A MISSING PROOFS

Lemma 1. Let f be a L-smooth function over a convex compact domain $\mathcal{D}$, and define $\text{diam}(\mathcal{D}) := \sup_{\mathbf{x},\mathbf{y}\in\mathcal{D}} ||\mathbf{x}-\mathbf{y}||$. Then $\bar{C}_f \leq \text{diam}^2(\mathcal{D})L$.

Proof. Let $\forall \mathbf{x}, \mathbf{s} \in \mathcal{D}, \gamma \in (0,1]$, and $\mathbf{y} = \mathbf{x} + \gamma(\mathbf{s}-\mathbf{x})$. The smoothness of f implies that f is continuously differentiable, hence we have:

$$\begin{aligned}
&\left| f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) \right| \\
&= \left| \int_0^1 \left(\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x})\right)^T (\mathbf{y}-\mathbf{x})\, dt \right| && \text{(Mean-value theorem)} \\
&\leq \int_0^1 \left| \left(\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x})\right)^T (\mathbf{y}-\mathbf{x}) \right| dt && \text{(Triangle inequality)} \\
&\leq \int_0^1 ||\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) - \nabla f(\mathbf{x})|| \cdot ||\mathbf{y}-\mathbf{x}||\, dt && \text{(Cauchy-Schwarz inequality)} \\
&\leq \int_0^1 tL\gamma^2 ||\mathbf{s}-\mathbf{x}||^2\, dt \leq \frac{L\gamma^2}{2}\text{diam}^2(\mathcal{D}) && \text{(Smoothness assumption of } f\text{)}
\end{aligned}$$

It immediately follows that

$$\bar{C}_f \leq \frac{2}{\gamma^2}\frac{L\gamma^2}{2}\text{diam}^2(\mathcal{D}) = \text{diam}^2(\mathcal{D})L$$

■

Theorem 2. Consider the problem (2) where f is a continuously differentiable function that is potentially nonconvex, but has a finite curvature constant $\bar{C}_f$ as defined by (10) over the compact convex domain $\mathcal{D}$. Consider running Frank-Wolfe (Algo. 1), then the minimal FW gap $\tilde{g}_T := \min_{0\leq t\leq T} g_t$ encountered by the iterates during the algorithm after T iterations satisfies:

$$\tilde{g}_T \leq \frac{\max\{2h_0\bar{C}_f, \sqrt{2h_0\bar{C}_f}\}}{\sqrt{T+1}}, \quad \forall T \geq 0 \tag{11}$$

where $h_0 := f(\mathbf{x}^{(0)}) - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x})$ is the initial global suboptimality. It thus takes at most $O(1/\varepsilon^2)$ iterations to find an approximate KKT point with gap smaller than ε.

Proof. Let $\mathbf{y} := \mathbf{x} + \gamma\mathbf{d}$, where $\mathbf{d} := \mathbf{s} - \mathbf{x}$ is the update direction found by the LMO in Alg. 1. Using the definition of $\bar{C}_f$, we have:

$$\begin{aligned}
f(\mathbf{y}) &= f(\mathbf{y}) - f(\mathbf{x}) - \gamma\nabla f(\mathbf{x})^T\mathbf{d} + f(\mathbf{x}) + \gamma\nabla f(\mathbf{x})^T\mathbf{d} \\
&\leq f(\mathbf{x}) + \gamma\nabla f(\mathbf{x})^T\mathbf{d} + \left| f(\mathbf{y}) - f(\mathbf{x}) - \gamma\nabla f(\mathbf{x})^T\mathbf{d} \right| \\
&\leq f(\mathbf{x}) + \gamma\nabla f(\mathbf{x})^T\mathbf{d} + \frac{\gamma^2}{2}\bar{C}_f
\end{aligned}$$

Now using the definition of the FW gap $g(\mathbf{x})$ and for $\forall C \geq \bar{C}_f$, we get:

$$f(\mathbf{y}) \leq f(\mathbf{x}) - \gamma g(\mathbf{x}) + \frac{\gamma^2}{2}\bar{C}_f, \quad \forall \gamma \in (0,1] \tag{15}$$

Depending on whether $C > 0$ or $C = 0$, the R.H.S. of (15) is a either a quadratic function with positive second order coefficient or an affine function. In the first case, the optimal γ^* that minimizes the R.H.S. is $\gamma^* = g(\mathbf{x})/C$. In the second case, $\gamma^* = 1$. Combining the constraint that $\gamma^* \leq 1$, we have $\gamma^* = \min\{1, g(\mathbf{x})/C\}$. Thus we obtain:

$$f(\mathbf{y}) \leq f(\mathbf{x}) - \min\left\{\frac{g^2(\mathbf{x})}{2C}, \left(g(\mathbf{x}) - \frac{C}{2}\right)\mathbb{I}_{g(\mathbf{x})>C}\right\} \tag{16}$$

(16) holds for each iteration in Alg. 1. A cascading sum of (16) through iteration step 1 to $T+1$ shows that:

$$f(\mathbf{x}^{(T+1)}) \leq f(\mathbf{x}^{(0)}) - \sum_{t=0}^{T} \min\left\{\frac{g^2(\mathbf{x}^{(t)})}{2C}, \left(g(\mathbf{x}^{(t)}) - \frac{C}{2}\right)\mathbb{I}_{g(\mathbf{x}^{(t)})>C}\right\} \tag{17}$$

Define $\tilde{g}_T := \min_{0\leq t\leq T} g(\mathbf{x}^{(t)})$ be the minimal FW gap in $T+1$ iterations. Then we can further bound inequality (17) as:

$$f(\mathbf{x}^{(T+1)}) \leq f(\mathbf{x}^{(0)}) - (T+1)\min\left\{\frac{\tilde{g}_T^2}{2C}, \left(\tilde{g}_T - \frac{C}{2}\right)\mathbb{I}_{\tilde{g}_T>C}\right\} \tag{18}$$

We discuss two subcases depending on whether $\tilde{g}_T > C$ or not. The main idea is to get an upper bound on $\tilde{g}_T$ by showing that $\tilde{g}_T$ cannot be too large, otherwise the R.H.S. of (18) can be smaller than the global minimum of f, which is a contradiction. For the ease of notation, define $h_0 := f(\mathbf{x}^{(0)}) - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x})$, i.e., the initial gap to the global minimum of f.

Case I. If $\tilde{g}_T > C$ and $\tilde{g}_T - \frac{C}{2} \leq \frac{\tilde{g}_T^2}{2C}$, from (18), then:

$$0 \leq f(\mathbf{x}^{(T+1)}) - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}) \leq f(\mathbf{x}^{(0)}) - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}) - (T+1)(\tilde{g}_T - \frac{C}{2}) = h_0 - (T+1)(\tilde{g}_T - \frac{C}{2})$$

which implies

$$C < \tilde{g}_T \leq \frac{h_0}{T+1} + \frac{C}{2} \Rightarrow \tilde{g}_T \leq \frac{2h_0 C}{T+1} = O(1/T)$$

On the other hand, solving the following inequality:

$$C - \frac{C}{2} \leq \tilde{g}_T - \frac{C}{2} \leq \frac{\tilde{g}_T^2}{2C} \leq \frac{4h_0^2 C^2}{(T+1)^2}\frac{1}{2C}$$

we get

$$T+1 \leq 2h_0$$

This means that $\tilde{g}_T$ decreases in rate $O(1/T)$ only for at most the first $2h_0$ iterations.

Case II. If $\tilde{g}_T \leq C$ or $\tilde{g}_T - \frac{C}{2} > \frac{\tilde{g}_T^2}{2C}$. Similarly, from (18), we have:

$$0 \leq f(\mathbf{x}^{(T+1)}) - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}) \leq f(\mathbf{x}^{(0)}) - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}) - (T+1)\frac{\tilde{g}_T^2}{2C} = h_0 - (T+1)\frac{\tilde{g}_T^2}{2C}$$

which yields

$$\tilde{g}_T \leq \sqrt{\frac{2h_0 C}{T+1}}$$

Combining the two cases together, we get $\tilde{g}_T \leq \frac{2h_0 C}{T+1}$ if $T+1 \leq 2h_0$; otherwise $\tilde{g}_T \leq \sqrt{\frac{2h_0 C}{T+1}}$. Note that for $T \geq 0$, $\sqrt{T+1} \leq T+1$, thus we can further simplify the upper bound of $\tilde{g}_T$ as:

$$\tilde{g}_T \leq \frac{\max\{2h_0 C, \sqrt{2h_0 C}\}}{\sqrt{T+1}}$$

■

Lemma 3. Let $f(W) = \frac{1}{4}||P - WW^T||_F^2$ and define $\nabla^2 f(W) := \partial \operatorname{vec} \nabla f(W) / \partial \operatorname{vec} W$. Then:

$$\begin{aligned}\nabla^2 f(W) &= W^T W \otimes I_n + I_k \otimes (WW^T - P) \\ &+ (W^T \otimes W)K_{nk}\end{aligned} \tag{12}$$

where K_{nk} is a commutation matrix such that $K_{nk} \operatorname{vec} W = \operatorname{vec} W^T$.

Proof. Using the theory of matrix differential calculus, the Hessian of a matrix-valued matrix function is defined as:

$$\nabla^2 f(W) := \frac{\partial \operatorname{vec} \nabla f(W)}{\partial \operatorname{vec} W}$$

Using the differential notation, we can compute the differential of $\nabla f(W)$ as:

$$\mathrm{d}\nabla f(W) = \mathrm{d}(WW^T - P)W = (\mathrm{d}W)W^T W + W(\mathrm{d}W)^T W + WW^T\,\mathrm{d}W - P\,\mathrm{d}W$$

Vectorize both sides of the above equation and make use of the identity that $\operatorname{vec}(ABC) = (C^T \otimes A)\operatorname{vec} B$ for A, B, C with appropriate shapes, we get:

$$\operatorname{vec}\mathrm{d}\,\nabla f(W) = (W^T W \otimes I_n)\operatorname{vec}\mathrm{d}W + (W^T \otimes W)\operatorname{vec}\mathrm{d}W^T + (I_k \otimes (WW^T - P))\operatorname{vec}\mathrm{d}W$$

Let K_{nk} be a commutation matrix such that $K_{nk}\operatorname{vec} W = \operatorname{vec} W^T$. We can further simplify the above equation as:

$$\operatorname{vec}\mathrm{d}\nabla f(W) = \left(W^T W \otimes I_n + (W^T \otimes W)K_{nk} + I_k \otimes (WW^T - P)\right)\operatorname{vec}\mathrm{d}W \tag{19}$$

It then follows from the first identification theorem [Magnus and Neudecker, 1985, Thm. 6] that the Hessian is given by

$$\nabla^2 f(W) = \left(W^T W \otimes I_n + I_k \otimes (WW^T - P) + (W^T \otimes W)K_{nk}\right) \in \mathbb{R}^{nk \times nk}$$

As a sanity check, the first two terms in $\nabla^2 f(W)$ are clearly symmetric. The third term can be verified as symmetric as well by realizing that $K_{nk}^{-1} = K_{nk}^T$, and

$$W \otimes W^T = K_{nk}(W^T \otimes W)K_{nk}$$

■

Lemma 4. $\sup_{\substack{W \geq 0, \\ W\mathbf{1}_k = \mathbf{1}_n}} ||W^T W||_2 = n.$

Proof. $\forall W \geq 0$, if $W\mathbf{1}_k = \mathbf{1}_n$, then by the Courant-Fischer theorem:

$$\begin{aligned}
||W^T W||_2 &:= \max_{\substack{\mathbf{v} \in \mathbb{R}^k, \\ ||\mathbf{v}||_2 = 1}} ||W^T W \mathbf{v}||_2 && \text{(Courant-Fischer theorem)} \\
&= \max_{\substack{\mathbf{v} \in \mathbb{R}^k_+, \\ ||\mathbf{v}||_2 = 1}} ||W^T W \mathbf{v}||_2 && \text{(Perron-Frobenius theorem)} \\
&\leq \max_{\substack{\mathbf{v} \in \mathbb{R}^k_+, \\ ||\mathbf{v}||_\infty \leq 1}} ||W^T W \mathbf{v}||_2 && (B_2(0,1) \subseteq B_\infty(0,1)) \\
&= ||W^T \mathbf{1}_n||_2 && (W \geq 0, W\mathbf{1}_k = \mathbf{1}_n) \\
&\leq ||W^T \mathbf{1}_n||_1 = n
\end{aligned}$$

To achieve this upper bound, consider $W = \mathbf{1}_n e_1^T$, where e_1 is the first column vector of the identity matrix I_k. In this case $W^T W = e_1 \mathbf{1}_n^T \mathbf{1}_n e_1^T = n e_1 e_1^T$, which is a rank one matrix with a positive eigenvalue n. Hence $\sup ||W^T W||_2 = n$. ■

Lemma 5. Let $c := ||P||_2$. $f = \frac{1}{4}||P - WW^T||_F^2$ is $(3n + c)$-smooth on $\mathcal{D} = \{W \in \mathbb{R}^{n \times k}_+ \mid W\mathbf{1}_k = \mathbf{1}_n\}$.

Proof. Recall that the spectral norm $||\cdot||_2$ is sub-multiplicative and the spectrum of $A \otimes B$ is the product of the spectrums of A and B. Using (12), we have:

$$\begin{aligned}
||\nabla^2 f(W)||_2 &= ||W^T W \otimes I_n + I_k \otimes (WW^T - P) + (W^T \otimes W)K_{nk}||_2 \\
&\leq ||W^T W \otimes I_n||_2 + ||I_k \otimes (WW^T - P)||_2 + ||(W^T \otimes W)K_{nk}||_2 && \text{(Triangle inequality)} \\
&= ||W^T W||_2 ||I_n||_2 + ||I_k||_2 ||WW^T - P||_2 + ||W^T \otimes W||_2 ||K_{nk}||_2 && \text{(submultiplicativity of } ||\cdot||_2) \\
&= ||W^T W||_2 + ||WW^T - P||_2 + ||W^T \otimes W||_2 && (||I_n||_2 = ||I_k||_2 = ||K_{nk}||_2 = 1) \\
&\leq 3||W^T W||_2 + ||P||_2 && \text{(Triangle inequality)} \\
&\leq 3n + c && \text{(Lemma 4)}
\end{aligned}$$

The result then follows from Lemma 2. ■

Lemma 6. Let $\mathcal{D} = \{W \in \mathbb{R}_+^{n \times k} \mid W\mathbf{1}_k = \mathbf{1}_n\}$. Then $\text{diam}^2(\mathcal{D}) = 2n$ with respect to the Frobenius norm.

Proof.

$$\begin{aligned}
\text{diam}^2(\mathcal{D}) &= \sup_{W,Z \in \mathcal{D}} ||W - Z||_F^2 \\
&= \sup_{W,Z \in \mathcal{D}} \sum_{ij} (W_{ij} - Z_{ij})^2 = \sup_{W,Z \in \mathcal{D}} \sum_{ij} W_{ij}^2 + Z_{ij}^2 - 2W_{ij}Z_{ij} \\
&\le \sup_{W,Z \in \mathcal{D}} \sum_{W,Z \in \mathcal{D}} W_{ij}^2 + Z_{ij}^2 \le \sup_{W,Z \in \mathcal{D}} \sum_{W,Z \in \mathcal{D}} W_{ij} + Z_{ij} \\
&= 2n
\end{aligned}$$

Note that choosing $W = \mathbf{1}e_1^T$ and $Z = \mathbf{1}_n e_2^T$ make all the equalities hold in the above inequalities. Hence $\text{diam}^2(\mathcal{D}) = 2n$. ■

Lemma 7. $\inf_{\substack{W \ge 0, \\ W\mathbf{1}_k = \mathbf{1}_n}} ||\nabla^2 f(W)||_2 \ge n/k^2 - c$.

Proof. For a matrix A, we will use $\sigma_i(A)$ to mean the ith largest singular value of A and $\lambda_{max}(A)$, $\lambda_{min}(A)$ to mean the largest and smallest eigenvalues of A, respectively. Recall $\nabla^2 f(W) = W^T W \otimes I_n + I_k \otimes (WW^T - P) + (W^T \otimes W)K_{nk}$. For $W \ge 0, W\mathbf{1}_k = \mathbf{1}_n$, let $r = \text{rank}(W)$. Clearly $r \ge 1$. We have the following inequalities hold:

$$\begin{aligned}
||\nabla^2 f(W)||_2 &= ||W^T W \otimes I_n + I_k \otimes (WW^T - P) + (W^T \otimes W)K_{nk}||_2 \\
&\ge \lambda_{max}\left(WW^T \otimes I_n + (W^T \otimes W)K_{nk}\right) + \lambda_{min}\left(I_k \otimes (WW^T - P)\right) && \text{(Weyl's inequality)} \\
&\ge \lambda_{max}(WW^T \otimes I_n) + \lambda_{min}\left((W^T \otimes W)K_{nk}\right) + \lambda_{min}\left(I_k \otimes (WW^T - P)\right) \\
&= \lambda_{max}(WW^T) + \lambda_{min}(W^T \otimes W) + \lambda_{min}(WW^T - P) \\
&\ge \lambda_{max}(WW^T) + \lambda_{min}(W^T \otimes W) + \lambda_{min}(WW^T) - \lambda_{max}(P) \\
&= \sigma_1^2(W) + 2\sigma_r^2(W) - \lambda_{max}(P) \\
&\ge \sigma_1^2(W) - c && (||P||_2 \le ||P||_F) \\
&\ge \frac{1}{r}||W||_F^2 - c && (r \cdot \sigma_1^2(W) \ge ||W||_F^2) \\
&\ge \frac{1}{k}||W||_F^2 - c && (\text{rank}(W) \le k) \\
&= \frac{1}{k}\sum_{i=1}^n \sum_{j=1}^k W_{ij}^2 - c \\
&\ge \frac{1}{k}\sum_{i=1}^n k\left(\frac{\sum_{j=1}^k W_{ij}}{k}\right)^2 - c && \text{(Cauchy ineq.)} \\
&= \frac{n}{k^2} - c
\end{aligned}$$

where the first three inequalities all follow from Weyl's inequality. ■