Supplementary Materials

Appendix A
On Model Uncertainties

In the signal processing [55, Chapter 1], [41] and automatic control [56, Chapter 9] communities (and also many other fields), a nominal model $\boldsymbol{O} = \mathcal{M}(\boldsymbol{I})$ is said to be uncertain if it is not guaranteed to be exactly the same as the true governing model $\boldsymbol{O} = \mathcal{M}_0(\boldsymbol{I})$, where $\boldsymbol{O}$ denotes the output and $\boldsymbol{I}$ the input. Other equivalent terms to "uncertain model" that are widely used include "mismatched model", "deviated model", and "perturbed model", etc. Possible cases are as follows.

1) **Parameter Uncertainty**. Suppose the nominal model $\boldsymbol{O} = \mathcal{M}(\boldsymbol{I};\boldsymbol{\beta})$ is parameterized by $\boldsymbol{\beta}$. If the model type is exact and only the parameter $\boldsymbol{\beta}$ is uncertain, the model uncertainty is reflected by "parameter uncertainty". In state estimation contexts, a possible example is that the true system model is guaranteed to be linear and the noises are guaranteed to be Gaussian, but we do not exactly know the system matrices and/or noise statistics.
2) **Type Uncertainty**. In state estimation contexts, an example might be the case that the true system model is nonlinear but we might use a linear nominal model. Another example might be the case that the true system model is known to be the one among candidate models. However, at one time instant, we do not exactly know which candidate model is governing the true plant [15], [36]. In this case, one may also call it "mode uncertainty".
3) **Measurement Outlier**. If outliers unexpectedly exist in measurements, the nominal measurement distribution might deviate from the true measurement distribution. In linear-system state estimation contexts, a possible example is that the nominal measurement noise model is Gaussian, whereas the true measurement noise model is fat-tailed (e.g., Laplacian, Student's t).

The list is not exhaustive, however, most common in practice.

Appendix B
Proof of Lemma 1

This lemma is a special case of [62, Theorem 1.3]. With the facts in [62, Remark 1.12], the statements in this lemma can be obtained. However, the proof of [62, Theorem 1.3] is rather complicated because it dealt with a more general problem and conducted many advanced analyses; it is not motivational for the contexts of this article. Below gives a new and concise proof because it is necessary for insights in Fig. 1.

First, by noting that $p(\boldsymbol{x}_{\mathbb{Q}}) = q(\boldsymbol{x}) = \sum_{i=1}^N q_i \delta_{\boldsymbol{x}^i}(\boldsymbol{x})$ and $\int q_i \delta_{\boldsymbol{x}^i}(\boldsymbol{x}) d\boldsymbol{x} = q_i$, we have

$$\begin{aligned}
& \inf_{\pi(\boldsymbol{x}_{\mathbb{P}},\boldsymbol{x}_{\mathbb{Q}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \pi(\boldsymbol{x}_{\mathbb{P}},\boldsymbol{x}_{\mathbb{Q}}) d\boldsymbol{x}_{\mathbb{P}} d\boldsymbol{x}_{\mathbb{Q}} \\
= & \inf_{I(\boldsymbol{x}_{\mathbb{Q}}|\boldsymbol{x}_{\mathbb{P}})} \iint \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}_{\mathbb{Q}}\| \tfrac{I(\boldsymbol{x}_{\mathbb{Q}}|\boldsymbol{x}_{\mathbb{P}}) p(\boldsymbol{x}_{\mathbb{P}})}{p(\boldsymbol{x}_{\mathbb{Q}})} p(\boldsymbol{x}_{\mathbb{Q}}) d\boldsymbol{x}_{\mathbb{P}} d\boldsymbol{x}_{\mathbb{Q}} \\
= & \inf_{I(\boldsymbol{x}^i|\boldsymbol{x}_{\mathbb{P}})} \sum_{i=1}^N \int \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}^i\| \tfrac{I(\boldsymbol{x}^i|\boldsymbol{x}_{\mathbb{P}}) p(\boldsymbol{x}_{\mathbb{P}})}{p(\boldsymbol{x}_{\mathbb{Q}})|_{\boldsymbol{x}_{\mathbb{Q}}=\boldsymbol{x}^i}} q_i d\boldsymbol{x}_{\mathbb{P}} \\
= & \inf_{I(\boldsymbol{x}^i|\boldsymbol{x}_{\mathbb{P}})} \sum_{i=1}^N \int \|\boldsymbol{x}_{\mathbb{P}} - \boldsymbol{x}^i\| I(\boldsymbol{x}^i|\boldsymbol{x}_{\mathbb{P}}) p(\boldsymbol{x}_{\mathbb{P}}) d\boldsymbol{x}_{\mathbb{P}} \\
= & \inf_{I(\boldsymbol{x}^i|\boldsymbol{x})} \sum_{i=1}^N \int \|\boldsymbol{x} - \boldsymbol{x}^i\| I(\boldsymbol{x}^i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}.
\end{aligned}$$

The first equality holds because when reformulating the Wasserstein distance, the marginals $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$ are fixed. The infimum optimization problem above has a clear physical meaning in transport theory: we aim to move all the resources (that are continuously distributed) in the whole region to some fixed facilities $\{\boldsymbol{x}^i\}_{i=1,2,\ldots,N}$. At every point $\boldsymbol{x}$, the normalized amount of resources are $p(\boldsymbol{x})$. The proportion of $p(\boldsymbol{x})$ to be moved from $\boldsymbol{x}$ to the facility $\boldsymbol{x}^i$ is $I(\boldsymbol{x}^i|\boldsymbol{x})$. The cost to move every unit of resources from $\boldsymbol{x}$ to $\boldsymbol{x}^i$ is $\|\boldsymbol{x}-\boldsymbol{x}^i\|$. Therefore, the Wasserstein distance denotes the minimum transport cost to move a distribution from one support set to another. Since $I(\boldsymbol{x}^i|\boldsymbol{x})$ are conditional distributions, implicit constraints are

$$\left\{\begin{array}{rl}
\int I(\boldsymbol{x}^i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} & = q_i, \quad \forall i \in [N], \\
\sum_{i=1}^N I(\boldsymbol{x}^i|\boldsymbol{x}) & = 1, \quad \forall \boldsymbol{x}, \\
I(\boldsymbol{x}^i|\boldsymbol{x}) & \geq 0, \quad \forall i \in [N], \forall \boldsymbol{x}.
\end{array}\right.$$

Second, we write the Lagrange dual problem

$$\begin{array}{ll}
\sup_{\lambda_i} \inf_{I(\boldsymbol{x}^i|\boldsymbol{x})} & \sum_{i=1}^N \int \|\boldsymbol{x} - \boldsymbol{x}^i\| I(\boldsymbol{x}^i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} + \\
& \sum_{i=1}^N \lambda_i \left[q_i - \int p(\boldsymbol{x}) I(\boldsymbol{x}^i|\boldsymbol{x}) d\boldsymbol{x}\right] \\
\textit{s.t.} & \sum_{i=1}^N I(\boldsymbol{x}^i|\boldsymbol{x}) = 1, \qquad \forall \boldsymbol{x}, \\
& I(\boldsymbol{x}^i|\boldsymbol{x}) \geq 0, \qquad \forall i \in [N], \forall \boldsymbol{x}.
\end{array}$$

The sup-inf objective function also writes

$$\begin{array}{ll}
\sup_{\lambda_i} \inf_{I(\boldsymbol{x}^i|\boldsymbol{x})} & \int \sum_{i=1}^N (\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i) I(\boldsymbol{x}^i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} + \\
& \sum_{i=1}^N \lambda_i q_i.
\end{array}$$

Now we recall the physical meaning of $I(\boldsymbol{x}^i|\boldsymbol{x})$ from perspective of optimal transport: it denotes the proportion of $p(\boldsymbol{x})$ to be moved to $\boldsymbol{x}^i$; i.e., $I(\boldsymbol{x}^i|\boldsymbol{x})$ are weights. As a result, we have

$$\min_i \{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\} \leq \sum_{i=1}^N (\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i) I(\boldsymbol{x}^i|\boldsymbol{x}), \quad \forall \boldsymbol{x},$$

where $I(\boldsymbol{x}^i|\boldsymbol{x}) = 1$ for the i letting the equality strictly hold, and $I(\boldsymbol{x}^i|\boldsymbol{x}) = 0$ otherwise. The above inequality holds because the weighted mean of a vector is no less than the minimum element in this vector. This gives the dual problem

$$\sup_{\lambda_i} \int \min_{i \in [N]} \{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\} p(\boldsymbol{x}) d\boldsymbol{x} + \sum_{i=1}^N \lambda_i q_i.$$

Note that the strong duality holds because the primal optimization problem is convex, and the relative interior point $p(\boldsymbol{x}_{\mathbb{Q}})$ satisfies the Slater's condition: when $p(\boldsymbol{x}_{\mathbb{P}}) := p(\boldsymbol{x}_{\mathbb{Q}})$, the optimal solution $I(\boldsymbol{x}^i|\boldsymbol{x}^i) = 1$ and $I(\boldsymbol{x}^i|\boldsymbol{x}^j) = 0, \forall j \neq i$. Since the value of $I(\boldsymbol{x}^i|\boldsymbol{x})$ is either one or zero, all $p(\boldsymbol{x})$ near $\boldsymbol{x}^i$ are moved to $\boldsymbol{x}^i$, and the cumulative at $\boldsymbol{x}^i$ is q_i (n.b., $\int I(\boldsymbol{x}^i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = q_i$). This implies a region-partition operation: the sub-region C_i is defined by such a set of $\boldsymbol{x}$ that satisfies $\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i \leq \|\boldsymbol{x} - \boldsymbol{x}^j\| - \lambda_j, \ \forall j \neq i$. In other words, $\int_{C_i} p(\boldsymbol{x}) d\boldsymbol{x} = q_i, \ \forall i \in [N]$. □

APPENDIX C
PROOF OF THEOREM 3

We first consider the case when $\theta > 0$. Let $g(\boldsymbol{x}, \boldsymbol{\lambda}) := \min_{i\in[N]} \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\right\}$. The Lagrange dual problem is

$$\begin{aligned}
&\min_{v_0\geq 0, v_1} \max_{p(\boldsymbol{x})} \int -p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} + v_0 \cdot \\
&\left\{\theta - \max_{\boldsymbol{\lambda}} \left[\int p(\boldsymbol{x}) \min_{i\in[N]} \left\{\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i\right\} d\boldsymbol{x} + \sum_{i=1}^N q_i\lambda_i\right]\right\} \\
&+ v_1 \left[1 - \int p(\boldsymbol{x}) d\boldsymbol{x}\right] \\
&= \min_{v_0\geq 0, v_1} \max_{p(\boldsymbol{x})} \min_{\boldsymbol{\lambda}} \quad v_0 \cdot \left(\theta - \sum_{i=1}^N q_i\lambda_i\right) + v_1 + \\
&\int -[\ln p(\boldsymbol{x}) + v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) + v_1] p(\boldsymbol{x}) d\boldsymbol{x}.
\end{aligned}$$

For every two bounded functions f_1 and f_2 that have the same support, $\min(f_1 + f_2) \geq \min f_1 + \min f_2$. Therefore, it is easy to verify that the objective function is convex in terms of $\boldsymbol{\lambda}$ and concave in terms of $p(\boldsymbol{x})$ by the original definitions of convexity and concavity. Since the objective function is concave and constraint-free in terms of $p(\boldsymbol{x})$, we use the varitional method to maximize it over $p(\boldsymbol{x})$. Let $\mathcal{L}[p(\boldsymbol{x})] := \int -[\ln p(\boldsymbol{x}) + v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) + v_1] p(\boldsymbol{x}) d\boldsymbol{x}$ be a functional of $p(\boldsymbol{x})$. The variation of $\mathcal{L}[p(\boldsymbol{x})]$ is

$$\begin{aligned}
\delta\mathcal{L}[p(\boldsymbol{x})] &= \left.\frac{\partial \mathcal{L}[p(\boldsymbol{x}) + \epsilon h(\boldsymbol{x})]}{\partial \epsilon}\right|_{\epsilon=0} \\
&= \int -\left[\ln p(\boldsymbol{x}) + 1 + v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) + v_1\right] h(\boldsymbol{x}) d\boldsymbol{x},
\end{aligned}$$

where $h(\boldsymbol{x}) \in L^1$ is an arbitrary function.

Let $\delta\mathcal{L}[p(\boldsymbol{x})] = 0$ and according to the fundamental lemma of calculus of variations, we have

$$[\ln p(\boldsymbol{x}) + 1 + v_0 g(\boldsymbol{x}, \boldsymbol{\lambda}) + v_1] \equiv 0,$$

almost everywhere. This gives the form of $p(\boldsymbol{x})$ in (16). Substituting $p(\boldsymbol{x})$ back into the objective of the Lagrange dual problem gives (17). The strong duality holds because (15) is concave and $\mathbb{Q}_{\boldsymbol{x}}$ is a relative interior point at which the inequality constraint in (15) is strictly satisfied (due to $\theta > 0$) and the equality constraint in (15) simultaneously holds (i.e., the Slater's conditions are met).

When $\theta = 0$, the gradient in (19) vanishes if and only if $\mathbb{P}_{\boldsymbol{x}} = \mathbb{Q}_{\boldsymbol{x}}$. Therefore, (16) and (17) also work for $\theta = 0$. In summary, this theorem works for all $\theta \geq 0$. □

APPENDIX D
PROOF OF LEMMA 2

This lemma is a special case of [62, Theorem 1.3]. One can also prove it using the standard Lagrange dual theory (cf. Appendix B). We do not give details due to necessity. □

APPENDIX E
PROOF OF THEOREM 4

The proof is straightforward by writing the Lagrange dual problem and differentiating with respect to P_{ij}. The strong duality holds: (25) is concave and $\{P^0_{ij}\}_{\forall i,\forall j}$ is assumed to be a relative interior point satisfying the Slater's conditions. In the special case when $M = N$, and $\mathbb{P}_{\boldsymbol{x}}$ and $\mathbb{Q}_{\boldsymbol{x}}$ have the same support, P^0_{ij} can be constructed as follow:

$$P^0_{ij} = \begin{cases} q_i, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

which is resulted from letting $\mathbb{P}_{\boldsymbol{x}} := \mathbb{Q}_{\boldsymbol{x}}$. In a general case when $M \neq N$ or they have different supports, to guarantee the existence of P^0_{ij}, we must let θ be strictly larger than $\min_{P_{ij}} \sum_{i=1}^N \sum_{j=1}^M \|\boldsymbol{x}^i - \boldsymbol{x}^j\| \cdot P_{ij}$ over all P_{ij} such that $\sum_{j=1}^M P_{ij} = q_i,\ \forall i \in [N]$. Note that unlike Theorem 3, we additionally require the existence of P^0_{ij}. This is because the reference distribution $\mathbb{Q}_{\boldsymbol{x}}$ in this case is no longer guaranteed to be a relative interior point that satisfies the Slater's conditions. □

APPENDIX F
PROOF OF THEOREM 6

If $\theta = 0$, the maximum entropy distribution solving (34) is $\boldsymbol{q}$ itself. Below discusses the case when $\theta > 0$. The Lagrange dual problem of (34) is

$$\begin{aligned}
\min_{\lambda_0\geq 0, \lambda_1} \max_{p_i} \quad & \textstyle\sum_{i=1}^N -p_i \ln p_i \\
& + \lambda_0 \cdot \left[\theta - \textstyle\sum_{i=1}^N p_i \ln\left(\frac{p_i}{q_i}\right)\right] \\
& + \lambda_1 \cdot \left[1 - \textstyle\sum_{i=1}^N p_i\right].
\end{aligned}$$

It is concave, smooth, and constraint-free with respect to p_i. Therefore, the optimal solution of p_i is obtained by the first-order optimality condition, i.e.,

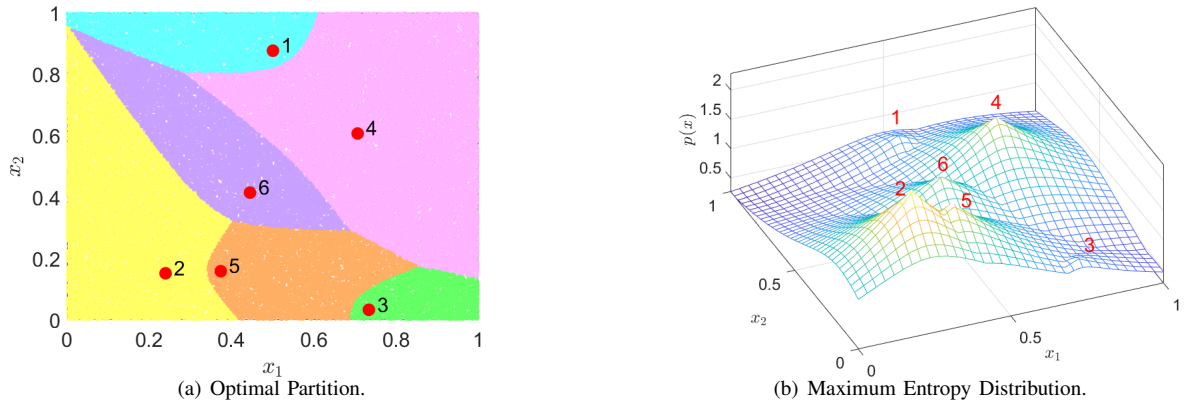
$$-(\lambda_0 + 1) \cdot [\ln(p_i) + 1] + \lambda_0 \ln(q_i) - \lambda_1 = 0.$$

This gives (35). Substituting (35) back into the objective of the Lagrange dual problem, we have (36). Since (34) is concave, and $\boldsymbol{q}$ is a relative interior point in the feasible region of (34) such that the inequality is strictly satisfied (due to $\theta > 0$) and the equality is met, the strong duality holds due to the Slater's condition. Namely, if λ_0 and λ_1 solve (36), p_i in (35) solves (34). When $\theta = 0$, the gradient (37) vanishes if and only if $\boldsymbol{p} = \boldsymbol{q}$; i.e., (35) and (36) also work for the case when $\theta = 0$. In summary, this theorem works for all $\theta \geq 0$. □

APPENDIX G
MAXIMUM ENTROPY DISTRIBUTIONS

A. Continuous Maximum Entropy Distribution Using Wasserstein Distance

We consider a two-dimensional continuous rectangular region $[0, 1]\times[0, 1]$. Let $\boldsymbol{x}$ be a 2-dimensional prior state particle: x_1 denote the horizontal axis and x_2 the vertical axis. Suppose the reference discrete prior state distribution $\boldsymbol{q}$ is supported on six points, which are randomly sampled from the rectangle.



(a) Optimal Partition. (b) Maximum Entropy Distribution.

Fig. 7. Optimal partition and maximum entropy distribution. The whole rectangular region is partitioned into six sub-regions. Red-filled circles in (a) indicate the supports of the reference distribution $\boldsymbol{q}$. Peaks in (b) correspond to the the supporting points of $\boldsymbol{q}$.

TABLE IV
THE REFERENCE DISTRIBUTION

	$\boldsymbol{x}^1$	$\boldsymbol{x}^2$	$\boldsymbol{x}^3$	$\boldsymbol{x}^4$	$\boldsymbol{x}^5$	$\boldsymbol{x}^6$
Points	0.5007	0.2397	0.7338	0.7065	0.3739	0.4450
	0.8763	0.1513	0.0323	0.6066	0.1581	0.4139
Weights	0.0583	0.2695	0.0340	0.3496	0.1453	0.1433

Their weights are also randomly determined. The points and their weights are displayed in Table IV.

We use Theorem 3 and its corresponding projected gradient descent method to find the continuous maximum entropy distribution. The uncertainty budget θ is set to $\theta := 0.025$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. The results are shown in Fig. 7. The Monte Carlo integration method is used to evaluate integrals in (19), (20), and (21); for every Monte Carlo sample $\boldsymbol{x}$, it belongs to C_i if

$$\|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i \leq \|\boldsymbol{x} - \boldsymbol{x}^j\| - \lambda_j, \ \forall j \neq i.$$

B. Discrete Maximum Entropy Distribution Using Kullback-Leibler Divergence

The reference distribution $\boldsymbol{q}$ and the induced maximum entropy distribution $\boldsymbol{p}$ are displayed in Table V and Fig. 8. $\boldsymbol{p}$ is calculated by Theorem 6. Since they have the same support set, we do not explicitly demonstrate what the particles $\boldsymbol{x}^i$ are. The uncertainty budget θ is set to $\theta := 0.0075$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. From Table V and Fig. 8, we can see that $\boldsymbol{p}$ are more balanced than $\boldsymbol{q}$: the minimum of $\boldsymbol{p}$ is larger than that of $\boldsymbol{q}$ (when $i = 4$), while the maximum of $\boldsymbol{p}$ is smaller than that of $\boldsymbol{q}$ (when $i = 2$).

C. Discrete Maximum Entropy Distribution Using Wasserstein Distance

We let the reference discrete distribution $\boldsymbol{q}$ explicitly be a likelihood distribution of one (worst-case) prior state particle

TABLE V
THE REFERENCE DISTRIBUTION AND ITS INDUCED MAXIMUM ENTROPY DISTRIBUTION (USING KULLBACK-LEIBLER DIVERGENCE)

	$\boldsymbol{x}^1$	$\boldsymbol{x}^2$	$\boldsymbol{x}^3$	$\boldsymbol{x}^4$	$\boldsymbol{x}^5$	$\boldsymbol{x}^6$
$\boldsymbol{q}$	0.1993	0.2907	0.0974	0.0492	0.1505	0.2128
$\boldsymbol{p}$	0.1934	0.2492	0.1196	0.0756	0.1602	0.2021

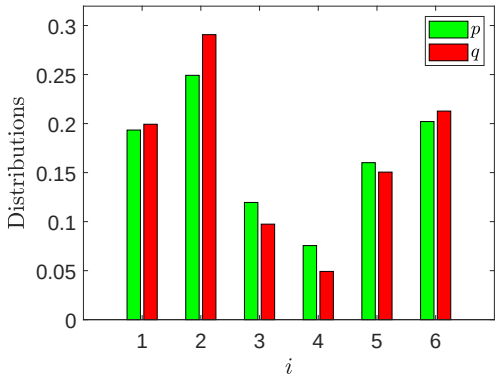


Fig. 8. The maximum entropy distribution $\boldsymbol{p}$ (left bar at each i) induced by the reference distribution $\boldsymbol{q}$ (right bar at each i) using the Kullback-Leibler Divergence.

$\boldsymbol{x}$. Suppose $\boldsymbol{q}$ and its induced maximum entropy distribution $\boldsymbol{p}$ have different support sets, as displayed in Fig. 9. The support set $\{\boldsymbol{y}^r|\boldsymbol{x}\}_{r\in[R]}$ of $\boldsymbol{q}$ consists of particles propagated from a 2-dimensional nonlinear measurement equation

$$\begin{cases} y_1^r &= |\sin(x_1 + x_2 + v_1^r)|, \\ y_2^r &= |\cos(e^{x_1 \times x_2 + v_2^r})|, \end{cases} \qquad \forall r \in [4]$$

where $\boldsymbol{x} := [x_1, x_2]^T := [0, 0]^T$ is the fixed prior state particle, and measurement noises v_1^r and v_2^r are sampled from a uniform distribution $\mathcal{U}[0, 1]$. The support set of $\boldsymbol{p}$, however, is constructed by five uniformly sampled points (i.e., green-filled squares No. $1 \sim 5$) and a new measurement (i.e., green-filled square No. 6). Randomly setting the reference distribution

$$\boldsymbol{q} := [0.3700,\ 0.3194,\ 0.0610,\ 0.2496]^T,$$

then the induced maximum entropy distribution $\boldsymbol{p}$ is given as

$$\boldsymbol{p} = [0.2641,\ 0.1272,\ 0.3440,\ 0.2513,\ 0.0071,\ 0.0064]^T,$$

where $\boldsymbol{p}$ is obtained by Theorem 5. The uncertainty budget θ is set to $\theta := 0.325$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. As expected, although the support sets are different, we can still calculate the weights of new supporting points of $\boldsymbol{p}$, and the worst-case likelihood of the new measurement is evaluated as 0.0064. This small-valued likelihood result coincides with our intuition because the new point No. 6 is far away from the supports (i.e., red-filled circles) of $\boldsymbol{q}$.

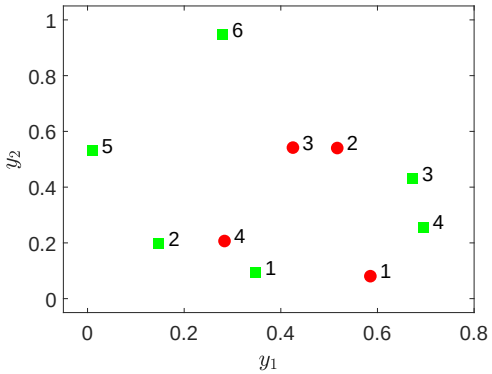


Fig. 9. The maximum entropy distribution $\boldsymbol{p}$ induced by the reference distribution $\boldsymbol{q}$ using the Wasserstein distance. Red-filled circles are supports of $\boldsymbol{q}$, while green-filed squares are supports of $\boldsymbol{p}$.

Alternatively, we may suppose the support set of $\boldsymbol{p}$ is constructed by the union of the support set of $\boldsymbol{q}$ and the new measurement. The supporting points of $\boldsymbol{q}$ are uniformly sampled from $[0,1] \times [0,1]$. We have the results in Table VI, in which $\boldsymbol{y}^5$ is a new measurement uniformly sampled from $[0,1] \times [0,1]$ as well. The uncertainty budget θ is set to $\theta := 0.01$ (only for a possible demonstration; other values also applicable). In the projected gradient descent procedure, the step size $\alpha := 0.05$ and the maximum allowed iteration steps $S := 500$. From Table VI, it can be seen that the likelihood (of the associated worst-case prior state particle) at this new measurement is 0.0260.

TABLE VI
THE REFERENCE DISTRIBUTION AND ITS INDUCED MAXIMUM ENTROPY DISTRIBUTION (USING WASSERSTEIN DISTANCE)

	$\boldsymbol{y}^1$	$\boldsymbol{y}^2$	$\boldsymbol{y}^3$	$\boldsymbol{y}^4$	$\boldsymbol{y}^5$
Points	0.4314	0.6146	0.0059	0.5459	**0.6206**
	0.5779	0.2699	0.8958	0.1993	**0.3924**
Weights ($\boldsymbol{q}$)	0.3438	0.1316	0.3191	0.2055	/
Weights ($\boldsymbol{p}$)	0.3372	0.1327	0.3191	0.1850	**0.0260**