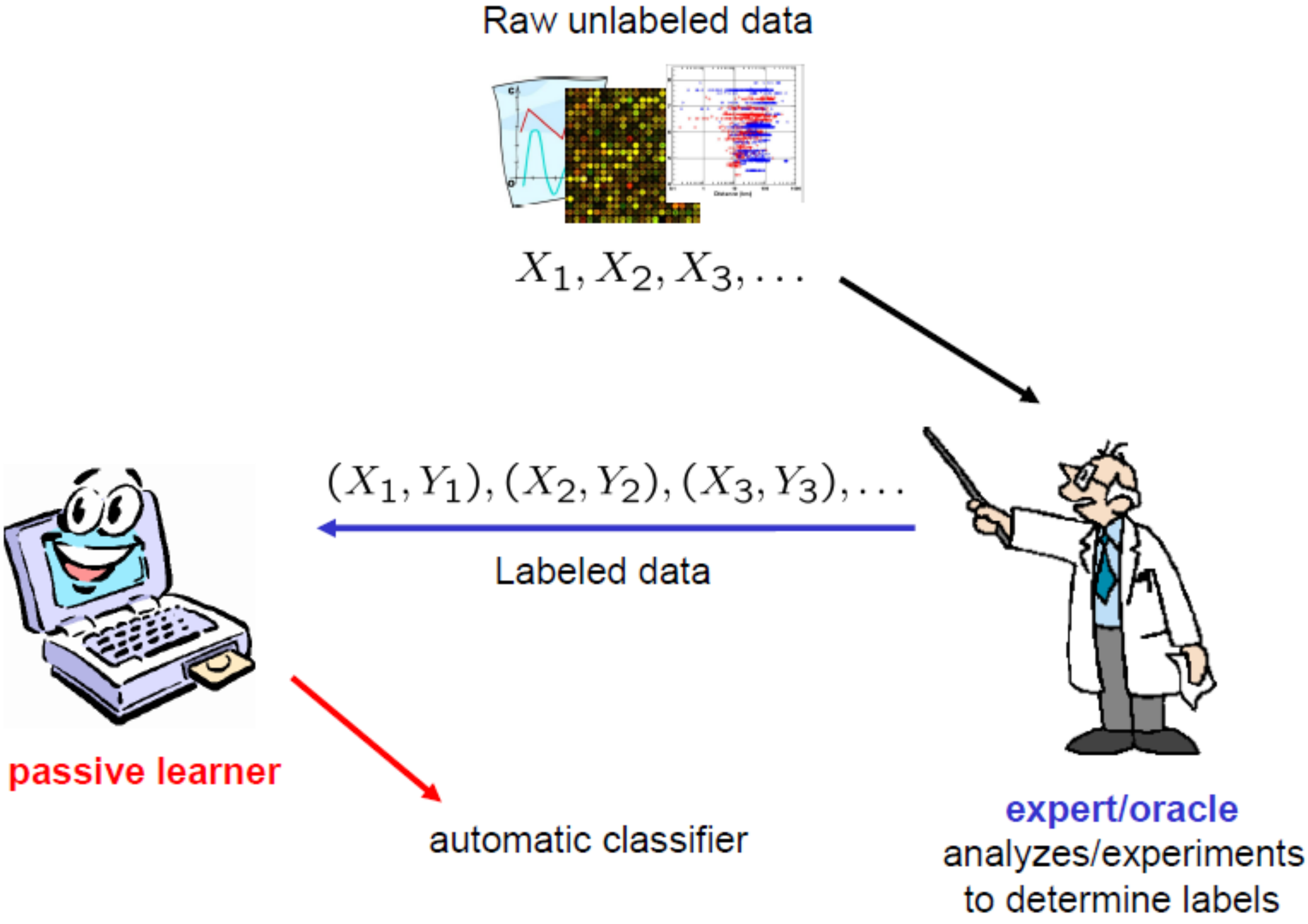
Tsybakov noise adaptive margin-based active learning

Aarti Singh
A. Nico Habermann Associate Professor

NIPS workshop on Learning Faster from Easy Data II
Dec 11, 2015

Passive Learning



Raw unlabeled data

$X_1, X_2, X_3, \ldots$

$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots$

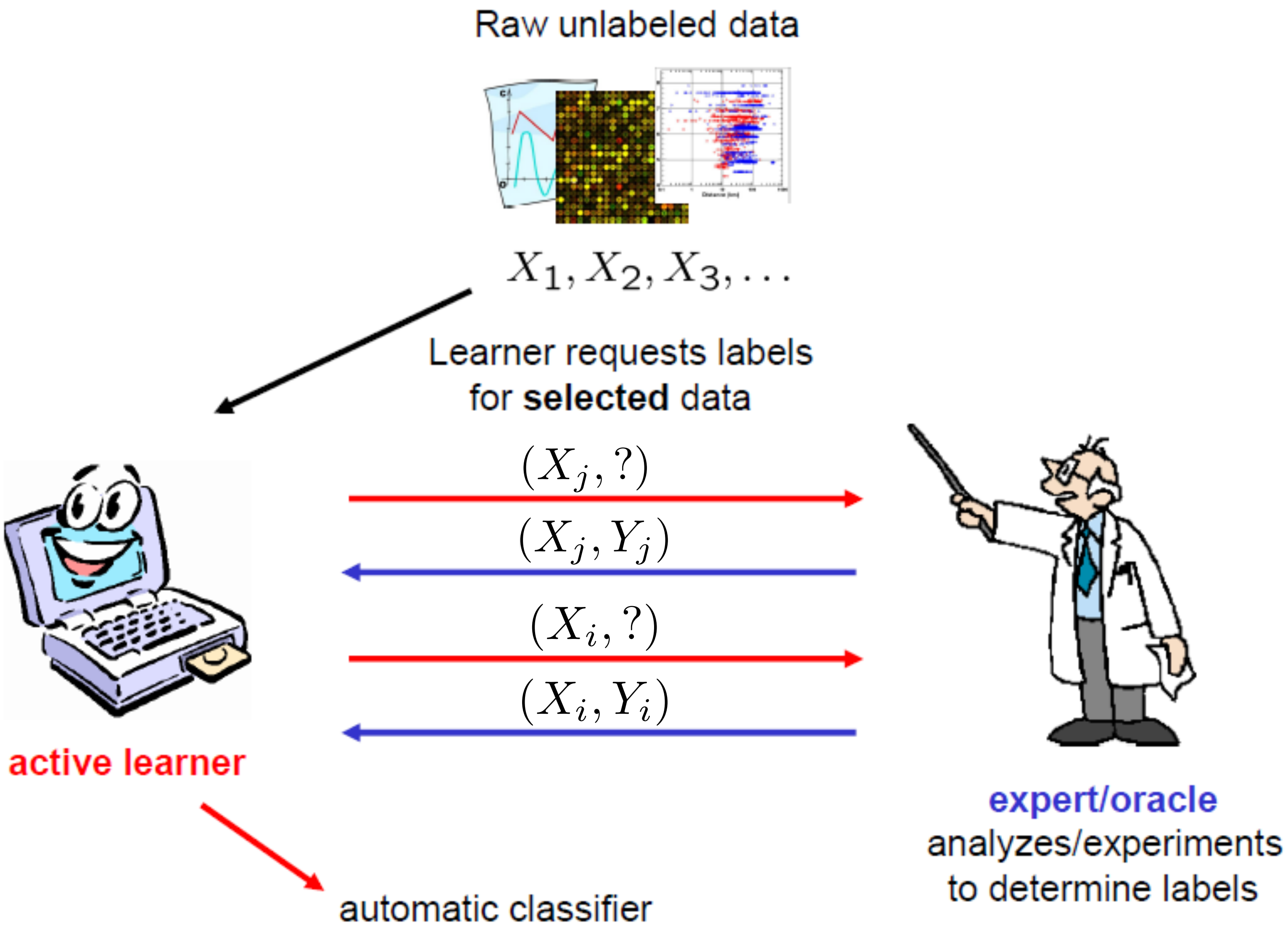
Labeled data

passive learner

automatic classifier

expert/oracle
analyzes/experiments
to determine labels

Active Learning



Raw unlabeled data

$X_1, X_2, X_3, \ldots$

Learner requests labels for **selected** data

$(X_j, ?)$

(X_j, Y_j)

$(X_i, ?)$

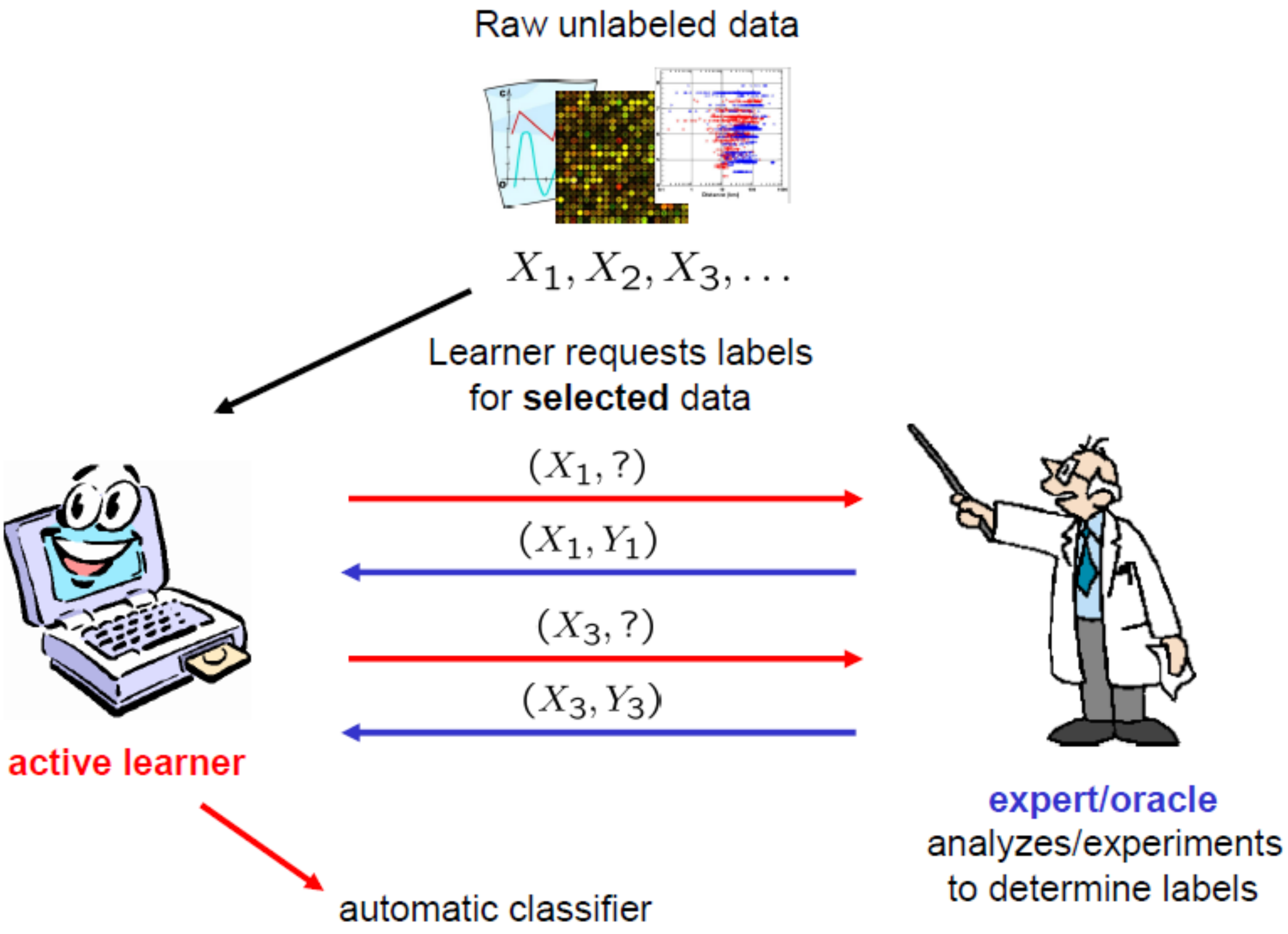
(X_i, Y_i)

active learner

automatic classifier

expert/oracle
analyzes/experiments
to determine labels

Streaming setting



Raw unlabeled data

$X_1, X_2, X_3, \ldots$

Learner requests labels
for **selected** data

$(X_1, ?)$

(X_1, Y_1)

$(X_3, ?)$

(X_3, Y_3)

active learner

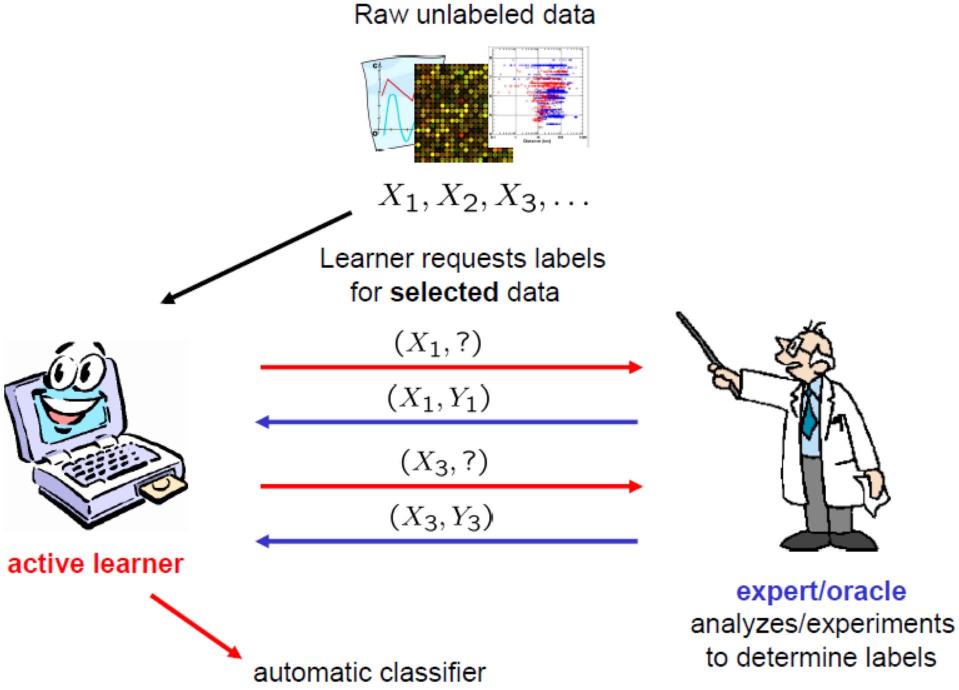
automatic classifier

expert/oracle
analyzes/experiments
to determine labels

Streaming setting

- Algorithm obtains X_t sampled iid from marginal distribution P_X

- Based on previous labeled and unlabeled data, the algorithm decides whether or not to accept X_t and query its label.

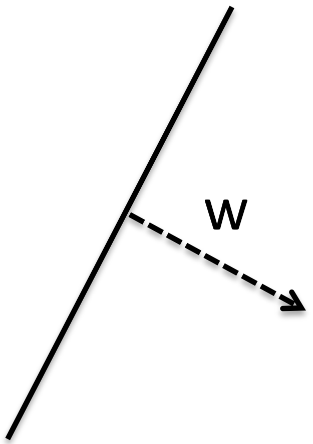


- If label is queried, algorithm receives Y_t sampled iid from conditional distribution $P(Y|X=X_t)$

Problem setup

- X is d-dimensional, P_X is uniform (or log-concave + isotropic)
- Binary classification: Labels Y in {+1, -1}
- Homogeneous linear classifiers sign(w. X)

  with $||w||_2 = 1$



- err(w) = P(sign(w.X) ≠ Y)
- Bayes optimal classifier is linear w*

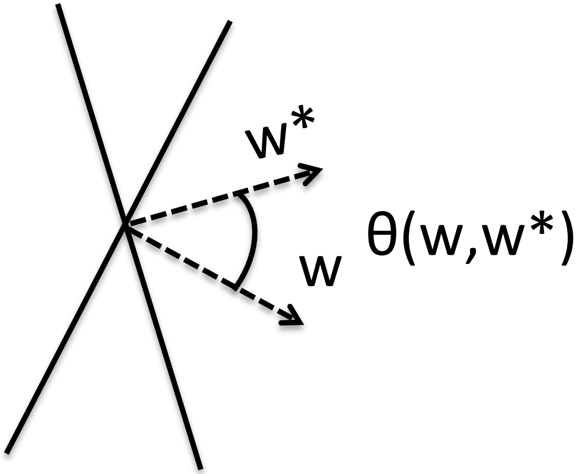
  $\arg\max_Y P(Y|X) = \text{sign}(w^*. X)$

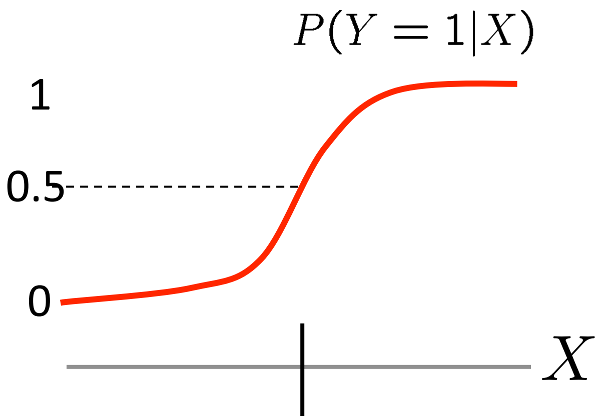
Tsybakov Noise Condition

For all linear classifiers w with $||w||_2 = 1$

$$\mu\, \theta(w,w^*)^{\kappa} \leq err(w) - err(w^*)$$

where κ in $[1,\infty)$ is the TNC exponent and $0 < \mu < \infty$ is a constant.





κ characterizes noise in label distribution

κ makes problem easy or hard – small κ implies easier problem

Minimax active learning rates

If Tsybakov Noise Condition (TNC) holds, then minimax optimal active learning rate is

$$E[\mathrm{err}(w_T) - \mathrm{err}(w^*)] = \tilde{O}((d/T)^{\kappa/(2\kappa-2)})$$

$\kappa = \infty$ passive rate $1/\sqrt{T}$

$\kappa = 1$ exponential rate e^{-T}

Lower bound: Castro-Nowak'06 (d=1), Hanneke-Yang'14 (d, P_X), Singh-Wang'14 (d, lower-bounded/uniform P_X)

Algorithms need to know κ!!

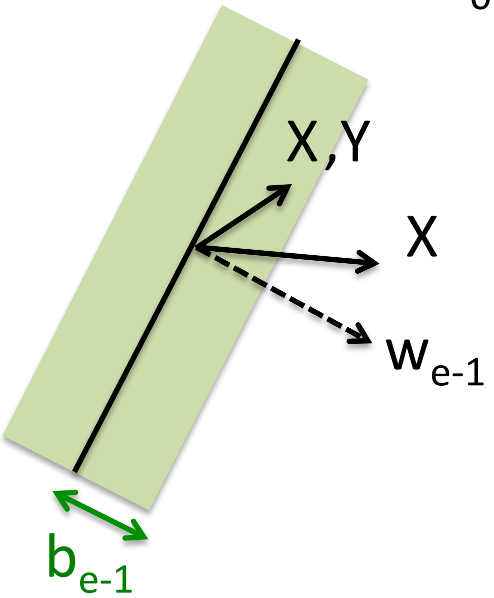
Model selection for active learning - Can we adapt to easy cases, while being robust to worst-case?

Balcan-Broder-Zhang’07

Margin-based active learning

- Input: Desired accuracy ε, Failure probability δ
- Initialize: E; For e = 1, …, E: epoch budgets T_e , search radii R_e , acceptance regions b_e , precision values ε_e; random classifier w_0
- For e = 1, …, E

```
    Until labeled examples < $T_e$
        Obtain a sample $X_t$ from $P_X$
        If $|w_{e-1}. X_t| \le b_{e-1}$, query label $Y_t$
    end
```

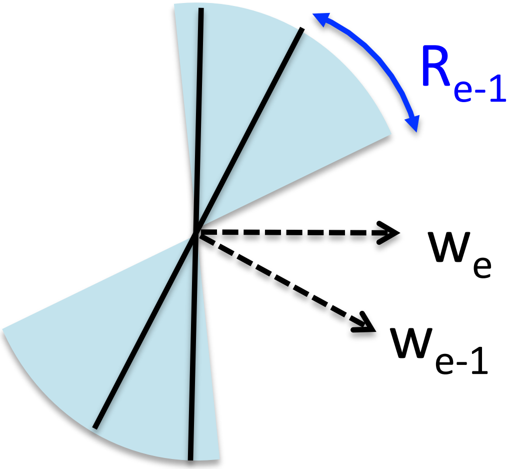


Balcan-Broder-Zhang’07

Margin-based active learning

- Input: Desired accuracy ε, Failure probability δ
- Initialize: E; For e = 1, …, E: epoch budgets T_e , search radii R_e , acceptance regions b_e , precision values ε_e; random classifier w_0
- For e = 1, …, E

```
    Until labeled examples < $T_e$
        Obtain a sample $X_t$ from $P_X$
        If $|w_{e-1}. X_t| \le b_{e-1}$, query label $Y_t$
    end
    Find $w_e$ that (approximately) minimizes training error up
    to precision $\varepsilon_e$ on the $T_e$ labeled examples among all w
    s.t. $\theta(w,w_{e-1}) \le R_{e-1}$
```



- Output: $w_T = w_E$

Balcan-Broder-Zhang'07

Margin-based active learning

- Input: Desired accuracy ε, Failure probability δ
- Initialize: E; For e = 1, …, E: epoch budgets T_e , search radii R_e , acceptance regions b_e , precision values ε_e; random classifier w_0

All depend on κ

- For e = 1, …, E

```
    Until labeled examples < $T_e$
        Obtain a sample $X_t$ from $P_X$
        If $|w_{e-1}. X_t| \leq b_{e-1}$, query label $Y_t$
    end
    Find $w_e$ that (approximately) minimizes training error up
    to precision $\varepsilon_e$ on the $T_e$ labeled examples among all w
    s.t. $\theta(w,w_{e-1}) \leq R_{e-1}$
```

- Output: $w_T = w_E$

Adaptive margin-based active learning

- Input: Query budget T, Failure probability δ, shrink rate r
- Initialize: $E = \log \sqrt{T}$; For e = 1, …, E: epoch budgets $T_e = T/E$, search radius $R_0 = \pi$, acceptance region $b_0 = \infty$; random classifier w_0
- For e = 1, …, E

```
    Until labeled examples < $T_e$
        Obtain a sample $X_t$ from $P_X$
        If $|w_{e-1}. X_t| \leq b_{e-1}$, query label $Y_t$
    end
    Find $w_e$ that ~~(approximately)~~ minimizes training error on
    the $T_e$ labeled examples among all w s.t. $\theta(w,w_{e-1}) \leq R_{e-1}$
    $R_e = r R_{e-1}$; $b_e = 2R_e\sqrt{[E(1+\log(1/r))/d]}$
```

No knowledge of κ

- Output: $w_T = w_E$

Adaptive margin-based active learning

Let $T \geq 4$, $d \geq 4$, r in (0,1/2), P_X is uniform on d-dim unit ball and $P_{Y|X}$ satisfies TNC(μ, κ). Then the streaming adaptive active learning algorithm achieves, with probability $\geq 1 - \delta$,

$$\mathrm{err}(w_T) - \mathrm{err}(w^*) = \tilde{O}((d+\log(1/\delta)/T)^{\kappa/(2\kappa-2)})$$

for all $1+ 1/(\log(1/r)) \leq \kappa < \infty$.

Minimax optimal rate without knowing μ, κ up to log factors!!

Adapt to easy cases, while being robust to worst-case!

Why does it work? (proof sketch)

Consider shrink rate $r = ½$. We will argue adaptivity to κ in $[2,\infty)$

Let w_e^* denote the best linear classifier among all w s.t. $\theta(w,w_{e-1}) \leq R_{e-1}$ in acceptance region b_{e-1}

For all e, with high probability

$\text{err}(w_e) - \text{err}(w_e^*) = \tilde{O}(R_{e-1} (d/T)^{1/2})$ passive rate

For e = 1,we have $(d/T)^{1/2}$

For e = E we have d/T since $R_E = R_0/2^E = R_0/\sqrt{T}$. (but $w^*_E \neq w^*$)

Therefore, there exists epoch e' s.t. with high probability

$\text{err}(w_{e'}) - \text{err}(w_{e'}^*) = \tilde{O}((d/T)^{\kappa/(2\kappa-2)})$

Why does it work? (proof sketch)

Consider shrink rate r = ½. We will argue adaptivity to κ in $[2,\infty)$

Let w_e^* denote the best linear classifier among all w s.t. $\theta(w,w_{e-1}) \leq R_{e-1}$ in acceptance region b_{e-1}

Therefore, there exists epoch e' s.t. with high probability

$$\mathrm{err}(w_{e'}) - \mathrm{err}(w_{e'}^*) = \tilde{O}((d/T)^{\kappa/(2\kappa-2)})$$

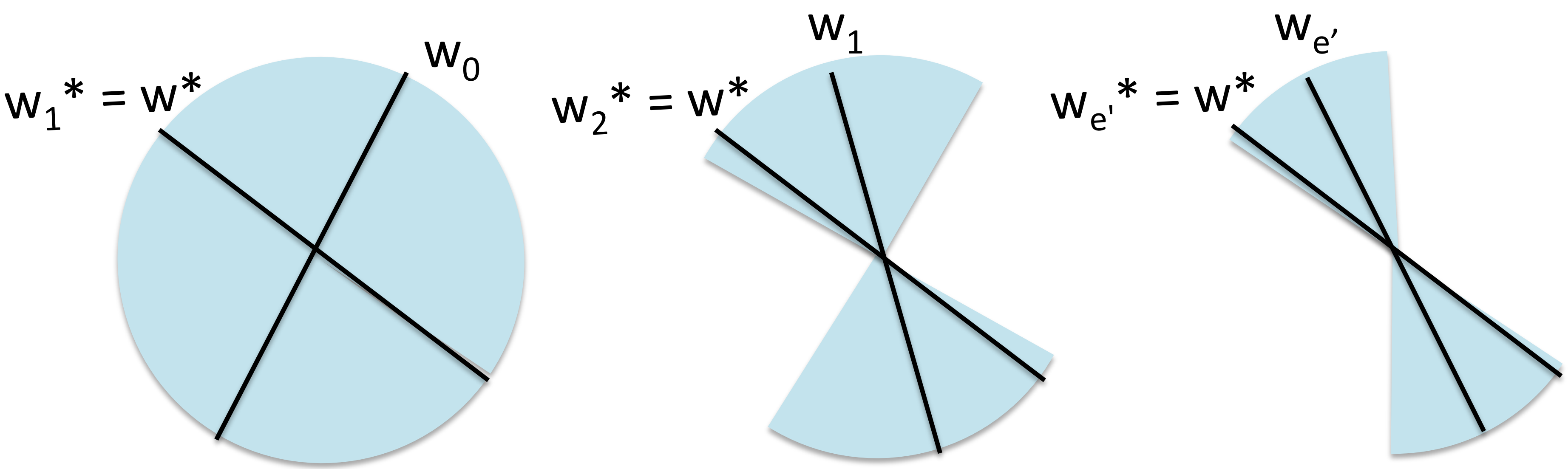
Also, $w_{e'}^* = w^*$ (using same argument as Balcan-Broder-Zhang'07)

Point of departure: They ensure $w^*_e = w^*$ for all e

we allow $w^*_e \neq w^*$ for all $e \geq e'$

Why does it work? (proof sketch)

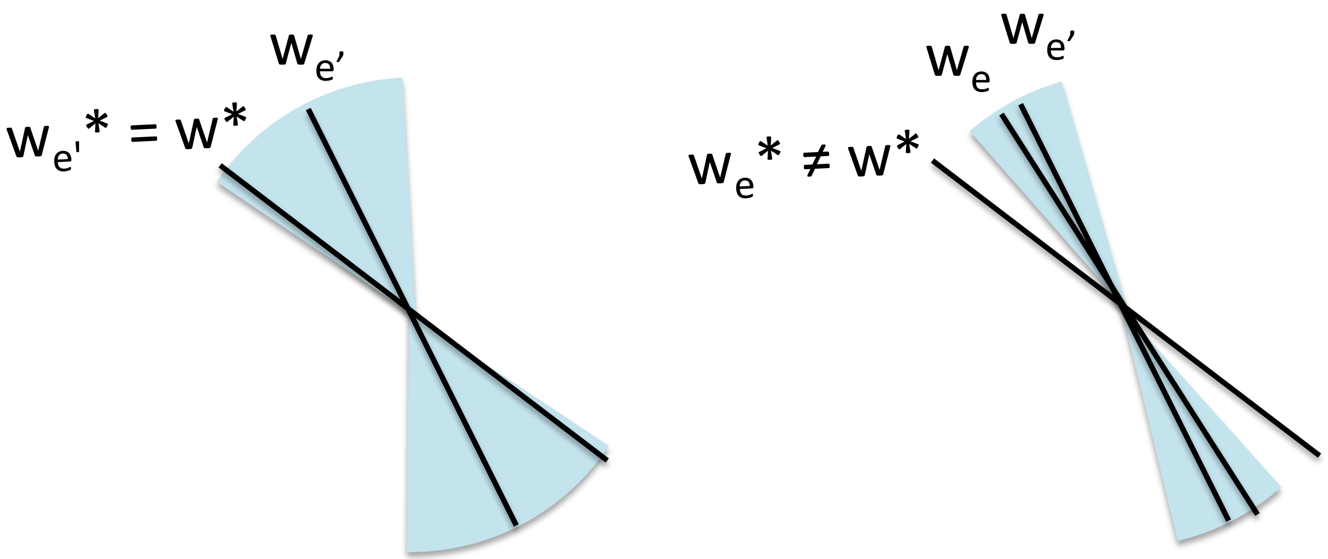
Let w_e^* denote the best linear classifier among all w s.t. $\theta(w,w_{e-1}) \leq R_{e-1}$ in acceptance region b_{e-1}



There exists epoch e’ s.t. $err(w_{e'}) - err(w_{e'}^*) = \tilde{O}((d/T)^{\kappa/(2\kappa-2)})$

Why does it work? (proof sketch)

Let w_e* denote the best linear classifier among all w s.t. $\theta(w, w_{e-1}) \leq R_{e-1}$ in acceptance region b_{e-1}



There exists epoch e' s.t. $\mathrm{err}(w_{e'}) - \mathrm{err}(w_{e'}^*) = \tilde{O}((d/T)^{\kappa/(2\kappa-2)})$

For all epochs e ≥ e', w_e stays close to $w_{e'}$

Why does it work? (proof sketch)

Consider shrink rate r = ½. We will argue adaptivity to κ in $[2,\infty)$

Let w_e* denote the best linear classifier among all w s.t. $\theta(w,w_{e-1}) \leq R_{e-1}$ in acceptance region b_{e-1}

Therefore, there exists epoch e' s.t. with high probability

$$err(w_{e'}) - err(w_{e'}^*) = \tilde{O}((d/T)^{\kappa/(2\kappa-2)})$$

Also, $w_{e'}^* = w^*$ (using same argument as Balcan-Broder-Zhang'07)

For all epochs e ≥ e' $\quad err(w_e) - err(w_{e'}) = \tilde{O}((d/T)^{\kappa/(2\kappa-2)})$

Therefore, $err(w_E) - err(w^*) = \tilde{O}((d/T)^{\kappa/(2\kappa-2)})$.

Adaptive margin-based active learning

Let $T \geq 4$, $d \geq 4$, r in $(0,1/2)$, P_X is **log-concave and isotropic** on d-dim unit ball and $P_{Y|X}$ satisfies TNC(μ, κ). Then the streaming adaptive active learning algorithm with achieves, with $b_e = C R_e \log T$ probability $\geq 1 - \delta$,

$$\mathrm{err}(w_T) - \mathrm{err}(w^*) = \tilde{O}((d+\log(1/\delta)/T)^{\kappa/(2\kappa-2)})$$

for all $1+ 1/(\log(1/r)) \leq \kappa < \infty$.

Minimax optimal rate without knowing μ, κ up to log factors!!

Adapt to easy cases, while being robust to worst-case!

Limitations/Open questions

- Constants become large as κ tends to 1, log factors

  $\mu^{-1/(\kappa-1)} r^{-(\kappa-2)/(\kappa-1)}$

- How to adapt to $\kappa = 1$? $1 + 1/(\log(1/r)) \leq \kappa < \infty$

- Adaptive active learning given desired accuracy ε (instead of query budget T)

- Agnostic setting (Bayes optimal classifier not in hypothesis space)

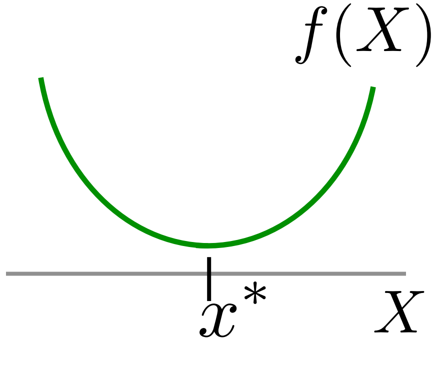
Related work

- Juditsky-Nesterov’14 – adaptive stochastic optimization of uniformly convex functions (κ ≥ 2)

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\lambda}{2}\|x - y\|^\kappa$$

- Our analysis extends to achieve adaptive optimization of TNC functions (κ ≥ 1)

$$f(x) - f(x^*) \geq \lambda \|x - x^*\|^\kappa$$



- For d = 1 $\|f(\widehat{x}) - f(x^*)\| \asymp T^{-\frac{\kappa}{2\kappa-2}}$

  Rates exactly same as 1-dim active learning!

Related work

- Same algorithm also studied by Awasthi et al’14 for a different question:

  Maximum amount of adversarial noise tolerated by algorithm for constant excess risk and polylog sample complexity (exponential rate for error)

We study convergence of excess risk to zero with increasing samples not restricted to be polylog.

References + Acknowledgements

- Noise-adaptive Margin-based Active Learning and Lower Bounds under Tsybakov Noise Condition, AAAI'16 To appear.
- Algorithmic connections between active learning and stochastic convex optimization, ALT'13.
- Optimal rates for stochastic convex optimization under Tsybakov noise condition, ICML'13.

Yining Wang

Aaditya Ramdas