Restarting Frank-Wolfe.

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

With Thomas Kerdreux (ENS) and Sebastian Pokutta (Georgia Tech.)

Jobs

Postdoc position in **optimization/ML**.

At INRIA - Ecole Normale Supérieure in Paris.

Introduction

Frank-Wolfe. Classical first order methods for solving

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

in $x \in \mathbb{R}^n$, with $C \subset \mathbb{R}^n$ convex.

Assumes that the **linear minimization oracle**

$$\begin{array}{ll}\text{minimize} & d^T x\\ \text{subject to} & x \in C\end{array}$$

can be solved efficiently for any $d \in \mathbb{R}^n$.

Franke-Wolfe

Algorithm 1 Franke-Wolfe **(FW)**

```
Inputs: $x_0 \in C$.
for $k = 1, \ldots, k^{max}$ do
    Solve the linear minimization oracle
        $x_d := \text{argmin} \quad x^T \nabla f(y_k)$
               $\text{subject to} \quad x \in C$
    Update the current point
        $x_{k+1} = x_k + \frac{2}{k+2}(x_d - x_k)$
end for
Output: approximate solution $\hat{x}$
```

Note that all iterates are feasible.

Franke-Wolfe

Complexity.

- Assume that f is differentiable. Define the curvature C_f of the function $f(x)$ as
$$C_f \triangleq \sup_{\substack{s,x\in\mathcal{M},\ \alpha\in[0,1],\\ y=x+\alpha(s-x)}} \frac{1}{\alpha^2}(f(y)-f(x)-\langle y-x,\nabla f(x)\rangle).$$
- The basic Frank-Wolfe algorithm will then produce an ϵ solution after at most
$$N_{\max}=\frac{4C_f}{\epsilon}$$
iterations.

Franke-Wolfe

Stopping criterion. At each iteration, we get a lower bound on the optimum as a byproduct of the affine minimization step.

- If x_d minimizes $\nabla f(x_k)^T x_d$ over C, we have by convexity

$$f(x_k) + \nabla f(x_k)^T (x_d - x_k) \leq f(x), \quad \text{for all } x \in C$$

- Calling f^* the optimal value of problem, we then get

$$f(x_k) - f^* \leq \nabla f(x_k)^T (x_k - x_d).$$

This allows us to bound the suboptimality of iterate at no additional cost.

Franke-Wolfe

Machine Learning Applications. See [Jaggi, 2013].

- When C is an atomic norm ball, each vertex is an atom and FW naturally produces "sparse" solutions.
- Linear minimization oracle is often easy to solve.
  - Complexity $O(n)$ for $\|\cdot\|_q$-balls.
  - Just an SVD for classical matrix norms (matrix completion, etc.)
  - Also works for structured atomic norms.
  - Idem for structured prediction [Lacoste-Julien et al., 2012].
- For some combinatorial polytopes with an exponential number of vertices, the linear minimization oracle is tractable, while projection is hard.

A Faster Franke-Wolfe

Faster convergence.

- **Linear convergence** with away steps when the optimum is inside the set [Guélat and Marcotte, 1986].
- **Linear convergence** for away step variants when the function is strongly convex [Garber and Hazan, 2013, Lacoste-Julien and Jaggi, 2015].
- Various **extensions** further improved upon these results for special cases, e.g. [Lacoste-Julien et al., 2013, Freund and Grigas, 2016, Garber and Meshi, 2016, Braun et al., 2017, Lan et al., 2017, Bashiri and Zhang, 2017, Garber et al., 2018, Kerdreux et al., 2018b, Braun et al., 2018],
- See also Joulin et al. [2014], Shah et al. [2015], Osokin et al. [2016], Freund et al. [2017], Miech et al. [2017] for applications of Frank-Wolfe to **machine learning problems**.

A Faster Franke-Wolfe

Today.

- Restarting accelerated gradient methods gives significantly improved performance.
- Complexity gains controlled by the “sharpness” of the optimum.
- Can we do the same for Frank-Wolfe?

Introduction

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

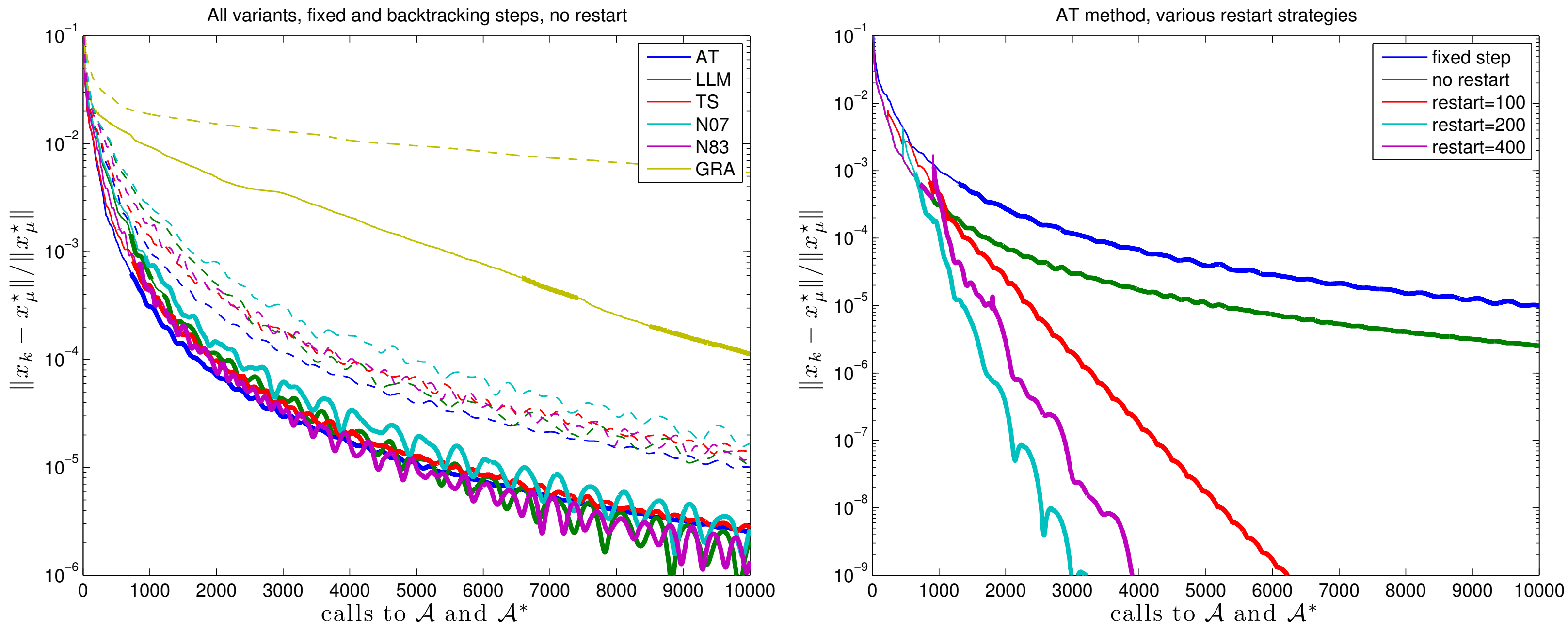


Figure 6: Comparing first order methods applied to a smoothed Dantzig selector model. Left: comparing all variants using a fixed step size (dashed lines) and backtracking line search (solid lines). Right: comparing various restart strategies using the AT method.

Restarting fast gradient methods yields linear convergence. . .

Outline

Today.

- Introduction
- **Sharpness & Łojasiewicz's growth condition**
- Optimal restart schemes for gradient methods
- Restarting Frank-Wolfe
- 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. the following local growth condition holds

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

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 gradient method (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

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

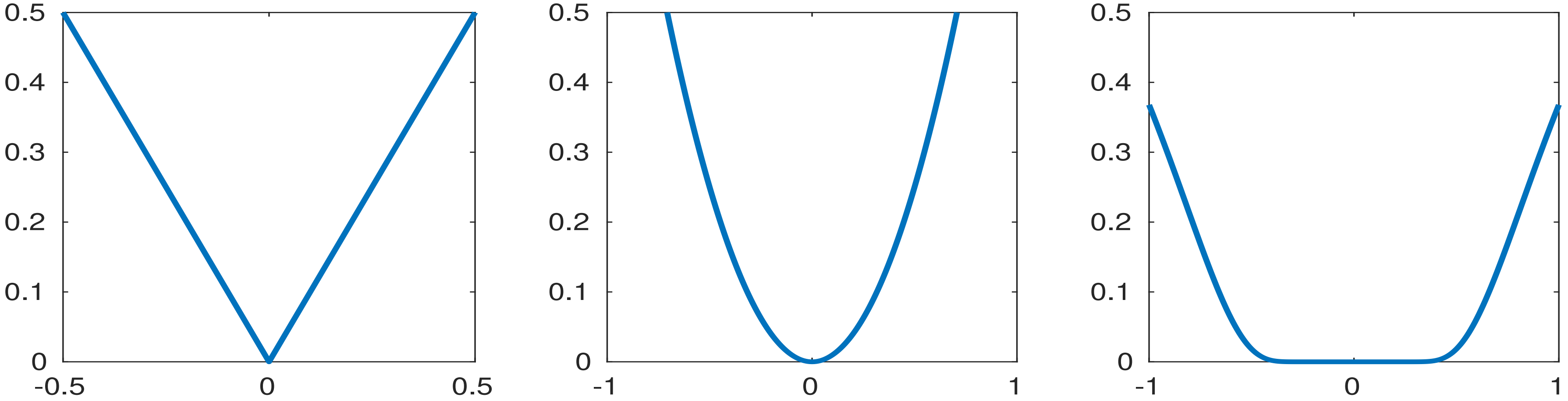
  at each outer iteration, after t_k inner iterations.
- Restart proves **linear convergence**.

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**, **Hölderian 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.

Outline

- Introduction
- Sharpness & Łojasiewicz's growth condition
- **Optimal restart schemes for gradient methods**
- Restarting Frank-Wolfe
- Numerical results

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 2 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

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

Restart schemes

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,

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

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. . .

- **Yes:** Grid of size $(\log_2 N)^2$ on restart parameters suffices. Fully adaptive if primal gap is known. See [Roulet and A., 2017].
- Can we prove something similar for Frank-Wolfe?

Outline

- Introduction
- Sharpness & Łojasiewicz's growth condition
- Optimal restart schemes for gradient methods
- **Restarting Frank-Wolfe**
- Numerical results

Restarting Frank-Wolfe

Strong Wolfe gap.

- Let f be a smooth convex function, $\mathcal{C}$ a polytope and let $x \in \mathcal{C}$ be arbitrary. Then the **strong Wolfe gap** $w(x)$ over $\mathcal{C}$ is defined as

$$w(x) \triangleq \left(\min_{S \in \mathcal{S}_x} \max_{y \in S, z \in \mathcal{C}} \nabla f(x)^T (y - z) \right)_+ \tag{1}$$

  where $x \in \mathbf{Co}(S)$ and $S_x = \{S \mid S \subset \mathbf{Ext}(\mathcal{C}), x \in \mathbf{Co}(S), |S| \text{ finite}\}$.

- We also write

$$w(x, S) \triangleq \left(\max_{y \in S, z \in \mathcal{C}} \nabla f(x)^T (y - z) \right)_+ \tag{2}$$

  given $S \in \mathcal{S}_x$.

Gap: $w(x)$ and $w(x, S)$ equal zero if and only if x is an optimal solution.

Restarting Frank-Wolfe

Definition [Kerdreux, A., and Pokutta, 2018a]

Strong Wolfe Primal Bound. *For any compact subset K of $\mathcal{C}$ such that $x^* \in K$, there are $\mu > 0$ and $r > 0$ such that*

$$f(x) - f^\star \leq \mu w(x)^r, \quad \textit{for } x \in K \tag{3}$$

[Lacoste-Julien and Jaggi, 2015, Theorem 6] shows $r = 2$ when f is strongly convex and $\mathcal{C}$ is a polytope.

Restarting Frank-Wolfe

Definition

Scaling inequality. *For all $x \in \mathcal{C} \setminus X^*$ and all differentiable convex function f,*

$$w(x) \geq \delta(\mathcal{C}) \max_{x^* \in X^*} \langle \nabla f(x); \frac{x - x^*}{||x - x^*||} \rangle. \qquad \text{(Scaling)}$$

- A convex polytope satisfies the δ-scaling inequality with $\delta(\mathcal{C}) = PWidth(\mathcal{C})$ [Lacoste-Julien and Jaggi, 2015].
- Łojasiewicz's factorization lemma then shows that the **strong Wolfe primal bound holds** when the scaling inequality holds.

Restarting Frank-Wolfe

Fractional Away-Step Frank-Wolfe Algorithm.

```
Given a smooth convex function $f$ with curvature $C_f^{\mathcal{A}}$. Starting point $x_0 = \sum_{v\in\mathcal{S}'} \alpha_0^v v \in \mathcal{C}$ with support $\mathcal{S}_0 \subset \mathbf{Ext}(\mathcal{C})$ and schedule parameter $\gamma > 0$.
Set $t := 0$
while $w(x_t, \mathcal{S}_t) > e^{-\gamma} w(x_0, \mathcal{S}_0)$ do
    $v_t := LP_{\mathcal{C}}(\nabla f(x_t))$ and $d_t^{FW} \triangleq v_t - x_t$
    $s_t := LP_{S_t}(-\nabla f(x_t))$ with $S_t$ current active set and $d_t^{Away} \triangleq x_t - s_t$
    if $-\nabla f(x_t)^T d_t^{FW} > e^{-\gamma} w(x_0, \mathcal{S}_0)/2$ then
        $d_t := d_t^{FW}$ with $\eta_{\text{max}} = 1$
    else
        $d_t := d_t^{Away}$ with $\eta_{\text{max}} = \frac{\alpha_t^{s_t}}{1-\alpha_t^{s_t}}$
    end if
    $x_{t+1} := x_t + \eta_t d_t$ with $\eta_t \in [0, \eta_{\text{max}}]$ via line-search
    Update active set $\mathcal{S}_{t+1}$ and coefficients $\{\alpha_{t+1}^v\}_{v\in\mathcal{S}_{t+1}}$
    $t := t + 1$
end while
Output: $x_t \in \mathcal{C}$ such that $w(x_t, \mathcal{S}_t) \leq e^{-\gamma} w(x_0, \mathcal{S}_0)$
```

Restarting Frank-Wolfe

Proposition [Kerdreux, A., and Pokutta, 2018a]

FAFW convergence. *Let f be a globally subanalytic, smooth convex function with away curvature C_f^A, satisfying the strong Wolfe primal bound on a compact set K for some $r \geq 1$ and $\mu > 0$. Let $\gamma > 0$ and assume $x_0 \in K$ is such that $e^{-\gamma} w(x_0)/2 \leq C_f^A$. The algorithm above outputs $x_T \in K$ such that*

$$w(x_T, \mathcal{S}_T) \leq w(x_0, \mathcal{S}_0) e^{-\gamma}$$

after at most

$$T \leq |\mathcal{S}_0| - |\mathcal{S}_T| + 16 e^{2\gamma} C_f^A \mu w(x_0, \mathcal{S}_0)^{r-2}$$

iterations, where $\mathcal{S}_0$ and $\mathcal{S}_T$ are the supports of respectively x_0 and x_T.

Restarting Frank-Wolfe

Proposition [Kerdreux, A., and Pokutta, 2018a]

FAFW with restarts. *Let f be a globally subanalytic, smooth convex function with away curvature C_f^A, satisfying the strong Wolfe primal bound on a compact set K with $r \geq 1$ and $\mu > 0$. Let $\gamma > 0$ and assume $x_0 \in K$ is such that $e^{-\gamma} w(x_0, \mathcal{S}_0)/2 \leq C_f^A$. With $\gamma_k = \gamma$, the output of FAFW with restarts satisfies*

$$\begin{cases} f(x_T) - f(x^\star) \leq w_0 \dfrac{1}{\left(1 + \tilde{T} C_\gamma^r\right)^{\frac{1}{2-r}}} & \text{when } 1 \leq r < 2 \\[2ex] f(x_T) - f(x^\star) \leq w_0 \exp\left(-\dfrac{\gamma}{e^{2\gamma}} \dfrac{\tilde{T}}{8 C_f^A \mu}\right) & \text{when } r = 2\ , \end{cases} \tag{4}$$

with $w_0 = w(x_0, \mathcal{S}_0)$ and $\tilde{T} \triangleq T - (|\mathcal{S}_0| - |\mathcal{S}_T|)$, where $C_\gamma^r \triangleq \frac{e^{\gamma(2-r)} - 1}{8 e^{2\gamma} C_f^A \mu w(x_0, \mathcal{S}_0)^{r-2}}$.

Restarting Frank-Wolfe

Proposition [Kerdreux, A., and Pokutta, 2018a]

Robustness in γ. *Suppose f satisfies strong Wolfe primal bound for $r > 0$ and write $\gamma^*(r)$ the optimal choice of $\gamma > 0$ in the complexity bound. Running FAFW with $\gamma = 1/2$ yields $\hat{x}$ satisfying*

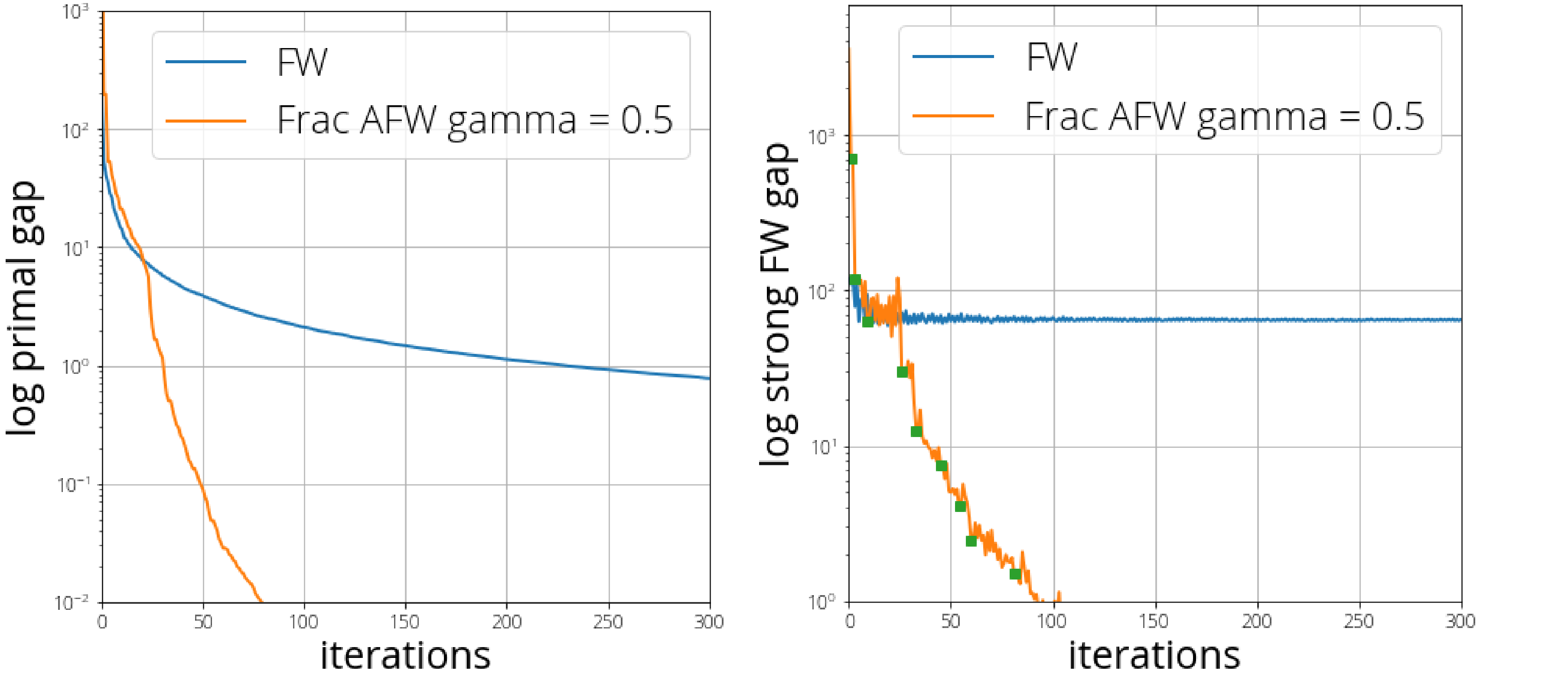
$$f(\hat{x}) - f^* \leq \sqrt{\frac{e}{e-1}} \frac{w(x_0, \mathcal{S}_0)}{\left(1 + \tilde{T} C^r_{\gamma^*(r)}\right)^{\frac{1}{2-r}}} \quad \text{when } 0 \leq r < 2 \; . \tag{5}$$

When $r = 2$, we have $\gamma^*(r) = 1/2$.

Outline

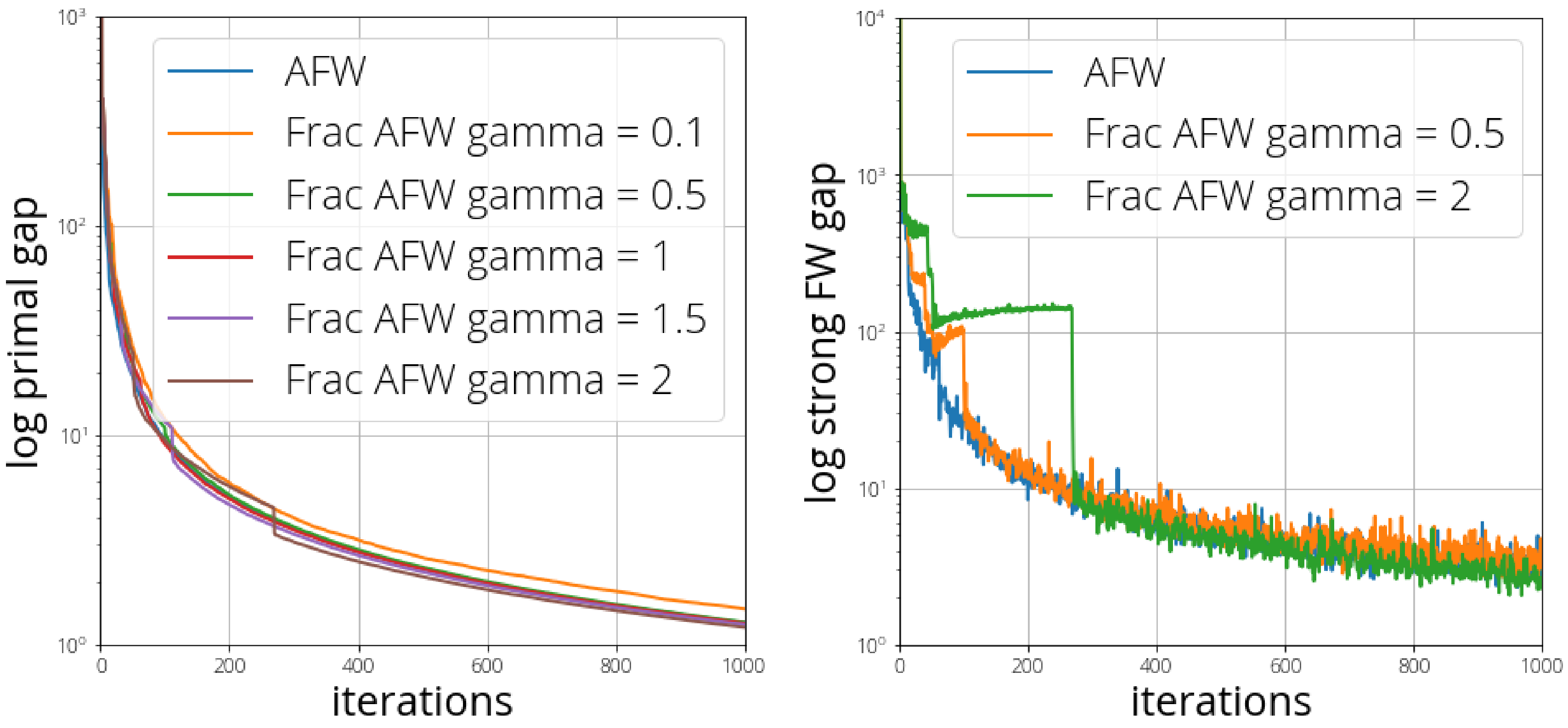
- Introduction
- Sharpness & Łojasiewicz's growth condition
- Optimal restart schemes for gradient methods
- Restarting Frank-Wolfe
- **Numerical results**

Numerical Results



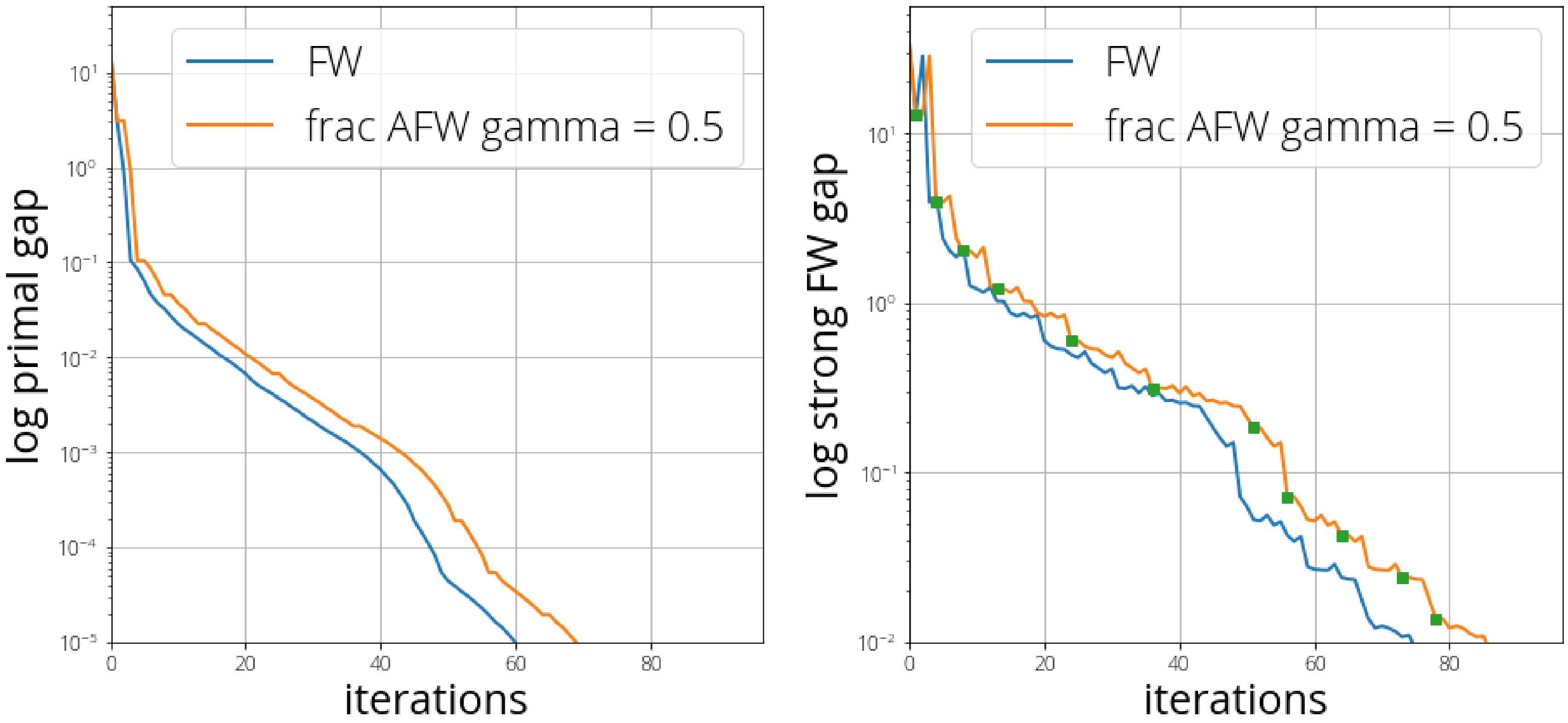
Comparing classical FW and FAFW with $\gamma = 0.5$ on a regression problem with loss power $\alpha = 1.5$, so that the classical geometric strong convexity condition does not hold. Green squares indicate restart times.

Numerical Results



Representative examples on Lasso with various values of γ in restart schemes of algorithm FAFW.

Numerical Results



Comparing classical FW and FAFW with $\gamma = 0.5$ on logistic regression with ℓ_1 constraint, where the constrained minimum lies in the interior of the ball. Here AFW and FW share the very same curve.

Conclusion

- Restarting Frank-Wolfe yields generically faster rates.
- Performance gains controlled by sharpness.
- Restart scheme is robust to growth condition/sharpness parameters.

Open problems.

- Fully adaptive bounds? Restart means we're missing something in the FAFW convergence proof. . .

*

References

Mohammad Ali Bashiri and Xinhua Zhang. Decomposition-invariant conditional gradient for general polytopes with line search. In *Advances in Neural Information Processing Systems*, pages 2687–2697, 2017.

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, 2011.

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.

Gábor Braun, Sebastian Pokutta, and Daniel Zink. Lazifying conditional gradient algorithms. *Proceedings of ICML*, 2017.

Gábor Braun, Sebastian Pokutta, Dan Tu, and Stephen Wright. Blended conditional gradients: the unconditioning of conditional gradients. *arXiv preprint arXiv:1805.07311*, 2018.

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.

Robert M. Freund and Paul Grigas. New analysis and results for the frank–wolfe method. *Mathematical Programming*, 155(1):199–230, 2016. ISSN 1436-4646. doi: 10.1007/s10107-014-0841-6.

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.

Robert M Freund, Paul Grigas, and Rahul Mazumder. An extended frank–wolfe method with "in-face" directions, and its application to low-rank matrix completion. *SIAM Journal on Optimization*, 27(1):319–346, 2017.

Dan Garber and Elad Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization. *arXiv preprint arXiv:1301.4666*, 2013.

Dan Garber and Ofer Meshi. Linear-memory and decomposition-invariant linearly convergent conditional gradient algorithm for structured polytopes. *arXiv preprint, arXiv:1605.06492v1*, May 2016.

Dan Garber, Shoham Sabach, and Atara Kaplan. Fast generalized conditional gradient method with applications to matrix recovery problems. *arXiv preprint arXiv:1802.05581*, 2018.

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.

Jacques Guélat and Patrice Marcotte. Some comments on wolfe's away step. *Mathematical Programming*, 35(1):110–119, 1986.

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

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*, pages 253–268. Springer, 2014.

Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Restarting frank-wolfe. *arXiv preprint arXiv:1810.02429*, 2018a.

Thomas Kerdreux, Fabian Pedregosa, and Alexandre d'Aspremont. Frank-wolfe with subsampling oracle. *arXiv preprint arXiv:1803.07348*, 2018b.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank–Wolfe optimization variants. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 496–504. Curran Associates, Inc., 2015.

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate frank-wolfe optimization for structural svms. *arXiv preprint arXiv:1207.4747*, 2012.

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate frank-wolfe optimization for structural svms. In *ICML 2013 International Conference on Machine Learning*, pages 53–61, 2013.

Guanghui Lan, Sebastian Pokutta, Yi Zhou, and Daniel Zink. Conditional accelerated lazy stochastic gradient descent. *Proceedings of ICML*, 2017.

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.

Antoine Miech, Jean-Baptiste Alayrac, Piotr Bojanowski, Ivan Laptev, and Josef Sivic. Learning from video and text via large-scale discriminative clustering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5267–5276. IEEE, 2017.

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.

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

Anton Osokin, Jean-Baptiste Alayrac, Isabella Lukasewitz, Puneet K Dokania, and Simon Lacoste-Julien. Minding the gaps for block frank-wolfe optimization of structured svms. *ICML 2016 International Conference on Machine Learning / arXiv preprint arXiv:1605.09346*, 2016.

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. In *Advances in Neural Information Processing Systems*, pages 1119–1129, 2017.

Neel Shah, Vladimir Kolmogorov, and Christoph H Lampert. A multi-plane block-coordinate frank-wolfe algorithm for training structural svms with a costly max-oracle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2737–2745, 2015.

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.