# Semi-Supervised Learning: the Case When Unlabeled Data is Equally Useful

**Jingge Zhu**
Department of Electrical and Electronic Engineering
University of Melbourne
jingge.zhu@unimelb.edu.au

## Abstract

Semi-supervised learning algorithms attempt to take advantage of relatively inexpensive unlabeled data to improve learning performance. In this work, we consider statistical models where the data distributions can be characterized by continuous parameters. We show that under certain conditions on the distribution, unlabeled data is equally useful as labeled date in terms of learning rate. Specifically, let $n, m$ be the number of labeled and unlabeled data, respectively. It is shown that the learning rate of semi-supervised learning scales as $O(1/n)$ if $m \sim n$, and scales as $O(1/n^{1+\gamma})$ if $m \sim n^{1+\gamma}$ for some $\gamma > 0$, whereas the learning rate of supervised learning scales as $O(1/n)$.

## 1 INTRODUCTION

It is known that in favorable situations, semi-supervised learning (SSL) is able to take advantage of unlabeled data to improve learning performance. In this work, we study how learning rate (defined to be the convergence rate of the excess risk in this paper) is improved by having additional unlabeled data under a parametrization assumption of the data distribution. Our main finding is that under our assumption and certain conditions on the data-generating distribution, unlabeled data is as useful as the labeled data in terms of learning rate.

Numerous works across the past few decades are devoted to understand the role of unlabeled data in learning problems. The early work of [Castelli and Cover, 1996] studied a simple mixture model and showed the relative value of labeled and unlabeled data under different assumptions of the model. The author in [Rigollet, 2007] formally formulated the notion of *cluster assumption* and proposed a method that takes advantage of unlabeled data to achieve fast convergence rates. A more sophisticated mixutre model was studied in [Singh et al., 2009], where different regimes of parameters are identified in which the unlabeled data help. The recent work [Göpfert et al., 2019] gave an overview of various assumptions in different works. The readers are referred to [Chapelle et al., 2006] [Zhu, 2008] for a comprehensive literature review on the topic. This paper uses a different assumption than most previous works, and we will comment on their differences in Section 6 after presenting our main results.

In this work, we view both supervised and semi-supervised learning problem as a variation of the universal prediction problem [Merhav and Feder, 1998]. In the classical setup of universal prediction, an observer sequentially receives a sequence of observations $x_1, x_2, \ldots$, and wishes to predict the next outcome $x_t$ based on all past observations up to time $t-1$. The exact underlying distribution that generates the data is unknown to the predictor, except that it comes from a family of parametrized distributions. The goal is to design a *universal predictor* that performs well in the absence of the exact knowledge of the distribution. The connection to the learning problem is that instead of considering a sequential prediction problem, we assume that all past observations (i. e. training data) are given, and only one prediction needs to be made (for the testing data). Importantly, we still assume that the data-generating distribution is not exactly known except that it comes from a parameterized family.

The main contributions of this paper are summarized as follows.

- For some widely used loss functions, we provide an upper bound on the excess risk (Lemma 1) characterized by a conditional mutual information term. This bound could be interesting on its own.

*Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, PMLR volume 124, 2020.

- Using the above upper bound, we obtain the learning rate of supervised and semi-supervised learning problems (Theorem 1). Let $n, m$ be the number of labeled and unlabeled data, respectively. We show that under certain conditions (to be specified in Section 4), the rate of semi-supervised learning scales as $O(1/n)$ if $m \sim n$, and scales as $O(1/n^{1+\gamma})$ if $m \sim n^{1+\gamma}$ for $\gamma > 0$, whereas the learning rate of supervised learning scales as $O(1/n)$. We also identify the corresponding constant in the leading term in each case. This shows that under appropriate conditions, the unlabeled data is equally useful as the labeled data insofar as the convergence rate is concerned.
- A lower bound on the learning rate of supervised learning algorithms with a certain type of loss function is given (Lemma 5), showing that our characterization of the learning rate is tight.

## 2 PROBLEM STATEMENT

Let $(X, Y)$ be a pair of random variables with the density function $p_\theta(x, y)$ where $\theta \in \Lambda$, and the set $\Lambda$ is a measurable set in $\mathbb{R}^d$. We assume $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ where $\mathcal{X}$ is an arbitrary feature space, and $\mathcal{Y}$ is a discrete set consisting of labels. With a slight abuse of notations, we use $p_\theta(x) := \sum_y p_\theta(x, y)$ and $p_\theta(y) := \int_x p_\theta(x, y)dx$ to denote the marginal distributions of $X$ and $Y$, respectively. The distribution $p_\theta(x, y)$ can also be seen as the conditional distribution of $(X, Y)$ given the parameter value $\theta$. Hence throughout the paper, we will use the notations $p_\theta(\cdot)$ and $p(\cdot|\theta)$ interchangeably. We point out that our results can be extended straightforwardly to the case when $Y$ is a continuous random variable.

Let $w : \mathcal{X} \to \mathcal{D}$ be a hypothesis (classifier/predictor) that maps each element in $\mathcal{X}$ to an element in the space $\mathcal{D}$. It is most natural to take $\mathcal{D}$ to be $\mathcal{Y}$, where the mapping $w$ returns a label for each $x$. However, we also allow $\mathcal{D}$ to be different from $\mathcal{Y}$. For example, we could take $\mathcal{D}$ to be the probability simplex of dimension $|\mathcal{Y}|$, where each element in $\mathcal{D}$ is a nonnegative vector summing up to unity. In this case, the mapping $w$ returns a *probability assignment* on $y$ for each $x$.

Given the hypothesis $w$ and a pair $(x, y)$, the *risk* is defined as $\ell(w(x), y)$ for some loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$. To lighten notations, we often use $Z$ to represent a pair $(X, Y)$, and write $\ell(w(X), Y)$ simply as $\ell(w, Z)$.

For a given hypothesis $w$, the expected risk is defined as

$$L_\theta(w) := \mathbb{E}_\theta \{\ell(w, Z)\}, \tag{1}$$

where the subscript denotes that the expectation is taken with respect to $Z \sim p_\theta(Z)$. We define $w^*$ to be the *Bayes hypothesis* that minimizes of the expected risk

$$w^* := \text{argmin}_w \mathbb{E}_\theta \{\ell(w, Z)\}.$$

The *excess risk* of a given hypothesis $w$ is defined to be

$$R_\theta(w) := L_\theta(w) - L_\theta(w^*).$$

Notice that in general $w^*$ depends on the distribution $p_\theta$, whereas $w$ only has access to a finite number of samples.

We consider two different learning scenarios.

- 1) **Supervised learning.** Let $w_{Z^n}$ denote the hypothesis generated by labeled data $Z^n = (Z_1, \ldots, Z_n)$. Assume $Z_i$ are i.i.d data distributed according to $p_{\theta_0}(z)$ where $\theta_0$ is the true parameter. The optimal expected excess risk of supervised learning is

  $$R_{SL}(\theta_0) := \min_w \mathbb{E} \{R_{\theta_0}(w_{Z^n})\},$$

  where the expectation is taken with respect to the labelled data $Z^n$.
- 2) **Semi-supervised learning (SSL).** Let $w_{Z^n, \tilde{X}^m}$ denote the hypothesis generated with the labeled data $Z^n = (Z_1, \ldots, Z_n)$ and additional unlabeled data $\tilde{X}^m = (\tilde{X}_1, \ldots, \tilde{X}_m)$. Also assume that $\tilde{X}_i$ are i.i.d data distributed according to $p_{\theta_0}(x)$ with $\theta_0$ being the true parameter. The optimal expected excess risk of supervised learning is

  $$R_{SSL}(\theta_0) := \min_w \mathbb{E} \left\{R_{\theta_0}(w_{Z^n, \tilde{X}^m})\right\},$$

  where the expectation is taken with respect to the labeled data $Z^n$ and unlabeled data $\tilde{X}^m$.

Throughout the paper, we assume that the density function $p_{\theta_0}(x, y)$ does *not* depend on the number of samples $n$ an $m$.

In this work, we state all our results for some given $\theta_0$. We point out that it is also possible to work with a minimax setup by defining the minimax excess risk as $R_{SL} := \min_w \max_\theta \mathbb{E} \{R_\theta(w_{Z^n})\}$ and so on. Similar results can be derived within the minimax problem formulations.

## 3 UPPER BOUNDS ON RISKS

In the following, we give upper bounds on $R_{SL}$ and $R_{SSL}$ in terms of (conditional) mutual information involving $Z^n, \tilde{X}^m$ and an auxiliary random variable $\Theta$ defined over $\Lambda$. Recall that the (conditional) mutual information $I(X; Y|Z)$ is defined as

$$I(X; Y|Z) := \int p_{XYZ}(x, y, z) \log \frac{p_{Y|XZ}(y|x, z)}{p_{Y|Z}(y|z)} dxdydz.$$

We will use the notation $I(X = x; Y|Z)$ to denote mutual information conditioned on $X = x$

$$I(X = x; Y|Z) := \int p_{YZ|X}(y, z|x) \log \frac{p_{Y|XZ}(y|x, z)}{p_{Y|Z}(y|z)} dydz, \quad (2)$$

which can also be written as

$$I(X = x; Y|Z) = D(p_{Y|X=x,Z}||p_{Y|Z}|p_{Z|X=x}), \quad (3)$$

where $D(p_{X|Y}||q_{X|Y}|r_Y)$ denotes the conditional Kullback-Leibler (KL) divergence

$$D(p_{X|Y}||q_{X|Y}|r_Y) := \int r_Y(y) D(p_{X|Y=y}||q_{X|Y=y}) dy.$$

The upper bounds to be derived are inspired by the classical universal prediction problem, where a non-negative quantity called *minimax redundancy* plays an important role. It is the smallest possible worst-case difference between the risk incurred by a universal predictor and that incurred by a predictor that *knowns* the true distribution of the data. It is well known that the minimax redundancy is equivalent to the maximin redundancy under some assumptions on the loss function [Gallager, 1974], which can be characterized as the capacity of a "channel" (hence in the form of mutual information), where the input is the parameter that characterizes the distribution, and the output is the generated data. The next two lemmas could also be interesting on their own due to their connection to the information-theoretic quantity.

We first give an upper bound when the loss function belongs to the class of exponentially concave functions, defined as follows.

**Definition 1 (Exponentially concave function)** *A function $f : \mathcal{D} \to \mathbb{R}$ is called a $\beta$-exponentially concave function in $x \in \mathcal{D}$ for some $\beta > 0$ if $\exp(-\beta f(x))$ is concave.*

The class of exponentially concave (exp-concave) functions are widely used as loss functions in machine learning problems. For example, it is easy to verify that the square loss $(b - x)^2$ is $1/(8a^2)$-exp-concave if the absolute value of $b, x$ are no larger than $a$. It is also shown in [Alirezaei and Mathar, 2018] that both discrete entropy and Renyi entropy, when appropriately scaled, are exp-concave functions. Another important 1-exp-concave function is the so-called *self-information* loss function [Merhav and Feder, 1998], is defined as $\ell(w(x), y) = -\log w(y)$ where $w$ is a probability assignment (depending on $x$) of $y$. In other words, $w$ can be thought as a length-$|\mathcal{Y}|$ nonnegative vector summing up to 1, and $w(y)$ returns the value of the entry corresponding to $y$. Furthermore, the cross entropy loss function can also be shown to be exponentially concave.

**Lemma 1 (Upper bound on risk for exp-concave loss)** *Assume that $\ell(w, z)$ is a $\beta$-exponentially concave function of $w$ for all $z$. Then for any true parameter $\theta_0 \in \Lambda$, it holds that*

$$R_{SL}(\theta_0) \leq \frac{1}{\beta} I(\Theta = \theta_0; Y'|X^n, Y^n, X'),$$

*where the distribution of $(\Theta, X', Y', X^n, Y^n)$ is given by $q(\theta)p_\theta(x', y') \prod_{i=1}^n p_\theta(x_i, y_i)$ for any choice of $q(\theta)$. It also holds that*

$$R_{SSL}(\theta_0) \leq \frac{1}{\beta} I(\Theta = \theta_0; Y'|X^n, Y^n, \tilde{X}^m, X'),$$

*where the distribution of $(\Theta, X', Y', X^n, Y^n, \tilde{X}^m)$ is given by $q(\theta)p_\theta(x', y') \prod_{i=1}^n p_\theta(x_i, y_i) \prod_{j=1}^m p_\theta(\tilde{x}_j)$ for any choice of $q(\theta)$.*

**Remark 1** *Instead of using the classical empirical risk minimization (ERM) approach to generate the hypothesis, we use a Bayes method for the prediction (cf. Equation (4) in the proof). In the context of universal prediction, this method is shown to produce an optimal universal predictor under appropriate conditions, in the sense that the average excess risk vanishes as the number of samples increases ([Merhav and Feder, 1998], [Clarke and Barron, 1990]).*

**Proof:** Let $Z^n$ denote $n$ pairs of i.i.d. data representing the training data and $Z' = (X', Y')$ another i.i.d. pair representing the test data. Recall that

$$R_{SL}(\theta_0) = \min_w \mathbb{E}_{\theta_0} \left\{ \ell(w, Z') - \ell(w^*, Z') \right\}.$$

To obtain an upper bound to the above quantity, for each $x'$, we choose the hypothesis $w_{Z^n}$ to be

$$\hat{w}_{Z^n}(x') := \text{argmin}_w \mathbb{E}_Q \left\{ \ell(w(x'), Y') | X^n, Y^n, X' = x' \right\}, \quad (4)$$

where the distribution $Q$ over $\mathcal{X}^{n+1} \times \mathcal{Y}^{n+1}$ is chosen to be

$$Q(x^{n+1}, y^{n+1}) := \int \prod_{i=1}^{n+1} p_\theta(x_i, y_i) q(\theta) d\theta \quad (5)$$

for some $q(\theta)$ that we can choose to suit our needs. More precisely, the term $\mathbb{E}_Q \left\{ \ell(w(x'), Y') | X^n, Y^n, X' = x' \right\}$ is given by

$$\mathbb{E}_Q \left\{ \ell(w(x'), Y') | X^n, Y^n, X' = x' \right\} = \sum_{y'} Q(y'|X^n, Y^n, x') \ell(w(x'), y'),$$

where the conditional distribution $Q(y'|x^n, y^n, x')$ is induced from $Q(x^{n+1}, y^{n+1})$ defined in (5). Notice that $\hat{w}_{z^n}$ does not depend on $\theta_0$. With this choice, we have

$$R_{SL}(\theta_0) \leq \mathbb{E}_{\theta_0}\left\{\ell(\hat{w}_{Z^n}, Z') - \ell(w^*, Z')\right\} =$$
$$\frac{1}{\beta_0}\sum_{z^n,x'} p_\theta(z^n, x') \sum_{y'} p_{\theta_0}(y'|z^n, x')(\beta\ell(\hat{w}_{Z^n}, z') - \beta\ell(w^*, z')).$$

Now we upper bound the term

$$\sum_{y'} p_{\theta_0}(y'|z^n, x')(\beta\ell(\hat{w}_{Z^n}, z') - \beta\ell(w^*, z'))$$
$$= \sum_{y'} p_{\theta_0}(y'|z^n, x')(\log \frac{e^{-\beta\ell(w^*(x'),y')}Q(y'|z^n, x')}{e^{-\beta\ell(\hat{w}_{Z^n}(x'),y')}p_{\theta_0}(y'|z^n, x')}$$
$$+ \log \frac{p_{\theta_0}(y'|z^n, x')}{Q(y'|z^n, x')})$$
$$\leq \log \sum_{y'} Q(y'|z^n, x')\frac{e^{-\beta\ell(w^*(x'),y')}}{e^{-\beta\ell(\hat{w}_{Z^n}(x'),y')}}$$
$$+ D(p_{\theta_0}(Y'|z^n, x')||Q(Y'|z^n, x'))$$
$$\leq D(p_{\theta_0}(Y'|z^n, x')||Q(Y'|z^n, x')).$$

The last inequality holds because

$$\sum_{y'} Q(y'|z^n, x')\frac{e^{-\beta\ell(w^*(x'),y')}}{e^{-\beta\ell(\hat{w}_{Z^n}(x'),y')}} \leq 1.$$

Indeed, as $\hat{w}_{Z^n}$ is chosen to be the minimizer of the expected value of $\ell(w(x'), y')$ under the distribution of $Q(y'|z^n, x')$, Lemma 3 (stated at the end of this section) shows that this expectation is smaller or equal to 1. Consequently,

$$\mathbb{E}\left\{\ell(\hat{w}_{Z^n}, Z') - \ell(w^*, Z')\right\}$$
$$\leq \frac{1}{\beta}\sum_{z^n,x'} p_{\theta_0}(z^n, x')D(p_{\theta_0}(Y'|z^n, x')||Q(Y'|z^n, x'))$$
$$= \frac{1}{\beta}D(p_{\theta_0}(Y'|Z^n, X')||Q(Y'|Z^n, X')|p_{\theta_0}(Z^n, X'))$$
$$= \frac{1}{\beta}I(Y'; \Theta = \theta_0|Z^n, X'),$$

where the last equality holds because the choice of the distribution $Q$ in (5). To see this, recall the representation in (3). In this expression, replace $X$ with $\theta$, $Y$ with $Y'$, and $Z$ with $Z^n, X'$ for our argument. It can be easily verified that due to the choice of $Q$ in (5), we have the claimed result.

The derivation of the upper bound on $R_{SSL}$ is similar to the above derivations, and we only highlight the difference. Similarly, the hypothesis$\hat{w}_{Z^n,\tilde{X}^m}$ in the SSL case is chosen to be

$$\hat{w}_{Z^n,\tilde{X}^m}(x')$$
$$:= \text{argmin}_w \mathbb{E}_Q\left\{\ell(w(x'), Y')|X^n, Y^n, \tilde{X}^m, X' = x'\right\}$$

where the distribution $Q$ over $\mathcal{X}^{n+1} \times \mathcal{Y}^{n+1} \times \mathcal{X}^m$ is chosen to be

$$Q(x^{n+1}, y^{n+1}, \tilde{x}^m) := \int \prod_{j=1}^{m} p_\theta(\tilde{x}_j) \prod_{i=1}^{n+1} p_\theta(x_i, y_i)q(\theta)d\theta.$$

We have

$$R_{SSL}(\theta_0) \leq \mathbb{E}_{\theta_0}\left\{\ell(\hat{w}_{Z^n,\tilde{X}^m}, Z') - \ell(w^*, Z')\right\}$$
$$= \frac{1}{\beta_0}\sum_{z^n,\tilde{x}^m,x'} p_{\theta_0}(z^n, \tilde{x}^m, x') \sum_{y'} p_\theta(y'|z^n, \tilde{x}^m, x')$$
$$\cdot (\beta\ell(\hat{w}_{Z^n,\tilde{X}^m}, z') - \beta\ell(w^*, z'))$$
$$\leq \frac{1}{\beta}D(p_{\theta_0}(Y'|Z^n, \tilde{X}^m, X')||Q(Y'|Z^n, \tilde{X}^m)|p_{\theta_0}(Z^n, \tilde{X}^m, X')).$$

The last inequality holds because we can show

$$\sum_{y'} p_{\theta_0}(y'|z^n, \tilde{x}^m, x')(\beta\ell(\hat{w}_{Z^n,\tilde{X}^m}, z') - \beta\ell(w^*, z'))$$
$$\leq D(p_{\theta_0}(Y'|z^n, \tilde{x}^m, x')||Q(Y'|z^n, \tilde{x}^m, x'))$$

in the same way as in the proof of $R_{SL}$ using Lemma 3. This concludes the proof. □

There is one important class of loss function which is not covered in the above lemma, namely the $0-1$ function. This loss function is mostly used in classification problems where the alphabet $\mathcal{Y}$ is a finite set, defined as

$$\ell(w(x), y) = \begin{cases} 0 & \text{if } w(x) = y \\ 1 & \text{otherwise} \end{cases} . \qquad (6)$$

The following lemma shows that the same result holds with one additional assumptions on the distribution $p_\theta(y|x)$ .

**Lemma 2 (Upper bound on risk for $0-1$ loss)** *Assume that $\ell(w(x), y)$ is the $0-1$ loss function defined in (6) where $\mathcal{Y}$ is a finite set. Also assume that $p_\theta(y|x) > 0$ for all $x, y$ and $\theta \in \Lambda$. Then the bounds on $R_{SL}$ and $R_{SSL}$ in Lemma 1 hold with some appropriately chosen $\beta > 0$.*

**Proof:** The proof proceeds in the same way as in the proof of Lemma 1. In particular, we choose the hypothesis $\hat{w}_{z^n}$ as in (4). With this choice, we can upper bound $R_{SL}$ as

$$R_{SL}(\theta_0) \leq \mathbb{E}_{\theta_0}\left\{\ell(\hat{w}_{Z^n}, Z') - \ell(w^*, Z')\right\}$$
$$= \sum_{z^n,x'} p_{\theta_0}(z^n, x') \sum_{y'} p_{\theta_0}(y'|z^n, x')(\ell(\hat{w}, z') - \ell(w^*, z')).$$

In the following bound, we take logarithm with the base $\alpha$ for some $\alpha > 1$ to be determined later. We use $D_\alpha$ and

$I_\alpha$ to denote the KL divergence and mutual information where the logarithm is with base $\alpha > 1$. It holds that

$$\begin{aligned}
&\sum_{y'} p_{\theta_0}(y'|z^n, x')(\ell(\hat{w}_{Z^n}, z') - \ell(w^*, z')) \\
&= \sum_{y'} p_{\theta_0}(y'|z^n, x') \Bigg( \log_\alpha \frac{\alpha^{-\ell(w^*,z')} Q(y'|z^n, x')}{\alpha^{-\ell(\hat{w}_{Z^n},z')} p_{\theta_0}(y'|z^n, x')} \\
&+ \log_\alpha \frac{p_{\theta_0}(y'|z^n, x')}{Q(y'|z^n, x')} \Bigg) \\
&\leq \log_\alpha \sum_{y'} Q(y'|z^n, x') \frac{\alpha^{-\ell(w^*(x'),y')}}{\alpha^{-\ell(\hat{w}_{Z^n}(x'),y')}} \\
&+ D_\alpha(p_{\theta_0}(Y'|z^n, x')||Q(Y'|z^n, x')) \\
&\leq D_\alpha(p_{\theta_0}(Y'|z^n, x')||Q(Y'|z^n, x')),
\end{aligned}$$

if we can show that

$$\sum_{y'} Q(y'|z^n, x') \frac{\alpha^{-\ell(w^*(x'),y')}}{\alpha^{-\ell(\hat{w}_{Z^n}(x'),y')}} \leq 1.$$

Notice that Lemma 3 does not apply here when $\ell(w(x), y) = \mathbf{1}_{w(x)\neq y}$. To show the above inequality, we use $\hat{y}$ to denote $\hat{w}_{Z^n}(x')$ and $y^*$ to denote $w^*(x')$ because both $\hat{w}$ and $w^*$ belongs to $\mathcal{Y}$. We can rewrite the LSH of the above inequality as

$$\sum_{y'\neq\hat{y}, y'\neq y^*} Q(y'|z^n, x') + Q(y^*|z^n, x')\alpha + Q(\hat{y}|z^n, x')\alpha^{-1}$$

because $\ell(y^*, y') = \ell(\hat{y}, y') = 1$ if $y' \neq y^*$ and $y' \neq \hat{y}$. Then the desired inequality is satisfied if it holds that

$$\begin{aligned}
\sum_{y'\neq\hat{y}, y'\neq y^*} & Q(y'|z^n, x') + Q(y^*|z^n, x')\alpha \\
&+ Q(\hat{y}|z^n, x')\alpha^{-1} \leq 1,
\end{aligned}$$

or equivalently

$$\frac{Q(\hat{y}|z^n, x')}{Q(y^*|z^n, x')} \leq \alpha. \tag{7}$$

In the following, we will show that there exists some $\alpha > 1$ independent of $n$ such that (7) holds. To this end, it is sufficient to show that the ratio on the LHS is upper bounded by some constant independent of $n$, so that choosing $\alpha$ to be larger than this constant will satisfy the desired inequality[1] . Furthermore, as it holds $Q(\hat{y}|z^n, x') \leq 1$, we only need to show that $Q(y^*|z^n, x')$ is lower bounded by some positive constant independent of $n$. This claim is proved in Section A in the Supplementary Materials.

Consequently,

$$\begin{aligned}
&\mathbb{E}\left\{\ell(\hat{w}_{Z^n}, Z') - \ell(w^*, Z')\right\} \\
&\leq \sum_{z^n, x'} p_{\theta_0}(z^n, x') D_\alpha(p_{\theta_0}(Y'|z^n, x')||Q(Y'|z^n, x')) \\
&= D_\alpha(p_{\theta_0}(Y'|Z^n, X')||Q(Y'|Z^n, X')|p_{\theta_0}(Z^n, X')) \\
&= I_\alpha(Y'; \Theta = \theta_0|Z^n, X') \\
&= I(Y'; \Theta = \theta_0|Z^n, X')/\log\alpha,
\end{aligned}$$

where in the last step we use the change of base again. Setting $\beta = \log\alpha > 0$ gives the claimed result for $R_{SL}$. The proof of $R_{SSL}$ follows an almost identical argument. □

The following result was used in the proof of Lemma 1.

**Lemma 3** *Let $w^*$ be the minimizer of $\mathbb{E}_Q\{\ell(w, Z)\}$ where the expectation over $Z$ is taken with respect to the distribution $Q$. Let $w'$ be any other choice of the hypothesis. If $\ell(w, z)$ is $\beta$-exp-concave for all $z$ with $w > 0$ then it holds that*

$$\mathbb{E}_Q\left\{\frac{g(w', Z)}{g(w^*, Z)}\right\} \leq 1,$$

*where $g(w, z) := \exp\{-\beta\ell(w, z)\}$.*

**Proof:** Let $w_\lambda = (1-\lambda)w^* + \lambda w'$ be a deviation from the minimizer $w^*$ to another predictor $w'$ characterized by $\lambda$. By the optimality condition of $w^*$, we have

$$\frac{d\mathbb{E}_Q\{-\beta\ell(w_\lambda, X)\}}{d\lambda}\Big|_{\lambda=0^+} \leq 0$$

Notice that

$$\begin{aligned}
&\frac{d\mathbb{E}_Q\{-\beta\ell(w_\lambda, X)\}}{d\lambda}\Big|_{\lambda=0^+} \\
&= \lim_{\lambda\to 0} \frac{1}{\lambda}\mathbb{E}\left\{-\beta\ell(w_\lambda, X) + \beta\ell(w^*, X)\right\} \\
&= \lim_{\lambda\to 0} \frac{1}{\lambda}\mathbb{E}\left\{\log\frac{g(w_\lambda, X)}{g(w^*, X)}\right\} \\
&\geq \lim_{\lambda\to 0} \frac{1}{\lambda}\mathbb{E}\left\{\log\frac{(1-\lambda)g(w^*, X) + \lambda g(w', X)}{g(w^*, X)}\right\} \\
&= \mathbb{E}\left\{\frac{g(w', X)}{g(w^*, X)}\right\} - 1
\end{aligned}$$

where the inequality holds because $g$ is concave in $w$. □

# 4 THE CASE WHEN UNLABELED DATA IS EQUALLY USEFUL

In this section, we evaluate the mutual information terms in Lemma 1 to derive the asymptotic expression for ex-

[1]Notice that we already have $\frac{Q(\hat{y}|z^n,x')}{Q(y^*|z^n,x')} \geq 1$ due to the well-known result that the $\hat{y}$ that minimizes the expectation in (4) with the $0-1$ loss is the *maximum a posteriori* (MAP) estimator $\hat{y} = \text{argmax}_y Q(y|z^n, x')$.

cess risks with additional assumptions on the distributions $p_\theta$.

Define Fisher information matrices as

$$I_{XY}(\theta) := \mathbb{E}\left\{\frac{\partial}{\partial\theta_j}\log p(X,Y|\theta)\frac{\partial}{\partial\theta_k}\log p(X,Y|\theta)\right\}_{j,k}$$

$$I_X(\theta) := \mathbb{E}\left\{\frac{\partial}{\partial\theta_j}\log p(X|\theta)\frac{\partial}{\partial\theta_k}\log p(X|\theta)\right\}_{j,k}$$

for $j,k = 1,\ldots,d$.

The main condition we need on the distribution is:

*Condition 1:* Let $\theta_0$ denote the true parameter. The density $p_\theta(x)$ and $p_\theta(x,y)$ are twice continuously differentiable at $\theta_0$. The Fisher information matrices $I_{XY}(\theta_0)$ and $I_X(\theta_0)$ are positive definite, and it holds that $I_{XY}(\theta_0) \succ I_X(\theta_0)$ with respect to the positive definite ordering[2].

We also need the following technical conditions.

*Condition 2:* Assume that for all $\theta$ in some neighbourhood of $\theta_0$, the (normalized) Renyi divergences of order $1+\lambda$

$$\log\int p(x|\theta_0)^{1+\lambda}p(x|\theta)^{-\lambda}dx$$

$$\log\int p(x,y|\theta_0)^{1+\lambda}p(x,y|\theta)^{-\lambda}dxdy$$

are bounded for some small enough $\lambda > 0$.

*Condition 3:* Assume that for all $\theta$ in some neighbourhood of $\theta_0$, the moment generating function

$$\mathbb{E}\left\{e^{\lambda\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p(X|\theta)}\right\}, \mathbb{E}\left\{e^{\lambda\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p(X,Y|\theta)}\right\}$$

exist for all $j,k = 1,\ldots,d$ with some small $\lambda > 0$.

*Condition 4:* Let $l := \nabla\log p(X,Y|\theta_0)$, $\tilde{l} := \nabla\log p(X|\theta_0)$, and $l', \tilde{l}'$ an independent copy of $l$ and $\tilde{l}$, respectively. The moment generating functions

$$\mathbb{E}\left\{e^{\lambda l^T(I_{XY}(\theta_0)+I_X(\theta_0))l}\right\}, \mathbb{E}\left\{e^{\lambda l^T(I_{XY}(\theta_0)+I_X(\theta_0))l'}\right\}$$

$$\mathbb{E}\left\{e^{\lambda\tilde{l}^T(I_{XY}(\theta_0)+I_X(\theta_0))\tilde{l}}\right\}, \mathbb{E}\left\{e^{\lambda\tilde{l}^T(I_{XY}(\theta_0)+I_X(\theta_0))\tilde{l}'}\right\}$$

exist for some small enough $\lambda > 0$.

A few words are in order with regard to the above conditions. Condition 1 is crucial for our results. Notice that the density functions need to be twice continuously differentiable so we are only dealing with continuous parameters. The positive definiteness of Fisher information matrices is also a key assumption. In particular, the matrix $I_X(\theta_0)$ being positive definite means that the unlabeled data contains non-trivial information about the whole parameter vector $\theta_0$. Condition 2, 3, and 4 are technical conditions to ensure that the reminding terms of the approximation to mutual information term in Lemma 1 decays in a fast enough rate. We point out that though complicated-looking, the existence requirement of divergence and moment generating functions are in general easy to satisfy if Condition 1 holds. Furthermore, we expect that with a refined analysis, it may be possible to prove the same result without Condition 2, 3, and 4. Indeed, a proof outline is given in [Clarke, 2012] for a similar result without additional assumptions.

Now we are ready to state the main result of this section.

**Theorem 1 (Learning rate)** *Let $n, m$ be the number of labeled and unlabeled data, respectively. Assume that the loss function $\ell(w,z)$ is $\beta$-exponentially concave in $w$ for all $z$. Assuming that Condition 1, 2, 3, and 4 above hold, we have the following statements.*

1) *(Semi-supervised learning) Let $m = \alpha n$ for some $\alpha > 0$. It holds that*

$$R_{SSL}(\theta_0) \le \frac{K_1(\theta_0)}{2n} + o(1/n)$$

*where*

$$K_1(\theta_0) := \frac{1}{\beta}\mathrm{Tr}((I_{XY}(\theta_0)+\alpha I_X(\theta_0))^{-1}I_{XY}(\theta_0))) - \mathrm{Tr}((I_{XY}(\theta_0)+\alpha I_X(\theta_0))^{-1}I_X(\theta_0))).$$

2) *(Supervised learning) Let $m = 0$ . It holds that*

$$R_{SL}(\theta_0) \le \frac{K_2(\theta_0)}{2n} + o(1/n)$$

*where $K_2(\theta_0) := \frac{1}{\beta}(d - \mathrm{Tr}(I_{XY}^{-1}(\theta_0)I_X(\theta_0)))$.*

3) *(Semi-supervised learning with many unlabeled data) Let $m = n^{1+\gamma}$ for some $\gamma > 0$. It holds that*

$$R_{SLL}(\theta_0) \le \frac{K_3(\theta_0)}{2n^{1+\gamma}} + o(1/n^{1+\gamma})$$

*where $K_3(\theta_0) := \frac{1}{\beta}\mathrm{Tr}(I_X^{-1}(\theta_0)I_{XY}(\theta_0)) - d)$.*

*If the loss function $\ell(w,z)$ is the $0-1$ function and it holds that $p_\theta(y|x) > 0$ for all $x, y$ and $\theta \in \Lambda$, the above bounds hold with some appropriately chosen $\beta > 0$.*

**Remark 2** *It can be checked straightforwardly that we have $0 < K_1(\theta_0) \le K_2(\theta_0) \le K_3(\theta_0)$. Item 1) and 2) show that if the number of unlabeled data $m$ grows as $O(n)$ with $n$ being the number of labeled data, then the*

[2]Notice it always holds that $I_{XY}(\theta_0) \succeq I_X(\theta_0)$ by the chain rule of information matrix.

*learning rate for both supervised and semi-supervised learning converges as $O(1/n)$ where additional unlabeled data only improves the constant from $K_2$ to $K_1$.*

*Item 3) shows that when the number of unlabeled data $m$ is dominating $n$, then the convergence rate is $O(1/m)$ with a larger constant $K_3$. Hence the learning rate can be improved from $O(1/n)$ to $O(1/n^{1+\gamma})$ if the number of unlabeled data grows superlinearly with respect to $n$. In other words, unlabeled data is equally useful in terms of the convergence rate in this case, and the loss due to not having all data labeled is only shown in the constant ($K_3 \geq K_2$). We point out that a similar observation has also been made in [Göpfert et al., 2019].*

**Remark 3** *It is interesting to exam the upper bound in Item 1) (semi-supervised learning) for $d = 1$ when both $I_{XY}$ and $I_X$ are scalars. In this case, the upper bound takes the form*

$$R_{SLL}(\theta_0) \leq O\left(\frac{1}{nI_{XY}(\theta_0) - mI_X(\theta_0)}\right).$$

*In other words, one labeled data is $I_{XY}(\theta_0)/I_X(\theta_0)$ more valuable than unlabeled data as far as the convergence rate is concerned (notice that $I_{XY}(\theta_0)/I_X(\theta_0) \geq 1$) for the regime $m \sim n$. The same result was obtained in [Castelli and Cover, 1996] for the simple mixture model with $d = 1$. Our theorem extend this result to more general cases when it is not necessarily a mixture model.*

The proof of Theorem 1 relies on the following asymptotic characterization of the KL divergence between $p(Z^n, \tilde{X}^m|\theta)$ and the "mixture" distribution $Q(\cdot)$ defined as $Q(\cdot) := \int q(\theta)p(\cdot|\theta)d\theta$.

**Lemma 4 (Asymptotic expression of KL-divergence)** *Assume that Condition 1, 2, 3, and 4 hold. Let both $m, n$ increase in a way that either $m = \alpha n$ for some $\alpha > 0$, or $m = n^{1+\gamma}$ for some $\gamma > 0$. Then there exists a prior $q(\theta)$ so that*

$$\begin{aligned}
&D(p(Y^n, X^n, \tilde{X}^m|\theta)||Q(Y^n, X^n, \tilde{X}^m)) \\
&= \frac{d}{2}\log\frac{1}{2\pi e} + \log\frac{1}{q(\theta)} \\
&+ \frac{1}{2}\log|nI_{XY}(\theta) + mI_X(\theta)| + o(1/\max\{n, m\})
\end{aligned}$$

*Under the same assumptions and let $m = 0$, we have*

$$\begin{aligned}
&D(p(Y^n, X^n|\theta)||Q(Y^n, X^n)) \\
&= \frac{d}{2}\log\frac{1}{2\pi e} + \log\frac{1}{q(\theta)} + \frac{1}{2}\log|nI_{XY}(\theta)| + o(1/n).
\end{aligned}$$

The same approximation result has been established in [Clarke and Barron, 1990] where the authors showed that the reminder term vanishes as $n \to \infty$. In our case, we need to show that it vanishes with a fast enough rate $o(1/n)$. This lemma is proved in Section B in the Supplementary Materials.

Equipped with Lemma 4, we are ready to give a proof of Theorem 1.

**Proof:** We first show the upper bound on $R_{SL}$. Using the chain rule of mutual information, we have

$$I(Y'; \Theta|Z^n, X') = I(Y', X', Z^n; \Theta) - I(Z^n, X'; \Theta).$$

Using Lemma 4, we have

$$\begin{aligned}
&I(\Theta = \theta_0; Z^n, X') = D(p_{\theta_0}(X', X^n, Y^n)||Q(X', X^n, Y^n)) \\
&= \frac{d}{2}\log\frac{1}{2\pi e} + \log\frac{\sqrt{|nI_{XY}(\theta_0) + I_X(\theta_0)|}}{q(\theta_0)} + o(1/n) \\
&= \frac{d}{2}\log\frac{1}{2\pi e} + \log\frac{\sqrt{|nI_{XY}(\theta_0)|}}{q(\theta_0)} \\
&+ \log\frac{\sqrt{|\mathbf{I} + \frac{1}{n}I_{XY}^{-1}(\theta_0)I_X(\theta_0)|}}{q(\theta_0)} + o(1/n).
\end{aligned}$$

By noticing that $(Y', X', Z^n)$ has the same distribution as $Z^{n+1}$, we have

$$\begin{aligned}
&I(\Theta = \theta_0; Y', X', Z^n) = I(\Theta = \theta_0; Z^{n+1}) \\
&= D(p_{\theta_0}(X^{n+1}, Y^{n+1})||Q(X^{n+1}, Y^{n+1})) \\
&= \frac{d}{2}\log\frac{1}{2\pi e} + \log\frac{\sqrt{|(n+1)I_{XY}(\theta_0)|}}{q(\theta_0)} + o(1/n).
\end{aligned}$$

Hence

$$\begin{aligned}
&I(\Theta = \theta_0; Y'|X^n, Y^n, X') \\
&= I(\Theta = \theta_0; Y', X', Z^n) - I(\Theta = \theta_0; X', Z^n) \\
&= \frac{d}{2}\log\frac{n+1}{n} - \frac{1}{2}\log|\mathbf{I} + \frac{1}{n}I_{XY}^{-1}(\theta_0)I_X(\theta_0)| + o(1/n).
\end{aligned}$$

Using the expansion of determinant:

$$\left|\mathbf{I} + \frac{1}{n}A\right| = 1 + \frac{1}{n}\mathrm{Tr}(A) + o(1/n), \tag{8}$$

we have

$$\begin{aligned}
&I(\Theta = \theta_0; Y'|X^n, Y^n, X') \\
&\leq \frac{d}{2}\log\frac{n+1}{n} - \frac{1}{2}\log\Big(1 + \frac{1}{n}\mathrm{Tr}(I_{XY}^{-1}(\theta_0)I_Y(\theta_0)) \\
&+ o(1/n)\Big) + o(1/n) \\
&= \frac{d}{2n} - \frac{\mathrm{Tr}(I_{XY}^{-1}(\theta_0)I_X(\theta_0))}{2n} + o(1/n)
\end{aligned}$$

where we use the fact $\log(1+\frac{C}{n}) = \frac{C}{n} + o(1/n)$ for some $C > 0$. The bound on $R_{SL}$ follows from Lemma 1 by defining $K_2$ as in the theorem.

To bound $R_{SSL}$ for the case $m = \alpha n$, we have

$$\begin{aligned}
&I(\Theta = \theta_0; Y'|X^n, Y^n, \tilde{X}^m, X') \\
&= I(\Theta = \theta_0; X', Y', X^n, Y^n, \tilde{X}^m) \\
&- I(\Theta = \theta_0; X^n, Y^n, \tilde{X}^m, X') \\
&= \frac{1}{2}\log|(n+1)I_{XY}(\theta_0) + \alpha n I_X(\theta_0)| \\
&- \frac{1}{2}\log|nI_{XY}(\theta_0) + (\alpha n+1)I_X(\theta_0)| + o(1/n) \\
&= \frac{1}{2}\log|\mathbf{I} + \frac{1}{n}A^{-1}I_{XY}(\theta_0)| - \frac{1}{2}\log|\mathbf{I} + \frac{1}{n}A^{-1}I_{XY}(\theta_0)| \\
&+ o(1/n)
\end{aligned}$$

where $A := I_{XY}(\theta_0) + \alpha I_X(\theta_0)$. Using (8) again together with $\log(1 + C/n) = C/n + o(1/n)$, we obtain the claimed constant $K_1(\theta_0)$.

For the case $m = n^{1+\gamma}$, Lemma 4 shows that

$$\begin{aligned}
&I(\Theta = \theta_0; Y'|X^n, Y^n, \tilde{X}^m, X') \\
&= I(\Theta = \theta_0; X', Y', X^n, Y^n, \tilde{X}^m) \\
&- I(\Theta = \theta_0; X^n, Y^n, \tilde{X}^m, X') \\
&= \frac{1}{2}\log|\mathbf{I} + \frac{n+1}{n^{1+\gamma}}I_X^{-1}(\theta_0)I_{XY}(\theta_0)| \\
&- \frac{1}{2}\log|\mathbf{I} + \frac{1}{n^{1+\gamma}}I_X^{-1}(\theta_0)(nI_{XY}(\theta_0)) + I_X(\theta_0)| \\
&+ o(1/n^{1+\gamma}).
\end{aligned}$$

Using (8) and $\log(1 + C/n) = C/n + o(1/n)$ again, the main terms in the above expression simplifies to

$$\begin{aligned}
&\frac{n+1}{2n^{1+\gamma}}\mathrm{Tr}(I_X^{-1}(\theta_0)I_{XY}(\theta_0)) - \frac{n\mathrm{Tr}(I_X^{-1}(\theta_0)I_{XY}(\theta_0)) + d}{2n^{1+\gamma}} \\
&= \frac{\mathrm{Tr}(I_X^{-1}(\theta_0)I_{XY}(\theta_0)) - d}{2n^{1+\gamma}},
\end{aligned}$$

which gives the desired constant $K_3(\theta_0)$. □

## 5 LOWER BOUND

Theorem 1 only gives upper bounds on the learning rate, and it is natural to ask whether the $O(1/n)$ rate for supervised learning is optimal. In other words, whether the rate improvement in semi-supervised learning is genuinely due to additional unlabeled data, and not possible with labeled data alone. The next result shows this is indeed the case. Specifically, we show for a certain type of loss function, the mutual information characterization in Lemma 1 is in fact exact for the worst-case $\theta_0$.

To formulate the result, define the maximin excess risk of the supervised learning algorithm to be

$$R_{SL}^* := \max_{\theta_0} R_{SL}(\theta_0) = \max_{\theta_0}\min_{w} \mathbb{E}\left\{R_{\theta_0}(w_{Z^n})\right\}.$$

We will consider the self-information loss function mentioned in Section 3. Formally, for the case when $|\mathcal{Y}|$ is finite, the self-information loss function can be defined as $\ell(w_{Z^n}(x), y) = -\log w_{Z^n}(x)^T \bar{y}$ where $w_{Z^n}(x)$ is a length-$|\mathcal{Y}|$ probability vector (a vector with nonnegative entries and sum to 1) depending on $Z^n$ and $x$, and $\bar{y}$ is the "one-hot" vector of length-$|\mathcal{Y}|$ that consisting 1 at the entry with value $y$ and 0 otherwise. Namely $\ell(w_{Z^n}(x), y)$ returns a value in $[0, 1]$ denoting the predicted probability that $Y = y$. In the case when $Y$ is continuous, the self-information loss can be written as $-\log w(y)$ where $w$ denotes a distribution on $Y$.

**Lemma 5 (Exact excess risk for self-information loss)** *For the self-information loss function defined above, we have*

$$R_{SL}^* = \max_{q(\theta)} I(Y'; \Theta|X^n, Y^n, X')$$

*where the distribution of* $(\Theta, X', Y', X^n, Y^n)$ *is given by* $q(\theta)p_\theta(x', y')\prod_{i=1}^n p_\theta(x_i, y_i)$.

This lemma follows directly from the classical "redundancy-capacity" result in universal prediction, and a proof can be found in e. g. [Gallager, 1974].

Combined with the result in Lemma 4, we see that the $O(1/n)$ convergence rate is optimal for the self-information loss function in the worst case, and cannot be improved.

## 6 COMPARISONS AND EXAMPLES

In this section, we compare our results with several existing results in the literature, and comment on the difference in the problem formulation.

As mentioned in Remark 3, [Castelli and Cover, 1996] studied the mixture model where the individual density functions are known but the mixing parameter is unknown. It was shown that in this case, unlabeled and labeled data play the same role in terms of convergence rate. Theorem 1 extends this result to more general cases. More precisely, [Castelli and Cover, 1996] studied the distribution (formulated with our notation)

$$p(x, y|\theta) = (\theta f_1(x))^{\mathbf{1}_{y=1}}(\bar{\theta} f_2(x))^{\mathbf{1}_{y=2}}$$

where $\bar{\theta} := 1-\theta$ and $f_1$ and $f_2$ are two density functions. Take the example in [Castelli and Cover, 1996, Sec. III,

A] where $f_1(x) = 2x$ and $f_2(x) = 2(1-x)$ for $x \in [0,1]$ and $\theta_0 = 1/2$. It is straightforward to show that the conditions used in Theorem 1 are satisfied, and that $I_{XY}(\theta_0) = 4$ and $I_X(\theta_0) = 4/3$. In the case when $n \sim m$, Theorem 1 states that

$$R_{SSL} \leq O\left(\frac{1}{4n + \frac{4}{3}m}\right)$$

which recovers the result in [Castelli and Cover, 1996]. However, it is also shown in this paper that if both the association and the mixing parameters are unknown, then the learning rate, to the first order, decays exponentially fast with the number of labeled data if $e^n m^{-1} \to 0$ (see [Castelli and Cover, 1996, Thm. 2] for details). This observation cannot be deduced from our results because it is not covered in our problem formulation. Indeed, the assumption with unknown association can only be converted into our model by introducing *discrete* parameters, where our current model assumes that the unknown parameter takes continuous values in $\mathbb{R}^d$. The same comment also applies to the problem formulation in [Rigollet, 2007], where the notion of "cluster" is in spirit similar to a model with discrete unknown parameters, which is not handled in our problem formulation.

[Göpfert et al., 2019] studied different problem formulations under which the unlabeled data can improve the learning rate. In particular, the authors discussed three different approaches in the literature, namely *"improvements via idealistic SSL"*, *"improvements via sample size dependent classes"*, and *"improvements via easy marginal estimation"* (see the paper for detailed information). Our problem formulation does not fall under the first two categories, as we neither assume that the true marginal distribution of the label is known to the learning algorithm, nor allow the distribution depend on the number of samples. The reason for the improvement in Theorem 1 is because $X$ contains non-trivial information about the whole parameter vector $\theta_0$ (Condition 1), hence is closer to the third category. Nevertheless, our result confirms the observation in [Göpfert et al., 2019] that we can have non-trivial rate change in SSL if $m$ grows faster than $n$ (see, e. g. Example 3 in the paper). Lastly, we point out that our result does not contradict with the lower bound $O(1/\sqrt{n})$ in [Göpfert et al., 2019, Appendix B], because they allow the distribution to depend on the number of samples.

We provide one additional example to illustrate our result. Consider a model where $X, Y$ is given by

$$X = Y + Z$$

with $Y \sim \mathcal{N}(0, \sigma^2)$ and $Z \sim \mathcal{N}(\mu, 1)$ being independent. The (unknown) parameter $\theta_0$ is a two-dimensional vector $\theta_0 = (\mu, \sigma^2)$ with some $\sigma^2 > 0$ and $\mu \in \mathbb{R}$. We would like to predict $Y$ from $X$ given labeled data $(X_i, Y_i), i = 1, \ldots, n$ and possible unlabeled data $\tilde{X}_i, i = 1, \ldots, m$ with the self-information loss function $-\log w(y)$ where $w$ is a distribution on $\mathbb{R}$. Notice this is not a mixture model so it is not clear from previous results that the unlabeled data is useful. However we expect that additional $\hat{X}_i$ should be helpful as it does provide information about the variance $\sigma^2$, which is essentially what we need for predicting $Y$ under the self-information loss. Indeed, our result in Theorem 1 confirms the intuition. Straightforward calculation shows that the Fisher information matrices are

$$I_X = \begin{pmatrix} \frac{1}{\sigma^2+1} & 0 \\ 0 & \frac{2}{\sigma^2+1} \end{pmatrix}, \quad I_{XY} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}.$$

which are both positive definite and Condition 1 holds. We can also verify straightforwardly that Condition 2, 3, and 4 hold for this Gaussian model.

The constant $K_1$ defined in Theorem 1 is (in this case $\beta = 1$ because the self-loss information is 1-exp-concave)

$$\begin{aligned} K_1 &= \frac{\sigma^2}{1+\sigma^2+\alpha} + \frac{1}{1+\sigma^2+\alpha\sigma^2} \\ &\leq \frac{\max\{1,\sigma^2\}}{1+\sigma^2+\min\{1,\sigma^2\}\alpha} \end{aligned}$$

hence the semi-supervised learning rate scales as

$$O\left(\frac{1}{(1+\sigma^2)n + \min\{1,\sigma^2\}m}\right).$$

In this case, both labeled and unlabeled examples contribute to the learning rate in the same order. Labeled data is $\frac{1+\sigma^2}{\min\{1,\sigma^2\}}$ times more valuable than unlabeled data in terms of the leading constant.

Similarly, it is straightforward to show that

$$K_2 = 1, \quad K_3 = \sigma^2 + \frac{1}{\sigma^2},$$

where indeed $K_1 < K_2 < K_3$.

**Acknowledgements**

The author would like to thank the anonymous reviewers for their helpful feedback. The author would also like to thank Prof. Bertrand Clarke for valuable discussions.

## References

[Alirezaei and Mathar, 2018] Alirezaei, G. and Mathar, R. (2018). On Exponentially Concave Functions and

Their Impact in Information Theory. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10, San Diego, CA. IEEE.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

[Castelli and Cover, 1996] Castelli, V. and Cover, T. (1996). The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6):2102–2117.

[Chapelle et al., 2006] Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-supervised learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass. OCLC: ocm64898359.

[Clarke, 1989] Clarke, B. (1989). *Asymptotic cumulative risk and Bayes risk under entropy loss, with applications*. PhD thesis.

[Clarke, 1999] Clarke, B. (1999). Asymptotic normality of the posterior in relative entropy. *IEEE Transactions on Information Theory*, 45(1):165–176.

[Clarke, 2012] Clarke, B. (2012). Comment on Article by Sancetta. *Bayesian Analysis*, 7(1):37–44.

[Clarke and Barron, 1994] Clarke, B. S. and Barron, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60.

[Clarke and Barron, 1990] Clarke, B. S. and Barron, A, R. (1990). Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471.

[Gallager, 1974] Gallager, R. G. (1974). Source coding with side information and universal coding. In *unpublished manuscript; also presetned at International Symposium on Information Theory (ISIT)*.

[Göpfert et al., 2019] Göpfert, C., Ben-David, S., Bousquet, O., Gelly, S., Tolstikhin, I., and Urner, R. (2019). When can unlabeled data improve the learning rate? In *Conference on Learning Theory*, pages 1500–1518.

[Merhav and Feder, 1998] Merhav, N. and Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147.

[Rigollet, 2007] Rigollet, P. (2007). Generalization Error Bounds in Semi-supervised Classification Under the Cluster Assumption. *The Journal of Machine Learning Research*, 8:1369–1392.

[Singh et al., 2009] Singh, A., Nowak, R., and Zhu, J. (2009). Unlabeled data: Now it helps, now it doesn't. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1513–1520. Curran Associates, Inc.

[Zhu, 2008] Zhu, X. (2008). Semi-supservised leraning literature survey. *Technical Report 1530, University of Wisconsin-Madison*.