To appear in *Optimization Methods & Software*
Vol. 00, No. 00, Month 20XX, 1–30

arXiv:1508.07607v6 [math.OC] 21 Dec 2020

OPTIMIZATION

Efficient Numerical Methods to Solve Sparse Linear Equations with Application to PageRank

Anton Anikin[a], Alexander Gasnikov[b,c,d], Alexander Gornov[a], Dmitry Kamzolov[b], Yury Maximov [e,f] and Yurii Nesterov [g]

[a] *Institute of System Dynamics and Control Theory (ISDCT SB RAS), Russia;*
[b] *Moscow Institute of Physics and Technology, Moscow, Russia;*
[c] *Higher School of Economics, Moscow, Russia;*
[d] *Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany;*
[e] *Skolkovo Institute of Science and Technology, Russia;*
[f] *Theoretical Division T-5, Los Alamos National Laboratory, USA;*
[g] *Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), Belgium.*

(*v1.0 released May 2019*)

Over the last two decades, the PageRank problem has received increased interest from the academic community as an efficient tool to estimate web-page importance in information retrieval. Despite numerous developments, the design of efficient optimization algorithms for the PageRank problem is still a challenge. This paper proposes three new algorithms with a linear-time complexity for solving the problem over a bounded-degree graph. The idea behind them is to set up the PageRank as a convex minimization problem over a unit simplex, and then solve it using iterative methods with small iteration complexity. Our theoretical results are supported by an extensive empirical justification using real-world and simulated data.

Keywords: PageRank, Sparsity, Randomization, Frank–Wolfe method, and ℓ_1-optimization

AMS Subject Classification: 90C25, 90C47, 90C60

1. Introduction

In this paper, we aim at solving a system of linear equations $Px = b$ for a sufficiently sparse matrix A and a vector b, $P \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, which is primarily motivated by finding a stationary distribution of a Markov Chain over a sparse large-scale communication graph. This problem, known as the PageRank, was pioneered by Brin and Page in [5, 36] in the early nineties; however, it still attracts significant interest from both the academic community and the industry.

In the PageRank we assume, that the (asymmetric) transition probability matrix P, associated with the web-graph, is known; so, the problem is to find a stationary distri-

Anton Anikin Email: anikin@icc.ru
Alexander Gasnikov. Email: gasnikov@yandex.ru
Alexander Gornov Email: gornov@icc.ru
Dmitry Kamzolov Email: kamzolov.dmitry@phystech.edu
Yury Maximov Email: yury@lanl.gov
Yurii Nesterov Email: nesterov@core.ucl.ac.be

bution or a vector x, $x \in \mathbb{R}^n$, with the coordinates corresponding to an expected portion of time which a random walk spends at a particular node, $x = P^\top x$.

The problem becomes especially challenging in high dimensions, since direct computations of an inverse matrix become inefficient due to non-linear time and memory efforts. Many studies have been devoted to the approximation of the PageRank vector based on a random walk analysis and Markov Chain Monte Carlo methods [2, 13, 20, 21, 38, 40]. Those methods are very attractive, both theoretically and practically, while the spectral gap, i.e., the difference between the two largest eigenvalues of the transition matrix, is sufficiently large. The latter is often not the case for sparse graphs with complex topologies, and, furthermore, estimating the gap requires significant time and effort, as well [23].

Another line of research aims at finding a stationary distribution of a Markov chain using convex optimization [14, 26, 31, 34]. According to these results, the PageRank can be equally stated as an ℓ_p-norm minimization, $p \geq 1$, constrained on a unit simplex:

$$\|P^\top x - x\|_p \to \min_{x \in \Delta_1^n}, \tag{1}$$

where $\Delta_1^n = \{x : \sum_{i=1}^n x^i = 1, x^i \geq 0\}$.

From the convex optimization perspective, similar problems appear in applied mathematics, statistics, and machine learning. Among these are LASSO [12], a traffic matrix estimation in large-scale communication networks [41], phase recovery in a linearized model of electric current [39], and a finite element method [24]. The high-dimensional nature of the problems above call into a question of the utility of traditional approaches, which do not devote sufficient attention to the problem structure and require non-linear time and memory efforts.

In this paper, we focus on the influence of the transition probability matrix sparsity on the computational complexity of the PageRank problem. We advocate particular efficiency of convex optimization methods for the simplex constrained $\|Ax\|_2^2$ and $\|Ax\|_\infty$ minimization problems [6, 7]. In particular, we prove that the time complexity of these problems is linear in the problem dimension if the number of non-zeros in each row/column is bounded above by some constant d. The key contribution is a set of efficient algorithms to update a function value, a gradient, and an argument in an (almost) dimension-independent manner. Later in the paper, we extend our results to a more general setup, where a limited number of dense rows or columns in the transition matrix is allowed.

1.1 *Contribution*

Our contribution is as follows. We propose:

(1) The `NL1` algorithm, a ℓ_1-proximal gradient descent method that supports sparse updates of the gradient and the function value. It allows us to solve the problem with overall time complexity $O(nd^2 \log^2 n + d^2 \log^2(n/\varepsilon)/\varepsilon^2)$, where d is the maximal in- and out- vertexes' degree of the graph, ε is required accuracy, and $\|Px - x\|_2 \leq \varepsilon$;
(2) The `S-FW` method, an extension of the Frank–Wolfe algorithm that allows efficient updates to the gradient and the objective value. The resulting algorithm `S-FW` has better time-complexity estimates compared to the `NL1` method, $O(n + d^2 \log(2 + n/d^2)/\varepsilon^2)$. Also, the algorithm often has remarkably better performance in practice;
(3) The `GK` algorithm, which aims to minimize $\|P^\top x - x\|_\infty$ over the unit simplex. We provide an equivalent saddle-point setup for this problem and, subsequently, solve it

subsequently by a version of the mirror descent with a randomized projecting. We prove that a randomized projection with KL-divergence guarantees the total running time of the algorithm to be bounded from above by $O(n + d \log n \log(n/\delta)/\varepsilon^2)$ with probability at least $1 - \delta$, for any $0 < \delta < 1$;

(4) Finally, we extend the linear-time complexity estimates to the sparse graphs with a small number of dense rows/columns containing more than d non-zeros.

Let us emphasize that a d-sparse matrix could have as many non-zero elements as $n \cdot d$. Time-complexity estimates of the proposed algorithms are yet sub-linear in the number of non-zero elements of the transition matrix if its sparsity pattern is known beforehand. The last statement implies that there is no need to read all non-zero elements of a stochastic matrix to arrive at approximate solution of the PageRank problem. In Table 1 we summarize the best known results for solving the PageRank problem using optimization techniques.

Algorithm	**Constraint**	**Time Complexity**	**Objective**
Nazin - Polyak [26]	no	$O\left(\frac{n \log(n/\delta)}{\varepsilon^2}\right)$	$\|P^\top x - x\|_2 \leq \varepsilon$
Nesterov [29]	d-sparse	$O\left(n + \frac{d^2 \log n}{\varepsilon^2}\right)$	$\mathbb{E}\|P^\top x - x\|_2 \leq \varepsilon$
Nesterov [31]	d-sparse	$O\left(dn \frac{\log n}{\varepsilon^2}\right)$ $O\left(\frac{d^{1/2} n^{3/2} \log n}{\varepsilon}\right)$	$\|P^\top x - x\|_\infty \leq \varepsilon$
Juditsky et al. [19, 27]	no	$O\left(\frac{n \log(n/\delta)}{\varepsilon^2}\right)$	$\|P^\top x - x\|_\infty \leq \varepsilon$
Nesterov - Nemirovski [34]	d-sparse Google page	$O\left(\frac{nd}{\alpha} \log 1/\varepsilon\right)$	$\|x - x^*\|_1 \leq \varepsilon$
Polyak - Tremba [37]	d-sparse on average	$O\left(\frac{dn}{\varepsilon}\right)$	$\|P^\top x - x\|_1 \leq \varepsilon$
Gasnikov - Dmitriev, [13]	d-sparse spectral gap β	$O\left(n + \frac{d \log n \log(n/\delta)}{\beta \varepsilon^2}\right)$	$\mathbb{E}\|x - x^*\|_2 \leq \varepsilon$
Gasnikov - Dmitriev, [13]	m non-zeros	$O\left(m + \left(n + \frac{m^2}{n^2}\right) \frac{\log n}{\varepsilon^2}\right)$	$\|P^T x - x\|_\infty \leq \varepsilon$
Langville - Meyer, [22]	d-sparse spectral gap β	$O\left(\frac{dn}{\beta} \log \frac{n}{\varepsilon}\right)$	$\|x - x^*\|_1 \leq \varepsilon$
Cohen et al. [9]	m non-zeros	$O(m + n^{1+o(1)}) \log^{O(1)}\left(\frac{n}{\beta\varepsilon}\right)$	$\mathbb{E}\|P^\top x - x\|_2 \leq \varepsilon$
This paper, Section 2	d-sparse	$O\left(nd^2 \log^2 n + \frac{d^2 \log(n/d^2+1)}{\varepsilon^2}\right)$	$\|P^\top x - x\|_2 \leq \varepsilon$
This paper, Section 3	d-sparse	$O\left(n + \frac{d^2 \log(n/d^2+2)}{\varepsilon^2}\right)$	$\|P^\top x - x\|_2 \leq \varepsilon$
This paper, Section 4	d-sparse	$O\left(n + \frac{d \log n \log(n/\delta)}{\varepsilon^2}\right)$	$\|P^\top x - x\|_\infty \leq \varepsilon$

Table 1. Time complexity of the PageRank problem. Time complexity of the algorithms proposed in this paper along with results of Nesterov [29] are the only sub-linear algorithms known to the authors. Algorithm [9] can be favourable in theory for high dimensions, but useless in practice due to a high degree of the logarithm.

The Google page condition, required in [34], implies the existence of a column j such that all $P_{ij} \geq \alpha$, $\alpha > 0$ for any $i : 1 \leq i \leq n$; and β corresponds to the spectral gap of the matrix P, e.g. $\beta = \lambda_1(P) - \lambda_2(P) = 1 - \lambda_2(P)$, the difference between the largest and the second-largest eigenvalues of the transition matrix P. Algorithm [29] runs on average for the randomized coordinate descent algorithm while the estimates [13, 26, 27] are correct with probability at least $1 - \delta$, for any $\delta : 0 < \delta < 1$.

1.2 *Paper structure and notation*

In Section 2 we introduce a ℓ_1 proximal gradient descent algorithm. The key idea behind this time-efficient algorithm, referenced below as `NL1`, is a sparse update to a gradient and a function value updates, $f(x) = \|Ax\|_2^2$, $A = P^\top - I$:

$$\nabla f(x+h) = \nabla f(x) + A^\top Ah, \quad f(x+h) = f(x) + h^\top \nabla f(x) + \|Ah\|_2^2.$$

The `NL1` method does not require computation of a (full) gradient on each step if matrix A is sufficiently sparse. We show that the update vector h on each step has only two non-zero coordinates, which correspond to the minimal and maximal components of the gradient. To update gradient coordinates and extract the minimal and the maximal value, we use a list of binary heaps [10], which admits logarithmic dependence of iteration complexity in dimension.

Algorithm `S-FW`, a revision of the Frank–Wolfe conditional gradient, is proposed in Section 3. The Frank–Wolfe algorithm has recently stimulated much interest, mainly due to the numerous Big Data problems to which it has been applied [17, 18, 32]. In this paper, we focus on an efficient gradient and a function value update for each iteration, which reduces the algorithm's overall time complexity. Accurate theoretical analysis results in a better time complexity estimate compared to the `NL1` algorithm. Also, the `S-FW` and `NL1` algorithms appear to be comparable from in practice.

In Section 4 we deal with the problem of the approximation of the PageRank vector in the ℓ_∞ norm, i.e., the minimization of $\|P^\top x - x\|_\infty$. Using a technique of Juditsky et al. [19], the minimization can be equally stated as a saddle-point problem over the product of two unit simplexes. A randomized mirror descent seems to be one of the most popular tools for solving the problem efficiently. Its application to the PageRank improves time-complexity of an iteration to $O(n \log(n/\delta)/\varepsilon^2)$, and does not depend on the problem sparsity [27].

Finally, in this paper, we investigate an approach pioneered by Grigoriadis and Khachiyan in [16]. The approach's idea is a randomized projection of the gradient on the simplex instead of a randomized approximation of the gradient itself. The gradient's projection on the unit simplex is carried out with the Kullback-Leibler divergence, corresponding to the exponential weighting of the gradient coordinates. The (sparse) randomized projection chooses one of the vertices of the unit simplex in such a way that the expected value gives an unbiased estimation of the projection itself [13].

To this end, we design an algorithm to update the gradient vector on each iteration in an (almost) dimension-independent manner, e.g., with an (almost) linear time-complexity

$$O(n + d \log n \log(n/\sigma)/\varepsilon^2)$$

for d-sparse transition matrices, with $\|P^\top x - x\|_\infty \leq \varepsilon$. Finally, we pay a special attention to its interpretation in terms of game theory [25].

In Sections 2 - 4, we summarize sparsification techniques allowing to reduce ranking problems over a dense graph to the problem on a sparse graph. The proposed algorithms deliver state-of-the-art time-complexity estimates for d-sparse optimization problems if d is sufficiently small. Faster methods are possible via the Markov Chain Monte Carlo method for several situations. We refer to recent results [13] for details.

We conclude in Section 5 with the implementation details and a case study. Finally, some technical proofs are given in the appendix.

In this paper, we use the following notation. By $\|x\|_p$, we denote the ℓ_p norm of vector

$x \in \mathbb{R}^n$, in particular, $\|x\|_2 = x^\top x$ denotes the Euclidean norm, $\|x\|_1 = \sum_{i=1}^n |x_i|$, and $\|x\|_\infty = \max\{|x_1|, \dots, |x_n|\}$. By $O(1)$, we denote a positive constant.

2. ℓ_1-Gradient descent for d-sparse problems

The PageRank problem can be equally stated as the following convex optimization problem for a sufficiently large positive constant γ

$$f_\gamma(x) = \frac{1}{2}\|Ax\|_2^2 + \sum_{i=1}^n \frac{\gamma}{2}(-x^i)_+^2 \to \min_{x^\top e=1}, \qquad \gamma > 0, \tag{2}$$

where $A = I - P^\top$, I is an $n \times n$ identity matrix, $P \in \mathbb{R}_+^{n\times n}$ is a column stochastic transition matrix, e is an n-dimensional vector with each coordinate equals to one. Notice, that $f_\gamma(x)$ is a convex, non-negative, and increasing function with $f_\gamma(x_*) = 0$, and $z_+ = \max\{z,0\}$.

To solve the problem, we use a proximal gradient descent with an ℓ_1 setup:

$$x_{k+1} = x_k + \operatorname*{argmin}_{h:h^\top e=0} \left\{ f_\gamma(x_k) + h^\top \nabla f_\gamma(x_k) + \frac{L_1}{2}\|h\|_1^2 \right\}, \tag{3}$$

where $f_\gamma(x)$ is L_1-smooth in ℓ_1 norm

$$\|\nabla f_\gamma(x) - \nabla f_\gamma(y)\|_\infty \le L_1 \|x - y\|_1.$$

In the case of the PageRank, $L_1 = 1 + \gamma$, as P is a stochastic matrix. Lemma 2.1 implies that the vector h_k, which solves the problem (3), is always sparse whatever the function $f_\gamma(x)$ is.

LEMMA 2.1 *A set of solutions of the following minimization problem*

$$\phi(h) = f_\gamma(x_k) + h^\top \nabla f_\gamma(x_k) + \frac{L_1}{2}\|h\|_1^2 \to \min_{h:h^\top e=0} \tag{4}$$

contains at least one vector h_k with only two non-zero coordinates

$$h_k^{i_+} = -\frac{1}{4L_1}\left(\frac{\partial f_\gamma(x_k)}{\partial x^{i_+}} - \frac{\partial f_\gamma(x_k)}{\partial x^{i_-}}\right) \text{ and } h_k^{i_-} = \frac{1}{4L_1}\left(\frac{\partial f_\gamma(x_k)}{\partial x^{i_+}} - \frac{\partial f_\gamma(x_k)}{\partial x^{i_-}}\right),$$

where $i_+ = \operatorname{argmax}_{1\le i\le n} \partial f_\gamma(x_k)/\partial x^i$ and $i_- = \operatorname{argmin}_{1\le i\le n} \partial f_\gamma(x_k)/\partial x^i$.

Proof. Let $\bar{h}$ be a solution to the problem 4, so that $\|\bar{h}\|_1 = c$. The minimum to the linear function $f_\gamma(x_k) + h^\top \nabla f_\gamma(x_k) + \frac{L_1}{2}c^2 \to \min_{h:h^\top e=0, \|h\|_1=c}$ is attained at h_k which has exactly two non-zero coordinates $h_k^{i^+} = -c/2$, $h_k^{i^-} = c/2$. The objective value is then

$$\phi(h_k) = f_\gamma(x_k) - (\max_{1\le i\le n} \partial f_\gamma(x_k)/\partial x^i - \min_{1\le i\le n} \partial f_\gamma(x_k)/\partial x^i)c/2 + L_1 c^2/2.$$

Taking minimum in $c \ge 0$ we get the statement of the lemma. ∎

By Lemma 2.1, the gradient $\nabla f_\gamma(x_{k+1})$ is

$$\begin{aligned}\nabla f_\gamma(x_{k+1}) =& \nabla f_\gamma(x_k + h_k) \\ =& A^\top A x_k + A^\top A h_k + \gamma \sum_{i=i_+,i_-} \left\{(-x_k^i - h_k^i)_+^2 - (-x_k^i)_+^2\right\}. \end{aligned} \tag{5}$$

The update vector, h_k, has at most $O(\min(n, d^2))$ non-zero coordinates. To efficiently update the gradient, we will use a doubly-linked list of binary heaps, so that the j-th heap is used to extract the minimal value and update coordinates from $(j-1)\lfloor n/d^2 \rfloor + 1$ to $\min\{j\lfloor n/d^2 \rfloor, n\}$ inclusively. We refer to a binary heap [10] as a *Max-Heap* (resp. *Min-Heap*), if the key stored in each node is greater (resp. less) than or equal to the keys in the node's children. We require both the minimal and the maximal coordinates of the gradient for an iteration of the algorithm. Using *Min-Heap* (*Max-Heap*) extracting the minimal (maximal) element from a heap of m items requires $O(\log m)$ time. An update of a single element to preserve the keys' order requires $O(\log m)$ time, as well [10]. That is why a gradient update (5) requires $O(d^2 \log(2 + n/d^2))$ time in total.

Algorithm 1: `NL1`: ℓ_1 Gradient Descent for PageRank

Input: d-sparse transition matrix P, starting point x_0,
objective $f_\gamma(x) = \frac{1}{2}\|Ax\|_2^2 + \frac{\gamma}{2}\sum_{i=1}^n (-x_i)_+^2$ with $A = I - P^\top$ and $\gamma > 0$,
number of iterations N required by Theorem 2.2 and accuracy ε.
Output: x_N, $\|Ax_N\|_2^2 \le \varepsilon^2$ where N is given by Theorem 2.2

while $k \le N$ *and* $f_\gamma(x) > \varepsilon^2$ **do**
  $i_+ = \operatorname{argmax}_{1\le i\le n} \partial f_\gamma(x_k)/\partial x^i, \quad i_- = \operatorname{argmin}_{1\le i\le n} \partial f_\gamma(x_k)/\partial x^i$
  $h_k^i = \begin{cases} \frac{1}{4L_1}\left(\frac{\partial f_\gamma(x_k)}{\partial x^{i_+}} - \frac{\partial f_\gamma(x_k)}{\partial x^{i_-}}\right), & i = i_+ \\ -\frac{1}{4L_1}\left(\frac{\partial f_\gamma(x_k)}{\partial x^{i_+}} - \frac{\partial f_\gamma(x_k)}{\partial x^{i_-}}\right), & i = i_- \\ 0, & \text{otherwise} \end{cases}$
  Update argument: $x_{k+1} = x_k + h_k$
  Update gradient: $\nabla f_\gamma(x_{k+1}) = \nabla f_\gamma(x_k) + A^\top A h_k + \gamma(-x_k + h_k)_+ - \gamma(-x_k)_+$
  Update $f_\gamma(x_{k+1})$: $f_\gamma(x_{k+1}) =$
   $f_\gamma(x_k) + {h_k}^\top A^\top A x_k + \|Ah_k\|_2^2/2 + \frac{\gamma}{2}\sum_{i:\, {h_k}^i \ne 0} (-{x_k}^i)_+^2 - (-x_k^i - {h_k}^i_k)_+^2$
  $k = k + 1$
return x_k

Algorithm 1 presents the `NL1` algorithm with convergence rate established in Theorem 2.2.

Theorem 2.2 *For a starting point x_0 in one of the vertices of the unit simplex, Algorithm 1 converges to $f_\gamma(x) \le \varepsilon^2$ for any constant $\gamma > 0$, $x \ge 0$, $e^\top x = 1$, with the overall time complexity*

$$T = O\left(nd^2 \log(n/d^2 + 2)\log n + \frac{d^2 \log(n/d^2 + 2)}{\varepsilon^2}\right).$$

The complete proof of the theorem in given in the Appendix. The bound established in Theorem 2.2 gives a sub-linear time complexity for the case $n\varepsilon^2 = o(1)$, while it is less practical for sufficiently large ε, $n\varepsilon^2 = \Omega(1)$.

Matrix Sparsification. In a number of practical problems, both column and row sparsity of the transition probability matrix seems to be very restrictive. Indeed, for the PageRank problem, search engines such as `Google` or `Yahoo!` may refer to a large number of sites simultaneously. In particular, that means that matrix $P^\top$ may have a few dense columns. Below we propose a method to sparsify the transition probability matrix P in order to improve the convergence of optimization methods without loss of quality.

Consider a single linear equation $a^\top x = b$, $x \in \mathbb{R}^n$, where a is a dense vector. The system can be equally stated as

$$a_i^\top x = bz^i, \quad 1 \le i \le \lceil n/d \rceil, \quad \sum_{i=1}^{\lceil n/d \rceil} a_i = a, \quad \sum_{i=1}^{\lceil n/d \rceil} z^i = 1, \, z \in \mathbb{R}^{\lceil n/d \rceil} \tag{6}$$

such that each a_i has no more than d non-zero elements. In order to guarantee that $\|a^\top x - b\|_2^2 \le \varepsilon^2$, one needs to find a vector $\psi = (x,\, z)^\top$, $\psi \in \mathbb{R}^{n+\lceil n/d \rceil}$, such that

$$\sum_{i=1}^{\lceil n/d \rceil} \lceil n/d \rceil \|a_i^\top x - bz^i\|_2^2 \le \varepsilon^2$$

as under Conditions (6) and the Cauchy inequality one has:

$$\|a^\top x - b\|_2^2 = \left\| \sum_{i=1}^{\lceil n/d \rceil} a_i^\top x - bz_i \right\|_2^2 \ge \lceil n/d \rceil^{-1} \sum_{i=1}^{\lceil n/d \rceil} \|a_i^\top x - bz_i\|_2^2.$$

Eqs. (6) correspond to a system of linear equations $\tilde{A}\psi = \tilde{A}(x,\, z)^\top = 0$ of the size $(n + d \cdot \lceil n/d \rceil) \times (n + \lceil n/d \rceil)$ with no more than $d+1$ non-zero elements in each row. For each i, row $\tilde{A}_i$ of $\tilde{A}$ is as such $\tilde{A}_{ij} = a_{ij}$ for $1 \le j \le n$; $\tilde{A}_{ij} = -b$ for $j = n + i$, and $\tilde{A}_{ij} = 0$ otherwise. For our convenience, we refer $\tilde{A}^x$ as the matrix consists of the first n columns of $\tilde{A}$, and $\tilde{A}^z$ and the matrix consists of the last n/d columns. In order to solve the problem, we apply the `NL1` algorithm, although it requires a minor correction:

$$f_\gamma(x,z) = \frac{1}{2}\|\tilde{A}\psi\|_2^2 + \frac{\gamma}{2}\sum_{i=1}^{n}(-x^i)_+^2 + \frac{\gamma}{2}\sum_{j=1}^{\lceil n/d \rceil}(-z^i)_+^2 \to \min_{\substack{x^\top e = 1,\, x \ge 0 \\ z^\top e = 1}}.$$

A step of the `NL1` algorithm is

$$\begin{pmatrix} x_{k+1} \\ z_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ z_k \end{pmatrix} + \operatorname*{argmin}_{\substack{h_x :\, h_x^\top e = 0 \\ h_z :\, h_z^\top e = 0}} \left\{ f_\gamma(x_k, z_k) + \nabla f_\gamma(x_k, z_k) \cdot \begin{pmatrix} x_k \\ z_k \end{pmatrix} + \frac{L}{2} \left\| \begin{pmatrix} h_x \\ h_z \end{pmatrix} \right\|_1^2 \right\}. \tag{7}$$

Notice, that Eq. (7) is separable in variables x and z, thus

$$x_{k+1} = x_k + \operatorname*{argmin}_{h_x : h_x^\top e = 0} \left\{ \frac{1}{2}\|\tilde{A}^x h_x\| + \frac{\gamma}{2}\sum_{i=1}^{n}(-x^i)_+^2 + \gamma \tilde{A}^x + \gamma \sum_{i=1}^{n}(-x_i)_+ + \frac{L}{2}\|h_x\|_1^2 \right\}$$

$$z_{k+1} = z_k + \operatorname*{argmin}_{h_z : h_z^\top e = 0} \left\{ \frac{1}{2}\|\tilde{A}^z h_z\| + \frac{\gamma}{2}\sum_{i=1}^{\lceil n/d \rceil}(-z^i)_+^2 + \gamma \tilde{A}^z + \gamma \sum_{i=1}^{n}(-z_i)_+ + \frac{L}{2}\|h_z\|_1^2 \right\}$$

Recall that the optimal $h = (h_x, h_z)^\top$ is such as it has only two non-zero coordinates in each of h_x and h_z, as was true for the `NL1` algorithm over the simplex, corresponding to

$$i_x^+ = \operatorname*{argmax}_{1\le i\le n} \frac{\partial f(x_k,z_k)}{\partial x^i} \quad \text{and} \quad i_x^- = \operatorname*{argmin}_{1\le i\le n} \frac{\partial f(x_k,z_k)}{\partial x^i},$$

if

$$\max_{1\le i\le n} \frac{\partial f_\gamma(x_k,z_k)}{\partial x^i} - \min_{1\le i\le n} \frac{\partial f_\gamma(x_k,z_k)}{\partial x^i} > \max_{1\le j\le \lceil \frac{n}{d} \rceil} \frac{\partial f_\gamma(x_k,z_k)}{\partial z^j} - \min_{1\le j\le \lceil \frac{n}{d} \rceil} \frac{\partial f_\gamma(x_k,z_k)}{\partial z^j},$$

while

$$i_z^+ = \operatorname*{argmax}_{1\le j\le \lceil \frac{n}{d} \rceil} \frac{\partial f_\gamma(x_k,z_k)}{\partial z^j} \quad \text{and} \quad i_z^- = \operatorname*{argmin}_{1\le j\le \lceil \frac{n}{d} \rceil} \frac{\partial f_\gamma(x_k,z_k)}{\partial z^j},$$

otherwise. Now, for efficient implementation of the `NL1` algorithm, one needs to store the values $\partial f_\gamma/\partial z^j$ and $\partial f_\gamma/\partial x^i$ in separate binary heaps. A similar strategy can be used if sparsification of multiple rows is required. Finally, we remind the reader that single-row sparsification increases the problem dimension by $\lceil n/d \rceil$, at most.

Column sparsification is slightly more involved. Consider a function $f(x)$ in more details:

$$\begin{aligned} f_\gamma(x) =& \frac{1}{2}\|Ax\|_2^2 + \frac{\gamma}{2}\sum_{i=1}^n (-x)_+^2, \qquad x = (x_1, x_2, \dots x_n) \\ \bar{f}_\gamma(x_1) =& \min_{x_{2:n}} \left\{ f(x) + I_{\{x\in\Delta_1^n\}} \right\}. \end{aligned}$$

The function $\bar{f}_\gamma(x_1)$ is convex in x_1 for any fixed $\gamma > 0$ [4, Section 3.2.5]. That is why the value of $\bar{f}_\gamma(x_1)$ can be computed using the `NL1` algorithm while the value

$$f_\gamma^* = \min_{0\le x_1\le 1} f_\gamma(x_1)$$

requires a one dimensional binary search so that the overall time complexity T of the `NL1` algorithm used to solve the PageRank problem with a single dense column is in $O(\log(n/\varepsilon))$ higher than the one established in Theorem 2.2.

The same approach can be applied to improving the efficiency of the `S-FW` and `GK` algorithms proposed later in Sections 3 and 4 on the PageRank instances with a few dense rows or columns.

3. Frank–Wolfe algorithm with sparse updates

The PageRank problem, according to Eq. (2), can be stated as

$$f(x) = \frac{1}{2}\|Ax\|_2^2 \to \min_{x\in\Delta_1^n}.$$

We use the Frank–Wolfe conditional gradient [11, 18] to solve the problem above.

We choose the starting point x_0 of the algorithm in an arbitrary vertex of the unit simplex. Then on each step we solve

$$h_k^\top \nabla f(x_k) \to \min_{y \in \Delta_1^n}. \tag{8}$$

Furthermore the solution y_k of Eq. (8) has only one non-zero coordinate $y_k^{i_k}$, corresponding to

$$i_k = \operatorname*{argmin}_{1 \le i \le k} \partial f(x)/\partial x^i.$$

Then, the update rule is

$$x_{k+1} = (1-\gamma_k)x_k + \gamma_k h_k, \quad \gamma_k = \frac{2}{k+1}, \quad k \ge 1.$$

According to [30], $f(x_k) - f(x)$ is bounded from above as:

$$f(x_k) - f^* = f(x_k) \le \frac{2L_1 \max_{x,y \in \Delta_1^n} \|x-y\|_1^2}{k+1} \le \frac{8L_1}{k+1}, \tag{9}$$

where $L_1^2 = \max_{x \in \Delta_1^n} \|Ax\|_2^2 \le 2$. Thus, in order to guarantee $f(x_k) \le \varepsilon^2/2$, one needs at most $32\varepsilon^{-2}$ iterations.

Consider a step of the algorithm in more detail. Denote β_k as

$$\beta_k = \prod_{r=1}^{k-1}(1-\gamma_r),\ z_k = x_k/\beta_k,\ \tilde{\gamma}_k = \gamma_k/\beta_{k+1}, \text{with } \beta_0 = 1.$$

A step of the Frank–Wolfe algorithm is

$$z_{k+1} = z_k + \tilde{\gamma}_k h_k, \text{ with } z_1 = x_1.$$

where h_k is a solution of Eq. (8). Moreover, only one coordinate of h_k:

$$i_k = \operatorname*{argmin}_{1 \le i \le k} \partial f(x_k)/\partial x^i = \operatorname*{argmin}_{1 \le i \le k} A^\top A z_k^i$$

is other than zero. Therefore, the update of the `S-FW` algorithm is similar to the Gauss-Southwell rule studied in detail in [35] for minimization of the strongly convex functions.

Using the doubly-linked list of binary heaps, described in Section 2, the minimal coordinate of $A^\top A z_k$ can be computed in $O(d^2 \log(2 + n/d^2))$ time. Then $A^\top A z_{k+1}$ is

$$A^\top A z_{k+1} = A^\top A z_k + \tilde{\gamma}_k A^\top A y_k, \tag{10}$$

and

$$\frac{\nabla f(x_{k+1})}{\beta_k} = \frac{\nabla f(x_k)}{\beta_k} + \frac{\tilde{\gamma}_k \beta_k A^\top A h_k}{\beta_k},$$

Algorithm 2: S-FW: Frank–Wolfe algorithm for PageRank

Input: d-sparse transition matrix P, starting point x_0 in one of the vertices of unit simplex and $A = P^\top - I$; Set $k = 1$ and $\beta_0 = 1$

Output: $x_k : \|Ax_k\|_2 \le \varepsilon$

while $f(z_k) > \beta_k^2 \varepsilon^2$ **do**
- $i_k = \operatorname{argmin}_{1 \le i \le n} \partial f(x_k)/\partial x^i$
- Set $\gamma_k = 2/(k+1)$, $\beta_{k+1} = \beta_k(1-\gamma_k)$, and $\tilde{\gamma}_k = \gamma_k/\beta_{k+1}$
- Update step direction $h_k = \tilde{\gamma}_k \cdot \delta$, where $\delta^{i_k} = 1$, and $\delta^i = 0$, $i \ne i_k$
- Update argument: $z_{k+1} = z_k + \tilde{\gamma}_k h_k$
- Update gradient: $\frac{\nabla f(x_{k+1})}{\beta_k} = \frac{\nabla f(x_k)}{\beta_k} + \tilde{\gamma}_k A^\top A h_k$
- Update function value $f(x_{k+1})$: $\frac{f(x_{k+1})}{\beta_k^2} = \frac{f(x_k)}{\beta_k^2} + \tilde{\gamma}_k h_k^\top \frac{\nabla f(x)}{\beta_k} + \|Ah_k\|_2^2 \tilde{\gamma}_k^2/2$
- $k = k + 1$

return $x_k = z_k \beta_k$ //Compute x_k on the last iteration only

and also

$$\frac{f(x_{k+1})}{\beta_k^2} = \frac{f(x_k)}{\beta_k^2} + \tilde{\gamma}_k h_k^\top \frac{\nabla f(x)}{\beta_k} + \|Ah_k\|_2^2 \tilde{\gamma}_k^2/2.$$

For a d-sparse matrix A, the time required to compute $A^\top A z_{k+1}$ from $A^\top A z_k$ is $O(d^2 \log(2 + n/d^2))$, as well. One can compute x_k having z_k as $x_k = \beta_k x_k$ in $O(n)$ time.

Combining Eqs. (9) and (10), we have the following complexity estimate for the algorithm.

Theorem 3.1 *Algorithm 2 requires at most $k = 32\varepsilon^{-2}$ iterations to guarantee $\|P^\top x - x\|_2 \le \varepsilon$ for any d-sparse transition matrix P. The overall time complexity of the algorithm does not exceed*

$$T = O\left(n + \frac{d^2 \log(2 + n/d^2)}{\varepsilon^2}\right).$$

Discussion. Theorem 3.1 implies sub-linear convergence in the number of non-zero elements of the transition matrix P. This is not surprising, since we assume that the underlying graph structure, along with required smoothness, are known a priori.

It remains an open question for the authors to improve convergence rate in terms of the accuracy and maximal degree of the transition graph while preserving linear dependence in the problem dimension.

4. Saddle point setup for PageRank

In the PageRank problem, one often requires an accurate approximation of a few of the largest coordinates representing the most relevant websites rather than the full PageRank vector. To this end, we propose an algorithm to approximate the PageRank vector in

ℓ_∞-norm. Below, we consider the problem

$$f(x) = \|(P^\top - I)x\|_\infty = \|Ax\|_\infty \to \min_{x \in \Delta_1^n}.$$

Following [27], we set up the problem as

$$\min_{x \in \Delta_1^n} \max_{\|\tilde{y}\|_1 \le 1} \langle Ax, \tilde{y} \rangle = \min_{x \in \Delta_1^n} \max_{y \in \Delta_1^{2n}} \langle Ax, Jy \rangle = \min_{x \in \Delta_1^n} \max_{y \in \Delta_1^{2n}} \langle x, \tilde{A}y \rangle, \tag{11}$$

where $\tilde{A} = A^\top J$, $J = [I_n, -I_n]$, and I_n is the $n \times n$ identity matrix.

We propose a sub-linear-time algorithm to approximate a bilinear matrix game representing the PageRank, Problem (11). Let $\tilde{A}_{ij}$ be a gain for Player A (loss of player B), if Player A plays strategy i and B plays strategy j, $1 \le i \le n$, and $1 \le j \le 2n$. Consider the loss function for Player B at step k:

$$f(x, y_k) = x^\top \tilde{A} y_k, \quad x \in \Delta_1^n,$$

where $y_k \in \Delta_1^{2n}$ is a vector with a single non-zero coordinate corresponding to the strategy of Player A. We also emphasize that y_k depends on the whole history of the game. Let C be the cost of the matrix game:

$$C = \max_{y \in \Delta_1^{2n}} \min_{x \in \Delta_1^n} y^\top \tilde{A} x = \min_{x \in \Delta_1^n} \max_{y \in \Delta_1^{2n}} y^\top \tilde{A} x = \min_{x \in \Delta_1^n} \|Ax\|_\infty = 0, \tag{12}$$

and

$$\min_{x \in \Delta_1^n} \frac{1}{N} \sum_{i=1}^N f(x, y_i) \ge C \ge \max_{y \in \Delta_1^{2n}} \frac{1}{N} \sum_{i=1}^N f(x_i, y), \tag{13}$$

for any sequences $\{x_i\}_{i=1}^N$, $\{y_i\}_{i=1}^N$ if for any i: $x_i \in \Delta_1^n$, $y_i \in \Delta_1^{2n}$, $1 \le i \le n$. In the subsequent of the section, we consider $\{(x_i, y_i)\}_{i=1}^N$ with a single non-zero coordinate each.

To solve the problem, we assume the following randomized strategy for the Player B played against any strategy of the Player A:

(1) Let $p_1 = (n^{-1}, \dots, n^{-1})$;
(2) Choose at random j_k, such that $\mathbb{P}(j_k = j) = p_j^k$;
(3) Assume $x_{j_k}^k = 1$ and $x_j^k = 0$ for all $j \ne j_k$;
(4) Update

$$p_j^{k+1} \propto p_j^k \exp\left(-\gamma_x \tilde{A}_{i_k, j}\right), \tag{14}$$

where i_k is a strategy that Player A chooses at step k.

The crucial gain in the efficiency of the algorithm is due to time-efficient updates at stage 3. Indeed, consider a binary tree with its leaves corresponding to the variables, and constructed in such a way that the value p_v assigned to a node v is a total probability of all leaves having v as a predecessor. If we update the weight of a leaf according to

Algorithm 3: UPDATE RULE FOR A PROBABILITY DISTRIBUTION

Input: (unnormalized) probability distribution p given by a binary tree with leafs values p_i such that the probability of any leaf j is $p_j / \sum_{i=1}^{n} p_i$, $p_i > 0$ for any $i: \ 1 \leq i \leq n$, $\sum_{i=1}^{n} p_i > 0$, with and update rule of coordinate k according to Eq. (14)

$$\tilde{p}_k \propto p_k \exp\left(-\psi_k\right),$$

Output: $x \sim \tilde{p}$, a sample x follows the updated distribution $\tilde{p}$

// Distribution update
$u = k$ // Start with the leaf k
while $u \neq$ *root* **do**
    $p_u \leftarrow p_u + p_k(\exp\left(-\psi_k\right) - 1)$;
    $u \leftarrow$ parent of u
// Sampling $x \sim \tilde{p}$
$u = k$ // Start with the leaf k
while $u \neq$ *leaf* **do**
    Let ν, ω be children of u

$$u = \begin{cases} \nu, & \text{with probability } p_\nu/(p_\nu + p_\omega), \\ \omega, & \text{otherwise.} \end{cases}$$

return u

Eq. (14), we also update each vertex u belonging to the path from the leaf to the root as

$$p_u = p_u + \xi, \quad \xi = p_j^{k+1} - p_j^k.$$

In order to sample $x \sim p^{k+1}$, we start from the root of the tree and proceed to its child a with probability $p_a/(p_a + p_b)$. Otherwise, we proceed to its sibling b, where p_a, and p_b are the values assigned to a and b, respectively. We repeat the same procedure for each node one a path from the root to one of the leafs of the tree. Algorithm 3 formalizes this argument.

Using the same strategy for the Player A, we establish the convergence rate to the Nash equilibrium (x^*, ω^*), which solves Problem (11) in Theorem 4.1. Algorithm 4 contains all necessary details. It is worth mentioning an interpretation of the algorithm, e.g., on each iteration it make an update following to a sparse projection of the gradient to the unit simplex according to the KL divergence.

THEOREM 4.1 *Algorithm 4 after* $N \geq 4\varepsilon^{-2}\Big(\ln(2n) + \ln(n) + 16\ln(1/\delta)\Big)$ *iterations with a constant step-size* $\gamma_x = \sqrt{2(\log n)/N}$ *and* $\gamma_y = \sqrt{2(\log(2n))/N}$ *results in a point* $(\bar{x}_N, \bar{y}_N)$ *such that, with probability at least* $1 - \delta$*, for any* $\delta > 0$ *one has:*

$$\|A\bar{x}_N\|_\infty \leq \varepsilon.$$

Algorithm 4: ℓ_∞ APPROXIMATION TO THE PAGERANK PROBLEM

Input: d-sparse transition matrix P, starting point x_0 in one of the vertices of the unit simplex, and learning rate γ

Output: x_k : $f(\bar{x}_k, \bar{y}_k) = \bar{x}_k \tilde{A} \bar{y}_k \leq \varepsilon$, implies $\|P^\top x_k - x_k\|_\infty \leq \varepsilon$

$\pi = ((2n)^{-1}, \ldots, (2n)^{-1}), \quad p = (n^{-1}, \ldots, n^{-1})$,

starting point (x_0, y_0) in one of the vertices of $\Delta_1^n \times \Delta_1^{2n}$

while $f(\bar{x}_k, \bar{y}_k) > \varepsilon$ **do**

// Player A turn

Choose at random i_k^A, such that $\mathbb{P}(i_k = j) = \pi_i$;

Assume $y_k^{i_k^A} = 1$, and $y_k^i = 0$ if $i \neq i_k^A$;

Update π_k// see Algorithm 3 for details

$$\pi_{k+1}^i \propto \pi_k^i \exp\left(\gamma_y \tilde{A}_{i, j_k^B}\right)$$

// Player B turn

Choose at random j_k^B, such that $\mathbb{P}(j_k^B = j) = p_j$;

Assume $x_k^{j_k^B} = 1$, and $x_k^j = 0$ if $j \neq j_k^B$;

Update p_k // see Algorithm 3 for details

$$p_{k+1}^j \propto p_k^j \exp\left(-\gamma_x \tilde{A}_{i_k^A, j}\right)$$

Update an average point $(\bar{x}_k, \bar{y}_k)$. Indeed, more time-efficient is to update $k \cdot (\bar{x}_k, \bar{y}_k)$ as this involves only sparse operations according to:

$$k \cdot \bar{x}_k = \sum_{t=1}^{k} x_t = (k-1)\bar{x}_{k-1} + x_k, \quad k \cdot \bar{y}_k = \sum_{t=1}^{k} y_t = (k-1)\bar{y}_{k-1} + y_k$$

an can be done in $O(d)$ time

Update the function value, $f(\bar{x}_k, \bar{y}_k)$, using sparse operations only:

$$\begin{aligned} k^2 f(\bar{x}_k, \bar{y}_k) &= (k\bar{x}_k)^\top \tilde{A}(k\bar{y}_k) \\ &= (k-1)^2 f(\bar{x}_{k-1}, \bar{y}_{k-1}) + (k-1)\bar{x}_{k-1}\tilde{A}y_k + x_k\tilde{A}((k-1)\bar{y}_k) + x_k\tilde{A}y_k \end{aligned}$$

as $k\bar{x}_k = (k-1)\bar{x}_{k-1} + x_k$, and $k\bar{y}_k = (k-1)\bar{y}_{k-1} + y_k$. The update requires $O(d)$ time.

return $(\bar{x}_k, \bar{y}_k)$

Moreover, the total running time of the algorithm is bounded from above as

$$T = O\left(n + \frac{d \log n \log \frac{n}{\delta}}{\varepsilon^2}\right).$$

The proof of the theorem is provided in Appendix B.

Discussion. The mirror descent and the dual averaging perspectives. The proposed algorithm is essentially a mirror descent with randomized projection of the gradient on

a unit simplex according to KL divergence. Let us consider the problem of minimizing of the left hand-side of Eq. (13) in more details:

$$\frac{1}{N}\sum_{i=1}^{N} f(x, y_i) = \frac{1}{N}\sum_{i=1}^{N}\langle x, A^\top J y_i\rangle \to \min_{x\in\Delta_1^n}, \tag{15}$$

where $\{y_i\}_{i=1}^n$ is a sequence of unit coordinate vectors. Denote $f_i(x) \doteq f(x, y_i)$ for $i \geq 1$. Recall the setup of the mirror descent algorithm [27]. Let $\omega(x) = \sum_{i=1}^n x^i \log x^i$ be the distance-generating function, which is is 1-strongly convex with respect to the ℓ_1 norm. A step of the dual averaging algorithm [28] with step-size γ_x is:

$$z_k = z_{k-1} - \gamma_x \nabla f_k(x_k), \quad x_{k+1} = \nabla\omega^*(z_k), \tag{16}$$

where $\omega^*(z) = \sup_{x\in\Delta_1^n}\left\{z^\top x - \omega(x)\right\} = \log\left\{\sum_{j=1}^n \exp(z_j)\right\}$. An update of x_{k+1} in Eq. (16) can be also viewed as a projection of z_{k+1} to the unit simplex in accordance with Kullback-Leibler divergence. Indeed Eq. (16) is a step of mirror descent algorithm for simplex constrained problems as well [1, Appendix A], [3], [19]. A randomized version of the update is then

$$x_{k+1} = e_i, \text{ with probability } p_{k+1}^i = \frac{p_k^i \exp\left(-\gamma_x \frac{\partial f_k(x_k)}{\partial x^i}\right)}{\sum_{t=1}^n p_k^t \exp\left(-\gamma_x \frac{\partial f_k(x_k)}{\partial x^t}\right)}, \tag{17}$$

where e_i is a unit vector with a single non-zero coordinate corresponding to index i. Since $f(x,y) = y^\top \tilde{A} x$, update (17) is the same as the update in Algorithm 4.

5. Implementation details and case study

All algorithms proposed in the paper are implemented in `C++`. We test our code with different versions of `GCC` (GNU Compiler Collection), `clang` (C language family front-end for LLVM), and `icc` (Intel C Compiler) compilers under GNU/Linux, Microsoft Windows, and Mac OS X. We conduct the experiments using:

- Ubuntu server 16.04.6 LTS, x86_64
- Intel Core i5-2500K, 16 Gb RAM
- `GCC`-5.4.0 to compile `C++` code,
- Assembly parameters: `-std=c++11 -O2 -mcmodel=small -DNDEBUG`

We test our algorithms in different dimensions using the following three test beds:

(1) d-diagonal matrix for n_d = 1,3,5,.... Each row/column of these matrices contains $(n_d - 1)/2 + 1 \leq d \leq n_d$ non-zero elements;
(2) randomly-generated matrices with d non-zero elements (on average);
(3) and web-graphs from the Stanford University graph collection [1]

We use accuracy $\varepsilon = 10^{-4}$ in each of our experiments; $x_0 = (1, 0, \ldots, 0)$ is used as a starting point for the `NL1` and `S-FW` algorithms, and we terminate the algorithms if $f(x_k) = \|Ax\|_2^2/2 \leq \varepsilon^2/2$. Computational time reported for the case study includes time

[1] `http://snap.stanford.edu/data/#web`

required by optimization method, A and $A^\top$ generating times, an initialization of all data structures used, as well as initial gradient/function computation.

web-graph		# non zeros				
		in a row		in a column		average
		min	max	min	max	
Stanford,	$n = 281\,903$	2	38 607	1	256	9.20
NotreDame,	$n = 325\,729$	2	10 722	1	3 445	5.51
BerkStan,	$n = 685\,230$	1	84 209	1	250	12.09
Google,	$n = 875\,713$	1	6 327	1	457	6.83

Table 2. Structure of the matrix A for the graphs from the Stanford web-graph collection. Columns of the transition matrix are more dense than the rows and on average both columns and rows contain a few non-zero elements only.

The numerical experiments described below allow the following conclusions to be drawn:

(1) In our experiments (see Figure 1) algorithm GK converges sufficiently fast for the desired precision. But after a large number of iterations, the value of $f(\bar{x})$ starts to grow, and the resulting point does not satisfy the accuracy condition. The reason for this is that the values of several probabilities p_i become extremely large and out of range after a number of iterations. That leads to significant numerical errors in estimating residual small probabilities and overall unsatisfactory performance of the algorithm. Rescaling the probability vectors does not change the behavior of the algorithm. The described effect decreases for larger n; refer to Fig. 1 for details.

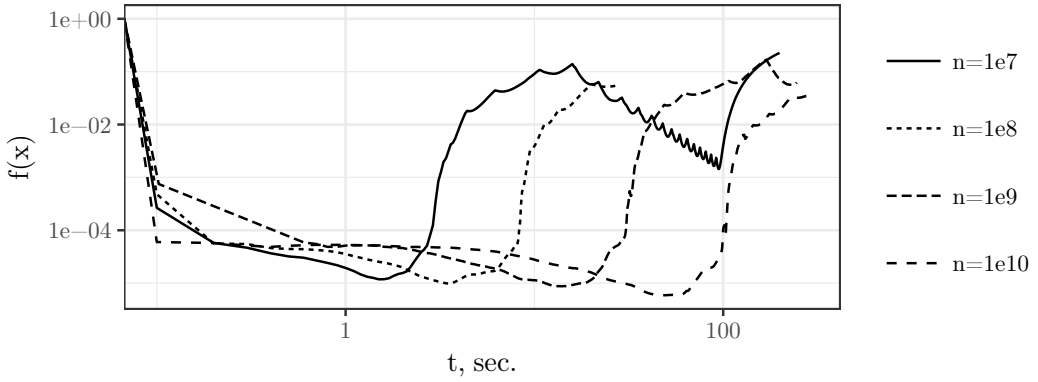


Figure 1. Convergence of the GK algorithm for various dimensions n. A is a random $n \times n$ matrix, with a number of non-zeros in each row and column $d = 3$. Practical performance of the GK algorithm is limited due to unavoidable errors in estimating small probabilities.

To summarize, the theoretical bounds for the GK algorithm differ markedly from from those seen in practical performance.

(2) Computational time for d-diagonal matrices A is much smaller than that for random matrices (see Figures 2 and 3 for details). This is due to the fast cache operations, which require far fewer memory reads for sequential data. Updates to computational trees/heaps are also performed in sequential elements, which improves time performance as well. Also, this permits the dramatic reduction of the dependence on the actual problem dimension in practice. Thus in Table 3 for $n_d = 3$ and accuracy

$\varepsilon = 10^{-4}$ the computational time has increased less than twice for $n = 10^8$ compared with $n = 10^2$.

Conversely, for the random matrices, caching does not give the same improvement in speed. This significantly decreases the actual performance of the algorithms; see Table 4 for details.

	NL1		S-FW	
n	time, sec.	iteration	time	iterations
$n_d = 3$; $2 \le d \le 3$				
10^2	4.089	3 948 632	0.007	14 142
10^3	4.221	3 950 392	0.008	14 142
10^4	4.575	3 950 392	0.009	14 142
10^5	4.814	3 950 392	0.010	14 142
10^6	5.143	3 950 392	0.010	14 142
10^7	5.566	3 950 392	0.010	14 142
10^8	6.021	3 950 392	0.010	14 142
$n_d = 11$; $6 \le d \le 11$				
10^2	14.655	2 100 964	0.041	14 749
10^3	37.796	5 101 072	0.041	16 956
10^4	39.170	5 101 072	0.062	19 995
10^5	39.897	5 101 072	0.064	24 495
10^6	41.004	5 101 072	0.065	24 495
10^7	43.917	5 101 072	0.068	24 495
$n_d = 51$; $26 \le d \le 51$				
10^3	529.240	5 216 119	1.552	46 447
10^4	535.348	5 216 119	1.045	29 991
10^5	537.419	5 216 119	1.741	49 235
10^6	549.782	5 216 119	1.758	49 235
10^7	552.271	5 216 119	1.789	49 235
$n_d = 101$; $51 \le d \le 101$				
10^4	1 935.198	5 175 085	6.464	49 925
10^5	1 962.307	5 175 085	9.097	68 646
10^6	1 940.331	5 175 085	9.134	68 646

Table 3. Time in seconds required to solve the PageRank problem. A is a d-diagonal matrix. The S-FW algorithm outperforms the NL1 algorithm for most of the instances and have better scalability with the dimension of the problem.

(3) Surprisingly, the time complexity of the S-FW algorithm for the Stanford web-graph collection is much less than that for the NL1 algorithm (see Table 5). Unfortunately, for two problems on the list, the NL1 algorithm performance is not sufficiently high. We propose that this is due to the fact that the NL1 algorithm modifies two variables per iteration and often involves very expensive, dense updates compared with to the S-FW algorithm (see Table 2 for the information about the sparsity of the transition matrices). Table 7 contains information about the average iteration complexity of the NL1 algorithm; it is much higher than that for the S-FW algorithm, which supports our conjecture, particularly for the web-BerkStan dataset. Recall that this property is true without additional matrix sparsification (refer to Section 2).

The performance of the S-FW and NL1 algorithms is shown in Figures 4–9.

We also implement and test some other non-sparse (NS) methods and non-sparse versions of our S-FW and NL1 methods. Non-sparse versions do not perform any sparse

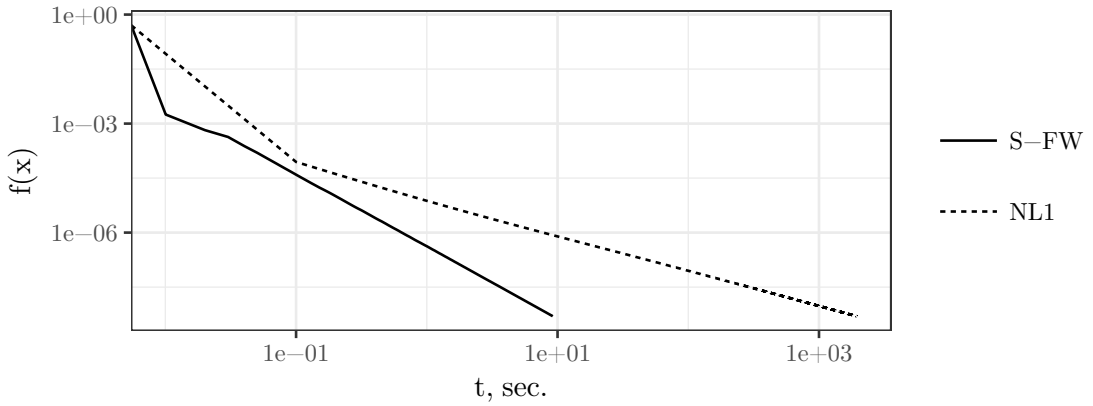


Figure 2. Computational time for the PageRank problem, A is a n_d-diagonal matrix, dimension $n = 10^6$, number of non-zero diagonals $n_d = 101$. The `S-FW` algorithm significantly outperforms the `NL1` algorithms for large scale problems.

	`NL1`		`S-FW`	
n	time	iterations	time	iterations
$d = 3$				
10^2	0.003	1 999	0.023	39 734
10^3	0.031	17 748	0.118	190 601
10^4	0.233	141 739	0.414	632 954
10^5	2.374	840 617	2.107	2 009 854
10^6	16.171	4 020 388	9.355	6 203 826
10^7	56.694	11 669 495	32.442	17 916 520
10^8	173.070	19 988 053	121.258	43 390 838
$d = 11$				
10^2	0.013	590	0.173	44 706
10^3	0.072	5 106	0.593	142 109
10^4	0.568	40 029	2.123	450 873
10^5	6.342	299 382	10.374	1 482 735
10^6	78.383	2 025 423	60.715	4 753 809
10^7	503.385	11 272 158	219.988	14 693 667
$d = 51$				
10^3	0.891	3 851	11.681	162 015
10^4	8.383	31 372	42.824	510 444
10^5	77.137	241 191	164.751	1 621 686
10^6	1 300.194	1 683 845	1 152.805	5 082 774
10^7	11 250.461	10 627 974	5 432.107	17 479 622
$d = 101$				
10^4	29.540	29 127	168.124	529 685
10^5	304.419	225 146	650.878	1 696 708
10^6	4 692.729	1 607 834	4 619.220	5 267 738

Table 4. Time in seconds required to solve the PageRank problem. A is a random matrix. The `NL1` algorithm outperforms the `S-FW` algorithm in most of the sparse and low-dimensional instances, while the `S-FW` algorithm is preferable for large scale cases.

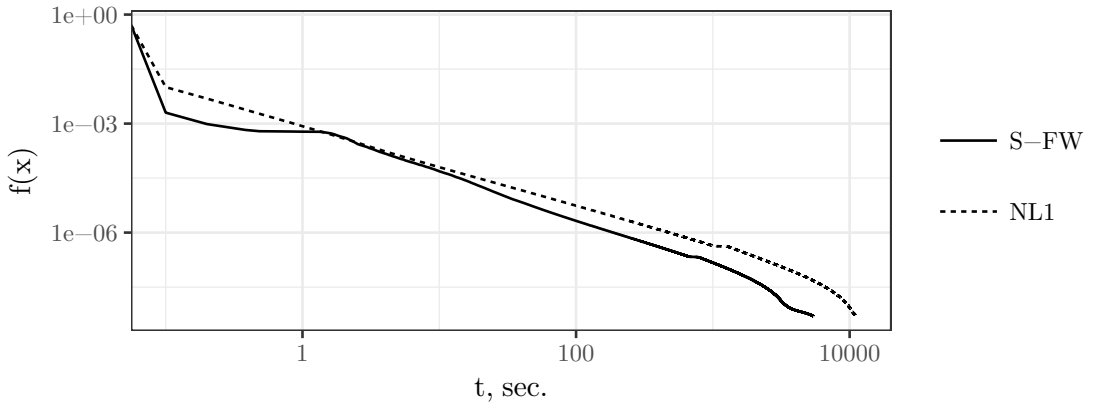


Figure 3. Time complexity for the PageRank problem. A is a random matrix, dimension $n = 10^7$, average number of non-zeros in each row and column $d = 51$. The `NL1` and `S-FW` algorithms have almost the same computational time.

		`NL1`		`S-FW`	
web-graph	n	time, sec.	iterations	time, sec.	iteration
Stanford	281 903	0.145	93 152	0.008	14 142
NotreDame	325 729	700.810	3 816 436	0.526	38 014
BerkStan	685 230	38 161.847	12 315 700	0.536	19 990
Google	875 713	113.643	1 083 996	0.278	37 313

Table 5. Time in seconds required to solve the PageRank problem for web-graphs from the Stanford graph collection. The `S-FW` algorithm achieves significantly better time performance compared to the `NL1` algorithm.

	Stanford, $n = 281903$		Google, $n = 875713$	
method	time, sec.	iterations	time, sec.	iteration
`S-FW`	0.008	14 142	0.278	37 313
`S-FW(NS)`	75.438	14 142	451.131	38 672
`NL1`	0.145	93 152	113.643	1 083 996
`NL1(NS)`	458.493	93 152	13 507.423	1 220 868
`FGM(NS)`	82.978	12 464	423.008	22 811

Table 6. Time in seconds required to solve the PageRank problem for web-graphs with sparse and non-sparse (NS) versions of `S-FW` and `NL1` methods. We compare sparse versions of the algorithms with the non-sparse ones (`NS`). We have compared our algorithms with the state-of-the-art Similar Triangles Algorithm [15, 33], , which is essentially an extension to the Fast Gradient method (FGM).

updates and replace them with "classic" full-update operations, so we can check not only the speedup of proposed sparse updates but also it's computational accuracy and stability.

We implement a standard non-sparse Projected Gradient Method (PG) [33], and the Similar Triangles algorithm [15, 33], which we further refer as `FGM-NS`. Test results prove the accuracy and stability of our sparse methods – in all cases proposed methods outperforms non-sparse ones and operates very closely it's "full" versions (`S-FW` vs. `S-FW(NS)` for example). Details are provided in Table 6, Figures 4, 5, 8, and 9.

# elements		Stanford		BerkStan	
		NL1	FW	NL1	FW
d_r	min	1.0	1.0	1.0	1.0
	max	34.0	4.0	84 209.0	84 209.0
	average	3.9	3.9	2 278.4	148.6
d_c	min	2.0	2.0	1.0	1.0
	max	37.0	3.0	244.0	83.0
	average	2.9	2.8	15.7	6.2
$d_r \cdot d_c$	min	3.0	3.0	2.0	2.0
	max	1 258.0	12.0	15 494 456.0	6 989 347.0
	average	11.7	11.3	84 304.3	7 507.5

Table 7. Iteration complexity for the Stanford graph collection. The `S-FW` algorithm has performed a much fewer number of updates, resulting in a higher performance compared to the `NL1` algorithm.

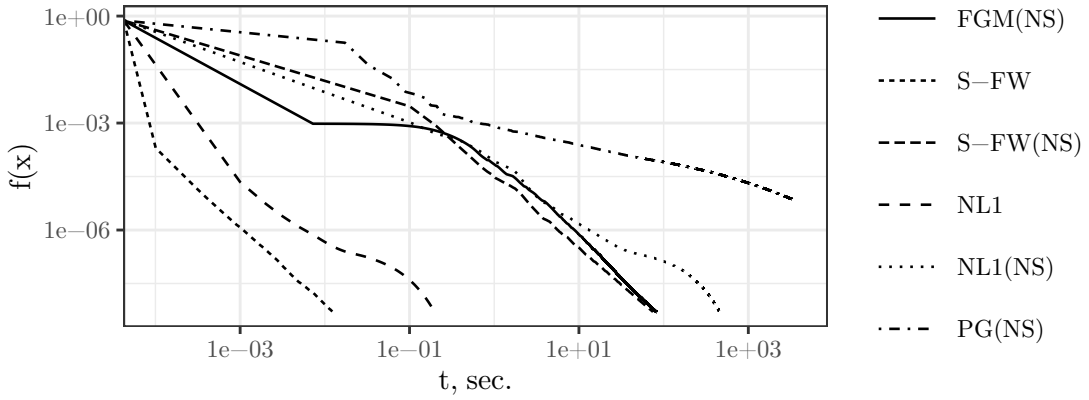


Figure 4. Time complexity for PageRank over the web-Stanford dataset. The `S-FW` is fastest one for both sparse and non-sparse (NS) method versions. Sparse algorithms significantly mostly outperforms non-sparse versions.

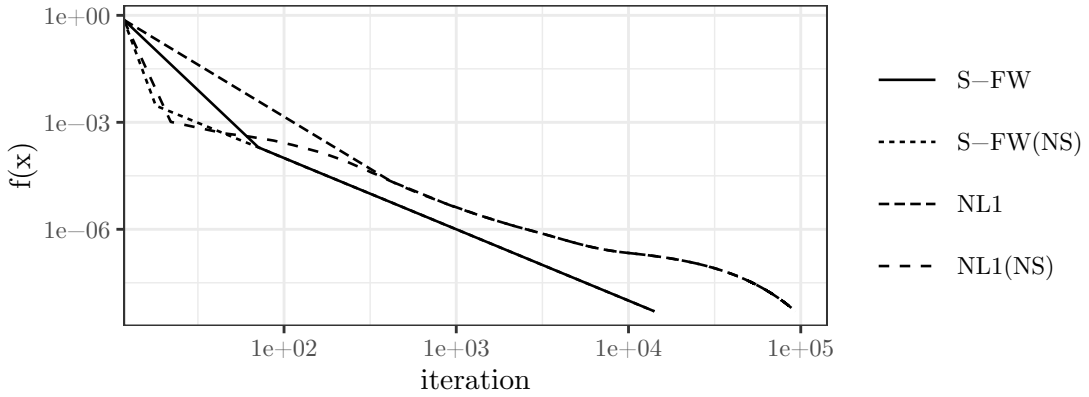


Figure 5. Iteration complexity for PageRank over the web-Stanford dataset. Sparse methods have almost the same iteration complexity as the non-sparse ones.

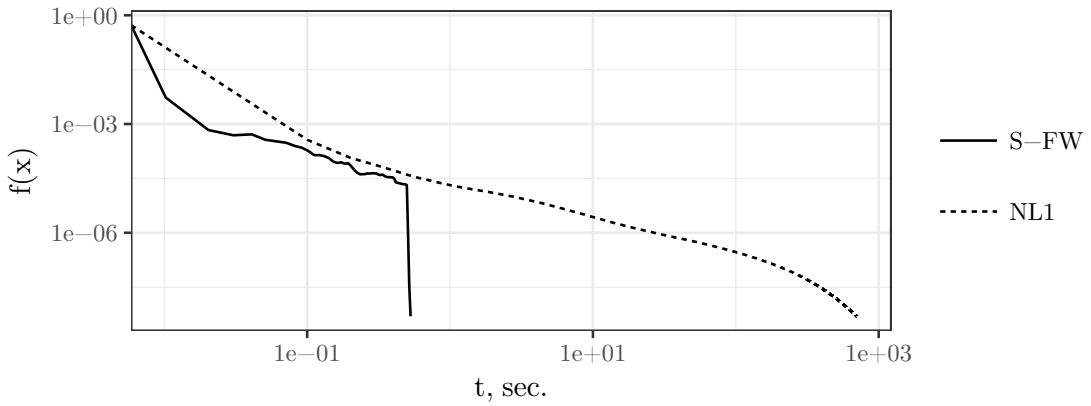


Figure 6. Time complexity for PageRank over the web-BerkStan dataset. The `S-FW` algorithm significantly outperforms the `NL1` algorithm and has a significant gain in the vicinity of the optimal point.

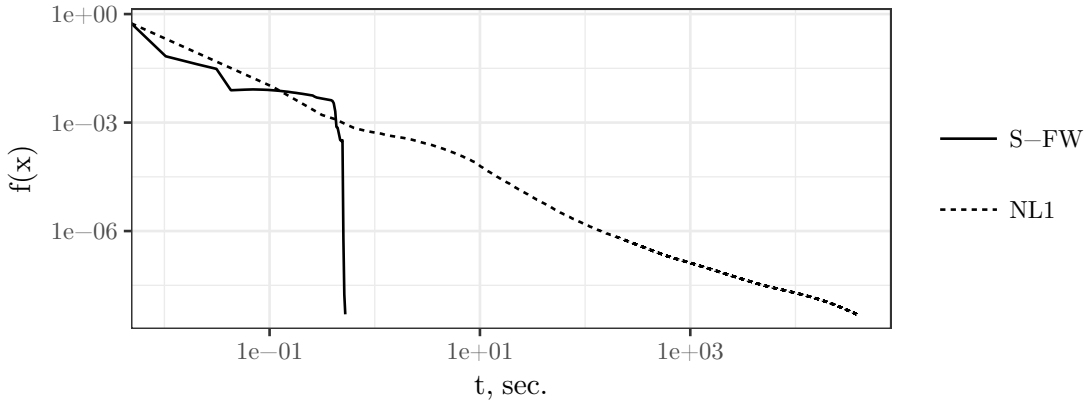


Figure 7. Time complexity for PageRank over the web-BerkStan dataset. The `S-FW` algorithm significantly outperforms the `NL1` algorithm and has a significant gain in the vicinity of the optimal point.

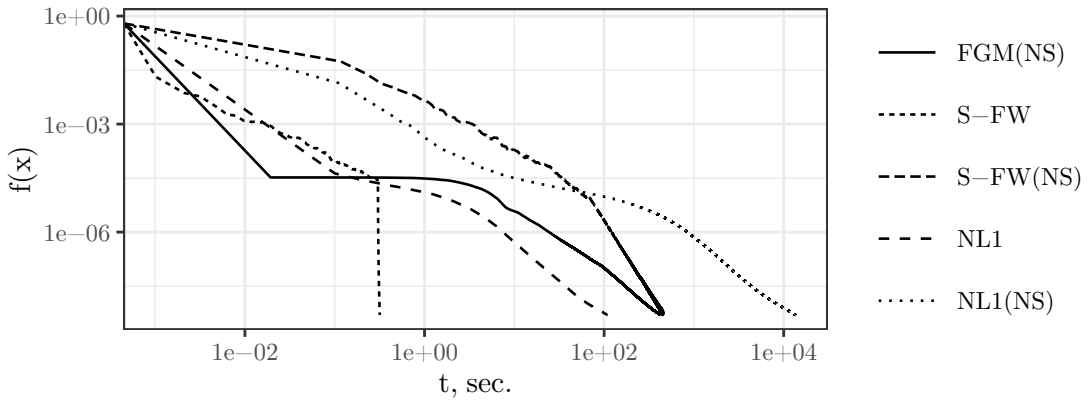


Figure 8. Time complexity for PageRank over the web-Google dataset. Sparse `S-FW` method is the fastest among compared. The convergence of the Similar triangles algorithm (`FGM-NS`) [15, 33] is faster in terms of the number of iterations, but slower in terms of the computational time.

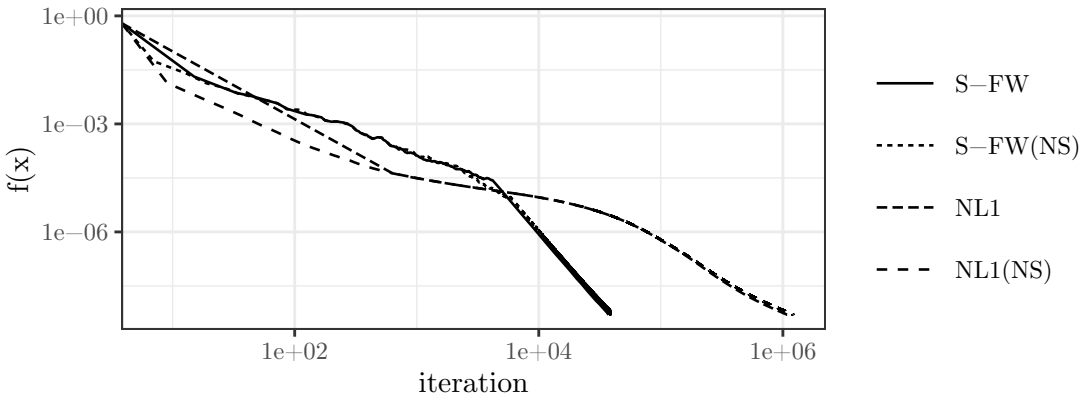


Figure 9. Iteration complexity for PageRank over the web-Google dataset. The convergence of method's sparse versions is very close to the non-sparse (NS) ones.

6. Conclusion

In this paper, we have proposed three novel algorithms to solve the PageRank problem. All the algorithms can be viewed as guided versions of coordinate or block-coordinate descent and demonstrate superior practical performance. In further works, the authors intend to devote more attention to sparsification techniques and interplay between the problem sparsity, dimension, and desired accuracy.

References

[1] Z. Allen-Zhu and L. Orecchia, *Linear coupling: An ultimate unification of gradient and mirror descent*, in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 2017, p. 3:1–3:22.

[2] K. Avrachenkov, N. Litvak, D. Nemirovsky, and N. Osipova, *Monte carlo methods in pagerank computation: When one iteration is sufficient*, SIAM Journal on Numerical Analysis 45 (2007), pp. 890–904.

[3] A. Banerjee, S. Merugu, I.S. Dhillon, and J. Ghosh, *Clustering with bregman divergences*, Journal of machine learning research 6 (2005), pp. 1705–1749.

[4] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[5] S. Brin and L. Page, *Reprint of: The anatomy of a large-scale hypertextual web search engine*, Computer networks 56 (2012), pp. 3825–3833.

[6] S. Bubeck, *et al.*, *Convex optimization: Algorithms and complexity*, Foundations and Trends® in Machine Learning 8 (2015), pp. 231–357.

[7] E.J. Candes, M.B. Wakin, and S.P. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, Journal of Fourier analysis and applications 14 (2008), pp. 877–905.

[8] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*, Cambridge university press, 2006.

[9] M.B. Cohen, J. Kelner, J. Peebles, R. Peng, A.B. Rao, A. Sidford, and A. Vladu, *Almost-linear-time algorithms for Markov chains and new spectral primitives for directed graphs*, in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 2017, pp. 410–419.

[10] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, *Introduction to algorithms, second edition* (2001).

[11] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics (NRL) 3 (1956), pp. 95–110.

[12] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, Vol. 1, Springer series in statistics. New York, 2001.

[13] A. Gasnikov and D. Dmitriev, *On efficient randomized algorithms for finding the pagerank vector*, Computational Mathematics and Mathematical Physics 55 (2015), p. 349.

[14] A. Gasnikov, Y. Nesterov, and V. Spokoiny, *On the efficiency of a randomized mirror descent algorithm in online optimization problems*, Computational Mathematics and Mathematical Physics 55 (2015), pp. 580–596.

[15] A.V. Gasnikov and Y.E. Nesterov, *Universal method for stochastic composite optimization problems*, Computational Mathematics and Mathematical Physics 58 (2018), pp. 48–64.

[16] M.D. Grigoriadis and L.G. Khachiyan, *A sublinear-time randomized approximation algorithm for matrix games*, Operations Research Letters 18 (1995), pp. 53–58.

[17] Z. Harchaoui, A. Juditsky, and A. Nemirovski, *Conditional gradient algorithms for norm-regularized smooth convex optimization*, Mathematical Programming 152 (2015), pp. 75–112.

[18] M. Jaggi, *Revisiting Frank-Wolfe: projection-free sparse convex optimization*, in *Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML)*, 2013, pp. 427–435.

[19] A. Juditsky and A. Nemirovski, *First order methods for nonsmooth convex large-scale optimization, I: general purpose methods*, Optimization for Machine Learning (2011), pp. 121–148.

[20] S. Kamvar, T. Haveliwala, and G. Golub, *Adaptive methods for the computation of pagerank*, Linear Algebra and its Applications 386 (2004), pp. 51–65.

[21] S.D. Kamvar, T.H. Haveliwala, C.D. Manning, and G.H. Golub, *Extrapolation methods for accelerating PageRank computations*, in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 261–270.

[22] A.N. Langville and C.D. Meyer, *Google's PageRank and beyond: The science of search engine rankings*, Princeton University Press, 2011.

[23] D.A. Levin and Y. Peres, *Markov chains and mixing times*, Vol. 107, American Mathematical Soc., 2017.

[24] D.L. Logan, *A first course in the finite element method*, Cengage Learning, 2011.

[25] R.B. Myerson, *Game theory*, Harvard university press, 2013.

[26] A.V. Nazin and B.T. Polyak, *Randomized algorithm to determine the eigenvector of a stochastic matrix with application to the pagerank problem*, Automation and Remote Control 72 (2011), pp. 342–352.

[27] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on optimization 19 (2009), pp. 1574–1609.

[28] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Mathematical programming 120 (2009), pp. 221–259.

[29] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization 22 (2012), pp. 341–362.

[30] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, Vol. 87, Springer Science & Business Media, 2013.

[31] Y. Nesterov, *Subgradient methods for huge-scale optimization problems*, Mathematical Programming 146 (2014), pp. 275–297.

[32] Y. Nesterov, *Complexity bounds for primal-dual methods minimizing the model of objective function*, Mathematical Programming (2015), pp. 1–20.

[33] Y. Nesterov, *Lectures on convex optimization*, Vol. 137, Springer, 2018.

[34] Y. Nesterov and A. Nemirovski, *Finding the stationary states of markov chains by iterative methods*, Applied Mathematics and Computation 255 (2015), pp. 58–65.

[35] J. Nutini, M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke, *Coordinate descent converges faster with the Gauss-Southwell rule than random selection*, in *International Conference on Machine*

Learning (ICML), 2015, pp. 1632–1641.

[36] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank citation ranking: Bringing order to the web.*, Tech. Rep., Stanford InfoLab, 1999.

[37] B.T. Polyak and A.A. Tremba, *Regularization-based solution of the pagerank problem for large matrices*, Automation and Remote Control 73 (2012), pp. 1877–1894.

[38] A.D. Sarma, A.R. Molla, G. Pandurangan, and E. Upfal, *Fast distributed pagerank computation*, Theoretical Computer Science 561 (2015), pp. 113–121.

[39] B. Stott, J. Jardim, and O. Alsaç, *Dc power flow revisited*, IEEE Transactions on Power Systems 24 (2009), pp. 1290–1300.

[40] H. Tong, C. Faloutsos, and J.Y. Pan, *Fast random walk with restart and its applications*, in *Sixth International Conference on Data Mining (ICDM'06)*, IEEE, 2006, pp. 613–622.

[41] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, *Fast accurate computation of large-scale IP traffic matrices from link loads*, in *ACM SIGMETRICS Performance Evaluation Review*, ACM, 2003, pp. 206–217.

Appendix A. Missing Proofs in Section 2

Recall the definition of $f_\gamma(x)$:

$$f_\gamma(x) = \frac{1}{2}\|Ax\|_2^2 + \sum_{i=1}^{n} \frac{\gamma}{2}\left(-x^i\right)_+^2.$$

LEMMA A.1 *Let x_* satisfies $e^\top x^* = 1$ and $f_\gamma(x_*) \le \varepsilon^2$ for some $\gamma > 0$. Then for $\hat{x} = (x_*)_+/e^\top(x_*)_+$ we have*

$$\|A\hat{x}\|_2^2 \le 4(1 + 4\gamma^{-1})\varepsilon^2$$

and $\hat{x} \in \Delta_1^n = \{x \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1,\ x_i \ge 0\}$

Proof. By the conditions of the lemma

$$\frac{1}{2}\|Ax_*\|_2^2 + \sum_{i=1}^{n} \frac{\gamma}{2}\left(-x_*^i\right)_+^2 \le \varepsilon^2. \tag{A1}$$

Let $x = (x_*)_+ - (-x_*)_+$, then by the triangle inequality we have

$$\|(A(x_*)_+)\|_2 \le \|Ax_*\|_2 + \|(A(-x_*)_+)\|_2 \le \sqrt{2}\varepsilon + \|(A(-x_*)_+)\|_2. \tag{A2}$$

By the Perron-Frobenius theorem, $|\lambda_i(P^\top)| \le 1$, $1 \le i \le n$. Thus $\lambda_{\max}(A) = \lambda_{\max}(I - P^\top) \le 2$. By Inequality A1 one has

$$\|A(-x_*)_+\|_2 \le \|(-x_*)_+\|_2 \le 2\sqrt{\frac{2}{\gamma}}\varepsilon. \tag{A3}$$

Using Inequalities A1, A2 and A3 we have the final estimate

$$\|A(x_*)_+\|_2 \le \sqrt{2}\varepsilon + \|A(-x_*)_+\|_2 \le \sqrt{2}\varepsilon + 2\sqrt{\frac{2}{\gamma}}\varepsilon.$$

By definition of $\hat{x}$, we have $\|A\hat{x}\|_2((x_*)_+^\top e) = \|A(x_*)_+\|_2$. Since $e^\top((x_*)_+ - (-x_*)_+) = 1$, $e^\top x_* = 1$, and $(-x_*)_+^\top e \ge 0$ we have $(x_*)_+^\top e \ge 1$ and $\|A\hat{x}\|_2 \le \|A(x_*)_+\|_2$. An application of the Cauchy-Schwartz inequality concludes the proof of the lemma. ■

Proposition B.1 of [1] establishes convergence rate of the gradient descent in arbitrary norm.

Proposition 1 (Proposition B.1 of [1]) Let $f(x)$ be a convex, differentiable function that is L-smooth with respect to $\|\cdot\|$ on $Q = \mathbb{R}^n$, and x_0 any initial point in Q. Consider the sequence of N gradient steps $x_{k+1} = \operatorname{argmin}_{y \in Q}\left\{\frac{L}{2}\|y - x\|^2 + \nabla f(x_k)^\top(y - x)\right\}$, then the last point x_N satisfies

$$f(x_N) - f(x_*) \le 2\frac{LR^2}{N},$$

where $R = \max_{x:\, f(x) \le f(x_0)} \|x - x_*\|$, and x_* is any minimizer of f.

Now we are ready to proof Theorem 2.2.

Proof of the Theorem 2.2. A single iteration of Algorithm 1 results in a sparse update vector h containing at most two non-zero coordinates which correspond to the minimal and the maximal coordinates of the gradient (see Lemma 2.1 for the details). Then the gradient update for $x_{k+1} = x_k + h_k$ is

$$\nabla f_\gamma(x_k + h_k) = \nabla f_\gamma(x_k) + \eta A^\top A h_k + \gamma(-x_k - h_k)_+ - \gamma(-x_k)_+$$

and requires $O(d^2 \log(n/d^2 + 2))$ arithmetic operations by using a set of $\lceil n/d^2 \rceil$ binary heaps described earlier in Section 2. An update to the function value is

$$f_\gamma(x_k + h_k) = f_\gamma(x_k) + 2h_k^\top A x_k + \|Ah_k\|_2^2/2 + \frac{\gamma}{2}(-x_k - h_k)_+^2 - \frac{\gamma}{2}(-x_k)_+^2$$

and similarly requires at most $O(d^2 \log(n/d^2 + 2))$ operations as h_k contains at most 2 non-zero coordinates. Notice that the size of the level set

$$R = \max_{x:\, f(x) \le f(x_0)} \|x - x_*\|_1$$

at a point x is bounded from above as $R \le 4\sqrt{2nf_\gamma(x_0)/\gamma} + 2$. Indeed

$$\begin{aligned} \frac{\gamma}{2}\frac{\|(-x)_+\|_1^2}{n} \le \frac{\gamma}{2}\|(-x)_+\|_2^2 &= \frac{\gamma}{2}\sum_{i=1}^n (-x^i)_+^2 \\ &\le \frac{1}{2}\|Ax\|_2^2 + \frac{\gamma}{2}\sum_{i=1}^n (-x^i)_+^2 = f_\gamma(x) \le f_\gamma(x_0) \end{aligned} \tag{A4}$$

and

$$\begin{aligned} R = \max_{x:\, f(x) \le f(x_0)} \|x - x_*\|_1 &\le \max_{x: f_\gamma(x) \le f_\gamma(x_0)} 2\|x\|_1 \\ &\le 4\|(-x)_+\|_1 + 2 \le 4\sqrt{\frac{2nf_\gamma(x_0)}{\gamma}} + 2, \end{aligned} \tag{A5}$$

where $\|x\|_1 \le \|(-x)_+\|_1 + \|(x)_+\|_1 \le 2\|(-x)_+\|_1 + 1$ since $\sum_{i=1}^n x^i = 1$.

The remainder of the proof will consists of two phases. First, we estimate the time complexity of the algorithm to achieve $f_\gamma(x_{(k)}) \le \gamma/(8n)$. After that, starting from $x_{(k)}$: $f_\gamma(x_{(k)}) \le \gamma/(8n)$ we find the complexity of the algorithm to achieve $f_\gamma(x_m) \le \varepsilon^2$.

Now fix any $\delta > 0$ such that $n\delta^2 < 1$, and let $\varepsilon_k^2 = (\delta^2 n)^k f_\gamma(x_0)$, $k \ge 1$. To achieve $f_\gamma(x_{(1)}) \le n\delta^2 f_\gamma(x_0)$ one needs according to Proposition 1 and Eq. (A5) at most:

$$T_{(1)} = 16\frac{(1+\gamma)(1 + 8nf_\gamma(x_0)/\gamma)}{\varepsilon_1^2} = 16\frac{(1+\gamma)(1/n + 8f_\gamma(x_0)/\gamma)}{\delta^2} \le 256\frac{(1+\gamma)f_\gamma(x_0)}{\gamma\delta^2}$$

iterations. Similarly, accuracy $f(x_{(k)}) \le \varepsilon_k^2$ can be achieved in

$$T_{(k)} \le 16\frac{(1+\gamma)(1 + 8nf_\gamma(x_{(k)})/\gamma)}{\varepsilon_k^2} \le 256\frac{(1+\gamma)nf_\gamma(x_{(k-1)})}{\gamma\varepsilon_k^2}$$

$$= 256\frac{(1+\gamma)n(\delta^2 n)^{k-1}f_\gamma(x_0)}{\gamma\delta^{2k}n^k} = 256\frac{(1+\gamma)f_\gamma(x_0)}{\gamma\delta^2} = T_{(1)}$$

starting from $x_{(k-1)}$, s.t. $f_\gamma(x_{(k-1)}) \le \varepsilon_{k-1}^2 f_\gamma(x_0)$.

Then $f_\gamma(x_{(k)}) \le \gamma/(8n)$ in at most k restarts:

$$8n\gamma f_\gamma(x_0)/\gamma \le 8n(\delta^2 n)^k f_\gamma(x_0)/\gamma \ \le \ 1,$$

e.g. $k = O(\log(nf_\gamma(x_0)/\gamma)/\log(n\delta^2))$. Thus the overall number of iterations is

$$T = \sum_{i=1}^{k} T_{(i)} = O\left(\frac{\log(nf_\gamma(x_0)/\gamma)}{\delta^2\log(n\delta^2)}\right).$$

The remaining time required to solve the problem starting with $x_{(k)}$ is bounded from above by Proposition 1 as $32(1+\gamma)/\varepsilon^2$. Thus, the overall complexity of the algorithm does not exceed

$$T = O\left(n + d^2\log(n/d^2+2)\left(\frac{1+\gamma}{\varepsilon^2} + \inf_{\delta: n\delta^2<1}\left[\frac{\log(nf_\gamma(x_0)/\gamma)}{\delta^2\log(n\delta^2)}\right]\right)\frac{(1+\gamma)^2}{\gamma}\right). \quad \text{(A6)}$$

by Lemma A.1. Taking $\gamma = O(1)$, and notice that $f_\gamma(x_0) = 1$ if x_0 is a one of the simplex vertices one has

$$\begin{aligned} T &= O\left(n + d^2\log(n/d^2+2)\left(\frac{1}{\varepsilon^2} + \inf_{\delta: n\delta^2<1}\left[\frac{\log n}{\delta^2\log(n\delta^2)}\right]\right)\right) \\ &= O\left(n + \frac{d^2\log(n/d^2+2)}{\varepsilon^2} + nd^2\log(n/d^2+2)\log n\right), \end{aligned}$$

which completes the proof of the theorem. ■

Appendix B. Missing Proofs in Section 4

Theorem 4.1 provides us with an upper bound on the efficiency of this strategy. Here we assume that the number of iterations N is known in advance. Let us emphasize that Algorithm 4 is a version of the Mirror Descent algorithm with randomized projecting aims to support sparse updates (see Section 4 for details).

Our main tool below is Proposition 2, which establishes the convergence rate of stochastic online optimization for linear functions f_k linear in x. We refer to recent results in [14] for a more general problem setup.

In the proof of the following proposition we mostly follow [8] and [14]. Our proof is based on the recent results for dual averaging method [28] that gives essentially the same sequence of steps as the mirror descent for simplex constrained convex optimization problems [1].

Proposition 2 Let $\{f_k(x) = x^\top A y_k\}_{k=1}^N$ be a set of functions $f : \mathbb{R}^n \to \mathbb{R}$ of variable x such that $\|\nabla f_k(x)\|_\infty \le M$ almost surely. Then for a constant step-size policy $\gamma =$

$M^{-1}\sqrt{2\log n/N}$ we have

$$\psi_N \doteq \frac{1}{N}\sum_{k=1}^{N} f_k(x_k)$$

$$\mathbb{E}_N\left[\psi_N\right] - \min_{x\in\Delta_1^n} \frac{1}{N}\sum_{k=1}^{N} f_k(x) \le M\sqrt{\frac{2\log n}{N}}.$$

where x_k is given as

$$x_k = e_i, \text{ with probability } p_k^i = \frac{p_{k-1}^i \exp\left(-\gamma \frac{\partial f_{k-1}(x_{k-1})}{\partial x^i}\right)}{\sum_{t=1}^n p_{k-1}^t \exp\left(-\gamma \frac{\partial f_{k-1}(x_{k-1})}{\partial x^t}\right)},\ k \ge 1 \tag{B1}$$

and $p_0^i = 1/n$ for all i, $1 \le i \le n$, and the expectation $\mathbb{E}_N$ is taken over the choice of $x_1, \ldots, x_N$. Moreover, for any $\Omega > 0$

$$Prob\left[\psi_N > M\sqrt{\frac{2}{N}}\left(\sqrt{\log n} + 2\sqrt{\Omega}\right)\right] \le \exp(-\Omega).$$

Proof. Let $\omega(x) = \sum_{i=1}^n x_i \log x_i$ be a distance generating functions, and $\omega^*(z) = \log\left\{\sum_{j=1}^n \exp(z^j)\right\}$ be its conjugate. Then the step of the dual averaging algorithm gives:

$$z_k = z_{k-1} - \gamma \nabla f_k(x_k), \tag{B2}$$
$$x_{k+1} = \nabla\omega^*(z_k), \tag{B3}$$

Instead, our update rule uses a randomised projection x_{k+1} of z_k on a unit simplex, according to Eq. (B1).

Let $z_{k,t} = z_k t + (1-t)z_{k-1}$, $t \in \mathbb{R}$, then

$$\begin{aligned}
\omega^*(z_k) &= \omega^*(z_{k-1}) + \int_0^1 (z_k - z_{k-1})^\top \nabla\omega^*(tz_k + (1-t)z_{k-1})dt \\
&= \omega^*(y_{k-1}) - \gamma \nabla f_k(x_k)^\top \nabla\omega^*(z_{k-1}) \\
&\quad - \gamma \nabla f_k(x_k) \int_0^1 \left[\nabla\omega^*(z_{k,t}) - \nabla\omega^*(z_{k-1})\right] dt \\
&\le \omega^*(z_{k-1}) - \gamma \nabla f_k(x_k)^\top \nabla\omega^*(z_{k-1}) \\
&\quad + \gamma \|\nabla f_k(x_k)\|_\infty \int_0^1 \left\|\nabla\omega^*(z_{k,t}) - \nabla\omega^*(z_{k-1})\right\|_1 dt,
\end{aligned} \tag{B4}$$

where the last is due to Hoelder's inequality. By the 1-strong convexity of $\omega(x)$ with respect to the ℓ_1-norm we have

$$\left\|\nabla\omega^*(z') - \nabla\omega^*(z)\right\|_1 \le \|z' - z\|_\infty.$$

Then by Inequality (B4) and (B2) we have

$$\begin{aligned}\omega^*(z_k) &\leq \omega^*(z_{k-1}) - \gamma \nabla f_k(x_k)^\top \nabla \omega^*(z_{k-1}) + \frac{\gamma^2 \|\nabla f_k(x_k)\|_\infty^2}{2} \\ &\overset{\text{(B3)}}{\leq} \omega^*(z_{k-1}) - \gamma \nabla f_k(x_k)^\top x_k + \frac{\gamma^2 \|\nabla f_k(x_k)\|_\infty^2}{2}.\end{aligned} \tag{B5}$$

Next summing up Ineq. (B5) for all k, $1 \leq k \leq N$, we have

$$\gamma \sum_{k=1}^{N} x_k^\top \nabla f_k(x_k) \leq -\omega^*(z_N) + \omega^*(z_0) + \frac{\gamma^2}{2} \sum_{k=1}^{N} \|\nabla f_k(x_k)\|_\infty^2 .$$

As $z_0 = 0$ and (B2) we get

$$\gamma \sum_{k=1}^{N} (x_k - x)^\top \nabla f_k(x_k) \leq -\omega^*(z_N) + \omega^*(z_0) + x^T y_N + \frac{\gamma^2}{2} \sum_{k=1}^{N} \|\nabla f_k(x_k)\|_\infty^2 .$$

From $\omega^*(z_0) = 0$ and Young's inequality $\omega^*(z) + \omega(x) \geq x^T z$ we get

$$\gamma \sum_{k=1}^{N} (x_k - x)^\top \nabla f_k(x_k) \leq \omega(x) + \frac{\gamma^2}{2} \sum_{k=1}^{N} \|\nabla f_k(x_k)\|_\infty^2 .$$

From convexity of f we get

$$\gamma \sum_{k=1}^{N} (f_k(x_k) - f_k(x)) \leq \omega(x) + \frac{\gamma^2}{2} \sum_{k=1}^{N} \|\nabla f_k(x_k)\|_\infty^2 .$$

Taking expectation $\mathbb{E}_N(\cdot)$ with respect to $x_1, \ldots, x_N$ (e.g. the choice of $i_1^A, \ldots, i_N^A$), we have

$$\gamma \sum_{k=1}^{N} \mathbb{E}_N \ [f_k(x_k) - f_k(x)] \leq \omega(x) + \frac{\gamma^2}{2} \sum_{k=1}^{N} \mathbb{E}_N \ \left[\|\nabla f_k(x)\|_\infty^2\right].$$

To finish the proof it remains to note that $\|\nabla f_k(x_k)\|_\infty \leq M$ and

$$\psi_N \leq \min_{\gamma > 0} \frac{\log n}{N\gamma} + \frac{M^2 \gamma}{2} = M\sqrt{\frac{2 \log n}{N}}.$$

The remainder of the proof relies on Azuma's inequality. Let

$$Z_j = \sum_{k=1}^{j} \gamma (x - x_k)^\top \nabla f_k(x_k)$$

is a Martingale satisfying $|Z_{j+1} - Z_j| \leq c_j \doteq 4M\gamma$ almost surely. By Azuma's inequality

we have

$$Prob[Z_N \geq t] \leq \exp\left(-\frac{t^2}{2\sum_{j=1}^{N} c_j^2}\right).$$

Setting $t = 4M\gamma\sqrt{2\Omega N}$ finishes the proof of the proposition. ■

Now we are ready to proof the Theorem 4.1.

Proof of the Theorem 4.1. For a point $(\bar{x}_N, \bar{y}_N)$ defined by Algorithm 4 we have:

$$\begin{aligned}
0 \leq ||A\overline{x}_N||_\infty &\overset{(11)}{=} \max_{y\in\Delta_1^{2n}} \langle y, \widetilde{A}\overline{x}_N\rangle \overset{(12)}{=} \max_{y\in\Delta_1^{2n}} \langle y, \widetilde{A}\overline{x}_N\rangle - \max_{y\in\Delta_1^{2n}} \min_{x\in\Delta_1^{n}} \langle y, \widetilde{A}x\rangle \\
&\leq \max_{y\in\Delta_1^{2n}} \langle y, \widetilde{A}\overline{x}_N\rangle - \min_{x\in\Delta_1^{n}} \langle \overline{y}_N, \widetilde{A}x\rangle \\
&= \left\{\max_{y\in\Delta_1^{2n}} \langle y, \widetilde{A}\overline{x}_N\rangle - \frac{1}{N}\sum_{k=1}^{N}\langle y_k, \widetilde{A}x_k\rangle\right\} + \left\{\frac{1}{N}\sum_{k=1}^{N}\langle y_k, \widetilde{A}x_k\rangle - \min_{x\in\Delta_1^{n}} \langle \overline{y}_N, \widetilde{A}x\rangle\right\} \\
&\leq \sqrt{\frac{2}{N}}\Big(\sqrt{\ln(2n)} + 2\sqrt{\ln(1/\delta)}\Big) + \sqrt{\frac{2}{N}}\Big(\sqrt{\ln n} + 2\sqrt{\ln(1/\delta)}\Big) \\
&= \sqrt{\frac{2}{N}}\Big(\sqrt{\ln(2n)} + \sqrt{\ln(n)} + 4\sqrt{\ln(1/\delta)}\Big),
\end{aligned}$$

where the last estimate is accurate owing to Proposition 2 with $M = 1$. That is, with probability at least $1 - \delta$, it is sufficient to have

$$\begin{aligned}
N &\geq \frac{2}{\varepsilon^2}\Big(\sqrt{\ln(2n)} + \sqrt{\ln(n)} + 4\sqrt{\ln(1/\delta)}\Big)^2, \\
N &\geq \frac{4}{\varepsilon^2}\Big(\ln(2n) + \ln(n) + 16\ln(1/\delta)\Big)
\end{aligned}$$

iterations of the `GK` algorithm in order to guarantee $\|A\overline{x}_N\|_\infty \leq \varepsilon$. Each update of x or y involves an update of no more than d probabilities in vectors p and π corresponding to non-zeros in the gradient. Algorithm 3 requires $O(\log n)$ time to update each. Therefore, the time-complexity of the algorithm is bounded from above as

$$O\left(n + \frac{d\ln n(\ln n + \ln(\sigma^{-1}))}{\varepsilon^2}\right).$$

■