

The Mixing method: coordinate descent for low-rank semidefinite programming

Po-Wei Wang

Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
poweiw@cs.cmu.edu

Wei-Cheng Chang

Language Technologies Inst
Carnegie Mellon University
Pittsburgh, PA 15213
wchang2@cs.cmu.edu

J. Zico Kolter

Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
zkolter@cs.cmu.edu

Abstract

In this paper, we propose a coordinate descent approach to low-rank structured semidefinite programming. The approach, which we call the Mixing method, is extremely simple to implement, has no free parameters, and typically attains an order of magnitude or better improvement in optimization performance over the current state of the art. We show that for certain problems, the method is strictly decreasing and guaranteed to converge to a critical point. We then apply the algorithm to three separate domains: solving the maximum cut semidefinite relaxation, solving a (novel) maximum satisfiability relaxation, and solving the GloVe word embedding optimization problem. In all settings, we demonstrate improvement over the existing state of the art along various dimensions. In total, this work substantially expands the scope and scale of problems that can be solved using semidefinite programming methods.

1 Introduction

This paper considers the solution of large-scale, structured semidefinite programming problems (SDPs). A generic semidefinite program can be written as the optimization problem

$$\begin{aligned} & \underset{X \in \mathbb{R}^{n \times n}}{\text{minimize}} && \langle C, X \rangle \\ & \text{subject to} && \langle A_i, X \rangle = b_i, \quad i = 1, \dots, m \\ & && X \succeq 0. \end{aligned} \tag{1}$$

Semidefinite programs can encode a huge range of practical problems, including relaxations of many combinatorial optimization tasks [Boyd and Vandenberghe, 2004], approximate probabilistic inference [Jordan and Wainwright, 2004], metric learning [Yang, 2006], matrix completion [Candes and Recht, 2012], and many others. Unfortunately, generic semidefinite programs involve optimizing a *matrix-valued* variable $X \in \mathbb{R}^{n \times n}$, which quickly increases the number of variables in the optimization problem to a size that is not solvable by solvers employing exact methods such as primal-dual interior point algorithms.

Fortunately, a property of these problems, which has been recognized for some time now, is that the solution to such problems is often *low-rank*; specifically, the problem always admits an optimal solution with at most rank $\lceil \sqrt{2m} \rceil$ [Barvinok, 1995, Pataki, 1998], and many SDPs are set up to often have even lower rank solutions in practice. This has motivated the development of non-convex low-rank solvers for these systems: that is, we can attempt to solve the equivalent (but now non-convex) formulation of the problem

$$\begin{aligned} & \underset{V \in \mathbb{R}^{k \times n}}{\text{minimize}} && \langle C, V^T V \rangle \\ & \text{subject to} && \langle A_i, V^T V \rangle = b_i, \quad i = 1, \dots, m. \end{aligned} \tag{2}$$

Here we are explicitly representing X by the matrix V of rank k (typically with $k \ll n$), $X = V^T V$. Note that because we are representing X in this way, we no longer need to explicitly enforce semidefiniteness, as it is implied by the change of variables. In a long series of works dating back several years, it has been shown that, somewhat surprisingly, this change to a non-convex problem does not cause as many difficulties as might be thought: in practice, local solutions to the problem tend to recover the optimal solution [Burer and Monteiro, 2003]; under some reasonable assumption all second order local optima of the problem are also global optima [Boumal et al., 2016b] (assuming sufficient rank k); and it even holds that some approximated local optima also sometimes have good approximation properties [Mei et al., 2017] for convex relaxations of some combinatorial problems.

Despite this, even solving the problem above over V often presents challenges in practice. This problem involves non-linear equality constraints, and is traditionally solved either through augmented Lagrangian methods, or via Riemannian manifold methods. While both these approaches look promising in many situations, they often suffer from slow convergence, difficulty in selecting step size, or other usual problems.

In this paper, we present an alternative approach to solving such problems: coordinate descent over columns of V . While this approach is difficult to apply to the fully generic semidefinite programming problem (since we cannot easily derive closed form expressions for the coordinate updates) for many special cases of structured semidefinite problems. Specifically, we focus primarily in this paper on the case of an SDP with unit diagonal constraints; that is, the semidefinite program

$$\underset{X \in \mathbb{R}^{n \times n}}{\text{minimize}} \langle C, X \rangle \quad \text{subject to } X \succeq 0, X_{ii} = 1. \quad (3)$$

This is clearly a very special case of the full semidefinite program, but it also captures some very important problems, such as the famous semidefinite relaxation of the maximum cut (MAXCUT) combinatorial optimization problem; indeed, the MAXCUT relaxation will be one of the primary applications in this paper. For this setting, we show that we can derive the coordinate descent updates in a very simple closed form, resulting in an algorithm several times faster than the existing state of the art. We call our approach the Mixing method, as for the MAXCUT SDP problem, the updates have a natural interpretation in terms of giving each v_i as a mixture of the remaining v_j terms. We will also show, however, that the method can be applied to other problems as well, such as a (novel, to the best of our knowledge) relaxation of the MAXSAT problem, as well as an unconstrained quadratic SDP very similar to the GloVe word embedding algorithm [Pennington et al., 2014]. From a theoretical standpoint, we show that (in the case of linear SDP objective with unit diagonal constraints) the method recovers most of the same guarantees as gradient methods for the non-convex SDP formulation: specifically, we show the Mixing method will converge to a stationary point of the non-convex problem. As a whole, our approach enables solving a number of large semidefinite programs substantially faster than the all existing approaches in this space, thereby significantly improving upon the state of the art in optimization on the applied problems.

2 Background and related work

Low-rank method for SDP problem Given a SDP problem with m constraints, it was proven by [Barvinok, 1995, Pataki, 1998] that, if the problem is solvable, it admits solutions of rank $k = \lceil \sqrt{2m} \rceil$. That is, we have solutions satisfying $X = V^T V$, such that $V \in \mathbb{R}^{k \times n}$. Thus, if we can solve the problem in the space of V , we can ignore the semidefinite constraint and have much fewer variables. The idea of using this low-rank structure during optimization method was first proposed by [Burer and Monteiro, 2003] in their solver SDPLR, in which they solve the low-rank problem with L-BFGS on the extended Lagrangian problem. Since then, many low-rank optimization algorithms have been developed. One of the most notable branch is the Riemannian trust region method introduced by [Absil et al., 2009]. They considered the Riemannian manifold of low-rank structures, and extended the non-linear conjugate gradient method to work on the manifold. Later, [Boumal and Absil, 2015] improved the method by including the preconditioned CG; these methods are implemented in the popular Manopt package [Boumal et al., 2014].

Somewhat surprisingly, all the above low-rank method are observed to converge to globally optimal solution in practice [Burer and Monteiro, 2003, Absil et al., 2009, Boumal et al., 2014]. However, to the best of the authors' knowledge, though there is not yet a general proof on convergence to the globally optimal solution without strong assumptions. Absil et al. [2009, Theorem 7.4.2] proved that their method converges to the critical point under a sufficient decrease condition, and super-linear convergence near isolated local minimizer when the Riemannian Hessian is positive definite.

However, the result do not apply to the linear objective in our problem. [Sato and Iwai \[2015\]](#) proved the convergence to critical point without assumptions for a modified Riemannian conjugate gradient method, and [\[Boumal et al., 2016a\]](#) proved that the Riemannian trust region method converges to a solution with Hessian larger than $-\epsilon I$ in $O(1/\epsilon^2)$. Further, [\[Boumal et al., 2016b, Theorem 2\]](#) proved that, for sufficiently large k , all second-order optimal solution is globally optimal for almost all cost matrix C . However, it is still not yet proven that the Riemannian trust region method converges to a second-order optimal solution. Note that, the term “global convergence” in the Riemannian optimization literature usually means converging to critical regardless of initialization rather than convergence to global optima; this point is emphasized specifically in [\[Boumal et al., 2016a\]](#).

Other results on global convergence to the optimal on low-rank do not apply to our problem setting. For example, [Bhojanapalli et al. \[2016\]](#) proved that under a certain spectral initialization, the gradient descent method converges to the global optima for unconstrained low-rank SDP. [Park et al. \[2016\]](#) further proved that the method works for norm-constrained low-rank SDP problem when the feasible space of V are convex. [Lee et al.](#) proved that, under random initialization, the gradient decent method converges to a local minimizer for unconstrained problems. However, these methods do not apply to the unit diagonal constraint because it is constrained and the feasible space is not convex.

Approximation algorithm for MAXCUT and MAXSAT Semidefinite programming has many applications in approximation algorithm for NP-complete problems. In particular, [\[Goemans and Williamson, 1995\]](#) proposed a classical SDP relaxation on the MAXCUT and MAX-2SAT problems, which has 0.878 approximation guarantee. Experiments on max-2-sat [\[Gomes et al., 2006\]](#) shows that the SDP upper bound and lower bound are much tighter than the classical linear programming relaxation for MAXSAT [\[Goemans and Williamson, 1994\]](#). However, traditionally the SDP is also more expensive to solve than linear programming, though we will show here that with our approach, SDP relaxations can achieve substantially better results than linear programming in less time.

The word embedding problem The word embedding is a feature learning technique used to embed the meaning of words as low-dimension vectors. One of the most popular embedding is Word2vec [\[Mikolov et al., 2013b,a\]](#), in which the authors trained a shallow neural network to obtain the embeddings, and show that the embedding can success model similarities and analogies between words. Another popular model, GloVe [\[Pennington et al., 2014\]](#), uses factorization-based formulation instead and is shown to have better accuracy in analogies tasks compared with Word2vec. The theoretical justification of these two models are discussed in RANDWALK [\[Arora et al., 2015\]](#). Here we will show that our coordinate descent approach can also be applied to learning word embedding with the GloVe objective, highlighting the fact that convex methods can be applied to problems typically considered the domain solely of non-convex optimization.

3 The mixing method

As mentioned above, the goal of the basic mixing method is to solve a semidefinite program in (3) with a unit diagonal constraint. As discussed, we can replace the $X \succeq 0$ constraint with $X = V^T V$ for some $V \in \mathbb{R}^{k \times n}$; when we do this, the constraint that $X_{ii} = 1$ translates to the constraint that $\|v_i\| = 1$, for v_i the i th column of V . This leads to the equivalent (nonconvex) optimization problem

$$\underset{V \in \mathbb{R}^{k \times n}}{\text{minimize}} \langle C, V^T V \rangle \quad \text{subject to } \|v_i\| = 1, \quad i = 1, \dots, n. \quad (4)$$

Although this the problem is nonconvex, it is known [\[Barvinok, 1995, Pataki, 1998\]](#) that for the case the rank of V larger than $\sqrt{2n}$, the optimal solution for V can recover the optimal solution of X .

Now let us consider how we can solve the problem (4) via a coordinate descent method. The resulting algorithm is extremely simple to implement, and as we will show, it performs substantially better than existing approaches for the semidefinite problems of interest. Specifically, the objective terms that depend on v_i are given by $v_i^T (\sum_{j=1}^n C_{ij} v_j)$. However, because $\|v_i\| = 1$ we can assume that $C_{ii} = 0$ without affecting the solution of the optimization problem. Thus, the problem is equivalent to simply minimizing the inner product $v_i^T g$ (where g is given by the above sum), subject to the constraint that $\|v_i\| = 1$; this problem has a closed form solution, simply given by $v_i = -g/\|g\|$. Put in terms of the original v_j variable, this is simply the update

$$v_i^{\text{next}} := \text{normalize} \left(- \sum_{j=1}^n C_{ij} v_j \right). \quad (5)$$

Algorithm 1: The Mixing method

```

1 Initialize  $v_i$  randomly on a unit sphere;
2 while not yet converged do
3   for  $i = 1, \dots, n$  do
4      $v_i := \text{normalize}(-\sum_{j=1}^n C_{ij}v_j)$ ;
5   end
6 end

```

This way, we can initialize v_i on unit sphere and perform cyclic update over all the $i = 1, \dots, n$ in closed-form. We called it the mixing method, because for each v_i it mixes and normalizes the remaining vectors v_j according to weight C_{ij} . Thus, in the case of sparse C (which is the normal case for any large data problem) the time complexity for updating all variable once is $O(k\#\text{nnz})$, which is significantly cheaper than the interior point method. However, the details

for efficient computation differ depending on the precise nature of the SDP, so we will describe these in more detail in the subsequent application sections. A complete description of the generic algorithm is shown in Algorithm 1.

3.1 Convergence properties

Despite it's simplicity, the Mixing method has a number of desirable properties. Specifically, we show the following results under virtually all conditions.

Theorem 1. *The mixing method is strictly decreasing and always converges to a critical point.*

Before starting the proof, we formulate the Mixing method in a slightly different manner. Let L be the strictly lower triangular part of C so that $C = L + L^T$ (because we can assume $C_{ii} = 0$). Then the update of the Mixing method can be written as

$$-V^r L = V^{r+1} (L^T + \text{diag}(y^{r+1})), \quad (6)$$

in which V^r is the solution at the r -th iteration and

$$y_i^{r+1} = \left\| \sum_{j < i} C_{ij} v_j^{r+1} + \sum_{j > i} C_{ij} v_j^r \right\|. \quad (7)$$

This formulation is similar to the analysis of block coordinate descent in [Sun and Hong, 2015]. Specifically, note that y^{r+1} is not a constant and thus the evolution is not linear. Further, we show that our method admit sufficient decrease for every cycle.

Lemma 2. *The function difference for each cycle of the Mixing method is*

$$\langle C, X^r \rangle - \langle C, X^{r+1} \rangle = \sum_{i=1}^n y_i^{r+1} \|v_i^r - v_i^{r+1}\|^2. \quad (8)$$

Proof. Left-multiplying (6) by V^{rT} and V^{r+1T} and taking the difference between these two equations (using the fact that $X^r = V^{rT} V^r$), we have

$$X^r L - X^{r+1} L = V^{r+1T} V^r L - V^{rT} V^{r+1} L^T + (X^{r+1} - V^{rT} V^{r+1}) \text{diag}(y^{r+1}). \quad (9)$$

Because

$$\text{tr}(XL) = \frac{1}{2} \text{tr}(XC) \quad \text{and} \quad \text{tr}(V^{r+1T} V^r L) = \text{tr}(V^{rT} V^{r+1} L^T),$$

taking trace on (9) gives

$$\frac{1}{2} (\text{tr}(CX^r) - \text{tr}(CX^{r+1})) = 0 + \sum_i y_i^{r+1} (1 - v_i^{rT} v_i^{r+1}). \quad (10)$$

The result follows from the fact that $1 - v_i^{rT} v_i^{r+1} = \frac{1}{2} \|v_i^r - v_i^{r+1}\|^2$. \square

Proof. (of Theorem 1) Lemma 2 means that the Mixing method admits a unique limit point. Let the limit be \bar{V} and the corresponding limit of y^r be \bar{y} . Then \bar{V} being a fix point of (6) implies

$$\bar{V}(C + \text{diag}(\bar{y})) = 0, \quad (11)$$

which also means

$$\bar{X}(C + \text{diag}(\bar{y})) = 0 \quad (12)$$

if we let $\bar{X} = \bar{V}^\top \bar{V}$. Remember that the KKT condition of (3) is

$$X^* \succeq 0, \quad X_{ii}^* = 1 \quad \text{primal feasibility} \quad (13)$$

$$X^*(C + \text{diag}(y^*)) = 0 \quad \text{complementary slackness} \quad (14)$$

$$C + \text{diag}(y^*) \succeq 0 \quad \text{dual feasibility.} \quad (15)$$

Thus, together with the feasibility of the Mixing method, the limit \bar{V} satisfies (13) and (14). \square

Finally, we note that although it is not yet proven, for cases where we can compute the globally optimal solution of the SDP, we observe that 1) the Mixing method always converges to this solution, provided the rank $k \geq \sqrt{2n}$; and 2) it converges to this global optimum at a linear rate. Proving either of these properties for *any* low-rank semidefinite programming approach remains an open problem (though this behavior has been observed repeatedly), so it is not surprising that a formal proof of this fact remains elusive. Thus, the method in practice seems to give the exact solutions of the semidefinite programming problem, while being many orders of magnitude faster.

4 Application: Maximum cut problem

The SDP MAXCUT relaxation is indeed the motivating example of the Mixing method, so we consider it first. In this section, we demonstrate how to apply our method to this problem, which originated from [Goemans and Williamson, 1995].

Problem description. The maximum cut problem is an NP-hard binary optimization problem, which seek a partition over a set of vertex $i = 1, \dots, n$, so that the sum of edge weight C_{ij} across the partition is maximized. If we denote the two partition as ± 1 , we can formulate the assignment v_i of vertex i as the following binary optimization problem

$$\underset{v_i \in \{\pm 1\}, \forall i}{\text{maximize}} \quad \frac{1}{2} \sum_{ij} C_{ij} \left(\frac{1 - v_i v_j}{2} \right). \quad (16)$$

Goemans and Williamson [1995] proposed that we can approximate the above solution by “lifting” the assignment to a unit sphere in \mathbb{R}^k for sufficient large k

$$\underset{v_i \in \mathbb{R}^k, \forall i}{\text{maximize}} \quad \frac{1}{2} \sum_{ij} C_{ij} \left(\frac{1 - v_i^\top v_j}{2} \right). \quad (17)$$

To recover the binary assignment, we can do a randomized rounding by picking a random vector $r \in \mathbb{R}^k$ on the unit sphere, and let the binary assignment of vertex i be $\text{sign}(r^\top v_i)$. The analysis shows that the approximation ratio for the NP-hard problem is 0.878, which means that the expected objective from the randomized rounding scheme is at least 0.878 times the optimal binary objective.

Because the problem can be solve by the unit diagonal SDP (3), we can apply the Mixing method directly, as presented in Algorithm 1. Further, for sparse adjacency matrix C , the coefficient $\sum_j C_{ij} v_j$ can be constructed in time proportional to the nonzeros in column i of C . Thus, the time complexity of running a round of update for all v_i is $O(k \cdot \#\text{edges})$, in which k is at most $\sqrt{2n}$.

Results Figure 1 shows the results of running the Mixing method on several instances of benchmark MAXCUT problems. These range in size from approximately 1000 nodes and 20000 edges to approximately 2 million nodes and 3 million edges. For this application, we are largely concerned with evaluating the runtime of our Mixing method versus another approaches for solving this same semidefinite program. Specifically, we compare to DSDP [Benson and Ye, 2005], an interior point methods; SDPLR [Burer and Monteiro, 2003], one of the first approaches to exploit low-rank structure; and Manopt [Boumal et al., 2014], a recent toolkit for optimization on Riemannian manifolds, with special solvers dedicated specifically to the MAXCUT problem. As the results show, in all cases the Mixing method is substantially faster than other approaches: for reaching modest accuracy (defined as 10^{-4} times the difference between the initial and optimal value), we are typically 10-100x faster than all competing approach; only the Manopt algorithm ever surpasses our approach, and this happens only once both methods have achieved very high accuracy. Crucially, on the largest problems, we remain about 10x (or more) faster than Manopt over the entire run. This allows the Mixing method to scale to substantially larger problems in less time. While only 3 illustrative cases are shown in the figure, additional results (which all look similar) are included in the appendix.

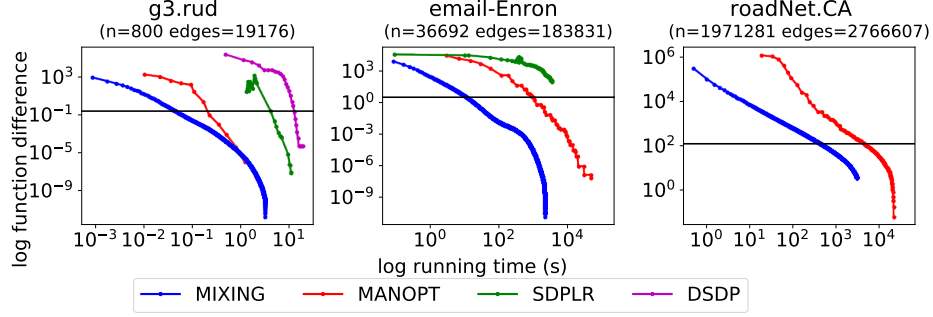


Figure 1: Relative objective value difference versus training time. The horizontal lines mark $1e^{-4}$ times the starting relative objective of the mixing method.

5 Application: Maximum satisfiability problem

Using similar ideas as in the previous section, [Goemans and Williamson, 1995] proposed that we can use SDP to solve the maximum 2-satisfiability problem. In this section, we propose a novel formulation that generalize this idea to general maximum satisfiability problem. The relaxation we propose here is novel, to the best of our knowledge, and, as we will show, achieves substantially better approximation results than existing relaxations.

Problem description. The MAXSAT problem is an extension of the well-known satisfiability problem, where the goal is to find an assignment that *maximizes* the number of satisfied clauses. Letting $v_i \in \{\pm 1\}$ be a binary variable and $s_{ij} \in \{-1, 0, 1\}$ be the sign of variable i in clause j . The goal of MAXSAT can then be written as the optimization problem

$$\underset{v \in \{-1, 1\}^n}{\text{maximize}} \quad \sum_{j=1}^m \bigvee_{i=1}^n \mathbf{1}\{s_{ij}v_i > 0\} \quad (18)$$

Note that most clauses will contain relatively few variables, so the s_j vectors will be sparse. To avoid the need for an additional bias term, we introduce an auxiliary “truth” variable v_0 , and define $z_j = \sum_{i=1}^n s_{ij}v_i - 1 = \sum_{i=0}^n s_{ij}v_i = Vs_j$. Then the MAXSAT problem can be approximated as:

$$\underset{v \in \{-1, 1\}^n}{\text{maximize}} \quad \sum_{j=1}^m 1 - \frac{\|Vs_j\|^2 - (|s_j| - 1)}{(|s_j| + 1)^2 - (|s_j| - 1)} \quad (19)$$

Although we will not derive it formally, for any v_i , this term represents an upper bound on the exact MAXSAT solution. We can now perform a similar trick to the MAXCUT SDP, and relax the v_i s to be vectors in \mathbb{R}^k with $\|v_i\| = 1$. This leads to the full MAXSAT semidefinite programming relaxation

$$\underset{X \succeq 0}{\text{minimize}} \quad \langle C, X \rangle, \quad \text{subject to } C = \sum_{j=1}^m w_j s_j s_j^T, \quad X_{ii} = 1, \quad i = 0, \dots, n. \quad (20)$$

where $w_j = 1/((|s_j| + 1)^2 - (|s_j| - 1))$.

Application of the Mixing method To apply the mixing method, need a slightly more involved approach than for MAXCUT, but the algorithm is still extremely simple. Specifically, we maintain $z_j = Vs_j$ for all clause j . Because in each subproblem only one variable v_i is changed, the z_j can be maintained in $O(k|s_{i\cdot}|)$ (where $|s_{i\cdot}|$ denotes the number of clauses that contain variable i). To sum up, we have time complexity $O(k\|S\|_1)$ for updating over all the variables in V once. Also, because applying arbitrary rotation $R \in \mathbb{R}^{k \times k}$ to V doesn’t change the objective value of our problem, we can avoid updating v_0 . See Algorithm 2 for complete algorithm. To recover the binary assignment, we consider the following classic rounding scheme: sample a random vector r from a unit sphere, then assign binary variable i as true if $\text{sign}(r^T v_i) = \text{sign}(r^T v_0)$ and false otherwise.

Unlike in the previous section (where the focus was solely on optimization performance), in this section we highlight the fact that with the Mixing method we are able to obtain MAXSAT results with high approximation ratio on challenging domains (as the problems are similar, relative optimization performance is similar to that of the MAXCUT evaluations). Specifically, we evaluate solving MAXSAT examples from the 2016 MaxSAT competition [Argelich et al., 2016].

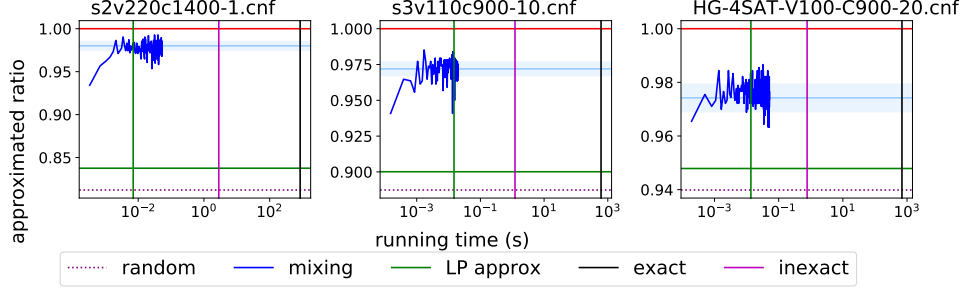


Figure 2: Approximated ratio versus training time

Algorithm 2: The Mixing method for MAXSAT problem

```

1 Initialize all  $v_i$  randomly on a unit sphere;
2 Let  $z_j = \sum_{i=0}^n s_{ij}v_i$  for  $j = 1, \dots, m$ ;
3 while not yet converged do
4   for  $i = 1, \dots, n$  do
5     foreach  $s_{ij} \neq 0$  do  $z_j := z_j - s_{ij}v_i$ ;
6      $v_i := \text{normalize} \left( - \sum_{j=1}^m \frac{s_{ij}}{(|s_j|+1)^2 - (|s_j|-1)} z_j \right)$ ;
7     foreach  $s_{ij} \neq 0$  do  $z_j := z_j + s_{ij}v_i$ ;
8   end
9 end

```

Out of 525 problems (every problem solved exactly by some solver within 30 minutes during the competition), our method achieved as average approximation ratio of 0.978, and usually found such solution within seconds or less. Figure 2 shows the progress of our approximate method versus the time for the exact solution as well as the well-known linear programming approximation (solved via the Gurobi solver). In all cases we

are able to converge to a reasonable solution after just a few passes over the variables.

6 Application: Word embedding problem

Finally, we end this paper with an application of the Mixing method to a different type of problem: an SDP with a quadratic objective and no constraint. Specifically, in this section we will discuss an application to a word embedding problem. Notably, we emphasize that these results demonstrate the potential for convex methods to apply to domains typically considered the realm of purely local non-convex optimization.

Problem description. A word embedding problem seek to represent the meaning of a word as a real-valued vector. It is ubiquitous in NLP and information retrieval. Specifically, let C_{ij} be the number of times word i and j co-occur within the same window in the corpus. We consider solving a slightly modified version of the GloVe objective function [Pennington et al., 2014]

$$\min_{V \in \mathbb{R}^{k \times n}} \frac{1}{2} \sum_{i \neq j} w_{ij} \left(v_i^T v_j + b_i + b_j - \log C_{ij} \right)^2, \quad (21)$$

where n is number of vocabulary, k is number of latent factor and $w_{ij} = \min\{C_{ij}^{3/4}, 100\}$ is a tuning factor to suppress the high-frequency words. The only difference with GloVe is that we do not include the self-loop, i.e., $i = j$ terms, in the formulation.

Application of Mixing method If we focus on one v_i at a time, we can see that the subproblem

$$\min_{d \in \mathbb{R}^k} f(v_i + d) = \frac{1}{2} \sum_j w_{ij} \left((v_i + d)^T v_j + b_i + b_j - \log C_{ij} \right)^2 \quad (22)$$

is a unconstrained quadratic program for direction $d \in \mathbb{R}^k$. After expansion, the above subproblem is equivalent to

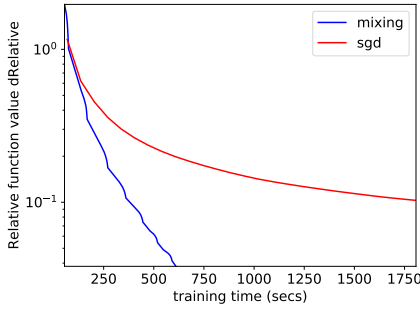
$$\min_{d \in \mathbb{R}^k} \frac{1}{2} d^T \left(\sum_j^H w_{ij} v_j v_j^T \right) d + \left(\sum_j^g e_{ij} w_{ij} v_j \right)^T d,$$

Algorithm 3: The Mixing method for Word embedding problem

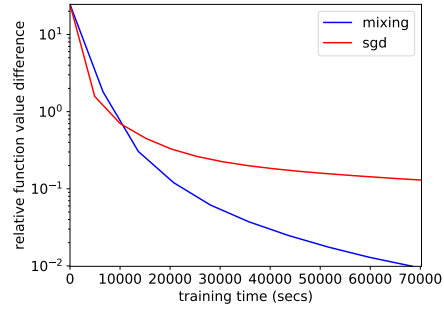
```

1 Initialize  $v_i$  randomly on a unit sphere;
2 Initialize  $b_i := 0$  for each  $i = 1, \dots, m$ ;
3 while not yet converged do
4   for  $i = 1, \dots, n$  do
5     Solve  $Hd + g = 0$  approximately by conjugate gradient method using
        $Hd := \sum_j w_{ij}(v_j^T d)v_j$  and  $g := \sum_j w_{ij}e_{ij}v_j$ , where
        $e_{ij} = v_i^T v_j + b_i + b_j - \log(C_{ij})$ ;
6     Let  $v_i := v_i + d$ ;
7     Update bias term  $b_i := b_i - (\sum_j w_{ij}e_{ij})/(\sum_j w_{ij}^2)$ ;
8   end
9 end

```



(a) wiki8 ($n=1332$, $\text{nnz}=1678046$, $k=50$)



(b) enwiki ($n=75317$, $\text{nnz}=875447516$, $k=300$)

Figure 3: $(f - f_{\min})$ v.s. training time, where f_{\min} is minimum objective value we have.

where $e_{ij} = v_i^T v_j + b_i + b_j - \log C_{ij}$. Although a closed-form solution exact method in $O(n^3)$, in practice, we apply the conjugate gradient method with stopping condition on $\|\nabla_d f(v + d)\|$ to get a good enough solution for each of the subproblem and cyclically update all the v_i . Each round of update through all the variables take $O(k \cdot \# \text{nnz} \cdot (\# \text{CG iteration}))$, in which $(\# \text{CG iteration})$ is typically below 10 in our settings. Algorithm 3 contains a complete description.

Results. Figure 3a shows the result of comparing the proposed mixing method with the stochastic gradient method, which is the default solver for GloVe. We consider the wiki8 and enwiki datasets, well-known benchmarks for word embedding. The corpus is pre-processed by standard approach [Pennington et al., 2014] (removing non-textual elements, sentence splitting, and tokenization), and words that appeared less than 1000 times in the corpus are ignored. Figure 3a shows the results of the Mixing method versus SGD on our two corpora. For both datasets, Mixing converges substantially faster to a lower function objective value than the SGD method.

7 Conclusion

In this paper we have presented the Mixing method, a low-rank coordinate descent approach for solving certain classes of structured semidefinite programming problems. The algorithm is extremely simple to implement, and involves no free parameters such as learning rates. We have proved that the algorithm converges everywhere to a critical point of the low-rank optimization problem. And we have demonstrated the method on three different application domains: the MAXCUT SDP, a (novel) MAXSAT relaxation, and a word embedding problem. In all cases we show positive results, that the method performs much faster than existing approaches from an optimization standpoint (for MAXCUT and word embeddings), and that the resulting solution have high quality from an application perspective (for MAXSAT). In total, this substantially raises the bar as to what applications can be feasibly addressed using semidefinite programming, and also advances the state of the art in structured low-rank optimization.

References

- P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.
- Josep Argelich, Chu Min Li, Felip Manyà, and Jordi Planes. Max-sat-2016 eleventh max-sat evaluation. <http://http://maxsat.ia.udl.cat/>, 2016.
- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Rand-walk: A latent variable model approach to word embeddings. *arXiv preprint arXiv:1502.03520*, 2015.
- Alexander I. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(2):189–202, 1995.
- Steven J. Benson and Yinyu Ye. DSDP5: Software for semidefinite programming. Technical Report ANL/MCS-P1289-0905, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, September 2005. URL <http://www.mcs.anl.gov/~benson/dsdp>. Submitted to ACM Transactions on Mathematical Software.
- Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582, 2016.
- Nicolas Boumal and P-A Absil. Low-rank matrix completion via preconditioned optimization on the grassmann manifold. *Linear Algebra and its Applications*, 475:200–239, 2015.
- Nicolas Boumal, Bamdev Mishra, Pierre-Antoine Absil, Rodolphe Sepulchre, et al. Manopt, a matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(1):1455–1459, 2014.
- Nicolas Boumal, P-A Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *arXiv preprint arXiv:1605.08101*, 2016a.
- Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765, 2016b.
- Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.
- Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.
- Michel X Goemans and David P Williamson. New 3/4-approximation algorithms for the maximum satisfiability problem. *SIAM Journal on Discrete Mathematics*, 7(4):656–666, 1994.
- Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6): 1115–1145, 1995.
- Carla P Gomes, Willem-Jan van Hoeve, and Lucian Leahu. The power of semidefinite programming relaxations for max-sat. In *International Conference on Integration of Artificial Intelligence (AI) and Operations Research (OR) Techniques in Constraint Programming*, pages 104–118. Springer, 2006.
- Christoph Helmberg and Franz Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696, 2000.
- Michael I Jordan and Martin J Wainwright. Semidefinite relaxations for approximate inference on graphs with cycles. In *Advances in Neural Information Processing Systems*, pages 369–376, 2004.
- Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers.

- Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.
- Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.
- Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.
- Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto I Oliveira. Solving sdps for synchronization and maxcut problems via the grothendieck inequality. *arXiv preprint arXiv:1703.08729*, 2017.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.
- Dohyung Park, Anastasios Kyrillidis, Srinadh Bhojanapalli, Constantine Caramanis, and Sujay Sanghavi. Provable burer-monteiro factorization for a class of norm-constrained matrix problems. *arXiv preprint arXiv:1606.01316*, 2016.
- Gábor Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research*, 23(2):339–358, 1998.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.
- Hiroyuki Sato and Toshihiro Iwai. A new, globally convergent riemannian conjugate gradient method. *Optimization*, 64(4):1011–1031, 2015.
- Ruoyu Sun and Mingyi Hong. Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. In *Advances in Neural Information Processing Systems*, pages 1306–1314, 2015.
- Liu Yang. Distance metric learning: A comprehensive survey. 2006.

A Full experiment result for MAXCUT problem

For completeness, we here include timing results for all MAXCUT optimization problems. These problems are taken from GSET [Helmberg and Rendl, 2000], which is a suite of standard benchmark problems for the MAXCUT problem. The horizontal line in the figures marks the stopping threshold for the Mixing method, which equals 10^{-4} times the initial difference to the optimal objective value. In the smaller GSET dataset, the Mixing method is 7.53x faster than Manopt, 43.62x faster than SDPLR, and 361x faster than DSDP on the time to reach the horizontal line. In the large dataset (email-Enron [Leskovec et al., 2005], ca-HepPh [Leskovec et al., 2007], and roadNet-CA [Leskovec et al., 2009]), the Mixing method is 40.01x faster than Manopt on all three data, and 114.83x faster than SDPLR when it is runnable. Some solvers are not presented in the experiments of large dataset because they crashed or didn't produce any iteration after an hour.

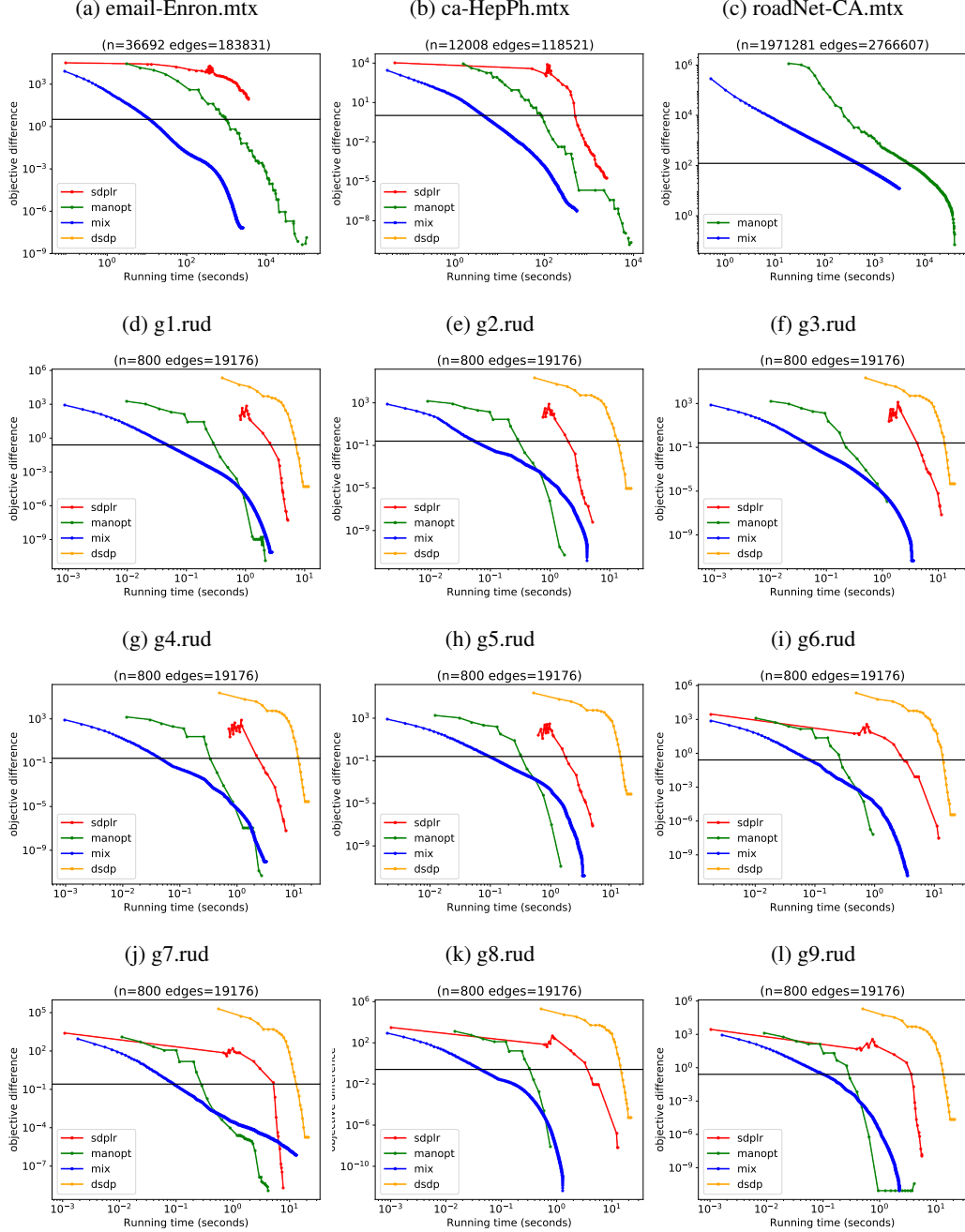


Figure 4: Function difference vs time

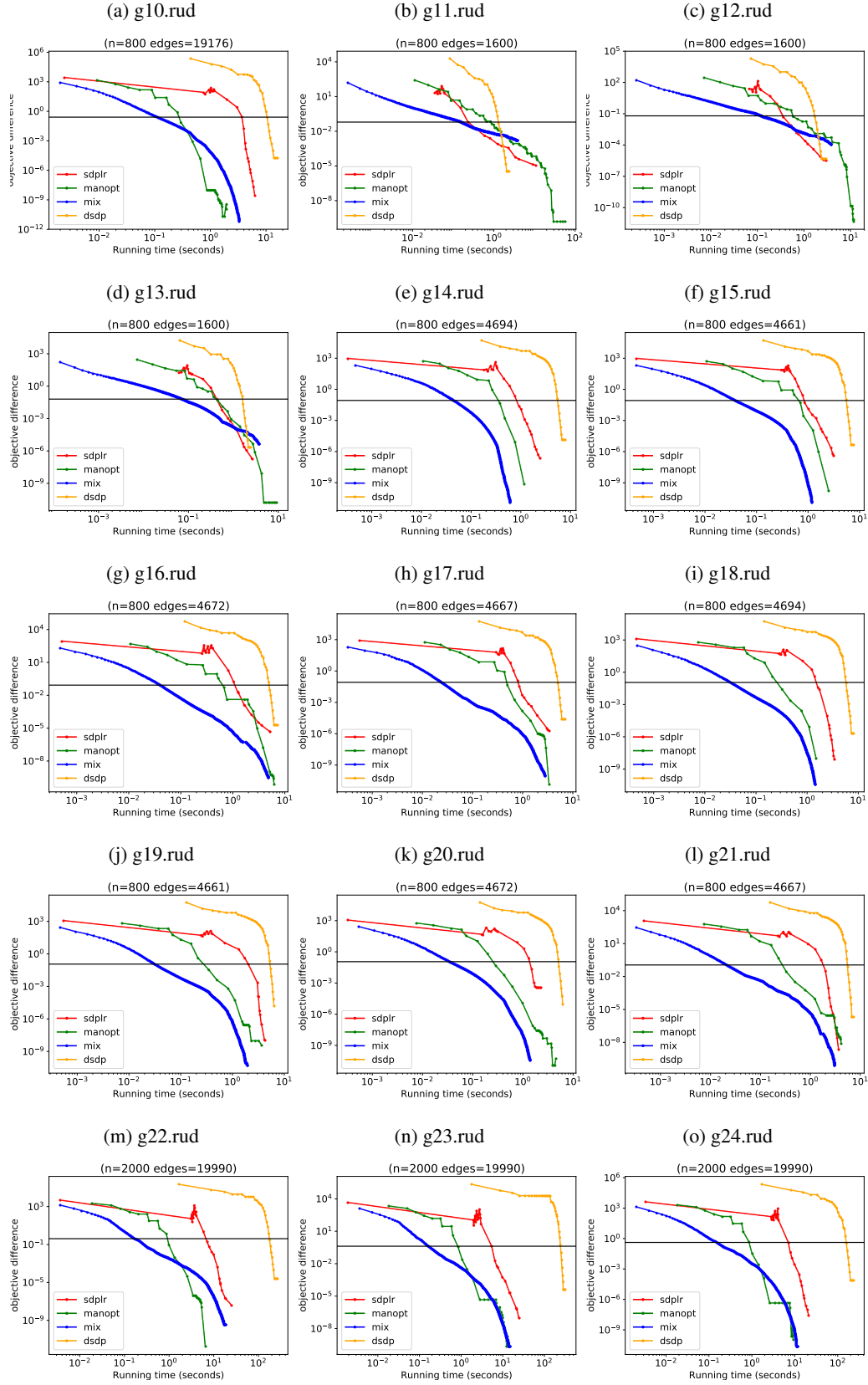


Figure 5: Function difference vs time

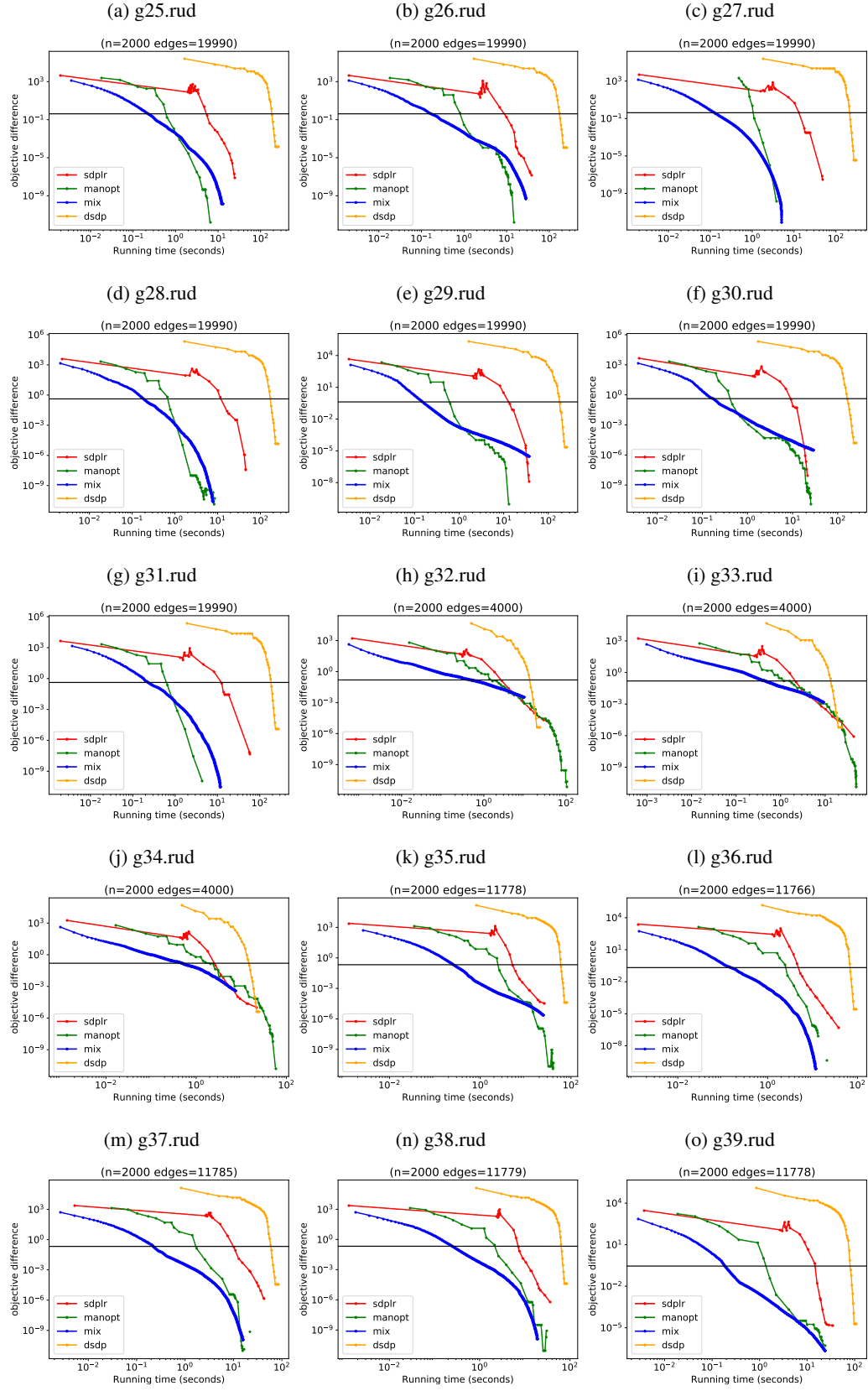


Figure 6: Function difference vs time

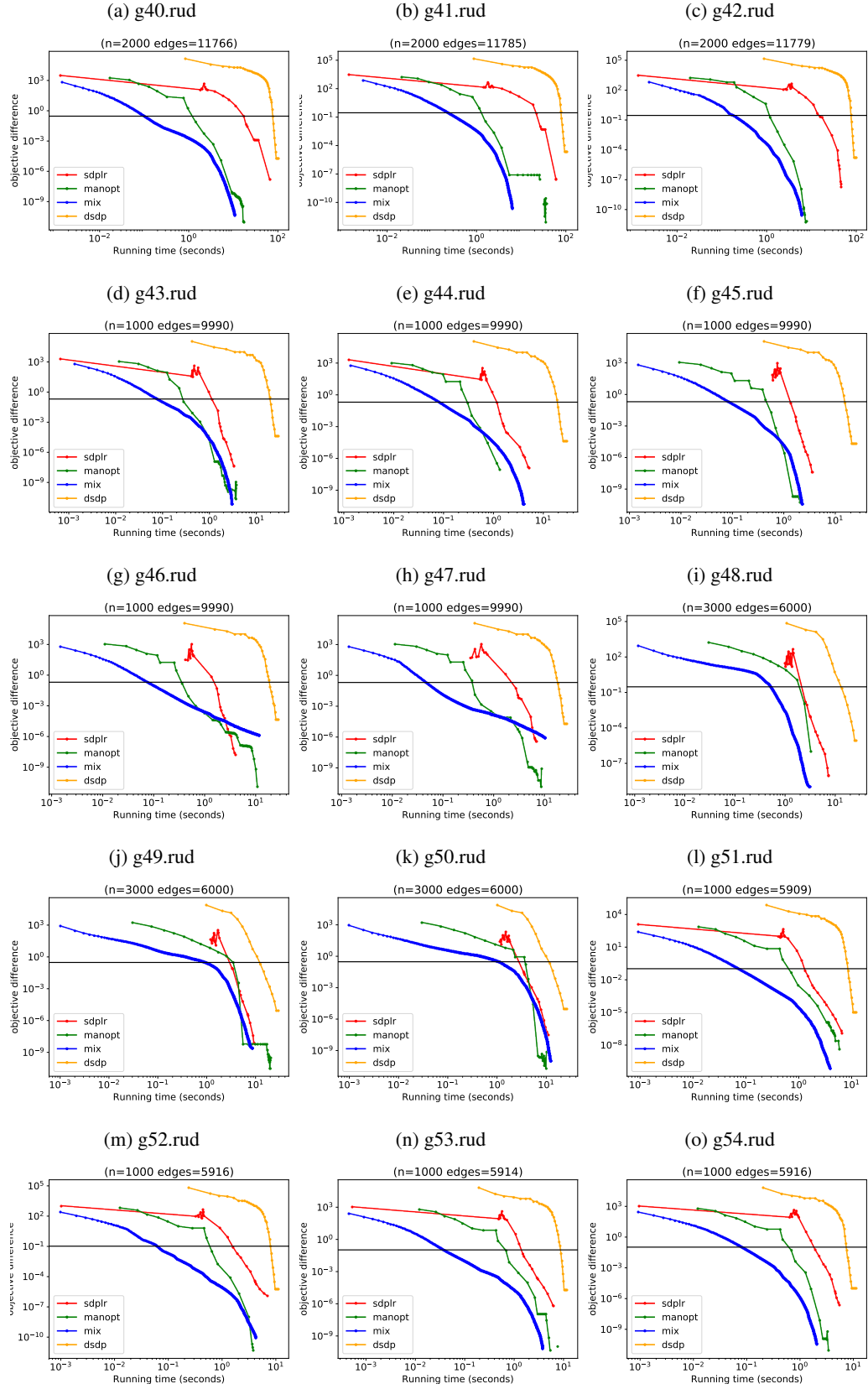


Figure 7: Function difference vs time

B Partial experiment result for MAXSAT problem

For completeness, we here include the approximation results for part of the 2016 MAXSAT competition [Argelich et al., 2016] from the 525 solved instances. The vertical lines denote the stopping time of Gurobi (LP approx), the best incomplete solver (inexact), and the best complete solver (exact) in the competition. The horizontal lines denote the approximate rate of the optimal solution (the red line), the rounded solution from Gurobi solver (LP approx), and the random solution (random). Our method (mixing) achieves as average approximation ratio of 0.978, and usually found such solution within seconds or less.

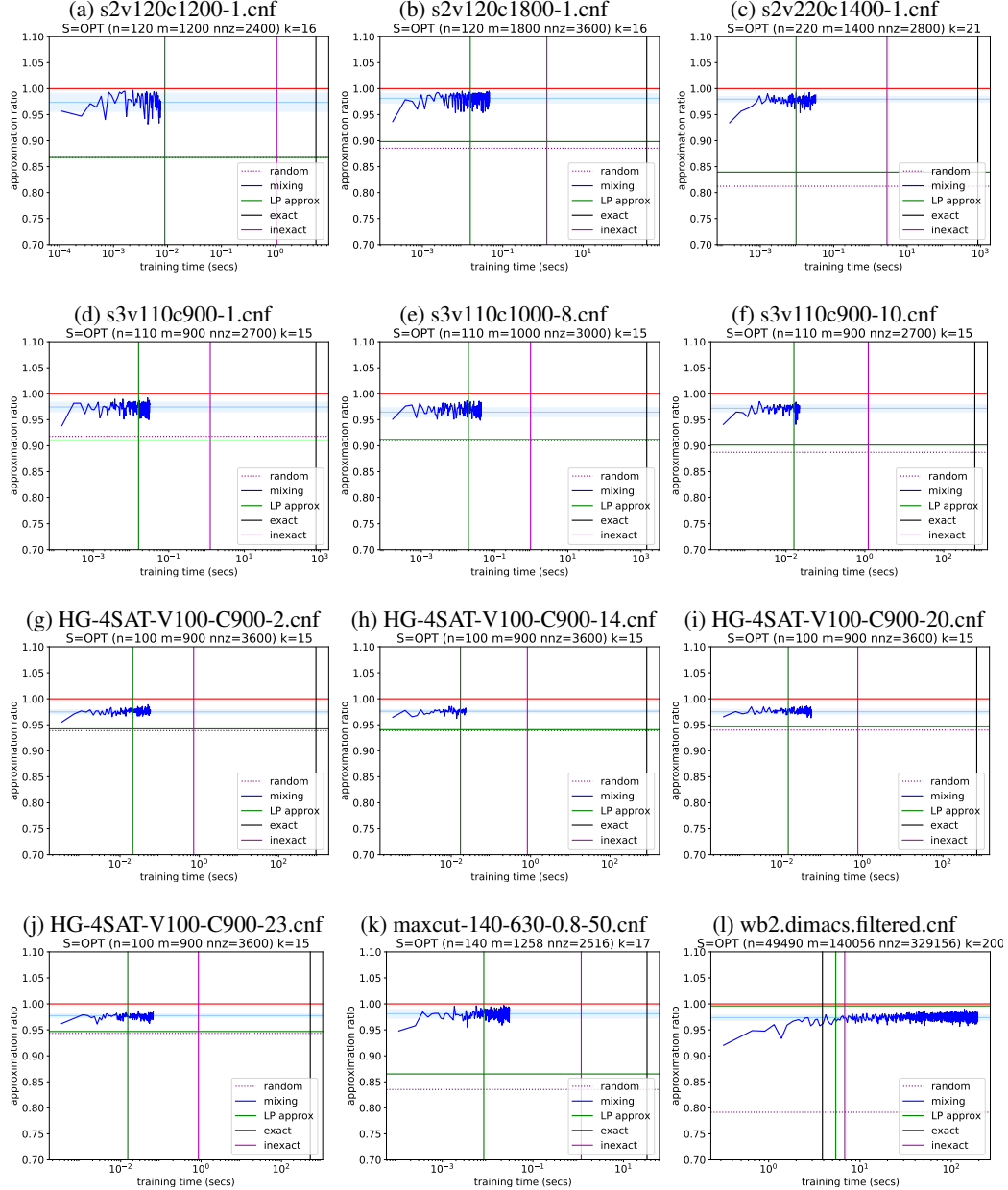


Figure 8: Approximation ratio vs time