A Brief Survey of Private and Online Learning

Matthew T. Regehr

April 25, 2022

1 Introduction

1.1 Private Learning

Differential privacy is a property of a randomized algorithm demanding that the distribution of its output is not too sensitive to local changes to the input [Dwo06]. This makes it more difficult to maliciously infer sensitive local information in the input to an algorithm when the output is publicly released. A typical formalization is as follows:

Definition 1. *For a measurable space $\mathcal{W}$, a randomized algorithm $A : \mathcal{Z}^m \to \mathcal{W}$ is called (ϵ, δ)-private if*

$$\mathbb{P}(A(z) \in W) \leq e^{\epsilon}\, \mathbb{P}(A(z') \in W) + \delta$$

holds for all measurable $W \subseteq \mathcal{W}$ and $z, z' \in \mathcal{Z}^m$ such that z and z' differ in one coordinate.

Learnability, on the other hand, may be treated as a property classes of hypotheses mapping features to labels demanding that an arbitrary hypothesis can be approximately recovered from random samples labelled by the hypothesis.

Definition 2. *Given a distribution $D \in \Delta(\mathcal{X} \times \mathbb{F}_2)$, the population loss of a hypothesis $h \in \mathbb{F}_2^{\mathcal{X}}$ with respect to D is $L_D(h) := \mathbb{P}_{(x,y)\sim D}(h(x) \neq y)$. The loss of a class $\mathcal{H} \subseteq \mathbb{F}_2^{\mathcal{X}}$ is $L_D(\mathcal{H}) := \inf_{h\in\mathcal{H}} L_D(h)$.*

Definition 3. *We say that a (possibly randomized) algorithm $A : (\mathcal{X} \times \mathbb{F}_2)^m$ is a (α, β)-PAC learner for a class $\mathcal{H} \subseteq \mathbb{F}_2^{\mathcal{X}}$ if, for any distribution $D \in \Delta(\mathcal{X} \times \mathbb{F}_2)$ s.t. $L_D(\mathcal{H}) = 0$, we have*

$$\mathbb{P}_{S\sim D^m}(L_D(A(S)) > \alpha) \leq \beta.$$

In this language, learnability just means that α and β can be made arbitrarily small, provided that m is large enough. The line of work on private learning begins with [RSL+08], who extend this notion of learning to account for privacy.

Definition 4. *We say that $A : (\mathcal{X} \times \mathbb{F}_2)^* \to \mathbb{F}_2^{\mathcal{X}}$ is a private realizable PAC learner[1] for a class $\mathcal{H} \subseteq \mathbb{F}_2^{\mathcal{X}}$ with sample complexity $m : (0,1)^4 \to \mathbb{N}$ when, for any $\alpha, \beta \in (0,1)$ and $\epsilon, \delta > 0$, $A|_{(\mathcal{X}\times\mathbb{F}_2)^{m(\alpha,\beta,\epsilon,\delta)}}$ is (ϵ, δ)-private and (α, β)-accurate for $\mathcal{H}$.*

[1] Known closure properties of learning allow us to boost these results to the agnostic setting [ABMS20].

1.2 Online Learning

Roughly, a class is online learnable when some algorithm that receives examples labelled by an unknown hypothesis one at a time and predicts a new hypothesis at every step has sublinear regret.

Definition 5. *A sample $S := ((x_1, y_1), \dots, (x_m, y_m)) \in (\mathcal{X} \times \mathbb{F}_2)^*$ is called $\mathcal{H}$-realizable if there is some $h \in \mathcal{H}$ s.t. $y_t = h(x_t)$ for all $t \in [m]$, in which case we define the regret of an algorithm $A : (\mathcal{X} \times \mathbb{F}_2)^* \to \mathbb{F}_2^{\mathcal{X}}$ as*

$$R_A(S) := \left|\{t \in [m] : A((x_1, y_1), \dots, (x_{t-1}, y_{t-1}))(x_t) \neq y_t\}\right|.$$

Definition 6. *A class $\mathcal{H}$ is realizably online learnable if there is some algorithm A such that*

$$\lim_{t\to\infty} \frac{R_A((x_1, h(x_1)), \dots, (x_t, h(x_t))}{t} = 0$$

for any $x_1, x_2, \dots \in \mathcal{X}$ and $h \in \mathcal{H}$.

It turns out that, much like other models of learnability, online learnability is characterized by a combinatorial measure [Lit88].

Definition 7. *Representing a full tree of depth d labelled by a feature space $\mathcal{X}$ as a map $T : \bigcup_{t=0}^{d-1} \mathbb{F}_2^t \to \mathcal{X}$, we say that $\mathcal{H} \subseteq \mathbb{F}_2^{\mathcal{X}}$ shatters T if, for every $y_1, \dots, y_d \in \mathbb{F}_2$, there is some $h \in \mathcal{H}$ for which $y_t = h(T(y_1, \dots, y_{t-1}))$ for every $t \in [d]$. The Littlestone dimension of $\mathcal{H}$, $\operatorname{LDim} \mathcal{H}$, is the largest positive integer d such that $\mathcal{H}$ shatters some $T : \bigcup_{t=0}^{d-1} \mathbb{F}_2^t \to \mathcal{X}$.*

Theorem 1. *A class $\mathcal{H}$ is realizably online learnable if and only if $\operatorname{LDim} \mathcal{H} < \infty$, in which case there is an online learner SOA satisfying $R_{\mathrm{SOA}}(S) \leq \operatorname{LDim} \mathcal{H}$ for every realizable S.*

2 Relationship Between Private and Online Learning

2.1 Private Learning Implies Online Learning

A recent line of work establishes a rich connection between these seemingly unrelated models of learning. The first paper, [ALMM19], shows that private learnability implies online learnability.

Theorem 2. *For any $\mathcal{H}$ and private realizable PAC learner A with sample complexity m,*

$$m := m(1/16, 1/16, 1/10, \delta) = \Omega(\log^*(\operatorname{LDim} \mathcal{H}))$$

for some $\delta = O\left(\frac{1}{m^2 \log m}\right)$ (provided such a δ exists for A). In particular, $\operatorname{LDim} \mathcal{H}$ must be finite and $\mathcal{H}$ must be realizably online learnable by Theorem 1.

We remark that the condition $\delta = O\left(\frac{1}{m^2 \log m}\right)$ is actually quite mild and indeed holds in the standard regime of differential privacy $\delta = m^{-\omega(1)}$ (see e.g. [Vad17]).

To prove this result, the authors first invoke a theorem of [She90] to reduce to the case where $\mathcal{H}$ is a class of thresholds.

The main remaining hurdle to overcome is that the given private learner may not be proper. The authors leverage Ramsey theory [ER52] to argue that even a randomized improper learner must obey satisfy some form of regularity over a sufficiently large subdomain.

Lemma 1. *For any ordered set $\mathcal{X}$ of size n, $A : (\mathcal{X} \times \mathbb{F}_2)^* \to \mathbb{F}_2^{\mathcal{X}}$, and even $m \in 2\mathbb{N}$, there is $\mathcal{X}' \subseteq \mathcal{X}$ with size at least $\frac{\log^{(m)}(n)}{2^{O(m \log m)}}$ as well as $p_0, \dots, p_m \in [0,1]$ such that*

$$\left|\mathbb{E}[A(((x_1,0),\dots,(x_{m/2},0),(x_{m/2+1},1),\dots,(x_m,1)))(x)] - p_i\right| = O(1/m)$$

for any $x_1 < \cdots < x_i < x < x_{i+1} < \cdots < x_m \in \mathcal{X}'$.

In the case of thresholds, this regularity essentially implies that the expected output of A is approximately proper. The authors then use binary search to construct an identification attack on the output of the private learner whenever n is too large, which contradicts privacy.

To conclude this subsection, it also worth noting that the $\log^*$ dependence on $\operatorname{LDim} \mathcal{H}$ is nearly tight for some classes. In particular, [KLM+20] show that for a class $\mathcal{H}$ of thresholds can be privately learned in $\tilde{O}((\log^*(\operatorname{LDim} \mathcal{H})^{1.5})$ samples.

2.2 Online Learning Implies Private Learning

A followup collaboration with Mark Bun [BLM20] addresses the converse.

Theorem 3. *Let $\mathcal{H} \subseteq \mathbb{F}_2^{\mathcal{X}}$ have finite Littlestone dimension d. There exists a realizable PAC learner for $\mathcal{H}$ with sample complexity*

$$m(\alpha, \beta, \epsilon, \delta) = O\left(\frac{2^{\tilde{O}(2^d)} - \log(\beta\delta)}{\alpha\epsilon}\right).$$

They prove this result via a new notion of stability:

Definition 8. *A (possibly randomized learning algorithm) $A : (\mathcal{X} \times \mathbb{F}_2)^* \to \mathbb{F}_2^{\mathcal{X}}$ is called globally (m, η)-stable with respect to $\mathcal{D} \in \Delta(\mathcal{X} \times \mathbb{F}_2)$*

$$\exists h^* \in \mathbb{F}_2^{\mathcal{X}} \quad \mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}(A(S) = h^*) \geq \eta.$$

The authors argue that, for a class $\mathcal{H}$ of finite Littlestone dimension d, there is a globally $(2^{2^{O(d)}}, 2^{-2^{O(d)}})$-stable learner with respect to any distribution $D \in \Delta(\mathcal{X} \times \mathbb{F}_2)$ for which $L_D(h^*) = 0$ for some $h^* \in \mathcal{H}$. Intuitively, this should not be suprising as SOA makes only finitely many mistakes and thus we should expect it to identify h^* with high probability on sufficiently large samples. Although the precise proof is much more involved, the idea is that the more mistakes an online learner mistakes, the more quickly it can identify h^*, so we should try to force SOA to make as many mistakes as early as possible. We can achieve this by inductively drawing pairs of samples from D, randomly choosing between them, and then appending an example that causes a mistake. Algorithm 1 gives a detailed account of this sampling procedure and the resulting globally stable learner.

Algorithm 1 A globally stable learner for Littlestone classes

Set N and n carefully
procedure SAMPLE(k)
  if $k = 0$ **then**
    return ()
  else if
    repeat then
      $(S_1, S_2) \leftarrow (\text{SAMPLE}(k-1), \text{SAMPLE}(k-1))$
      $(T_1, T_2) \sim D^n \otimes D^n$
      if Total number of calls to D so far exceeds N **then**
        return "fail"
      end if
    until $y_1 := \text{SOA}(S_1 \circ T_1)(x) \neq \text{SOA}(S_2 \circ T_2)(x) =: y_2$ for some $x \in \mathcal{X}$
    return $S_1 \circ T_1 \circ ((x, y_2))$ or $S_2 \circ T_2 \circ ((x, y_1))$, each with probability $1/2$
  end if
end procedure
procedure LEARNER
  $k \sim \mathcal{U}(\{0, \ldots, d\})$
  $S \leftarrow \text{SAMPLE}(k)$
  $T \sim D^n$
  return $\text{SOA}(S \circ T)$
end procedure

Globally (m, η)-stable learner G in hand, the next step is to construct a private learner. Given a random sample S, we would like to privatize ERM via the exponential mechanism [MT07]. In particular, we would like to return a random hypothesis $h \in \mathcal{H}$ with probability proportional to $e^{-cL_S(h)}$, where c is some appropriately chosen constant to ensure privacy. However, the exponential mechanism will not succeed until we narrow the search space (it must be finite at least). This is where global stability comes in. The trick is to draw k samples $(S_1, \ldots, S_k) \sim (D^m)^k$ (for sufficiently large k), form a private histogram [KKMN09] from the candidate hypotheses $G(S_1), \ldots, G(S_k)$ that is accurate to within $O(\eta)$ (with high probability), and discard those with estimated frequency much less than η. Using the standard statistical toolbelt (e.g. uniform convergence, Chernoff), one can show that the resulting finite list of hypotheses contains h^* with high probability and so sampling it via the exponential mechanism with empirical risk as its score function achieves good generalization.

The sampling procedure SAMPLE has been significantly streamlined by [GGKM21] to reduce the dependence on d from doubly exponential to polynomial. They also show how to boost a private improper learner to a private proper learner via some notion of irreducibility.

Theorem 4. *Let $\mathcal{H} \subseteq \mathbb{F}_2^{\mathcal{X}}$ have finite Littlestone dimension d. There exists a private, proper realizable PAC learner for $\mathcal{H}$ with sample complexity*

$$m(\alpha, \beta, \epsilon, \delta) = O\left(\frac{d^6 \log\left(\frac{d}{\epsilon\delta\alpha\beta}\right)^2}{\epsilon\alpha^2}\right).$$

2.3 Complexity

Understanding on the computational complexity of the reduction from online to private learning and vice-versa remains only partial. [Bun20] shows that, under typical cryptographic assumptions, efficient private learnability does not imply efficient online learnability:

Theorem 5. *Assuming the existence of one-way functions, there is a hypothesis class that is privately PAC learnable in polynomial-time but is not online learnable by any polytime algorithm with a polynomial mistake bound.*

The proof involves the same hypothesis class used by [Blu94] to show an analogous computational separation from PAC learnability to online learnability.

2.4 Open Problems

Two open problems of interest to me are:

1. The gap in sample complexity from $\log^*(d)$ to d^6 is still very large. Can this dependence on d be tightened further? The result of [KLM$^+$20] suggests that the optimal dependence on d may be closer to $\log^*(d)$, although this remains to be seen.

2. While the existence of a computationally efficient reduction from private to online learning has been settled by [Bun20], the converse remains open. Is there a polytime reduction from online to private learning?

References

[ABMS20] Noga Alon, Amos Beimel, Shay Moran, and Uri Stemmer. Closure properties for private classification and online prediction. In *Conference on Learning Theory*, pages 119–152. PMLR, 2020.

[ALMM19] Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860, 2019.

[BLM20] Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402. IEEE, 2020.

[Blu94] Avrim L Blum. Separating distribution-free and mistake-bound learning models over the boolean domain. *SIAM Journal on Computing*, 23(5):990–1000, 1994.

[Bun20] Mark Bun. A computational separation between private learning and online learning. *Advances in Neural Information Processing Systems*, 33:20732–20743, 2020.

[Dwo06] Cynthia Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Verlag, July 2006.

[ER52] Paul Erdos and Richard Rado. Combinatorial theorems on classifications of subsets of a given set. *Proceedings of the London mathematical Society*, 3(1):417–439, 1952.

[GGKM21] Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Sample-efficient proper pac learning with approximate differential privacy. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 183–196, 2021.

[KKMN09] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. In *Proceedings of the 18th international conference on World wide web*, pages 171–180, 2009.

[KLM+20] Haim Kaplan, Katrina Ligett, Yishay Mansour, Moni Naor, and Uri Stemmer. Privately learning thresholds: Closing the exponential gap. In *Conference on Learning Theory*, pages 2263–2285. PMLR, 2020.

[Lit88] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[RSL+08] Sofya Raskhodnikova, Adam Smith, Homin K Lee, Kobbi Nissim, and Shiva Prasad Kasiviswanathan. What can we learn privately. In *Proceedings of the 54th Annual Symposium on Foundations of Computer Science*, pages 531–540, 2008.

[She90] Saharon Shelah. *Classification theory: and the number of non-isomorphic models*. Elsevier, 1990.

[Vad17] Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer, 2017.