A. Proofs

Our main results in this section make the following assumptions.

(A1) the predictor f is unconstrained.

(A2) both the loss and deviation are squared errors.

(A3) $|\mathcal{B}(x_i)| = m, \forall x_i \in \mathcal{D}_x$.

(A4) $x_j \in \mathcal{B}(x_i) \implies x_i \in \mathcal{B}(x_j), \forall x_i, x_j \in \mathcal{D}_x$.

(A5) $\cup_{x_i \in \mathcal{D}_x} \mathcal{B}(x_i) = \mathcal{D}_x$.

We note that **(A3)** and **(A4)** are not technically necessary but simplify the presentation. We denote the predictor in the uniform criterion (Eq. (2)), the symmetric game (Eq. (3)), and the asymmetric (Eq. (4)) game as f_U, f_S, and f_A, respectively. We use $X_i \in \mathbb{R}^{m \times d}$ to denote the neighborhood $\mathcal{B}(x_i) = \{x'_1, \ldots, x'_m\}$ ($X_i = [x'_1, \ldots, x'_m]^\top$), and $f(X_i) \in \mathbb{R}^m$ to denote the vector $[f(x'_1), \ldots, f(x'_m)]^\top$. $X_j^\dagger$ denotes the pseudo-inverse of X_j. Then we have

Theorem 2. *If* **(A1-5)** *hold and the witness is in the linear family, the optimal f_S satisfies*

$$f_S^*(x_i) = \frac{1}{1+\lambda}\left[y_i + \frac{\lambda}{m}\Big(\sum_{x_j \in \mathcal{B}(x_i)} X_j^\dagger f_S^*(X_j)\Big)^\top x_i\right],$$

and the optimal f_A, at every equilibrium, is the fixed point

$$f_A^*(x_i) = \frac{1}{1+\lambda}\left[y_i + \lambda (X_i^\dagger f_A^*(X_i))^\top x_i\right], \forall x_i \in \mathcal{D}_x.$$

Proof. We first re-write the symmetric criterion explicitly as a game:

$$\min_f \sum_i (f(x_i) - y_i)^2 + \frac{\lambda}{m} \sum_{x_j \in \mathcal{B}(x_i)} (f(x_j) - \hat{g}_{x_i}(x_j))^2,$$

where $\hat{g}_{x_i}$ is the best response strategy from the local witness.

Since f is unconstrained and the objective in convex in it, we can treat each $f(x_i)$ as a distinct variable, and use the derivative to find its optimum:

$$\begin{aligned} f_S^*(x_i) &= \frac{1}{1+\lambda}\left[y_i + \frac{\lambda}{m} \sum_{x_j \in \mathcal{B}^{-1}(x_i)} \hat{g}_{x_j}(x_i)\right] \\ &= \frac{1}{1+\lambda}\left[y_i + \frac{\lambda}{m} \sum_{x_j \in \mathcal{B}(x_i)} \hat{g}_{x_j}(x_i)\right], \end{aligned} \tag{7}$$

where $\mathcal{B}^{-1}(x_i) = \{x_j \in \mathcal{D}_x : x_i \in \mathcal{B}(x_j)\}$. Note that we only have to collect witnesses $\hat{g}_{x_j}$ that are relevant to $f(x_i)$ for the first equality, and the second equality is due to **(A4)**. On the other hand, the objective for f in the asymmetric game is:

$$\min_f \sum_i (f(x_i) - y_i)^2 + \lambda (f(x_i) - \hat{g}_{x_i}(x_i))^2,$$

The corresponding optimum is:

$$f_A^*(x_i) = \frac{1}{1+\lambda}\left[y_i + \lambda \hat{g}_{x_i}(x_i)\right] \tag{8}$$

For both games, the objective for g_{x_i} can be described as:

$$\begin{aligned} &\min_{g_{x_i}} \frac{\lambda}{m} \sum_{x_j \in \mathcal{B}(x_i)} (f(x_j) - g_{x_i}(x_j))^2 \\ &= \min_{\theta_i} \frac{\lambda}{m} \|f(X_i) - X_i \theta_i\|_2^2, \end{aligned} \tag{9}$$

Then Eq. (10) is an optimal witness $g_{x_i}^*$ at x_i.

$$g_{x_i}^*(x_j) = \theta_i^\top x_j = (X_i^\dagger f(X_i))^\top x_j, \forall x_j \in \mathcal{X}, \tag{10}$$

and we note that every optimal witness $g_{x_i}^*$ has the same values on $\mathcal{B}(x_i)$

Since the optimal $g_{x_i}^*$ is functionally dependent to f. we put Eq. (10) back to Eq. (7) to obtain the optimal condition for f_S^* (at equilibrium) as

$$f_S^*(x_i) = \frac{1}{1+\lambda}\left[y_i + \frac{\lambda}{m}\Big(\sum_{x_j \in \mathcal{B}(x_i)} X_j^\dagger f_S^*(X_j)\Big)^\top x_i\right].$$

Again, putting Eq. (10) back to Eq. (8), we obtain the optimal condition for f_A^* at equilibrium as

$$f_A^*(x_i) = \frac{1}{1+\lambda}\left[y_i + \lambda (X_i^\dagger f_A^*(X_i))^\top x_i\right].$$

□

Note that the equilibrium for the linear class is not unique when the solution of Eq. (9) is not unique: there may be infinitely many optimal solution to the witness in a neighborhood due to degeneracy. In this case, Theorem 2 adopts the minimum norm solution as used in the pseudo-inverse in Eq. (10). In this case, one may use Ridge regression instead to establish a strongly convex objective for the witness to ensure a unique solution, where the objective for the witness is rewritten as

$$\min_{\theta_i} \frac{\lambda}{m} \|f(X_i) - X_i \theta_i\|_2^2 + \alpha \|\theta_i\|_2^2, \tag{11}$$

with a positive α.

Theorem 3. *If* **(A1-5)** *hold and the witness is in the linear family, the optimal f_U satisfies*

$$f_U^*(x_i) = \begin{cases} \alpha(x_i, f_U^*), & \text{if } \alpha(x_i, f_U^*) > y_i, \\ \beta(x_i, f_U^*), & \text{if } \beta(x_i, f_U^*) < y_i, \\ y_i, & \text{otherwise,} \end{cases}$$

for $x_i \in \mathcal{D}_x$, where

$$\alpha(x_i, f_U^*) = \max_{x_j \in \mathcal{B}(x_i)} \Bigg[(X_j^\dagger f_U^*(X_j))^\top x_i - \sqrt{\delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f_U^*(x_k) - (X_j^\dagger f_U^*(X_j))^\top x_k)^2} \Bigg];$$

$$\beta(x_i, f_U^*) = \min_{x_j \in \mathcal{B}(x_i)} \Bigg[(X_j^\dagger f_U^*(X_j))^\top x_i + \sqrt{\delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f_U^*(x_k) - (X_j^\dagger f_U^*(X_j))^\top x_k)^2} \Bigg].$$

Proof. The objective for the uniform criterion is:

$$\min_f \sum_{i=1}^N (f(x_i) - y_i)^2 \tag{12}$$
$$s.t. \ \min_{g \in \mathcal{G}} \frac{1}{m} \sum_{x_j \in \mathcal{B}(x_i)} (f(x_j) - g(x_j))^2 \le \delta, \forall x_i \in \mathcal{D}_x.$$

Our strategy is to temporarily treat each g as a fixed function, and then replace it with its best response strategy.

Since f is unconstrained (in capacity), we can treat each $f(x_i)$ as a distinct variable for optimization. For each $f(x_i)$, we first filter its relevant criteria:

$$\min_{f(x_i)} (f(x_i) - y_i)^2$$
$$s.t. \ (f(x_i) - g_{x_j}(x_i))^2, \le \delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f(x_k) - g_{x_j}(x_k))^2, \forall x_j \in \mathcal{B}(x_i).$$

For any feasible f, we can further rewrite the constraint of $f(x_i)$ with respect to each x_j as:

$$g_{x_j}(x_i) - \sqrt{\delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f(x_k) - g_{x_j}(x_k))^2} \le f(x_i) \le g_{x_j}(x_i) + \sqrt{\delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f(x_k) - g_{x_j}(x_k))^2}.$$

Collectively, we can fold all the upper bounds of $f(x_i)$ as

$$f(x_i) \le \min_{x_j \in \mathcal{B}(x_i)} \Bigg[g_{x_j}(x_i) + \sqrt{\delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f(x_k) - g_{x_j}(x_k))^2} \Bigg].$$

All the lower bounds can be folded similarly.

Finally, since the objective for $f(x_i)$ is simply a squared error with an interval constraint, evidently if y_i satisfies the lower bounds and upper bounds, then $f_U^*(x_i) = y_i$. If

$$y_i > \min_{x_j \in \mathcal{B}(x_i)} \Bigg[g_{x_j}(x_i) + \sqrt{\delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f(x_k) - g_{x_j}(x_k))^2} \Bigg],$$

then we have

$$f_U^*(x_i) = \min_{x_j \in \mathcal{B}(x_i)} \Bigg[g_{x_j}(x_i) + \sqrt{\delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f(x_k) - g_{x_j}(x_k))^2} \Bigg].$$

Otherwise, we have

$$f_U^*(x_i) = \max_{x_j \in \mathcal{B}(x_i)} \Bigg[g_{x_j}(x_i) - \sqrt{\delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f(x_k) - g_{x_j}(x_k))^2} \Bigg].$$

For each g_{x_i} is in the linear class, Eq. (13) is an optimal solution.

$$g_{x_j}^*(x_i) = (X_j^\dagger f(X_j))^\top x_i, \forall x_i \in \mathcal{X}, \tag{13}$$

and we note that every optimal witness $g_{x_j}^*$ has the same values on $\mathcal{B}(x_j)$.

Since the optimal $g_{x_i}^*$ is functionally dependent to f, to obtain the optimal f_U^*, we combine our previous result with $g_{x_i}^*$ such that the optimality conditions for f and g_{x_i} are both satisfied. Finally, we have

$$f_U^*(x_i) = \begin{cases} \alpha(x_i, f_U^*), & \text{if } \alpha(x_i, f_U^*) > y_i, \\ \beta(x_i, f_U^*), & \text{if } \beta(x_i, f_U^*) < y_i, \\ y_i, & \text{otherwise,} \end{cases}$$

for $x_i \in \mathcal{D}_x$, where

$$\alpha(x_i, f_U^*) = \max_{x_j \in \mathcal{B}(x_i)} \Bigg[(X_j^\dagger f_U^*(X_j))^\top x_i - \sqrt{\delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f_U^*(x_k) - (X_j^\dagger f_U^*(X_j))^\top x_k)^2} \Bigg];$$

$$\beta(x_i, f_U^*) = \min_{x_j \in \mathcal{B}(x_i)} \Bigg[(X_j^\dagger f_U^*(X_j))^\top x_i + \sqrt{\delta m - \sum_{x_k \in \mathcal{B}(x_j)\setminus\{x_i\}} (f_U^*(x_k) - (X_j^\dagger f_U^*(X_j))^\top x_k)^2} \Bigg].$$

□

Lemma 4. *If $d(\cdot,\cdot)$ is squared error, $\mathcal{L}(\cdot,\cdot)$ is differentiable, f is sub-differentiable, and* **A(4-5)** *hold, then*

$$\sum_{(x_i,y_i)\in\mathcal{D}} \mathcal{L}(f(x_i),y_i) + \frac{\lambda}{\bar{N}_i}\left[\bar{N}_i f(x_i) - \sum_{x_t\in\mathcal{B}(x_i)} \frac{\hat{g}_{x_t}(x_i)}{|\mathcal{B}(x_t)|}\right]^2, \tag{14}$$

where $\bar{N}_i := \sum_{x_t\in\mathcal{B}(x_i)} \frac{1}{|\mathcal{B}(x_t)|}$, induces the same equilibrium as the symmetric game.

Proof. Since the criteria for the witness g_{x_i} are the same in the symmetric game and the proposed asymmetric criterion here, we only have to check for the optimality condition for the predictor f. Here we use $\nabla_\theta f(x)$ to denote the subgradient of f at x with respect to the underlying parameter θ, the optimality condition for Eq. (14) is

$$\begin{aligned}
0 \in & \sum_{(x_i,y_i)\in\mathcal{D}} \Bigg[\frac{\partial}{\partial f(x_i)}\mathcal{L}(f(x_i),y_i) \\
& + 2\lambda\Big(\sum_{x_t\in\mathcal{B}(x_i)} \frac{f(x_i)}{|\mathcal{B}(x_t)|} - \sum_{x_t\in\mathcal{B}(x_i)} \frac{\hat{g}_{x_t}(x_i)}{|\mathcal{B}(x_t)|}\Big)\Bigg]\nabla_\theta f(x_i) \\
= & \sum_{(x_i,y_i)\in\mathcal{D}} \Bigg[\frac{\partial}{\partial f(x_i)}\mathcal{L}(f(x_i),y_i)\nabla_\theta f(x_i) \\
& + \sum_{x_t\in\mathcal{B}(x_i)} \frac{2\lambda}{|\mathcal{B}(x_t)|}(f(x_i) - \hat{g}_{x_t}(x_i))\nabla_\theta f(x_i)\Bigg]
\end{aligned}$$

For the symmetric game, the optimality condition is

$$\begin{aligned}
0 \in & \sum_{(x_i,y_i)\in\mathcal{D}} \Bigg[\frac{\partial}{\partial f(x_i)}\mathcal{L}(f(x_i),y_i)\nabla_\theta f(x_i) \\
& + \sum_{x_t\in\mathcal{B}(x_i)} \frac{2\lambda}{|\mathcal{B}(x_i)|}(f(x_t) - \hat{g}_{x_i}(x_t))\nabla_\theta f(x_t)\Bigg]
\end{aligned}$$

It is evident that the two conditions coincide if Eq. (15) is equal to Eq. (16).

$$\sum_{(x_i,y_i)\in\mathcal{D}} \sum_{x_t\in\mathcal{B}(x_i)} \frac{1}{|\mathcal{B}(x_i)|}(f(x_t) - \hat{g}_{x_i}(x_t))\nabla_\theta f(x_t) \tag{15}$$

$$\begin{aligned}
= & \sum_{x_t\in\cup_{x_i\in\mathcal{D}_x}\mathcal{B}(x_i)} \sum_{x_i\in\mathcal{B}^{-1}(x_t)} \\
& \qquad \frac{1}{|\mathcal{B}(x_i)|}(f(x_t) - \hat{g}_{x_i}(x_t))\nabla_\theta f(x_t) \\
= & \sum_{x_t\in\mathcal{D}_x} \sum_{x_i\in\mathcal{B}(x_t)} \frac{1}{|\mathcal{B}(x_i)|}(f(x_t) - \hat{g}_{x_i}(x_t))\nabla_\theta f(x_t) \\
= & \sum_{(x_i,y_i)\in\mathcal{D}} \sum_{x_t\in\mathcal{B}(x_i)} \frac{1}{|\mathcal{B}(x_t)|}(f(x_i) - \hat{g}_{x_t}(x_i))\nabla_\theta f(x_i),
\end{aligned} \tag{16}$$

where the first equality is simply re-ordering of the two summations, and the second equality is due to $x_t \in \mathcal{B}(x_i) \iff x_i \in \mathcal{B}(x_t)$ and $\cup_{x_i\in\mathcal{D}_x}\mathcal{B}(x_i) = \mathcal{D}_x$. □

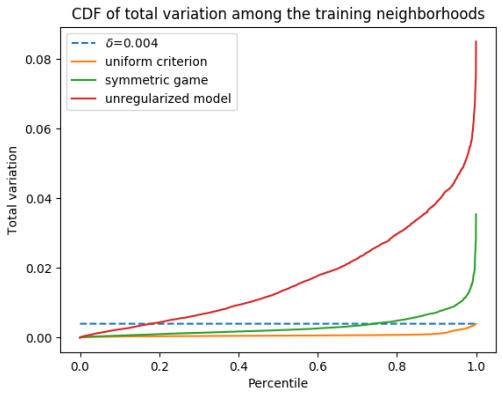

Figure 5. The cumulative distribution function of the total variation loss between the predictor f and the local witness g in each training neighborhood.

B. Supplementary Materials for Molecule Property Prediction

Implementation. To conduct training, we use GCNs as the predictor with 6 layers of graph convolution with 1800 hidden dimension. We use a $80\%/10\%/10\%$ split for training / validation / testing.

Evaluation Measures. We use the `roc_auc_score` in `scikit-learn` (Pedregosa et al., 2011) to compute the AUC score. Note that for each criterion, we evaluate the model with respect to each label, and then report the average score across the 12 labels. Here N denotes the number of testing data.

- AUC(f, y): we compare $f(\mathcal{M}_i)$ with the labels y_i among the testing data $\{(\mathcal{M}_i, y_i)\}_{i=1}^N$ in AUC.
- AUC$(\hat{g}_\mathcal{M}, y)$: we compare $\hat{g}_{\mathcal{M}_i}(x(\mathcal{M}_i))$ with the labels y_i among the testing data $\{(\mathcal{M}_i, y_i)\}_{i=1}^N$ in AUC.
- AUC$_\mathcal{B}(\hat{g}_\mathcal{M}, f)$: for each testing data $(\mathcal{M}, y)$, we evaluate the following score among the neighborhood $\mathcal{B}(\mathcal{M}) = \{\mathcal{M}_1, \ldots, \mathcal{M}_{N_\mathcal{M}}\}$, where $N_\mathcal{M} := |\mathcal{B}(\mathcal{M})|$, around $\mathcal{M}$:

  $$\frac{\sum_{i=1}^{N_\mathcal{M}}\sum_{j=1}^{N_\mathcal{M}} \mathbb{I}(f(\mathcal{M}_i) > f(\mathcal{M}_j))\mathbb{I}(\hat{g}_\mathcal{M}(\mathcal{M}_i) > \hat{g}_\mathcal{M}(\mathcal{M}_j))}{\sum_{i=1}^{N_\mathcal{M}}\sum_{j=1}^{N_\mathcal{M}} \mathbb{I}(f(\mathcal{M}_i) > f(\mathcal{M}_j))}.$$

  The average score across all the testing neighborhood is then reported.
- AUC$_\mathcal{D}(\hat{g}_\mathcal{M}, f)$: we evaluate the following score among the testing data $\{(\mathcal{M}_i, y_i)\}_{i=1}^N$:

  $$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{I}(f(\mathcal{M}_i) > f(\mathcal{M}_j))\mathbb{I}(\hat{g}_{\mathcal{M}_i}(\mathcal{M}_i) > \hat{g}_{\mathcal{M}_j}(\mathcal{M}_j))}{\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{I}(f(\mathcal{M}_i) > f(\mathcal{M}_j))}.$$

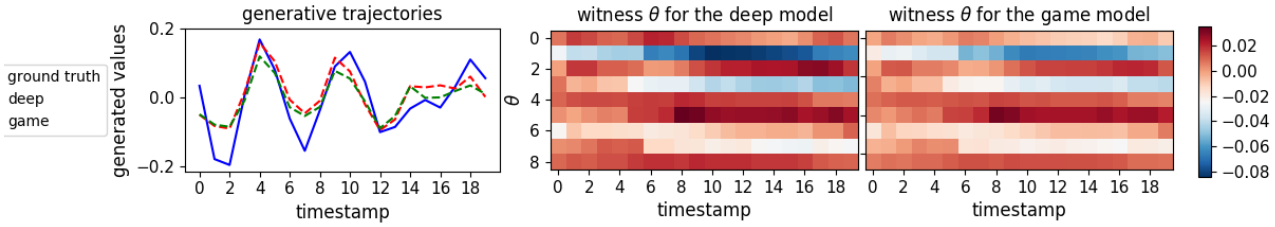

Figure 6. Visualization of the witnesses with the their parameters (middle and right plots) for *teacher-forced* predictions on the first channel (left plot) along each timestamp (x-axis) on the bearing dataset. The y-axis of the parameters from 0 to 8 denotes the bias $(\theta_0)_1$ and weights $(\theta_1)_{1,1:4}, (\theta_2)_{1,1:4}$.

Visualization. To investigate the behavior of the models, we plot their total variation loss from the local witness among the training neighborhoods in Figure 5. The uniform criterion imposes a strict functional constraint, while the symmetric game allows a more flexible model, exhibiting a tiny fraction of high deviation among the training neighborhoods.

C. Supplementary Materials for Physical Component Modeling

Implementation. We randomly sample 85%, 5%, and 10% of the data for training, validation, and testing. We set the learning rate as 10^{-5} with the Adam optimizer (Kingma & Ba, 2015). The batch size is set to 128. All the hidden dimensions are set to 128. We use the `MultivariateNormalTriL` function in Tensorflow (Abadi et al., 2016) to parametrize the multivariate Gaussian distribution. Specifically, we let the network output a $N + \frac{(N+1)(N)}{2}$ dimensional vector. The first N dimensions are treated as the mean. The second part is transformed to a lower triangular matrix, where the diagonal is further processed with a softplus nonlinearity. Such representation satisfies the Cholesky decomposition for covariance matrix.

For fitting the linear witness, we use Ridge regression in `scikit-learn` (Pedregosa et al., 2011) with the default hyperparameter. The usage of Ridge regression instead of vanilla linear regression is justified by our analysis of the equilibrium for linear witnesses.

Visualization. The visualization for the teacher-forced generative trajectory is in Figure 6.

Neighborhood size analysis

Here we investigate the effect of neighborhood radius ϵ. The results are shown in Figure 7. The impact of the neighborhood size is quite monotonic to deviation and TV, but in a reverse way. As ϵ increases, the weight of the witness on fitting the current point x_i among the neighborhood

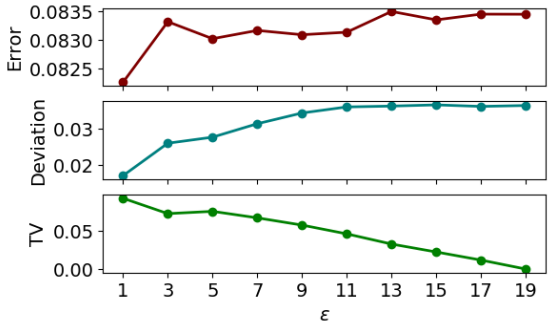

Figure 7. Parameter analysis of ϵ on the GAME model with $\lambda = 1$.

$\mathcal{B}(x_i)$ decreases, so the deviation of the witness $\hat{g}_{x_i}(x_i)$ from $f(x_i)$ increases. In contrast, as more points are overlapped between the neighborhoods of consecutive points, the resulting witnesses are more similar and thus yield smaller TV. In terms of prediction error, as the neighborhood radius ϵ determines the region to impose coherency, a larger region leads to greater restriction on the predictive model. All the arguments are well supported by the empirical results. We suggest users to trade off faithfulness (deviation) and smooth transition of functional properties (TV) based on the application at hand. We note that, however, smooth transition of functional properties is not equivalent to smoothness of f.

Finally, we remark that our sample complexity analysis for the linear class suggests that the neighborhood size is guaranteed to be effective for $2\epsilon + 1 > d = 2c + 1 = 9$. However, since the result is an sufficient condition, the regularization may still happens for $\epsilon < 5$ if the matrix rank of each neighborhood $X_i = [x_{i-\epsilon}, \ldots, x_{i+\epsilon}]^\top$ is less than $\min\{d, m\} = \min\{2c + 1, 2\epsilon + 1\}$.