Convex Optimization for Data Science.

Big Data applications

Gasnikov Alexander (MIPT, IITP RAS)

gasnikov@yandex.ru

Двенадцатая Международная Школа-семинар «Проблемы оптимизации сложных систем»

Новосибирск, Академгородок, ИВМиМГ СО РАН

12-16 декабря 2016 г.

Main books:

Nemirovski A. Lectures on modern convex optimization analysis, algorithms, and engineering applications. – Philadelphia: SIAM, 2013.

Nesterov Yu., Shpirko S. Primal-dual subgradient method for huge-scale linear conic problem // SIAM Journal on Optimization. – 2014. – V. 24. –no. 3. – P. 1444–1457.

Bubeck S. Convex optimization: algorithms and complexity // In Foundations and Trends in Machine Learning. – 2015. – V. 8. – no. 3-4. – P. 231–357.

Blum A., Hopcroft J., Kannan R. Foundations of Data Science. Draft, June 2016.
http://www.cs.cornell.edu/jeh/book2016June9.pdf

Gasnikov A. Searching equilibriums in large transport networks. Doctoral Thesis. MIPT, 2016.
https://arxiv.org/ftp/arxiv/papers/1607/1607.03142.pdf

http://www.maths.ed.ac.uk/~prichtar/

Structure of Lecture

- Google problem (Page Rank)
- Inverse problems: traffic demand matrix estimation from link loads
- Empirical Risk Minimization (ERM)
- Maximum Likelihood Estimation (MLE)
- Bayesian inference
- L1-optimization (sparse solution)
- Typical Data Science problem formulation (as optimization problem)
- Dual problem

Google problem (Page Rank)

Let there are $N \gg 1$ users that independently walk at random on the web-graph (n vertexes). Assume that transitional probability matrix of random walks P is irreducible. Let's denote $n_i\left(t\right)$ – the number of users at the i-th web-page at the moment of time t. Using Gordon–Newell's theorem one can obtain $\exists!\, p \in S_n\left(1\right):\ p^T = p^T P$ (p – Page Rank) and

$$\lim_{t\to\infty} P\left(\vec{n}\left(t\right)=\vec{n}\right)=\frac{N!}{n_1!\cdot\ldots\cdot n_n!}\, p_1^{n_1}\cdot\ldots\cdot p_n^{n_n}.$$

Hence, using Hoeffding's inequality in a Hilbert space one can obtain

$$\lim_{t\to\infty} P\left(\left\|\frac{n\left(t\right)}{N}-p\right\|_2 \geq \frac{2\sqrt{2}+4\sqrt{\ln\left(\sigma^{-1}\right)}}{\sqrt{N}}\right)\leq \sigma.$$

How to find Page Rank via Convex Optimization?

According to Frobenius' theory for nonnegative matrix we have the following equivalent optimization's type reformulations of Google problem:

$$\frac{1}{2}\|Ax\|_2^2 \to \min_{x\in S_n(1)}; \text{ (smooth representation)}$$

$$\|Ax\|_\infty \to \min_{x\in S_n(1)}; \text{ (not smooth representation)}$$

$$\min_{x\in S_n(1)} \max_{\omega\in S_{2n}(1)} \left\langle \omega, \tilde{A}x \right\rangle; \text{ (saddle point representation)}$$

$$\frac{1}{2}\|x\|_2^2 \to \min_{\bar{A}x=b}, \text{ (required dual representation)}$$

where $A = P^T - I_n$, $\tilde{A} = J^T A$, $J = [I_n; -I_n]$, $\bar{A} = \left[P - I_n; \vec{1}\right]^T$, $b = (0,\dots,0,1)^T$.

Inverse problems: traffic demand matrix estimation from link loads

In the problem of traffic demand matrix estimation the goal is to recover traffic demand matrix represented as a vector $x \geq 0$ from known route matrix A (the element $A_{i,j}$ is equal 1 iff the demand with number j goes through link with number i and equals 0 otherwise) and link loads b (amount of traffic which goes through every link). This leads to the problem of finding the solution of linear system $Ax = b$. Also we assume that we have some $x_g \geq 0$ which reflects our prior assumption about x. Thus we consider x to be a projection of x_g on a simplex-type set $\left\{x \geq 0 : \quad Ax = b\right\}$

$$\min_{\substack{Ax=b \\ x\geq 0}} \left\{ g(x) := \left\| x - x_g \right\|_2^2 \right\} = \min_{\substack{\left\| Ax-b \right\|_2^2 \leq 0 \\ x \geq 0}} g(x).$$

Slater’s relaxation of this problem leads to the problem (denote x_* the solution of this problem)

$$\left\|x-x_g\right\|_2^2 \to \min_{\substack{\left\|Ax-b\right\|_2^2\le\varepsilon^2 \\ x\ge 0}} .$$

This problem can be reduced to the problem (unfortunately without explicit dependence $\bar{\lambda}\left(\varepsilon\right)$)

$$\tilde{f}\left(x\right)=\left\|x-x_g\right\|_2^2+\bar{\lambda}\left\|Ax-b\right\|_2^2\to\min_{x\ge 0},$$

where $\bar{\lambda}$ – dual multiplier to the convex inequality $\left\|Ax-b\right\|_2^2\le\varepsilon^2$.

$$\tilde{f}\left(x\right)=\left\|x-x_g\right\|_2^2+\bar{\lambda}\left\|Ax-b\right\|_2^2\to\min_{x\geq 0}$$

One might expect that $\bar{\lambda}\gg\left\|x_*-x_g\right\|_2^2\big/\varepsilon^2$, but in reality $\bar{\lambda}$ can be chosen much smaller ($\bar{\lambda}\sim\varepsilon^{-1}-\varepsilon^{-2}$) if we restrict ourselves only by approximate solution. Let's reformulate the problem

$$f\left(x\right)=\left\|Ax-b\right\|_2^2+\lambda\left\|x-x_g\right\|_2^2\to\min_{x\geq 0},$$

where $\lambda=\bar{\lambda}^{-1}$. But sometimes it is worth to consider more general cases:

$$f\left(x\right)=\frac{1}{2}\left\|Ax-b\right\|_2^2+g\left(x\right)\to\min_{x\in Q}.$$

Hastie T., *Tibshirani R.*, *Friedman R*. The Elements of statistical learning: Data mining, Inference and Prediction. Springer, 2009.

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 + g(x) \to \min_{x \in Q}$$

Possible variants for choosing $g(x)$ are:

1. (Ridge Regression / Tomogravity model)

$$g(x) = \lambda \left\| x - x^g \right\|_2^2,\ Q = \mathbb{R}_+^n;$$

2. (Mimimal mutual information model)

$$g(x) = \lambda \sum_{k=1}^{n} x_k \ln\left(x_k / x_k^g \right),\ x, x^g \in Q = S_n(R) = \left\{ x \geq 0 : \sum_{k=1}^{n} x_k = R \right\}.$$

3. (LASSO)

$$g(x) = \lambda \|x\|_1,\ Q = \mathbb{R}_+^n;$$

Empirical Risk Minimization (ERM)

Suppose we have observation $\left\{x_i, y_i\right\}_{i=1}^{n}$ and we have some loss function $l\left(\hat{f}, X, Y\right)$. For example,

$$l\left(\hat{f}, X, Y\right) = I\left\{\hat{f}\left(X\right) \neq Y\right\} \text{ – binary classification;}$$

$$l\left(\hat{f}, X, Y\right) = \left(\hat{f}\left(X\right) - Y\right)^2 \text{ – regression;}$$

$$l\left(\hat{f}, X, Y\right) = \max\left\{0, 1 - Y\hat{f}\left(X\right)\right\} \text{ – hinge loss.}$$

Let's introduce V – VC-dimension of class F,

$$L\left(\hat{f}\right) = E_{X,Y}\left[l\left(\hat{f}_{\left\{x_i, y_i\right\}_{i=1}^{n}}, X, Y\right) \middle| \left\{x_i, y_i\right\}_{i=1}^{n}\right], \text{ for } \hat{f} = \hat{f}_{\left\{x_i, y_i\right\}_{i=1}^{n}} \in F,$$

$$\hat{f}_{ERM} = \arg\min_{f \in F} \sum_{i=1}^{n} l\left(f, x_i, y_i \right),\; L\left(f_* \right) = \inf_{f \in F} L\left(f \right).$$

Then (Vapnik–Chervonenkis, Zauer, Hausler for binary classification)

$$P\left(L\left(\hat{f}_{ERM} \right) - L\left(f_* \right) \le C\sqrt{\frac{V}{n}} + \sqrt{\frac{2\ln\left(\sigma^{-1}\right)}{n}} \right) \ge 1 - \sigma,$$

where C is universal constant.

Now Statistical Learning Theory (SLT) is a big branch of research where ERM approach (and its penalized versions) is the main tools.

Shalev-Shwartz S., Ben-David S. Understanding Machine Learning: From theory to algorithms. Cambridge University Press, 2014.

Sridharan K. Learning from an optimization viewpoint. PhD Thesis, 2011.

Maximum Likelihood Estimation (Fisher, Le Kam, Spokoiny)

Let $\mathrm{x}_k, k = 1,...,n$ – i.i.d. with density function $p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right)$ (supp. doesn't depend on θ), depends on unknown vector of parameters θ. Then for all statistics $\tilde{\theta}\left(\mathrm{x}\right)$ (with $E_{\mathrm{x}}\left[\tilde{\theta}\left(\mathrm{x}\right)^2\right] < \infty$):

$$E_{\mathrm{x}}\left[\left(\tilde{\theta}\left(\mathrm{x}\right)-\theta\right)\left(\tilde{\theta}\left(\mathrm{x}\right)-\theta\right)^T\right] \succ \left[I_{p,n}\right]^{-1}, \text{ (Rao–Cramer inequality)}$$

$$I_{p,n} \overset{def}{=} E_{\mathrm{x}}\left[\nabla_\theta \ln p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right)\left(\nabla_\theta \ln p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right)\right)^T\right] = nI_{p,1},$$

$$\tilde{\theta}_{MLE}\left(\mathrm{x}\right) = \arg\max_\theta p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right) = \arg\max_\theta \ln p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right) =$$

$$= \arg\max_\theta \ln \prod_{i=1}^n p_{\mathrm{x}_i}\left(\mathrm{x}_i\middle|\theta\right) = \arg\max_\theta \sum_{i=1}^n \ln p_{\mathrm{x}_i}\left(\mathrm{x}_i\middle|\theta\right),$$

$$\theta_* = \arg\max_\theta \sum_{i=1}^n E_{\mathrm{x}_i}\left[\ln p_{\mathrm{x}_i}\left(\mathrm{x}_i \middle| \theta\right)\right] = \arg\max_\theta \sum_{i=1}^n \int p_{\mathrm{x}_i}\left(x_i \middle| \theta_*\right) \ln p_{\mathrm{x}_i}\left(x_i \middle| \theta\right) dx_i .$$

Le Kam theory (Fisher's theorem): *When* $n \to \infty$ *then* $\tilde{\theta}_{MLE}\left(\mathrm{x}\right)$ *is asymptotically normal and optimal in sense of Rao–Cramer inequality ("=").*

Recently V. Spokoiny've proposed non asymptotic variant of this theory. In particular his theory allows to answer for the question: how fast could $m \to \infty$ ($m = \dim\theta$) with $n \to \infty$ for asymptotic optimality of $\tilde{\theta}_B\left(\mathrm{x}\right)$. He also considered closely connected result – Wilks' phenomenon.

Example (Least Squares). $y_i = kx_i + b + \varepsilon_i \;\; \varepsilon_i \in N\left(0, \sigma^2\right)$, $\theta = \left(k, b\right)^T$,

$$\mathrm{x} = A\theta + \varepsilon,\; \mathrm{x} = \left\{y_i\right\}_{i=1}^n,\; A = \begin{pmatrix} x_1 & \dots & x_n \\ 1 & \dots & 1 \end{pmatrix}^T,\; \tilde{\theta}_{MLE}\left(\mathrm{x}\right) = \arg\min_\theta \left\| A\theta - \mathrm{x} \right\|_2^2 .$$

Van Trees inequality (generalization of Rao–Cramer inequality)

Let $\mathrm{x}_k, k = 1,\ldots,n$ – i.i.d. with density function $p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right)$ (supp. doesn't depend on θ), depends on unknown vector of parameters θ with prior distribution $\pi\left(\theta\right)$. Then for all statistics $\tilde{\theta}\left(\mathrm{x}\right)$ (with $E_{\mathrm{x}}\left[\tilde{\theta}\left(\mathrm{x}\right)^2\right] < \infty$):

$$E_{\mathrm{x},\theta}\left[\left(\tilde{\theta}\left(\mathrm{x}\right)-\theta\right)\left(\tilde{\theta}\left(\mathrm{x}\right)-\theta\right)^T\right] \succ \left[I_{p,n} + I_{\pi}\right]^{-1}, \qquad (*)$$

$$I_{p,n} \overset{def}{=} E_{\mathrm{x},\theta}\left[\nabla_\theta \ln p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right)\left(\nabla_\theta \ln p_{\mathrm{x}}\left(\mathrm{x}\middle|\theta\right)\right)^T\right] = nI_{p,1},$$

$$I_{\pi} \overset{def}{=} E_{\theta}\left[\nabla_\theta \ln \pi\left(\theta\right)\left(\nabla_\theta \ln \pi\left(\theta\right)\right)^T\right].$$

Bayesian inference

Bayesian estimation:

$$\tilde{\theta}_B(\mathrm{x}) = \arg\min_{\breve{\theta}} \int I\left(\breve{\theta},\theta\right) p_{\mathrm{x}}\left(\mathrm{x}|\theta\right)\pi(\theta)\,d\theta,$$

$$I\left(\breve{\theta},\theta\right) = \frac{1}{2}\left\|\breve{\theta}-\theta\right\|_2^2.$$

Le Kam theory: *When* $n \to \infty$ *then* $\tilde{\theta}_B(\mathrm{x})$ *is asymptotically normal and optimal in sense of (*) ("=").*

Recently V. Spokoiny've proposed non asymptotic variant of this theory. In particular his theory allows to answer for the question: how fast could $m \to \infty$ ($m = \dim\theta$) with $n \to \infty$ for asymptotic optimality of $\tilde{\theta}_B(\mathrm{x})$.

Van Trees inequality $\to$ Rao–Cramer inequality and $\tilde{\theta}_B(\mathrm{x}) \to \tilde{\theta}_{MLE}(\mathrm{x})$ when $\pi(\theta) \in N(0, \sigma^2 I)$ with $\sigma \to \infty$.

Berstein–von Mises theorem say that $\tilde{\theta}_B(\mathrm{x})$ is $n^{-1/2}$-normaly concentrated around $\tilde{\theta}_{MLE}(\mathrm{x})$ when $n \to \infty$. Recently V. Spokoiny've proposed non asymptotic variant of this theorem.

Example. Assume that

$$\mathrm{x} = A\theta + \xi,\ \xi \in N(0, \sigma^2 I), \text{ prior on } \theta \in N(\theta_g, \tilde{\sigma}^2 I)$$

Then (compare to the traffic demand matrix estimation problem)

$$\tilde{\theta}_B(\mathrm{x}) = \arg\min_{\theta} \left\{ \left\| A\theta - \mathrm{x} \right\|_2^2 + \lambda \left\| \theta - \theta_g \right\|_2^2 \right\},\ \lambda = \sigma^2 / \tilde{\sigma}^2 .$$

Compressed Sensing and L1-optimization (Donoho, Candes, Tao)

There are many areas where linear systems arise in which a sparse solution is unique. One is in plant breading. Consider a breeder who has a number of apple trees and for each tree observes the strength of some desirable feature. He wishes to determine which genes are responsible for the feature so he can cross bread to obtain a tree that better expresses the desirable feature. This gives rise to a set of equations $Ax = b$ where each row of the matrix A corresponds to a tree and each column to a position on the genome. The vector b corresponds to the strength of the desired feature in each tree. The solution x tells us the position on the genome corresponding to the genes that account for the feature. So one can hope that NP-hard problem $\|x\|_0 \to \min\limits_{Ax=b}$ can be replaced by convex problem $\|x\|_1 \to \min\limits_{Ax=b}$. Due to Lagrange multipliers principle we can relax this problem as

$$\frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_1 \to \min_x .$$

What are the sufficient conditions for: $\|x\|_0 \to \min\limits_{Ax=b} \Leftrightarrow \|x\|_1 \to \min\limits_{Ax=b}$?

Restricted Isometry Property (RIP)

$$(1-\delta_s)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\delta_s)\|x\|_2^2 \text{ \textit{for any} } s\text{-\textit{sparse} } x.$$

Sufficient condition. Suppose that x_0 (solution of $\|x\|_0 \to \min\limits_{Ax=b}$) has at most s nonzero coordinates, matrix A satisfy RIP with $\delta_s \le \left(5\sqrt{s}\right)^{-1}$, then x_0 is the unique solution of the convex optimization problem $\|x\|_1 \to \min\limits_{Ax=b}$.

Example RIP matrix: for all $x \in \mathbb{R}^n \to P\left((1-\varepsilon)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\varepsilon)\|x\|_2^2\right) \ge 1-2\exp\left(-\varepsilon^2 n/6\right)$, where i.i.d. $A_{ij} \in N\left(0,n^{-1}\right)$ ($0<\varepsilon<1$). If A is $(\varepsilon,2s)$-RIP and $\|\tilde{x}\|_0 \le s$ satisfy $Ax=b$ then $\tilde{x}=x_0$.

Examples of Data Science problems

Typically Data Science problems lead to the optimization problems:

$$\sum_{k=1}^{m} f_k\left(A_k^T x\right) + g\left(x\right) \to \min_{x \in Q}.$$

At least one of the dimensions is huge m (sample size), n (parameters).

Ridge Regression

$f_k\left(y_k\right) = C \cdot \left(y_k - b_k\right)^2,\ g\left(x\right) = \frac{1}{2}\|x\|_2^2$. (smooth, strictly convex)

Support Vector Machine (SVM has Bayesian nature, V.V. Mottl’)

$f_k\left(y_k\right) = C \max\left\{0, 1 - b_k y_k\right\},\ g\left(x\right) = \frac{1}{2}\|x\|_2^2$. (non smooth, strictly convex)

Dual problem (convex case)

Sometimes it is proper to solve dual problem instead of primal one:

$$\sum_{k=1}^{m} f_k\left(\langle A_k, x\rangle\right) + g\left(x\right) \to \min_{x\in Q},$$

$$\min_{x\in Q}\left\{\sum_{k=1}^{m} f_k\left(A_k^T x\right) + g\left(x\right)\right\} = \min_{\substack{x\in Q\\ z=Ax}}\left\{\sum_{k=1}^{m} f_k\left(z_k\right) + g\left(x\right)\right\} =$$

$$= \min_{\substack{x\in Q\\ z=Ax, z'}} \max_{y}\left\{\langle z - z', y\rangle + \sum_{k=1}^{m} f_k\left(z'_k\right) + g\left(x\right)\right\} =$$

$$= \max_{y\in\mathbb{R}^m}\left\{-\max_{\substack{x\in Q\\ z=Ax}}\left\{\langle -z, y\rangle - g\left(x\right)\right\} - \max_{z'}\left\{\langle z', y\rangle - \sum_{k=1}^{m} f_k\left(z'_k\right)\right\}\right\} =$$

$$= \max_{y\in\mathbb{R}^m}\left\{-\max_{x\in Q}\left(\left\langle -A^T y, x\right\rangle - g(x)\right) - \sum_{k=1}^{m}\max_{z'_k}\left(z'_k y_k - f_k(z'_k)\right)\right\} =$$

$$= \max_{y\in\mathbb{R}^m}\left\{-g^*\left(-A^T y\right) - \sum_{k=1}^{m} f_k^*(y_k)\right\} = -\min_{y\in\mathbb{R}^m}\left\{g^*\left(-A^T y\right) + \sum_{k=1}^{m} f_k^*(y_k)\right\}.$$

1) $\frac{L}{2}\|Ax-b\|_2^2 + \frac{\mu}{2}\|x-x_g\|_2^2 \to \min\limits_{x\in\mathbb{R}^n}$, 2) $\frac{L}{2}\|Ax-b\|_2^2 + \mu\sum\limits_{k=1}^{n} x_k \ln x_k \to \min\limits_{x\in S_n(1)}$.

1) $\frac{1}{2\mu}\left(\left\|x_g - A^T y\right\|_2^2 - \left\|x_g\right\|_2^2\right) + \frac{1}{2L}\left(\|y+b\|_2^2 - \|b\|_2^2\right) \to \min\limits_{y\in\mathbb{R}^m}$, (dual for 1))

2) $\frac{1}{\mu}\ln\left(\sum\limits_{i=1}^{n}\exp\left(\frac{\left[-A^T y\right]_i}{\mu}\right)\right) + \frac{1}{2L}\left(\|y+b\|_2^2 - \|b\|_2^2\right) \to \min\limits_{y\in\mathbb{R}^m}$. (dual for 2))

Good and Bad News

Good news: In convex case even with huge *m* and *n* these type of the problems

$$\sum_{k=1}^{m} f_k\left(\langle A_k, x\rangle\right) + g(x) \to \min_{x\in Q}$$

are fast solvable numerically (often by accelerated primal or dual coordinate descent methods).

Bad news: Real Data Science problems often lead to non convex optimization problems. Typical example is probabilistic topic modeling (see K.V. Vorontsov). With MLE-approach one can obtain only non convex problem

$$f\left(\left\{\varphi_{\cdot t}\right\}_{t=1}^{|T|}, \left\{\theta_{\cdot d}\right\}_{d=1}^{|D|}\right) = -\sum_{d\in D}\sum_{w\in W} n_{wd} \ln\left(\sum_{t\in T}\varphi_{wt}\theta_{td}\right) \to \min_{\left\{\varphi_{\cdot t}\in S_{|W|}(1)\right\};\left\{\theta_{\cdot d}\in S_{|T|}(1)\right\}}.$$

Спасибо за внимание.