Model-Agnostic Private Learning

Raef Bassily *The Ohio State University*

Om Thakkar *Boston University*

Abhradeep Thakurta *UC Santa Cruz*

DP Learning: Standard Model

Private Training set

$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

features vector

label

Learning Alg.

for a class of models $\mathcal{H}$

(ϵ, δ) - DP Algorithm

A model h_S :

a map $x \mapsto y$

DP Learning: Standard Model

Private Training set

S = $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

features vector

label

Learning Alg.

for a class of models

$\mathcal{H}$

(ϵ, δ) - DP Algorithm

A model h_S :

a map $x \mapsto y$

Main issues with this approach:

- Requires white-box modification of standard *non-private* learners.
- Often requires some knowledge about structure of $\mathcal{H}$.
- Sometimes, yields error with necessary dependence on dimensions or size of $\mathcal{H}$ even for simple classes, *e.g., learning thresholds [Bun et al. 2015].*

DP Learning: Standard Model

Private
Training set

$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

features
vector

label

Learning Alg.
for a class of models
$\mathcal{H}$

(ϵ, δ) - DP Algorithm

A model h_S :

a map $x \mapsto y$

Main issues with this approach:

- Requires white-box modification of standard *non-private* learners.
- Often requires some knowledge about structure of $\mathcal{H}$.
- Sometimes, yields error with necessa
  size of $\mathcal{H}$ even for simple classes, *e.g*

Become more challenging with the rise of modern over-parameterized machine learning.

DP Learning: Alternative Approach

Private
Training set

$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

(ϵ,δ)-DP Algorithm

Classification queries

$\{\tilde{x}_1, \ldots, \tilde{x}_m\}$

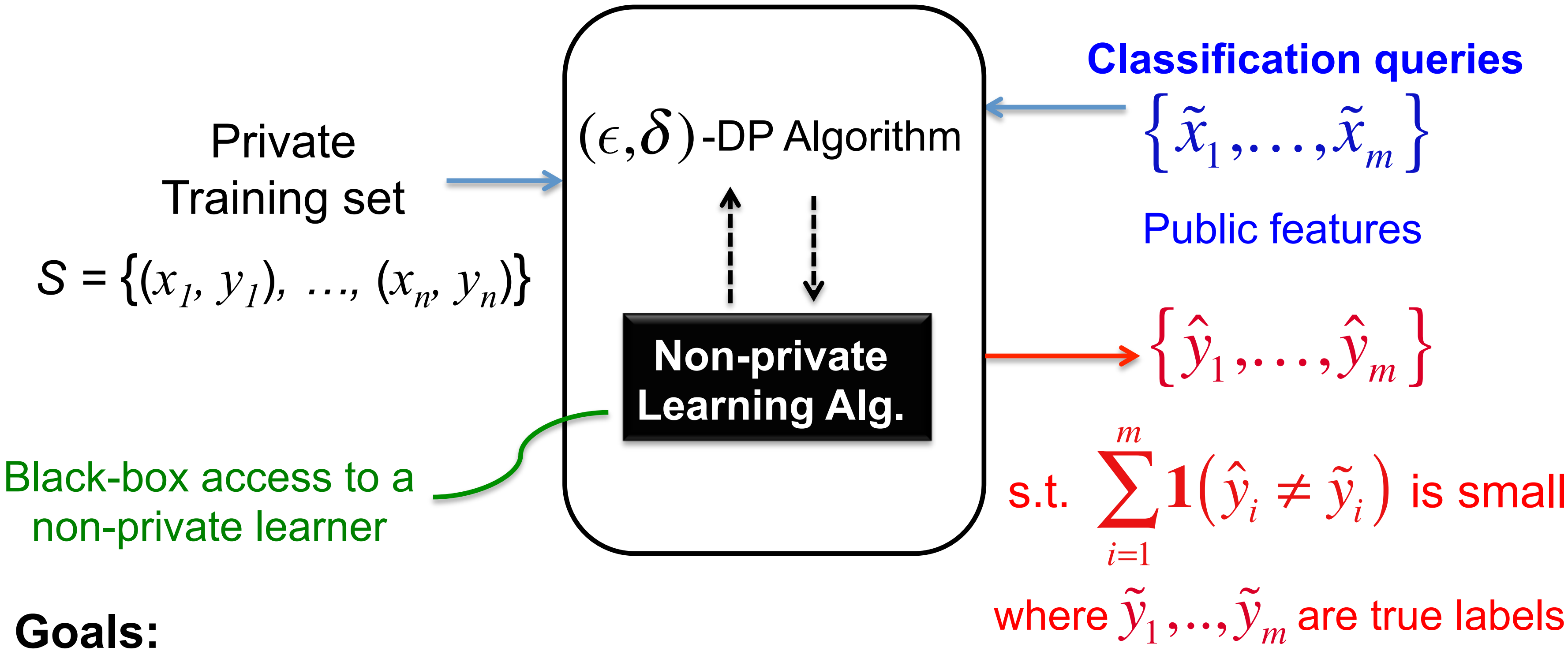
Public features

$\{\hat{y}_1, \ldots, \hat{y}_m\}$

s.t. $\sum_{i=1}^{m} \mathbf{1}(\hat{y}_i \neq \tilde{y}_i)$ is small

where $\tilde{y}_1, \ldots, \tilde{y}_m$ are true labels.

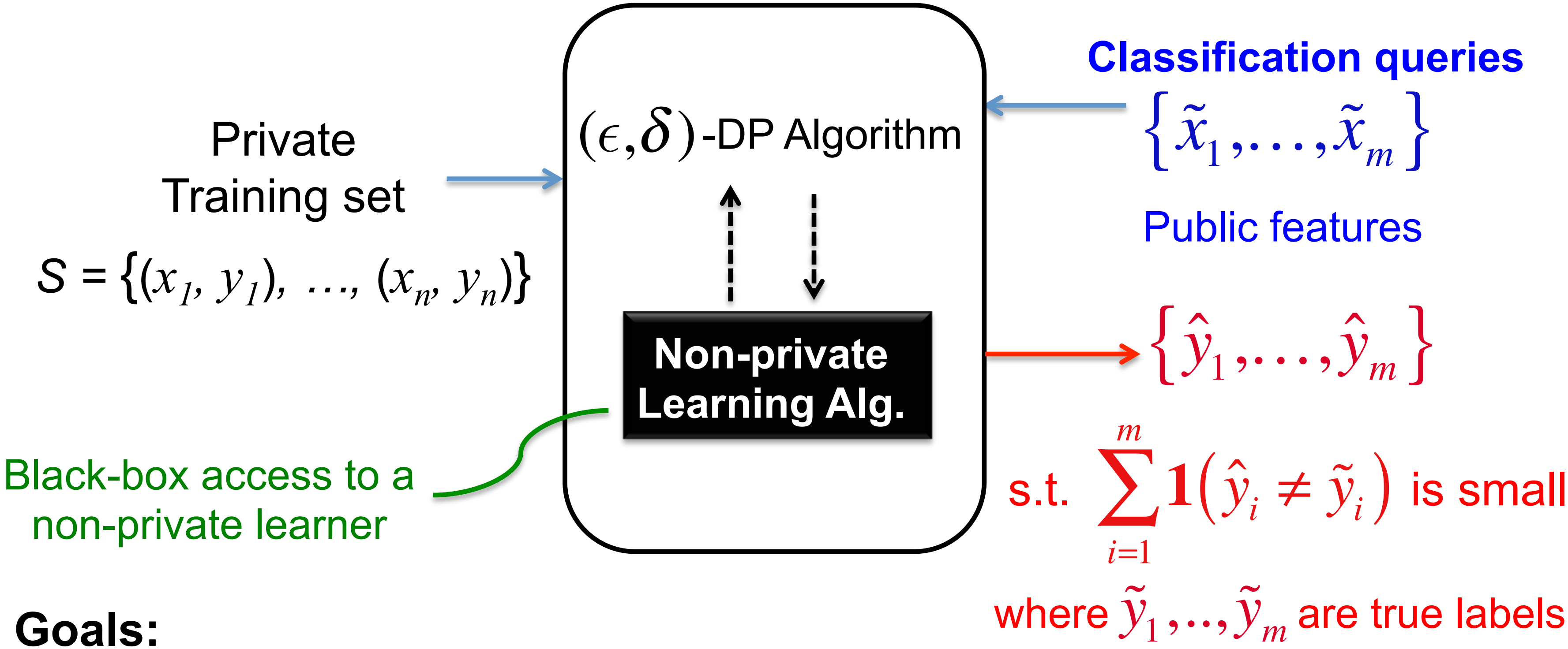
DP Learning: *Alternative* Approach



Goals:

- *Black-box* use of any non-private learner.

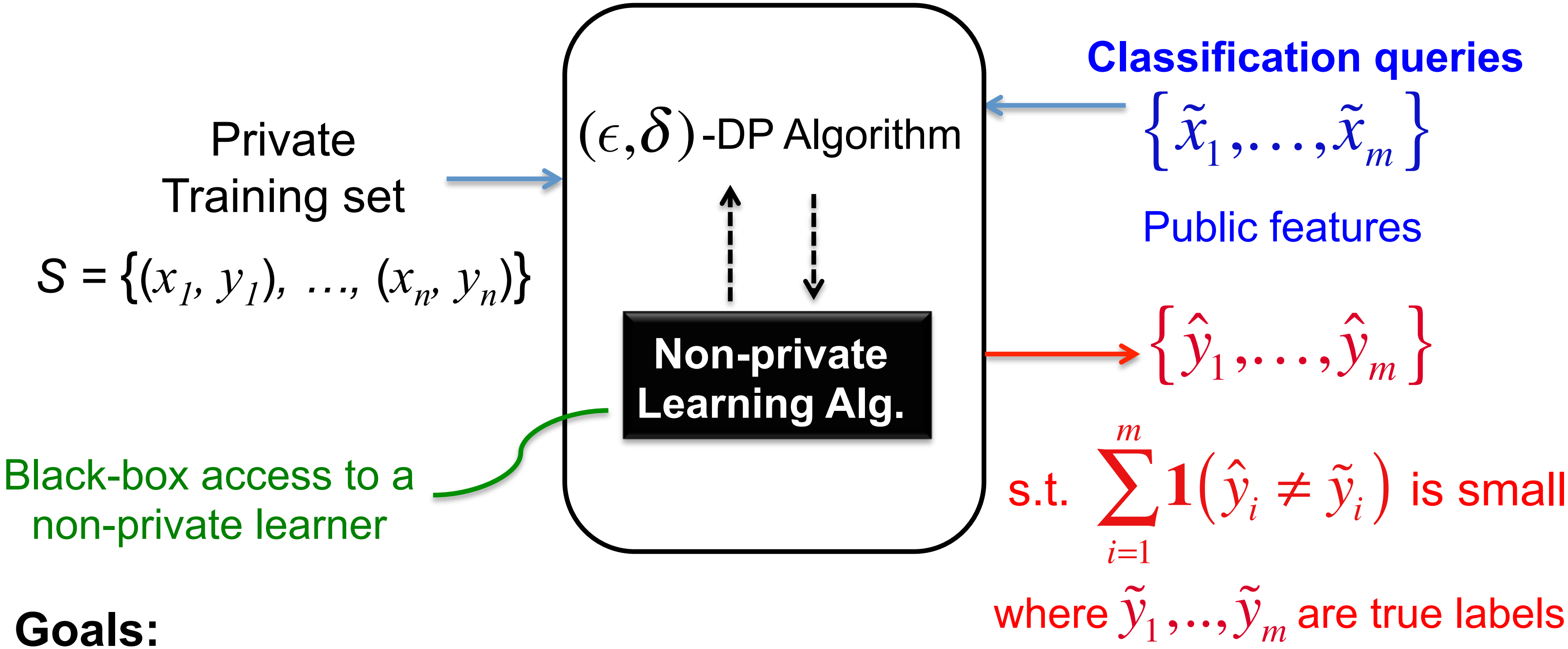
DP Learning: Alternative Approach



Goals:

- *Black-box* use of any non-private learner.
- *Answer lots of queries: conservative use of the privacy budget.*

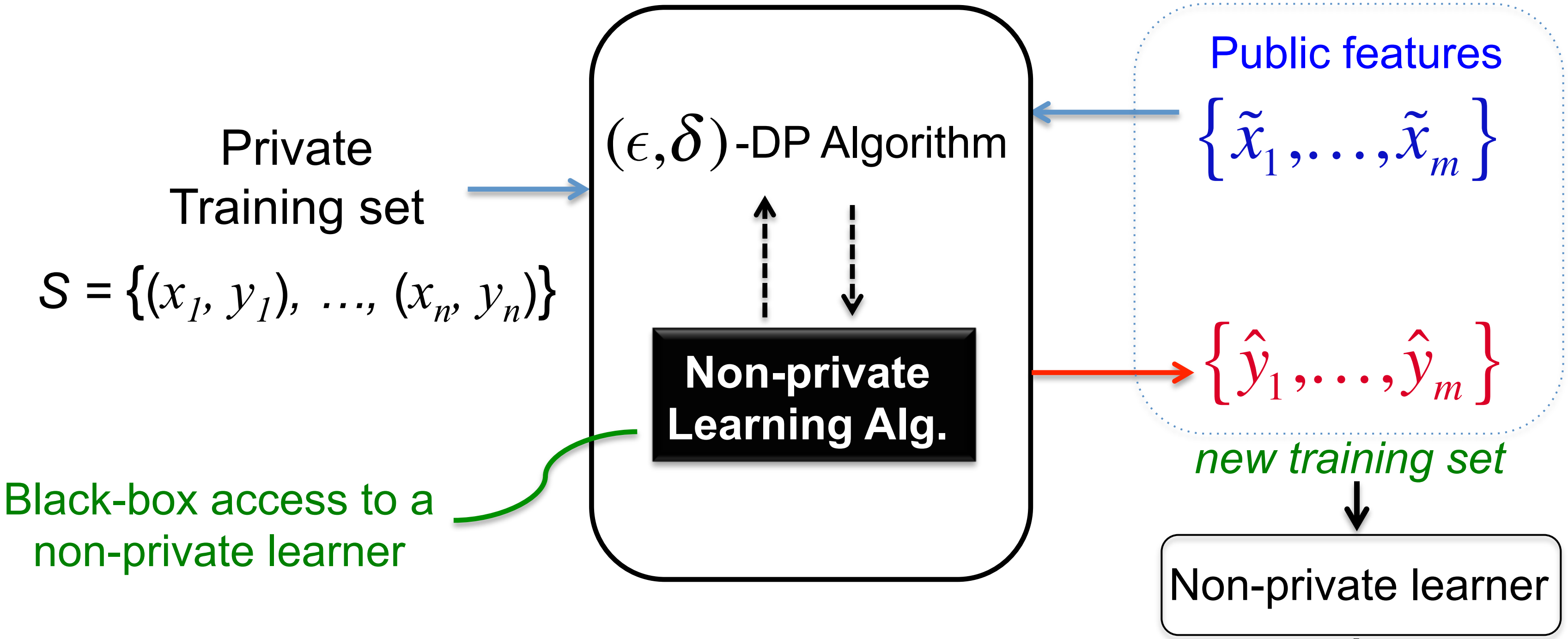
DP Learning: Alternative Approach



Goals:

- *Black-box* use of any non-private learner.
- *Answer lots of queries: conservative use of the privacy budget.*
- *Transferrable guarantees*: non-private accuracy → private accuracy.

DP Learning: Alternative Approach



Goals:

- *Black-box* use of any non-private learner.
- *Answer lots of queries: conservative use of the privacy budget.*
- *Transferrable guarantees*: non-private accuracy → private accuracy.
- *Knowledge transfer*: public features + private labels used to train a final private classifier.

Related Work

- Knowledge transfer based on aggregated classifiers ensemble dates back to [Breiman'94].

Related Work

- Knowledge transfer based on aggregated classifiers ensemble dates back to [Breiman'94].
- In privacy literature, general subsample-and-aggregate framework was introduced in [NRS'07].

Related Work

- Knowledge transfer based on aggregated classifiers ensemble dates back to [Breiman'94].
- In privacy literature, general subsample-and-aggregate framework was introduced in [NRS'07].
- Subsample-and-aggregate for label prediction [Bilenko-Dwork-Muthukrishnan-Rothblum-Thakurta-Wang'12]

Related Work

- Knowledge transfer based on aggregated classifiers ensemble dates back to [Breiman’94].
- In privacy literature, general subsample-and-aggregate framework was introduced in [NRS’07].
- Subsample-and-aggregate for label prediction [Bilenko-Dwork-Muthukrishnan-Rothblum-Thakurta-Wang’12]
- Knowledge transfer for private classification was first explored in [Hamm-Cao-Belkin’16]:
  - White-box construction with weaker guarantees.

Related Work

- Knowledge transfer based on aggregated classifiers ensemble dates back to [Breiman'94].
- In privacy literature, general subsample-and-aggregate framework was introduced in [NRS'07].
- Subsample-and-aggregate for label prediction [Bilenko-Dwork-Muthukrishnan-Rothblum-Thakurta-Wang'12]
- Knowledge transfer for private classification was first explored in [Hamm-Cao-Belkin'16]:
  - White-box construction with weaker guarantees.
- Better constructions were given in [Papernot et al.'17, Papernot et al'18] (*PATE* framework), but without formal accuracy guarantees.
  - [Papernot et al.'18]: report-noisy-max + sparse-vector

Related Work

- Knowledge transfer based on aggregated classifiers ensemble dates back to [Breiman'94].
- In privacy literature, general subsample-and-aggregate framework was introduced in [NRS'07].
- Subsample-and-aggregate for label prediction [Bilenko-Dwork-Muthukrishnan-Rothblum-Thakurta-Wang'12]
- Knowledge transfer for private classification was first explored in [Hamm-Cao-Belkin'16]:
  - White-box construction with weaker guarantees.
- Better constructions were given in [Papernot et al.'17, Papernot et al'18] (*PATE* framework), but without formal accuracy guarantees.
  - [Papernot et al.'18]: report-noisy-max + sparse-vector
- Very recently, [Dwork-Feldman'18] considers the problem of private prediction (focuses on *the single-query case*):
  - Different constructions, more general settings

Results

1. A new general paradigm for answering *“stable” queries*:
   - Based on a new approach combining *distance-to-instability* *[Smith-Thakurta’13]* with *sparse-vector [DNRRV’09, DR’14]* techniques.

Results

1. A new general paradigm for answering *“stable” queries*:
   - Based on a new approach combining *distance-to-instability* *[Smith-Thakurta’13]* with *sparse-vector [DNRRV’09, DR’14]* techniques.
2. New construction for *privately answering classification queries*:
   - Bounds on misclassification rate in the standard PAC model: *better than what is implied by advanced composition*.

Results

1. A new general paradigm for answering *“stable” queries*:
   - Based on a new approach combining *distance-to-instability* *[Smith-Thakurta’13]* with *sparse-vector [DNRRV’09, DR’14]* techniques.
2. New construction for *privately answering classification queries*:
   - Bounds on misclassification rate in the standard PAC model: *better than what is implied by advanced composition*.
3. A black-box construction for a private learner via knowledge transfer with rigorous guarantees
   - Sample complexity bounds in terms of VC-dimension.

Results

1. A new general paradigm for answering *"stable" queries*:
   - Based on a new approach combining *distance-to-instability* *[Smith-Thakurta'13]* with *sparse-vector [DNRRV'09, DR'14]* techniques.
2. New construction for *privately answering classification queries*:
   - Bounds on misclassification rate in the standard PAC model: *better than what is implied by advanced composition*.
3. A black-box construction for a private learner via knowledge transfer with rigorous guarantees
   - Sample complexity bounds in terms of VC-dimension.
4. Extension: construction for privately answering soft-label queries

Generic paradigm for answering stable queries

A stable query is a function (of the dataset) whose outcome does not change unless we change a "relatively large" number of points in the dataset.

Generic paradigm for answering stable queries

A stable query is a function (of the dataset) whose outcome does not change unless we change a “relatively large” number of points in the dataset.

Any *good learner* can be used in natural way to achieve this notion, e.g., via *aggregating ensemble of classifiers* (Bagging [Breiman’94]).

Generic paradigm for answering stable queries

A stable query is a function (of the dataset) whose outcome does not change unless we change a "relatively large" number of points in the dataset.

Any *good learner* can be used in natural way to achieve this notion, e.g., via *aggregating ensemble of classifiers* (Bagging [Breiman'94]).

Idea: Combining *distance-to-instability* and *sparse-vector* techniques:

- Distance-to-instability [ST'13] exploits stability to produce *noiseless outputs for stable queries.*
- Sparse-vector [DNRRV'09, DR14] enables us to pay a privacy cost only for unstable queries → *efficient use of privacy budget* → *answer more queries than what advanced composition suggests.*

Privately answering classification queries

Inputs:

- Private training set $S \in \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d.
- Queries $\tilde{x}_1, .., \tilde{x}_m$: public feature-vectors i.i.d. (from same distribution).
- Privacy parameters ϵ, δ
- Cut-off T : number of *unstable* queries we allow before terminating.

Privately answering classification queries

Inputs:

- Private training set $S \in \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d.
- Queries $\tilde{x}_1, .., \tilde{x}_m$: public feature-vectors i.i.d. (from same distribution).
- Privacy parameters ϵ, δ
- Cut-off T : number of *unstable* queries we allow before terminating.

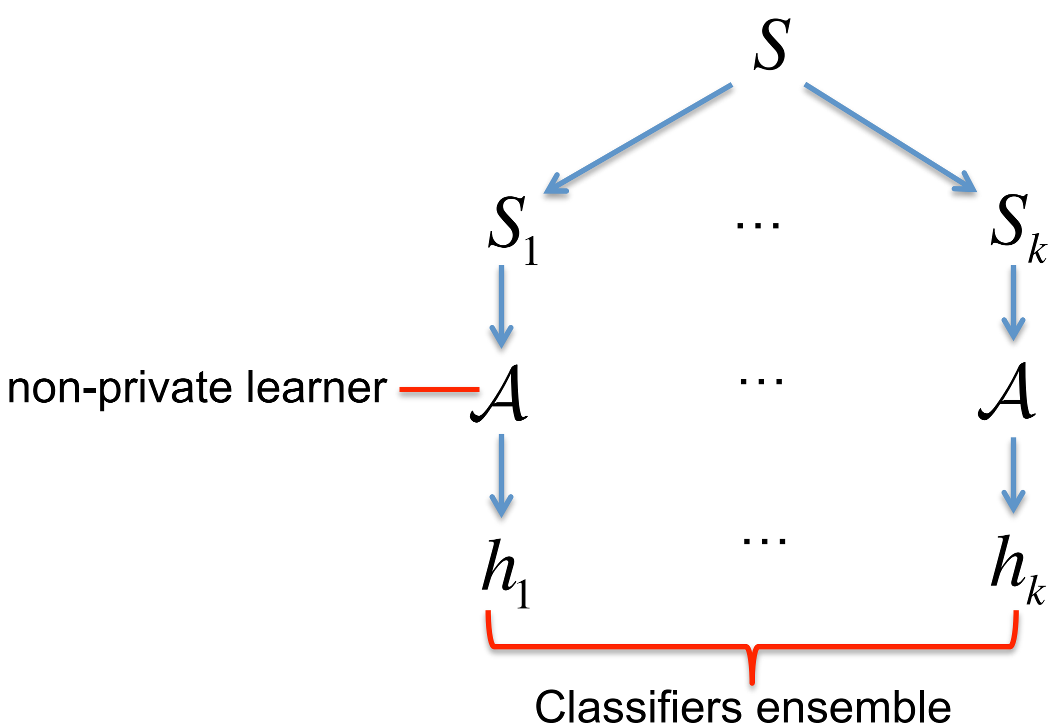
0) Initialize counter for unstable queries: *counter = 0.*

Privately answering classification queries

Inputs:

- Private training set $S \in \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d.
- Queries $\tilde{x}_1, .., \tilde{x}_m$: public feature-vectors i.i.d. (from same distribution).
- Privacy parameters ϵ, δ
- Cut-off T : number of *unstable* queries we allow before terminating.

1) Split S into k chunks; each used to train a non-private learner $\mathcal{A}$



Privately answering classification queries

Inputs:

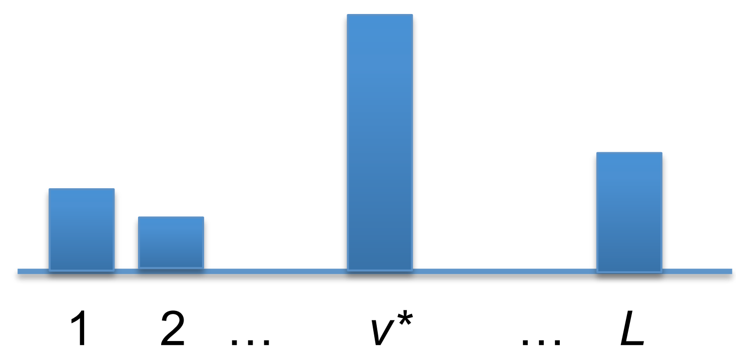
- Private training set $S \in \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d.
- Queries $\tilde{x}_1, .., \tilde{x}_m$: public feature-vectors i.i.d. (from same distribution).
- Privacy parameters ϵ, δ
- Cut-off T : number of *unstable* queries we allow before terminating.

2) For each query $\tilde{x}_j$, construct histogram of the votes $h_1(\tilde{x}_j), \ldots, h_k(\tilde{x}_j)$

Classification query

$$\tilde{x}_j, \ j = 1, \ldots, m \longrightarrow h_1 \ \ldots \ h_k$$

$$h_1(\tilde{x}_j) \cdots h_k(\tilde{x}_j) \longrightarrow$$



for L-class problem

Privately answering classification queries

Inputs:

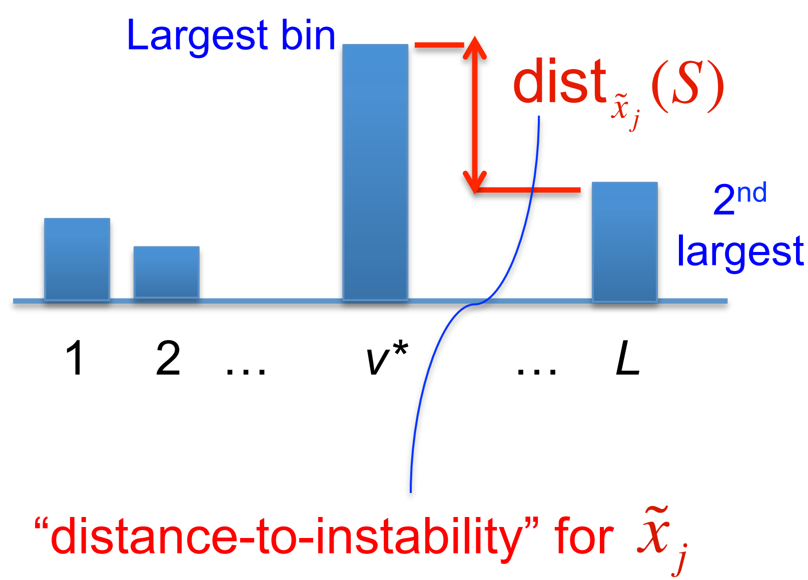
- Private training set $S \in \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d.
- Queries $\tilde{x}_1, .., \tilde{x}_m$: public feature-vectors i.i.d. (from same distribution).
- Privacy parameters $\epsilon,\ \delta$
- Cut-off T : number of *unstable* queries we allow before terminating.

2) For each query $\tilde{x}_j$, construct histogram of the votes $h_1(\tilde{x}_j), \ldots, h_k(\tilde{x}_j)$

Classification query

$\tilde{x}_j,\ \ j = 1, \ldots, m \longrightarrow h_1 \ \ldots \ h_k$

$h_1(\tilde{x}_j) \cdots h_k(\tilde{x}_j) \longrightarrow$



Privately answering classification queries

Inputs:

- Private training set $S \in \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d.
- Queries $\tilde{x}_1, .., \tilde{x}_m$: public feature-vectors i.i.d. (from same distribution).
- Privacy parameters $\epsilon, \ \delta$
- Cut-off T : number of *unstable* queries we allow before terminating.

2) For each query $\tilde{x}_j$, construct histogram of the votes $h_1(\tilde{x}_j), \ldots, h_k(\tilde{x}_j)$

Classification query

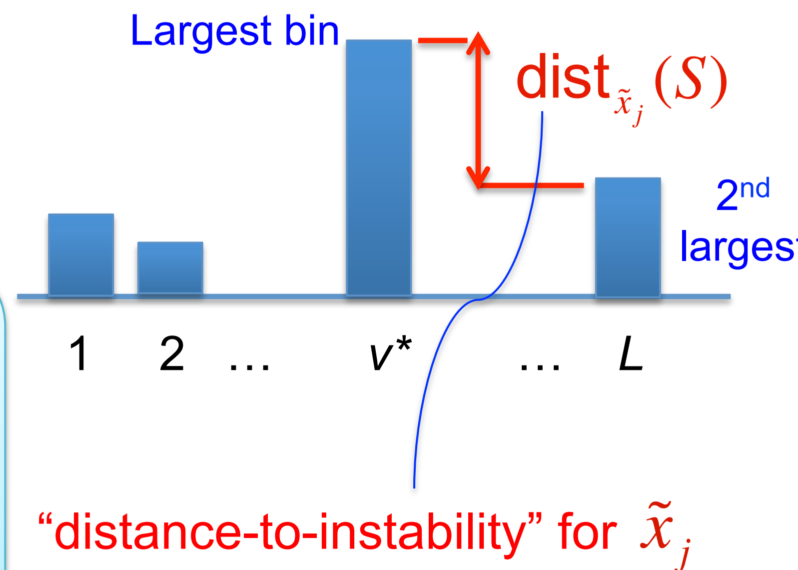
$\tilde{x}_j, \ \ j = 1, \ldots, m \longrightarrow h_1 \ \ldots \ h_k$

We say $\tilde{x}_j$ is stable query if

$\text{dist}_{\tilde{x}_j}(S)$ is sufficiently large

i.e., $>$ *some fixed threshold*

$Thres \approx \sqrt{T \log(1/\delta)} \log(m/\delta)/\epsilon$



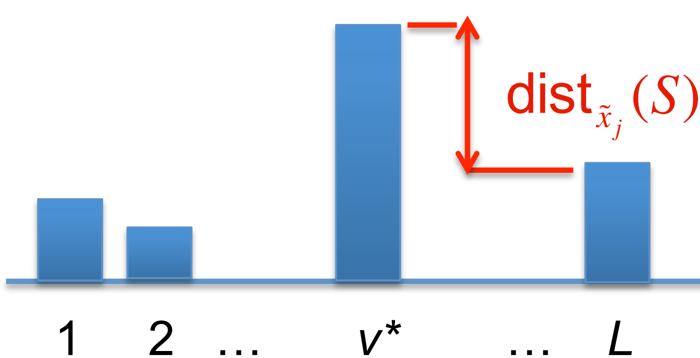
Privately answering classification queries

Inputs:

- Private training set $S \in \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d.
- Queries $\tilde{x}_1, \ldots, \tilde{x}_m$: public feature-vectors i.i.d. (from same distribution).
- Privacy parameters $\epsilon,\ \delta$
- Cut-off T : number of *unstable* queries we allow before terminating.

3) Private stability test: $\widehat{\text{dist}}_{\tilde{x}_j}(S) > \widehat{Thres}$?

$$\text{dist}_{\tilde{x}_j}(S) + \text{Lap}(2/\epsilon') \qquad \widehat{Thres} = Thres + \text{Lap}(1/\epsilon')$$



$$Thres \approx \log(m/\delta)/\epsilon'$$

$$\epsilon' \approx \epsilon / \sqrt{T \log(1/\delta)}$$

Privately answering classification queries

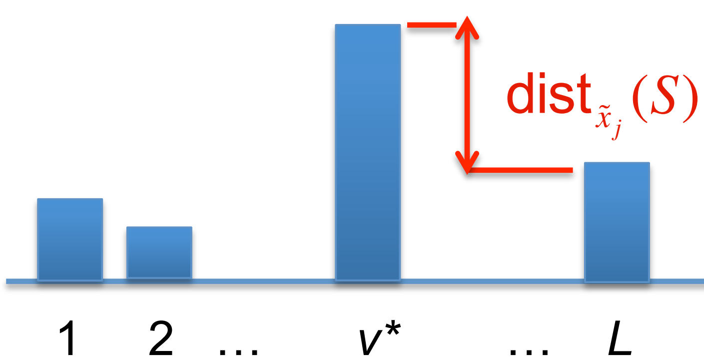
Inputs:

- Private training set $S \in \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d.
- Queries $\tilde{x}_1, .., \tilde{x}_m$: public feature-vectors i.i.d. (from same distribution).
- Privacy parameters $\epsilon, \ \delta$
- Cut-off T : number of *unstable* queries we allow before terminating.

3) Private stability test: $\widehat{\mathrm{dist}_{\tilde{x}_j}(S)} > \widehat{Thres}$?

YES

- Output *v* = label with largest # of votes.*
- *Go to next query.*



$Thres \approx \log(m / \delta) / \epsilon'$

$\epsilon' \approx \epsilon / \sqrt{T \log(1/\delta)}$

Privately answering classification queries

Inputs:

- Private training set $S \in \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn i.i.d.
- Queries $\tilde{x}_1, .., \tilde{x}_m$: public feature-vectors i.i.d. (from same distribution).
- Privacy parameters $\epsilon, \ \delta$
- Cut-off T : number of *unstable* queries we allow before terminating.

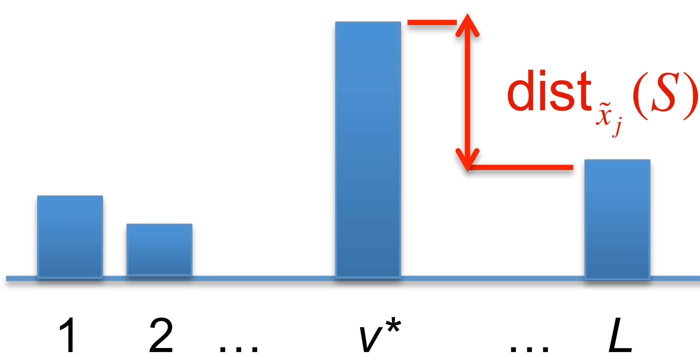
3) Private stability test: $\widehat{\mathrm{dist}_{\tilde{x}_j}}(S) > \widehat{Thres}$?

YES

- Output *v* = label with largest # of votes.*
- *Go to next query.*

NO

- Output $\perp$
- *counter = counter + 1*
- *If counter > T, then Abort.*
- *Go to next query.*



$Thres \approx \log(m / \delta) / \epsilon'$

$\epsilon' \approx \epsilon / \sqrt{T \log(1 / \delta)}$

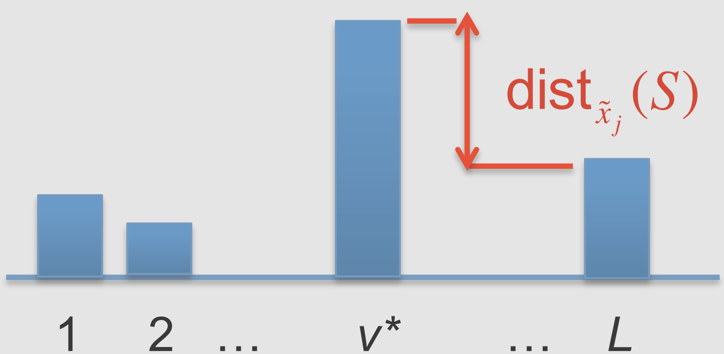
Privately answering classification queries

Theorem: This algorithm is (ϵ, δ)-DP.

Proof idea: The construction can be viewed as a composition of a $(\epsilon, \delta/2)$-DP sparse-vector algorithm [DR'14] and a $(0, \delta/2)$-DP distance-to-instability algorithm [ST'13].

- *Output v* = label with largest # of votes.*
- *Go to next query.*

- Output $\perp$
- *counter = counter + 1*
- *If counter > T, then Abort.*



$$Thres \approx \log(m/\delta)/\epsilon'$$

$$\epsilon' \approx \epsilon / \sqrt{T \log(1/\delta)}$$

Privately answering classification queries

Accuracy depends on how *accurate* and *consistent* are the predictions of the classifiers ensemble $h_1(\tilde{x}_j),\ldots,h_k(\tilde{x}_j)$ for each query $\tilde{x}_j$

- Output *v* = label with largest # of votes.*
- *Go to next query.*

- Output $\perp$
- *counter = counter + 1*
- *If counter > T, then Abort.*

$$Thres \approx \log(m/\delta)/\epsilon'$$

$$\epsilon' \approx \epsilon / \sqrt{T \log(1/\delta)}$$

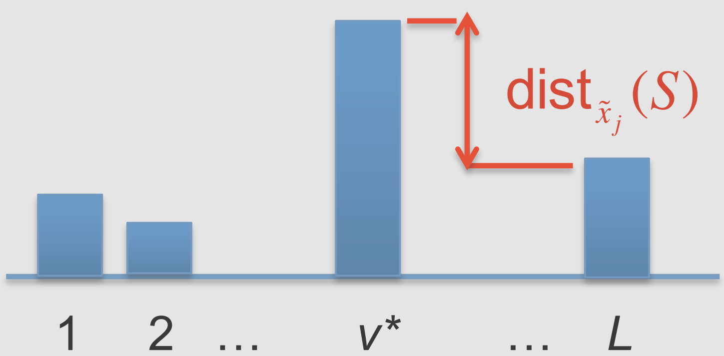
Privately answering classification queries

Accuracy depends on how *accurate* and *consistent* are the predictions of the classifiers ensemble $h_1(\tilde{x}_j),\ldots,h_k(\tilde{x}_j)$ for each query $\tilde{x}_j$

Intuition: If $\mathcal{A}$ is a *good* non-private learner, then *most of the ensemble predictions will agree* (consistency) *on the correct label* (accuracy).

- Output *v* = label with largest # of votes.*
- *Go to next query.*

- Output $\perp$
- *counter = counter + 1*
- *If counter > T, then Abort.*



$Thres \approx \log(m/\delta)/\epsilon'$

$\epsilon' \approx \epsilon/\sqrt{T\log(1/\delta)}$

Privately answering classification queries

Analysis of misclassification rate (*binary labels case*)

Idea: If each of $h_1, \ldots, h_k$ has *classification error* α,

$$\mathbb{E}_{x,y}\left[\mathbf{1}(h_\ell(x) \neq y)\right] \leq \alpha, \;\; \forall \ell \in [k]$$

Privately answering classification queries

Analysis of misclassification rate (*binary labels case*)

Idea: If each of $h_1,\ldots,h_k$ has *classification error* α,

$$\mathbb{E}_{x,y}\left[\mathbf{1}(h_\ell(x)\neq y)\right]\leq\alpha,\ \ \forall\ell\in[k]$$

then except for at most $\approx 3m\alpha$ queries, at least $2k/3$ classifiers will *agree on the correct label.*

\# of ✗ in each row $\approx m\alpha$

Total # of ✗ $\approx km\alpha$

\# of columns w/ more than $\approx k/3$ ✗ is $< 3m\alpha$

	$\tilde{x}_1$	$\tilde{x}_2$			$\tilde{x}_m$
h_1	✔	✗	✔ ✔ ✗ ⋯ ✔ ✔	✗	✔
h_2	✔	✔	✗ ✔ ✔ ⋯ ✗ ✔	✔	✔
	.	.	… … …	.	.
	.	.	… … …	.	.
h_k	✗	✔	✔ ✗ ✔ ⋯ ✔ ✗	✗	✔

Privately answering classification queries

Analysis of misclassification rate (*binary labels case*)

Idea: If each of $h_1,\ldots,h_k$ has *classification error* α,

$$\mathbb{E}_{x,y}\left[\mathbf{1}(h_\ell(x) \neq y)\right] \leq \alpha, \quad \forall \ell \in [k]$$

then except for at most $\approx 3m\alpha$ queries, at least $2k/3$ classifiers will *agree on the correct label.*

$$\Downarrow$$

Setting $T \approx 3m\alpha$ and $k \approx Thres \approx \sqrt{T}/\epsilon$, then our construction yields a misclassification rate $T/m \approx 3\alpha$

Privately answering classification queries

Hence, we can give the following guarantees in the ***standard PAC model****.*

Setup:

- Training set (of size n) and queries set (of size m) are i.i.d.
- True labels generated by a hypothesis from a class $\mathcal{H}$ of VC-dim V.

Privately answering classification queries

Hence, we can give the following guarantees in the ***standard PAC model****.*

Setup:

- Training set (of size n) and queries set (of size m) are i.i.d.
- True labels generated by a hypothesis from a class $\mathcal{H}$ of VC-dim V.

Let $\mathcal{A}$ *be any non-private PAC learner for* $\mathcal{H}$, then (*ignoring logs!*),

i) can privately answer up to m $\approx$ n/V binary classification queries with the *optimal non-private misclassification rate* $\approx V/n$ *(privacy for free).*

ii) Beyond n/V *queries,* our misclassification rate is $\approx m\, V^2/n^2$

Privately answering classification queries

Hence, we can give the following guarantees in the ***standard PAC model****.*

Setup:

- Training set (of size n) and queries set (of size m) are i.i.d.
- True labels generated by a hypothesis from a class $\mathcal{H}$ of VC-dim V.

Let $\mathcal{A}$ *be any non-private PAC learner for* $\mathcal{H}$, then (*ignoring logs!*),

i) can privately answer up to m $\approx$ n/V binary classification queries with the *optimal non-private misclassification rate* $\approx$ *V/n (privacy for free).*

ii) Beyond *n/V queries,* our misclassification rate is $\approx m V^2/n^2$

Standard advanced composition would have led to error $\approx \sqrt{m}\ V/n$ for all m.

Privately answering classification queries

Hence, we can give the following guarantees in the ***standard PAC model****.*

Setup:

- Training set (of size *n*) and queries set (of size *m*) are i.i.d.
- True labels generated by a hypothesis from a class $\mathcal{H}$ of VC-dim *V*.

Let $\mathcal{A}$ *be any non-private PAC learner for* $\mathcal{H}$, then (*ignoring logs!*),

i) can privately answer up to m $\approx$ *n/V* binary classification queries with the *optimal non-private misclassification rate* $\approx$ *V/n (privacy for free).*

ii) Beyond *n/V queries,* our misclassification rate is $\approx$ *m* V^2/n^2

Standard advanced composition would have led to error $\approx \sqrt{m}$ *V/n* for all m.

We also obtain analogous bounds for the agnostic setting.

Private learner via Knowledge Transfer

A black-box construction for a private learner (*outputs a classifier*) for any of the following settings:

- Training set is private but we can access public unlabeled data.
- Only the labels of the training set are considered private

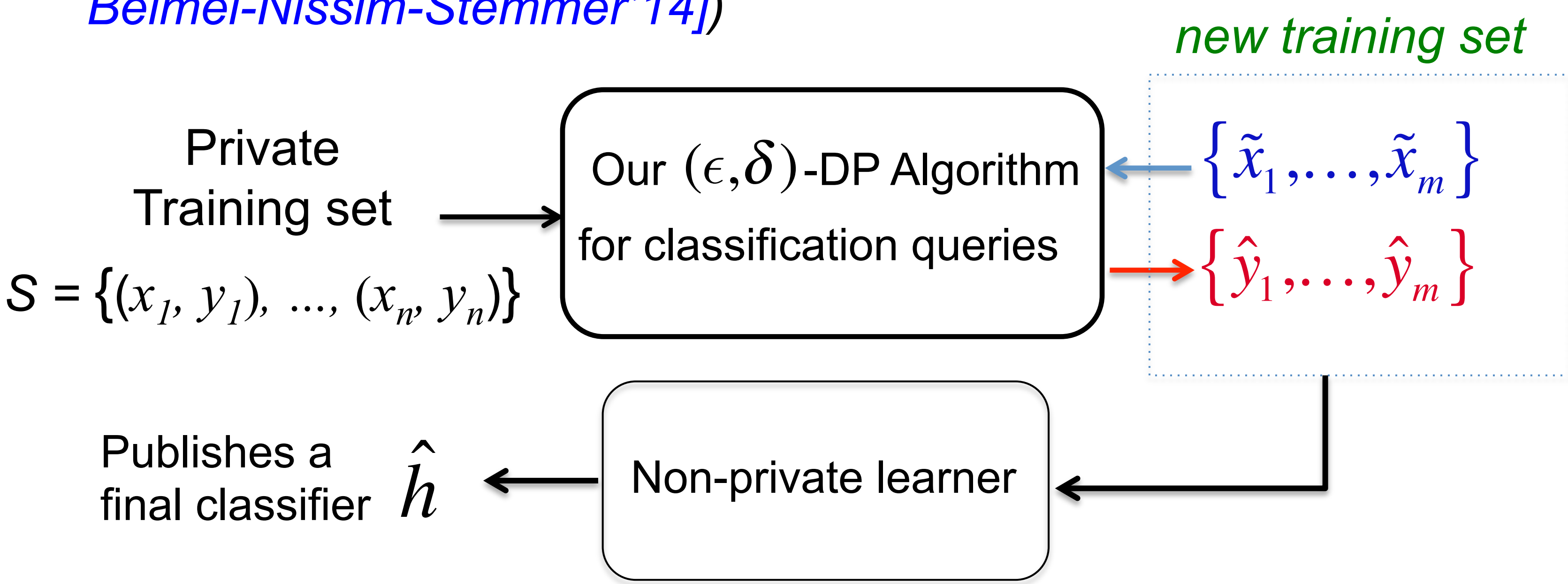
  (known as label-private learning [Chaudhuri-Hsu'11, Beimel-Nissim-Stemmer'14])

Private learner via Knowledge Transfer

A black-box construction for a private learner (*outputs a classifier*) for any of the following settings:

- Training set is private but we can access public unlabeled data.
- Only the labels of the training set are considered private

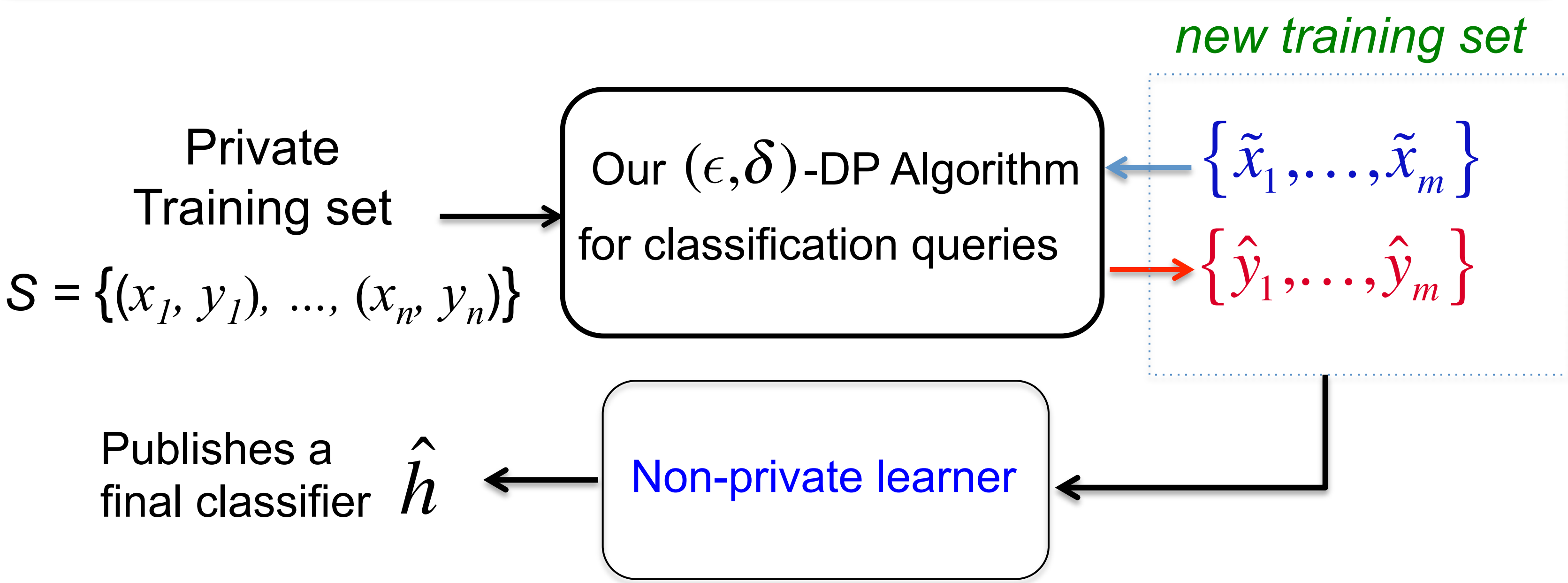
  (known as label-private learning [Chaudhuri-Hsu'11, Beimel-Nissim-Stemmer'14])



Private learner via Knowledge Transfer

- *This construction:*

  - is efficient as long as the non-private learner is efficient.
  - allows for transferring accuracy guarantees of the non-private learner to accuracy guarantees of the private learner.

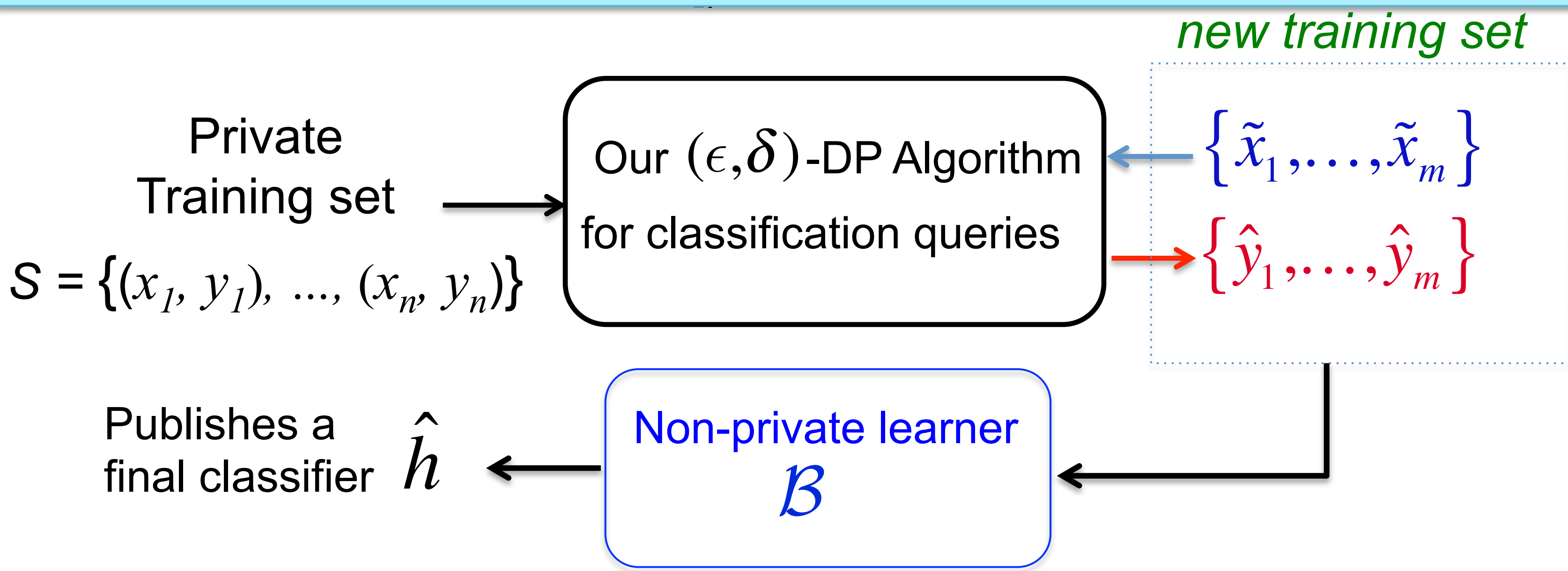


*Private learner via **Knowledge Transfer***

We obtain formal accuracy guarantees for the final learner.

Idea:

1) The labels of the *new* training set are generated by our previous algorithm.
2) We can bound the classification error for our previous algorithm.
3) A good non-private learner $\mathcal{B}$ will yield h whose classification error is close to the classification error in the *new* training set.



Private learner via ***Knowledge Transfer***

Let $\mathcal{H}$ be of VC-dim V. Let $\mathcal{B}$ be an agnostic PAC learner for $\mathcal{H}$.

For any $\alpha > 0$, let $m = \tilde{O}\left(V/\alpha^2\right)$.

- *Realizable case:* if $n = \tilde{O}\left(V^{3/2}/\alpha^{3/2}\right)$, then w.h.p. the output $\hat{h} \in \mathcal{H}$ has
$$\mathbb{E}_{x,y}\left[\mathbf{1}(\hat{h}(x) \neq y)\right] = O(\alpha)$$
- *Agnostic case:* if $n = \tilde{O}\left(V^{3/2}/\alpha^{5/2}\right)$, then w.h.p. output $\hat{h} \in \mathcal{H}$ has
$$\mathbb{E}_{x,y}\left[\mathbf{1}(\hat{h}(x) \neq y)\right] = O(e^* + \alpha) \text{ , where } e^* = \min_{h \in \mathcal{H}} \mathbb{E}_{x,y}\left[\mathbf{1}(h(x) \neq y)\right]$$

Private Training set

$S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ → Our (ϵ, δ)-DP Algorithm for classification queries

$\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ → Our (ϵ, δ)-DP Algorithm

Our (ϵ, δ)-DP Algorithm → $\{\hat{y}_1, \ldots, \hat{y}_m\}$

$\{\tilde{x}_1, \ldots, \tilde{x}_m\}$, $\{\hat{y}_1, \ldots, \hat{y}_m\}$ → (Agnostic) PAC learner $\mathcal{B}$ for $\mathcal{H}$

$\hat{h} \in \mathcal{H}$ ← (Agnostic) PAC learner $\mathcal{B}$ for $\mathcal{H}$

Private learner via Knowledge Transfer

- Prior work on label-privacy *[CH'11, BNS'14]*:
  - Pure DP, white-box constructions.
  - [CH'11] obtains sample complexity bounds: involves smoothness assumptions on the distribution.
  - [BNS'14] obtains upper bounds for the realizable setting only.

Extension: privately answering soft-label queries

- A soft-label $\in [0,1]$ for a feature-vector x is an estimate for the conditional probability $p(y=1 \mid x)$
- Applications: ranking and product recommendation.

Extension: privately answering soft-label queries

- A soft-label $\in [0,1]$ for a feature-vector x is an estimate for the conditional probability $p(y=1 \mid x)$
- Applications: ranking and product recommendation.

- A construction with private predictions nearly as accurate as the non-private ones with a small cost $\tilde{O}\left(\sqrt{T} / \epsilon\right)$ in sample size assuming:
  - # queries with low label-noise (high margin) $\geq m - T$

$$| \, p(y=1|x) - 0.5 \, |$$

Extension: privately answering soft-label queries

- A soft-label $\in [0,1]$ for a feature-vector x is an estimate for the conditional probability $p(y=1 \mid x)$
- Applications: ranking and product recommendation.

- A construction with private predictions nearly as accurate as the non-private ones with a small cost $\tilde{O}\left(\sqrt{T} / \epsilon\right)$ in sample size assuming:
  - # queries with low label-noise (high margin) $\geq m - T$
  - the non-private learner satisfies a weak notion of stability

    (*on-average stability*), *satisfied by SGD.*

Summary

1. A new general paradigm for answering *“stable” queries*:
   - Based on a new approach combining *distance-to-instability* *[Smith-Thakurta’13]* with *sparse-vector [DNRRV’09, DR14]* techniques.

2. New construction for *privately answering classification queries*:
   - Bounds on misclassification rate in the standard PAC model: *better than what is implied by advanced composition*.

3. A black-box construction for a private learner via knowledge transfer with rigorous guarantees
   - Sample complexity bounds in terms of VC-dimension.
   - also, serves as label-private learner.

4. Extension: construction for privately answering soft-label queries