Beyond convexity: solving nonlinear underdetermined equations

Boris Polyak, Andrey Tremba
Inst. Contr. Sci., Moscow

3 Oracles 2, ScolTech

December 28, 2016

$$P(x) = 0, \quad P\colon \mathbb{R}^n \to \mathbb{R}^m, \quad m \le n.$$

P is nonlinear smooth.

Two problems:

1. Existence of solutions: if $P(0) = 0$, when $P(x) = y$ is solvable?
2. Algorithm of finding a solution.
   The problem is nonconvex! E.g. $f(x) = \|P(x)\|^2$ is nonconvex for nonlinear $P(\cdot)$; $P(x)$ is *not* a gradient for $m < n$.

Basic Algorithm

k-th iteration

$$z^k = \arg\min\{\|z\| \colon P'(x^k)z = P(x^k)\},$$

$$x^{k+1} = x^k - \alpha_k z^k,$$

- $\|z\|$ is arbitrary norm,
- various ways to choose stepsize α_k.

Newton method

$$\alpha_k = 1, \quad x^{k+1} = x^k - (P'(x^k))^{-1} P(x^k)$$

Newton 1669 $\rightarrow$ Raphson 1690 $\rightarrow$ Fourier 1818 $\rightarrow$ Cauchy 1829 $\rightarrow$ Bennet 1916 $\rightarrow$ Kantorovich 1948 $\rightarrow$ Smale 1986 $\rightarrow$ Nesterov-Nemirovski 1994

Newton method	Damped Newton method
$\alpha_k = 1$	$\alpha_k < 1$
Local convergence	Global convergence
Fast convergence	Slow convergence

Combined methods

E.g. $f(x)$ is self-concordant function, $P(x) = \nabla f(x), \; f(x) \rightarrow \min$

History 2. $m < n$

Graves 1950. Assume $P(0) = 0$, and for $x \in S_\rho = \{\|x\| \le \rho\}$

$$\|P(x^a) - P(x^b) - A(x^a - x^b)\| \le c\|x^a - x^b\|,$$

also A is $\mathbb{R}^n$ *onto* $\mathbb{R}^m$ ($\sigma(A) = \mu > c > 0$), thus $\exists Q : AQ = I$. If r is small, then $\forall \|y\| \le r, \exists x^* : P(x^*) = y$, and it is achieved by method

$$x^{k+1} = x^k - \alpha Q P(x^k), \quad x^k \to x^*$$

Magaril-Ilyaev, Tikhomirov 2008

Ben-Israel 1966

$$x^{k+1} = x^k - (P'(x^k))^\dagger P(x^k),$$

Polyak 1964

Assumptions

$x^0 \in \mathbb{R}^n, \quad P : \mathbb{R}^n \to \mathbb{R}^m, \quad S_\rho = \{x : \|x - x^0\| \leq \rho\},$
$\|x\|_X \in \mathbb{R}^n, \; \|y\|_Y \in \mathbb{R}^m,$

1. P' exists, P' is Lipschitz on S_ρ,
2. $\|P'(x)^T h\| \geq \mu \|h\|, \; \mu > 0$ on S_ρ.

Lemma

$A : \mathbb{R}^n \to_{onto} \mathbb{R}^m \implies \exists \mu > 0 : \|A^T h\| \geq \mu \|h\|$, and $Ax = y$ has a solution $x : \|x\| \leq \frac{\|y\|}{\mu}$.

Need uniformly for $A = P'(x), x \in S_\rho$.

Algorithm

$$x^{k+1} = x^k - \alpha_k z^k, \quad z^k = \arg\min\{\|z\| : P'(x^k)z = P(x^k)\}$$

Note: z^k is well-defined, as $P'(x^k)z = P(x^k)$ has a solution $\|z^k\| \le \frac{\|P(x^k)\|}{\mu}$.

Main estimate:

$$\|P(x^{k+1})\| = \|P(x^k - \alpha z^k)\| \le |1-\alpha| \cdot \|P(x^k)\| + \frac{L\alpha^2}{2}\|z^k\|^2$$

Algorithm (cont)

Denote $u_k = \|P(x^k)\|$:

$$u_{k+1} \leq |1-\alpha|u_k + \frac{L\alpha^2}{2}\|z^k\|^2 \leq |1-\alpha|u_k + \frac{L\alpha^2}{2\mu^2}u_k^2$$

Choice of α:

- α_k is small $\implies$ existence of solution,
- α_k is "optimal" $\implies$ fast convergence.

Solvability

Theorem

If $\frac{\|P(x^0)\|}{\mu} = \frac{u_0}{\mu} < \rho$, *then* x^* *exists, and* $\|x^0 - x^*\| \leq \frac{u_0}{\mu}$.

Proof.

$\alpha \leq 1, \quad u_{k+1} \leq u_k(1 - \alpha(1 - \frac{L\alpha}{2\mu^2}u_0)) \leq qu_k, \; q < 1.$

Then

$u_k \to 0, \|x^{k+1} - x^k\| = \alpha\|z^k\| \leq \frac{\alpha}{\mu}u_k \leq cq^k, \; x^k \to x^*.$ □

Corollary

If $\rho = \infty$, *then* $P(x) = y$ *has a solution* $\forall y$*!*

Adaptive Newton

$$x^{k+1} = x^k - \alpha_k z^k, \quad z^k = \arg\min\{\|z\| : P'(x^k)z = P(x^k)\}$$

$$\alpha_k = \min\{1, \frac{\|P(x^k)\|}{L\|z^k\|^2}\}$$

is well-defined: $z^k = 0$ iff $P(x^k) = 0$, $\quad L = 0 \Rightarrow x^1 = x^*$.

Stage 1

$\alpha < 1 \Rightarrow u_{k+1} \leq u_k - \dfrac{\mu^2}{2L}$, finite number of iterations,

Stage 2

$\alpha = 1 \Rightarrow u_{k+1} \leq \dfrac{u_k^2}{2L}$, quadratic convergence if $\frac{u_k}{2L} < 1$.

For simplicity $\rho = \infty$. Advantage: μ is unknown!

Convergence rate

$$\|P(x^k)\| \leq \left[\|P(x^0)\| - \frac{\mu^2}{2L}k\right]_+ + \frac{2\mu^2}{L}2^{-(2^{[k-k_{\max}]_+})}, \ k \geq 0.$$

where

$$k_{\max} = \max\{0, \left\lceil \frac{2L}{\mu^2}\|P(x^0)\| - 2 \right\rceil\}.$$

How many steps to achieve accuracy ε:

$$N(\varepsilon) \leq k_{\max} + \log_2(\log_2\left(\frac{2\mu^2}{L\varepsilon}\right)), \quad \varepsilon < \frac{L}{4\mu^2}$$

A version of the algorithm

$$\alpha_k = \min\{1, \frac{\mu^2}{2L\|P(x^k)\|}\}$$

Then bound for $\|P(x^k)\|$ is the same, also

$$\|x^k - x^*\| \leq \frac{\mu}{L}\left([k_{\max} - k]_+ + 2^{\left(-2^{([k-k_{\max}]_+ - 1)} + 2\right)}\right), \; k \geq 0.$$

Note that μ is explicitly needed, and this algorithm in practice behaves worse.

Structured problems

Solve $P(x) = y,\ P(x)_i = \varphi(c_i^T x),\ c_i \in R^n, i = 1, \dots m$
Here $\varphi(t)$ is scalar function,

$$|\varphi'(t)| \geq \mu_\varphi > 0, \ \ |\varphi''(t)| \leq M, \ \ \forall t$$

Then $P'(x) = D(x)C,\ D(x) = \operatorname{diag}(c_i^T x)$ and the main inequality can be proved to be

$$u_{k+1} \leq (1-\alpha)u_k + \gamma \frac{\alpha^2 u_k^2}{2}, \gamma = \frac{M}{\mu_\varphi^2}.$$

Hence $u_{k+1} \leq u_k - \dfrac{1}{2\gamma}$ at Stage 1, it does not depend on C!

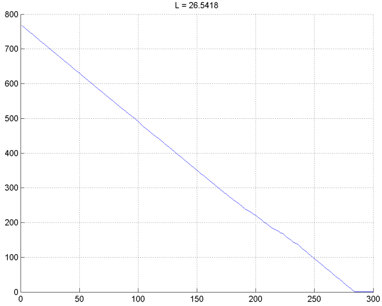
Numerical Example 1

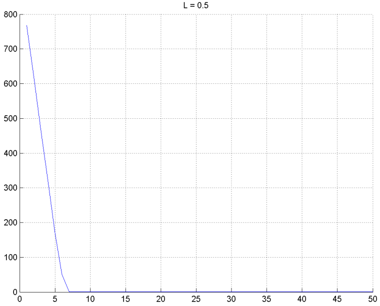
$$P(x) = \varphi(Cx), \;\; \varphi(t) = \frac{t}{1 + e^{-|t|}}$$

$|\varphi'(t)| \geq \frac{1}{2}, \;\; |\varphi''(t)| \leq \frac{1}{2},$
If consider $P(x)$ directly, then $L = \sup_t |\phi''(t)| \|C\|$, while in structured setup $M = \sup_t |\phi''(t)|$.

Numerical Example 1

$$n = 60, m = 21,\ L = 26.54,\ M = 0.5.$$





Solving subproblem for z^k

$$z^k = \arg\min\{\|z\| : P'(x^k)z = P(x^k)\}$$

$P'(x^k) = A, P(x^k) = b,\ Az = b$

1. Euclidean norm $\|\cdot\|_2$

$$z = A(AA^T)^{-1}b, \quad z^k = (P'(x^k))^\dagger P(x^k),$$

2. ℓ_1-norm $\|\cdot\|_1$

$$\min_{Az=b} \|z\|_1 \text{ is an LP,}$$

   with $\|z^*\|_0 \leq m$ - sparse solutions for $m \leq n$!

3. For ℓ_∞-norm the problem is also LP.

Scalar case. $m = 1$

$$f(x) = 0, \quad f : \mathbb{R}^n \to \mathbb{R}, \quad \rho = \infty.$$

$\nabla f(x)$ has Lipschitz constant L, $\|\nabla f(x)\| \geq \mu > 0, \forall x$.

Euclidean norm

$$z^k = \frac{f(x^k)}{\|\nabla f(x^k)\|^2} \nabla f(x^k)$$

$$x^{k+1} = \begin{cases} x^k - \dfrac{\operatorname{sign} f(x^k)}{L} \nabla f(x^k), & \text{if } \|\nabla f(x^k)\|^2 \leq L|f(x^k)|, \\ x^k - \dfrac{f(x^k)}{\|\nabla f(x^k)\|^2} \nabla f(x^k), & \text{otherwise.} \end{cases}$$

Scalar case (cont)

ℓ_1-norm: $z^k = \dfrac{f(x^k)}{\|\nabla f(x^k)\|_\infty} e^i,\ i = \arg\max_i |\nabla f(x^k)_i|,$

If $x^0 = 0$, then $\|x^k\|_0 \leq k$.

ℓ_∞-norm: $z^k = \dfrac{f(x^k)}{\|\nabla f(x^k)\|_1} \operatorname{sign}(\nabla f(x^k)).$

Scalar case: Inequality

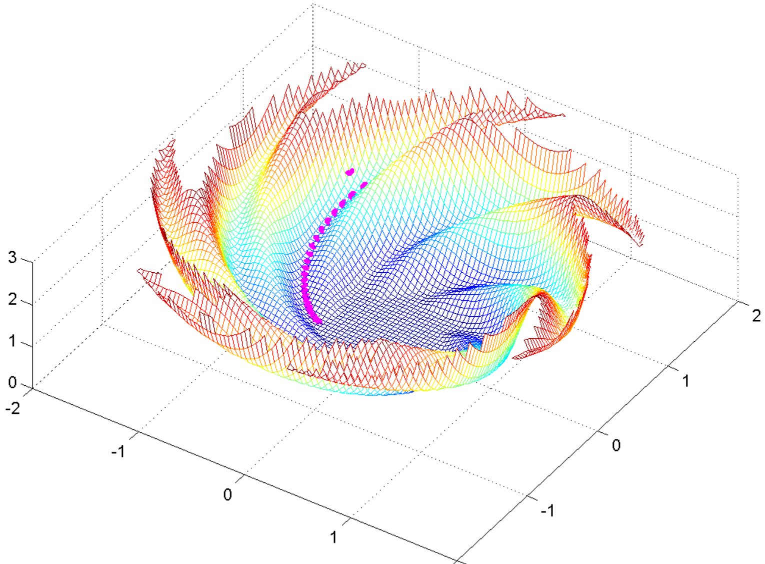
Solve $f(x) \leq 0$, $\;S = \{x : f(x) \geq 0\}$, $\nabla f(x)$ is Lipschitz on $\mathbb{R}^n$, $\|\nabla f(x)\| \geq \mu > 0$ on S, ℓ_2 norm.

Algorithm stops if $f(x^k) < 0$.

$$x^{k+1} = \begin{cases} x^k - \dfrac{\operatorname{sign} f(x^k)}{L} \nabla f(x^k), & \text{if } \|\nabla f(x^k)\|^2 \leq L|f(x^k)|, \\ x^k - \dfrac{f(x^k)}{\|\nabla f(x^k)\|^2} \nabla f(x^k), & \text{otherwise.} \end{cases}$$

New: nonconvex $f(\cdot)$.

Numerical Example 2



$P'(x)$ is non-singular at a point

Assumption $\|P'(x)^T h\| \geq \mu\|h\|, \;\; \forall x \in S_\rho$ is hard to check.

Suppose

$$\|P'(x^0)^T h\| \geq \mu_0\|h\|, \; \mu_0 > 0.$$

Then $\|P'(x)^T h\| \geq (\mu_0 - L\rho)\|h\|$ for $x \in S_\rho$.

Theorem

If $P(0) = 0, \sigma_{\min}(P'(0)) \geq \mu_0 > 0, \; P'(x)$ *is Lipschitz on* S_ρ*, and* $\rho > \frac{\mu_0}{2L}$*, then* $P(x) = y$ *has a solution for all* y*:*

$$\|y\| \leq \frac{\mu_0^2}{4L}$$

Quadratic equations

$$P_i(x) = \frac{1}{2}(A_i x, x) + (b_i, x), \quad P(x) = (P_1(x); \dots; P_m(x)),$$

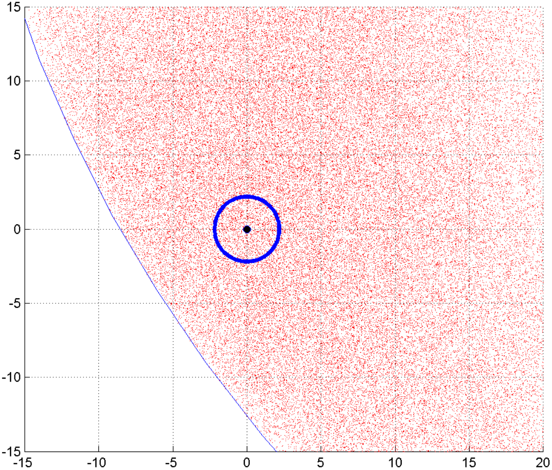
Solve $P(x) = y, x \in \mathbb{R}^n$

$$P'(x) = \begin{bmatrix} x^T A_1 + b_1^T \\ \vdots \\ x^T A_m + b_m^T \end{bmatrix}, \quad P'(0) = \begin{bmatrix} b_1^T \\ \vdots \\ b_m^T \end{bmatrix},$$

Use Euclidean norm for x, denote $A = \sum_{i=1}^m A_i^T A_i$, then $L = \|A\|^{1/2}, \mu = \sigma_{\min}(P'(0))$, and the Theorem above can be applied.

Numerical Example 3

$$n = 4,\ m = 2,\ \mu_0 = 9.6,\ \ L = 10.7,\ \ \|y\| \le \frac{\mu_0^2}{4L} = 2.2$$



Extensions

1. Approximate solution of the subproblem.
2. Stepsize if L is unknown (Armijo-like).
3. Not arbitrary solution, but a close to x^0 one, e.g. $\min_{P(x)=0} \|x - x^0\|$. (Too strong assumptions, local resutls.)
4. Analog of self-concordant operators, compare Example 1.
5. Regularization for equations $=$?

Thank you for attention!