# Supplementary material: Speeding up the Frank-Wolfe method using the Orthogonal Jacobi polynomials

Robin Francis and Sundeep Prabhakar Chepuri

This document presents the 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 $\gamma$
2: **for** $k = 1, 2, \ldots$ **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: $\quad \gamma_k \leftarrow \frac{2}{k+2}$

**Theorem 1.** *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a $L$-smooth and convex function, $\mathcal{C} \subseteq \mathbf{dom}(f)$ be closed and convex, and $\mathbf{x}^\star$ be a minimizer of $f$ over $\mathcal{C}$. For a given $\alpha$ and $\beta$ with $\alpha \geq \beta > -1$ and $\beta \neq 0$, `JFW` in Algorithm 1 satisfies*

$$f(\mathbf{x}_k) - f(\mathbf{x}^\star) \leq \left|\frac{\alpha}{\beta}\right| \frac{2LD^2}{(k+2)(k+3)}, \tag{1}$$

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

### A. Proof of Theorem 1

To show the convergence of the Jacobi FW iterates, we define a few parameters and results that aid the proof. We define the duality gap at each iterate of the FW algorithm and is denoted by $g_k$.

**Definition 1.** *The duality gap for kth iterate of the FW algorithm is defined as,*

$$g(\mathbf{x}^k) = \max_{\mathbf{s}\in\mathcal{C}} \left\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{s} \right\rangle. \tag{2}$$

Once we have defined the duality gap, we try to bound the improvement in each iteration for the family of FW algorithms. From the definition of the curvature constant $M$, the improvement in each iterate can be bounded by the current duality gap.

**Lemma 1.** *For an update of the form $\mathbf{x}^{k+1} = \mathbf{x}^k + \gamma(\mathbf{s}^k - \mathbf{x}^k)$, where step size $\gamma \in [0,1]$ and $\mathbf{x}, \mathbf{s} \in \mathcal{C}$ satisfies*

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \gamma_k g(\mathbf{x}^k) + \frac{\gamma_k^2}{2}M$$

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

*Proof: proof can be found in [1].*

The Lemma: 1 bound the improvement at each iterate and now we try to derive the bound for Jacobi FW algorithm.

Proof: We prove the Theorem 1 by induction. Consider an $L$-smooth and convex function $f$. For the base case, $h_0 = f(\mathbf{x}^0) - f^*$

$$\begin{aligned} h_0 &\leq \frac{L}{2}||\mathbf{x}^0 - \mathbf{x}^*||_2^2 \\ &\leq \left|\frac{\alpha}{\beta}\right| LD^2 \end{aligned}$$

The 1st inequality follows from the $L$-smoothness of the function $f$ and then from the compactness of the constraint, the diameter is bounded by $D$. To bound the error at $k = 1$, we rely on Lemma: 1. From Lemma: 1, we get an upper bound on the error $h_{k+1} = f(\mathbf{x}^{k+1}) - f^*$ at $k$th iterate assuming

$$\begin{aligned} f(\mathbf{x}^{k+1}) - f^* &\leq f(\mathbf{x}^k) - f^* - \gamma_k g(\mathbf{x}^k) + \frac{\gamma_k^2}{2}M \\ h_{k+1} &\leq h_k - \gamma_k g(\mathbf{x}^k) + \frac{\gamma_k^2}{2}M \\ &\leq (1-\gamma_k)h_k + \frac{\gamma_k^2}{2}M. \end{aligned} \tag{3}$$

The final simplified expression for $h_{k+1}$ is obtained by using the fact that dual error will be greater than or equal to the primal error, i.e., $h_k \leq g(\mathbf{x}^k)$.

For $k = 0$, we perform normal FW update with $\gamma_0 = 1$, (3) reduces to,

$$\begin{aligned} h_1 &\leq \frac{1}{2}M \\ &\leq \left|\frac{\alpha}{\beta}\right| \frac{2LD^2}{6}. \end{aligned}$$

The above follows by the assumption that $M = LD^2$ and $\alpha \geq \beta$. By induction, we assume the above holds up to $k$

$$h_k \leq \left|\frac{\alpha}{\beta}\right| \frac{2M}{(k+1)(k+2)}.$$

Now, from Step 5 from Algorithm 1,

$$\begin{aligned} \mathbf{x}^{k+1} &= (a_k(1-\gamma_1) + b_k)\mathbf{y}^{k+1} - (c_k - a_k\gamma_1)\mathbf{x}^k \\ f(\mathbf{x}^{k+1}) &\leq (a_k(1-\gamma_1) + b_k) f(\mathbf{y}^k) - (c_k - a_k\gamma_1) f(\mathbf{x}^k) \end{aligned}$$

In the above equation, we use the convexity of $f$ and the $L$-smoothness property of $f$, to obtain

$$\leq (a_k(1-\gamma_1)+b_k)\bigg(f(\mathbf{x}^k)+\left\langle \nabla f(\mathbf{x}^k), \mathbf{y}^{k+1}-\mathbf{x}^k\right\rangle + \frac{L}{2}||\mathbf{y}^{k+1}-\mathbf{x}^k||_2^2\bigg) - (c_k - a_k\gamma_1)f(\mathbf{x}^k)$$

We substitute $\mathbf{y}^{k+1}-\mathbf{x}^k = \gamma_k(\mathbf{s}^k-\mathbf{x}^k)$ from Step 4 of Algorithm 1 and rewrite the above equation as

$$f(\mathbf{x}^{k+1}) \leq (a_k(1-\gamma_1)+b_k)(f(\mathbf{x}^k) - \gamma_k \left\langle \nabla f(\mathbf{x}^k), \mathbf{x}^k - \mathbf{s}^k\right\rangle + \frac{L}{2}\gamma_k^2||\mathbf{s}^k-\mathbf{x}^k||_2^2) - (c_k - a_k\gamma_1)f(\mathbf{x}^k) \quad (4)$$

We can bound the distance between $\mathbf{s}^k, \mathbf{x}^k$ using boundedness of the constraint set. and we can obtain the expression in terms of $h_k$ and $h_{k+1}$, by subtracting the optimal function value $f^*$ from both sides

$$h_{k+1} \leq (a_k(1-\gamma_1)+b_k)(h_k - \gamma_k g_k + \frac{L}{2}\gamma_k^2 D^2) - (c_k - a_k\gamma_1)h_k \quad (5)$$

The above expression follows as the coefficients of $f(\mathbf{x}^k)$ sum to one i.e., $a_k(1-\gamma_1)+b_k+c_k-a_k\gamma_1 = 1$ and substituting $h_k = f(\mathbf{x}^k) - f^*$. Now to further simplify the expression, we have to find a lower bound for the duality gap at the $k$th iterate for the Jacobi FW.

**Lemma 2.** *For Jacobi FW algorithm with step size $\gamma_k = \frac{2}{k+2}$, the duality gap $g_k$ can be bounded as*

$$g_k \geq \frac{2M}{k+2}.$$

Proof: Assuming optimality at the $k$th iterate i.e., $\mathbf{x}^* = (a_k(1-\gamma_1)+b_k)\mathbf{y}^{k+1} - (c_k - a_k\gamma_1)\mathbf{x}^k$ and substituting $\mathbf{x}^{k+1} = \mathbf{x}^k + \gamma(\mathbf{s}^k - \mathbf{x}^k)$ in the expression for $\mathbf{x}^*$, we get $\mathbf{x}^* = \mathbf{x}^k + (a_k(1-\gamma_1)+b_k)\gamma_k(\mathbf{x}^k - \mathbf{s}^k)$. From the convexity of the objective function $f$

$$\begin{aligned}
f^* \geq & f(\mathbf{x}^k) + \left\langle \nabla f(\mathbf{x}^k), \mathbf{x}^* - \mathbf{x}^k\right\rangle \\
f^* - f(\mathbf{x}^k) \geq & (a_k(1-\gamma_1)+b_k)\gamma_k \left\langle \nabla f(\mathbf{x}^k), \mathbf{s}^k - \mathbf{x}^k\right\rangle \\
-h_k \geq & -(a_k(1-\gamma_1)+b_k)\gamma_k g_k.
\end{aligned}$$

The above expression was obtained from the definitions of $h_k$ and $g_k$. Now using the error bound for $k$th iterate of Jacobi FW

$$\begin{aligned}
g_k \geq & \frac{k+2}{2(a_k(1-\gamma_1)+b_k)}\left|\frac{\alpha}{\beta}\right|\frac{2LD^2}{(k+1)(k+2)} \\
= & \frac{1}{2(a_k(1-\gamma_1)+b_k)}\left|\frac{\alpha}{\beta}\right|\frac{2LD^2}{k+1} \\
\geq & \frac{2LD^2}{k+2}.
\end{aligned}$$

As $\alpha \geq \beta$ and $(a_k(1-\gamma_1)+b_k) \leq 1$, we could bound $g_k$ as given above. From induction hypothesis, we can substitute the bound on error at $k$th iterate $h_k$ and from Lemma: 2, we can substitute for $g_k$ in (5)

$$h_{k+1} \leq \left|\frac{\alpha}{\beta}\right|\frac{2LD^2}{(k+1)(k+2)} + (a_k(1-\gamma_1)+b_k)\left(-\frac{2}{k+2}\frac{2LD^2}{k+2} + \frac{2}{(k+2)^2}LD^2\right)$$

On simplifying further, by taking few terms common

$$\begin{aligned}
h_{k+1} \leq & \left|\frac{\alpha}{\beta}\right|\frac{2LD^2}{(k+1)(k+2)} + (a_k(1-\gamma_1)+b_k)\left(-\frac{2}{k+2}\frac{2LD^2}{k+2} + \frac{2}{(k+2)^2}M\right) \\
\leq & \left|\frac{\alpha}{\beta}\right|\frac{2LD^2}{(k+2)}\left(\frac{1}{k+1} - (a_k(1-\gamma_1)+b_k)\left|\frac{\beta}{\alpha}\right|\frac{1}{k+2}\right)
\end{aligned}$$

We simply further by bounding the above equation as

$$\begin{aligned}
\leq & \left|\frac{\alpha}{\beta}\right|\frac{2LD^2}{(k+2)}\frac{1}{k+3} \\
= & \left|\frac{\alpha}{\beta}\right|\frac{2LD^2}{(k+2)(k+3)}.
\end{aligned}$$

## REFERENCES


[1] M. Jaggi, "Revisiting frank-wolfe: Projection-free sparse convex optimization," in *Proc. of the 30th International Conference on International Conference on Machine Learning*, 2013.