Appendix

A. Details of the Proof

Proof of Theorem 2 We assume the optimization starts with an initialized weights w^0. t is denoted as the iteration index. Let w_g^t and w_s^t be the model parameter updated by our omniscient teacher and SGD, respectively. We first consider the case where $t = 1$. For SGD, the first gradient update w_s^1 is

$$w_s^1 = w^0 - \eta_t \frac{\partial \ell(\langle w^0, x_s \rangle, y_s)}{\partial w^0}. \tag{10}$$

Then we compute the difference between w_s^1 and w^*:

$$\begin{aligned} \left\|w_s^1 - w^*\right\|_2^2 &= \left\| w^0 - \eta_t \frac{\partial \ell(\langle w^0, x \rangle, y)}{\partial w^0} - w^* \right\|_2^2 \\ &= \left\|w^0 - w^*\right\|_2^2 + \eta_t^2 \left\| \frac{\partial \ell(\langle w^0, x \rangle, y)}{\partial w^0} \right\|_2^2 - 2\eta_t \left\langle w^0 - w^*, \frac{\partial \ell(\langle w^0, x \rangle, y)}{\partial w^0} \right\rangle \end{aligned} \tag{11}$$

Because the omniscient teacher is to minimize last two term, so we are guaranteed to have

$$\left\|w_g^1 - w^*\right\|_2^2 \leq \left\|w_s^1 - w^*\right\|_2^2. \tag{12}$$

So with the same initialization $w_g^0 = w_s^0$, $\left\|w_g^1 - w^*\right\|_2^2 \leq \left\|w_s^1 - w^*\right\|_2^2$ is always true. Then we consider the case where $t = k, k \geq 1$. We first compute the difference between w_g^{k+1} and w^*:

$$\begin{aligned} \left\|w_g^{k+1} - w^*\right\|_2^2 &= \left\| w_g^k - \eta_t \frac{\partial \ell(\langle w_g^k, x \rangle, y)}{\partial w^{k+1}} - w^* \right\|_2^2 \\ &= \left\|w_g^k - w^*\right\|_2^2 + \min_{\{x,y\}} \left\{ \eta_t^2 \left\| \frac{\partial \ell(\langle w_g^k, x \rangle, y)}{\partial w_g^k} \right\|_2^2 - 2\eta_t \left\langle w_g^k - w^*, \frac{\partial \ell(\langle w_g^k, x \rangle, y)}{\partial w_g^k} \right\rangle \right\} \\ &= \left\|w_g^k - w^*\right\|_2^2 + \eta_t^2 \left\| \frac{\partial \ell(\langle w_g^k, x_*^k \rangle, y_*^k)}{\partial w_g^k} \right\|_2^2 - 2\eta_t \left\langle w_g^k - w^*, \frac{\partial \ell(\langle w_g^k, x_*^k \rangle, y_*^k)}{\partial w_g^k} \right\rangle \\ &= \left\|w_g^k - w^*\right\|_2^2 - TV(w_g^k) \end{aligned} \tag{13}$$

where x_*^k, y_*^k is the sample selected by the omniscient teacher in the k-th iteration. Using the given conditions, we can bound the difference between w_s^{k+1} and w^* from below:

$$\begin{aligned} \left\|w_s^{k+1} - w^*\right\|_2^2 &= \left\| w_s^k - \eta_t \frac{\partial \ell(\langle w_s^k, x^s \rangle, y^s)}{\partial w_s^k} - w^* \right\|_2^2 \\ &= \left\|w_s^k - w^*\right\|_2^2 + \eta_t^2 \left\| \frac{\partial \ell(\langle w_s^k, x_s^k \rangle, y_s^k)}{\partial w_s^k} \right\|_2^2 - 2\eta_t \left\langle w_s^k - w^*, \frac{\partial \ell(\langle w_s^k, x_s^k \rangle, y_s^k)}{\partial w_s^k} \right\rangle \\ &\geq \left\|w_s^k - w^*\right\|_2^2 - TV(w_s^k) \end{aligned} \tag{14}$$

where x_s^k, y_s^k is the sample selected by the random teacher in the k-th iteration. Comparing Eq. 13 and Eq. 14 and using the condition in the theorem, the following inequality always holds under the condition $\left\|w_g^k - w^*\right\|_2^2 \leq \left\|w_s^k - w^*\right\|_2^2$:

$$\left\|w_s^{k+1} - w^*\right\|_2^2 = \left\|w_s^k - w^*\right\|_2^2 - TV(w_s^k) \geq \left\|w_g^k - w^*\right\|_2^2 - TV(w_g^k) = \left\|w_g^{k+1} - w^*\right\|_2^2. \tag{15}$$

Further because we already know that $\left\|w_g^1 - w^*\right\|_2^2 \leq \left\|w_s^1 - w^*\right\|_2^2$, using induction we can conclude that $\left\|w_g^t - w^*\right\|_2^2$ will be always not larger than $\left\|w_s^t - w^*\right\|_2^2$ (t can be any iteration). Therefore, in each iteration the omniscient teacher can always converge not slower than random teacher (SGD). ■

Proof of Proposition 3 Consider the square loss $\ell(\langle w,x\rangle,y)=(\langle w,x\rangle-y)^2$, we have $\frac{\partial\ell(\langle w,x\rangle,y)}{\partial w}=2(\langle w,x\rangle-y)x$. Suppose we are given two initializations w_1,w_2 satisfying $\|w_1-w_*\|_2^2\le\|w_2-w_*\|_2^2$. For square loss, we first write out

$$\begin{aligned}&\|w_1-w^*\|^2-TV(w_1)=\|w_1-w^*\|^2+\min_{x\in\mathcal{X},y\in\mathcal{Y}}\{\eta_t^2T_1(x,y|w_1)-2\eta_tT_2(x,y|w_1)\}\\&=\|w_1-w^*\|^2+\min_{\{x,y\}}\left\{\eta_t^2\left\|\frac{\partial\ell(\langle w_1,x\rangle,y)}{\partial w_1}\right\|_2^2-2\eta_t\left\langle w_1-w^*,\frac{\partial\ell(\langle w_1,x^*\rangle,y^*)}{\partial w_1}\right\rangle\right\}\\&=\|w_1-w^*\|^2+\begin{cases}2(\frac{R}{\|w_1-w^*\|})^2\|w_1-w^*\|^2(w_1-w^*),&\text{if }\frac{R}{\|w_1-w^*\|}<\frac{1}{\eta_t}\\-\|w_1-w^*\|^2,&\text{if }\frac{R}{\|w_1-w^*\|}\ge\frac{1}{\eta_t}\end{cases}\end{aligned}\tag{16}$$

Similarly for w_2, we have

$$\begin{aligned}&\|w_2-w^*\|^2-TV(w_2)\\&=\|w_2-w^*\|^2+\begin{cases}2(\frac{R}{\|w_2-w^*\|})^2\|w_2-w^*\|^2(w_2-w^*),&\text{if }\frac{R}{\|w_2-w^*\|}<\frac{1}{\eta_t}\\-\|w_2-w^*\|^2,&\text{if }\frac{R}{\|w_2-w^*\|}\ge\frac{1}{\eta_t}\end{cases}\end{aligned}\tag{17}$$

There will be three scenarios to consider: (1) $R\eta_t\le\|w_1-w^*\|\le\|w_2-w^*\|$; (2) $\|w_1-w^*\|\le R\eta_t\le\|w_2-w^*\|$; (3) $\|w_1-w^*\|\le\|w_2-w^*\|\le R\eta_t$. It is easy to verify that under all three scenarios, we have

$$\|w_1-w^*\|^2-TV(w_1)\le\quad\|w_2-w^*\|^2-TV(w_2)\tag{18}$$

■

To simplify notations, we denote $\beta_{(\langle w,x\rangle,y)}=\nabla_{\langle w,x\rangle}\ell(\langle w,x\rangle,y)$ for a loss function $\ell(\cdot,\cdot)$ in the following proof. For omniscient teacher, $(\hat{x},\hat{y})$ denotes a specific construction of (x,y). Notice that $(\tilde{x},\tilde{y})$ will not be used in omniscient teacher case to avoid ambiguity, since the student and the teacher use the same representation space.

Proof of Theorem 4 At t-step, the omniscient teacher selects the samples via optimization

$$\min_{x\in\mathcal{X},y\in\mathcal{Y}}\eta^2\|\nabla_{w^t}\ell\left(\left\langle w^t,x\right\rangle,y\right)\|^2-2\eta\left\langle w^t-w^*,\nabla_{w^t}\ell\left(\left\langle w^t,x\right\rangle,y\right)\right\rangle.$$

We denote $\hat{x}=\gamma(w^t-w^*)$ and $\hat{y}\in\mathcal{Y}$, since $\gamma(w-w^*)\in\mathcal{X}$, we have

$$\min_{x\in\mathcal{X},y\in\mathcal{Y}}\eta^2\|\nabla_{w^t}\ell\left(\left\langle w^t,x\right\rangle,y\right)\|^2-2\eta\left\langle w^t-w^*,\nabla_{w^t}\ell\left(\left\langle w^t,x\right\rangle,y\right)\right\rangle\tag{19}$$

$$\le\quad\left(\eta^2\beta_{(\langle w^t,\hat{x}\rangle,\hat{y})}^2\gamma^2-2\eta\beta_{(\langle w^t,\hat{x}\rangle,\hat{y})}\gamma\right)\|w^t-w^*\|_2^2.\tag{20}$$

Plug Eq. (19) into the recursion Eq. (3), we have

$$\begin{aligned}\left\|w^{t+1}-w^*\right\|_2^2&=\min_{x\in\mathcal{X},y\in\mathcal{Y}}\left\|w^t-\eta\frac{\partial\ell(\langle w,x\rangle,y)}{\partial w}-w^*\right\|_2^2\\&=\left\|w^t-w^*\right\|_2^2+\min_{x\in\mathcal{X},y\in\mathcal{Y}}\eta^2\left\|\frac{\partial\ell(\langle w^t,x\rangle,y)}{\partial w^t}\right\|_2^2-2\eta\left\langle w^t-w^*,\frac{\partial\ell(\langle w^t,x\rangle,y)}{\partial w^t}\right\rangle\\&\le\left(1+\eta^2\beta_{(\langle w^t,\hat{x}\rangle,\hat{y})}^2\gamma^2-2\eta\beta_{(\langle w^t,\hat{x}\rangle,\hat{y})}\gamma\right)\|w^t-w^*\|_2^2=\left(1-\eta\beta_{(\langle w^t,\gamma(w^t-w^*)\rangle,\hat{y})}\gamma\right)^2\|w^t-w^*\|_2^2.\end{aligned}\tag{21}$$

First we let $\nu(\gamma)=\min_{w,y}\gamma\nabla_{\langle w,\gamma(w-w^*)\rangle}\ell(\langle w,\gamma(w-w^*)\rangle,y)$. Then we have the condition $0<\nu(\gamma)\le\gamma\beta_{(\langle w,\gamma(w-w^*)\rangle,\hat{y})}\le\frac{1}{\eta}<\infty$ for any w,y, so we can obtain

$$0\le1-\gamma\eta\beta_{(\langle w,\gamma(w-w^*)\rangle,\hat{y})}\le1-\eta\nu(\gamma),$$

after simplifying $\nu(\gamma)$ to ν, we therefore have the following inequality from Eq. (21):

$$\left\|w^{t+1}-w^*\right\|_2^2\le(1-\eta\nu)^2\left\|w^t-w^*\right\|_2^2,$$

Thus we can have the exponential convergence:

$$\left\|w^t-w^*\right\|_2\le(1-\eta\nu)^t\left\|w^0-w^*\right\|_2,$$

in other words, the student needs $\left(\log\frac{1}{1-\eta\nu}\right)^{-1}\log\frac{\|w^0-w^*\|}{\epsilon}$ samples to achieve an ϵ-approximation of w^*.

■

Proof of Proposition 5 Because $\ell\left(\langle w,x\rangle,y\right)$ is ζ_1-strongly convex w.r.t. w, we have

$$\zeta_1\left(\ell\left(\langle w,x\rangle,y\right)-\min_w \ell\left(\langle w,x\rangle,y\right)\right)\leq\left\|\nabla_w\ell\left(\langle w,x\rangle,y\right)\right\|^2=\beta^2_{(\langle w,x\rangle,y)}\left\|x\right\|^2,\quad\forall\{x,y\}\in\mathcal{X}\times\mathcal{Y},$$

where $\mathcal{X}=\left\{x\in\mathbb{R}^d,\|x\|\leq R\right\}$. Using $\hat{x}=\gamma(w-w^*),\gamma\geq 0$, we have

$$\sqrt{\zeta_1\left(\ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)-\min_w \ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)\right)}\leq\beta_{(\langle w,\gamma(w-w^*)\rangle,y)}\gamma\|w-w^*\|.$$

We assume the loss function is always non-negative, i.e., $\ell\left(\langle w,x\rangle,y\right)\geq 0$. Therefore we have

$$\sqrt{\zeta_1\left(\ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)\right)}\leq\beta_{(\langle w,\gamma(w-w^*)\rangle,y)}\gamma\|w-w^*\|.$$

Because $\ell\left(\langle w,x\rangle,y\right)$ is ζ-strongly convex w.r.t. w, it is also ζ_2-strongly convex w.r.t. $\langle w,x\rangle$. Then we perform Taylor expansion to $\ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)$ w.r.t. $\langle w,x\rangle$ at the point $\langle w^*,x\rangle$ and obtain

$$\ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)\geq\ell\left(\langle w,\gamma(w^*-w^*)\rangle,y\right)+\nabla_{\langle w,x\rangle}\ell\left(\langle w,\gamma(w^*-w^*)\rangle,y\right)(w-w^*)^Tx+\frac{\zeta_2}{2}\|(w-w^*)^Tx\|^2$$

which leads to

$$\ell\left(\langle w,\gamma(w-w^*)\rangle,y\right)\geq\frac{\zeta_2}{2}\gamma^2\|w-w^*\|^4$$

Combining pieces, we have

$$\sqrt{\frac{\zeta_1\zeta_2}{2}}\gamma\|w-w^*\|\leq\beta_{(\langle w,\gamma(w-w^*)\rangle,y)}\gamma.$$

Then if we set $\gamma=\min\left\{\sqrt{\frac{2}{\zeta_1\zeta_2}}\frac{1}{\|w-w^*\|\eta},\frac{R}{\|w-w^*\|}\right\}$, we can have $\frac{1}{\eta}\leq\beta_{(\langle w,\gamma(w-w^*)\rangle,y)}\gamma$. Because $\ell\left(\langle w,x\rangle,y\right)$ is Lipschitz smooth w.r.t. $\langle w,x\rangle$ with parameter L, we have

$$\left\|\beta_{(\langle w,x\rangle,y)}-\beta_{(\langle w^*,x\rangle,y)}\right\|\leq LR\left\|w-w^*\right\|$$

Because $\beta_{(\langle w^*,x\rangle,y)}=0$, we have the following inequality:

$$\left\|\beta_{(\langle w,x\rangle,y)}\right\|\leq LR\left\|w-w^*\right\|$$

If we multiply both side with γ, we can have

$$\beta_{(\langle w,x\rangle,y)}\gamma\leq LR\left\|w-w^*\right\|\gamma$$

By setting γ as $\frac{1}{LR\eta\|w-w^*\|}$, we arrive at $\beta_{(\langle w,x\rangle,y)}\gamma<\frac{1}{\eta}$. Combining pieces, as long as we set

$$\gamma=\min\left\{\sqrt{\frac{2}{\zeta_1\zeta_2}}\frac{1}{\eta\|w-w^*\|},\frac{R}{\|w-w^*\|},\frac{1}{LR\eta\left\|w-w^*\right\|}\right\},$$

then we can have

$$0<c\leq\beta_{(\langle w,\gamma\hat{x}\rangle,\hat{y})}\gamma\leq\frac{1}{\eta}.$$

where c is a non-zero positive constant. Therefore, we achieve the condition for the exponential synthesis-based teaching.

■

By the Proposition 5, the absolute loss and sqaure loss are exponentially teachable in synthesis-based case, and we can obtain γ by plugging into the general form. We will tighten the γ up by analyzing absolute loss and square loss separately. Besides that, we also show the commonly used loss functions for classification, *e.g.*, hinge loss and logistic loss, are also exponentially teachable in synthesis-based teaching if $\|w^*\|$ can be bounded.

Proposition 9 *Absolute loss is exponentially teachable in synthesis-based teaching.*

Proof To show one loss function is exponentially teachable in synthesis-based case, we just need to find the appropriate γ such that the learning intensity is bounded below and above, according to Theorem 4. For the absolute loss, *i.e.*,

$$\ell\left(\langle w,x\rangle,y\right)=\left|\langle w,x\rangle-y\right|,$$

its sub-gradient is

$$\nabla_w\ell(\langle w,x\rangle,y)=\operatorname{sign}(\langle w,x\rangle-y)x,$$

and thus, the learning intensity $\beta_{(\langle w,x\rangle,y)}=\operatorname{sign}\left(\langle w,x\rangle-y\right)$. For $w\neq w^*$, plugging $\hat{x}=\gamma\left(w-w^*\right)$ and

$\hat{y} = \langle w^*, \gamma (w - w^*) \rangle$ into the learning intensity, we have

$$\beta_{\gamma\langle w,\hat{x}\rangle,\hat{y}}\gamma = \text{sign}\left(\gamma^2 \langle w - w^*, w - w^* \rangle\right)\gamma = \gamma.$$

Recall that $\gamma \neq 0$, $|\gamma| \leq \frac{R}{\|w^t - w^*\|}$, $\forall t \in \mathbb{N}$, we have

$$\gamma \leq \min_{t\in\mathbb{N}} \frac{R}{\|w^t - w^*\|} := C.$$

Set $\gamma = \min\{C, \frac{1}{\eta}\}$, we have $\nu = \min\{C, \frac{1}{\eta}\}$. Therefore, we obtain the exponential decay. In fact, since the $\|w^t - w^*\|$ decreases in every step, we have $C = \frac{R}{\|w^0 - w^*\|}$. In following proof, we will follow the same argument to use this fact. ■

Proposition 10 *Square loss is exponentially teachable in synthesis-based teaching.*

Proof For square loss, *i.e.*,

$$\ell\left(\langle w, x\rangle, y\right) = \left(\langle w, x\rangle - y\right)^2,$$

its gradient is

$$\nabla_w \ell\left(\langle w, x\rangle, y\right) = 2\left(\langle w, x\rangle - y\right)x,$$

and thus, the learning intensity $\beta_{\langle w,x\rangle,y} = 2\left(\langle w, x\rangle - y\right)$. For $w \neq w^*$, plugging $\hat{x} = \gamma(w - w^*)$ and $\hat{y} = \langle w^*, \gamma(w - w^*)\rangle$ into the learning intensity, we have

$$\beta_{(\langle w,\hat{x}\rangle,\hat{y})}\gamma = 2\gamma^2 \|w - w^*\|^2.$$

Set $\gamma = \min\left\{\frac{1}{\sqrt{2\eta}\|w^t - w^*\|}, \frac{R}{\|w^t - w^*\|}\right\}$, we achieve the exponential teachable condition. ■

Proposition 11 *Hinge loss is exponentially teachable in synthesis-based teaching if* $\|w^*\| \leq 1$.

Proof For hinge loss, *i.e.*,

$$\ell\left(\langle w, x\rangle, y\right) = \max\left(1 - y\langle w, x\rangle, 0\right),$$

as long as $1 - y\langle w, x\rangle > 0$, its subgradient will be

$$\nabla_w \ell\left(\langle w, x\rangle, y\right) = -yx.$$

Denote $\hat{x} = \gamma(w - w^*)$, we have $\beta_{\langle w,\hat{x}\rangle,\hat{y}} = -\hat{y}$ where $\hat{y} \in \{-1, 1\}$. To satisfy the exponential teachable condition, we need to select $\hat{y}$ and γ such that

$$\begin{cases} 1 - \hat{y}\langle w, \hat{x}\rangle > 0 \\ 0 < -\hat{y}\gamma \leq \frac{1}{\eta} \\ |\gamma| \leq \frac{R}{\|w - w^*\|} \end{cases} \Rightarrow \begin{cases} \hat{y}\gamma\langle w, w - w^*\rangle < 1 \\ -\frac{1}{\eta} \leq \hat{y}\gamma < 0 \\ |\gamma| \leq \frac{R}{\|w - w^*\|} \end{cases} \Rightarrow \begin{cases} \langle w, w - w^*\rangle > -1 \\ -\frac{1}{\eta} \leq \hat{y}\gamma < 0 \\ |\gamma| \leq \frac{R}{\|w - w^*\|} \end{cases}.$$

If $\|w^*\| \leq 1$, we can show

$$\langle w, w^*\rangle \leq \|w\|\|w^*\| \leq \|w\| < 1 + \|w\|^2,$$

where the last inequality comes from the fact $1 + a^2 - a > 0$, and thus, we have $\langle w, w - w^*\rangle > -1$. Therefore, we select any configuration of $\hat{y}$ and γ satisfying

$$-\frac{1}{\eta} \leq \hat{y}\gamma < 0, \quad \text{and} \quad |\gamma| \leq \frac{R}{\|w - w^*\|}.$$

Particularly, we set $\hat{y} = -1$ and $\gamma = \min\left\{\frac{1}{\eta}, \frac{R}{\|w^0 - w^*\|}\right\}$.

■

Proposition 12 *Logistic loss is exponentially teachable in synthesis-based teaching if* $\|w^*\| \leq 1$.

Proof For the logistic loss, *i.e.*,

$$\ell\left(\langle w, x\rangle, y\right) = \log\left(1 + \exp(-y\langle w, x\rangle)\right),$$

its gradient is

$$\nabla_w \ell\left(\langle w, x\rangle, y\right) = -\frac{yx}{1+\exp(y\langle w, x\rangle)}.$$

Denote $\hat{x} = \gamma(w - w^*)$, we have $\beta_{\langle w,\hat{x}\rangle,\hat{y}} = -\frac{\hat{y}}{1+\exp(\hat{y}\langle w,\hat{x}\rangle)}$ where $\hat{y} \in \{-1, 1\}$. To satisfy the exponential teachable condition, we need to select $\hat{y}$ and γ such that

$$\begin{cases} 0 < -\frac{\hat{y}\gamma}{1+\exp(\hat{y}\langle w,\hat{x}\rangle)} \leq \frac{1}{\eta} \\ |\gamma| \leq \frac{R}{\|w-w^*\|} \end{cases}.$$

Particularly, we set $\hat{y} = -1$, we can fix the γ by

$$0 < \frac{\gamma}{1+\exp(\gamma)} < \frac{\gamma}{1+\exp(\hat{y}\langle w,\hat{x}\rangle)} \leq \gamma \leq \frac{1}{\eta}, \quad \text{and} \quad |\gamma| \leq \frac{R}{\|w-w^*\|}.$$

The $\frac{\gamma}{1+\exp(\gamma)} < \frac{\gamma}{1+\exp(\hat{y}\langle w,\hat{x}\rangle)}$ is obtained by the monotonicity of $\exp(\cdot)$ and $\langle w, w - w^*\rangle > -1$ when $\|w^*\|$. Therefore, we can choose $\gamma = \min\left\{\frac{1}{\eta}, \frac{R}{\|w^0-w^*\|}\right\}$, and thus, the lower bound $\nu = \frac{\gamma}{1+\exp(\gamma)}$. ■

Proof of Corollary 6 In each update, given the training sample $x \in \text{span}(\mathcal{X})$, we have $w^{t+1} = w^t - \eta\beta_{\langle w,x\rangle,y}x$, therefore, the $\Delta_{t+1}w := w^{t+1} - w^0 \in \text{span}(\mathcal{X})$. If $w^0 - w^* \in \text{span}(\mathcal{X})$, $w^{t+1} - w^* \in \text{span}(\mathcal{X})$, which means by linear combination, we can construct $\hat{\gamma}\sum_{i=1}^n \alpha_i^t x_i = \gamma(w^t - w^*)$. With the condition that the loss function is exponentially synthesis-based teachable, we achieve the conclusion that the combination-based omniscient teacher will converge at least exponentially with the same rate to the synthesis-based teaching. ■

Proof of Theorem 8 The proof is similar to the synthesis-based case. However, we introduce the consideration of the effect of pool-based teaching. Specifically, we first obtain a virtual training sample in full space, and then, we generate the sample from the candidate pool to mimic the virtual sample.

With the condition $w^0 - v^* \in \text{span}(\mathcal{D})$, as we discussed in the proof of Corollary 6, in every iteration, $w^t - v^* \in \text{span}(\mathcal{D})$. Therefore, we only need to consider in the space of span $(\mathcal{D})$. Meanwhile, since the teacher can rescale the sample, without loss of generality, we assume if $x \in \mathcal{X}$, then $-x \in \mathcal{X}$ to make the rescaling is always positive.

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 $x_v = \gamma(w^t - w^*)$ with γ satisfying the condition of exponentially teachable in synthesis-based settings, we first rescale the candidate pool $\mathcal{X}$ such that

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

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

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

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

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

Under the condition

$$0 < \gamma\beta_{\langle w,\gamma\frac{w-w^*}{\hat{x}}\rangle,\hat{y}} < \frac{2\mathcal{V}(\mathcal{X})}{\eta},$$

we denote $\nu(\gamma) = \min_{w,\hat{x}\in\mathcal{X},\hat{y}\in\mathcal{Y}}\gamma\beta_{\langle w,\gamma\frac{w-w^*}{\hat{x}}\rangle,\hat{y}} > 0$ and $\mu(\gamma) = \max_{w,\hat{x}\in\mathcal{X},\hat{y}\in\mathcal{Y}}\gamma\beta_{\langle w,\gamma\frac{w-w^*}{\hat{x}}\rangle,\hat{y}} < \frac{2\mathcal{V}(\mathcal{X})}{\eta}$.

we have the recursion

$$\left\|w^{t+1}-w^*\right\|_2^2 \leq r(\eta,\gamma)\left\|w^t-w^*\right\|_2^2,$$

with $r(\eta,\gamma,\mathcal{V}(\mathcal{X})) := \max\left\{1+\eta^2\mu\left(\gamma\right)^2-2\eta\mu\left(\gamma\right)\mathcal{V}(\mathcal{X}), 1+\eta^2\nu\left(\gamma\right)^2-2\eta\nu\left(\gamma\right)\mathcal{V}(\mathcal{X})\right\}$ and $0 \leq r(\eta,\gamma) < 1$. Therefore, the algorithm converges exponentially

$$\left\|w^t-w^*\right\|_2 \leq r\left(\eta,\gamma\right)^{t/2}\left\|w^0-w^*\right\|_2,$$

in other words, the student needs $2\left(\log\frac{1}{r(\eta,\gamma,\mathcal{V}(\mathcal{X}))}\right)^{-1}\log\frac{\|w^0-w^*\|}{\epsilon}$ samples to achieve an ϵ-approximation of w^*. For clearity, we define the constant term as $C_2^{\eta,\gamma,\mathcal{V}(\mathcal{X})} = 2\left(\log\frac{1}{r(\eta,\gamma,\mathcal{V}(\mathcal{X}))}\right)^{-1}$. ■

B. Detailed Experimental Setting

Layer	CNN-6	CNN-9	CNN-12
Conv1.x	[3×3, 16]×2	[3×3, 16]×3	[3×3, 16]×4
Pool1	2×2 Max, Stride 2		
Conv2.x	[3×3, 32]×2	[3×3, 32]×3	[3×3, 32]×4
Pool2	2×2 Max, Stride 2		
Conv3.x	[3×3, 64]×2	[3×3, 64]×3	[3×3, 64]×4
Pool3	2×2 Max, Stride 2		
FC1	32	32	32

Table 1. Our standard CNN architectures for CIFAR-10. Conv1.x, Conv2.x and Conv3.x denote convolution units that may contain multiple convolution layers. E.g., [3×3, 16]×3 denotes 3 cascaded convolution layers with 16 filters of size 3×3. The CNNs learning ends at 20K iterations with multi-step rate decay.

General Settings We have used three linear models in the experiments. In specific, the formulation of ridge regression (RR) is

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(w^T x_i + b - y_i)^2 + \frac{\lambda}{2}\|w\|^2$$

The formulation of logistic regression (LR) is

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp\{-y_i(w^T x_i + b)\}) + \frac{\lambda}{2}\|w\|^2$$

The formulation of support vector machine (SVM) is

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\max(1 - y_i(w^T x_i + b), 0) + \frac{\lambda}{2}\|w\|^2$$

Comparison of different teaching strategies We use a linear regression model (ridge regression with $\lambda = 0$) for this experiment. We set R as 1 and uniformly generate 30 data points as our knowledge pool for the teacher. In this first case, we set the feature dimension as 2, while in the second case, feature dimension is 70. The learning rate is set as 0.0001 for pool-based teaching, same as BGD and SGD.

Experiments on Gaussian data Specifically, RR is run on training data $(\boldsymbol{x}_i, y)$ where each entry in $\boldsymbol{x}_i$ is Gaussian distributed and $y = \langle \boldsymbol{w}^*, \boldsymbol{x}_i\rangle + \epsilon$. LR and SVM are run on $\{\mathcal{X}_1, +1\}$ 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 10-dimension data that is Gaussian distributed with $(0.5, \cdots, 0.5)$ (label $+1$) and $(-0.5, \cdots, -0.5)$ (label -1) as mean and identity matrix as covariance matrix. We generate 1000 training data points for each class. Learning rate for the same feature space is 0.0001, while learning rate for different feature spaces are 0.00001. λ is set as 0.00005.

Experiments on uniform spherical data We first generate the training data that are uniformly distributed on a unit sphere $\|\boldsymbol{x}_i\|_2 = 1$. Then we set the data points on half of the sphere ($(0, \pi]$) as label $+1$ and the other half ($(\pi, 2\pi]$) as label -1. All the generated data points are 2D. For the scenario of different features, we use a random orthogonal projection matrix to generate the teacher's feature space from student's. Learning rate for the same feature space is 0.001, while learning rate for different feature spaces are 0.0001. λ is set as 0.00005.

Experiments on MNIST dataset We use 24D random features (projected by a random matrix $\mathbb{R}^{784\times 24}$) for the MNIST dataset. The learning rate for all the compared methods are 0.001. Note that, we generate the teacher's features using a random projection matrix ($\mathbb{R}^{24\times 24}$) from the original 24D student's features. λ is set as 0.00005.

Experiments on CIFAR-10 dataset The learning rate for all the compared methods are 0.001. λ is set as 0.00005. The goal is to learn the $\mathbb{R}^{32\times 10}$ fully connected layer, which is also the classifiers for 10 classes. The three network we use in the experiments are shown as follows:

Experiments on infant ego-centric dataset We manually crop and label all the objects that the child is holding for this experiments. For feature extraction, we use VGG-16 network that is pre-trained on Imagenet dataset. Then we use PCA to reduce the 4096 dimension to 64 dimension. We train a multi-class logistic regression to classify the objects. Note that, the omniscient teacher is also applied to train the logistic regression model. The learning rate is set to 0.001 for both SGD and omniscient teacher.

C. Comparison of different teaching strategies

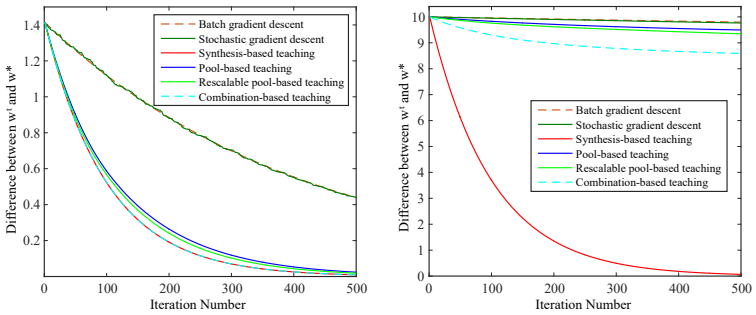


Figure 8. Comparison of different teaching strategies.

We first compare four different teaching strategies for the omniscient teacher. We consider two scenarios. One is that the dimension of feature space is smaller than the number of samples (the given features are sufficient to represent the entire feature), and the other is that the feature dimension is greater than the number of samples (the given features are not sufficient to represent the entire feature). In these two scenarios, we find that synthesis-based teaching usually works the best and always achieves exponential convergence. The combination-based teaching is exactly the same as the synthesis-based teaching in the first scenario, but it is much worse than synthesis in the second scenario. Rescalable pool-based teaching is also better than pool-based teaching. Empirically, the experiment verifies our theoretical findings: the more flexible the teaching strategy is, the more convergence gain we may obtain.

D. More experiments on MNIST dataset

We provide more experimental results on MNIST dataset. Fig. 9 shows the selected examples from 7/9 binary digit classification. The results further verify the teacher models tend to select easy examples at first and gradually shift their focuses to difficult examples, very much resembling the human learning. Fig. 10 shows the difference between the current model parameter and the optimal model parameter over iterations. It also shows that our teachers achieve faster convergence.

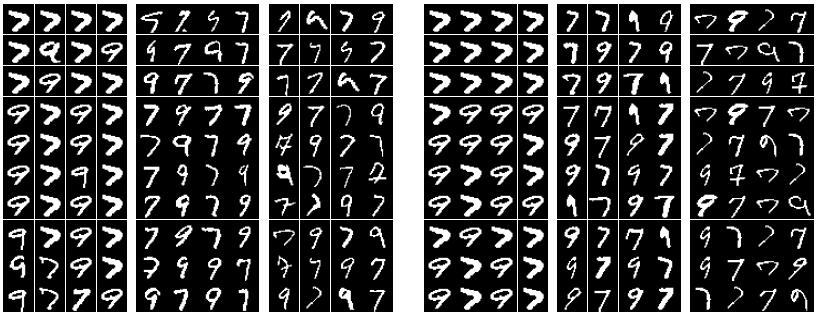


Figure 9. Selected training examples during iteration. (7/9 classification)

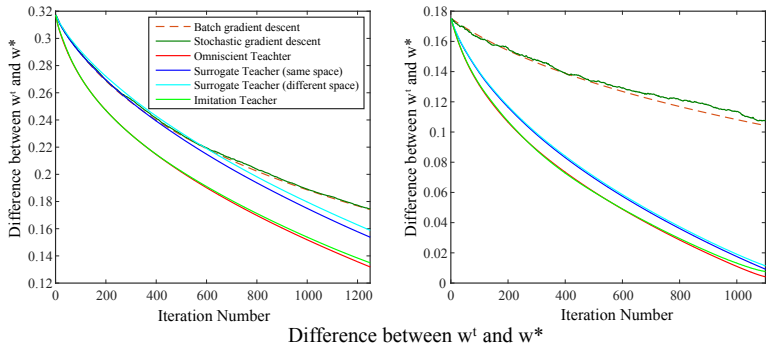


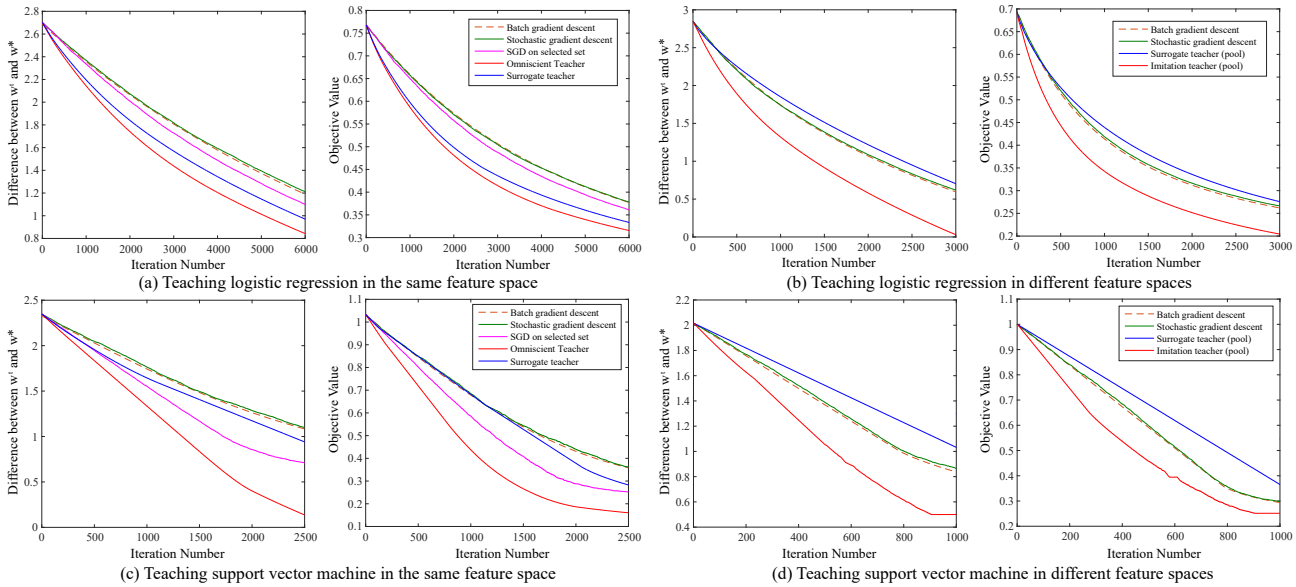
Figure 10. Teaching logistic regression on MNIST dataset. Left column: 0/1 classification. Right column: 3/5 classification

E. Teaching linear models on uniform spherical data

In this experiment, we use a different data distribution to further evaluate the teacher models. We will examine LR and SVM by classifying uniform spherical data.

Teaching in the same feature space. From Fig. 11, one can observe that the convergence is consistently improved while using omniscient teacher to provide examples to learners. We find that the significance of improvement is related to the training data distribution and loss function, as indicated by our theoretical results. The surrogate teacher produces less convergence gain in SVM, because the convexity lower bound becomes very loose in this case. Overall, omniscient teacher still presents strong teaching capability. More interestingly, we use simple SGD run on the sample set selected by the omniscient teacher and also get faster convergence, showing that the selected example set is better than the entire set in terms of convergence.

Teaching in different feature spaces. While the teacher and student use different feature spaces, one can observe from Fig. 11 that the surrogate teacher performs very poorly, even worse than the original SGD and BGD. The imitation teacher works much better and achieves consistent and significant convergence speedup, showing its superiority while the teacher and the student use different features.



(a) Teaching logistic regression in the same feature space

(b) Teaching logistic regression in different feature spaces

(c) Teaching support vector machine in the same feature space

(d) Teaching support vector machine in different feature spaces

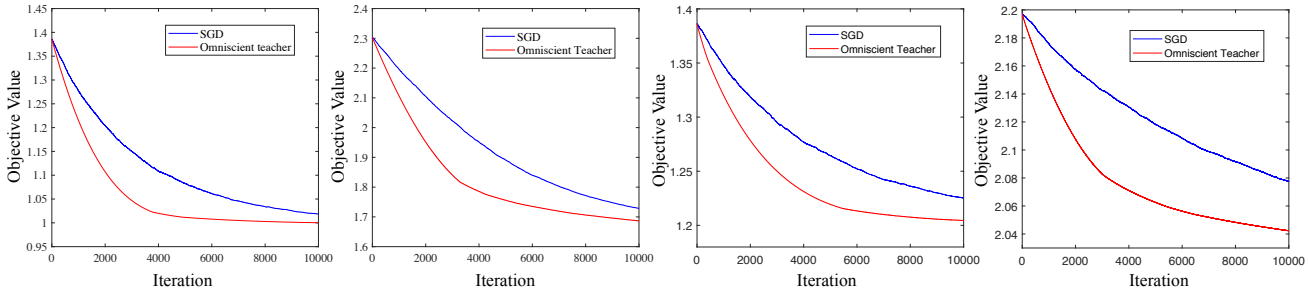
Figure 11. Convergence results on uniform spherical data.

F. Object learning experiment on children's ego-centric visual data

We experiment with a dataset capturing children and parents interacting with toys in a naturalistic setting (Yurovsky et al., 2013). These interactions are recorded for around 10.5 minutes with a camera worn low on the child's forehead. The head-camera's visual field was 90 degrees wide, providing a broad view of objects visible to the infant. The camera was attached to a headband that was tightened so that it did not move once set on the child. To calibrate the camera, the experimenter noted when the child focused on an object and adjusted the camera until the object was in the center of the image in the control monitor.

For our experiments, we selected interactions of 4 one year old infants. For each parent-child dyad, we annotated the bounding box location and category of the toy attended to by the infant at each frame. There are 10 objects in total: doll (34 frames), toy (53 frames), duck (335 frames), frog (2108 frames), helicopter (169 frames), horse (42 frames), mickey (472 frames), phone (394 frames), sheep (119 frames) and tiger (266 frames). We use a VGG-16 network that is pre-trained on Imagenet dataset as our feature extraction. We first extract the 4096D features from these images and then use PCA to reduce the dimension to 64D. Finally, we run our omniscient teacher on these ego-centric data.

One can observe from Fig. 12 that our omniscient teacher achieves faster convergence than the random teacher. Moreover, we give part of the selected training examples of random teacher and omniscient teacher in Fig. 14 and Fig. 15, respectively. We visualize the selected samples every 50 iterations from the first iteration to the 10000th iteration. Interestingly, we find that the training samples that are selected by the omniscient teacher consist of contiguous bouts of experience with the same object instance, unlike the random teacher. The adjacent samples are similar and the object changes in a smooth way. These inputs are qualitatively similar in their ordering to the actual visual experiences of infants in our study, as illustrated in Fig. 13. This can be seen as partial algorithmic confirmation of the desirable structural properties of children's natural learning environment, which emphasizes a smooth and continuous evolution of visual experience, in sharp contrast to random sample selection.



(a) Video of infant 1 (b) Video of infant 2 (c) Video of infant 3 (d) Video of infant 4

Figure 12. Convergence comparison on infant ego-centric visual data.

Figure 13. Training examples corresponding to the natural sequence of objects experienced by a single infant in our study.

Figure 14. Training examples selected by the random teacher (Stochastic gradient descent).

Figure 15. Training examples selected by the omniscient teacher.