Squared-loss Mutual Information Regularization: A Novel Information-theoretic Approach to Semi-supervised Learning

Gang Niu GANG@SG.CS.TITECH.AC.JP
Wittawat Jitkrittum WITTAWATJ@GMAIL.COM
Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8552, Japan

Bo Dai BDAI6@GATECH.EDU
College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA

Hirotaka Hachiya HACCHAN@GMAIL.COM
Masashi Sugiyama SUGI@CS.TITECH.AC.JP
Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8552, Japan

Abstract

We propose *squared-loss mutual information regularization* (SMIR) for multi-class probabilistic classification, following the *information maximization principle*. SMIR is convex under mild conditions and thus improves the nonconvexity of mutual information regularization. It offers all of the following four abilities to semi-supervised algorithms: Analytical solution, out-of-sample/multi-class classification, and probabilistic output. Furthermore, novel generalization error bounds are derived. Experiments show SMIR compares favorably with state-of-the-art methods.

1. Introduction

Semi-supervised learning, which utilizes both labeled and unlabeled data for training, has attracted much attention over the last decade. Many semi-supervised assumptions have been made to extract information from unlabeled data. Among them, the *manifold assumption* (Belkin et al., 2006) is of vital importance. Its origin is the *low-density separation principle*.

However, this low-density separation principle is not the only way to go. A useful alternative is the *information maximization principle* (IMP). IMP comes from information maximization clustering (Agakov & Barber, 2006; Gomes et al., 2010; Sugiyama et al., 2011), where a probabilistic classifier is trained in an unsupervised manner, so that a given information measure between data and cluster assignments is maximized. These clustering methods have shown IMP is reasonable and powerful.

Following IMP, we propose an information-theoretic approach to semi-supervised learning. Specifically, the *squared-loss mutual information* (SMI) (Suzuki et al., 2009) is designated as the information measure to be maximized. Then, we introduce an SMI approximator with no logarithm inside (Sugiyama et al., 2011), and propose the model of *SMI regularization* (SMIR). Unlike maximizing the mutual information, SMIR is strictly convex under mild conditions and the unique globally optimal solution is accessible. Albeit we can employ any convex loss in principle, SMIR can get rid of logarithm in the involved optimization and guarantees the analytic expression of the globally optimal solution if we use the *squared difference of two probabilities* (Sugiyama, 2010). SMIR aims at *multi-class probabilistic classifiers* that possess the innate ability of multi-class classification with the probabilistic output, and no reduction from the multi-class case to the binary case (cf. Allwein et al., 2000) is needed. These classifiers can also naturally handle unseen data and need no explicit out-of-sample extension. To the best of our knowledge, SMIR is the only framework up to the present which leads to semi-supervised algorithms equipped with all these properties.

Furthermore, we establish two *data-dependent generalization error bounds* for a reduced SMIR algorithm based on the theory of *Rademacher averages* (Bartlett & Mendelson, 2002). Our error bounds can consider

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

not only labeled data but also unlabeled data. Thus, they can reflect the properties of the particular mechanism generating the data. Thanks to the analytical solution, our bounds also have closed-form expression even though they depend on the data in terms of the Rademacher complexity. Notice that previous bounds (Belkin et al., 2004; Cortes et al., 2008) just focus on the regression error, and none of semi-supervised algorithms hitherto have similar theoretical results.

The rest of this paper is organized as follows. First of all, we present preliminaries, and propose the model and algorithm of SMIR in Section 2. In Section 3, we derive the generalization error bounds. The comparisons to related works are in Section 4, and then the experiments are in Section 5.

2. Squared-loss Mutual Information Regularization (SMIR)

In this section, we propose the SMIR approach.

2.1. Preliminaries

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{1, \ldots, c\}$ where d and c are natural numbers, $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ have an underlying $p(\boldsymbol{x}, y)$ and $p(\boldsymbol{x}) > 0$ over $\mathcal{X}$. Given i.i.d. $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^l$ and $\{\boldsymbol{x}_i\}_{i=l+1}^n$ where $n = l + u$ and $l \ll u$, we aim at estimating $p(y \mid \boldsymbol{x})$. Then, we can classify any $\boldsymbol{x} \in \mathcal{X}$ to $\hat{y} = \arg\max_{y \in \mathcal{Y}} p(y \mid \boldsymbol{x})$.

As an information measure, *squared-loss mutual information* (SMI) (Suzuki et al., 2009) between random variables X and Y is defined by

$$\mathrm{SMI} := \frac{1}{2} \int_{\mathcal{X}} \sum\nolimits_{y \in \mathcal{Y}} p(\boldsymbol{x}) p(y) \left(\frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x}) p(y)} - 1 \right)^2 \mathrm{d}\boldsymbol{x}.$$

SMI is the *Pearson divergence* (Pearson, 1900) from $p(\boldsymbol{x}, y)$ to $p(\boldsymbol{x})p(y)$, while the *mutual information* (Shannon, 1948) is the *Kullback-Leibler divergence* (Kullback & Leibler, 1951) from $p(\boldsymbol{x}, y)$ to $p(\boldsymbol{x})p(y)$. They both belong to f-divergence (Ali & Silvey, 1966; Csiszár, 1967), and thus share similar properties. For instance, both of them are nonnegative, and take zero if and only if X and Y are independent.

In Sugiyama et al. (2011), a computationally-efficient unsupervised SMI approximator was proposed. By assuming a uniform class-prior probability $p(y) = 1/c$, SMI becomes

$$\mathrm{SMI} = \frac{c}{2} \int_{\mathcal{X}} \sum\nolimits_{y \in \mathcal{Y}} (p(y \mid \boldsymbol{x}))^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \frac{1}{2}. \quad (1)$$

Then, $p(y \mid \boldsymbol{x})$ is approximated by a kernel model:

$$q(y \mid \boldsymbol{x}; \boldsymbol{\alpha}) := \sum\nolimits_{i=1}^n \alpha_{y,i} k(\boldsymbol{x}, \boldsymbol{x}_i), \quad (2)$$

where $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_c\}$ and $\boldsymbol{\alpha}_y = (\alpha_{y,1}, \ldots, \alpha_{y,n})^\top$ are model parameters, and $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel. After approximating the expectation w.r.t. $p(\boldsymbol{x})$ in Eq. (1) by the empirical average, an SMI approximator is derived as

$$\widehat{\mathrm{SMI}} = \frac{c}{2n} \sum\nolimits_{y \in \mathcal{Y}} \boldsymbol{\alpha}_y^\top \boldsymbol{K}^2 \boldsymbol{\alpha}_y - \frac{1}{2},$$

where $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ is the kernel matrix.

2.2. Basic model

Instead of Eq. (2), we introduce an alternative kernel model for SMIR (the reason will be explained in Remark 1). Let the empirical kernel map (Schölkopf & Smola, 2001) be

$$\Phi_n : \mathcal{X} \mapsto \mathbb{R}^n, \boldsymbol{x} \mapsto (k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n))^\top,$$

the degree of $\boldsymbol{x}_i$ be $d_i = \sum_{j=1}^n k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, and the degree matrix be $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_n)$. We approximate the class-posterior probability $p(y \mid \boldsymbol{x})$ by[1]

$$q(y \mid \boldsymbol{x}; \boldsymbol{\alpha}) := \langle \boldsymbol{K}^{-1/2} \Phi_n(\boldsymbol{x}), \boldsymbol{D}^{-1/2} \boldsymbol{\alpha}_y \rangle, \quad (3)$$

where $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{j=1}^n a_j b_j$ is the inner product. Plugging (3) into Eq. (1) gives us an alternative SMI approximator:

$$\widehat{\mathrm{SMI}} = \frac{c}{2n} \mathrm{tr} \left(\boldsymbol{A}^\top \boldsymbol{D}^{-1/2} \boldsymbol{K} \boldsymbol{D}^{-1/2} \boldsymbol{A} \right) - \frac{1}{2}, \quad (4)$$

where $\boldsymbol{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_c) \in \mathbb{R}^{n \times c}$ is the matrix representation of model parameters.

Subsequently, we employ Eq. (4) to regularize a loss function $\Delta(p, q)$ that is convex w.r.t. q. More specifically, we have three objectives: (i) Minimize $\Delta(p, q)$; (ii) Maximize $\widehat{\mathrm{SMI}}$; (iii) Regularize $\boldsymbol{\alpha}$. Therefore, we formulate the optimization problem of SMIR as

$$\min_{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_c \in \mathbb{R}^n} \Delta(p, q) - \gamma \widehat{\mathrm{SMI}} + \lambda \sum\nolimits_{y \in \mathcal{Y}} \frac{1}{2} \|\boldsymbol{\alpha}_y\|_2^2, \quad (5)$$

where $\gamma, \lambda > 0$ are regularization parameters.

A remarkable characteristic of optimization (5) is its convexity, as long as the kernel function k is nonnegative and $\lambda > \gamma c / n$:

Theorem 1. *Assume that $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+$ and $\lambda > \gamma c / n$. Then optimization (5) is strictly convex, and there exists a unique globally optimal solution.*[2]

[1] Assume that $\boldsymbol{K}$ is full-rank, and then $\boldsymbol{K}^{-1/2}$ is well-defined. The Gaussian kernel matrix is full-rank as long as $\forall i \neq j, \boldsymbol{x}_i \neq \boldsymbol{x}_j$.

[2] In the rest of this paper, we will assume that k is nonnegative and $\lambda > \gamma c / n$. See Appendix A for the proof.

Remark 1. We introduced Eq. (3) due to the following reasons: (i) In principle, any kernel model linear w.r.t. $\boldsymbol{\alpha}_y$ may be used to approximate $p(y \mid \boldsymbol{x})$, and maximizing $\widehat{\mathrm{SMI}}$ alone must be non-convex. However, optimization (5) becomes convex if λ is large enough. Hence, only λ above a certain threshold is acceptable: The threshold of (3) is $\gamma c/n$. The threshold of (2) is $\|K\|_2^2 \cdot \gamma c/n$ where $\|K\|_2$ is the spectral norm of K. It depends upon all the training data thoroughly and is usually much larger than $\gamma c/n$. (ii) We found that (3) experimentally outperformed (2).

2.3. Proposed algorithm

Due to limited space, we give a brief derivation here.

We choose the squared difference of probabilities p and q as the loss function (Sugiyama, 2010):

$$\Delta_2(p,q) := \frac{1}{2}\int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} (p(y \mid \boldsymbol{x}) - q(y \mid \boldsymbol{x};\boldsymbol{\alpha}))^2 p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

It enables the analytical solution and facilitates our future theoretical analysis. Its empirical version is

$$\widehat{\Delta}_2 = \text{Const.} - \frac{1}{l}\sum_{i=1}^{l} q(y_i \mid \boldsymbol{x}_i) + \frac{1}{2l}\sum_{i=1}^{l}\sum_{y=1}^{c} (q(y \mid \boldsymbol{x}_i))^2. \tag{6}$$

Let $\boldsymbol{Y} \in \mathbb{R}^{l\times c}$ be the class indicator matrix for l labeled data and $\boldsymbol{B} = (\boldsymbol{I}_l; \boldsymbol{0}_{u\times l}) \in \mathbb{R}^{n\times l}$. Subsequently, Eq. (6) can be expressed by

$$\begin{aligned}\widehat{\Delta}_2 = \text{Const.} &- \frac{1}{l}\operatorname{tr}(\boldsymbol{Y}^\top \boldsymbol{B}^\top \boldsymbol{K}^{1/2}\boldsymbol{D}^{-1/2}\boldsymbol{A}) \\ &+ \frac{1}{2l}\operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{D}^{-1/2}\boldsymbol{K}^{1/2}\boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{K}^{1/2}\boldsymbol{D}^{-1/2}\boldsymbol{A}).\end{aligned} \tag{7}$$

Substituting Eq. (7) into optimization (5), we will get the following objective function:

$$\begin{aligned}\mathcal{F}(\boldsymbol{A}) = &-\frac{1}{l}\operatorname{tr}(\boldsymbol{Y}^\top \boldsymbol{B}^\top \boldsymbol{K}^{1/2}\boldsymbol{D}^{-1/2}\boldsymbol{A}) \\ &+ \frac{1}{2l}\operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{D}^{-1/2}\boldsymbol{K}^{1/2}\boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{K}^{1/2}\boldsymbol{D}^{-1/2}\boldsymbol{A}) \\ &- \frac{\gamma c}{2n}\operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{D}^{-1/2}\boldsymbol{K}\boldsymbol{D}^{-1/2}\boldsymbol{A}) + \frac{\lambda}{2}\operatorname{tr}(\boldsymbol{A}^\top\boldsymbol{A}).\end{aligned}$$

At last, by equating $\nabla\mathcal{F}$ to the zero matrix, we obtain the analytical solution to unconstrained optimization problem (5):

$$\begin{aligned}\boldsymbol{A}^*_{\mathcal{F}} = n\Big(&n\boldsymbol{D}^{-1/2}\boldsymbol{K}^{1/2}\boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{K}^{1/2}\boldsymbol{D}^{-1/2} + \lambda nl\boldsymbol{I}_n \\ &-\gamma lc\boldsymbol{D}^{-1/2}\boldsymbol{K}\boldsymbol{D}^{-1/2}\Big)^{-1}\boldsymbol{D}^{-1/2}\boldsymbol{K}^{1/2}\boldsymbol{B}\boldsymbol{Y}.\end{aligned}$$

We recommend to post-process the model parameters as

$$\boldsymbol{\beta}_y = n\pi_y \cdot \frac{\boldsymbol{K}^{-1/2}\boldsymbol{D}^{-1/2}\boldsymbol{\alpha}^*_y}{\boldsymbol{1}_n^\top \boldsymbol{K}^{1/2}\boldsymbol{D}^{-1/2}\boldsymbol{\alpha}^*_y},$$

where $\boldsymbol{\beta}_y$ is a normalized version of $\boldsymbol{\alpha}^*_y$, and π_y is an estimate of $p(y)$ based on labeled data. In addition, probability estimates should be nonnegative and thus our final solution can be expressed as follows (cf. Yamada et al., 2011):

$$\hat{p}(y \mid \boldsymbol{x}) = \frac{\max(0, \langle \Phi_n(\boldsymbol{x}), \boldsymbol{\beta}_y\rangle)}{\sum_{y'=1}^{c}\max(0, \langle \Phi_n(\boldsymbol{x}), \boldsymbol{\beta}_{y'}\rangle)}.$$

Although $q(y \mid \boldsymbol{x};\boldsymbol{\alpha}^*)$ might be negative or unnormalized, Kanamori et al. (2012) implies that minimizing Δ_2 could achieve the optimal non-parametric convergence rate from q to p, and when we have enough data q is automatically a probability (i.e., non-negative and normalized).

3. Generalization Error Bounds

To elucidate the generalization capability, we reduce SMIR to binary classification. Now, a class label y is ± 1, a single vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ is enough to construct a discriminative model, and we classify any $x \in \mathcal{X}$ to

$$\hat{y} = \operatorname{sign}(\langle \boldsymbol{K}^{-1/2}\Phi_n(x), \boldsymbol{D}^{-1/2}\boldsymbol{\alpha}\rangle).$$

Let us encode the information of class labels into $\boldsymbol{y} = (y_1, \ldots, y_l)^\top \in \mathbb{R}^l$. The solution is then

$$\begin{aligned}\boldsymbol{\alpha}^*_{\mathcal{F}} = n\Big(&n\boldsymbol{D}^{-1/2}\boldsymbol{K}^{1/2}\boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{K}^{1/2}\boldsymbol{D}^{-1/2} + \lambda nl\boldsymbol{I}_n \\ &-\gamma lc\boldsymbol{D}^{-1/2}\boldsymbol{K}\boldsymbol{D}^{-1/2}\Big)^{-1}\boldsymbol{D}^{-1/2}\boldsymbol{K}^{1/2}\boldsymbol{B}\boldsymbol{y},\end{aligned} \tag{8}$$

and for convenience, we define the decision function

$$f(\boldsymbol{x}) = \langle \boldsymbol{K}^{-1/2}\Phi_n(\boldsymbol{x}), \boldsymbol{D}^{-1/2}\boldsymbol{\alpha}^*_{\mathcal{F}}\rangle. \tag{9}$$

Let $\mathbb{E}$ and $\hat{\mathbb{E}}$ stand for the true and empirical expectations, $\ell(z) = (1 - \operatorname{sign}(z))/2$ be the *indicator loss*, and $\ell_\eta(z) = \min(1, \max(0, 1 - z/\eta))$ be the *surrogate loss*. We bound $\mathbb{E}\ell(yf)$ using the theory of *Rademacher averages* (Bartlett & Mendelson, 2002). If all labels are available for evaluation, we can evaluate $\hat{\mathbb{E}}\ell_\eta(yf)$ over all training data and bound $\mathbb{E}\ell(yf)$ more tightly. We state the theoretical result in Theorem 2 and prove it in Appendix B.

Theorem 2. *Assume that*

$$\exists B_k > 0, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}, k(\boldsymbol{x}, \boldsymbol{x}') \le B_k^2.$$

Let $\boldsymbol{\alpha}^*_{\mathcal{F}}$ and $f(\boldsymbol{x})$ be the optimal solution and the decision function defined in Eqs. (8) and (9) respectively, and

$$B_{\mathcal{F}} = \|\boldsymbol{D}^{-1/2}\boldsymbol{\alpha}^*_{\mathcal{F}}\|_2, B'_{\mathcal{F}} = \|\boldsymbol{K}^{-1/2}\boldsymbol{D}^{-1/2}\boldsymbol{\alpha}^*_{\mathcal{F}}\|_1.$$

For any $\eta > 0$ and $0 < \delta < 1$, with probability at least $1 - \delta$, we have

$$\mathbb{E}\ell(yf(\boldsymbol{x})) \leq \frac{1}{l}\sum_{i=1}^{l}\ell_\eta(y_i f(\boldsymbol{x}_i)) + \frac{2B_k B_{\mathcal{F}}}{\eta\sqrt{l}} + \min\left(3, 1 + \frac{4B_k^2 B'_{\mathcal{F}}}{\eta}\right)\sqrt{\frac{\ln(2/\delta)}{2l}}. \tag{10}$$

If the ground truth class labels $y_{l+1}, \ldots, y_n$ are also available for evaluation, with probability at least $1 - \delta$, we have

$$\mathbb{E}\ell(yf(\boldsymbol{x})) \leq \frac{1}{n}\sum_{i=1}^{n}\ell_\eta(y_i f(\boldsymbol{x}_i)) + \frac{2B_k B_{\mathcal{F}}}{\eta\sqrt{n}} + \min\left(3, 1 + \frac{4B_k^2 B'_{\mathcal{F}}}{\eta}\right)\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{11}$$

Theorem 2 gives the tightest upper bounds (i.e., the coefficients of $1/\sqrt{l}$ and $1/\sqrt{n}$ are smallest under each given scenario) based on the inductive Rademacher complexity. The bound in Eq. (10) is asymptotically $O(1/\sqrt{l})$, if we only know the first l labels. In such cases, we may benefit from unlabeled data by a lower empirical error. It becomes $O(1/\sqrt{n})$ in Eq. (11) if we can access the other u labels, even though they are not used for training. Due to the smaller deviation of the empirical error and the empirical Rademacher complexity when they are estimated over all training data, we can improve the order from $O(1/\sqrt{l})$ to $O(1/\sqrt{n})$. Nevertheless, there is no free lunch: In (11), the empirical error is evaluated over all training data, and it may be significantly higher than that evaluated over labeled data. Basically, (10) or (11) which right-hand side is smaller reflects whether the information maximization principle befits the data set or not.

4. Related Works

Information-theoretic semi-supervised approaches directly constrain $p(y \mid \boldsymbol{x})$ by unlabeled data or some $p(\boldsymbol{x})$ given as the prior knowledge. *Information regularization* (IR; Szummer & Jaakkola, 2002) is the pioneer for this purpose. Compared with later information maximization methods, IR minimizes the mutual information (MI) based on a key observation: Within a small region $Q \subset \mathcal{X}$, MI_Q is low/high if the label information is pure/chaotic. Subsequently, IR estimates a cover $\mathcal{C}$ of $\mathcal{X}$ from $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, and minimizes the maximal MI_Q for $Q \in \mathcal{C}$, subject to class constraints provided by labeled data. The advantage of IR is its flexibility and convexity, while the drawback is that it is unclear how to estimate $\mathcal{C}$ properly. Each region should be small enough to preserve the locality of the label information in a single region; each pair of regions should be connected to ensure the dependence of $p(y \mid \boldsymbol{x})$ over all regions, and this implies a great number of tiny regions.

By employing the Shannon entropy of $p(y \mid \boldsymbol{x})$ as a measure of class overlap, *entropy regularization* (ER; Grandvalet & Bengio, 2004) minimizes the entropy from a viewpoint of *maximum a posteriori* estimation. More specifically, ER regularizes the maximum log-likelihood estimation of a logistic regression or kernel logistic regression model by an entropy term:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{l}\ln q(y_i \mid \boldsymbol{x}_i; \boldsymbol{\alpha}) + \gamma\sum_{i=l+1}^{n}\sum_{y\in\mathcal{Y}} q(y \mid \boldsymbol{x}_i; \boldsymbol{\alpha})\ln q(y \mid \boldsymbol{x}_i; \boldsymbol{\alpha}).$$

ER favors low-density separations, since the low/high entropy means that the class overlap is mild/intensive. ER and IR seem opposite at a first glance, because MI equals the difference of the entropies of class prior and posterior. However, IR minimizes MI *locally* and ER minimizes the entropy *globally*, so both of them highly penalize the variations of the class-posterior probability in high-density regions. A recent framework called *regularized information maximization* (RIM; Gomes et al., 2010) follows ER and further maximizes the entropy of the class-prior probability to encourage balanced classes. ER and RIM do not model $p(\boldsymbol{x})$ explicitly which is a major improvement, but the disadvantage is the non-convexity of their optimizations.

Expectation regularization (XR; Mann & McCallum, 2007) goes one step further such that it does not use $p(\boldsymbol{x})$ at all. Therefore, XR does not favor low-density separations and can handle highly overlapped classes. XR encourages the predictions on unlabeled data to match a designer-provided expectation by minimizing the KL-divergence between the expectations predicted by the model and provided as the prior knowledge. If there is no prior knowledge, XR will match the class prior of unlabeled data with that of labeled data:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{l}\ln q(y_i \mid \boldsymbol{x}_i; \boldsymbol{\alpha}) - \lambda\sum_{y\in\mathcal{Y}}\frac{1}{2}\|\boldsymbol{\alpha}_y\|_2^2 + \gamma\sum_{y\in\mathcal{Y}}\pi_y \ln\left(\sum_{i=l+1}^{n} q(y \mid \boldsymbol{x}_i; \boldsymbol{\alpha})\right),$$

where π_y is an estimate of $p(y)$ through labeled data, and $q(y \mid \boldsymbol{x}; \boldsymbol{\alpha})$ is a logistic or kernel logistic regression model. Unlike IR and ER, XR does not prefer low-density separations. As a result, XR cannot deal with low-dimensional data with nonlinear structures (such as the famous *two-moons* or *two-circles*), if there are not enough labeled data.

On the other hand, there are lots of geometric methods for semi-supervised learning. Please see Table 1

Table 1. Summary of existing semi-supervised learning methods.

	AS	OC	MC	PO
Geometric				
Transductive SVM (Joachims, 1999)	×	◯	△	×
Semi-supervised SVM (Bennett & Demiriz, 1998)	×	◯	△	×
Laplacian SVM (Belkin et al., 2006)	×	◯	△	×
Laplacian Regularized Least Squares (Belkin et al., 2006)	◯	◯	△	×
Markov Random Walks (Szummer & Jaakkola, 2001)	×	×	◯	◯
Local and Global Consistency (Zhou et al., 2003)	◯	△	◯	×
Spectral Graph Transducer (Joachims, 2003)	◯	×	×	×
Harmonic Energy Minimization (Zhu et al., 2003)	◯	×	×	◯
Sparse Eigenfunction Bases (Sinha & Belkin, 2009)	×	◯	×	×
Information-theoretic				
Information Regularization (Szummer & Jaakkola, 2002)	×	◯	◯	◯
Entropy Regularization (Grandvalet & Bengio, 2004)	×	◯	◯	◯
Expectation Regularization (Mann & McCallum, 2007)	×	◯	◯	◯
Regularized Information Maximization (Gomes et al., 2010)	×	◯	◯	◯
Squared-loss Mutual Information Regularization	◯	◯	◯	◯

AS: analytical solution OC: out-of-sample classification MC: multi-class classification PO: probabilistic output
◯: Yes ×: No △: Extension has been proposed

as a list of representative methods. Note that all geometric methods in Table 1 are in the style of either large margins or similarity graphs. According to Table 1, we could know that many methods based on similarity graphs (Szummer & Jaakkola, 2001; Zhou et al., 2003; Joachims, 2003; Zhu et al., 2003) are transductive, while the information-theoretic methods are all inductive; only two geometric methods (Szummer & Jaakkola, 2001; Zhou et al., 2003) could deal with multi-class data directly, while it is an inherent property of all information-theoretic methods. However, none of previous information-theoretic methods have analytical solutions, due to the logarithms in the entropy, MI or KL-divergence. Thanks to SMI, the proposed SMIR involves a strictly convex optimization problem with no logarithm inside and consequently it has the analytic expression of the unique globally optimal solution.

The similarity between ER and SMIR is intriguing. RIM followed ER historically. Nonetheless, if we start from MI maximization with the uniform $p(y)$, we will get ER as

$$\max_{\boldsymbol{\alpha}} \int_{\mathcal{X}} \sum\nolimits_{y\in\mathcal{Y}} q(y \mid \boldsymbol{x};\boldsymbol{\alpha}) \ln q(y \mid \boldsymbol{x};\boldsymbol{\alpha}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

Recall that SMI maximization under the assumption of the uniform $p(y)$ is expressed by

$$\max_{\boldsymbol{\alpha}} \int_{\mathcal{X}} \sum\nolimits_{y\in\mathcal{Y}} q(y \mid \boldsymbol{x};\boldsymbol{\alpha}) q(y \mid \boldsymbol{x};\boldsymbol{\alpha}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

As a consequence, they have the similar preference as the logarithm is strictly monotonically increasing. The vital difference is the convexity and the analytical solution: SMIR is convex and the globally optimal solution can be obtained analytically, whereas ER is non-convex so any locally optimal solution has to be found numerically.[3]

5. Experiments

In this section, we numerically evaluate SMIR. The specification of benchmark data sets is summarized in Table 2. Besides the four well-tried benchmarks in the first block (i.e., USPS, MNIST, 20Newsgroups and Isolet), there are eight benchmarks from a book entitled *Semi-Supervised Learning* (Chapelle et al., 2006)[4] in the second block, and eight benchmarks from the *UCI machine learning repository*[5] in the third block except that Senseval-2 is from a workshop for *word sense disambiguation*[6]. Detailed explanation of benchmarks is omitted due to lack of space. Our experiments consist of three parts:

Firstly, we compare SMIR with entropy regularization (ER; Grandvalet & Bengio, 2004) and expectation regularization (XR; Mann & McCallum, 2007). The probabilistic models are the logistic regression

$$q(y \mid \boldsymbol{x};\boldsymbol{\alpha}) \propto \exp\langle \boldsymbol{x}, \boldsymbol{\alpha}_y \rangle, \boldsymbol{\alpha}_y \in \mathbb{R}^d,$$

and the kernel logistic regression (Ker)

$$q(y \mid \boldsymbol{x};\boldsymbol{\alpha}) \propto \exp\langle \Phi_n(\boldsymbol{x}), \boldsymbol{\alpha}_y \rangle, \boldsymbol{\alpha}_y \in \mathbb{R}^n,$$

[3] SMIR may also be solved numerically in consideration of the computational efficiency for large n in practice.

[4] http://olivier.chapelle.cc/ssl-book/benchmarks.html.

[5] http://archive.ics.uci.edu/ml/.

[6] http://www.senseval.org/.

Table 2. Specification of benchmark data sets.

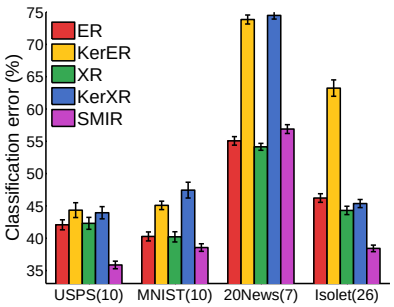
	# Classes	# Dimensions	# Data	Balance of classes (in %)
USPS	10	256	11000	10 per class
MNIST	10	784	70000	11.3 / 10.0 / 10.2 / 9.8 / 9.0 / 9.8 / 10.4 / 9.8 / 9.9 / 9.9
20Newsgroups	7	53975	11269	4.3 / 25.8 / 5.2 / 21.1 / 21.0 / 5.3 / 17.3
Isolet	26	617	7797	3.85 per class
g241c	2	241	1500	50.0 / 50.0
g241n	2	241	1500	50.1 / 49.9
Digit1	2	241	1500	51.1 / 48.9
USPS	2	241	1500	80.0 / 20.0
COIL	6	241	1500	16.7 per class
COIL2	2	241	1500	50.0 / 50.0
BCI	2	117	400	50.0 / 50.0
Text	2	11960	1500	50.0 / 50.0
Diabetes	2	8	768	65.1 / 34.9
Wine	3	13	178	33.1 / 39.9 / 27.0
Vowel	11	13	990	9.1 per class
Image	2	18	1155	42.9 / 57.1
Vehicle	4	18	846	25.1 / 25.7 / 25.8 / 23.5
German	2	20	1000	70.0 / 30.0
Satimage	6	36	6435	23.8 / 10.9 / 21.1 / 9.7 / 11.0 / 23.4
Senseval-2	3	50	534	33.3 per class

where $\langle \cdot, \cdot \rangle$ is the inner product, Φ_n is the empirical kernel map for the Gaussian kernel. SMIR also applies the Gaussian kernel, so there are three kernel methods which allow nonlinear decision boundaries in $\mathbb{R}^d$. The two-fold cross-validation is performed to select the hyperparameters. The kernel width is the median of all pairwise distances times the best value among $\{1/15, 1/10, 1/5, 1/2, 1\}$. A Gaussian prior of parameters, which is same as the third term of optimization (5), is included for XR and KerXR (Mann & McCallum, 2007). No extra prior is added to ER or KerER, since ER itself is a prior from a viewpoint of maximum a posteriori estimation (Grandvalet & Bengio, 2004). Therefore, ER/KerER has one regularization parameter whereas XR/KerXR and SMIR have two. The candidate list of regularization parameters is $10\hat{}\{-7, -3, -1, 1, 3\}$, except that λ is chosen from $\gamma c/n + 10\hat{}\{-10, -8, -6, -4, -2\}$ for SMIR to ensure the convexity. The *minFunc*[7] package for unconstrained optimization using line-search methods (the quasi-Newton limited-memory BFGS updates, by default) is utilized to solve ER/KerER and XR/KerXR. Since minimizing the entropy is non-convex, we initialize ER/KerER with the globally optimal solution of its supervised part.

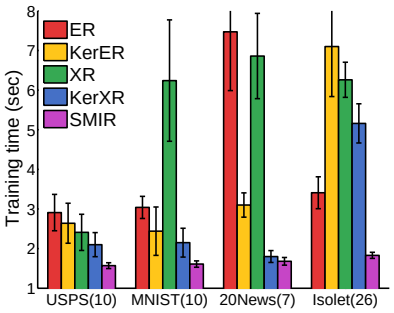
We evaluated them on USPS, MNIST, 20Newsgroups and Isolet. Pearson's correlation (Hall, 2000) was used to select 1000 most informative features for 20Newsgroups. For each data set, we prepared a multi-class task, namely, the tasks using 10 classes of USPS and MNIST, 7 classes of 20Newsgroups, and 26 classes of Isolet. In addition, extensive experiments of simple classification tasks were conducted, including 45 binary tasks of USPS, 45 binary tasks of MNIST and 21 binary tasks of 20Newsgroups. Isolet may lead to too many binary tasks and these tasks are often too easy, and thus we combined 26 letters into 13 groups (e.g., 'a' with 'b', 'c' with 'd' etc.) and treated each group as a single class resulting in 78 simple classification tasks. For each task, we repeatedly ran all methods on 100 random samplings, where the sample size was fixed to 500. Each random sampling was partitioned into a training set and a test set with 80% and 20% data, and 10% class labels of training data were revealed to construct labeled data.

Figure 1 reports the experimental results of the multi-class tasks, Figure 2 reports the experimental results of the simple tasks, and Table 3 summarizes the experimental results. We can see from Figure 1 that SMIR outperformed others on the multi-class tasks of USPS, MNIST and Isolet. Likewise Figure 1 indicates that SMIR was the most computationally-efficient algorithm on all four multi-class tasks. According to Figure 2, SMIR was the best on the simple tasks of USPS, 20Newsgroups and Isolet, but was slightly inferior to plain ER on MNIST. Note that there were 12 highly imbalanced tasks among 21 simple tasks of 20Newsgroups, which implies that the uniform class-prior assumption will not affect the performance of SMIR essentially, if the tasks are not so complicated. The experiments of Isolet further imply that SMIR is

[7] http://www.di.ens.fr/~mschmidt/Software/minFunc/.

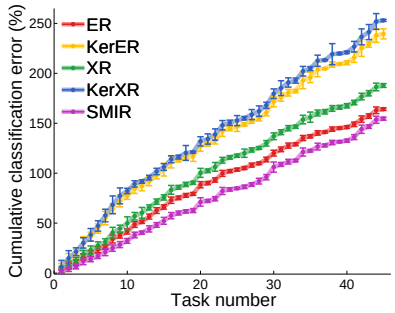


(a) Classification error

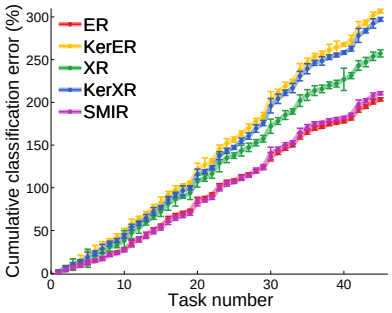


(b) Training time

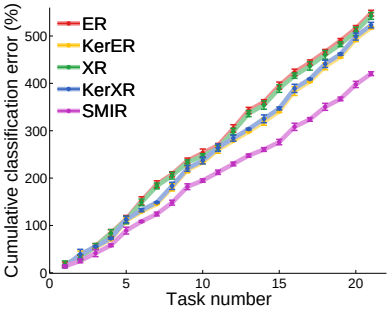
Figure 1. Experimental results of the multi-class classification tasks. Means with standard errors are shown by bar charts.



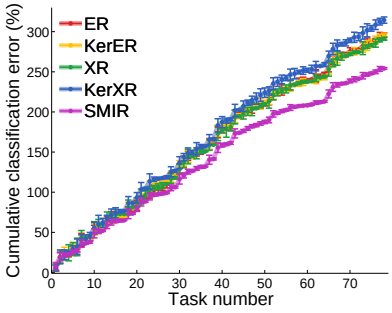
(a) USPS (45 tasks)



(b) MNIST (45 tasks)



(c) 20News (21 tasks)



(d) Isolet (78 tasks)

Figure 2. Experimental results of the simple classification tasks. The cumulative classification error at the k-th task is the sum of classification errors from the first to k-th tasks. Non-cumulative standard deviations are shown along the curves.

Table 3. Summary of all experimental results on USPS, MNIST, 20Newsgroups and Isolet. For each method, we measure how frequently it is the best or a comparable method based on the unpaired t-test at the significance level 5%, and the training time is averaged over all samplings of all tasks. The most accurate method and the most computationally-efficient method are highlighted in boldface.

	ER	KerER	XR	KerXR	SMIR
USPS, best or comparable (%)	45.65	15.22	21.74	17.39	**73.91**
MNIST, best or comparable (%)	**86.95**	0.00	19.57	2.17	80.43
20News, best or comparable (%)	36.36	18.18	36.36	18.18	**63.64**
Isolet, best or comparable (%)	60.76	62.03	68.35	48.10	**81.01**
USPS, training time (sec)	1.545	1.906	**1.149**	1.770	1.608
MNIST, training time (sec)	2.367	1.676	2.060	**1.536**	1.575
20News, training time (sec)	3.987	2.023	4.144	1.917	**1.654**
Isolet, training time (sec)	2.377	1.842	2.194	1.728	**1.723**

fairly good at multi-modal data, since all classes there had two clusters. Compared with KerER and KerXR, the plain ER and XR were better on USPS, MNIST and Isolet, but worse on 20Newsgroups. Nonetheless, ER/XR always outperformed KerER/KerXR in Table 3. Even though other algorithms often converged quite quickly on the simple tasks, SMIR was still a computationally-efficient algorithm after taking these simple tasks into account.

Secondly, we compare SMIR with two well-known geometric methods: Laplacian regularized least squares (LapRLS; Belkin et al., 2006) with a multi-class exten-

Table 4. Comparisons of LapRLS, LGC and SMIR, by means with standard errors of the classification error (in %) on the multi-class tasks. The best method and comparable ones based on the 5% unpaired t-test are highlighted in boldface.

	LapRLS	LGC	SMIR
USPS	39.64 ± 0.55	$\mathbf{36.53 \pm 0.53}$	$\mathbf{35.87 \pm 0.59}$
MNIST	42.34 ± 0.67	42.70 ± 0.60	$\mathbf{38.56 \pm 0.59}$
20News	64.85 ± 0.61	73.03 ± 0.24	$\mathbf{56.90 \pm 0.68}$
Isolet	39.98 ± 0.56	40.62 ± 0.47	$\mathbf{38.43 \pm 0.51}$

Table 5. Means with standard errors of the classification error (in %) on benchmarks from Chapelle et al. (2006). The best method and comparable ones based on the unpaired t-test at the significance level 5% are highlighted in boldface.

	ER	KerER	XR	KerXR	LapRLS	LGC	SMIR
g241c	30.14 ± 0.55	$\mathbf{24.86 \pm 0.66}$	31.66 ± 0.81	$\mathbf{24.42 \pm 0.69}$	34.12 ± 0.69	36.53 ± 0.74	31.69 ± 0.66
g241n	$\mathbf{33.07 \pm 0.58}$	35.65 ± 0.98	$\mathbf{33.90 \pm 0.83}$	36.67 ± 0.99	35.07 ± 0.65	38.15 ± 0.72	$\mathbf{33.76 \pm 0.65}$
Digit1	12.12 ± 0.41	$\mathbf{9.31 \pm 0.32}$	12.47 ± 0.49	$\mathbf{9.68 \pm 0.62}$	11.44 ± 0.43	11.87 ± 0.46	10.23 ± 0.40
USPS	26.60 ± 0.59	17.58 ± 0.33	27.07 ± 0.90	18.02 ± 0.72	12.45 ± 0.34	$\mathbf{10.27 \pm 0.33}$	12.23 ± 0.40
COIL	46.16 ± 0.78	38.58 ± 0.98	50.55 ± 1.20	39.81 ± 1.06	37.03 ± 0.81	$\mathbf{32.95 \pm 0.88}$	$\mathbf{33.62 \pm 0.82}$
COIL2	28.83 ± 0.72	25.81 ± 0.75	31.54 ± 1.02	27.73 ± 0.98	26.52 ± 0.65	$\mathbf{23.39 \pm 0.71}$	$\mathbf{24.12 \pm 0.69}$
BCI	$\mathbf{40.58 \pm 0.67}$	47.76 ± 0.45	43.21 ± 0.70	48.35 ± 0.46	43.46 ± 0.63	48.70 ± 0.44	47.27 ± 0.57
Text	$\mathbf{34.92 \pm 0.56}$	44.36 ± 0.58	$\mathbf{35.38 \pm 0.54}$	43.79 ± 0.65	44.50 ± 0.54	49.53 ± 0.18	38.80 ± 0.64

Table 6. Means with standard errors of the classification error (in %) on seven UCI benchmarks and Senseval-2. The best method and comparable ones based on the unpaired t-test at the significance level 5% are highlighted in boldface.

	ER	KerER	XR	KerXR	LapRLS	LGC	SMIR
Diabetes	$\mathbf{27.26 \pm 0.41}$	29.70 ± 0.50	28.41 ± 0.53	30.16 ± 0.72	32.01 ± 0.62	32.32 ± 0.42	29.87 ± 0.57
Wine	8.09 ± 0.44	$\mathbf{4.21 \pm 0.44}$	10.56 ± 1.21	6.56 ± 0.95	8.21 ± 0.45	7.71 ± 0.44	6.91 ± 0.54
Vowel	70.65 ± 0.78	63.03 ± 0.78	69.70 ± 0.77	$\mathbf{61.32 \pm 0.68}$	63.90 ± 0.65	64.13 ± 0.66	$\mathbf{62.77 \pm 0.65}$
Image	27.32 ± 0.64	22.38 ± 0.67	26.91 ± 0.75	23.07 ± 0.90	$\mathbf{18.80 \pm 0.66}$	$\mathbf{19.45 \pm 0.65}$	$\mathbf{19.82 \pm 0.67}$
Vehicle	39.43 ± 0.90	45.61 ± 0.78	48.44 ± 1.10	46.86 ± 0.91	$\mathbf{38.22 \pm 0.79}$	43.01 ± 0.54	$\mathbf{37.48 \pm 0.74}$
German	32.30 ± 0.55	$\mathbf{29.31 \pm 0.31}$	32.76 ± 0.65	$\mathbf{29.45 \pm 0.35}$	30.96 ± 0.42	30.94 ± 0.33	30.62 ± 0.43
Satimage	31.01 ± 0.73	22.59 ± 0.58	34.79 ± 0.68	25.12 ± 1.43	20.15 ± 0.40	$\mathbf{18.75 \pm 0.34}$	$\mathbf{18.96 \pm 0.39}$
Senseval-2	$\mathbf{32.72 \pm 0.62}$	35.56 ± 0.73	37.14 ± 1.10	36.37 ± 0.83	34.66 ± 0.71	37.77 ± 0.67	$\mathbf{33.11 \pm 0.74}$

sion, as well as learning with local and global consistency (LGC; Zhou et al., 2003) with an out-of-sample extension. They represent the state-of-the-art manifold regularization and similarity graph transduction respectively. Similarly to SMIR, their optimizations are convex and can be solved analytically. LapRLS is extended using the one-vs-rest trick, and LGC is extended via the Nadaraya-Watson estimator (Delalleau et al., 2005). The experimental setup and the candidates of hyperparameters for LapRLS and LGC are same as SMIR, except that the regularization parameter α of LGC is chosen from $\{0.2, 0.4, 0.6, 0.8, 0.99\}$. SMIR was always best or tie in Table 4, and thus it is fairly competitive with those pure geometric methods on these benchmarks.

Finally, we take all seven methods and compare their performance on the sixteen benchmarks listed in Table 2. The experimental results are reported in Tables 5 and 6, where the experimental setup and the candidates of hyperparameters are same as previous experiments. To be clear, there are two benchmarks, BCI and Wine, whose sample size is less than 500. As a result, each of their random samplings included the whole set, and the randomness or the difference of the classification error was actually from how the training, test and cross-validation data were split and also how labeled data were selected. We can see from Table 5 that ER, LGC and SMIR were best or comparable on three benchmarks, and KerER, XR and KerXR were best or comparable on two benchmarks. Moreover, in Table 6, SMIR won or tied five times, while all other methods except XR won or tied twice. Therefore, it is reasonable and practical to maximize SMI following the information maximization principe, and SMIR is a promising information-theoretic approach to semi-supervised learning.

6. Conclusions

In this paper, we proposed squared-loss mutual information regularization (SMIR). Compared with other information-theoretic regularization, SMIR is convex with no logarithm in the involved optimization problem, and thus enables the analytic expression of the globally optimal solution. We established novel data-dependent generalization error bounds that even incorporate the information of unlabeled data. We then evaluated SMIR on twenty benchmark data sets, and the results demonstrated that SMIR compared favorably with entropy regularization, expectation regularization, manifold regularization, and similarity graph transduction.

Acknowledgments

GN was supported by the MEXT scholarship 103250, WJ was supported by the Okazaki Kaheita International Scholarship Foundation, HH was supported by the FIRST program, and MS was supported by the MEXT KAKENHI 25700022.

References

Agakov, F. and Barber, D. Kernelized infomax clustering. In *NIPS*, 2006.

Ali, S. M. and Silvey, S. D. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28(1):131–142, 1966.

Allwein, E., Schapire, R., and Singer, Y. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

Bartlett, P. and Mendelson, S. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Belkin, M., Matveeva, I., and Niyogi, P. Regularization and semi-supervised learning on large graphs. In *ALT*, 2004.

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Bennett, K. and Demiriz, A. Semi-supervised support vector machines. In *NIPS*, 1998.

Chapelle, O., Schölkopf, B., and Zien, A. (eds.). *Semi-Supervised Learning*. MIT Press, 2006.

Cortes, C., Mohri, M., Pechyony, D., and Rastogi, A. Stability of transductive regression algorithms. In *ICML*, 2008.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

Delalleau, O., Bengio, Y., and Le Roux, N. Efficient non-parametric function induction in semi-supervised learning. In *AISTATS*, 2005.

El-Yaniv, R. and Pechyony, D. Transductive Rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

Gomes, R., Krause, A., and Perona, P. Discriminative clustering by regularized information maximization. In *NIPS*, 2010.

Grandvalet, Y. and Bengio, Y. Semi-supervised learning by entropy minimization. In *NIPS*, 2004.

Hall, M. Correlation-based feature selection for discrete and numeric class machine learning. In *ICML*, 2000.

Joachims, T. Transductive inference for text classification using support vector machines. In *ICML*, 1999.

Joachims, T. Transductive learning via spectral graph partitioning. In *ICML*, 2003.

Kanamori, T., Suzuki, T., and Sugiyama, M. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.

Koltchinskii, V. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

Kullback, S. and Leibler, R. A. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

Mann, G. and McCallum, A. Simple, robust, scalable semi-supervised learning via expectation regularization. In *ICML*, 2007.

McDiarmid, C. On the method of bounded differences. In Siemons, J. (ed.), *Surveys in Combinatorics*, pp. 148–188. Cambridge University Press, 1989.

Meir, R. and Zhang, T. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

Schölkopf, B. and Smola, A. *Learning with Kernels*. MIT Press, 2001.

Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 & 623–656, 1948.

Sinha, K. and Belkin, M. Semi-supervised learning using sparse eigenfunction bases. In *NIPS*, 2009.

Sugiyama, M. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, E93-D(10): 2690–2701, 2010.

Sugiyama, M., Yamada, M., Kimura, M., and Hachiya, H. On information-maximization clustering: Tuning parameter selection and analytic solution. In *ICML*, 2011.

Suzuki, T., Sugiyama, M., Kanamori, T., and Sese, J. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52, 2009.

Szummer, M. and Jaakkola, T. Partially labeled classification with Markov random walks. In *NIPS*, 2001.

Szummer, M. and Jaakkola, T. Information regularization with partially labeled data. In *NIPS*, 2002.

Yamada, M., Sugiyama, M., Wichern, G., and Simm, J. Improving the accuracy of least-squares probabilistic classifiers. *IEICE Transactions on Information and Systems*, E94-D(6):1337–1340, 2011.

Zhou, D., Bousquet, O., Navin Lal, T., Weston, J., and Schölkopf, B. Learning with local and global consistency. In *NIPS*, 2003.

Zhu, X., Ghahramani, Z., and Lafferty, J. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, 2003.

Appendix: Supplementary Material

A. Proof of Theorem 1

Proof. Denote the *unnormalized graph Laplacian* by $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$ and the *normalized graph Laplacian* by

$$\boldsymbol{L}^* = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2} = \boldsymbol{I}_n - \boldsymbol{D}^{-1/2}\boldsymbol{K}\boldsymbol{D}^{-1/2},$$

where $\boldsymbol{I}_n$ is the identity matrix of size n. Optimization (5) can be rewritten as

$$\min_{\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_c \in \mathbb{R}^n} \Delta(p,q) + \gamma' \sum\nolimits_{y\in\mathcal{Y}} \boldsymbol{\alpha}_y^\top \boldsymbol{L}^* \boldsymbol{\alpha}_y + \lambda' \sum\nolimits_{y\in\mathcal{Y}} \frac{1}{2}\|\boldsymbol{\alpha}_y\|_2^2, \tag{12}$$

where $\gamma' = \gamma c/2n > 0$ and $\lambda' = \lambda - \gamma c/n > 0$ are regularization parameters. Notice that $\forall y \in \mathcal{Y}$,

$$\boldsymbol{\alpha}_y^\top \boldsymbol{L}^* \boldsymbol{\alpha}_y = \frac{1}{2}\sum_{i,j=1}^{n} \left(\frac{\alpha_{y,i}}{d_i} - \frac{\alpha_{y,j}}{d_j}\right)^2 \boldsymbol{K}_{i,j},$$

and then the second term of (12) is convex since $\boldsymbol{K}_{i,j} \geq 0$.

The loss function $\Delta(p,q)$ is convex w.r.t. $q(y \mid \boldsymbol{x};\boldsymbol{\alpha})$, and $q(y \mid \boldsymbol{x};\boldsymbol{\alpha})$ is linear w.r.t. $\boldsymbol{\alpha}_y$, so $\Delta(p,q)$ is convex w.r.t. $\boldsymbol{\alpha}_y$. The ℓ_2-norm of $\boldsymbol{\alpha}_y$ is strictly convex w.r.t. $\boldsymbol{\alpha}_y$, i.e., it takes zero if and only if $\boldsymbol{\alpha}_y$ is identically zero. Therefore, optimization (12) is strictly convex and there exists a unique globally optimal solution. □

B. Derivation of the Error Bounds

B.1. Definitions

To begin with, we state the inductive definition of Rademacher complexity following El-Yaniv & Pechyony (2009).

Definition 1. *Suppose that $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ are independent observations according to $p(\boldsymbol{x})$. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}$, and $\sigma_1,\dots,\sigma_n$ be independent uniformly $\{\pm 1\}$-valued random variables, i.e., Rademacher variables. Subsequently, the empirical Rademacher complexity conditioned on $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ is defined as*

$$\widehat{\mathcal{R}}_n(\mathcal{F}) := \mathbb{E}_{\sigma_1,\dots,\sigma_n}\left\{\sup_{f\in\mathcal{F}} \frac{2}{n}\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)\right\},$$

and the inductive Rademacher complexity is defined as

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n}\left\{\widehat{\mathcal{R}}_n(\mathcal{F})\right\}.$$

There exist various definitions of $\widehat{\mathcal{R}}_n(\mathcal{F})$: The definition in Bartlett & Mendelson (2002) is

$$\widehat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_{\sigma_1,\dots,\sigma_n}\left\{\sup_{f\in\mathcal{F}} \frac{2}{n}\left|\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)\right|\right\},$$

the definition in Koltchinskii (2001) uses

$$\widehat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_{\sigma_1,\dots,\sigma_n}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)\right|\right\},$$

while the definition in Meir & Zhang (2003) adopt

$$\widehat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_{\sigma_1,\dots,\sigma_n}\left\{\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)\right\}.$$

The definition in El-Yaniv & Pechyony (2009) is consistent with Bartlett & Mendelson (2002) for function classes that are closed under negation, and is always equal to or less than the one in Bartlett & Mendelson (2002).

Nevertheless, a vital disagreement arises when considering comparison theorems and thus the famous *contraction principle* of Rademacher averages. If $\psi : \mathbb{R} \mapsto \mathbb{R}$ is Lipschitz continuous with a Lipschitz constant L_ψ and satisfies $\psi(0) = 0$, then

$$\widehat{\mathcal{R}}_n(\psi \circ \mathcal{F}) \leq L_\psi \widehat{\mathcal{R}}_n(\mathcal{F})$$

for El-Yaniv & Pechyony (2009) and

$$\widehat{\mathcal{R}}_n(\psi \circ \mathcal{F}) \leq 2L_\psi \widehat{\mathcal{R}}_n(\mathcal{F})$$

for Bartlett & Mendelson (2002). When all involved error bounds are single-sided concentration results, those definitions without the absolute value in the argument of the supremum (El-Yaniv & Pechyony, 2009; Meir & Zhang, 2003) are more natural and powerful.

B.2. Proof of Theorem 2

Let $\boldsymbol{\beta}_{\mathcal{F}} = \boldsymbol{K}^{-1/2}\boldsymbol{D}^{-1/2}\boldsymbol{\alpha}^*_{\mathcal{F}}$, then

$$B^2_{\mathcal{F}} = \|\boldsymbol{D}^{-1/2}\boldsymbol{\alpha}^*_{\mathcal{F}}\|_2^2 = \boldsymbol{\beta}_{\mathcal{F}}^\top \boldsymbol{K}\boldsymbol{\beta}_{\mathcal{F}},$$
$$B'_{\mathcal{F}} = \|\boldsymbol{K}^{-1/2}\boldsymbol{D}^{-1/2}\boldsymbol{\alpha}^*_{\mathcal{F}}\|_1 = \|\boldsymbol{\beta}_{\mathcal{F}}\|_1.$$

Define the class of functions $\mathcal{F}$ as

$$\mathcal{F} := \left\{ \boldsymbol{x} \mapsto \sum_{i=1}^n \beta_i k(\boldsymbol{x}, \boldsymbol{x}'_i) \;\middle|\; \boldsymbol{x}'_i \in \mathcal{X}, \beta_i \in \mathbb{R}, \sum_{i=1}^n |\beta_i| \leq B'_{\mathcal{F}}, \sum_{i,j=1}^n \beta_i\beta_j k(\boldsymbol{x}'_i, \boldsymbol{x}'_j) \leq B^2_{\mathcal{F}} \right\}.$$

It is easy to verify that $f(\boldsymbol{x}) = \langle \Phi_n(\boldsymbol{x}), \boldsymbol{\beta}_{\mathcal{F}} \rangle \in \mathcal{F}$, where $f(\boldsymbol{x})$ is the decision function defined in Eq. (9). By Lemma 22 of Bartlett & Mendelson (2002), we get

$$\widehat{\mathcal{R}}_n(\mathcal{F}) \leq \frac{2B_{\mathcal{F}}}{n}\left(\sum_{i=1}^n k(\boldsymbol{x}_i, \boldsymbol{x}_i)\right)^{1/2} \leq \frac{2B_k B_{\mathcal{F}}}{\sqrt{n}}. \tag{13}$$

Applying Lemma 22 of Bartlett & Mendelson (2002) again gives us

$$\widehat{\mathcal{R}}_l(\mathcal{F}) \leq \frac{2B_{\mathcal{F}}}{l}\left(\sum_{i=1}^l k(\boldsymbol{x}_i, \boldsymbol{x}_i)\right)^{1/2} \leq \frac{2B_k B_{\mathcal{F}}}{\sqrt{l}}. \tag{14}$$

where $\widehat{\mathcal{R}}_l(\mathcal{F})$ is the empirical Rademacher complexities of $\mathcal{F}$ conditioned only on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l$.

In the following, we only focus on the proof of inequality (11) based on inequality (13). Inequality (10) can be derived by the exactly same way based on inequality (14). Let

$$\ell_\eta \circ \mathcal{F} := \{(\boldsymbol{x}, y) \mapsto \ell_\eta(yf(\boldsymbol{x})) \mid f \in \mathcal{F}\},$$

which is a class of functions mapping from $\mathcal{X} \times \mathcal{Y}$ to the interval $[0, 1]$. The rest of the proof consists of two steps. The first step bounds $\mathcal{R}_n(\ell_\eta \circ \mathcal{F})$ from above, and the second step bounds $\mathbb{E}\ell(yf(\boldsymbol{x}))$ using $\mathcal{R}_n(\ell_\eta \circ \mathcal{F})$.

B.2.1. Step 1

The following lemma relates the inductive Rademacher complexity of a class of bounded functions to the corresponding empirical Rademacher complexity.

Lemma 3 (Concentration Lemma)**.** *Let $\mathcal{F}_C$ be a class of functions mapping to the interval $[-C, C]$. With probability at least $1 - \delta/2$, we have*

$$\mathcal{R}_n(\mathcal{F}_C) \leq \widehat{\mathcal{R}}_n(\mathcal{F}_C) + 4C\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Similarly, let $\mathcal{F}^+_C$ be a class of functions mapping to the interval $[0, C]$. With probability at least $1 - \delta/2$, we have

$$\mathcal{R}_n(\mathcal{F}^+_C) \leq \widehat{\mathcal{R}}_n(\mathcal{F}^+_C) + 2C\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Proof. Recall that $\widehat{\mathcal{R}}_n(\mathcal{F}_C)$ conditioned on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is a random variable defined as

$$\widehat{\mathcal{R}}_n(\mathcal{F}_C) = \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left\{\sup_{f\in\mathcal{F}_C} \frac{2}{n}\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)\right\}.$$

When an observation $\boldsymbol{x}_i$ changes to $\boldsymbol{x}_i'$, the change of $\widehat{\mathcal{R}}_n(\mathcal{F}_C)$ is no more than $4C/n$, and thus *McDiarmid's inequality* (McDiarmid, 1989) implies that

$$\Pr\left\{\mathcal{R}_n(\mathcal{F}_C) - \widehat{\mathcal{R}}_n(\mathcal{F}_C) \geq \epsilon\right\} \leq \exp\left(-\frac{\epsilon^2 n}{8C^2}\right).$$

The first bound can be obtained by equating the right-hand side of the above inequality to $\delta/2$.

For $\mathcal{F}_C^+$, when an observation $\boldsymbol{x}_i$ changes to $\boldsymbol{x}_i'$, the change of $\widehat{\mathcal{R}}_n(\mathcal{F}_C^+)$ is no more than $2C/n$. The lemma follows by the same argument as above. □

The next lemma is a variation of the comparison lemma in Meir & Zhang (2003), where the comparison is done for two sets of functions under a Bayesian framework, and its validity follows Lemma 5 of El-Yaniv & Pechyony (2009) by setting $p = 1/2$.

Lemma 4 (Comparison Lemma)**.** *Let*

$$\mathcal{H} := \{\boldsymbol{h} = (h_1, \ldots, h_n)^\top \mid h_i = y_i f(\boldsymbol{x}_i), f \in \mathcal{F}\},$$

and $\psi, \psi' : \mathbb{R} \mapsto \mathbb{R}$ be real-valued functions. If for all $\boldsymbol{h}, \boldsymbol{h}' \in \mathcal{H}$ and $i = 1, \ldots, n$,

$$|\psi(h_i) - \psi(h_i')| \leq |\psi'(h_i) - \psi'(h_i')|,$$

then

$$\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left\{\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i\psi(h_i)\right\} \leq \mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left\{\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\sigma_i\psi'(h_i)\right\}.$$

Now $\widehat{\mathcal{R}}_n(\ell_\eta \circ \mathcal{F})$ and $\mathcal{R}_n(\ell_\eta \circ \mathcal{F})$ can be bounded from above by $\widehat{\mathcal{R}}_n(\mathcal{F})$ and $\mathcal{R}_n(\mathcal{F})$ based on the comparison lemma.

Lemma 5 (Contraction Lemma)**.** *For any $\eta > 0$, we have*

$$\widehat{\mathcal{R}}_n(\ell_\eta \circ \mathcal{F}) \leq \frac{1}{\eta}\widehat{\mathcal{R}}_n(\mathcal{F}),$$
$$\mathcal{R}_n(\ell_\eta \circ \mathcal{F}) \leq \frac{1}{\eta}\mathcal{R}_n(\mathcal{F}).$$

Proof. Note that $\ell_\eta(z)$ satisfies the Lipschitz condition

$$|\ell_\eta(z) - \ell_\eta(z')| \leq \frac{1}{\eta}|z - z'|, \quad \forall z, z' \in \mathbb{R}.$$

Let $\psi(h_i) = \ell_\eta(y_i f(\boldsymbol{x}_i))$ and $\psi'(h_i) = y_i f(\boldsymbol{x}_i)/\eta$, then

$$\begin{aligned}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left\{\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i\ell_\eta(y_i f(\boldsymbol{x}_i))\right\} &\leq \frac{1}{\eta}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left\{\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i y_i f(\boldsymbol{x}_i)\right\} \\ &= \frac{1}{\eta}\mathbb{E}_{\sigma_1,\ldots,\sigma_n}\left\{\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)\right\},\end{aligned}$$

where the first step is a corollary of the comparison lemma, and the second step is due to the same distribution of each $\sigma_i y_i$ and σ_i. This completes the proof. □

As a result, if we contract $\widehat{\mathcal{R}}_n(\mathcal{F})$ and then concentrate $\widehat{\mathcal{R}}_n(\ell_\eta \circ \mathcal{F})$, we could know

$$\begin{aligned}\mathcal{R}_n(\ell_\eta \circ \mathcal{F}) &\leq \widehat{\mathcal{R}}_n(\ell_\eta \circ \mathcal{F}) + 2\sqrt{\frac{\ln(2/\delta)}{2n}} \\ &\leq \frac{2B_k B_{\mathcal{F}}}{\eta\sqrt{n}} + 2\sqrt{\frac{\ln(2/\delta)}{2n}},\end{aligned} \tag{15}$$

since ℓ_η maps to the interval $[0, 1]$. On the other hand, for any $f \in \mathcal{F}$,

$$\|f\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} \left| \sum_{i=1}^{n} \beta_i k(\boldsymbol{x}, \boldsymbol{x}'_i) \right| \leq B_k^2 B'_{\mathcal{F}},$$

which says that $\mathcal{F}$ is a class of functions mapping to the interval $[-B_k^2 B'_{\mathcal{F}}, B_k^2 B'_{\mathcal{F}}]$. Thus, if we concentrate $\widehat{\mathcal{R}}_n(\mathcal{F})$ before contract $\mathcal{R}_n(\mathcal{F})$, we can obtain

$$\begin{aligned}\mathcal{R}_n(\ell_\eta \circ \mathcal{F}) &\leq \frac{1}{\eta}\mathcal{R}_n(\mathcal{F}) \\ &\leq \frac{1}{\eta}\left(\frac{2B_k B_{\mathcal{F}}}{\sqrt{n}} + 4B_k^2 B'_{\mathcal{F}}\sqrt{\frac{\ln(2/\delta)}{2n}}\right).\end{aligned} \tag{16}$$

Combining inequalities (15) and (16) finalizes the first step of the proof, that is,

$$\mathcal{R}_n(\ell_\eta \circ \mathcal{F}) \leq \frac{2B_k B_{\mathcal{F}}}{\eta\sqrt{n}} + \min\left(2, \frac{4B_k^2 B'_{\mathcal{F}}}{\eta}\right)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

B.2.2. Step 2

This step is composed of a single concentration inequality, that is, with probability at least $1 - \delta/2$,

$$\mathbb{E}\ell(yf(\boldsymbol{x})) \leq \hat{\mathbb{E}}_n \ell_\eta(yf(\boldsymbol{x})) + \mathcal{R}_n(\ell_\eta \circ \mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{17}$$

Since $\forall z \in \mathbb{R}$, $\ell(z)$ is always equal to or less than $\ell_\eta(z)$, for any $f \in \mathcal{F}$ we can write

$$\begin{aligned}\mathbb{E}\ell(yf(\boldsymbol{x})) &\leq \mathbb{E}\ell_\eta(yf(\boldsymbol{x})) \\ &\leq \hat{\mathbb{E}}_n \ell_\eta(yf(\boldsymbol{x})) + \sup_{\psi\in\ell_\eta\circ\mathcal{F}} (\mathbb{E}\psi - \hat{\mathbb{E}}_n\psi).\end{aligned}$$

Any function $\psi(\boldsymbol{x}, y) = \ell_\eta(yf(x)) \in \ell_\eta \circ \mathcal{F}$ satisfies $0 \leq \psi(\boldsymbol{x}, y) \leq 1$, so when $(\boldsymbol{x}_i, y_i)$ changes to $(\boldsymbol{x}'_i, y'_i)$, the change of $\sup_{\psi\in\ell_\eta\circ\mathcal{F}}(\mathbb{E}\psi - \hat{\mathbb{E}}_n\psi)$ cannot be more than $1/n$. Hence, McDiarmid's inequality implies that

$$\Pr\left\{\sup_{\psi\in\ell_\eta\circ\mathcal{F}} (\mathbb{E}\psi - \hat{\mathbb{E}}_n\psi) - \mathbb{E}_{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} (\mathbb{E}\psi - \hat{\mathbb{E}}_n\psi) \geq \epsilon\right\} \leq \exp(-2\epsilon^2 n),$$

or equivalently, with probability at least $1 - \delta/2$,

$$\sup_{\psi\in\ell_\eta\circ\mathcal{F}} (\mathbb{E}\psi - \hat{\mathbb{E}}_n\psi) \leq \mathbb{E}_{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} (\mathbb{E}\psi - \hat{\mathbb{E}}_n\psi) + \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

It remains to bound the expectation $\mathbb{E}_{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)} \sup_{\psi\in\ell_\eta\circ\mathcal{F}}(\mathbb{E}\psi - \hat{\mathbb{E}}_n\psi)$ by the complexity $\mathcal{R}_n(\ell_\eta \circ \mathcal{F})$. Suppose that

$$\{(\boldsymbol{x}'_1, y'_1), \ldots, (\boldsymbol{x}'_n, y'_n) \mid (\boldsymbol{x}'_i, y'_i) \sim p(\boldsymbol{x}, y)\}$$

is a ghost sample for symmetrization, then

$$
\begin{aligned}
\mathbb{E}_{(\boldsymbol{x}_i,y_i)} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} (\mathbb{E}\psi - \hat{\mathbb{E}}_n\psi) &= \mathbb{E}_{(\boldsymbol{x}_i,y_i)} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} \left(\mathbb{E}_{(\boldsymbol{x}'_i,y'_i)}[\hat{\mathbb{E}}_n\psi(\boldsymbol{x}'_i,y'_i)] - \hat{\mathbb{E}}_n\psi(\boldsymbol{x}_i,y_i)\right) \\
&= \mathbb{E}_{(\boldsymbol{x}_i,y_i)} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} \left(\mathbb{E}_{(\boldsymbol{x}'_i,y'_i)}[\hat{\mathbb{E}}_n\psi(\boldsymbol{x}'_i,y'_i) - \hat{\mathbb{E}}_n\psi(\boldsymbol{x}_i,y_i)]\right) \\
&\le \mathbb{E}_{(\boldsymbol{x}_i,y_i),(\boldsymbol{x}'_i,y'_i)} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} \left(\hat{\mathbb{E}}_n\psi(\boldsymbol{x}'_i,y'_i) - \hat{\mathbb{E}}_n\psi(\boldsymbol{x}_i,y_i)\right) \qquad (18) \\
&= \mathbb{E}_{(\boldsymbol{x}_i,y_i),(\boldsymbol{x}'_i,y'_i)} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} (\psi(\boldsymbol{x}'_i,y'_i) - \psi(\boldsymbol{x}_i,y_i)) \\
&= \mathbb{E}_{\sigma_i,(\boldsymbol{x}_i,y_i),(\boldsymbol{x}'_i,y'_i)} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i(\psi(\boldsymbol{x}'_i,y'_i) - \psi(\boldsymbol{x}_i,y_i)) \qquad (19) \\
&\le \mathbb{E}_{(\boldsymbol{x}'_i,y'_i),\sigma_i} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\psi(\boldsymbol{x}'_i,y'_i) + \mathbb{E}_{(\boldsymbol{x}_i,y_i),\sigma_i} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} (-\sigma_i)\psi(\boldsymbol{x}_i,y_i) \\
&= 2\mathbb{E}_{(\boldsymbol{x}_i,y_i),\sigma_i} \sup_{\psi\in\ell_\eta\circ\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\psi(\boldsymbol{x}_i,y_i) \qquad (20) \\
&= \mathcal{R}_n(\ell_\eta\circ\mathcal{F}),
\end{aligned}
$$

where (18) uses the fact that the supremum is a convex function and then we apply *Jensen's inequality*, (19) is due to the symmetry of the ghost sample and the original sample and thus the same distribution of $\psi(\boldsymbol{x}'_i,y'_i)-\psi(\boldsymbol{x}_i,y_i)$ and $\sigma_i(\psi(\boldsymbol{x}'_i,y'_i)-\psi(\boldsymbol{x}_i,y_i))$, and (20) is valid since σ_i and $-\sigma_i$ have the same distribution while the original and ghost samples also have the same distribution. □