Restarts of Accelerated Gradient Methods: Generic Theoretical Speed-up

Vincent Roulet
University of Washginton

West Coast Optimization Meeting Fall 2019

with Alexandre d'Aspremont CNRS, Ecole Normale Superieure

Convex Optimization

Consider for $f : \mathbb{R}^d \to \mathbb{R}$ closed convex,

$$\min_x \quad f(x)$$

Convex Optimization

Consider for $f : \mathbb{R}^d \to \mathbb{R}$ closed convex,

$$\min_x \quad f(x)$$

Accelerated Gradient method convergence ingredients:

- Smoothness $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2$ for all $x, y \in \operatorname{dom} f$

  $$f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|x-y\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f$$

  $\rightarrow$ upper bound at each iterate
- Convexity

  $$f(y) \geq f(x) + \nabla f(x)^\top (y-x) \quad \text{for all } x, y \in \operatorname{dom} f$$

  $\rightarrow$ lower bound on previous iterates

Provides convergence at rate $\mathcal{O}(1/k^2)$

[Nesterov, 1983; Diakonikolas and Orecchia, 2019]

Additional assumptions

Strong convexity

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|x - y\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f$$

$\rightarrow$ provides linear rate of convergence to the minimum

Additional assumptions

Strong convexity

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|x-y\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f$$

$\rightarrow$ provides linear rate of convergence to the minimum

Can we relax strong convexity assumption
and still get faster rates than plain convexity ?

Additional assumptions

Strong convexity

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|x - y\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f$$

$\rightarrow$ provides linear rate of convergence to the minimum

Can we relax strong convexity assumption
and still get faster rates than plain convexity ?

Here using **error bounds** and **restarts** of accelerated gradient

Error Bounds

Restarts of Smooth Functions

Restart for Non-smooth Functions

Plan

Error Bounds

Restarts of Smooth Functions

Restart for Non-smooth Functions

Hölderian Error Bounds

Definition

A function f satisfies a *Hölderian error bound* on a set K if there exist $r \geq 1$, $\mu > 0$, s.t.

$$\frac{\mu}{r} d(x, X^*)^r \leq f(x) - f^*, \quad \text{for all } x \in K, \qquad (\mathrm{HEB_{r,\,\mu}(K)})$$

where $f^* = \min f$, $X^* = \arg\min f$, $d(x, X^*) = \min_{y \in X^*} \|x - y\|_2$

Hölderian Error Bounds

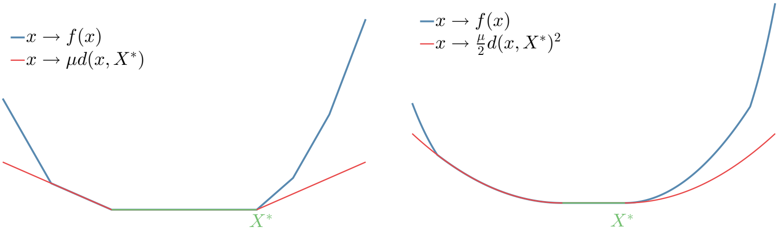
Definition

A function f satisfies a *Hölderian error bound* on a set K if there exist $r \geq 1$, $\mu > 0$, s.t.

$$\frac{\mu}{r} d(x, X^*)^r \leq f(x) - f^*, \quad \text{for all } x \in K, \qquad (\mathrm{HEB}_{r,\,\mu}(K))$$

where $f^* = \min f$, $X^* = \arg\min f$, $d(x, X^*) = \min_{y \in X^*} \|x - y\|_2$

Lower bound on the function around minimizers



$r = 1$ (sharp) $\qquad$ $r = 2$ ($\approx$ strongly convex)

Hölderian Error Bounds

Hölderian error bound

$$\frac{\mu}{r} d(x, X^*)^r \leq f(x) - f^*, \quad \text{for all } x \in K, \qquad (\mathrm{HEB}_{\mathrm{r},\,\mu}(\mathrm{K}))$$

Remarks

- covers strong convexity ($r = 2$)
- covers $\ell_{1,p}$ regularization of Least-Squares ($r = 2$)

$$\min_x \|Ax - b\|_2^2 + \|x\|_{1,p}$$

  [Zhou et al., 2015; Drusvyatskiy and Lewis, 2018]
- covers zero-sum game problems ($r = 1$)

$$\min_{x \in \Delta} \{f(x) = \max_{y \in \Delta} x^\top Ay\}$$

  [Gilpin et al., 2012]
- equivalent to Łojasiewicz inequality (gradient dominated) [Bolte et al., 2017]
- generically satisfied by subanalytic functions (r unknown) [Łojasiewicz, 1963; Bolte et al., 2007]

Error Bound and Smoothness

Combining ($\mathrm{HEB}_{\mathrm{r},\ \mu}$(K)) lower bound and smoothness upper bound,

$$\frac{\mu}{r} d(x, X^*)^r \leq f(x) - f^* \leq \frac{L}{2} d(x, X^*)^2$$

We get

$$0 < \frac{2\mu}{rL} \leq \frac{d(x, X^*)^2}{d(x, X^*)^r}$$

Error Bound and Smoothness

Combining ($\mathrm{HEB}_{\mathrm{r},\,\mu}(\mathrm{K})$) lower bound and smoothness upper bound,

$$\frac{\mu}{r} d(x, X^*)^r \leq f(x) - f^* \leq \frac{L}{2} d(x, X^*)^2$$

We get

$$0 < \frac{2\mu}{rL} \leq \frac{d(x, X^*)^2}{d(x, X^*)^r}$$

Consequences:

- Necessarily $2 \leq r$ (take $x \to X^*$)
- If $2 < r$, only valid on subset of dom f ,here

$$K = S_0 \triangleq \{x : f(x) \leq x_0\}$$

Plan

Error Bounds

Restarts of Smooth Functions

Restart for Non-smooth Functions

Restarts

Principle:
Run accelerated algo on the cvx pb, stop it, restart from last iterate.

Question: When must the algorithm be stopped ?

This talk [R. and d'Aspremont, 2017]:

- schedule the restarts in advance
  $\rightarrow$ requires all parameters to be known
- stop when gap has decreased by constant factor
  $\rightarrow$ requires knowing f^*

Scheduled restarts

Accelerated gradient [Nesterov, 1983]
Starting from $\bar{x}$, outputs after t iterations

$$\hat{x} = \mathcal{A}(\bar{x}, t) \quad \text{s.t.} \quad f(\hat{x}) - f^* \leq \frac{4L}{t^2} d(\bar{x}, X^*)^2,$$

Scheduled restart
Schedule restarts in advance at times t_k and build from $x_0 \in \mathbb{R}^d$

$$x_k = \mathcal{A}(x_{k-1}, t_k)$$

Scheduled restarts

Accelerated gradient [Nesterov, 1983]

Starting from $\bar{x}$, outputs after t iterations

$$\hat{x} = \mathcal{A}(\bar{x}, t) \quad \text{s.t.} \quad f(\hat{x}) - f^* \leq \frac{4L}{t^2} d(\bar{x}, X^*)^2,$$

Scheduled restart

Schedule restarts in advance at times t_k and build from $x_0 \in \mathbb{R}^d$

$$x_k = \mathcal{A}(x_{k-1}, t_k)$$

Ingredients

Combine convergence bound and sharpness

$$f(x_k) - f^* \leq \frac{4L}{t_k^2} d(x_{k-1}, X^*)^2 \quad \text{and} \quad \frac{\mu}{r} d(x_{k-1}, X^*)^r \leq f(x_{k-1}) - f^*$$

So $$f(x_k) - f^* \leq \tfrac{c_{L,\mu,r}}{t_k^2} (f(x_{k-1}) - f^*)^{2/r}$$

Scheduled restarts

Accelerated gradient [Nesterov, 1983]

Starting from $\bar{x}$, outputs after t iterations

$$\hat{x} = \mathcal{A}(\bar{x}, t) \quad \text{s.t.} \quad f(\hat{x}) - f^* \leq \frac{4L}{t^2} d(\bar{x}, X^*)^2,$$

Scheduled restart

Schedule restarts in advance at times t_k and build from $x_0 \in \mathbb{R}^d$

$$x_k = \mathcal{A}(x_{k-1}, t_k)$$

Ingredients

Combine convergence bound and sharpness

$$f(x_k) - f^* \leq \frac{4L}{t_k^2} d(x_{k-1}, X^*)^2 \quad \text{and} \quad \frac{\mu}{r} d(x_{k-1}, X^*)^r \leq f(x_{k-1}) - f^*$$

So $$f(x_k) - f^* \leq \frac{c_{L,\mu,r}}{t_k^2} (f(x_{k-1}) - f^*)^{2/r}$$

1. Fix $0 < \gamma < 1$, find $(t_k)_{k \geq 1}$ s.t. $f(x_k) - f^* \leq \gamma (f(x_{k-1}) - f^*)$
2. Optimize γ for optimal rate w.r.t. $N = \sum_{i=1}^R t_k$ after R restarts

Optimal Schedule

Proposition [R. and d'Aspremont, 2017]

For f convex, L-smooth satisfying $(\text{HEB}_{r,\,\mu}(S_0))$, denote

$$\tau = 1 - 2/r \in [0,1) \quad \text{and} \quad \kappa = L/\mu^{2/r}$$

Run scheduled restarts with

$$t_k = C_{\tau,\kappa} e^{\tau k}$$

After R restarts and $N = \sum_{i=1}^{R} t_k$ total iterations, we get $\hat{x}$ s.t.

$$f(\hat{x}) - f^* = O\left(\exp(-\kappa^{-1/2} N)\right) \qquad \text{when } \tau = 0 \qquad (1)$$

$$f(\hat{x}) - f^* = O\left(N^{-2/\tau}\right) \qquad \text{when } \tau > 0 \qquad (2)$$

Remarks

- Retrieve accelerated rate for strongly convex functions,
- Optimal for this class of problems [Nemirovskii and Nesterov, 1985]

Optimal Schedule

Proposition [R. and d'Aspremont, 2017]

For f convex, L-smooth satisfying ($\mathrm{HEB}_{\mathrm{r},\,\mu}(S_0)$), denote

$$\tau = 1 - 2/r \in [0, 1) \quad \text{and} \quad \kappa = L/\mu^{2/r}$$

Run scheduled restarts with

$$t_k = C_{\tau,\kappa} e^{\tau k}$$

After R restarts and $N = \sum_{i=1}^{R} t_k$ total iterations, we get $\hat{x}$ s.t.

$$f(\hat{x}) - f^* = O\left(\exp(-\kappa^{-1/2} N)\right) \qquad \text{when } \tau = 0 \qquad (1)$$

$$f(\hat{x}) - f^* = O\left(N^{-2/\tau}\right) \qquad \text{when } \tau > 0 \qquad (2)$$

Technical detail

- *Detailed bound continuous in τ:*
  for $\tau \to 0$, right hand side of (2) $\to$ right hand side of (1)

Parameter-free strategy

Adaptive strategy (log-scale grid search)

Given a fixed budget of iterations N, search with schedules like

$$t_k = Ce^{\tau k} \tag{3}$$

- Grid on C limited by N
- Grid on C limited by continuity of the bounds in τ

Parameter-free strategy

Adaptive strategy (log-scale grid search)
Given a fixed budget of iterations N, search with schedules like

$$t_k = Ce^{\tau k} \tag{3}$$

- Grid on C limited by N
- Grid on C limited by continuity of the bounds in τ

Proposition [R. and d'Aspremont, 2017]

For f convex, L-smooth satisfying $(\mathrm{HEB}_{\mathrm{r},\ \mu}(S_0))$, run restart schemes with schedules of the form (3) on a $\log_2$-scale grid for a budget of N iterations.
Get one scheme nearly optimal up to a factor 4, costs $(\log_2 N)^2$ times more than running optimal schedule for N iterations

Restarts with sufficient gap decrease

Scheme
Assume f^* is known, run accelerated algo from x_{k-1}, stop for y_t s.t.

$$f(y_t) - f^* \leq \gamma(f(x_{k-1}) - f^*) \tag{4}$$

where $\gamma < 1$ and iterate the process with $x_k = y_t$

Restarts with sufficient gap decrease

Scheme

Assume f^* is known, run accelerated algo from x_{k-1}, stop for y_t s.t.

$$f(y_t) - f^* \leq \gamma(f(x_{k-1}) - f^*) \tag{4}$$

where $\gamma < 1$ and iterate the process with $x_k = y_t$

Proposition [R. and d'Aspremont, 2017]

For f convex, L-smooth satisfying ($\mathrm{HEB}_{\mathrm{r},\,\mu}(S_0)$), restarts monitoring the decrease gap (4) with $\gamma = e^{-2}$ do not worse than the optimal scheduled restart.

Remark:

- Does not need any knowledge of the parameters

Gradient Descend Analysis

Gradient descend convergence rate can be written

$$f(x_{k+t}) - f^* \leq \frac{L}{t} d(x_k, X^*)^2, \quad \text{for any } t, k \geq 0$$

$\rightarrow$ same analysis can be done under the HEB property

Gradient Descend Analysis

Gradient descend convergence rate can be written

$$f(x_{k+t}) - f^* \leq \frac{L}{t} d(x_k, X^*)^2, \quad \text{for any } t, k \geq 0$$

$\rightarrow$ same analysis can be done under the HEB property

Proposition [R. and d'Aspremont, 2017]

For f convex, L-smooth satisfying ($\mathrm{HEB}_{\mathrm{r},\,\mu}(S_0)$), denote

$$\tau = 1 - 2/r \in [0,1) \quad \text{and} \quad \kappa = L/\mu^{2/r}$$

After N iterations of the gradient descend, we get $\hat{x}$ s.t.

$$f(\hat{x}) - f^* = O\left(\exp(-\kappa N)\right) \qquad \text{when } \tau = 0$$

$$f(\hat{x}) - f^* = O\left(N^{-1/\tau}\right) \qquad \text{when } \tau > 0$$

Plan

Error Bounds

Restarts of Smooth Functions

Restart for Non-smooth Functions

Composite problems

Consider

$$\min_x f(x) \triangleq h(x) + g(x) \tag{5}$$

with h smooth convex, g prox-friendly convex.

Accelerated algorithm [Nesterov, 2013]

Started at x_{k-1}, outputs after t_k iterations x_k s.t.

$$f(x_k) - f^* \leq \frac{4L}{t_k^2} d(x_{k-1}, X^*)^2$$

Composite problems

Consider

$$\min_x f(x) \triangleq h(x) + g(x) \tag{5}$$

with h smooth convex, g prox-friendly convex.

Accelerated algorithm [Nesterov, 2013]

Started at x_{k-1}, outputs after t_k iterations x_k s.t.

$$f(x_k) - f^* \leq \frac{4L}{t_k^2} d(x_{k-1}, X^*)^2$$

Same bound $\rightarrow$ same analysis.

Corollary

For f defined as in (5) with h L-smooth, g prox-friendly, if f satisfies ($\text{HEB}_{\text{r},\,\mu}(S_0)$), then scheduled restarts and restarts on decreasing gap have same complexities as presented before.

Remark:

- Captures $\ell_{1,p}$ regularization

Non-smooth and Hölder smooth

Generic smoothness For f convex,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \frac{L}{s}\|x - y\|_2^s \quad \text{for all } x, y \in \operatorname{dom} f \qquad (\mathrm{S}_{\mathrm{s},\ \mathrm{L}})$$

and any $\nabla f(x) \in \partial f(x)$, $\nabla f(y) \in \partial f(y)$.

- for $s = 1$ retrieves assumption for non-smooth cvx functions
- for $1 < s < 2$ gets definition of Hölder smooth functions

Non-smooth and Hölder smooth

Generic smoothness For f convex,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \frac{L}{s}\|x - y\|_2^s \quad \text{for all } x, y \in \operatorname{dom} f \qquad (\mathrm{S_{s,\ L}})$$

and any $\nabla f(x) \in \partial f(x)$, $\nabla f(y) \in \partial f(y)$.

- for $s = 1$ retrieves assumption for non-smooth cvx functions
- for $1 < s < 2$ gets definition of Hölder smooth functions

Combined with Hölderian error bounds
Assume f satisfies $(\mathrm{S_{s,\ L}})$ and $(\mathrm{HEB_{r,\ \mu}}(S_0))$ then necessary

$$s \leq r$$

Schedule restarts

Universal fast gradient method [Nesterov, 2015]
Starting from $\bar{x}$, given a target accuracy ϵ, outputs after t iterations

$$\hat{x} = \mathcal{U}(\bar{x}, t, \epsilon) \quad \text{s.t.} \quad f(\hat{x}) - f^* \leq \frac{\epsilon}{2} + \frac{cL^{\frac{2}{s}} d(\bar{x}, X^*)^2}{\epsilon^{\frac{2}{s}} t^{\frac{2\rho}{s}}} \frac{\epsilon}{2}$$

with $\rho = 3s/2 - 1$

Schedule restarts

Universal fast gradient method [Nesterov, 2015]
Starting from $\bar{x}$, given a target accuracy ϵ, outputs after t iterations

$$\hat{x} = \mathcal{U}(\bar{x}, t, \epsilon) \quad \text{s.t.} \quad f(\hat{x}) - f^* \leq \frac{\epsilon}{2} + \frac{cL^{\frac{2}{s}} d(\bar{x}, X^*)^2}{\epsilon^{\frac{2}{s}} t^{\frac{2\rho}{s}}} \frac{\epsilon}{2}$$

with $\rho = 3s/2 - 1$

Universal scheduled restart
Schedule restarts at times t_k with precision ϵ_k, i.e.

$$x_k = \mathcal{U}(x_{k-1}, t_k, \epsilon_k)$$

starting from $x_0 \in \operatorname{dom} f$ and $\epsilon_0 \geq f(x_0) - f^*$

Optimal Schedule

Proposition [R. and d'Aspremont, 2017]

For f convex, satisfying $(\mathrm{S}_{\mathrm{s},\ \mathrm{L}})$ and $(\mathrm{HEB}_{\mathrm{r},\ \mu}(S_0))$ denote

$$\tau = 1 - s/r \in [0, 1) \quad \text{and} \quad \kappa = L^{2/s}/\mu^{2/r}$$

Run scheduled restarts with

$$t_k = C_{\tau,\kappa} e^{\tau k}, \quad \epsilon_k = e^{-\rho}\epsilon_{k-1}$$

After R restarts and $N = \sum_{i=1}^{R} t_k$ total iterations, we get $\hat{x}$ s.t.

$$f(\hat{x}) - f^* = O\left(\exp(-\kappa^{-s/2\rho} N)\right) \qquad \text{when } \tau = 0$$

$$f(\hat{x}) - f^* = O\left(N^{-\rho/\tau}\right) \qquad \text{when } \tau > 0$$

Remarks

- Optimal for this class of problems [Nemirovskii and Nesterov, 1985]
- Log-scale grid-search fails to get nearly optimal rate
- Needs to stay in the initial sub-level set, which can be enforced

Restarts with sufficient gap decrease

Scheme

Assume f^* is known, run universal fast algo from x_{k-1}, with precision $\epsilon_k = \gamma(f(x_{k-1}) - f^*)$, stop when it outputs x_k s.t.

$$f(x_k) - f^* \leq \epsilon_k \qquad (6)$$

where $\gamma < 1$ and iterate the process

Restarts with sufficient gap decrease

Scheme

Assume f^* is known, run universal fast algo from x_{k-1}, with precision $\epsilon_k = \gamma(f(x_{k-1}) - f^*)$, stop when it outputs x_k s.t.

$$f(x_k) - f^* \leq \epsilon_k \tag{6}$$

where $\gamma < 1$ and iterate the process

Proposition [R. and d'Aspremont, 2017]

For f convex, satisfying ($\mathrm{S}_{\mathrm{s,\ L}}$) and ($\mathrm{HEB}_{\mathrm{r},\ \mu}(S_0)$), restarts monitoring the decrease gap (6) with $\gamma = e^{-\rho}$ do not worse than the optimal scheduled restart.

Remark:

- Does not need any knowledge of the parameters
- Taking $\gamma = e^{-1}$ is suboptimal by a factor at most 1.3

Smoothable objectives

Consider

$$\min_x f(x) \triangleq \phi(Ax) + g(x)$$

with ϕ non-smooth cvx with analytically computable Moreau envelope, g cvx prox-friendly, e.g., matrix sum-game

Smoothable objectives

Consider

$$\min_x f(x) \triangleq \phi(Ax) + g(x)$$

with ϕ non-smooth cvx with analytically computable Moreau envelope, g cvx prox-friendly, e.g., matrix sum-game

Smoothing and acceleration [Nesterov, 2005]
Starting from $\bar{x}$, given a target accuracy ϵ, outputs after t iterations

$$\hat{x} = \mathcal{S}(\bar{x}, \epsilon, t) \quad \text{s.t.} \quad f(\hat{x}) - f^* \leq \frac{\epsilon}{2} + \frac{cL^2_{\psi^*,A} D_h(\bar{x}, X^*)}{\epsilon^2 t^2} \frac{\epsilon}{2},$$

Similar bound $\rightarrow$ same analysis

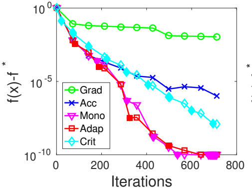
Remark

- Retrieves algorithm of [Gilpin et al., 2012] for zero-sum games

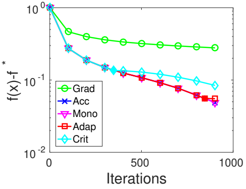
Numerical Illustrations

Classification on UCI dataset ($n = 206$ samples, $d = 60$ features), compare

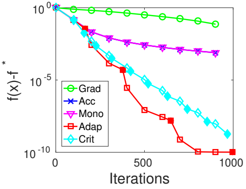
- Restarts enforcing monotonicity, Mono
  i.e., stop and restart when $f(y_{t+1}) \geq f(y_k)$
- Best scheduled restart found by grid-search Adap
- Restart with f^* known Crit



Least-squares,



Logistic,



Dual SVM.

Conclusion

Restarts get fastest rates for convex problem with error bounds

Yet, needs **adaptivity**,

- adaptivity to unknown μ for $r = 2$ [Fercoq and Qu, 2016, 2017]
- here for smooth problems and any μ, r
- universal restart scheme any r, s, μ, L [Renegar and Grimmer, 2018]

Extensions

- applies also to conditional gradient [Kerdreux et al., 2019]

Conclusion

Restarts get fastest rates for convex problem with error bounds

Yet, needs **adaptivity**,

- adaptivity to unknown μ for $r = 2$ [Fercoq and Qu, 2016, 2017]
- here for smooth problems and any μ, r
- universal restart scheme any r, s, μ, L [Renegar and Grimmer, 2018]

Extensions

- applies also to conditional gradient [Kerdreux et al., 2019]

Thanks ! Questions ?

Sparse Recovery Problems

Recovery objective

Recover a s-sparse signal $\bar{x} \in \mathbb{R}^d$ from $n < d$ linear observations

$$b_i = a_i^T \bar{x}, \quad i \in \{1, \ldots, n\}$$

Decoding procedure

$$\begin{aligned} \min_x \quad & \|x\|_1 \\ \text{subject to} \quad & Ax = b \end{aligned}$$

Recovery threshold

Given $A \in \mathbb{R}^{n \times d}$, denote $s_{\max}(A)$ its recovery threshold s.t. for any $\bar{x}$ is s-sparse, if $s < s_{\max}(A)$, then is the unique solution of

$$\begin{aligned} \min_x \quad & \|x\|_1 \\ \text{subject to} \quad & Ax = A\bar{x} \end{aligned}$$

Recovery performance

Proposition [R., Boumal and d'Aspremont, 2019]

Given $A \in \mathbb{R}^{n \times d}$ and $\bar{x}$, s-sparse, with $s < s_{\max}(A)$,

$$\|x\|_1 - \|x^*\|_1 > (1 - \sqrt{s/s_{\max}(A)})\|x - x^*\|_1 \quad \forall x : Ax = Ax^*, x \neq x^*$$

so the decoding problem satisfies $(\mathrm{HEB}_{1,\,\mu})$

Rate of convergence of optimal restart scheme reads

$$\|\hat{x}\|_1 - \|x^*\|_1 = O\left(\exp\left(-\left(1 - \sqrt{s/s_{\max}(A)}\right) N\right)\right)$$

Illustration

For random observation matrix A, $s_{\max}(A) \approx n / \log d$

So to recover s-sparse signals, needs $n \approx s \log d$

Convergence rate of optimal restart

$$\|\hat{x}\|_1 - \|x^*\|_1 = O\left(\exp\left(-\left(1 - c\sqrt{s \log d / n}\right) N\right)\right)$$

Bolte, J., Daniilidis, A. and Lewis, A. [2007], 'The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems', *SIAM Journal on Optimization* **17**(4), 1205–1223.

Bolte, J., Nguyen, T. P., Peypouquet, J. and Suter, B. W. [2017], 'From error bounds to the complexity of first-order descent methods for convex functions', *Mathematical Programming* **165**(2), 471–507.

Diakonikolas, J. and Orecchia, L. [2019], 'The approximate duality gap technique: A unified theory of first-order methods', *SIAM Journal on Optimization* **29**(1), 660–689.

Drusvyatskiy, D. and Lewis, A. S. [2018], 'Error bounds, quadratic growth, and linear convergence of proximal methods', *Mathematics of Operations Research* **43**(3), 919–948.

Fercoq, O. and Qu, Z. [2016], 'Restarting accelerated gradient methods with a rough strong convexity estimate', *arXiv preprint arXiv:1609.07358* .

Fercoq, O. and Qu, Z. [2017], 'Adaptive restart of accelerated gradient methods under local quadratic growth condition', *arXiv preprint arXiv:1709.02300* .

Gilpin, A., Pena, J. and Sandholm, T. [2012], 'First-order algorithm with ln(1/epsilon) convergence for epsilon-equilibrium in two-person zero-sum games', *Mathematical programming* **133**(1-2), 279–298.

Kerdreux, T., d'Aspremont, A. and Pokutta, S. [2019], Restarting frank-wolfe, *in* 'The 22nd International Conference on Artificial Intelligence and Statistics', pp. 1275–1283.

Łojasiewicz, S. [1963], 'Une propriété topologique des sous-ensembles analytiques réels', *Les équations aux dérivées partielles* pp. 87–89.

Nemirovskii, A. and Nesterov, Y. [1985], 'Optimal methods of smooth convex minimization', *USSR Computational Mathematics and Mathematical Physics* **25**(2), 21–30.

Nesterov, Y. [2005], 'Smooth minimization of non-smooth functions', *Mathematical programming* **103**(1), 127–152.

Nesterov, Y. [2013], 'Gradient methods for minimizing composite functions', *Mathematical Programming* **140**(1), 125–161.

Nesterov, Y. [2015], 'Universal gradient methods for convex optimization problems', *Mathematical Programming* **152**(1-2), 381–404.

Nesterov, Y. E. [1983], A method for solving the convex programming problem with convergence rate o (1/kˆ 2), *in* 'Dokl. akad. nauk Sssr', Vol. 269, pp. 543–547.

R., V., Boumal, N. and d'Aspremont, A. [2019], 'Complexity verus statistacal performance on sparse recovery problems', *Information and Inference : A Journal of the IMA* .

R., V. and d'Aspremont, A. [2017], Sharpness, restart and acceleration, *in* 'Advances in Neural Information Processing Systems', pp. 1119–1129.

Renegar, J. and Grimmer, B. [2018], 'A simple nearly-optimal restart scheme for speeding-up first order methods', *arXiv preprint arXiv:1803.00151* .

Zhou, Z., Zhang, Q. and So, A. M.-C. [2015], l 1, p-norm regularization: error bounds and convergence rate analysis of first-order methods, *in* 'Proceedings of the 32nd International

Conference on International Conference on Machine Learning-Volume 37', JMLR. org, pp. 1501–1510.