Adaptive regularization without function values

Sadok Jerad[1,2] with Serge Gratton [1,2] and Philippe L.Toint[3]

[1]IRIT-APO

[2]Toulouse INP-ENSEEIHT

[3]University of Namur

Siam OP, May 2023

Table of Contents

Problem motivation

Approximately solving the non convex problem

$$\boxed{\min_{x \in \mathbb{R}^n} f(x)}$$

with a **robust** deterministic algorithm.

Problem motivation

Approximately solving the non convex problem

$$\boxed{\min_{x\in\mathbb{R}^n} f(x)}$$

with a **robust** deterministic algorithm. **Advantages:**

- Inexact arithmetic.
- Results from simulations.

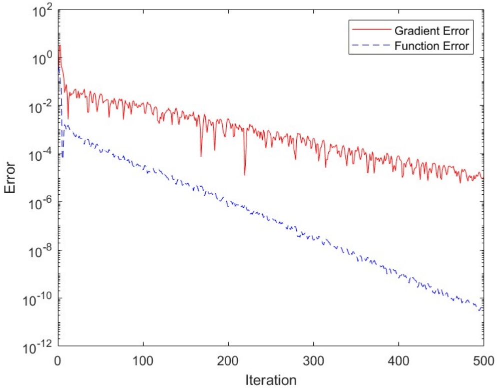
Problem motivation

Approximately solving the non convex problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

with a **robust** deterministic algorithm. **Advantages:**

- Inexact arithmetic.
- Results from simulations.



$\rightarrow$ Devise an algorithm that do not evaluate the function.

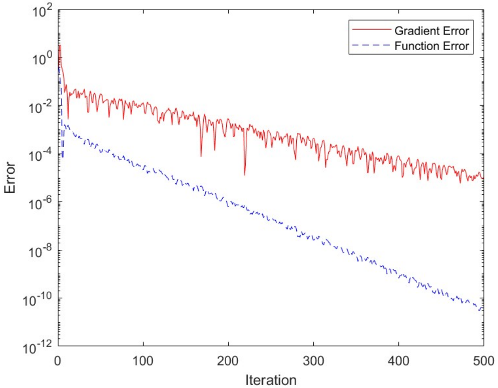
Problem motivation

Approximately solving the non convex problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

with a **robust** deterministic algorithm. **Advantages:**

- Inexact arithmetic.
- Results from simulations.



$\rightarrow$ Devise an algorithm that do not evaluate the function. Adaptive gradient methods :

- Adagrad [?], WNGrad [?].
- Adam [?].

Contributions

Determistic OFFO

- pth order information.
- Optimal complexity rate.

to approximately solve the non-convex unconstrained problem

$$\min_{x \in \mathbb{R}^n} f(x).$$

Contributions

Determistic OFFO

- pth order information.
- Optimal complexity rate.

to approximately solve the non-convex unconstrained problem

$$\min_{x \in \mathbb{R}^n} f(x).$$

OFFO : Objective Free Function Optimization Current Higher Order OFFO :

Trust region of **?**. Matches $\mathcal{O}\left(\epsilon^{-2}\right)$ rate of Vanilla TR when searching

$$\|\nabla f(x_\epsilon)\| \leq \epsilon.$$

Table of Contents

ARp methods

Assumptions

AS.1 f is a differentiable function.

AS.2 $\nabla^p f$ is a Lipschitz function with constant L_p.

AS.3 f is lower bounded by f_{low}.

AS.4 $\min_{\|d\|\leq 1} \nabla_x^i f(x)[d]^i \geq -\kappa_{\text{high}}$ for $i \in \{2, \dots, p\}$.

$$T_{f,p}(x,s) \stackrel{\text{def}}{=} f(x) + \sum_{i=1}^{p} \frac{1}{i!} \nabla_x^i f(x)[s]^i,$$

$$m_k(s) \stackrel{\text{def}}{=} T_{f,p}(x_k, s) + \frac{\sigma_k}{(p+1)!} \|s_k\|^{p+1}.$$

By adapting σ_k **with f values** and under AS.1-AS.3, complexity rate in $\mathcal{O}\left(\epsilon^{-p+1/p}\right)$ [?].

Algorithm presentation

Algorithm 1: OFFO-ARp Algorithm

Input: x_0, $\epsilon \in (0, 1]$, $\sigma_0 > 0$, $\theta_1 > 1$ and $k = 0$.
Output: x_ϵ: an approximate first order point.
Evaluate $\{\nabla_x^i f(x_0)\}_{i=1}^p$.

ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

Université Fédérale Toulouse Midi-Pyrénées

Algorithm presentation

Algorithm 2: $\mathrm{OFFO\text{-}AR}p$ Algorithm

Input: x_0, $\epsilon \in (0,1]$, $\sigma_0 > 0$, $\theta_1 > 1$ and $k = 0$.
Output: x_ϵ: an approximate first order point.
Evaluate $\{\nabla_x^i f(x_0)\}_{i=1}^p$.
while $\|\nabla_x^1 f(x_k)\| > \epsilon$ **do**

Compute s_k

$$\|\nabla_s^1 T_{f,p}(x_k, s_k)\| \le \frac{\theta_1 \sigma_k}{p!}\|s_k\|^p, \quad m_k(s_k) \le m_k(0)$$

$x_{k+1} \leftarrow x_k + s_k$

$\sigma_{k+1} \leftarrow \sigma_k + \sigma_k \|s_k\|^{p+1}$

$k \leftarrow k + 1$

Generate new tensors $\{\nabla_x^i f(x_k)\}_{i=1}^p$.

Comments

- s_k approximate local minimum of m_k.
- $p = 1$,

$$\sigma_{k+1} = \sigma_k + \frac{\|g_k\|^2}{\sigma_k}$$

retrives WNGRAD [?].
- σ_k is a non-decreasing sequence.
- No decrease ratio computation $\rho_k = \frac{f(x_k)-f(x_k+s_k)}{m_k(0)-m_k(s_k)}$.

Basic properties

Recall that

$$T_{f,p}(x_k, s) \stackrel{\text{def}}{=} f(x_k) + \sum_{i=1}^{p} \frac{1}{i!} \nabla_x^i f(x_k)[s]^i.$$

Lipschitz error bound [**?**]:

$$\boxed{f(x_{k+1}) - T_{f,p}(x_k, s) \leq \frac{L_p}{(p+1)!} \|s_k\|^{p+1},}$$

$$\boxed{\|\nabla_x^1 f(x_{k+1}) - \nabla_s^1 T_{f,p}(x_k, s)\| \leq \frac{L_p}{(p+1)!} \|s_k\|^{p+1}.}$$

Basic properties

Recall that

$$T_{f,p}(x_k, s) \stackrel{\text{def}}{=} f(x_k) + \sum_{i=1}^{p} \frac{1}{i!} \nabla_x^i f(x_k)[s]^i.$$

Lipschitz error bound [?]:

$$\boxed{f(x_{k+1}) - T_{f,p}(x_k, s) \leq \frac{L_p}{(p+1)!} \|s_k\|^{p+1},}$$

$$\boxed{\|\nabla_x^1 f(x_{k+1}) - \nabla_s^1 T_{f,p}(x_k, s)\| \leq \frac{L_p}{(p+1)!} \|s_k\|^{p+1}.}$$

Model's decrease

$$f(x_k) - T_{f,p}(x_k, s_k) \geq \frac{\sigma_k}{(p+1)!} \|s_k\|^{p+1}.$$

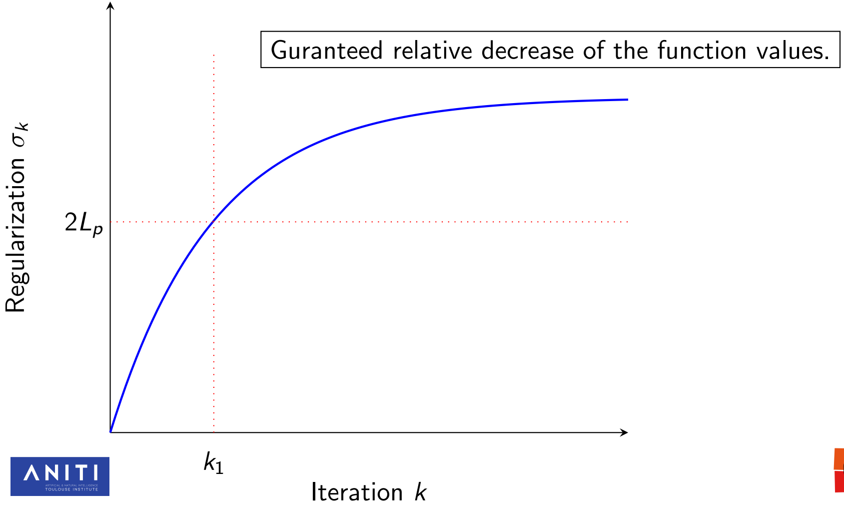
If $\sigma_p \geq 2L_p$,

$$f(x_p) - f(x_{p+k+1}) \geq \sum_{j=p}^{p+k} \frac{\sigma_j}{2(p+1)!} \|s_j\|^{p+1}.$$

Evolution of the regularization

Define

$$k_1 \stackrel{\text{def}}{=} \inf\{k \geq 0 \mid \sigma_k \geq 2L_p\}.$$



Proof's steps

Recall

$$k_1 \overset{\text{def}}{=} \inf\{k \geq 0 \mid \sigma_k \geq 2L_p\}.$$

Steps

- Show that $k_1 = \mathcal{O}\left(\epsilon^{-(p+1)/p}\right)$.
- Bound σ_{k_1}.
- Bound $f(x_{k_1})$.
- Proof that $\sigma_k \leq \sigma_{\max}$ when $k \geq k_1$.

Can reuse classical arguments from [?] and derive $\mathcal{O}\left(\epsilon^{-(p+1)/p}\right)$.

$k_1 = \mathcal{O}\left(\epsilon^{-(p+1)/p}\right)$

Lipschitz error + Inexact condition on the step

$$\boxed{\|\nabla_x^1 f(x_{k+1})\| \leq \frac{L_p + \theta_1 \sigma_k}{p!} \|s_k\|^p.}$$

$k_1 = \mathcal{O}\left(\epsilon^{-(p+1)/p}\right)$

Lipschitz error + Inexact condition on the step

$$\boxed{\|\nabla_x^1 f(x_{k+1})\| \leq \frac{L_p + \theta_1 \sigma_k}{p!} \|s_k\|^p.}$$

Proof

$$2L_p \geq \sigma_{k_1-1} = \sigma_0 + \sum_{j=0}^{k_1-2} \sigma_j \|s_j\|^{p+1} \geq \sum_{j=0}^{k_1-2} \sigma_j \left(\frac{p!\|\nabla_x^1 f(x_{j+1})\|}{L_p + \theta_1 \sigma_j}\right)^{\frac{p+1}{p}}$$

$$\geq \sum_{j=0}^{k_1-2} \frac{1}{\sigma_j^{\frac{1}{p}}} \left(\frac{p!\|\nabla_x^1 f(x_{j+1})\|}{\frac{L_p}{\sigma_j} + \theta_1}\right)^{\frac{p+1}{p}}$$

$$\geq \sum_{j=0}^{k_1-2} \frac{1}{(2L_p)^{\frac{1}{p}}} \left(\frac{p!\|\nabla_x^1 f(x_{j+1})\|}{\frac{L_p}{\sigma_0} + \theta_1}\right)^{\frac{p+1}{p}} \geq \frac{k_1 - 1}{(2L_p)^{\frac{1}{p}}} \left(\frac{p!\epsilon}{\frac{L_p}{\sigma_0} + \theta_1}\right)^{\frac{p+1}{p}}.$$

$k_1 = \mathcal{O}\left(\epsilon^{-(p+1)/p}\right)$

Lipschitz error + Inexact condition on the step

$$\boxed{\|\nabla_x^1 f(x_{k+1})\| \leq \frac{L_p + \theta_1\sigma_k}{p!}\|s_k\|^p.}$$

Proof

$$2L_p \geq \sigma_{k_1-1} = \sigma_0 + \sum_{j=0}^{k_1-2}\sigma_j\|s_j\|^{p+1} \geq \sum_{j=0}^{k_1-2}\sigma_j\left(\frac{p!\|\nabla_x^1 f(x_{j+1})\|}{L_p+\theta_1\sigma_j}\right)^{\frac{p+1}{p}}$$

$$\geq \sum_{j=0}^{k_1-2}\frac{1}{\sigma_j^{\frac{1}{p}}}\left(\frac{p!\|\nabla_x^1 f(x_{j+1})\|}{\frac{L_p}{\sigma_j}+\theta_1}\right)^{\frac{p+1}{p}}$$

$$\geq \sum_{j=0}^{k_1-2}\frac{1}{(2L_p)^{\frac{1}{p}}}\left(\frac{p!\|\nabla_x^1 f(x_{j+1})\|}{\frac{L_p}{\sigma_0}+\theta_1}\right)^{\frac{p+1}{p}} \geq \frac{k_1-1}{(2L_p)^{\frac{1}{p}}}\left(\frac{p!\epsilon}{\frac{L_p}{\sigma_0}+\theta_1}\right)^{\frac{p+1}{p}}.$$

$$k_1 = \mathcal{O}\left(\epsilon^{\frac{-(p+1)}{p}}\right).$$

Bounding σ_{k_1} and $f(x_{k_1})$

$$k_1 \stackrel{\text{def}}{=} \inf\{k \geq 0 \mid \sigma_k \geq 2L_p\}.$$

As for $i \in \{2, \ldots, p\}$

$$\min_{\|d\| \leq 1} \nabla_x^i f(x)[d]^i \geq -\kappa_{\text{high}}.$$

Step's Bound

$$\|s_k\| \leq 2\eta + 2\left(\frac{(p+1)!\|\nabla_x^1 f(x_k)\|}{\sigma_k}\right)^{\frac{1}{p}},$$

with η depending on κ_{high} only.

Bounding σ_{k_1} and $f(x_{k_1})$

$$k_1 \stackrel{\text{def}}{=} \inf\{k \geq 0 \mid \sigma_k \geq 2L_p\}.$$

As for $i \in \{2, \ldots, p\}$

$$\min_{\|d\| \leq 1} \nabla_x^i f(x)[d]^i \geq -\kappa_{\text{high}}.$$

Step's Bound

$$\|s_k\| \leq 2\eta + 2\left(\frac{(p+1)!\|\nabla_x^1 f(x_k)\|}{\sigma_k}\right)^{\frac{1}{p}},$$

with η depending on κ_{high} only.
Since

$$\sigma_{k_1} = \sigma_{k_1-1} + \sigma_{k_1-1}\|s_{k_1-1}\|^{p+1}$$

$+$ other properties

$$\sigma_{k_1} \leq \kappa_c.$$

Elements of the rest of the proof

Bound on $f(x_{k_1})$

$$k_1 \stackrel{\text{def}}{=} \inf\{k \geq 0 \mid \sigma_k \geq 2L_p\},\ \sigma_{k_1} \leq \kappa_c,$$

$$\kappa_c \geq \sigma_{k_1} = \sigma_0 + \sum_{j=0}^{k_1-1} \sigma_j \|s_j\|^{p+1} \geq \sigma_0 + \sigma_0 \sum_{j=0}^{k_1-1} \|s_j\|^{p+1}$$

Lipschitz error bound in $\|s_k\|^{p+1}$, $f(x_{k_1})$ is bounded.

Elements of the rest of the proof

Bound on $f(x_{k_1})$

$$k_1 \stackrel{\text{def}}{=} \inf\{k \geq 0 \mid \sigma_k \geq 2L_p\},\ \sigma_{k_1} \leq \kappa_c,$$

$$\kappa_c \geq \sigma_{k_1} = \sigma_0 + \sum_{j=0}^{k_1-1} \sigma_j \|s_j\|^{p+1} \geq \sigma_0 + \sigma_0 \sum_{j=0}^{k_1-1} \|s_j\|^{p+1}$$

Lipschitz error bound in $\|s_k\|^{p+1}$, $f(x_{k_1})$ is bounded.

Bound on σ_k

$$f(x_{k_1}) - f_{\text{low}} \geq f(x_{k_1}) - f(x_{k_1+k+1}) \geq \sum_{j=k_1}^{k_1+k} \frac{\sigma_j}{2(p+1)!}\|s_j\|^{p+1} = \frac{\sigma_{k_1+k+1} - \sigma_{k_1}}{2(p+1)!}.$$

First order complexity

Standard arguments of [?] + derived results,

$\mathcal{O}\left(\epsilon^{\frac{-(p+1)}{p}}\right)$ iterations required to reach $\|\nabla_x^1 f(x)\| \leq \epsilon$.

First order complexity

Standard arguments of [?] + derived results,

$\mathcal{O}\left(\epsilon^{\frac{-(p+1)}{p}}\right)$ iterations required to reach $\|\nabla_x^1 f(x)\| \leq \epsilon$.

Comments

- Same rate as vanilla ARp [?].
- For $p = 2$, equal to ARC method [?]. Optimal rate for second order methods [??].
- Optimal rates for exact pth order methods [?].

ANITI

Université Fédérale Toulouse Midi-Pyrénées

Algorithm 3: MOFFO-ARp Algorithm

Input: x_0, ϵ_1, $\epsilon_2 \in (0,1]$, $\sigma_0 > 0$, $\theta_1, \theta_2 > 1$ and $k = 0$.

Output: x_ϵ: an approximate second order point.

Evaluate $\{\nabla_x^i f(x_0)\}_{i=1}^p$.

Algorithm 4: MOFFO-ARp Algorithm

Input: x_0, ϵ_1, $\epsilon_2 \in (0,1]$, $\sigma_0 > 0$, $\theta_1, \theta_2 > 1$ and $k = 0$.

Output: x_ϵ: an approximate second order point.

Evaluate $\{\nabla_x^i f(x_0)\}_{i=1}^p$.

while $\|\nabla_x^1 f(x_k)\| > \epsilon$ *and* $\max\left(0, -\lambda_{\min}[\nabla_x^2 f(x_k)]\right) > \epsilon_2$ **do**

Compute s_k

$$\|\nabla_s^1 T_{f,p}(x_k, s_k)\| \leq \frac{\theta_1 \sigma_k}{p!}\|s_k\|^p, \quad m_k(s_k) \leq m_k(0)$$

$$\max\left(0, -\lambda_{\min}[\nabla_s^2 T_{f,p}(x_k, s_k)]\right) \leq \frac{\theta_2 \sigma_k}{(p-1)!}\|s_k\|^{p-1}$$

$x_{k+1} \leftarrow x_k + s_k$

$\sigma_{k+1} \leftarrow \sigma_k + \sigma_k \|s_k\|^{p+1}$

$k \leftarrow k + 1$

Generate new tensors $\{\nabla^i f(x_k)\}_{i=1}^p$

Complexity rate derivation

Recall

$$k_1 \stackrel{\text{def}}{=} \inf\{k \geq 0 \mid \sigma_k \geq 2L_p\}.$$

Steps

- Show that $k_1 = \mathcal{O}\left(\epsilon_1^{-(p+1)/p}, \epsilon_2^{-(p+1)/(p-1)}\right)$.
- Bound σ_{k_1} and $f(x_{k_1})$.
- Proof that $\sigma_k \leq \sigma_{\max}$ when $k \geq k_1$.

Complexity rate derivation

Recall

$$k_1 \stackrel{\text{def}}{=} \inf\{k \geq 0 \mid \sigma_k \geq 2L_p\}.$$

Steps

- Show that $k_1 = \mathcal{O}\left(\epsilon_1^{-(p+1)/p}, \epsilon_2^{-(p+1)/(p-1)}\right)$.
- Bound σ_{k_1} and $f(x_{k_1})$.
- Proof that $\sigma_k \leq \sigma_{\max}$ when $k \geq k_1$.

Comments

- Same complexity results as standard ARp [**?**].
- For $p = 2$, $\mathcal{O}\left(\epsilon_1^{-3/2}, \epsilon_2^{-3}\right)$.

Table of Contents

Regularization parameter in practice

Recall that

$$\sigma_{k+1} \leftarrow \sigma_k + \sigma_k \|s_k\|^{p+1}.$$

No local adaptivity

Regularization parameter in practice

Recall that

$$\sigma_{k+1} \leftarrow \sigma_k + \sigma_k \|s_k\|^{p+1}.$$

No local adaptivity

$$\nu_{k+1} \leftarrow \nu_k + \nu \|s_k\|^{p+1}.$$

Propose

$$\sigma_k \in [\vartheta \sigma_k, \max(\nu_k, \mu_k)] \quad , \vartheta \in (0, 1].$$

With $\mu_k \lesssim L_p$. Still same complexity rates. See [**?**].

Numerical experiments framework

- A set of 117 small dimensional CUTEst problems (OPM : Matlab library).
- Baseline : standard AR2 of **?** with function values.
- Increasing levels of relative Gaussian noise (both in function values and derivatives): 0%, 5%, 15%, 25%, 50%. E.g: $[g_k]_i = [\nabla_x(x_k)]_i(1 + \phi\mathcal{N}(0,1))$.
- Approximate first order point with $\epsilon = 1\mathrm{e}{-6}$.

Metrics

- a performance measure : π_{algo} (See [**?**]).
- a reliability ratio : ρ_{algo}.

Numerical performances

With exact values

Metric / Algorithm	π_{algo}	ρ_{algo}
AR2	0.97	95.8
OFFOAR2	0.79	88.24

When f is available and no noise, better use standard algorithm.

Numerical performances

With exact values

Algorithm \ Metric	π_{algo}	ρ_{algo}
AR2	0.97	95.8
OFFOAR2	0.79	88.24

When f is available and no noise, better use standard algorithm.

ρ_{algo} for noisy experiences

Algorithm \ Noise level	5%	10%	25%	50%
AR2	43.7	31.26	26.3	10
OFFOAR2	86.05	82.1	79.83	68.99

Better performance for noisy problems without specific adaptation.

Table of Contents

Conclusion

Recap

- Adaptive regularization methods with no function evaluation.
- Optimal complexity.
- Promising initial numerical results in noisy settings.
- Sharp Bounds [?].

Conclusion

Recap

- Adaptive regularization methods with no function evaluation.
- Optimal complexity.
- Promising initial numerical results in noisy settings.
- Sharp Bounds [?].

Extensions

- β Hölder pth derivative.
- Non euclidean norm [?].

Accepted by Siopt.

Perspectives

Probabilistic framework [**??**] ?

For more details; See `https://arxiv.org/pdf/2203.09947.pdf`.

Perspectives

Probabilistic framework [??] ?

For more details; See `https://arxiv.org/pdf/2203.09947.pdf`. Thank for your attention.

References I