Stochastic Reformulations of Linear Systems: Algorithms and Convergence Theory

Peter Richtárik

Modern Convex Optimization and Applications
Workshop in honour of Arkadi Nemirovski's 70th birthday
Fields Institute, Toronto
July 4-7, 2017

Arkadi Nemirovski

Follow

Professor and Jh. Hunter Academic Chair, School of Industrial and Systems Engineering, Georgia

optimization, operations research, convex optimization, nonparametric statistics

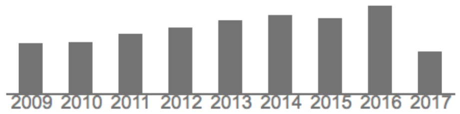
Verified email at isye.gatech.edu - Homepage

Title 1–20	Cited by	Year
LMI Control Toolbox for use with Matlab. 1995. The MATH Works P Gahinet, A Nemirovski, AJ Laub, M Chilali Inc., Natick, MA	5763 *	
Interior-point polynomial algorithms in convex programming Y Nesterov, A Nemirovskii Society for industrial and applied mathematics	5125	1994
Lectures on modern convex optimization analysis, algorithms and engineering applications (MPS/SIAM series on optimization, 2) A BEN-TAL, A NEMIROVSKI	2198 *	2001
Robust convex optimization A Ben-Tal, A Nemirovski Mathematics of operations research 23 (4), 769-805	1998	1998
Robust optimization A Ben-Tal, L El Ghaoui, A Nemirovski Princeton University Press	1916	2009
Robust solutions of uncertain linear programs A Ben-Tal, A Nemirovski Operations research letters 25 (1), 1-13	1595	1999
Problem Complexity and Method Efficiency in Optimization AS Nemirovsky, DB Yudin J. Wiley @ Sons, New York	1365 *	1983
Robust solutions of linear programming problems contaminated with uncertain data A Ben-Tal, A Nemirovski Mathematical programming 88 (3), 411-424	1298	2000
Robust optimization–methodology and applications A Ben-Tal, A Nemirovski Mathematical programming 92 (3), 453-480	999	2002
Adjustable robust solutions of uncertain linear programs A Ben-Tal, A Goryashko, E Guslitzer, A Nemirovski Mathematical Programming 99 (2), 351-376	830	2004
Stochastic approximation approach to stochastic programming A Nemirovski, A Juditsky, G Lan, A Shapiro SIAM Journal on Optimization 19 (4), 1574-1609	800	2009

Google Scholar

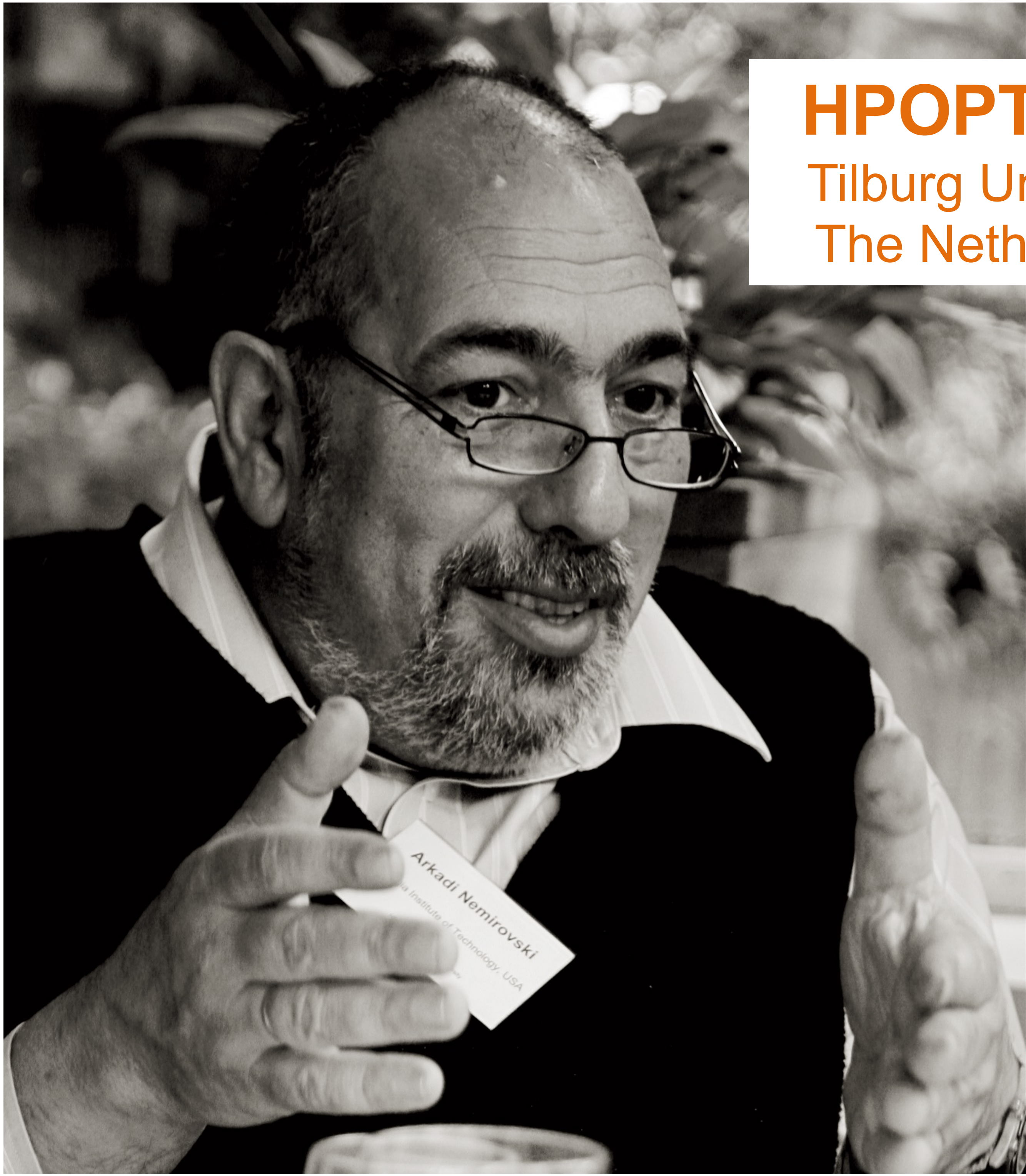
Get my own profile

Citation indices	All	Since 2012
Citations	35475	16344
h-index	63	44
i10-index	158	104



HPOPT 2008

Tilburg University
The Netherlands



A System of Linear Equations

m equations with n unknowns

$$A \in \mathbb{R}^{m\times n}, \quad x \in \mathbb{R}^{n}, \quad b \in \mathbb{R}^{m}$$

$$m\left\{\overbrace{A}^{n}\right. x = b$$

Assumption: The system is consistent (i.e., a solution exists)

Part I
Stochastic Reformulations

P.R. and Martin Takáč
Stochastic Reformulations of Linear Systems: Algorithms and Convergence Theory
arXiv:1706.01108, 2017

Stochastic Reformulations of Linear Systems

$n \times n$ pos def

distribution over $m \times q$ matrices

$$B, \mathcal{D}$$

$$Ax = b$$

1. Stochastic Optimization
2. Stochastic Linear System
3. Stochastic Fixed Point
4. Probabilistic Intersection

Example: $B =$ identity
$\mathcal{D} =$ uniform over $e_1, \ldots, e_m$ (unit basis vectors in $\mathbb{R}^m$)

Theorem

a) These 4 problems have the same solution sets

b) Necessary & sufficient conditions for the solution set to be equal to $\{x \;:\; Ax = b\}$

Reformulation 1: Stochastic Optimization

$$\text{Minimize } f(x) \overset{\text{def}}{=} \mathbf{E}_{S\sim\mathcal{D}}[f_S(x)]$$

$$f_S(x) = \frac{1}{2}\|x - \Pi^B_{\mathcal{L}_S}(x)\|^2_B = \frac{1}{2}(Ax-b)^\top H(Ax-b)$$

$$\mathcal{L}_S = \{x \ : \ S^\top Ax = S^\top b\}$$

$$H = S(S^\top AB^{-1}A^\top S)^\dagger S^\top$$

Reformulation 2: Stochastic Linear System

Instead of $Ax = b$ we solve the preconditioned system:

$$H = S(S^\top A B^{-1} A^\top S)^\dagger S^\top$$

$$\text{Solve } \underbrace{B^{-1}A^\top \mathbf{E}_{S\sim\mathcal{D}}[H]}_{\text{preconditioner}} A x = B^{-1}A^\top \mathbf{E}_{S\sim\mathcal{D}}[H] b$$

Instead of $B^{-1}A^\top \mathbf{E}[H] A$ we have access to $B^{-1}A^\top H A$

Unbiased estimate of the preconditioner

Reformulation 3: Stochastic Fixed Point Problem

$$\text{Solve } x = \mathbf{E}_{S\sim\mathcal{D}}\left[\Pi^B_{\mathcal{L}_S}(x)\right]$$

Projection in B-norm onto $\mathcal{L}_S = \{x \,:\, S^\top A x = S^\top b\}$

Reformulation 4: Probabilistic Intersection Problem

$$\text{Find } x \in \mathbb{R}^n \text{ such that } \mathbf{P}(x \in \mathcal{L}_S) = 1$$

$$\mathcal{L}_S = \{x \; : \; S^\top A x = S^\top b\}$$

Sketched system

S discrete $\Rightarrow$ $\{x \; : \; \mathbf{P}(x \in \mathcal{L}_S) = 1\} = \bigcap_S \mathcal{L}_S$

Part II
Randomized Algorithms

Viewpoint 1:
Stochastic Optimization

Stochastic Gradient Descent

constant stepsize

$S \sim \mathcal{D}$

$$x^{t+1} = x^t - \omega \nabla f_S(x^t)$$

"stochastic gradient"

A key method in machine learning

Stochastic "Newton" Descent

Constant stepsize

"stochastic gradient"

$$x^{t+1} = x^t - \omega(\nabla^2 f_S)^{\dagger_B} \nabla f_S(x^t)$$

B- pseudoinverse of the "stochastic Hessian"

$S \sim \mathcal{D}$

Stochastic Proximal Point Method

$$S \sim \mathcal{D}$$

$$x^{t+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f_S(x) + \frac{\omega - 1}{2\omega} \|x - x^t\|_B^2 \right\}$$

Viewpoint 3: Stochastic Fixed Point Method

Stochastic Fixed Point Method

Stochastic fixed point mapping

$$x^{t+1} = \omega \Pi^{B}_{\mathcal{L}_S}(x^t) + (1-\omega)x^t$$

Relaxation parameter

$S \sim \mathcal{D}$

Part III
Complexity

Basic Method

Basic Method: Complexity

$$\mathbf{E}[U^\top B^{1/2}(x^t - x^*)] = (I - \omega\Lambda)^t U^\top B^{1/2}(x^0 - x^*)$$

stepsize / relaxation parameter

$$W = B^{-1/2}A^\top \mathbf{E}_{S\sim\mathcal{D}}[H]AB^{-1/2} = U\Lambda U^\top$$

$$H = S(S^\top AB^{-1}A^\top S)^\dagger S^\top$$

Basic Method: Complexity

Convergence of Expected Iterates

$$t \geq \frac{1}{\lambda_{\min}^{+}} \log\left(\frac{1}{\epsilon}\right) \quad \xrightarrow{\omega = 1} \quad \|\mathbf{E}[x^t - x^*]\|_B^2 \leq \epsilon$$

$$t \geq \frac{\lambda_{\max}}{\lambda_{\min}^{+}} \log\left(\frac{1}{\epsilon}\right) \quad \xrightarrow{\omega = 1/\lambda_{\max}} \quad \|\mathbf{E}[x^t - x^*]\|_B^2 \leq \epsilon$$

L2 Convergence

$$t \geq \frac{1}{\lambda_{\min}^{+}} \log\left(\frac{1}{\epsilon}\right) \quad \xrightarrow{\omega = 1} \quad \mathbf{E}\left[\|x^t - x^*\|_B^2\right] \leq \epsilon$$

Parallel Method

Parallel Method

"Run 1 step of the basic method from x^t
several times independently,
and average the results."

i.i.d.

$$x^{t+1} = \frac{1}{\tau}\sum_{i=1}^{\tau} \underbrace{\phi_\omega(x^t, S_i^t)}_{\text{One step of the basic method from } x^t}$$

Parallel Method: Complexity

L2 Convergence

$$\tau = 1 \qquad\qquad\qquad \tau = +\infty$$

$$t \geq \frac{1}{\lambda_{\min}^{+}} \log\left(\frac{1}{\epsilon}\right) \quad \text{or} \quad t \geq \frac{\lambda_{\max}}{\lambda_{\min}^{+}} \log\left(\frac{1}{\epsilon}\right)$$

$$\Downarrow$$

$$\mathbf{E}\left[\|x^t - x^*\|_B^2\right] \leq \epsilon$$

Accelerated Method

Accelerated Method

Acceleration parameter
(between 1 and 2)

$S^t, S^{t-1} \sim \mathcal{D}$ (independent)

$$x^{t+1} = \gamma \underbrace{\phi_\omega(x^t, S^t)}_{\text{One step of the basic method from } x^t} + (1-\gamma)\underbrace{\phi_\omega(x^{t-1}, S^{t-1})}_{\text{One step of the basic method from } x^{t-1}}$$

Accelerated Method: Complexity

Convergence of Iterates

$$t \geq \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}^{+}}} \log\left(\frac{1}{\epsilon}\right) \quad \Rightarrow \quad \|\mathbf{E}[x^t - x^*]\|_B^2 \leq \epsilon$$

Basic Method depends on $\frac{\lambda_{\max}}{\lambda_{\min}^{+}}$!

Detailed Complexity Results

Alg.	ω	τ	γ	Quantity	Rate	Complexity	Theorem
1	1	-	-	$\\|\mathrm{E}[x_k - x_*]\\|_{\mathbf{B}}^2$	$(1-\lambda_{\min}^+)^{2k}$	$1/\lambda_{\min}^+$	4.3, 4.4, 4.6
1	$1/\lambda_{\max}$	-	-	$\\|\mathrm{E}[x_k - x_*]\\|_{\mathbf{B}}^2$	$(1-1/\zeta)^{2k}$	ζ	4.3, 4.4, 4.6
1	$\frac{2}{\lambda_{\min}^+ + \lambda_{\max}}$	-	-	$\\|\mathrm{E}[x_k - x_*]\\|_{\mathbf{B}}^2$	$(1-2/(\zeta+1))^{2k}$	ζ	4.3, 4.4, 4.6
1	1	-	-	$\mathrm{E}\left[\\|x_k - x_*\\|_{\mathbf{B}}^2\right]$	$(1-\lambda_{\min}^+)^{k}$	$1/\lambda_{\min}^+$	4.8
1	1	-	-	$\mathrm{E}[f(x_k)]$	$(1-\lambda_{\min}^+)^{k}$	$1/\lambda_{\min}^+$	4.10
2	1	τ	-	$\mathrm{E}\left[\\|x_k - x_*\\|_{\mathbf{B}}^2\right]$	$\left(1-\lambda_{\min}^+\left(2-\xi(\tau)\right)\right)^k$		5.1
2	$1/\xi(\tau)$	τ	-	$\mathrm{E}\left[\\|x_k - x_*\\|_{\mathbf{B}}^2\right]$	$\left(1-\frac{\lambda_{\min}^+}{\xi(\tau)}\right)^k$	$\xi(\tau)/\lambda_{\min}^+$	5.1
2	$1/\lambda_{\max}$	∞	-	$\mathrm{E}\left[\\|x_k - x_*\\|_{\mathbf{B}}^2\right]$	$(1-1/\zeta)^k$	ζ	5.1
3	1	-	$\frac{2}{1+\sqrt{0.99\lambda_{\min}^+}}$	$\\|\mathrm{E}[x_k - x_*]\\|_{\mathbf{B}}^2$	$\left(1-\sqrt{0.99\lambda_{\min}^+}\right)^{2k}$	$\sqrt{1/\lambda_{\min}^+}$	5.3
3	$1/\lambda_{\max}$	-	$\frac{2}{1+\sqrt{0.99/\zeta}}$	$\\|\mathrm{E}[x_k - x_*]\\|_{\mathbf{B}}^2$	$\left(1-\sqrt{0.99/\zeta}\right)^{2k}$	$\sqrt{\zeta}$	5.3

Table 1: Summary of the main complexity results. In all cases, $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$ (the projection of the starting point onto the solution space of the linear system). "Complexity" refers to the number of iterations needed to drive "Quantity" below some error tolerance $\epsilon > 0$ (we suppress a $\log(1/\epsilon)$ factor in all expressions in the "Complexity" column). In the table we use the following expressions: $\xi(\tau) = \frac{1}{\tau} + (1-\frac{1}{\tau})\lambda_{\max}$ and $\zeta = \lambda_{\max}/\lambda_{\min}^+$.

Part IV

Conclusion

Contributions

- 4 Equivalent stochastic reformulations of a linear system
  - Stochastic optimization
  - Stochastic fixed point problem
  - Stochastic linear system
  - Probabilistic intersection
- 3 Algorithms
  - Basic (SGD, stochastic Newton method, stochastic fixed point method, stochastic proximal point method, stochastic projection method, …)
  - Parallel
  - Accelerated
- Iteration complexity guarantees for various measures of success
  - Expected iterates (closed form)
  - L1 / L2 convergence
  - Convergence of f; ergodic …

Related Work

Basic method with unit stepsize and full rank A:

Robert Mansel Gower and P.R.
Randomized Iterative Methods for Linear Systems
SIAM J. Matrix Analysis & Applications 36(4):1660-1690, 2015

- 2017 IMA Fox Prize (2^{nd} Prize) in Numerical Analysis
- Most downloaded SIMAX paper

Removal of full rank assumption + duality:

Robert Mansel Gower and P.R.
Stochastic Dual Ascent for Solving Linear Systems
arXiv:1512.06890, 2015

Inverting matrices & connection to Quasi-Newton updates:

Robert Mansel Gower and P.R.
Randomized Quasi-Newton Methods are Linearly Convergent Matrix Inversion Algorithms
arXiv:1602.01768, 2016

Computing the pseudoinverse:

Robert Mansel Gower and P.R.
Linearly Convergent Randomized Iterative Methods for Computing the Pseudoinverse
arXiv:1612.06255, 2016

Application in machine learning:

Robert Mansel Gower, Donald Goldfarb and P.R.
Stochastic Block BFGS: Squeezing More Curvature out of Data
ICML 2016

THE END