Yale University CPSC 683
Stability in Machine Learning: Generalization, Privacy & Replicability

Lecture 8: Replicable and DP PAC Learning

Instructor: Alkis Kalavasis, alkis.kalavasis@yale.edu

In Lecture 8, we discuss about connections between replicability, total variation (TV) indistinguishability, and differential privacy in the context of PAC learning.

1 Pseudo-determinism and Replicability

In Lecture 7, we introduced the notion of *global stability* (Bun et al., 2020) that was crucial in order to show that online learnability (i.e., finite Littlestone dimension) implies DP PAC learnability of a concept class. Another way to view global stability is in the context of *pseudo-deterministic algorithms* (Gat and Goldwasser, 2011; Chen et al., 2023; Grossman and Liu, 2019; Goldwasser et al., 2017).

A pseudo-deterministic algorithm is a randomized algorithm which yields some fixed output with high probability (i.e., it looks like a deterministic algorithm to a computationally bounded observer). Thinking of a realizable distribution $\mathcal{D}$ as an instance on which PAC learning algorithm has oracle access, a globally-stable learner is one which is ”weakly” pseudo-deterministic: it produces some fixed output with probability bounded away from zero (recall that the global stability parameter that we could achieve was quite small – doubly exponentially small in the Littlestone dimension).

The definition of pseudo-deterministic algorithms inspired the work of Impagliazzo et al. (2022) to introduce the following definition to capture replicability for ML algorithms.

> **Definition 1: Replicability (Impagliazzo et al., 2022)**
>
> Fix some source of randomness $\mathcal{R}$. An algorithm A is ρ-replicable if for any $S, S' \sim \mathcal{D}^n$, it holds that
>
> $$\Pr_{S,S' \sim \mathcal{D}^n, r \sim \mathcal{R}}[A(S,r) \neq A(S',r)] \leq \rho\,.$$

This definition says that with probability at least $1-\rho$, the outputs of the algorithm trained on two i.i.d. datasets S and S' with shared randomness (i.e., Python seed) will be *exactly* the same. The shared random string can be seen as a way to achieve a *coupling* between two executions of the algorithm A. An interesting aspect of this definition is that replicability is verifiable; replicability under Definition 1 can be tested using polynomially many samples, random seeds r and queries to A.

Impagliazzo et al. (2022) proved an interesting positive result: any SQ algorithm can be made replicable.

Recall that a statistical query oracle for a distribution $\mathcal{D}$ over Z with accuracy τ is an oracle that gets as input a statistical query $q : Z \to [-1,1]$ and returns a value $v \in [\mathbf{E}_{\mathcal{D}}\, q - \tau, \mathbf{E}_{\mathcal{D}}\, q + \tau]$.

> **Exercise 1**
>
> Prove that there is an algorithm that simulates the statistical query oracles using samples from $\mathcal{D}$. Then prove that this algorithm can be made replicable with a blow-up in the sample complexity.

2 A Weaker Requirement: Total Variation Indistinguishability

At first glance, replicability seems like a rather strong requirement. We will now see a weaker definition, studied in Kalavasis et al. (2023). Total variation distance between two distributions P and Q over the probability space (Ω, Σ_Ω) can be expressed as

$$\begin{aligned} \mathrm{TV}(P,Q) &= \sup_{A \in \Sigma_\Omega} P(A) - Q(A) \\ &= \inf_{(X,Y) \sim \Pi(P,Q)} \mathbf{Pr}[X \neq Y]\,, \end{aligned} \tag{1}$$

where the infimum is over all couplings between P and Q so that the associated marginals are P and Q respectively. A *coupling* between the distributions P and Q is a set of variables (X, Y) on some common probability space with the given marginals, i.e., $X \sim P$ and $Y \sim Q$. We think of a coupling as a construction of random variables X, Y with the prescribed laws.

This definition leads to the following relaxation of replicability.

Definition 2: Total Variation Indistinguishability

A learning rule A is n-sample ρ-TV indistinguishable if for any distribution over inputs $\mathcal{D}$ and two independent sets $S, S' \sim \mathcal{D}^n$ it holds that

$$\mathop{\mathbf{E}}_{S,S' \sim \mathcal{D}^n}[\mathrm{TV}(A(S), A(S'))] \leq \rho\,.$$

This information-theoretic definition of TV indistinguishability seems to put weaker restrictions on learning rules than the notion of replicability in two ways:

- it allows for *arbitrary* couplings between the two executions of the algorithm (recall the coupling definition of TV distance), and,
- it allows for *different* couplings between every pair of datasets S, S' (the optimal coupling in the definition of TV distance will depend on S, S').

In short, our definition allows for *arbitrary data-dependent* couplings, instead of just sharing the randomness across two executions. We note that this definition is not new and it turns out that it corresponds to one-way perfect generalization (Cummings et al., 2016).

3 Privacy Reminder

The notion of algorithmic replicability that we have discussed so far has close connections with the classical definition of approximate differential privacy. For $a, b, \epsilon, \delta \in [0, 1]$, let $a \approx_{\epsilon,\delta} b$ denote the statement $a \leq e^\epsilon b + \delta$ and $b \leq e^\epsilon a + \delta$. We say that two probability distributions P, Q are (ϵ, δ)-indistinguishable if $P(E) \approx_{\epsilon,\delta} Q(E)$ for any measurable event E.

Definition 3: Approximate Differential Privacy

A learning rule A is an n-sample (ϵ, δ)-differentially private if for any pair of samples $S, S' \in (\mathcal{X} \times \{0,1\})^n$ that disagree on a single example, the induced posterior distributions $A(S)$ and $A(S')$ are (ϵ, δ)-indistinguishable.

As we saw in previous lectures, in the context of PAC learning, any hypothesis class $\mathcal{H}$ can be PAC-learned by an approximate differentially-private algorithm if and only if it has a finite Littlestone dimension $\mathrm{Ldim}(\mathcal{H})$, i.e., there is a qualitative equivalence between online learnability and private PAC learnability.

The results of the next sections appear in Kalavasis et al. (2023). Similar results independently appear in Bun et al. (2023).

4 Replicability $\iff$ TV Indistinguishability

We will prove the following result.

Theorem 1: Replicability ⇒ TV Indistinguishability

If a learning rule A is n-sample ρ-replicable, then it is also n-sample ρ-TV indistinguishable.

Proof. Fix some distribution $\mathcal{D}$ over inputs. Let A be n-sample ρ-replicable with respect to $\mathcal{D}$. For the random variables $A(S), A(S')$ where $S, S' \sim \mathcal{D}^n$ are two independent samples and using Eq.(1), we have

$$\mathop{\mathbf{E}}_{S,S' \sim \mathcal{D}^n}[\mathrm{TV}(A(S), A(S'))] = \mathop{\mathbf{E}}_{S,S' \sim \mathcal{D}^n}\left[\inf_{(h,h') \sim \Pi(A(S),A(S'))} \mathbf{Pr}[h \neq h']\right] . \tag{2}$$

Let $\mathcal{R}$ be the source of randomness that A uses. The expected optimal coupling of Eq.(2) is at most $\mathbf{E}_{S,S' \sim \mathcal{D}^n}\left[\mathbf{Pr}_{r \sim \mathcal{R}}[A(S,r) \neq A(S',r)]\right]$. This inequality follows from the fact that using shared randomness between the two executions of A is a particular way to couple the two random variables. To complete the proof, it suffices to notice that this upper bound is equal to

$$\mathop{\mathbf{Pr}}_{S,S' \sim \mathcal{D}^n, r \sim \mathcal{R}}[A(S,r) \neq A(S',r)] \leq \rho\,.$$

The last inequality follows since A is ρ-replicable. □

We now deal with the opposite direction, i.e., we show that TV indistinguishability implies replicability. In order to be formal, we need to discuss some measure theoretic properties first. Let us recall the definition of absolute continuity for two measures.

Definition 4: Absolute Continuity

Consider two measures P, Q on a σ-algebra $\mathcal{B}$ of subsets of Ω. We say that P is absolutely continuous with respect to Q if for any $E \in \mathcal{B}$ such that $Q(E) = 0$, it holds that $P(E) = 0$.

Since the learning rules induce posterior distributions over hypotheses, this definition extends naturally to such rules.

Definition 5

Given learning rule A, distribution over inputs $\mathcal{D}$ and reference probability measure $\mathcal{P}$, we say that A is absolutely continuous with respect to $\mathcal{P}$ on inputs from $\mathcal{D}$ if, for almost every sample S drawn from $\mathcal{D}$, the posterior distribution $A(S)$ is absolutely continuous with respect to $\mathcal{P}$.

In the previous definition, we fixed the data-generating distribution $\mathcal{D}$. We next consider its distribution-free version.

Definition 6

Given learning rule A and reference probability measure $\mathcal{P}$, we say that A is absolutely continuous with respect to $\mathcal{P}$ if, for any distribution over inputs $\mathcal{D}$, A is absolutely continuous with respect to $\mathcal{P}$ on inputs from $\mathcal{D}$.

If $\mathcal{X}$ is finite, then one can take $\mathcal{P}$ to be the uniform probability measure over $\{0,1\}^{\mathcal{X}}$ and any learning rule is absolutely continuous with respect to $\mathcal{P}$. We now show how we can find such a prior $\mathcal{P}$ in the case where $\mathcal{X}$ is countable.

Lemma 1: Reference Probability Measure for Countable Domains

Let $\mathcal{X}$ be a countable domain and A be a learning rule. Then, there is a reference probability measure $\mathcal{P}$ such that A is absolutely continuous with respect to $\mathcal{P}$.

Proof. Since $\mathcal{X}$ is countable, for a fixed n, we can consider an enumeration of all the n-tuples $\{S_i\}_{i \in \mathbb{N}}$. Then, we can take $\mathcal{P}$ to be a countable mixture of these probability measures, i.e., $\mathcal{P} = \sum_{i=1}^{\infty} \frac{1}{2^i} A(S_i)$. Notice that since, each $A(S_i)$ is a measure and $1/2^i > 0$ for $i \in \mathbb{N}$, and, $\sum_{i=1}^{\infty} 1/2^i = 1$, we have that $\mathcal{P}$ is indeed a probability measure. We now argue that each $A(S_i)$ is absolutely continuous with respect to $\mathcal{P}$. Assume towards contradiction that this is not the case and let $E \in \mathcal{B}$ be a set such that $\mathcal{P}(E) = 0$ but $A(S_j)(E) \neq 0$, for some $j \in \mathbb{N}$. Notice that $A(S_j)$ appears with coefficient $1/2^j > 0$ in the mixture that we consider, hence if $A(S_j)(E) > 0 \implies 1/2^j A(S_j)(E) > 0$. Moreover $A(S_i)(E) \geq 0, \forall i \in \mathbb{N}$, which means that $\mathcal{P}(E) > 0$, so we get a contradiction. □

We next define when two learning rules A, A' are equivalent.

Definition 7: Equivalent Learning Rules

Two learning rules A, A' are equivalent if for every sample S it holds that $A(S) = A'(S)$, i.e., for the same input they induce the same distribution over hypotheses.

In the next result, we show that for every TV indistinguishable algorithm A, that is absolutely continuous with respect to some reference probability measure $\mathcal{P}$, there exists an equivalent learning rule which is replicable.

Theorem 2: TV Indistinguishability ⇒ Replicability

Let $\mathcal{P}$ be a reference probability measure over $\{0,1\}^{\mathcal{X}}$, and let A be a learning rule that is n-sample ρ-TV indistinguishable and absolutely continuous with respect to $\mathcal{P}$. Then, there exists an equivalent learning rule A' that is n-sample $\frac{2\rho}{1+\rho}$-replicable.

This implies the following:

Theorem 3

Let $\mathcal{X}$ be a countable domain and let A be a learning rule that is n-sample ρ-TV indistinguishable. Then, there exists an equivalent learning rule A' that is n-sample $\frac{2\rho}{1+\rho}$-replicable.

Proof Sketch. Let us consider a learning rule A satisfying the conditions of Theorem 4. Fix a distribution $\mathcal{D}$ over inputs. The crux of the proof is that given two random variables X, Y whose TV distance is bounded by ρ, we can couple them using only a carefully designed source of shared randomness $\mathcal{R}$ so that the probability that the realizations of these random variables differ is at most $2\rho/(1+\rho)$. We can instantiate this observation with $X = A(S)$ and $Y = A(S')$. Crucially, in the countable $\mathcal{X}$ setting, we can pick the shared randomness $\mathcal{R}$ in a way that only depends on the learning rule A, but not on S or S'. Let us now describe how this coupling works. Essentially, it can be thought of as a generalization of the von Neumann rejection-based sampling which does not necessarily require that the distribution has bounded density. Following Angel and Spinka (2019), we pick $\mathcal{R}$ to be a Poisson point process which generates points of the form (h, y, t) with intensity[1] $\mathcal{P} \times \mathrm{Leb} \times \mathrm{Leb}$, where $\mathcal{P}$ is a reference probability measure with respect to which A is absolutely continuous and Leb is the Lebesgue measure over $\mathbb{R}_+$. Intuitively, $h \sim \mathcal{P}$ lies in the hypotheses' space, y is a non-negative real value and t corresponds to a time value. The coupling mechanism performs *rejection sampling* for each distribution we would like to couple (here $A(S)$ and $A(S')$): it checks (in the ordering indicated by the time parameter) for each point (h, y, t) whether $f(h) > y$ (i.e., if y falls below the density curve f at h) and accepts the first point that satisfies this condition. In the formal proof, there will be two density functions; f (resp. f') for the density function of $A(S)$ (resp. $A(S')$). We also refer to 1. One can show that $\mathcal{R}$ gives rise to a coupling between $A(S)$ and $A(S')$ under the condition that both measures are absolutely continuous with respect to the reference probability measure $\mathcal{P}$.

This coupling technique appears in Angel and Spinka (2019), which shows:

Theorem 4: Pairwise Optimal Coupling

Let $\mathcal{S}$ be any collection of random variables that are absolutely continuous with respect to a common probability measure[a] μ. Then, there exists a coupling of the variables in $\mathcal{S}$ such that, for any $X, Y \in \mathcal{S}$,

$$\mathbf{Pr}[X \neq Y] \leq \frac{2\mathrm{TV}(X, Y)}{1 + \mathrm{TV}(X, Y)} .$$

Moreover, this coupling requires sample access to a Poisson point process with intensity $\mu \times \mathrm{Leb} \times \mathrm{Leb}$, where Leb is the Lebesgue measure over $\mathbb{R}_+$, and full access to the densities of all the random variables in $\mathcal{S}$ with respect to μ.

[a] This result extends to the setting where μ is a σ-finite measure, but it is not needed for the purposes of our work.

Insight 1: Intuition of the above Result

Crucially, the points generated by the Poisson point process $\mathcal{R}$ are generated once. However, we may accept a different point for X and a different point for Y. The key property is that the probability that we pick a different point is roughly speaking the total variation distance between the two distributions.

Insight 2: Correlated Sampling

The above rejection sampling process is one way to couple the random variables. Another way to couple the two random variables is via **correlated sampling.** We refer to Bavarian et al. (2016) and Bun et al. (2023) for details.

We can then apply it and get

$$\mathop{\mathbf{Pr}}_{r \sim \mathcal{R}}[A(S, r) \neq A(S', r)] \leq \frac{2\mathrm{TV}(A(S), A(S'))}{1 + \mathrm{TV}(A(S), A(S'))} .$$

Taking the expectation with respect to the draws of S, S', we show (after some algebraic manipulations) that $\mathbf{Pr}_{S,S' \sim \mathcal{D}^n r \sim \mathcal{R}}[A(S, r) \neq A(S', r)] \leq 2\rho/(1 + \rho)$. We conclude this section with the following remarks.

[1] Roughly speaking, a point process is a (general) Poisson point process with intensity λ if (i) the number of points in a bounded Borel set E is a Poisson random variable with mean $\lambda(E)$ and (ii) the numbers of points in n disjoint Borel sets forms n independent random variables.

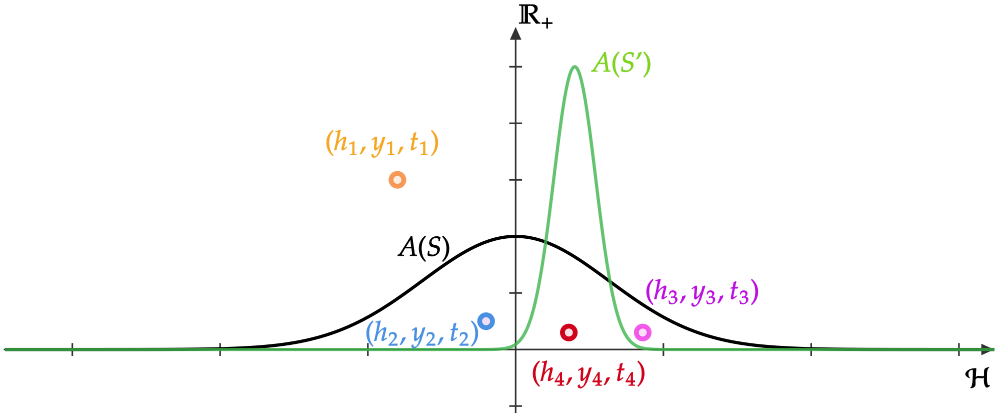

Figure 1: Taken from Kalavasis et al. (2023). Our goal is to couple $A(S)$ with $A(S')$, where these two distributions are absolutely continuous with respect to the reference probability measure $\mathcal{P}$. A sequence of points of the form (h, y, t) is generated by the Poisson point process with intensity $\mathcal{P} \times \text{Leb} \times \text{Leb}$ where $h \sim \mathcal{P}, (y, t) \in \mathbb{R}_+^2$ and Leb is the Lebesgue measure over $\mathbb{R}_+$ (note that we do not have upper bounds for the densities). Intuitively, h lies in the hypotheses' space, y is a non-negative real value and t corresponds to a time value. Let f be the Radon-Nikodym derivate of $A(S)$ with respect to $\mathcal{P}$. We assign the first (the one with minimum t) value h to $A(S)$ that satisfies the property that $f(h) > y$, i.e., y falls below the density curve of $A(S)$. We assign a hypothesis to $A(S')$ in a similar manner. This procedure defines a data-independent way to couple the two random variables and naturally extends to multiple ones. In the example, we set $A(S) = h_2$ and $A(S') = h_4$ given that $t_1 < t_2 < t_3 < t_4$.

5 TV Indistinguishability $\iff$ DP

In this section we study the connections with differential privacy. We start with the following result.

Theorem 5: (ϵ, δ)-DP $\Rightarrow$ TV Indistinguishability

Let $\mathcal{X}$ be a (possibly infinite) domain and $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$. Let $\gamma \in (0, 1/2), \alpha, \beta, \rho \in (0,1)^3$. Assume that $\mathcal{H}$ is learnable by an n-sample $(1/2 - \gamma, 1/2 - \gamma)$-accurate $(0.1, 1/(n^2 \log(n)))$-differentially private learner. Then, it is also learnable by an (α, β)-accurate ρ-TV indistinguishable learning rule.

Proof Sketch The proof goes through the notion of global stability. The existence of an (ϵ, δ)-DP learner implies that the hypothesis class $\mathcal{H}$ has finite Littlestone dimension (Alon et al., 2019). Thus, we know that there exists a ρ-globally stable learner for $\mathcal{H}$ (Bun et al., 2020). The next step is to use the replicable heavy-hitters algorithm (Impagliazzo et al., 2022) with frequency parameter $O(\rho)$ and replicability parameter $O(\rho')$, where $\rho' \in (0, 1)$ is the desired TV indistinguishability parameter of the learning rule.

Global stability implies that the list of heavy-hitters will be non-empty and it will contain at least one hypothesis with small error rate, with high probability. Finally, since the list of heavy-hitters is finite and has bounded size, we feed the output into the replicable agnostic learner. Thus, we have designed a replicable learner for $\mathcal{H}$, and Theorem 4 shows that this learner is also TV indistinguishable.

We proceed to the opposite direction where we provide an algorithm that takes as input a TV indistinguishable learning rule for $\mathcal{H}$ and outputs a learner for $\mathcal{H}$ which is (ϵ, δ)-DP. In this direction countability of $\mathcal{X}$ is crucial.

Theorem 6: TV Indistinguishability $\Rightarrow$ (ϵ, δ)-DP

Let $\mathcal{X}$ be a countable domain. Assume that $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ is learnable by an (α, β)-accurate ρ-TV indistinguishable learner A, for some $\rho \in (0,1), \alpha \in (0,1/2), \beta \in \left(0, \frac{1-\rho}{1+\rho}\right)$. Then, for any $(\alpha', \beta', \varepsilon, \delta) \in (0,1)^4$, it is also learnable by an $(\alpha + \alpha', \beta')$-accurate (ε, δ)-differentially private learner A'.

On a high level, the proof resembles the approaches of Bun et al. (2020); Ghazi et al. (2021); Bun et al. (2023) to show that (pseudo-)global stability implies differential privacy. We consider k' different batches of k datasets of size n. For each such batch, our goal is to couple the k different executions of the algorithm on an input of size n, so that most of these outputs are, with high probability, the same.

One first approach would be to use a random variable as a "pivot" element in each batch: we first draw $A(S_1^1)$ according to its distribution and the remaining $\{A(S_i^1)\}_{i \in [k] \setminus \{1\}}$ from their optimal coupling with $A(S_1^1)$, given its realized value. Even though this coupling has the property that, in expectation, most of the outputs will be the same, it is not robust at all. If the adversary changes a point of S_1^1, then the values of all the outputs will change! This is not privacy preserving. For this reason, we use the coupling that we used before. We use the fact that $\mathcal{X}$ is countable to design a reference probability measure $\mathcal{P}$ that is independent of the data. This is the key step that leads to privacy-preservation. Then, we can argue that if we follow this approach for multiple batches, there will be a classifier whose frequency and performance are non-trivial. The next step is to feed all these hypotheses into the Stable Histograms algorithm, which will output a list of frequent hypotheses that includes the non-trivial one we mentioned above. Finally, we feed these hypotheses into the Generic Private Learner and we get the desired result.

Hence the algorithm looks as follows:

1. Since $\mathcal{X}$ is countable, we define $\mathcal{P} = \sum A(S_i)/2^i$ for some enumeration of the n-element sets.
2. Now we perform the von-Neumann coupling $\Pi_{\mathcal{R}}$ in each batch to get classifiers
$$(h_1^j, ..., h_k^j) \leftarrow \Pi_{\mathcal{R}}(A(S_1^j), ..., A(S_k^j)), j \in [k'].$$
3. Next, we apply to each batch the stable histograms algorithm (e.g., check Lecture 7) which outputs the most frequent elements of a list in a private manner. We get that
$$L^j = \mathrm{StableHist}(h_1^j, ..., h_k^j)\,.$$
4. Remove rare elements from the lists and get $\widetilde{L}^j$.
5. Apply a standard DP PAC learner for the finite concept class $\cup_{j \in [k']} \ \widetilde{L}^j$.

References

Alon, N., Livni, R., Malliaris, M., and Moran, S. (2019). Private pac learning implies finite littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 852–860.

Angel, O. and Spinka, Y. (2019). Pairwise optimal coupling of multiple random variables. *arXiv preprint arXiv:1903.00632*.

Bavarian, M., Ghazi, B., Haramaty, E., Kamath, P., Rivest, R. L., and Sudan, M. (2016). Optimality of correlated sampling strategies. *arXiv preprint arXiv:1612.01041*.

Bun, M., Gaboardi, M., Hopkins, M., Impagliazzo, R., Lei, R., Pitassi, T., Sivakumar, S., and Sorrell, J. (2023). Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 520–527, New York, NY, USA. Association for Computing Machinery.

Bun, M., Livni, R., and Moran, S. (2020). An equivalence between private classification and online prediction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 389–402. IEEE.

Chen, L., Lu, Z., Oliveira, I. C., Ren, H., and Santhanam, R. (2023). Polynomial-time pseudodeterministic construction of primes. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1261–1270. IEEE.

Cummings, R., Ligett, K., Nissim, K., Roth, A., and Wu, Z. S. (2016). Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory*, pages 772–814. PMLR.

Gat, E. and Goldwasser, S. (2011). Probabilistic search algorithms with unique answers and their cryptographic applications. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 18, pages 1–3.

Ghazi, B., Kumar, R., and Manurangsi, P. (2021). User-level private learning via correlated sampling. *arXiv preprint arXiv:2110.11208*.

Goldwasser, S., Grossman, O., and Holden, D. (2017). Pseudo-deterministic proofs. *arXiv preprint arXiv:1706.04641*.

Grossman, O. and Liu, Y. P. (2019). Reproducibility and pseudo-determinism in log-space. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 606–620. SIAM.

Impagliazzo, R., Lei, R., Pitassi, T., and Sorrell, J. (2022). Reproducibility in learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 818–831.

Kalavasis, A., Karbasi, A., Moran, S., and Velegkas, G. (2023). Statistical indistinguishability of learning algorithms. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15586–15622. PMLR.