High-order Accumulative Regularization Methods for Gradient Minimization

Yao Ji

Joint work with Guanghui (George) Lan

H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology

2025 Informs Annual Meeting

Problem.

$$\min_{x \in X} f(x) + h(x)$$

f is convex, p-times differentiable, with ν-Hölderian continuous p-th derivative.

$$\|D^p f(x) - D^p f(y)\| \leq L_p(\nu)\|x - y\|^\nu, \quad \nu \in [0\ 1],$$

where $\|D^p f(x) - D^p f(y)\| = \max_h \{|D^p f(x)[h]^p - D^p f(y)[h]^p| : \|h\| \leq 1\}$.

Problem.

$$\min_{x \in X} f(x) + h(x)$$

f is convex, p-times differentiable, with ν-Hölderian continuous p-th derivative.

$$\|D^p f(x) - D^p f(y)\| \leq L_p(\nu)\|x - y\|^\nu, \quad \nu \in [0\ 1],$$

where $\|D^p f(x) - D^p f(y)\| = \max_h \{|D^p f(x)[h]^p - D^p f(y)[h]^p| : \|h\| \leq 1\}$.

Examples: $p = 1, \nu = 1$, L-smoothness

$p = 2, \nu = 0$, bounded Hessian

$p \geq 2, \nu = 1$, p-th Lipschitz continuous derivative

Problem.

$$\min_{x \in X} f(x) + h(x)$$

f is convex, p-times differentiable, with ν-Hölderian continuous p-th derivative.

$$\|D^p f(x) - D^p f(y)\| \leq L_p(\nu)\|x - y\|^\nu, \quad \nu \in [0\ 1],$$

where $\|D^p f(x) - D^p f(y)\| = \max_h \{|D^p f(x)[h]^p - D^p f(y)[h]^p| : \|h\| \leq 1\}$.

h is convex, maybe nonsmooth; X is convex, for simplicity, assume $h = 0, X = \mathbb{R}^n$.

Problem.

$$\min_{x \in X} f(x) + h(x)$$

f is convex, p-times differentiable, with ν-Hölderian continuous p-th derivative.

$$\|D^p f(x) - D^p f(y)\| \leq L_p(\nu)\|x - y\|^\nu, \quad \nu \in [0\ 1],$$

where $\|D^p f(x) - D^p f(y)\| = \max_h \{|D^p f(x)[h]^p - D^p f(y)[h]^p| : \|h\| \leq 1\}$.
h is convex, maybe nonsmooth; X is convex, for simplicity, assume $h = 0, X = \mathbb{R}^n$.

Question.

How to make the gradient small?

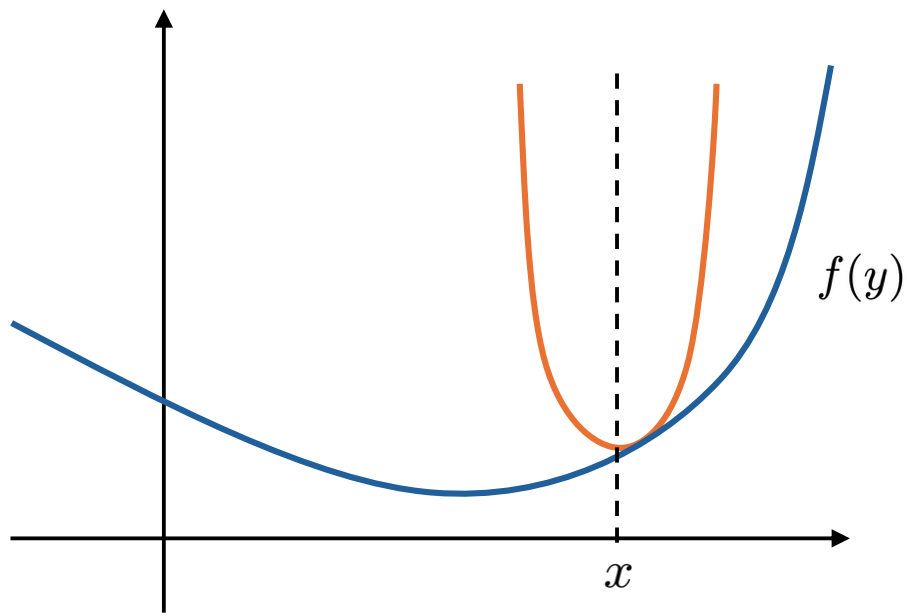
$$\|\nabla f(x)\| \leq \varepsilon$$

Smoothness of f: upper curvature

ν-Hölderian continuous p-th derivative

$$\|D^p f(x) - D^p f(y)\| \leq L_p(\nu)\|x - y\|^\nu, \quad \nu \in [0\ 1].$$

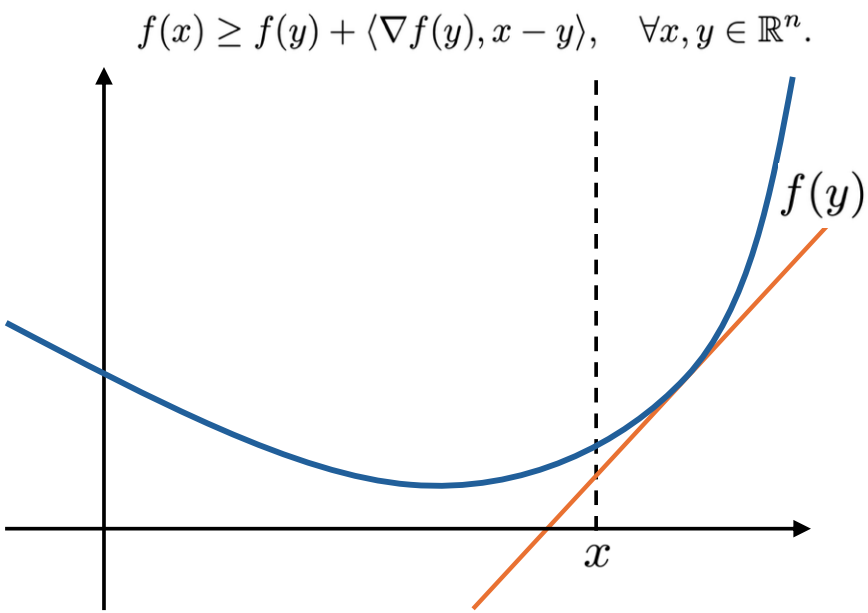
$$\Longrightarrow f(y) \leq T_p(x, y - x) + \frac{L}{p!}\|x - y\|^{p+\nu}, \quad L \geq L_p(\nu.)$$



Regularity of f: lower curvature

Convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall\, x, y \in \mathbb{R}^n.$$



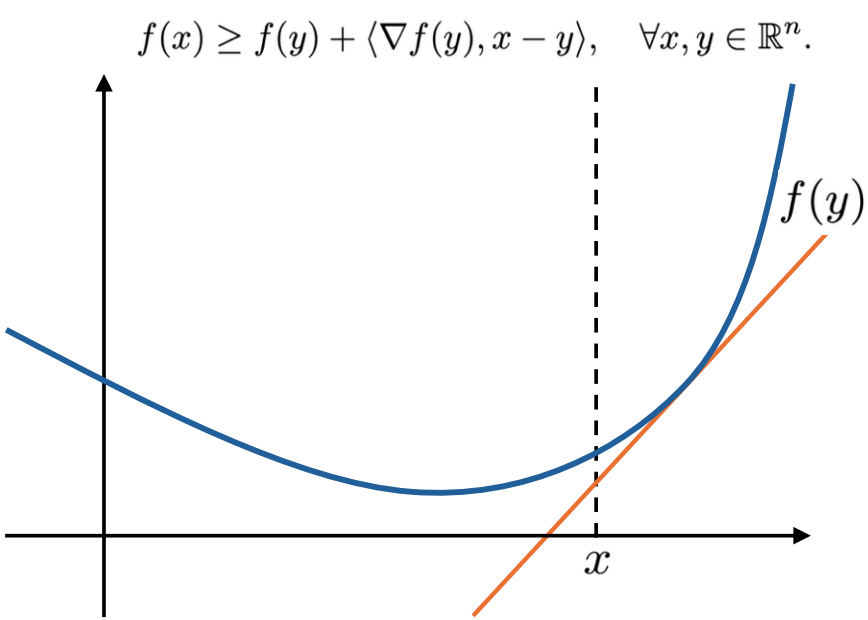
Regularity of f: lower curvature

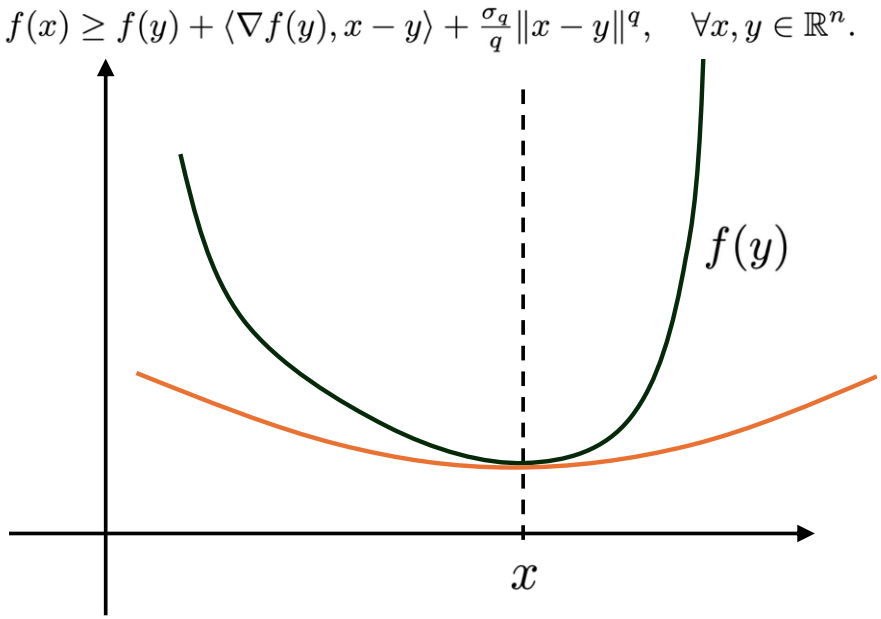
Convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall\, x, y \in \mathbb{R}^n.$$

Uniform convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\sigma_q}{q} \|x - y\|^q, \quad \forall\, x, y \in \mathbb{R}^n, q \geq 2.$$





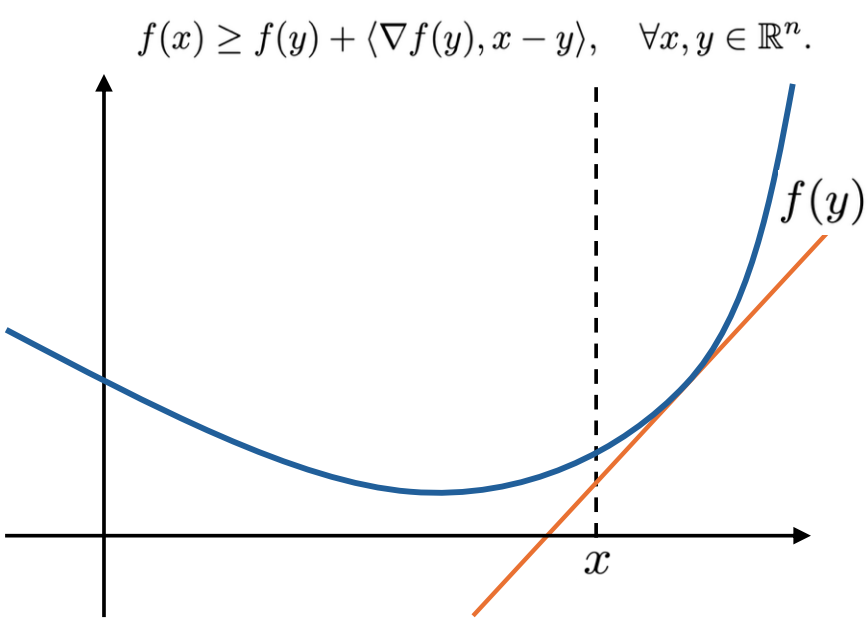
Regularity of f: lower curvature

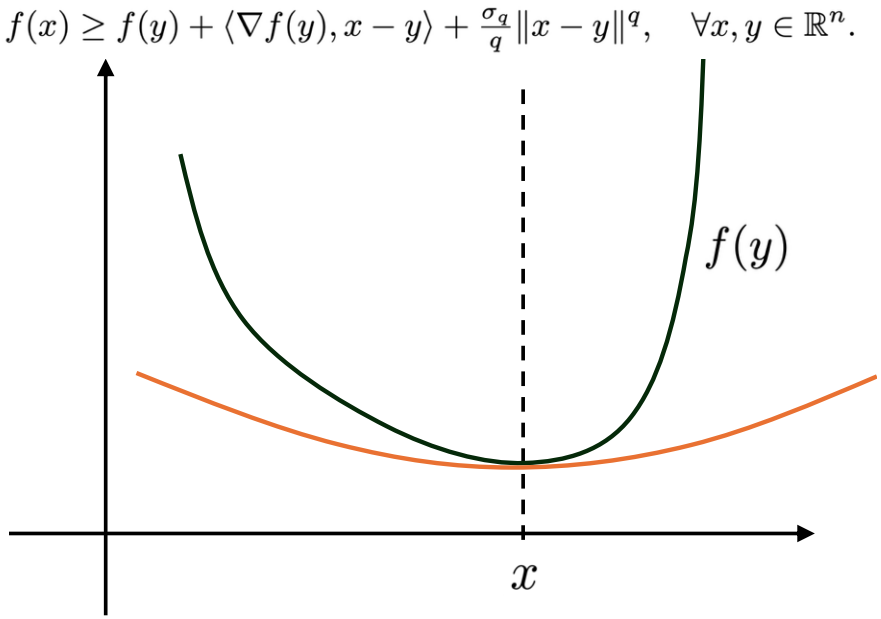
Convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall\, x, y \in \mathbb{R}^n. \quad \text{This talk.}$$

Uniform convexity

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \tfrac{\sigma_q}{q} \|x - y\|^q, \quad \forall\, x, y \in \mathbb{R}^n, q \geq 2.$$





What is the status of p-th order methods?

Different Approaches.

1. p-th order tensor methods and its accelerated version: Nesterov 06' 08', Baes 09', Nesterov 19' ...

$$f(\widehat{x}) - f^* \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(L_p/\varepsilon^{\frac{1}{p+1}}\right) \quad \text{iterations.}$$

What is the status of p-th order methods?

Different Approaches.

1. p-th order tensor methods and its accelerated version: Nesterov 06' 08', Baes 09', Nesterov 19' ...

$$f(\widehat{x}) - f^* \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(L_p/\varepsilon^{\frac{1}{p+1}}\right) \quad \text{iterations.}$$

2. Accelerated Hybrid Proximal Extragradient: Monteiro & Svaiter 13', Jiang et al. 19', Bubeck et al. 19', Gasnikov et al. 19...

$$f(\widehat{x}) - f^* \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(L_p \log \tfrac{1}{\varepsilon}/\varepsilon^{\frac{2}{3p+1}}\right) \quad \text{iterations.}$$

What is the status of p-th order methods?

Different Approaches.

1. p-th order tensor methods and its accelerated version: Nesterov 06' 08', Baes 09', Nesterov 19' ...

$$f(\widehat{x}) - f^* \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(L_p/\varepsilon^{\frac{1}{p+1}}\right) \quad \text{iterations.}$$

2. Accelerated Hybrid Proximal Extragradient: Monteiro & Svaiter 13', Jiang et al. 19', Bubeck et al. 19', Gasnikov et al. 19...

$$f(\widehat{x}) - f^* \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(L_p \log \tfrac{1}{\varepsilon}/\varepsilon^{\frac{2}{3p+1}}\right) \quad \text{iterations.}$$

3. Adaptivity and inexactness: Cartis, Gould & Toint 11', Jiang, lin, & Zhang 20', Gragpilia & Nesterov 19' 20'...

$$f(\widehat{x}) - f^* \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(\max\{L_0, pL_p, \delta\}/\varepsilon^{\frac{1}{p+1}}\right) \quad \text{iterations.}$$

What is the status to drive gradients small?

Different Approach.

1. p-th order tensor methods and its accelerated version: Nesterov 06' 08' 19', ...

$$\|\nabla f(\widehat{x})\| \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(L_p \log \tfrac{1}{\varepsilon} / \varepsilon^{\frac{1}{p+1}}\right) \quad \text{iterations.}$$

2. Accelerated Hybrid Proximal Extragradient: Monteiro & Svaiter 13', ...

$$\|\nabla f(\widehat{x})\| \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(L_p \log \tfrac{1}{\varepsilon} / \varepsilon^{\frac{2}{3p}}\right) \quad \text{iterations.}$$

3. Adaptivity and inexactness: Gragpilia & Nesterov 19' 20' 23'

$$\|\nabla f(\widehat{x})\| \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(\max\{L_0, pL_p, \delta\} / \varepsilon^{\frac{p+1}{p(p+2)}}\right) \quad \text{iterations.}$$

Function residual and gradient norm gap still cannot match — they remain $\log \frac{1}{\varepsilon}, \varepsilon^{\frac{2}{3p+1}-\frac{2}{3p}}, \varepsilon^{\frac{1}{p+1}-\frac{p+1}{p(p+2)}}$ off.

A formal question

Question

Is there a unified way to translate function-residual rates into fast (matching) gradient-norm rates?

A formal question

Question

Is there a unified way to translate function-residual rates into fast (matching) gradient-norm rates?

Answer

Yes, for most algorithm of interest with fast function residual and slow gradient norm.

Orders up! Accumulative Regularization for Gradient minimization

Algorithm AR framework for gradient minimization

Initialize S, strictly increasing $\{\sigma_s\}_{s=0}^{S}$ with $\sigma_0 = 0$ and initial point $x_0 \in \mathbb{R}^n$.

1: **for** $s = 1, \ldots, S$ **do**

2: Compute an approximate solution x_s of the proximal subproblem

$$x_s \approx \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ f_s(x) := f(x) + \sum_{i=1}^{s} \frac{\sigma_i - \sigma_{i-1}}{p+\nu} \|x - x_{i-1}\|^{p+\nu} \right\},$$

where $p + \nu \geq 2$, by running some subroutine $\mathcal{A}$ with the initialization x_{s-1} for N_s iterations.

3: **output** x_S

Orders up! Accumulative Regularization for Gradient minimization

Algorithm AR framework for gradient minimization

Initialize S, strictly increasing $\{\sigma_s\}_{s=0}^{S}$ with $\sigma_0 = 0$ and initial point $x_0 \in \mathbb{R}^n$.

1: **for** $s = 1, \ldots, S$ **do**

2: Compute an approximate solution x_s of the proximal subproblem

$$x_s \approx \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ f_s(x) := f(x) + \sum_{i=1}^{s} \frac{\sigma_i - \sigma_{i-1}}{p+\nu} \|x - x_{i-1}\|^{p+\nu} \right\},$$

where $p + \nu \geq 2$, by running some subroutine $\mathcal{A}$ with the initialization x_{s-1} for N_s iterations.

3: **output** x_S

Inspired by Lan, Ouyang & Zhang 23' for optimal gradient minimization using *first order* method.

Accumulative Regularization (AR) uses two ingredients:

1. Regularization Approach.
   Nesterov 12'...
2. Accelerated proximal point method.
   Güler 92', Nesterov 23'...

Recap: A formal question

Question

Is there a unified way to translate function-residual rates into fast (matching) gradient-norm rates?

A formal answer

Yes, using **AR** with algorithm of interest $\mathcal{A}$ as subroutine, where $\mathcal{A}$ has fast function residual decrease and slow gradient norm decrease.

Recap: A formal question

Question

Is there a unified way to translate function-residual rates into fast (matching) gradient-norm rates?

A formal answer

Yes, using **AR** with algorithm of interest $\mathcal{A}$ as subroutine, where $\mathcal{A}$ has fast function residual decrease and slow gradient norm decrease.

Showcase:
Using AR with subroutine $\mathcal{A}$: p-th order accelerated tensor methods Gragpilia & Nesterov 19' 20'.
Goal: to show

$$\|\nabla f(\widehat{x})\| \leq \varepsilon \quad \text{in} \quad \underbrace{\mathcal{O}\left(1/\varepsilon^{\frac{1}{p+\nu}}\right)}_{\text{match function residual}} \quad \text{iterations.}$$

Subroutine $\mathcal{A}$ Assumption

Assumption (Sublinear convergence)

After N_s iterations of the subroutine $\mathcal{A}(f, \{\sigma_i\}_{i\le s}, \{x_{i-1}\}_{i\le s})$, there holds

$$f_s(x_s) - f_s(x_s^*) \le \frac{C_{\mathcal{A}} L_{p,\nu} \|x_s^* - x_{s-1}\|^{p+\nu}}{N_s^{p+\nu}}, \quad \forall\ 1 \le s \le S$$

fast part

Subroutine $\mathcal{A}$ Assumption

Assumption (Sublinear convergence)

After N_s iterations of the subroutine $\mathcal{A}(f, \{\sigma_i\}_{i\le s}, \{x_{i-1}\}_{i\le s})$, there holds

$$f_s(x_s) - f_s(x_s^*) \le \frac{C_{\mathcal{A}} L_{p,\nu} \|x_s^* - x_{s-1}\|^{p+\nu}}{N_s^{p+\nu}}, \quad \forall\ 1 \le s \le S$$

fast part

Subroutine $\mathcal{A}$ Assumption

Assumption (Sublinear convergence)

After N_s iterations of the subroutine $\mathcal{A}(f, \{\sigma_i\}_{i\le s}, \{x_{i-1}\}_{i\le s})$, there holds

$$f_s(x_s) - f_s(x_s^*) \le \frac{C_{\mathcal{A}} L_{p,\nu} \|x_s^* - x_{s-1}\|^{p+\nu}}{N_s^{p+\nu}}, \quad \forall\ 1 \le s \le S \qquad \textit{fast part}$$

For the S-th epoch, there exists k, $N_S \le k \le 2N_S$, such that

$$\min_{k=N_S+1,\dots,2N_S} \|\nabla f_S(x_S^k)\| \le \frac{C_{\mathcal{A}} L_{p,\nu} \|x_S^* - x_{S-1}\|^{p+\nu-1}}{N_s^{p+\nu-1}}, \qquad \textit{slow part}$$

where $C_{\mathcal{A}} > 1$ is a universal constant.

Convergence for accumulative regularization

Algorithm AR framework for gradient minimization

Initialize S, strictly increasing $\{\sigma_s\}_{s=0}^{S}$ with $\sigma_0 = 0$ and initial point $x_0 \in \mathbb{R}^n$.

1: **for** $s = 1, \ldots, S$ **do**

2: Compute an approximate solution x_s of the proximal subproblem

$$x_s \approx \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ f_s(x) := f(x) + \sum_{i=1}^{s} \frac{\sigma_i - \sigma_{i-1}}{p+\nu} \|x - x_{i-1}\|^{p+\nu} \right\},$$

where $p + \nu \geq 2$, by running some subroutine $\mathcal{A}$ with the initialization x_{s-1} for N_s iterations.

3: **output** x_S

Thm: Ji-Lan 25' Setting $S \approx \log_{c_p} \frac{L_{p,\nu} D^{p+\nu-1}}{\varepsilon}$, $\sigma_s \approx \frac{c_p^{s-1} \varepsilon}{D^{p+\nu-1}}$, $N_s \approx \left(\frac{c_p L_{p,\nu}}{\sigma_s}\right)^{\frac{1}{p+\nu}}$ guarantees

$$\|\nabla f(\widehat{x})\| \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(\frac{c_p}{c_p - 1} \frac{L_{p,\nu}^{\frac{1}{p+\nu}} D^{\frac{p+\nu-1}{p+\nu}}}{\varepsilon^{\frac{1}{p+\nu}}} \right) \quad \text{iterations,}$$

where $D \geq \min_{x^* \in X^*} \|x_0 - x^*\|$, and universal constant $c_p > 1$.

Consequence: match function residual!

Convergence for accumulative regularization

Algorithm AR framework for gradient minimization

Initialize S, strictly increasing $\{\sigma_s\}_{s=0}^{S}$ with $\sigma_0 = 0$ and initial point $x_0 \in \mathbb{R}^n$.

1: **for** $s = 1, \ldots, S$ **do**

2: Compute an approximate solution x_s of the proximal subproblem

$$x_s \approx \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ f_s(x) := f(x) + \sum_{i=1}^{s} \frac{\sigma_i - \sigma_{i-1}}{p+\nu} \|x - x_{i-1}\|^{p+\nu} \right\},$$

where $p + \nu \geq 2$, by running some subroutine $\mathcal{A}$ with the initialization x_{s-1} for N_s iterations.

3: **output** x_S

Thm: Ji-Lan 25' Setting $S \approx \log_{c_p} \frac{L_{p,\nu} D^{p+\nu-1}}{\varepsilon}$, $\sigma_s \approx \frac{c_p^{s-1} \varepsilon}{D^{p+\nu-1}}$, $N_s \approx \left(\frac{c_p L_{p,\nu}}{\sigma_s}\right)^{\frac{1}{p+\nu}}$ guarantees

$$\|\nabla f(\widehat{x})\| \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(\frac{c_p}{c_p - 1} \frac{L_{p,\nu}^{\frac{1}{p+\nu}} D^{\frac{p+\nu-1}{p+\nu}}}{\varepsilon^{\frac{1}{p+\nu}}} \right) \quad \text{iterations,}$$

where $D \geq \min_{x^* \in X^*} \|x_0 - x^*\|$, and universal constant $c_p > 1$.

Parameter-free subroutine $\mathcal{A}_s$

Assumption

For any $k > 1$, there holds

$$f_s(x_s^k) - f_s(x_s^*) \leq \frac{L_{s,k}\|x_{s-1} - x_s^*\|^{p+\nu}}{(k-1)^{p+\nu}}, \quad \forall\ 1 \leq s \leq S, \quad \textit{fast part}$$

where $L_{s,k}$ is an local estimate of Hölder constant of f at (s,k) such that

$$L_{s,k} \leq c_{\mathcal{A}} \max\{pL_{p,\nu}, L_0, \theta\},$$

where θ is a user-defined subproblem inexactness parameter. For each s-th epoch, there exists $\nabla f_s(z_s^k)$, $N_s \leq k \leq 2N_s$, such that

$$\min_{k=N_s+1,\ldots,2N_s} \|\nabla f_s(z_s^k)\| \leq \frac{L_{s,2N_s}\|x_{s-1} - x_s^*\|^{p+\nu-1}}{N_s^{p+\nu-1}}, \quad \textit{slow part}$$

where $L_{s,2N_s}$ is the local estimate of Hölder constant at $(s,2N_s)$ such that

$$L_{s,2N_s} \leq c_{\mathcal{A}} \max\{pL_{p,\nu}, L_0, \theta\}.$$

Satisfied e.g. by Gragpilia & Nesterov 20'

Parameter-free accumulative regularization

Algorithm Accumulative regularization without the knowledge of $L_{p,\nu}$

Initialize $\sigma_0 = 0, \sigma_1 > 0$, and x_0, L_0.

1: **for** $s = 1, 2 \dots$ **do**

2: If $s > 1$, set $\sigma_s = c_p \sigma_{s-1}$, $c_p > 1$

3: Compute an approximate solution x_s of the proximal subproblem

$$x_s \approx \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ f_s(x) := f(x) + \sum_{i=1}^{s} \frac{\sigma_i - \sigma_{i-1}}{p+\nu} \|x - x_{i-1}\|^{p+\nu} \right\}, \ p + \nu \geq 2$$

by running $\mathcal{A}_s$ with initialization (x_{s-1}, L_{s-1}).
$\mathcal{A}_s$ will output the k-th iterate as x_s and its line search value $L_{s,k}$ when

$$k \geq 8 \left[\frac{L_{s,k}(p+\nu)}{4\sigma_s} \right]^{\frac{1}{p+\nu}} + 1.$$

Denote $N_s := k$. Continue the same number of surplus iterations to record $L_{s,2N_s}$.

4: If $\sigma_s \geq \frac{(L_{s,2N_s})^{p+\nu}}{(L_{s,N_s})^{p+\nu-1}}$, then **terminate** with $\widehat{x} = \operatorname*{argmin}_{N_s < k \leq 2N_s} \{\|\nabla f(z_s^k)\|\}$.

5: **output** $\widehat{x}$

Remark: Without additional line search, just the line search from the subroutines.
Unknown D can be handled by doubling trick (omitted).

Convergence for parameter-free accumulative regularization

Thm: Ji-Lan 25' Parameter free AR guarantees

$$\|\nabla f(\widehat{x})\| \leq \varepsilon \quad \text{in} \quad \mathcal{O}\left(\frac{\max\{pL_{p,\nu}, L_0, \theta\}^{\frac{1}{p+\nu}} \operatorname{dist}(x_0, X^*)^{\frac{p+\nu-1}{p+\nu}}}{\varepsilon^{\frac{1}{p+\nu}}}\right) \qquad \text{iterations,}$$

where $\operatorname{dist}(x_0, X^*) = \min_{x^* \in X^*} \|x_0 - x^*\|$, and θ is the inexactness level

Consequence:

Inexact and parameter free!

Match the function residual Gragpilia & Nesterov 19'.

Improve over: $\mathcal{O}\left(\frac{\max\{pL_{p,\nu}, L_0, \theta\}^{\frac{1}{p+\nu}} \operatorname{dist}(x_0, X^*)^{\frac{p+\nu-1}{p+\nu}}}{\varepsilon^{\frac{p+\nu}{(p+\nu-1)(p+\nu+1)}}}\right)$ Gragpilia & Nesterov 20'.

Summary

$\mathcal{A}$ → AR → Gradient norm matching function residual