Parameter-free and optimal restart schemes for first-order methods via approximate sharpness

Maksym Neyra-Nesterenko

Department of Mathematics
Simon Fraser University
mneyrane@sfu.ca

Sep 21, 2024

Collaborators

Ben Adcock
(Simon Fraser U.)

Matthew Colbrook
(U. of Cambridge)

Paper: *Restarts subject to approximate sharpness: A parameter-free and optimal scheme for first-order methods*. Foundations of Computational Mathematics (in press, 2024).

Key contribution

We use **approximate sharpness** to design a **meta-algorithm** that **accelerates** the convergence of **any** first-order optimization method.

Remarks:

1. Our approach, based on restarts, can be used with essentially any first-order method
2. It applies to broad classes of convex problems, e.g. ℓ^1-minimization
3. We guarantee fast decay of the objective function error down to an underlying error level

Outline

Outline

General setup

Problem: Let $f : \mathbb{R}^N \to \mathbb{R}$ be proper, closed and convex, and $Q \subseteq \mathbb{R}^N$ be a closed, convex set. Consider the problem

$$\min_{\boldsymbol{x} \in Q} f(\boldsymbol{x}) \qquad (\star)$$

and let $\hat{\boldsymbol{X}}$ be its set of minimizers with function value $\hat{f}$.

Approximate sharpness: We assume that $(\star)$ satisfies

$$\operatorname{dist}(\boldsymbol{x}, \hat{\boldsymbol{X}}) \leq \left(\frac{f(\boldsymbol{x}) - \hat{f} + g_Q(\boldsymbol{x}) + \eta}{\alpha} \right)^{1/\beta}, \quad \forall \boldsymbol{x} \in \mathbb{R}^N,$$

where $\alpha > 0$, $\eta \geq 0$ and $\beta \geq 1$, $\operatorname{dist}(\boldsymbol{x}, \hat{\boldsymbol{X}}) = \inf_{\boldsymbol{z} \in \hat{\boldsymbol{X}}} d(\boldsymbol{x}, \boldsymbol{z})$ for some metric d on $\mathbb{R}^N$ and g_Q is a known function satisfying if $\operatorname{dist}(\boldsymbol{x}_i, Q) \to 0$, then $g_Q(\boldsymbol{x}_0) \to 0$.

General setup

Problem: Let $f : \mathbb{R}^N \to \mathbb{R}$ be proper, closed and convex, and $Q \subseteq \mathbb{R}^N$ be a closed, convex set. Consider the problem

$$\min_{\boldsymbol{x} \in Q} f(\boldsymbol{x}) \tag{$\star$}$$

and let $\hat{\boldsymbol{X}}$ be its set of minimizers with function value $\hat{f}$.

Approximate sharpness: We assume that $(\star)$ satisfies

$$\mathrm{dist}(\boldsymbol{x}, \hat{\boldsymbol{X}}) \leq \left(\frac{f(\boldsymbol{x}) - \hat{f} + g_Q(\boldsymbol{x}) + \eta}{\alpha} \right)^{1/\beta}, \quad \forall \boldsymbol{x} \in \mathbb{R}^N,$$

where $\alpha > 0$, $\eta \geq 0$ and $\beta \geq 1$, $\mathrm{dist}(\boldsymbol{x}, \hat{\boldsymbol{X}}) = \inf_{\boldsymbol{z} \in \hat{\boldsymbol{X}}} d(\boldsymbol{x}, \boldsymbol{z})$ for some metric d on $\mathbb{R}^N$ and g_Q is a known function satisfying if $\mathrm{dist}(\boldsymbol{x}_i, Q) \to 0$, then $g_Q(\boldsymbol{x}_0) \to 0$.

Related work - sharpness

Approximate sharpness generalizes the well-known condition

$$\operatorname{dist}(\boldsymbol{x}, \hat{\boldsymbol{X}}) \leq \left(\frac{f(\boldsymbol{x}) - \hat{f}}{\alpha}\right)^{1/\beta}, \quad \forall \boldsymbol{x} \in \mathbb{R}^N,$$

dubbed sharpness, Hölderian growth or Lojasiewicz-type inequality.

Hoffman (1952), Lojasiewicz (1963), Robinson (1975), Mangasarian (1985), Auslender & Crouzeix (1988), Burke & Ferris (1993), Burke & Deng (2002), Bolte, A. Daniilidis & Lewis (2007), ...

Various works have used these conditions to quantify/accelerate convergence:

Nemirovskii & Nesterov (1985), Attouch, Bolte, Redont & Soubeyran (2010), Bolte, Nguyen, Peypouquet & Suter (2017), Bolte, Sabah & Teboulle (2014), Frankel, Garrigos & Peypouquet (2015), Karimi, Nutini & Schmidt (2016), Kerdreux, d'Aspremont & Pokutta (2019), ...

Recent works employing restart schemes specifically:

Roulet & d'Aspremont (2020), Roulet, Boumal & d'Aspremont (2020), Renegar & Grimmer (2021)

Approximate sharpness and constants

$$\operatorname{dist}(\boldsymbol{x}, \hat{\boldsymbol{X}}) \leq \left(\frac{f(\boldsymbol{x}) - \hat{f} + g_Q(\boldsymbol{x}) + \eta}{\alpha} \right)^{1/\beta}, \qquad \forall \boldsymbol{x} \in \mathbb{R}^N.$$

Generalizations:

- We allow $\eta > 0$.
- We incorporate a feasibility gap function g_Q, which means the optimization method need not produce feasible iterates.
- We do not assume the constants α, β and η are known to apply our restart scheme.

Motivation from compressed sensing

Compressed sensing concerns the recovery of (approximately) sparse vectors from incomplete sets of noisy, linear measurements.

Typical setup:

- The vector $\boldsymbol{x} \in \mathbb{C}^N$ to recover
- Measurement matrix $\boldsymbol{A} \in \mathbb{C}^{m \times N}$ (often with $m \ll N$)
- Linear measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} \in \mathbb{C}^m$, where $\boldsymbol{e} \in \mathbb{C}^m$ is noise
- Goal: Recover the vector $\boldsymbol{x}$ from the measurements $\boldsymbol{y}$

Sparsity and ℓ^1-minimization

Sparsity: $\boldsymbol{x}$ is s-sparse if it has at most s nonzero entries.

Approximate sparsity: "$\sigma_s(\boldsymbol{x})_1 := \min\{\|\boldsymbol{x}-\boldsymbol{z}\|_1 : \boldsymbol{z} \text{ is } s\text{-sparse}\}$ is small".

Standard approaches to recover (approximately) sparse $\boldsymbol{x}$ in compressed sensing involve ℓ^1-minimization, e.g. *Quadratically Constrained Basis Pursuit (QCBP)*

$$\min_{\boldsymbol{z}\in\mathbb{C}^N} \|\boldsymbol{z}\|_1 \text{ subject to } \|\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y}\|_2 \leq \varsigma.$$

Equivalent to:

$$\min_{\boldsymbol{z}\in Q} f(\boldsymbol{z}), \qquad f(\boldsymbol{z}) = \|\boldsymbol{z}\|_1, \; Q = \{\boldsymbol{z} : \|\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y}\|_2 \leq \varsigma\}.$$

Sparsity and ℓ^1-minimization

Sparsity: $\boldsymbol{x}$ is s-sparse if it has at most s nonzero entries.

Approximate sparsity: "$\sigma_s(\boldsymbol{x})_1 := \min\{\|\boldsymbol{x}-\boldsymbol{z}\|_1 : \boldsymbol{z}$ is s-sparse$\}$ is small".

Standard approaches to recover (approximately) sparse $\boldsymbol{x}$ in compressed sensing involve ℓ^1-minimization, e.g. *Quadratically Constrained Basis Pursuit (QCBP)*

$$\min_{\boldsymbol{z}\in\mathbb{C}^N} \|\boldsymbol{z}\|_1 \text{ subject to } \|\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y}\|_2 \le \varsigma.$$

Equivalent to:

$$\min_{z\in Q} f(\boldsymbol{z}), \qquad f(\boldsymbol{z}) = \|\boldsymbol{z}\|_1,\ Q = \{\boldsymbol{z} : \|\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y}\|_2 \le \varsigma\}.$$

Sparsity and ℓ^1-minimization

Sparsity: $\boldsymbol{x}$ is s-sparse if it has at most s nonzero entries.

Approximate sparsity: "$\sigma_s(\boldsymbol{x})_1 := \min\{\|\boldsymbol{x}-\boldsymbol{z}\|_1 : \boldsymbol{z}$ is s-sparse$\}$ is small".

Standard approaches to recover (approximately) sparse $\boldsymbol{x}$ in compressed sensing involve ℓ^1-minimization, e.g. *Quadratically Constrained Basis Pursuit (QCBP)*

$$\min_{\boldsymbol{z}\in\mathbb{C}^N} \|\boldsymbol{z}\|_1 \text{ subject to } \|\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y}\|_2 \leq \varsigma.$$

Equivalent to:

$$\min_{\boldsymbol{z}\in Q} f(\boldsymbol{z}), \qquad f(\boldsymbol{z}) = \|\boldsymbol{z}\|_1,\ Q = \{\boldsymbol{z} : \|\boldsymbol{A}\boldsymbol{z}-\boldsymbol{y}\|_2 \leq \varsigma\}.$$

Compressed sensing theory

Definition (Restricted Isometry Property)

Let $1 \leq s \leq N$. The sth *Restricted Isometry Constant (RIC)* δ_s of a matrix $\boldsymbol{A} \in \mathbb{C}^{m \times N}$ is the smallest $\delta \geq 0$ such that

$$(1-\delta)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq (1+\delta)\|\boldsymbol{x}\|_2^2,$$

for all s-sparse vectors $\boldsymbol{x}$. If $0 \leq \delta_s \leq 1$, then $\boldsymbol{A}$ is said to have the *Restricted Isometry Property (RIP)* of order s.

Intuition: $\boldsymbol{A}$ approximately preserves the norm of any s-sparse vector.

Adcock & Hansen (2021), Foucart & Rauhut (2013)

Approximate sharpness in compressed sensing

Lemma

Suppose that $\boldsymbol{A} \in \mathbb{C}^{m\times N}$ *has the RIP of order* $2s$ *with constant* $\delta = \delta_{2s} < \sqrt{2} - 1$*. Then the QCBP problem satisfies*

$$\operatorname{dist}(\boldsymbol{x}, \hat{\boldsymbol{X}}) \leq \left(\frac{f(\boldsymbol{x}) - \hat{f} + g_Q(\boldsymbol{x}) + \eta}{\alpha} \right)^{1/\beta}, \qquad \forall \boldsymbol{x} \in \mathbb{C}^N,$$

where $g_Q(\boldsymbol{z}) = \sqrt{s}\max\{\|\boldsymbol{Az} - \boldsymbol{y}\|_2 - \varsigma, 0\}$, $\alpha = C_1\sqrt{s}$, $\beta = 1$, $\eta = C_2\sigma_s(\boldsymbol{x})_1 + C_3\sqrt{s}\varsigma$, *and the constants* C_1, C_2, C_3 *depend on* δ *only.*

Approximate sharpness: the distance to $\hat{\boldsymbol{X}}$ is bounded by:

- the error in the objective function $f(\boldsymbol{x}) - \hat{f}$
- the feasibility gap $g_Q(\boldsymbol{x})$
- the underlying compressed sensing error η

Approximate sharpness in compressed sensing

$$\operatorname{dist}(\boldsymbol{x}, \hat{\boldsymbol{X}}) \leq \left(\frac{f(\boldsymbol{x}) - \hat{f} + g_Q(\boldsymbol{x}) + \eta}{\alpha} \right)^{1/\beta}, \qquad \forall \boldsymbol{x} \in \mathbb{R}^N$$

In the compressed sensing example with

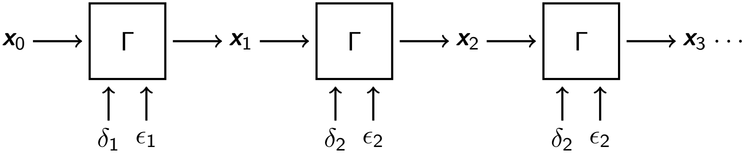
$$\alpha = C_1 \sqrt{s}, \qquad \beta = 1, \qquad \eta = C_2 \sigma_s(\boldsymbol{x})_1 + C_3 \sqrt{s} \varsigma.$$

- the order s of the RIP may be unknown
- $\sigma_s(\boldsymbol{x})_1$ is typically unknown
- C_1, C_2, C_3 depend on the RIC δ
- moreover, given $\boldsymbol{A}$ and s, finding the RIC δ is NP-hard

Outline

Restart scheme

Let Γ be a first-order method that takes input $(\boldsymbol{x}, \delta, \epsilon) \in \mathbb{C}^N \times \mathbb{R}_+ \times \mathbb{R}_+$.



- Run multiple instances of Γ, where the output $\boldsymbol{x}_k$ of the kth instance is used as the input of the $(k+1)$th instance
- Update the parameters $(\delta, \epsilon) = (\delta_{k+1}, \epsilon_{k+1})$ using the approximate sharpness condition
- Restarts can be extended to perform a grid search over α or β if their values are unknown, while preserving the order of convergence

Restart scheme for unknown α, β, η

Algorithm 2: Restart scheme for unknown α, β and η in (1.2) via grid search.

Input : Optimization algorithm Γ for (1.1), bijection ϕ as in Definition 3.1, initial vector $x^{(0)} \in D$, upper bound ϵ_0 such that $f(x^{(0)}) - \hat{f} + g_Q(x^{(0)}) \leq \epsilon_0$, constants $a, b > 1$, $r \in (0,1)$, $\alpha_0 > 0$, $\beta_0 \geq 1$ and total number of inner iterations $t \in \mathbb{N}$.

Output: Final iterate $x^{(t)}$ approximating a solution to (1.1).

Initialize $x^{(0)} = x_0$, $U_{i,j} = 0$, $V_{i,j} = 0$, $\epsilon_{i,j,0} = \epsilon_0$ for all $i \in \mathbb{Z}, j \in \mathbb{N}_0$;
for $m = 0, 1, \ldots, t-1$ **do**
  $(i, j, k) \leftarrow \phi(m+1)$;
  $\alpha_i \leftarrow a^i \alpha_0,\ \beta_j \leftarrow b^j \beta_0,\ U \leftarrow U_{i,j},\ V \leftarrow V_{i,j}$;
  $\epsilon_{i,j,U+1} \leftarrow r\epsilon_{i,j,U}$;
  if $2\epsilon_{i,j,U} > \alpha_i$ **then**
    $\delta_{i,j,U+1} \leftarrow \left(\frac{2\epsilon_{i,j,U}}{\alpha_i}\right)^{\min\{b/\beta_j, 1/\beta_0\}}$;
  else
    $\delta_{i,j,U+1} \leftarrow \left(\frac{2\epsilon_{i,j,U}}{\alpha_i}\right)^{1/\beta_j}$;
  end
  if $V + \mathcal{C}_\Gamma(\delta_{i,j,U+1}, \epsilon_{i,j,U+1}) \leq k$ **then**
    $z^{(m)} \leftarrow \Gamma(\delta_{i,j,U+1}, \epsilon_{i,j,U+1}, x^{(m)})$;
    $x^{(m+1)} \leftarrow \operatorname{argmin}\{f(x) + g_Q(x) : x = z^{(m)} \text{ or } x = x^{(m)}\}$;
    $V_{i,j} \leftarrow V + \mathcal{C}_\Gamma(\delta_{i,j,U+1}, \epsilon_{i,j,U+1})$;
    $U_{i,j} \leftarrow U + 1$;
  else
    $x^{(m+1)} = x^{(m)}$;
  end
end

Adcock, Colbrook & Neyra-Nesterenko (2024)

Theorem (Unknown α, β, η)

Suppose the number of iterations computed by Γ is at most $C\delta^{d_1}/\epsilon^{d_2}+1$, for all $\delta,\epsilon>0$. Then there is a restart scheme such that after at most

$$\hat{C}\cdot\varepsilon^{d_1/\beta_*-d_2}\cdot\begin{cases}\lceil\log(1/\varepsilon)\rceil, & \text{if } d_2\le d_1/\beta_*,\\ 1, & \text{if } d_2>d_1/\beta_*,\end{cases}$$

iterations of Γ, where β_* is the scheme's closest grid point to β, the restart scheme produces an output $\boldsymbol{x}^\star$ satisfying

$$f(\boldsymbol{x}^\star)-\hat{f}+g_Q(\boldsymbol{x}^\star)\le\max\{\varepsilon,\eta\}.$$

Here $\hat{C}$ depends on C,α,β_*,d_1 and d_2.

Remark: There are restart schemes for the special cases when α or β are known, and analogous results can be stated.

Adcock, Colbrook & Neyra-Nesterenko (2024)

Rates for different problem classes

Objective function class/structure	Asymptotic bound for $K(\varepsilon)$	Example method
$L-$smooth See Definition 4.2 (NB: must have $\beta \geq 2$)	$\beta = 2$: $\sqrt{L/\alpha} \cdot \log(1/\varepsilon)$ $\beta > 2$: $\frac{\sqrt{L}}{\alpha^{1/\beta_*}} \cdot \frac{1}{\varepsilon^{1/2-1/\beta_*}}$	Nesterov's method $d_1 = 1$, $d_2 = 1/2$ See Section 4.1
$(u, v)-$smoothable See Definition 4.5	$\beta = 1$: $\frac{\sqrt{ab}}{\alpha} \cdot \log(1/\varepsilon)$ $\beta > 1$: $\frac{\sqrt{ab}}{\alpha^{1/\beta_*}} \cdot \frac{1}{\varepsilon^{1-1/\beta_*}}$	Nesterov's method with smoothing $d_1 = 1$, $d_2 = 1$ See Section 4.2
Hölder smooth, parameter $\nu \in [0, 1]$ See Definition 4.8 (NB: must have $\beta \geq 1 + \nu$)	$\beta = 1+\nu$: $\frac{M_\nu^{\frac{2}{1+3\nu}}}{\alpha^{\frac{2}{(1+3\nu)}}} \cdot \log(1/\varepsilon)$ $\beta > 1+\nu$: $\frac{M_\nu^{\frac{2}{1+3\nu}}}{\alpha^{\frac{2+2\nu}{\beta_*(1+3\nu)}}} \cdot \frac{1}{\varepsilon^{\frac{2(\beta_*-1-\nu)}{\beta_*(1+3\nu)}}}$	Universal fast gradient method $d_1 = (2 + 2\nu)/(1 + 3\nu)$ $d_2 = 2/(1 + 3\nu)$ See Section 4.3
$f(x) = q(x) + g(x) + h(Bx)$, q is L_q-smooth, $\sup_{z \in \mathrm{dom}(h)} \inf_{y \in \partial h(z)} \|y\| \leq L_h$, $\|B\| \leq L_B$	$\beta = 1$: $\frac{L_B L_h + L_q}{\alpha} \cdot \log(1/\varepsilon)$ $\beta > 1$: $\frac{L_B L_h + L_q}{\alpha^{1/\beta_*}} \cdot \frac{1}{\varepsilon^{1-1/\beta_*}}$	Primal-dual algorithm $d_1 = 1$, $d_2 = 1$ See Section 4.4
$f(x) = q(x) + g(x) + h(Bx)$, q is L_q-smooth, $\sup_{z \in \mathrm{dom}(h)} \inf_{y \in \partial h(z)} \|y\| \leq L_h$, $\|A\| \leq L_A$, $\|B\| \leq L_B$, $Q = \{x : Ax \in C\}$, $g_Q(x) = \kappa \inf_{z \in C} \|Ax - z\|$	$\beta = 1$: $\frac{\kappa L_A + L_B L_h + L_q}{\alpha} \cdot \log(1/\varepsilon)$ $\beta > 1$: $\frac{\kappa L_A + L_B L_h + L_q}{\alpha^{1/\beta_*}} \cdot \frac{1}{\varepsilon^{1-1/\beta_*}}$	Primal-dual algorithm with constraints $d_1 = 1$, $d_2 = 1$ See Section 4.5

Table 1: Asymptotic cost bounds (as $\varepsilon \downarrow 0$ for $\eta \lesssim \varepsilon$) and suitable first-order methods for Algorithm 2 when applied to different classes of objective functions. Note also that whenever the bound is a polynomial in $\log(1/\varepsilon)$, we have $\beta_* = \beta$.

Remark: The first three lines constitute optimal rates.

Adcock, Colbrook & Neyra-Nesterenko (2024)

Outline

Motivations for approximate sharpness

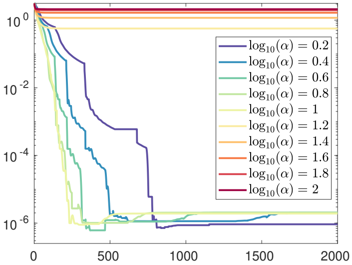
Restart schemes for approximately sharp optimization problems

Numerical example

Conclusions

Compressed sensing example

If $d_1 = d_2/\beta$, then the cost bound reduces to $\hat{C} \cdot \log(1/\varepsilon)$, yielding linear decay to η.



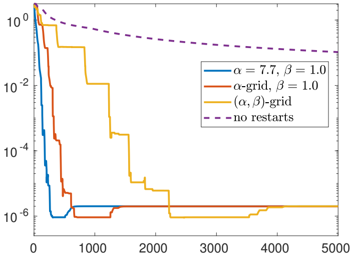
Recovery error vs. # iterations

Example: Γ is primal-dual iteration applied to QCBP, where $\beta = d_1 = d_2 = 1$.

This is applied to our compressed sensing problem, where $\boldsymbol{A}$ is a Gaussian random matrix and $\varsigma = 10^{-6}$. The ground truth $\boldsymbol{x}$ is exactly sparse, hence $\eta \approx \varsigma$.

Issue: Restarts are brittle with respect to fixed α.

Compressed sensing example



Recovery error vs. # iterations

A direct comparison of restart schemes with tuned constants and the nonrestarted optimization method Γ (primal-dual iterations).

Grid searching maintains linear decay and still outperforms the non-restarted optimization method.

Outline

Conclusions

Convex optimization problems arising in applications, such as image and signal reconstruction, matrix completion, feature selection) satisfy an approximate sharpness condition with unknown constants.

In this setting, our goal is to obtain fast convergence down to the (unknown) approximate sharpness constant η.

We introduced an algorithm for accelerating any convex optimization method, based on restarts and grid searching.

This leads to optimal rates for various convex optimization problems and competitive practical performance.

Paper: *Restarts subject to approximate sharpness: A parameter-free and optimal scheme for first-order methods*. Foundations of Computational Mathematics (in press, 2024).