Transductive Active Learning: Theory and Applications

Jonas Hübotter, Bhavya Sukhija, Lenart Treven, Yarden As, Andreas Krause

Background

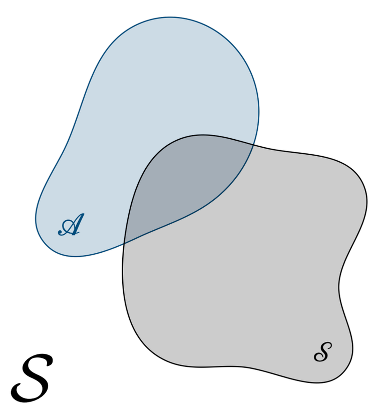
- Active learning is a powerful paradigm for data selection that commonly aims to learn f globally on $\mathcal{X}$
- In many real-world problems,
  - *i* the domain is so large that learning f globally is hopeless; or
  - *ii* agents have limited information / access to $\mathcal{X}$

Can we solve tasks *efficiently* and *extrapolate* beyond limited information?

Transductive Active Learning

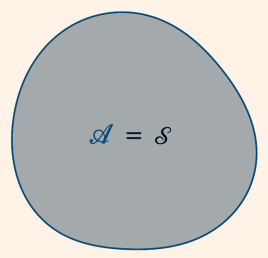
"only learn what is needed to solve a task" [Vapnik, adapted]

- Sample space $\mathcal{S} \subseteq \mathcal{X}$
- Target space $\mathcal{A} \subseteq \mathcal{X}$
- Unknown function f over $\mathcal{X}$



Goal: Learn f within $\mathcal{A}$ by sampling from $\mathcal{S}$

Example: (Inductive) Active Learning
"learn everything"
⇝ TAL generalizes AL to goal-orientation ($\mathcal{A}$) and extrapolation ($\mathcal{S}$)



Probabilistic model of f:

- prior $p(f)$
- likelihood $p(D \mid f)$ of data D
- posterior $p(f \mid D) \propto p(f)p(D \mid f)$

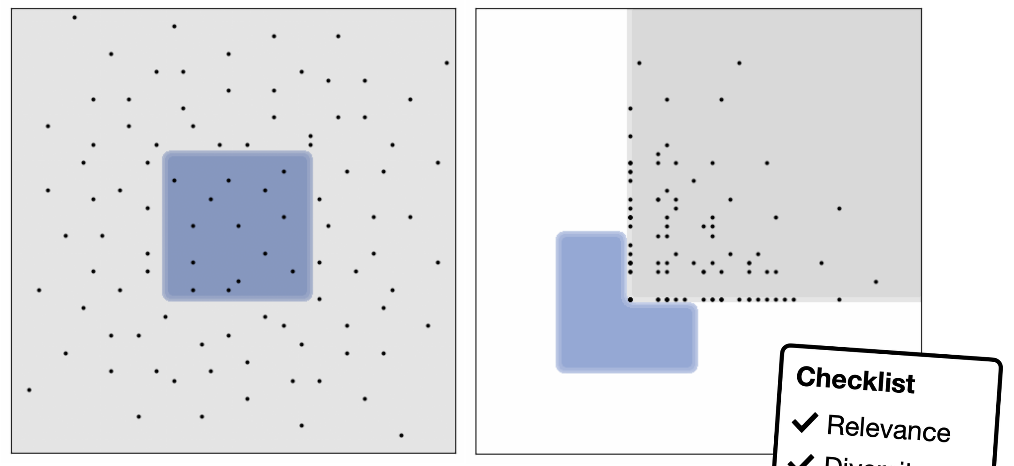
Algorithms [MacKay, 1992]
Select data to minimize *posterior* uncertainty within $\mathcal{A}$

⇝ quantifying "uncertainty", e.g., by entropy:

$$\boldsymbol{x}_n = \arg\min_{\boldsymbol{x} \in \mathcal{S}} \mathrm{H}(\boldsymbol{f}(\mathcal{A}) \mid D_{n-1}, (\boldsymbol{x}, f(\boldsymbol{x}) + \varepsilon)) = \arg\max_{\boldsymbol{x} \in \mathcal{S}} \mathrm{I}(\boldsymbol{f}(\mathcal{A}); (\boldsymbol{x}, f(\boldsymbol{x}) + \varepsilon) \mid D_{n-1})$$

Tractable Transductive Active Learning

Key assumption: f is a Gaussian process



Example with RBF kernel

Theory: How much can be learned about $\mathcal{A}$ from $\mathcal{S}$?

Informally: For every $\boldsymbol{x} \in \mathcal{A}$:

$$\underbrace{\mathrm{Var}(f(\boldsymbol{x}) \mid D_n)}_{\text{posterior uncertainty}} - \underbrace{\mathrm{Var}(f(\boldsymbol{x}) \mid \boldsymbol{f}(\mathcal{S}))}_{\text{irreducible uncertainty}} \leq \underbrace{C\gamma_{\mathcal{A},\mathcal{S}}(n)/\sqrt{n}}_{\to 0 \text{ for many kernels}}$$

where $\gamma_{\mathcal{A},\mathcal{S}}(n) = \max_{X \subseteq \mathcal{S}, |X| = n} \mathrm{I}(\boldsymbol{f}(\mathcal{A}); \boldsymbol{y}(X))$

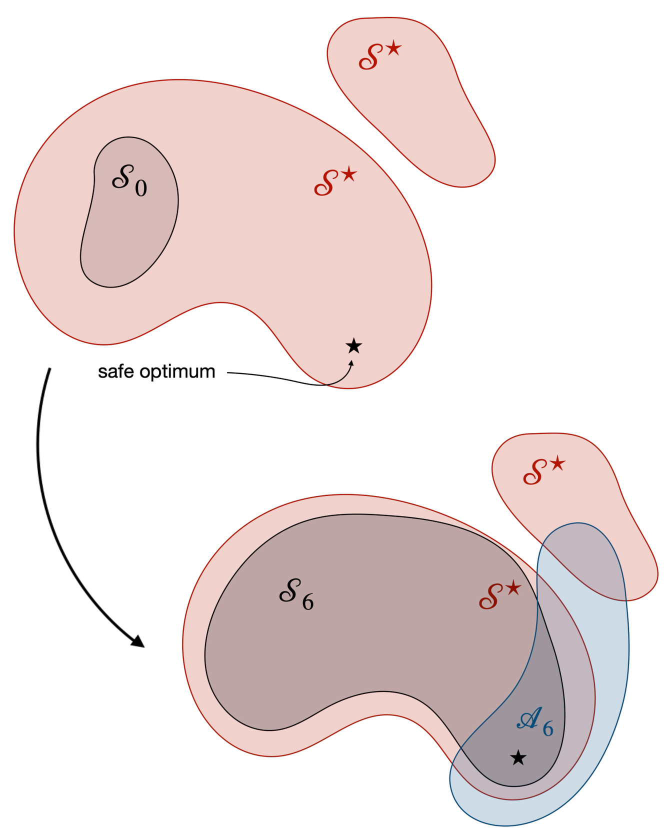
⇝ implies agnostic error bound if f is in RKHS

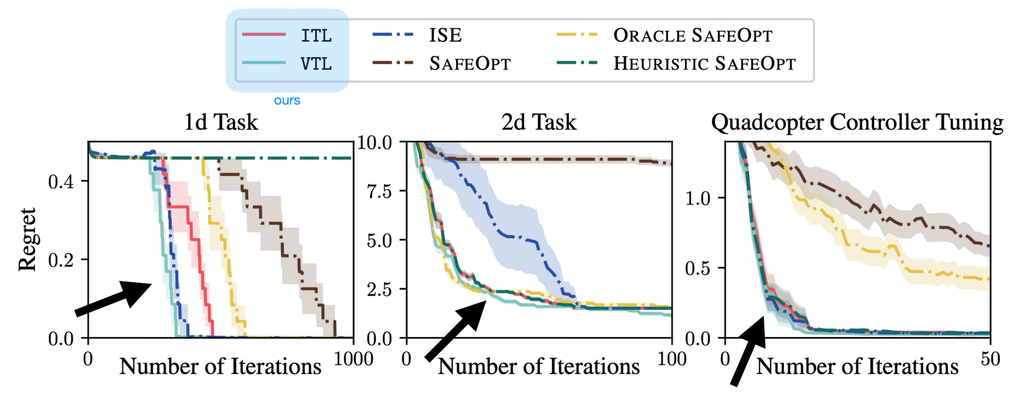
Application: Safe Bayesian Optimization

Task: Optimize *unknown* function under *unknown* constraints that have to be satisfied at all times.

- $\mathcal{S}_n$ - pessimistic safe set
- $\mathcal{A}_n$ - potential safe optima

Theory: *Tighter* convergence guarantees that *generalize* to continuous domains.



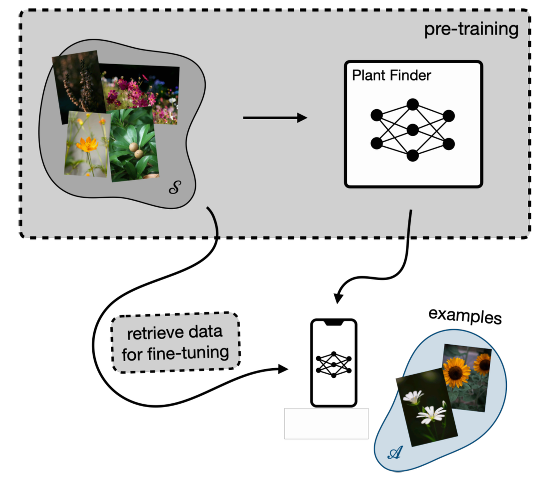


⇝ framing Safe BO as TAL samples only the information needed to find the safe optimum

Application: Active Fine-Tuning

Motivating example: learning a good plant classifier for a user's local biome $\mathcal{A}$

⇝ Automatically find informative (that is, relevant & diverse) examples for $\mathcal{A}$ in dataset $\mathcal{S}$

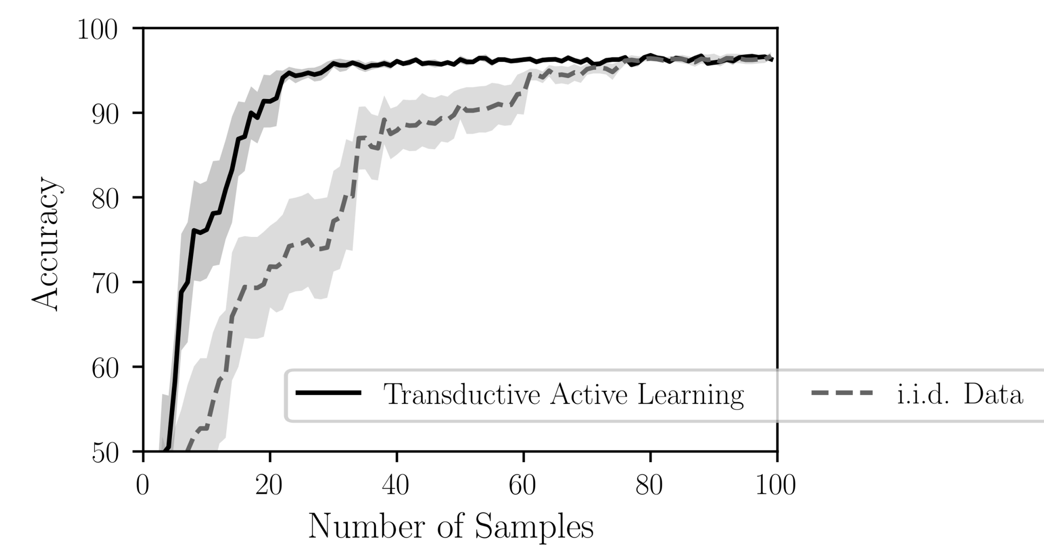


Can we exploit the **latent (unknown) structure**?

- **Inner loop (batch selection):** Approximate NN as a logit-linear function of its (fixed) latent embeddings $\phi(\cdot)$
- **Outer loop (model update):** Train model on batch and improve latent embeddings

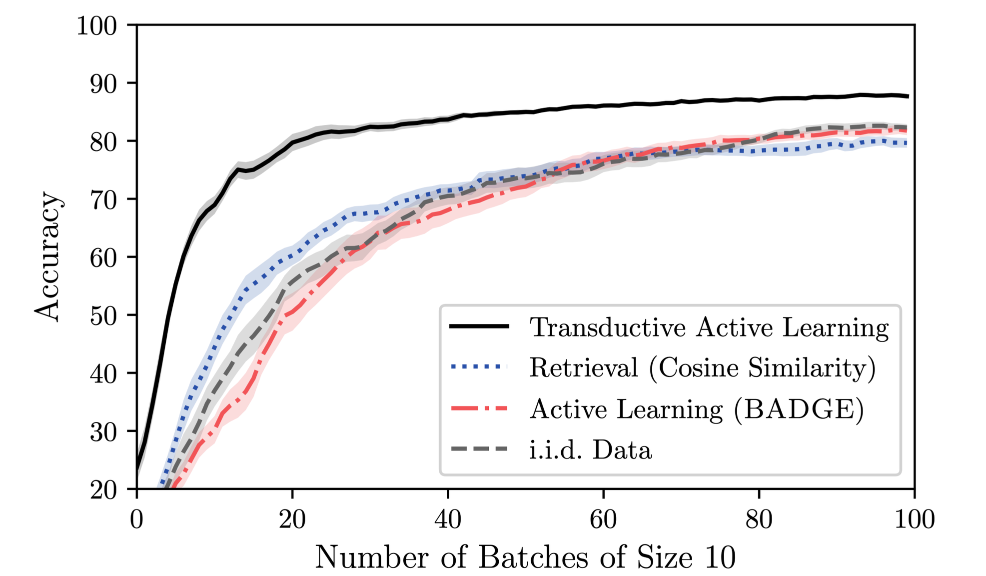
Pre-training: Can **bootstrap** strong representations!

MNIST with target classes $\{3, 6, 9\}$, randomly initialized ConvNet



Fine-tuning: Unifying **retrieval** & **active learning**

CIFAR-100 with target classes $\{1, \ldots, 10\}$, pre-trained EfficientNet-B0



Key Takeaway

Transductive active learning is a promising paradigm for efficiently solving specific tasks under resource constraints.