Supplementary material: Faster rates for the Frank-Wolfe algorithm using Jacobi polynomials

Robin Francis and Sundeep Prabhakar Chepuri

This document presents convergence results for the Jacobi accelerated Frank-Wolfe (JFW) method summarized as Algorithm 1.

Algorithm 1 Jacobi accelerated Frank-Wolfe (JFW)

1: Initialize $\mathbf{x}_0 \in \mathcal{C}$, $\alpha \geq \beta > -1$, and γ
2: **for** $k = 0, 1, \ldots$ **do**
3: $\quad \mathbf{s}_k \leftarrow \arg\min_{\mathbf{s}\in\mathcal{C}} \ \langle \nabla f(\mathbf{x}_k), \mathbf{s}\rangle$ ▷ FW direction finding
4: $\quad \mathbf{y}_{k+1} \leftarrow \mathbf{x}_k + \gamma_k(\mathbf{s}_k - \mathbf{x}_k)$ ▷ FW update
5: $\quad \mathbf{z}_{k+1} \leftarrow (a_k(1-\gamma) + b_k)\mathbf{y}_{k+1} - c_k\mathbf{x}_k$ ▷ Jacobi recursion
6: $\quad \mathbf{x}_{k+1} \leftarrow \mathbf{z}_{k+1} + \gamma a_k \mathbf{x}_k$ ▷ Correction step
7: $\quad \gamma_k \leftarrow \frac{2}{k+2}$

We first provide the following definitions and results before presenting the convergence results.

Definition 1. *We say that a function $f : \mathbf{dom}(f) \to \mathbb{R}$ is L smooth over a convex set $\mathcal{C} \subseteq \mathbf{dom}(f)$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ it holds that*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Lemma 1 (Lower bound on $g(\mathbf{x}_k)$)**.** *For the Jacobi FW algorithm with step size $\gamma_k = \frac{2}{k+2}$, $g(\mathbf{x}_k) = \max_{\mathbf{s}\in\mathcal{C}} \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{s}\rangle$ can be bounded as*

$$g(\mathbf{x}_k) \geq \frac{4LD^2}{k+2}$$

if $f(\mathbf{x}_k) - f(\mathbf{x}^\star) \leq \left|\frac{\alpha}{\beta}\right| \frac{4LD^2}{(k+1)(k+2)}$ holds. Here, $D = \sup_{\mathbf{x},\mathbf{y}\in\mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_2$ is the diameter of the constraint set

Proof. Assuming optimality at the kth JFW iterate, i.e., $\mathbf{x}^\star = \omega_k \mathbf{y}_{k+1} + (1-\omega_k)\mathbf{x}_k$ and substituting $\mathbf{y}_{k+1} = \mathbf{x}_k + \gamma_k(\mathbf{s}_k - \mathbf{x}_k)$, we get

$$\mathbf{x}^\star - \mathbf{x}_k = \gamma_k\omega_k(\mathbf{s}_k - \mathbf{x}_k),$$

where $\omega_k = a_k(1-\gamma) + b_k$. From convexity of f, we have the following

$$\begin{aligned} f(\mathbf{x}^\star) &\geq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k\rangle \\ f(\mathbf{x}^\star) - f(\mathbf{x}_k) &\geq \omega_k\gamma_k \langle \nabla f(\mathbf{x}_k), \mathbf{s}_k - \mathbf{x}_k\rangle \\ &\overset{(a)}{\geq} -\omega_k\gamma_k g_k \end{aligned}$$

where in (a) we use the definition of $g(\mathbf{x}_k)$. When $f(\mathbf{x}_k) - f(\mathbf{x}^\star) \leq \left|\frac{\alpha}{\beta}\right| \frac{4LD^2}{(k+1)(k+2)}$, we have

$$\begin{aligned} g_k \geq& \frac{k+2}{2\omega_k}\left|\frac{\alpha}{\beta}\right| \frac{4LD^2}{(k+1)(k+2)} \\ =& \frac{1}{2\omega_k}\left|\frac{\alpha}{\beta}\right| \frac{4LD^2}{k+1} \\ \overset{(a)}{\geq}& \frac{4LD^2}{k+2} \end{aligned}$$

The authors are with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India. Email:{robinfrancis;spchepuri}@iisc.ac.in.

where we use the fact that $\alpha \geq \beta$ and $\omega_k \leq 1$ in (a). Thus we have the above lower bound. □

We next provide the descent lemma for `JFW`.

Lemma 2. *The Jacobi accelerated Frank-Wolfe algorithm with* $\gamma, \gamma_k \in [0, 1]$ *satisfies*

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \gamma_k \omega_k g(\mathbf{x}_k) + \frac{L}{2}\omega_k \gamma_k^2 D^2,$$

where $\omega_k = a_k(1-\gamma) + b_k = 1 + c_k - \gamma a_k$ *with* (a_k, b_k, c_k) *being the recurrence weights that characterize the second-order recursion of the Jacobi polynomials.*

Proof. Recall the Jacobi recursion update in Step 6 of Algorithm 1 with $\omega_k = a_k(1-\gamma) + b_k$:

$$\mathbf{x}_{k+1} = \omega_k \mathbf{y}_{k+1} + (1-\omega_k)\mathbf{x}_k.$$

Due to convexity of f, we have

$$f(\mathbf{x}_{k+1}) \leq \omega_k f(\mathbf{y}_k) + (1-\omega_k) f(\mathbf{x}_k).$$

Since f is L-smooth, we apply Definition 1 to obtain

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq \omega_k(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{y}_{k+1} - \mathbf{x}_k\rangle + \frac{L}{2}||\mathbf{y}_{k+1} - \mathbf{x}_k||^2) + (1-\omega_k) f(\mathbf{x}_k) \\ &\overset{(a)}{\leq} f(\mathbf{x}_k) + \omega_k(\langle \nabla f(\mathbf{x}_k), \mathbf{y}_{k+1} - \mathbf{x}_k\rangle + \frac{L}{2}\|\mathbf{y}_{k+1} - \mathbf{x}_k\|^2) \\ &\overset{(b)}{\leq} f(\mathbf{x}_k) - \omega_k(\gamma_k\langle \nabla f(\mathbf{x}_k), \mathbf{s}_k - \mathbf{x}_k\rangle + \frac{L}{2}\gamma_k^2\|\mathbf{s}_k - \mathbf{x}_k\|^2) \\ &\overset{(c)}{\leq} f(\mathbf{x}_k) - \omega_k\gamma_k g(\mathbf{x}_k) + \frac{L}{2}\omega_k\gamma_k^2 D^2, \end{aligned} \tag{1}$$

where we use the fact $a_k + b_k - c_k = 1$ that to arrive at (a). To arrive at (b), we Step 4 of Algorithm 1 and (c) is obtained directly using definitions $g(\mathbf{x}_k) = \max_{\mathbf{s} \in \mathcal{C}} \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{s}\rangle$ and $D = \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_2$. □

Using this lower bound from Lemma 1 with $\gamma_k = 2/(k+2)$, we have

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{6\omega_k L D^2}{(k+2)^2},$$

which asserts that `JFW` is a descent method as ω_k and L are nonnegative.

We now present the main theorem on the convergence rate of `JFW`.

Theorem 1. *Let* $f : \mathbf{dom}(f) \to \mathbb{R}$ *be a L-smooth and convex function,* $\mathcal{C} \subseteq \mathbf{dom}(f)$ *be compact and convex, and* $\mathbf{x}^\star$ *be a minimizer of* f *over* $\mathcal{C}$*. For appropriately chosen parameters* (α, β, γ) *with* $\alpha \geq \beta > -1$*,* `JFW` *in Algorithm 1 satisfies*

$$f(\mathbf{x}_k) - f(\mathbf{x}^\star) \leq \left|\frac{\alpha}{\beta}\right| \frac{4LD^2}{(k+1)(k+2)}, \tag{2}$$

where $D = \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_2$ *is the diameter of the constraint set.*

Proof. The proof is based on mathematical induction.

Let us define the suboptimality gap $h_k = f(\mathbf{x}_k) - f(\mathbf{x}^\star)$. Let us first consider the base case $k = 0$. From Definition 1 with $\mathbf{y} = \mathbf{x}_0$ and $\mathbf{x} = \mathbf{x}^\star$, we have

$$h_0 \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \leq \frac{1}{2}LD^2 \leq \left|\frac{\alpha}{\beta}\right| 2LD^2$$

as D is the diameter of the set and $\alpha \geq \beta$ by assumption.

Subtracting $f(\mathbf{x}^\star)$ from both sides of (1), we obtain

$$h_{k+1} \leq h_k - \omega_k\gamma_k g(\mathbf{x}_k) + \frac{L}{2}\omega_k\gamma_k^2 D^2. \tag{3}$$

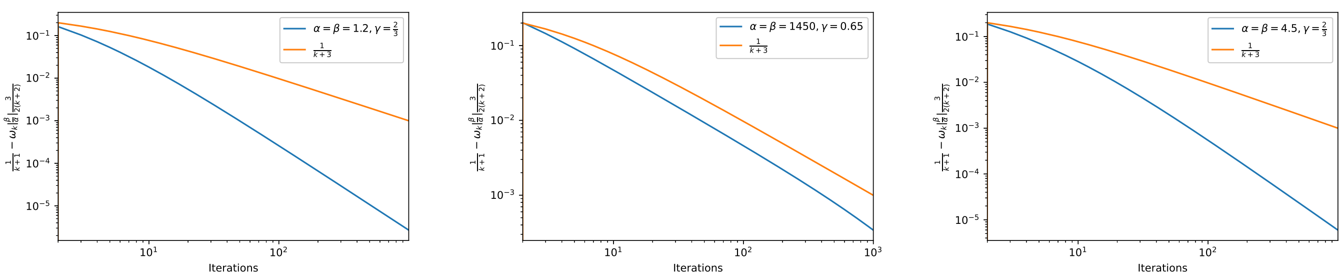
From induction hypothesis, we assume that the upper bound on the suboptimality gap holds up to k iterations. Substituting the upper bound on h_k and lower bound on $g(\mathbf{x}_k)$ from Lemma 1 in (3), we have

$$\begin{aligned}h_{k+1} &\leq \left|\frac{\alpha}{\beta}\right| \frac{4LD^2}{(k+1)(k+2)} + \omega_k \left(-\frac{8LD^2}{(k+2)^2} + \frac{2LD^2}{(k+2)^2}\right) \\ &= \left|\frac{\alpha}{\beta}\right| \frac{4LD^2}{(k+2)} \left(\frac{1}{k+1} - \omega_k \left|\frac{\beta}{\alpha}\right| \frac{3}{2(k+2)}\right) \\ &\overset{(a)}{\leq} \left|\frac{\alpha}{\beta}\right| \frac{4LD^2}{(k+2)(k+3)},\end{aligned}$$

where (a) holds for carefully tuned values of (α, β, γ).

$\square$

The above bound does not hold for arbitrary values of (α, β, γ) and requires hyperparameter tuning. See in the figure below, a few example (α, β, γ) values for which the above bound is valid, where we show that $\frac{1}{k+1} - \omega_k \left|\frac{\beta}{\alpha}\right| \frac{3}{2(k+2)}$ is bounded from above by $\frac{1}{k+3}$.



(a) (b) (c)

Fig. 1: A few examples to illustrate that $\frac{1}{k+3}$ upper bounds $\frac{1}{k+1} - \omega_k \left|\frac{\beta}{\alpha}\right| \frac{3}{2(k+2)}$.