Always Asking for Advice is Often Optimal

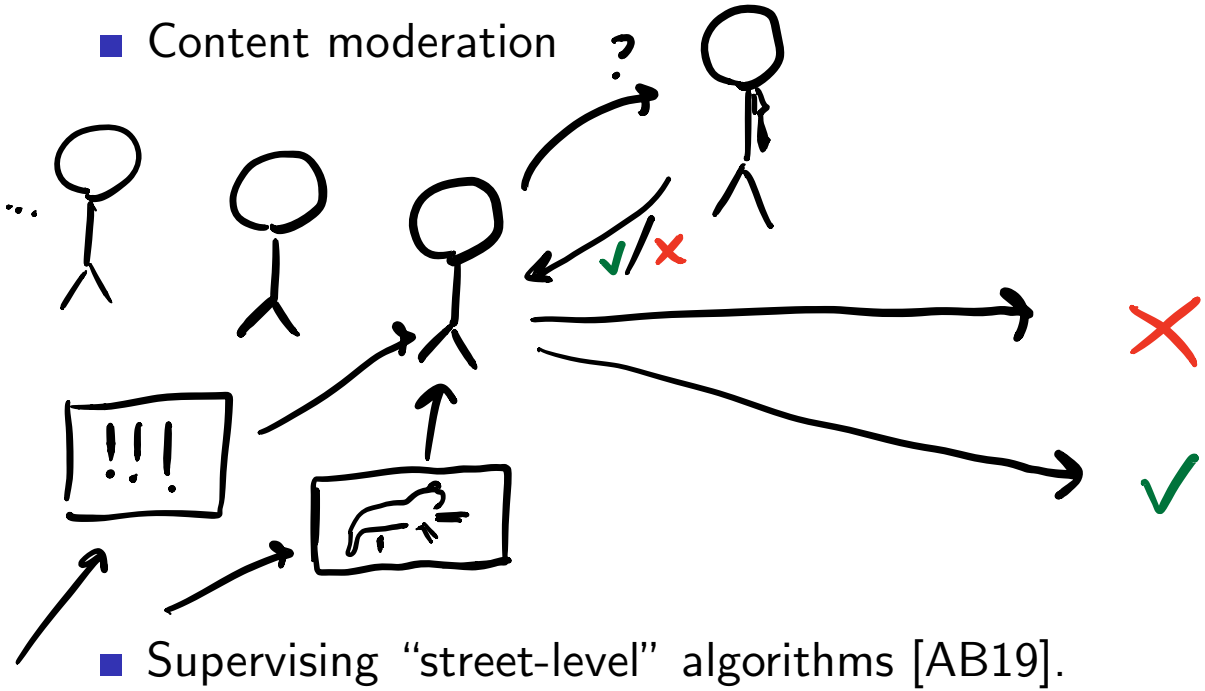
Spencer Peters

Joint work with Sid Banerjee and Joe Halpern

April 9, 2021

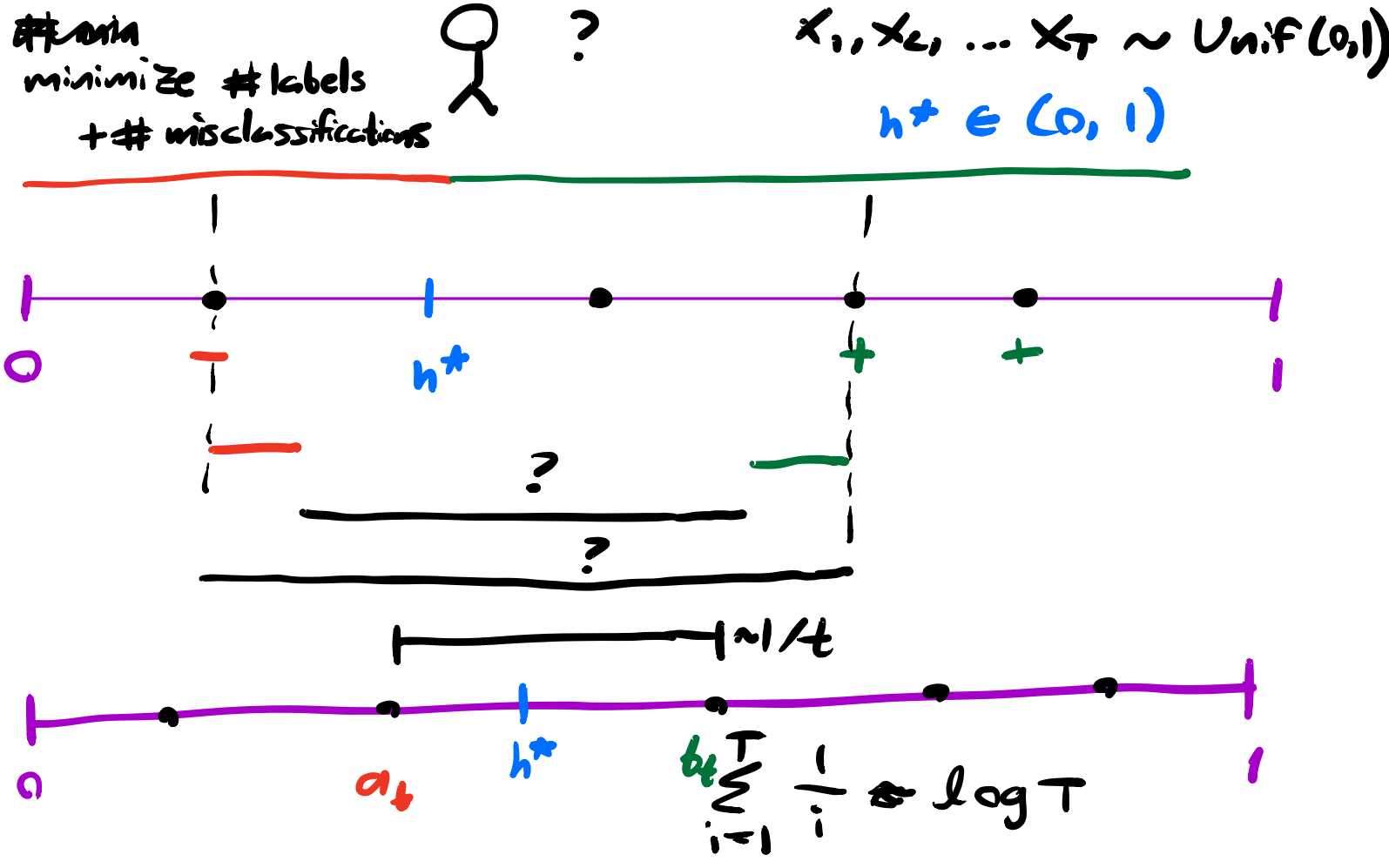
Context

- Content moderation



- Supervising "street-level" algorithms [AB19].
- Legal systems
- Selective sampling + online decision-making
- Consistency matters

Thresholds I



Thresholds II

$$\Pr(\text{ask for } x_t\text{'s label}) = \Pr(x_t = c_+ \vee x_t = b_t)$$

$$\leq \frac{2}{t}$$

$$\mathbb{E}[\#\text{labels}] = \sum_{t=1}^{T} \Pr(\text{ask for } x_t\text{'s label})$$

$$\leq \sum_{t=1}^{T} \frac{2}{t} \leq 2(\log T + 1)$$

Preliminaries

The Setting

Instance space $\mathcal{X}$, distribution $\mathcal{D}$ over $\mathcal{X}$. Hypothesis space $\mathcal{H}$.

The Task

Algorithm gets i.i.d. samples $X_1, X_2, \ldots, X_T$ one at a time.
Guaranteed that labels are $h^*(X_t)$ for some $h^* \in \mathcal{H}$ (realizability).
After seeing X_t, can either ask for its label $h^*(X_t)$, or try to guess it.

The (Informal) Goal

Minimize sum of # wrong guesses and # label requests.

Main Result

Definition (Version Space $\mathcal{H}_t$)

$\mathcal{H}_t = \{h \in \mathcal{H} \mid h(X_i) = h^*(X_i) \text{ for all } 1 \leq i \leq t-1\}$.

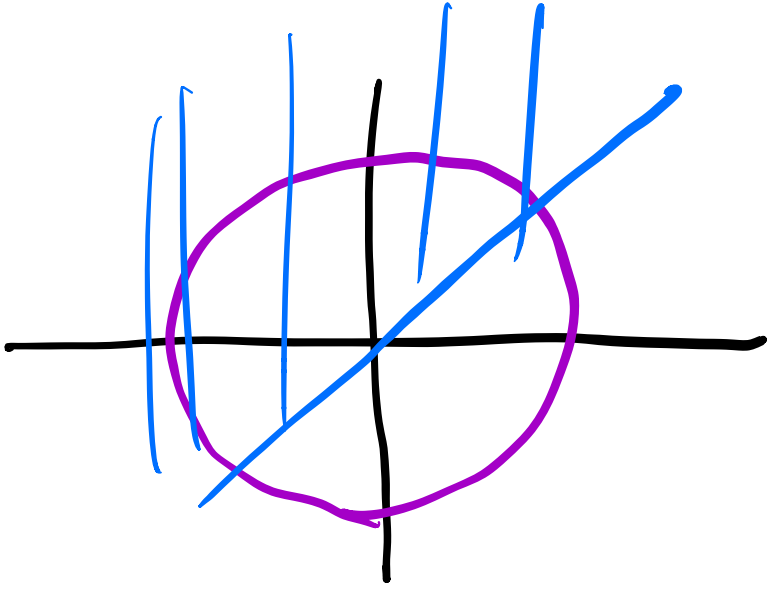
CAL algorithm A_{CAL} [CAL94]

Guess y_t only if $h(X_t) = y_t$ for all $h \in \mathcal{H}_t$. Otherwise ask for X_t's label.

Main Result

Theorem

If $d \geq 1$, $\mathcal{D}$ is uniform on the unit hypersphere $S^d \subset \mathbb{R}^d$, and $\mathcal{H}$ is the set of halfspaces through the origin, then A_{CAL} makes $\Theta(d^{3/2} \log T)$ label requests in expectation (and no wrong guesses).



Main Result

Theorem

If $d \geq 1$, $\mathcal{D}$ is uniform on the unit hypersphere in $\mathbb{R}^d$, and $\mathcal{H}$ is the set of halfspaces through the origin, then A_{CAL} makes $\Theta(d^{3/2} \log T)$ label requests in expectation (and no wrong guesses).

- This is the optimal dependence on T; known PAC lower bounds imply $\Omega(d \log T)$.
- Moreover, we show that in this setting,

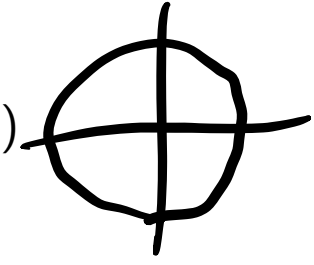
A_{CAL} can be implemented in (amortized expected) time $O(d^{3.6})$ per instance X_t, independent of t.

This implementation has reasonable wall-clock time performance (~ 0.3 seconds per instance) even for moderately large numbers of features ($d = 50$).

Related Work

- *Disagreement coefficient*-based analysis due to Hanneke [Han07, Han14], specialized to our setting: with probability $1-\delta$, A_{CAL} makes

$$O(d^{3/2}\log\sqrt{d}\log T + \sqrt{d}\log T\log\frac{\log T}{\delta})$$



label requests with high probability.

- Our result, although not a high-probability bound, removes the $\log\log T$ term, thus showing A_{CAL} eventually performs better than previously known when d is fixed.
- Hanneke's analysis did not show that A_{CAL} can be implemented efficiently.

Related Work

- Prior *efficient* algorithms for selective sampling include the Active Modified Perceptron algorithm [DKM05] and a class of *margin-based* algorithms [BBZ07]. All of these algorithms make mistakes.
- Moreover, their error bounds all have $O(\log T \log\log T)$ dependence on T. (Some of them have the optimal $O(d)$ dependence on dimension.)
- However, these algorithms apply in more general settings. Margin-based algorithms have been shown to work for more general distributions over $\mathbb{R}^d$, namely, log-concave distributions [BL13]. Also, an analogue of A_{CAL} has been designed for the *agnostic* case where the labels need not correspond to some $h^* \in \mathcal{H}$ [BBL09].

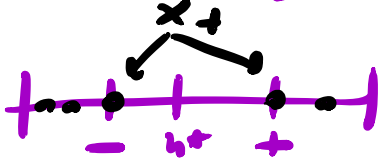
A^2

Undetermined Observations

Similar to H_t, define

$$H^t_{-i} = \{h \in H \mid h(x_j) = h^*(x_j),\ j \in [t],\ j \neq i\}$$

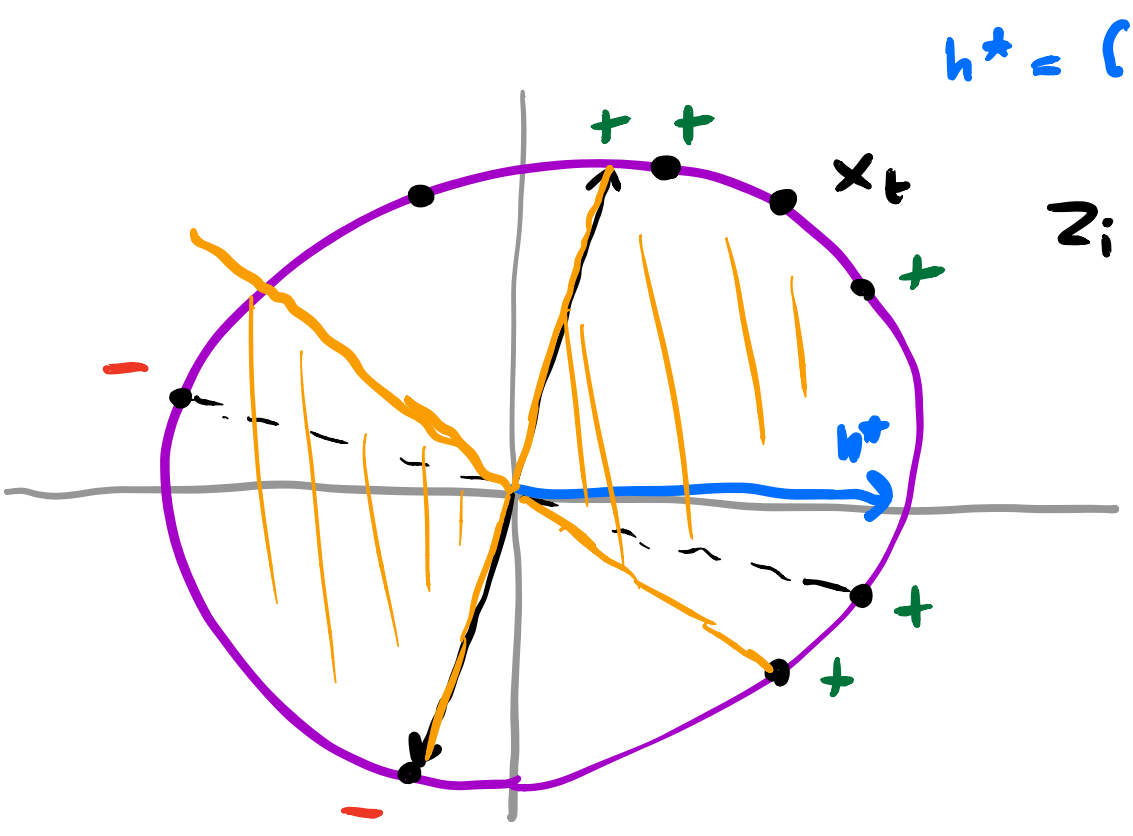
$$B_t = \{x_i \mid i \in [t],\ \exists h, h' \in H^t_{-i},\ h(x_i) \neq h'(x_i)\}$$



B_t analogous to $\{a_t, b_t\}$

$$\Pr(\text{ask for label of } x_t) = \frac{|B_t|}{t}.$$

Equivalent Sample



Undetermined Observations are Vertices

Lemma $x_i \in \mathcal{B}_t$ iff

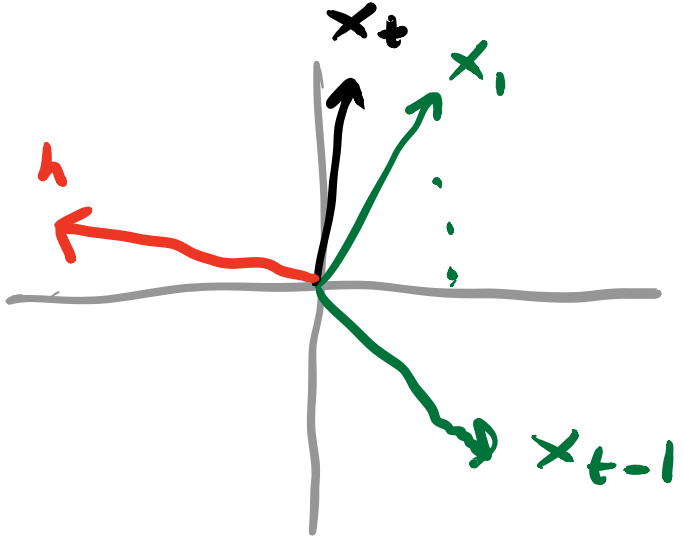
$x_i \in \text{cone}(z_1, \ldots z_{i-1}, z_{i+1}, \ldots z_t)$

or $-x_i \in \text{cone}($ " $)$

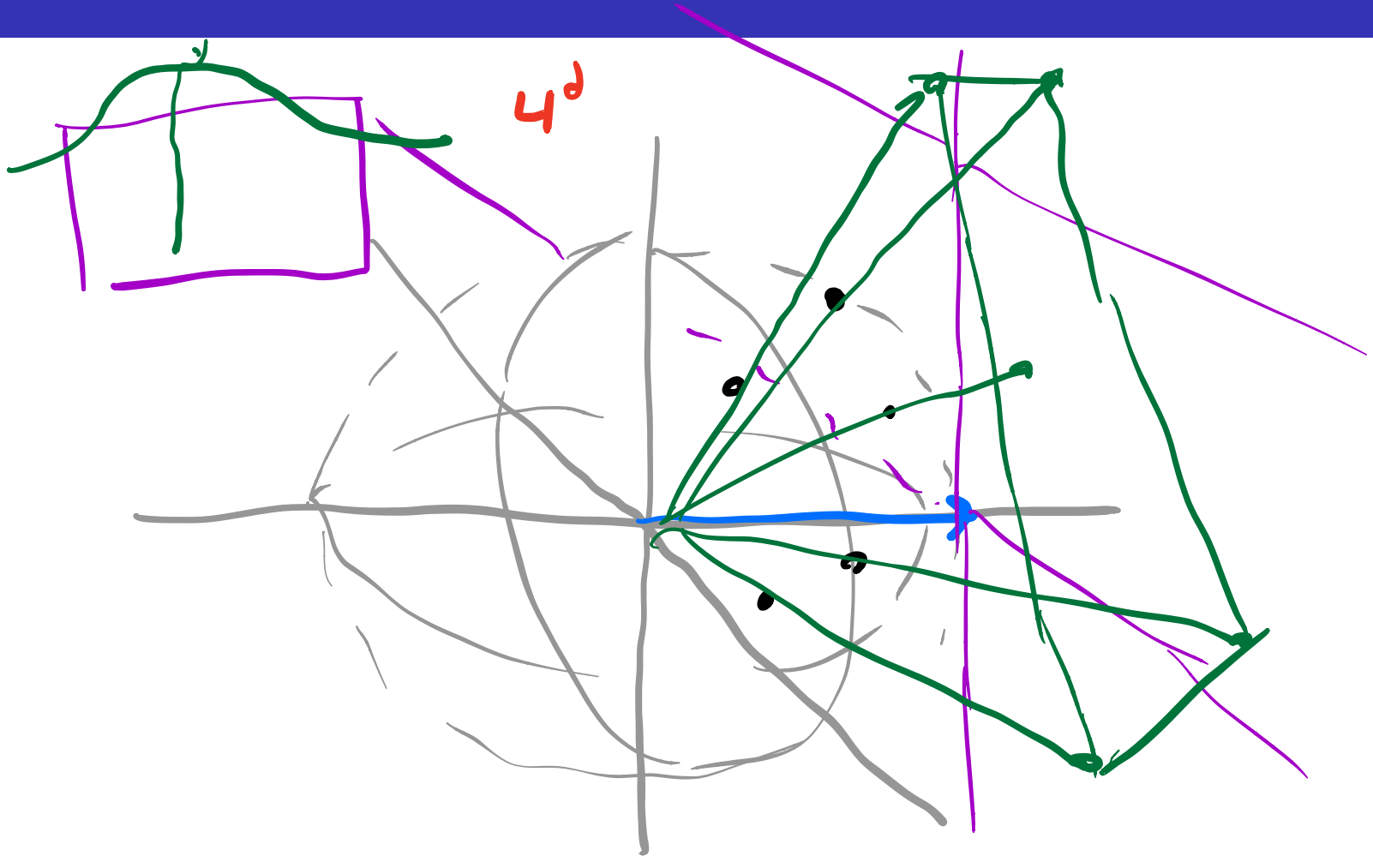
Farkas's Lemma

$h \cdot x_t > 0$

$h \cdot x_i \leq 0$



Cone Size is Constant



Cone Size is Constant

Power of email! Result based on [Kab20] after I found [KMTT19].

Let $Z_1, \ldots Z_t \underset{\text{iid}}{\sim} S^d_+$

Let $C_t = \text{cone}\{Z_1, \ldots Z_t\}$

$N_t = \#$ edges of C_t

$$\lim_{d\to\infty} \lim_{t\to\infty} \mathbb{E}\, N_t / d^{3/2} \sim \sqrt{\frac{2\pi}{3}}$$

Putting it Together

$$\Pr(\text{Asc for } x_t\text{'s label}) = \frac{|\mathcal{B}_t|}{t}$$

$$|\mathcal{B}_t| = N_t \approx \sqrt{\frac{2\pi}{3}}\, d^{3/2}$$

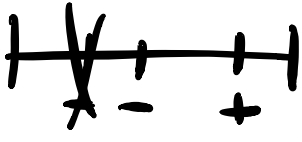
$$\mathbb{E}[\#\text{ label requests}] = \sum_{t=1}^{T} \frac{|\mathcal{B}_t|}{t}$$

$$= \sum_{t=1}^{T} \frac{\Theta(d^{3/2})}{t} = \Theta(d^{3/2} \log T)$$

Implementing CAL

Algorithm 1 CAL

$\mathcal{B} \leftarrow \emptyset$
$N = 0$ ▷ $|\mathcal{B}|$ after last call to RRP.
for $t = 1, 2, \ldots$ **do**
    if $X_t \in \text{cone}(\mathcal{B})$ **then**
        Classify X_t as positive.
    else if $-X_t \in \text{cone}(\mathcal{B})$ **then**
        Classify X_t as negative.
    else
        Ask for the label Y_t.
        Compute $Z_t = Y_t X_t$.
        $\mathcal{B} \leftarrow \mathcal{B} \cup \{Z_t\}$
        if $|\mathcal{B}| \geq 2N$ **then**
            $\mathcal{B} \leftarrow \text{RRP}(\mathcal{B})$
            $N \leftarrow |\mathcal{B}|$.

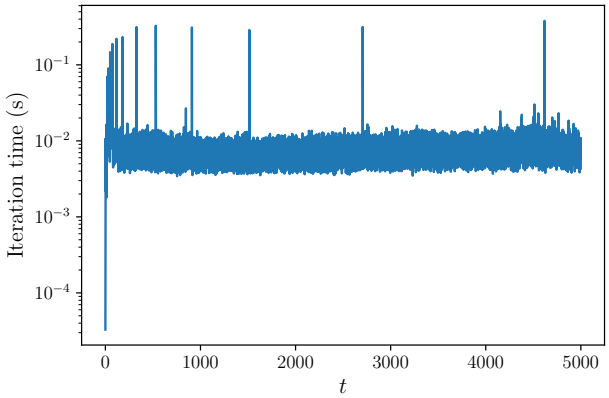


Implementing CAL with Linear Programming

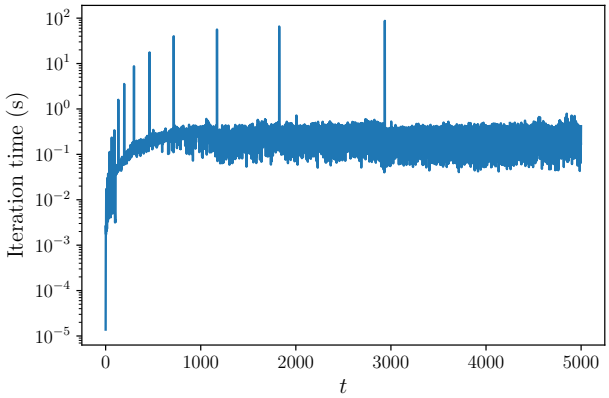
LP has $O(d^{3/2})$ $O(n^w \log n) < O(n^{2.37})$

$O(d^{3.5})$

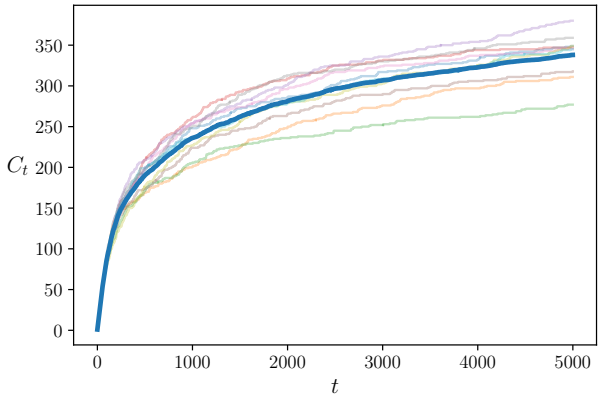
Results

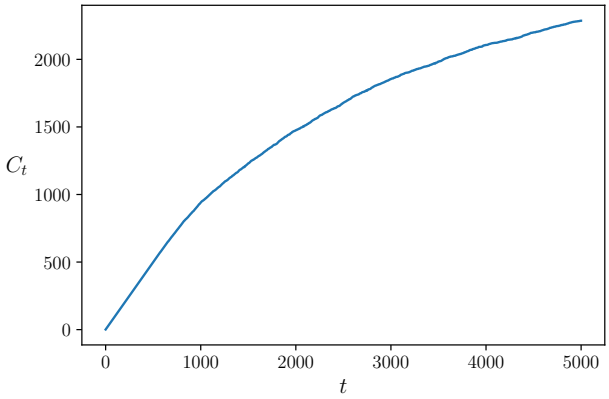


(a) Time per iteration, $d = 10$

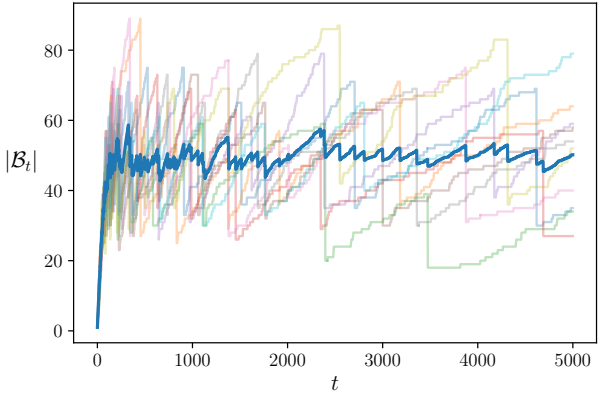


(b) Time per iteration, $d = 50$

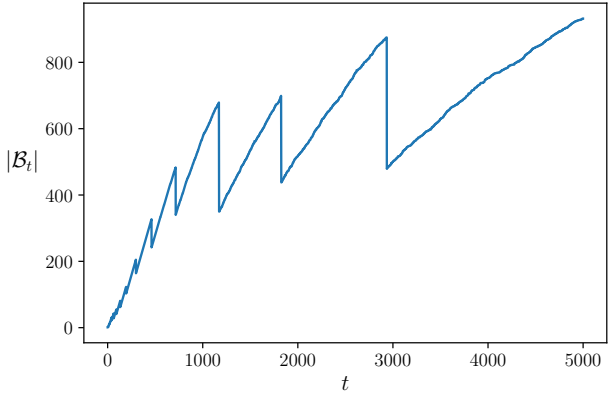




Results



(a) Size of the set $\mathcal{B}$, S^{10}.



(b) Size of the set $\mathcal{B}$, S^{50}.

Summary

- A_{CAL} [CAL94] is optimal among selective sampling algorithms that make no mistakes.
- We show that, in the commonly studied setting of $\mathcal{D} = \text{Unif} S^d$ and $\mathcal{H}$ halfspaces through the origin, it is more label-efficient than previously thought, and it can be implemented in a computationally efficient manner.
- This suggests that "safe" decision-making rules similar to A_{CAL} deserve a second look for applications such as content moderation, legal processes, and the supervision of algorithms.

$\Omega(d \log T)$

Future Directions

- Real world decision making is noisy and complicated. A_{CAL} has been extended to an algorithm A^2 for the "agnostic" setting, which can model noisy labels and tolerate simplified hypothesis classes [BBL09]. Can our analysis be extended to A^2?
- What about richer classes of distributions, such as log-concave distributions [BL13]?
- Modeling hypothesis classes that change over time (evolving norms)?
- What are the obstacles to using A_{CAL} in practice?

Questions?

References I

- A. Alkhatib and M. Bernstein, *Street-level algorithms: A theory at the gaps between policy and decisions*, Proc 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–13.
- M.-F. Balcan, A. Beygelzimer, and J. Langford, *Agnostic active learning*, Journal of Computer and System Sciences **75** (2009), no. 1, 78–89.
- M.-F. Balcan, A. Broder, and T. Zhang, *Margin based active learning*, 20th Annual Conference on Learning Theory (COLT 2007), 2007, pp. 35–50.
- Gyora M Benedek and Alon Itai, *Learnability with respect to fixed distributions*, Theoretical Computer Science **86** (1991), no. 2, 377–389.

References II

- M.-F. Balcan and P. Long, *Active and passive learning of linear separators under log-concave distributions*, 26 Annual Conference on Learning Theory (COLT 2013), 2013, pp. 288–316.
- D. Cohn, L. Atlas, and R. Ladner, *Improving generalization with active learning*, Machine Learning **15** (1994), no. 2, 201–221.
- S. Dasgupta, A. T. Kalai, and C. Monteleoni, *Analysis of perceptron-based active learning*, International Conference on Computational Learning Theory, 2005, pp. 249–263.
- S. Hanneke, *A bound on the label complexity of agnostic active learning*, Proc. 24th International Conference on Machine learning, 2007, pp. 353–360.

References III

- Steve Hanneke, *Theory of active learning*, Foundations and Trends in Machine Learning **7** (2014), no. 2-3, 131–309.
- Zakhar Kabluchko, *Expected f-vector of the poisson zero polytope and random convex hulls in the half-sphere*, Mathematika **66** (2020), no. 4, 1028–1053.
- Z. Kabluchko, A. Marynych, D. Temesvari, and C. Thäle, *Cones generated by random points on half-spheres and convex hulls of poisson point processes*, Probability Theory and Related Fields **175** (2019), no. 3, 1021–1061.
- P. M Long, *On the sample complexity of PAC learning half-spaces against the uniform distribution*, IEEE Transactions on Neural Networks **6** (1995), no. 6, 1556–1559.

References IV

 L. G. Valiant, *A theory of the learnable*, Communications of the ACM **27** (1984), no. 11, 1134–1142.

Lower Bounds from PAC

Definition (PAC learning sample complexity $m_{\mathcal{H},D}(\epsilon,\delta)$, informal [Val84] [BI91])

$m_{\mathcal{H},\ D}(\epsilon,\delta)$ is the minimum number of labeled samples needed to (with probability at least $1-\delta$) return a classifier with error probability at most ϵ, given that instances are drawn from $\mathcal{D}$ and labeled by some hypothesis in $\mathcal{H}$.

Proposition

Let L_T be the total number of label requests, M_T be the total number of misclassifications, and $C_T = L_T + M_T$. If $m_{\mathcal{H},D}(\epsilon,\delta) \in \Omega(f(\epsilon,\delta))$, then for some $c > 0, \Delta > 0$, all selective sampling algorithms have $\mathbb{E}C_T \in \Omega\left(\sum_{t=1}^{T} f'^{-1}(ct)\right)$, where $f' = f(\epsilon,\Delta)$.

Our Lower Bound

[Lon95, Theorem 1]

$m_{\mathcal{H},\text{Unif}S^d}(\epsilon, \delta) = \Omega(\frac{d}{\epsilon} + \frac{1}{\epsilon} \log \frac{1}{\delta})$.

Plugging this into last slide's Proposition gives

Proposition

If $\mathcal{X} = \mathbb{R}^d$ and the X_t are drawn i.i.d. from the uniform distribution on S^d, then all algorithms for selective sampling incur expected cost $\Omega(d \log T)$.

No Upper Bounds from PAC

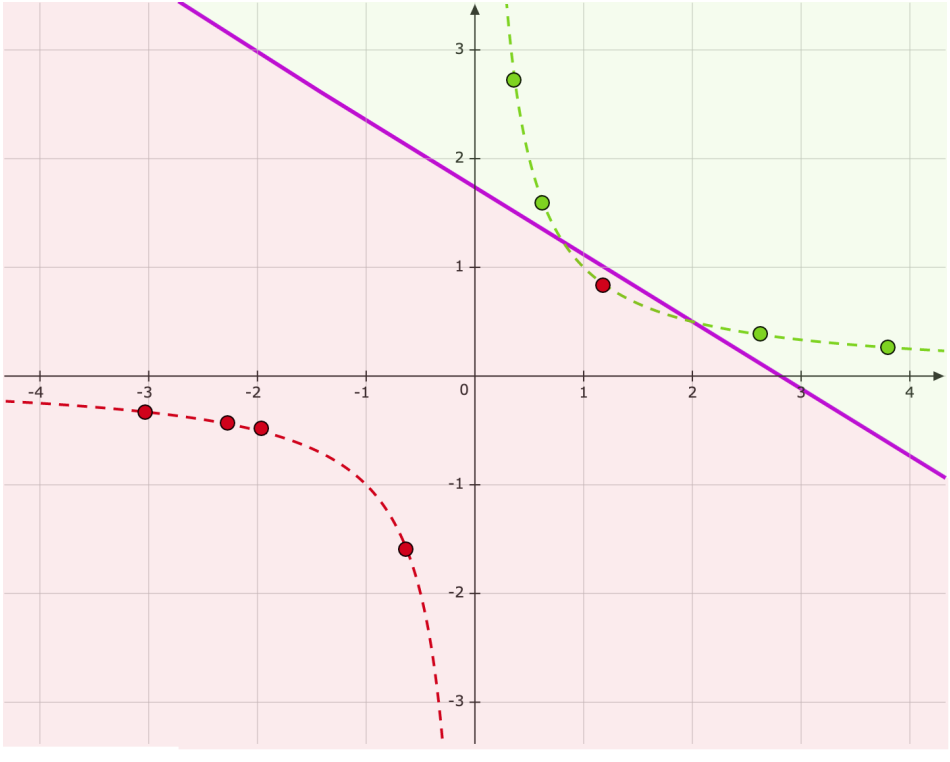


Figure: All points always undetermined