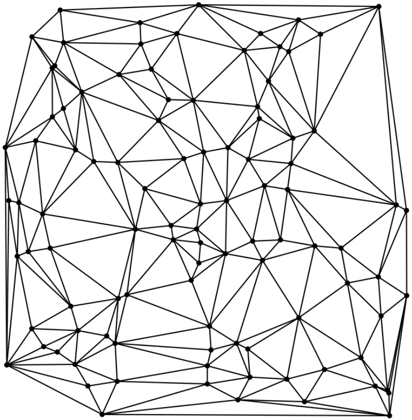
Mismatched Estimation in the Distance Geometry Problem

Mahmoud Abdelkhalek, Dror Baron, and Chau-Wai Wong

Electrical and Computer Engineering, North Carolina State University

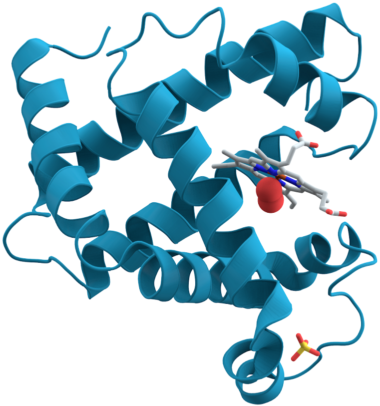
2022 Asilomar Conference on Signals, Systems, and Computers

Overview



- The Distance Geometry Problem (DGP) involves determining the locations of points in Euclidean space given measurements of the distances between these points.

Overview



- The Distance Geometry Problem (DGP) involves determining the locations of points in Euclidean space given measurements of the distances between these points.
- Some of the applications of the DGP include:
    - Computational biology

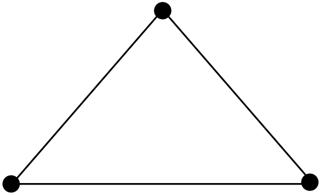
Overview

- The Distance Geometry Problem (DGP) involves determining the locations of points in Euclidean space given measurements of the distances between these points.
- Some of the applications of the DGP include:
  - Computational biology
  - Wireless networks

Overview

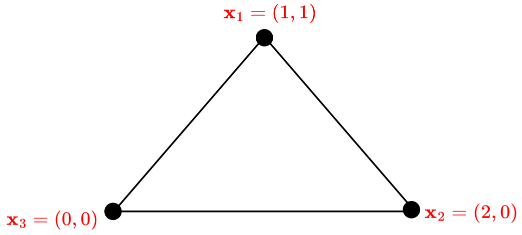
- The Distance Geometry Problem (DGP) involves determining the locations of points in Euclidean space given measurements of the distances between these points.
- Some of the applications of the DGP include:
  - Computational biology
  - Wireless networks
  - Robotics

Problem Formulation



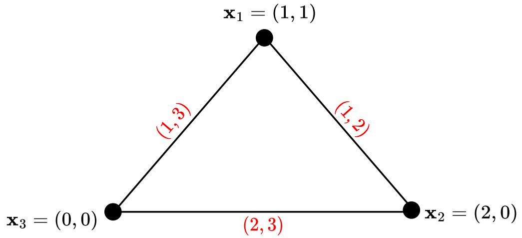
- The DGP consists of:

Problem Formulation



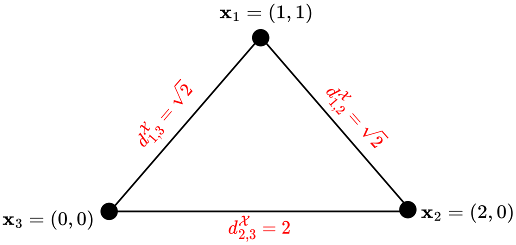
- The DGP consists of:
  - $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{K \times N}$

Problem Formulation



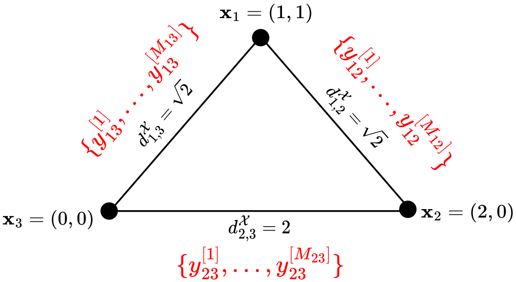
- The DGP consists of:
  - $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \in \mathbb{R}^{K \times N}$
  - $\mathcal{E}(\mathcal{X}) = \{(i,j) \mid (i,j) \in \{1,\dots,N\}^2, i < j\}$

Problem Formulation

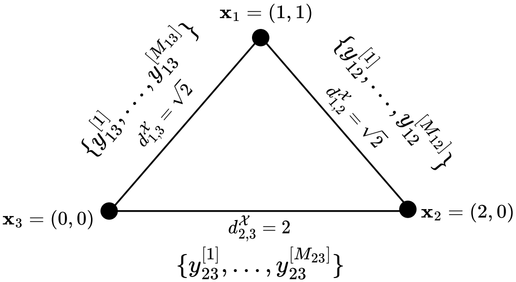


- The DGP consists of:
  - $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{K \times N}$
  - $\mathcal{E}(\mathcal{X}) = \{(i,j) \mid (i,j) \in \{1, \ldots, N\}^2, i < j\}$
  - $d^{\mathcal{X}}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$

Problem Formulation

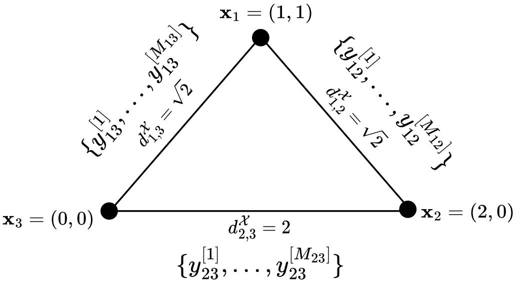


- The DGP consists of:
  - $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \in \mathbb{R}^{K \times N}$
  - $\mathcal{E}(\mathcal{X}) = \{(i,j) \mid (i,j) \in \{1, \dots, N\}^2, i < j\}$
  - $d^{\mathcal{X}}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$
  - $\left\{ \left\{ y_{ij}^{[1]}, \dots, y_{ij}^{[M_{ij}]} \right\} \;\middle|\; (i,j) \in \mathcal{E}(\mathcal{X}) \right\}$



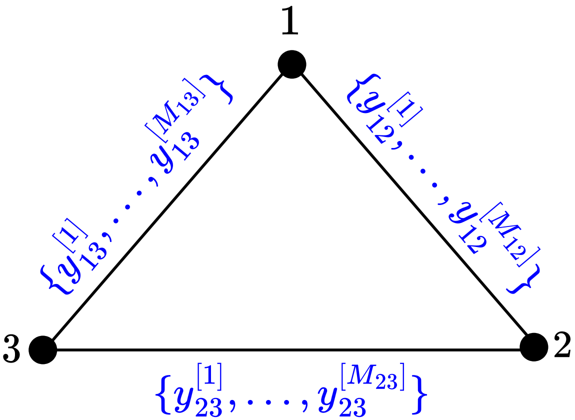
- The DGP consists of:
  - $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{K \times N}$
  - $\mathcal{E}(\mathcal{X}) = \{(i,j) \mid (i,j) \in \{1,\ldots,N\}^2, i < j\}$
  - $d_{ij}^{\mathcal{X}} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$
  - $\left\{\left\{y_{ij}^{[1]}, \ldots, y_{ij}^{[M_{ij}]}\right\} \,\middle|\, (i,j) \in \mathcal{E}(\mathcal{X})\right\}$
- The objective is to determine $\mathcal{X}$ given noisy measurements of lengths of edges in $\mathcal{E}(\mathcal{X})$.

Problem Formulation



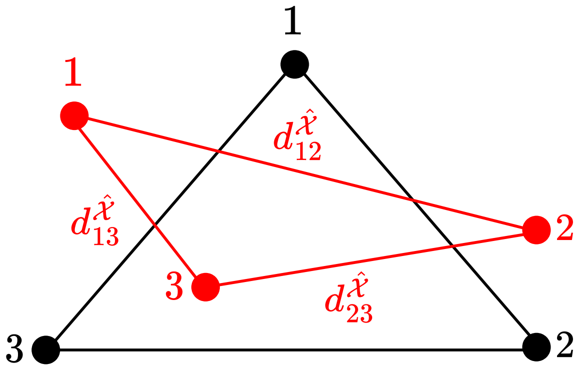
- The DGP consists of:
  - $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} \in \mathbb{R}^{K \times N}$
  - $\mathcal{E}(\mathcal{X}) = \{(i,j) \mid (i,j) \in \{1, \dots, N\}^2, i < j\}$
  - $d_{ij}^{\mathcal{X}} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$
  - $\left\{ \left\{ y_{ij}^{[1]}, \dots, y_{ij}^{[M_{ij}]} \right\} \,\middle|\, (i,j) \in \mathcal{E}(\mathcal{X}) \right\}$
- The objective is to determine $\mathcal{X}$ given noisy measurements of lengths of edges in $\mathcal{E}(\mathcal{X})$.
- Measurements are often assumed to be Gaussian.

Estimating the locations of points



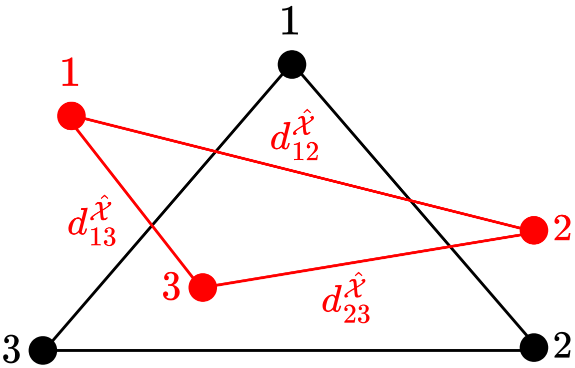
- Given noisy measurements for each edge, the common approach in the literature to compute an estimate $\widehat{\mathcal{X}}$ of $\mathcal{X}$ is to:

Estimating the locations of points



- Given noisy measurements for each edge, the common approach in the literature to compute an estimate $\widehat{\mathcal{X}}$ of $\mathcal{X}$ is to:
  - Make an initial guess and compute its edge lengths.

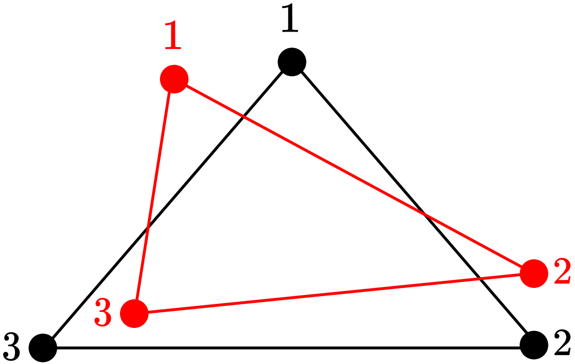
Estimating the locations of points



- Given noisy measurements for each edge, the common approach in the literature to compute an estimate $\widehat{\mathcal{X}}$ of $\mathcal{X}$ is to:
  - Make an initial guess and compute its edge lengths.
  - Compute the sum of squared errors (SSE) as

$$\sum_{(i,j)\in\mathcal{E}(\mathcal{X})} \sum_{m=1}^{M_{ij}} \left(y_{ij}^{[m]} - d_{ij}^{\hat{\mathcal{X}}} \right)^2$$

Estimating the locations of points

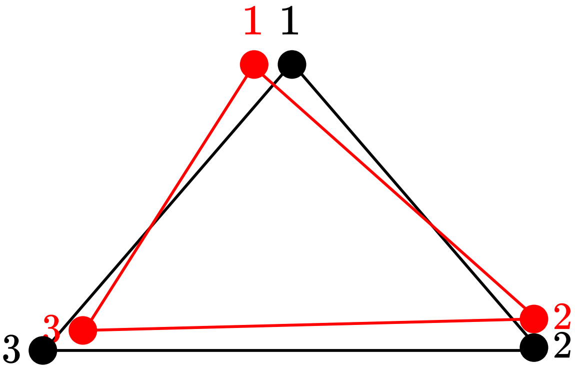


- Given noisy measurements for each edge, the common approach in the literature to compute an estimate $\widehat{\mathcal{X}}$ of $\mathcal{X}$ is to:
  - Make an initial guess and compute its edge lengths.
  - Compute the sum of squared errors (SSE) as

$$\sum_{(i,j)\in\mathcal{E}(\mathcal{X})} \sum_{m=1}^{M_{ij}} \left(y_{ij}^{[m]} - d_{ij}^{\hat{\mathcal{X}}} \right)^2$$

  - Adjust the locations of the points to minimize this SSE cost function (e.g. using a black-box optimizer).

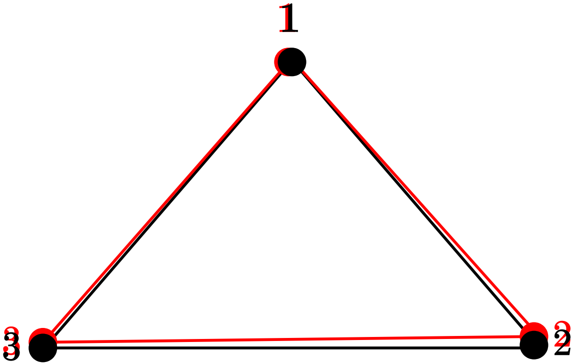
Estimating the locations of points



- Given noisy measurements for each edge, the common approach in the literature to compute an estimate $\widehat{\mathcal{X}}$ of $\mathcal{X}$ is to:
  - Make an initial guess and compute its edge lengths.
  - Compute the sum of squared errors (SSE) as

$$\sum_{(i,j)\in\mathcal{E}(\mathcal{X})} \sum_{m=1}^{M_{ij}} \left(y_{ij}^{[m]} - d_{ij}^{\hat{\mathcal{X}}} \right)^2$$

  - Adjust the locations of the points to minimize this SSE cost function (e.g. using a black-box optimizer).
  - Repeat until convergence.

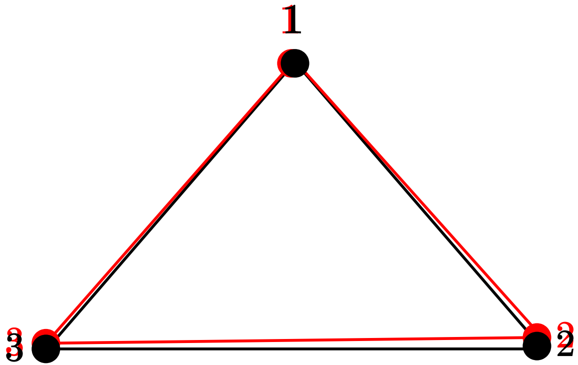


- Given noisy measurements for each edge, the common approach in the literature to compute an estimate $\widehat{\mathcal{X}}$ of $\mathcal{X}$ is to:
  - Make an initial guess and compute its edge lengths.
  - Compute the sum of squared errors (SSE) as

$$\sum_{(i,j)\in\mathcal{E}(\mathcal{X})} \sum_{m=1}^{M_{ij}} \left(y_{ij}^{[m]} - d_{ij}^{\hat{\mathcal{X}}} \right)^2$$

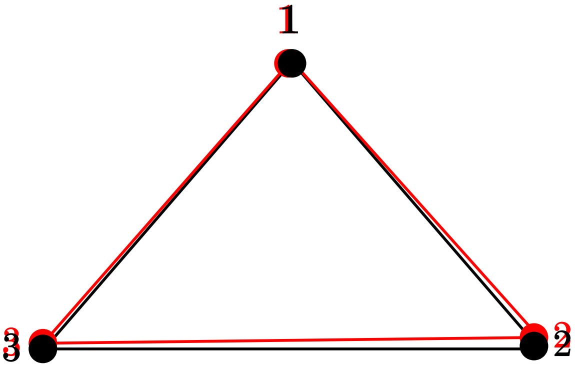
  - Adjust the locations of the points to minimize this SSE cost function (e.g. using a black-box optimizer).
  - Repeat until convergence.

Contributions



- Minimizing the SSE cost function implies Gaussian noise assumption, but what if the measurement noise isn't Gaussian?

Contributions



- Minimizing the SSE cost function implies Gaussian noise assumption, but what if the measurement noise isn't Gaussian?
- Our work: choice of cost function depends directly on distribution of noisy measurements $\implies$ approximately *half* the number of noisy edge length measurements per edge needed for the same estimation error.

Results

- An estimate $\widehat{\mathcal{X}}$ of $\mathcal{X}$ will need to be evaluated up to rotation, translation, and reflection.

Results

- An estimate $\widehat{\mathcal{X}}$ of $\mathcal{X}$ will need to be evaluated up to rotation, translation, and reflection.
- To do so, we first translate all points in $\widehat{\mathcal{X}}$ and $\mathcal{X}$ so that their centroids are aligned at the origin.

Results

- An estimate $\widehat{\mathcal{X}}$ of $\mathcal{X}$ will need to be evaluated up to rotation, translation, and reflection.
- To do so, we first translate all points in $\widehat{\mathcal{X}}$ and $\mathcal{X}$ so that their centroids are aligned at the origin.
- The *OPP loss*, which is a measure of how close an estimate $\widehat{\mathcal{X}}$ is to $\mathcal{X}$, is then the solution of the optimization problem:

$$\begin{aligned} \min_{\mathbf{R}} \quad & \left\| \mathbf{R}\widehat{\mathbf{X}}_c - \mathbf{X}_c \right\|_F, \\ \text{s.t.} \quad & \mathbf{R}^T\mathbf{R} = \mathbf{I}, \end{aligned} \tag{1}$$

Results

- An estimate $\widehat{\mathcal{X}}$ of $\mathcal{X}$ will need to be evaluated up to rotation, translation, and reflection.
- To do so, we first translate all points in $\widehat{\mathcal{X}}$ and $\mathcal{X}$ so that their centroids are aligned at the origin.
- The *OPP loss*, which is a measure of how close an estimate $\widehat{\mathcal{X}}$ is to $\mathcal{X}$, is then the solution of the optimization problem:

$$\begin{aligned} \min_{\mathbf{R}} \quad & \left\| \mathbf{R}\widehat{\mathbf{X}}_c - \mathbf{X}_c \right\|_F , \\ \text{s.t.} \quad & \mathbf{R}^T\mathbf{R} = \mathbf{I}, \end{aligned} \tag{1}$$

- For the purposes of comparison, the OPP loss is normalized by the number of points in $\mathcal{X}$.

Results

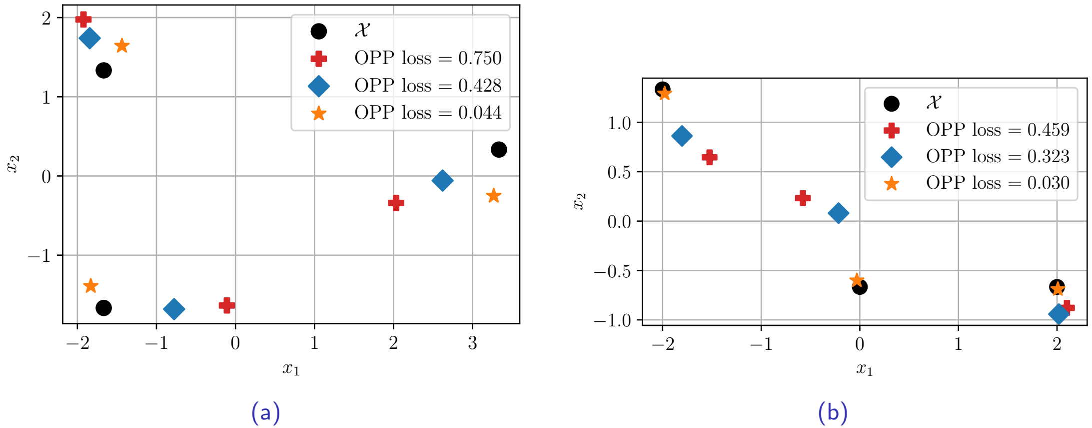


Figure: OPP loss values for example structures. Ground truth structures are labeled $\mathcal{X}$, while example estimates, together with their associated OPP losses, are also shown. We observe that estimates with lower OPP losses more closely approximate the structure $\mathcal{X}$.

Results

- Hypothesis: the cost function that corresponds to maximizing a likelihood function improves performance over using the SSE cost function.

Results

- Hypothesis: the cost function that corresponds to maximizing a likelihood function improves performance over using the SSE cost function.
- We tested this hypothesis on 8 triangles and 30 10-point structures in 2D.

Results

- Hypothesis: the cost function that corresponds to maximizing a likelihood function improves performance over using the SSE cost function.
- We tested this hypothesis on 8 triangles and 30 10-point structures in 2D.
- We let the noisy edge length measurements for these structures follow a Laplace or non-standardized Student's t (NSST) distribution.

Results

- Hypothesis: the cost function that corresponds to maximizing a likelihood function improves performance over using the SSE cost function.
- We tested this hypothesis on 8 triangles and 30 10-point structures in 2D.
- We let the noisy edge length measurements for these structures follow a Laplace or non-standardized Student's t (NSST) distribution.
- We then computed estimates for these structures using maximum likelihood (LL) estimation, where the LL function was either Gaussian (SSE) or follows the distribution of the measurements (Laplace or NSST).

Results

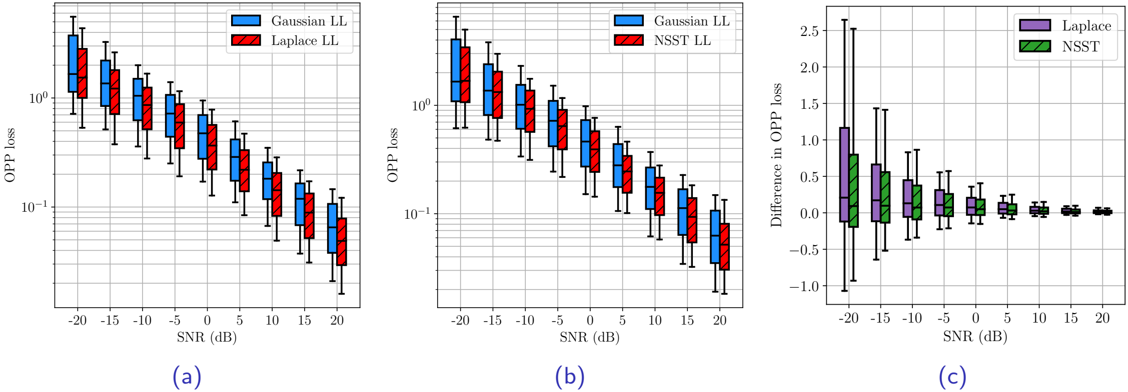


Figure: Distributions of OPP losses for 8 triangles when the noisy measurements follow a (a) Laplace or (b) NSST distribution, and when matched and mismatched (Gaussian) likelihood (LL) functions are used. Each box represents the percentiles (bottom to top): 10, 25, 50, 75, and 90.

Results

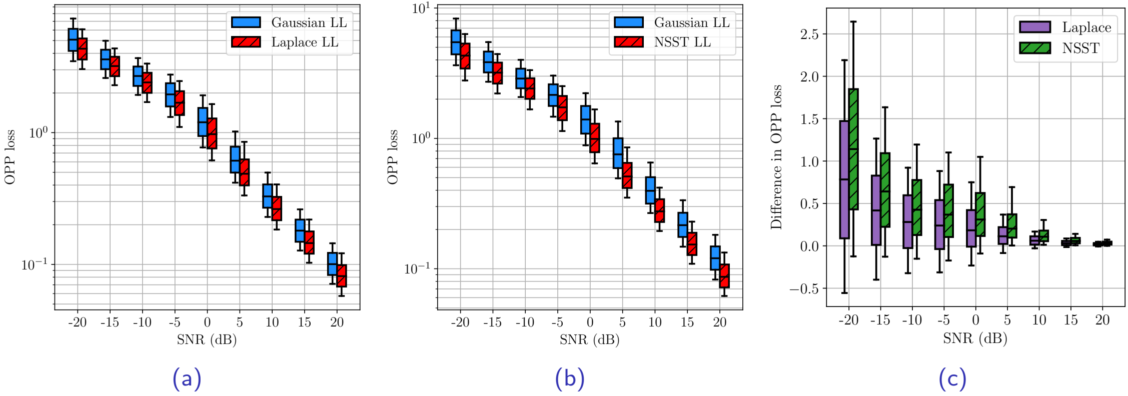


Figure: Distributions of OPP losses for the 30 10-point structures when the noisy measurements follow a (a) Laplace or (b) NSST distribution, and when matched and mismatched (Gaussian) LL functions are used.

Results

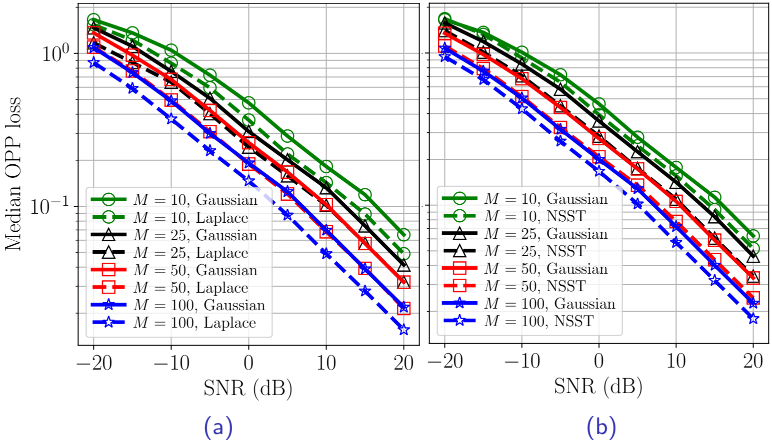


Figure: Median OPP losses for the 8 triangles and for different M values when the noisy measurements follow a (a) Laplace or (b) NSST distribution, and when matched or mismatched (Gaussian) likelihood functions are used.

Future Work

Future Work

- Dependence of measurements.

Future Work

- Dependence of measurements.
- Prior knowledge about structure (e.g. molecules).

Future Work

- Dependence of measurements.
- Prior knowledge about structure (e.g. molecules).
- Number of measurements per edge.

Future Work

- Dependence of measurements.
- Prior knowledge about structure (e.g. molecules).
- Number of measurements per edge.
- Theoretical evaluation of mismatch.