Fuzzy Iterative Machine Teaching

Pradyot Prakash
Department of Computer Sciences
UW-Madison
pradyot@cs.wisc.edu

Ankit Pensia
Department of Computer Sciences
UW-Madison
ankitp@cs.wisc.edu

Abstract

In this paper, we explore various practical extensions of the Iterative Teaching Model proposed by Liu et al. [1]. Instead of considering an omniscient teacher, we impose the restriction on the teacher that she doesn't know the learning rate (η) precisely but only knows the interval that $\eta \in [L, U]$. In real life, it corresponds to the situation where the student may learn at different rate on different days. If the teacher is able to achieve exponential convergence in this setting, we call it *robustly* exponentially teachable. We devise strategies for teaching in such situations as well as analyze them.

1 Warmup: Iterative Teaching Framework

We here briefly describe the Iterative Teaching Framework and review the already known results for this framework. We refer the interested reader to the original paper by Liu et al. [1] for further details.

At each iteration t, teacher chooses an example (x_t, y_t) to be given to the student. Given an example (x_t, y_t), the students uses gradient descent (with the loss function $\tilde{\ell}(w, x, y)$) to update her current parameter vector w_t, i.e,

$$w_t = w_{t-1} - \eta \cdot \frac{\partial \tilde{\ell}(w, x, y)}{\partial w}$$

The student is specified by three parameters: $(w_0, \eta, \tilde{l}(\cdot, \cdot, \cdot))$. The goal for the teacher is to find a *short* sequence of examples $\{(x_t, y_t)\}_{t=1}^{T}$, i.e., minimize T, as well as ensure parameter convergence, i.e., $\|w_T - w_*\|$ is small. We restrict our attention here to linear class of models, where $\tilde{\ell}(w, x, y) = \ell(\langle w, x \rangle, y)$. Some examples of such losses are (1) linear regression: $\ell_{sq}(\langle w, x \rangle, y) = {}^1\!/_2(\langle w, x \rangle - y)^2$, (2) hinge loss: $\ell_{hi}(\langle w, x \rangle, y) = \max(0, 1 - y\langle w, x \rangle)$.

The Teaching rule proposed by Liu et al. [1] is to choose an example that satisfies the following condition

$$(x_t, y_t) = \operatorname*{argmin}_{x,y} \eta^2 T_1(w, x) - 2\eta T_2(w, x) \tag{1}$$

where $T_1(w, x, y) = \left\| \frac{\partial \tilde{\ell}(w,x,y)}{\partial w} \right\|^2$ and $T_2(w, x) = \left\langle w_t - w_*, \frac{\partial \tilde{\ell}(w,x,y)}{\partial w} \right\rangle$. This rule is motivated by the following observation:

$$\|w_{t+1} - w_*\|^2 = \|w_t - w_*\|^2 + \eta^2 \left\| \frac{\partial \tilde{\ell}(w, x, y)}{\partial w} \right\|^2 - 2\eta \left\langle w_t - w_*, \frac{\partial \tilde{\ell}(w, x, y)}{\partial w} \right\rangle \tag{2}$$

As we want the term on the left hand side to be small, we should pick the example that minimizes the right hand side.

Theorem 4 of [1] states that if the loss function and learning rate are nice enough, then it is possible to show exponential convergence.

Course project for CS861, UW-Madison

Theorem 1 (Theorem 4, [1]). *For a student with fixed learning rate $\eta \neq 0$, if the loss function satisfies that for any $w \in \mathbb{R}^d$, there exists $\gamma \neq 0$, $\gamma \leq \frac{R}{w-w_*}$, such that while $\hat{x} = \gamma(w - w_*)$ and $\hat{y} \in \mathcal{Y}$, we have*

$$0 < \gamma \nabla_{\langle w, \hat{x} \rangle} \ell(\langle w, \hat{x} \rangle, \hat{y}) \leq \frac{1}{\eta} \tag{3}$$

then the student can learn an ϵ-approximation of w_* with $\mathcal{O}(C_1^{\gamma,\eta} \log(\frac{1}{\epsilon}))$ samples where $C_1^{\gamma,\eta} = (-\log(1 - \eta\nu(\gamma))^{-1}$ and $\nu(\gamma) = \min_{w,y} \gamma \nabla_{\langle w, \hat{x} \rangle} \ell(\langle w, \hat{x} \rangle, \hat{y}) > 0$.

Absolute loss, hinge loss are exponentially teachable by this definition [1]. Note that for the squared-loss, $\nu(\gamma) = 0$ leading to vacuous bounds by Theorem 1. However, it is still exponentially teachable, see Section 2.2.

1.1 Perceptrons with lipschitz activations are exponentially machine teachable

If we have a perceptron model with $y = \sigma(w^T x)$ where $\sigma(\cdot)$ is a Lipschitz activation function, then by the Theorem 1, if the following condition is satisfied then, then it is exponentially machine teachable:

$$0 < \gamma \nabla_{\sigma(\langle w, \hat{x} \rangle)} \ell(\sigma(\langle w, \hat{x} \rangle), \hat{y}) \leq \frac{1}{\eta L} \tag{4}$$

1.2 Extension: Unknown (fixed) η and Omniscient Teacher

Consider a setting where the teacher doesn't know the (fixed) learning rate but can observe the parameter w_t of the student for every t. This setting might seem artificial but it is an extra layer of uncertainty for the teacher.

Since the updates are gradient-based and teacher knows w_t for every t, then teacher can calculate η_t after observing the updated parameter w_{t+1}.

$$\begin{aligned} w_{t+1} &= w_t - \eta_t \frac{\partial \ell(w_t, x_t)}{\partial w_t} \\ \implies \eta_t &= \frac{(w_t - w_{t-1})}{\frac{\partial \ell(w_t, x_t)}{\partial w_t}} \qquad \text{element-wise division should return same } \eta \end{aligned} \tag{5}$$

Therefore, if η is fixed, then we can calculate η exactly with just one update. With η known, we have now converted this problem to the previous problem. Hence, the iterative teaching dimension of this problem is at-most 1 more than the omniscient teacher case of Liu et al. [1].

Note that, we assumed that the example (x, y) was chosen such that gradient of the loss function was non-zero. We can always find such an example (x, y) unless $w_0 \neq w^*$.

2 Robustly Exponential machine Teachable

In this section, we consider the scenario that the learning rate η_t is unknown and changes at every iteration. Moreover, no distribution is assumed on η_t. We would still like to achieve exponential teaching in this setting where the teacher knows w_t but not the learning rate η_t. We can't follow the strategy used in Sec. 1.2 because the learning rate η_t changes at every iteration. However, we restrict the learning rate to be bounded in the region $[L, U]$ (known to the teacher) where $L \geq 0$.

Robustly Exponentially teachable: We say a strategy for a loss function is robustly exponentially teachable when only $\mathcal{O}(\log(\frac{1}{\epsilon}))$ samples are required for $\|w_t - w_*\| \leq \epsilon$ accuracy, even in the worst case choice of $\eta_t \in [L, U]$ at each step. That is, $\|w_t - w_*\| \leq \epsilon$, for any choice of $(\eta_1, \eta_2, \ldots, \eta_t) \in [L, U]^t$ where $w_{t+1} = w_t - \eta_t \nabla_{w_t} \tilde{\ell}(w_t, x_t, y_t)$.

Robust Teaching Rule: We consider the following teaching rule:

$$x_t^* = \operatorname*{argmin}_{x} \max_{\eta \in [L, U]} \eta^2 T_1(\delta_t, x) - 2\eta T(\delta_t, x) \tag{6}$$

The rest of the report is focused on showing that if the loss function is exponentially teachable with the teaching rule (Eq. 1), then it is robustly exponentially teachable with robust teaching rule (Eq. 6).

2.1 Special case: Squared loss and single η

We begin by analyzing the simpler problem of squared loss where we know the true learning rate η. This would correspond to the setting where $\eta \in [L, U]$ and $L = U$. We show that the iterative teaching dimension is exactly 1 in this case. This analysis provides us with insights which we use in the case of noisy η.

Lemma 2. *The Iterative teaching dimension for Squared-Loss is* 1.

Proof. For the square loss, $\tilde{\ell}(w, x, y) = \frac{1}{2}(w^T x - y)^2$. For the realizable setting, $y = w_*^T x$. Throughout this proof, $t = 0$.

$$\begin{aligned} w_{t+1} &= w_t - \eta \nabla_w \left(\frac{1}{2}(w_t^T x - w_*^T x)^2 \right) \\ &= w_t - \eta((w_t - w_*)^T x) x \end{aligned}$$

$$\begin{aligned} \|w_{t+1} - w_*\|^2 &= \|w_t - w_* \eta((w_t - w_*)^T x) x\|^2 \\ &= \|w_t - w_*\|^2 + \eta^2((w_t - w_*)^T x)^2 \|x\|^2 - 2\eta(w_t - w_*)^T((w_t - w_*)^T x) x \\ \|\delta_{t+1}\|^2 &= \|\delta_t\|^2 + \eta^2(\delta_t^T x)^2 \|x\|^2 - 2\eta(\delta_t^T x)^2 \end{aligned}$$

where $\delta_t = w_t - w_*$. If we take $x = \frac{1}{\sqrt{\eta}\|\delta_t\|}\delta_t$, then $\delta_t^T x = \frac{1}{\sqrt{\eta}}\|\delta_t\|$

$$\begin{aligned} \|\delta_{t+1}\|^2 &= \delta_t^2 + \eta^2 \left(\frac{1}{\sqrt{\eta}} \|\delta_t\| \right)^2 \left(\frac{1}{\sqrt{\eta}} \right)^2 - 2\eta(\frac{1}{\sqrt{\eta}}\|\delta_t\|)^2 \\ &= \delta_t^2 + \delta_t^2 - 2\delta_t^2 = 0 \end{aligned}$$

□

Therefore, the iterative teaching dimension is 1 for any $\epsilon \geq 0$. This analysis suggests that the best example x_t is of the form $x = c\delta_t$ for some scalar c depending on the learning rate η.

2.2 Square Loss is robustly exponentially teachable

We now generalize the above case in the setting where η could vary from $\eta \in [L, U]$. This could corresponding to teaching a student whose learning rate is dynamic and unknown to the teacher. Teacher would still like to teach her student with the exponential convergence to the true w_*.

Lemma 3. *The square loss is robustly exponentially machine teachable for* $\eta \in [L, U]$.

Proof. According to the teaching rule that we consider,

$$x_t^* = \underset{x}{\operatorname{argmin}} \max_{\eta \in [L,U]} \eta^2 T_1(\delta_t, x) - 2\eta T(\delta_t, x) \tag{7}$$

$$= \underset{x}{\operatorname{argmin}} \max_{\eta \in [L,U]} \eta^2 (\delta_t^T x)^2 \|x\|^2 - 2\eta(\delta_t^T x)^2 \tag{8}$$

Therefore,

$$\begin{aligned}
\eta^2 T_1(\delta_t, x_t^*) - 2\eta T(\delta_t, x_t^*) &= \min_x \max_{\eta \in [L,U]} \eta^2 (\delta_t^T x)^2 \|x\|^2 - 2\eta (\delta_t^T x)^2 \\
&\le \max_{\eta \in [L,U]} \eta^2 (\delta_t^T \tilde{x})^2 \|\tilde{x}\|^2 - 2\eta (\delta_t^T \tilde{x})^2 && \text{take } \tilde{x} = \frac{\gamma}{\|\delta_t\|} \delta_t \\
&= \max_{\eta \in [L,U]} \eta^2 (\gamma \|\delta_t\|)^2 \gamma^2 - 2\eta (\gamma \|\delta_t\|)^2 \\
&= \max_{\eta \in [L,U]} \eta \gamma^2 \|\delta_t\|^2 \left(\eta \gamma^2 - 2\right) \\
&= \gamma^2 \|\delta_t\|^2 \max_{\eta \in \{L,U\}} \eta \left(\eta \gamma^2 - 2\right) && \text{convex in } \eta \\
&= \gamma^2 \|\delta_t\|^2 \max \left(L(L\gamma^2 - 2), U(U\gamma^2 - 2)\right) \\
&= \frac{2}{L+U} \|\delta_t\|^2 \max \left(\frac{2L^2}{L+U} - 2L, \frac{2U^2}{L+U} - 2U \right) \\
&= \frac{2}{L+U} \|\delta_t\|^2 \max \left(\frac{-2UL}{(U+L)}, \frac{-2UL}{(U+L)} \right) \\
&= -\frac{4UL}{(L+U)^2} \|\delta_t\|^2
\end{aligned}$$

For the case of square loss and the above rule, it becomes

$$\begin{aligned}
\|\delta_{t+1}\|^2 &= \|\delta_t\|^2 + \eta^2 (\delta_t^T x_t^*)^2 \|x_t^*\|^2 - 2\eta (\delta_t^T x_t^*)^2 \\
&\le \|\delta_t\|^2 - \frac{4UL}{(L+U)^2} \|\delta_t\|^2 \\
&= \left(1 - \frac{4UL}{(L+U)^2}\right) \|\delta_t\|^2 \\
&= \left(\frac{U-L}{U+L}\right)^2 \|\delta_t\|^2
\end{aligned}$$

Therefore, exponential convergence for any learning rate $\eta \in [L, U]$. We also recover that the iterative teaching dimension for squared loss is 1 by setting $L = U$.

□

2.3 Robust Exponential Teaching for general loss functions

Theorem 4. *If the problem with a given loss l is exponentially teachable with learning rate $\eta = U$, then it is also robustly exponentially teachable for $\forall\ \eta \le U$ with $\mathcal{O}(C_2 \log(\frac{1}{\epsilon}))$ samples.*

Proof. Based on the discussion in the previous section, it's easy to extend the idea to a more general loss function. Assume we have a loss $\tilde{\ell} = \ell(w^T x, y)$, then $\nabla_w \ell(\langle w, x \rangle, y) = \ell' x$ where ℓ' is the derivative of the loss with respect to the first argument.

We follow the approach taken in the previous section and set $\tilde{x} = \gamma \delta_t$ and since we require that y be consistent with the target concept, w_*. This gives us that $\ell' = \ell'(w^T \tilde{x}, w_*^T \tilde{x})$ and since $\tilde{x}$ is a function of γ, let's call $\psi(w, w_*, \gamma) = \ell'(\gamma w^T \delta, \gamma w_*^T \delta)$. In the following derivation, w, w_* remain fixed so for brevity, let's call $\psi(w, w_*, \gamma)$ simply $\psi(\gamma)$.

With this idea in mind, we wish to choose our x which solves the following program and correspondingly γ according to the rule,

$$\begin{aligned}
&\min_x \max_{\eta \in [L,U]} \|\delta_t\|^2 + \eta^2 T_1(\delta_t, x) - 2\eta T(\delta_t, x) \\
&\le \min_\gamma \max_{\eta \in [L,U]} \|\delta_t\|^2 + \eta^2 \|\psi(\gamma) \gamma \delta_t\|^2 - 2\eta \left\langle \delta_t, \psi(\gamma) \gamma \delta_t \right\rangle && \text{set } x = \gamma \delta_t \\
&= \min_\gamma \max_{\eta \in [L,U]} \|\delta_t\|^2 + \eta^2 \gamma^2 \psi(\gamma)^2 \|\delta_t\|^2 - 2\eta \gamma \psi(\gamma) \|\delta_t\|^2
\end{aligned}$$

$$= \min_{\gamma} \max_{\eta \in [L,U]} (1 - \eta\gamma\psi(\gamma))^2 \|\delta_t\|^2 \tag{9}$$

We note that this objective is a quadratic as a function of η and hence the maximum occurs at the endpoints of $[L, U]$. With this, the inner maximization problem simplifies to

$$\max\left((1 - L\gamma\psi(\gamma))^2 \|\delta_t\|^2, (1 - U\gamma\psi(\gamma))^2 \|\delta_t\|^2\right). \tag{10}$$

To analyse this, let's define,

$$\nu(\gamma) = \min_{w,w_*} \gamma\psi(w, w_*, \gamma) \tag{11}$$

Also, since the loss function is exponentially teachable with learning rate $\eta = U$, it satisfies the following property,

$$\exists\, \gamma \text{ such that } 0 < \gamma\psi(w, w_*, \gamma) \leq \frac{1}{U}\ \forall\, w, w_*. \tag{12}$$

With these, we get

- $\nu(\gamma) > 0$
- $\nu(\gamma) \leq \gamma\psi(\gamma) \leq \frac{1}{U}$

This directly implies,

$$\begin{array}{rcccccl} & 0 & < & \nu(\gamma) & \leq & \gamma\psi(\gamma) & \leq \dfrac{1}{U} \\ \implies & 0 & < & L\nu(\gamma) & \leq & L\gamma\psi(\gamma) & \leq \dfrac{L}{U} \\ \implies & 1 & > & 1 - L\nu(\gamma) & \geq & 1 - L\gamma\psi(\gamma) & \geq 1 - \dfrac{L}{U} > 0 \end{array}$$

and

$$\begin{array}{rcccccl} & 0 & < & \nu(\gamma) & \leq & \gamma\psi(\gamma) & \leq \dfrac{1}{U} \\ \implies & 0 & < & U\nu(\gamma) & \leq & U\gamma\psi(\gamma) & \leq 1 \\ \implies & 1 & > & 1 - U\nu(\gamma) & \geq & 1 - U\gamma\psi(\gamma) & \geq 0 \end{array}$$

With this,

$$\begin{aligned} \max\left((1 - L\gamma\psi(\gamma))^2, (1 - U\gamma\psi(\gamma))^2\right) &\leq \max\left((1 - L\nu(\gamma))^2, (1 - U\nu(\gamma))^2\right) \\ &= (1 - L\nu(\gamma))^2 \end{aligned}$$

because $\nu(\gamma) > 0$ and hence $0 < 1 - U\nu(\gamma) < 1 - L\nu(\gamma) < 1$. This establishes,

$$\|\delta_{t+1}\|^2 \leq (1 - L\nu(\gamma))^2 \|\delta_t\|^2 \tag{13}$$

and we get exponential convergence. This recursion can be simplified further to get the constants. Below we show that this analysis works for the l_1 loss. □

2.3.1 Example: ℓ_1 loss

We can verify that the above property holds for the ℓ_1 loss as well. Note that $\ell_1(\alpha, \beta) = |\alpha - \beta|$.

$$\begin{aligned} \ell_1' &= \frac{\partial l_1(\alpha, \beta)}{\partial \alpha} \\ &= \operatorname{sgn}(\alpha - \beta) \end{aligned}$$

Therefore,

$$\begin{aligned} \psi(\gamma) &= \operatorname{sgn}\left(\langle w, \gamma\delta\rangle - \langle w_*, \gamma\delta\rangle\right) \\ &= \operatorname{sgn}\left(\gamma\|\delta\|^2\right) \\ &= \operatorname{sgn}(\gamma) \end{aligned}$$

Hence we want,

$$\begin{aligned}\gamma\psi(\gamma) &\leq \frac{1}{U}\\ \implies \gamma \cdot \operatorname{sgn}(\gamma) &\leq \frac{1}{U}\\ \implies |\gamma| &\leq \frac{1}{U}\end{aligned}$$

and by choosing some arbitrarily small γ, we can conclude that the l_1 loss guarantees exponential teachability. Also, $\nu(\gamma) = \gamma\psi(\gamma) = |\gamma|$ and we can reduce the error $\|w_t - w_*\|$ to ϵ in $\mathcal{O}(\log \frac{1}{\epsilon})$ number of steps.

2.4 Alternate analysis for l_2 loss

An observant reader might note that $\nu(\gamma) = 0$ for the squared loss and hence, we don't get exponential convergence (we require $\nu(\gamma)$ to be strictly more than 0). We provide an alternate condition (satisfied by square loss) that also ensures exponential convergence.

Theorem 5. *If the loss function satisfies Eq. 14 for every w, w^* for an interval of $\eta \in [L, U]$ for some $\gamma = \gamma^*$, then it is also robustly exponentially teachable.*

Proof. The key equation of interest in the analysis is equation 10. We equate the two terms, and try to find a specific value of γ which makes them equal (with the intention of making both of them small at the same time using a symmetry argument),

$$\begin{aligned}&(1 - L\gamma\psi(\gamma))^2 \|\delta_t\|^2 &&= (1 - U\gamma\psi(\gamma))^2 \|\delta_t\|^2\\ \implies &L^2\gamma^2\psi(\gamma)^2 - 2L\gamma\psi(\gamma) &&= U^2\gamma^2\psi(\gamma)^2 - 2U\gamma\psi(\gamma)\\ \implies &L^2\gamma\psi(\gamma) - 2L &&= U^2\gamma\psi(\gamma) - 2U\\ \implies &\left(L^2 - U^2\right)\gamma\psi(\gamma) &&= 2\left(L - U\right)\\ \implies &\gamma\psi(\gamma) &&= \frac{2}{L+U} \end{aligned} \tag{14}$$

Let γ^* be the solution to the above equation. For convergence, we want the multiplicative scalar in equation 9 < 1. This imposes the condition that γ^* should satisfy,

$$\eta^2\gamma^{*2}\psi(\gamma^*)^2 - 2\eta\gamma^*\psi(\gamma^*) < 0.$$

where $\eta \in \{L, U\}$. Plugging in γ^* from equation 14, we have,

$$\begin{aligned}\eta^2\gamma^{*2}\psi(\gamma^*)^2 - 2\eta\gamma^*\psi(\gamma^*) &= \eta^2\left(\frac{2}{L+U}\right)^2 - 2\eta\left(\frac{2}{L+U}\right)\\ &= \eta\left(\frac{2}{L+U}\right)\left(\frac{2\eta}{L+U} - 2\right)\\ &= \left(\frac{4\eta}{L+U}\right)\left(\frac{\eta - L - U}{L+U}\right)\\ &< 0.\end{aligned}$$

Hence, the solution to equation 14 gives us a contractive γ^*. Plugging in the value for $\gamma^*\psi(\gamma^*)$ into equation 9, we get the value,

$$\begin{aligned}\|\delta_{t+1}\|^2 &\leq \left(1 - L\frac{2}{L+U}\right)^2 \|\delta_t\|^2\\ &= \left(\frac{U-L}{U+L}\right)^2 \|\delta_t\|^2\end{aligned}$$

which gives us exponential convergence for any $\eta \in [L, U]$ and is considerably tighter than the one obtained in the section 2.3.

Thus, we conclude that as long as one can find a solution to equation 14, this approach will guarantee exponential teachability. □

2.4.1 l_1, l_2 losses

It is easy to verify that this property holds for the l_1 loss. See section 2.2 for a proof of l_2 loss. This is easily extendable for l_p^p losses in the same vein.

3 Conclusion and Future Work

We built upon the ideas of Liu et al. [1] in the context of iterative machine teaching. In this report, we showed that a slight modification of the teaching rule proposed by the authors of the paper can be generalized to perceptrons with lipschitz activations, thus making them exponentially teachable. More importantly, we also looked at the setting where the learning rate is not fixed a priori but bounded. We were able to prove that this setting is robustly exponentially machine teachable, and hence generalized the findings of [1]. This immediately implied that our ideas are robust to the choice of learning rate η. One are of extension would be to teach a collection of students in a ‘classroom’ each with her own different (but bounded) learning rate with the same set of examples. We believe that a slightly modified teaching rule could be used to achieve this for smooth loss functions like squared loss.

Acknowledgement

Authors thank Ayon Sen for helpful discussions and guidance.

References

[1] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B. Smith, James M. Rehg, and Le Song. Iterative machine teaching. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2149–2158, Sydney, Australia, 06–11 Aug 2017. PMLR.