identification*, International Journal of Control, 50 (1989), pp. 1873–1896.
[52] C. Song, S. Cui, Y. Jiang, and S.-T. Xia, *Accelerated stochastic greedy coordinate descent by soft thresholding projection onto simplex*, in Advances in Neural Information Processing Systems, 2017.
[53] D. Spielman and S.-H. Teng, *Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time*, in Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing, 2001, p. 296–305.
[54] Y. Sun and F. Bach, *Safe screening for the generalized conditional gradient method*, Open Journal of Mathematical Optimization, 3 (2022), pp. 1–35.
[55] A. B. Taylor, J. Hendrickx, and F. a. Glineur, *Exact worst-case convergence rates of the proximal gradient method for composite convex minimization*, Journal of Optimization Theory and Applications, 178 (2018), pp. 455–476.
[56] A. Tewari, P. Ravikumar, and I. Dhillon, *Greedy algorithms for structurally constrained high dimensional problems*, in Advances in Neural Information Processing Systems, 2011.
[57] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, Journal of the Royal Statistical Society. Series B, 58 (1996), pp. 267–288.
[58] J. A. Tropp, *Greed is good: algorithmic results for sparse approximation*, IEEE Transactions on Information Theory, 50 (2004), pp. 2231–2242.
[59] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, Journal of Optimization Theory and Applications, 109 (2001), p. 475–494.
[60] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, B (2009), pp. 387–423.
[61] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Series in Statistical and Probabilistic Mathematics, 2018.
[62] T. Zhang, *Adaptive forward-backward greedy algorithm for learning sparse representations*, IEEE Transactions on Information Theory, 57 (2011), pp. 4689–4708.
[63] T. Zhang, *Sparse recovery with orthogonal matching pursuit under RIP*, IEEE Transactions on Information Theory, 57 (2011), pp. 6215–6221.
[64] X. Zhang, D. Schuurmans, and Y.-l. Yu, *Accelerated training for matrix-norm regularization: A boosting approach*, in Advances in Neural Information Processing Systems, vol. 25, 2012.

Appendix.

Appendix A. Linear convergence Propositions 2.4, 2.7.

Let us prove linear convergence of gradient descent with fixed step size and coordinate descent with GS rule in a more general framework. This proof leads to the results of Propositions 2.4, 2.7 for the ℓ_2, ℓ_1 norm respectively.

Let F be L^F-smooth with respect to a norm $\|\cdot\|$, (possibly) μ-strongly convex with respect to $\|\cdot\|$ and verify the Łojasiewicz inequality with parameter $\mu^{L,F}$. We consider a method (α_k) starting from $\alpha_0 \in \mathbb{R}^d$, and obtained by minimizing the smoothness quadratic upper bound:

$$F(\alpha_{k+1}) \leqslant F(\alpha_k) + \min_{\alpha \in \mathbb{R}^d} \left(\langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L^F}{2} \|\alpha - \alpha_k\|^2 \right) \leqslant F(\alpha_k) - \frac{1}{2L^F} \|\nabla F(\alpha_k)\|_\star^2.$$

If F is μ^F-strongly convex, then F verifies the Łojasiewicz inequality with parameter μ^F: for all $\alpha \in \mathbb{R}^d$, $\mu^F(F(\alpha) - F_\star) \leqslant \frac{1}{2}\|\nabla F(\alpha)\|_\star^2$. Thus, substracting $F_\star$ on each side of the smoothness inequality, we have

$$F(\alpha_{k+1}) - F_\star \leqslant \left(1 - \frac{\mu^F}{L^F}\right) (F(\alpha_k) - F_\star).$$

Similarly, if F satisfies the Łojasiewicz inequality with parameter $\mu^{L,F}$, but is not strongly convex ($\mu^F = 0$).

Appendix B. Approximate linear convergence of coordinate descent for least-squares.

In Proposition 2.7, a sequence (α_k) generated by steepest coordinate descent for a linear regression problem has a linear convergence rate in function values,

$$F(\alpha_k) - F_\star \leqslant \left(1 - \frac{\max(\mu_1^F, \mu_1^{L,F})}{L_1^F}\right)^{2k} (F(\alpha_0) - F_\star).$$

We first derive inequalities connecting $\mu_1^{L,F}$ and μ_1^F to L_1^F. Then, we prove concentration inequalities for $\mu_1^{L,F}$, μ_1^F and L_1^F, and derive approximate convergence guarantees of steepest coordinate descent.

B.1. SDP relaxations for μ_1^F and μ_1^{PL}; proof for Proposition 2.8.

We look for exact lower bounds for μ_1^F and μ_1^{PL}. Both in the overparametrized and underparametrized regime, we are going to reformulate the optimization problems defining $\mu_1, \mu_1^{L,F}$ into SDPs, and relax some rank constraints.

SDP relaxation for μ_1^F in the underparametrized regime. Recall from Lemma 2.3 that μ_1^F is non zero in this regime and given by $\frac{1}{\sqrt{n}\mu_1^F} = \max_{\alpha, \|\alpha\|_\infty \leqslant 1} \max_{\nu, \|P\nu\|_2 \leqslant 1} \alpha^\top \nu$. Since