# Greedy Algorithms for Structurally Constrained High Dimensional Problems

**Authored by:**

Inderjit S. Dhillon
Pradeep K. Ravikumar
Ambuj Tewari

**Abstract**

A hallmark of modern machine learning is its ability to deal with high dimensional problems by exploiting structural assumptions that limit the degrees of freedom in the underlying model. A deep understanding of the capabilities and limits of high dimensional learning methods under specific assumptions such as sparsity, group sparsity, and low rank has been attained. Efforts (Negahban et al., 2010, Chandrasekaran et al., 2010} are now underway to distill this valuable experience by proposing general unified frameworks that can achieve the twin goals of summarizing previous analyses and enabling their application to notions of structure hitherto unexplored. Inspired by these developments, we propose and analyze a general computational scheme based on a greedy strategy to solve convex optimization problems that arise when dealing with structurally constrained high-dimensional problems. Our framework not only unifies existing greedy algorithms by recovering them as special cases but also yields novel ones. Finally, we extend our results to infinite dimensional problems by using interesting connections between smoothness of norms and behavior of martingales in Banach spaces.

## 1 Paper Body

Increasingly in modern settings, in domains across science and engineering, one is faced with the challenge of working with high-dimensional models where the number of parameters is large, particularly when compared to the number of observations. In such high-dimensional regimes, a growing body of literature in machine learning and statistics has shown that it is typically iropossible to obtain

consistent estimators unless some low-dimensional "structure" is imposed on the high dimensional object that is being estimated from the data. For instance, the sigoal could be sparse in some basis, could lie on some manifold, have some graphical model structure, or be matrix-structured with a low rank. Indeed,

given the variety of high dimensional problems that researchers face, it is natural that many novel notions of such low-dimensional structure will continue to appear in the future. There are a variety of issues that researchers have grappled with in this area but two themes stand out. First, there is the statistical problem of identifyiog the miniroum amount of data needed to accurately estimate high-¡limensional objects that are structurally constrained. Second is the computational issue of desigoing efficient algurithms that, in the ideal case, can recover high dimensional objects from a limited amount of data. Both of these themes have spurred a huge amount of work over the past decade. For each of the specific structures, a large body of work has studied regularized and

1

constrained M -estimators, where some loss function such as the negative log-likelihood of the data which measures goodness of fit to the data, is regularized by a function appropriate to the assumed structure, or constrained to lie within an appropriately chosen set. In recent years, researchers [I, 2] studying the statistical properties of such estimators have started discovering commonalities among proofs and analyses and have proposed unified frameworks that take advantage of such commonalities. Specifically, using a single theorem, they are able to rederive a wide range of known results on high-dimensional consisteney and error bounds for the various regularized and constrained estimators. The potential benefits are obvious: distillation of existing ideas and knowledge and the enabling of novel applications that are unexplored to date. In this paper, we consider the computational facet of such high-dimensional estimation, and propose a general computational scheme that can be used for recovering objects with low-dimensional structure in the high dimensional setting. A key feature of our general method is that, at each step, it greedily chooses to add a single "simple element" or "atom" to the current representation. The idea, of course, is not new. Indeed we show that our general framework yields several existing greedy algorithms if we specialize it appropriately. It also yields novel algorithms that, to the best of our knowledge, have not appeared in the literature so far. Greedy algorithms for optimizing smooth convex functions over the ii-ball [3,4,5], the probability simplex [6] and the trace norm ball [7] have appeared in the recent literature. Other recent references on greedy leaming algorithm for high-dimensional problems include [8, 9]. Greedy algorithms have also been studied in approximation theory [10, II] to approximate a given function, viewed as an element of a Banach space of functions, using convex combinations of "simple" functions. There is also the well-known viewpoint of seeing boosting algorithms as greedy minimization algorithms in function space (see, for example, [12, Section 3], and the references therein). Often, the proofs and results in these various settings resemble each other to a great extent. There is thus clearly a need for unification of ideas and proofs. In this paper, we focus on the underlying similarities between the greedy algorithms mentioned above. All these algorithms can be seen as specializations of a general computational scheme, with specific choices of the loss function, regularization or constraint set, and assumptions on the lowdimensional structure. Is there a commonality in their analyses of convergence rates, and are there key properties that

inform such analyses? Here, we identify two such key properties. The first is a restricted smoothness property (RSP) parameter (see also [13], for a similar quantity), which relates the smoothness of the function when restricted to sets with low-dimensional structure, and which depends on the ambient space norm, as well as a potentially distinct norm in which smoothness is established. The other, established in [I, 2], measures the size of the low-dimensional structured object with respect to an "atomic" norm. Using these two quantities, we are able to provide a general theorem that yields convergence rates for general greedy methods. We recover a wide range of existing results, as well as some potentially novel ones, such as for block-sparse matrices, lowrank tensors, and permutation matrices. In certain cases, most notably for low rank tensors, the scheme appears to lead to a greedy step that is intractable, which leads to intriguing questions about tractable approximations that we hope will be adequately addressed in the future. We then show how to extend these results to a general infinite-dimensional setting, by extending our definition of the restricted smoothness property (RSP) parameter, which allows us to obtain rates for L. spaces as well Banach spaces with Martingale type p. For the latter, the RSP parameter binges on the rate at which martingale difference sequences concentrate in that space, which provides yet another connection to the folklore statement that the "curse of dimensionality" in high dimensional settings is sometimes accompanied with the "blessings of concentration of measure".

2 PreUminaries 2.1

Atnms, Norms, and Structure

In Negahban et al.'s work[I], any specific structure such as sparsity is related to a low-dimensional subspace of structured vectors. In Chandrasekaran et al.'s work [2], this notion of structure is distilled further by the use of "atoms." Specifically, given a set A of very "simple" objects, called atoms, we can say that a vector x is simple (with some low-dimensional structure) if it can be written as a linear combination of few atoms: x = LZI C,8;, where k is small relative to the ambient 2

dimensionality. They then use these atoms to generalize the idea behind the use of i , -norm for sparsity, trace or nuclear norm for low rank, etc. Let A be a collection of atoms. We start by assuming [2] that these atoms lie in a finite-dimensional space, and that in particular A is a compact subset of some Euclidean space RP. Later, in Section 6, we will extend our treatment to include the case where the atoms belong to an infinite-dimensional space. Let CA denote the convex hull of A and define the gauge:

IlxiiA := inf{t ¡': a : x E tCA)} .

(I)

Note that the gauge II . IIA is not a norm in general, unless for instance A satisfies a technical condition, namely that it be centrally symmetric: x E A iff - x E A. Also, define the support function, IlxiiA := sup{ (x, a) : a E A}. If II . IIA happens to be a norm, then this is just the dual norm of I . IIA.

2.2 Examples Example 1. (Sparse vectors) A huge amount of recent literature deals with the notion of sparsity of high-dimensional vectors. Here, the set A c RP of atoms is finite and consists of the 2p vectors ?ei. This is a centrally

symmetric set and hence II? IIA becomes a norm, viz. the i , -norm. Example 2. (Sparse non-negative vectors) Using a slight variation on the previous example, the atoms can be the p non-negative basis vectors ei. The convex hull CA is the (p - I)-dimensional probability simplex. This is not centrally symmetric and hence I . IIA is not a norm. Example 3. (Group sparse matrices) Here the structure we have in mind for a p x k matrix is that it only has a few non-zero rows. This generalizes Example I which can be thought of as the case when k = 1. There are an infinite number of atoms: all matrices with a single non-zero row where that row has i.-norm I for some q ¿ 1. The convex hull CA becomes the unit ball of the I . 11.,1 group normonRPxk that is defined to be the sum ofthel.-norms of the rows of its matrix argument. Example 4. (Low rank matrices) This is another example that has attracted a huge amount of attention in the recent literature. The set I A E RPxp of atoms here is infinite and consists of rankone matrices with Frobenius norm 1. This is centrally symmetric and II?IIA becomes the trace norm (also called the nucleaTor Schatten-I norm, it is equal to the sum of the singular values of a matrix). Example 5. (Low rank tensors) This is a generalization of the previous example to higher order tensors. Considering order three tensors, the set A of atoms can be taken to be all rank-one tensors of the form U1 ¡8l U2 ¡8l Ua E Rpxpxp for Ui E RP, 1111;112 = 1. Their convex hull is the unit ball of II . IIA which can thought of as the tensor nuclear norm. Unfortunately, the tensor nuclear norm is intractable to compute and hence there is a need to consider relaxations to retain tractability. Example 6. (Permutation matrices) Here, we consider permutation matrices2 of size p x p as the set A of atoms. Even though there are p! of them, their convex hull has a succinct description thanks to the BiTklwff-von Newnann theorem: the convex hull of permutation matrices is the set of doubly stochastic matrices. As we shall see later, this fact will be crucial for the greedy algorithm to be tractable in this case.

3 Problem. Setup We consider the general optimization problem

min

x: IIxIlA1t

f(x),

(2)

where f is a convex and smooth function, and {x: IlxiiA ::; ,,} is the atomic norm constraint set that encourages some specific structure. This is a convex optimiwtion problem that is a constrained version of the usual regularized problem, minx f(x) + I'llxIlA. A line of recent work (see, for example, [2], and the references therein) has focused on different cases, with different atomic norms, i For simplicity we consider square matrices. It is definitely also possible to consider rectangular matrices in lRP1 XP2 for PI =f:. P2 2A pennutation matrix is one consisting only of O's & l's such that there is exactly a single 1 in each row & column. A non-negative matrix with every row & column sum equal to 1 is called a doubly stochastic matrix.

3

but largely on the linear case, where f(x) = lly - q.xll, for a given y E Rn and a linear map ¡P : RP -¿ Rn. ¡P is typically a linear measurement operator that

generates a noisy measurement y E Rn from an underlying "simple" signal Xt, and 11?112 is the standard Euclidean norm in Rn. For the linear case, projected gradient type methods have been suggested [2]. In this paper, we consider the general problem in (2), with a general loss function f(x), and a general constraint set induced by a structure-inducing atomic "norm" II? IIA. 3_1

Smoothness

We now discuss our assumptions on the loss function f in (2). We start by defining a restricted smoothness property that we require for our analysis. Consider a convex function f : RP -¿ R that is differentiable on some convex subset S of RP. Given a norm 11?11 on RP, we would like to measure how "smooth" the function f is on S with respect to 11?11. Towards this end, we define the following: Definition 1. Given a set S, and nonn II ,11, we define the Restricted Smoothness Property (RSP) constant of a function f : RP -¿ R as

L [[?11 (f ; S) :=

sup

x,yES,a:E(O,l]

f((l - a)x + ay) - f(x) - (V f(x), a(y - x)) 211 112 a Y- x

(3)

Since f is convex, it is clear that LII'II (f; S) ¡': O. The larger it is, the larger the function f "curves up" on the set S. Remark 1. (Connection to Lipschitz continuity of the gradient) Recall that a function f : RP -¿ R is said to have L-Upschitz continuous gradients w.r.t. II ?11 if for all x, y E RP, we have IIV f(x) V f(y)ll* ::; L 'lIx - yll where II . 11* is the norm dual to II . II. Using the mean value theorem it is easy to see that if f has L-Upschitz continuous gradient w.r.t. 11?11 then LII'II (f; S) ::; L. However, LII'II (f; S) can be much smaller since it only looks at the behavior of f on S and cares less about the global smoothness of f. Remark 2. (Connection to boundetiness of the Hessian) If the function f is twice differentiable on S, using second order Taylor expansion, L II ' II (f; S) can be bounded as LII'II (f; S)::;

sup

x,y,zES

(V2 f(z)(y - x),y - x) II _ xl12 Y

(4)

Again, suppose we have global control on V 2 f(x) in the form'lz E RP, IIIV2 fez) III ::; H where 111?111 is the II . II -¿ II . 11* operator norm of the matrix M defined as 111M III := sUPllxIl91IMxll*. Then, we immediately have LII'II (f; S) ::; H but this inequality might be loose in general. In the statement of our results, we will derive convergence rates that would depend on this Restricted Smoothness Property (RSP) constant of the loss function f in (2).

4 Greedy Algorithm and Analysis In this section, we consider a general greedy scheme to solve the general optimization problem in (2) where f is a convex, smooth function. The idea is to add one atom to our representation at a time in a way that the stucture of the set of atoms can be exploited to perform the greedy step efficiently. Our greedy method is applicable to any constrained problem where the objective is sufficiently smooth.

Algorithm 1 A general greedy algorithm to minimize a convex function f over the ",-scaled atomicnorm "ball" 1: Xo +- ",ao for an arbitrary atom ao E A 2: for t = 0, 1, 2, 3, ... do 3: a, +- argminaEA (V f(x,), a) 4: at +- argrninaE[o,l] f(x, + a(",at - Xt)) 5: Xt+l +- Xt + G't(x;at - Xt) 6: end for

4

Theorem 1. Assume that ! is convex and differentiable and let II . II be any norm. Then, for any 1. the iterates generated by Algorithm 1 lie in !¡CA and satisfy.

T

(5)

for any solution x* o! (2). Here LII'II (/; !¡CA) is the smoothness constant as defined in (3) and IIAII := sUP.EA II all?

Proof. Let us use the abbreviations L and R for LII'II (/; S) and IIAII respeetively. The fact that the iterates lie in !¡CA follows inunediately from the definition of the algorithm and a simple induction. Now assuming Xt E "CA. we have, by definition of L, for any a E [0,1], 2 !(Xt + a("at - Xt)) :S !(Xt) + a (V!(Xt), "at - Xt} + a2 LII"at - Xtl1 :S !(Xt) + a (V!(Xt), "at - Xt} + a2 L (211"atI1 2 + 211Xt112) :S !(Xt) - a( - (V!(Xt), "at} + (V!(Xt),Xt)) + 2a2L,,2 R2 . (6) The last inequality holds because II"atll, IIXtl1 :S "R. Now, for any minimizer x* of!, we have, by convexity of !,

8t := !(Xt) - !(x*) :S (V !(Xt), Xt - x*) = (V!(Xt), Xt} - (V !(Xt), x*) :S (V !(Xt), Xt) - (V!(Xt), "at} .

(7)

The last inequality holds because, at is the minimizer of the linear function (V!(Xt), -} over A (and Plugging (7) into (6), hence also over CA ) and x* /" E CA. Thus, (V!(Xt), at} :S (V !Xt), x* we have, for any a 0, !(xt+a("at -Xt)) :S !(Xt) -a8t +2a L,,2R . SinceXt+1 is chosen by minimizing theLHS over a E [0,1]. we have !(Xt+1) :S !(Xt) + mina E[o,1] (-a8 t + 2a2L,,2R2). Thus, we have, for all t O,8t +1 :S 8t + mina E[O,1] (-a8t + 2a2L,,2R2). For t = 0, choose a = 1 on the RHS to get 8, :S 2L,,2 R2. Since 8t's are decreasing, this shows 8t :S 2L,,2 R2 for all t 1. Hence, for t 1, we can choose a = 8';4L,,2 R2 E [0, ] on the RHS to get 1ft I, 8t+1 :S

t,,}.

8t -

8L!i R?. Solving this recursion easily gives, for all t I, !(Xt+1) - !(x*) :S 8.'..R'.

D

Remark 3. We emphasize that the norm II . II appears only in the analysis and not in the algorithm. Since the bound of Theorem 1 is simultaneously true for all norms II ,11, the best bound is achieved by choosing a norm that minimizes the product of IIAI12 and LII'[[ (/; !¡CA). Remark 4. We make the simple but useful observation that the iterate Xt can be written as a convex combination of at most t + 1 atoms, namely Ro, a" ... , at. Remark 5. Given ", Algorithm 1 is completely parameter free. This is a nice feature from a practical perspective as it frees the practitioner from the task of tuning parameters.

5 Special Cases Let us revisit the examples from Section 2.2 to see what concrete algorithms and accuracy bounds we get by speeializing Algorithm 1 and its bound (Theorem I) to them. Sparse vectors

The greedy step reduces to at

¡-

argmin aE?{el, ... ,ep

(V!(Xt), a} . }

Clearly, assuming that the gradient is already available, this can be done in O(P) time by finding j E {I, ... ,p} such thatj = argmaxj' I[V!(Xt)];.I and setting at = - sign([V!(xt)]j)ej. This actually gives a well-known algorithm whose roots go back to the 1950s [3]. More recently, a vatiant appeared as the Forward Greedy Selection algorithm in [5] (see also [4]). In fact, the original FrankWolfe algorithm can be applied whenever the set CA is polyhedral. If we choose the norm 11?11 to be iq then II All is 1 irrespective of q E [1,00] and the smoothness constant LII'II (/; !¡CA) is an increasing function of q. Hence to minimize the boond, we should choose p = 1 and measure smoothness of! over the It-scaled i , -ball using the i , -norm. When !(x) = Ily - ef?xll. we can use the connection to Hessian bounds (Remark 2) and inunediately get the upper bound 8,,2 . lief? T ef?111oo/T where the norm IIMII¿oo := sUPllxll,¡;1 IIMxlloo is simply maxi,j IMi,j I.

5

Sparse non-negative vectors

The greedy step becomes

a, ¡-

argmin

('il J(x,), aJ

.

aE{el, ... ,e p }

& in the previous example, this can be done in O(P) time given the gradient entries by computing j = argmin;'E{l, ... ,p} ['ilJ(x,)I;' and setting a, = e;. This particular algorithm to optimize a smooth function over the (scaled) probability simplex appears in [6]. Following the same reasoning as above, we get the best (among all i.-norms) bound if we choose II . II to be II . 111 and then our smoothness constant becomes similar to Clarkson's "nonlinearity measure" that he denotes by C f.

Group sparse matrices But still the greedy step

This is an interesting case since there are an infinite number of atoms.

a, ¡-

('ilJ(x,), aJ

argmin a: nnzrows(a)=l,IIBllq,l=l

(where nozrows counts the number of non-zero rows of a matrix) can be computed easily as follows. Let ri be the dual exponent of q that satisfies I/q + I/ri = I and find the row j of 'il J(x,) with maximal norm. Then, set a, to be the matrix all of whose rows are zero except row j. In row j, place the vector u T where u E Rkx1 is such that' (u, ['ilJ(xtllD = -11['ilJ(x,)lIII ?? and lIull. = 1. Such a vector u can be found in closed form. For the case J(x,) = Ily -

il?xll, choosing the norm 11?11 in Theorem I to be 11,11.,1 (and this gives the optinlal bound among allll?II.,r norms for r ¿ I), we get the accuracy bound: 8,,2 '11iI?T il?11.,1 .,=/Twhere the q, I -+ q, 00 norm of the operator il?T iI? is defined as sup{lliI? T il?MII.,= : M E RPx., IIMII.,l :0; I}. This algorithm and its analysis are novel to the best of our knowledge. However, we note that a related greedy algorithm (that does not directly optimize the objective (2? called Group-OMP appears in [14, 15].

i..

!

Low rank matrices As in the previous case, we have an infinite number of atoms: all rank-I matrices with Frobenius norm I. Yet, the greedy step

a, ¡-
argmin
('il J(x,),
aJ
a: rank(a)=l,lIaIlF=l

can be done in polynomial time by computing the SVD, 'ilJ(x,) = UEV T and setting a = -U1V! where U" v, are the left, right singular vectors corresponding to the largest singular value 0'1. Since we only need the singular vectors corresponding to the largest singular value, the computation of a, can be done much faster than the time it takes to compute a full SVD. For the case J(x) = Illy - il?xll, the bound of Theorem I is minimized, among all Schatten-pnorms" by using 11?11 = Ils(l), i.e. the trsce or nuclear norm. Since the objective is twice differentiable, using Remark 2 we get the following upper bound on the accuracy: 8,,2 . IliI?T il?lls(l)s(=)/T which depends on the S(I) -+ operator norm of iI? T iI? which is defined as sup{lIiI? T il?Mlls(=) : M E RPxp, IIMlls(l) :0; I}. This algorithm was recently independently discovered and analyzed in [7].

ff?
stool
Low rank tensors
Here, the greedy step
a, ¡-
argmin
('il J(x,), aJ
a: a=ul@u2I8lus,lIuiI12=1

appears intrsctable. Indeed, the above problem is closely related to the problem of finding the best rank-one approximLltion to a given tensor which is known to be NP-hard [16] already for order-3 tensors. However, as described in [2], it is possible to construct a fanilly of outer approximLltions CA ¡;; ??? ¡;; THk+! ¡;; THk such that, for any fixed k, THk can be described by a semidefinite program of size polynomial in k. So, even though the exact greedy step above may not be tractable, we can use these "theta bodies" (whence the notation 'TH") to approximate the greedy step. The iterates will no longer lie strictly in the tensor nuclear ball of the given radius. Understanding the implications of such approximations and their analysis are interesting questions

to pursue but lie beyond the scope of the current paper. 3We use MATLAB notation M j ,: to denote row j of a matrix M. 4The Schatten-q norm of a matrix is the i q norm of its singular values. 6

Permutation matrices

Here, fortunately, we again do not face intractability: the step axgmin

aT ¡-

(V/(xt),a)

a : a is a permutation matrix

reduces to solving a linear assignment problem with costs C(i,j) = [V/(xtlk;. This can be efficiently done using, for example, the Hungarian algorithm. Another way to see that the above step does not involve combinatorial explosion is to appeal to the Birkhoff-von Neumann theorem that statea that the convex hull of permutation matrices is the set of doubly stochastic matrices. As a result, the above reduces to minimizing a linear objective (V I(xtl, M) subject to polynomially many constraints: M 2': 0, M1 = 1 and MT 1 = 1.

6 Extension to Infinite Dimensional Banach Spaces 1n !hi section, we consider an extension of the framework behind Algorithm 1 to the case when the set of atoms are in some infinite dimensional (real) Banach space (V, II . II). For example, the atoms could be some "simple" real valued functions on some interval [a, b] JR. The two ingredients in our framework were the atomic norms, and the Restricted Smoothness Property (RSP) parameters. 1n [2], and in Section 2.1, the atoms were considered as belonging to a finite dimensional Euclidean space. Note however that the definition of the atomic norms in (1) did not make use of the topology of the ambient space, and hence is applicable even when the atoms belong to some Banach space (V, 11?11). However, our definition of the RSP parameter in (3) relied critically on the Euclidean inner product, whence we will now extend this to the infinite dimensional case in the sequel. Consider a convex continuous Frechet differentiable function I : V –+ JR, and let V I(x) denote the Frechet derivative of I at x. Let (.,.) : V* x V –+ JR denotea the bilinear function (which is not an inner product in general) (X, x) := X(x) for x E V and X in the dual space V* (consisting of bounded linear functions on V). Definition 2. Given a Banach space (V, II . II), and a set S V, and some r E [1,2], we define the Restricted Uniform Smoothness Property (RUSP) constant of a convex continuous Frechet differentiable function I : V –+ JR as 1((1 - a)x + ay) - I(x) - (V I(x), a(y - x)} Lr (/; S) := sup (8) (l/r) arlly - xll r x,yES,o:E[O,l] This need not be bounded in general, but would be bounded for instance if the function I were runiformly smooth (though this would be a far stronger condition). Suppose the set of atoms A V is such that max.,E.A (X, a) is defined for any X E V*. Then, we can define a straightforward extension of Algorithm I given as Algorithm 2. Algorithm 2 A general greedy algorithm to minimize a continuous Frechet differentiable convex function lover the convex hull of a set of atoms A in a Banach space (V, II ,11) 1: Xo ¡- ao for an arbitrary atom ao E A 2: for t = 0, 1, 2, 3, ... do 3: X t E V* +- V I(xt), the Frechet derivative of I at Xt 4: at ¡- argmaxaE.A (-Xt,a) 5: at +- argminaE[O,l] I(xt + a(at - Xt)) 6: Xt+l +- Xt + at( at - Xt) 7: end for The following result proves a general rate of convergence for Algorithm 2. Since the proof follows the proof

of Theorem I very closely, we defer it to the appendix. Theorem 2. Suppose that (V, II . II) is a Banach space and let I : V –+ JR be a convex continuous Frechet differentiable function. Let A be a set of atoms such that II all :¡; Rfor all a E A, and let S = conv(A). Suppose the Restricted Uniform Smoothness Property (RUSP) constant Lr (/; S) of I is boundedfor some r E [1,2]. Then,

I(xt) - inf I(x)

=

xES

where the hidden constant depends on r only. 7

0 (Lr (/; S)

tr

1

RT)

6.1 Rates of Convex Approximation in Lp spaces For p E (1,00) the space Lp([a, b]) consists of all functions 9 : [a, b] –¿ IR such that the (Lebesgue) integral Ig(x)IPda; is fioite. The space Lp is a Banach space once we eqoip it with the norm

J:

(J:

IlgiIL. := Ig(x)IPda;) 1/? . LetA be a set of atoms in L. with bounded norm and let hE L. be a function that we wish to approximate using convex combinations of the atoms. Since, the function 9 r-¿ Ilgll is 1" = min{p, 2} uoiformly smooth for p E(1,00), we can use Algorithm 2 to generate a sequence? of functions 91, !J2, ... such that gt is a convex combination of only t atoms. Moreover,

(,.t, ).

we will have the guarantee: Ilgt+1 - hll( - infgEconv(A) Ilg - hll( = 0 Such rates of convex approximation in non-Hilbert spaces have been studied earlier (see, for example, [10, II]). Note that, unlike [10), we do not assume that h E conv(A). If that is the case, the above rate simplifies to the rates given in [10): O(t-1+) for p E (1,2), and orr!) for p ¿ 2.

6.2 Rates of Convex Approximation in Spaces with Martingale Type p Note the fact that, in the previous subsection, the only property of L. spaces that we used to get rates was the fact that the norm to some power was a uoiformly smooth convex function. It tums out that the existence ofuoiformly smooth functions in a given Banach space is intimately connected to the behavior of martingale difference sequences in that space. To precisely state the connection, we need to define the notion of martingale type (also called Haar type) [17, p. 320). A Banach space (V, 11?11) is said to have martingale type p (M-type p in short) if there exists a constant K. such that, for all T ;::: I, and any V -valued martingale difference sequence d" ... ,dT, we have

Note that, by triangle inequality for norms, any Banach space always has M-type I while a Hilbert space (i.e. the norm II . II comes from an inner product) has M-type 2. Hilbert space essentially have the best M-type in the sense that no Banach space has M-type p for p ¿ 2. The conoection of M-type to uoiform smoothness is made precise by the following remarkable theorem (see also [18]).

Theorem (Pisier, [19). A Banach space has M-type p iff there is an equivalent norm' 11?11# such thar the function II'II is p-uniformly smooth. Consider the setting of the previous subsection where we have some hE conv(A) for some set A of atoms in an arbitrary Banach space (V, 11?11). Using Pisier's theorem, we get the following corollary. Corollary3. Suppose A is a seto/atoms in a Banach space (V, 11?11) that has M-typepand let hE conv(A). Suppose Algorithm 2 generates iterates g" g2, ... when run on the function 9 r-¿ Ilgll whose existence is guaranteed byPisier's theorem Then, we have,

Ilg'+1 - hll =

0

(C1+*).

7 Future Work First, we envisage the algorithm being used to compute the entire regularization parh corresponding to all values of the constraint parameter 1¡. Using a warm start strategy, where the algorithm for higher values of I¡ is ioitialized with the solution for lower values, can be very helpful here. Exploring this to get a general practical algorithm to compute the entire path would be very oice. Third, linear convergence guaraotees for projected gradient type methods have been obtained by [13) where they make the additional assumption of (generali2ed) restricted strong convexity. It should be possible to derive similar faster rates for our greedy algorithm.

Acknowledgments We gratefully acknowledge the support of NSF under grant IIS-1018426. ISD acknowledges support from the Moncrief Grand Challenge Award.

'Thatis,cLII?1I

11'11# cull' II forsomecL,cu ¿ O. 8

# 2 References

[I] S. Negahban, P. Ravikumar, M. Wainwrigbt, and B. Yu. A unified framework for higb-dimensional analysis of M-estimators with decomposable regularizers. In Advances in Neural In/onnation Processing Systems 22, pages 1341¿-1356,2009.

[2] V. Chandrasekaran, B. Rech P. A. Parri1o, and A. S. Willsky. The convex geometry of linear inverse problems. In Proceedings of the 48th Annual Allerton Conference on Communication, Control and Computing, pages 699-703, 2010. [3] M. Frank and P. Wolfe. An algorithm for quadratic programming. Naval Research Logistics Quarterly, 3(1-2):95-110,1956. [4] T. Zhang. Sequential greedy approximation for certain convex optimization problems. IEEE Transactions

on Information T!reory, 49(3):682-{;91, 2003. [5] S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. SIAM Journal on Optimiuztion, 20(6):2807-2832, 2010.

[6] K. Clarkson. Coresets, sparse greedy approximation" and the Frank-Wolfe algorithm. In Proceedings of the nineteenth annual ACM-SIAM symposiwn on discrete algorithms, pages 922-931. Society for Industrial and Applied Matherustics, 2008.

[7] S. Shalev-Shwartz, A. Gonen, and O. Shamir. Large-scale convex minimization with a low-rank constraint. In Proceedings o/the 28th International Conference on Machine Learning. pages 329-336, 2011. [8] H. Liu and X. Chen. Nonpararnetric greedy algorithms for the sparse learning problem. In Advances in Neuralinformation Processing Systems 22, pages 1141-1149,2009. [9] A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection" sparse approximation and dictionary selection. In Proceedings of the 28th International Conference on Machine Learning, pages 1057-1064,2011. [10] M. 1. Donahue, C. Darken, L. Gurvits, and E. Sontag. Rates of convex approximation in non-Hilbert spaces. Constructive Approximation, 13(2):187-220,1997. [II] V. N. Thmlyakov. Greedy approximation. Acta Numerica. 17:235-409,2008. [12] R. E. Schapire. The boosting approach to machine learning: an overview. In D. D. Denison, M. H. Hansen" C. C. Holmes, B. Mallick, and B. Yu, editors, Nonlinear estimation and classification, volume 171 of Lecture Notes in Statistics, pages 149-172. Springer, 2003.

[13] A. Agarwal, S. Negahban, and M. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In Advances in Neural Information Processing Systems 23, pages 37-45,2010. [14] A. C. Lozano, G. Swirszcz, and N. Abe. Grouped orthogonal matching pursuit for variable selection and prediction. In Advances in Neural Information Processing Systems 22, pages 1150–1158, 2009. [15] A. C. Lozano, G. Swirszcz, and N. Abe. Grouped orthogonal matching pursuit for logistic regression. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, volume 15 of JMLR Workshop and Conference Proceedings, 2011. [16] C. Hillar andL.-H. Lim. Most tensor problems areNF hard, 2010. available at http:/ / arxiv. org/ abs/0911.1393v2.

[17] A. Pietsch. History of Banach spaces and linear operators. Birkhiuser, 2007. [18] 1. Borwein, A. 1. Guirao, P. Hljek, and 1. Vanderwerff. Uniformly convex functions in Banach spaces. Proceedings of the American Marhematical Society, 137(3):1081-1091, 2009.

[19] G. Pisier. Martingales with values in uniformly convex spaces. Israel Journal of Mathematics, 20(34):326-350, 1975.

9