Sharpness, Restart and Compressed Sensing Performance.

Alexandre d'Aspremont,
CNRS & D.I., Ecole normale supérieure.

With Vincent Roulet (U. Washington) and Nicolas Boumal (Princeton U.).
Support from ERC SIPA.

Introduction

Statistical performance vs. computational complexity.

- Clear empirical link between statistical performance and computational complexity.
- Quantities describing computational complexity lack statistical meaning.

Today: Two minor enigmas. . .

Introduction

"Fast solution of ℓ_1-norm minimization problems when the solution may be sparse" by [Donoho and Tsaig, 2008].

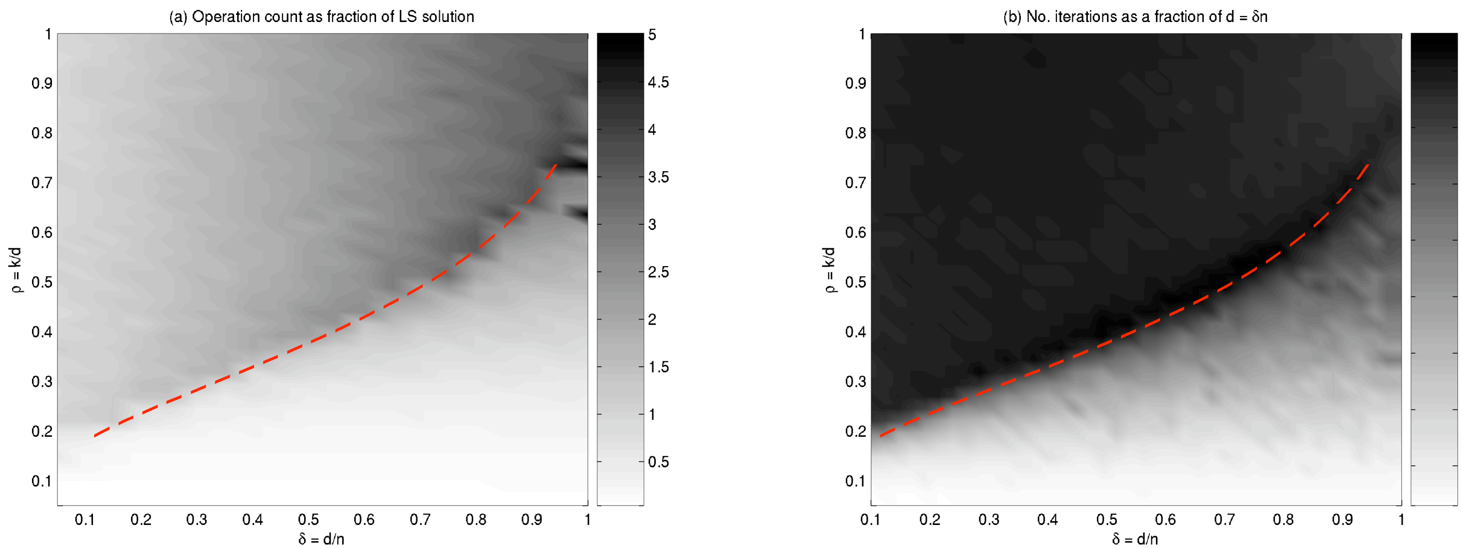


Figure 3: Computational Cost of Homotopy. Panel (a) shows the operation count as a fraction of one least-squares solution on a ρ-δ grid, with $n = 1000$. Panel (b) shows the number of iterations as a fraction of $d = \delta \cdot n$. The superimposed dashed curve depicts the curve ρ_W, below which Homotopy recovers the sparsest solution with high probability.

First enigma: Phase transition for computation and recovery match. . .

Introduction

"Templates for convex cone problems with applications to sparse signal recovery." (TFOCS) by [Becker, Candès, and Grant, 2011b].

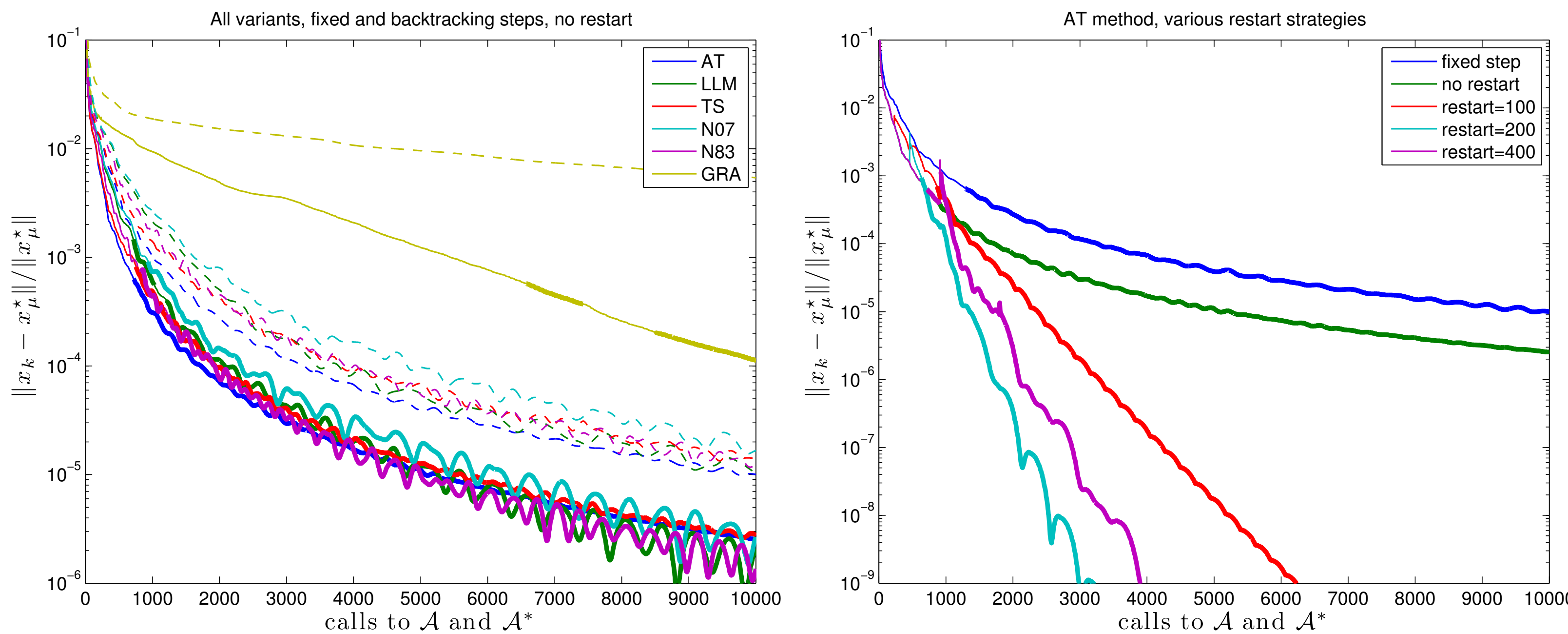


Figure 6: Comparing first order methods applied to a smoothed Dantzig selector model. Left: comparing all variants using a fixed step size (dashed lines) and backtracking line search (solid lines). Right: comparing various restart strategies using the AT method.

Second enigma: Restarting yields linear convergence. . .

Outline

Today.

- **Sharpness**
- Optimal restart schemes, adaptation
- Compressed Sensing Performance
- Numerical results

Sharpness

Consider

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in Q \end{array}$$

where $f(x)$ is a **convex** function, $Q \subset \mathbb{R}^n$.

- Assume ∇f **is Hölder continuous,**

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|^{s-1}, \quad \text{for every } x, y \in \mathbb{R}^n,$$

- Assume **sharpness**, i.e.

$$\mu d(x, X^*)^r \leq f(x) - f^*, \quad \text{for every } x \in K,$$

  where f^* is the minimum of f, $K \subset \mathbb{R}^n$ is a compact set, $d(x, X^*)$ the distance from x to the set $X^* \subset K$ of minimizers of f, and $r \geq 1$, $\mu > 0$ are constants.

Sharpness, Restart

Strong convexity is a particular case of sharpness.

$$\mu d(x, X^*)^2 \leq f(x) - f^*$$

If f is also **smooth**, an optimal algorithm (ignoring strong convexity), will produce a point x satisfying

$$f(x) - f^* \leq \frac{cL}{t^2} d(x_0, X^*)^2,$$

after t iterations.

- Restarting the algorithm, we thus get

$$f(x_{k+1}) - f^* \leq \frac{cL}{\mu t_k^2} \left(f(x_k) - f^*\right), \quad k = 1, \ldots, N$$

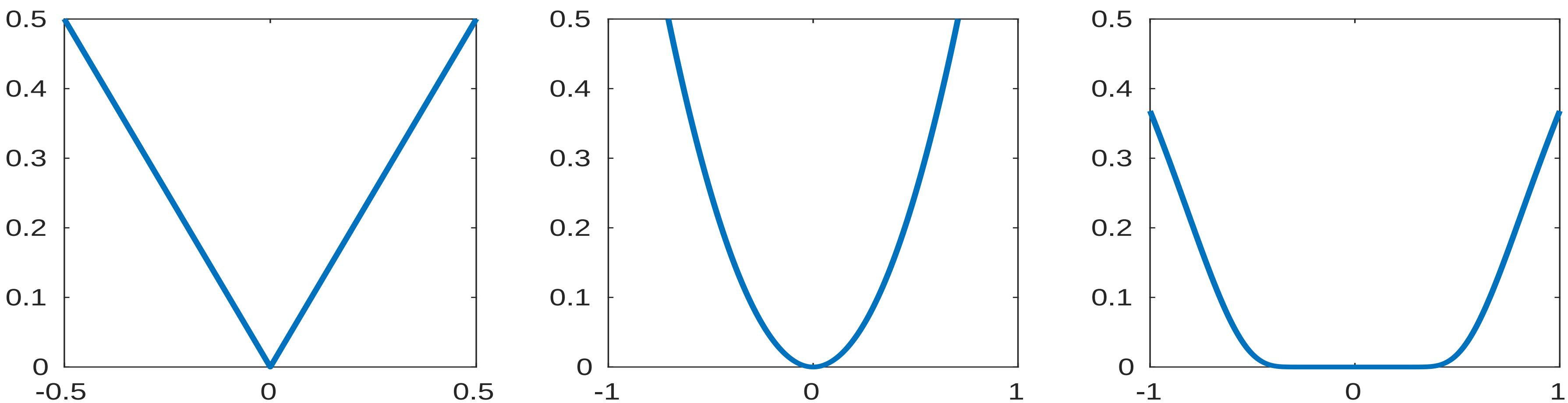
  at each outer iteration, after t_k inner iterations.
- Restart yields **linear convergence**, without explicitly modifying the algorithm.

Sharpness

Smoothness is classical [Nesterov, 1983, 2005], sharpness less so. . .

$$\mu d(x, X^*)^r \leq f(x) - f^*, \quad \text{for every } x \in K.$$

- Real analytic functions all satisfy this locally, a result known as Łojasiewicz's inequality [Lojasiewicz, 1963].
- Generalizes to a much wider class of non-smooth functions [Lojasiewicz, 1993, Bolte et al., 2007]
- Conditions of this form are also known as **sharp minimum**, **error bound**, etc. [Polyak, 1979, Burke and Ferris, 1993, Burke and Deng, 2002].



The functions $|x|$, x^2 and $\exp(-1/x^2)$.

Sharpness & Smoothness

- Gradient ∇f Hölder continuous ensures

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

an **upper bound** on suboptimality.

- If in addition f sharp on a set K with parameters (r, μ), we have

$$\frac{s\mu}{rL} \leq d(x, X^*)^{s-r}$$

hence $s \leq r$.

In the following, we write

$$\kappa \triangleq L^{\frac{2}{s}} / \mu^{\frac{2}{r}} \qquad \text{and} \qquad \tau \triangleq 1 - \frac{s}{r}$$

If $r = s = 2$, κ matches the classical condition number of the function.

Sharpness & Complexity

- Restart schemes were studied for strongly or uniformly convex functions [Nemirovskii and Nesterov, 1985, Nesterov, 2007, Iouditski and Nesterov, 2014, Lin and Xiao, 2014]
- In particular, Nemirovskii and Nesterov [1985] link sharpness with (optimal) faster convergence rates using restart schemes.
- Weaker versions of this strict minimum condition used more recently in restart schemes by [Renegar, 2014, Freund and Lu, 2015].
- Several heuristics [O'Donoghue and Candes, 2015, Su et al., 2014, Giselsson and Boyd, 2014] studied adaptive restart schemes to speed up convergence.
- The robustness of restart schemes was also studied by Fercoq and Qu [2016] in the strongly convex case.
- Sharpness used to prove linear converge matrix games by Gilpin et al. [2012].

Restart schemes

Algorithm 1 Scheduled restarts for smooth convex minimisation **(RESTART)**

Inputs : $x_0 \in \mathbb{R}^n$ and a sequence t_k for $k = 1, \ldots, R$.
for $k = 1, \ldots, R$ **do**

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

end for
Output : $\hat{x} := x_R$

Here, the number of inner iterations t_k satisfies

$$t_k = Ce^{\alpha k}, \quad k = 1, \ldots, R.$$

for some $C > 0$ and $\alpha \geq 0$ and will ensure

$$f(x_k) - f^* \leq \nu e^{-\gamma k}.$$

Proposition [Roulet and A., 2017]

Restart. *Let f be a smooth convex function with parameters $(2, L)$, sharp with parameters (r, μ) on a set K. Restart with iteration schedule $t_k = C^*_{\kappa,\tau} e^{\tau k}$, for $k = 1, \ldots, R$, where $C^*_{\kappa,\tau} \triangleq e^{1-\tau}(c\kappa)^{\frac{1}{2}}(f(x_0) - f^*)^{-\frac{\tau}{2}}$, with $c = 4e^{2/e}$ here. The precision reached at the last point $\hat{x}$ is given by,*

$$f(\hat{x}) - f^* \leq e^{-2e^{-1}(c\kappa)^{-\frac{1}{2}}N}(f(x_0) - f^*) = O\left(\exp(-\kappa^{-\frac{1}{2}}N)\right), \quad \textit{when } \tau = 0,$$

while,

$$f(\hat{x}) - f^* \leq \frac{f(x_0) - f^*}{\left(\tau e^{-1}(f(x_0) - f^*)^{\frac{\tau}{2}}(c\kappa)^{-\frac{1}{2}}N + 1\right)^{\frac{2}{\tau}}} = O\left(N^{-\frac{2}{\tau}}\right), \quad \textit{when } \tau > 0,$$

where $N = \sum_{k=1}^{R} t_k$ is the total number of iterations.

Adaptation

- The sharpness constant μ and exponent r in

$$\mu d(x, X^*)^r \leq f(x) - f^*, \quad \text{for every } x \in K.$$

  are of course **never observed.**

- Can we make restart schemes **adaptive?** Otherwise, sharpness is useless. . .

- Solves robustness problem for accelerated methods on strongly convex functions.

Adaptation

Proposition [Roulet and A., 2017]

Adaptation. *Assume* $N \geq 2C^*_{\kappa,\tau}$, *and if* $\frac{1}{N} > \tau > 0$, $C^*_{\kappa,\tau} > 1$.
If $\tau = 0$, *there exists* $i \in [1, \ldots, \lfloor \log_2 N \rfloor]$ *such that scheme* $\mathcal{S}_{i,0}$ *achieves a precision given by*

$$f(\hat{x}) - f^* \leq \exp\left(-e^{-1}(c\kappa)^{-\frac{1}{2}}N\right)(f(x_0) - f^*).$$

If $\tau > 0$, *there exist* $i \in [1, \ldots, \lfloor \log_2 N \rfloor]$ *and* $j \in [1, \ldots, \lceil \log_2 N \rceil]$ *such that scheme* $\mathcal{S}_{i,j}$ *achieves a precision given by*

$$f(\hat{x}) - f^* \leq \frac{f(x_0) - f^*}{\left(\tau e^{-1}(c\kappa)^{-\frac{1}{2}}(f(x_0) - f^*)^{\frac{\tau}{2}}(N-1)/4 + 1\right)^{\frac{2}{\tau}}}.$$

Overall, running the logarithmic grid search has a complexity $(\log_2 N)^2$ **times** higher than running N iterations using the optimal (oracle) scheme.

Hölder smooth case

The generic Hölder smooth case $s \neq 2$ is harder.

- When f is smooth with parameters (s, L) and $s \neq 2$, the restart scheme is more complex.
- The universal fast gradient method in [Nesterov, 2015], outputs after t iterations a point $x \triangleq \mathcal{U}(x_0, \epsilon, t)$, such that

$$f(x) - f^* \leq \frac{\epsilon}{2} + \left(\frac{cL^{\frac{2}{s}} d(x_0, X^*)^2}{\epsilon^{\frac{2}{s}} t^{\frac{2\rho}{s}}} \right) \frac{\epsilon}{2},$$

  where c is a constant ($c = 8$) and $\rho \triangleq \frac{3s}{2} - 1$ is the optimal rate of convergence for s-smooth functions.
- Contrary to the case $s = 2$ above, we need to schedule *both* the target accuracy ϵ_k used by the algorithm *and* the number of iterations t_k.
- We **lose adaptivity when $s \neq 2$.**

Outline

- Sharpness
- Optimal restart schemes, adaptation
- **Compressed Sensing Performance**
- Numerical results

Compressed Sensing

Sparse Recovery. Given $A \in \mathbb{R}^{n \times p}$ and observations $b = Ax^*$ on $x^* \in \mathbb{R}^p$, solve

$$\begin{array}{ll} \mbox{minimize} & \|x\|_1 \\ \mbox{subject to} & Ax = b \end{array} \qquad (\ell_1 \text{ recovery})$$

in the variable $x \in \mathbb{R}^p$.

Definition [Cohen et al., 2009]

Nullspace Property. *The matrix A satisfies the Null Space Property (NSP)* on support $S \subset [1, p]$ *with constant $\alpha \geq 1$ if for any $z \in Null(A) \setminus \{0\}$,*

$$\alpha \|z_S\|_1 < \|z_{S^c}\|_1. \qquad \text{(NSP)}$$

The matrix A satisfies the Null Space Property at order s *with constant $\alpha \geq 1$ if it satisfies it on every support S of cardinality at most s.*

NSP & Shaprness

Sharpness. Sharpness for ℓ_1-recovery of a sparse signals x^* means

$$\|x\|_1 - \|x^*\|_1 > \gamma \|x - x^*\|_1 \qquad \text{(Sharp)}$$

for any $x \neq x^*$ such that $Ax = b$, and some $0 \leq \gamma < 1$.

Proposition [Roulet, Boumal, and A., 2017]

NSP & Sharpness. *Given a coding matrix $A \in \mathbb{R}^{n \times p}$ satisfying* (NSP) *at order s with constant $\alpha \geq 1$, if the original signal x^* is s-sparse, then for any $x \in \mathbb{R}^p$ satisfying $Ax = b$, $x \neq x^*$, we have*

$$\|x\|_1 - \|x^*\|_1 > \frac{\alpha - 1}{\alpha + 1} \|x - x^*\|_1.$$

This implies signal recovery, i.e. optimality of x^* for (ℓ_1 recovery), *and the sharpness bound* (Sharp) *with $\gamma = (\alpha - 1)/(\alpha + 1)$.*

Computational Complexity

Restart scheme.

Restart Scheme **(Restart)**

Input: $y_0 \in \mathbb{R}^p$, gap $\epsilon_0 \geq \|y_0\|_1 - \|\hat{x}\|_1$, decreasing factor ρ, restart clock t

For $k = 1 \ldots, K$ compute

$$\epsilon_k = \rho \epsilon_{k-1}, \qquad y_k = \mathcal{A}(y_{k-1}, \epsilon_k, t) \quad \textbf{(NESTA)} \tag{Restart}$$

Output: A point $\hat{y} = y_K$ approximately solving (ℓ_1 recovery).

Restart NESTA by [Becker et al., 2011a] with geometrically increasing precision targets.

Computational Complexity

Proposition [Roulet, Boumal, and A., 2017]

Complexity. *Given a Gaussian design matrix* $A \in \mathbb{R}^{n\times p}$ *and a signal* x^* *with sparsity* $s < n/(c^2 \log p)$, *the optimal* (Restart) *scheme outputs a point* $\hat{y}$ *such that*

$$\|\hat{y}\|_1 - \|x^*\|_1 \leq \exp\left(-\left(1 - c\sqrt{\frac{s\log p}{n}}\right)\frac{e}{2\sqrt{p}}N\right)\epsilon_0,$$

with high probability, where $c > 0$ *is a universal constant and* N *is the total number of iterations.*

- The **iteration complexity** of solving the (ℓ_1 recovery) problem is controlled by the **oversampling ratio** n/s.
- Directly generalizes to other decomposable norms.
- Similar result involving Renegar's condition number and cone restricted eigenvalues.

Outline

- Sharpness
- Optimal restart schemes, adaptation
- Compressed Sensing Performance
- **Numerical results**

Numerical results

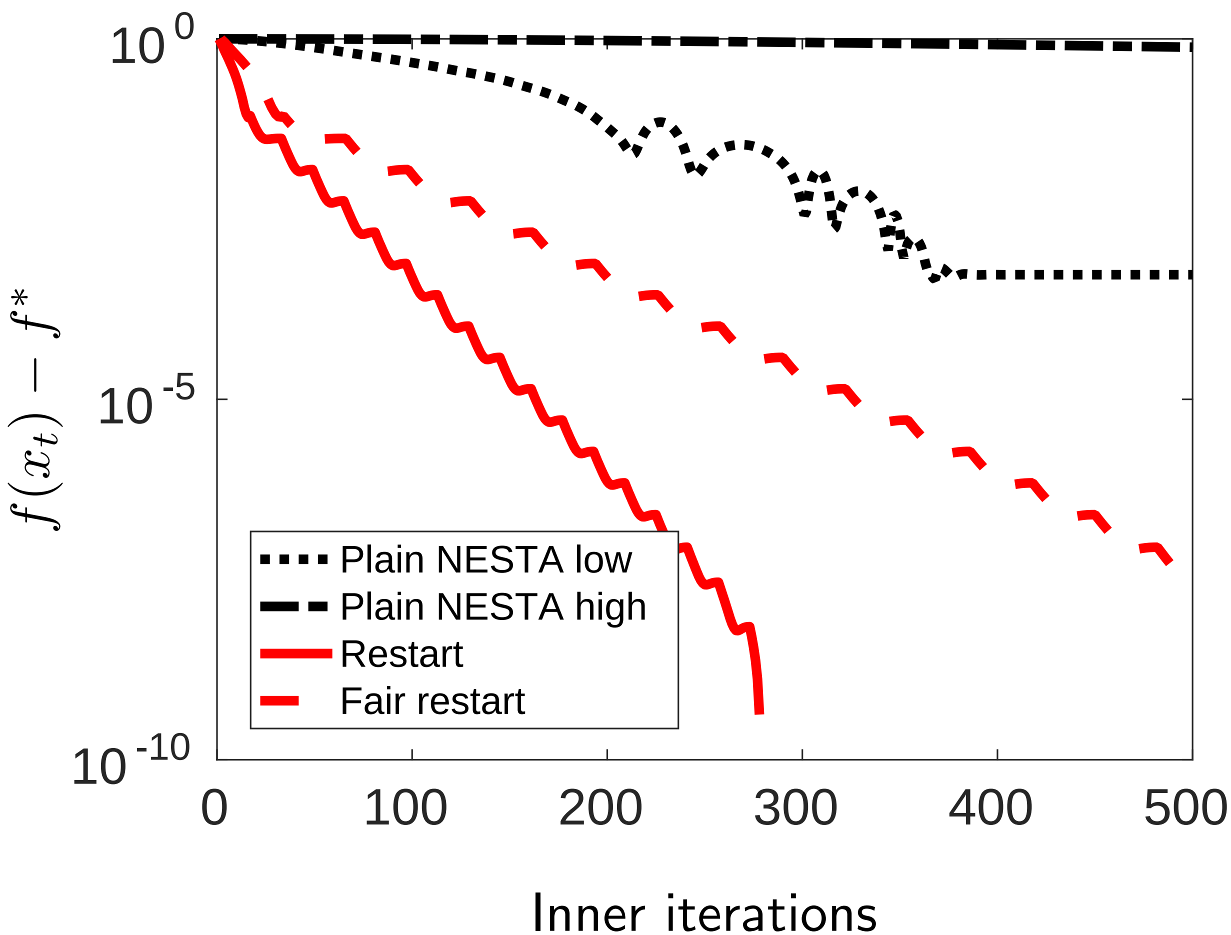
Sample ℓ_1-recovery problems. Solve

$$\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & Ax = b \end{array}$$

using NESTA and restart scheme.

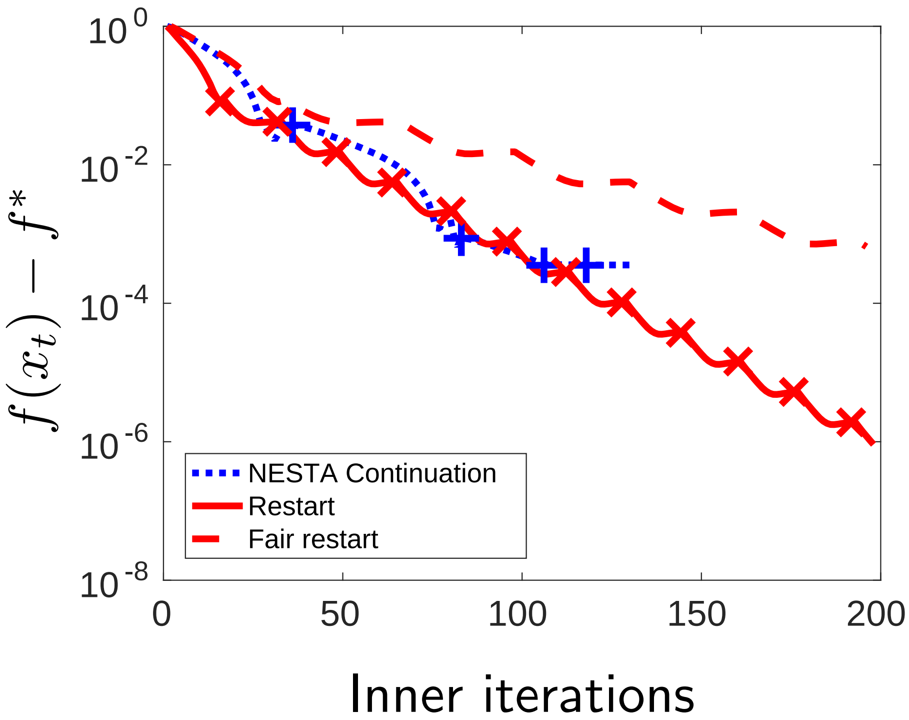
- Generate random design matrix $A \in \mathbb{R}^{n \times p}$ with i.i.d. Gaussian coefficients.
- Normalize A so that $AA^T = \mathbf{I}$ (to fit NESTA's format)
- Generate observations $b = Ax^*$ where $x^* \in \mathbb{R}^p$ is an s-sparse vector whose nonzero coefficients are all ones.

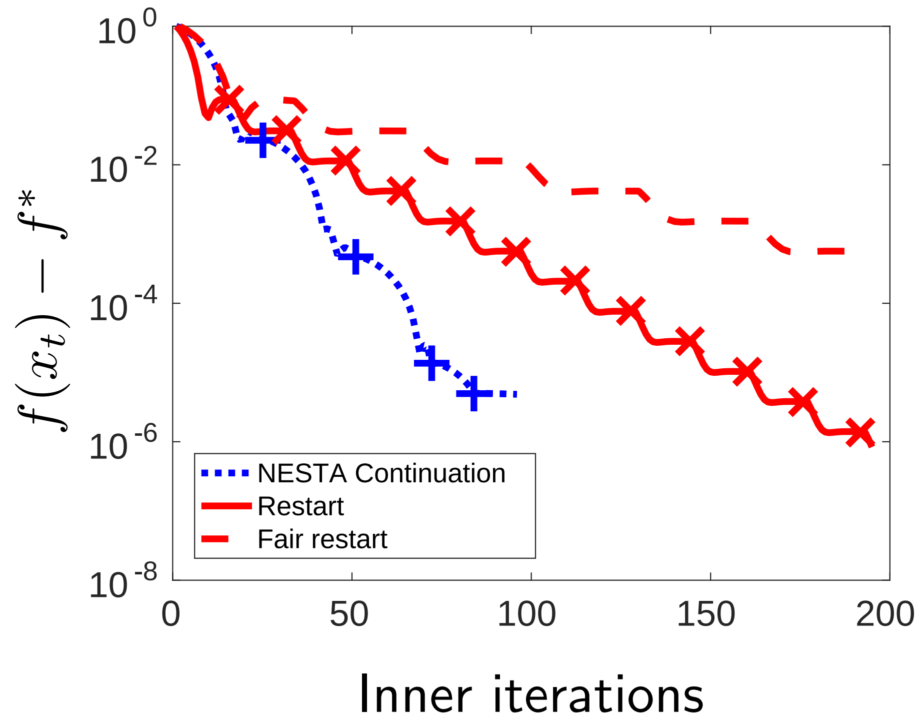
Numerical results



Best restarted NESTA **(solid red line)**, overall cost of the adaptive restart scheme **(dashed red line)** versus plain NESTA implementation with low accuracy $\epsilon = 10^{-1}$ **(dotted black line)**, and higher accuracy $\epsilon = 10^{-3}$ **(dash-dotted black line)**. Total budget of 500 iterations.

Numerical results

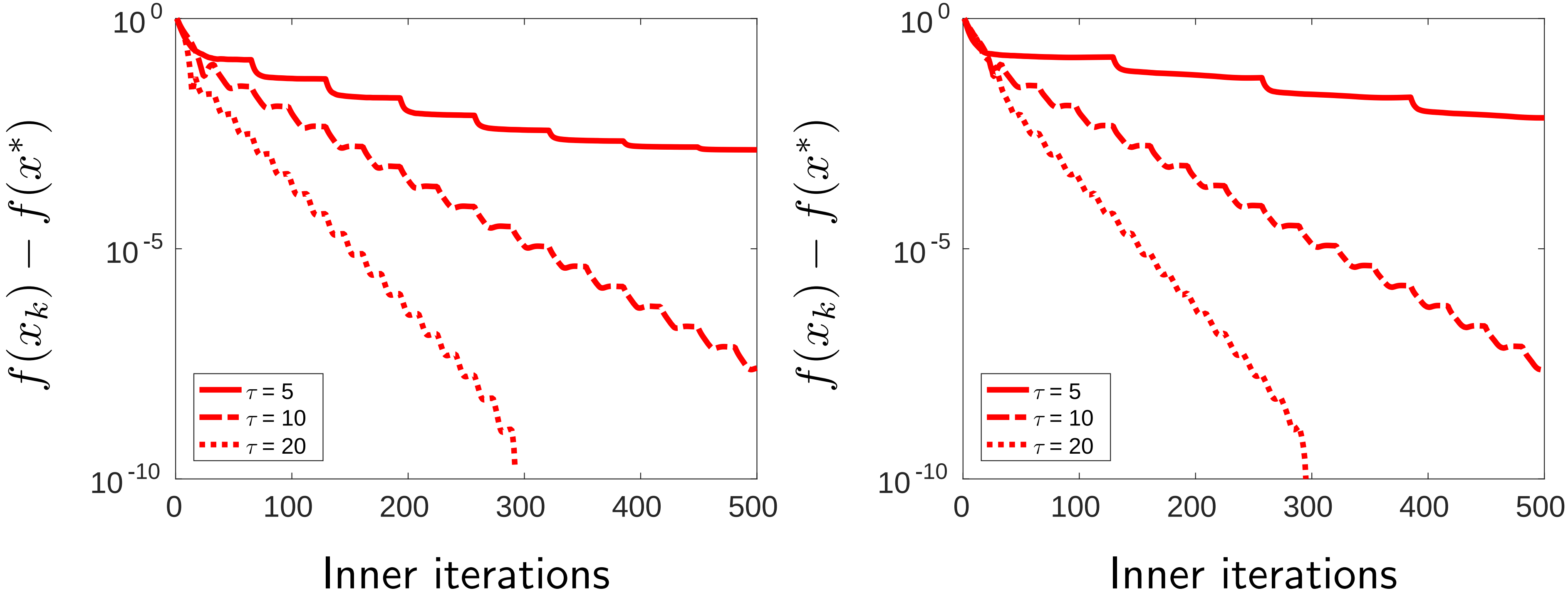




Best restarted NESTA **(solid red line)** and overall cost of the practical restart schemes **(dashed red line)** versus NESTA with 5 continuation/restart steps **(dotted blue line)** for a total budget of 500 iterations.

Crosses at restart occurrences. Left: $n = 200$. Right : $n = 300$.

Numerical results



Best restart scheme found by grid search for increasing values of the oversampling ratio $\tau = n/s$.

Left: Constant sparsity $s = 20$. Right: constant number of samples $n = 200$.

Numerical results

Number of samples n	100	200	400
Time in seconds for $f(x_t) - f^* < 10^{-2}$	$5.07 \cdot 10^{-2}$	$3.07 \cdot 10^{-2}$	$1.66 \cdot 10^{-2}$

Time to achieve $\epsilon = 10^{-2}$ by the best restart scheme for increasing number of samples n

More data less work (ignoring cost of adaptation).

Conclusion

- Sharpness holds generically.
- Restarting then accelerates convergence, cost of adaptation is marginal.
- Shows better conditioned recovery problems are faster to solve.

Open problems.

- Adaptation in generic Hölder gradient case.
- Optimal algorithm for sharp problems without restart.
- Local adaptation to sharpness.

*

References

Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. Nesta: a fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011a.

Stephen R Becker, Emmanuel J Candès, and Michael C Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011b.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

James Burke and Sien Deng. Weak sharp minima revisited part i: basic theory. *Control and Cybernetics*, 31:439–469, 2002.

JV Burke and Michael C Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5): 1340–1359, 1993.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the AMS*, 22(1):211–231, 2009.

David L Donoho and Yaakov Tsaig. Fast solution of ℓ_1-norm minimization problems when the solution may be sparse. *Information Theory, IEEE Transactions on*, 54(11):4789–4812, 2008.

Olivier Fercoq and Zheng Qu. Restarting accelerated gradient methods with a rough strong convexity estimate. *arXiv preprint arXiv:1609.07358*, 2016.

Robert M Freund and Haihao Lu. New computational guarantees for solving convex optimization problems with first order methods, via a function growth condition measure. *arXiv preprint arXiv:1511.02974*, 2015.

Andrew Gilpin, Javier Pena, and Tuomas Sandholm. First-order algorithm with $\mathcal{O}(\log 1/\epsilon)$ convergence for ϵ-equilibrium in two-person zero-sum games. *Mathematical programming*, 133(1-2):279–298, 2012.

Pontus Giselsson and Stephen Boyd. Monotonicity and restart in fast gradient methods. In *53rd IEEE Conference on Decision and Control*, pages 5058–5063. IEEE, 2014.

Anatoli Iouditski and Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792*, 2014.

Qihang Lin and Lin Xiao. An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. In *ICML*, pages 73–81, 2014.

Stanislas Lojasiewicz. Sur la géométrie semi-et sous-analytique. *Annales de l'institut Fourier*, 43(5):1575–1595, 1993.

Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, pages 87–89, 1963.

AS Nemirovskii and Yu E Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2): 372–376, 1983.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Y. Nesterov. Gradient methods for minimizing composite objective function. *CORE DP2007/96*, 2007.

Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

Brendan O'Donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

BT Polyak. Sharp minima institute of control sciences lecture notes, moscow, ussr, 1979. In *IIASA workshop on generalized Lagrangians and their applications, IIASA, Laxenburg, Austria*, 1979.

James Renegar. Efficient first-order methods for linear programming and semidefinite programming. *arXiv preprint arXiv:1409.5832*, 2014.

Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. *ArXiv preprint arXiv:1702.03828*, 2017.

Vincent Roulet, Nicolas Boumal, and Alexandre d'Aspremont. Computational complexity versus statistical performance on sparse recovery problems. *arXiv preprint arXiv:1506.03295*, 2017.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.