# Active learning for maximum likelihood estimation

## Theory and neural implementation

Geelon So, agso@eng.ucsd.edu
Time series reading group — September 9, 2021

# I. Theory

# Active learning framework

In **active learning**, the learner:

- has access to a large pool of *unlabelled samples*
- can *interactively ask for labels*

# Maximum likelihood estimation (MLE)

Consider a model class $\Theta$ where each $\theta \in \Theta$ corresponds to the model:

$$p(y|x;\theta),$$

where $x$ is a sample and $y$ is a labelling.

- Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, identify the model $\theta$ that most likely generated it.

# Question

Is there an active learning strategy to do MLE?

- If an assumption* is made, then yes (Chaudhuri et al., 2015).
- Idealized selection strategy is intractable.
- There is a SDP relaxation that is almost statistically optimal.
  - SDP infeasible in high dimensions—algorithm in the neural regime? (Ash et al., 2021).

*__Assumption:__ the Fisher information matrix for $\theta$ at any $(x, y)$ depends only on $x$ and $\theta$.

# Formal setting

**Notation**

- instance space $\mathcal{X}$
- label space $\mathcal{Y}$
- family of models $p(y|x;\theta)$ parametrized over $\theta \in \Theta$
- unlabelled pool $U = \{x_1, \ldots, x_n\}$
- label oracle $\textsc{label}(x) \sim p(y|x;\theta^*)$ where $\theta^* \in \Theta$ is unknown
- loss function $\ell(x, y; \theta) = -\log p(y|x;\theta)$

**Problem**

Minimize: $$\mathrm{loss}(\hat{\theta}) = \mathop{\mathbb{E}}_{X \sim \mathrm{Unif}(U)} \mathop{\mathbb{E}}_{Y \sim p(y|X;\theta^*)} \ell(x, y; \hat{\theta})$$

- $\hat{\theta}$ is the parameter estimated from data obtained through active sampling strategy
- the *fixed design* or *transductive* setting: we treat $U$ as the entire distribution

# Fisher information matrix

## Definition (Fisher information)

*Let $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $\theta \in \Theta \subset \mathbb{R}^d$. The **Fisher information matrix** is:*

$$I(x, y; \theta) = \nabla^2_\theta \ell(x, y; \theta).$$

## Assumption

*For any $x, y, \theta$, the Fisher information matrix is a function of only $x$ and $\theta$,*

$$I(x, y; \theta) = I_{x;\theta}.$$

- Notation: let $I_{\mu;\theta} = \mathop{\mathbb{E}}_{X \sim \mu} I_{x;\theta}$ and $I_{U;\theta} = I_{\mathrm{Unif}(U);\theta}$.

# Intuition for the Fisher information matrix

## Claim (Duchi (2019))

*The Taylor expansion is on the order:*

$$\mathop{\mathbb{E}}_{Y \sim p_{y|x;\theta^*}} [\ell(x, Y; \theta^* + d\theta)] \approx \text{constant} + \frac{1}{2} \text{tr} \left( I_{x;\theta} d\theta d\theta^\top \right).$$

- The claim makes use of the identity: $\mathop{\mathbb{E}}_{Y \sim p_{y|x;\theta}} [\nabla_\theta \ell(x, Y; \theta)] = 0.$
- This means that in expectation w.r.t. $\theta^*$, the signal for $\ell(x, y; \theta)$ is greatest along the largest eigendirections of $I_{x;\theta}$.
  - Learning $p(y|x; \theta^*)$ for $x$ will yields information about $\theta^*$ along certain directions of $\theta$.

# Fisher information bounds error

## Lemma (Informal, (Chaudhuri et al., 2015))

*Let $\mu$ be a distribution of $U$. Let $\hat{\theta}_{\mu}^{(m)}$ be the MLE estimator for $\theta^*$,*

$$\hat{\theta}_{\mu}^{(m)} = \arg\min_{\theta \in \Theta} \frac{1}{m} \sum_{i=1}^{m} \ell(X_i, Y_i; \theta),$$

*where $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} \mu$ and $Y_i \sim p(y|X_i; \theta^*)$. Then:*

$$\mathbb{E}\left[\text{loss}(\hat{\theta}_{\mu}^{(m)}) - \text{loss}(\theta^*)\right] \sim \frac{1}{m} \text{tr}\left(I_{\mu;\theta^*}^{-1} I_{U;\theta^*}\right).$$

- **Read:** the error of the estimator learned from samples drawn from $\mu$ versus the true distribution $\text{Unif}(U)$ is controlled by $\text{tr}\left(I_{\mu;\theta^*}^{-1} I_{U;\theta^*}\right)$.

# Active learning strategy

Iteratively do the following:

- Use an estimate $\theta_t$ of $\theta^*$ to construct sampling distribution $\mu_{t+1}$,

$$\mu_{t+1} \leftarrow \underset{\mu \text{ distribution over } U}{\arg\min} \operatorname{tr}\left(I^{-1}_{\mu;\theta_t} I_{U;\theta_t}\right).$$

- Sample batch of unlabelled data points $X_1, \ldots, X_B \sim \mu_{t+1}$.
- Query labels $Y_i \sim p(y|X_i;\theta^*)$.
- Update estimator $\theta_{t+1} \leftarrow \text{MLE}\left(\{X_i, Y_i\}_{i=1}^B\right)$.

# Theoretical guarantee

**Theorem (Near-optimal strategy, Chaudhuri et al. (2015))**

*Under regularity assumptions, there is a polynomial active learning strategy with excess loss:*

$$\text{loss}(\hat{\theta}^{(m)}) - \text{loss}(\theta^*) = \widetilde{O}\left(\frac{1}{m}\text{tr}\left(I^{-1}_{\mu^*;\theta^*} I_{U;\theta^*}\right)\right),$$

*where $\mu^*$ is the optimal sampling distribution on $U$.*

# Remaining questions about computation

Though the algorithm in Chaudhuri et al. (2015) is polynomial, it involves solving SDPs where the dimensionality $d$ corresponds to $\theta \in \mathbb{R}^d$. Ideally, finding $\mu_{t+1}$ means solving:

$$\underset{S \subset U, |S| \leq B}{\arg\min} \ \operatorname{tr}\left[ \left( \sum_{x \in S} I_{x;\theta_t} \right)^{-1} I_{U;\theta_t} \right].$$

- Since this is infeasible if $d$ is large, this motivates Ash et al. (2021).

# II. Neural implementation

# Neural networks as the model class

Let $\theta \in \mathbb{R}^d$ correspond to the parameters of a neural network.

- Each neural net $\theta$ computes the conditional distribution $p(y|x;\theta)$.
- Since $d$ is typically very large, when computing the Fisher information, just consider the last layer:

$$I_{x;\theta^L} = \mathop{\mathbb{E}}_{Y \sim p(y|x;\theta)} \nabla^2_{\theta^L} \ell(x, Y; \theta).$$

# Greedy sample selection

Recall the goal is to find a set $S \subset U$ minimizing:

$$\operatorname{tr}\left[\left(\sum_{x \in S} I_{x;\theta_t^L}\right)^{-1} I_{U;\theta_t^L}\right].$$

- In a greedy sample selection process, at iteration $n$, add the point $S_n \leftarrow S_{n-1} \cup \{x_n\}$ that minimizes:

$$\operatorname{tr}\left[\left(\sum_{x \in S_n} I_{x;\theta_t^L}\right)^{-1} I_{U;\theta_t^L}\right].$$

- This function is not submodular, so it is not amenable to a greedy method.
  - Selection procedure improves empirically if first you greedily oversample $2B$ points, then greedily reject the worst $B$ points.

# **B**atch **a**ctive learning via **i**nformation ma**t**rices (BAIT)

**Algorithm** *BAIT*

(∗ Batch active learning with neural networks, Ash et al. (2021) ∗)

**Initialize:** $S$ a random labeled dataset of size $B$, $\theta_1 \leftarrow \arg\min_\theta \mathbb{E}_S[\ell(x, y; \theta)]$.

1. **for** $t = 1, 2, \ldots,$
2. **do** compute $I_{U;\theta_t^L}$
3. initialize $M_0 \leftarrow I_{S;\theta_t^L} + \lambda \mathrm{Id}$
4. initialize $\tilde{S} \leftarrow \varnothing$
5. greedily select points $x_1, \ldots, x_B$ by sequentially optimizing:

$$x_b = \underset{x \in U}{\arg\min}\ \mathrm{tr}\left[\left(I_{x;\theta_t^L} + M_{b-1}\right)^{-1} I_{U;\theta_t^L}\right],$$

   setting $M_t \leftarrow I_{x_b;\theta_t^L} + M_{b-1}$ and $\tilde{S} \leftarrow \tilde{S} \cup \{x_b\}$

6. query labels for $\tilde{S}$ and save data $S \leftarrow S \cup \tilde{S}$
7. train model on data $\theta_{t+1} \leftarrow \arg\min_\theta \mathbb{E}_S[\ell(x, y; \theta)]$

# Existing state-of-the-art algorithm

Batch active learning by diverse gradient embeddings (BADGE):

- Represent each candidate $x \in U$ by $g_x \in \mathbb{R}^d$:

$$g_x = \nabla \ell(x, y_x^*; \theta^L),$$

  the last-layer gradient obtained if the most likely label $y_x^*$ is observed.
- Select a batch $S$ of samples that has large Gram determinant.
  - Requires $\|g_x\|$ to be large (learning about $x$ is informative).
  - Requires $S$ to be spread out (the batch is diverse; the $x$'s give different information).

# Experiments

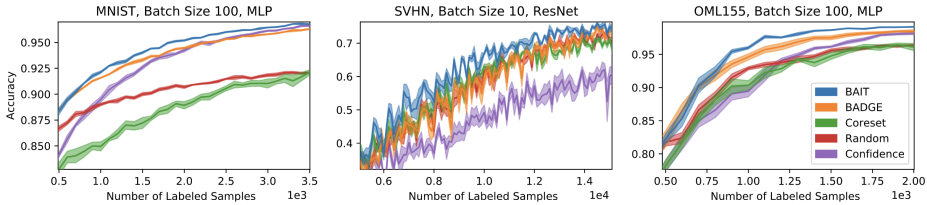


Figure 1: Experiments with MLPs had single hidden ReLu layer of 128 dimensions. Otherwise, a 18-layer ResNet.

# Experiments

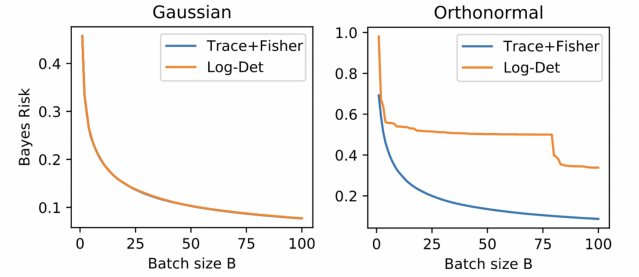


Figure 2: Comparison of BAIT and BADGE for Bayesian linear regression. Gaussian data is generated on $\mathbb{R}^{100}$ with spectral decay $\Sigma_{ii} \propto 1/i^2$. Orthonormal data is supported only on standard basis, $P(x = e_i) \propto 1/i^2$.

- "BADGE does not exploit the occurrence probabilities, and in fact simply selects the coordinates in a cyclic fashion. On the other hand, the optimal strategy focuses effort on the high-probability coordinates, which is captured in the Fisher matrix."

# III. Discussion

# Assumption on Fisher information

## Assumption

*For any $x, y, \theta$, the Fisher information matrix is a function of only $x$ and $\theta$,*

$$I(x, y; \theta) = I_{x;\theta}.$$

- What does this assumption mean?

# Generalize linear models

Generalized linear models (GLM) are examples that satisfy the assumption.

- In a GLM, the response variable has exponential family distribution:

$$p(y|\eta) = \exp\left(\eta^\top t(y) - A(\eta)\right),$$

where $\eta = \theta^\top x$, $t$ is a sufficient statistic, and $A$ is the log-partition function, and:

$$\nabla_\theta \log p(y|x,\theta) = xt(y) - xA'(\theta^\top x)$$
$$\nabla^2_\theta \log p(y|x;\theta) = -xx^\top A''(\theta^\top x)$$

# References


Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with fisher embeddings. *arXiv preprint arXiv:2106.09675*, 2021.

Kamalika Chaudhuri, Sham Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. *arXiv preprint arXiv:1506.02348*, 2015.

John Duchi. Fisher information. *https://web.stanford.edu/class/stats311/Lectures/lec-09.pdf*, 2019. Lecture notes for Information Theory and Statistics.