Appendix

A. Details of the Proofs

We analyze the sample complexity by separating the teaching procedure into two stages in each iteration, *i.e.*, the active query stage by conducting examination for the student and the teaching stage by providing samples to the student.

A.1. Error Decomposition

Recall that there is a mapping $\mathcal{G}$ from the feature space of the teacher to that of the student, and we have $\langle w, \tilde{x}\rangle = \langle w, \mathcal{G}(x)\rangle = \left\langle \mathcal{G}^\top(w), x\right\rangle$ where $\mathcal{G}^\top$ denotes the conjugate mapping of $\mathcal{G}$. We also denote the $\sigma_{\max} = \max_{x^\top x=1} \mathcal{G}^\top(x)\mathcal{G}(x)$, $\sigma_{\min} = \min_{x^\top x=1} \mathcal{G}^\top(x)\mathcal{G}(x) > 0$ since the operator $\mathcal{G}$ is invertible, and $\kappa\left(\mathcal{G}^\top\mathcal{G}\right) = \frac{\sigma_{\max}}{\sigma_{\min}}$. To involve the inconsistency between the student's parameters w^t, and the teacher's estimator v^t, at t-th iteration into the analysis, we first provide the recursion with error decomposition. For simplicity, we denote $\beta(\langle w, x\rangle, y) := \nabla_{\langle w,x\rangle}\ell\left(\langle w, x\rangle, y\right)$. Then, we have the update rule of student as

$$w^{t+1} = w^t - \eta\beta\big(\left\langle w^t, \mathcal{G}(x^t)\right\rangle, y^t\big)\mathcal{G}(x^t),$$

where $x^t = \gamma\left(v^t - v^*\right)$ is constructed by teacher with the estimator v^t. Plug into the difference, we have

$$\begin{aligned}
&\left\|\mathcal{G}^\top(w^{t+1}) - v^*\right\|^2 \\
&= \left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 + \eta^2\beta^2\left(\left\langle w^t, \mathcal{G}(x^t)\right\rangle, y^t\right)\left\|\mathcal{G}^\top\mathcal{G}(x^t)\right\|^2 - 2\eta\beta\left(\left\langle w^t, \mathcal{G}(x^t)\right\rangle, y^t\right)\left\langle\mathcal{G}^\top\mathcal{G}(x^t), \mathcal{G}^\top(w^t) - v^*\right\rangle \\
&= \left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 + \eta^2\beta^2\left(\left\langle v^t, x^t\right\rangle, y^t\right)\left\|\mathcal{G}^\top\mathcal{G}(x^t)\right\|^2 - 2\eta\beta\left(\left\langle v^t, x^t\right\rangle, y^t\right)\left\langle\mathcal{G}^\top\mathcal{G}(x^t), \mathcal{G}^\top(w^t) - v^*\right\rangle \\
&\qquad + \eta^2\left\|\mathcal{G}^\top\mathcal{G}(x^t)\right\|^2\left(\beta^2\left(\left\langle\mathcal{G}^\top(w^t), x^t\right\rangle, y^t\right) - \beta^2\left(\left\langle v^t, x^t\right\rangle, y^t\right)\right) \\
&\qquad - 2\eta\left\langle\mathcal{G}^\top\mathcal{G}(x^t), \mathcal{G}^\top(w^t) - v^*\right\rangle\left(\beta\left(\left\langle\mathcal{G}^\top(w^t), x^t\right\rangle, y^t\right) - \beta\left(\left\langle v^t, x^t\right\rangle, y^t\right)\right).
\end{aligned}$$

Suppose the loss function is L-Lipschitz smooth and $x \in \mathcal{X} = \left\{x \in \mathbb{R}^d, \|x\| \le R\right\}$,

$$\left|\beta\left(\langle v_1, x\rangle, y\right) - \beta\left(\langle v_2, x\rangle, y\right)\right| \le LR\left\|v_1 - v_2\right\|,$$

which implies

$$\beta\left(\langle v_2, x\rangle, y\right) - LR\left\|v_1 - v_2\right\| \le \beta\left(\langle v_1, x\rangle, y\right) \le \beta\left(\langle v_2, x\rangle, y\right) + LR\left\|v_1 - v_2\right\|.$$

We have the error decomposition as follows,

$$\begin{aligned}
\left\|\mathcal{G}^\top(w^{t+1}) - v^*\right\|^2 \le{}& \left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 + \eta^2\beta^2\left(\left\langle v^t, \gamma(v^t - v^*)\right\rangle, y^t\right)\gamma^2\left\|\mathcal{G}^\top\mathcal{G}(v^t - v^*)\right\|^2 \\
&- 2\eta\beta\left(\left\langle v^t, \gamma(v^t - v^*)\right\rangle, y^t\right)\gamma\left\langle\mathcal{G}^\top\mathcal{G}(v^t - v^*), \mathcal{G}^\top(w^t) - v^*\right\rangle \\
&+ \eta^2\gamma^2 LR\left\|\mathcal{G}^\top\mathcal{G}(v^t - v^*)\right\|^2\left\|\mathcal{G}^\top(w^t) - v^t\right\|\left(\beta\left(\left\langle\mathcal{G}^\top(w^t), x^t\right\rangle, y^t\right) + \beta\left(\left\langle v^t, x^t\right\rangle, y^t\right)\right) \\
&+ 2\eta\gamma LR\left\langle\mathcal{G}^\top\mathcal{G}(v^t - v^*), \mathcal{G}^\top(w^t) - v^*\right\rangle\left\|\mathcal{G}^\top(w^t) - v^t\right\| \\
\le{}& \left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 + \eta^2\gamma^2\sigma_{\max}^2\beta^2\left(\left\langle v^t, \gamma(v^t - v^*), y^t\right\rangle\right)\left\|v^t - v^*\right\|^2 \\
&- 2\eta\beta\left(\left\langle v^t, \gamma(v^t - v^*)\right\rangle, y^t\right)\gamma\left(\sigma_{\min}\left\|v^t - v^*\right\|^2 - \sigma_{\max}\left\|\mathcal{G}^\top(w^t) - v^t\right\|\left\|v^t - v^*\right\|\right) \\
&+ \eta^2\gamma^2 LR\left\|\mathcal{G}^\top\mathcal{G}(v^t - v^*)\right\|^2\left\|\mathcal{G}^\top(w^t) - v^t\right\|\left(2\beta\left(\left\langle v^t, x^t\right\rangle, y^t\right) + LR\left\|\mathcal{G}^\top(w^t) - v^t\right\|\right) \\
&+ 2\eta\gamma LR\left(\left\|\mathcal{G}^\top\mathcal{G}\right\|\left\|v^t - v^*\right\|^2 + \left\|\mathcal{G}^\top\mathcal{G}\right\|\left\|v^t - v^*\right\|\left\|\mathcal{G}^\top(w^t) - v^t\right\|\right)\left\|\mathcal{G}^\top(w^t) - v^t\right\|
\end{aligned} \tag{5}$$

where the last two terms represent the inconsistency on the teacher's side and the student's side in computing β.

A.2. Exact Recovery of $\mathcal{G}^\top(w)$

Theorem 2 *Suppose the teacher is able to recover $\mathcal{G}^\top(w^t)$ exactly using m samples at each iteration. If for any $v \in \mathbb{R}^d$, there exists $\gamma \neq 0$ and $\hat{y}$ such that $\hat{x} = \gamma\left(v - v^*\right)$ and*

$$0 < \gamma\nabla_{\langle v^t, \hat{x}\rangle}\ell\left(\left\langle v^t, \hat{x}\right\rangle, \hat{y}\right) < \frac{2\sigma_{\min}}{\eta\sigma_{\max}^2},$$

then $(\ell, \mathcal{G})$ *is ET with* $\mathcal{O}\left((m+1)\log\frac{1}{\epsilon}\right)$ *samples.*

Proof Plug $\left\|\mathcal{G}^\top(w^t) - v^t\right\| = 0$ into the error decomposition (5), we have

$$\begin{aligned}\left\|\mathcal{G}^\top(w^{t+1}) - v^*\right\|^2 &\leq \left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 + \eta^2\gamma^2\sigma_{\max}^2\beta^2\left(\left\langle v^t, \gamma(v^t - v^*), y^t\right\rangle\right)\left\|v^t - v^*\right\|^2 \\ &\quad -2\eta\beta\left(\left\langle v^t, \gamma(v^t - v^*)\right\rangle, y^t\right)\gamma\left(\sigma_{\min}\left\|v^t - v^*\right\|^2\right) \\ &\leq \left(1 + \eta^2\gamma^2\sigma_{\max}^2\beta^2\left(\left\langle v^t, \gamma(v^t - v^*), y^t\right\rangle\right) - 2\eta\beta\left(\left\langle v^t, \gamma(v^t - v^*)\right\rangle, y^t\right)\gamma\sigma_{\min}\right)\left\|\mathcal{G}^\top(w^t) - v^*\right\|^2.\end{aligned}$$

Denote $\nu(\gamma) = \min_{\hat{x}\in\mathcal{X},\hat{y}\in\mathcal{Y}}\gamma\beta\left(\left\langle v^t, \gamma(v^t - v^*)\right\rangle, y^t\right) > 0$, and $\mu(\gamma) = \max_{\hat{x}\in\mathcal{X},\hat{y}\in\mathcal{Y}}\gamma\beta\left(\left\langle v^t, \gamma(v^t - v^*)\right\rangle, y^t\right) < \frac{2\sigma_{\min}}{\eta\sigma_{\max}^2}$, we have the recursion

$$\left\|\mathcal{G}^\top(w^{t+1}) - v^*\right\|^2 \leq r(\eta,\gamma)\left\|\mathcal{G}^\top(w^t) - v^*\right\|^2,$$

where $r(\eta,\gamma) = \max\left\{1 + \eta^2\sigma_{\max}^2\mu(\gamma) - 2\eta\sigma_{\min}\mu(\gamma), 1 + \eta^2\sigma_{\max}^2\nu(\gamma) - 2\eta\sigma_{\min}\nu(\gamma)\right\}$ and $0 \leq r(\eta,\gamma) \leq 1$. Therefore, the algorithm converges exponentially,

$$\left\|\mathcal{G}^\top(w^t) - v^*\right\| \leq r(\eta,\gamma)^{t/2}\left\|\mathcal{G}^\top(w^0) - v^*\right\|.$$

In other words, the students needs $2\left(\log\frac{1}{r(\eta,\gamma)}\right)^{-1}\log\frac{\left\|\mathcal{G}^\top(w^0)-v^*\right\|}{\epsilon}$ samples for updating. Consider that at each iteration, if the teacher first uses m samples for estimating $\mathcal{G}^\top(w)$, then the total number of samples is no larger than $(m+1)\,2\left(\log\frac{1}{r(\eta,\gamma)}\right)^{-1}\log\frac{\left\|\mathcal{G}^\top(w^0)-v^*\right\|}{\epsilon}$. ■

Lemma 3 *If* $F(\cdot)$ *is bijective, then we can exactly recover* $\mathcal{G}^\top(w) \in \mathbb{R}^d$ *with* d *samples.*

Proof We prove the theorem by construction. Denote d independent samples as $Z = \{z_i\}_{i=1}^d \in \mathbb{R}^d$. We can exactly recover arbitrary v with these samples by solving the linear system,

$$\langle v, Z\rangle = b, \tag{6}$$

where $b = F^{-1}\left(F\left(\langle w, \mathcal{G}(x)\rangle\right)\right)$ are provided by the student. F^{-1} exists because F is bijective. Since $\mathrm{rank}(Z) = d$, the linear system (6) has a unique solution. ■

Lemma 4 *If* $F(\cdot) = \max(0,\cdot)$*, then we can exactly recover* $\mathcal{G}^\top(w) \in \mathbb{R}^d$ *with* $2d$ *samples.*

Proof We prove the lemma by construction. Notice that $\forall a \in \mathbb{R}$, either $\max(0,a) = a$ and $\max(0,-a) = 0$, or $\max(0,a) = 0$ and $\max(0,-a) = -a$. Then, we can first construct d independent samples as $\{z_i\}_{i=1}^d \in \mathbb{R}^d$, and then, extend the set with $\{-z_i\}_{i=1}^d$. We construct the linear system by picking one of the linear equations from $\langle v, z_i\rangle = \max(0, \langle w, \mathcal{G}(z_i)\rangle)$ or $\langle v, -z_i\rangle = \max(0, -\langle w, \mathcal{G}(z_i)\rangle)$ which does not equal to zero. Denote the linear system $\langle v, Z'\rangle = b$, since we select either z_i or $-z_i$ to form Z, then, $\mathrm{rank}(Z') = d$, therefore, the linear system has a unique solution. ■

In both regression and classification scenarios, if the student answers the questions in the query phase with $F(\cdot) = I(\cdot)$, $F(\cdot) = S(\cdot)$, or $F(\cdot) = \max(0,\cdot)$, where I denotes the identity mapping and S denotes some sigmoid function, *e.g.*, logistic function, hyperbolic tangent, error function and so on, we can exactly recover $v = \mathcal{G}^\top(w) \in \mathbb{R}^d$ with arbitrary $\mathcal{O}(d)$ independent data, omitting the numerical error and consider the solution as exact recovery. Recall we can reuse these $\mathcal{O}(d)$ independent data in each iteration, we have

Corollary 5 *Suppose the student answers questions in query phase via* $F(\cdot) = I(\cdot)$*,* $F(\cdot) = S(\cdot)$*, or* $F(\cdot) = \max(0,\cdot)$*, then* $(\ell, \mathcal{G})$ *is ET with* $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ *teaching samples and* $\mathcal{O}(d)$ *query samples via exact recovery.*

A.3. Approximate Recovery of $\mathcal{G}^\top(w)$

Theorem 6 *Suppose the loss function* ℓ *is* L*-Lipschitz smooth in a compact domain* $\Omega_v \subset \mathbb{R}^d$ *of* v *containing* v^* *and sample candidates* (x,y) *are from bounded* $\mathcal{X}\times\mathcal{Y}$*, where* $\mathcal{X} = \left\{x \in \mathbb{R}^d, \|x\| \leq R\right\}$*. Further suppose at* t*-th iteration, the teacher*

estimates the student $\epsilon_{est} := \left\|\mathcal{G}^\top(w^t) - v^t\right\| = \mathcal{O}(\epsilon)$ *with probability at least* $1-\delta$ *using* $m(\epsilon_{est}, \delta)$ *samples. If for any* $v \in \Omega_v$*, there exists* $\gamma \neq 0$ *and* $\hat{y}$ *such that for* $\hat{x} = \gamma(v - v^*)$*, we have*

$$0 < \gamma \nabla_{\langle v^t, \hat{x}\rangle} \ell\left(\left\langle v^t, \hat{x}\right\rangle, \hat{y}\right) < \frac{2(1-\lambda)\sigma_{\min}}{\eta \sigma_{\max}^2},$$
$$\text{with } 0 < \lambda < \min\left(\frac{\kappa\left(\mathcal{G}^\top \mathcal{G}\right)}{\sqrt{2}}, 1\right),$$

then the student can achieve ϵ*-approximation of* v^* *with* $\mathcal{O}\left(\log \frac{1}{\epsilon}\left(1 + m\left(\lambda\epsilon, \frac{\delta}{\log \frac{1}{\epsilon}}\right)\right)\right)$ *samples with probability at least* $1-\delta$*. If* $m(\epsilon_{est}, \delta) = \mathcal{O}(\log \frac{1}{\epsilon})$*, then* $(\ell, \mathcal{G})$ *is ET.*

Proof Assume that in each iteration, the teacher will estimate the w^t at least satisfying $\epsilon_{\text{est}} := \left\|\mathcal{G}^\top(w^t) - v^t\right\| \leq \lambda \frac{\sigma_{\min}}{\sigma_{\max}} \|v^t - v^*\|$. Plugging into the error decomposition (5), we obtain

$$\begin{aligned}
\left\|\mathcal{G}^\top(w^{t+1}) - v^*\right\|^2 \leq{}& \left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 + \eta^2\gamma^2\sigma_{\max}^2\beta^2\left(\left\langle v^t, \gamma(v^t - v^*), y^t\right\rangle\right)\left\|v^t - v^*\right\|^2 \\
&- 2\eta\beta\left(\left\langle v^t, \gamma_t(v^t - v^*)\right\rangle, y^t\right)\gamma\sigma_{\min}(1-\lambda)\left\|v^t - v^*\right\|^2 \\
&+ \eta^2 LR\epsilon_{\text{est}}\sigma_{\max}^2\gamma^2\left\|(v^t - v^*)\right\|^2\left(2\beta\left(\left\langle v^t, x^t\right\rangle, y^t\right) + LR\epsilon_{\text{est}}\right) \\
&+ 2\eta LR\epsilon_{\text{est}}\left(\gamma\sigma_{\max}\left\|v^t - v^*\right\|^2 + \gamma\sigma_{\max}\left\|\mathcal{G}^\top(w^t) - v^t\right\|\left\|v^t - v^*\right\|\right) \\
\leq{}& \left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 + \eta^2\gamma^2\sigma_{\max}^2\beta^2\left(\left\langle v^t, \gamma(v^t - v^*), y^t\right\rangle\right)\left\|v^t - v^*\right\|^2 \\
&- 2\eta\beta\left(\left\langle v^t, \gamma_t(v^t - v^*)\right\rangle, y^t\right)\gamma\sigma_{\min}(1-\lambda)\left\|v^t - v^*\right\|^2 \\
&+ \eta^2 LR^3\epsilon_{\text{est}}\sigma_{\max}^2\left(2\beta\left(\left\langle v^t, x^t\right\rangle, y^t\right) + LR\epsilon_{\text{est}}\right) \\
&+ 2\eta LR^2\epsilon_{\text{est}}\left(\sigma_{\max}\left\|v^t - v^*\right\| + \sigma_{\max}\left\|\mathcal{G}^\top(w^t) - v^t\right\|\right)
\end{aligned}$$

The last inequality due to the fact that $x^t = \gamma(v^t - v^*) \in \mathcal{X}$, implying $\gamma\|v^t - v^*\| \leq R$. On the other hand, we have

$$\begin{aligned}
&\left\|v^t - v^*\right\|^2 = \left\|v^t - \mathcal{G}^\top(w^t) + \mathcal{G}^\top(w^t) - v^*\right\|^2 \leq 2\left\|\mathcal{G}^\top(w^t) - v^t\right\|^2 + 2\left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 \\
&\leq 2\lambda^2\frac{\sigma_{\min}^2}{\sigma_{\max}^2}\left\|v^t - v^*\right\|^2 + 2\left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 \\
&\Rightarrow \left\|v^t - v^*\right\|^2 \leq \frac{2}{1 - 2\lambda^2\frac{\sigma_{\min}^2}{\sigma_{\max}^2}}\left\|\mathcal{G}^\top(w^t) - v^*\right\|^2.
\end{aligned}$$

Combine this into the recursion,

$$\left\|\mathcal{G}^\top(w^{t+1}) - v^*\right\|^2 \leq C_0\left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 + C_1\left(\beta\left(\left\langle v^t, x^t\right\rangle, y^t\right) + \left\|v^t - v^*\right\|\right)\epsilon_{\text{est}} + C_2\epsilon_{\text{est}}^2, \tag{7}$$

where $C_0 := \left(1 + \frac{2}{1 - 2\lambda^2\frac{\sigma_{\min}^2}{\sigma_{\max}^2}}\left(\eta^2\beta^2\left(\left\langle v^t, v^t - v^*\right\rangle, y^t\right)\gamma^2\sigma_{\max}^2 - 2\eta\beta\left(\left\langle v^t, v^t - v^*\right\rangle, y^t\right)\gamma\sigma_{\min}(1-\lambda)\right)\right)$, $C_1 := \eta^2 LR^3\sigma_{\max}^2 + 2\eta LR^2\sigma_{\max}$, and $C_2 := 2\eta LR^2\sigma_{\max} + \eta^2 L^2 R^4\sigma_{\max}^2$.

Under the ET condition, we are able to pick $\hat{x}$ and $\hat{y}$ so that $0 < \gamma\nabla_{\langle v^t, \hat{x}\rangle}\ell\left(\left\langle v^t, \hat{x}\right\rangle, \hat{y}\right) < 2\frac{(1-\lambda)\sigma_{\min}}{\eta\sigma_{\max}^2}$, we obtain,

$$C_0 = 1 + \frac{2}{1 - 2\lambda^2}\left(\eta^2\beta^2\left(\left\langle v^t, v^t - v^*\right\rangle, y^t\right)\gamma^2\sigma_{\max}^2 - 2\eta\beta\left(\left\langle v^t, v^t - v^*\right\rangle, y^t\right)\gamma\sigma_{\min}(1-\lambda)\right) \leq 1.$$

With the condition $\forall v \in \Omega_v$, $\|v\| \leq C_v$ and $\beta\left(\left\langle v, x^t\right\rangle, y^t\right) \leq C_\beta$ holds, as long as we can obtain $\epsilon_{\text{est}} = \mathcal{O}\left(\frac{1}{t^2}\right)$, $\left\|\mathcal{G}^\top(w^{t+1}) - v^*\right\|^2$ converges in rate $\mathcal{O}\left(\frac{1}{t}\right)$ (Nemirovski et al., 2009). In fact, we can achieve better converges rate, *i.e.*,

less sample complexity, with more accurate estimation in each iteration. Specifically, we expand the recursion (7),

$$\begin{aligned}
\left\|\mathcal{G}^\top(w^{t+1}) - v^*\right\|^2 &\leq C_0 \left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 + \underbrace{C_1\left(C_\beta + 2C_v\right)}_{C_1'}\epsilon_{\text{est}} + C_2\epsilon_{\text{est}}^2 \\
&\leq C_0^2 \left\|\mathcal{G}^\top(w)^{t-1} - v^*\right\|^2 + C_0\left(C_1'\epsilon_{\text{est}} + C_2\epsilon_{\text{est}}^2\right) + C_1'\epsilon_{\text{est}} + C_2\epsilon_{\text{est}}^2 \\
&\leq \cdots \\
&\leq C_0^{t+1}\left\|\mathcal{G}^\top(w^0) - v^*\right\|^2 + \left(\sum_{i=1}^{t} C_0^i\right)\left(C_1'\epsilon_{\text{est}} + C_2\epsilon_{\text{est}}^2\right) \\
&= C_0^{t+1}\left\|\mathcal{G}^\top(w^0) - v^*\right\|^2 + \frac{C_0\left(1 - C_0^t\right)}{1 - C_0}\left(C_1'\epsilon_{\text{est}} + C_2\epsilon_{\text{est}}^2\right).
\end{aligned}$$

To achieve ϵ-approximation of v^* for student, we may need the number of teaching samples to be

$$T = \left(\log\frac{1}{\sqrt{C_0}}\right)^{-1}\log\frac{2\left\|\mathcal{G}^\top(w^0) - v^*\right\|}{\epsilon} \tag{8}$$

so that $C_0^{t+1}\left\|\mathcal{G}^\top(w^0) - v^*\right\|^2 \leq \frac{\epsilon}{2}$, while the number of query samples in each iteration m should satisfy

$$\begin{cases} \frac{C_0\left(1-C_0^T\right)}{1-C_0}C_1'\epsilon_{\text{est}} \leq \frac{C_0}{1-C_0}C_1'\epsilon_{\text{est}} \leq \min\left(\frac{\epsilon}{4}, \frac{\lambda\sigma_{\min}}{\sigma_{\max}}\epsilon\right) \\ \epsilon_{\text{est}} \leq \frac{C_1'}{C_2} \end{cases} \Rightarrow \epsilon_{\text{est}} \leq \min\left(\frac{1 - C_0}{C_0C_1'}\min\left(\frac{1}{4}, \frac{\lambda\sigma_{\min}}{\sigma_{\max}}\right)\epsilon, \frac{C_1'}{C_2}\right). \tag{9}$$

Then, the total number of samples will be

$$T\left(1 + m\left(\epsilon_{\text{est}}, \frac{\delta}{T}\right)\right) = \mathcal{O}\left(\log\frac{1}{\epsilon}\left(1 + m\left(\lambda\epsilon, \frac{\delta}{\log\frac{1}{\epsilon}}\right)\right)\right).$$

■

Theorem 7 *Suppose that Assumption 1 holds. Then with probability at least $1-\delta$, then we can recover $\mathcal{G}^\top(w) \in \mathbb{R}^d$ with $\tilde{\mathcal{O}}\left(\left(d^2 + d\log\frac{1}{\delta}\right)\log\frac{1}{\epsilon}\right)$ query samples.*

Proof Similarly, we prove this claim by construction. Basically, we first approximate the $\tilde{\alpha} = \frac{\mathcal{G}^\top(w)}{\left\|\mathcal{G}^\top(w)\right\|}$ within $\Omega_\alpha = \left\{\alpha \in \mathbb{R}^d, \|\alpha\| = 1\right\}$, and then, rescale it by $\left\|\mathcal{G}^\top(w)\right\|$.

In the first stage, we exploit active learning (Balcan et al., 2009). Obvisouly, $\|v\| = 1$, therefore, after t-iteration in examination phase, we have

$$\left\|\alpha_t - \tilde{\alpha}\right\|^2 = \left\|\alpha_t\right\|^2 + \left\|\tilde{\alpha}\right\|^2 - 2\left\langle\alpha_t, \tilde{\alpha}\right\rangle = 2\left(1 - \cos\left(\alpha_t, \tilde{\alpha}\right)\right) = 2\left(1 - \sqrt{1 - \sin^2\left(\alpha_t, \tilde{\alpha}\right)}\right),$$

therefore,

$$\left\|\alpha_t - \tilde{\alpha}\right\|^2 \leq 2\sin\left(\alpha_t, \tilde{\alpha}\right).$$

which is obtained by applying $\sqrt{1 - x^2} \geq (1 - x)$ when $0 \leq x \leq 1$. Recall $\sin\left(\alpha_t, \tilde{\alpha}\right) = \mathcal{O}\left(\frac{1}{2^t\sqrt{d}}\right)$, we have

$$\left\|\alpha_t - \tilde{\alpha}\right\|^2 = \mathcal{O}\left(\frac{1}{2^t\sqrt{d}}\right),$$

which is equivalent that we can approximate $\left\|\alpha_t - \tilde{\alpha}\right\|^2 \leq \epsilon$ with $t = \mathcal{O}\left(\log\frac{1}{\epsilon}\right)$. In each iteration, the active learning make $\tilde{\mathcal{O}}\left(d^2\log d + d\log\frac{1}{\delta}\right)$ queries, implying the total sample complexity is $\tilde{\mathcal{O}}\left(\left(d^2 + d\log\frac{1}{\delta}\right)\log\frac{1}{\epsilon}\right)$.

When rescaling, we increase the error by $\left\|\mathcal{G}^\top(w)\right\|^2$, then, we can set $\epsilon' = \frac{\epsilon}{\|\mathcal{G}^\top(w)\|^2}$. When $\left\|\mathcal{G}^\top(w)\right\|$ is bounded by some constant C, which is the case, the sample we needed will be $\tilde{\mathcal{O}}\left(\left(d^2 + d\log\frac{1}{\delta}\right)\log\frac{C^2}{\epsilon}\right)$ which does not affect the asymptotic sample complexity. ■

Plug Theorem 6 with Theorem 7, we have

Corollary 8 *Suppose that Assumption 1 holds. Then then $(\ell, \mathcal{G})$ is ET with $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ teaching samples and $\tilde{\mathcal{O}}\left(\log\frac{1}{\epsilon}\log\frac{1}{\lambda\epsilon}\left(d^2 + d\log\frac{\log\frac{1}{\epsilon}}{\delta}\right)\right)$ query samples.*

A.4. Estimation Error Preservation

Lemma 9 *Suppose that $\mathcal{G}$ is a unitary operator. If $\left\|\mathcal{G}^\top(w^0) - v^0\right\| \leq \epsilon$, then $\left\|\mathcal{G}^\top(w^{t+1}) - v^{t+1}\right\| \leq \epsilon$.*

Proof This can be checked by induction, assume in t-th step, $\left\|\mathcal{G}^\top(w^t) - v^t\right\| \leq \epsilon$,

$$\begin{aligned}\left\|\mathcal{G}^\top(w^{t+1}) - v^{t+1}\right\| &= \left\|\mathcal{G}(w^t) - \eta\beta_{\langle v^t, x_t\rangle}\mathcal{G}^\top\mathcal{G}(x)_t - v^t + \eta\beta_{\langle v^t, x_t\rangle}x_t\right\| \\ &= \left\|\mathcal{G}^\top(w^t) - v^t\right\| \leq \epsilon.\end{aligned}$$

■

A.5. Extension to Combination-based and Pool-based Active Teaching

In this section, we mainly discuss the results for synthesis-based active teaching to combination-based and pool-based active learning.

For **combination-based active teaching**, where both the training samples and query samples are constructed by linear combination of k samples $\mathcal{D} = \{x_i\}_{i=1}^k$, we have the following results for exact recovery and approximate recovery in the sense of

$$\langle v_1, v_2\rangle_{\mathcal{D}} := \sqrt{v_1^\top \mathcal{D}\left(\mathcal{D}^\top\mathcal{D}\right)^+\mathcal{D}^\top v_2}, \text{ and } \|v\|_{\mathcal{D}} := \langle v, v\rangle_{\mathcal{D}}.$$

Note that with the introduced metric, for $v \in \mathbb{R}^d$, we only consider its component in $\operatorname{span}(\mathcal{D})$ and the components in the null space will be ignored. Therefore, $\forall\, v_1, v_2 \in \operatorname{span}(\mathcal{D})$ such that $\|v_1\|_{\mathcal{D}} = \|v_2\|_{\mathcal{D}}$, we have $v_1^\top x = v_2^\top x = \langle v_1, x\rangle_{\mathcal{D}}$ for all $x \in \mathbb{R}^d$. For notational convenience, we omit the subscript $\mathcal{D}$ for the analysis in this section.

Corollary 10 *Suppose the student answers questions in query phase via $F(\cdot) = I(\cdot)$ or $F(\cdot) = S(\cdot)$ and $\mathcal{G}^\top(w^0), v^* \in \operatorname{span}(\mathcal{D})$. Then $(\ell, \mathcal{G})$ is ET with $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ teaching samples and $\operatorname{rank}(\mathcal{D})$ query samples via exact recovery.*

Corollary 11 *Suppose Assumption 1 holds, the student answers questions in query phase via $F(\cdot) = I(\cdot)$ or $F(\cdot) = S(\cdot)$ and $\mathcal{G}^\top(w^0), v^* \in \operatorname{span}(\mathcal{D})$. Then $(\ell, \mathcal{G})$ is ET with $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ teaching samples and $\tilde{\mathcal{O}}\left(\log\frac{1}{\epsilon}\log\frac{1}{\lambda\epsilon}\left(d^2 + d\log\frac{\log\frac{1}{\epsilon}}{\delta}\right)\right)$ query samples via approximate recovery.*

The proof for these two corollaries are straightforward since under the condition that $\mathcal{G}^\top(w^0), v^* \in \operatorname{span}(\mathcal{D})$, every teaching sample will be in $\operatorname{span}(\mathcal{D})$, so that the $\mathcal{G}^\top(w^t)$ and v^t are also in $\operatorname{span}(\mathcal{D})$. Therefore, we can reduce such setting to synthesis-based active teaching with essential dimension as $\operatorname{rank}(\mathcal{D})$. Then, the conclusions are achieved.

For **rescaled pool-based active teaching**, where the teacher can only pick samples from a prefixed sample candidates pool, $\mathcal{D} = \{x_i\}_{i=1}^k$, for teaching and query. We will still evaluate using the same metric $\|\cdot\|_{\mathcal{D}}$ defined above (omit subscript $\mathcal{D}$ for convenience). We first discuss the exact recovery case.

Theorem 13 *Suppose the student answers questions in the exam phase via $F(\cdot) = I(\cdot)$ or $F(\cdot) = S(\cdot)$ and $\mathcal{G}^\top(w^0), v^* \in \operatorname{span}(\mathcal{D})$. If $\forall\, \mathcal{G}^\top(w) \in \operatorname{span}(D)$, there exist $(x, y) \in \mathcal{D} \times \mathcal{Y}$ and γ such that for $\hat{x} = \frac{\gamma\left\|\mathcal{G}^\top(w) - v^*\right\|_{\mathcal{D}}}{\|x\|_{\mathcal{D}}}x$, $\hat{y} = y$, we have*

$$0 \leq \gamma\nabla_{\langle v^t, \hat{x}\rangle}\ell\left(\left\langle v^t, \hat{x}\right\rangle, \hat{y}\right) \leq \frac{2\mathcal{V}(\mathcal{X})\,\sigma_{\min}}{\eta\sigma_{\max}^2},$$

then $(\ell, \mathcal{G})$ *is ET with* $\mathcal{O}\left(\log \frac{1}{\epsilon}\right)$ *teaching samples and* $\operatorname{rank}(\mathcal{D})$ *query samples.*

Proof Under the conditions that $\mathcal{G}^\top(w^0), v^* \in \operatorname{span}(\mathcal{D})$, with the same argument, in each iteration, both $\mathcal{G}^\top(w^t)$ and v^t are in $\operatorname{span}(\mathcal{D})$. Therefore, as long as we pick $\operatorname{rank}(\mathcal{D})$ independent samples from $\mathcal{D}$ as query samples, we can recover any $v \in \operatorname{span}(\mathcal{D})$ in the sense of the introduced metric. For the training sample, due to the restriction in selecting samples, we need to recheck the recursion. We follow the proof for rescaled pool-based omniscient teaching in (Liu et al., 2017a). Specifically, at t-step, as the loss is exponentially synthesis-based teachable with γ, therefore, we have the virtually constructed sample $\{x_v, y_v\}$ where $x_v = \gamma\left(\mathcal{G}^\top(w^t) - v^*\right)$ with γ satisfying the condition of exponentially synthesis-based active teachability, we first rescale the candidate pool $\mathcal{X}$ such that

$$\forall x \in \mathcal{X}, \gamma_x \|x\| = \|x_v\| = \gamma \left\|\mathcal{G}^\top(w^t) - v^*\right\|.$$

We denote the rescaled candidate pool as $\mathcal{X}_t$, under the condition of rescalable pool-based teachability, there is a sample $\{\hat{x}, \hat{y}\} \in \mathcal{X} \times \mathcal{Y}$ with scale factor $\hat{\gamma}$ such that

$$\begin{aligned}
&\min_{(x,y)\in\mathcal{X}_t\times\mathcal{Y}} \eta^2 \|\mathcal{G}^\top \nabla_{w^t} \ell\left(\left\langle w^t, \hat{\gamma}\mathcal{G}(x)\right\rangle, y\right)\|^2 - 2\eta\left\langle \mathcal{G}^\top(w^t) - v^*, \mathcal{G}^\top \nabla_{w^t} \ell\left(\left\langle w^t, \hat{\gamma}\mathcal{G}(x)\right\rangle, y\right)\right\rangle \\
&\leq \eta^2 \left\|\beta\left(\left\langle w^t, \hat{\gamma}\mathcal{G}(\hat{x})\right\rangle, \hat{y}\right)\mathcal{G}^\top\mathcal{G}(\hat{\gamma}\hat{x})\right\|^2 - 2\eta\beta\left(\left\langle w^t, \hat{\gamma}\mathcal{G}(\hat{x})\right\rangle, \hat{y}\right)\left\langle \hat{\gamma}\mathcal{G}^\top\mathcal{G}\hat{x}, \mathcal{G}^\top(w^t) - v^*\right\rangle.
\end{aligned}$$

We decompose the $\hat{\gamma}\hat{x} = a x_v + x_{v\perp}$ with $a = \frac{\langle \hat{\gamma}\hat{x}, x_v\rangle}{\|x_v\|^2}$. and $x_{v\perp} = \hat{\gamma}\hat{x} - a x_v$. Then, we have

$$\begin{aligned}
&\min_{(x,y)\in\mathcal{X}_t\times\mathcal{Y}} \eta^2 \|\mathcal{G}^\top \nabla_{w^t} \ell\left(\left\langle w^t, \mathcal{G}(x)\right\rangle, y\right)\|^2 - 2\eta\left\langle \mathcal{G}^\top(w^t) - v^*, \mathcal{G}^\top \nabla_{w^t} \ell\left(\left\langle w^t, \mathcal{G}(x)\right\rangle, y\right)\right\rangle \\
&\leq \eta^2\beta^2\left(\left\langle w^t, \hat{\gamma}\mathcal{G}(\hat{x})\right\rangle, \hat{y}\right)\left\|\hat{\gamma}\mathcal{G}^\top\mathcal{G}(\hat{x})\right\|^2 - 2\eta\beta\left(\left\langle w^t, \hat{\gamma}\mathcal{G}(\hat{x})\right\rangle, \hat{y}\right)\left\langle \hat{\gamma}\mathcal{G}^\top\mathcal{G}\hat{x}, \mathcal{G}^\top(w^t) - v^*\right\rangle \\
&\leq \eta^2\beta^2\left(\left\langle w^t, \hat{\gamma}\mathcal{G}(\hat{x})\right\rangle, \hat{y}\right)\gamma^2\sigma_{\max}^2\left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 - 2\eta\beta\left(\left\langle w^t, \hat{\gamma}\mathcal{G}(\hat{x})\right\rangle, \hat{y}\right)\sigma_{\min}\left\langle a x_v + x_{v\perp}, \mathcal{G}^\top(w^t) - v^*\right\rangle \\
&= \eta^2\beta^2\left(\left\langle w^t, \hat{\gamma}\mathcal{G}(\hat{x})\right\rangle, \hat{y}\right)\gamma^2\sigma_{\max}^2\left\|\mathcal{G}^\top(w^t) - v^*\right\|^2 - 2\eta\beta\left(\left\langle w^t, \hat{\gamma}\mathcal{G}(\hat{x})\right\rangle, \hat{y}\right)\sigma_{\min} a\left\|\mathcal{G}^\top(w^t) - v^*\right\|^2
\end{aligned}$$

Under the condition

$$0 \leq \gamma\beta\left(\left\langle w^t, \gamma\frac{\left\|\mathcal{G}^\top(w^t) - v^*\right\|}{\|x\|}\mathcal{G}(x)\right\rangle, y\right) \leq \frac{2\mathcal{V}(\mathcal{X})\sigma_{\min}}{\eta\sigma_{\max}^2},$$

we have the recursion

$$\left\|\mathcal{G}^\top(w^{t+1}) - v^*\right\|^2 \leq r\left(\eta, \gamma, \mathcal{G}, \mathcal{V}(\mathcal{X})\right)\left\|\mathcal{G}^\top(w^t) - v^*\right\|^2,$$

where $r\left(\eta, \gamma, \mathcal{G}, \mathcal{V}(\mathcal{X})\right) = \max\left\{1 + \eta^2\mu(\gamma)^2\sigma_{\max}^2 - 2\eta\mu(\gamma)\sigma_{\min}\mathcal{V}(\mathcal{X}), 1 + \eta^2\nu(\gamma)^2\sigma_{\max}^2 - 2\eta\nu(\gamma)\sigma_{\min}\mathcal{V}(\mathcal{X})\right\}$ and $0 \leq r\left(\eta, \gamma, \mathcal{G}, \mathcal{V}(\mathcal{X})\right) < 1$, with $\nu(\gamma) = \min_{w,\hat{x}\in\mathcal{X},\hat{y}\in\mathcal{Y}} \gamma\beta\left(\left\langle w^t, \gamma\frac{\left\|\mathcal{G}^\top(w^t) - v^*\right\|}{\|x\|}\mathcal{G}(x)\right\rangle, y\right) > 0$ and $\mu(\gamma) = \max_{w,\hat{x}\in\mathcal{X},\hat{y}\in\mathcal{Y}} \gamma\beta\left(\left\langle w^t, \gamma\frac{\left\|\mathcal{G}^\top(w^t) - v^*\right\|}{\|x\|}\mathcal{G}(x)\right\rangle, y\right) < \frac{2\mathcal{V}(\mathcal{X})\sigma_{\min}}{\eta\sigma_{\max}^2}$. Therefore, the algorithm converges exponentially

$$\left\|\mathcal{G}^\top(w^t) - v^*\right\|_2 \leq r\left(\eta, \gamma, \mathcal{G}, \mathcal{V}(\mathcal{X})\right)^{t/2}\left\|\mathcal{G}^\top(w^t) - v^*\right\|_2.$$

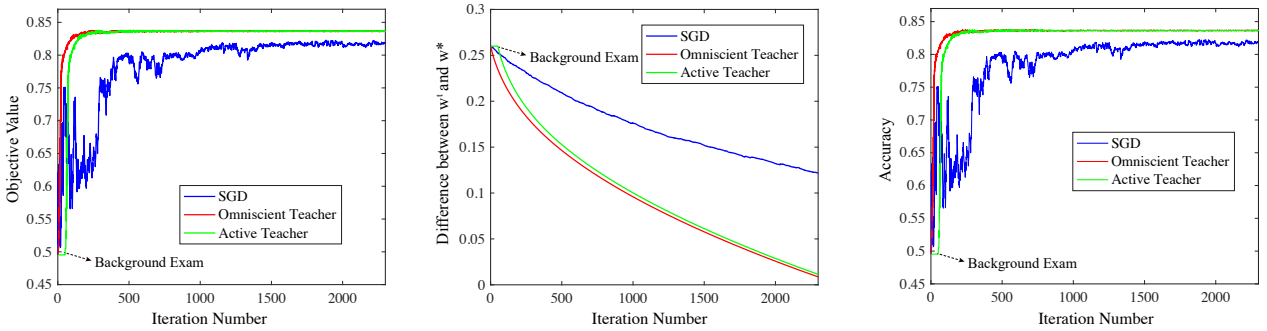
In sum, the student needs $2\left(\log\frac{1}{r(\eta,\gamma,\mathcal{G},\mathcal{V}(\mathcal{X}))}\right)^{-1}\log\frac{\|\mathcal{G}^\top(w^0) - v^*\|}{\epsilon}$ teaching samples and $\operatorname{rank}(\mathcal{D})$ query samples to achieve an ϵ-approximation of v^*. ■

For approximate recovery case, the active learning is no longer able to achieve the required accuracy for estimating of the student parameters with the restricted sample pool. Therefore, the algorithm may not achieve exponential teaching. We will leave this as an open problem.

B. Experimental Details

For synthetic data, we generate training data $(\boldsymbol{x}_i, y)$ where each entry in $\boldsymbol{x}_i$ is Gaussian distributed and $y = \langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + \epsilon$ where ϵ is a Gaussian distributed noise for the LSR learner. For the LR learner, $\{\mathcal{X}_1, +1\}$ and $\{\mathcal{X}_2, -1\}$ where $\boldsymbol{x}_i \in \mathcal{X}_1$ is Gaussian distributed in each entry and $+1, -1$ are the labels. Specifically, we use the 50-dimension data that is Gaussian distributed with $(0.5, \cdots, 0.5)$ (label +1) and $(-0.5, \cdots, -0.5)$ (label -1) as the mean and identity matrix as the covariance matrix. We generate 1000 training data points for each class. Learning rate for the same feature space is 0.0001, λ for regularization term is set as 0.00005. For the operator $\mathcal{G}$ that maps between teacher's and student's spaces, we mostly use an orthogonal transformation in experiments. In MNIST dataset, we use full training set of digits 7 and 9 and extract 24-dim projective random features from the raw 32×32 images. We use the full testing set to evaluate the 7/9 classification accuracy.

C. More Experiments: LR Learner with $F(z) = S(z)$



(a) Logistic Regression (F(z) is sigmoid) (b) Logistic Regression (F(z) is sigmoid) (c) Logistic Regression (F(z) is sigmoid)

Figure 5: The convergence performance of random teacher (SGD), omniscient teacher and active teacher in MNIST 7/9 classification. We evaluate the LR learner with $F(z) = S(z)$ here.

For the LR learner that uses the sigmoid function as feedbacks, one could clearly see that the experimental results match our theoretical analysis in case of the exact recovery of the ideal virtual learner. The active teacher is able to achieve the same performance as the omniscient teacher after the "background exam", and converges much faster than the SGD. In fact, the active teacher and the omniscient teacher should achieve the same convergence speed without consideration of numerical errors. Moreover, the empirical results indicate that the teacher tends to pick easy examples first and difficult examples later. In iterative machine teaching, the difficulty level of an example is essentially the distance between the example and the decision boundary. Interestingly, deeply learned features also exhibit similar difficulty level in terms of the norm of the feature (Liu et al., 2018; 2017b), which may be useful for improving the convergence of deep models (*e.g.*, the norm fo deeply learned features can be used as a form of difficulty indicator in curriculum learning and iterative machine teaching).

D. Analysis and Experiments of the Learner with Forgetting Behavior

D.1. Modeling the forgetting behavior

We model the forgetting behavior of the learner by adding a deviation to the learned parameter in each iteration of updating the learner. Specifically in each iteration, the learner will update its model in its feature space with

$$w^{t+1} = w^t + \nabla_w \ell(\langle w^t, x \rangle, y) + \epsilon_t \tag{10}$$

where ϵ_t is a random deviation vector. The larger the deviation is, the more the learner forgets. ϵ_t can be modeled in a time-variant fashion, or simply using a fixed probability distribution. There will be a number of ways to model the deviation. For simplicity, we only consider a Gaussian distribution with zero mean and fixed variance here. Throughout this section, we mainly study the case where the teacher and learner share the same feature space when the learner has the forgetting behavior. It could help us simplify the problem, but it also more clearly shows the superiority of the active teacher because the setting is comparable to the omniscient teacher.

D.2. The exponential teachability of the learner with forgetting behavior

Before delving deep into the exponential teachability of the learner with forgetting behavior, we first define a lazy teacher model. The lazy teacher model works essentially similar to the omniscient teacher, except that the lazy teacher will first construct a virtual learner before the teaching and will not observe the status of the learner during iteration. Specifically, the lazy teacher will first construct a virtual learner without forgetting behavior based on the initial status (information) from the real learner. Then the lazy teacher will pick samples based on the observation from the virtual learner and will feed the same samples to the real learner. One can notice that if the real learner has no forgetting behavior, the lazy teacher will be identical to the omniscient teacher. An overview of the lazy teacher is given in Fig. 6.

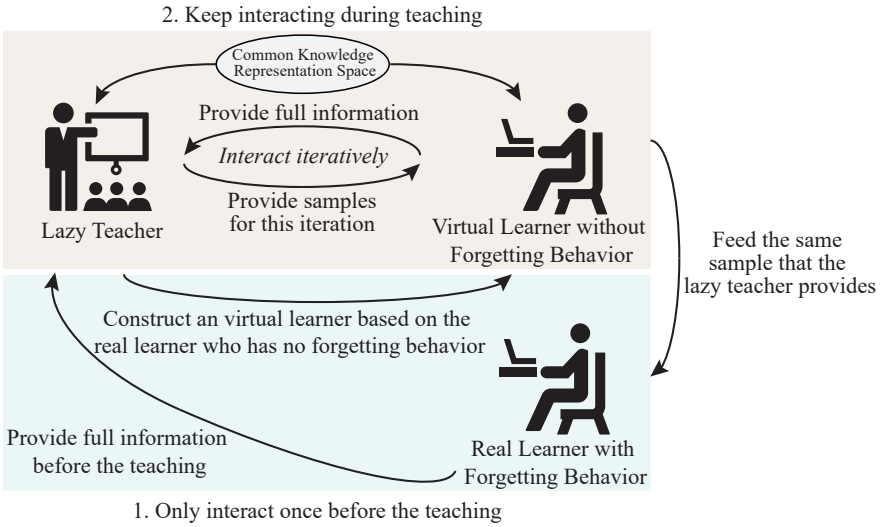


Figure 6: An illustrative overview of the lazy teacher.

For the learner guided by the active teacher to achieve ET, it requires the sample complexity of the active learning to be $\mathcal{O}(\log \frac{1}{\epsilon})$, as shown in Theorem 6. It is obvious that the deviation error ϵ_t of a forgetting learner can not converge to a small enough value. Therefore, the forgetting learner can not achieve ET with the lazy teacher, because the the deviation error can not be controlled by the lazy teacher. In contrast, the forgetting learner can still achieve ET with our proposed active teacher, because the deviation error can also be estimated by the active query strategy. In other words, the active teacher is still able to estimate accurate enough current parameters of the forgetting learner, which also prevents the deviation error to propagate over iterations.

D.3. Experiments

We perform an experiment on MNIST dataset to show how the forgetting behavior will affect the fast convergence, and also compare our active teacher with the lazy teacher. We still use the binary classification for digit 7 and 9 for our experiment.

The experimental setting for the MNIST dataset is similar to Section 7.2 except that we only use one random projection to extract the features, which means that the teacher and the learner share the same feature space. We could see from Fig. 7 that the forgetting behavior will greatly affect the convergence of the lazy learner, while the lazy learner have the same convergence speedup with the omniscient teacher if the learner has no forgetting behavior. Most importantly, our active teacher can well address this forgetting behavior and provide significant convergence speedup. This experiment also partially validates that it is very reasonable in real-world education to make students take exam. If the teacher model can not well estimate or have access to the current parameter of the student model, then the entire teaching will very possibly fail (*i.e.*, similar to or even worse than the random teacher).

Experimental settings. We perform the experiment on MNIST dataset with digit 7/9 binary classification. The 24-dim feature is computed by random projection from raw pixels. The learner will provide $F(z) = \text{sign}(z)$ as feedbacks. For fairness, the learning rates for all method are the same.

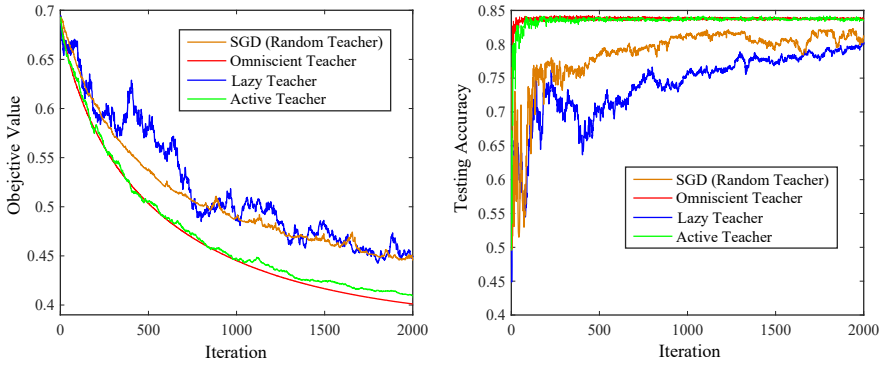


Figure 7: The convergence performance of random teacher (SGD), omniscient teacher, lazy teacher and active teacher in MNIST 7/9 binary classification.