Slides available! https://ide-research.net/papers/2021_AAA_Ide_presentation.pdf

IBM **Research**

Anomaly Attribution with Likelihood Compensation

Tsuyoshi ("Ide-san") Idé, Amit Dhurandhar, Jiří Navrátil, Moninder Singh, Naoki Abe
{tide, adhuran, jiri, moninder, nabe}@us.ibm.com
IBM T. J. Watson Research Center

- To be presented at the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21, Feb 2-9, 2021)

Contents

Technical task: anomaly attribution for black-box regression

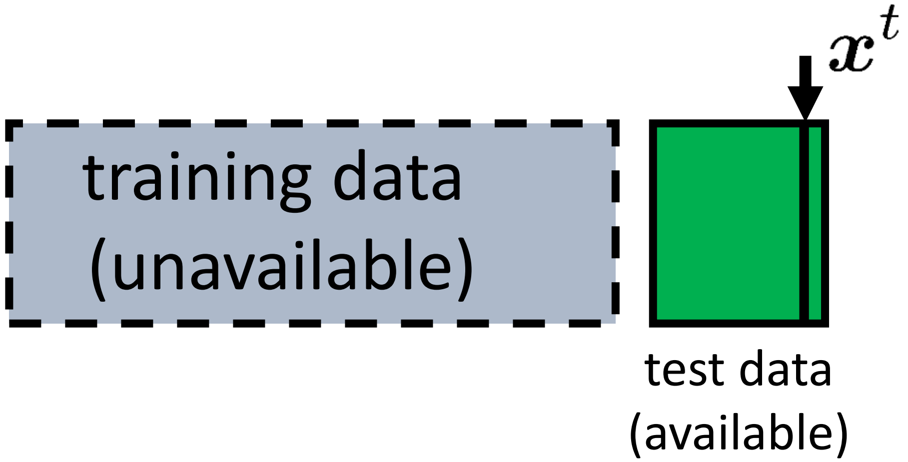
- **Task**: Attribute deviation from black-box prediction *f*(***x***) to each input variable
- **Background**: Most of XAI methods are designed to explain *f*(***x***), not deviations
- **Solution**: New notion of "likelihood compensation"
  - Define the responsibility through perturbation to achieve the highest possible likelihood

Black-box regression function

real-valued

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

Multivariate vector in $\mathbb{R}^M$



Slides available! https://ide-research.net/papers/2021_AAA_Ide_presentation.pdf

Technical task: anomaly attribution for black-box regression

Input and output

Input

Test sample(s) showing anomaly/deviation

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

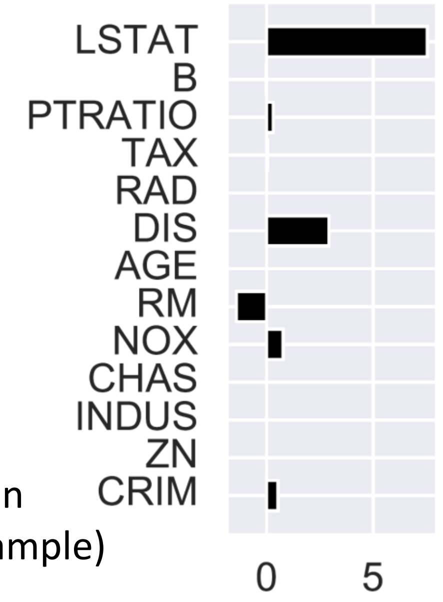
Black-box regression function

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

Likelihood compensation algorithm

Output

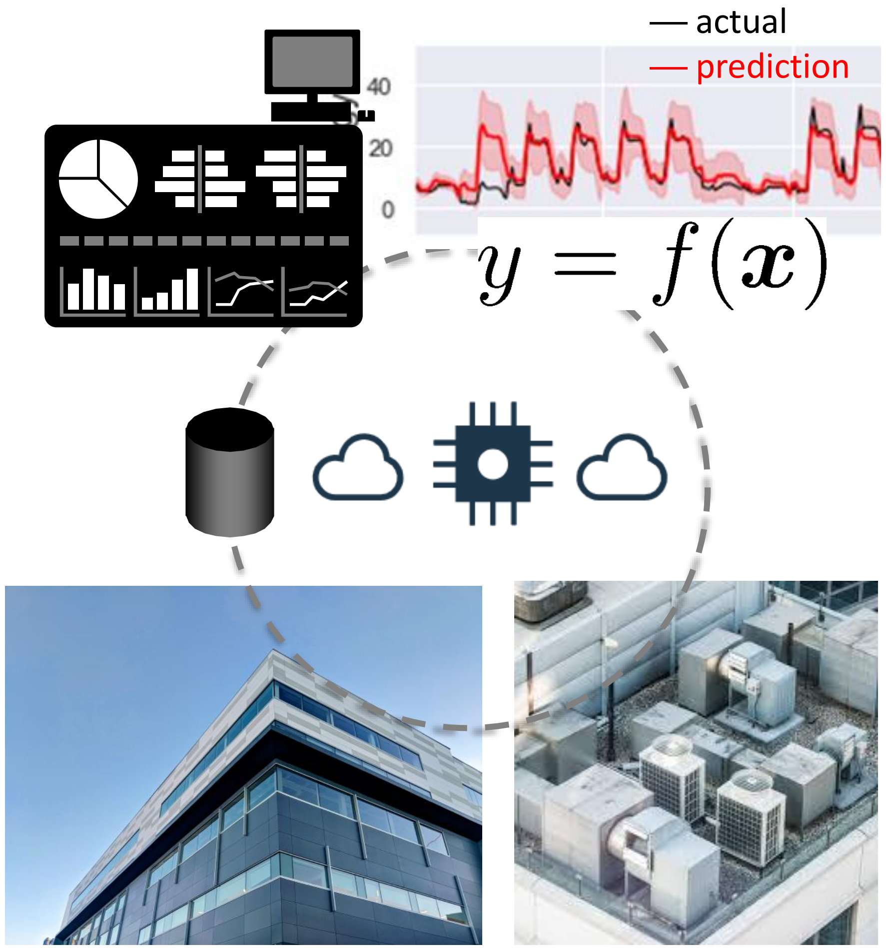
responsibility score computed locally at $(\mathbf{x}^t, y^t)$: $\delta_1, \ldots, \delta_M$



(from Boston Housing example)

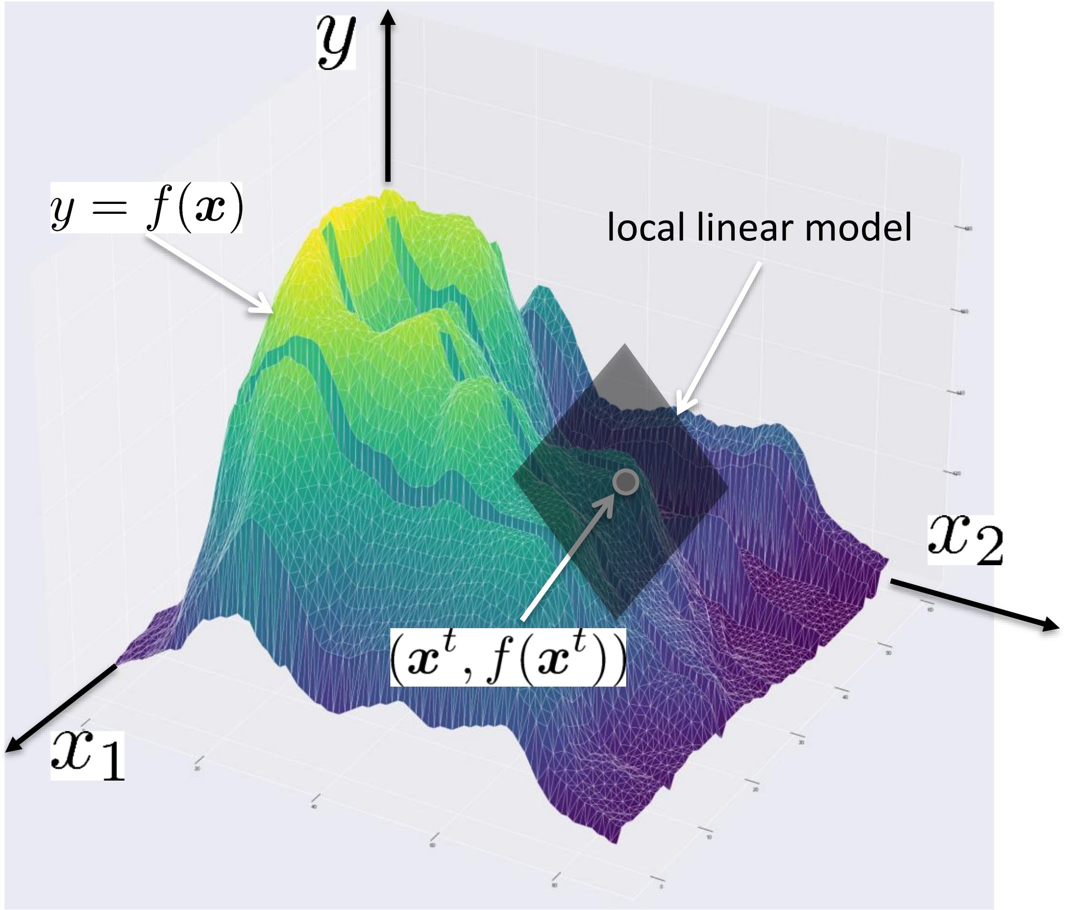
Use-case example: Building energy management

- Use case example: building management
  - y: building energy consumption
  - **x**: Temperature, humidity, day of week, month, room occupancy, etc.
- Building admin (primary end-user) does not have full visibility of the model *f*, training data, and sensing system
  - AI vendor/SIer/HVAC constructor often use proprietary technologies
  - Only some amount of test data is accessible

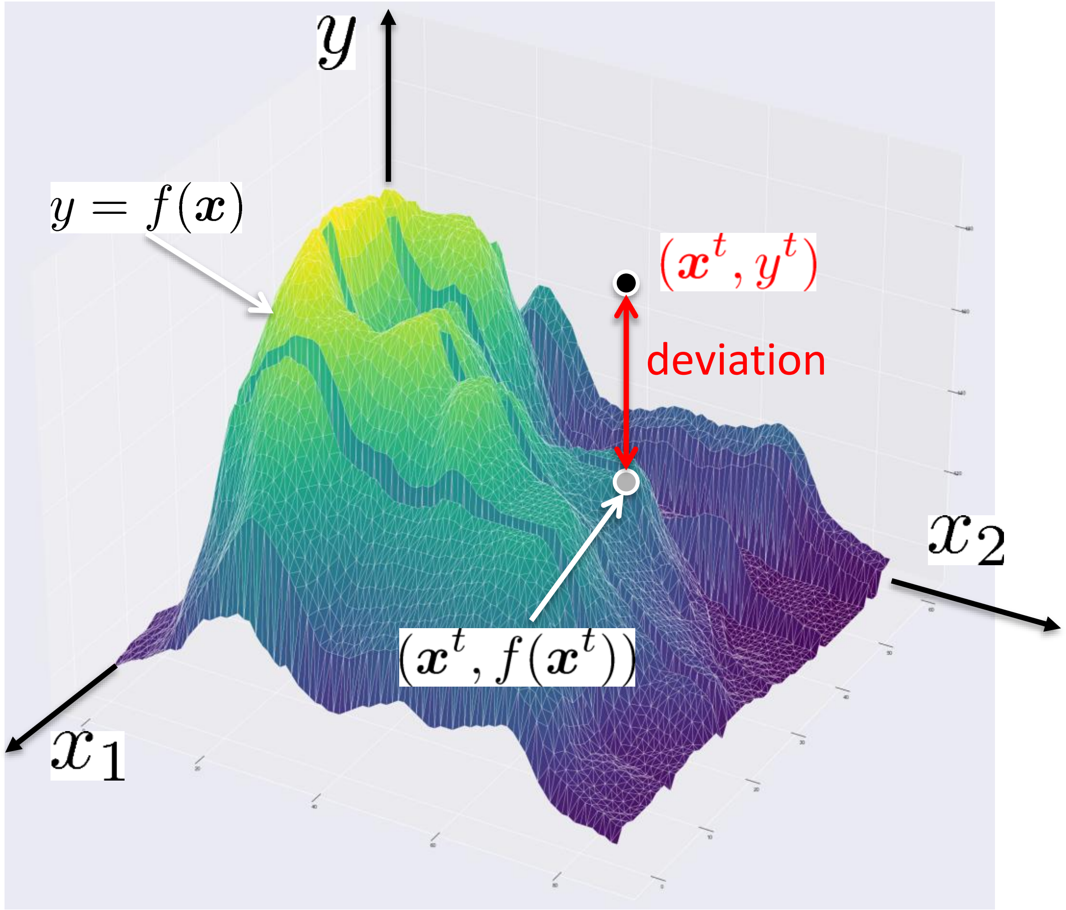


Local surrogate model for *f*(*x*) alone cannot explain deviations. We need a new idea.

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



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

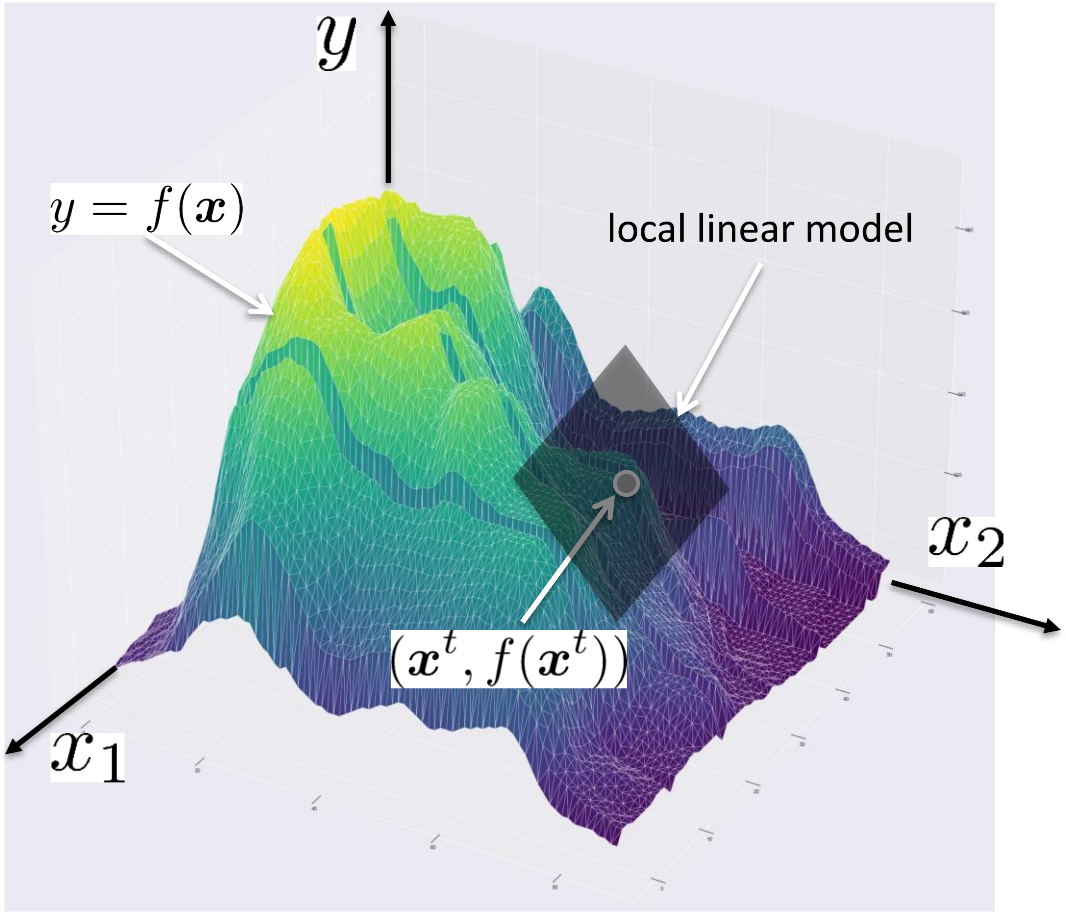


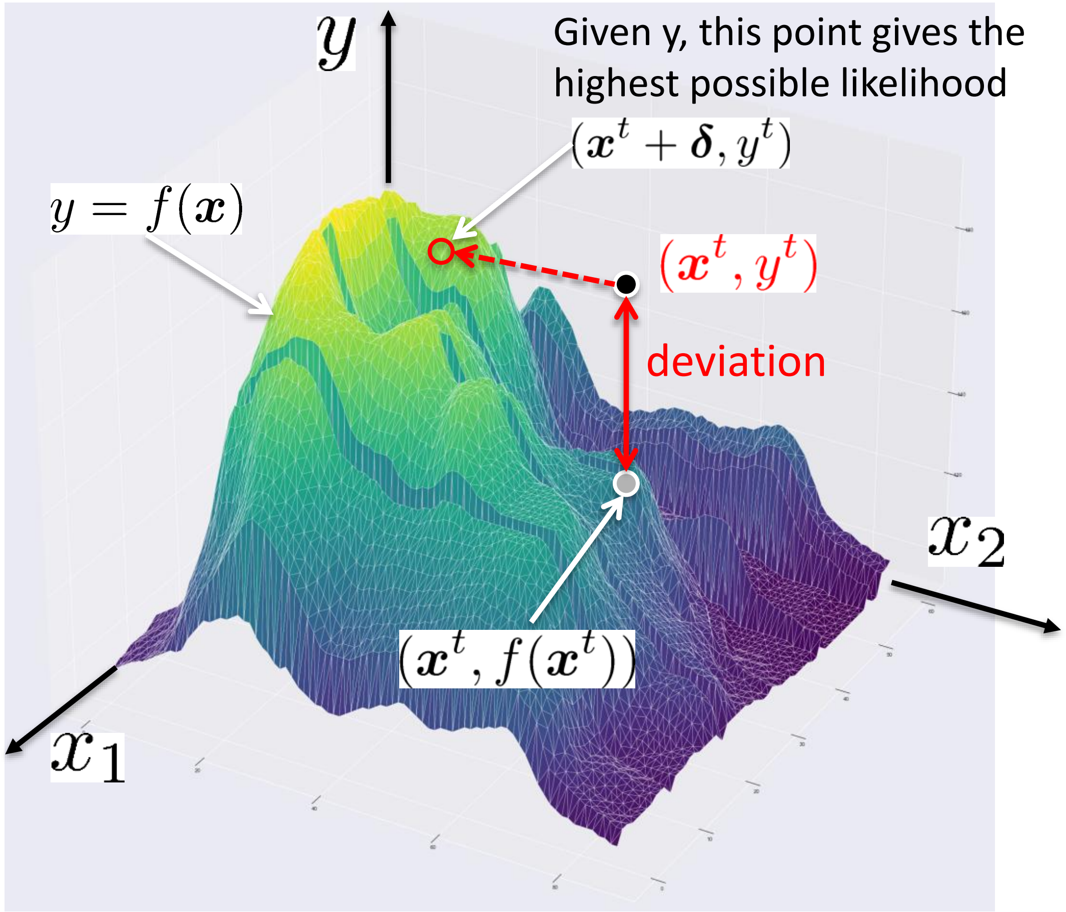
Contents

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

δ : responsibility score ("likelihood compensation")

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

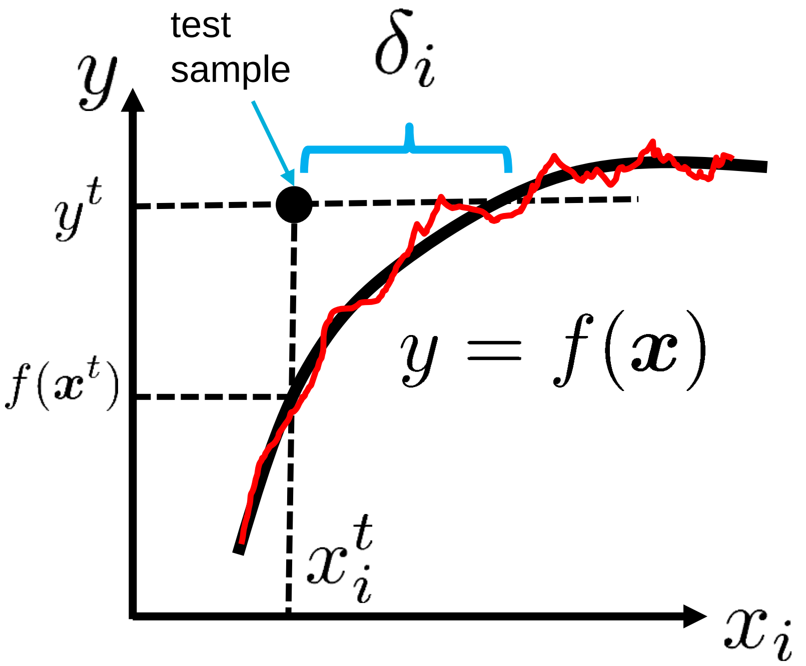




Defining Likelihood Compensation (LC) as optimal perturbation

- Likelihood compensation $\boldsymbol{\delta}$: A perturbation to $\boldsymbol{x}^t$ such that $\boldsymbol{x}^t + \boldsymbol{\delta}$ achieves the best possible fit to the model
  - The log likelihood $\ln p(y^t \mid f(\boldsymbol{x}^t))$ is a measure of goodness-of-fit of a test sample ($\mathbf{x}^t$,y^t)
  - LC seeks a best possible fit by correcting $\boldsymbol{x}^t$ under a certain regularization
    - ✓ $\boldsymbol{\delta} = \arg\max_{\boldsymbol{\delta}} \left[\ln \left\{ \underbrace{p(y^t \mid f(\boldsymbol{x}^t + \boldsymbol{\delta}))}_{\text{Gaussian}} \ \underbrace{p(\boldsymbol{\delta})}_{\text{elastic net}} \right\} \right],$

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



- The main optimization problem

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

(λ and ν are constant)

Iterating local smooth approximation and proximal gradient

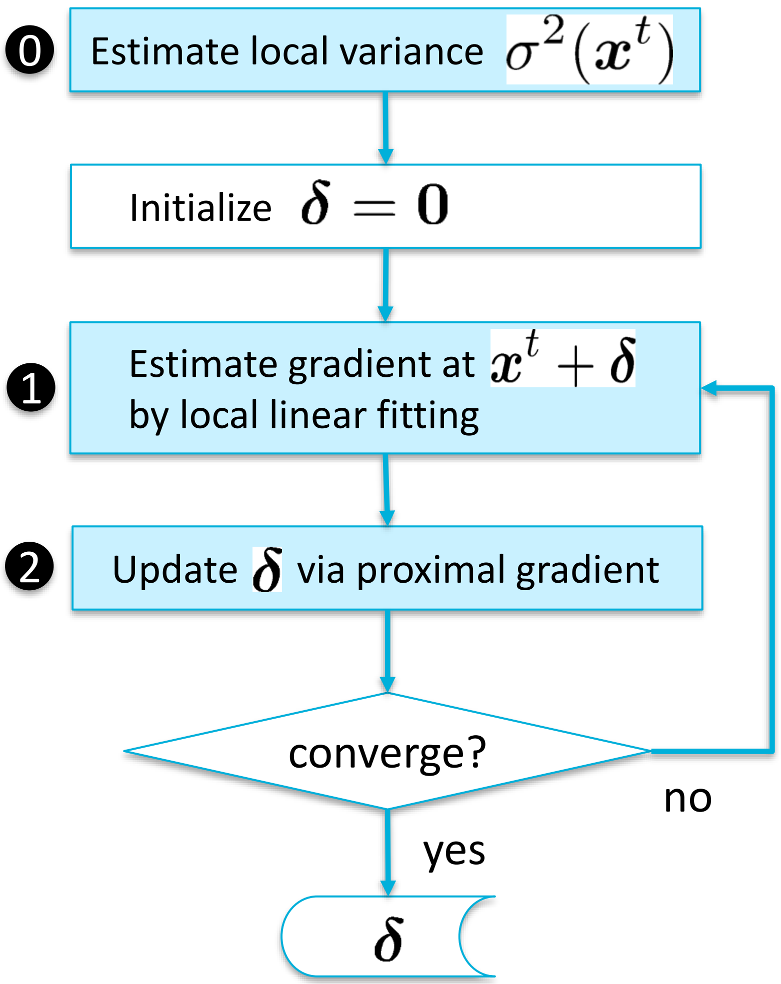
- *f*(***x***) is black-box. It may not be even smooth or continuous

- 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 $\boldsymbol{\delta}$

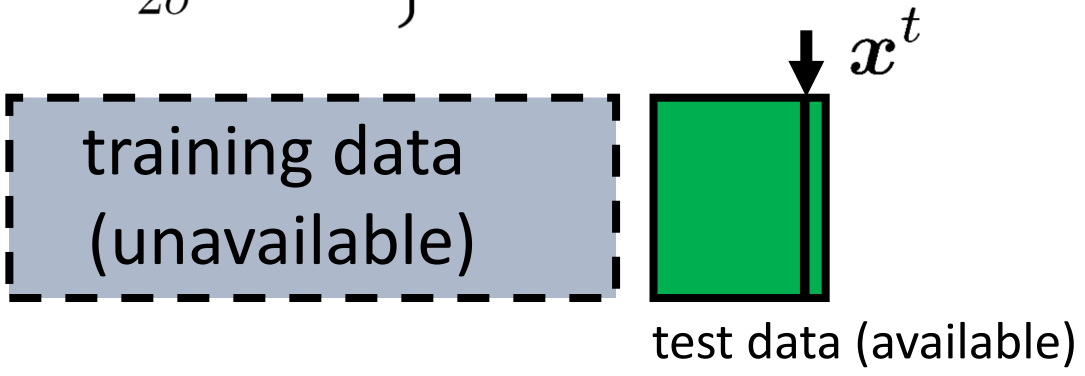
iterate



0. Local variance estimation

- If available test samples 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
  - $$\sigma^2(\boldsymbol{x}^t) = \max_{\sigma^2} \sum_{n=1}^{N_{\mathrm{heldout}}} w_n(\boldsymbol{x}^t) \left\{ \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y^{(n)} - f(\boldsymbol{x}^{(n)}))^2}{2\sigma^2} \right\},$$

Gaussian kernel defined for the specific test sample $\boldsymbol{x}^t$



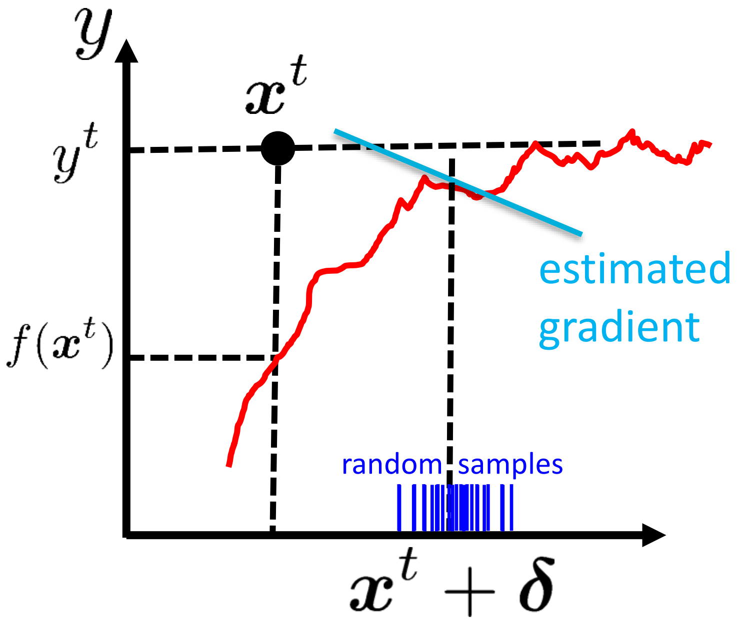
1. Local gradient estimation of *f*

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

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

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

convex-*ish* function with the smoothed gradient

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

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

smooth quadratic approximation of *J*

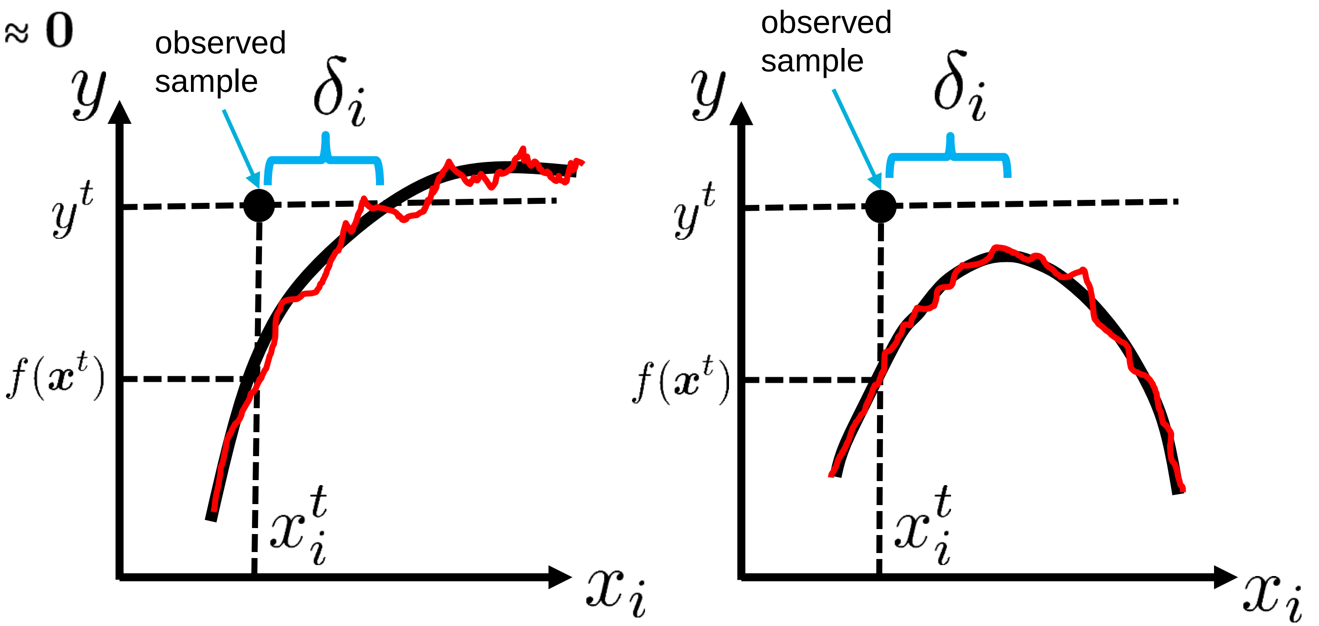
$$= \mathrm{prox}_{\kappa\nu\|\boldsymbol{\delta}\|_1} \left(\boldsymbol{\delta}^{\mathrm{old}} - \kappa \langle\!\langle \nabla J(\boldsymbol{\delta}^{\mathrm{old}}) \rangle\!\rangle \right)$$

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 $N_{\text{test}} = 1$



(a) |deviation| = 0 (b) |gradient| = 0

Contents

Existing methods for anomaly attribution

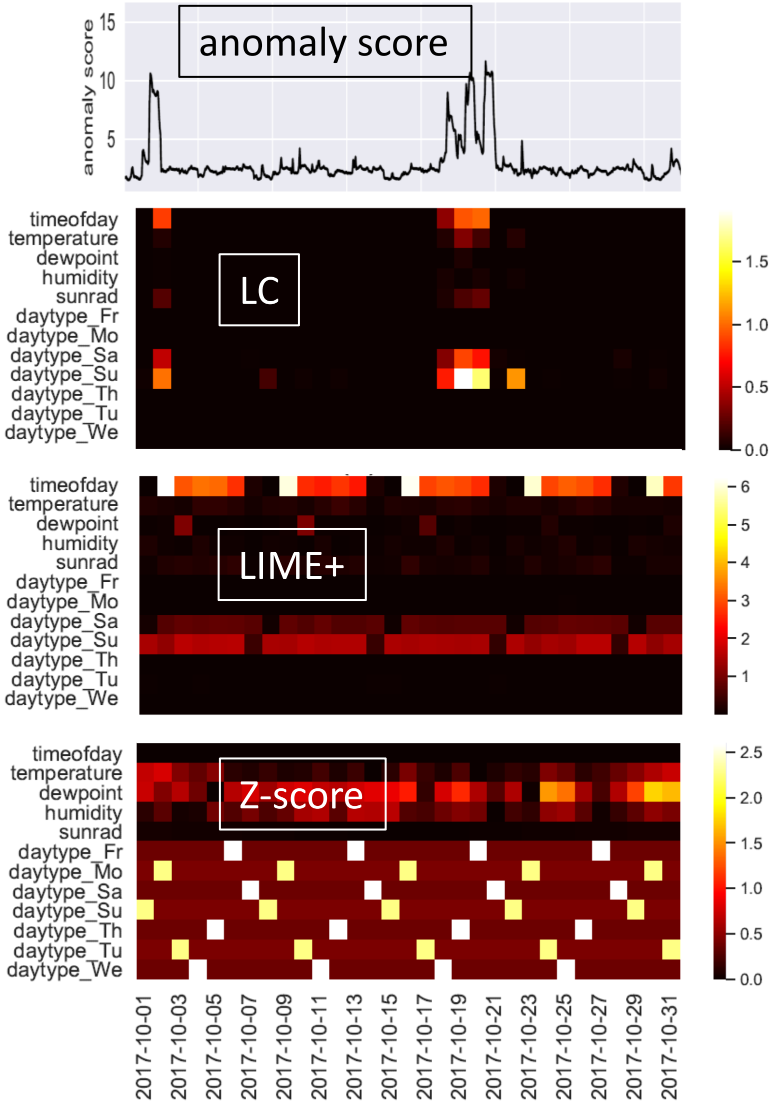
- Few applicable approaches in the regression setting
  - Most methods are for classification (especially for images)
  - Very limited choice for explaining deviations/anomalies in the black-box regression setting
- Possible baselines
  - Z-score
    - ✓ $a_i(\boldsymbol{x}^t) = (x_i^t - \mathtt{mean}_i)/\mathtt{stddev}_i$
    - ✓ Does not depend on y^t

- LIME* [Ribeiro 18], extended
  - ✓ Sampling-based local lasso fitting for of $f(\boldsymbol{x}^t) - y^t$ rather than $f(\boldsymbol{x}^t)$
    - To be able to explain deviations
  - ✓ Regression coefficient (≒ gradient) is the score
- Shapley value [Strumbelj+ 14], extended
  - ✓ SV computed for $f(\boldsymbol{x}^t) - y^t$
  - ✓ Requires the true distribution for $\boldsymbol{x}$ or the training data set to evaluate conditional local means

* Local Interpretable Model-agnostic Explanations

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

- Problem setting
- Introducing *Likelihood Compensation*
- Experimental results
- Summary

Summary

- LC is a principled framework designed to explain deviations from black-box regression function

- We empirically showed that LIME and Shapley values are insensitive to deviations
  - Is there any theoretical justification on this ? --- Yes.

Backup

Anomaly attribution as inverse problem of anomaly detection

- This is a statistical inverse problem
  - Forward problem: Given ($\boldsymbol{x}^t$, y^t), tell whether it is anomalous
    - ✓ Simple: Just check the amount of deviation $|f(\boldsymbol{x}) - y|$ to see if it is too big
  - Inverse is challenging: Quantify how each of $\mathbf{x}$ contributes to a large $|f(\boldsymbol{x}) - y|$

- Existing black-box explainability methods are not directly applicable
  - They are either:
  - (1) designed specifically for (image) classification, or
  - (2) focused only on characterizing $f(\mathbf{x})$, not the deviation between y and f,
    - ✓ We are interested in anomaly diagnosis
    - ✓ Anomalies are defined by large $|f(\boldsymbol{x}) - y|$ values, not $f(\boldsymbol{x})$ alone

Summarizing practical features of LC

- LC is directly interpretable (c.f. LIME)
  - It is defined as the amount of correction required to fit the observed y^t
  - 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"

- LC is model-agnostic
  - c.f. most of existing anomaly diagnosis methods, which assume full access to the model

- LC can characterize $f(\boldsymbol{x}^t) - y^t$, thus can produce outlier-specific explanations

(For ref.) Algorithm for LIME+ and SV+

- LIME+ (extended LIME)
  - For a given test sample ($\boldsymbol{x}^t$,y^t), populate N_s samples around $\boldsymbol{x}^t$ as { $\boldsymbol{x}^{t[1]}$, …, $\boldsymbol{x}^{t[Ns]}$ }
  - Create a data set D^t = { ($z^{t[1]}$, $\mathbf{x}^{t[1]}$), …, ($z^{t[Ns]}$, $\mathbf{x}^{t[Ns]}$) }, where $z^{t[n]} = f(\boldsymbol{x}^{t[n]}) - y^t$
  - Fit lasso regression to the data
  - Your explainability score is the regression coefficients
- SV+ (extended Shapley value)
  - For a given test sample ($\boldsymbol{x}^t$,y^t), the SV+ score for the j-th variable is

$$\mathtt{SV}_j(\boldsymbol{x}^t) \triangleq \sum_{|\mathcal{S}_j|=0}^{M-1} \frac{(M-|\mathcal{S}_j|-1)!\,|\mathcal{S}_j|!}{M!}\left[\langle f-y^t \mid x_j = x_j^t, \boldsymbol{x}_{\mathcal{S}_j} = \boldsymbol{x}_{\mathcal{S}_j}^t\rangle - \langle f-y^t \mid \boldsymbol{x}_{\mathcal{S}_j} = \boldsymbol{x}_{\mathcal{S}_j}^t\rangle\right]$$

✓ where S_j is the set of all the variable indices excluding j, and

✓ for an M-variate function g, $\langle g \mid x_j = x_j^t, \boldsymbol{x}_{\mathcal{S}_j} = \boldsymbol{x}_{\mathcal{S}_j}^t\rangle \triangleq \int \mathrm{d}\boldsymbol{x}\; P(\boldsymbol{x}) g(x_j = x_j^t, \boldsymbol{x}_{\mathcal{S}_j} = \boldsymbol{x}_{\mathcal{S}_j}^t, \boldsymbol{x}_{\bar{\mathcal{S}}_j})$

$$\langle g \mid \boldsymbol{x}_{\mathcal{S}_j} = \boldsymbol{x}_{\mathcal{S}_j}^t\rangle \triangleq \int \mathrm{d}\boldsymbol{x}\; P(\boldsymbol{x}) g(x_j, \boldsymbol{x}_{\mathcal{S}_j} = \boldsymbol{x}_{\mathcal{S}_j}^t, \boldsymbol{x}_{\bar{\mathcal{S}}_j})$$

$P(\boldsymbol{x})$: true (or empirical) distribution (problematic)