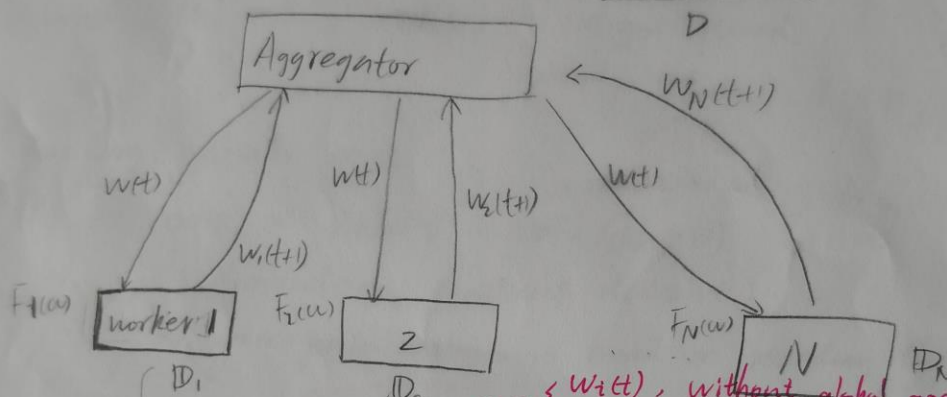
ECE 696/596 Lecture #27

Review. ① Introduction to Federated Learning.
② Design and detailed implementation.

Today: ① Theory part and proof

System Setup: (Distributed Gradient Descent, DGD for short)

global aggregation: $w(t+1) = \dfrac{\sum_{i=1}^{N} D_i w_i(t+1)}{D}$



local update: $w_i(t+1) = \tilde{w}_i(t) - \eta \nabla F_i(\tilde{w}_i(t))$, where $\tilde{w}_i(t) = \begin{cases} w_i(t), & \text{without global agg.} \\ w(t), & \text{with global agg.} \end{cases}$

Local loss: $F_i(w) \triangleq \dfrac{1}{|\mathbb{D}_i|} \sum_{j \in \mathbb{D}_i} f_j(w)$, $f_j(w) \triangleq f(w, x_j, y_j)$

$f_j(w)$ → loss on sample x_j

$\mathbb{D}_i$ → Local dataset for worker i.

Overall loss: $F(w) \triangleq \dfrac{\sum_{i=1}^{N} D_i F_i(w)}{D}$

$$D_i = |\mathbb{D}_i|, \quad D = \sum_{i=1}^{N} D_i$$

Objective: $w^* = \underset{w}{\operatorname{argmin}}\, F(w)$.

K : total # of global aggregations

τ : # of local updates between two global agg.

$T = K\tau$: # of updates.

c: each step of local update at all workers consumes c units of resource

b: each global update consumes b units of resource

$\boxed{\text{Goal}}$

$$\min_{\tau, T} F(w(T))$$

s.t. $Tc + Kb = Tc + \frac{T}{\tau} b = T(c + \frac{b}{\tau}) \le R$

R: resource budget

Denote $[k] = [(k-1)\tau, k\tau]$ integer interval $(\tau+1)$ integers $\{(k-1)\tau, (k-1)\tau+1, \cdots, k\tau\}$

Auxiliary parameter vector:

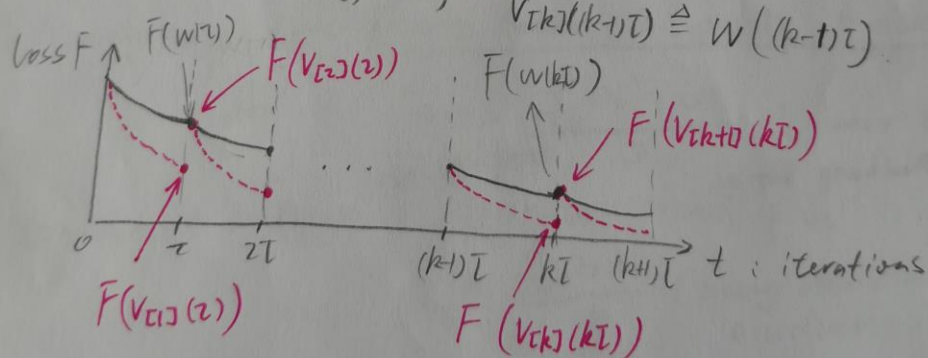
— $v_{[k]}(t) = v_{[k]}(t-1) - \eta \nabla F(v_{[k]}(t-1))$ (∇F: full gradient)

(Centralized gradient descent)

$\Rightarrow$ convergence analysis is based on bounding the difference

— $v_{[k]}(t)$ is $\boxed{\text{synchronized}}$ with $w(t)$ at the beginning of each interval $[k]$, i.e., $v_{[k]}((k-1)\tau) \triangleq w((k-1)\tau)$

($w(t)$: averaged global update)



Remark: $v_{[k]}(t\tau) \neq v_{[k+1]}(t\tau)$, $t = 1, 2, \cdots, T$.

Assumptions: 1) $F_i(w)$ is CVX. (individual loss is convex).

2) $F_i(w)$ is ρ-Lipschitz, i.e.,

$$\|F_i(w) - F_i(w')\| \le \rho \|w - w'\|, \ \forall w, w'$$

absolute value

(Function value cannot change too much between two points w and w')

3) $F_i(w)$ is β-smooth, i.e.,

$$\|\nabla F_i(w) - \nabla F_i(w')\| \le \beta \|w - w'\|, \ \forall w, w'$$

$\Rightarrow$ two properties ① ②

① $f(w) - f(w') \le \nabla f(w)^T (w-w') - \frac{1}{2\beta} \|\nabla f(w) - \nabla f(w')\|^2$

X

② $|f(w) - f(w') - \nabla f(w')^T (w-w')| \le \frac{\beta}{2} \|w - w'\|^2$

plus convexity: $f(w) \ge f(w') + \nabla f(w')^T (w-w')$

$$f(w) - \underbrace{f(w') - \nabla f(w')^T (w-w')}_{\text{linear approximation of } f(w) \text{ at } w'} \le \frac{\beta}{2} \|w - w'\|^2$$

Remarks: 1) pretty common assumptions, strong convexity et.

It's easy to see that $F(w)$ is convex, ρ-Lipschitz, β-smooth

due to $F(w) = \frac{\sum_{i=1}^{N} D_i F_i(w)}{D}$. (linear combination)

[def] Gradient Divergence: $\|\underbrace{\nabla F_i(w)}_{\text{local gradient}} - \underbrace{\nabla F(w)}_{\text{full gradient}}\| \le \delta_i$

$$\delta \triangleq \frac{\sum_{i=1}^{N} D_i \delta_i}{D}$$ (averge deviation of local gradients)

GOAL: Bound $F(w(T)) - F(w^*)$

$\downarrow$

analyze the convergenc behavior of $F(v_{[k]}(t))$

$\downarrow$

bound $F(w(T)) - F(v_{[k]}(T))$

$\downarrow$

bound by $\|w(T) - v_{[k]}(T)\|$.

Remark: compare the difference between the Distributed GD (DGD) with the centralized version $v_{[k]}(t)$. ④

Theorem 1: For any interval $[k]$ and $t \in [k]$, we have

$$\|w(t) - v_{[k]}(t)\| \le h\big(t-(k-1)\tau\big)$$

centralized param.

averaged param. from global agg.

where

$$h(x) \triangleq \frac{\delta}{\beta}\left[(\eta\beta+1)^x - 1\right] - \eta\delta x.$$

Furthermore, as $F(\cdot)$ is ρ-Lipschitz, we have

$$F(w(t)) - F(v_{[k]}(t)) \le \cancel{\rho h(t-(k-1)\tau)}$$

$$F(w) - F(v) \le |F(w(t)) - F(v_{[k]}(t))| \le \rho\|w-v\| \quad \rho\|w(t) - v_{[k]}(t)\|$$

$$\Rightarrow F(w) - F(v) \le \rho h(\cdot) \quad \le \rho h(\cdot) \quad \le \rho h\big(t-(k-1)\tau\big).$$

Remarks ① $h(0) = h(1) = 1$.

$\Rightarrow w(t) - v_{[k]}(t) = 0$, if $\tau = 1$, that is, when $\tau = 1$, we perform global agg. after each local update, then DGD is equivalent to centralized GD:

[proof] Need to prove $w(t) = w(t-1) - \eta\nabla F(w(t-1))$

$\tau = 1 \Rightarrow \hat{w}_i(t) = w(t)$

$$w(t) = \frac{\sum_{i=1}^N D_i w_i(t)}{D} \quad w_i(t) = \hat{w}_i(t-1) - \eta\nabla F_i(\hat{w}_i(t-1))$$

$$= \frac{\sum_i D_i\left[\hat{w}_i(t-1) - \eta\nabla F_i(\hat{w}_i(t-1))\right]}{D}$$

$$= \frac{\sum_i D_i w_i(t-1)}{D} - \eta \cdot \frac{\sum_i D_i \nabla F_i(w(t-1))}{D}$$

$$= w_i(t-1) - \eta\nabla F(w(t-1)) \qquad \eta\nabla\left(\frac{\sum_i D_i F_i(w(t-1))}{D}\right) = \eta\nabla F(w(t-1))$$

(linearity of gradient operator).

$\square$

✗ ② When x is large, $h(x) \simeq \frac{\delta}{\beta}(\eta\beta+1)^x$: exponential w.r.t. x

proof of Thm 1.

local para. at iteration t

$\tilde{w}_i(t) = \begin{cases} w(t), & \text{global agg.} \\ w_i(t) \end{cases}$

[Lemma 2] For any interval $[k]$, and $t \in [(k-1)\tau, k\tau]$, we have

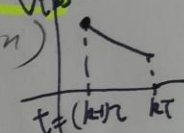
$$\|\tilde{w}_i(t) - v_{[k]}(t)\| \le g_i(t-(k-1)\tau)$$

where $g_i(x) \triangleq \frac{\delta_i}{\beta}\left[(\eta\beta+1)^x - 1\right]$.

(using β-smoothness, triangle inequality, induction)

$\tilde{w}_i(t) = w(t)$

$v_{[k]}(t)$ is synchronized



[proof] By induction on $t \in [k]$.

The local para. is not too far from the centralized para, ⇒ the average of local paras. ($w(t)$) should also be not too far from $v_{[k]}(t)$.

First, when $t=(k-1)\tau$, we know that $\tilde{w}_i(t) = v_{[k]}(t)$ due to global aggregation and sychronization of $v_{[k]}(t)$ with $w(t)$.

That is, $\|w_i(t) - v_{[k]}(t)\| = g_i(0) = 0$ when $t=(k-1)\tau$ — holds for the initial value

For the induction, we assume that

$$\|\tilde{w}_i(t-1) - v_{[k]}(t-1)\| \le g_i(t-1-(k-1)\tau) \text{ hold for some } t \in ((k-1)\tau, k\tau)$$

We have $\|\tilde{w}_i(t) - v_{[k]}(t)\|$

$$= \left\|\left(\tilde{w}_i(t-1) - \eta\nabla F_i(\tilde{w}_i(t-1))\right) - \left(v_{[k]}(t-1) - \eta\nabla F(v_{[k]}(t-1))\right)\right\|$$

$$= \left\|\left(\tilde{w}_i(t-1) - v_{[k]}(t-1)\right) - \eta\left(\nabla F_i(\tilde{w}_i(t-1)) \overbrace{- \nabla F_i(v_{[k]}(t-1)) + \nabla F_i(v_{[k]}(t-1))}^{\text{zero trick}} - \nabla F(v_{[k]}(t-1))\right)\right\|$$

rearrange the terms

$$= \left\|\underbrace{\left(\tilde{w}_i(t-1) - v_{[k]}(t-1)\right)}_{①} - \underbrace{\eta\left(\nabla F_i(\tilde{w}_i(t-1)) - \nabla F_i(v_{[k]}(t-1))\right)}_{②} - \underbrace{\eta\left(\nabla F_i(v_{[k]}(t-1)) - \nabla F(v_{[k]}(t-1))\right)}_{③}\right\|$$

triangle inequality

$$\le \|①\| + \|②\| + \|③\|$$

$$= \|\tilde{w}_i(t-1) - v_{[k]}(t-1)\| + \eta\underbrace{\|\nabla F_i(\tilde{w}_i(t-1)) - \nabla F_i(v_{[k]}(t-1))\|}_{\le \eta\beta\|\tilde{w}_i(t-1) - v_{[k]}(t-1)\|}$$

$$\eta \| \underline{\nabla F_i(v_{[k]}(t-1)) - \nabla F(v_{[k]}(t-1))} \| \le \delta_i$$ gradient divergence

$$\le \underbrace{(\eta\beta+1)}_{\beta\text{-smoothness}} \| \hat{w}_i(t-1) - v_{[k]}(t-1) \| + \eta\delta_i$$

individual gradient divergence

$(\ \|\nabla F_i(w) - F(w)\| \le \delta_i, \forall w\ )$.

~~inductio~~

induction assumption at t-1

$$\le (\eta\beta+1)\, \underline{g_i(t-1-(k-1)\tau)} + \eta\delta_i$$

$$= (\eta\beta+1)\left(\frac{\delta_i}{\beta}\left[(\eta\beta+1)^{t-1-(k-1)\tau} - 1\right]\right) + \eta\delta_i$$

... math manipulation

$$= g_i(t-(k-1)\tau)$$

$$\Rightarrow \| \hat{w}_i(t) - v_{[k]}(t) \| \le g_i(t-(k-1)\tau) \quad \forall t \in [k]. \quad \square$$

proof of Thm 1 continued :

Since $w_i(t) = \hat{w}_i(t-1) - \eta \nabla F_i(\hat{w}_i(t-1))$

$$w(t) = \frac{\sum_{i=1}^{N} D_i w_i(t)}{D}, \text{ then}$$

$$w(t) = \frac{\sum_i D_i \hat{w}_i(t-1)}{D} - \eta \frac{\sum_i D_i \nabla F_i(\hat{w}_i(t-1))}{D}$$

$$= w(t-1) - \eta \frac{\sum_i D_i \nabla F_i(\hat{w}_i(t-1))}{D}$$

then for $t \in ((k-1)\tau, k\tau]$,

$$\| w(t) - v_{[k]}(t) \| = \Big\| w(t-1) - \eta \frac{\sum_i D_i \nabla F_i(\hat{w}_i(t-1))}{D} - v_{[k]}(t-1) + \eta \nabla F(v_{[k]}(t-1)) \Big\|$$

rearrange the terms

$$= \Big\| w(t-1) - v_{[k]}(t-1) - \eta \left(\frac{\sum_i D_i \nabla F_i(\hat{w}_i(t-1))}{D} - \nabla F(v_{[k]}(t-1)) \right) \Big\|$$

$$= \left\| w(t-1) - v_{[k]}(t-1) - \eta \frac{\sum_i D_i \left(\nabla F_i(\tilde{w}_i(t-1)) - \nabla F_i(v_{[k]}(t-1)) \right)}{D} \right\|$$

triangle inequality

$$\leq \| w(t-1) - v_{[k]}(t-1) \| + \eta \cdot \left(\frac{\sum_i D_i \| \nabla F_i(\tilde{w}_i(t-1)) - \nabla F_i(v_{[k]}(t-1)) \|}{D} \right)$$

$$\| \nabla F_i(\tilde{w}_i(t-1)) - \nabla F_i(v_{[k]}(t-1)) \| \leq \beta \| \tilde{w}_i(t-1) - v_{[k]}(t-1) \|$$

β-smooth

$$\leq \underbrace{\| w(t-1) - v_{[k]}(t-1) \|}_{①} + \eta\beta \left(\frac{\sum_i D_i \| \tilde{w}_i(t-1) - v_{[k]}(t-1) \|}{D} \right)$$

$$\leq ① + \eta\beta \frac{\sum_i D_i \, g_i(t-1-(k-1)\tau)}{D}$$

Lemma 2

$$= ① + \eta\beta \frac{\sum_i D_i \frac{\delta_i}{\beta}\left[(\eta\beta+1)^{t-1-(k-1)\tau} - 1 \right]}{D}$$

$$= ① + \eta \underbrace{\left(\frac{\sum_i D_i \delta_i}{D} \right)}_{\triangleq \delta} \left[(\eta\beta+1)^{t-1-(k-1)\tau} - 1 \right]$$

def of $\delta = \frac{\sum_i D_i \delta_i}{D}$

$$= ① + \eta\delta \left[(\eta\beta+1)^{t-1-(k-1)\tau} - 1 \right]$$

$$\Rightarrow \| w(t) - v_{[k]}(t) \| - \| w(t-1) - v_{[k]}(t-1) \| \leq \eta\delta \left[(\eta\beta+1)^{t-1-(k-1)\tau} - 1 \right] \quad (*)$$

If $t = (k-1)\tau$, we have $w(t) = v_{[k]}(t)$ (synchronization)

$$\Rightarrow \| w(t) - v_{[k]}(t) \| = 0.$$

If $t \in ((k-1)\tau, k\tau]$, then by summing $(*)$ from $(k-1)\tau+1$, $(k-1)\tau+2, \ldots, t$, then we have

$$\| w(t) - v_{[k]}(t) \| = \sum_{y=(k-1)\tau+1}^{t} \| w(y) - v_{[k]}(y) \| - \| w(y-1) - v_{[k]}(y-1) \|$$

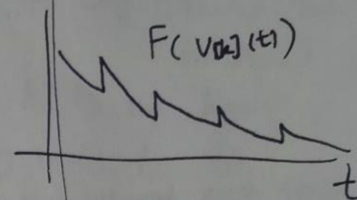
$$\leq \eta\delta \sum_{y=(k-1)\tau+1}^{t} \left[(\eta\beta+1)^{y-1-(k-1)\tau} - 1 \right]$$

geometric sum

$$= \eta\delta \sum_{z=1}^{t-(k-1)\tau} \left[(\eta\beta+1)^{z-1} - 1 \right]$$

let $z = y - (k-1)\tau$

X $$= \eta\delta \sum_{z=1}^{t-(k-1)\tau} (\eta\beta+1)^{z-1} - \eta\delta\left(t-(k-1)\tau\right)$$



geometric sum $$= \eta\delta \, \frac{1-(\eta\beta+1)^{t-(k-1)\tau}}{-\eta\beta} - \eta\delta\left(t-(k-1)\tau\right)$$

$$= \frac{\delta}{\beta}\left[(\eta\beta+1)^{t-(k-1)\tau} - 1 \right] - \eta\delta\left(t-(k-1)\tau\right)$$

$$= h\left(t-(k-1)\tau\right)$$ recall $h(x) \triangleq \frac{\delta}{\beta}\left[(\eta\beta+1)^{x} - 1\right] - \eta\delta x$.

Theorem 2: When all the followy conditions are satisfied,

1) $\eta \le \frac{1}{\beta}$ (small enough step size).
2) $\omega\eta\left(1-\frac{\beta\eta}{2}\right) - \frac{\rho h(\tau)}{\tau\varepsilon^2} > 0$
3) $F(v_{[k]}(t)) - F(w^*) \ge \varepsilon$. $\forall t, k$
4) $F(w(T)) - F(w^*) \ge \varepsilon$.

} weired.

Where $\varepsilon > 0$, and $\omega \triangleq \min_k \frac{1}{\|v_{[k]}((k-1)\tau) - w^*\|^2}$

Then the convergence upper bound is given by

$$F(w(T)) - F(w^*) \le \frac{1}{T\left(\omega\eta\left(1-\frac{\beta\eta}{2}\right) - \frac{\rho h(\tau)}{\tau\varepsilon^2}\right)}$$

Remarks ① Unreasonable assumptions 3) & 4)

② Smaller ε, loose (larger) bound on convergence.

roof of Thm 2.

Denote $\Theta_{[k]}(t) = F(v_{[k]}(t)) - F(w^*)$, $(k-1)\tau \le t \le k\tau$

Assume $\Theta_{[k]}(t) > 0$. (I didn't see why $\Theta_{[k]}(t)$ cannot be zero)

[lemma 3]. When $\eta \le \frac{1}{\beta}$, $\forall k$, $t \in [k] \triangleq [(k-1)\tau, k\tau]$, we have that $\|v_{[k]}(t) - w^*\|$ is non-increasing w.r.t. t.

getting closer to the optimal para. since $v_{[k]}(t)$ is the full GD. (intuition) Each step will lead the para. $v_{[k]}(t)$ closer to its optimal.

[proof] $\|v_{[k]}(t+1) - w^*\|^2$

$= \|v_{[k]}(t) - \eta \nabla F(v_{[k]}(t)) - w^*\|^2$

$= \|v_{[k]}(t) - w^*\|^2 - 2\eta \nabla F(v_{[k]}(t))^T (v_{[k]}(t) - w^*) + \eta^2 \|\nabla F(v_{[k]}(t))\|^2$

✓

Since $F(\cdot)$ is β-smooth, we have

$$0 < \Theta_{[k]}(t) = F(v_{[k]}(t)) - F(w^*) \le \nabla F(v_{[k]}(t))^T (v_{[k]}(t) - w^*) - \frac{\|\nabla F(v_{[k]}(t))\|^2}{2\beta}$$

(due to $f(x) - f(y) \le \nabla f(x)^T (x-y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2$ and $\nabla F(w^*) = 0$)

$$\Rightarrow -\nabla F(v_{[k]}(t))^T (v_{[k]}(t) - w^*) < -\frac{\|\nabla F(v_{[k]}(t))\|^2}{2\beta}.$$

Therefore,

$$\|v_{[k]}(t+1) - w^*\|^2 = \|v_{[k]}(t) - \eta \nabla F(v_{[k]}(t)) - w^*\|^2$$

$$= \underbrace{\|v_{[k]}(t) - w^*\|^2}_{①} - 2\eta \nabla F(v_{[k]}(t))^T (v_{[k]}(t) - w^*) + \eta^2 \underbrace{\|\nabla F(v_{[k]}(t))\|^2}_{③}$$

$$< ① - \eta \frac{\|\nabla F(v_{[k]}(t))\|^2}{\beta} + ③$$

$$= ① + \eta\left(\frac{1}{\beta} - \eta\right)\|\nabla F(v_{[k]}(t))\|^2$$

as long as $\frac{1}{\beta} - \eta > 0$ ($\eta \le \frac{1}{\beta}$), we have

$$\|v_{[k]}(t+1) - w^*\|^2 < \|v_{[k]}(t) - w^*\|^2 \quad \square$$

[Lemma 4] For any k, when $\eta \le \frac{1}{\beta}$ and $t \in [(k-1)\tau, k\tau]$, we have

$$\frac{1}{\theta_{[k]}(t+1)} - \frac{1}{\theta_{[k]}(t)} \ge \omega\eta\left(1-\frac{\beta\eta}{2}\right) \rightarrow \text{positive term}$$

where $\omega = \min_k \frac{1}{\|v_{[k]}((k-1)\tau) - w^*\|^2}$ $\quad \theta_{[k]}(t) \triangleq F(v_{[k]}(t)) - F(w^*)$

$\Rightarrow F(v_{[k]}(t+1)) - F(w^*) < F(v_{[k]}(t)) - F(w^*)$, i.e. $F(v_{[k]}(t))$ is decreasing w.r.t. t.

[proof] using convexity, Cauchy-Schwarz inequality $\square$

skip the proof of this one.

proof of Thm 2 continued. (real proof)

Using Lemma 4 and $t \in [(k-1)\tau, k\tau]$, then

$$\frac{1}{\theta_{[k]}(k\tau)} - \frac{1}{\theta_{[k]}((k-1)\tau)} = \sum_{t=(k-1)\tau}^{k\tau-1}\left(\frac{1}{\theta_{[k]}(t+1)} - \frac{1}{\theta_{[k]}(t)}\right)$$

$$\overset{\text{Lemma 4.}}{\ge} \tau\omega\eta\left(1-\frac{\beta\eta}{2}\right)$$

Summing up from $k=1$ to K, we obtain

$$\sum_{k=1}^{K}\left(\frac{1}{\theta_{[k]}(k\tau)} - \frac{1}{\theta_{[k]}((k-1)\tau)}\right) \ge \sum_{k=1}^{K} \tau\omega\eta\left(1-\frac{\beta\eta}{2}\right)$$

$$= K\tau\omega\eta\left(1-\frac{\beta\eta}{2}\right)$$

$$\Leftrightarrow \frac{1}{\theta_{[K]}(T)} - \frac{1}{\theta_{[1]}(0)} \ge \underbrace{\sum_{k=1}^{K-1}\left(\frac{1}{\theta_{[k+1]}(k\tau)} - \frac{1}{\theta_{[k]}(k\tau)}\right)}_{\text{each term here can be bounded from below as:}} + K\tau\omega\eta\left(1-\frac{\beta\eta}{2}\right) \quad (\star\star)$$

$$\overset{T=K\tau}{=} T\omega\eta\left(1-\frac{\beta\eta}{2}\right)$$

$$\frac{1}{\theta_{[k+1]}(k\tau)} - \frac{1}{\theta_{[k]}(k\tau)} = \frac{\theta_{[k]}(k\tau) - \theta_{[k+1]}(k\tau)}{\theta_{[k+1]}(k\tau)\,\theta_{[k]}(k\tau)}$$

$\theta_{[k]}(t) = F(v_{[k]}(t)) - F(w^*)$

$$= \frac{F(v_{[k]}(k\tau)) - F(v_{[k+1]}(k\tau))}{\text{denominator}}$$

equal by definition

$$= \frac{F(v_{[k]}(k\tau)) - F(w(k\tau))}{\text{denominator}}$$

$$\geq \frac{-\rho\,\| w(k\tau) - v_{[k]}(k\tau)\|}{\text{denom}}$$

ρ-Lipschitz

$$\geq -\frac{\rho\, h(k\tau - (k-1)\tau)}{\theta_{[k]}(k\tau)\,\theta_{[k+1]}(k\tau)} = \frac{-\rho h(\tau)}{\theta_{[k]}(\cdot)\,\theta_{[k+1]}(\cdot)}$$

Thm 1.

Since it is assumed that $\theta_{[k]}(t) = F(v_{[k]}(t)) - F(w^*) \geq \varepsilon \;\forall k, t$,

$$\theta_{[k]}(k\tau)\,\theta_{[k+1]}(k\tau) \geq \varepsilon^2$$

$$\frac{-1}{\theta_{[k]}(\cdot)\,\theta_{[k+1]}(\cdot)} \geq -\frac{1}{\varepsilon^2} \quad (\star\star\star)$$

go back to $(\star\star)$, we get

$$\sum_{k=1}^{K-1}\left(\frac{1}{\theta_{[k+1]}(k\tau)} - \frac{1}{\theta_{[k]}(k\tau)}\right) \geq -\sum_{k=1}^{K-1} \frac{\rho h(\tau)}{\varepsilon^2} = -(K-1)\frac{\rho h(\tau)}{\varepsilon^2}$$

and

$$\frac{1}{\theta_{[K]}(T)} - \frac{1}{\theta_{[1]}(0)} \geq T\omega\eta\left(1 - \frac{\beta\eta}{2}\right) - (K-1)\frac{\rho h(\tau)}{\varepsilon^2} \quad (\mathrm{I})$$

Similarly, it's assumed that $F(w(T)) - F(w^*) \geq \varepsilon$, then

$$\frac{-1}{\big(F(w(T)) - F(w^*)\big)\,\underset{\geq \varepsilon}{\theta_{[k]}(T)}} \geq -\frac{1}{\varepsilon^2} \qquad (\star\star\star)$$

assuming $\theta_{[k]}(t) = F(v_{[k]}(t)) - F(w^*) \geq \varepsilon$

Then,

$$\frac{1}{F(w(T)) - F(w^*)} - \frac{1}{\theta_{[k]}(T)}$$

$$= \frac{\theta_{[k]}(T) - \big(F(w(T)) - F(w^*)\big)}{\text{denominator}}$$

$$= \frac{F(v_{[k]}(T)) - F(w(T))}{\text{denom}}$$

ρ-Lipschitz $$\geq \frac{-\rho\,\| v_{[k]}(T) - w(T) \|}{\text{denom}}$$

$$\geq \frac{-\rho h\,(T - (k-1)\tau)}{\text{demon}}$$

(Thm 1). $\| w(t) - v_{[k]}(t) \| \leq h\,(t - (k-1)\tau)$

$$= \frac{-\rho h(\tau)}{\varepsilon^2} \qquad (\mathrm{II})$$

Summing up (I), (II), we have

$$\frac{1}{F(w(T)) - F(w^*)} - \frac{1}{\underset{>0}{\theta_{[1]}(0)}} \geq \left(T\omega\eta\left(1 - \frac{\beta\eta}{2}\right) - (K-1)\frac{\rho h(\tau)}{\varepsilon^2}\right) - \frac{\rho h(\tau)}{\varepsilon^2}$$

$$= T\omega\eta\left(1 - \frac{\beta\eta}{2}\right) - \frac{T\rho h(\tau)}{\tau\varepsilon^2}$$

$$= T\left(\omega\eta\left(1 - \frac{\beta\eta}{2}\right) - \frac{\rho h(\tau)}{\tau\varepsilon^2}\right)$$

Then $\dfrac{1}{F(w(T)) - F(w^*)} \geq \dfrac{1}{F(w(T)) - F(w^*)} - \dfrac{1}{\theta_{[1]}(0)}$

$$\geq T\left(\omega\eta\left(1 - \frac{\beta\eta}{2}\right) - \frac{\rho h(\tau)}{\tau\varepsilon^2}\right) > 0$$

because $\theta_{[1]}(0) = F(v_{[1]}(0)) - F(w^*) > 0$.

If we further assume $\omega\eta\left(1 - \frac{\beta\eta}{2}\right) - \frac{\rho h(\tau)}{\tau\varepsilon^2} > 0$,

then

$$F(w(T)) - F(w^*) \leq \frac{1}{T\left(\omega\eta\left(1 - \frac{\beta\eta}{2}\right) - \frac{\rho h(\tau)}{\tau\varepsilon^2}\right)}.$$

Remarks: $\omega \triangleq \min_k \dfrac{1}{\|v_{[k]}((k-1)\tau) - w^*\|^2}$ is used in Lemma 4, which we omitted the proof. $\square$