IBM **Research**

Explaining Anomalies of Black-Box Regression Function

Tsuyoshi ("Ide-san") Idé / 井手 剛
tide@us.ibm.com
IBM T. J. Watson Research Center

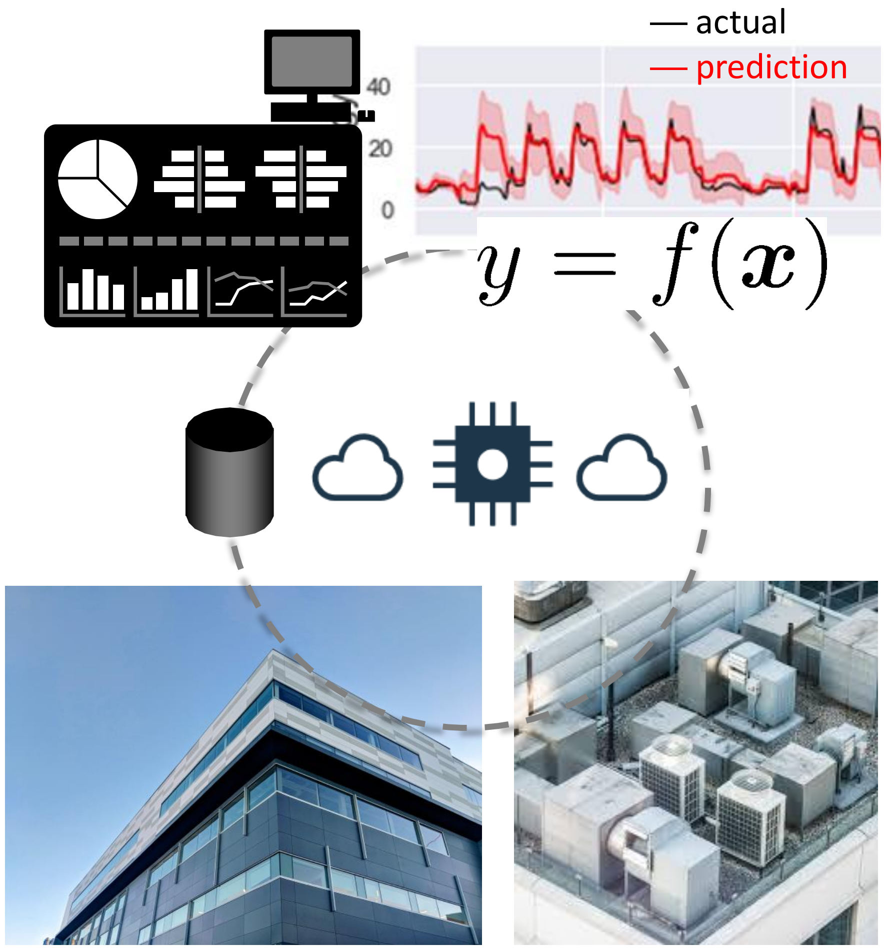
Contents

- Problem setting
- Review of existing attribution approach
- Introducing *Likelihood Compensation*
- Experimental results
- Summary

The main part of this talk has been published as:
T. Idé, A. Dhurandhar, J. Navratil, M. Singh, N. Abe, "Anomaly Attribution with Likelihood Compensation," Proc. AAAI 21, pp.4131-4138.

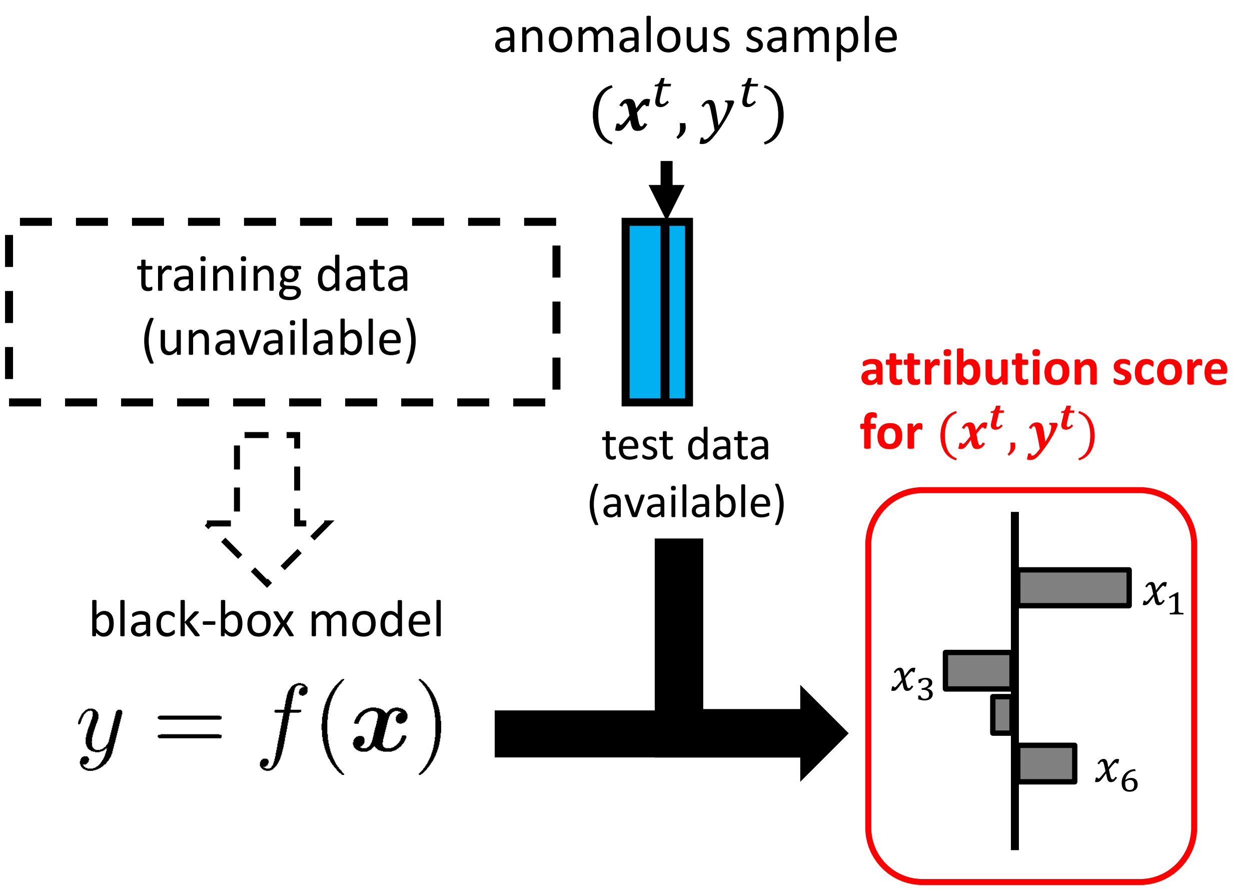
Motivating use-case: Building energy management. Deviations from the model need be explained.

- Building admin wants to keep healthy condition of building’s air-conditioning system
- He got prediction model built on data <u>under normal working conditions.</u>
  - y: building energy consumption
  - **x**: Temperature, humidity, day of week, month, room occupancy, etc.
- Then, any large deviations imply a suboptimal situation.



"Doubly black-box" is the most common industrial setting.

- Our task: Anomaly attribution
  - Compute responsibility score of each input variable
- Constraint: "doubly black-box"
  - Able to access model's API
  - Not able to access training data
  - Not able to access internal model parameters
- Note: typical end-users are not ML researchers!
  - Even you have access to the source code, the model can be a black-box



Technical task: Compute responsibility score of each input variable, given test sample(s).

Input

Test sample(s) showing anomaly/deviation

$(\mathbf{x}^t, y^t)$

t-th test sample

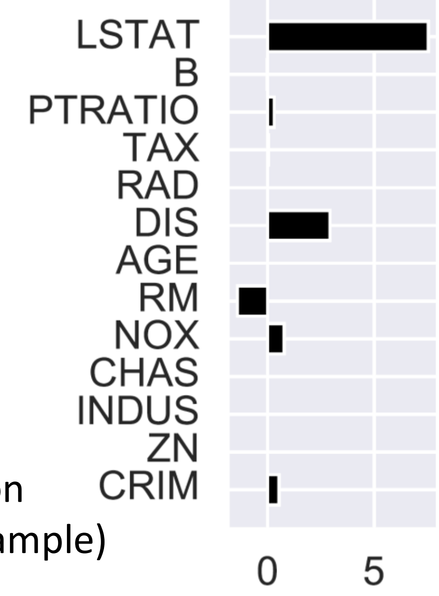
Black-box regression function

$$y = f(\boldsymbol{x})$$

anomaly attribution algorithm

Output

responsibility score computed locally at $(\mathbf{x}^t, y^t)$: $\delta_1, \ldots, \delta_M$



(from Boston Housing example)

Contents

The main part of this talk has been published as:
T. Idé, A. Dhurandhar, J. Navratil, M. Singh, N. Abe, "Anomaly Attribution with Likelihood Compensation," Proc. AAAI 21, pp.4131-4138.

Major attribution approaches: LIME, Shapley value (SV), and Integrated Gradient (IG)

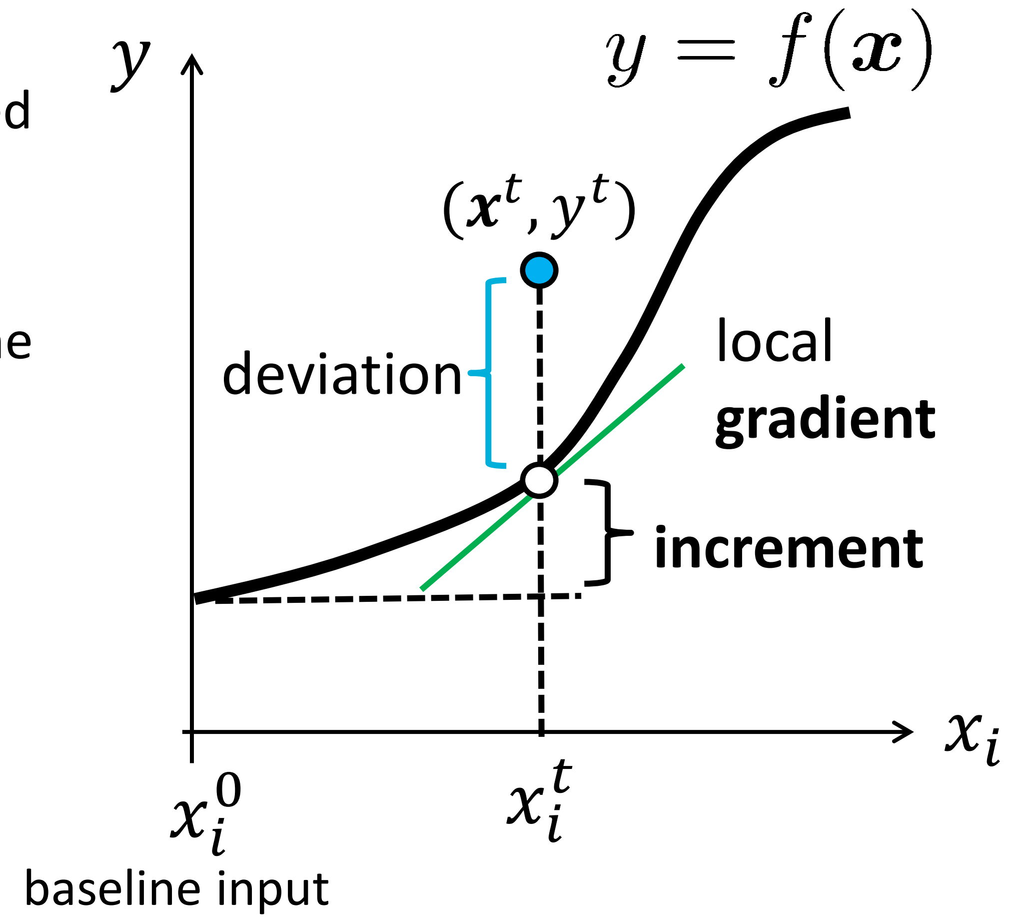
- Local linear surrogate modeling (LIME)
  - Attribution score = i-th variable's gradient estimated locally at $\boldsymbol{x}^t$
- Integrated gradient (IG)
  - $$\mathrm{IG}_i(\boldsymbol{x}^t \mid \boldsymbol{x}^0) \triangleq (x_i^t - x_i^0) \int_0^1 \mathrm{d}\alpha \left. \frac{\partial f}{\partial x_i} \right|_{\boldsymbol{x}^0 + (\boldsymbol{x}^t - \boldsymbol{x}^0)\alpha},$$
- Shapley value (SV)
  - $$\mathrm{SV}_i(\boldsymbol{x}^t) = \frac{1}{M} \sum_{k=0}^{M-1} \binom{M-1}{k}^{-1} \sum_{\mathcal{S}_i : |\mathcal{S}_i| = k} \left[\langle f \mid x_i^t, \boldsymbol{x}_{\mathcal{S}_i}^t \rangle - \langle f \mid \boldsymbol{x}_{\mathcal{S}_i}^t \rangle \right].$$

- Not intuitive.
- Tend to be used as a black-box!

conditional expectation, given S_i, a subset of the variables

Major attribution approaches: LIME, Shapley value (SV), and Integrated Gradient (IG)

- Local linear surrogate modeling (LIME)
  - Attribution score = i-th variable's gradient estimated locally at $\boldsymbol{x}^t$
- Integrated gradient (IG)
  - Attribution score = i-th variable's contribution to the increment from the baseline input $\boldsymbol{x}^0$
    - ✓ $\sum_i \mathrm{IG}_i(\boldsymbol{x}^t) = f(\boldsymbol{x}^t) - f(\boldsymbol{x}^0)$
- Shapley value (SV)
  - Attribution score = i-th variable's contribution to expected increment
    - ✓ $\sum_i \mathrm{SV}_i(\boldsymbol{x}^t) = f(\boldsymbol{x}^t) - \langle f \rangle$



Hidden secret: Existing "anomaly attribution" methods do not explain deviations!

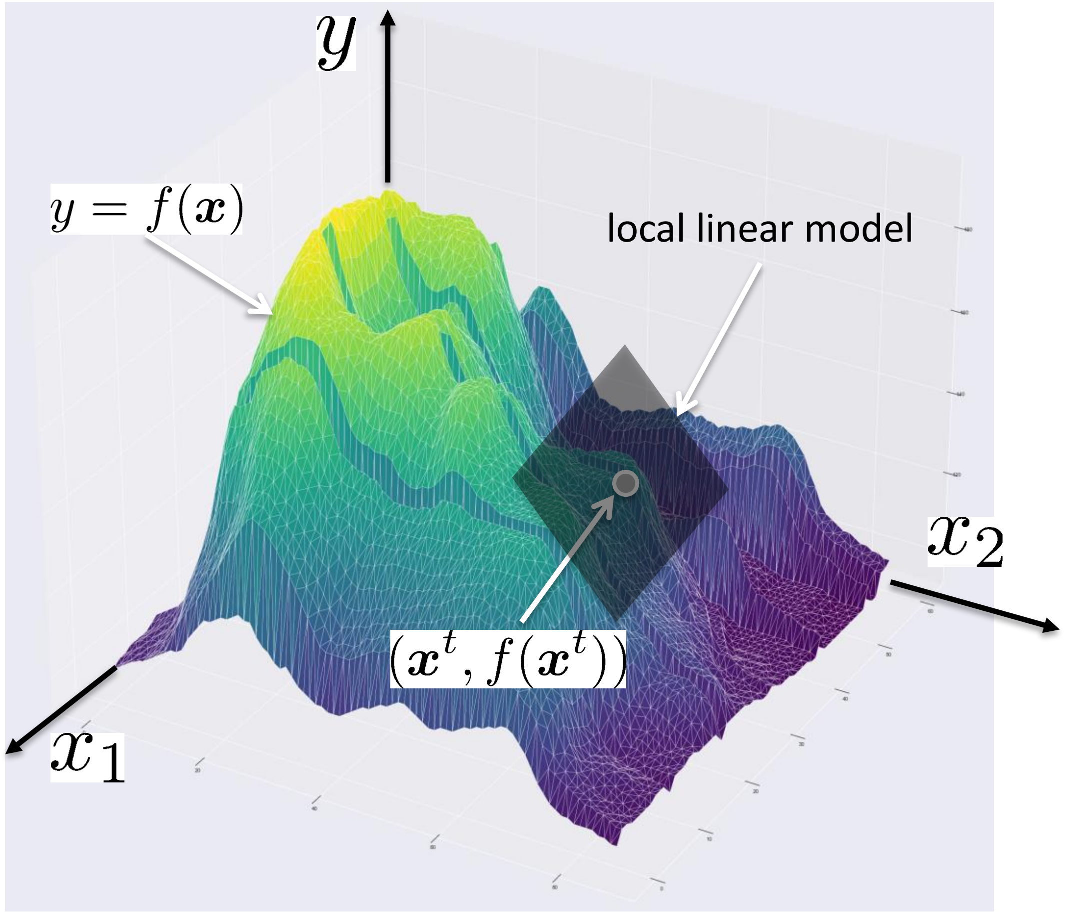
- Local linear surrogate modeling (LIME)
  - Attribution score = i-th variable's gradient estimated locally at $\boldsymbol{x}^t$
- Integrated gradient (IG)
  - Attribution score = i-th variable's contribution to the increment from the baseline input $\boldsymbol{x}^0$
    - ✓ $\sum_i \mathrm{IG}_i(\boldsymbol{x}^t) = f(\boldsymbol{x}^t) - f(\boldsymbol{x}^0)$
- Shapley value (SV)
  - Attribution score = i-th variable's contribution to expected increment
    - ✓ $\sum_i \mathrm{SV}_i(\boldsymbol{x}^t) = f(\boldsymbol{x}^t) - \langle f \rangle$

Provides local gradient

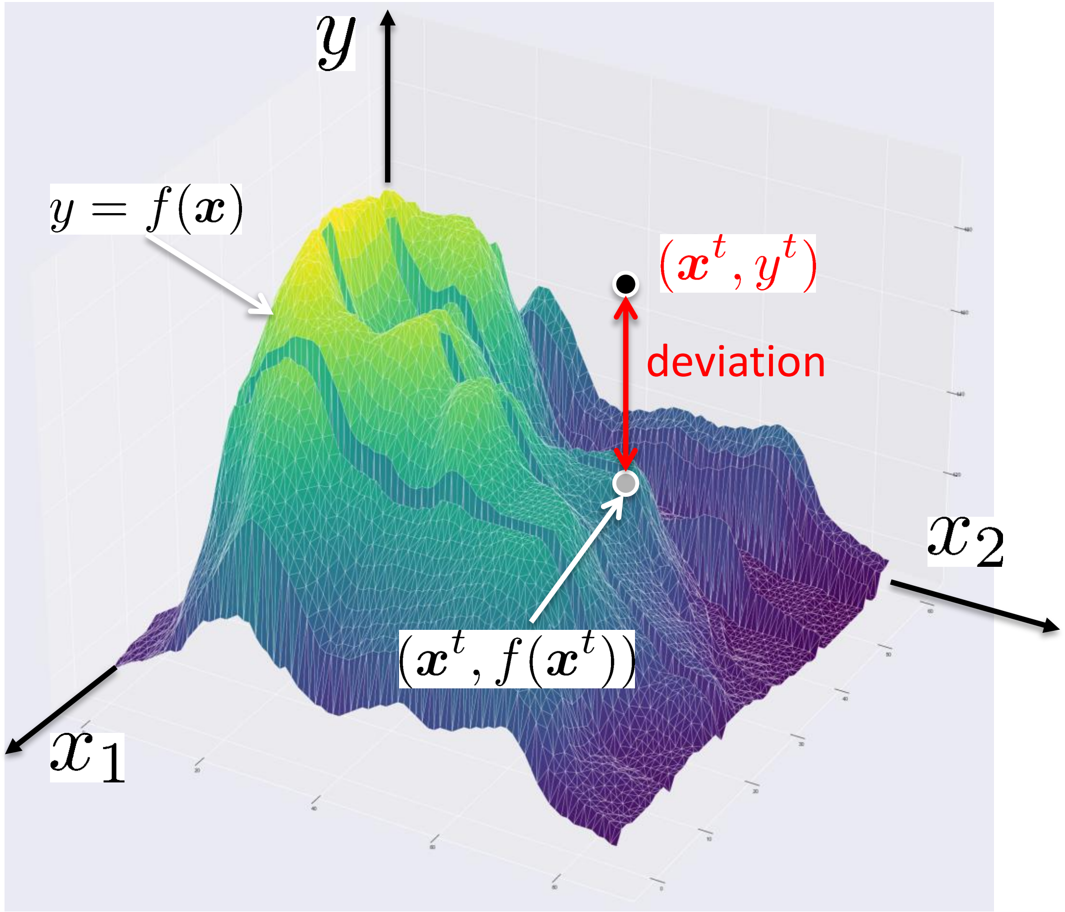
Explain increment

Hidden secret: Existing "anomaly attribution" methods do not explain deviations! Example of LIME

Local surrogate model to explain *f*(**x**)



Anomaly attribution needs to explain *f*(***x***) - *y*



Hidden secret: Existing "anomaly attribution" methods do not explain deviations! LIME, IG, SV are deviation-agnostic.

- Local linear surrogate modeling (LIME)
  - Attribution score = i-th variable's gradient estimated locally at $\boldsymbol{x}^t$
- Integrated gradient (IG)
  - Attribution score = i-th variable's contribution to the increment from the baseline input $\boldsymbol{x}^0$
    - ✓ $\sum_i \mathrm{IG}_i(\boldsymbol{x}^t) = f(\boldsymbol{x}^t) - f(\boldsymbol{x}^0)$
- Shapley value (SV)
  - Attribution score = i-th variable's contribution to expected increment
    - ✓ $\sum_i \mathrm{SV}_i(\boldsymbol{x}^t) = f(\boldsymbol{x}^t) - \langle f \rangle$

deviation-agnostic

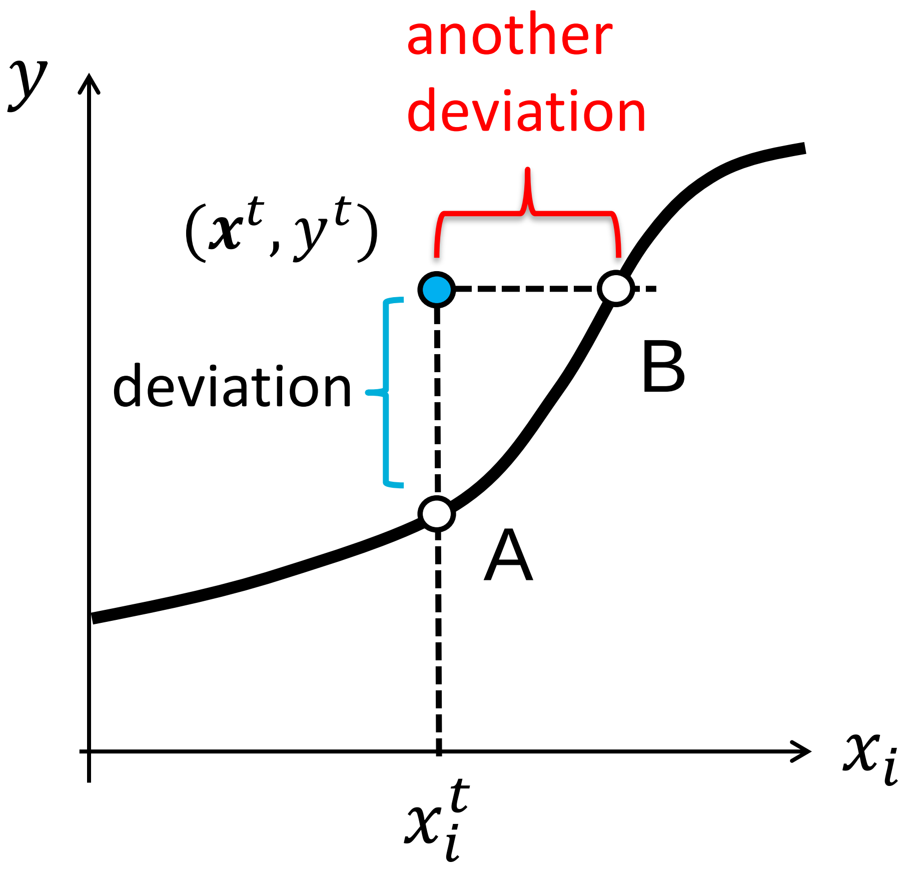
Contents

- Problem setting
- Review of existing attribution approach
- Introducing *Likelihood Compensation*
- Experimental results
- Summary

The main part of this talk has been published as:
T. Idé, A. Dhurandhar, J. Navratil, M. Singh, N. Abe, “Anomaly Attribution with Likelihood Compensation,” Proc. AAAI 21, pp.4131-4138.

High-level idea: Focus on another deviation

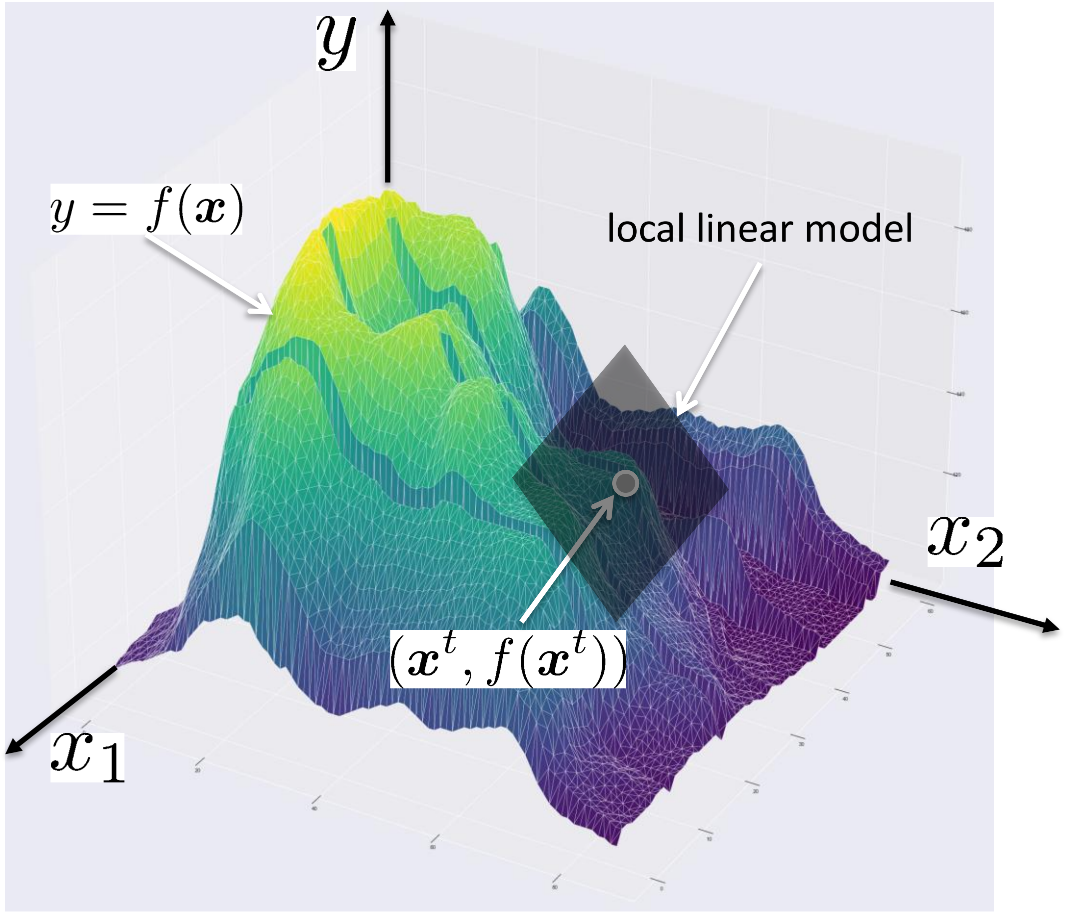
- An anomaly implies a deviation from a certain reference point.
- Point A is typically used to determine the anomalousness.
  - But is not useful for attribution purposes.
- What if point B is used?
  - How do we characterize Point B?

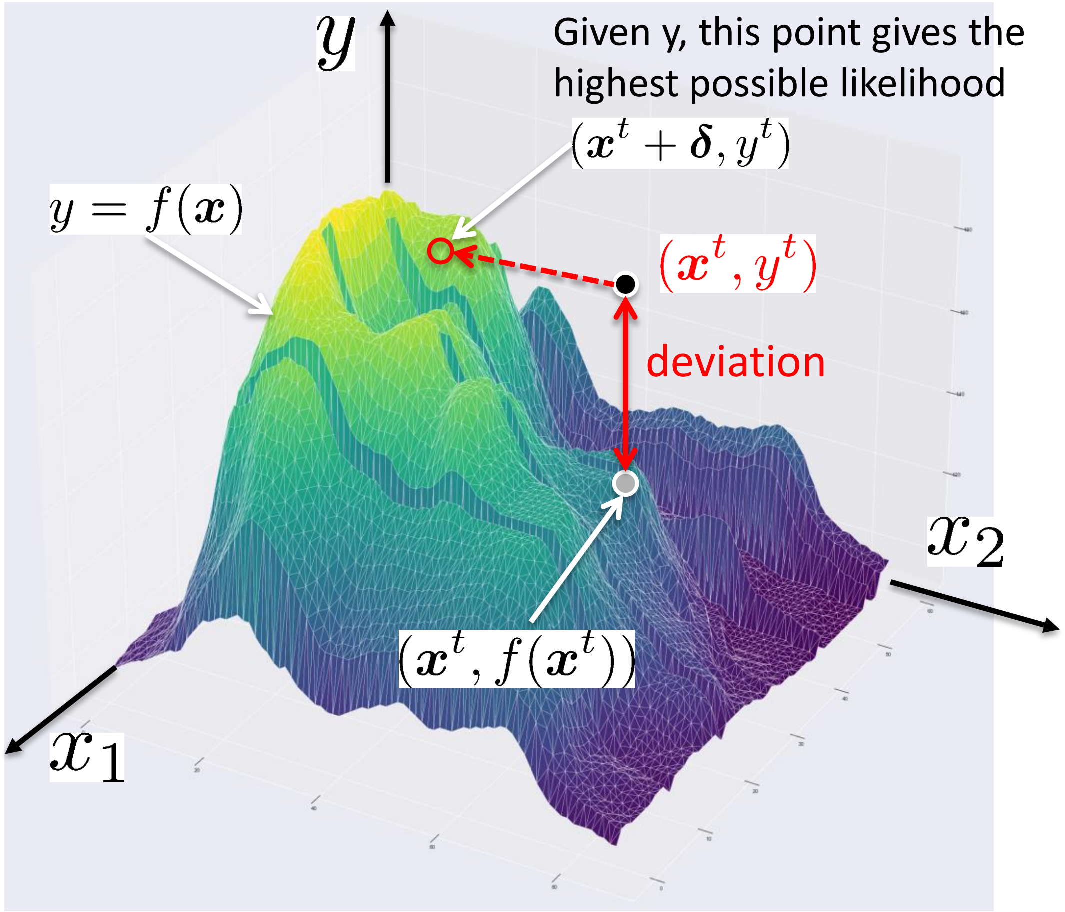


High-level idea: Defining responsibility score through local perturbation as "horizontal deviation"

$\boldsymbol{\delta}$: responsibility score
("likelihood compensation")

Local surrogate model to explain *f*(**x**)





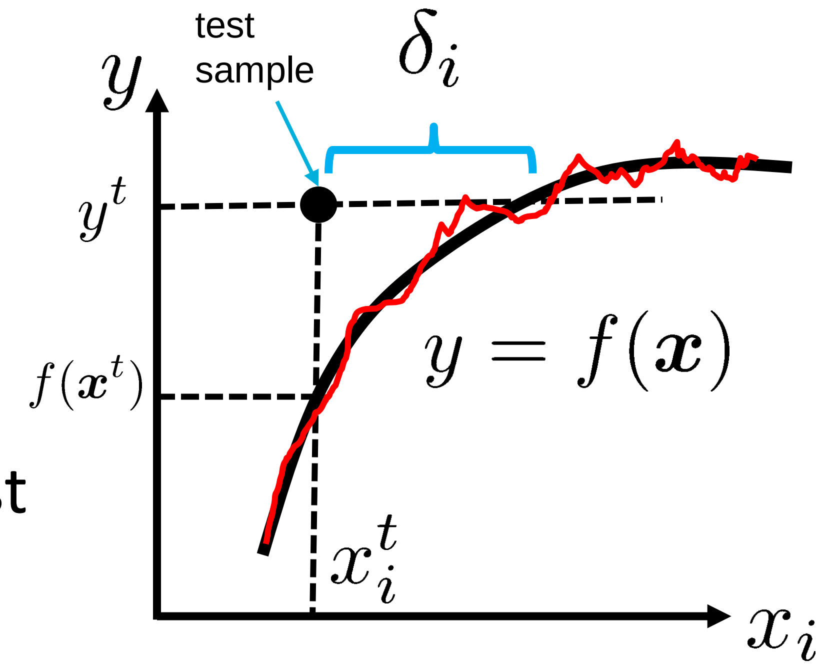
Likelihood Compensation (LC): Seeking a perturbation that achieves highest possible likelihood in the vicinity

- We use the horizontal deviation as the attribution score
  - i.e., A measure of responsibility of each variable.

- Likelihood compensation $\boldsymbol{\delta}$:
  - $\delta^* = \underset{\boldsymbol{\delta}}{\mathrm{argmax}} \{\ln p(\, y^t \mid \boldsymbol{x}^t + \boldsymbol{\delta}\,)\}$
    - ✓ $\mathrm{s.t.}\quad \boldsymbol{x}^t + \boldsymbol{\delta} \in$ vicinity of $\boldsymbol{x}^t$

- $\boldsymbol{\delta}$ is a perturbation such that $\boldsymbol{x}^t + \boldsymbol{\delta}$ achieves the best possible fit to the model
  - $\boldsymbol{\delta}$ compensates for the loss in likelihood incurred by an anomalous prediction.

LC can be thought of as the '**deviation measured horizontally**'



Gaussian-based representation of the LC problem

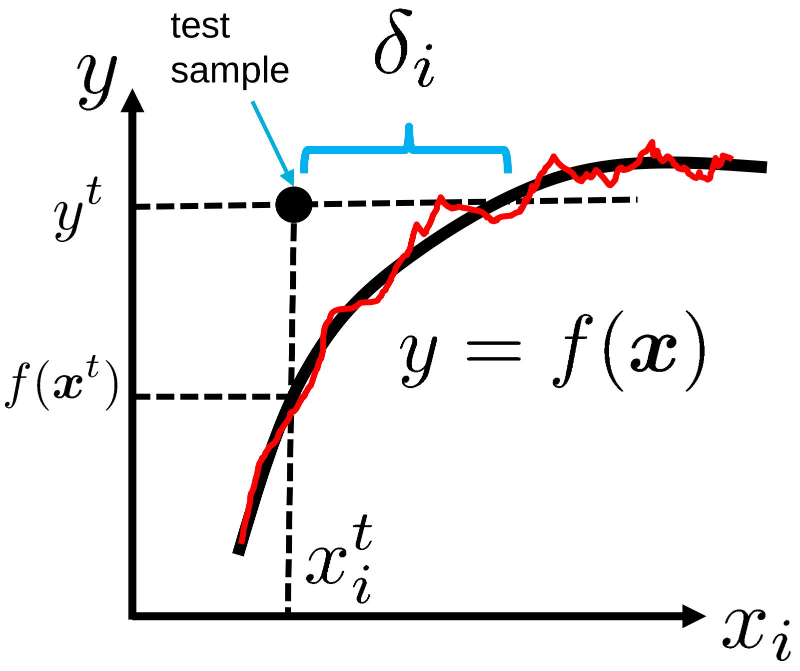
- When $p(\, y \mid x\,)$ is Gaussian, LC's optimization problem can be written as

$$\min_{\boldsymbol{\delta}} \left\{ \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2\sigma_t^2} + \frac{1}{2}\lambda \|\boldsymbol{\delta}\|_2^2 + \nu \|\boldsymbol{\delta}\|_1 \right\},$$

  - σ_t^2 : local variance at $\boldsymbol{x}^t$
  - λ, ν: regularization parameters (hyper parameters)

- Looks simple but challenging to solve when $f(\boldsymbol{x})$ is a black-box function with potential non-smoothness.

LC can be thought of as the
'**deviation measured horizontally**'



Generalization and interesting contrast to adversarial training

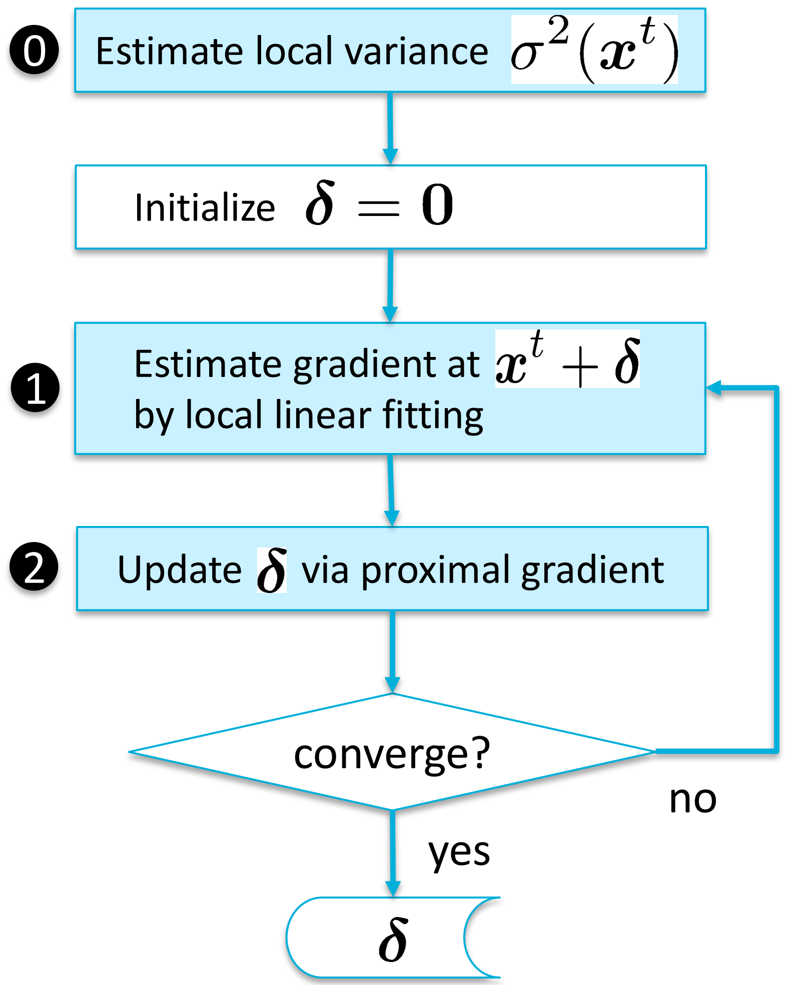
- LC's optimization problem can be generalizaed as
  - $\min_{\boldsymbol{\delta}} \langle \mathrm{Loss}(y^t, \boldsymbol{x}^t + \boldsymbol{\delta}) \rangle \quad \mathrm{s.t.} \quad \boldsymbol{x}^t + \boldsymbol{\delta} \in \text{vicinity of } \boldsymbol{x}^t$
  - $\langle \cdots \rangle$ is empirical average over test samples and Loss is $-\ln p(\, y^t \mid \boldsymbol{x}^t \,)$
  - This is to compute the attribution scores for a collection of test samples

- Adversarial training
  - $\min_{\theta} \max_{\boldsymbol{\delta}} \langle \mathrm{Loss}(y^t, \boldsymbol{x}^t + \boldsymbol{\delta} \mid \theta) \rangle \quad \mathrm{s.t.} \quad \boldsymbol{x}^t + \boldsymbol{\delta} \in \text{vicinity of } \boldsymbol{x}^t$
    - ✓ θ is the model parameter (unavailable in the doubly black-box setting)
  - In Adversarial training, $\boldsymbol{x}^t$ is normal. $\boldsymbol{x}^t + \boldsymbol{\delta}$ is abnormal (adversarial)
  - In LC, $\boldsymbol{x}^t$ is abnormal. $\boldsymbol{x}^t + \boldsymbol{\delta}$ is normal

Solving optimization problem by iterating local smooth approximation and proximal gradient

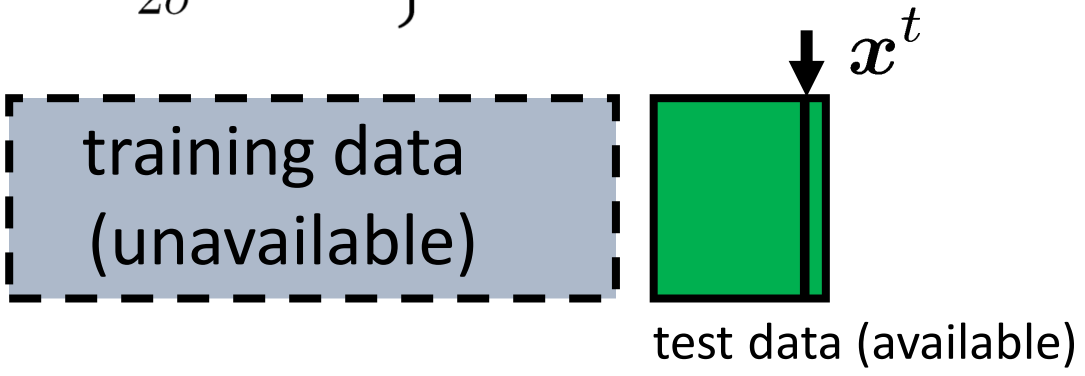
- $f(\boldsymbol{x})$ is black-box. It may not be even smooth or continuous
- 0. Local variance estimation (only once)
  - Leverage available test data or prior knowledge
- 1. Local gradient estimation of f
  - Amounts to smooth approximation of f
- 2. Proximal gradient update for $\boldsymbol{\delta}$

iterate



0. Local variance estimation

- If test samples available are too few, use a constant variance to define a Gaussian observation model
  - $p(y \mid \boldsymbol{x}) = \mathcal{N}(y \mid f(\boldsymbol{x}), \sigma^2)$
- If some amount of test samples are available, use locally weighted maximum likelihood to estimate an input-dependent variance
  - $$\sigma^2(\boldsymbol{x}^t) = \max_{\sigma^2} \sum_{n=1}^{N_{\text{heldout}}} \underbrace{w_n(\boldsymbol{x}^t)}_{\text{Gaussian kernel defined for the specific test sample } \boldsymbol{x}^t} \left\{ \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y^{(n)} - f(\boldsymbol{x}^{(n)}))^2}{2\sigma^2} \right\},$$



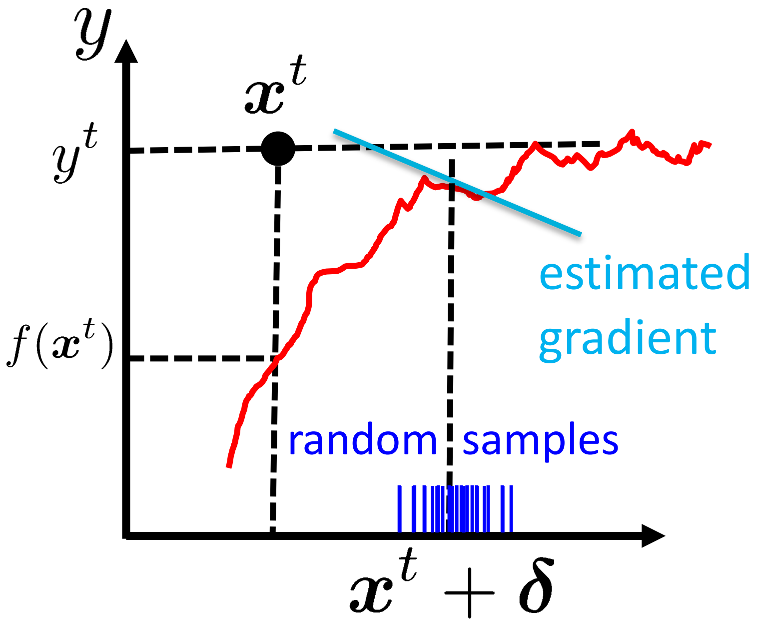
1. Local gradient estimation of *f*

- We solve the problem with gradient descent
  - $$\min_{\boldsymbol{\delta}} \left\{ \underbrace{\frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2\sigma_t^2} + \frac{1}{2}\lambda \|\boldsymbol{\delta}\|_2^2}_{} + \nu \|\boldsymbol{\delta}\|_1 \right\},$$

gradient: $$\frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\sigma_t^2} \left\langle\!\!\left\langle \frac{\partial f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \right\rangle\!\!\right\rangle + \lambda \boldsymbol{\delta}$$

Smooth surrogate of gradient at $\boldsymbol{x}^t + \boldsymbol{\delta}$

- We use a simple sampling-based algorithm
  - At a given test location $\boldsymbol{x}^t$, we random-sample N_s samples in the vicinity of $\boldsymbol{x}^t$, and fit a linear regression model
    - ✓ N_s ~ 1000.
    - ✓ Assumption: evaluation of $f(\boldsymbol{x})$ can be done cheaply
  - The gradient is obtained as the regression coefficient.



2. Proximal gradient update for δ

- The objective now looks like L_1-regularized convex-*ish* optimization
  - $$\min_{\boldsymbol{\delta}} \left\{ \underbrace{\frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2\sigma_t^2} + \frac{1}{2}\lambda \|\boldsymbol{\delta}\|_2^2}_{\text{convex-ish function with the smoothed gradient}} + \nu \|\boldsymbol{\delta}\|_1 \right\},$$

convex-*ish* function with the smoothed gradient

$$J(\boldsymbol{\delta}) \triangleq \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{[y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})]^2}{2\sigma_t^2} + \frac{1}{2}\lambda \|\boldsymbol{\delta}\|_2^2$$

- Building an updating rule from δ^{old} using prox gradient-like algorithm
  - $$\boldsymbol{\delta} = \arg\min_{\boldsymbol{\delta}} \left\{ \underbrace{J(\boldsymbol{\delta}^{\text{old}}) + (\boldsymbol{\delta} - \boldsymbol{\delta}^{\text{old}})\langle\!\langle \nabla J(\boldsymbol{\delta}^{\text{old}}) \rangle\!\rangle + \frac{1}{2\kappa} \|\boldsymbol{\delta} - \boldsymbol{\delta}^{\text{old}}\|_2^2}_{\text{smooth quadratic approximation of } J} + \nu \|\boldsymbol{\delta}\|_1 \right\}$$

smooth quadratic approximation of *J*

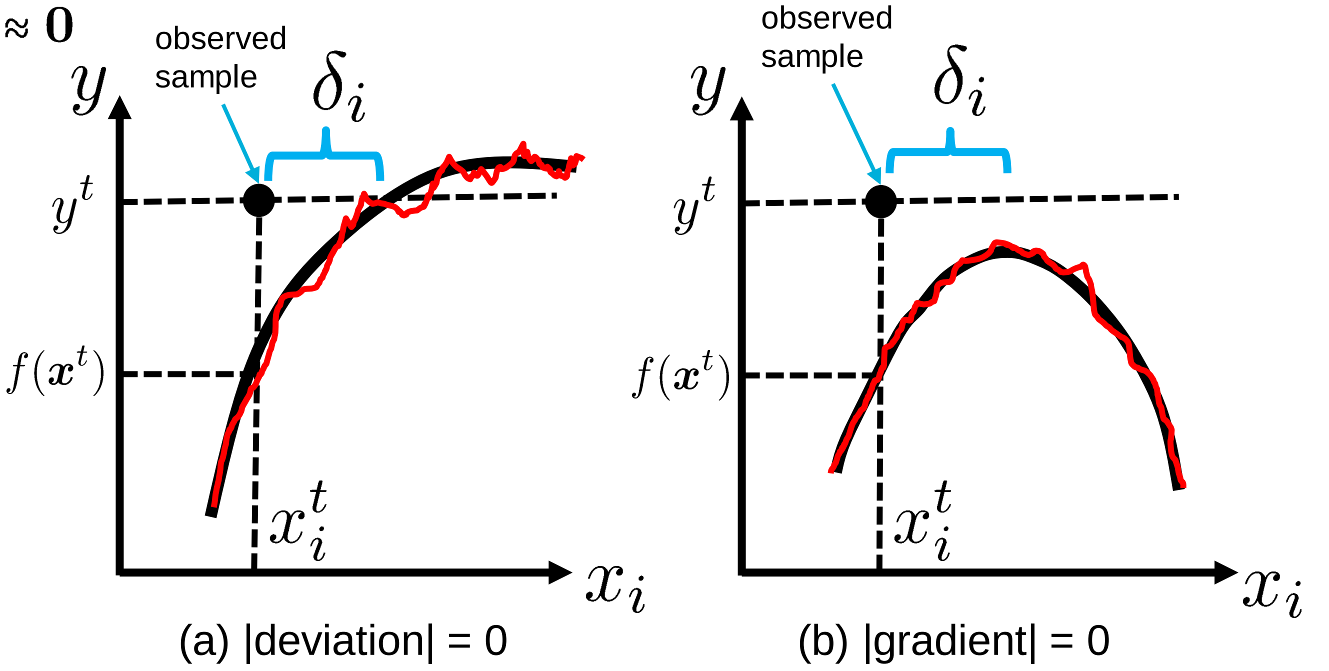
$$= \text{prox}_{\kappa\nu\|\boldsymbol{\delta}\|_1} \left(\boldsymbol{\delta}^{\text{old}} - \kappa \langle\!\langle \nabla J(\boldsymbol{\delta}^{\text{old}}) \rangle\!\rangle \right)$$

The L_1 prox operator has an analytic solution! (→ paper)

Condition of convergence – where the intuition of "horizontal deviation" comes from

- The prox gradient-like update converges when
  - $$\frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{y^t - f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\sigma_t^2} \left\langle\!\!\left\langle \frac{\partial f(\boldsymbol{x}^t + \boldsymbol{\delta})}{\partial \boldsymbol{\delta}} \right\rangle\!\!\right\rangle \approx \mathbf{0}$$
- Condition (a): |deviation|= 0
  - Met when $y^t = f(\boldsymbol{x}^t + \boldsymbol{\delta})$
  - "Keep the height, move horizontally until you hit *f* "
- Condition (b): |gradient|= 0
  - In case there is no horizontal intersection, this warrants convergence

Illustration for N_{test} = 1



Contents

- Problem setting
- Review of existing attribution approach
- Introducing *Likelihood Compensation*
- Experimental results
- Summary

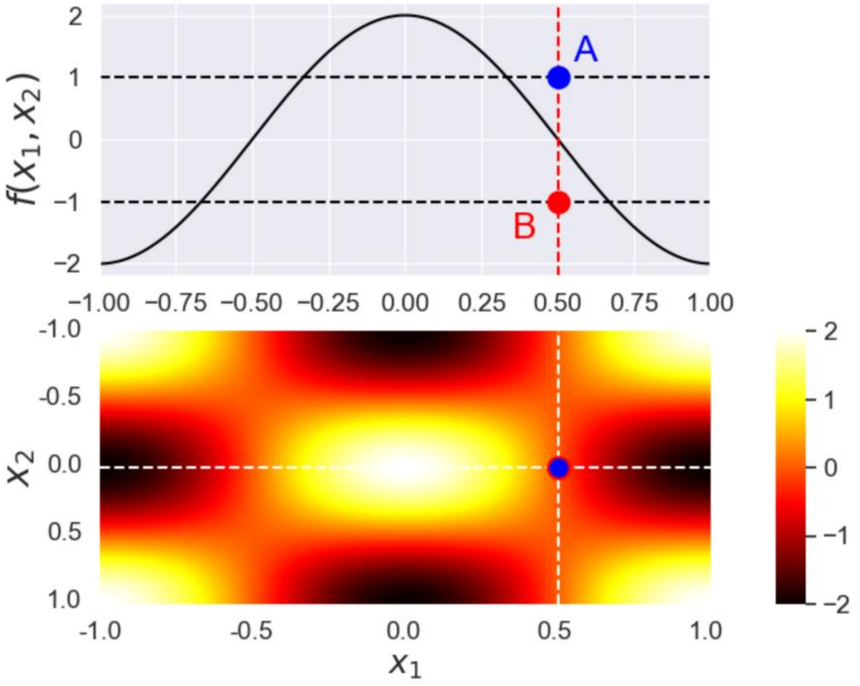
The main part of this talk has been published as:
T. Idé, A. Dhurandhar, J. Navratil, M. Singh, N. Abe, “Anomaly Attribution with Likelihood Compensation,” Proc. AAAI 21, pp.4131-4138.

Baseline methods

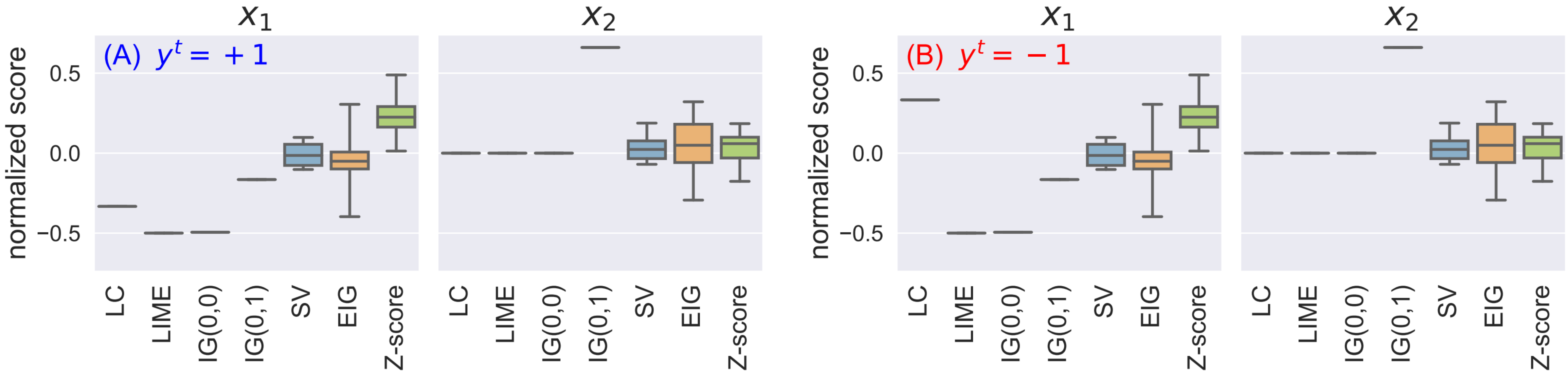
- Existing methods either need training data or is deviation-agnostic.

	training-data-free	baseline-free	y-sensitive	reference point
LIME	yes	yes	no	infinitesimal vicinity
SV	no	yes	no	globally distributional
IG	yes	no	no	arbitrary
EIG	no	yes	no	globally distributional
Z-score	no	yes	no	global mean of predictors
LC	**yes**	**yes**	**yes**	maximum likelihood point

Two-dimensional synthetic data: Existing methods are “deviation-agnostic”

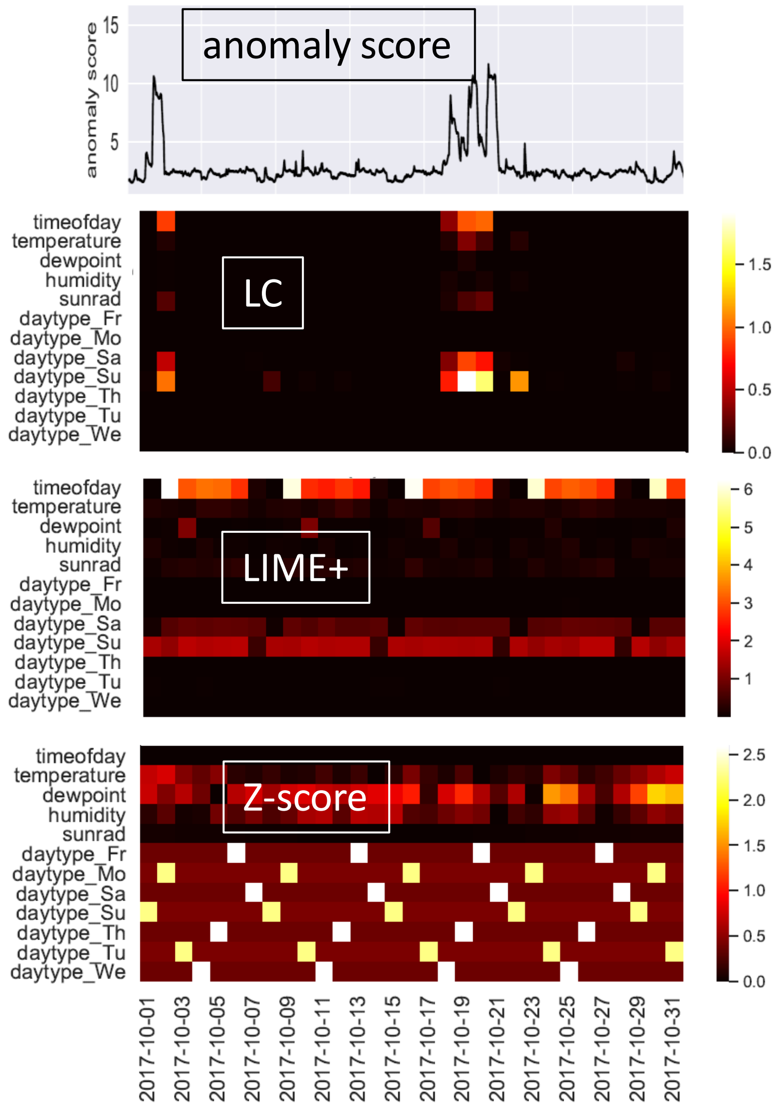


- x1 should be responsible for the outliers A, B
  - LC, LIME IG successfully identified x1
  - But only LC can distinguish points A and B.
- SV, expected IG (EIG), Z-score suffer significant variability issue due to the need for the true distribution P(x).



Comparison with LIME+ and Z-score in building energy use-case

- One month-worth building energy data
  - *y*: energy consumption
  - ***x***: time of day, temperature, humidity, sunrad, day of week (one-hot encoded)
- The score is computed based on hourly 24 test points for each day
  - The mean of the absolute values are visualized
  - SV+ was not computable due to lack of training data
- LIME+ is insensitive to outliers
  - LIME score remain the same for any outliers, making it less useful in anomaly attribution
- Z-score does not depend on *y* (by definition)
  - The artifact for the day-of-week variables is due to one-hot encoding



Contents

Summarizing practical features of LC

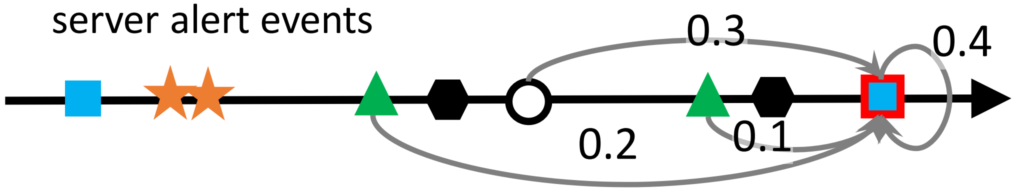
- LC is deviation-sensitive
- LC is model-agnostic

- LC is directly interpretable
  - LC represents “what you could have done for the best fit” for each input
  - Naturally provides counterfactual explanations
    - ✓ LC > 0 for a temperature variable, for example, reads “To be consistent to the observed y^t, the temperature could have been higher.”
    - ✓ Or simply, “Your temperature was too low for y^t”

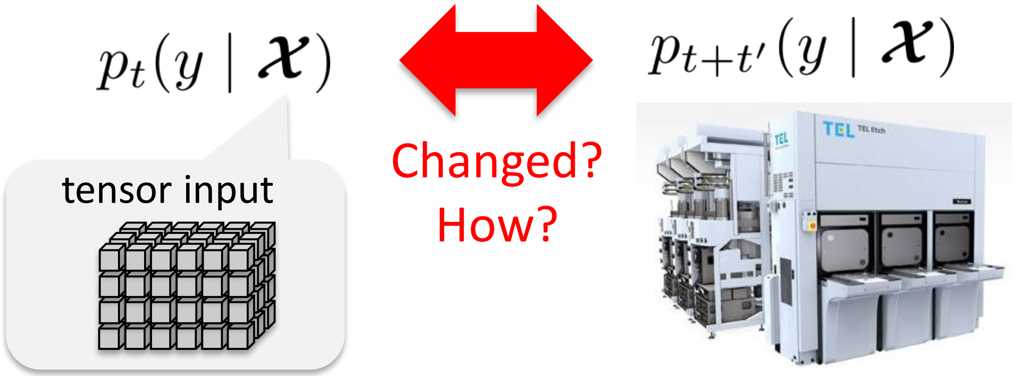
My other recent works on XAI for business actionability

(Full publication list → https://ide-research.net)

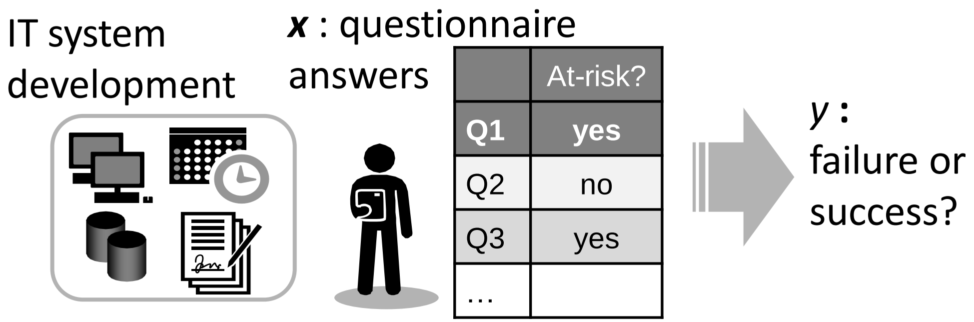
- Causal diagnosis from event data
  - For AIOps application (event grouping)
  - Idé et al., "Cardinality-Regularized Hawkes-Granger Model," NeurIPS 21 [slides, paper].



- Actionable change detection for tensor inputs
  - For semiconductor tool monitoring
  - Idé, "Tensorial Change Analysis using Probabilistic Tensor Regression," AAAI 19 [poster, paper].



- Project failure risk prediction through psychometric analysis of questionnaire data
  - For risk management of IBM projects
  - Idé & Dhurandhar, "Informative Prediction based on Ordinal Questionnaire Data," ICDM 15 [slides, paper].



Thank you!