Fast Frank–Wolfe Algorithms with Adaptive Bregman Step-Size for Weakly Convex Functions

The 14th International Conference on Learning Representations

Shota Takahashi,[1] Sebastian Pokutta,[2,3] and Akiko Takeda[1,4]

[1]University of Tokyo [2]Zuse Institute Berlin [3]TU Berlin [4]Riken

April 25th, 2026

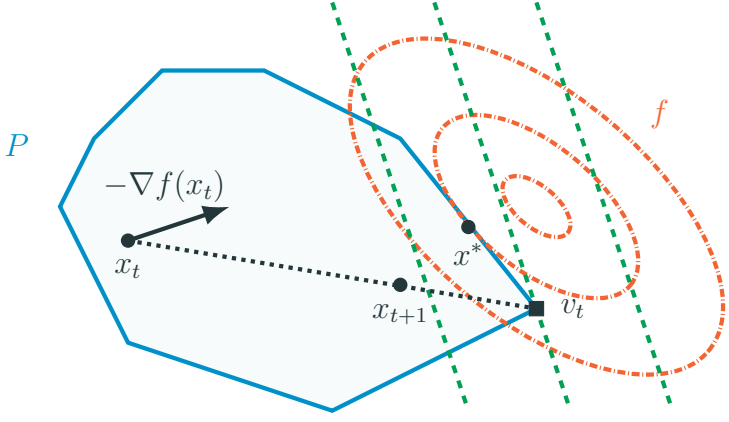
Frank–Wolfe (FW) Algorithm [Frank and Wolfe, 1956]

$$\min_{x \in P} \quad f(x)$$

where $P \subset \mathbb{R}^n$ is compact and convex

Fast even if computing projection is not easy

$$v_t \leftarrow \operatorname*{argmin}_{v \in P} \langle \nabla f(x_t), v \rangle, \quad x_{t+1} \leftarrow (1 - \gamma_t) x_t + \gamma_t v_t, \qquad \gamma_t \in [0, 1]$$



Propose FW algorithm with adaptive Bregman step-size strategy

	Assumptions				Convergence rate	
FW	f convex	f growth	$x^* \in \operatorname{int} P$	P polytope	L-smooth	L-smad
any	convex	✘	✘	✘	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-1/\nu})$
adaptive	convex	q-HEB	✔	✘	$\mathcal{O}(\log \epsilon^{-1})$	$\mathcal{O}(\log \epsilon^{-1})$
AFW	convex	q-HEB	✘	✔	$\mathcal{O}(\log \epsilon^{-1})$	$\mathcal{O}(\log \epsilon^{-1})$
any	weak	2-HEB[1]	✘	✘	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-1/\nu})$
adaptive	weak	2-HEB[1]	✔	✘	$\mathcal{O}(\log \epsilon^{-1})$	$\mathcal{O}(\log \epsilon^{-1})$
AFW	weak	2-HEB[1]	✘	✔	$\mathcal{O}(\log \epsilon^{-1})$	$\mathcal{O}(\log \epsilon^{-1})$
any	✘	✘	✘	✘	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1-1/\nu})$

- Nonconv. $\langle \nabla f(x_t), x_t - v_t \rangle \leq \epsilon$; (weakly) convex $f(x_t) - f(x^*) \leq \epsilon$
- HEB: Hölder error bound; AWF: away-step FW algorithm

[1] Assume local 2-HEB

Adaptive Bregman Step-Size Strategy

Bregman distance: $D_\phi(x,y) := \phi(x) - \phi(y) - \langle \nabla\phi(x), x-y\rangle$; ϕ is strictly convex

$$
\begin{aligned}
L\textbf{-smad} &\overset{\text{def.}}{\iff} \exists L>0 \text{ such that } L\phi - f, L\phi + f \text{ are convex} \\
&\iff |f(x) - f(y) - \langle \nabla f(y), x-y\rangle| \le L D_\phi(x,y) \\
&\implies f(x) - f((1-\gamma)x+\gamma v) \ge \underbrace{\gamma\langle \nabla f(x), x-v\rangle - L\gamma^{1+\nu} D_\phi(v,x)}_{\text{max at } \tilde{\gamma} = \left(\frac{\langle \nabla f(x), x-v\rangle}{L(1+\nu)D_\phi(v,x)}\right)^{1/\nu}} \qquad (1)
\end{aligned}
$$

Procedure `step_size`$(f, \phi, x, v, \tilde{L}, \gamma_{\max})$
- Choose $\beta \in (0,1)$, $\eta \in (0,1]$, $\tau > 1$
- $\kappa \leftarrow 1$, $M \leftarrow \eta\tilde{L}$
- **while** $f(x) - f((1-\gamma)x+\gamma v) < \gamma\langle \nabla f(x), x-v\rangle - M\gamma^{1+\kappa} D_\phi(v,x)$ **do**
  - $\gamma \leftarrow \min\left\{ \left(\frac{\langle \nabla f(x), x-v\rangle}{M(1+\kappa)D_\phi(v,x)}\right)^{\frac{1}{\kappa}}, \gamma_{\max}\right\}$
  - $M \leftarrow \tau M$, $\kappa \leftarrow \beta\kappa$
- $\tilde{L}^* \leftarrow M$, $\tilde{\nu}^* \leftarrow \kappa$
- **return** $\tilde{L}^*$, $\tilde{\nu}^*$, γ

FW zigzags near optimal face; AFW avoids zigzagging [Wolfe, 1970]

1 $\mathcal{S}_0 \leftarrow \{x_0\}$, $\lambda_{x_0,0} \leftarrow 1$
2 **for** $t = 0, \ldots$ **do**
3 $\quad v_t^{\mathsf{FW}} \leftarrow \operatorname{argmin}_{v \in P} \langle \nabla f(x_t), v \rangle$
4 $\quad v_t^{\mathsf{A}} \leftarrow \operatorname{argmax}_{v \in \mathcal{S}_t} \langle \nabla f(x_t), v \rangle$
5 $\quad$ **if** $\langle \nabla f(x_t), x_t - v_t^{\mathsf{FW}} \rangle \geq \langle \nabla f(x_t), v_t^{\mathsf{A}} - x_t \rangle$ **then**
6 $\quad\quad v_t \leftarrow v_t^{\mathsf{FW}}$, $d_t \leftarrow x_t - v_t^{\mathsf{FW}}$, $\gamma_{t,\max} \leftarrow 1$ ▷ Frank-Wolfe step
7 $\quad$ **else**
8 $\quad\quad v_t \leftarrow v_t^{\mathsf{A}}$, $d_t \leftarrow v_t^{\mathsf{A}} - x_t$, $\gamma_{t,\max} \leftarrow \frac{\lambda_{v_t^{\mathsf{A}},t}}{1-\lambda_{v_t^{\mathsf{A}},t}}$ ▷ Away step
9 $\quad \gamma_t \leftarrow \min\left\{ \left(\frac{\langle \nabla f(x_t), d_t \rangle}{L(1+\nu) D_\phi(v_t, x_t)} \right)^{1/\nu}, \gamma_{t,\max} \right\}$
10 $\quad x_{t+1} \leftarrow x_t - \gamma_t d_t$
11 $\quad$ Update the active set $\mathcal{S}_t$

Assume f convex & ϕ strictly convex & (f, ϕ) is L-smad

Theorem (Linear Convergence (Inner Optimum))

f is q-Hölder error bound (HEB) & $B(x^*, r) \subset P$

$$f(x_t) - f^* \leq \begin{cases} \max\left\{\frac{1}{1+\nu}, 1-\alpha\right\}^{t-1} LD^2 & \text{if } q = 1+\nu \\ LD^2/(1+\nu)^{t-1} & \text{if } 1 \leq t \leq t_0, q > 1+\nu \\ \mathcal{O}(1/t^{\nu q/(q-1-\nu)}) & \text{if } t \geq t_0, q > 1+\nu \end{cases}$$

where $\alpha = \frac{\nu}{1+\nu} \frac{r^{1+1/\nu}}{c^{1+1/\nu} D^{2/\nu}}$, $c = (q/\mu)^{1/q}$, $t_0 \in \mathbb{N}$

Theorem (Linear Convergence (AWF))

f is q-HEB & ϕ strongly convex & P polytope

$$f(x_t) - f^* \leq \begin{cases} \left(1 - \frac{\mu}{32L} \frac{\delta^2}{D^2}\right)^{\lceil (t-1)/2 \rceil} LD^2 & \text{if } (\nu, q) = (1, 2) \\ LD^2/(1+\nu)^{\lceil (t-1)/2 \rceil} & \text{if } 1 \leq t \leq t_0, q > 1+\nu \\ \mathcal{O}(1/t^{\nu q/(q-1-\nu)}) & \text{if } t \geq t_0, q > 1+\nu \end{cases}$$

$\mathcal{O}(\epsilon^{-2})$ for nonconvex [Lacoste-Julien, 2016]; $\mathcal{O}(\epsilon^{-1-1/\nu})$ for L-smad nonconvex

f is weakly convex & local quadratic growth $\implies$ Local linear convergence

Definition (Weakly Convex & Local Quadratic Growth)

- f is ρ-**weakly convex** $\overset{\text{def.}}{\iff}$ $\exists \rho > 0$ such that $f + \frac{\rho}{2}\|\cdot\|^2$ is convex
- f is **local quadratic growth** $\overset{\text{def.}}{\iff}$ $\exists \mu > 0$ such that

$$\operatorname{dist}(x, \mathcal{X}^*)^2 \leq \frac{2}{\mu}(f(x) - f^*), \quad \forall x \in [f \leq f^* + \zeta] \cap P \tag{2}$$

$$\implies f(x) - f^* \leq \max_{v \in P} \langle \nabla f(x), x - v \rangle + \underbrace{\frac{\rho}{2}\|x - x^*\|^2}_{\leq \frac{\rho}{\mu}(f(x) - f^*)}$$

$$\therefore \quad \left(1 - \frac{\rho}{\mu}\right) \underbrace{(f(x) - f^*)}_{\text{Primal gap}} \leq \underbrace{\max_{v \in P} \langle \nabla f(x), x - v \rangle}_{\text{FW gap}}$$

f is ρ-weakly convex & local quadratic growth; $M := (L(1+\nu))^{1/\nu}$

Theorem (Local Linear Convergence (Inner Optimum))

$B(x^*, r) \subset P$ & $\rho/\mu < 1$

$$f(x_t) - f^* \leq \begin{cases} \max\left\{\frac{1}{2} + \frac{\rho}{2\mu}, 1 - \frac{\nu}{1+\nu}\frac{r^{1+1/\nu}}{Mc^{1+1/\nu}D^{2/\nu}}\right\}^{t-1} LD^2 & \text{if } \nu = 1 \\ \left(\frac{1}{1+\nu}\left(1 + \frac{\nu\rho}{\mu}\right)\right)^{t-1} LD^2 & \text{if } 1 \leq t \leq t_0, \nu \in (0,1) \\ \mathcal{O}(1/t^{2\nu/(1-\nu)}) & \text{if } t \geq t_0, \nu \in (0,1) \end{cases}$$

Theorem (Local Linear Convergence (AFW))

ϕ strongly convex & P polytope & $\rho < \mu \leq L$

$$f(x_t) - f^* \leq \begin{cases} \left(1 - \frac{(\mu-\rho)^2}{32\mu L}\frac{\delta^2}{D^2}\right)^{\lceil (t-1)/2 \rceil} LD^2 & \text{if } \nu = 1 \\ 2\left(\frac{1}{1+\nu}\left(1 + \frac{\nu\rho}{\mu}\right)\right)^{\lceil (t-1)/2 \rceil} LD^2 & \text{if } 1 \leq t \leq t_0, \nu \in (0,1) \\ \mathcal{O}(1/t^{2\nu/(1-\nu)}) & \text{if } t \geq t_0, \nu \in (0,1) \end{cases}$$

Numerical Experiments

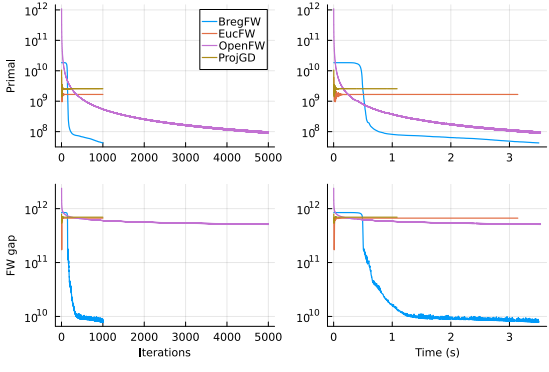


Figure 1: ℓ_p Loss Problem: Gas Sensor Data

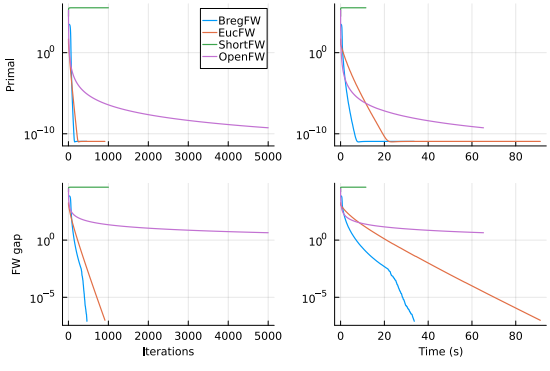


Figure 2: Phase Retrieval

- **BregFW**: Proposed
- EucFW [Pedregosa et al., 2020]: search L, then $\gamma_t = \frac{\langle \nabla f(x_t), x_t - v_t \rangle}{L \|x_t - v_t\|^2}$
- ShortFW: $\gamma_t = \frac{\langle \nabla f(x_t), x_t - v_t \rangle}{L \|x_t - v_t\|^2}$
- OpenFW: $\gamma_t = \frac{2}{2+t}$
- ProjGD: Projected gradient descent

References i

[Frank and Wolfe, 1956] Frank, M. and Wolfe, P. (1956).
An algorithm for quadratic programming.
Nav. Res. Logist. Q., 3(1-2):95–110.

[Lacoste-Julien, 2016] Lacoste-Julien, S. (2016).
Convergence rate of Frank-Wolfe for non-convex objectives.
arXiv preprint arXiv:1607.00345.

[Pedregosa et al., 2020] Pedregosa, F., Negiar, G., Askari, A., and Jaggi, M. (2020).
Linearly convergent Frank-Wolfe with backtracking line-search.
In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *PMLR*, pages 1–10.

[Wolfe, 1970] Wolfe, P. (1970).
Convergence theory in nonlinear programming.
Integer and nonlinear programming, pages 1–36.