Supplementary Material of Lifted Proximal Operator Machines

Optimality Conditions of (Zeng et al., 2018)

The optimality conditions of (Zeng et al., 2018) are (obtained by differentiating the objective function w.r.t. X^n, $\{X^i\}_{i=2}^{n-1}$, $\{W^i\}_{i=1}^{n-1}$, and $\{U^i\}_{i=2}^{n}$, respectively):

$$\frac{\partial \ell(X^n, L)}{\partial X^n} + \mu(X^n - \phi(U^n)) = \mathbf{0}, \tag{1}$$

$$(W^i)^T(W^iX^i - U^{i+1}) + (X^i - \phi(U^i)) = \mathbf{0},\ i = 2, \cdots, n-1, \tag{2}$$

$$(W^iX^i - U^{i+1})(X^i)^T = \mathbf{0},\ i = 1, \cdots, n-1, \tag{3}$$

$$(U^i - W^{i-1}X^{i-1}) + (\phi(U^i) - X^i) \circ \phi'(U^i) = \mathbf{0},\ i = 2, \cdots, n \tag{4}$$

where $\circ$ denotes the element-wise multiplication.

Proof of Theorem 2

If $f(x)$ is contractive: $\|f(x) - f(y)\| \leq \rho\|x - y\|$, for all x, y, where $0 \leq \rho < 1$. Then the iteration $x_{k+1} = f(x_k)$ is convergent and the convergence rate is linear (Kreyszig, 1978). If $f(x)$ is continuously differentiable, then $\|\nabla f(x)\| \leq \rho$ ensures that $f(x)$ is contractive.

Now we need to estimate the Lipschitz coefficient ρ for the mapping $X^{i,t+1} = f(X^{i,t}) = \phi\left(W^{i-1}X^{i-1} - \frac{\mu_{i+1}}{\mu_i}(W^i)^T(\phi(W^iX^i) - X^{i+1})\right)$. Its Jacobian matrix is:

$$\begin{aligned}
J_{kl,pq} &= \frac{\partial [f(X^{i,t})]_{kl}}{\partial X^{i,t}_{pq}} \\
&= \frac{\partial \phi\left([W^{i-1}X^{i-1}]_{kl} - \frac{\mu_{i+1}}{\mu_i}[(W^i)^T(\phi(W^iX^{i,t}) - X^{i+1})]_{kl}\right)}{\partial X^{i,t}_{pq}} \\
&= -\frac{\mu_{i+1}}{\mu_i}\phi'(c^{i,t}_{kl})\frac{\partial [(W^i)^T(\phi(W^iX^{i,t}) - X^{i+1})]_{kl}}{\partial X^{i,t}_{pq}} \\
&= -\frac{\mu_{i+1}}{\mu_i}\phi'(c^{i,t}_{kl})\frac{\partial \sum_r W^i_{rk}[\phi\left((W^iX^{i,t})_{rl}\right) - X^{i+1}_{rl}]}{\partial X^{i,t}_{pq}} \\
&= -\frac{\mu_{i+1}}{\mu_i}\phi'(c^{i,t}_{kl})\sum_r W^i_{rk}\phi'((W^iX^{i,t})_{rl})\frac{\partial (W^iX^{i,t})_{rl}}{\partial X^{i,t}_{pq}} \\
&= -\frac{\mu_{i+1}}{\mu_i}\phi'(c^{i,t}_{kl})\sum_r W^i_{rk}\phi'((W^iX^{i,t})_{rl})\frac{\partial \sum_s W^i_{rs}X^{i,t}_{sl}}{\partial X^{i,t}_{pq}} \\
&= -\frac{\mu_{i+1}}{\mu_i}\phi'(c^{i,t}_{kl})\sum_r W^i_{rk}\phi'((W^iX^{i,t})_{rl})\sum_s W^i_{rs}\delta_{sp}\delta_{lq} \\
&= -\frac{\mu_{i+1}}{\mu_i}\phi'(c^{i,t}_{kl})\sum_r W^i_{rk}\phi'((W^iX^{i,t})_{rl})W^i_{rp}\delta_{lq},
\end{aligned} \tag{5}$$

where $c^{i,t}_{kl} = [W^{i-1}X^{i-1}]_{kl} - \frac{\mu_{i+1}}{\mu_i}[(W^i)^T(\phi(W^iX^{i,t}) - X^{i+1})]_{kl}$, δ_{sp} is the Kronecker delta function, it is 1 if s and p are equal, and 0 otherwise. Its l_1 norm is upper bounded by:

$$\begin{aligned}
\|J\|_1 &= \max_{pq}\sum_{kl}|J_{kl,pq}| \\
&= \frac{\mu_{i+1}}{\mu_i}\max_{pq}\sum_{kl}\left|\phi'(c^{i,t}_{kl})\sum_r W^i_{rk}\phi'((W^iX^{i,t})_{rl})W^i_{rp}\delta_{lq}\right| \\
&\leq \frac{\mu_{i+1}}{\mu_i}\gamma^2\max_p\sum_k\sum_r |W^i_{rk}||W^i_{rp}| \\
&\leq \frac{\mu_{i+1}}{\mu_i}\gamma^2\max_p\sum_k\left(|(W^i)^T||W^i|\right)_{kp} \\
&= \frac{\mu_{i+1}}{\mu_i}\gamma^2\left\||(W^i)^T||W^i|\right\|_1.
\end{aligned} \tag{6}$$

Its l_∞ norm is upper bounded by

$$\begin{aligned}
\|J\|_\infty &= \max_{kl}\sum_{pq}|J_{kl,pq}| \\
&= \frac{\mu_{i+1}}{\mu_i}\max_{kl}\sum_{pq}\left|\phi'(c^{i,t}_{kl})\sum_r W^i_{rk}\phi'((W^iX^{i,t})_{rl})W^i_{rp}\delta_{lq}\right| \\
&\leq \frac{\mu_{i+1}}{\mu_i}\gamma^2\max_k\sum_p\sum_r |W^i_{rk}||W^i_{rp}| \\
&\leq \frac{\mu_{i+1}}{\mu_i}\gamma^2\max_k\sum_p\left(|(W^i)^T||W^i|\right)_{kp} \\
&= \frac{\mu_{i+1}}{\mu_i}\gamma^2\left\||(W^i)^T||W^i|\right\|_\infty.
\end{aligned} \tag{7}$$

Therefore, by using $\|A\|_2 \leq \sqrt{\|A\|_1\|A\|_\infty}$ (Golub and Van Loan, 2012), the l_2 norm of its Jacobian matrix is upper bounded by

$$\|J\|_2 \leq \frac{\mu_{i+1}}{\mu_i}\gamma^2\sqrt{\||(W^i)^T||W^i|\|_1\,\||(W^i)^T||W^i|\|_\infty}, \tag{8}$$

which is the Lipschitz coefficient ρ.

Proof of Theorem 3

The proof of the first part is the same as that of Theorem 2. So we only detail how to estimate the Lipschitz coefficient τ for the mapping $X^{n,t+1} = f(X^{n,t}) = \phi\left(W^{n-1}X^{n-1} - \frac{1}{\mu_n}\frac{\partial \ell(X^{n,t}, L)}{\partial X^{n,t}}\right)$. Its Jacobian matrix is:

$$\begin{aligned}
J_{kl,pq} &= \frac{\partial [f(X^{n,t})]_{kl}}{\partial X^{n,t}_{pq}} \\
&= \frac{\partial \phi\left((W^{n-1}X^{n-1})_{kl} - \frac{1}{\mu_n}\frac{\partial \ell(X^{n,t}, L)}{\partial X^{n,t}_{kl}}\right)}{\partial X^{n,t}_{pq}} \\
&= -\frac{1}{\mu_n}\phi'(d^{n,t}_{kl})\frac{\partial \frac{\partial \ell(X^{n,t}, L)}{\partial X^{n,t}_{kl}}}{\partial X^{n,t}_{pq}} \\
&= -\frac{1}{\mu_n}\phi'(d^{n,t}_{kl})\frac{\partial^2 \ell(X^{n,t}, L)}{\partial X^{n,t}_{kl}\partial X^{n,t}_{pq}},
\end{aligned} \tag{9}$$

where $d_{kl}^{n,t} = (W^{n-1}X^{n-1})_{kl} - \frac{1}{\mu_n}\left(\frac{\partial \ell(X^{n,t},L)}{\partial X^{n,t}}\right)_{kl}$. Its l_1 norm is upper bounded by:

$$\begin{aligned}\|J\|_1 &= \max_{pq}\sum_{kl}|J_{kl,pq}| \\ &= \frac{1}{\mu_n}\max_{pq}\sum_{kl}\left|\phi'(d_{kl}^{n,t})\frac{\partial^2\ell(X^{n,t},L)}{\partial X_{kl}^{n,t}\partial X_{pq}^{n,t}}\right| \\ &\leq \frac{\gamma}{\mu_n}\max_{pq}\sum_{kl}\left|\frac{\partial^2\ell(X^{n,t},L)}{\partial X_{kl}^{n,t}\partial X_{pq}^{n,t}}\right| \\ &= \frac{\gamma}{\mu_n}\left\|\left|\frac{\partial^2\ell(X^{n,t},L)}{\partial X_{kl}^{n,t}\partial X_{pq}^{n,t}}\right|\right\|_1 \\ &\leq \frac{\gamma\eta}{\mu_n}.\end{aligned} \tag{10}$$

Its l_∞ norm is upper bounded by:

$$\begin{aligned}\|J\|_\infty &= \max_{kl}\sum_{pq}|J_{kl,pq}| \\ &= \frac{1}{\mu_n}\max_{kl}\sum_{pq}\left|\phi'(d_{kl}^{n,t})\frac{\partial^2\ell(X^{n,t},L)}{\partial X_{kl}^{n,t}\partial X_{pq}^{n,t}}\right| \\ &\leq \frac{\gamma}{\mu_n}\max_{kl}\sum_{pq}\left|\frac{\partial^2\ell(X^{n,t},L)}{\partial X_{kl}^{n,t}\partial X_{pq}^{n,t}}\right| \\ &= \frac{\gamma}{\mu_n}\left\|\left|\frac{\partial^2\ell(X^{n,t},L)}{\partial X_{kl}^{n,t}\partial X_{pq}^{n,t}}\right|\right\|_1 \\ &\leq \frac{\gamma\eta}{\mu_n}.\end{aligned} \tag{11}$$

Therefore, the l_2 norm of J is upper bounded by

$$\|J\|_2 \leq \sqrt{\|J\|_1\|J\|_\infty} \leq \frac{\gamma\eta}{\mu_n} = \tau. \tag{12}$$

Proof of Theorem 4

The L_φ-smoothness of φ:

$$\|\nabla\varphi(x)-\nabla\varphi(y)\| \leq L_\varphi\|x-y\|, \forall x, y$$

enables the following inequality (Nesterov, 2004):

$$\varphi(z) \leq \varphi(y) + \langle\nabla\varphi(y), z-y\rangle + \frac{L_\varphi}{2}\|z-y\|^2, \forall x, y. \tag{13}$$

By putting $z = Ax$ and $y = Ay_k$, where y_k is yet to be chosen, we have

$$\varphi(Ax) \leq \varphi(Ay_k) + \langle\nabla\varphi(Ay_k), A(x-y_k)\rangle + \frac{L_\varphi}{2}\|A(x-y_k)\|^2. \tag{14}$$

As assumed,

$$x_{k+1} = \operatorname*{argmin}_x \langle\nabla\varphi(Ay_k), A(x-y_k)\rangle + \frac{L_\varphi}{2}\|A(x-y_k)\|^2 + h(x) \tag{15}$$

is easy to solve. This gives

$$-L_\varphi A^T A(x_{k+1}-y_k) \in A^T\nabla\varphi(Ay_k) + \partial h(x_{k+1}). \tag{16}$$

Then by (14) and the convexity of h, we have

$$\begin{aligned}&F(x_{k+1}) = \varphi(Ax_{k+1}) + h(x_{k+1}) \\ &\leq \varphi(Ay_k) + \langle\nabla\varphi(Ay_k), A(x_{k+1}-y_k)\rangle + \frac{L_\varphi}{2}\|A(x_{k+1}-y_k)\|^2 \\ &\quad + h(u) - \langle\xi, u-x_{k+1}\rangle \\ &\leq \varphi(Au) + \langle\nabla\varphi(Ay_k), A(u-y_k)\rangle + \langle\nabla\varphi(Ay_k), A(x_{k+1}-y_k)\rangle \\ &\quad + \frac{L_\varphi}{2}\|A(x_{k+1}-y_k)\|^2 + h(u) - \langle\xi, u-x_{k+1}\rangle \\ &= F(u) - \langle A^T\nabla\varphi(Ay_k) + \xi, u-x_{k+1}\rangle + \frac{L_\varphi}{2}\|A(x_{k+1}-y_k)\|^2 \\ &= F(u) + L_\varphi\langle A^TA(x_{k+1}-y_k), u-x_{k+1}\rangle + \frac{L_\varphi}{2}\|A(x_{k+1}-y_k)\|^2 \\ &= F(u) + L_\varphi\langle A(x_{k+1}-y_k), A(u-x_{k+1})\rangle + \frac{L_\varphi}{2}\|A(x_{k+1}-y_k)\|^2,\end{aligned} \tag{17}$$

where ξ is any subgradient in $\partial h(x_{k+1})$, u is any point, and the third equality used (16). Thus

$$\begin{aligned}F(x_{k+1}) \leq\ & F(u) + L_g\langle A(x_{k+1}-y_k), A(u-x_{k+1})\rangle \\ & + \frac{L_g}{2}\|A(x_{k+1}-y_k)\|^2, \quad \forall u.\end{aligned} \tag{18}$$

Let $u = x_k$ and $u = x^*$ in (18), respectively. Then multiplying the first inequality with θ_k and the second with $1-\theta_k$ and adding them together, we have

$$\begin{aligned}&F(x_{k+1}) \leq \theta_k F(x_k) + (1-\theta_k)F(x^*) \\ &\quad + L_\varphi\langle A(x_{k+1}-y_k), A[\theta_k(x_k-x_{k+1}) + (1-\theta_k)(x^*-x_{k+1})]\rangle \\ &\quad + \frac{L_\varphi}{2}\|A(x_{k+1}-y_k)\|^2 \\ &= \theta_k F(x_k) + (1-\theta_k)F(x^*) \\ &\quad + L_\varphi\langle A(x_{k+1}-y_k), A[\theta_k x_k - x_{k+1} + (1-\theta_k)x^*]\rangle \\ &\quad + \frac{L_\varphi}{2}\|A(x_{k+1}-y_k)\|^2 \\ &= \theta_k F(x_k) + (1-\theta_k)F(x^*) \\ &\quad + \frac{L_\varphi}{2}\left\{\|A[(x_{k+1}-y_k) + (\theta_k x_k - x_{k+1} + (1-\theta_k)x^*)]\|^2\right. \\ &\quad \left. - \|A(x_{k+1}-y_k)\|^2 - \|A[\theta_k x_k - x_{k+1} + (1-\theta_k)x^*]\|^2\right\} \\ &\quad + \frac{L_\varphi}{2}\|A(x_{k+1}-y_k)\|^2 \\ &= \theta_k F(x_k) + (1-\theta_k)F(x^*) \\ &\quad + \frac{L_\varphi}{2}\left\{\|A[\theta_k x_k - y_k + (1-\theta_k)x^*]\|^2\right. \\ &\quad \left. - \|A[\theta_k x_k - x_{k+1} + (1-\theta_k)x^*]\|^2\right\}.\end{aligned} \tag{19}$$

In order to have a recursion, we need to have:

$$\theta_k x_k - y_k + (1-\theta_k)x^* = \sqrt{\theta_k}[\theta_{k-1}x_{k-1} - x_k + (1-\theta_{k-1})x^*].$$

By comparing the coefficient of x^*, we have

$$1-\theta_k = \sqrt{\theta_k}(1-\theta_{k-1}). \tag{20}$$

Accordingly,

$$y_k = \theta_k x_k - \sqrt{\theta_k}(\theta_{k-1}x_{k-1} - x_k). \tag{21}$$

With the above choice of $\{\theta_k\}$ and y_k, (19) can be rewritten as

$$\begin{aligned} &F(x_{k+1})-F(x^*)+\frac{L_\varphi}{2}\|z_{k+1}\|^2 \\ \leq &\theta_k\left(F(x_k)-F(x^*)+\frac{L_\varphi}{2}\|z_k\|^2\right), \end{aligned} \tag{22}$$

where $z_k = A[\theta_{k-1}x_{k-1} - x_k + (1-\theta_{k-1})x^*]$. Then by recursion, we have

$$\begin{aligned} &F(x_k)-F(x^*)+\frac{L_\varphi}{2}\|z_k\|^2 \\ \leq &\left(\prod_{i=1}^{k-1}\theta_i\right)\left(F(x_1)-F(x^*)+\frac{L_\varphi}{2}\|z_1\|^2\right). \end{aligned} \tag{23}$$

It remains to estimate $\prod_{i=1}^{k-1}\theta_i$. We choose $\theta_0 = 0$ and prove

$$1-\theta_k < \frac{2}{k+1} \tag{24}$$

by induction. (24) is true for $k=0$. Suppose (24) is true for $k-1$, then by $1-\theta_k = \sqrt{\theta_k}(1-\theta_{k-1})$, we have

$$1-\theta_k = \sqrt{\theta_k}(1-\theta_{k-1}) < \sqrt{\theta_k}\frac{2}{k}. \tag{25}$$

Let $\tilde{\theta}_k = 1-\theta_k$, then the above becomes $k^2\tilde{\theta}_k^2 < 4(1-\tilde{\theta}_k)$. So

$$\tilde{\theta}_k < \frac{-4+\sqrt{16+16k^2}}{2k^2} = \frac{2}{1+\sqrt{1+k^2}} < \frac{2}{k+1}. \tag{26}$$

Thus (24) is proven.

Now we are ready to estimate $\prod_{i=1}^{k-1}\theta_i$. From $1-\theta_k = \sqrt{\theta_k}(1-\theta_{k-1})$, we have

$$1-\theta_{k-1} = \sqrt{\prod_{i=1}^{k-1}\theta_i(1-\theta_0)} = \sqrt{\prod_{i=1}^{k-1}\theta_i}.$$

So $\prod_{i=1}^{k-1}\theta_i = (1-\theta_{k-1})^2 < \frac{4}{k^2}$. Hence

$$F(x_k)-F(x^*)+\frac{L_\varphi}{2}\|z_k\|^2 \leq \frac{4}{k^2}\left(F(x_1)-F(x^*)+\frac{L_\varphi}{2}\|z_1\|^2\right).$$

The three equations, (20), (21), and (15) constitute the major steps in Algorithm 2.

Convergence Analysis of Algorithm 1

If the loss function is differentiable and both ϕ and ϕ^{-1} are strictly increasing, then the objective function of LPOM is differentiable and the block coordinate descent in Algorithm 1 converges to stationary points by subsequence (Bertsekas, 1999). The results in (Xu and Yin, 2013) can also be applied to obtain the convergence of Algorithm 1.

References

Bertsekas, D. P. 1999. *Nonlinear Programming: 2nd Edition*. Athena Scientific.

Golub, G. H., and Van Loan, C. F. 2012. *Matrix Computations*, volume 3. The Johns Hopkins University Press.

Kreyszig, E. 1978. *Introductory Functional Analysis with Applications*, volume 1. Wiley New York.

Nesterov, Y. 2004. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer.

Xu, Y., and Yin, W. 2013. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences* 6(3):1758–1789.

Zeng, J.; Ouyang, S.; Lau, T. T.-K.; Lin, S.; and Yao, Y. 2018. Global convergence in deep learning with variable splitting via the Kurdyka-łojasiewicz property. *arXiv preprint arXiv:1803.00225*.