A Richer Theory of Convex Constrained Optimization with Reduced Projections and Improved Rates

Tianbao Yang [1] **Qihang Lin** [1] **Lijun Zhang** [2]

Abstract

This paper focuses on convex constrained optimization problems, where the solution is subject to a convex inequality constraint. In particular, we aim at challenging problems for which both projection into the constrained domain and a linear optimization under the inequality constraint are time-consuming, which render both projected gradient methods and conditional gradient methods (a.k.a. the Frank-Wolfe algorithm) expensive. In this paper, we develop projection reduced optimization algorithms for both smooth and non-smooth optimization with improved convergence rates under a certain regularity condition of the constraint function. We first present a general theory of optimization with only one projection. Its application to smooth optimization with only one projection yields $O(1/\ )$ iteration complexity, which improves over the $O(1/\ ^2)$ iteration complexity established before for non-smooth optimization and can be further reduced under strong convexity. Then we introduce a local error bound condition and develop faster algorithms for non-strongly convex optimization at the price of a logarithmic number of projections. In particular, we achieve an iteration complexity of $\widetilde{O}(1/\ ^{2(1-\theta)})$ for non-smooth optimization and $\widetilde{O}(1/\ ^{1-\theta})$ for smooth optimization, where $\ \in (0, 1]$ appearing the local error bound condition characterizes the functional local growth rate around the optimal solutions. Novel applications in solving the constrained ℓ_1 minimization problem and a positive semi-definite constrained distance metric learning problem demonstrate that the proposed algorithms achieve significant speed-up compared with previous algorithms.

[1]The University of Iowa, Iowa City, IA 52242, USA [2]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. Correspondence to: Tianbao Yang <tianbao-yang@uiowa.edu>.

This is the long version of our paper appearing in the *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

1. Introduction

In this paper, we aim at solving the following convex constrained optimization problem:

$$\min_{\in \mathbb{R}^d} \quad f(\mathbf{x}); \quad s.t. \ c(\mathbf{x}) \quad 0; \tag{1}$$

where $f(\mathbf{x})$ is a smooth or non-smooth convex function and $c(\mathbf{x})$ is a lower-semicontinuous and convex function. The problem can find applications in machine learning, signal processing, statistics, marketing optimization, and etc. For example, in distance metric learning one needs to learn a positive semi-definite (PSD) matrix such that similar examples are close to each other and dissimilar examples are far from each other (Weinberger et al., 2006; Xing et al., 2003), where the positive semi-definite constraint can be cast into a convex inequality constraint. Another example arising in compressive sensing is to minimize the ℓ_1 norm of high-dimensional vector subject to a measurement constraint (Candès & Wakin, 2008). Although general interior-point methods can be applied to solve the problem with linear convergence, they suffer from exceedingly high computational cost per-iteration. Another solution is to employ the projected gradient (PG) method (Nesterov, 2004) or the conditional gradient (CG) method (Frank & Wolfe, 1956), where the PG method needs to compute the projection into the constrained domain at each iteration and CG needs to solve a linear optimization problem under the constraint. However, for many constraints (e.g., PSD, quadratic constraints) both projection into the constrained domain and the linear optimization under the constraint are time-consuming, which restrict their capabilities to solving these problems.

Recently, there emerges a new direction towards addressing the challenge of expensive projection that is to reduce the number of projections. In the seminal paper (Mahdavi et al., 2012), the authors have proposed two algorithms with only one projection at the end of iterations for non-smooth convex and strongly convex optimization, respectively. The idea of both algorithms is to move the constraint function into the objective function and to control the violation of constraint for intermediate solutions. While their developed algorithms enjoy an optimal convergence rate for non-smooth optimization (i.e., $O(1/\ ^2)$ iteration complexity)

and a close-to-optimal convergence rate for strongly convex optimization (i.e., $\widetilde{O}(1/\epsilon)$ [1]), *there still lack of theory and algorithms with reduced projections and faster rates for smooth convex optimization and for convex optimization without strong convexity assumptions*.

In this paper, we make significant contributions by developing a richer theory of convex constrained optimization with reduced projections and faster rates. To be specific,

- we develop a general framework and theory of optimization with only one projection, where any favorable smooth or non-smooth convex optimization algorithms can be employed to solve the intermediate augmented unconstrained objective function. We discuss in full details the applicability of the proposed algorithms to problems with polyhedral, quadratic or PSD constraints.
- Applying the general theory to smooth convex optimization [2] with Nesterov's accelerated gradient methods yields an iteration complexity of $O(1/\sqrt{\epsilon})$ with only one projection. In addition, when equipped with an optimal algorithm for strongly convex optimization the general theory implies the *optimal* iteration complexity of $O(1/\sqrt{\epsilon})$ for strongly convex optimization with only one projection. For smooth and strongly convex optimization, the general theory implies an iteration complexity of $O(1/\epsilon^{\beta})$ where $\beta \geq 2(1/2, 1)$ with only one projection and a sufficiently large number of iterations.
- Building on the general framework and theory, we further develop an improved theory with faster convergence rates for non-strongly convex optimization at the price of a logarithmic number of projections. In particular, we show that under a mild local error bound condition, the iteration complexities can be reduced to $\widetilde{O}(1/\epsilon^{2(1-\theta)})$ for non-smooth optimization and $\widetilde{O}(1/\epsilon^{1-\theta})$ for smooth optimization, where $\theta \in (0, 1]$ is a constant in the local error bound condition that characterizes the local growth rate of functional values. To our knowledge, these are the best convergence results with only a logarithmic number of projections for non-strongly convex optimization. We also demonstrate their effectiveness for solving compressive sensing and distance metric learning problems.

2. Related Work

The issue of high projection cost in projected gradient descent has received increasing attention in recent years. Most studies are based on the Frank-Wolfe technique that eschews the projection in favor of a linear optimization over the constrained domain (Jaggi, 2013; Hazan & Kale, 2012; Lacoste-Julien et al., 2013; Garber & Hazan, 2015). It happens that for many bounded domains (e.g., bounded balls for vectors and matrices, a PSD constraint with a bounded trace norm) the linear optimization over the constrained domain is much cheaper than projection into the constrained domain (Jaggi, 2013). However, there still exist many constraints that render both projection into the constrained domain and linear optimization under the constraint are comparably expensive. Examples include polyhedral constraints, quadratic constraints and a PSD constraint [3].

To tackle these complex constraints, the idea of optimization with a reduced number of projections was explored in several studies since (Mahdavi et al., 2012). In a recent paper (Chen et al., 2016), the authors show that for stochastic strongly convex optimization, the optimal convergence rate can be achieved using a logarithmic number of projections. In contrast, the developed theory in this paper implies that only one projection is sufficient to achieve the optimal convergence rate for strongly convex optimization, and a logarithmic number of projections can be used to accelerate convergence rates for non-strongly convex optimization. Cotter et al. (2016) proposed a stochastic algorithm for solving heavily constrained problems with many constraint functions by extending the work of (Mahdavi et al., 2012). Nonetheless, their focus is not to improve the convergence rates. Zhang et al. (2013) studied the smooth and strongly convex optimization and they proposed a stochastic algorithm with $O(\kappa \log(T))$ projections and proved an $O(1/T)$ convergence rate, where κ is the condition number and T is the total number of iterations. Nonetheless, if the condition number is high the number of projections could be very large. In addition, their algorithm utilizes the mini-batch to avoid frequent projections in stochastic optimization, which is different from the present paper.

We note that several recent works also exploit different forms of error bound conditions to improve the convergence (Wang & Lin, 2014; So, 2013; Hou et al., 2013; Zhou et al., 2015; Yang & Lin, 2016; Xu et al., 2016). Most notably, the technique used in our work is closely related to (Yang & Lin, 2016). However, for constrained optimization problems the methods in (Yang & Lin, 2016) still need to conduct projections at each iteration.

Finally, we comment on the differences between the proposed methods and the classical penalty methods that also move the constraint into the objective using a penalty function (Bertsekas, 1996). The major differences are that (i) the classical penalty methods typically require solving each subproblem exactly while our methods do not require that; and (ii) the classical penalty methods typically guarantee asymptotic convergence while our methods have explicit convergence rates.

[1] where $\widetilde{O}()$ suppresses a logarithmic factor.

[2] where the constraint function is assumed to be smooth.

[3] Indeed, a linear optimization over a PSD constraint is ill-posed because the PSD domain is unbounded.

3. Preliminaries

Let $= f\mathbf{x} \in \mathbb{R}^d : c(\mathbf{x}) \quad 0g$ denote the constrained domain, $_*$ denote the optimal solution set and f_* denote the optimal objective value. We denote by $r f(\mathbf{x})$ the gradient and by $@f(\mathbf{x})$ the subgradient of a smooth or non-smooth function, respectively. When $f(\mathbf{x})$ is a non-smooth function, we consider the problem as non-smooth constrained optimization. When both $f(\mathbf{x})$ and $c(\mathbf{x})$ are smooth, we consider the problem as smooth constrained optimization. A function $f(\mathbf{x})$ is L-smooth if it has a Lipschitz continuous gradient, i.e., $k r f(\mathbf{x}) \quad r f(\mathbf{y})k \quad Lk\mathbf{x} \quad \mathbf{y}k$, where $k \quad k$ denotes the Euclidean norm. A function $f(\mathbf{x})$ is - strongly convex if it satisfies $f(\mathbf{x}) \quad f(\mathbf{y}) + @f(\mathbf{y})^\top(\mathbf{x} \quad \mathbf{y}) + \frac{\mu}{2}k\mathbf{x} \quad \mathbf{y}k^2$.

In the sequel, $dist(\mathbf{x}; \ )$ denotes the distance of $\mathbf{x}$ to a set , i.e., $dist(\mathbf{x}; \ ) = \min_{\in\Omega} k\mathbf{x} \quad \mathbf{u}k$. Let $[s]_+$ be a hinge operator that is defined as $[s]_+ = s$ if $s \quad 0$, and $[s]_+ = 0$ if $s < 0$.

Throughout the paper, we make the the following assumptions to facilitate the development of our algorithms and theory.

Assumption 1. *For a convex minimization problem (1), we assume (i) there exists a positive value* > 0 *such that*

$$\min_{\substack{c(\mathbf{x})=0 \\ \mathbf{v}\in@c(\mathbf{x});\mathbf{v}\neq 0}} k\mathbf{v}k \quad ; \tag{2}$$

or more generally there exists a constant > 0 *for any* $\mathbf{x} \in$

Remark: It is worth mentioning that we omit some constant factors in the convergence bound $B_T(\cdot;x;x_1)$ that are irrelevant to our discussions. The notation $B_T(\cdot;x;x_1)$ emphasizes that it is a function of $\cdot$ and depends on x_1 and a target solution x and it will be referred to as B_T. In the next several subsections, we will see that by carefully choosing the penalty function $h(x)$ we are able to provide nice convergence for smooth and non-smooth optimization with only one projection. In the above theorem, we assume the optimization algorithm $\mathcal{A}$ is deterministic. However, a similar result can be easily extended to a stochastic optimization algorithm $\mathcal{A}$.

Proof. First, we consider $c(\hat{x}_T) \le 0$, which implies that $\hat{x}_T = \tilde{x}_T$. Due to the certificate of $h(x)$, $F(\tilde{x}_T) \le f(\tilde{x}_T)$ and $F(x) \le f(x) + C$. Hence $f(\tilde{x}_T) - F(\hat{x}_T) \le F(x) + B_T(\cdot;x_1;x) - f(x) + C + B_T(\cdot;x_1;x)$. Then (7) follows due to $= (G)$ 1. Next, we assume $c(\hat{x}_T) > 0$. Inequality (6) implies that

$$f(\hat{x}_T) + [c(\hat{x}_T)]_+ \le f(x) + C + B_T(\cdot;x;x_1): \quad (8)$$

By Assumption 1(i), we have $[c(\hat{x}_T)]_+ \ge \|\hat{x}_T - \tilde{x}_T\|$. Combined with (8) we have

$$\begin{aligned}\|\hat{x}_T - \tilde{x}_T\| &\le f(x) - f(\hat{x}_T) + C + B_T(\cdot;x;x_1)\\ &\le G\|\hat{x}_T - \tilde{x}_T\| + C + B_T(\cdot;x;x_1);\end{aligned}$$

where the last inequality follows that fact $f(x) - f(\hat{x}_T) \le f(x) - f(\tilde{x}_T) + f(\tilde{x}_T) - f(\hat{x}_T) \le G\|\hat{x}_T - \tilde{x}_T\|$ because the Lipschitz property and $f(x) \le f(\tilde{x}_T)$. Therefore we have

$$\|\hat{x}_T - \tilde{x}_T\| \le \frac{C + B_T(\cdot;x;x_1;)}{G}:$$

Finally, we obtain

$$\begin{aligned}f(\tilde{x}_T) - f(x) &\le f(\tilde{x}_T) - f(\hat{x}_T) + f(\hat{x}_T) - f(x)\\ &\le G\|\hat{x}_T - \tilde{x}_T\| + C + B_T(\cdot;x;x_1)\\ &\le \frac{}{G}(C + B_T(\cdot;x;x_1)):\end{aligned}$$

□

4.1. Non-smooth Optimization

Since an optimal convergence rate for general non-smooth optimization with only one projection has been attained in (Mahdavi et al., 2012), in this subsection we present an optimal convergence result for strongly convex problems. For non-smooth optimization, we can choose

$$h(x) = [c(x)]_+;$$

and hence $= 0$. We will use deterministic subgradient descent as an example to demonstrate the convergence for $f(x)$, though many other optimization algorithms designed for non-smooth optimization are applicable (e.g., the stochastic subgradient method). The update of subgradient descent method is given by the following

$$x_{t+1} = x_t - {}_t \partial F(x_t); \quad t = 1;:::;T; \quad (9)$$

where ${}_t$ is an appropriate step size. If $f(x)$ is -strongly convex, the step size can be set as $= 1=(t)$ and the final solution can be computed by the -suffix averaging $\hat{x}_T = \frac{1}{T} \sum_{t=(1-)T+1}^{T} x_t$ with > 0 (Rakhlin et al., 2012), or by the polynomial decay averaging with $\hat{x}_t = (1 - \frac{s+1}{s+t})\hat{x}_{t-1} + \frac{s+1}{s+t} x_t$ and $s \ge 1$ (Shamir & Zhang, 2013). Both schemes can attain $B_T = O(1=(T))$ for the convergence of $F(x)$ when $f(x)$ is -strongly convex. Combining this with Theorem 1, we have the following convergence result with the proof omitted due to its simplicity.

Corollary 2. Suppose that Assumption 1 holds and $f(x)$ is -strongly convex. Set $F(x) = f(x) + [c(x)]_+$ with $G = $. Let (9) run for T iterations with ${}_t = 1=(t)$. Let $\hat{x}_T$ be computed by -suffix averaging or the polynomial decay averaging. Then with only one projection $\tilde{x}_T = (\hat{x}_T)$, we achieve

$$f(\tilde{x}_T) - f \le \frac{(G + G_c)^2 O(1)}{G \quad T}:$$

Remark: We note that the $O(1=(T))$ is also achieved for strongly convex optimization in (Zhang et al., 2013; Chen et al., 2016) but with a logarithmic number of projections. In contrast, Corollary 2 implies only one projection is sufficient to achieve the optimal convergence for strongly convex optimization.

4.2. Smooth Optimization

For smooth optimization, we consider both $f(x)$ and $c(x)$ to be smooth[4]. Let the smoothness parameter of $f(x)$ and $c(x)$ be L_f and L_c, respectively. In order to ensure the augmented function $F(x)$ to be still a smooth function, we construct the following penalty function

$$h(x) = \ln(1 + \exp(-c(x)=)): \quad (10)$$

The following proposition shows that $h(x)$ is a smooth function and obeys the condition in (4).

Proposition 1. Suppose $c(x)$ is L_c-smooth and G_c-Lipschitz continuous. The penalty function in (10) is a $(L_c + \frac{^2 G_c^2}{4})$-smooth function and satisfies (i) $h(x) - [c(x)]_+$ and (ii) $h(x) \le \ln 2$, $\forall x$ such that $c(x) \le 0$.

Then $F(x)$ is a smooth function and its smoothness parameter is given by $L_F = L_f + L_c + \frac{^2 G_c^2}{4}$. Next, we will

[4] it can be extended to when $f(x)$ is non-smooth but its proximal mapping can be easily solved.

Quadratic constraint. A quadratic constraint can take the form of $\|A\mathbf{x} - \mathbf{y}\|^2 \leq \epsilon$, where $A \in \mathbb{R}^{m\times d}$ and $\mathbf{y} \in \mathbb{R}^m$. Such a constraint appears in compressive sensing (Candès & Wakin, 2008)[5], where the goal is to reconstruct a sparse high-dimensional vector $\mathbf{x}$ from a small number of noisy measurements $\mathbf{y} = A\mathbf{x} + \epsilon \in \mathbb{R}^m$ with $m \ll d$. The corresponding optimization problem is

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{x}\|_1, \quad s.t. \ \|A\mathbf{x} - \mathbf{y}\|^2 \leq \epsilon \qquad (14)$$

where $\|\epsilon\|^2$ is an upper bound on the magnitude of the noise. To check the Assumption 1(i), we note that $c(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|^2 - \epsilon$ and $\nabla c(\mathbf{x}) = A^\top(A\mathbf{x} - \mathbf{y})$. Let us consider that A has a full row rank [6] and denote by $\mathbf{v} = A\mathbf{x} - \mathbf{y}$, then on the boundary $c(\mathbf{x}) = 0$ we have $\|\mathbf{v}\| = \sqrt{\epsilon}$ and $\|A^\top\mathbf{v}\| \geq \sqrt{\epsilon}\sigma_{\min}(AA^\top)$, where $\sigma_{\min}(AA^\top) > 0$ is the minimum eigenvalue of $AA^\top \in \mathbb{R}^{m\times m}$. Therefore the Assumption 1(i) is satisfied with $\rho = \sqrt{\epsilon}\sigma_{\min}(AA^\top)$. It is notable that the projection and the linear optimization under the quadratic constraint require solving a quadratic programming problem and therefore could be expensive.

PSD constraint. A PSD constraint $X \succeq 0$ for $X \in \mathbb{R}^{d\times d}$ can be written as an inequality constraint $\lambda_{\min}(X) \geq 0$, where $\lambda_{\min}(X)$ denotes the minimum eigen-value of X. The subgradient of $c(X) = -\lambda_{\min}(X)$ when $\lambda_{\min}(X) = 0$ is given by $\text{Conv}\{\mathbf{u}\mathbf{u}^\top : \|\mathbf{u}\| = 1, X\mathbf{u} = 0\}$, i.e., the convex hull of the outer products of normalized vectors in the null space of the matrix X. In (Yang et al., 2017), we show that if the dimension of the null space of X is r with $1 \leq r \leq d$, the norm of the subgradient of $c(X)$ on the boundary $c(X) = 0$ is lower bounded by $\rho = \frac{1}{\sqrt{r}} \geq \frac{1}{\sqrt{d}}$. Finally, we note that computing a subgradient of $[c(X)]_+$ only needs to compute one eigen-vector corresponding to the smallest eigen-value. In contrast, both projection and linear optimization under a PSD constraint could be very expensive for high-dimensional problems. In particular, the projection onto a PSD domain needs to conduct a singular value decomposition. The linear optimization over a PSD cone is ill-posed due to that PSD cone is not compact (the solution is either 0 or infinity). One may add an artificial constraint on the upper bound of the eigen-values. According to (Jaggi, 2013), the time complexity for solving this linear optimization problem approximately up to an accuracy level ϵ' is $O(Nd^{1.5}\sqrt{1/\epsilon'^{2.5}})$ with N being the number of non-zeros in the gradient and ϵ' decreasing iteratively required in the Frank-Wolfe method, which could be much more expensive especially for high-dimensional problems and in later iterations than computing the first eigen-pairs at each iteration in our methods.

[5] Here we use the square constraint to make it a smooth function so that the proposed algorithms for smooth optimization are applicable by using proximal gradient mapping to handle the ℓ_1 norm.

[6] which is reasonable because $m \ll d$.

7. Applications

7.1. Compressive Sensing

We first consider a compressive sensing problem in (14). Becker et al. (2011) proposed an optimization algorithm based on the Nesterov's smoothing and the Nesterov's optimal method for the smoothed problem, known as NESTA. It needs to perform the projection into the domain $\|A\mathbf{x} - \mathbf{y}\|^2 \leq \epsilon$ at every iteration and has an iteration complexity of $O(1/\epsilon)$. In contrast, the presented algorithm with *only one projection* in Section 4.2 using Nesterov's accelerated *proximal* gradient method (Beck & Teboulle, 2009) to solve the unconstrained problem enjoys an iteration complexity of $O(1/\epsilon)$. Moreover, we present a theorem below showing that the problem (14) satisfies the local error bound condition with $\theta = 1/2$, and hence the presented LoPNAG enjoys an $\widetilde{O}(1/\sqrt{\epsilon})$ iteration complexity with only a logarithmic number of projections.

Theorem 6. *Let $f(\mathbf{x}) = \|\mathbf{x}\|_1, c(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|^2 - \epsilon$, $\mathcal{X}_*$ denote the optimal set and f_* be the optimal solution to (14). Assume that there exists $\mathbf{x}_0$ such that $\|A\mathbf{x}_0 - \mathbf{y}\|^2 < \epsilon$ and $0 \notin \partial f(\mathbf{x}_*)$. Then for any $\epsilon > 0$, $\mathbf{x} \in \mathbb{R}^d$ such that $c(\mathbf{x}) \leq 0$ and $f(\mathbf{x}) \leq f_* + \epsilon$, there exists $0 < \gamma < 1$ such that $dist(\mathbf{x}, \mathcal{X}_*) \leq (f(\mathbf{x}) - f_*)^{1/2}$. Hence, LoPNAG can have an iteration complexity of $\widetilde{O}(1/\sqrt{\epsilon})$ with only $O(\log(1/\epsilon))$ projections.*

Next, we demonstrate the effectiveness of the LoPNAG for solving the compressive sensing problem in (14) by comparing with NESTA. We generate a synthetic data for testing. In particular, we generate a random measurement matrix $A \in \mathbb{R}^{m\times d}$ with $m = 1000$ and $d = 5000$. The entries of the matrix A are generated independently with the uniform distribution over the interval $[-1, +1]$. The vector $\mathbf{x}_* \in \mathbb{R}^d$ is generated with the same distribution at 100 randomly chosen coordinates. The noise $\epsilon \in \mathbb{R}^m$ is a dense vector with independent random entries with the uniform distribution over the interval $[-\sigma, \sigma]$, where σ is the noise magnitude and is set to 0.01. Finally the vector $\mathbf{y}$ was obtained as $\mathbf{y} = A\mathbf{x}_* + \epsilon$.

We use the Matlab package of NESTA [7]. For fair comparison, we also use the projection code in the NESTA package for conducting projection. To handle the unknown smoothness parameter in the proposed algorithm, we use the backtracking technique (Beck & Teboulle, 2009). The parameter γ is initially set to 0.001 and decreased by half every 5000 iterations after a projection and the target smoothing parameter in NESTA is set to 10^{-5}. For the value of θ in LoPNAG, we tune it from its theoretical value to several smaller values and choose the one that yields the fastest convergence. We report the results in Table 7.1, which include different number of iterations, the corresponding

[7] http://statweb.stanford.edu/~candes/nesta/

Table 1. LoPNAG vs. NESTA for solving the compressive sensing problem.

LoPNAG				NESTA			
Iters - Projs	Rec. Err.	Objective	Time (s)	Iters - Projs	Rec. Err.	Objective	Time (s)
5000 - 1	0:018017	52:042878	18:04	1000 - 2000	0:137798	52:703275	48:49
10000 - 2	0:018038	52:042418	35:88	3000 - 6000	0:018669	52:050051	93:84
15000 - 3	0:018043	52:042358	53:09	5000 - 10000	0:018659	52:050046	245:23
20000 - 4	0:018043	52:042358	70:24	8000 - 16000	0:018657	52:050045	404:72
25000 - 5	0:018043	52:042358	87:32	10000 - 20000	0:018657	52:050044	501:65

Table 2. LoPGD vs. OPGD and PGD for solving the considered distance metric learning problem.

LoPGD			OPGD			PGD		
Iters - Projs	Objective	Time (h)	Iters - Projs	Objective	Time (h)	Iters - Projs	Objective	Time (h)
1000 - 1	0:0953	0:22	1000 - 1	0:1707	0:20	1000 - 1000	0:1491	7:97
2000 - 2	0:0695	0:43	2000 - 1	0:1583	0:40	2000 - 2000	0:1278	15:46
4000 - 4	0:0494	0:87	4000 - 1	0:1469	0:80	4000 - 4000	0.1072	29.39
6000 - 6	0:0428	1:33	6000 - 1	0:1398	1:22	6000 - 6000	0:0957	43:36
8000 - 8	0:0405	1:89	8000 - 1	0:1343	1:64	8000 - 8000	0:0879	57:43

number of projections, the recovery error of the found solution compared to the underlying true sparse solution, the objective value (i.e., the ℓ_1 norm of the found solution) and the running time. Note that each iteration of NESTA requires two projections because it maintains two extra sequence of solutions. From the results, we can see that LoPNAG converges significantly faster than NESTA. Even with only one projection, we are able to obtain a better solution than that of NESTA after running 10000 iterations.

7.2. High-dimensional Distance Metric Learning

Consider the following distance metric learning problem:

$$\min_{A \succeq 0} \frac{1}{2|E|} \sum_{(i,j)\in\mathcal{E}} (1 \quad y_{ij} \quad k\mathbf{x}_i \quad \mathbf{x}_j k_A^2)^2 + \quad kAk_1^{\text{off}}; \quad (15)$$

Acknowledgements

We are grateful to all anonymous reviewers for their helpful comments. T. Yang is partially supported by National Science Foundation (IIS-1463988, IIS-1545995). L. Zhang thanks the support from NSFC (61603177) and JiangsuSF (BK20160658).

References

Rockafellar, R.T. *Convex Analysis*. Princeton mathematical series. Princeton University Press, 1970.

Shamir, Ohad and Zhang, Tong. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 71–79, 2013.

So, Anthony Man-Cho. Non-asymptotic convergence analysis of inexact gradient methods for machine learning without strong convexity. *CoRR*, abs/1309.0113, 2013.

Wang, Po-Wei and Lin, Chih-Jen. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research*, 15(1):1523–1548, 2014.

Weinberger, Kilian Q., BlitzyeiJohnQ., and