Active Learning of Deep Neural Networks via Gradient-Free Cutting Planes

Erica Zhang*, Fangzhao Zhang*, Mert Pilanci

Stanford | ENGINEERING

Proceedings of the 42nd International Conference on Machine Learning
Vancouver, Canada, 2025

June 16, 2025

Overview

- Introduction
  - Active Learning of Deep Networks
  - The Cutting-plane Method
  - Linear Cutting-Plane Active Learning
- The Cutting-Plane Active Learning (CPAL) Method
  - The CPAL Algorithm
  - Theorem 4.2 (linear to non-linear)
  - Theorem 6.3 (Convergence Guarantees)
- Experiments
  - Synthetic
  - Sentiment Analysis
- Summary

Introduction

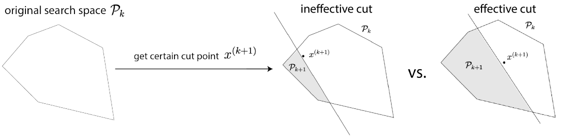
Active Learning of Deep Networks

Active Learning (AL) is a data acquisition paradigm for supervised learning. Active learning aims at training a model $\mathcal{X} \to \mathcal{Y} : f(x; \theta) = y$, characterized by its parameters θ, with the *most informative* data $\subseteq \mathcal{X} \times \mathcal{Y}$.

- As deep networks scale, training efficiency increasingly hinges on selecting data that contributes the most to learning, rather than relying on random samples.
  - A key class of active learning algorithms selects the next point $(x, y)_{k+1}$ to *maximize model uncertainty reduction*, leveraging the model trained on the current labeled set $\{(x_i, y_i)\}_{i=1}^{k}$.
- However, the *nonconvexity* of deep neural networks poses challenges for active learning.
  - **Challenge 1:** Existing deep AL methods lack convergence guarantees and many rely heavily on heuristics.

The Cutting-Plane Method

Cutting-plane methods are first introduced by Gomory, 1958 [1] for LPs and now heavily used in commercial MILP solvers.



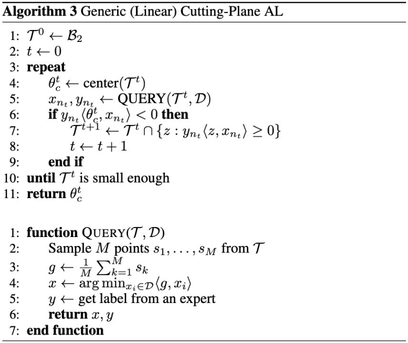
(adapted from Boyd and Vandenberghe, 2018)

Figure 2: The Cutting-plane method

- Consider any minimization problem with objective $f(\theta)$ whose solution set Θ is a convex set.
- Assuming existence of an oracle such that given any input θ_0, we receive a cutting plane that cuts between the current input θ_0 and the desired solution set Θ.
- Such cut is given by subgradient of f at query point θ_0.

Linear Cutting-Plane Active Learning

- **The linear cutting-plane AL.** Louche & Ralaivola, 2015 [2] proposes the first use of cutting-plane method in the context of active learning.
  - **Challenge 2:** Restricted to AL for shallow (linear) models.
  - Allows only linear cuts and binary classifications.
  - *Our proposed Cutting-Plane Active Learning (CPAL) method addresses these challenges while preserving the theoretical convergence guarantees inherent to cutting-plane methods.*



Algorithm 3 Generic (Linear) Cutting-Plane AL

```
$\mathcal{T}^0 \leftarrow \mathcal{B}_2$
$t \leftarrow 0$
repeat
    $\theta_c^t \leftarrow$ center($\mathcal{T}^t$)
    $x_{n_t}, y_{n_t} \leftarrow$ QUERY($\mathcal{T}^t, \mathcal{D}$)
    if $y_{n_t}\langle \theta_c^t, x_{n_t}\rangle < 0$ then
        $\mathcal{T}^{t+1} \leftarrow \mathcal{T}^t \cap \{z : y_{n_t}\langle z, x_{n_t}\rangle \geq 0\}$
        $t \leftarrow t + 1$
    end if
until $\mathcal{T}^t$ is small enough
return $\theta_c^t$
```

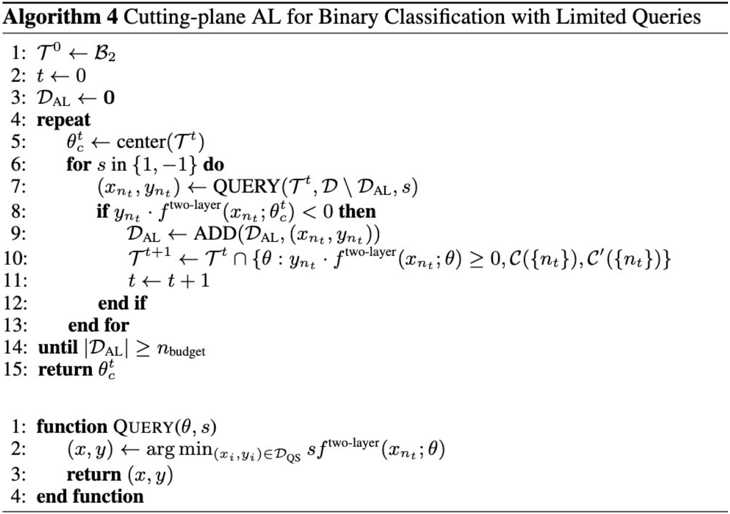
```
function QUERY($\mathcal{T}, \mathcal{D}$)
    Sample $M$ points $s_1, \ldots, s_M$ from $\mathcal{T}$
    $g \leftarrow \frac{1}{M}\sum_{k=1}^{M} s_k$
    $x \leftarrow \arg\min_{x_i \in \mathcal{D}} \langle g, x_i\rangle$
    $y \leftarrow$ get label from an expert
    return $x, y$
end function
```

The Cutting-Plane Active Learning (CPAL) Method

The Cutting-Plane Active Learning (CPAL) Method

- **The Cutting-Plane Active Learning (CPAL) method** addresses the **two** aforementioned key challenges:
  - **Theorem 4.2 (linear to non-linear):** Reformulates deep ReLU network training as a linear program via activation patterns.
    - Extends to *multi-class classification* (via one-vs-rest decomposition) and *regression* (by adapting constraints to bound prediction error).
    - **Contribution 1:** Generalizes beyond the linearity of traditional cutting-plane methods to nonlinear models.
  - **Theorem 6.3 (convergence guarantees):** Establishes volumetric convergence guarantees in parameter space.
    - Preserves the classic $(1 - 1/e)$ volumetric shrinkage guarantee of cutting-plane methods with cuts through center of gravity [3].
    - Two-layer NN prediction functions converge in norm to the optimal decision boundary under CPAL (Corollary 6.4).
    - **Contribution 2:** Presents the first deep AL method with convergence guarantees.

The CPAL Algorithm



Algorithm 4 Cutting-plane AL for Binary Classification with Limited Queries

```
$\mathcal{T}^0 \leftarrow \mathcal{B}_2$
$t \leftarrow 0$
$\mathcal{D}_{\text{AL}} \leftarrow \mathbf{0}$
repeat
    $\theta_c^t \leftarrow \text{center}(\mathcal{T}^t)$
    for $s$ in $\{1, -1\}$ do
        $(x_{n_t}, y_{n_t}) \leftarrow \text{QUERY}(\mathcal{T}^t, \mathcal{D} \setminus \mathcal{D}_{\text{AL}}, s)$
        if $y_{n_t} \cdot f^{\text{two-layer}}(x_{n_t}; \theta_c^t) < 0$ then
            $\mathcal{D}_{\text{AL}} \leftarrow \text{ADD}(\mathcal{D}_{\text{AL}}, (x_{n_t}, y_{n_t}))$
            $\mathcal{T}^{t+1} \leftarrow \mathcal{T}^t \cap \{\theta : y_{n_t} \cdot f^{\text{two-layer}}(x_{n_t}; \theta) \geq 0, \mathcal{C}(\{n_t\}), \mathcal{C}'(\{n_t\})\}$
            $t \leftarrow t + 1$
        end if
    end for
until $|\mathcal{D}_{\text{AL}}| \geq n_{\text{budget}}$
return $\theta_c^t$
```

```
function QUERY($\theta$, $s$)
    $(x, y) \leftarrow \arg\min_{(x_i, y_i) \in \mathcal{D}_{\text{QS}}} s f^{\text{two-layer}}(x_{n_t}; \theta)$
    return $(x, y)$
end function
```

Theorem 4.2 (Linear to Non-Linear)

Theorem 4.2 (Linear to Non-Linear)

Consider a ReLU network with n hidden layers for binary classification:

$$\begin{aligned} \text{find} \quad & W_1, W_2, \ldots, W_{n+1} \\ \text{s.t.} \quad & y \odot (\cdots(((XW_1)_+ W_2)_+ \cdots)_+ W_n)_+ W_{n+1} \geq 1. \end{aligned} \tag{1}$$

Using the notations established in the paper, then when $m_i \geq \prod_{i=1}^{n} 2P_i$ for each $i \in [n]$, Problem (1) is equivalent to:

$$\begin{aligned} \text{find} \quad & u_{j_n j_{n-1}\cdots j_1}^{c_n c_{n-1}\cdots c_1} \\ \text{s.t.} \quad & y \odot \sum_{j_n=1}^{P_n} D_{j_n}^{(n)} \left(\mathcal{T}_1^{(n-1)}(D^{(n-1)}) - \mathcal{T}_2^{(n-1)}(D^{(n-1)}) \right) \geq 1 \\ & (2D_{j_i}^{(i)} - I)\mathcal{T}_{c_{n-1}\cdots c_{i-1}}^{(n-1)\ldots(i-1)}(D^{(i-1)}) \geq 0, \quad 2 \leq i \leq n \\ & (2D_{j_1}^{(1)} - I)Xu_{j_n\cdots j_1}^{c_n\cdots c_1} \geq 0. \end{aligned} \tag{2}$$

where $c_i \in \{1, 2\}$, and

$$\mathcal{T}_{c_{n-1}\cdots c_i}^{(n-1)\ldots(i)}(D^{(i)}) = \sum_{j_i=1}^{P_i} D_{j_i}^{(i)} \left(\mathcal{T}_{c_{n-1}\cdots c_i 1}^{(n-1)\ldots(i)(i-1)}(D^{(i-1)}) - \mathcal{T}_{c_{n-1}\cdots c_i 2}^{(n-1)\ldots(i)(i-1)}(D^{(i-1)}) \right), \quad \forall i \leq n-1$$

$$\mathcal{T}_{c_{n-1}\cdots c_1}^{(n-1)\ldots(1)}(D^{(1)}) = \sum_{j_1=1}^{P_1} D_{j_1}^{(1)} X \left(u_{j_n\cdots j_1}^{1c_{n-1}\cdots c_1} - u_{j_n\cdots j_1}^{2c_{n-1}\cdots c_1} \right).$$

Theorem 6.3 (Convergence Guarantees)

Theorem 6.3 (Convergence via Center of Gravity)

Let $\mathcal{T} \subseteq \mathbb{R}^d$ be a convex body and let θ_G denote its center of gravity. The polyhedron cut given in Algorithm 4 (CPAL), i.e.,

$$\mathcal{T} \cap \left\{ \theta : y_n \cdot f^{\text{two-layer}}(x_n; \theta) \geq 0,\ \mathcal{C}(\{n\}),\ \mathcal{C}'(\{n\}) \right\},$$

where the pair (x_n, y_n) is the data point returned by the cutting-plane oracle after receiving the query point θ_G, partitions $\mathcal{T}$ into two subsets:

$$\mathcal{T}_1 := \left\{ \theta \in \mathcal{T} : y_n \cdot f^{\text{two-layer}}(x_n; \theta) \geq 0,\ \mathcal{C}(\{n\}),\ \mathcal{C}'(\{n\}) \right\},$$
$$\mathcal{T}_2 := \left\{ \theta \in \mathcal{T} : y_n \cdot f^{\text{two-layer}}(x_n; \theta) < 0\ \vee\ \neg\mathcal{C}(\{n\})\ \vee\ \neg\mathcal{C}'(\{n\}) \right\},$$

where $\neg$ denotes the complement. Then $\mathcal{T}_1$ satisfies:

$$\textbf{vol}(\mathcal{T}_1) < \left(1 - \frac{1}{e}\right) \cdot \textbf{vol}(\mathcal{T}).$$

Experiments

Experiments (Synthetic)

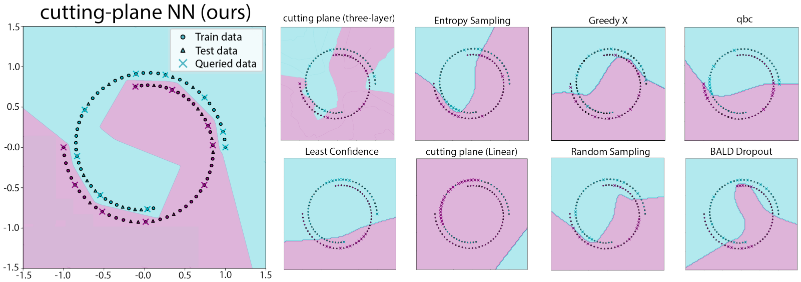


Figure 3: Decision boundaries for binary classification on the spiral dataset for the cutting-plane AL method using a two-layer ReLU neural network, alongside various deep AL baselines. For compactness, we also include the decision boundaries for the cutting-plane AL method with a three-layer ReLU network in the collage to demonstrate its feasibility. For fairness of comparison, we use the same two-layer ReLU network structure and embedding size of 623 for all methods. We enforce the same hyperparameters for all deep AL baselines and select the best performing number of training epochs at 2000 and a learning rate at 0.001 to ensure optimal performance.

Experiments (Sentiment Analysis)

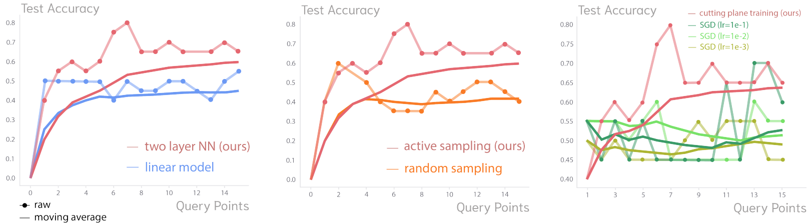


Figure 4: Sentiment analysis on IMDB movie review dataset with two-layer ReLU model. We take Phi-2 embedding as our training features and compare with various baselines. The result shows that the introduction of non-linearity improves upon linear model performance, our active sampling scheme effectively identifies valuable training points compared to random sampling, and our cutting-plane training scheme is more effective than SGD in this setting. Linear and our reframed two-layer models are initialized to predict zero while two-layer NN trained with SGD has random weight initialization, thus starting from non-zero prediction.

Summary

Summary

- We generalize classic cutting-plane methods from linear models to nonlinear deep ReLU networks via activation pattern enumeration.
- We propose the first deep active learning algorithm with formal convergence guarantees.
- We establish the viability of gradient-free training through cutting-plane techniques.
- Our cutting-plane active learning method (CPAL) outperforms popular deep AL baselines on both synthetic and real-world datasets.
- While GPU-based CPAL currently faces scalability limitations, it lays a theoretical foundation for deep AL and offers a promising direction as LP solvers and activation pattern sampling continue to improve.

Thank you!

Bibliography

[1] R. E. Gomory, "An algorithm for integer solutions to linear programs,", 1958. [Online]. Available: https://api.semanticscholar.org/CorpusID:116324171.

[2] L. Ralaivola and U. Louche, *From cutting planes algorithms to compression schemes and active learning*, 2015. arXiv: 1508.02986 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1508.02986.

[3] B. Grünbaum, "Partitions of mass-distributions and of convex bodies by hyperplanes," *Journal of Mathematical Analysis and Applications*, vol. 10, no. 4, pp. 1257–1261, 1960, Received January 22, 1960. This research was supported by the United States Air Force through the Air Force Office of Scientific Research of the Air Research and Development Command, under contract No. AF49(638)-253.