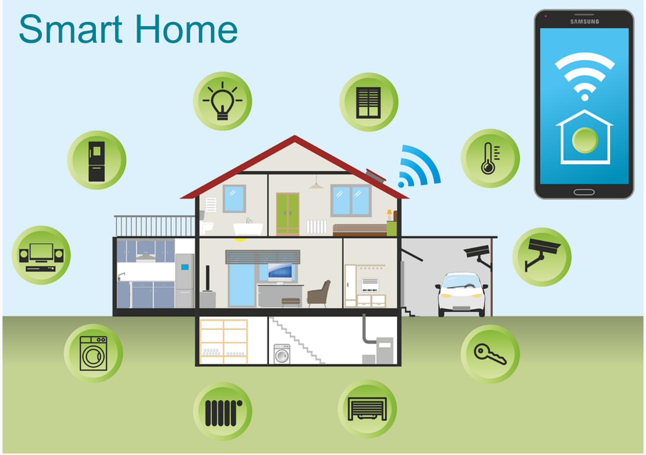
FUNDAMENTAL PERFORMANCE LIMITS OF STATISTICAL PROBLEMS: FROM DETECTION THEORY TO SEMI-SUPERVISED LEARNING

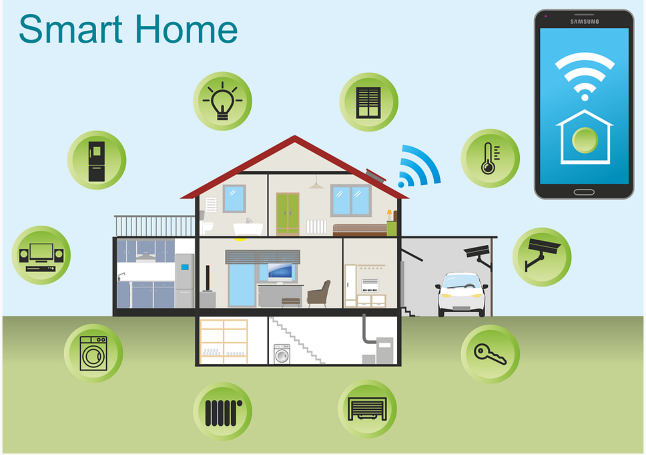
Ph.D. Thesis Defense

Candidate: Haiyun He

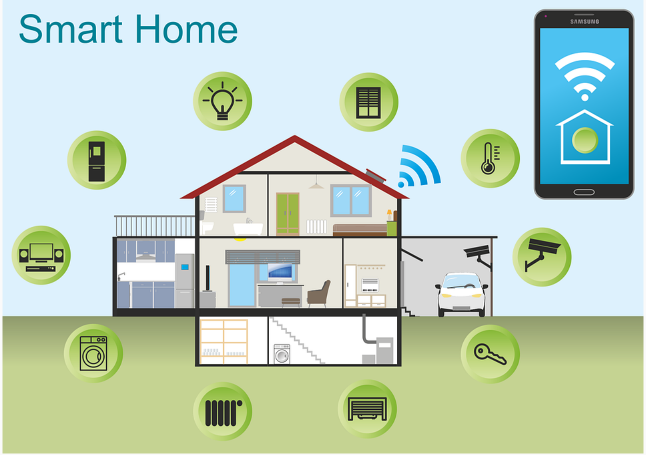
Department of Electrical and Computer Engineering

Motivation



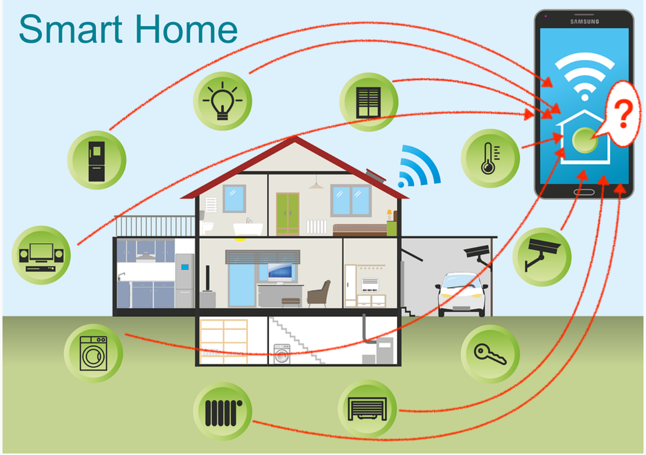


One core problem: to design good mechanisms to infer or learn useful information from the raw data.



One core problem: to design good mechanisms to infer or learn useful information from the raw data.

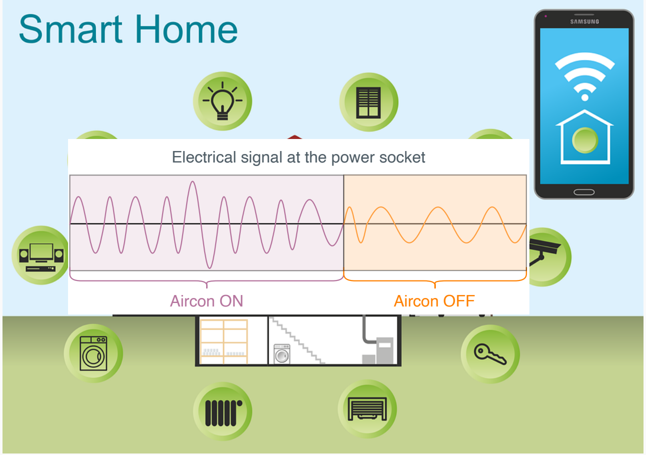
Statistical viewpoint:



One core problem: to design good mechanisms to infer or learn useful information from the raw data.

Statistical viewpoint:

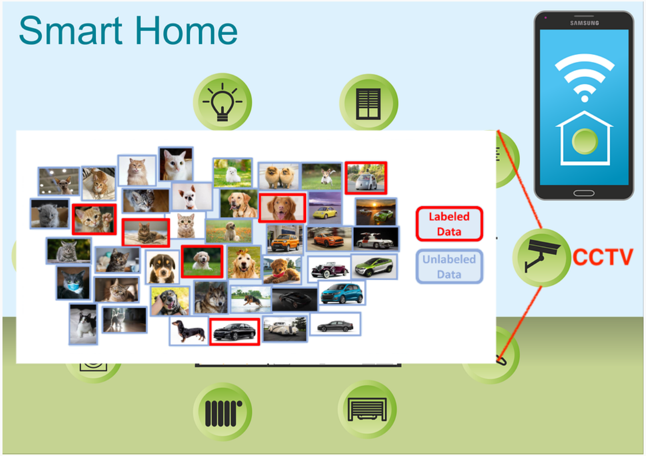
- Distributed detection



One core problem: to design good mechanisms to infer or learn useful information from the raw data.

Statistical viewpoint:

- Distributed detection
- Change-point detection



One core problem: to design good mechanisms to infer or learn useful information from the raw data.

Statistical viewpoint:

- Distributed detection
- Change-point detection
- Semi-supervised learning

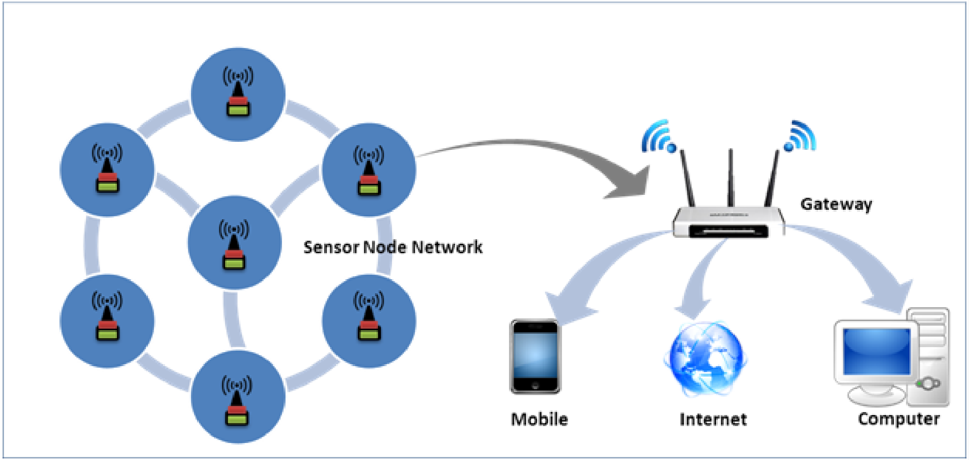
Outline

1. Distributed Detection with Empirically Observed Statistics
2. Change-Point Detection with Training Sequences
3. Information-Theoretic Generalization Error for Iterative Semi-Supervised Learning

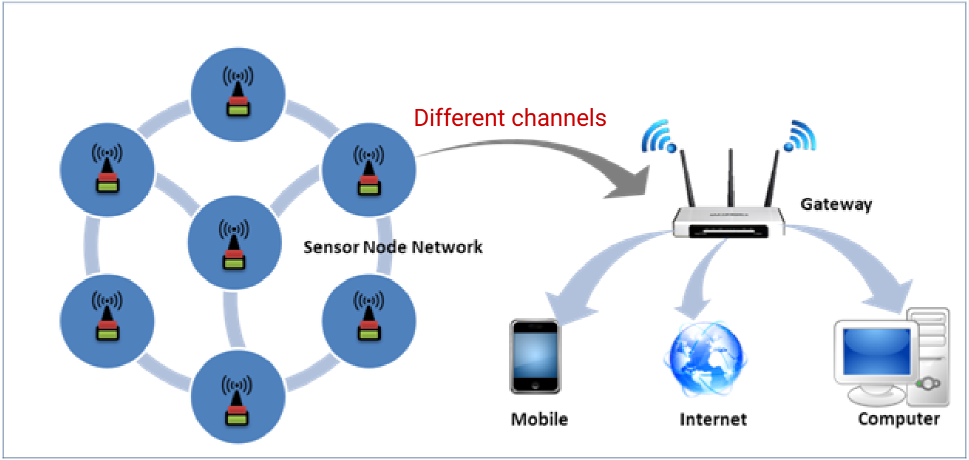
Outline

1. **Distributed Detection with Empirically Observed Statistics**
2. Change-Point Detection with Training Sequences
3. Information-Theoretic Generalization Error for Iterative Semi-Supervised Learning

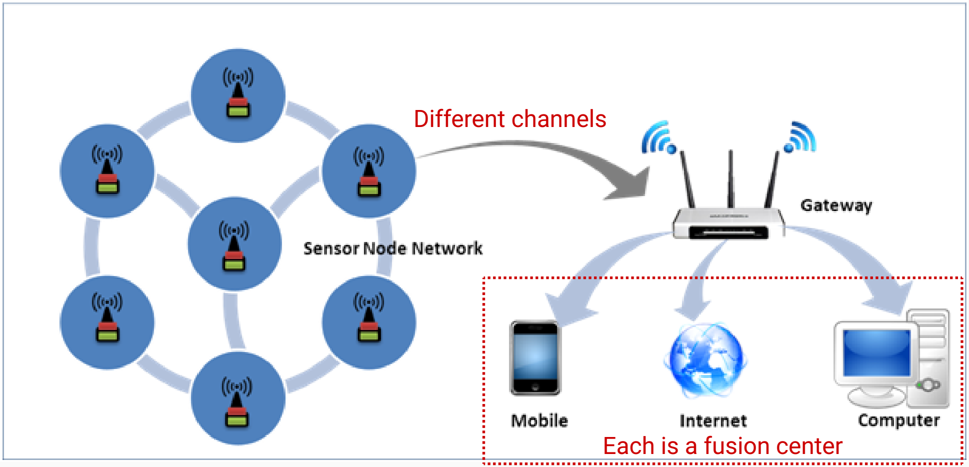
Background: Distributed Detection



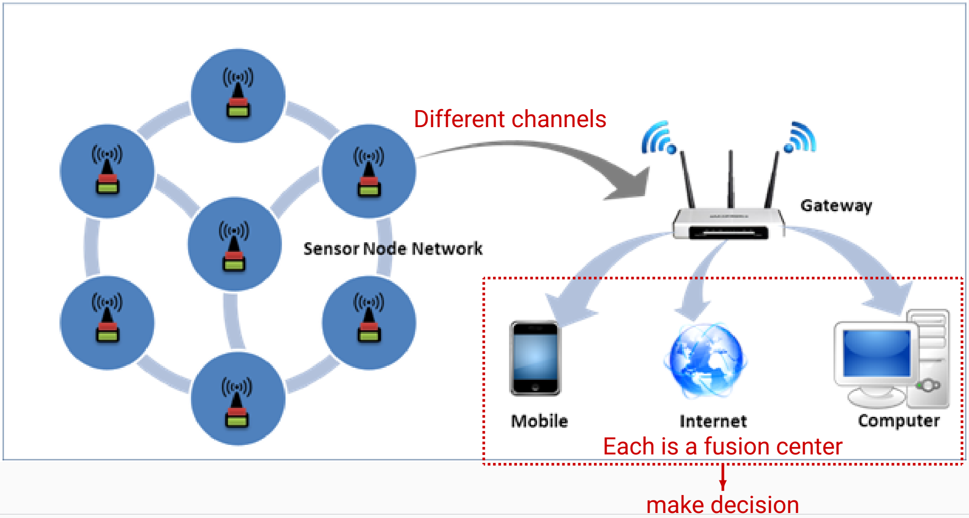
Background: Distributed Detection



Background: Distributed Detection



Background: Distributed Detection



Distributed Detection: Related Works

- Inspired by Tsitsiklis who considered distributed detection with known distributions

Math. Control Signals Systems (1988) 1: 167–182

Mathematics of Control, Signals, and Systems

Decentralized Detection by a Large Number of Sensors*

John N. Tsitsiklis†

Abstract. We consider the decentralized detection problem, in which N independent, identical sensors transmit a finite-valued function of their observations to a fusion center which then decides which one of M hypotheses is true. For the case where the number of sensors tends to infinity, we show that it is asymptotically optimal to divide the sensors into $M(M-1)/2$ groups, with all sensors in each group using the same decision rule in deciding what to transmit. We also show how the optimal number of sensors in each group may be determined by solving a mathematical programming problem. For the special case of two hypotheses and

NUS National University of Singapore

- Inspired by Tsitsiklis who considered distributed detection with known distributions
  - A key result: Using a single type of compressor for binary hypothesis testing is optimal

Math. Control Signals Systems (1988) 1: 167–182

Mathematics of Control, Signals, and Systems

Decentralized Detection by a Large Number of Sensors*

John N. Tsitsiklis†

Abstract. We consider the decentralized detection problem, in which N independent, identical sensors transmit a finite-valued function of their observations to a fusion center which then decides which one of M hypotheses is true. For the case where the number of sensors tends to infinity, we show that it is asymptotically optimal to divide the sensors into $M(M-1)/2$ groups, with all sensors in each group using the same decision rule in deciding what to transmit. We also show how the optimal number of sensors in each group may be determined by solving a mathematical programming problem. For the special case of two hypotheses and

Let $R_N = \inf_{\gamma^N \in \Gamma^N} r_N(\gamma^N)$ be the optimal exponent. Let Γ_0^N be the set of all $\gamma^N \in \Gamma^N$ with the property that the set $\{\gamma_1, \ldots, \gamma_N\}$ has at most $M(M-1)/2$ different elements. Let $Q_N = \inf_{\gamma^N \in \Gamma_0^N} r_N(\gamma^N)$ be the optimal exponent, when we restrict to sets of decision rules in Γ_0^N. The following result shows that, asymptotically, optimality is not lost, if we restrict to Γ_0^N.

Theorem 1. *Subject to Assumption* 1 *below*, $\lim_{N\to\infty}(Q_N - R_N) = 0$.

Distributed Detection: Related Works

- Inspired by Tsitsiklis who considered distributed detection with known distributions
  - A key result: Using a single type of compressor for binary hypothesis testing is optimal
- Also inspired by Gutman who adopted an information-theoretic approach to statistical classification

Math. Control Signals Systems (1988) 1: 167–182

Mathematics of Control, Signals, and Systems

Decentralized Detection by a Large Number of Sensors*

John N. Tsitsiklis†

Abstract. We consider the decentralized detection problem, in which N independent, identical sensors transmit a finite-valued function of their observations to a fusion center which then decides which one of M hypotheses is true. For the case where the number of sensors tends to infinity, we show that it is asymptotically optimal to divide the sensors into $M(M-1)/2$ groups, with all sensors in each group using the same decision rule in deciding what to transmit. We also show how the optimal number of sensors in each group may be determined by solving a mathematical programming problem. For the special case of two hypotheses and

Let $R_N = \inf_{\gamma^N \in \Gamma^N} r_N(\gamma^N)$ be the optimal exponent. Let Γ_0^N be the set of all $\gamma^N \in \Gamma^N$ with the property that the set $\{\gamma_1, \ldots, \gamma_N\}$ has at most $M(M-1)/2$ different elements. Let $Q_N = \inf_{\gamma^N \in \Gamma_0^N} r_N(\gamma^N)$ be the optimal exponent, when we restrict to sets of decision rules in Γ_0^N. The following result shows that, asymptotically, optimality is not lost, if we restrict to Γ_0^N.

Theorem 1. *Subject to Assumption* 1 *below*. $\lim_{N\to\infty}(Q_N - R_N) = 0$.

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 35, NO. 2, MARCH 1989 401

Asymptotically Optimal Classification for Multiple Tests with Empirically Observed Statistics

MICHAEL GUTMAN, MEMBER, IEEE

Abstract—The decision problem of testing M hypotheses, when the source is Kth-order Markov and there are M (or fewer) training sequences of length N and a single test sequence of length n, is considered. K, M, n, N are all given. Answers to the following questions are given: What are the requirements on M, n, N to achieve vanishing (exponential) error probabilities? In such cases, how can we determine or bound the exponent? A likelihood ratio test which is allowed to produce a "no-match" decision is shown to provide asymptotically optimal error probabilities and minimum no-match decisions. As an important special case, the binary hypotheses problem without rejection is discussed. It is shown that, for this configuration, only one training sequence is needed to achieve an asymptotically optimal test.

An approach which is closely related to information theory is to investigate asymptotically optimum tests, i.e., those tests with error exponents that asymptotically achieve optimal performance.

The optimum test for the case where the sources are known and where one of the misclassification probabilities is prescribed (case 1) was derived by Neyman and Pearson [2]. Its asymptotic behavior is described in [3] for independent identical distribution (i.i.d.) functions.

For case 2, in which one hypothesis is simple, Hoeffding [4] has shown that for i.i.d. discrete sources with finite

Distributed Detection: Related Works

- Inspired by Tsitsiklis who considered distributed detection with known distributions
  - A key result: Using a single type of compressor for binary hypothesis testing is optimal
- Also inspired by Gutman who adopted an information-theoretic approach to statistical classification
  - Derived an asymptotically optimal type-based test

Math. Control Signals Systems (1988) 1: 167–182

Mathematics of Control, Signals, and Systems

Decentralized Detection by a Large Number of Sensors*

John N. Tsitsiklis†

Abstract. We consider the decentralized detection problem, in which N independent, identical sensors transmit a finite-valued function of their observations to a fusion center which then decides which one of M hypotheses is true. For the case where the number of sensors tends to infinity, we show that it is asymptotically optimal to divide the sensors into $M(M-1)/2$ groups, with all sensors in each group using the same decision rule in deciding what to transmit. We also show how the optimal number of sensors in each group may be determined by solving a mathematical programming problem. For the special case of two hypotheses and

Let $R_N = \inf_{\gamma^N \in \Gamma^N} r_N(\gamma^N)$ be the optimal exponent. Let Γ_0^N be the set of all $\gamma^N \in \Gamma^N$ with the property that the set $\{\gamma_1, \ldots, \gamma_N\}$ has at most $M(M-1)/2$ different elements. Let $Q_N = \inf_{\gamma^N \in \Gamma_0^N} r_N(\gamma^N)$ be the optimal exponent, when we restrict to sets of decision rules in Γ_0^N. The following result shows that, asymptotically, optimality is not lost, if we restrict to Γ_0^N.

Theorem 1. *Subject to Assumption 1 below.* $\lim_{N\to\infty}(Q_N - R_N) = 0$.

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 35, NO. 2, MARCH 1989 401

Asymptotically Optimal Classification for Multiple Tests with Empirically Observed Statistics

MICHAEL GUTMAN, MEMBER, IEEE

Abstract—The decision problem of testing M hypotheses, when the source is Kth-order Markov and there are M (or fewer) training sequences of length N and a single test sequence of length n, is considered. K, M, n, N are all given. Answers to the following questions are given: What are the requirements on M, n, N to achieve vanishing (exponential) error probabilities? In such cases, how can we determine or bound the exponent? A likelihood ratio test which is allowed to produce a "no-match" decision is shown to provide asymptotically optimal error probabilities and minimum no-match decisions. As an important special case, the binary hypotheses problem without rejection is discussed. It is shown that, for this configuration, only one training sequence is needed to achieve an asymptotically optimal test.

An approach which is closely related to information theory is to investigate asymptotically optimum tests, i.e., those tests with error exponents that asymptotically achieve optimal performance.

The optimum test for the case where the sources are known and where one of the misclassification probabilities is prescribed (case 1) was derived by Neyman and Pearson [2]. Its asymptotic behavior is described in [3] for independent identical distribution (i.i.d.) functions.

For case 2, in which one hypothesis is simple, Hoeffding [4] has shown that for i.i.d. discrete sources with finite

Distributed Detection: Related Works

- ♠ Inspired by Tsitsiklis who considered distributed detection with known distributions
  - A key result: Using a single type of compressor for binary hypothesis testing is optimal
- ♣ Also inspired by Gutman who adopted an information-theoretic approach to statistical classification
  - Derived an asymptotically optimal type-based test

Math. Control Signals Systems (1988) 1: 167–182

Mathematics of Control, Signals, and Systems

Decentralized Detection by a Large Number of Sensors*

John N. Tsitsiklis†

Abstract. We consider the decentralized detection problem, in which N independent, identical sensors transmit a finite-valued function of their observations to a fusion center which then decides which one of M hypotheses is true. For the case where the number of sensors tends to infinity, we show that it is asymptotically optimal to divide the sensors into $M(M-1)/2$ groups, with all sensors in each group using the same decision rule in deciding what to transmit. We also show how the optimal number of sensors in each group may be determined by solving a mathematical programming problem. For the special case of two hypotheses and

Let $R_N = \inf_{\gamma^N \in \Gamma^N} r_N(\gamma^N)$ be the optimal exponent. Let Γ_0^N be the set of all $\gamma^N \in \Gamma^N$ with the property that the set $\{\gamma_1, \ldots, \gamma_N\}$ has at most $M(M-1)/2$ different elements. Let $Q_N = \inf_{\gamma^N \in \Gamma_0^N} r_N(\gamma^N)$ be the optimal exponent, when we restrict to sets of decision rules in Γ_0^N. The following result shows that, asymptotically, optimality is not lost, if we restrict to Γ_0^N.

Theorem 1. *Subject to Assumption 1 below.* $\lim_{N\to\infty}(Q_N - R_N) = 0$.

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 35, NO. 2, MARCH 1989 401

Asymptotically Optimal Classification for Multiple Tests with Empirically Observed Statistics

MICHAEL GUTMAN, MEMBER, IEEE

Abstract—The decision problem of testing M hypotheses, when the source is Kth-order Markov and there are M (or fewer) training sequences of length N and a single test sequence of length n, is considered. K, M, n, N are all given. Answers to the following questions are given: What are the requirements on M, n, N to achieve vanishing (exponential) error probabilities? In such cases, how can we determine or bound the exponent? A likelihood ratio test which is allowed to produce a "no-match" decision is shown to provide asymptotically optimal error probabilities and minimum no-match decisions. As an important special case, the binary hypotheses problem without rejection is discussed. It is shown that, for this configuration, only one training sequence is needed to achieve an asymptotically optimal test.

An approach which is closely related to information theory is to investigate asymptotically optimum tests, i.e., those tests with error exponents that asymptotically achieve optimal performance.

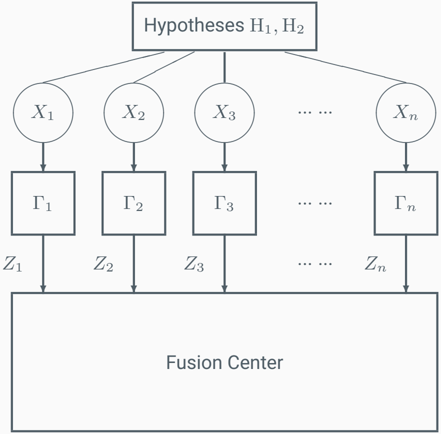
The optimum test for the case where the sources are known and where one of the misclassification probabilities is prescribed (case 1) was derived by Neyman and Pearson [2]. Its asymptotic behavior is described in [3] for independent identical distribution (i.i.d.) functions.

For case 2, in which one hypothesis is simple, Hoeffding [4] has shown that for i.i.d. discrete sources with finite

★ Question: What is the optimal design of the channels and the decision rule at the fusion center?

Distributed Detection: System Model

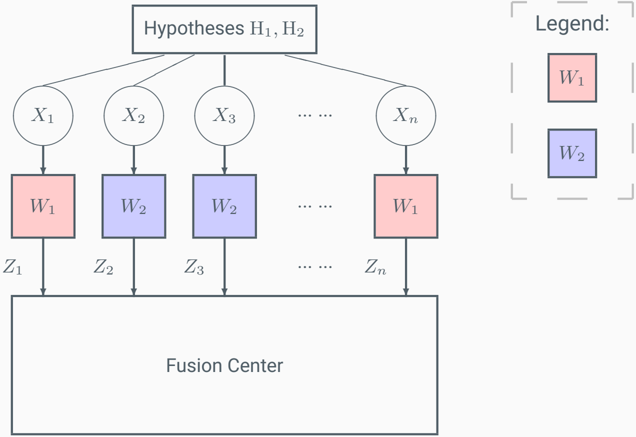
n sensors:



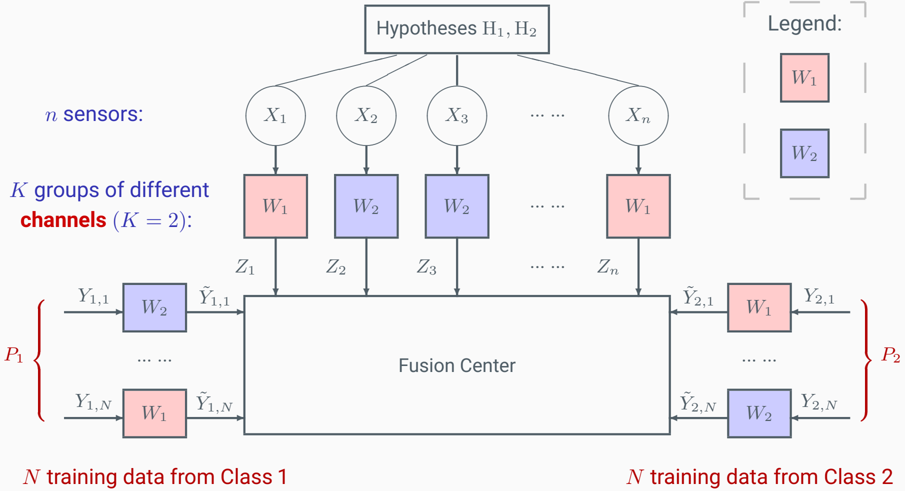
Distributed Detection: System Model

n sensors:

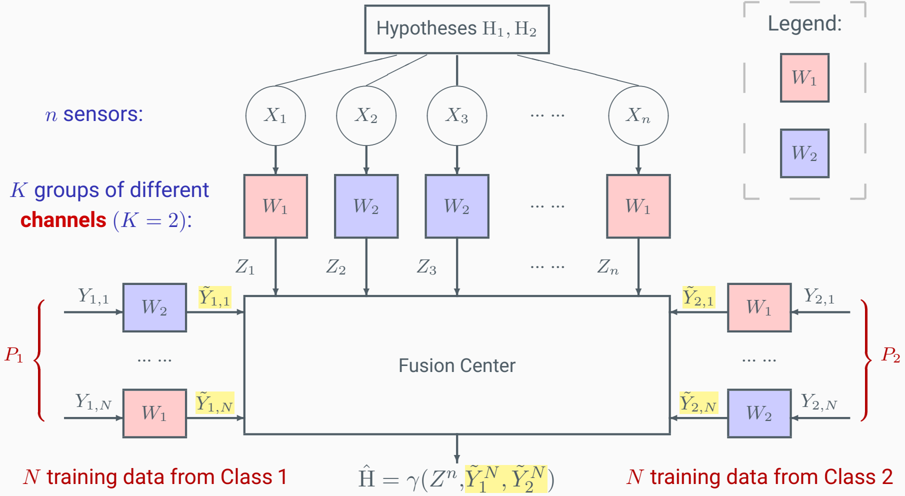
K groups of different **channels** $(K = 2)$:



Distributed Detection: System Model



Distributed Detection: System Model



Distributed Detection: System Model

- Ratio between lengths: $\alpha = \frac{N}{n}$

Distributed Detection: System Model

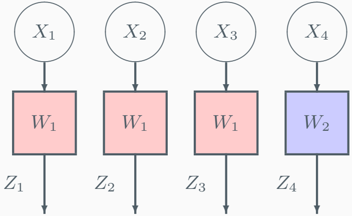
- Ratio between lengths: $\alpha = \frac{N}{n}$
- K different channels: $\mathcal{W} := \{W_i\}_{i \in [K]}$

Distributed Detection: System Model

- Ratio between lengths: $\alpha = \frac{N}{n}$
- K different channels: $\mathcal{W} := \{W_i\}_{i \in [K]}$

Example: $K = 2$, $n = 4$, $N = 5$, $\alpha = \frac{5}{4}$ (to show proportions of different channels)

Test Sequence (Sensor Data)

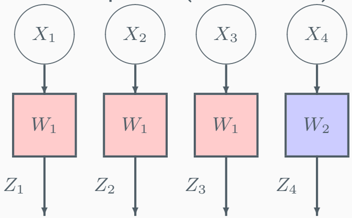


Distributed Detection: System Model

- Ratio between lengths: $\alpha = \frac{N}{n}$
- K different channels: $\mathcal{W} := \{W_i\}_{i \in [K]}$

Example: $K = 2$, $n = 4$, $N = 5$, $\alpha = \frac{5}{4}$ (to show proportions of different channels)

Test Sequence (Sensor Data)



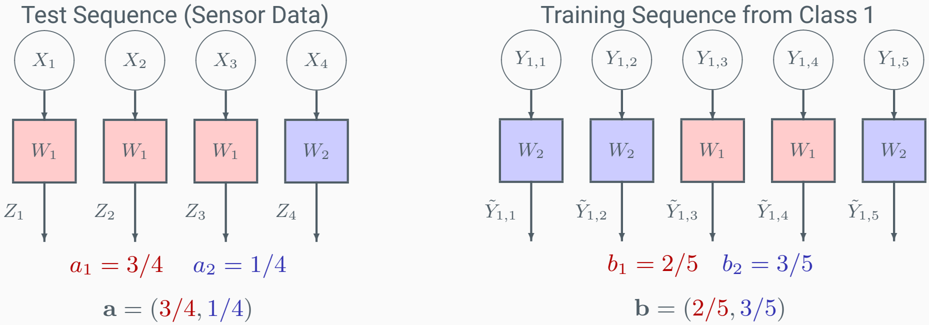
$a_1 = 3/4$ $\quad a_2 = 1/4$

$$\mathbf{a} = (3/4, 1/4)$$

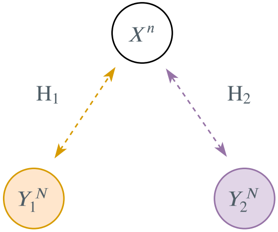
Distributed Detection: System Model

- Ratio between lengths: $\alpha = \frac{N}{n}$
- K different channels: $\mathcal{W} := \{W_i\}_{i \in [K]}$

Example: $K = 2$, $n = 4$, $N = 5$, $\alpha = \frac{5}{4}$ (to show proportions of different channels)

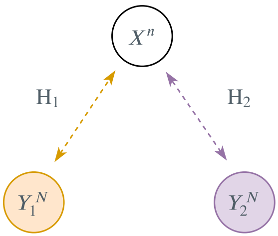


Fusion center decision rule γ: decide between the two hypotheses



Distributed Detection: System Model

Fusion center decision rule γ: decide between the two hypotheses

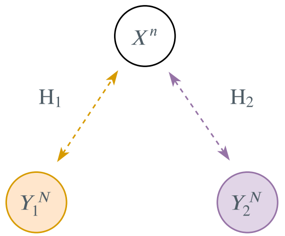


Questions

★ Q1: Optimal fusion center decision rule γ given X^n, Y_1^N, Y_2^N and the channels $\{W_i\}_{i=1}^K$?

Distributed Detection: System Model

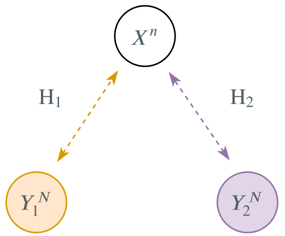
Fusion center decision rule γ: decide between the two hypotheses



Questions

- ★ Q1: Optimal fusion center decision rule γ given X^n, Y_1^N, Y_2^N and the channels $\{W_i\}_{i=1}^K$?
- ★ Q2: Optimal error exponent?

Fusion center decision rule γ: decide between the two hypotheses



Questions

- ★ Q1: Optimal fusion center decision rule γ given X^n, Y_1^N, Y_2^N and the channels $\{W_i\}_{i=1}^K$?
- ★ Q2: Optimal error exponent?
- ★ Q3: Optimal proportions of different channels, i.e., $\mathbf{a} = (a_1, \ldots, a_K)$, $\mathbf{b} = (b_1, \ldots, b_K)$?

Distributed Detection: Setup

- Type-I and type-II error probabilities:

$$\beta_j(\gamma, P_1, P_2) := \Pr\{\gamma(Z^n, \tilde{Y}_1^N, \tilde{Y}_2^N) \neq \mathrm{H}_j \mid \mathrm{H}_j\},\ j \in [2]$$

Distributed Detection: Setup

- Type-I and type-II error probabilities:

$$\beta_j(\gamma, P_1, P_2) := \Pr\{\gamma(Z^n, \tilde{Y}_1^N, \tilde{Y}_2^N) \neq \mathrm{H}_j \mid \mathrm{H}_j\},\ j \in [2]$$

- **Objective:** Consider the family $\Gamma_n(\lambda)$ of all tests γ s.t.

$$\max_{(\tilde{P}_1, \tilde{P}_2)} \beta_1(\gamma, \tilde{P}_1, \tilde{P}_2) \leq \exp(-n\lambda).$$

Given P_1, P_2, we want to derive the optimal type-II error exponent

$$E^* := \liminf_{n\to\infty} \sup_{\gamma \in \Gamma_n(\lambda)} -\frac{1}{n} \log \beta_2(\gamma; P_1, P_2).$$

E^* depends on train/test ratio $\alpha = \frac{N}{n}$, type-I error exponent λ, ratios of channels $\mathbf{a} = (a_1, \ldots, a_K), \mathbf{b} = (b_1, \ldots, b_K)$, and distributions P_1, P_2 (which will be suppressed).

Distributed Detection: Some Definitions

- Linear combinations of KL-divergences

$$\mathrm{LD}(\mathbf{Q},\tilde{\mathbf{Q}},P,\tilde{P}|\alpha,\mathbf{a},\mathbf{b},\mathcal{W}) := \sum_{k\in[K]} \left(a_k D(Q_k\|PW_k) + \alpha b_k D(\tilde{Q}_k\|\tilde{P}W_k)\right),$$

- Linear combinations of KL-divergences

$$\mathrm{LD}(\mathbf{Q},\tilde{\mathbf{Q}},P,\tilde{P}|\alpha,\mathbf{a},\mathbf{b},\mathcal{W}) := \sum_{k\in[K]} \left(a_k D(Q_k\|PW_k) + \alpha b_k D(\tilde{Q}_k\|\tilde{P}W_k)\right),$$

- Set of distributions:

$$\mathcal{Q}_\lambda(\alpha,\mathbf{a},\mathbf{b},\mathcal{W}) := \left\{(\mathbf{Q},\tilde{\mathbf{Q}}) : \min_{\tilde{P}\in\mathcal{P}(\mathcal{X})} \mathrm{LD}(\mathbf{Q},\tilde{\mathbf{Q}},\tilde{P},\tilde{P}) \le \lambda\right\}.$$

When $K = 1$ and $W_1 = I_{|\mathcal{X}|\times|\mathcal{X}|} \Longrightarrow$ recovers to Gutman's classification problem setup

Theorem 1 (Asymptotically optimal type-II error exponent)

Given any pair of target distributions (P_1, P_2)*, we have*

$$E^*(\lambda, \alpha, \mathbf{a}, \mathbf{b}) = \min_{(\mathbf{Q}, \tilde{\mathbf{Q}}) \in \mathcal{Q}_\lambda(\alpha, \mathbf{a}, \mathbf{b}, V, \mathcal{W})} \mathrm{LD}(\mathbf{Q}, \tilde{\mathbf{Q}}, P_2, P_1).$$

Theorem 1 (Asymptotically optimal type-II error exponent)

Given any pair of target distributions (P_1, P_2)*, we have*

$$E^*(\lambda, \alpha, \mathbf{a}, \mathbf{b}) = \min_{(\mathbf{Q}, \tilde{\mathbf{Q}}) \in \mathcal{Q}_\lambda(\alpha, \mathbf{a}, \mathbf{b}, V, \mathcal{W})} \mathrm{LD}(\mathbf{Q}, \tilde{\mathbf{Q}}, P_2, P_1).$$

In the achievability proof, we use the asymptotically optimal fusion center type-based test:

$$\gamma(Z^n, \tilde{Y}_1^N, \tilde{Y}_2^N) = \begin{cases} \mathrm{H}_1 & \text{if } \min_{\tilde{P}} \mathrm{LD}\Big(\{T_{Z^{na_k}}\}_{k\in[K]}, \{T_{\tilde{Y}_1^{Nb_k}}\}_{k\in[K]}, \tilde{P}, \tilde{P}\Big) \le \lambda, \\ \mathrm{H}_2 & \text{otherwise.} \end{cases}$$

Theorem 1 (Asymptotically optimal type-II error exponent)

Given any pair of target distributions (P_1, P_2), *we have*

$$E^*(\lambda, \alpha, \mathbf{a}, \mathbf{b}) = \min_{(\mathbf{Q}, \tilde{\mathbf{Q}}) \in \mathcal{Q}_\lambda(\alpha, \mathbf{a}, \mathbf{b}, V, \mathcal{W})} \mathrm{LD}(\mathbf{Q}, \tilde{\mathbf{Q}}, P_2, P_1).$$

In the achievability proof, we use the asymptotically optimal fusion center type-based test:

$$\gamma(Z^n, \tilde{Y}_1^N, \tilde{Y}_2^N) = \begin{cases} \mathrm{H}_1 & \text{if } \min_{\tilde{P}} \mathrm{LD}\Big(\{T_{Z^{na_k}}\}_{k\in[K]}, \{T_{\tilde{Y}_1^{Nb_k}}\}_{k\in[K]}, \tilde{P}, \tilde{P}\Big) \le \lambda, \\ \mathrm{H}_2 & \text{otherwise.} \end{cases}$$

Do NOT make use of $\tilde{Y}_2^N$!

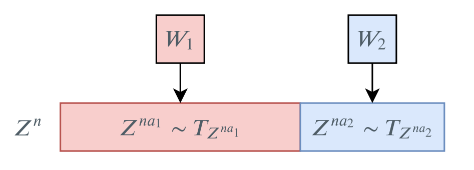
Theorem 1 (Asymptotically optimal type-II error exponent)

Given any pair of target distributions (P_1, P_2), *we have*

$$E^*(\lambda, \alpha, \mathbf{a}, \mathbf{b}) = \min_{(\mathbf{Q}, \tilde{\mathbf{Q}}) \in \mathcal{Q}_\lambda(\alpha, \mathbf{a}, \mathbf{b}, V, \mathcal{W})} \mathrm{LD}(\mathbf{Q}, \tilde{\mathbf{Q}}, P_2, P_1).$$

In the achievability proof, we use the asymptotically optimal fusion center type-based test:

$$\gamma(Z^n, \tilde{Y}_1^N, \tilde{Y}_2^N) = \begin{cases} \mathrm{H}_1 & \text{if } \min_{\tilde{P}} \mathrm{LD}\Big(\{T_{Z^{na_k}}\}_{k\in[K]}, \{T_{\tilde{Y}_1^{Nb_k}}\}_{k\in[K]}, \tilde{P}, \tilde{P}\Big) \le \lambda, \\ \mathrm{H}_2 & \text{otherwise.} \end{cases}$$



Do NOT make use of $\tilde{Y}_2^N$!

Further discussions on (a, b)

Tsitsiklis' result (known P_1, P_2): binary hypothesis testing, only 1 type of channel optimizes the type-II error exponent and Bayesian error exponent.

Further discussions on $(\mathbf{a}, \mathbf{b})$

Tsitsiklis' result (known P_1, P_2): binary hypothesis testing, only 1 type of channel optimizes the type-II error exponent and Bayesian error exponent.

Ours: E^* depends on $\mathbf{a}, \mathbf{b} \Longrightarrow$ can further maximize over $\mathbf{a}, \mathbf{b}$

Further discussions on (a, b)

Tsitsiklis' result (known P_1, P_2): binary hypothesis testing, only 1 type of channel optimizes the type-II error exponent and Bayesian error exponent.

Ours: E^* depends on $\mathbf{a}, \mathbf{b} \Longrightarrow$ can further maximize over $\mathbf{a}, \mathbf{b}$

- Let $f_\alpha(\mathbf{a}, \mathbf{b}, \lambda) := \min_{\substack{(\mathbf{Q}, \tilde{\mathbf{Q}}) \\ \in \mathcal{Q}_\lambda(\alpha, \mathbf{a}, \mathbf{b}, V, \mathcal{W})}} \mathrm{LD}\Big(\mathbf{Q}, \tilde{\mathbf{Q}}, P_2, P_1 | \alpha, \mathbf{a}, \mathbf{b}, \mathcal{W}\Big)$ (i.e. type-II error exponent)

Further discussions on (a, b)

Tsitsiklis' result (known P_1, P_2): binary hypothesis testing, only 1 type of channel optimizes the type-II error exponent and Bayesian error exponent.

Ours: E^* depends on $\mathbf{a}, \mathbf{b} \Longrightarrow$ can further maximize over $\mathbf{a}, \mathbf{b}$

- Let $f_\alpha(\mathbf{a}, \mathbf{b}, \lambda) := \min_{\substack{(\mathbf{Q}, \tilde{\mathbf{Q}}) \\ \in \mathcal{Q}_\lambda(\alpha, \mathbf{a}, \mathbf{b}, V, \mathcal{W})}} \mathrm{LD}\Big(\mathbf{Q}, \tilde{\mathbf{Q}}, P_2, P_1 | \alpha, \mathbf{a}, \mathbf{b}, \mathcal{W}\Big)$ (i.e. type-II error exponent)
- Maximized over $(\mathbf{a}, \mathbf{b})$

$$f_\alpha^*(\lambda) = \max_{(\mathbf{a}, \mathbf{b})} f_\alpha(\mathbf{a}, \mathbf{b}, \lambda)$$

Further discussions on (a, b)

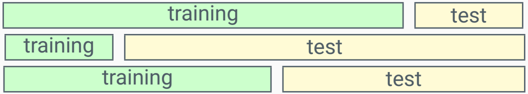
Tsitsiklis' result (known P_1, P_2): binary hypothesis testing, only 1 type of channel optimizes the type-II error exponent and Bayesian error exponent.

Ours: E^* depends on $\mathbf{a}, \mathbf{b} \Longrightarrow$ can further maximize over $\mathbf{a}, \mathbf{b}$

- Let $f_\alpha(\mathbf{a}, \mathbf{b}, \lambda) := \min\limits_{\substack{(\mathbf{Q}, \tilde{\mathbf{Q}}) \\ \in \mathcal{Q}_\lambda(\alpha, \mathbf{a}, \mathbf{b}, V, \mathcal{W})}} \mathrm{LD}\Big(\mathbf{Q}, \tilde{\mathbf{Q}}, P_2, P_1 | \alpha, \mathbf{a}, \mathbf{b}, \mathcal{W}\Big)$ (i.e. type-II error exponent)
- Maximized over $(\mathbf{a}, \mathbf{b})$

$$f_\alpha^*(\lambda) = \max_{(\mathbf{a}, \mathbf{b})} f_\alpha(\mathbf{a}, \mathbf{b}, \lambda)$$

- **Three cases:**
  - $\alpha \to \infty$:
  - $\alpha \to 0$:
  - α moderate:



(Details in full thesis)

Further discussions on (a, b): $\alpha \to \infty$

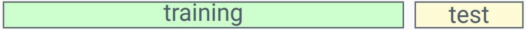
training | test

Corollary 1

Given any $\lambda \in \mathbb{R}_+$, as $\alpha \to \infty$, we have

$$f^*_\infty(\lambda) = \max_{k \in [K]} f_\infty(\mathbf{e}_k, \mathbf{e}_k, \lambda),$$

and thus the maximizers $(\mathbf{a}^*, \mathbf{b}^*)$ for $f_\infty(\mathbf{a}, \mathbf{b}, \lambda)$ satisfies that $(\mathbf{a}^*, \mathbf{b}^*)$ are both deterministic and $\mathbf{a}^* = \mathbf{b}^*$. (e.g. $\mathbf{a} = (1, 0, 0, \dots, 0)$, $\mathbf{b} = (1, 0, 0, \dots, 0)$)



standard basis vector: $\mathbf{e}_k$

1	⋯⋯	$k-1$	k	$k+1$	⋯⋯	K
0	……	0	1	0	……	0

Corollary 1

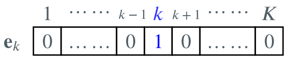
Given any $\lambda \in \mathbb{R}_+$, as $\alpha \to \infty$, we have

$$f_\infty^*(\lambda) = \max_{k \in [K]} f_\infty(\mathbf{e}_k, \mathbf{e}_k, \lambda),$$

and thus the maximizers $(\mathbf{a}^*, \mathbf{b}^*)$ for $f_\infty(\mathbf{a}, \mathbf{b}, \lambda)$ satisfies that $(\mathbf{a}^*, \mathbf{b}^*)$ are both deterministic and $\mathbf{a}^* = \mathbf{b}^*$. (e.g. $\mathbf{a} = (1, 0, 0, \ldots, 0)$, $\mathbf{b} = (1, 0, 0, \ldots, 0)$)

training | test

standard basis vector:



Corollary 1

Given any $\lambda \in \mathbb{R}_+$, as $\alpha \to \infty$, we have

$$f_\infty^*(\lambda) = \max_{k \in [K]} f_\infty(\mathbf{e}_k, \mathbf{e}_k, \lambda),$$

and thus the maximizers $(\mathbf{a}^*, \mathbf{b}^*)$ for $f_\infty(\mathbf{a}, \mathbf{b}, \lambda)$ satisfies that $(\mathbf{a}^*, \mathbf{b}^*)$ are both deterministic and $\mathbf{a}^* = \mathbf{b}^*$. (e.g. $\mathbf{a} = (1, 0, 0, \ldots, 0)$, $\mathbf{b} = (1, 0, 0, \ldots, 0)$)

Explanation: optimal to use only one identical channel to process both test and training sequences.

Further discussions on $(\mathbf{a}, \mathbf{b})$: $\alpha \to \infty$

training | test

standard basis vector: $\mathbf{e}_k$

1	······	$k-1$	k	$k+1$	······	K
0	……	0	1	0	……	0

Corollary 1

Given any $\lambda \in \mathbb{R}_+$, as $\alpha \to \infty$, we have

$$f_\infty^*(\lambda) = \max_{k \in [K]} f_\infty(\mathbf{e}_k, \mathbf{e}_k, \lambda),$$

and thus the maximizers $(\mathbf{a}^*, \mathbf{b}^*)$ for $f_\infty(\mathbf{a}, \mathbf{b}, \lambda)$ satisfies that $(\mathbf{a}^*, \mathbf{b}^*)$ are both deterministic and $\mathbf{a}^* = \mathbf{b}^*$. (e.g. $\mathbf{a} = (1, 0, 0, \ldots, 0)$, $\mathbf{b} = (1, 0, 0, \ldots, 0)$)

Explanation: optimal to use only one identical channel to process both test and training sequences.

$\Longrightarrow$ analogous to Tsitsiklis' result

Further discussions on (a, b): $\alpha \to 0$

training | test

Lemma 1

Given any $(\mathbf{a}, \mathbf{b}) \in \mathcal{P}([K])^2$ *and any* $\lambda \in \mathbb{R}_+$, $\exists \alpha_0(\mathbf{a}, \mathbf{b}, \lambda) > 0$, *if* $\alpha \leq \alpha_0(\mathbf{a}, \mathbf{b}, \lambda)$, *then*

$$f_\alpha(\mathbf{a}, \mathbf{b}, \lambda) = 0.$$

Further discussions on (a, b): $\alpha \to 0$

training | test

Lemma 1

Given any $(\mathbf{a}, \mathbf{b}) \in \mathcal{P}([K])^2$ *and any* $\lambda \in \mathbb{R}_+$, $\exists \alpha_0(\mathbf{a}, \mathbf{b}, \lambda) > 0$, *if* $\alpha \leq \alpha_0(\mathbf{a}, \mathbf{b}, \lambda)$, *then*

$$f_\alpha(\mathbf{a}, \mathbf{b}, \lambda) = 0.$$

Explanation: When the training data are too scarce compared to the test data, if we require the type-I error decays exponentially fast, the decision rule γ always declares H_1 and the type-II error$= \exp(-n f_\alpha(\mathbf{a}, \mathbf{b}, \lambda)) = 1$ all the time.

Further discussions on $(\mathbf{a}, \mathbf{b})$: $\alpha \to 0$

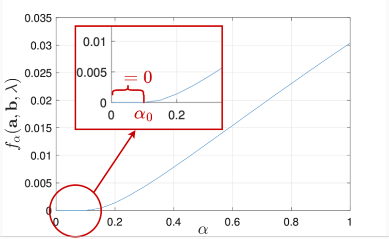
training | test

Lemma 1

Given any $(\mathbf{a}, \mathbf{b}) \in \mathcal{P}([K])^2$ *and any* $\lambda \in \mathbb{R}_+$, $\exists \alpha_0(\mathbf{a}, \mathbf{b}, \lambda) > 0$, *if* $\alpha \leq \alpha_0(\mathbf{a}, \mathbf{b}, \lambda)$, *then*

$$f_\alpha(\mathbf{a}, \mathbf{b}, \lambda) = 0.$$

Explanation: When the training data are too scarce compared to the test data, if we require the type-I error decays exponentially fast, the decision rule γ always declares H_1 and the type-II error$= \exp(-n f_\alpha(\mathbf{a}, \mathbf{b}, \lambda)) = 1$ all the time.



Distributed Detection: Summary

★ Problem setup: distributed detection with test and training data

Distributed Detection: Summary

★ Problem setup: distributed detection with test and training data

★ Main results: $n \to \infty$

- Optimal fusion center type-based test γ

Distributed Detection: Summary

★ Problem setup: distributed detection with test and training data

★ Main results: $n \to \infty$

- Optimal fusion center type-based test γ
- Optimal type-II error exponent $E^*(\lambda, \alpha, \mathbf{a}, \mathbf{b})$

Distributed Detection: Summary

★ Problem setup: distributed detection with test and training data

★ Main results: $n \to \infty$

- Optimal fusion center type-based test γ
- Optimal type-II error exponent $E^*(\lambda, \alpha, \mathbf{a}, \mathbf{b})$
- Optimal design of $(\mathbf{a}, \mathbf{b})$ when $\alpha \to \infty$: one identical channel at all test and training data

Distributed Detection: Summary

★ Problem setup: distributed detection with test and training data

★ Main results: $n \to \infty$

- Optimal fusion center type-based test γ
- Optimal type-II error exponent $E^*(\lambda, \alpha, \mathbf{a}, \mathbf{b})$
- Optimal design of $(\mathbf{a}, \mathbf{b})$ when $\alpha \to \infty$: one identical channel at all test and training data

★ We also generalized the results to distributed detection problem with $m \geq 2$ hypotheses and a rejection option .

Distributed Detection: Summary

★ Problem setup: distributed detection with test and training data

★ Main results: $n \to \infty$
- Optimal fusion center type-based test γ
- Optimal type-II error exponent $E^*(\lambda, \alpha, \mathbf{a}, \mathbf{b})$
- Optimal design of $(\mathbf{a}, \mathbf{b})$ when $\alpha \to \infty$: one identical channel at all test and training data

★ We also generalized the results to distributed detection problem with $m \geq 2$ hypotheses and a rejection option .

H. He, L. Zhou, and V. Y. F. Tan, "Distributed detection with empirically observed statistics", *IEEE Transactions on Information Theory*, vol. 66, pp. 4349–4367, 2020.

Outline

Change-Point Detection with Training Sequences Motivation

- Change-Point Detection (CPD) with Training Sequences:
  Example – Room light change detection

Change-Point Detection with Training Sequences Motivation

- Change-Point Detection (CPD) with Training Sequences:
  Example – Room light change detection

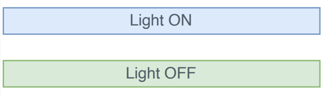
- Sensor detects when room light changes: given a test sequence of sensor data
  $\Longrightarrow$ **Offline CPD**

- Change-Point Detection (CPD) with Training Sequences:
  Example – Room light change detection

- Sensor detects when room light changes: given a test sequence of sensor data $\Longrightarrow$ **Offline CPD**
- Unknown distributions

- Change-Point Detection (CPD) with Training Sequences:
  Example – Room light change detection

- Sensor detects when room light changes: given a test sequence of sensor data
  $\Longrightarrow$ **Offline CPD**
- Unknown distributions
- **Training sequences:** collect sensor data when light is on or off, respectively



Change-Point Detection with Training Sequences Motivation

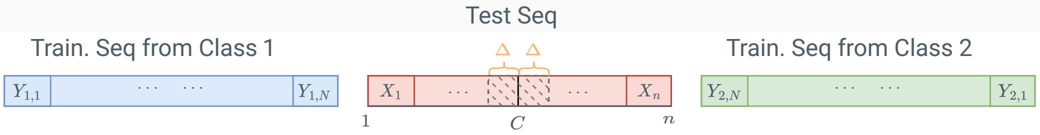
- Change-Point Detection (CPD) with Training Sequences:
  Example – Room light change detection

- Sensor detects when room light changes: given a test sequence of sensor data
  $\Longrightarrow$ **Offline CPD**
- Unknown distributions
- **Training sequences:** collect sensor data when light is on or off, respectively

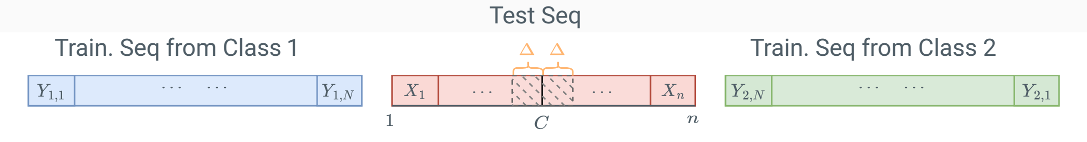
Light ON

Light OFF

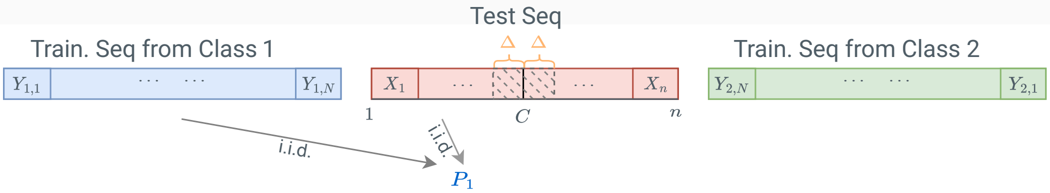
- Change-point detector: test $+$ training sequences



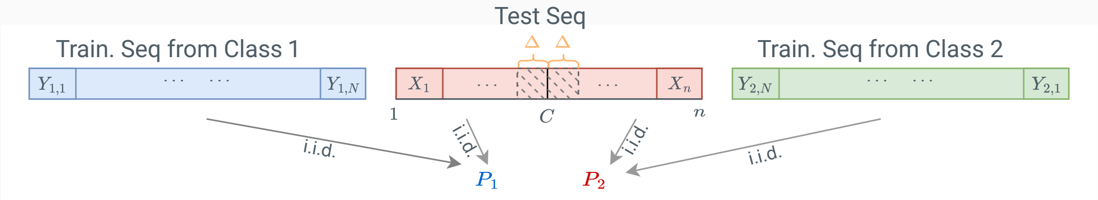
- A sequence of observations $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$



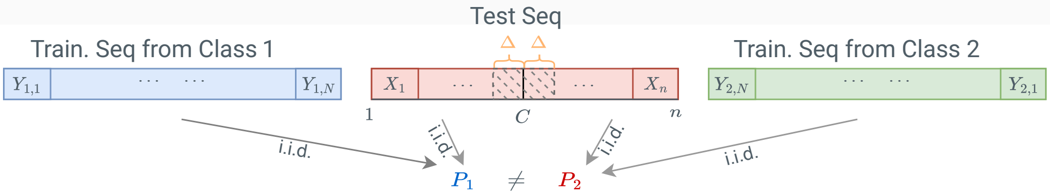
- A sequence of observations $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$
- A single change-point $C = \lceil \alpha n \rceil \in [1 : n]$



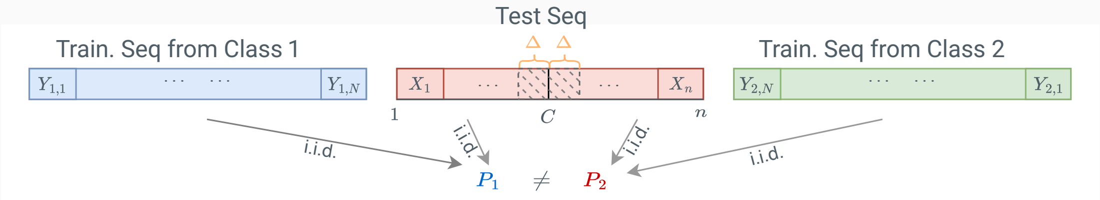
- A sequence of observations $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$
- A single change-point $C = \lceil \alpha n \rceil \in [1 : n]$



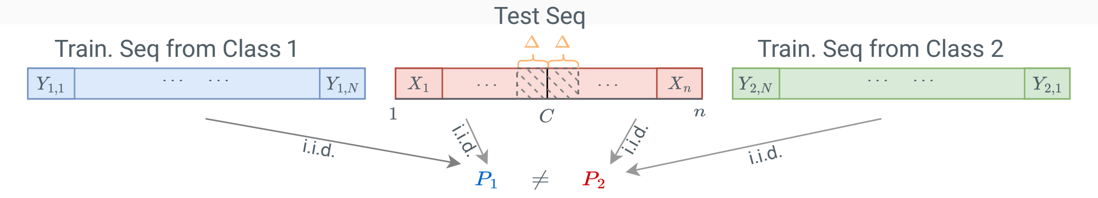
- A sequence of observations $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$
- A single change-point $C = \lceil \alpha n \rceil \in [1 : n]$



- A sequence of observations $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$
- A single change-point $C = \lceil \alpha n \rceil \in [1 : n]$

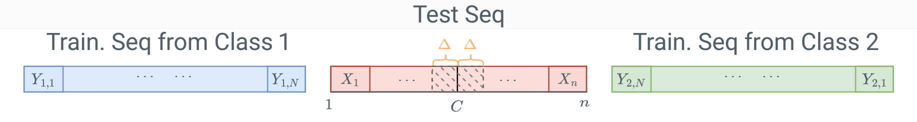


- A sequence of observations $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$
- A single change-point $C = \lceil \alpha n \rceil \in [1:n]$
- $N = \lceil rn \rceil$ for some $r \in \mathbb{R}_+$

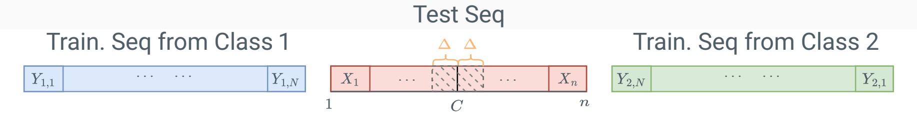


- A sequence of observations $X^n = (X_1, \dots, X_n) \in \mathcal{X}^n$
- A single change-point $C = \lceil \alpha n \rceil \in [1:n]$
- $N = \lceil rn \rceil$ for some $r \in \mathbb{R}_+$
- An estimator $\gamma : \mathcal{X}^{n+2N} \mapsto [n] \cup \{\mathrm{e}\}$:

$$\begin{cases} \text{either declare one of } n \text{ points in the test sequence} \\ \text{or declare that an ``erasure'' has occurred} \end{cases}$$



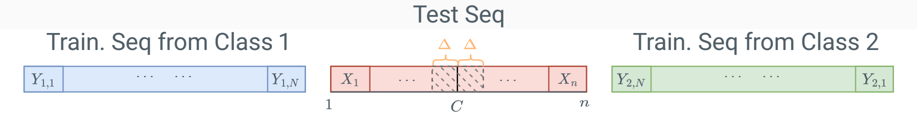
- Performance metrics: given any true change-point $C \in [n]$, (X^n, Y_1^N, Y_2^N) is distributed as $X^C \sim P_1^C, X_{C+1}^n \sim P_2^{n-C}, Y_1^N \sim P_1^N$, and $Y_2^N \sim P_2^N$



- Performance metrics: given any true change-point $C \in [n]$, (X^n, Y_1^N, Y_2^N) is distributed as $X^C \sim P_1^C, X_{C+1}^n \sim P_2^{n-C}, Y_1^N \sim P_1^N$, and $Y_2^N \sim P_2^N$
  Undetected error probability:

$$\mathbb{P}_C\{\mathcal{E}_C\} := \Pr\left\{\gamma(X^n, Y_1^N, Y_2^N) \notin [C \pm \Delta] \cup \{\mathrm{e}\}\right\},$$

  where Δ represents the confidence width between the output and the true change-point and $[a \pm b] := [a - b, a + b]$.



- Performance metrics: given any true change-point $C \in [n]$, (X^n, Y_1^N, Y_2^N) is distributed as $X^C \sim P_1^C, X_{C+1}^n \sim P_2^{n-C}, Y_1^N \sim P_1^N$, and $Y_2^N \sim P_2^N$

  Undetected error probability:

$$\mathbb{P}_C\{\mathcal{E}_C\} := \Pr\left\{\gamma(X^n, Y_1^N, Y_2^N) \notin [C \pm \Delta] \cup \{\mathrm{e}\}\right\},$$

  where Δ represents the confidence width between the output and the true change-point and $[a \pm b] := [a-b, a+b]$.

  Erasure probability:

$$\mathbb{P}_C\{\mathcal{E}_\mathrm{e}\} := \Pr\left\{\gamma(X^n, Y_1^N, Y_2^N) = \mathrm{e}\right\}.$$

Definition 1 (Good Estimator)

For any $\Delta \in [0, n/2)$, any $r \in \mathbb{R}_+$, any $(\lambda, \epsilon) \in \mathbb{R}_+ \times [0, 1)$, and any $t \in [0, 1/2)$, given any particular pair $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$, an estimator $\gamma : \mathcal{X}^{n+2N} \mapsto [n] \cup \{\mathrm{e}\}$ is said to be $(n, \Delta, r, \lambda, \epsilon, t)$-good if

$$\max_{C \in [n]} \mathbb{P}_C\{\mathcal{E}_\mathrm{e}\} \le \epsilon,$$

and for all $(\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}(\mathcal{X}^2)$,

$$\max_{C \in [n]} \tilde{\mathbb{P}}_C\{\mathcal{E}_C\} \le \exp(-n^{1-t}\lambda).$$

Definition 1 (Good Estimator)

For any $\Delta \in [0, n/2)$, any $r \in \mathbb{R}_+$, any $(\lambda, \epsilon) \in \mathbb{R}_+ \times [0, 1)$, and any $t \in [0, 1/2)$, given any particular pair $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$, an estimator $\gamma : \mathcal{X}^{n+2N} \mapsto [n] \cup \{\mathrm{e}\}$ is said to be $(n, \Delta, r, \lambda, \epsilon, t)$-good if

$$\max_{C \in [n]} \mathbb{P}_C\{\mathcal{E}_\mathrm{e}\} \le \epsilon,$$

and for all $(\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}(\mathcal{X}^2)$,

$$\max_{C \in [n]} \tilde{\mathbb{P}}_C\{\mathcal{E}_C\} \le \exp(-n^{1-t}\lambda).$$

- $t = 0$: decay **exponentially fast**, large deviations regime

Definition 1 (Good Estimator)

For any $\Delta \in [0, n/2)$, any $r \in \mathbb{R}_+$, any $(\lambda, \epsilon) \in \mathbb{R}_+ \times [0, 1)$, and any $t \in [0, 1/2)$, given any particular pair $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$, an estimator $\gamma : \mathcal{X}^{n+2N} \mapsto [n] \cup \{e\}$ is said to be $(n, \Delta, r, \lambda, \epsilon, t)$-good if

$$\max_{C \in [n]} \mathbb{P}_C\{\mathcal{E}_e\} \le \epsilon,$$

and for all $(\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}(\mathcal{X}^2)$,

$$\max_{C \in [n]} \tilde{\mathbb{P}}_C\{\mathcal{E}_C\} \le \exp(-n^{1-t}\lambda).$$

- $t = 0$: decay **exponentially fast**, large deviations regime
- $t \in (0, 1/2)$: decay **subexponentially fast**, moderate deviations regime

Definition 1 (Good Estimator)

For any $\Delta \in [0, n/2)$, any $r \in \mathbb{R}_+$, any $(\lambda, \epsilon) \in \mathbb{R}_+ \times [0, 1)$, and any $t \in [0, 1/2)$, given any particular pair $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$, an estimator $\gamma : \mathcal{X}^{n+2N} \mapsto [n] \cup \{\mathrm{e}\}$ is said to be $(n, \Delta, r, \lambda, \epsilon, t)$-good if

$$\max_{C \in [n]} \mathbb{P}_C\{\mathcal{E}_\mathrm{e}\} \le \epsilon,$$

and for all $(\tilde{P}_1, \tilde{P}_2) \in \mathcal{P}(\mathcal{X}^2)$,

$$\max_{C \in [n]} \tilde{\mathbb{P}}_C\{\mathcal{E}_C\} \le \exp(-n^{1-t}\lambda).$$

- $t = 0$: decay **exponentially fast**, large deviations regime
- $t \in (0, 1/2)$: decay **subexponentially fast**, moderate deviations regime

- **Goal: what is the smallest Δ a good estimator can achieve?**

Theorem 2 (Optimal confidence width)

For any $r \in \mathbb{R}_+$, $\epsilon \in [0,1)$, any pair of distributions $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$, the optimal NCW is

$$\bar{\Delta}^*(r, \lambda, P_1, P_2) = \begin{cases} \mathrm{G}_{\min}^{-1}(\lambda), & \lambda \in \left(0, \mathrm{G}_{\min}\left(\frac{1}{2}\right)\right), \\ \frac{1}{2}, & \text{otherwise}; \end{cases}$$

($\mathrm{G}_{\min}$ *is based on Jensen-Shannon divergence and* P_1, P_2)

(λ *is the undetected error exponent*)

In the moderate deviations regime, the t-optimal NCW for any $t \in (0, 1/2)$ and $\lambda > 0$ is

$$\bar{\Delta}_t^*(r, \lambda, P_1, P_2) = \max_{\alpha \in [0,1]} \frac{\sqrt{\lambda}\left(\sqrt{\alpha(\alpha + r)\chi_2(P_1\|P_2)} + \sqrt{(1-\alpha)(1-\alpha+r)\chi_2(P_2\|P_1)}\,\right)}{\sqrt{2r\chi_2(P_1\|P_2)\chi_2(P_2\|P_1)}}.$$

Theorem 2 (Optimal confidence width)

For any $r \in \mathbb{R}_+$, $\epsilon \in [0,1)$, any pair of distributions $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$, the optimal NCW is

$$\bar{\Delta}^*(r, \lambda, P_1, P_2) = \begin{cases} \mathrm{G}_{\min}^{-1}(\lambda), & \lambda \in \left(0, \mathrm{G}_{\min}\left(\frac{1}{2}\right)\right), \\ \frac{1}{2}, & \text{otherwise}; \end{cases}$$

($\mathrm{G}_{\min}$ *is based on Jensen-Shannon divergence and* P_1, P_2)
(λ *is the undetected error exponent*)

In the moderate deviations regime, the t-optimal NCW for any $t \in (0, 1/2)$ and $\lambda > 0$ is

$$\bar{\Delta}_t^*(r, \lambda, P_1, P_2) = \max_{\alpha \in [0,1]} \frac{\sqrt{\lambda}\left(\sqrt{\alpha(\alpha + r)\chi_2(P_1\|P_2)} + \sqrt{(1-\alpha)(1-\alpha+r)\chi_2(P_2\|P_1)}\right)}{\sqrt{2r\chi_2(P_1\|P_2)\chi_2(P_2\|P_1)}}.$$

For any $t \in [0, 1/2)$, $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ is independent of $\epsilon \implies$ strong converses hold.

Theorem 2 (Optimal confidence width)

For any $r \in \mathbb{R}_+$, $\epsilon \in [0,1)$, *any pair of distributions* $(P_1, P_2) \in \mathcal{P}(\mathcal{X})^2$, *the optimal NCW is*

$$\bar{\Delta}^*(r, \lambda, P_1, P_2) = \begin{cases} \mathrm{G}_{\min}^{-1}(\lambda), & \lambda \in \left(0, \mathrm{G}_{\min}\left(\frac{1}{2}\right)\right), \\ \frac{1}{2}, & \text{otherwise}; \end{cases}$$

($\mathrm{G}_{\min}$ *is based on Jensen-Shannon divergence and* P_1, P_2)

(λ *is the undetected error exponent*)

In the moderate deviations regime, the t*-optimal NCW for any* $t \in (0, 1/2)$ *and* $\lambda > 0$ *is*

$$\bar{\Delta}_t^*(r, \lambda, P_1, P_2) = \max_{\alpha \in [0,1]} \frac{\sqrt{\lambda}\left(\sqrt{\alpha(\alpha+r)\chi_2(P_1\|P_2)} + \sqrt{(1-\alpha)(1-\alpha+r)\chi_2(P_2\|P_1)}\right)}{\sqrt{2r\chi_2(P_1\|P_2)\chi_2(P_2\|P_1)}}.$$

For any $t \in [0, 1/2)$, $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ is independent of $\epsilon \Longrightarrow$ strong converses hold.

※ Refer to the full thesis for the asymptotically optimal estimator

Optimal NCW $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ INCREASES when

Optimal NCW $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ INCREASES when

★ λ (error decaying rate) increases;

Optimal NCW $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ INCREASES when

★ λ (error decaying rate) increases;

Explanations: the requirement $\max_{C\in[n]} \tilde{\mathbb{P}}\{\mathcal{E}_C\} \leq \exp(-n^{1-t}\lambda)$ becomes more stringent.

Optimal NCW $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ INCREASES when

★ λ (error decaying rate) increases;

Explanations: the requirement $\max_{C \in [n]} \tilde{\mathbb{P}}\{\mathcal{E}_C\} \leq \exp(-n^{1-t}\lambda)$ becomes more stringent.

★ r (train/test ratio) decreases;

Optimal NCW $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ INCREASES when

- ★ λ (error decaying rate) increases;

  Explanations: the requirement $\max_{C\in[n]} \tilde{\mathbb{P}}\{\mathcal{E}_C\} \leq \exp(-n^{1-t}\lambda)$ becomes more stringent.

- ★ r (train/test ratio) decreases;

  Explanations: $r = \frac{N}{n} \downarrow$, and thus less knowledge about distributions P_1 and P_2 can be learned from the training sequences.

Optimal NCW $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ INCREASES when

- ★ λ (error decaying rate) increases;

  Explanations: the requirement $\max_{C \in [n]} \tilde{\mathbb{P}}\{\mathcal{E}_C\} \leq \exp(-n^{1-t}\lambda)$ becomes more stringent.

- ★ r (train/test ratio) decreases;

  Explanations: $r = \frac{N}{n} \downarrow$, and thus less knowledge about distributions P_1 and P_2 can be learned from the training sequences.

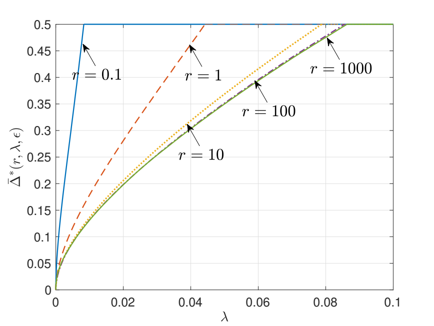


Fig: Large deviations regime.

 (⁕ Refer to the thesis for more figures in moderate deviations regime.)

Optimal NCW $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ INCREASES when

★ the distance between P_1 and P_2 decreases;

Optimal NCW $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ INCREASES when

★ the distance between P_1 and P_2 decreases;

Explanations: it is harder to distinguish between them and thus the accuracy of detection decreases, leading to a larger confidence width.

Optimal NCW $\bar{\Delta}_t^*(r, \lambda, P_1, P_2)$ INCREASES when

★ the distance between P_1 and P_2 decreases;

Explanations: it is harder to distinguish between them and thus the accuracy of detection decreases, leading to a larger confidence width.

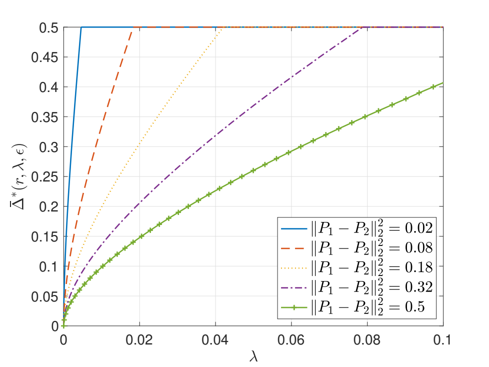


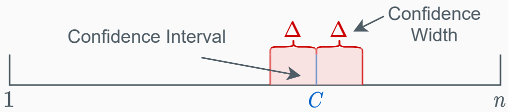
Fig: Large deviations regime

(⁜ Refer to the thesis for more figures in moderate deviations regime.)

Main contributions: offline single-CPD with training sequences

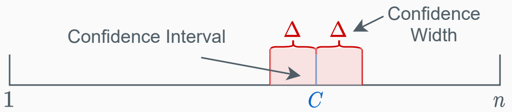
Main contributions: offline single-CPD with training sequences

- The asymptotically optimal confidence width between the estimated and true change-points under



Main contributions: offline single-CPD with training sequences

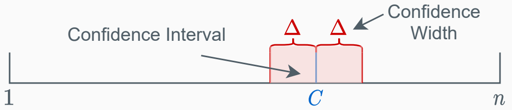
- The asymptotically optimal confidence width between the estimated and true change-points under



- **Large deviations regime:** the undetected error probability decays **exponentially** fast

Main contributions: offline single-CPD with training sequences

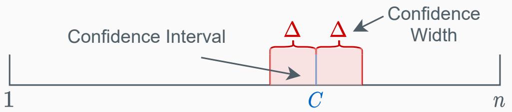
- The asymptotically optimal confidence width between the estimated and true change-points under



  - **Large deviations regime:** the undetected error probability decays **exponentially** fast
  - **Moderate deviations regime:** ——decays **sub-exponentially** fast

Main contributions: offline single-CPD with training sequences

- The asymptotically optimal confidence width between the estimated and true change-points under

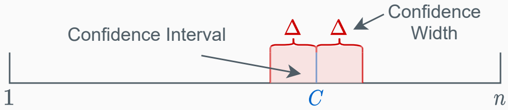


  - **Large deviations regime:** the undetected error probability decays **exponentially** fast
  - **Moderate deviations regime:** ——decays **sub-exponentially** fast

- An asymptotically optimal estimator based on test and training sequences under both regimes

Main contributions: offline single-CPD with training sequences

- The asymptotically optimal confidence width between the estimated and true change-points under

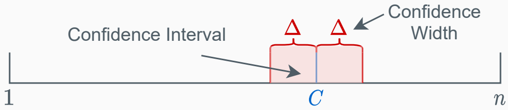


  - **Large deviations regime:** the undetected error probability decays **exponentially** fast
  - **Moderate deviations regime:** ——decays **sub-exponentially** fast

- An asymptotically optimal estimator based on test and training sequences under both regimes
- The dependence of the optimal confidence width on various parameters

Main contributions: offline single-CPD with training sequences

- The asymptotically optimal confidence width between the estimated and true change-points under



  - **Large deviations regime:** the undetected error probability decays **exponentially** fast
  - **Moderate deviations regime:** ——decays **sub-exponentially** fast

- An asymptotically optimal estimator based on test and training sequences under both regimes
- The dependence of the optimal confidence width on various parameters

H. He, Q. Zhang, and V. Y. F. Tan, "Optimal change-point detection with training sequences in the large and moderate deviations regimes", *IEEE Transactions on Information Theory*, vol. 67, no. 10, pp. 6758–6784, 2021.

Outline

1. Distributed Detection with Empirically Observed Statistics
2. Change-Point Detection with Training Sequences
3. **Information-Theoretic Generalization Error for Iterative Semi-Supervised Learning**

Semi-supervised learning (SSL) algorithms

a small amount of labelled data + a large amount of unlabelled data

Figure: An example of SSL.[1,2]

[1] Hu, Zijian, et al. Simple: similar pseudo label exploitation for semi-supervised classification. Proceedings of the IEEE/CVF Conference. (2021).
[2] Peikari, M., Salama, S., Nofech-Mozes, S. et al. A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification. Sci Rep 8, 7193 (2018).

Semi-supervised learning (SSL) algorithms

a small amount of labelled data + a large amount of unlabelled data

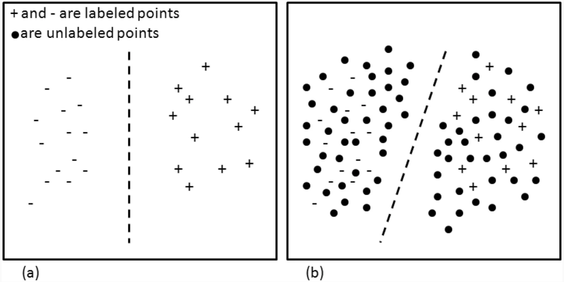
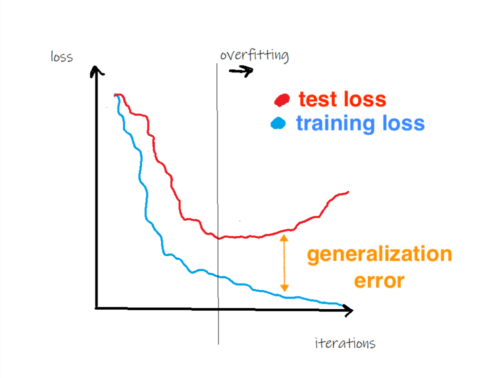


Figure: An example of SSL.[1,2]

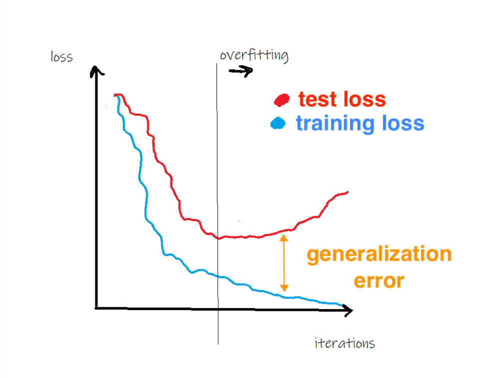
[1] Hu, Zijian, et al. Simple: similar pseudo label exploitation for semi-supervised classification. Proceedings of the IEEE/CVF Conference. (2021).
[2] Peikari, M., Salama, S., Nofech-Mozes, S. et al. A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification. Sci Rep 8, 7193 (2018).

Generalization error:



test loss=training loss+generalization error

♠ Generalization error:



test loss=training loss+generalization error

♣ Information-theoretic bound:

Theorem 3 (Bu et al. 2020)

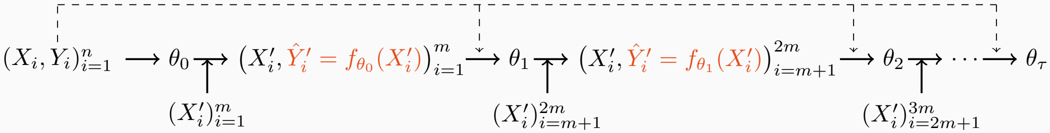
Suppose $l(\theta, Z)$ is R-sub-Gaussian under $Z \sim P_Z$ for all $\theta \in \Theta$, then

$$|\mathrm{gen}| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2R^2 I(W; Z_i)}.$$

▲ Iterative semi-supervised learning (SSL) algorithms:

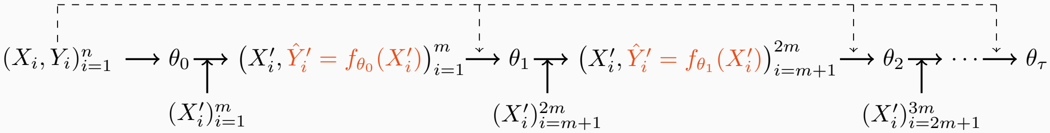
$$(X_i, Y_i)_{i=1}^{n} \longrightarrow \theta_0 \xrightarrow[(X'_i)_{i=1}^{m}]{} \left(X'_i, \hat{Y}'_i = f_{\theta_0}(X'_i)\right)_{i=1}^{m} \longrightarrow \theta_1 \xrightarrow[(X'_i)_{i=m+1}^{2m}]{} \left(X'_i, \hat{Y}'_i = f_{\theta_1}(X'_i)\right)_{i=m+1}^{2m} \longrightarrow \theta_2 \xrightarrow[(X'_i)_{i=2m+1}^{3m}]{} \cdots \longrightarrow \theta_\tau$$

▲ Iterative semi-supervised learning (SSL) algorithms:



- Labelled training dataset $S_l = \{Z_1, \ldots, Z_n\} = \{(X_i, Y_i)\}_{i=1}^n$, $X_i \overset{\text{i.i.d.}}{\sim} P_X$, Y_i is the label

▲ Iterative semi-supervised learning (SSL) algorithms:

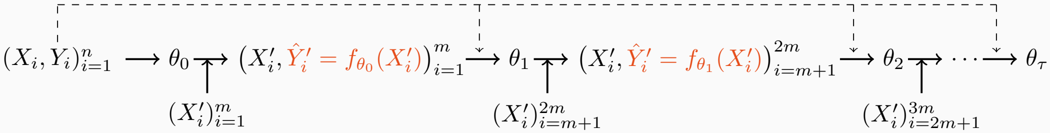


- Labelled training dataset $S_l = \{Z_1, \ldots, Z_n\} = \{(X_i, Y_i)\}_{i=1}^n$,
  $X_i \overset{\text{i.i.d.}}{\sim} P_X$, Y_i is the label
- Unlabelled training dataset $S_u = \{X'_1, \ldots, X'_{\tau m}\}$, maximum iteration $\tau \in \mathbb{N}$
  $X'_i \overset{\text{i.i.d.}}{\sim} P_X$, $m \gg n$

Information-Theoretic Gen-Error for Iterative SSL Problem Setup

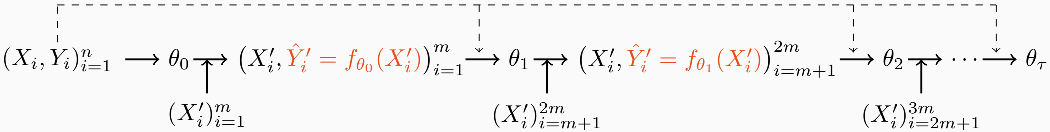
NUS National University of Singapore

▲ Iterative semi-supervised learning (SSL) algorithms:



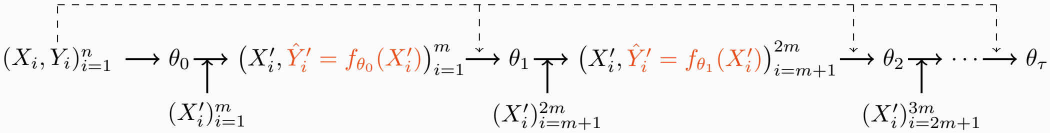
- Labelled training dataset $S_\mathrm{l} = \{Z_1, \ldots, Z_n\} = \{(X_i, Y_i)\}_{i=1}^n$,
  $X_i \overset{\text{i.i.d.}}{\sim} P_X$, Y_i is the label
- Unlabelled training dataset $S_\mathrm{u} = \{X'_1, \ldots, X'_{\tau m}\}$, maximum iteration $\tau \in \mathbb{N}$
  $X'_i \overset{\text{i.i.d.}}{\sim} P_X$, $m \gg n$
- $\{S_{\mathrm{u},t}\}_{t=1}^\tau$, where $S_{\mathrm{u},t} = \{X'_{(t-1)m+1}, \ldots, X'_{tm}\}$

Iterative semi-supervised learning (SSL) algorithms:



- Labelled training dataset $S_{\mathrm{l}} = \{Z_1, \ldots, Z_n\} = \{(X_i, Y_i)\}_{i=1}^n$, $X_i \overset{\text{i.i.d.}}{\sim} P_X$, Y_i is the label
- Unlabelled training dataset $S_{\mathrm{u}} = \{X_1', \ldots, X_{\tau m}'\}$, maximum iteration $\tau \in \mathbb{N}$ $X_i' \overset{\text{i.i.d.}}{\sim} P_X$, $m \gg n$
- $\{S_{\mathrm{u},t}\}_{t=1}^{\tau}$, where $S_{\mathrm{u},t} = \{X_{(t-1)m+1}', \ldots, X_{tm}'\}$
- Iterative pseudo-labelling: a predictor $f_{\theta_{t-1}} : \mathcal{X} \mapsto \mathcal{Y}$, $\hat{Y}_i' = f_{\theta_{t-1}}(X_i')$

Iterative semi-supervised learning (SSL) algorithms:



- Labelled training dataset $S_l = \{Z_1, \ldots, Z_n\} = \{(X_i, Y_i)\}_{i=1}^n$,
  $X_i \overset{\text{i.i.d.}}{\sim} P_X$, Y_i is the label
- Unlabelled training dataset $S_u = \{X'_1, \ldots, X'_{\tau m}\}$, maximum iteration $\tau \in \mathbb{N}$
  $X'_i \overset{\text{i.i.d.}}{\sim} P_X$, $m \gg n$
- $\{S_{u,t}\}_{t=1}^{\tau}$, where $S_{u,t} = \{X'_{(t-1)m+1}, \ldots, X'_{tm}\}$
- Iterative pseudo-labelling: a predictor $f_{\theta_{t-1}} : \mathcal{X} \mapsto \mathcal{Y}$, $\hat{Y}'_i = f_{\theta_{t-1}}(X'_i)$
  $S_{u,t} \Longrightarrow \hat{S}_{u,t} = \{(X'_i, \hat{Y}'_i)\}_{i \in \mathcal{I}_t}$, where $\mathcal{I}_t = [(t-1)m + 1 : tm]$

- Goal: minimize the *population risk*

$$L_{P_Z}(\theta_t) := \mathbb{E}_{Z\sim P_Z}[l(\theta_t, Z)].$$

- Goal: minimize the *population risk*

$$L_{P_Z}(\theta_t) := \mathbb{E}_{Z\sim P_Z}[l(\theta_t, Z)].$$

P_Z unknown $\Longrightarrow$ Goal: instead minimize the *empirical risk*

- Goal: minimize the *population risk*

$$L_{P_Z}(\theta_t) := \mathbb{E}_{Z\sim P_Z}[l(\theta_t, Z)].$$

P_Z unknown $\Longrightarrow$ Goal: instead minimize the *empirical risk* of labelled and pseudo-labelled data:

$$L_{S_1}(\theta_t) := \frac{1}{n}\sum_{i=1}^{n} l(\theta_t, Z_i), \quad L_{\hat{S}_{\mathrm{u},t}}(\theta_t) := \frac{1}{m}\sum_{i\in\mathcal{I}_t} l(\theta_t, (X_i', \hat{Y}_i')).$$

- Goal: minimize the *population risk*

$$L_{P_Z}(\theta_t) := \mathbb{E}_{Z\sim P_Z}[l(\theta_t, Z)].$$

P_Z unknown $\Longrightarrow$ Goal: instead minimize the *empirical risk* of labelled and pseudo-labelled data:

$$L_{S_1}(\theta_t) := \frac{1}{n}\sum_{i=1}^{n} l(\theta_t, Z_i), \quad L_{\hat{S}_{\mathrm{u},t}}(\theta_t) := \frac{1}{m}\sum_{i\in\mathcal{I}_t} l(\theta_t, (X_i', \hat{Y}_i')).$$

Total empirical risk: $w = \frac{n}{n+m}$

$$\begin{aligned} L_{S_1,\hat{S}_{\mathrm{u},t}}(\theta_t) &:= wL_{S_1}(\theta_t) + (1-w)L_{\hat{S}_{\mathrm{u},t}}(\theta_t) \\ &= \frac{1}{n+m}\left(\sum_{i=1}^{n} l(\theta_t, Z_i) + \sum_{i\in\mathcal{I}_t} l(\theta_t, (X_i', \hat{Y}_i'))\right). \end{aligned}$$

Generalization error at the t-th iteration: the expected gap between the population risk of θ_t and the empirical risk on the training data

$$\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_1,S_u}\}_{k=0}^{t}, \{f_{\theta_k}\}_{k=0}^{t-1}) := \mathbb{E}[L_{P_Z}(\theta_t) - L_{S_1,\hat{S}_{u,t}}(\theta_t)]$$

$$= w\left(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z) \mid \theta_t]] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta_t, Z_i}[l(\theta_t, Z_i)]\right)$$

$$+ (1-w)\left(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z) \mid \theta_t]] - \frac{1}{m}\sum_{i\in\mathcal{I}_t}\mathbb{E}_{\theta_t, X_i', \hat{Y}_i'}[l(\theta_t, (X_i', \hat{Y}_i'))]\right).$$

Generalization error at the t-th iteration: the expected gap between the population risk of θ_t and the empirical risk on the training data

$$\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_1,S_u}\}_{k=0}^{t}, \{f_{\theta_k}\}_{k=0}^{t-1}) := \mathbb{E}[L_{P_Z}(\theta_t) - L_{S_1,\hat{S}_{u,t}}(\theta_t)]$$

$$= w\left(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z) \mid \theta_t]] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta_t, Z_i}[l(\theta_t, Z_i)]\right)$$

$$+ (1-w)\left(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z) \mid \theta_t]] - \frac{1}{m}\sum_{i \in \mathcal{I}_t}\mathbb{E}_{\theta_t, X_i', \hat{Y}_i'}[l(\theta_t, (X_i', \hat{Y}_i'))]\right).$$

- gap for the labelled training data

NUS National University of Singapore

Generalization error at the t-th iteration: the expected gap between the population risk of θ_t and the empirical risk on the training data

$$\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_1,S_u}\}_{k=0}^{t}, \{f_{\theta_k}\}_{k=0}^{t-1}) := \mathbb{E}[L_{P_Z}(\theta_t) - L_{S_1,\hat{S}_{u,t}}(\theta_t)]$$

$$= w\left(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z) \mid \theta_t]] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta_t, Z_i}[l(\theta_t, Z_i)]\right)$$

$$+ (1-w)\left(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z) \mid \theta_t]] - \frac{1}{m}\sum_{i\in\mathcal{I}_t} \mathbb{E}_{\theta_t, X_i', \hat{Y}_i'}[l(\theta_t, (X_i', \hat{Y}_i'))]\right).$$

- gap for the labelled training data
- gap for the pseudo-labelled training data

Generalization error at the t-th iteration: the expected gap between the population risk of θ_t and the empirical risk on the training data

$$\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_1,S_u}\}_{k=0}^{t}, \{f_{\theta_k}\}_{k=0}^{t-1}) := \mathbb{E}[L_{P_Z}(\theta_t) - L_{S_1,\hat{S}_{u,t}}(\theta_t)]$$

$$= w\left(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z) \mid \theta_t]] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta_t, Z_i}[l(\theta_t, Z_i)]\right)$$

$$+ (1-w)\left(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z) \mid \theta_t]] - \frac{1}{m}\sum_{i\in\mathcal{I}_t}\mathbb{E}_{\theta_t, X_i', \hat{Y}_i'}[l(\theta_t, (X_i', \hat{Y}_i'))]\right).$$

- gap for the labelled training data
- gap for the pseudo-labelled training data

Questions

★ How does gen_t evolve as the iteration count t increases?

★ Do the unlabelled data examples in S_u help to improve the generalization error?

Theorem 1.A (Gen-error upper bound for iterative SSL)

Suppose $l(\theta, Z) \sim \text{subG}(R)$ under $Z \sim P_Z$ for all $\theta \in \Theta$, then for any $t \in [0:\tau]$,

$$|\text{gen}_t| \leq \frac{w}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta^{(t-1)}}\left[\sqrt{2R^2 I_{\theta^{(t-1)}}(\theta_t; Z_i)}\right]$$
$$+\frac{1-w}{m}\sum_{i=(t-1)m+1}^{tm}\mathbb{E}_{\theta^{(t-1)}}\left[\sqrt{2R^2\left(I_{\theta^{(t-1)}}(\theta_t; X_i', \hat{Y}_i') + D_{\theta^{(t-1)}}(P_{X_i', \hat{Y}_i'} \| P_Z)\right)}\right].$$

General Results for Generalization Error

Theorem 1.A (Gen-error upper bound for iterative SSL)

Suppose $l(\theta, Z) \sim \mathrm{subG}(R)$ under $Z \sim P_Z$ for all $\theta \in \Theta$, then for any $t \in [0:\tau]$,

$$|\mathrm{gen}_t| \leq \frac{w}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta^{(t-1)}} \left[\sqrt{2R^2 I_{\theta^{(t-1)}}(\theta_t; Z_i)} \right] + \frac{1-w}{m} \sum_{i=(t-1)m+1}^{tm} \mathbb{E}_{\theta^{(t-1)}} \left[\sqrt{2R^2 \left(I_{\theta^{(t-1)}}(\theta_t; X_i', \hat{Y}_i') + D_{\theta^{(t-1)}}(P_{X_i', \hat{Y}_i'} \| P_Z) \right)} \right].$$

- The term depends on the labelled training data.

General Results for Generalization Error

Theorem 1.A (Gen-error upper bound for iterative SSL)

Suppose $l(\theta, Z) \sim \mathrm{subG}(R)$ under $Z \sim P_Z$ for all $\theta \in \Theta$, then for any $t \in [0:\tau]$,

$$|\mathrm{gen}_t| \leq \frac{w}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t-1)}}\left[\sqrt{2R^2 I_{\theta^{(t-1)}}(\theta_t; Z_i)}\right] + \frac{1-w}{m}\sum_{i=(t-1)m+1}^{tm} \mathbb{E}_{\theta^{(t-1)}}\left[\sqrt{2R^2\left(I_{\theta^{(t-1)}}(\theta_t; X_i', \hat{Y}_i') + D_{\theta^{(t-1)}}(P_{X_i', \hat{Y}_i'} \| P_Z)\right)}\right].$$

- The term depends on the labelled training data.
- The term depends on the pseudo-labelled training data.

General Results for Generalization Error

Theorem 1.A (Gen-error upper bound for iterative SSL)

Suppose $l(\theta, Z) \sim \text{subG}(R)$ under $Z \sim P_Z$ for all $\theta \in \Theta$, then for any $t \in [0:\tau]$,

$$|\text{gen}_t| \le \frac{w}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta^{(t-1)}} \left[\sqrt{2R^2 I_{\theta^{(t-1)}}(\theta_t; Z_i)} \right]$$

$$+ \frac{1-w}{m} \sum_{i=(t-1)m+1}^{tm} \mathbb{E}_{\theta^{(t-1)}} \left[\sqrt{2R^2 \left(I_{\theta^{(t-1)}}(\theta_t; X_i', \hat{Y}_i') + D_{\theta^{(t-1)}}(P_{X_i', \hat{Y}_i'} \| P_Z) \right)} \right].$$

- The term depends on the labelled training data.
- The term depends on the pseudo-labelled training data. The divergence is caused by pseudo-labelling.

General Results for Generalization Error

Theorem 1.A (Gen-error upper bound for iterative SSL)

Suppose $l(\theta, Z) \sim \mathrm{subG}(R)$ under $Z \sim P_Z$ for all $\theta \in \Theta$, then for any $t \in [0:\tau]$,

$$|\mathrm{gen}_t| \leq \frac{w}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta^{(t-1)}} \left[\sqrt{2R^2 I_{\theta^{(t-1)}}(\theta_t; Z_i)} \right] + \frac{1-w}{m} \sum_{i=(t-1)m+1}^{tm} \mathbb{E}_{\theta^{(t-1)}} \left[\sqrt{2R^2 \left(I_{\theta^{(t-1)}}(\theta_t; X'_i, \hat{Y}'_i) + D_{\theta^{(t-1)}}(P_{X'_i, \hat{Y}'_i} \| P_Z) \right)} \right].$$

- The term depends on the labelled training data.
- The term depends on the pseudo-labelled training data. The divergence is caused by pseudo-labelling.

♣ Follows from Bu et al. (2020, Theorem 1) and Wu et al. (2020, Theorem 1)

Theorem 1.B (EXACT gen-error for iterative SSL)

Consider the NLL loss function $l(\theta, Z) = -\log p_\theta(Z)$, where $p_\theta(Z)$ is the likelihood of Z under parameter θ. For any $t \in [0:\tau]$,

$$\mathrm{gen}_t = \mathbb{E}_{\theta(t)}\left[\frac{w}{n}\sum_{i=1}^{n}\Delta\mathrm{h}_{\theta_t}^{(i)} + \frac{1-w}{m}\sum_{i\in\mathcal{I}_t}\left(\Delta\mathrm{h}_{\theta(t)}^{\prime(i)} + \widetilde{\Delta\mathrm{h}}_{\theta(t)}^{\prime(i)}\right)\right],$$

where $\Delta\mathrm{h}_{\theta_t}^{(i)} := \Delta\mathrm{h}(P_Z \| P_{Z_i|\theta_t} | p_{\theta_t})$, $\Delta\mathrm{h}_{\theta(t)}^{\prime(i)} := \Delta\mathrm{h}(P_Z \| P_{X_i', \hat{Y}_i'|\theta(t-1)} | p_{\theta_t})$, and $\widetilde{\Delta\mathrm{h}}_{\theta(t)}^{\prime(i)} := \Delta\mathrm{h}(P_{X_i', \hat{Y}_i'|\theta(t-1)} \| P_{X_i', \hat{Y}_i'|\theta(t)} | p_{\theta_t})$. (i.e., cross-entropies)

Theorem 1.B (EXACT gen-error for iterative SSL)

Consider the NLL loss function $l(\theta, Z) = -\log p_\theta(Z)$, where $p_\theta(Z)$ is the likelihood of Z under parameter θ. For any $t \in [0:\tau]$,

$$\mathrm{gen}_t = \mathbb{E}_{\theta(t)}\left[\frac{w}{n}\sum_{i=1}^{n}\Delta\mathrm{h}_{\theta_t}^{(i)} + \frac{1-w}{m}\sum_{i\in\mathcal{I}_t}\left(\Delta\mathrm{h}'^{(i)}_{\theta(t)} + \widetilde{\Delta\mathrm{h}}'^{(i)}_{\theta(t)}\right)\right],$$

where $\Delta\mathrm{h}_{\theta_t}^{(i)} := \Delta\mathrm{h}(P_Z \| P_{Z_i|\theta_t} | p_{\theta_t})$, $\Delta\mathrm{h}'^{(i)}_{\theta(t)} := \Delta\mathrm{h}(P_Z \| P_{X_i', \hat{Y}_i' | \theta(t-1)} | p_{\theta_t})$, and $\widetilde{\Delta\mathrm{h}}'^{(i)}_{\theta(t)} := \Delta\mathrm{h}(P_{X_i', \hat{Y}_i' | \theta(t-1)} \| P_{X_i', \hat{Y}_i' | \theta(t)} | p_{\theta_t})$. (i.e., cross-entropies)

- The term depends on the labelled training data.

General Results for Generalization Error

Theorem 1.B (EXACT gen-error for iterative SSL)

Consider the NLL loss function $l(\theta, Z) = -\log p_\theta(Z)$, where $p_\theta(Z)$ is the likelihood of Z under parameter θ. For any $t \in [0:\tau]$,

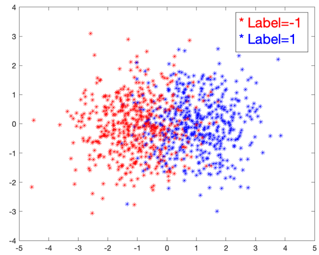
$$\text{gen}_t = \mathbb{E}_{\theta(t)}\left[\frac{w}{n}\sum_{i=1}^{n}\Delta\text{h}_{\theta_t}^{(i)} + \frac{1-w}{m}\sum_{i\in\mathcal{I}_t}\left(\Delta\text{h}'^{(i)}_{\theta(t)} + \widetilde{\Delta\text{h}}'^{(i)}_{\theta(t)}\right)\right],$$

where $\Delta\text{h}_{\theta_t}^{(i)} := \Delta\text{h}(P_Z \| P_{Z_i|\theta_t} | p_{\theta_t})$, $\Delta\text{h}'^{(i)}_{\theta(t)} := \Delta\text{h}(P_Z \| P_{X_i',\hat{Y}_i'|\theta(t-1)} | p_{\theta_t})$, and $\widetilde{\Delta\text{h}}'^{(i)}_{\theta(t)} := \Delta\text{h}(P_{X_i',\hat{Y}_i'|\theta(t-1)} \| P_{X_i',\hat{Y}_i'|\theta(t)} | p_{\theta_t})$. (i.e., cross-entropies)

- The term depends on the labelled training data.
- The term depends on the pseudo-labelled training data. The divergence is caused by pseudo-labelling.

Main Results on Binary Gaussian Mixture Model

♠ **Iterative SSL under bGMM**: Under the bGMM with mean $\boldsymbol{\mu}$ and standard deviation σ (bGMM($\boldsymbol{\mu}, \sigma$)), assume $\mathcal{Y} = \{-1, +1\}$, $Y \sim P_Y = \mathrm{unif}\{-1, +1\}$, and $X|Y \sim \mathcal{N}(Y\boldsymbol{\mu}, \sigma^2 \mathbf{I}_d)$



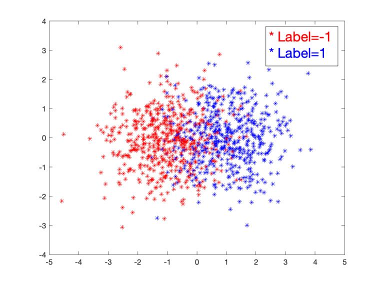
Main Results on Binary Gaussian Mixture Model

♠ **Iterative SSL under bGMM**: Under the bGMM with mean μ and standard deviation σ (bGMM(μ, σ)), assume $\mathcal{Y} = \{-1, +1\}$, $Y \sim P_Y = \mathrm{unif}\{-1, +1\}$, and $X|Y \sim \mathcal{N}(Y\mu, \sigma^2 \mathbf{I}_d)$

NLL loss:

$$l(\theta, (X, Y)) = -\log p_\theta(X, Y)$$

$$= -\log \frac{1}{2\sqrt{(2\pi)^d}\sigma^d} + \frac{1}{2\sigma^2}(X - Y\theta)^\top (X - Y\theta)$$

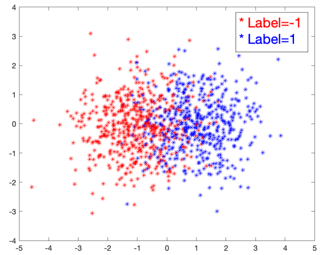


Main Results on Binary Gaussian Mixture Model

♠ **Iterative SSL under bGMM**: Under the bGMM with mean μ and standard deviation σ (bGMM(μ, σ)), assume $\mathcal{Y} = \{-1, +1\}, Y \sim P_Y = \mathrm{unif}\{-1, +1\}$, and $X|Y \sim \mathcal{N}(Y\mu, \sigma^2 \mathbf{I}_d)$

NLL loss:

$$l(\theta, (X, Y)) = -\log p_\theta(X, Y)$$
$$= -\log \frac{1}{2\sqrt{(2\pi)^d}\sigma^d} + \frac{1}{2\sigma^2}(X - Y\theta)^\top (X - Y\theta)$$



Pseudo-labelling function: for any $t \in [0 : \tau]$,

$$\hat{Y}_i' = f_{\theta_{t-1}}(X_i') = \mathrm{sgn}(\theta_{t-1}^\top X_i')$$

- **Step 1: Initial round $t=0$ with S_1:** Estimate θ using labelled dataset $S_1=\{(X_i,Y_i)\}_{i=1}^n$, i.e.,

$$\theta_0=\frac{1}{n}\sum_{i=1}^n Y_iX_i.$$

- **Step 1: Initial round $t = 0$ with S_l**: Estimate θ using labelled dataset $S_l = \{(X_i, Y_i)\}_{i=1}^n$, i.e.,

$$\theta_0 = \frac{1}{n}\sum_{i=1}^n Y_i X_i.$$

- **Step 2: Pseudo-label each unlabelled data using previous parameter θ_{t-1}**: At each $t \in [1:\tau]$, for any $i \in \mathcal{I}_t$,

$$\hat{Y}_i' = \mathrm{sgn}(\theta_{t-1}^\top X_i').$$

- **Step 1: Initial round $t=0$ with S_1:** Estimate θ using labelled dataset $S_1 = \{(X_i, Y_i)\}_{i=1}^n$, i.e.,

$$\theta_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i.$$

- **Step 2: Pseudo-label each unlabelled data using previous parameter θ_{t-1}:** At each $t \in [1:\tau]$, for any $i \in \mathcal{I}_t$,

$$\hat{Y}_i' = \mathrm{sgn}(\theta_{t-1}^\top X_i').$$

- **Step 3: Refine the model:** Estimate new parameter using augmented dataset $S_1 \cup \{(X_i', \hat{Y}_i')\}_{i \in \mathcal{I}_t}$, i.e.,

$$\theta_t = \frac{1}{n+m}\left(\sum_{i=1}^{n} Y_i X_i + \sum_{i \in \mathcal{I}_t} \hat{Y}_i' X_i'\right)$$

If $t < \tau$, go back to Step 2.

Main Results on Binary Gaussian Mixture Model (Continued)

Theorem 2 (Exact gen-error for iterative SSL under bGMM)

We derived exact characterization of gen-error gen_t for iterative SSL under bGMM as a function of standard deviation σ when the number of unlabelled data is large enough.

⋇ Refer to the thesis for the full theorem.

Main Results on Binary Gaussian Mixture Model (Continued)

Theorem 2 (Exact gen-error for iterative SSL under bGMM)

We derived exact characterization of gen-error gen_t for iterative SSL under bGMM as a function of standard deviation σ when the number of unlabelled data is large enough.

⁂ Refer to the thesis for the full theorem.

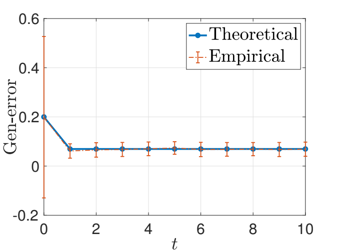


Fig.1.1 $\sigma = 0.6$

Main Results on Binary Gaussian Mixture Model (Continued)

Theorem 2 (Exact gen-error for iterative SSL under bGMM)

We derived exact characterization of gen-error gen_t for iterative SSL under bGMM as a function of standard deviation σ when the number of unlabelled data is large enough.

⋇ Refer to the thesis for the full theorem.

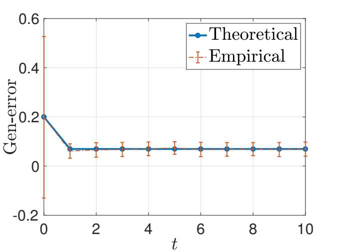


Fig.1.1 $\sigma = 0.6$

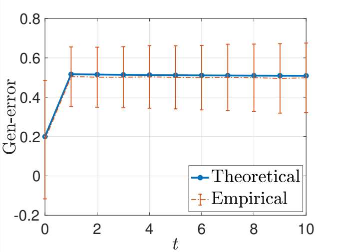


Fig.1.2 $\sigma = 3$

Theorem 2 (Exact gen-error for iterative SSL under bGMM)

We derived exact characterization of gen-error gen_t for iterative SSL under bGMM as a function of standard deviation σ when the number of unlabelled data is large enough.

⋇ Refer to the thesis for the full theorem.

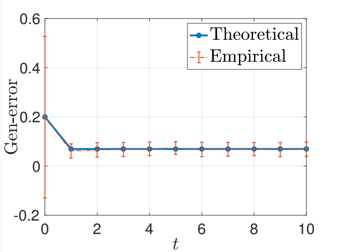


Fig.1.1 $\sigma = 0.6$

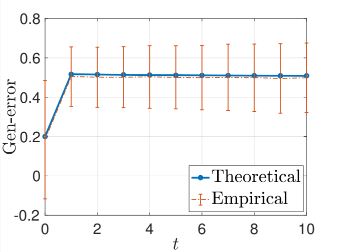


Fig.1.2 $\sigma = 3$

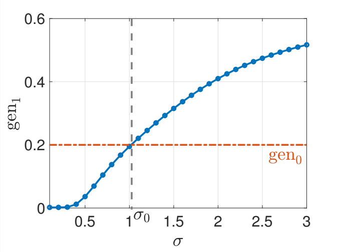


Fig.1.3 Gen-error at $t = 1$ vs σ

Main Results on Binary Gaussian Mixture Model (Continued)

To mitigate the undesirable increase of gen-error across the pseudo-labelling iterations, we prove that adding l_2-regularization (add $\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ to loss function) to the loss function can help.

To mitigate the undesirable increase of gen-error across the pseudo-labelling iterations, we prove that adding l_2-regularization (add $\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ to loss function) to the loss function can help.

Theorem 4 (Gen-error with regularization)

Fix any $d \in \mathbb{N}$, and $\sigma, \lambda \in \mathbb{R}_+$. The gen-error at any $t \in [1:\tau]$ is

$$\mathrm{gen}_t^{\mathrm{reg}} = \frac{\mathrm{gen}_t}{1+\sigma^2\lambda}.$$

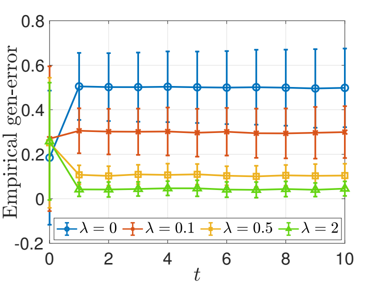


Fig.2.1 Gen-error for $\sigma = 3$, different λ

To mitigate the undesirable increase of gen-error across the pseudo-labelling iterations, we prove that adding l_2-regularization (add $\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ to loss function) to the loss function can help.

Theorem 4 (Gen-error with regularization)

Fix any $d \in \mathbb{N}$, and $\sigma, \lambda \in \mathbb{R}_+$. The gen-error at any $t \in [1:\tau]$ is

$$\mathrm{gen}_t^{\mathrm{reg}} = \frac{\mathrm{gen}_t}{1+\sigma^2\lambda}.$$

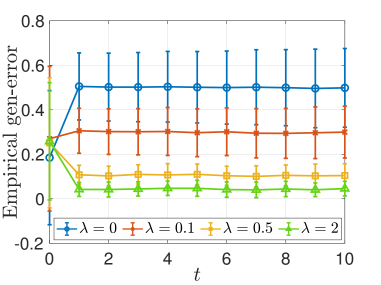


Fig.2.1 Gen-error for $\sigma = 3$, different λ

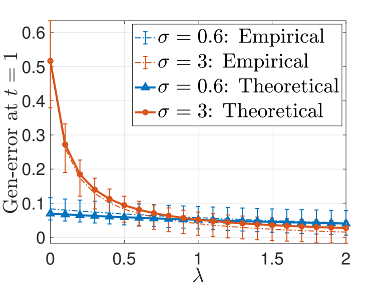


Fig.2.2 Gen-error at $t = 1$ versus λ

Experiments on CIFAR-10 and MNIST

Easy-to-distinguish pairs: ”horse-ship” & “automobile-truck” , and multi-class (Repeat 10 times)

Experiments on CIFAR-10 and MNIST

NUS National University of Singapore

Easy-to-distinguish pairs: "horse-ship" & "automobile-truck" , and multi-class (Repeat 10 times)

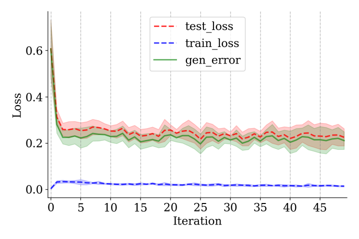


Fig.3.1 "horse-ship": gen-error

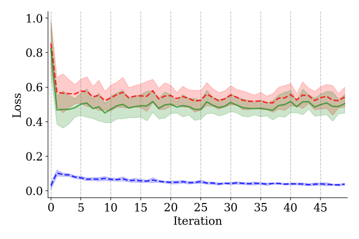


Fig.3.2 "automobile-truck": gen-error

Experiments on CIFAR-10 and MNIST

Easy-to-distinguish pairs: ”horse-ship” & “automobile-truck” , and multi-class (Repeat 10 times)

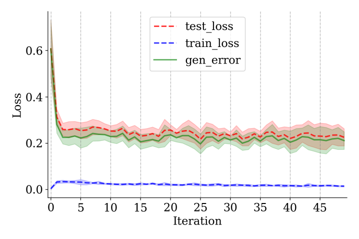


Fig.3.1 “horse-ship”: gen-error

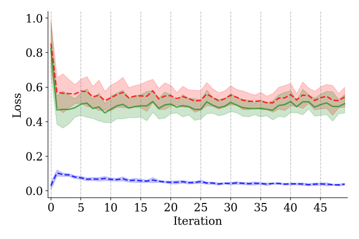


Fig.3.2 “automobile-truck”: gen-error

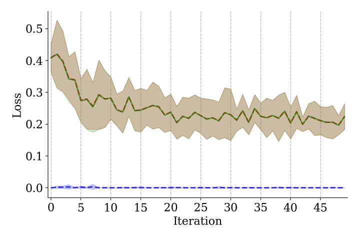


Fig.3.3 MNIST: gen-error

Difficult-to-distinguish pairs: "cat-dog" (Repeat 10 times)

Fig.4.1 "cat-dog": gen-error

Difficult-to-distinguish pairs: “cat-dog” (Repeat 10 times)

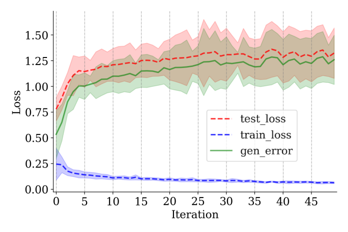


Fig.4.1 “cat-dog”: gen-error

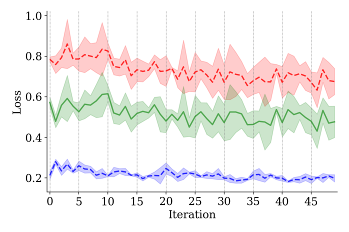


Fig.4.2 “cat-dog”: gen-error with weight decay 0.0005

Experiments on CIFAR-10 and MNIST

Difficult-to-distinguish pairs: "cat-dog" (Repeat 10 times)

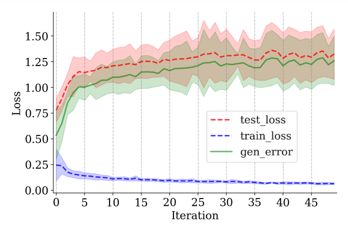


Fig.4.1 "cat-dog": gen-error

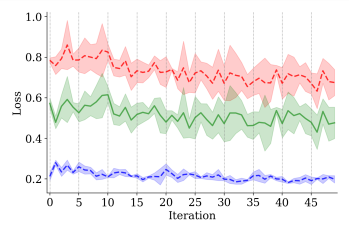


Fig.4.2 "cat-dog": gen-error with weight decay 0.0005

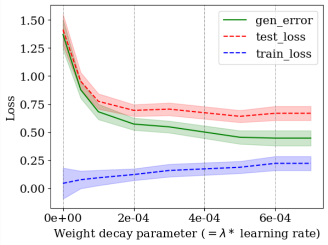


Fig.4.3 "cat-dog": gen-error after convergence versus weight decay

★ Problem setup: SSL with iterative pseudo-labelling

★ **Problem setup:** SSL with iterative pseudo-labelling

★ **Main contributions:**

Answers to previous questions

★ *How does gen_t evolve as the iteration count t increases?*

Information-Theoretic Gen-Error for Iterative SSL Summary

★ **Problem setup:** SSL with iterative pseudo-labelling

★ **Main contributions:**

Answers to previous questions

★ *How does gen_t evolve as the iteration count t increases?*

♠ Exact information-theoretic characterization for gen-error across the iterations. First decreases when the class-overlap is small (or increases when the class-overlap is large) and then converges rapidly.

★ **Problem setup:** SSL with iterative pseudo-labelling

★ **Main contributions:**

Answers to previous questions

★ *How does gen_t evolve as the iteration count t increases?*

♠ Exact information-theoretic characterization for gen-error across the iterations. First decreases when the class-overlap is small (or increases when the class-overlap is large) and then converges rapidly.

★ *Do the unlabelled data examples in S_{u} help to improve the generalization error?*

★ **Problem setup:** SSL with iterative pseudo-labelling

★ **Main contributions:**

Answers to previous questions

★ *How does gen_t evolve as the iteration count t increases?*

♠ Exact information-theoretic characterization for gen-error across the iterations. First decreases when the class-overlap is small (or increases when the class-overlap is large) and then converges rapidly.

★ *Do the unlabelled data examples in S_{u} help to improve the generalization error?*

♠ Specialize to bGMM case: for large data variance, the unlabelled data DO NOT help, but adding l_2 regularization can help to improve.

★ **Problem setup:** SSL with iterative pseudo-labelling

★ **Main contributions:**

Answers to previous questions

★ *How does gen_t evolve as the iteration count t increases?*

♠ Exact information-theoretic characterization for gen-error across the iterations. First decreases when the class-overlap is small (or increases when the class-overlap is large) and then converges rapidly.

★ *Do the unlabelled data examples in S_{u} help to improve the generalization error?*

- ♠ Specialize to bGMM case: for large data variance, the unlabelled data DO NOT help, but adding l_2 regularization can help to improve.
- ♠ Extensive experiments on CIFAR-10 and MNIST: corroborate theoretical results on bGMM.

★ Problem setup: SSL with iterative pseudo-labelling

★ Main contributions:

Answers to previous questions

★ *How does gen_t evolve as the iteration count t increases?*

♠ Exact information-theoretic characterization for gen-error across the iterations. First decreases when the class-overlap is small (or increases when the class-overlap is large) and then converges rapidly.

★ *Do the unlabelled data examples in S_u help to improve the generalization error?*

- ♠ Specialize to bGMM case: for large data variance, the unlabelled data DO NOT help, but adding l_2 regularization can help to improve.
- ♠ Extensive experiments on CIFAR-10 and MNIST: corroborate theoretical results on bGMM.

H. He, H. Yan, and V. Y. F. Tan, "Information-Theoretic Characterization of the Generalization Error for Iterative Semi-Supervised Learning", *Journal of Machine Learning Research* (accepted with minor revisions), 2022+

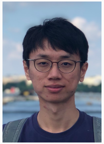
Acknowledgements

Thesis advisor:
Prof. Vincent Tan

Acknowledgements

Co-authors:

Dr. Lin Zhou



Dr. Qiaosheng Zhang

Hanshu Yan

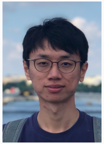
Thesis advisor:
Prof. Vincent Tan

All collaborators

Acknowledgements

Co-authors:

Dr. Lin Zhou



Dr. Qiaosheng Zhang

Hanshu Yan

Thesis advisor:
Prof. Vincent Tan

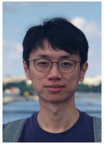
All collaborators

Thesis committee

Acknowledgements

Co-authors:

Dr. Lin Zhou



Dr. Qiaosheng Zhang

Hanshu Yan

All collaborators

Thesis committee

Thesis advisor:
Prof. Vincent Tan

SG BUAA Alumni Assoc.

Labmates in E4-06-12

Dear friends
&
My parents


THANKS FOR LISTENING
ANY QUESTIONS?