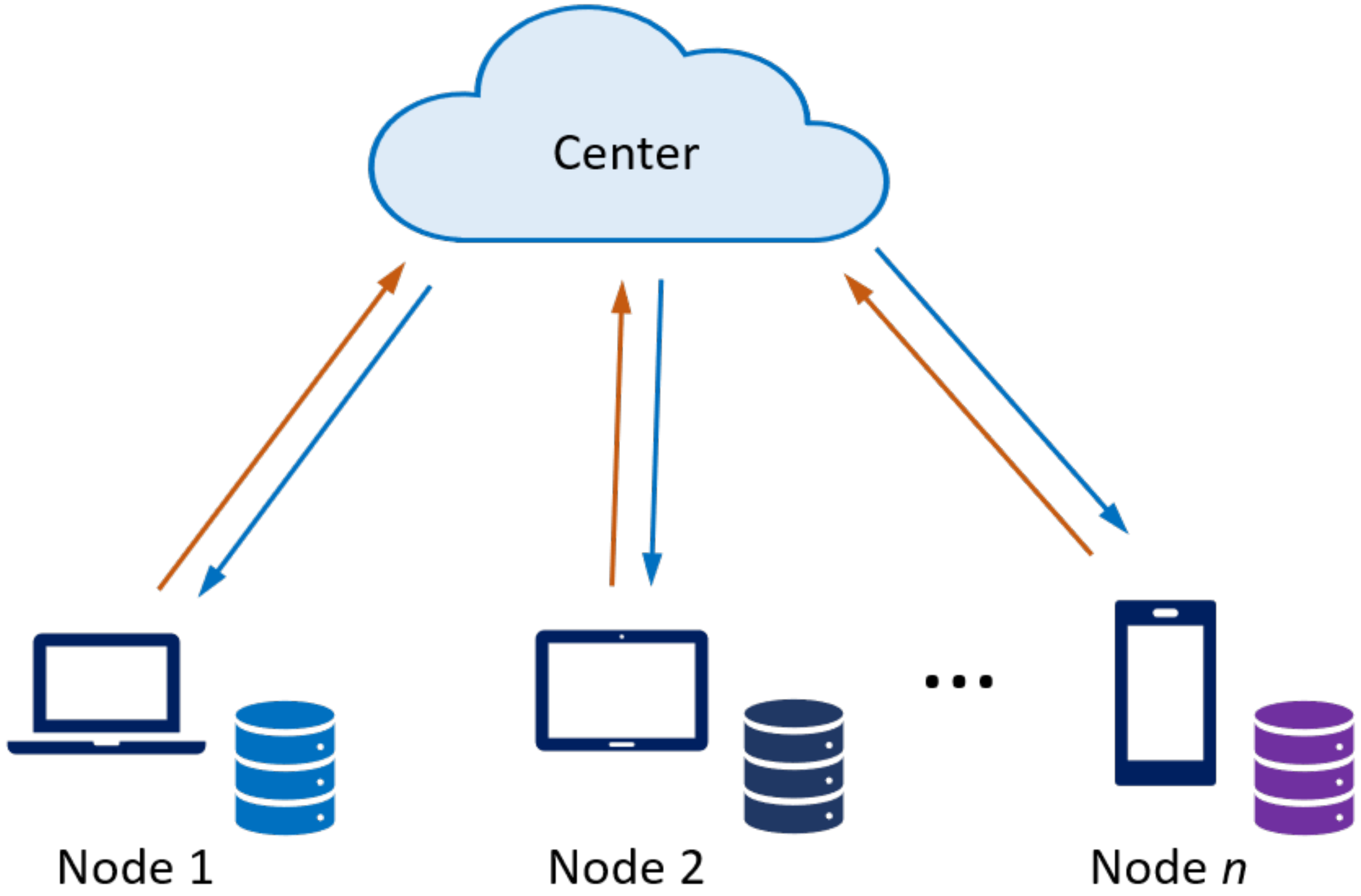
Feature Extractions in Distributed Parameter Estimation: A Local Information Geometric Approach

Shao-Lun Huang
Tsinghua-Berkeley Shenzhen Institute (TBSI)

2024/3/11@BIRS Workshop, Banff
Algorithmic Structures for Uncoordinated Communications and Statistical Inference in Exceedingly Large Spaces

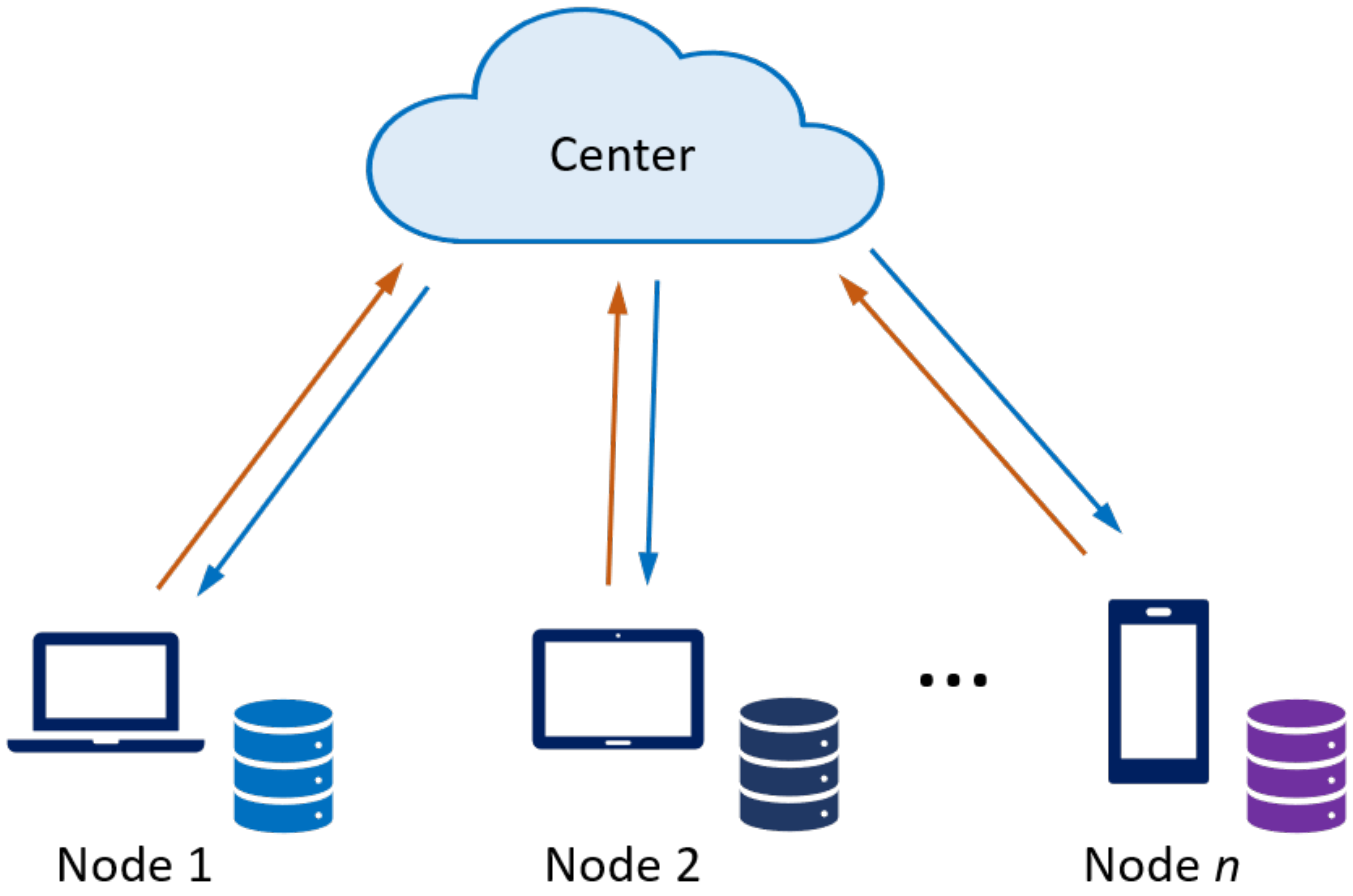
Joint work with Xiniyi Tong

Distributed Learning



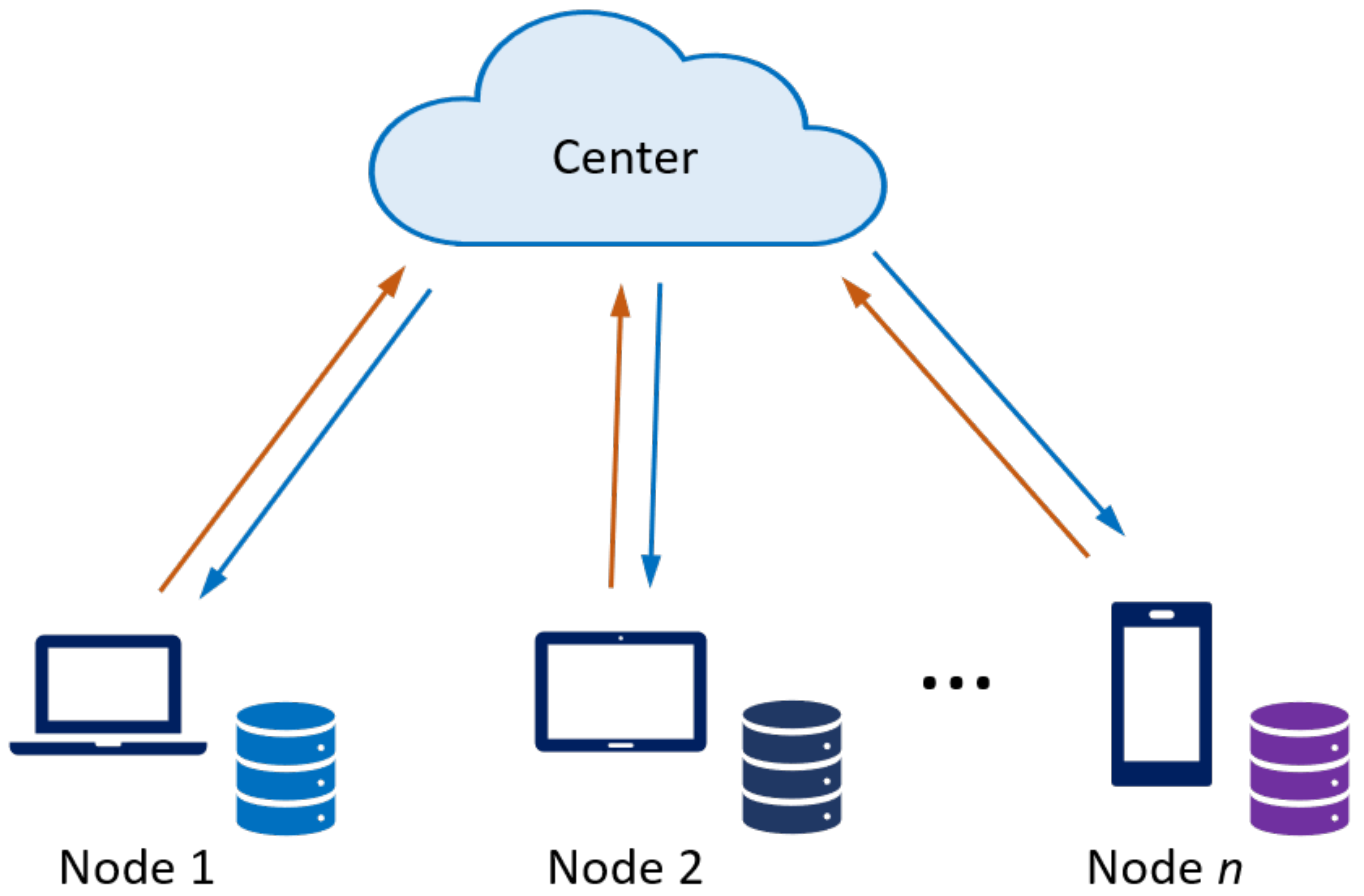
- Distributed nodes collected data and communicate with the center.
  - Label prediction, parameter estimation, model training,…

Distributed Learning



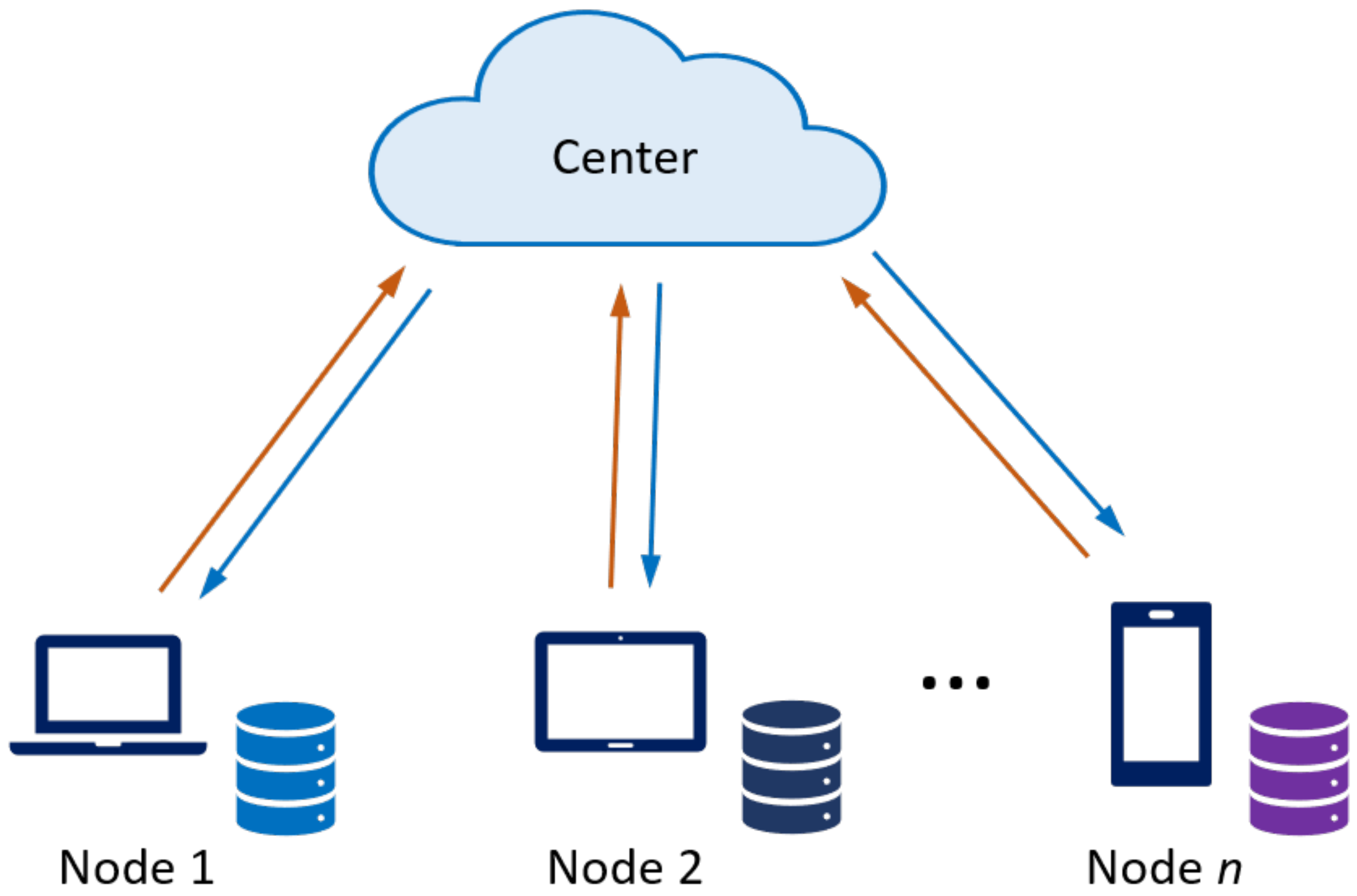
- Distributed nodes collected data and communicate with the center.
  - Label prediction, parameter estimation, model training,…
- Restricted communication between nodes and center.

Motivation



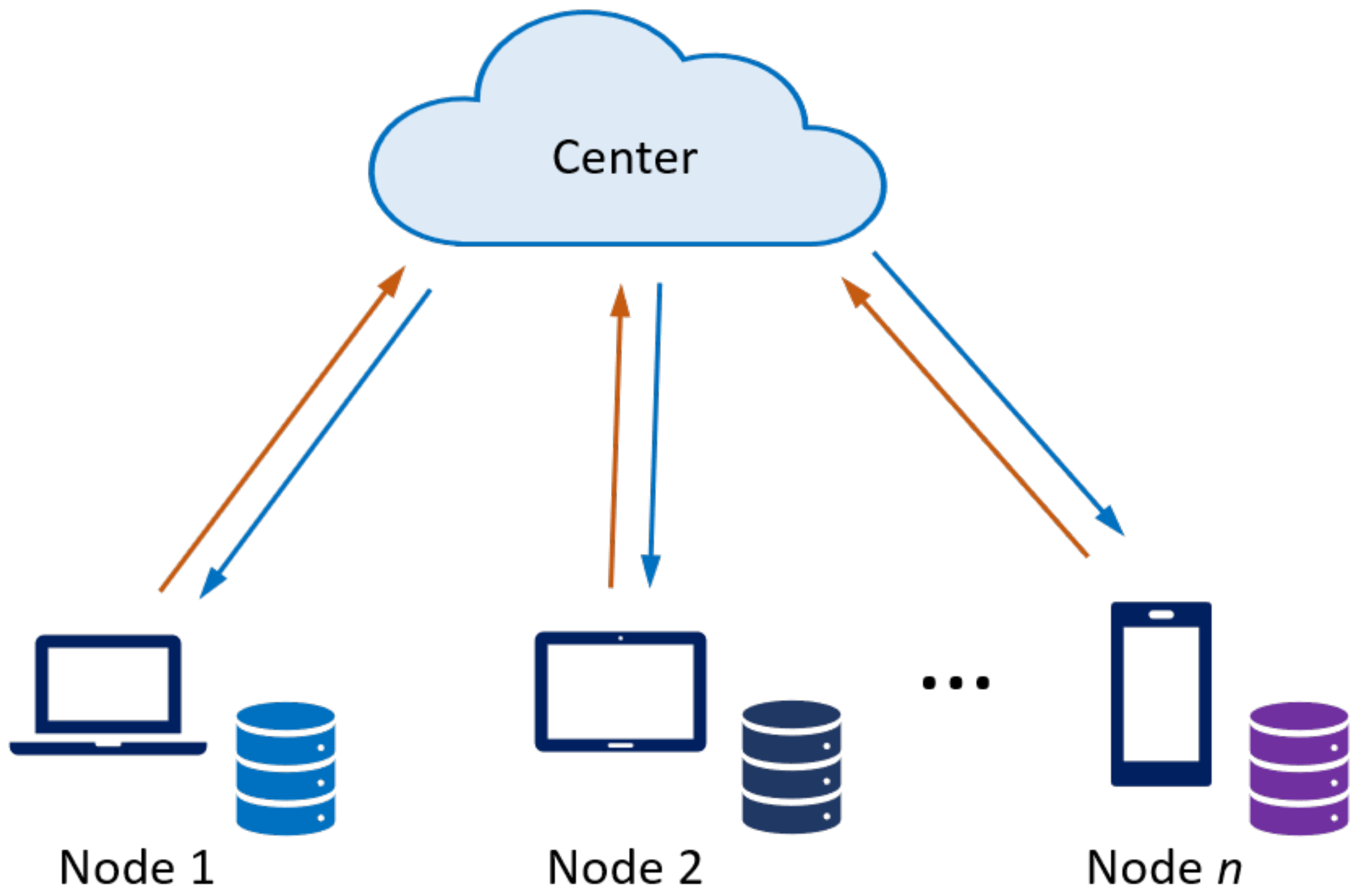
- Suppose that the nodes can communicate the with the statistics of the data:
  - Observe $x_1, x_2, \cdots, x_n$, transmit $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ to other nodes for some function $f(x)$.

Motivation



- Suppose that the nodes can communicate the with the statistics of the data:
  - Observe $x_1, x_2, \cdots, x_n$, transmit $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ to other nodes for some function $f(x)$.
- Computationally efficient for high-dimensional data.

Motivation

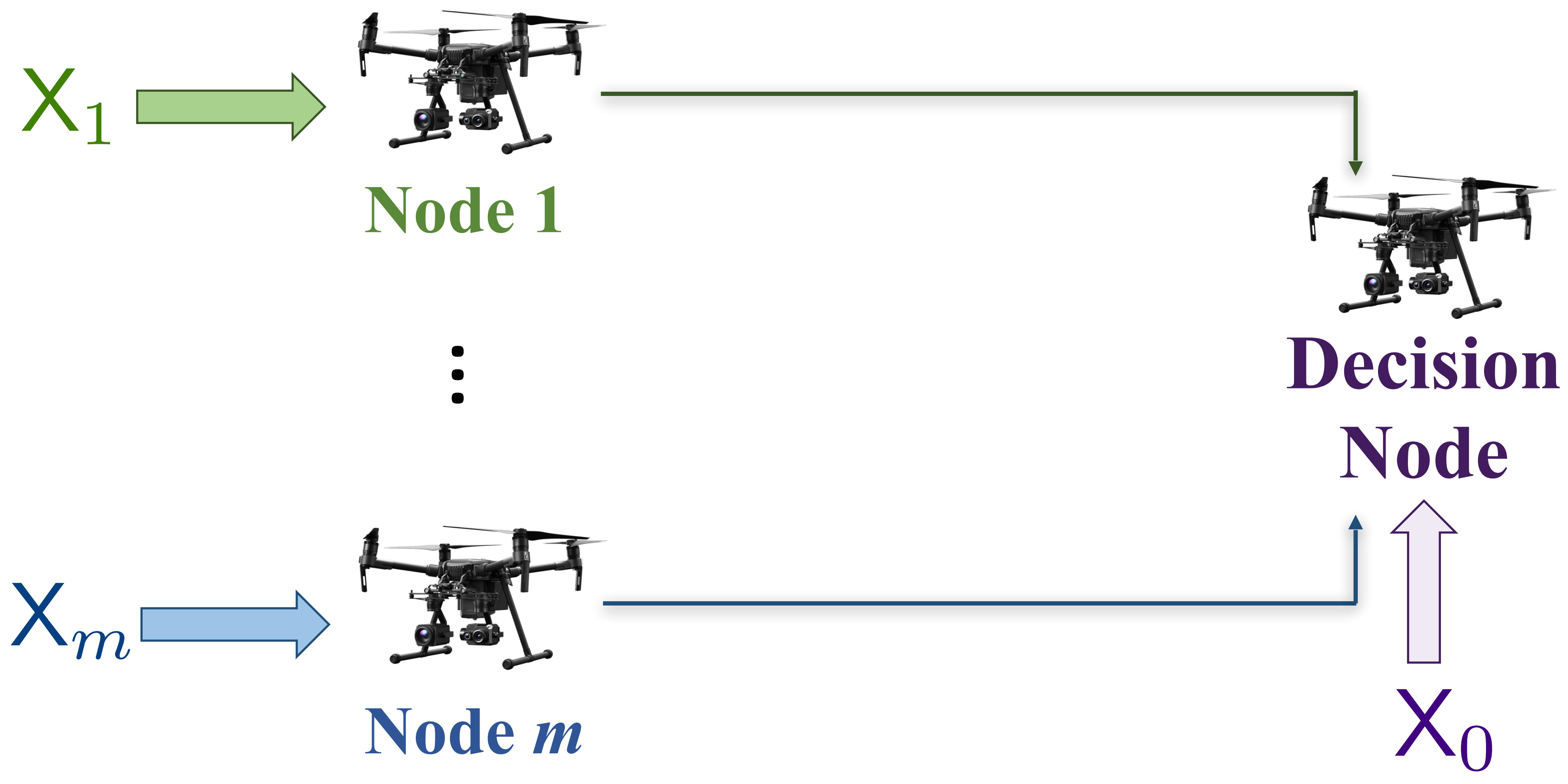


- Suppose that the nodes can communicate the with the statistics of the data:
  - Observe $x_1, x_2, \cdots, x_n$, transmit $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ to other nodes for some function $f(x)$.
- Computationally efficient for high-dimensional data.
- Communication constraints = Dimensionality constraints of the features.

Collaborative Distributed Parameter Estimation

$$\mathsf{X}_i = (x_i^{(1)}, \ldots, x_i^{(n)}), \quad i = 1, \ldots, m,$$

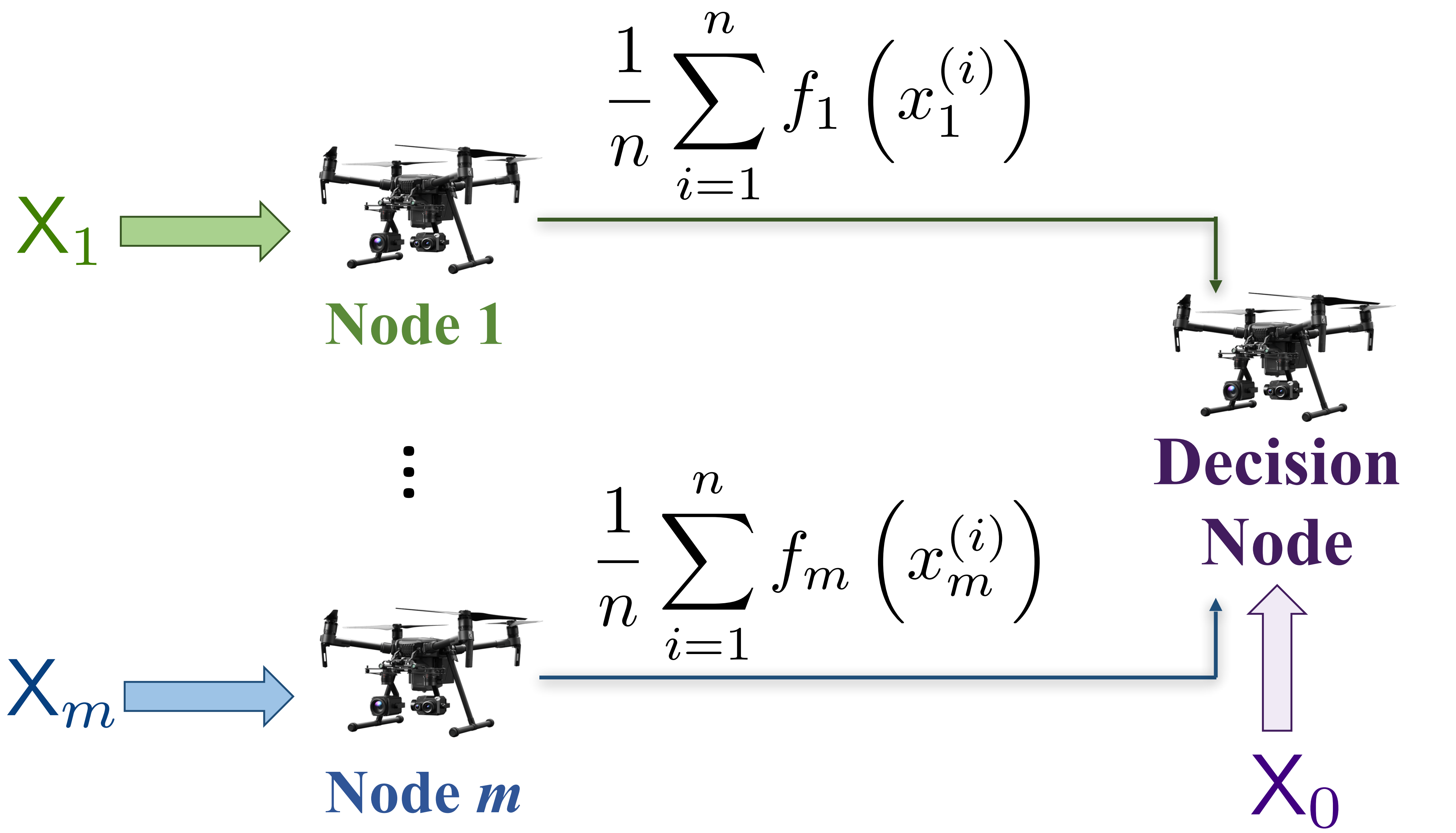
$$(x_0^{(j)}, \ldots, x_m^{(j)}) \overset{\text{i.i.d.}}{\sim} P_{X_0 \cdots X_m}(x_1, \ldots, x_m; \theta), \quad j = 1, \ldots, n.$$



Collaborative Distributed Parameter Estimation

$$\mathsf{X}_i = (x_i^{(1)}, \ldots, x_i^{(n)}), \quad i = 1, \ldots, m,$$

$$(x_0^{(j)}, \ldots, x_m^{(j)}) \overset{\text{i.i.d.}}{\sim} P_{X_0 \cdots X_m}(x_1, \ldots, x_m; \theta), \quad j = 1, \ldots, n.$$

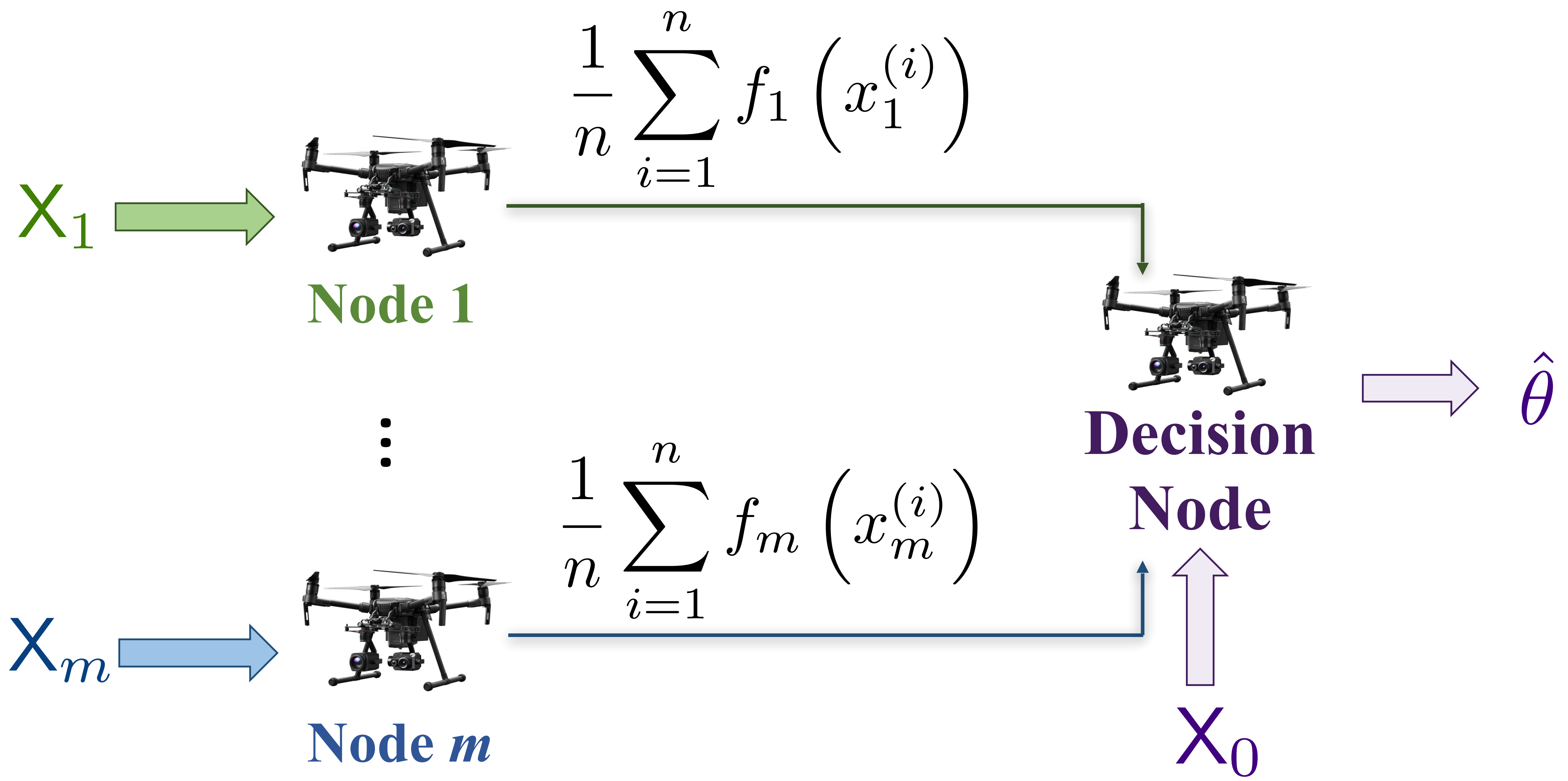


- Each node *i* transmit a statistic of its own data to the decision center.

Collaborative Distributed Parameter Estimation

$$\mathsf{X}_i = (x_i^{(1)}, \ldots, x_i^{(n)}), \quad i = 1, \ldots, m,$$

$$(x_0^{(j)}, \ldots, x_m^{(j)}) \overset{\text{i.i.d.}}{\sim} P_{X_0 \cdots X_m}(x_1, \ldots, x_m; \theta), \quad j = 1, \ldots, n.$$

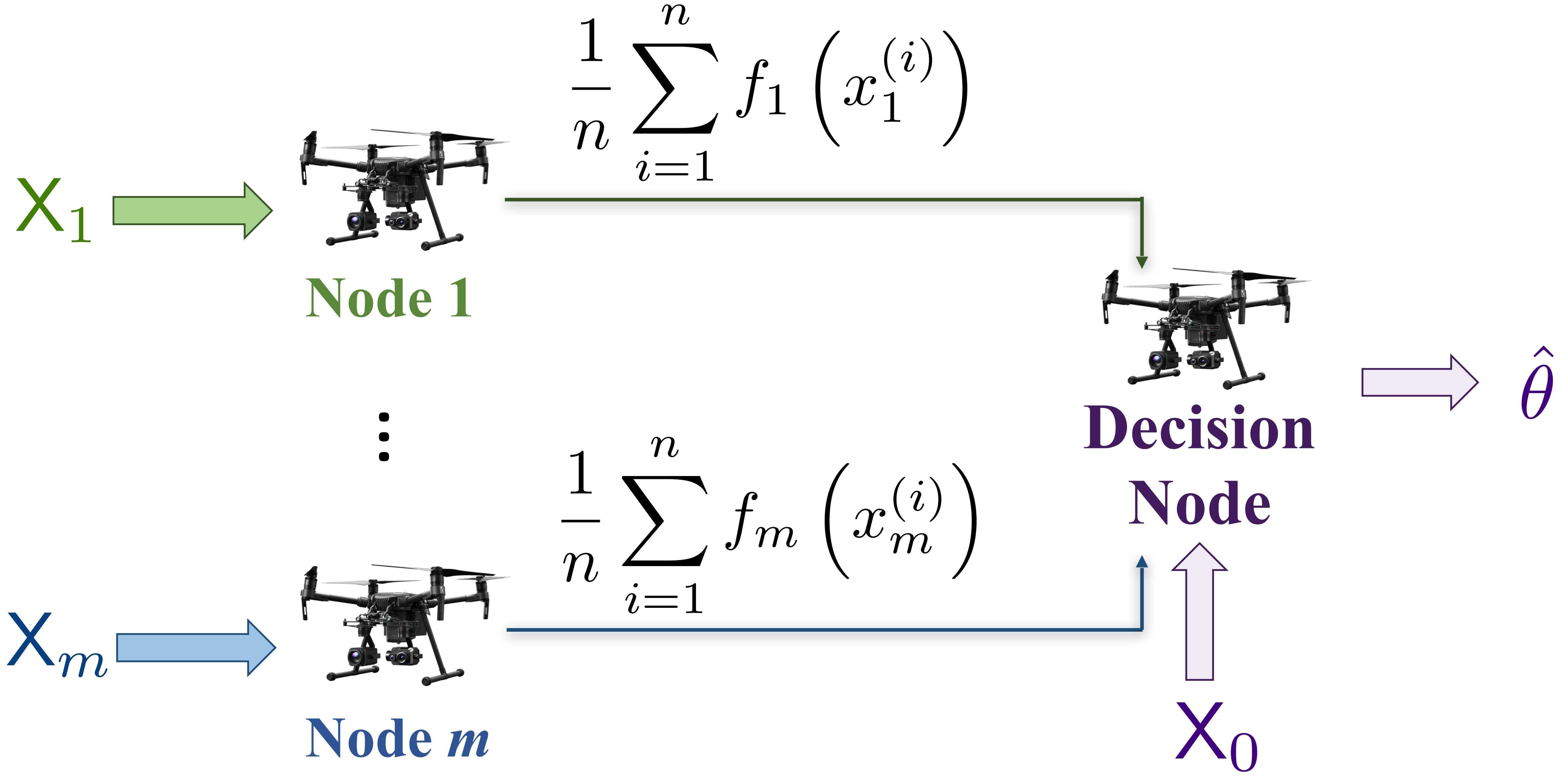


- Each node *i* transmit a statistic of its own data to the decision center.
  - The decision node estimate the parameter based on the statistics and its data.

Collaborative Distributed Parameter Estimation

$$\mathsf{X}_i = (x_i^{(1)}, \ldots, x_i^{(n)}), \quad i = 1, \ldots, m,$$

$$(x_0^{(j)}, \ldots, x_m^{(j)}) \overset{\text{i.i.d.}}{\sim} P_{X_0 \cdots X_m}(x_1, \ldots, x_m; \theta), \quad j = 1, \ldots, n.$$



- Each node *i* transmit a statistic of its own data to the decision center.
  - The decision node estimate the parameter based on the statistics and its data.
- What are informative feature functions $f_i : \mathcal{X}_i \mapsto \mathbb{R}^{k_i}$ the nodes should extract?

Cramér–Rao Lower Bound

- Given $x^n = (x_1, x_2, \ldots, x_n)$, i.i.d. generated from a distribution $P_X(x;\theta)$ parametrized by $\theta \in \mathbb{R}^K$, then for any *unbiased* estimator $\hat{\theta} : \mathcal{X}^n \mapsto \mathbb{R}^K$

$$\mathbb{E}\left[\left\|\hat{\theta}(x^n) - \theta\right\|^2\right] \geq \frac{1}{n}\mathrm{tr}\left\{J_X^{-1}(\theta)\right\}$$

where $J_X(\theta)$ is the Fisher information matrix defined as

$$J_X(\theta) = \tilde{S}_X^{\mathbf{T}}(\theta) \cdot \tilde{S}_X(\theta)$$

and the scaled score function defined as

$$\left[\tilde{S}_X(\theta)\right]_{x,\ell} = \sqrt{P_X(x;\theta)} \cdot \frac{\partial}{\partial\theta_\ell} \log P_X(x;\theta) = \sqrt{P_X(x;\theta)} \cdot \frac{\frac{\partial}{\partial\theta_\ell} P_X(x;\theta)}{P_X(x;\theta)}$$

Maximal Likelihood Estimator

- The maximal likelihood estimator (MLE) to estimate the parameter is defined as

$$\hat{\theta}_{\mathrm{MLE}}(x^n) = \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \log P_X(x_i;\theta) = \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{\hat{P}_X}\left[\log P_X(X;\theta)\right]$$

  - Depend only on the empirical distribution $\rightarrow$ sufficient statistic.

Maximal Likelihood Estimator

- The maximal likelihood estimator (MLE) to estimate the parameter is defined as

$$\hat{\theta}_{\mathrm{MLE}}(x^n) = \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \log P_X(x_i; \theta) = \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{\hat{P}_X} [\log P_X(X; \theta)]$$

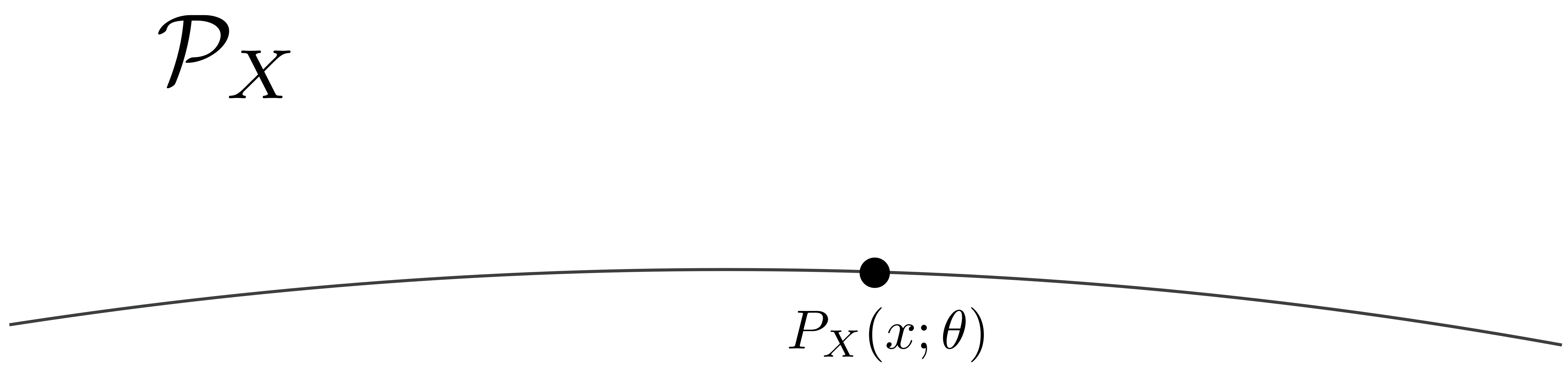
  - Depend only on the empirical distribution → sufficient statistic.

- The asymptotic normality of the MLE:

$$\sqrt{n} \cdot \left(\hat{\theta}_{\mathrm{MLE}}(x^n) - \theta \right) \overset{n \to \infty}{\longrightarrow} \mathcal{N}\left(\underline{0}, J_X^{-1}(\theta) \right)$$

  - The MLE asymptotically achieves the Cramér–Rao lower bound.

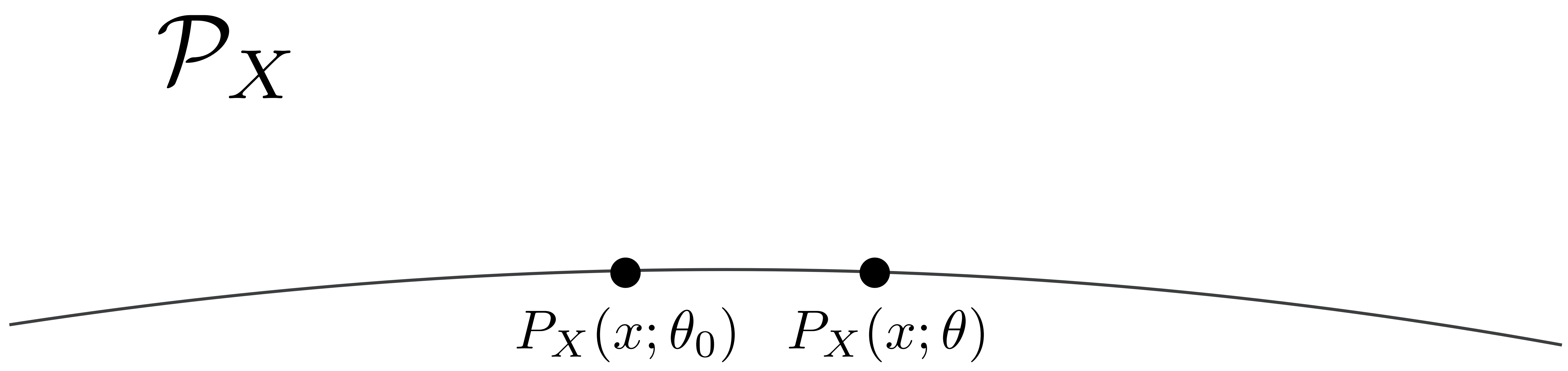
The Local Geometric Interpretation



$\boxed{\theta \in \mathbb{R}}$

$P_X(x;\theta)$: true distribution

The Local Geometric Interpretation

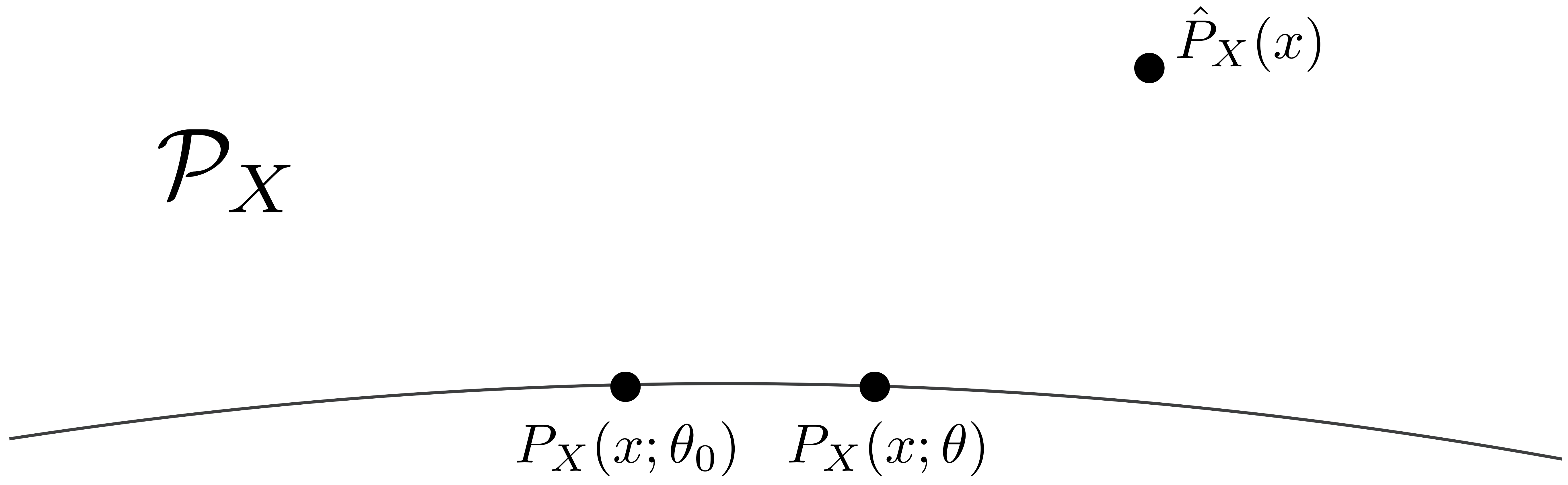


$\theta \in \mathbb{R}$

$P_X(x;\theta)$: true distribution

$P_X(x;\theta_0)$: reference distribution

The Local Geometric Interpretation



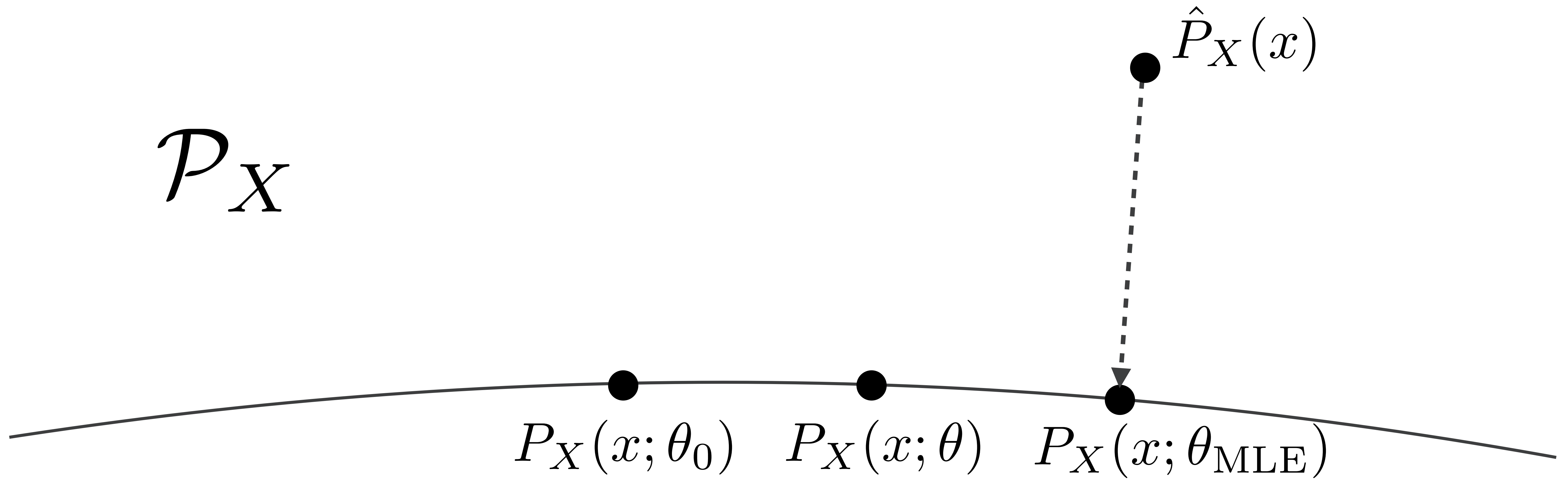
$\theta \in \mathbb{R}$

$P_X(x;\theta)$: true distribution

$P_X(x;\theta_0)$: reference distribution

$\hat{P}_X$: empirical distribution

The Local Geometric Interpretation



$\theta \in \mathbb{R}$

$P_X(x;\theta)$: true distribution

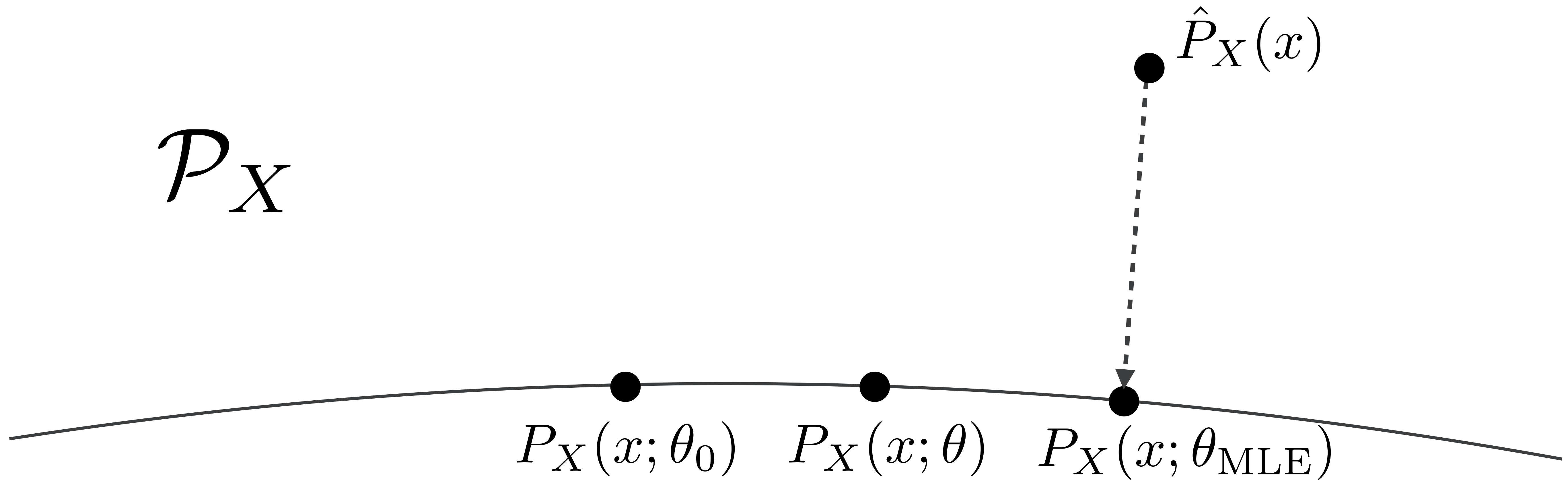
$P_X(x;\theta_0)$: reference distribution

$\hat{P}_X$: empirical distribution

$P_X(x;\theta_{\mathrm{MLE}})$: estimated distribution

The Local Geometric Interpretation

$\mathcal{P}_X$



$\theta \in \mathbb{R}$

$P_X(x;\theta)$: true distribution

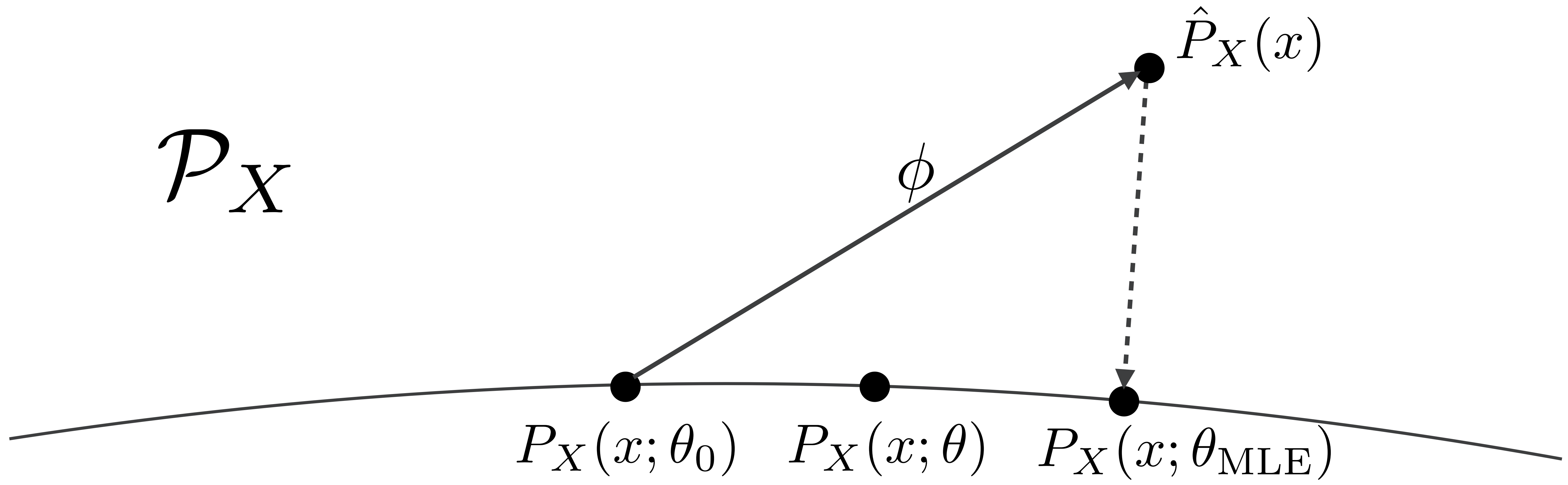
$P_X(x;\theta_0)$: reference distribution

$\hat{P}_X$: empirical distribution

$P_X(x;\theta_{\mathrm{MLE}})$: estimated distribution

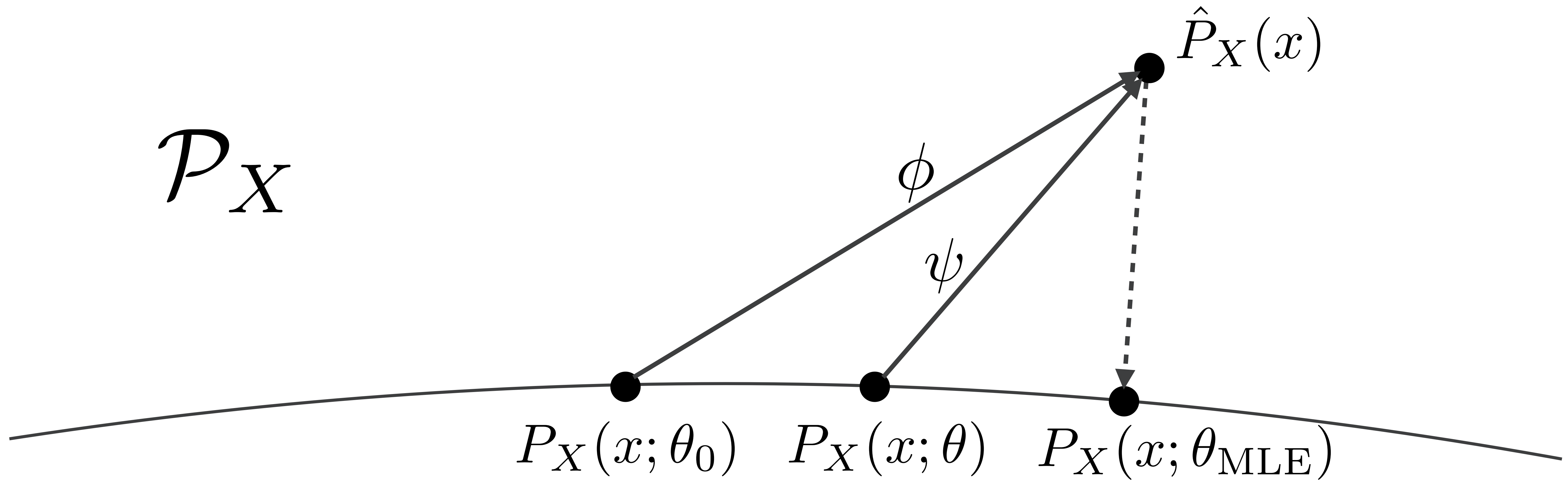
One-to-one correspondence between θ and $P_X(x;\theta)$.

The Local Geometric Interpretation



$$D(\hat{P}_X \| P_X(x;\theta_0)) \simeq \frac{1}{2}\sum_x \Bigg(\underbrace{\frac{\hat{P}_X(x) - P_X(x;\theta_0)}{\sqrt{P_X(x;\theta_0)}}}_{\phi(x)}\Bigg)^2 = \frac{1}{2}\|\phi\|^2$$

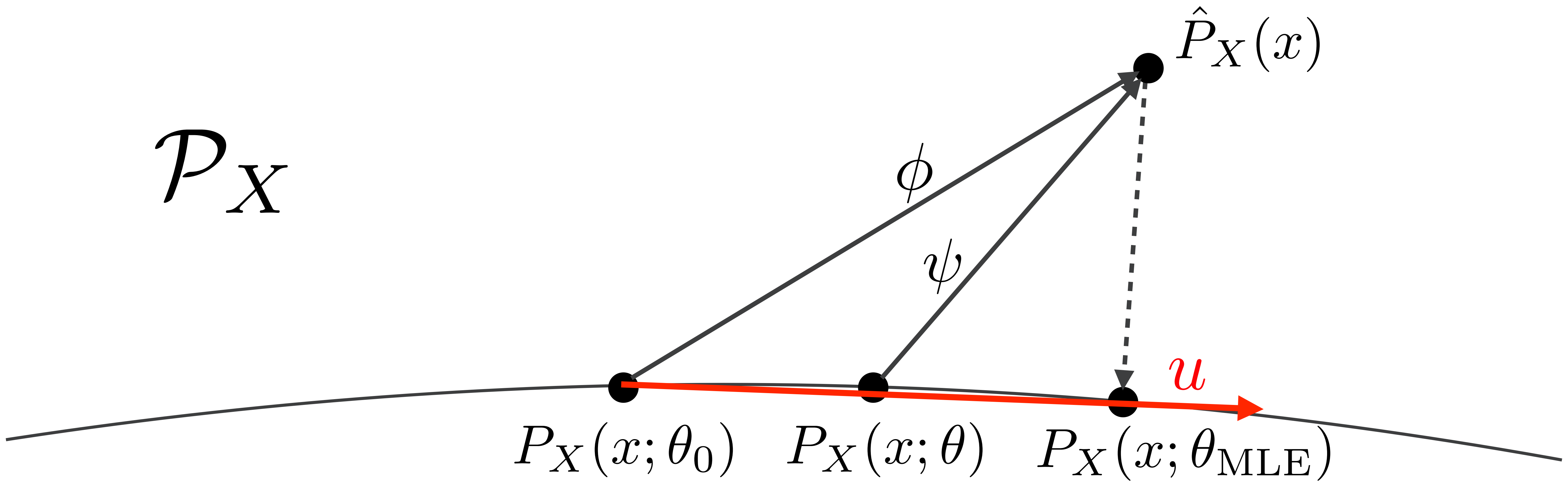
The Local Geometric Interpretation



$$D(\hat{P}_X \| P_X(x;\theta_0)) \simeq \frac{1}{2} \sum_x \Bigg(\underbrace{\frac{\hat{P}_X(x) - P_X(x;\theta_0)}{\sqrt{P_X(x;\theta_0)}}}_{\phi(x)} \Bigg)^2 = \frac{1}{2} \|\phi\|^2$$

$$D(\hat{P}_X \| P_X(x;\theta)) \simeq \frac{1}{2} \sum_x \Bigg(\underbrace{\frac{\hat{P}_X(x) - P_X(x;\theta)}{\sqrt{P_X(x;\theta)}}}_{\psi(x)} \Bigg)^2 = \frac{1}{2} \|\psi\|^2$$

The Local Geometric Interpretation

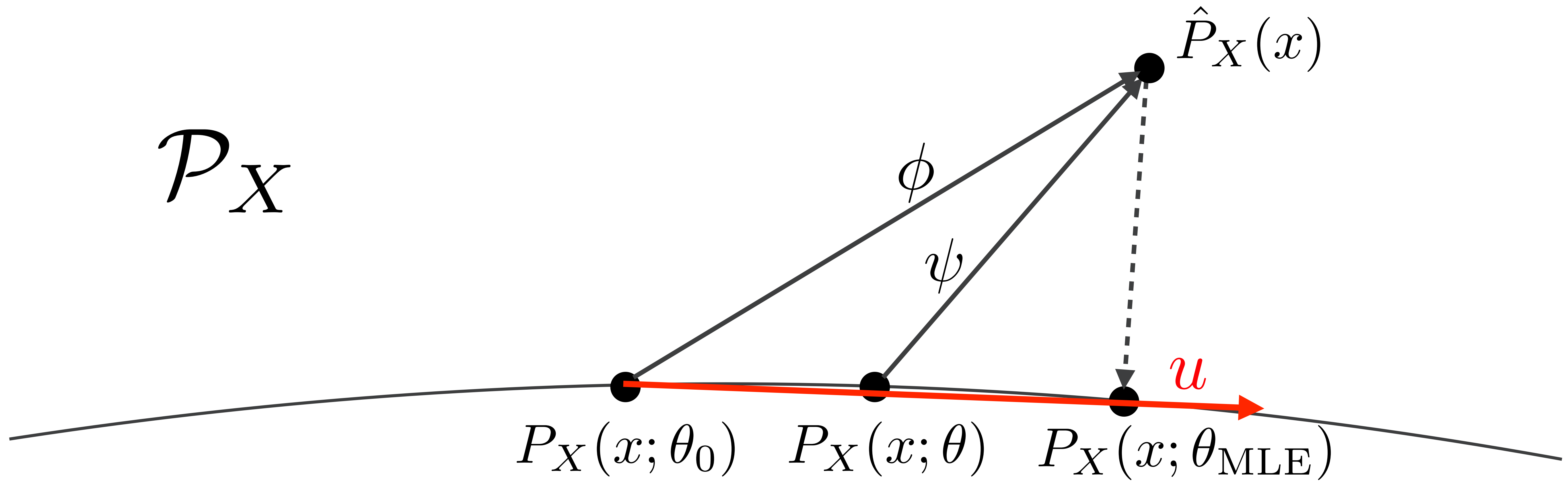


$$D(\hat{P}_X \| P_X(x;\theta_0)) \simeq \frac{1}{2}\sum_x \Bigg(\underbrace{\frac{\hat{P}_X(x) - P_X(x;\theta_0)}{\sqrt{P_X(x;\theta_0)}}}_{\phi(x)}\Bigg)^2 = \frac{1}{2}\|\phi\|^2$$

$$D(\hat{P}_X \| P_X(x;\theta)) \simeq \frac{1}{2}\sum_x \Bigg(\underbrace{\frac{\hat{P}_X(x) - P_X(x;\theta)}{\sqrt{P_X(x;\theta)}}}_{\psi(x)}\Bigg)^2 = \frac{1}{2}\|\psi\|^2$$

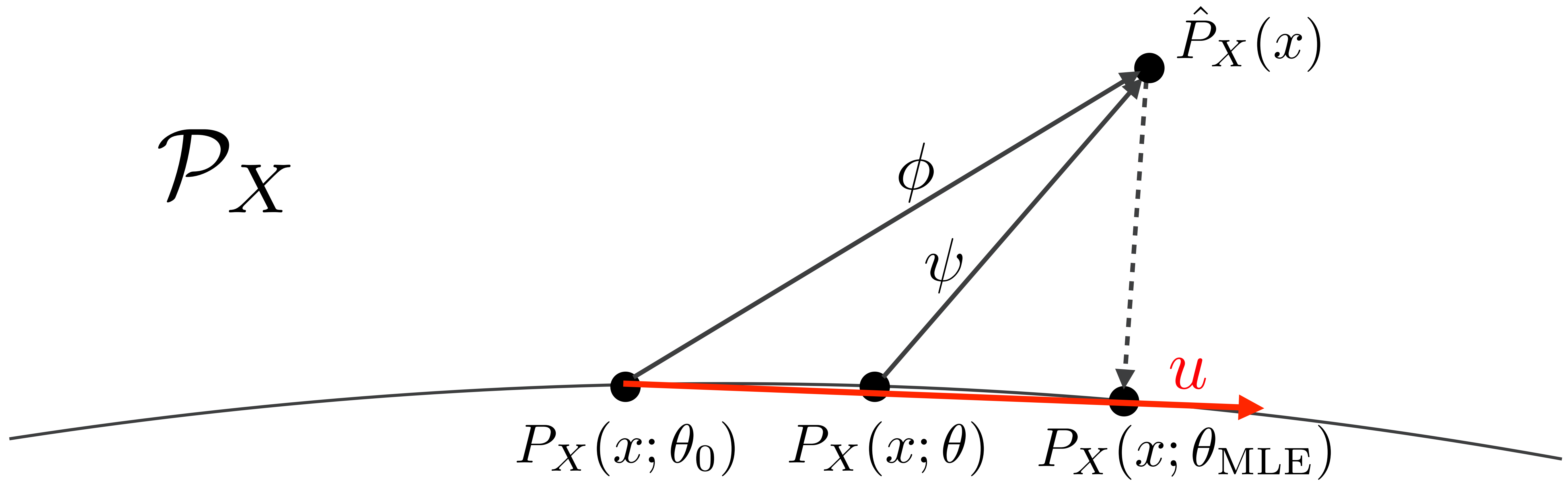
$$\tilde{u}(x) = \frac{P_X(x;\theta) - P_X(x;\theta_0)}{\sqrt{P_X(x;\theta_0)}} \simeq \frac{\frac{\partial}{\partial\theta}P_X(x;\theta_0)}{\sqrt{P_X(x;\theta_0)}} \cdot (\theta - \theta_0) \;\Rightarrow\; u(x) \triangleq \frac{\tilde{u}(x)}{\|\tilde{u}(x)\|} = J_X^{-\frac{1}{2}}(\theta_0)\frac{\frac{\partial}{\partial\theta}P_X(x;\theta_0)}{\sqrt{P_X(x;\theta_0)}}$$

The Local Geometric Interpretation



$$(\theta_{\text{MLE}} - \theta_0) \cdot \frac{\frac{\partial}{\partial \theta} P_X(x;\theta_0)}{\sqrt{P_X(x;\theta_0)}} = \langle \phi, u \rangle \cdot u \;\Rightarrow\; \theta_{\text{MLE}} = \theta_0 + J_X^{-\frac{1}{2}}(\theta_0) \cdot \langle \phi, u \rangle$$

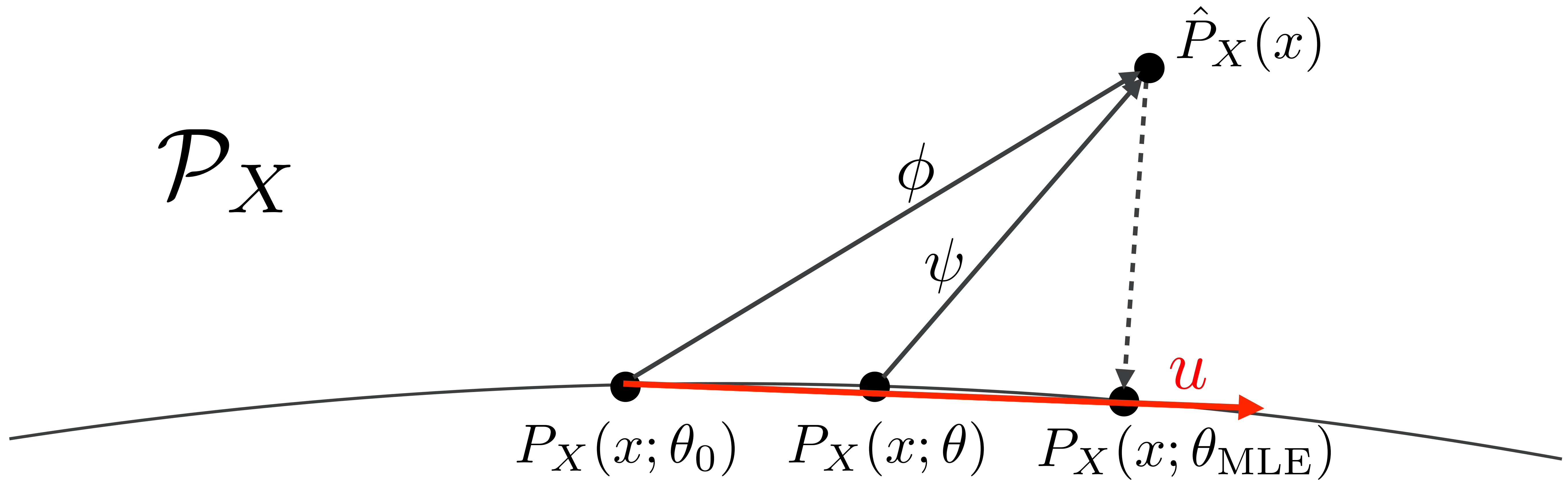
The Local Geometric Interpretation



$$(\theta_{\mathrm{MLE}} - \theta_0) \cdot \frac{\frac{\partial}{\partial\theta} P_X(x;\theta_0)}{\sqrt{P_X(x;\theta_0)}} = \langle \phi, u \rangle \cdot u \;\Rightarrow\; \theta_{\mathrm{MLE}} = \theta_0 + J_X^{-\frac{1}{2}}(\theta_0) \cdot \langle \phi, u \rangle$$

$$\langle \phi, u \rangle = \mathbb{E}_{\hat{P}_X}\Bigg[\underbrace{J_X^{-\frac{1}{2}}(\theta_0) \frac{\frac{\partial}{\partial\theta} P_X(X;\theta_0)}{P_X(X;\theta_0)}}_{f(X)} \Bigg] = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \;\Rightarrow\; \text{estimate based on the statistic of } f$$

The Local Geometric Interpretation

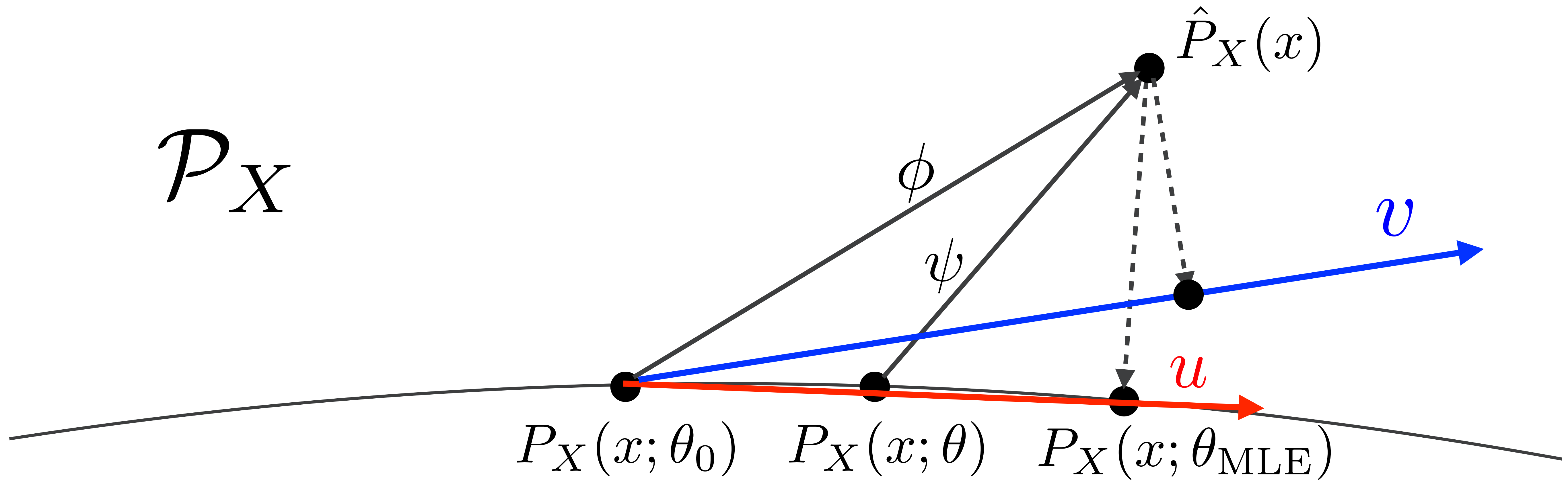


$$(\theta_{\text{MLE}} - \theta_0) \cdot \frac{\frac{\partial}{\partial\theta} P_X(x;\theta_0)}{\sqrt{P_X(x;\theta_0)}} = \langle \phi, u \rangle \cdot u \;\Rightarrow\; \theta_{\text{MLE}} = \theta_0 + J_X^{-\frac{1}{2}}(\theta_0) \cdot \langle \phi, u \rangle$$

$$\langle \phi, u \rangle = \mathbb{E}_{\hat{P}_X}\Bigg[\underbrace{J_X^{-\frac{1}{2}}(\theta_0) \frac{\frac{\partial}{\partial\theta} P_X(X;\theta_0)}{P_X(X;\theta_0)}}_{f(X)} \Bigg] = \frac{1}{n}\sum_{i=1}^{n} f(x_i) \;\Rightarrow\; \text{estimate based on the statistic of } f$$

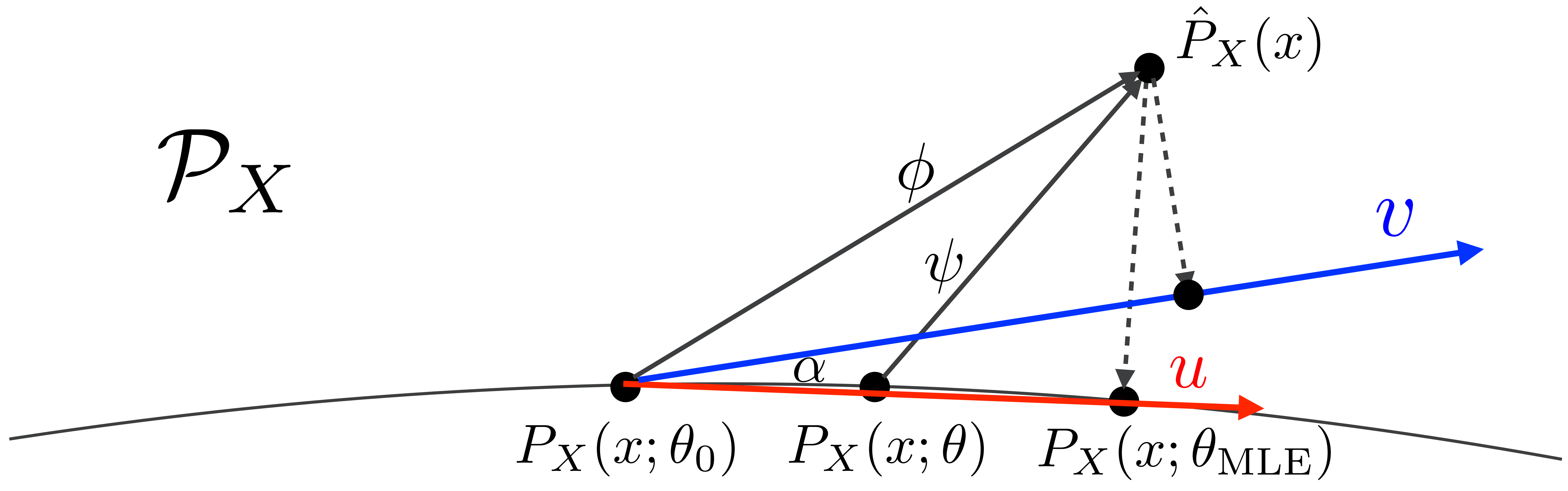
$$\text{MSE} = \mathbb{E}\left[(\theta_{\text{MLE}} - \theta)^2\right] = \mathbb{E}\left[\left(J_X^{-\frac{1}{2}}(\theta_0) \cdot \langle \psi, u \rangle\right)^2\right] = \frac{1}{n} J_X^{-1}(\theta_0)$$

The Mismatched Estimator



Mismatched statistic: $\hat{\theta} = \theta_0 + J_X^{-\frac{1}{2}}(\theta_0) \cdot \langle \phi, v \rangle$

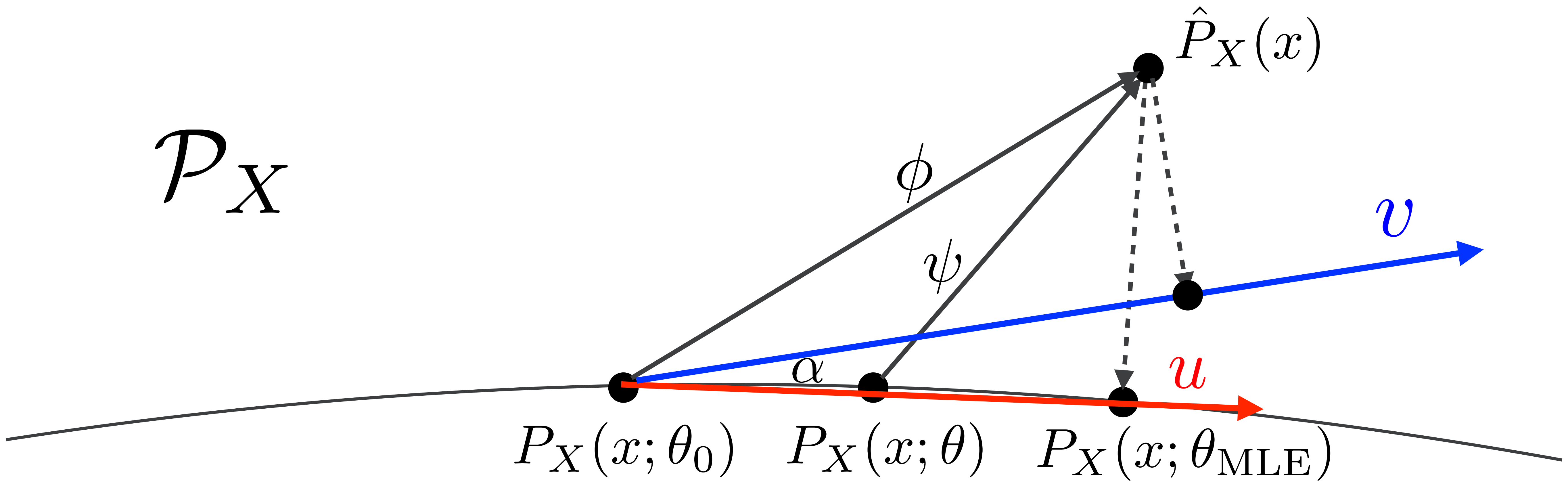
The Mismatched Estimator



Mismatched statistic: $\hat{\theta} = \theta_0 + J_X^{-\frac{1}{2}}(\theta_0) \cdot \langle \phi, v \rangle$

The degraded MSE $= \dfrac{1}{n} J_X^{-1}(\theta_0) \cdot \dfrac{1}{\langle u, v \rangle^2} = \dfrac{1}{n} J_X^{-1}(\theta_0) \cdot \dfrac{1}{\cos^2 \alpha}$

The Mismatched Estimator



Mismatched statistic: $\hat{\theta} = \theta_0 + J_X^{-\frac{1}{2}}(\theta_0) \cdot \langle \phi, v \rangle$

The degraded MSE $= \dfrac{1}{n} J_X^{-1}(\theta_0) \cdot \dfrac{1}{\langle u, v \rangle^2} = \dfrac{1}{n} J_X^{-1}(\theta_0) \cdot \dfrac{1}{\cos^2 \alpha}$

For general $\theta \in \mathbb{R}^K$, estimate θ based on $V = [v_1 \ v_2 \ \ldots \ v_k]^{\mathbf{T}}$,

the degraded MSE $= \dfrac{1}{n} \text{tr} \left\{ \left(\tilde{S}_X^{\mathbf{T}}(\theta_0) V^{\mathbf{T}} \left(V V^{\mathbf{T}} \right)^{-1} V \tilde{S}_X(\theta_0) \right)^{-1} \right\}$

$\Rightarrow$ Project $\tilde{S}_X(\theta_0)$ onto the subspace spanned by V .

清华-伯克利深圳学院 TBSI Tsinghua-Berkeley Shenzhen Institute

Distributed Scenarios

- When the parameter estimation is based statistics of data in a functional subspace.
  - MSE is characterized by the projection of the score function onto the subspace.

Distributed Scenarios

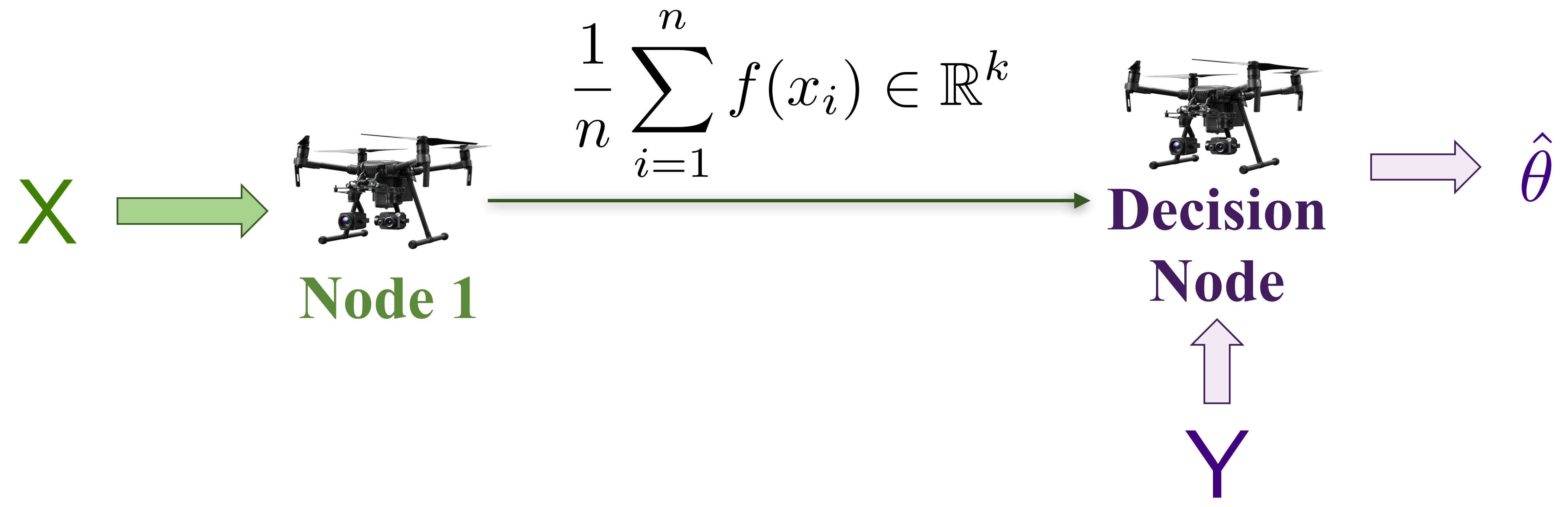
- When the parameter estimation is based statistics of data in a functional subspace.
  - MSE is characterized by the projection of the score function onto the subspace.
- In distributed parameter estimation problems, each node observes part of the data.
  - k-dimensional statistics = k-dimensional functional subspaces of data.
  - The computable estimators form a functional subspace.

Distributed Scenarios

- When the parameter estimation is based statistics of data in a functional subspace.
  - MSE is characterized by the projection of the score function onto the subspace.
- In distributed parameter estimation problems, each node observes part of the data.
  - k-dimensional statistics = k-dimensional functional subspaces of data.
  - The computable estimators form a functional subspace.
- Challenge: where to get the reference distributions?
  - Collaborative distributed parameter estimation.

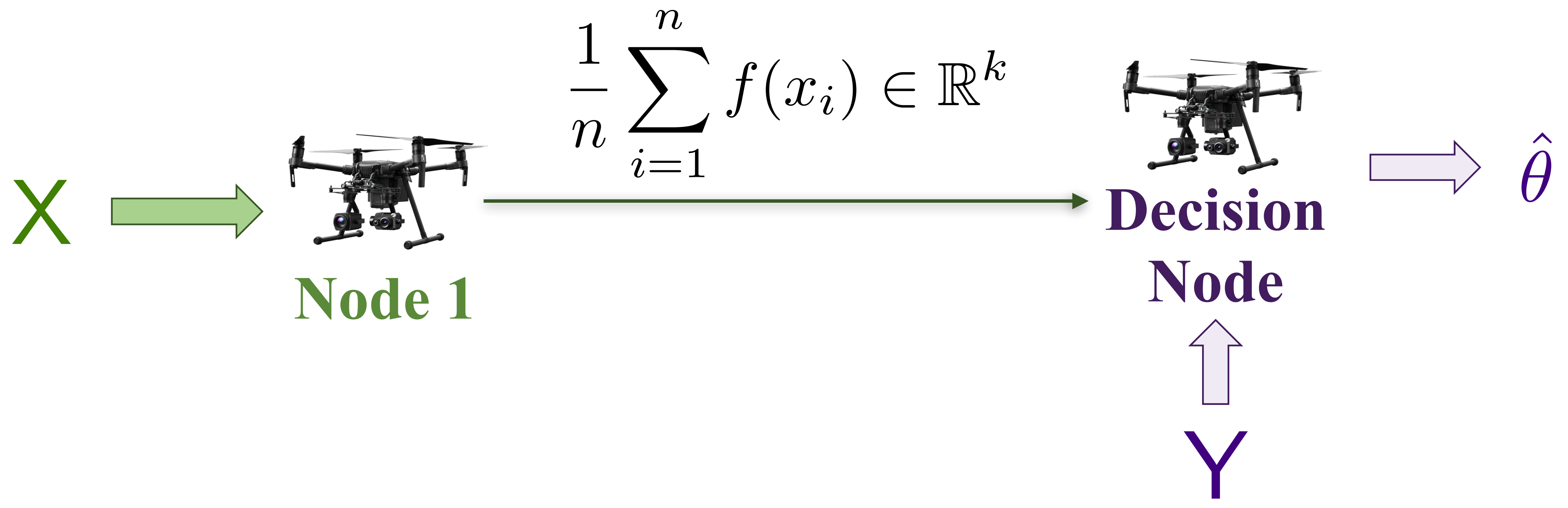
The Two Nodes Case

$$\mathsf{X} = (x_1, \ldots, x_n),\ \mathsf{Y} = (y_1, \ldots, y_n), \quad (x_i, y_i) \overset{\text{i.i.d.}}{\sim} P_{XY}(x, y; \theta)$$



The Two Nodes Case

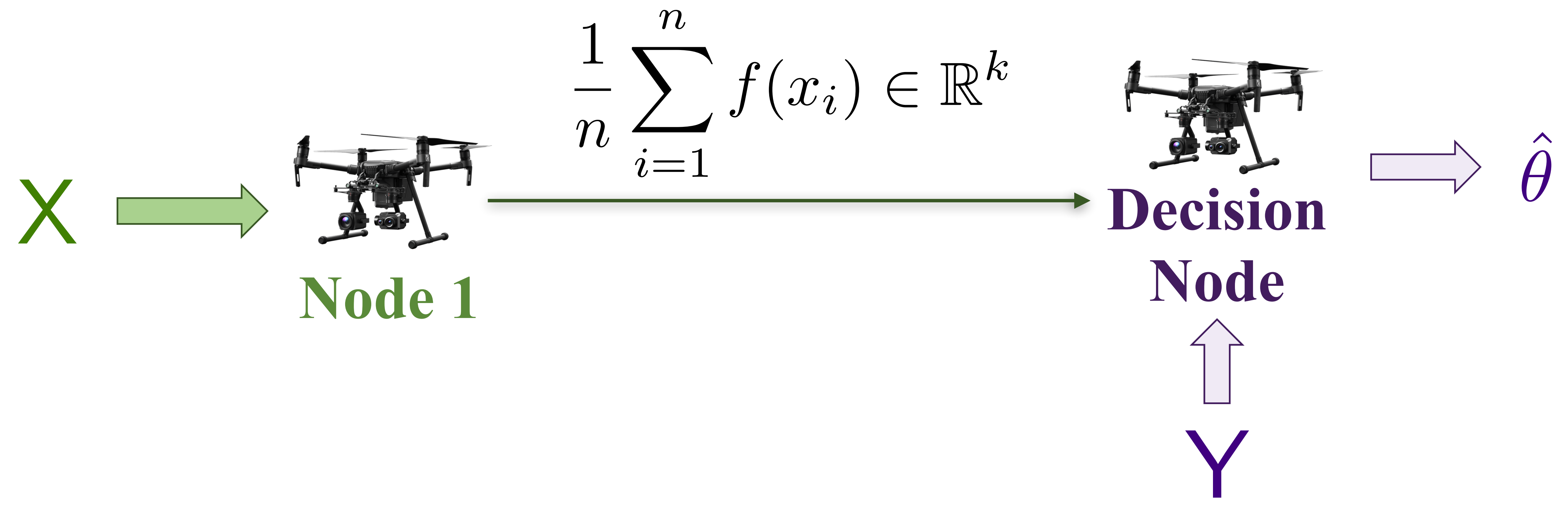
$$\mathsf{X} = (x_1, \ldots, x_n),\ \mathsf{Y} = (y_1, \ldots, y_n), \quad (x_i, y_i) \overset{\text{i.i.d.}}{\sim} P_{XY}(x, y; \theta)$$



- Node 1 transmits a statistic of its own data to the decision center.

The Two Nodes Case

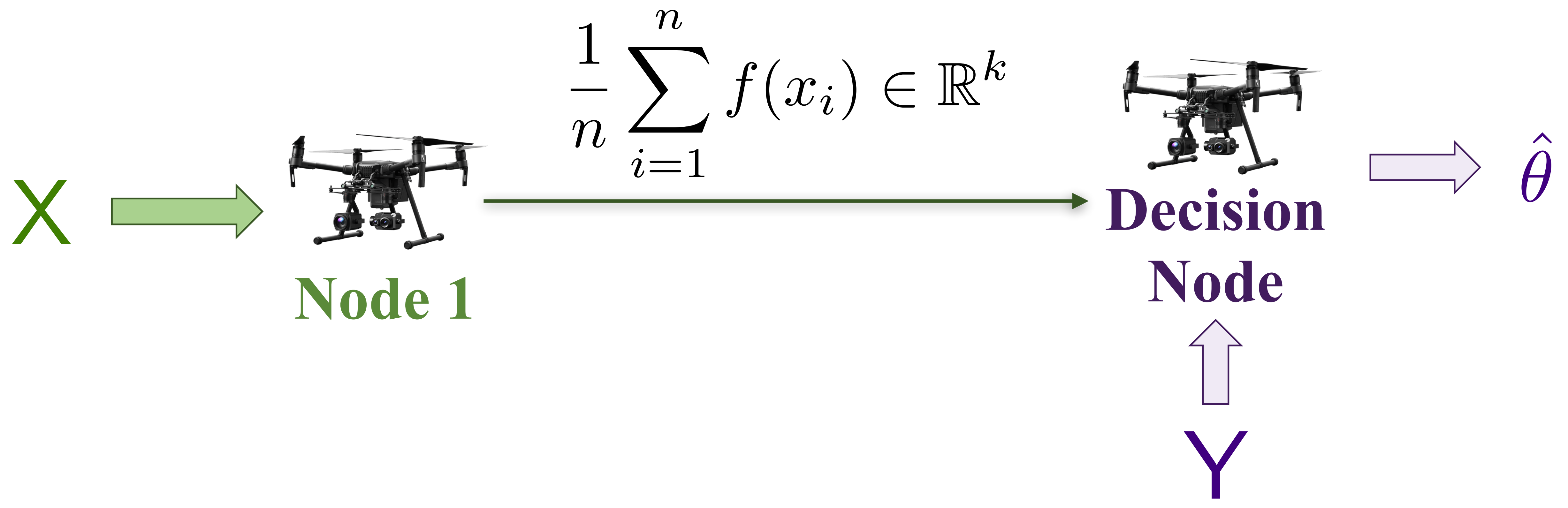
$$\mathsf{X} = (x_1, \ldots, x_n), \ \mathsf{Y} = (y_1, \ldots, y_n), \quad (x_i, y_i) \overset{\text{i.i.d.}}{\sim} P_{XY}(x, y; \theta)$$



- Node 1 transmits a statistic of its own data to the decision center.
  - The decision node estimates the parameter based on the statistic and its data.

The Two Nodes Case

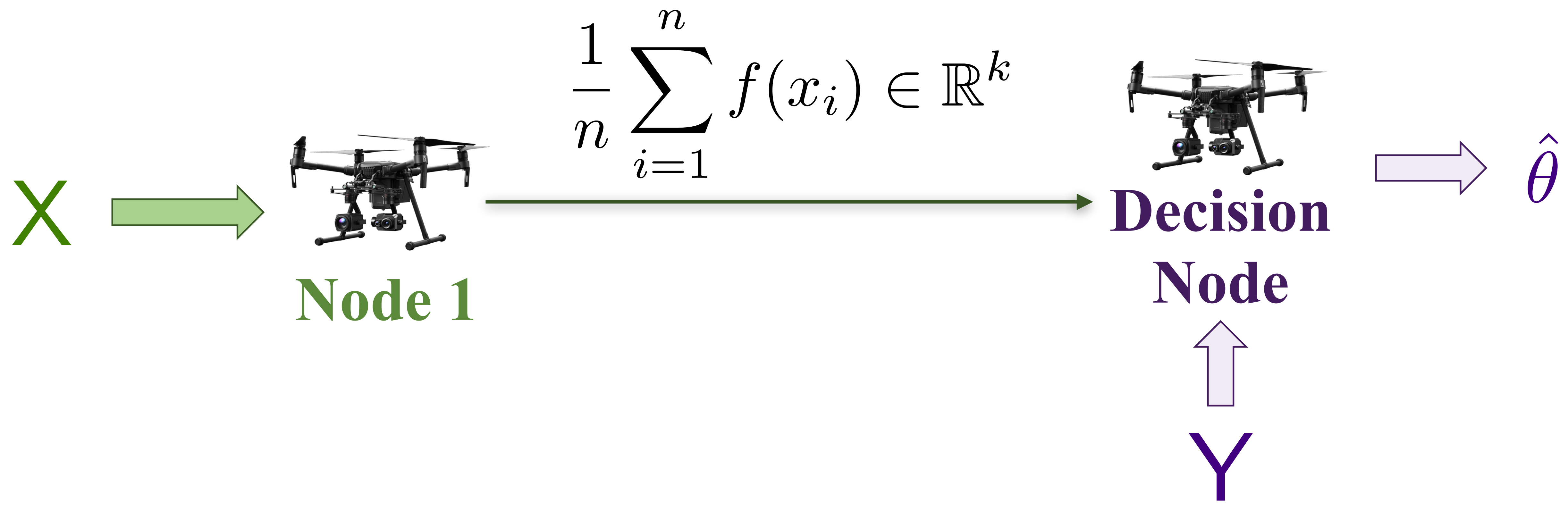
$$\mathsf{X} = (x_1, \ldots, x_n),\ \mathsf{Y} = (y_1, \ldots, y_n), \quad (x_i, y_i) \overset{\text{i.i.d.}}{\sim} P_{XY}(x, y; \theta)$$



- Node 1 transmits a statistic of its own data to the decision center.
  - The decision node estimates the parameter based on the statistic and its data.
  - The empirical distribution of the observation Y provides a reference distribution.

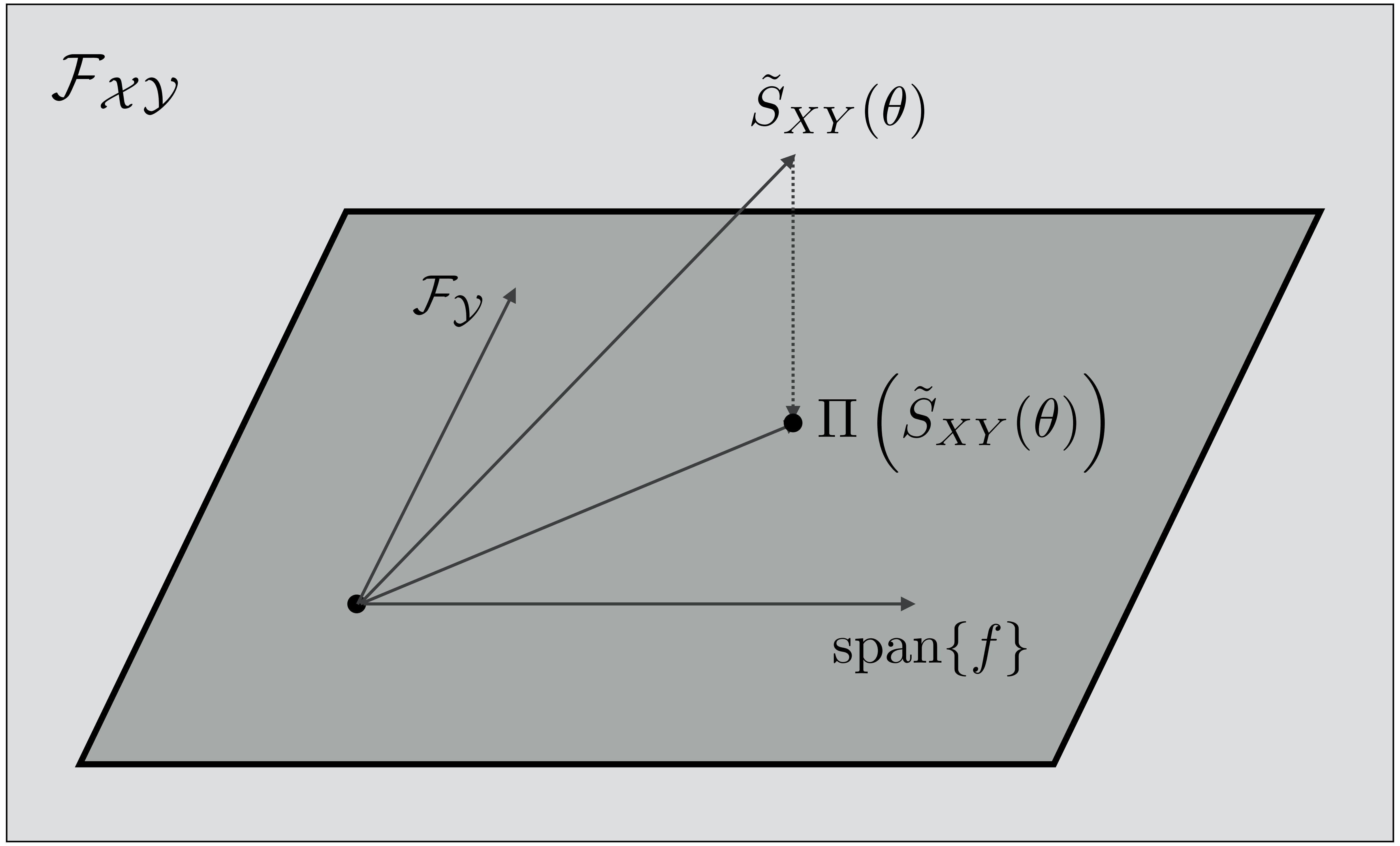
The Two Nodes Case

$$\mathsf{X} = (x_1, \ldots, x_n),\ \mathsf{Y} = (y_1, \ldots, y_n), \quad (x_i, y_i) \overset{\text{i.i.d.}}{\sim} P_{XY}(x, y; \theta)$$



- Node 1 transmits a statistic of its own data to the decision center.
  - The decision node estimates the parameter based on the statistic and its data.
  - The empirical distribution of the observation Y provides a reference distribution.
- What is the informative k-dimensional feature function node 1 should extract?

The Geometric Interpretation



$\mathcal{F}_{\mathcal{X}\mathcal{Y}}$: The functional subspace of X, Y, $\mathcal{F}_{\mathcal{Y}}$: The functional subspace of Y.

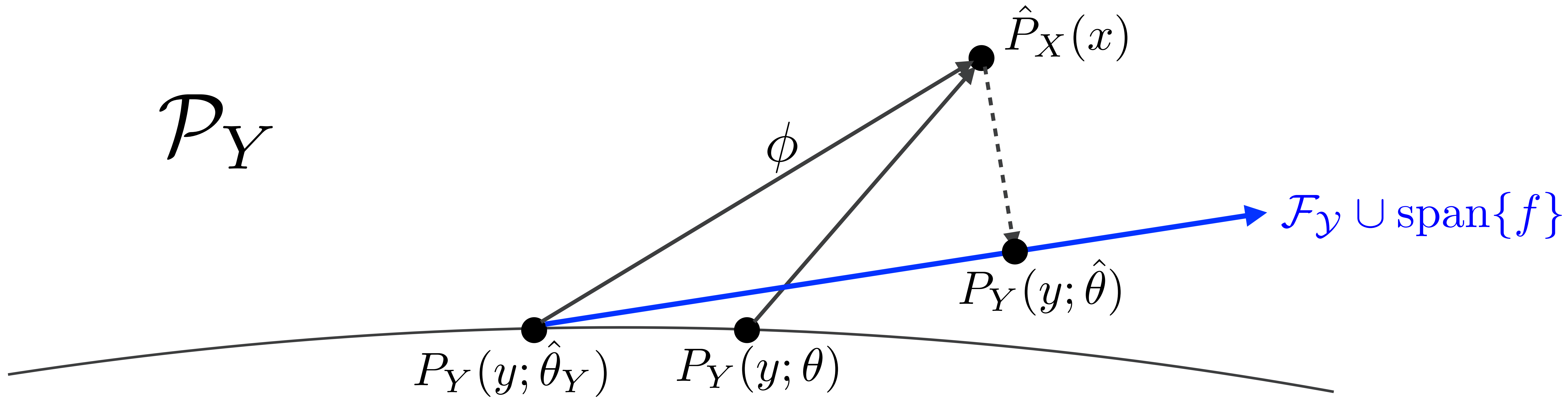
span$\{f\}$: The functional space spanned by each dimension of f.

$$\text{MSE} = \frac{1}{n}\text{tr}\left\{\left(\tilde{S}_X^{\mathbf{T}}(\theta)V^{\mathbf{T}}\left(VV^{\mathbf{T}}\right)^{-1}V\tilde{S}_X(\theta)\right)^{-1}\right\}, \text{ where } V = [\mathbf{B}_Y \;\; \mathbf{F}]$$

The projection operator from (X, Y) onto Y

[1] X. Tong, S.-L. Huang, "An Information Theoretic Approach for Collaborative Distributed Parameter Estimation," IEEE International Symposium on Information Theory (ISIT), June, 2023.

The Estimator To Achieve Optimal MSE



$$\hat{\theta}_Y = \arg\max_{\theta} \frac{1}{n}\sum_{i=1}^{n} \log P_Y(y_i;\theta)$$

The estimator: $$\hat{\theta} = \hat{\theta}_Y + J^{-1}(\hat{\theta}_Y)\underbrace{\tilde{S}_X^{\mathbf{T}}(\hat{\theta}_Y)V^{\mathbf{T}}\left(VV^{\mathbf{T}}\right)^{-1}}_{\text{projection operation}} \cdot \begin{bmatrix} \phi \\ b \end{bmatrix}$$

Eliminate the bias from the difference between $\hat{\theta}_Y$ and θ.

$$b = \left(\mathbf{F}^{\mathbf{T}}\mathbf{F}\right)^{-\frac{1}{2}} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}_{P_X(\cdot;\hat{\theta}_Y)}[f(X)]\right)$$

[1] X. Tong, S.-L. Huang, "An Information Theoretic Approach for Collaborative Distributed Parameter Estimation," IEEE International Symposium on Information Theory (ISIT), June, 2023.

The Special Case

- Suppose that $P_{XY}(x,y;\theta) = P_X(x;\theta)P_Y(y;\theta)$ then the MSE can be reduced to

$$\mathrm{MSE} = \frac{1}{n}\mathrm{tr}\left\{\left(\underbrace{J_Y(\theta)}_{\text{Fisher information of } Y} + \underbrace{\tilde{S}_X^{\mathbf{T}}(\theta)\mathbf{F}^{\mathbf{T}}\left(\mathbf{F}\mathbf{F}^{\mathbf{T}}\right)^{-1}\mathbf{F}\tilde{S}_X(\theta)}_{\text{Fisher information of } f}\right)^{-1}\right\}$$

[1] X. Tong, S.-L. Huang, "An Information Theoretic Approach for Collaborative Distributed Parameter Estimation," IEEE International Symposium on Information Theory (ISIT), June, 2023.

The Special Case

- Suppose that $P_{XY}(x,y;\theta) = P_X(x;\theta)P_Y(y;\theta)$ then the MSE can be reduced to

$$\mathrm{MSE} = \frac{1}{n}\mathrm{tr}\left\{\left(\underbrace{J_Y(\theta)}_{\text{Fisher information of } Y} + \underbrace{\tilde{S}_X^{\mathbf{T}}(\theta)\mathbf{F}^{\mathbf{T}}\left(\mathbf{F}\mathbf{F}^{\mathbf{T}}\right)^{-1}\mathbf{F}\tilde{S}_X(\theta)}_{\text{Fisher information of } f}\right)^{-1}\right\}$$

- The performance gain provided by the statistics of the *x* sequence.

[1] X. Tong, S.-L. Huang, "An Information Theoretic Approach for Collaborative Distributed Parameter Estimation," IEEE International Symposium on Information Theory (ISIT), June, 2023.

The Special Case

- Suppose that $P_{XY}(x,y;\theta) = P_X(x;\theta)P_Y(y;\theta)$ then the MSE can be reduced to

$$\mathrm{MSE} = \frac{1}{n}\mathrm{tr}\left\{\left(\underbrace{J_Y(\theta)}_{\text{Fisher information of } Y} + \underbrace{\tilde{S}_X^{\mathbf{T}}(\theta)\mathbf{F}^{\mathbf{T}}\left(\mathbf{F}\mathbf{F}^{\mathbf{T}}\right)^{-1}\mathbf{F}\tilde{S}_X(\theta)}_{\text{Fisher information of } f}\right)^{-1}\right\}$$

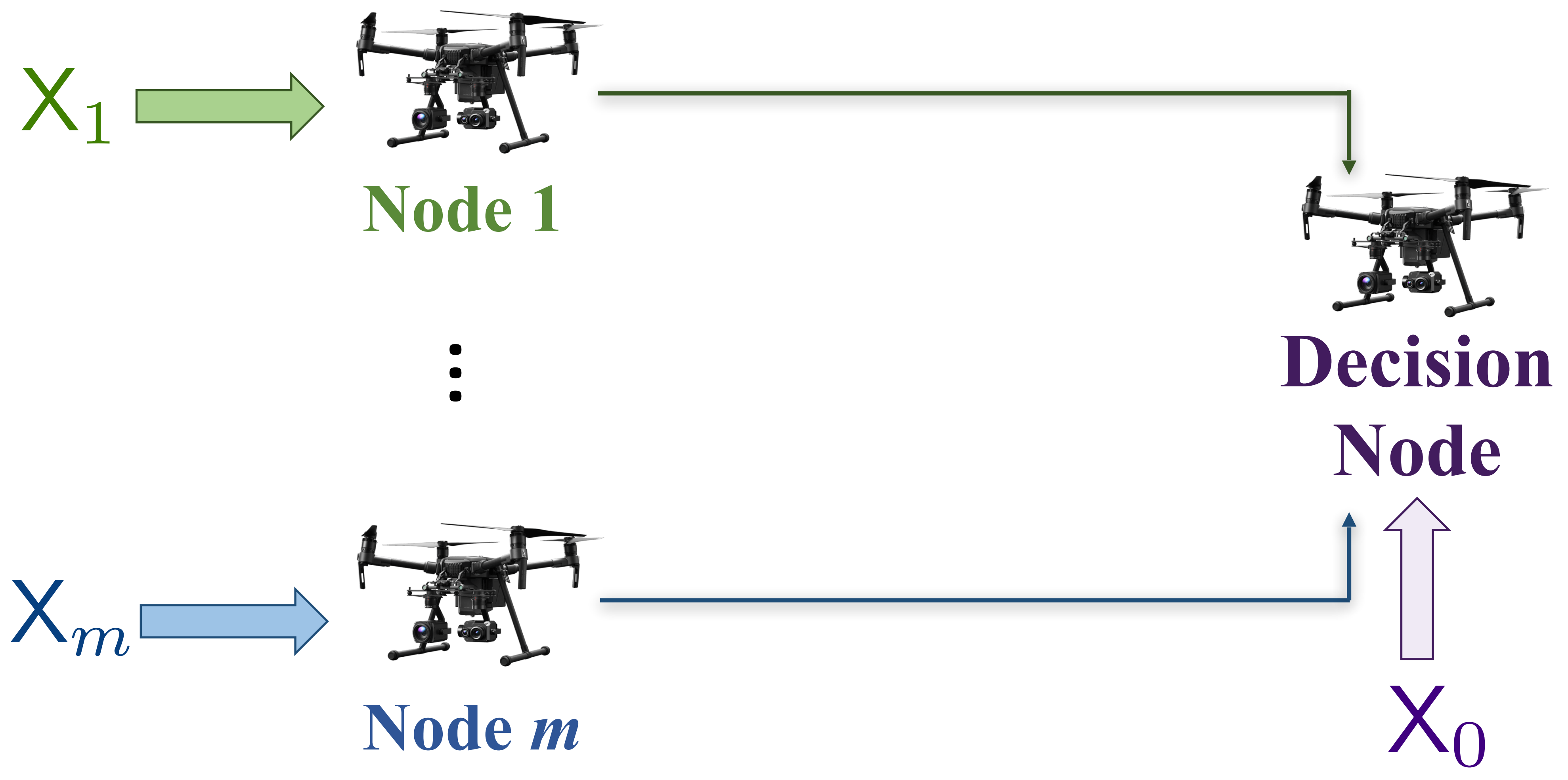
  - The performance gain provided by the statistics of the x sequence.
  - The top-k singular vectors of the Fisher information matrix = The most informative feature functions.

[1] X. Tong, S.-L. Huang, "An Information Theoretic Approach for Collaborative Distributed Parameter Estimation," IEEE International Symposium on Information Theory (ISIT), June, 2023.

Collaborative Distributed Parameter Estimation

$$\mathsf{X}_i = (x_i^{(1)}, \ldots, x_i^{(n)}), \quad i = 1, \ldots, m,$$

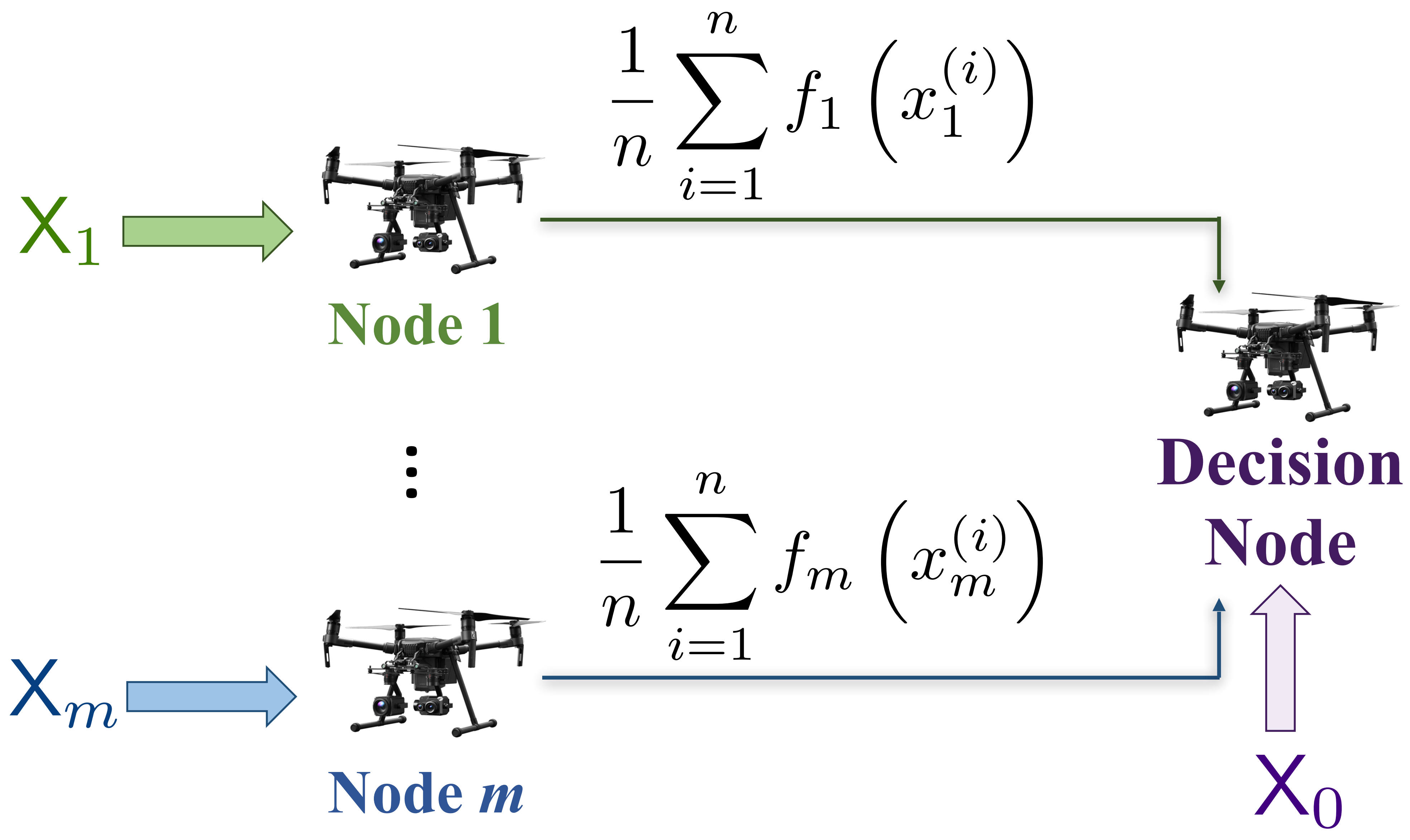
$$(x_0^{(j)}, \ldots, x_m^{(j)}) \overset{\text{i.i.d.}}{\sim} P_{X_0 \cdots X_m}(x_1, \ldots, x_m; \theta), \quad j = 1, \ldots, n.$$



Collaborative Distributed Parameter Estimation

$$\mathsf{X}_i = (x_i^{(1)}, \ldots, x_i^{(n)}), \quad i = 1, \ldots, m,$$

$$(x_0^{(j)}, \ldots, x_m^{(j)}) \overset{\text{i.i.d.}}{\sim} P_{X_0 \cdots X_m}(x_1, \ldots, x_m; \theta), \quad j = 1, \ldots, n.$$

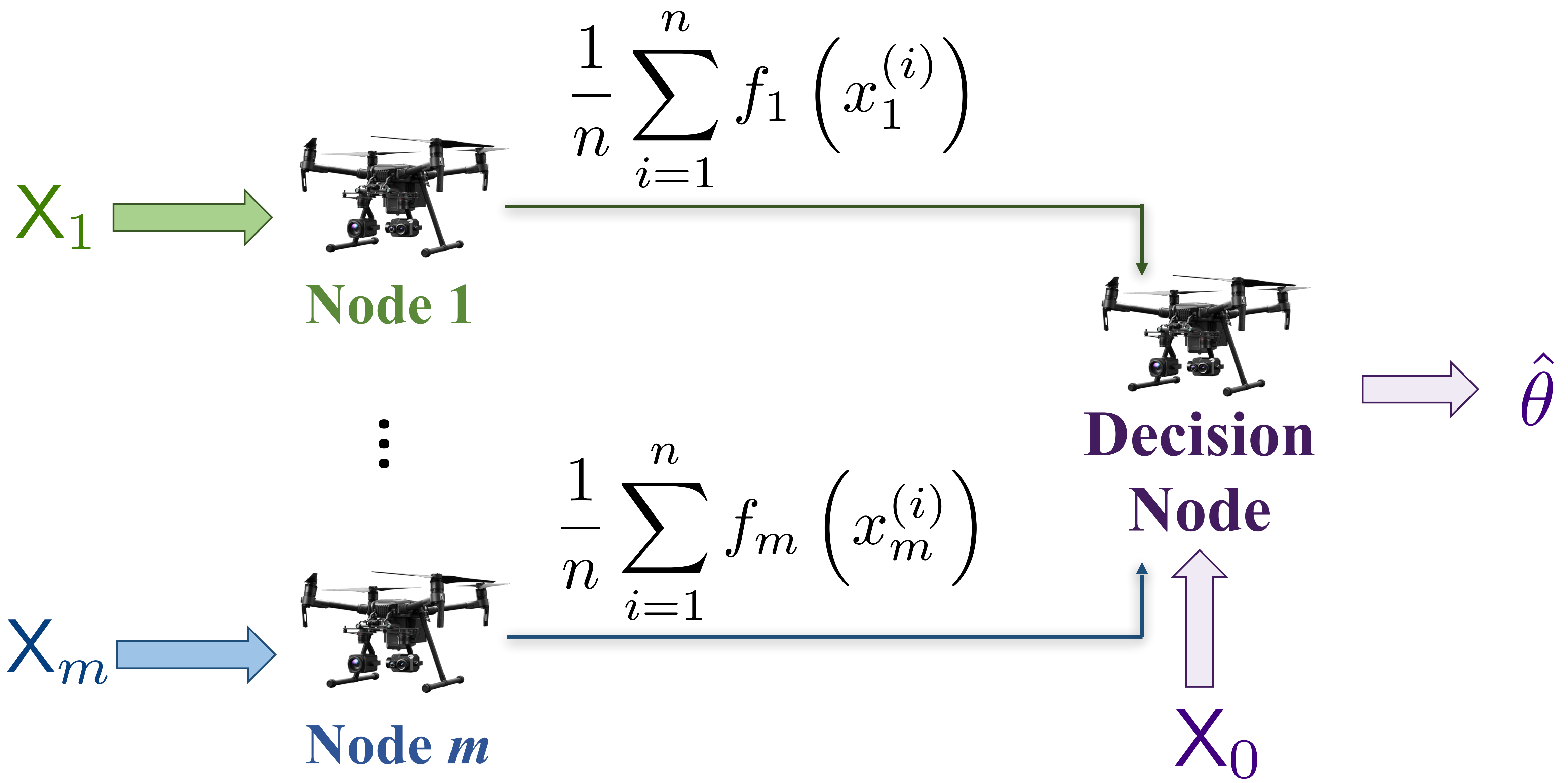


- Each node *i* transmit a statistic of its own data to the decision center.

Collaborative Distributed Parameter Estimation

$$\mathsf{X}_i = (x_i^{(1)}, \ldots, x_i^{(n)}), \quad i = 1, \ldots, m,$$

$$(x_0^{(j)}, \ldots, x_m^{(j)}) \overset{\text{i.i.d.}}{\sim} P_{X_0 \cdots X_m}(x_1, \ldots, x_m; \theta), \quad j = 1, \ldots, n.$$

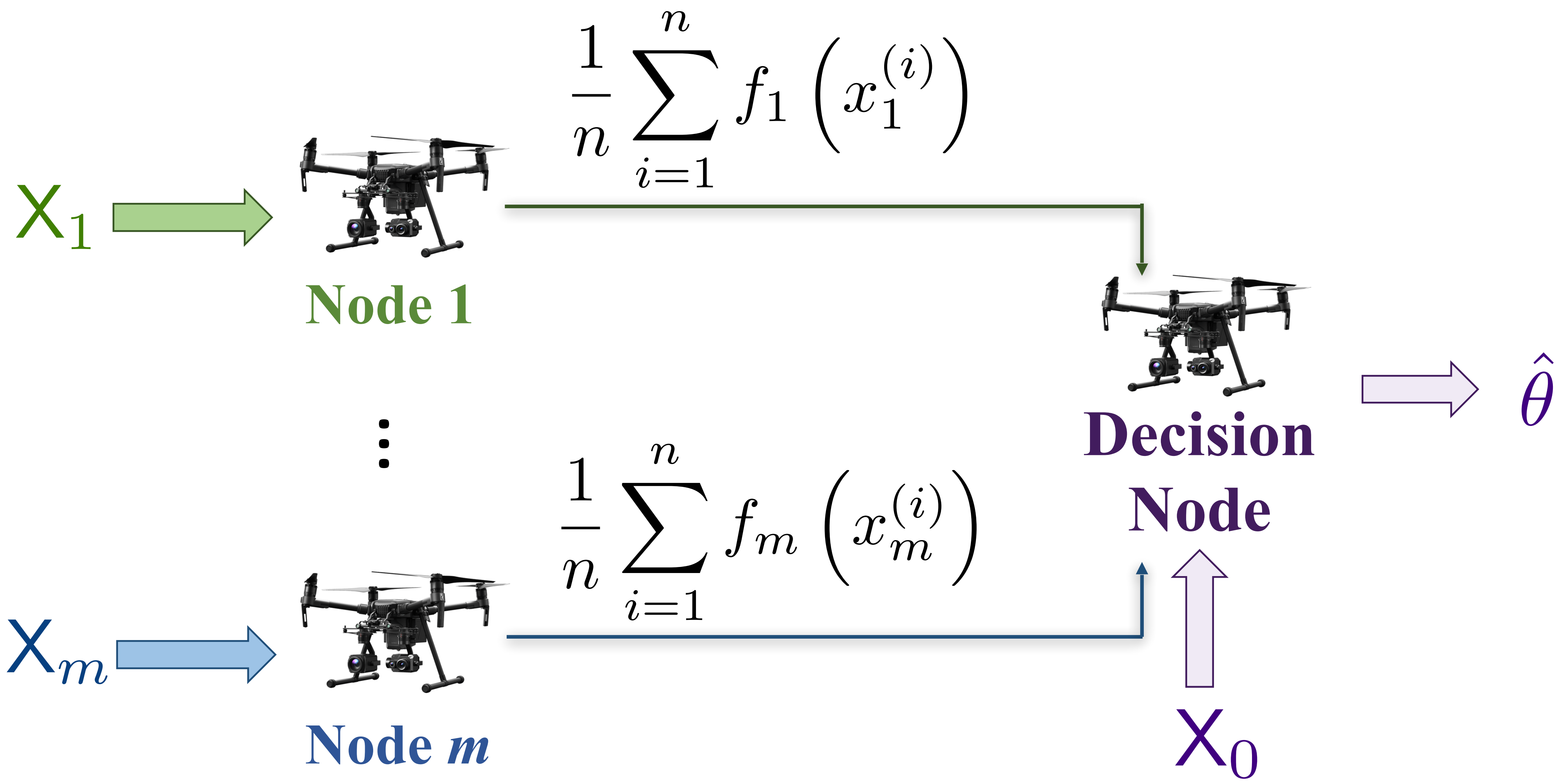


- Each node i transmit a statistic of its own data to the decision center.
  - The decision node estimate the parameter based on the statistics and its data.

Collaborative Distributed Parameter Estimation

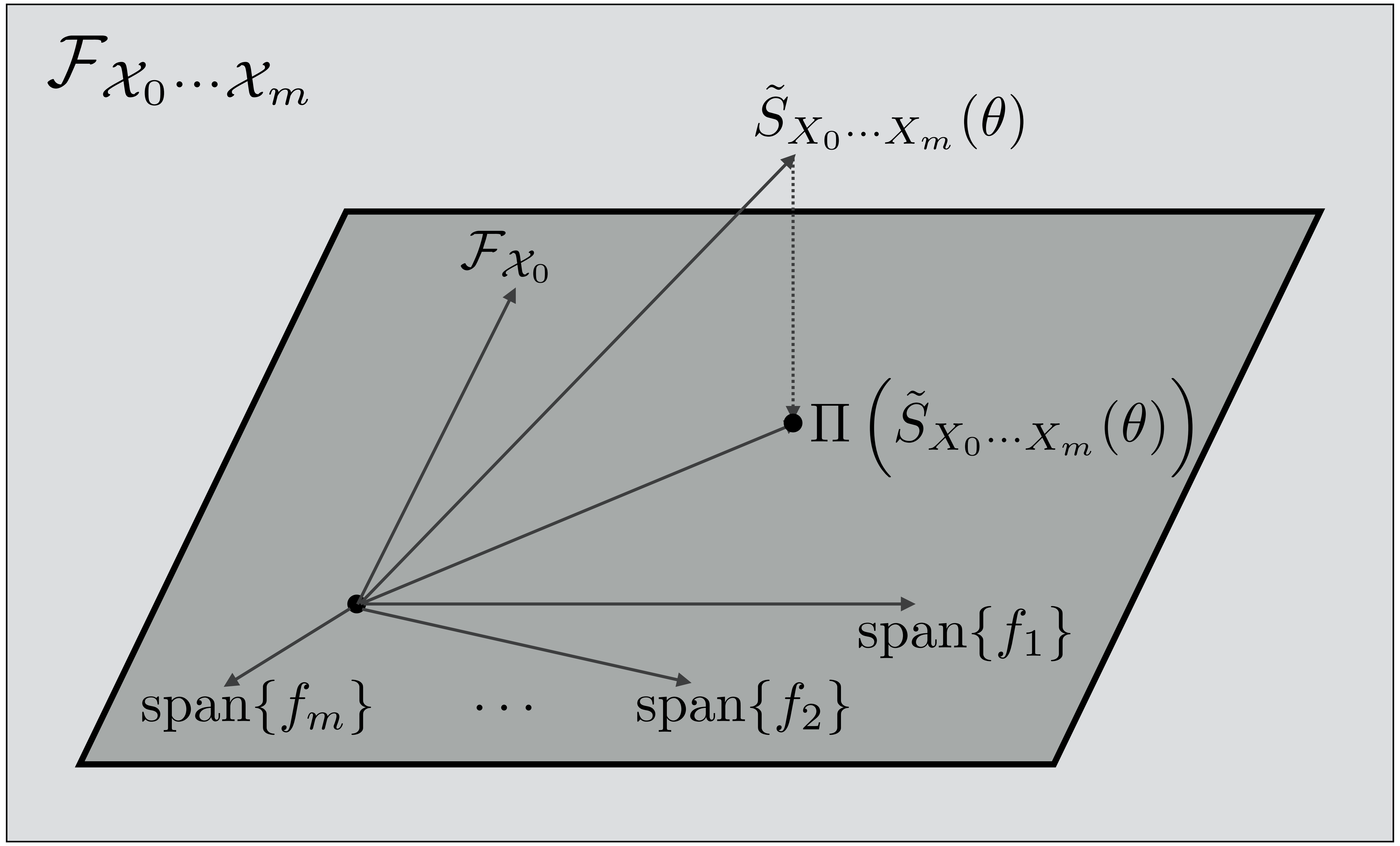
$$\mathsf{X}_i = (x_i^{(1)}, \ldots, x_i^{(n)}), \quad i = 1, \ldots, m,$$

$$(x_0^{(j)}, \ldots, x_m^{(j)}) \overset{\text{i.i.d.}}{\sim} P_{X_0 \cdots X_m}(x_1, \ldots, x_m; \theta), \quad j = 1, \ldots, n.$$



- Each node *i* transmit a statistic of its own data to the decision center.
  - The decision node estimate the parameter based on the statistics and its data.
- What are informative feature functions $f_i : \mathcal{X}_i \mapsto \mathbb{R}^{k_i}$ the nodes should extract?

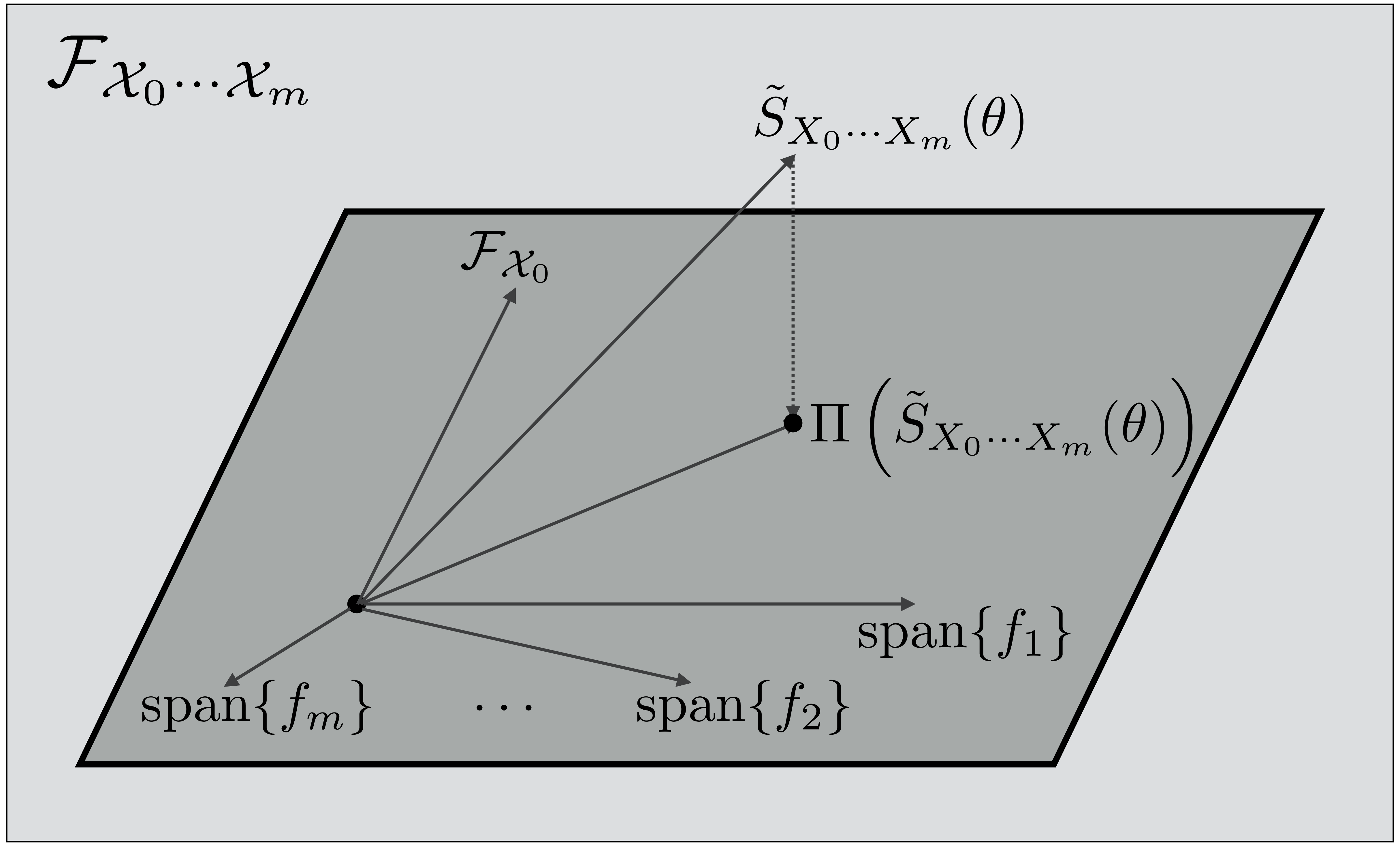
The Geometric Interpretation



$$\mathrm{MSE} = \mathrm{tr}\left\{\left(\tilde{S}_X^{\mathbf{T}}(\theta)V^{\mathbf{T}}\left(VV^{\mathbf{T}}\right)^{-1}V\tilde{S}_X(\theta)\right)^{-1}\right\}, \text{ where } V = [\mathbf{B}_Y\ \mathbf{F}_1\ \cdots\ \mathbf{F}_{m-1}]$$

[1] X. Tong, S.-L. Huang, "An Information Theoretic Approach for Collaborative Distributed Parameter Estimation," IEEE International Symposium on Information Theory (ISIT), June, 2023.

The Geometric Interpretation

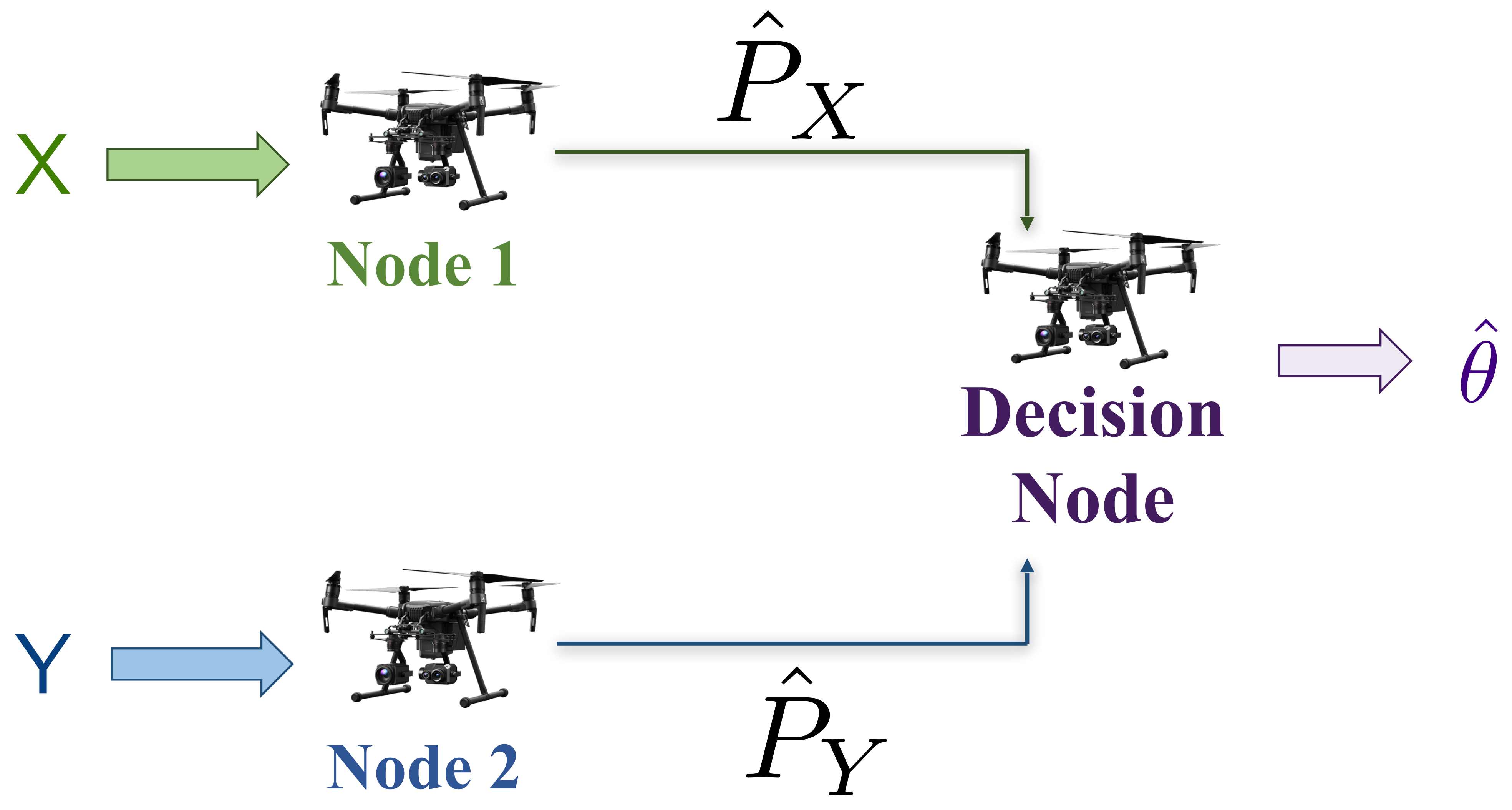


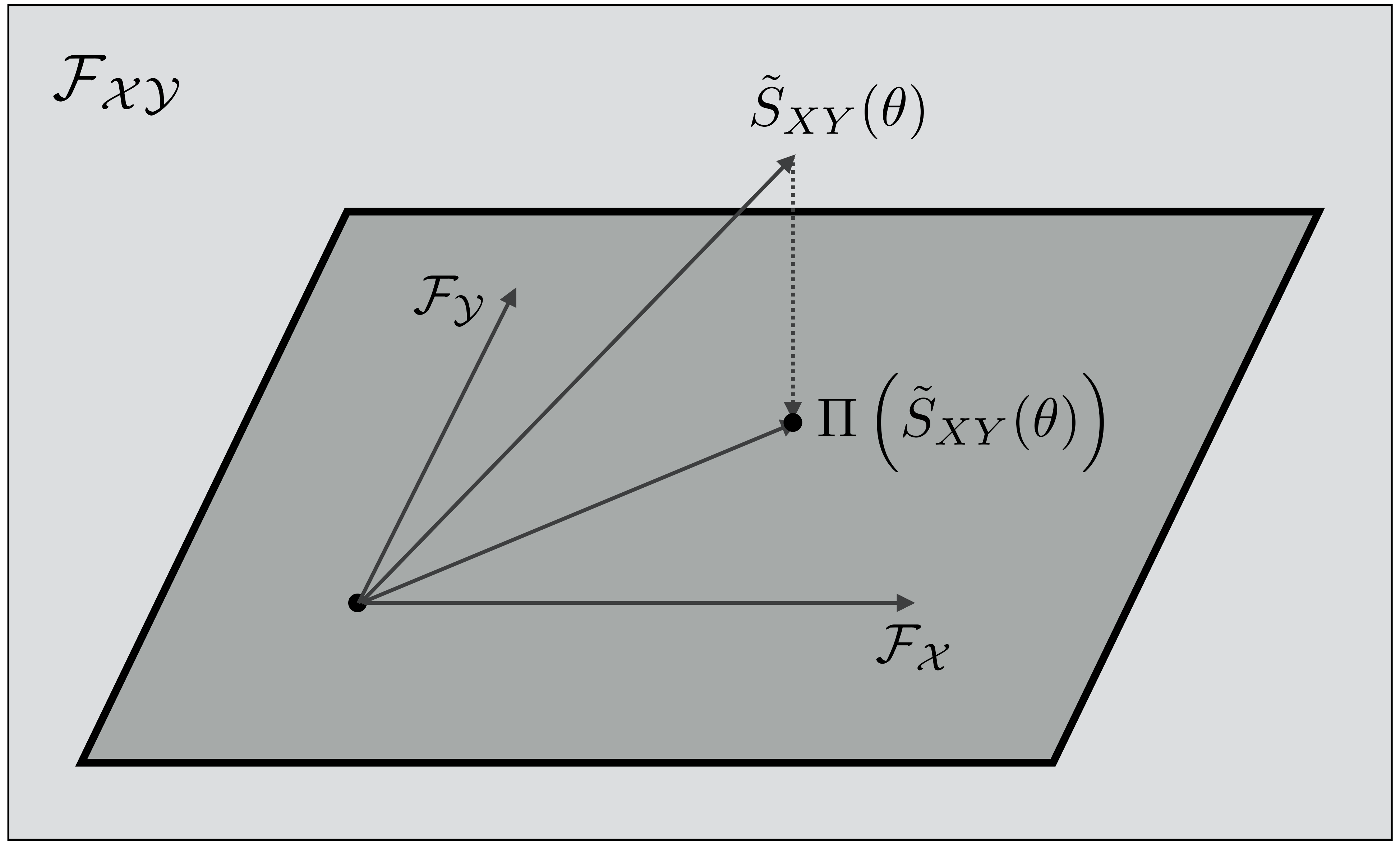
$$\text{MSE} = \text{tr}\left\{\left(\tilde{S}_X^{\mathbf{T}}(\theta)V^{\mathbf{T}}\left(VV^{\mathbf{T}}\right)^{-1}V\tilde{S}_X(\theta)\right)^{-1}\right\}, \text{ where } V = [\mathbf{B}_Y \ \mathbf{F}_1 \ \cdots \ \mathbf{F}_{m-1}]$$

- When X is continuous, parametrizing the functions by neural networks to optimize the MSE loss.

[1] X. Tong, S.-L. Huang, "An Information Theoretic Approach for Collaborative Distributed Parameter Estimation," IEEE International Symposium on Information Theory (ISIT), June, 2023.

Communicating The Type of Data





- Statistics of data = linear functions of types
  - Communicating type of data = projecting to the functional subspace.
  - Distributed parameter estimation with O(log n) bits communication constraint [1].
    - Geometric interpretation to the classical result.

[1] T. Han and S. Amari, "Statistical Inference Under Multiterminal Data Compression," IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2300–2324, 1998.

TBSI 清华–伯克利深圳学院 Tsinghua-Berkeley Shenzhen Institute

Thank you!