Robust Distributed Learning

Challenges of Data Heterogeneity and Privacy

Nirupam Gupta and Rafael Pinot
Distributed Computing Laboratory
EPFL

Workshop on **Principles of Distributed Learning (PODL)** at **DISC'23**

Content

- General Lower Bound under Heterogeneity
    - Implications in learning
    - Optimal robustness strategy
    - Applicability to robust state estimation
- Characterizing Heterogeneity for First-Order Methods
    - (G, B)-Gradient dissimilarity
    - Impact of condition number
- Differential Privacy in Distributed SGD
    - Distributed (ϵ, δ)-DP
    - Synthesis with robustness

Challenge of Data Heterogeneity

Resilience Property

(f, ε) – Resilience

Despite f adversarial nodes, output an ε-suboptimal solution to ERM over the training samples of honest nodes.

$$\mathscr{L}_H\left(\hat{\theta}\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}_H(\theta) \leq \varepsilon$$

Resilience Property

(f, ε) – Resilience

Despite f adversarial nodes, output an ε-suboptimal solution to ERM over the training samples of honest nodes.

$$\mathscr{L}_H\left(\hat{\theta}\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}_H(\theta) \leq \varepsilon$$

When the loss is defined by the indicator function, in the case of classification, ε is the additional fraction of misclassified samples

Impossibility under Non-identical Data

Impossibility under Non-identical Data

In general, local training samples of the nodes are different

$$\mathscr{L}_i \neq \mathscr{L}_j$$

Impossibility under Non-identical Data

In general, local training samples of the nodes are different

$$\mathcal{L}_i \neq \mathcal{L}_j$$

It is impossible to achieve (f, ε)-Resilience, due to the anonymity of adversarial nodes

Impossibility under Non-identical Data

In general, local training samples of the nodes are different

$$\mathscr{L}_i \neq \mathscr{L}_j$$

"Approximate Fault-Tolerance in Distributed Optimization." S. Liu et al., PODC'21

It is impossible to achieve (f, ε)-Resilience, due to the anonymity of adversarial nodes

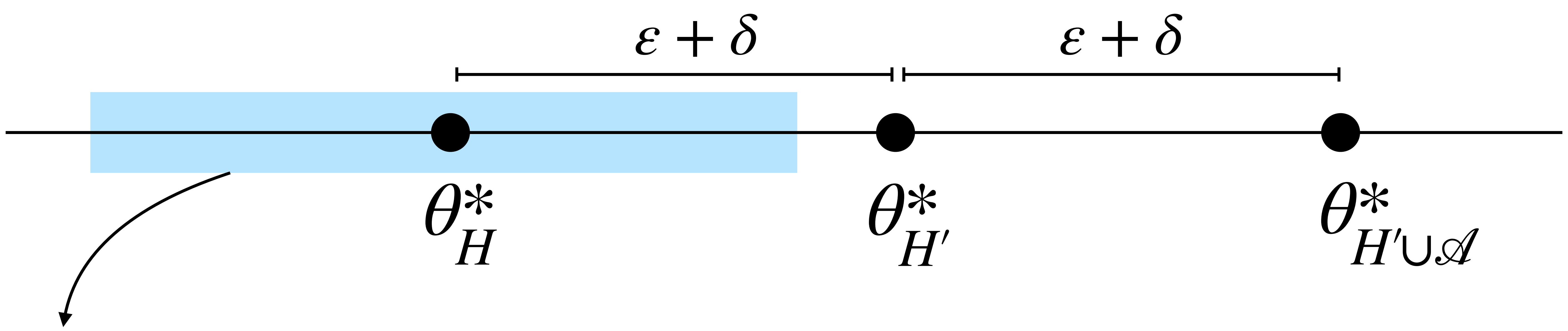
Impossibility under Non-identical Data

In general, local training samples of the nodes are different

$$\mathscr{L}_i \neq \mathscr{L}_j$$

"Approximate Fault-Tolerance in Distributed Optimization." S. Liu et al., PODC'21

It is impossible to achieve (f, ε)-Resilience, due to the anonymity of adversarial nodes



$[\theta^*_H - \varepsilon, \theta^*_H + \varepsilon]$

$H \rightarrow (n - f)$ nodes

$H' \rightarrow (n - 2f)$ nodes

$\mathscr{A} \rightarrow f$ nodes

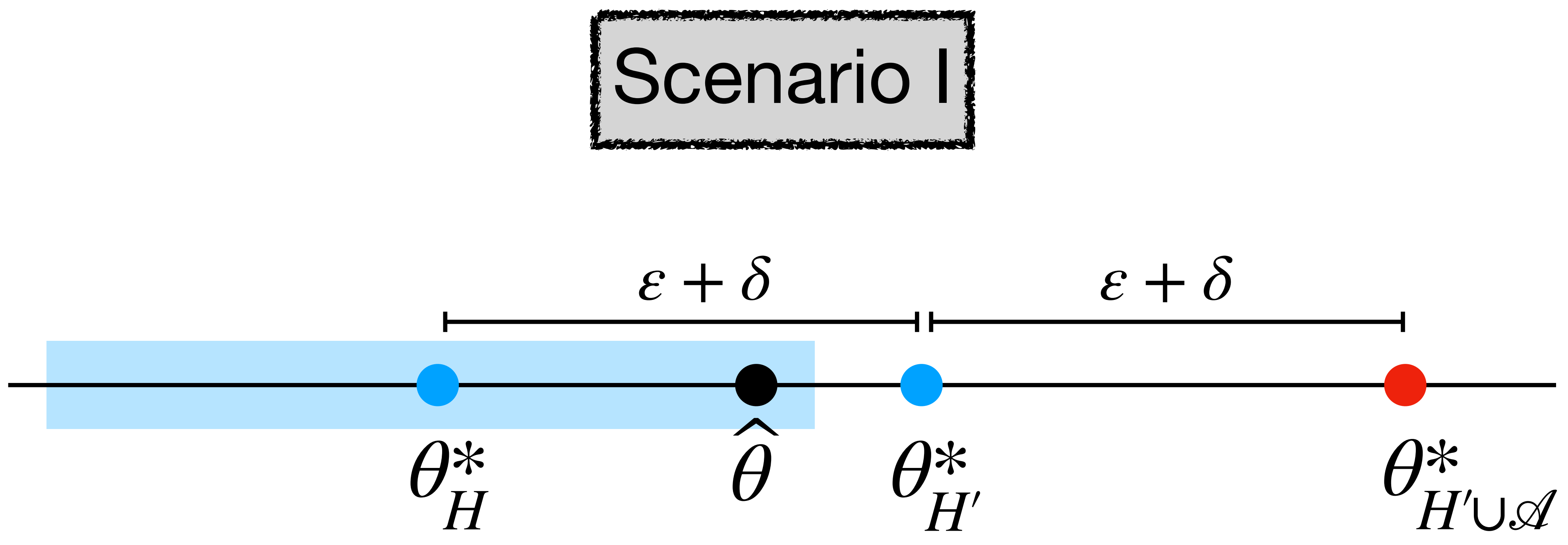
$\theta^*_S := \arg\min \mathscr{L}_S(\theta)$

Indistinguishable Executions

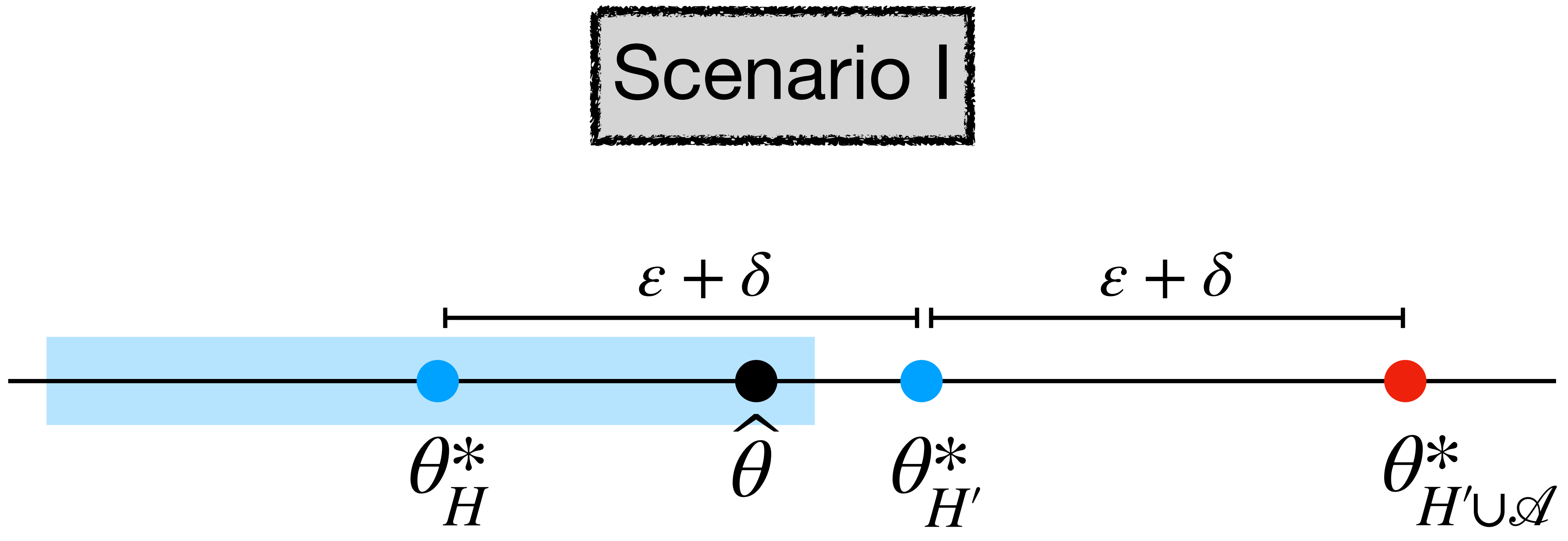
Indistinguishable Executions

Scenario I

Indistinguishable Executions



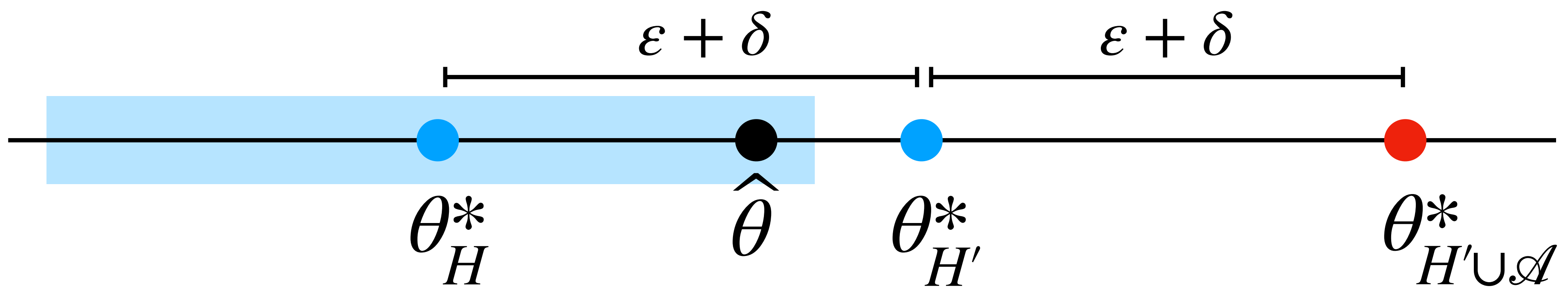
Indistinguishable Executions



Set of honest nodes is H

Indistinguishable Executions

Scenario I

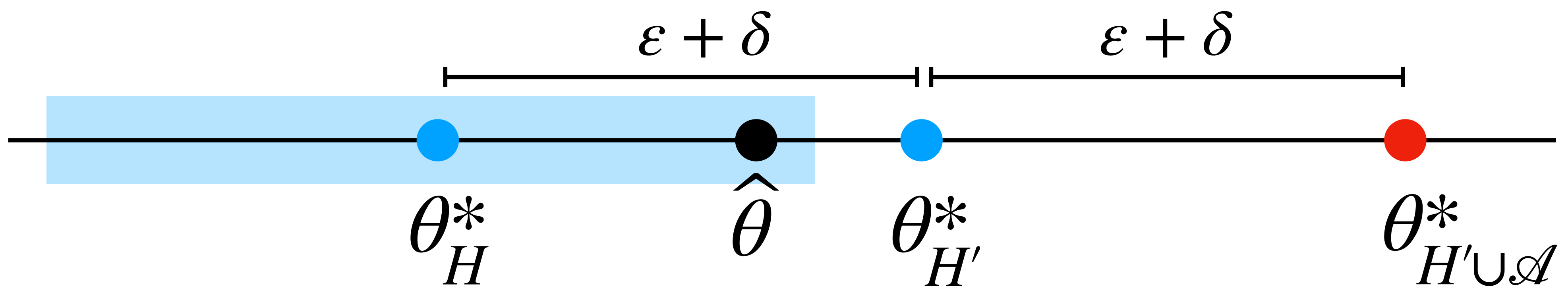


Set of honest nodes is H

$\widehat{\theta} \in [\theta^*_H - \varepsilon,\, \theta^*_H + \varepsilon]$

Indistinguishable Executions

Scenario I

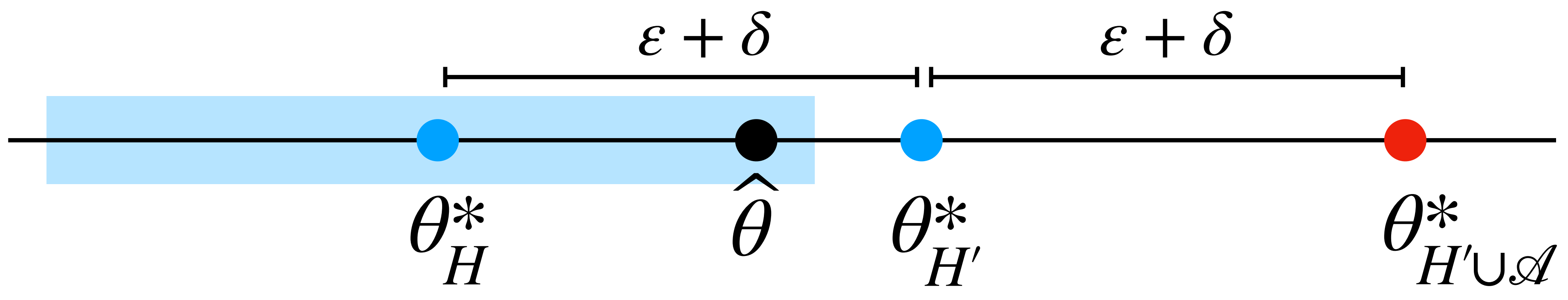


Set of honest nodes is H

$$\widehat{\theta} \in [\theta^*_H - \varepsilon, \theta^*_H + \varepsilon]$$

Indistinguishable Executions

Scenario I



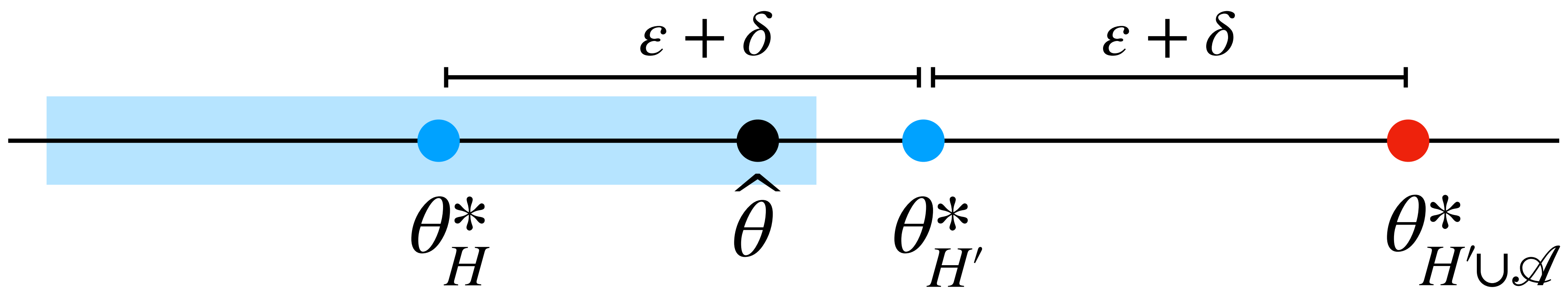
Set of honest nodes is H

$$\widehat{\theta} \in [\theta^*_H - \varepsilon, \theta^*_H + \varepsilon]$$

Scenario II

Indistinguishable Executions

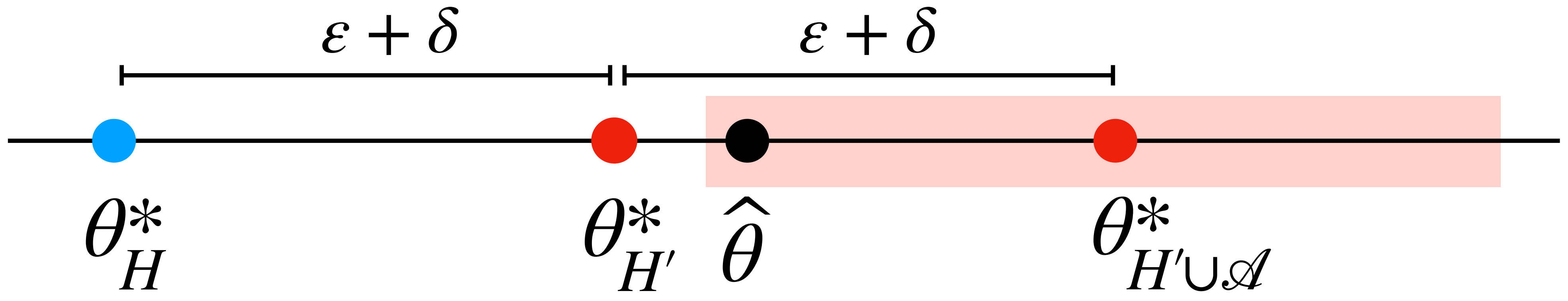
Scenario I



Set of honest nodes is H

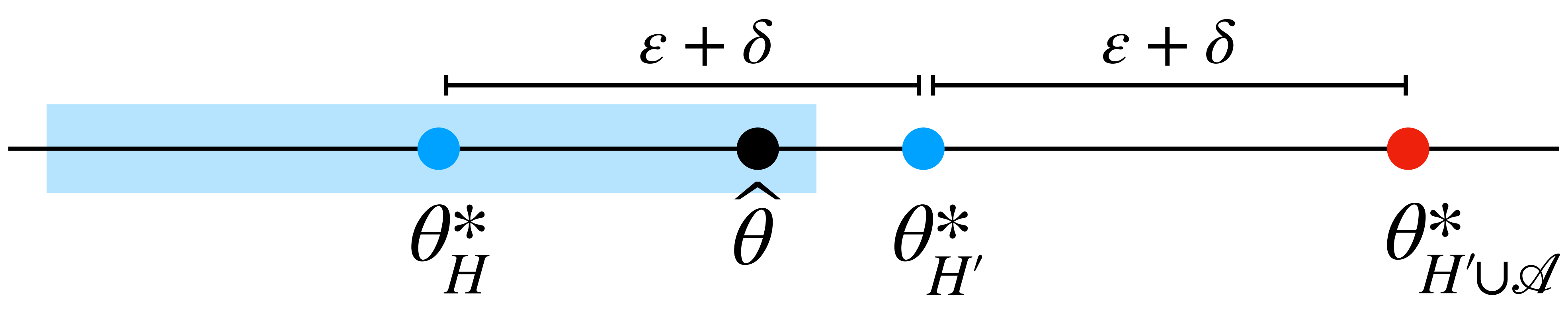
$$\widehat{\theta} \in [\theta^*_H - \varepsilon,\, \theta^*_H + \varepsilon]$$

Scenario II



Indistinguishable Executions

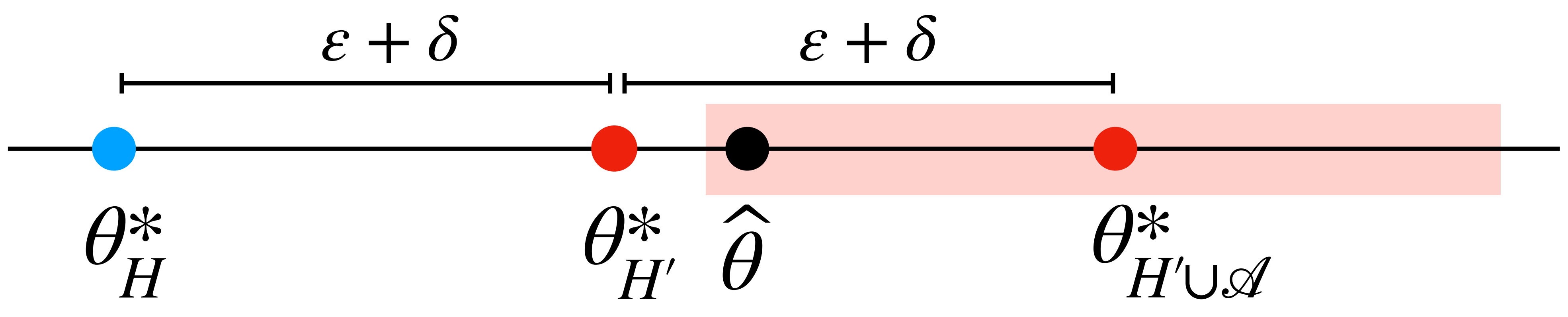
Scenario I



Set of honest nodes is H

$\widehat{\theta} \in [\theta^*_H - \varepsilon, \theta^*_H + \varepsilon]$

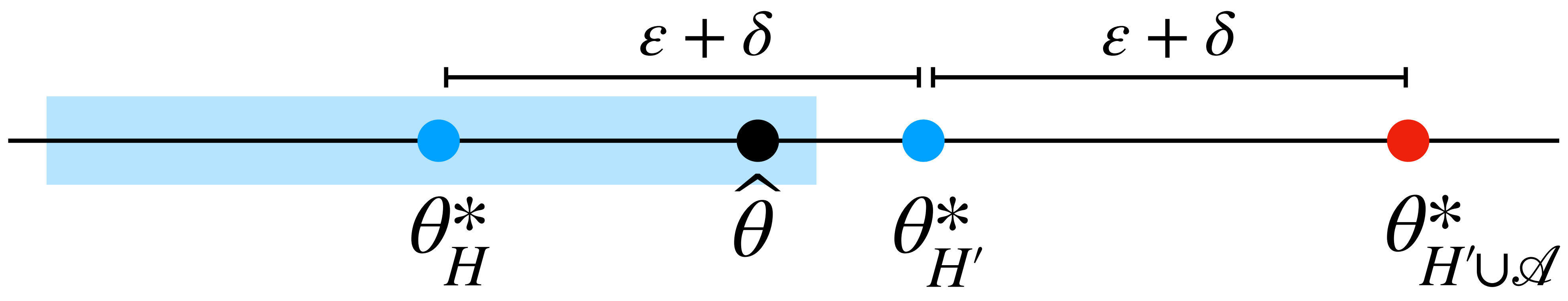
Scenario II



Set of honest nodes is $H' \cup \mathcal{A}$

Indistinguishable Executions

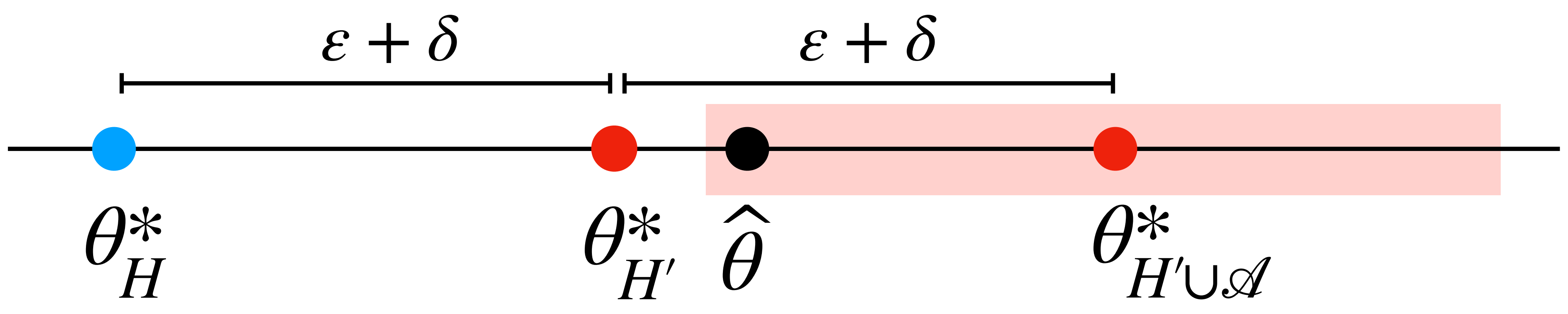
Scenario I



Set of honest nodes is H

$\hat{\theta} \in [\theta^*_H - \varepsilon, \, \theta^*_H + \varepsilon]$

Scenario II

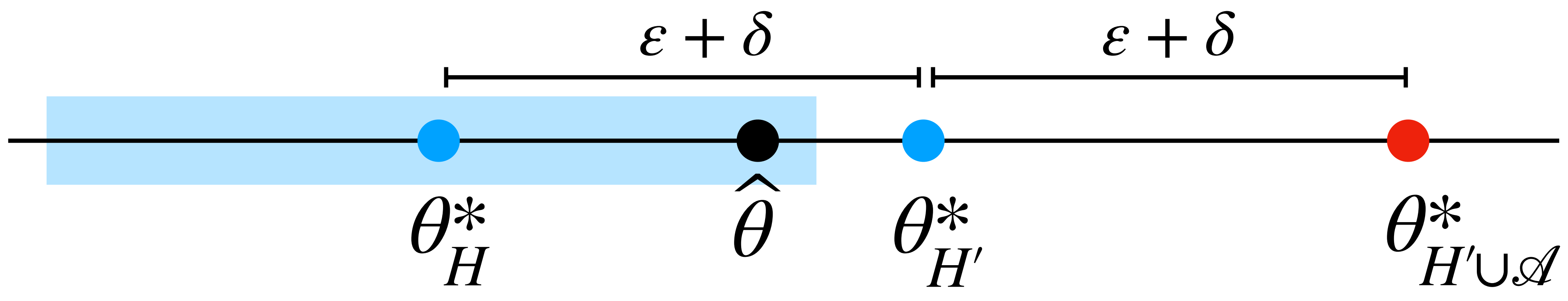


Set of honest nodes is $H' \cup \mathscr{A}$

$\hat{\theta} \in [\theta^*_{H' \cup \mathscr{A}} - \varepsilon, \, \theta^*_{H' \cup \mathscr{A}} + \varepsilon]$

Indistinguishable Executions

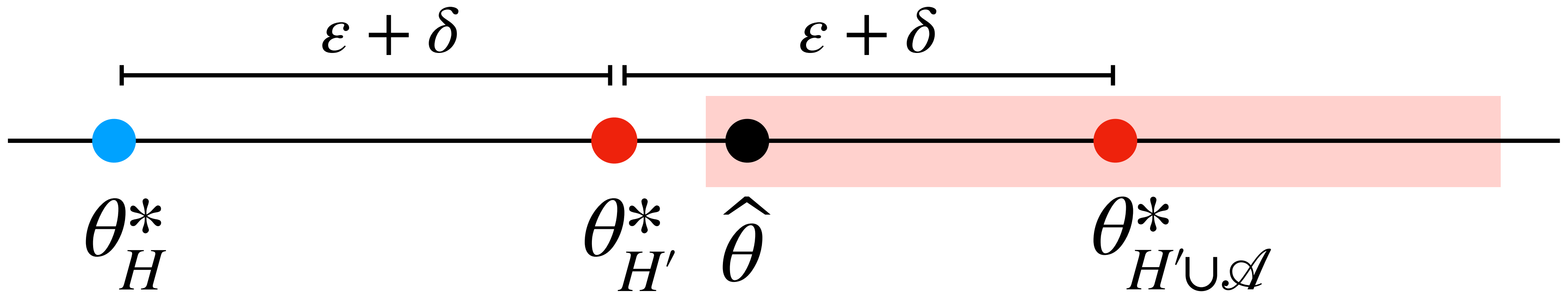
Scenario I



Set of honest nodes is H

$\widehat{\theta} \in [\theta^*_H - \varepsilon, \theta^*_H + \varepsilon]$

Scenario II



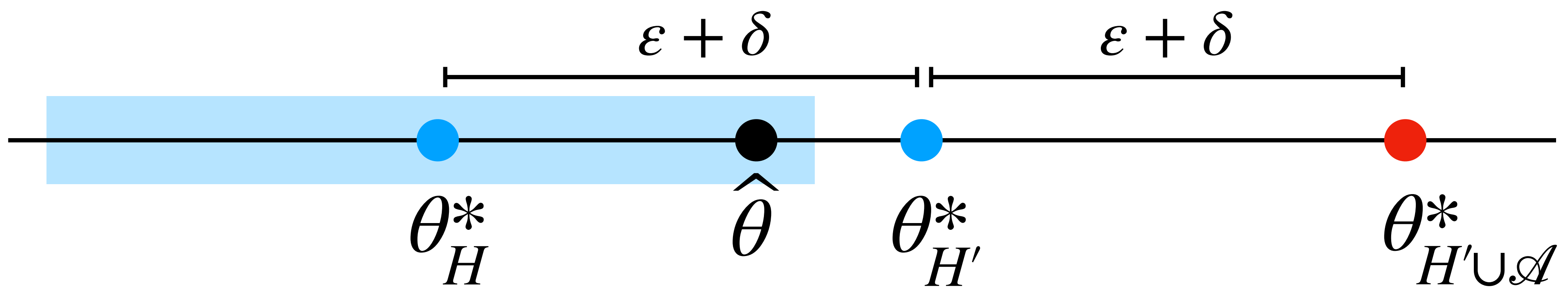
Set of honest nodes is $H' \cup \mathscr{A}$

$\widehat{\theta} \in [\theta^*_{H' \cup \mathscr{A}} - \varepsilon, \theta^*_{H' \cup \mathscr{A}} + \varepsilon]$

Which scenario is the correct one?

Indistinguishable Executions

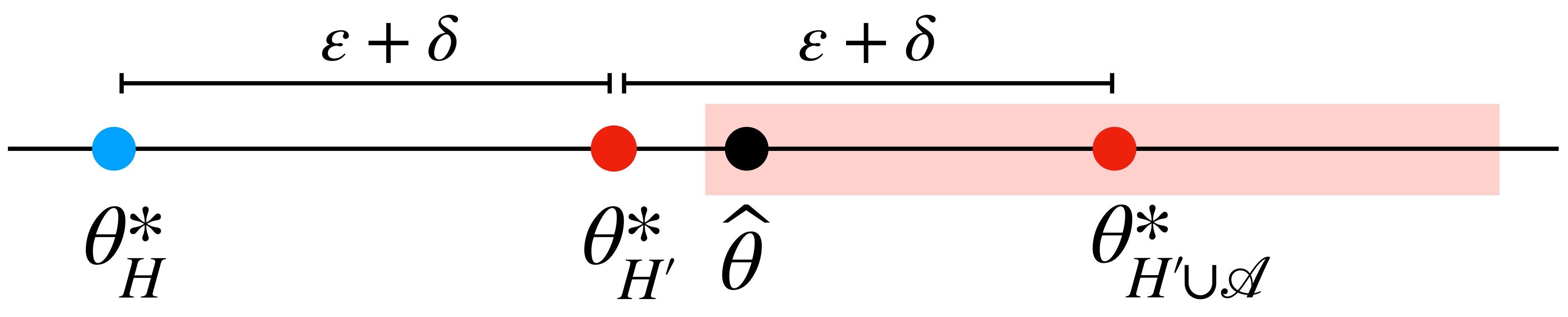
Scenario I



Set of honest nodes is H

$\widehat{\theta} \in [\theta^*_H - \varepsilon,\, \theta^*_H + \varepsilon]$

Scenario II



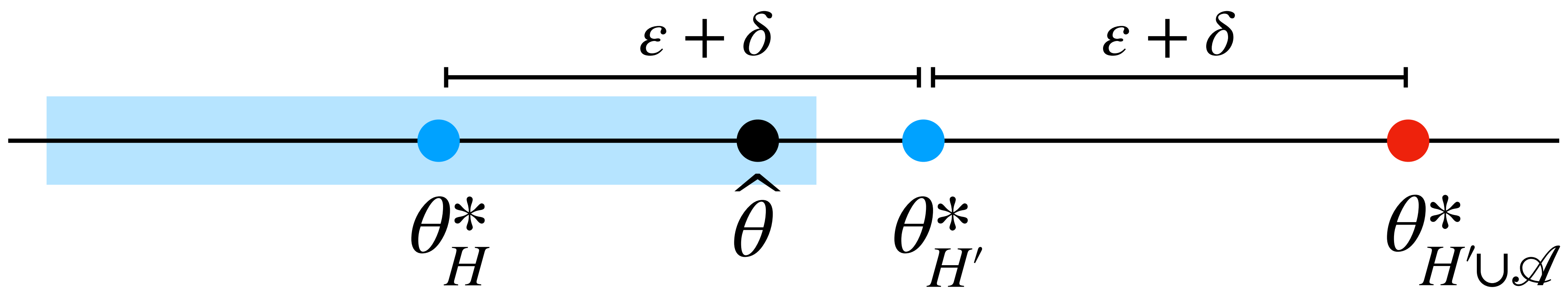
Set of honest nodes is $H' \cup \mathscr{A}$

$\widehat{\theta} \in [\theta^*_{H' \cup \mathscr{A}} - \varepsilon,\, \theta^*_{H' \cup \mathscr{A}} + \varepsilon]$

Which scenario is the correct one? We cannot know

Indistinguishable Executions

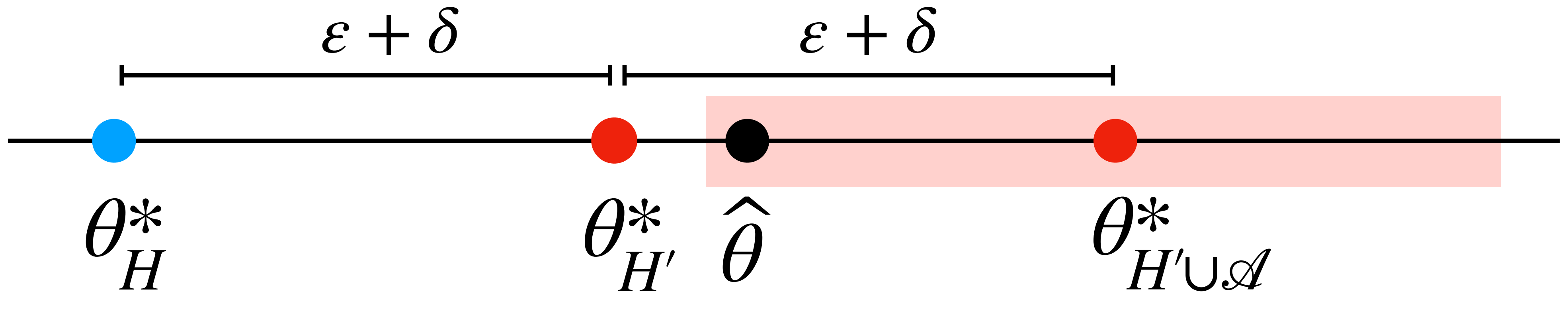
Scenario I



Set of honest nodes is H

$\widehat{\theta} \in [\theta^*_H - \varepsilon,\ \theta^*_H + \varepsilon]$

Scenario II



Set of honest nodes is $H' \cup \mathscr{A}$

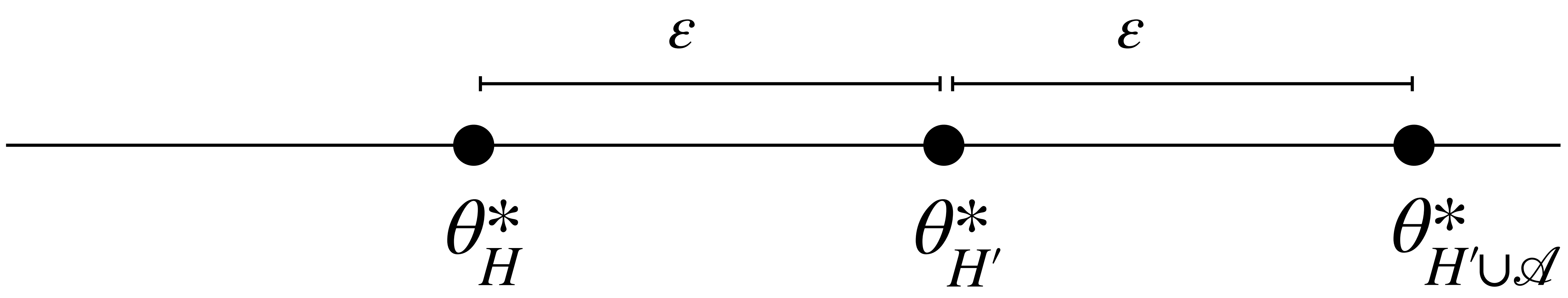
$\widehat{\theta} \in [\theta^*_{H' \cup \mathscr{A}} - \varepsilon,\ \theta^*_{H' \cup \mathscr{A}} + \varepsilon]$

Which scenario is the correct one? We cannot know

Satisfying (f, ε)-Resilience in Scenario I (or II) violates the condition in Scenario II (or I)

Bound the Differences in Segmented Solutions

Bound the Differences in Segmented Solutions

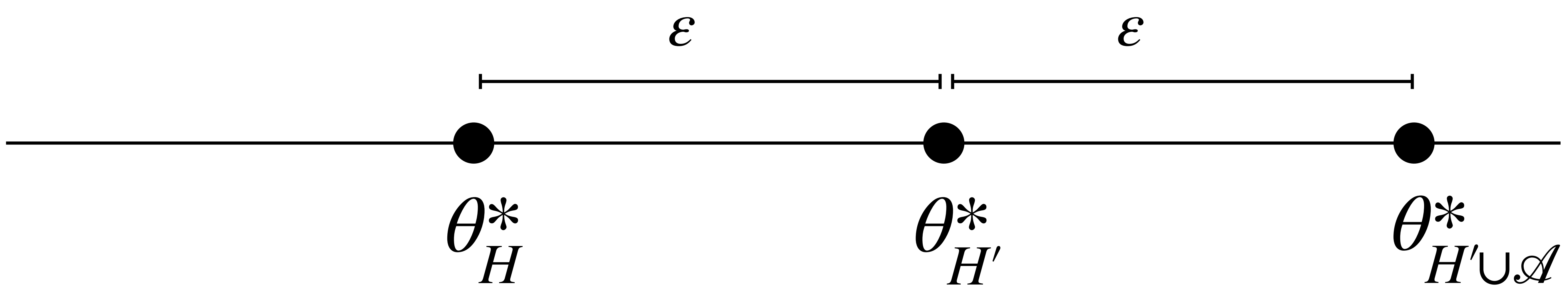


$H \rightarrow (n-f)$ nodes

$H' \rightarrow (n-2f)$ nodes

$\mathscr{A} \rightarrow f$ nodes

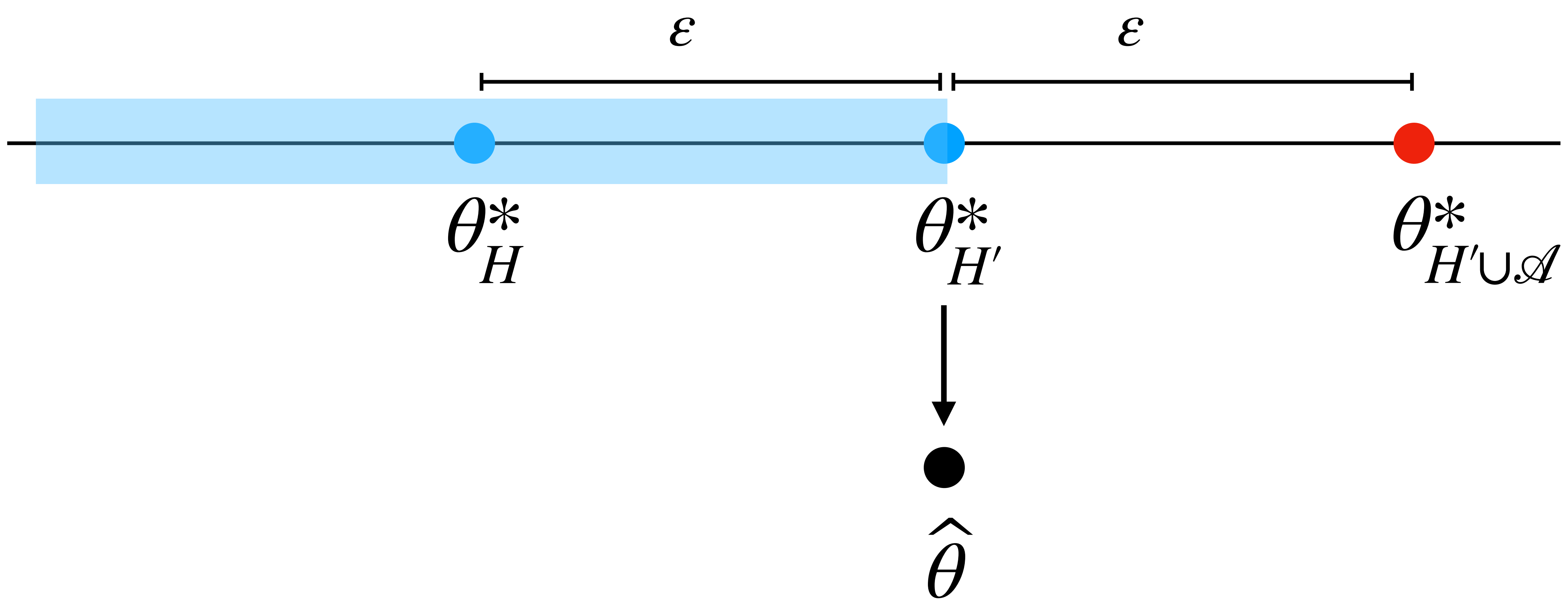
Bound the Differences in Segmented Solutions



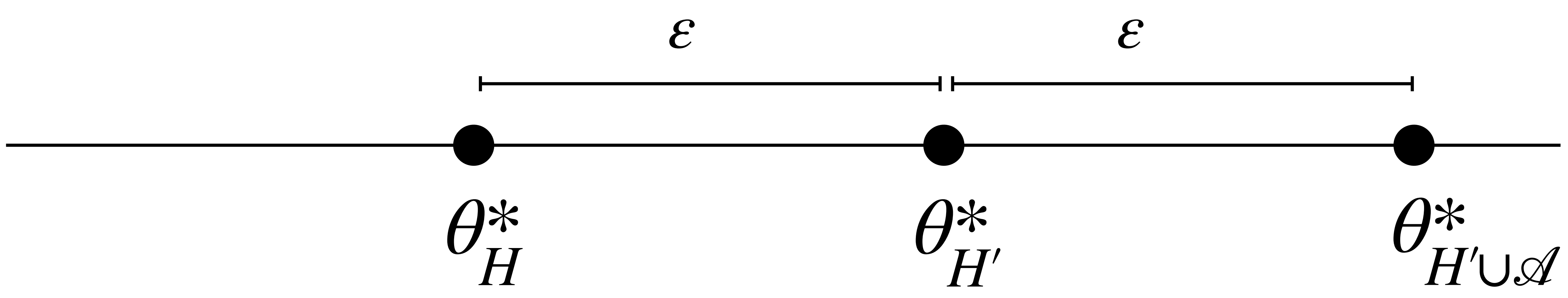
$H \rightarrow (n - f)$ nodes

$H' \rightarrow (n - 2f)$ nodes

$\mathscr{A} \rightarrow f$ nodes



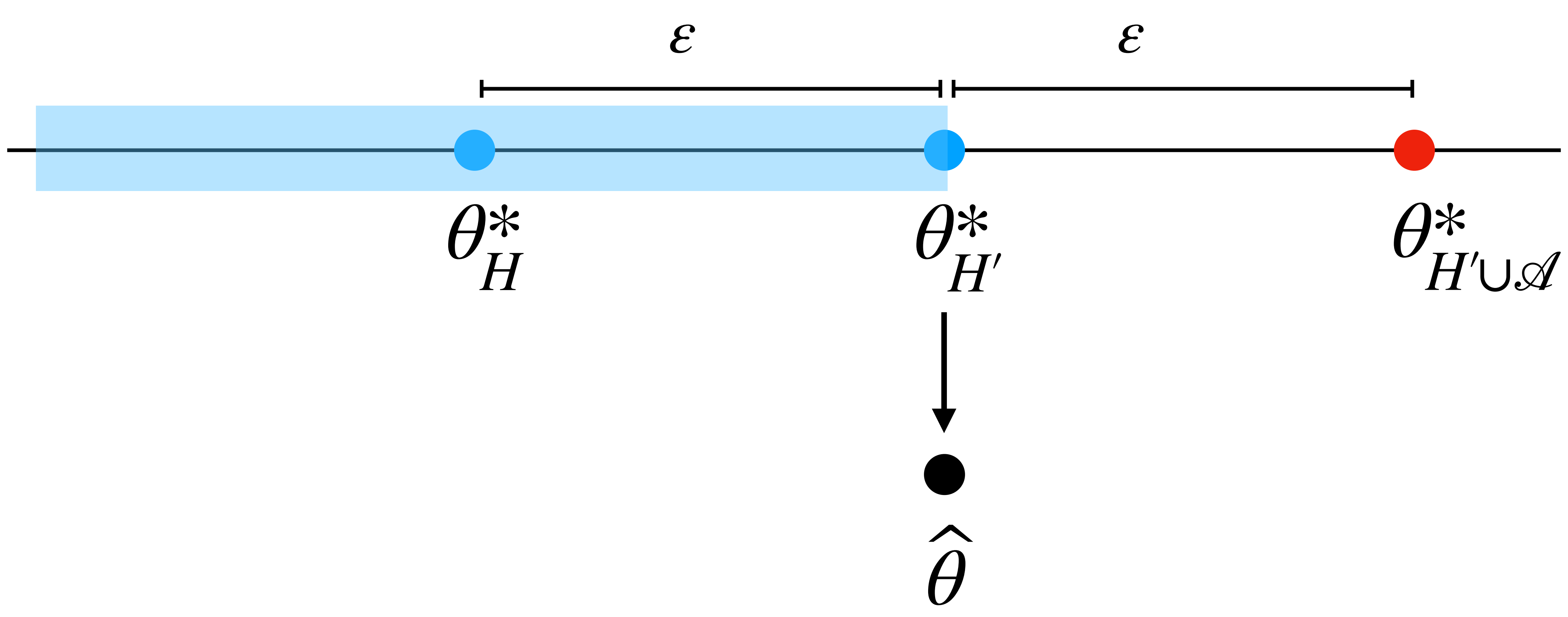
Bound the Differences in Segmented Solutions

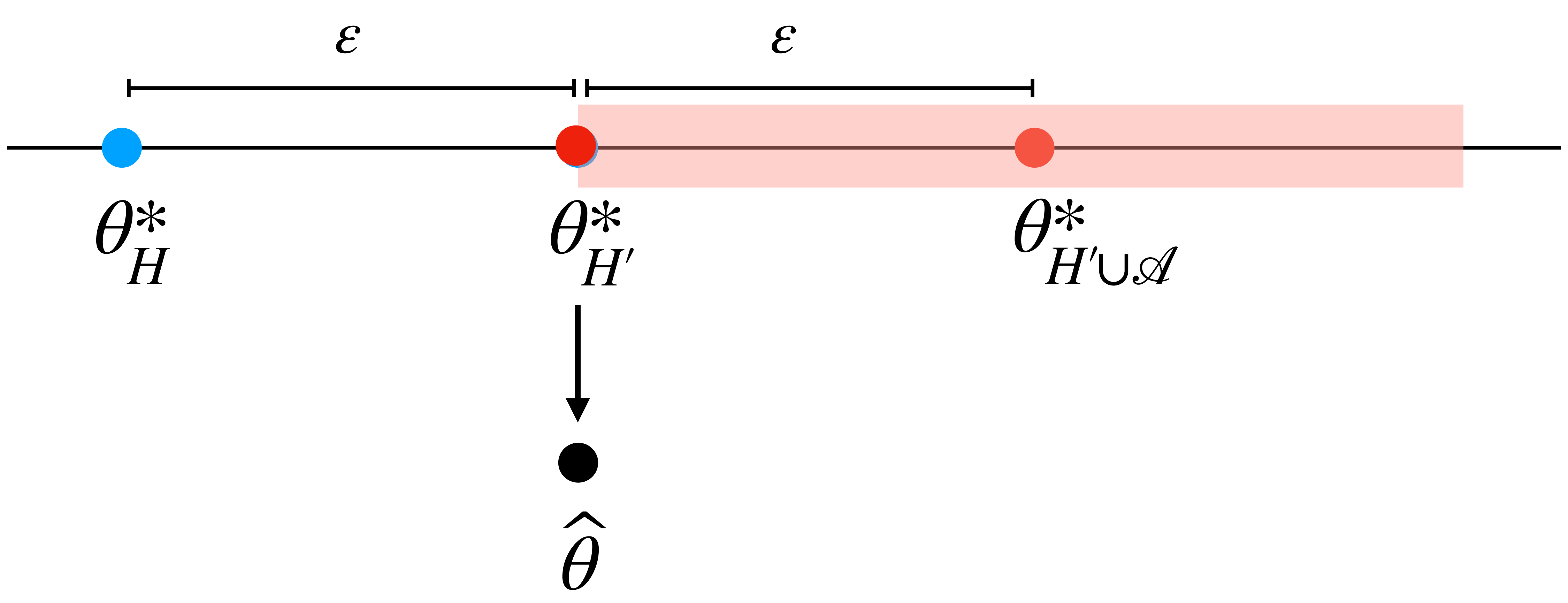


$H \to (n-f)$ nodes

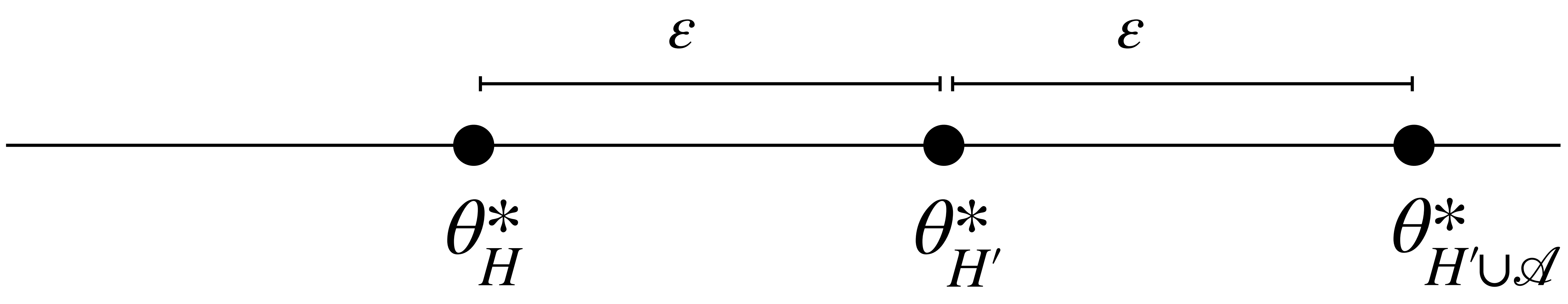
$H' \to (n-2f)$ nodes

$\mathscr{A} \to f$ nodes





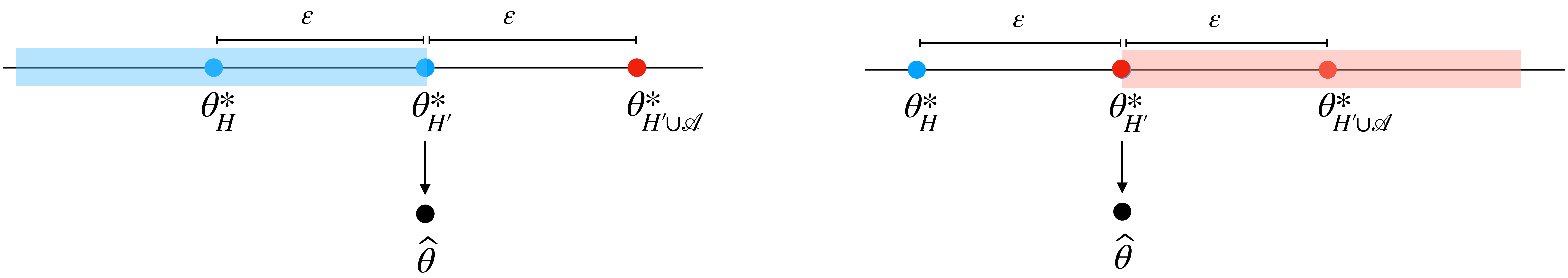
Bound the Differences in Segmented Solutions



$H \rightarrow (n-f)$ nodes

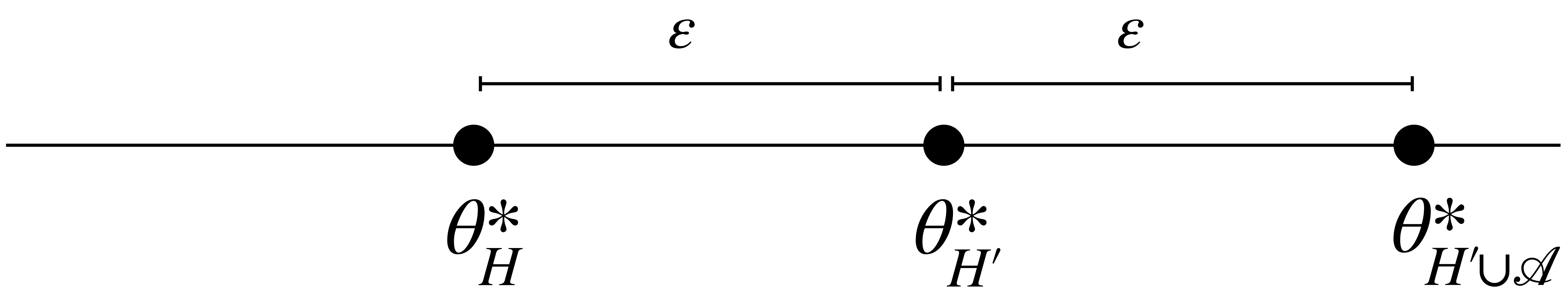
$H' \rightarrow (n-2f)$ nodes

$\mathscr{A} \rightarrow f$ nodes



satisfies (f, ε)-Resilience in both scenarios

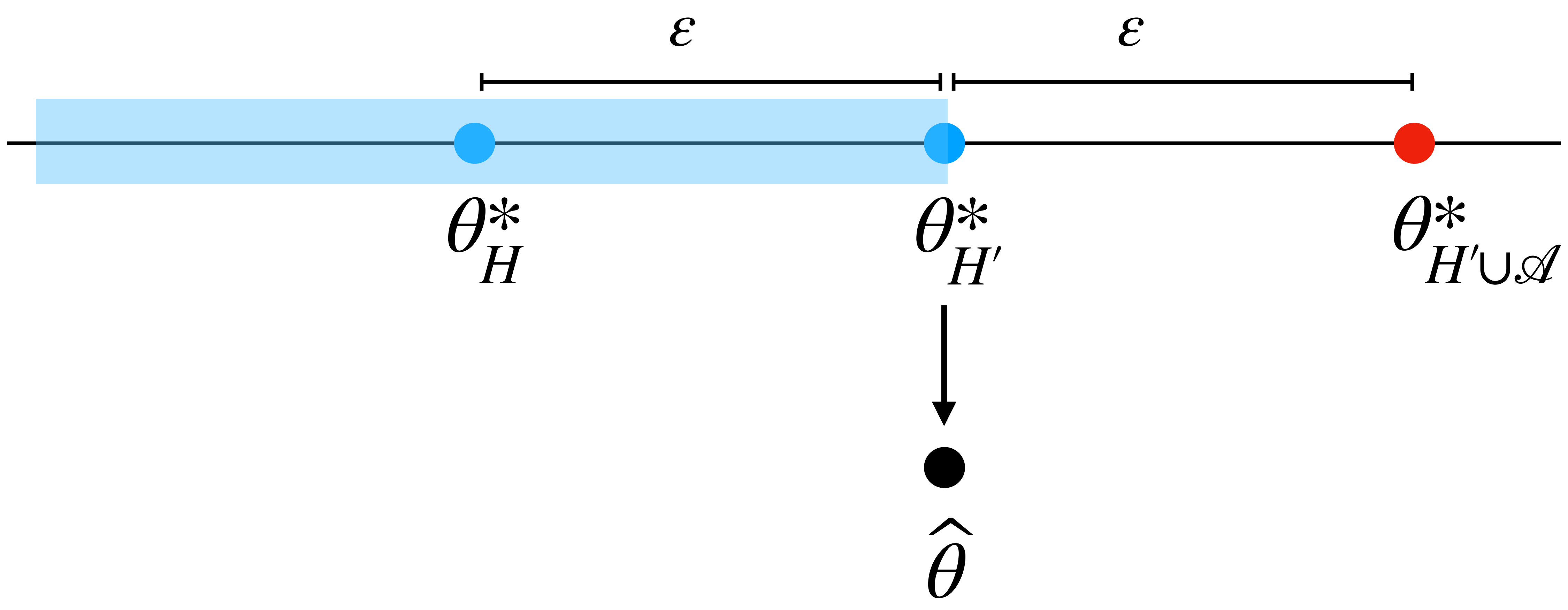
Bound the Differences in Segmented Solutions

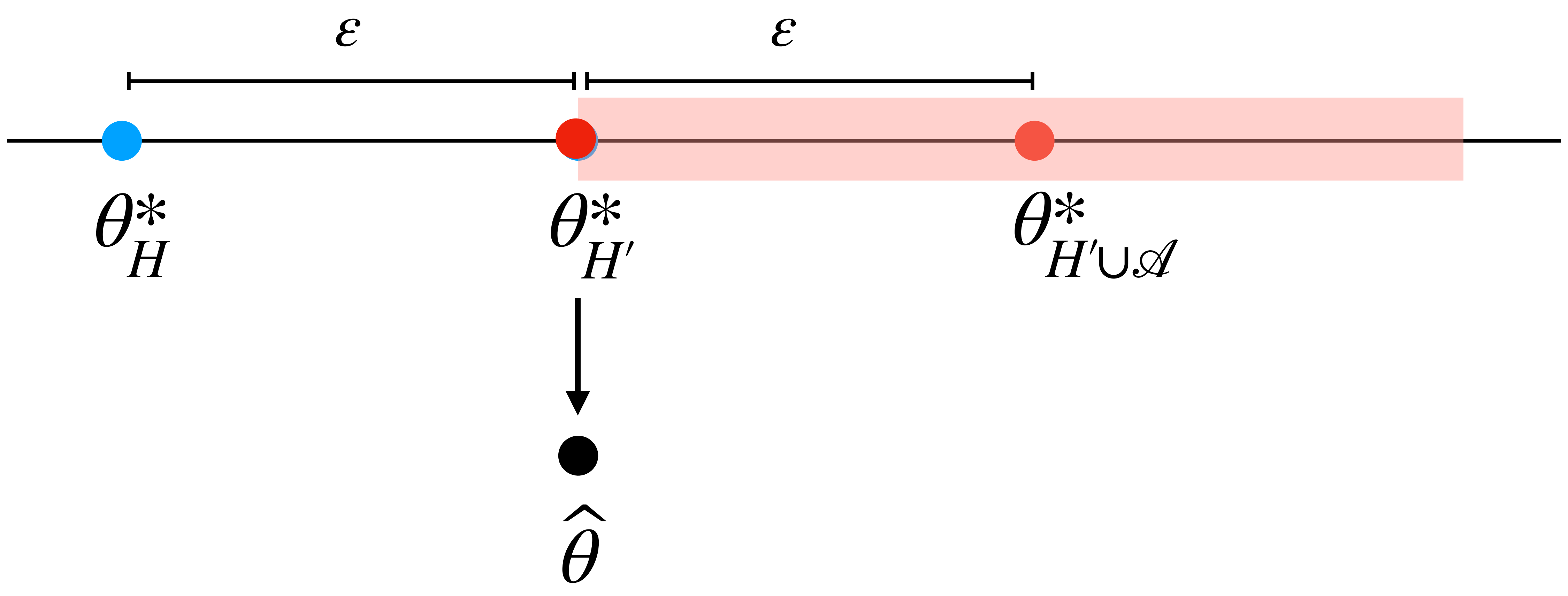


$H \rightarrow (n - f)$ nodes

$H' \rightarrow (n - 2f)$ nodes

$\mathcal{A} \rightarrow f$ nodes

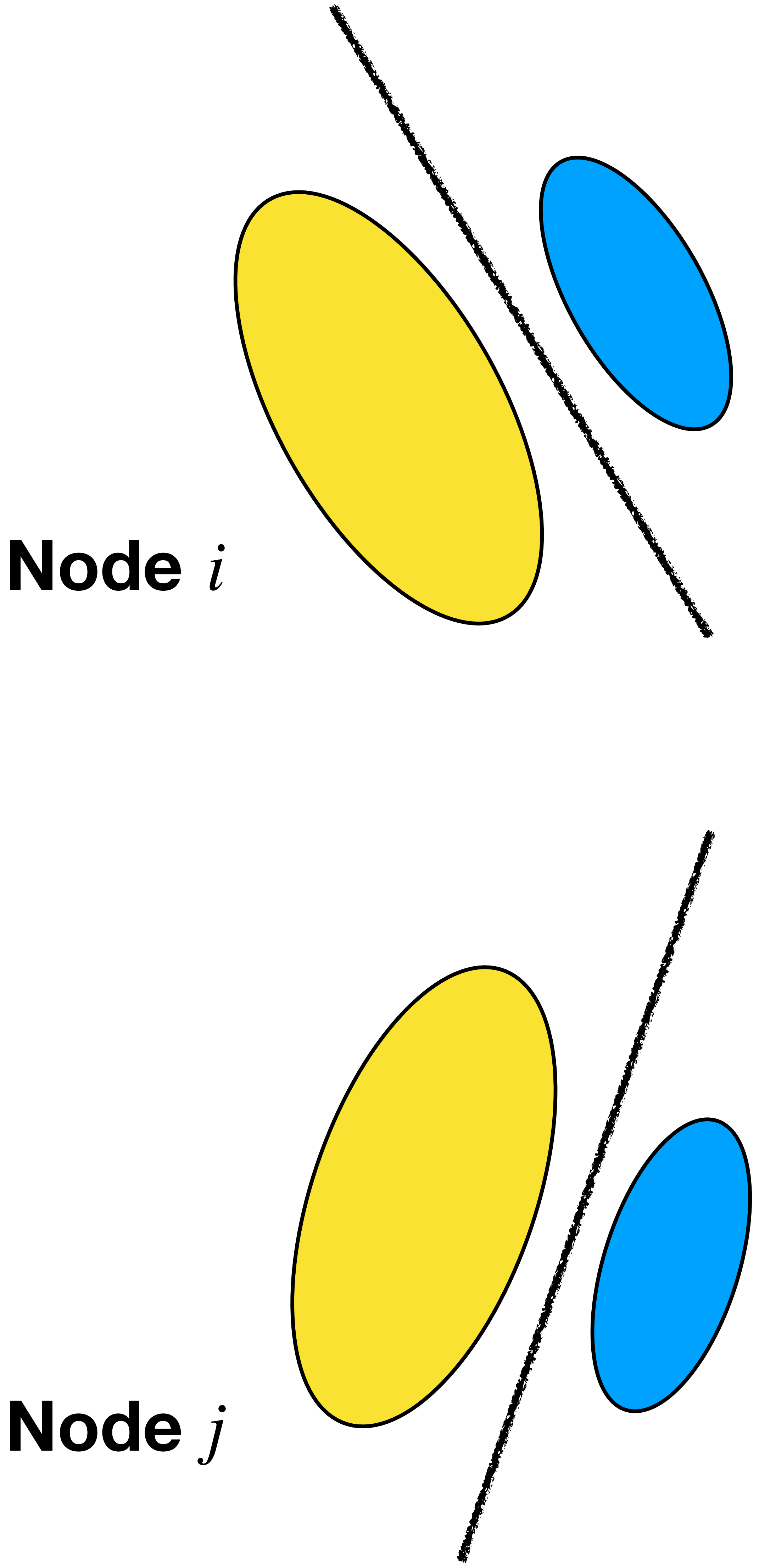




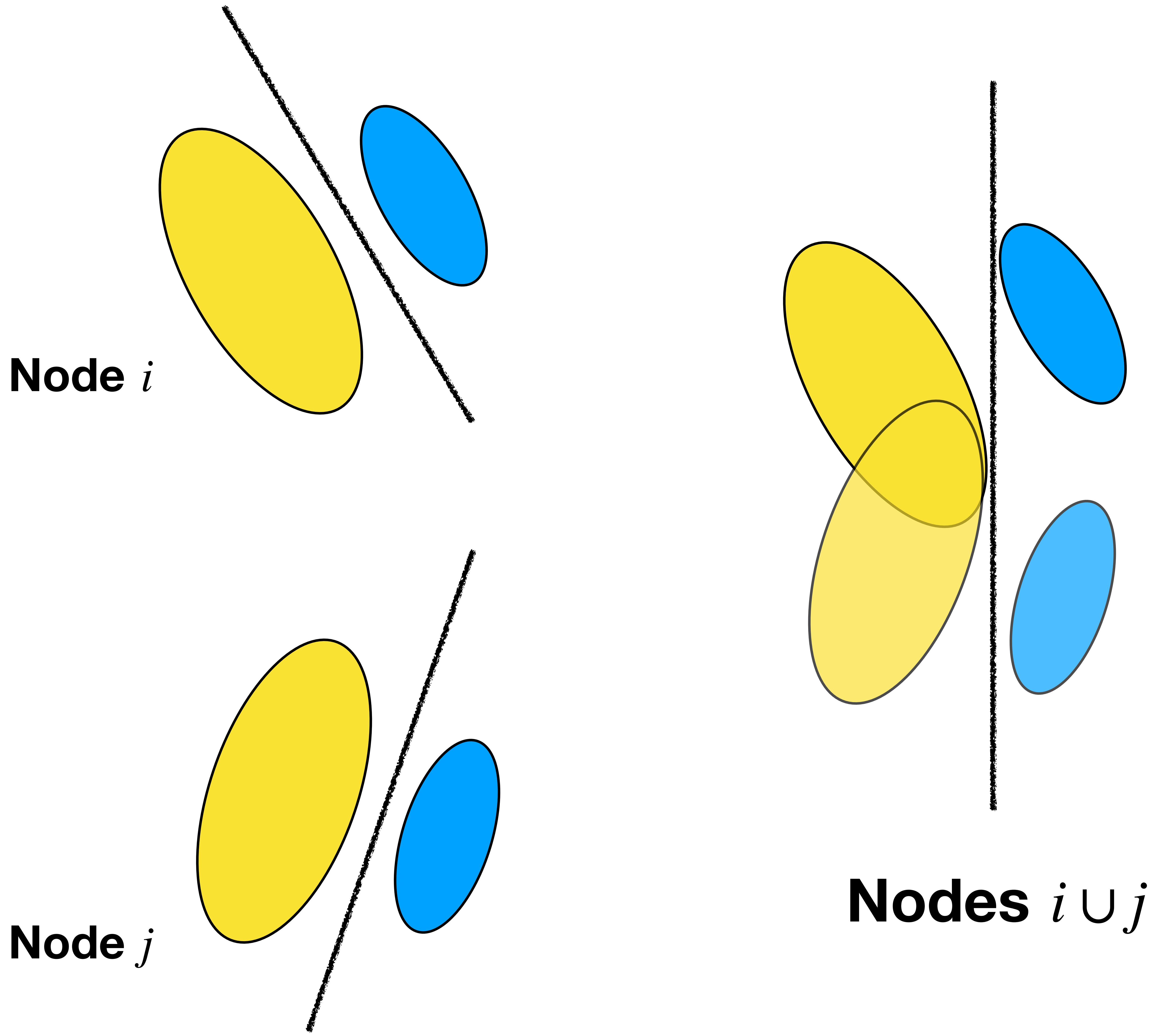
satisfies (f, ε)-Resilience in both scenarios

Bounded "Heterogeneity" is Critical to Robustness

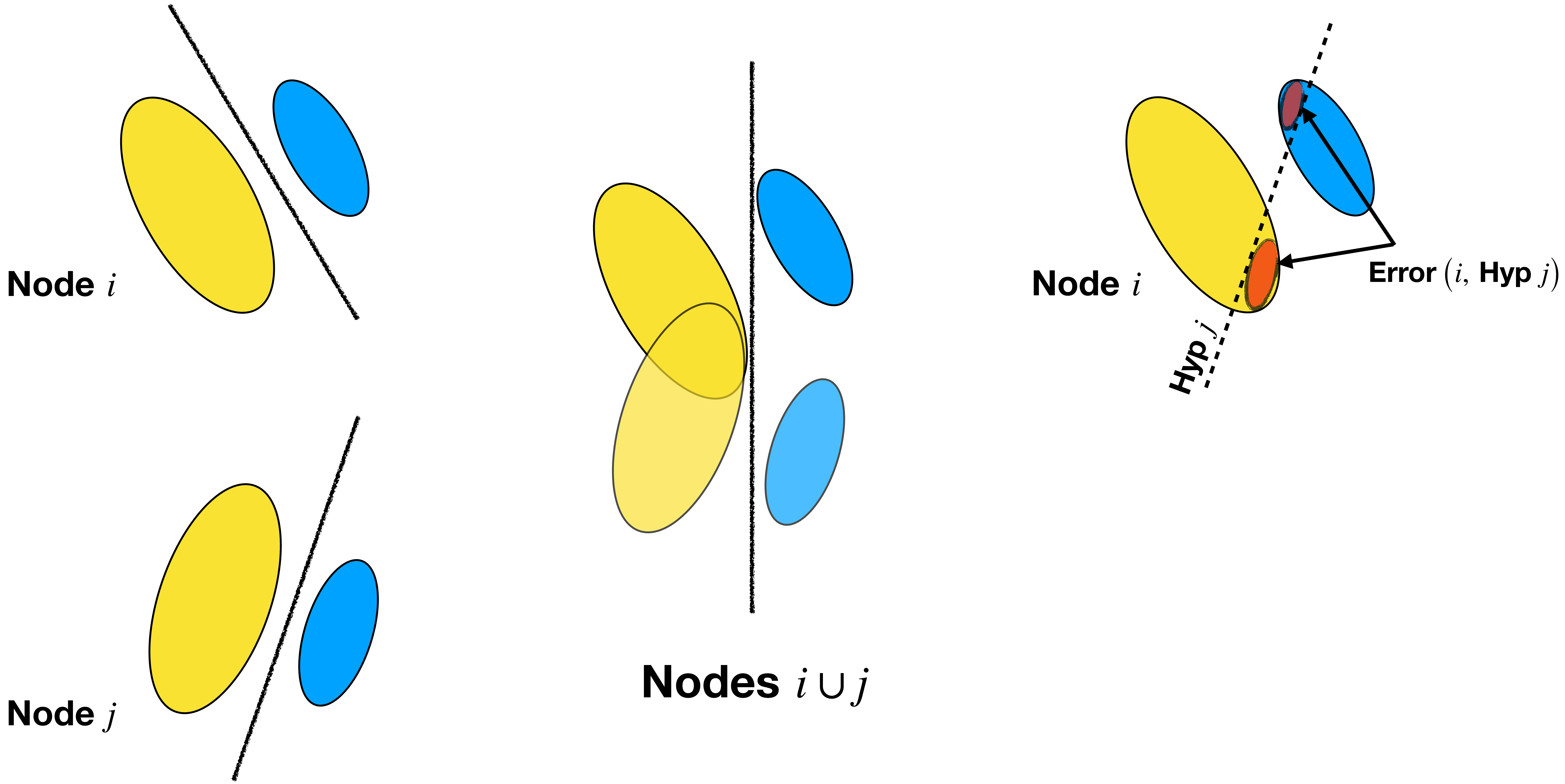
How to Characterize Heterogeneity?



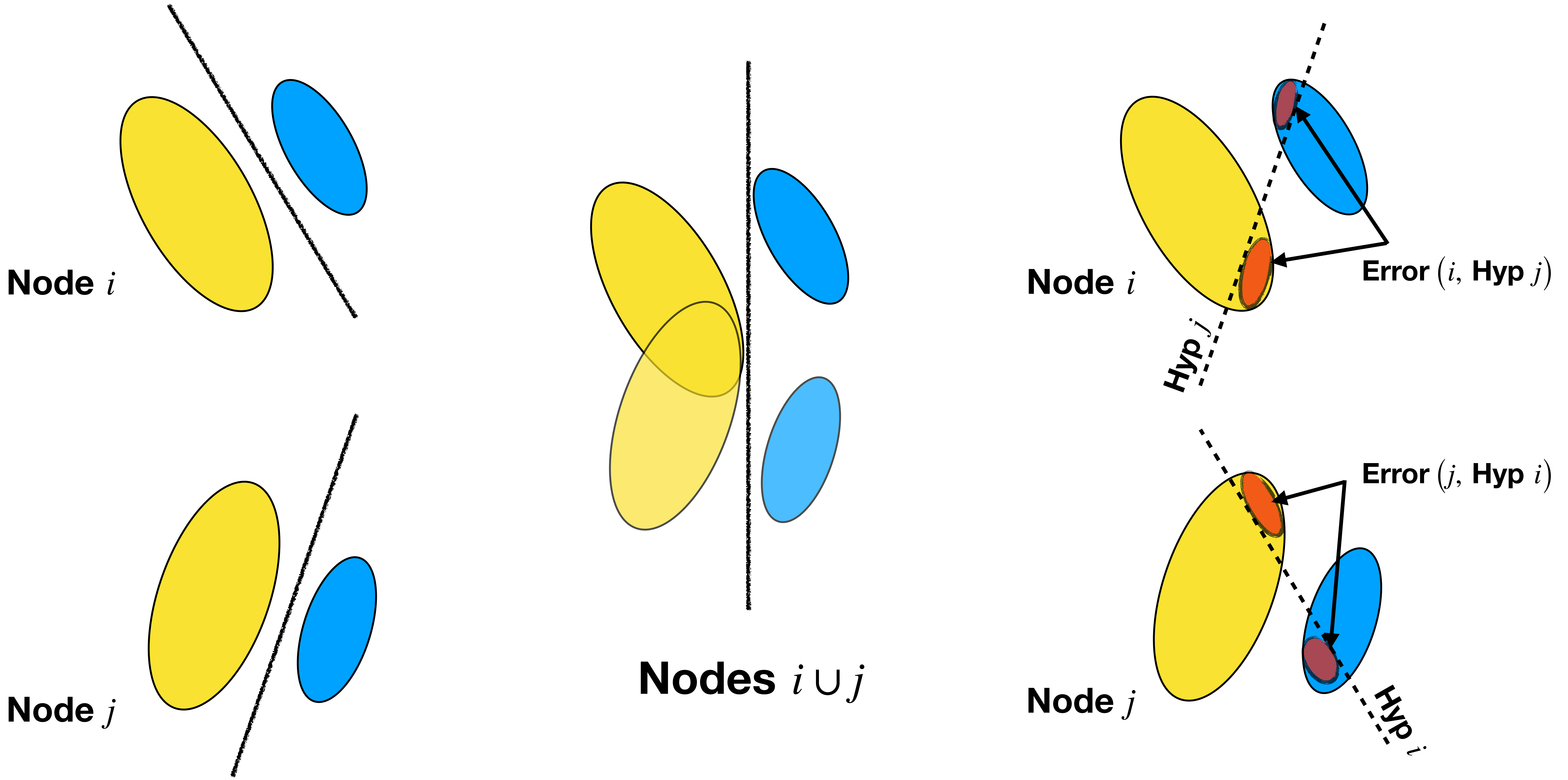
How to Characterize Heterogeneity?



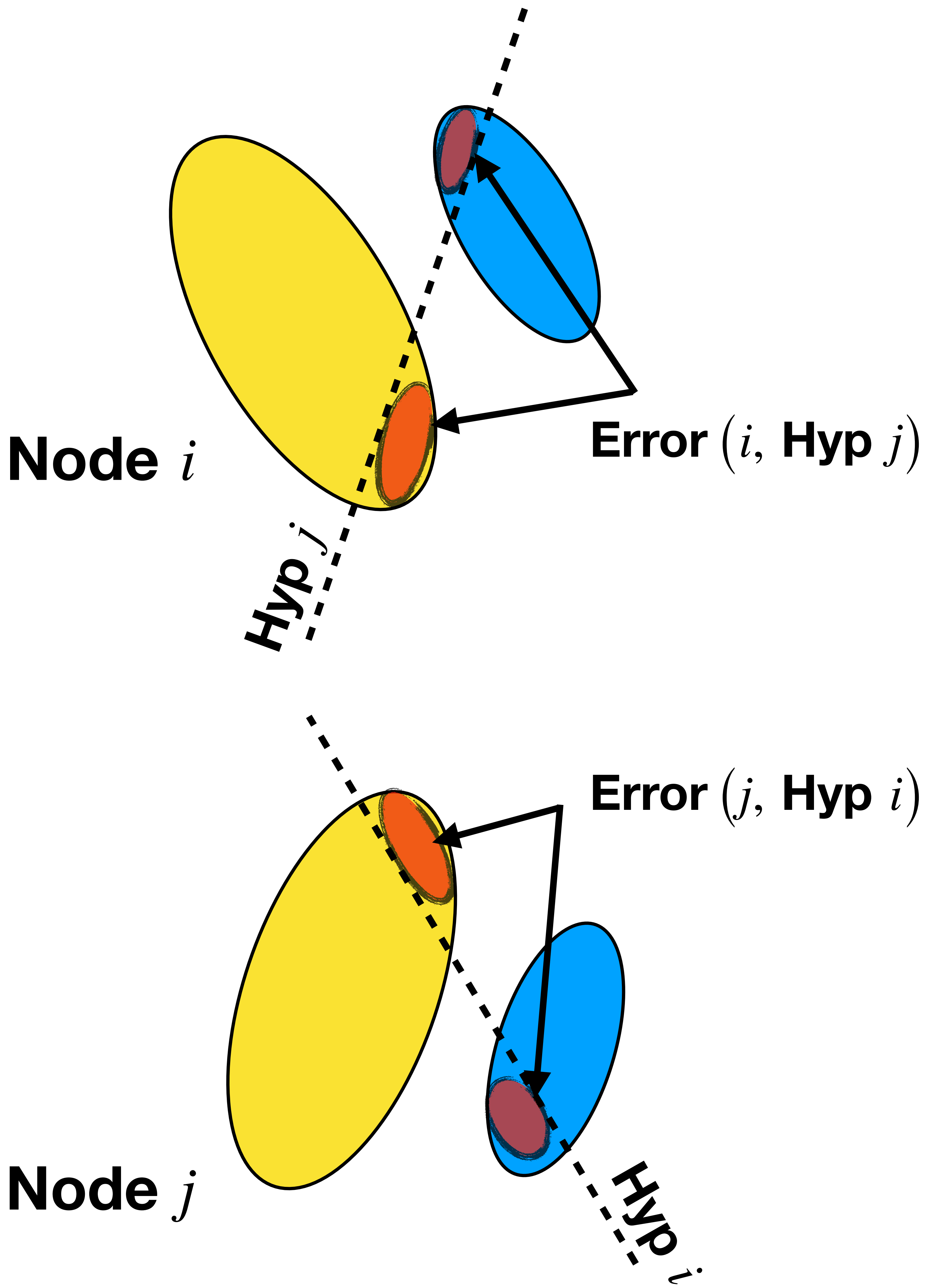
How to Characterize Heterogeneity?



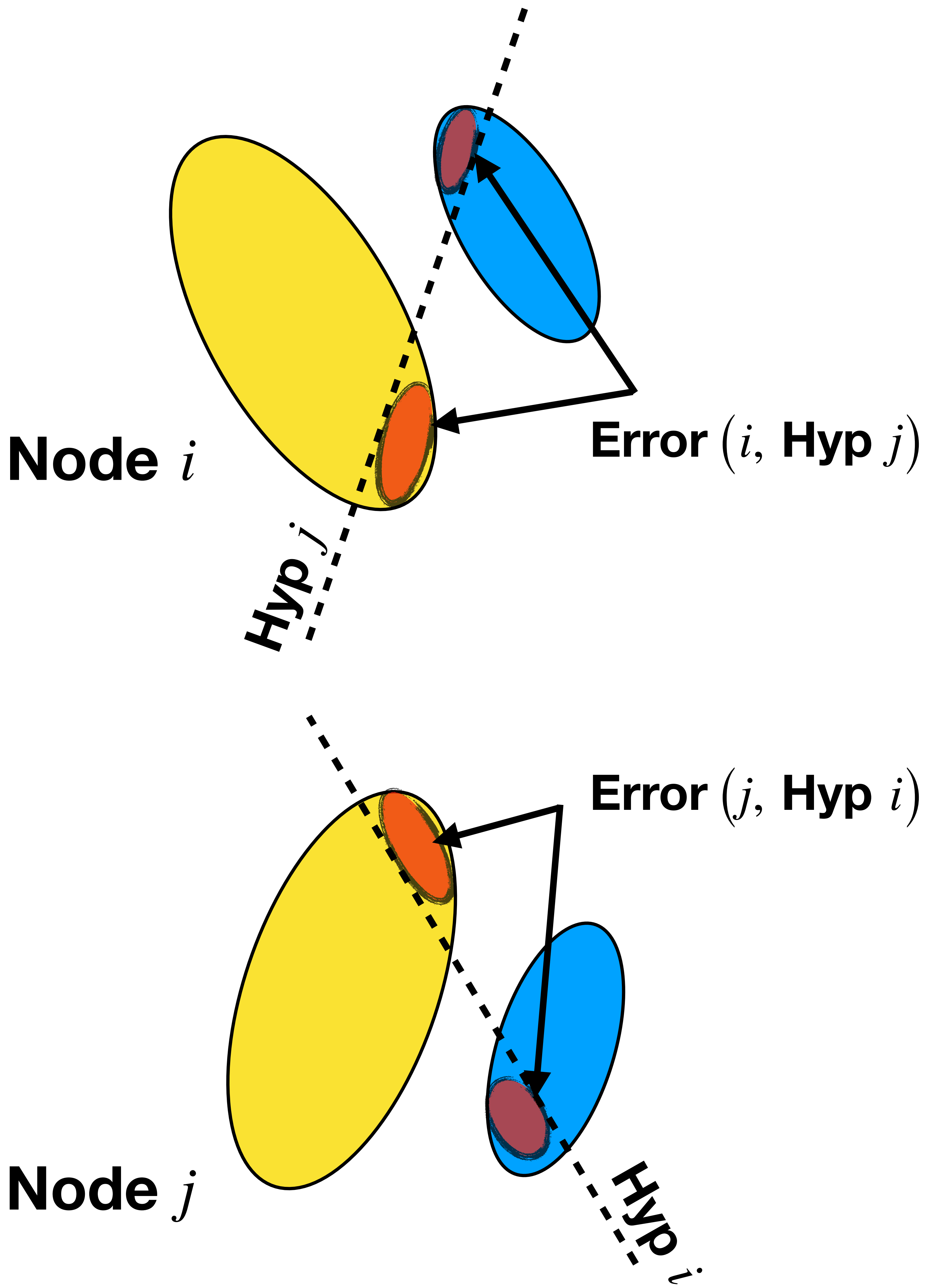
How to Characterize Heterogeneity?



Heterogeneity $\triangleq$ Variation Between Solutions

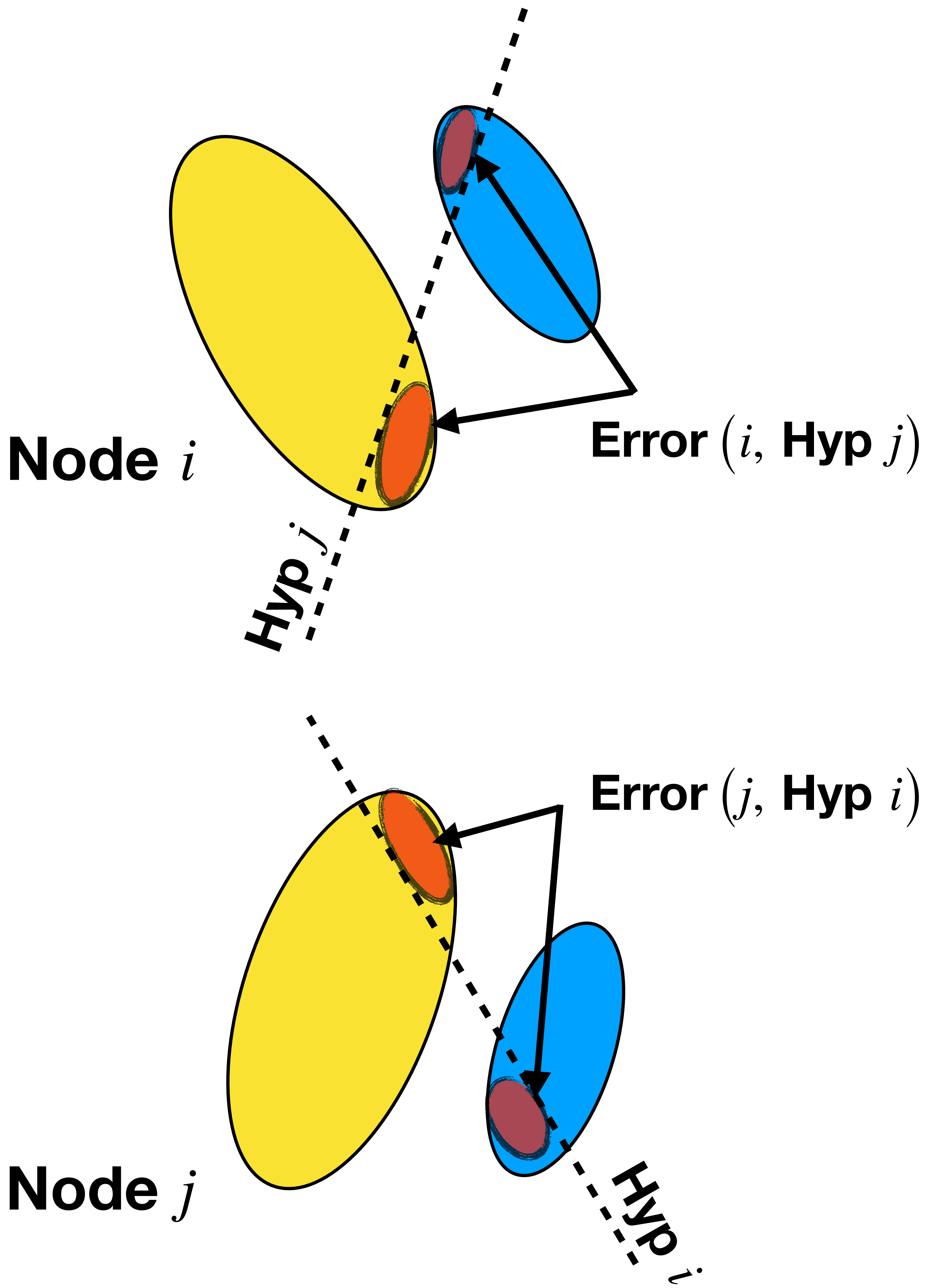


Heterogeneity $\triangleq$ Variation Between Solutions



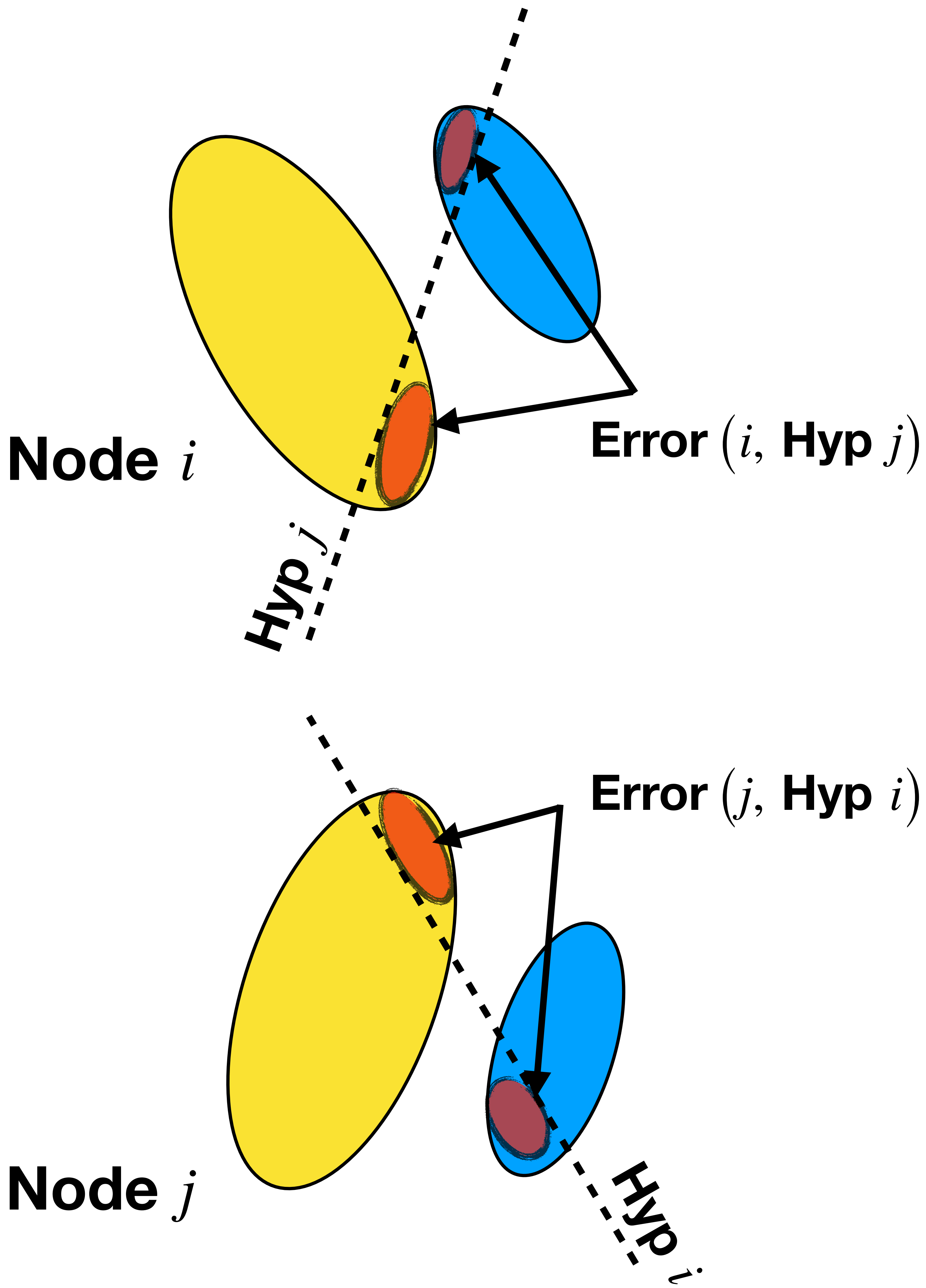
(f, ε)-Redundancy

Heterogeneity $\triangleq$ Variation Between Solutions



(f, ε)-Redundancy

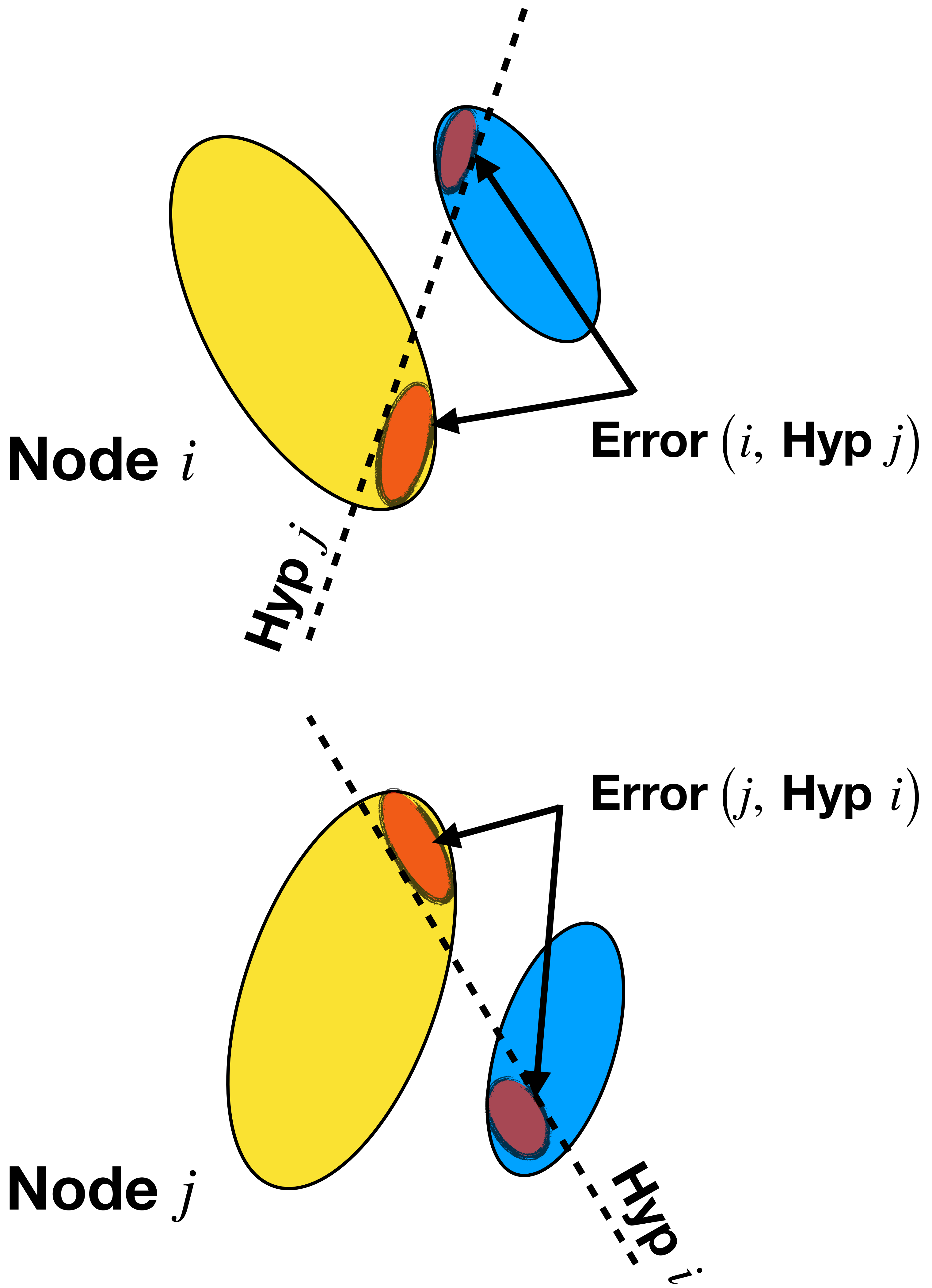
Heterogeneity $\triangleq$ Variation Between Solutions



(f, ε)-Redundancy

Ignoring f nodes leads to sub-optimality of value less than ε :

Heterogeneity $\triangleq$ Variation Between Solutions

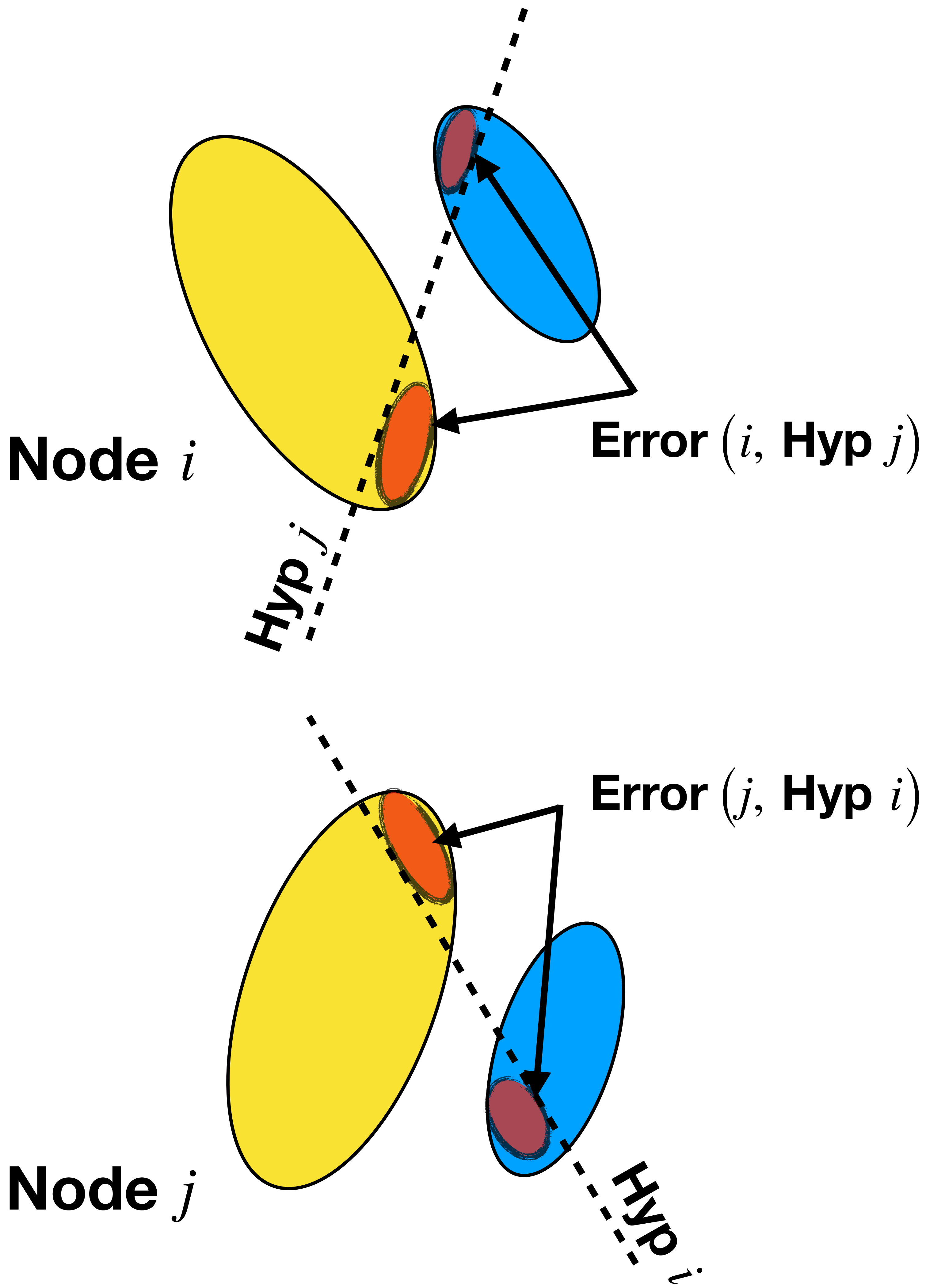


(f, ε)-Redundancy

Ignoring f nodes leads to sub-optimality of value less than ε :

$S \rightarrow n - f$ nodes

Heterogeneity $\triangleq$ Variation Between Solutions

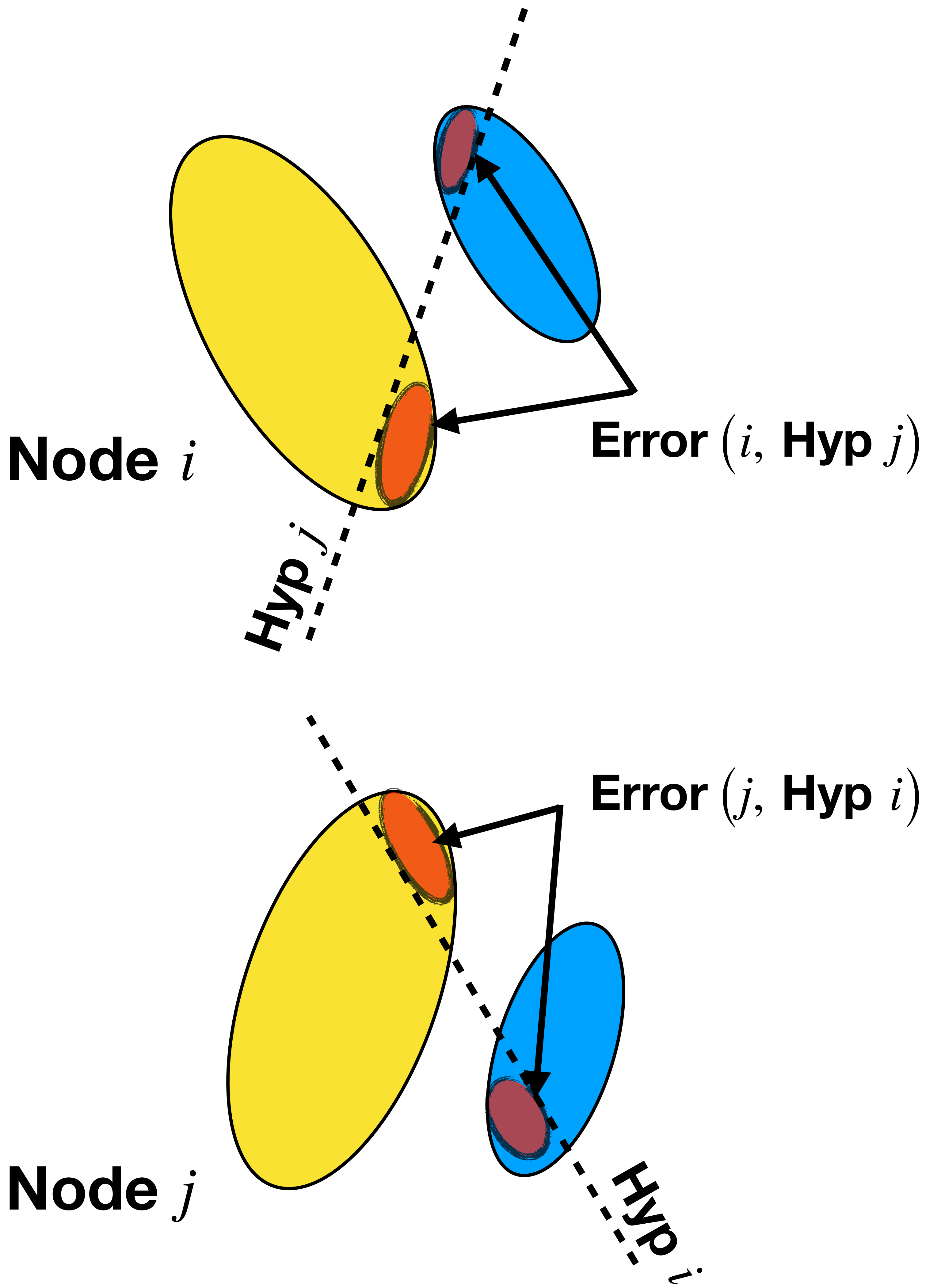


(f, ε)-Redundancy

Ignoring f nodes leads to sub-optimality of value less than ε :

$S \rightarrow n - f$ nodes $\quad$ $S' \subseteq S$ of $n - 2f$ nodes

Heterogeneity $\triangleq$ Variation Between Solutions



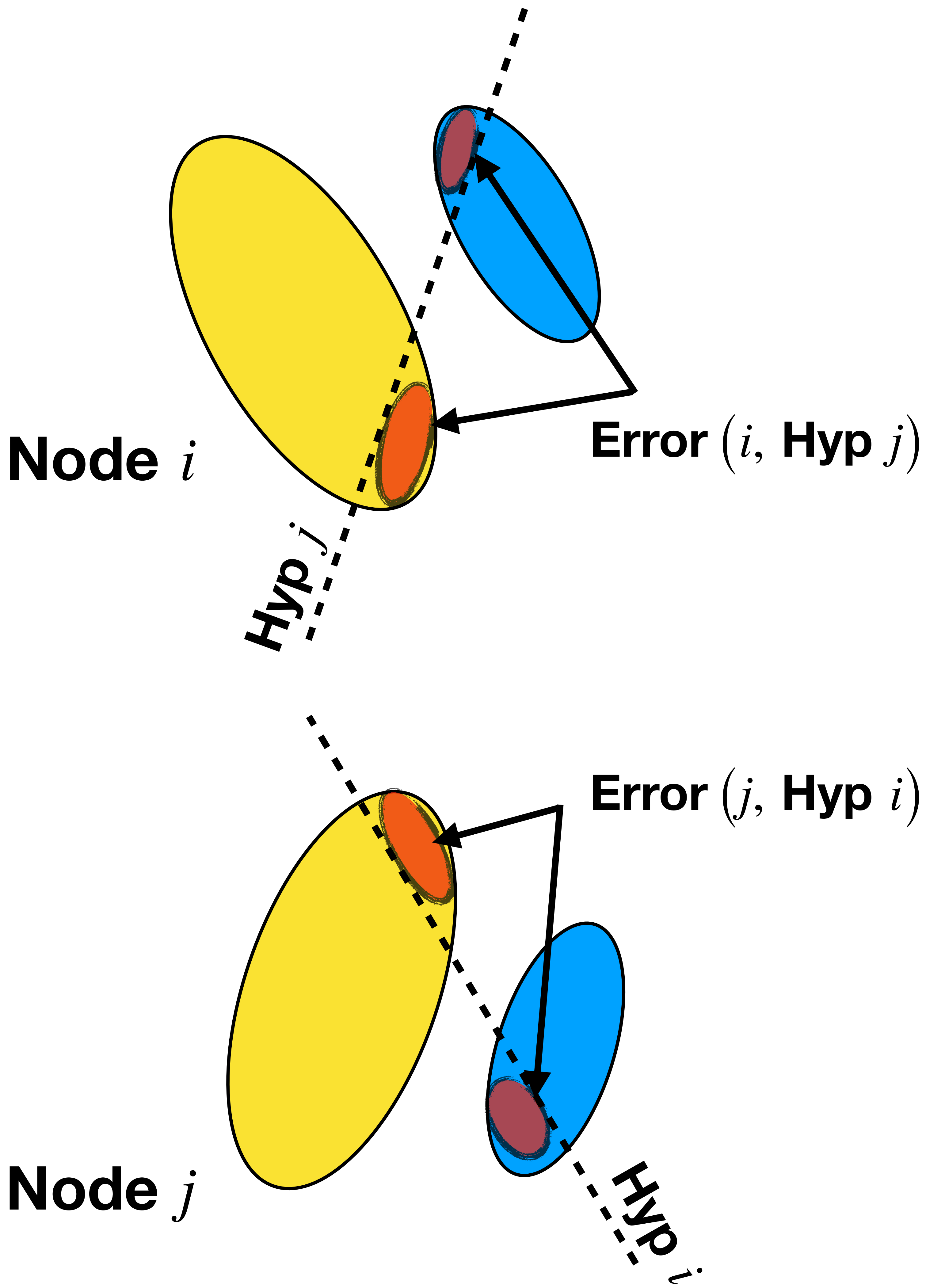
(f, ε)-Redundancy

Ignoring f nodes leads to sub-optimality of value less than ε :

$S \rightarrow n - f$ nodes $\qquad$ $S' \subseteq S$ of $n - 2f$ nodes

$$\mathcal{L}_S\left(\theta^*_{S'}\right) - \min \mathcal{L}_S \leq \varepsilon$$

Heterogeneity $\triangleq$ Variation Between Solutions



(f, ε)-Redundancy

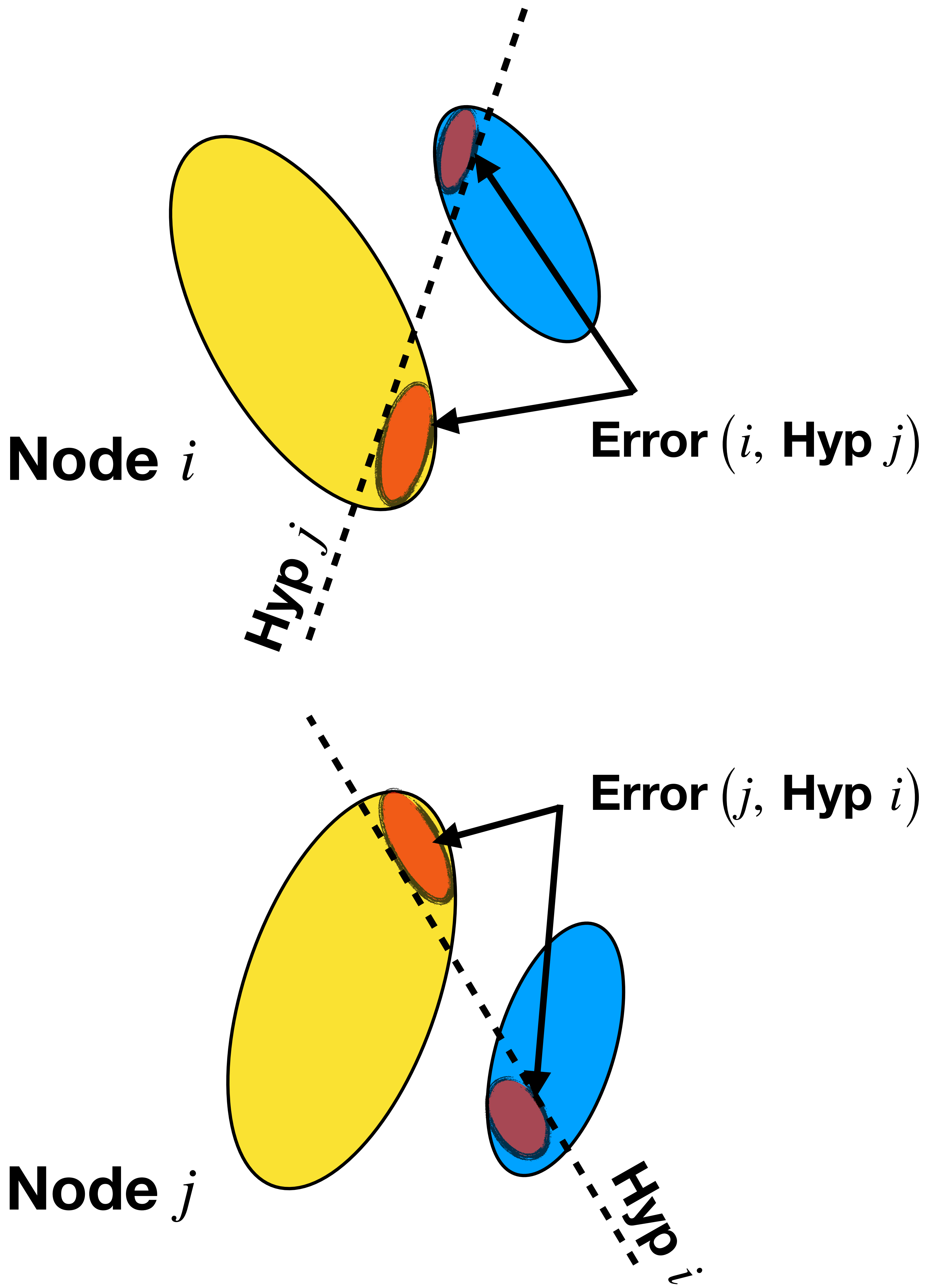
Ignoring f nodes leads to sub-optimality of value less than ε :

$S \rightarrow n - f$ nodes $\quad S' \subseteq S$ of $n - 2f$ nodes

$$\mathscr{L}_S\left(\theta^*_{S'}\right) - \min \mathscr{L}_S \leq \varepsilon$$

$\theta^*_{S'} := \arg\min \mathscr{L}_{S'}(\theta)$

Heterogeneity $\triangleq$ Variation Between Solutions



(f, ε)-Redundancy

Ignoring f nodes leads to sub-optimality of value less than ε :

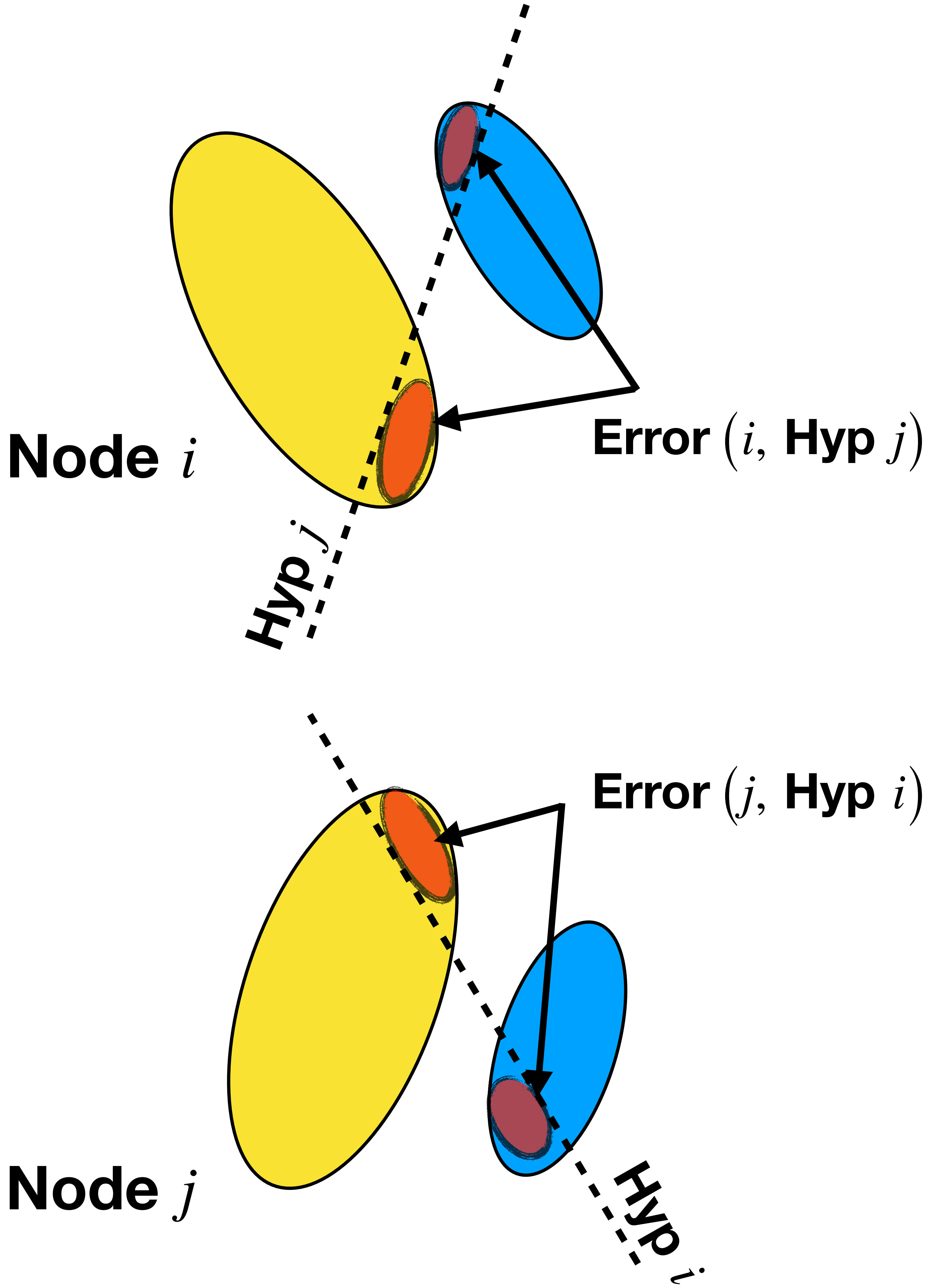
$S \to n - f$ nodes $\quad S' \subseteq S$ of $n - 2f$ nodes

$$\mathscr{L}_S\left(\theta^*_{S'}\right) - \min \mathscr{L}_S \leq \varepsilon$$

$\theta^*_{S'} := \arg\min \mathscr{L}_{S'}(\theta)$

(f, ε)-resilience $\iff (f, \varepsilon)$-redundancy

Heterogeneity $\triangleq$ Variation Between Solutions



(f, ε)-Redundancy

Ignoring f nodes leads to sub-optimality of value less than ε :

$S \to n - f$ nodes $\qquad S' \subseteq S$ of $n - 2f$ nodes

$$\mathscr{L}_S\left(\theta^*_{S'}\right) - \min \mathscr{L}_S \leq \varepsilon$$

$\theta^*_{S'} := \arg\min \mathscr{L}_{S'}(\theta)$

(f, ε)-resilience $\Longleftrightarrow$ (f, ε)-redundancy

"Approximate Fault-Tolerance in Distributed Optimization." S. Liu et al., PODC'21

Optimal Robust Distributed Learning

Optimal Robust Distributed Learning

Choose a set S such that $|S| = n - f$

Optimal Robust Distributed Learning

Choose a set S such that $|S| = n - f$

For all $S' \subseteq S$ such that $|S'| = n - 2f$

Optimal Robust Distributed Learning

Choose a set S such that $|S| = n - f$

For all $S' \subseteq S$ such that $|S'| = n - 2f$

Compute error$(S, S') \triangleq \mathscr{L}_S\left(\theta^*_{S'}\right) - \min \mathscr{L}_S$

Optimal Robust Distributed Learning

Choose a set S such that $|S| = n - f$

For all $S' \subseteq S$ such that $|S'| = n - 2f$

Compute $\text{error}(S, S') \triangleq \mathscr{L}_S\left(\theta^*_{S'}\right) - \min \mathscr{L}_S$

Output $\arg\min \mathscr{L}_{S*}(\theta)$ such that

$$S* \in \arg\min_{S} \left\{ \max_{S' \subseteq S} \ \text{error}\left(S, S'\right) \right\}$$

Optimal Robust Distributed Learning

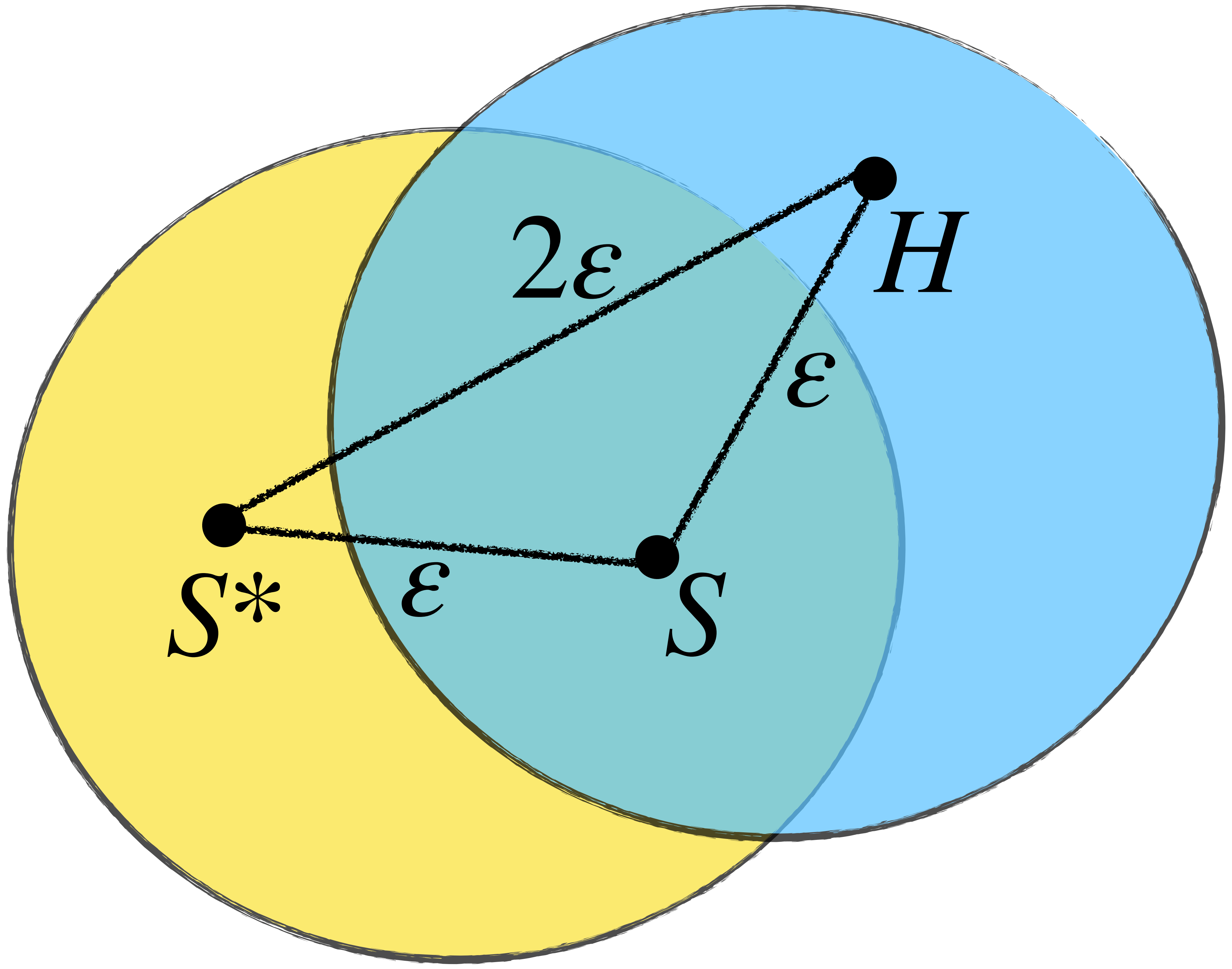
Choose a set S such that $|S| = n - f$

For all $S' \subseteq S$ such that $|S'| = n - 2f$

Compute $\text{error}(S, S') \triangleq \mathscr{L}_S\left(\theta^*_{S'}\right) - \min \mathscr{L}_S$

Output $\arg\min \mathscr{L}_{S*}(\theta)$ such that

$$S^* \in \arg\min_{S} \left\{ \max_{S' \subseteq S} \text{error}\,(S, S') \right\}$$



Optimal Robust Distributed Learning

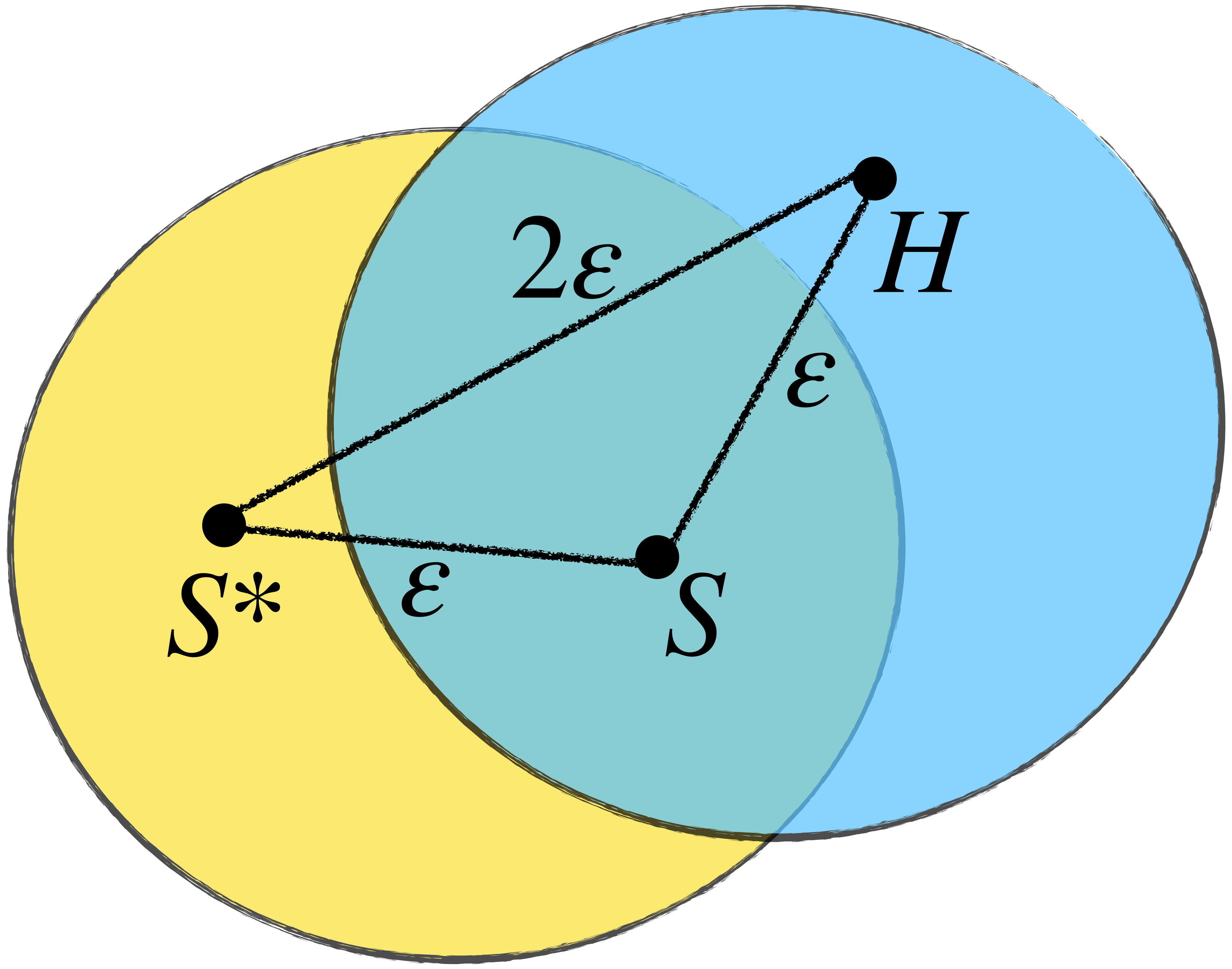
Choose a set S such that $|S| = n - f$

For all $S' \subseteq S$ such that $|S'| = n - 2f$

Compute error$(S, S') \triangleq \mathscr{L}_S\left(\theta^*_{S'}\right) - \min \mathscr{L}_S$

Output $\arg\min \mathscr{L}_{S^*}(\theta)$ such that

$$S^* \in \arg\min_S \left\{ \max_{S' \subseteq S} \text{ error}(S, S') \right\}$$



(f, ε)-redundancy $\Longrightarrow (f, 2\varepsilon)$-resilience

Robust Decoding or State Estimation

Robust Decoding or State Estimation

Consider $y = Ax$ where $A \in \mathbb{R}^{n \times m}$, and observation $y \in \mathbb{R}^n$

Robust Decoding or State Estimation

Consider $y = Ax$ where $A \in \mathbb{R}^{n \times m}$, and observation $y \in \mathbb{R}^n$

Suppose that up to f of the observations are arbitrarily corrupted

Robust Decoding or State Estimation

Consider $y = Ax$ where $A \in \mathbb{R}^{n \times m}$, and observation $y \in \mathbb{R}^n$

Suppose that up to f of the observations are arbitrarily corrupted

We can recover x from $\tilde{y} \triangleq y + \delta$ where $||\delta||_0 \leq f$

Robust Decoding or State Estimation

Consider $y = Ax$ where $A \in \mathbb{R}^{n \times m}$, and observation $y \in \mathbb{R}^n$

Suppose that up to f of the observations are arbitrarily corrupted

We can recover x from $\tilde{y} \triangleq y + \delta$ where $||\delta||_0 \leq f$

As long as $\text{rank}\left(A^S\right) = m$ where $S \subseteq [n]$ such that $|S| = n - 2f$

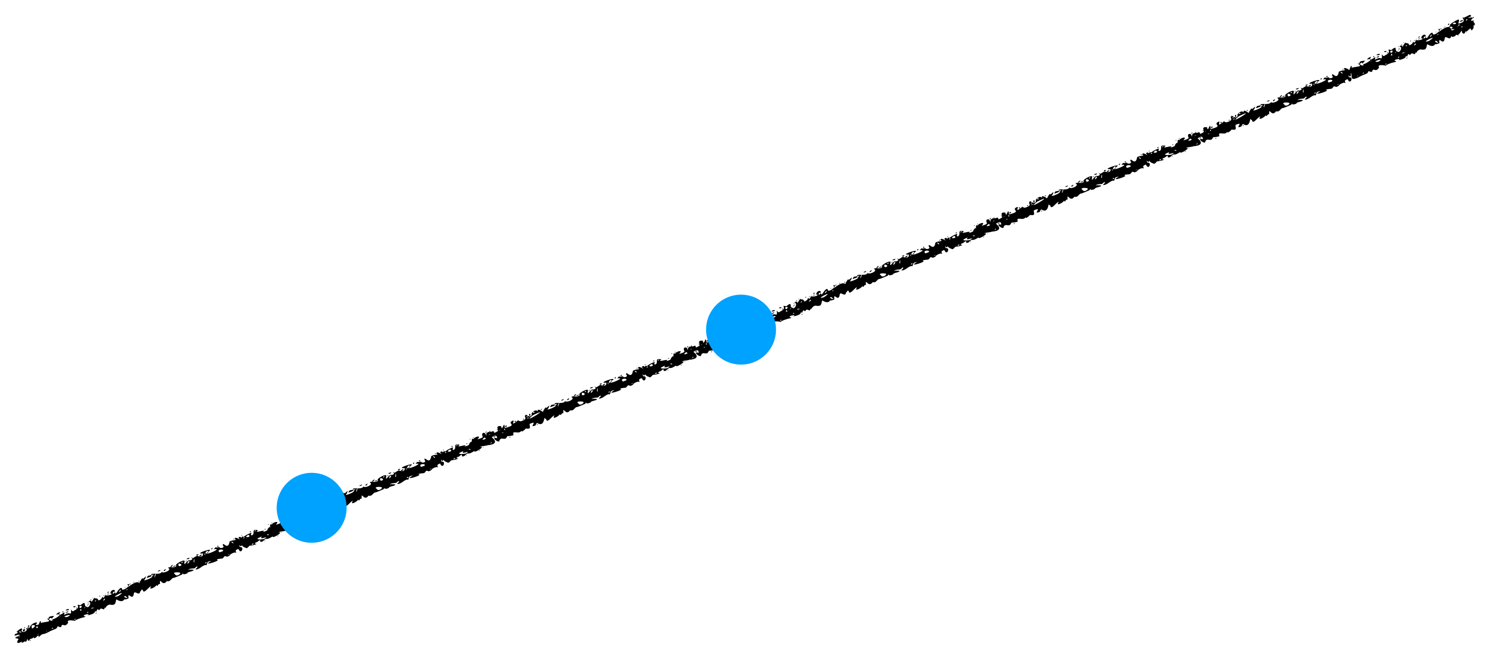
Robust Decoding or State Estimation

Consider $y = Ax$ where $A \in \mathbb{R}^{n \times m}$, and observation $y \in \mathbb{R}^n$

Suppose that up to f of the observations are arbitrarily corrupted

We can recover x from $\tilde{y} \triangleq y + \delta$ where $||\delta||_0 \leq f$

As long as $\operatorname{rank}\left(A^S\right) = m$ where $S \subseteq [n]$ such that $|S| = n - 2f$



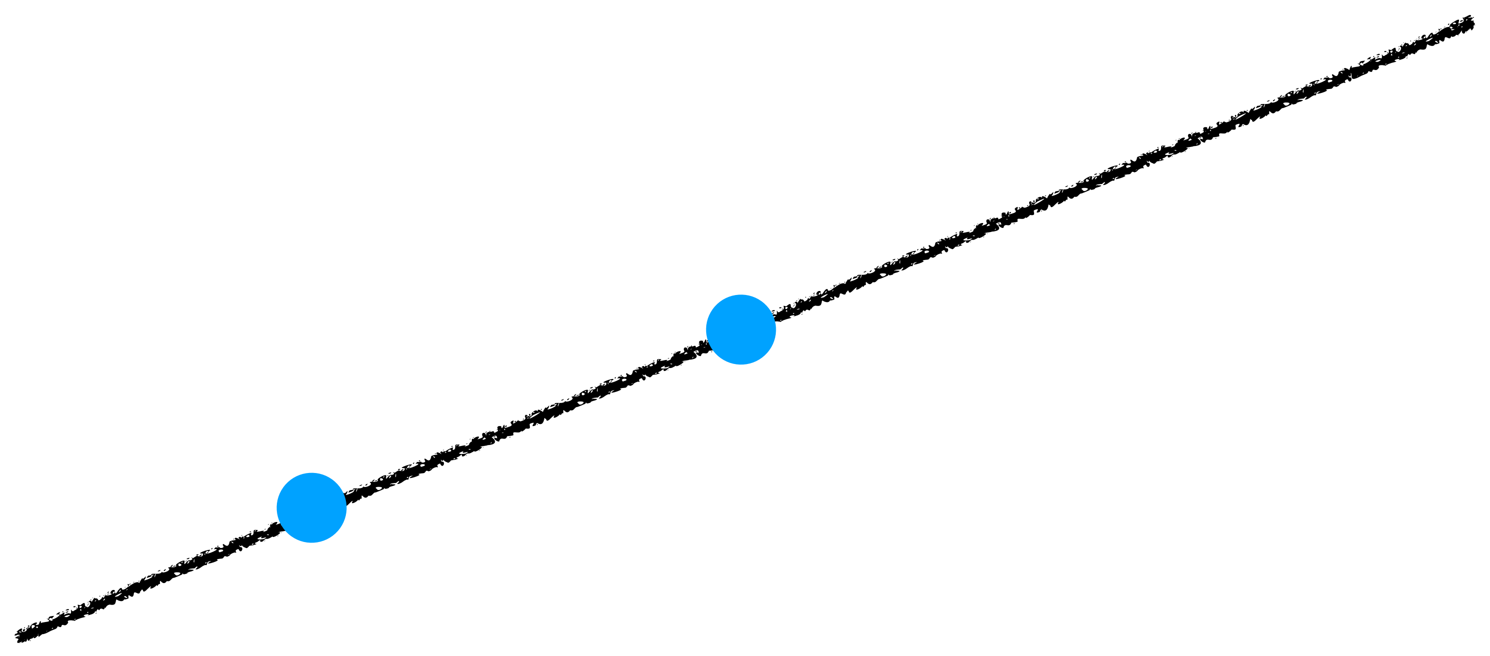
Robust Decoding or State Estimation

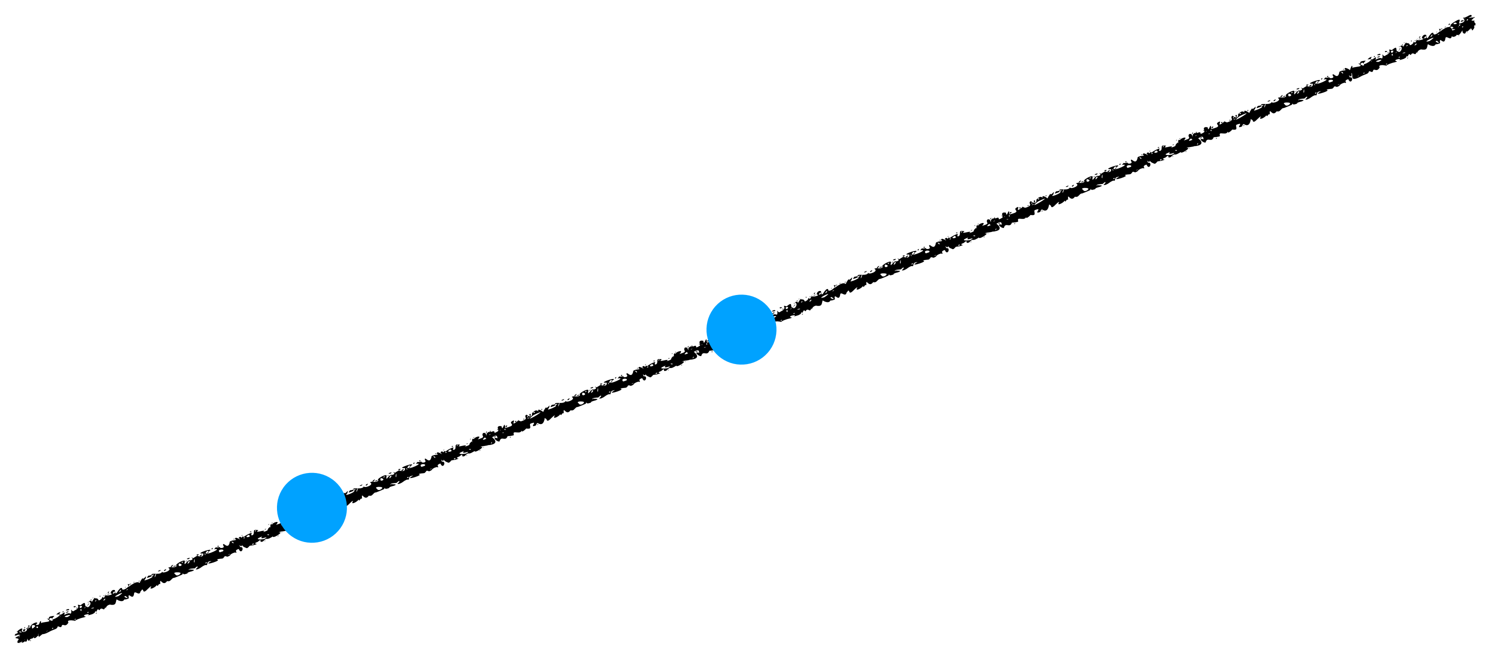
Consider $y = Ax$ where $A \in \mathbb{R}^{n \times m}$, and observation $y \in \mathbb{R}^n$

Suppose that up to f of the observations are arbitrarily corrupted

We can recover x from $\tilde{y} \triangleq y + \delta$ where $||\delta||_0 \leq f$

As long as $\operatorname{rank}\left(A^S\right) = m$ where $S \subseteq [n]$ such that $|S| = n - 2f$





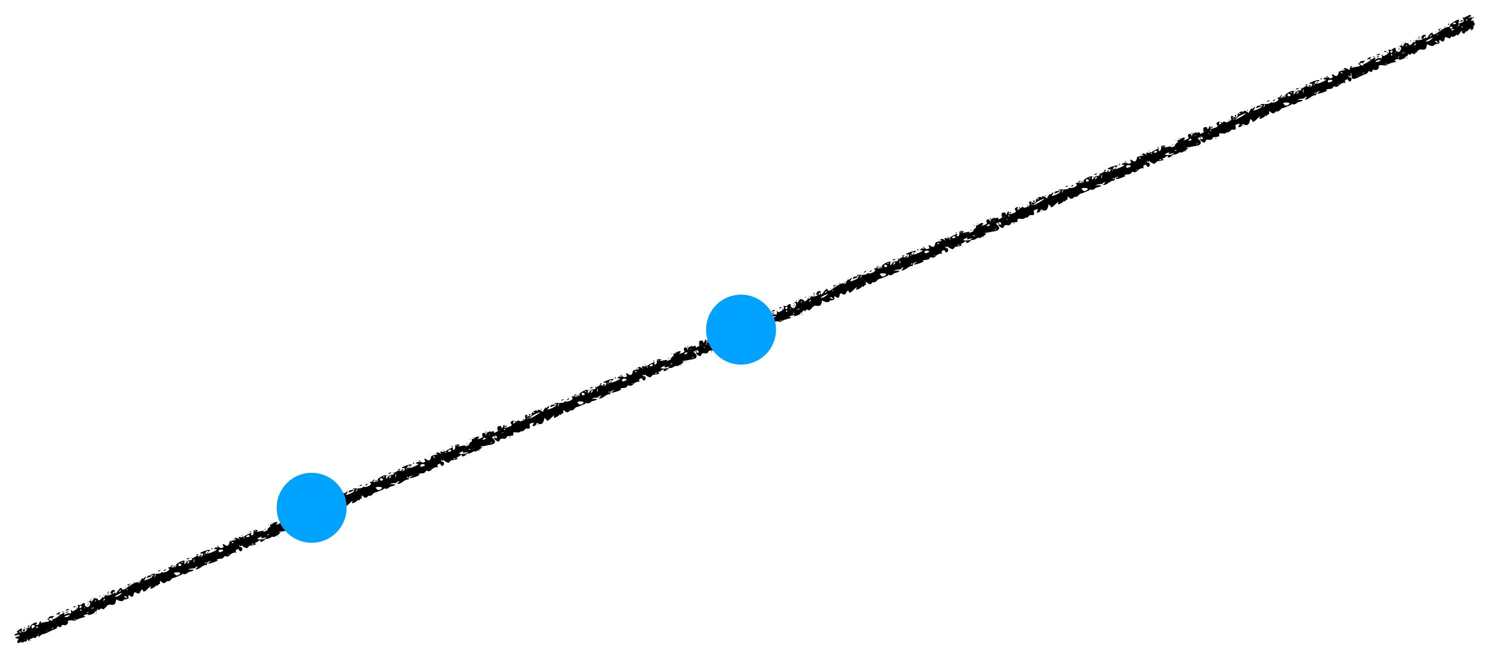
Robust Decoding or State Estimation

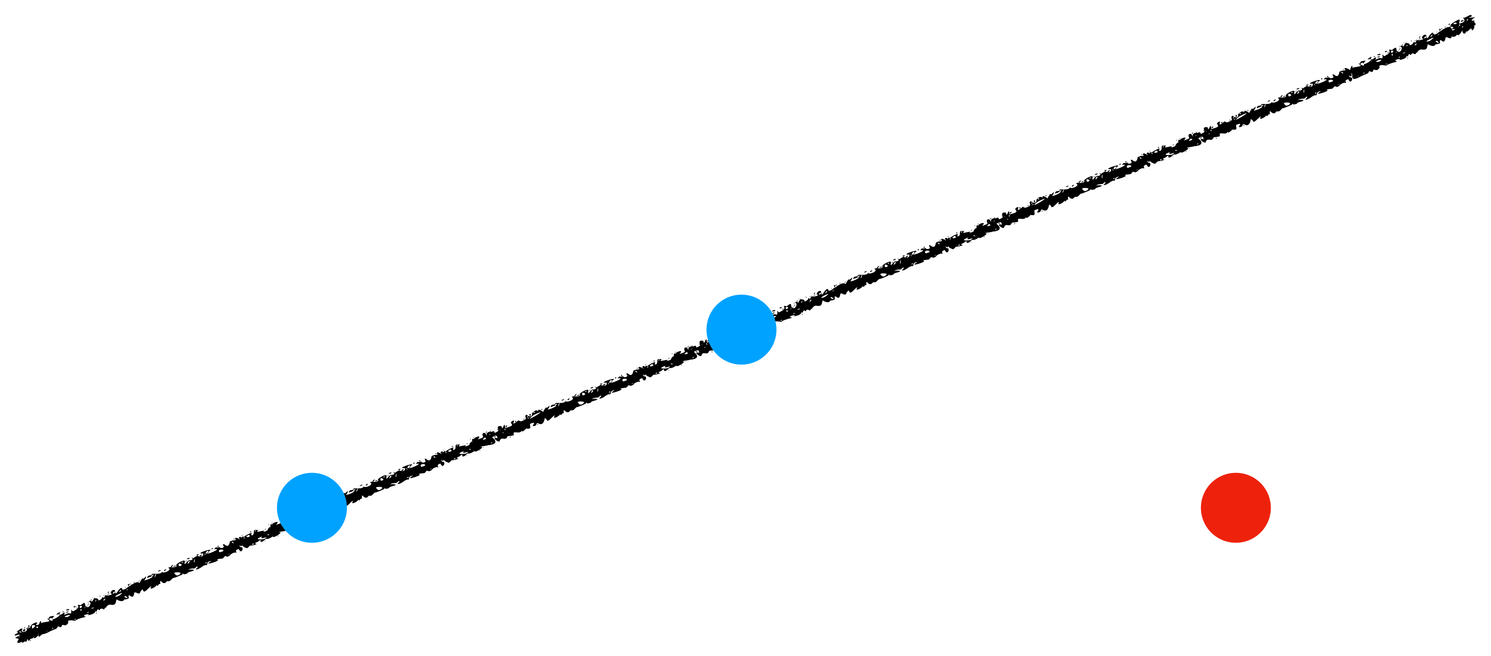
Consider $y = Ax$ where $A \in \mathbb{R}^{n \times m}$, and observation $y \in \mathbb{R}^n$

Suppose that up to f of the observations are arbitrarily corrupted

We can recover x from $\tilde{y} \triangleq y + \delta$ where $||\delta||_0 \leq f$

As long as $\operatorname{rank}\left(A^S\right) = m$ where $S \subseteq [n]$ such that $|S| = n - 2f$





Robust Decoding or State Estimation

Consider $y = Ax$ where $A \in \mathbb{R}^{n\times m}$, and observation $y \in \mathbb{R}^n$

Suppose that up to f of the observations are arbitrarily corrupted

We can recover x from $\tilde{y} \triangleq y + \delta$ where $||\delta||_0 \leq f$

As long as $\operatorname{rank}\left(A^S\right) = m$ where $S \subseteq [n]$ such that $|S| = n - 2f$

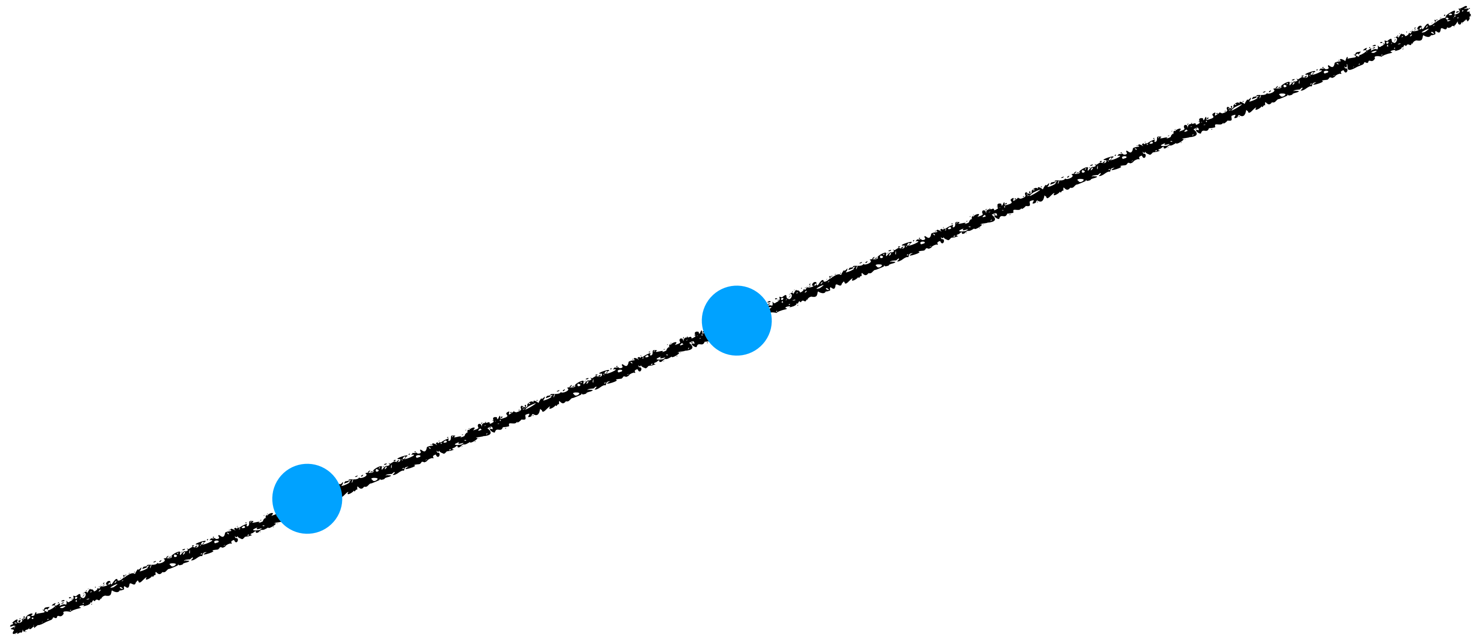
Robust Decoding or State Estimation

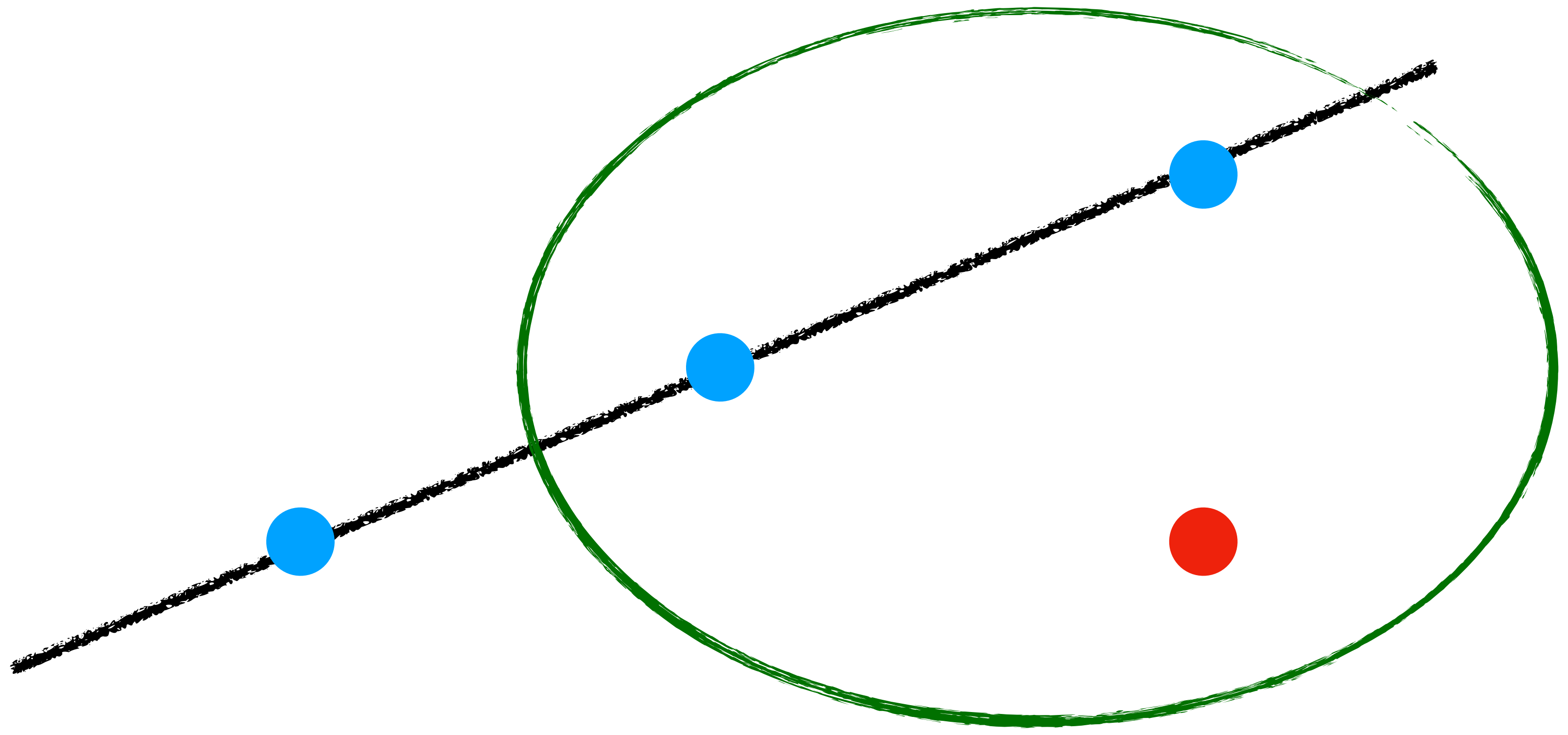
Consider $y = Ax$ where $A \in \mathbb{R}^{n \times m}$, and observation $y \in \mathbb{R}^n$

Suppose that up to f of the observations are arbitrarily corrupted

We can recover x from $\tilde{y} \triangleq y + \delta$ where $||\delta||_0 \leq f$

As long as $\mathrm{rank}\left(A^S\right) = m$ where $S \subseteq [n]$ such that $|S| = n - 2f$





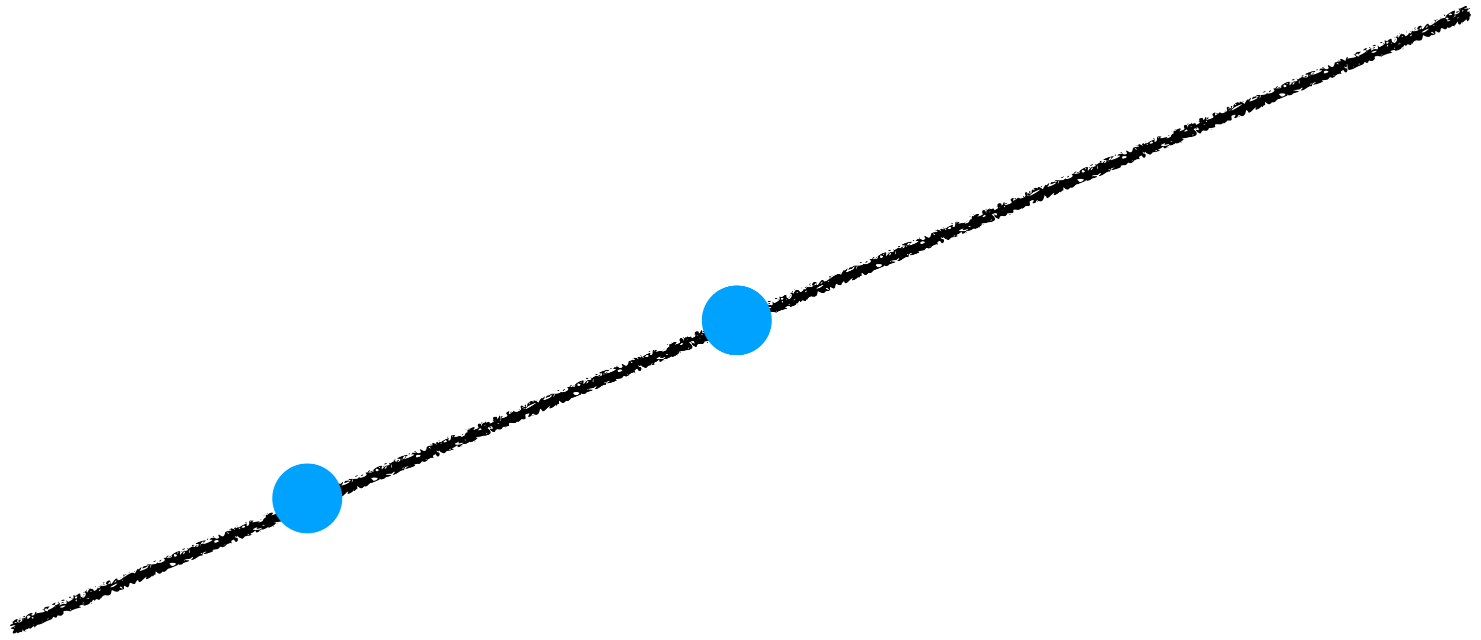
Robust Decoding or State Estimation

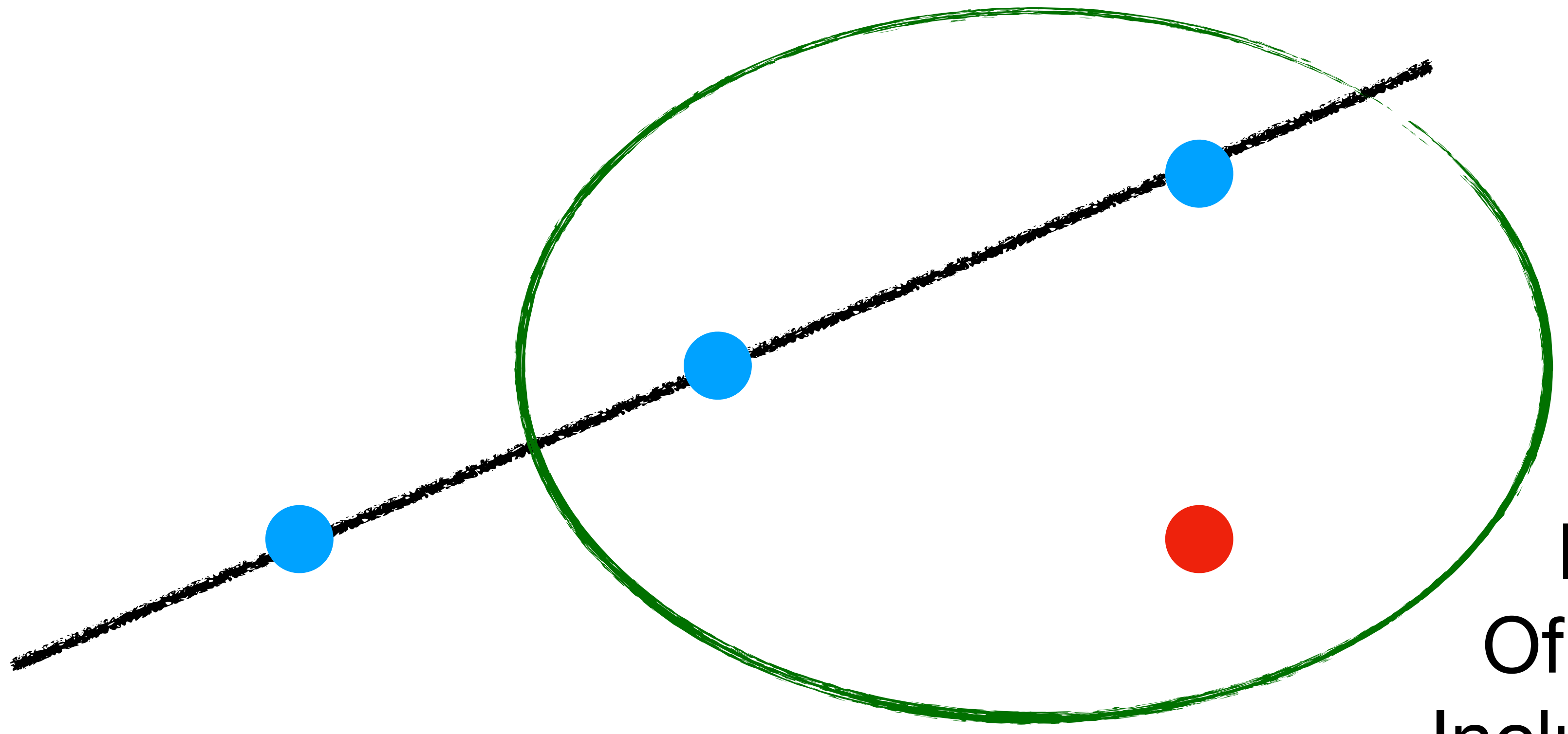
Consider $y = Ax$ where $A \in \mathbb{R}^{n \times m}$, and observation $y \in \mathbb{R}^n$

Suppose that up to f of the observations are arbitrarily corrupted

We can recover x from $\tilde{y} \triangleq y + \delta$ where $||\delta||_0 \leq f$

As long as $\mathrm{rank}\left(A^S\right) = m$ where $S \subseteq [n]$ such that $|S| = n - 2f$





Each quorum
Of 3 observations
Includes the solution

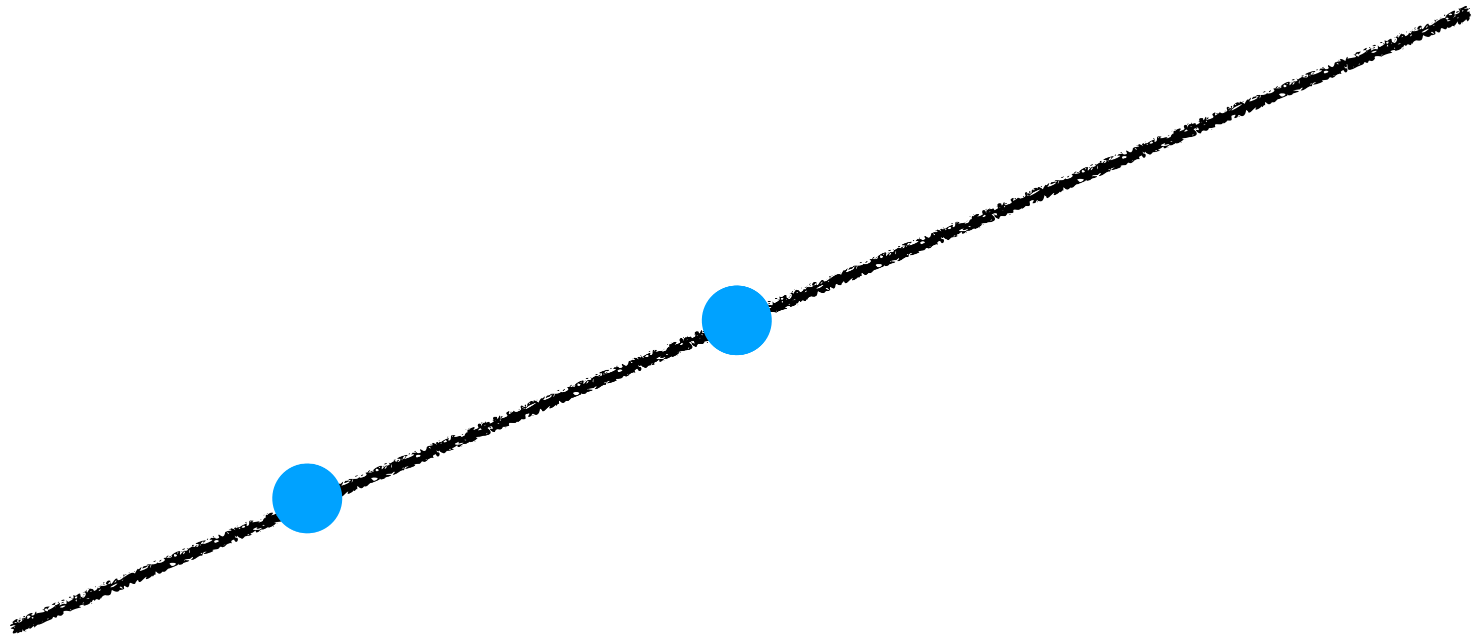
Robust Decoding or State Estimation

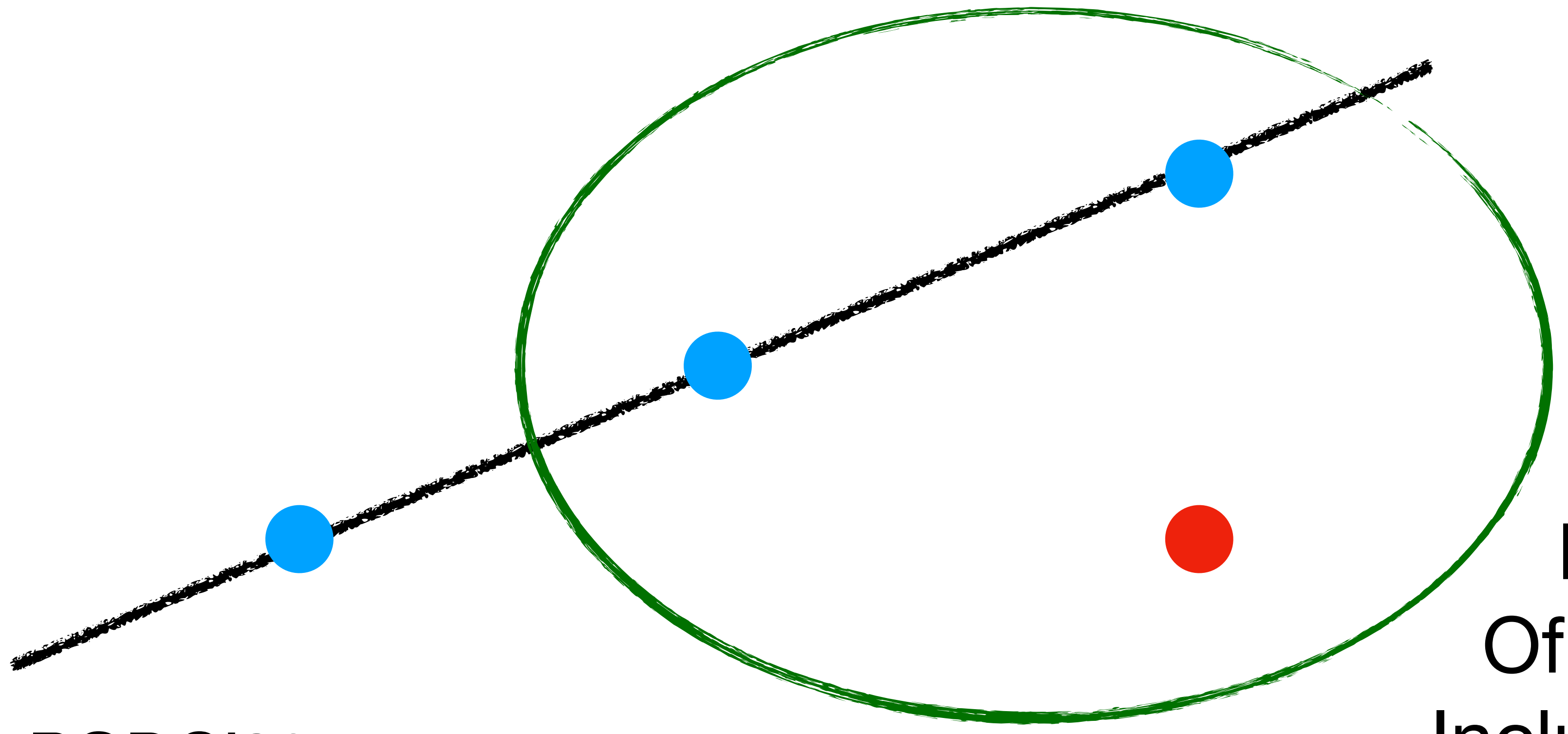
Consider $y = Ax$ where $A \in \mathbb{R}^{n \times m}$, and observation $y \in \mathbb{R}^n$

Suppose that up to f of the observations are arbitrarily corrupted

We can recover x from $\tilde{y} \triangleq y + \delta$ where $||\delta||_0 \leq f$

As long as $\text{rank}\left(A^S\right) = m$ where $S \subseteq [n]$ such that $|S| = n - 2f$





Each quorum
Of 3 observations
Includes the solution

“Fault-Tolerance in Distributed Optimization: The Case of Redundancy” **Gupta** and Vaidya, PODC’20

How about Gradient Descent

Applicability to Distributed Gradient-Descent

Applicability to Distributed Gradient-Descent

How to *adapt* functional redundancy for *analyzing* resilience of *robust DGD*

Applicability to Distributed Gradient-Descent

How to *adapt* functional redundancy for *analyzing* resilience of *robust DGD*

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

Applicability to Distributed Gradient-Descent

How to *adapt* functional redundancy for *analyzing* resilience of *robust DGD*

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

$$g_t^i := \nabla \mathcal{L}_i(\theta_t)$$

Applicability to Distributed Gradient-Descent

How to *adapt* functional redundancy for *analyzing* resilience of *robust DGD*

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

$$g_t^i := \nabla \mathscr{L}_i(\theta_t)$$

Global Phase: Receiving gradients $g_t^1, \ldots, g_t^n$ the server "robustly" aggregates them, i.e., compute

Applicability to Distributed Gradient-Descent

How to *adapt* functional redundancy for *analyzing* resilience of *robust DGD*

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

$$g_t^i := \nabla \mathscr{L}_i(\theta_t)$$

Global Phase: Receiving gradients $g_t^1, \ldots, g_t^n$ the server "robustly" aggregates them, i.e., compute

$$\widehat{g}_t := F\left(g_t^1, \ldots, g_t^n\right) \ ,$$

Applicability to Distributed Gradient-Descent

How to *adapt* functional redundancy for *analyzing* resilience of *robust DGD*

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

$$g_t^i := \nabla \mathscr{L}_i(\theta_t)$$

Global Phase: Receiving gradients $g_t^1, \ldots, g_t^n$ the server "robustly" aggregates them, i.e., compute

$$\widehat{g}_t := F\left(g_t^1, \ldots, g_t^n\right) \quad ,$$

And updates the current parameters: $\theta_{t+1} = \theta_t - \gamma_t \ \widehat{g}_t$

Bounded Gradient Dissimilarity

Bounded Gradient Dissimilarity

L-smooth local losses, i.e., $\|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_i(\theta')\| \leq L\|\theta - \theta'\|$

μ-PL (Polyak-Lojasiewicz) average loss function, i.e., $\|\nabla \mathscr{L}_H(\theta)\|^2 \geq 2\mu \left(\mathscr{L}_H(\theta) - \min \mathscr{L}_H\right)$

$(2f, \varepsilon)$-redundancy replaced by bounded gradient dissimilarity:

Bounded Gradient Dissimilarity

L-smooth local losses, i.e., $\|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_i(\theta')\| \leq L\|\theta - \theta'\|$

μ-PL (Polyak-Lojasiewicz) average loss function, i.e., $\|\nabla \mathscr{L}_H(\theta)\|^2 \geq 2\mu \left(\mathscr{L}_H(\theta) - \min \mathscr{L}_H\right)$

$(2f, \varepsilon)$-redundancy replaced by bounded gradient dissimilarity:

$$\frac{1}{|H|}\sum_{i \in H}\|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_H(\theta)\|^2 \;\leq\; G^2 + B^2\;\|\nabla \mathscr{L}_H(\theta)\|^2$$

Bounded Gradient Dissimilarity

L-smooth local losses, i.e., $\|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_i(\theta')\| \leq L\|\theta - \theta'\|$

μ-PL (Polyak-Lojasiewicz) average loss function, i.e., $\|\nabla \mathscr{L}_H(\theta)\|^2 \geq 2\mu \left(\mathscr{L}_H(\theta) - \min \mathscr{L}_H\right)$

$(2f, \varepsilon)$-redundancy replaced by bounded gradient dissimilarity:

$$\frac{1}{|H|} \sum_{i \in H} \|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_H(\theta)\|^2 \;\leq\; G^2 + B^2 \;\|\nabla \mathscr{L}_H(\theta)\|^2$$

$$G^2 = \frac{2L}{|H|} \sum_{i \in H} \left(\mathscr{L}_i(\theta^*) - \min \mathscr{L}_i\right)$$

Bounded Gradient Dissimilarity

L-smooth local losses, i.e., $\|\nabla \mathcal{L}_i(\theta) - \nabla \mathcal{L}_i(\theta')\| \leq L\|\theta - \theta'\|$

μ-PL (Polyak-Lojasiewicz) average loss function, i.e., $\|\nabla \mathcal{L}_H(\theta)\|^2 \geq 2\mu \left(\mathcal{L}_H(\theta) - \min \mathcal{L}_H\right)$

$(2f, \varepsilon)$-redundancy replaced by bounded gradient dissimilarity:

$$\frac{1}{|H|}\sum_{i \in H}\|\nabla \mathcal{L}_i(\theta) - \nabla \mathcal{L}_H(\theta)\|^2 \;\leq\; G^2 + B^2\;\|\nabla \mathcal{L}_H(\theta)\|^2$$

$$G^2 = \frac{2L}{|H|}\sum_{i \in H}\left(\mathcal{L}_i(\theta^*) - \min \mathcal{L}_i\right) \qquad B^2 = 2K_{\mathcal{L}} - 1 \;;\; K_{\mathcal{L}} := \frac{L}{\mu}$$

Bounded Gradient Dissimilarity

L-smooth local losses, i.e., $\|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_i(\theta')\| \leq L\|\theta - \theta'\|$

μ-PL (Polyak-Lojasiewicz) average loss function, i.e., $\|\nabla \mathscr{L}_H(\theta)\|^2 \geq 2\mu \left(\mathscr{L}_H(\theta) - \min \mathscr{L}_H\right)$

$(2f, \varepsilon)$-redundancy replaced by bounded gradient dissimilarity:

$$\frac{1}{|H|}\sum_{i \in H}\|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_H(\theta)\|^2 \;\leq\; G^2 + B^2\;\|\nabla \mathscr{L}_H(\theta)\|^2$$

$$G^2 = \frac{2L}{|H|}\sum_{i \in H}\left(\mathscr{L}_i(\theta^*) - \min \mathscr{L}_i\right) \qquad B^2 = 2K_{\mathscr{L}} - 1\;;\;\; K_{\mathscr{L}} := \frac{L}{\mu}$$

$$\theta^* := \arg\min \mathscr{L}_H(\theta)$$

Bounded Gradient Dissimilarity

L-smooth local losses, i.e., $\|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_i(\theta')\| \leq L\|\theta - \theta'\|$

μ-PL (Polyak-Lojasiewicz) average loss function, i.e., $\|\nabla \mathscr{L}_H(\theta)\|^2 \geq 2\mu \left(\mathscr{L}_H(\theta) - \min \mathscr{L}_H\right)$

$(2f, \varepsilon)$-redundancy replaced by bounded gradient dissimilarity:

$$\frac{1}{|H|}\sum_{i \in H}\|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_H(\theta)\|^2 \;\leq\; G^2 + B^2 \;\|\nabla \mathscr{L}_H(\theta)\|^2$$

$$G^2 = \frac{2L}{|H|}\sum_{i \in H}\left(\mathscr{L}_i(\theta^*) - \min \mathscr{L}_i\right) \qquad B^2 = 2K_{\mathscr{L}} - 1 \;;\; K_{\mathscr{L}} := \frac{L}{\mu}$$

$$\theta^* := \arg\min \mathscr{L}_H(\theta)$$

Bounded Gradient Dissimilarity

L-smooth local losses, i.e., $\|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_i(\theta')\| \leq L\|\theta - \theta'\|$

μ-PL (Polyak-Lojasiewicz) average loss function, i.e., $\|\nabla \mathscr{L}_H(\theta)\|^2 \geq 2\mu \left(\mathscr{L}_H(\theta) - \min \mathscr{L}_H\right)$

$(2f, \varepsilon)$-redundancy replaced by bounded gradient dissimilarity:

$$\frac{1}{|H|} \sum_{i \in H} \|\nabla \mathscr{L}_i(\theta) - \nabla \mathscr{L}_H(\theta)\|^2 \;\leq\; G^2 + B^2 \;\|\nabla \mathscr{L}_H(\theta)\|^2$$

$$G^2 = \frac{2L}{|H|} \sum_{i \in H} \left(\mathscr{L}_i(\theta^*) - \min \mathscr{L}_i\right) \qquad B^2 = 2K_{\mathscr{L}} - 1 \;;\; K_{\mathscr{L}} := \frac{L}{\mu}$$

$$\theta^* := \arg\min \mathscr{L}_H(\theta)$$

Condition number

Resilience under (G, B)-Gradient Dissimilarity

Resilience under (G, B)-Gradient Dissimilarity

Suppose that aggregation F is (f, κ)-robust averaging

Resilience under (G, B)-Gradient Dissimilarity

Suppose that aggregation F is (f, κ)-robust averaging

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^H\|^2 \leq \kappa \frac{1}{|H|} \sum_{i \in H} \|g_t^i - g_t^H\|^2$$

Resilience under (G, B)-Gradient Dissimilarity

Suppose that aggregation F is (f, κ)-robust averaging

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^H\|^2 \leq \kappa \frac{1}{|H|} \sum_{i \in H} \|g_t^i - g_t^H\|^2$$

Robust DGD is (f, ε)-resilient with

Resilience under (G, B)-Gradient Dissimilarity

Suppose that aggregation F is (f, κ)-robust averaging

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^H\|^2 \leq \kappa \frac{1}{|H|} \sum_{i \in H} \|g_t^i - g_t^H\|^2$$

Robust DGD is (f, ε)-resilient with

$$\varepsilon \in \mathcal{O}\left(\frac{\kappa G^2}{1 - \kappa B^2} + \exp\left(-\frac{\left(1 - \kappa B^2\right)}{K_{\mathcal{L}}} T\right)\right)$$

Resilience under (G, B)-Gradient Dissimilarity

Suppose that aggregation F is (f, κ)-robust averaging

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^H\|^2 \leq \kappa \frac{1}{|H|} \sum_{i \in H} \|g_t^i - g_t^H\|^2$$

Robust DGD is (f, ε)-resilient with

$$\varepsilon \in \mathcal{O}\left(\frac{\kappa G^2}{1 - \kappa B^2} + \exp\left(-\frac{\left(1 - \kappa B^2\right)}{K_{\mathscr{L}}} T\right)\right)$$

Resilience under (G, B)-Gradient Dissimilarity

Suppose that aggregation F is (f, κ)-robust averaging

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^H\|^2 \leq \kappa \frac{1}{|H|} \sum_{i \in H} \|g_t^i - g_t^H\|^2$$

Robust DGD is (f, ε)-resilient with

$$\varepsilon \in \mathcal{O}\left(\frac{\kappa G^2}{1 - \kappa B^2} + \exp\left(-\frac{\left(1 - \kappa B^2\right)}{K_{\mathscr{L}}} T\right)\right)$$

Resilience under (G, B)-Gradient Dissimilarity

Suppose that aggregation F is (f, κ)-robust averaging

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^H\|^2 \leq \kappa \frac{1}{|H|} \sum_{i \in H} \|g_t^i - g_t^H\|^2$$

Robust DGD is (f, ε)-resilient with

$$\varepsilon \in \mathcal{O}\left(\frac{\kappa G^2}{1 - \kappa B^2} + \exp\left(-\frac{\left(1 - \kappa B^2\right)}{K_{\mathscr{L}}} T\right)\right)$$

Limits the robustness parameter f

Resilience under (G, B)-Gradient Dissimilarity

Suppose that aggregation F is (f, κ)-robust averaging

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^H\|^2 \leq \kappa \frac{1}{|H|} \sum_{i \in H} \|g_t^i - g_t^H\|^2$$

Robust DGD is (f, ε)-resilient with

$$\varepsilon \in \mathcal{O}\left(\frac{\kappa G^2}{1 - \kappa B^2} + \exp\left(-\frac{\left(1 - \kappa B^2\right)}{K_{\mathcal{L}}} T\right)\right)$$

Limits the robustness parameter f

Provided that $\kappa B^2 < 1$

Resilience under (G, B)-Gradient Dissimilarity

Suppose that aggregation F is (f, κ)-robust averaging

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^H\|^2 \leq \kappa \frac{1}{|H|} \sum_{i \in H} \|g_t^i - g_t^H\|^2$$

Robust DGD is (f, ε)-resilient with

$$\varepsilon \in \mathcal{O}\left(\frac{\kappa G^2}{1 - \kappa B^2} + \exp\left(-\frac{\left(1 - \kappa B^2\right)}{K_{\mathcal{L}}} T\right)\right)$$

Limits the robustness parameter f

Provided that $\kappa B^2 < 1$

"Robust Distributed Learning: Tight Error Bounds and Breakdown Point under Data Heterogeneity." Y. Allouah et al. NeurIPS'23 [Spotlight]

Condition Number Strikes Again

Condition Number Strikes Again

Assuming (G, B)-gradient dissimilarity, it is generally impossible to tolerate f adversarial nodes if $\frac{f}{n} \geq \frac{1}{2 + B^2}$

Condition Number Strikes Again

Assuming (G, B)-gradient dissimilarity, it is generally impossible to tolerate f adversarial nodes if $\dfrac{f}{n} \geq \dfrac{1}{2 + B^2}$

"Robust Distributed Learning: Tight Error Bounds and Breakdown Point under Data Heterogeneity." Y. Allouah at al. NeurIPS'23 [Spotlight]

Condition Number Strikes Again

Assuming (G, B)-gradient dissimilarity, it is generally impossible to tolerate f adversarial nodes if $\dfrac{f}{n} \geq \dfrac{1}{2 + B^2}$

"Robust Distributed Learning: Tight Error Bounds and Breakdown Point under Data Heterogeneity." Y. Allouah at al. NeurIPS'23 [Spotlight]

Under *homogeneity* we can tolerate up to $\dfrac{n}{2}$ adversarial nodes

Condition Number Strikes Again

Assuming (G, B)-gradient dissimilarity, it is generally impossible to tolerate f adversarial nodes if $\frac{f}{n} \geq \frac{1}{2+B^2}$

"Robust Distributed Learning: Tight Error Bounds and Breakdown Point under Data Heterogeneity." Y. Allouah at al. NeurIPS'23 [Spotlight]

Under *homogeneity* we can tolerate up to $\frac{n}{2}$ adversarial nodes

Recall that $B^2 = 2K_{\mathscr{L}} - 1$

Condition Number Strikes Again

Assuming (G, B)-gradient dissimilarity, it is generally impossible to tolerate f adversarial nodes if $\frac{f}{n} \geq \frac{1}{2 + B^2}$

"Robust Distributed Learning: Tight Error Bounds and Breakdown Point under Data Heterogeneity." Y. Allouah at al. NeurIPS'23 [Spotlight]

Under *homogeneity* we can tolerate up to $\frac{n}{2}$ adversarial nodes

Recall that $B^2 = 2K_{\mathscr{L}} - 1$

We cannot tolerate $\frac{f}{n} \geq \frac{1}{1 + 2K_{\mathscr{L}}}$, where recall that $K_{\mathscr{L}} \geq 1$

Lower Bound on Training Error

Lower Bound on Training Error

Under (G, B)-gradient dissimilarity, it is generally impossible to achieve (f, ε)-resilience for

Lower Bound on Training Error

Under (G, B)-gradient dissimilarity, it is generally impossible to achieve (f, ε)-resilience for

$$\varepsilon < \frac{1}{8\mu}\left(\frac{f}{n-(2+B^2)\,f}\;G^2\right)$$

Lower Bound on Training Error

Under (G, B)-gradient dissimilarity, it is generally impossible to achieve (f, ε)-resilience for

$$\varepsilon < \frac{1}{8\mu}\left(\frac{f}{n-(2+B^2)\,f}\;G^2\right)$$

Recall that $G^2 = \dfrac{2L}{|H|}\sum_{i \in H}\left(\mathcal{L}_i(\theta^*) - \min \mathcal{L}_i\right)$

Lower Bound on Training Error

Under (G, B)-gradient dissimilarity, it is generally impossible to achieve (f, ε)-resilience for

$$\varepsilon < \frac{1}{8\mu}\left(\frac{f}{n-(2+B^2)\,f}\;G^2\right)$$

Recall that $G^2 = \dfrac{2L}{|H|}\displaystyle\sum_{i\in H}\left(\mathscr{L}_i(\theta^*) - \min \mathscr{L}_i\right)$

$$\varepsilon \in \Omega\left(\frac{f}{n}\;K_{\mathscr{L}}\right)$$

Lower Bound on Training Error

Under (G, B)-gradient dissimilarity, it is generally impossible to achieve (f, ε)-resilience for

$$\varepsilon < \frac{1}{8\mu}\left(\frac{f}{n - (2 + B^2)\, f}\; G^2\right)$$

Recall that $G^2 = \dfrac{2L}{|H|} \displaystyle\sum_{i \in H} \left(\mathscr{L}_i(\theta^*) - \min \mathscr{L}_i\right)$

$$\varepsilon \in \Omega\left(\frac{f}{n}\; K_{\mathscr{L}}\right)$$

“Robust Distributed Learning: Tight Error Bounds and Breakdown Point under Data Heterogeneity.” Y. Allouah et al. NeurIPS’23 [Spotlight]

Rendering Optimal Robustness to DGD

Rendering Optimal Robustness to DGD

In general, (f, κ)-robust averaging is impossible for $\kappa < \dfrac{f}{n - 2f}$

Rendering Optimal Robustness to DGD

In general, (f, κ)-robust averaging is impossible for $\kappa < \dfrac{f}{n - 2f}$

Coordinate-wise Trimmed Mean (CWTM) matches this bound, up to a *small* constant factor

Rendering Optimal Robustness to DGD

In general, (f, κ)-robust averaging is impossible for $\kappa < \dfrac{f}{n - 2f}$

Coordinate-wise Trimmed Mean (CWTM) matches this bound, up to a *small* constant factor

When $\kappa \leq c \dfrac{f}{n - 2f}$ we have $\varepsilon \in \mathcal{O}\left(\dfrac{f}{n - (2 + B^2)\, f}\, G^2 + e^{-\frac{T}{K_{\mathscr{L}}}}\right)$

Nearest Neighbor Mixing: Order-Optimal Robustness

Nearest Neighbor Mixing: Order-Optimal Robustness

We can have efficient rules that are order optimal, i.e., $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

Nearest Neighbor Mixing: Order-Optimal Robustness

We can have efficient rules that are order optimal, i.e., $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

NNM is a *pre-aggregation scheme* that imparts order-optimal robustness to many robust aggregation rules

Nearest Neighbor Mixing: Order-Optimal Robustness

We can have efficient rules that are order optimal, i.e., $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

NNM is a *pre-aggregation scheme* that imparts order-optimal robustness to many robust aggregation rules

If F is (f, κ)-robust averaging with $\kappa \in \mathcal{O}(1)$ then
$F \cdot$ NNM is (f, κ)-robust averaging with $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

Nearest Neighbor Mixing: Order-Optimal Robustness

We can have efficient rules that are order optimal, i.e., $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

NNM is a *pre-aggregation scheme* that imparts order-optimal robustness to many robust aggregation rules

If F is (f, κ)-robust averaging with $\kappa \in \mathcal{O}(1)$ then

$F \cdot$ NNM is (f, κ)-robust averaging with $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

"Fixing by Mixing: A Recipe for Optimal Byzantine ML under Heterogeneity." Y. Allouah at al. AISTATS'23

NNM Pre-aggregation Scheme

NNM Pre-aggregation Scheme

For each input vector v_i determine $n - f$ nearest neighbors in the set of input vectors $\{v_1, \ldots, v_n\}$

NNM Pre-aggregation Scheme

For each input vector v_i determine $n - f$ nearest neighbors in the set of input vectors $\{v_1, \ldots, v_n\}$

Let N_i be the set of $n - f$ vectors nearest to v_i

NNM Pre-aggregation Scheme

For each input vector v_i determine $n - f$ nearest neighbors in the set of input vectors $\{v_1, \ldots, v_n\}$

Let N_i be the set of $n - f$ vectors nearest to v_i

Map v_i to $z_i := \frac{1}{n-f} \sum_{v \in N_i} v$

NNM Pre-aggregation Scheme

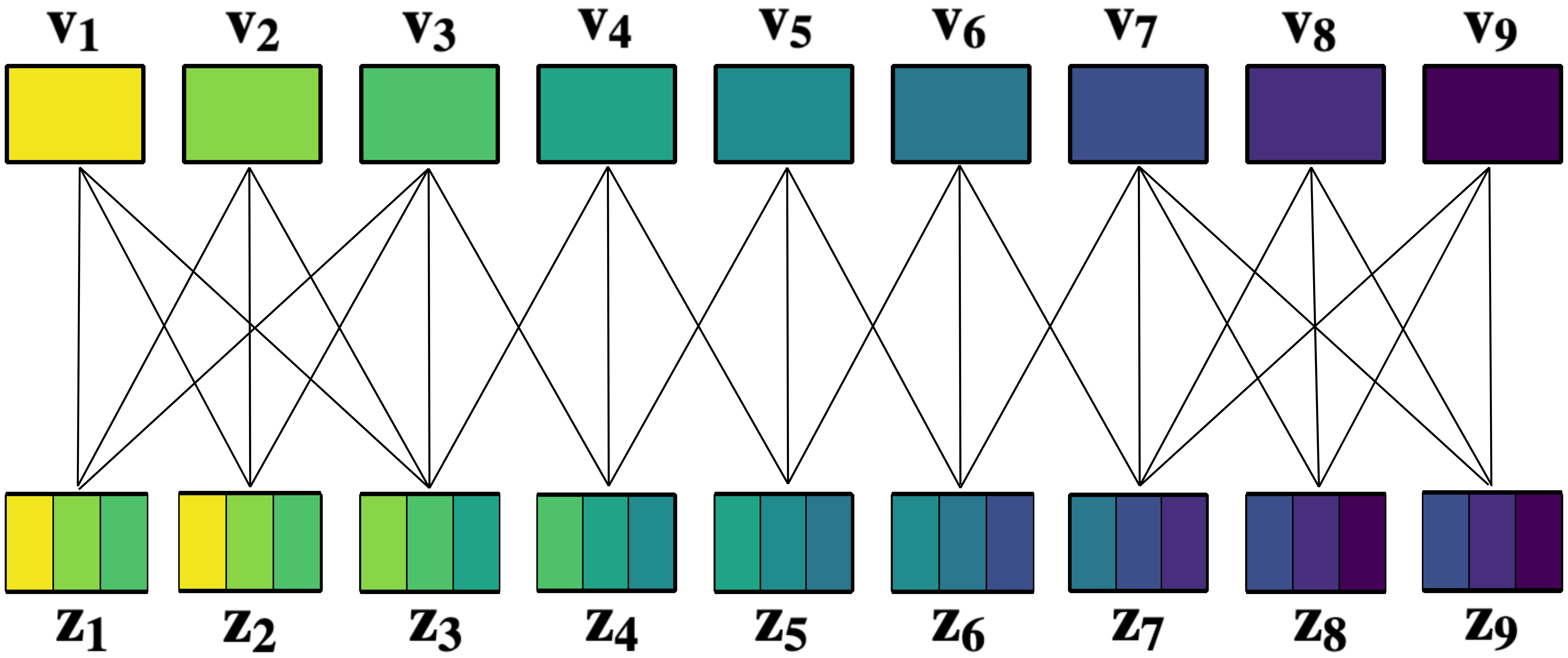
For each input vector v_i determine $n - f$ nearest neighbors in the set of input vectors $\{v_1, \ldots, v_n\}$

Let N_i be the set of $n - f$ vectors nearest to v_i

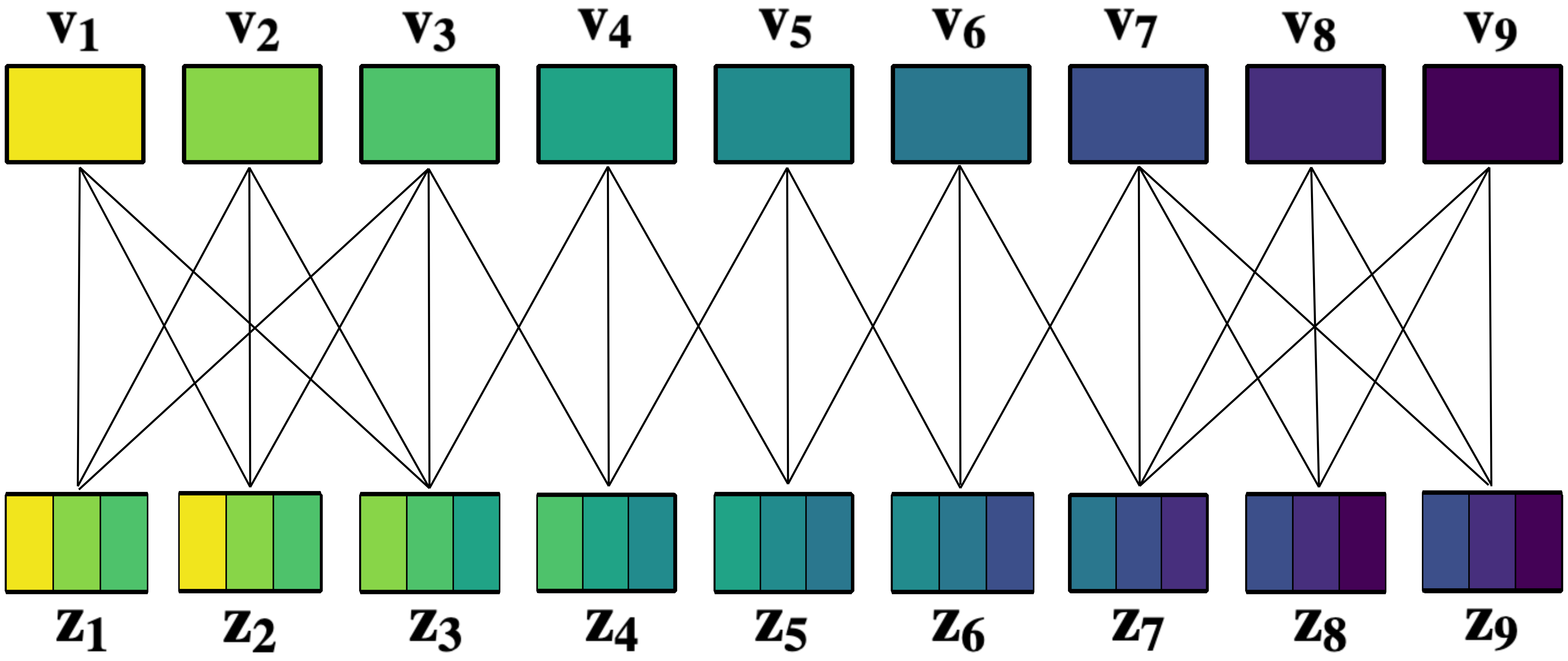
Map v_i to $z_i := \dfrac{1}{n-f} \sum_{v \in N_i} v$

Define $F \cdot NNM\left(v_1, \ldots, v_n\right) = F\left(z_1, \ldots, z_n\right)$

Intuition on Why NNM Works

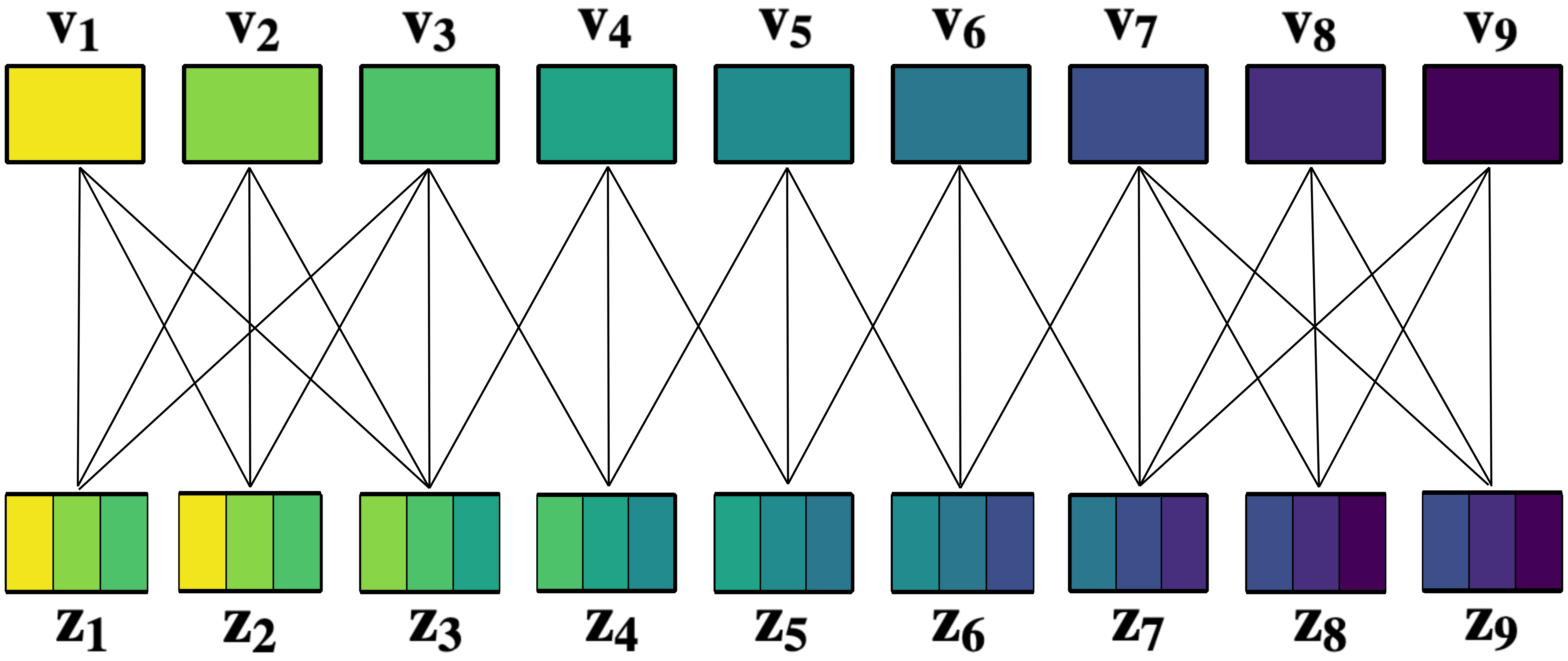


Intuition on Why NNM Works



Variance of z_i's is less than v_i's by factor $\mathcal{O}\left(\frac{f}{n}\right)$

Intuition on Why NNM Works



Variance of z_i's is less than v_i's by factor $\mathcal{O}\left(\frac{f}{n}\right)$

"Fixing by Mixing: A Recipe for Optimal Byzantine ML under Heterogeneity." Y. Allouah et al. AISTATS'23

Empirical Observations

Agg. Rule	ALIE	FOE	SF	Worst-Case
GeoMed	92.01 ± 04.35	65.61 ± 12.17	57.86 ± 10.42	57.86 ± 10.42
+ NNM	81.26 ± 08.91	75.27 ± 02.69	86.32 ± 03.77	75.27 ± 02.69
+ Bucketing	39.83 ± 11.35	44.73 ± 16.47	91.30 ± 03.91	44.73 ± 16.47

Agg. Rule	ALIE	FOE	SF	Worst-Case
CWTM	76.16 ± 07.68	69.96 ± 16.57	27.45 ± 08.83	27.45 ± 08.83
+ NNM	79.04 ± 09.19	79.91 ± 03.94	84.78 ± 05.78	79.04 ± 09.19
+ Bucketing	55.86 ± 10.00	42.80 ± 21.25	50.96 ± 16.53	42.80 ± 21.25

CNN trained on MNIST dataset, distributed among 13 honest nodes with Dirichlet parameter of 0.1 (extreme heterogeneity). There are 4 additional adversarial nodes executing attacks: ALIE, FOE and SF. We run 800 iterations, with local batch-size of 25.

Challenge of Privacy

Differential Privacy in Distributed Mini-batch GD

Differential Privacy in Distributed Mini-batch GD

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

Differential Privacy in Distributed Mini-batch GD

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

$$g_t^i := \frac{1}{b} \sum_{z \in S_t^{(i)}} \mathrm{Clip}\left(\nabla_\theta \ell(\theta_t, z) \ , \ C \right) + \eta_t$$

Differential Privacy in Distributed Mini-batch GD

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

$$g_t^i := \frac{1}{b} \sum_{z \in S_t^{(i)}} \text{Clip}\left(\nabla_\theta \ell(\theta_t, z) \ , \ C \right) + \eta_t$$

with $\eta_t \sim \mathcal{N}\left(0, \ \sigma_{\text{DP}}^2 \ I_d\right)$, where $\text{Clip}(v, \ C) = \min\left\{1, \ \frac{C}{\|v\|}\right\} \ v$

Differential Privacy in Distributed Mini-batch GD

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

$$g_t^i := \frac{1}{b} \sum_{z \in S_t^{(i)}} \text{Clip}\left(\nabla_\theta \ell(\theta_t, z) \ , \ C \right) + \eta_t$$

with $\eta_t \sim \mathcal{N}\left(0, \ \sigma_{\text{DP}}^2 \ I_d\right)$, where $\text{Clip}(v, \ C) = \min\left\{ 1, \ \frac{C}{\|v\|} \right\} \ v$

Global Phase: Receiving gradients $g_t^1, \ldots, g_t^n$ the server "robustly" aggregates them, i.e., compute

Differential Privacy in Distributed Mini-batch GD

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

$$g_t^i := \frac{1}{b} \sum_{z \in S_t^{(i)}} \mathrm{Clip}\left(\nabla_\theta \ell(\theta_t, z)\ ,\ C \right) + \eta_t$$

with $\eta_t \sim \mathcal{N}\left(0,\ \sigma_{\mathrm{DP}}^2\ I_d\right)$, where $\mathrm{Clip}(v,\ C) = \min\left\{1,\ \frac{C}{\|v\|}\right\}\ v$

Global Phase: Receiving gradients $g_t^1, \ldots, g_t^n$ the server "robustly" aggregates them, i.e., compute

$$\widehat{g}_t := F\left(g_t^1, \ldots, g_t^n\right)\ ,$$

Differential Privacy in Distributed Mini-batch GD

Local Phase: In each iteration t, each *honest* node i computes the local gradient:

$$g_t^i := \frac{1}{b} \sum_{z \in S_t^{(i)}} \text{Clip}\left(\nabla_\theta \ell(\theta_t, z) \ , \ C \right) + \eta_t$$

with $\eta_t \sim \mathcal{N}\left(0, \ \sigma_{\text{DP}}^2 \ I_d\right)$, where $\text{Clip}(v, \ C) = \min\left\{1, \ \frac{C}{\|v\|}\right\} \ v$

Global Phase: Receiving gradients $g_t^1, \ldots, g_t^n$ the server "robustly" aggregates them, i.e., compute

$$\widehat{g}_t := F\left(g_t^1, \ldots, g_t^n\right) \ ,$$

And updates the current parameters: $\theta_{t+1} = \theta_t - \gamma_t \ \widehat{g}_t$

Distributed Differential Privacy

Distributed Differential Privacy

(ϵ, δ)-Distributed DP

Distributed Differential Privacy

(ϵ, δ)-Distributed DP

"Is Interaction Necessary Distributed Private Learning?" A. Smith et al. IEEE S&P 2017.

Distributed Differential Privacy

(ϵ, δ)-Distributed DP

The transcript of communication between each node i and the server is $(\epsilon,\ \delta)$-DP w.r.t. the data held by node i

“Is Interaction Necessary Distributed Private Learning?” A. Smith et al. IEEE S&P 2017.

Distributed Differential Privacy

(ϵ, δ)-Distributed DP

The transcript of communication between each node i and the server is $(\epsilon,\ \delta)$-DP w.r.t. the data held by node i

"Is Interaction Necessary Distributed Private Learning?" A. Smith et al. IEEE S&P 2017.

A randomized algorithm $\mathscr{A} : \mathscr{X}^m \to \mathscr{Y}$ is $(\epsilon,\ \delta)$-DP if for any adjacent datasets $D,\ D' \in \mathscr{X}^m$ and any subset $S \subseteq \mathscr{Y}$,

Distributed Differential Privacy

(ϵ, δ)-Distributed DP

The transcript of communication between each node i and the server is $(\epsilon,\ \delta)$-DP w.r.t. the data held by node i

"Is Interaction Necessary Distributed Private Learning?" A. Smith et al. IEEE S&P 2017.

A randomized algorithm $\mathscr{A} : \mathscr{X}^m \to \mathscr{Y}$ is $(\epsilon,\ \delta)$-DP if for any adjacent datasets $D,\ D' \in \mathscr{X}^m$ and any subset $S \subseteq \mathscr{Y}$,

$$\Pr\left(\mathscr{A}(D) \in S\right) \leq e^{\epsilon} \Pr\left(\mathscr{A}(D') \in S\right) + \delta$$

Distributed Differential Privacy

(ϵ, δ)-Distributed DP

The transcript of communication between each node i and the server is $(\epsilon,\ \delta)$-DP w.r.t. the data held by node i

"Is Interaction Necessary Distributed Private Learning?" A. Smith et al. IEEE S&P 2017.

A randomized algorithm $\mathscr{A} : \mathscr{X}^m \to \mathscr{Y}$ is $(\epsilon,\ \delta)$-DP if for any adjacent datasets $D,\ D' \in \mathscr{X}^m$ and any subset $S \subseteq \mathscr{Y}$,

$$\Pr\left(\mathscr{A}(D) \in S\right) \leq e^{\epsilon} \Pr\left(\mathscr{A}(D') \in S\right) + \delta$$

"Our Data, Ourselves: Privacy via Distributed Noise Generation" C. Dwork et al. Eurocrypt 2006.

Privacy by DP-DMGD

Privacy by DP-DMGD

Consider T iterations of DP-DMGD

Privacy by DP-DMGD

Consider T iterations of DP-DMGD

By RDP composition and subsampling amplification theorems, we get

Privacy by DP-DMGD

Consider T iterations of DP-DMGD

By RDP composition and subsampling amplification theorems, we get

"Rényi Differential Privacy.” *Mironov, Ilya.* IEEE CSF,2017.

Privacy by DP-DMGD

Consider T iterations of DP-DMGD

By RDP composition and subsampling amplification theorems, we get

"Rényi Differential Privacy." *Mironov, Ilya.* IEEE CSF,2017.

Suppose $\epsilon \leq \log(1/\delta)$. There exists $k > 0$ such that, for sufficiently small batch-size b, if $\sigma_{\mathrm{DP}} \geq k \dfrac{2C}{b} \max\left\{ 1 \,,\, \dfrac{b\sqrt{T\log(1/\delta)}}{m\,\epsilon} \right\}$

then DP-DMGD satisfies $(\epsilon,\ \delta)$-Distributed DP

Privacy by DP-DMGD

Consider T iterations of DP-DMGD

By RDP composition and subsampling amplification theorems, we get

"Rényi Differential Privacy." *Mironov, Ilya.* IEEE CSF,2017.

Suppose $\epsilon \leq \log(1/\delta)$. There exists $k > 0$ such that, for sufficiently small batch-size b, if $\sigma_{\mathrm{DP}} \geq k \dfrac{2C}{b} \max \left\{ 1 \,,\, \dfrac{b\sqrt{T \log(1/\delta)}}{m\,\epsilon} \right\}$

then DP-DMGD satisfies $(\epsilon,\ \delta)$-Distributed DP

"On the Privacy-Robustness-Utility Trilemma in Distributed Learning." *Allouah, Youssef et al.* ICML, 2023.

Training Error by DP-DMGD

Training Error by DP-DMGD

Suppose we provide (ϵ, δ)-distributed DP

Training Error by DP-DMGD

Suppose we provide $(\epsilon,\ \delta)$-distributed DP

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d\sigma_{\mathrm{DP}}^2}{T}\right)$$

Training Error by DP-DMGD

Suppose we provide (ϵ, δ)-distributed DP

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d\sigma_{\mathrm{DP}}^2}{T}\right)$$

Assuming NO clipping

Training Error by DP-DMGD

Suppose we provide $(\epsilon,\ \delta)$-distributed DP

$$\mathscr{L}\ \left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}\,(\theta) \in \mathscr{O}\left(K_{\mathscr{L}} \frac{d\sigma_{\mathrm{DP}}^2}{T}\right)$$

Assuming NO clipping

Substituting $\sigma_{\mathrm{DP}} = k\,\frac{2C}{b} \max\left\{1\ ,\ \frac{b\sqrt{T\log(1/\delta)}}{m\,\epsilon}\right\}$

Training Error by DP-DMGD

Suppose we provide $(\epsilon,\ \delta)$-distributed DP

$$\mathscr{L}\ \left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}\left(\theta\right) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d\sigma_{\mathrm{DP}}^2}{T}\right)$$

Assuming NO clipping

Substituting $\sigma_{\mathrm{DP}} = k\,\frac{2C}{b} \max\left\{1\ ,\ \frac{b\sqrt{T \log(1/\delta)}}{m\,\epsilon}\right\}$

$$\mathscr{L}\ \left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}\left(\theta\right) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{m^2 \epsilon^2}\right)$$

Training Error by DP-DMGD

Suppose we provide (ϵ, δ)-distributed DP

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d\sigma_{\mathrm{DP}}^2}{T}\right)$$

Assuming NO clipping

Substituting $\sigma_{\mathrm{DP}} = k\,\frac{2C}{b} \max\left\{1\,,\, \frac{b\sqrt{T\log(1/\delta)}}{m\,\epsilon}\right\}$

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d\log(1/\delta)}{m^2\epsilon^2}\right)$$

“On the Privacy-Robustness-Utility Trilemma in Distributed Learning.” *Allouah, Youssef et al.* ICML, 2023.

Robustness with (f, κ)-Robust Averaging

Robustness with $(f,\ \kappa)$-Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

Robustness with $(f,\ \kappa)$-Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \frac{1}{|S|} \sum_{i \in S} \|g_t^i - g_t^S\|^2$$

Robustness with (f, κ)-Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \frac{1}{|S|} \sum_{i \in S} \|g_t^i - g_t^S\|^2$$

Without Distributed Polyak's Momentum

Robustness with (f, κ)-Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \frac{1}{|S|} \sum_{i \in S} \|g_t^i - g_t^S\|^2$$

Without Distributed Polyak's Momentum

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{n m^2 \epsilon^2} + \kappa T \frac{d \log(1/\delta)}{m^2 \epsilon^2} + \frac{\kappa G^2}{1 - \kappa B^2}\right)$$

Robustness with (f, κ)-Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \frac{1}{|S|} \sum_{i \in S} \|g_t^i - g_t^S\|^2$$

Without Distributed Polyak's Momentum

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{n m^2 \epsilon^2} + \kappa T \frac{d \log(1/\delta)}{m^2 \epsilon^2} + \frac{\kappa G^2}{1 - \kappa B^2}\right)$$

Robustness with (f, κ)-Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \frac{1}{|S|} \sum_{i \in S} \|g_t^i - g_t^S\|^2$$

Without Distributed Polyak's Momentum

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{n m^2 \epsilon^2} + \kappa T \frac{d \log(1/\delta)}{m^2 \epsilon^2} + \frac{\kappa G^2}{1 - \kappa B^2}\right)$$

Robustness with $(f,\, \kappa)$-Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \frac{1}{|S|} \sum_{i \in S} \|g_t^i - g_t^S\|^2$$

Without Distributed Polyak's Momentum

Grows with T !!

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{n m^2 \epsilon^2} + \kappa T \frac{d \log(1/\delta)}{m^2 \epsilon^2} + \frac{\kappa G^2}{1 - \kappa B^2}\right)$$

Robustness with (f, κ)-RA and Polyak's Momentum

Robustness with (f, κ)-RA and Polyak's Momentum

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{m^2 \epsilon^2} \left(\frac{1}{n} + \kappa \right) + \frac{\kappa G^2}{1 - \kappa B^2} \right)$$

Robustness with (f, κ)-RA and Polyak's Momentum

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{m^2 \epsilon^2} \left(\frac{1}{n} + \kappa \right) + \frac{\kappa G^2}{1 - \kappa B^2} \right)$$

Recall that we can achieve $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

Robustness with (f, κ)-RA and Polyak's Momentum

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{m^2 \epsilon^2} \left(\frac{1}{n} + \kappa \right) + \frac{\kappa G^2}{1 - \kappa B^2} \right)$$

Recall that we can achieve $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

Lower Bound:

Robustness with (f, κ)-RA and Polyak's Momentum

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{m^2 \epsilon^2} \left(\frac{1}{n} + \kappa \right) + \frac{\kappa G^2}{1 - \kappa B^2} \right)$$

Recall that we can achieve $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

Lower Bound:

$$\mathscr{L}\left(\widehat{\theta}\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \Omega\left(\frac{d \log(1/\delta)}{n m^2 \epsilon^2} + \frac{f}{n} \cdot \frac{\log(1/\delta)}{m^2 \epsilon^2} + \frac{f}{n} G^2 \right)$$

Robustness with $(f,\, \kappa)$-RA and Polyak's Momentum

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{m^2 \epsilon^2} \left(\frac{1}{n} + \kappa \right) + \frac{\kappa G^2}{1 - \kappa B^2} \right)$$

Recall that we can achieve $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

Lower Bound:

$$\mathscr{L}\left(\widehat{\theta}\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \Omega\left(\frac{d \log(1/\delta)}{n m^2 \epsilon^2} + \frac{f}{n} \cdot \frac{\log(1/\delta)}{m^2 \epsilon^2} + \frac{f}{n} G^2 \right)$$

Assuming $(G,0)$-Dissimilarity

Robustness with $(f,\, \kappa)$-RA and Polyak's Momentum

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(K_{\mathscr{L}} \frac{d \log(1/\delta)}{m^2 \epsilon^2} \left(\frac{1}{n} + \kappa \right) + \frac{\kappa G^2}{1 - \kappa B^2} \right)$$

Recall that we can achieve $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

Lower Bound:

$$\mathscr{L}\left(\widehat{\theta}\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \Omega\left(\frac{d \log(1/\delta)}{n m^2 \epsilon^2} + \frac{f}{n} \cdot \frac{\log(1/\delta)}{m^2 \epsilon^2} + \frac{f}{n} G^2 \right)$$

Assuming $(G, 0)$-Dissimilarity

"On the Privacy-Robustness-Utility Trilemma in Distributed Learning." *Allouah, Youssef et al.* ICML, 2023.

Optimality with (f, κ)-Spectral Robust Averaging

Optimality with $(f,\ \kappa)$-Spectral Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

Optimality with (f, κ)-Spectral Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \lambda_{\max}\left(\frac{1}{|H|}\sum_{i \in S}\left(g_t^i - g_t^S\right)\left(g_t^i - g_t^S\right)^{\mathrm{T}}\right)$$

Optimality with $(f,\ \kappa)$-Spectral Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \lambda_{\max}\left(\frac{1}{|H|}\sum_{i \in S}\left(g_t^i - g_t^S\right)\left(g_t^i - g_t^S\right)^{\mathrm{T}}\right)$$

Optimality with (f, κ)-Spectral Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \lambda_{\max}\left(\frac{1}{|H|}\sum_{i \in S}\left(g_t^i - g_t^S\right)\left(g_t^i - g_t^S\right)^{\mathrm{T}}\right)$$

Optimality with (f, κ)-Spectral Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \lambda_{\max}\left(\frac{1}{|H|}\sum_{i \in S}\left(g_t^i - g_t^S\right)\left(g_t^i - g_t^S\right)^{\mathrm{T}}\right)$$

Transpose

Optimality with (f, κ)-Spectral Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \lambda_{\max}\left(\frac{1}{|H|}\sum_{i \in S}\left(g_t^i - g_t^S\right)\left(g_t^i - g_t^S\right)^{\mathrm{T}}\right)$$

Transpose

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(\frac{d \log(1/\delta)}{nm^2\epsilon^2} + \kappa \cdot \frac{\log(1/\delta)}{m^2\epsilon^2} + \kappa G^2\right)$$

Optimality with (f, κ)-Spectral Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F(g_t^1, \ldots, g_t^n) - g_t^S\|^2 \leq \kappa \lambda_{\max}\left(\frac{1}{|H|}\sum_{i \in S}(g_t^i - g_t^S)(g_t^i - g_t^S)^{\mathrm{T}}\right)$$

Transpose

$$\mathscr{L}(\theta_T) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(\frac{d\log(1/\delta)}{nm^2\epsilon^2} + \kappa \cdot \frac{\log(1/\delta)}{m^2\epsilon^2} + \kappa G^2\right)$$

Matches LB if
$\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

Optimality with (f, κ)-Spectral Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \lambda_{\max}\left(\frac{1}{|H|}\sum_{i \in S}\left(g_t^i - g_t^S\right)\left(g_t^i - g_t^S\right)^{\mathrm{T}}\right)$$

Transpose

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(\frac{d\log(1/\delta)}{nm^2\epsilon^2} + \kappa \cdot \frac{\log(1/\delta)}{m^2\epsilon^2} + \kappa G^2\right)$$

Matches LB if

$\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

"On the Privacy-Robustness-Utility Trilemma in Distributed Learning." *Allouah, Youssef et al.* ICML, 2023.

Optimality with (f, κ)-Spectral Robust Averaging

Suppose that aggregation F is (f, κ)-robust averaging,

for all $S \subseteq [n]$ with $|S| = n - f$,

$$\|F\left(g_t^1, \ldots, g_t^n\right) - g_t^S\|^2 \leq \kappa \lambda_{\max}\left(\frac{1}{|H|}\sum_{i \in S}\left(g_t^i - g_t^S\right)\left(g_t^i - g_t^S\right)^{\mathrm{T}}\right)$$

Transpose

$$\mathscr{L}\left(\theta_T\right) - \min_{\theta \in \mathbb{R}^d} \mathscr{L}(\theta) \in \mathcal{O}\left(\frac{d \log(1/\delta)}{nm^2\epsilon^2} + \kappa \cdot \frac{\log(1/\delta)}{m^2\epsilon^2} + \kappa G^2\right)$$

Matches LB if
$\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

“On the Privacy-Robustness-Utility Trilemma in Distributed Learning.” *Allouah, Youssef et al.* ICML, 2023.

SMEA: $(f,\ \kappa)$-Spectral Robust Averaging

SMEA: $(f,\ \kappa)$-Spectral Robust Averaging

Smallest Maximum Eigenvalue Averaging

SMEA: (f, κ)-Spectral Robust Averaging

Smallest Maximum Eigenvalue Averaging

$$S^* \in \arg\min_{S} \lambda_{\max}\left(\frac{1}{|S|}\sum_{i\in S}\left(g_t^i - g_t^S\right)\left(g_t^i - g_t^S\right)^{\mathrm{T}}\right)$$

SMEA: (f, κ)-Spectral Robust Averaging

Smallest Maximum Eigenvalue Averaging

$$S^* \in \arg\min_{S} \lambda_{\max}\left(\frac{1}{|S|} \sum_{i \in S} \left(g_t^i - g_t^S\right) \left(g_t^i - g_t^S\right)^{\mathrm{T}} \right)$$

$$F\left(g_t^1, \ldots, g_t^n\right) \triangleq g_t^{S^*}$$

SMEA: $(f,\ \kappa)$-Spectral Robust Averaging

Smallest Maximum Eigenvalue Averaging

$$S^* \in \arg\min_{S} \lambda_{\max}\left(\frac{1}{|S|}\sum_{i\in S}\left(g_t^i - g_t^S\right)\left(g_t^i - g_t^S\right)^{\mathrm{T}}\right)$$

$$F\left(g_t^1, \ldots, g_t^n\right) \triangleq g_t^{S^*}$$

SMEA is $(f,\ \kappa)$-Spectral Robust with $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

SMEA: $(f,\ \kappa)$-Spectral Robust Averaging

Smallest Maximum Eigenvalue Averaging

$$S^* \in \arg\min_{S} \lambda_{\max}\left(\frac{1}{|S|}\sum_{i\in S}\left(g_t^i - g_t^S\right)\left(g_t^i - g_t^S\right)^{\mathrm{T}}\right)$$

$$F\left(g_t^1, \ldots, g_t^n\right) \triangleq g_t^{S^*}$$

SMEA is $(f,\ \kappa)$-Spectral Robust with $\kappa \in \mathcal{O}\left(\frac{f}{n}\right)$

"On the Privacy-Robustness-Utility Trilemma in Distributed Learning." *Allouah, Youssef et al.* ICML, 2023.

Other Interesting Results

- Su, Lili, and Nitin H. Vaidya. "Fault-Tolerant Multi-agent Optimization: Optimal Iterative Distributed Algorithms." *Proceedings of the 2016 ACM Symposium On Principles Of Distributed Computing.* 2016.
- Charikar, Moses, Jacob Steinhardt, and Gregory Valiant. "Learning from untrusted data." *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 2017.
- Karimireddy, Sai Praneeth, Lie He, and Martin Jaggi. "Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing." *International Conference on Learning Representations*. 2021.
- Farhadkhani, Sadegh, et al. "Byzantine machine learning made easy by resilient averaging of momentums." *International Conference on Machine Learning*. PMLR, 2022.

For an Overview on Robust Machine-Learning

SPRINGER NATURE

Robust Machine-Learning

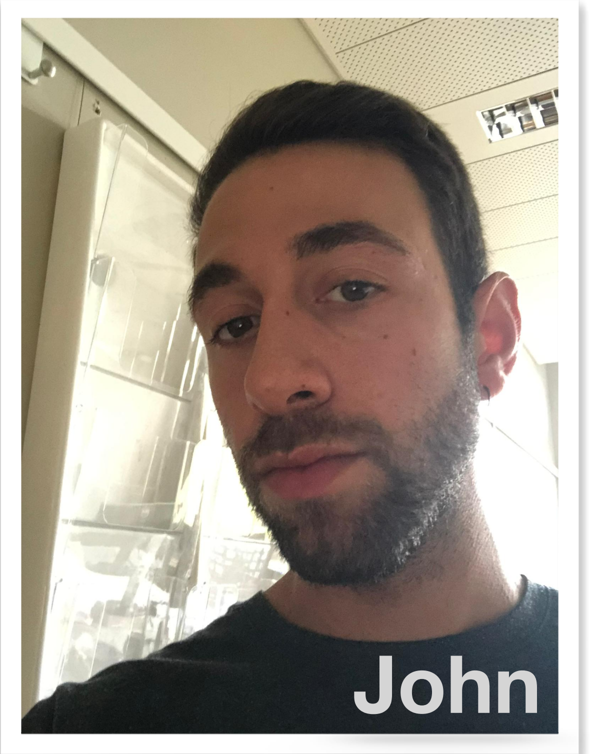
Distributed Methods for Safe AI

Byzantine Machine Learning: A Primer

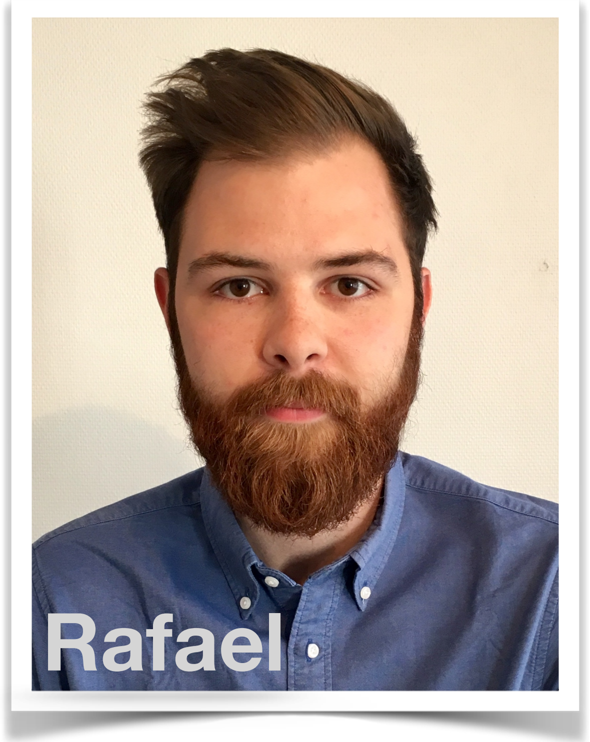
ACM Computing Surveys, 2023

Rachid Guerraoui, Nirupam Gupta & Rafael Pinot

Thanks to



Shuo Liu



EPFL

1789

GEORGETOWN UNIVERSITY


Thank you