Sharpness, Restart & Acceleration.

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

With Vincent Roulet. Support from ERC SIPA.

Introduction

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.

Introduction, 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

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

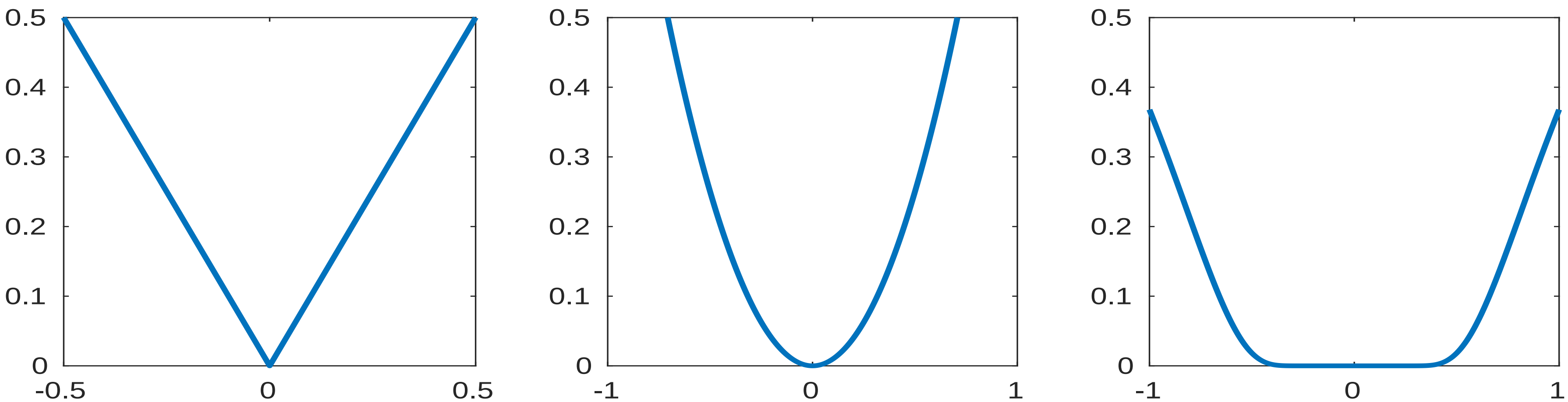
  at each outer iteration, after t_k inner iterations.
- Restart yields **linear convergence**, without explicitly modifying the algorithm.

Introduction, 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 $\exp(-1/x^2)$.

Introduction, 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.

Introduction, 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].
- Sharpness was also used to characterize the convergence of alternating and splitting methods [Attouch et al., 2010, Frankel et al., 2014]
- 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].

Introduction, Adaptation

Today.

- 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. . .
- Solve robustness problem for accelerated methods on strongly convex functions.
- What happens when we have an explicit termination criterion?

Outline

Today.

- **Sharpness & optimal restart schemes**
- Adaptation
- Restart with termination criterion
- Composite and constrained problems
- Numerical results

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 d'Aspremont, 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

Adaptation. When $s = 2$, a log-scale grid search on τ and κ works.

Run several schemes with a fixed number of inner iterations N.

$$\begin{cases} \mathcal{S}_{i,0} : \text{Restart scheme with } t_k = C_i, \\ \mathcal{S}_{i,j} : \text{Restart scheme with } t_k = C_i e^{\tau_j k}, \end{cases}$$

where $C_i = 2^i$ and $\tau_j = 2^{-j}$.

Adaptation

Proposition [Roulet and d'Aspremont, 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.

Adaptation

Proof sketch. Need to show robustness w.r.t. τ.

Split in two regimes.

- If $\frac{1}{N} \leq \tau$, show that we only lose a constant factor with respect to the polynomial bound.
- If $\frac{1}{N} > \tau > 0$, show that we are a constant factor away from linear convergence bound.

Accelerated algorithms are much less robust to strong convexity parameter.

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 $s \neq 2$.**

Restart with criterion

Termination criterion. We stop the algorithm after t_ϵ inner iterations, using a termination criterion to ensure $x = \mathcal{U}(x_0, \epsilon, t_\epsilon)$ satisfies $f(x) - f^* \leq \epsilon$, and write

$$x \triangleq \mathcal{C}(x_0, \epsilon).$$

Algorithm 2 Restart on criterion **(ε-RESTART)**

Inputs : $x_0 \in \mathbb{R}^n, f^*, \gamma \geq 0, \epsilon_0 = f(x_0) - f^*$
for $k = 1, \ldots, R$ **do**

$$\epsilon_k := e^{-\gamma}\epsilon_{k-1}, \qquad x_k := \mathcal{C}(x_{k-1}, \epsilon_k)$$

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

[Roulet and d'Aspremont, 2017]: very robust in γ.

- Given $\rho = \frac{3s}{2} - 1$, the algorithm automatically adapts to the optimal values of the sharpness parameters (r, μ).
- If ρ is not known, we lose a factor $e/2$ at worst.

Composite and constrained problems

Composite and constrained problems. Consider

$$\text{minimize } f(x) \triangleq \phi(x) + g(x), \qquad \text{(Composite)}$$

- **Prox function** h with $\mathbf{dom}(f) \subset \mathbf{dom}(h)$, strongly convex with respect to the norm $\|\cdot\|$ with convexity parameter equal to one. We define the Bregman divergence associated to h as

$$D_h(y,x) = h(y) - h(x) - \nabla h(x)^T(y-x), \quad \text{for } x, y \in \mathbf{dom}(h).$$

so that $D_h(y,x) \geq \frac{1}{2}\|x-y\|^2$.

- Given $x, y \in \mathbf{dom}(f)$ and $\lambda \geq 0$ we assume that

$$\min_z \left\{ y^T z + g(z) + \lambda D_h(z,x) \right\}$$

can be solved either in closed form or by some fast computational procedure.

Composite and constrained problems

Definition [Roulet and d'Aspremont, 2017]

Relative sharpness. *A convex function f is called relatively sharp with respect to a strictly convex function h on a set $K \subset \mathbf{dom}(f)$ iff there exist $r \geq 1$, $\mu > 0$ such that*

$$\frac{\mu}{r} D_h(x, X^*)^{\frac{r}{2}} \leq f(x) - f^* \quad \text{for any } x \in K \qquad \text{(Relative Sharpness)}$$

where $D_h(x, X^*) = \min_{x^* \in X^*} D_h(x, x^*)$ and D_h is the Bregman divergence associated to h.

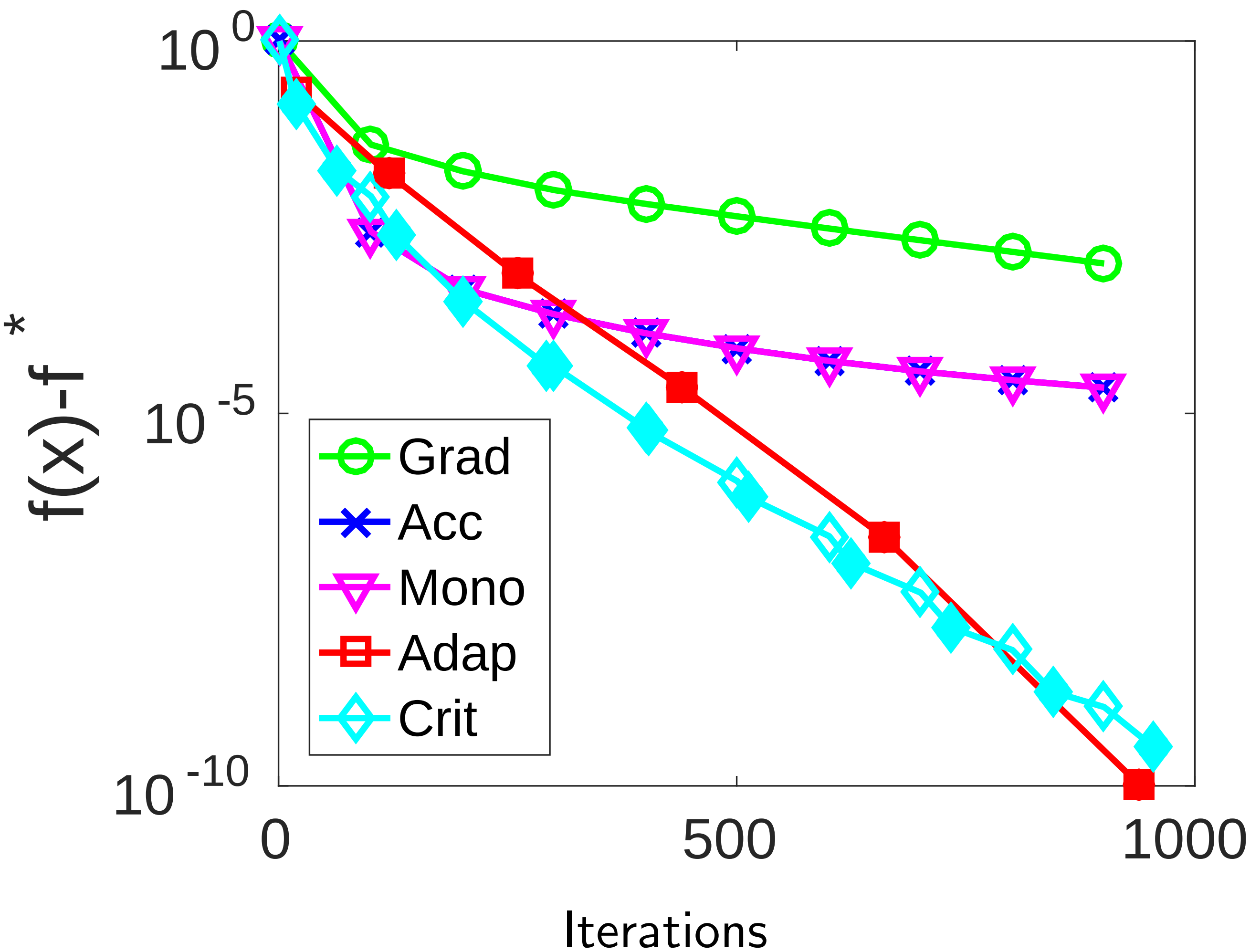
- In the spirit of relative-smoothness [Bauschke et al., 2016, Lu et al., 2016].
- **As generic as sharpness:** Satisfied if f and h are subanalytic [Bierstone and Milman, 1988, Th. 6.4].
- All previous results transpose directly to this setting, using the complexity bound

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

Outline

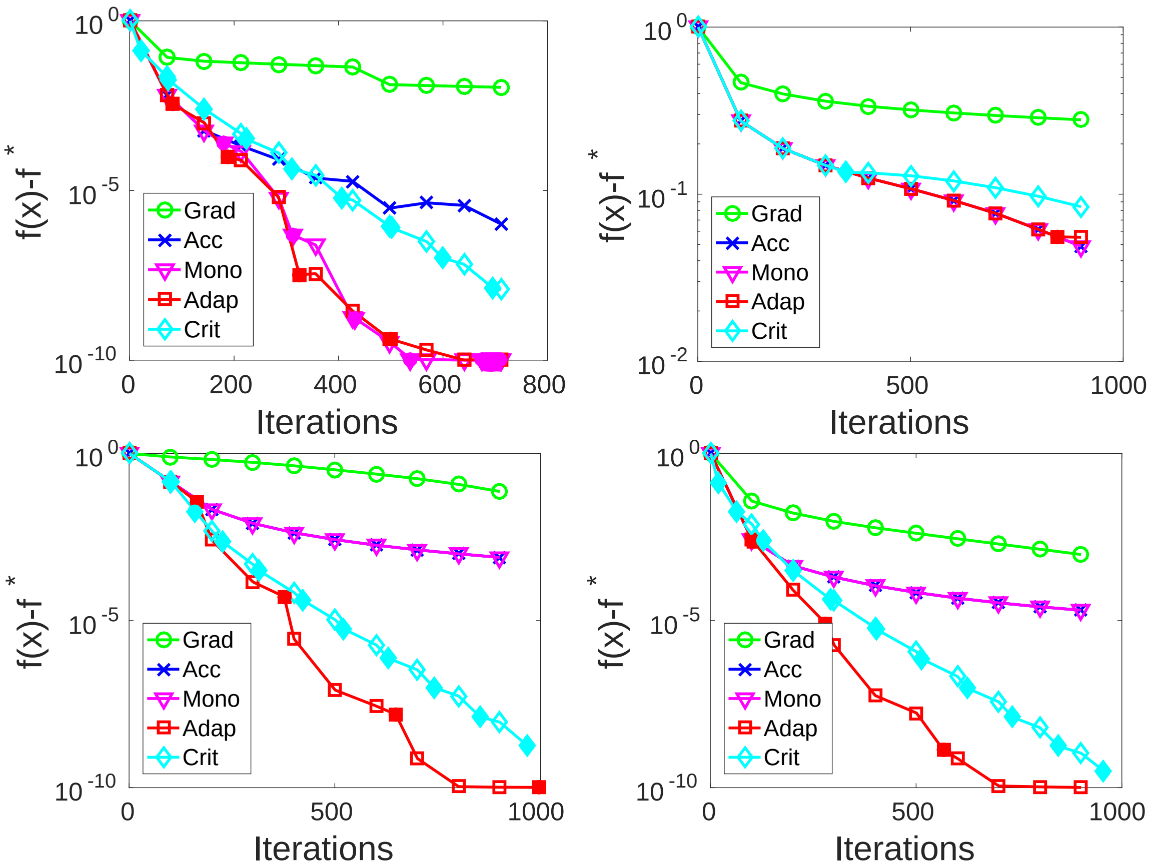
- Sharpness & optimal restart schemes
- Adaptation
- Restart with termination criterion
- Composite and constrained problems
- **Numerical results**

Numerical results



Comparison of the methods for the LASSO problem on the Sonar dataset where number of iterations of the Adaptive method is multiplied by the size of the grid. Large dots represent the restart iterations. Grid search size is set to 4.

Numerical results



Sonar data set. From top to bottom and left to right: least square loss, logistic loss, dual SVM problem and LASSO. We use adaptive restarts (Adap), gradient descent (Grad), accelerated gradient (Acc) and restart heuristic enforcing monotonicity (Mono). Large dots represent the restart iterations.

Conclusion

- Restart performance directly linked to sharpness.
- Restarting almost always works.
- In practice, testing a few schemes is enough to guarantee optimal complexity.

Open problems.

- Adaptation in generic Hölder gradient case.
- Optimal bounds for sharp problems without restart.
- Equivalently, local adaptation to sharpness.

*

References

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

Edward Bierstone and Pierre D Milman. Semianalytic and subanalytic sets. *Publications Mathématiques de l'IHÉS*, 67:5–42, 1988.

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.

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

Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for kl functions. *arXiv preprint arXiv:1405.1357*, 2014.

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.

H. Lu, R. M. Freund, and Y. Nesterov. Relatively-Smooth Convex Optimization by First-Order Methods, and Applications. *ArXiv e-prints*, October 2016.

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.

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.