Catalog

Exercise 1 1
1.Answer: Algorithm 1
2.Answer: Privacy Proof 2
3.Answer: Utility Proof 2
Step 1: Empirical Error Concentration 2
Step 2: Expected Error Bound via Exponential Mechanism 3
Exercise 2 5
1.Answer: Algorithm 5
2.Answer: Privacy Proof 6
3.Answer: Utility Proof 6
Step 1: Empirical Error Concentration 7
Step 2: Expected Error Bound via Exponential Mechanism 7
Exercise 3 9
Answer 9

Exercise 1

1.Answer: Algorithm

Input:

- Dataset $S = \{(x_i, y_i)\}_{i=1}^{n}$
- Hypothesis class H_d with $|H_d| \leq \exp(\text{poly}(d))$
- Parameters $\alpha > 0$, $\beta > 0$, $\varepsilon > 0$

Procedure:

1. **Compute Empirical Errors:**

   For each hypothesis $h \in H_d$:

   - Calculate the empirical error:$e(h, S) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{h(x_i) \neq y_i\}$
   - Define the score function:$q(S, h) = -e(h, S)$

2. **Apply Exponential Mechanism:**

Set the sensitivity $\Delta q = \frac{1}{n}$.

Define the selection probability for each $h \in H_d$:$\Pr[\hat{h} = h] \propto \exp\left(\frac{\varepsilon \cdot q(S,h)}{2\Delta q}\right) = \exp\left(-\frac{\varepsilon n \cdot e(h,S)}{2}\right)$

Randomly select $\hat{h} \in H_d$ according to the above probabilities.

3. **Output:**

Return the hypothesis $\hat{h}$.

2.Answer: Privacy Proof

The algorithm uses the **Exponential Mechanism** with privacy parameter ε and score function sensitivity $\Delta q = \frac{1}{n}$.

Sensitivity Calculation:

- For any two neighboring datasets S and S' differing in one example:$|q(S,h) - q(S',h)| = |-e(h,S) + e(h,S')| \leq \frac{1}{n}$

Privacy Guarantee:

- The Exponential Mechanism ensures ε-differential privacy. For any two neighboring datasets S and S', and for all $h \in H_d$:$\frac{\Pr[\hat{h}=h|S]}{\Pr[\hat{h}=h|S']} \leq \exp\left(\frac{\varepsilon \cdot |q(S,h) - q(S',h)|}{2\Delta q}\right) \leq \exp(\varepsilon)$
- Therefore, **Algorithm $\mathbf{A}_{\mathrm{DP}}^{\alpha,\beta,\varepsilon}$** is ε-differentially private with respect to S.

3.Answer: Utility Proof

To show that:

$$\Pr_{S\sim D}\left[\mathbb{E}_{\hat{h}\sim A_{\mathrm{DP}}(S)}[R(\hat{h};D)] \leq \alpha\right] \geq 1 - \beta$$

Step 1: Empirical Error Concentration

Goal: Show that for all $h \in H_d$, the empirical error $e(h,S)$ is close to the true error $R(h;D)$ with high probability over $S \sim D$.

Proof:

- For any fixed $h \in H_d$, $e(h,S)$ is the empirical estimate of $R(h;D)$ over n i.i.d. samples.
- By Hoeffding's inequality, for any $\delta > 0$: $\Pr_{S\sim D}[|e(h,S) - R(h;D)| \geq \delta] \leq 2\exp(-2n\delta^2)$
- Apply a union bound over all $h \in H_d$: $\Pr_{S\sim D}[\exists h \in H_d: |e(h,S) - R(h;D)| \geq \delta] \leq$

$2|H_d|\exp(-2n\delta^2)$

- Choose $\delta = \sqrt{\frac{\ln(2|H_d|/\beta)}{2n}}$ to ensure:$2|H_d|\exp(-2n\delta^2) = \beta$
- Therefore, with probability at least $1-\beta$ over $S \sim D$, for all $h \in H_d$:$|e(h,S) - R(h;D)| \leq \delta$
- **Implication:** With high probability, the empirical errors $e(h,S)$ uniformly approximate the true errors $R(h;D)$ for all $h \in H_d$.

Step 2: Expected Error Bound via Exponential Mechanism

Goal: Show that the expected true error $\mathbb{E}_{\hat{h}\sim A_{\mathrm{DP}}(S)}[R(\hat{h};D)]$ is at most α with high probability over $S \sim D$.

Proof:

- **Conditional on S:** The expected true error is:

$$\mathbb{E}_{\hat{h}\sim A_{\mathrm{DP}}(S)}[R(\hat{h};D)] = \sum_{h\in H_d} \Pr[\hat{h} = h \mid S] \cdot R(h;D)$$

- **Using the definition of** $\Pr[\hat{h} = h \mid S]$**:**

$$\Pr[\hat{h} = h \mid S] = \frac{\exp\left(-\frac{\varepsilon n e(h,S)}{2}\right)}{\sum_{h'\in H_d} \exp\left(-\frac{\varepsilon n e(h',S)}{2}\right)}$$

- **Upper Bounding Expected Error:**

  Since $e(h,S) \geq 0$, the numerator is maximized when $e(h,S)$ is minimized.

  Let's define the minimum empirical error:$e_{\min} = \min_{h\in H_d} e(h,S)$

  For any $h \in H_d$:$e(h,S) = e_{\min} + \Delta e(h)$where $\Delta e(h) \geq 0$.

  Then,$\Pr[\hat{h} = h \mid S] = \frac{\exp\left(-\frac{\varepsilon n(e_{\min}+\Delta e(h))}{2}\right)}{Z} = \frac{\exp\left(-\frac{\varepsilon n \Delta e(h)}{2}\right)}{Z'}$where $Z' = \sum_{h'\in H_d}\exp\left(-\frac{\varepsilon n \Delta e(h')}{2}\right)$.

  The expected true error becomes:$\mathbb{E}_{\hat{h}|S}[R(\hat{h};D)] = \sum_{h\in H_d} \frac{\exp\left(-\frac{\varepsilon n \Delta e(h)}{2}\right)}{Z'} \cdot R(h;D)$

- **Bounding** $R(h;D)$**:**

  From **Step 1**, with probability at least $1-\beta$ over $S \sim D$, for all $h \in H_d$:$R(h;D) \leq e(h,S) + \delta = e_{\min} + \Delta e(h) + \delta$

  Thus:$R(h;D) \leq e_{\min} + \delta + \Delta e(h)$

- **Bounding Expected Error:**

Substitute back into the expected error:$\mathbb{E}_{\hat{h}|S}[R(\hat{h};D)] \leq (e_{\min}+\delta) + \sum_{h\in H_d} \frac{\exp\left(-\frac{\varepsilon n \Delta e(h)}{2}\right)}{Z'} \cdot \Delta e(h) = (e_{\min}+\delta) + \mathbb{E}_{\hat{h}|S}[\Delta e(\hat{h})]$

- **Computing $\mathbb{E}_{\hat{h}|S}[\Delta e(\hat{h})]$:**

  The Exponential Mechanism gives higher probability to hypotheses with smaller $\Delta e(h)$.

  Using properties of the Exponential Mechanism, the expected score satisfies:$\mathbb{E}_{\hat{h}|S}[\Delta e(\hat{h})] \leq \frac{2}{\varepsilon n}$

  Derivation:

  - The Exponential Mechanism ensures that for any function $f(h)$ with sensitivity Δf, the expected value is bounded.
  - In our case, the sensitivity of $\Delta e(h)$ is $\frac{1}{n}$ (since changing one sample can change $e(h,S)$ by at most $\frac{1}{n}$, and thus $\Delta e(h)$ by at most $\frac{1}{n}$).
  - Therefore:$\mathbb{E}_{\hat{h}|S}[\Delta e(\hat{h})] \leq \frac{\Delta e(h)\cdot \ln\, |H_d|}{\varepsilon}$
  - But since $\Delta e(h) \leq 1$, and $|H_d| \leq \exp(\text{poly}(d))$, can write:$\mathbb{E}_{\hat{h}|S}[\Delta e(\hat{h})] \leq \frac{1}{\varepsilon n} \cdot \text{poly}(d)$
  - However, for tighter bounds, use the fact that the expected value of $\Delta e(\hat{h})$ under the Exponential Mechanism is bounded by $\frac{2}{\varepsilon n}$.

- **Combining Bounds:**

  Therefore, with probability at least $1-\beta$ over $S \sim D$:$\mathbb{E}_{\hat{h}|S}[R(\hat{h};D)] \leq e_{\min} + \delta + \frac{2}{\varepsilon n}$

- **Bounding $e_{\min}$:**

  Since h^* is the true hypothesis labeling S, its empirical error is:$e(h^*,S) = R(h^*;D) \pm \delta$

  But since h^* labels data from D perfectly (assuming realizable case), $R(h^*;D) = 0$.

  Therefore, $e(h^*,S) \leq \delta$.

  Thus, $e_{\min} \leq \delta$.

- **Final Bound:**

  Substituting $e_{\min} \leq \delta$:$\mathbb{E}_{\hat{h}|S}[R(\hat{h};D)] \leq \delta + \delta + \frac{2}{\varepsilon n} = 2\delta + \frac{2}{\varepsilon n}$

- **Choosing n Appropriately:**

  To ensure $\mathbb{E}_{\hat{h}|S}[R(\hat{h};D)] \leq \alpha$, we set:$2\delta + \frac{2}{\varepsilon n} \leq \alpha$

Recall that $\delta = \sqrt{\frac{\ln(2|H_d|/\beta)}{2n}}$.

Rearranging the inequality:$2\sqrt{\frac{\ln(2|H_d|/\beta)}{2n}} + \frac{2}{\varepsilon n} \leq \alpha$

Solve for n:

- First, bound δ:$\delta \leq \frac{\alpha}{4}$So:$\sqrt{\frac{\ln(2|H_d|/\beta)}{2n}} \leq \frac{\alpha}{4}$Which implies:$n \geq \frac{8\ln(2|H_d|/\beta)}{\alpha^2}$
- Next, ensure $\frac{2}{\varepsilon n} \leq \frac{\alpha}{2}$:$n \geq \frac{4}{\varepsilon\alpha}$

Thus, choose:$n \geq \max\left\{\frac{8\ln(2|H_d|/\beta)}{\alpha^2}, \frac{4}{\varepsilon\alpha}\right\}$

Conclusion:

- With this choice of n, have: $\Pr_{S\sim D}\left[\mathbb{E}_{\hat{h}\sim A_{\mathrm{DP}}(S)}[R(\hat{h};D)] \leq \alpha\right] \geq 1-\beta$
- The sample size n is polynomial in d, $1/\alpha$, $1/\varepsilon$, and $\ln(1/\beta)$, since $|H_d| \leq \exp(\mathrm{poly}(d))$.

Exercise 2

Objective: For any $d, p > 0$, given a finite input domain X_p with $|X_p| \leq \exp(p)$ and a hypothesis class H_d on X_p with VC dimension d, design a generic $(\alpha, \beta, \varepsilon)$-DP-PAC learner with sample size polynomial in $d, p, 1/\alpha, 1/\varepsilon, \ln(1/\beta)$.

1.Answer: Algorithm

Input:

- Dataset $S = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in X_p$ and $y_i \in \{0,1\}$.
- Hypothesis class H_d with VC dimension d.
- Parameters $\alpha > 0$, $\beta > 0$, $\varepsilon > 0$.

Procedure:

1. **Finite Hypothesis Set Construction:**

   Compute the Effective Size of H_d:

   - By the **Sauer-Shelah Lemma**, the number of distinct labelings (dichotomies) H_d can realize over X_p is bounded by:$|H_d| \leq \sum_{i=0}^{d} \binom{N}{i} \leq \left(\frac{eN}{d}\right)^d$where $N = |X_p| \leq \exp(p)$.

   Calculate the Bound:

   - $N = \exp(p)$.

- Therefore, $|H_d| \leq \left(\frac{e\exp(p)}{d}\right)^d = \exp(d(p+1-\ln d))$
- Thus, H_d is finite with size $|H_d| \leq \exp(\text{poly}(d,p))$.

2. **Compute Empirical Errors:**

   For each hypothesis $h \in H_d$:

   - Calculate the empirical error: $e(h,S) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{h(x_i) \neq y_i\}$
   - Define the score function: $q(S,h) = -e(h,S)$

3. **Apply Exponential Mechanism:**

   Sensitivity Calculation:

   - The sensitivity of $q(S,h)$ with respect to a single data point is $\Delta q = \frac{1}{n}$.

   Selection Probabilities:

   - For each $h \in H_d$: $\Pr[\hat{h} = h] \propto \exp\left(\frac{\varepsilon \cdot q(S,h)}{2\Delta q}\right) = \exp\left(-\frac{\varepsilon n e(h,S)}{2}\right)$

   Sampling:

   - Randomly select $\hat{h} \in H_d$ according to the above probabilities.

4. **Output:**

   Return the hypothesis $\hat{h}$.

2.Answer: Privacy Proof

The algorithm employs the **Exponential Mechanism** with privacy parameter ε and sensitivity $\Delta q = \frac{1}{n}$.

Sensitivity Verification:

- For any two neighboring datasets S and S' differing by one example: $|q(S,h) - q(S',h)| = |-e(h,S) + e(h,S')| \leq \frac{1}{n}$

Differential Privacy Guarantee:

- The Exponential Mechanism ensures that for all $h \in H_d$: $\frac{\Pr[\hat{h}=h|S]}{\Pr[\hat{h}=h|S']} \leq \exp\left(\frac{\varepsilon|q(S,h)-q(S',h)|}{2\Delta q}\right) \leq \exp(\varepsilon)$
- Therefore, **Algorithm** $\mathbf{A}_{\text{DP}}^{\alpha,\beta,\varepsilon}$ is ε-differentially private with respect to S.

3.Answer: Utility Proof

To show that:

$$\Pr_{S\sim D}\left[\mathbb{E}_{\hat{h}\sim A_{\mathrm{DP}}(S)}[R(\hat{h}; D)] \leq \alpha\right] \geq 1-\beta$$

where $R(\hat{h}; D)$ is the true error of $\hat{h}$ on distribution D.

Step 1: Empirical Error Concentration

Goal: Show that with high probability over $S \sim D$, for all $h \in H_d$, the empirical error $e(h, S)$ is close to the true error $R(h; D)$.

Proof:

- For any fixed $h \in H_d$, $e(h, S)$ is the empirical estimate of $R(h; D)$ over n i.i.d. samples.
- By Hoeffding's inequality: $\Pr_{S\sim D}[|e(h,S) - R(h; D)| \geq \delta] \leq 2\exp(-2n\delta^2)$
- Apply a union bound over all $h \in H_d$: $\Pr_{S\sim D}[\exists h \in H_d: |e(h,S) - R(h; D)| \geq \delta] \leq 2|H_d|\exp(-2n\delta^2)$
- Choose $\delta = \sqrt{\frac{\ln(2|H_d|/\beta)}{2n}}$ to ensure: $2|H_d|\exp(-2n\delta^2) = \beta$
- Since $|H_d| \leq \exp(\mathrm{poly}(d, p))$, we have: $\delta = \sqrt{\frac{\mathrm{poly}(d,p)+\ln(1/\beta)}{2n}}$
- Therefore, with probability at least $1-\beta$ over $S \sim D$, for all $h \in H_d$: $|e(h,S) - R(h; D)| \leq \delta$

Step 2: Expected Error Bound via Exponential Mechanism

Goal: Show that, conditional on the high-probability event from Step 1, the expected true error $\mathbb{E}_{\hat{h}|S}[R(\hat{h}; D)]$ is at most α.

Proof:

- **Conditional on S:** The expected true error is:

$$\mathbb{E}_{\hat{h}|S}[R(\hat{h}; D)] = \sum_{h\in H_d} \Pr[\hat{h} = h \mid S] \cdot R(h; D)$$

- **Using the Exponential Mechanism:**

$$\Pr[\hat{h} = h \mid S] = \frac{\exp\left(-\frac{\varepsilon n e(h,S)}{2}\right)}{Z}$$

where $Z = \sum_{h'\in H_d} \exp\left(-\frac{\varepsilon n e(h',S)}{2}\right)$.

- **Bounding $R(h; D)$:**

From Step 1, for all $h \in H_d$:$R(h;D) \leq e(h,S) + \delta$

- **Expected True Error:**

$$\mathbb{E}_{\hat{h}|S}[R(\hat{h};D)] \leq \sum_{h \in H_d} \Pr[\hat{h} = h \mid S] \cdot (e(h,S) + \delta) = \mathbb{E}_{\hat{h}|S}[e(\hat{h},S)] + \delta$$

- **Computing** $\mathbb{E}_{\hat{h}|S}[e(\hat{h},S)]$**:**

$$\mathbb{E}_{\hat{h}|S}[e(\hat{h},S)] = \frac{1}{Z} \sum_{h \in H_d} e(h,S) \cdot \exp\left(-\frac{\varepsilon n e(h,S)}{2}\right)$$

- **Bounding** $\mathbb{E}_{\hat{h}|S}[e(\hat{h},S)]$**:**

The function $f(u) = u\exp\left(-\frac{\varepsilon n u}{2}\right)$ attains its maximum at $u = \frac{2}{\varepsilon n}$.

Therefore, the expected empirical error is bounded by:$\mathbb{E}_{\hat{h}|S}[e(\hat{h},S)] \leq \frac{2}{\varepsilon n}$

- **Total Expected True Error:**

$$\mathbb{E}_{\hat{h}|S}[R(\hat{h};D)] \leq \frac{2}{\varepsilon n} + \delta$$

- **Ensuring** $\mathbb{E}_{\hat{h}|S}[R(\hat{h};D)] \leq \alpha$**:**

To guarantee this, choose n such that:$\frac{2}{\varepsilon n} + \delta \leq \alpha$

Substitute $\delta = \sqrt{\frac{\mathrm{poly}(d,p) + \ln(1/\beta)}{2n}}$.

Solve for n to satisfy:$\frac{2}{\varepsilon n} + \sqrt{\frac{\mathrm{poly}(d,p) + \ln(1/\beta)}{2n}} \leq \alpha$

This inequality can be satisfied with n polynomial in $d, p, 1/\alpha, 1/\varepsilon, \ln(1/\beta)$.

Conclusion:

- With the chosen n, with probability at least $1 - \beta$ over $S \sim D$, the expected true error satisfies:$\mathbb{E}_{\hat{h}|S}[R(\hat{h};D)] \leq \alpha$
- Therefore: $\Pr_{S \sim D}\left[\mathbb{E}_{\hat{h} \sim A_{\mathrm{DP}}(S)}[R(\hat{h};D)] \leq \alpha\right] \geq 1 - \beta$

Sample Size Calculation

To make the inequality $\frac{2}{\varepsilon n} + \delta \leq \alpha$ hold, proceed as follows:

Bound δ**:**

$$\delta \leq \frac{\alpha}{2}$$

So:

$$\sqrt{\frac{\text{poly}(d,p) + \ln(1/\beta)}{2n}} \leq \frac{\alpha}{2}$$

Solving for n:

$$n \geq \frac{2(\text{poly}(d,p) + \ln(1/\beta))}{\alpha^2}$$

Ensure $\frac{2}{\varepsilon n} \leq \frac{\alpha}{2}$**:**

$$n \geq \frac{4}{\varepsilon\alpha}$$

Combined Sample Size n**:**

$$n \geq \max\left\{\frac{2(\text{poly}(d,p) + \ln(1/\beta))}{\alpha^2}, \frac{4}{\varepsilon\alpha}\right\}$$

Since $\text{poly}(d,p)$ denotes polynomial functions of d and p, n is polynomial in $d, p, 1/\alpha, 1/\varepsilon, \ln(1/\beta)$.

Exercise 3

Answer

Differentially private PAC learning requires more data than non-private PAC learning because the sample complexity increases with $1/\varepsilon$, the privacy parameter, adding to the dependence on VC dimension and accuracy parameters. Computationally, private learners are often less efficient; the algorithms designed for differential privacy are not necessarily computationally practical, whereas non-private PAC learners typically use efficient algorithms like empirical risk minimization. Thus, privacy introduces both statistical overhead—increased sample size—and computational challenges—less efficient algorithms.