# Adaptive Step-Size Rule for Conditional Gradient Methods Minimizing Weakly Smooth Objective Functions

*Masaru Ito (Nihon University, Japan)

Joint work with Zhaosong Lu and Chuan He (University of Minnesota)

SIAM Conference on Optimizaiton
July 22, 2021

# Problem setting

## Composite optimization

$$\varphi^* := \min_{x \in \mathbb{R}^n} \varphi(x), \quad \varphi(x) := f(x) + g(x),$$

$f : \mathbb{R}^n \to \mathbb{R}$ is a $C^1$ function and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is an lsc convex function.

- Example: $g$ is the indicator function of a compact convex set
- $f$ is possibly non-convex.
- Assumption 1: $\mathrm{dom}\, g$ is bounded ($\mathrm{dom}\, g = \{x : g(x) < +\infty\}$)
- Assumption 2: $f$ is weakly smooth, i.e., $\nabla f$ is Hölder continuous: $\exists \nu \in (0, 1]$, $\exists L_f^{(\nu)} > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\|_* \le L_f^{(\nu)} \|x - y\|^{\nu}, \quad \forall x, y \in \mathrm{dom}\, g,$$

where $\| \cdot \|_*$ is the dual of $\| \cdot \|$.

# Approximate solutions

For the problem $\varphi^* = \min_x[\varphi(x) = f(x) + g(x)]$, we define the quantity

$$\delta(x) := \max_{v}\{\langle \nabla f(x), x - v\rangle + g(x) - g(v)\} \geq 0.$$

It is often called the 'Frank-Wolfe gap' at $x$.

(1) $\delta(x) = 0$ if and only if $0 \in \nabla f(x) + \partial g(x)$.
(2) $\varphi(x) - \varphi^* \leq \delta(x)$ if $f$ is convex.

- Assumption 3: For any fixed $x$, we can solve the following convex optimization (i.e., $\delta(x)$ is computable)

$$\min_{v \in \mathbb{R}^n}\{\langle \nabla f(x), v\rangle + g(v)\}$$

Example: When $g = \mathrm{ind}_C$ for a compact convex set $C$, the above problem is $\min\{\langle \nabla f(x), v\rangle : v \in C\}$

# Conditional gradient method

## Frank-Wolfe method [Frank-Wolfe '56] with regularization [Bach '15]

$x_0 \in \operatorname{dom} g$
For $t = 0, 1, 2, \ldots$:
(1) $v_t \in \mathrm{Argmin}_{v \in \mathbb{R}^n} \{\langle \nabla f(x_t), v \rangle + g(v)\}$ (Convex optimization)
(2) Terminate if $\delta_t := \delta(x_t)$ is sufficiently small
(3) $x_{t+1} := x_t + \tau_t (v_t - x_t)$ $(\in \operatorname{dom} g)$, for some step size $\tau_t \in [0, 1]$.

# Conditional gradient method

**Frank-Wolfe method [Frank-Wolfe '56] with regularization [Bach '15]**

$x_0 \in \operatorname{dom} g$
For $t = 0, 1, 2, \ldots$:
(1) $v_t \in \operatorname{Argmin}_{v \in \mathbb{R}^n} \{\langle \nabla f(x_t), v \rangle + g(v)\}$ (Convex optimization)
(2) Terminate if $\delta_t := \delta(x_t)$ is sufficiently small
(3) $x_{t+1} := x_t + \tau_t(v_t - x_t)$ $(\in \operatorname{dom} g)$, for some step size $\tau_t \in [0, 1]$.

- Complexity per iteration: Gradient $\nabla f(x_t)$ and convex minimization $\min\{\langle \nabla f(x_t), v \rangle + g(v)\}$
- Cheap iteration cost compared to proximal gradient methods (Minimizing linear function$+g$ vs Minimizing quadratic function$+g$).
  $\longrightarrow$ Large scale optimization: Machine learning, Data mining, etc.
- Computable termination criterion $\delta_t \leq \varepsilon$
- Slower convergence rate than (accelerated) proximal gradient method ($O(1/t)$ vs $O(1/t^2)$ for smooth convex $f$).

Key: Step size rule affects the rate of convergence

# Some existing step size rules

The basic tool is the "Descent lemma" :

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f^{(\nu)}}{1+\nu} \|y - x\|^{1+\nu}.$$

(1) Exact line search $\tau_t \in \mathrm{Argmin}_{\tau \in [0,1]} \varphi(x_t + \tau(v_t - x_t))$.
Convergence results follows for many cases.

# Some existing step size rules

The basic tool is the "Descent lemma" :

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f^{(\nu)}}{1+\nu} \|y - x\|^{1+\nu}.$$

(1) Exact line search $\tau_t \in \mathrm{Argmin}_{\tau \in [0,1]} \varphi(x_t + \tau(v_t - x_t))$.
Convergence results follows for many cases.

(2) $\tau_t = \min \left\{ 1, \frac{\delta_t}{L_f^{(1)} \|x_t - v_t\|^2} \right\}$ [Frank & Wolfe '56] for the case $\nu = 1$
When $f$ is smooth, same convergence guarantee as (1) follows.

# Some existing step size rules

The basic tool is the "Descent lemma" :

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f^{(\nu)}}{1+\nu}\|y - x\|^{1+\nu}.$$

(1) Exact line search $\tau_t \in \mathrm{Argmin}_{\tau \in [0,1]} \varphi(x_t + \tau(v_t - x_t))$.
Convergence results follows for many cases.

(2) $\tau_t = \min\left\{1, \frac{\delta_t}{L_f^{(1)}\|x_t - v_t\|^2}\right\}$ [Frank & Wolfe '56] for the case $\nu = 1$
When $f$ is smooth, same convergence guarantee as (1) follows.

(3) $\tau_t = 2/(t+2)$ [Clarkson 2008], [Hazan 2008], [Nesterov 2018]
Convergence results follows when $f$ is convex.
Advantage: It is parameter free.

# Some existing step size rules

The basic tool is the "Descent lemma" :

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_f^{(\nu)}}{1+\nu} \|y - x\|^{1+\nu}.$$

(1) Exact line search $\tau_t \in \mathrm{Argmin}_{\tau \in [0,1]} \varphi(x_t + \tau(v_t - x_t))$.
Convergence results follows for many cases.

(2) $\tau_t = \min \left\{ 1, \dfrac{\delta_t}{L_f^{(1)} \|x_t - v_t\|^2} \right\}$ [Frank & Wolfe '56] for the case $\nu = 1$
When $f$ is smooth, same convergence guarantee as (1) follows.

(3) $\tau_t = 2/(t+2)$ [Clarkson 2008], [Hazan 2008], [Nesterov 2018]
Convergence results follows when $f$ is convex.
Advantage: It is parameter free.

(4) $\tau_t = \min \left\{ 1, \left( \dfrac{\delta_t}{L_f^{(\nu)} \|x_t - v_t\|^{1+\nu}} \right)^{1/\nu} \right\}$ [Zhao & Freund, 2020]
Same convergence guarantee as (1) follows.
It is parameter dependent.

# Proposed method: Adaptive step size rule

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L(\varepsilon)}{2}\|y - x\|^2 + \varepsilon, \quad L(\varepsilon) = \left(\frac{1+\nu}{1-\nu}\frac{1}{2\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} (L_f^{(\nu)})^{\frac{2}{1+\nu}}$$

## Proposed method: Adaptive step size rule

$x_0 \in \operatorname{dom} g$, $L_{-1} > 0$

For $t = 0, 1, 2, \ldots$:

(1) $v_t \in \operatorname{Argmin}_{v \in \mathbb{R}^n} \{\langle \nabla f(x_t), v\rangle + g(v)\}$ (Convex optimization)

(2) Terminate if $\delta_t := \delta(x_t)$ is sufficiently small.

(3) Adaptive line search to compute $\tau_t \in [0, 1]$:

(3a) Repeat $i = 0, 1, 2, \ldots$:

$L_t^{(i)} := 2^{i-1} L_{t-1}$

$\tau_t^{(i)} := \min\left\{1, \frac{\delta_t/2}{L_t^{(i)}\|x_t - v_t\|^2}\right\}$

$x_{t+1}^{(i)} := x_t + \tau_t^{(i)}(v_t - x_t)$

Until $\varphi(x_{t+1}^{(i)}) \le \varphi(x_t) - \tau_t^{(i)}\delta_t/2 + \frac{1}{2}L_t^{(i)}(\tau_t^{(i)})^2\|x_t - v_t\|^2$

(3b) $\tau_t := \tau_t^{(i)}$, $\quad L_t := L_t^{(i)}$.

(4) $x_{t+1} := x_t + \tau_t(v_t - x_t)$ $\;(\in \operatorname{dom} g)$

# Main result: Rate of convergence of proposed method

## Theorem

(i) The number of iterations $T_\varepsilon$ to attain $\delta_t \le \varepsilon$ is bounded as follows.

$$T_\varepsilon \le O(1) \left( \frac{L_f^{(\nu)} D_{\mathrm{dom}\, g}^{1+\nu}}{\varepsilon} \right)^{\frac{1}{\nu}} \frac{\Delta_0}{\varepsilon},$$

where $\Delta_0 = \varphi(x_0) - \varphi^*$, $D_{\mathrm{dom}\, g} = \mathrm{diam}(\mathrm{dom}\, g)$, and $O(1)$ is an absolute constant.

(ii) When we further assume $f$ is convex,

$$T_\varepsilon \le O(1) \left( \frac{L_f^{(\nu)} D_{\mathrm{dom}\, g}^{1+\nu}}{\varepsilon} \right)^{\frac{1}{\nu}}.$$

- We can prove the same result for the exact line search (1) or the step size rule (4) of [Zhao-Freund 2020].
- Advantage of proposed method: It is parameter free.

# Faster convergence under error bound

- Some conditions for linear convergence:
  (1) $f$ is smooth convex, $g = \mathrm{ind}_C$ for a strongly convex set $C$ which does not contain stationary points of $f$ [Levitin & Polyak, 1966]
  $$\lambda x + (1-\lambda)y + \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2 u \in C, \quad \forall x, y \in C,\ \lambda \in [0,1],\ u \in B(0,1)$$
  (2) $f$ is smooth convex, $g$ is a strongly convex function [Ghadimi, 2019]
  $$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|^2$$
- Existing step size rules are parameter dependent or analyzed for $\nu = 1$.
- We introduce an error bound condition and observe the convergence rate of our proposed method.

## Error bound of subproblems

Assume that there exists $\mu > 0$ and $\rho \geq 2$ such that any solution $v^* \in \mathrm{Argmin}_v[\langle \nabla f(x), v\rangle + g(v)]$ satisfies

$$[\langle \nabla f(x), v\rangle + g(v)] - [\langle \nabla f(x), v^*\rangle + g(v^*)] \geq \frac{\mu}{\rho}\|v - v^*\|^{\rho}, \quad \forall v \in \operatorname{dom} g.$$

## Error bound of subproblems

Assume that there exists $\mu > 0$ and $\rho \geq 2$ such that any solution $v^* \in \mathrm{Argmin}_v[\langle \nabla f(x), v\rangle + g(v)]$ satisfies

$$[\langle \nabla f(x), v\rangle + g(v)] - [\langle \nabla f(x), v^*\rangle + g(v^*)] \geq \frac{\mu}{\rho}\|v - v^*\|^{\rho}, \quad \forall v \in \mathrm{dom}\, g.$$

Examples:

(1) $g = \mathrm{ind}_C$ for a uniformly convex set $C$ which does not contain stationary points of $f$

$$\lambda x + (1-\lambda)y + \tfrac{c}{2}\lambda(1-\lambda)\|x-y\|^{\rho} u \in C,$$
$$\forall x, y \in C,\ \lambda \in [0,1],\ u \in B(0,1)$$

(2) $g$ is a uniformly convex function

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y) - \tfrac{\mu}{2}\lambda(1-\lambda)[\lambda^{\rho} + (1-\lambda)^{\rho}]\|x-y\|^{\rho}$$

# Main result 2: Faster convergence under error bound

**Theorem**

Suppose that the error bound condition holds.
(i) The number of iterations $T_\varepsilon$ to attain $\delta_t \le \varepsilon$ is bounded by

$$T_\varepsilon \le O(1)\left(\frac{\rho^{1+\nu}(L_f^{(\nu)})^\rho}{\mu^{1+\nu}\varepsilon^{\rho-1-\nu}}\right)^{\frac{1}{\nu\rho}}\frac{\Delta_0}{\varepsilon},$$

where $\Delta_0 = \varphi(x_0) - \varphi^*$.
(ii) When we further assume $f$ is convex,

$$T_\varepsilon \le \begin{cases} O(1)\dfrac{L_f^{(1)}}{\mu}\log\dfrac{\Delta_0}{\varepsilon} & (\rho=\nu+1=2) : \text{linear convergence}, \\ O(1)\left(\dfrac{\rho^{1+\nu}(L_f^{(\nu)})\rho}{\mu^{1+\nu}\varepsilon^{\rho-1-\nu}}\right)^{\frac{1}{\nu\rho}} & (\text{otherwise}). \end{cases}$$

# Summary

- Proposed step size rule does not rely on parameters in the problem.
- The iteration complexity bound is the same as the one for the exact line search.

Further interests

- Improvements of oracle complexity
- Analysis under more general setting than Hölder condition

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L(\varepsilon)}{2}\|y - x\|^2 + \varepsilon, \quad L(\varepsilon) = \left( \frac{1+\nu}{1-\nu} \frac{1}{2\varepsilon} \right)^{\frac{1-\nu}{1+\nu}} (L_f^{(\nu)})^{\frac{2}{1+\nu}}$$

# Summary

- Proposed step size rule does not rely on parameters in the problem.
- The iteration complexity bound is the same as the one for the exact line search.

Further interests

- Improvements of oracle complexity
- Analysis under more general setting than Hölder condition

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L(\varepsilon)}{2}\|y - x\|^2 + \varepsilon, \quad L(\varepsilon) = \left(\frac{1+\nu}{1-\nu}\frac{1}{2\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} (L_f^{(\nu)})^{\frac{2}{1+\nu}}$$

Thank you for your attention!

# References I


- Francis Bach, Duality between subgradient and conditional gradient methods, SIAM J. Optim., **25**(1):115-129, 2015.
- Jonathan M. Borwein, Guoyin Li, and Liangjin Yao, Analysis of the convergence rate for the cyclic projection algorithm applied to basic semi-algebraic convex sets, SIAM J. Optim., **24**(1):498–527, 2014.
- J. C. Dunn, Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals, SIAM J. Control Optim., **17**:187–211, 1979.
- Marguerite Frank Philip Wolfe, An algorithm for quadratic programming, Naval Research Logistics Quarterly, **3**:95–110, 1956.
- Robert M. Freund and Paul Grigas, New analysis and results for the Frank-Wolfe method, Math. Program., **155**:199–230, 2016.
- Dan Garber and Elad Hazan, Faster Rates for the Frank-Wolfe Method over Strongly-Convex Sets, Proceedings of the 32nd International Conference on Machine Learning, **37**:541–549, 2015.

# References II

Zaid Harchaoui, Anatoli Juditsky, Arkadi Nemirovski, Conditional gradient algorithms for norm-regularized smooth convex optimization, Math. Program., **152**:75–112, 2015.

Saeed Ghadimi, Conditional gradient type methods for composite nonlinear and stochastic optimization, Math. Program., **173**:431–464, 2019.

Cristóbal Guzmán and Arkadi Nemirovski, On lower complexity bounds for large-scale smooth convex optimization, J. Complexity **31**:1–14, 2015.

Martin Jaggi, Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization, Proceedings of the 30th International Conference on Machine Learning, **28**:427-435, 2013.

Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta, Projection-free optimization on uniformly convex sets, ArXiv preprint, arXiv:2004.11053v2, 2020.

E. S. Levitin and B. T. Polyak, Constrained Minimization Methods, *USSR Computational Mathematics and Mathematical Physics*, **6**(5):1–50, 1966.

# References III


Yu Nesterov, Universal gradient methods for convex optimization problems, Math. Program., **152**:381–404, 2015.

Yu. Nesterov, Complexity bounds for primal-dual methods minimizing the model of objective function, Math. Program., **171**:311–330, 2018.

Y. Nesterov and B.T. Polyak, Cubic regularization of Newton method and its global performance, Math. Program., **108**:177–205, 2006.

C. Zălinescu, Convex analysis in general vector spaces, World Scientific, Singapore, 2002.

Renbo Zhao and Robert M. Freund, Analysis of the Frank-Wolfe method for logarithmically-homogeneous barriers, with an extension, ArXiv preprint, arXiv:2010.08999v1, 2020.

# Numerical example

$$\begin{array}{ll} \min & f(x) = \frac{1}{p}\|Ax - b\|_p^p \\ \text{s.t.} & \|x\|_q \leq 1. \end{array}$$

- $p > 1$ and $q > 1$.
- $A \in \mathbb{R}^{n \times n}$ is symmetric, $n = 1000$, $\lambda_{\min}(A) = 1$, $\lambda_{\max}(A) = 100$.
- $b = A\bar{x}$ with $\|\bar{x}\|_q = 10$.
- Initial point $x_0 = 0$; Termination criterion: $\delta_t \leq 10^{-5}\delta_0$
- Compared three step size rules:
  1. Proposed method with the Euclidean norm $\|\cdot\|_2$.
  2. ZF20: $\tau_t = \min\left\{1, \left(\frac{\delta_t}{L_f^{(\nu)}\|x_t - v_t\|^{1+\nu}}\right)^{1/\nu}\right\}$ [Zhao & Freund 2020] with the Euclidean norm $\|\cdot\|_2$, $\nu = p - 1$,
     $L_f^{(\nu)} = 2^{2-p} n^{\frac{(p-1)(2-p)}{2p}} \lambda_{\max}(A)^p$ when $p \in (1, 2]$.
     $L_f^{(\nu)}$ is unclear for $p > 2$.
  3. $\tau_t = 2/(t+2)$.
- Implemented by Matlab on an Apple desktop with the 3.0GHz Intel Xeon E5-1680v2 processor and 64GB of RAM.

# Numerical example

| $q$ | $p$ | Average of CPU time (sec) Proposed alg | ZF20 | $\frac{2}{t+2}$ | Average of number of iterations Proposed alg | ZF20 | $\frac{2}{t+2}$ |
|---|---|---|---|---|---|---|---|
| 1.5 | 1.3 | 0.016 | 0.55 | 0.21 | 24.0 | 1129.3 | 437.9 |
| | 1.6 | 0.0044 | 0.004 | 0.2 | 5.2 | 7.1 | 442.0 |
| | 2.0 | 0.0033 | 0.0029 | 0.18 | 6.0 | 4.9 | 407.2 |
| | 3.0 | 0.0086 | NA | 0.15 | 11.3 | NA | 363.8 |
| 2.0 | 1.3 | 0.038 | 0.29 | 0.2 | 64.4 | 677.5 | 452.3 |
| | 1.6 | 0.0053 | 0.0032 | 0.18 | 6.2 | 5.1 | 422.6 |
| | 2.0 | 0.0023 | 0.002 | 0.16 | 4.0 | 4.0 | 411.1 |
| | 3.0 | 0.0028 | NA | 0.15 | 5.2 | NA | 378.9 |
| 3.0 | 1.3 | 0.25 | 2.0 | 0.37 | 413.4 | 4419.7 | 776.9 |
| | 1.6 | 0.01 | 0.017 | 0.21 | 12.9 | 33.7 | 418.8 |
| | 2.0 | 0.0041 | 0.0031 | 0.18 | 6.7 | 6.3 | 408.2 |
| | 3.0 | 0.0061 | NA | 0.16 | 6.3 | NA | 381.2 |

Table: Numerical results (average over 10 instances). RED indicates the best.

# Upper bound of the total number of line search iterations

As long as $\min_{0\le i\le t}\delta_i \ge \varepsilon$, the total number of inner iterations in the line search until $t$-th outer iteration is bounded by

$$2t + 2 + \left[\log_2 \frac{2\overline{L}(\varepsilon)}{L_{-1}}\right]_+ ,$$

where $[\alpha]_+ = \max(0,\alpha)$ and

$$\overline{L}(\varepsilon) = \begin{cases} \max\left\{ \left(\frac{1-\nu}{1+\nu}\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} (L_f^{(\nu)})^{\frac{2}{1+\nu}},\quad \left(\frac{2(1-\nu)}{1+\nu}\right)^{\frac{1-\nu}{2\nu}} (L_f^{(\nu)})^{\frac{1}{\nu}} \left(\frac{D_{\mathrm{dom}\, g}}{\varepsilon}\right)^{\frac{1-\nu}{\nu}} \right\} \\ \qquad \text{if } \mathrm{dom}\, g \text{ is bounded}, \\ \max\left\{ \left(\frac{1-\nu}{1+\nu}\frac{1}{\varepsilon}\right)^{\frac{1-\nu}{1+\nu}} (L_f^{(\nu)})^{\frac{2}{1+\nu}},\quad \left(\frac{2(1-\nu)}{1+\nu}\right)^{\frac{1-\nu}{2\nu}} (L_f^{(\nu)})^{\frac{1}{\nu}} \left(\frac{\rho}{\kappa\varepsilon^{\rho-1}}\right)^{\frac{1-\nu}{\rho\nu}} \right\} \\ \qquad \text{if error bound condition holds.} \end{cases}$$