# Supplementary Material (Appendix)

## Appendix A. Auxiliary Lemmas

We begin with some helpful lemmas as follows:

**Lemma 16 (Berge's Maximum Theorem (Aliprantis et al., 2006))** *Let $\phi : \mathcal{X} \to \mathcal{Y}$ be a continuous correspondence between topological spaces with nonempty compact values, and suppose $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous. Define the "value function" $m : \mathcal{X} \to \mathbb{R}$ by*

$$m(x) = \max_{y \in \phi(x)} f(x, y) \ ,$$

*and the correspondence $\mu : \mathcal{X} \to \mathcal{Y}$ of maximizers by*

$$\mu(x) = \{y \in \phi(x) : f(x, y) = m(x)\} \ .$$

*Then*

- *The value function $m$ is continuous.*
- *The "$\arg\max$" correspondence $\mu$ is upper hemicontinuous and has nonempty compact values.*

**Lemma 17 (Kakutani's Fixed Point Theorem (Kakutani, 1941))** *Let $S$ be a compact nonempty convex subset of $\mathbb{R}^n$, and let $\phi : S \to 2^S$ be an upper hemicontinuous correspondence, and $\phi(x)$ is nonempty, closed and convex for every $x \in S$, then $\phi$ has a fixed point s.t. $x \in \phi(x)$.*

**Lemma 18 (Bubeck (Bubeck, 2015))** *Let $f : \mathbb{R} \to \mathbb{R}$ be a $\beta$-smooth, $\gamma$-strongly convex function, then for every $x, y \in \mathbb{R}$,*

$$(\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x - y) \geq \frac{\beta\gamma}{\beta + \gamma}\|x - y\|_2^2 + \frac{1}{\beta + \gamma}\|\nabla f(x) - \nabla f(y)\|_2^2 \ .$$

**Lemma 19 (Kantorovich-Rubinstein (Villani, 2009))** *A distribution map $\mathcal{D}(\cdot)$ is $\epsilon$-Lipschitz continuous if and only if for every $\theta, \theta' \in \Theta$,*

$$sup_g \left\{|E_{\boldsymbol{z} \sim D(\theta)} g(\boldsymbol{z}) - E_{\boldsymbol{z} \sim D(\theta')} g(\boldsymbol{z})|\right\} \leq \epsilon\|\theta - \theta'\|_2 \ ,$$

*where $g : \mathbb{R}^d \to \mathbb{R}, g$ is 1-Lipschitz function.*

**Lemma 20 (Banach Fixed Point Theorem (Aliprantis et al., 2006))** *Let $(X, d)$ be a complete metric space and for every function $f : X \to X$ if there exists a constant $0 \leq c < 1$ such that for $\forall x, y \in X, d(f(x), f(y)) \leq cd(x, y)$ then $f$ has a unique fixed point $x$. Moreover, for any choice $x_0 \in X$ the sequence defined recursively by*

$$x_{n+1} = f(x_n), \ n = 0, 1, 2, \cdots \ ,$$

*converges to the fixed point $x$ and satisfies that for every $n \in N$, we have*

$$d(x_n, x) \leq c^n d(x_0, x).$$

**Lemma 21 (Fournier&Guillin (Fournier and Guillin, 2015))** *Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of all probability measures on $\mathbb{R}^d$, define $\xi_{\alpha,\mu}(\mathcal{D}) = \int_{\mathbb{R}^d} e^{\mu\|\boldsymbol{z}\|^\alpha} d\mathcal{D}$. Assume that when $\exists\alpha > 1$, $\exists\mathcal{D} \in \mathcal{P}(\mathbb{R}^d), \exists\mu > 0$, $\xi_{\alpha,\mu}(\mathcal{D}) < \infty$ holds, then for every $N \geq 1$, $t \in (0, \infty)$,*

$$\mathbb{P}(W_1(\mathcal{D}^N, \mathcal{D}) \geq t) \leq a(N,t)1_{\{t\leq 1\}} + b(N,t) \ ,$$

*where $N$ is the number of sampled instances, and we define $a(N,t), b(N,t)$ as follows*

$$\begin{aligned} a(N,t) &= Cexp(-cNt^d) \ , \\ b(N,t) &= Cexp(-cNt^\alpha)1_{\{t>1\}} \ , \end{aligned}$$

*where $C, c$ are constants, and only depend on $d, \alpha, \mu, \xi_{\alpha,\mu}(\mathcal{D})$.*

## Appendix B. Proofs of Main Results

This section presents the detailed proofs of Example 1 and Theorems 6-15.

### B.1. Proof of the Lipschitz continuity for Example 1

Denote by $\mu(x;\theta^{\hat{1}},\theta^{\hat{2}}) = \frac{\epsilon}{2|x|}\left(|x\theta_* - h_{\hat{1}}(x;\theta^{\hat{1}})| - |x\theta_* - h_{\hat{2}}(x;\theta^{\hat{2}})|\right)$. In Example 1, the transport distance for each instance $(x, y)$ is bounded by the Euclidean distance between the shifted $y$. Specifically, for every $\theta^{\hat{1}}, \theta^{\hat{2}}, \theta^{\hat{1}'}, \theta^{\hat{2}'} \in \Theta$, $y$ would shift from $x\theta_* + \mu(x;\theta^{\hat{1}},\theta^{\hat{2}})$ to $x\theta_* + \mu(x;\theta^{\hat{1}'},\theta^{\hat{2}'})$, and it can be bounded as follows:

$$\begin{aligned} &|x\theta_* + \mu(x;\theta^{\hat{1}},\theta^{\hat{2}}) - x\theta_* - \mu(x;\theta^{\hat{1}'},\theta^{\hat{2}'})| \\ &= \frac{\epsilon}{2|x|}\big||x\theta_* - x\theta^{\hat{1}} + b| - |x\theta_* - x\theta^{\hat{2}} - b| - |x\theta_* - x\theta^{\hat{1}'} + b| + |x\theta_* - x\theta^{\hat{2}'} - b|\big| \\ &\leq \frac{\epsilon}{2|x|}\left(|x\theta^{\hat{1}'} - x\theta^{\hat{1}}| + |x\theta^{\hat{2}} - x\theta^{\hat{2}'}|\right) \\ &\leq \frac{\epsilon}{2|x|}\left(|\theta^{\hat{1}'} - \theta^{\hat{1}}||x| + |\theta^{\hat{2}'} - \theta^{\hat{2}}||x|\right) \\ &\leq \epsilon\big\|(\theta^{\hat{1}'},\theta^{\hat{2}'}) - (\theta^{\hat{1}},\theta^{\hat{2}})\big\|_2 \ . \end{aligned}$$

Therefore, the optimal transport distance between any pair of distributions $\mathcal{D}(\theta^{\hat{1}},\theta^{\hat{2}})$ and $\mathcal{D}(\theta^{\hat{1}'},\theta^{\hat{2}'})$ can be upper bounded in the same way, i.e.,

$$W_1(\mathcal{D}(\theta^{\hat{1}},\theta^{\hat{2}}), \mathcal{D}(\theta^{\hat{1}'},\theta^{\hat{2}'})) \leq \epsilon\big\|(\theta^{\hat{1}'},\theta^{\hat{2}'}) - (\theta^{\hat{1}},\theta^{\hat{2}})\big\|_2 \ .$$

Hence, the distribution in the Example 1 follows the $\epsilon$-Lipschitz continuity. □

### B.2. Proof of Theorem 6

We begin with model $\hat{1}$. We denote by $g^{\hat{1}}(\theta) = \arg\min_{\theta^{\hat{1}}\in\Theta} \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta)}\ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}})$ and $f^{\hat{1}}(\theta,\theta^{\hat{1}}) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta)}\ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}})$, for $\theta \in \Theta^2, \theta^{\hat{1}} \in \Theta$, for simplicity.

Firstly, with the $\epsilon$-Lipschitz continuity for the distribution $\mathcal{D}(\theta)$ and the continuity for the loss function $\ell_{\hat{1}}$, it is obvious that $f^{\hat{1}}(\theta,\theta^{\hat{1}})$ is also continuous. Since the parameter space $\Theta$ is nonempty and compact, by applying Berge's Maximum Theorem (Lemma 16), we can conclude that $g^{\hat{1}}(\theta) = \arg\min_{\theta_{\hat{1}} \in \Theta} f^{\hat{1}}(\theta, \theta^{\hat{1}})$ is upper hemicontinuous with nonempty compact values for every $\theta \in \Theta^2$. The same statement also holds for model $\hat{2}$. Let $g(\theta) = g^{\hat{1}}(\theta) \times g^{\hat{2}}(\theta)$ be the Cartesian product of $g^{\hat{1}}$ and $g^{\hat{2}}$, and apparently $g(\theta)$ is also upper hemicontinuous.

Therefore loss functions $\ell_{\hat{1}}$ and $\ell_{\hat{2}}$ are convex, $g(\theta)$ is non-empty and convex set for every $\theta$, and is also upper hemicontinuous. With Kakutani's Fixed Point Theorem (Lemma 17), we can conclude that $g(\theta)$ has a fixed point, which completes the proof. □

### B.3. Proof of Theorem 7

For simplicity, we first denote by $G(\theta) = (G_{\hat{1}}(\theta), G_{\hat{2}}(\theta))$ with

$$G_{\hat{i}}(\theta) = \arg\min_{\theta^{\hat{i}} \in \Theta} \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}(\theta)} \left[\ell_{\hat{i}}(\boldsymbol{z}; \theta^{\hat{i}})\right] \quad \text{for} \quad i \in [2].$$

Here, $G_{\hat{i}}(\theta)$ denotes the output in one-step by RRM with input $\theta = (\theta^{\hat{1}}, \theta^{\hat{2}})$. For every $\theta_1, \theta_2 \in \Theta^2$, denote by

$$f_1^{\hat{i}}(\theta) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}(\theta_1)}[\ell_{\hat{i}}(\boldsymbol{z}, \theta)] \text{ and } f_2^{\hat{i}}(\theta) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{D}(\theta_2)}[\ell_{\hat{i}}(\boldsymbol{z}, \theta)] \quad \text{for} \quad i \in [2].$$

Since $f_1^{\hat{1}}$ is $\gamma_{\hat{1}}$-strongly convex, we have

$$\begin{aligned} f_1^{\hat{1}}(G_{\hat{1}}(\theta_1)) - f_1^{\hat{1}}(G_{\hat{1}}(\theta_2)) \geq\ & G_{\hat{1}}(\theta_1)^{\mathrm{T}} \nabla f_1^{\hat{1}}(G_{\hat{1}}(\theta_2)) - G_{\hat{1}}(\theta_2)^{\mathrm{T}} \nabla f_1^{\hat{1}}(G_{\hat{1}}(\theta_2)) \\ & + \frac{\gamma_{\hat{1}}}{2} \|G_{\hat{1}}(\theta_1) - G_{\hat{1}}(\theta_2)\|_2^2 \ , \end{aligned} \tag{1}$$

$$\begin{aligned} f_1^{\hat{1}}(G_{\hat{1}}(\theta_2)) - f_1^{\hat{1}}(G_{\hat{1}}(\theta_1)) \geq\ & G_{\hat{1}}(\theta_2)^{\mathrm{T}} \nabla f_1^{\hat{1}}(G_{\hat{1}}(\theta_1)) - G_{\hat{1}}(\theta_1)^{\mathrm{T}} \nabla f_1^{\hat{1}}(G_{\hat{1}}(\theta_1)) \\ & + \frac{\gamma_{\hat{1}}}{2} \|G_{\hat{1}}(\theta_2) - G_{\hat{1}}(\theta_1)\|_2^2 \ . \end{aligned} \tag{2}$$

Because $G_{\hat{1}}(\theta_1)$ minimizes $f_1^{\hat{1}}(\theta)$, with the first-order optimality condition, we can infer that

$$(G_{\hat{1}}(\theta_2) - G_{\hat{1}}(\theta_1))^{\mathrm{T}} \nabla f_1^{\hat{1}}(G_{\hat{1}}(\theta_1)) \geq 0 \ . \tag{3}$$

Therefore, with Eqns. (2) and (3), we have

$$\begin{aligned} f_1^{\hat{1}}(G_{\hat{1}}(\theta_2)) - f_1^{\hat{1}}(G_{\hat{1}}(\theta_1)) &\geq (G_{\hat{1}}(\theta_2) - G_{\hat{1}}(\theta_1))^{\mathrm{T}} \nabla f_1^{\hat{1}}(G_{\hat{1}}(\theta_1)) + \frac{\gamma_{\hat{1}}}{2} \|G_{\hat{1}}(\theta_2) - G_{\hat{1}}(\theta_1)\|_2^2 \\ &\geq \frac{\gamma_{\hat{1}}}{2} \|G_{\hat{1}}(\theta_2) - G_{\hat{1}}(\theta_1)\|_2^2 \ . \end{aligned} \tag{4}$$

Together with Eqns. (1) and (4), it follows that

$$-\gamma_{\hat{1}} \|G_{\hat{1}}(\theta_1) - G_{\hat{1}}(\theta_2)\|_2^2 \geq (G_{\hat{1}}(\theta_1) - G_{\hat{1}}(\theta_2))^{\mathrm{T}} \nabla f_1^{\hat{1}}(G_{\hat{1}}(\theta_2)) \ .$$

It also holds for model $\hat{2}$ that

$$-\gamma_{\hat{2}} \|G_{\hat{2}}(\theta_1) - G_{\hat{2}}(\theta_2)\|_2^2 \geq (G_{\hat{2}}(\theta_1) - G_{\hat{2}}(\theta_2))^{\mathrm{T}} \nabla f_1^{\hat{2}}(G_{\hat{2}}(\theta_2)) \ .$$

Adding two inequalities together, we have

$$\begin{aligned}
&-\gamma_{\min}\|G(\theta_1)-G(\theta_2)\|_2^2\\
&\geq (G_{\hat{1}}(\theta_1)-G_{\hat{1}}(\theta_2))^{\mathrm{T}}\nabla f_1^{\hat{1}}(G_{\hat{1}}(\theta_2))+(G_{\hat{2}}(\theta_1)-G_{\hat{2}}(\theta_2))^{\mathrm{T}}\nabla f_1^{\hat{2}}(G_{\hat{2}}(\theta_2))\\
&=(G(\theta_1)-G(\theta_2))^{\mathrm{T}}\nabla f_1(G(\theta_2))\ ,
\end{aligned}\tag{5}$$

where $\gamma_{\min}=\min\{\gamma_{\hat{1}},\gamma_{\hat{2}}\}$, and $\nabla f_1(G(\theta_2))=(\nabla f_1^{\hat{1}}(G_{\hat{1}}(\theta_2));\nabla f_1^{\hat{2}}(G_{\hat{2}}(\theta_2)))$ is the concatenation of two gradient vectors.

Then, to prove the target, we only need to find the relationship between $\|\theta_1-\theta_2\|_2$ and $(G(\theta_1)-G(\theta_2))^{\mathrm{T}}\nabla f_1(G(\theta_2))$ . With Cauchy–Schwarz inequality and the $\beta$-smoothness of losses $\ell_{\hat{1}},\ell_{\hat{2}}$, we can infer that $(G(\theta_1)-G(\theta_2))^{\mathrm{T}}\nabla_\theta\ell(\boldsymbol{z};G(\theta_2))$ is $\tilde{\beta}\|G(\theta_1)-G(\theta_2)\|_2$-Lipschitz continuous in $\boldsymbol{z}$, where $\tilde{\beta}=(\beta_{\hat{1}}^2+\beta_{\hat{2}}^2)^{1/2}$. Denote by $g(\boldsymbol{z})=(G(\theta_1)-G(\theta_2))^{\mathrm{T}}\nabla_\theta\ell(\boldsymbol{z};G(\theta_2))$, with Lemma 19 we can derive that

$$(G(\theta_1)-G(\theta_2))^{\mathrm{T}}(\nabla f_1(\boldsymbol{z};G(\theta_2))-\nabla f_2(\boldsymbol{z};G(\theta_2)))\geq-\epsilon\tilde{\beta}\|G(\theta_1)-G(\theta_2)\|_2\|\theta_1-\theta_2\|_2\ .$$

While $G_{\hat{i}}$ minimizes $f^{\hat{i}}$ for $i\in[2]$, using the first-order optimality condition, we have

$$(G(\theta_1)-G(\theta_2))^{\mathrm{T}}\nabla f_2(G(\theta_2))\geq 0\ .$$

This follows that

$$(G(\theta_1)-G(\theta_2))^{\mathrm{T}}\nabla f_1(\boldsymbol{z};G(\theta_2))\geq-\epsilon\tilde{\beta}\|G(\theta_1)-G(\theta_2)\|_2\|\theta_1-\theta_2\|_2\ .\tag{6}$$

Finally, with Eqns. (5) and (6), we have

$$\begin{aligned}
-\gamma_{\min}\|G(\theta_1)-G(\theta_2)\|_2^2\geq&(G(\theta_1)-G(\theta_2))^{\mathrm{T}}\nabla f_1(G(\theta_2))\\
\geq&-\epsilon\tilde{\beta}\|G(\theta_1)-G(\theta_2)\|_2\|\theta_1-\theta_2\|_2\ .
\end{aligned}$$

By eliminating $\|G(\theta_1)-G(\theta_2)\|_2$ from both sides, we further have

$$\|G(\theta_1)-G(\theta_2)\|_2\leq\epsilon\frac{\tilde{\beta}}{\gamma_{\min}}\|\theta_1-\theta_2\|_2\ .$$

Finally, we only need to put $\theta_{\mathrm{DS}}$ and $\theta_{T-1}$ into the inequality above and get

$$\begin{aligned}
\|\theta_T-\theta_{\mathrm{DS}}\|_2&=\|(G_{\hat{1}}(\theta_{T-1}),G_{\hat{2}}(\theta_{T-1}))-(G_{\hat{1}}(\theta_{\mathrm{DS}}),G_{\hat{2}}(\theta_{\mathrm{DS}}))\|_2\\
&\leq\left(\epsilon(\beta_{\hat{1}}^2+\beta_{\hat{2}}^2)^{\frac{1}{2}}\big/\min\{\gamma_{\hat{1}},\gamma_{\hat{2}}\}\right)\|\theta_{T-1}-\theta_{\mathrm{DS}}\|_2\\
&\leq\left(\epsilon(\beta_{\hat{1}}^2+\beta_{\hat{2}}^2)^{\frac{1}{2}}\big/\min\{\gamma_{\hat{1}},\gamma_{\hat{2}}\}\right)^T\|\theta_0-\theta_{\mathrm{DS}}\|_2\ ,
\end{aligned}$$

which completes the proof. □

### B.4. Proof of Theorem 9

The proof sketch follows that we consider two cases: when $\|\theta_t-\theta_{\mathrm{DS}}\|_2>r$ and when $\|\theta_t-\theta_{\mathrm{DS}}\|_2\leq r$.

Let $\mathcal{D}^{n_t}(\theta)$ denotes the empirical distribution consists of training sample $S_{n_t}$ drawn i.i.d. from distribution $\mathcal{D}(\theta)$, then the update function for two models in RERM can be written as

$$G_i^{n_t}(\theta)=\arg\min\nolimits_{\theta^i\in\Theta}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta)}\ell_i(\boldsymbol{z};\theta^{\hat{i}})\quad\text{for}\quad i\in[2].$$

**For $\|\theta_t - \theta_{\mathbf{DS}}\|_2 > r$:**
With the triangle inequality for Wasserstein-1 distance, we have

$$W_1(D^{n_t}(\theta_t), D(\theta_{\text{DS}})) \leq W_1(D^{n_t}(\theta_t), D(\theta_t)) + W_1(D(\theta_t), D(\theta_{\text{DS}})) \ . \tag{7}$$

For $W_1(D(\theta_t), D(\theta_{\text{DS}}))$, since $\mathcal{D}(\cdot)$ is $\epsilon$-Lipschitz continuous, which reveals that

$$W_1(D(\theta_t), D(\theta_{\text{DS}})) \leq \epsilon \|\theta_t - \theta_{\text{DS}}\|_2 \ . \tag{8}$$

As for $W_1(D^{n_t}(\theta_t), D(\theta_t))$, by applying Lemma 21, when $\epsilon r = x < 1$, we have

$$\bar{\delta} = \mathbb{P}(W_1(D^{n_t}(\theta_t), D(\theta_t)) \geq \epsilon r) \leq c_1 exp(-c_2 n_t (\epsilon r)^{2m}) \ .$$

A short calculation infers that

$$n_t = \frac{1}{c_2(\epsilon r)^{2m}} log(\frac{c_1}{\bar{\delta}}) \ .$$

Denote by $\bar{\delta} = 1 - \sqrt{\pi^2 t^2 - 6\delta}/(\pi t)$, when $n_t \geq \log\left(\pi t c_1/(\pi t - \sqrt{\pi^2 t^2 - 6\delta})\right)/c_2(\epsilon r)^{2m}$, we have

$$\mathbb{P}(W_1(D^{n_t}(\theta_t), D(\theta_t)) \geq \epsilon r) \leq \frac{6\delta}{\pi^2 t^2} \ , \tag{9}$$

where $c_1, c_2$ are constants and only rely on $m, \alpha, \mu, \xi_{\alpha,\mu}(\mathcal{D})$.
Therefore, by Eqns. (7)-(9), with probability at least $\sqrt{\pi^2 t^2 - 6\delta}/(\pi t)$, we have that

$$W_1(D^{n_t}(\theta_t), D(\theta_{\text{DS}})) \leq \epsilon r + \epsilon \|\theta_t - \theta_{\text{DS}}\|_2 \leq 2\epsilon \|\theta_t - \theta_{\text{DS}}\|_2 \ . \tag{10}$$

For model $\hat{1}$, with the first-order optimality condition, we have

$$\begin{aligned} (G_{\hat{1}}^{n_t}(\theta_t) - G_{\hat{1}}(\theta_{\text{DS}}))^{\text{T}} \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\text{DS}})} \nabla_\theta \ell_{\hat{1}}(\boldsymbol{z}; G_{\hat{1}}(\theta_{\text{DS}})) &\geq 0 \ , \\ (G_{\hat{1}}^{n_t}(\theta_t) - G_{\hat{1}}(\theta_{\text{DS}}))^{\text{T}} \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)} \nabla_\theta \ell_{\hat{1}}(\boldsymbol{z}; G_{\hat{1}}^{n_t}(\theta_t)) &\leq 0 \ , \end{aligned}$$

and this follows that,

$$(G_{\hat{1}}^{n_t}(\theta_t) - G_{\hat{1}}(\theta_{\text{DS}}))^{\text{T}} \left( \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)} \nabla_\theta \ell_{\hat{1}}(\boldsymbol{z}; G_{\hat{1}}^{n_t}(\theta_t)) \ - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\text{DS}})} \nabla_\theta \ell_{\hat{1}}(\boldsymbol{z}; G_{\hat{1}}(\theta_{\text{DS}})) \right) \leq 0 \ .$$

A further calculation reveals that

$$\begin{aligned} &(G_{\hat{1}}^{n_t}(\theta_t) - G_{\hat{1}}(\theta_{\text{DS}}))^{\text{T}} \left( \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)} \nabla_\theta \ell_{\hat{1}}(\boldsymbol{z}; G_{\hat{1}}^{n_t}(\theta_t)) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\text{DS}})} \nabla_\theta \ell_{\hat{1}}(\boldsymbol{z}; G_{\hat{1}}^{n_t}(\theta_t)) \right) \\ &+ (G_{\hat{1}}^{n_t}(\theta_t) - G_{\hat{1}}(\theta_{\text{DS}}))^{\text{T}} \left( \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\text{DS}})} \nabla_\theta \ell_{\hat{1}}(\boldsymbol{z}; G_{\hat{1}}^{n_t}(\theta_t)) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\text{DS}})} \nabla_\theta \ell_{\hat{1}}(\boldsymbol{z}; G_{\hat{1}}(\theta_{\text{DS}})) \right) \\ &\leq 0 \ . \end{aligned}$$

For model $\hat{2}$, we have the same conclusion, i.e. there is probability at least $\sqrt{\pi^2 t^2 - 6\delta}/(\pi t)$ that

$$\begin{aligned} &(G_{\hat{2}}^{n_t}(\theta_t) - G_{\hat{2}}(\theta_{\text{DS}}))^{\text{T}} \left( \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)} \nabla_\theta \ell_{\hat{2}}(\boldsymbol{z}; G_{\hat{2}}^{n_t}(\theta_t)) \ - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\text{DS}})} \nabla_\theta \ell_{\hat{2}}(\boldsymbol{z}; G_{\hat{2}}^{n_t}(\theta_t)) \right) \\ &+ (G_{\hat{2}}^{n_t}(\theta_t) - G_{\hat{2}}(\theta_{\text{DS}}))^{\text{T}} \left( \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\text{DS}})} \nabla_\theta \ell_{\hat{2}}(\boldsymbol{z}; G_{\hat{2}}^{n_t}(\theta_t)) \ - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\text{DS}})} \nabla_\theta \ell_{\hat{2}}(\boldsymbol{z}; G_{\hat{2}}(\theta_{\text{DS}})) \right) \\ &\leq 0 \ . \end{aligned}$$

Adding two inequalities together, we have

$$\begin{aligned}&(G^{n_t}(\theta_t)-G(\theta_{\mathrm{DS}}))^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta\ell(\boldsymbol{z};G^{n_t}(\theta_t))-\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};G^{n_t}(\theta_t))\right)\\&+(G^{n_t}(\theta_t)-G(\theta_{\mathrm{DS}}))^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};G^{n_t}(\theta_t))-\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};G(\theta_{\mathrm{DS}}))\right)\le 0\ .\end{aligned}\tag{11}$$

The probability of that Eqn. (11) holds is at least $(\sqrt{\pi^2t^2-6\delta}/(\pi t))^2=1-6\delta/(\pi^2t^2)$. With Cauchy–Schwarz inequality and the $\beta$-smoothness of loss $\ell_{\hat{1}},\ell_{\hat{2}}$, we can easily infer that $(G^{n_t}(\theta_t)-G(\theta_{\mathrm{DS}}))^{\mathrm{T}}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta\ell(\boldsymbol{z};G^{n_t}(\theta_t))$ is $\tilde{\beta}\|G^{n_t}(\theta_t)-G(\theta_{\mathrm{DS}})\|_2$-Lipschitz continuous in $\boldsymbol{z}$, where $\tilde{\beta}=(\beta_{\hat{1}}^2+\beta_{\hat{2}}^2)^{1/2}$. By applying Eqn. (10) and Lemma 19, we have that

$$\begin{aligned}&(G^{n_t}(\theta_t)-G(\theta_{\mathrm{DS}}))^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta\ell(\boldsymbol{z};G^{n_t}(\theta_t))\ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};G^{n_t}(\theta_t))\right)\\&\ge -2\epsilon\tilde{\beta}\|G^{n_t}(\theta_t)-G(\theta_{\mathrm{DS}})\|_2\|\theta_t-\theta_{\mathrm{DS}}\|_2\ .\end{aligned}\tag{12}$$

As for the second term in Eqn. (11), since $\ell_{\hat{1}}$ is $\gamma_{\hat{1}}$-strongly convex and $\ell_{\hat{2}}$ is $\gamma_{\hat{2}}$-strongly convex, we have

$$\begin{aligned}&(G_{\hat{1}}^{n_t}(\theta_t)-G_{\hat{1}}(\theta_{\mathrm{DS}}))^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell_{\hat{1}}(\boldsymbol{z};G_{\hat{1}}^{n_t}(\theta_t))\ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell_{\hat{1}}(\boldsymbol{z};G_{\hat{1}}(\theta_{\mathrm{DS}}))\right)\\&\ge\gamma_{\hat{1}}\|G_{\hat{1}}^{n_t}(\theta_t)-G_{\hat{1}}(\theta_{\mathrm{DS}})\|_2^2\ .\\&(G_{\hat{2}}^{n_t}(\theta_t)-G_{\hat{2}}(\theta_{\mathrm{DS}}))^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell_{\hat{2}}(\boldsymbol{z};G_{\hat{2}}^{n_t}(\theta_t))\ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell_{\hat{2}}(\boldsymbol{z};G_{\hat{2}}(\theta_{\mathrm{DS}}))\right)\\&\ge\gamma_{\hat{2}}\|G_{\hat{2}}^{n_t}(\theta_t)-G_{\hat{2}}(\theta_{\mathrm{DS}})\|_2^2\ .\end{aligned}$$

Thus, we further have

$$\begin{aligned}&(G^{n_t}(\theta_t)-G(\theta_{\mathrm{DS}}))^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};G^{n_t}(\theta_t))\ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};G(\theta_{\mathrm{DS}}))\right)\\&\ge\gamma_{\min}\|G^{n_t}(\theta_t)-G(\theta_{\mathrm{DS}})\|_2^2\ .\end{aligned}\tag{13}$$

With Eqns. (11)-(13), we can conclude with probability at least $1-6\delta/(\pi^2t^2)$ that

$$\|G_{\hat{1}}^{n_t}(\theta_t)-G_{\hat{1}}(\theta_{\mathrm{DS}})\|_2\le\frac{2\epsilon\tilde{\beta}}{\gamma_{\min}}\|\theta_t-\theta_{\mathrm{DS}}\|_2\ .\tag{14}$$

Note that the probability of which for all $t\in N^*$, Eqn. (14) holds at the same time can be calculated as follows:

$$\prod_{t=1}^{\infty}(1-\frac{6\delta}{\pi^2t^2})\ge 1-\sum_{t=1}^{\infty}\frac{6\delta}{\pi^2t^2}=1-\delta\ .$$

**For $\|\theta_t-\theta_{\mathbf{DS}}\|_2\le r$:**
Similarly, by applying triangle inequality, we have

$$\begin{aligned}W_1(D^{n_t}(\theta_t),D(\theta_{\mathrm{DS}}))&\le W_1(D^{n_t}(\theta_t),D(\theta_t))+W_1(D(\theta_t),D(\theta_{\mathrm{DS}}))\\&\le W_1(D^{n_t}(\theta_t),D(\theta_t))+\epsilon r\ .\end{aligned}$$

By Lemma 21, with the probability at least $\sqrt{\pi^2t^2-6\delta}/(\pi t)$,

$$W_1(D^{n_t}(\theta_t),D(\theta_{\mathrm{DS}}))\le\epsilon r+\epsilon r=2\epsilon r\ .$$

With similarly proof, by Lemma 19, we have

$$\|G^{n_t}(\theta_t)-G(\theta_{\mathrm{DS}})\|_2=\|\theta_{t+1}-\theta_{\mathrm{DS}}\|_2\le\frac{2\epsilon\tilde{\beta}}{\gamma_{\min}}r\ .\tag{15}$$

This completes the proof by combine Eqn. (14) with Eqn. (15). □

### B.5. Proof of Theorem 11

We can firstly prove that for every $\theta_1, \theta_2 \in \Theta^2$, the following holds.

$$\|G_{\mathrm{gd}}(\theta_1) - G_{\mathrm{gd}}(\theta_2)\|_2 \leq \Big(1 - \eta\big(\frac{\beta\gamma}{\beta+\gamma} - \epsilon\beta(\sqrt{2} + 2\eta\beta + \eta\epsilon\beta)\big)\Big)\|\theta_1 - \theta_2\|_2 \ ,$$

where $G_{\mathrm{gd},i}(\theta) = \Pi_\Theta\big(\theta^{\hat{i}} - \eta\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta)}\nabla_\theta \ell_{\hat{i}}(\boldsymbol{z};\theta^{\hat{i}})\big)$ for $i \in [2]$ denote the update of the parameter for each model in RGD.

Because projecting onto a convex set can only bring two iterates closer together, we would ignore the projection operator in the following proof. Thus, the update formulas for each model can be shown as follows.

$$\begin{aligned}
\theta^{\hat{1}}_{t+1} &= G_{\mathrm{gd},\hat{1}}(\theta_t) = \theta^{\hat{1}}_t - \eta\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_t)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}}_t) \ , \\
\theta^{\hat{2}}_{t+1} &= G_{\mathrm{gd},\hat{2}}(\theta_t) = \theta^{\hat{2}}_t - \eta\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_t)}\nabla_\theta \ell_{\hat{2}}(\boldsymbol{z};\theta^{\hat{2}}_t) \ .
\end{aligned}$$

Firstly, we consider model $\hat{1}$, and we have

$$\begin{aligned}
&\|G_{\mathrm{gd},\hat{1}}(\theta_1) - G_{\mathrm{gd},\hat{1}}(\theta_2)\|_2^2 \\
&= \left\|\theta^{\hat{1}}_1 - \eta\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}}_1) - \theta^{\hat{1}}_2 + \eta\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}}_2)\right\|_2^2 \\
&= \|\theta^{\hat{1}}_1 - \theta^{\hat{1}}_2\|_2^2 - 2\eta(\theta^{\hat{1}}_1 - \theta^{\hat{1}}_2)^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}}_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}}_2)\right) \\
&\quad + \eta^2\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}}_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}}_2)\right\|_2^2 \ .
\end{aligned}$$

The equation above would also holds for model $\hat{2}$. Adding them together, we have

$$\begin{aligned}
\|G_{\mathrm{gd}}(\theta_1) - G_{\mathrm{gd}}(\theta_2)\|_2^2 &= \|G_{\mathrm{gd},\hat{1}}(\theta_1) - G_{\mathrm{gd},\hat{1}}(\theta_2)\|_2^2 + \|G_{\mathrm{gd},\hat{2}}(\theta_1) - G_{\mathrm{gd},\hat{2}}(\theta_2)\|_2^2 \\
&= \|\theta_1 - \theta_2\|_2^2 - 2\eta(\theta_1 - \theta_2)^{\mathrm{T}}\big(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\big) \\
&\quad + \eta^2\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2^2 \\
&= T_1 - 2\eta T_2 + \eta^2 T_3 \ .
\end{aligned}$$

where $\nabla_\theta \ell(\boldsymbol{z};\theta) = (\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}}); \nabla_\theta \ell_{\hat{2}}(\boldsymbol{z};\theta^{\hat{2}}))$ is the concatenation of two gradient vectors, and $T_1, T_2, T_3$ is defined as below

$$\begin{aligned}
T_1 &= \|\theta_1 - \theta_2\|_2^2 \ , \\
T_2 &= (\theta_1 - \theta_2)^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right) \ , \\
T_3 &= \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2^2 \ .
\end{aligned}$$

Next, we analyze each term individually. Starting with $T_2$, we have

$$\begin{aligned}
T_2 =&(\theta_1 - \theta_2)^{\mathrm{T}}\big(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) \\
&+ \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\big) \\
=&(\theta_1 - \theta_2)^{\mathrm{T}}\big(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\big) \\
&+ (\theta_1 - \theta_2)^{\mathrm{T}}\big(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\big) \ .
\end{aligned} \tag{16}$$

Applying Cauchy–Schwarz inequality to the first term of Eqn. (16), we have

$$\begin{aligned} T_2 \geq & -\|\theta_1-\theta_2\|_2 \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2 \\ & + (\theta_1-\theta_2)^{\mathrm{T}} \left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right) \ . \end{aligned} \tag{17}$$

With the $\epsilon$-Lipschitz continuity of distribution $\mathcal{D}(\theta)$ and $\beta$-smoothness of losses $\ell_{\hat{1}}, \ell_{\hat{2}}$, by Lemma 19, it reveals that

$$\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_1)\right\|_2 \leq \tilde{\beta}\epsilon\|\theta_1-\theta_2\|_2 \ ,$$

where $\tilde{\beta} = (\beta_{\hat{1}}^2 + \beta_{\hat{2}}^2)^{1/2}$. Put it back to Eqn. (17), we have

$$T_2 \geq -\tilde{\beta}\epsilon\|\theta_1-\theta_2\|_2^2 + (\theta_1-\theta_2)^{\mathrm{T}} \left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right) \ .$$

As for the second term of Eqn. (16), because loss $\ell_{\hat{1}}, \ell_{\hat{2}}$ are both $\beta_{\max}$-jointly smooth and $\gamma_{\min}$-strongly convex, where $\beta_{\max} = \max\{\beta_{\hat{1}}, \beta_{\hat{2}}\}$ and $\gamma_{\min} = \min\{\gamma_{\hat{1}}, \gamma_{\hat{2}}\}$. With Lemma 18, we have

$$\begin{aligned} & (\theta_1-\theta_2)^{\mathrm{T}} \left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right) \\ & \quad = \quad (\theta_1^{\hat{1}}-\theta_2^{\hat{1}})^{\mathrm{T}} \left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta_1^{\hat{1}}) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta_2^{\hat{1}})\right) \\ & \qquad\quad + (\theta_1^{\hat{2}}-\theta_2^{\hat{2}})^{\mathrm{T}} \left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell_{\hat{2}}(\boldsymbol{z};\theta_1^{\hat{2}}) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell_{\hat{2}}(\boldsymbol{z};\theta_2^{\hat{2}})\right) \\ & \quad \geq \quad \frac{\beta_{\max}\gamma_{\min}}{\beta_{\max}+\gamma_{\min}}\|\theta_1^{\hat{1}}-\theta_2^{\hat{1}}\|_2^2 + \frac{1}{\beta_{\max}+\gamma_{\min}}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\left\|\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta_1^{\hat{1}}) - \nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta_2^{\hat{1}})\right\|_2^2 \\ & \qquad\quad + \frac{\beta_{\max}\gamma_{\min}}{\beta_{\max}+\gamma_{\min}}\|\theta_1^{\hat{2}}-\theta_2^{\hat{2}}\|_2^2 + \frac{1}{\beta_{\max}+\gamma_{\min}}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\left\|\nabla_\theta \ell_{\hat{2}}(\boldsymbol{z};\theta_1^{\hat{2}}) - \nabla_\theta \ell_{\hat{2}}(\boldsymbol{z};\theta_2^{\hat{2}})\right\|_2^2 \\ & \quad \geq \quad \frac{\beta_{\max}\gamma_{\min}}{\beta_{\max}+\gamma_{\min}}\|\theta_1-\theta_2\|_2^2 + \frac{1}{\beta_{\max}+\gamma_{\min}}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\|\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \nabla_\theta \ell(\boldsymbol{z};\theta_2)\|_2^2 \\ & \quad \geq \quad \frac{\beta_{\max}\gamma_{\min}}{\beta_{\max}+\gamma_{\min}}\|\theta_1-\theta_2\|_2^2 + \frac{1}{\beta_{\max}+\gamma_{\min}}\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\left[\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \nabla_\theta \ell(\boldsymbol{z};\theta_2)\right]\right\|_2^2 \ . \end{aligned}$$

Hence, we finally can get the following inequality for $T_2$:

$$T_2 \geq (-\epsilon\tilde{\beta} + \frac{\beta_{\max}\gamma_{\min}}{\beta_{\max}+\gamma_{\min}})\|\theta_1-\theta_2\|_2^2 + \frac{1}{\beta_{\max}+\gamma_{\min}}\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\left[\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \nabla_\theta \ell(\boldsymbol{z};\theta_2)\right]\right\|_2^2 \ .$$

As for $T_3$, we have

$$\begin{aligned} T_3 = & \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right. \\ & \left. + \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2^2 \\ = & \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2^2 \\ & + \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2^2 \\ & + 2\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right)^{\mathrm{T}} \\ & \cdot \left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right) \ . \end{aligned} \tag{18}$$

For the second term in Eqn. (18), with the same calculation in Eqn. (B.5), we can infer that

$$\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2^2 \le \epsilon^2\tilde{\beta}^2\|\theta_1-\theta_2\|_2^2 \ .$$

As for the third term in Eqn. (18), we can define the unit vector $v$ as follows.

$$v = \frac{\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)}{\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2} \ .$$

This follows that

$$\begin{aligned}
&2\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right)^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right)\\
&\ = 2\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2\\
&\quad \cdot\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right)^{\mathrm{T}} v\\
&\ = 2\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2\\
&\quad \cdot\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right) \ .
\end{aligned} \tag{19}$$

Similarly, with $\beta$-smoothness, $v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta)$ is $\tilde{\beta}$-Lipschitz continuous. By the $\epsilon$-Lipschitz continuity of $\mathcal{D}(\theta)$, with Lemma 19, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_1^{\hat{1}}) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_1^{\hat{1}})\\
&= \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_1^{\hat{1}}) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_1^{\hat{1}})\right\|_2\\
&\le \tilde{\beta}\epsilon\|\theta_1-\theta_2\|_2 \ .
\end{aligned} \tag{20}$$

By the definition of $\beta$-smoothness, we have

$$\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2 \le \beta_{\max}\|\theta_1-\theta_2\|_2 \ . \tag{21}$$

Therefore, put Eqns. (20) and (21) back to Eqn. (19), we can infer that

$$\begin{aligned}
&2\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right)^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_2)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right)\\
&\ \le 2\epsilon\beta_{\max}\tilde{\beta}\|\theta_1-\theta_2\|_2^2 \ .
\end{aligned}$$

Hence, we can bound $T_3$ as

$$T_3 \le (\epsilon^2\tilde{\beta}^2 + 2\epsilon\tilde{\beta}\beta_{\max})\|\theta_1-\theta_2\|_2^2 + \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2^2 \ .$$

Having bounded all the terms, we can finally conclude that

$$\begin{aligned}
\|G_{\mathrm{gd}}(\theta_1) - G_{\mathrm{gd}}(\theta_2)\|_2^2 \le &\left(1 - \frac{2\eta\beta_{\max}\gamma_{\min}}{\beta_{\max}+\gamma_{\min}} + 2\eta\epsilon\tilde{\beta} + 2\eta^2\epsilon\tilde{\beta}\beta_{\max} + \eta^2\epsilon^2\beta_{\max}^2\right)\|\theta_1-\theta_2\|_2^2\\
&- \left(\frac{2\eta}{\beta_{\max}+\gamma_{\min}} - \eta^2\right)\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_1) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_1)}\nabla_\theta \ell(\boldsymbol{z};\theta_2)\right\|_2^2 .
\end{aligned}$$

Because $\tilde{\beta}/\sqrt{2} \leq \beta_{\max} \leq \tilde{\beta}$, if the step size $\eta$ small enough, satisfying $\eta \leq 2/(\beta_{\max}+\gamma_{\min}) \leq 2\sqrt{2}/(\tilde{\beta}+\sqrt{2}\gamma_{\min})$, we can then further bound it as follows:

$$\begin{aligned}\|G_{\mathrm{gd}}(\theta_1) - G_{\mathrm{gd}}(\theta_2)\|_2^2 &\leq \left(1 - \frac{2\eta\beta_{\max}\gamma_{\min}}{\beta_{\max}+\gamma_{\min}} + 2\eta\epsilon\tilde{\beta} + 2\eta^2\epsilon\tilde{\beta}\beta_{\max} + \eta^2\epsilon^2\beta_{\max}^2\right)\|\theta_1-\theta_2\|_2^2 \\ &\leq \left(1 - \frac{2\eta\tilde{\beta}\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}} + 2\eta\epsilon\tilde{\beta} + 2\eta^2\epsilon\tilde{\beta}^2 + \eta^2\epsilon^2\tilde{\beta}^2\right)\|\theta_1-\theta_2\|_2^2 \ .\end{aligned}$$

To ensure the contraction, we need $2\eta\tilde{\beta}\gamma_{\min}/(\tilde{\beta}+\sqrt{2}\gamma_{\min}) - 2\eta\epsilon\tilde{\beta} - 2\eta^2\epsilon\tilde{\beta}^2 - \eta^2\epsilon^2\tilde{\beta}^2 > 0$. When $\epsilon \leq 1$, we have

$$\begin{aligned}&\frac{2\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}} - 2\epsilon - 2\eta\epsilon\tilde{\beta} - \eta\epsilon^2\tilde{\beta} \geq \frac{2\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}} - 2\epsilon - 3\eta\epsilon\tilde{\beta} > 0 \ , \\ &\therefore \epsilon < \frac{2\gamma_{\min}}{(\tilde{\beta}+\sqrt{2}\gamma_{\min})(3\eta\tilde{\beta}+2)} \leq 1 \ .\end{aligned}$$

Therefore, with a short calculation we have

$$\|G_{\mathrm{gd}}(\theta_1) - G_{\mathrm{gd}}(\theta_2)\|_2 \leq \sqrt{1 - \eta\left(\frac{2\tilde{\beta}\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}} - \epsilon\tilde{\beta}(2+2\eta\tilde{\beta}+\eta\epsilon\tilde{\beta})\right)}\|\theta_1-\theta_2\|_2 \ .$$

For $x \in [0,1]$, we have $\sqrt{1-x} \leq 1 - \frac{x}{2}$, thus we can further simplify it as

$$\|G_{\mathrm{gd}}(\theta_1) - G_{\mathrm{gd}}(\theta_2)\|_2 \leq \left(1 - \eta\left(\frac{\tilde{\beta}\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}} - \epsilon\tilde{\beta}(1+\eta\tilde{\beta}+0.5\eta\epsilon\tilde{\beta})\right)\right)\|\theta_1-\theta_2\|_2 \ . \quad (22)$$

According to Banach fixed point theorem (Lemma 20), we can prove that when $\epsilon < 2\gamma_{\min}/[(\tilde{\beta}+\sqrt{2}\gamma_{\min})(3\eta\tilde{\beta}+2)] \leq 1$, there exists a unique fixed point, i.e., there exists a unique decision-dependent stable point and two models will converge to this point. Recursively apply Eqn. (22), we can get the convergence rate

$$\begin{aligned}\|\theta_T - \theta_{\mathrm{DS}}\|_2 &\leq \left[1 - \eta\left(\frac{\tilde{\beta}\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}} - \epsilon\tilde{\beta}(1+\eta\tilde{\beta}+0.5\eta\epsilon\tilde{\beta})\right)\right]^T \|\theta_0 - \theta_{\mathrm{DS}}\|_2 \\ &\leq \exp\left[-T\eta\left(\frac{\tilde{\beta}\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}} - \epsilon\tilde{\beta}(1+\eta\tilde{\beta}+0.5\eta\epsilon\tilde{\beta})\right)\right]\|\theta_0 - \theta_{\mathrm{DS}}\|_2 \ ,\end{aligned}$$

which completes the proof. □

### B.6. Proof of Theorem 13

Here we follow the proof of Theorem 9, which considers two cases: $\|\theta_t - \theta_{\mathrm{DS}}\|_2 > r$ and $\|\theta_t - \theta_{\mathrm{DS}}\|_2 \leq r$. Let $\mathcal{D}^{n_t}(\theta)$ denotes the empirical distribution consists of training sample $S_{n_t}$ drawn i.i.d. from distribution $\mathcal{D}(\theta)$, then the update function for two models in REGD can be written as

$$G_{\mathrm{gd},i}^{n_t}(\theta) = \Pi_\Theta\left(\theta_t^i - \eta\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta)}\nabla_\theta\ell_i(\boldsymbol{z};\theta^i)\right) \quad \text{for} \quad i \in [2].$$

**For $\|\theta_t - \theta_{\mathbf{DS}}\|_2 > r$:**
When $n_t \geq \log(\frac{c_1}{\delta})/(c_2(\epsilon r)^{2m})$, with the probability at least $\sqrt{\pi^2 t^2 - 6\delta}/(\pi t)$ we have

$$\begin{aligned} W_1(D^{n_t}(\theta_t), D(\theta_{\mathrm{DS}})) &\leq \epsilon r + \epsilon\|\theta_t - \theta_{\mathrm{DS}}\|_2 \\ &\leq 2\epsilon\|\theta_t - \theta_{\mathrm{DS}}\|_2 \ . \end{aligned}$$

Firstly, we consider model $\hat{1}$,

$$\begin{aligned} &\|G^{n_t}_{\mathrm{gd},\hat{1}}(\theta_t) - G_{\mathrm{gd},\hat{1}}(\theta_{\mathrm{DS}})\|_2^2 \\ &= \left\|\theta_t^{\hat{1}} - \eta\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta_t^{\hat{1}}) \ -\theta_{\mathrm{DS}}^{\hat{1}} + \eta\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta_{\mathrm{DS}}^{\hat{1}})\right\|_2^2 \\ &= \|\theta_t^{\hat{1}} - \theta_{\mathrm{DS}}^{\hat{1}}\|_2^2 - 2\eta(\theta_t^{\hat{1}} - \theta_{\mathrm{DS}}^{\hat{1}})^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta_t^{\hat{1}}) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta_{\mathrm{DS}}^{\hat{1}})\right) \\ &\quad + \eta^2\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta_t^{\hat{1}}) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta_{\mathrm{DS}}^{\hat{1}})\right\|_2^2 \ . \end{aligned}$$

The equation above would also holds for model $\hat{2}$. Adding them together, we have

$$\begin{aligned} &\|G^{n_t}_{\mathrm{gd}}(\theta_t) - G_{\mathrm{gd}}(\theta_{\mathrm{DS}})\|_2^2 \\ &= \|G^{n_t}_{\mathrm{gd},\hat{1}}(\theta_t) - G_{\mathrm{gd},\hat{1}}(\theta_{\mathrm{DS}})\|_2^2 + \|G^{n_t}_{\mathrm{gd},\hat{2}}(\theta_t) - G_{\mathrm{gd},\hat{2}}(\theta_{\mathrm{DS}})\|_2^2 \\ &= \|\theta_t - \theta_{\mathrm{DS}}\|_2^2 - 2\eta(\theta_t - \theta_{\mathrm{DS}})^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right) \\ &\quad + \eta^2\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right\|_2^2 \ . \\ &= T_1 - 2\eta T_2 + \eta^2 T_3 \ . \end{aligned}$$

where $\nabla_\theta \ell(\boldsymbol{z};\theta) = (\nabla_\theta \ell_{\hat{1}}(\boldsymbol{z};\theta^{\hat{1}}), \nabla_\theta \ell_{\hat{2}}(\boldsymbol{z};\theta^{\hat{2}}))$ is the concatenation of two gradient vectors, and $T_1, T_2, T_3$ is defined as below

$$\begin{aligned} T_1 &= \|\theta_t - \theta_{\mathrm{DS}}\|_2^2 \ , \\ T_2 &= (\theta_t - \theta_{\mathrm{DS}})^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) \ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right) \ , \\ T_3 &= \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) \ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right\|_2^2 \ . \end{aligned}$$

Next, we individually bound each of the three terms. For $T_2$, we have

$$\begin{aligned} T_2 =&(\theta_t - \theta_{\mathrm{DS}})^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right. \\ &\left. + \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t) \ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right) \\ =&(\theta_t - \theta_{\mathrm{DS}})^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) \ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right) \\ &+ (\theta_t - \theta_{\mathrm{DS}})^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t) \ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right) \ . \end{aligned}$$

By Cauchy–Schwarz inequality,

$$\begin{aligned} T_2 \geq& - \|\theta_t - \theta_{\mathrm{DS}}\|_2\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) \ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right\|_2 \\ &+ (\theta_t - \theta_{\mathrm{DS}})^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t) \ -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right) \ . \end{aligned}$$

Since loss $\ell\hat{1}, \ell\hat{2}$ are $\beta_{\hat{1}}$-smooth and $\beta_{\hat{2}}$-smooth respectively, $\nabla_\theta \ell(\boldsymbol{z};\theta)$ is $\tilde{\beta}$-Lipschitz continuous in $\boldsymbol{z}$ with $\tilde{\beta} = (\beta_{\hat{1}}^2 + \beta_{\hat{2}}^2)^{1/2}$. By Lemma 19, we have

$$\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right\|_2 \leq 2\tilde{\beta}\epsilon\|\theta_t - \theta_{\mathrm{DS}}\|_2 \ ,$$

which follows that

$$T_2 \geq -2\tilde{\beta}\epsilon\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 + (\theta_t - \theta_{\mathrm{DS}})^{\mathrm{T}} \left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t) \; - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right) \; .$$

Because $\ell_{\hat{1}}, \ell_{\hat{2}}$ are both $\beta_{\max}$-smooth and $\gamma_{\min}$-strongly convex, with Lemma 18 we have

$$\begin{aligned} T_2 \geq & -2\tilde{\beta}\epsilon\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 + \frac{\beta_{\max}\gamma_{\min}}{\beta_{\max}+\gamma_{\min}}\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 \\ & + \frac{1}{\beta_{\max}+\gamma_{\min}}\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\left[\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right]\right\|_2^2 \; . \end{aligned}$$

As for $T_3$, we have

$$\begin{aligned} T_3 = & \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right. \\ & \left. +\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right\|_2^2 \\ = & \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right\|_2^2 \\ & + \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\left[\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right]\right\|_2^2 \\ & + 2\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right)^{\mathrm{T}} \\ & \cdot \left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right) . \end{aligned}$$

For the first term, we can similarly get

$$\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right\|_2^2 \leq 4\tilde{\beta}^2\epsilon^2\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 \; .$$

For the third term, we can define unit vector $v$ as shown below

$$v = \frac{\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\left[\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right]}{\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\left[\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right]\right\|_2} \; .$$

A simple calculation reveals that

$$\begin{aligned} & 2\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right)^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right) \\ & = 2\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\left[\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right]\right\|_2 \left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right)^{\mathrm{T}} v \\ & = 2\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\left[\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \nabla_\theta \ell(\boldsymbol{z};\theta_2)\right]\right\|_2\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)} v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})} v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right) . \end{aligned}$$

By the smoothness of $\ell_{\hat{1}}$ and $\ell_{\hat{2}}$, $v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta)$ is $\tilde{\beta}$-Lipschitz continuous. With Lemma 19, we have

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)} v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})} v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_t) \\ & = \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)} v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})} v^{\mathrm{T}}\nabla_\theta \ell(\boldsymbol{z};\theta_t)\right\|_2 \\ & \leq 2\tilde{\beta}\epsilon\|\theta_t - \theta_{\mathrm{DS}}\|_2 \; . \end{aligned}$$

With the definition of $\beta$-smoothness, we have

$$\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right\|_2 \leq \beta_{\max}\|\theta_t - \theta_{\mathrm{DS}}\|_2 \; ,$$

where $\beta_{\max} = \max\{\beta_{\hat{1}}, \beta_{\hat{2}}\}$. Hence, we can conclude that

$$
\begin{aligned}
&2\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta\ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};\theta_t)\right)^{\mathrm{T}}\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right)\\
&= 2\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\left[\nabla_\theta\ell(\boldsymbol{z};\theta_t) - \nabla_\theta\ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right]\right\|_2\left(\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}^{n_t}(\theta_t)}\nabla_\theta\ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};\theta_t)\right)^{\mathrm{T}} v\\
&\le 4\tilde{\beta}\beta_{\max}\epsilon\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 \ .
\end{aligned}
$$

This follows that

$$
T_3 \le 4\tilde{\beta}^2\epsilon^2\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 + 4\tilde{\beta}\beta_{\max}\epsilon\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 + \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right\|_2^2 .
$$

Therefore, we can finally prove that, if the step size $\eta$ small enough, i.e. $\eta \le 2/(\beta_{\max} + \gamma_{\min}) \le 2\sqrt{2}/(\tilde{\beta} + \gamma_{\min}\sqrt{2})$. With probability at least $(\frac{\sqrt{\pi^2t^2-6\delta}}{\pi t})^2 = 1 - 6\delta/(\pi^2t^2)$, we have

$$
\begin{aligned}
&\|G_{\mathrm{gd}}^{n_t}(\theta_t) - G_{\mathrm{gd}}(\theta_{\mathrm{DS}})\|_2^2\\
&\le \left(1 - \frac{2\eta\beta_{\max}\gamma_{\min}}{\beta_{\max} + \gamma_{\min}} + 4\eta\epsilon\tilde{\beta} + 4\eta^2\epsilon^2\tilde{\beta}^2 + 4\eta^2\epsilon\tilde{\beta}\beta_{\max}\right)\|\theta_t - \theta_{\mathrm{DS}}\|_2^2\\
&\quad - \left(\frac{2\eta}{\beta_{\max} + \gamma_{\min}} - \eta^2\right)\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta\ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right\|_2^2\\
&\le \left(1 - \frac{2\eta\beta_{\max}\gamma_{\min}}{\beta_{\max} + \gamma_{\min}} + 4\eta\epsilon\tilde{\beta} + 4\eta^2\epsilon^2\tilde{\beta}^2 + 4\eta^2\epsilon\tilde{\beta}\beta_{\max}\right)\|\theta_t - \theta_{\mathrm{DS}}\|_2^2\\
&\le \left(1 - \frac{2\eta\tilde{\beta}\gamma_{\min}}{\tilde{\beta} + \sqrt{2}\gamma_{\min}} + 4\eta\epsilon\tilde{\beta} + 4\eta^2\epsilon\tilde{\beta}^2 + 4\eta^2\epsilon^2\tilde{\beta}^2\right)\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 \ .
\end{aligned}
$$

When $\epsilon \le 1$, to ensure the contraction, we have

$$
\begin{aligned}
&\frac{2\gamma_{\min}}{\tilde{\beta} + \sqrt{2}\gamma_{\min}} - 4\epsilon - 4\eta\epsilon\tilde{\beta} - 4\eta\epsilon^2\tilde{\beta} \ge \frac{\gamma_{\min}}{\tilde{\beta} + \sqrt{2}\gamma_{\min}} - 2\epsilon - 4\eta\epsilon\tilde{\beta} > 0,\\
&\therefore \epsilon < \frac{\gamma_{\min}}{(\tilde{\beta} + \sqrt{2}\gamma_{\min})(4\eta\tilde{\beta} + 2)} \le 1 \ .
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
\|G_{\mathrm{gd}}^{n_t}(\theta_t) - G_{\mathrm{gd}}(\theta_{\mathrm{DS}})\|_2 &= \|\theta_{t+1} - \theta_{\mathrm{DS}}\|_2\\
&\le \sqrt{1 - 2\eta\left(\frac{\tilde{\beta}\gamma_{\min}}{\tilde{\beta} + \sqrt{2}\gamma_{\min}} - 2\epsilon(\tilde{\beta} + \eta\tilde{\beta}^2 + \eta\epsilon\tilde{\beta}^2)\right)}\|\theta_t - \theta_{\mathrm{DS}}\|_2\\
&\le \left(1 - \eta\left(\frac{\tilde{\beta}\gamma_{\min}}{\tilde{\beta} + \sqrt{2}\gamma_{\min}} - 2\epsilon(\tilde{\beta} + \eta\tilde{\beta}^2 + \eta\epsilon\tilde{\beta}^2)\right)\right)\|\theta_t - \theta_{\mathrm{DS}}\|_2 \ .
\end{aligned}
$$

**For $\|\theta_t - \boldsymbol{\theta_{\mathrm{DS}}}\|_2 \le r$:**
Similarly, by triangle inequality, we have with the probability at least $\sqrt{\pi^2t^2 - 6\delta}/(\pi t)$ that,

$$
\begin{aligned}
W_1(D_{\mathrm{gd}}^{n_t}(\theta_t), D_{\mathrm{gd}}(\theta_{\mathrm{DS}})) &\le W_1(D_{\mathrm{gd}}^{n_t}(\theta_t), D_{\mathrm{gd}}(\theta_t)) + W_1(D_{\mathrm{gd}}(\theta_t), D_{\mathrm{gd}}(\theta_{\mathrm{DS}}))\\
&\le W_1(D_{\mathrm{gd}}^{n_t}(\theta_t), D_{\mathrm{gd}}(\theta_t)) + \epsilon r\\
&\le 2\epsilon r \ .
\end{aligned}
$$

With similarly procedures in the proof above, we have

$$\begin{aligned}
T_1 &= \|\theta_t - \theta_{\mathrm{DS}}\|_2^2 \ , \\
T_2 &\geq - 2\tilde{\beta}\epsilon r \|\theta_t - \theta_{\mathrm{DS}}\|_2 + \frac{\beta_{\max}\gamma_{\min}}{\beta_{\max}+\gamma_{\min}}\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 \\
&\quad + \frac{1}{\beta_{\max}+\gamma_{\min}}\left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t) \ - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right\|_2^2 , \\
T_3 &\leq 4\tilde{\beta}^2\epsilon^2 r^2 + 4\tilde{\beta}\beta_{\max}\epsilon r\|\theta_t - \theta_{\mathrm{DS}}\|_2 + \left\|\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_t) - \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}(\theta_{\mathrm{DS}})}\nabla_\theta \ell(\boldsymbol{z};\theta_{\mathrm{DS}})\right\|_2^2 .
\end{aligned}$$

With probability at least $1 - 6\delta/(\pi^2 t^2)$, for $r < 1$ we have

$$\begin{aligned}
&\|G_{\mathrm{gd}}^{n_t}(\theta_t) - G_{\mathrm{gd}}(\theta_{\mathrm{DS}})\|_2^2 \\
&\leq (1 - \frac{2\eta\beta_{\max}\gamma_{\min}}{\beta_{\max}+\gamma_{\min}})\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 + 4\eta^2\epsilon^2\tilde{\beta}^2 r^2 + (4\eta\tilde{\beta}\epsilon r + 4\eta^2\tilde{\beta}\beta_{\max}\epsilon r)\|\theta_t - \theta_{\mathrm{DS}}\|_2 \\
&\leq (1 - \frac{2\eta\tilde{\beta}\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}})\|\theta_t - \theta_{\mathrm{DS}}\|_2^2 + 4\eta^2\epsilon^2\tilde{\beta}^2 r^2 + (4\eta\tilde{\beta}\epsilon r + 4\eta^2\tilde{\beta}^2\epsilon r)\|\theta_t - \theta_{\mathrm{DS}}\|_2 \\
&\leq (1 - \frac{2\eta\tilde{\beta}\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}})r^2 + 4\eta^2\epsilon^2\tilde{\beta}^2 r^2 + (4\eta\tilde{\beta}\epsilon r + 4\eta^2\tilde{\beta}^2\epsilon r)r \\
&\leq \left(1 - \frac{2\eta\tilde{\beta}\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}} + 4\eta^2\epsilon^2\tilde{\beta}^2 + 4\eta\tilde{\beta}\epsilon + 4\eta^2\tilde{\beta}^2\epsilon\right) r^2.
\end{aligned}$$

Similarly, to ensure a contraction, we further have

$$\begin{aligned}
\|G_{\mathrm{gd}}^{n_t}(\theta_t) - G_{\mathrm{gd}}(\theta_{\mathrm{DS}})\|_2 &= \|\theta_{t+1} - \theta_{\mathrm{DS}}\|_2 \\
&\leq \sqrt{1 - 2\eta\left(\frac{\tilde{\beta}\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}} - 2\epsilon(\tilde{\beta} + \eta\tilde{\beta}^2 + \eta\epsilon\tilde{\beta}^2)\right)}\, r \\
&\leq \left(1 - \eta\left(\frac{\tilde{\beta}\gamma_{\min}}{\tilde{\beta}+\sqrt{2}\gamma_{\min}} - 2\epsilon(\tilde{\beta} + \eta\tilde{\beta}^2 + \eta\epsilon\tilde{\beta}^2)\right)\right) r \ ,
\end{aligned}$$

which completes the proof. □

### B.7. Proof of Theorem 15

For model $\hat{1}$, because $\Theta$ is convex and closed, projection can only bring iterates closer to the stable point, i.e.

$$\|\theta_{t+1}^{\hat{1}} - \theta_{\mathrm{DS}}^{\hat{1}}\|_2^2 = \|\Pi_\Theta(\theta_t^{\hat{1}} - \eta_t \nabla \ell_{\hat{1}}(\boldsymbol{z}_{\hat{1}}^{(t)};\theta_t^{\hat{1}})) - \theta_{\mathrm{DS}}^{\hat{1}}\|_2^2 \leq \|\theta_t^{\hat{1}} - \eta_t \nabla \ell_{\hat{1}}(\boldsymbol{z}_{\hat{1}}^{(t)};\theta_t^{\hat{1}}) - \theta_{\mathrm{DS}}^{\hat{1}}\|_2^2 \ .$$

Splitting the expectation of the square into three terms, we have

$$\begin{aligned}
\mathbb{E}\left[\|\theta_t^{\hat{1}} - \eta_t \nabla \ell_{\hat{1}}(\boldsymbol{z}_{\hat{1}}^{(t)};\theta_t^{\hat{1}}) - \theta_{\mathrm{DS}}^{\hat{1}}\|_2^2\right] &= \mathbb{E}\left[\|\theta_t^{\hat{1}} - \theta_{\mathrm{DS}}^{\hat{1}}\|_2^2\right] - 2\eta_t \mathbb{E}\left[\nabla \ell_{\hat{1}}(\boldsymbol{z}_{\hat{1}}^{(t)};\theta_t^{\hat{1}})^{\mathrm{T}}(\theta_t^{\hat{1}} - \theta_{\mathrm{DS}}^{\hat{1}})\right] \\
&\quad + \eta_t^2 \mathbb{E}\left[\|\nabla \ell_{\hat{1}}(\boldsymbol{z}_{\hat{1}}^{(t)};\theta_t^{\hat{1}})\|_2^2\right] .
\end{aligned}$$

Hence, for two models, we have:

$$\begin{aligned}\mathbb{E}\left[\|\theta_{t+1}-\theta_{\mathrm{DS}}\|_2^2\right] &= \mathbb{E}\left[\|\theta_{t+1}^{\hat{1}}-\theta_{\mathrm{DS}}^{\hat{1}}\|_2^2\right]+\mathbb{E}\left[\|\theta_{t+1}^{\hat{2}}-\theta_{\mathrm{DS}}^{\hat{2}}\|_2^2\right]\\ &\le \mathbb{E}\left[\|\theta_t-\theta_{\mathrm{DS}}\|_2^2\right]-2\eta_t\mathbb{E}\left[\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})\right]+\eta_t^2\mathbb{E}\left[\|\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)\|_2^2\right]\\ &= T_1-2\eta_t T_2+\eta_t^2 T_3\ ,\end{aligned}$$

where $\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)$ denotes $(\nabla\ell_{\hat{1}}(\boldsymbol{z}_{\hat{1}}^{(t)};\theta_t);\nabla\ell_{\hat{2}}(\boldsymbol{z}_{\hat{2}}^{(t)};\theta_t))$, the concatenation of two gradient vector for models. Then, we can bound each term respectively. To begin with $T_2$, since $\theta_{\mathrm{DS}}$ is the optimal parameter for two models, we have $\mathbb{E}\left[\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_{\mathrm{DS}})^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})\right]\ge 0$ with first-order optimality condition. We can bound $T_2$ as:

$$\begin{aligned}T_2 &\ge \mathbb{E}\left[\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})\right]-\mathbb{E}\left[\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_{\mathrm{DS}})^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})\right]\\ &= \mathbb{E}\left[(\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)-\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_t)+\nabla\ell(\boldsymbol{z}^{(\theta_t)};\theta_{\mathrm{DS}})-\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_{\mathrm{DS}}))^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})\right]\\ &= \mathbb{E}\left[(\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)-\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_t))^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})\right]\\ &+\mathbb{E}\left[(\nabla\ell(\boldsymbol{z}^{(\theta_t)};\theta_{\mathrm{DS}})-\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_{\mathrm{DS}}))^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})\right].\end{aligned}\tag{23}$$

For the first term in Eqn. (23), by applying law of total expectation, we have

$$\begin{aligned}&\mathbb{E}\left[(\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)-\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_t))^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})\right]\\ &=\mathbb{E}\left[\mathbb{E}\left[(\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)-\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_t))^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})|\theta_t\right]\right],\end{aligned}$$

Note that when it is conditional on $\theta_t$, with Cauchy-Schwarz inequality and $\tilde{\beta}$-jointly smoothness of $\ell$, $\nabla\ell(\boldsymbol{z};\theta_t)^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})$ is $\tilde{\beta}\|\theta_t-\theta_{\mathrm{DS}}\|_2$-Lipschitz continuous in $\boldsymbol{z}$. Therefore, by applying Lemma 19 with $\epsilon$-Lipschitz continuity of distribution $\mathcal{D}$, we have

$$\begin{aligned}&\mathbb{E}\left[(\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)-\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_t))^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})\right]\\ &=\mathbb{E}\left[\mathbb{E}\left[(\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)-\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_t))^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})|\theta_t\right]\right]\\ &\ge -\epsilon\tilde{\beta}\mathbb{E}\left[\|\theta_t-\theta_{\mathrm{DS}}\|_2^2\right].\end{aligned}$$

For the second term in Eqn. (23), we use the $\gamma_{\min}$-strongly convexity of $\ell$ and get:

$$\begin{aligned}&\mathbb{E}\left[(\nabla\ell(\boldsymbol{z}^{(\theta_t)};\theta_{\mathrm{DS}})-\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_{\mathrm{DS}}))^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})\right]\\ &=\mathbb{E}\left[\mathbb{E}\left[(\nabla\ell(\boldsymbol{z}^{(\theta_t)};\theta_{\mathrm{DS}})-\nabla\ell(\boldsymbol{z}^{(\theta_{\mathrm{DS}})};\theta_{\mathrm{DS}}))^{\mathrm{T}}(\theta_t-\theta_{\mathrm{DS}})|\theta_t\right]\right]\\ &\ge \gamma_{\min}\mathbb{E}\left[\|\theta_t-\theta_{\mathrm{DS}}\|_2^2\right].\end{aligned}$$

Adding the two inequality together, we can finally bound $T_2$ as:

$$T_2\ge(\gamma_{\min}-\epsilon\tilde{\beta})\mathbb{E}\left[\|\theta_t-\theta_{\mathrm{DS}}\|_2^2\right].$$

Now we consider $T_3$. With the assumptions on the variance on the expected norm of gradients, we have:

$$\begin{aligned}\mathbb{E}\left[\|\nabla\ell(\boldsymbol{z}^{(t)};\theta_t)\|_2^2\right] &\leq \tilde{\sigma}^2 + L_{\max}^2\mathbb{E}\left[\|\theta_t - G(\theta_t)\|_2^2\right]\\ &= \tilde{\sigma}^2 + L_{\max}^2\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{DS}} + \theta_{\mathrm{DS}} - G(\theta_t)\|_2^2\right]\\ &\leq \tilde{\sigma}^2 + L_{\max}^2\left(\mathbb{E}\left[(\|\theta_t - \theta_{\mathrm{DS}}\|_2 + \|\theta_{\mathrm{DS}} - G(\theta_t)\|_2)^2\right]\right)\\ &\leq \tilde{\sigma}^2 + L_{\max}^2\left(1 + \epsilon\frac{\tilde{\beta}}{\gamma_{\min}}\right)^2\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{DS}}\|_2^2\right],\end{aligned}$$

where $\tilde{\sigma} = \sqrt{\sigma_{\hat{1}}^2 + \sigma_{\hat{2}}^2}$, $L_{\max} = \max\{L_{\hat{1}}, L_{\hat{2}}\}$, and the last step follows the first statement in the proof B.3. Adding all the three terms together, we can get that

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{DS}}\|_2^2\right] \leq \left(1 - 2\eta_t(\gamma_{\min} - \epsilon\tilde{\beta}) + \eta_t^2 L_{\max}^2\left(1 + \epsilon\frac{\tilde{\beta}}{\gamma_{\min}}\right)^2\right)\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{DS}}\|_2^2\right] + \eta_t^2\tilde{\sigma}^2.$$

Since $\epsilon < \gamma_{\min}/\tilde{\beta}$, we can further get

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{DS}}\|_2^2\right] \leq \left(1 - 2\eta_t(\gamma_{\min} - \epsilon\tilde{\beta}) + 4\eta_t^2 L_{\max}^2\right)\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{DS}}\|_2^2\right] + \eta_t^2\tilde{\sigma}^2 .$$

We now can give the rest of the proof by induction on $t$. Let $t_0 = 8L_{\max}^2/(\gamma_{\min} - \epsilon\tilde{\beta})^2$, and $\eta_t = 1/(\gamma_{\min} - \epsilon\tilde{\beta})(t + t_0)$. Assume that for $t \in \mathbb{N}$, it holds that

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{DS}}\|_2^2\right] \leq \frac{\max\{2\tilde{\sigma}^2, 8L_{\max}^2\|\theta_1 - \theta_{\mathrm{DS}}\|_2^2\}}{(\gamma_{\min} - \epsilon\tilde{\beta})^2(t + t_0)} ,$$

then it follows that

$$\begin{aligned}&\mathbb{E}\left[\|\theta_{t+1} - \theta_{\mathrm{DS}}\|_2^2\right]\\ &\leq \left(1 - 2\eta_t(\gamma_{\min} - \epsilon\tilde{\beta}) + 4\eta_t^2 L_{\max}^2\right)\mathbb{E}\left[\|\theta_t - \theta_{\mathrm{DS}}\|_2^2\right] + \eta_t^2\tilde{\sigma}^2\\ &\leq \frac{1}{(\gamma_{\min} - \epsilon\tilde{\beta})^2}\left(\frac{t + t_0 - 2 + \frac{4L_{\max}^2}{(\gamma_{\min} - \epsilon\tilde{\beta})^2 t_0}}{(t + t_0)^2}\max\left\{2\tilde{\sigma}^2, 8L_{\max}^2\|\theta_1 - \theta_{\mathrm{DS}}\|_2^2\right\} + \frac{\tilde{\sigma}^2}{(t + t_0)^2}\right)\\ &\leq \frac{1}{(\gamma_{\min} - \epsilon\tilde{\beta})^2}\left(\frac{t + t_0 - 1.5}{(t + t_0)^2}\max\left\{2\tilde{\sigma}^2, 8L_{\max}^2\|\theta_1 - \theta_{\mathrm{DS}}\|_2^2\right\} + \frac{\tilde{\sigma}^2}{(t + t_0)^2}\right)\\ &\leq \frac{1}{(\gamma_{\min} - \epsilon\tilde{\beta})^2}\left(\frac{t + t_0 - 1}{(t + t_0)^2}\max\left\{2\tilde{\sigma}^2, 8L_{\max}^2\|\theta_1 - \theta_{\mathrm{DS}}\|_2^2\right\} - \frac{0.5 * 2\tilde{\sigma}^2 - \tilde{\sigma}^2}{(t + t_0)^2}\right)\\ &= \frac{1}{(\gamma_{\min} - \epsilon\tilde{\beta})^2}\frac{t + t_0 - 1}{(t + t_0)^2}\max\left\{2\tilde{\sigma}^2, 8L_{\max}^2\|\theta_1 - \theta_{\mathrm{DS}}\|_2^2\right\}\\ &\leq \frac{1}{(\gamma_{\min} - \epsilon\tilde{\beta})^2}\frac{1}{1 + t + t_0}\max\left\{2\tilde{\sigma}^2, 8L_{\max}^2\|\theta_1 - \theta_{\mathrm{DS}}\|_2^2\right\} ,\end{aligned}$$

where the last step follows $(t + t_0)^2 > (t + t_0)^2 - 1 = (t + t_0 + 1)(t + t_0 - 1)$. This completes the proof by setting $t = T - 1$. □