Journal of Machine Learning Research 26 (2025) 1-50 Submitted 2/24; Revised 6/25; Published 09/25

Geometry-Dependent Matching Pursuit: a Transition Phase for Convergence on Linear Regression and LASSO

Celine Moucer CELINE.MOUCER@INRIA.FR
Inria, Département d'Informatique de l'Ecole Normale Supérieure, PSL Research University.
Ecole Nationale des Ponts et Chaussées, Marne-la-Vallée, France.

Adrien B. Taylor ADRIEN.TAYLOR@INRIA.FR
Inria, Département d'Informatique de l'Ecole Normale Supérieure, PSL Research University.

Francis Bach FRANCIS.BACH@INRIA.FR
Inria, Département d'Informatique de l'Ecole Normale Supérieure, PSL Research University.

Editor: Silvia Villa

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 tendency to propose 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 in the convergence of (regularized) matching pursuit from underparametrized to overparametrized models.

Keywords: optimization, first-order methods, matching pursuit, linear regression, LASSO

1. Introduction

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

$$\min_{\alpha\in\mathbb{R}^d} \{F(\alpha) + H(\alpha) \triangleq f(P\alpha) + \lambda\|\alpha\|_1\}, \tag{1}$$

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 the dimension of the problem. Typically, in the vanilla least-squares regression problem, $H(\alpha) = 0$ and $F(\alpha) = f(P\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2$, and 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) is exactly the LASSO problem (Tibshirani, 1996), that belongs to more general variational problems appearing in Fenchel duality theory (Bauschke and Combettes, 2017, Section 15.3). Problem (1) is often compared to its constrained counterpart,

$$\min_{\alpha\in\mathbb{R}^d} F(\alpha), \text{ such that } \|\alpha\|_1 \leqslant R, \tag{2}$$

where λ may be seen as the Lagrange multiplier associated to the constraint $\|\alpha\|_1 \leqslant R$ with $R > 0$. Problems (1) and (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 (Dudik et al., 2012). In particular, Problem (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 can be seen as a convex substitute for ℓ_0-penalization problems (Candès and Tao, 2005, Section1.2) for performing feature selection.

First-order methods have become popular to solve optimization Problems (1) and (2), due to their low cost per iteration and to the limited accuracy requirements in machine learning (Bottou et al., 2018, Section 7 and 8). Within these methods, different algorithms might be used, whose choice depends on the properties of functions F and H. For instance, a first-order method may rely on the gradient or the proximal operator (Parikh and Boyd, 2013) as in the proximal gradient method, or the linear minimization oracle as in the Frank-Wolfe algorithm of Jaggi (2013) for a recent constrained version (2). These methods often benefit from convergence guarantees.

In the context of sparsity, traditional first-order methods, such as the proximal gradient, often identify sparse structures but often lacks quantitative guarantees on the number of iterations needed to achieve such solutions (Iutzeler and Malick, 2020, Section 5). As an alternative, boosting strategies (also known as matching pursuit) have been developed to ensure sparse representations of approximate solutions (Mallat and Zhang, 1993; Tropp, 2004). 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 (Tropp, 2004), from an optimization perspective, their convergence analyses often rely on extra statistical assumptions (Zhang, 2011a). More recently, randomized and greedy coordinate descent methods have gained interest due to their low-cost per iteration even in high dimension (Nesterov, 2012) and to their implicit induced sparsity (Beck and Tetruashvili, 2013; Fang et al., 2020).

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 (Locatello et al., 2018), as well as Frank-Wolfe algorithms (Jaggi, 2013; Locatello et al., 2017) for constrained Problems (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 (Friedlander et al., 2014). To our knowledge, this comparison was only drawn for non-regularized problems for which $\lambda = 0$ (Locatello et al., 2018) and for constrained problems (Sun and Bach, 2022).

Problem (1) in $\mathbb{R}^d$ can be naturally formulated as an optimization problem in $\mathbb{R}^n$, letting the gauge geometry appear,

$$\begin{aligned} &\min_{\alpha\in\mathbb{R}^d, x\in\mathbb{R}^n} \; f(x) + \lambda\|\alpha\|_1, \text{ such that } x = P\alpha, \\ &= \min_{x\in\mathbb{R}^n} \; f(x) + \lambda \inf_{\alpha\in\mathbb{R}^d,\; x=P\alpha} \|\alpha\|_1, \\ &= \min_{x\in\mathbb{R}^n} f(x) + \lambda\gamma_{\mathcal{P}}(x), \end{aligned} \tag{3}$$

where the gauge function is defined as $\gamma_{\mathcal{P}}(x) \triangleq \inf_{\alpha \in \mathbb{R}^d,\ x = P\alpha} \|\alpha\|_1$ 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 geometry or to the other, and to formulate geometry-adapted convergence guarantees. For the ℓ_1-geometry, Nutini et al. (2018b, 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 (Nutini et al., 2018b, Appendix 4.1). More generally, d'Aspremont et al. (2018, Section 2) extended smoothness and strong convexity with respect to the gauge $\gamma_{\mathcal{P}}$, which led to formulations of the smoothness parameter as a maximization problem for linear models in the work of Sun and Bach (2022, Section 2.5). These optimization problems are often hard to solve; however they have closed-form reformulation in some cases.

Main results. 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) (Zhang, 2011b), 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) in $\mathbb{R}^d$ and (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 deterministic estimates using SDP relaxations (Goemans and Williamson, 1995), and on the other hand high probability bounds using random matrix theory. As a result, we observe a transition phase 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 (Belkin et al., 2019) 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*, enjoying linear convergence 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).

Organization of the paper. Section 2 examines the convergence properties of gradient descent in the ℓ_2-geometry and coordinate descent with the Gauss-Southwell rule in the ℓ_1-geometry. We interpret these methods as optimization problems within specific geometric frameworks, leading to linear convergence in both underparametrized and overparametrized regimes. The paper introduces two estimation techniques: one based on an SDP relaxation for deterministic bounds, and another using statistical assumptions for high-probability bounds. These techniques reveal a transition phase between the two regimes, illustrated through random feature experiments. In Section 3, we derive a new matching pursuit algorithm adapted to ℓ_1-regularized model, that may be compared with proximal coordinate descent and proximal coordinate descent. Building on the results from Section 2, we establish convergence guarantees dependent on the properties of functions at hand, revealing a strong connection to the proximal coordinate descent with the GS rule. However, in the overparametrized regime, neither proximal gradient nor regularized matching pursuit achieves linear convergence, and we experimentally explore the role of λ in the LASSO as a continuous mapping between low-rank and full-rank solutions to least-squares.

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 (Mallat and Zhang, 1993), and extended to the fitting of weak-learners with 'gradient boosting' techniques such as Adaboost by Freund and Schapire (1999). Typically, when solving problems of the form:

$$\min_{x\in\operatorname{conv}(\mathcal{P})} f(x),$$

where f is a convex function and $\operatorname{conv}(\mathcal{P})$ is the convex hull of some atoms $p \in \mathcal{P}$, matching pursuit (MP) algorithms produce sparse combinations of atoms by picking a direction from a set of atoms using information on the gradient. Typically, it aims at solving problems of the form:

$$\min_{x\in\operatorname{conv}\mathcal{P}} f(x),$$

where f is a convex function and $\operatorname{conv}\mathcal{P}$ is the convex hull of some atoms $p \in \mathcal{P}$. Boosting algorithms are suited to both constrained models, with for example orthogonal matching pursuit (Sheng Chen and Luo, 1989; Tropp, 2004; Zhang, 2011b) or greedy algorithms (Tewari et al., 2011), as well as to unconstrained (penalized) optimization problems, with for example the vanilla boosting strategy (Zhang, 2011a), that minimizes a well-chosen quadratic upper-bound. Recently, Locatello et al. (2017) have unified the framework for matching pursuit and Frank-Wolfe algorithms (Frank and Wolfe, 1956) 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 (2001) opened the path to convergence guarantees for proximal coordinate descent on composite minimization problems (Tseng and Yun, 2009). Nesterov (2012) derived global guarantees for coordinate gradient descent applied on convex objectives, paving the way to families of randomized coordinate updates (Richtárik and Takáč, 2014), and greedy updates (Beck and Tetruashvili, 2013). Yet, these analyses often lead to dimension-dependent convergence guarantees. Nutini et al. (2018b) 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 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 et al. (2019), 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. (2017) 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 produce solutions sparse with respect to atoms, as Sun and Bach (2022) did with the generalized conditional gradient method (Bach, 2015).

Refined 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, 2011b) proposed a forward-backward greedy algorithm for which he derived convergence guarantees under RIP. Similarly, Agarwal et al. (2010) analyzed the proximal gradient and the projected gradient under restricted strong convexity and smoothness, that comes directly from restricted eigenvalue conditions (Raskutti et al., 2010), 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 and Scieur (2020), coming from the convergence analysis of the simplex method (Borgwardt, 1987; Spielman and Teng, 2001). On the contrary, other works improved global convergence guarantees considering well-chosen geometries. For separable quadratics, Nutini et al. (2018b, 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 and Bach (2022, 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 (Nutini et al., 2018b, Appendix 4) or from the geometry such as the pyramidal width (Lacoste-Julien and Jaggi, 2015) or the directional width (Locatello et al., 2017) in Frank-Wolfe techniques, or to the Hoffman constant (Hoffman, 1957) for linear mappings with strongly convex functions (Necoara et al., 2019; Karimi et al., 2016; Guille-Escuret et al., 2021). From a computational perspective, Massias et al. (2017) first introduced variants of coordinate descent with GS rule using working set strategy (focusing on certain relevant features), showing improvement over advanced solvers. Massias et al. (2018) introduced advanced coordinate descent solvers that incorporate screening (eliminating provably irrelevant features), working set techniques and stopping time strategies, which were later improved by Bertrand et al. (2022).

1.2 Assumptions

Convex optimization framework. In this work, functions f into consideration are convex, differentiable and Problem (3) admits 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 (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 $x, y \in \mathbb{R}^n$,

$$f(y) \leqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{L^f}{2}\|y - x\|^2_{\mathbb{R}^n}. \tag{4}$$

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

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu^f}{2}\|y - x\|^2_{\mathbb{R}^n}. \tag{5}$$

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 2.5, F however inherits the Łojasiewicz property with parameter $\mu^F_L > 0$, such that for all $\beta \in \mathbb{R}^d$,

$$\frac{1}{2}\|\nabla F(\beta)\|^2_{\mathbb{R}^d,\star} \geqslant \mu^F_L(F(\beta) - F_\star), \tag{6}$$

where $\|\cdot\|_{\mathbb{R}^d,\star}$ is the dual norm for $\|\cdot\|_{\mathbb{R}^d}$.

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 P (i.e., geometry). On the other hand, we propose high probability bounds of these parameters, relying on random matrix theory. Random matrices often appears in statistical assumptions, such as with restricted isometry property (Candès and Tao, 2005) or the restricted eigenvalue condition (Raskutti et al., 2010). In the machine learning literature, random matrices appear in average-case analysis for quadratics (Pedregosa and Scieur, 2020) with the Marchenko-Pastur distribution (Marchenko and Pastur, 1967), or when studying the double descent phenomena for the generalization error (Belkin et al., 2019; Mei and Montanari, 2022; Bach, 2024) 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. Transition Phase for Linear Regression

We begin with the study of a linear regression problem, where problem (7) is a special case of the optimization Problem (1) with $\lambda = 0$,

$$\min_{\alpha \in \mathbb{R}^d} \{F(\alpha) = f(P\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2\}, \tag{7}$$

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 (Karimi et al., 2016; Nutini et al., 2018b) 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, that is, as optimization problems in the geometry under consideration. This interpretation leads to linear convergence both in the underparametrized and the overparametrized regime, letting smoothness and strong convexity parameters appear, that are adapted to the geometry. For characterizing convergence properties of these methods, we provide estimates of these quantities. A first technique developed in this work is based on an SDP relaxation, and leads to deterministic estimates. A second technique, inspired from statistical and average-case analyses, leads to high probability bounds under statistical assumptions on the data. These estimates let a transition phase 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.

2.1 Strong Convexity and Smoothness Constants as Solutions to Optimization Problems

First, let us compute estimates of smoothness and strong convexity parameters by formulating their computation as optimization problems in a generic norm for the least-squares minimization (7). 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 1 and possibly equal to 0 (especially with dimension $d > n$).

Lemma 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 (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 $\beta \in \mathbb{R}^d$, $\|P\beta\|_2^2 \geqslant \mu^F\|\beta\|^2$, from which both formulations follow. ■

In Lemma 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 (6) with μ^{F_L}. Again μ_L^F is formulated as an optimization problem.

Lemma 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* (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 Let $\alpha_\star$ be a minimizer for F such that $P^\top y = P^\top P\alpha_\star$ and $F(\alpha_\star) = 0$. Then, for all $\alpha \in \mathbb{R}^d$: $\|\nabla F(\alpha)\|_\star^2 = \frac{1}{n^2}\|P^\top P(\alpha - \alpha_\star)\|_\star^2 \geqslant \frac{1}{n}\frac{\|P^\top P(\alpha-\alpha_\star)\|_\star}{\|P(\alpha-\alpha_\star)\|_2^2}\frac{1}{n}\|P(\alpha-\alpha_\star)\|_2^2$. Thus, $\frac{\|\nabla F(\alpha)\|_\star^2}{F(\alpha)} \geqslant \mu_L^F = \frac{1}{n}\inf_{\beta\in\mathbb{R}^d, \|P\beta\|_2^2\geqslant 1}\|P^\top P\beta\|_\star$, which is the Łojasiewicz inequality with parameter μ_L^F. ■

Again in Lemma 2, μ_L^F is formulated as a (nonconvex) minimization problem. The two quantities μ^F and μ_L^F are compared in Lemma 3, with equality in the underparametezed regime in which $P^\top P$ is invertible. Again, this result holds in different norms.

Lemma 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. μ_L^F) defined in Lemma 1 (resp. Lemma 2). If $P^\top P$ is invertible, $\mu_L^F = \mu^F$.*

Proof Let us consider the squared-root formulations of μ^F and μ_L^F given in Lemma 1 and Lemma 2.

$$\begin{aligned}\frac{1}{\sqrt{n\mu^F}} &= \sup_{\|P\beta\|_2\leqslant 1}\|\beta\| = \sup_{\|z\|_\star\leqslant 1, \|P\beta\|_2\leqslant 1}\langle\beta, z\rangle,\\ \frac{1}{\sqrt{n\mu_L^F}} &= \sup_{\|P^\top P\nu\|_\star\leqslant 1}\|P\nu\|_2 = \sup_{\|P^\top P\nu\|_\star\leqslant 1, \|P\beta\|_2\leqslant 1}\langle P\beta, P\nu\rangle = \sup_{\|P^\top P\nu\|_\star\leqslant 1, \|P\beta\|_2\leqslant 1}\langle\beta, P^\top P\nu\rangle.\end{aligned}$$

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 $\mu_L^F \geqslant \mu^F$. In the special case where $P^\top P$ is invertible, $\mathrm{Im}(P^\top P) = \mathbb{R}^d$, and $\mu_L^F = \mu^F$. ■

In the next sections, we study the role of these parameters in the convergence guarantees of gradient descent and steepest coordinate descent, both in the underparametrized and overparametrized regime. We then propose deterministic estimates for μ^F and μ_L^F, as well as high probability bounds based on a simple random model for P. Finally, we observe a transition phase, that appear for both gradient descent and steepest coordinate descent, from the underparametrized to the overparametrized regime as highlighted in Figure 1.

2.2 Gradient Descent in the ℓ_2-geometry

We are interested in the 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

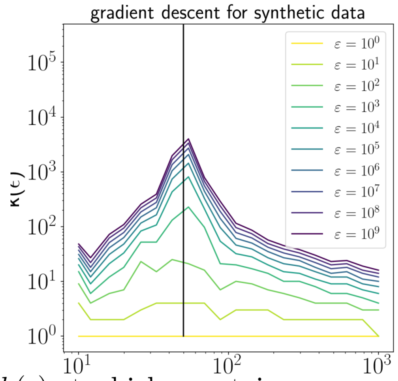

Figure 1: Iteration number $k(\epsilon)$ at which a certain accuracy ϵ is reached for gradient descent on synthetic quadratics with $n = 50$ and for different values of the dimension.

interpretation of gradient descent with fixed step size $\gamma = \frac{1}{L_2^F}$ comes from the minimization of a quadratic (smoothness) upper bound on F:

$$\alpha_1 = \alpha_0 - \frac{1}{L_2^F} \nabla F(\alpha_0) = \alpha_0 - \frac{1}{L_2^F} P^\top (P\alpha_0 - y). \tag{8}$$

In the underparametrized regime, the function F is μ_2^F-strongly convex with respect to the ℓ_2-norm, with $\mu_2^F = \lambda_{\min}(\frac{P^\top P}{n}) \geqslant 0$. In the underparametrized regime, we assume in addition that $\mu_2^F = \lambda_{\min}(\frac{P^\top P}{n}) > 0$. As a result, gradient descent (8) converges linearly. However, in the overparametrized regime in which $d > n$, $\mu_2^F = 0$, F is not strongly convex. Yet, gradient descent still converges linearly (Bolte et al., 2010), since quadratics benefit from the Łojasiewicz inequality, with $\mu_{L,2}^F = \frac{1}{n} \lambda_{\min}(PP^\top) \geqslant 0$. As defined earlier, in the overparametrized regime, we assume in addition that $\mu_{L,2}^F = \frac{1}{n} \lambda_{\min}(PP^\top) > 0$.

Proposition 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 $0 \leqslant \mu_2^F \leqslant L_2^F$ and $0 \leqslant \mu_{2,L}^F \leqslant L_2^F$. Let $(\alpha_k)_{k \in \mathrm{N}}$ be generated by gradient descent in (8) 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 B. ■

Convergence speeds obtained in Proposition 4 depend on $\lambda_{\max}(P^\top P/n)$, $\lambda_{\max}(P^\top P/n)$ and $\lambda_{\min}(PP^\top/n)$. In the case where P is generated randomly, we can derive estimates of these extremal eigenvalues, avoiding a full computation of the extremal eigenvalues, and thus, an approximate convergence guarantee of the method. In the following, we consider random data P, with i.i.d. entries having the same variance, so that $P^\top P$ and $PP^\top$ have a limiting Marchenko-Pastur distribution (Marchenko and Pastur, 1967), whose extremal eigenvalues are known. This distribution generalizes the Wishart distribution of $P^\top P$ and $PP^\top$, obtained from Gaussian data P.

Theorem 5 (Limits of extreme eigenvalues - Theorem 5.11 (Bai and Silverstein, 2010)) *Assume $P \in \mathbb{R}^{n\times d}$, where each entry is an i.i.d. random variable with mean 0, variance σ^2, $\mathbb{E}[P_{i,j}^4] < +\infty$ and let $H = \frac{1}{n}P^\top P$. If $\frac{d}{n} \to r \in (0,\infty)$, then we have almost surely that*

$$\mu \to \sigma^2(1-\sqrt{r})^2,$$
$$L \to \sigma^2(1+\sqrt{r})^2,$$

where μ and L are respectively the minimal and maximal non-zero eigenvalues for H.

Combining Theorems 5 with Proposition 4, we obtain natural estimates of the convergence properties in the underparametrized and the overparametrized regimes for random P.

Corollary 6 *Under the same assumptions than Theorem 5 and assuming $\mathbb{E}[\|P\|^4] < +\infty$, gradient descent with step size $\gamma = \frac{1}{L_2^F}$ converges linearly to the optimum. Then, if $\frac{d}{n} \to r \in (0,\infty)$, we almost surely have that*

- *In the underparametrized regime, assuming in addition that $r \ll 1$: $1 - \frac{\lambda_{\min}(P^\top P)}{\lambda_{\max}(P^\top P)} \to 1 - \frac{(1-\sqrt{r})^2}{(1+\sqrt{r})^2}$,*
- *In the overparametrized regime, assuming in addition that $r \gg 1$: $1 - \frac{\lambda_{\min}(PP^\top)}{\lambda_{\max}(P^\top P)} \to 1 - r\frac{(1-\sqrt{1/r})^2}{(1+\sqrt{r})^2} = 1 - \frac{(\sqrt{r}-1)^2}{(1+\sqrt{r})^2}$.*

Proof First, we apply the Marchenko-Pastur Theorem 5 $H = \frac{P^\top P}{n}$ with $\frac{d}{n} \to r$ and to $I = \frac{PP^\top}{d}$ with $\frac{n}{d} \to u = \frac{1}{r}$. Together with Proposition 4, it leads to the result. ■

Given that $\frac{d}{n} \to r$, we deduce approximate convergence guarantees from Corollary 6: assuming $\frac{d}{n} \to r \ll 1$ (which is included in the underparametrized regime), we have $1 - \frac{\lambda_{\min}(P^\top P)}{\lambda_{\max}(P^\top P)} = 4\sqrt{r} + o(\sqrt{r}) \approx 4\sqrt{\frac{d}{n}}$, and assuming $\frac{d}{n} \to r \gg 1$ (which is included in the overparametrized regime), $1 - \frac{\lambda_{\min}(PP^\top)}{\lambda_{\max}(P^\top P)} = 4\sqrt{1/r} + o(\sqrt{1/r}) \approx 4\sqrt{\frac{n}{d}}$. These approximate convergence rates should be compared to the average-case analysis of Pedregosa and Scieur (2020) for least-squares problems, and to the polynomial-based analysis for convergence of gossip developed by Berthier et al. (2020). Depending on the distribution under consideration, Scieur et Pedregosa developed average-case optimal accelerated methods, whose limit in the number of iterations happens to be the Polyak-Momentum (Scieur and Pedregosa, 2020). Its worst-case convergence guarantee verifies $\frac{\sqrt{L_2^F}-\sqrt{\mu_2^F}}{\sqrt{L_2^F}+\sqrt{\mu_2^F}} \approx \sqrt{\frac{d}{n}}$, can be compared to Nesterov's accelerated gradient method (Nesterov, 1983) with $1 - \sqrt{\frac{\mu_2^F}{L_2^F}} \approx 2\sqrt{\frac{d}{n}}$ and to gradient descent with $1 - \frac{\mu_2^F}{L_2^F} \approx 4\sqrt{\frac{d}{n}}$. When considering their average-case guarantees, only a polynomial sublinear term is added (Paquette et al., 2023, Table 2) to the worst-case guarantee, without major modifications in the linear term. In other words, Polyak-Momentum (resp. Nesterov's accelerated gradient method) converges four times (resp. twice) as fast as gradient descent with fixed step sizes.

We conclude from Corollary 6 that convergence of gradient descent depends on the degree of underparametrization (resp. overparametrization). The more independent samples (resp. features), the better the convergence. While the advantage of adding independent samples is well known for improving both learning and convergence speed, it appears that the larger the set of features, the better the convergence. We are now going the study approximate convergence guarantees of coordinate descent, that appears in the ℓ_1-geometry.

2.3 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 minimization of a smoothness upper bound with respect to the ℓ_1-norm, as shown by Nutini et al. (2018b, Section 4), for all $\alpha_0, \alpha \in \mathbb{R}^d$,

$$F(\alpha) \leqslant F(\alpha_0) + \langle \nabla F(\alpha_0), \alpha - \alpha_0 \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_0\|_1^2. \tag{9}$$

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 an extremal point of the simplex). Gauss-Southwell coordinate descent follows by minimizing over $\alpha \in \mathbb{R}^d$, for a fixed $\alpha_0 \in \mathbb{R}^d$,

$$\begin{aligned} i_0 &= \operatorname*{arg\,max}_{k=1,\dots,d} |\nabla_{i_k} F(\alpha_0)|, \\ \alpha_1 &= \alpha_0 - \frac{1}{L_1^F} \nabla_{i_0} F(\alpha_0) e_{i_0}. \end{aligned} \tag{10}$$

As for gradient descent, its convergence speed depends on the parametrization regime. Depending on n and d, F may be μ_1^F-strongly convex, or verify the Łojasiewicz inequality with parameter $\mu_{1,L}^F$. Both μ_1^F and $\mu_{1,L}^F$ can be formulated as optimization problems for computing explicit estimates. It follows from the strong convexity characterization given in Lemma 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 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 7 states that coordinate descent converges linearly, as already proven by Karimi et al. (2016, Theorem 1).

Proposition 7 *(Karimi et al., 2016, Theorem 1) Let F be convex, L_1^F-smooth with respect to the norm $\|\cdot\|_1$, be μ_1^F-strongly convex and verify the Łojasiewicz inequality with $\mu_{1,L}^F$, where $0 \leqslant \mu_{1,L}^F \leqslant L_1^F$ and $0 \leqslant \mu_1^F \leqslant L_1^F$. Let (α_k) be generated by coordinate gradient descent* (10) *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 B. ■

The convergence guarantee provided in Proposition 7 depends on μ_1^F and $\mu_{1,L}^F$ and is hence complicated to compute. 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.

In the following, we construct estimates to these quantities, so that they may be computed a priori. First, we provide SDP relaxations for the optimization problems defining μ_1^F and $\mu_{1,L}^F$, that may differ from the exact solution. We thus propose to construct high probability bounds for μ_1^F and $\mu_{1,L}^F$, assuming randomly generated data.

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 8 *Let $P \in \mathbb{R}^{n\times d}$, and let us define $\mu_1^F = \frac{1}{n}\inf_{\|z\|_1^2 \geqslant 1} \|Pz\|_2^2$ and $\mu_{1,L}^F = \inf_{\|P\beta\|_2^2 \geqslant 1} \|P^\top P\beta\|_\infty^2$. Then the following inequality holds*

- *in the underparametrized regime, let $\tilde{\mu}_1^F$ be defined by $\frac{1}{n\tilde{\mu}_1^F} = \sup_{X\succcurlyeq 0} \operatorname{Tr}((P^\top P)^{-1}X), \;\text{ s.t. } \operatorname{diag}(X) \leqslant 1$. Then,*
$$1 - \frac{\pi}{2}\frac{\tilde{\mu}_1^F}{L_1^F} \leqslant 1 - \frac{\mu_1^F}{L_1^F} \leqslant 1 - \frac{\tilde{\mu}_1^F}{L_1^F},$$
- *in the overparametrized regime, let $\tilde{\mu}_{1,L}^F$ be defined by $\frac{1}{n\tilde{\mu}_{1,L}^F} = \sup_{X\succcurlyeq 0} \operatorname{Tr}(P^\top PX) \;\text{ s.t. } \|P^\top PXP^\top P\|_\infty \leqslant 1$. Then, i*
$$1 - \frac{\mu_{1,L}^F}{L_1^F} \leqslant 1 - \frac{\tilde{\mu}_{1,L}^F}{L_1^F},$$

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

Proof See Appendix C.1. ■

In Proposition 8, we find out SDP relaxations that yield a deterministic estimate 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.

High probability bounds. We now assume that P is randomly generated, as in the ℓ_2-geometry. Under subgaussian assumptions, we derive in Proposition 9 high probability bounds for 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 around $\frac{\sigma^2}{d}$ and $\mu_{1,L}^F$ around $\frac{\sigma^2}{n}$ with subgaussian tails.

Proposition 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 absolute constants $C, C_1, C_2, C_3, C_4, K > 0$ such that,*

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

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

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

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

3. *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 C.2).*

Proof See Appendix C.2. ■

Compared with Proposition 4, Proposition 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.

From Proposition 7, we have seen that coordinate descent with GS-rule (10) converges in function values with a rate $1-\frac{\max(\mu_{1,L}^F,\mu_1^F)}{L_1^F}$. In the overparametrized regime (resp. underparametrized), we conclude in Coroallary 10 with limiting concentration of the convergence rate for large dimensions (resp. large number of samples).

Corollary 10 *Let* $P\in\mathbb{R}^{n\times d}$, *with* $P_i\in\mathbb{R}^d$ *i.i.d. subgaussian such that* $\mathbb{E}[P_{i,j}]=0$, $\mathbb{E}[P_{i,j}]=\sigma^2$. *Then,*

- *in the underparametrized regime, when* $n\to\infty$, *the quantity* $1-\frac{\mu_1^F}{L_1^F}$ *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 quantity* $1-\frac{1}{\mu_{1,L}^F}$ *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 C.3. ■

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 Appendix B. Unlike the approximate convergence guarantees for gradient descent in the underparametrized regime (resp. overparametrized) detailed in Corollary 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 to its linear convergence in both the underparametrized and overparametrized regime. For a linear regression problem, neither the strong convexity parameter nor 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 their expectations with subgaussian tails. In the next section, we perform numerical experiments showing the transition phase between the two regimes.

2.4 A Transition Phase Phenomenon: Experimental Results

We compare approximate convergence guarantees to numerical experimental convergence for gradient descent from Corollary 6 and for coordinate descent from Corollary 10. More precisely, we verify the expected transition phase 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 a few experiments on several datasets: least-squares problems obtained either with synthetic Gaussian vectors or from the Leukemia dataset or from random features, described below.

Synthetic quadratics. We consider several least-squares problems (7), where the number of samples $n = 50$ is fixed, and dimension d varies so that both the overparametrized and the underparametrized regimes are explored. In this model, the feature matrix P into consideration 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 (that is, the number of non-zero entries) equal to $s = 8 < d$, and $y = P\alpha_\star + \epsilon$, where $\epsilon_i \sim \mathcal{N}(0, \sigma)$.

Leukemia dataset. We consider the standard Leukemia dataset (Golub et al., 1999), where $n = 72$ and $d = 7129$. Again, we consider submatrices, so that the dimensions vary from the underparametrized regime to the overparametrized one. For each model, P was centered and scaled to have a unit variance.

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

In Figure 2, we plot the iteration number $k(\epsilon)$ at which a certain accuracy ϵ is reached for the three models described above, both for gradient descent and coordinate descent with GS-rule. We consider accuracies $\epsilon = \{10^{-0}, \ldots, 10^{-9}\}$ and we refer to these curves by ϵ-curves. For the three models, both steepest coordinate descent and gradient descent converge faster

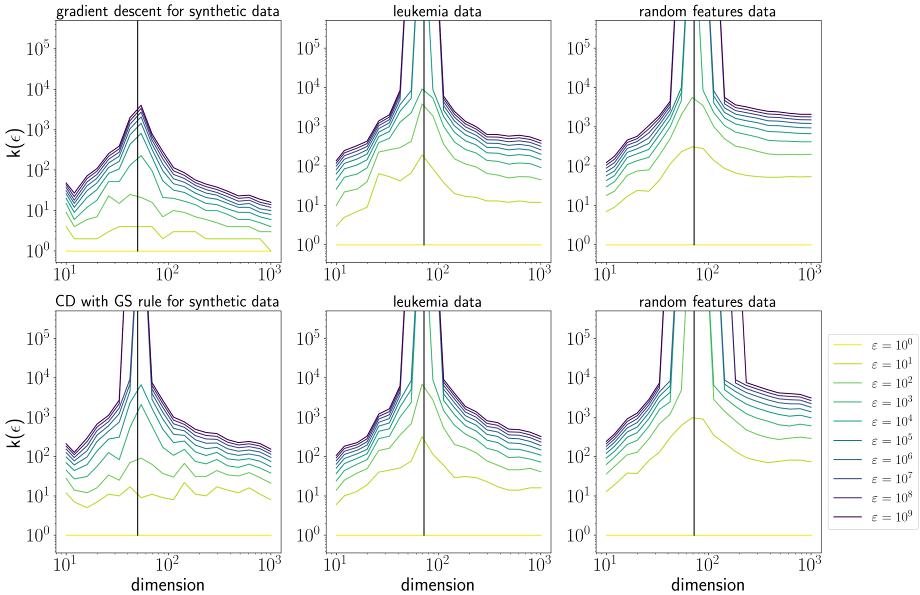

Figure 2: ϵ-curve for gradient descent (top) and coordinate descent with the GS-rule (bottom), for the three models: synthetic quadratics (on the left) with $n = 50$, the leukemia dataset (in the middle) with $n = 72$, a random feature model (on the right) with $n = 72$.

when $n \gg d$ (resp. $d \gg n$) in the underparametrized (resp. overparametrized) regime in Figure 2. 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. (2019), and formalized by Mei and Montanari (2022). For a fixed prediction model, as the number of features increases, the excess risk is U-shaped for underparametrized optimization models and goes down for overparametrized models. As for the excess risk, we observe a transition at $d \approx n$ as well as a better precision for overparametrized models. Contrary to the generalization error, underparametrized models ($d \ll n$) perform well even when $\frac{d}{n} \to 0$ and are not U-shaped. We refer to this phenomenon as a transition phase for gradient and coordinate descent.

In Figure 3, 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 6 matches the observed 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 8 for 'small' values of d and n, and to its high probability bound otherwise. In both cases, we numerically recover that convergence is improved as the dimension increases.

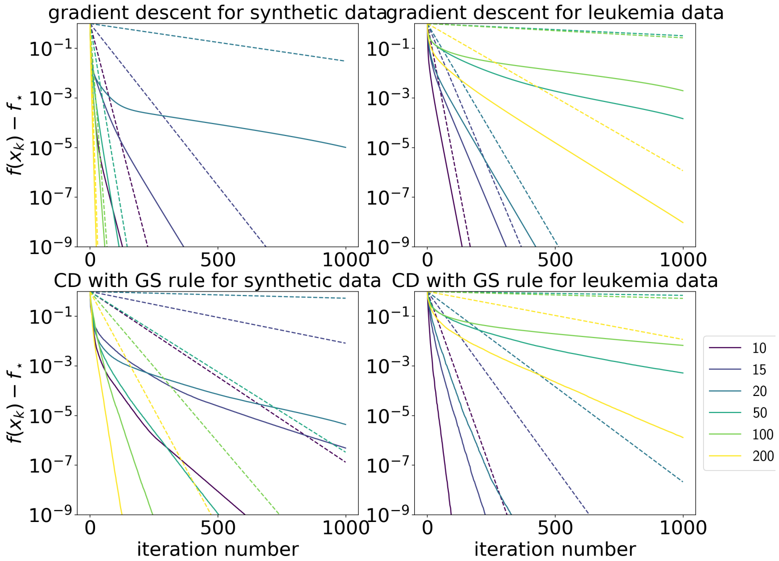

Figure 3: Convergence in function value for gradient descent and coordinate descent with GS rule, on synthetic quadratics ($n = 20$) and on the leukemia dataset ($n = 72$), for different values for d. Dashed lines: comparison to the approximate convergence guarantees from Corollary 6 for synthetic quadratics, and to high probability estimates for the leukemia dataset from Proposition 9.

2.5 Coordinate Descent Is an Instance of Matching Pursuit

Coordinate descent, as well as gradient descent, converge linearly both in the underparametrized and overparametrized regime, as provided by Proposition 7, respectively due to strong convexity and the Łojasiewicz property. Given a certain structure on f, we prove that $F(\cdot) = f(P\cdot)$ inherits some regularity properties from f and that coordinate descent can be interpreted as a matching pursuit algorithm in either $\mathbb{R}^d$ or $\mathbb{R}^n$. Now, let us consider the more general formulation,

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

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

Underparametrized regime. F is μ_1^F-strongly convex (since $P^\top P$ is invertible). The connection between coordinate descent with GS-rule (10) and matching pursuit was highlighted by Locatello et al. (2018). Considering the set of unitary direction $\mathcal{A} = \operatorname{conv}(\{\pm e_i, i = 1, \ldots, d\})$, that is the ℓ_1 unit ball, coordinate descent may be rewritten as a matching pursuit:

$$\begin{aligned} e_{i_0} &\in -\text{LMO}_{\mathcal{A}}(\nabla F(\alpha_0)) = -\operatorname*{arg\,min}_{i=1,\ldots,d} \nabla F(\alpha_0)^\top e_i, \\ \alpha_1 &= \alpha_0 - \frac{1}{L_1^F} \nabla F(\alpha_0)^\top e_{i_0}, \end{aligned}$$

where $\mathrm{LMO}_{\mathcal{A}}(\nabla F(\alpha_0)) = \inf_{z \in \mathcal{A}} \langle \nabla F(\alpha_0), z \rangle$. Steepest coordinate descent converges linearly from Proposition 7, with the same convergence guarantee as in the context of matching pursuit (Locatello et al., 2018, Theorem 5).

Overparametrized regime. F does not inherit strong convexity. Yet, for least-squares, F but does inherit some structure from f (see Lemma 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 11 ensures $\gamma_{\mathcal{P}}(\cdot)$ is a norm in the overparametrized regime.

Lemma 11 *Let $\alpha \in \mathbb{R}^d \mapsto P\alpha \in \mathbb{R}^n$ be a surjection in $\mathbb{R}^d$, and $\mathcal{P} = \mathrm{conv}(P)$ 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 D.1. ■

Let f be convex, L_2^f-smooth and μ_2^f-strongly convex with respect to the ℓ_2-norm. We define $L_{\mathcal{P}}^f = L_2^f \sup_{j=1,\ldots,d} \|P_j\|_2^2$ and $\mu_{\mathcal{P}}^f = \mu_2^f \inf_{z \in \mathbb{R}^n} \|P^\top z\|_\infty^2$, such that $\|z\|_2^2 = 1$. Then, f is convex, $L_{\mathcal{P}}^f$-smooth and $\mu_{\mathcal{P}}^f$-strongly convex with respect to the norm $\gamma_{\mathcal{P}}(\cdot)$. Using the surjection of $\alpha \in \mathbb{R}^d \mapsto P\alpha \in \mathbb{R}^n$, for all $x \in \mathbb{R}^n$, it holds from the definition that $L_2^f \|x\|_2^2 \leqslant L_{\mathcal{P}}^f \|x\|_2^2$ and that $L_2^f \|x\|_2^2 \leqslant L_{\mathcal{P}}^f \|x\|_1^2$ from the norm equivalence. Thus, $L_2^f \|x\|_2^2 \leqslant L_{\mathcal{P}}^f \gamma_{\mathcal{P}}(x)^2$. Similarly, for all $x \in \mathbb{R}^n$, $\mu_2^f \|x\|_2^2 \geqslant \mu_{\mathcal{P}}^f \gamma_{\mathcal{P}}(x)^2$. Thus f is $L_{\mathcal{P}}^f$-smooth and $\mu_{\mathcal{P}}^f$-strongly convex with respect to $\gamma_{\mathcal{P}}$ (that is a norm in this regime).

As before, our estimates for smoothness (resp. strong convexity) parameters are obtained by an optimization problem. They appear to be closely related to the parameters for least-squares from Lemma 1 and 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, we have that $L_{\mathcal{P}}^f = L_1^F$ as defined in (9), and $\mu_{\mathcal{P}}^f = \mu_{1,L}^F$ as soon as $PP^\top$ is invertible (which is the case here). Multiplying (10) by P, noticing that $\min_{e, \|e\|_1 = 1} \langle \nabla F(\alpha), e \rangle = \min_{p \in \mathcal{P}} \langle \nabla f(x), p \rangle$, coordinate descent with the GS-rule on F can be formulated as matching pursuit on f,

$$
\begin{aligned}
& z_0 \in \mathrm{LMO}_{\mathcal{P}}(\nabla f(x_0)), \\
& x_1 = x_0 - \frac{1}{L_{\mathcal{P}}^f} \langle \nabla f(x_0), z_0 \rangle z_0.
\end{aligned}
\tag{11}
$$

Let x_k be generated by matching pursuit (11), starting from $x_0 \in \mathbb{R}^n$ for $L_{\mathcal{P}}^f$-smooth and $\mu_{\mathcal{P}}^f$-strongly convex functions, then, Locatello et al (Locatello et al., 2018, Theorem 5) proved linear convergence of the sequence with

$$
f(x_k) - f_\star \leqslant \left(1 - \frac{\mu_{\mathcal{P}}^f}{L_{\mathcal{P}}^f}\right)(f(x_0) - f_\star). \tag{12}
$$

By construction, since $x_k = P\alpha_k$, we have that $F(\alpha_k) - F_\star \leqslant (1 - \frac{\mu_{\mathcal{P}}^f}{L_{\mathcal{P}}^f})(F(\alpha_0) - F_\star) = (1 - \frac{\mu_{1,L}^F}{L_1^F})(F(\alpha_0) - F_\star)$. The same result could have been derived from Proposition 7 and the observation that strongly convex functions composed with a linear mapping verify a Łojasiewicz-inequality.

Lemma 12 *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 $\alpha \in \mathbb{R}^d$,*

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

Proof Let $x \in \mathbb{R}^n$. By minimizing in y both sides in the strong convexity equation, we get $f_\star \geqslant f(x) - \sup_y \langle -\nabla f(x), y - x\rangle - \frac{\mu^f_{\mathcal{P}}}{2}\gamma^2_{\mathcal{P}}(y - x) \geqslant f(x) -$ $(\frac{\mu^f_{\mathcal{P}}}{2}\gamma^2_{\mathcal{P}}(\cdot))^\star(-\nabla f(x)) \geqslant f(x) - \frac{1}{2\mu^f_{\mathcal{P}}}\|P^\top \nabla f(x)\|^2_\infty$, by definition of the Fenchel dual. 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 12 corresponds to the result of Karimi et al. (2016, 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 (Necoara et al., 2019, Section 3 and 4.1) by $\theta(P) = \max_{z, \|P^\top z\|_\infty = 1} \|z\|^2_2$. They indeed proved that F verifies a Łojasiewicz inequality, for all $\alpha \in \mathbb{R}^d$, $\frac{1}{2}\|\nabla F(\alpha)\|^2_2 \geqslant \theta(P)\mu^F(F(\alpha) - F_\star)$.

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 regularity 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, global values of the smoothness and strong-convexity parameters 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 generalize this framework for analyzing penalized linear models.

3. Transition Phase for Penalized Linear Models

We now consider the penalized linear model,

$$\min_{\alpha \in \mathbb{R}^d} \{G(\alpha) \triangleq f(P\alpha) + \lambda\|\alpha\|_1\}, \tag{13}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is L^f_2-smooth, μ^f_2-strongly convex, (and thus, the function $F : \mathbb{R}^d \to \mathbb{R}$ such that $F(\cdot) = f(P\cdot)$, is L^F_2 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 results 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. (Combettes and Wajs, 2005)) was developed for such 'composite' convex optimization problems. Given a starting point $\alpha_0 \in \mathbb{R}^d$, each iterate is obtained by minimizing a smooth quadratic upper bound on F:

$$G(\alpha) \leqslant F(\alpha_k) + \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L_2^F}{2} \|\alpha_k - \alpha\|_2^2 + \lambda \|\alpha\|_1. \tag{14}$$

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

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

The proximal gradient method converges sublinearly if F is smooth and 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 ((Taylor et al., 2018, Theorem 2.1)),

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

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. (2018b) analyzed coordinate descent with the Gauss-Southwell selection rule, that tends to perform better than randomized coordinate descent. The GS rule corresponds to choosing $i_k = \arg\min_l \min_{t \in \mathbb{R}} \ \nabla_l F(P\alpha_k)(t - \alpha^{(l)}) + \frac{L_2^F}{2}(t - \alpha^{(l)})^2 + \lambda|t|$, and then

$$\alpha_{k+1} = \underset{\alpha \in \mathbb{R}^d}{\arg\min} \ \nabla_{i_k} F(\alpha_k)(\alpha^{(i_k)} - \alpha_k^{(i_k)}) + \frac{L_2^F}{2}(\alpha^{(i_k)} - \alpha_k^{(i_k)})^2 + \lambda|\alpha^{(i_k)}|. \tag{15}$$

Nutini et al. (Nutini et al., 2018b, 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 lets a sublinear dependence in the parameter μ_1^F appear, is mentioned in (Nutini et al., 2018b, Appendix 8). Coordinate descent with GS-rule is closely related to matching pursuit as for nonpenalized models, as detailed in Appendix E, 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 (13), that we compare to classical boosting algorithm and coordinate descent with GS-rule. After that, we compute convergence guarantees for smooth (possibly strongly) convex functions. Finally, we interpret the convergence regimes as a function of 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 (14) 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:

$$\alpha_{k+1} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^d} \langle P^\top \nabla f(P\alpha), \alpha - \alpha_k \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_k\|_1^2 + \lambda \|\alpha\|_1. \tag{16}$$

Whereas the optimization steps in proximal gradient descent (14) and proximal coordinate descent with the GS rule (15) can be decomposed coordinate-wise, the function $\alpha \mapsto \|\alpha\|_1^2$ is not separable. Based on the same upper bound (16), Song et al. (2017, Algorithm 1) generalized greedy coordinate descent with the "SOft ThresOlding PrOjection" (SOTOPO) algorithm using a reweighted least-squares formulation (Appendix A). 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 (16) as a matching pursuit algorithm, that only calls for a linear minimization oracle. Using a variational trick detailed in Appendix A to approach $\|\beta\|_1^2$, we begin by formulating Problem (16) starting from $\alpha_k \in \mathbb{R}^d$ as a separable 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 separable in each coordinate $\beta^{(i)}$, and can be reduced to an optimization problem in $z \geqslant 0$ in Lemma 13.

Lemma 13 *The optimization step* (16) *can be reformulated as*

$$V_\star = \max_{z_{\min} \leqslant z} \left\{ h(z) \triangleq -\frac{z^2}{2L_1^F} + \sum_{i \in I} \min\left(\lambda |\alpha_k^{(i)}|, -\nabla_i F(\alpha_k) \alpha_k^{(i)} + z|\alpha_k^{(i)}|\right) \right\}, \tag{17}$$

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 and only 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 13 leads to a convex constrained optimization problem in $\mathbb{R}^+$, whose objective is the sum of a quadratic and piecewise linear functions whose slope coefficients change 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 (16) by constructing a solution as sparse as possible, we introduce Assumption 14.

Assumption 14 *The algorithm only selects 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 the minimum of h explicitly. Under Assumption 14, Lemma 15 first deals with the situation in which the objective is quadratic, that is for all $i \in I$, $z_i \geqslant z_{\min}$.

Lemma 15 *Let* $(z_\star, \beta_\star)$ *be a solution to* (17) $\lambda + \nabla_i F(\alpha_k)\alpha_k^{(i)}/|\alpha_k^{(i)}|$. *Assume* $\{i, z_i \geqslant z_{\min}\} = \emptyset$ *and verify Assumption 14. Then* $z_\star = z_{\min}$, $\beta_\star^{(i_{\min})} = -\mathrm{sign}(\nabla_{i_{\min}} F(\alpha_k))\frac{z_{\min}}{L_1^F}$ *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 15 and Assumption 14, 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 16 *Let* $z_\star, \beta_\star$ *be a solution of* (17) *and assume* $\mathcal{S} = \{i, z_i \geqslant z_{\min}\} \neq \emptyset$ *and verify Assumption 14. There are four possible solutions to Problem* (17),

- *If* $\frac{dh}{dz}(z_{\min}) \leqslant 0$, *then* $z_\star = z_{\min}$.
  In addition, $\beta_\star^{(i)} = \begin{cases} -\alpha_k^{(i)} & \text{if } z_i \geqslant z_\star, \\ 0 & \text{if } z_i \leqslant z_\star, \\ -\mathrm{sign}(\nabla_{i_{\min}} F(\alpha_k))\frac{z_{\min} - \sum_{i\in\mathcal{S}} |\alpha_k^{(i)}|}{L_1^F} & \text{if } i = i_{\min}. \end{cases}$

- *If there exists* $k \in \mathcal{S}$ *such that* $\frac{dh}{dz}(z_k^+) \geqslant 0$ *and* $\frac{dh}{dz}(z_{k+1}^-) \leqslant 0$, *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* $\frac{dh}{dz}(z_k^-) \geqslant 0$ *and* $\frac{dh}{dz}(z_k^+) \leqslant 0$ *then* $z_\star = z_k$. *In addition,*
  $\beta_\star^{(i)} = \begin{cases} -\alpha_k^{(i)} & \text{if } z_i > z_\star, \\ 0 & \text{if } z_i < z_\star, \\ -\mathrm{sign}(\alpha_k^{(i)})\left(\frac{z_k}{L_1^F} - \sum_{i, z_i > z_k} |\alpha_k^{(i)}|\right) & \text{if } i = k. \end{cases}$

- *If* $\frac{dh}{dz}(z_{|I|}) > 0$, *then for all* $i \in I$, $z_\star > z_i$ *and* $\beta_\star^{(i)} = \begin{cases} -\mathrm{sign}(\alpha_k^{(i)})\frac{z_i}{L_1^F} & \text{if } i = |I|, \\ 0 & \text{otherwise.} \end{cases}$

Proof The function h is strictly concave and piecewise quadratic on $[z_i, z_{i+1}]$. The solution to the optimization Problem (17) is thus obtained by studying the sign of $\frac{dh}{dz}(\cdot)$ at z_i^- and z_i^+. By construction of the solution given in the proof of Lemma 13, 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 15 and 16 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 1. In short, at each iteration, Algorithm 1 performs one of the three possible actions: either one new atom is added (at most) by calling the $\mathrm{LMO}(\nabla F(\alpha_k)) = \arg\max_i |\nabla_i F(\alpha_k)| = \arg\max_{p \in \mathcal{P}} p^\top \nabla f(P\alpha_k)$ 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

Algorithm 1 Regularized matching pursuit (RMP)

```
$\alpha \in \mathbf{R}^d$, $N \in \mathbf{N}$
for $k \in [0, \ldots, N]$ 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
        $\beta^{(i_{\min})} = -\mathrm{sign}(\nabla_{i_{\min}} F(\alpha_k)) \frac{z_{\min}}{L_1^F}$
    else
        Compute $u = \mathrm{argmin}_i \{z_i \geqslant z_{\min}\}$ and for $i \in [u, v]$, compute $\frac{dh}{dz}(z_i)$
        if $\frac{dh}{dz}(z_{\min}) \leqslant 0$ or $\frac{dh}{dz}(z_u) \leqslant 0$ then
            For $i \in [u, v]$, $\beta^{(i)} = -\alpha_k^{(i)}$
            If $\frac{dh}{dz}(z_{\min}) \leqslant 0$, then $\beta^{(i_{\min})} = -\mathrm{sign}(\nabla_{i_{\min}} F(\alpha_k))(\frac{z_{\min}}{L_1^F} - \sum_{i=u}^{v} |\alpha_k^{(i)}|)$
        else
            $n = \arg\max\{i, i \in [u, v-1], \frac{dh}{dz}(z_i^+), \frac{dh}{dz}(z_{i+1}^-) \geqslant 0\}$
            if $n = v - 1$ then
                $\beta^{(v)} = -\mathrm{sign}(\alpha_v) \frac{1}{L_1^F}(\lambda + \frac{\alpha_v}{|\alpha_v|} \nabla_v F(\alpha_k))$
            else
                For $i \in [n+1, v]$, $\beta^{(i)} = -\alpha_k^{(i)}$
                If $\frac{dh}{dz}(z_n^+) \leqslant 0$, then $\beta^{(n)} = -\mathrm{sign}(\alpha_k^{(n)})(\frac{1}{L_1^F}\left(\lambda + \frac{\alpha_k^{(n)}}{|\alpha_k^{(n)}|} \nabla_n F(\alpha_k)\right) - \sum_{i=n+1}^{v} |\alpha_k^{(i)}|)$
            end if
        end if
    end if
    $\alpha_{k+1} = \alpha_k + \beta$
end for
```

possibly a new one generated by the LMO. Therefore, Algorithm 1 belongs to the family of boosting algorithms. We refer to it as the regularized matching pursuit (RMP).

Compared to the boosting approach of Zhang et al. (2012) for metric-norm regularization or to the generalized conditional gradient (Bach, 2015; Sun and Bach, 2022), active atoms are not modified uniformly since only some of them may be reduced to zero. The SOTOPO method of Song et al. (2017) minimizes the same upper bound with respect to the ℓ_1-norm (16). 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 (15) 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 4. This suggests a connection between regularized matching pursuit and proximal coordinate descent, as proven by Locatello et al. (2018) for gradient descent and steepest coordinate descent.

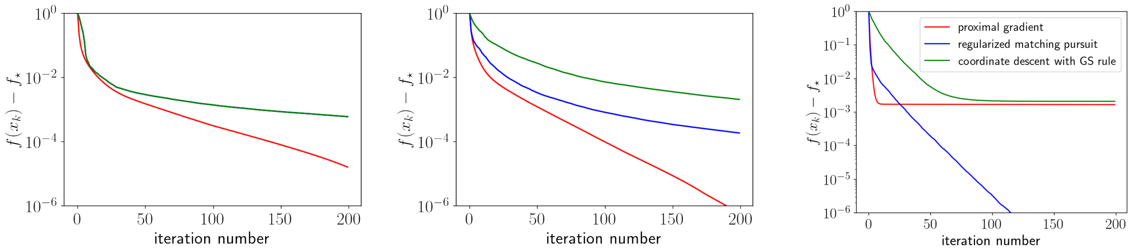

Figure 4: Convergence in function value of the proximal gradient descent, coordinate descent with Gauss-Soutwhell rule and with $L = L_1^F$ (instead of $L = L_2^F$) and of the regularized matching pursuit, for synthetic quadratics (see Section 2.4) with $n = 50$, $s = 8$, $\lambda = 0.001$, $\sigma = 0.5$ 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 match exactly in these examples.

In Figure 4, the RMP appears to converge linearly in the underparametrized regime, and sublinearly in the overparametrized regime. We compute some convergence guarantees in the next section.

3.2 Convergence Guarantee

We now establish convergence guarantees for the RMP, both for strongly convex and non-strongly convex functions. We consider a more general composite minimization problem,

$$\min_{\alpha \in \mathbb{R}^d} \left\{ G(\alpha) \triangleq F(\alpha) + H(\alpha) \right\}, \tag{18}$$

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). We evaluate the convergence guarantees of a generalized version of the RMP (the GRMP), that is not always a boosting method,

$$\alpha_{k+1} = \underset{\alpha \in \mathbb{R}^d}{\arg\min} \ \langle \nabla F(\alpha_k), \alpha - \alpha_k \rangle + \frac{L_1^F}{2} \|\alpha - \alpha_k\|_1^2 + H(\alpha). \tag{19}$$

As we will see, our proofs are similar to those for randomized coordinate descent (Richtárik and Takáč, 2014, Theorem 5, 7).

3.2.1 Strongly convex functions

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 (Nutini et al., 2018b), 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 17 *(Nutini, 2018, Appendix A.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* (19) *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 is taken from Nutini (2018, Appendix A.8.) and consists in an optimization step over all trajectories. The argument is inspired from randomized coordinate descent (Richtárik and Takáč, 2014)). ■

The RMP is a special case of method (19), and verifies the convergence guarantee of Proposition 17. 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 adapted 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 8 and 9 still apply. In the overparametrized regime however, Figure 4 suggests that the method does not converge linearly (since it is stuck at an accuracy of about around 10^{-5}).

3.2.2 Smooth convex functions

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 18, we prove sublinear convergence for the GRMP. To our knowledge, there is no such result for sublinear convergence for SOTOPO (Song et al., 2017) or for coordinate descent with the Gauss-Southwell rule, which is very close to the GRMP.

Proposition 18 *Let (α_k) be generated by the generalized regularized matching pursuit* (19), *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 a proof for sublinear convergence of randomized proximal coordinate descent established by Richtarik and Takac (Richtárik and Takáč, 2014, Theorem 5). Let $\alpha_{k+1} \in \mathbb{R}^d$ be a minimizer of the smooth upper bound:

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

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 $G(\alpha_k) - G_\star - L_1^F\|\alpha_k - \alpha_\star\|_1^2$:

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

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 as in the proof for sublinear convergence of steepest coordinate descent, detailed in Appendix D, we arrive to a convergence guarantee $G(\alpha_k) - G_\star \leqslant \frac{2L_1^F \mathcal{R}_{\alpha_0}^2}{k+1}$. ■

Proposition 18 provides a sublinear convergence guarantee for the GRMP 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 see from the numerical experiments in the next section that the RMP does converge linearly in certain regimes in the case of the ℓ_1-regularized model.

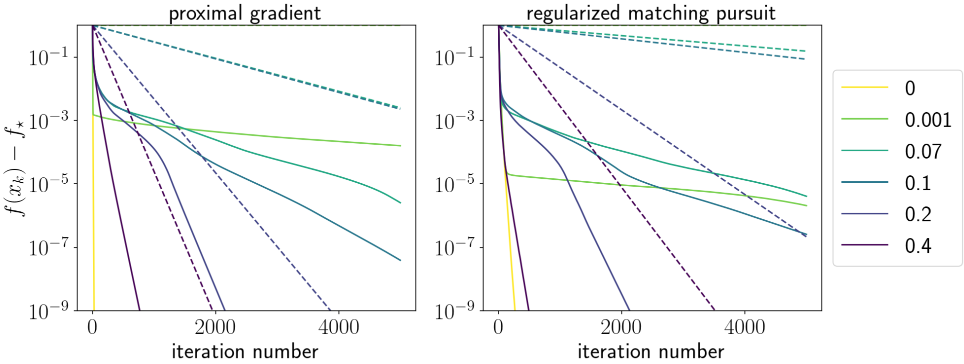

Figure 5: Convergence in function values for the proximal gradient on the left and the regularized matching pursuit on the right for $n = 50$, $d = 500$ and a sparsity $s = 8$ and for several penalty λ. Convergence is compared in dashed lines to local convergence guarantee, taken on the support S on the last iterates and the SDP relaxation from Proposition 8.

3.3 A Transition Phase Depending on λ: Experimental rResults

The RMP algorithm benefits from convergence guarantees similar to those for the proximal gradient: these methods converge linearly under strong convexity assumptions (underparametrized regime) but have sublinear guarantees for smooth 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 (Raskutti et al., 2010), and for a well-chosen parameter λ (Agarwal et al., 2010, Theorem 2). In this section, our experiments reveals a transition phenomenon driven by λ on a LASSO problem $F(\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2 + \lambda\|x\|_1$, where P are synthetic Gaussian data in the overparametrized setting as in Section 2.4.

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 (2020, Theorem 1) prove the proximal gradient 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. (2018a) 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 5, based on the sparsity of the solution to the optimization problem. In Figure 5, local strong convexity parameters are given by the estimated of Corollary 6 and Proposition 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 6 by plotting the ϵ-curve (see Section 2.4) 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' values of λ for which the support is quickly identified, the convergence is linear

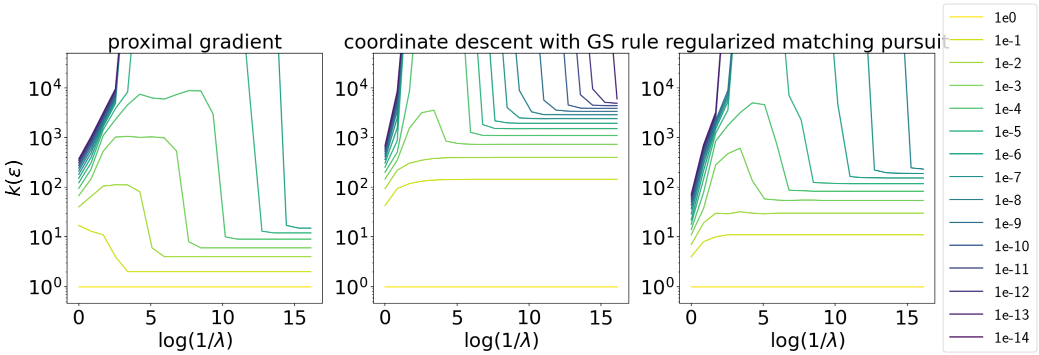

Figure 6: ϵ-curve of the proximal gradient, coordinate descent with the GS rule and regularized matching pursuit for a LASSO problem with $d = 500$, $n = 50$, a sparsity level $s = 8$, $\sigma = 0.5$, after $k = 10000$ iterations for several values of λ.

too. In the intermediary 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 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 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. 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 6): 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 6.

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 1 was derived from the ℓ_1-geometry. In Section 2.5, for non-regularized models, coordinate descent with GS-rule was interpreted as a matching pursuit algorithm in both the underparametrized and overparametrized regime. In what follow, we see that the regularized matching pursuit as developed above does not benefit from this formulation in the overparametrized regime. Instead, we 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) and (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 2.5, 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 in $\mathbb{R}^n$, of lower dimension.

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 $x_0\in\mathbb{R}^n$:

$$x_{k+1}=\underset{x\in\mathbb{R}^n}{\arg\min}\langle\nabla f(x_k),x-x_k\rangle+\frac{L^f_{\gamma_{\mathcal{P}}}}{2}\gamma_{\mathcal{P}}(x-x_k)^2+\lambda\gamma_{\mathcal{P}}(x). \tag{20}$$

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. Let us reformulate the minimization Problem (20) 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 (1). However, in the overparametrized regime, $P\alpha=P\nu$ does not imply $\alpha=\nu$ in general. This method does not belong to boosting algorithms due to the evaluation of the gauge function in x and in $x-x_k$ in (20). In addition, this minimization problem admits neither a simple closed-form solution in general nor a solution based on a variational formulation of the ℓ_1-norm (as we did for regularized matching pursuit). While not directly computable in general, the minimization step (20) converges linearly to the optimum, as proven below in Proposition 19.

Proposition 19 *Let f be $L^f_{\gamma_{\mathcal{P}}}$-smooth and $\mu^f_{\gamma_{\mathcal{P}}}$-strongly convex with respect to the norm $\gamma_{\mathcal{P}}(\cdot)$. The ultimate method* (20) (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 17, replacing the function F by f and the norm $\|\cdot\|_1$ by $\gamma_{\mathcal{P}}(\cdot)$. ■

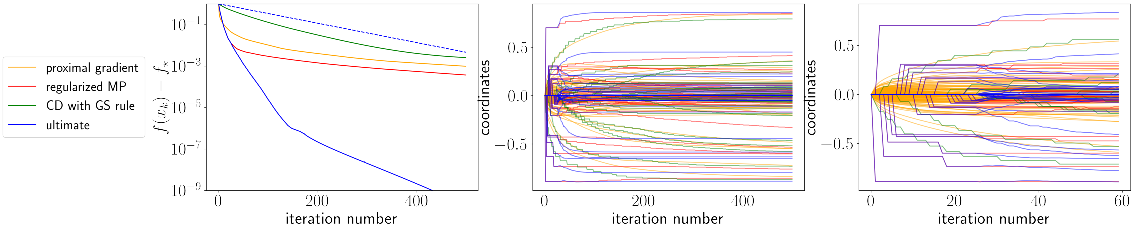

Figure 7: Convergence in function value and coordinates as a function of the iteration number for the proximal gradient descent, the proximal coordinate descent with GS rule, the regularized matching pursuit and the ultimate method on a LASSO problem, with $d = 500$, $n = 50$, $s = 8$, $\lambda = 0.2$. The approximate guarantee in provided in dashed lines.

Proposition 19 provides a linear convergence guarantee for the ultimate algorithm. We recover the convergence guarantee of coordinate descent with GS rule in the non regularized model (12).

Remark 20 *In Figure 7, we compare three methods having comparable one-step complexity $O(d)$: proximal gradient descent, proximal coordinate descent with GS rule and regularized matching pursuit. Yet, the one-step complexity of the ultimate method is directly connected to the inner loop complexity and accuracy.*

In Figure 7, we solve the optimization step for the ultimate method for the gauge with the solver MOSEK (ApS, 2022) 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 (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 8 apply here. In the Appendix F, 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.

4. 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 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 does not converge linearly in this regime.

We finally formulate an ultimate method adapted to overparametrized geometries. Yet, this method lacks a closed-form formulation. In numerical experiments, we approximate it using 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 upper bounds (resp. high probability bounds) for convergence guarantees using SDP relaxations (resp. under statistical assumptions on the data). 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, we believe it could be of interest to extend this principled approach to accelerated matching pursuit algorithms (and thus, to accelerated coordinate descent algorithms). Some accelerated techniques have already been developed relying on randomly selected coordinates, such as those of Nesterov and Stich (2017) for nonregularized minimization and Fercoq and Richtárik (2015) or Locatello et al. (2018, Section 3) for composite minimization problems, and the techniques of Bertrand and Massias (2021) of Lin et al. (2015). Another interesting line of research could be to understand the connections between the observed transition phase 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). We thank the reviewers for their careful review and constructive comments.

Codes

All codes for numerical results are provided at `https://github.com/CMoucer/Geometry_Dependent_Matching_Pursuit`.

References

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, 2010.

MOSEK ApS. *The MOSEK Optimization Toolbox for MATLAB Manual. Version 10.0.*, 2022. URL `http://docs.mosek.com/9.0/toolbox/index.html`.

Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

Francis Bach, Rodolph Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1): 1–106, 2012.

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer, 2010.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2017.

Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 32(10):15849–15854, 2019.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Accelerated gossip in networks of given dimension using Jacobi polynomial iterations. *SIAM Journal on Mathematics of Data Science*, 2(1):24–47, 2020.

Quentin Bertrand and Mathurin Massias. Anderson acceleration of coordinate descent. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1288–1296, 2021.

Quentin Bertrand, Quentin Klopfenstein, Pierre-Antoine Bannier, Gauthier Gidel, and Mathurin Massias. Beyond l1: Faster and better sparse models with SKGLM. In *Advances in Neural Information Processing Systems*, 2022.

Jérôme Bolte, Aris Daniilinis, Olivier Ley, and Laurent Mazet. Characterization of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

Karl-Heinz Borgwardt. The average number of pivot steps required by the simplex-method is polynomial. *Zeitschrift für Operations Research*, 26:157–177, 1987.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Stéphane Boucheron, Gabòr Lugosi, and Pascal Massart. *Concentration Inequalities. A Nonasymptotic Theory of Independence*. Oxford University Press. 2013.

Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Patrick L. Combettes and Valérie R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

Alexandre d'Aspremont, Cristóbal Guzmán, and Martin Jaggi. Optimal affine-invariant smooth minimization algorithms. *SIAM Journal on Optimization*, 28(3):2384–2405, 2018.

Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. *Acceleration Methods*, volume 5 of *Foundations and Trends in Optimization*. 2021.

Jelena Diakonikolas and Lorenzo Orecchia. Alternating randomized block coordinate descent. In *Proceedings of the International Conference on Machine Learning*, 2018.

Miroslav Dudik, Zaid Harchaoui, and Jerome Malick. Lifted coordinate descent for learning with trace-norm regularization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2012.

Huang Fang, Zhenan Fan, Yifang Sun, and Michael 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.

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

Yoav Freund and Robert E. Schapire. A short introduction to boosting. *Japanese Society For Artifical Intelligence*, 14:771–780, 1999.

Michael P. Friedlander, Ives Macêdo, and Ting Kei Pong. Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014.

Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42:1115–1145, 1995.

T.R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. 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*, 286:531–537, 1999.

Charles Guille-Escuret, Baptiste Goujaud, Manuela Girotti, and Ioannis Mitliagkas. A study of condition numbers for first-order optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2021.

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal Research of the National Bureau of Standards*, 49:263–264, 1957.

Franck Iutzeler and Jérôme Malick. Nonsmoothness in machine learning: specific structure, proximal identification, and applications. *Set-Valued and Variational Analysis*, 28:661–678, 2020.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning*, pages 427–435, 2013.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, 2016.

Sai Praneeth Karimireddy, Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Efficient greedy coordinate descent for composite problems. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2019.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems*, 2015.

Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015.

Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A unified optimization view on generalized matching pursuit and Frank-Wolfe. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2017.

Francesco Locatello, Anant Raj, Sai Praneeth Karimireddy, Gunnar Raetsch, Bernhard Schölkopf, Sebastian Stich, and Martin Jaggi. On matching pursuit and coordinate descent. In *Proceedings of the International Conference on Machine Learning*, 2018.

Stéphane .G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

Vladimir A. Marchenko and Loenid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 1(4):457 – 483, 1967.

Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. From safe screening rules to working sets for faster lasso-type solvers. *NIPS Workshop on Optimization for Machine Learning*, 2017.

Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. Celer: a fast solver for the lasso with dual extrapolation. In *Proceedings of the International Conference on Machine Learning*, 2018.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(75):667–766, 2022.

Ion Necoara, Yurii Nesterov, and François Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

Yuri Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Yurii Nesterov and Sebastian U. Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1): 110–123, 2017.

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*, 13:645–655, 2018a.

Julie Nutini. *Greed is Good: Greedy Optimization Methods for Large-Scale Structured Problems.* PhD thesis, University of British Columbia, 2018.

Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *Proceedings of the International Conference on Machine Learning*, 2018b.

Courtney Paquette, Bart van Merriënboer, Elliot Courtney, and Fabian Pegregosa. Halting time is predictable for large models: A universality property and average-case analysis. *Foundations of Computational Mathematics*, 23:597–673, 2023.

Neal Parikh and Stephen P. Boyd. Proximal algorithms. *Foundations and Trends Optimization*, 1:127–239, 2013.

Fabian Pedregosa and Damien Scieur. Acceleration through spectral density estimation. In *Proceedings of the International Conference on Machine Learning*, 2020.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144: 1–38, 2014.

Damien Scieur and Fabian Pedregosa. Universal asymptotic optimality of Polyak momentum. In *Proceedings of the International Conference on Machine Learning*, 2020.

Stephen A. Billings Sheng Chen and Wan Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5): 1873–1896, 1989.

Chaobing Song, Shaobo Cui, Yong Jiang, and Shu-Tao Xia. Accelerated stochastic greedy coordinate descent by soft thresholding projection onto simplex. In *Advances in Neural Information Processing Systems*, 2017.

Daniel Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, 2001.

Yifan Sun and Francis Bach. Safe screening for the generalized conditional gradient method. *Open Journal of Mathematical Optimization*, 3:1–35, 2022.

Adrien B. Taylor, Julien Hendrickx, and François and Glineur. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *Journal of Optimization Theory and Applications*, 178(2):455–476, 2018.

Ambuj Tewari, Pradeep Ravikumar, and Inderjit Dhillon. Greedy algorithms for structurally constrained high dimensional problems. In *Advances in Neural Information Processing Systems*, 2011.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, B(117):387–423, 2009.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. 2018.

Tong Zhang. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Transactions on Information Theory*, 57(9):6215–6221, 2011a.

Tong Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Transactions on Information Theory*, 57(7):4689–4708, 2011b.

Xinhua Zhang, Dale Schuurmans, and Yao-liang Yu. Accelerated training for matrix-norm regularization: A boosting approach. In *Advances in Neural Information Processing Systems*, 2012.

Appendix

In this document, we provide proofs for the main theorems of the paper as well as additional experiments offering a comprehensive overview of the main paper's results. Table 1 summarizes the main contributions and results of the Appendix.

Table 1: Content of the appendices

Appendix A	Mathematical tools appearing in the proofs: η-tricks and computation of the basis of a kernel.
Appendix B	Proof for linear convergence of matching pursuit (Proposition 7) and gradient descent (Proposition 4).
Appendix C	Estimates for the convergence of steepest coordinate descent for least-squares: SDP relaxations and high-probability bounds for μ_1^F, $\mu_{1,L}^F$ and L_1^F, with numerical comparisons to μ_2^F and $\mu_{2,L}^F$.
Appendix D	Matching pursuit in the gauge geometry: properties of the gauge and sublinear convergence guarantee.
Appendix E	Formulation of steepest coordinate descent as a 'nearly' matching pursuit algorithm.
Appendix F	Efficiently computing the ultimate method: an inner loop strategy.

Appendix A. Mathematical Tools

A.1 The η-tricks: Reweighted Least-Squares Formulations

Due to non-smoothness, the ℓ_1-norm is often seen as difficult to optimize. A common way to simplify regularization term containing an ℓ_1-term, such as the LASSO, consists in formulating it as a reweighted least-square problems. We refer to the work of Bach et al. (2012) There are three common formulations:

- the variational formulation (Bach et al., 2012, Section 5)

$$\|x\|_1 = \min_{\eta \in \mathbb{R}^n, \eta \geqslant 0} \frac{1}{2} \sum_{i=1}^n \frac{x_i^2}{\eta_i} + \frac{1}{2} \sum_{i=1}^n \eta_i,$$

- the variational constrained formulation (Bach et al., 2012, Section 1)

$$\|x\|_1^2 = \min_{\eta \in \Delta_n} \sum_{i=1}^n \frac{x_i^2}{\eta_i},$$

  where $\Delta_n = \{\eta \in \mathbb{R}^n, \eta \geqslant 0, \sum_{i=1}^n \eta_i = 1\}$ is the simplex,

- the maximization problem,
$$\|x\|_1 = \max_{\|s\|_\infty \leqslant 1} \langle s, x\rangle.$$

Finally, we notice a useful variational trick, in dimension 1. For $x \in \mathbb{R}^n$,

$$\frac{L}{2}\|x\|_1^2 = \max_{z \geqslant 0} \frac{-z^2}{2L} + z\|x\|_1, \tag{21}$$

with the optimum $z_\star = L\|x\|_1$.

A.2 Computing the Basis of a Kernel

To compute the basis of $\mathrm{Ker}(P)$, we perform a QR decomposition on P such that $Q_1^\top Q_1 = I_n$, $Q_2^\top Q_2 = I_{d-n}$ and $Q_1^\top Q_2 = 0$: $P^\top = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix}$, where Q_2 is a basis for the nullspace of P. Indeed, let $z \in \mathbb{R}^{d-n}$, and $AQ_2 z = R^\top Q_1^\top Q_2 z = 0$.

Appendix B. Proof for Propositions 4, 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 4, 7 for the ℓ_2, ℓ_1 norm respectively.

Let F be L^F-smooth with respect to a norm $\|\cdot\|$, (possibly) μ^F-strongly convex with respect to $\|\cdot\|$ and verify the Łojasiewicz inequality with parameter μ_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 μ_L^F, but is not strongly convex ($\mu^F = 0$).

Appendix C. Estimates for the Convergence of Steepest Coordinate Descent for Least-Squares

In Proposition 7, a sequence (α_k) generated by steepest coordinate descent for a linear regression problem has a linear convergence rate in function values,

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

We assume that F is a quadratic, that is $F(\alpha) = \frac{1}{2n}\|P\alpha - y\|_2^2$. We first derive SDP relaxations for the optimization problems characterizing μ_1^F, $\mu_{1,L}^F$ and L_1^F, and numerically compare these estimates to μ_2^F and $\mu_{2,L}^F$. Then, we compute inequalities connecting $\mu_{1,L}^F$ and μ_1^F to L_1^F. Thanks to these inequalities, 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.

C.1 SDP Relaxations for μ_1^F and $\mu_{1,L}^F$

We look for exact lower bounds for μ_1^F and $\mu_{1,L}^F$. Both in the overparametrized and underparametrized regime, we are going to reformulate the optimization problems defining $\mu_1^F, \mu_{1,L}^F$ into SDPs, and relax some rank constraints. Then, we compare these estimates to μ_2^F and $\mu_{2,L}^F$ in numerical experiments.

C.1.1 Proof for Proposition 8

SDP relaxation for μ_1^F in the underparametrized regime. Recall from Lemma 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 $P^\top P$ is invertible, we proceed to a change of variable in $P^\top P$:

$$\begin{aligned}\frac{1}{\sqrt{n\mu_1^F}} &= \max_{\alpha,\|P^\top P\alpha\|_\infty\leqslant 1}\ \max_{\nu,\|Pz\|_2\leqslant 1}(P^\top Px)^\top z,\\ &= \max_{\alpha,\|P^\top P\alpha\|_\infty\leqslant 1}\ \max_{\nu,\|P\nu\|_2\leqslant 1}(P\alpha)^\top(P\nu),\\ &= \max_{\alpha,\|P^\top P\alpha\|_\infty\leqslant 1}\|P\alpha\|_2,\\ &= \max_{\alpha}\|P(P^\top P)^{-1}\alpha\|_2,\ \text{s.t. } \|\alpha\|_\infty \leqslant 1.\end{aligned}$$

A first attempt to compute an exact solution to this problem is to consider $\alpha \in \{-1,1\}^d$ and compute all possible values for the $\|\cdot\|_\infty$-norm. However this computation would require 2^d configurations. Instead, we compute an SDP relaxation of $\frac{1}{\sqrt{\mu_1^F}}$ in Lemma 21.

Lemma 21 *Let the regime be underparametrized ($n \geqslant d$). μ_1^F has a $\frac{\pi}{2}$-approximation:*

$$\tilde{\mu}_1^F \leqslant \mu_1^F \leqslant \frac{\pi}{2}\tilde{\mu}_1^F.$$

where $\tilde{\mu}_1^F$ has an SDP-formulation $\frac{1}{\tilde{\mu}_1^F} = n\max_{X\succcurlyeq 0,\mathrm{diag}(X)\leqslant 1}\mathrm{Tr}(CX)$, with $C = (P^\top P)^{-1}$.

Proof Let us reformulate an SDP relaxation to this problem. The problem $OPT = \frac{1}{\mu_1^F} = n\max_{\nu,\|\nu\|_\infty^2\leqslant 1}\|P(P^\top P)^{-1}\nu\|_2^2 = n\max_{\nu,\nu_i^2\leqslant 1}\nu^\top(P^\top P)^{-1}\nu$ can be relaxed into a SDP,

$$SDP = \frac{1}{\tilde{\mu}_1^F} = n\max_{Z\succcurlyeq 0}\mathrm{Tr}(CZ), \quad \text{s.t. } \mathrm{diag}(Z) \leqslant 1, \tag{22}$$

where $C = (P^\top P)^{-1} \succ 0$. This is exactly the Max-Cut SDP relaxation, for which Goemans and Williamson (1995) proved the approximation $\frac{2}{\pi}SDP \leqslant OPT \leqslant SDP$. ■

SDP relaxation for $\mu^F_{1,L}$ in the overparametrized regime. For $\mu^F_{1,L}$, the maximization problem does not correspond to Max-Cut. So it is possible to derive an SDP lower bound, but no SDP-approximation. Recall the formulation of $\mu^F_{1,L}$ as an optimization problem $\mu^F_{1,L} = \frac{1}{n}\inf_{\nu\in\mathbb{R}^d} \|P^\top P\nu\|^2_\infty$, s.t. $\|P\nu\|^2_2 = 1$. In Lemma 22, we derive a lower bound for $\mu^F_{1,L}$ formulated as a SDP.

Lemma 22 *Let $PP^\top$ be invertible, then,*

$$\mu^F_{1,L} \geqslant \frac{1}{n}\inf_{Z\succcurlyeq 0} \|P^\top PZP^\top P\|_\infty, \quad \text{s.t. } \mathrm{Tr}(P^\top PZ) = 1,$$

Proof We introduce $Z = \nu\nu^\top \in \mathbb{R}^{d\times d}$, where $\mathrm{rank}(Z) = 1$, and reformulate the problem into $\mu^F_{1,L} = \frac{1}{n}\inf_{Z\succcurlyeq 0,\mathrm{rank}(Z)=1,\mathrm{Tr}(P^\top PZ)=1} \|P^\top PZP^\top P\|_\infty$, which can be relaxed as an SDP $\mu^F_{1,L} \geqslant \frac{1}{n}\inf_{Z\succcurlyeq 0,\mathrm{Tr}(P^\top PZ)=1} \|P^\top PZP^\top P\|_\infty$. ■

Comparison of $\tilde\mu^F_1$ and $\tilde\mu^F_{1,L}$. First, let us notice that norm $\|\cdot\|_\infty$ corresponds to the infinite norm on the diagonal of the matrices. Consequently, the constraint $\mathrm{diag}(X) \leqslant 1$ is equivalent with $\|X\|_\infty \leqslant 1$. In addition, in the underparametrized regime, a change of variable $X \to P^\top PZP^\top P$ leads to the reformulation of (22) as,

$$\frac{1}{\tilde\mu^F_1} = n\max_{X\succcurlyeq 0}\ \mathrm{Tr}(P^\top PX), \quad \text{s.t. } \|P^\top PXP^\top P\|_\infty \leqslant 1.$$

Thus, we conclude with the fact that $(P^\top P)$ is not invertible in the overparametrized regime, and thus that $\frac{1}{\tilde\mu^F_1} \geqslant \frac{1}{\tilde\mu^F_{1,L}}$.

C.1.2 Numerical comparison

In this section, we compare the estimate $\tilde\mu^F_1$ (resp. $\tilde\mu^F_{1,L}$) for μ^F_1 (resp. $\mu^F_{1,L}$) to μ^F_2 and μ^F_2/d (resp. $\mu^F_{2,L}$ and $\mu^F_{2,L}/d$). We consider Gaussian matrices $X \in \mathbb{R}^{n,d}$ such that $X_i \sim \mathcal{N}(0,\Sigma)$ are i.i.d., given four different diagonal variances: a uniform variance $\Sigma = I_d$, a non-uniform variance $\Sigma = \mathrm{Diag}(1,\ldots,1/d)$, a variance with only one small value $\Sigma = \mathrm{Diag}(1,\cdots,1,\frac{1}{100})$, a variance with only one large value $\Sigma = \mathrm{Diag}(1,\ldots,1,100)$. These variances are inspired from the work of Nutini et al. (2018b, Section 4.1), who computed explicitly μ^F_1 for separable quadratics.

Underparametrized regime: comparison for μ^F_1. In this regime, μ^F_1 has a SDP approximation given by $\tilde\mu^F_1$, as proven in Appendix C.1. By norm equivalence, $\frac{\mu^F_2}{d} \leqslant \mu^F_1 \leqslant \mu^F_2$, as proven by Nutini et al. (2018b, Appendix 4). The value of the SDP approximation for μ^F_1 in Figure 8 is very close to its lower bound $\frac{\mu^F_2}{d}$, except for the variance where only one diagonal element of the variance is very large.

Overparametrized regime: comparison for $\mu^F_{1,L}$. In this regime, $\mu^F_{1,L}$ is lower bounded by $\tilde\mu^F_{1,L}$ defined by a SDP (see Appendix C.1. By norm equivalence, we have that $\frac{\mu^F_{2,L}}{d} \leqslant \mu^F_{1,L} \leqslant \mu^F_{2,L}$. The SDP relaxation provides a lower bound for $\mu^F_{1,L}$, that is always close to its lower bound $\frac{\mu^F_{2,L}}{d}$ as observed in Figure 8. We observe similar results for a random features model generated from the Leukemia dataset (libsvm) in Figure 9.

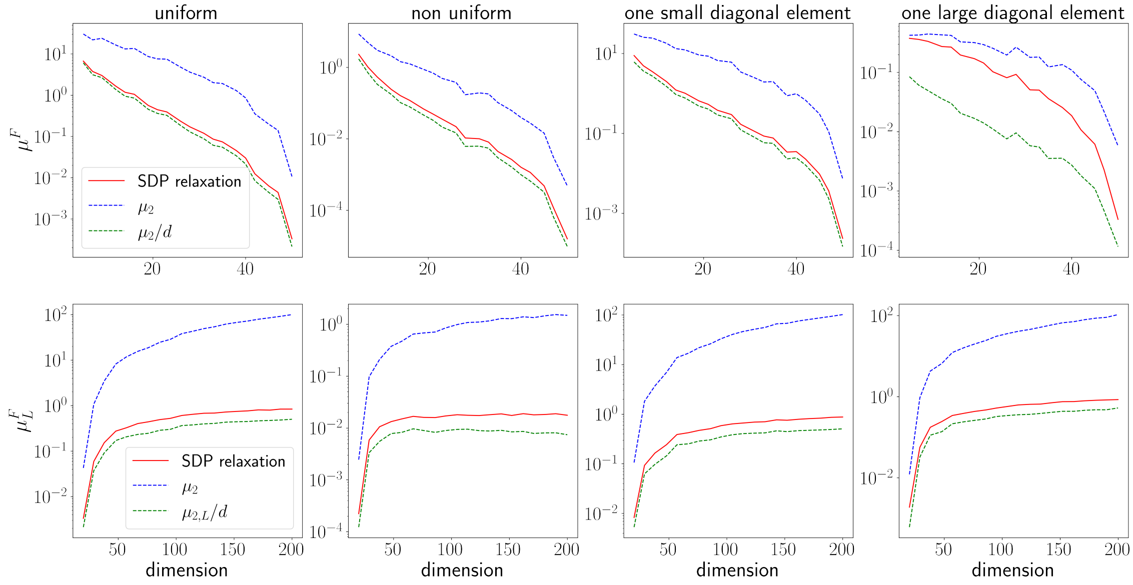

Figure 8: Comparison of μ_1^F on the top (resp. $\mu_{1,L}^F$ on the bottom) with μ_2^F and μ_2^F/d (resp. $\mu_{2,L}^F$ and $\mu_{2,L}^F$) for $n = 50$ (resp. $n = 20$) averaged on 5 trials, for Gaussian random data with a variance equal from the left to the right to $\Sigma = I_d$, $\Sigma = \text{diag}(1, 1/2, \ldots, 1/d)$, $\Sigma = \text{diag}(1, \ldots, 1, 1/100)$ and $\text{diag}(1, \ldots, 1, 100)$.

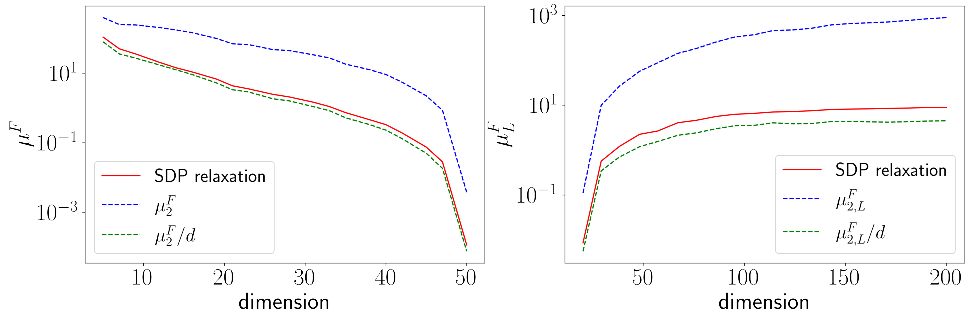

Figure 9: Comparison of the SDP relaxation for μ_1^F on the left (resp. $\mu_{1,L}^F$ on the right) with μ_2^F and μ_2^F/d (resp. $\mu_{2,L}^F$ and $\mu_{2,L}^F/d$) for a random feature model generated from the Leukemia dataset with a number of samples $n = 50$ (resp. $n = 20$).

C.2 High-Probability Bounds for μ_1^F, $\mu_{1,L}^F$ and L_1^F: Proof for Proposition 9

In this proof, we look for concentration inequalities on μ_1^F, $\mu_{1,L}^F$ and L_1^F. To this end, we first establish deterministic inequalities connecting μ_1^F, $\mu_{1,L}^F$ and L_1^F. In a second part, we assume the data P are generated randomly, such that each row is subgaussian. Under some mild assumptions, we derive concentration inequalities on these parameters.

C.2.1 Deterministic inequalities connecting μ_1^F, $\mu_{1,L}^F$ and L_1^F

In this proof, we look for inequalities connecting L_1^F, μ_1^F and $\mu_{1,L}^F$ to minimal and maximal eigenvalues of $P^\top P$ and $PP^\top$. We first propose to establish deterministic inequalities characterizing these parameters.

We consider the special case of least-squares, for which $F(\alpha) = \frac{1}{n}\|P\alpha - y\|_2^2$ with $P \in \mathbb{R}^{n\times d}$. We have seen that:

- $L_1^F = \frac{1}{n}\max_{i=1,\dots,d} \|P_{:,i}\|_2^2$,
- $\mu_1^F = \frac{1}{n}\inf_{\eta\in\mathbb{R}^d} \|P\eta\|_2^2$ such that $\|\eta\|_1^2 \geqslant 1$,
- $\mu_{1,L}^F = \frac{1}{n}\sup_{\eta\in\mathbb{R}^d} \|P^\top P\eta\|_\infty^2$such that $\|P\eta\|_2^2 = 1$.

In Lemma 23, we establish deterministic lower and upper bounds for both μ_1^F and $\mu_{1,L}^F$.

Lemma 23 *Let $F(\alpha) = \frac{1}{n}\|P\alpha - y\|_2^2$ with $P \in \mathbb{R}^{n\times d}$. Denote μ_1^F (resp. $\mu_{1,L}^F$) the strongly convex (resp. Łojasiewicz) parameter of F with respect to the ℓ_1-norm. Then, μ_1^F verifies,*

$$\begin{aligned} \frac{1}{n}\frac{\lambda_{\min}(P^\top P)}{d} &\leqslant \mu_1^F &\leqslant \frac{1}{n}\frac{\mathbf{1}_d^\top P^\top P\mathbf{1}_d}{d^2}, \\ \frac{1}{n}\frac{\lambda_{\min}(PP^\top)}{d} &\leqslant \mu_{1,L}^F &\leqslant \frac{L_1^F}{n}. \end{aligned}$$

Proof From result of Nutini et al.(Nutini et al., 2018b, Appendix 4), we have by the norm equivalence $\frac{\mu_2^F}{d} \leqslant \mu_1^F$. In the special case of least-squares, $\mu_2^F = \frac{\lambda_{\min}(P^\top P)}{n}$. In addition, we have that $\mu_1^F = \frac{1}{n}\inf_{\eta\in\mathbb{R}^d, \|\eta\|_1^2=1} \|P\eta\|_2^2 \leqslant \frac{1}{n}\frac{\mathbf{1}_d^\top P^\top P\mathbf{1}_d}{d^2}$ by taking $\eta = \frac{\mathbf{1}_d}{d}$.

Let us reformulate $\mu_{1,L}^F$. Using the trick $\forall \nu \in \mathbb{R}^d$, $\|\nu\|_1^2 = \inf_{\gamma\in\Delta_d}\sum_{i=1}^d \frac{\nu_i^2}{\gamma_i}$,

$$\begin{aligned} \mu_{1,L}^F &= \frac{1}{n}\inf_{\nu\in\mathbb{R}^n} \|P^\top\nu\|_\infty^2, \quad \text{s.t. } \|\nu\|_2^2 = 1, \\ &= \frac{1}{n}\inf_{\nu\in\mathbb{R}^n}\max_{\eta\in\Delta_d} \nu^\top P\mathrm{Diag}(\eta)P^\top\nu, \quad \text{s.t. } \|\nu\|_2^2 = 1, \\ &= \frac{1}{n}\max_{\eta\in\Delta_d}\inf_{\nu\in\mathbb{R}^n} \nu^\top P\mathrm{Diag}(\eta)P^\top\nu, \quad \text{s.t. } \|\nu\|_2^2 = 1, \\ &= \frac{1}{n}\max_{\eta\in\Delta_d}\lambda_{\min}\left(P\mathrm{Diag}(\eta)P^\top\right). \end{aligned}$$

For $\eta = \frac{1}{d}\mathbf{1}_d$, we have that $\mu_{1,L}^F \geqslant \frac{1}{n}\frac{\lambda_{\min}(PP^\top)}{d}$. In addition, we can reformulate $\mu_1^{L,F}$ as follows: $\mu_1^{L,F} = \frac{1}{n}\max_{\eta\in\Delta_d}\lambda_{\min}\left(P\mathrm{Diag}(\eta)P^\top\right) = \frac{1}{n}\max_{\eta\in\mathbb{R}^d}\sup_{M\succcurlyeq 0,\mathrm{Tr}(M)=1}\mathrm{Tr}(P\mathrm{Diag}(\eta)P^\top M) = \frac{1}{n}\sup_{M\succcurlyeq 0,\mathrm{Tr}(M)=1}\|P^\top MP\|_\infty \leqslant \frac{1}{n^2}\|P^\top P\|_\infty$, for $M = \frac{I_d}{n} = \frac{L_1^F}{n}$. ■

As expected by the norm equivalence (and already highlighted by Nutini et al. (Nutini et al., 2018b, Appendix 4)), strong convexity parameters verify $\mu_1^F \geqslant \frac{\mu_2^F}{d}$. Lemma 23 states a similar comparison for Łojasiewicz parameters with $\mu_{1,L}^F \geqslant \frac{\mu_{2,L}^F}{d}$. We recover that

$\mu_1^F \geqslant 0$ (resp. $\mu_{1,L}^F \geqslant 0$) in the overparametrized (resp. underparametrized) regime. Under subgaussian data, we now build high probability bounds for $\mu_{1,L}^F$, μ_1^F and L_1^F based on these lower and upper bounds.

C.2.2 Generalities on subgaussian data

Before deriving concentration bounds, we detail the subgaussian assumptions on P.

Definition 24 *(Vershynin, 2018, Definition 2.5.2, Proposition 2.5.2) Let X be a subgaussian random with $\tau > 0$, and $\mathbb{E}[X] = 0$. Then, there exists absolute constants $c_1, c_2 > 0$ such that,*

$$\forall t \in \mathbb{R},\ \mathbb{E}[e^{tX}] \leqslant e^{c_1 \|X\|_{\Psi_2}^2 t^2},$$
$$\forall t \in \mathbb{R},\ \mathbb{P}(|X| \geqslant t) \leqslant 2\exp(-\frac{c_2}{\|X\|_{\Psi_2}^2} t^2),$$

In addition, $\|X\|_{\Psi_2} = \inf\{t > 0, \mathbb{E}[e^{X^2/t^2}] \leqslant 2\}$ is the subgaussian norm of X.

Usually, a subgaussian variable X with $\mathbb{E}[X] = 0$, $\mathbb{E}[x^2] = \sigma^2$ is defined with a parameter $\tau > 0$ such that for all $t \in \mathbb{R}$, $\mathbb{E}[\exp(tX)] \leqslant \exp(\frac{t^2\tau^2}{2})$. Then, $\sigma \leqslant \tau = \sqrt{2c_1}\|X\|_{\Psi_2}$. As for gaussian data, Definition 24 can be extended to vectors.

Definition 25 *(Vershynin, 2018, Definition 3.4.1) A random vector $X \in \mathbb{R}^d$ is called subgaussian if the one-dimensional marginals $\langle X, x\rangle$ are subgaussian random variables for all $x \in \mathbb{R}^d$. The subgaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{x \in \mathbb{S}^{d-1}} \|\langle X, x\rangle\|_{\Psi_2}$.*

Throughout this section, we assume the data P to be subgaussian as follows:

Assumption 26 (Subgaussian data) *$P_1, \ldots, P_n \in \mathbb{R}^d$ are i.i.d. subgaussian random vectors with $\mathbb{E}[P_i] = 0$, $\mathbb{E}[P_{i,j}^2] = \sigma^2$ and $K = \max_i \|P_i\|_{\Psi_2}$.*

C.2.3 A concentration inequality for $\sqrt{L_1^F}$

Our goal is to obtain a concentration inequality for $L_1^F = \frac{1}{n}\max_{i=1,..,d} \|P_{:,i}\|_2^2$, where the data P is generated as in Assumption 26. We rather focus on $\sqrt{L_1^F} = \frac{1}{\sqrt{n}}\max_{i=1,...,d} \|P_{:,i}\|_2$.

Theorem 27 *(Vershynin, 2018, Theorem 3.1.1, equation (3.3)) Let $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$ be a random vector with independent subgaussian coordinates p_i that satisfy $\mathbb{E}[p_i^2] = 1$, and $K = \max_i \|p_i\|_{\Psi_2}$. Then, there exists $C > 0$ an absolute constant such that, for all $t \geqslant 0$,*

$$\mathbb{P}(|\|p\|_2 - \sqrt{n}| \geqslant t) \leqslant 2\text{exp}\left(-\frac{C}{K^4}t^2\right).$$

Equivalently, the concentration result of Theorem 27 can be extended by a linearity argument to random variables with variance $\mathbb{E}[p_i^2] = \sigma^2$: for all $t \in R$,

$$\mathbb{P}(|\|p\|_2 - \sigma\sqrt{n}| \geqslant t) \leqslant 2\text{exp}\left(-\frac{C}{\sigma^2 K^4}t^2\right).$$

In Theorem 27, vectors p correspond to columns of P as defined in Assumption 26. We conclude with the concentration of each norm of the columns $(P_{:,i})$ around $\sigma\sqrt{n}$. More precisely, according to Definition 24, $\|P_{:,i}\|_2^2 - \sigma\sqrt{n}$ are i.i.d. subgaussian variables. In Lemma 28, we provide a lower and an upper bound for the expectation of $\max_{i=1,\ldots,d} \|P_{:,i}\|_2$.

Lemma 28 *Let $P \in \mathbb{R}^{n\times d}$ be a collection of subgaussian elements as in Assumption 26. There exist $C_1, C_2 > 0$ absolute constant such that,*

$$C_2K^2\sigma \leqslant \mathbb{E}[\max_{i=1,\ldots,d}\left(\|P_{:,i}\|_2 - \sigma\sqrt{n}\right)] \leqslant 2\sigma K^2\sqrt{C_1\log(d)}.$$

Proof From Theorem 27 and Definition 24, there exists $C_1 > 0$ an absolute constant such that $Y = \|P\|_2 - \sigma\sqrt{n}$ is subgaussian with for all $t \in \mathbb{R}$, $\mathbb{E}[\exp(tY] \leqslant \exp(C_1t^2K^4\sigma^2) = \exp(v^2t^2/2)$. Boucheron et al. (Boucheron et al., 2013, Theorem 2.5) derived an upper bound for the expectation to the maximum of independent subgaussian random variables (and more generally, to subgamma random variables): $\mathbb{E}[Y] \leqslant \sqrt{2\log(d)}$. For the left-hand side, first we have the following inequality $\mathbb{E}[\max_{i=1,\ldots,d}\|P_{:,i}\|_2] \geqslant \mathbb{E}[\|P_{:,i}\|_2]$. This expectation can be lower bounded using (Vershynin, 2018, Theorem 3.1.1) and $1 + x \leqslant e^x$ for all $x > 0$. ■

Proposition 29 *Let P be a subgaussian matrix generated as in Assumption 26 and $L_1^F = \frac{1}{n}\max_{i=1,\ldots}\|P_{:,i}\|_2^2$. Then, there exists absolute constant $C, C_1, C_2 > 0$ such that for all $t \geqslant 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}}$,*

$$\mathbb{P}\left(|\sqrt{L_1^F} - \mathbb{E}[\sqrt{L_1^F}]| \geqslant t\right) \leqslant e^{-\frac{C}{\sigma^2K^4}\min(u_1(t),u_2(t))}.$$

where $u_1(t) = \log(d)(t + \frac{C_2K^2\sigma}{\sqrt{n}})^2$ and $u_2(t) = d(t - 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}})^2$.

Proof Recall that $\sqrt{L_1^F} = \max_{i=1,\ldots,d}\|P_{:,i}\|_2$, where $\frac{1}{\sqrt{n}}\|P_{:,i}\|_2 - \sigma$ is subgaussian. For the right side event, notice that $\mathbb{P}\left(\sqrt{L_1^F} \geqslant \mathbb{E}[\sqrt{L_1^F}] + t\right) \leqslant d\mathbb{P}\left(\frac{\|P_{:,i}\|_2}{\sqrt{n}} \geqslant \mathbb{E}[\sqrt{L_1^F}] + t\right)$ by the union bound. In addition, using the bounds on $\mathbb{E}(\max_{i=1,\ldots,d}\|P_{:,i}\|_2)$ from Lemma 28 and concentration from Theorem 27, there exists $C_2, C > 0$ such that $\mathbb{P}\left(\frac{\|P_{:,i}\|_2}{\sqrt{n}} \geqslant \mathbb{E}[\sqrt{L_1^F}] + t\right) \leqslant d\mathbb{P}\left(\frac{\|P_{:,i}\|_2}{\sqrt{n}} - \sigma \geqslant \frac{C_2K^2\sigma}{\sqrt{n}} + t\right) \leqslant \exp\left(-\frac{C\log(d)}{\sigma^2K^4}(t + \frac{C_2K^2\sigma}{\sqrt{n}})^2\right)$. For the left side event, using independence of the $P_{i,j}$, we conclude from Theorem 27 that there exists an absolute constant $C_1 > 0$ such that $\mathbb{P}\left(\sqrt{L_1^F} \leqslant -t + \mathbb{E}[\sqrt{L_1^F}]\right) \leqslant \mathbb{P}\left(\sqrt{L_1^F} \leqslant -t + \sigma + 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}}\right) \leqslant \mathbb{P}\left(\frac{\|P_{:,i}\|_2}{\sqrt{n}} - \sigma \leqslant -t + 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}}\right)^d \leqslant \exp\left(-\frac{Cd}{\sigma^2K^4}(t - 2\sigma K^2\sqrt{\frac{C_1\log(d)}{n}})^2\right)$. ■

We conclude from Proposition 29 that $\sqrt{L_1^F}$ concentrates around its mean, that admits lower and upper bounds as in Lemma 28. More precisely, there exist absolute constant

$C, C_1, C_2 > 0$ such that for all $t \geqslant 2\sigma K^2 \sqrt{\frac{C_1 \log(d)}{n}}^2$,

$$C_2 K^2 \sigma \frac{1}{\sqrt{n}} + \sigma - t \leqslant \sqrt{L_1^F} \leqslant \sigma + 2\sigma K^2 \sqrt{\frac{C_1 \log(d)}{n}} + t,$$

holds with probability $1 - \exp(-\frac{C}{\sigma^2 K^4} u(t))$, where $u(\cdot)$ is quadratic by part.

C.2.4 Concentration inequality for $\mu_{1,L}^F$ and μ_1^F under subgaussian data.

As we have seen in Lemma 23, μ_1^F and $\mu_{1,L}^F$ have deterministic approximants, closely related to the minimal eigenvalues of $PP^\top$ and $P^\top P$. Under subgaussian Assumption 26, we provide concentration inequalities for μ_1^F and $\mu_{1,L}^F$.

More precisely, recall from Lemma 23 that μ_1^F (resp. $\mu_{1,L}^F$) is lower bounded by $\frac{1}{d}\frac{\lambda_{\min(P^\top P)}}{n}$ (resp. $\frac{1}{n}\frac{\lambda_{\min}(PP^\top)}{d}$). Let us begin by calling a concentration inequality (Vershynin, 2018, Theorem 4.6.1) for eigenvalues of such subgaussian matrices.

Theorem 30 *(Vershynin, 2018, Theorem 4.6.1) Let $P \in \mathbb{R}^{n \times d}$ be a subgaussian matrix generated as in Assumption 26. Then, there exists an absolute constant $C_3 > 0$ such that, for all $t \geqslant 0$,*

$$\sqrt{n} - C_3 K^2(\sqrt{d} + t) \leqslant \sqrt{\lambda_{\min}(P^\top P)} \leqslant \sqrt{\lambda_{\max}(P^\top P)} \leqslant \sqrt{n} + C_3 K^2(\sqrt{d} + t),$$

with probability at least $1 - 2\exp(-t^2)$, with $K = \max_i \|P_i\|_{\phi_2}$.

Similarly, we deduce from Theorem 30 that there exists $C_4 > 0$ such that for all $t \geqslant 0$,

$$\sqrt{d} - C_4 K^2(\sqrt{n} + t) \leqslant \sqrt{\lambda_{\min}(PP^\top)} \leqslant \sqrt{\lambda_{\max}(PP^\top)} \leqslant \sqrt{d} + C_4 K^2(\sqrt{n} + t),$$

holds with probability at least $1 - 2\exp(-t^2)$. In particular, it is possible to derive bounds for quadratics of subgaussian data from Theorem 30. We provide a concentration result in Proposition 31 for $\mathbf{1}_d^\top P^\top P \mathbf{1}_d$ that appears in Lemma 23.

Proposition 31 *Let P be generated as in Assumption 26. Then, there exists $C_3 > 0$ such that, for all $t \geqslant 0$,*

$$\frac{1}{\sqrt{d}} - C_3 K^2 \left(\sqrt{\frac{1}{nd}} + \frac{t}{d\sqrt{n}}\right) \leqslant \sqrt{\frac{\mathbf{1}_d^\top P^\top P \mathbf{1}_d}{d^2 n}} \leqslant \frac{1}{\sqrt{d}} + C_3 K^2 \left(\sqrt{\frac{1}{nd}} + \frac{t}{d\sqrt{n}}\right),$$

holds with probability $1 - 2\exp(-t^2)$.

Proof This result is directly obtained from Vershynin (Vershynin, 2018, Theorem 4.6.1). Under the same assumptions than in Theorem 30, there exists an absolute constant $C_3 > 0$ such that, for all $t \geqslant 0$, $\|\frac{1}{n}P^\top P - I_d\| \leqslant K^2 \max(\delta, \delta^2)$, where $\delta = C_3 K^2(\sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}})$. From this, Vershynin concludes an approximate isometry for P (Vershynin, 2018, Lemma 4.1.5). For all $x \in \mathbb{R}^d$, $(1 - \delta)\|x\|_2 \leqslant \|Px\|_2 \leqslant (1 + \delta)\|x\|_2$. Note that this is exactly the same constant than in Theorem 30. ■

Using Theorem 30, Proposition 31 and Lemma 23, we get the expected concentration bounds n μ_1^F and $\mu_{1,L}^F$.

C.3 Proof for Corollary 10

Given the convergence guarantee for coordinate descent with GS-rule from Theorem 7, we conclude concentration for the convergence rate. We provide the proof for convergence in the underparametrized regime, where the convergence guarantee is given by $1-\frac{\mu_1^F}{L_1^F}$, and let the overparametrized regime to the reader. Recall from Proposition C.2 that, there exists absolute constant $C, C_1, C_2, C_3, K > 0$ such that,

$$1-\frac{1}{d}\frac{\Big(1+C_3K^2(\frac{1}{\sqrt{n}}+\frac{t}{\sqrt{nd}}\Big)^2}{\Big(1+C_2K^2\frac{1}{\sqrt{n}}+\frac{t}{\sqrt{n}}\Big)^2} \leqslant 1-\frac{\mu_1^F}{L_1^F} \leqslant 1-\frac{1}{d}\frac{\Big(1-C_3K^2(\frac{1}{\sqrt{n}}+\frac{t}{\sqrt{nd}}\Big)^2}{\Bigg(1+2K^2\sqrt{C_1\frac{\log(d)}{n}}+\frac{t}{\sqrt{n}}\Bigg)^2},$$

with probability $p(t) = 1-4\exp\Big(-\min(t^2, \frac{d\sigma^2}{n}(t-2K^2\sqrt{C_1\log(d)})^2, \frac{\log(d)\sigma^2}{n}(t+C_2^2K^2)^2)\Big)$. Applying a limited development in $\frac{1}{\sqrt{n}}$, the left term is equal to

$$1-\frac{1}{d}\left(1+\frac{2}{\sqrt{n}}[(C_3-C_2)K^2+\frac{t}{\sqrt{d}}-t]+o(\frac{1}{\sqrt{n}})\right),$$

and the right term,

$$1-\frac{1}{d}\left(1-\frac{2}{\sqrt{n}}[K^2(C_3+2\sqrt{C_1\log(d)})+\frac{t}{\sqrt{d}}+t]+o(\frac{1}{\sqrt{n}})\right).$$

These two limited development allows to conclude to the limiting convergence rate for coordinate descent with GS rule in the underparametrized regime.

Appendix D. Matching Pursuit in the Gauge Geometry

D.1 The Gauge Is a Norm: Proof for Lemma 11

Since $\alpha \to P\alpha$ is surjective, the function $\gamma_{\mathcal{P}}(x)$ is well-defined. Let us prove subbaditivity, positive definiteness and absolute homogeneity.

- Let $t > 0$ and $x \in \mathrm{R}^n$,

  $$\gamma(tx) = \inf_{\alpha\in\mathrm{R}^d, tx=P\alpha} \|\alpha\|_1 = \inf_{\tilde{\alpha}(=\frac{\alpha}{t})\in\mathrm{R}^d, x=P\tilde{\alpha}} \|t\tilde{\alpha}\|_1 = t \inf_{\tilde{\alpha}\in\mathrm{R}^d, x=P\tilde{\alpha}} \|\tilde{\alpha}\|_1 = t\gamma(x).$$

  Since $\mathcal{P} = \mathrm{conv}(P)$ is centrally symmetric, we conclude that for all $t \neq 0$, $\gamma(tx) = |t|\gamma(x)$. Finally, letting $t \to 0$, we conclude that $\gamma(0) = 0$.

- Let $x \in \mathrm{R}^n$ be such that $\gamma(x) = 0$. We have $0 = \inf_{\alpha\in\mathrm{R}^d} \|\alpha\|_1, \text{s.t. } x = P\alpha$. There exists (α_k) a sequence in R^d such that $x = P\alpha_k$ and $\|\alpha_k\|_1 \to 0$, meaning that $\alpha_k \to 0$. By linearity of $P\alpha$, we obtain $x = 0$.

- Let $x, y \in \mathrm{R}^n$, $\gamma(x+y) = \inf_\eta \|\eta\|_1, \text{s.t. } x+y = P\eta$. Let α, β be the minimal representation for x, y, such that $x = P\alpha$ and $y = P\beta$. We have that

  $$\gamma(x+y) = \inf_{\eta, P(\alpha+\beta)=P\eta,} \|\eta\|_1 \leqslant \|\alpha+\beta\|_1 \leqslant \|\alpha\|_1 + \|\beta\|_1 = \gamma(x) + \gamma(y).$$

If $\gamma(\cdot)$ is a norm, we compute its dual norm $\gamma^\star(z) = \sup_{x,\gamma(x)\leqslant 1}\langle z, x\rangle = \sup_x \inf_{\lambda\geqslant 0}\langle z, x\rangle + \lambda - \lambda\gamma(x) = \sup_x \inf_{\lambda\geqslant 0}\sup_{\alpha, x=P\alpha}\langle z, x\rangle + \lambda - \lambda\|\alpha\|_1 = \inf_{\lambda\geqslant 0}\lambda + \sup_\alpha\langle \alpha, P^\top z\rangle - \lambda\|\alpha\|_1 = \inf_{\lambda\geqslant 0, \|P^\top z\|_\infty\leqslant\lambda}\lambda = \|P^\top z\|_\infty$.

D.2 Proof for Sublinear Convergence of Matching Pursuit

We have seen in Section 2.5 that matching pursuit converges linearly in both the underparametrized and overparametrized regime. This result improves the sublinear guarantee of matching pursuit proven by Locatello et al. (2018, Theorem 3) letting a sublevel set radius appear (as usual in the coordinate descent literature).

Theorem 32 *Let f be convex, $L^f_{\gamma_{\mathcal{P}}}$-smooth with respect to the norm $\gamma_{\mathcal{P}}(\cdot)$. Then, the sequence verifies for $\mathcal{R} = \max_{x_\star\in X_\star}\max_{x\in\mathbb{R}^d}\gamma_{\mathcal{P}}(x - x_\star),\ \text{s.t. } f(x)\leqslant f(x_0) < +\infty$,*

$$f(x_k) - f_\star \leqslant \frac{2L^f_{\gamma_{\mathcal{P}}}\mathcal{R}^2}{k-1}.$$

Proof The sequence (x_k) verifies the descent lemma:

$$f(x_{k+1}) - f(x_k) \leqslant -\frac{1}{2L^f_{\gamma_{\mathcal{P}}}}\sigma_{\mathcal{P}}(\nabla f(x_k))^2.$$

We introduce $\delta_k = f(x_k) - f_\star$, so that $\delta_{k+1} \leqslant \delta_k - \frac{1}{L^f_{\gamma_{\mathcal{P}}}}\sigma_{\mathcal{P}}(\nabla f(x_k))^2$. By convexity and Cauchy-Schwarz inequality, we have $\delta_k \leqslant \langle x_k - x_\star, \nabla f(x_k)\rangle \leqslant \gamma_{\mathcal{P}}(x_k - x_\star)\sigma_{\mathcal{P}}(\nabla f(x_k))$. Then,

$$\delta_{k+1} \leqslant \delta_k - \frac{1}{L^f_{\gamma_{\mathcal{P}}}\gamma_{\mathcal{P}}(x_k - x_\star)^2}\delta_k^2 \leqslant \delta_k - \frac{1}{L^f_{\gamma_{\mathcal{P}}}\mathcal{R}^2}, \delta_k^2.$$

where $\mathcal{R} = \max_{x_\star\in X_\star, x\in\mathbb{R}^d}\gamma_{\mathcal{P}}(x_k - x_\star), \text{s.t.} f(x)\leqslant f(x_0)$ is assumed to be finite. Denoting $\omega = \frac{1}{L_{\mathcal{P}}\mathcal{R}^2}$, and dividing by δ_k, we have that $\frac{1}{\delta_k} + \omega\frac{\delta_k}{\delta_{k+1}} \leqslant \frac{1}{\delta_{k+1}}$. Since δ_k is nonincreasing with k, $\frac{1}{\delta_{k+1}} \geqslant \frac{1}{\delta_k} + \omega$. By summation, $\frac{1}{\delta_k} \geqslant \omega(t-1)$ and the result follows. ■

Appendix E. Steepest Coordinate Descent Is 'Nearly' a Matching Pursuit Algorithm

Coordinate descent with a Gauss-Southwell rule can be formulated using an LMO. Indeed, recall its formulation,

$$\alpha_{k+1} = \arg\min_{\alpha\in\mathbb{R}^d}\langle\nabla_{i_k}F(\alpha_k)e_{i_k}, \alpha - \alpha_k\rangle + \frac{L_2^F}{2}\|\alpha - \alpha_k\|_2^2 + \lambda|\alpha_{i_k}|,$$

where $i_k = \arg\min_k\min_{t\in\mathbb{R}} P^\top_{:,k}\nabla f(P\alpha)(t-\alpha_k) + \frac{L_2^F}{2}(t-\alpha_k)^2 + \lambda|t|$. We introduce the gauge function $\gamma_{\mathcal{P}}(x) = \inf_{\alpha\in(\mathbb{R}^+)^d}\sum_{i=1}^d \alpha_i P_i$. Then, applying the GS-rule can be formulated as,

$$\begin{aligned} i_k &= \arg\min_k \min_{u+\alpha_k\geqslant 0} P^\top_{:,k}\nabla f(P\alpha)u + \frac{L_2^F}{2}u^2 + \lambda|u + \alpha_k|, \\ &= \arg\min_k \Delta_k = \frac{1}{2L_2^F}(P^\top_{:,k}\nabla f(P\alpha) + \lambda - L_2^F\alpha_k)_+^2 - \frac{1}{2L_2^F}(P^\top_{:,k}\nabla f(P\alpha) + \lambda)^2. \end{aligned}$$

Let $\mathcal{P}$ be the set of atoms, $\mathcal{P}_k = \{k, \alpha_k \neq 0\}$ the set of visited atoms at iteration k. The algorithm consists in computing:

- for non visited atoms $k \in \mathcal{P} \setminus \mathcal{P}$ ($\alpha_k = 0$), $\Delta_k = -\frac{1}{2L_2^F}(-P_{:,k}^\top \nabla f(P\alpha) - \lambda)_+^2$ and this value can be computed using the $\mathrm{LMO}_{\mathcal{P}\setminus\mathcal{P}_k}(\nabla f(P\alpha_k)) = p_{\text{out}}$, leading to $\Delta_{\text{out}} = \frac{1}{2L_2^F}(-p_{\text{out}}^\top \nabla f(P\alpha_k) - \lambda)_+^2$, which costs $O(|\mathcal{P} \setminus \mathcal{P}_k|)$,

- for visited atoms, the objective needs to be computed completely in at most $|\mathcal{P}_k|$ iterations: $\Delta_{\text{in}} = \frac{1}{2L_2^F} \sup_{p\in\mathcal{P}_k} (p^\top \nabla f(P\alpha_k) + \lambda)^2 - (p^\top \nabla f(P\alpha_k) + \lambda - L\alpha_k^p)_+^2$.

Then, we compute i_k corresponding to the minimizer of $\min(\Delta_{\text{in}}, \Delta_{\text{out}})$, and compute the update $\alpha_{k+1}^{i_k} = (\alpha_k^{i_k} - \frac{1}{L_2^F}(p_{i_k}^\top \nabla f(P\alpha_k) + \lambda))_+$.

Appendix F. The Ultimate Method: an Inner Loop Strategy

The ultimate method is defined as a minimization problem, that has no closed-form solution. To overcome this issue, at each iteration k, we will solve iteratively the inner minimization problem using a randomized alternating minimization technique:

$$\min_{\eta,\beta\in\mathbb{R}^d} \langle P^\top \nabla f(P\alpha), \beta - \alpha\rangle + \frac{L_{\gamma_\mathcal{P}}^f}{2}\|\beta - \alpha\|_1^2 + \lambda\|\nu\|_1, \text{ such that } P\beta = P\nu.$$

First, we give the guarantees of an inner loop strategy, and in a second part, the result when applying a randomized alternating minimization technique to the inner loop.

Let us provide the convergence guarantees of the inner loop strategy, following the example of d'Aspremont et al. (2021, Section 5.2). Let $L(\beta, \nu) = \langle P^\top \nabla f(P\alpha), \beta - \alpha\rangle + \frac{L_{\gamma_\mathcal{P}}^f}{2}\|\beta - \alpha\|_1^2 + \lambda\|\nu\|_1$ and $L_\star^\alpha = \min_{\eta,\beta\in\mathbb{R}^d} L(\beta, \nu)$, such that $P\beta = P\nu$. Given an approximate solution to this problem at iteration k, Theorem 33 provides a guarantee on the outer loop.

Theorem 33 *Let $(\tilde{\beta}_k, \tilde{\nu}_k)$ with $\tilde{x}_k = P\tilde{\beta}_k = P\tilde{\nu}_k$ be an approximate solution of x_k produced by the ultimate method (3.8) such that $L^k(\tilde{\beta}_k, \tilde{\nu}_k) - L_\star^k \leqslant \epsilon_k$, then we have*

$$f(\tilde{x}_k) - f_\star \leqslant \left(1 - \frac{\mu_{\gamma_\mathcal{P}}^f}{L_{\gamma_\mathcal{P}}^f}\right)^k (f(x_0) - f_\star) + \sum_{i=0}^{k-1} \left(1 - \frac{\mu_{\gamma_\mathcal{P}}^f}{L_{\gamma_\mathcal{P}}^f}\right)^i \epsilon_{k-i}.$$

Proof

$$\begin{aligned}
f(\tilde{x}_{k+1}) &\leqslant f(\tilde{x}_k) + L^k(\tilde{\eta}_{k+1}, \tilde{\beta}_{k+1}), \\
&\leqslant f(\tilde{x}_k) + L_\star^k + \epsilon_k, \\
&\leqslant f(\tilde{x}_k) - \frac{\mu_{\gamma_\mathcal{P}}^f}{L_{\gamma_\mathcal{P}}^f}(f(\tilde{x}_k) - f_\star) + \epsilon_k, \\
f(\tilde{x}_{k+1}) - f_\star &\leqslant \left(1 - \frac{\mu_{\gamma_\mathcal{P}}^f}{L_{\gamma_\mathcal{P}}^f}\right)(f(\tilde{x}_k) - f_\star) + \epsilon_k.
\end{aligned}$$

The result is obtained by a direct summation. ■

The precision of the outer loop depends on the precision ϵ_k of the inner loop at each iteration k. Given an iterative method with iteration number t to obtain $(\tilde{\beta}_k^t, \tilde{\nu}_k^t)$, Theorem 33 ensures that the better the precision of the inner loop, the better the convergence guarantee of the outer loop. More precisely, if the error is constant $\epsilon_k = \epsilon$, the global convergence guarantee becomes $f(\tilde{x}_k) - f_\star \leqslant \left(1 - \frac{\mu^f_{\gamma_\mathcal{P}}}{L^f_{\gamma_\mathcal{P}}}\right)^k (f(x_0) - f_\star) + \epsilon \frac{L^f_{\gamma_\mathcal{P}}}{\mu^f_{\gamma_\mathcal{P}}} \left(1 - (1 - \frac{\mu^f_{\gamma_\mathcal{P}}}{L^f_{\gamma_\mathcal{P}}})^k\right)$. For a linearly decreasing inner precision $\epsilon_i = \left(1 - \frac{\mu^f_{\gamma_\mathcal{P}}}{L^f_{\gamma_\mathcal{P}}}\right)^i$, the global convergence rate is exactly $1 - \frac{\mu^f_{\gamma_\mathcal{P}}}{L^f_{\gamma_\mathcal{P}}}$. Finally, for a sublinearly inner precision, the global convergence guarantee is given by $F(\tilde{x}_k) - F_\star \leqslant \left(1 - \frac{\mu^f_{\gamma_\mathcal{P}}}{L^f_{\gamma_\mathcal{P}}}\right)^k (F(x_0) - F_\star) + \sum_{i=0}^k -1 \left(1 - \frac{\mu^f_{\gamma_\mathcal{P}}}{L^f_{\gamma_\mathcal{P}}}\right)^i \frac{1}{(k-i+1)^\alpha}$. The inner precision ϵ_k can slow down the global convergence of the ultimate method. However, achieving a low precision in the inner loop can be very costly, especially if the inner method converges sublinearly.

F.1 Alternating Randomized Block Coordinate Descent

We propose to solve this minimization problem that defines matching pursuit using an alternating minimization technique. Diakonikolas and Orecchia (2018) developed the alternating randomized block coordinate descent (AR-BCD), that generalizes alternating minimization to more than two blocks.

Theorem 34 *(Diakonikolas and Orecchia, 2018, Theorem 3.4) Let x_k be generated by AR-BCD with a distribution $(p_i)_{i=1,\ldots,n-1}$ over $n-1$ $L^{f,i}$-smooth blocks, and n be the non-smooth block. Then, for $R_i = \max_{x \in \mathbb{R}^d, f(x) \leqslant f(x_0)} \|x_{\star,i} - x_i\|^2$,*

$$\mathbb{E}[f(x_{k+1})] - f_\star \leqslant \frac{2}{k+3} \left(\sum_{i=1}^{n-1} \frac{L^{f,i}}{p_i} R_i \right).$$

The AR-BCD algorithm converges sublinearly for a non-smooth objective, according to Theorem 34. Let us first reformulate the minimization problem as an unconstrained problem, calling $\alpha = \alpha_k$:

$$\begin{aligned}
&\min_{\eta,\beta \in \mathbb{R}^d} \langle P^\top \nabla f(P\alpha), \beta - \alpha \rangle + \frac{L^f_{\gamma_\mathcal{P}}}{2} \|\beta - \alpha\|_1^2 + \lambda \|\eta\|_1, \text{ such that } P\beta = P\eta, \\
&= \min_{\eta,k \in \mathbb{R}^d} \langle P^\top \nabla f(P\alpha), \eta - \alpha \rangle + \frac{L^f_{\gamma_\mathcal{P}}}{2} \|\eta + k - \alpha\|_1^2 + \lambda \|\eta\|_1, \text{ such that } k \in \mathrm{Ker}(P), \\
&= \min_{\eta \in \mathbb{R}^d, z \in \mathbb{R}^{d_Q}} \langle P^\top \nabla f(P\alpha), \eta - \alpha \rangle + \frac{L^f_{\gamma_\mathcal{P}}}{2} \|\eta + Qz - \alpha\|_1^2 + \lambda \|\eta\|_1,
\end{aligned}$$

where Q is a basis for $\mathrm{Ker}(P)$ and $d_Q = \dim(\mathrm{Ker(Q)})$, obtained using a QR decomposition. We now use the eta-trick on $\|\cdot\|_1^2$ and an other η-trick on $\|\cdot\|_1$ (see A.1) , which leads to the equivalent minimization problem,

$$\min_{\eta\in\mathbb{R}^d, z\in\mathbb{R}^r}\min_{\gamma\in\Delta_d,\theta\geqslant 0} G(\eta,z,\gamma,\theta)=\langle P^\top\nabla f(P\alpha),\eta-\alpha\rangle+\frac{\lambda}{2}\left(\eta^\top\mathrm{Diag}(\theta)^{-1}\eta+\mathrm{Diag}(\theta)1\right)$$
$$+\frac{L^f_{\gamma_{\mathcal{P}}}}{2}(\eta+Qz-\alpha)^\top\mathrm{Diag}(\gamma)^{-1}(\eta+Qz-\alpha).$$

Remark 35 *Note that without applying the second η-trick on $\|\cdot\|_1$, the objective function is smooth with respect to z, no non-smooth with respect to η,γ, which prevents us from using the alternating minimization technique.*

Each coordinate can be solved as follows:

$$\begin{aligned}
z_{opt}&=\left(Q^\top\mathrm{Diag}(\gamma_k)^{-1}Q\right)^{-1}Q^\top\mathrm{Diag}(\gamma_k)^{-1}(\alpha-\eta),\\
\eta_{opt}&=\mathrm{Diag}(L^f_{\gamma_{\mathcal{P}}}/\gamma+\lambda/\theta)^{-1}\left(L^f_{\gamma_{\mathcal{P}}}\mathrm{Diag}(\gamma)^{-1}(\alpha-Qz)-P^\top\nabla f(P\alpha)\right),\\
\gamma_{opt}&=\frac{|\eta^i+(Qz)^i-\alpha^i|}{\|\eta+(Qz)-\alpha\|_1},\\
\theta_{opt}&=|\eta|.
\end{aligned}$$

Let us rewrite our problem into $\min_{\eta\in\mathbb{R}^d,z\in\mathbb{R}^r}\min_{\gamma\in\Delta_d,\theta\geqslant 0}G(\eta,z,\gamma,\theta)=G(\eta,z,\beta)=G(\xi)$ where $G(\cdot)$ is smooth with respect to η,z but non smooth with respect to $\beta=(\gamma,\theta)^\top$. We perform AR-BCD with probabilities p_1 for η (respectively $p_2=1-p_1$ for z). Let us rewrite $(\eta,z,\beta)=(\eta_1,\eta_2,\eta_3)$, and let $S_i(\xi)$ be the set of points that differs from ξ only over block i. AR-BCD is given at iteration k by:

$$\begin{aligned}
&\text{Pick } i_k\in\{1,2\} \text{ with probability } p_{i_k},\\
&\tilde{\xi}_{k+1}=\underset{\xi\in S_{i_k}(\xi_k)}{\arg\min}\,G(\xi),\\
&\xi_{k+1}=\underset{\xi\in S_3(\tilde{\xi}_{k+1})}{\arg\min}\,G(\xi).
\end{aligned}$$

F.2 Alternating Minimization

$$\min_{\eta\in\mathbb{R}^d,z\in\mathbb{R}^{d_Q}}\min_{\gamma\in\Delta_d}\langle P^\top\nabla f(P\alpha),\eta-\alpha\rangle+\frac{L^f_{\gamma_{\mathcal{P}}}}{2}(\eta+Qz-\alpha)^\top\mathrm{Diag}(\gamma)^{-1}(\eta+Qz-\alpha)+\lambda\|\eta\|_1.$$

The objective function $F(\eta,z,\gamma)$ is jointly convex in (η,z,γ) (for $\gamma>0$). This problem can be solved using alternating minimization (which is stronger than coordinate descent). Starting from $\eta_0\in\mathbb{R}^d, z_0\in\mathbb{R}^r(r=d-\mathrm{rg}(P)=r,\gamma_0\in\Delta_d$,

$$\begin{aligned}
z_{k+1}&=\arg\min_z F(\eta_k,z,\gamma_k),\\
\gamma_{k+1}&=\arg\min_\gamma F(\eta_k,z_{k+1},\gamma),\\
\eta_{k+1}&=\arg\min_\eta F(\eta,z_{k+1},\gamma_{k+1}).
\end{aligned}$$

In this context, it corresponds to computing

$$
\begin{aligned}
z_{k+1} &= \left(Q^\top \mathrm{Diag}(\gamma_k)^{-1} Q\right)^{-1} Q^\top \mathrm{Diag}(\gamma_k)^{-1}(\alpha - \eta_k),\\
\eta_{k+1} &= S_{\lambda/L^f_{\gamma_\mathcal{P}}\gamma_k}\left(\alpha - \frac{1}{L^f_{\gamma_\mathcal{P}}}\mathrm{Diag}(\gamma_k)P^\top \nabla f(P\alpha) - Qz\right),\\
\gamma_{k+1} &= \frac{|\eta^i_{k+1} + (Qz_{k+1})^i - \alpha^i|}{\|\eta_{k+1} + (Qz_{k+1}) - \alpha\|_1}.
\end{aligned}
$$

F.3 Experimental Results

In Figure 10, we apply an alternating randomized block coordinate descent (AR-BCD) technique to solve a well-chosen inner-loop optimization problem, which was developed by Diakonikolas and Orecchia (2018). The inner loop strategy is developed in F.1, as well as its convergence guarantee. We also derive an alternating minimization method on three blocks, as detailed in F.2, simpler than AR-BCD, for which there is no convergence guarantee.

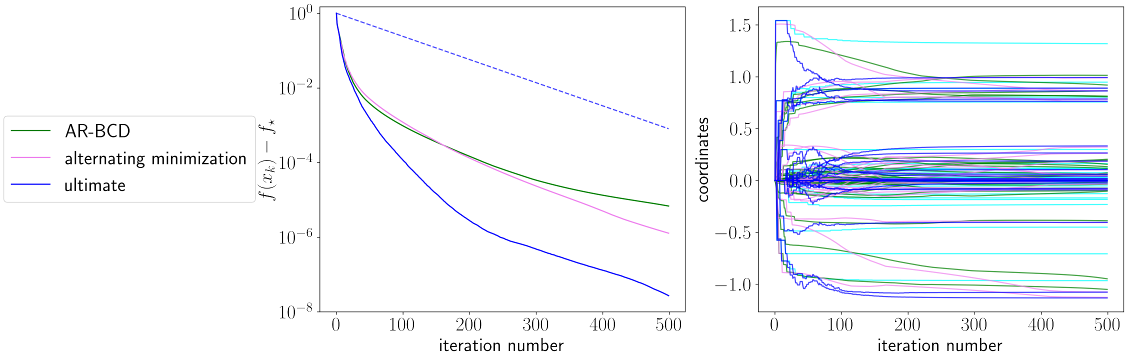

Figure 10: Convergence in function value for the ultimate method computed with an alternating minimization technique, AR-BCD, and with the solver MOSEK, compared to the regularized matching pursuit. Parameters are given by $d = 50$, $n = 20$, $s = 8$ and $\lambda = 0.001$, with an inner loop with exponential precision.