Exploiting the Structure via Sketched Gradient Algorithms

Citation for published version:

Published In:
Proceedings of the 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
EXPLOITING THE STRUCTURE VIA SKETCHED GRADIENT ALGORITHMS

Jungi Tang, Mohammad Golbabaee, Mike Davies

Institute for Digital Communications, University of Edinburgh. EH9 3JL, UK

ABSTRACT

Sketched gradient algorithms [1] have been recently introduced for efficiently solving the large-scale constrained Least-squares regressions. In this paper we provide novel convergence analysis for the basic method Gradient Projection Classical Sketch (GPCS) to reveal the fast linear convergence rate of GPCS thanks to the intrinsic low-dimensional geometric structure of the solution prompted by constraint set. Similar to our analysis we observe computational and sketch size trade-offs in numerical experiments. Hence we justify that the combination of gradient methods and the sketching technique is a way of designing efficient algorithms which can actively exploit the low-dimensional structure to accelerate computation in large scale data regression and signal processing applications.

Index Terms— Large data optimization, Inverse problems, Sketching, Gradient projection

1. INTRODUCTION

In many signal processing, data mining and machine learning applications, people encounter large scale optimization tasks which involve regression on a huge amount of data. In recent years there has been a growth in the amount of research on fast first order optimization algorithms such as the variance-reduced stochastic gradient descent [2][3] which has scalability with respect to the amount of the data points to be processed by constructing a stochastic estimate of the truth gradient each iteration. Such novel algorithms have been shown to be effective in various large scale regression problems when compared to conventional full gradient methods such as FISTA [4].

Classical randomized sketching techniques [5][6] and iterative sketching techniques [7][8] have been recently introduced to construct subprograms with a reduced sample-size. Meanwhile the maximum sample size reduction (aka. the smallest sketch size) is directly related to the intrinsic dimension of the desired solution prompted by low-dimensional structure such as sparsity, group sparsity and low-rank. Such sketching techniques open the door for designing optimization algorithms which are able to exploit these low-dimensional structures and hence achieve scalability and acceleration. Very recently, the sketched gradient algorithms [1] have been proposed based on a combination of iterative sketching and projected gradient descent for constrained Least-squares regression. The sketched gradient algorithms merit the strength of both full gradient algorithms and stochastic gradient algorithms and appear to be efficient in practice.

This paper focuses on exploring how to optimally exploit the constraint set via randomized sketched gradient algorithms to speedily and approximately solving large scale constrained Least-square estimation problems.

2. BACKGROUND

In many signal processing and image reconstruction applications the forward model is linear, which involves the "ground truth" \( x^\star \) which a priori is assumed to belong to model (constraint) set \( \mathcal{K} \), with measurement operator \( A \) and the noise vector \( w \):

\[
y = Ax^\star + w, \quad x^\star \in \mathcal{K}, \quad A \in \mathbb{R}^{n \times d}.
\] (1)

In this paper we focus on the large data setting \( n \gg d \). We wish to estimate \( x^\star \) via a constrained Least-squares (LS) estimator:

\[
x_{LS} = \arg \min_{x \in \mathcal{K}} \{ f(x) := \|y - Ax\|_2^2 \}.
\] (2)

A popular approach for (2) is the Projected Gradient Descent (PGD) [9]:

\[
x_{j+1} = \mathcal{P}_\mathcal{K}(x_j - \eta A^T (Ax_j - y)).
\] (3)

The PGD/ISTA algorithm and its accelerated variant FISTA [4] have been of great interest in signal processing and compressed sensing applications since they cope with various type of signal models prompted by hard constraints or non-smooth regularization. We refer to these methods as full gradient methods (FG). The main drawback of FG is when the linear systems’ size is large, each iteration of FG methods becomes costly to compute. Alternatively, recent works on sketching techniques [10][11][6][8] and consequently the sketched gradient methods [1] which was built on both the Classical Sketch [6] as the fast initialization step and Iterative Hessian Sketch [8] for solving the Least-squares regression to machine precision. In this paper we focus particularly on the novel convergence analysis of the initialization step which is called Gradient Projection Classical Sketch.

2.1. Main contributions

- For the sketched Least-square problem where very often we may wish to have a sketch size \( m \) which is smaller than the ambient dimension \( d \), from an optimization point of view, it is unknown that when a first order method can be efficiently used and whether the convergence speed can be fast since the global strong convexity assumption is vacuous in this setting. We show that if a Gaussian sketching matrix is used to construct the sketched problem and we choose to use the standard projected gradient descent as the solver, this method converges towards the vicinity of \( x^\star \) with a linear convergence rate, as long as the sketch size \( m \) is larger than a certain measure (Gaussian width) of the intrinsic dimension of \( x^\star \) associated with the constraint set \( \mathcal{K} \). Moreover, the convergence speed is also a function of the sketch size choice and the statistical dimension. Hence we demonstrate the fact that by combining the projected gradient methods and randomized sketching techniques, one can design accelerated algorithms

\[
\text{for efficiently solving the large-scale constrained Least-squares regression.}
\] (GPCS) to
which can actively exploit the statistical structure of the signal/image to be estimated. While in previous work [1], the GPCS is also analyzed but in a conservative way which does not include the case of choosing a sketch size smaller than the ambient dimension.

- We study the computational trade-off between the sketch size and the computational cost with respect to the structure of the ground truth and the constrained set for the sketched problem solved by the projected gradient descent.

### 3. GRADIENT PROJECTION CLASSICAL SKETCH

We first introduce the Gradient Projection Classical Sketch (GPCS) which is the initialization loop of our previously proposed Gradient Projection Iterative Sketch algorithm. This simple algorithm is basically running the PGD to solve the sketched Least-square problem (4). We denote $S \in \mathbb{R}^{m \times n} (m \ll n)$ as the randomized sketching operator which significantly reduces the sample dimension of the LS estimator.

$$\hat{x} = \arg\min_{x \in \mathcal{K}} \left\{ f_0(x) := \|Sy - SAx\|_2^2 \right\}, \quad (4)$$

In next section, we describe the convergence behavior of the GPCS algorithm.

**Algorithm 1: Gradient Projection Classical Sketch**

**Initialization:** $x_0 = 0$; Given $A \in \mathbb{R}^{n \times d}$, sketch size $m \ll n$; Prior knowledge: the true solution $z$ belongs to set $\mathcal{K}$; Generate a random sketching matrix $S \in \mathbb{R}^{m \times n}$; Calculate $SA, Sy$;

while $i = 0: k - 1$ do

\[ x_{i+1} = P_K(x_i - \eta(SA)^T(SAx_i - Sy)); \]

end

Return $x = x_k$;

### 4. CONVERGENCE ANALYSIS

We start by defining some certain properties of the operator $A$, sketching scheme $S$, and its interaction with the signal model (constraint set), in a similar manner to the analysis in [1][12] and [6]:

**Definition 1.** Let $\mathcal{C}$ be the smallest closed cone at $x^*$ containing the set $\mathcal{K} = x^*, S^{d-1}$ be the unit sphere in $\mathbb{R}^d$, $\mathbb{R}^d$ be the unit ball in $\mathbb{R}^d$, $z$ be an arbitrary fixed unit-norm vector in $\mathbb{R}^n$. The contraction factor $\alpha(\eta, SA)$, the error amplification factor $\beta(S, A)$ are defined as:

\[
\alpha(\eta, SA) = \sup_{u,v \in \mathcal{C} \cap \mathbb{R}^d} v^T(I - \eta A^T S^T SA)u, \quad (5)
\]

\[
\beta(S, A) = \sup_{v \in \mathcal{A} \cap \mathbb{R}^n} v^T S^T S/m z. \quad (6)
\]

For the convenience of the presentation of our main theorem, we shall denote these two key factors defined in Def.1 as $\alpha := \alpha(\eta, SA)$ and $\beta := \beta(S, A)$, respectively.

**Definition 2.** The cone-restricted strong convexity constant $\mu_c$ is defined as the largest positive constant which satisfies: for all $z_c \in \mathcal{C}$

\[
\|Az_c\|_2^2 \geq \mu_c\|z_c\|_2^2. \quad (7)
\]

The cone restricted strong-convexity defined here can be viewed as a recovery/inversion stability measure which ensures that the original Least-squares estimator is reliable and robust to noise: $\|x_{LS} - x^*\|_2 \leq \frac{\|w\|_2}{\mu_c}$, see [13, Proposition 2.2].

**Definition 3.** The Lipschitz constant $L$ for the LS (2) is defined as the largest singular value of the Hessian matrix $A^T A$: for all $z \in \mathbb{R}^d$

\[
\|Az\|_2^2 \leq L\|z\|_2^2, \quad (8)
\]

Now we are ready to present our main result on linear convergence of the GPCS:

**Theorem 1.** Starting from $x_0$, if the step size $\eta$, the sketching operator $S \in \mathbb{R}^{m \times n}$ and sketch size $m$ are properly chosen such that $\alpha(\eta, S^T A) < 1$, the following error bound holds:

\[
\|x_k - x^*\|_2 \leq \alpha^k\|x_0 - x^*\|_2 + \frac{mnL\beta}{1 - \alpha}\|w\|_2. \quad (9)
\]

**Proof.** The proof of Theorem 1 can be found in appendix. \[\square\]

**Theorem 1** reveals that as long as the step size, the sketching matrix and sketch size are chosen properly, the updates sequence $x_1, x_2, ..., x_k$ generated by the GPCS algorithm converge exponentially towards the vicinity of the ground truth $x^*$ with a distance scales with the noise energy $\|w\|_2$. In the forthcoming subsection we explicitly quantify the two pillar factors $\alpha$ and $\beta$ in Theorem 1 when $S$ is a Gaussian sketch, and hence reveals the structure-exploiting property of the GPCS algorithm.

#### 4.1. Explicit analysis results for Gaussian sketches

Theorem 1 has provided us a general framework to describe the convergence of GPCS in terms of $\alpha$ and $\beta$. We can derive expressions of these terms in terms of the Gaussian Width of the set $\mathcal{W} := \mathcal{W}(\mathcal{A} \cap S^{d-1})$ and the sketch dimension $m$ when we choose $S$ to be a Gaussian sketching matrix. The Gaussian Width is defined as:

**Definition 4.** [13, Definition 3.1] The Gaussian width of a convex set $\Omega$ is defined as:

\[
\mathcal{W}(\Omega) = E_g \left( \sup_{v \in \Omega} v^T g \right), \quad (10)
\]

where $g \in \mathbb{R}^n$ is draw from distribution $\mathcal{N}(0, I_n)$.

The square of Gaussian width $\mathcal{W}^2(\mathcal{C} \cap S^{d-1})$ is a well-known tool of measuring the statistical intrinsic dimension of $x^*$ enforced by the constrained set $\mathcal{C}$. If $x^*$ has only $s$ non-zero entries, and we construct the Least-square estimator with a 1-constraint, the width $\mathcal{W}^2(\mathcal{C} \cap S^{d-1})$ can be upper bounded by $2s\log(\frac{d}{s}) + \frac{2}{n} s$ [13, Proposition 3.10], which means $\mathcal{W}(\mathcal{C} \cap S^{d-1})$ scales with the sparsity of $x^*$.

In parallel to the step size choices described in [12] (which is about the PGD convergence analysis on solving the original LS problems where $A$ is a Gaussian map), our analysis covers a greedy choice and a conservative choice of the step size $\eta$. We denote $b_m = \sqrt{\frac{1}{\eta} \frac{1}{\mathcal{W}(\Omega)^2}} \approx \sqrt{m}$ as the same in [1] and [12]. For the greedy choice, we can bound $\alpha$ as described in Lemma 1.

1 In optimization community, this convergence rate also referred to as linear convergence.
Lemma 1. (Greedy step size) If the step-size $\eta = \frac{\mu_e}{L}$, and the entries of the sketching matrix $S$ are i.i.d drawn from Normal distribution, then:

$$\alpha(\eta, SA) \leq (1 - \frac{\mu_e}{L})(1 + \frac{(W + \theta)^2}{b_m^2}) + \frac{\sqrt{S(W + \theta)}}{b_m},$$

(11)

with probability at least $1 - 8e^{-\frac{\alpha^2}{2}}$.

This result for the greedy step size is reminiscent of the result for greedy step size described in [12, Theorem 2.2], in a sense that when $\frac{L}{\mu_e} \to 1$, we recover their convergence result. However, Lemma 1 does not ensure $\alpha < 1$ since it inherently demands a sketch size $m \gtrsim \left(\frac{W}{\mu_e}\right)^2$. This sketchsize requirement can be moderated, at the cost of a more conservative stepsize:

Lemma 2. (Conservative step size) If the step-size $\eta = \frac{\mu_e}{L(b_m + \sqrt{d} + \theta)}$, and the entries of the sketching matrix $S$ are i.i.d drawn from Normal distribution, then:

$$\alpha(\eta, SA) \leq \left(1 - \frac{\mu_e}{L(b_m + \sqrt{d} + \theta)}\right)^2,$$

(12)

with probability at least $1 - 2e^{-\frac{\alpha^2}{2}}$.

From Lemma 2 we can see that when the conservative step size is used, the sketch size we need to ensure $\alpha < 1$ is only $m \gtrsim W^2$.

Lemma 3. (Bound on noise amplification factor $\beta$) If the entries of the sketching matrix $S$ are i.i.d drawn from Normal distribution, then:

$$\beta \leq 1 + \frac{2b_m(W + \theta)}{m} + \frac{b_m^2}{m} - 1,$$

(13)

with probability at least $1 - 2e^{-\frac{\alpha^2}{2}}$.

Proof. The proofs of Lemma 1, 2 and 3 can be found in appendix. \hfill \Box

Lemma 2 and 3 reveals that the sketch size $m$ has an impact on both the convergence speed and the noise amplification. For the convergence speed, the larger the sketch size $m$ w.r.t the Gaussian width $W$ is, the steeper this convergence can be, but on the other hand we should not choose the sketch size to be too large since each iteration will become more expensive to compute hence there exists a trade-off between sketch size and computation. The noise amplification is a decreasing function w.r.t the sketch size $m$, which means the larger sketch size we choose, the accuracy of the GPCS output will increase. We next explore these theoretical findings through numerical experiments.

5. NUMERICAL EXPERIMENTS

5.1. Synthetic constrained LS estimation example

We first test the behaviour of GPCS algorithm with various sketch sizes for sparse linear inversion task which recovers a sparse vector $x^*$ from noiseless measurements $y = Ax$. The $l_1$ norm of $x^*$ is assumed known as a prior hence we construct the constraint set $K = \{v \in \mathbb{V} : \|v\|_1 \leq \|x^*\|_1\}$.

The experiments are executed on a PC (2.6 GHz CPU, 1.6 GB RAM) with MATLAB R2015b.

### Table 1. Experiment settings. (*) $s$ denotes sparsity of the ground truth $x^*$

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>SIZE</th>
<th>(*)$s$</th>
<th>$\frac{\mu_e}{L}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYN-SLI</td>
<td>(100000, 1500)</td>
<td>50</td>
<td>$10^6$</td>
</tr>
<tr>
<td>SYN1</td>
<td>(10000, 100)</td>
<td>5</td>
<td>$10^6$</td>
</tr>
<tr>
<td>SYN2</td>
<td>(100000, 100)</td>
<td>5</td>
<td>$10^6$</td>
</tr>
<tr>
<td>SYN3</td>
<td>(200000, 100)</td>
<td>5</td>
<td>$10^6$</td>
</tr>
</tbody>
</table>

The details of the experimental setting can be found in Table 1, and meanwhile the procedure of generating the synthetic data matrix $A$ with a condition number $c$ follows [1]:

1) Generate a random matrix $A$ sized $n$ by $d$ using MATLAB command Randn.

2) Compute the Singular Value Decomposition of $A$: $A = U\Sigma V^T$ and modify the singular values $e_i = diag(\Sigma)$, by:

$$e_i = e_{i-1} \frac{\kappa}{\kappa^2},$$

(14)

to achieve the condition number of $\frac{L}{\mu_e} = 10^6$.

The performance of the Projected Gradient and fast gradient method (FISTA) for SYN-SLI is shown in figure 1. The step sizes for all the algorithms are generated by the line-search given by [14]. Although in this paper we analyze the explicit convergence speed and noise amplification for the Gaussian Sketch as a motivational theory, this type of sketches in costly to compute. In practice, instead of using directly the Gaussian sketch, people use faster sketches such as the Fast Johnson-Lindenstrauss Transform (FJLT) [15][16], the sparse JL [17] and the Count Sketch [5]. We choose to use the Count-Sketch [5] for the GPCS to speedily produce the sketch matrix $S_A$.

This experimental result confirms that for noiseless inversion $y = Ax$, the GPCS converges towards the ground truth with an exponential rate as our theory predicts, and also the best convergence is given by appropriate median sketch size choices ($m = 700, 1100$, which are some factors larger than the sparsity $s = 50$, but less than the ambient dimension $d = 1500$. The GPCS provides significant computational benefits over the full gradient methods PGD and FISTA on this large scale inversion task.

We then turn to the noisy set up for SYN-SLI $y = Ax + w$ (Fig. 2). The Gaussian noise vector $w$ satisfies $\frac{\|Ax\|_2}{\|w\|_2} = 30$. Here we find out that as our theory predicted, the GPCS converges to a vicinity of the ground truth $x^*$, meanwhile the larger the sketch size is, the more accurate the solution is. That is, the GPCS converge to an approximated solution of the Least-squares, and the approximation accuracy is determined by the noise level and the sketch size $m$. As discussed in [1], if one needs to solve the LS to machine precision in the presence of noise $\|w\|_2 > 0$, one can use the GPCS as a fast initialization step and then continue by GPHS which is based on the “iterative” sketches [8].

5.2. Computational and sketch size trade-off

We finally examine the computational cost of GPCS on the linear inversion task $y = Ax$ through different choices of sketch size on three synthetic examples (Syn1, Syn2 and Syn3). Here we test GPCS on both the Gaussian Sketch and the Count-Sketch which is more computationally efficient. We run GPCS with sketch size from 10 to 1000 until a fixed accuracy $\|x^f - x^*\|_2 \leq 10^{-3}$ is achieved, or when
the computational budget has run out. We average the results of 20 random trials.

From the experimental results (Fig. 3) we can observe a sharp trade-off phenomenon as our theory predicted. When the sketch size is too near to the intrinsic dimension, the GPCS takes a huge cost to achieve the targeted accuracy; but if we increase the sketch size by a small amount, the computational cost drops radically to a "sweet spot", then if we continue to increase the sketch size, the cost increases again since each iteration’s cost is more expensive for a large sketch size. We can also observe that the optimal choice of sketch size is not a function of the data sample size $n$, as our theory predicted. Surprisingly we see that the larger the $n$ is, the behavior of the Gaussian Sketch and Count Sketch become more similar as shown in the right hand figures in Fig 3. It would be an interesting future research direction to investigate in theory the relationship between the performance of the practical sparse embedding schemes such as the Count Sketch, and the properties of $A$, e.g. the sample size $n$, the parameter dimension $d$ and the conditioning $\frac{L}{\mu}$, the distribution of its singular values and singular vectors, etc.

6. CONCLUSION

We prove for the first time that when the original constrained Least-square problem satisfies the robust recovery condition, the exponential convergence holds for the GPCS algorithm towards a vicinity of the ground truth without any assumption on the strong convexity of the sketched Least-squares, thanks to the concentration results for the Gaussian sketch and the intrinsic dimension of the ground truth. The numerical experiments vindicate our theoretical findings such as the convergence speed and noise amplification of GPCS algorithm. We also demonstrate the huge potential of the sketched gradient method based on exploiting the low dimensional structure (such as sparsity) via sketching techniques, by comparing with the well known full gradient methods PGD and FISTA.
7. ACKNOWLEDGEMENTS

JT, MG and MD would like to acknowledge the support from H2020-MSCA-ITN Machine Sensing Training Network (MacSeNet), project 642685; EPSRC Compressed Quantitative MRI grant, number EP/M019802/1; and ERC Advanced grant, project 694888, C-SENSE, respectively. MD is also supported by a Royal Society Wolfson Research Merit Award.

8. REFERENCES


9. APPENDIX

9.1. Proof for Theorem 1

Proof:

\[ \|h_{i+1}\|_2 = \|x_{i+1} - x^*\|_2 \]

= \[ \|P_K(x_i - \eta(A^T S^T S^t Ax_i - A^T S^T S^t y)) - x^*\|_2, \]

then because of the distance preservation translation [12, Lemma 6.3], we have:

\[ \|P_{K-x^*}(x_i - x^* - \eta(A^T S^T S^t Ax_i - A^T S^T S^t y))\|_2, \]

then we apply [12, Lemma 6.4], where \( C \) is the smallest close cone containing the set \( K - x^* \):

\[ \leq \|P_C(h_i + \eta A^T S^T S^t y - \eta(A^T S^T S^t A(h_i - x^*))\|_2, \]

= \[ \|P_C((I - \eta A^T S^T S^t A)h_i + \eta A^T S^T S^t (y - Ax^*))\|_2, \]

then because of the definition of the cone-projection operator [12, Lemma 6.2] we have:

\[ \leq \{ \sup_{v \in \mathbb{C}^d} \|v^T((I - \eta A^T S^T S^t A)h_i + \eta A^T S^T S^t w)\|_2, \]

next we tidy up the supremums terms by the definition of \( \alpha(\eta, S^t A) \)

and \( \beta(S^t, A) \) in Definition 1:

\[ \leq \{ \sup_{v \in \mathbb{C}^d} \|v^T((I - \eta A^T S^T S^t A)h_i + \eta A^T S^T S^t w)\|_2, \]

+ \[ \eta \|v^T((I - \eta A^T S^T S^t A)h_i + \eta A^T S^T S^t w)\|_2 \}

\[ \leq \{ \sup_{u, v \in \mathbb{C}^d} \|v^T((I - \eta A^T S^T S^t A)h_i + \eta A^T S^T S^t w)\|_2, \]

+ \[ \eta L \|v^T((I - \eta A^T S^T S^t A)h_i + \eta A^T S^T S^t w)\|_2 \}

Then we do recursive substitution:

\[ \|h_{i+1}\|_2 \leq \alpha^t(\eta, S^t A)\|h_0\|_2 \]

+ \[ m\eta \beta(S^t, A)\|h_0\|_2 \]

\[ \leq \alpha^t(\eta, S^t A)\|h_0\|_2 + m\eta \beta(S^t, A)\|h_0\|_2, \]

9.2. Proofs for the explicit bounds

9.2.1. Proof for Lemma 1

Proof. From the results in [12, Lemma 6.8] we can have the following bounds with probability at least \( 1 - 4e^{-\frac{\alpha^2}{2}} \):

\[ \|SA(u + v)\|_2 \geq b_m\|A(u + v)\|_2 - 2\sqrt{L}(W + \theta), \]

\[ \|SA(u - v)\|_2 \leq b_m\|A(u - v)\|_2 + 2\sqrt{L}(W + \theta), \]

then we can have:

\[ \alpha \leq \frac{1}{4}(\|u + v\|_2^2 - \eta\|SA(u + v)\|_2^2 - \|u - v\|_2^2 + \eta\|SA(u - v)\|_2^2) \]

\[ \leq \frac{1}{4}(\|u + v\|_2^2 - \eta\|b_m\|\|u - v\|_2^2 - 2(W + \theta)^2) \]

\[ \leq \frac{1}{4}(\|u + v\|_2^2 - \eta\|b_m\|\|u - v\|_2^2 + 4(W + \theta)^2 - 2b_m(W + \theta)\|u + v\|_2) \]

\[ \leq \frac{1}{4}(\|u + v\|_2^2 + \eta\|b_m\|\|u - v\|_2^2 + 4(W + \theta)^2 - 2b_m(W + \theta)\|u + v\|_2) \]

\[ \leq \frac{1}{4}(\|u + v\|_2^2 + \eta\|b_m\|\|u - v\|_2^2 + 4(W + \theta)^2 + 2b_m(W + \theta)\|u - v\|_2) \]

Now we set \( \eta = \frac{1}{b_m L}, \) since \( L \geq \mu_s, \|u + v\|_2 \leq 4, \|u - v\|_2 \leq 4 \)

and \( \|u - v\|_2 + \|u + v\|_2 \leq 2\sqrt{L}, \) we have:

\[ \alpha \leq \frac{1}{4}(4(1 - \frac{\mu_s}{L}) + \frac{4(W + \theta)^2}{b_m^2}(1 - \frac{\mu_s}{L}) + 8\sqrt{L}(W + \theta)) \]

\[ = (1 - \frac{\mu_s}{L})(1 + \frac{(W + \theta)^2}{b_m^2}) + 2\sqrt{L}(W + \theta), \]

with probability at least \( 1 - 8e^{-\frac{\alpha^2}{2}} \), Thus finishes the proof.

9.3. The proof for Lemma 2

Proof. In this proof we first state two inequality from [1, Lemma 3] which is extented the from [12, Lemma 6.7]:

\[ \|SA(u + v)\|_2 \geq \sqrt{m}\|b_m - W - \theta\|u + v\|_2, \] (15)

\[ \|SA(u - v)\|_2 \leq \sqrt{L}(b_m + \sqrt{d + \theta})\|u - v\|_2, \] (16)

with probability at least \( 1 - e^{-\frac{\alpha^2}{2}} \), then we can have:

\[ \alpha \leq \frac{1}{4}(\|u + v\|_2^2 - \eta\|SA(u + v)\|_2^2 - \|u - v\|_2^2 + \eta\|SA(u - v)\|_2^2) \]

\[ \leq \frac{1}{4}(\|u + v\|_2^2 - \eta\|b_m - W - \theta\|^2\|u + v\|_2^2) \]

\[ - \|u - v\|_2^2 + \eta L(b_m + \sqrt{d + \theta})^2\|u - v\|_2^2 \]

\[ \leq \frac{1}{4}(\|u + v\|_2^2 - \eta\|b_m - W - \theta\|^2\|u + v\|_2^2) \]

\[ - \|u - v\|_2^2 + \eta L(b_m + \sqrt{d + \theta})^2\|u - v\|_2^2 \]
Let $\eta = \frac{1}{L(b_m + \sqrt{d} + \theta)^2}$ we have:

$$
\alpha \leq \frac{1}{4} \left\{ 1 - \frac{\mu_c(b_m - W - \theta)^2}{L(b_m + \sqrt{d} + \theta)^2} \right\} \| u + v \|_2^2
\leq \left\{ 1 - \frac{\mu_c(b_m - W - \theta)^2}{L(b_m + \sqrt{d} + \theta)^2} \right\}
$$

with probability at least $1 - 2e^{-\frac{\theta^2}{16}}$.

The proof for Lemma 3 is almost the same as the proof of [1, Proposition 2] hence is not needed to be included here.