Seminars 1-2
PageRank-ing

Andrey Tremba, Alexander Gasnikov

November 14, 17. 2016

Lecture Plan

1. Ranking-by-graph problem
2. Markov Chain process problem
3. Problem formulation
4. Methods
5. Special

Node Ranking

Consider directed graph $G = (V, U)$, where V - set of nodes (points), U - set of edges (links) (v_1, v_2), $v_1, v_2 \in V$.

For simplicity let's enumerate all nodes by numbers $1, \ldots, n$. For now suppose all edges are the same and non-weighted.

Problem: Given graph G assign weights p_i, $i = 1, \ldots, n$ for all nodes in a rational way, so the weights reflect importance of nodes.

Recurrent weight definition

Consider i-th node. Let its weight be defined by weights of nodes, pointing to the chosen one.

But if a node is pointing to few, say, 3 other nodes, its weight is "distributed" among these nodes.

For example suppose there are only vertices a, b, j with $1, 2, n_j$ outgoing arcs have links to i-th vertice.

Then

$$p_i = p_a + \frac{1}{2}p_b + \frac{1}{n_j}p_j.$$

In general we have correspondence

$$p_i = \sum_{(j,i)\in U} \frac{p_j}{n_j}, \quad i = 1, \ldots, n,$$

or, in matrix form

$$A_1 p = p.$$

where each j-th column contains n_j nonzero elements, all equal to $1/n_j$, if $n_j \neq 0$, and all zeros otherwise.

Empty rows and/or empty columns. The latter are more important.

The resulting matrix A is stochastic (or - column-stochastic) matrix: it has non-negative entries and sum of elements in each column is equal to 1.

$$Ap = p.$$

It is an eigenvalue* problem!

Markov Chains

Consider a graph (maybe the same) which is viewed as representation of a random switching process of a system with n states: 1st, 2nd, etc. At each discrete time instant its state randomly changes from current, say, i-th state to other states, following rule:

$$Prob(s_{t+1} = j | s_t = i) = p_{i,j}, \quad \sum_{j=1}^{n} p_{i,j} = 1.$$

If we assume initial probability distribution on states π^0: $\pi_i^0 = Prob(s_0 = i)$, $\sum_{k=i}^{n} \pi_i^0 = 1$, then we can predict probability distribution on next step:

$$\pi_j^{t+1} = \sum_{i=1}^{n} Prob(s_{t+1} = j | s_t = i) Prob(s_t = i) = \sum_{i=1}^{n} p_{i,j} \pi_i^t$$

When considering vector π as row vector, the iterations can be written as

$$\pi^{t+1} = \pi^t P \tag{1}$$

with (row) stochastic matrix P.

The transitions form stationary Markov chain (next state depends on the previous state only), which can be represented by a directed weighted graph.

A question of interest if there exists stationary distribution

$$\pi^* = \pi^* P,$$

and whether (1) converges.

Theoretically we should look at eigenvalues with unit absolute value and its multiplicities.

Robust PageRank

Modification of the ranking problem is *robust* PageRank, proposed by Anatoli Juditsky and Boris Polyak.

Key idea is to optimize function

$$\min_{x \in S_n} \|Ax - x\|_2 + \lambda \|x\| \qquad (2)$$

with small λ representing *matrix disturbance* magnitude.

Page Ranking problem

In any way, we got main problem

$$Ax = x, \; x \in \Re^n.$$

Let's consider few reformulations:

As smooth optimization on unit simplex

$$\min_{x \in S_n} \|Ax - x\|_2^2 \tag{3}$$

As non-smooth optimization on unit simplex

$$\min_{x \in S_n} \|Ax - x\|_2. \tag{4}$$

As non-smooth problem on unit simplex (ℓ_∞-norm)

$$\min_{x \in S_n} \|Ax - x\|_\infty, \tag{5}$$

As non-smooth problem on unit simplex (ℓ_1-norm)

$$\min_{x \in S_n} \|Ax - x\|_1, \tag{6}$$

As solution of linear equation system problem: suppose that $x_n \neq 0$, then solve

$$(A - I) \left[\begin{array}{c} \widehat{x} \\ 1 \end{array} \right] = 0, \tag{7}$$

then take as solution $x^* = \begin{bmatrix} \widehat{x} \\ 1 \end{bmatrix} / \| \begin{bmatrix} \widehat{x} \\ 1 \end{bmatrix} \|_1$.

As saddle-point problem (based on (5))

$$\min_{x \in S_n} \max_{y \in S_{2n}} y^T \begin{bmatrix} A - I \\ -A + I \end{bmatrix} x, \tag{8}$$

or (based on (4))

$$\min_{x \in S_n} \max_{y \in B_n} y^T (A - I) x, \tag{9}$$

As projection on linear space

$$\min_{\substack{Ax - x = 0 \\ (e, x) = 1}} \|x\|_2^2 \tag{10}$$

with vector $e = (1, 1, \ldots, 1)$.

As unconstrained minimization*

$$\min_x \left\| \begin{bmatrix} A - I \\ e^T \end{bmatrix} x - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\| \tag{11}$$

Methods

- Non-optimization
- (Sub)gradient Methods
- Fast Gradient Methods
- Mirror Descent Methods
- Frank-Wolfe Method
- Stochastic Mirror Descent

Non-optimization methods

a) Power Iterations

$$x^{k+1} = Ax^k$$

PageRank statement: $\widehat{A} = (1-\alpha)A + \alpha\frac{1}{n}ee^T$

$$\|\widehat{A}x^k - x^k\| \sim \alpha^k$$

Brin S., Page L. The anatomy of a large-scale hypertextual web search engine, Comput. Networks ISDN Syst. 30(1–7):107–117, 1998.

Nesterov Yu., Nemirovski A. Finding the stationary states of Markov chains by iterative methods, Applied Mathematics and Computation 255:58–65, 2015.

b) Averaging Power Iterations

$$\overline{x}^k = \sum_{i=0}^{k-1} x^i$$

$$\|A\overline{x}^k - \overline{x}^k\| \sim \frac{1}{k}$$

c) Regularized Power Iterations

$$x^{k+1} = (1 - \alpha_k)Ax^k + \alpha_k \frac{1}{n} ee^T,\ \alpha_k \to 0,$$

$$\|Ax^k - x^k\| \sim \frac{1}{k}$$

Polyak B.T. and Tremba A.A. Regularization-based solution of the PageRank problem for large matrices, Automation and Remote Control, 2012, vol. 73, issue 11, pp. 1877–1894.

Gradient Descent (with Projection)

Applicable to (3)

$$f(x) = \frac{1}{2}\|Ax - x\|_2^2, \; x \in S_n.$$

$$\nabla f(x) = (A - I)^T (A - I)x$$

Bound* on Lipschitz constant for the gradient is

$$L = \sigma_1((A - I)^T (A - I)) \leq 4$$

$$x^{k+1} = P_{S_n}(x^k - \frac{1}{4}\nabla f(x^k)).$$

$$f(x^k) \leq \frac{8R^2}{k}, \quad \|Ax^k - x^k\|_2 \leq \frac{4R}{\sqrt{k}}.$$

Prox-Function

cf. Lecture 3-4 slides. Norm $\|\cdot\|$ + set Q $\rightarrow$ strongly convex prox-function $d(x)$ $\rightarrow$ Bregman divergence

$$V(x, y) = d(x) - d(y) - (\nabla d(y), x - y)$$

"Mirror Projection"

$$\mathrm{Mirr}_y(v) = \arg\min_{x \in Q} (v, x - y) + V(x, y)$$

Essentially

$$\mathrm{Mirr}_y(v) = \arg\min_{x \in Q} (v - \nabla d(y), x) + d(x)$$

Prox-Function Samples

$Q = B_n = \{x : \|x\|_2 \leq 1\}$

$$d(x) = \frac{1}{2}\|x\|_2^2$$

Exercise 1

$$\mathrm{Mirr}_y(v) = ?$$

Prox-Function Samples (2)

$Q = S_n = \{x : x_i \geq 0, \sum_{i=1}^{n} x_i = 1\}$

$$d(x) = \log(n) + \sum_{i=1}^{n} x_i \ln x_i$$

Exercise 2

$$\mathrm{Mirr}_y(v) = ?$$

Fast Gradient Method (Similar Triangles Method)

Applicable to (3), cf. Lecture 3-4 slides. Choose $u^0 = x^0 = (\frac{1}{n}, \ldots, \frac{1}{n})$, $\alpha_0 = 1/L$, $A_k = \sum_{i=0}^k \alpha_i$, $\alpha_{k+1} = \frac{1+\sqrt{1+4L^2\alpha_k^2}}{2L}$,

$$
\begin{aligned}
y^{k+1} &= \tfrac{\alpha_{k+1}u^k + A_k x^k}{A_{k+1}}, \\
u^{k+1} &= \mathrm{Mirr}_{u^k}(\alpha_{k+1}\nabla f(y^{k+1})), \\
x^{k+1} &= \tfrac{\alpha_{k+1}u^{k+1} + A_k x^k}{A_{k+1}}.
\end{aligned}
$$

Convergence Speed of FGM/STM

$$f(x^k) \leq \frac{4LR^2}{(k+1)^2}$$

$$\|Ax^k - x^k\|_2 \leq \frac{4R}{k+1}$$

Almost the same speed as of Power-like methods with averaging, but with higher step computation cost.

Mirror Descent Method

Applicable to (3), cf. Lecture 3-4 slides, example 2.

While $\|\nabla f(x)\|_\infty \leq 4$, put $h = \sqrt{\frac{ln(n)}{4N}}$

$$x^{k+1} = \mathrm{Mirr}_{x^k}(h \nabla f(x^k)),$$

and collect

$$\overline{x}^k = \frac{1}{k+1} \sum_{i=0}^{k} x^i.$$

$$f(\overline{x}^N) \sim \frac{1}{\sqrt{N}}, \quad \|A\overline{x}^N - \overline{x}^N\|_2 \sim \frac{1}{N^{1/4}}.$$

Frank-Wolfe Method

Solve (3) by conditional gradient method:

$$y^k = \arg\min_{x \in S_n} (\nabla f(x^k), x) = e_{i_k},$$

where i_k-th axis vector $(0, \ldots, 0, 1, 0, \ldots, 0)$ has 1 at position $i_k = \arg\min_i \frac{\partial f(x^k)}{\partial x_i}$.

Step point resembles averaging:

$$x^{k+1} = \frac{k-1}{k+1} x^k + \frac{2}{k+1} y^k.$$

$$f(x^k) \le \frac{16}{k+1}, \quad \|Ax^k - x^k\|_2 \le \frac{4}{\sqrt{k+1}}$$

Parametrized Projection Step

$$\mathrm{Mirr}(\beta, v) = \arg\min_{x \in S_n} (v^T x + \beta d(x)) = \mathrm{Mirr}_0(v/\beta).$$

The mapping can be calculated in explicit form. Denote $z = \mathrm{Mirr}_x(\beta, v)$, then (check, also cf. Lecture 3-4 slides and Exercise 2)

$$z_i = \frac{e^{-v_i/\beta}}{\sum_{j=1}^{n} e^{-v_j/\beta}}.$$

Stochastic Mirror Descent

Put $\beta_k = \beta_0\sqrt{k+1}$, $\beta_0 = 2/\sqrt{ln(n)}$, $x^0 = v_0 = (0,0,\ldots,0)^T \in \mathbb{R}^n, x_0 = (1/n, 1/n, \ldots, 1/n)$

$$\begin{aligned} v^{k+1} &= v^k + \xi^k, \\ x^{k+1} &= \mathrm{Mirr}(\beta_k, v^{k+1}), \\ \overline{x}^{k+1} &= \overline{x}^k - \tfrac{1}{k+1}(\overline{x}^k - x^{k+1}) \end{aligned}$$

Where $\mathbb{E}\xi^k \in \partial f(x^k)$. If $\mathbb{E}\|\xi^k(x)\|_\infty^2 \leq L^2$, then

$$\mathbb{E}f(\overline{x}^k) \leq 2L\frac{\sqrt{k+1}}{k} \sim \frac{1}{\sqrt{k}}$$

Adaptive Parameter Choice

$$
\begin{aligned}
v^{k+1} &= v^k + \xi^k, \\
\beta_{k+1} &= \left(\beta_k^2 + \tfrac{\|\xi^k\|_*^2}{\ln n}\right)^{1/2}, \\
x^{k+1} &= \mathrm{Mirr}(\beta_k, v^{k+1}), \\
\overline{x}^{k+1} &= \overline{x}^k - \tfrac{1}{k+1}(\overline{x}^k - x^{k+1})
\end{aligned}
$$

$$
\mathbb{E} f(\overline{x}^k) \leq \frac{L}{k} O(\mathbb{E}\beta_k)
$$

Gradient Calculation Cost

For smooth function (3):

$$\nabla f(x) = (A - I)^T (A - I)x,$$

Two matrix-vector multiplicatons $\Theta(2n^2)$ operations for dense matrices and $\Theta(2sn)$ for s-sparse matrices.

For non-convex problem (4) the same order.

$$\nabla f(x) = \frac{(A - I)^T (A - I)x}{\|Ax - x\|_2}$$

Randomization

The idea is to introduce randomness in deterministic problem. We need vectors $\xi(x)$ with:

- mean value being (sub)gradient of target function
$$\mathbb{E}\xi(x) \in \partial f(x),$$
- uniformly bounded
$$\mathbb{E}\|\xi(x)\|_* \leq M,$$
- or with bounded second moment
$$\mathbb{E}\|\xi(x)\|_*^2 \leq M^2.$$

Randomizing Matrix-vector Multiplication

Consider a) column-stochastic matrix A,
a) row-stochastic matrix A,
and gradient being

$$\nabla f(x) = Ax.$$

Introduce random index

$$\eta : Prob(\eta = j | x) = x_j,$$

Then take η-th (random) column $A^{(\eta)}$ of matrix A

$$\xi = A^{(\eta)}, \;\; \mathbb{E}(\xi | x) = Ax, \;\; \|\xi\|_\infty \leq 1.$$

$$\nabla f(x) = A^T A x - A^T x - A x + x.$$

Introduce second index χ

$$P(\chi = i | x, \eta) = a_{i,\eta}$$

and take

$$\xi = A^T_{(\chi)} - A^T_{(\eta)} - A^{(\eta)} + x.$$

where $A_{(i)}$ denotes i-th row of matrix A.

Exercise 3

Prove

$$\mathbb{E}(\xi | x) = \nabla f(x)$$

For saddle-point problem (9)

$$\begin{aligned} \partial_x f(x,y) &= (A-I)^T y, \quad y \in B_n \\ \partial_y f(x,y) &= (A-I)x, \quad x \in S_n \end{aligned}$$

Exercise 4

Propose a randomization for $\partial_x f(x,y)$ and estimate constant M.

Randomizing for Robust PageRank

$$\|Ax - x\|_2 + \lambda \|x\|_2 \to \min$$

Exercise 5

- propose reasonable randomization,
- estimate constant M

Hint: saddle-point representation worth trying.