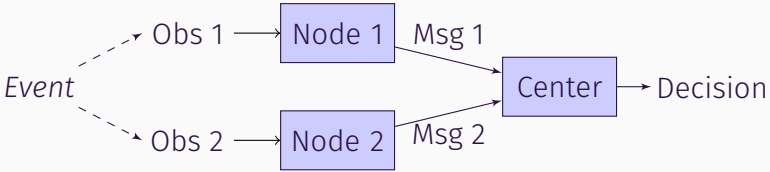
An Information Theoretic Framework for Distributed Learning Algorithms

Xiangxiang Xu Shao-Lun Huang

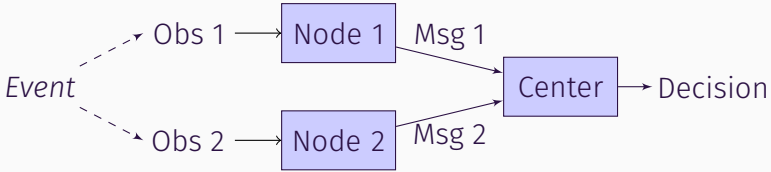
ISIT 2021

Tsinghua–Berkeley Shenzhen Institute, Tsinghua University

Background: Distributed Learning

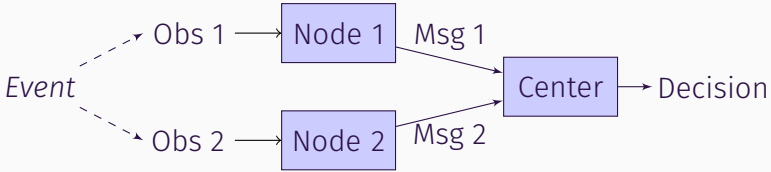


Background: Distributed Learning



Related: Federated Learning, Sensor Fusion, ...

Background: Distributed Learning

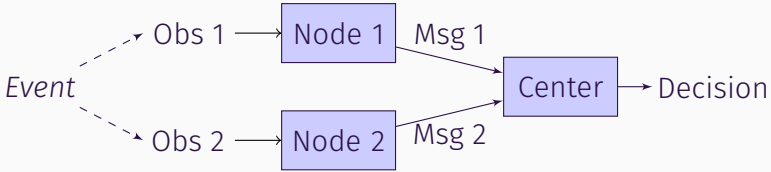


Related: Federated Learning, Sensor Fusion, ...

Information-Theoretic Characterizations

- Multiterminal Statistical Inference/Estimation (Ahlswede, Csiszár, Han, ...)

Background: Distributed Learning

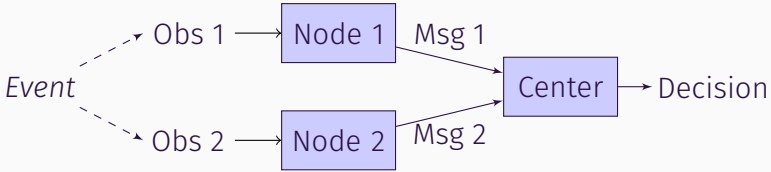


Related: Federated Learning, Sensor Fusion, ...

Information-Theoretic Characterizations

- Multiterminal Statistical Inference/Estimation (Ahlswede, Csiszár, Han, ...)
- CEO problem (Berger et al.)

Background: Distributed Learning



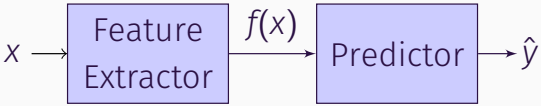
Related: Federated Learning, Sensor Fusion, ...

Information-Theoretic Characterizations

- Multiterminal Statistical Inference/Estimation (Ahlswede, Csiszár, Han, ...)
- CEO problem (Berger et al.)
- Multi-user information theory ...

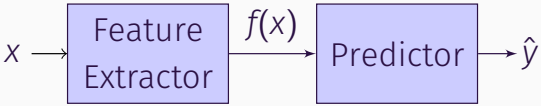
Use data X to predict label Y

Use data X to predict label Y



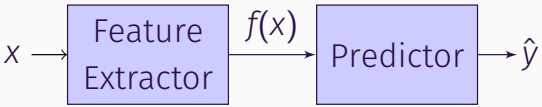
- Extract low-dim feature for prediction

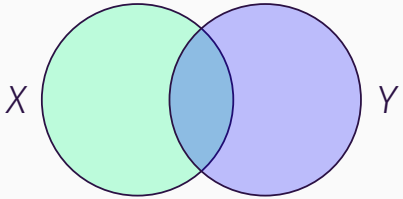
Use data X to predict label Y



- Extract low-dim feature for prediction
- f should be "informative" about Y

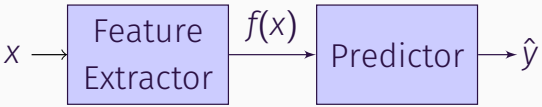
Use data X to predict label Y

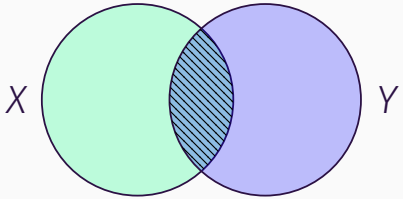




- Extract low-dim feature for prediction
- f should be "informative" about Y

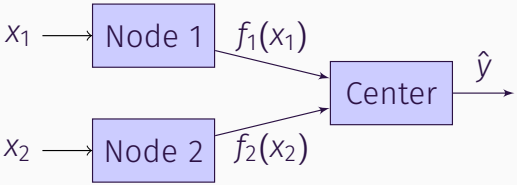
Use data X to predict label Y





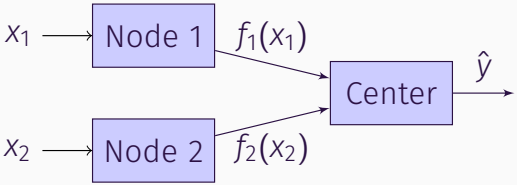
- Extract low-dim feature for prediction
- f should be "informative" about Y

Prediction by Distributed Nodes



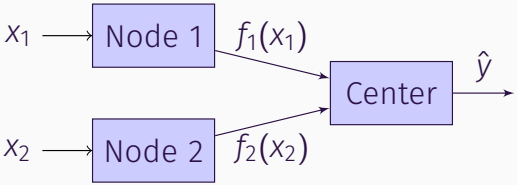
- Feature extraction required due to communication constraints

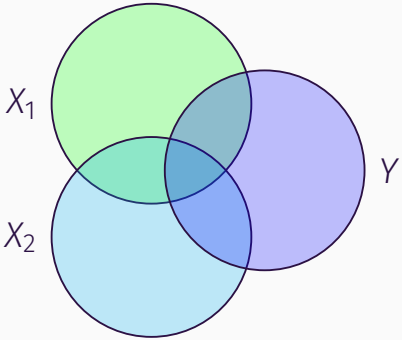
Prediction by Distributed Nodes



- Feature extraction required due to communication constraints

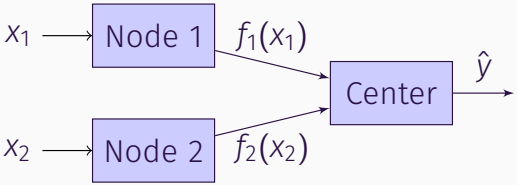
Prediction by Distributed Nodes

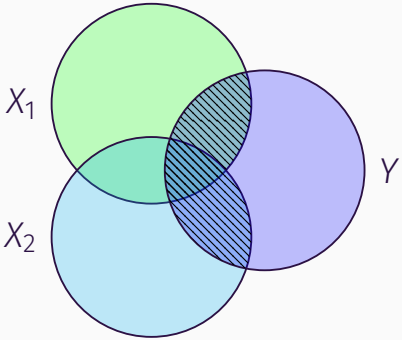




- Feature extraction required due to communication constraints
- (f_1, f_2) should be "informative" for prediction

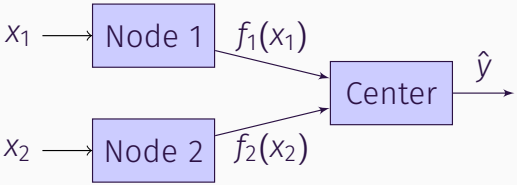
Prediction by Distributed Nodes



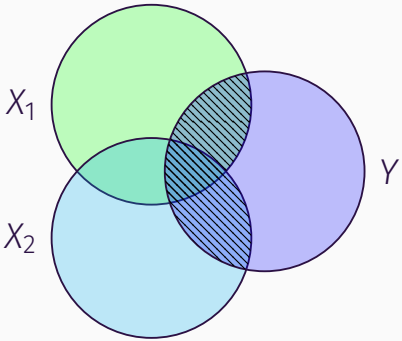


- Feature extraction required due to communication constraints
- (f_1, f_2) should be "informative" for prediction

Prediction by Distributed Nodes

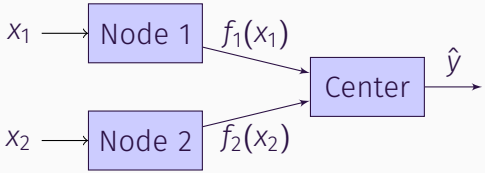


- Feature extraction required due to communication constraints
- (f_1, f_2) should be "informative" for prediction

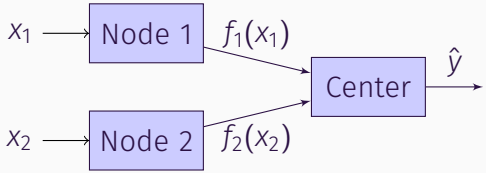


- Overlap between distributed nodes

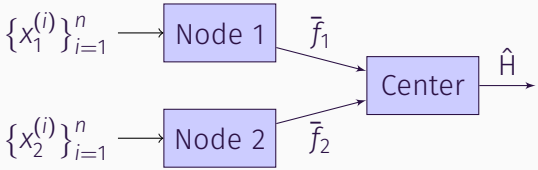
Distributed Classification



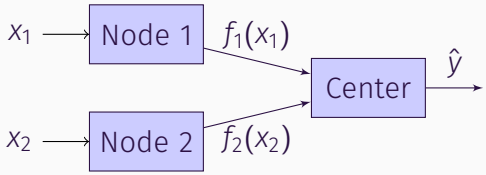
Distributed Classification



Distributed Hypothesis Testing

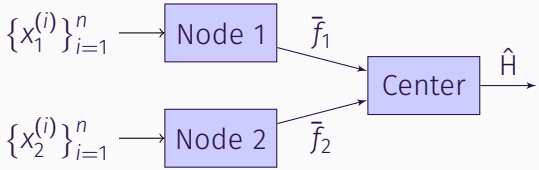


Distributed Classification



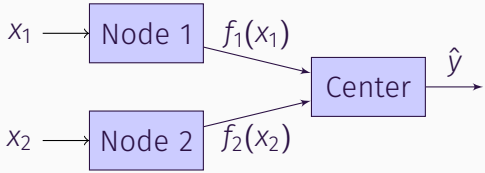
- Binary label $Y \in \{0, 1\}$

Distributed Hypothesis Testing



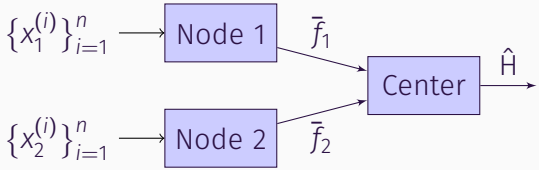
- Binary hypothesis $H \in \{0, 1\}$

Distributed Classification



- Binary label $Y \in \{0, 1\}$
- Sample pairs $\sim P_{\underline{X}|Y=y}$

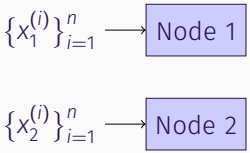
Distributed Hypothesis Testing



- Binary hypothesis $H \in \{0, 1\}$
- Sample pairs $\overset{\text{i.i.d.}}{\sim} P_{\underline{X}}^{(H)}$

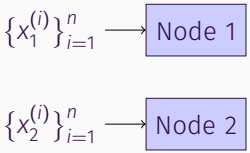
$$\underline{X} \triangleq \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

Hypothesis Testing with Distributed Nodes



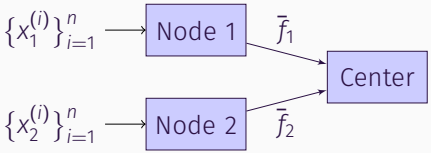
- (X_1, X_2) samples i.i.d. drawn from $P_{\underline{X}}^{(\mathrm{H})}$

Hypothesis Testing with Distributed Nodes



- (X_1, X_2) samples i.i.d. drawn from $P_{\underline{X}}^{(\mathrm{H})}$
- Node 1 observes only X_1 samples
- Node 2 observes only X_2 samples

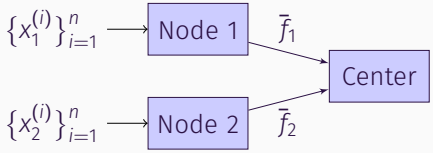
Hypothesis Testing with Distributed Nodes



- (X_1, X_2) samples i.i.d. drawn from $P_{\underline{X}}^{(\mathrm{H})}$
- Node 1 observes only X_1 samples
- Node 2 observes only X_2 samples
- Each node k sends center the statistic

$$\bar{f}_k \triangleq \frac{f_k(x_k^{(1)}) + \cdots + f_k(x_k^{(n)})}{n}$$

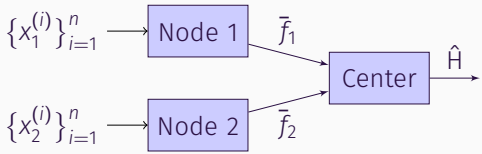
Hypothesis Testing with Distributed Nodes



- (X_1, X_2) samples i.i.d. drawn from $P_{\underline{X}}^{(\mathrm{H})}$
- Node 1 observes only X_1 samples
- Node 2 observes only X_2 samples
- Each node k sends center the statistic

$$\bar{f}_k \triangleq \frac{f_k(x_k^{(1)}) + \cdots + f_k(x_k^{(n)})}{n}$$

- *zero-rate* communication

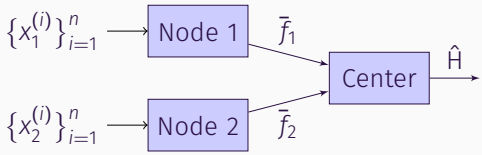


- (X_1, X_2) samples i.i.d. drawn from $P_{\underline{X}}^{(\mathsf{H})}$
- Node 1 observes only X_1 samples
- Node 2 observes only X_2 samples
- Each node k sends center the statistic

$$\bar{f}_k \triangleq \frac{f_k(x_k^{(1)}) + \cdots + f_k(x_k^{(n)})}{n}$$

- The center decides $\hat{\mathsf{H}}$ by MAP:

$$\frac{\mathbb{P}\left\{\bar{f}_1, \bar{f}_2 \middle| \mathsf{H}=1\right\}}{\mathbb{P}\left\{\bar{f}_1, \bar{f}_2 \middle| \mathsf{H}=0\right\}} \underset{\hat{\mathsf{H}}=0}{\overset{\hat{\mathsf{H}}=1}{\gtrless}} \frac{P_{\mathsf{H}}(0)}{P_{\mathsf{H}}(1)}$$



Error exponent E of the decision:

$$\mathbb{P}\left\{\hat{\mathsf{H}} \neq \mathsf{H}\right\} \doteq \exp(-nE)$$

Design Goal: maximize E

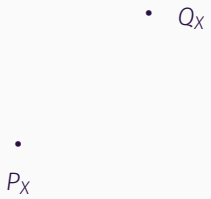
- (X_1, X_2) samples i.i.d. drawn from $P_{\underline{X}}^{(\mathsf{H})}$
- Node 1 observes only X_1 samples
- Node 2 observes only X_2 samples
- Each node k sends center the statistic

$$\bar{f}_k \triangleq \frac{f_k(x_k^{(1)}) + \cdots + f_k(x_k^{(n)})}{n}$$

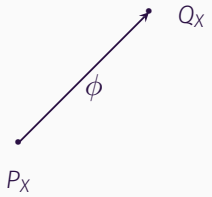
- The center decides $\hat{\mathsf{H}}$ by MAP:

$$\frac{\mathbb{P}\left\{\bar{f}_1, \bar{f}_2 \middle| \mathsf{H}=1\right\}}{\mathbb{P}\left\{\bar{f}_1, \bar{f}_2 \middle| \mathsf{H}=0\right\}} \underset{\hat{\mathsf{H}}=0}{\overset{\hat{\mathsf{H}}=1}{\gtrless}} \frac{P_{\mathsf{H}}(0)}{P_{\mathsf{H}}(1)}$$

Local Geometric Structure



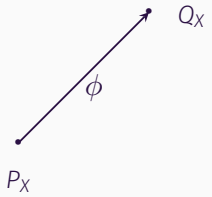
Local Geometric Structure



Distribution Q_X $\leftrightarrow$ *Information Vector*

$$\phi(x) \triangleq \frac{Q_X(x) - P_X(x)}{\sqrt{P_X(x)}}$$

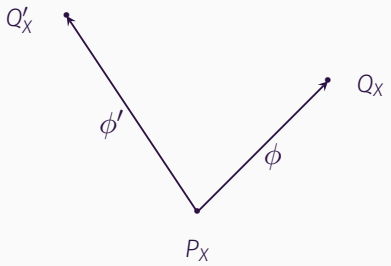
Local Geometric Structure



Distribution		*Information Vector*		*Feature Function*
Q_X	$\leftrightarrow$	$\phi(x) \triangleq \frac{Q_X(x) - P_X(x)}{\sqrt{P_X(x)}}$	$\leftrightarrow$	$f(x) \triangleq \frac{\phi(x)}{\sqrt{P_X(x)}}$

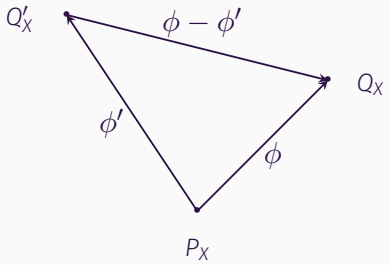
Local Geometric Structure

- $Q_X \leftrightarrow \phi, Q'_X \leftrightarrow \phi'$



Distribution		*Information Vector*		*Feature Function*
Q_X	$\leftrightarrow$	$\phi(x) \triangleq \frac{Q_X(x) - P_X(x)}{\sqrt{P_X(x)}}$	$\leftrightarrow$	$f(x) \triangleq \frac{\phi(x)}{\sqrt{P_X(x)}}$

Local Geometric Structure

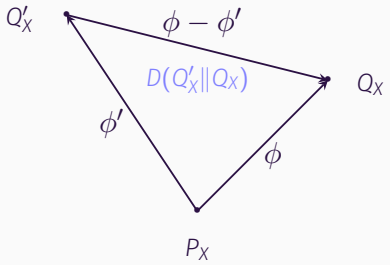


- $Q_X \leftrightarrow \phi, Q'_X \leftrightarrow \phi'$
- LLR (log-likelihood ratio):

$$\log \frac{Q'_X(x)}{Q_X(x)} \leftrightarrow \phi - \phi'$$

Distribution		*Information Vector*		*Feature Function*
Q_X	$\leftrightarrow$	$\phi(x) \triangleq \frac{Q_X(x) - P_X(x)}{\sqrt{P_X(x)}}$	$\leftrightarrow$	$f(x) \triangleq \frac{\phi(x)}{\sqrt{P_X(x)}}$

Local Geometric Structure



- $Q_X \leftrightarrow \phi, Q'_X \leftrightarrow \phi'$
- LLR (log-likelihood ratio):

$$\log \frac{Q'_X(x)}{Q_X(x)} \leftrightarrow \phi - \phi'$$

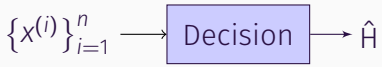
- K-L divergence:

$$D(Q'_X \| Q_X) = \frac{1}{2}\|\phi - \phi'\|^2$$

Distribution		*Information Vector*		*Feature Function*
Q_X	$\leftrightarrow$	$\phi(x) \triangleq \frac{Q_X(x) - P_X(x)}{\sqrt{P_X(x)}}$	$\leftrightarrow$	$f(x) \triangleq \frac{\phi(x)}{\sqrt{P_X(x)}}$

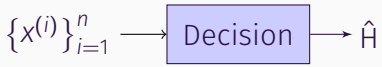
$\{x^{(i)}\}_{i=1}^{n} \longrightarrow$ Decision

- n samples $\{x^{(1)}, \ldots, x^{(n)}\} \overset{\text{i.i.d.}}{\sim} P_X^{(\mathrm{H})}$



- n samples $\{x^{(1)}, \ldots, x^{(n)}\} \overset{\text{i.i.d.}}{\sim} P_X^{(\mathrm{H})}$
- LLRT with opt. feature ℓ

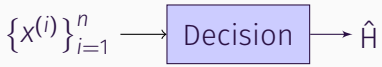
$$\bar{\ell} \underset{\hat{\mathrm{H}}=0}{\overset{\hat{\mathrm{H}}=1}{\gtrless}} 0, \quad \ell(x) \triangleq \log \frac{P_X^{(1)}(x)}{P_X^{(0)}(x)}$$



- n samples $\{x^{(1)}, \ldots, x^{(n)}\} \overset{\text{i.i.d.}}{\sim} P_X^{(H)}$
- LLRT with opt. feature ℓ

$$\bar{\ell} \underset{\hat{H}=0}{\overset{\hat{H}=1}{\gtrless}} 0, \quad \ell(x) \triangleq \log \frac{P_X^{(1)}(x)}{P_X^{(0)}(x)}$$

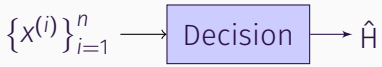
- Opt. error exponent E^*:
  $=$ Chernoff Information between $P_X^{(0)}$ and $P_X^{(1)}$



- n samples $\{x^{(1)}, \ldots, x^{(n)}\} \overset{\text{i.i.d.}}{\sim} P_X^{(\mathrm{H})} \leftrightarrow \psi^{(\mathrm{H})}$
- LLRT with opt. feature ℓ

$$\bar{\ell} \underset{\hat{\mathrm{H}}=0}{\overset{\hat{\mathrm{H}}=1}{\gtrless}} 0, \quad \ell(x) \triangleq \log \frac{P_X^{(1)}(x)}{P_X^{(0)}(x)}$$

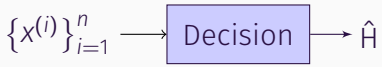
- Opt. error exponent E^*:
  $=$ Chernoff Information between $P_X^{(0)}$ and $P_X^{(1)}$



- n samples $\{x^{(1)}, \ldots, x^{(n)}\} \overset{\text{i.i.d.}}{\sim} P_X^{(\mathrm{H})} \leftrightarrow \psi^{(\mathrm{H})}$
- LLRT with opt. feature $\ell \leftrightarrow \phi_\ell = \psi^{(1)} - \psi^{(0)}$

$$\bar{\ell} \underset{\hat{\mathrm{H}}=0}{\overset{\hat{\mathrm{H}}=1}{\gtrless}} 0, \quad \ell(x) \triangleq \log \frac{P_X^{(1)}(x)}{P_X^{(0)}(x)}$$

- Opt. error exponent E^*:
  $=$ Chernoff Information between $P_X^{(0)}$ and $P_X^{(1)}$



- n samples $\{x^{(1)}, \ldots, x^{(n)}\} \overset{\text{i.i.d.}}{\sim} P_X^{(\mathrm{H})} \leftrightarrow \psi^{(\mathrm{H})}$
- LLRT with opt. feature $\ell \leftrightarrow \phi_\ell = \psi^{(1)} - \psi^{(0)}$

$$\bar{\ell} \underset{\hat{\mathrm{H}}=0}{\overset{\hat{\mathrm{H}}=1}{\gtrless}} 0, \quad \ell(x) \triangleq \log \frac{P_X^{(1)}(x)}{P_X^{(0)}(x)}$$

- Opt. error exponent E^*:
  $=$ Chernoff Information between $P_X^{(0)}$ and $P_X^{(1)}$
  $= \|\phi_\ell\|^2/8$

- n samples $\{x^{(1)}, \ldots, x^{(n)}\} \overset{\text{i.i.d.}}{\sim} P_X^{(\mathsf{H})} \leftrightarrow \psi^{(\mathsf{H})}$
- LLRT with opt. feature $\ell \leftrightarrow \phi_\ell = \psi^{(1)} - \psi^{(0)}$

$$\bar{\ell} \underset{\hat{\mathsf{H}}=0}{\overset{\hat{\mathsf{H}}=1}{\gtrless}} 0, \quad \ell(x) \triangleq \log \frac{P_X^{(1)}(x)}{P_X^{(0)}(x)}$$

- Opt. error exponent E^*:
  $=$ Chernoff Information between $P_X^{(0)}$ and $P_X^{(1)}$
  $= \|\phi_\ell\|^2/8$

Decision with feature $f \leftrightarrow \phi$

- Decision rule: $\bar{f} \underset{\hat{\mathsf{H}}=0}{\overset{\hat{\mathsf{H}}=1}{\gtrless}} 0$

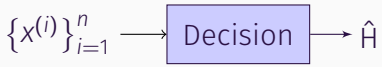
- n samples $\{x^{(1)}, \ldots, x^{(n)}\} \overset{\text{i.i.d.}}{\sim} P_X^{(\mathsf{H})} \leftrightarrow \psi^{(\mathsf{H})}$
- LLRT with opt. feature $\ell \leftrightarrow \phi_\ell = \psi^{(1)} - \psi^{(0)}$

$$\bar{\ell} \underset{\hat{\mathsf{H}}=0}{\overset{\hat{\mathsf{H}}=1}{\gtrless}} 0, \quad \ell(x) \triangleq \log \frac{P_X^{(1)}(x)}{P_X^{(0)}(x)}$$

- Opt. error exponent E^*:
  $=$ Chernoff Information between $P_X^{(0)}$ and $P_X^{(1)}$
  $= \|\phi_\ell\|^2/8$

Decision with feature $f \leftrightarrow \phi$

- Decision rule: $\bar{f} \underset{\hat{\mathsf{H}}=0}{\overset{\hat{\mathsf{H}}=1}{\gtrless}} 0$
- Exponent $E = \frac{1}{8} \cdot \frac{\langle \phi, \phi_\ell \rangle^2}{\|\phi\|^2}$



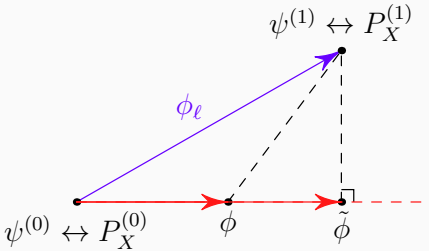
- n samples $\{x^{(1)}, \ldots, x^{(n)}\} \overset{\text{i.i.d.}}{\sim} P_X^{(H)} \leftrightarrow \psi^{(H)}$
- LLRT with opt. feature $\ell \leftrightarrow \phi_\ell = \psi^{(1)} - \psi^{(0)}$

$$\bar{\ell} \underset{\hat{H}=0}{\overset{\hat{H}=1}{\gtrless}} 0, \quad \ell(x) \triangleq \log \frac{P_X^{(1)}(x)}{P_X^{(0)}(x)}$$

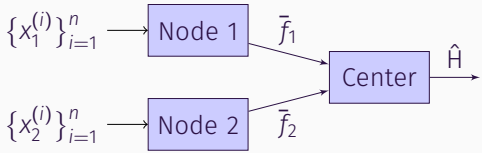
- Opt. error exponent E^*:
  $=$ Chernoff Information between $P_X^{(0)}$ and $P_X^{(1)}$
  $= \|\phi_\ell\|^2/8$

Decision with feature $f \leftrightarrow \phi$

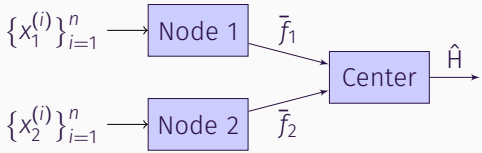
- Decision rule: $\bar{f} \underset{\hat{H}=0}{\overset{\hat{H}=1}{\gtrless}} 0$
- Exponent $E = \frac{1}{8} \cdot \frac{\langle \phi, \phi_\ell \rangle^2}{\|\phi\|^2} = \frac{1}{8}\|\tilde{\phi}\|^2$



Distributed Hypothesis Testing: Local Geometric Analysis

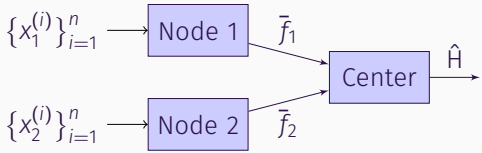


Distributed Hypothesis Testing: Local Geometric Analysis

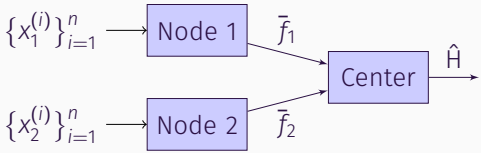


- f_1^*, f_2^* are both one-dim

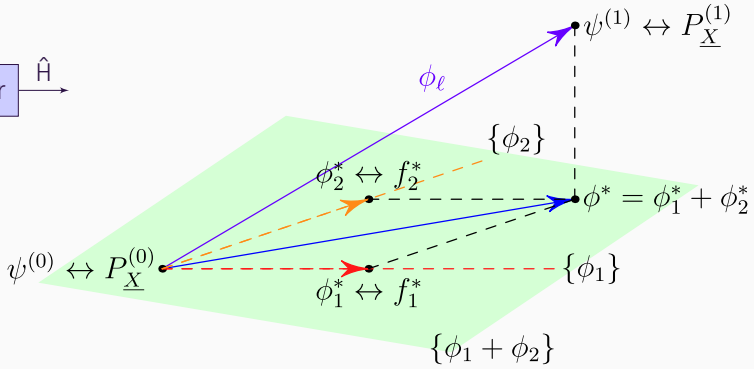
Distributed Hypothesis Testing: Local Geometric Analysis

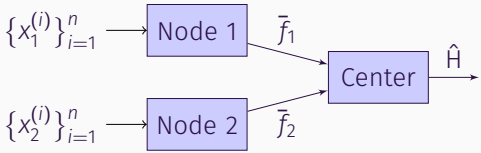


- f_1^*, f_2^* are both one-dim
- $f_1^* \leftrightarrow \phi_1^*, f_2^* \leftrightarrow \phi_2^*$

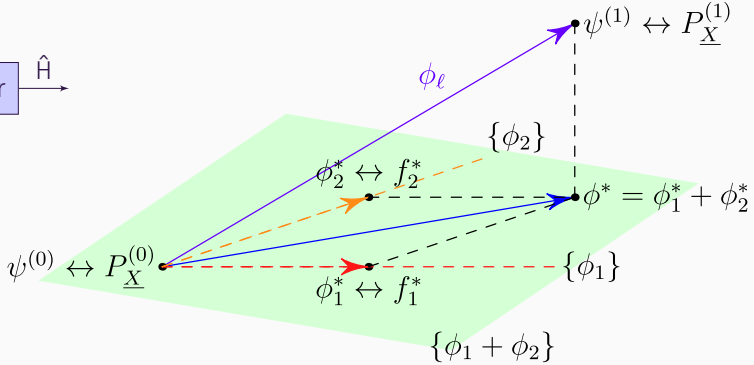


- f_1^*, f_2^* are both one-dim
- $f_1^* \leftrightarrow \phi_1^*, f_2^* \leftrightarrow \phi_2^*$
- $\phi^* = \phi_1^* + \phi_2^*$:
  Proj. of ϕ_ℓ onto $\{\phi_1 + \phi_2\}$
- $\{\phi_1 + \phi_2\} \leftrightarrow \{f_1 + f_2\}$, features of the form $f_1(x_1) + f_2(x_2)$



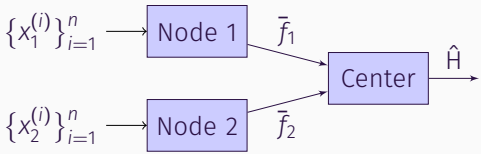


- f_1^*, f_2^* are both one-dim
- $f_1^* \leftrightarrow \phi_1^*, f_2^* \leftrightarrow \phi_2^*$
- $\phi^* = \phi_1^* + \phi_2^*$:
  Proj. of ϕ_ℓ onto $\{\phi_1 + \phi_2\}$
- $\{\phi_1 + \phi_2\} \leftrightarrow \{f_1 + f_2\}$, features of the form $f_1(x_1) + f_2(x_2)$

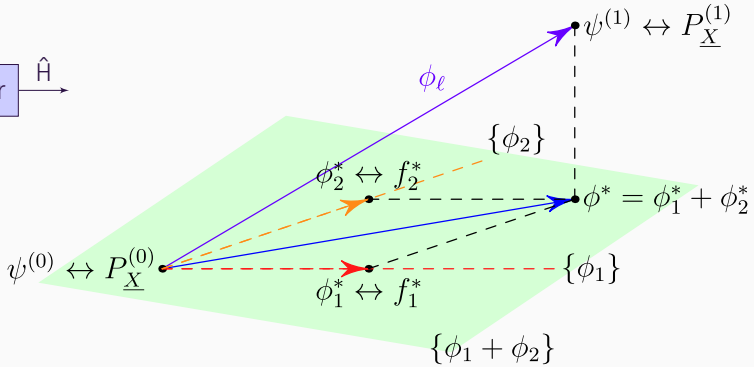


Decision based only on X_1 samples:

$\iff$ Proj. of ϕ_ℓ onto $\{\phi_1\}$



- f_1^*, f_2^* are both one-dim
- $f_1^* \leftrightarrow \phi_1^*, f_2^* \leftrightarrow \phi_2^*$
- $\phi^* = \phi_1^* + \phi_2^*$:
  Proj. of ϕ_ℓ onto $\{\phi_1 + \phi_2\}$
- $\{\phi_1 + \phi_2\} \leftrightarrow \{f_1 + f_2\}$, features of the form $f_1(x_1) + f_2(x_2)$

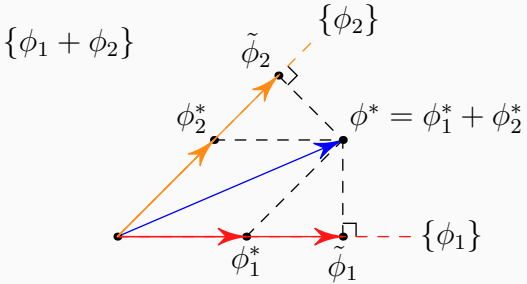


Decision based only on X_1 samples:

$\iff$ Proj. of ϕ_ℓ onto $\{\phi_1\}$

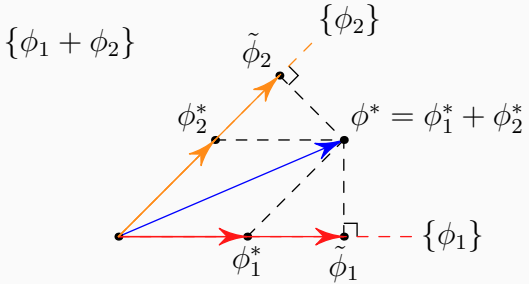
$\iff$ Proj. of ϕ^* onto $\{\phi_1\}$

Information Decomposition among Distributed Nodes



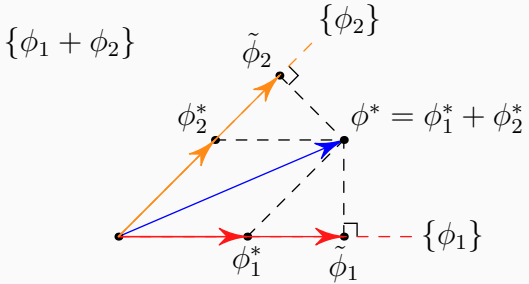
- $\{\phi_1\} \leftrightarrow \{\text{feature } f_1 \text{ of } x_1\}$
- $\{\phi_2\} \leftrightarrow \{\text{feature } f_2 \text{ of } x_2\}$
- $\{\phi_1 + \phi_2\} \leftrightarrow \{f_1 + f_2\}$

Information Decomposition among Distributed Nodes



- $\{\phi_1\} \leftrightarrow \{\text{feature } f_1 \text{ of } x_1\}$
- $\{\phi_2\} \leftrightarrow \{\text{feature } f_2 \text{ of } x_2\}$
- $\{\phi_1 + \phi_2\} \leftrightarrow \{f_1 + f_2\}$
- $\tilde{\phi}_1$: proj. of ϕ^* onto $\{\phi_1\}$
  $\leftrightarrow$ opt. feature in X_1-based decision
- $\tilde{\phi}_2$: proj. of ϕ^* onto $\{\phi_2\}$
  $\leftrightarrow$ opt. feature in X_2-based decision
- Performances: $\|\tilde{\phi}_1\|^2, \|\tilde{\phi}_2\|^2, \|\phi^*\|^2$

Information Decomposition among Distributed Nodes

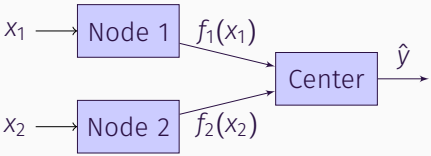


- $\{\phi_1\} \leftrightarrow \{\text{feature } f_1 \text{ of } x_1\}$
- $\{\phi_2\} \leftrightarrow \{\text{feature } f_2 \text{ of } x_2\}$
- $\{\phi_1 + \phi_2\} \leftrightarrow \{f_1 + f_2\}$
- $\tilde{\phi}_1$: proj. of ϕ^* onto $\{\phi_1\}$
  $\leftrightarrow$ opt. feature in X_1-based decision
- $\tilde{\phi}_2$: proj. of ϕ^* onto $\{\phi_2\}$
  $\leftrightarrow$ opt. feature in X_2-based decision
- Performances: $\|\tilde{\phi}_1\|^2, \|\tilde{\phi}_2\|^2, \|\phi^*\|^2$

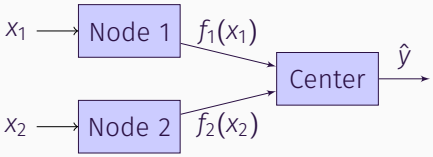
Due to the correlations between distributed nodes, orthogonal decomposition is in general not optimal.

Classification with Distributed Nodes

- Each node k transmits feature $f_k(x_k)$



Classification with Distributed Nodes

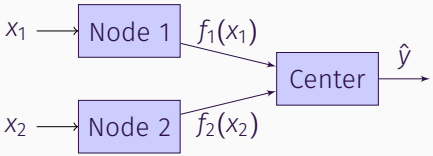


$$\underline{x} \triangleq \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \underline{f}(\underline{x}) \triangleq \begin{bmatrix} f_1(x_1) \\ f_2(x_2) \end{bmatrix}$$

- Each node k transmits feature $f_k(x_k)$
- The center learns a softmax classifier

$$\tilde{P}_{Y|\underline{X}}(y|\underline{x}) = \frac{\exp\left(\underline{f}^{\mathrm{T}}(\underline{x})g(y)\right)}{\sum_{y'} \exp\left(\underline{f}^{\mathrm{T}}(\underline{x})g(y)\right)}$$

Classification with Distributed Nodes



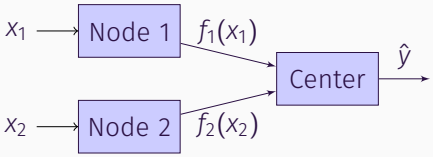
$$\underline{x} \triangleq \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \underline{f}(\underline{x}) \triangleq \begin{bmatrix} f_1(x_1) \\ f_2(x_2) \end{bmatrix}$$

- Each node k transmits feature $f_k(x_k)$
- The center learns a softmax classifier

$$\tilde{P}_{Y|\underline{X}}(y|\underline{x}) = \frac{\exp\left(\underline{f}^{\mathrm{T}}(\underline{x}) g(y)\right)}{\sum_{y'} \exp\left(\underline{f}^{\mathrm{T}}(\underline{x}) g(y)\right)}$$

- Predict $\hat{y}$ by MAP:

$$\hat{y} = \arg\max_{y'} \tilde{P}_{Y|\underline{X}}(y'|\underline{x})$$



$$\underline{x} \triangleq \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \underline{f}(\underline{x}) \triangleq \begin{bmatrix} f_1(x_1) \\ f_2(x_2) \end{bmatrix}$$

- Each node k transmits feature $f_k(x_k)$
- The center learns a softmax classifier

$$\tilde{P}_{Y|\underline{X}}(y|\underline{x}) = \frac{\exp\left(\underline{f}^{\mathrm{T}}(\underline{x}) g(y)\right)}{\sum_{y'} \exp\left(\underline{f}^{\mathrm{T}}(\underline{x}) g(y)\right)}$$

- Predict $\hat{y}$ by MAP:

$$\hat{y} = \arg\max_{y'} \tilde{P}_{Y|\underline{X}}(y'|\underline{x})$$

- **Goal**: minimize the logarithm loss

$$L \triangleq -\frac{1}{n} \sum_{i=1}^{n} \log \tilde{P}_{Y|\underline{X}}(y^{(i)}|\underline{x}^{(i)})$$

f_1^* and f_2^* are also optimal, under the correspondences

Hypothesis Testing	Classification
$P_{\underline{X}}^{(0)}$	$P_{\underline{X}\|Y=0}$
$P_{\underline{X}}^{(1)}$	$P_{\underline{X}\|Y=1}$
P_{H}	P_Y

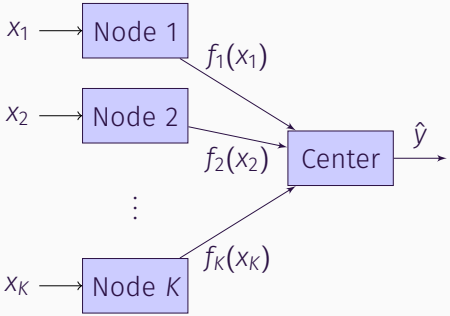
f_1^* and f_2^* are also optimal, under the correspondences

Hypothesis Testing	Classification
$P_{\underline{X}}^{(0)}$	$P_{\underline{X}\|Y=0}$
$P_{\underline{X}}^{(1)}$	$P_{\underline{X}\|Y=1}$
P_{H}	P_Y

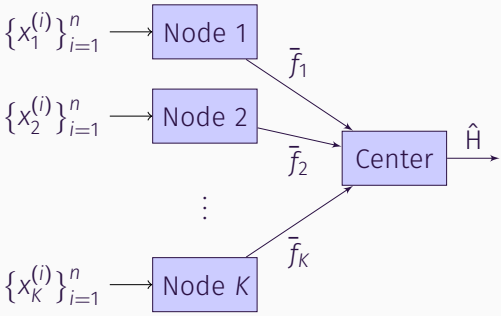
- Extension of [Huang et al., ISIT 2019]

Generalization: Learning with K Distributed Nodes

Distributed Classification

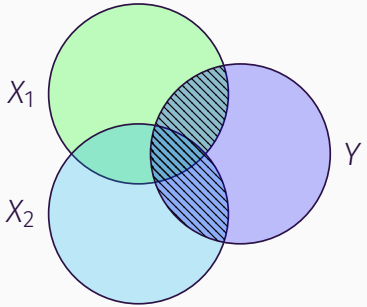


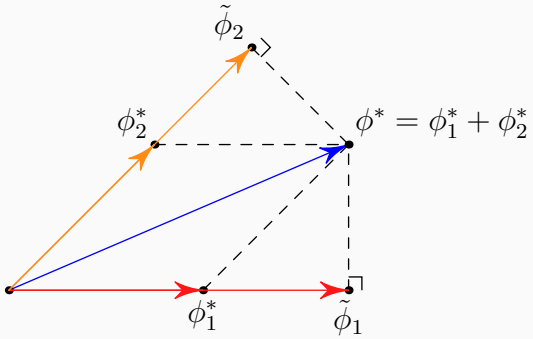
Distributed Hypothesis Testing



Summary

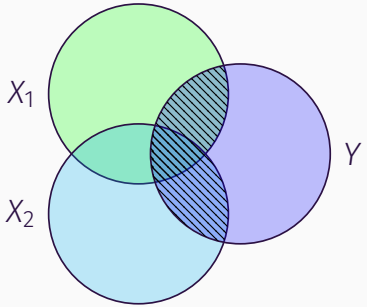
- Information-theoretic framework

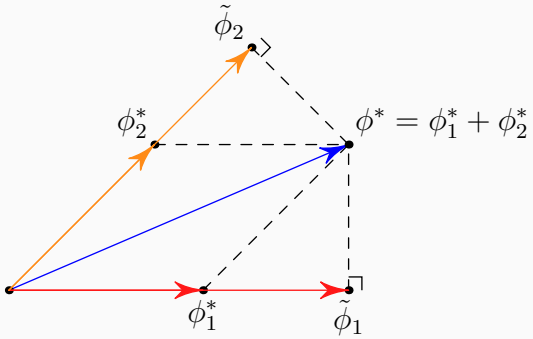




Summary

- Information-theoretic framework
- Local geometric analyses: information decomposition structure





Summary

- Information-theoretic framework
- Local geometric analyses: information decomposition structure
- Connection between information theory and machine learning