Average Curvature FISTA for Nonconvex Smooth Composite Optimization Problems

Jiaming Liang

School of Industrial and Systems Engineering
Georgia Institute of Technology

Joint work with Renato Monteiro

MOPTA, Lehigh University - August 2, 2021

- J. Liang and R. D. C. Monteiro. An average curvature accelerated composite gradient method for nonconvex smooth composite optimization problems. SIAM Journal on Optimization, 31(1):217-243, 2021.
- J. Liang and R. D. C. Monteiro. Average Curvature FISTA for Nonconvex Smooth Composite Optimization Problems. Available on arXiv:2105.06436, 2021.

The main problem:

$$(P) \qquad \min\left\{f(z)+h(z) : z \in \mathbb{R}^n\right\}$$

where

- $h \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$ and $\mathrm{dom}\, h$ has a finite diameter $D_{\mathcal{H}}$
- there exist scalars $m, M, L \geq 0$ and a compact convex set $\Omega \supset \mathrm{dom}\, h$ such that f is nonconvex and differentiable on Ω, and for every $z, z' \in \Omega$

$$\|\nabla f(z) - \nabla f(z')\| \leq L\|z - z'\|,$$

$$-\frac{m}{2}\|z - z'\|^2 \leq f(z) - \ell_f(z; z') \leq \frac{M}{2}\|z - z'\|^2$$

where $\ell_f(z'; z) := f(z) + \langle \nabla f(z), z' - z\rangle$.

A necessary condition for $\bar{z}$ to be a local minimizer of (P) is that

$$0 \in \nabla f(\bar{z}) + \partial h(\bar{z})$$

Goal: for given $\hat{\rho} > 0$, find a $\hat{\rho}$-approximate solution of (P), i.e., a pair $(\hat{z}, \hat{v})$ such that

$$\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}), \quad \|\hat{v}\| \leq \hat{\rho}$$

There are a couple of ACG methods which accomplishes the above goal (e.g., Ghadimi and Lan's AG method [1]). This talk describes a different and novel ACG method for doing that.

ACG methods compute the next iterate as

$$z_{k+1} = z(\tilde{x}_k; M_k) := \mathrm{argmin}_z \left\{ \ell_f(z; \tilde{x}_k) + h(z) + \frac{M_k}{2}\|z - \tilde{x}_k\|^2 \right\}$$

where

$$M_k \geq \mathcal{C}(z_{k+1}; \tilde{x}_k) := \frac{2[f(z_{k+1}) - \ell_f(z_{k+1}; \tilde{x}_k)]}{\|z_{k+1} - \tilde{x}_k\|^2} \quad (*)$$

Fact: small M_k leads to fast convergence

- Constant estimate: $M_k = L$
- Adaptive estimate: $M_k \leftarrow \tau M_k$ for some $\tau > 1$, if (*) is not satisfied

  Matrix completion: function, gradient and resolvent evaluations require SVD

ACG methods compute the next iterate as

$$z_{k+1} = z(\tilde{x}_k; M_k) := \text{argmin}_z \left\{ \ell_f(z; \tilde{x}_k) + h(z) + \frac{M_k}{2} \|z - \tilde{x}_k\|^2 \right\}$$

where

$$M_k \geq \mathcal{C}(z_{k+1}; \tilde{x}_k) := \frac{2[f(z_{k+1}) - \ell_f(z_{k+1}; \tilde{x}_k)]}{\|z_{k+1} - \tilde{x}_k\|^2} \quad (*)$$

Fact: small M_k leads to fast convergence

- Constant estimate: $M_k = L$
- Adaptive estimate: $M_k \leftarrow \tau M_k$ for some $\tau > 1$, if (*) is not satisfied
  Matrix completion: function, gradient and resolvent evaluations require SVD

We will exploit the novel idea of choosing M_k as

$$M_k = \frac{\sum_{i=0}^{k-1} \mathcal{C}(z_{i+1}; \tilde{x}_i)}{k\,\alpha}$$

where $\alpha \in (0, 1)$

Note: No search for M_k is involved here!
$M_k \geq \mathcal{C}(z_{k+1}; \tilde{x}_k)$ might be violated.

Average Curvature ACG (AC-ACG) Method

0. Let $\alpha, \gamma \in (0,1)$, tolerance $\hat{\rho} > 0$ and initial point $z_0 \in \operatorname{dom} h$ be given; set $A_0 = 0$, $x_0 = z_0$, $M_0 = \gamma M$ and $k = 0$
1. compute

$$a_k = \frac{1+\sqrt{1+4M_kA_k}}{2M_k}, \quad A_{k+1} = A_k + a_k, \quad \tilde{x}_k = \frac{A_k z_k + a_k x_k}{A_{k+1}};$$

2. compute

$$x_{k+1} = \operatorname{argmin}_u \left\{ a_k \left(\ell_f(u;\tilde{x}_k) + h(u) \right) + \frac{1}{2}\|u - x_k\|^2 \right\}$$

$$z^g_{k+1} = \operatorname{argmin}_u \left\{ \ell_f(u;\tilde{x}_k) + h(u) + \frac{M_k}{2}\|u - \tilde{x}_k\|^2 \right\}$$

$$v_{k+1} = M_k(\tilde{x}_k - z^g_{k+1}) + \nabla f(z^g_{k+1}) - \nabla f(\tilde{x}_k)$$

if $\|v_{k+1}\| \le \hat{\rho}$ then output $(\hat{z}, \hat{v}) = (z^g_{k+1}, v_{k+1})$ and **stop**;

3. compute

$$C_k = \max\left\{\mathcal{C}(z^g_{k+1};\tilde{x}_k), \mathcal{L}(z^g_{k+1};\tilde{x}_k)\right\}, \quad \mathcal{L}(z;\tilde{x}) = \frac{\|\nabla f(z) - \nabla f(\tilde{x})\|}{\|z-\tilde{x}\|}$$

$$M_{k+1} = \max\left\{\gamma M, \frac{\sum_{j=0}^{k} C_j}{\alpha(k+1)}\right\},$$

4. set

$$z_{k+1} = \begin{cases} z^g_{k+1} & \text{if } C_k \le 0.9 M_k \quad \leftarrow \text{good iteration} \\ \frac{A_k z_k + a_k x_{k+1}}{A_{k+1}} & \text{otherwise} \quad \leftarrow \text{bad iteration} \end{cases}$$

and $k \leftarrow k+1$, and go to step 1

Remarks:

- Both good and bad iterations perform well-known types of acceleration steps
  Good: mimics the condition in standard ACG methods
  $M_k \geq \mathcal{C}(z_{k+1}; \tilde{x}_k)$
- If
  $$\frac{1}{\alpha} \geq 1 + \frac{1}{\gamma}$$
  then it can be shown that the proportion of good iterations is at least $2/3$
- Every iteration performs two resolvent evaluations
- Numerical results suggest removing the curvature $\mathcal{L}(z^g_{k+1}; \tilde{x}_k)$

AC-FISTA in contrast to AC-ACG

- AC-ACG computes:

$$x_{k+1} = \operatorname{argmin}_u \left\{ a_k \left(\ell_f(u;\tilde{x}_k) + h(u) \right) + \frac{1}{2}\|u - x_k\|^2 \right\}$$

$$z^g_{k+1} = z(\tilde{x}_k; M_k) = \operatorname{argmin}_u \left\{ \ell_f(u;\tilde{x}_k) + h(u) + \frac{M_k}{2}\|u - \tilde{x}_k\|^2 \right\}$$

$$C_k = \max \left\{ \mathcal{C}(z^g_{k+1};\tilde{x}_k), \mathcal{L}(z^g_{k+1};\tilde{x}_k) \right\}$$

- AC-FISTA computes:

$$z^g_{k+1} = z(\tilde{x}_k; M_k), \quad C_k = \mathcal{C}(z^g_{k+1};\tilde{x}_k),$$

Good and bad iterations

If $C_k \leq 0.9 M_k$, compute

$$x^g_{k+1} = P_\Omega\left(a_k M_k z^g_{k+1} - \frac{A_k}{a_k} z_k\right), \quad \leftarrow \text{good iteration}$$

and set $x_{k+1} = x^g_{k+1}$ and $z_{k+1} = z^g_{k+1}$; **otherwise**, compute

$$x^b_{k+1} = \operatorname*{argmin}_{u \in \mathbb{R}^n} \left\{ a_k[\ell_f(u; \tilde{x}_k) + h(u)] + \frac{1}{2}\|u - x_k\|^2 \right\}, \quad \leftarrow \text{bad iteration}$$

and set $x_{k+1} = x^b_{k+1}$ and

$$z_{k+1} = \frac{A_k z_k + a_k x^b_{k+1}}{A_{k+1}}$$

Good and bad iterations

If $C_k \le 0.9 M_k$, compute

$$x^g_{k+1} = P_\Omega\left(a_k M_k z^g_{k+1} - \frac{A_k}{a_k} z_k\right), \quad \leftarrow \text{good iteration}$$

and set $x_{k+1} = x^g_{k+1}$ and $z_{k+1} = z^g_{k+1}$; **otherwise**, compute

$$x^b_{k+1} = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ a_k[\ell_f(u; \tilde{x}_k) + h(u)] + \frac{1}{2}\|u - x_k\|^2 \right\}, \quad \leftarrow \text{bad iteration}$$

and set $x_{k+1} = x^b_{k+1}$ and

$$z_{k+1} = \frac{A_k z_k + a_k x^b_{k+1}}{A_{k+1}}$$

Remarks:

- If
$$\frac{1}{\alpha} \geq 1 + \frac{1}{\gamma}$$
then it can be shown that the proportion of good iterations is at least $2/3$
- The good iteration is performing a FISTA update
- Good iterations: one resolvent evaluation + one projection onto Ω
Bad iterations: two resolvent evaluations
- AC-FISTA uses curvature $\mathcal{C}(z^g_{k+1}; \tilde{x}_k)$ instead of $\max\left\{\mathcal{C}(z^g_{k+1}; \tilde{x}_k), \mathcal{L}(z^g_{k+1}; \tilde{x}_k)\right\}$

- Define the index sets for the good and bad iterations as

$$\mathcal{G} := \{k \geq 0 : C_k \leq 0.9M_k\}, \quad \mathcal{B} := \{k \geq 0 : C_k > 0.9M_k\}.$$

- **Condition A:** There exist $k_0 \in \mathbb{N}_+$ such that $|\mathcal{B}_k| \leq k/3$ for every $k \geq k_0$ where $\mathcal{B}_k := \{i \in \mathcal{B} : i \leq k-1\}$ for every $k \geq 1$.
- Define

$$M_k^{hm} := \frac{k}{\sum_{i=0}^{k-1} \frac{1}{M_i}}, \quad L_k^{avg} := \frac{1}{k}\sum_{i=0}^{k-1} \mathcal{L}(y_{i+1}^g; \tilde{x}_i),$$

and let

$$\theta_k := \frac{M_k}{M_k^{hm}}, \quad \tau_k := \frac{L_k^{avg}}{M_k}.$$

Theorem

Then, the following statements hold:

(a) *for every* $k \geq 1$, *we have* $v_k \in \nabla f(z_k^g) + \partial h(z_k^g)$;

(b) *if Condition A holds, then for every* $k \geq \max\{12, k_0\}$,

$$\min_{1 \leq i \leq k} \|v_i\| =$$

$$\mathcal{O}\left((1+\tau_k)\left[\frac{M_k d_0}{k^{3/2}} + \left(\sqrt{\bar{M}} + \sqrt{\bar{m}}\right)\frac{\sqrt{M_k \theta_k} D_\Omega}{k} + \frac{\sqrt{\bar{m} M_k \theta_k} D_{\mathcal{H}}}{\sqrt{k}}\right]\right)$$

where D_Ω *and* $D_{\mathcal{H}}$ *denote the diameters of* Ω *and* $\mathrm{dom}\, h$, *respectively.*

Corollary

If Condition A holds, then for every $k \geq 1$,

$$\frac{k-1}{2k} \leq \theta_k \leq \frac{M_k}{\gamma M} = \mathcal{O}\left(\frac{1}{\alpha\gamma}\right), \quad \tau_k \leq \frac{\bar{L}}{\gamma M},$$

and, as a consequence,

$$\min_{1\leq i\leq k} \|v_i\| = \mathcal{O}\left(\left(1 + \frac{\bar{L}}{\gamma M}\right)\left[\frac{Md_0}{\alpha k^{3/2}} + \left(\sqrt{\bar{M}} + \sqrt{\bar{m}}\right)\frac{\sqrt{M D_\Omega}}{\alpha\sqrt{\gamma}k} + \frac{\sqrt{\bar{m}MD_{\mathcal{H}}}}{\alpha\sqrt{\gamma}\sqrt{k}}\right]\right).$$

Dependence on α, γ is

$$\mathcal{O}\left(\frac{1}{\alpha\gamma^{3/2}}\right).$$

- Interesting case: α large (e.g., 0.5), and γ small (e.g., 10^{-6}) $(0.5, 10^{-6})$-AC-FISTA forces M_k to be small.

$$M_{k+1} = \max\left\{\gamma M\,,\, \frac{\sum_{j=0}^{k} C_j}{\alpha(k+1)}\right\}$$

- $\mathcal{O}(\alpha^{-1}\gamma^{-3/2})$ is obtained by using conservative estimates $\theta_k = \mathcal{O}(\alpha^{-1}\gamma^{-1})$ and $\tau_k = \mathcal{O}(\gamma^{-1})$
- In practice, $\theta_k = \mathcal{O}(1)$ and $\tau_k = \mathcal{O}(1)$, and the convergence rate bound reduces to

$$\min_{1\le i\le k} \|v_i\| = \mathcal{O}\left(\frac{M_k d_0}{k^{3/2}} + \left(\sqrt{\bar{M}} + \sqrt{\bar{m}}\right)\frac{\sqrt{M_k}D_\Omega}{k} + \frac{\sqrt{\bar{m}M_k}D_{\mathcal{H}}}{\sqrt{k}}\right)$$

Computational Results

Restart variant of AC-FISTA: whenever $k \in \mathcal{G}$ and $\phi(z_{k+1}) \geq \phi(z_k)$, rejects z_{k+1}, sets $x_k = z_k$ and $A_k = 0$, and repeats the k-th iteration. AC-FISTA (AF) and its restart variant AF(R) described above were benchmarked against

- UPFAG (UP) [2] by Ghadimi, Lan and Zhang (backtracking)
- ADAP-NC-FISTA (AD) [4] by L., Monteiro and Sim (backtracking)
    - and its restart variant, AD(R)
- AC-ACG (AC) [3] by L. and Monteiro (average curvature)
    - and its restart variant, AC(R)

on three problems.

All methods stop with a pair (z, v) satisfying

$$v \in \nabla f(z) + \partial h(z), \qquad \frac{\|v\|}{\|\nabla f(z_0)\| + 1} \leq \hat{\rho}$$

1st Problem (Constrained matrix completion)

$$\min \left\{ \frac{1}{2} \|\Pi_{\mathcal{Q}}(Z - O)\|_F^2 + \mu \sum_{i=1}^{r} p(\sigma_i(Z)) : Z \in \mathcal{B}_R \right\}$$

where $O \in \mathbb{R}^{\Omega}$ is an incomplete observed matrix, $r := \min\{l, n\}$, $\sigma_i(Z)$ is the i-th singular value of Z, $\mathcal{B}_R = \{Z \in \mathbb{R}^{l \times n} : \|Z\|_F \leq R\}$ and

$$p(t) = p_{\beta,\theta}(t) := \beta \log \left(1 + \frac{|t|}{\theta} \right).$$

Table: Matrix completion datasets

Dataset	Users (l)	Items (n)	Ratings	Density	Scale
MovieLens 100K	943	1682	100000	6.30%	[1,5]
FilmTrust	1508	2071	35497	1.14%	[0.5,4.0]

Table: Solving MC with *MovieLens 100K*

m	Function Value / Iteration Count / Running Time (s)						
	UP	AD	AC	AF	AD(R)	AC(R)	AF(R)
4.4	2605	2625	2288	**1836**	2625	2240	1912
	521	1674	765	375	1674	718	305
	1545	1946	923	287	1946	851	**245**
8.9	4261	4203	3884	**3617**	4203	3914	3797
	576	1794	968	291	1794	896	241
	1621	1930	1173	233	1930	1057	**208**
20	4637	4582	4267	**4098**	4582	4358	4164
	676	2209	1079	260	2209	1028	304
	1914	2364	1236	**212**	2364	1210	267
30	6753	6293	5975	**5333**	6293	5958	5524
	606	1963	1085	505	1963	1205	413
	1628	2104	1263	417	2104	1687	**349**

Table: Solving MC with *FilmTrust*

m	Function Value / Iteration Count / Running Time (s)						
	UP	AD	AC	AF	AD(R)	AC(R)	AF(R)
4.4	1050 584 6460	1069 2025 9063	959 836 3559	849 347 **991**	1069 2025 9063	981 796 3321	**804** 586 1753
8.9	1814 634 7130	1854 2410 11171	1769 1050 4617	1538 469 **1334**	1854 2410 11171	1701 1198 4939	**1516** 753 2198
20	2120 630 7214	2064 2665 12701	2016 1015 4656	**1739** 676 1959	2064 2665 12701	2008 1100 4582	1777 528 **1617**
30	2980 559 6244	2917 2365 11205	2845 1086 4824	**2593** 533 **1582**	2917 2365 11205	2845 1086 4518	**2593** 533 **1582**

Table: Statistics of $\bar{\theta}_k$, $\bar{\tau}_k$ and $|\mathcal{B}_k|$ for *MovieLens 100K*

m	AF			AF(R)		
	$\bar{\theta}_k$	$\bar{\tau}_k$	$\lvert\mathcal{B}_k\rvert/k$	$\bar{\theta}_k$	$\bar{\tau}_k$	$\lvert\mathcal{B}_k\rvert/k$
4.4	1.07	1.23	6%	1.12	1.20	4%
8.9	1.04	1.53	8%	1.02	1.48	10%
20	0.97	2.16	9%	1.00	1.88	13%
30	1.02	2.49	7%	1.02	2.40	11%

Table: Statistics of $\bar{\theta}_k$, $\bar{\tau}_k$ and $|\mathcal{B}_k|$ for *FilmTrust 100K*

m	AF			AF(R)		
	$\bar{\theta}_k$	$\bar{\tau}_k$	$\lvert\mathcal{B}_k\rvert/k$	$\bar{\theta}_k$	$\bar{\tau}_k$	$\lvert\mathcal{B}_k\rvert/k$
4.4	1.09	1.25	10%	1.11	1.21	9%
8.9	1.02	1.55	6%	0.99	1.61	6%
20	1.04	2.07	8%	1.06	2.07	9%
30	1.04	2.59	11%	1.04	2.59	11%

2nd Problem (Nonconvex QP)

$$\min \left\{ f(Z) := -\frac{\xi}{2}\|D\mathcal{B}(Z)\|^2 + \frac{\tau}{2}\|\mathcal{A}(Z) - b\|^2 : z \in P_n \right\}$$

where P_n is the unit spectraplex, i.e.,

$$P_n := \left\{ Z \in S^n_+ : \mathrm{tr}(Z) = 1 \right\},$$

$\mathcal{A} : S^n_+ \to \mathbb{R}^\ell$ and $\mathcal{B} : S^n_+ \to \mathbb{R}^p$ are linear operators, $D \in \mathbb{R}^{p\times p}$ is a positive diagonal matrix, and $b \in \mathbb{R}^\ell$ is a vector.

Table: Quadratic programming datasets

Dataset	l	n	Density d
QP-1	50	200	2.5%
QP-2	50	400	0.5%

Table: Solving QP with *QP-1*

m	Iteration Count / Running Time (s)						
	UP	AD	AC	AF	AD(R)	AC(R)	AF(R)
10^6	9	12	10	18	23	10	15
	0.7	0.7	**0.5**	0.6	0.7	**0.5**	**0.5**
10^5	2633	2206	1054	947	787	1054	419
	261	89	43	30	33	43	**14**
10^4	7203	2591	1678	1744	1573	1678	601
	705	104	68	55	66	68	**20**
10^3	5429	2637	1464	2000	1552	1464	773
	540	109	60	63	65	60	**26**
10^2	6891	2639	1420	1687	1666	1420	736
	653	116	58	52	69	58	**25**
10	6479	2640	1424	1804	1675	1424	785
	613	116	58	56	69	58	**26**

Table: Solving QP with *QP-2*

m	Iteration Count / Running Time (s)						
	UP	AD	AC	AF	AD(R)	AC(R)	AF(R)
10^6	10	12	10	17	23	10	14
	1.9	1.8	**1.2**	1.5	1.8	**1.2**	1.3
10^5	56	530	142	140	292	142	140
	13	56	16	**12**	30	16	**12**
10^4	105	868	320	195	364	320	182
	26	93	36	**17**	38	36	**17**
10^3	115	900	271	187	384	271	164
	29	103	29	16	40	29	**15**
10^2	119	904	300	216	385	300	179
	32	103	33	19	40	33	**16**
10	113	904	274	221	385	274	177
	31	104	30	19	40	30	**16**

Table: Statistics of $\bar{\theta}_k$, $\bar{\tau}_k$ and $|\mathcal{B}_k|$ for *QP-1* and *QP-2*

m	AF			AF(R)		
	$\bar{\theta}_k$	$\bar{\tau}_k$	$\lvert\mathcal{B}_k\rvert/k$	$\bar{\theta}_k$	$\bar{\tau}_k$	$\lvert\mathcal{B}_k\rvert/k$
10^5	0.92	1.22	13%	1.04	1.24	15%
10^4	1.07	1.05	7%	1.07	1.05	13%
10^3	0.99	1.14	5%	0.99	1.14	13%
10^2	1.02	1.07	5%	1.02	1.07	18%
10	1.00	1.10	5%	1.00	1.10	10%

m	AF			AF(R)		
	$\bar{\theta}_k$	$\bar{\tau}_k$	$\lvert\mathcal{B}_k\rvert/k$	$\bar{\theta}_k$	$\bar{\tau}_k$	$\lvert\mathcal{B}_k\rvert/k$
10^5	0.60	3.08	13%	0.60	3.08	13%
10^4	0.68	2.29	18%	0.72	2.16	15%
10^3	0.69	2.38	16%	0.74	2.14	15%
10^2	0.69	2.40	14%	0.73	2.17	15%
10	0.69	2.40	14%	0.73	2.17	15%

3rd Problem (SVM)

$$\min_{z\in\mathbb{R}^n}\left\{ f(z) := \frac{1}{p}\sum_{i=1}^{p} \ell(x_i, y_i; z) + \frac{\lambda}{2}\|z\|^2 : z \in B_r \right\},$$

for some $\lambda, r > 0$, where $x_i \in \mathbb{R}^n$ is a feature vector, $y_i \in \{1, -1\}$ denotes the corresponding label, $\ell(x_i, y_i; \cdot) = 1 - \tanh(y_i\langle\cdot, x_i\rangle)$ is a nonconvex sigmoid loss function and $B_r := \{z \in \mathbb{R}^n : \|z\| \le r\}$.

Table: SVM datasets

Dataset	n	p	λ	M
SVM-1	1000	500	0.002	13
SVM-2	2000	1000	0.001	25
SVM-3	3000	1000	0.001	38
SVM-4	4000	500	0.002	50

Table: Solving SVM with *SVM-1, 2, 3, & 4*

Dataset	Iteration Count / Running Time (s)						
	UP	AD	AC	AF	AD(R)	AC(R)	AF(R)
SVM-1	130	12274	546	545	12274	342	342
	8	188	6	6	188	5	**4**
SVM-2	278	21127	1131	1130	21127	392	366
	39	1836	81	67	1836	30	**26**
SVM-3	401	71991	1035	1034	71991	290	291
	97	8957	109	92	8957	34	**31**
SVM-4	247	12450	665	664	12450	210	210
	44	1033	47	37	1033	11	**10**

Table: Statistics of $\bar{\theta}_k$, $\bar{\tau}_k$ and $|\mathcal{B}_k|$

Dataset	AF			AF(R)		
	$\bar{\theta}_k$	$\bar{\tau}_k$	$\lvert\mathcal{B}_k\rvert/k$	$\bar{\theta}_k$	$\bar{\tau}_k$	$\lvert\mathcal{B}_k\rvert/k$
SVM-1	2.71	0.57	32%	2.32	0.56	42%
SVM-2	8.51	0.60	35%	1.93	0.54	40%
SVM-3	1.86	0.59	37%	1.75	0.53	41%
SVM-4	2.32	0.55	32%	1.83	0.56	41%

Features of AC-FISTA

- AC-FISTA is a FISTA-type ACG variant of the AC-ACG method proposed in [3], which is an ACG method based on the average of the previously observed curvatures.
- AC-FISTA does not require any line search for M_k.
- Using $\mathcal{C}(z^g_{k+1}; \tilde{x}_k)$ instead of $\max\left\{\mathcal{C}(z^g_{k+1}; \tilde{x}_k), \mathcal{L}(z^g_{k+1}; \tilde{x}_k)\right\}$
- Good iterations: one resolvent evaluation, bad iterations: two resolvent evaluations
  In practice, one resolvent evaluation per iteration on average

Results

- A practical AC-FISTA variant substantially outperforms previous ACG variants as well as the theoretical and practical AC-ACG variants
- Establish a convergence rate bound in terms of the average observed curvatures (novel result)
- Convergence rate analysis of the restart variant

THE END

Thanks!

S. Ghadimi and G. Lan.
Accelerated gradient methods for nonconvex nonlinear and stochastic programming.
Math. Programming, 156:59–99, 2016.

S. Ghadimi, G. Lan, and H. Zhang.
Generalized uniformly optimal methods for nonlinear programming.
Journal of Scientific Computing, 79(3):1854–1881, 2019.

J. Liang and R. D. C. Monteiro.
An average curvature accelerated composite gradient method for nonconvex smooth composite optimization problems.
SIAM Journal on Optimization, 31(1):217–243, 2021.

J. Liang, R. D. C. Monteiro, and C.-K. Sim.
A FISTA-type accelerated gradient algorithm for solving smooth nonconvex composite optimization problems.

Computational Optimization and Applications, 79(3):649–679, 2021.