On the Convergence of Loss and Uncertainty-based Active Learning Algorithms

Milan Vojnovic

Department of Statistics

Joint work with Daniel Haimovich, Dima Karamshuk, Fridolin Linder, and Niek Tax

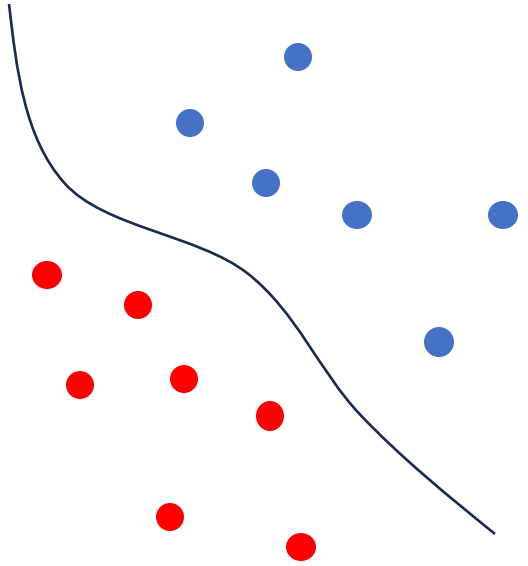
LSE Department of Statistics Research Showcase, 7-8 April 2025

Our problem

- For training ML models, one often has an abundance of unlabeled data, but obtaining the corresponding labels can be costly.
- Active learning (AL): Train an accurate ML model while minimizing labelling cost.
- A well-designed data acquisition strategy is essential for active learning !

Some unlabelled data examples:

Data acquisition strategies



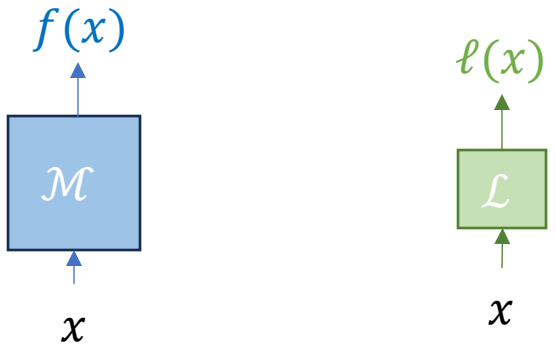
- Different data acquisition strategies:
  - Query-by-committee
  - Expected model change
  - Expected error reduction
  - Expected variance reduction
  - Mutual information
  - Uncertainty (used commonly, e.g. margin of confidence)
  - ⋮

- **Key question**: How does performance vary under different data acquisition strategies?

- **Related work**: Theoretical guarantees for the margin of confidence data acquisition criteria (Raj and Bach, 2022).

Loss-based data acquisition

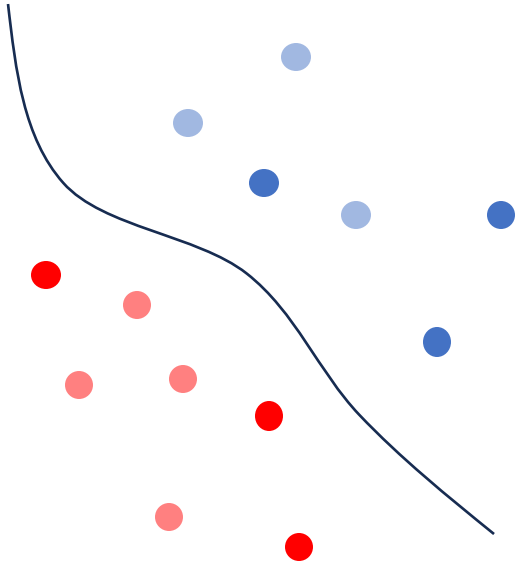
- Loss-based data acquisition strategies have been increasingly used in recent years.
  - In research: Yoo and Kweon (2019), Lahlou et al (2022), …
  - In industry: Applied at Meta for integrity violation classifiers.
- **Approach**: Utilise a loss value predictor to select points with high predicted loss value.



- **Key question**: How does model training converge under loss-based data acquisition?
- **Our work**: A theoretical analysis under the assumption that the loss predictor provides an unbiased estimate of the conditional expected loss of a data point.

Subset selection problem

- Also known also as **coreset selection** or **data pruning**.
- Given a labelled training dataset, the goal is to train a model using only a small subset of the training data.
- Loss-based strategies may also be relevant in this context!



Algorithm

- **Stochastic gradient decent (SGD) algorithm with adaptive filtering**

  The update rule:

$$\theta_{t+1} = \theta_t - z_t \nabla_\theta \ell(x_t, y_t, \theta_t)$$

stochastic step size (adaptive filtering)

Case 1: Bernoulli sampling w constant step size

$$z_t = \begin{cases} \gamma & \text{w. p. } \pi(x_t, y_t, \theta_t) \\ 0 & \text{otherwise} \end{cases}$$

sampling probability

Case 2: Bernoulli sampling w adaptive step size

$$z_t = \begin{cases} \zeta(x_t, y_t, \theta_t)/\pi(x_t, y_t, \theta_t) & \text{w. p. } \pi(x_t, y_t, \theta_t) \\ 0 & \text{otherwise} \end{cases}$$

expected step size

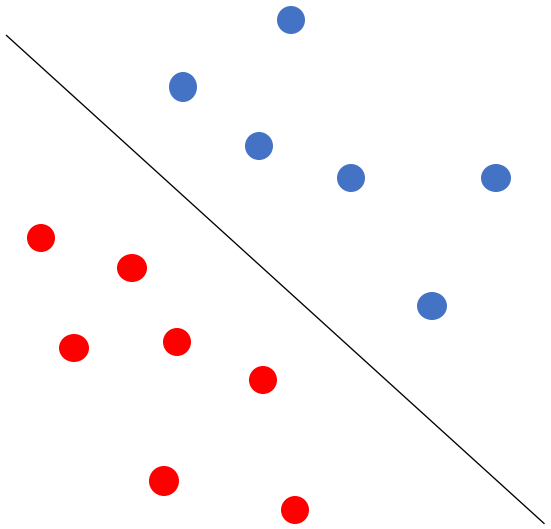
- **Key question**: What conditions on π guarantee a convergence rate?

Outline

- Convergence rates for linear classifiers and the hinge loss family of functions
- Convergence rates for more general cases
- An adaptive step-size algorithm

Linearly separable data and linear classifiers

- Binary classification: points in $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbf{R}^d$ and $\mathcal{Y} = \{-1,1\}$
- Linear separation: for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, there exists a $\theta^* \in \mathbf{R}^d$ such that $y = \text{sgn}(x^\top \theta^*)$
- Separation margin: for some $\rho^* \in \mathbf{R}_+$, $|x^\top \theta^*| \geq \rho^*$, for every $x \in \mathcal{X}$
- Bounded L2-norm points: for some $R \in \boldsymbol{R}$, $\|x\| \leq R$ for every $x \in \mathcal{X}$

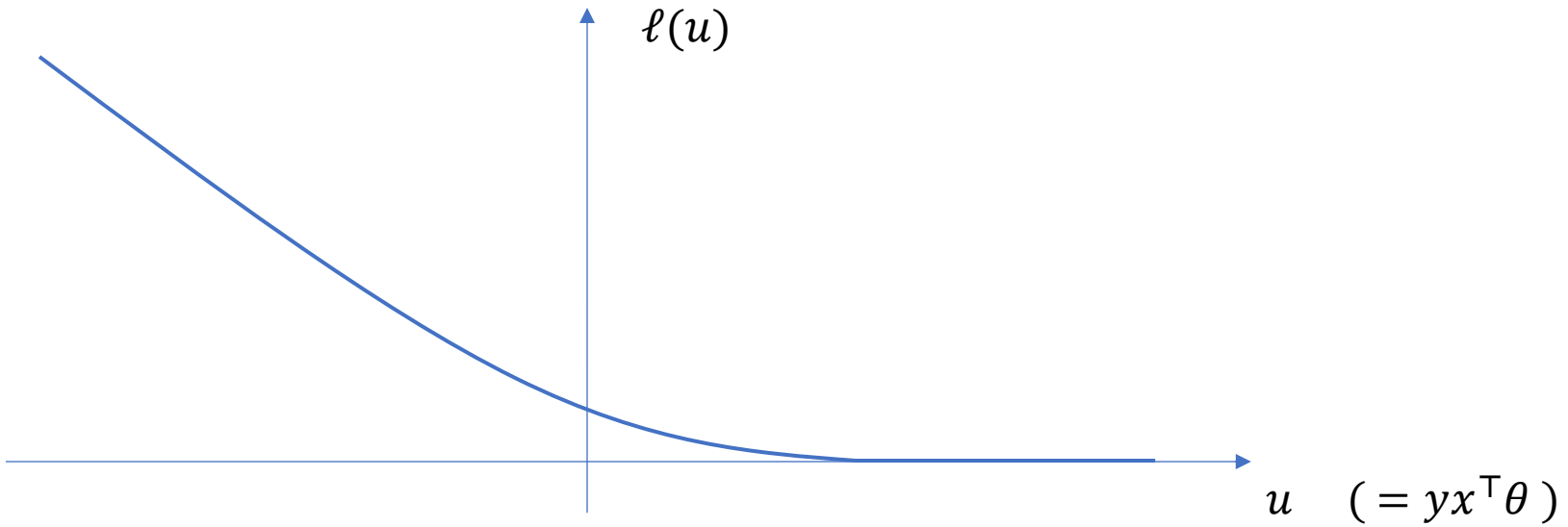


Linear classifiers: classification based on the value of the **margin** $x^\top \theta$

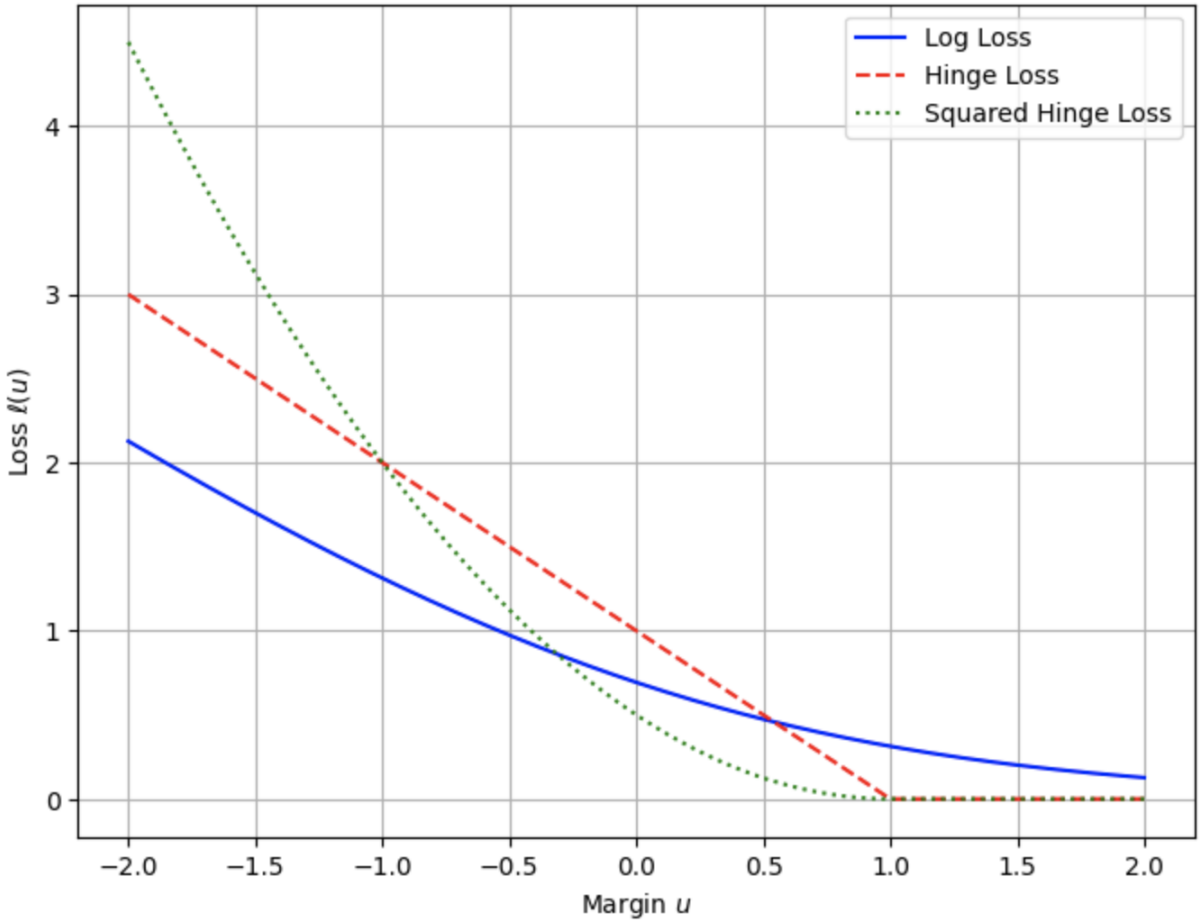
model parameter

A set of loss function conditions

- Loss function ℓ is assumed to satisfy the following conditions:
  - Continuously differentiable and convex on $(-\infty, 0]$
  - $\ell'(0) \leq -c_1$ for some constant $c_1 > 0$
  - $\lim_{u \to -\infty} \ell'(u) \geq -c_2$ for some constant $c_2 > 0$



Examples of loss functions



Log loss:

$$\ell(u) = \log(1 + e^{-u})$$

Hinge loss:

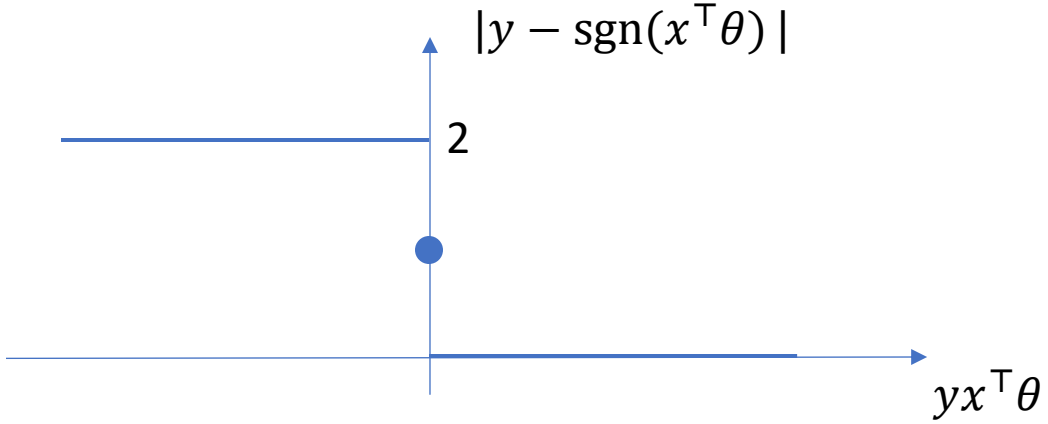
$$\ell(u) = \max\{1 - u, 0\}$$

Squared-hinge loss:

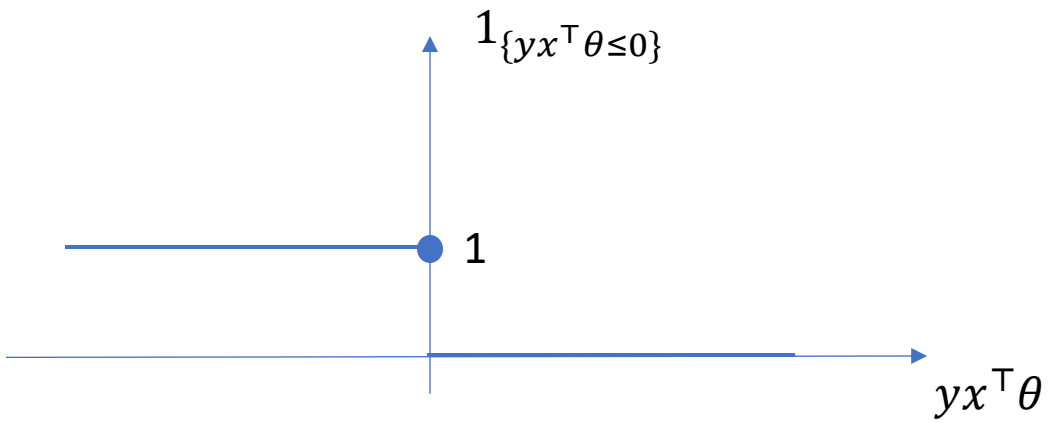
$$\ell(u) = \frac{1}{2}\max\{1 - u, 0\}^2$$

A warm-up: two loss-based strategies

Absolute error loss



Zero-one loss



A convergence rate bound

- **Theorem**. For sampling proportional to the zero-one loss with a constant factor ω, and setting $\gamma = c_1\rho^*/(c_2^2R^2)$, we have

$$\mathbf{E}[\sum_{t=1}^{n} \ell(y_t x_t^\top \theta_t)] \leq \frac{c_2^2 R^2 S^2}{c_1^2 \omega} \frac{1}{\rho^{*2}}.$$

- The same bound holds with an additional factor of 2 for sampling proportional to the absolute error loss with factor $\omega \in (0,1/2]$ and $\gamma = 2c_1\rho^*/(c_2^2R^2)$.

- Hereinafter, S is such that $\|\theta_1 - \theta^*\| \leq S$.

A key lemma

For all $u \in \mathbf{R}$:

- **C1**: $\pi(u)\ell'(u)^2 R^2 \leq \alpha\tilde{\ell}(u)$
- **C2**: $\pi(u)\big(-\ell'(u)\big)(\rho^* - u) \geq \beta\tilde{\ell}(u)$

ℓ : the training loss function

$\tilde{\ell}$: the evaluation loss function

Then, by setting $\gamma = \beta/\alpha$,

$$\mathbf{E}\left[\sum_{t=1}^{n} \tilde{\ell}(y_t x_t^\top \theta_t)\right] \leq S^2 \frac{\alpha}{\beta^2}$$

For a convex $\tilde{\ell}$ and $\bar{\theta}_t := (1/t)\sum_{s=1}^{t} \theta_s$, it holds: $\mathbf{E}\left[\tilde{\ell}(yx^\top \bar{\theta}_n)\right] \leq S^2 \frac{\alpha}{\beta^2}\frac{1}{n}$

Squared hinge loss function

Theorem. Assume that $\rho^* > 1$ and the training and evaluation loss functions are squared hinge loss functions. Furthermore, assume that the sampling probability function satisfies, for some constant $\beta \in (0,2]$, the following conditions: $\pi(u) \leq \beta/2$ for all $u \leq 1$ and

$$\pi(u) \geq \pi^*\big(\ell(u)\big) \coloneqq \frac{\beta}{2}\left(1 - \frac{1}{1+\mu\sqrt{\ell(u)}}\right)$$

where $\mu \geq \sqrt{2}/(\rho^* - 1)$.

Then, by setting $\gamma = 1/R^2$, we have

$$\mathbf{E}[\ell(yx^\top \bar{\theta}_t)] \leq \mathbf{E}\left[\frac{1}{n}\sum_{t=1}^{n} \ell(y_t x_t^\top \theta_t)\right] \leq R^2 S^2 \frac{1}{\beta}\frac{1}{n}.$$

Squared hinge loss function (cont'd)

- For sampling according to π^*, the expected total number of sampled points is bounded as:

$$\mathbf{E}\left[\sum_{t=1}^{n} \pi^*(\ell(y_t x_t^\top \theta_t))\right] \leq \min\left\{\frac{1}{2} RS\mu\sqrt{\beta}\sqrt{n}, \frac{1}{2}\beta n\right\}$$

- Hence, we have a sublinear expected number of sampled points of order $\sqrt{n}$.

Outline

- Convergence rates for linear classifiers and hinge loss family of functions
- Convergence rates for more general cases
- An adaptive step size algorithm

Covering a larger class of loss functions

- Following Liu and Li (2023), the algorithm is an SGD algorithm with the "equivalent" loss function, whose gradient is:

$$\nabla_\theta \tilde{\ell}(\theta) = \mathbf{E}[\pi(x, y, \theta) \nabla_\theta \ell(x, y, \theta)]$$

- Assume that π is a function of the expected conditional loss:

$$\ell(x, \theta) \coloneqq \mathbf{E}[\, \ell(x, y, \theta) \mid x \,]$$

- Then, $\tilde{\ell}(\theta) = \mathbf{E}[\Pi(\ell(x, \theta))]$, where Π is the primitive function of π.

A convergence rate bound

Assume that:

- ℓ is a convex function,
- $\tilde{\ell}$ is L-smooth, and
- $\mathbf{E}\left[\pi\left(\ell(x,\theta)\right)\|\nabla_\theta \ell(x,y,\theta)\|^2\right] - \left\|\nabla_\theta \tilde{\ell}(\theta)\right\|^2 \leq \sigma_\pi^2.$

Then, with $\gamma = 1/(L + \left(\frac{\sigma_\pi}{R}\right)\sqrt{n/2})$, we have

$$\mathbf{E}[\ell(\bar{\theta}_n)] \leq \mathbf{E}[\textstyle\sum_{t=1}^n \ell(\theta_t)] \leq \inf_\theta \Pi^{-1}\left(\tilde{\ell}(\theta)\right) + \Pi^{-1}\left(\frac{\sqrt{2}S\sigma_\pi}{\sqrt{n}}\right) + \Pi^{-1}\left(\frac{LS^2}{n}\right).$$

Examples of sampling probability functions

$\pi(x)$	$\Pi(x)$	$\Pi^{-1}(x)$
$1-e^{-x}$	$x+e^{-x}-1$	$\approx \sqrt{2x}$ for small x
$\min\{x,1\}$	$\begin{cases} \frac{1}{2}x^2 & x \le 1 \\ x-\frac{1}{2} & x \ge 1 \end{cases}$	$\begin{cases} \sqrt{2x} & x \le 1/2 \\ x+\frac{1}{2} & x \ge 1/2 \end{cases}$
$\min\{(x/b)^a,1\}, a>0, b>0$	$\begin{cases} \frac{1}{b^a(1+a)}x^{1+a} & x \le a \\ x-\frac{a}{1+a} & x \ge a \end{cases}$	$\begin{cases} b^{\frac{a}{1+a}}(1+a)^{\frac{1}{1+a}}x^{\frac{1}{1+a}} & x \le \frac{b}{1+a} \\ x+\frac{a}{1+a}b & x \ge \frac{b}{1+a} \end{cases}$
$1-\frac{1}{1+\mu x}$	$x-\frac{1}{\mu}\log(1+\mu x)$	$\approx \sqrt{(2/\mu)x}$ for small x
$1-\frac{1}{1+\mu\sqrt{x}}$	$x-\frac{2}{\mu}\sqrt{x}+\frac{2}{\mu^2}\log(1+\mu\sqrt{x})$	$\approx (((3/2)/\mu)x)^{2/3}$ for small x

Outline

- Convergence rates for linear classifiers and hinge loss family of functions
- Convergence rates for more general cases
- An adaptive step size algorithm

Sampling with stochastic Polyak's step size

- Step size: $\gamma = \zeta(x, y, \theta)/\pi(x, y, \theta)$ w. p. $\pi(x, y, \theta)$, 0 otherwise.

- $\zeta(x, y, \theta) = \beta \min\left\{\frac{\ell(x,y,\theta)}{\|\nabla_\theta \ell(x,y,\theta)\|^2}, \rho\right\}$

- When $\pi(x, y, \theta) \equiv 1$, then $\zeta(x, y, \theta)$ is the stochastic Polyak's step size (Loizou et al 2021), which has been shown to be efficient both theoretically and experimentally.

- Note: For simplicity, in the slides, we assume that $\inf_{\theta'} \ell(x, y, \theta') = 0$ for every x, y.

Convergence rate guarantee

Theorem. Assume that ℓ is a convex and L-smooth function, and

$$\pi(x,y,\theta) \geq \frac{\beta}{2(1-c)} \min\left\{\rho \frac{\|\nabla_\theta \ell(x,y,\theta)\|^2}{\ell(x,y,\theta)}, 1\right\}.$$

Then,

$$\mathbf{E}\left[\frac{1}{n}\sum_{t=1}^{n} \ell(x_t,y_t,\theta_t)\right] \leq \frac{\rho\beta}{c\kappa}\mathbf{E}[\ell(x,y,\theta^*)] + \frac{1}{2c\kappa} S^2 \frac{1}{n}$$

where $\kappa = \beta \min\{1/(2L), \rho\}$.

Binary classification: linear classifier case

- Sufficient condition for the sampling probability function:

$$\pi(x, y, \theta) \geq \frac{\beta}{2(1-c)} \min\{\rho R^2 h(y x^\top \theta), 1\}$$

  where $h(u) = \ell'(u)/\ell(u)$.

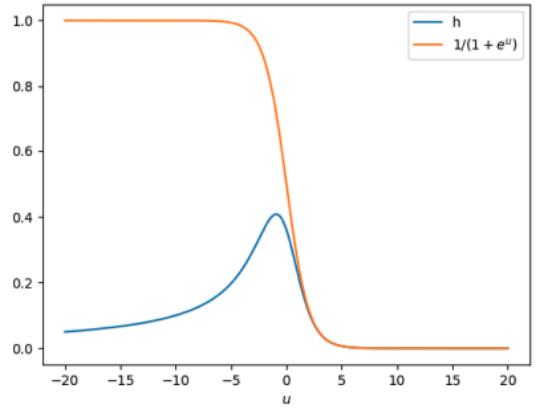
- For logistic regression and cross-entropy loss:

$$h(u) = \frac{1}{(1+e^u)^2 \log(1+e^{-u})}$$

Cases when the condition holds true

- **Loss-based sampling according to the absolute error loss**:

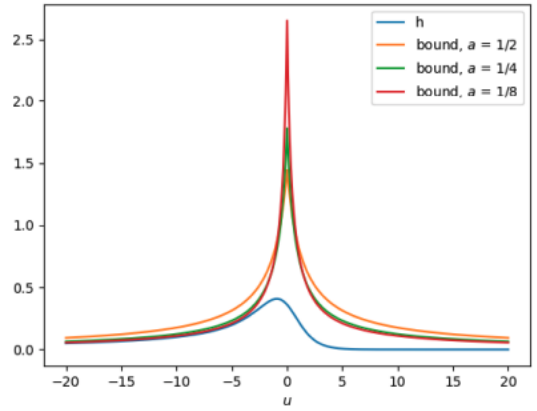
$$\pi^*(u) = \omega(1 - \sigma(u))$$



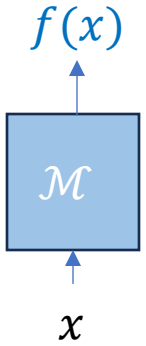
- **Uncertainty-based sampling according to**:

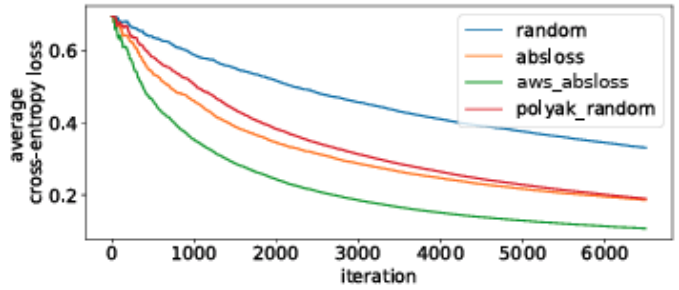
$$\pi^*(u) = \frac{\beta}{2(1-c)} \min\left\{\rho R^2 \frac{1}{c_a + (1-a)|u|}, 1\right\}$$

where $a \in (0, 1/2]$ is a hyper-parameter.

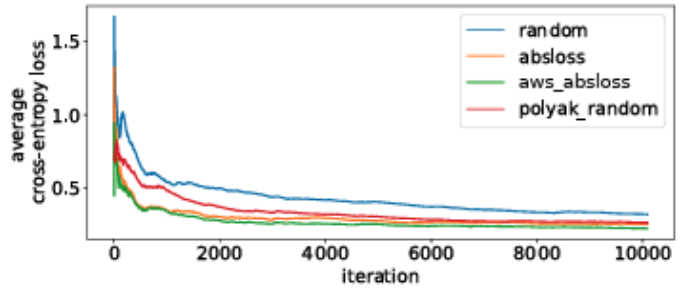


Experimental results: using true loss values

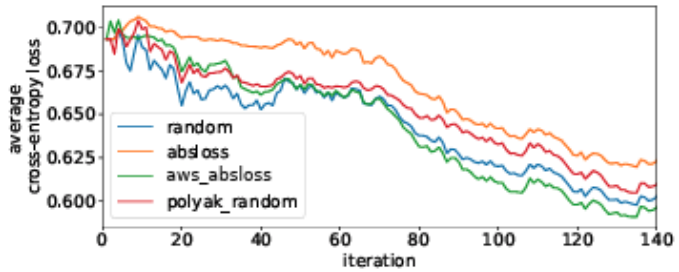




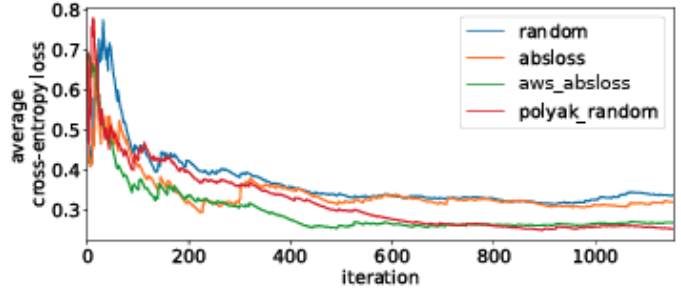
(a) Mushrooms dataset



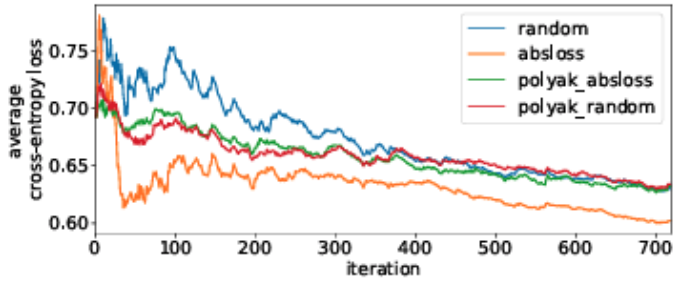
(b) MNIST binary classification of 3s vs 5s



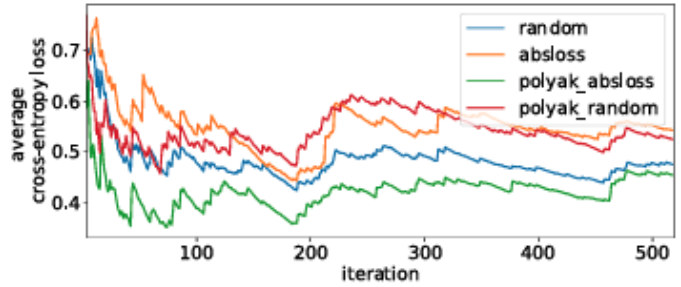
(c) Parkinsons dataset



(d) Splice dataset

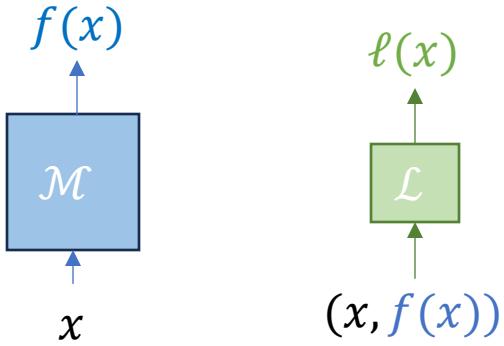


(e) Tictactoe dataset

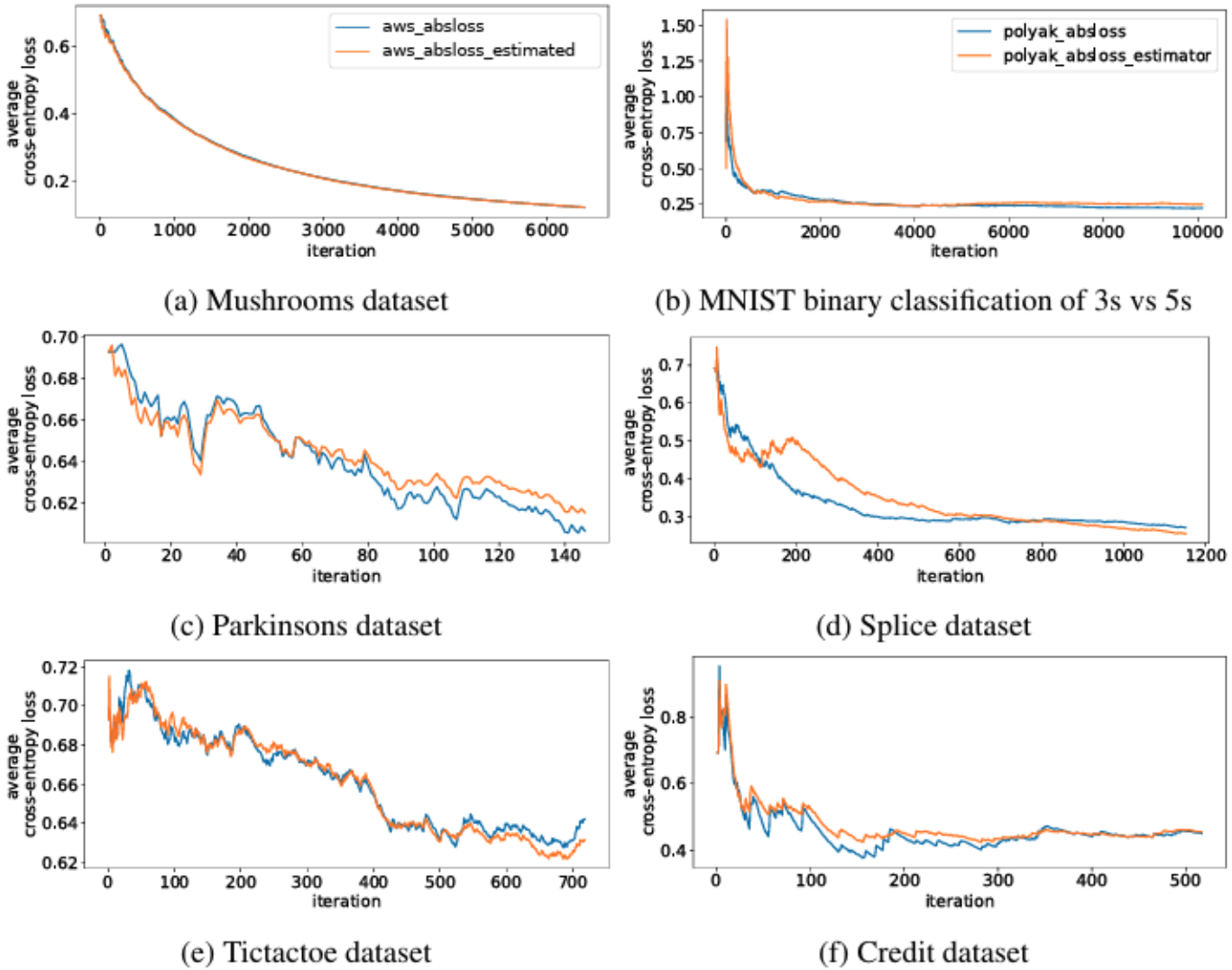


(f) Credit dataset

Experimental results: using predicted loss values



Loss predictor:
random forest



(a) Mushrooms dataset

(b) MNIST binary classification of 3s vs 5s

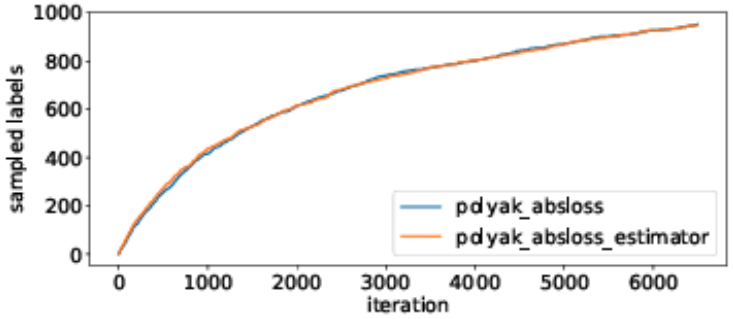
(c) Parkinsons dataset

(d) Splice dataset

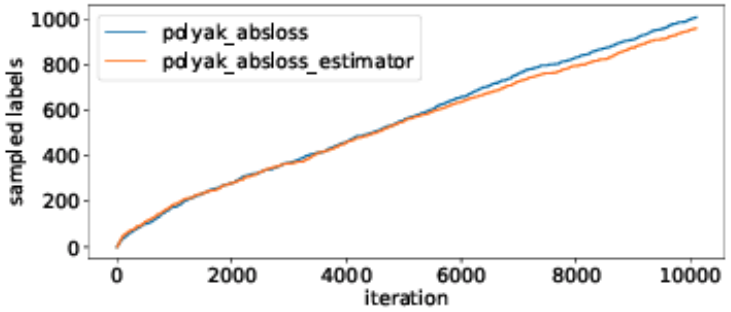
(e) Tictactoe dataset

(f) Credit dataset

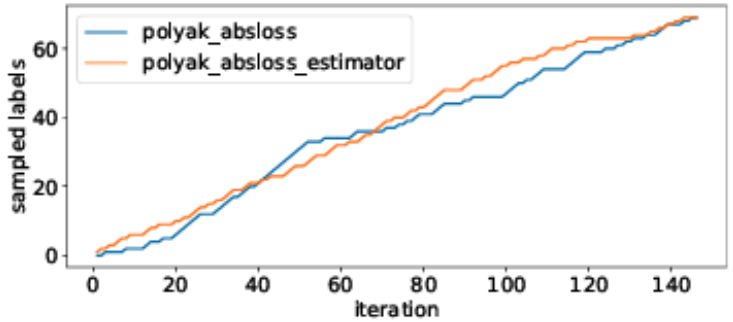
Sampling efficiency



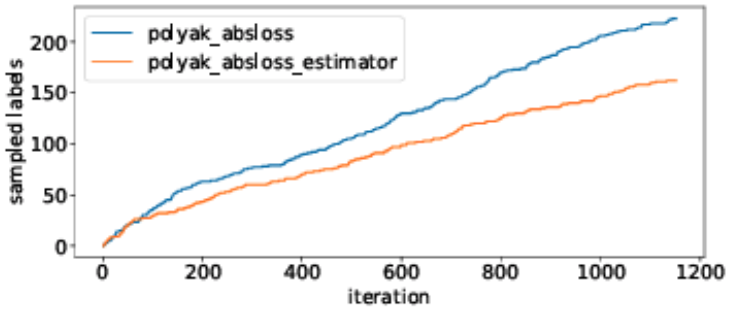
(a) Mushrooms dataset



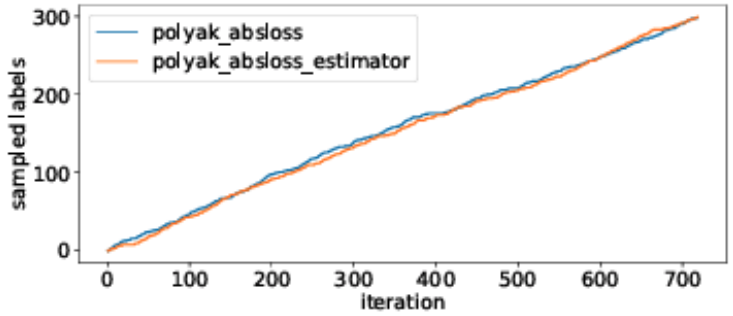
(b) MNIST binary classification of 3s vs 5s



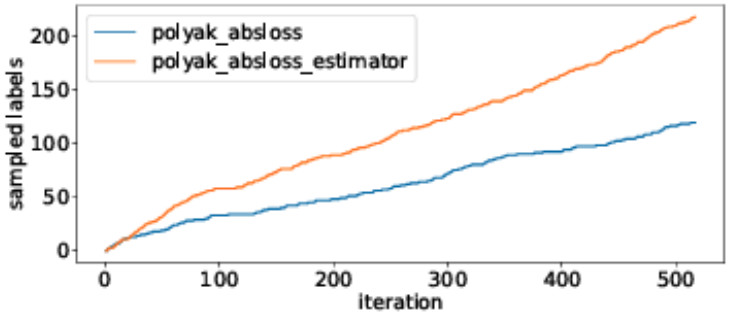
(c) Parkinsons dataset



(d) Splice dataset



(e) Tictactoe dataset



(f) Credit dataset

Conclusion

- Showed convergence rates for loss- and uncertainty-based active learning strategies under various conditions.
- Showed an algorithm using stochastic Polyak's step size in expectation.
- The results provide insights into sufficient conditions for loss- and uncertainty-based strategies to guarantee convergence rates.
- Future work: Tighter bounds on convergence rates? The effect of loss prediction noise? Generalisation bounds?