Flexible Gauss-Newton Methods[1]

Optimization Without Borders Conference

Nikita Yudin

MIPT, FRC CSC RAS

Sochi, 17 July 2021

[1] Yudin N., Gasnikov A. Flexible Modification of Gauss-Newton Method and Its Stochastic Extension //WIAS Preprint No. 2813. — 2021. doi: 10.20347/WIAS.PREPRINT.2813

The problem

System of nonlinear equations

Consider a smooth multidimensional mapping $F:\ \mathbb{R}^n \to \mathbb{R}^m$:

$$F(x) = \mathbf{0}_m,\ \mathbf{0}_m = (0,\ \ldots, 0)^T.$$

The problem

System of nonlinear equations

Consider a smooth multidimensional mapping $F: \mathbb{R}^n \to \mathbb{R}^m$:

$$F(x) = \mathbf{0}_m, \; \mathbf{0}_m = (0, \; \ldots, 0)^T.$$

Solving equations via optimization

The next minimization problem of merit function is considered as a relaxation of the problem of solving systems of equations:

$$\min_{x \in \mathbb{R}^n} \left\{ f_1(x) \stackrel{\text{def}}{=} \|F(x)\| \right\},$$

where $\| \; \|$ is the standard Euclidean norm.

Settings

Assumption (1)

$F(x)$ is smooth with Lipschitz continuous Jacobian:

$$\exists L_F > 0: \left\| F^{'}(y) - F^{'}(x) \right\|_F \leq L_F \|y - x\|, \ \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n,$$

where $\| \ \|_F$ is Frobenius norm.

Settings

Assumption (1)

$F(x)$ is smooth with Lipschitz continuous Jacobian:

$$\exists L_F > 0 : \left\| F'(y) - F'(x) \right\|_F \leq L_F \|y - x\|, \ \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n,$$

where $\| \ \|_F$ is Frobenius norm.

Local model[b]:

$$f_1(y) \leq \psi_{x,L}(y) \stackrel{\text{def}}{=} \frac{f_1(x)}{2} + \frac{(\phi(x, y))^2}{2 f_1(x)} + \frac{L}{2}\|y - x\|^2, \ L \geq L_F,$$

$$\phi(x, y) \stackrel{\text{def}}{=} \left\| F(x) + F'(x)(y - x) \right\|, \ (x, y) \in \mathbb{R}^n \times \mathbb{R}^n.$$

[b] Yu Nesterov. Flexible modification of Gauss-Newton method //CORE Discussion Papers. — 2021.

Proposed algorithm

Local model

$$f_1(y) \leq \psi_{x,L,\tau}(y) \stackrel{\text{def}}{=} \frac{\tau}{2} + \frac{(\phi(x,y))^2}{2\tau} + \frac{L}{2}\|y - x\|^2, \ (x,y) \in \mathbb{R}^n \times \mathbb{R}^n,$$
$$L \geq L_F, \ \tau > 0.$$

Proposed algorithm

Local model

$$f_1(y) \leq \psi_{x,L,\tau}(y) \stackrel{\text{def}}{=} \frac{\tau}{2} + \frac{(\phi(x,y))^2}{2\tau} + \frac{L}{2}\|y-x\|^2, \ (x,y) \in \mathbb{R}^n \times \mathbb{R}^n,$$
$$L \geq L_F, \ \tau > 0.$$

Exact update

$$T_{L,\tau}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{y\in\mathbb{R}^n} \psi_{x,L,\tau}(y) = x - \left(F'(x)^T F'(x) + \tau L\right)^{-1} F'(x)^T F(x),$$

Proposed algorithm

Algorithm 1 Flexible Gauss-Newton Method

1: **Input:** $x_0 \in \mathbb{R}^n$, $L_0 = L \in (0, L_F]$.
2: **for** $k = 0, 1,$ **to** $N - 1$ **do**
3: define $\tau_k > 0$, $\varepsilon_k \geq 0$.
4: compute $x_{k+1} : \ \psi_{x_k, L_k, \tau_k}(x_{k+1}) - \psi_{x_k, L_k, \tau_k}(T_{L_k, \tau_k}(x_k)) \leq \varepsilon_k$.
5: **if** $f_1(x_{k+1}) > \psi_{x_k, L_k, \tau_k}(x_{k+1})$ **then**
6: set $L_k := 2L_k$ and return to line 3.
7: **end if**
8: $L_{k+1} = \max\left\{\frac{L_k}{2},\ L\right\}$.
9: **end for**
10: **Output:** x_N.

Properties

We established global convergence properties for this method in listed below terms:

- norm of the proximal gradient mapping: $\left\| L_k \left(T_{L_k,\tau_k}(x_k) - x_k \right) \right\|$;
- local decrease:
$\Delta_r(x_k) \stackrel{\text{def}}{=} f_2(x_k) - \min\limits_{y \in \mathbb{R}^n} \left\{ (\phi(x_k, y))^2 : \ \|y - x_k\| \le r \right\}, r > 0,$
$f_2(x) \stackrel{\text{def}}{=} (f_1(x))^2, \ x \in \mathbb{R}^n.$

Properties

Theorem (1)

Suppose that assumption 1 is satisfied, $k \in \mathbb{N}$, $r > 0$. Then algorithm 1 with $\tau_k = f_1(x_k)$, $\varepsilon_k = \varepsilon \geq 0$, has the following estimates:

$$\begin{cases} \frac{8L_F^2}{L}\left(\varepsilon + \frac{(f_1(x_0) - f_1(x_k))}{k}\right) \geq \min\limits_{i \in \overline{0,k-1}} \left\{ \left\| 2L_F \left(T_{2L_F, f_1(x_i)}(x_i) - x_i \right) \right\|^2 \right\}; \\ L_F \left(\varepsilon + \frac{(f_1(x_0) - f_1(x_k))}{k}\right) \geq \min\limits_{i \in \overline{0,k-1}} \left\{ 2 (rL_F)^2 \varkappa \left(\frac{\Delta_r(x_i)}{4 f_1(x_i) r^2 L_F} \right) \right\}; \end{cases}$$

where $\varkappa(t) = \frac{t^2}{2} \mathbb{1}_{\{t \in [0,1]\}} + \left(t - \frac{1}{2}\right) \mathbb{1}_{\{t > 1\}},\ \mathbb{1}_{\{\cdot\}}$ *— set indicator function.*

Properties

Theorem (2)

Suppose that assumption 1 is satisfied, Jacobian is bounded: $\left\|F^{'}(x)\right\| \leq M_F$ *for all* $x \in \mathbb{R}^n$, *and the solution* $x^* \in \mathbb{R}^n,\ F(x^*) = \mathbf{0}_m$ *with* $\sigma_{\min}\left(F^{'}(x^*)\right) \geq \varsigma > 0$ *exists. Then algorithm 1 with* $\tau_k = f_1(x_k)$, $\varepsilon_k = 0$ *in region*

$$\|x_k - x^*\| \leq \min\left\{\frac{2\varsigma}{5L_F},\ \frac{1}{12L_F}\left((3M_F + 5\varsigma) - \sqrt{(3M_F + 5\varsigma)^2 - 24\varsigma^2}\right)\right\},\ k \in \mathbb{Z}_+$$

superlinearly converges

$$\|x_{k+1} - x^*\| \leq \frac{\frac{3L_F\|x_k - x^*\|^2}{2} + \|x_k - x^*\|\sqrt{f_1(x_k)L_k + \frac{L_F^2\|x_k - x^*\|^2}{4}}}{\varsigma - L_F\|x_k - x^*\|} \leq \|x_k - x^*\|,$$
$$x_{k+1} \in \mathbb{R}^n,\ f_1(x_k) = O\left(\|x_k - x^*\|\right).$$

Properties

Assumption (2)

Suppose the following PL condition is satisfied:

$$\exists \mu > 0,\ \sigma_{\min}(F^{'}(x)^{T}) \geq \sqrt{\mu},\ \forall x \in \mathbb{R}^{n},$$

where $\sigma_{\min}(\cdot)$ *is the smallest singular value of a matrix.*

Properties

Assumption (2)

Suppose the following PL condition is satisfied:

$$\exists \mu > 0,\ \sigma_{\min}(F'(x)^T) \geq \sqrt{\mu},\ \forall x \in \mathbb{R}^n,$$

where $\sigma_{\min}(\cdot)$ is the smallest singular value of a matrix.

Theorem (3)

Assume that assumptions 1 and 2 are held for algorithm 1 with $\tau_k = f_1(x_k)$. Then any sequence $\{x_k\}_{k \in \mathbb{Z}_+}$ has the property:

$$f_1(x_{k+1}) \leq \varepsilon_k + \begin{bmatrix} \frac{f_1(x_k)}{2} + \frac{L_F}{\mu} f_2(x_k) \leq \frac{3}{4} f_1(x_k), \text{ if } f_1(x_k) \leq \frac{\mu}{4L_F}; \\ f_1(x_k) - \frac{\mu}{16 L_F}, \text{ otherwise.} \end{bmatrix}$$

Experiments

Define $x^T \overset{\text{def}}{=} (x^1, \dots, x^n)$, $x \in \mathbb{R}^n$, $n = 100$. We consider two benchmark systems to test main features of presented methods:

- The system of equations based on Rosenbrock-Skokov function: $F_R(x) = \mathbf{0}_m$, where $m = 2n - 2$ and $F_{R_{2i-1}}(x) \overset{\text{def}}{=} i(x^i - (x^{i+1})^2)$, $F_{R_{2i}}(x) \overset{\text{def}}{=} 1 - x^{i+1}$, $i = \overline{1, n-1}$.
- Hat system: $F_H(x) \overset{\text{def}}{=} \nabla(\|x\|^2 - 1)^2 = \mathbf{0}_m$.

Experiments

We use three variants of proposed Gauss-Newton method with $\varepsilon_k = 0$ differing by τ_k:

1) $\tau_k = f_1(x_k)$;

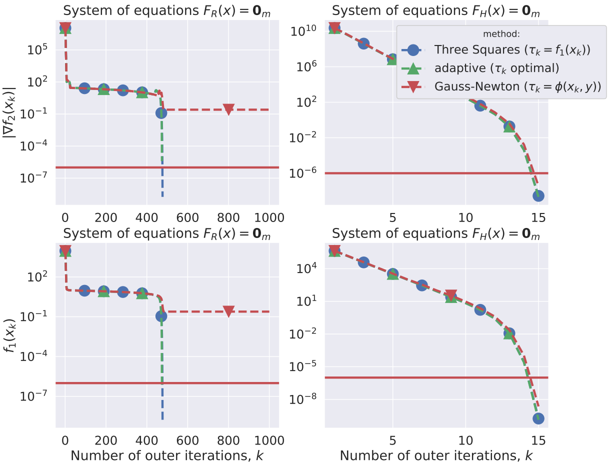
2) near-optimal τ_k:

$$\tau_k \approx \operatorname*{argmin}_{\tau>0} \psi_{x_k, L_k, \tau}(T_{L_k,\tau}(x_k)) = \operatorname*{argmin}_{\tau>0} \left\{ \frac{\tau}{2} + \frac{f_2(x_k)}{2\tau} - \right.$$

$$\left. - \frac{1}{2\tau} \left\langle \left(F'(x_k)^T F'(x_k) + \tau L_k \right)^{-1} F'(x_k)^T F(x_k),\ F'(x_k)^T F(x_k) \right\rangle \right\};$$

3) $\tau_k = \phi(x_k, y)$.

Experiments



The performance of solving systems of non-linear equations. Horizontal line — desired tolerance 10^{-6}.

Happy birthday to Yurii Evgen’evich and Vladimir Yur’evich!

Thank you for your attention! Any questions?