Representation Learning Theory with Unlabeled Data

Shengchao Liu

MILA-UdeM

2021 April

Overview

Motivation

For representation learning with unlabeled data:

- It is a general topic, including semi-supervised learning, self-supervised learning, self-training, etc.
- Empirically with huge success, like the recently widely discussed contrastive self-supervised learning.
- Despite the empirical successes, theoretical progress in understanding how to use unlabeled data has lagged.

Notations

- Hypothesis/Concept space $\mathcal{H}$ (here we do not distinguish the two)
- True error: $err(h)$
- Empirical error: $\widehat{err}(h)$

Uniform Convergence Bounds: Finite Hypothesis

Theorem 1 (Realizable Case)

With probability $1-\delta$, for any $h \in \mathcal{H}$ with $\widehat{err}(h) = 0$, we have that the following bound for m examples:

$$err(h) \leq \frac{1}{m}\Big(\log|\mathcal{H}| + \log\frac{1}{\delta}\Big). \tag{1}$$

Proof: Use union bound.

Theorem 2 (Agnostic Case)

With probability $1-\delta$, for any $h \in \mathcal{H}$, we have that the following bound for m examples:

$$err(h) \leq \widehat{err}(h) + \sqrt{\frac{\log|\mathcal{H} + \log 2/\delta|}{2m}}. \tag{2}$$

Proof: Use Hoeffding's inequality.

Uniform Convergence Bounds: Infinite Hypothesis

This is not applicable in general, since we have infinite size of the hypothesis complexity. How to solve this?
We propose the following complexity measures.

- Binary Classification Setting
  - Rademacher Complexity
  - Growth Function
  - Shattering and VC-Dimension
- More General Setting
  - Covering numbers
  - Packing numbers, etc.

Definition 3 (Empirical Rademacher complexity)

Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[a, b]$ and $S = (z_1, \dots z_m)$ a fixed sample of size m with elements in $\mathcal{Z}$. Then, the empirical Rademacher complexity of $\mathcal{G}$ with respect to the sample S is defined as:

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma \Big[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \Big], \tag{3}$$

where $\sigma = (\sigma_1, \dots, \sigma_m)^T$, with σ_i independent uniform random variable taking values in $\{\pm 1\}$. The random variables σ_i are called Rademacher variables.

If we write g_s as $g_S = (g(z_1), \dots, g(z_m))^T$. Then the empirical Rademacher complexity can be rewritten as

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma \Big[\sup_{g \in \mathcal{G}} \frac{\sigma \cdot g_s}{m} \Big]. \tag{4}$$

Definition 4 (Rademacher complexity)

Let D be the data distribution. For any integer $m \geq 1$, the Rademacher complexity of $\mathcal{G}$ is the expectation of the empirical Rademacher complexity over all samples of size m drawn from D:

$$\mathfrak{R}_m(\mathcal{G}) = \mathbb{E}_{S \sim D^m}[\widehat{\mathfrak{R}}_S(\mathcal{G})] = \mathbb{E}_{S \sim D^m, \sigma}[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i)] \tag{5}$$

Theorem 5 (Uniform convergence bounds with Rademacher complexity)

Let $\mathcal{G}$ be a family of functions mapping from $\mathcal{Z}$ to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over an i.i.d. sample S of size m, each of the following holds for all $g \in \mathcal{G}$:

$$\mathbb{E}[g(z)] \leq \underbrace{\frac{1}{m}\sum_{i=1}^{m} g(z_i)}_{\widehat{\mathbb{E}}_S[g(z)]} + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2m}} \tag{6}$$

and

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{\log 1/\delta}{2m}}. \tag{7}$$

Proof sketch:

- Define $\Phi(S) = \sum_{g \in \mathcal{G}} \left(\mathbb{E}[g] - \widehat{\mathbb{E}}_S[g] \right)$.
- Use the McDiarmid's inequality to get $\Phi(S) \leq \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\log 1/\delta}{2m}}$.
- By definitions and properties of Rademacher complexity, we can then have $\Phi(S) \leq 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2m}}$.
- Use McDiarmid's Inequality to have $\mathfrak{R}_m(\mathcal{G}) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2m}}$.

Lemma 6

Let $\mathcal{H}$ be a family of functions taking values in $\{\pm 1\}$ and let $\mathcal{G}$ be a family of loss functions associated to $\mathcal{H}$ for the zero-one loss: $\mathcal{G} = \{(x, y) \to 1_{h(x)\neq y} : h \in \mathcal{H}\}$. For any sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of elements in $\mathcal{X} \times \{\pm 1\}$, let $S_{\mathcal{X}}$ denote its projection over $\mathcal{X}$: $S_{\mathcal{X}} = (x_1, \ldots, x_m)$. Then, the following relation holds between the empirical Rademacher complexities of $\mathcal{G}$ and $\mathcal{H}$:

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2}\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H})$$

Proof sketch:
If in the binary setting, the loss is $L(h(x), y) = 1_{h(x)\neq y}$ we have:

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_\sigma\Big[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i 1_{h(x)\neq y}\Big] \quad = \frac{1}{2}\mathbb{E}_\sigma\Big[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} -\sigma_i h(x_i)\Big] = \frac{1}{2}\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H}).$$

Note 1: Notice that the lemma implies, by taking the expectation, we have $\mathfrak{R}_S(\mathcal{G}) = \frac{1}{2}\mathfrak{R}_{S_{\mathcal{X}}}(\mathcal{H})$.

Note 2: $\widehat{\mathfrak{R}}_S(\mathcal{G})$ measures how the loss correlates with random noise, and $\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\mathcal{H})$ measures how the prediction correlates with random noise (no label here).

With Theorem 5 and Lemma 6, we can easily get the following theorem:

Theorem 7 (Uniform convergence bounds with Rademacher complexity - Binary Classification)

Let $\mathcal{H}$ be a family of functions mapping from $\mathcal{X}$ to $\{\pm 1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over an i.i.d. sample S of size m, each of the following holds for all $f \in \mathcal{H}$:

$$R(h) \leq \widehat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2m}} \tag{8}$$

and

$$R(h) \leq \widehat{R}_S(h) + \widehat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log 1/\delta}{2m}}. \tag{9}$$

Note: Notice that this bound can be interpreted as

$$\underbrace{R(h)}_{\text{true loss}} \leq \underbrace{\widehat{R}_S(h)}_{\text{empirical loss}} + \underbrace{\mathfrak{R}_m(\mathcal{H})}_{\text{model complexity}} + \underbrace{\sqrt{\frac{\log 1/\delta}{2m}}}_{\text{margin}}.$$

Comments: The computation of $\mathfrak{R}_m(\mathcal{H})$ is equivalent to an *empirical risk minimization* problem, which is computationally expensive. In the next sections, we will relate the Rademacher complexity to combinatorial measures that are easier to compute.

Definition 8 (Growth function)

The growth function $\prod_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for a hypothesis set $\mathcal{H}$ is defined by:

$$\forall m \in \mathbb{N}, \prod_{\mathcal{H}}(m) = \max_{\{x_1,\ldots,x_m\} \subseteq X} \left|\left\{\left(h(x_1), \ldots, h(x_m)\right) : h \in \mathcal{H}\right\}\right|. \tag{10}$$

Notice 1: $\prod_{\mathcal{H}}$ is the maximum number of distinct ways in which m points can be classified using hypothesis in $\mathcal{H}$. Thus, the growth function (shattering number) provides another way to measure the richness of a hypothesis set $\mathcal{H}$.
Notice 2: Unlike Rademacher complexity, this measure does not depend on the distribution, it is purely combinatorial.

Lemma 9 (Massart's lemma)

Let $A \subseteq \mathbb{R}^m$ be a finite set, with $r = \max_{x \in \mathcal{A}} \|x\|_2$, then the following holds:

$$\mathbb{E}_\sigma \Big[\frac{1}{m} \sup_{x \in \mathcal{A}} \sum_{i=1}^{m} \sigma_i x_i \Big] \leq \frac{r\sqrt{2 \log |\mathcal{A}|}}{m}, \tag{11}$$

where σ_i are independent uniform random variables taking values in $\{\pm 1\}$ and $x_1, \ldots, x_m$ are the components of vector x.

Massart's lemma relates the Rademacher complexity and growth function.

Corollary 10

Let $\mathcal{G}$ be a family of functions taking values in $\{\pm 1\}$. Then the following holds:

$$\mathfrak{R}_m(\mathcal{G}) \leq \sqrt{\frac{2 \log \prod_{\mathcal{G}}(m)}{m}}. \tag{12}$$

Corollary 11 (Growth function generalization bound)

$$R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{2 \log \prod_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log 1/\delta}{2m}} \tag{13}$$

Proof sketch: Can be directly obtained from Theorem 7 and Corollary 10. **Notice 1:** The

computation of growth function may not be always convenient since it requires $\prod_{\mathcal{H}}(m), \forall m \geq 1$. The next section introduces an alternative measure of the complexity of $\mathcal{H}$ that is based on a single scalar instead.

Lemma 12 (Sauer's lemma)

Let $\mathcal{H}$ be a hypothesis with $VCdim(\mathcal{H}) = d$. Then for all $m \in \mathbb{N}$, the following inequality holds:

$$\prod_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

Proof sketch: By induction.

Corollary 13

Let $\mathcal{H}$ be a hypothesis set with $VCdim(\mathcal{H}) = d$. Then for all $m \geq d$,

$$\prod_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^{d} = \mathcal{O}(m^{d}).$$

Proof sketch: Use Lemma 12.
This is good because the sum when multiplied out becomes $\sum_{i=0}^{d} \binom{m}{i} = \sum_{i=0}^{d} \frac{m \cdot (m-1) \ldots}{i!} = \mathcal{O}(m^{d})$. When we plug this into the learning error limits, we have $\log(\prod_{\mathcal{H}}(2m)) = \log(\mathcal{O}(m^{d})) = \mathcal{O}(d \log m)$. And this leads to the following bound with VC-Dimension.

Corollary 14 (VC-dimension generalization bounds)

Let $\mathcal{H}$ be a family of functions taking values in $\{\pm 1\}$ with VC-dimension d. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $h \in \mathcal{H}$:

$$R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log 1/\delta}{2m}}. \tag{14}$$

Proof sketch: This can be directly obtained by combining Corollary 11 and Corollary 13.

Definition 1:
A legal notion of compatibility is a function $\chi : C \times \mathcal{X} \to [0, 1]$ where we (overloading notation) define $\chi(f, D) = \mathbb{E}_{x \sim D}[\chi(f, x)]$. Given a sample S, we define $\chi(f, S)$ to be the empirical average of χ over the sample, *i.e.*, $\chi(f, S) = \frac{1}{|S|} \sum_{i=1}^{|S|} \chi(f, x_i)$. Here χ is the measure/function for compatibility.

Definition 2:
Given compatibility notion χ, the incompatibility of f with D is $1 - \chi(f, D)$. We will also call this its **unlabeled error rate**, $err_{unl}(f) \triangleq 1 - \chi(f, D)$ when χ and D are clear from context. For a given sample S, we use $\widehat{err}_{unl}(f) = 1 - \chi(f, S)$ to denote the empirical average over S.

Definition 3:
Given value τ, we define $C_{D,\chi}(\tau) = \{f \in C : err_unl(f) \leq \tau\}$. So, *i.e.*, $C_{D,\chi}(1) = C$. Similarly, for a sample S, we define $C_{S,\chi}(\tau) = \{f \in C : \widehat{err}_{unl}(f) \leq \tau\}$.

Notations

$$err(f) = err_D(f) = Pr_{x\sim D}[f(x) \neq c^*(x)]$$
$$d(f_1, f_2) = d_D(f_1, d_2) = Pr_{x\sim D}[f_1(x) \neq f_2(x)]$$
$$\widehat{err}(f) = err_S(f) = \frac{1}{|S|}\sum_{i=1}^{|S|}\delta[f(x_i) \neq c^*(x_i)]$$
$$\widehat{d}(f, f_2) = d_S(f_1, d_2) = \frac{1}{|S|}\sum_{i=1}^{|S|}\delta[f_1(x_i) \neq f_2(x_i)].$$

Finite Hypothesis Space

Theorem 4. (Finite hypothesis space, realizable unsupervised learning, realizable supervised learning)
If $c^* \in C$ and $err_{unl}(c^*) = 0$, then m_u unlabeled examples and m_l examples are sufficient to learn to error ϵ with probability $1 - \delta$, where

$$m_u = \frac{1}{\epsilon}(\log|\mathcal{H}| + \log\frac{2}{\delta}), \quad \textit{and} \quad m_l = \frac{1}{\epsilon}(\log|C_{D,\chi}(\epsilon)| + \log\frac{2}{\delta}).$$

In particular, with probability at least $1 - \delta$, all $f \in C$ with $\widehat{err}(f) = 0$ and $\widehat{err}_{unl}(f) = 0$ have $err(f) \leq \epsilon$.

Theorem 5. (Finite hypothesis space, agnostic unsupervised learning, realizable supervised learning)
If $c^* \in C$ and $err_{unl}(c^*) = t$, then m_u unlabeled examples and m_l labeled examples are sufficient to learn to error ϵ with probability $1 - \delta$, for

$$m_u = \frac{1}{2\epsilon^2}(\log(|\mathcal{H}|) + \log\frac{4}{\delta}), \quad \textit{and} \quad m_l = \frac{1}{\epsilon}(\log|C_{D,\chi}(\epsilon)| + \log\frac{2}{\delta}).$$

(Typo in the paper.) In particular, with probability at least $1 - \delta$, the $f \in C$ that optimizes $\widehat{err}_{unl}(f)$ subject to $\widehat{err}(f) = 0$ has $err(f) \leq \epsilon$.

Proof sketch: Both are using standard uniform convergence bounds and the union bound between supervised and unsupervised part.

Infinite Hypothesis Space

The main logic is we bound the sample complexity for the unlabeled part using VC-dim, and the reduced hypothesis space is bounded by the *number of splits/partitions*, *i.e.*, $C[m, D]$ or $C[m, \bar{S}]$.

Theorem 7. (Infinite hypothesis space, PAC learnable $\mathcal{H}$ for unsupervised learning, realizable supervised learning)
If $c^* \in C$ and $err_{unl}(c^*) = t$, then m_u unlabeled examples and m_l labeled examples are sufficient to learn to error ϵ with probability $1 - \delta$, for

$$m_u = \mathcal{O}\Big(\frac{VCdim(\chi(C))}{\epsilon^2}\log\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\log\frac{2}{\delta}\Big)$$

and (15)

$$m_l = \frac{2}{\epsilon}\Big[\log\big(C_{D,\chi}(t+2\epsilon)[2m_l, D]\big) + \log\frac{4}{\delta}\Big],$$

where recall $C_{D,\chi}(t+2\epsilon)[2m_l, D]$ is the expected number of splits of $2m_l$ points drawn from D using concepts in C of unlabeled error rate $\leq t + 2\epsilon$. In particular, with probability at least $1 - \delta$, the $f \in C$ that optimizes $\widehat{err}_{unl}(f)$ subject to $\widehat{err}(f) = 0$ has $err(f) \leq \epsilon$.
Proof sketch: We can directly obtain this with VC-dim uniform convergence bound.

$$\begin{aligned}
&err_{unl}(c^*) = t && \text{//by assumption}\\
&\widehat{err}_{unl}(c^*) \leq t + \epsilon\\
&\widehat{err}_{unl}(h) \leq \widehat{err}_{unl}(c^*) \leq t + \epsilon && \text{//by optimizing } h\\
&err_{unl}(h) \leq \widehat{err}_{unl}(h) \leq t + 2\epsilon.
\end{aligned}$$

Notice 1: We want to highlight the difference between C and $\chi(C)$. C is the concept class, and $\chi(C)$ is the set of compatibility functions for each hypothesis in C.

Infinite Hypothesis Space

Theorem 10. (Infinite hypothesis space, PAC learnable $\mathcal{H}$ for unsupervised learning, agnostic supervised learning)
Let $f_t^* = \arg\min_{f \in C}[err(f)|err_{unl}(f) \leq t]$. Then an unlabeled sample of size

$$m_u = \mathcal{O}\Big(\frac{\max[VCdim(\chi(C)), VCdim(C)]}{\epsilon^2}\log\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\log\frac{2}{\delta}\Big)$$

and a labeled sample of size

$$m_l = \frac{8}{\epsilon^2}\Big[\log\left(C_{D,\chi}(t+2\epsilon)[2m_l, D]\right) + \log\frac{16}{\delta}\Big]$$

is sufficient so that with probability $\geq 1-\delta$, the $f \in C$ that optimizes $\widehat{err}(f)$ subject to $\widehat{err}_{unl}(f) \leq t+\epsilon$ has $err(f) \leq err(f_t^*) + \epsilon + \sqrt{\log(4/\delta)/(2m_l)} \leq err(f_t^*) + 2\epsilon$.

Interpretation: One can also state Theorem 10 in the form more commonly used in statistical learning theory: given any number of labeled examples (m_l) and given $t > 0$, Theorem 10 implies that with high probability, the function f that optimizes $\widehat{err}(f)$ subject to $\widehat{err}_{unl}(f) \leq t+\epsilon$ satisfies

$$err(f) \leq \widehat{err}(f) + \epsilon_t \leq err(f_t^*) + \epsilon_t + \sqrt{\frac{\log 4/\delta}{2m_l}}$$

where $\epsilon_t = \sqrt{\frac{8}{m_l}\log\left(16C_{D,\chi}(t+2\epsilon)[2m_l, D]/\delta\right)}$.

Short Summary

$$err(f) \leq err(f_t^*) + \epsilon_t + \sqrt{\frac{\log 4/\delta}{2m_l}}$$

where $\epsilon_t = \sqrt{\frac{8}{m_l} \log \left(16 C_{D,\chi}(t + 2\epsilon)[2m_l, D]/\delta\right)}$.

- This is agnostic of the compatibility function on the unlabeled data.
- Yingyu's paper discuss what compatibility (regression) loss would look like in the self-supervised learning setting.
- Sanjeev's paper discuss what f_t^* looks like in the contrastive self-supervised learning case.
- Kento's paper discuss a potential drawback of Sanjeev's paper.

Notations

- labeled data $S = \{(x_i, y_i)\}_{i=1}^{m_l}$ from a distribution $\mathcal{D}$ over the domains $\mathcal{X} \times \mathcal{Y}$.
- $\mathcal{X} \subseteq \mathbb{R}^d$ is input feature space
- $\mathcal{Y}$ is the label space
- representation function $\phi = h(x) \in \mathbb{R}^r$, where $h \in \mathcal{H}$
- predictor $y = f(\phi) \in \mathcal{Y}$, where $f \in \mathcal{F}$
- loss function $\ell_c(f(h(x)), y) \in [0, 1]$
- unlabeled data $U = \{\tilde{x}_i\}_{i=1}^{m_u}$ from a distribution $\mathcal{U}_X$, which can be same or different from the marginal distribution $\mathcal{D}_X$ of $\mathcal{D}$.

Definitions

Definition 1. Given a loss function $L_r(h, g; x)$ for an input x involving a representation h and a regularization function g, the regularization loss of h and g on a distribution $\mathcal{U}_X$ over $\mathcal{X}$ is defined as

$$L_r(h, g; \mathcal{U}_X) \triangleq \mathbb{E}_{x \sim \mathcal{U}_X} \left[L_r(h, g; x) \right].$$

The regularization loss of a representation h on $\mathcal{U}_X$ is defined as

$$L_r(h; \mathcal{U}_X) \triangleq \min_{g \in \mathcal{G}} L_f(h, g; \mathcal{U}_X).$$

Definition 2. Given $\tau \in [0, 1]$, the τ-regularization-loss subset of representation hypotheses $\mathcal{H}$ is:

$$\mathcal{H}_{\mathcal{D}_X, L_r}(\tau) \triangleq \{h \in \mathcal{H} : L_r(h; \mathcal{D}_X) \leq \tau\}.$$

Same Domain, Realizable, Finite Hypothesis Class.
Theorem 1. Suppose there exist $h^* \in \mathcal{H}, f^* \in \mathcal{F}, g^* \in \mathcal{G}$ such that $L_c(f^*, h^*; \mathcal{D}) = 0$ and $L_r(h^*, g^*; \mathcal{D}_X) = 0$. For any $\epsilon_0, \epsilon_1 \in (0, 1/2)$, a set U of m_u unlabeled examples and a set S of m_l labeled examples are sufficient to learn 4o an error ϵ_1 with probability $1 - \delta$, where

$$m_u \geq \frac{1}{\epsilon_0}\left[\ln|\mathcal{G}| + \ln|\mathcal{H}| + \ln\frac{2}{\delta}\right], \qquad m_l \geq \frac{1}{\epsilon_1}\left[\ln|\mathcal{F}| + \ln|\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)| + \ln\frac{2}{\delta}\right].$$

In particular, with probability at least $1 - \delta$, all hypotheses $h \in \mathcal{H}, f \in \mathcal{F}$ with $L_c(h, f; S) = 0$ and $L_r(h; U) = 0$ will have $L_c(f, h; \mathcal{D}) \leq \epsilon_1$.
Proof sketch: Same as before.

Same Domain, Unrealizable, Infinite Hypothesis Class. Let $\mathcal{N}_{\mathcal{H}}(\epsilon)$ denote the ϵ-covering number of $\mathcal{H}$.
Standard bound. With the size of

$$m \geq \frac{C}{\epsilon^2} \ln \frac{1}{\delta} \ln \epsilon = \mathcal{O}\big(\ln \frac{1}{\delta} \ln \epsilon \big) = \mathcal{O}\big(\log_{\frac{1}{\epsilon}} \frac{1}{\delta} \big)$$

then we can have $P[|L_c(h, f; D) - L_c(h, f; S)| \leq \epsilon] > 1 - \delta$.

Theorem 4. Suppose there exist $h^* \in \mathcal{H}, f^* \in \mathcal{F}, g^* \in \mathcal{G}$ such that $L_c(f^*, h^*; \mathcal{D}) = 0$ and $L_r(h^*, g^*; \mathcal{D}_X) \leq \epsilon_r$. For any $\epsilon_0, \epsilon_1 \in (0, 1/2)$, a set U of m_u unlabeled examples and a set S of m_l labeled examples are sufficient to learn to an error ϵ_1 with probability $1 - \delta$, where

$$m_u \geq \frac{C}{\epsilon_0^2} \ln \frac{1}{\delta} \Big[\ln \mathcal{N}_{\mathcal{G}} \Big(\frac{\epsilon_0}{4L} \Big) + \ln \mathcal{N}_{\mathcal{H}} \Big(\frac{\epsilon_0}{4L} \Big) \Big],$$

$$m_l \geq \frac{C}{\epsilon_1^2} \ln \frac{1}{\delta} \Big[\ln \mathcal{N}_{\mathcal{F}} \Big(\frac{\epsilon_1}{4L} \Big) + \ln \mathcal{N}_{\mathcal{H}_{\mathcal{D}_X}, L_r(\epsilon_r + \epsilon_0)} \Big(\frac{\epsilon_1}{4L} \Big) \Big]$$

for some absolute constant C. In particularly, with probability at least $1 - \delta$, the hypothesis $f \in \mathcal{F}, h \in \mathcal{H}$ with $L_c(h, f; S) = 0$ and $L_r(h, g; U) \leq \epsilon_r + \epsilon_0$ for some $g \in \mathcal{G}$ satisfy $L_c(f, h; \mathcal{D}) \leq \epsilon_1$.
Proof sketch: With the standard bound by covering numbers, the bound can be easily derived.

Different Domain, Unrealizable, Infinite Hypothesis Class.

Theorem 3. Suppose the unlabeled data U is from a distribution $\mathcal{U}_X$ different from $\mathcal{D}_X$. Suppose there exist $h^* \in \mathcal{H}, f^* \in \mathcal{F}, g^* \in \mathcal{G}$ such that $L_c(f^*, h^*; \mathcal{D} \leq \epsilon_c$ and $L_r(h^*, g^*; \mathcal{U}_X) \leq \epsilon_r$. Then the same sample complexity bounds as in Theorem 2 hold as follows (replacing $\mathcal{D}_X$ with $\mathcal{U}_X$ in the Equation 7):

$$m_u \geq \frac{C}{\epsilon_0^2} \ln \frac{1}{\delta} \Big[\ln \mathcal{N}_{\mathcal{G}} \Big(\frac{\epsilon_0}{4L} \Big) + \ln \mathcal{N}_{\mathcal{H}} \Big(\frac{\epsilon_0}{4L} \Big) \Big],$$

$$m_l \geq \frac{C}{\epsilon_1^2} \ln \frac{1}{\delta} \Big[\ln \mathcal{N}_{\mathcal{F}} \Big(\frac{\epsilon_1}{4L} \Big) + \ln \mathcal{N}_{\mathcal{H}_{\mathcal{U}_X, L_r}(\epsilon_r + 2\epsilon_0)} \Big(\frac{\epsilon_1}{4L} \Big) \Big].$$

Proof sketch: Same as Theorem 4, except by replacing $\mathcal{D}_X$ with $\mathcal{U}_X$.

Discussions of Functional Regularization

When is functional regularization not helpful? The theorems and analysis also provide implications for cases when the auxiliary self-supervised task may *not* help the target prediction task.

1. The regularization may not lead to a significant reduction in the size of hypothesis class. Namely, it can be too easy to be "compatible" on the self-supervised learning tasks. **To make it useful, we need to get $\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_0)$ significantly (exponentially) smaller than the entire class $\mathcal{H}$.**
2. The auxiliary task can fail if the regularization loss threshold may contain no hypothesis with a small prediction loss.
3. Inability of the optimization to lead to a good solution.

Is uniform convergence suitable for our analysis?

- Existing work with failed uniform convergence are for supervised learning without auxiliary tasks: it is generally believed that the hypothesis class is larger than statistically necessary, and the optimization has an implicit regularization during training. Thus the uniform convergence fails to explain the generalization.
- In the setting with auxiliary tasks, **functional regularization has a regularization effect of restricting the learning to a smaller subset of the hypothesis space, as will be shown next**.
- Once with the functional regularization, the regularization effect is more significant than the implicit ones, thus the generalization can be explained by uniform convergence.

Note. Another story here: explicit regularization (L_2 norm, data augmentation, momentum, functional regularization) has more significant effect than the implicit regularization.

An Example of Functional Regularization via Auto-encoder

Auto-Encoder as functional regularization

Learning Without Functional Regularization. Let $\mathcal{H}$ be the class of linear functions from $\mathbb{R}^d$ to $\mathbb{R}^r$, where $r < d/2$. $\mathcal{F}$ be the class of linear functions over some activations.

$$\phi = h_W(x) = Wx, \quad y = f_a(\phi) = \sum_{i=1}^{r} a_i \sigma(\phi_i), \quad \text{where } W \in \mathbb{R}^{r \times d}, a \in \mathbb{R}^r$$

Here $\sigma(t)$ is an activation function, the rows of W and a have ℓ_2 norms bounded by 1. We consider the MSE prediction loss, *i.e.*, $L_c(f, h; x) = \|y - f(h(x))\|_2^2$. Without prior knowledge on data, no functional regularization corresponds to end-to-end training on $\mathcal{F} \circ \mathcal{H}$.

Data Property. Assume data consists of a signal and noise. Let columns of $B \in \mathbb{R}^{d \times d}$ be eigenvectors of $\Sigma \triangleq \mathbb{E}\left[xx^T\right]$, then the prediction labels are largely determined by the signal in the first r directions: $y = \sum_{i=1}^{r} a_i^* \sigma(\phi_i^*) + \nu$ and $\phi^* = B_{1:r}^T x$, where $a^* \in \mathbb{R}^r$ is a ground-truth parameter with $\|a^*\|2 \leq 1$, $B_{1:r}$ is the set of first r eigenvectors of Σ, and ν is a small Gaussian noise.

An Example of Functional Regularization via Auto-encoder

Learning With Functional Regularization. Then we show that $\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon}{4L}\right) \geq \mathcal{N}_{\mathcal{D}_X, L_r(\epsilon_r)}\left(\frac{\epsilon}{4L}\right)$ (see Lemma 6 below) since

$$\mathcal{H}_{\mathcal{D}_X, L_r}(\epsilon_r) = \{h_W(x) : W = OB_{1:r}^T, \quad O \in \mathbb{R}_{r\times r}, O \text{ is orthonormal}\},$$
$$\mathcal{H} \supseteq \{h_W(x) : W = OB_S^T, \quad O \in \mathbb{R}_{r\times r}, O \text{ is orthonormal}\},$$

where B_S refers to the sub-matrix of columns in B having indices in S. Therefore, the label sample complexity bound is reduced by $\frac{C}{\epsilon^2} \ln \binom{d-r}{r}$, *i.e.*, the error bound is reduced by $\frac{C}{\sqrt{m_l}} \ln \binom{d-r}{r}$ when using m_l labeled points.

An Example of Functional Regularization via Auto-encoder

Lemma 6 For $\epsilon/4L < 1/2$,

$$\mathcal{N}_{\mathcal{H}}\left(\frac{\epsilon}{4L}\right) \geq \binom{d-r}{r}\mathcal{N}_{\mathcal{H}_{\mathcal{D}_X,L_r(\epsilon_r)}}\left(\frac{\epsilon}{4L}\right) \tag{16}$$

Proof sketch.

1. We first have $\mathcal{H}_{D_x,L_r}$ and $\mathcal{H}_S$.

$$\mathcal{H}_{\mathcal{D}_X,L_r}(\epsilon_r) = \{h_W(x) : W = OB_{1:r}^T, \quad O \in \mathbb{R}_{r\times r}, O \text{ is orthonormal}\},$$
$$\mathcal{H} \supseteq \{h_W(x) : W = OB_S^T, \quad O \in \mathbb{R}_{r\times r}, O \text{ is orthonormal}\},$$

2. We say that the covering number for $\mathcal{H}_{D_x,L_r}$ and H_S are the same (with same tau).
3. We prove that $\mathcal{H}_S$ and $\mathcal{H}'_S$ are far away.

$$\begin{aligned}\|OB_S^T - O'B_{S'}^T\|_F^2 &= \text{trace}((OB_S^T - O'B_{S'}^T)^T(OB_S^T - O'B_{S'}^T)) \\ &= \|B_S^T\|_F^2 + \|B_{S'}\|_F^2 - \text{trace}((O'B_{S'}^T)^T(OB_S^T)) \\ &\geq r + r - (r-1) - (r-1) = 2.\end{aligned}$$

For different S ans S', they do not overlap, so there are $\binom{d-r}{r}$ so many different S, and all of the corresponding $\mathcal{H}_S$ do not overlap.

Note that $\ln\binom{d-r}{r} = \Theta(r\ln(d))$ when r is small, and thus the reduction is roughly linear initially and then grows slower with r.

A Theoretical Analysis of Contrastive Unsupervised Representation Learning, ICML'19

Some notations:

- $\mathcal{X}$ denotes the set of all possible data points.
- $(x, x^+) \sim D_{sim}$ similar data in the form of pairs that come from a distribution D_{sim}.
- $(x_1^-, x_2^-, \ldots, x_k^-) \sim D_{neg}$ as k iid negative samples from distribution D_{neg}.
- The goal is to learn the representation function $f : \mathcal{X} \to \mathbb{R}^d$ such that $\|f(\cdot)\| \leq R$ for some $R > 0$.
- To formalize the notion of semantically similar pairs (x, x^+), we introduce the concept of *latent class*: let C denote the set of all latent classes. Associated with each class $c \in C$ is a probability distribution D_c over $\mathcal{X}$.
- We assume a distribution ρ over the classes that characterizes how these classes occur on the unlabeled data.
- Based on this, we define the semantic similarity and dissimilarity:
  - $D_{sim}(x, x^+) = \mathbb{E}_{c\sim\rho} D_c(x) D_c(x^+)$
  - $D_{neg}(x^-) = \mathbb{E}_{c\sim\rho} D_c(x^-)$

  Currently empirically works heuristically identify such similar pairs from co-occurring image or text data.
- The supervised labeled dataset for the task T consists of this process: A label c is picked according to a distribution D_T. Then, a sample is drawn from D_c. Together they form a labeled pair (x, c) with distribution $D_T(x, c) = D_c(x) D_T(c)$.
  Notice that D_c may or mat not relate to ρ, the previous one is on supervised label, while the latter is the *latent class* on the unlabeled data.

Notations

Define classifier $g : \mathcal{X} \to \mathbb{R}^{k+1}$ (for $k+1$z classes) and loss as $\ell(\{g(x)_y - g(x)_{y'}\}_{y' \neq y})$. We consider two losses:

- hinge loss: $\ell(v) = \max\{0, 1 + \max_i\{-v_i\}\}$ for $v \in \mathbb{R}^k$
- logistic loss: $\ell(v) = \log(1 + \sum_i \exp(-v_i))$ for $v \in \mathbb{R}^k$

Notice that here the loss is operated on a vector.
Supervised loss of the classifier g is

$$L_{sup}(T, g) = \mathbb{E}_{(x,c)\sim D_T}\left[\ell(\{g(x)_c - g(x)_{c'}\}_{c' \neq c})\right]$$

Often we use $g(x) = Wf(x)$, where $W \in \mathbb{R}^{(k+1)d}$. In the fine-tuning case, the supervised loss is to learn W with f fixed:

$$L_{sup}(T, f) = \inf_{W \in \mathbb{R}^{(k+1)d}} L_{sup}(T, Wf)$$

Definitions

Definition 2.1 (Mean classifier) For a function f and a task $T = \{c_1, \ldots, c_{k+1}\}$, the mean classifier is W^μ whose c^{th} row is the mean μ_c of representations of inputs with label c: $\mu_c = \mathbb{E}_{x \sim D_c}[f(x)]$. We use $L^\mu_{sup}(T, f) = L_{sup}(T, W^\mu f)$ as shorthand for its loss.

Definition 2.2 (Average Supervised Loss) Average loss for a function f on $(k+1)$-way task is defined as

$$L_{sup}(f) = \mathbb{E}_{\{c_i\}_{i=1}^{k+1} \sim \rho^{k+1}} \left[L_{sup}(\{c_i\}_{i=1}^{k+1}, f) | c_i \neq c_j \right]$$

The average supervised loss of its mean classifier is

$$L_{sup}(f)^\mu = \mathbb{E}_{\{c_i\}_{i=1}^{k+1} \sim \rho^{k+1}} \left[L^\mu_{sup}(\{c_i\}_{i=1}^{k+1}, f) | c_i \neq c_j \right]$$

Definition 2.3 (Unsupervised Loss) The population loss if

$$L_{un}(f) = \mathbb{E}\Big[\ell\big(\{ f(x)^T (f(x^+) - f(x_i^-)) \}_{i=1}^k \big) \Big]$$

and its empirical counterpart with M examples $(x_j, x_j^+, x_{j1}^-, \ldots, x_{jk}^-)_{k=1}^M$ from $D_{sim} \times D_{neg}^k$ is

$$\widehat{L}_{un}(f) = \frac{1}{M} \sum_{j=1}^{M} \ell\big(\{ f(x_j)^T (f(x_j^+) - f(x_{ij}^-)) \}_{i=1}^k \big)$$

Theorem A.3 [Vector-contraction inequality. Corollary 4 in paper A vector-contraction inequality for Rademacher complexities] Let Z be any set, and $S = \{z_j\}_{j=1}^{M} \in Z^M$. Let $\tilde{\mathcal{F}}$ be a class of functions $\tilde{f} : Z \to \mathbb{R}^n$ and $h : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz. For all $\tilde{f} \in \tilde{\mathcal{F}}$, let $g_{\tilde{f}} = h \circ \tilde{f}$. Then

$$\mathbb{E}_{\sigma \sim \{\pm 1\}^M}\left[\sup_{\tilde{f} \in \tilde{\mathcal{F}}} \langle \sigma, (g_{\tilde{f}})_{|S} \rangle\right] \leq \sqrt{2} L \mathbb{E}_{\sigma \sim \{\pm 1\}^{nM}}\left[\sum_{\tilde{f} \in \tilde{\mathcal{F}}} \langle \sigma, \tilde{f}_{|S} \rangle\right]$$

where $\tilde{f}_{|S} = \left(\tilde{f}_t(z_j)\right)_{t \in [n], j \in [M}$.

Interpretation: $\tilde{f}$ is like representation function, h is like loss function. This Theorem is essentially saying that $\mathfrak{R}_S(\{h \circ \tilde{f} : \tilde{f} \in \tilde{\mathcal{F}}\}) \leq 2L\mathfrak{R}_S(\tilde{\tilde{\mathcal{F}}})$ (an informal version).

Lemma 4.2 With probability at least $1-\delta$ over the training set S, for all $f \in \mathcal{F}$

$$L_{un}(\hat{f}) \leq L_{un}(f) + Gen_M$$

Lemma A.2 Let $\ell : \mathbb{R}^k \to \mathbb{R}$ be η-Lipschitz and bounded by B. Then with probability at least $1-\delta$ over the training set $S = \{(x_j, x_j^+, x_{j1}^- \dots x_{jk}^-)\}_{j=1}^M$, for all $f \in \mathcal{F}$

$$L_{un}(\hat{f}) \leq L_{un}(f) + \mathcal{O}\Big(\frac{\eta R\sqrt{k}\mathfrak{R}_S(\mathcal{F})}{M} + B\sqrt{\frac{\log 1/\delta}{M}}\Big),$$

where

$$\mathfrak{R}_S(\mathcal{F}) = \mathbb{E}_{\sigma\sim\{\pm 1\}^{(k+2)dM}}[\sup_{f\in\mathcal{F}}\langle\sigma, f_{|S}\rangle]$$

and $f_{|S} = \Big(f_t(x_j), f_t(x_j^+), f_t(x_{j1}^-), \dots, f_t(x_{jk}^-)\Big)_{j\in[M],t\in[d]}$. Note that for $(k+1)$-way classification, for hinge loss we have $\eta = 1$ and $B = \mathcal{O}(R^2)$, while for logistic loss $\eta = 1$ and $B = \mathcal{O}(R^2 + \log k)$. With $k = 1$, we get Lemma 4.2.

Proof sketch of Lemma A.2: Start with the uniform convergence bound with Rademacher complexity and definitions before. One key step is to calculate the bound for the Lipschitz coefficient.

Lemma 4.3 For all $f \in \mathcal{F}$

$$L^{\mu}_{sup}(f) \leq \frac{1}{1-\tau}(L_{un}(f) - \tau)$$

Proof.

The key idea of the proof is the use of Jensen's inequality. Unlike the unsupervised loss which uses a random point from a class as a classifier, using the mean of the class as the classifier should only make the loss lower. Let $\mu_c = \mathbb{E}_{x \sim D_c} f(x)$ be the mean representation of class c.

$$\begin{aligned}
L_{un}(f) &= \mathbb{E}_{\substack{(x,x^+)\sim D_{sim} \\ x^- \sim D_{neg}}} \left[\ell(f(x)^T(f(x^+) - f(x^-)))\right] \\
&= \mathbb{E}_{c^+,c^- \sim \rho^2} \mathbb{E}_{\substack{x^+ \sim D_{c^+} \\ x \sim D_{c^+} \quad x^- \sim D_{c^-}}} \left[\ell(f(x)^T(f(x^+) - f(x^-)))\right] \\
&\geq \mathbb{E}_{c^+,c^- \sim \rho^2} \mathbb{E}_{x^+ \sim D_{c^+}} \left[\ell(f(x)^T(\mu_{c^+} - \mu_{c^-}))\right] && \text{// Jensen's Inequality} \\
&= (1-\tau)\mathbb{E}_{c^+,c^- \sim \rho^2}\left[L^{\mu}_{sup}(\{c^+, c^-\}, f) | c^+ \neq c^-\right] + \tau \\
&= (1-\tau)L^{\mu}_{sup}(f) + \tau && \text{// Def 2.2}
\end{aligned}$$

□

Binary Classification

Theorem 4.1 With probability at least $1-\delta$, for all $f \in \mathcal{F}$

$$L^{\mu}_{sup}(\hat{f}) \leq \frac{1}{1-\tau}(L_{un}(f)-\tau) + \frac{1}{1-\tau} Gen_M,$$

where

$$Gen_M = \mathcal{O}\Big(R\frac{\mathfrak{R}_S(\mathcal{F})}{M} + R^2\sqrt{\frac{\log 1/\delta}{M}}\Big)$$

Proof.

The result follows directly by applying Lemma 4.3 for $\hat{f}$ and followed with Lemma 4.2. First according to Lemma 4.3, $\hat{f} \in \mathcal{F}$, we have:

$$L^{\mu}_{sup}(\hat{f}) \leq \frac{1}{1-\tau}(L_{un}(\hat{f})-\tau)$$

Then plug in Lemma 4.2, we have

$$L^{\mu}_{sup}(\hat{f}) \leq \frac{1}{1-\tau}(L_{un}(f)-\tau+Gen_M)$$

□

We can say that if $\mathcal{F}$ is rich enough and L_{un} can be made small, then Theorem 4.1 suffices. In the next section we explain that this may not always be possible unless $\tau \ll 1$. And we show one way to alleviate this.

Price of Negative Sampling: Class Collision

The inherent limitation of contrastive learning: negative samples can be from the same class as similar pair $\Longrightarrow L_{un}(f)$ can be large. Need to understand when L_{un} can be made small:

$$L_{un}(f) - \tau = \underbrace{(1-\tau)L_{un}^{\neq}(f)}_{\substack{c^+ \neq c^- \\ \text{need contrastive } f}} + \underbrace{\tau(L_{un}^{=}(f) - 1)}_{\substack{c^+ = c^- \\ \text{need intraclass concentration}}}$$

$$\begin{aligned}
L_{un}^{\neq}(f) &= \mathbb{E}_{\substack{c^+, c^- \sim \rho^2 \\ x, x^+ \sim D_{c^+}^2 \\ x^- \sim D_{c^-}}} \left[\ell(f(x)^T (f(x^+) - f(x^-))) | c^+ \neq c^- \right] \\
L_{un}^{=}(f) &= \mathbb{E}_{\substack{c \sim \nu \\ x, x' x^- \sim D_c^3}} \left[\ell(f(x)^T (f(x^+) - f(x^-))) \right] \\
&\geq \mathbb{E}_{c \sim \nu, x \sim D_c} \left[\ell(f(x)^T (\mu_c - \mu_c)) \right] \\
&= 1
\end{aligned}$$

We will show next is that the magnitude of $L_{un}^{=}(f)$ can be controlled by the intraclass deviation of f. Let $\Sigma(f, c)$ be the covariance matrix of $f(x)$ when $x \sim D_c$. We define a notion of intraclass deviation as follows:

$$s(f) = \mathbb{E}_{c \sim \nu} \left[\sqrt{\|\Sigma(f, c)\|_2} \mathbb{E}_{x \sim D_c} \|f(x)\| \right] \tag{17}$$

Price of Negative Sampling: Class Collision

Lemma A.1 Let $c \in C$ and $\ell : \mathbb{R}^t \to \mathbb{R}$ be either the t-way hinge loss or t-way logistic loss, as defined in Section 2. Let $x, x^+, x_1^-, \ldots, x_t^-$ be iid draws from D_c. For all $f \in \mathcal{F}$, let

$$L_{un,c}^{=}(f) = \mathbb{E}_{x,x^+,x_i^-}\left[\ell\left(\{f(x)^T(f(x^+) - f(x^-))\}_{i=1}^t\right)\right]$$

Then

$$L_{un,c}^{=}(f) - \ell(0) \leq c' t \sqrt{\|\Sigma(f,c)\|_2}\mathbb{E}_{x\sim D_c}[\|f(x)\|]$$

Theorem 4.5 With probability at least $1-\delta$, $\forall f \in \mathcal{F}$

$$L_{sup}(\hat{f}) \leq L_{sup}^{\mu}(\hat{f}) \leq L_{un}^{\neq}(f) + \beta s(f) + \eta Gen_M$$

where $\beta = c' \frac{\tau}{1-\tau}$, $\eta = \frac{1}{1-\tau}$ and c' is a constant.

Towards Competitive Guarantees

Limitations of contrastive learning.

- Inter-class representation not well aligned and distributed.
- Intra-class representation not well bounded.

Limitations of contrastive learning

(a) Mean is bad (b) High intraclass variance

Figure: Figure 1

		supervised	unsupervised
Figure 1.a	$f_0(x_i) = (0, 0)$ $f_1(x_1) = (1, r)$ $f_1(x_2) = (-1, 2r)$	①: $L_{sup}(f_0) = 0, L_{sup}(f_1) = 1$	②: $L_{un}(f_0) = 1, L_{un}(f_1) = \Omega(r^2)$
Figure 1.b	$f_0(x_i) = (0, 0)$ $f_1(x_1) = (0, 1)$ $f_1(x_2) = (0, r)$ $f_1(x_3) = (0, -1)$ $f_1(x_4) = (0, -r)$	③: $L_{sup}(f_0) = 0, L_{sup}(f_1) = 1$	④: $L_{un}(f_0) = 1, L_{un}(f_1) = \Omega(r^2)$

Table: Illustrations of two examples in Figure 1.

Limitations of contrastive learning

First of all, because it's the binary case (suppose $y = \{\pm 1\}$), so the length of v is 1, and the loss becomes

$$\ell = \max(0, 1 - (g(x)_y - g(x)_{-y}))$$

- ②, for the unsupervised learning on Figure 1.a.

$$\begin{aligned}
L_{un}(f_0) &= 0.5 \max(0, 1 - f_0(x_1)^T (f_0(x_1) - f_0(x_2)) \\
&\quad + 0.5 \max(0, 1 - f_0(x_2)^T (f_0(x_2) - f_0(x_1)) \\
&= 0.5 \max(0, 1 - 0) + 0.5 \max(0, 1 - 0) \\
&= 1 \\
L_{un}^{=}(f_1) &= 0 \\
L_{un}^{\neq}(f_1) &= 0.5 \max(0, 1 - f_1(x_1)^T (f_1(x_1) - f_1(x_2)) \\
&\quad + 0.5 \max(0, 1 - f_1(x_2)^T (f_1(x_2) - f_1(x_1)) \\
&= 0.5 \max(0, 1 - \langle (1, r), (2, -r) \rangle) + 0.5 \max(0, 1 - \langle (-1, 2r), (-2, r) \rangle \\
&= 0.5 \max(0, r^2 - 1) + 0.5 \max(0, -1 - 2r^2) \\
&= \Omega(r^2)
\end{aligned}$$

- ④, for the unsupervised learning on Figure 1.b.

$$
\begin{aligned}
L_{un}(f_0) &= 0.25 * 4 \max(0, 1 - (0,0)^T (0,0)) \\
&= 1 \\
L_{un}^{\neq}(f_1) &= 0.25 \max(0, 1 - f_1(x_1)^T (0, r+1)) + 0.25 \max(0, 1 - f_1(x_2)^T (0, r+1)) \\
&\quad + 0.25 \max(0, 1 - f_1(x_3)^T (0, -r-1)) + 0.25 \max(0, 1 - f_1(x_4)^T (0, -r-1)) \\
&= 0.5 \max(0, -r) + 0.5 \max(0, 1 - r^r - r) \\
&= 0
\end{aligned}
$$

For $L_{un}^{=}(f_1)$, WLOG, let's only check the first half of them.

$$
\begin{aligned}
L_{un}^{=}(f_1) &= \frac{1}{8} \max(0, 1 - \langle (0,1), (0, 1-r) \rangle) + \frac{1}{8} \max(0, 1 - \langle (0,1), (0, r-1) \rangle) + \dots \\
&= \frac{1}{8} \max(0, r) + \frac{1}{8} \max(0, r^2 - r) \\
&= \Omega(r^2)
\end{aligned}
$$

Limitations of contrastive learning

(a) Mean is bad (b) High intraclass variance

Figure: Figure 1

		supervised	unsupervised
Figure 1.a	$f_0(x_i) = (0, 0)$ $f_1(x_1) = (1, r)$ $f_1(x_2) = (-1, 2r)$	①: $L_{sup}(f_0) = 0, L_{sup}(f_1) = 1$	②: $L_{un}(f_0) = 1, L_{un}(f_1) = \Omega(r^2)$
Figure 1.b	$f_0(x_i) = (0, 0)$ $f_1(x_1) = (0, 1)$ $f_1(x_2) = (0, r)$ $f_1(x_3) = (0, -1)$ $f_1(x_4) = (0, -r)$	③: $L_{sup}(f_0) = 0, L_{sup}(f_1) = 1$	④: $L_{un}(f_0) = 1, L_{un}(f_1) = \Omega(r^2)$

Table: Illustrations of two examples in Figure 1.

Competitive Bound via Intraclass Concentration

Lemma 5.1 For $f \in \mathcal{F}$, if the random variable $f(X)$, where $X \sim D_c$, is σ^2-sub-Gaussian in every direction for every class c and has maximum norm $R = \max_{x \in X} \|f(x)\|$, then for all $\epsilon > 0$,

$$L_{un}^{\neq}(f) \leq \gamma L_{\gamma,sup}^{\mu}(f) + \epsilon$$

where $\gamma = 1 + c' R\sigma\sqrt{\log \frac{R}{\epsilon}}$ and c' is some constant.

Corollary 5.1.1 For all $\epsilon > 0$, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$,

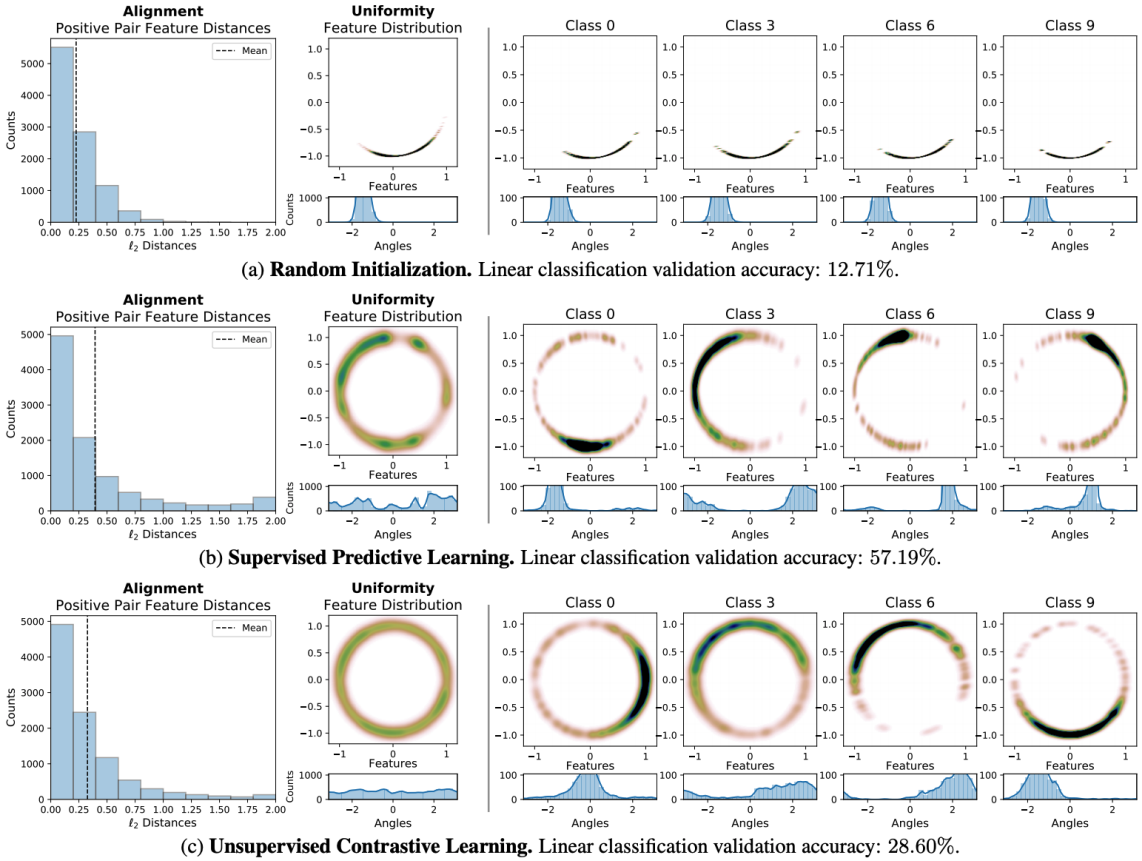
$$L_{sup}^{\mu}(\hat{f}) \leq \gamma(f) L_{\gamma(f),sup}^{\mu}(f) + \beta s(f) + \eta Gen_M + \epsilon$$

where $\gamma(f)$ is as defined in Lemma 5.1, $\beta = c' \frac{\tau}{1-\tau}$, $\eta = \frac{1}{1-\tau}$, and c' is a constant.

Proof sketch: This can be obtained directly by combining Theorem 4.5 and Lemma 5.1.

Some Insights

Another relevant paper we discussed before: [1] Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere, ICML'20



(a) **Random Initialization.** Linear classification validation accuracy: 12.71%.

(b) **Supervised Predictive Learning.** Linear classification validation accuracy: 57.19%.

(c) **Unsupervised Contrastive Learning.** Linear classification validation accuracy: 28.60%.

Figure: 3 from [1].

- The competitive bound needs high-margin mean classifier (well-distributed and well-aligned) and strong/low intraclass concentration (well-aligned).
- The first condition can match with [1].
- Yet, the second condition hasn't attracted enough attention yet.

Definitions

Definition 1 (Data Generation Process).

1. Draw latent classes: $c, \{c_k^-\}_{k=1}^K \sim \rho^{K+1}$
2. Draw input example: $x \sim \mathcal{D}_c$
3. Draw data augmentations: $a, a' \sim \mathcal{A}^2$
4. Apply data augmentations: $a(x), a'(x)$
5. Draw negative examples: $\{x_k^-\}_{k=1}^K \sim \mathcal{D}_{c_k^-}^K$
6. Draw data augmentations: $\{a_k^-\}_{k=1}^K \sim \mathcal{A}^K$
7. Apply data augmentations: $\{a_k^-(x_k^-)\}_{k=1}^K$

Definition 2 (Self-supervised Loss). Expected self-supervised loss is defined as

$$L_{info}(f) \triangleq \mathop{\mathbb{E}}_{c,\{c_k^-\}_{k=1}^K \sim \rho^{K+1}} \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_c \\ a,a'\sim\mathcal{A}^2}} \mathop{\mathbb{E}}_{\substack{\{x_k^-\sim\mathcal{D}_{c_k^-}\}_{k=1}^K \\ \{a_k^-\}_{k=1}^K\sim\mathcal{A}^K}} \ell_{info}(z, Z)$$

$$\triangleq \mathop{\mathbb{E}}_{c,\{c_k^-\}_{k=1}^K \sim \rho^{K+1}} \mathop{\mathbb{E}}_{\substack{x\sim\mathcal{D}_c \\ a,a'\sim\mathcal{A}^2}} \mathop{\mathbb{E}}_{\substack{\{x_k^-\sim\mathcal{D}_{c_k^-}\}_{k=1}^K \\ \{a_k^-\}_{k=1}^K\sim\mathcal{A}^K}} -\log \frac{\exp(z \cdot z/t)}{\sum_{z_k \in Z} \exp(z \cdot z_k/t)}$$

Definition 3 (Mean Classifier's Supervised Loss).

$$L^{\mu}_{sup}(\widehat{f}) \triangleq \mathop{\mathbb{E}}_{\substack{x,y\sim\mathcal{D}\\ a\sim\mathcal{A}}} -\ln \frac{\exp\left(\widehat{f}(a(x))\cdot\mu_y\right)}{\sum_{j\in\mathcal{Y}}\exp\left(\widehat{f}(a(x))\cdot\mu_j\right)}$$

Definition 4 (Mean Classifier's Supervised Sub-class Loss).

$$\begin{aligned} L^{\mu}_{sub}(\widehat{f},\mathcal{Y}_{sub}) &\triangleq \mathop{\mathbb{E}}_{\substack{x,y\sim\mathcal{D}_{sub}\\ a\sim\mathcal{A}}} \ell^{\mu}_{sub}(\widehat{f}(a(x)),y,\mathcal{Y}_{sub}) \\ &\triangleq \mathop{\mathbb{E}}_{\substack{x,y\sim\mathcal{D}_{sub}\\ a\sim\mathcal{A}}} -\ln \frac{\exp\left(\widehat{f}(a(x))\cdot\mu_y\right)}{\sum_{j\in\mathcal{Y}_{sub}}\exp\left(\widehat{f}(a(x))\cdot\mu_j\right)} \end{aligned}$$

Step 1. Introduce a lower bound We derive a lower bound of unsupervised loss $L_{info}(f)$.

$$
\begin{aligned}
L_{info}(f) &\geq \mathop{\mathbb{E}}_{\substack{c,\{c_k^-\}_{k=1}^K\sim\rho^{K+1} \\ x\sim\mathcal{D}_c \\ a\sim\mathcal{A} \\ \{x_k^-\sim\mathcal{D}_{c_k^-}\}_{k=1}^K}} \ell_{info}\left(z, \{\mu(x), \mu(x_1^-), \ldots, \mu(x_K^-)\}\right) \\
&\geq \mathop{\mathbb{E}}_{\substack{c,\{c_k^-\}_{k=1}^K\sim\rho^{K+1} \\ x\sim\mathcal{D}_c \\ a\sim\mathcal{A}}} \ell_{info}\left(z, \{\mu(x), \mu_{c_1^-}, \ldots, \mu_{c_K^-}\}\right) \\
&\geq \mathop{\mathbb{E}}_{\substack{c,\{c_k^-\}_{k=1}^K\sim\rho^{K+1} \\ x\sim\mathcal{D}_c \\ a\sim\mathcal{A}}} \ell_{info}\left(z, \{\mu_c, \mu_{c_1^-}, \ldots, \mu_{c_K^-}\}\right) + d(f), \qquad (7)
\end{aligned}
$$

where $\mu(x) = \mathbb{E}_{a\sim\mathcal{A}} f(a(x))$ and $d(f) = \frac{1}{t}\mathbb{E}_{c\sim\rho}[-\mathbb{E}_{x\sim\mathcal{D}_c}\|\mu(x)\|_2^2]$.

Step 2. Decomposition into the averaged sub-class loss Follow Arora's paper, we introduce collision probability: $\tau_K = \mathbb{P}(Col(c, \{c_k^-\}_{k=1}^K) \neq 0)$, where $Col(c, \{c_k^-\}_{k=1}^K = \sum_{k=1}^K 1_{c=c_k^-}$. We omit the arguments of Col for simplicity: let $C_{sub} = C_{sub}(\{c, c_1^-, \ldots, c_K^-\})$ be a function to remove duplicated latent classes given sampled latent classes.

Proposition 6 (CURL Lower Bound of Self-supervised Loss). For all f,

$$
\begin{aligned}
L_{info}(f) &\geq \mathop{\mathbb{E}}_{\substack{c,\{c_k^-\}_{k=1}^K \sim \rho^{K+1} \\ x\sim\mathcal{D}_c \\ a\sim\mathcal{A}}} \ell_{info}\left(z, \{\mu_c, \mu_{c_1^-}, \ldots, \mu_{c_K^-}\}\right) + d(f) \\
&\geq (1-\tau_K) \mathop{\mathbb{E}}_{c,\{c_k^-\}_{k=1}^K \sim \rho^{K+1}} [\underbrace{L_{sub}^{\mu}(f, C_{sub})}_{\text{sub-class loss}} | Col = 0] \\
&+ \tau_K \mathop{\mathbb{E}}_{c,\{c_k^-\}_{k=1}^K \sim \rho^{K+1}} [\underbrace{\ln(Col+1)}_{\text{collision}} | Col \neq 0] + d(f).
\end{aligned}
$$

Limitations of CURL Lower Bound

Proposition 6 (CURL Lower Bound of Self-supervised Loss). For all f,

$$
\begin{aligned}
L_{info}(f) &\geq \mathop{\mathbb{E}}_{\substack{c,\{c_k^-\}_{k=1}^K\sim\rho^{K+1} \\ x\sim\mathcal{D}_c \\ a\sim\mathcal{A}}} \ell_{info}\left(z, \{\mu_c, \mu_{c_1^-}, \ldots, \mu_{c_K^-}\}\right) + d(f) \\
&\geq (1-\tau_K) \mathop{\mathbb{E}}_{c,\{c_k^-\}_{k=1}^K\sim\rho^{K+1}} [\underbrace{L_{sub}^{\mu}(f, C_{sub})}_{\text{sub-class loss}} | Col = 0] \\
&+ \tau_K \mathop{\mathbb{E}}_{c,\{c_k^-\}_{k=1}^K\sim\rho^{K+1}} [\underbrace{\ln(Col+1)}_{\text{collision}} | Col \neq 0] + d(f).
\end{aligned}
$$

We can easily observe that the lower bound converges to the collision term by increasing K since the collision probability τ_K converges to 1. As a result, the sub-class loss rarely contributes to the lower bound.

For example in CIFAR-10, the number of supervised classes is 10, and let's assume the latent classes are the same as the supervised classes. Then $K = 32$ means we have $K = 32$ i.i.d. draws for unsupervised/self-supervised learning. When $K = 32$,

$$
(1-\tau_{K=32}) = \frac{9}{10}^{32} \approx 0.034, \quad \tau_{K=32} \approx 0.967,
$$

i.e., the only 3.4% training examples contribute to the sub-class loss, the others fall into the collision term. In Arora's paper, it argues that large negative samples degrade performance. However, empirically it's not the case, where larger K can be better.

Proposed Lower Bound

Definition 7 (Probability to Draw All Latent Classes). Assume that ρ is a uniform distribution over latent classes C. The probability such that drawn K latent classes from ρ contain all latent classes is defined as

$$v_k \triangleq \sum_{n=1}^{K} \sum_{m=0}^{|C|-1} \binom{|C|-1}{m} (-1)^m \left(1 - \frac{m}{|C|}\right)^{n-1},$$

where the first summation is a probability such that n drawn latent samples contain all latent classes.

Proof sketch. Classic combinatorial problems.

Theorem 8 (Lower Bound of Self-supervised Loss). For all f,

$$\begin{aligned} L_{info}(f) \geq \frac{1}{2}\Big\{ & v_{K+1} \mathop{\mathbb{E}}_{c,\{c_k^-\}_{k=1}^K \sim \rho^{K+1}} [\underbrace{L_{sub}^{\mu}(f, C)}_{\text{sup loss}} | C_{sub} = C] \\ & + (1 - v_{K+1}) \mathop{\mathbb{E}}_{c,\{c_k^-\}_{k=1}^K \sim \rho^{K+1}} [\underbrace{L_{sub}^{\mu}(f, C_{sub})}_{\text{sub-class loss}} | C_{sub} \neq C] \\ & + \mathop{\mathbb{E}}_{c,\{c_k^-\}_{k=1}^K \sim \rho^{K+1}} \underbrace{\ln(Col + 1)}_{\text{collision}} \Big\} + d(f). \end{aligned}$$

Proof sketch. Based on definition and Eq (7).

Conclusions and Future Directions

Conclusions:

- We discuss two general frameworks for theoretical analysis on the effect of how unlabeled training can help supervised learning.
  - By taking the unlabeled training as a functional regularization to reduce the hypothesis space to be exponentially smaller.
  - By introducing the notion of mean-class classifier learned from the contrastive self-supervised learning.

Future Directions:

- Finer-grained analysis, like the effect of different transformations.
- Another recent research direction in non-contrastive self-supervised learning: BYOL and SimSiam.
  - Ongoing project to illustrate how to generalize this to the standard supervised learning with MLE.

  Besides, some recent work have extended SimSiam to GNN:
  - Paper: Bootstrap Representation Learning on Graphs
  - Paper: SelfGNN: Self-supervised Graph Neural Networks without explicit negative sampling