Learning with Relaxed Supervision

Jacob Steinhardt **Percy Liang**

{jsteinhardt,pliang}@cs.stanford.edu

Intractable Supervision

For weakly-supervised tasks, inference can be intractable:

input x: What is the largest city in California?

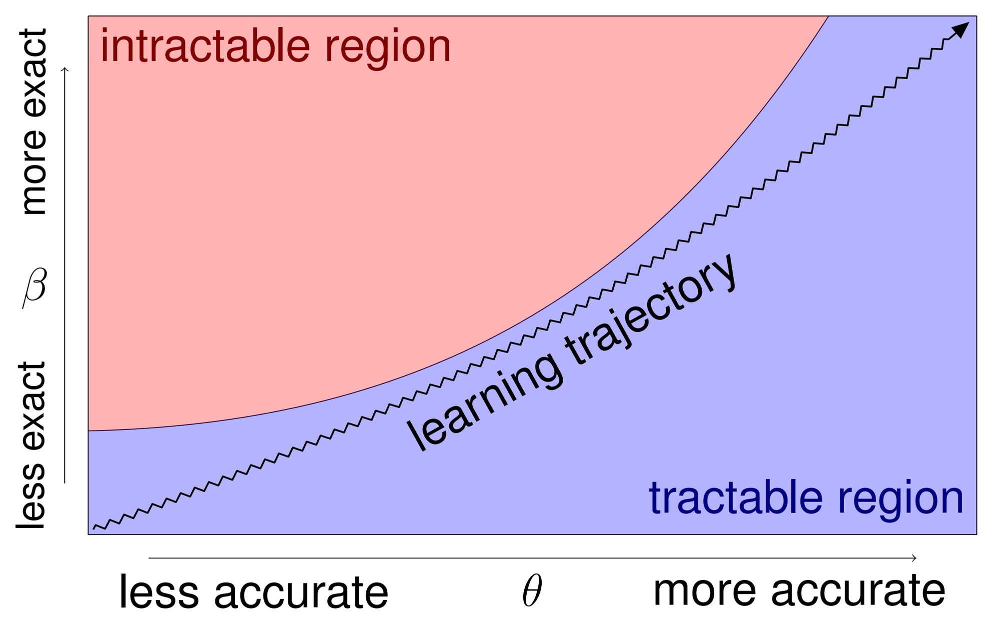
latent z: $\mathrm{argmax}(\lambda x.\text{CITY}(x) \wedge \text{LOC}(x, \text{CA}), \lambda x.\text{POPULATION}(x))$

output y: Los Angeles

Computing $p(z \mid x, y)$ requires inverting arbitrary logical forms!

- Still want to exploit likely statistical relationships (CITY and Los Angeles)
- Need a way to relax the supervision so we can learn tractably.
- Want to maintain good statistical properties (**asymptotic consistency**).

Our Approach



- Start with intractable supervision $q_\infty(y \mid z)$
- Replace with family of relaxed supervision functions $q_\beta(y \mid z)$
- Derive constraints on (θ, β) that ensure tractability of inference
- Optimize likelihood within the tractable region

Intuition:

- Supervision is intractable if too harsh relative to model accuracy.
- Initially need very forgiving supervision, can eventually incorporate full supervision (done adaptively over course of optimization).

The Relaxation

Assume relationship between z and y given by constraints $\mathbb{S}_j$, $j = 1, \ldots, k$ (think machine translation, checking that each word of the output is correct).

Relaxation based on weighted count of constraint violations:

$$p_\beta(y \mid z) \propto \exp\left(-\sum_{j=1}^{k} \beta_j (1 - \mathbb{S}_j(z, y))\right) \qquad (\dagger)$$

When $\beta = 0$, p_β is uniform; when $\beta = \infty$, p_β is original supervision.

Challenges: normalization constant of p_β; ensuring tractable inference.

Framework

Assumptions:

- $x \to z \to y$, where $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is observed and $z \in \mathcal{Z}$ is unobserved.
- Parameterized family $p_\theta(z \mid x)$.
- $z \to y$ is a known deterministic function $y = f(z)$.

Hence, letting $\mathbb{S}(z, y) \in \{0, 1\}$ denote the constraint $[y = f(z)]$:

$$p_\theta(y \mid x) = \sum_z \mathbb{S}(z, y) p_\theta(z \mid x).$$

Goal: decompose $\mathbb{S}$ into smaller components $\mathbb{S}_j$.

- Define projections $\pi_j : \mathcal{Y} \to \mathcal{Y}_j$.
- Projected constraint: $\mathbb{S}_j(z, y) \stackrel{\text{def}}{=} [\pi_j(f(z)) = \pi_j(y)]$.
- If $\pi_1 \times \cdots \times \pi_k$ is one-to-one, then can decompose $\mathbb{S}$ as $\mathbb{S} = \bigwedge_{j=1}^{k} \mathbb{S}_j$.

Example Decompositions

Translation from Unordered Supervision

Goal: infer substitution cipher

input x: a b a a

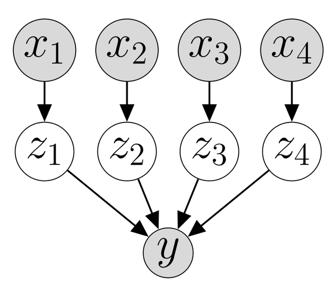
latent z: d c d d

output y: $\{c : 1, d : 3\}$

(cipher: $\{a \mapsto d, b \mapsto c, \ldots\}$)

Supervision: $y = \mathrm{multiset}(z)$

Model $p_\theta(z \mid x)$: soft substitutions



Decomposition (y and z match if all counts match):

$$\underbrace{[y = \overbrace{\mathrm{multiset}(z)}^{f(z)}]}_{\mathbb{S}(z,y)} \Leftrightarrow \bigwedge_{j=1}^{V} \underbrace{[\mathrm{count}(z, j) = \overbrace{\mathrm{count}(y, j)}^{\pi_j(y)}]}_{\mathbb{S}_j(z,y)}$$

Conjunctive Semantic Parsing

Side information: *predicates* $Q_1, \ldots, Q_m$.

- e.g. $Q_6 = [\![\text{DOG}]\!] =$ set of all dogs

input x:	brown dog	(input utterance)
latent z:	(Q_{11}, Q_6)	(set of all brown objects, set of all dogs)
output y:	$Q_{11} \cap Q_6$	(denotation, observed as a set)

For $z = (Q_{j_1}, \ldots, Q_{j_L})$, define the denotation $[\![z]\!] = Q_{j_1} \cap \cdots \cap Q_{j_L}$.

Decomposition (y and z match if contained in same predicates):

$$\underbrace{y = [\![z]\!]}_{\mathbb{S}(z,y)} \Leftrightarrow \bigwedge_{j=1}^{m} \underbrace{\mathbb{I}[[\![z]\!] \subseteq Q_j] = \overbrace{\mathbb{I}[y \subseteq Q_j]}^{\pi_j(y)}}_{\mathbb{S}_j(z,y)}$$

Theory

Lemma (normalization constant). For any z, the log-normalization constant of $p_\beta(y \mid z)$ is bounded above by

$$A(\beta) \stackrel{\text{def}}{=} \sum_{j=1}^{k} \log(1 + (|\mathcal{Y}_j| - 1)\exp(-\beta_j)).$$

Lemma (asymptotic consistency). Suppose that we use $A(\beta)$ above as a surrogate normalization constant for p_β. Then, the MLE of (θ, β) asymptotically recovers the true model parameters.

Tractability Constraints

Typical expression for gradient (for some features $\phi(x, z, y)$):

$$\nabla \log p_{\theta,\beta}(y \mid x) = \underbrace{\mathbb{E}_{z \mid x,y}[\phi(x, z, y)]}_{\text{model + supervision}} - \underbrace{\mathbb{E}_{z \mid x}[\phi(x, z, y)]}_{\text{model}}.$$

To learn, need to sample $p_{\theta,\beta}(z \mid x, y) \propto p_\theta(z \mid x) \exp(\beta^\top \mathbb{S}_{1:k}(z, y))$ (see (†)).

- For large β, this is as intractable as the original supervision.
- Need a way to constrain β to yield tractable inference.

Inference algorithm: rejection sampling.

- Sample from $p_\theta(z \mid x)$, accept with probability $p_\beta(y \mid z)$.

Constrain expected number of rejections based on computational budget τ:

minimize	$\mathbb{E}_{x,y}[-\log p_{\theta,\beta}(y \mid x)]$	$(\mathcal{L})$
subject to	$\mathbb{E}_{x,y}[\mathrm{Rejections}(x, y)] \leq \tau$	$(\mathcal{C})$

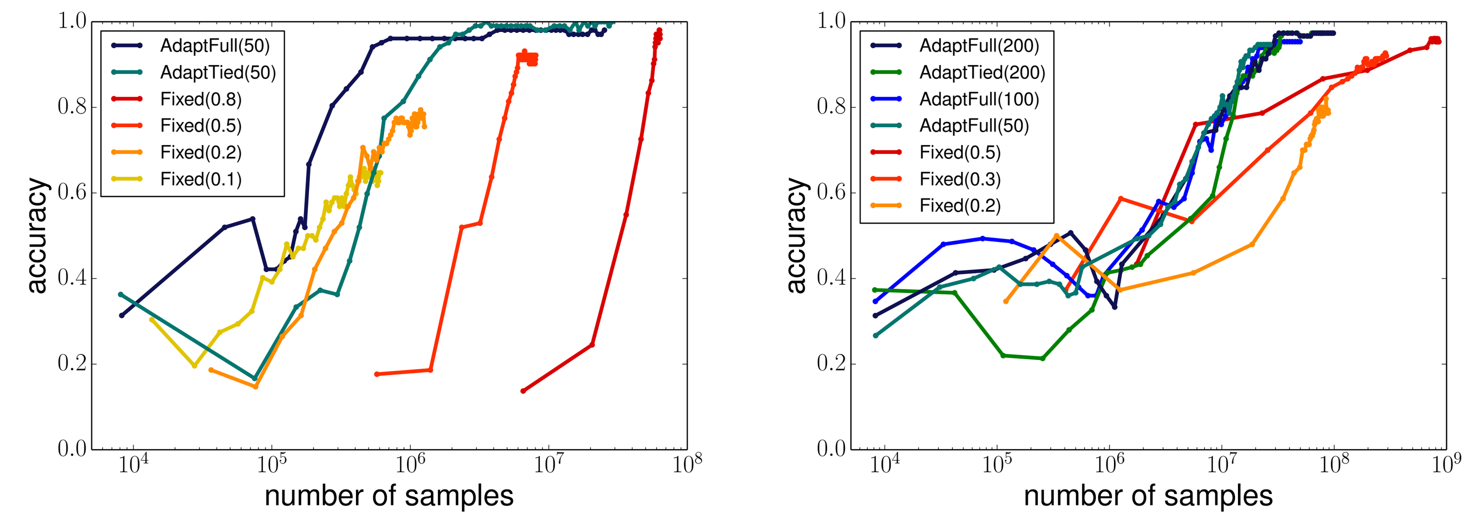
Amazingly, $(\mathcal{C})$ is well-behaved enough to admit an EM-like procedure for constrained optimization! (See paper for full details.)

Experiments

Implemented our relaxed supervision algorithm on both the unordered translation and conjunctive semantic parsing tasks.

Compared a fixed value of relaxation β (FIXED) to optimizing β subject to our tractability constraints (ADAPT).

Our tractability constraints improve efficiency by orders of magnitude while also improving accuracy:



(a) unordered translation (b) conjunctive semantic parsing

Reproducible experiments on CodaLab: worksheets.codalab.org

The first author was supported by the Hertz foundation and by the NSF.