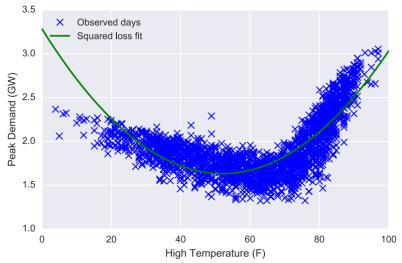
Big Data Technology and its Applications



Linear / Nonlinear Regression

张宁 ningzhang@tsinghua.edu.cn

Basic concepts of machine learning

Feature vs. Label

- The features are the descriptive attributes, and the label is what you're attempting to predict or forecast.
- Given: Training data: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, $\boldsymbol{x}_i \in \mathbb{R}^m$ is the feature and y_i is the label.

example $x_1 \rightarrow$	x_{11}	x_{12}	$\ldots$	x_{1m}	$y_1 \leftarrow$ label
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
example $x_i \rightarrow$	x_{i1}	x_{i2}	$\ldots$	x_{im}	$y_i \leftarrow$ label
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
example $x_n \rightarrow$	x_{n1}	x_{n2}	$\ldots$	x_{nm}	$y_n \leftarrow$ label

n : Number of training examples

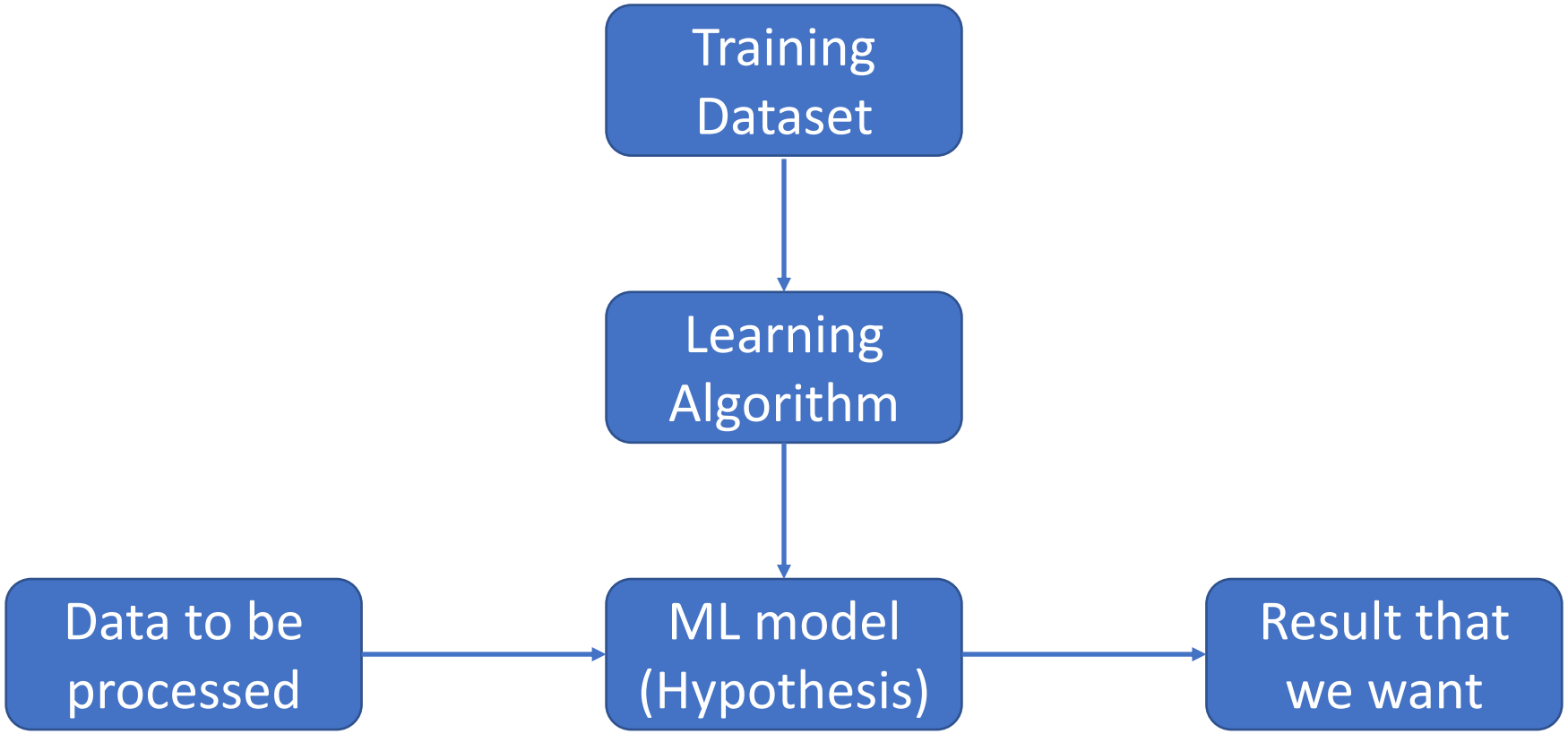
m : Number of features

- Example:

features

fruit	length	width	weight	label
fruit 1	165	38	172	Banana
fruit 2	218	39	230	Banana
fruit 3	76	80	145	Orange
fruit 4	145	35	150	Banana
fruit 5	90	88	160	Orange
...				
fruit n	...	...	...	...

Machine learning



Supervised vs. Unsupervised

- **Unsupervised learning:**
  Learning a model from un**labeled** data, e.g. predicting the category of fruit only using $\boldsymbol{x}_i = [length, width, weight]$, without knowing the specific label

- **Supervised learning:**
  Learning a model from **labeled** data. e.g. predicting the category of fruit using $(\boldsymbol{x}_i = [length, width, weight], y_i = label)$

-

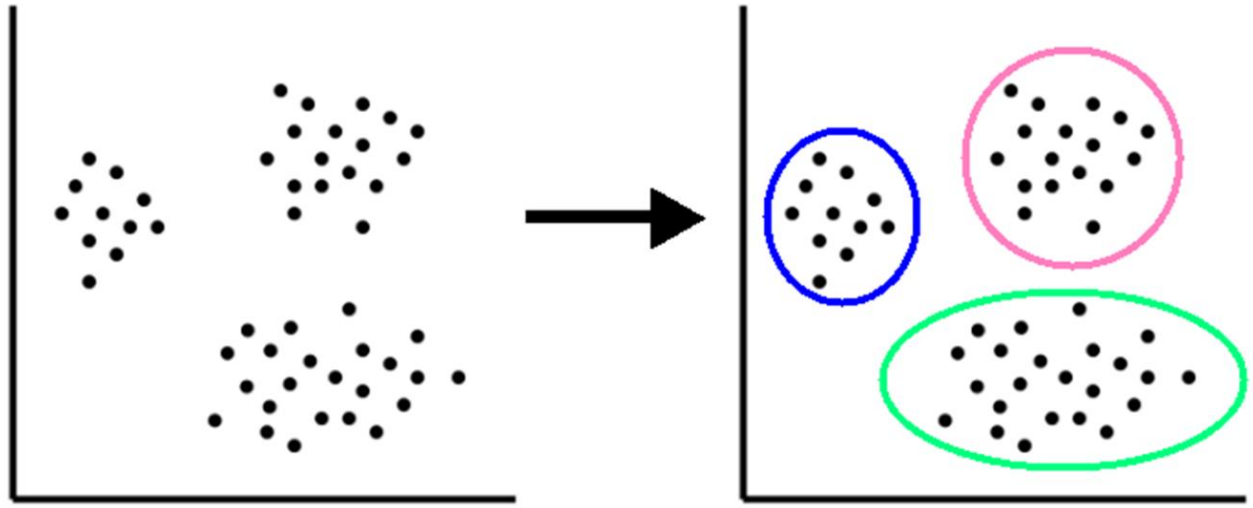
fruit	length	width	weight	label
fruit 1	165	38	172	Banana
fruit 2	218	39	230	Banana
fruit 3	76	80	145	Orange
fruit 4	145	35	150	Banana
fruit 5	90	88	160	Orange
…				
fruit n	…	…	…	…

Unsupervised learning

- **Clustering/segmentation**:

$$f : R^m \rightarrow \{C_1, \ldots, C_k\}$$

set of clusters



- Methods: K-means, Gaussian mixtures, hierarchical clustering, spectral clustering, etc.

Unsupervised learning problem in power systems

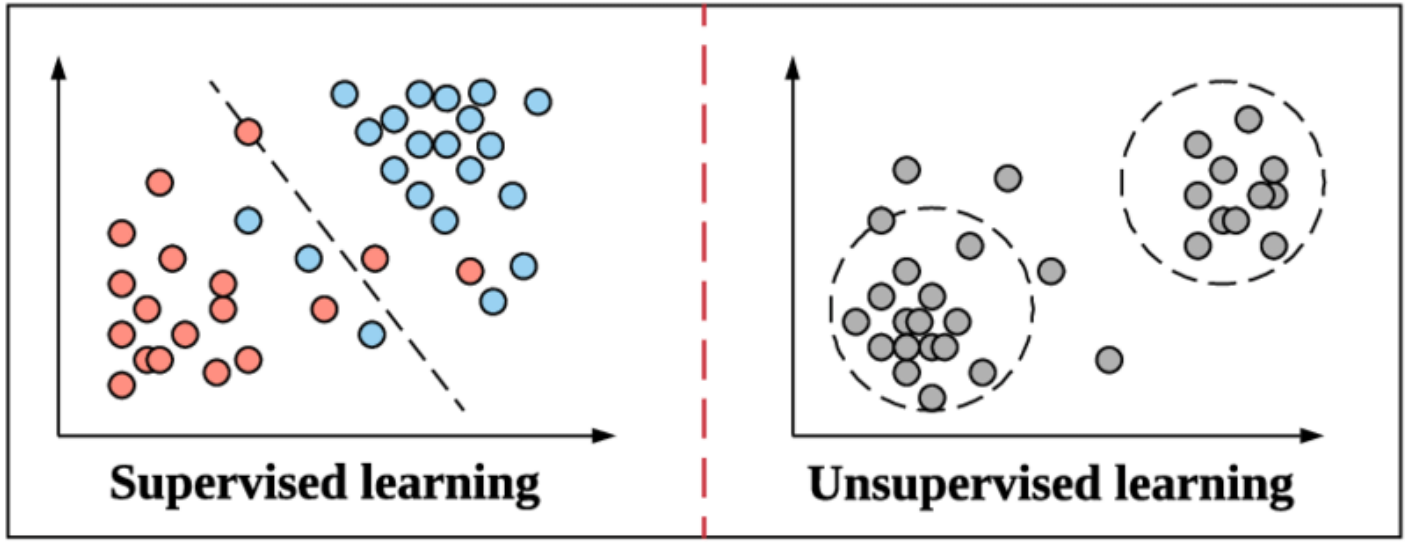
- Power system operation mode analysis / visualization
- Anomaly detection for power equipment
- Anomaly detection for energy consumers (energy theft, elderly care)
- Power system data generation
- Anomaly detection for power market

Supervised learning

(1) Classification: y is discrete. To simplify, $y \in \{-1, +1\}$

$$f : R^m \rightarrow \{-1, +1\} \qquad f \text{ is called a } \textbf{binary classifier}.$$

- Example: Approve credit yes/no, spam/ham, banana/orange.



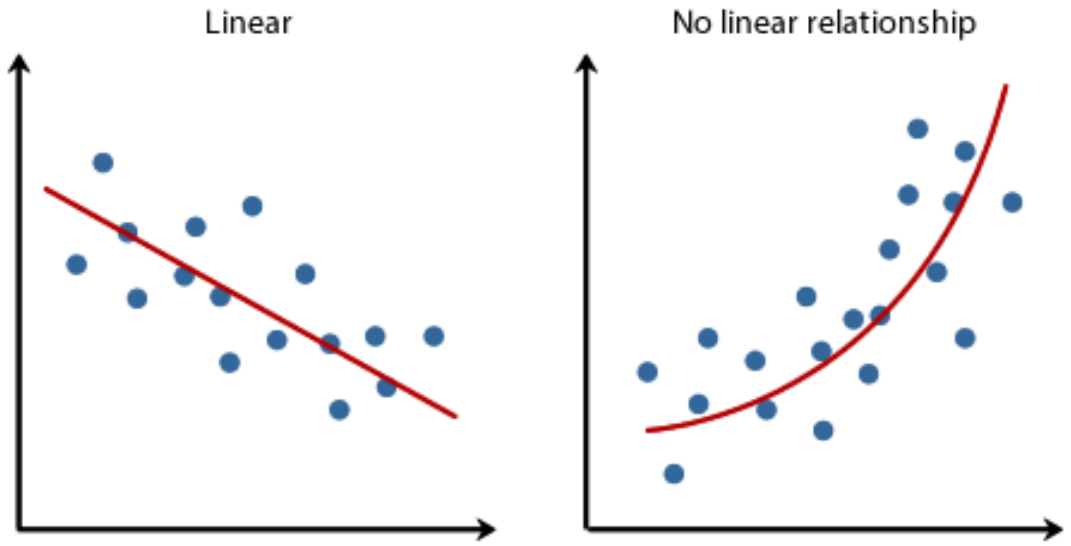
- **Methods:** Support Vector Machines, neural networks, decision trees, K-nearest neighbors, naive Bayes, etc.

Supervised learning

(2) Regression: y is a real value

$$f : R^m \rightarrow R \qquad\qquad f \text{ is called a } \textbf{regressor}$$

- Example: electricity demand, weight of fruit.

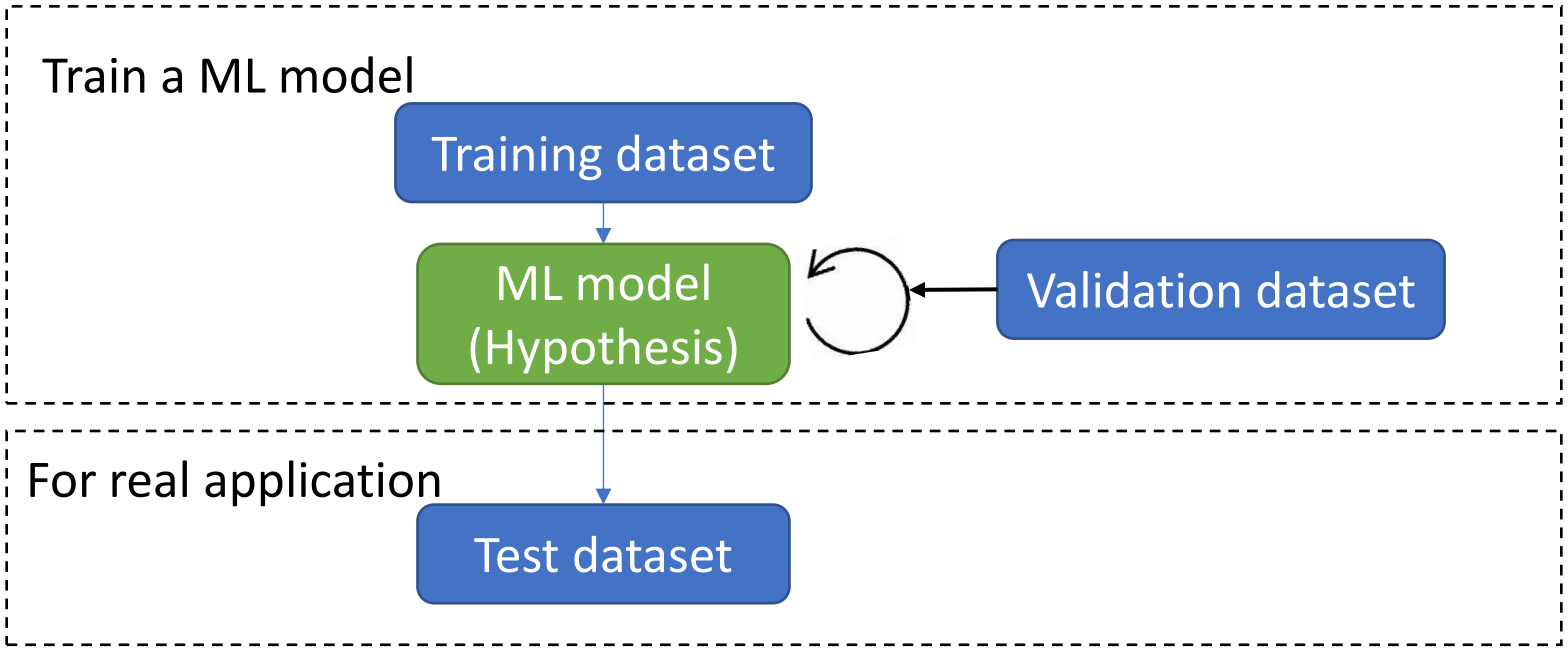


- **Methods:** Linear/nonlinear regression, Lasso, Support Vector Machines, neural networks, decision trees, etc.

Supervised learning problem in power systems

- Load / renewable energy forecasting
- Power flow calculation
- Power system security / stability assessment
- Power system dispatch / control optimization
- Topology identification
- Fault detection / classification
- Fault diagnosis of power equipment
- Power market simulation

Training, Validation, and Testing



- **A training dataset** is a dataset of examples used for learning, that is to fit the parameters of ML model.
- **A validation dataset** is a dataset of examples used to tune the hyperparameters (i.e. the architecture) of ML model
- **A test dataset** is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset.

Model Evaluation Metrics

- **For regression task,** we usually use mean-square error (MSE) or root-mean-square error (RMSE) to evaluate the model.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **For classification task,** the case is more complex.

Confusion Matrix (混淆矩阵):

		Actual Label	
		Positive	Negative
Predicted Label	Positive	**True Positive (TP)**	**False Positive (FP)**
	Negative	**False Negative (FN)**	**True Negative (TN)**

Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that are correct
Precision	TP / (TP + FP)	The percentage of positive predictions that are correct
Sensitivity (Recall)	TP / (TP + FN)	The percentage of positive cases that were predicted as positive
Specificity	TN / (TN + FP)	The percentage of negative cases that were predicted as negative

Ref: https://blog.csdn.net/anliu1/article/details/112794538

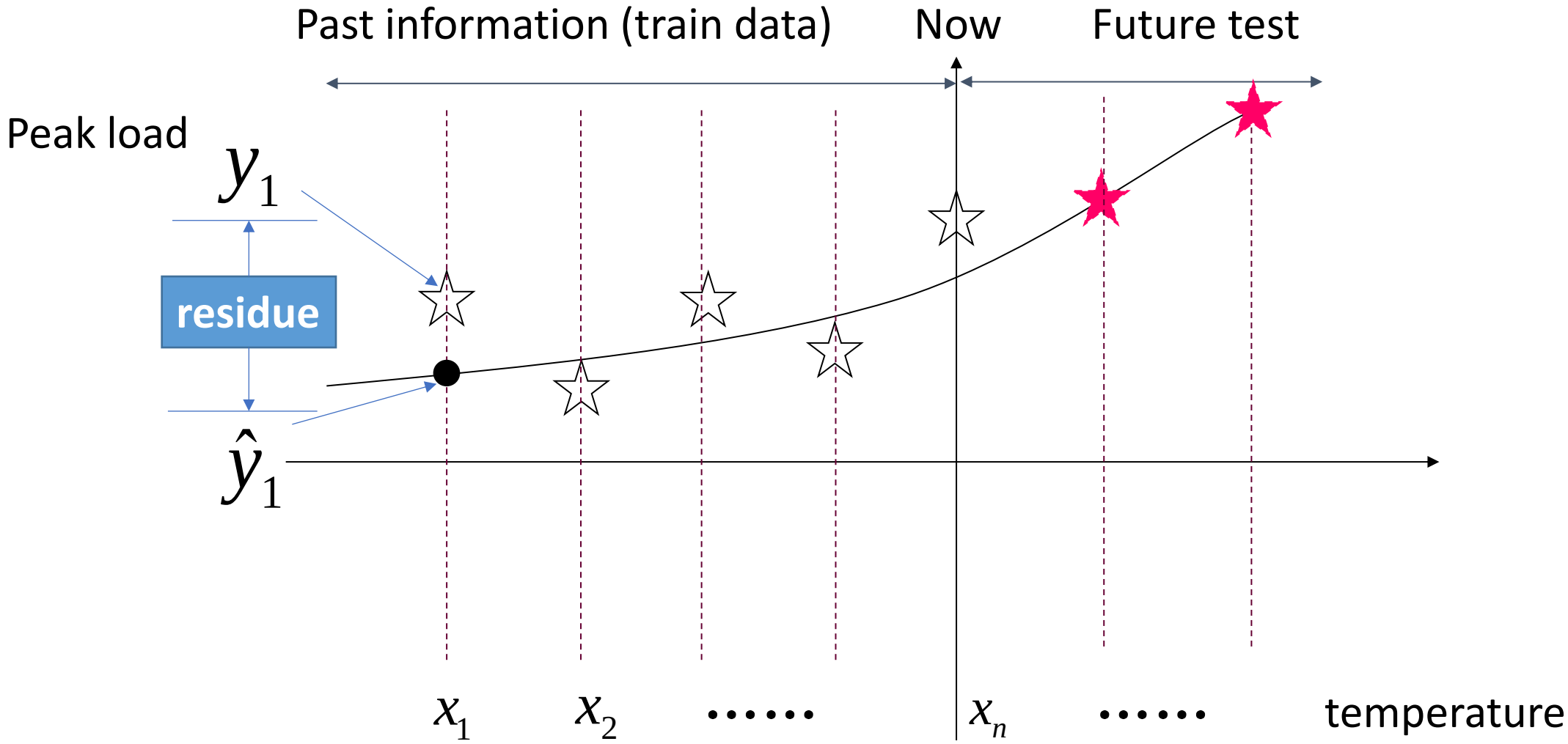
Linear Regression

Example: Predicting Electricity Demand

- Problem: What will the peak power consumption be tomorrow?
- Challenge: Difficult to build an a priori model from first principles (electric laws) to answer this question
- Idea: Easy to record past days of consumption, plus additional features that affect consumption (i.e., weather, weekdays)

Date	High Temperature (F)	Peak Demand (GW)
2011-06-01	84.0	2.651
2011-06-02	73.0	2.081
2011-06-03	75.2	1.844
2011-06-04	84.9	1.959
	…	…

Example: Predicting Electricity Demand



Univariate Linear Regression

- The objective in unary linear regression is to find the optimal parameter $S = (a, b)$ that minimize :

$$\min_{a,b} \sum_{i=1}^{n} \left(\hat{y}_i - ax_i - b \right)^2$$

- and use the model $f(x, S) = ax + b$ to predict the y_i in the future

What is Known:

- Independent variable x in the past: $x_1, x_2, \cdots, x_n$
- Dependent variable y in the past: $y_1, y_2, \cdots, y_n$
- Independent variable x in the future time: $x_{n+1}, x_{n+2}, \cdots, x_N$

- Question: What is the type of this optimization?

Univariate Linear Regression

- How to find the optimal parameter $S = (a, b)$?
- The sum of residues is:

$$Q = \sum_{i=1}^{n} v_i^{\,2} = \sum_{i=1}^{n} \left[y_i - (a \cdot x_i + b) \right]^2$$

- The function reaches the minimum when directives equal zero.

$$\frac{\partial Q}{\partial a} = -2 \sum_{i=1}^{n} x_i (y_i - a x_i - b) = 0$$

$$\frac{\partial Q}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a x_i - b) = 0$$

- The solution is:

$$a = \left[\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) \right] \Big/ \left[\sum_{i=1}^{n} (x_i - \overline{x})^2 \right] \qquad b = \overline{y} - a \cdot \overline{x}$$

- Where

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Multiple Linear Regression

- The objective in multiple linear regression is to find the optimal parameter $S \in \mathbb{R}^m, (s_0, s_1, s_2, \dots, s_m)$ that minimize :

$$\min_S \sum_{i=1}^{n} \left(\hat{y}_i - Sx \right)^2$$

where m is the dimension of feature x.

- For example, in the problem of predicting electric demand. If we only use the measure of temperature at one location to predict, the problem is unary linear regression. If we use the measure of data at multiple locations, the problem is multiple linear regression.

- Question: How to get the best parameters of multiple linear regression?

Multiple Linear Regression

- Introduce the matrix and vectors:

$$\boldsymbol{A}=\begin{bmatrix}1 & x_{11} & \cdots & x_{m1}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{1n} & \cdots & x_{mn}\end{bmatrix} \qquad \boldsymbol{Y}=\left[y_1,y_2,\cdots,y_n\right]^T$$

$$\boldsymbol{S}=\left[s_0,s_1,\cdots,s_m\right]^T$$

- The model can be represented as:

$$y=f(\boldsymbol{S},\boldsymbol{X})=s_0+\sum_{i=1}^{m}s_i\cdot x_i \Rightarrow \boldsymbol{Y}=\boldsymbol{AS}$$

- The sum of residue is:

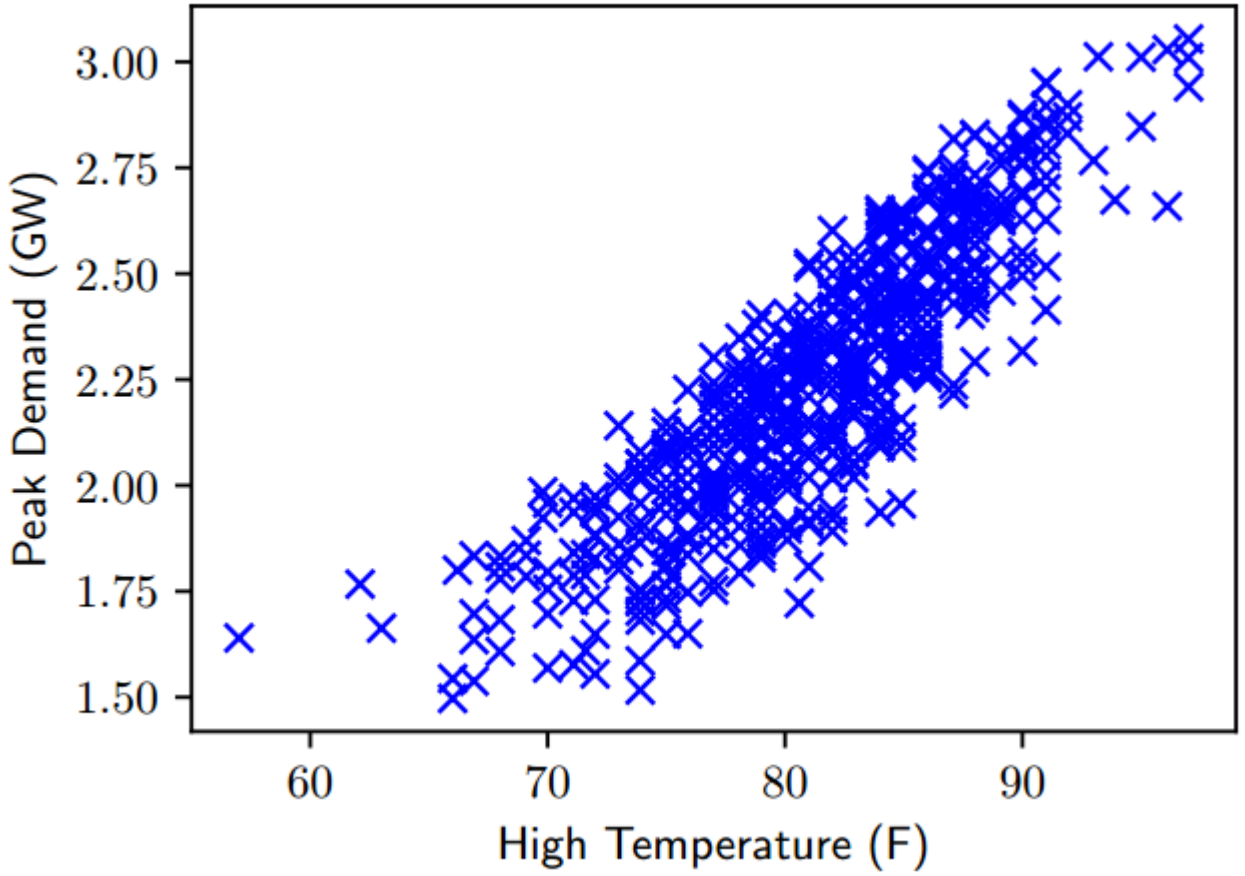
$$Q=\sum_{t=1}^{n}v_t^2=(\boldsymbol{AS}-\boldsymbol{Y})^T(\boldsymbol{AS}-\boldsymbol{Y})=\left\|\boldsymbol{AS}-\boldsymbol{Y}\right\|_2^2$$

- Let the derivative to be zero and get:

$$\frac{\partial Q}{\partial \mathbf{S}^T}=0 \quad \Rightarrow \quad \boldsymbol{S}=(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Y}$$

Example: Predicting Electricity Demand

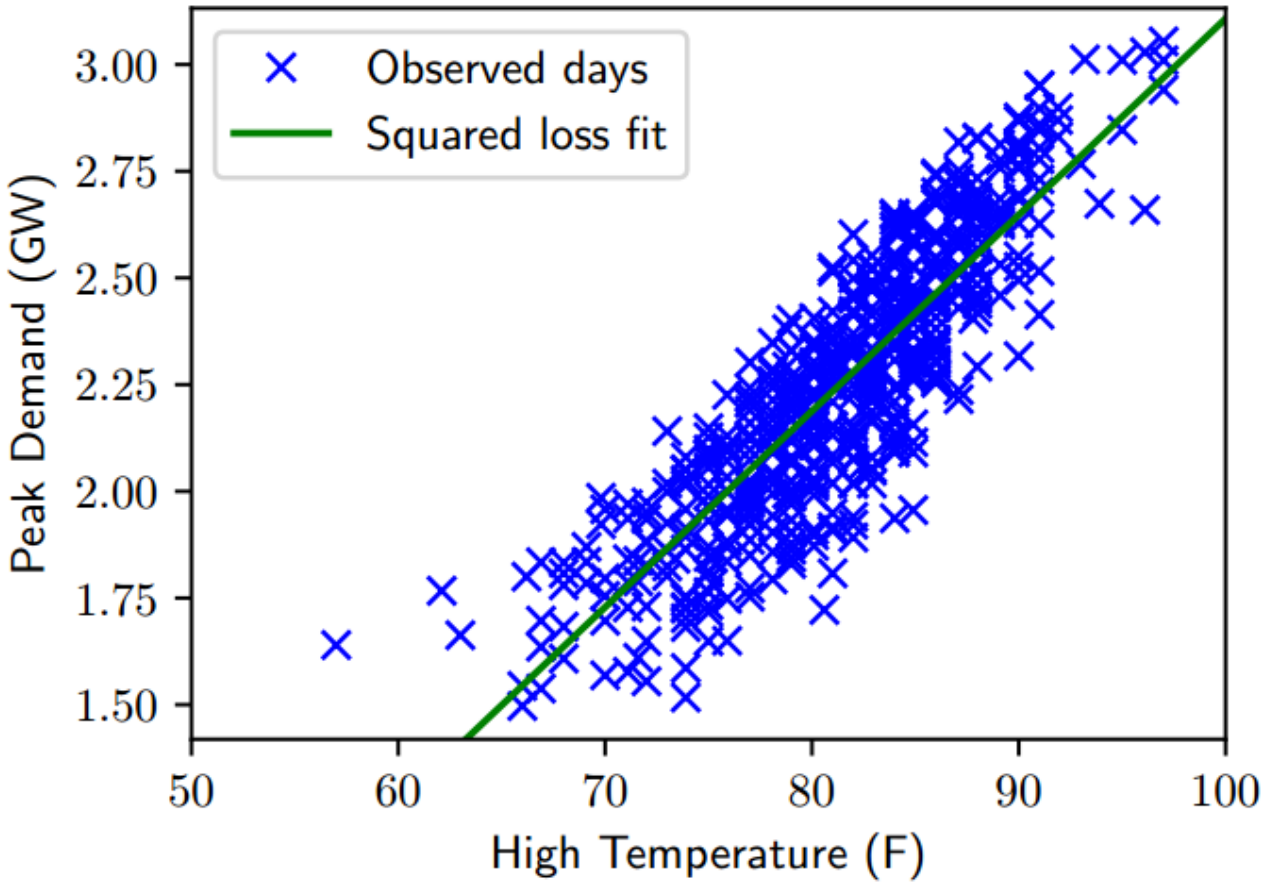
- Plot of high temperature vs. peak demand
  - **For summer months (June – August)** for past six years
  - Hypothesis: Linear Model



Example: Predicting Electricity Demand

- Unary linear regression: using the measurement of temperature at one location.

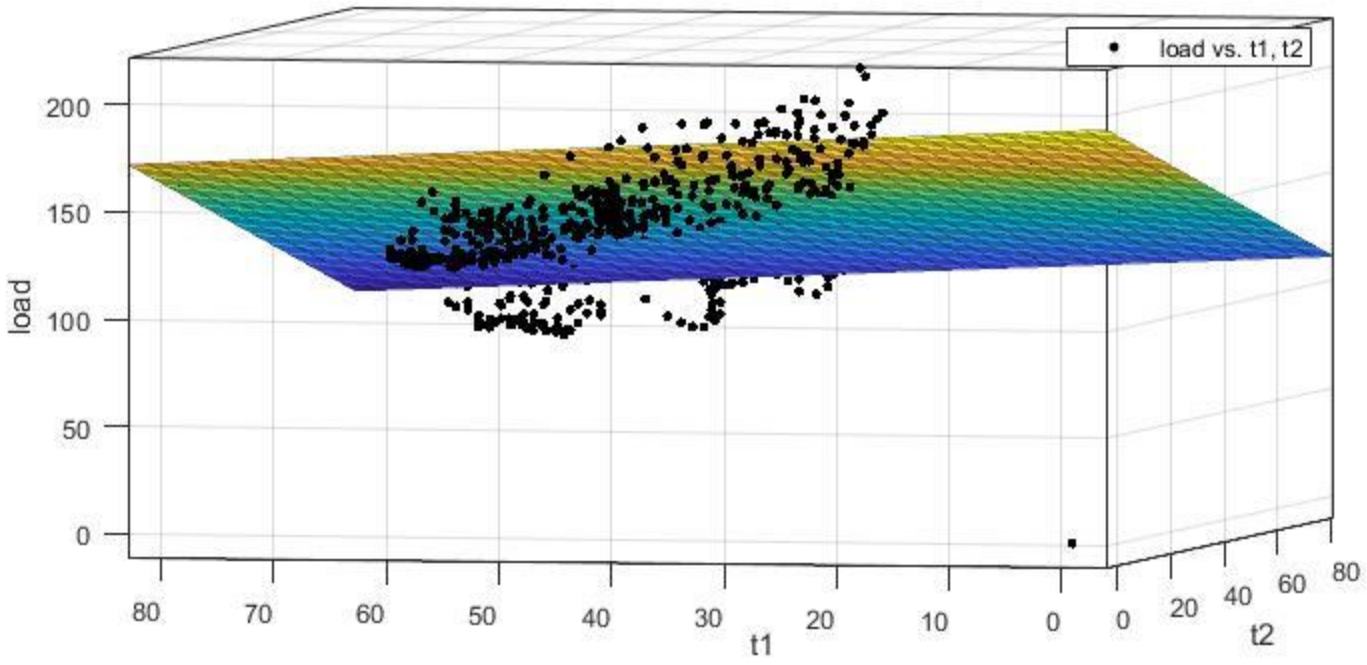
$$D = s_0 + s_1 T$$



Example: Predicting Electricity Demand

- Multiple linear regression: using the measurement of temperature at two different locations.

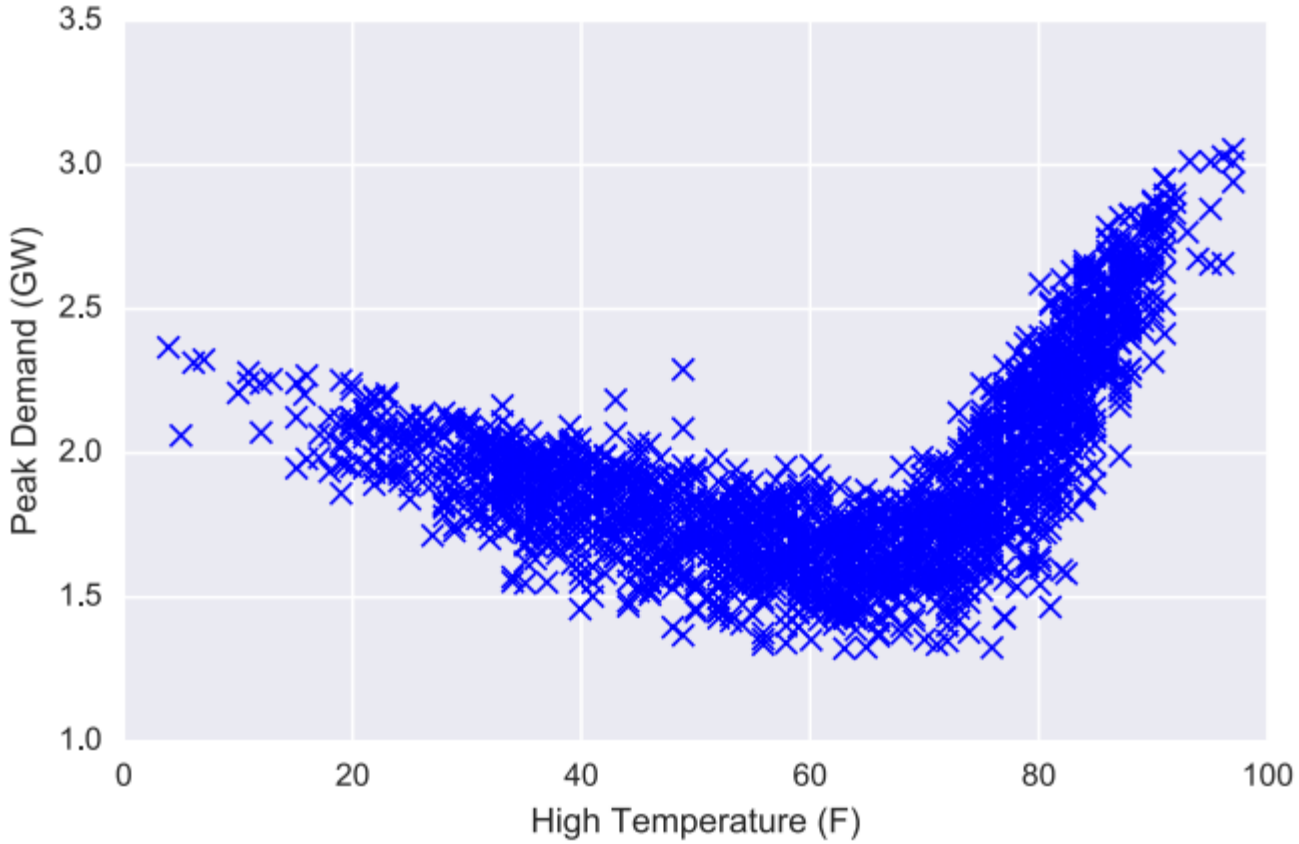
$$D = s_0 + s_1 T_1 + s_2 T_2$$



Nonlinear Regression

Example: Predicting Electricity Demand

- Plot of high temperature vs. peak demand
  - **All months** for past six years
  - Nonlinear Model



Nonlinear Regression

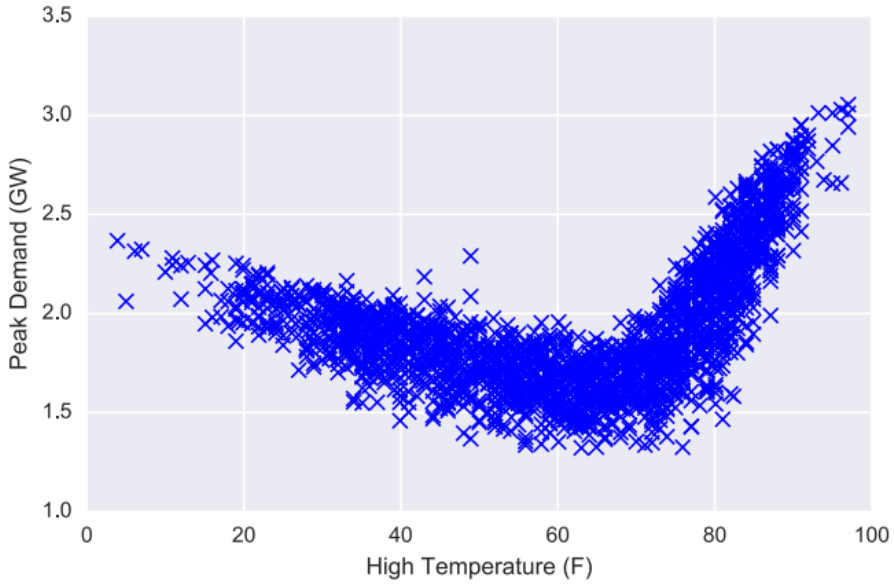
- Nonlinear regression is easy to transform to linear regression and is thus called generalized linear regression.
- For example, if we want to use quadratic polynomial to predict the power demand, we can simply let $x = T^2$ and make linear regression on the variable x and T.

Nonlinear regarding to T

$$D = s_0 + s_1 T + s_2 T^2$$

Linear regarding to T and x

$$D = s_0 + s_1 T + s_2 x$$



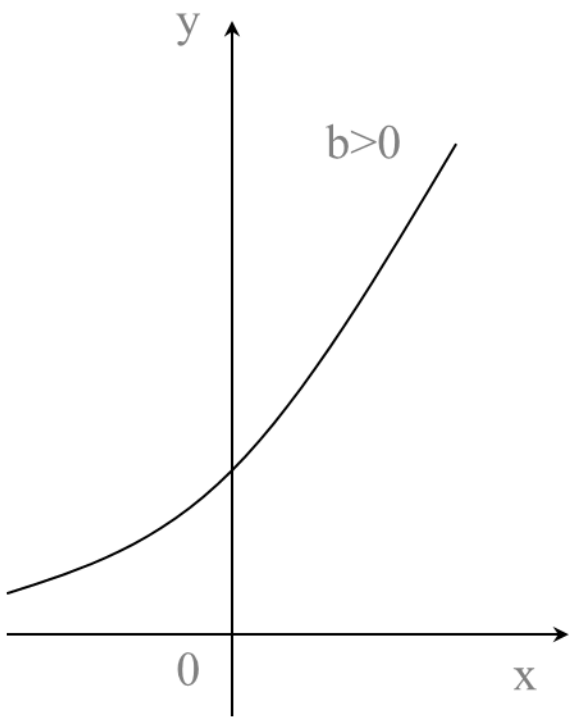
Nonlinear Regression

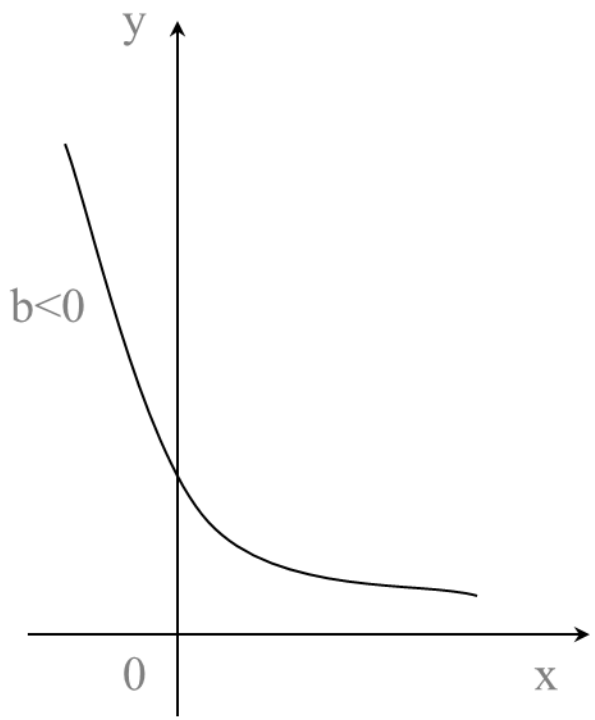
Exponential model

$$y = ae^{bx} \ (a > 0)$$

Variable substitution $y' = \ln(y)$

Corresponding linear model $y' = a' + bx$ $a' = \ln(a)$





Nonlinear Regression

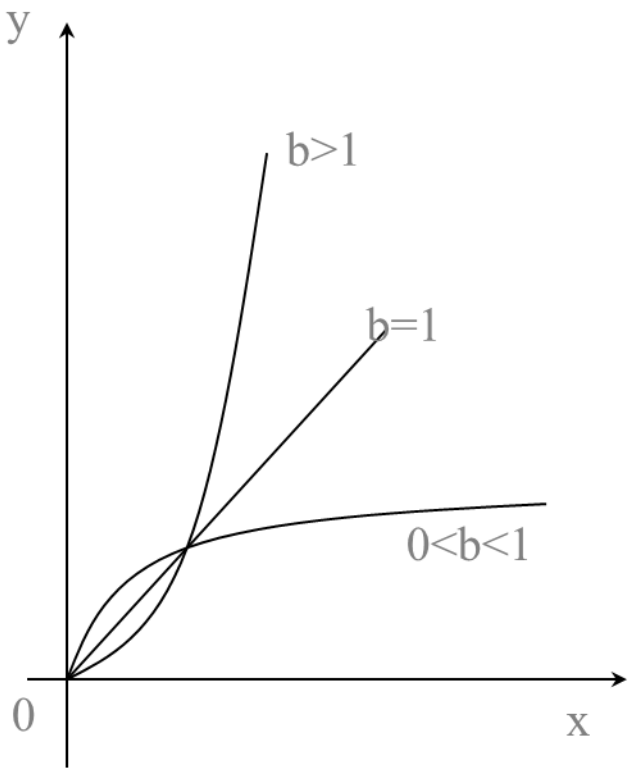
Power function model

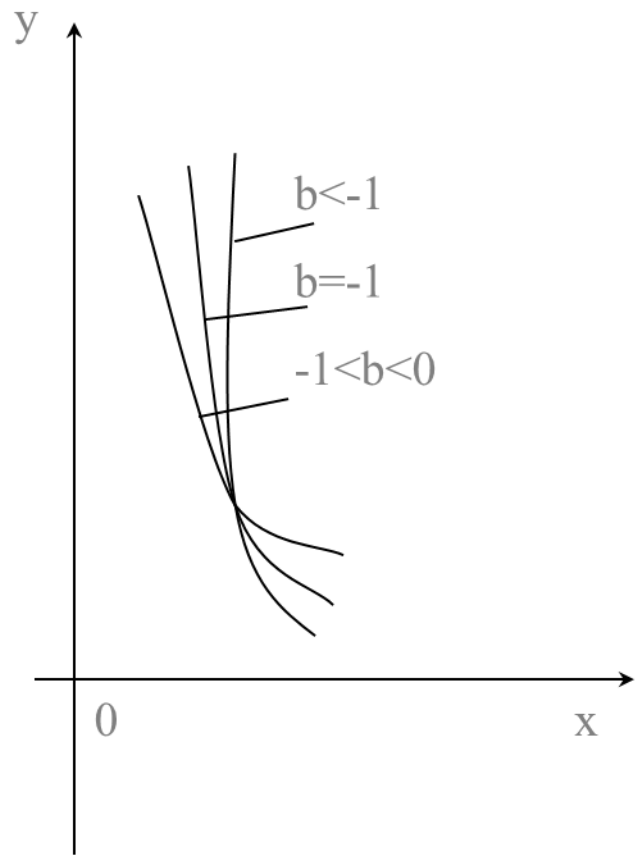
$$y = ax^b$$

$$(x > 0, a > 0)$$

Variable substitution $y' = \ln(y) \quad x' = \ln(x)$

Corresponding linear model $y' = a' + bx' \quad a' = \ln(a)$





Nonlinear Regression

S-function model

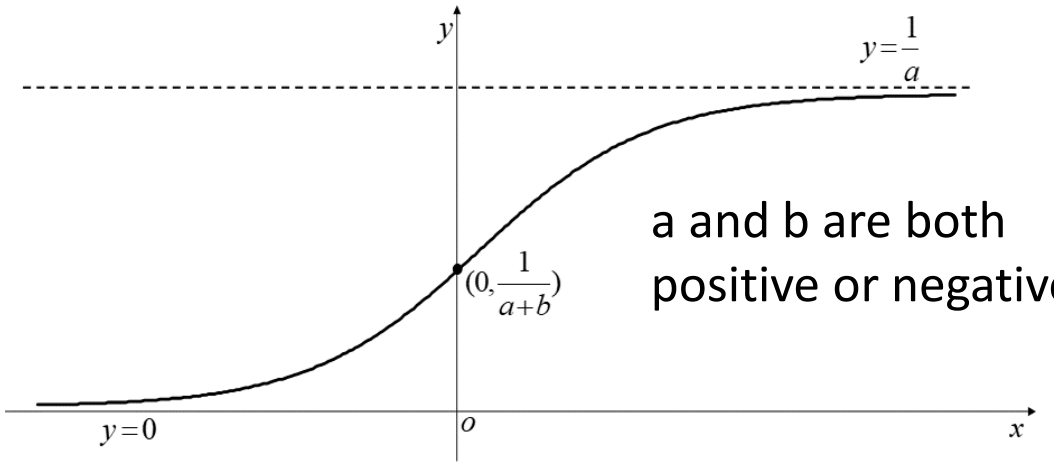
$$y = \frac{1}{a + be^{-x}}$$

Variable substitution

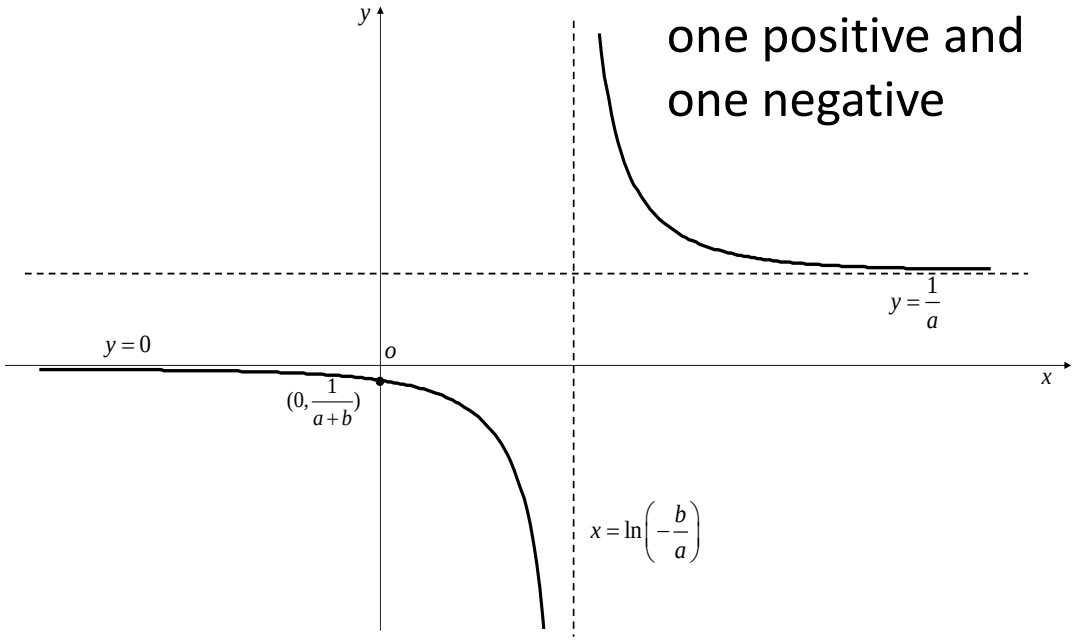
$$y' = \frac{1}{y}$$

Corresponding linear model

$$y' = a + b \cdot e^{-x}$$



a and b are both positive or negative



one positive and one negative

Finding the Best Parameters

- Apart from using the analytical solution for linear regression, we can also use gradient descent (GD) to find the best parameters in nonlinear regression.
- For example, the objective function after linear transformation is:

$$Q = \sum_{t=1}^{n} v_t^2 = (\boldsymbol{AS} - \boldsymbol{Y})^T (\boldsymbol{AS} - \boldsymbol{Y}) = \left\| \boldsymbol{AS} - \boldsymbol{Y} \right\|_2^2$$

- Then the parameters S can be updated as:

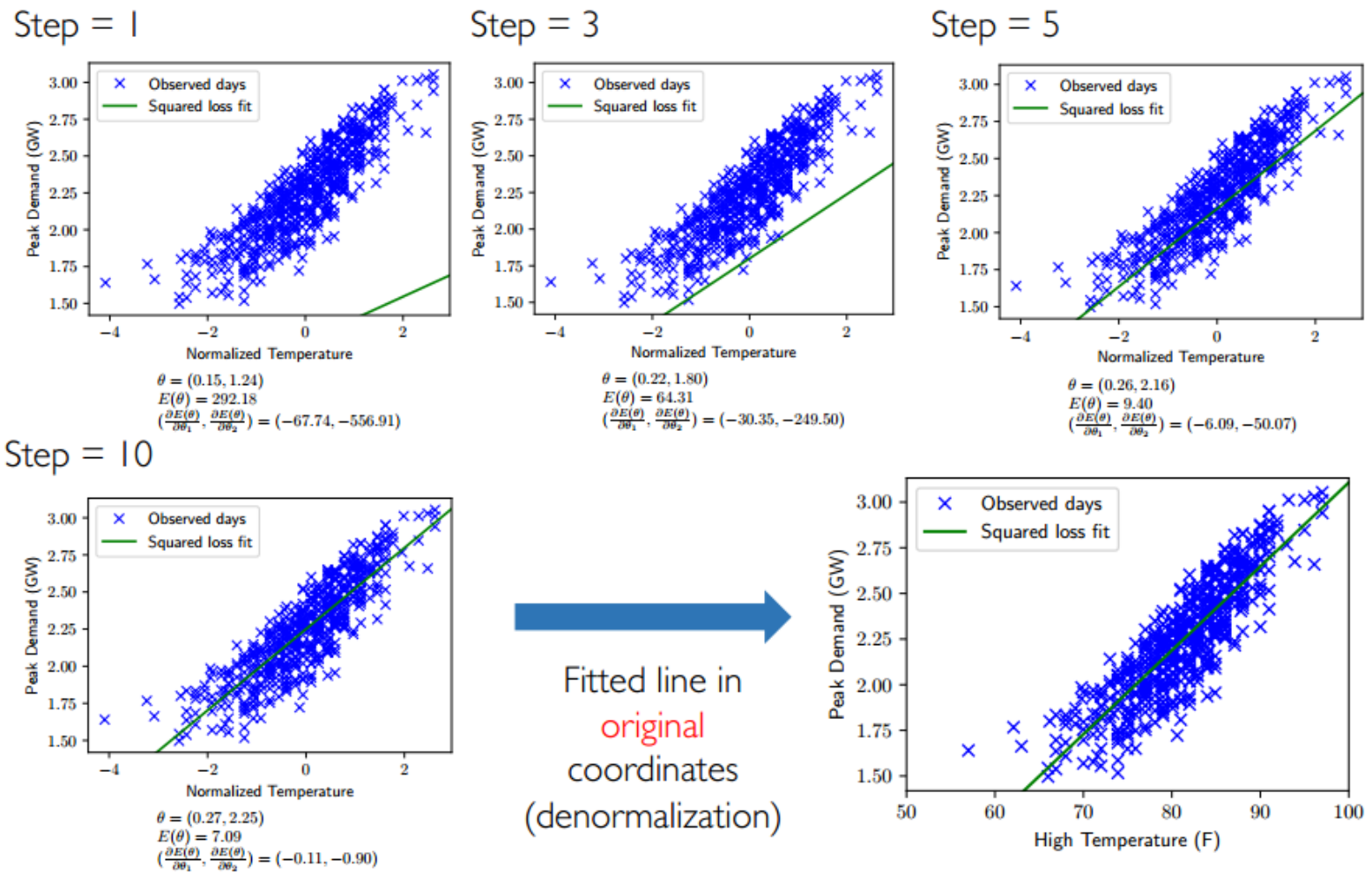
$$\mathbf{S}^{k+1} = \mathbf{S}^k - \alpha \frac{\partial Q^k}{\partial \mathbf{S}^T}$$

where α is a small positive number called step size (learning rate)

- This is the gradient decent (GD) algorithm, it is the workhorse of all modern machine learning.

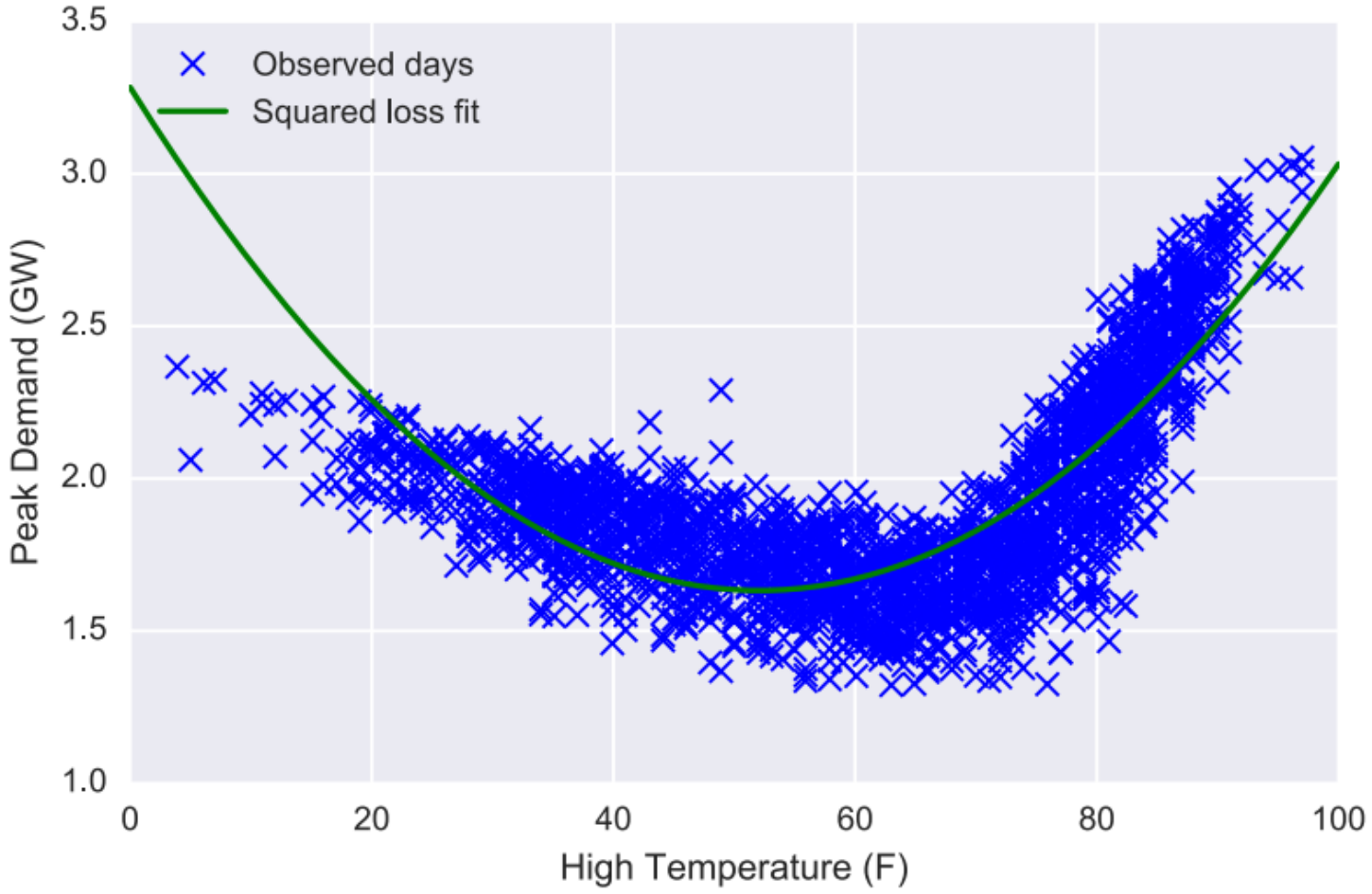
Finding the Best Parameters

- **Example of GD in first-order polynomial regression**



Example: Predicting Electricity Demand

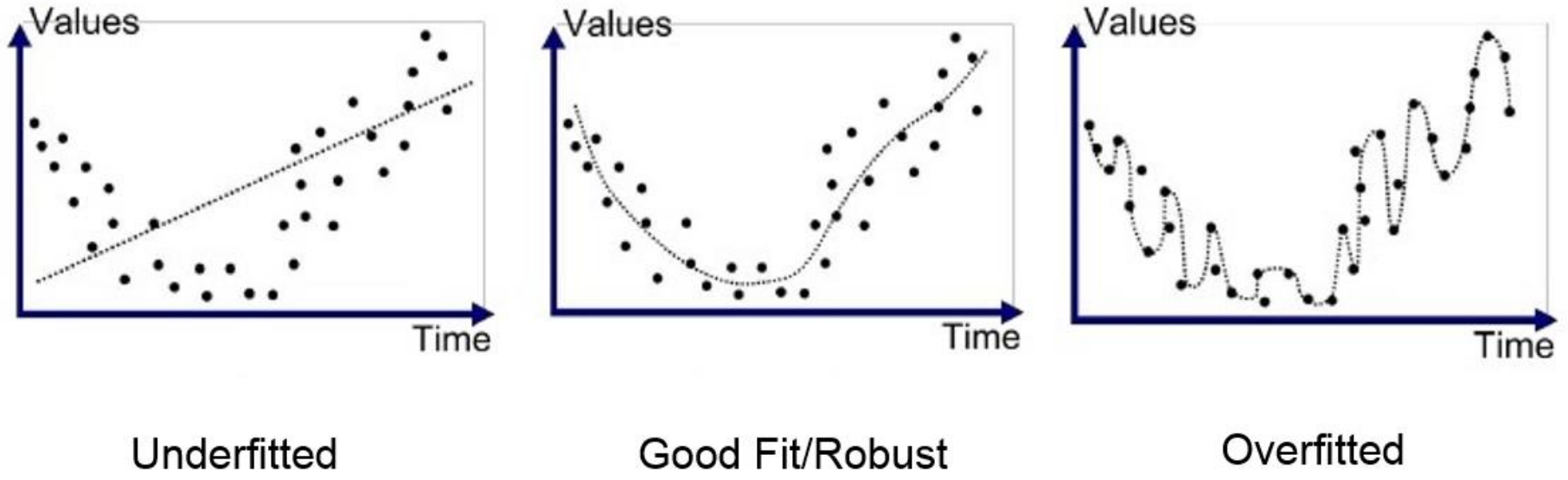
- Polynomial Features of Degree 2



Overfitting and underfitting

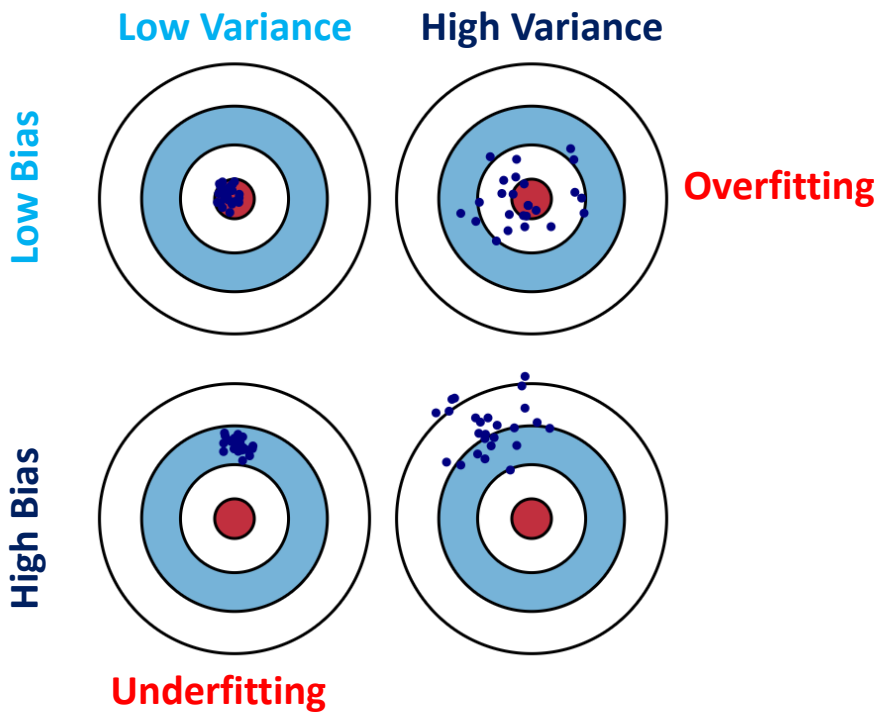
Model Generalization

- **Overfitting** refers to a model that was trained too much on the particulars of the training data (when the model learns the noise in the dataset). A model that is overfit will not perform well on new, unseen data.
- **Underfitting** typically refers to a model that has not been trained sufficiently. This could be due to insufficient training time or a model that was simply not trained properly.
- **Generalization** refers to the model's ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.



Bias and Variance

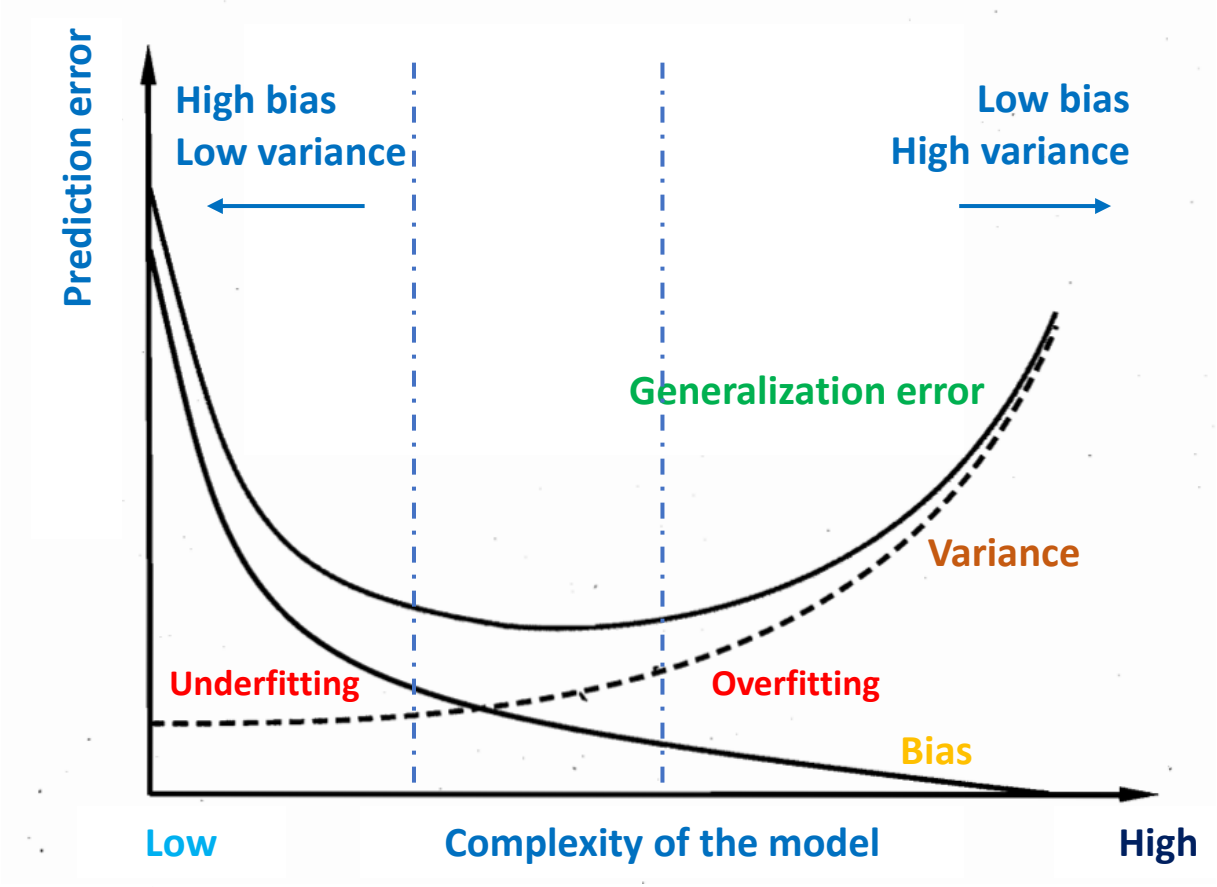
- **Bias:** the error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict.
- **Variance:** the error due to variance is taken as the variability of a model prediction for a given data point.
- **Generalization error** = Bias^2 + Variance + Noise



Ref: http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias and Variance

- **The bias–variance dilemma** is the property of a set of predictive models whereby models with a lower bias in parameter estimation have a higher variance of the parameter estimates across samples, and vice versa.



Question: How to identify / avoid overfitting?

Overfitting and Regularization

- Multiple linear regression may have serious problems of overfitting in case of large numbers of features or in case of multicollinearity.
- **Multicollinearity**: If two or more independent variables have an approximate linear relationship between them then we have multicollinearity.

$$y = s_0 + s_1 x_1 + s_2 x_2 + s_3 x_3 + \varepsilon$$

$$x_3 = a_0 + a_1 x_1 + a_2 x_2 + \delta$$

- With multicollinearity, the variances and the standard errors of the regression coefficient estimates will increase.
- Regression coefficients will be sensitive to specifications. Regression coefficients can change substantially when variables are added or dropped.
- Question: how to interpret? $\boldsymbol{S} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{Y}$

Overfitting and Regularization

- Consider a simple liner regression problem:

$$y = \{2, 4.1, 6, 8\}$$
$$x_1 = \{1, 2.01, 3, 4\}$$
$$x_2 = \{1, 2, 3, 4\}$$

- The “most fit” (minimum variance) formulation is:

$$y = 10x_1 - 8x_2$$

- However, if the feature has a small change:

$$y = \{2, 4.1, 6, 8\}$$
$$x_1 = \{1, 2.001, 3, 4\}$$
$$x_2 = \{1, 2, 3, 4\}$$

- The “most fit” (minimum variance) formulation turns to:

$$y = 100x_1 - 98x_2$$

- Regression coefficients will be sensitive to specifications.

$\boldsymbol{A}^T\boldsymbol{A}$ near singular matrix (ill condition)

Regularization

- A common approach to tackle multicollinearity is using $l1$-norm or $l2$-norm penalty on the coefficients. These two types of multiple linear regression called **Lasso** and **Ridge regression.**
- **Basic idea of Lasso**: Allow at most k coefficients to be nonzero.

$$\min \ \|\boldsymbol{AS}-\boldsymbol{Y}\|_2^2$$
$$s.t. \quad \|\boldsymbol{S}\|_0 \le k$$

- This problem is a Mixed-integer programming (NP-hard). Hard to solve for feature dimensions > 30.

- Convex relaxation: Lasso was proposed in 1996 by Tibshirani (24,000+ citations). Using $l1$-norm to approximate $l0$-norm.

$$\min \ \|\boldsymbol{AS}-\boldsymbol{Y}\|_2^2 \quad s.t. \quad \|\boldsymbol{S}\|_0 \le k \quad \Rightarrow \quad \min \ \|\boldsymbol{AS}-\boldsymbol{Y}\|_2^2 \quad s.t. \quad \|\boldsymbol{S}\|_1 \le k \quad \Rightarrow \quad \min \ \|\boldsymbol{AS}-\boldsymbol{Y}\|_2^2 + \lambda\|\boldsymbol{S}\|_1$$

Lasso

Regularization

- Convex relaxation: **Lasso** was proposed in 1996 by Tibshirani (24,000+ citations). Using $l1$-norm to approximate $l0$-norm.

$$\min \ \|\boldsymbol{AS}-\boldsymbol{Y}\|_2^2 \quad s.t. \ \ \|\boldsymbol{S}\|_0 \le k \ \Rightarrow \ \min \ \|\boldsymbol{AS}-\boldsymbol{Y}\|_2^2 \quad s.t. \ \ \|\boldsymbol{S}\|_1 \le k \ \Rightarrow \ \min \ \|\boldsymbol{AS}-\boldsymbol{Y}\|_2^2 + \lambda\|\boldsymbol{S}\|_1$$

Lasso

- **Basic idea of Ridge regression:** Using $l2$-norm to reduce the norm of coefficients.

$$\min \ \|\boldsymbol{AS}-\boldsymbol{Y}\|_2^2 + \lambda\|\boldsymbol{S}\|_2^2$$

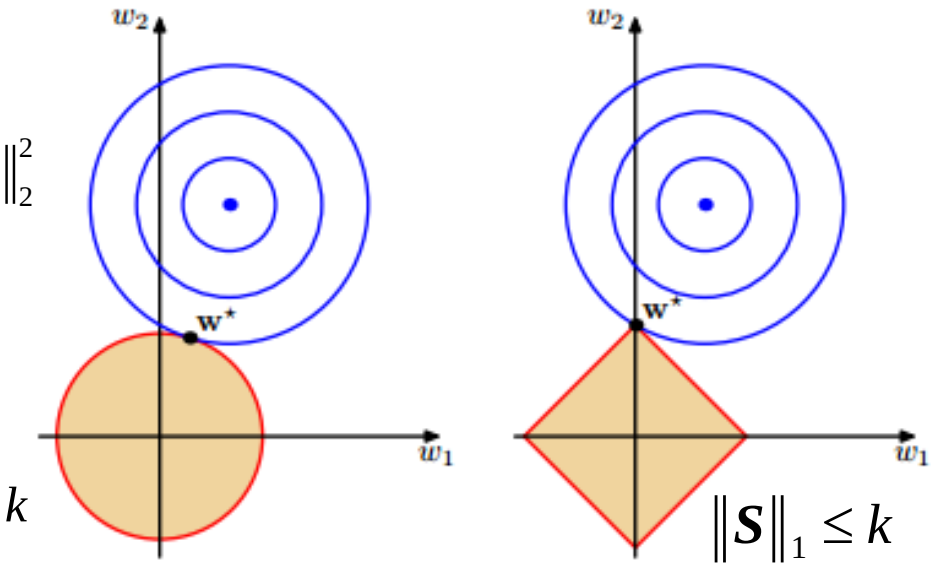
- The optimal parameters of ridge regression can be reached analytically:

$$\boldsymbol{S} = (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{Y}$$

Regularization

- Geometric interpretation

$$\min \ \|\boldsymbol{AS}-\boldsymbol{Y}\|_2^2$$

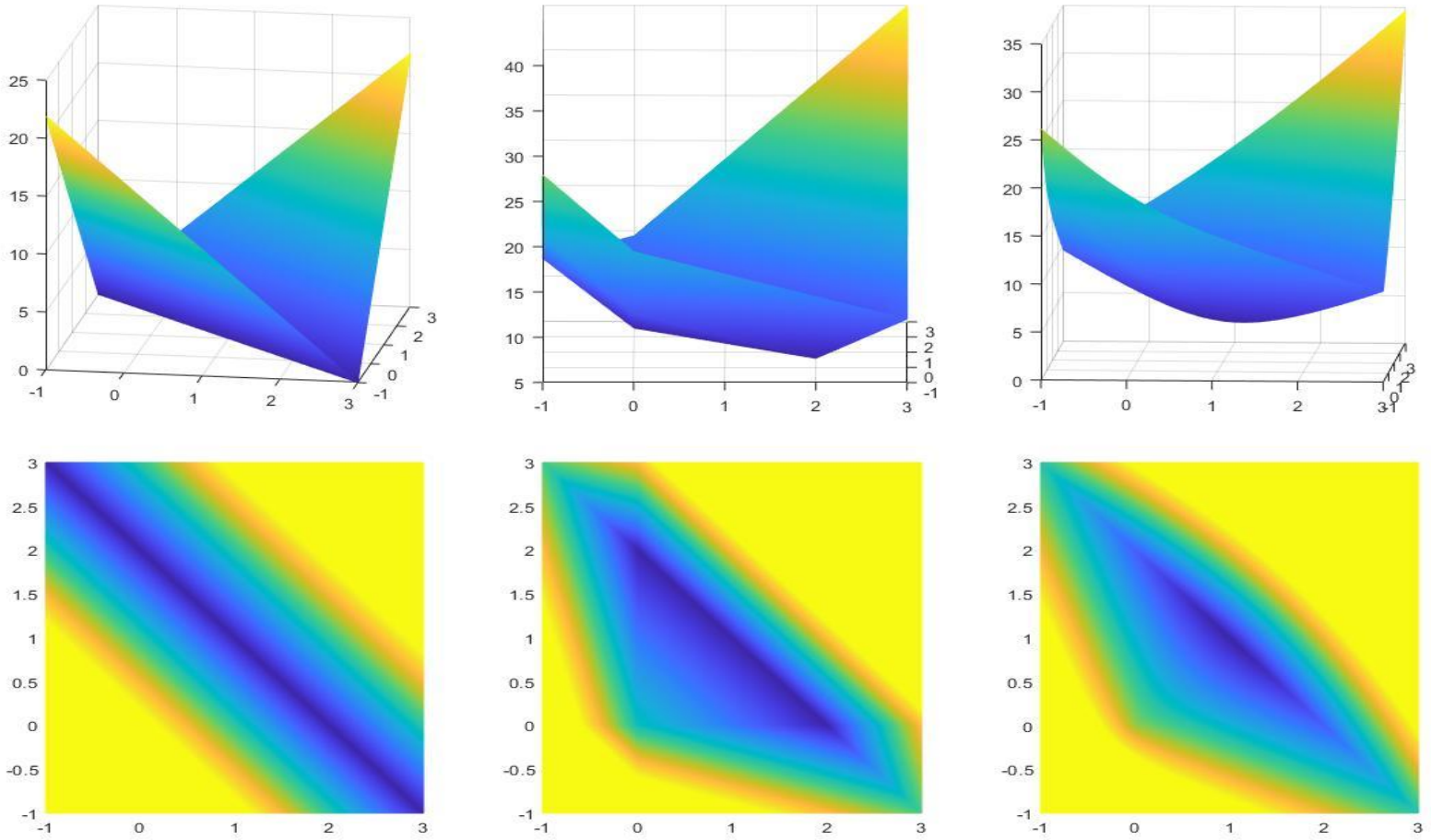


- Such idea is named **regularization**.
- Further reading: How to understand regularization.
  - https://blog.csdn.net/qq_20412595/article/details/81636105?utm_medium=distribute.pc_relevant.none-task-blog-title-1&spm=1001.2101.3001.4242
  - https://www.cnblogs.com/jianxinzhou/p/4083921.html

Regularization

- Geometric interpretation of the simple example

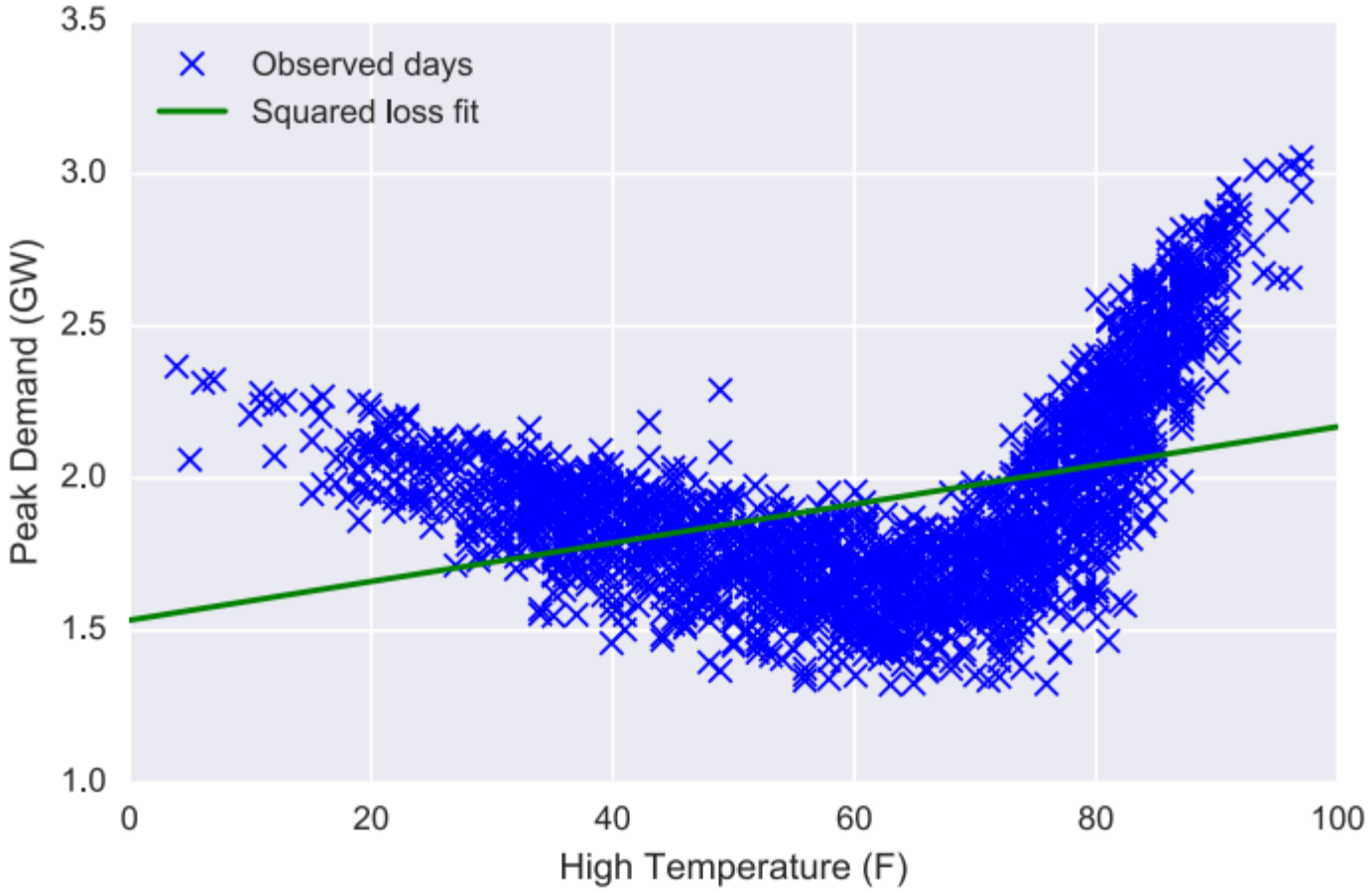
$$y = \{2, 4.1, 6, 8\}$$
$$x_1 = \{1, 2.01, 3, 4\}$$
$$x_2 = \{1, 2, 3, 4\}$$



$$\min \left\| \boldsymbol{AS} - \boldsymbol{Y} \right\|_2^2 \qquad \min \left\| \boldsymbol{AS} - \boldsymbol{Y} \right\|_2^2 + \lambda \left\| \boldsymbol{S} \right\|_1 \qquad \min \left\| \boldsymbol{AS} - \boldsymbol{Y} \right\|_2^2 + \lambda \left\| \boldsymbol{S} \right\|_2^2$$

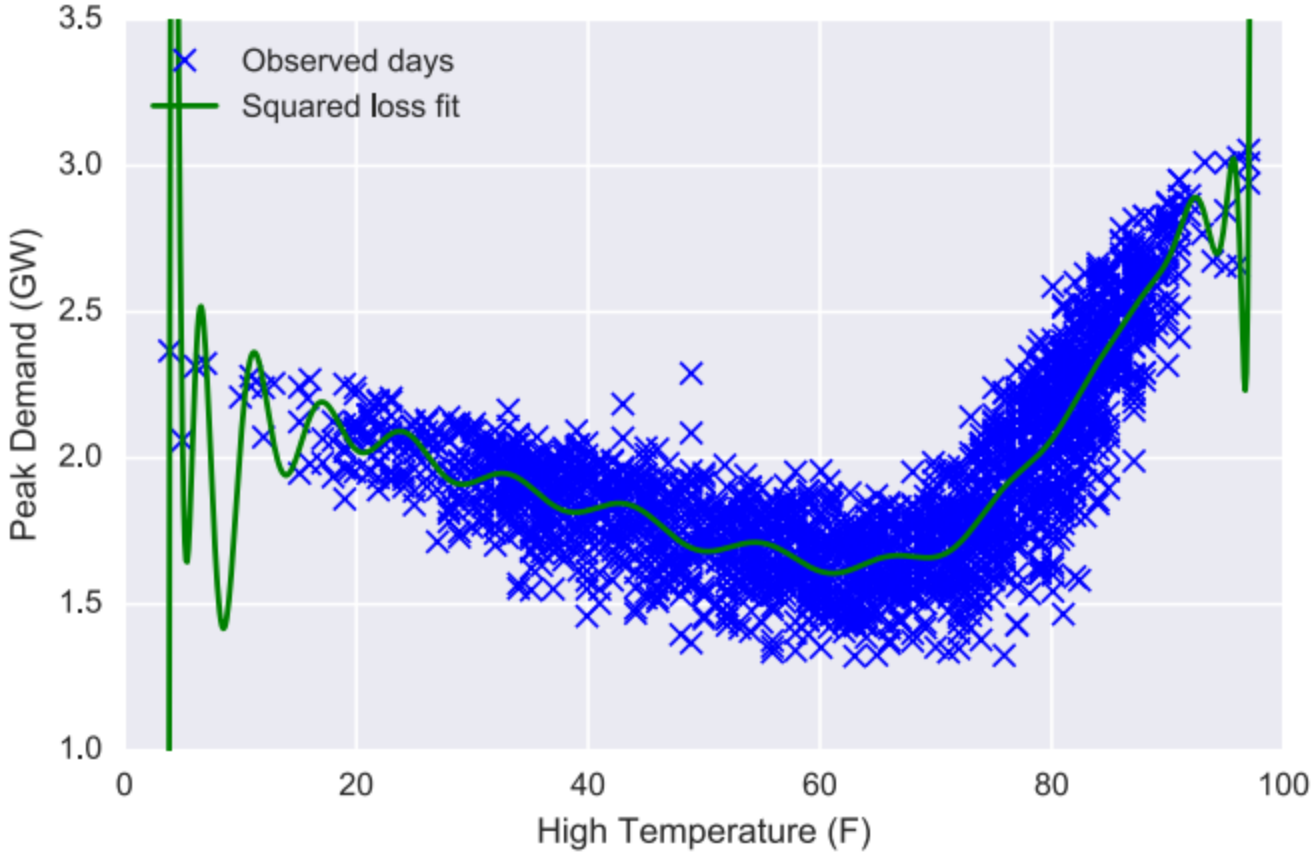
Example: Predicting Electricity Demand

- Linear Regression Fit: Under fit



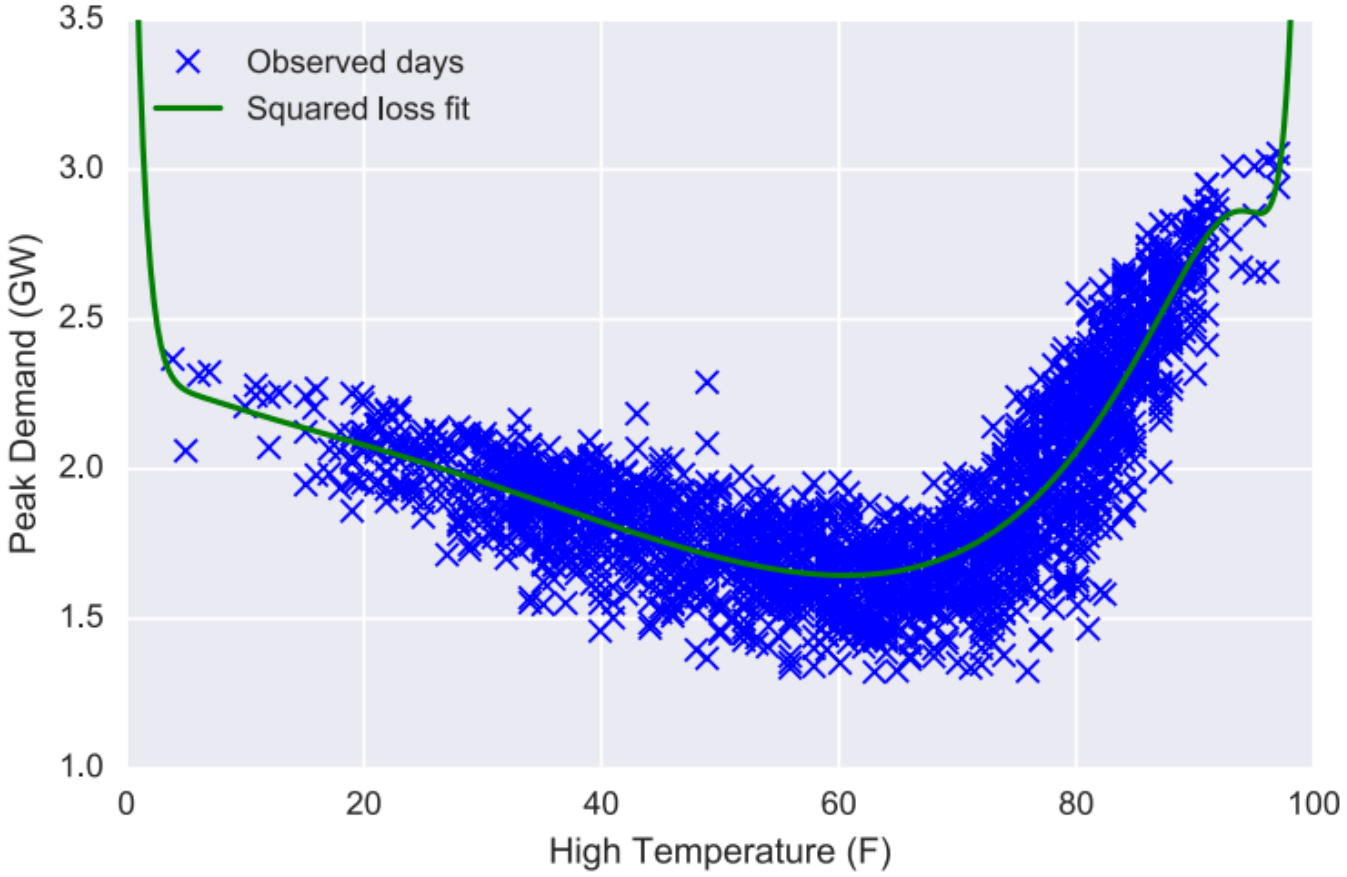
Example: Predicting Electricity Demand

- Polynomial Features of Degree 50
- An example of overfitting.



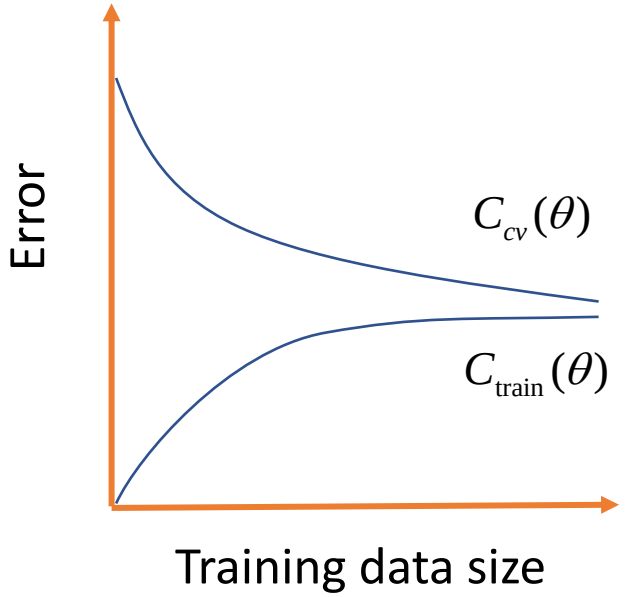
Example: Predicting Electricity Demand

- Polynomial Features of Degree 50
- But add $l2$-norm as Ridge regression!

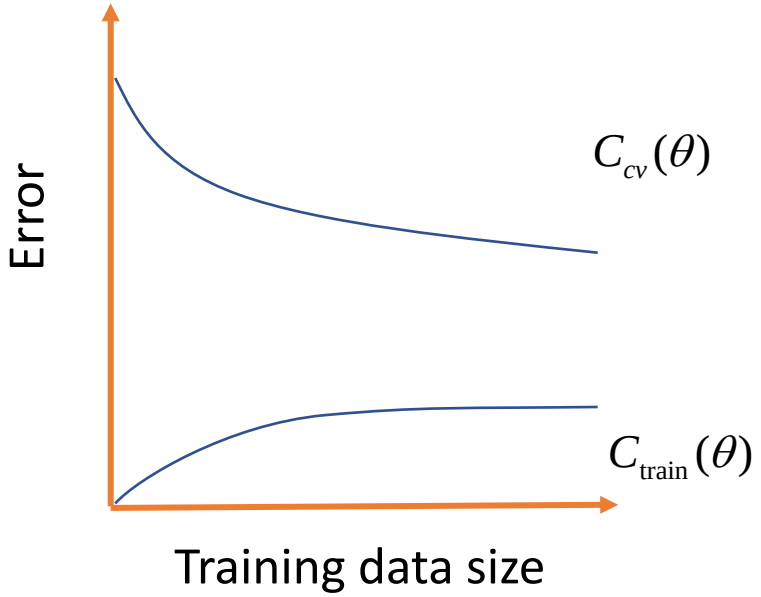


Learning curve

- **High bias:** if the cross validation error is still high and similar to the training error, the problem may have high bias problem, adding more data will not help, adding more features will help.



- **High variance:** if the cross validation error is high but the training error is low, the problem may have high variance problem, adding more data and reduce the numbers of features will help.



Homework 2

In recent researches, big data and machine learning are widely applied on the power system operation field. The power system operation data (e.g. power flow, bus voltage, generation data, loads' power demand, etc.) are enormous in terms of both dimensions and samples. With these data, system operators want to know the stability of a complex power system fast and accurately by the data-driven methods.

After learning this lecture on linear/nonlinear regression , please:

- Given the dataset, apply the linear/nonlinear regression approaches to estimate the stability margin of the system.
- Try different regression methods to find the best fitting results. Try to understand the overfitting & underfitting phenomenon when you adjust the regression models and hyperparameters.

Q&A