Supplementary Materials

A Proof of Theorem 1

Theorem 1 (Asymptotic consistency)**.** *Let Assumption 1 and 2 hold, and apply* msPG *to problem* (P)*. If the step size* $\eta < (L_f + 2Ls)^{-1}$*, then the global model and local models satisfy:*

1. $\sum_{t=0}^{\infty} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 < \infty$;
2. $\lim_{t\to\infty} \|\mathbf{x}(t+1) - \mathbf{x}(t)\| = 0$, $\lim_{t\to\infty} \|\mathbf{x}(t) - \mathbf{x}^i(t)\| = 0$;
3. *The limit points* $\omega(\{\mathbf{x}(t)\}) = \omega(\{\mathbf{x}^i(t)\}) \subseteq \operatorname{crit} F$.

Proof. We start from bounding the difference between the global model $\mathbf{x}$ and the local model $\mathbf{x}^i$ (on any machine i). Indeed, at iteration t, by the definition of the global and local models in msPG:

$$\|\mathbf{x}(t) - \mathbf{x}^i(t)\| = \sqrt{\sum_{j=1}^{p} \|x_j(t) - x_j(\tau_j^i(t))\|^2} \tag{21}$$

$$(\text{ triangle inequality }) \le \sqrt{\sum_{j=1}^{p} \left(\sum_{k=\tau_j^i(t)}^{t-1} \|x_j(k+1) - x_j(k)\| \right)^2} \tag{22}$$

$$(\text{ Assumption 2.1 }) \le \sqrt{\sum_{j=1}^{p} \left(\sum_{k=(t-s)_+}^{t-1} \|x_j(k+1) - x_j(k)\| \right)^2} \tag{23}$$

$$= \left\| \Big(\sum_{k=(t-s)_+}^{t-1} \|x_1(k+1) - x_1(k)\|, \ \cdots, \ \sum_{k=(t-s)_+}^{t-1} \|x_p(k+1) - x_p(k)\| \Big) \right\| \tag{24}$$

$$= \left\| \sum_{k=(t-s)_+}^{t-1} \big(\|x_1(k+1) - x_1(k)\|, \ \cdots, \ \|x_p(k+1) - x_p(k)\| \big) \right\| \tag{25}$$

$$(\text{ triangle inequality }) \le \sum_{k=(t-s)_+}^{t-1} \left\| \big(\|x_1(k+1) - x_1(k)\|, \ \cdots, \ \|x_p(k+1) - x_p(k)\| \big) \right\| \tag{26}$$

$$= \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|, \tag{27}$$

where in the last equality we used the following property of the Euclidean norm:

$$\|\mathbf{x}\| = \|(x_1, \ \ldots, \ x_p)\| = \|(\|x_1\|, \ \ldots, \ \|x_p\|)\|. \tag{28}$$

Equation (27) bounds the inconsistency between the global model and the local models. We will repeatedly use it in the following, as it provides a bridge to jump from the global model and the local models back and forth.

Next we bound the progress of the global model $\mathbf{x}(t)$. If $t \in T_i$ (i.e., machine i updates at iteration t), then using the definition of the update operator $U_i(\mathbf{x}^i(t))$ in Equation (9) we can rewrite $x_i(t+1)$ as

$$x_i(t+1) = \operatorname{prox}_{g_i}^{\eta}\big(x_i(t) - \eta \nabla_i f(\mathbf{x}^i(t))\big), \tag{29}$$

where we recall the proximal map $\operatorname{prox}_{g_i}^{\eta}$ from Definition 3. Thus, for all $z \in \mathbb{R}^{d_i}$:

$$g_i\big(x_i(t+1)\big) + \frac{1}{2\eta} \left\| x_i(t+1) - x_i(t) + \eta \nabla_i f\big(\mathbf{x}^i(t)\big) \right\|^2 \le g_i\big(z\big) + \frac{1}{2\eta} \left\| z - x_i(t) + \eta \nabla_i f\big(\mathbf{x}^i(t)\big) \right\|^2. \tag{30}$$

Substituting with $z = x_i(t)$ and simplifying yields

$$g_i\big(x_i(t+1)\big) - g_i\big(x_i(t)\big) \le -\frac{1}{2\eta} \|x_i(t+1) - x_i(t)\|^2 - \left\langle \nabla_i f\big(\mathbf{x}^i(t)\big), x_i(t+1) - x_i(t) \right\rangle. \tag{31}$$

(If g_i is convex, we can replace $\frac{1}{2\eta}$ with $\frac{1}{\eta}$.) Note that if $t \notin T_i$, then $x_i(t+1) = x_i(t)$ hence Equation (31) still trivially holds. On the other hand, Assumption 1.2 implies

$$f\big(\mathbf{x}(t+1)\big) - f\big(\mathbf{x}(t)\big) \le \langle \mathbf{x}(t+1) - \mathbf{x}(t), \nabla f\big(\mathbf{x}(t)\big)\rangle + \frac{L_f}{2}\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2. \tag{32}$$

Adding Equation (32) and Equation (31) (for all i) and recalling $F(\mathbf{x}) = f(\mathbf{x}) + \sum_i g_i(x_i)$, we have

$$F\big(\mathbf{x}(t+1)\big) - F\big(\mathbf{x}(t)\big) \le \tfrac{1}{2}(L_f - 1/\eta)\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 + \sum_{i=1}^{p} \big\langle x_i(t+1) - x_i(t), \nabla_i f(\mathbf{x}(t)) - \nabla_i f\big(\mathbf{x}^i(t)\big)\big\rangle \tag{33}$$

$$(\text{ Cauchy-Schwarz }) \le \tfrac{1}{2}(L_f - 1/\eta)\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 + \sum_{i=1}^{p} \|x_i(t+1) - x_i(t)\| \cdot \|\nabla_i f(\mathbf{x}(t)) - \nabla_i f\big(\mathbf{x}^i(t)\big)\| \tag{34}$$

$$(\text{ Assumption 1.2 }) \le \tfrac{1}{2}(L_f - 1/\eta)\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 + \sum_{i=1}^{p} \|x_i(t+1) - x_i(t)\| \cdot L_i\|\mathbf{x}(t) - \mathbf{x}^i(t)\| \tag{35}$$

$$(\text{ Equation (27) }) \le \tfrac{1}{2}(L_f - 1/\eta)\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 + \sum_{i=1}^{p} L_i\|x_i(t+1) - x_i(t)\| \cdot \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \tag{36}$$

$$(\text{ Assumption 1.2 }) \le \tfrac{1}{2}(L_f - 1/\eta)\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 + L\|\mathbf{x}(t+1) - \mathbf{x}(t)\| \cdot \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \tag{37}$$

$$(\ ab \le \tfrac{a^2+b^2}{2}\ ) \le \tfrac{1}{2}(L_f - 1/\eta)\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 + \frac{L}{2}\sum_{k=(t-s)_+}^{t-1} \Big[\|\mathbf{x}(k+1) - \mathbf{x}(k)\|^2 + \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2\Big] \tag{38}$$

$$\le \tfrac{1}{2}(L_f + Ls - 1/\eta)\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 + \frac{L}{2}\sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|^2. \tag{39}$$

Summing the above inequality from m to $n-1$ we have

$$F\big(\mathbf{x}(n)\big) - F\big(\mathbf{x}(m)\big) \le \tfrac{1}{2}(L_f + Ls - 1/\eta)\sum_{t=m}^{n-1} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 + \frac{L}{2}\sum_{t=m}^{n-1}\sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|^2 \tag{40}$$

$$\le \tfrac{1}{2}(L_f + 2Ls - 1/\eta)\sum_{t=m}^{n-1} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2. \tag{41}$$

Therefore, as long as $\eta < 1/(L_f + 2Ls)$, letting $m = 0$ we deduce

$$\sum_{t=0}^{n-1} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \le \frac{2}{1/\eta - L_f - 2Ls}[F\big(\mathbf{x}(0)\big) - F\big(\mathbf{x}(n)\big)] \le \frac{2}{1/\eta - L_f - 2Ls}[F\big(\mathbf{x}(0)\big) - \inf_{\mathbf{z}} F(\mathbf{z})]. \tag{42}$$

By Assumption 1.1, F is bounded from below hence the right-hand side is finite and independent of n. Letting n goes to infinity completes the proof of item 1.

Item 2 follows immediately from item 1 and (27), whence it is clear that for all i the limit points satisfy $\omega(\{\mathbf{x}(k)\}) = \omega(\{\mathbf{x}^i(k)\})$.

To prove item 3, let $\mathbf{x}^*$ be a limit point of $\{\mathbf{x}(t)\}_t$, i.e., there exists a subsequence $\mathbf{x}(t_m) \to \mathbf{x}^*$. Since the objective function F is closed we know $\mathbf{x}^* \in \operatorname{dom} F$. To show $\mathbf{x}^* \in \operatorname{crit} F$ we need to exhibit a sequence say $\mathbf{x}(k_m + 1)$ such that[3]

$$\mathbf{x}(k_m+1) \to \mathbf{x}^*,\ F(\mathbf{x}(k_m+1)) \to F(\mathbf{x}^*),\ \mathbf{0} \leftarrow \mathbf{u}(k_m+1) \in \partial F(\mathbf{x}(k_m+1)). \tag{43}$$

[3]Technically, from Definition 1 we should have the Frechét subdifferential $\hat{\partial} F$ in Equation (43), however, a usual diagonal argument allows us to use the more convenient (limiting) subdifferential.

This is the most difficult part of the proof, and the argument differs substantially from previous work (e.g. [8]).

We first prove the subdifferential goes to zero. Observe from Assumption 2.3 that the iterations $\{t, t \in T_i\}$ when machine i updates is infinite. Let $\hat{t} \in T_i$ and by the optimality condition of $x_i(\hat{t}+1)$ in Equation (29):

$$-\tfrac{1}{\eta}\left[x_i(\hat{t}+1) - x_i(\hat{t}) + \eta \nabla_i f\big(\mathbf{x}^i(\hat{t})\big)\right] \in \partial g_i(x_i(\hat{t}+1)), \tag{44}$$

i.e. there exists $u_i(\hat{t}+1) \in \partial g_i(x_i(\hat{t}+1))$ such that

$$\|u_i(\hat{t}+1) + \nabla_i f(\mathbf{x}(\hat{t}+1))\| \leq \|u_i(\hat{t}+1) + \nabla_i f(\mathbf{x}(\hat{t}))\| + \|\nabla_i f(\mathbf{x}(\hat{t}+1)) - \nabla_i f(\mathbf{x}(\hat{t}))\| \tag{45}$$

$$(\text{ Equation (44), Assumption 1.2 }) \leq \left\|\tfrac{1}{\eta}(x_i(\hat{t}+1) - x_i(\hat{t}) + \nabla_i f\big(\mathbf{x}^i(\hat{t})\big) - \nabla_i f\big(\mathbf{x}(\hat{t})\big)\right\| + L_i\|\mathbf{x}(\hat{t}+1) - \mathbf{x}(\hat{t})\| \tag{46}$$

$$(\text{ triangle inequality, Assumption 1.2 }) \leq \tfrac{1}{\eta}\|x_i(\hat{t}+1) - x_i(\hat{t})\| + L_i\|\mathbf{x}^i(\hat{t}) - \mathbf{x}(\hat{t})\| + L_i\|\mathbf{x}(\hat{t}+1) - \mathbf{x}(\hat{t})\| \tag{47}$$

$$(\text{ Equation (27) }) \leq \tfrac{1}{\eta}\|x_i(\hat{t}+1) - x_i(\hat{t})\| + L_i \sum_{k=(\hat{t}-s)_+}^{\hat{t}} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|. \tag{48}$$

We now use a chaining argument to remove the condition $\hat{t} \in T_i$ above. For each $t \notin T_i$ let $\hat{t}_i$ be the *largest* element in $\{k \leq t : k \in T_i\}$. Thanks to Assumption 2.3 $\hat{t}_i$ always exists and $t - \hat{t}_i \leq s$. Therefore, for any $t \notin T_i$, since $x_i(t+1) = x_i(\hat{t}_i+1)$ we can certainly choose $u_i(t+1) \in \partial g_i(x_i(t+1))$ to coincide with $u_i(\hat{t}_i+1) \in \partial g_i(x_i(\hat{t}_i+1))$. Then:

$$\|u_i(t+1) + \nabla_i f(\mathbf{x}(t+1)) - u_i(\hat{t}_i+1) - \nabla_i f(\mathbf{x}(\hat{t}_i+1))\| = \|\nabla_i f(\mathbf{x}(t+1)) - \nabla_i f(\mathbf{x}(\hat{t}_i+1))\| \tag{49}$$

$$(\text{ triangle inequality }) \leq \sum_{k=\hat{t}_i+1}^{t} \|\nabla_i f(\mathbf{x}(k+1)) - \nabla_i f(\mathbf{x}(k))\| \tag{50}$$

$$(\text{ Assumption 2.3 }) \leq \sum_{k=(t-s+1)_+}^{t} \|\nabla_i f(\mathbf{x}(k+1)) - \nabla_i f(\mathbf{x}(k))\| \tag{51}$$

$$(\text{ Assumption 1.2 }) \leq L_i \sum_{k=(t-s+1)_+}^{t} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|. \tag{52}$$

Combining the two separate cases in Equation (48) and Equation (52) above we have for all t:

$$\|\mathbf{u}(t+1) + \nabla f(\mathbf{x}(t+1))\| \leq (\sqrt{p}/\eta + 2L) \sum_{k=(t-2s)_+}^{t} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|, \tag{53}$$

where of course $\mathbf{u}(t+1) = \big(u_1(t+1),\ \ldots\ ,\ u_p(t+1)\big) \in \partial g(\mathbf{x}(t+1))$ and we artificially introduce $\sqrt{p}$ for convenience of subsequent proof. Therefore, from item 2 we deduce

$$\lim_{t\to\infty} \operatorname{dist}_{\partial F(\mathbf{x}(t+1))}(\mathbf{0}) \to 0. \tag{54}$$

Next we deal with the function value convergence in Equation (43). For any $\hat{t}_i \in T_i$, using Equation (30) with $z = x_i^*$ we have

$$g_i(x_i(\hat{t}_i+1)) + \frac{1}{2\eta}\left\|x_i(\hat{t}_i+1) - x_i(\hat{t}_i) + \eta\nabla_i f\big(\mathbf{x}^i(\hat{t}_i)\big)\right\|^2 \leq g_i\big(x_i^*\big) + \frac{1}{2\eta}\left\|x_i^* - x_i(\hat{t}_i) + \eta\nabla_i f\big(\mathbf{x}^i(\hat{t}_i)\big)\right\|^2, \tag{55}$$

which, after rearrangement, yields

$$g_i(x_i(\hat{t}_i+1)) \leq g_i\big(x_i^*\big) + \frac{1}{2\eta}\|x_i^* - x_i(\hat{t}_i)\|^2 - \frac{1}{2\eta}\|x_i(\hat{t}_i+1) - x_i(\hat{t}_i)\|^2 + \langle x_i^* - x_i(\hat{t}_i+1), \nabla_i f(\mathbf{x}^i(\hat{t}_i))\rangle \tag{56}$$

$$= g_i\big(x_i^*\big) + \frac{1}{2\eta}\|x_i^* - x_i(\hat{t}_i)\|^2 - \frac{1}{2\eta}\|x_i(\hat{t}_i+1) - x_i(\hat{t}_i)\|^2 + \langle x_i^* - x_i(\hat{t}_i+1), \nabla_i f(\mathbf{x}^*)\rangle$$
$$+ \langle x_i^* - x_i(\hat{t}_i+1), \nabla_i f(\mathbf{x}^i(\hat{t}_i)) - \nabla_i f(\mathbf{x}^*)\rangle. \tag{57}$$

We wish to deduce from the above inequality that $g_i(x_i(\hat{t}_i+1)) \to g(x_i^*)$, but we need a uniformization device to remove the dependence on i (hence removing the condition $\hat{t}_i \in T_i$). Observing from item 2 that

$$\lim_{m\to\infty} \max_{t\in[t_m-s,t_m+s]} \|\mathbf{x}(t+1)-\mathbf{x}^*\| \to 0, \qquad \lim_{m\to\infty} \max_{t\in[t_m-s,t_m+s]} \|\mathbf{x}^i(t+1)-\mathbf{x}^*\| \to 0. \tag{58}$$

By Assumption 2.3, $[t_m-s,t_m+s]\cap T_i \neq \emptyset$ for all i, using item 2 again and the Lipschitz continuity of ∇f, we deduce from Equation (57) that

$$\limsup_{m\to\infty} \max_{t\in[t_m-s,t_m+s]\cap T_i} g_i(x_i(t+1)) \leq g_i(x_i^*). \tag{59}$$

Since each machine must update at least once on the intervals $[t_m-s,t_m]$ and $[t_m,t_m+s]$, let $\hat{t}_m^i$ be the largest element of $[t_m-s,t_m]\cap T_i$. Then from the previous inequality we have

$$\limsup_{m\to\infty} \max_{t\in[\hat{t}_m^i,t_m+s]\cap T_i} g_i(x_i(t+1)) \leq g_i(x_i^*). \tag{60}$$

Since $g_i(x_i(t+1)) = g_i(x_i(t))$ if $t \notin T_i$ and $\hat{t}_m^i \in T_i$, it follows that

$$\max_{t\in[t_m,t_m+s]} g_i(x_i(t+1)) \leq \max_{t\in[\hat{t}_m^i,t_m+s]\cap T_i} g_i(x_i(t+1)), \tag{61}$$

hence

$$\limsup_{m\to\infty} \max_{t\in[t_m,t_m+s]} g_i(x_i(t+1)) \leq g_i(x_i^*). \tag{62}$$

Choose any sequence k_m such that $k_m \in [t_m,t_m+s]$. Since $\mathbf{x}(t_m)\to\mathbf{x}^*$, from item 2 it is clear that

$$\mathbf{x}(k_m+1)\to\mathbf{x}^*. \tag{63}$$

From Equation (62) we know for all i, $\limsup_{m\to\infty} g_i(x_i(k_m+1)) \leq g_i(x_i^*)$ while using closedness of the function g_i we have $\liminf_{m\to\infty} g_i(x_i(k_m+1)) \geq g_i(x_i^*)$, thus in fact $\lim_{m\to\infty} g_i(x_i(k_m+1)) = g_i(x_i^*)$. Since f is continuous, we know

$$\lim_{m\to\infty} F(\mathbf{x}(k_m+1)) = F(\mathbf{x}^*). \tag{64}$$

Lastly, combining Equation (54), Equation (63) and Equation (64), it follows from Definition 1 that $\mathbf{x}^* \in \operatorname{crit} F$. □

B Proof of Theorem 2

Theorem 2 (Finite Length)**.** *Let Assumption 1, 2, 3 and 4 hold, and apply* msPG *to problem* (P)*. If the step size* $\eta < (L_f+2Ls)^{-1}$ *and* $\{\mathbf{x}(t)\}$ *is bounded, then*

$$\sum_{t=0}^{\infty} \|\mathbf{x}(t+1)-\mathbf{x}(t)\| < \infty, \tag{11}$$

$$\forall i=1,\dots,p,\ \ \sum_{t=0}^{\infty} \|\mathbf{x}^i(t+1)-\mathbf{x}^i(t)\| < \infty. \tag{12}$$

Furthermore, $\{\mathbf{x}(t)\}$ *and* $\{\mathbf{x}^i(t)\}, i=1,...,p$*, converge to the same critical point of* F*.*

Our proof requires the following simple uniformization of the KŁ inequality in Definition 4:

Lemma 2 (Uniformized KŁ inequality, [11, Lemma 6])**.** *Let* h *be a KŁ function and* $\Omega \subset \operatorname{dom} h$ *be a compact set. If* h *is constant on* Ω*, then there exist* $\varepsilon,\lambda>0$ *and a function* φ *as in Definition 4, such that for all* $\bar{\mathbf{x}}\in\Omega$ *and all* $\mathbf{x}\in\{\mathbf{x}\in\mathbb{R}^d : \operatorname{dist}_\Omega(\mathbf{x})<\varepsilon\}\cap[\mathbf{x}: h(\bar{\mathbf{x}})<h(\mathbf{x})<h(\bar{\mathbf{x}})+\lambda]$*, one has*

$$\varphi'(h(\mathbf{x})-h(\bar{\mathbf{x}}))\cdot \operatorname{dist}_{\partial h(\mathbf{x})}(\mathbf{0}) \geq 1.$$

The proof of this lemma is the usual covering argument.

Proof of Theorem 2. We first show that if the global sequence has finite length (i.e. (11)) then the local sequences also have finite lenggth (i.e. (12)). Indeed,

$$\|\mathbf{x}^i(t+1)-\mathbf{x}^i(t)\| \le \|\mathbf{x}^i(t+1)-\mathbf{x}(t+1)\| + \|\mathbf{x}(t+1)-\mathbf{x}(t)\| + \|\mathbf{x}(t)-\mathbf{x}^i(t)\| \tag{65}$$

$$(\text{ Equation (27) }) \le \|\mathbf{x}(t+1)-\mathbf{x}(t)\| \ + \sum_{k=(t+1-s)_+}^{t} \|\mathbf{x}(k+1)-\mathbf{x}(k)\| \ + \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1)-\mathbf{x}(k)\|. \tag{66}$$

Therefore, summing from $t=0$ to $t=n$:

$$\sum_{t=0}^{n} \|\mathbf{x}^i(t+1)-\mathbf{x}^i(t)\| \le \sum_{t=0}^{n} \left[\sum_{k=(t+1-s)_+}^{t} \|\mathbf{x}(k+1)-\mathbf{x}(k)\| \ + \sum_{k=(t-s)_+}^{t} \|\mathbf{x}(k+1)-\mathbf{x}(k)\| \right] \tag{67}$$

$$\le (2s+1)\sum_{t=0}^{n} \|\mathbf{x}(t+1)-\mathbf{x}(t)\|. \tag{68}$$

Letting n tend to infinity we have (11) $\implies$ (12).

From Theorem 1 we know the limit points of $\{\mathbf{x}(t)\}$ and $\{\mathbf{x}^i(t)\}, i=1,\dots,p$, coincide, and they are critical points of F.

The only thing left to prove is the finite length property of the global sequence $\mathbf{x}(t)$. If for all large t we have $\mathbf{x}(t+1)=\mathbf{x}(t)$ then the conclusion is trivial. On the other hand, we can remove all iterations t with $\mathbf{x}(t+1)=\mathbf{x}(t)$, without affecting the length of the trajectory. Thus, in the following we assume for all (large) t we have $\mathbf{x}(t+1)\neq\mathbf{x}(t)$. Thanks to Assumption 3 and Assumption 1.1, it is then clear that the objective value $F(\mathbf{x}(t))$ is strictly decreasing to a limit F^*. Since $\{\mathbf{x}(t)\}$ is assumed to be bounded, the limit point set $\Omega := \omega(\{\mathbf{x}(t)\})$ is nonempty and compact. Obviously for any $\mathbf{x}^* \in \Omega$ we have $F(\mathbf{x}^*)=F^*$. Fix any $\epsilon>0$, clearly for t sufficiently large we have[4] $\text{dist}_\Omega(\mathbf{x}(t)) \le \epsilon$. We now have all ingredients to apply the uniformized KŁ inequality in Lemma 2, which implies that for all sufficiently large t, there exists a continuous and concave function φ (with additional properties listed in Definition 4) such that

$$\varphi'\big(F(\mathbf{x}(t))-F^*\big)\cdot \text{dist}_{\partial F(\mathbf{x}(t))}(\mathbf{0}) \ge 1. \tag{69}$$

Since φ is concave, we obtain

$$\Delta_{t,t+1} := \varphi\big(F(\mathbf{x}(t))-F^*\big) - \varphi\big(F(\mathbf{x}(t+1))-F^*\big) \ge \varphi'\big(F(\mathbf{x}(t))-F^*\big)\big(F(\mathbf{x}(t))-F(\mathbf{x}(t+1))\big) \tag{70}$$

$$(\text{ Assumption 3 and Equation (69) }) \ge \frac{\alpha\|\mathbf{x}(t+1)-\mathbf{x}(t)\|^2}{\text{dist}_{\partial F(\mathbf{x}(t))}(\mathbf{0})}. \tag{71}$$

It is clear that the function φ (composed with F) serves as a Lyapunov function. To proceed, we need to upper bound the subdifferential $\partial F(\mathbf{x}(t))$, which has been painstakingly dealt with in the proof of Theorem 1.

Using the inequality $2\sqrt{ab} \le a+b$ for positive numbers we obtain from Equation (71): for t sufficiently large,

$$2\|\mathbf{x}(t+1)-\mathbf{x}(t)\| \le \tfrac{\delta}{\alpha}\Delta_{t,t+1} + \tfrac{1}{\delta}\text{dist}_{\partial F(\mathbf{x}(t))}(\mathbf{0}), \tag{72}$$

[4]This is true for any bounded sequence, and we provide a proof for completeness: Suppose not, then there exists $\epsilon>0$ such that for all n there exists a $t\ge n$ such that $\text{dist}_\Omega(\mathbf{x}(t))>\epsilon$. Thus, we can extract a subsequence $\{\mathbf{x}(t_m)\}$ such that $\text{dist}_\Omega(\mathbf{x}(t_m))>\epsilon$. However, since $\{\mathbf{x}(t)\}$ is bounded, we can extract a further subsequence, say $\{\mathbf{x}(t_{m_n})\}$, that converges, i.e. $\text{dist}_\Omega(\mathbf{x}(t_{m_n}))\to 0$, contradiction.

where $\delta > 0$ will be fixed later. Summing the above inequality over t from m (sufficiently large) to n:

$$2\sum_{t=m}^{n}\|\mathbf{x}(t+1)-\mathbf{x}(t)\| \le \sum_{t=m}^{n}\frac{\delta}{\alpha}\Delta_{t,t+1} + \sum_{t=m}^{n}\frac{1}{\delta}\mathrm{dist}_{\partial F(\mathbf{x}(t))}(\mathbf{0}) \tag{73}$$

$$(\text{ telescoping and Equation (53) }) \le \frac{\delta}{\alpha}\varphi\big(F(\mathbf{x}(m))-F^*\big) + \sum_{t=m}^{n}\frac{\sqrt{p}/\eta+2L}{\delta}\sum_{k=(t-2s)_+}^{t}\|\mathbf{x}(k+1)-\mathbf{x}(k)\| \tag{74}$$

$$\le \frac{\delta}{\alpha}\varphi\big(F(\mathbf{x}(m))-F^*\big) + \frac{(2s+1)(\sqrt{p}/\eta+2L)}{\delta}\sum_{k=(m-2s)_+}^{m-1}\|\mathbf{x}(k+1)-\mathbf{x}(k)\| + \frac{(2s+1)(\sqrt{p}/\eta+2L)}{\delta}\sum_{t=m}^{n}\|\mathbf{x}(t+1)-\mathbf{x}(t)\|. \tag{75}$$

Setting $\delta = (2s+1)(\sqrt{p}/\eta+2L)$ and rearranging:

$$\sum_{t=m}^{n}\|\mathbf{x}(t+1)-\mathbf{x}(t)\| \le \frac{(2s+1)(\sqrt{p}/\eta+2L)}{\alpha}\varphi\big(F(\mathbf{x}(m))-F^*\big) + \sum_{k=(m-2s)_+}^{m-1}\|\mathbf{x}(k+1)-\mathbf{x}(k)\| \tag{76}$$

Since the right-hand side is finite and does not depend on n, letting n tend to infinity completes our proof for Equation (11). □

C Proof of Lemma 1

Lemma 1. *Assume $\forall t, i, t \in T_i$. Let the step size $\eta < \frac{\rho-1}{4C\rho}\frac{\sqrt{\rho}-1}{\sqrt{\rho}^{s+1}-1}$ for any $\rho > 1$ and all $U_i, i = 1, ..., p$ be eventually Lipschitz continuous, then the sequences $\{\mathbf{x}(t)\}$ and $\{\mathbf{x}^i(t)\}, i = 1, ..., p$, have finite length.*

Follow the same argument of equation (27), one can bound $\|\mathbf{x}^i(t) - \mathbf{x}^i(t+1)\|$ similarly as

$$\|\mathbf{x}^i(t)-\mathbf{x}^i(t+1)\| \le \sum_{k=(t-s)_+}^{t}\|\mathbf{x}(k+1)-\mathbf{x}(k)\|. \tag{77}$$

For simplicity we omit the details.

Since Assumption 5 holds for all $t > t_L$, we prove the lemma by considering two complementary cases.

Case 1: There exists a $\hat{t} > t_L$ such that

$$\sum_{k=(\hat{t}-s)_+}^{\hat{t}}\|\mathbf{x}(k+1)-\mathbf{x}(k)\| \le \frac{\sqrt{\rho}^{s+1}-1}{\sqrt{\rho}-1}\|\mathbf{x}(\hat{t}+1)-\mathbf{x}(\hat{t})\|$$

Case 2: For all $t > t_L$ case 1 fails.

We will show that case 1 leads to the sufficient decrease property in Assumption 3 for all large t, case 2 leads to the finite length of the models.

Case 1: $\hat{t}$ exists.

We start by proving the following lemma.

Lemma 3. *With Assumption 5 and the existence of $\hat{t}$. Set $\eta^{-1} > \frac{4C\rho}{\rho-1}\frac{\sqrt{\rho}^{s+1}-1}{\sqrt{\rho}-1}$, then it holds for all $t > \hat{t}$ that*

$$\|\mathbf{x}(\hat{t}+1)-\mathbf{x}(\hat{t})\| \le \sqrt{\rho}\|\mathbf{x}(\hat{t}+2)-\mathbf{x}(\hat{t}+1)\|.$$

Proof. Using the inequality $\|a\|_2^2 - \|b\|_2^2 \leq 2\|a\|\|a-b\|$, we have for all $t > \hat{t} > t_L$

$$\begin{aligned}
\|\mathbf{x}(t+1) - \mathbf{x}(t)\|_2^2 - \|\mathbf{x}(t+2) - \mathbf{x}(t+1)\|_2^2 &\leq 2\|\mathbf{x}(t+1) - \mathbf{x}(t)\| \|(\mathbf{x}(t+1) - \mathbf{x}(t)) - (\mathbf{x}(t+2) - \mathbf{x}(t+1))\| \\
\text{(no skip of update)} &= 2\|\mathbf{x}(t+1) - \mathbf{x}(t)\| \left\| \sum_{i=1}^{p} U_i(\mathbf{x}^i(t)) - \sum_{i=1}^{p} U_i(\mathbf{x}^i(t+1)) \right\| \\
&\leq 2\|\mathbf{x}(t+1) - \mathbf{x}(t)\| \sum_{i=1}^{p} \left\| U_i(\mathbf{x}^i(t)) - U_i(\mathbf{x}^i(t+1)) \right\| \\
\text{(Assumption 5)} &\leq 2\|\mathbf{x}(t+1) - \mathbf{x}(t)\| \left(\sum_{i=1}^{p} C_i \eta \|\mathbf{x}^i(t) - \mathbf{x}^i(t+1)\| \right) \\
\text{(equation (77))} &\leq 2\|\mathbf{x}(t+1) - \mathbf{x}(t)\| \left(\sum_{i=1}^{p} C_i \eta \left[\sum_{k=(t-s)_+}^{t} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \right] \right) \\
&= 2C\eta \|\mathbf{x}(t+1) - \mathbf{x}(t)\| \left[\sum_{k=(t-s)_+}^{t} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \right]. \qquad (78)
\end{aligned}$$

Now we use an induction argument. Since there exists $\hat{t} > t_L$ such that $\sum_{k=(\hat{t}-s)_+}^{\hat{t}} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \leq \frac{\sqrt{\rho}^{s+1}-1}{\sqrt{\rho}-1} \|\mathbf{x}(\hat{t}+1) - \mathbf{x}(\hat{t})\|$, then set $t = \hat{t}$ in the above inequality, we obtain

$$\begin{aligned}
\left\|\mathbf{x}(\hat{t}+1) - \mathbf{x}(\hat{t})\right\|_2^2 - \left\|\mathbf{x}(\hat{t}+2) - \mathbf{x}(\hat{t}+1)\right\|_2^2 &\leq 2C\eta \frac{\sqrt{\rho}^{s+1}-1}{\sqrt{\rho}-1} \left\|\mathbf{x}(\hat{t}+1) - \mathbf{x}(\hat{t})\right\|_2^2 \\
\text{(choice of } \eta) &\leq \left(1 - \frac{1}{\rho}\right) \left\|\mathbf{x}(\hat{t}+1) - \mathbf{x}(\hat{t})\right\|^2. \qquad (79)
\end{aligned}$$

After rearranging terms we conclude $\|\mathbf{x}(\hat{t}+1) - \mathbf{x}(\hat{t})\| \leq \sqrt{\rho}\|\mathbf{x}(\hat{t}+2) - \mathbf{x}(\hat{t}+1)\|$. Now we assume this relationship holds up to t ($t > \hat{t}$), then (78) becomes

$$\begin{aligned}
\|\mathbf{x}(t+1) - \mathbf{x}(t)\|_2^2 - \|\mathbf{x}(t+2) - \mathbf{x}(t+1)\|_2^2 &\leq 2C\eta \frac{\sqrt{\rho}^{s+1}-1}{\sqrt{\rho}-1} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|_2^2 \\
\text{(choice of } \eta) &\leq \left(1 - \frac{1}{\rho}\right) \|\mathbf{x}(t+1) - \mathbf{x}(t)\|_2^2
\end{aligned}$$

we obtain $\|\mathbf{x}(t+1) - \mathbf{x}(t)\| \leq \sqrt{\rho}\|\mathbf{x}(t+2) - \mathbf{x}(t+1)\|$. This completes the lemma. □

With this bound, inequality (37) can be further bounded for $t > \hat{t}$ as

$$\begin{aligned}
F\big(\mathbf{x}(t+1)\big) - F\big(\mathbf{x}(t)\big) &\leq \tfrac{1}{2}(L_f - 1/\eta)\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 + L\|\mathbf{x}(t+1) - \mathbf{x}(t)\| \cdot \sum_{k=(t-s)_+}^{t-1} \|\mathbf{x}(k+1) - \mathbf{x}(k)\|. \\
&\leq -\alpha\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2, \qquad (80)
\end{aligned}$$

where

$$\begin{aligned}
\alpha &\geq \frac{\eta^{-1} - L_f}{2} - \frac{L\sqrt{\rho}(1 - \sqrt{\rho}^s)}{1 - \sqrt{\rho}} \\
(C > L, \rho > 1) &\geq \frac{2L\rho}{\rho - 1} \frac{\sqrt{\rho}^{s+1} - 1}{\sqrt{\rho} - 1} - \frac{L(\sqrt{\rho}^{s+1} - 1)}{\sqrt{\rho} - 1} - \frac{L_f}{2} \\
&\geq \frac{L(\sqrt{\rho}^{s+1} - 1)}{\sqrt{\rho} - 1} \left(\frac{2\rho}{\rho - 1} - 1\right) - \frac{L_f}{2} \\
(L > L_f, \rho > 1) &> 0,
\end{aligned}$$

This proves the sufficient decrease for all $t > \hat{t}$ of the objective value. Hence, the finite length property of the models follows from Theorem 2.

Case 2: $\hat{t}$ does not exist

In this case we have for all $t > t_L$ it holds that $\sum_{k=(t-s)_+}^{t} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \geq \frac{\sqrt{\rho}^{s+1}-1}{\sqrt{\rho}-1} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|$. Set $D = \frac{\sqrt{\rho}^{s+1}-1}{\sqrt{\rho}-1}$ and sum the inequality over t from t_L to n yields

$$\begin{aligned}\sum_{k=t_L}^{n} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| &< \frac{1}{D} \sum_{t=t_L}^{n} \sum_{k=(t-s)_+}^{t} \|\mathbf{x}(k+1) - \mathbf{x}(k)\| \\ &< \frac{s+1}{D} \sum_{t=(t_L-s)_+}^{n} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|,\end{aligned}$$

which after rearranging terms becomes

$$(1 - \frac{s+1}{D}) \sum_{t=t_L}^{n} \|\mathbf{x}(t+1) - \mathbf{x}(t)\| \leq \frac{s+1}{D} \sum_{t=(t_L-s)_+}^{t_L-1} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|.$$

Since $D = \frac{\sqrt{\rho}^{s+1}-1}{\sqrt{\rho}-1} > s+1$ for $\rho > 1$ and t_L is finite, the right hand side of the above inequality is finite, and the left hand side has positive coefficient. Thus the above inequality implies

$$\sum_{t=0}^{n} \|\mathbf{x}(t+1) - \mathbf{x}(t)\| < +\infty.$$

Enlarge $n \to \infty$ gives the finite length property of the global model. By the proof of Appendix B, we know the finite length of global model implies the finite length of all local models.

D Proof of Example 1

We proof case by case, and the scaled version $\gamma g(\mathbf{x}), \gamma > 0$ trivially follows from the same argument.

Cases $g = 0,\ g = \frac{1}{2}\|\cdot\|^2$:

When $g = 0$, the update operator in eq. (9) becomes $U_i(\mathbf{x}^i(t)) = -\eta \nabla_i f(\mathbf{x}^i(t))$, which is ηL_i Lipschitz due to Assumption 1.2 .

When $g = \frac{1}{2}\|\cdot\|^2$, the update operator becomes for $i = 1, ..., p$

$$U_i(\mathbf{x}^i(t)) = \mathrm{prox}^{\eta}_{\frac{1}{2}\|\cdot\|_2^2}(x_i(t) - \eta \nabla_i f(\mathbf{x}^i(t))) - x_i(t) = -\frac{1}{1+\eta^{-1}} \left(x_i(t) + \nabla_i f(\mathbf{x}^i(t))\right).$$

With which we have

$$\begin{aligned}\|U_i(\mathbf{x}^i(t+1)) - U_i(\mathbf{x}^i(t))\| &\leq \frac{1}{1+\eta^{-1}} \|(x_i(t+1) - x_i(t)) + (\nabla_i f(\mathbf{x}^i(t+1)) - \nabla_i f(\mathbf{x}^i(t)))\| \\ &\leq \eta(1 + L_i)\|\mathbf{x}^i(t+1) - \mathbf{x}^i(t)\|\end{aligned}$$

Cases $g = \|\cdot\|_0,\ \|\cdot\|_0 + \|\cdot\|^2,\ \|\cdot\|_{0,2},\ \|\cdot\|_{0,2} + \|\cdot\|^2$: For the non-overlapping group norms, we assign each machine a subset of groups of coordinates.

Consider $g = \|\cdot\|_0$, its proximal map on i-th coordinate can be expressed as

$$\mathrm{prox}^{\eta}_{g_i}(z_i) = \begin{cases} z_i, \text{ if } |z_i| > \sqrt{2\eta} \\ 0, \text{ otherwise} \end{cases}$$

The mapping contains a hard threshold, i.e., it filters out those coordinates with magnitude less than $\sqrt{2\eta}$. This implies that any change of the support set of $\mathrm{prox}^{\eta}_{g_i}(z_i)$ will induce a jump of magnitude of at least $\sqrt{2\eta}$. On

the other hand, the second assertion of Theorem 1 imply that $\lim_{t\to\infty} \|x_i(t+2) - x_i(t+1)\| = 0$, which by local update can be expressed as

$$\lim_{t\to\infty} \|\text{prox}_{g_i}^{\eta}(x_i(t+1) - \eta\nabla_i f(\mathbf{x}^i(t+1))) - \text{prox}_{g_i}^{\eta}(x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t)))\| = 0.$$

Hence by the above equation and the jump of proximal map, the support Ω of $\text{prox}_{g_i}^{\eta}(x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t)))$ (i.e., $x_i(t+1)$) must remain stable for all t sufficiently large. Moreover, the proximal map reduces to identity operator on the support set Ω. Thus, for all t sufficiently large we have

$$\begin{aligned}
\|U_i(\mathbf{x}^i(t+1)) - U_i(\mathbf{x}^i(t))\| &= \|\text{prox}_{g_i}^{\eta}(x_i(t+1) - \eta\nabla_i f(\mathbf{x}^i(t+1))) - x_i(t+1) - \text{prox}_{g_i}^{\eta}(x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t))) - x_i(t)\| \\
\text{(support on } \Omega) &= \|[\text{prox}_{g_i}^{\eta}(x_i(t+1) - \eta\nabla_i f(\mathbf{x}^i(t+1))) - x_i(t+1) - \text{prox}_{g_i}^{\eta}(x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t))) - x_i(t)]_\Omega\| \\
(\text{prox}_g^{\eta} \text{ is identity on } \Omega) &\le \|\eta\nabla_i f(\mathbf{x}^i(t)) - \eta\nabla_i f(\mathbf{x}^i(t+1))\| \\
&\le \eta L_i \|\mathbf{x}^i(t+1) - \mathbf{x}^i(t)\|.
\end{aligned}$$

Hence the operator is eventually $\mathcal{O}(\eta)$ Lipschitz.

Next we consider (without loss of generality) $g = \|\cdot\|_0 + \frac{\lambda}{2}\|\cdot\|^2$ where $\lambda > 0$. The proximal map on i-th coordinate is

$$\text{prox}_{g_i}^{\eta}(z_i) = \begin{cases} z_i, \text{ if } |z_i| > \sqrt{2(\eta + \eta^2\lambda)} \\ 0, \text{ otherwise} \end{cases}$$

Thus, the mapping also contains a hard threshold. Following similar argument as previous case, we conclude that the support Ω of $\text{prox}_{g_i}^{\eta}(x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t)))$ (i.e., $x_i(t+1)$) must remain stable for all t sufficiently large, and the proximal map reduces to identity operator on Ω. Consequently, the operator $U_i(\mathbf{x}^i(t))$ is $\mathcal{O}(\eta)$ Lipschitz for all t large.

The proof of group norms $g = \|\cdot\|_{0,2}, \|\cdot\|_{0,2} + \|\cdot\|^2$ then follows by realizing that the proximal maps have hard threshold on group support.

Cases $g = \|\cdot\|_1,\ \|\cdot\|_1 + \|\cdot\|^2$ with eventual stable support set of $\{\mathbf{x}(t)\}$:

For these two cases we assume that the support set of $\{\mathbf{x}(t)\}$ remains unchanged for all large t.

We just need to consider $g = \|\cdot\|_1 + \frac{\lambda}{2}\|\cdot\|^2, \lambda \ge 0$. Its proximal map on vector z_i has the form

$$\text{prox}_g^{\eta}(z_i) = \tfrac{1}{1+\eta\lambda}\text{sgn}(z_i)\left(|z_i| - \eta\right)_+.$$

Since the support set Ω of $\mathbf{x}(t)$ (i.e. $\text{prox}_g^{\eta}\left(x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t))\right)$) is assumed to be stable after some t_L, the above soft-thresholding operator ensures that $|x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t))|_\Omega > \eta$ for all large t, and we obtain

$$\begin{aligned}
U_i(\mathbf{x}^i(t)) &= [x_i(t+1) - x_i(t)]_\Omega = [\text{prox}_g^{\eta}\left(x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t))\right) - x_i(t)]_\Omega \\
&= (1+\eta\lambda)^{-1}\left[-\eta\nabla_i f(\mathbf{x}^i(t)) - \eta\text{sgn}\left(x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t))\right)\right]_\Omega - \frac{\eta\lambda}{1+\eta\lambda}[x_i(t)]_\Omega
\end{aligned} \tag{81}$$

On the other hand, Theorem 1.2 and the Lispchitz gradient of f implies

$$\lim_{t\to\infty} \left\| \left[x_i(t+1) - \eta\nabla_i f(\mathbf{x}^i(t+1))\right] - \left[x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t))\right] \right\| = 0.$$

Then $[\text{sgn}(x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t))]_\Omega$ must eventually remain constant, since otherwise the condition $|x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t))|_\Omega > \eta$ will induce a change of $|x_i(t) - \eta\nabla_i f(\mathbf{x}^i(t)|$ to be at least 2η and violate the above asymptotic condition. In summary, for all large t we have

$$U_i(\mathbf{x}^i(t)) = (1+\eta\lambda)^{-1}\left[-\eta\nabla_i f(\mathbf{x}^i(t)) - \text{Const}\right]_\Omega - \frac{\eta\lambda}{1+\eta\lambda}[x_i(t)]_\Omega$$

which further implies that

$$\begin{aligned}
\|U_i(\mathbf{x}^i(t+1)) - U_i(\mathbf{x}^i(t))\| &\le \|(1+\eta\lambda)^{-1}\left[\eta\nabla_i f(\mathbf{x}^i(t+1)) - \eta\nabla_i f(\mathbf{x}^i(t))\right]_\Omega + \frac{\eta\lambda}{1+\eta\lambda}[x_i(t+1) - x_i(t)]_\Omega\| \\
&\le \eta L_i\|\mathbf{x}^i(t+1) - \mathbf{x}^i(t)\| + \eta\lambda\|\mathbf{x}^i(t+1) - \mathbf{x}^i(t)]\| \\
&\le \eta(L_i + \lambda)\|\mathbf{x}^i(t+1) - \mathbf{x}^i(t)\|.
\end{aligned}$$

E Proof of Theorem 3

Theorem 3 (Global rate of convergence)**.** *If the finite length property in Theorem 2 holds, then*

1. $\sum_{t=0}^{\infty} \|\mathbf{e}(t)\| < \infty$;
2. $F(\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}(k)) - \inf F \le O\left(t^{-1}\right)$.

Proof. For the first assertion, note that for any n:

$$\sum_{t=0}^{n}\|\mathbf{e}(t)\| = \eta\sum_{t=0}^{n}\left\|\left(\nabla_1 f(\mathbf{x}^1(t)) - \nabla_1 f(\mathbf{x}(t)),\ \dots\ ,\ \nabla_p f(\mathbf{x}^p(t)) - \nabla_p f(\mathbf{x}(t))\right)\right\| \tag{82}$$

$$\text{(triangle inequality, Assumption 1.2)} \ \le \eta\sum_{t=0}^{n}\sum_{i=1}^{p} L_i\|\mathbf{x}(t) - \mathbf{x}^i(t)\| \tag{83}$$

$$\text{(Equation (27))} \ \le \eta\sum_{t=0}^{n}\left(\sum_{i=1}^{p}L_i\right)\sum_{k=(t-s)_+}^{t-1}\|\mathbf{x}(k+1)-\mathbf{x}(k)\|$$

$$= L\eta\sum_{t=0}^{n}\sum_{k=(t-s)_+}^{t-1}\|\mathbf{x}(k+1)-\mathbf{x}(k)\| \tag{84}$$

$$\le Ls\eta\sum_{t=0}^{n-1}\|\mathbf{x}(t+1)-\mathbf{x}(t)\|\,. \tag{85}$$

Letting n tend to infinity we obtain

$$\sum_{t=0}^{\infty}\|\mathbf{e}(t)\| \le Ls\eta\sum_{t=0}^{\infty}\|\mathbf{x}(t+1)-\mathbf{x}(t)\| < \infty. \tag{86}$$

For the second assertion, we first recall from [30] that the inexact proximal gradient algorithm in Equation (14) has the following bound, provided that F is convex:

$$F\left(\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}(k)\right) - F^* \le \frac{(\|\mathbf{x}(0)-\mathbf{x}^*\| + 2A_t)^2}{2t\eta}, \quad \text{where} \quad A_t = \sum_{k=0}^{t}\eta\|\mathbf{e}(k)\|. \tag{87}$$

The second assertion thus follows from the first one (assuming convexity). □

F Experiments Specifications

Specifications for $\|\cdot\|_{0,2}$ Lasso:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\ \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|_2^2 + \sum_{i=1}^{20}\gamma_i\mathbb{I}(\|x_i\|).$$

Here $A \in \mathbb{R}^{1000\times 2000}, \mathbf{b} \in \mathbb{R}^{1000}$, and $\mathbf{x} \in \mathbb{R}^2000$ is divided into 20 equal groups of features. Matrix A is generated from $\mathcal{N}(0,1)$ with normalized columns. We set $\mathbf{b} = A\tilde{\mathbf{x}} + \varepsilon$, where ε is generated from $\mathcal{N}(0, 10^{-2})$ and $\tilde{\mathbf{x}}$ is a normalized vector with 8 non-zero groups of features generated from $\mathcal{N}(0,1)$. For the non-zero groups of $\tilde{\mathbf{x}}$, we set the corresponding $\gamma_i = 10^{-4}$, and for the remaining groups we set $\gamma_i = 10^{-2}$.

We implement msPG on four cores with each core assigned five group of features. Each core stores the corresponding column blocks of A.

Specifications for $\|\cdot\|_1$ Lasso:

$$\min_{\mathbf{x}}\ \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|_2^2 + \gamma\|\mathbf{x}\|_1.$$

Data Generation

We generate the data column-wise. Starting from first column, we randomly pick 10^4 samples to have non-zero in column 1 and sample each value from Uniform$(-1, 1)$. We normalize it such that the ℓ_2-norm of the column is 1. We denote these values as $\mathbf{v}_1 \in \mathbb{R}^n$. To generate column i, with probability 0.5 we randomly pick a new set of samples to have non-zero values at column i (otherwise we use the same samples from column $i-1$). This simulates the correlations between each column. Once the samples are chosen, we assign values from Unif$(-1, 1)$. $\mathbf{v}_i$ is again normalized. We generate ground truth regressor $\beta \in \mathbb{R}^d$ from $\mathcal{N}(0, 1)$ with 1% non-zero entries, and obtain the regressed value from $\mathbf{b} = A\beta$ where A is the design matrix.