Cambridge AI+

Linear Regression and Logistic Regression

Thomas Sauerwald

University of Cambridge, Department of Computer Science and Technology
email: tsauerwa@gmail.com

2021

Outline

Introduction

Linear Regression

Linear Regression with More than One Feature

Solving Linear Regression: Gradient Descent

Logistic Regression

Applications of Regression

Grade Prediction: Predict exam score of a student based on some data

Applications of Regression

Grade Prediction: Predict exam score of a student based on some data

Economics: Find a trend line that matches stock price time series

Applications of Regression

Grade Prediction: Predict exam score of a student based on some data

Astronomy: Determine the orbit of a planet (Gauss 1822)

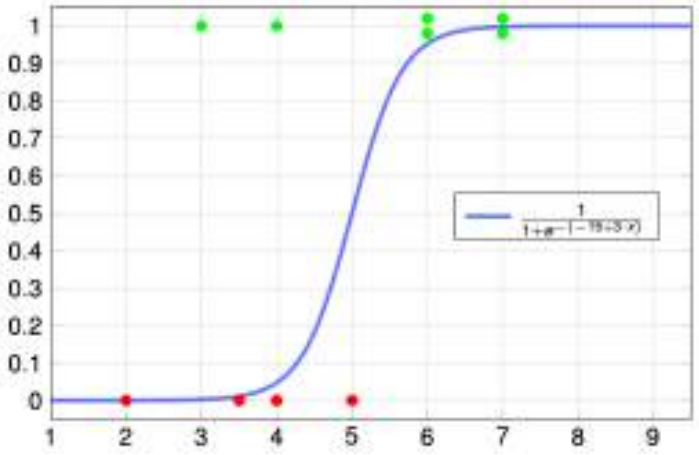
Economics: Find a trend line that matches stock price time series

Applications of Regression

Grade Prediction: Predict exam score of a student based on some data

Astronomy: Determine the orbit of a planet (Gauss 1822)

Economics: Find a trend line that matches stock price time series



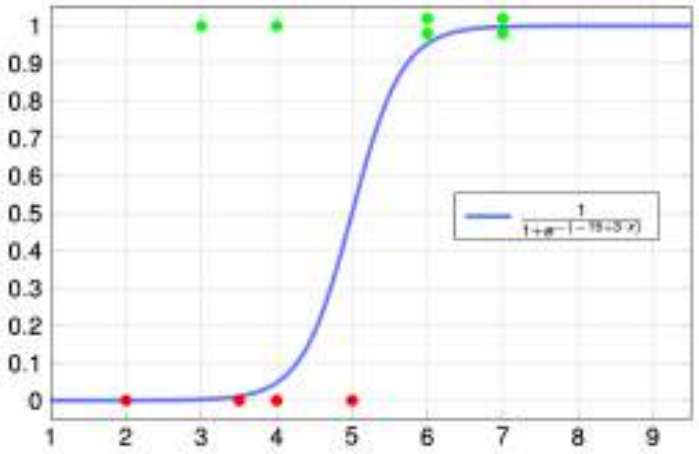
Medicine: Estimate likelihood of patient suffering from some disease

Applications of Regression

Linear Regression (this lecture)

Grade Prediction: Predict exam score of a student based on some data

Astronomy: Determine the orbit of a planet (Gauss 1822)



Economics: Find a trend line that matches stock price time series

Medicine: Estimate likelihood of patient suffering from some disease

Applications of Regression

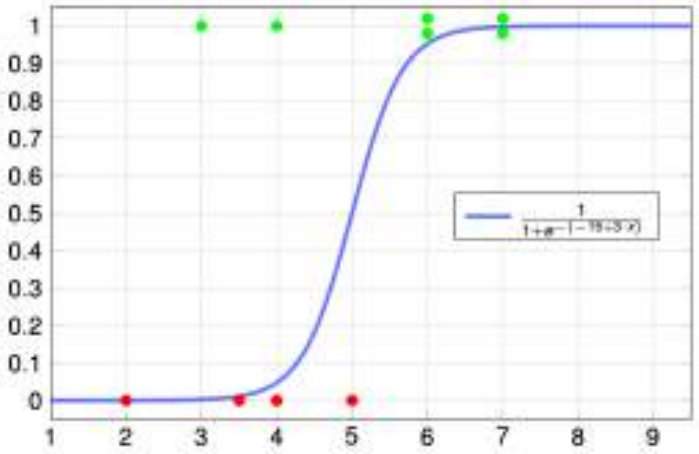
Linear Regression (this lecture)

Grade Prediction: Predict exam score of a student based on some data

≈ Linear Regression

Astronomy: Determine the orbit of a planet (Gauss 1822)

Economics: Find a trend line that matches stock price time series



Medicine: Estimate likelihood of patient suffering from some disease

Applications of Regression

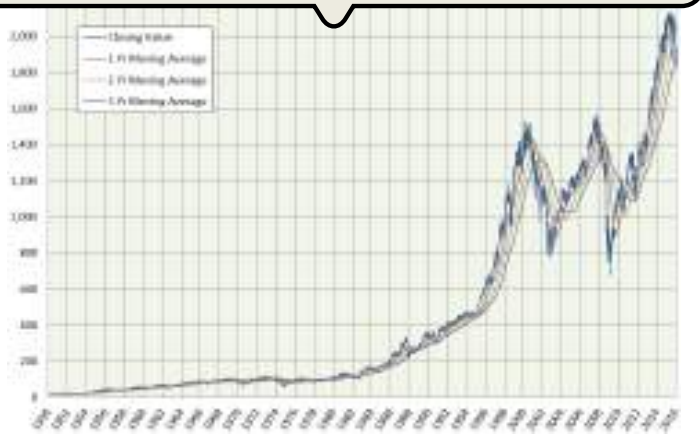
Linear Regression (this lecture)

Grade Prediction: Predict exam score of a student based on some data

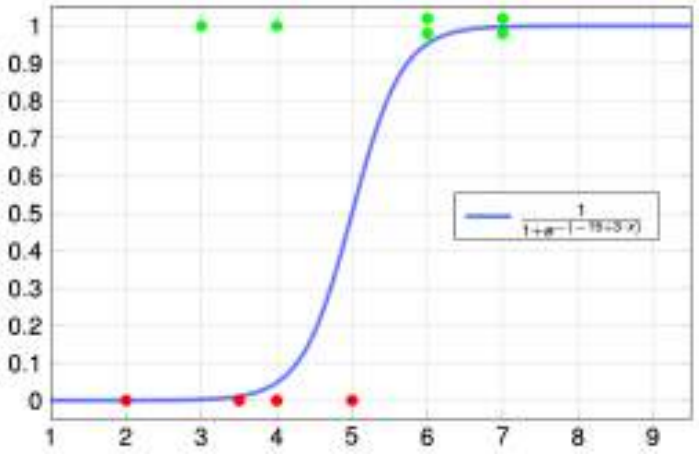
≈ Linear Regression

Astronomy: Determine the orbit of a planet (Gauss 1822)

Exponential Regression and Auto-Regressive Models (supervision)



Economics: Find a trend line that matches stock price time series



Medicine: Estimate likelihood of patient suffering from some disease

Applications of Regression

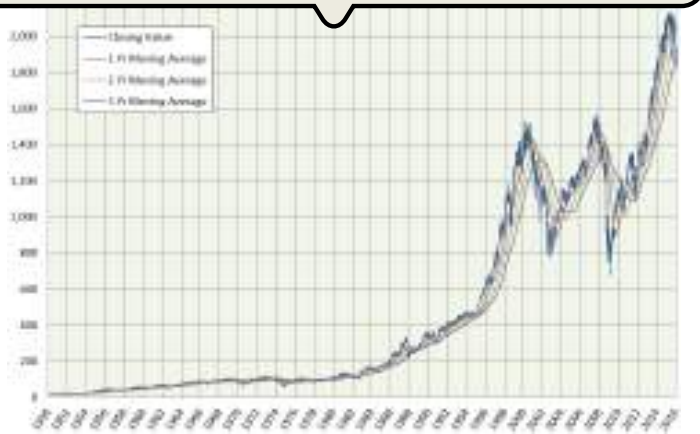
Linear Regression (this lecture)

Grade Prediction: Predict exam score of a student based on some data

≈ Linear Regression

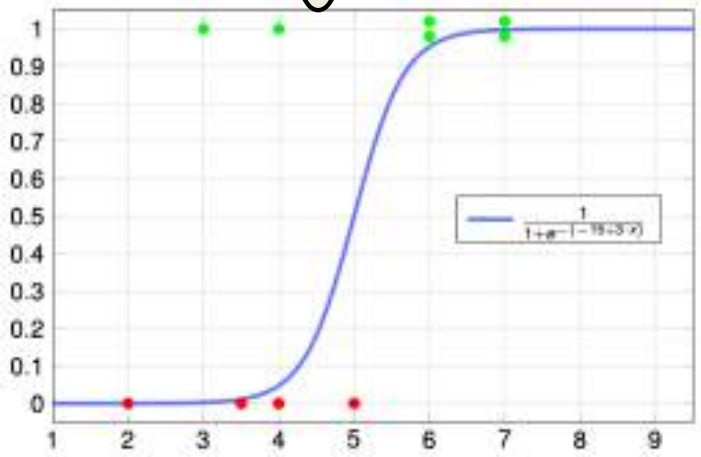
Astronomy: Determine the orbit of a planet (Gauss 1822)

Exponential Regression and Auto-Regressive Models (supervision)



Economics: Find a trend line that matches stock price time series

Logistic Regression (next lecture)



Medicine: Estimate likelihood of patient suffering from some disease

Exercise Question 2

Consider the following data set, which records several phone calls of clients to an e-commerce company.

Client	Call Length	Bought Items
1	13	2
2	17	1
3	32	2
4	36	4
5	45	4
6	53	3
7	67	6
8	74	5
9	83	7
10	132	11

Exercise Question 2

Consider the following data set, which records several phone calls of clients to an e-commerce company.

Client	Call Length	Bought Items
1	13	2
2	17	1
3	32	2
4	36	4
5	45	4
6	53	3
7	67	6
8	74	5
9	83	7
10	132	11

1. Fit the data to a linear regression model and compute the R^2-value.

Exercise Question 2

Consider the following data set, which records several phone calls of clients to an e-commerce company.

Client	Call Length	Bought Items
1	13	2
2	17	1
3	32	2
4	36	4
5	45	4
6	53	3
7	67	6
8	74	5
9	83	7
10	132	11

1. Fit the data to a linear regression model and compute the R^2-value.
2. Consider a client that calls the company for 90 minutes. Can you estimate the number of purchased items?

Exercise Question 2

Consider the following data set, which records several phone calls of clients to an e-commerce company.

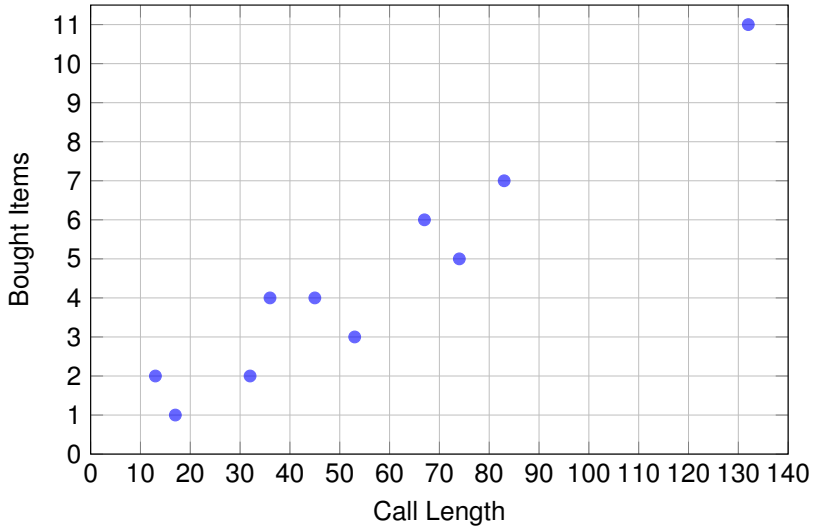
Client	Call Length	Bought Items
1	13	2
2	17	1
3	32	2
4	36	4
5	45	4
6	53	3
7	67	6
8	74	5
9	83	7
10	132	11

1. Fit the data to a linear regression model and compute the R^2-value.
2. Consider a client that calls the company for 90 minutes. Can you estimate the number of purchased items?
3. Somebody tells you that a recent client purchased 3 items, but you do not know the length of the call. Can you estimate that number too?

Exercise Question 2

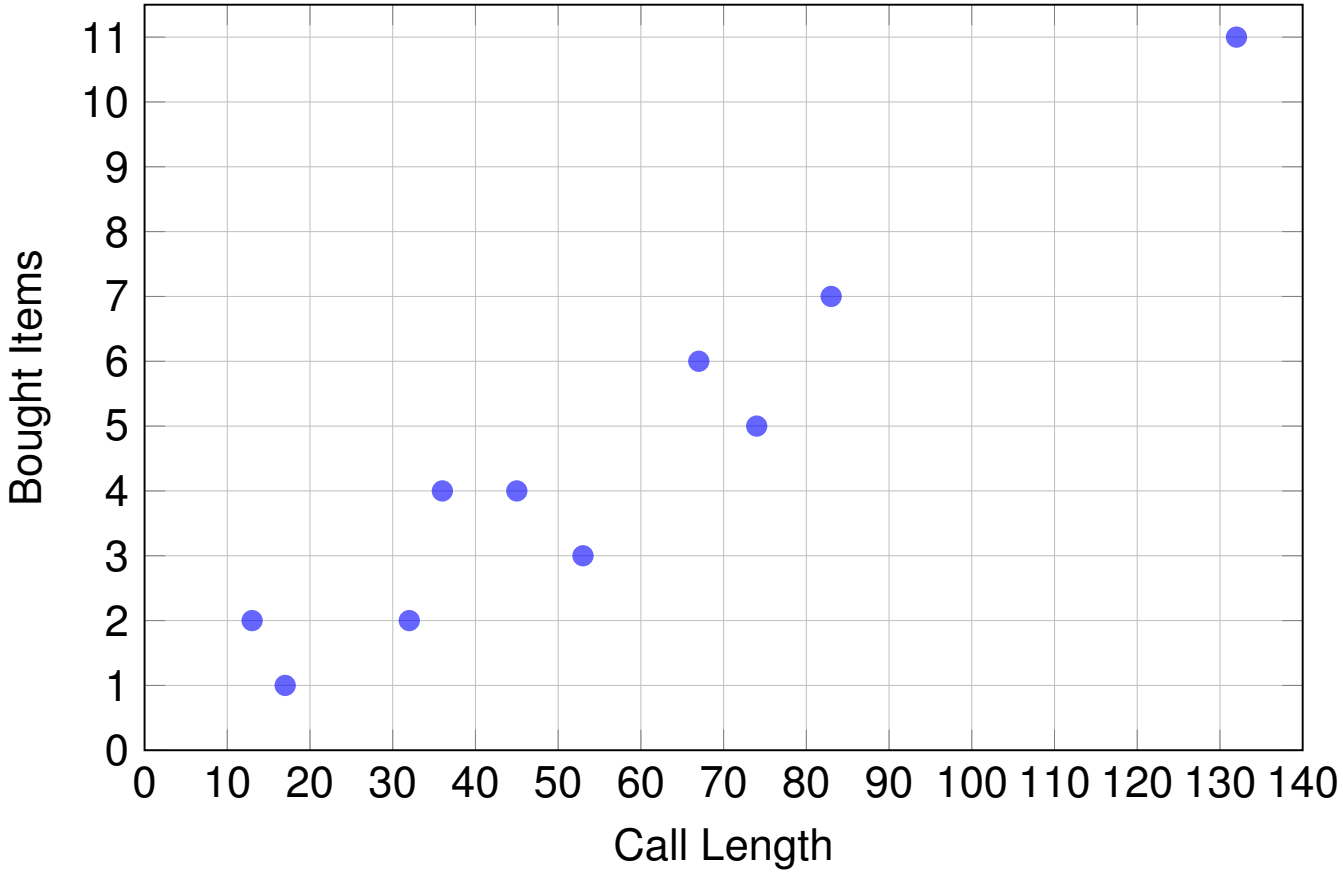
Consider the following data set, which records several phone calls of clients to an e-commerce company.

Client	Call Length	Bought Items
1	13	2
2	17	1
3	32	2
4	36	4
5	45	4
6	53	3
7	67	6
8	74	5
9	83	7
10	132	11

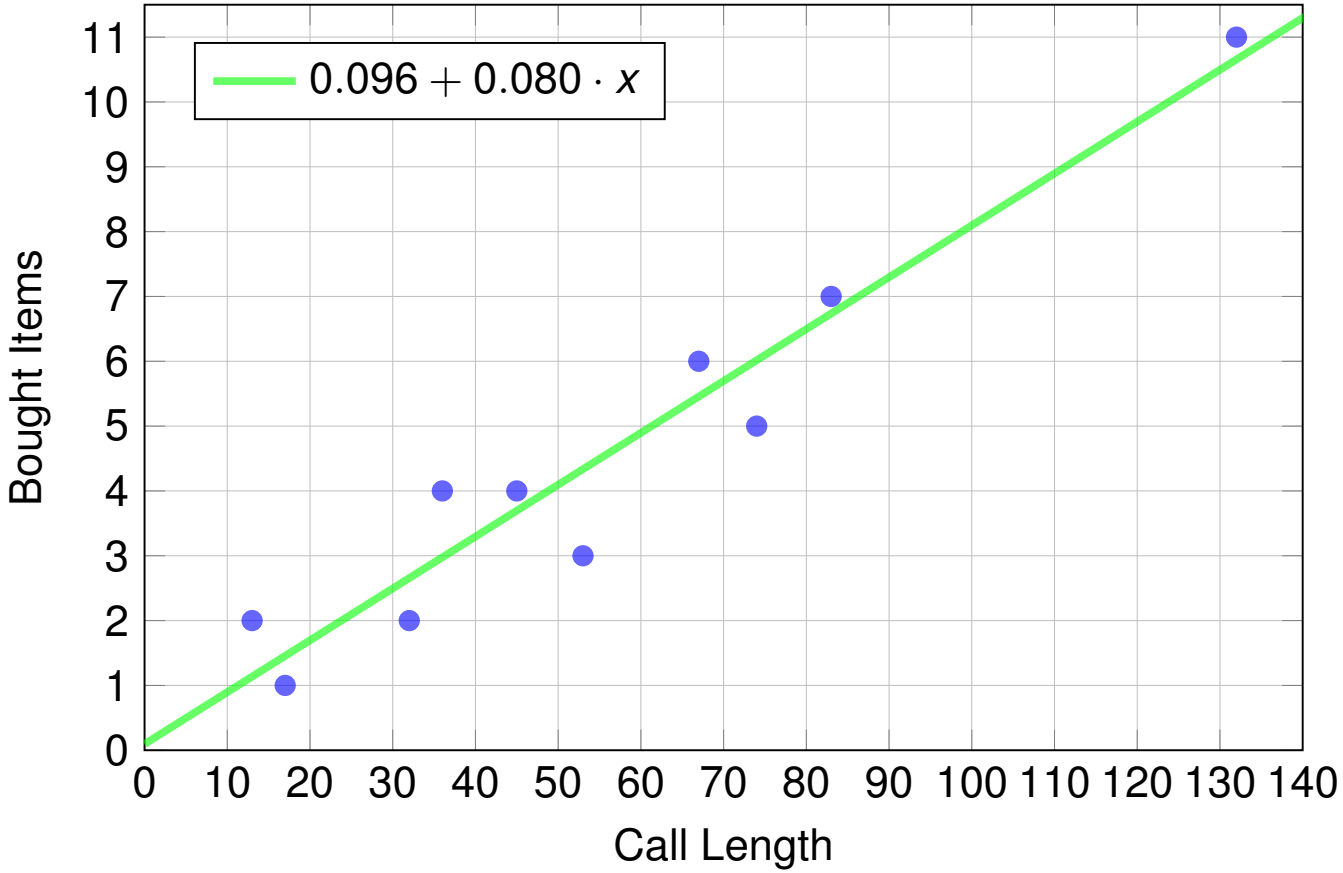


1. Fit the data to a linear regression model and compute the R^2-value.
2. Consider a client that calls the company for 90 minutes. Can you estimate the number of purchased items?
3. Somebody tells you that a recent client purchased 3 items, but you do not know the length of the call. Can you estimate that number too?

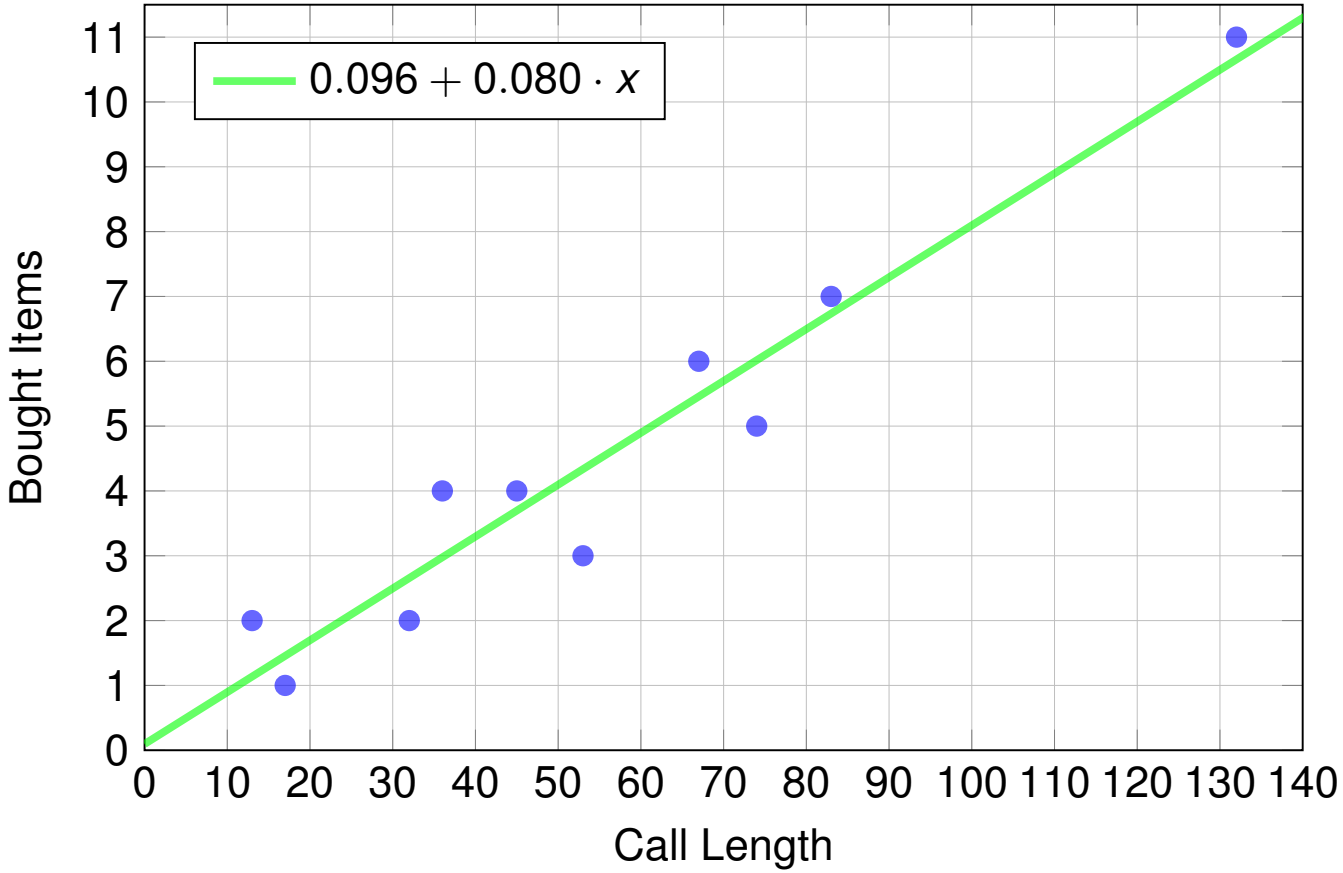
Step 1: Finding the Regression Line and R^2-Value



Step 1: Finding the Regression Line and R^2-Value

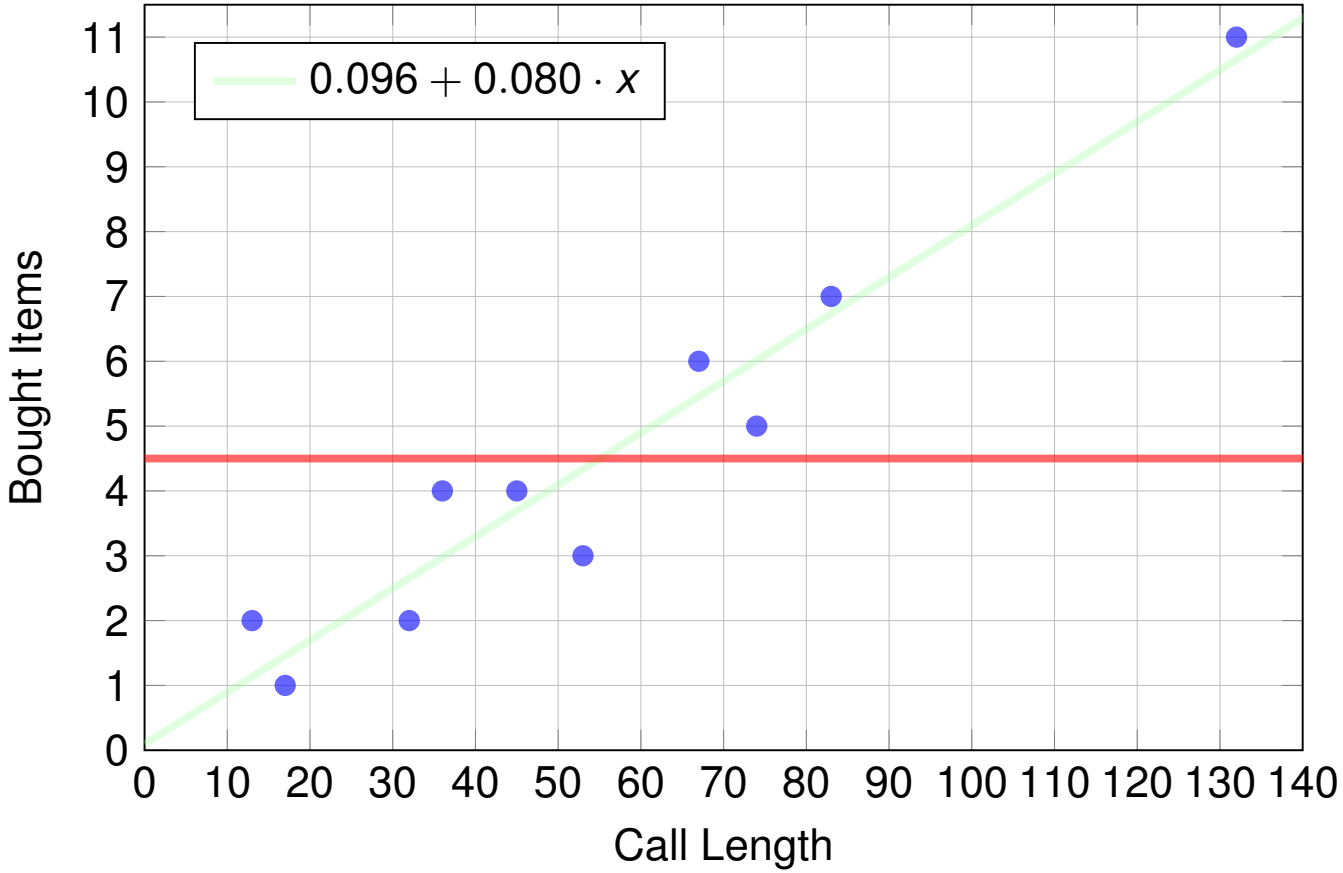


Step 1: Finding the Regression Line and R^2-Value



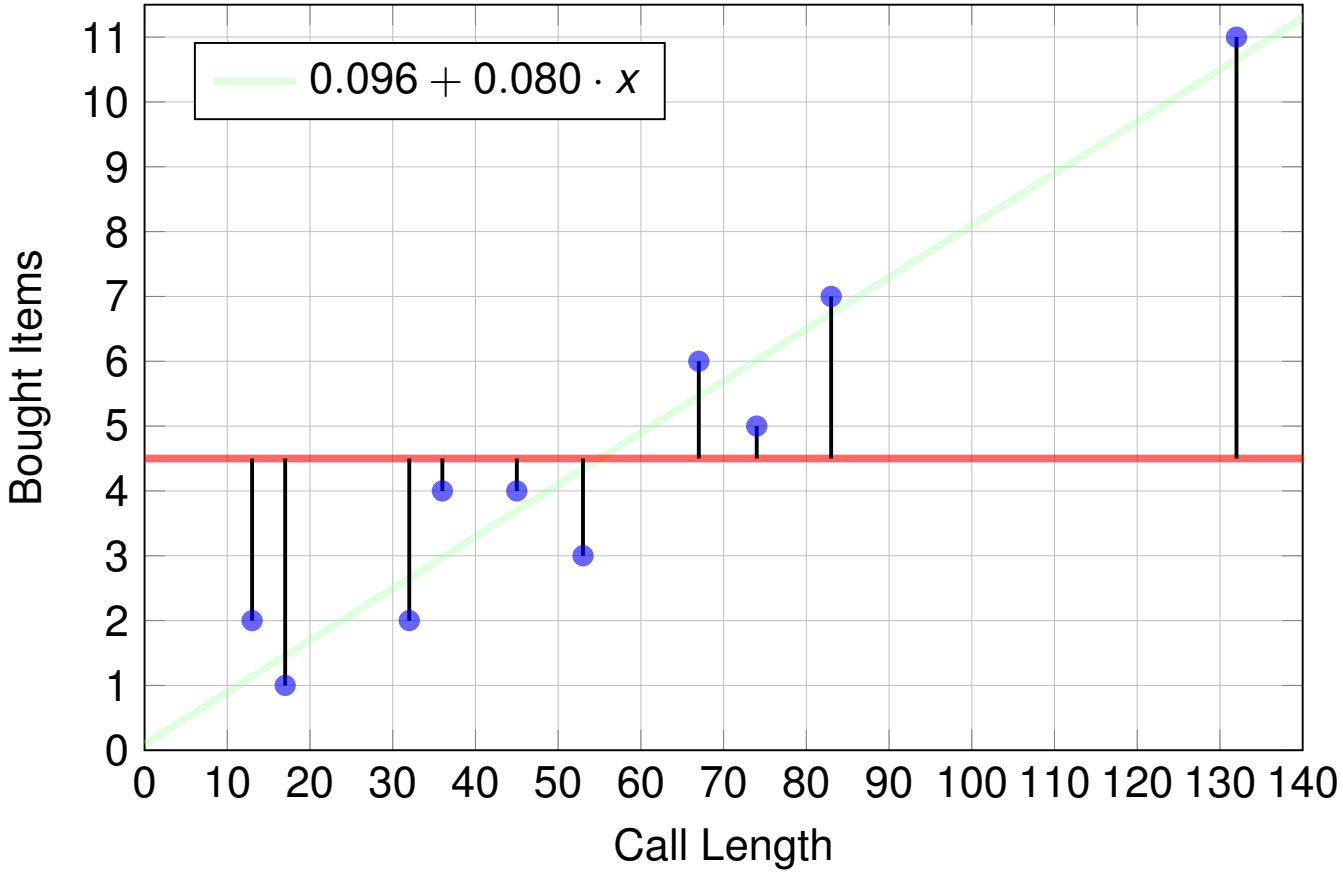
- Empirical Variance: $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2 =$

Step 1: Finding the Regression Line and R^2-Value



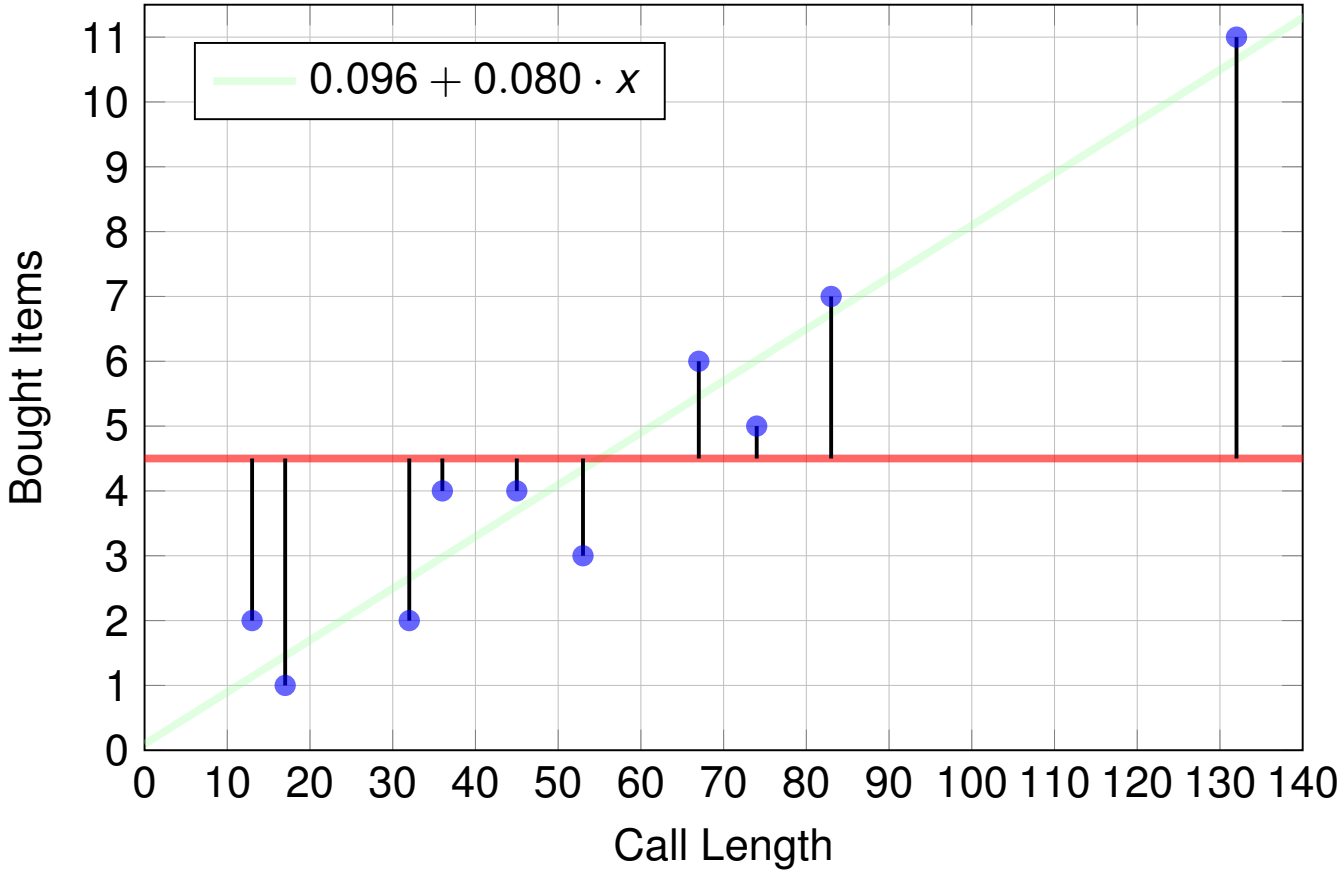
- Empirical Variance: $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2 =$

Step 1: Finding the Regression Line and R^2-Value



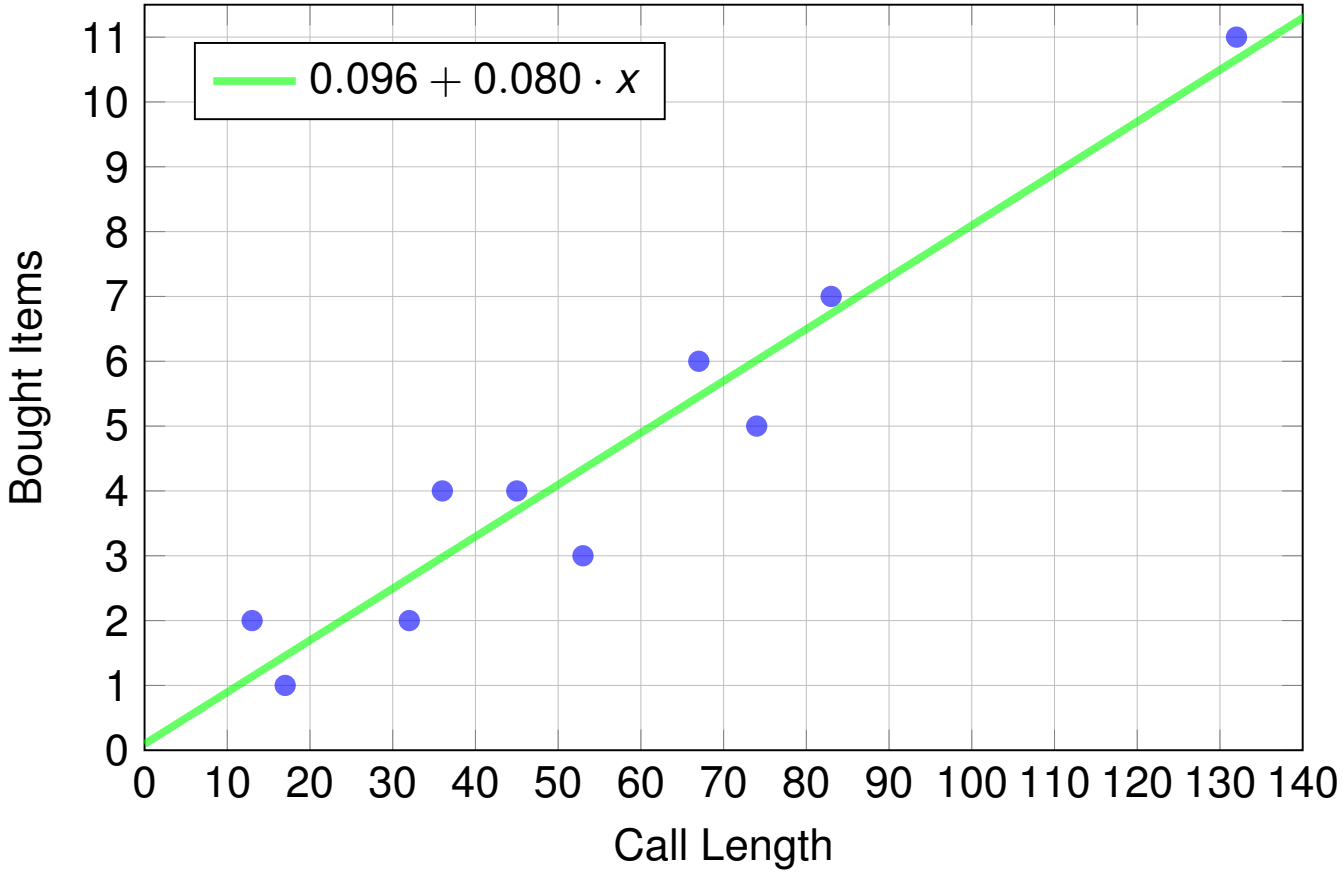
- Empirical Variance: $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2 =$

Step 1: Finding the Regression Line and R^2-Value



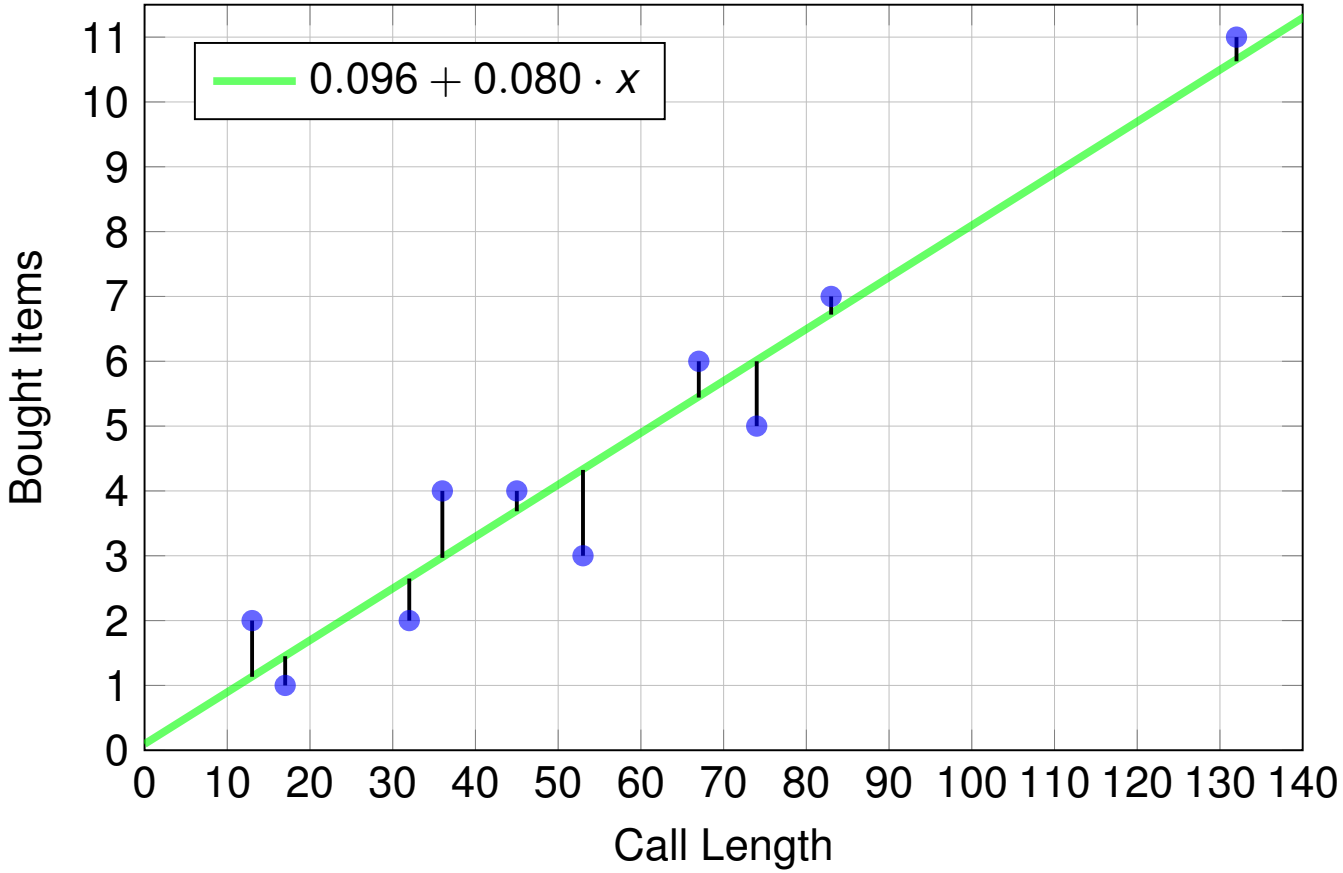
- Empirical Variance: $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2 = 78.5$

Step 1: Finding the Regression Line and R^2-Value



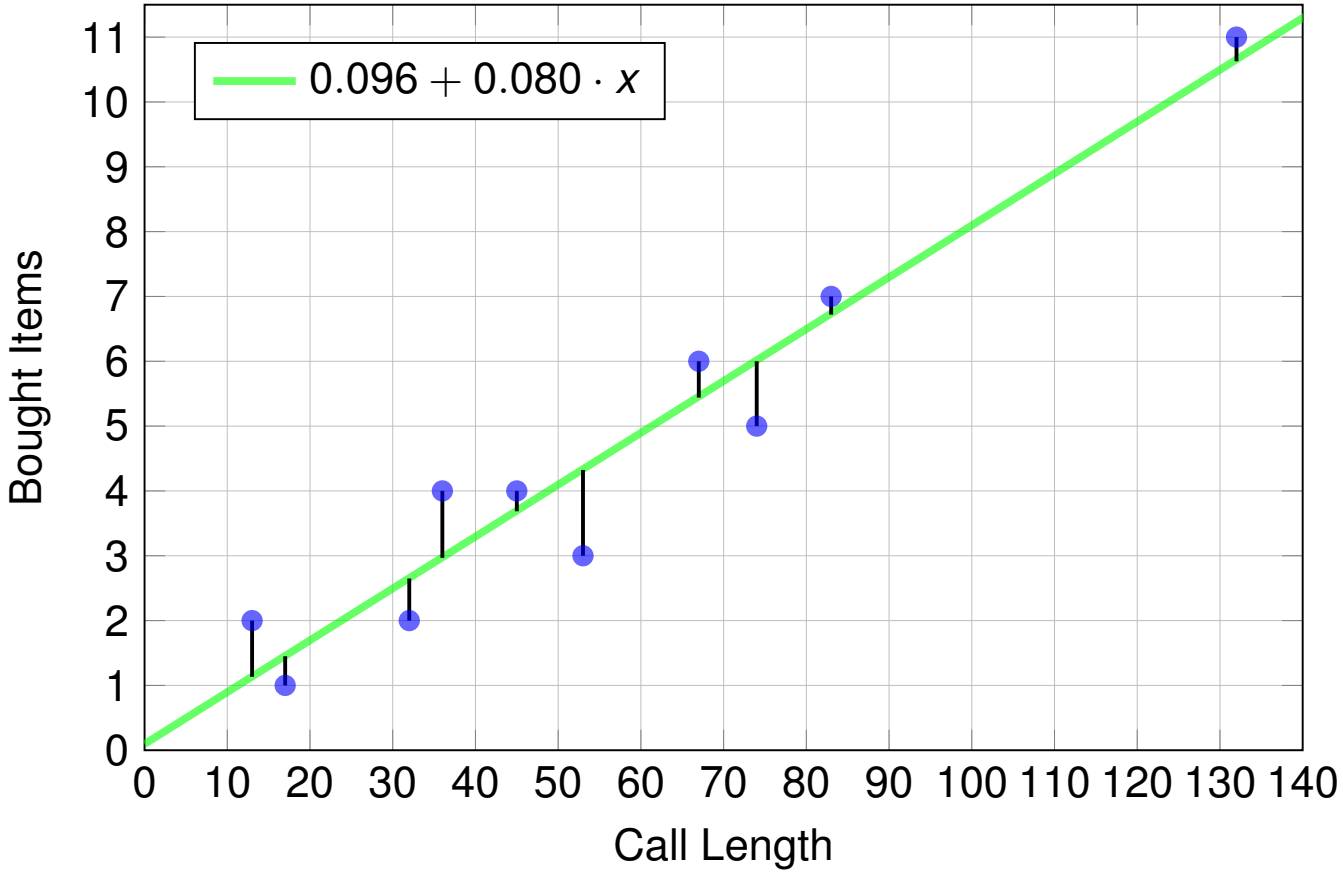
- Empirical Variance: $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2 = 78.5$
- Squared Error: $SS_{res} = \sum_{i=1}^{10} (y_i - \widetilde{y}_i)^2$

Step 1: Finding the Regression Line and R^2-Value



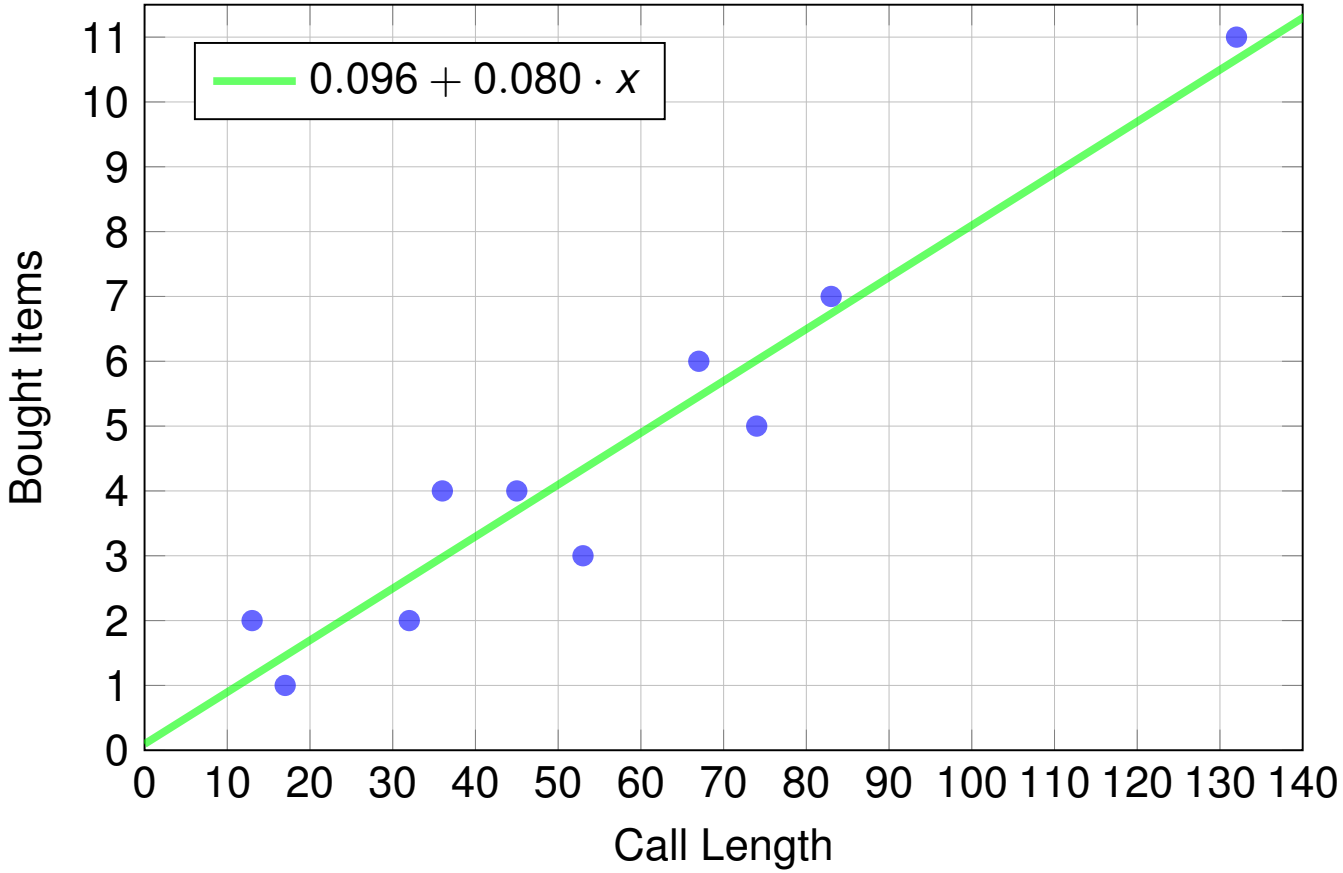
- Empirical Variance: $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2 = 78.5$
- Squared Error: $SS_{res} = \sum_{i=1}^{10} (y_i - \widetilde{y}_i)^2$

Step 1: Finding the Regression Line and R^2-Value



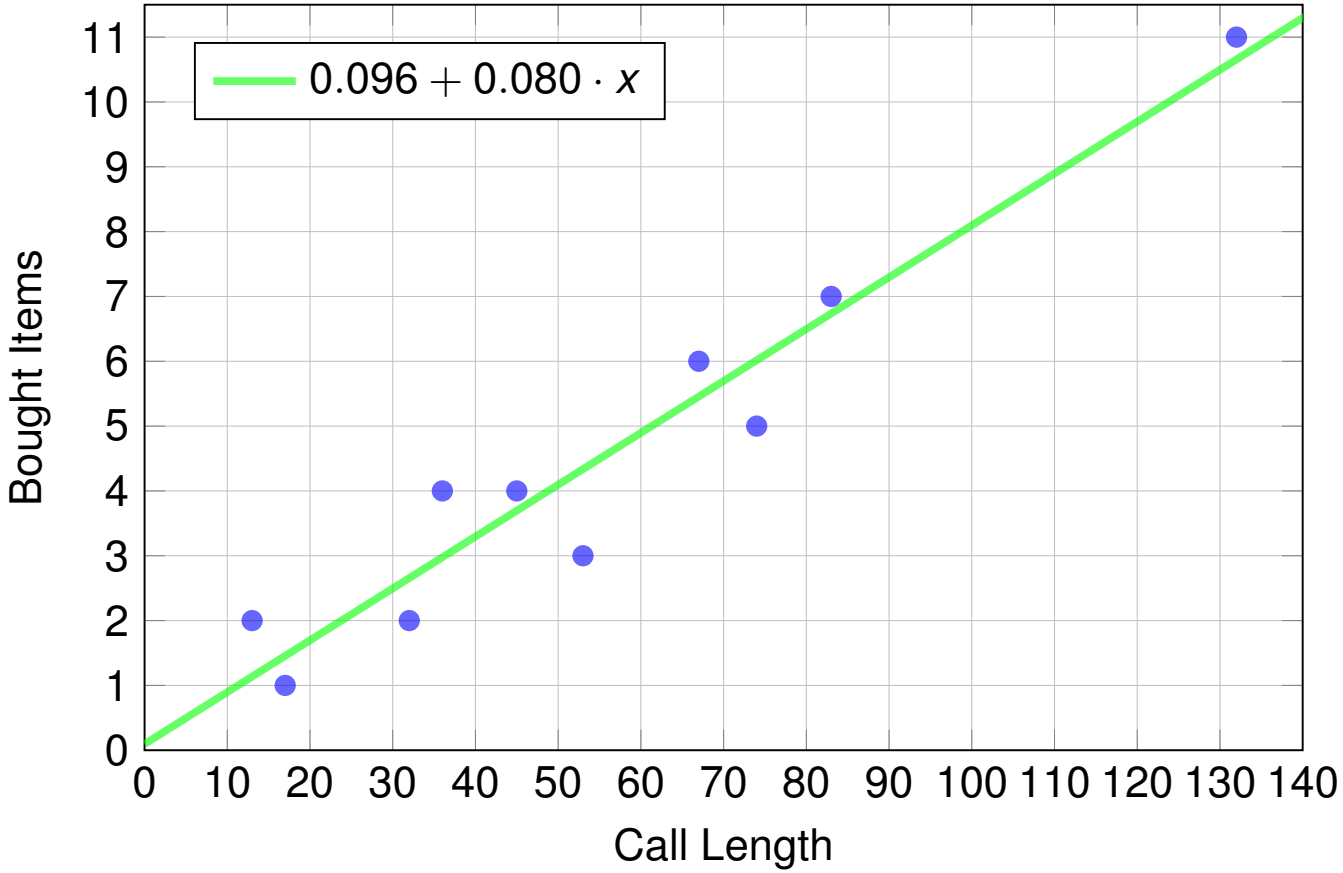
- Empirical Variance: $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2 = 78.5$
- Squared Error: $SS_{res} = \sum_{i=1}^{10} (y_i - \widetilde{y}_i)^2 = 5.825$

Step 1: Finding the Regression Line and R^2-Value



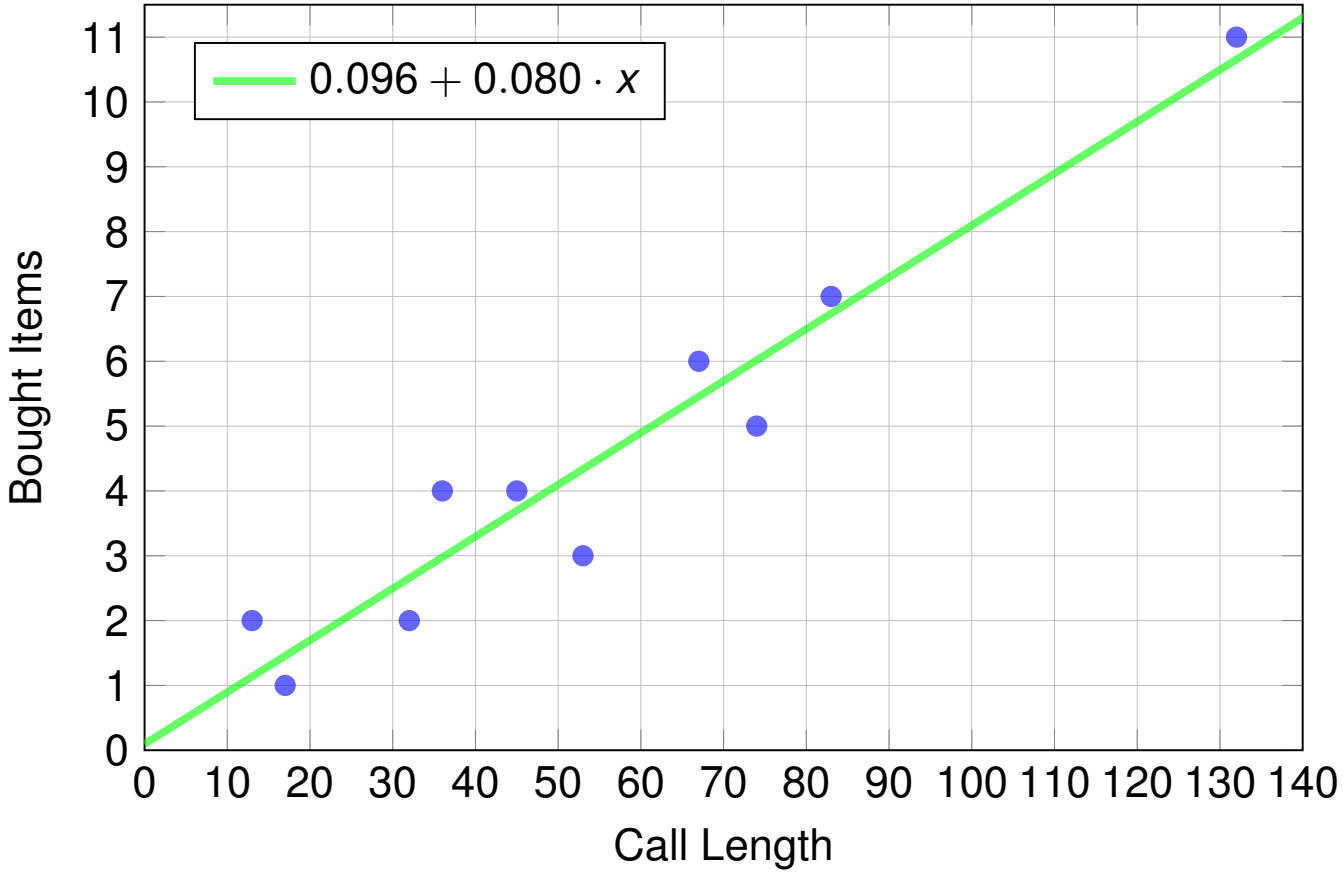
- Empirical Variance: $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2 = 78.5$
- Squared Error: $SS_{res} = \sum_{i=1}^{10} (y_i - \widetilde{y}_i)^2 = 5.825$
- R^2-value: $1 - \frac{SS_{res}}{SS_{tot}} =$

Step 1: Finding the Regression Line and R^2-Value



- Empirical Variance: $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2 = 78.5$
- Squared Error: $SS_{res} = \sum_{i=1}^{10} (y_i - \widetilde{y}_i)^2 = 5.825$
- R^2-value: $1 - \frac{SS_{res}}{SS_{tot}} = 0.926$

Step 1: Finding the Regression Line and R^2-Value

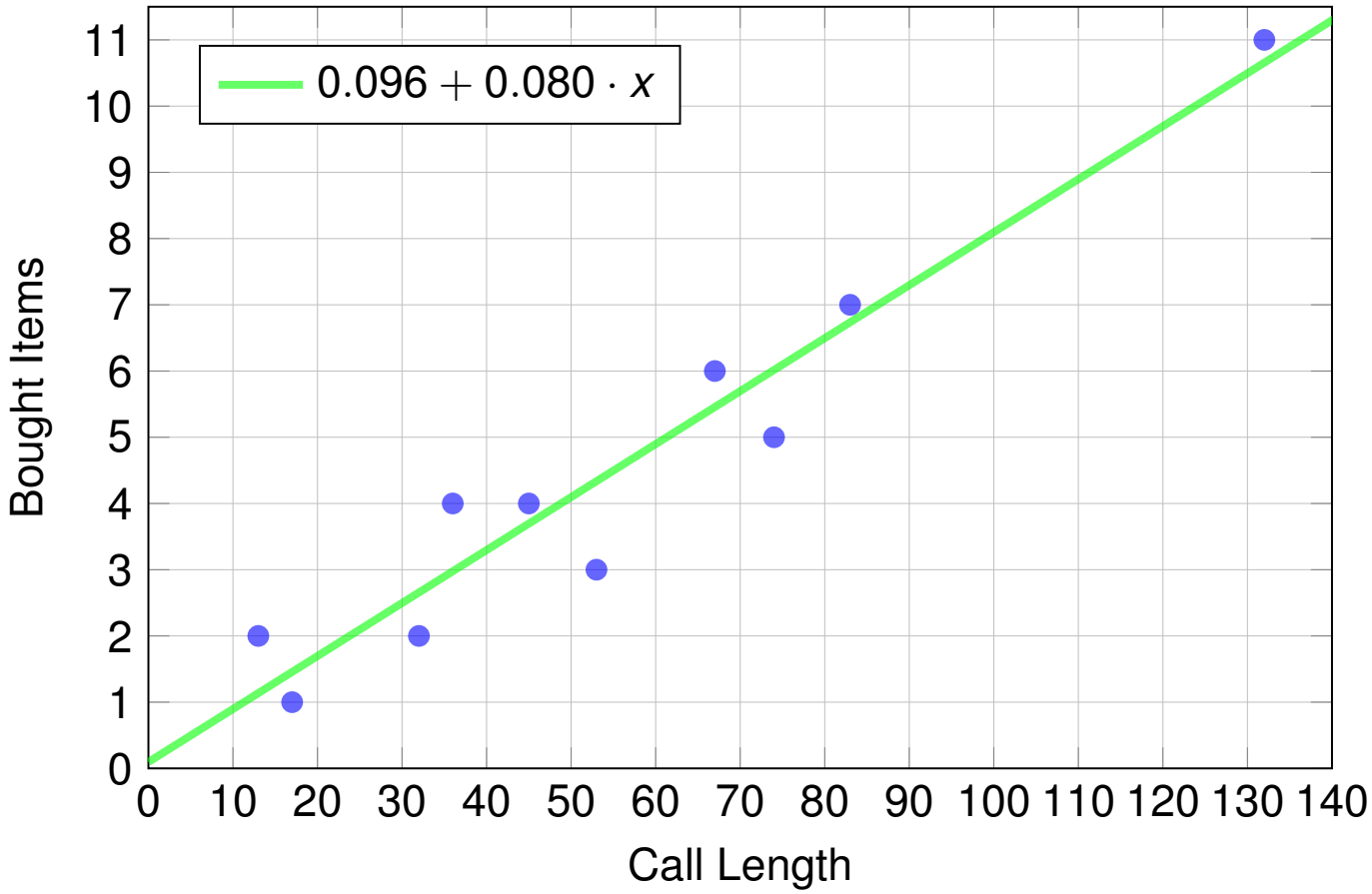


- Empirical Variance: $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2 = 78.5$
- Squared Error: $SS_{res} = \sum_{i=1}^{10} (y_i - \widetilde{y}_i)^2 = 5.825$
- R^2-value: $1 - \frac{SS_{res}}{SS_{tot}} = 0.926$

High R^2-value: our regression line is a much better fit than the constant predictor $y = 4.5$!

Step 2: Predicting y for a client

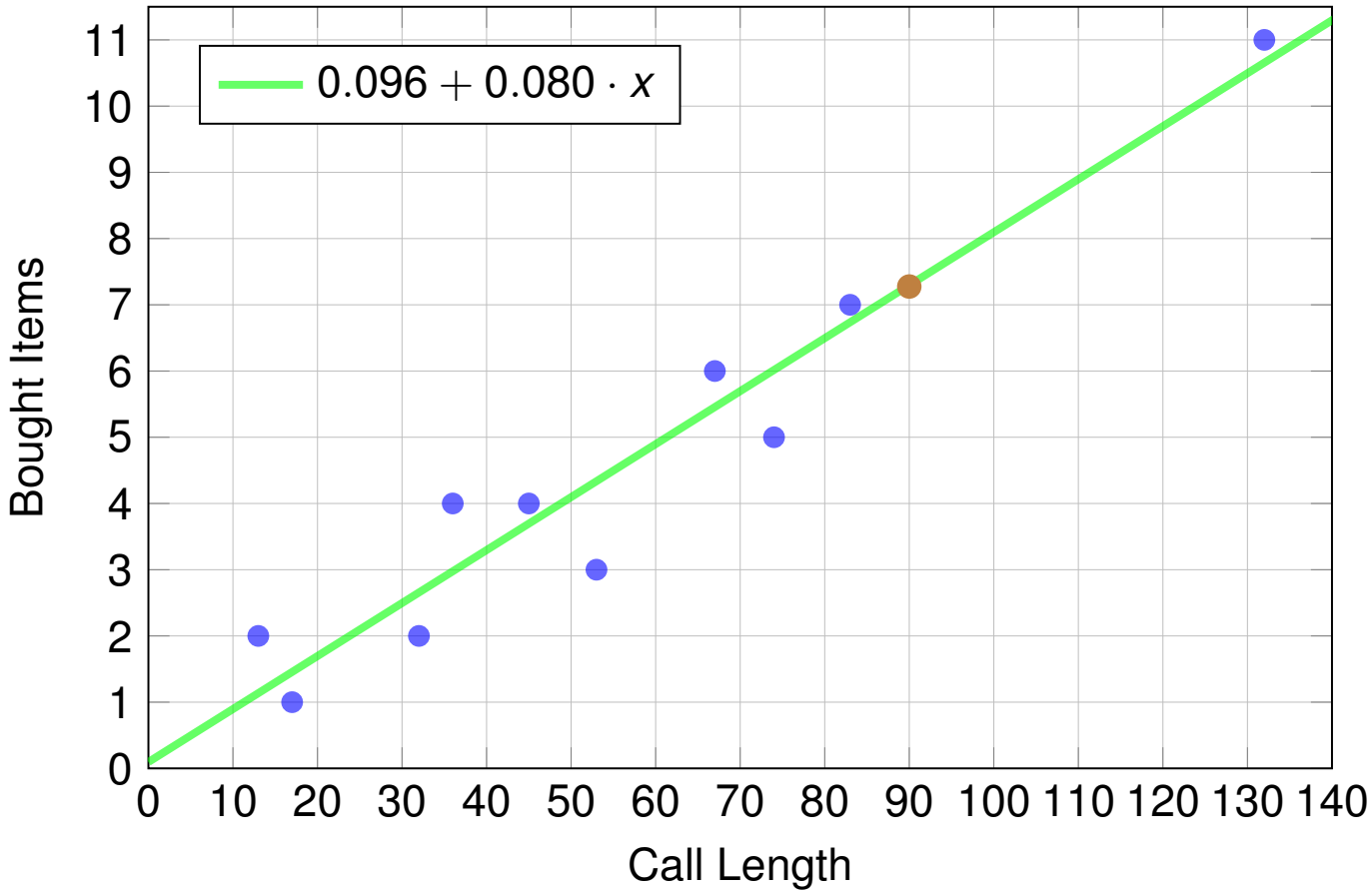
$$y = 0.0964 + 0.0798 \cdot x$$



Step 2: Predicting y for a client

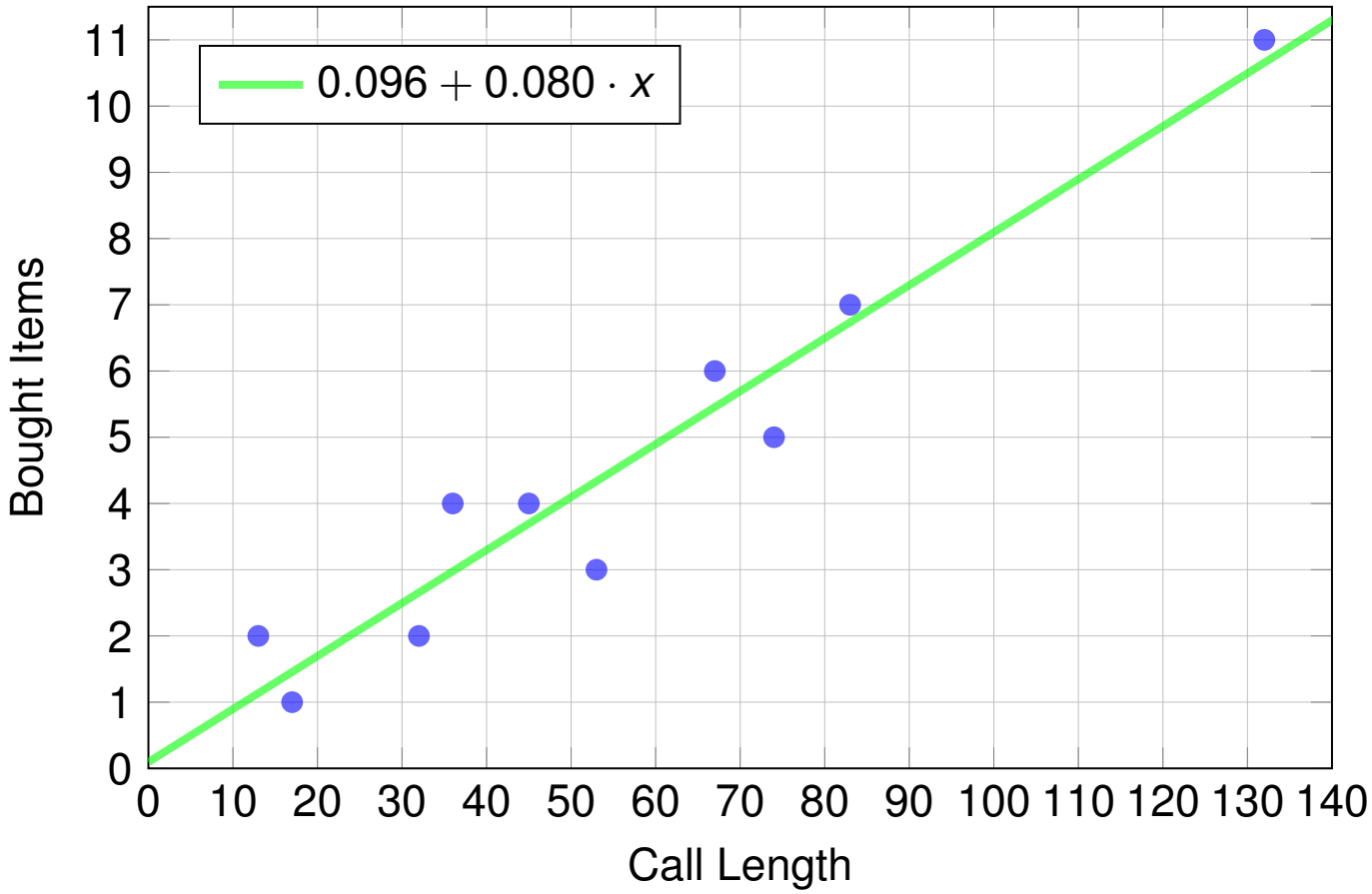
$$y = 0.0964 + 0.0798 \cdot x$$

For $x = 90$, we obtain $y(x) = 7.27 \approx 7$



Step 3: Recovering x for a known y

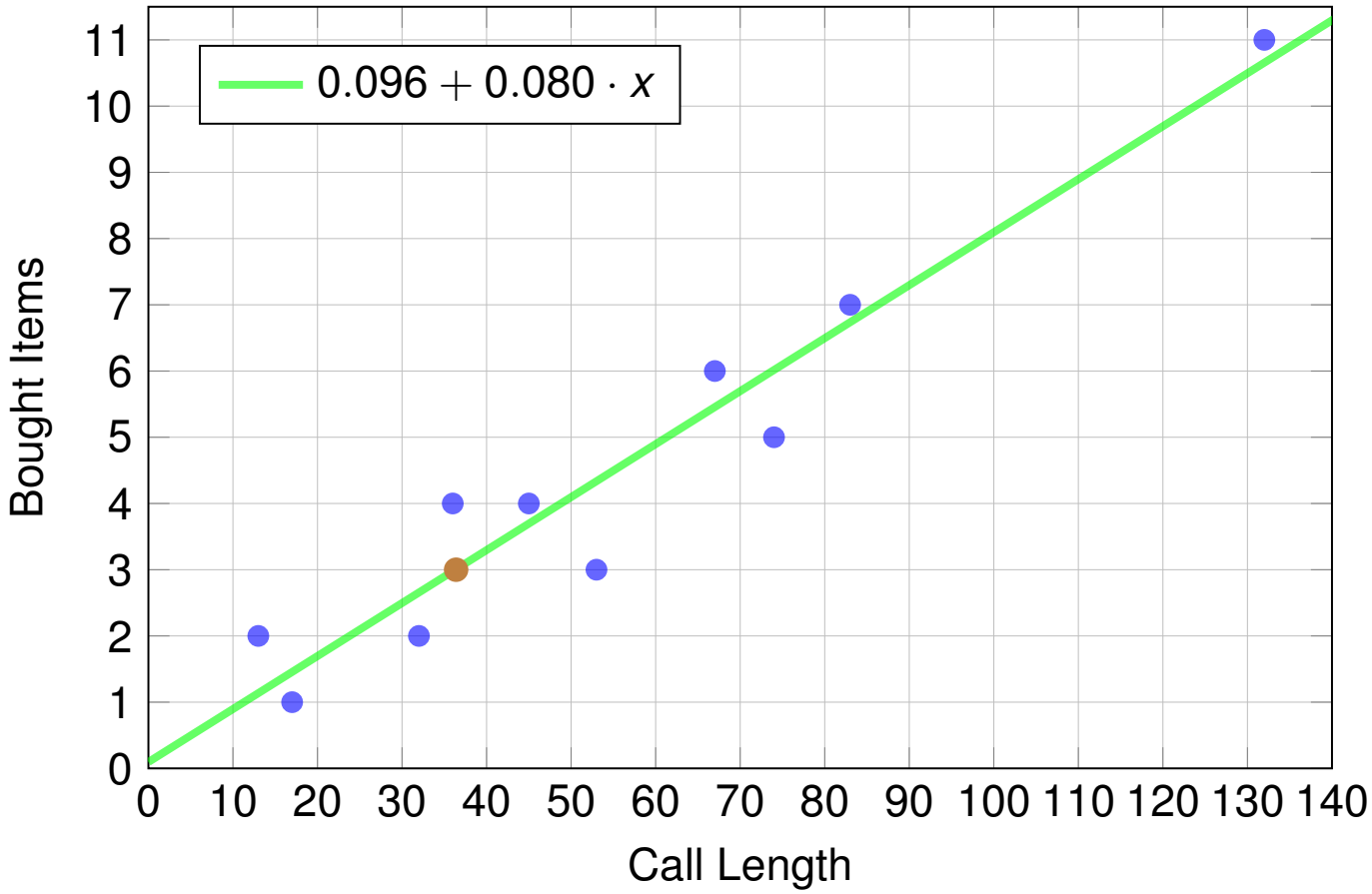
$$y = 0.0964 + 0.0798 \cdot x$$



Step 3: Recovering x for a known y

$$y = 0.0964 + 0.0798 \cdot x$$

For $y = 3$, we obtain $x(y) = 36.397 \approx 36$



Step 3: Another Approach

Question: What if we do regression with the roles of x and y swapped?

Step 3: Another Approach

Question: What if we do regression with the roles of x and y swapped?

$$x = 2.977 + 11.605 \cdot y$$

Step 3: Another Approach

Question: What if we do regression with the roles of x and y swapped?

$$x = 2.977 + 11.605 \cdot y$$

For $y = 3$, we obtain $x = 37.792$!

Step 3: Another Approach

Question: What if we do regression with the roles of x and y swapped?

$$x = 2.977 + 11.605 \cdot y$$

For $y = 3$, we obtain $x = 37.792$!

Recall that our previous regression line was:

Step 3: Another Approach

Question: What if we do regression with the roles of x and y swapped?

$$x = 2.977 + 11.605 \cdot y$$

For $y = 3$, we obtain $x = 37.792$!

Recall that our previous regression line was:

$$y = 0.0964 + 0.0798 \cdot x$$

Step 3: Another Approach

Question: What if we do regression with the roles of x and y swapped?

$$x = 2.977 + 11.605 \cdot y$$

For $y = 3$, we obtain $x = 37.792$!

Recall that our previous regression line was:

$$y = 0.0964 + 0.0798 \cdot x$$

This is not the same regression line!

Step 3: Another Approach

Question: What if we do regression with the roles of x and y swapped?

$$x = 2.977 + 11.605 \cdot y$$

For $y = 3$, we obtain $x = 37.792$!

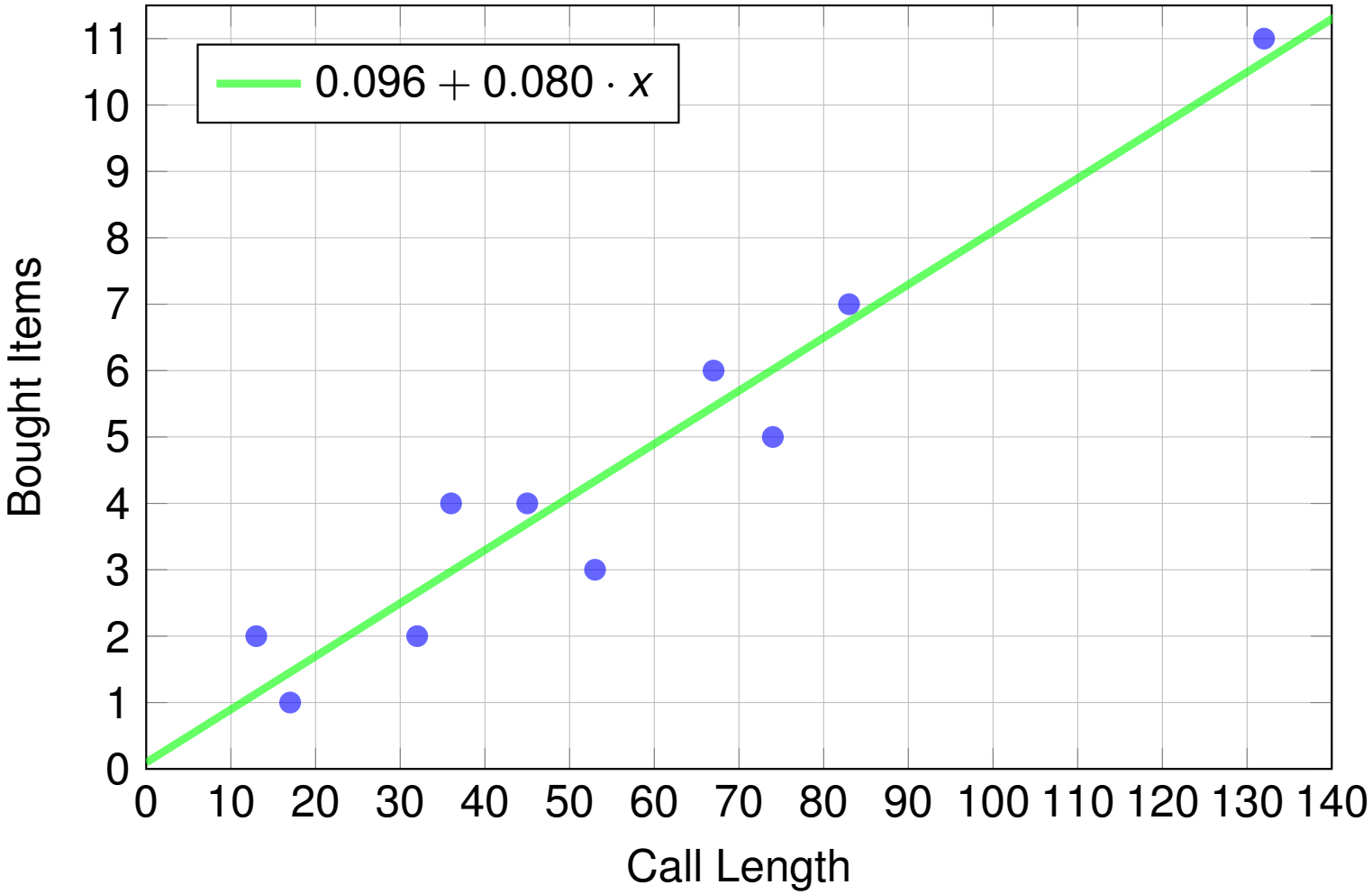
Recall that our previous regression line was:

$$y = 0.0964 + 0.0798 \cdot x$$

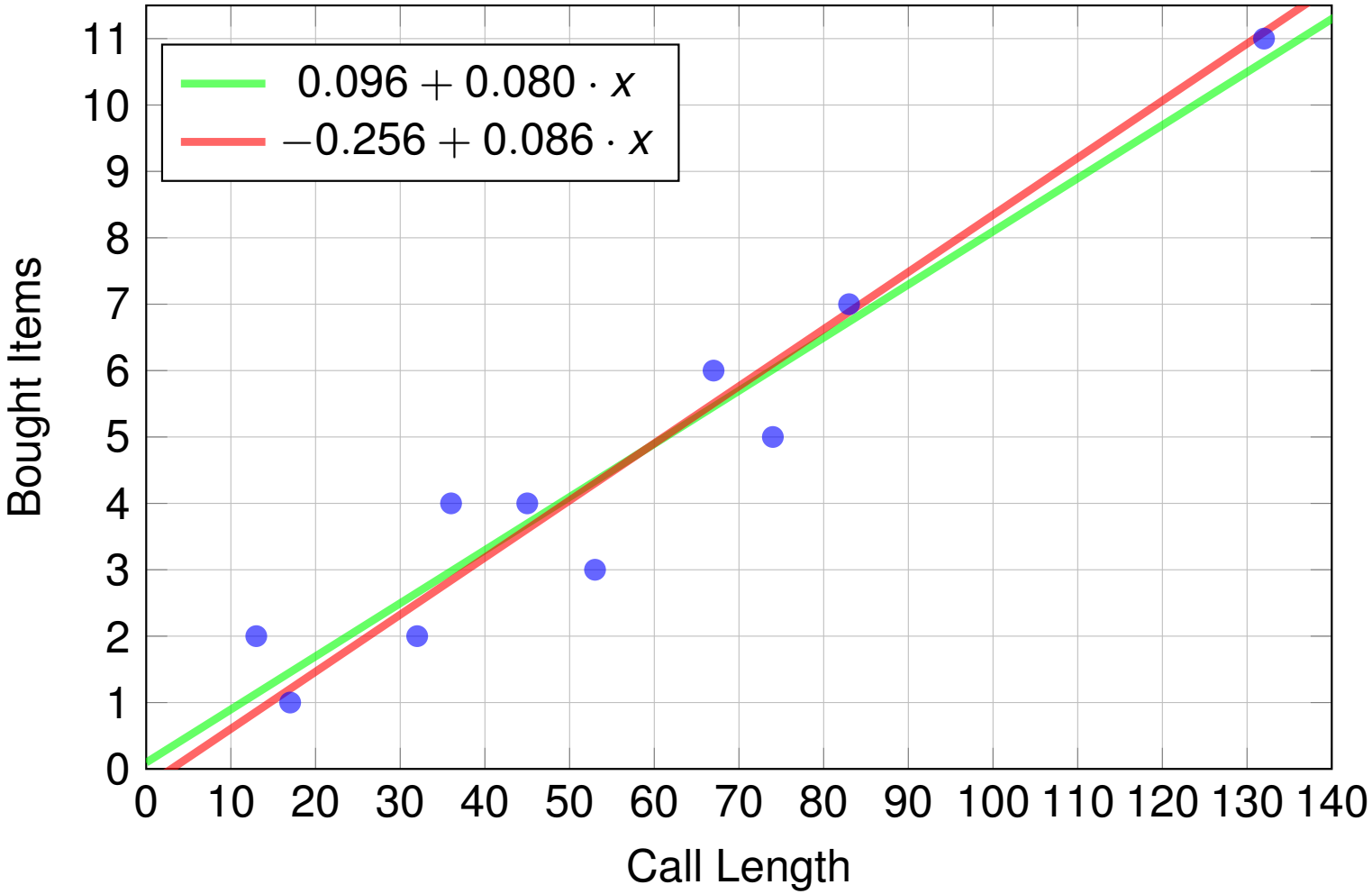
This is not the same regression line!

For $y = 3$, we get $x = 36.397$.

The two Regression Lines

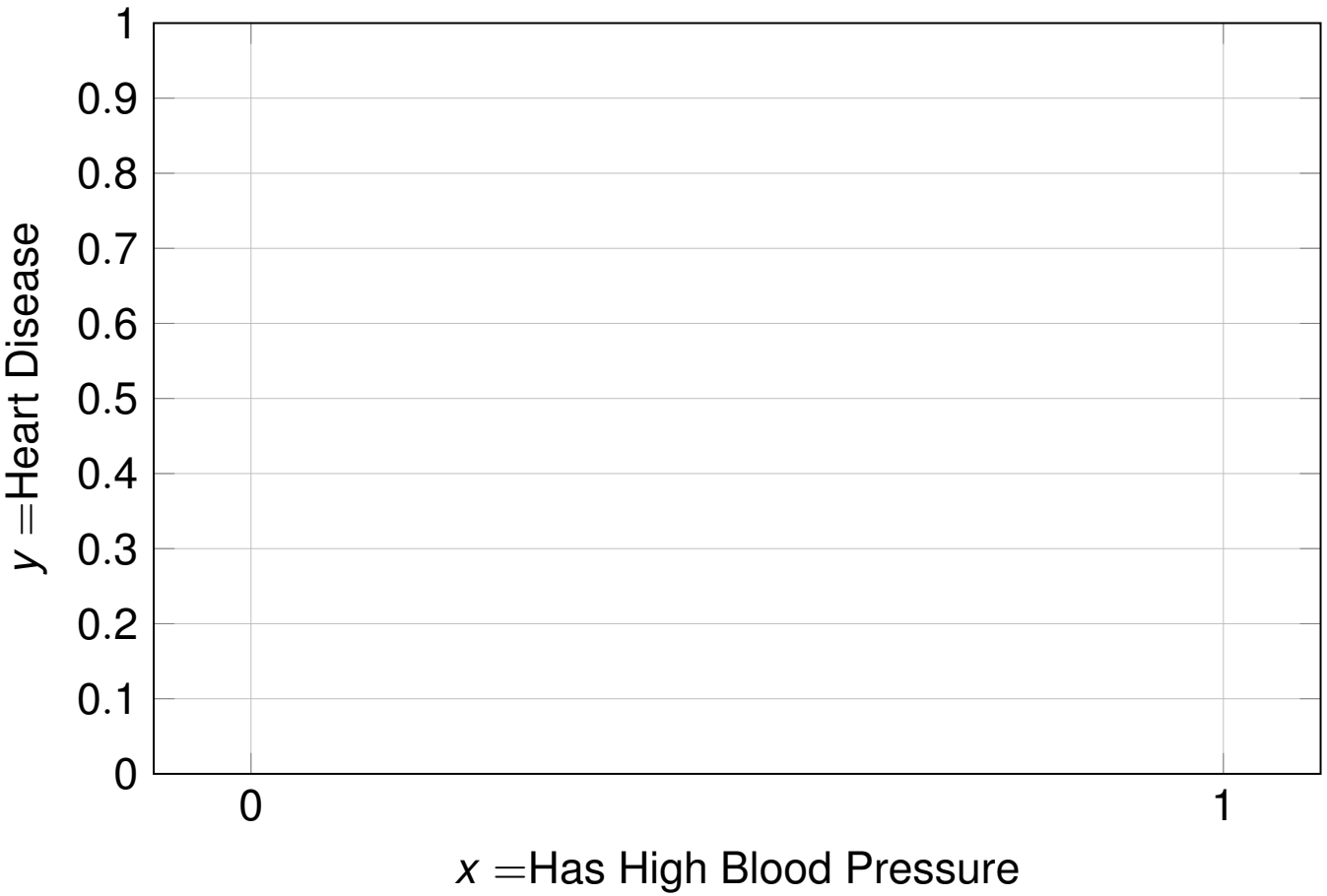


The two Regression Lines



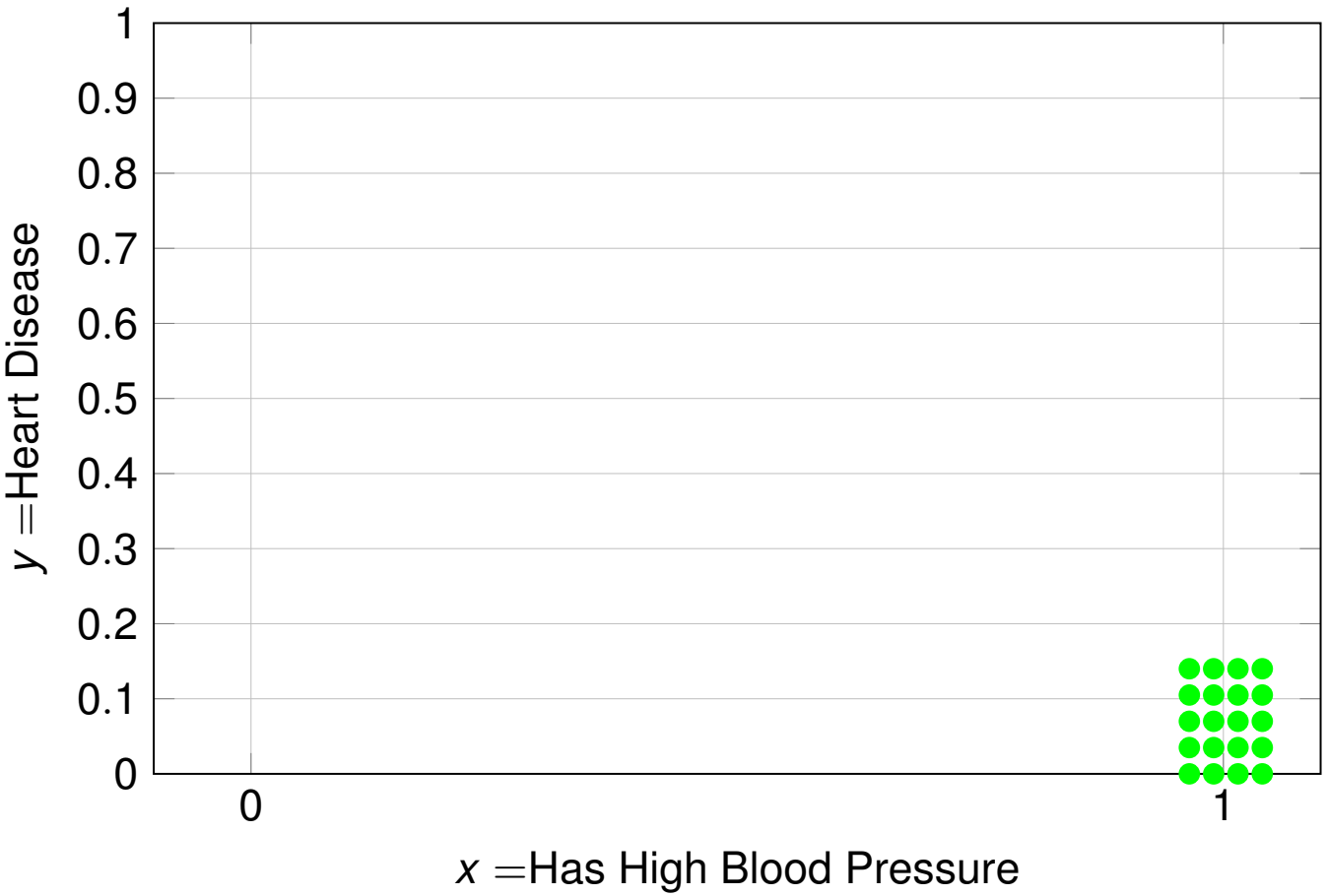
Question 3: Visualising the Data Set

- We have 200 people with **high** blood pressure and no heart disease



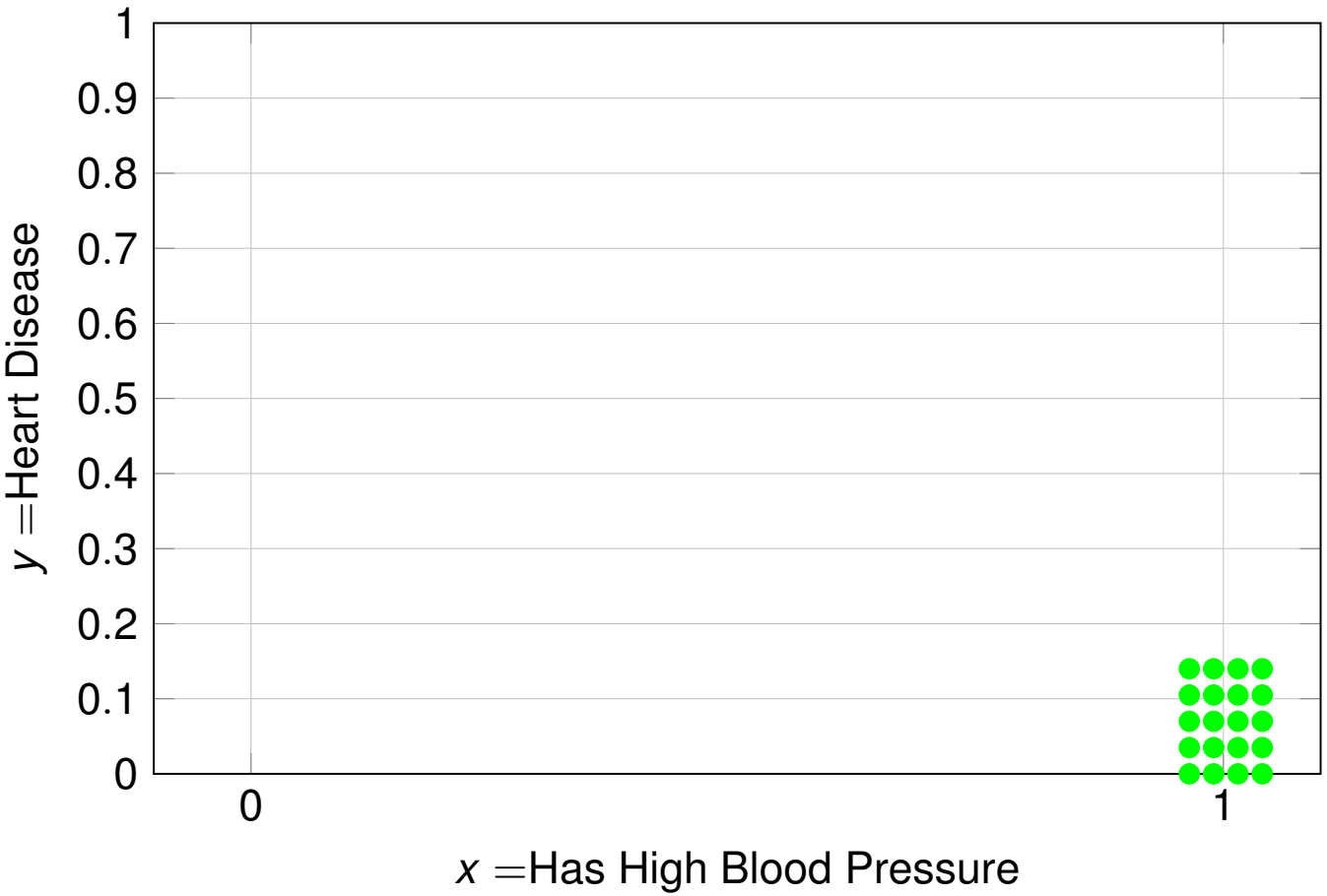
Question 3: Visualising the Data Set

- We have 200 people with **high** blood pressure and no heart disease



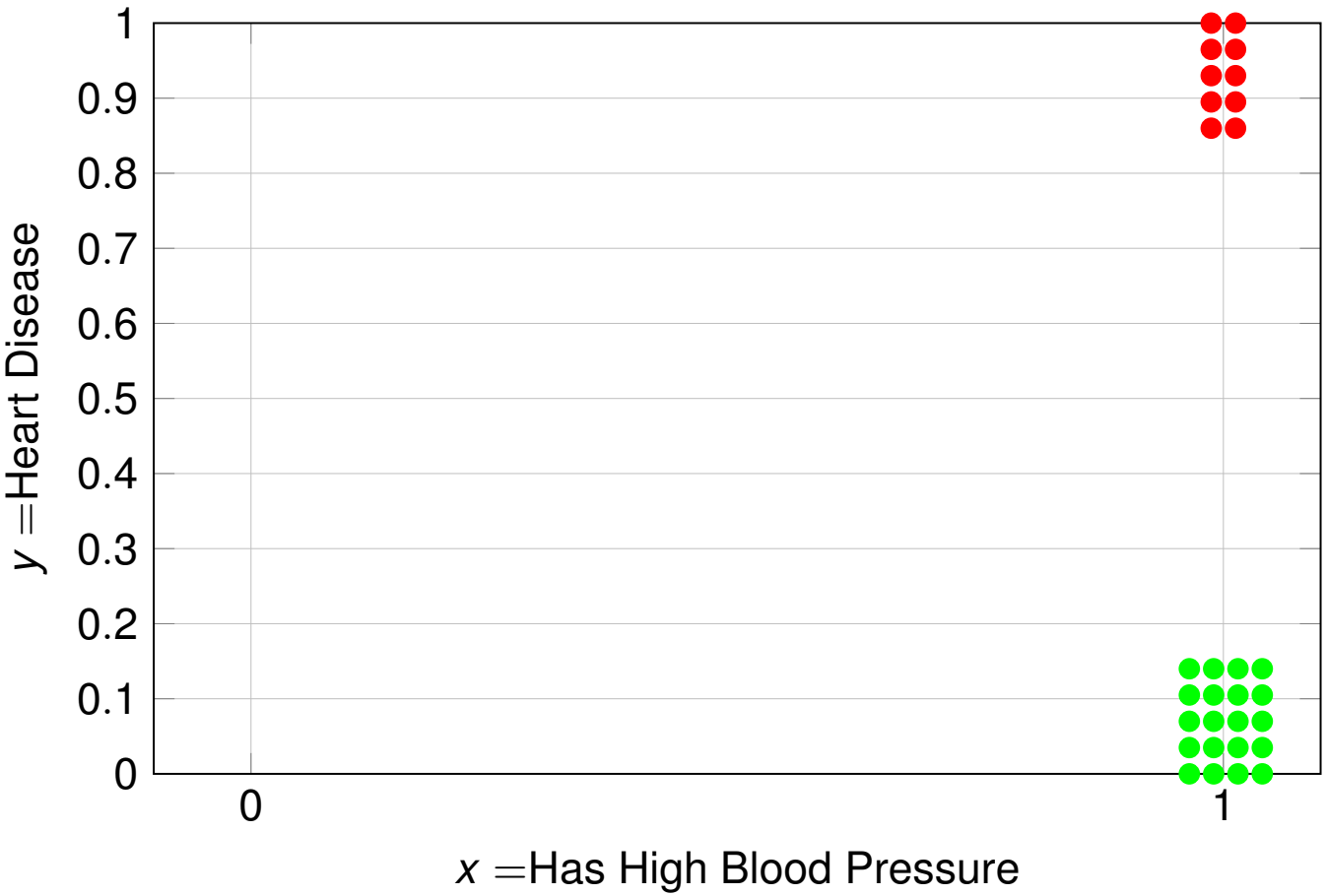
Question 3: Visualising the Data Set

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease



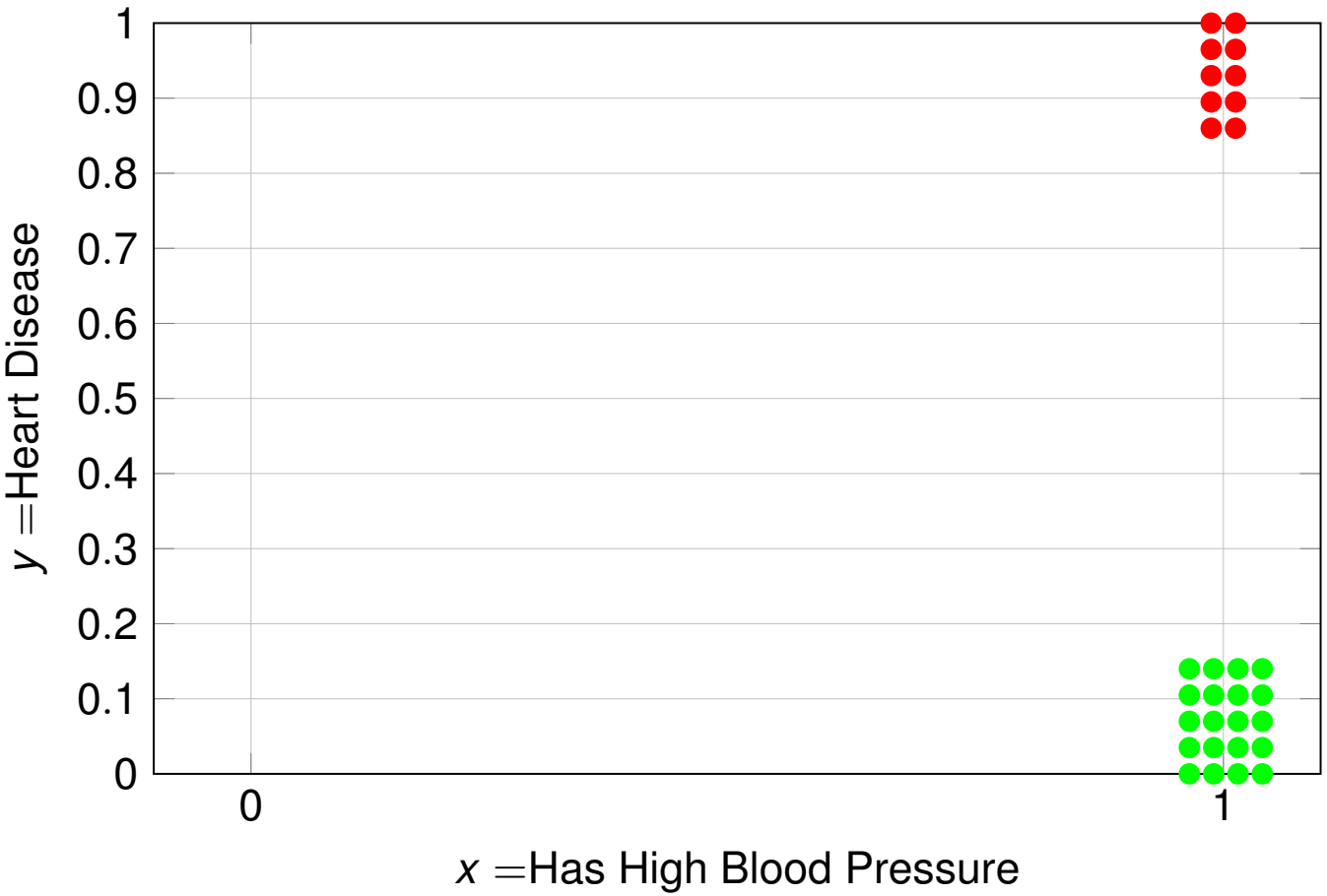
Question 3: Visualising the Data Set

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease



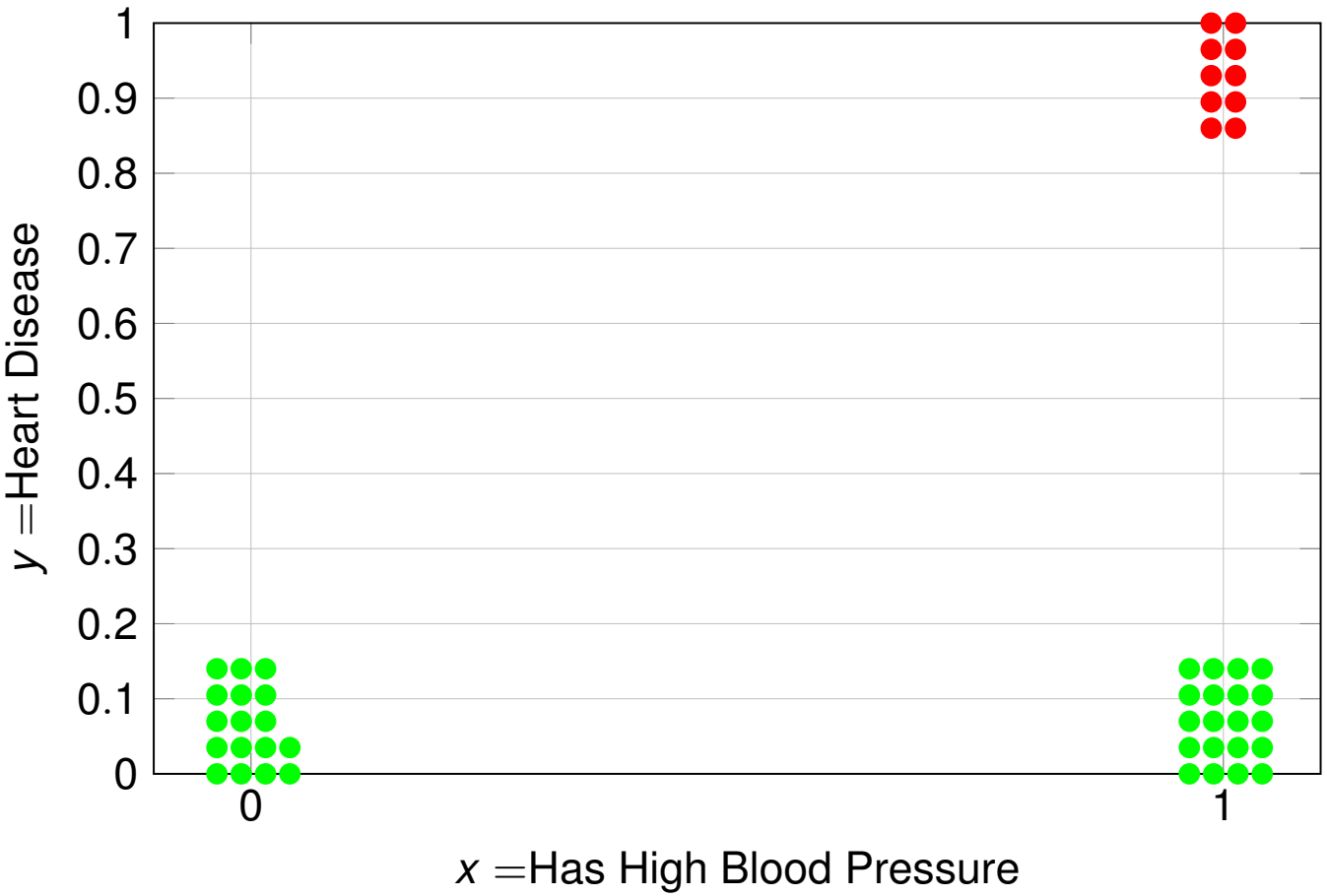
Question 3: Visualising the Data Set

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease
- We have 170 people with **low** blood pressure and no heart disease



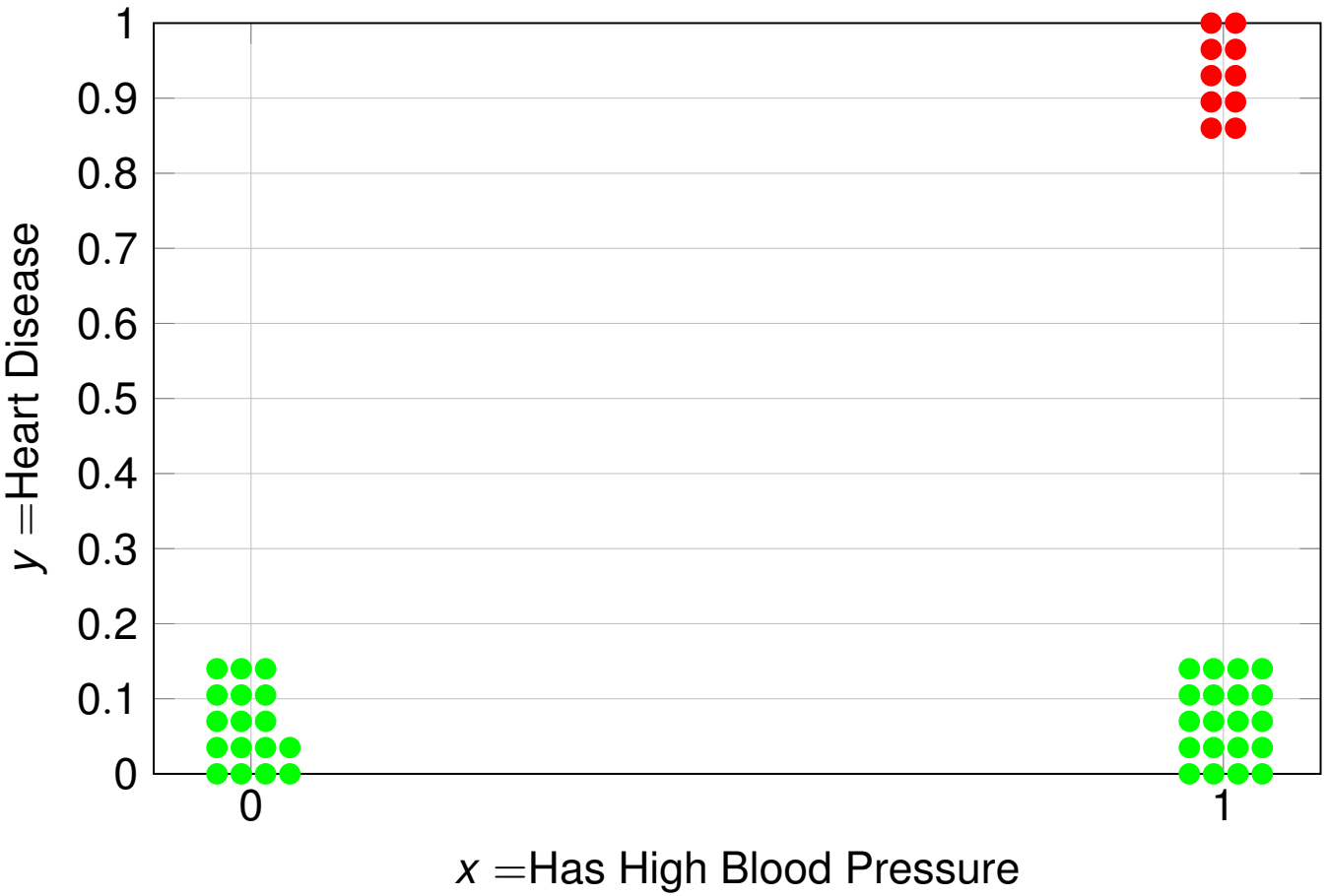
Question 3: Visualising the Data Set

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease
- We have 170 people with **low** blood pressure and no heart disease



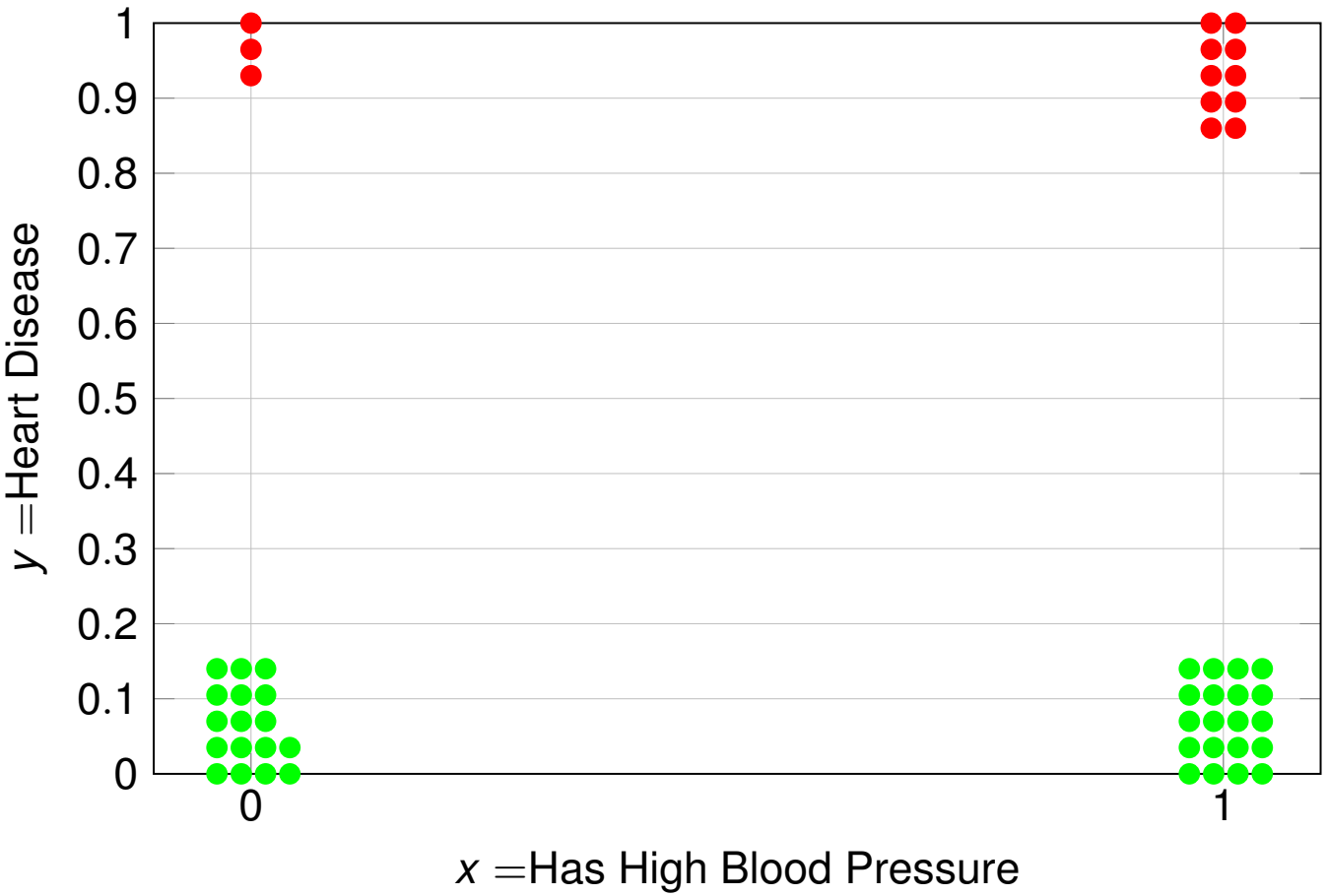
Question 3: Visualising the Data Set

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease
- We have 170 people with **low** blood pressure and no heart disease
- We have 30 people with **low** blood pressure and heart disease



Question 3: Visualising the Data Set

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease
- We have 170 people with **low** blood pressure and no heart disease
- We have 30 people with **low** blood pressure and heart disease



The Optimisation Problem

Our goal is to find w and b so that we maximize likelihood:

The Optimisation Problem

Our goal is to find w and b so that we maximize likelihood:

$$L(w, b) = \left(1 - \frac{1}{1 + \exp\left(-(1 \cdot w + b)\right)}\right)^{200}$$

The Optimisation Problem

Our goal is to find w and b so that we maximize likelihood:

$$L(w, b) = \left(1 - \frac{1}{1 + \exp\left(-(1 \cdot w + b)\right)}\right)^{200} \times \left(\frac{1}{1 + \exp\left(-(1 \cdot w + b)\right)}\right)^{100}$$

The Optimisation Problem

Our goal is to find w and b so that we maximize likelihood:

$$L(w, b) = \left(1 - \frac{1}{1 + \exp\left(-(1 \cdot w + b)\right)}\right)^{200} \times \left(\frac{1}{1 + \exp\left(-(1 \cdot w + b)\right)}\right)^{100} \times \left(1 - \frac{1}{1 + \exp\left(-(0 + b)\right)}\right)^{170}$$

The Optimisation Problem

Our goal is to find w and b so that we maximize likelihood:

$$
\begin{aligned}
L(w,b) = &\left(1 - \frac{1}{1+\exp\left(-(1\cdot w + b)\right)}\right)^{200} \\
&\times \left(\frac{1}{1+\exp\left(-(1\cdot w + b)\right)}\right)^{100} \\
&\times \left(1 - \frac{1}{1+\exp\left(-(0 + b)\right)}\right)^{170} \\
&\times \left(\frac{1}{1+\exp\left(-(0 + b)\right)}\right)^{30}
\end{aligned}
$$

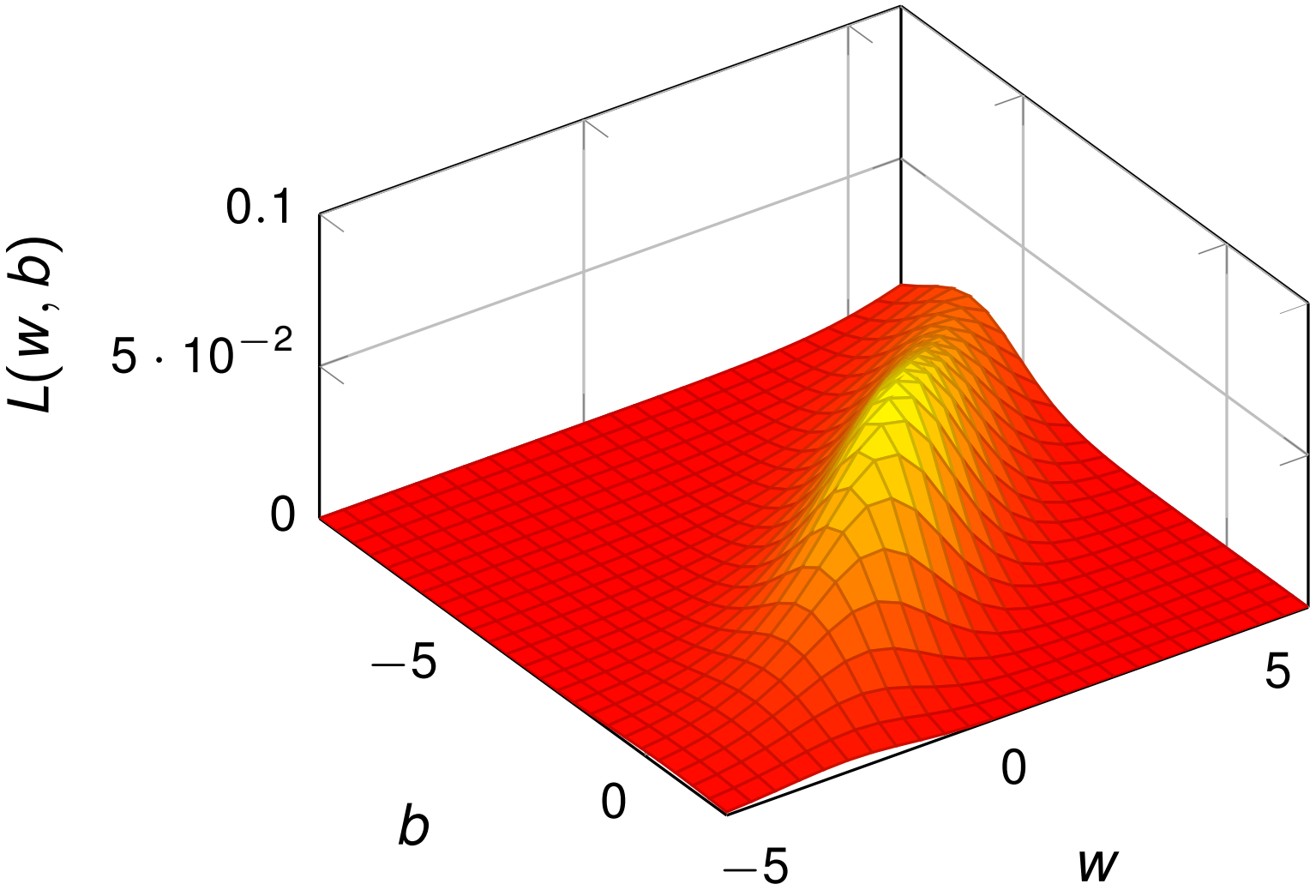
The Optimisation Problem

Our goal is to find w and b so that we maximize likelihood:

$$L(w,b) = \left(1 - \frac{1}{1+\exp\left(-(1\cdot w + b)\right)}\right)^{200} \times \left(\frac{1}{1+\exp\left(-(1\cdot w + b)\right)}\right)^{100} \times \left(1 - \frac{1}{1+\exp\left(-(0 + b)\right)}\right)^{170} \times \left(\frac{1}{1+\exp\left(-(0 + b)\right)}\right)^{30}$$

Apply an iterative solver such as Gradient Descent!

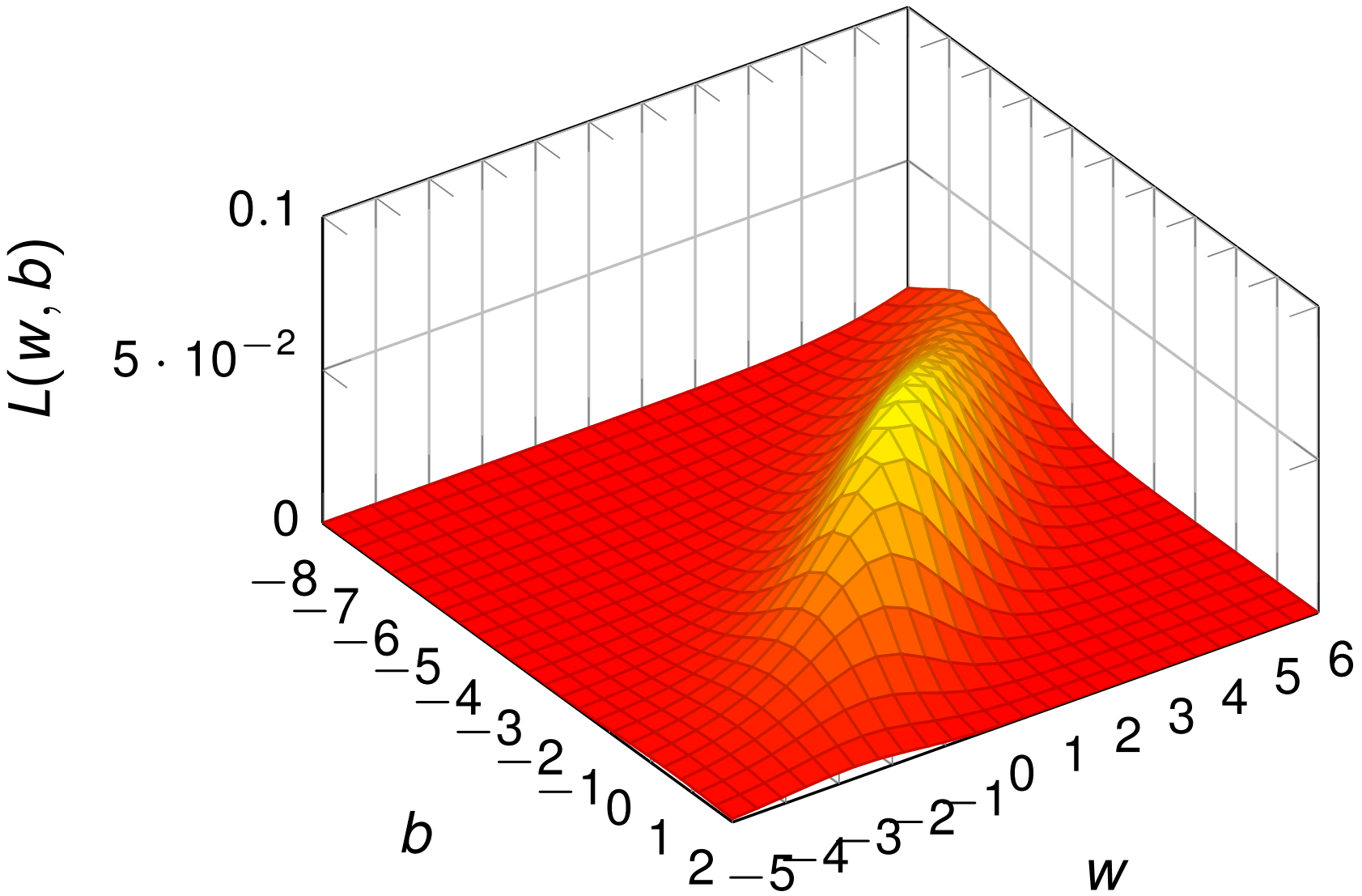
Illustration of the Likelihood Function



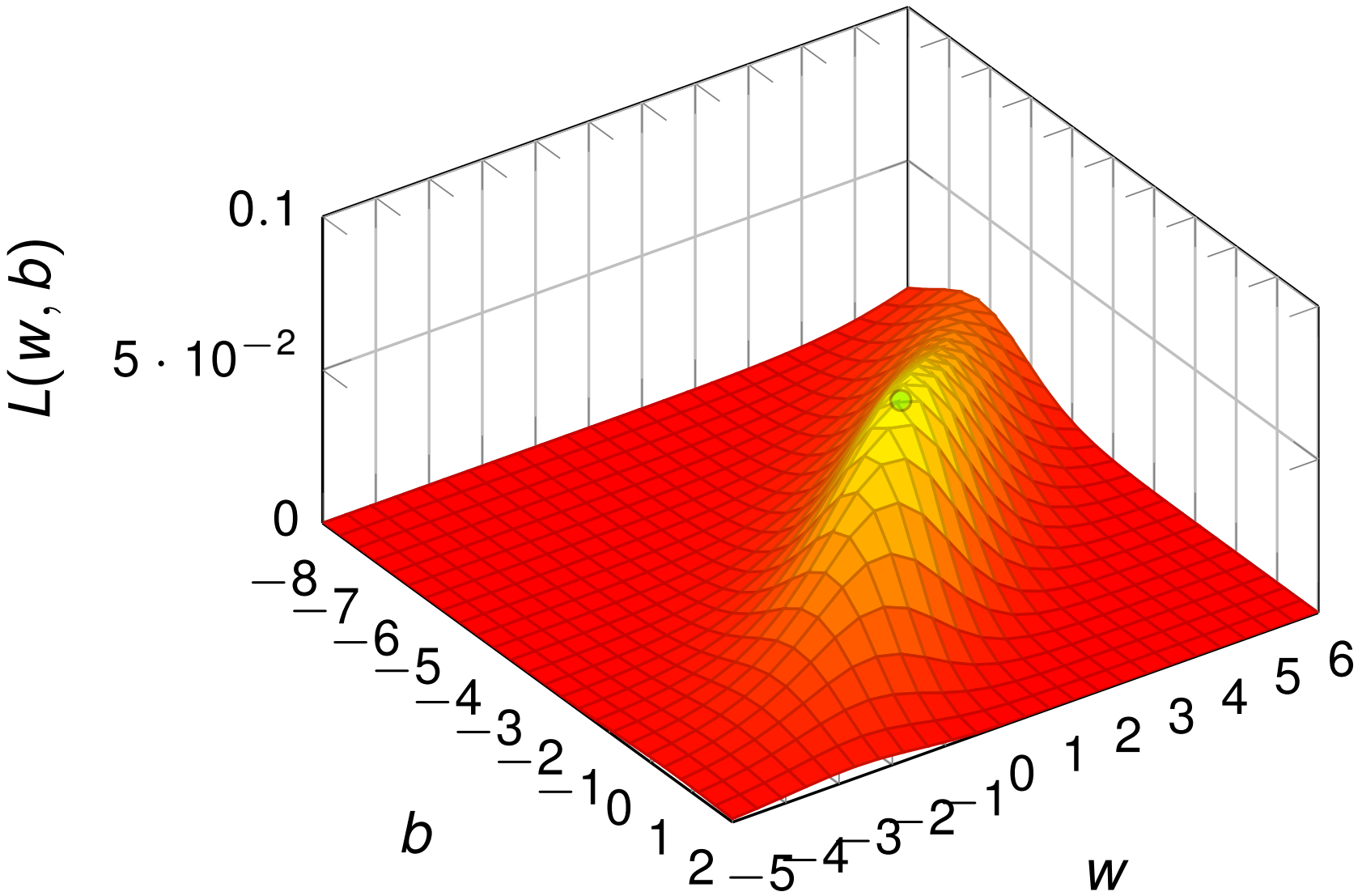
Run of Gradient Descent

Iteration	b	w	Likelihood
1	1	1	0.0012
2	0.93	0.96	0.0015
3	0.847	0.91	0.0021
4	0.734	0.843	0.0031
5	0.575	0.75	0.005
10	-1.14	0.28	0.060
20	-1.62	0.90	0.0635
30	-1.718	1.015	0.0636
40	-1.735	1.036	0.0636
50	-1.738	1.040	0.0636
60	-1.739	1.041	0.0636
100	-1.739	1.041	0.0636

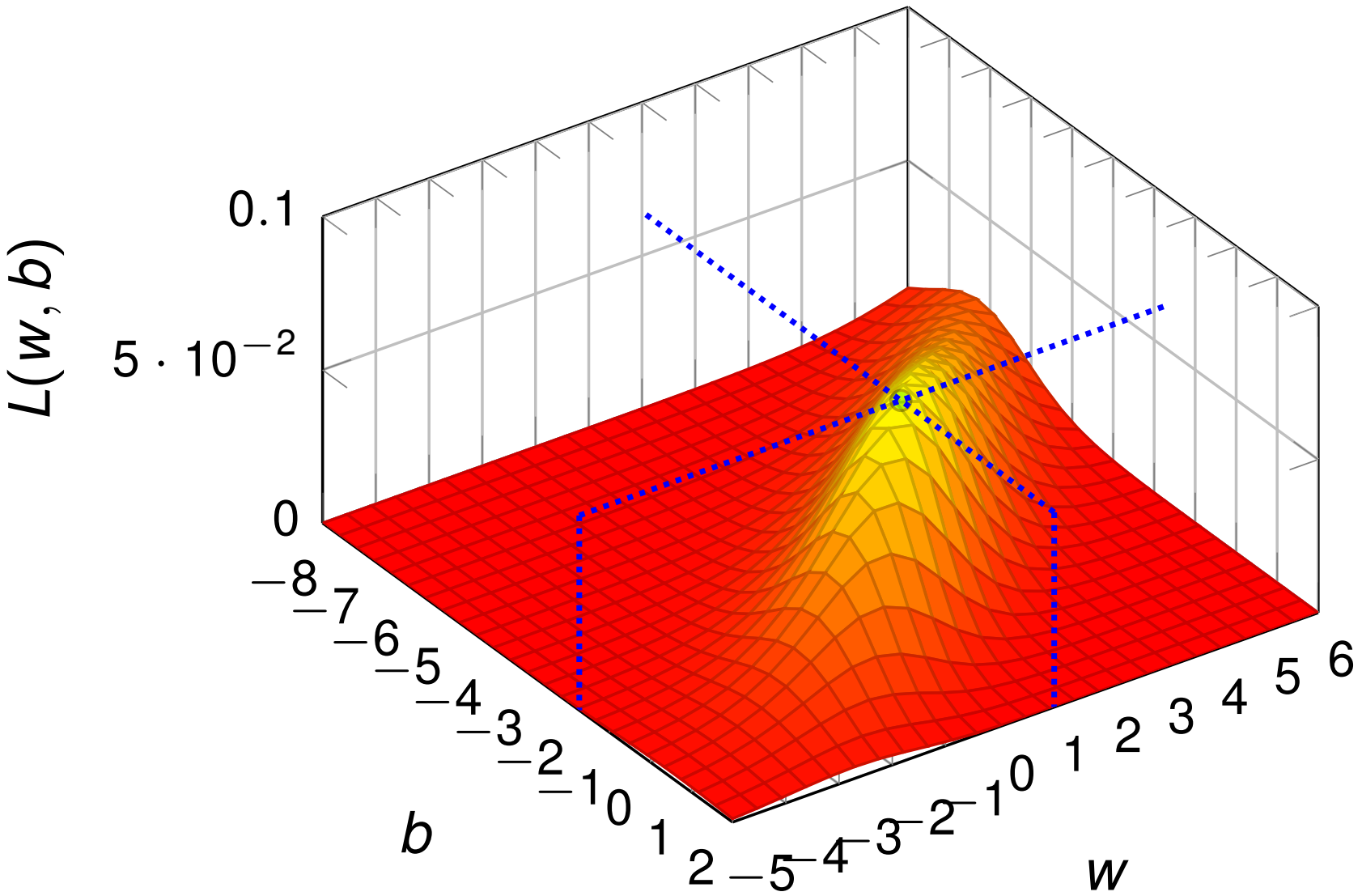
Displaying the Optimum



Displaying the Optimum



Displaying the Optimum

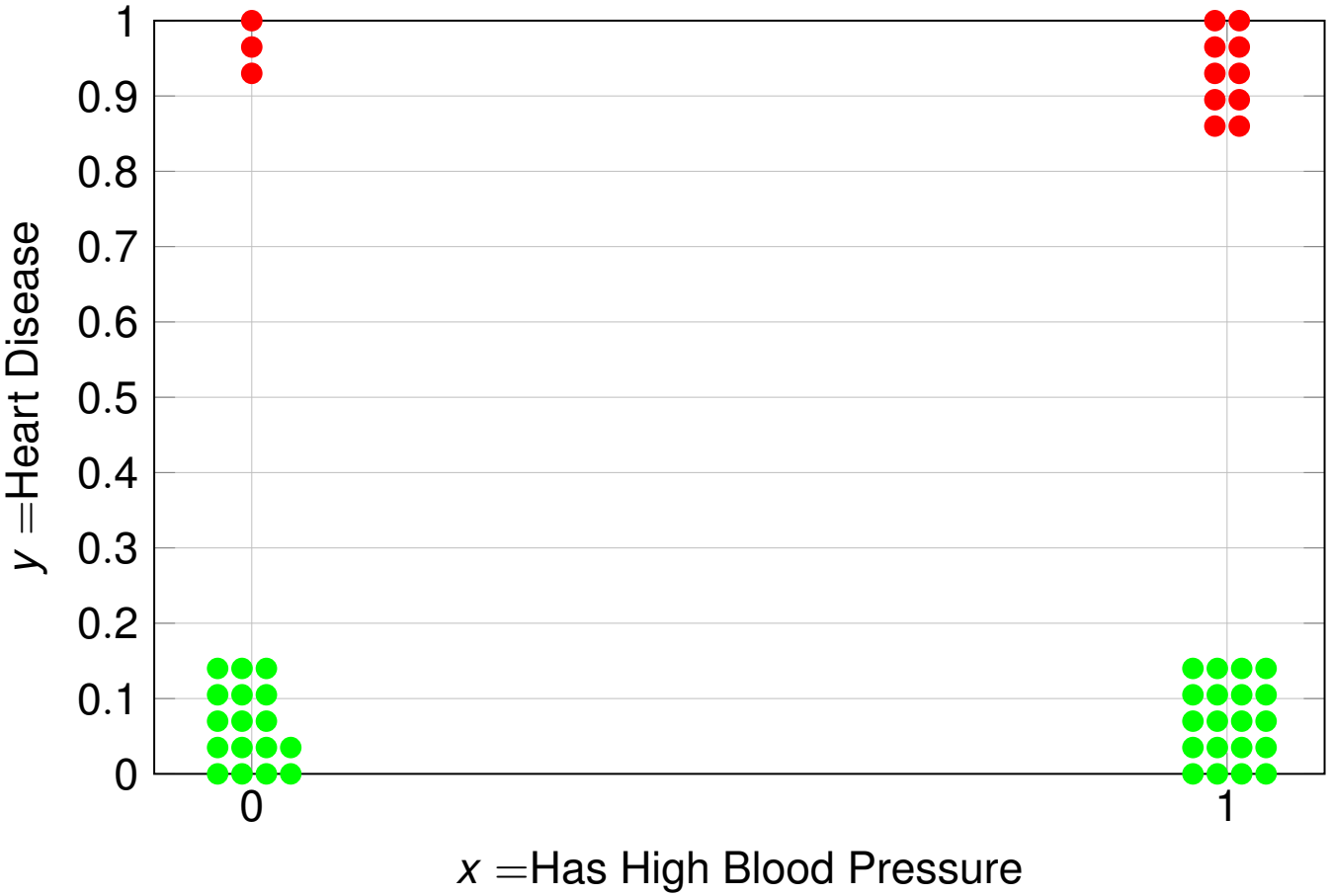


Solution of Logistic Regression

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease
- We have 170 people with **low** blood pressure and no heart disease
- We have 30 people with **low** blood pressure and heart disease

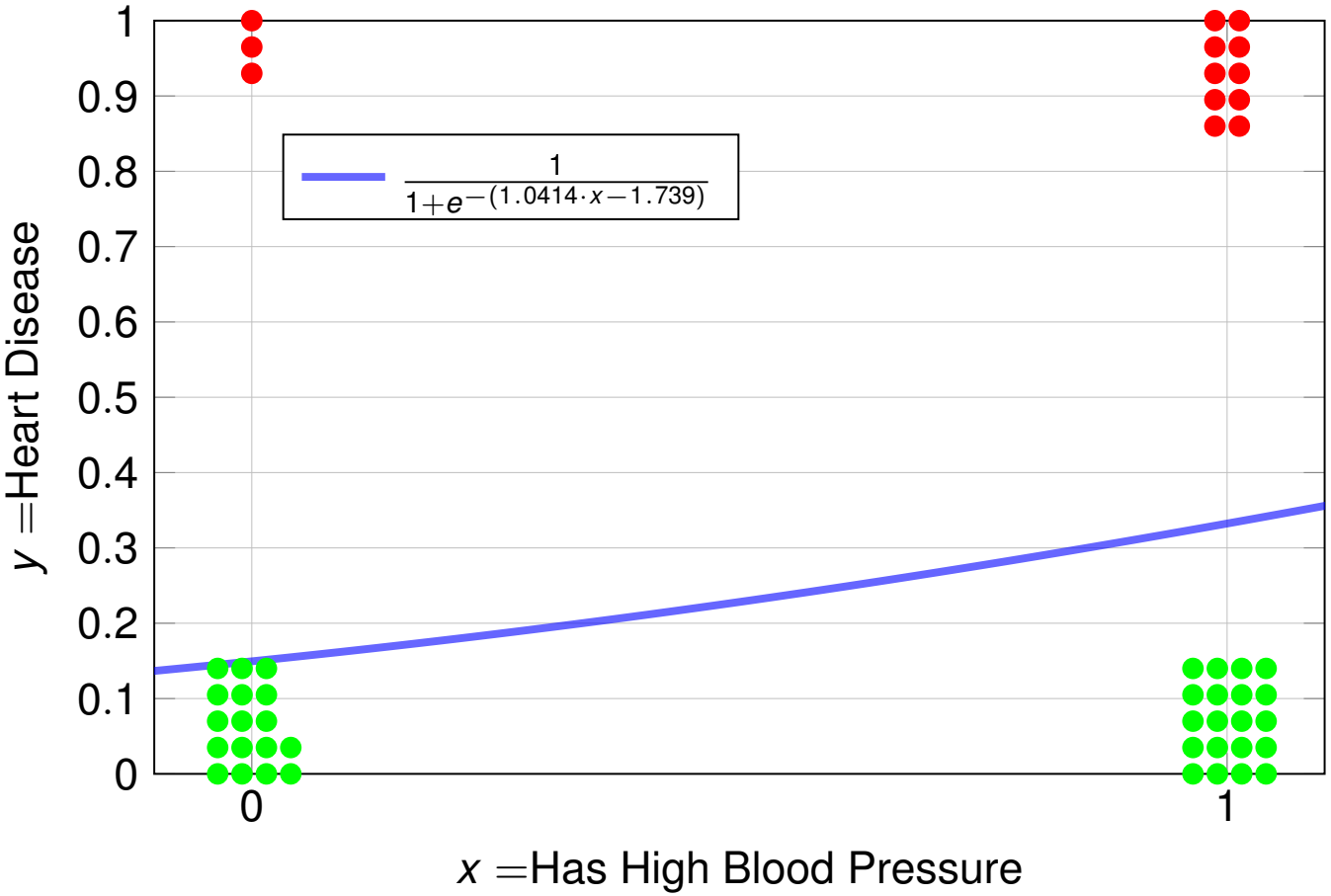
Solution of Logistic Regression

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease
- We have 170 people with **low** blood pressure and no heart disease
- We have 30 people with **low** blood pressure and heart disease



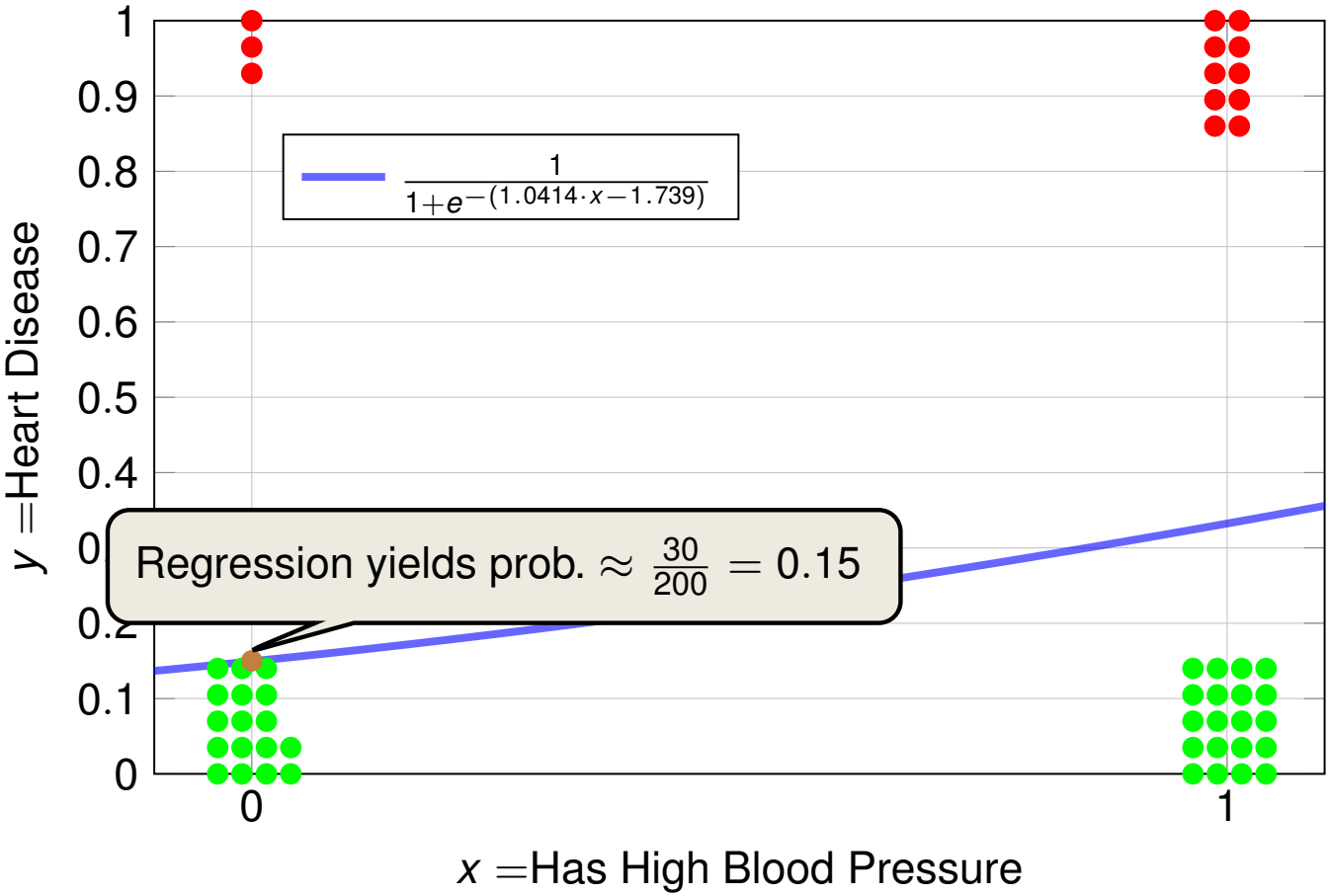
Solution of Logistic Regression

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease
- We have 170 people with **low** blood pressure and no heart disease
- We have 30 people with **low** blood pressure and heart disease



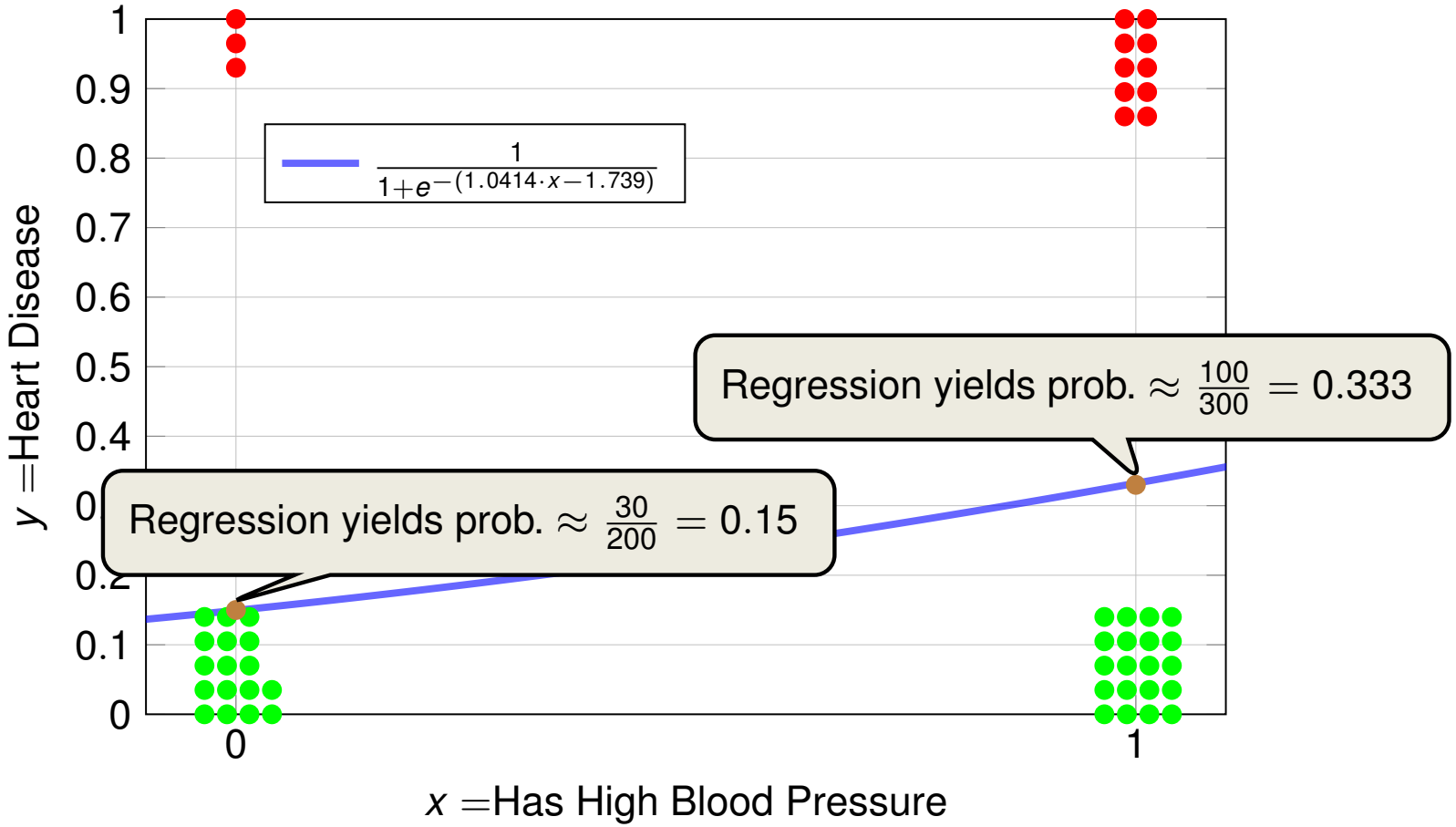
Solution of Logistic Regression

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease
- We have 170 people with **low** blood pressure and no heart disease
- We have 30 people with **low** blood pressure and heart disease



Solution of Logistic Regression

- We have 200 people with **high** blood pressure and no heart disease
- We have 100 people with **high** blood pressure and heart disease
- We have 170 people with **low** blood pressure and no heart disease
- We have 30 people with **low** blood pressure and heart disease



A Simple Example

Experiment

A Simple Example

Experiment

- Goal: predict exam grade of a student based on their learning habits

A Simple Example

Experiment

- Goal: predict exam grade of a student based on their learning habits
- You ask each student to count the number of hours watching TV

A Simple Example

Experiment

- Goal: predict exam grade of a student based on their learning habits
- You ask each student to count the number of hours watching TV
- Then record the exam grade (a number between 0 and 10)

A Simple Example

Experiment

- Goal: predict exam grade of a student based on their learning habits
- You ask each student to count the number of hours watching TV
- Then record the exam grade (a number between 0 and 10)

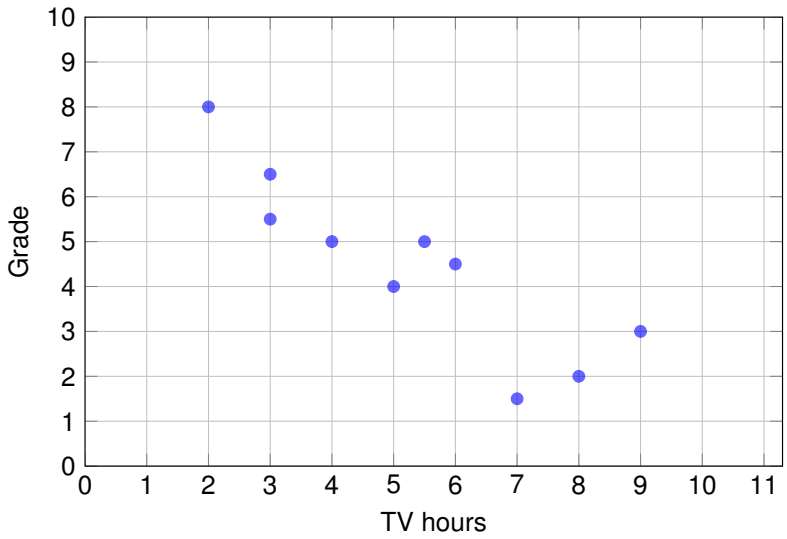
Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3

A Simple Example

Experiment

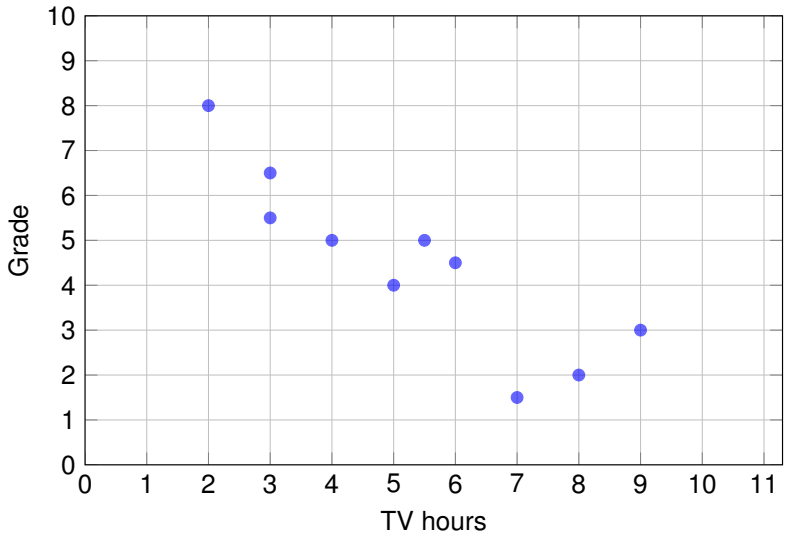
- Goal: predict exam grade of a student based on their learning habits
- You ask each student to count the number of hours watching TV
- Then record the exam grade (a number between 0 and 10)

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3



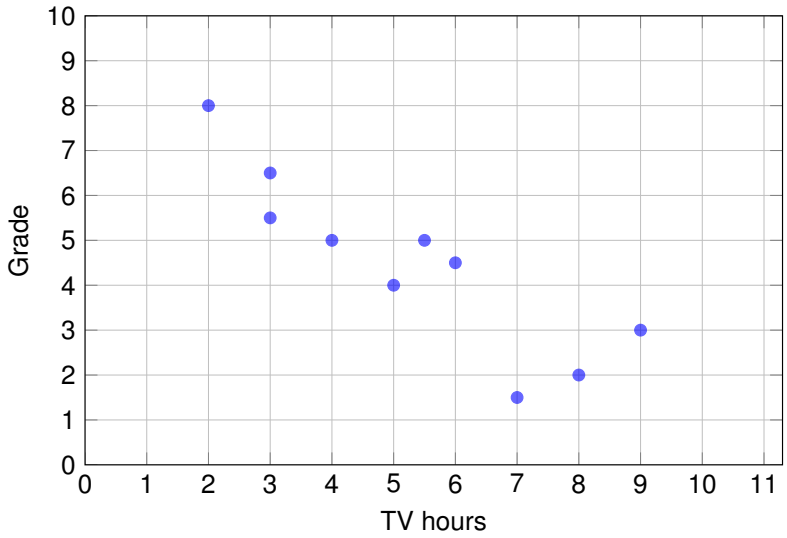
Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3



Simple Example continued

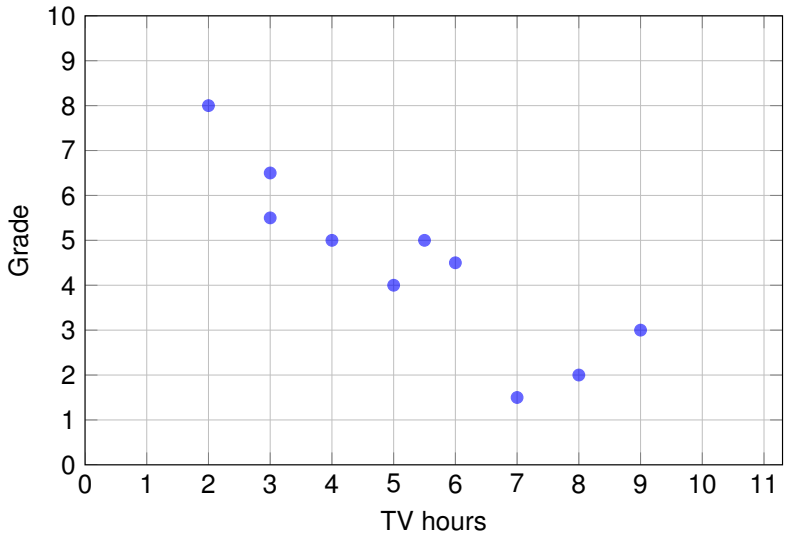
Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3



Comments

Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3

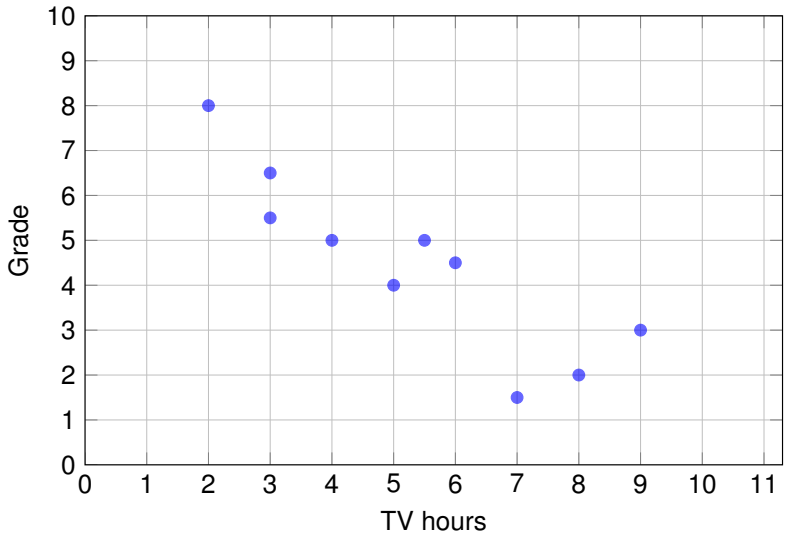


Comments

- Two data points with the same TV hours, but different exam grades

Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3

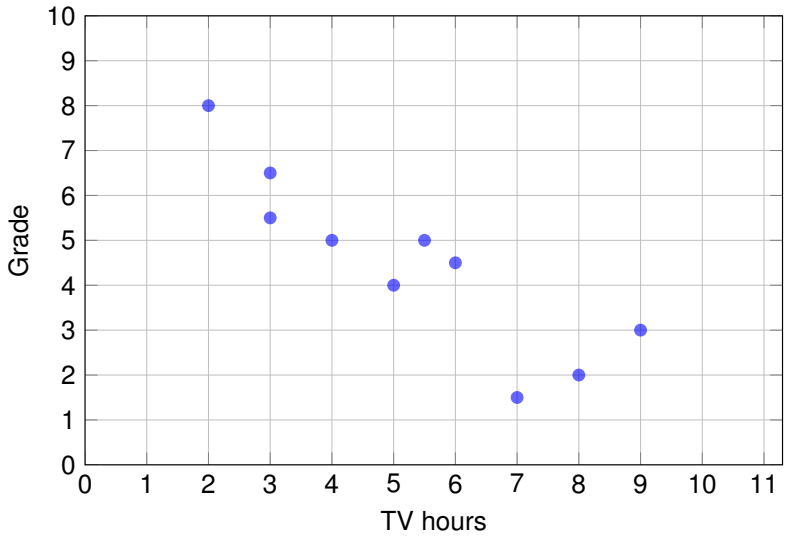


Comments

- Two data points with the same TV hours, but different exam grades
- The student IDs (1,2,3,. . .) should not be used for the data analysis

Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3

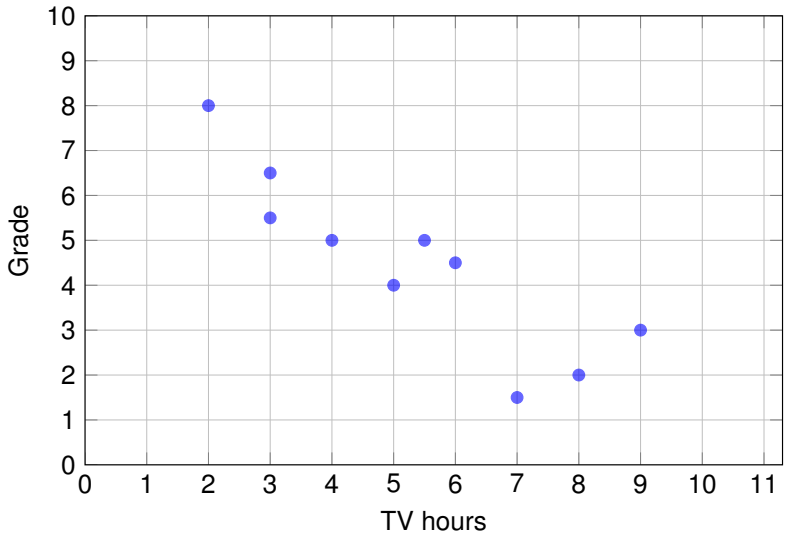


Comments

- Two data points with the same TV hours, but different exam grades
- The student IDs (1,2,3,. . .) should not be used for the data analysis
- Data set is quite small:
  - low-dimensional: only one feature
  - few examples: only 10 data points

Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3



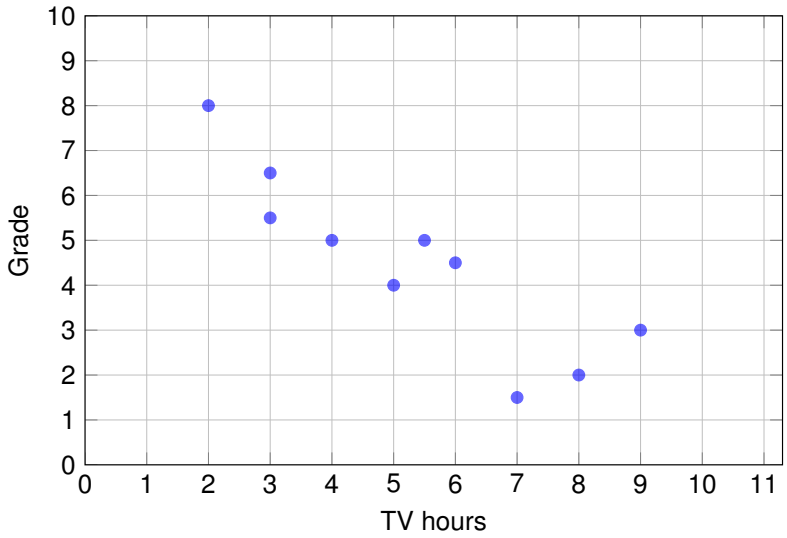
Comments

- Two data points with the same TV hours, but different exam grades
- The student IDs (1,2,3,. . .) should not be used for the data analysis
- Data set is quite small:
  - low-dimensional: only one feature
  - few examples: only 10 data points

Any prediction comes with risk and potential error!

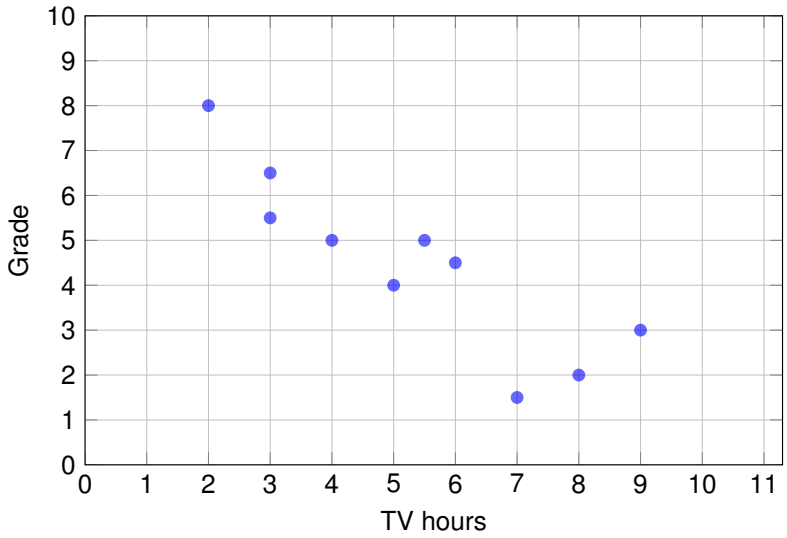
Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3



Simple Example continued

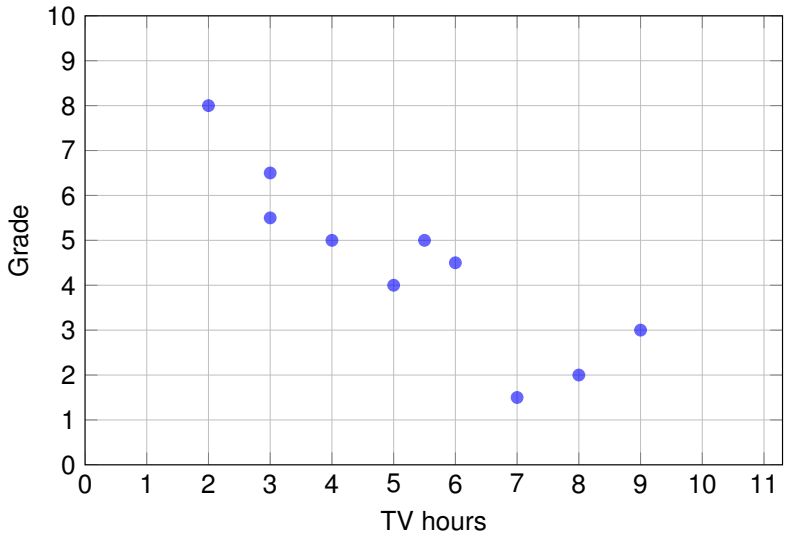
Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3



Ways to Proceed

Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3

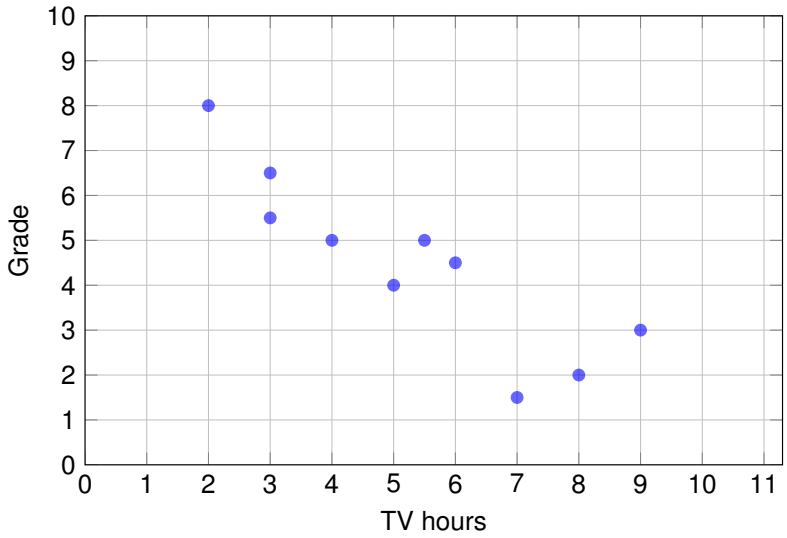


Ways to Proceed

- We could do a statistical analysis (e.g., compute covariance or correlation coefficient)

Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3

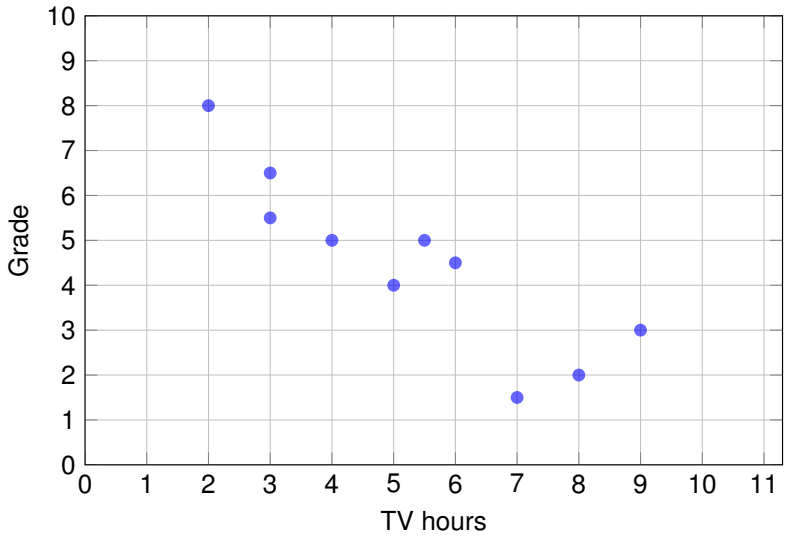


Ways to Proceed

- We could do a statistical analysis (e.g., compute covariance or correlation coefficient)
- In Machine Learning, we are usually interested in **predicting** unseen/unclassified data points

Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	??

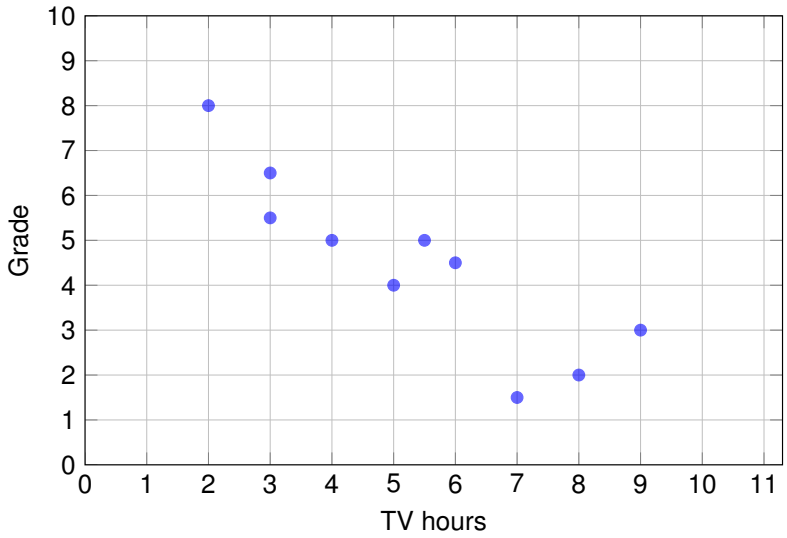


Ways to Proceed

- We could do a statistical analysis (e.g., compute covariance or correlation coefficient)
- In Machine Learning, we are usually interested in **predicting** unseen/unclassified data points

Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	??



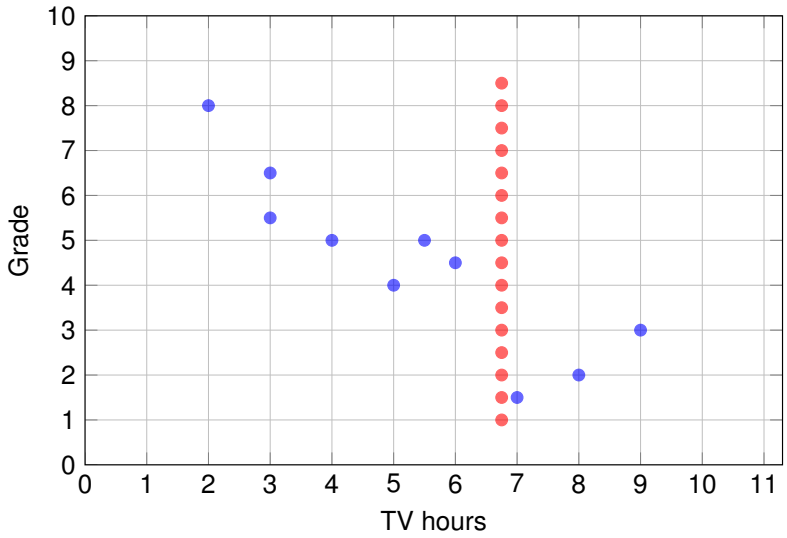
Ex 1: Guess grade for Student 11

Ways to Proceed

- We could do a statistical analysis (e.g., compute covariance or correlation coefficient)
- In Machine Learning, we are usually interested in **predicting** unseen/unclassified data points

Simple Example continued

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	??



Ex 1: Guess grade for Student 11

Ways to Proceed

- We could do a statistical analysis (e.g., compute covariance or correlation coefficient)
- In Machine Learning, we are usually interested in **predicting** unseen/unclassified data points

Outline

Introduction

Linear Regression

Linear Regression with More than One Feature

Solving Linear Regression: Gradient Descent

Logistic Regression

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!

  $\Rightarrow$ Either the more TV hours, the better the exam grade
  Or the more TV hours, the worse the exam grade

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!
  - $\Rightarrow$ Either the more TV hours, the better the exam grade
    Or the more TV hours, the worse the exam grade

$$\boxed{y = b + w \cdot x}$$

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!
  - $\Rightarrow$ Either the more TV hours, the better the exam grade
    Or the more TV hours, the worse the exam grade

$$\boxed{y = b + w \cdot x}$$

- x is an explanatory variable/feature

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!
  - $\Rightarrow$ Either the more TV hours, the better the exam grade
    Or the more TV hours, the worse the exam grade

$$\boxed{y = b + w \cdot x}$$

- x is an explanatory variable/feature (number of TV hours)

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!
  - $\Rightarrow$ Either the more TV hours, the better the exam grade
    Or the more TV hours, the worse the exam grade

$$\boxed{y = b + w \cdot x}$$

- x is an explanatory variable/feature (number of TV hours)
- y is the outcome/response

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!
  - $\Rightarrow$ Either the more TV hours, the better the exam grade
    Or the more TV hours, the worse the exam grade

$$\boxed{y = b + w \cdot x}$$

- x is an explanatory variable/feature (number of TV hours)
- y is the outcome/response (exam grade)

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!
  - $\Rightarrow$ Either the more TV hours, the better the exam grade
    Or the more TV hours, the worse the exam grade

$$\boxed{y = b + w \cdot x}$$

- x is an explanatory variable/feature (number of TV hours)
- y is the outcome/response (exam grade)
- w is the slope

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!
  - $\Rightarrow$ Either the more TV hours, the better the exam grade ($w > 0$)
    Or the more TV hours, the worse the exam grade

$$\boxed{y = b + w \cdot x}$$

- x is an explanatory variable/feature (number of TV hours)
- y is the outcome/response (exam grade)
- w is the slope

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!
  - $\Rightarrow$ Either the more TV hours, the better the exam grade $(w > 0)$
    Or the more TV hours, the worse the exam grade $(w < 0)$

$$\boxed{y = b + w \cdot x}$$

- x is an explanatory variable/feature (number of TV hours)
- y is the outcome/response (exam grade)
- w is the slope

Idea of Linear Regression

- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!
  - $\Rightarrow$ Either the more TV hours, the better the exam grade $(w > 0)$
    Or the more TV hours, the worse the exam grade $(w < 0)$

$$\boxed{y = b + w \cdot x}$$

- x is an explanatory variable/feature (number of TV hours)
- y is the outcome/response (exam grade)
- w is the slope
- b is the intercept

Idea of Linear Regression

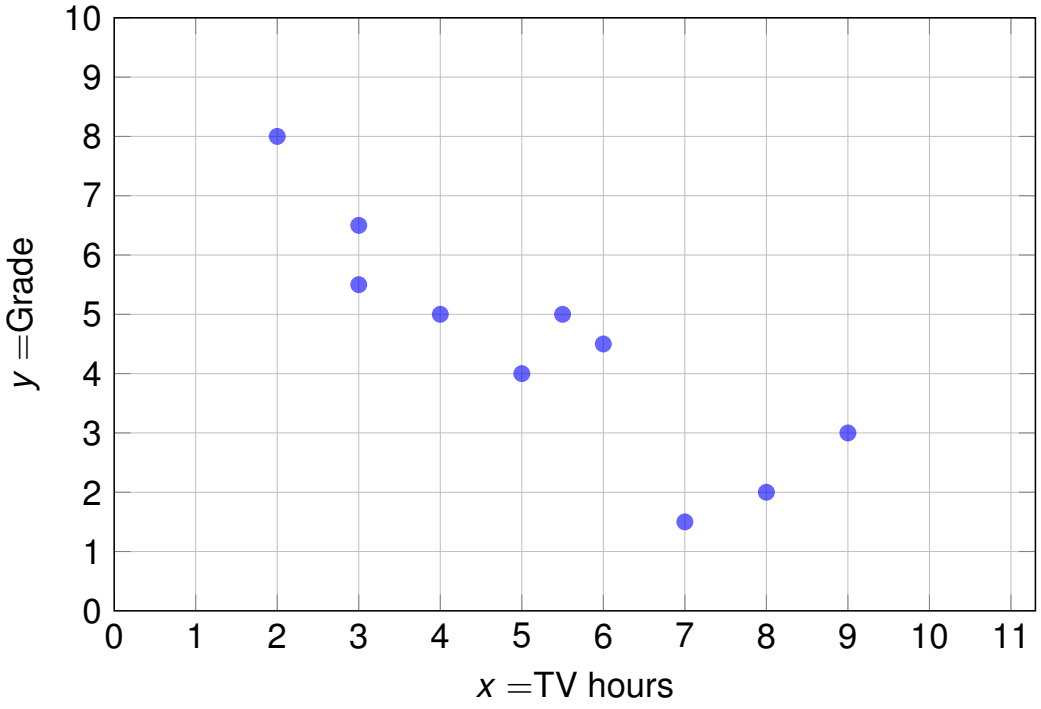
- We are interested in using the TV hours to predict the Exam Grade
- Regression assumes that there is a linear relation/trend (simple model)!
  - $\Rightarrow$ Either the more TV hours, the better the exam grade $(w > 0)$
    Or the more TV hours, the worse the exam grade $(w < 0)$

$$y = b + w \cdot x$$

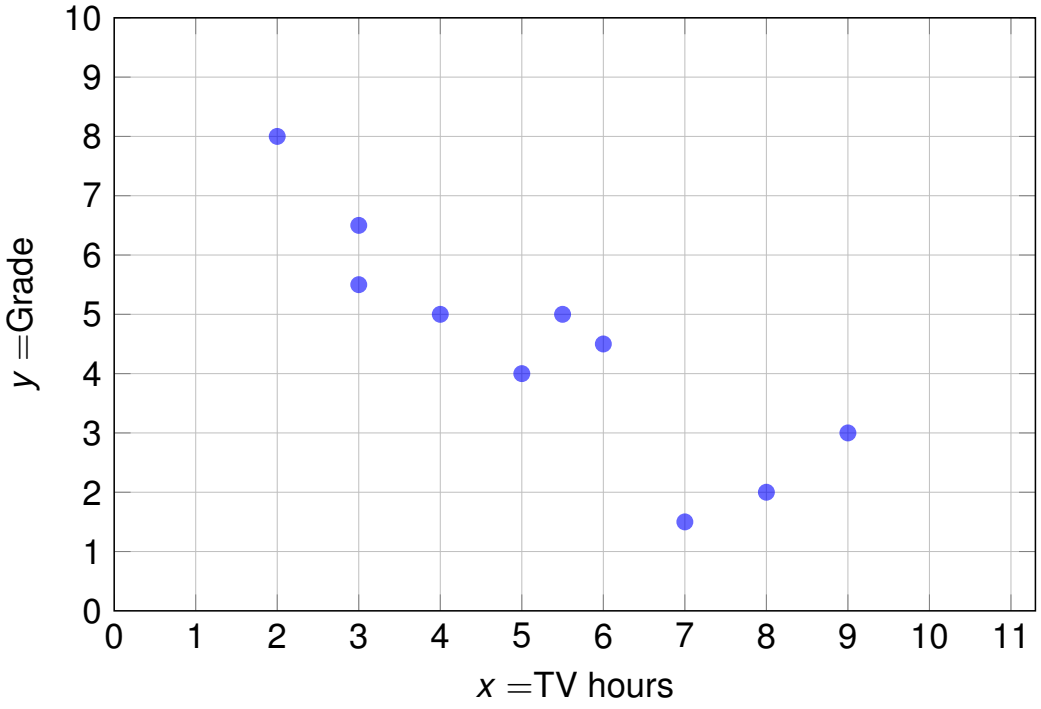
- x is an explanatory variable/feature (number of TV hours)
- y is the outcome/response (exam grade)
- w is the slope
- b is the intercept

Goal: Find b and w so that $y \approx b + w \cdot x$ for most points.

Example of different Linear Models

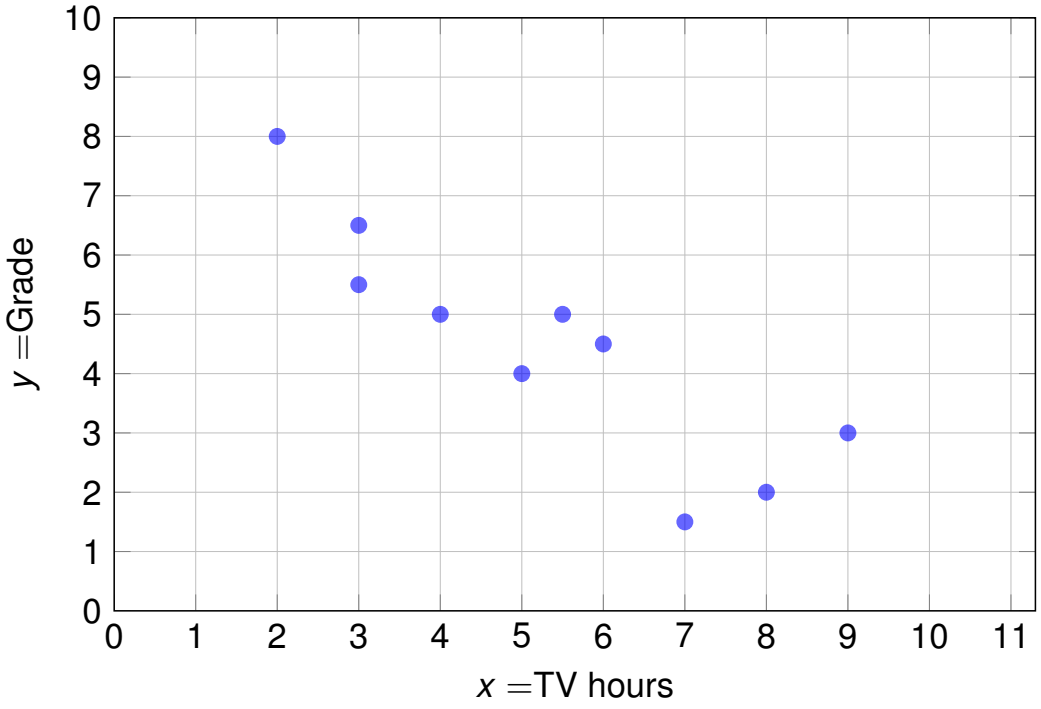


Example of different Linear Models



$$y = b + w \cdot x$$

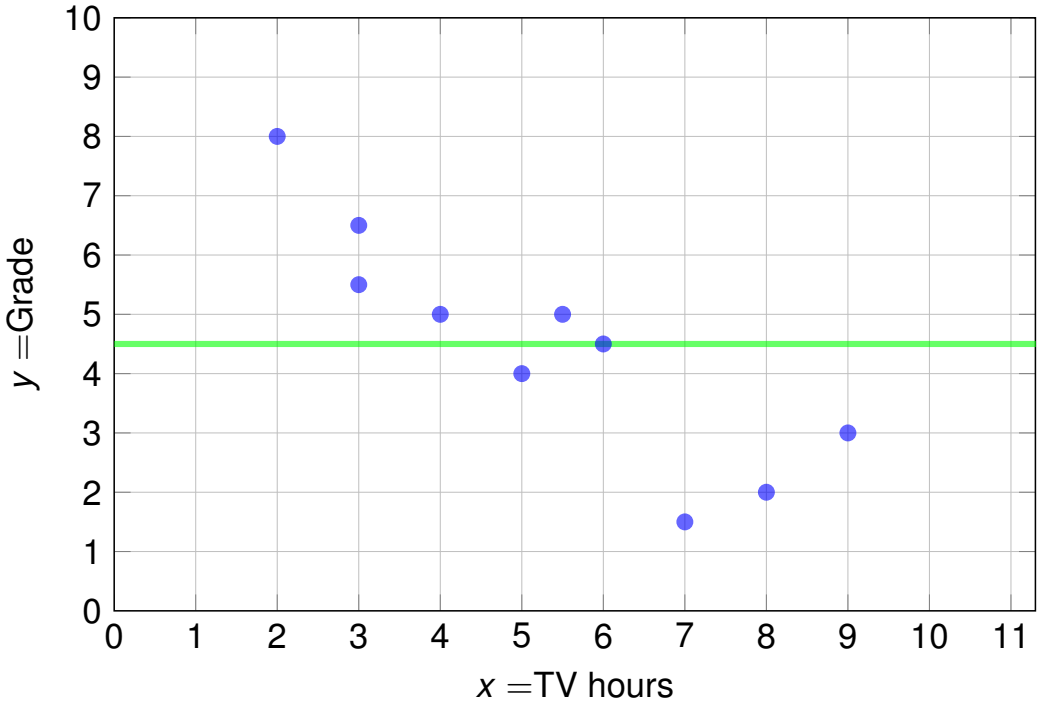
Example of different Linear Models



$$y = b + w \cdot x$$

1. $y = 4.5 + 0 \cdot x$

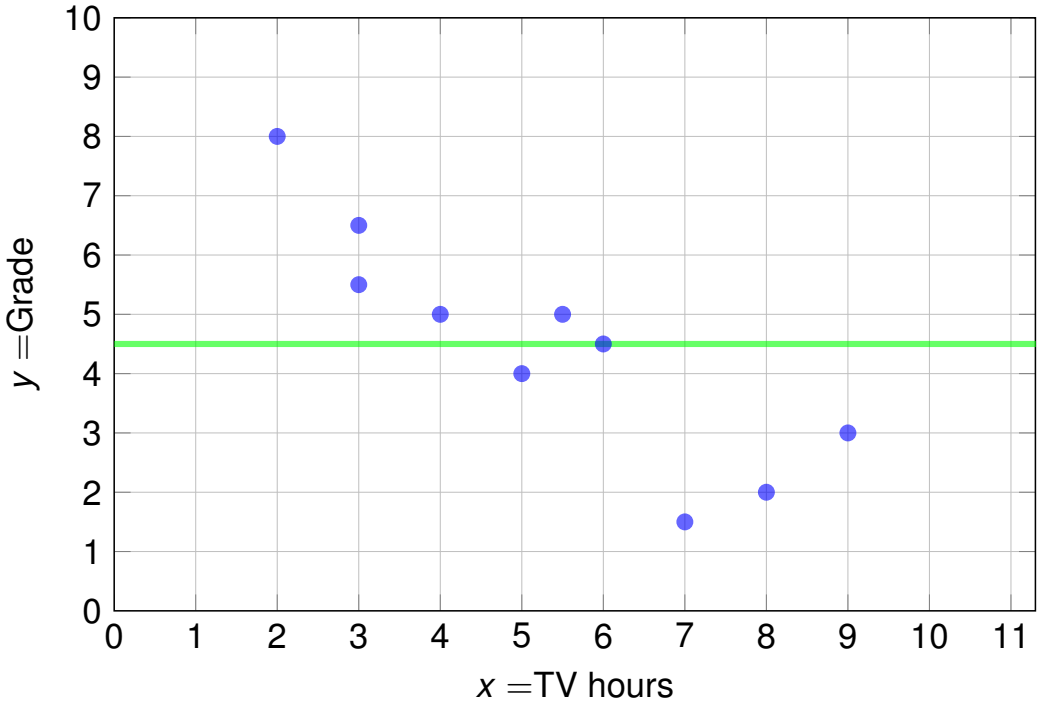
Example of different Linear Models



$$y = b + w \cdot x$$

1. $y = 4.5 + 0 \cdot x$

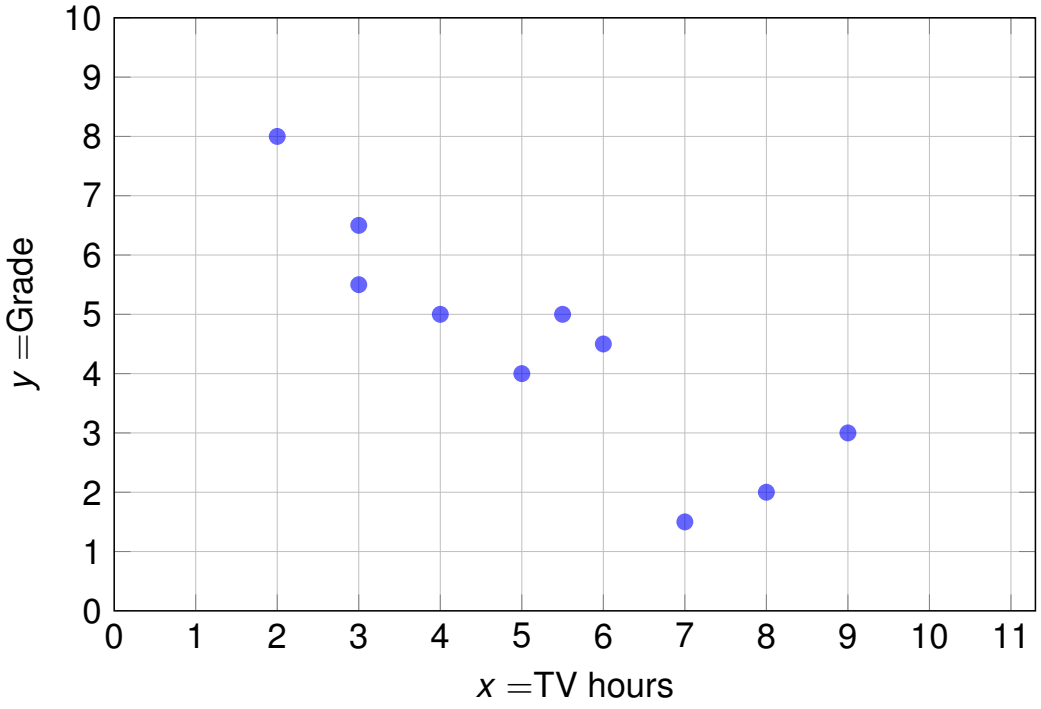
Example of different Linear Models



$$y = b + w \cdot x$$

1. $y = 4.5 + 0 \cdot x$ (missing slope!)

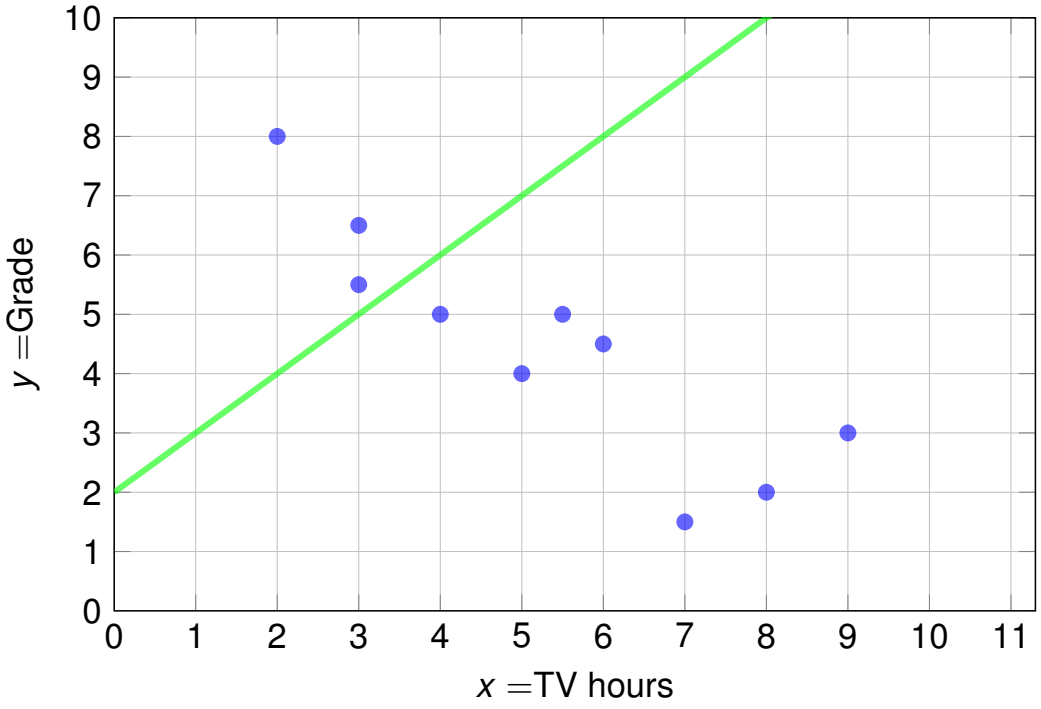
Example of different Linear Models



$$y = b + w \cdot x$$

1. $y = 4.5 + 0 \cdot x$ (missing slope!)
2. $y = 2 + x$

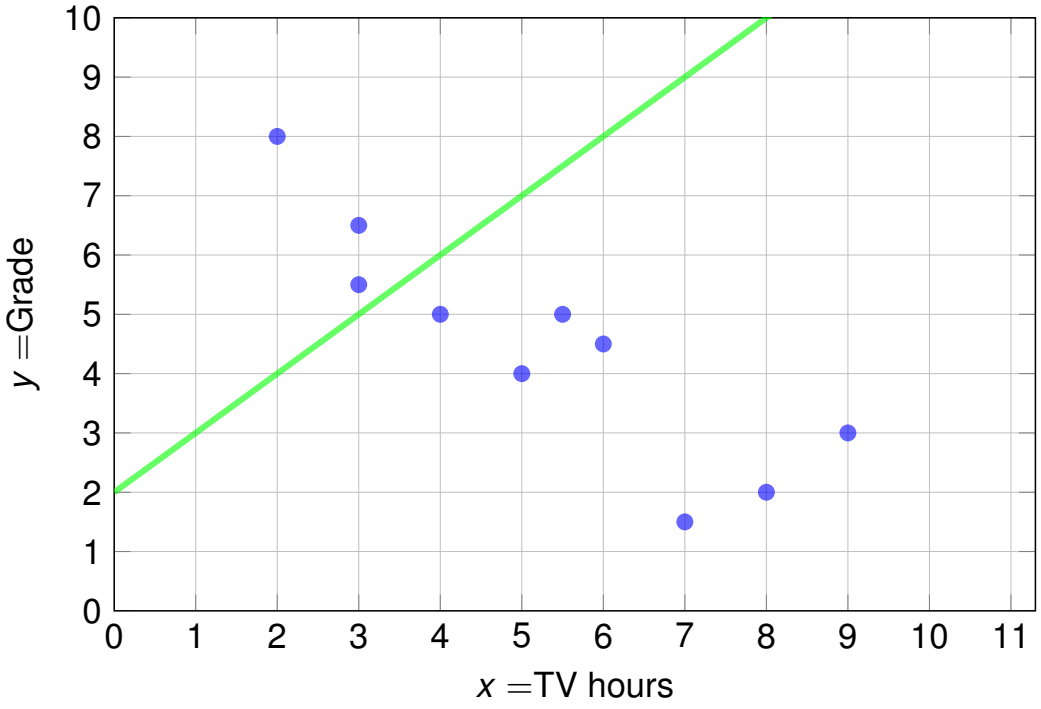
Example of different Linear Models



$$y = b + w \cdot x$$

1. $y = 4.5 + 0 \cdot x$ (missing slope!)
2. $y = 2 + x$

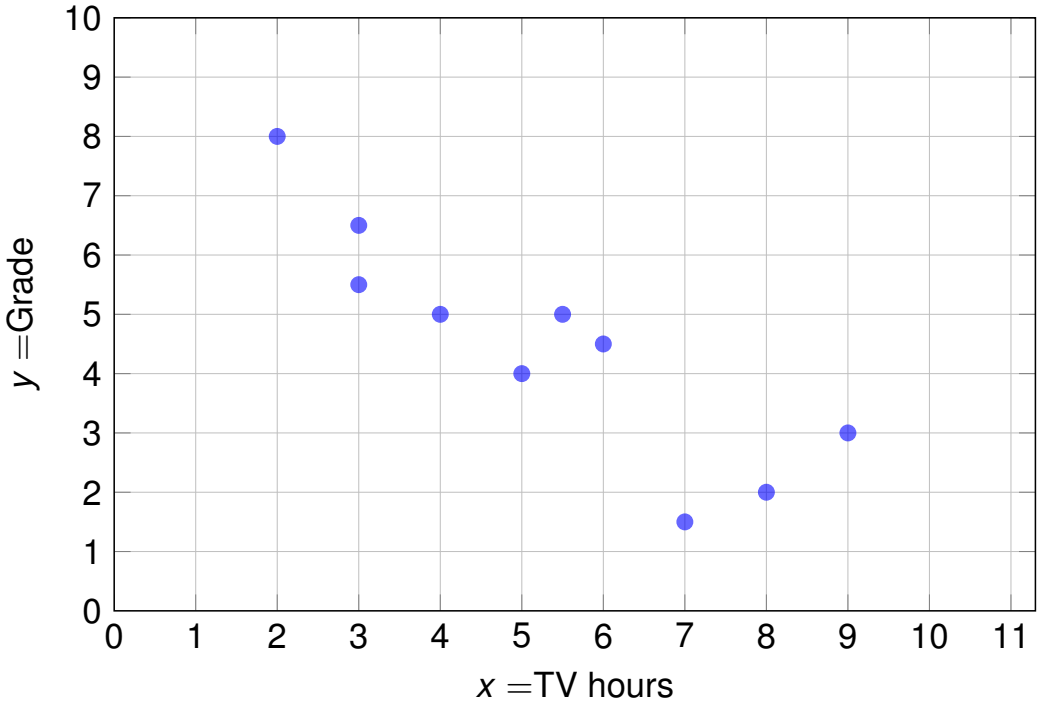
Example of different Linear Models



$$y = b + w \cdot x$$

1. $y = 4.5 + 0 \cdot x$ (missing slope!)
2. $y = 2 + x$ (wrong slope!)

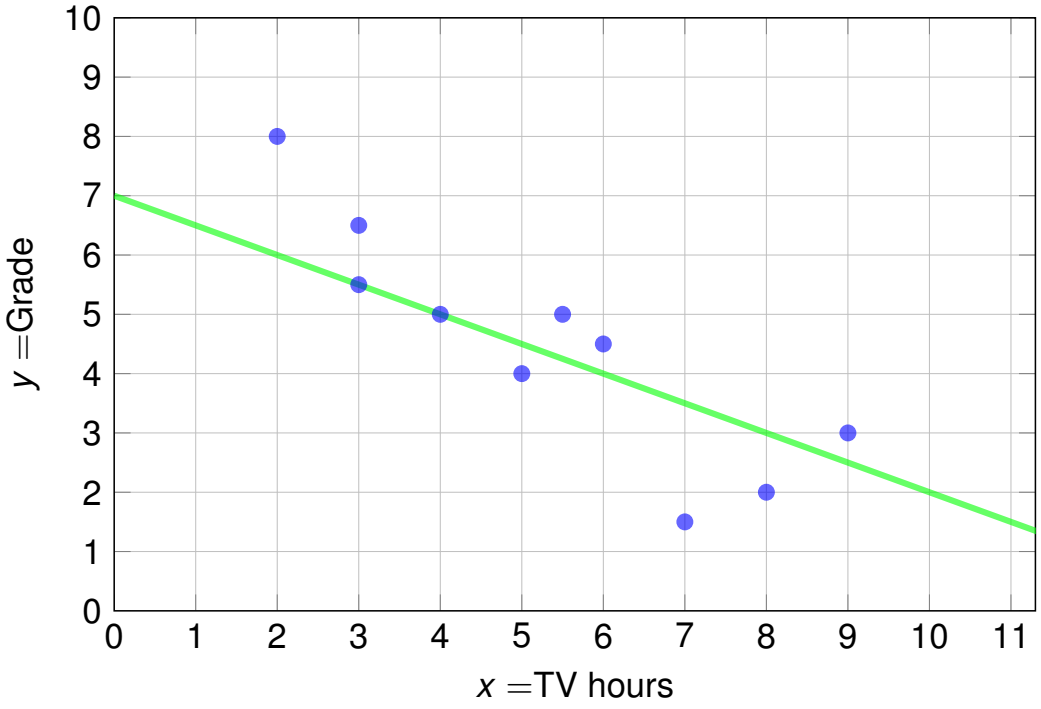
Example of different Linear Models



$$y = b + w \cdot x$$

1. $y = 4.5 + 0 \cdot x$ (missing slope!)
2. $y = 2 + x$ (wrong slope!)
3. $y = 7 - 0.5 \cdot x$

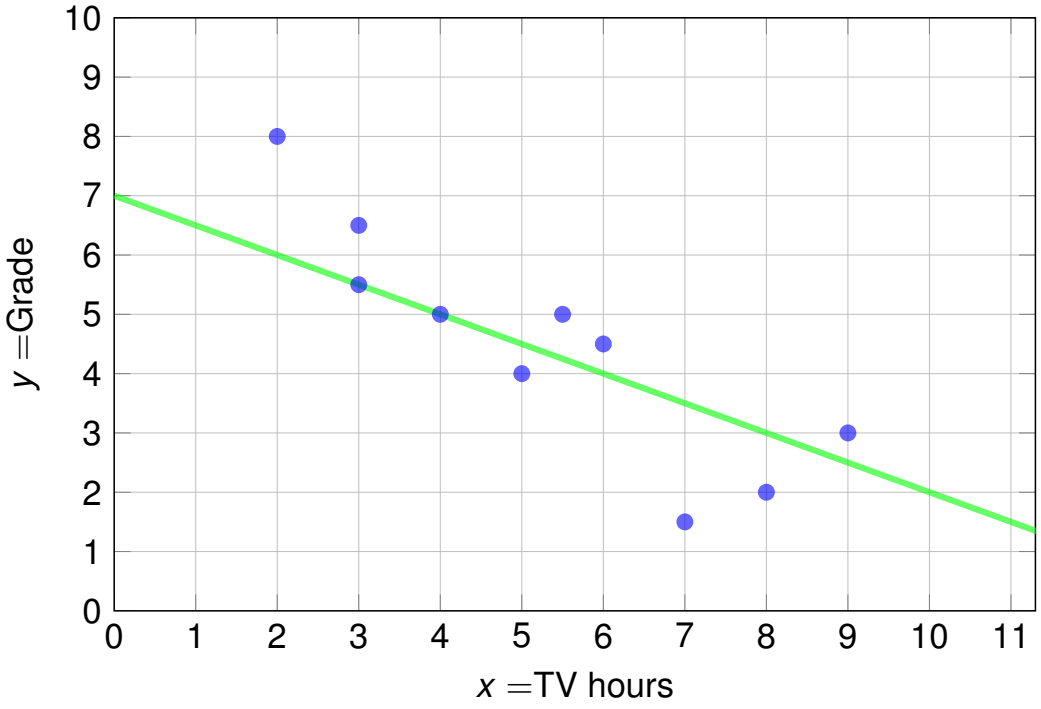
Example of different Linear Models



$$y = b + w \cdot x$$

1. $y = 4.5 + 0 \cdot x$ (missing slope!)
2. $y = 2 + x$ (wrong slope!)
3. $y = 7 - 0.5 \cdot x$

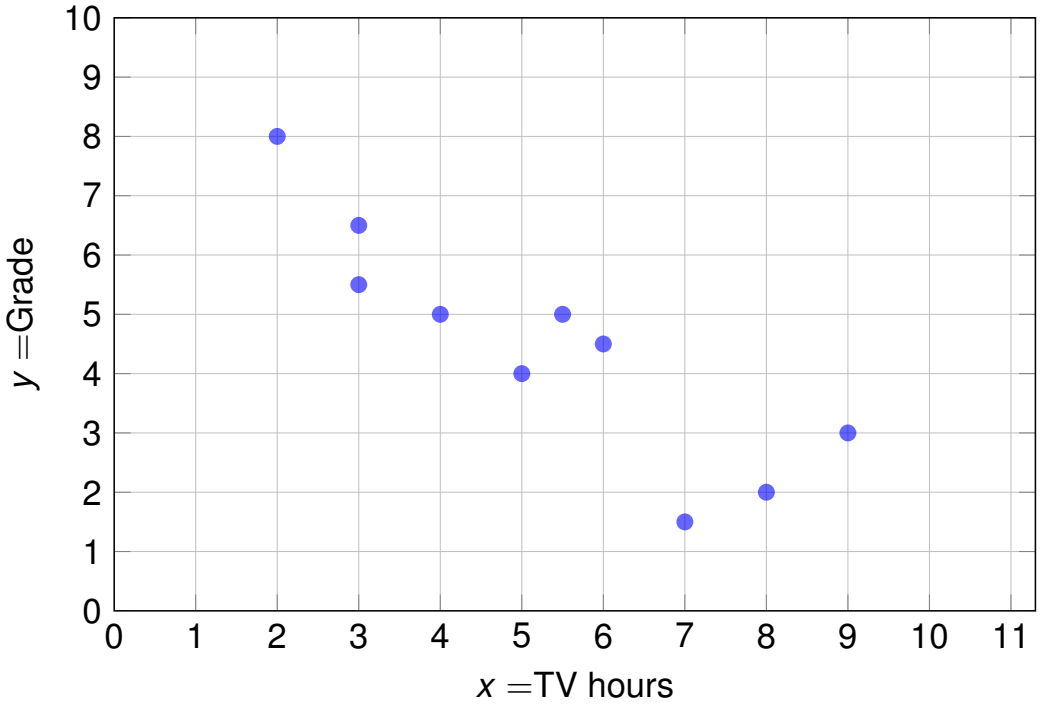
Example of different Linear Models



$$y = b + w \cdot x$$

1. $y = 4.5 + 0 \cdot x$ (missing slope!)
2. $y = 2 + x$ (wrong slope!)
3. $y = 7 - 0.5 \cdot x$ (much better!)

Example of different Linear Models



$$y = b + w \cdot x$$

1. $y = 4.5 + 0 \cdot x$ (missing slope!)
2. $y = 2 + x$ (wrong slope!)
3. $y = 7 - 0.5 \cdot x$ (much better!)

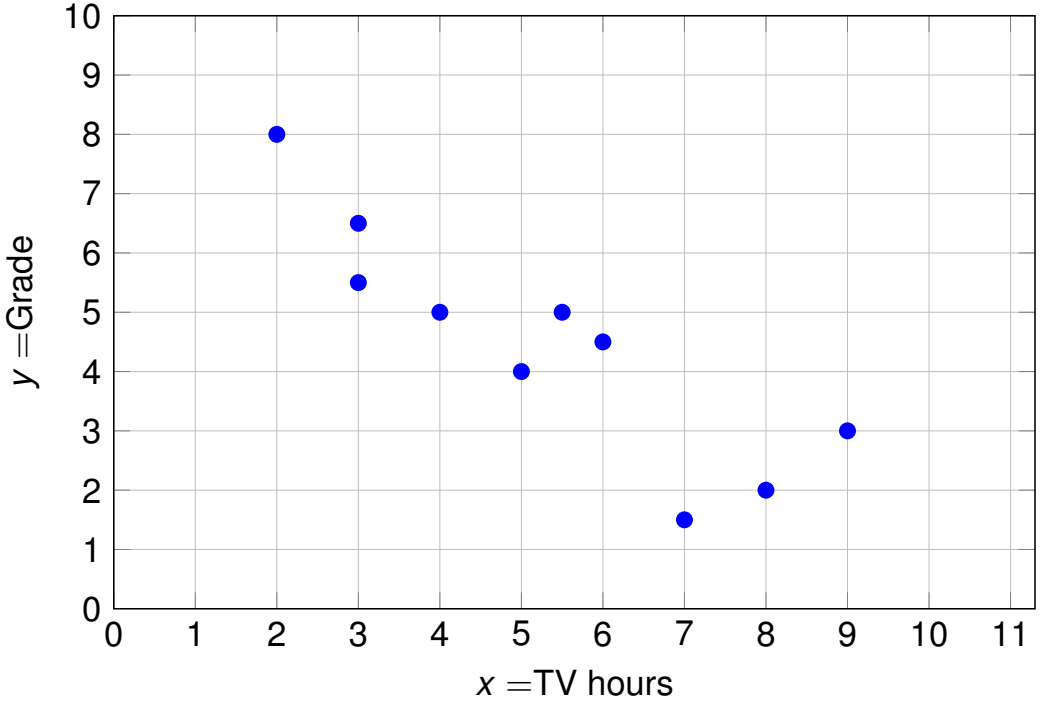
How do we find a good linear model?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.

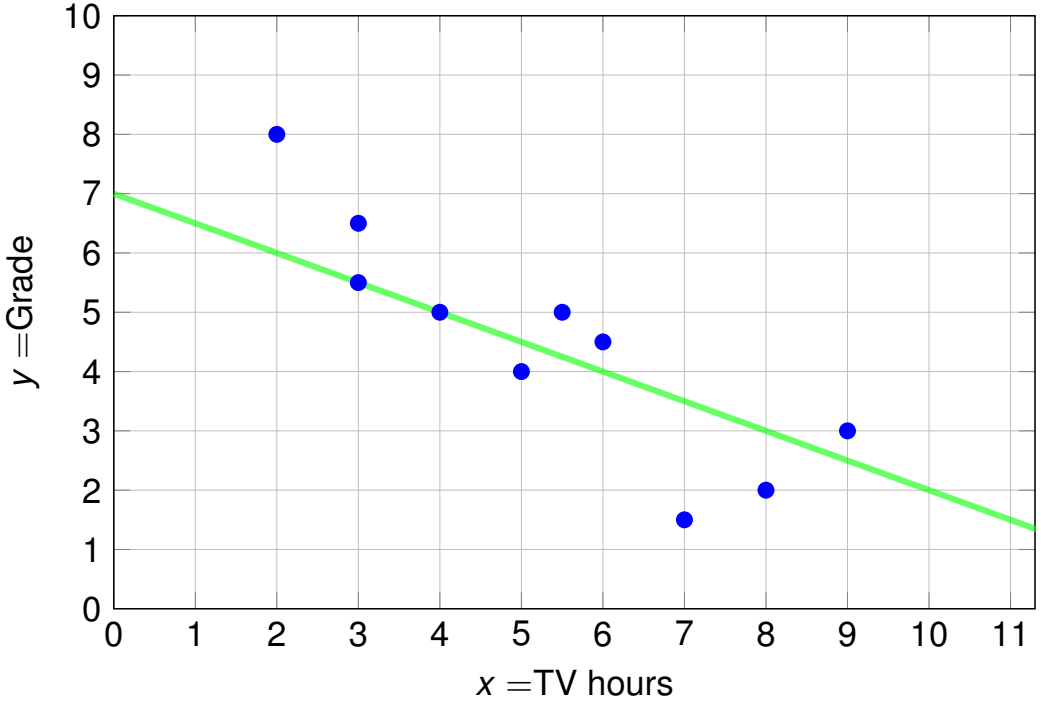
Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



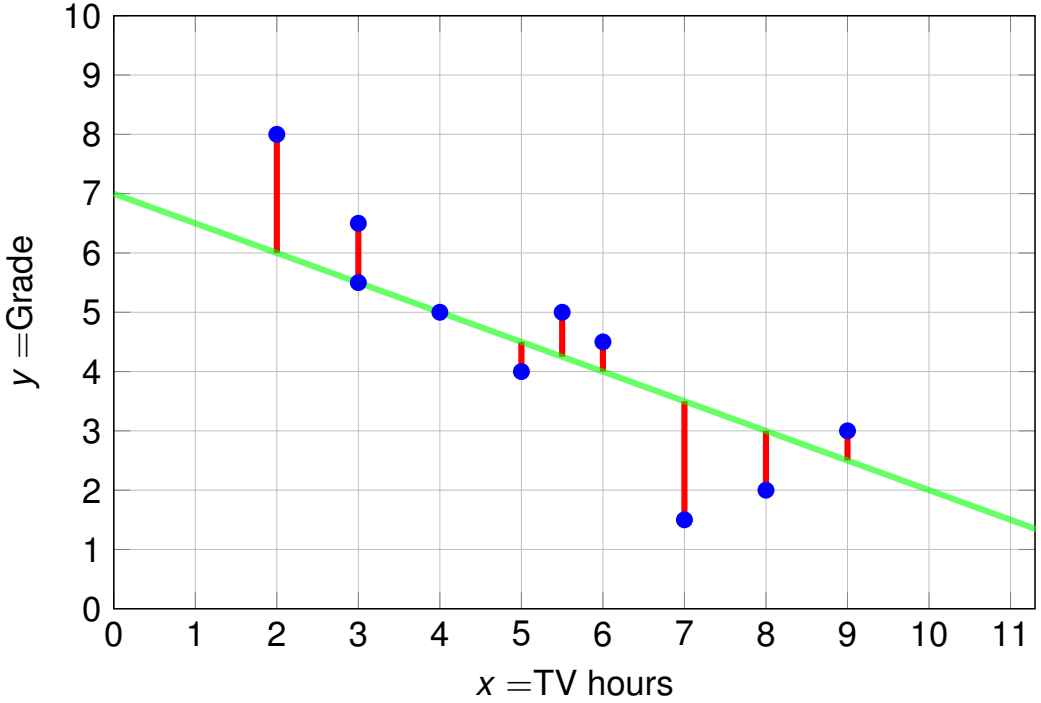
Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



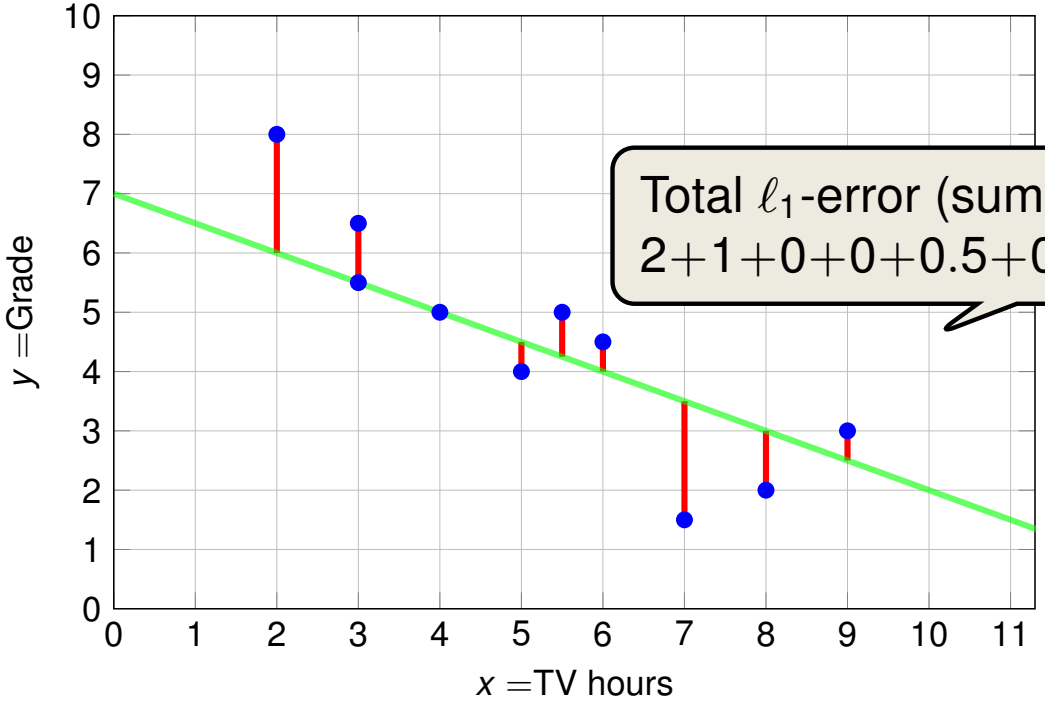
Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



Error of a Linear Model

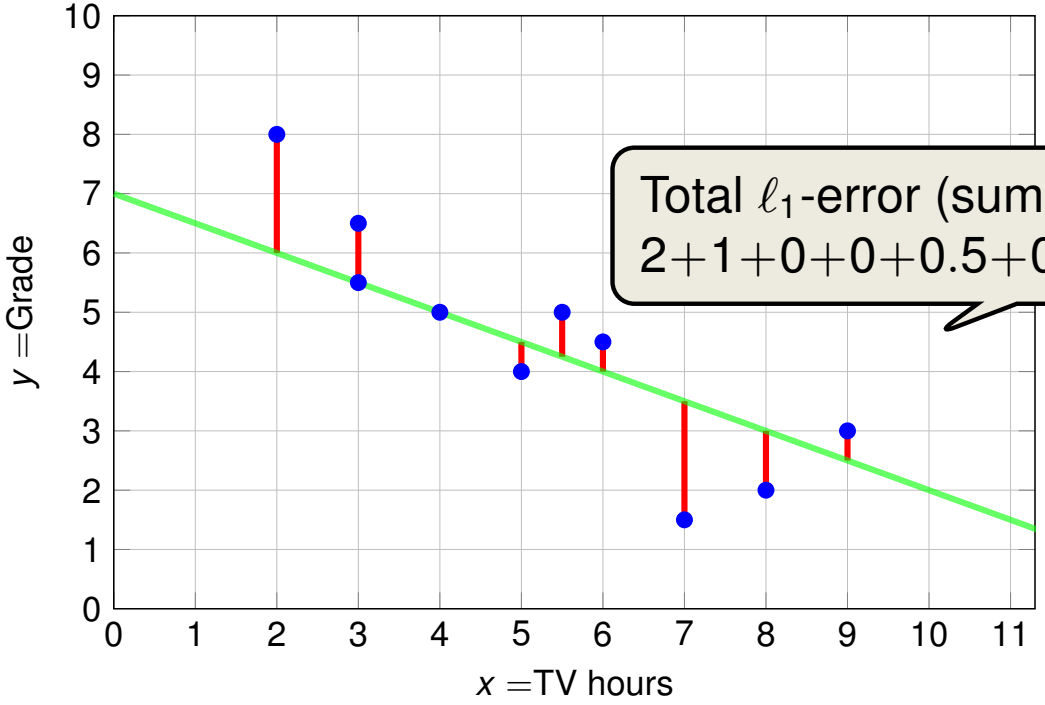
Find the linear model which achieves the smallest error for the input points.



Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.

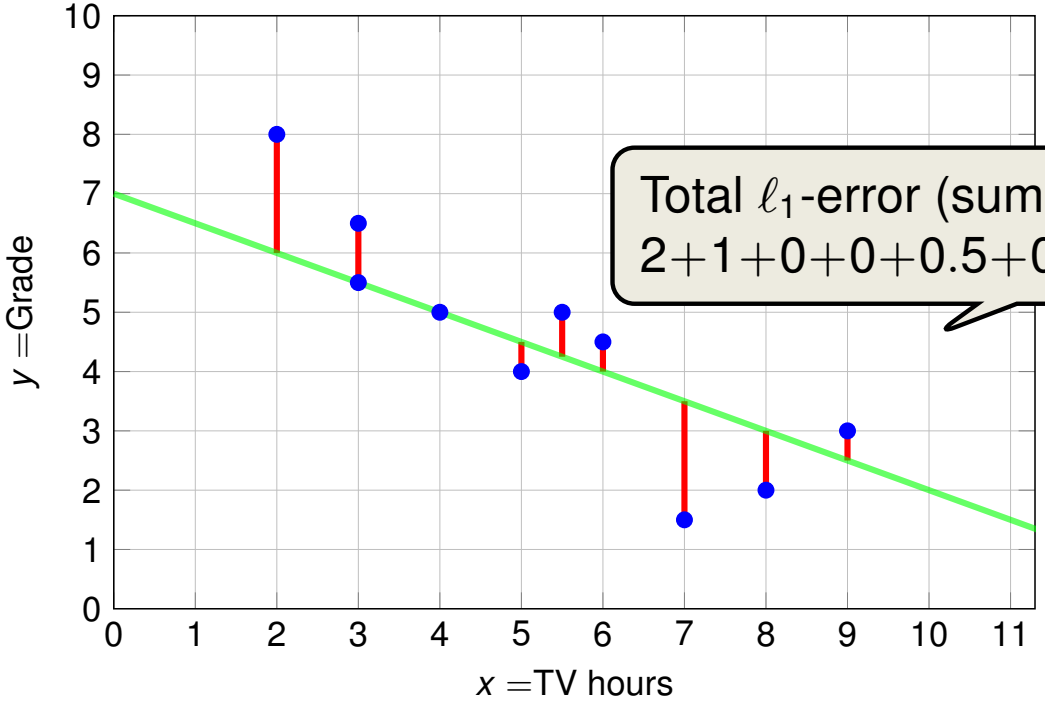


Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



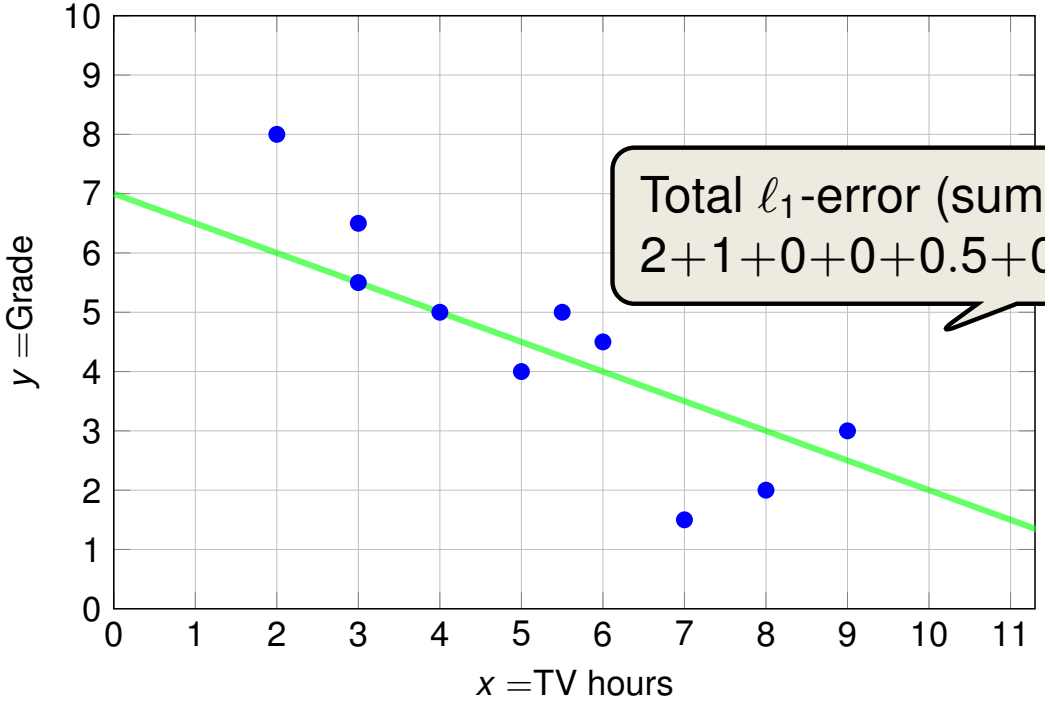
Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



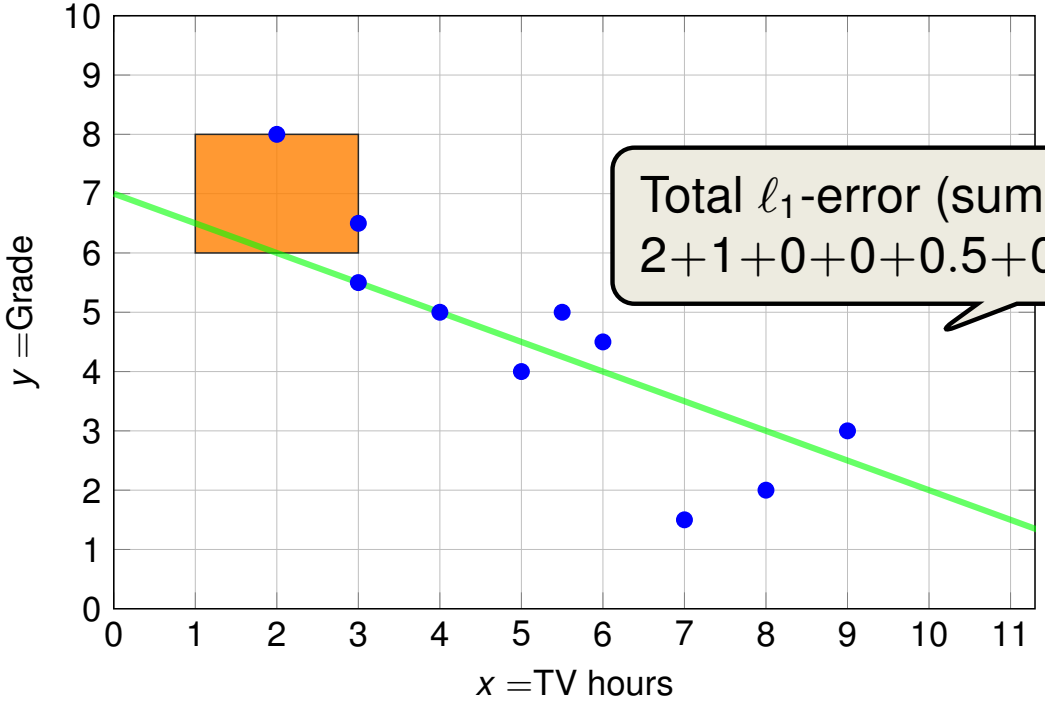
Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



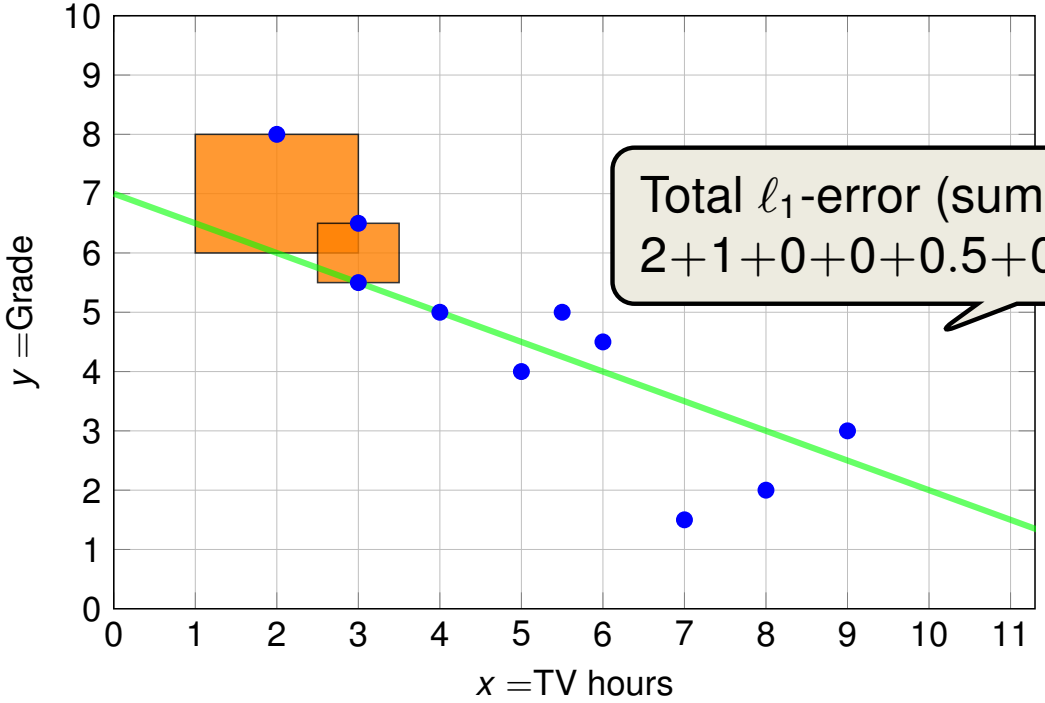
Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



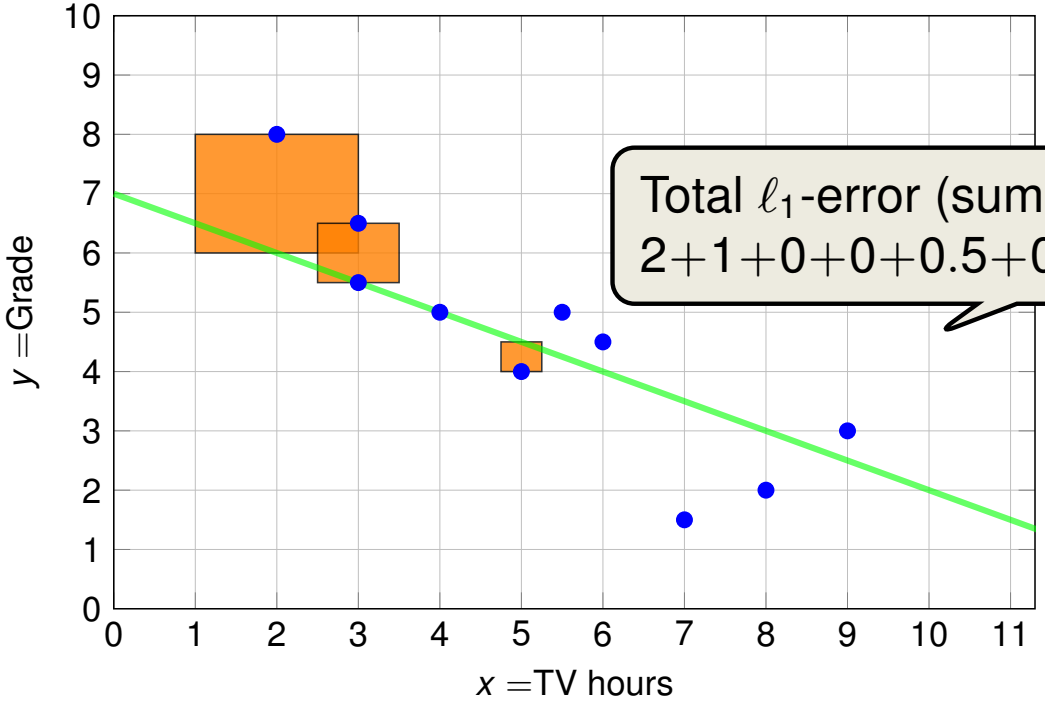
Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



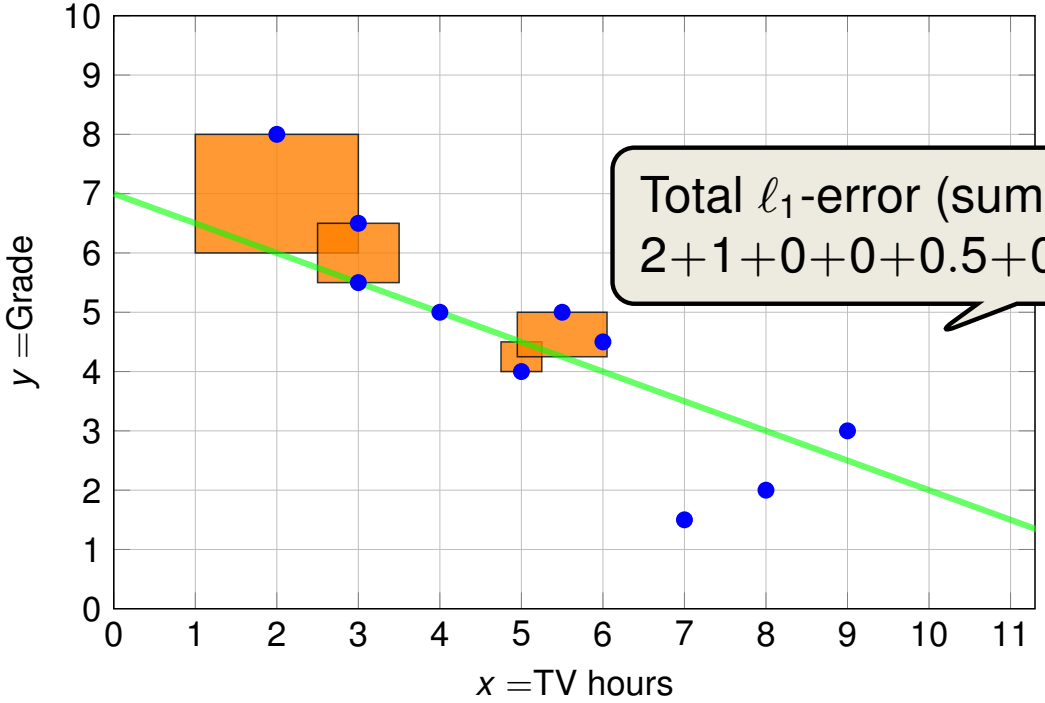
Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



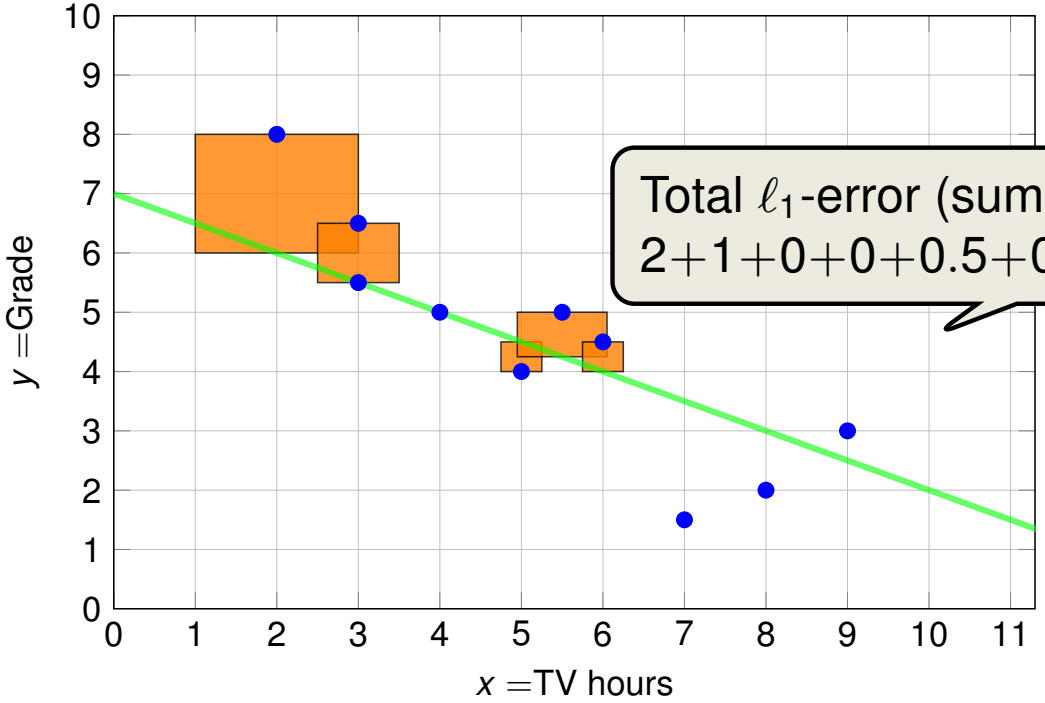
Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



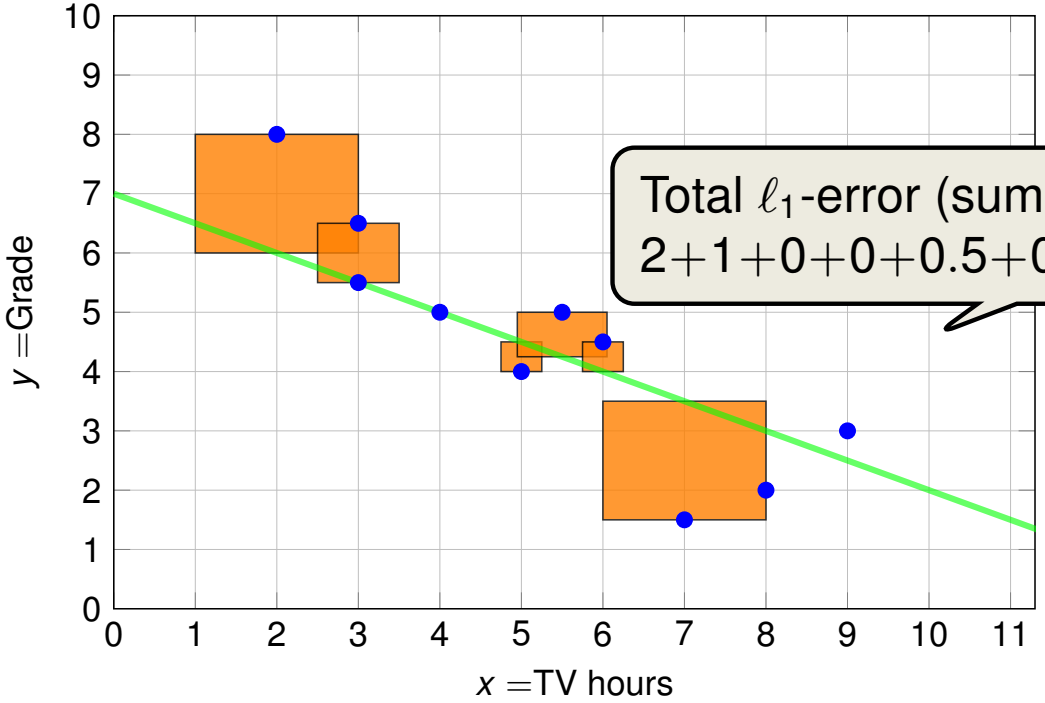
Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



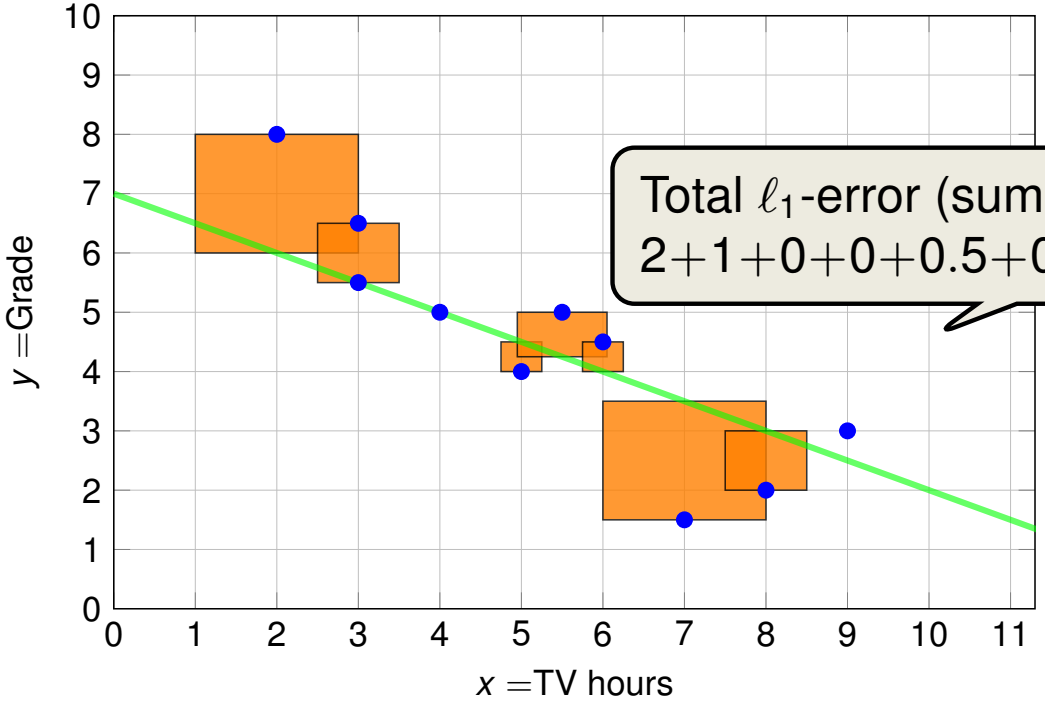
Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



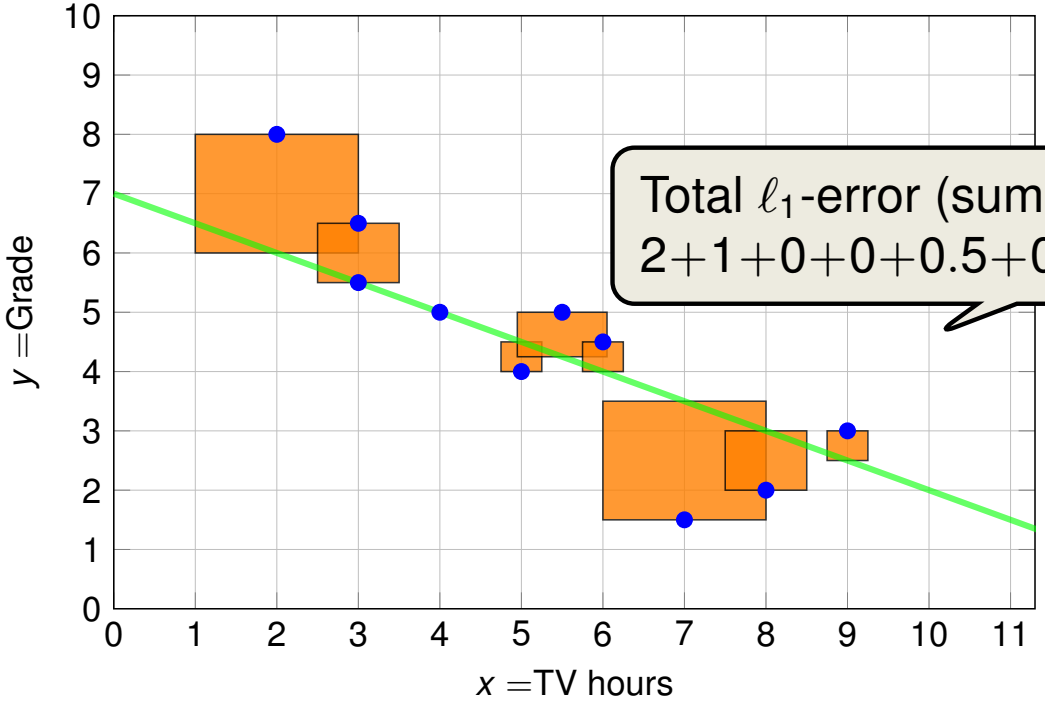
Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



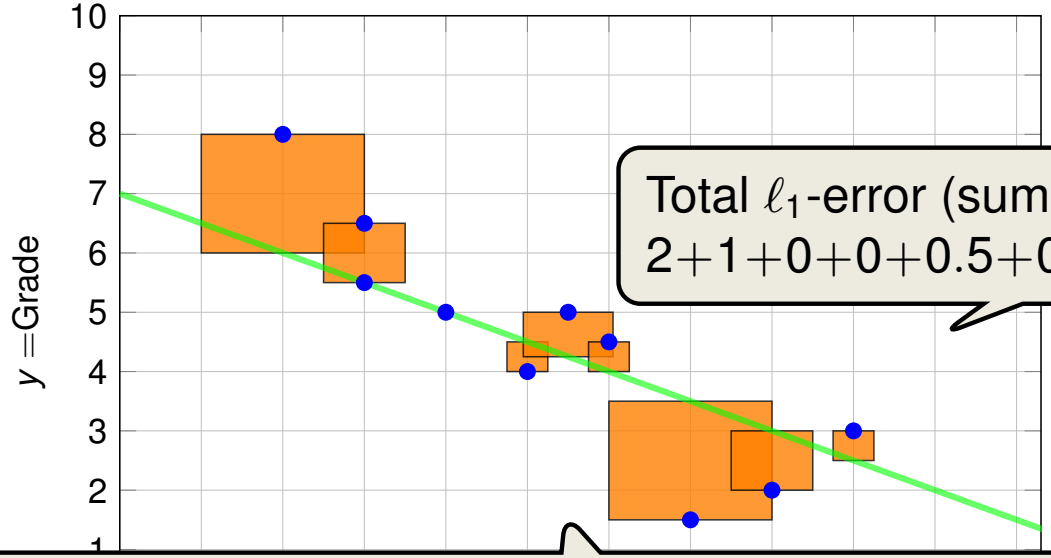
Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Error of a Linear Model

Find the linear model which achieves the smallest error for the input points.



Total ℓ_1-error (sum of residuals) is:
$2+1+0+0+0.5+0.75+0.5+2+1+0.5 = 8.25$

Total ℓ_2-error (quadratic error) is:
$2^2+1^2+0+0+0.5^2+0.75^2+0.5^2+2^2+1^2+0.5^2 = 11.3125$

Good fit apart from the points with 2 and 7 hours!

Exercise 2: How can we avoid/reduce these large deviations?

Method of Least Squares

C. Gauss (1777-1855)

In 1809 Gauss used the method of least squares for calculating the orbits of celestial bodies.

Method of Least Squares

C. Gauss (1777-1855)

In 1809 Gauss used the method of least squares for calculating the orbits of celestial bodies.

$$y = b + w \cdot x$$

Method of Least Squares

C. Gauss (1777-1855)

In 1809 Gauss used the method of least squares for calculating the orbits of celestial bodies.

$$y = b + w \cdot x$$

- Input: A data set of n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Method of Least Squares

C. Gauss (1777-1855)

In 1809 Gauss used the method of least squares for calculating the orbits of celestial bodies.

$$y = b + w \cdot x$$

- Input: A data set of n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Output: A linear model which minimises the squared error:

$$\sum_{i=1}^{n} \left(y_i - \widetilde{y}_i\right)^2$$

Method of Least Squares

C. Gauss (1777-1855)

In 1809 Gauss used the method of least squares for calculating the orbits of celestial bodies.

$$y = b + w \cdot x$$

- Input: A data set of n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Output: A linear model which minimises the squared error:

$$\sum_{i=1}^{n} \left(y_i - \widetilde{y}_i\right)^2$$

Method of Least Squares

C. Gauss (1777-1855)

In 1809 Gauss used the method of least squares for calculating the orbits of celestial bodies.

$$y = b + w \cdot x$$

- Input: A data set of n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Output: A linear model which minimises the squared error:

$$\sum_{i=1}^{n} (y_i - \widetilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b + w \cdot x_i))^2$$

Method of Least Squares

C. Gauss (1777-1855)

In 1809 Gauss used the method of least squares for calculating the orbits of celestial bodies.

$$y = b + w \cdot x$$

- Input: A data set of n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Output: A linear model which minimises the squared error:

$$\sum_{i=1}^{n} (y_i - \widetilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b + w \cdot x_i))^2$$

Direct Solution

- Let $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Conclusion of Simple Example (1/2)

Direct Solution

Conclusion of Simple Example (1/2)

Direct Solution

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

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Stud.	TV	Exam
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Stud.	TV	Exam
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
Mean		

- Step 1: Compute $\overline{x}$ and $\overline{y}$:

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Stud.	TV	Exam
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
Mean	5.25	4.5

- Step 1: Compute $\overline{x}$ and $\overline{y}$:

$$\overline{x} = 5.25,$$

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Stud.	TV	Exam
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
Mean	5.25	4.5

- Step 1: Compute $\overline{x}$ and $\overline{y}$:

$$\overline{x} = 5.25, \qquad \overline{y} = 4.5$$

- Step 2: Compute slope w:

$$w = \frac{\sum_{i=1}^{10}(x_i - 5.25) \cdot (y_i - 4.5)}{\sum_{i=1}^{10}(x_i - 5.25)^2}$$

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Stud.	TV	Exam
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
Mean	5.25	4.5

- Step 1: Compute $\overline{x}$ and $\overline{y}$:

$$\overline{x} = 5.25, \qquad \overline{y} = 4.5$$

- Step 2: Compute slope w:

$$w = \frac{\sum_{i=1}^{10}(x_i - 5.25) \cdot (y_i - 4.5)}{\sum_{i=1}^{10}(x_i - 5.25)^2} \approx -0.7611$$

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Stud.	TV	Exam
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
Mean	5.25	4.5

- Step 1: Compute $\overline{x}$ and $\overline{y}$:

$$\overline{x} = 5.25, \qquad \overline{y} = 4.5$$

- Step 2: Compute slope w:

$$w = \frac{\sum_{i=1}^{10}(x_i - 5.25) \cdot (y_i - 4.5)}{\sum_{i=1}^{10}(x_i - 5.25)^2} \approx -0.7611$$

- Step 3: Compute intercept b:

$$b = 5.25 + 0.7611 \cdot 5.25$$

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Stud.	TV	Exam	Pred.
1	2	8	
2	4	5	
3	3	5.5	
4	5	4	
5	6	4.5	
6	8	2	
7	3	6.5	
8	5.5	5	
9	7	1.5	
10	9	3	
Mean	5.25	4.5	

- Step 1: Compute $\overline{x}$ and $\overline{y}$:

$$\overline{x} = 5.25, \qquad \overline{y} = 4.5$$

- Step 2: Compute slope w:

$$w = \frac{\sum_{i=1}^{10}(x_i - 5.25) \cdot (y_i - 4.5)}{\sum_{i=1}^{10}(x_i - 5.25)^2} \approx -0.7611$$

- Step 3: Compute intercept b:

$$b = 5.25 + 0.7611 \cdot 5.25 \approx 8.496$$

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Stud.	TV	Exam	Pred.
1	2	8	6.97
2	4	5	5.45
3	3	5.5	6.21
4	5	4	4.69
5	6	4.5	3.93
6	8	2	2.41
7	3	6.5	6.21
8	5.5	5	4.31
9	7	1.5	3.16
10	9	3	1.65
Mean	5.25	4.5	

- Step 1: Compute $\overline{x}$ and $\overline{y}$:

$$\overline{x} = 5.25, \qquad \overline{y} = 4.5$$

- Step 2: Compute slope w:

$$w = \frac{\sum_{i=1}^{10}(x_i - 5.25) \cdot (y_i - 4.5)}{\sum_{i=1}^{10}(x_i - 5.25)^2} \approx -0.7611$$

- Step 3: Compute intercept b:

$$b = 5.25 + 0.7611 \cdot 5.25 \approx 8.496$$

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Stud.	TV	Exam	Pred.
1	2	8	6.97
2	4	5	5.45
3	3	5.5	6.21
4	5	4	4.69
5	6	4.5	3.93
6	8	2	2.41
7	3	6.5	6.21
8	5.5	5	4.31
9	7	1.5	3.16
10	9	3	1.65
Mean	5.25	4.5	

Total ℓ_2-error is only

- Step 1: Compute $\overline{x}$ and $\overline{y}$:

$$\overline{x} = 5.25, \qquad \overline{y} = 4.5$$

- Step 2: Compute slope w:

$$w = \frac{\sum_{i=1}^{10}(x_i - 5.25) \cdot (y_i - 4.5)}{\sum_{i=1}^{10}(x_i - 5.25)^2} \approx -0.7611$$

- Step 3: Compute intercept b:

$$b = 5.25 + 0.7611 \cdot 5.25 \approx 8.496$$

Conclusion of Simple Example (1/2)

Direct Solution

- Let $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Then

$$w = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

and $b = \overline{y} - w\overline{x}$.

Stud.	TV	Exam	Pred.
1	2	8	6.97
2	4	5	5.45
3	3	5.5	6.21
4	5	4	4.69
5	6	4.5	3.93
6	8	2	2.41
7	3	6.5	6.21
8	5.5	5	4.31
9	7	1.5	3.16
10	9	3	1.65
Mean	5.25	4.5	

Total ℓ_2-error is only 7.908!

- Step 1: Compute $\overline{x}$ and $\overline{y}$:

$$\overline{x} = 5.25, \qquad \overline{y} = 4.5$$

- Step 2: Compute slope w:

$$w = \frac{\sum_{i=1}^{10}(x_i - 5.25) \cdot (y_i - 4.5)}{\sum_{i=1}^{10}(x_i - 5.25)^2} \approx -0.7611$$

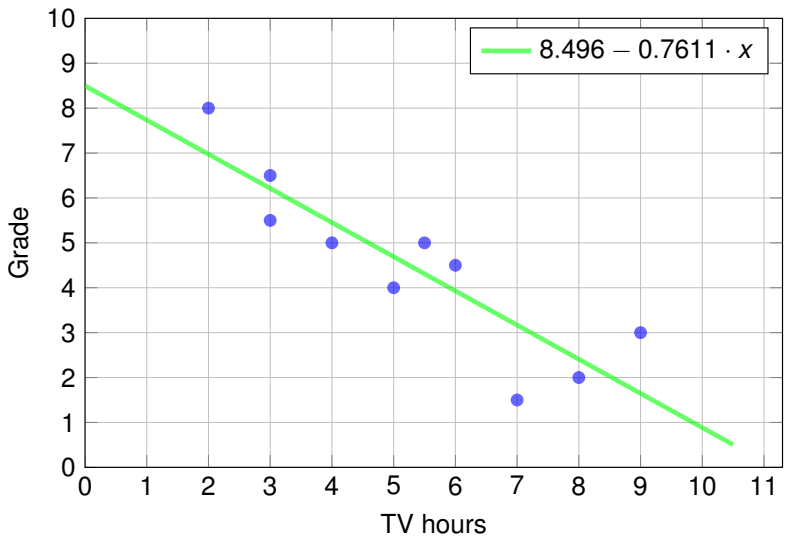
- Step 3: Compute intercept b:

$$b = 5.25 + 0.7611 \cdot 5.25 \approx 8.496$$

Conclusion of Simple Example (2/2)

$$y = 8.496 - 0.7611 \cdot x$$

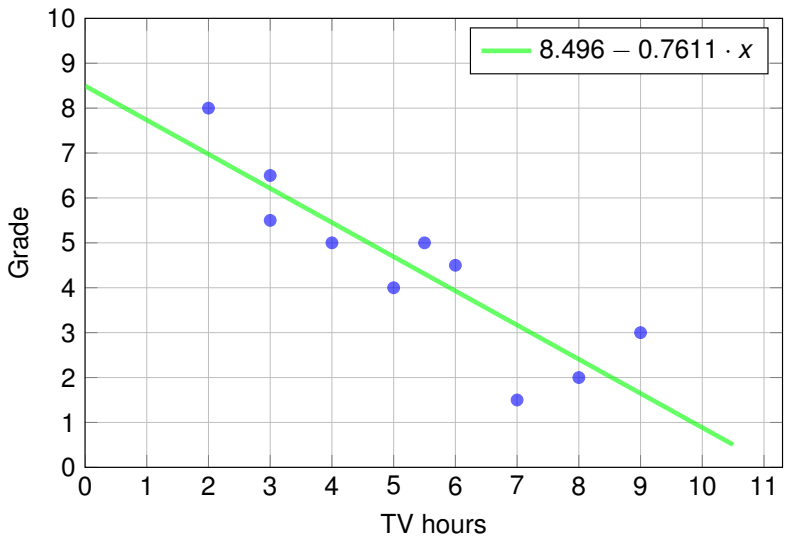
Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3



Conclusion of Simple Example (2/2)

$$y = 8.496 - 0.7611 \cdot x$$

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	??

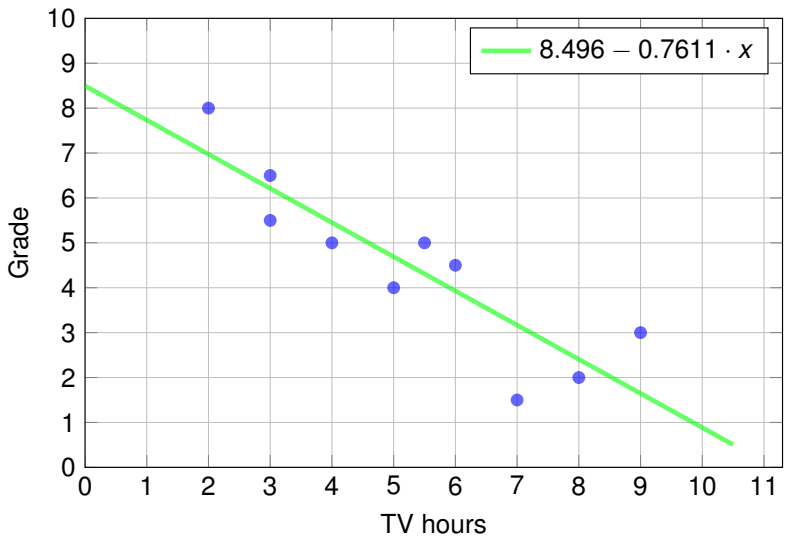


Which grade do you predict for Student 11?

Conclusion of Simple Example (2/2)

$$y = 8.496 - 0.7611 \cdot x$$

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	3.36

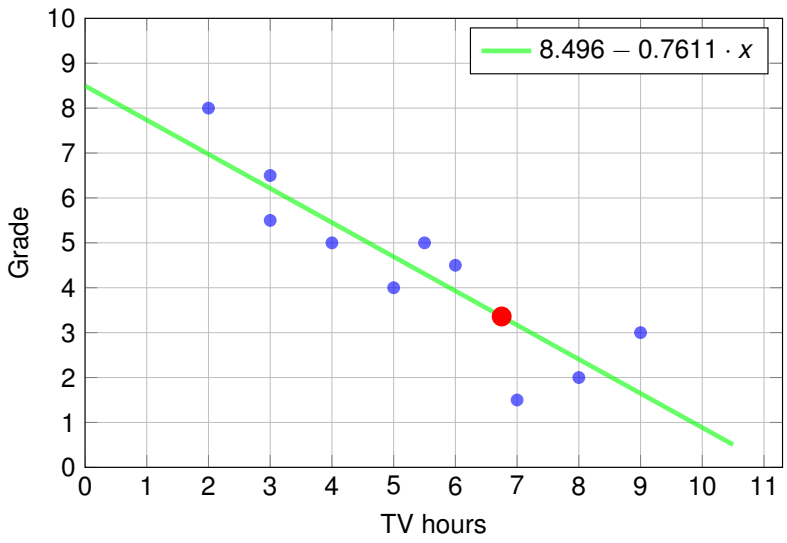


Which grade do you predict for Student 11?

Conclusion of Simple Example (2/2)

$$y = 8.496 - 0.7611 \cdot x$$

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	3.36

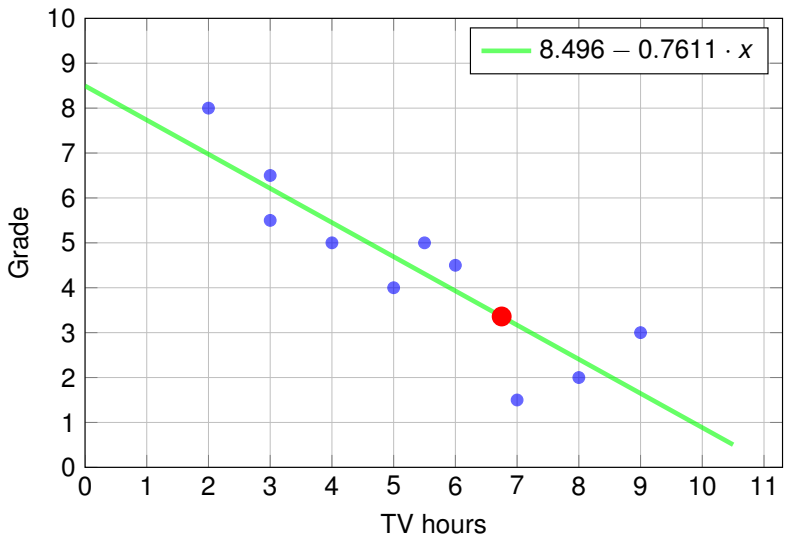


Which grade do you predict for Student 11?

Conclusion of Simple Example (2/2)

$$y = 8.496 - 0.7611 \cdot x$$

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	3.36



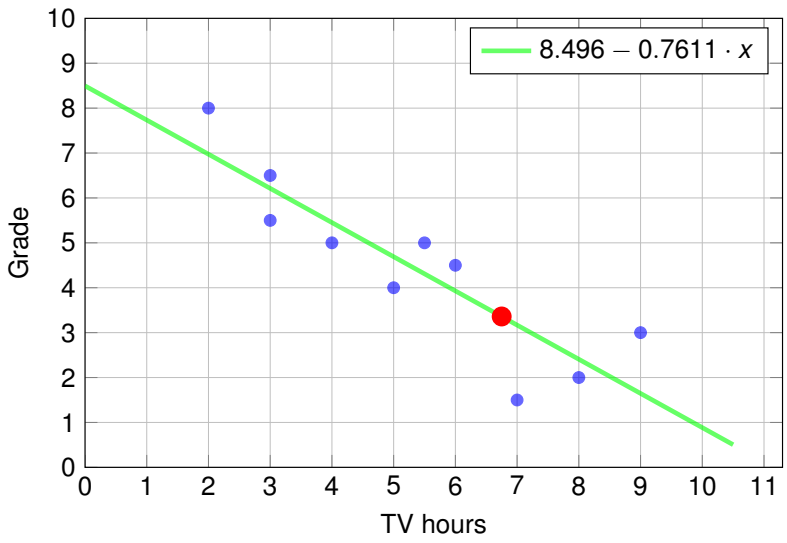
Which grade do you predict for Student 11?

Is this an accurate answer?

Conclusion of Simple Example (2/2)

$$y = 8.496 - 0.7611 \cdot x$$

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	3.36



Which grade do you predict for Student 11?

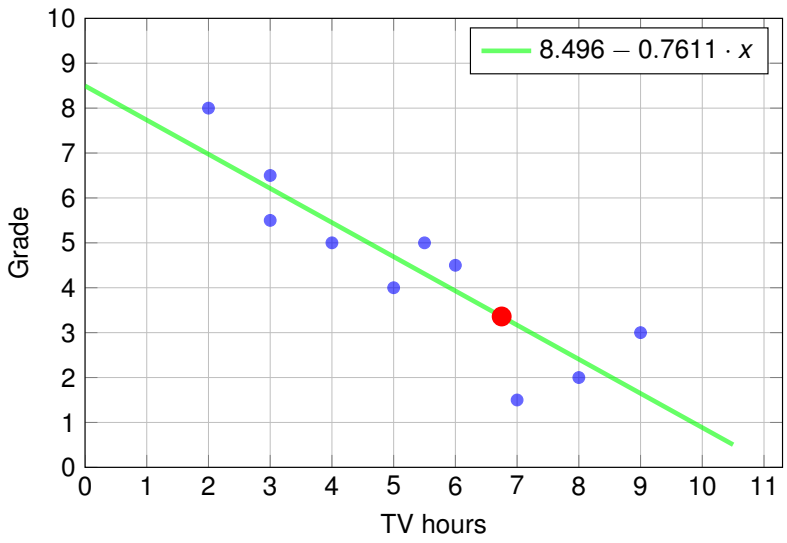
Is this an accurate answer?

- Yes (based on the current data and linearity assumption!)

Conclusion of Simple Example (2/2)

$$y = 8.496 - 0.7611 \cdot x$$

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	3.36



Which grade do you predict for Student 11?

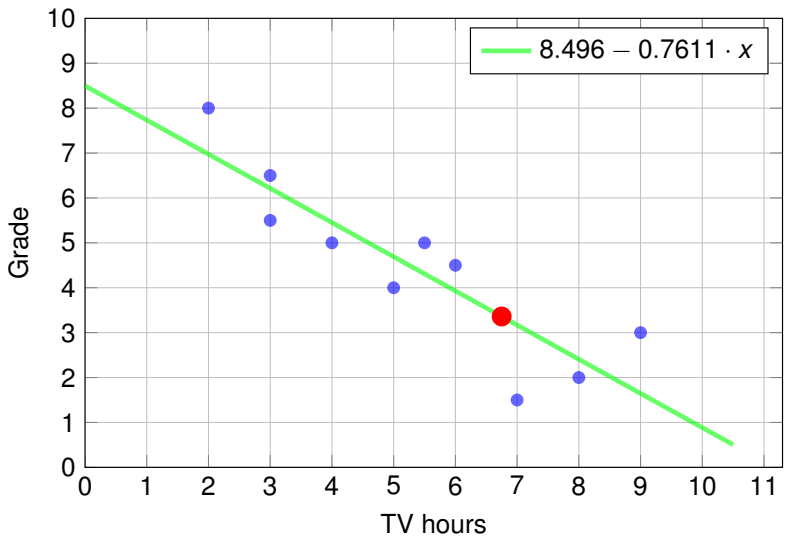
Is this an accurate answer?

- Yes (based on the current data and linearity assumption!)
- Methods: A) use test data set for validation

Conclusion of Simple Example (2/2)

$$y = 8.496 - 0.7611 \cdot x$$

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	3.36



Which grade do you predict for Student 11?

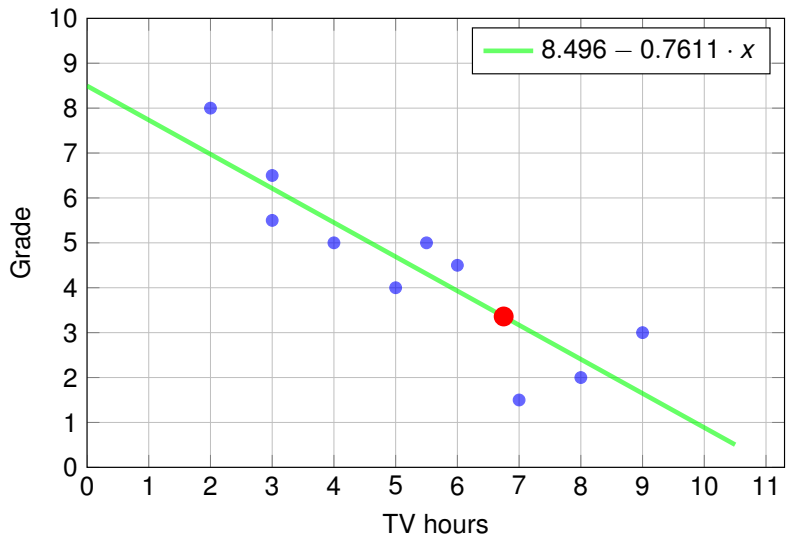
Is this an accurate answer?

- Yes (based on the current data and linearity assumption!)
- Methods: A) use test data set for validation or, B) use probabilistic model $y = b + wx + \mathcal{N}(0, \sigma)$ for some variance σ^2

Conclusion of Simple Example (2/2)

$$y = 8.496 - 0.7611 \cdot x$$

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	3.36



Which grade do you predict for Student 11?

Is this an accurate answer?

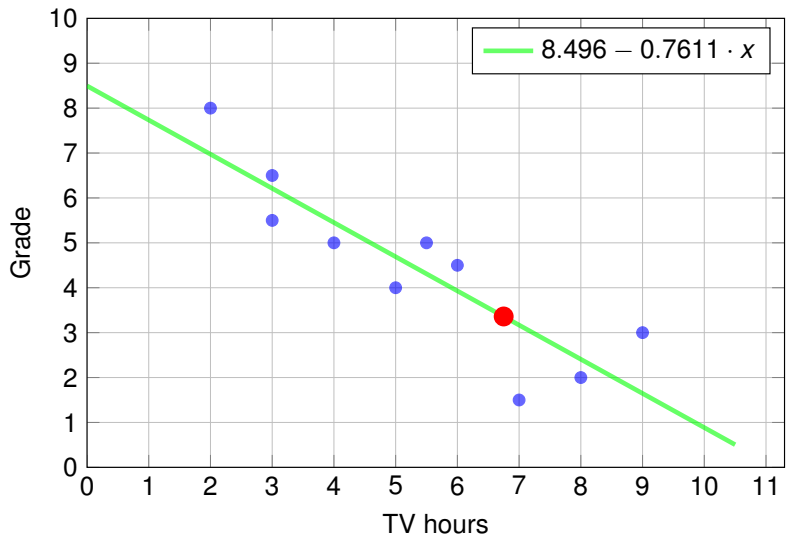
- Yes (based on the current data and linearity assumption!)
- Methods: A) use test data set for validation or, B) use probabilistic model $y = b + wx + \mathcal{N}(0, \sigma)$ for some variance σ^2

A) Is our model biased?

Conclusion of Simple Example (2/2)

$$y = 8.496 - 0.7611 \cdot x$$

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3
11	6.75	3.36



Which grade do you predict for Student 11?

Is this an accurate answer?

- Yes (based on the current data and linearity assumption!)
- Methods: A) use test data set for validation or, B) use probabilistic model $y = b + wx + \mathcal{N}(0, \sigma)$ for some variance σ^2

A) Is our model biased?

B) Is the data too noisy?

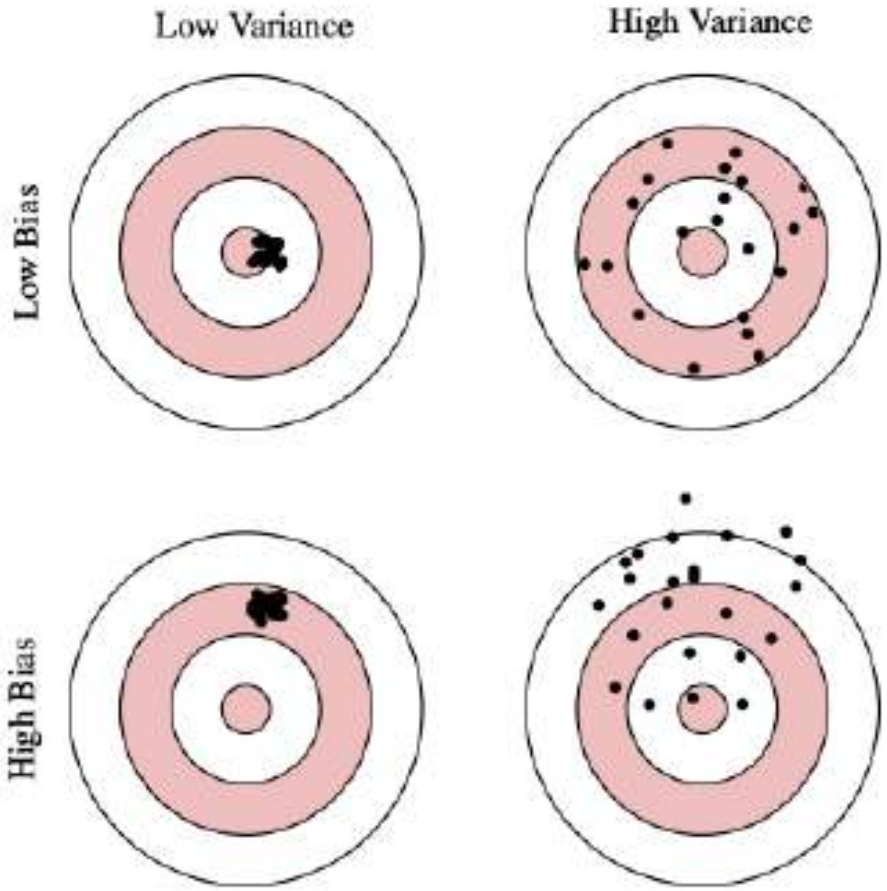
Bias versus Variance

In general, predictions by a model could be wrong due to:

Bias versus Variance

In general, predictions by a model could be wrong due to:

- High Bias: our model does not match the data well

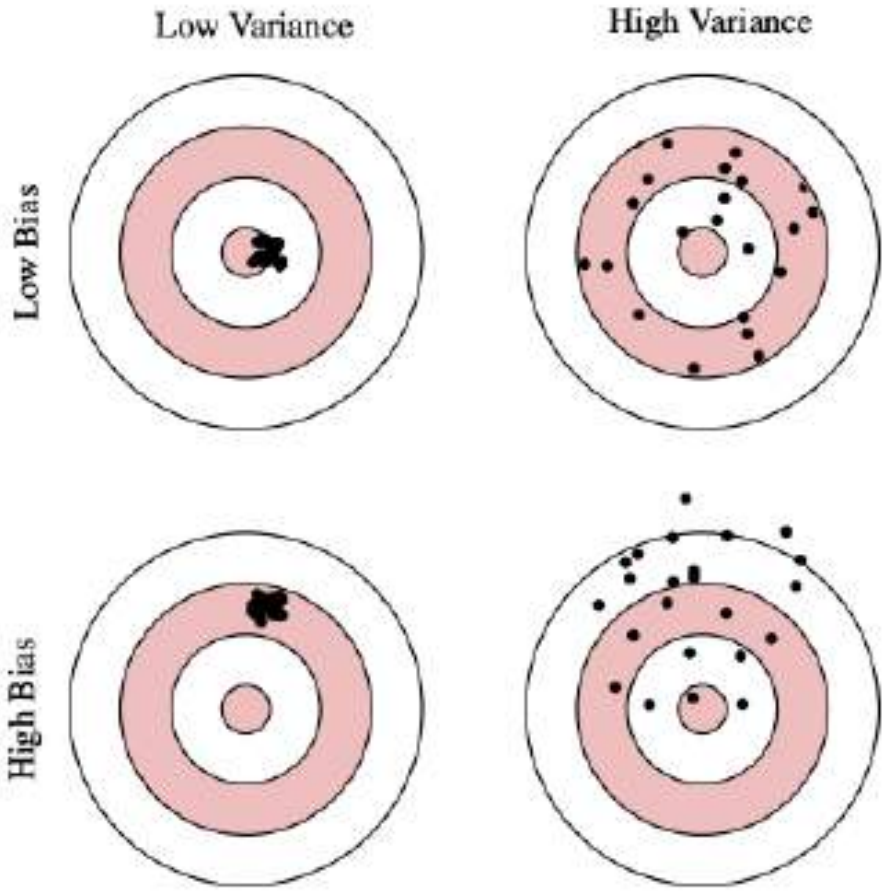


Source: Edwin Leuven (Point Estimation)

Bias versus Variance

In general, predictions by a model could be wrong due to:

- High Bias: our model does not match the data well
- High Variance: data set is inherently noisy, accurate prediction impossible

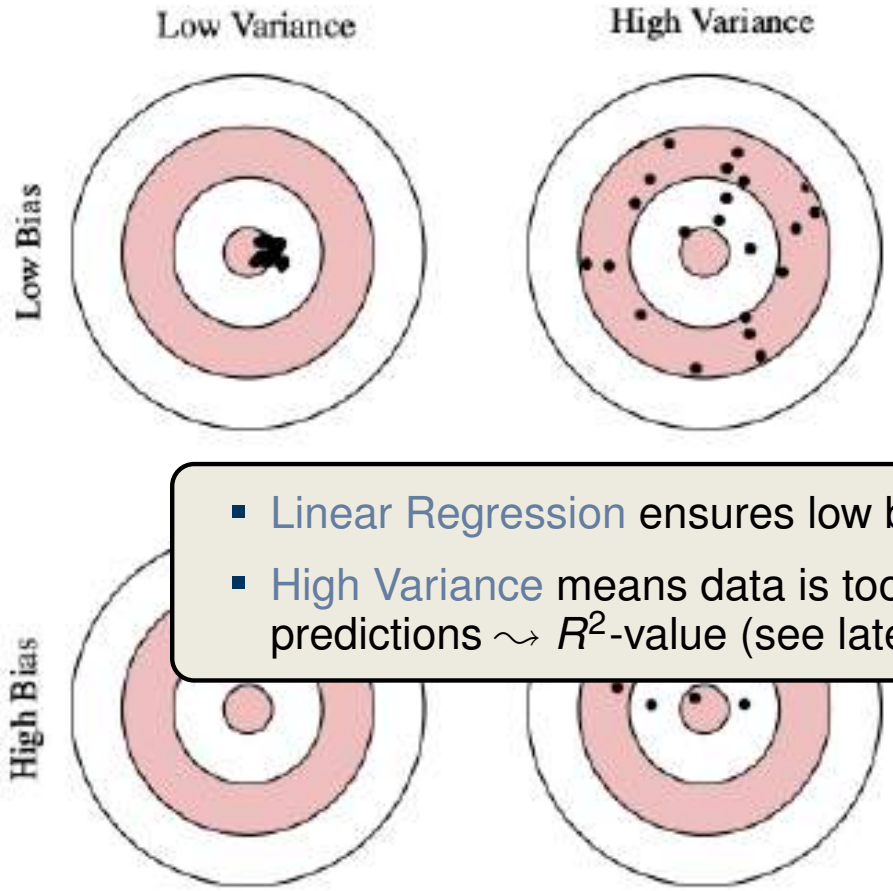


Source: Edwin Leuven (Point Estimation)

Bias versus Variance

In general, predictions by a model could be wrong due to:

- High Bias: our model does not match the data well
- High Variance: data set is inherently noisy, accurate prediction impossible



- Linear Regression ensures low bias
- High Variance means data is too noisy for predictions $\rightsquigarrow$ R^2-value (see later)!

Source: Edwin Leuven (Point Estimation)

Outline

Introduction

Linear Regression

Linear Regression with More than One Feature

Solving Linear Regression: Gradient Descent

Logistic Regression

Expanding the Data Set

Student	TV hours	Exam Grade
1	2	8
2	4	5
3	3	5.5
4	5	4
5	6	4.5
6	8	2
7	3	6.5
8	5.5	5
9	7	1.5
10	9	3

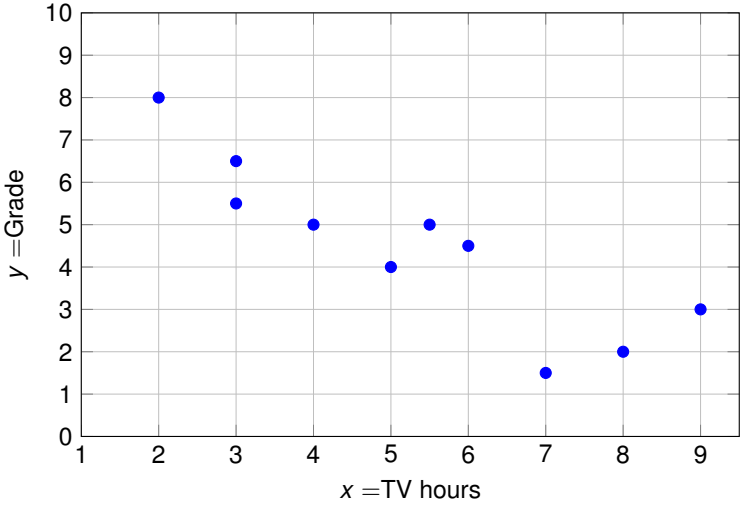
Expanding the Data Set

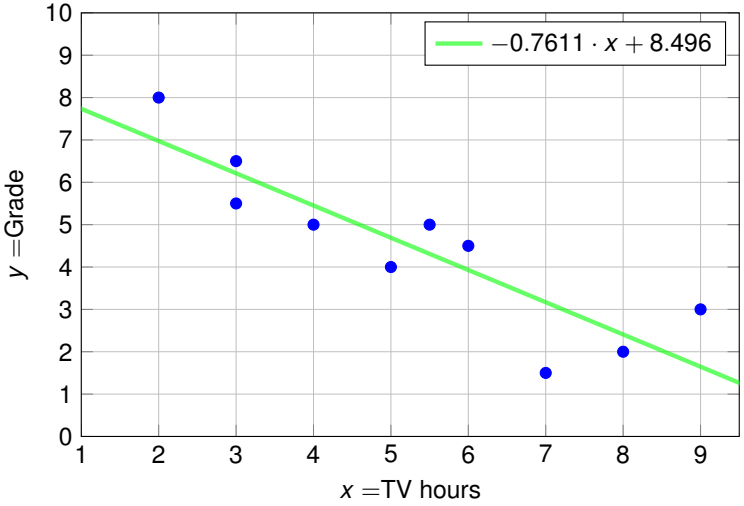
Student	TV hours	Homework Perf.	Exam Grade
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

Expanding the Data Set

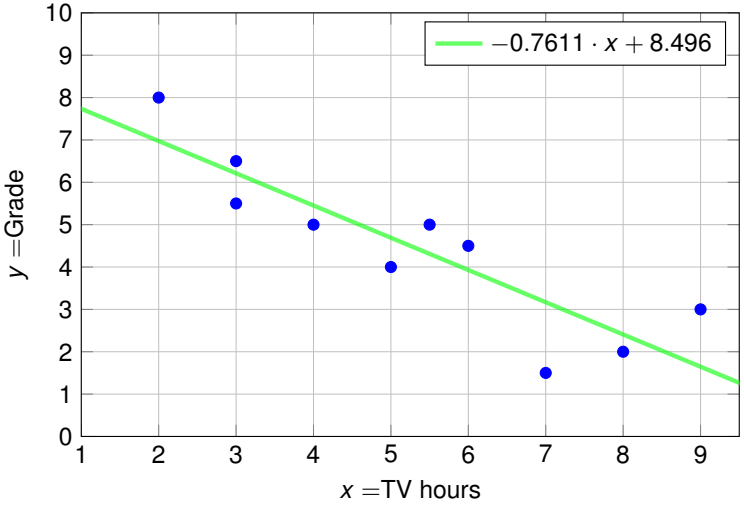
Student	TV hours	Homework Perf.	Exam Grade
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

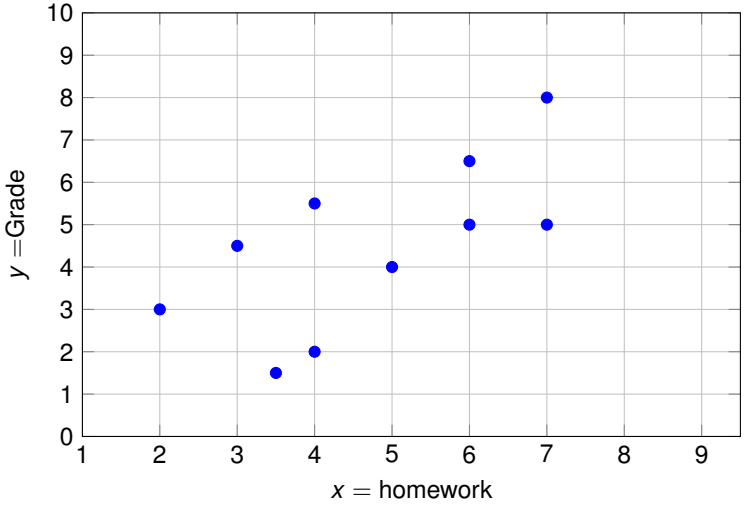
Can we get a better fit with more data?


10
9
8
7
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
y =Grade
x =TV hours

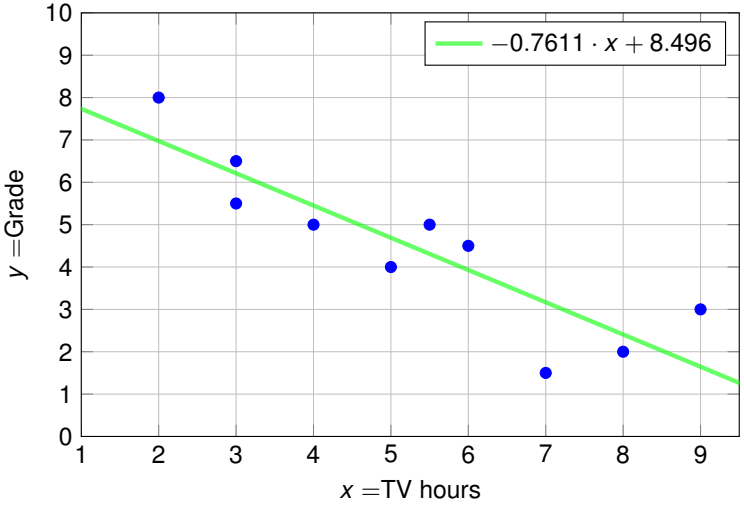


Square Error: 7.908

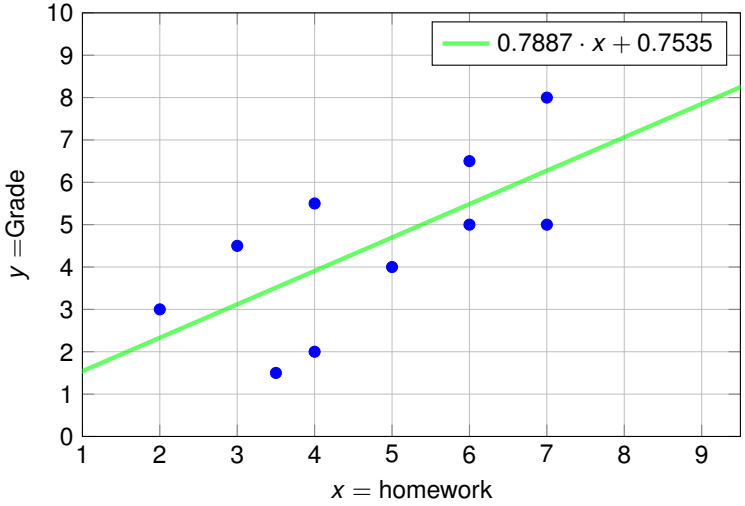




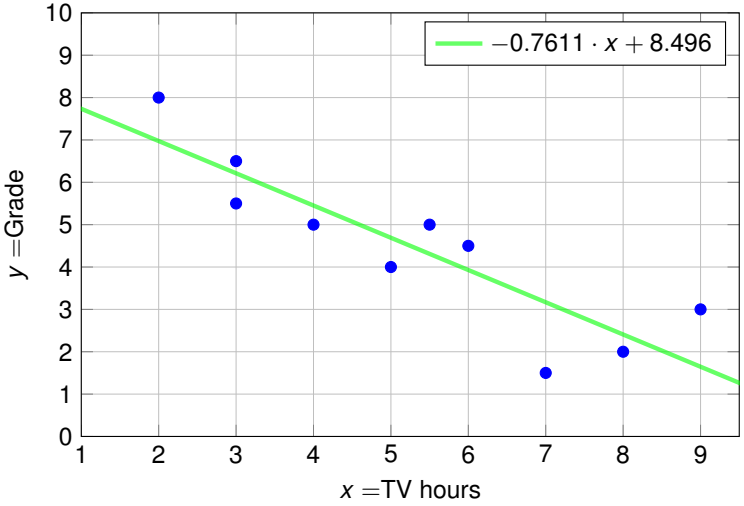
Square Error: 7.908



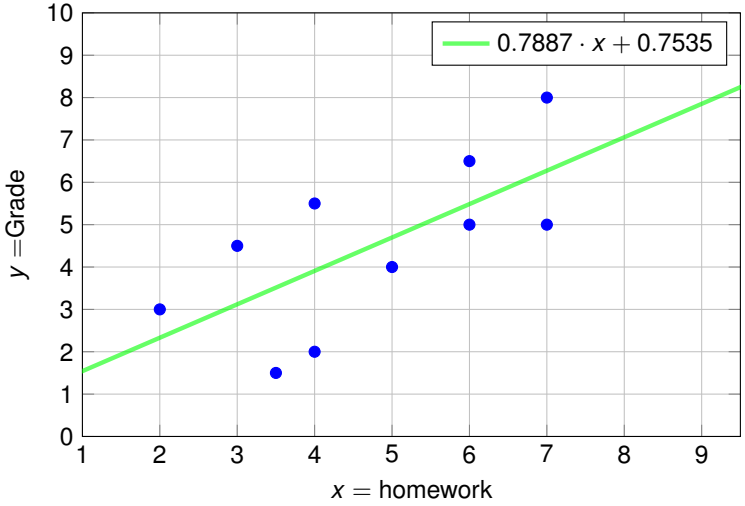
Square Error: 7.908



Square Error: 18.93

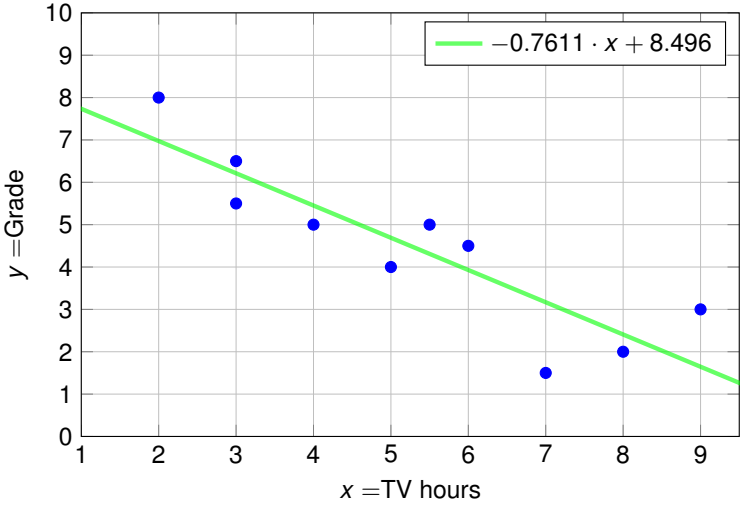


Square Error: 7.908

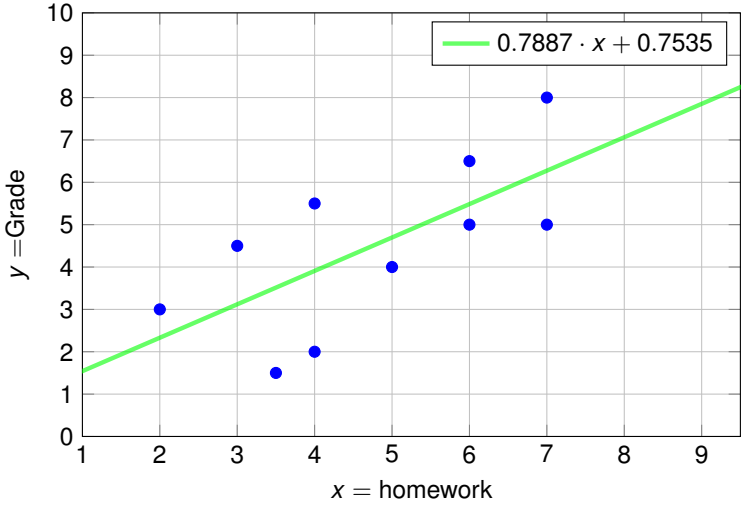


Square Error: 18.93

How to proceed?



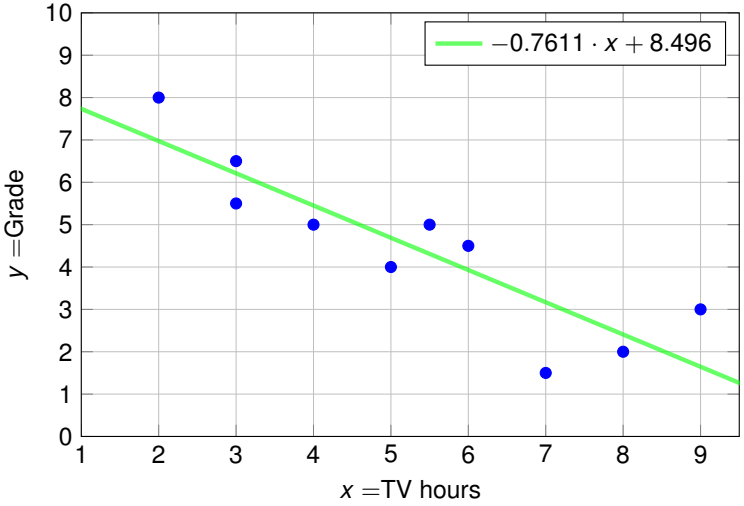
Square Error: 7.908



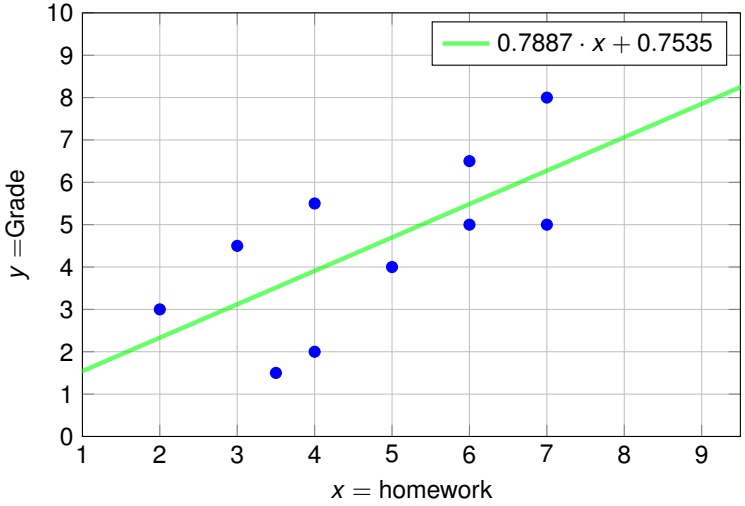
Square Error: 18.93

How to proceed?

- take the better of the two fits(?)



Square Error: 7.908



Square Error: 18.93

How to proceed?

- take the better of the two fits(?)
- take the average over the two lines(?) (yields square error: 10.00)

Multiple (Multi-Dimensional) Regression

$$\boxed{y = b + w_1 \cdot x_1 + w_2 \cdot x_2}$$

Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade

Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$

Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

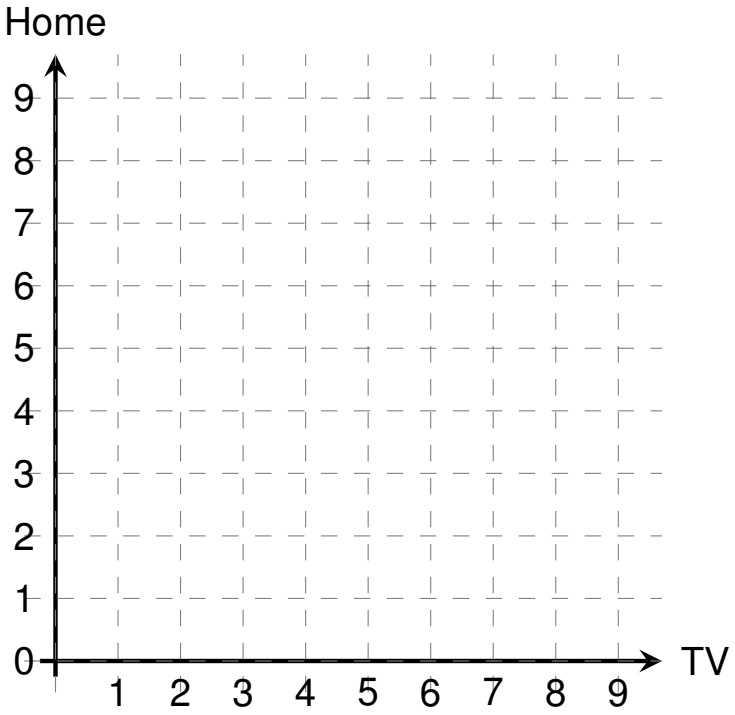
Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

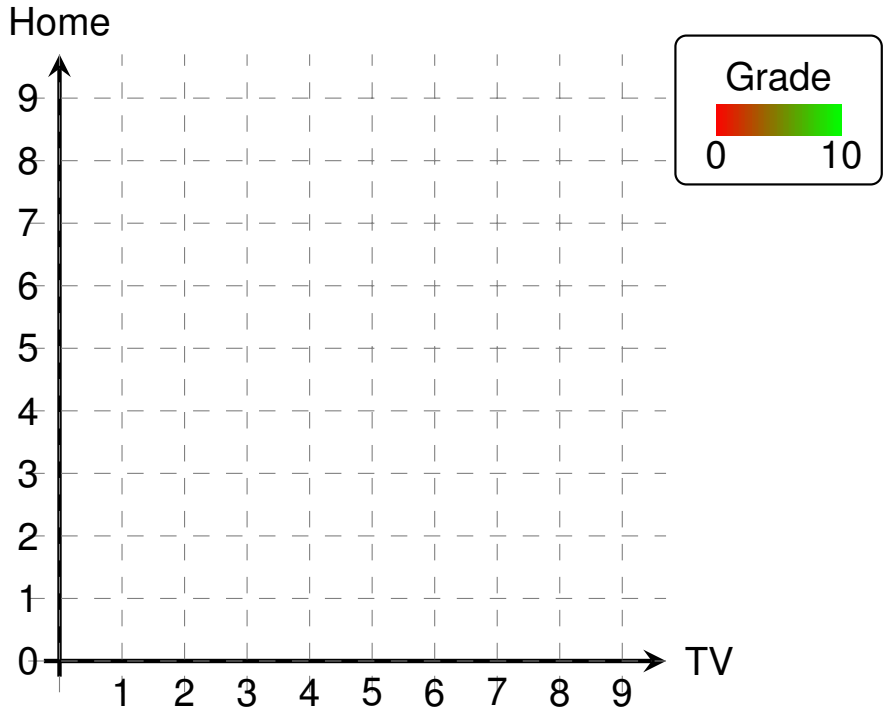


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

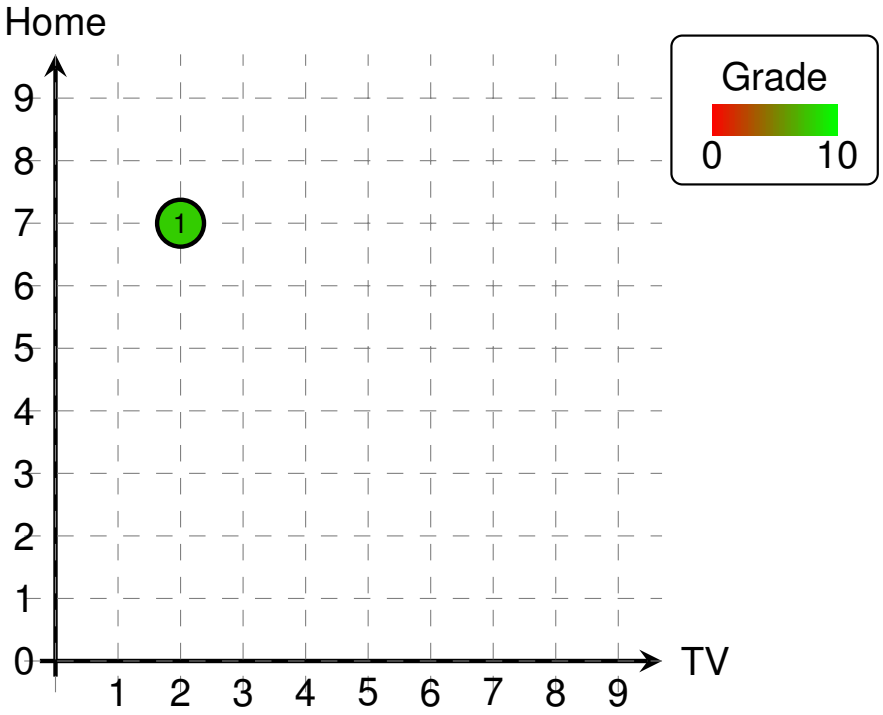


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

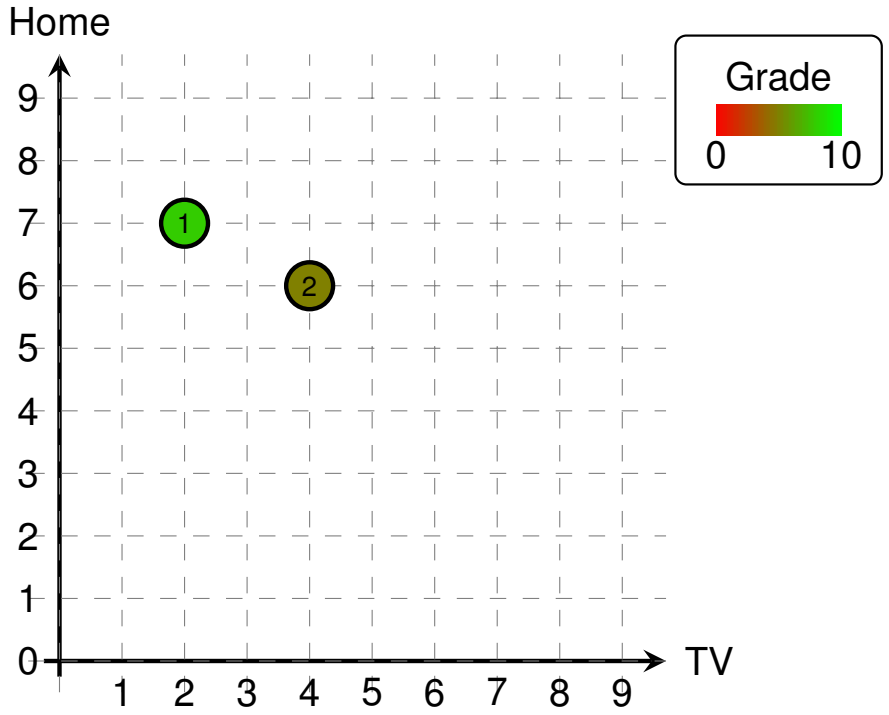


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

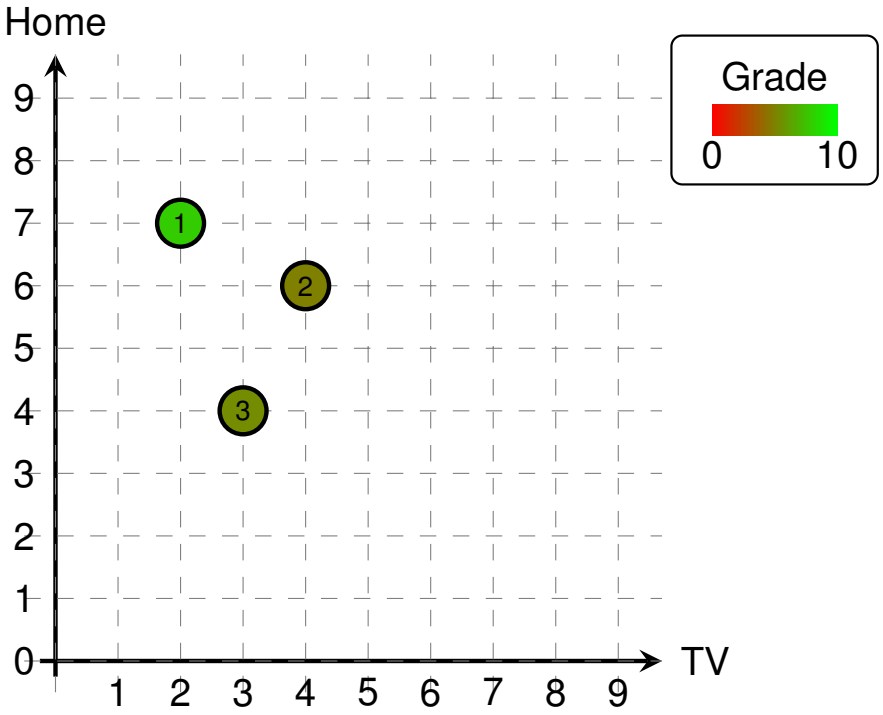


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

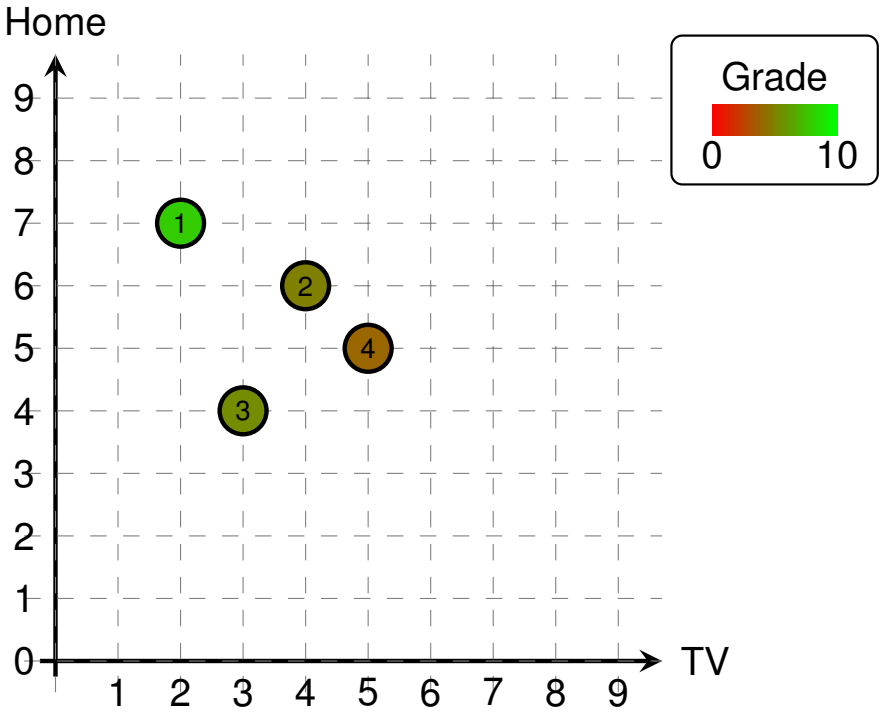


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

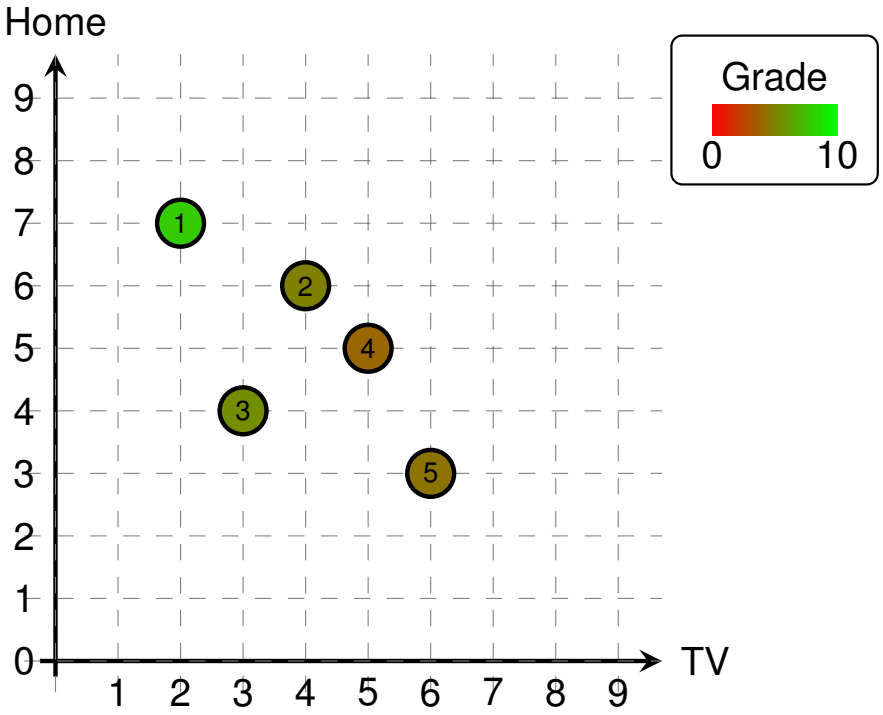


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

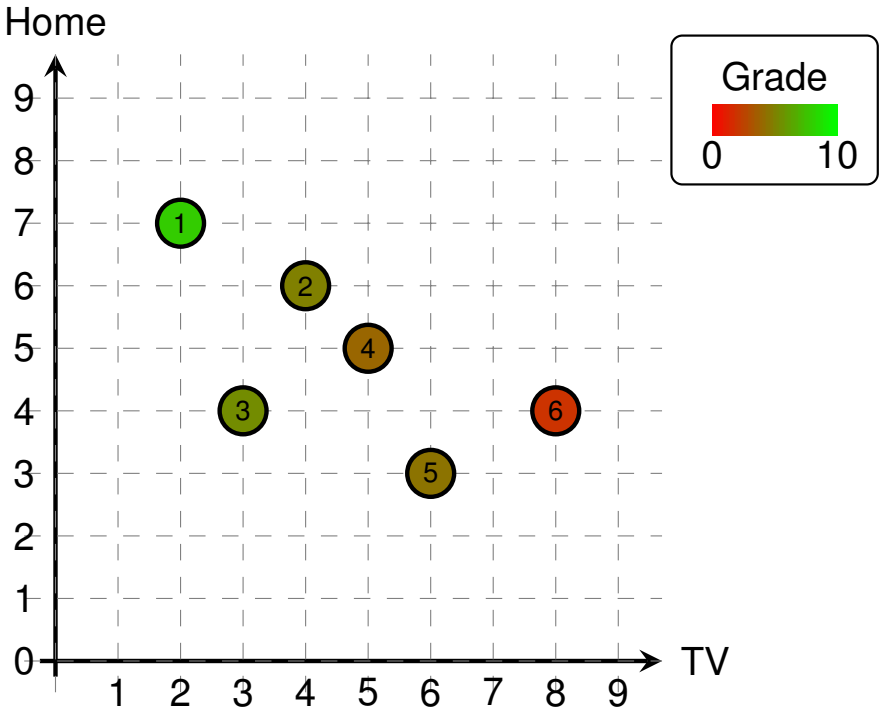


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

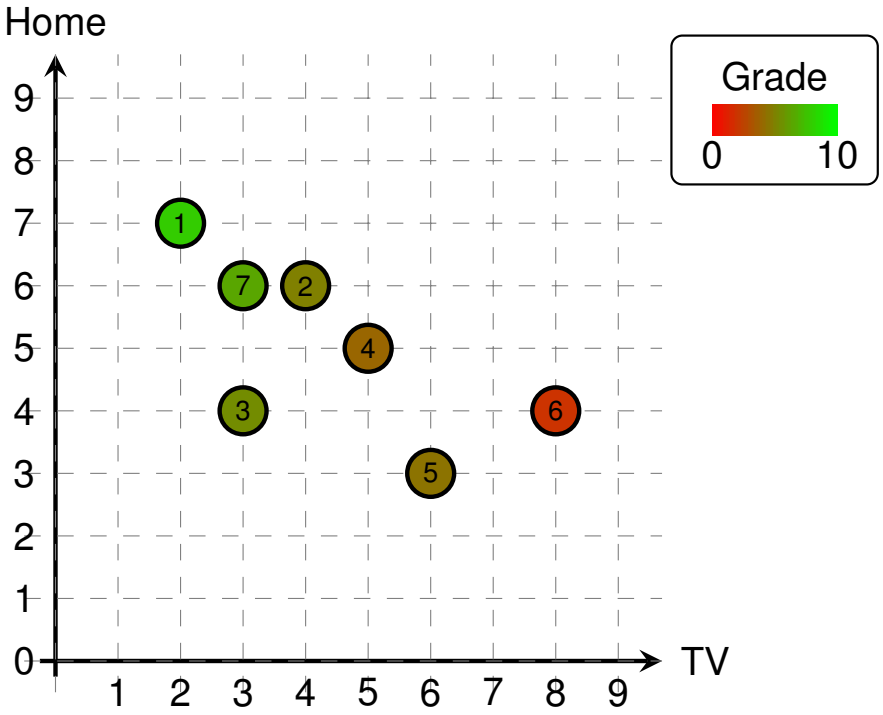


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

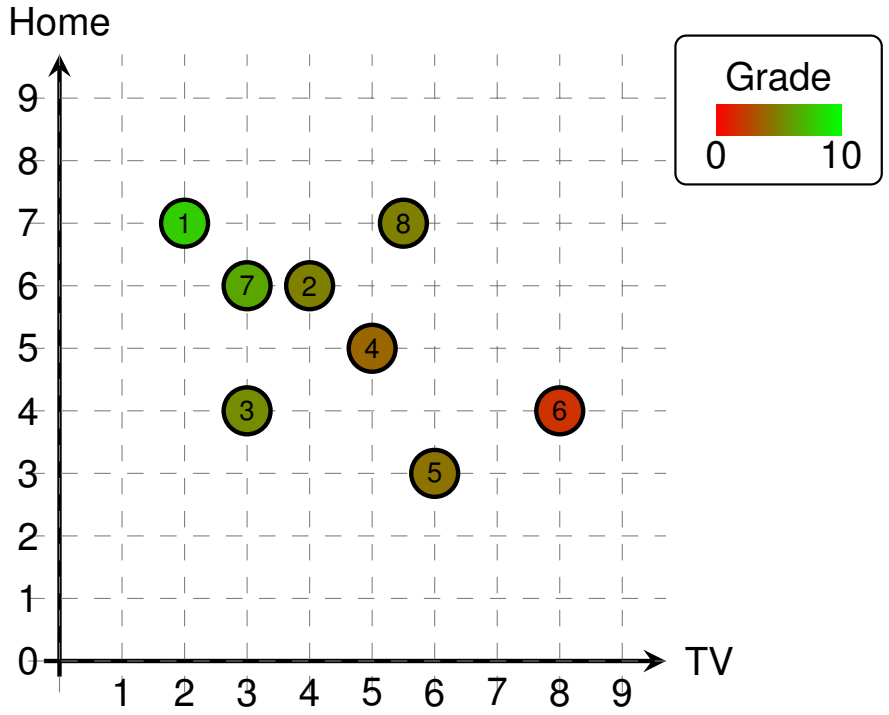


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

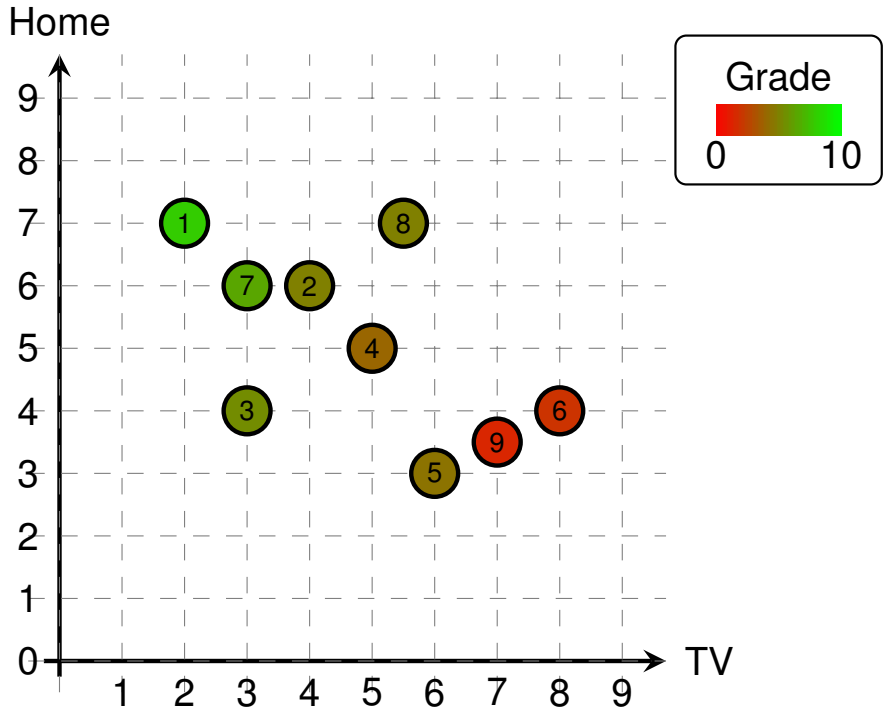


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

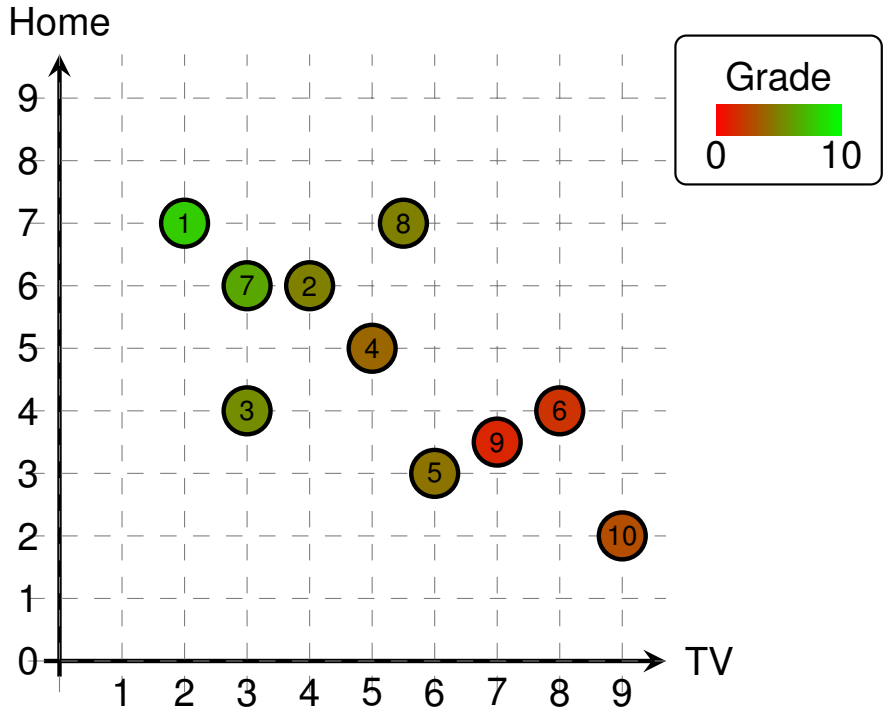


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

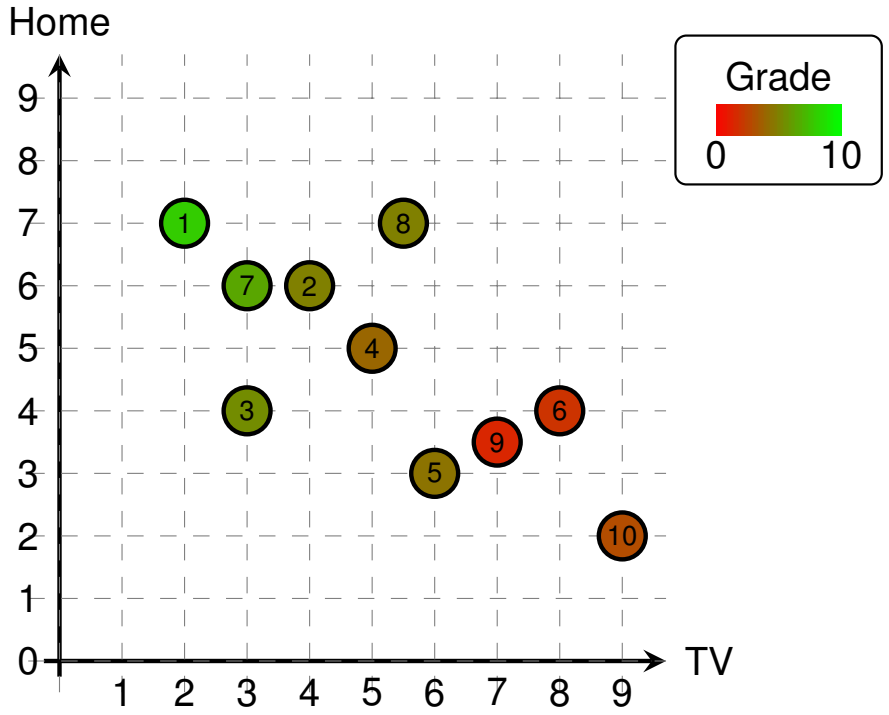


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam	Pred.
1	2	7	8	7.04
2	4	6	5	5.52
3	3	4	5.5	5.96
4	5	5	4	4.75
5	6	3	4.5	3.75
6	8	4	2	2.50
7	3	6	6.5	6.22
8	5.5	7	5	4.62
9	7	3.5	1.5	3.13
10	9	2	3	1.55

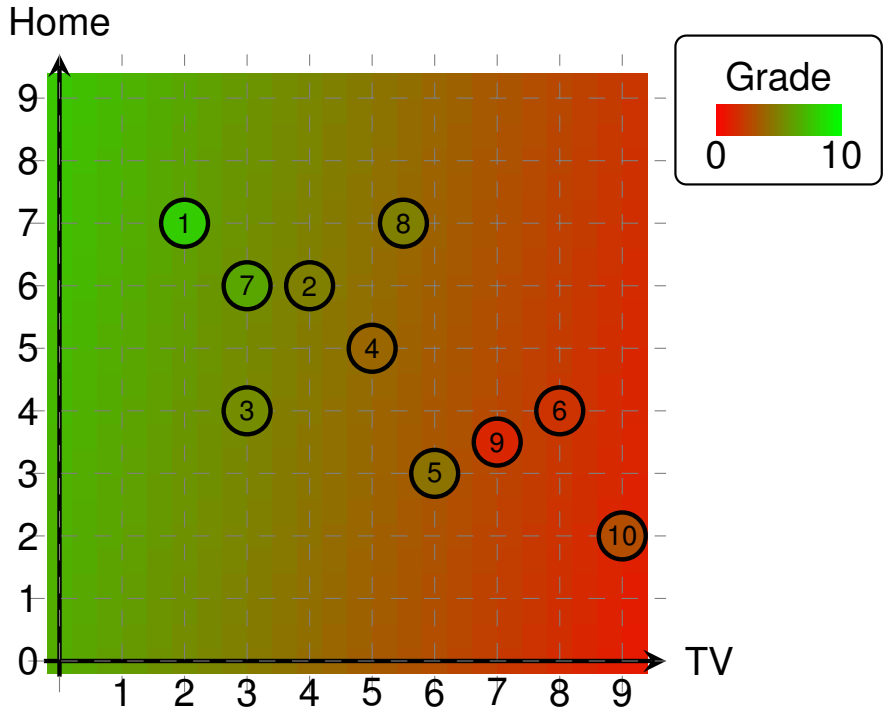


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam	Pred.
1	2	7	8	7.04
2	4	6	5	5.52
3	3	4	5.5	5.96
4	5	5	4	4.75
5	6	3	4.5	3.75
6	8	4	2	2.50
7	3	6	6.5	6.22
8	5.5	7	5	4.62
9	7	3.5	1.5	3.13
10	9	2	3	1.55

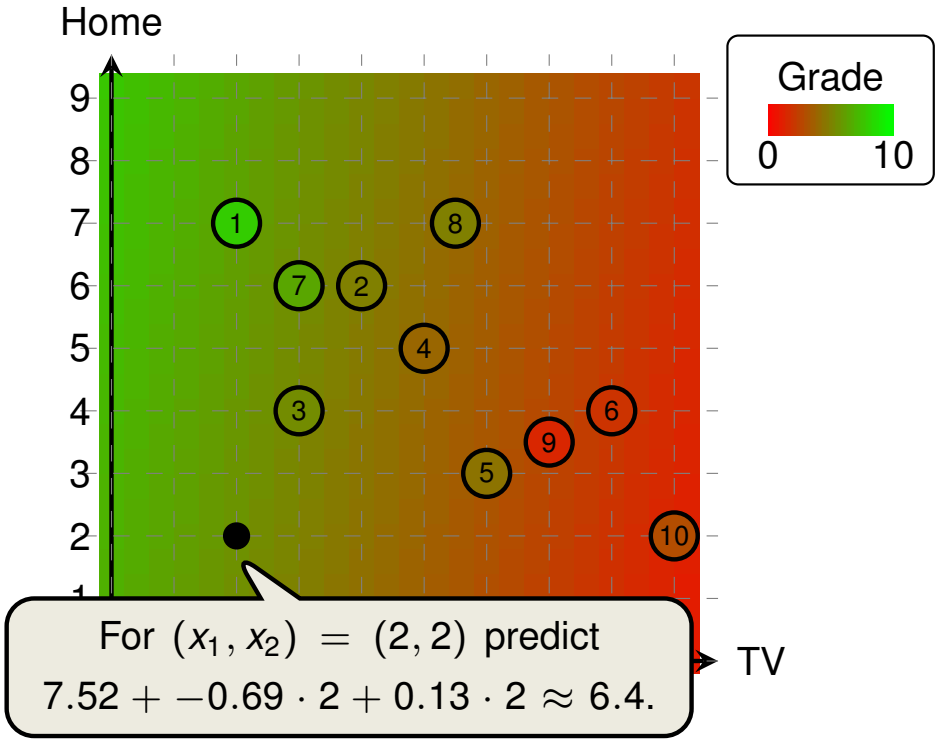


Multiple (Multi-Dimensional) Regression

$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

Stud.	TV	Home	Exam	Pred.
1	2	7	8	7.04
2	4	6	5	5.52
3	3	4	5.5	5.96
4	5	5	4	4.75
5	6	3	4.5	3.75
6	8	4	2	2.50
7	3	6	6.5	6.22
8	5.5	7	5	4.62
9	7	3.5	1.5	3.13
10	9	2	3	1.55



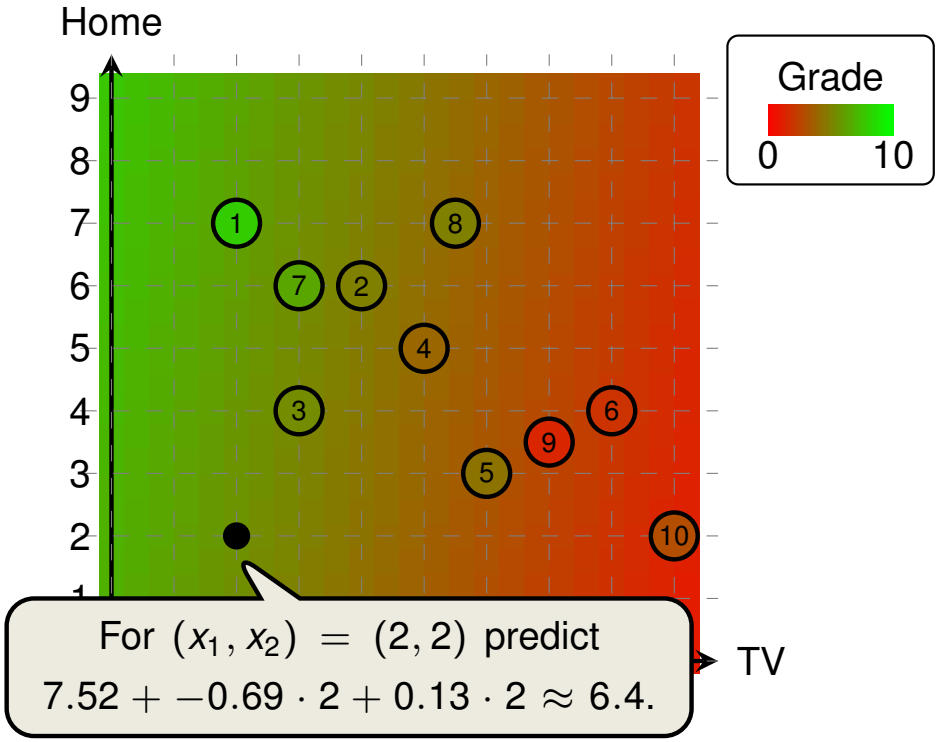
Multiple (Multi-Dimensional) Regression

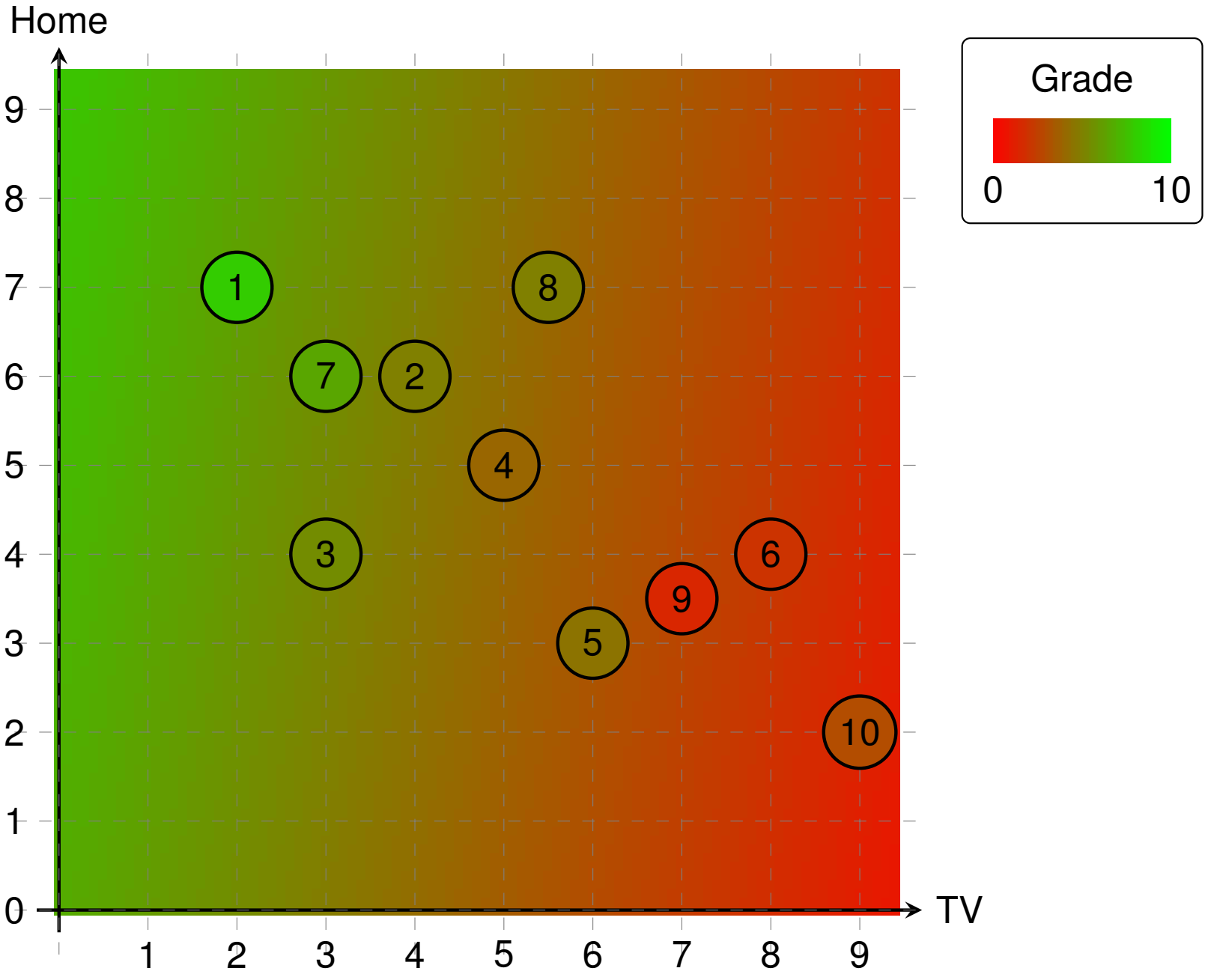
$$y = b + w_1 \cdot x_1 + w_2 \cdot x_2$$

- x_1 is TV hours, x_2 is homework performance and y is exam grade
- Solution is: $b = 7.52526$, $w_1 = -0.69262$, $w_2 = 0.12863$
- Square Error is: 7.691 (was 7.908 for the one-dim. regression)

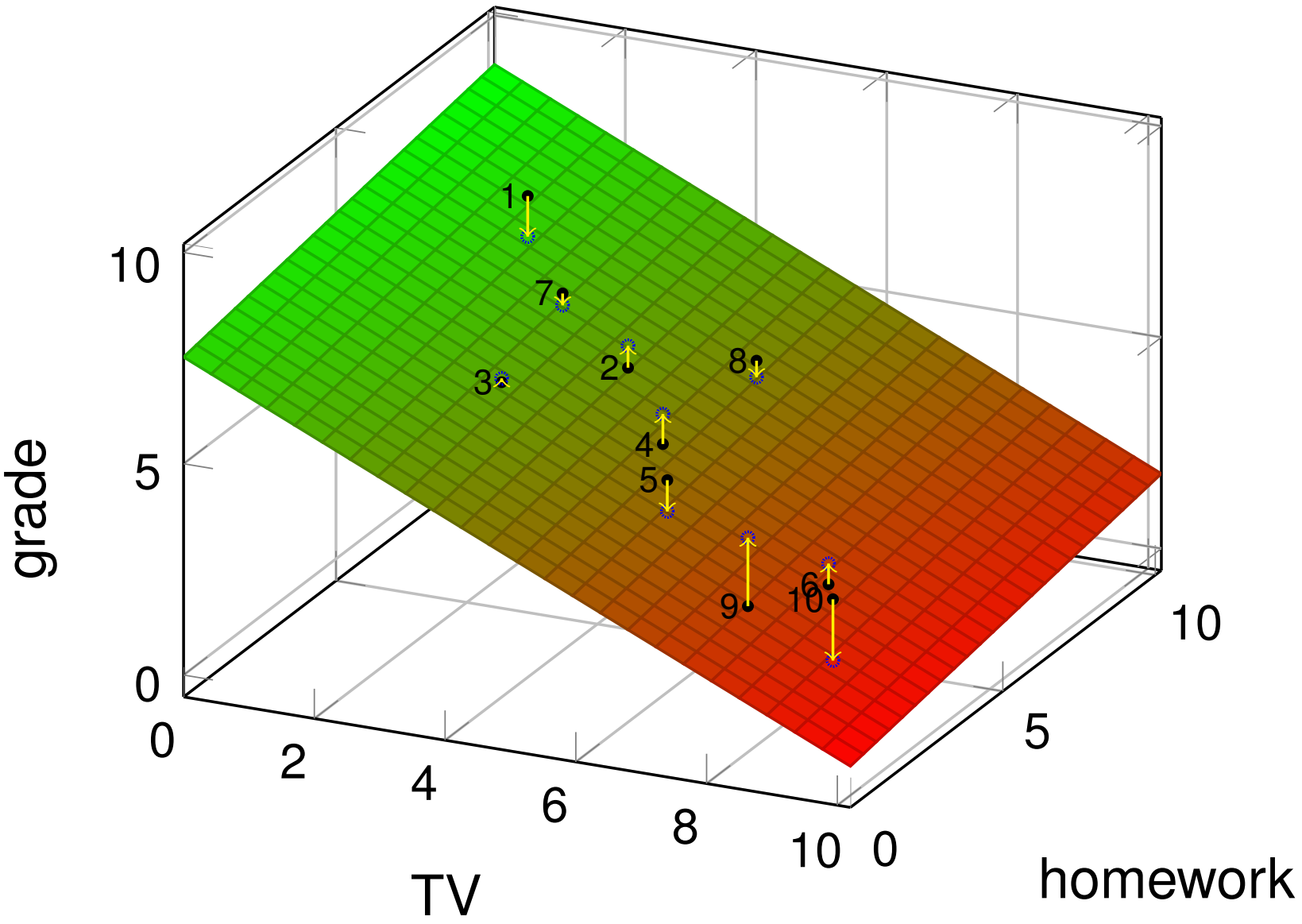
Stud.	TV	Home	Exam	Pred.
1	2	7	8	7.04
2	4	6	5	5.52
3	3	4	5.5	5.96
4	5	5	4	4.75
5	6	3	4.5	3.75
6	8	4	2	2.50
7	3	6	6.5	6.22
8	5.5	7	5	4.62
9	7	3.5	1.5	3.13
10	9	2	3	1.55

Largest error for Students 9 and 10!

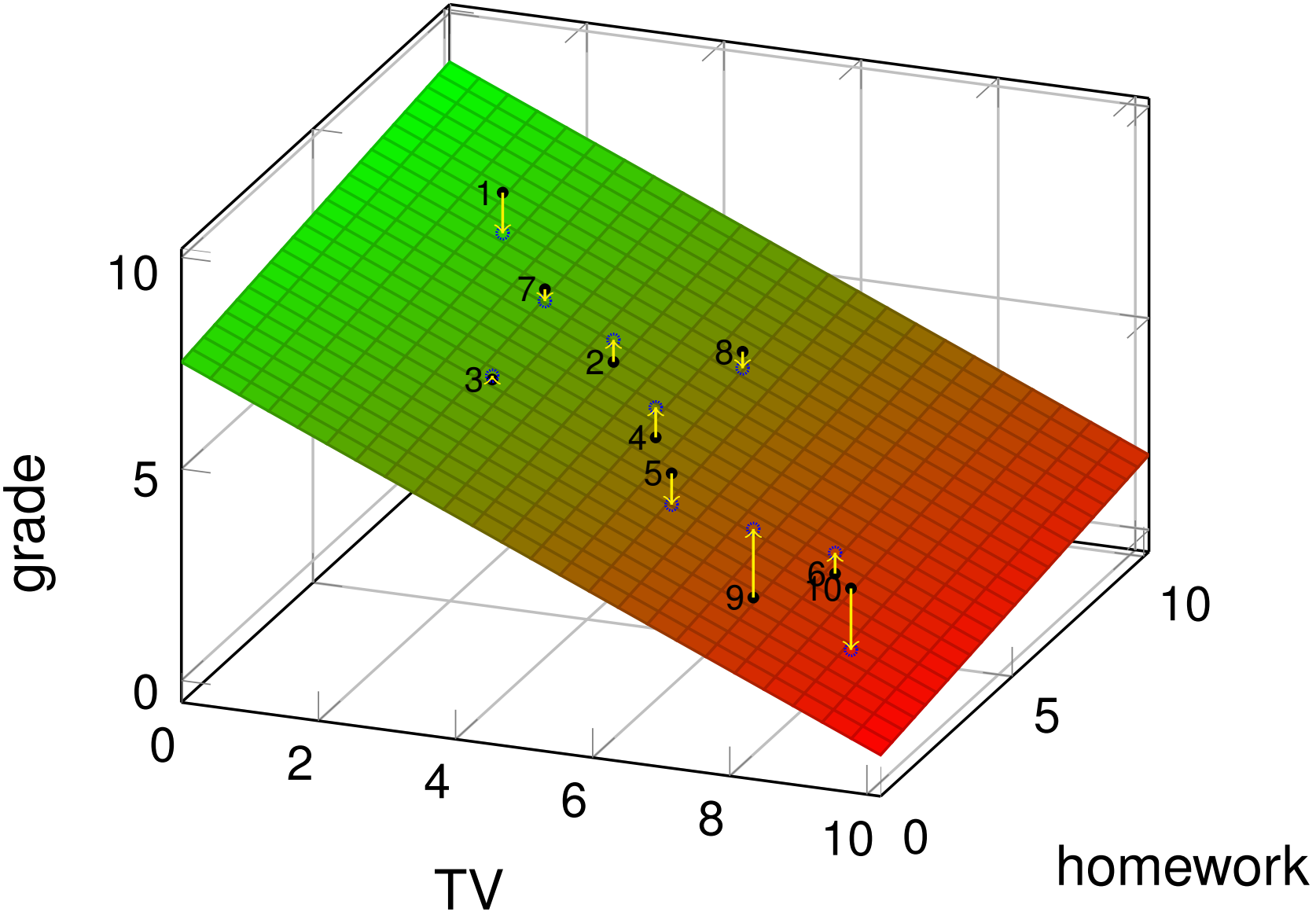



Home
TV
Grade
0
10
0
1
2
3
4
5
6
7
8
9
1
2
3
4
5
6
7
8
9
10

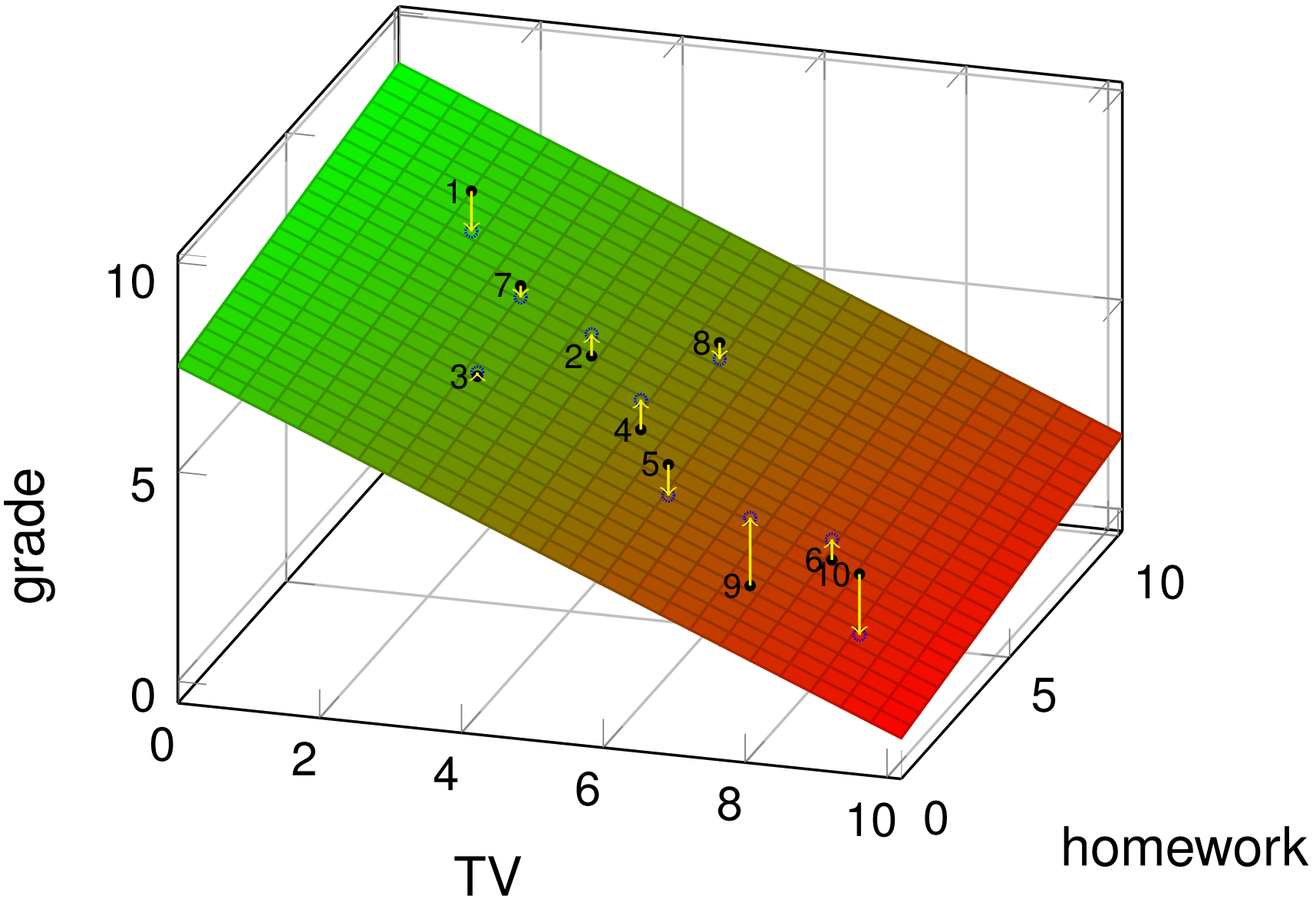
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



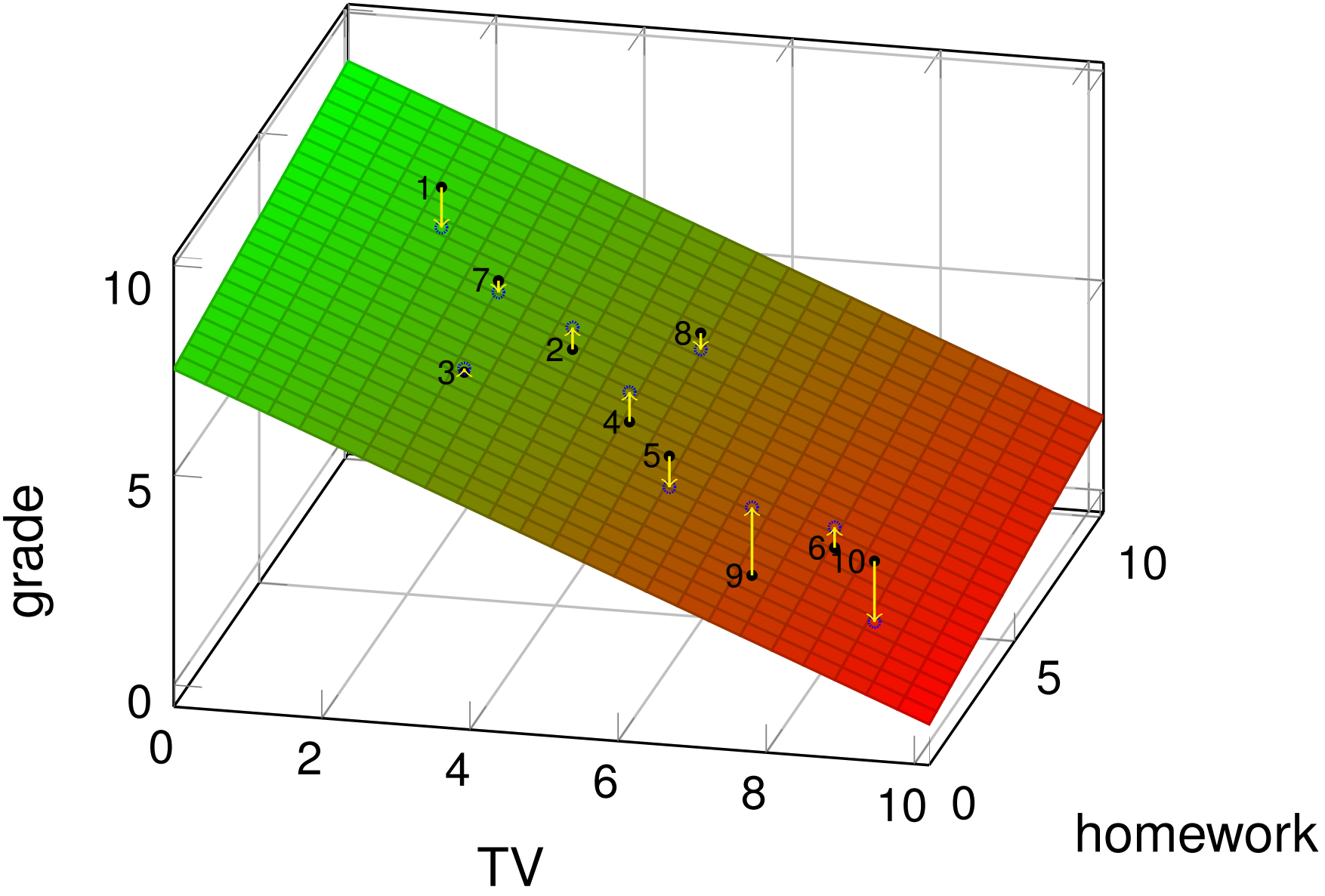
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



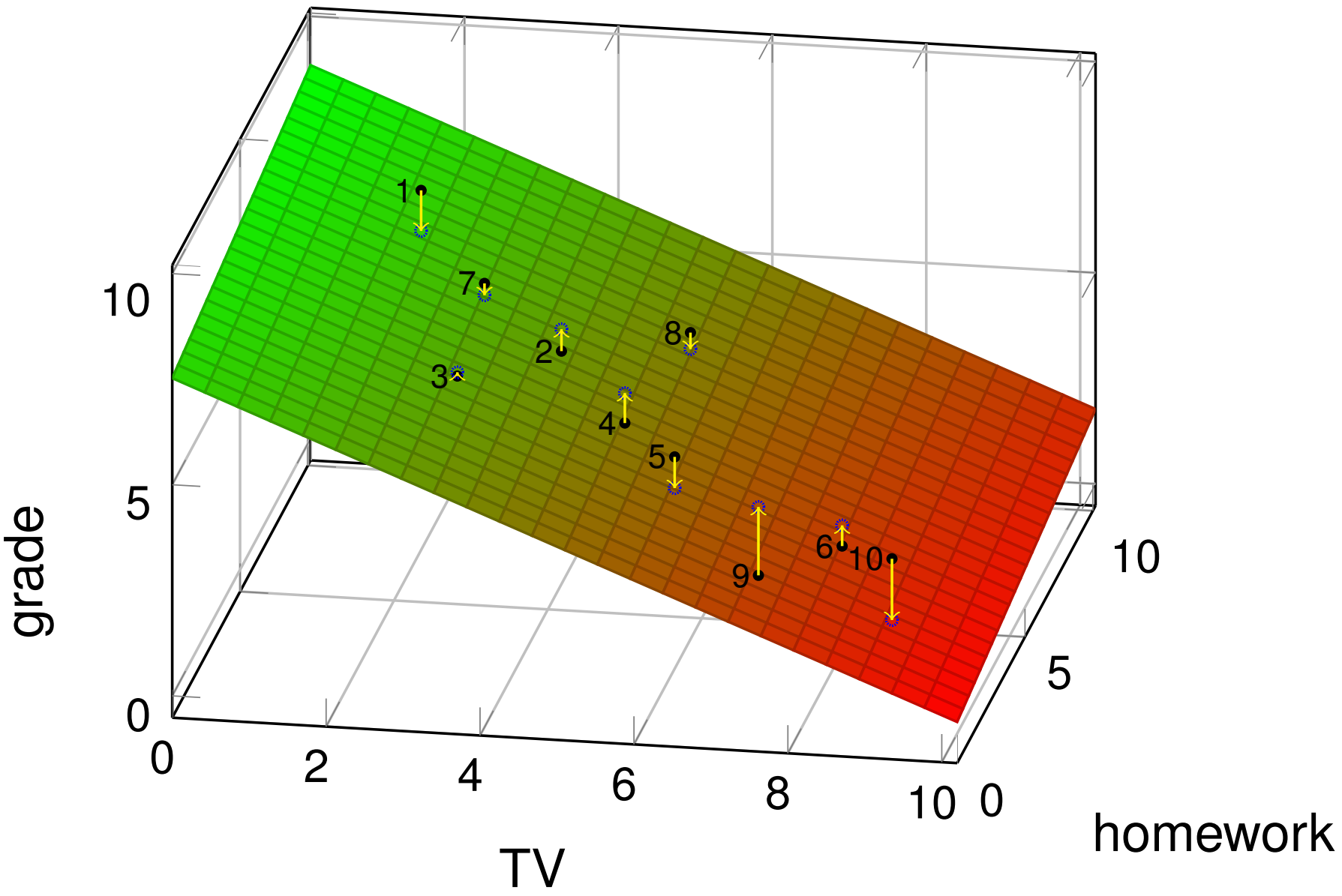
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



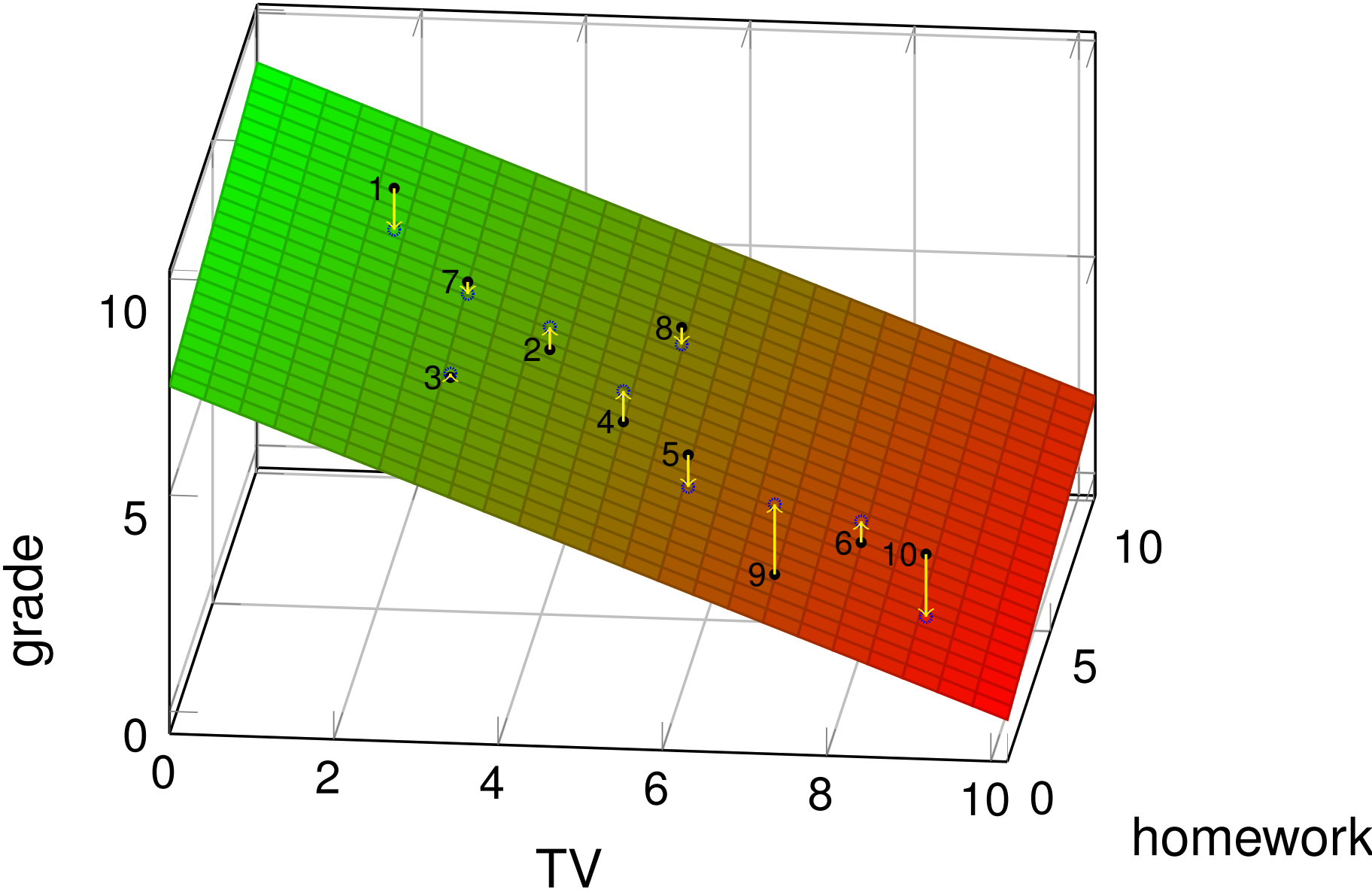
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



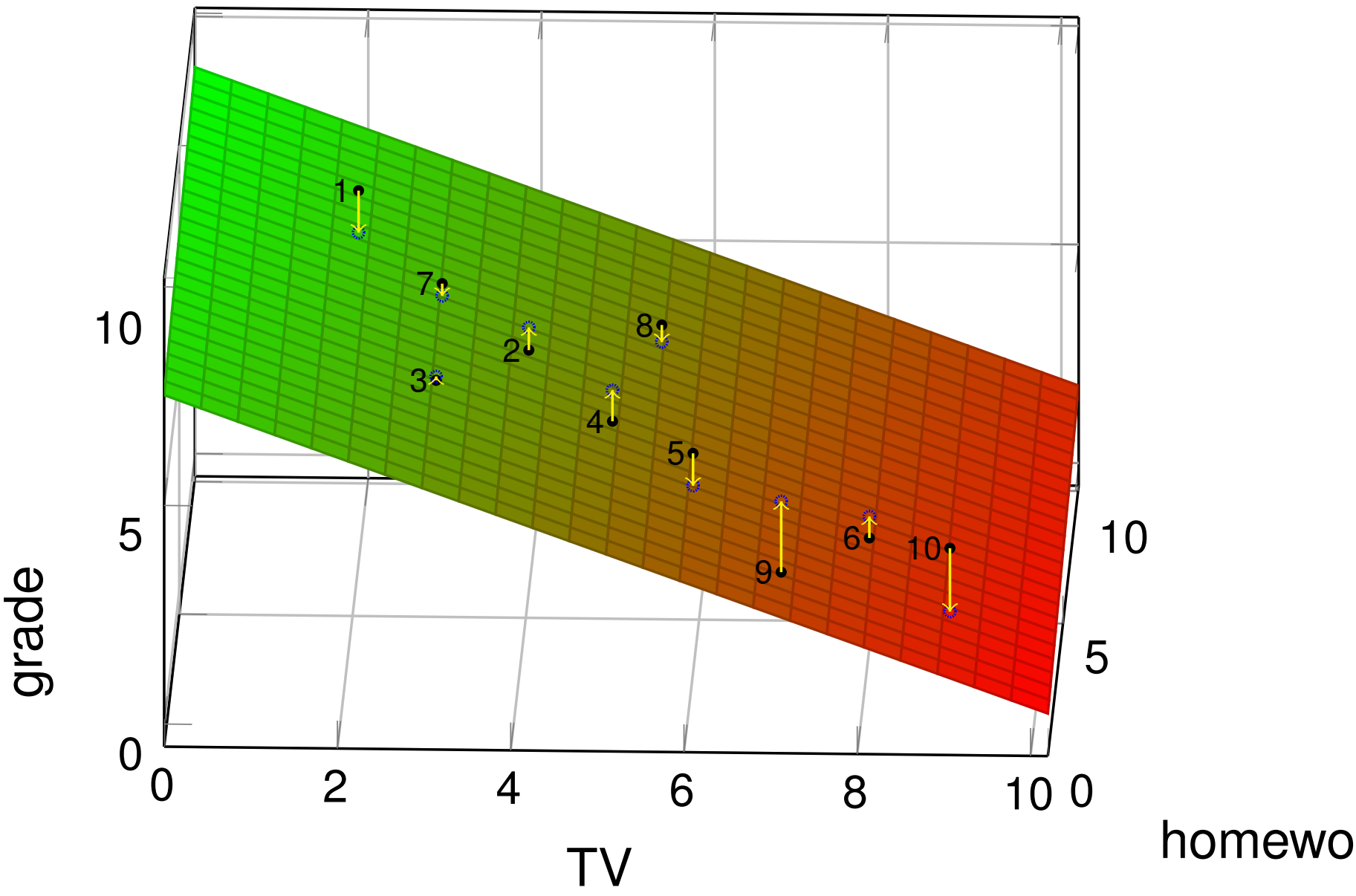
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



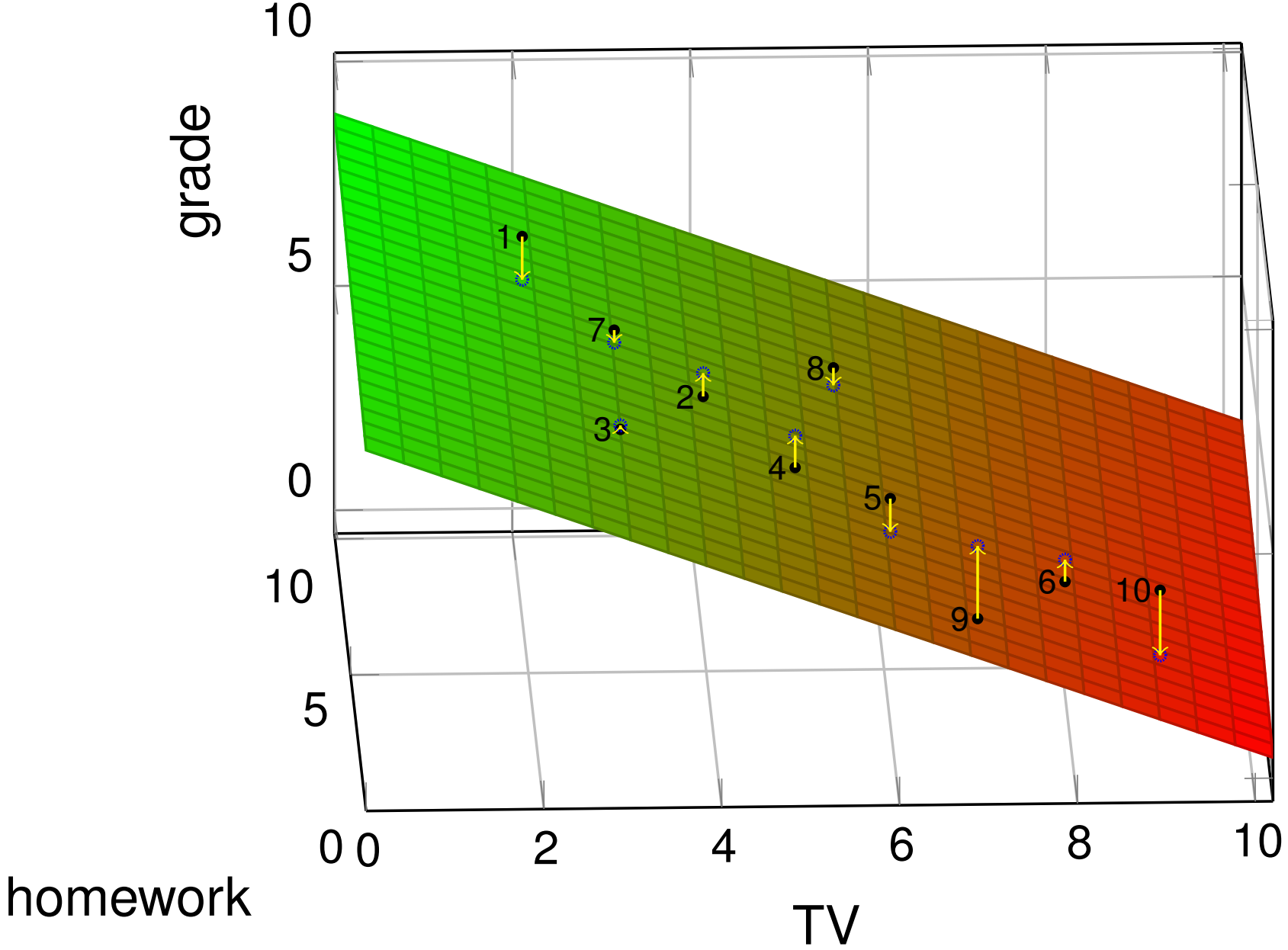
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



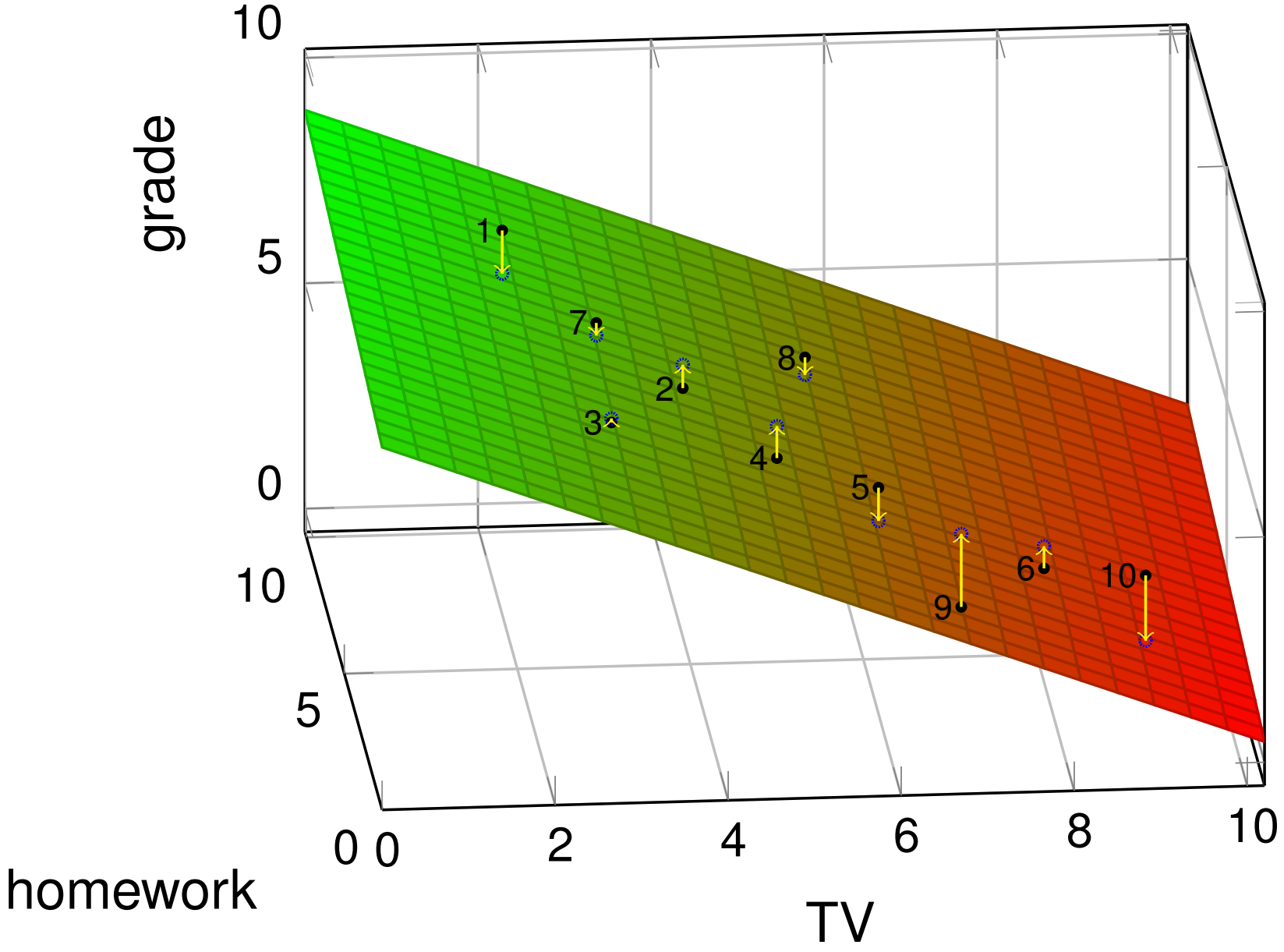
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



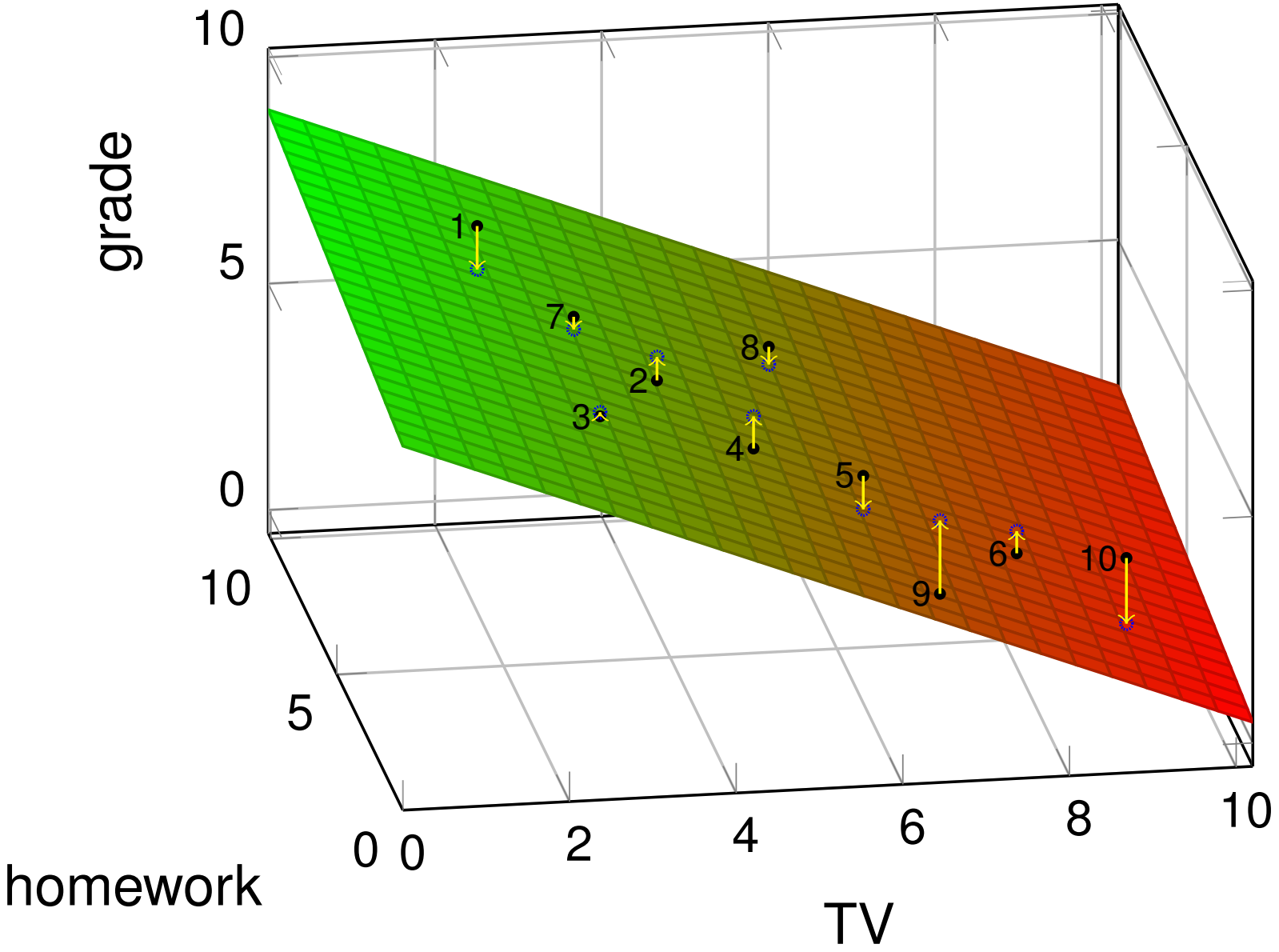
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



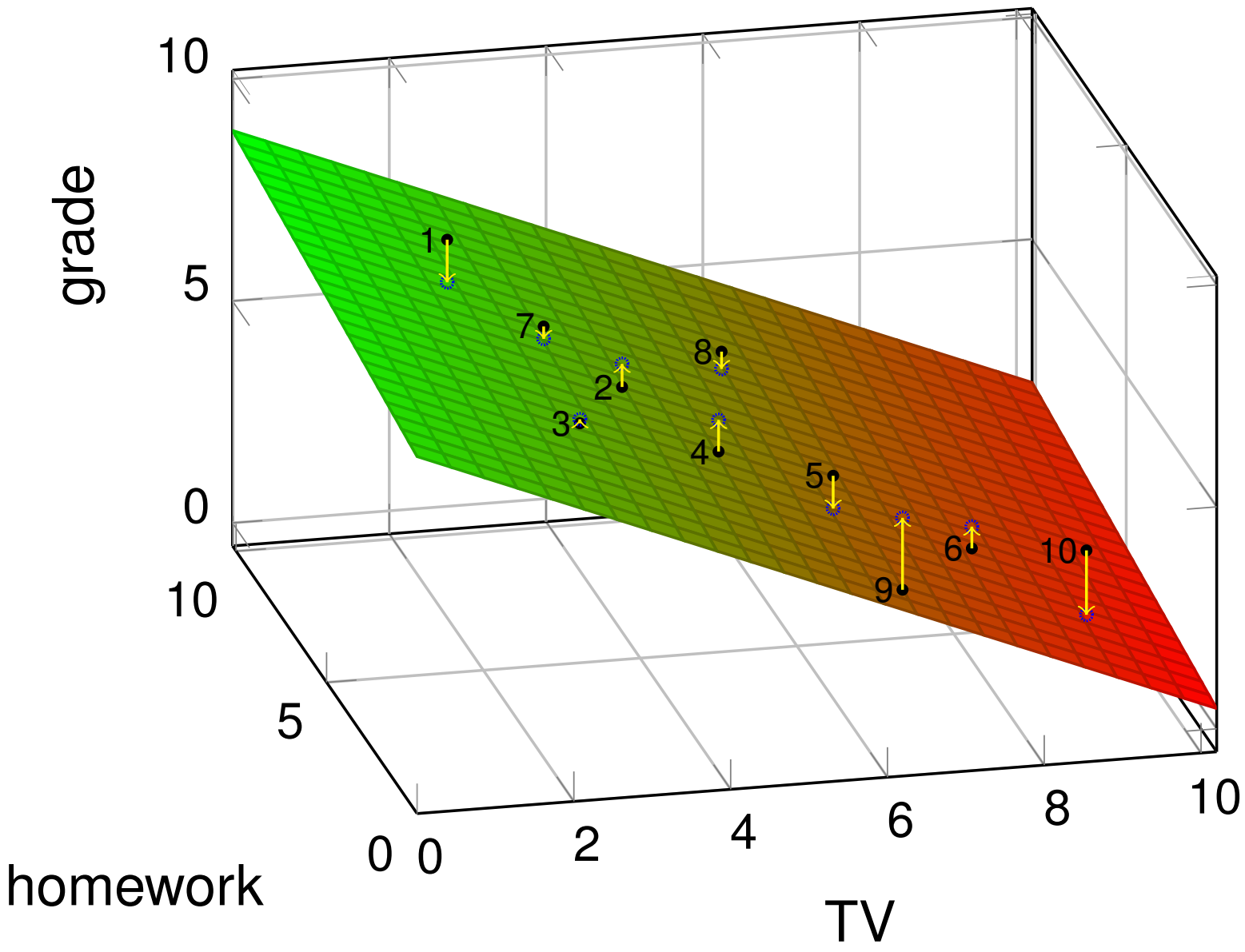
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



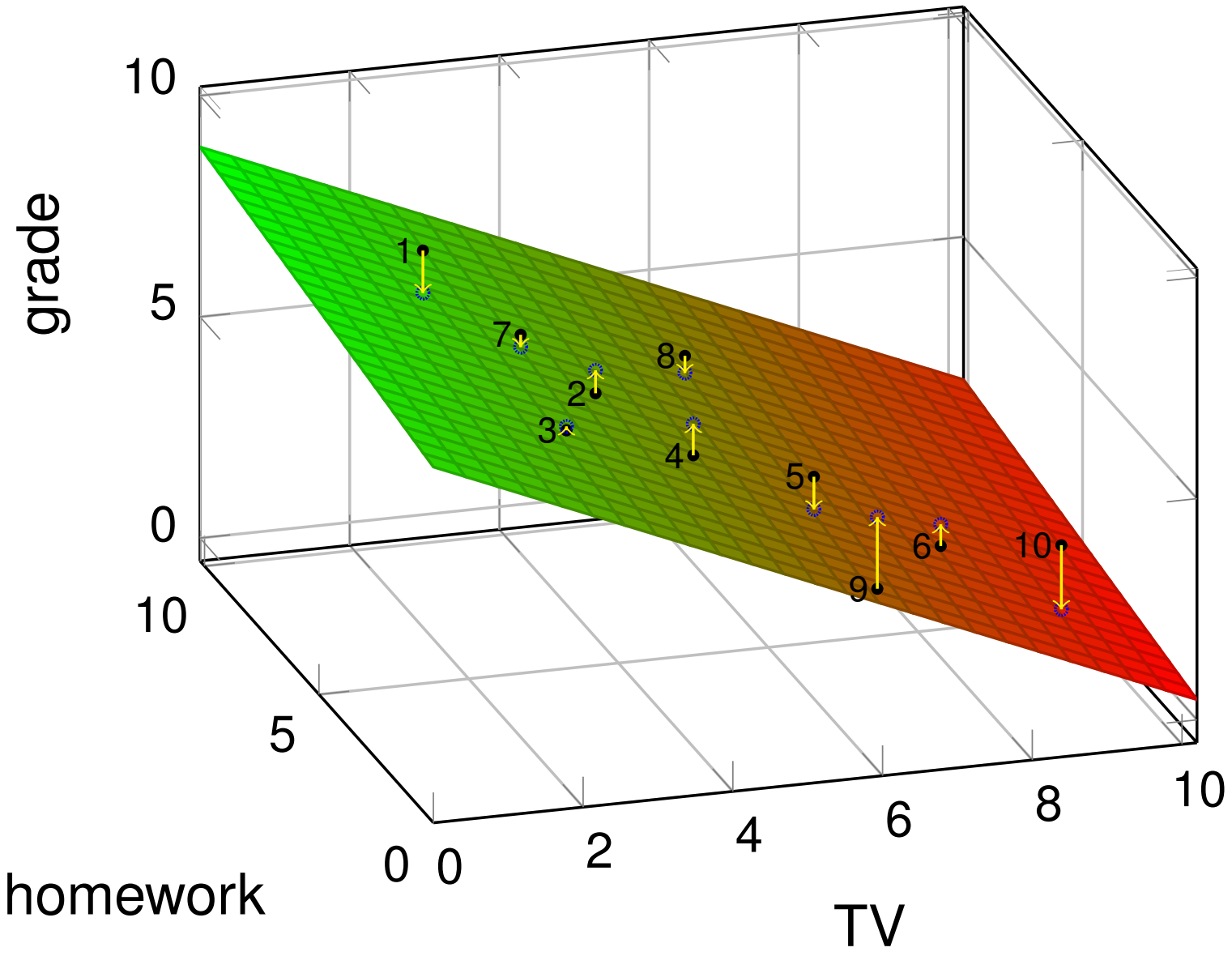
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



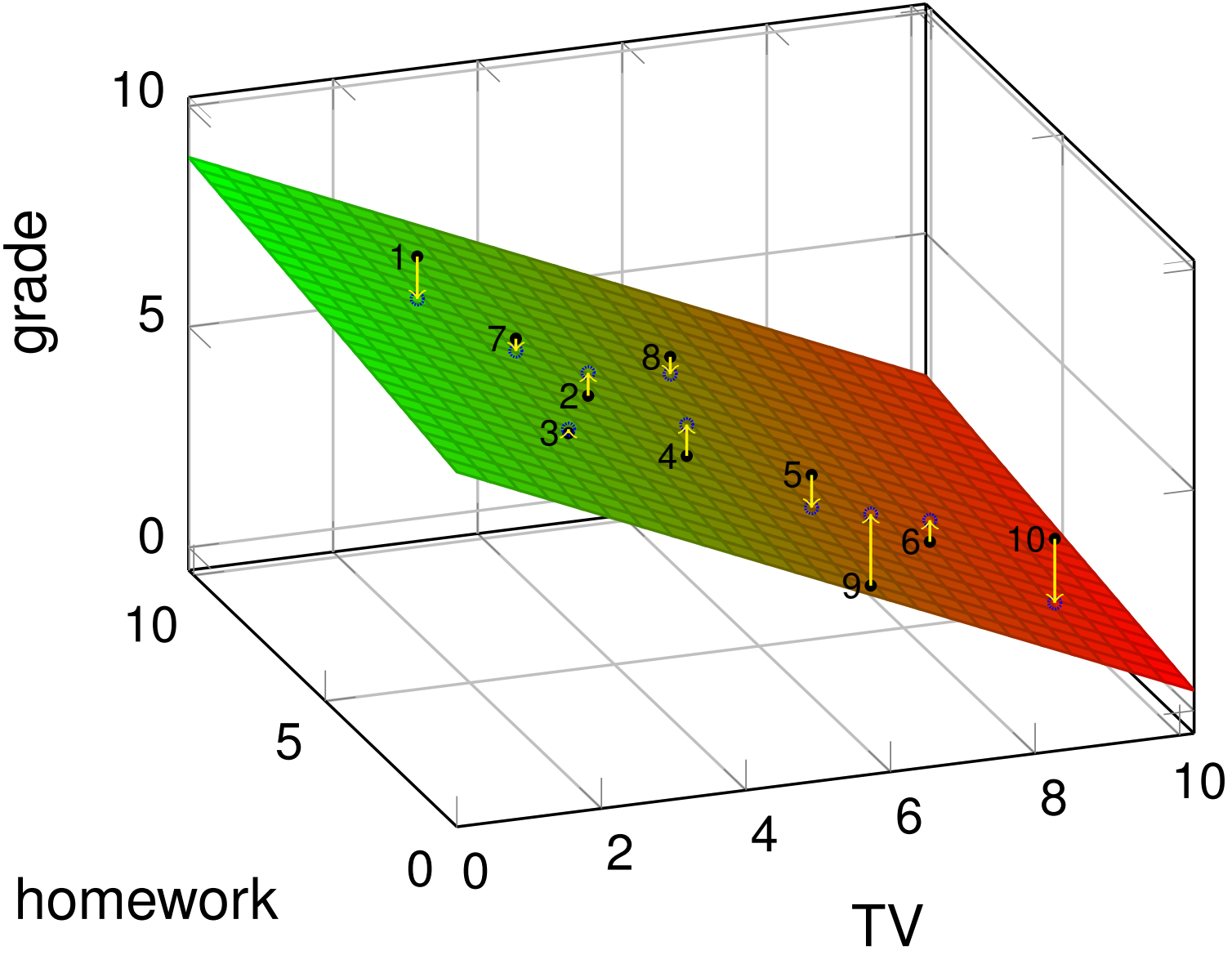
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



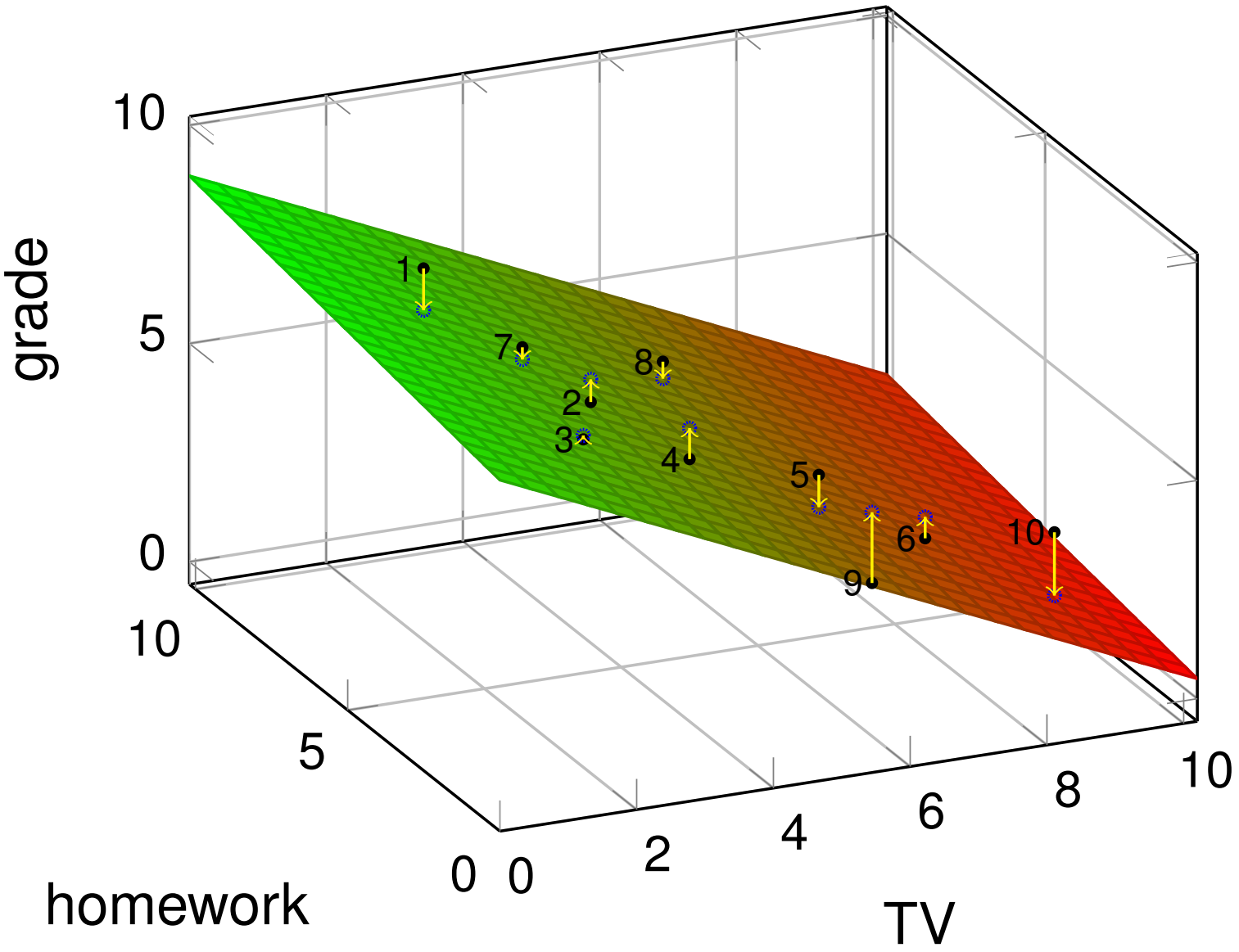
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



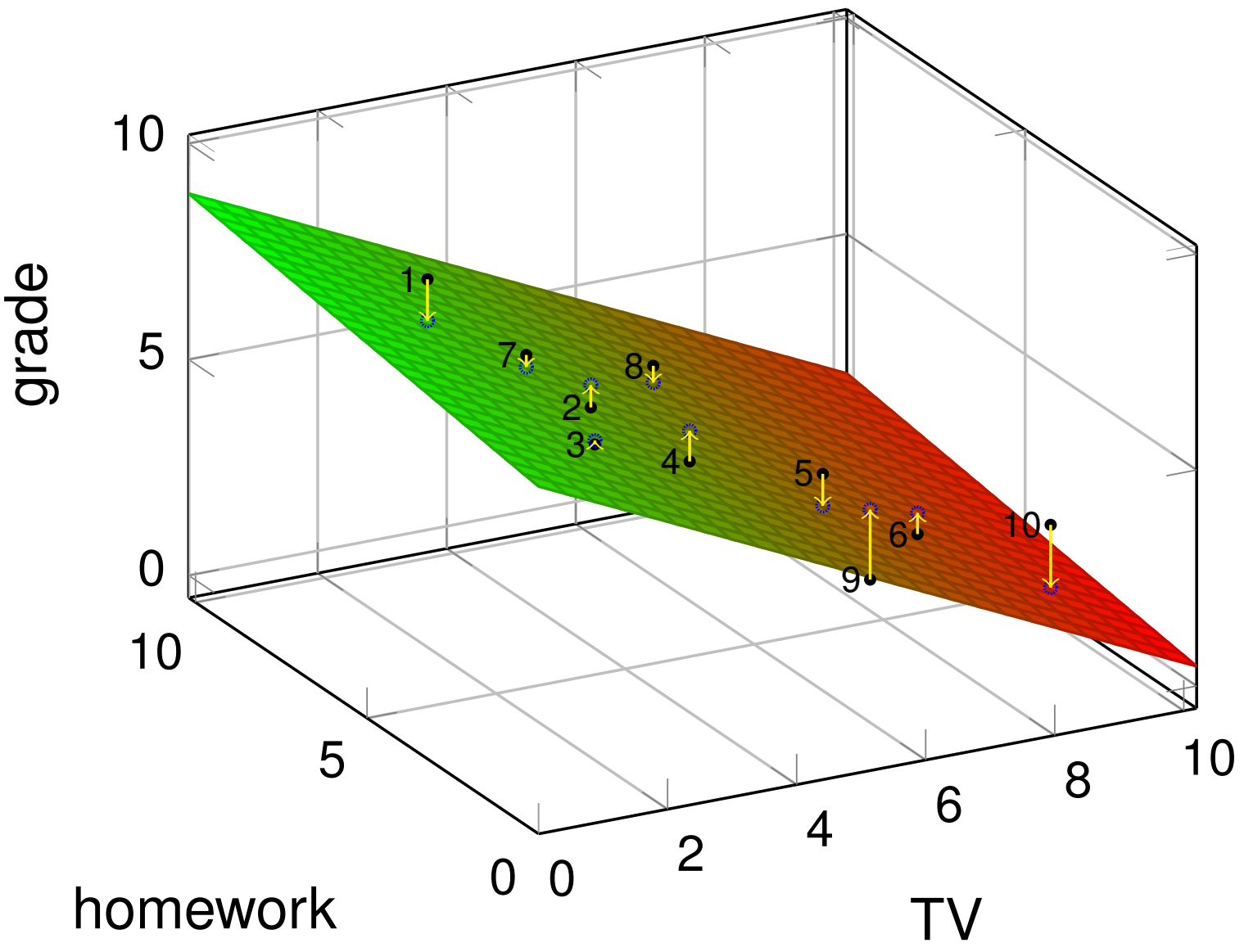
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



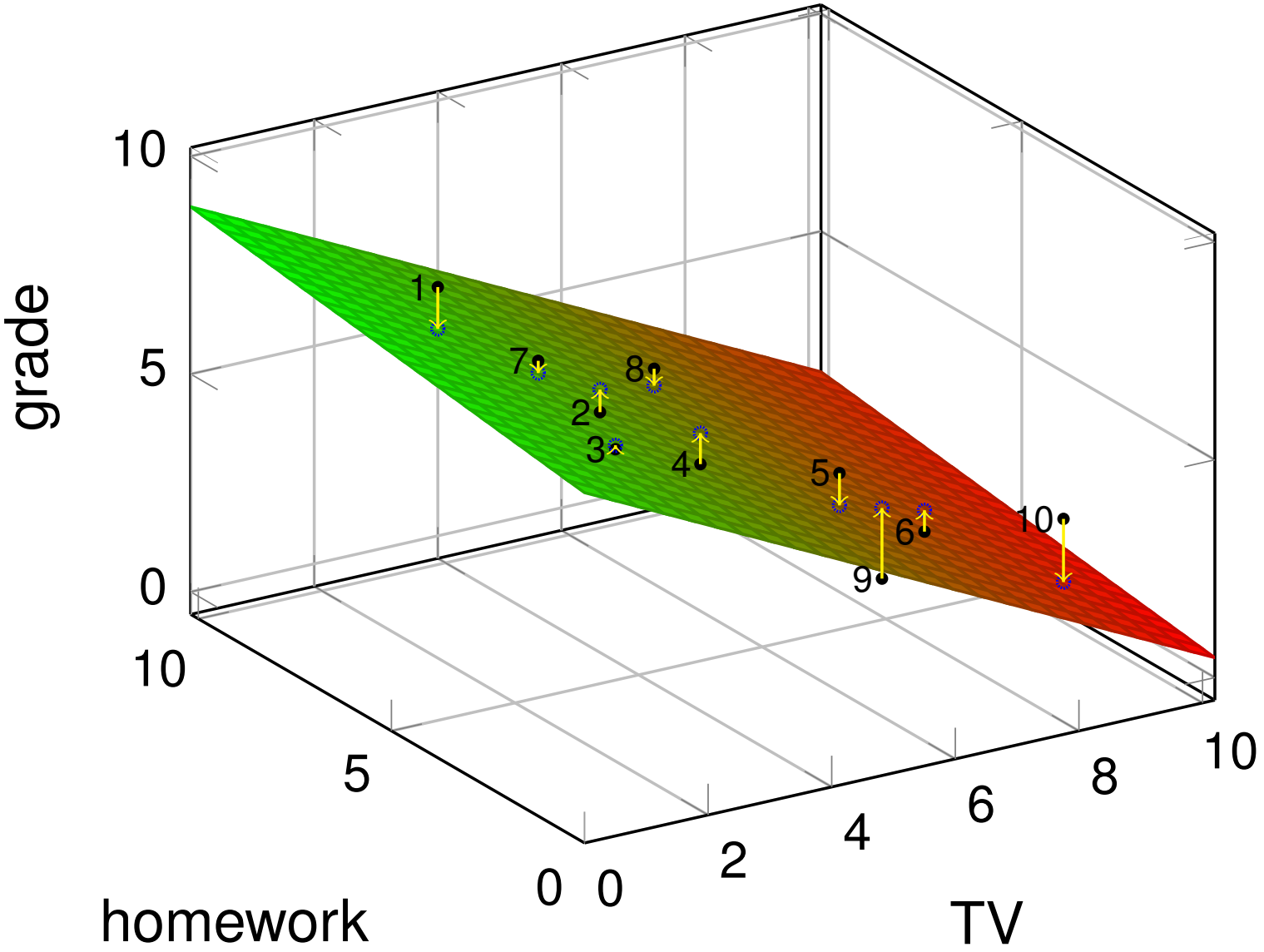
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



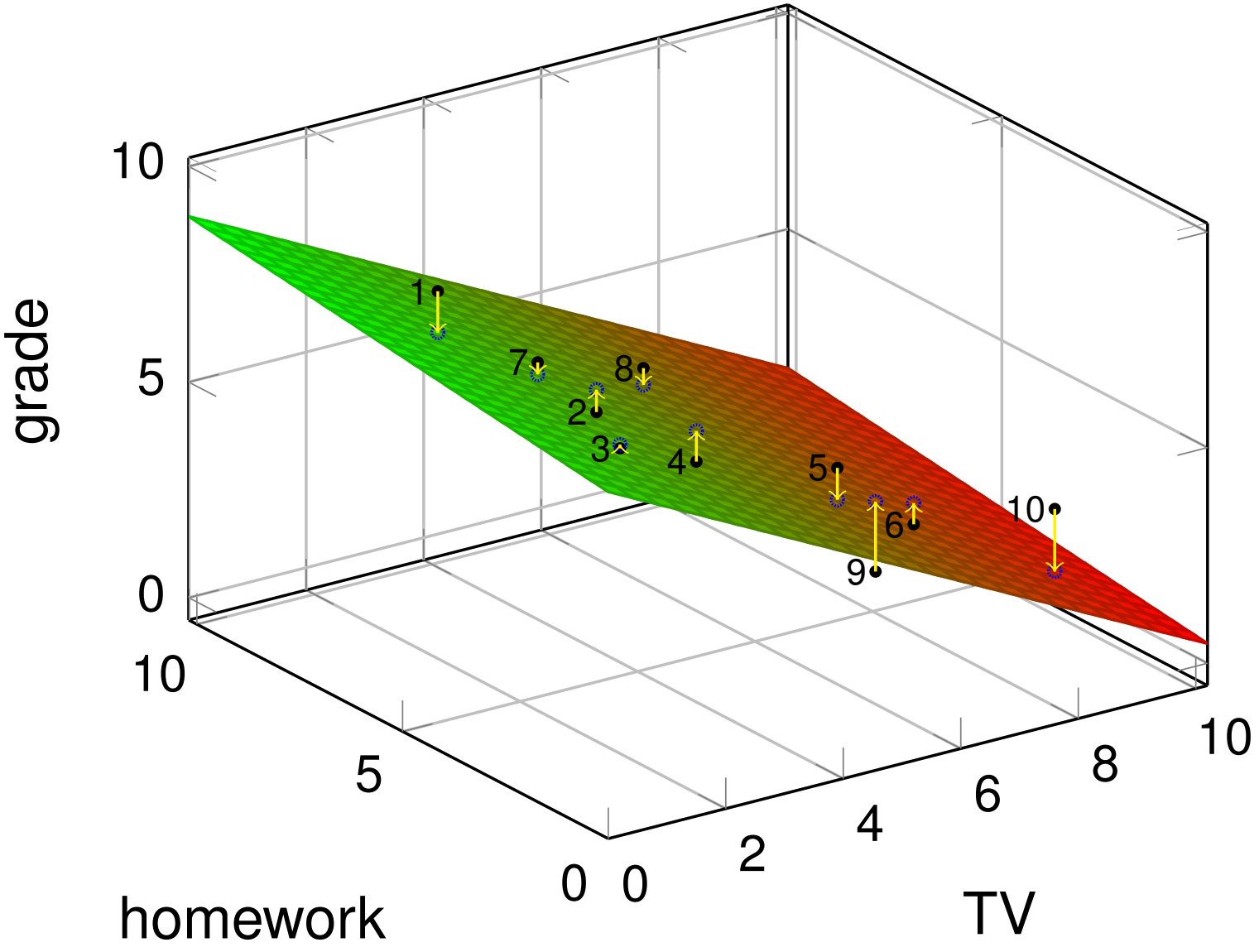
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



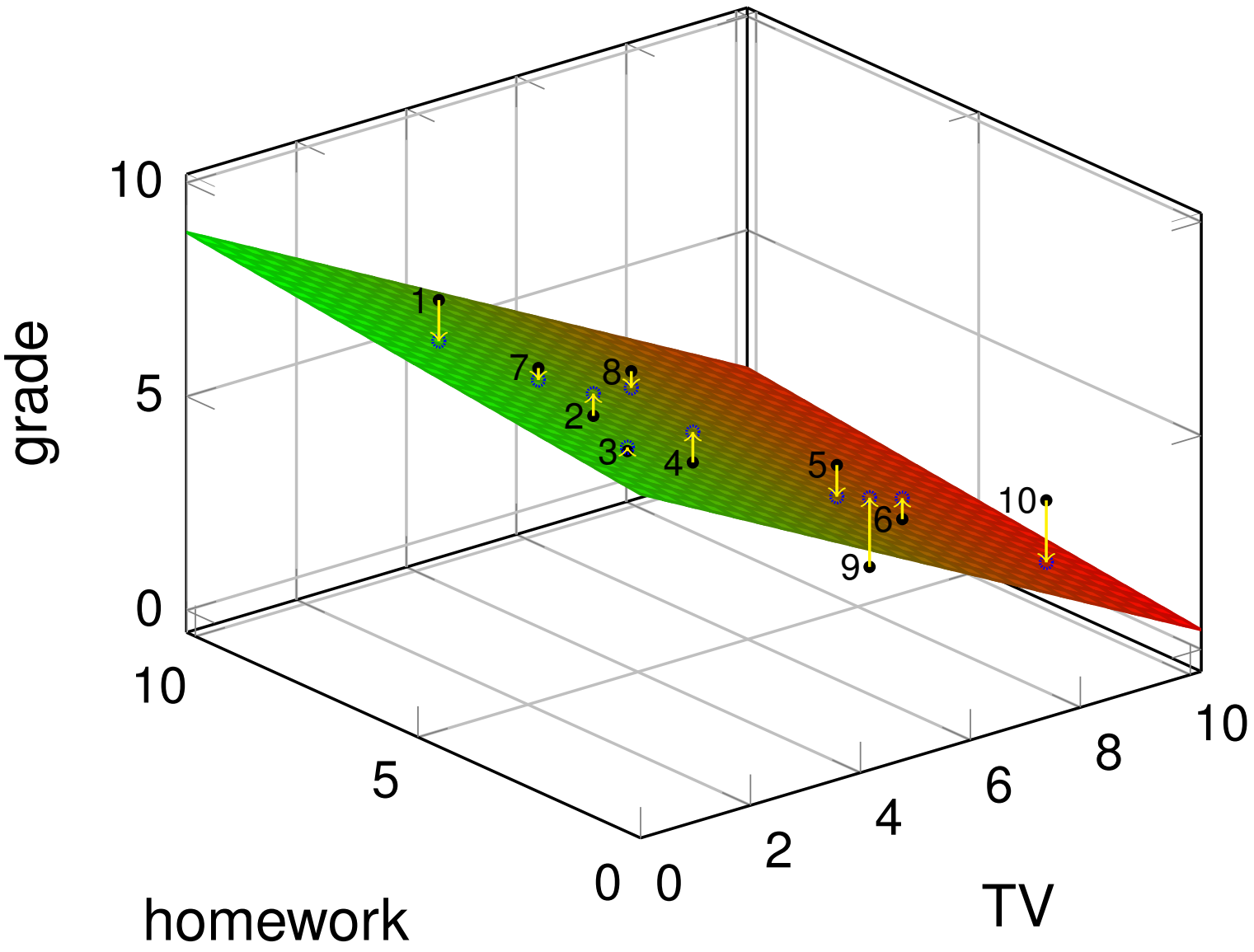
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



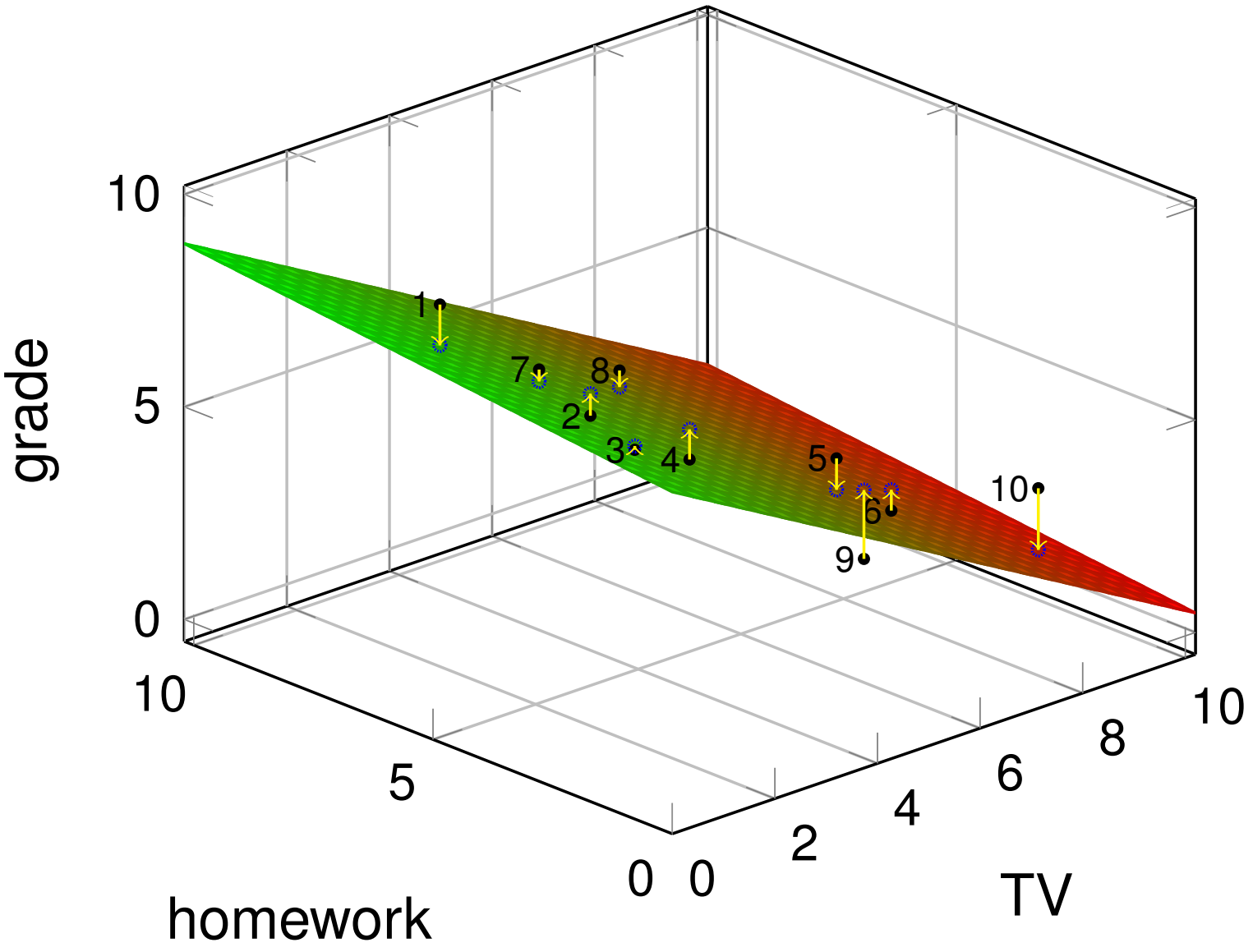
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



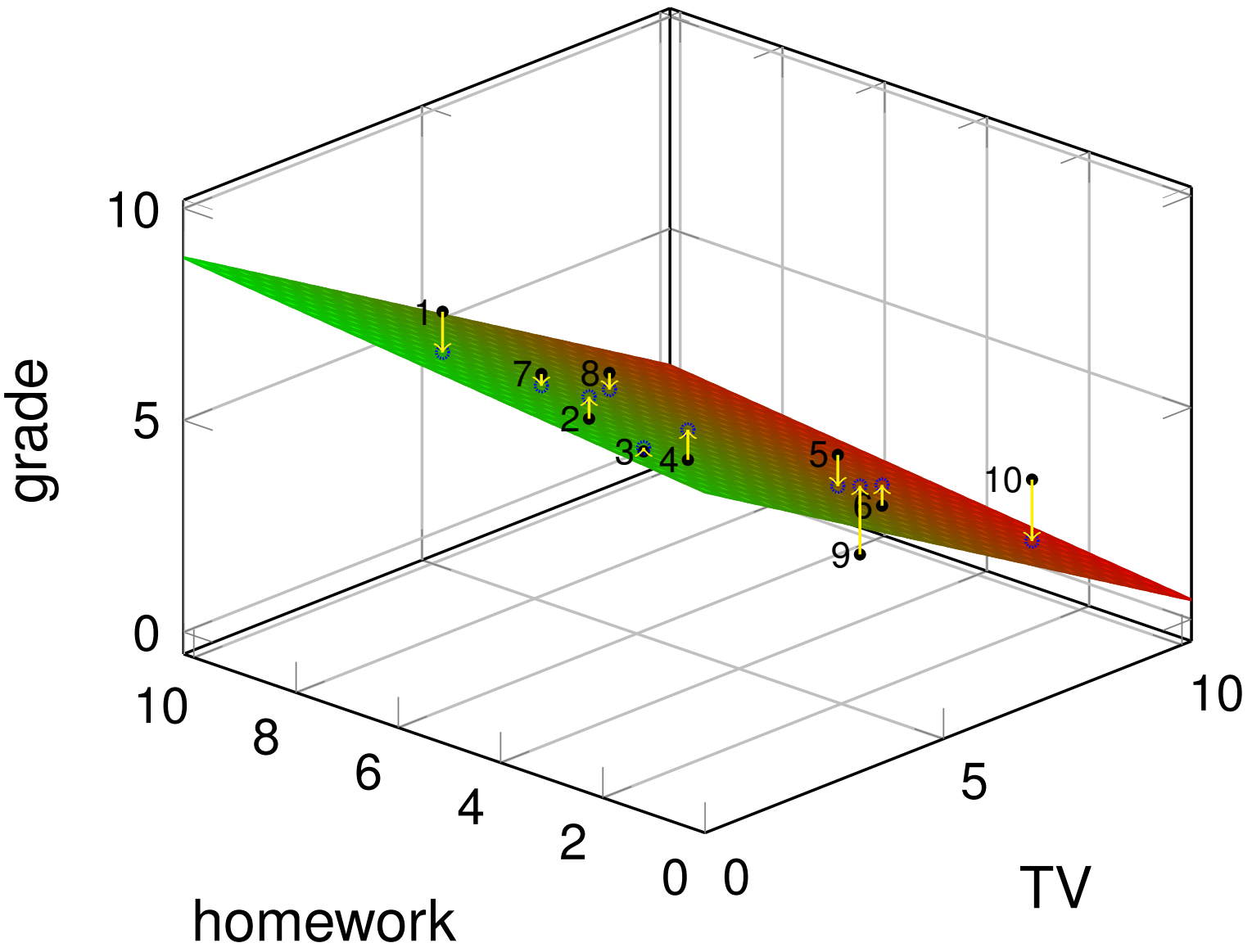
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



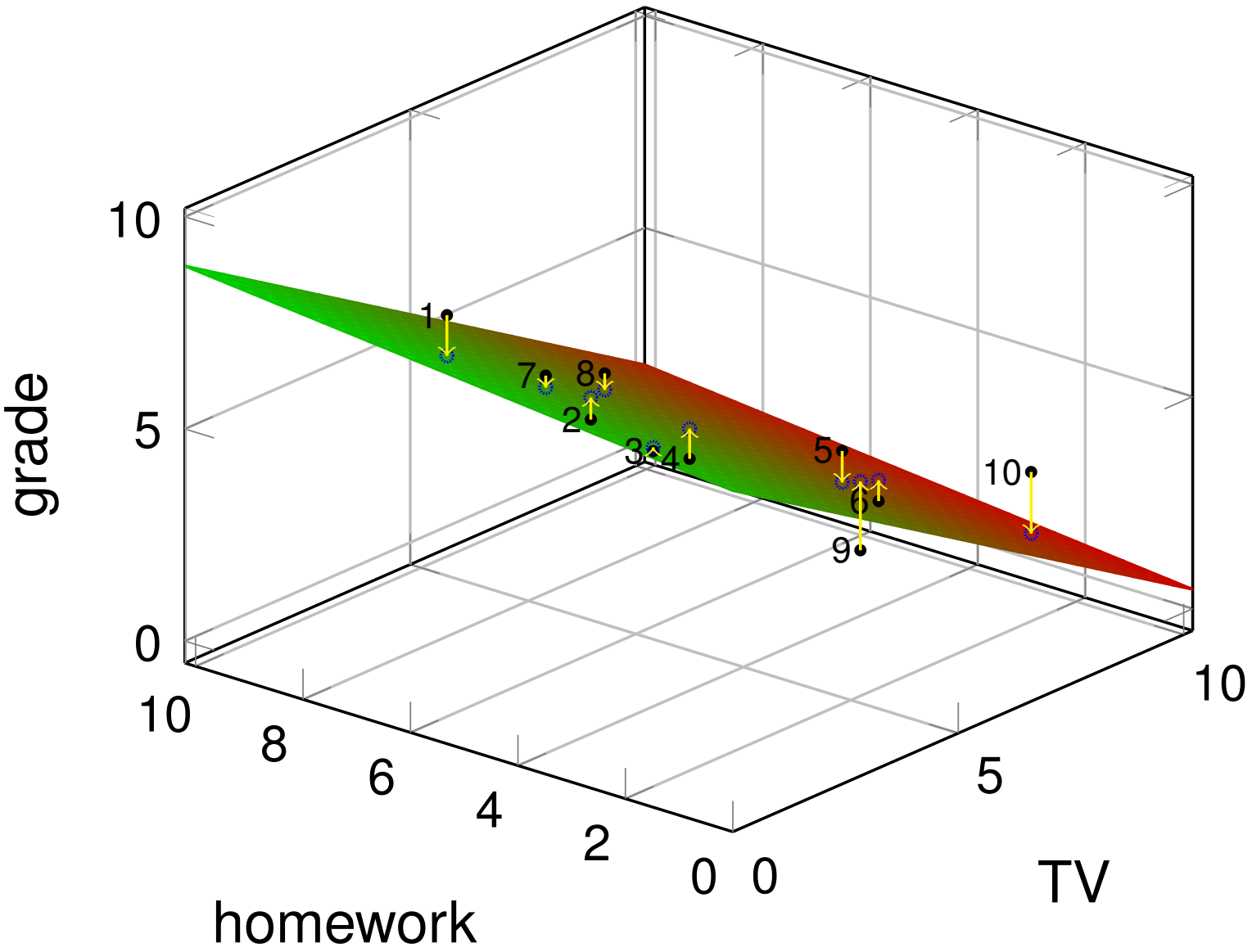
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



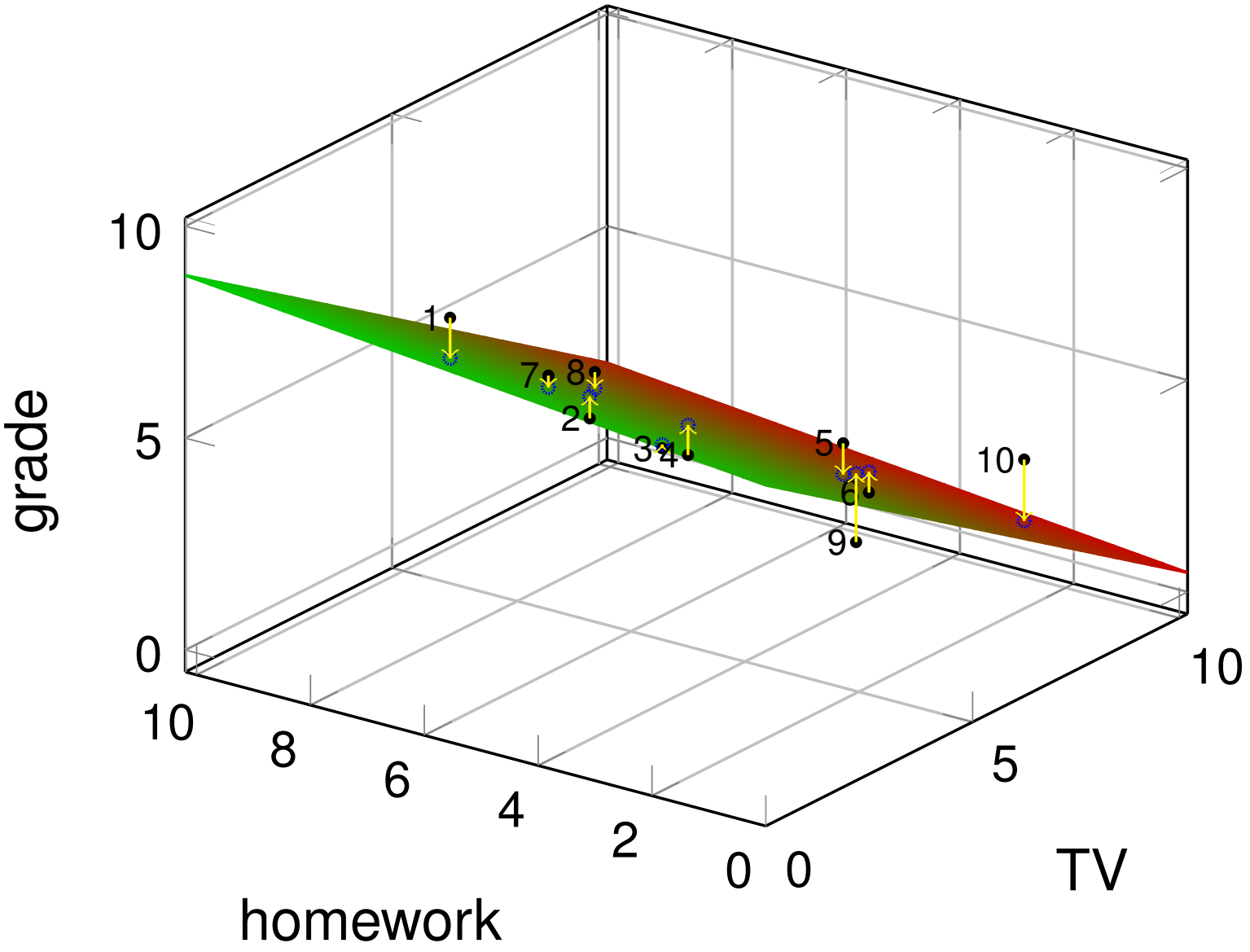
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



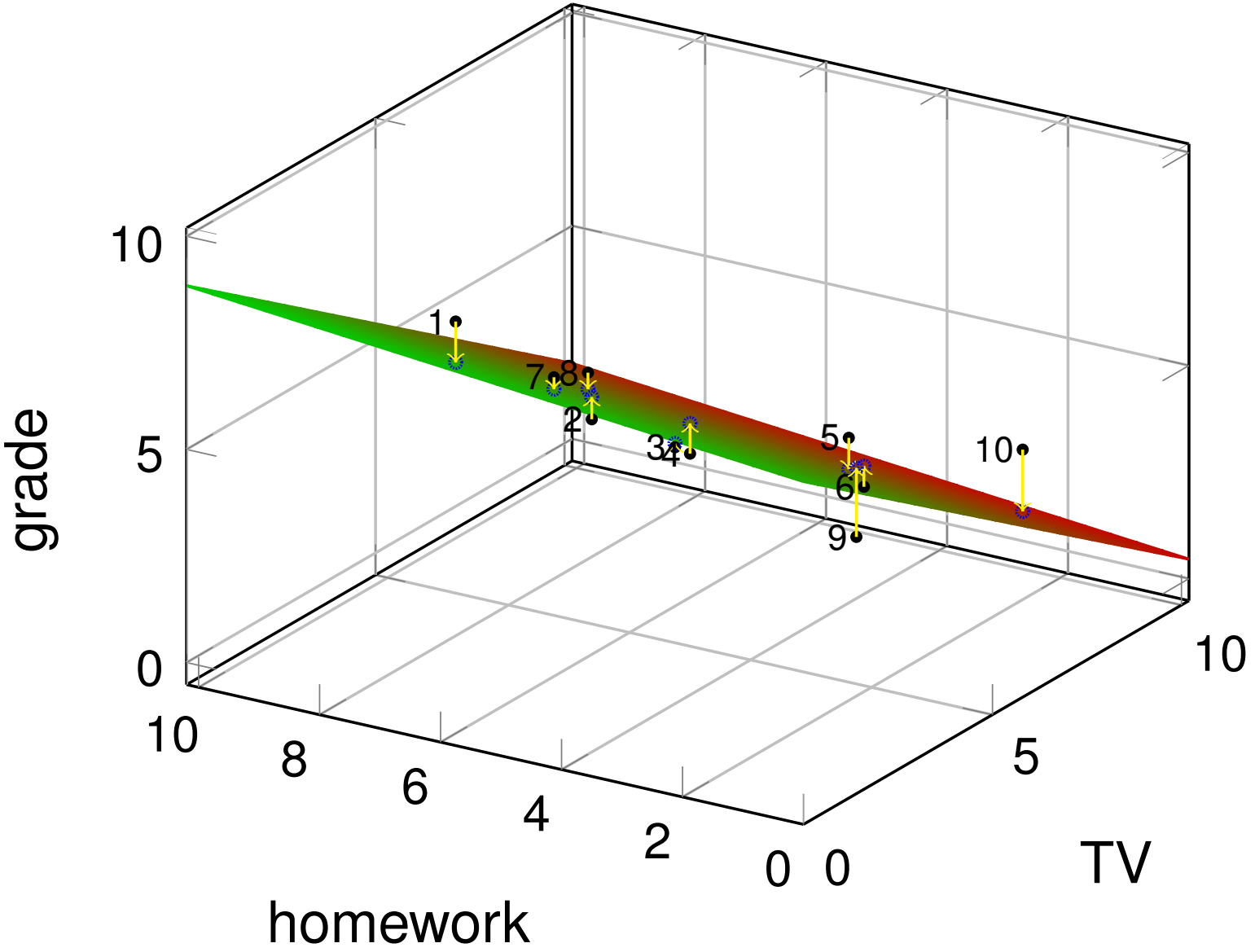
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



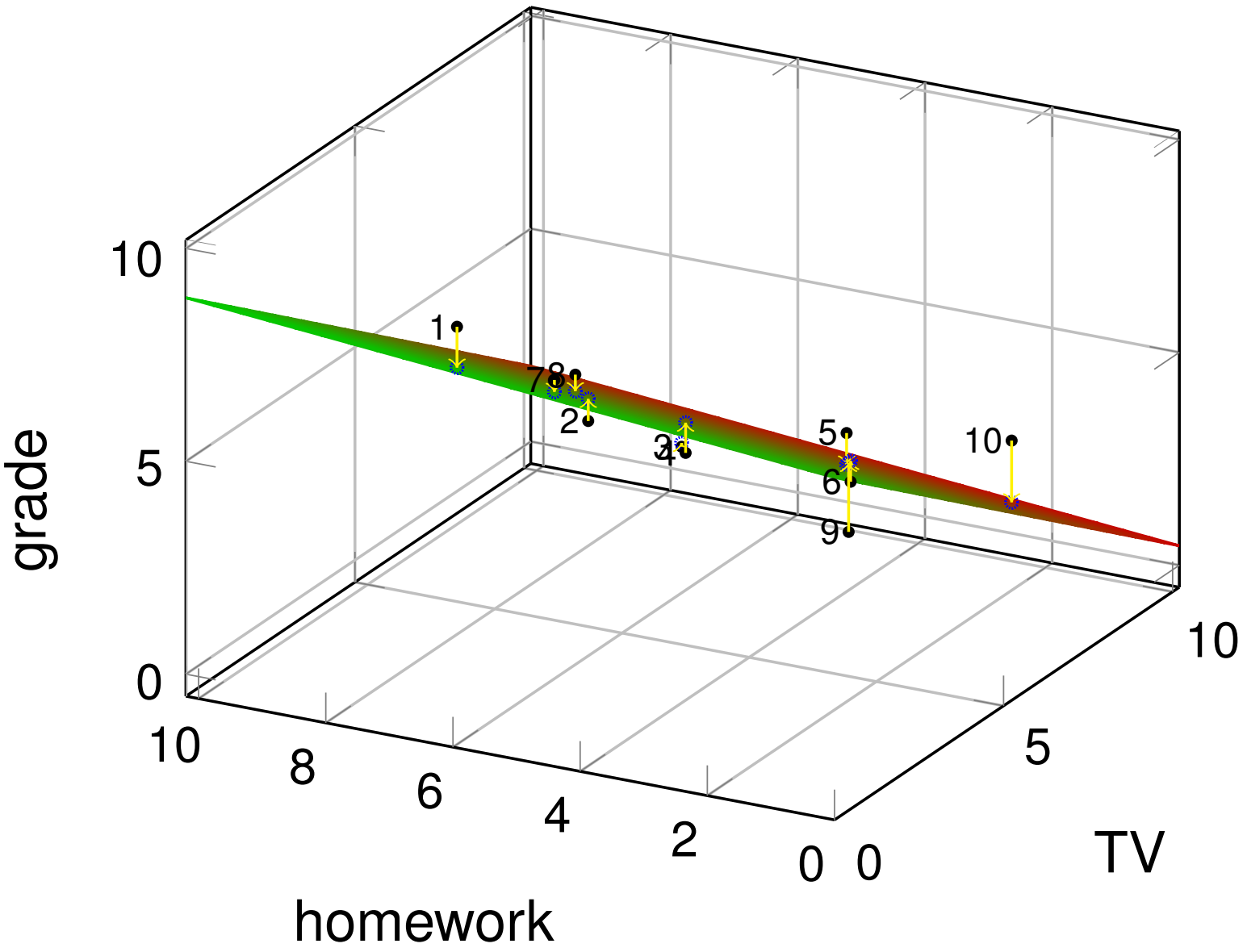
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



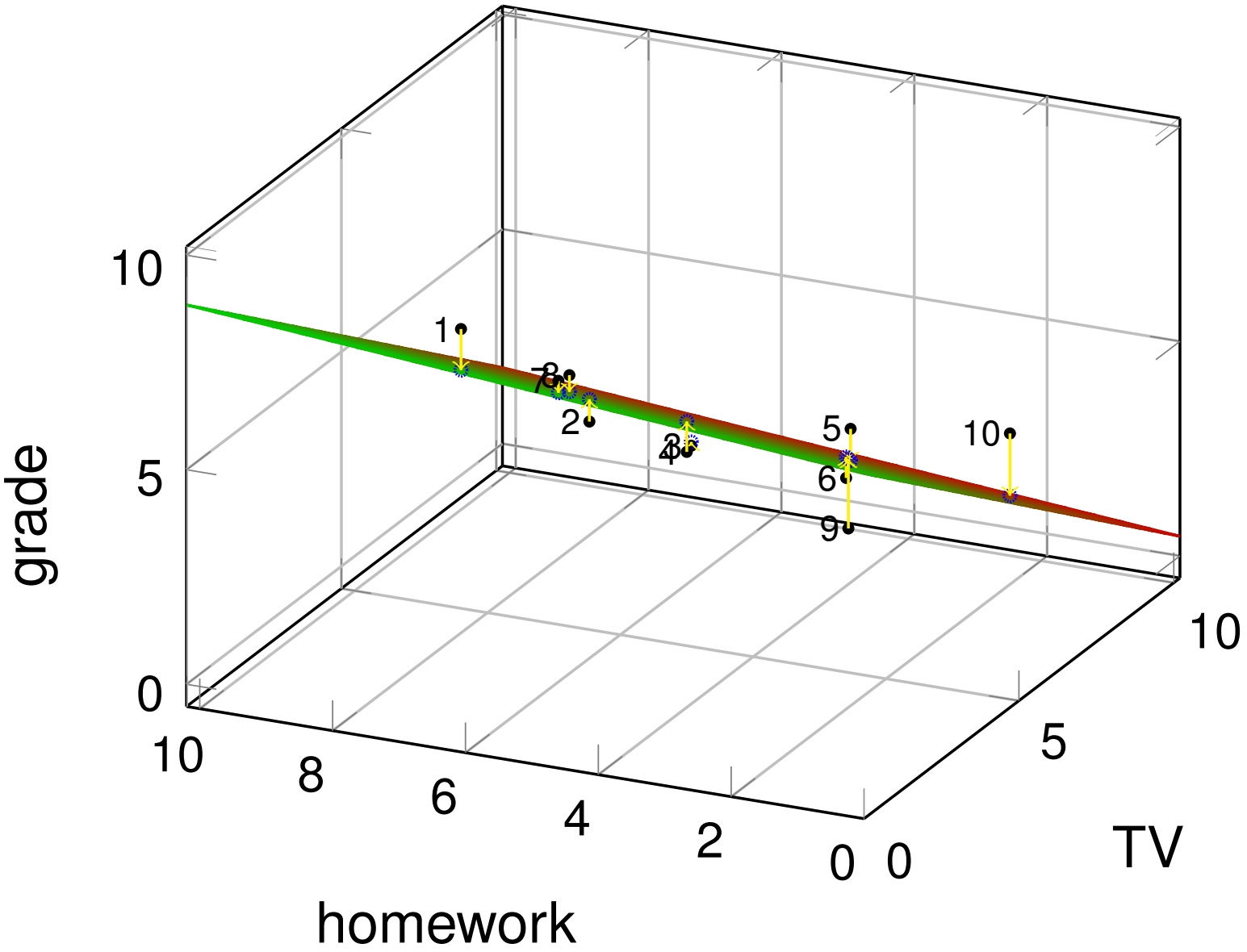
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



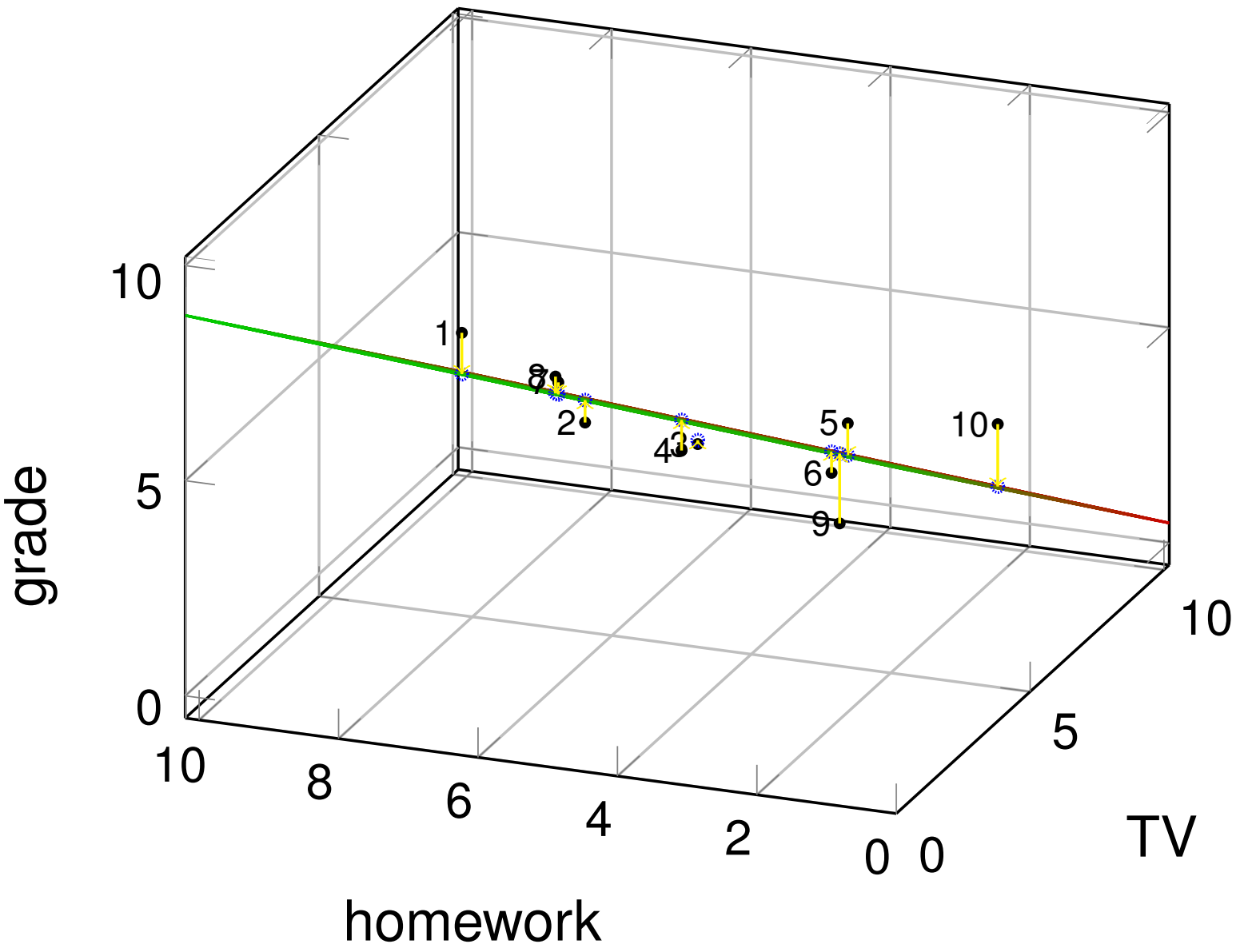
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



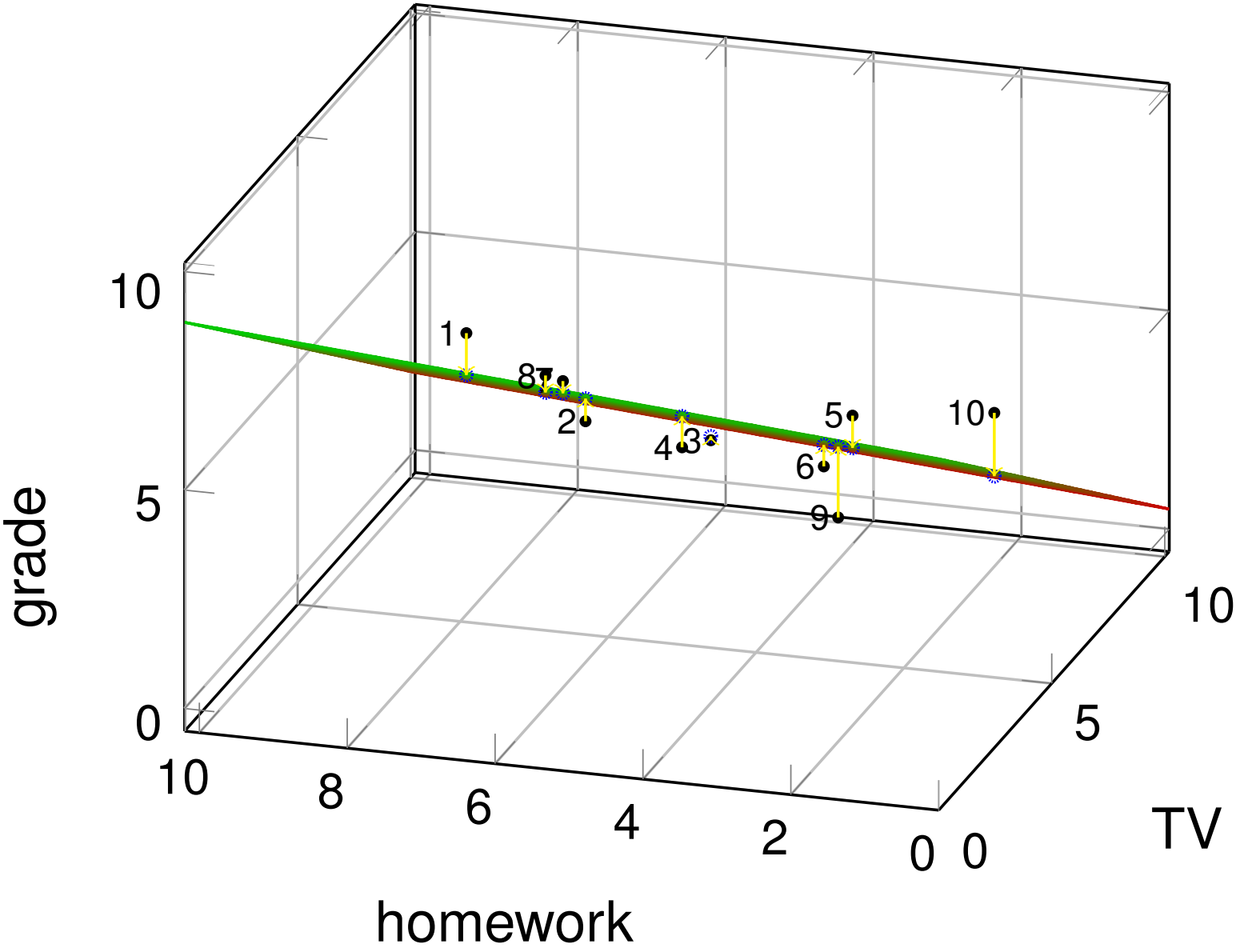
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



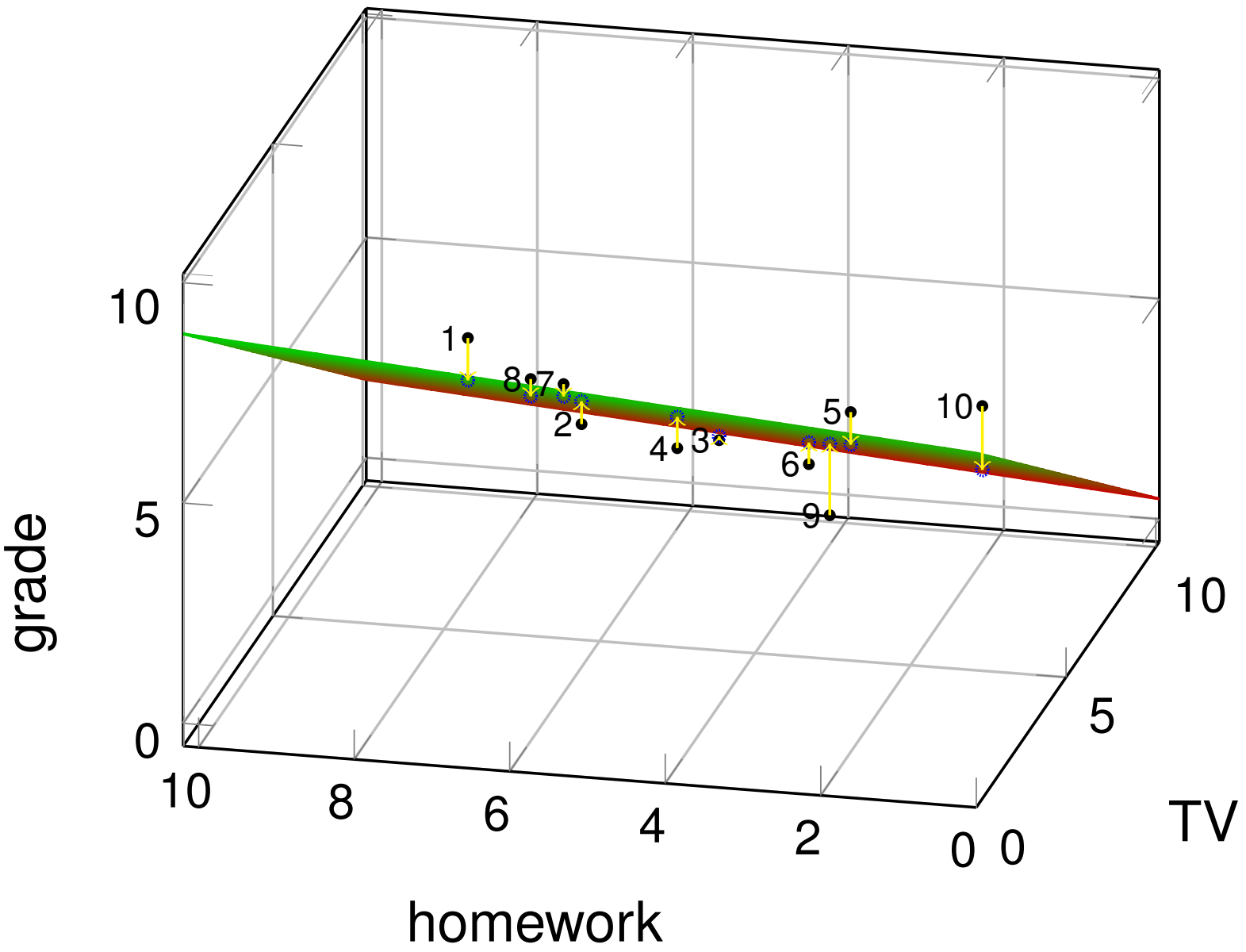
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



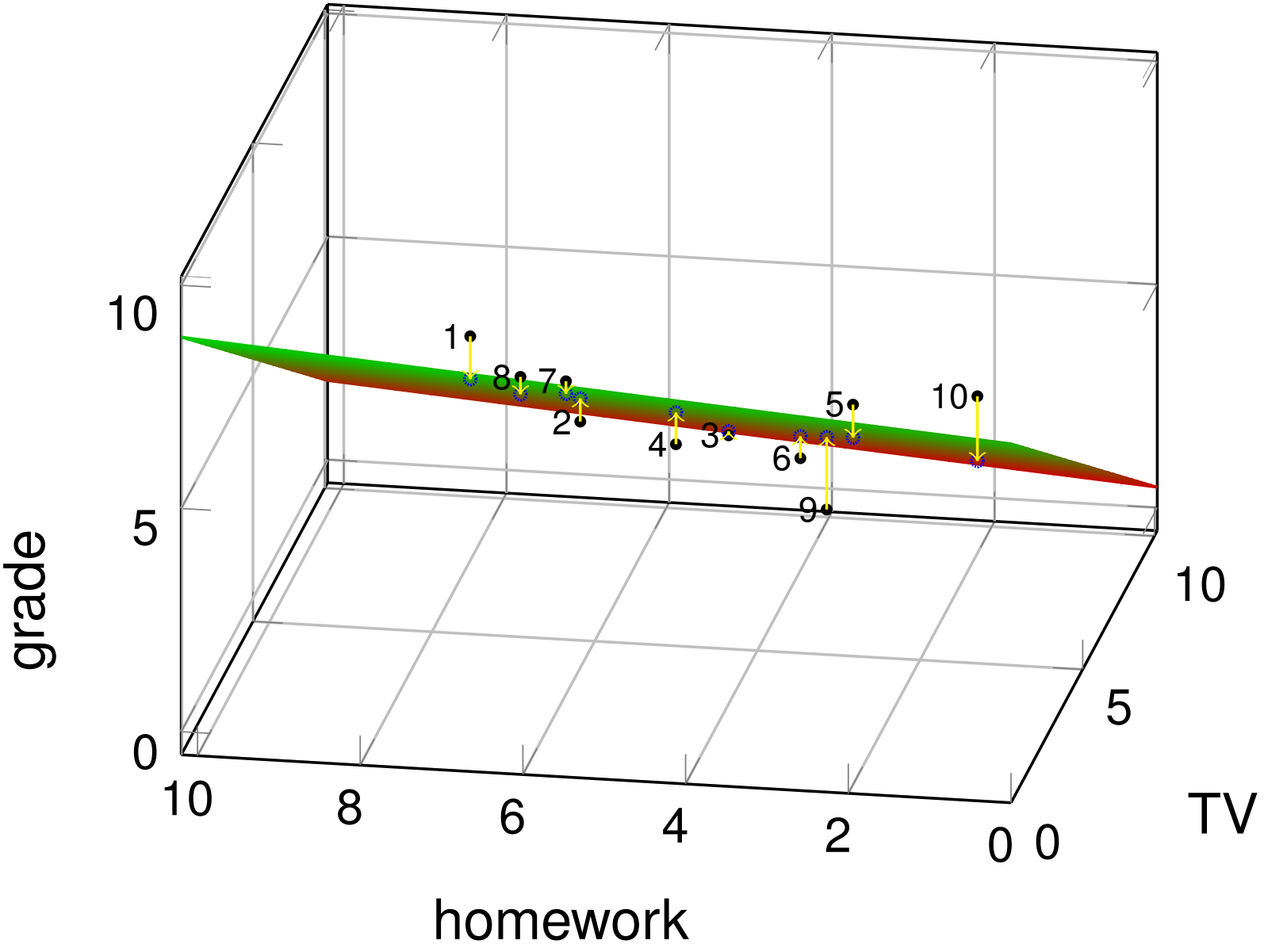
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



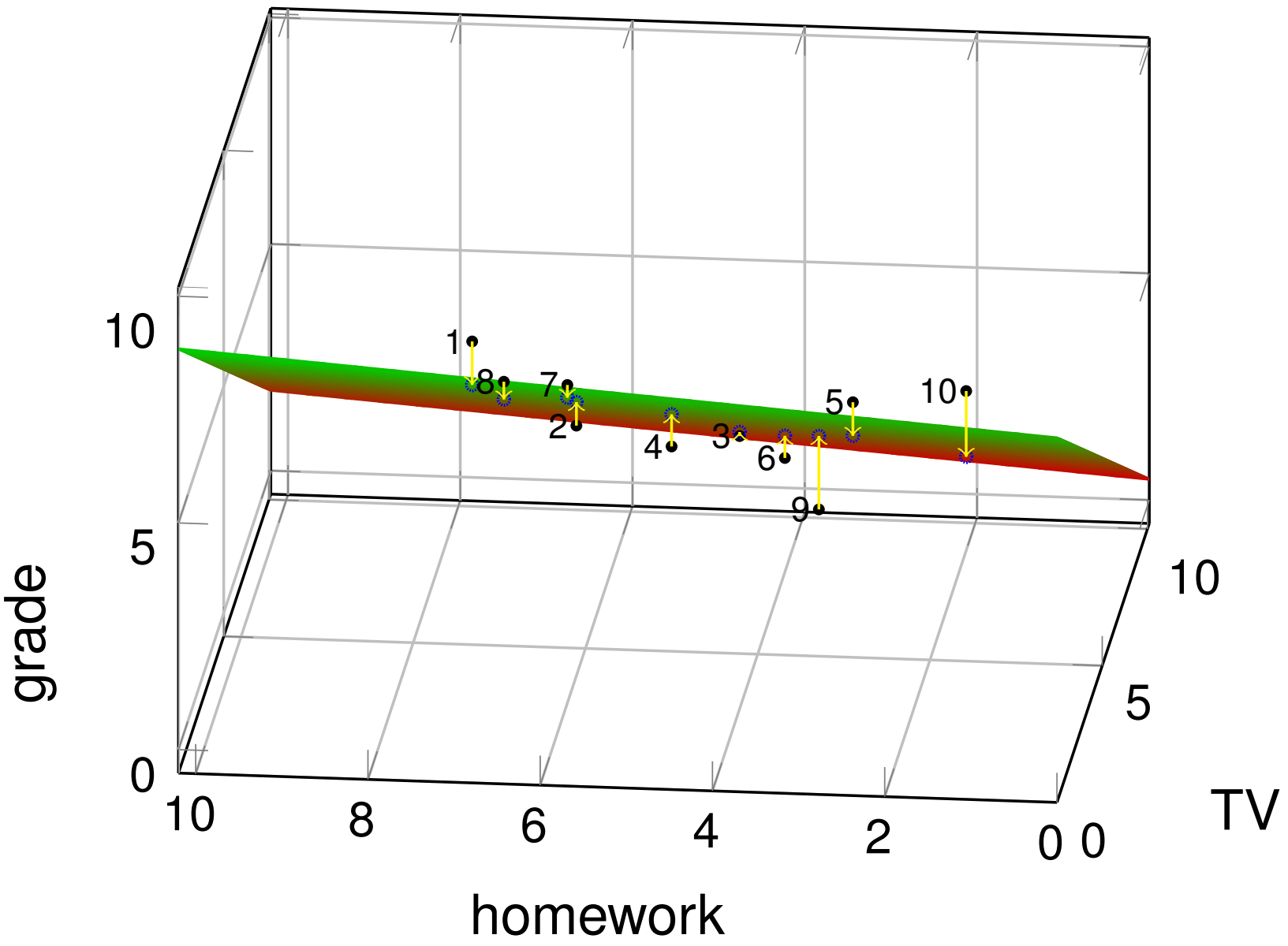
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



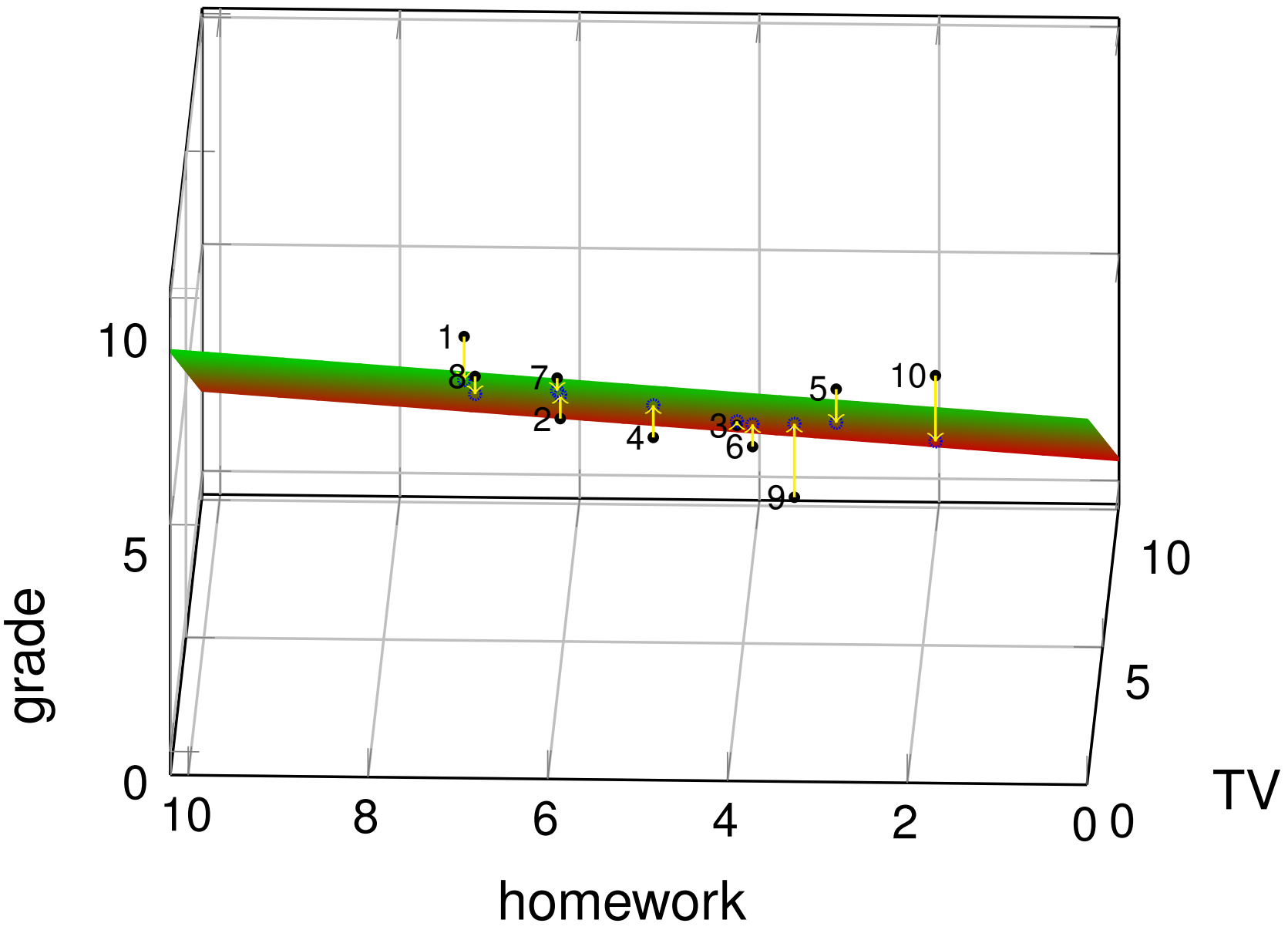
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



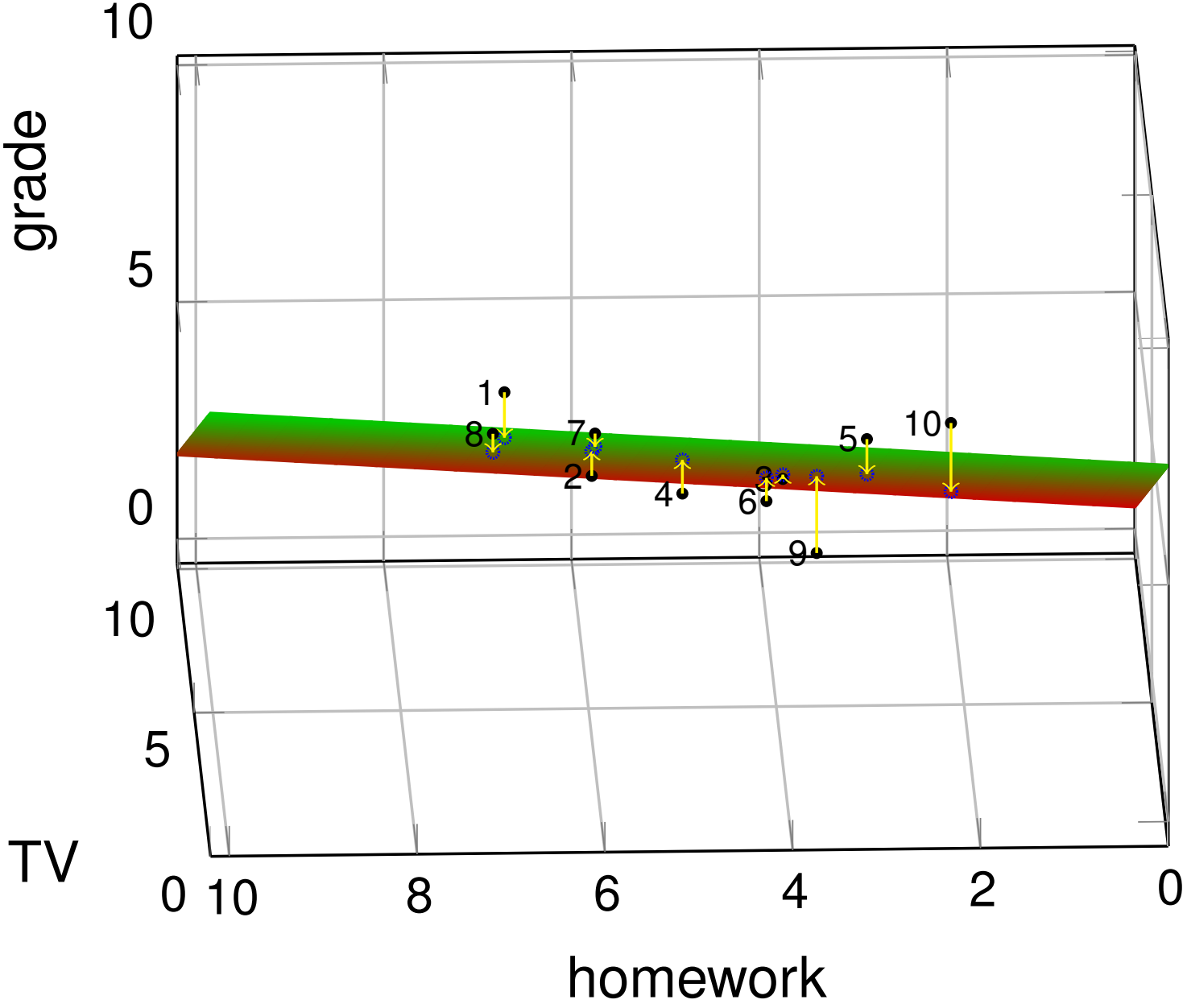
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



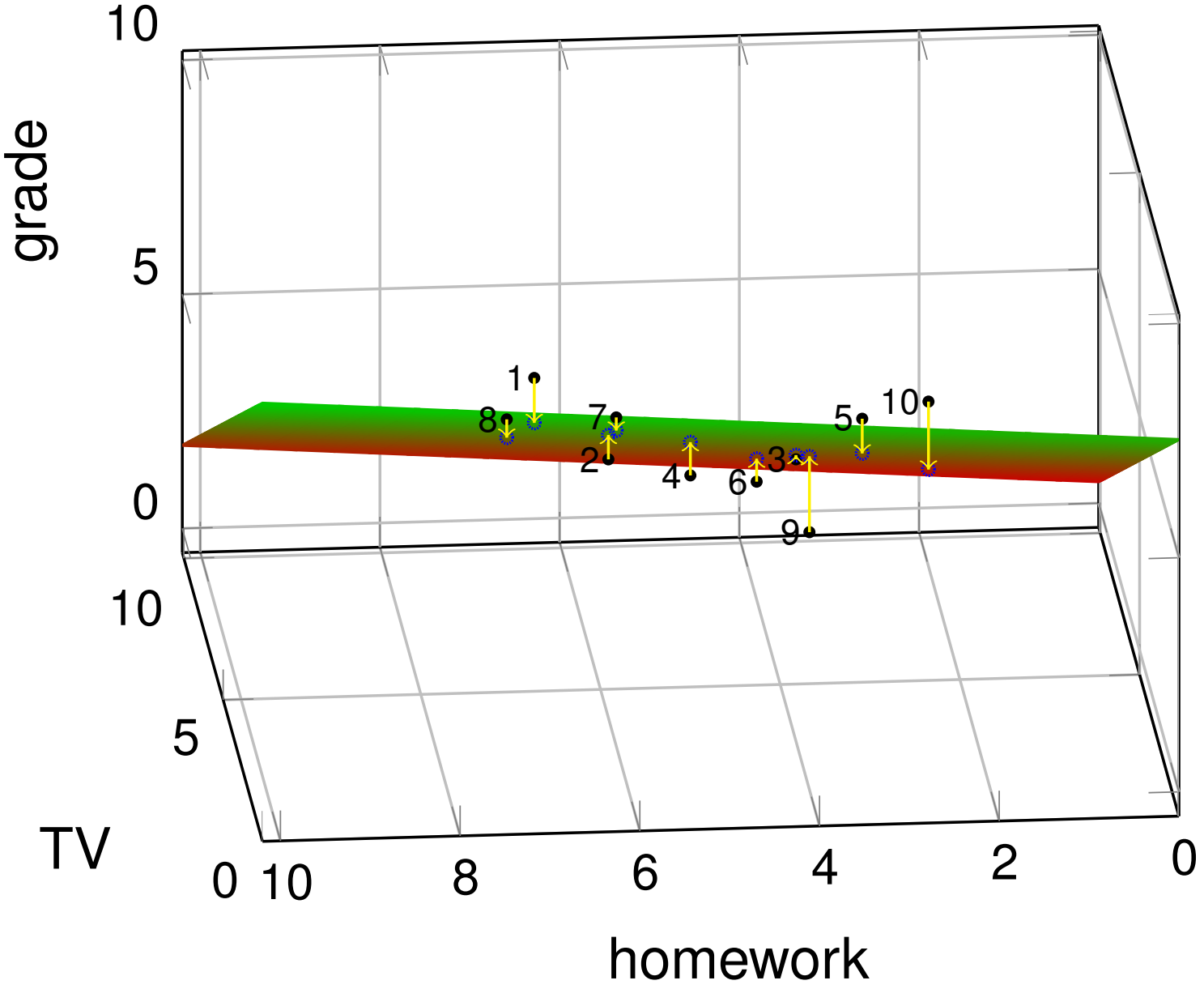
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



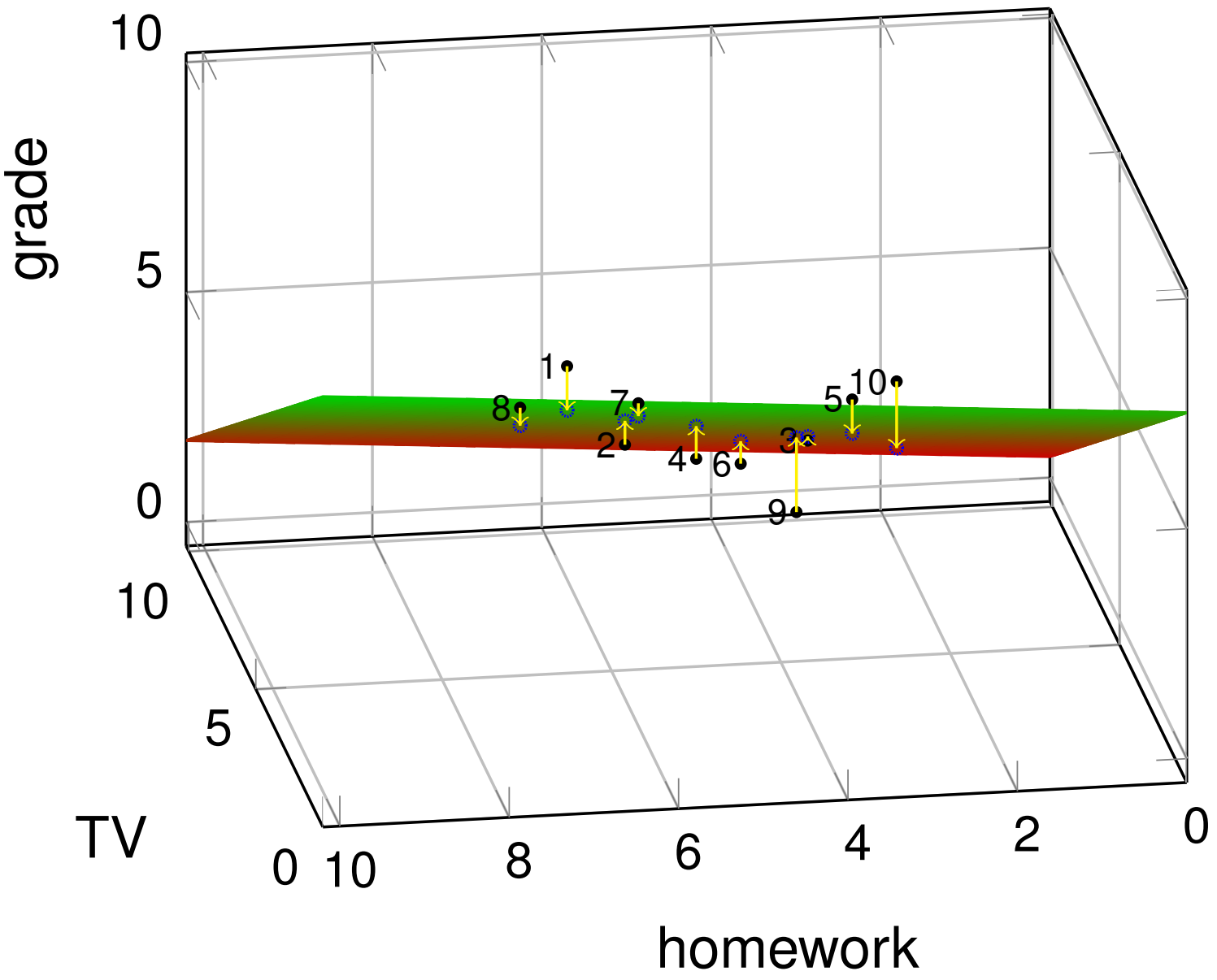
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



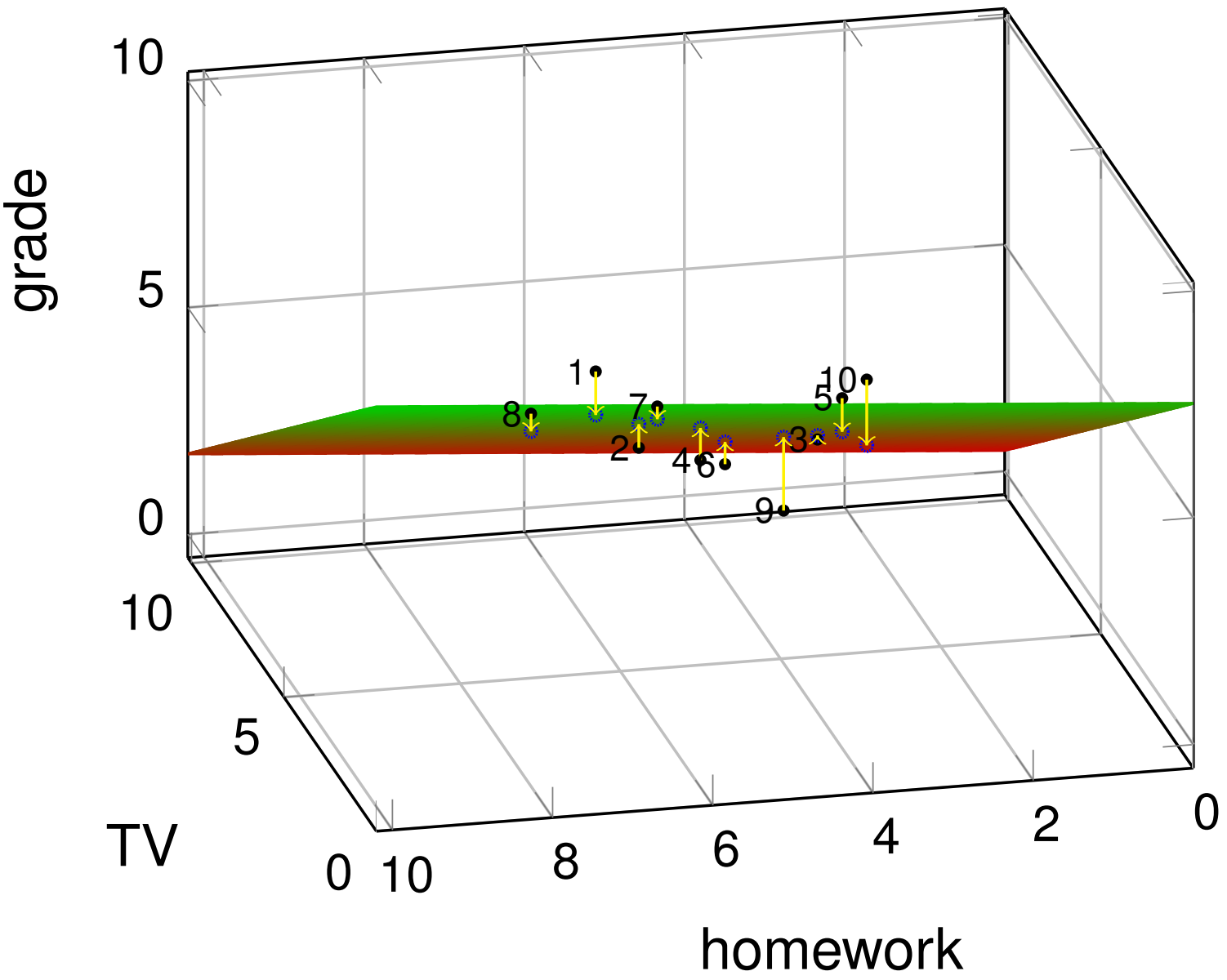
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



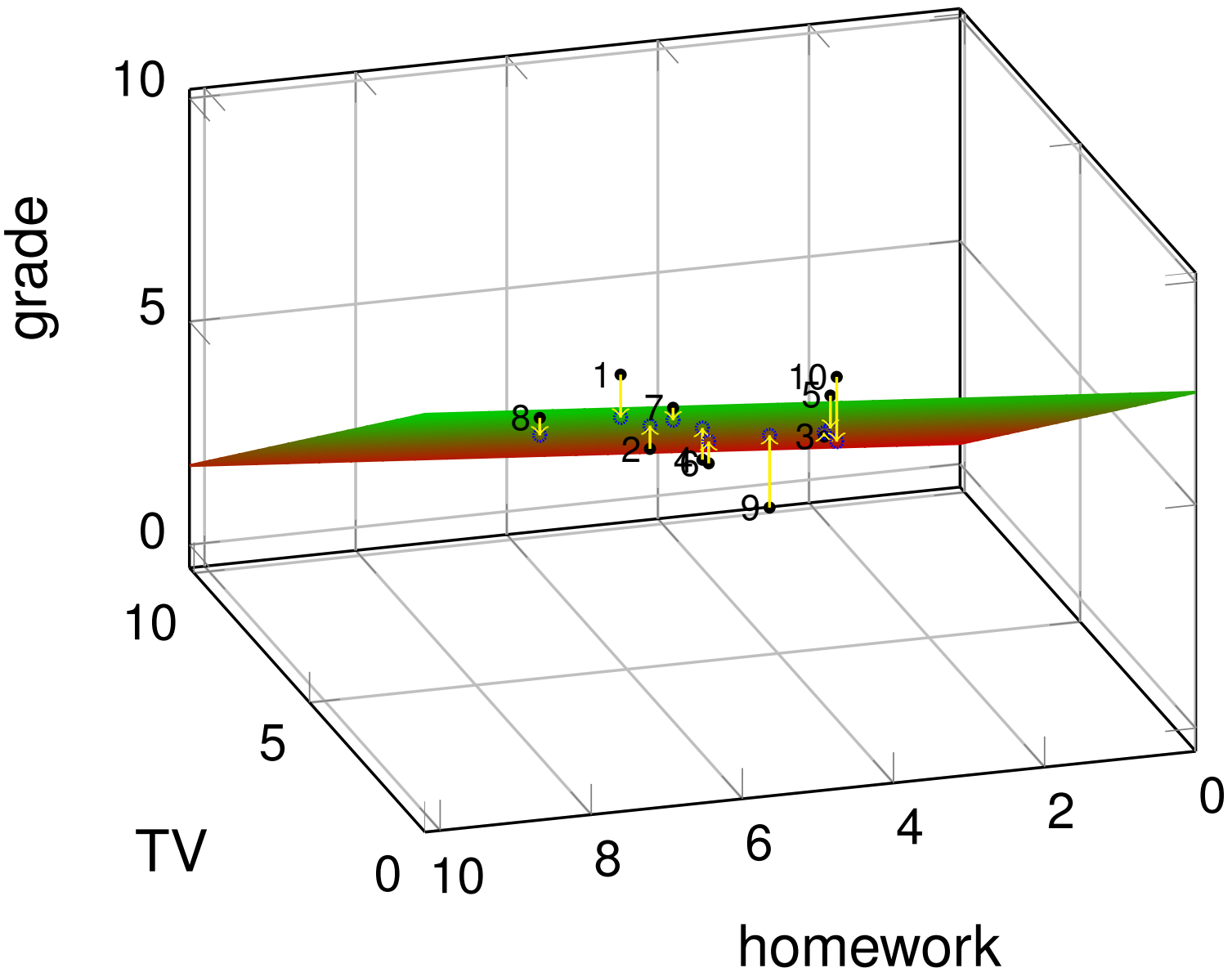
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



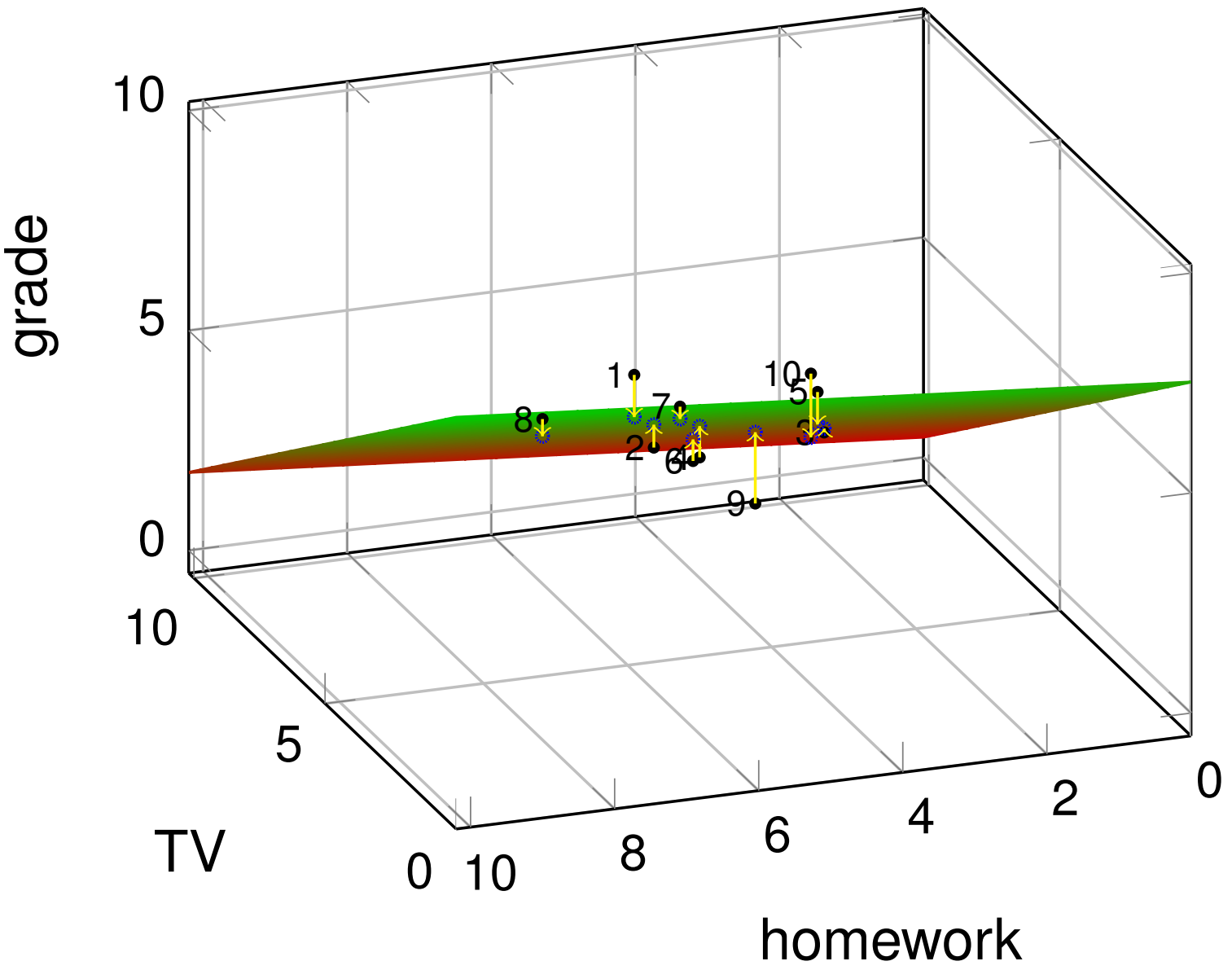
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



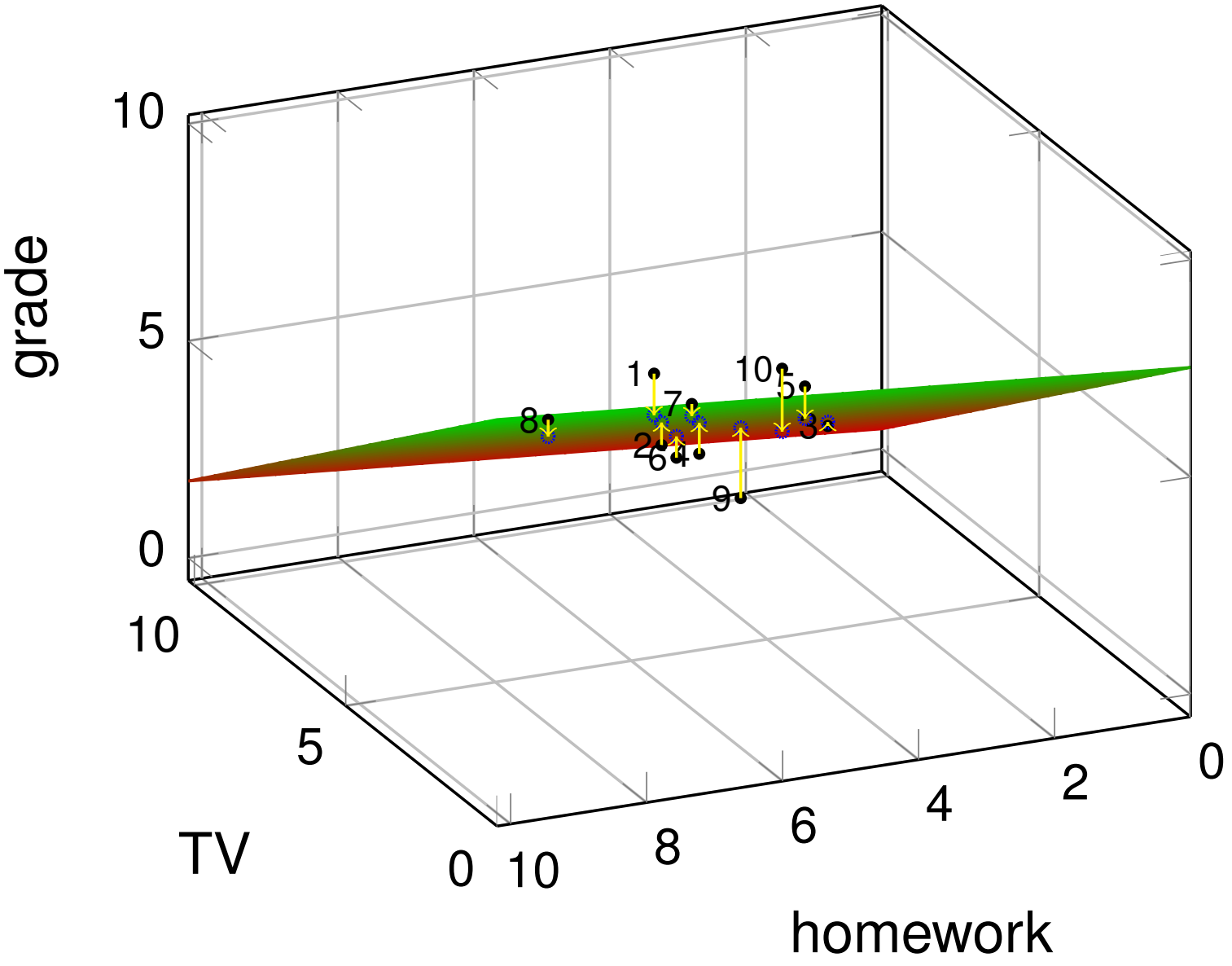
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



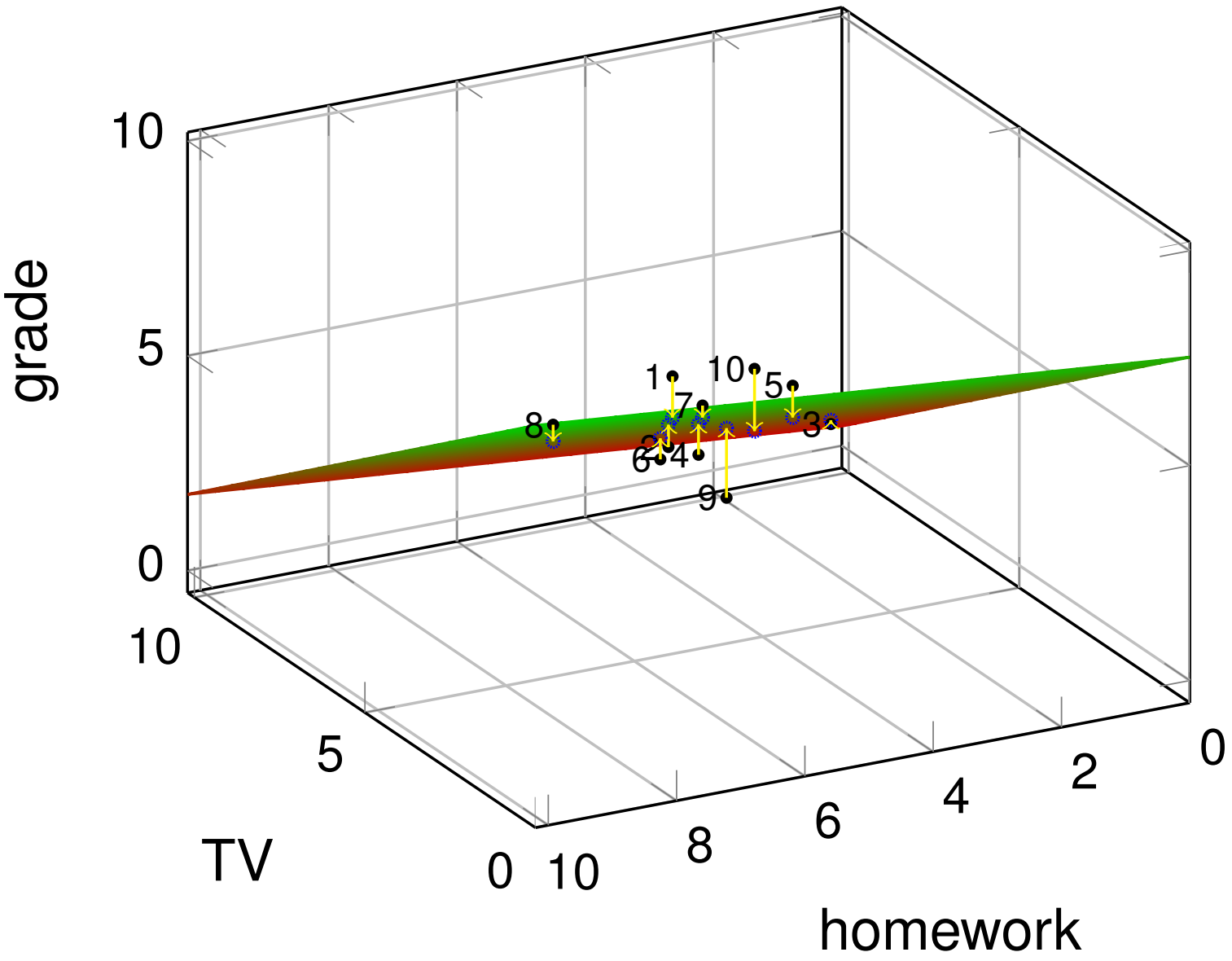
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



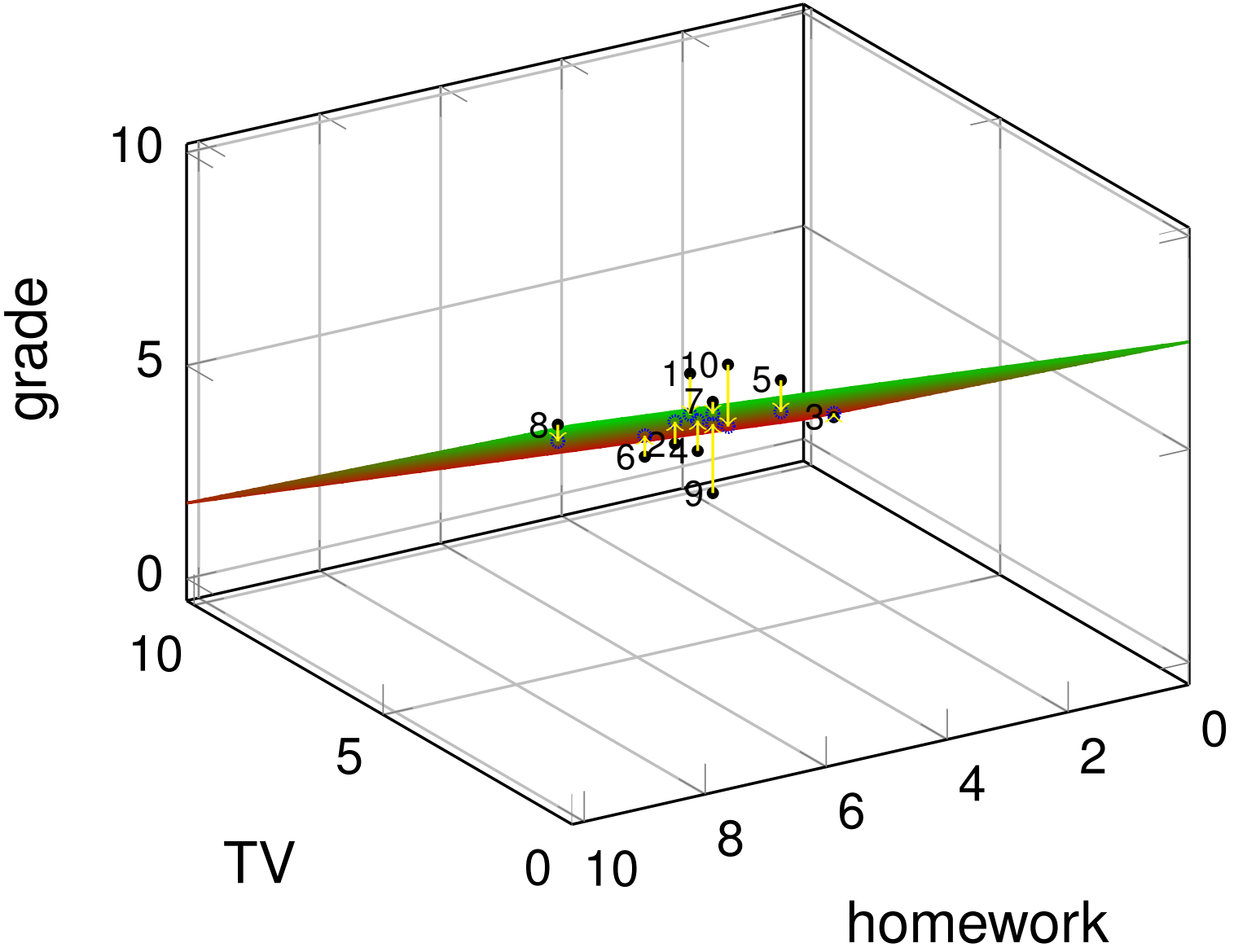
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



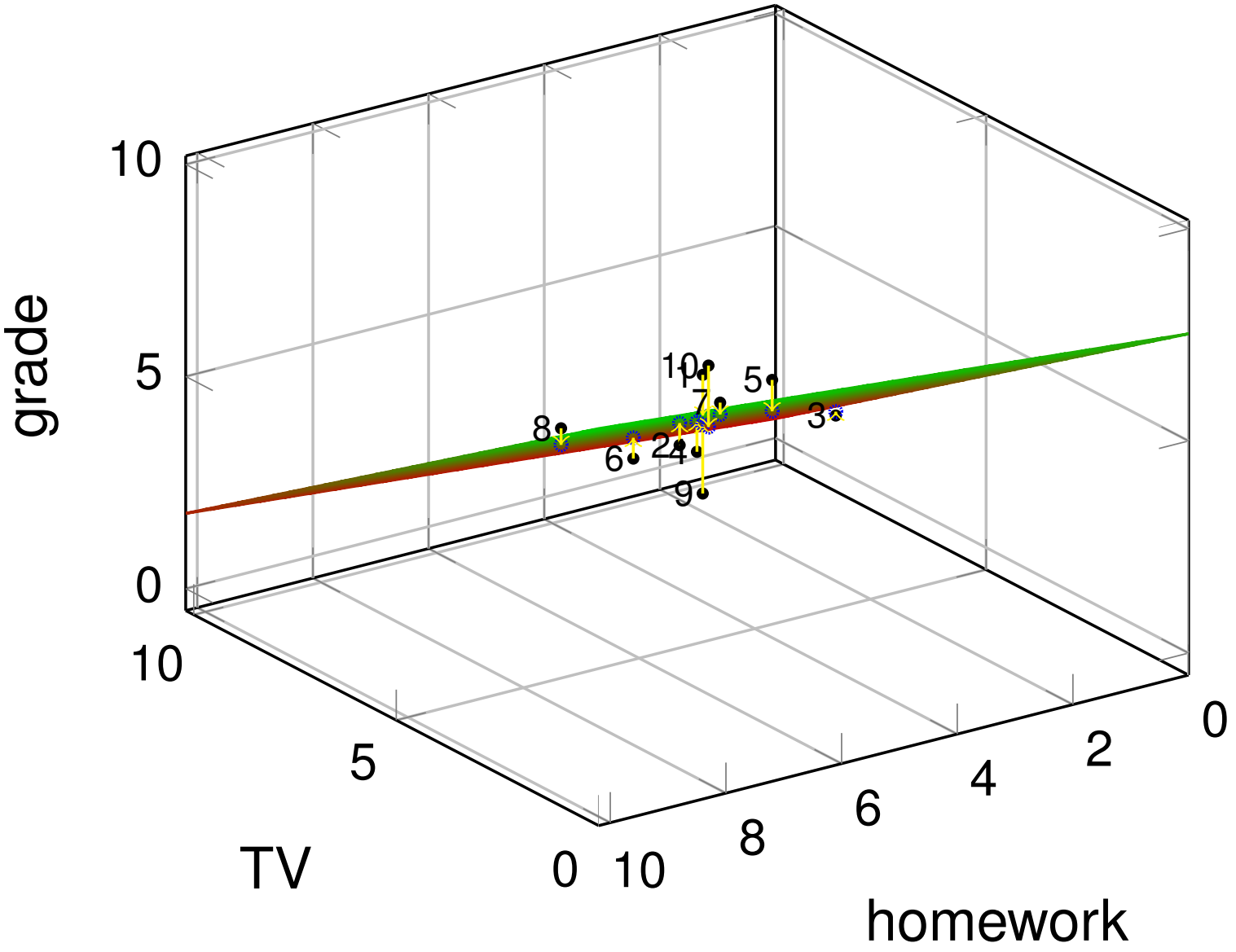
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



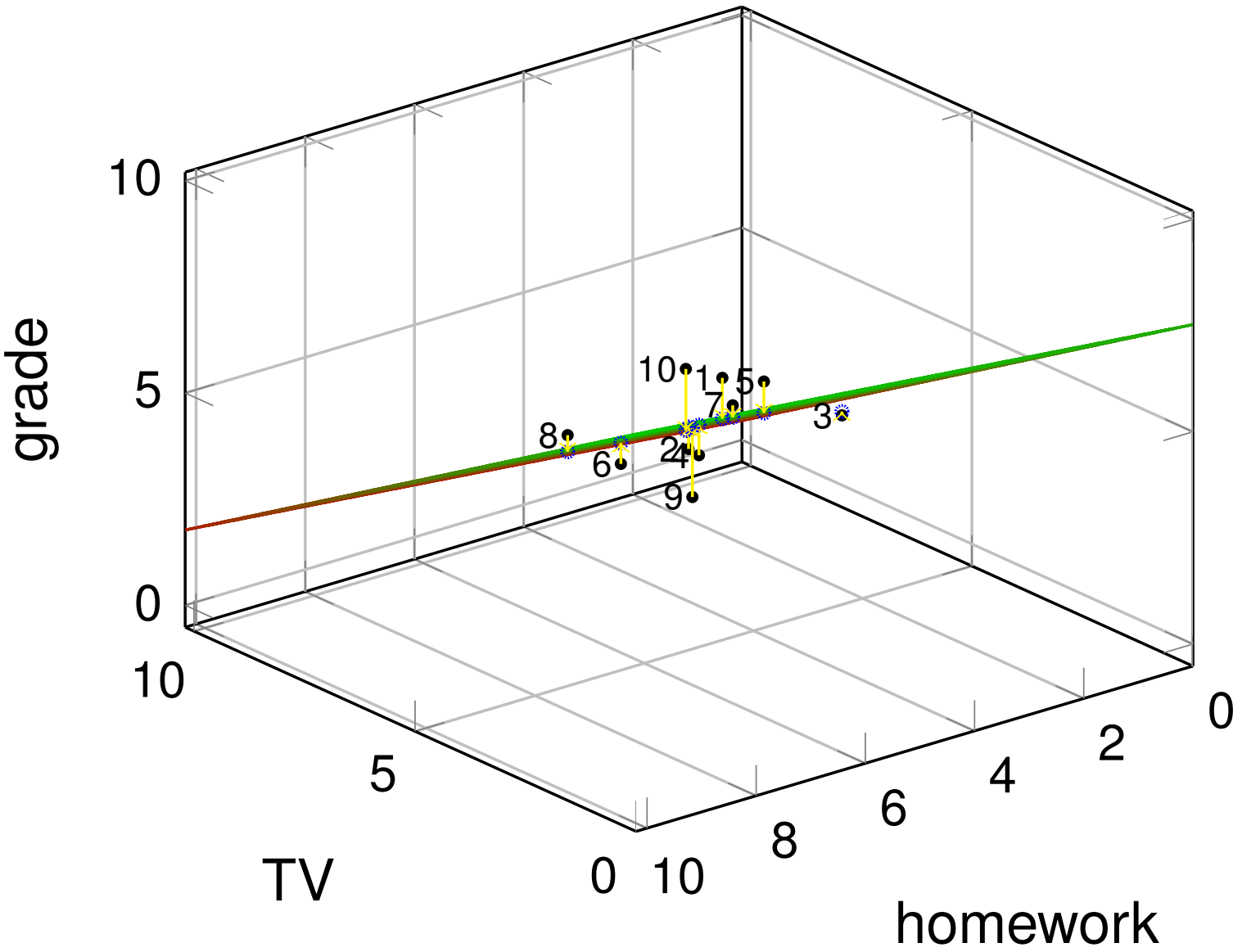
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



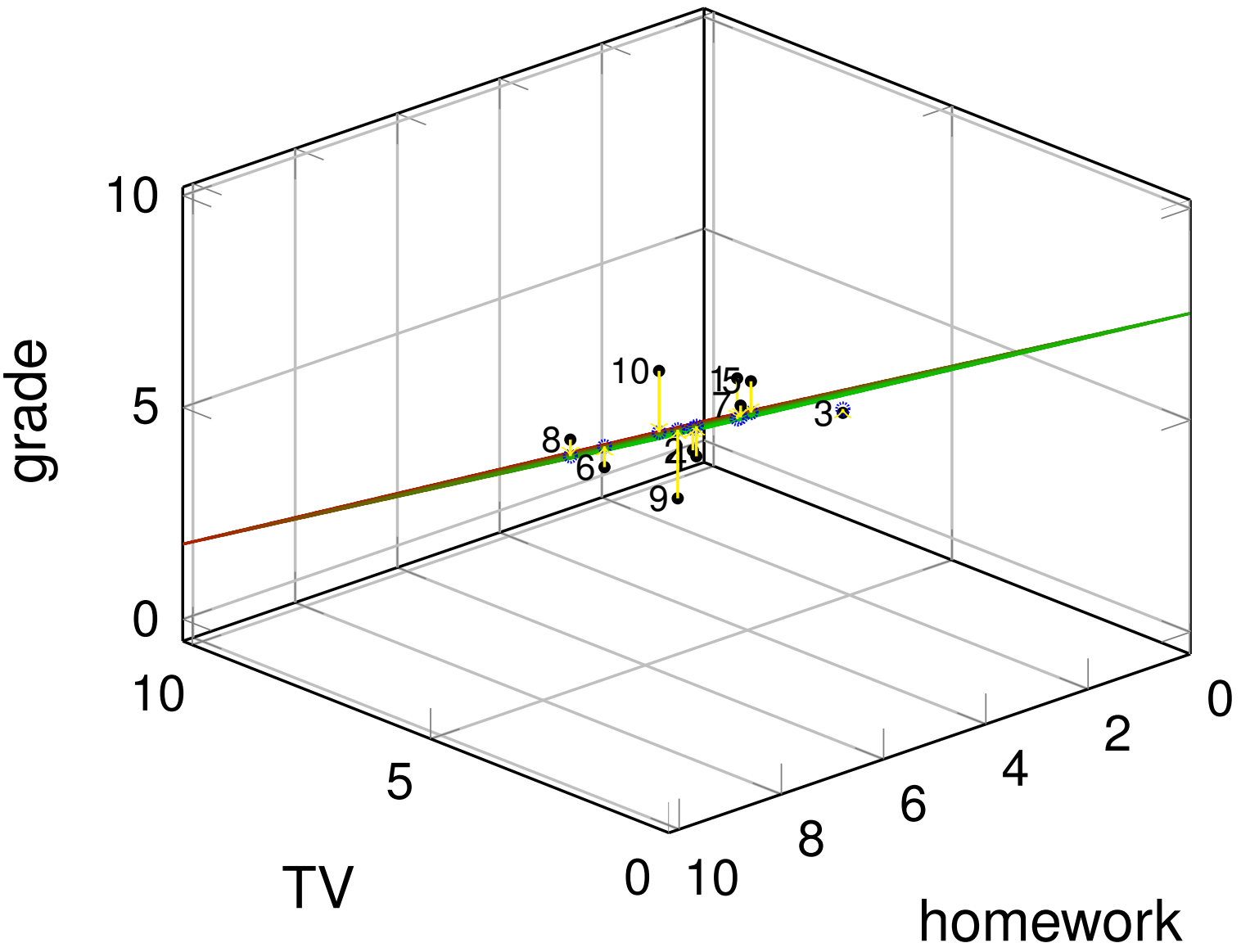
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



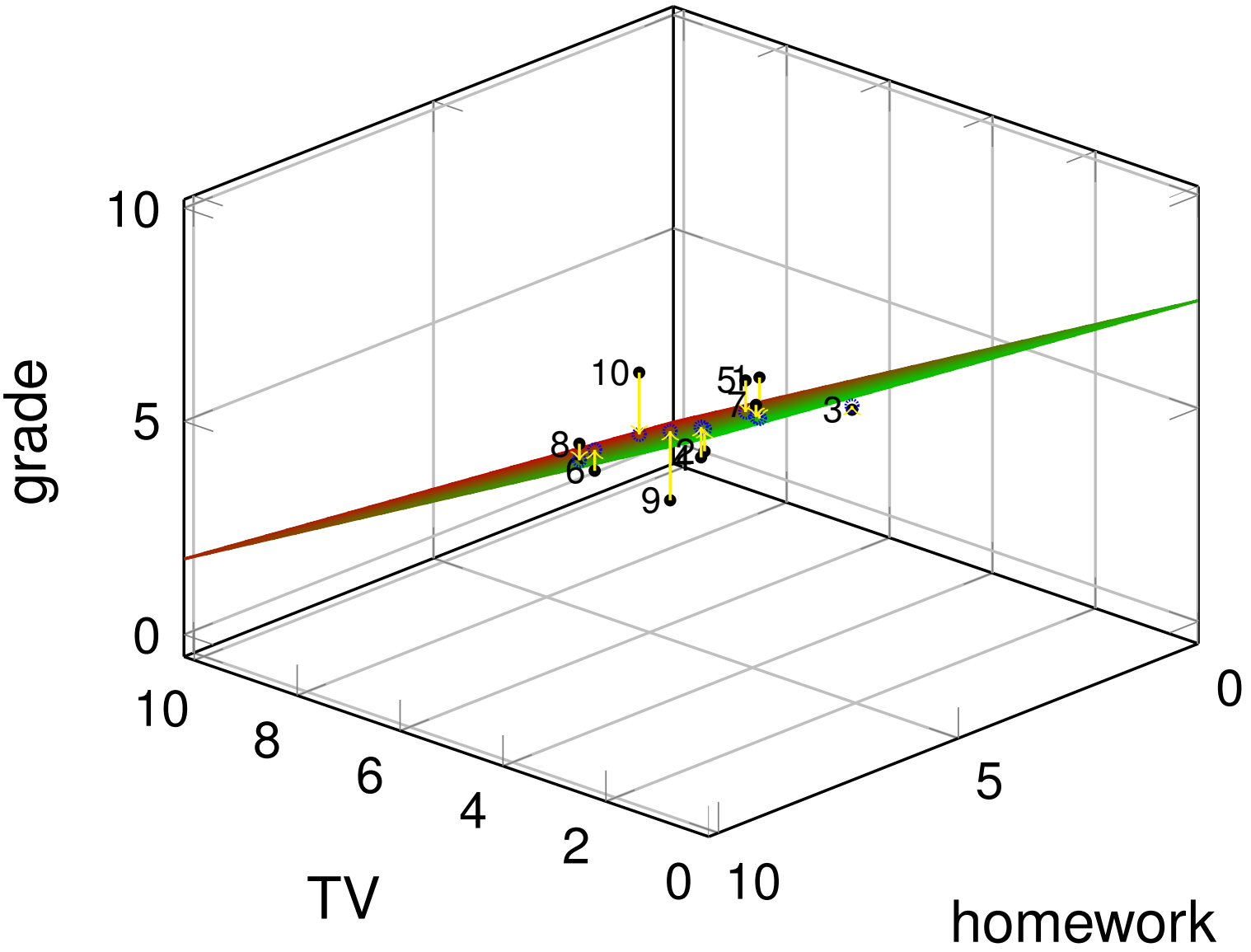
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



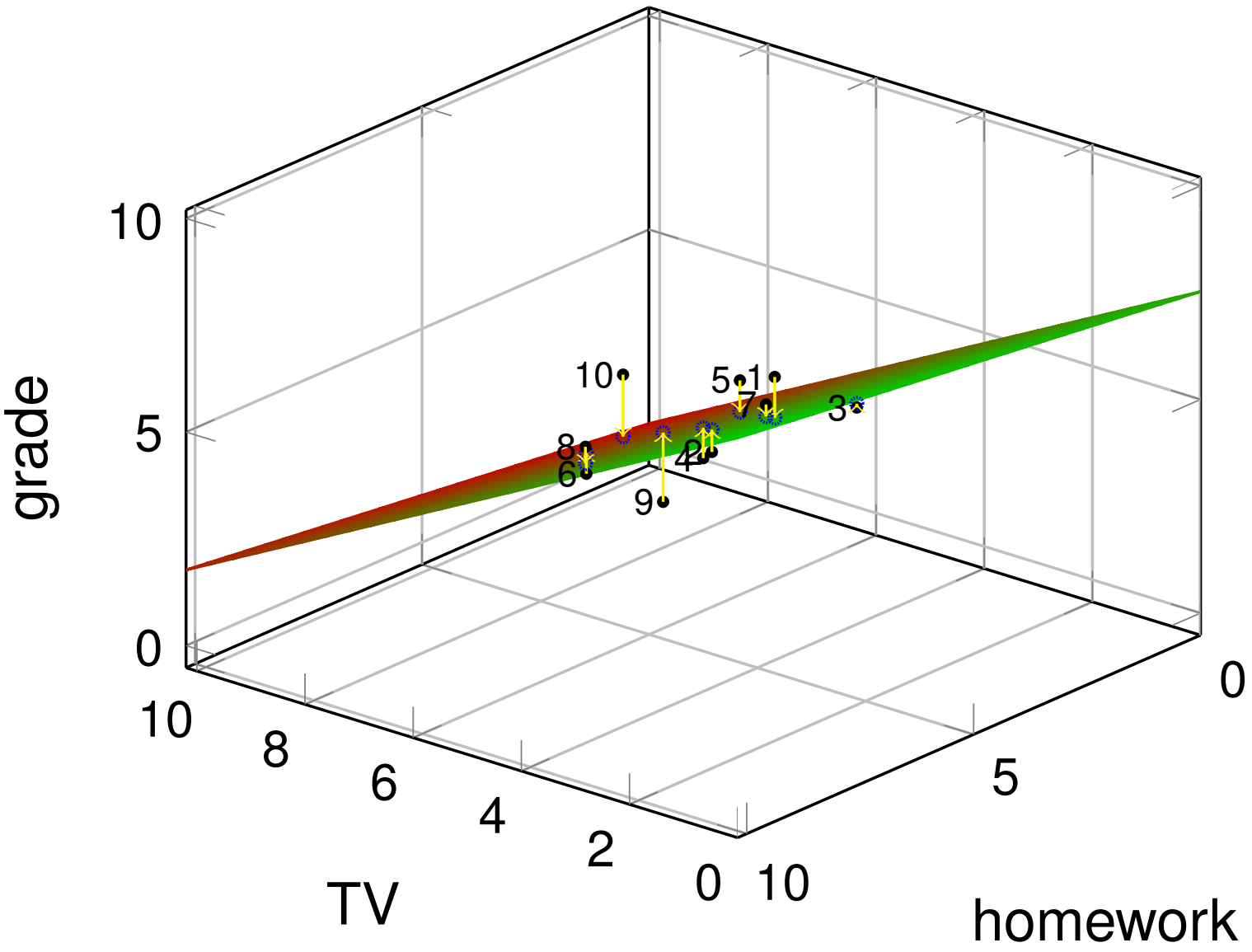
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



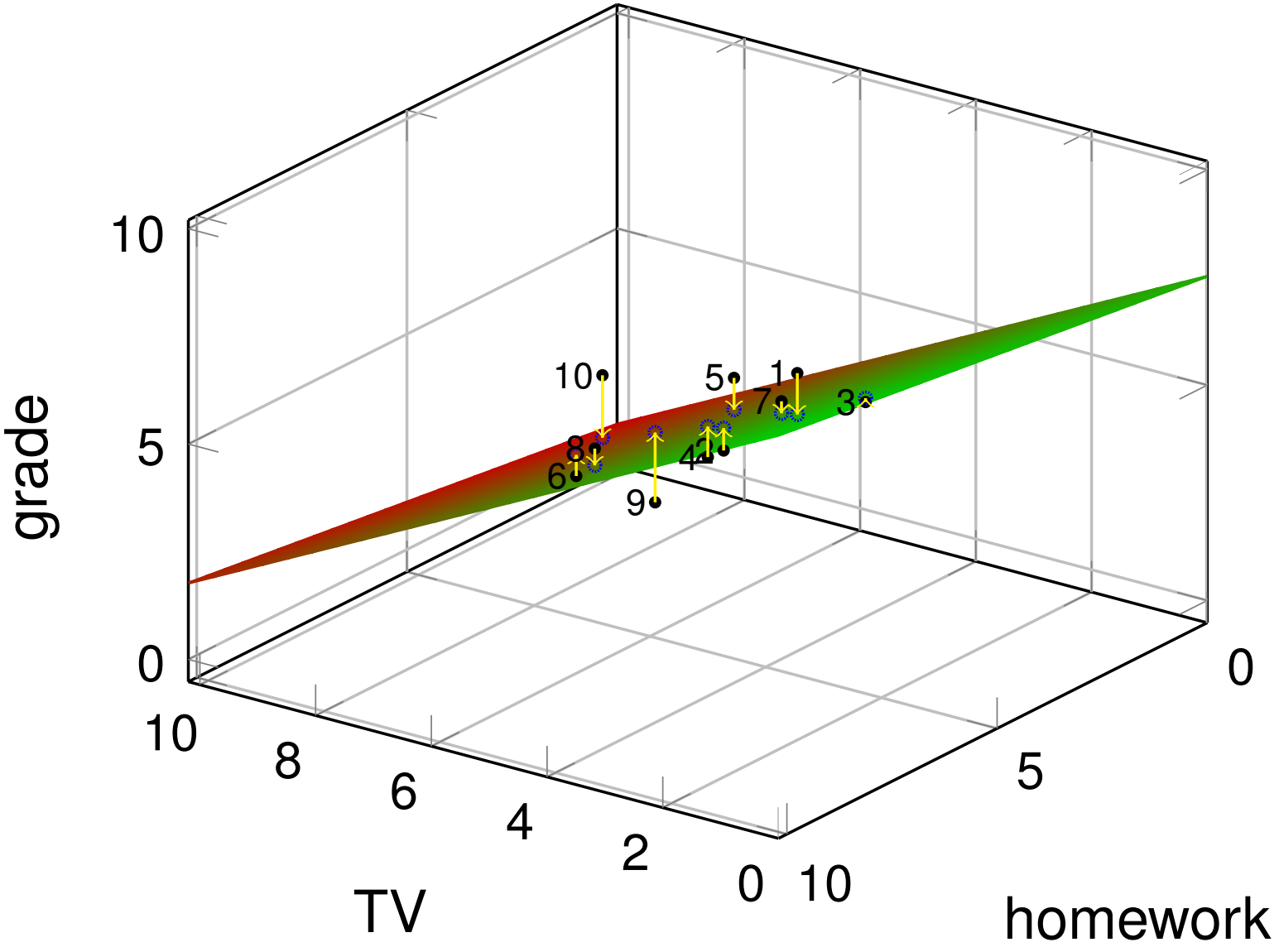
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



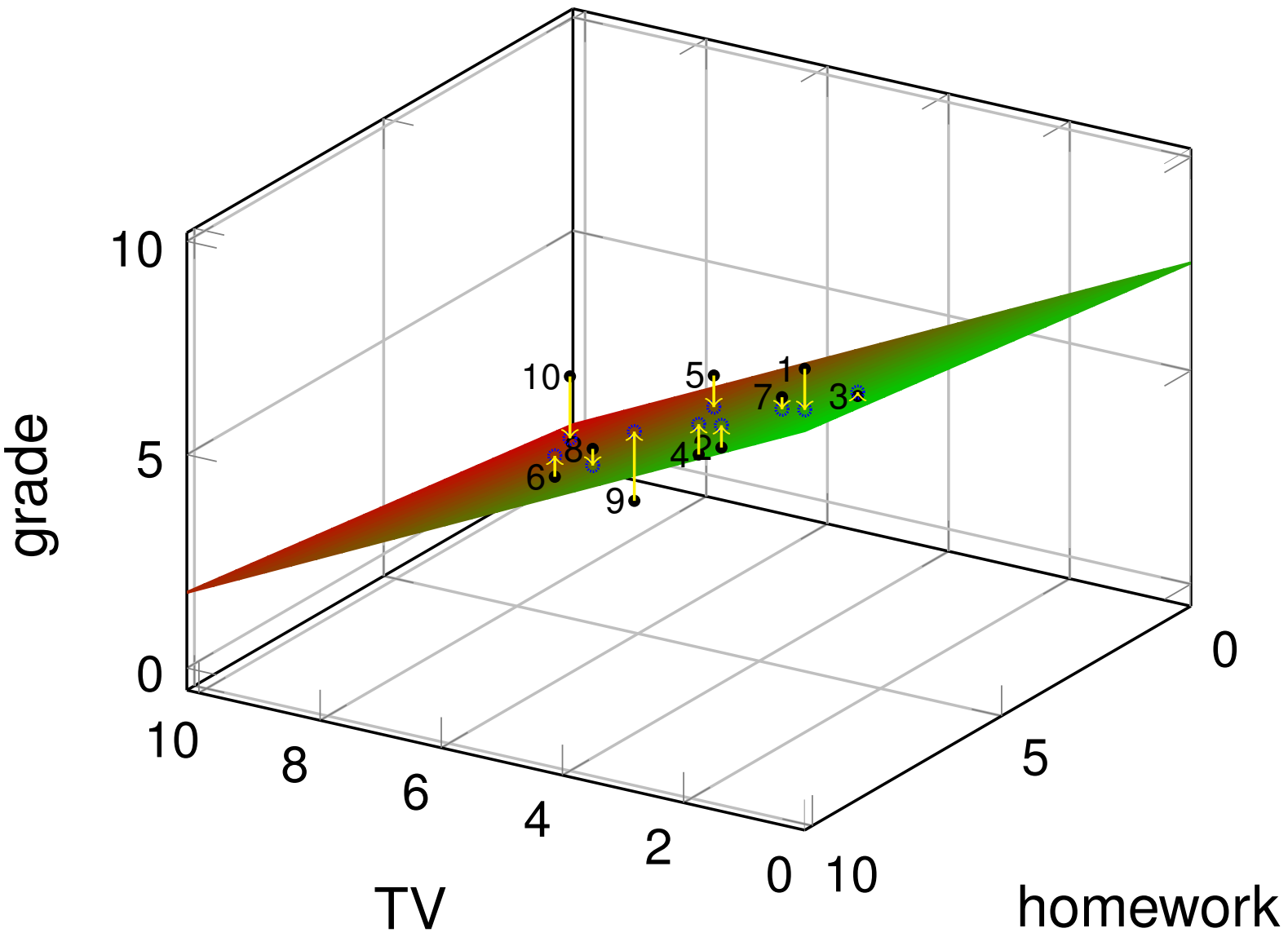
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



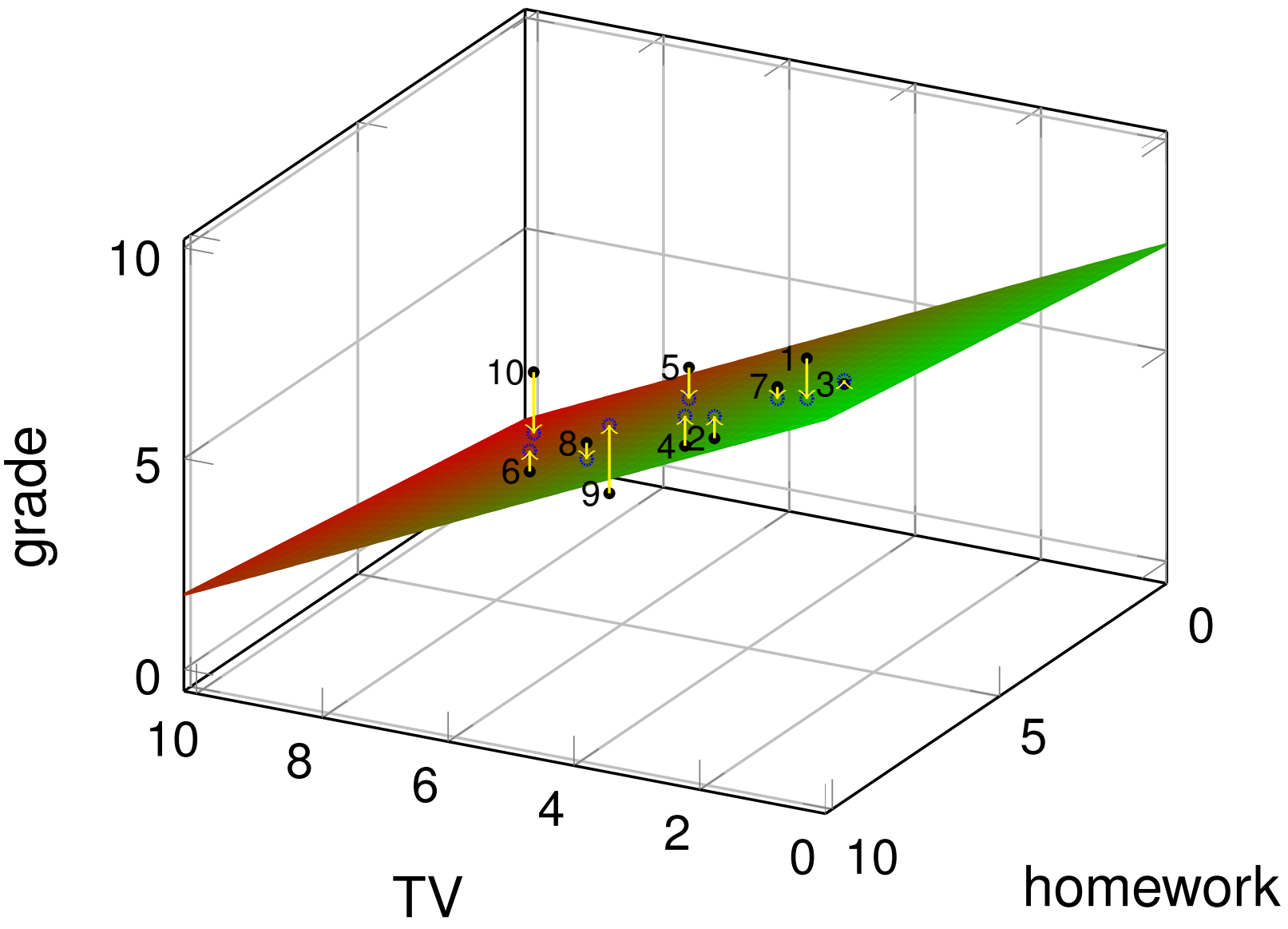
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



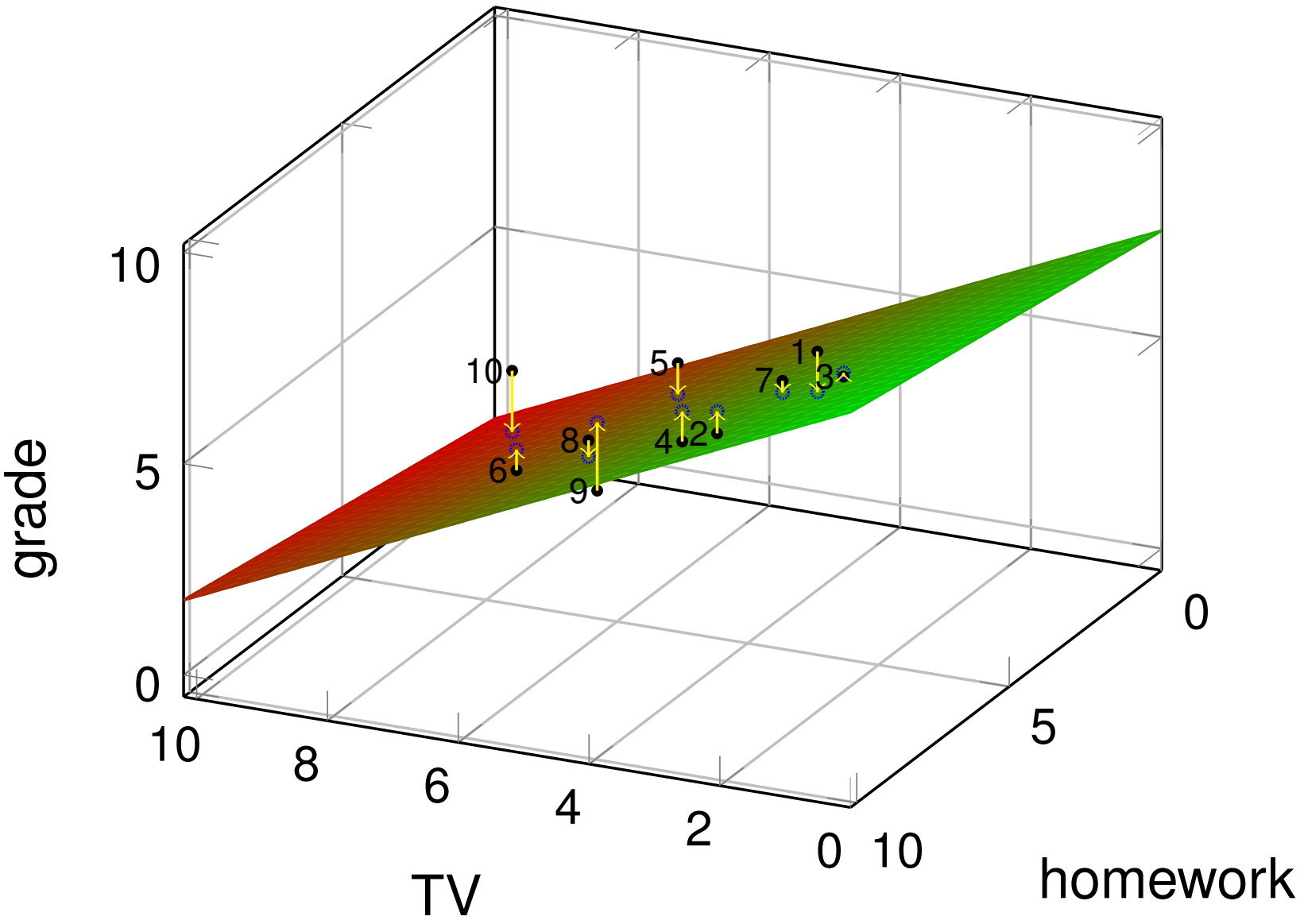
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



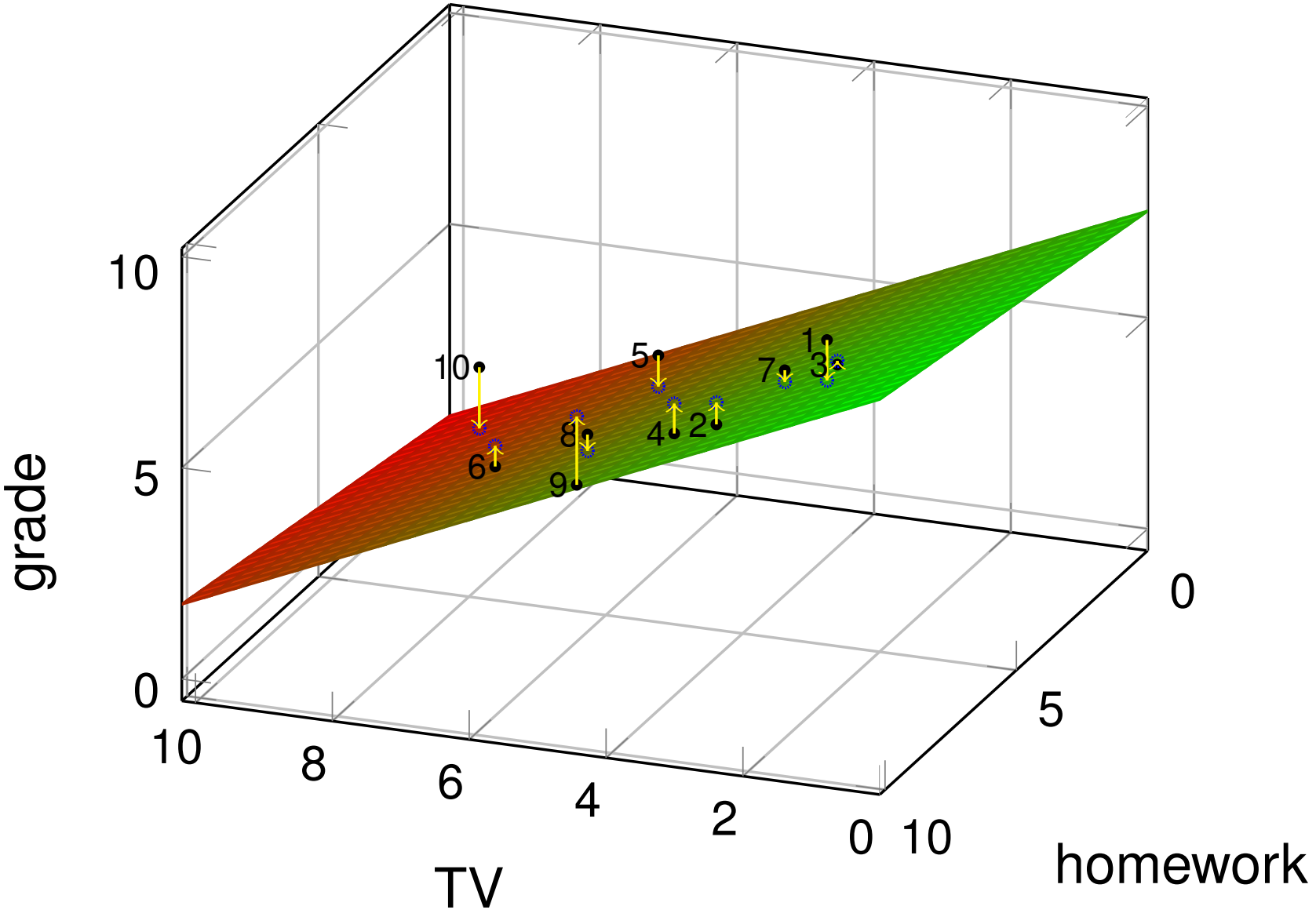
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



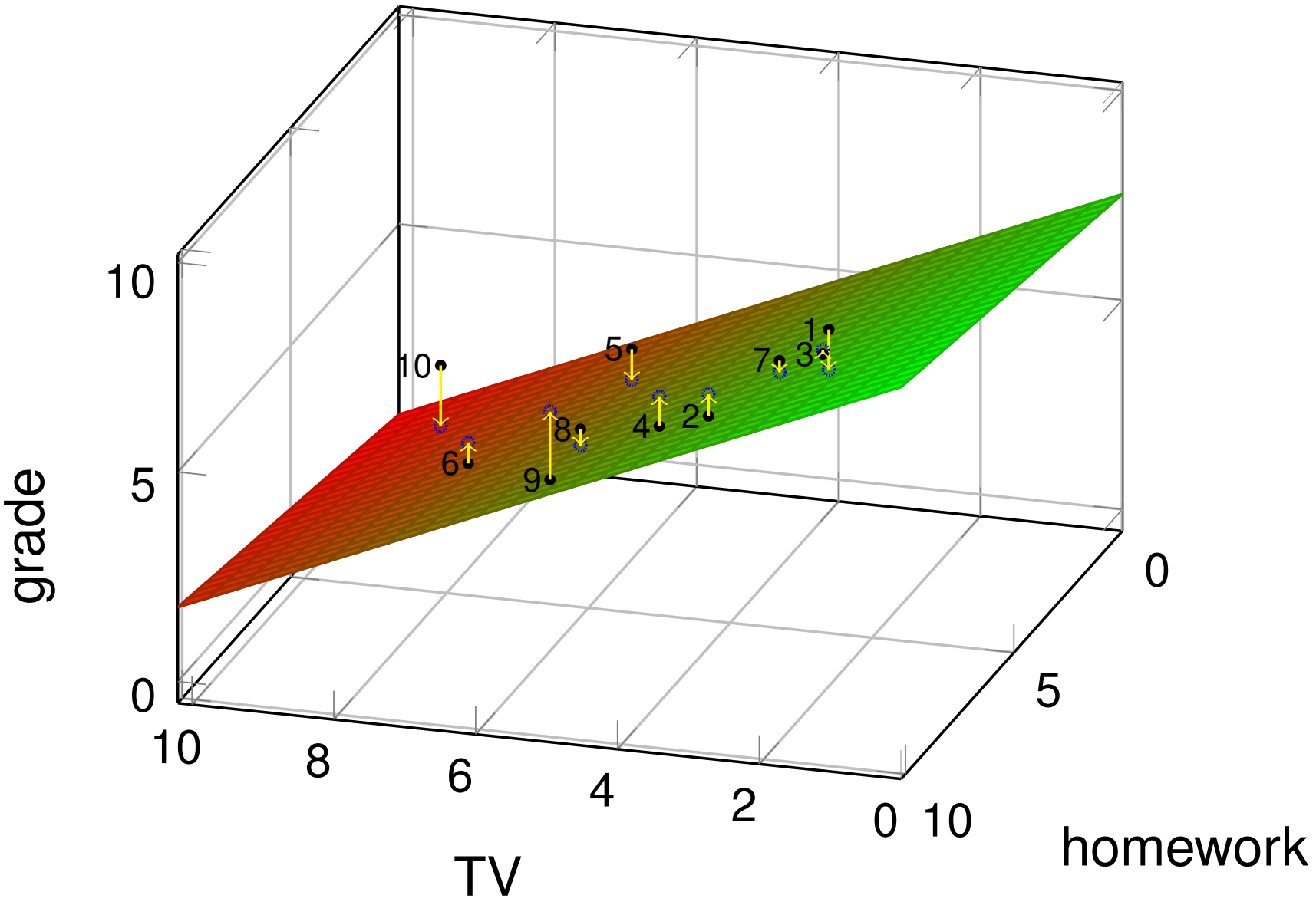
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



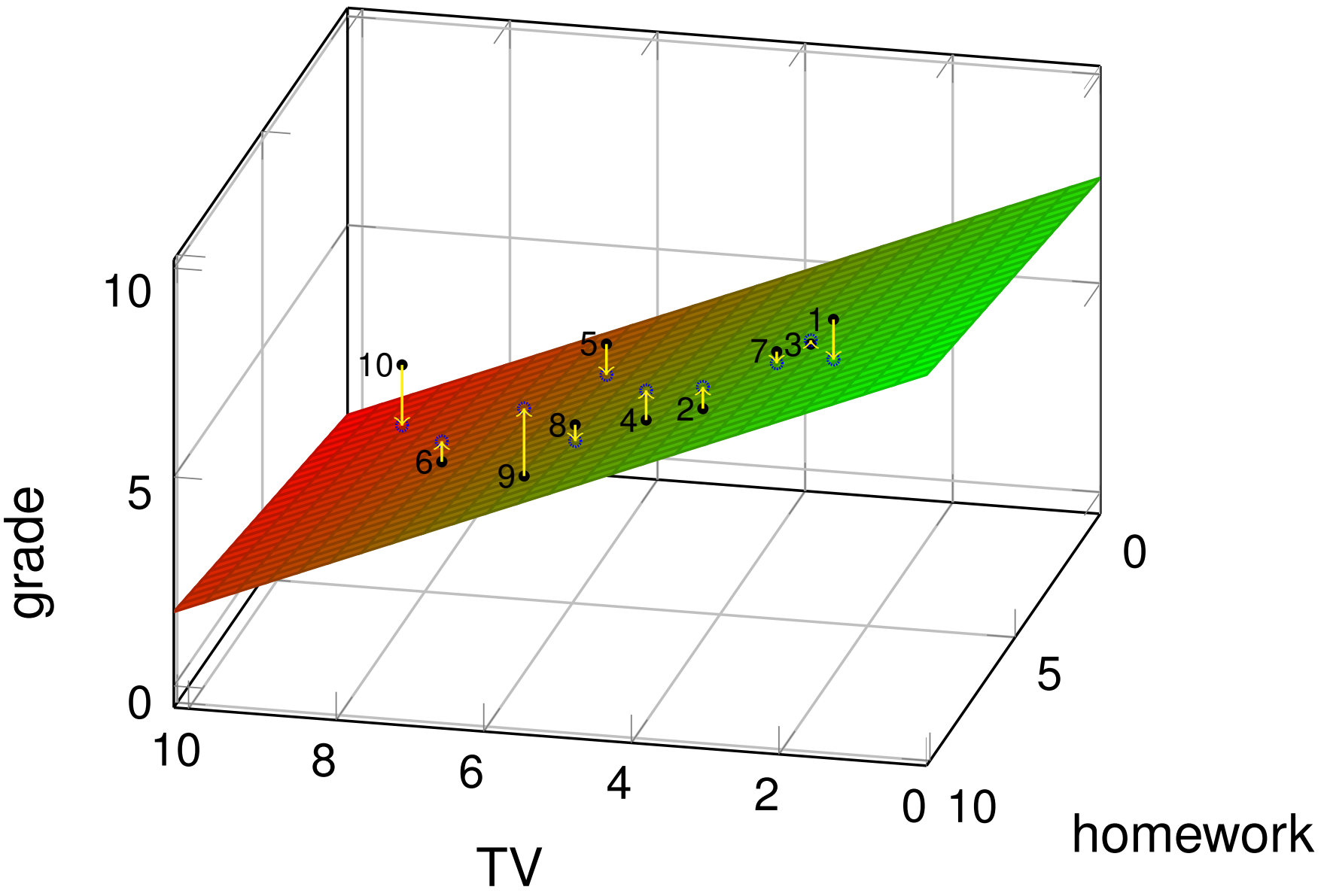
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



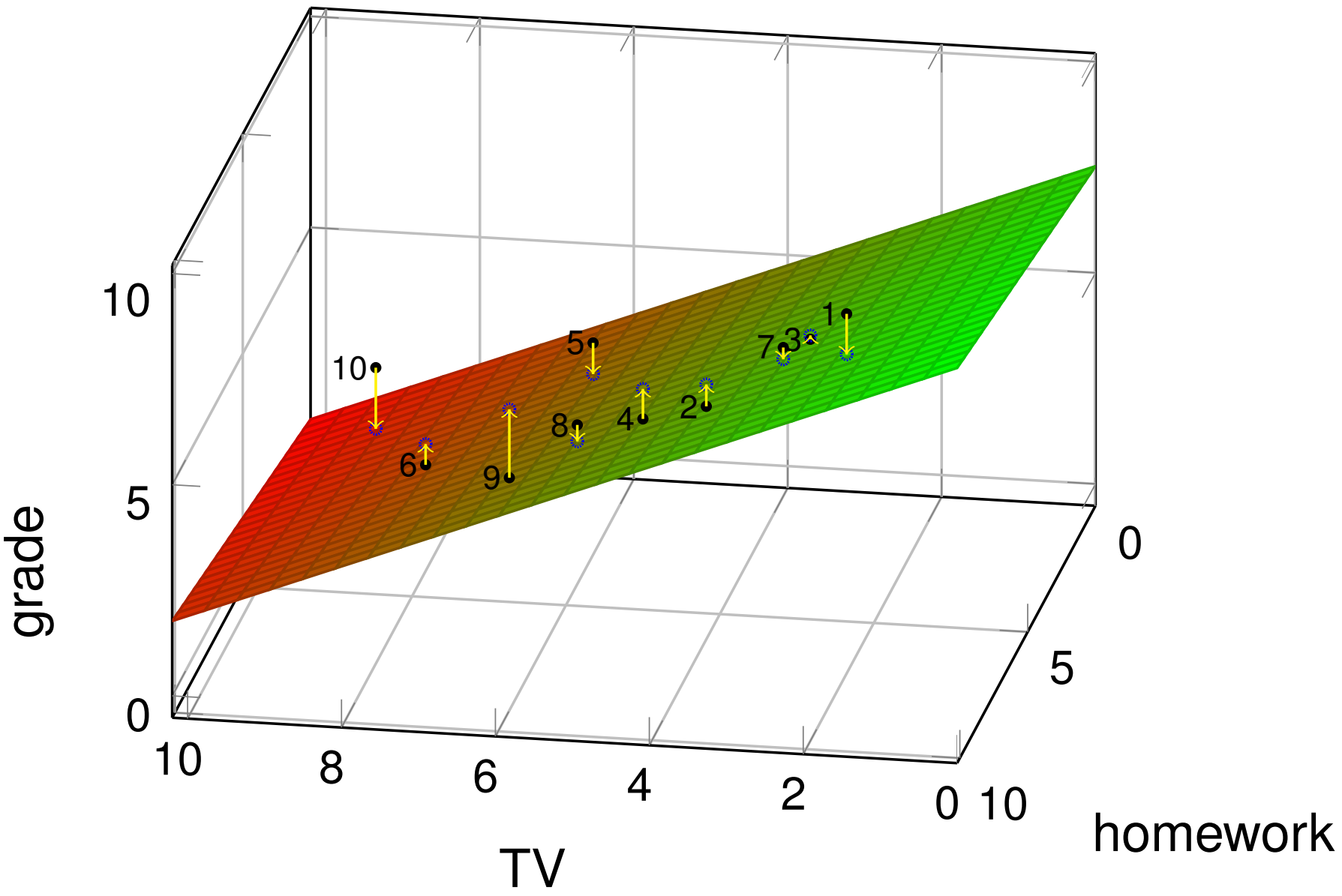
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



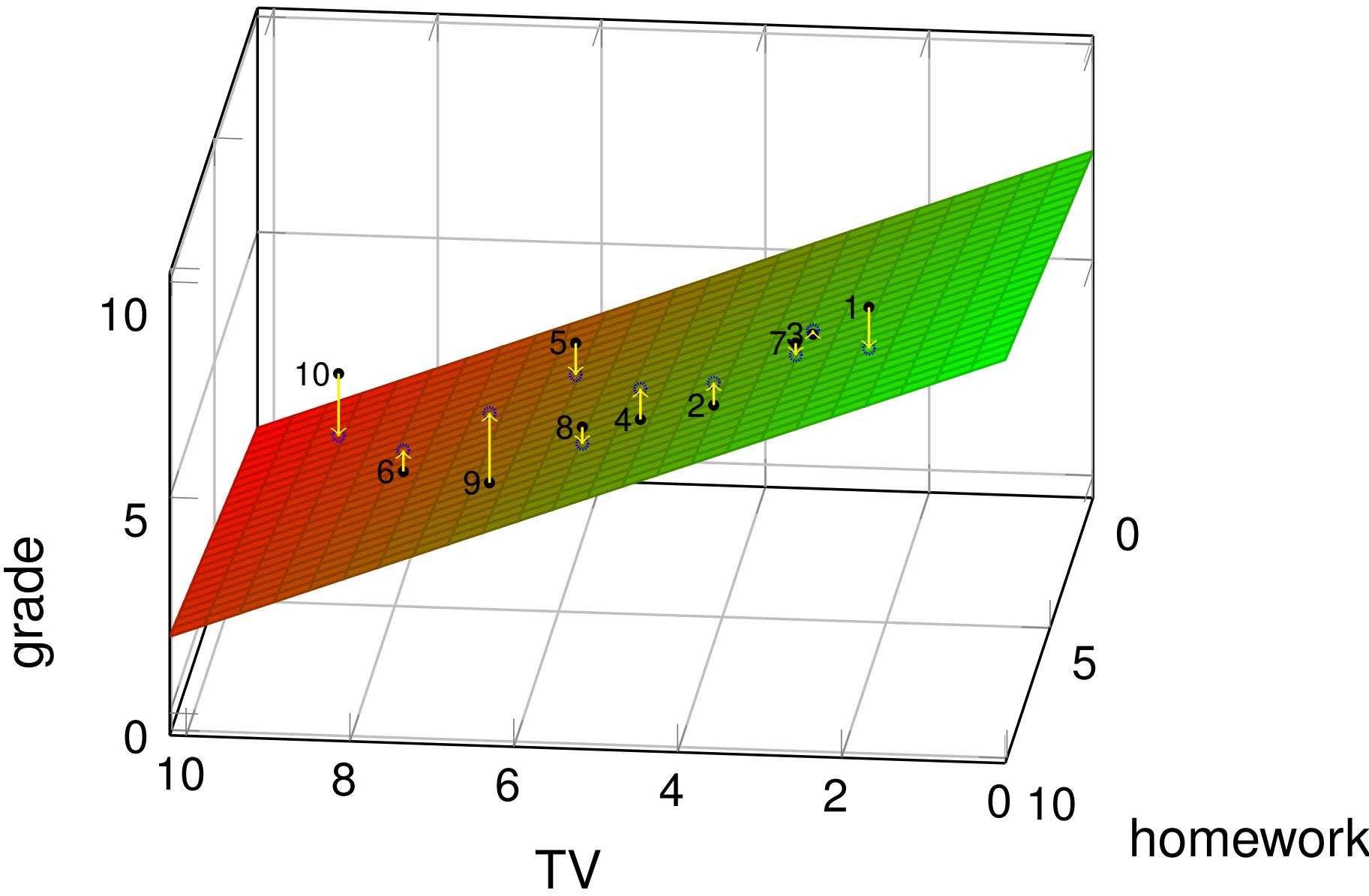
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



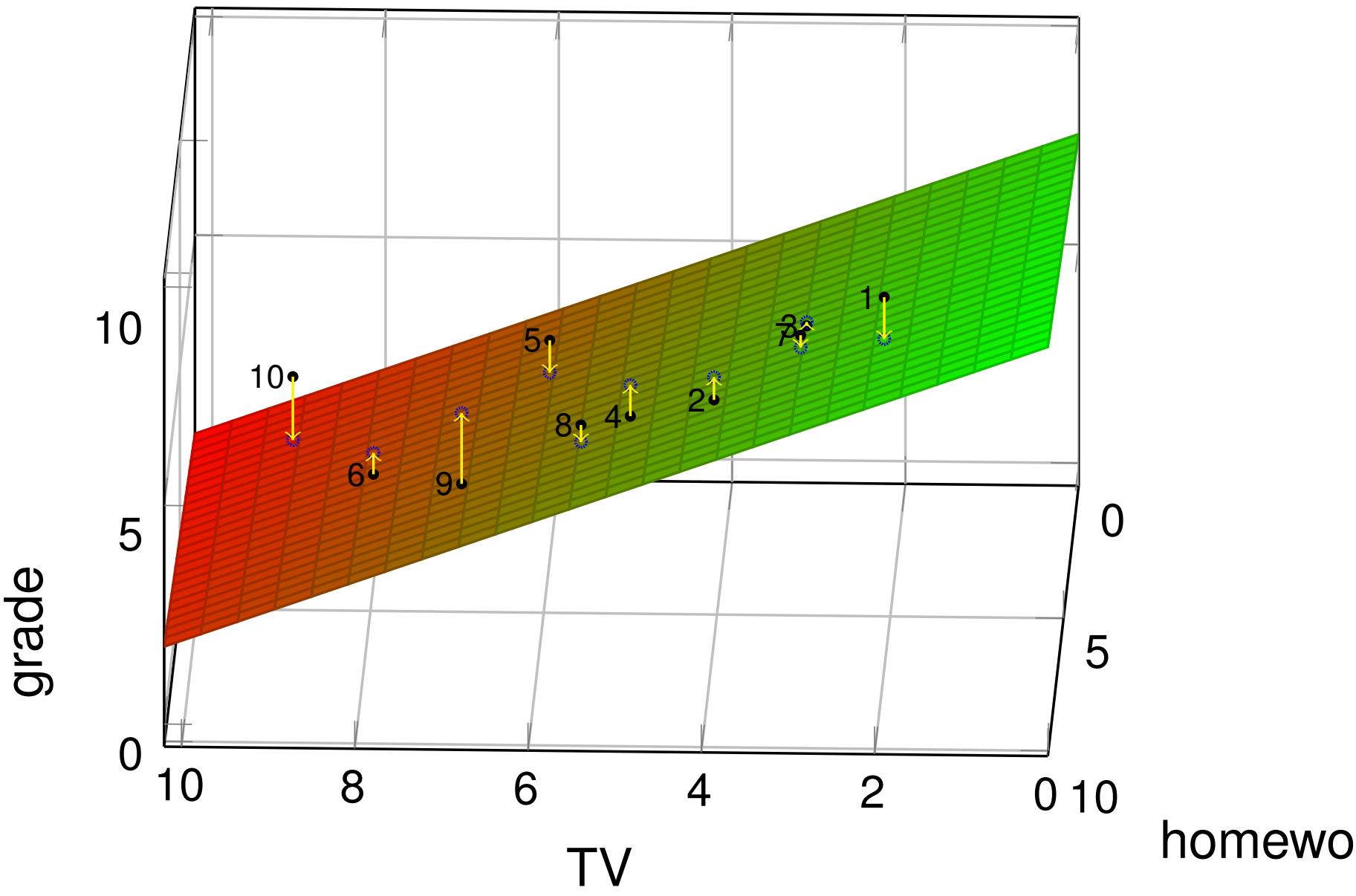
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



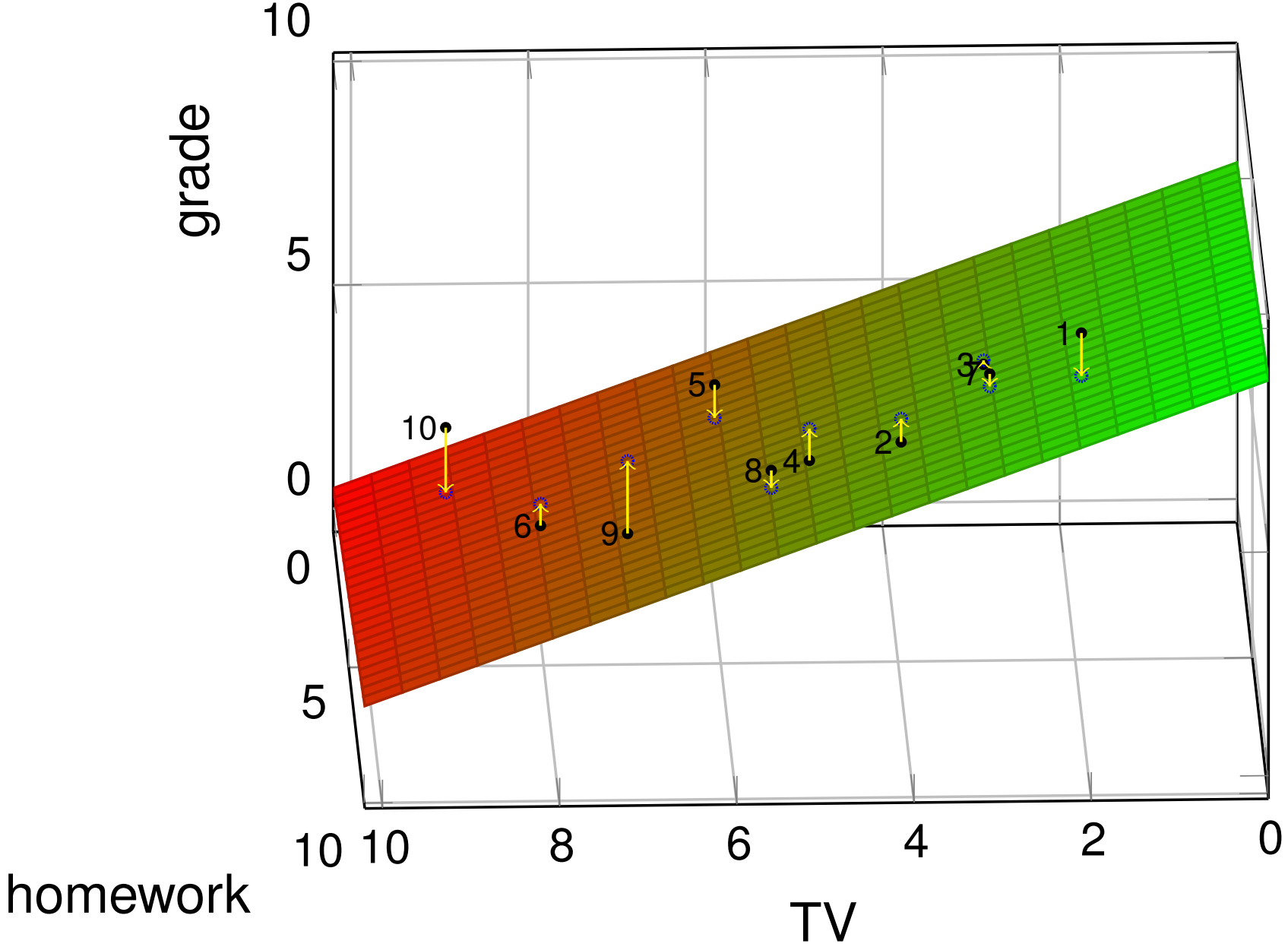
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



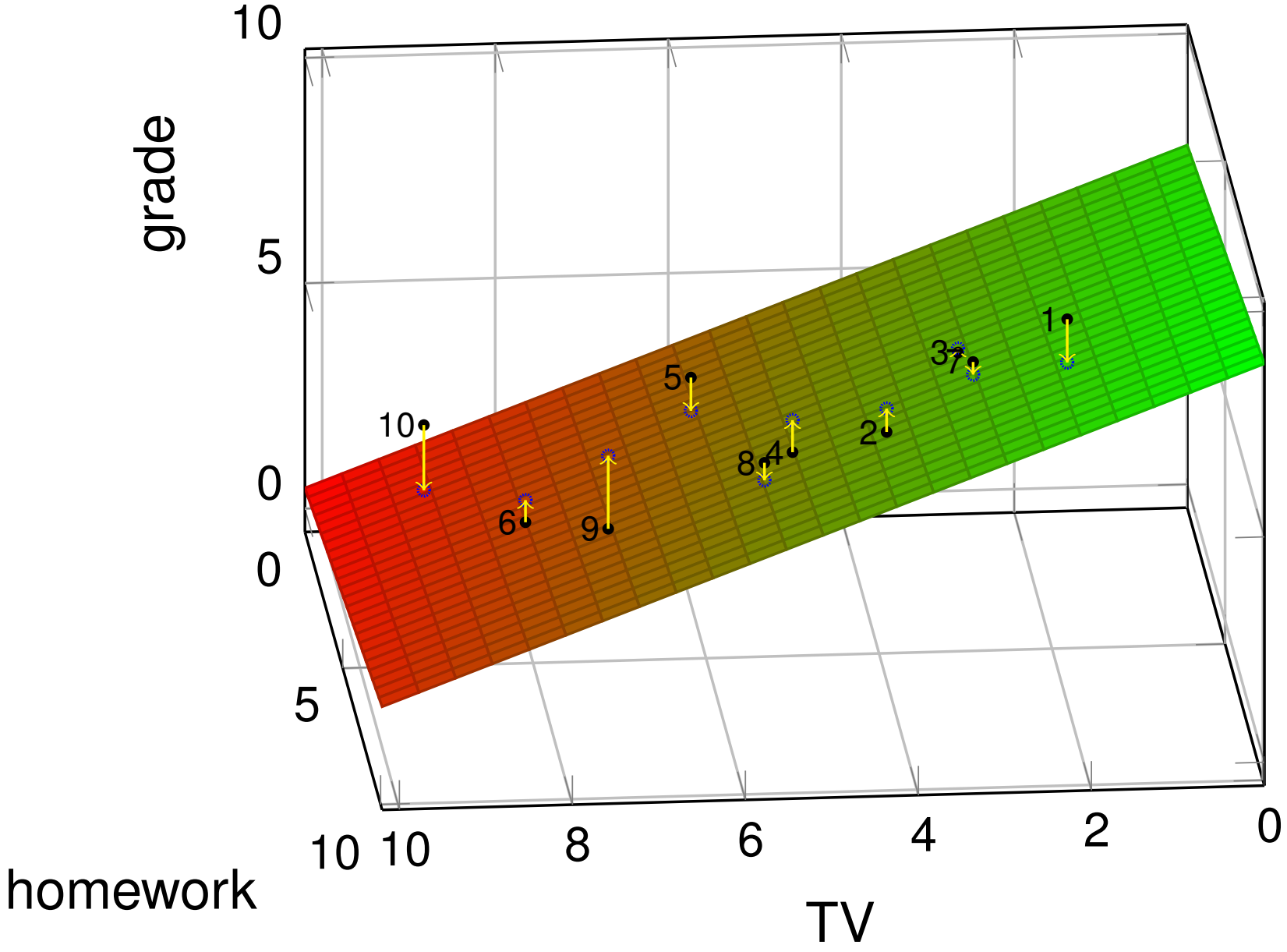
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



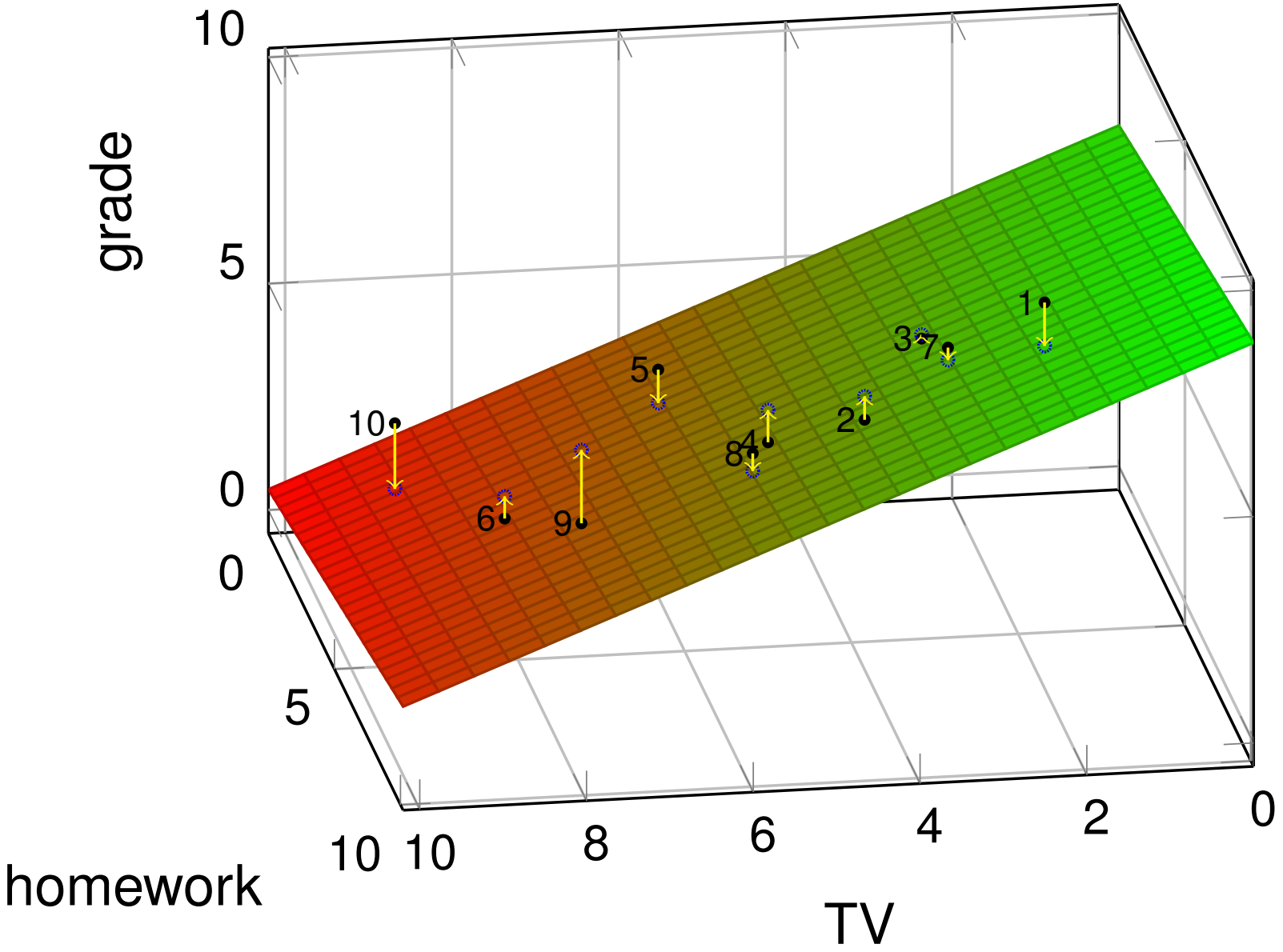
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



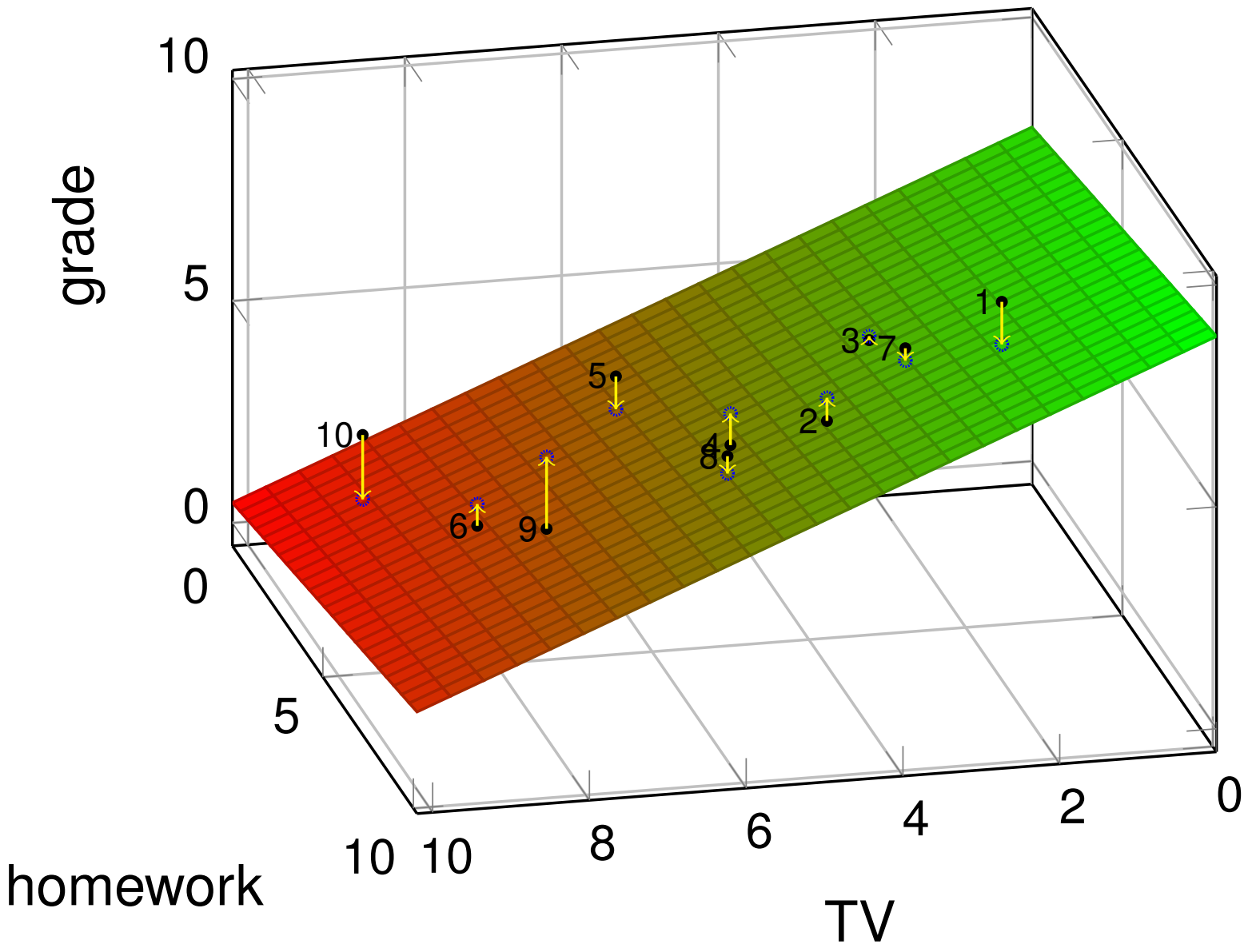
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



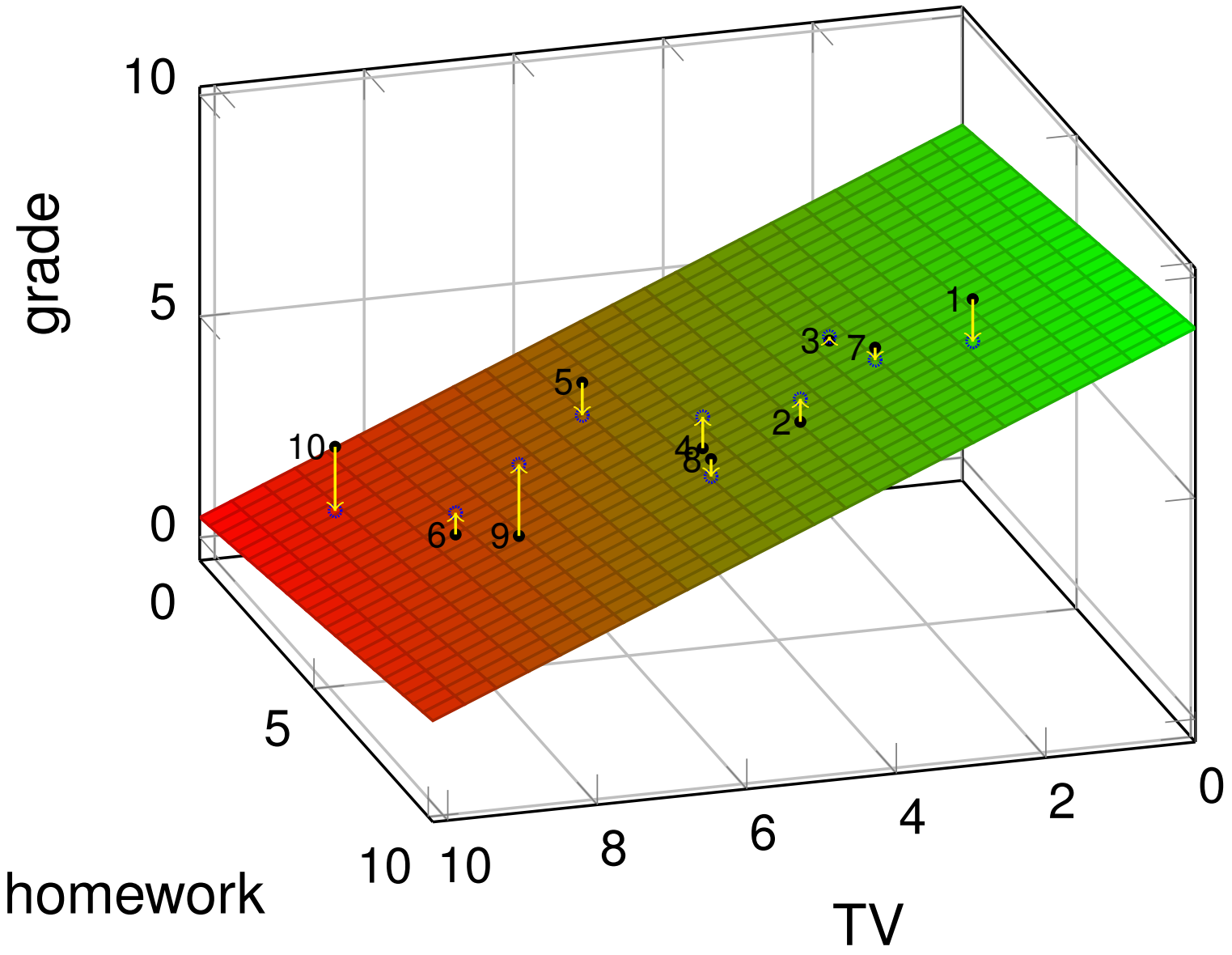
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



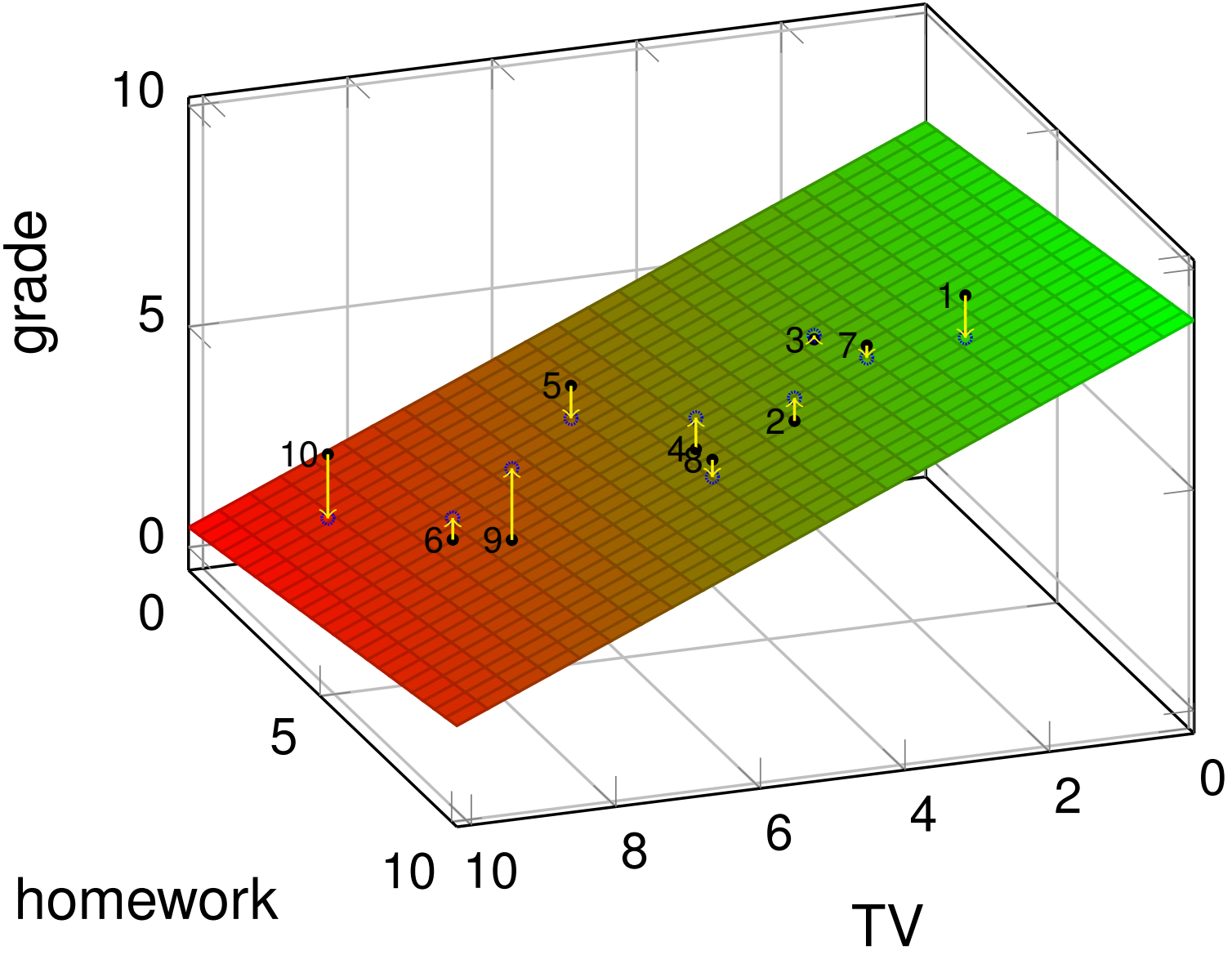
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



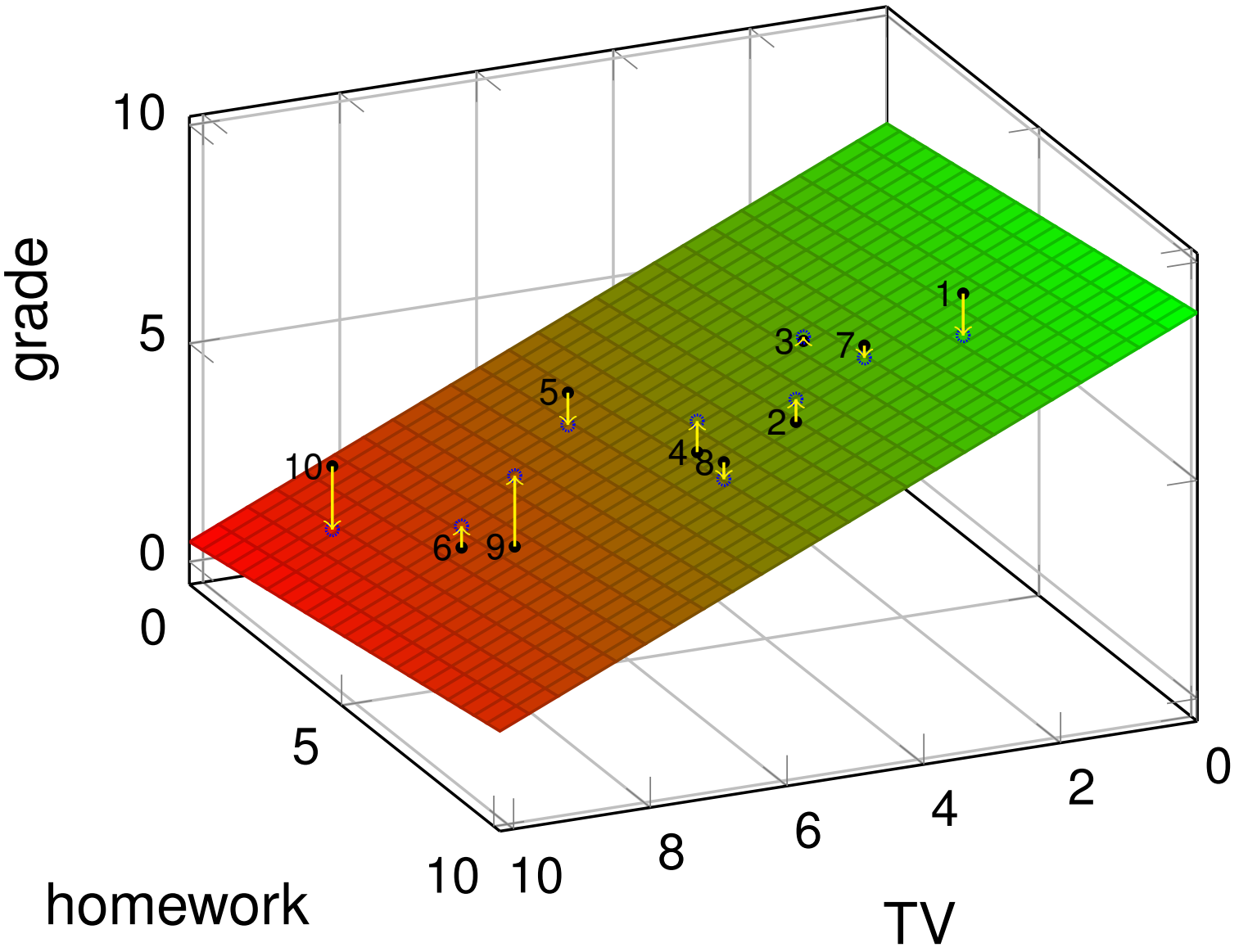
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



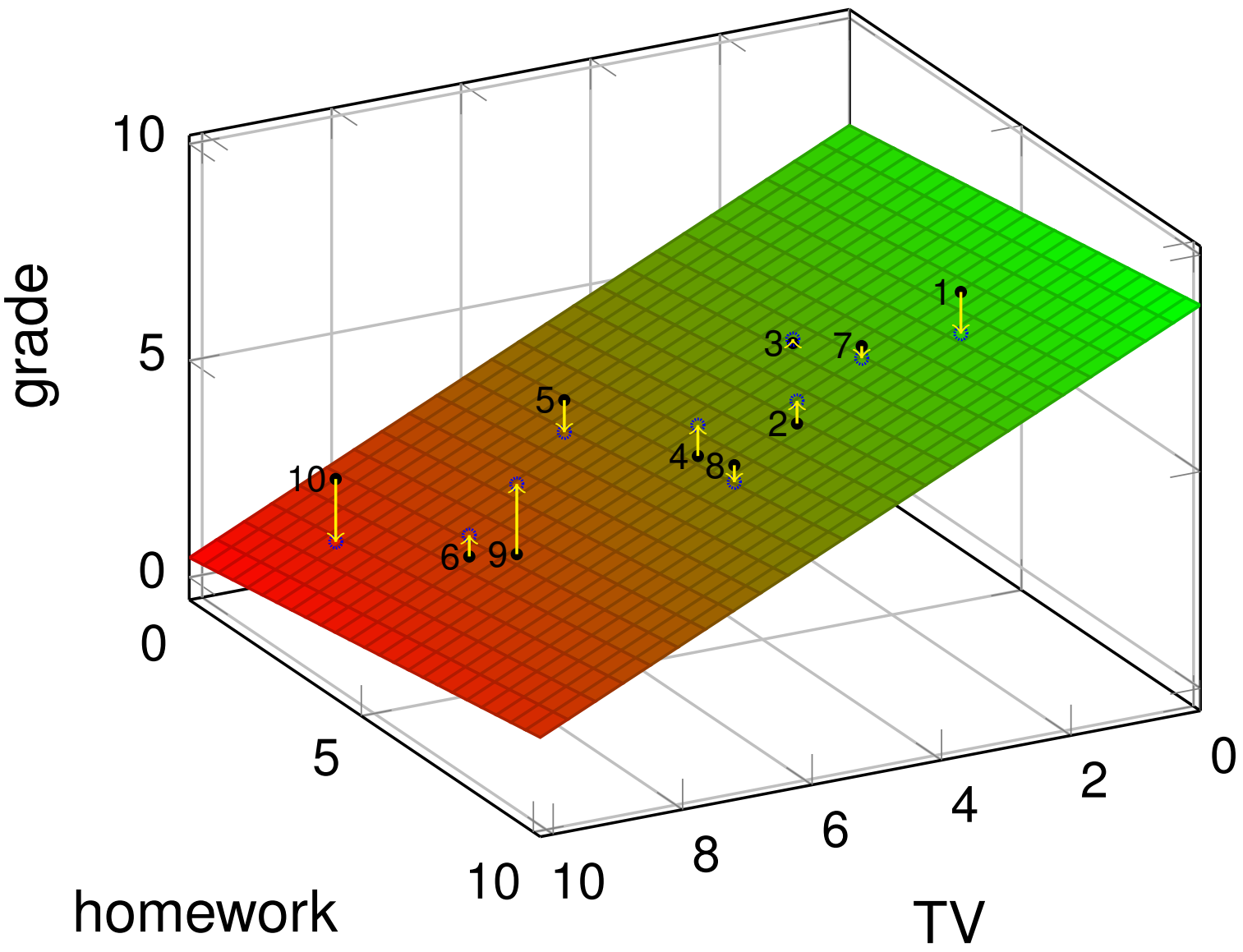
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



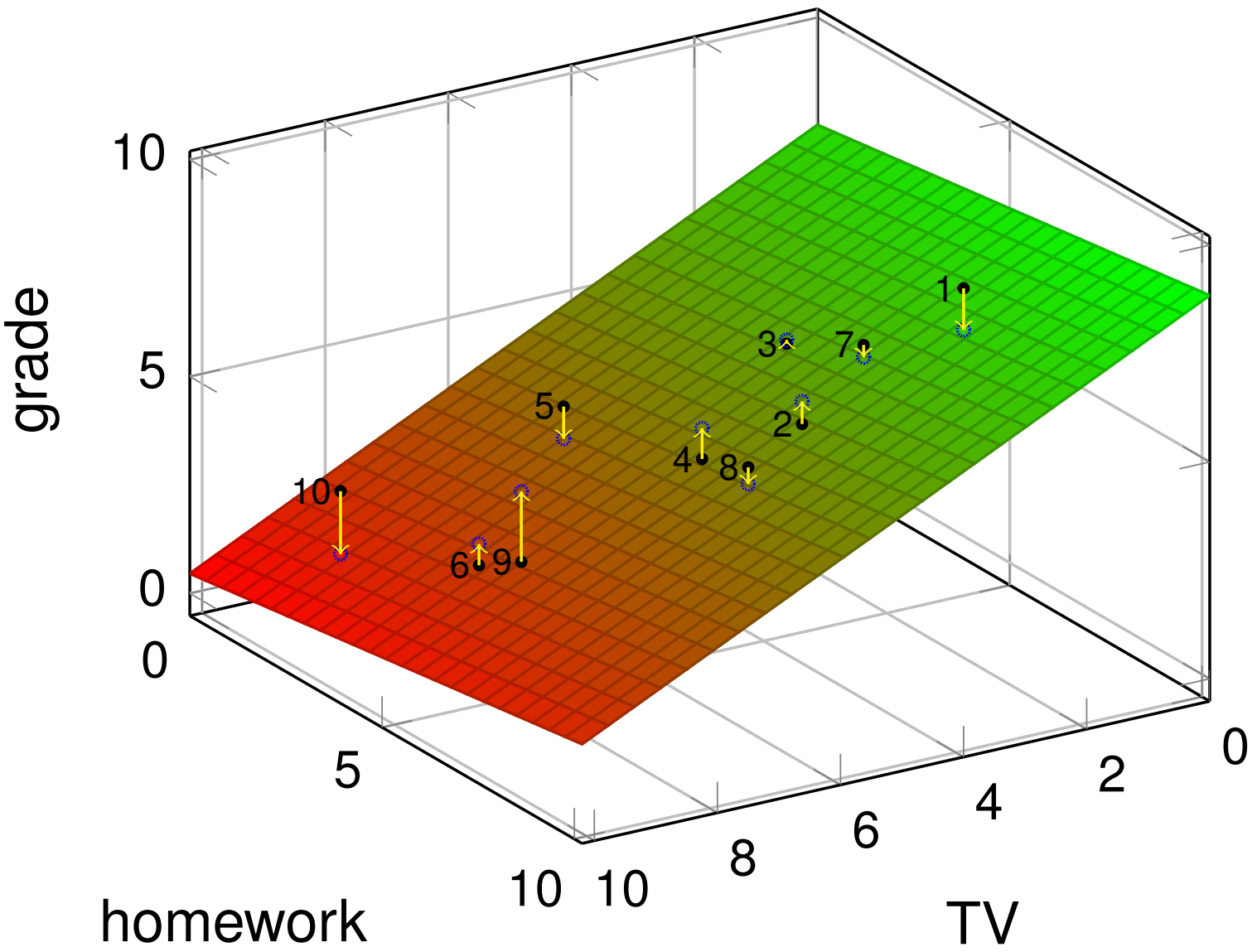
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



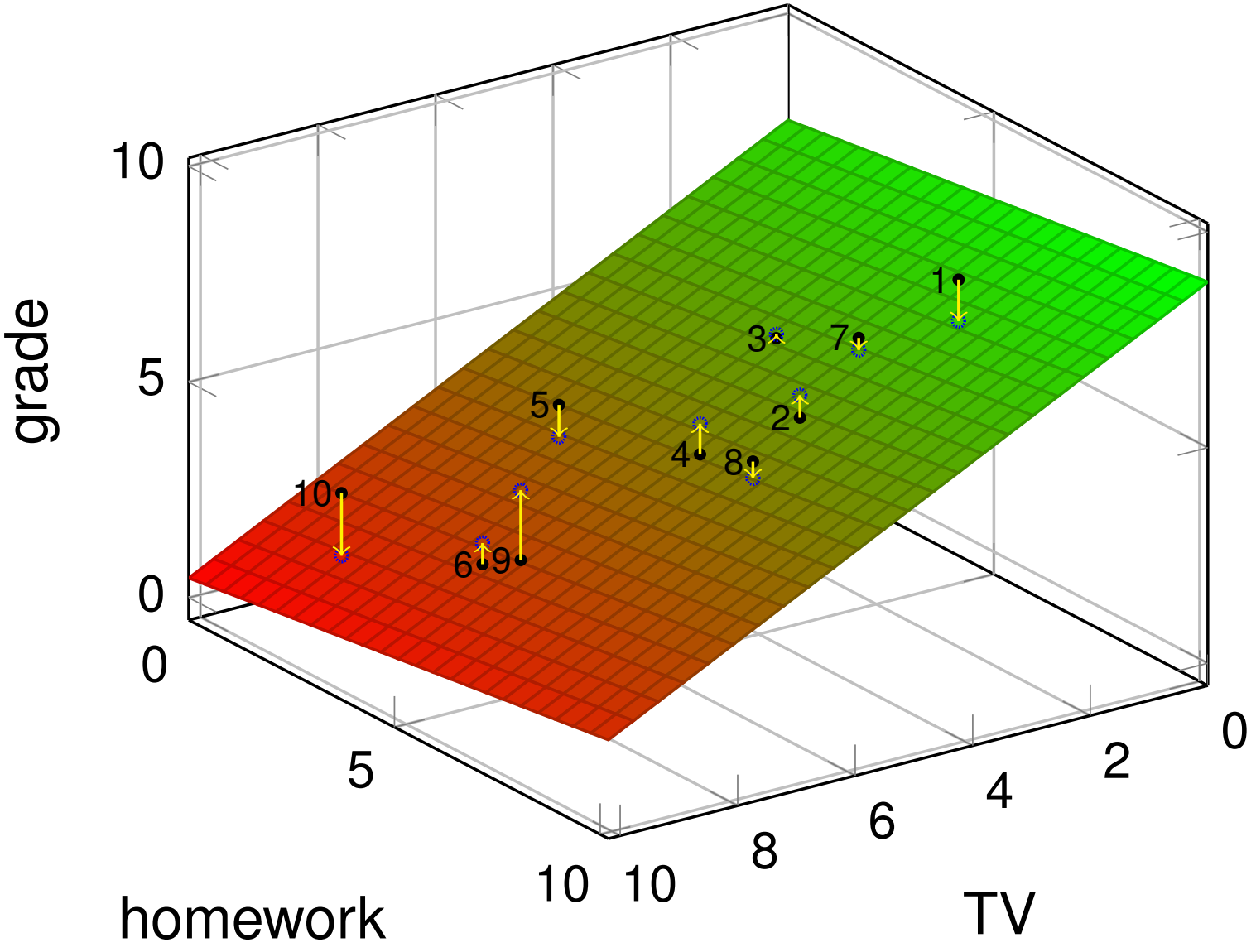
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



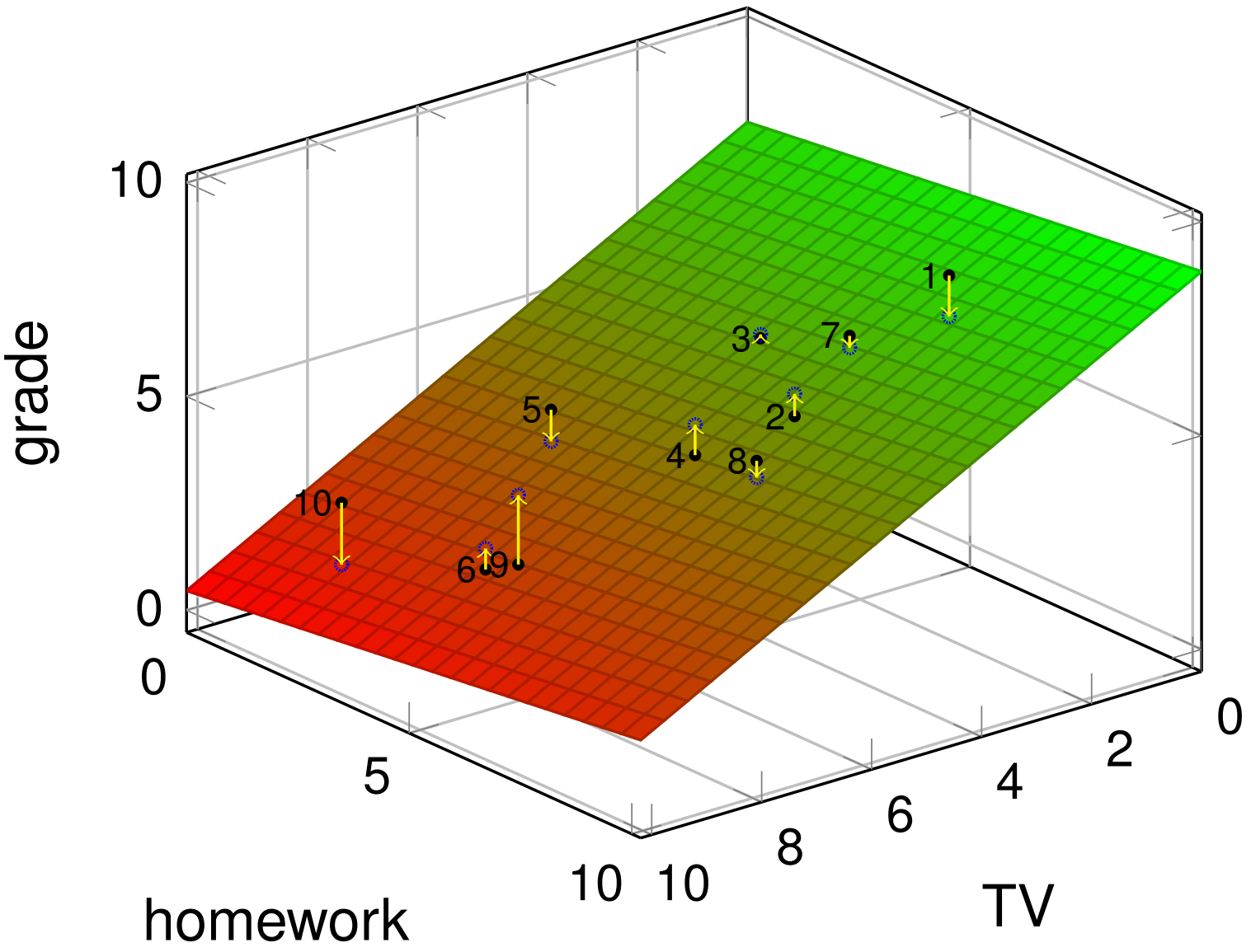
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



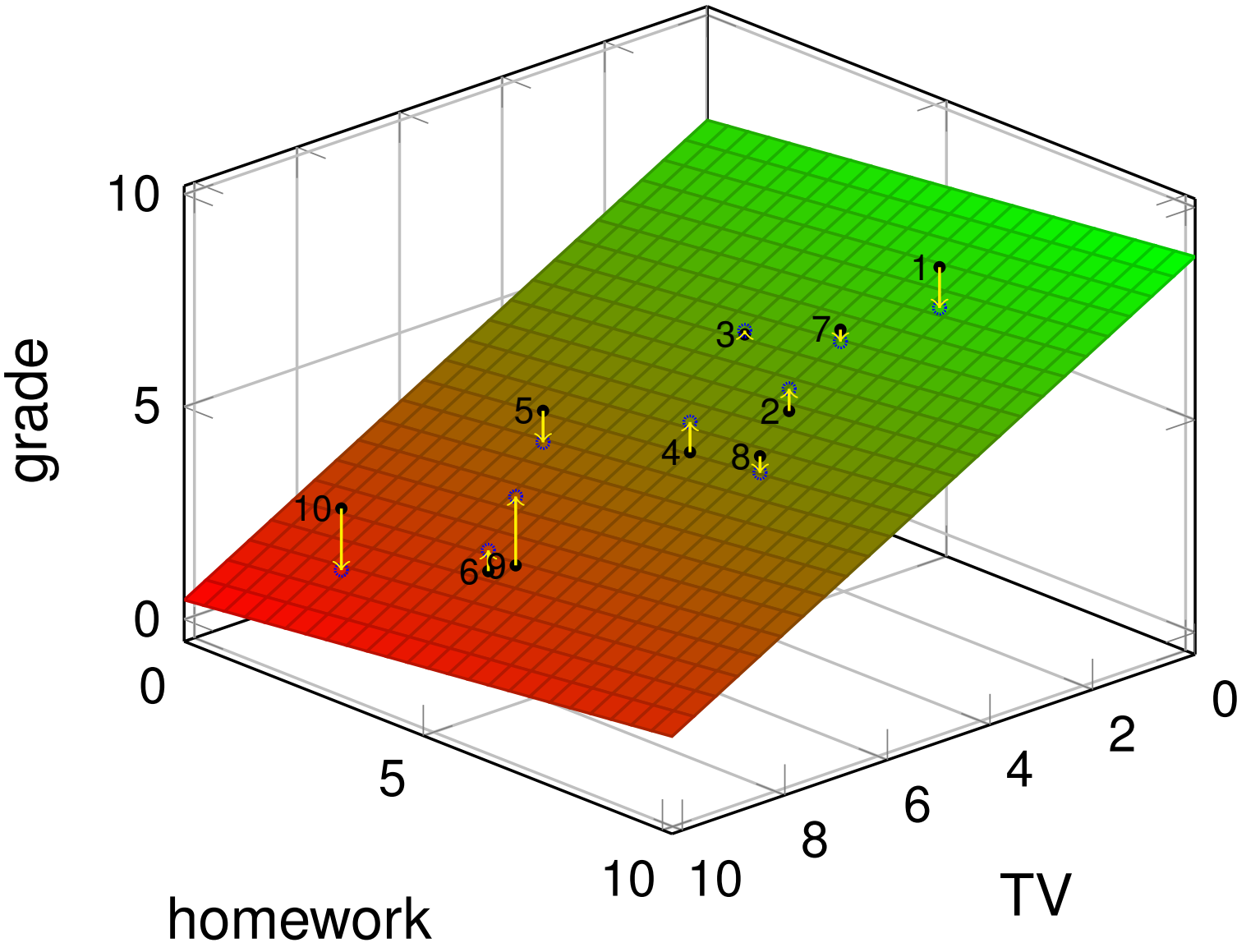
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



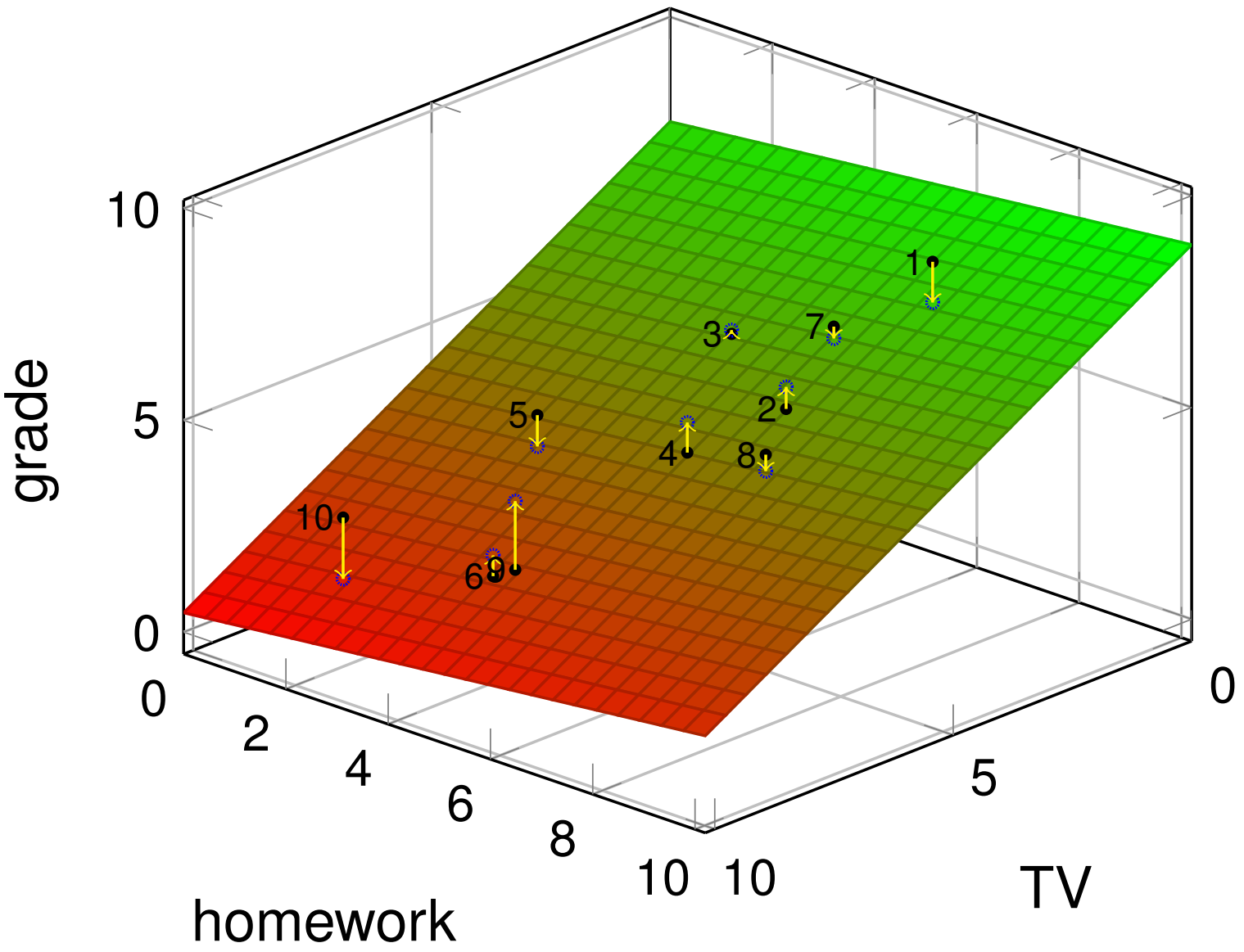
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



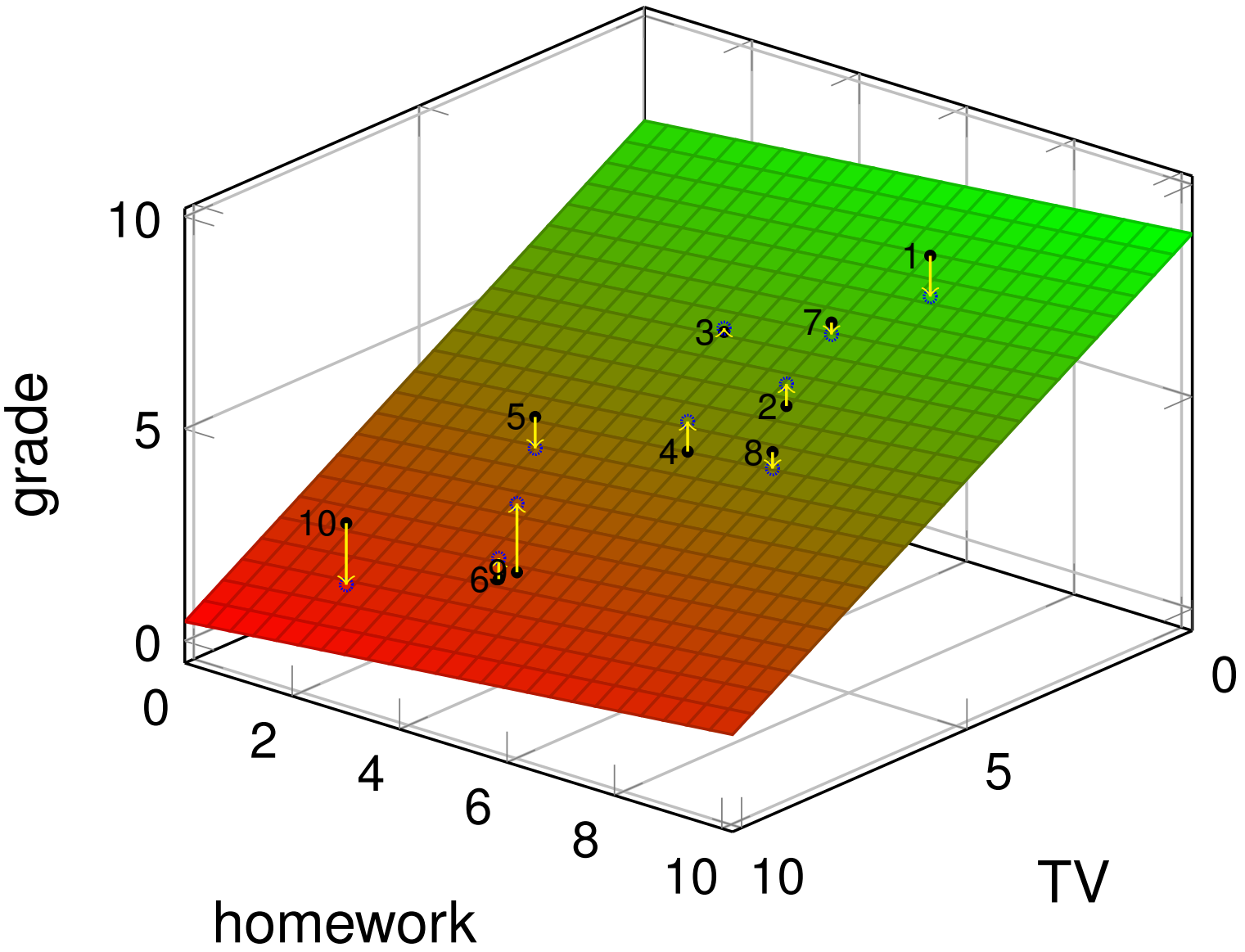
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



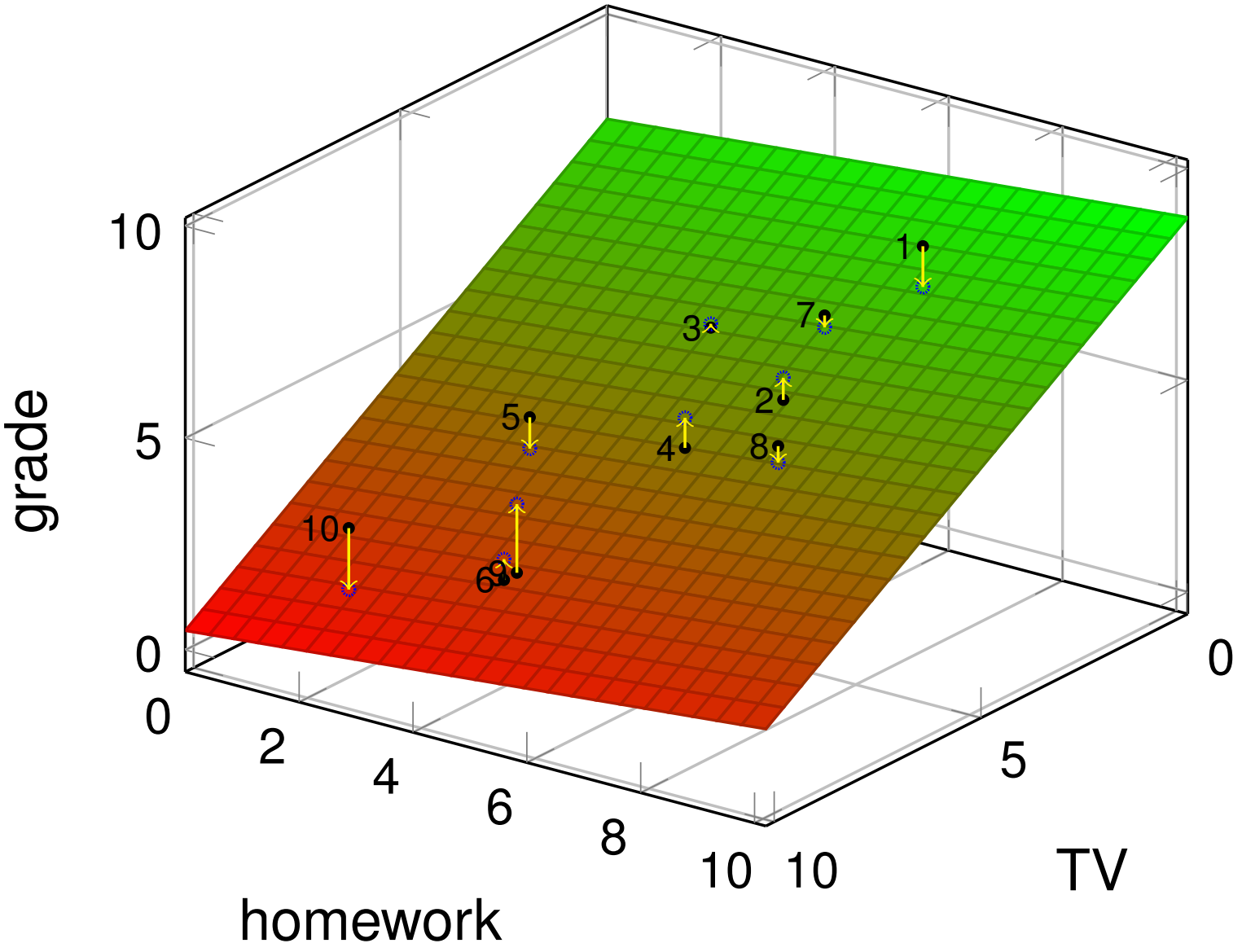
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



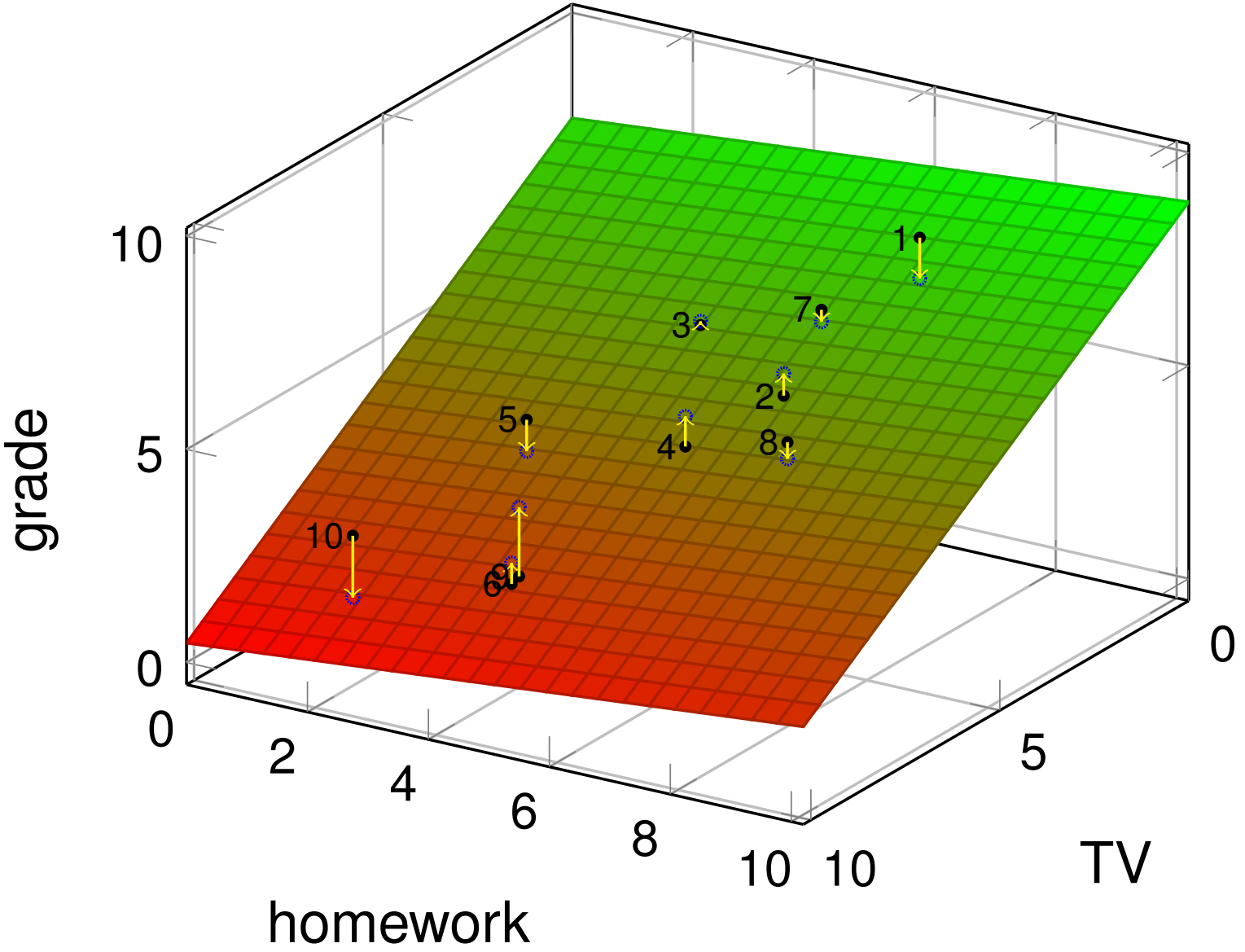
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



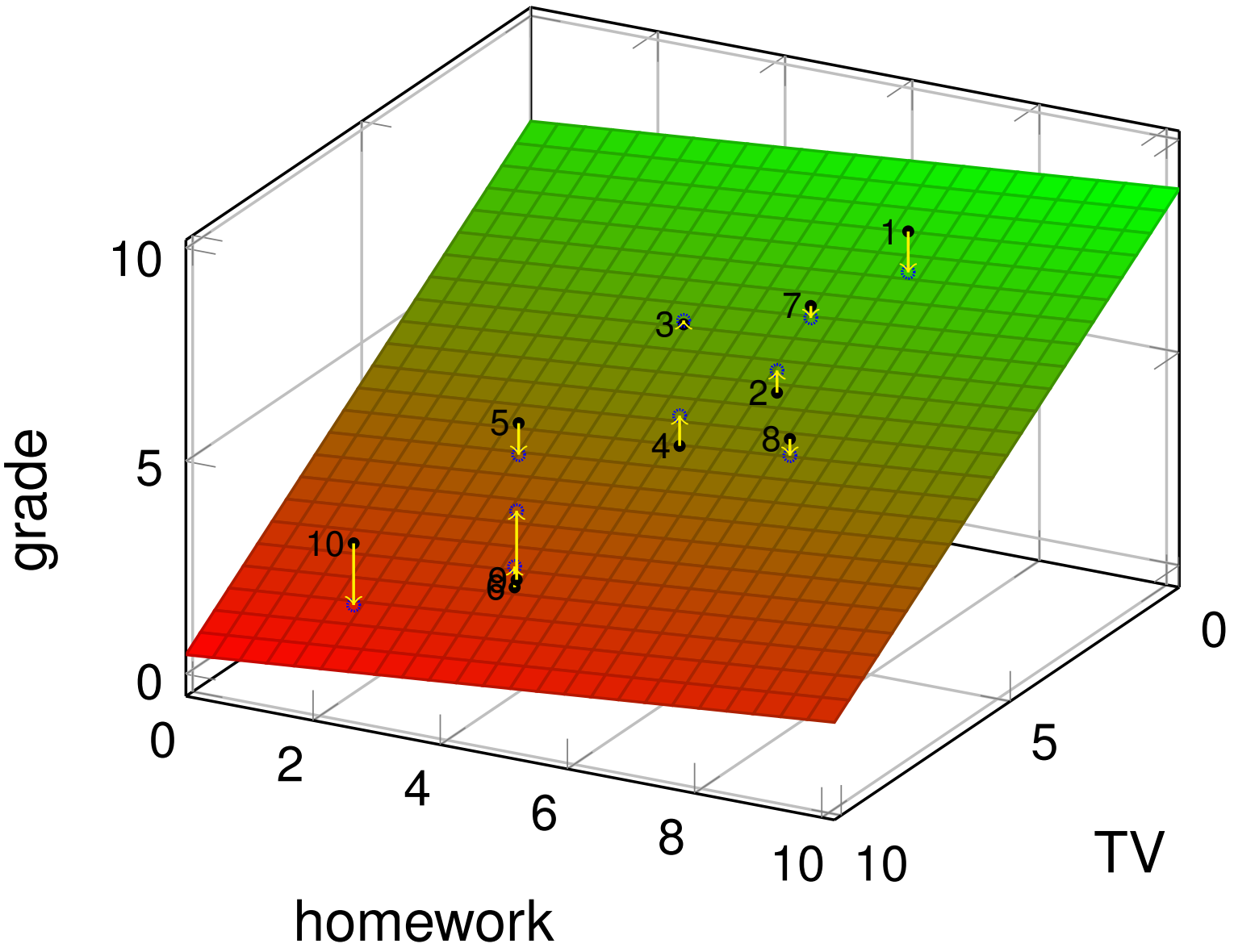
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



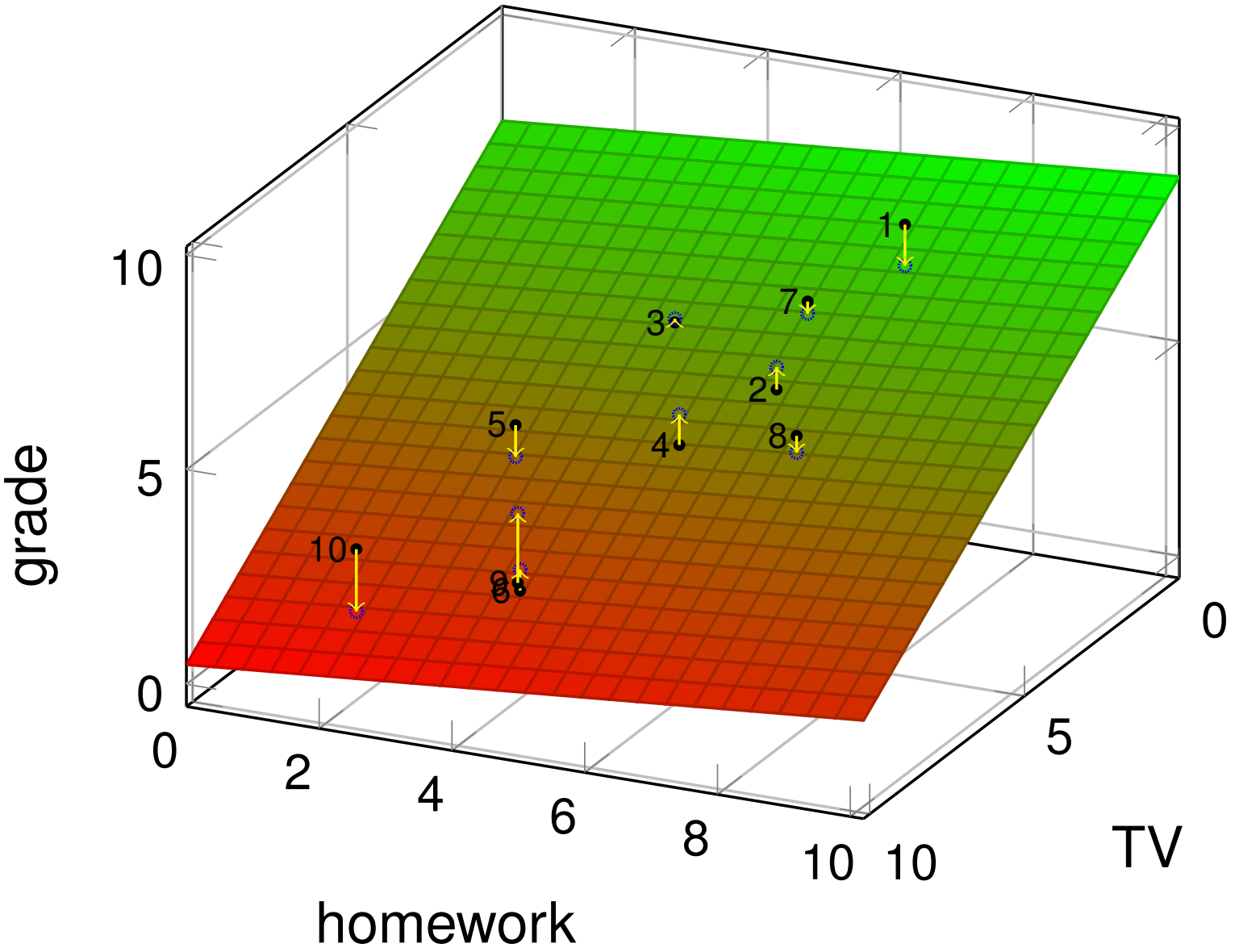
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



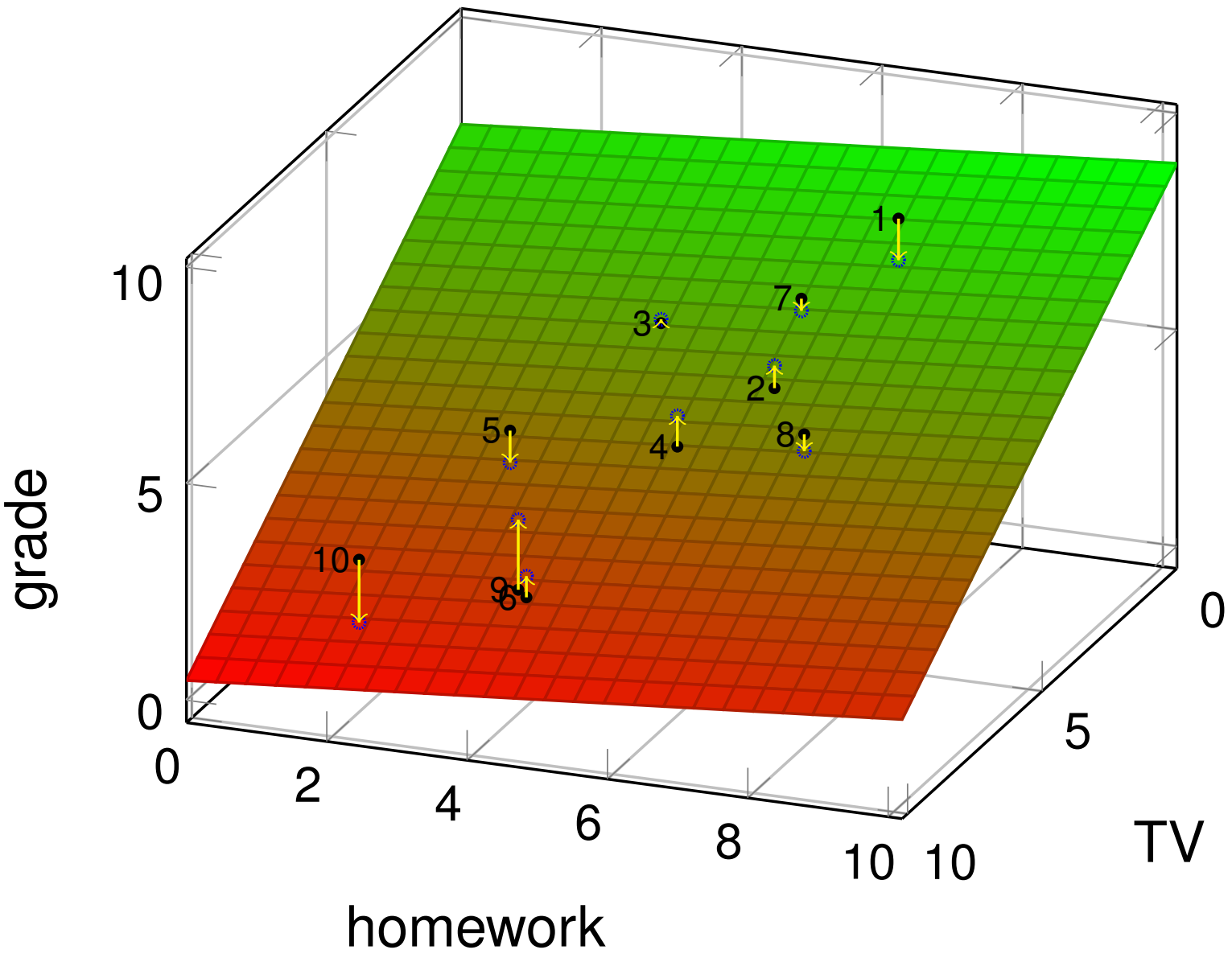
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



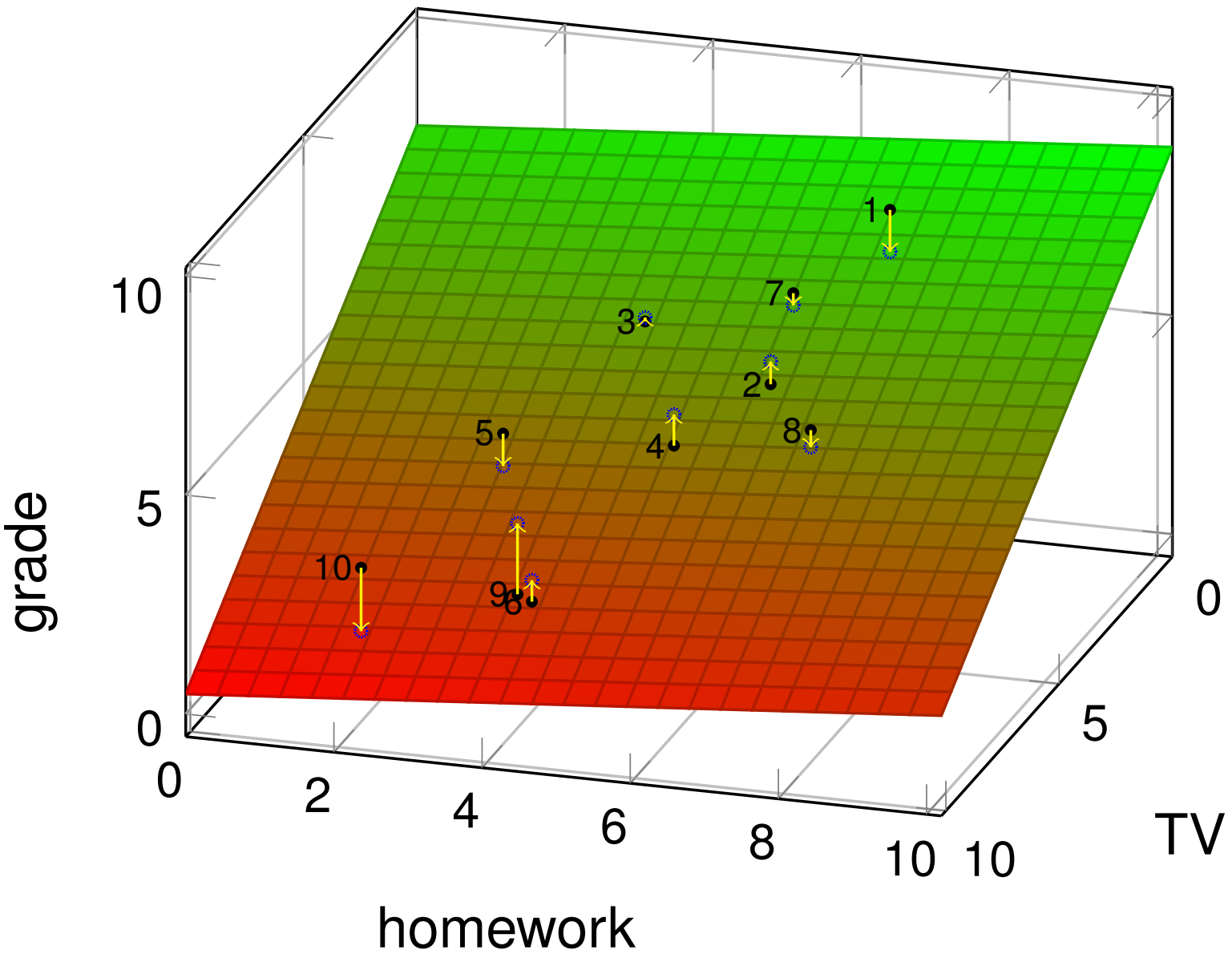
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



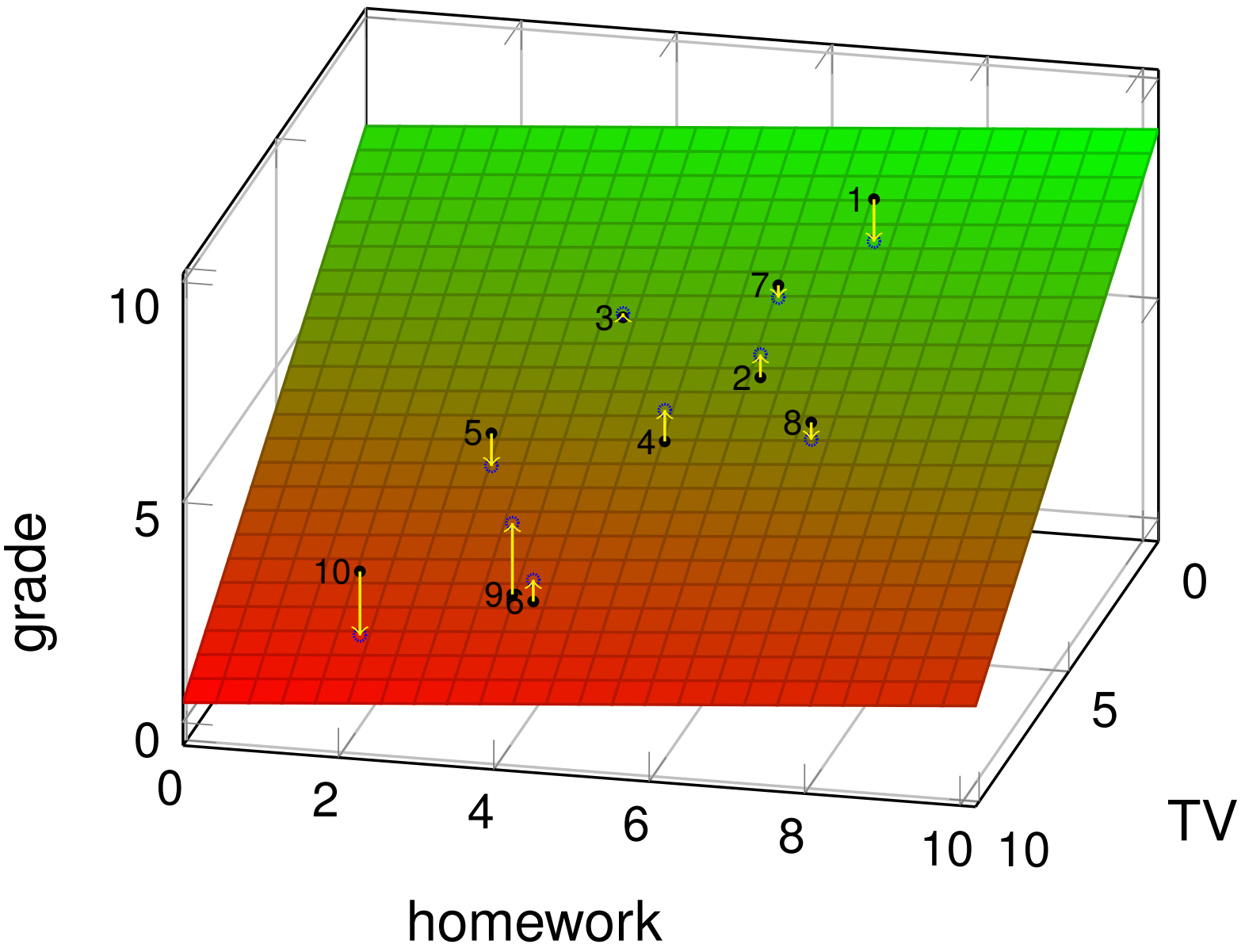
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



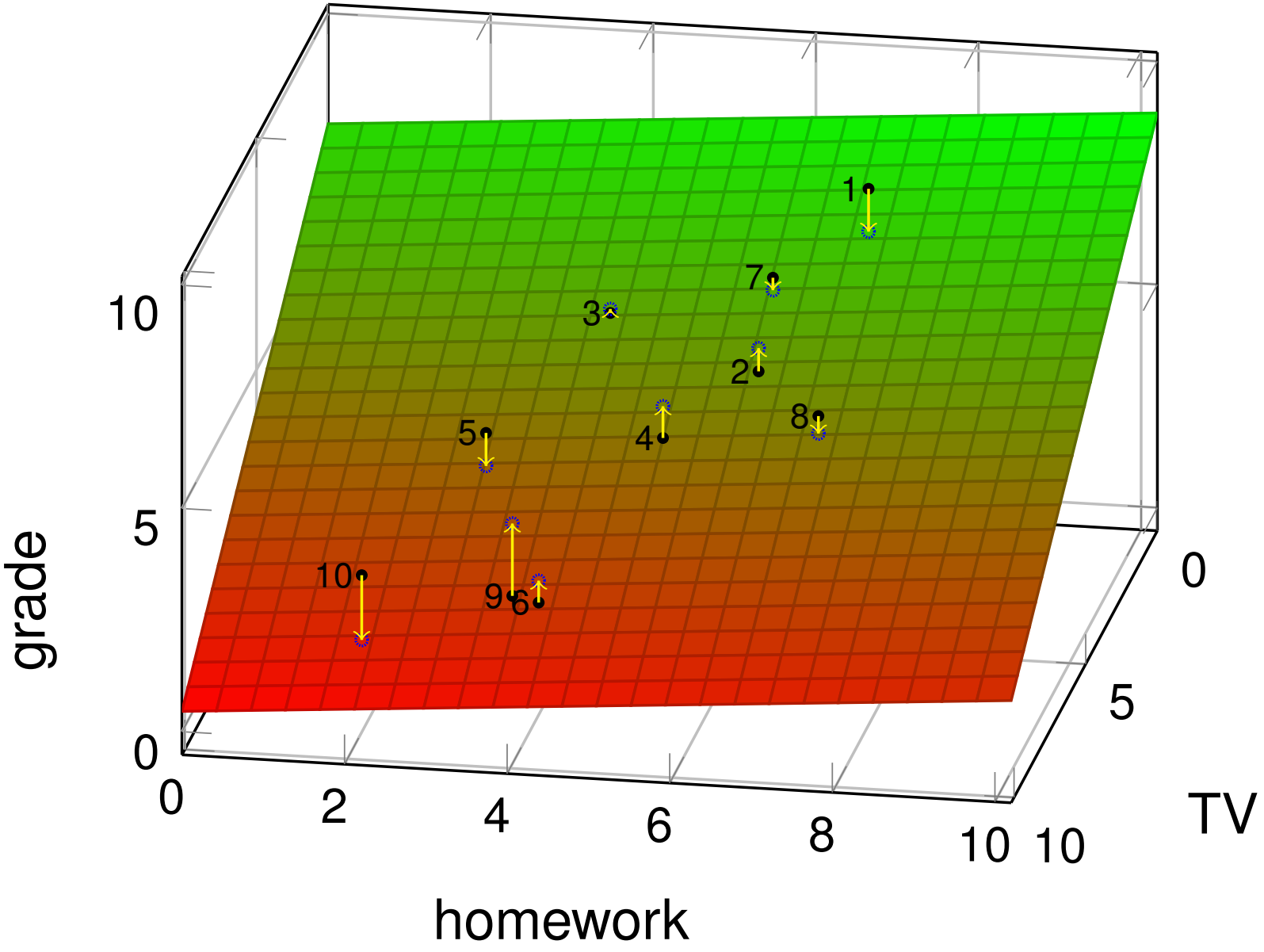
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



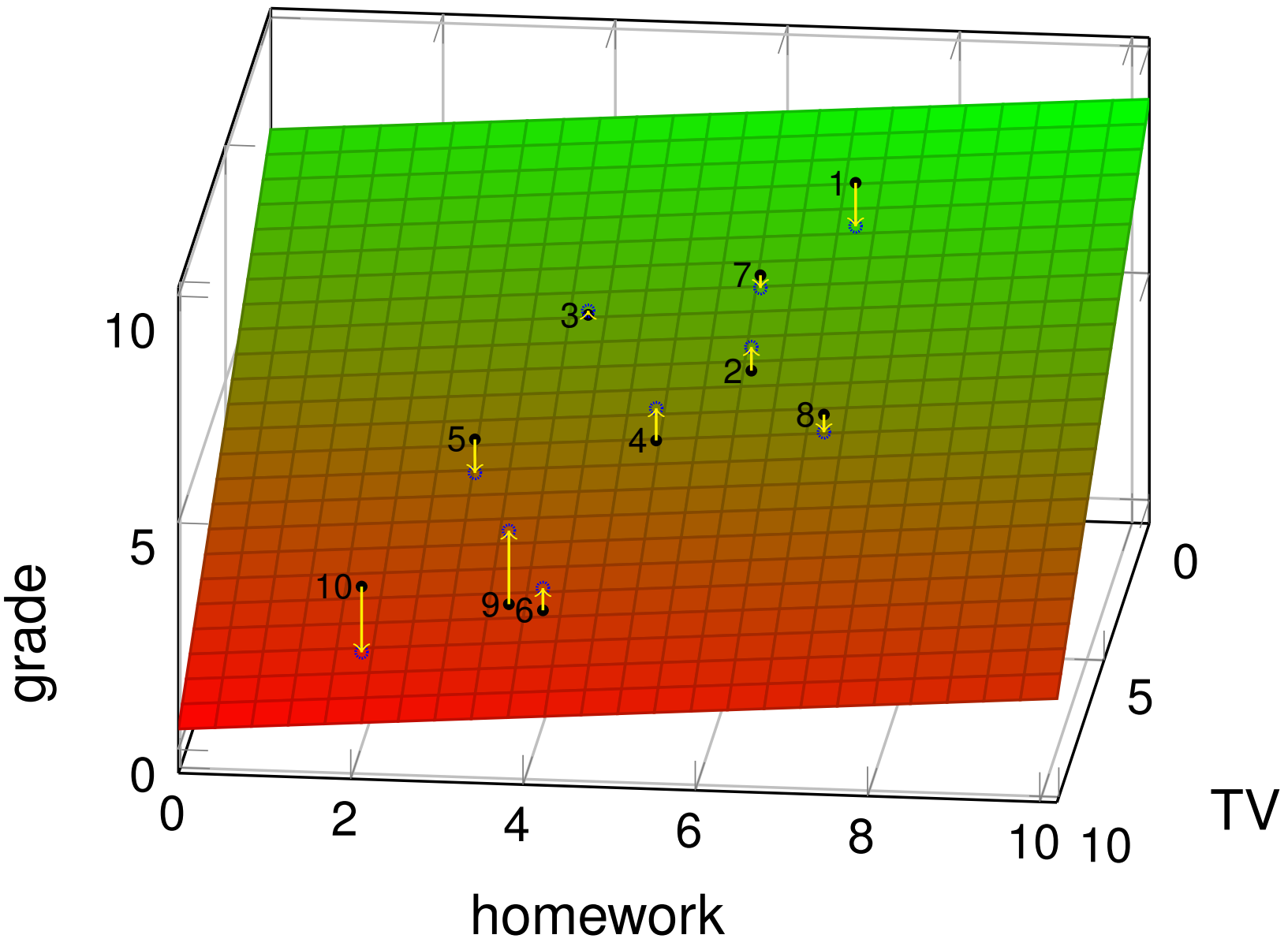
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



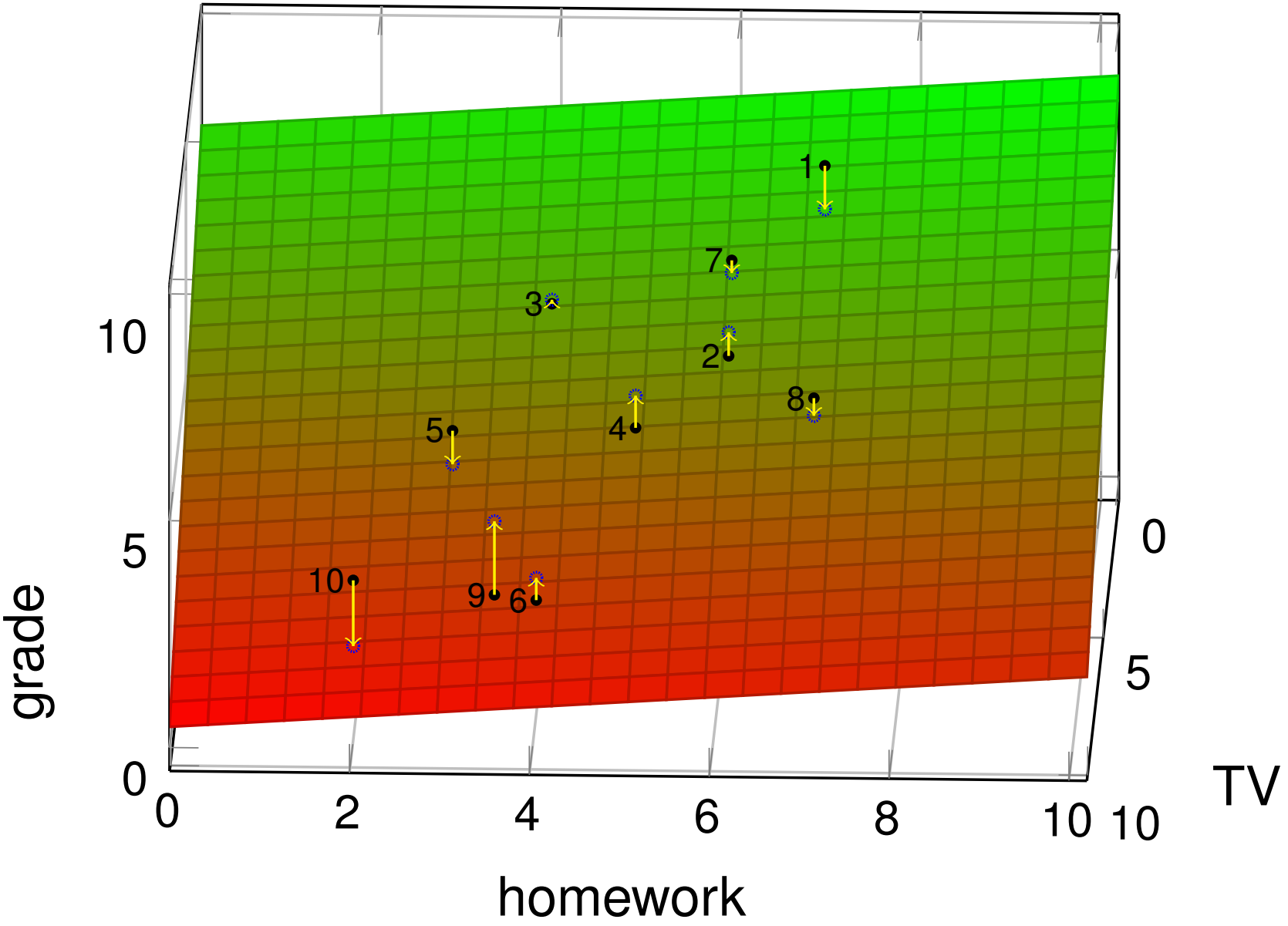
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



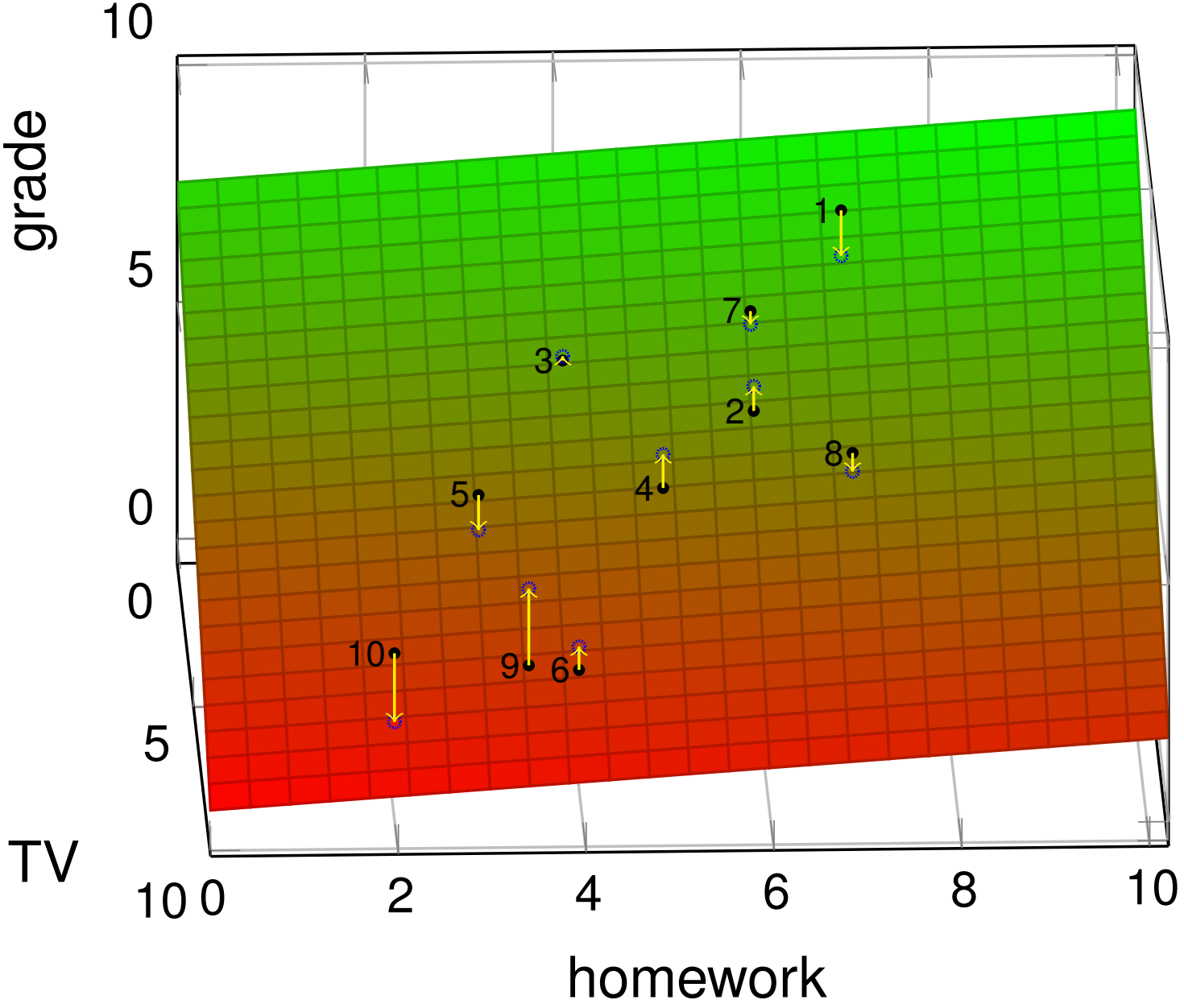
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



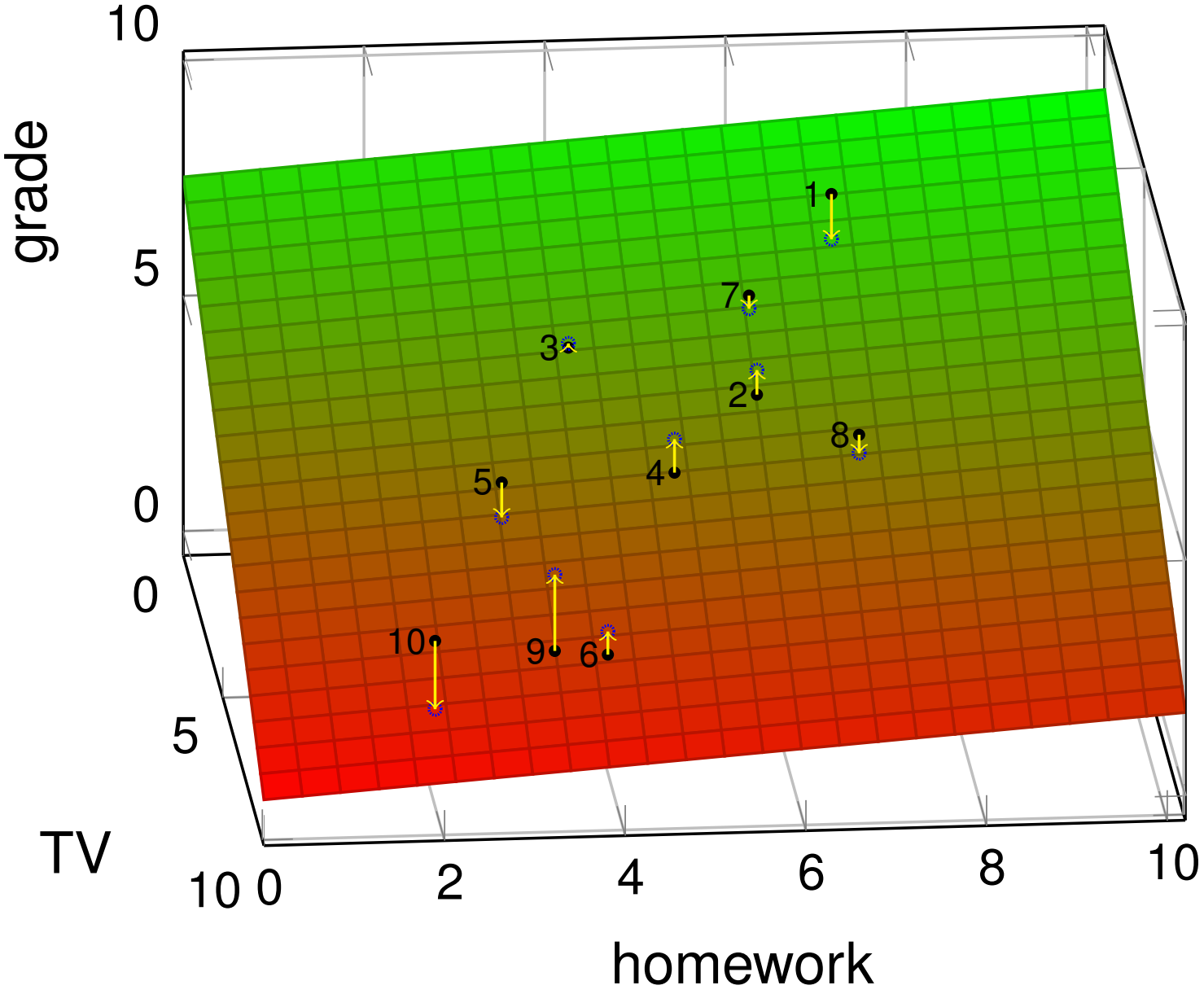
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



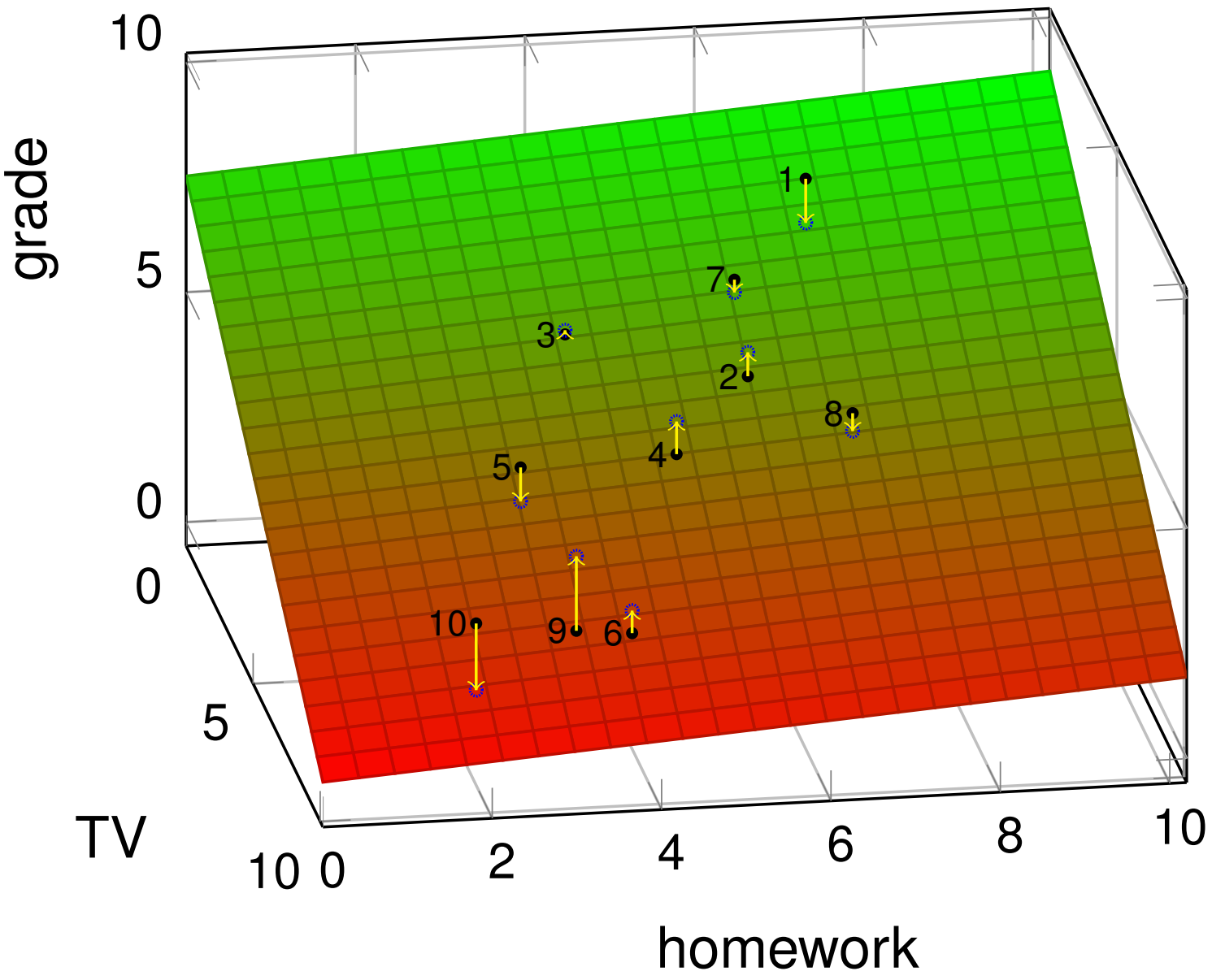
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



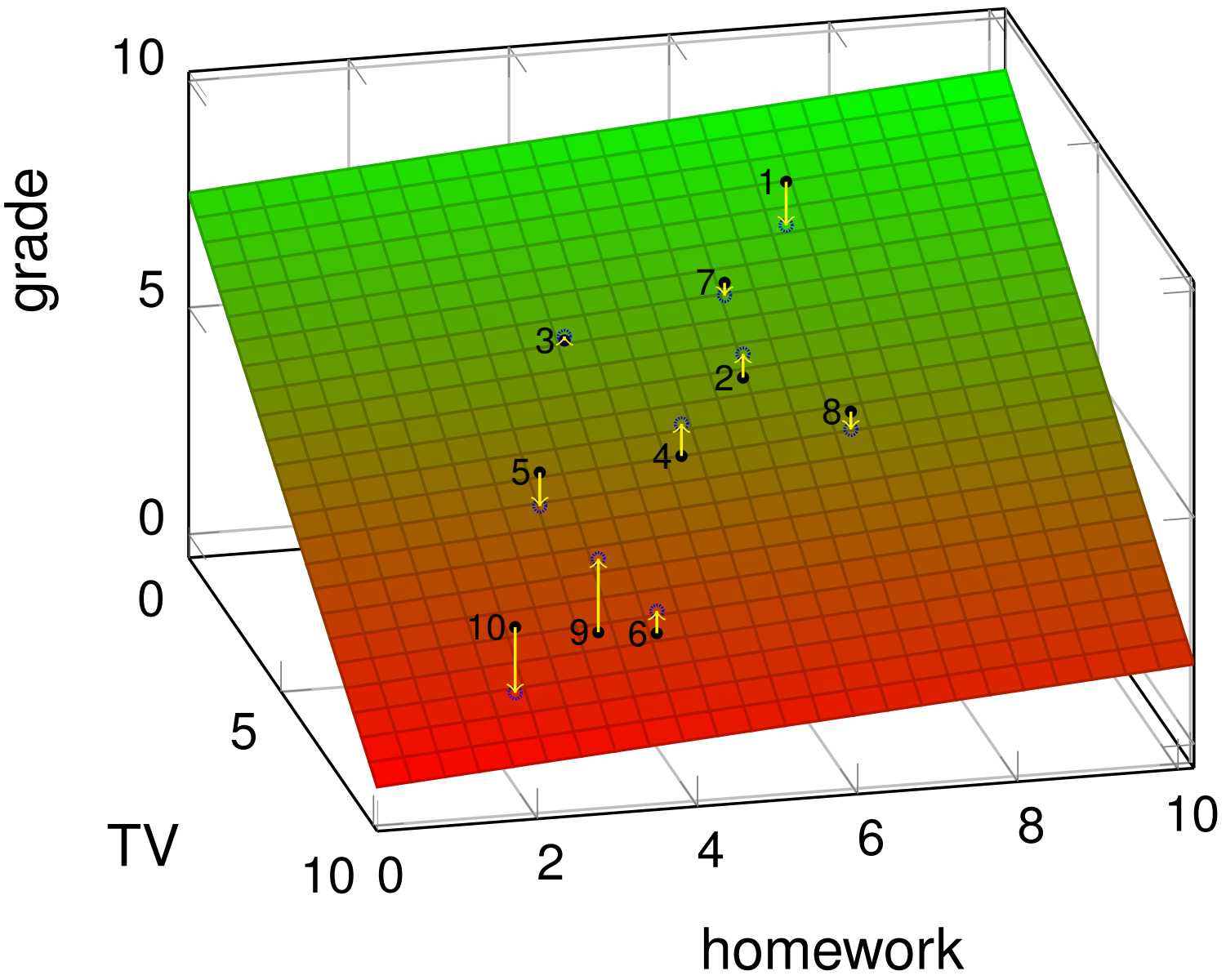
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



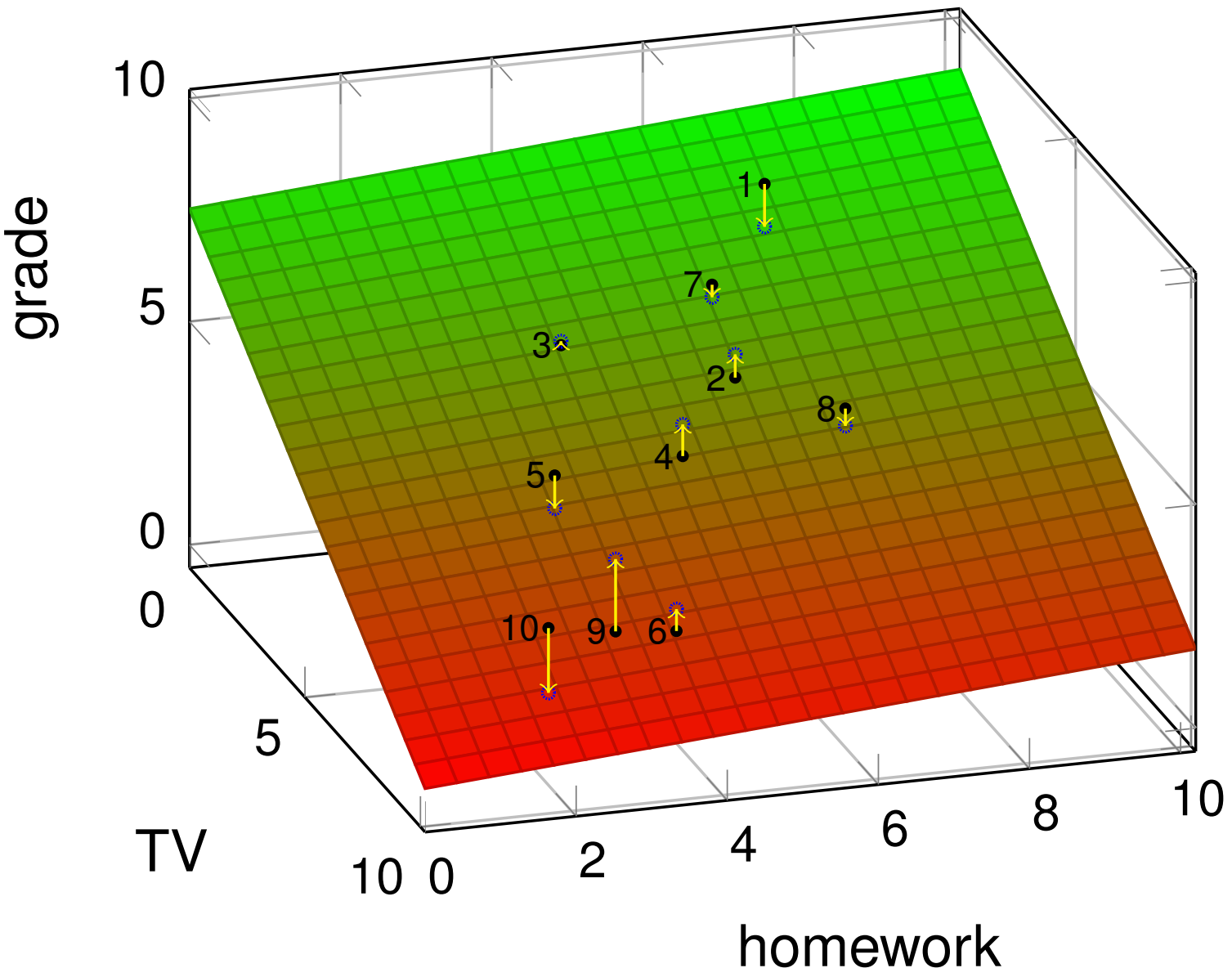
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



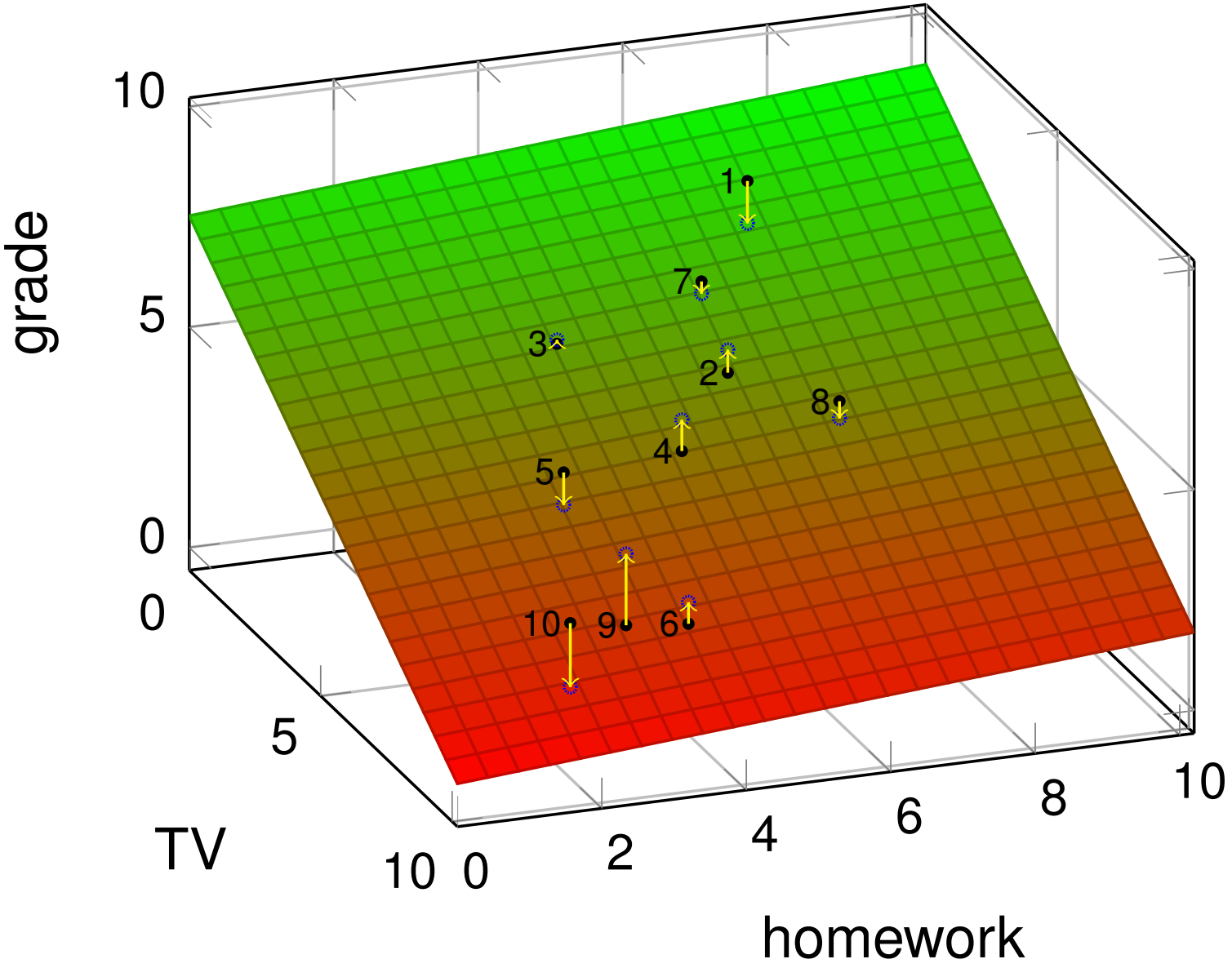
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



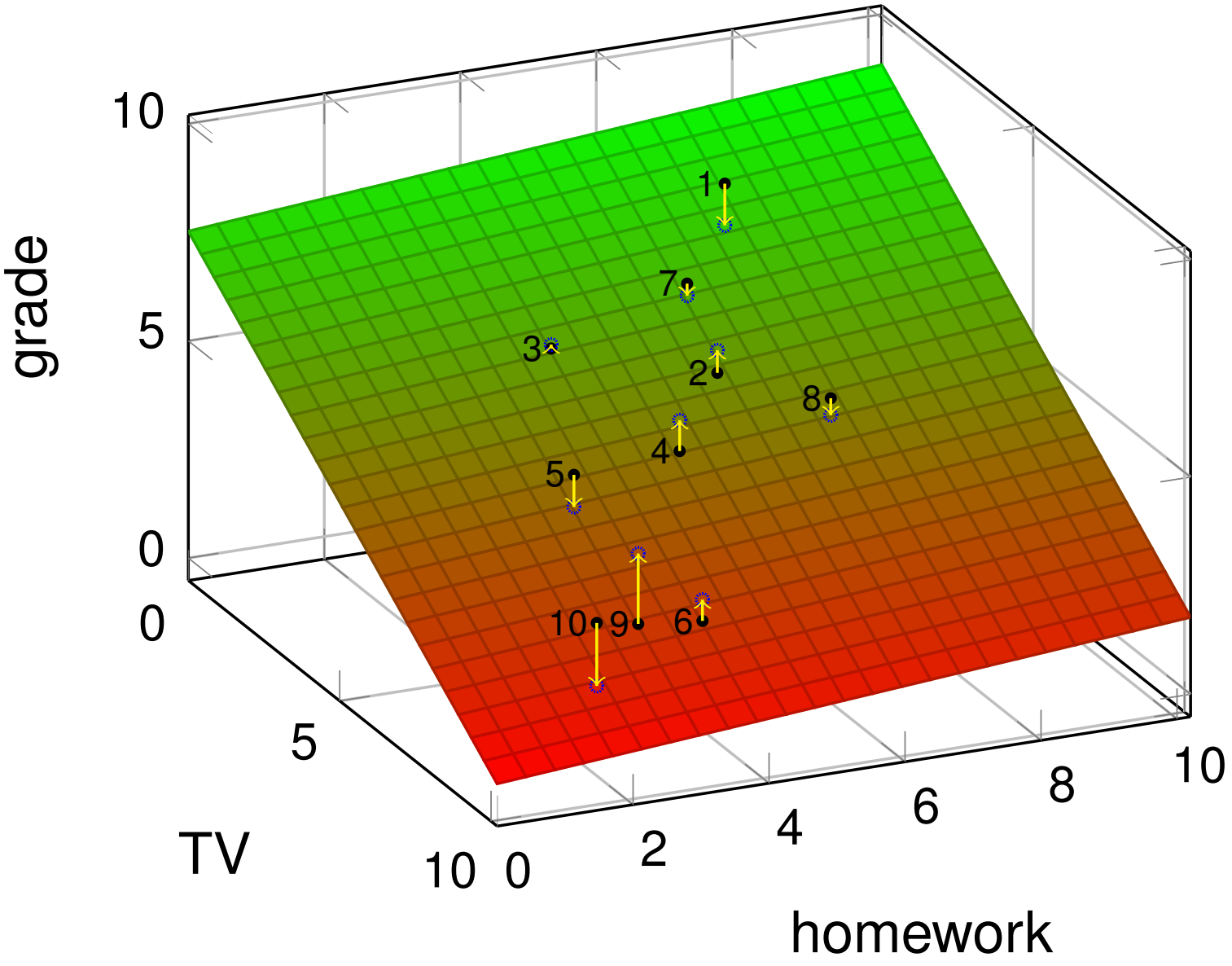
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



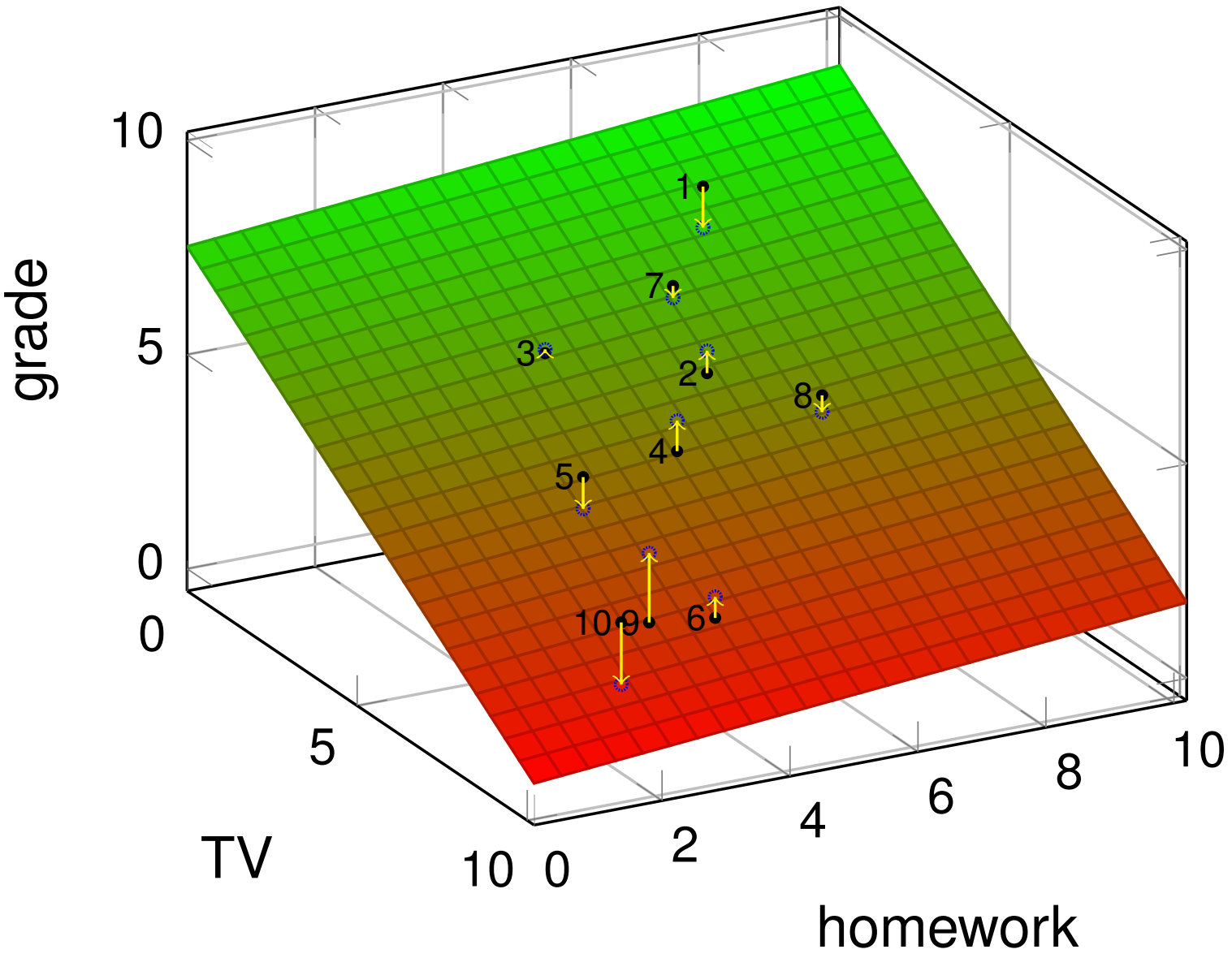
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



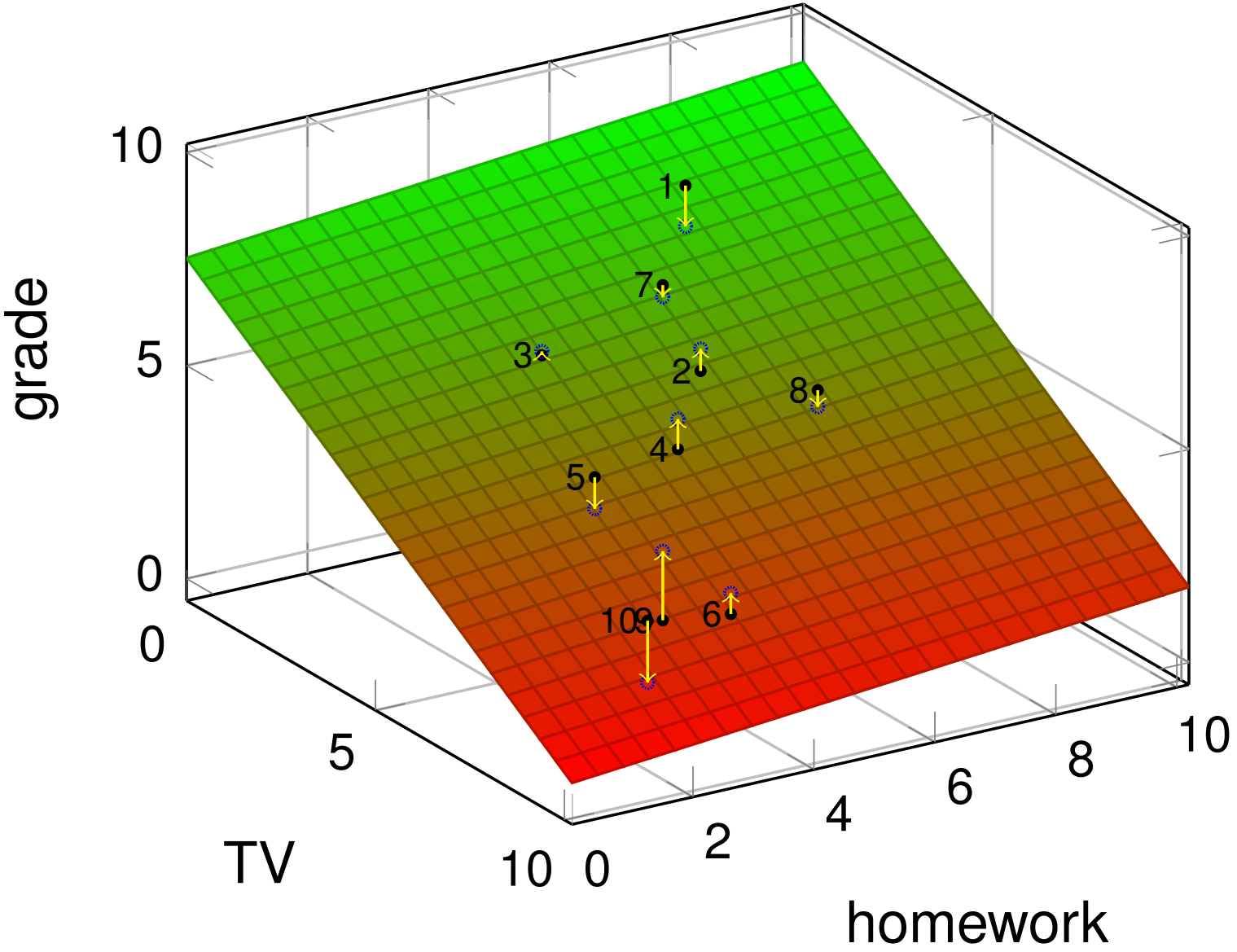
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



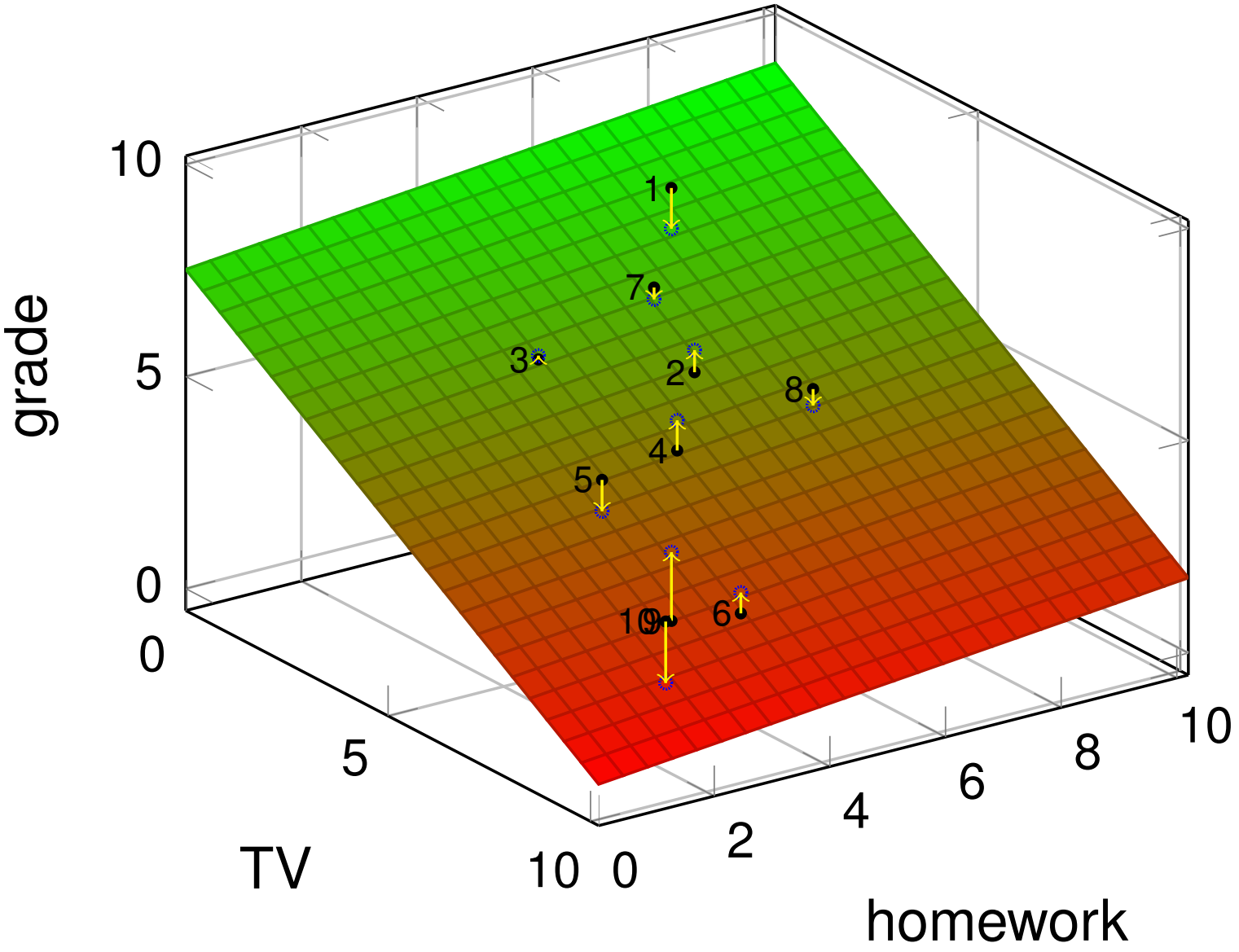
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



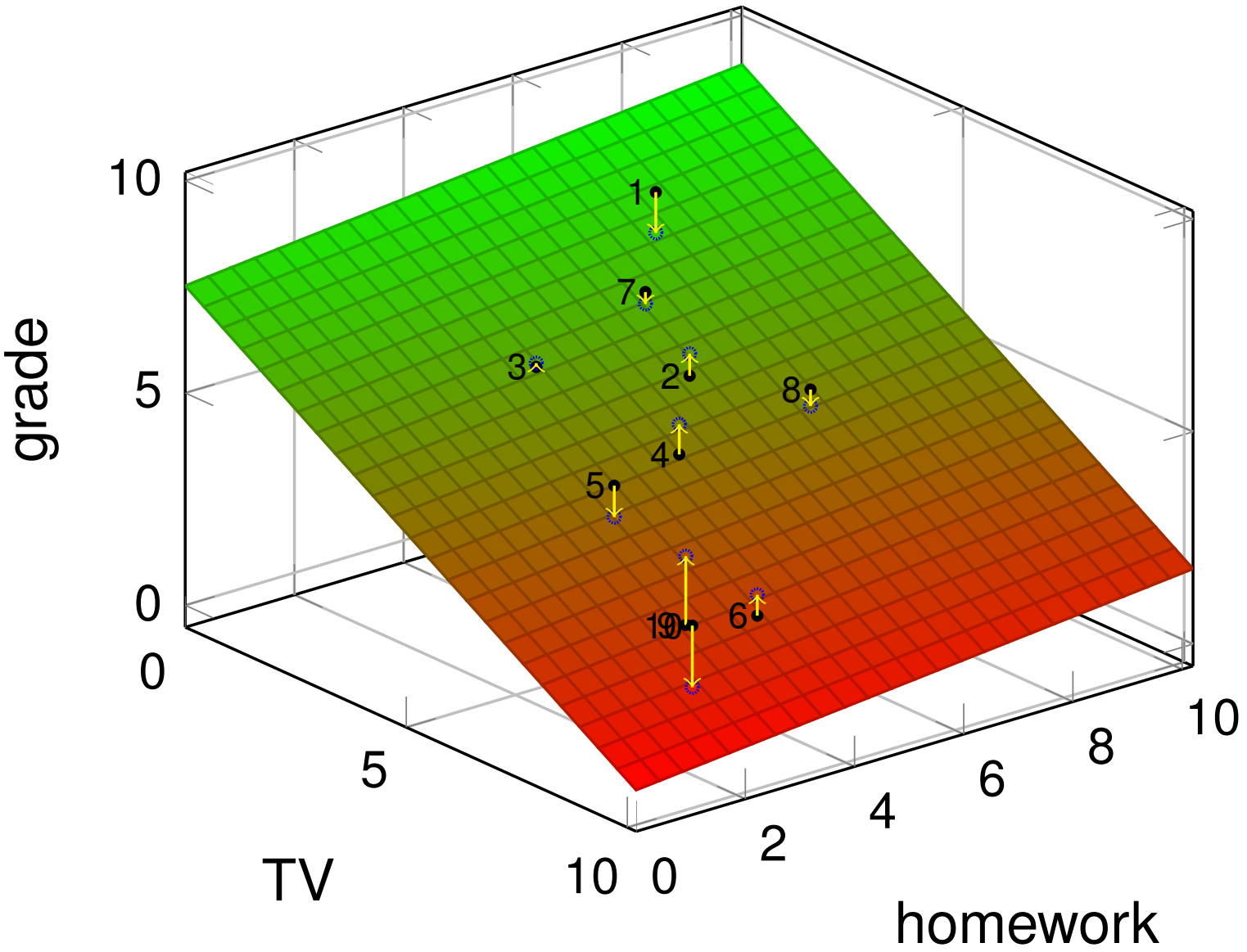
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



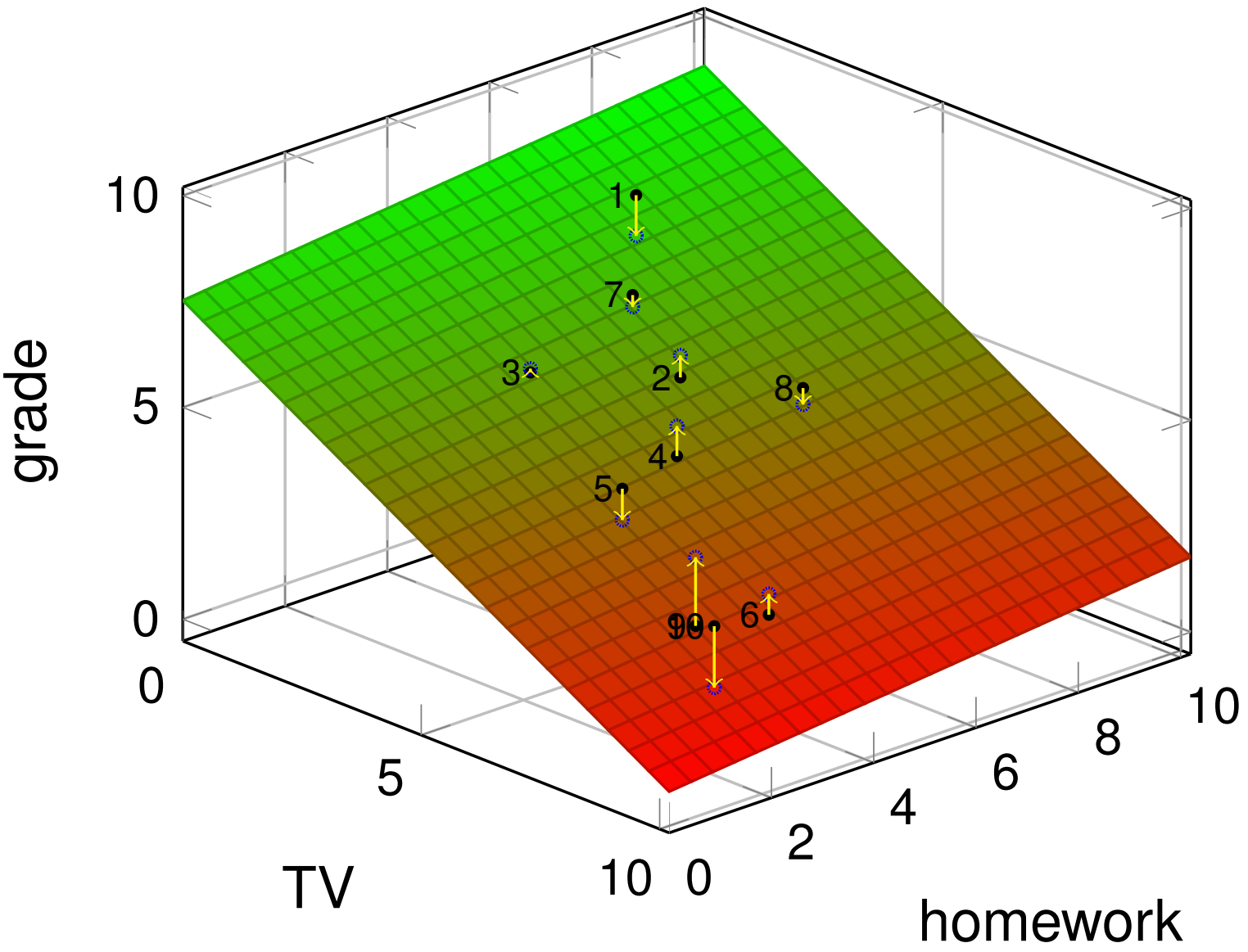
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



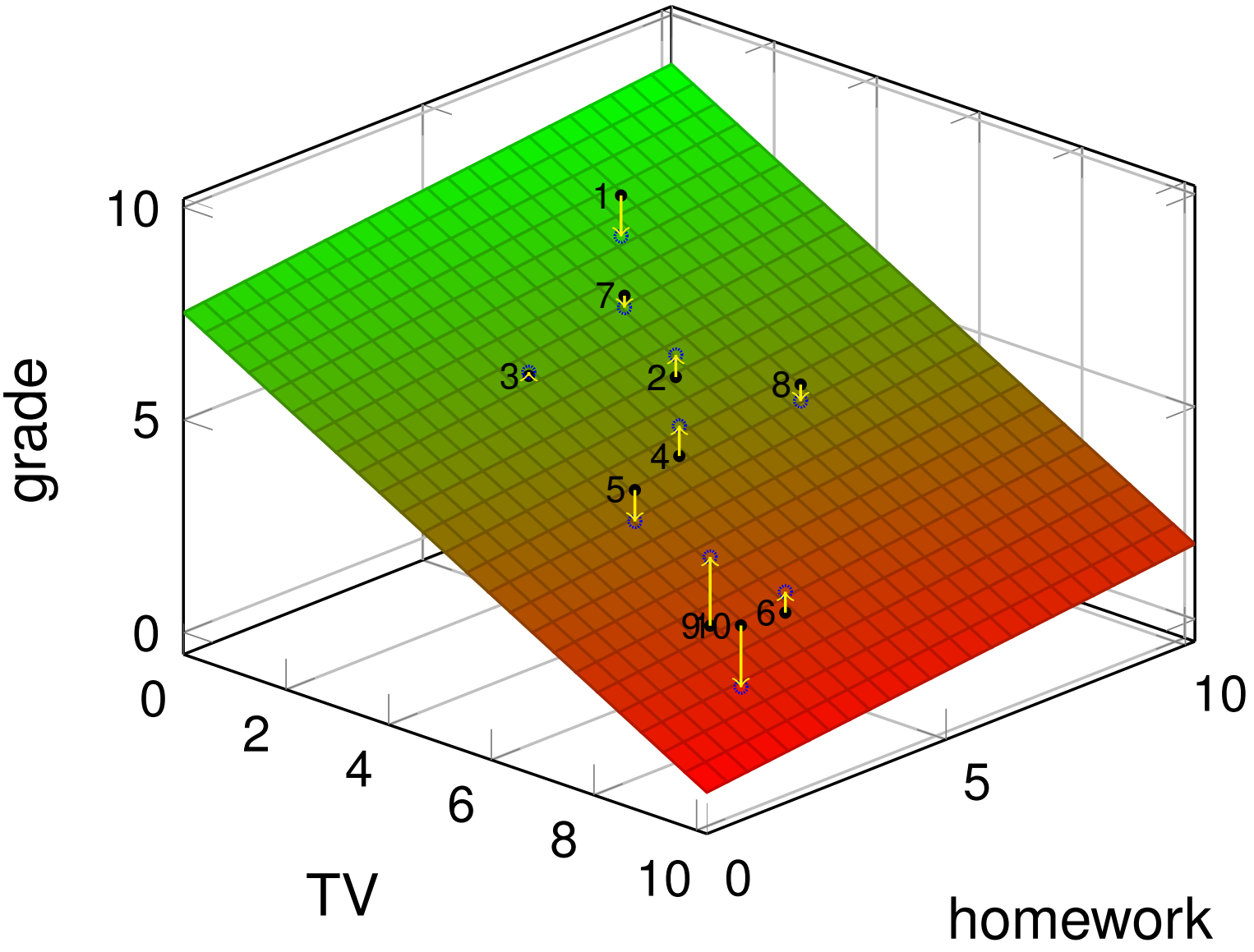
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



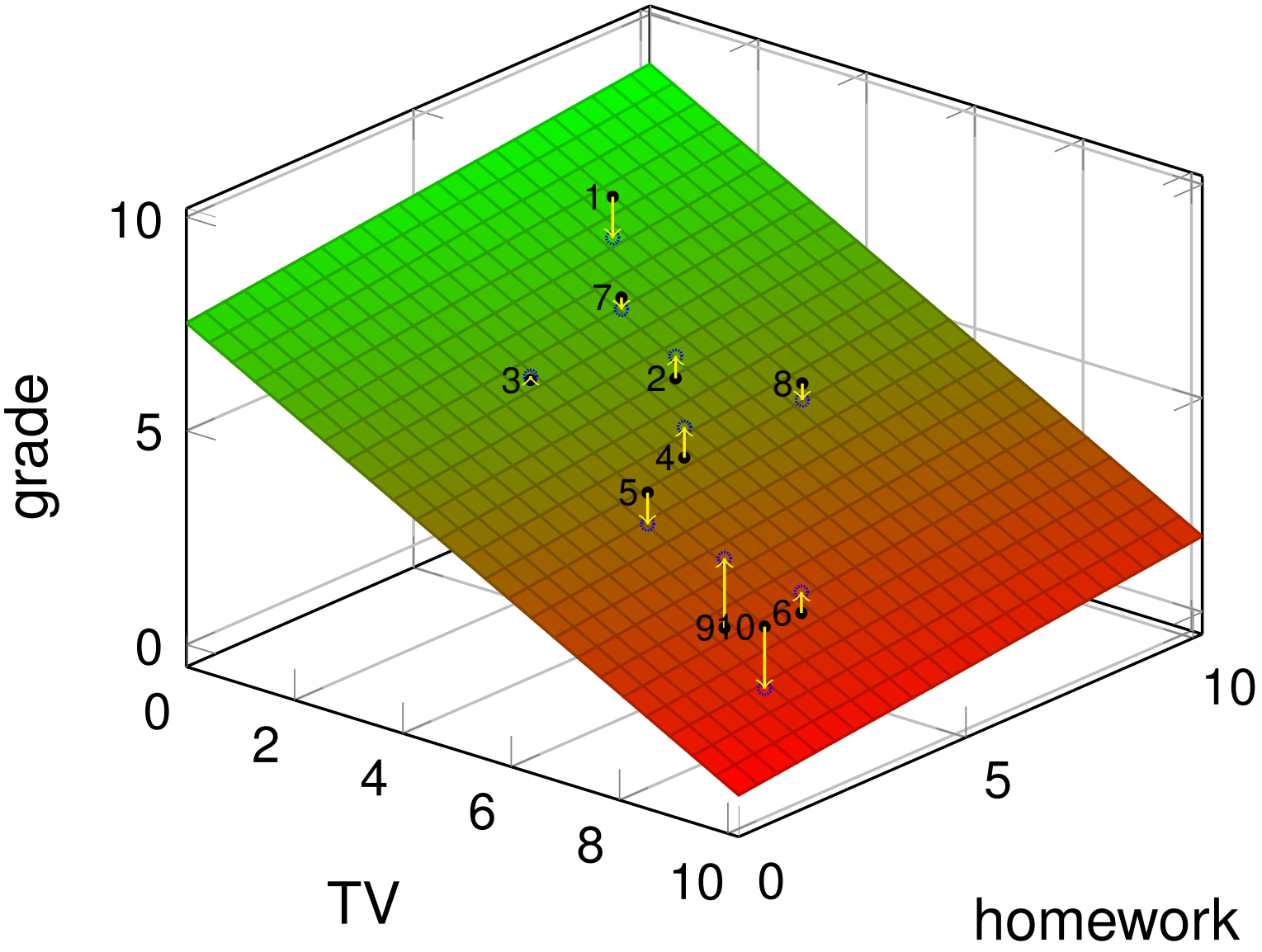
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



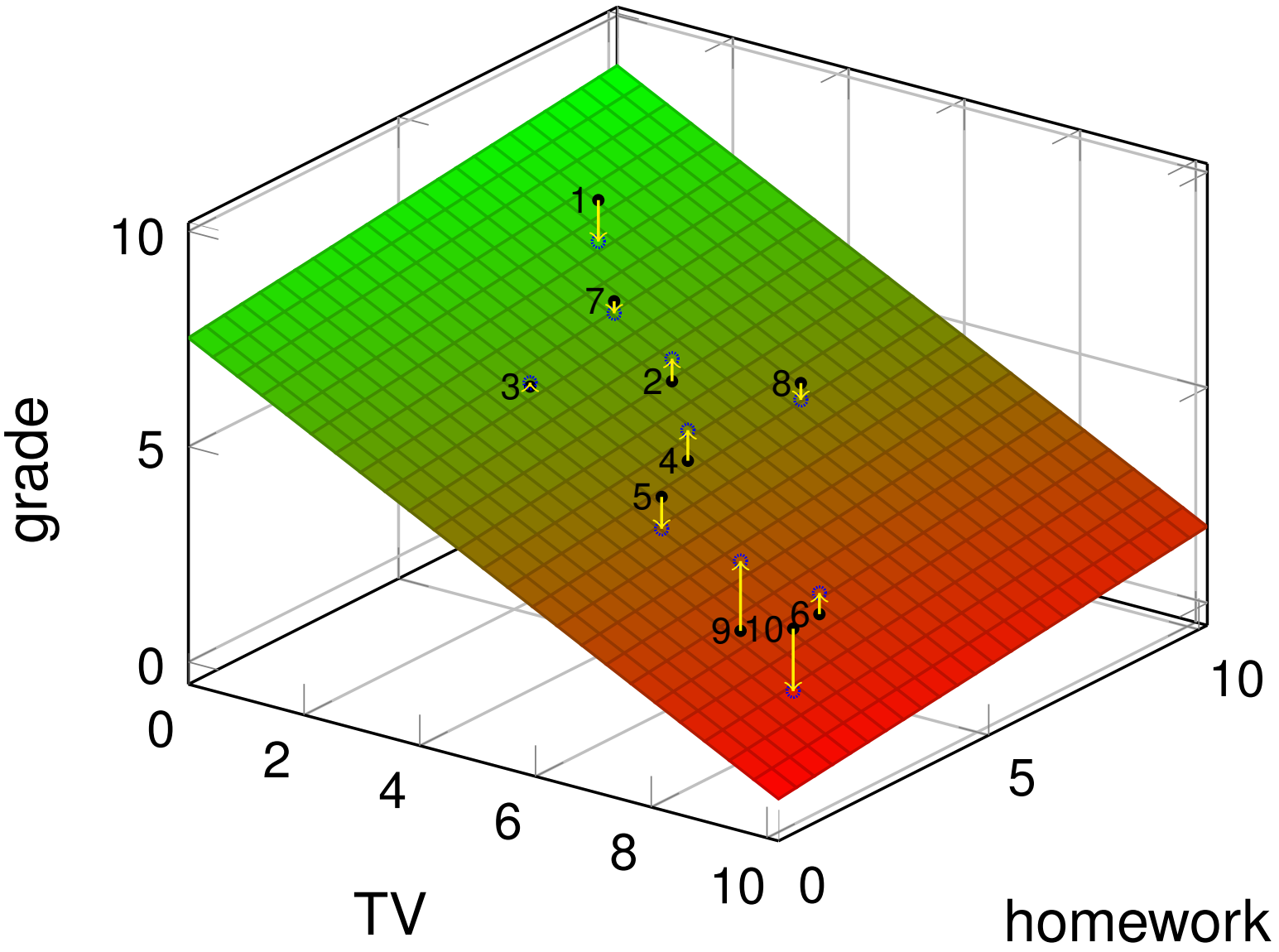
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



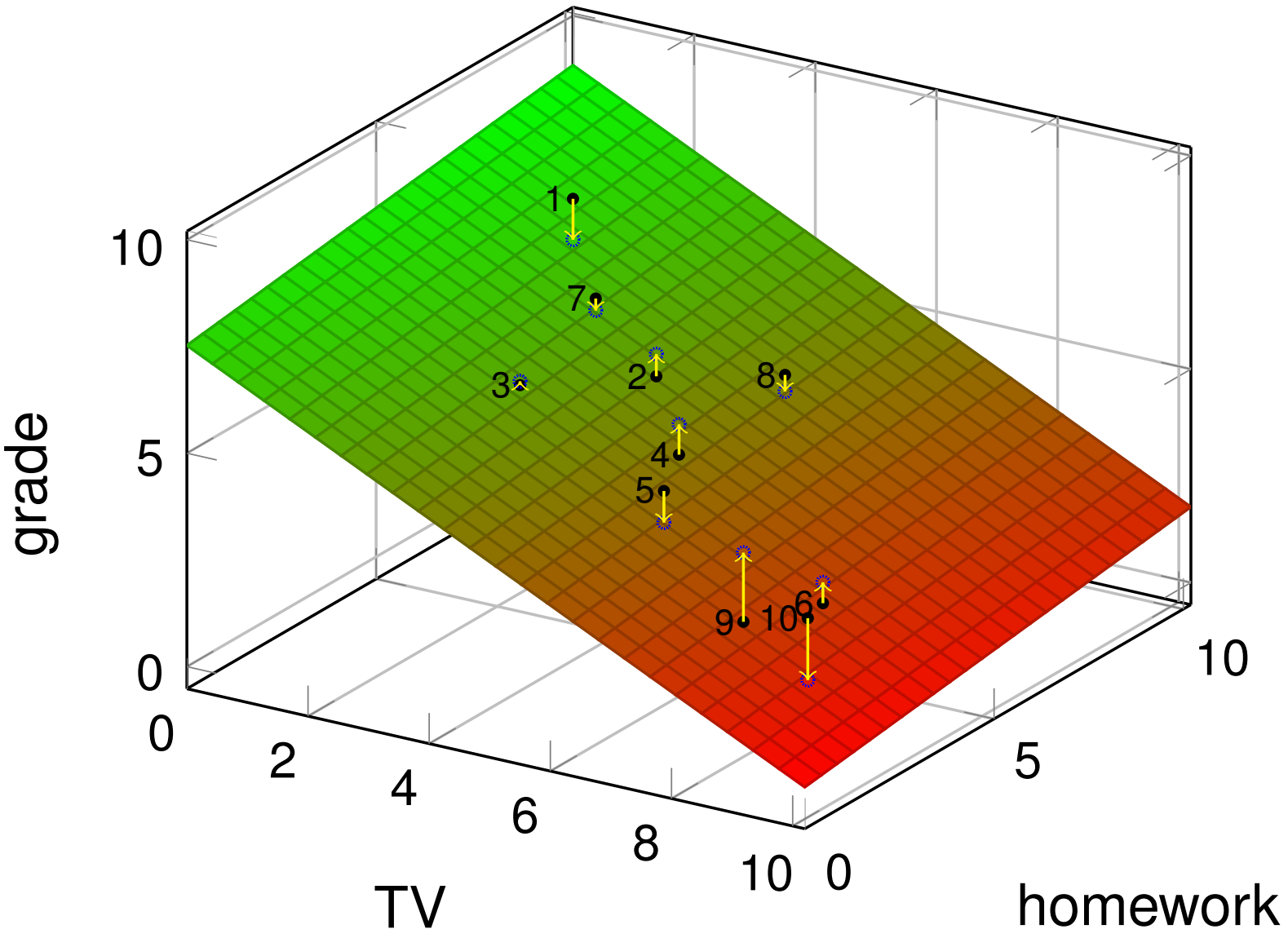
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



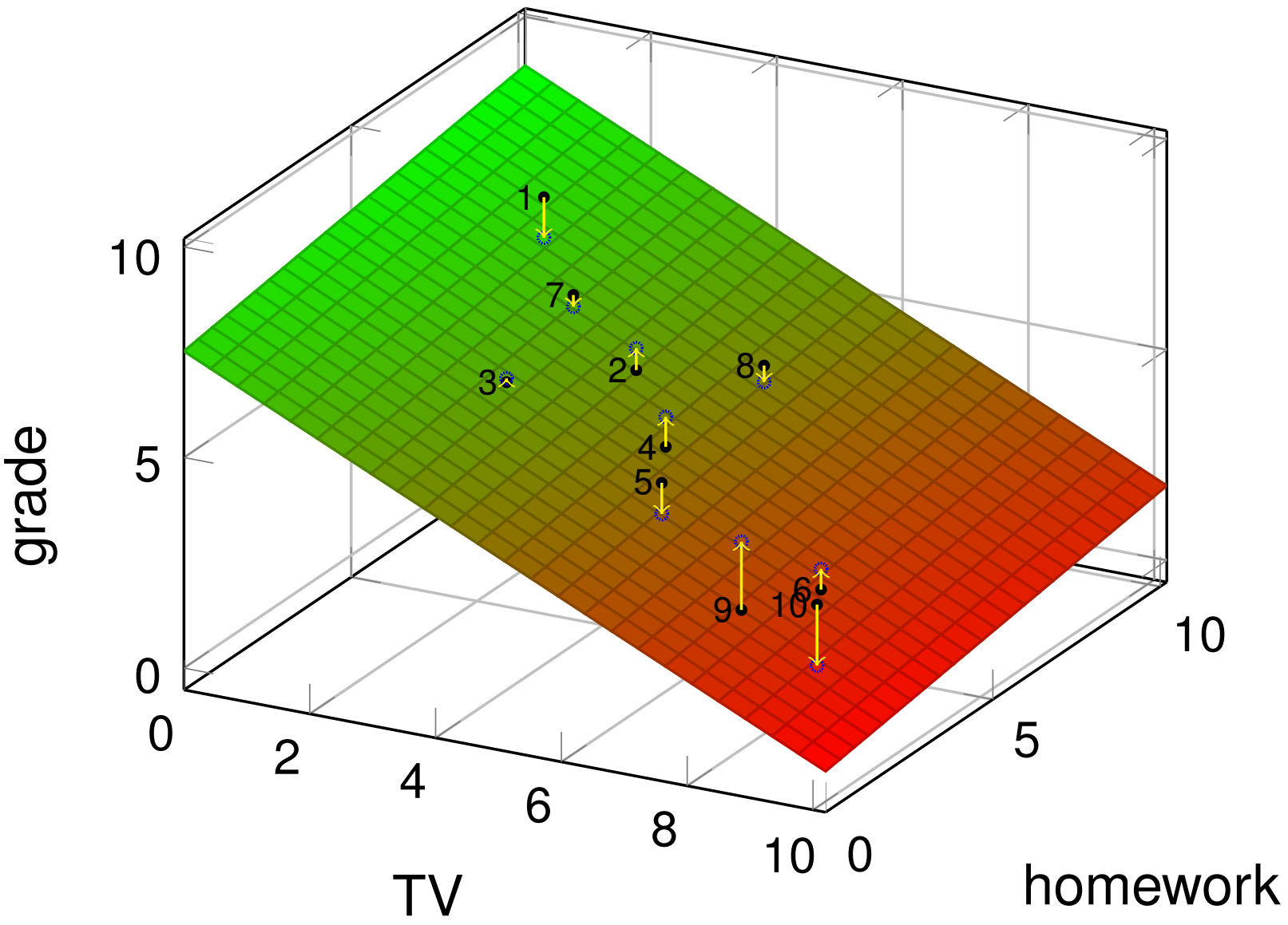
$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$



$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$

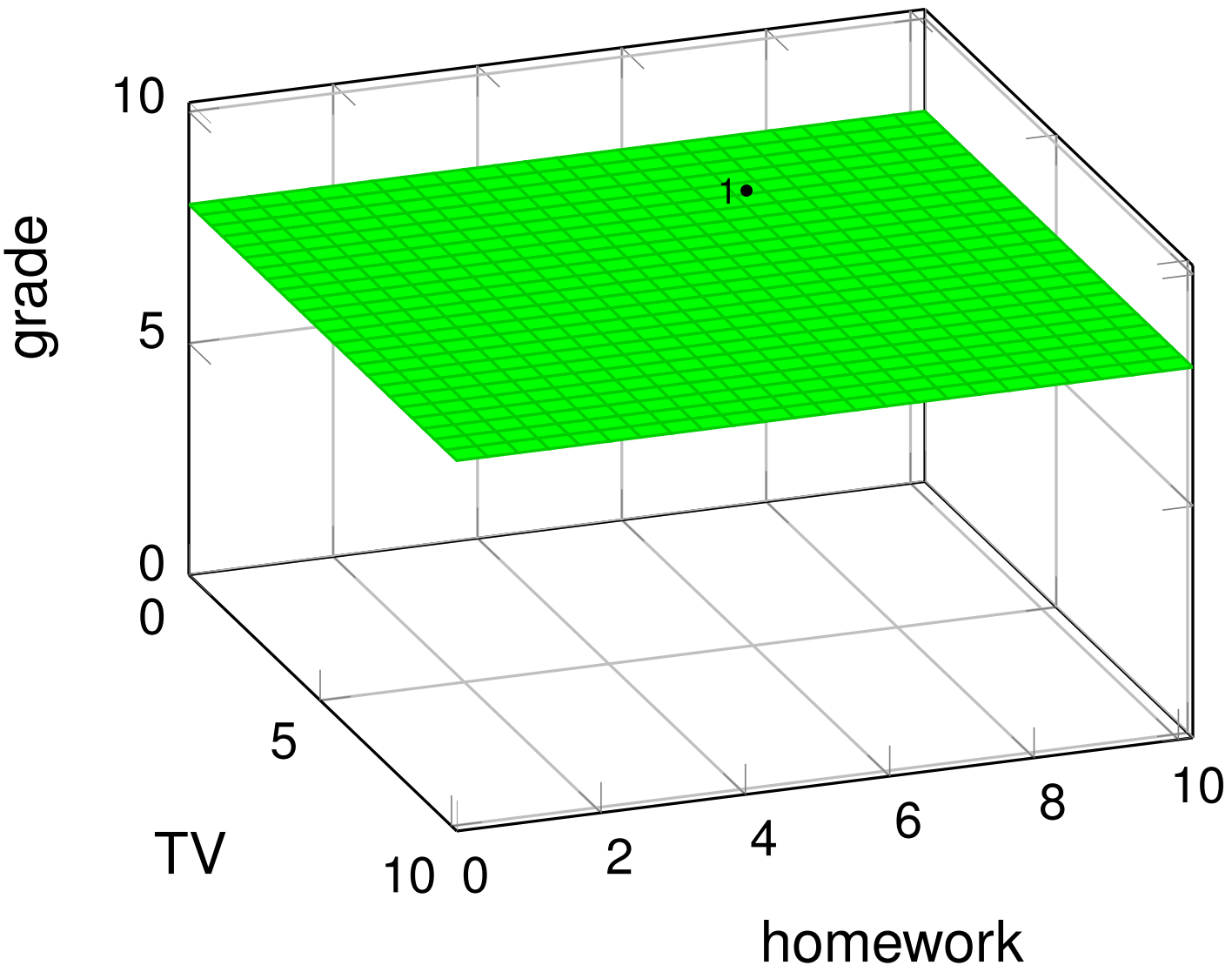


$$y = 7.525 - 0.693 \cdot x_1 + 0.129 \cdot x_2$$

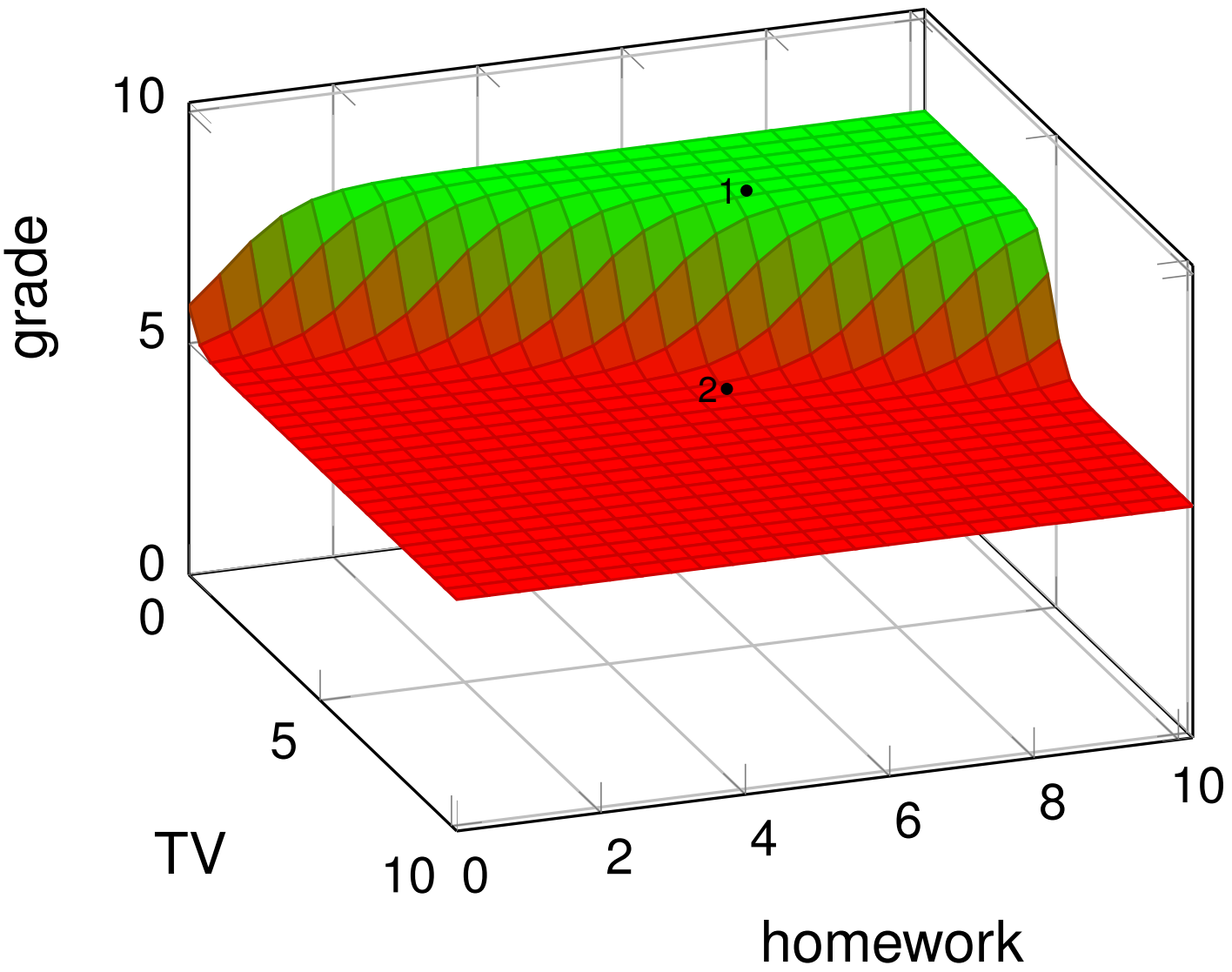


Let us now use k-NN with Gaussian Weighting

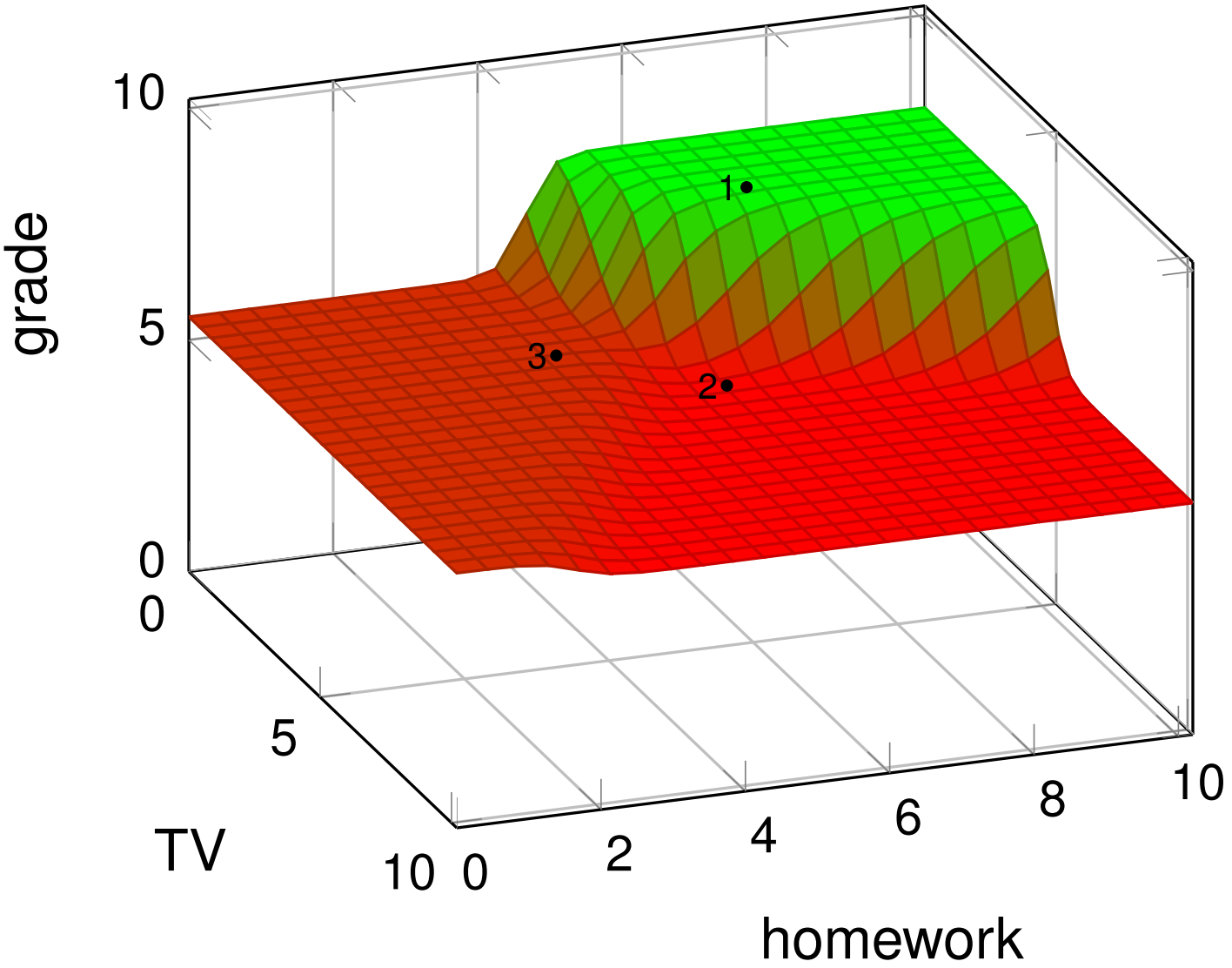
$$y(x) = \frac{\sum_{i=1}^{1} y_i \cdot \exp(-\|x - x_i\|^2)}{\sum_{i=1}^{1} \exp(-\|x - x_i\|^2)}$$



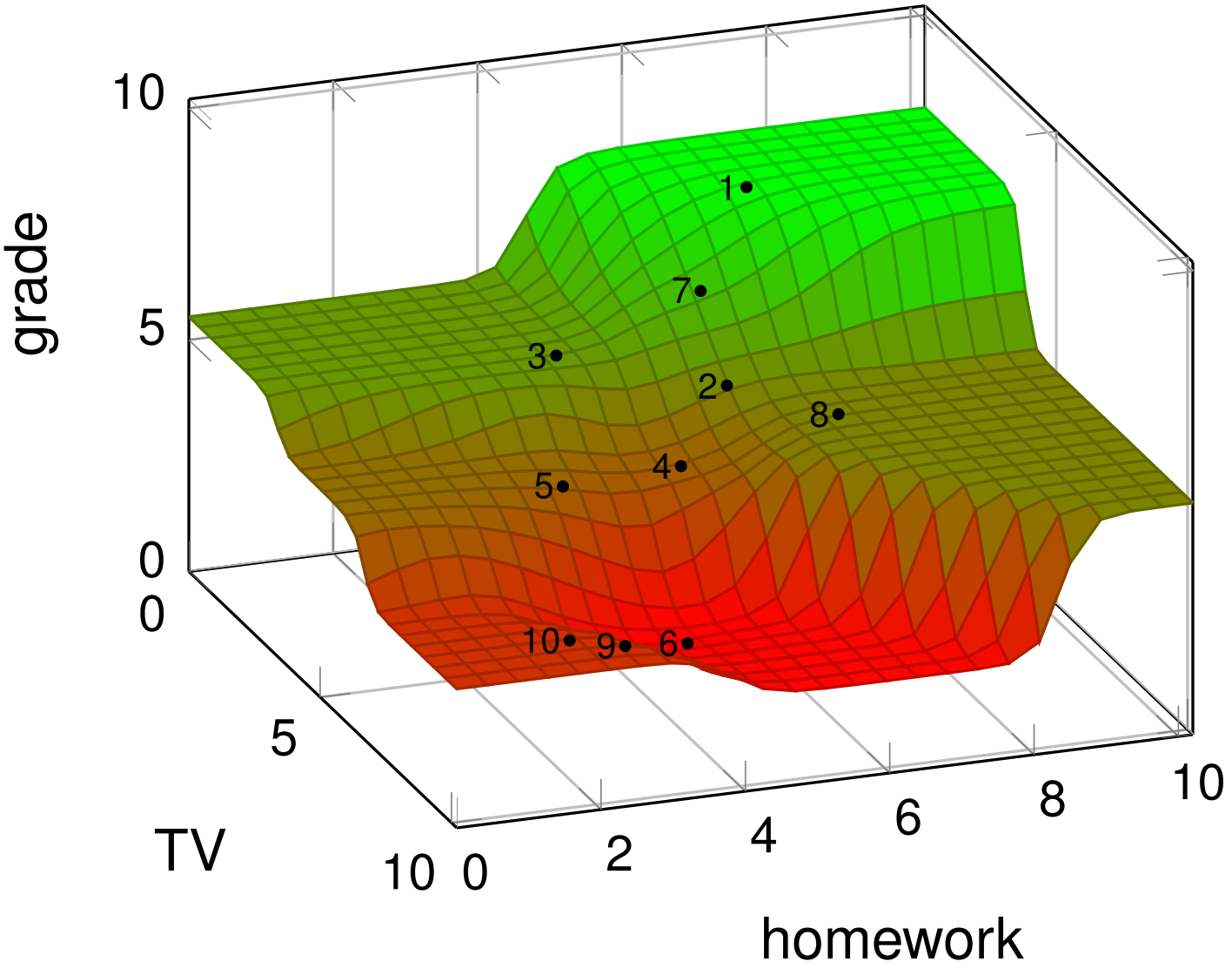
$$y(x) = \frac{\sum_{i=1}^{2} y_i \cdot \exp(-\|x - x_i\|^2)}{\sum_{i=1}^{2} \exp(-\|x - x_i\|^2)}$$



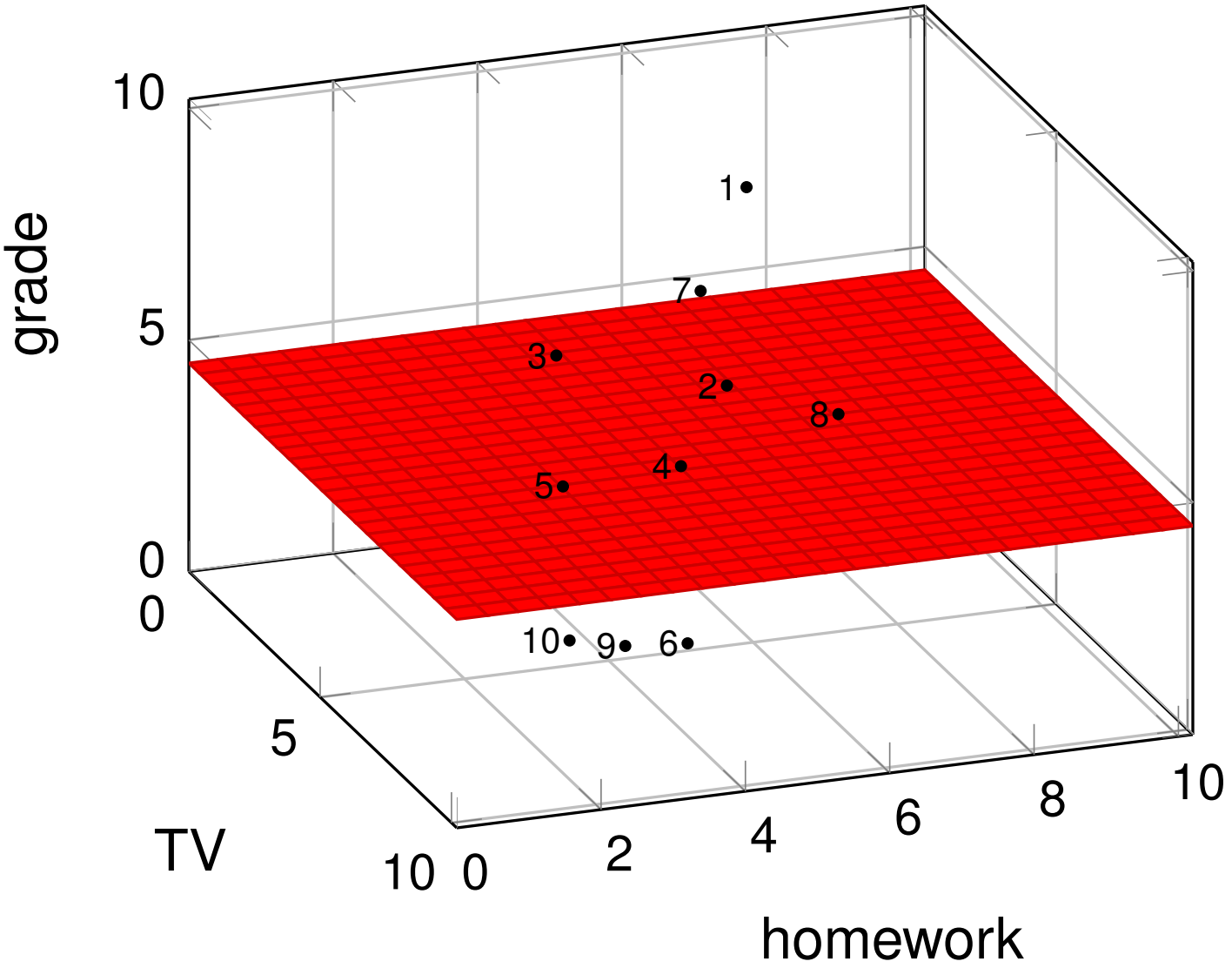
$$y(x) = \frac{\sum_{i=1}^{3} y_i \cdot \exp(-\|x - x_i\|^2)}{\sum_{i=1}^{3} \exp(-\|x - x_i\|^2)}$$



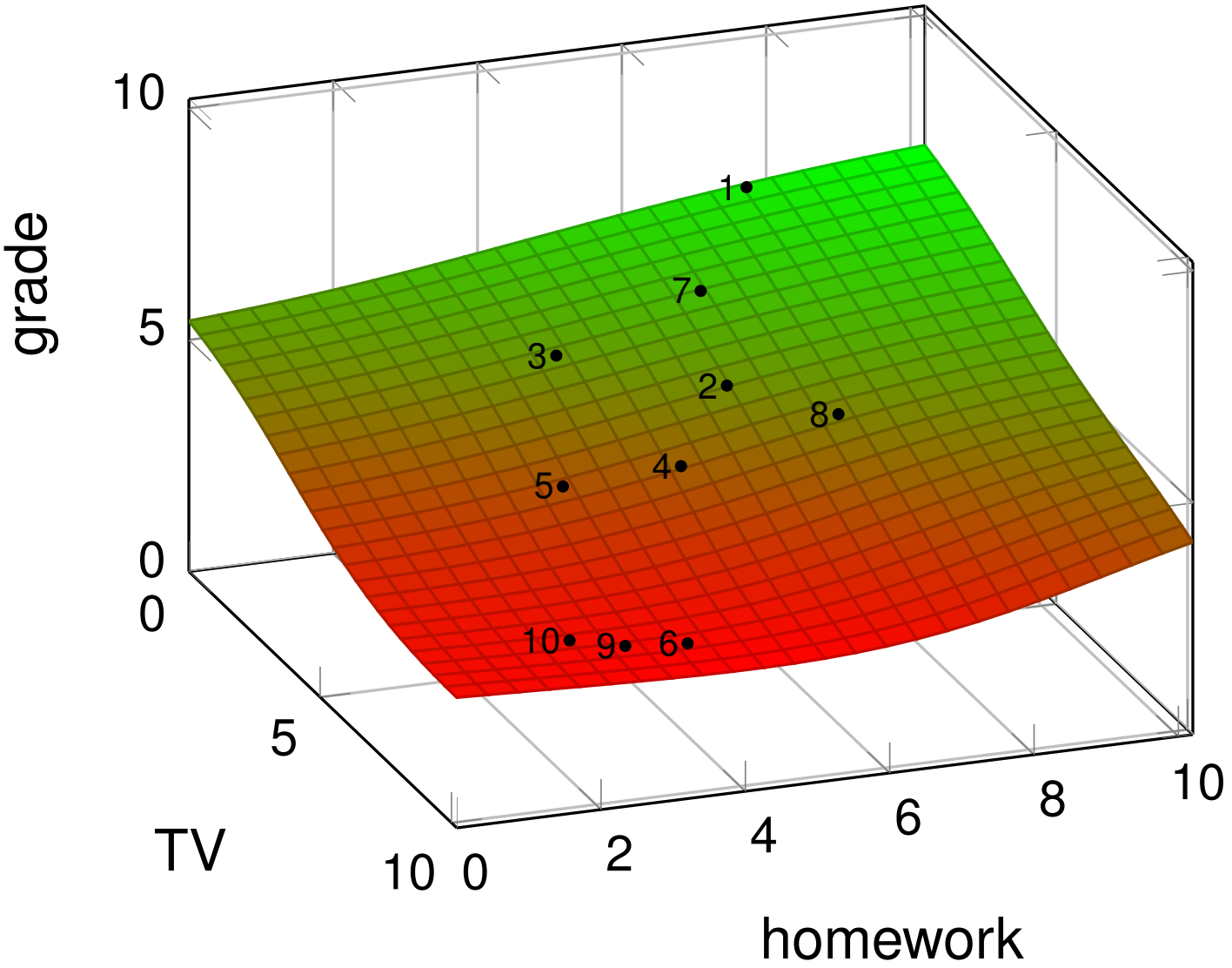
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-\|x - x_i\|^2)}{\sum_{i=1}^{10} \exp(-\|x - x_i\|^2)}$$



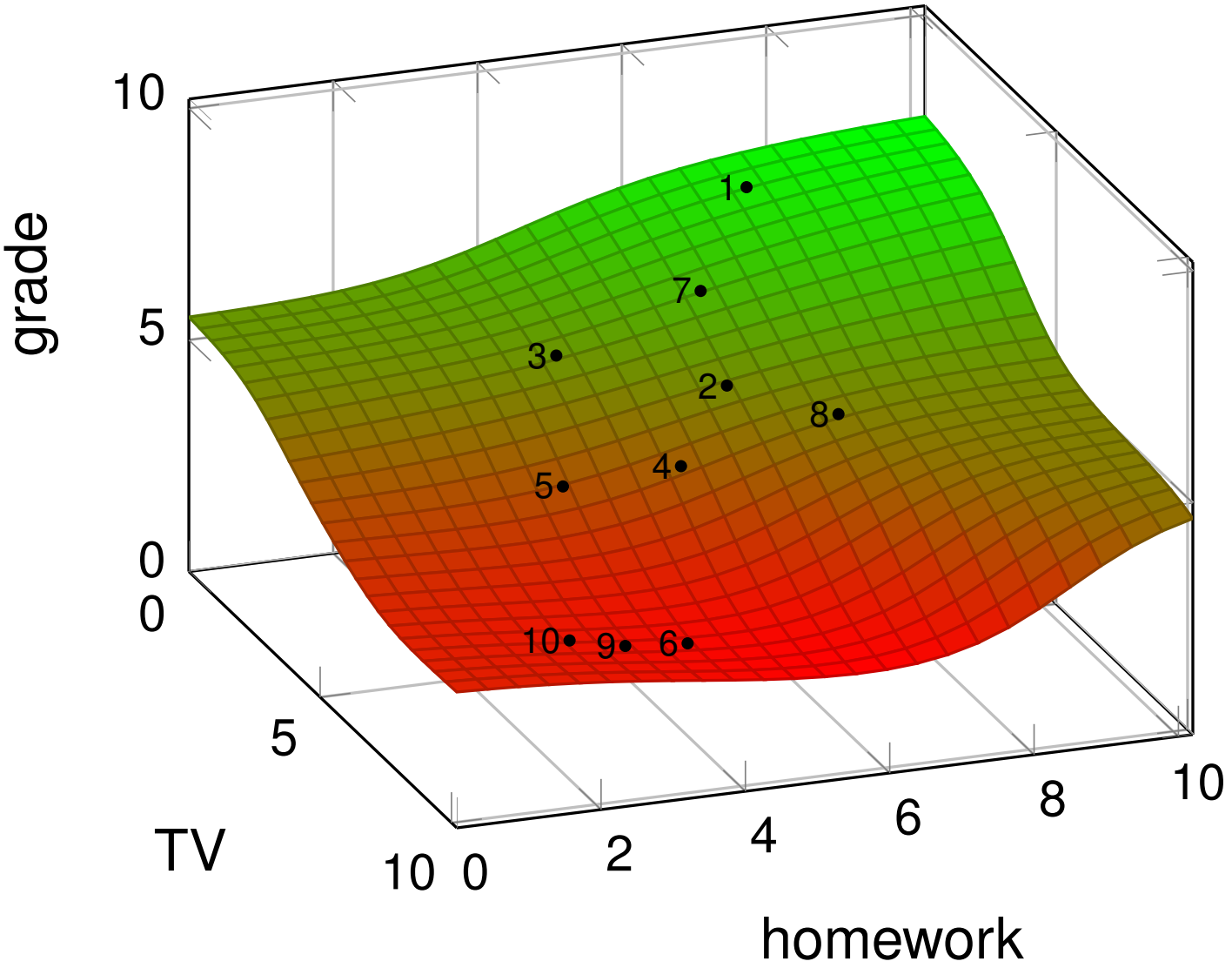
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 0.0$$



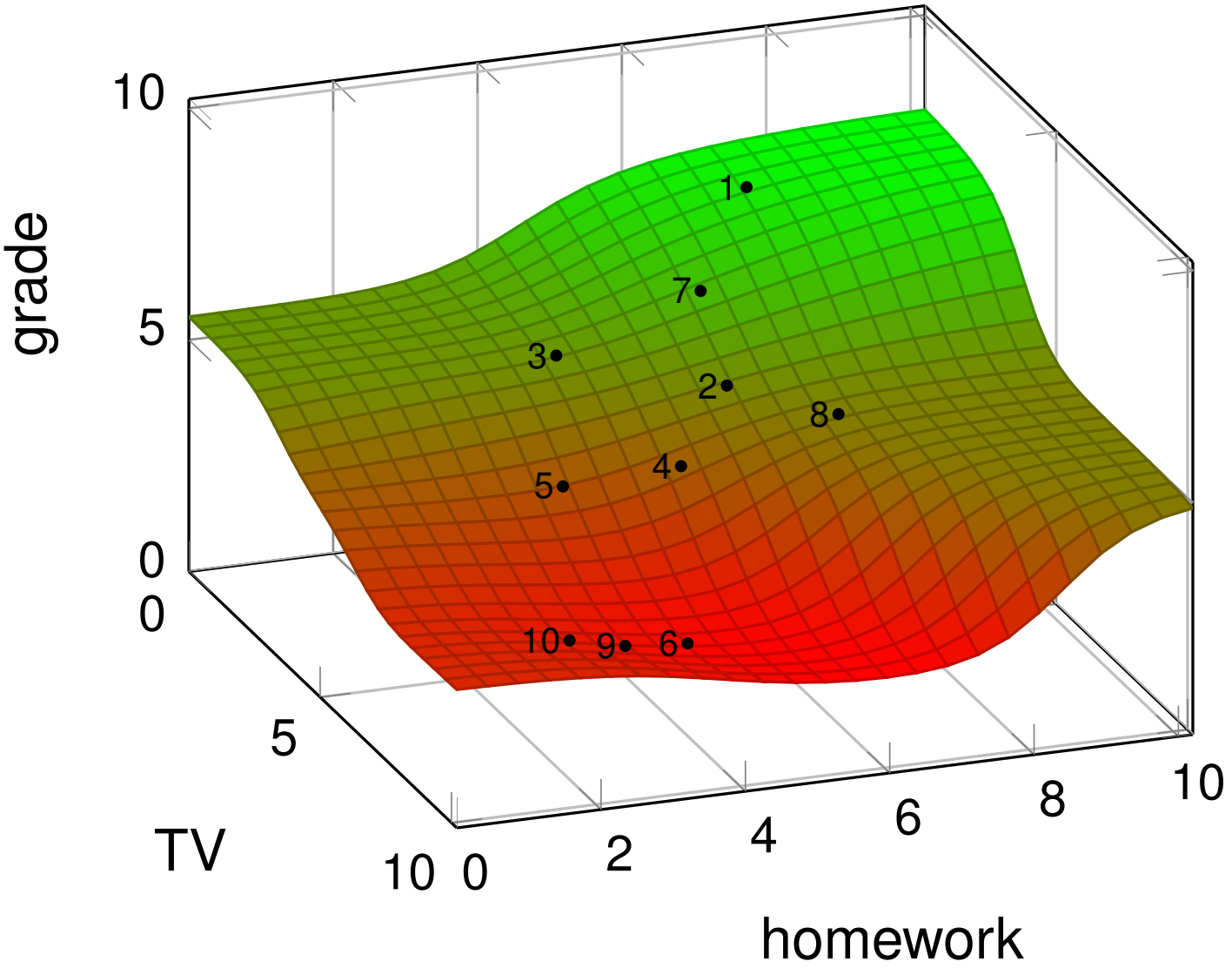
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 0.1$$



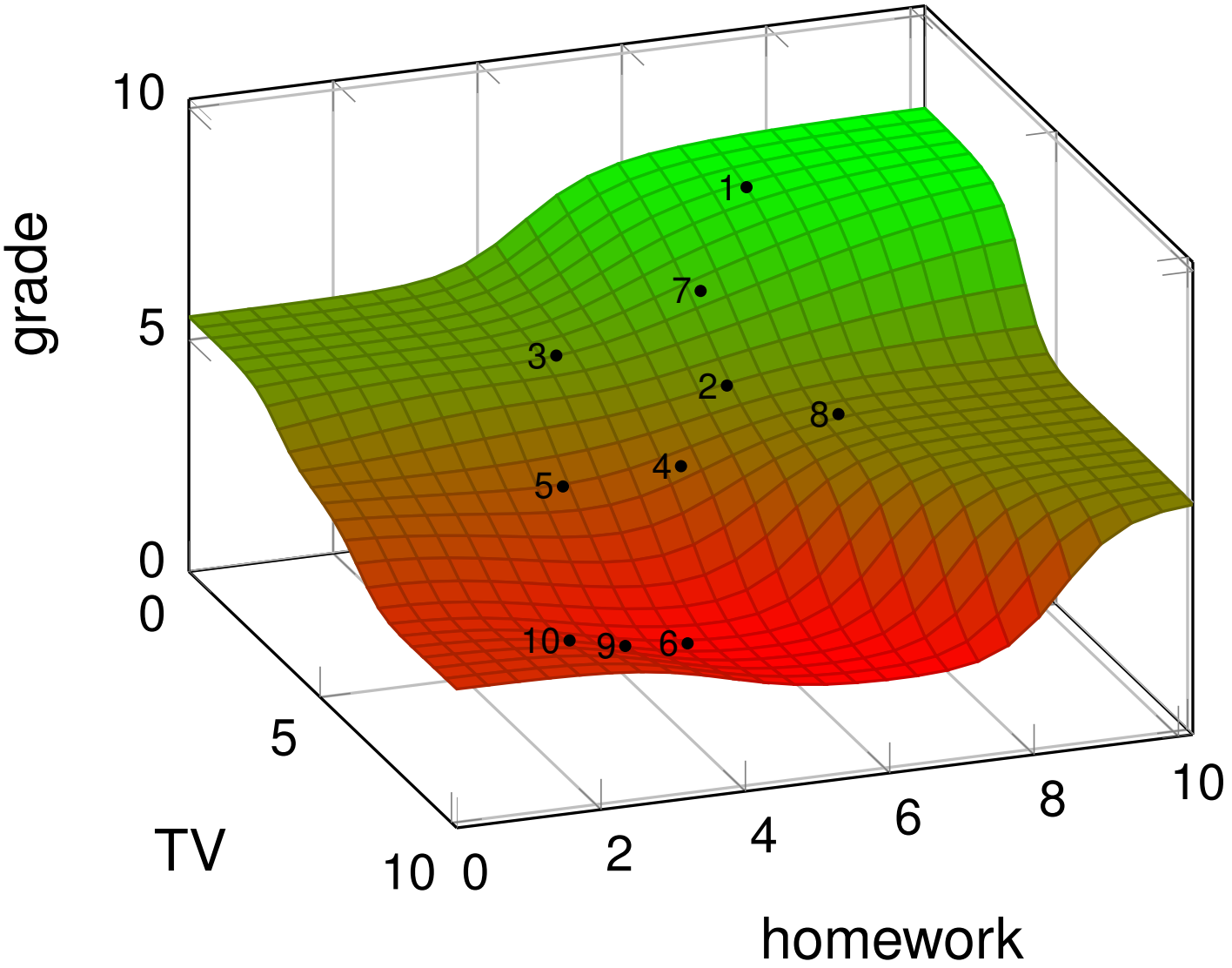
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 0.2$$



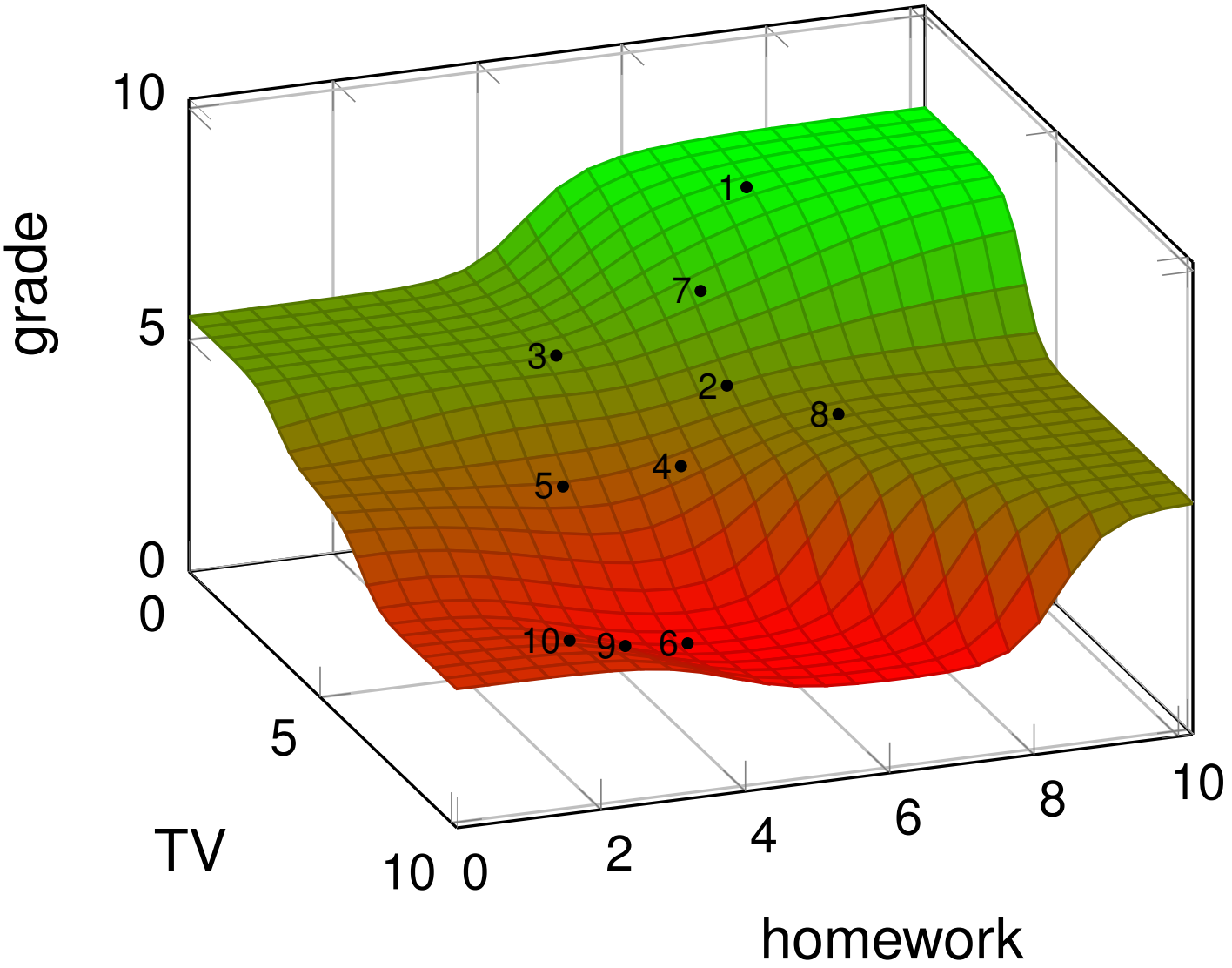
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 0.3$$



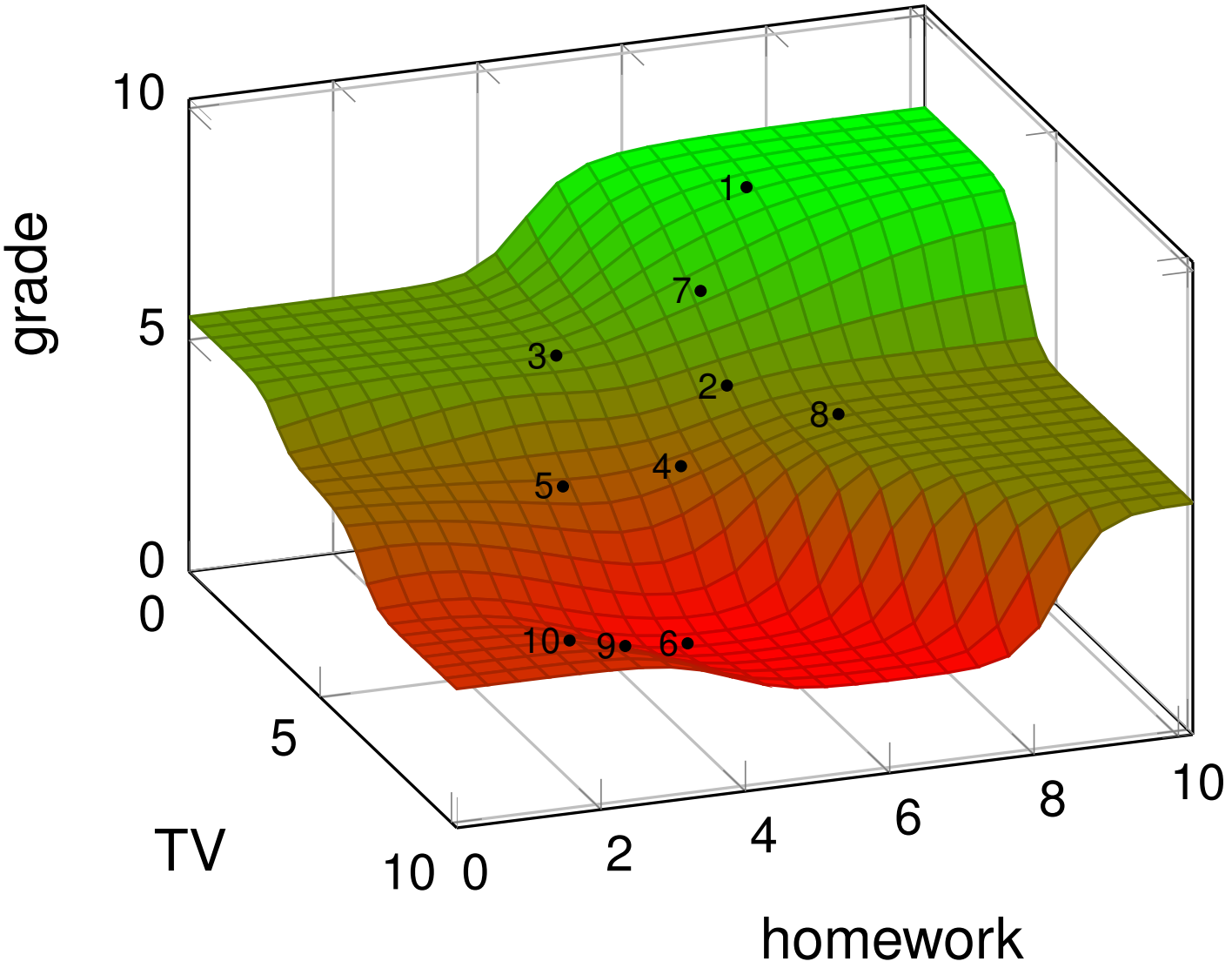
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 0.4$$



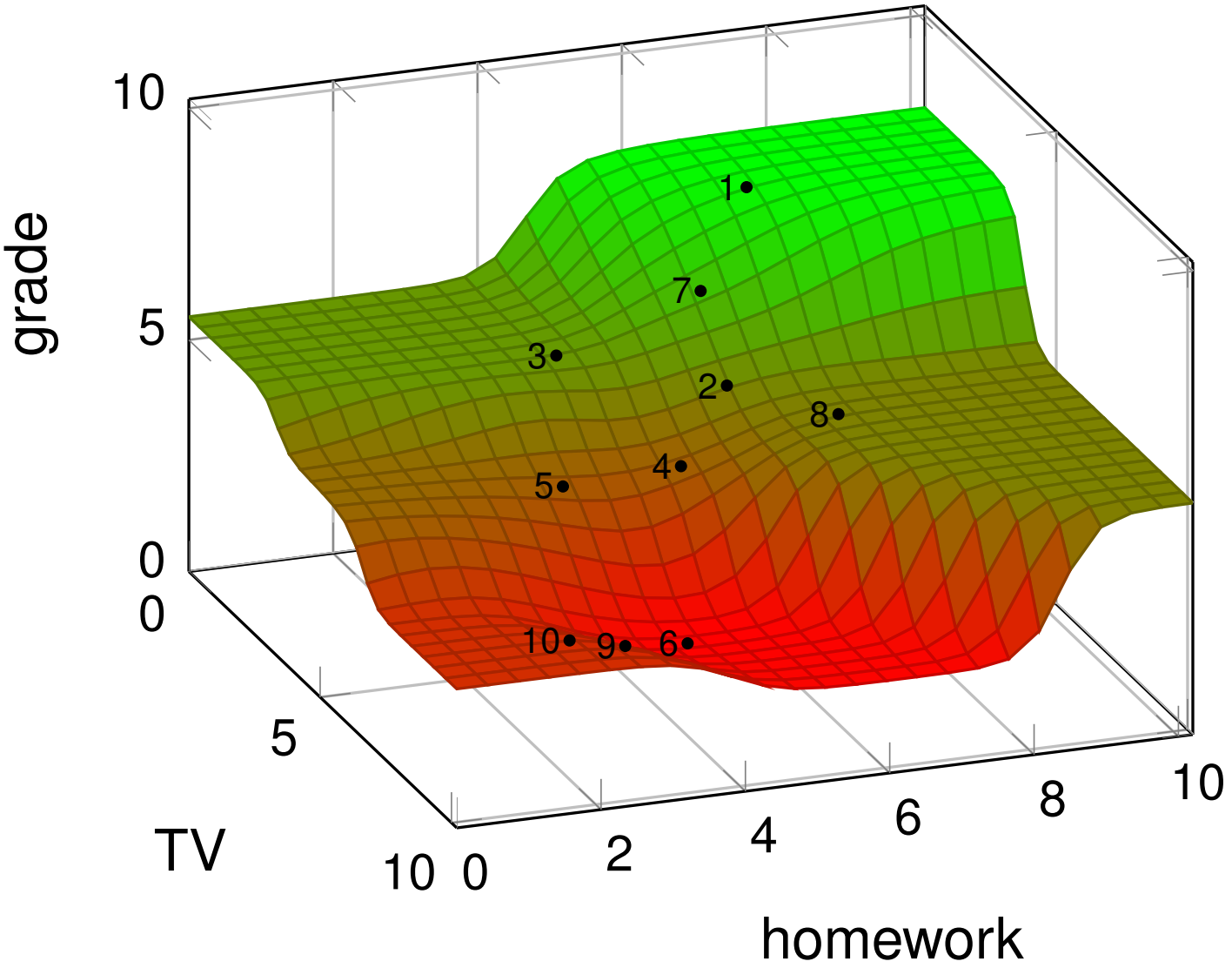
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 0.5$$



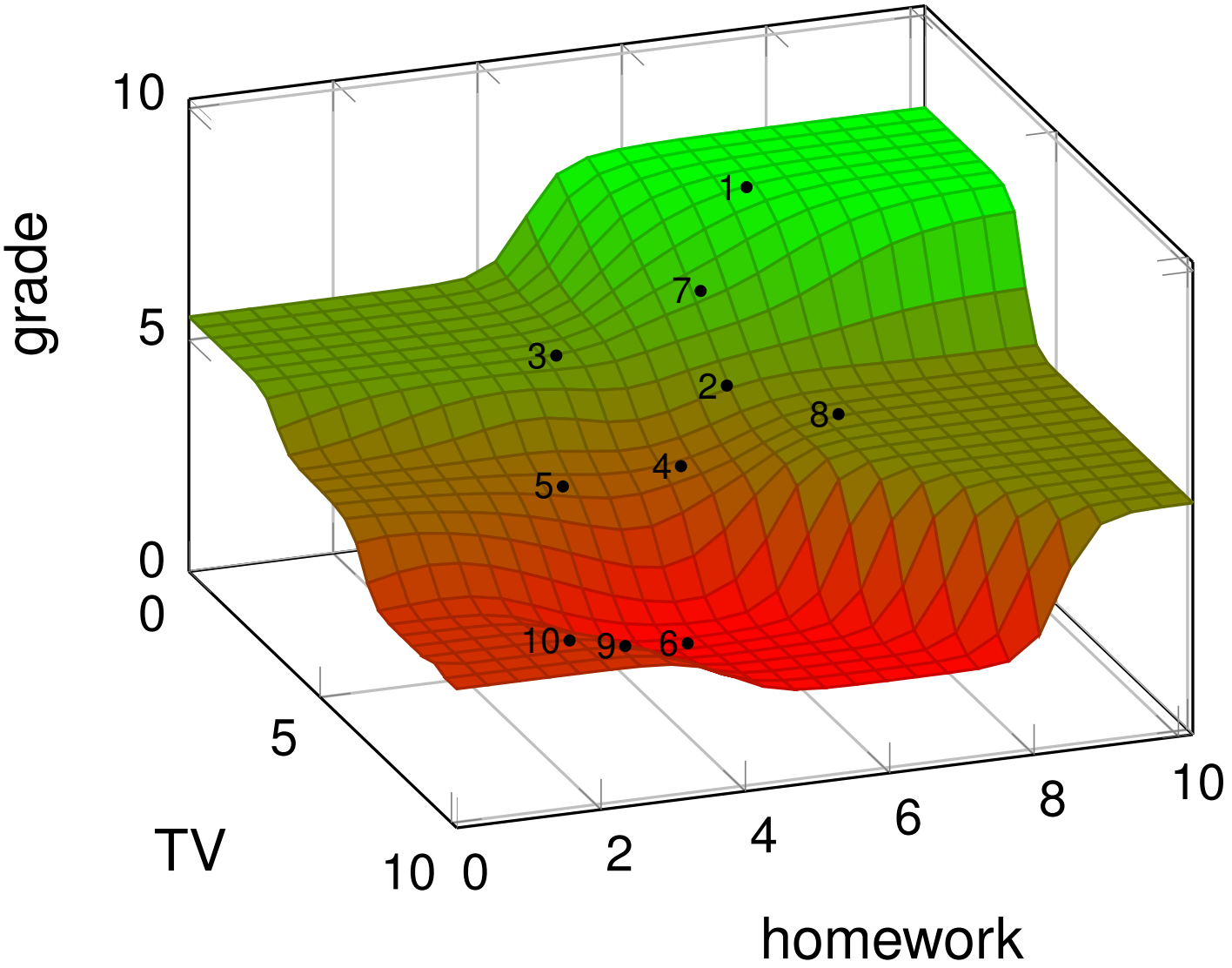
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 0.6$$



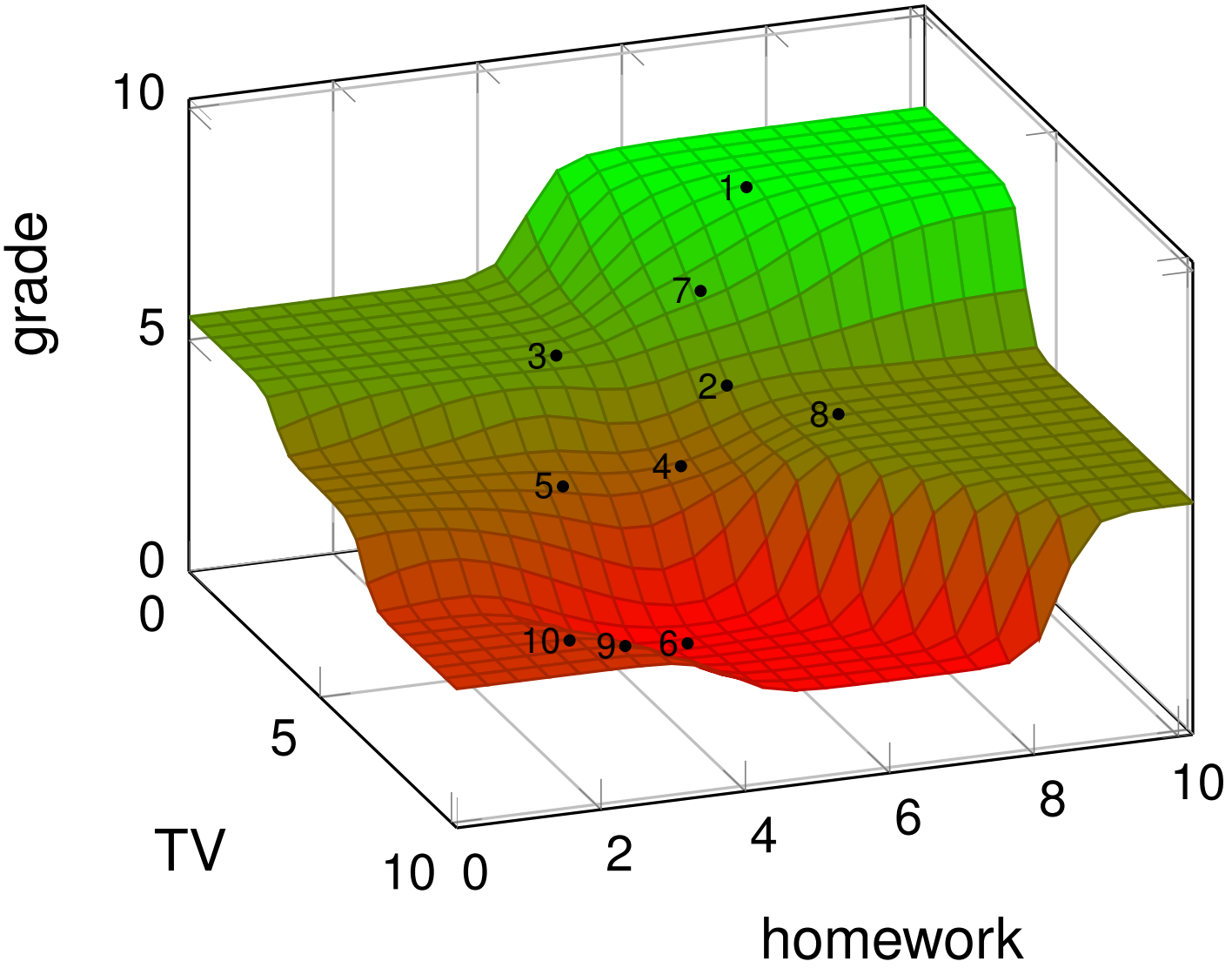
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 0.7$$



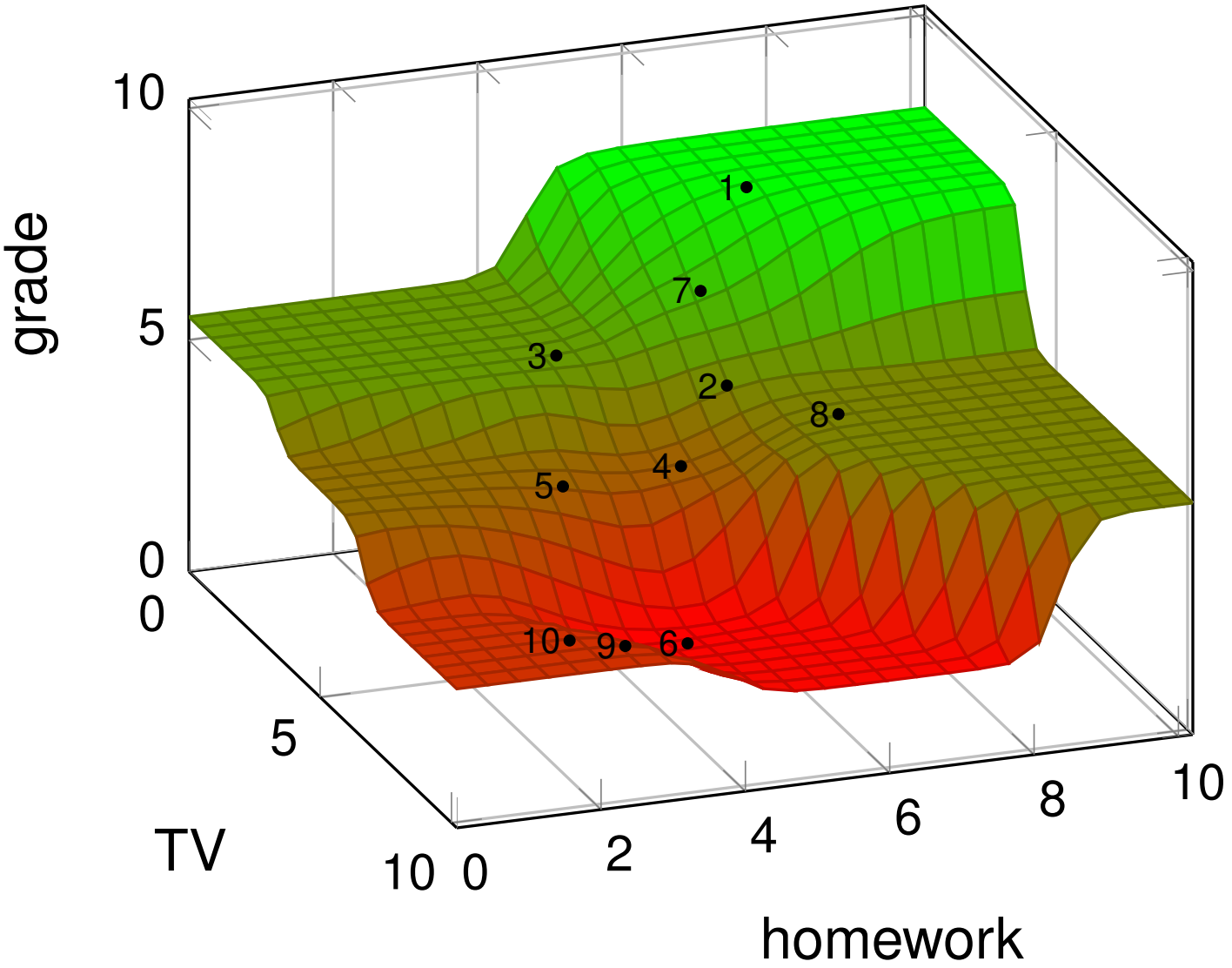
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 0.8$$



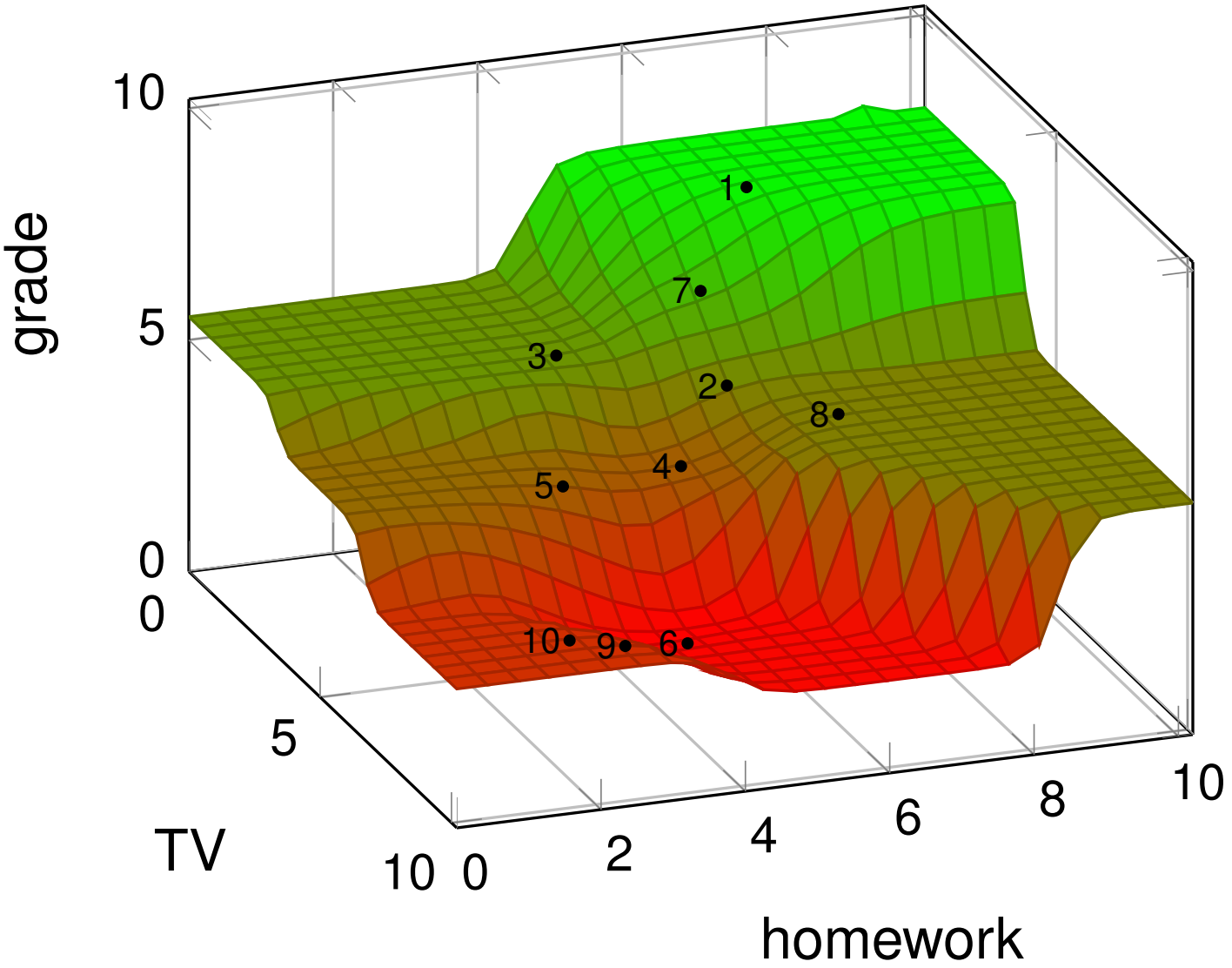
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 0.9$$



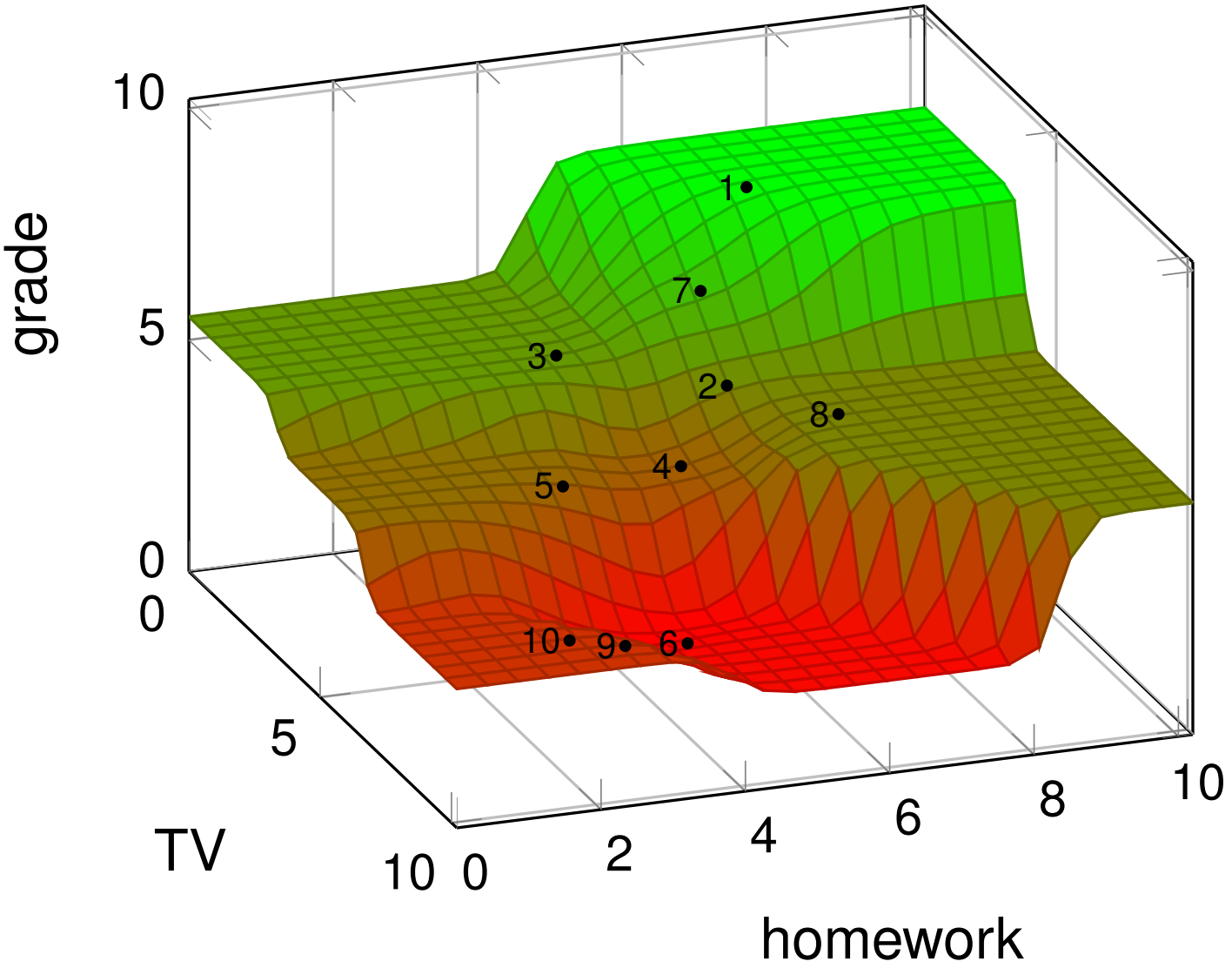
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 1.0$$



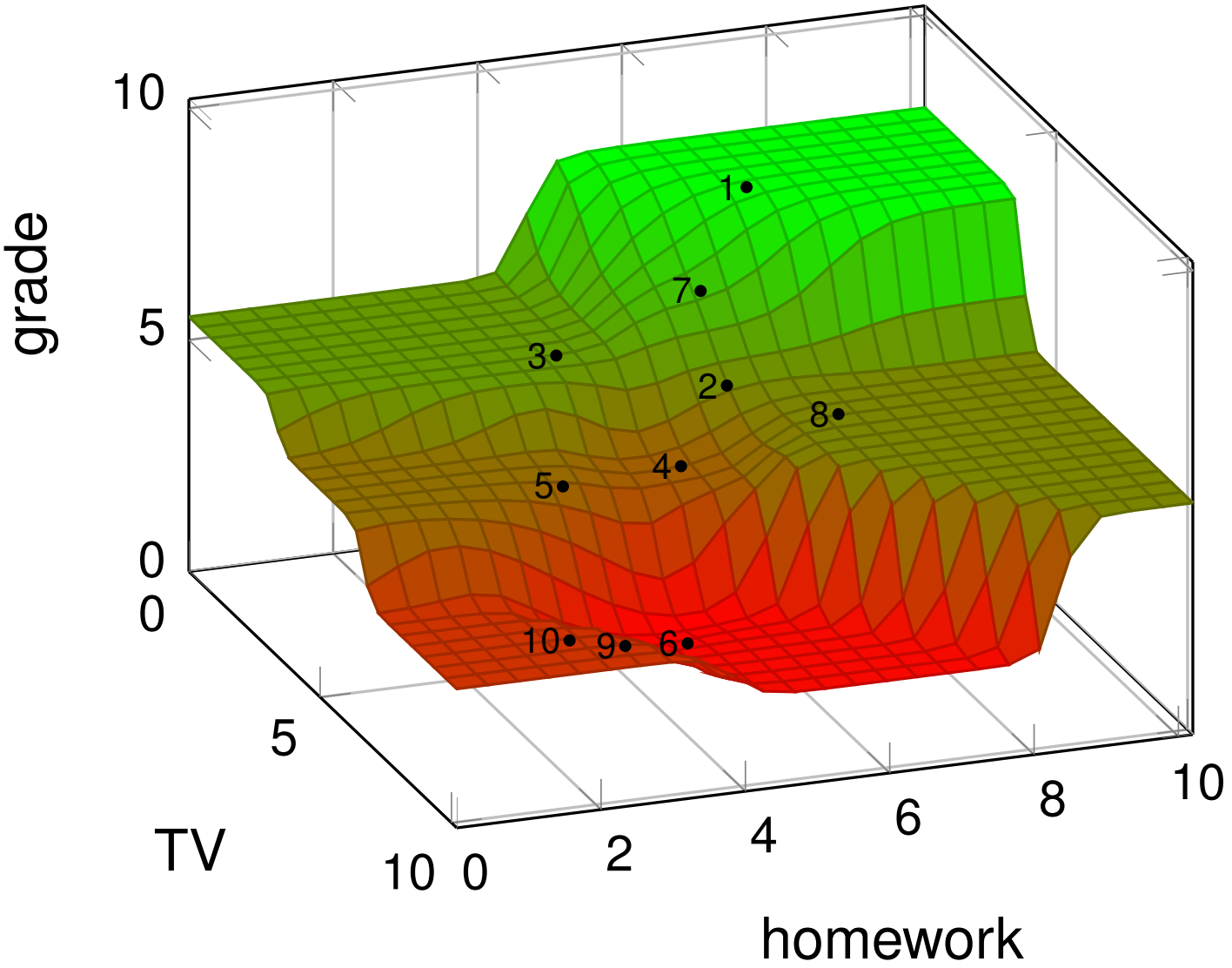
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 1.1$$



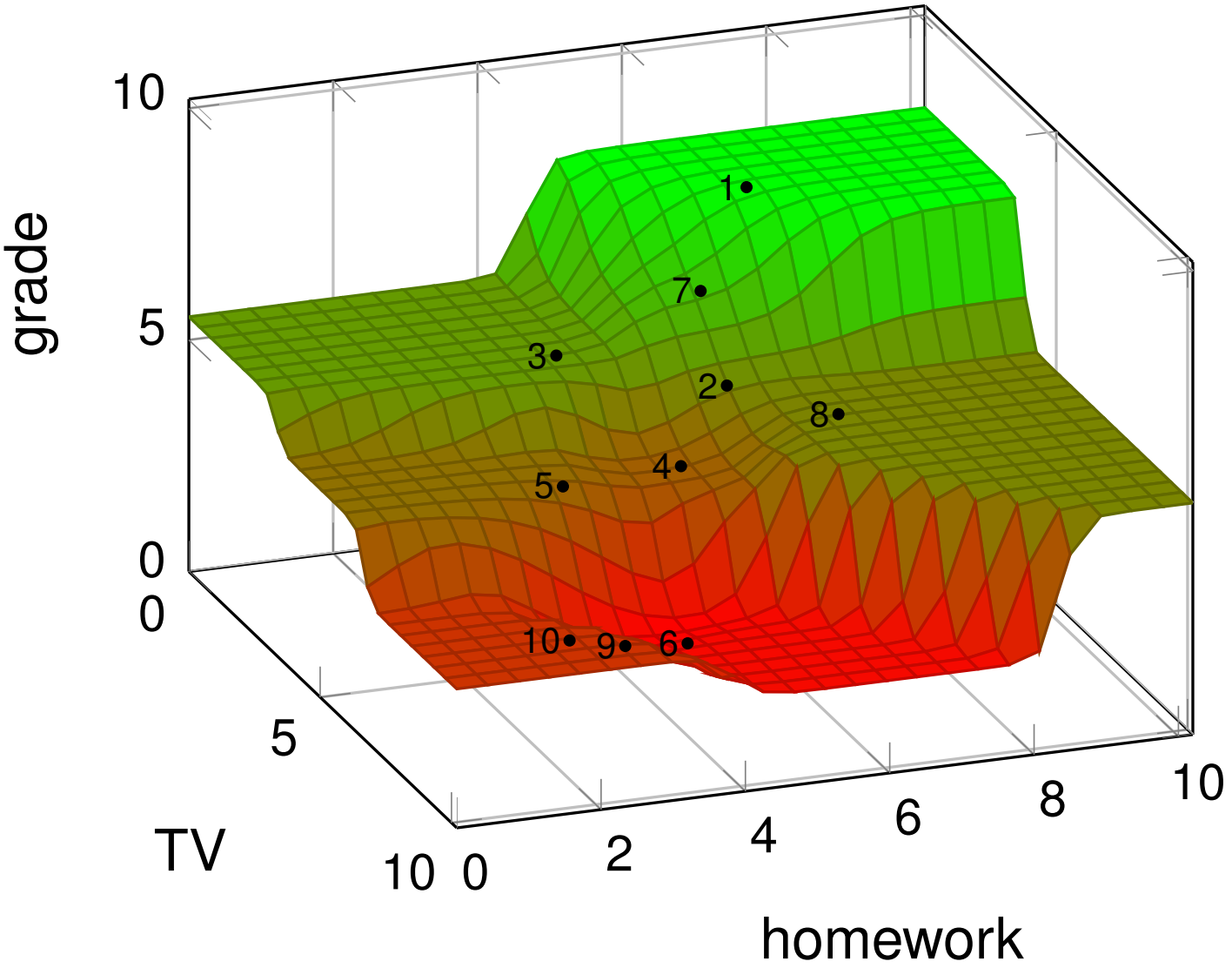
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 1.2$$



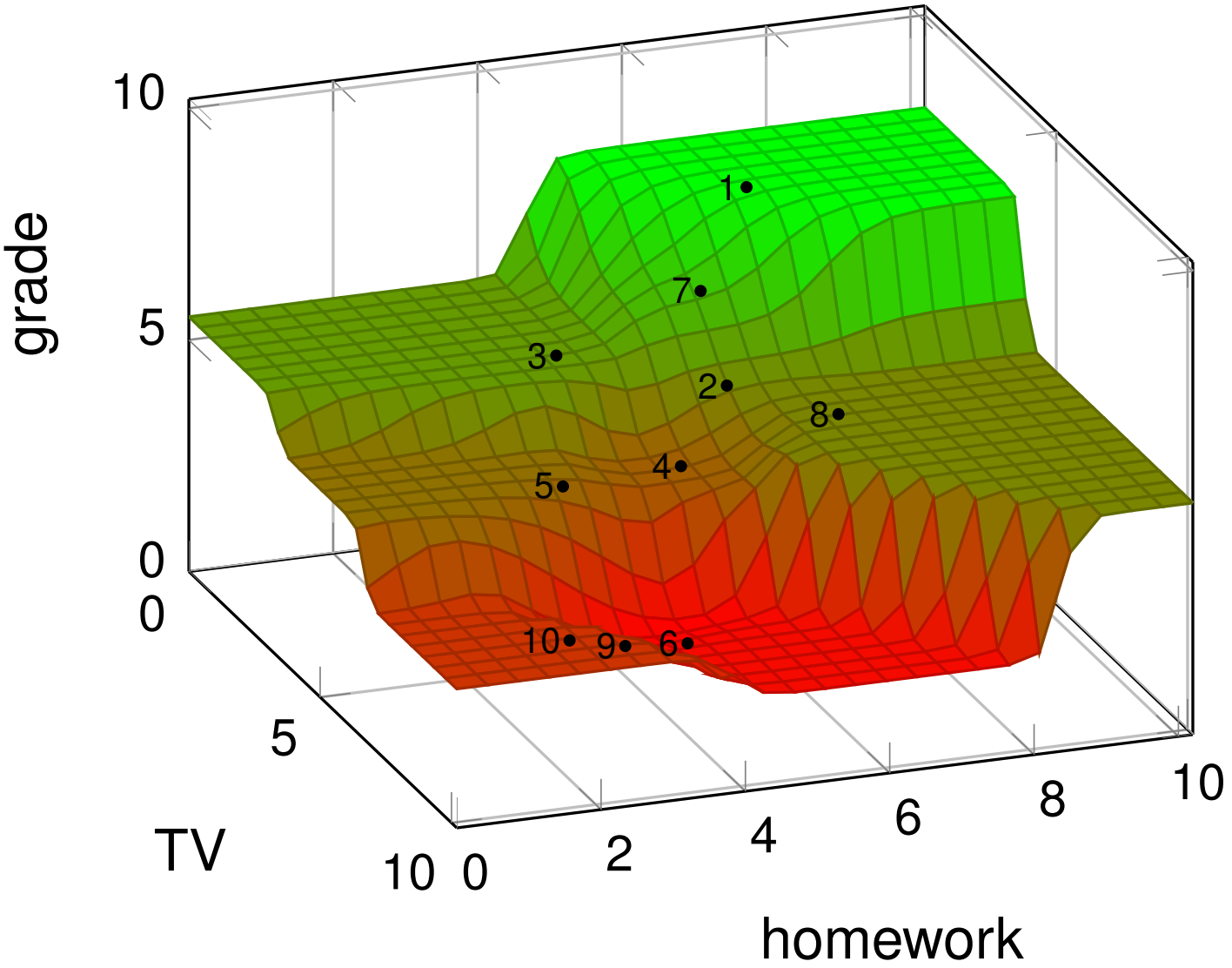
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 1.3$$



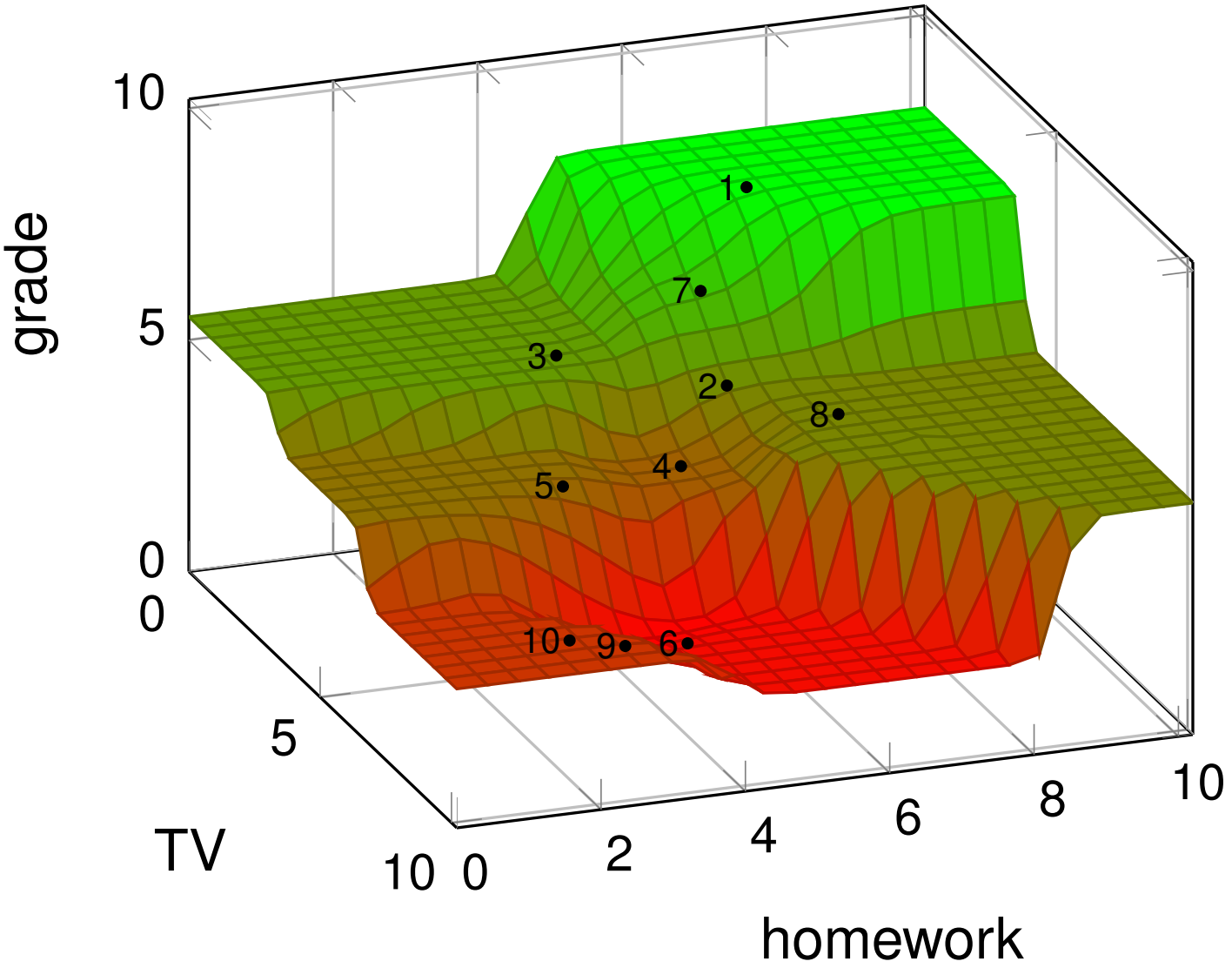
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 1.4$$



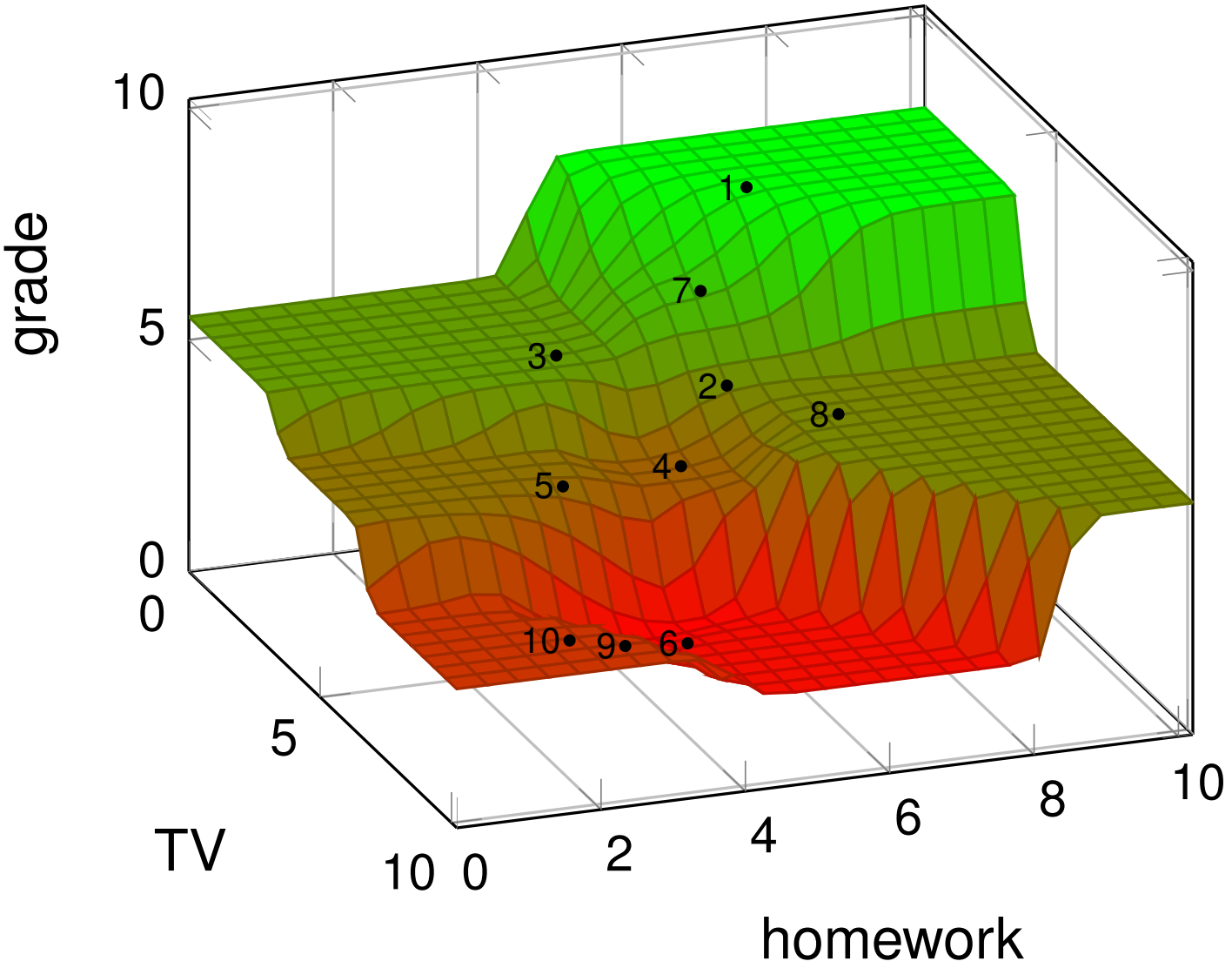
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 1.5$$



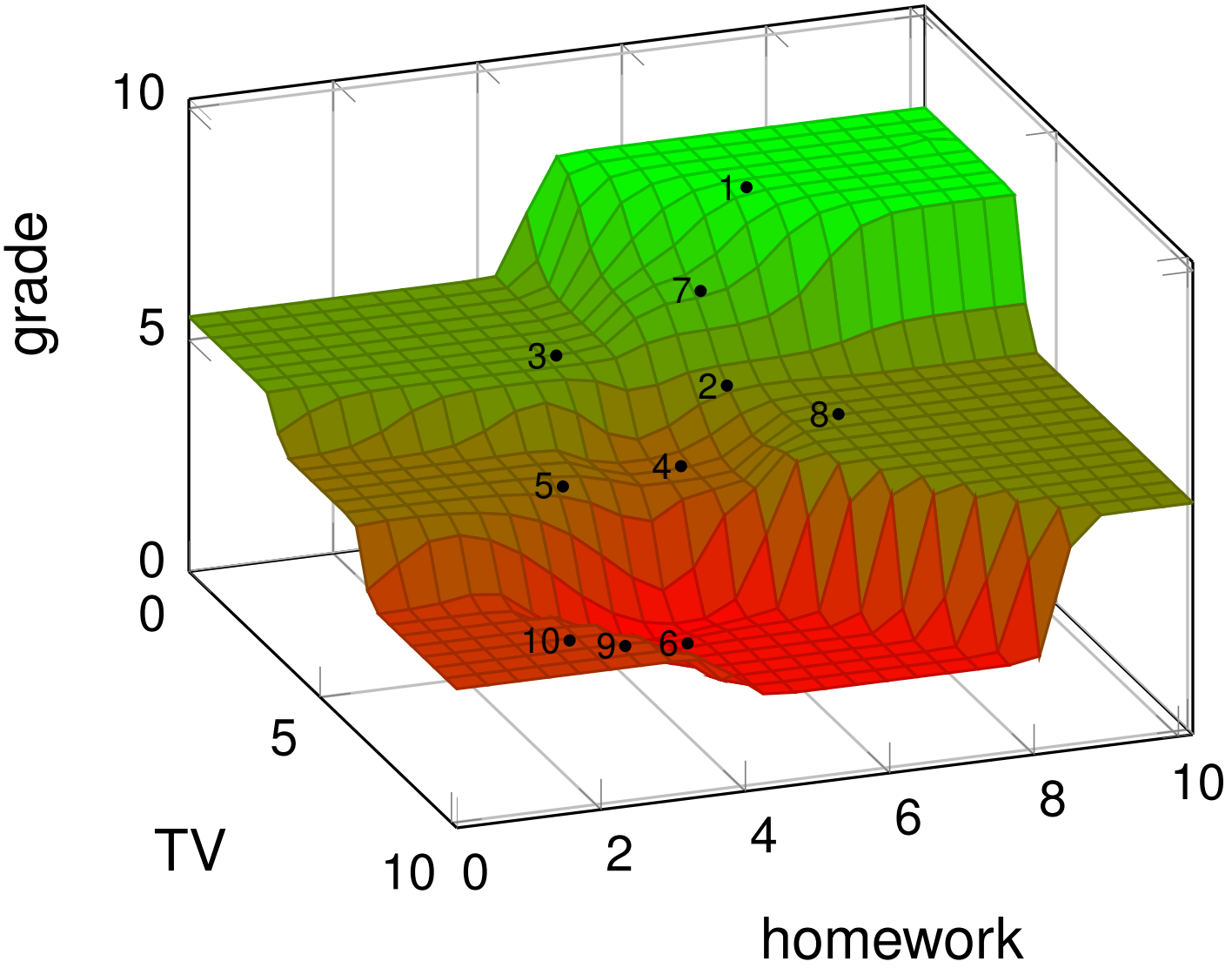
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 1.6$$



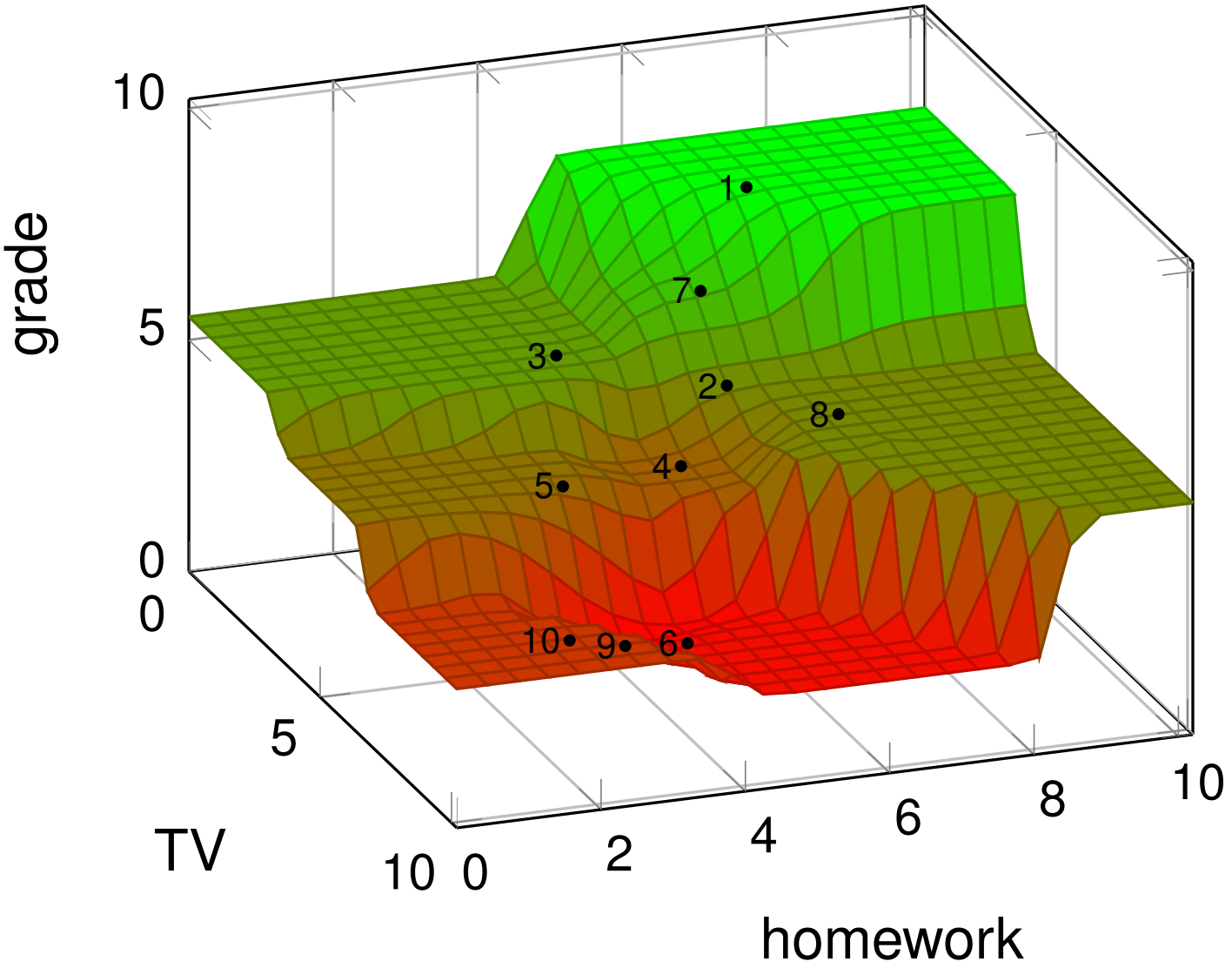
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 1.7$$



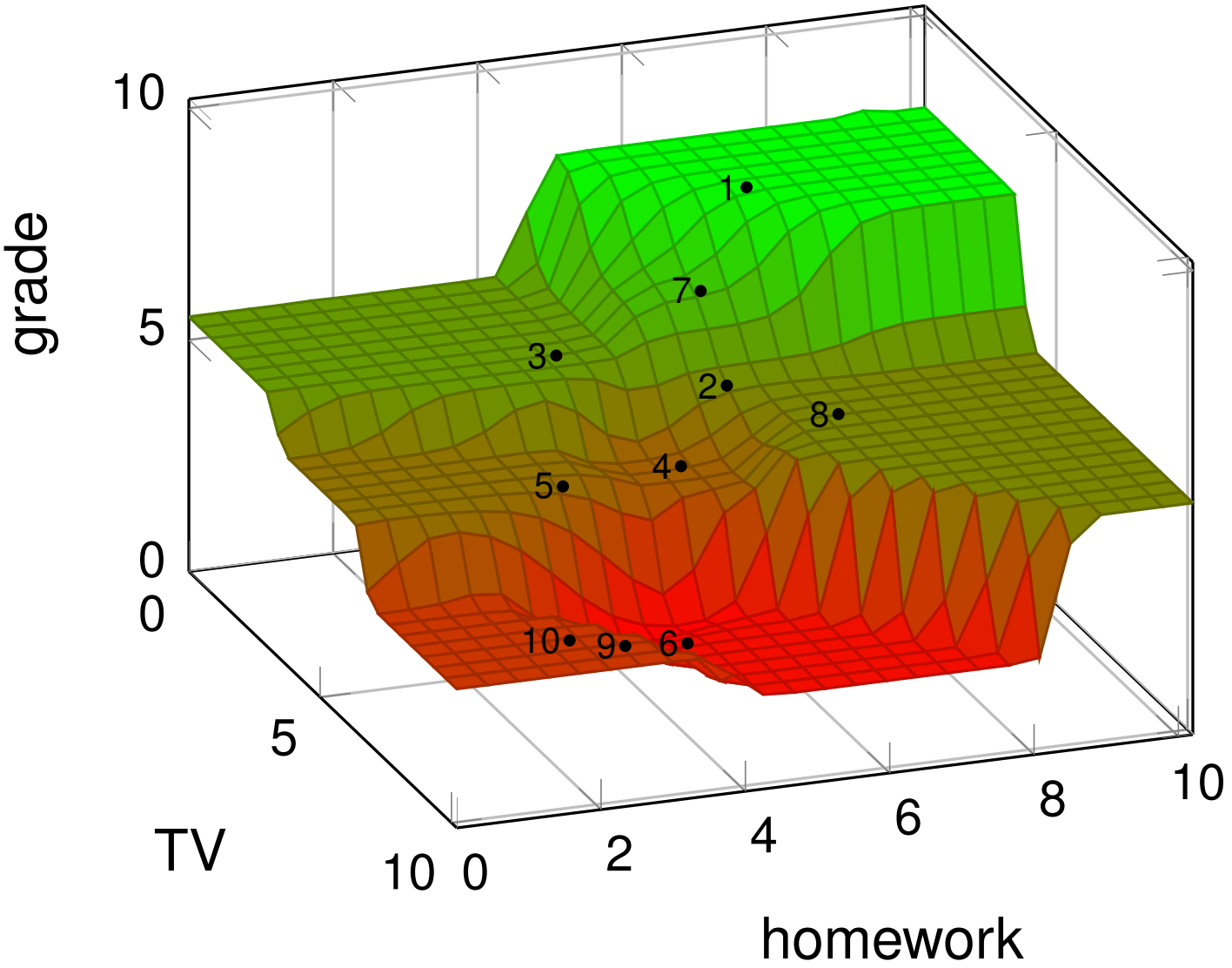
$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 1.8$$



$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 1.9$$



$$y(x) = \frac{\sum_{i=1}^{10} y_i \cdot \exp(-c\|x-x_i\|^2)}{\sum_{i=1}^{10} \exp(-c \cdot \|x-x_i\|^2)}, c = 2.0$$



Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2$ be the empirical variance

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10} (y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10} (y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10} (y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 large $\Leftrightarrow$ S_{res} small (makes sense!)

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10} (y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10} (y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 large $\Leftrightarrow$ S_{res} small (makes sense!)
- R^2 close to 1: linear model good fit, R close to 0 bad fit

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10}(y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10}(y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 large $\Leftrightarrow$ S_{res} small (makes sense!)
- R^2 close to 1: linear model good fit, R close to 0 bad fit
- R^2 smaller than 0: linear model very bad (something went wrong!)

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10}(y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10}(y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 large $\Leftrightarrow$ S_{res} small (makes sense!)
- R^2 close to 1: linear model good fit, R close to 0 bad fit
- R^2 smaller than 0: linear model very bad (something went wrong!)

R^2 for our 3 Models

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10}(y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10}(y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 large $\Leftrightarrow$ S_{res} small (makes sense!)
- R^2 close to 1: linear model good fit, R close to 0 bad fit
- R^2 smaller than 0: linear model very bad (something went wrong!)

R^2 for our 3 Models

- M1: Use TV to predict exams: $SS_{res} = 7.91$

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10}(y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10}(y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 large $\Leftrightarrow$ S_{res} small (makes sense!)
- R^2 close to 1: linear model good fit, R close to 0 bad fit
- R^2 smaller than 0: linear model very bad (something went wrong!)

$\overline{y} = 4.5$, $SS_{tot} = 35.5$

R^2 for our 3 Models

- M1: Use TV to predict exams: $SS_{res} = 7.91$

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10}(y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10}(y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 large $\Leftrightarrow$ S_{res} small (makes sense!)
- R^2 close to 1: linear model good fit, R close to 0 bad fit
- R^2 smaller than 0: linear model very bad (something went wrong!)

$\overline{y} = 4.5$, $SS_{tot} = 35.5$

R^2 for our 3 Models

- M1: Use TV to predict exams: $SS_{res} = 7.91 \Rightarrow R^2 = 0.777$
- M2: Use homework to predict exams: $SS_{res} = 18.93$

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10}(y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10}(y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 large $\Leftrightarrow$ S_{res} small (makes sense!)
- R^2 close to 1: linear model good fit, R close to 0 bad fit
- R^2 smaller than 0: linear model very bad (something went wrong!)

$\overline{y} = 4.5$, $SS_{tot} = 35.5$

R^2 for our 3 Models

- M1: Use TV to predict exams: $SS_{res} = 7.91 \Rightarrow R^2 = 0.777$
- M2: Use homework to predict exams: $SS_{res} = 18.93 \Rightarrow R^2 = 0.467$

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10}(y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10}(y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 large $\Leftrightarrow$ S_{res} small (makes sense!)
- R^2 close to 1: linear model good fit, R close to 0 bad fit
- R^2 smaller than 0: linear model very bad (something went wrong!)

$\overline{y} = 4.5$, $SS_{tot} = 35.5$

R^2 for our 3 Models

- M1: Use TV to predict exams: $SS_{res} = 7.91 \Rightarrow R^2 = 0.777$
- M2: Use homework to predict exams: $SS_{res} = 18.93 \Rightarrow R^2 = 0.467$
- M3: TV & homework to predict exams: $SS_{res} = 7.69$

Coefficient of Determination

How to figure out whether our regression model is good? Need a quantity!

R^2 (coefficient of determination)

- Let $SS_{tot} = \sum_{i=1}^{10}(y_i - \overline{y})^2$ be the empirical variance
- Let $SS_{res} = \sum_{i=1}^{10}(y_i - \widetilde{y}_i)^2$ be the squared error
  ($\widetilde{y}_i$ is the value predicted by regression)
- Define

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 large $\Leftrightarrow$ S_{res} small (makes sense!)
- R^2 close to 1: linear model good fit, R close to 0 bad fit
- R^2 smaller than 0: linear model very bad (something went wrong!)

$\overline{y} = 4.5$, $SS_{tot} = 35.5$

R^2 for our 3 Models

- M1: Use TV to predict exams: $SS_{res} = 7.91 \Rightarrow R^2 = 0.777$
- M2: Use homework to predict exams: $SS_{res} = 18.93 \Rightarrow R^2 = 0.467$
- M3: TV & homework to predict exams: $SS_{res} = 7.69 \Rightarrow R^2 = 0.783$

Digression: Polynomial Regression Tasks (1/2)

In some learning problems, a polynomial predictor may be more suitable:

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

Digression: Polynomial Regression Tasks (1/2)

In some learning problems, a polynomial predictor may be more suitable:

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

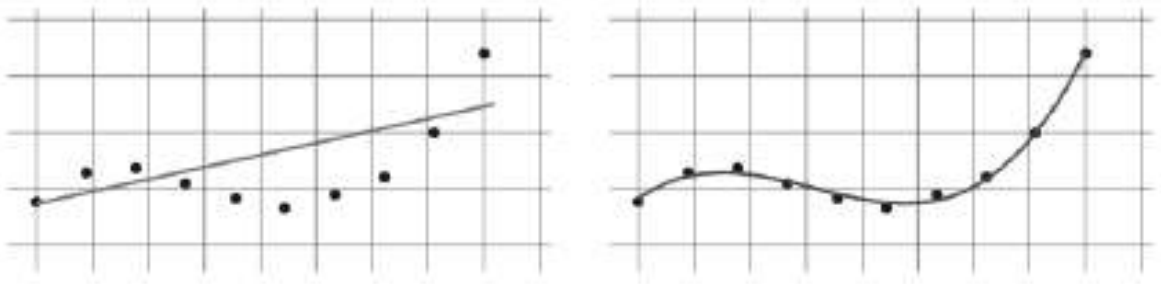
This is a polynomial of degree n, parametrised by $n + 1$ coefficients $(a_0, a_1, \ldots, a_n)$.

Digression: Polynomial Regression Tasks (1/2)

In some learning problems, a polynomial predictor may be more suitable:

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

This is a polynomial of degree n, parametrised by $n + 1$ coefficients $(a_0, a_1, \ldots, a_n)$.



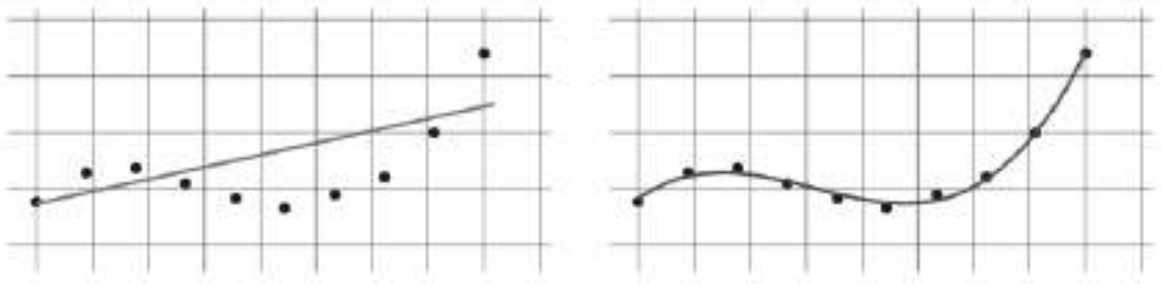
Source: Shalev-Shwartz & Ben-David

Digression: Polynomial Regression Tasks (1/2)

In some learning problems, a polynomial predictor may be more suitable:

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

This is a polynomial of degree n, parametrised by $n+1$ coefficients $(a_0, a_1, \ldots, a_n)$.



Source: Shalev-Shwartz & Ben-David

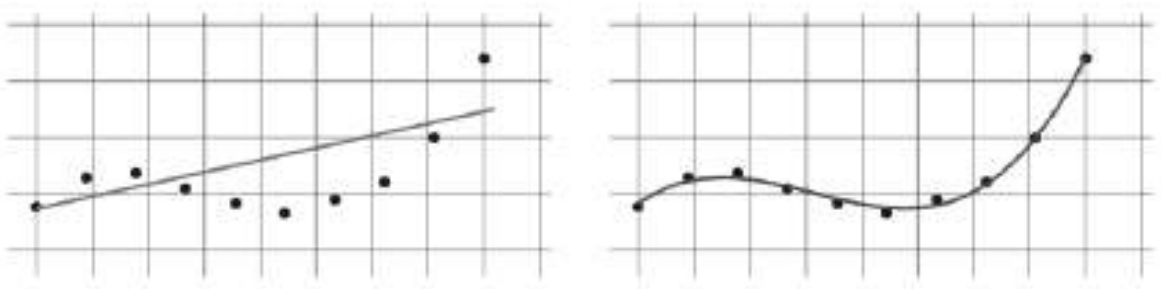
A linear predictor does not seem a good fit to the data set!

Digression: Polynomial Regression Tasks (1/2)

In some learning problems, a polynomial predictor may be more suitable:

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

This is a polynomial of degree n, parametrised by $n+1$ coefficients $(a_0, a_1, \ldots, a_n)$.

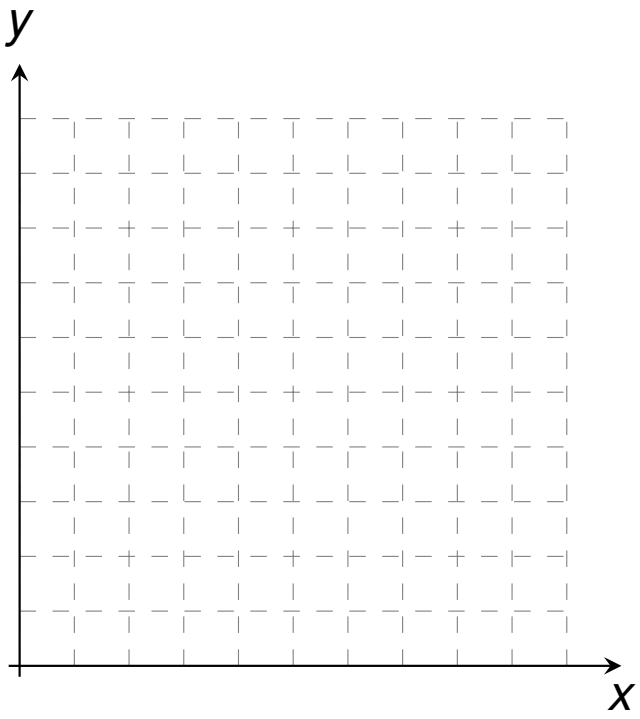


Source: Shalev-Shwartz & Ben-David

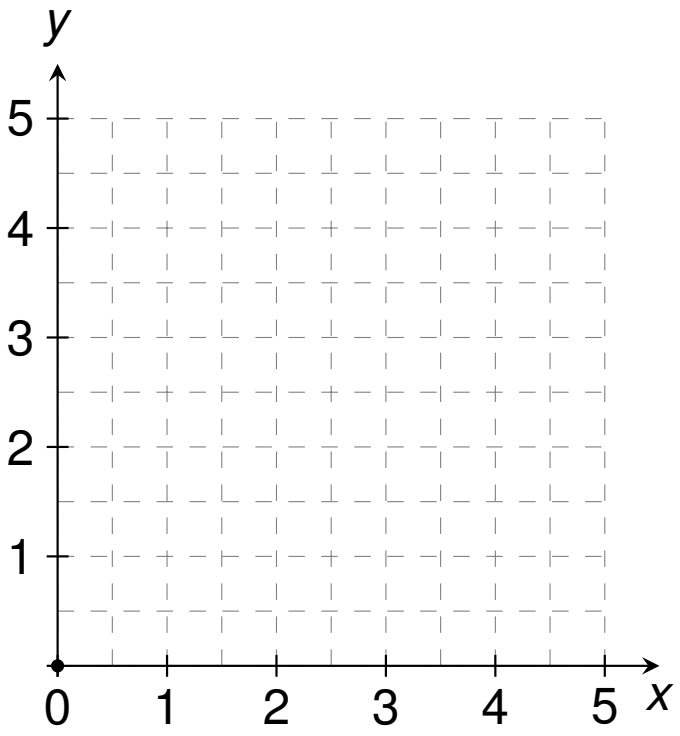
A linear predictor does not seem a good fit to the data set!

How can we learn these coefficients? (looks much harder than linear case!)

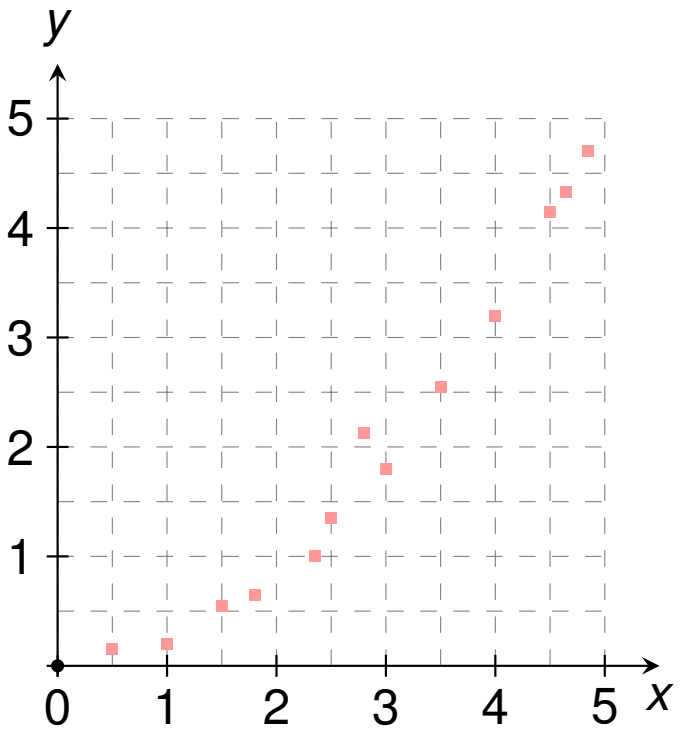
Digression: Polynomial Regression Tasks (2/2)



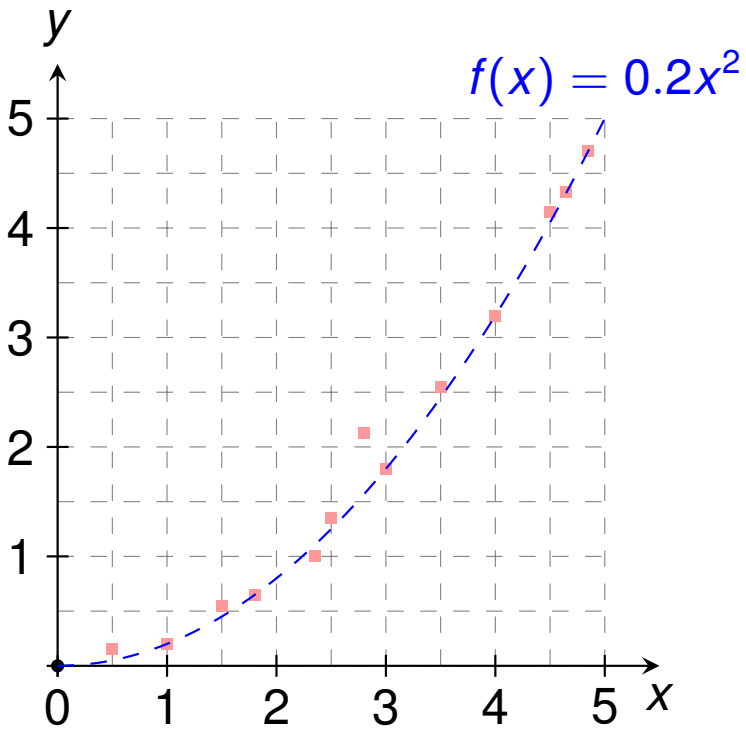
Digression: Polynomial Regression Tasks (2/2)



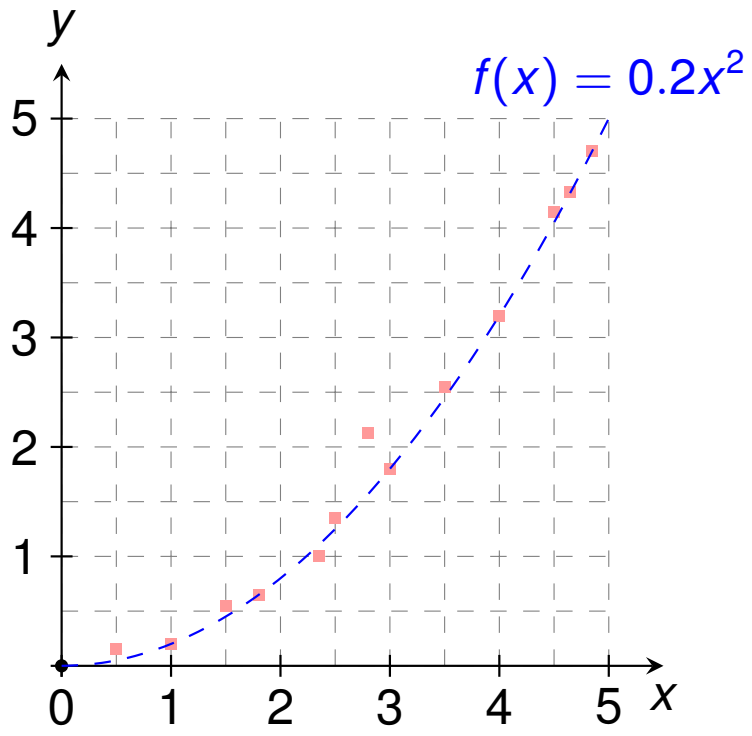
Digression: Polynomial Regression Tasks (2/2)



Digression: Polynomial Regression Tasks (2/2)



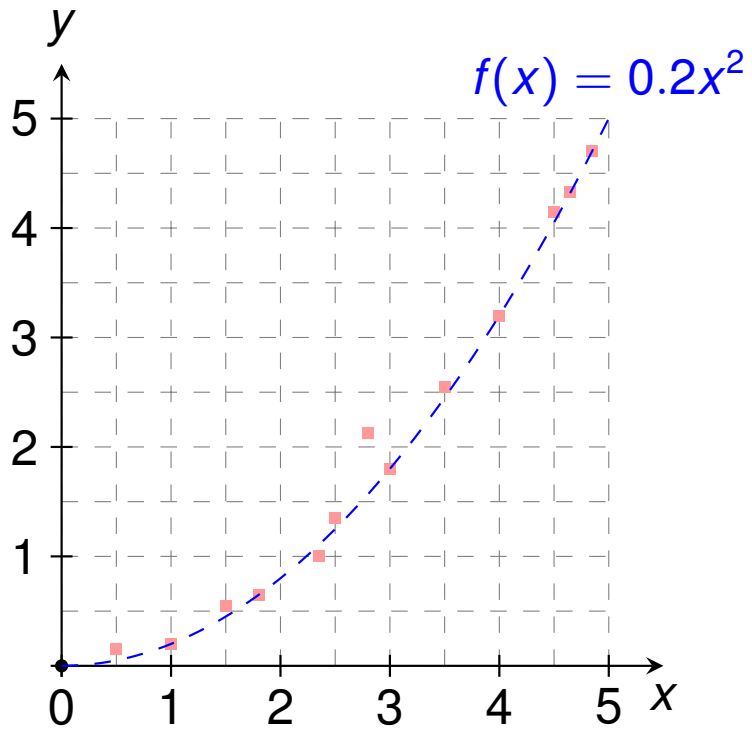
Digression: Polynomial Regression Tasks (2/2)

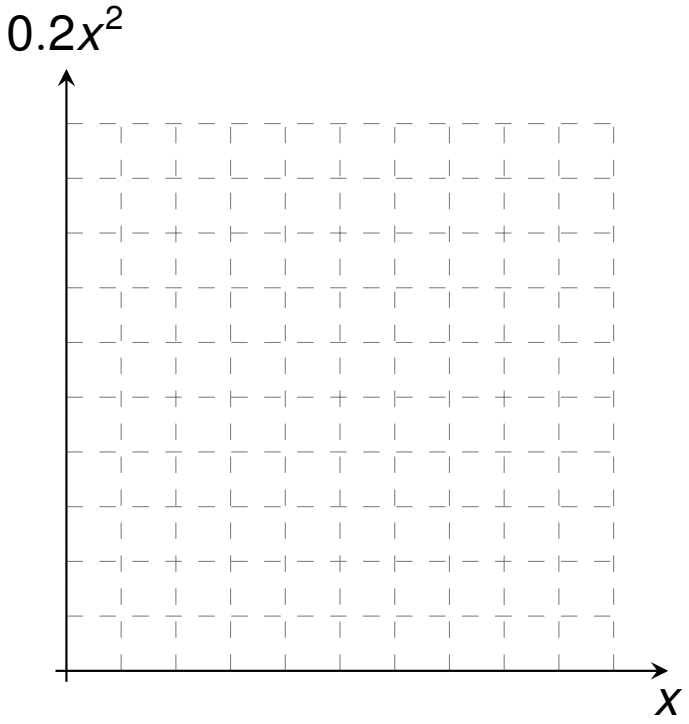


Try to fit data to a polynomial of degree 2:

$$p(x) = a_0 + a_1 x + a_2 x^2.$$

Digression: Polynomial Regression Tasks (2/2)



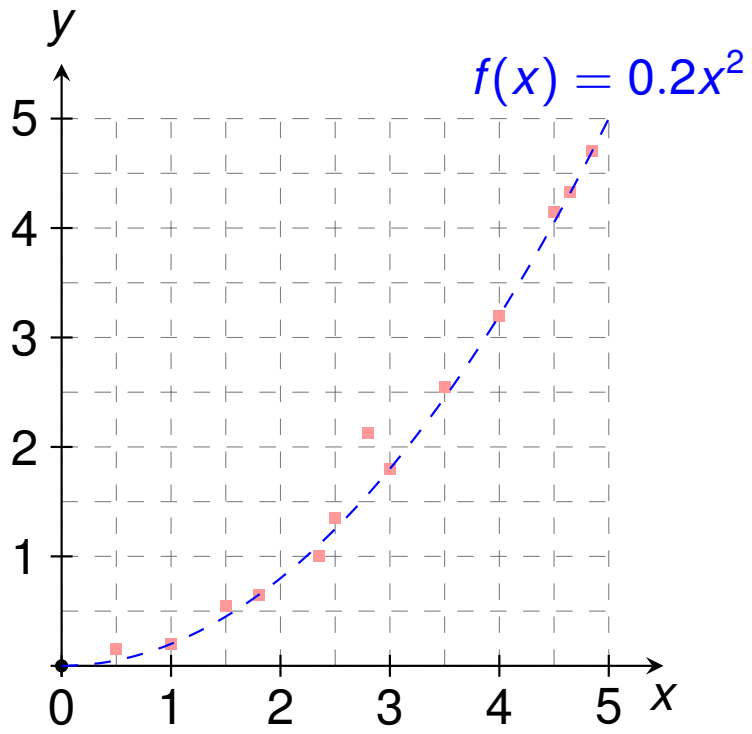


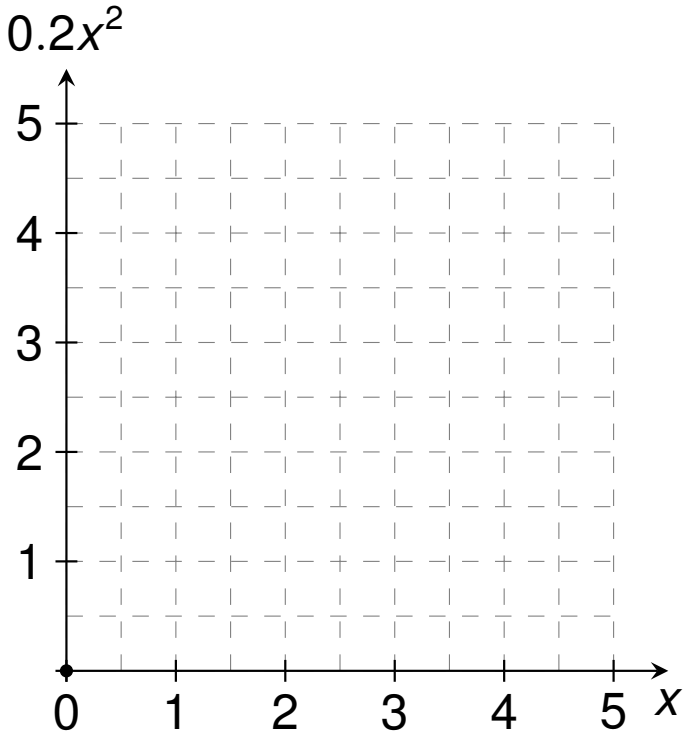
Try to fit data to a polynomial of degree 2:

$$p(x) = a_0 + a_1 x + a_2 x^2.$$

Create for each point x two features (x, x^2)!

Digression: Polynomial Regression Tasks (2/2)



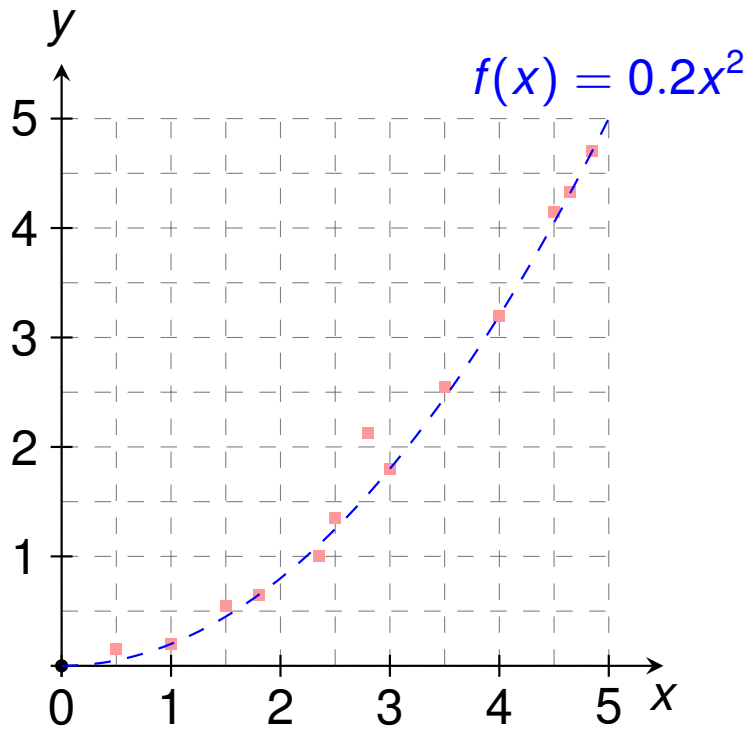


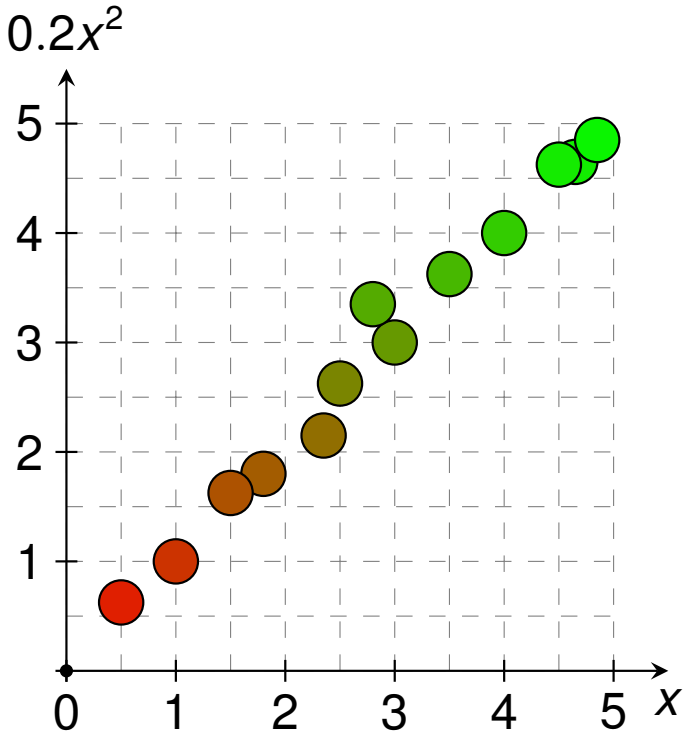
Try to fit data to a polynomial of degree 2:

$$p(x) = a_0 + a_1 x + a_2 x^2.$$

Create for each point x two features (x, x^2)!

Digression: Polynomial Regression Tasks (2/2)



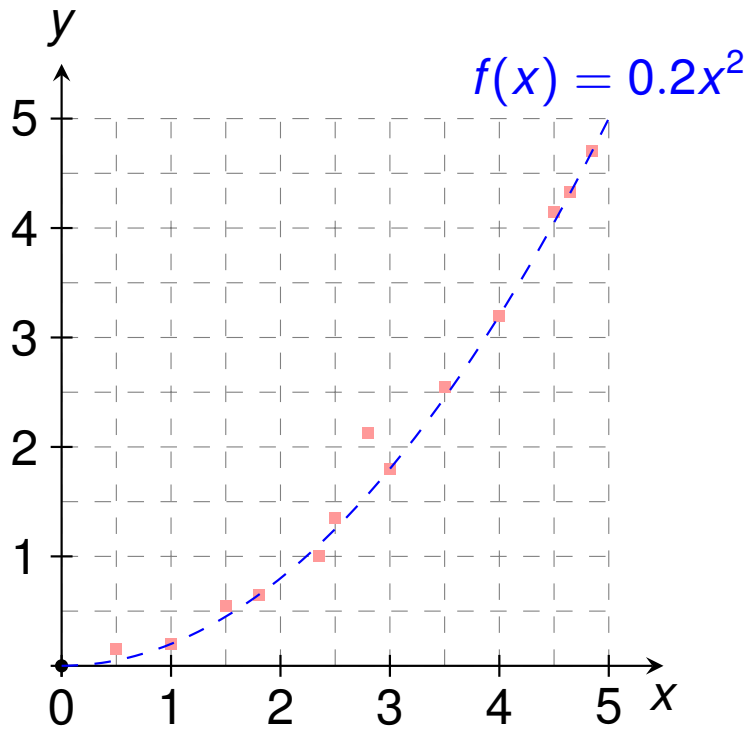


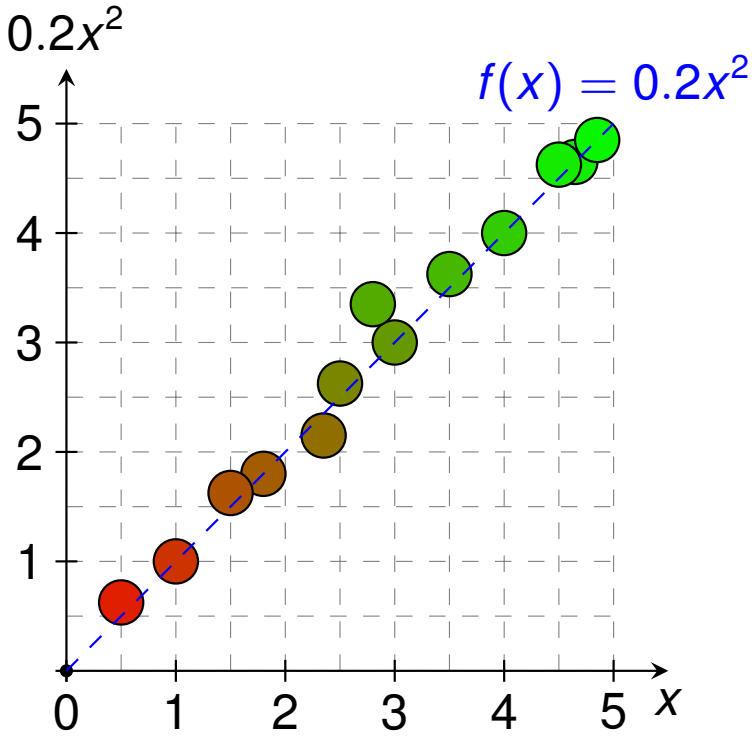
Try to fit data to a polynomial of degree 2:

$$p(x) = a_0 + a_1 x + a_2 x^2.$$

Create for each point x two features (x, x^2)!

Digression: Polynomial Regression Tasks (2/2)





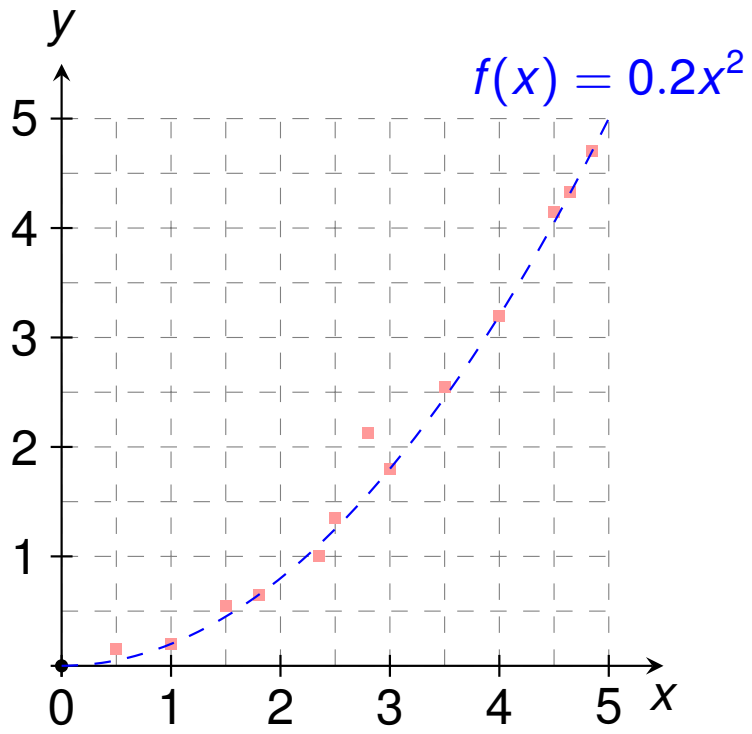
Try to fit data to a polynomial of degree 2:

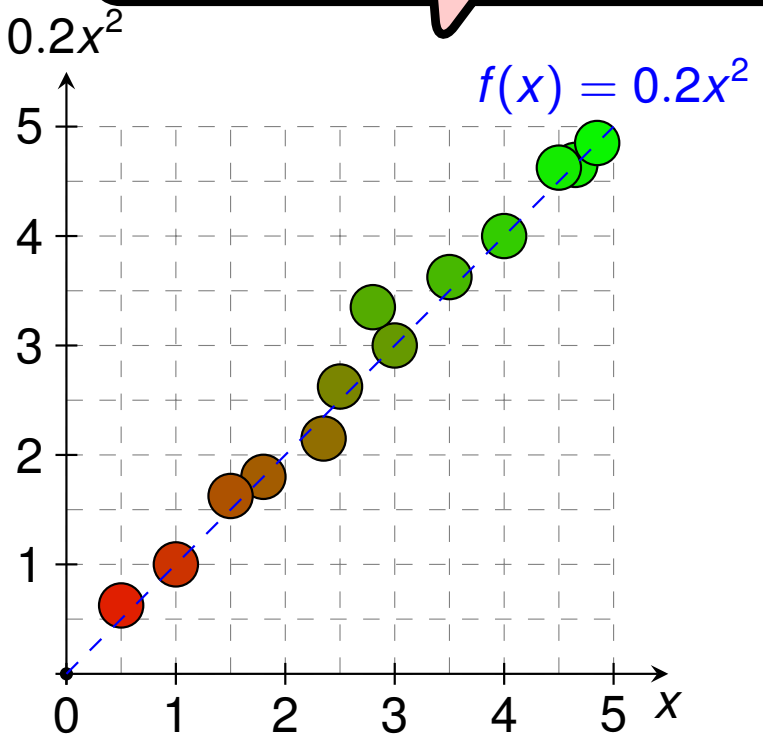
$$p(x) = a_0 + a_1 x + a_2 x^2.$$

Create for each point x two features (x, x^2)!

Digression: Polynomial Regression Tasks (2/2)

In this new feature space we apply linear regression in two dimensions!





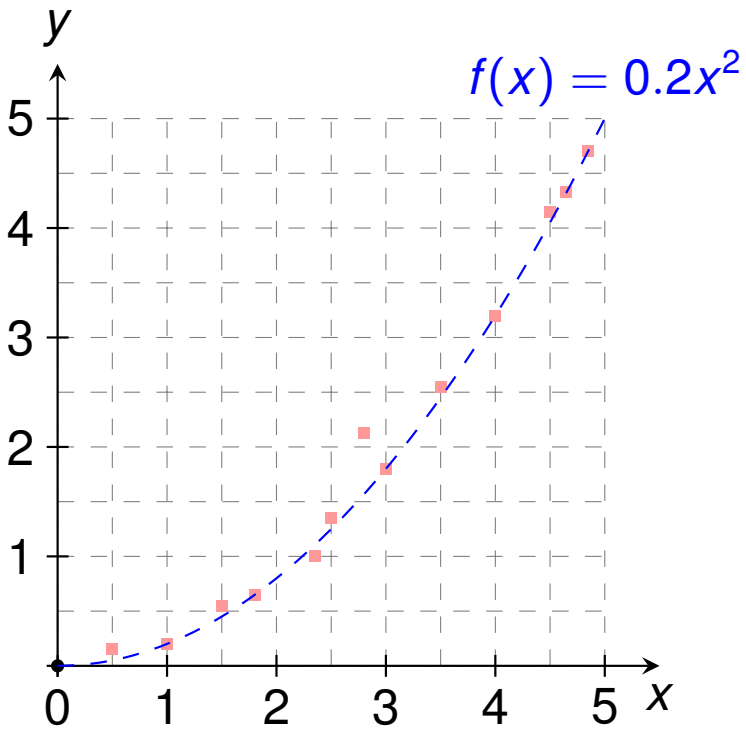
Try to fit data to a polynomial of degree 2:

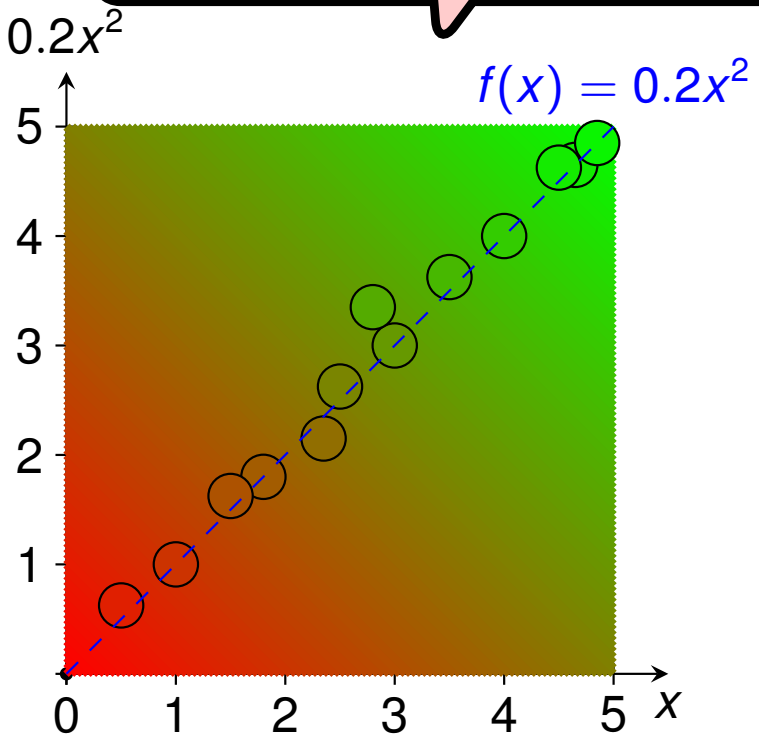
$$p(x) = a_0 + a_1 x + a_2 x^2.$$

Create for each point x two features (x, x^2)!

Digression: Polynomial Regression Tasks (2/2)

In this new feature space we apply linear regression in two dimensions!



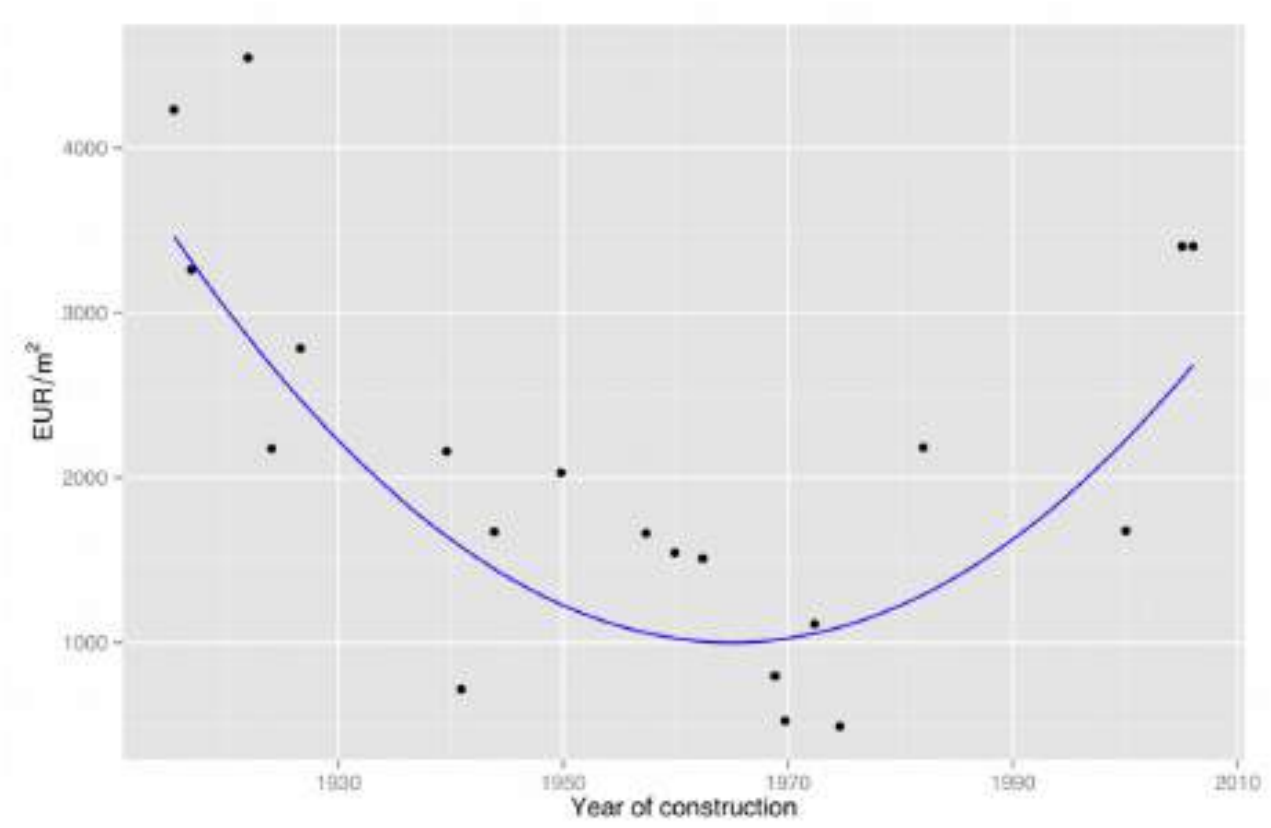


Try to fit data to a polynomial of degree 2:

$$p(x) = a_0 + a_1 x + a_2 x^2.$$

Create for each point x two features (x, x^2)!

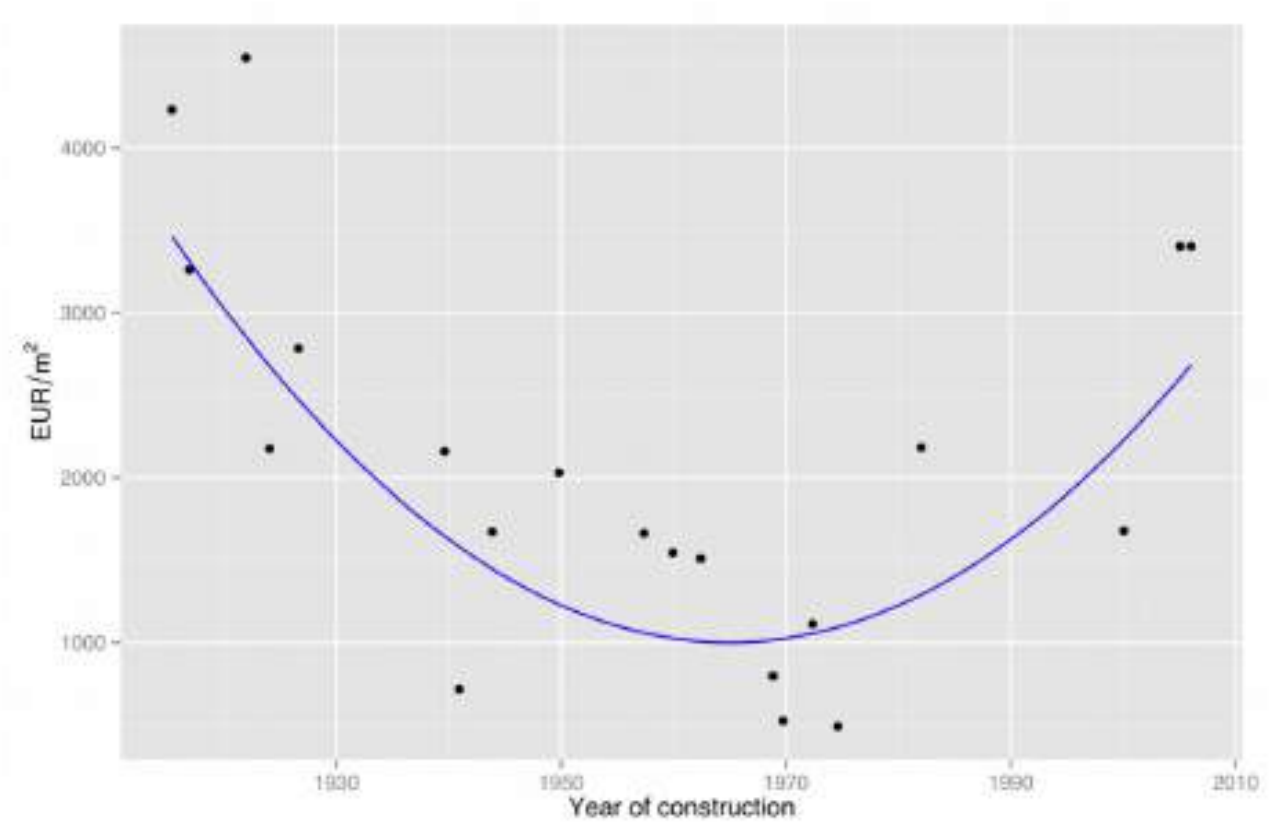
Example of a Non-Linear Relation



Source: Slides by I. Schuster and P. Jähnichen

Flat Prices in Berlin

Example of a Non-Linear Relation

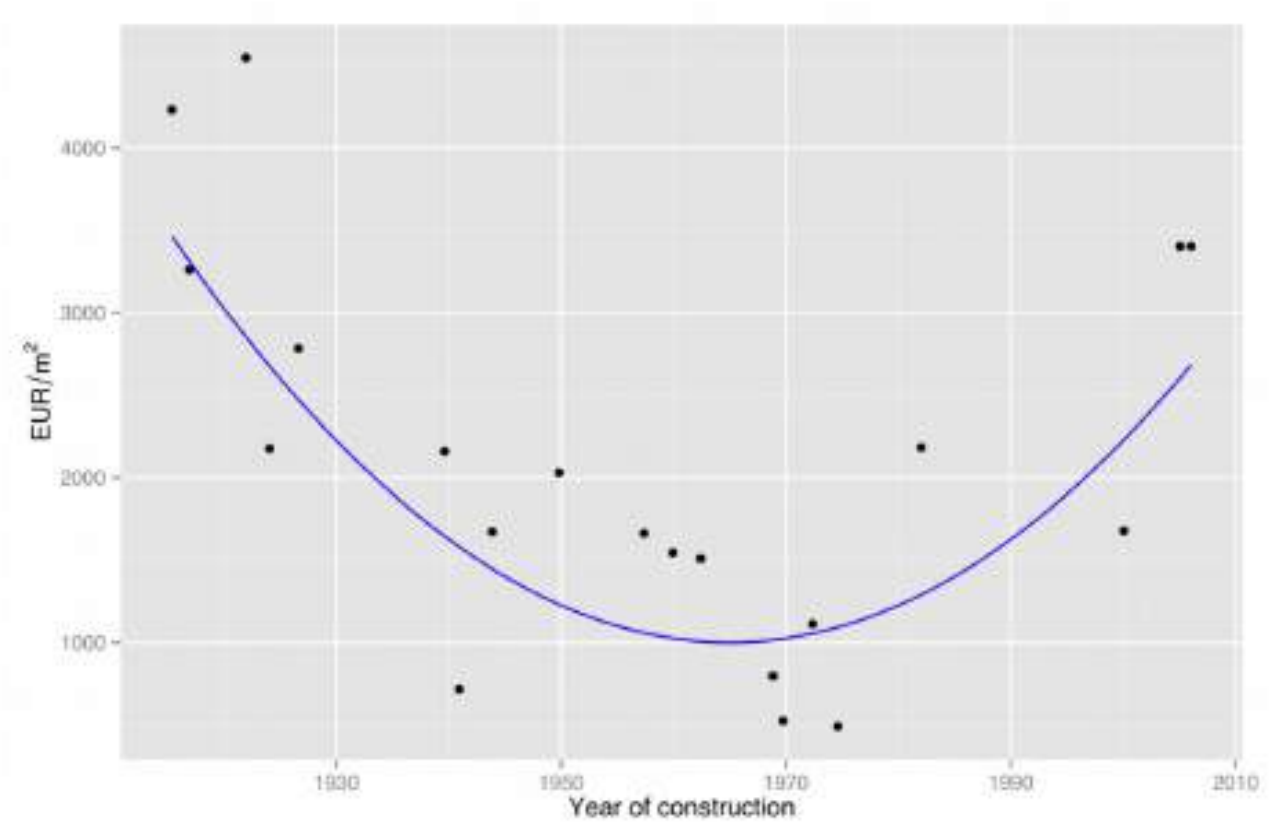


Source: Slides by I. Schuster and P. Jähnichen

Flat Prices in Berlin

- modern flats are expensive

Example of a Non-Linear Relation

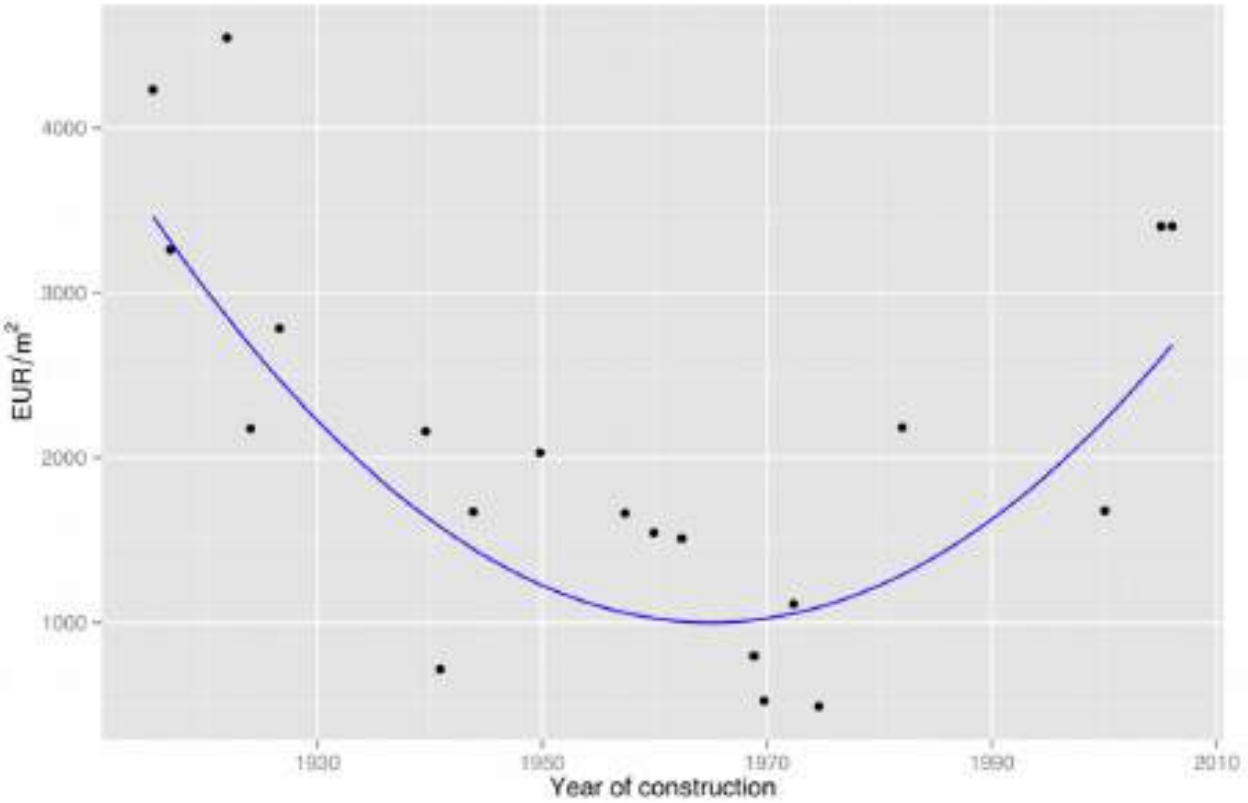


Source: Slides by I. Schuster and P. Jähnichen

Flat Prices in Berlin

- modern flats are expensive
- but: old flats are also expensive (unique, attractive design, high ceilings)

Example of a Non-Linear Relation



Source: Slides by I. Schuster and P. Jähnichen

Flat Prices in Berlin

- modern flats are expensive
- but: old flats are also expensive (unique, attractive design, high ceilings)
- flats constructed in the 70s are least attractive (unpopular design?)

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

St.	TV	Homework	Exam
1	2	7	8
2	4	6	5
3	3	4	5.5
4	5	5	4
5	6	3	4.5
6	8	4	2
7	3	6	6.5
8	5.5	7	5
9	7	3.5	1.5
10	9	2	3

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

St.	TV	Homework	Degree	Exam
1	2	7	Economics	8
2	4	6	Politics	5
3	3	4	Physics	5.5
4	5	5	Physics	4
5	6	3	Economics	4.5
6	8	4	Politics	2
7	3	6	Economics	6.5
8	5.5	7	Economics	5
9	7	3.5	Economics	1.5
10	9	2	Physics	3

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

St.	TV	Homework	Degree	Exam
1	2	7	Economics	8
2	4	6	Politics	5
3	3	4	Physics	5.5
4	5	5	Physics	4
5	6	3	Economics	4.5
6	8	4	Politics	2
7	3	6	Economics	6.5
8	5.5	7	Economics	5
9	7	3.5	Economics	1.5
10	9	2	Physics	3

Linear Regression needs to work with numerical values!

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

Use encoding 0 = Economics, 1 = Physics and 2 = Politics.

St.	TV	Homework	Degree	Exam
1	2	7	Economics	8
2	4	6	Politics	5
3	3	4	Physics	5.5
4	5	5	Physics	4
5	6	3	Economics	4.5
6	8	4	Politics	2
7	3	6	Economics	6.5
8	5.5	7	Economics	5
9	7	3.5	Economics	1.5
10	9	2	Physics	3

Linear Regression needs to work with numerical values!

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

Use encoding 0 = Economics, 1 = Physics and 2 = Politics.

St.	TV	Homework	Degree	Exam
1	2	7	0	8
2	4	6	2	5
3	3	4	1	5.5
4	5	5	1	4
5	6	3	0	4.5
6	8	4	2	2
7	3	6	0	6.5
8	5.5	7	0	5
9	7	3.5	0	1.5
10	9	2	1	3

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

Use encoding 0 = Economics, 1 = Physics and 2 = Politics.

St.	TV	Homework	Degree	Exam
1	2	7	0	8
2	4	6	2	5
3	3	4	1	5.5
4	5	5	1	4
5	6	3	0	4.5
6	8	4	2	2
7	3	6	0	6.5
8	5.5	7	0	5
9	7	3.5	0	1.5
10	9	2	1	3

Is Physics really the average of Economics and Politics ☺?

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

St.	TV	Homework	Degree	Exam
1	2	7	0	8
2	4	6	2	5
3	3	4	1	5.5
4	5	5	1	4
5	6	3	0	4.5
6	8	4	2	2
7	3	6	0	6.5
8	5.5	7	0	5
9	7	3.5	0	1.5
10	9	2	1	3

Is Physics really the average of Economics and Politics ☺?

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

Solution: Use one-hot-encoding: one-feature for each category value.

St.	TV	Homework	Degree	Exam
1	2	7	0	8
2	4	6	2	5
3	3	4	1	5.5
4	5	5	1	4
5	6	3	0	4.5
6	8	4	2	2
7	3	6	0	6.5
8	5.5	7	0	5
9	7	3.5	0	1.5
10	9	2	1	3

Is Physics really the average of Economics and Politics ☺?

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

Solution: Use one-hot-encoding: one-feature for each category value.

St.	TV	Homework	Economics	Physics	Politics	Exam
1	2	7	1	0	0	8
2	4	6	0	0	1	5
3	3	4	0	1	0	5.5
4	5	5	0	1	0	4
5	6	3	1	0	0	4.5
6	8	4	0	0	1	2
7	3	6	1	0	0	6.5
8	5.5	7	1	0	0	5
9	7	3.5	1	0	0	1.5
10	9	2	0	1	0	3

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

Solution: Use one-hot-encoding: one-feature for each category value.

St.	TV	Homework	Economics	Physics	Politics	Exam
1	2	7	1	0	0	8
2	4	6	0	0	1	5
3	3	4	0	1	0	5.5
4	5	5	0	1	0	4
5	6	3	1	0	0	4.5
6	8	4	0	0	1	2
7	3	6	1	0	0	6.5
8	5.5	7	1	0	0	5
9	7	3.5	1	0	0	1.5
10	9	2	0	1	0	3

Confounded Features: We can drop one category!

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

Solution: Use one-hot-encoding: one-feature for each category value.

St.	TV	Homework	Economics	Physics	Exam
1	2	7	1	0	8
2	4	6	0	0	5
3	3	4	0	1	5.5
4	5	5	0	1	4
5	6	3	1	0	4.5
6	8	4	0	0	2
7	3	6	1	0	6.5
8	5.5	7	1	0	5
9	7	3.5	1	0	1.5
10	9	2	0	1	3

How to Work with Categorical Variables

Most data sets in real life contain both categorical and numerical features.

Suppose we know the degree of each student: Economics, Physics and Politics

Solution: Use one-hot-encoding: one-feature for each category value.

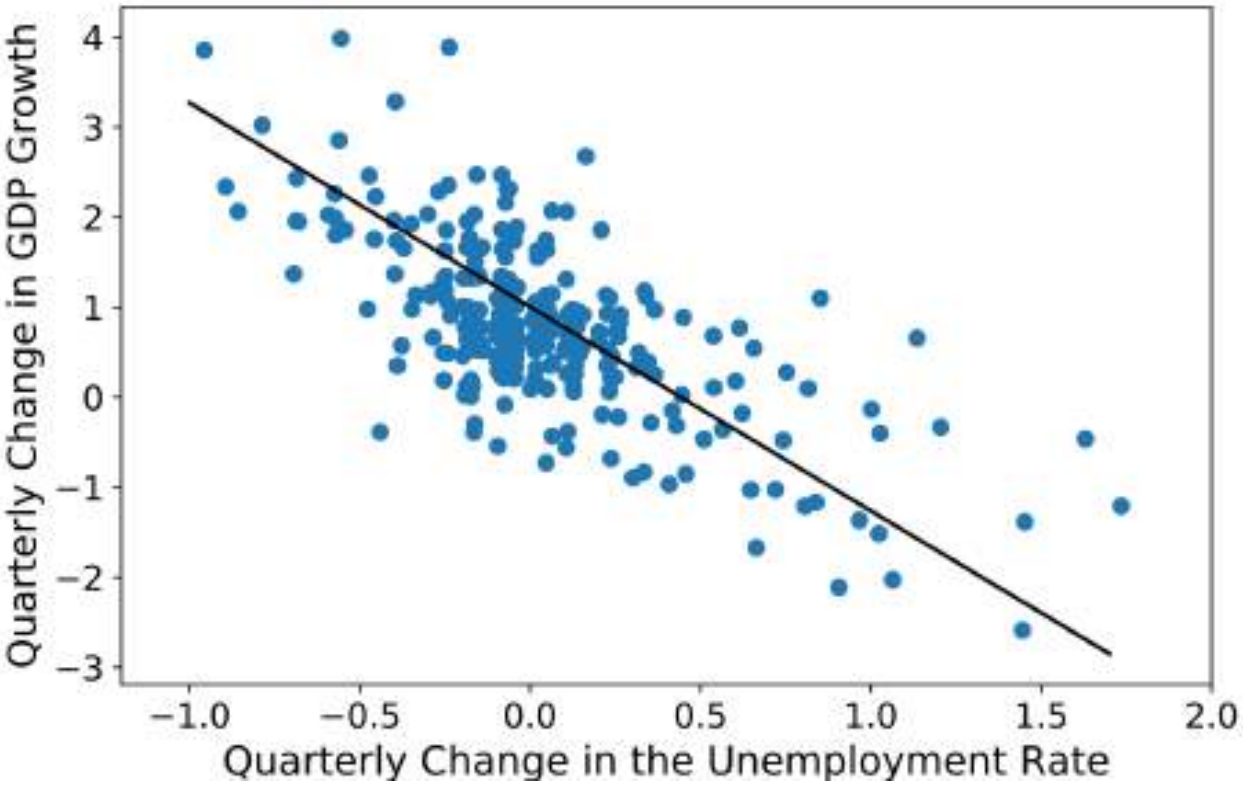
St.	TV	Homework	Economics	Physics	Exam
1	2	7	1	0	8
2	4	6	0	0	5
3	3	4	0	1	5.5
4	5	5	0	1	4
5	6	3	1	0	4.5
6	8	4	0	0	2
7	3	6	1	0	6.5
8	5.5	7	1	0	5
9	7	3.5	1	0	1.5
10	9	2	0	1	3

If a student does not study Economics or Physics, then he/she studies Politics ✓

Linear Regression in Practice: Economics

Okun's Law

For every 1% increase in the unemployment rate, a country's GDP will be roughly an additional 2% lower.



Source: Wikipedia

Outline

Introduction

Linear Regression

Linear Regression with More than One Feature

Solving Linear Regression: Gradient Descent

Logistic Regression

The Linear Regression Problem (in General)

Regression:

- n data points $(\mathbf{x}^i, y^i)$, $1 \leq i \leq n$

The Linear Regression Problem (in General)

Regression:

- n data points $(\mathbf{x}^i, y^i)$, $1 \leq i \leq n$
- Find a map $f : \mathbb{R}^{d+1} \to \mathbb{R}$ with

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + \underbrace{w_0}_{=b} .$$

The Linear Regression Problem (in General)

Regression:

- n data points $(\mathbf{x}^i, y^i)$, $1 \leq i \leq n$
- Find a map $f : \mathbb{R}^{d+1} \to \mathbb{R}$ with

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + \underbrace{w_0}_{=\mathrm{b}} .$$

- Loss-Function is the Mean Squared Error,

$$L(f) = \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 .$$

The Linear Regression Problem (in General)

Regression:

- n data points $(\mathbf{x}^i, y^i)$, $1 \leq i \leq n$
- Find a map $f : \mathbb{R}^{d+1} \to \mathbb{R}$ with

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + \underbrace{w_0}_{=b} .$$

- Loss-Function is the Mean Squared Error,

$$L(f) = \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 .$$

Objective: Find $(w_0, w_1, \ldots, w_d)$ such that $L(f_{\mathbf{w},b})$ is minimised.

How to solve the Linear Regression Problem?

We wish to solve:

$$\min_{w} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{d} w_i x_j^{(i)} - y^{(i)} \right)^2$$

How to solve the Linear Regression Problem?

We wish to solve:

$$\min_{w} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{d} w_i x_j^{(i)} - y^{(i)} \right)^2$$

- This is a convex function in $\mathbf{w} = (w_0, \ldots, w_d)$!

How to solve the Linear Regression Problem?

We wish to solve:

$$\min_{w} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{d} w_i x_j^{(i)} - y^{(i)} \right)^2$$

- This is a convex function in $\mathbf{w} = (w_0, \ldots, w_d)$!
- Convex functions are convenient since:

How to solve the Linear Regression Problem?

We wish to solve:

$$\min_{w} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{d} w_i x_j^{(i)} - y^{(i)} \right)^2$$

- This is a convex function in $\mathbf{w} = (w_0, \ldots, w_d)$!
- Convex functions are convenient since:

$$f(\mathbf{w}) \text{ global minimum}$$

How to solve the Linear Regression Problem?

We wish to solve:

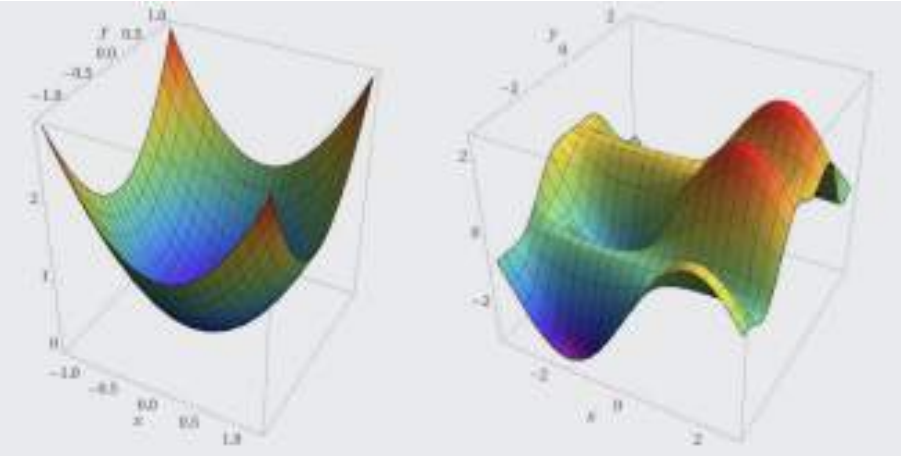
$$\min_{w} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{d} w_i x_j^{(i)} - y^{(i)} \right)^2$$

- This is a convex function in $\mathbf{w} = (w_0, \ldots, w_d)$!
- Convex functions are convenient since:

$$f(\mathbf{w}) \text{ global minimum } \Leftrightarrow \frac{df}{dw} = 0$$

We will see how to apply **gradient methods** to solve this!

How to solve the Linear Regression Problem?



We wish to solve:

$$\min_{w} \sum_{i=1}^{n} \left(w_0 + \sum_{j=1}^{d} w_i x_j^{(i)} - y^{(i)} \right)^2$$

- This is a convex function in $\mathbf{w} = (w_0, \ldots, w_d)$!
- Convex functions are convenient since:

$$f(\mathbf{w}) \text{ global minimum } \Leftrightarrow \frac{df}{dw} = 0$$

We will see how to apply **gradient methods** to solve this!

Gradient Descent

- $L : \mathbb{R}^{d+1} \to \mathbb{R}$ is any function in $\mathbf{w} = (w_0, w_1, \ldots, w_d)$

Gradient Descent

- $L : \mathbb{R}^{d+1} \to \mathbb{R}$ is any function in $\mathbf{w} = (w_0, w_1, \ldots, w_d)$ (loss function)
- The gradient is:

$$\nabla L(\mathbf{w}) = \left(\frac{dL(\mathbf{w})}{dw_0}, \frac{dL(\mathbf{w})}{dw_1}, \ldots, \frac{dL(\mathbf{w})}{dw_d} \right)$$

Gradient Descent

- $L : \mathbb{R}^{d+1} \to \mathbb{R}$ is any function in $\mathbf{w} = (w_0, w_1, \ldots, w_d)$ (loss function)
- The gradient is:

$$\nabla L(\mathbf{w}) = \left(\frac{dL(\mathbf{w})}{dw_0}, \frac{dL(\mathbf{w})}{dw_1}, \ldots, \frac{dL(\mathbf{w})}{dw_d} \right)$$

Gradient Descent for minimizing $L(\mathbf{w})$

Parameters: Learning Rate $\delta > 0$
Initialize: $\mathbf{w}^{(1)}$
for $t = 1, 2, \ldots, T$
  Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \delta \nabla L(\mathbf{w}^{(t)})$
Return $w^{(T)}$

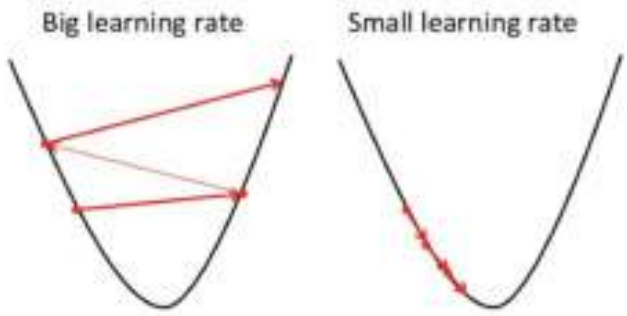
Gradient Descent

- $L : \mathbb{R}^{d+1} \to \mathbb{R}$ is any function in $\mathbf{w} = (w_0, w_1, \ldots, w_d)$ (loss function)
- The gradient is:

$$\nabla L(\mathbf{w}) = \left(\frac{dL(\mathbf{w})}{dw_0}, \frac{dL(\mathbf{w})}{dw_1}, \ldots, \frac{dL(\mathbf{w})}{dw_d} \right)$$

Gradient Descent for minimizing $L(\mathbf{w})$

Parameters: Learning Rate $\delta > 0$
Initialize: $\mathbf{w}^{(1)}$
for $t = 1, 2, \ldots, T$
  Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \delta \nabla L(\mathbf{w}^{(t)})$
Return $w^{(T)}$



Gradient Descent

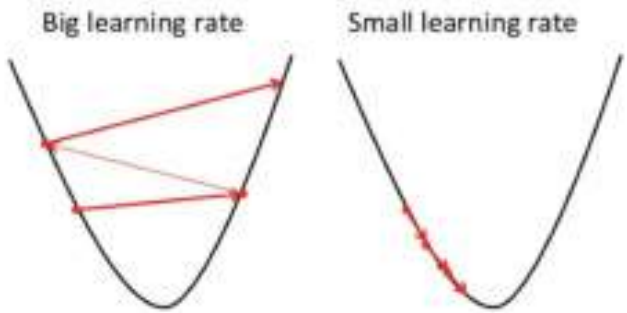
- $L : \mathbb{R}^{d+1} \to \mathbb{R}$ is any function in $\mathbf{w} = (w_0, w_1, \ldots, w_d)$ (loss function)
- The gradient is:

$$\nabla L(\mathbf{w}) = \left(\frac{dL(\mathbf{w})}{dw_0}, \frac{dL(\mathbf{w})}{dw_1}, \ldots, \frac{dL(\mathbf{w})}{dw_d} \right)$$

Gradient Descent for minimizing $L(\mathbf{w})$

Parameters: Learning Rate $\delta > 0$
Initialize: $\mathbf{w}^{(1)}$
for $t = 1, 2, \ldots, T$
  Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \delta \nabla L(\mathbf{w}^{(t)})$
Return $w^{(T)}$

Also possible to return average: $\frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)}$.



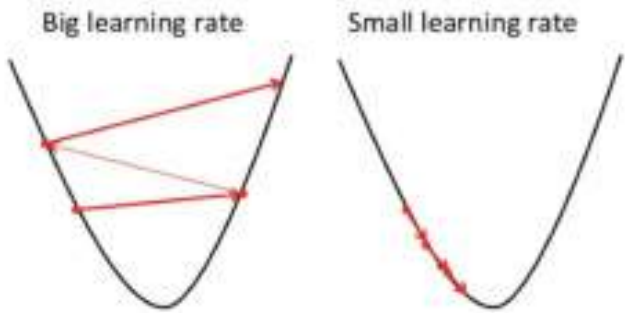
Gradient Descent

- $L : \mathbb{R}^{d+1} \to \mathbb{R}$ is any function in $\mathbf{w} = (w_0, w_1, \ldots, w_d)$ (loss function)
- The gradient is:

$$\nabla L(\mathbf{w}) = \left(\frac{dL(\mathbf{w})}{dw_0}, \frac{dL(\mathbf{w})}{dw_1}, \ldots, \frac{dL(\mathbf{w})}{dw_d} \right)$$

Gradient Descent for minimizing $L(\mathbf{w})$

Parameters: Learning Rate $\delta > 0$
Initialize: $\mathbf{w}^{(1)}$
for $t = 1, 2, \ldots, T$
  Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \delta \nabla L(\mathbf{w}^{(t)})$
Return $w^{(T)}$



Also possible to return average: $\frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)}$.

Problem: In every iteration, we need to recompute $L(\mathbf{w}^{(t)})$, a function over all training examples!

Stochastic Gradient Descent

Stochastic Gradient Descent for Loss Functions

Parameters: Learning rate $\delta > 0$, Loss function $L(\mathbf{w}) = \sum_{i=1}^{n} L_i(\mathbf{w})$
Initialize: $\mathbf{w}^{(1)}$
for $t = 1, 2, \ldots, T$
  Pick one training example $i = 1, 2, \ldots, n$ uniformly at random
  Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \delta \nabla L_i(\mathbf{w}^{(t)}, b)$
Return $w^{(T)}$

Stochastic Gradient Descent

Stochastic Gradient Descent for Loss Functions

Parameters: Learning rate $\delta > 0$, Loss function $L(\mathbf{w}) = \sum_{i=1}^{n} L_i(\mathbf{w})$
Initialize: $\mathbf{w}^{(1)}$
for $t = 1, 2, \ldots, T$
  Pick one training example $i = 1, 2, \ldots, n$ uniformly at random
  Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \delta \nabla L_i(\mathbf{w}^{(t)}, b)$
Return $w^{(T)}$

- In practice, stochastic gradient descent is often executed in epochs:
  - in each epoch, first shuffle all m training examples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$
  - then perform gradient descent on f with training example $\mathbf{x}_i$ only

Stochastic Gradient Descent

Stochastic Gradient Descent for Loss Functions

Parameters: Learning rate $\delta > 0$, Loss function $L(\mathbf{w}) = \sum_{i=1}^{n} L_i(\mathbf{w})$
Initialize: $\mathbf{w}^{(1)}$
for $t = 1, 2, \ldots, T$
  Pick one training example $i = 1, 2, \ldots, n$ uniformly at random
  Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \delta \nabla L_i(\mathbf{w}^{(t)}, b)$
Return $w^{(T)}$

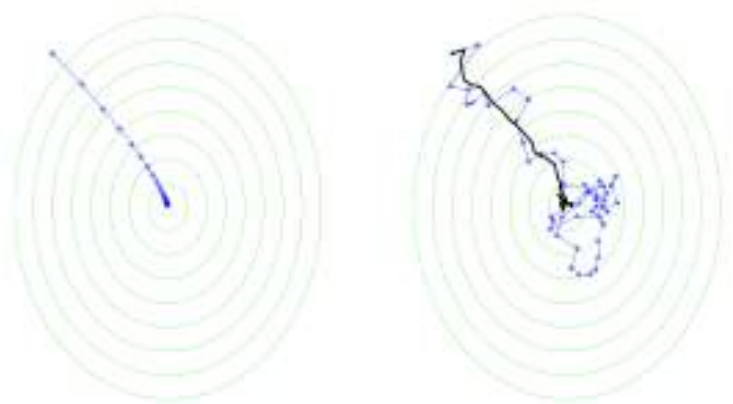
- In practice, stochastic gradient descent is often executed in epochs:
  - in each epoch, first shuffle all m training examples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$
  - then perform gradient descent on f with training example $\mathbf{x}_i$ only
- Also possible to return the uniform average over $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)}$:

Stochastic Gradient Descent

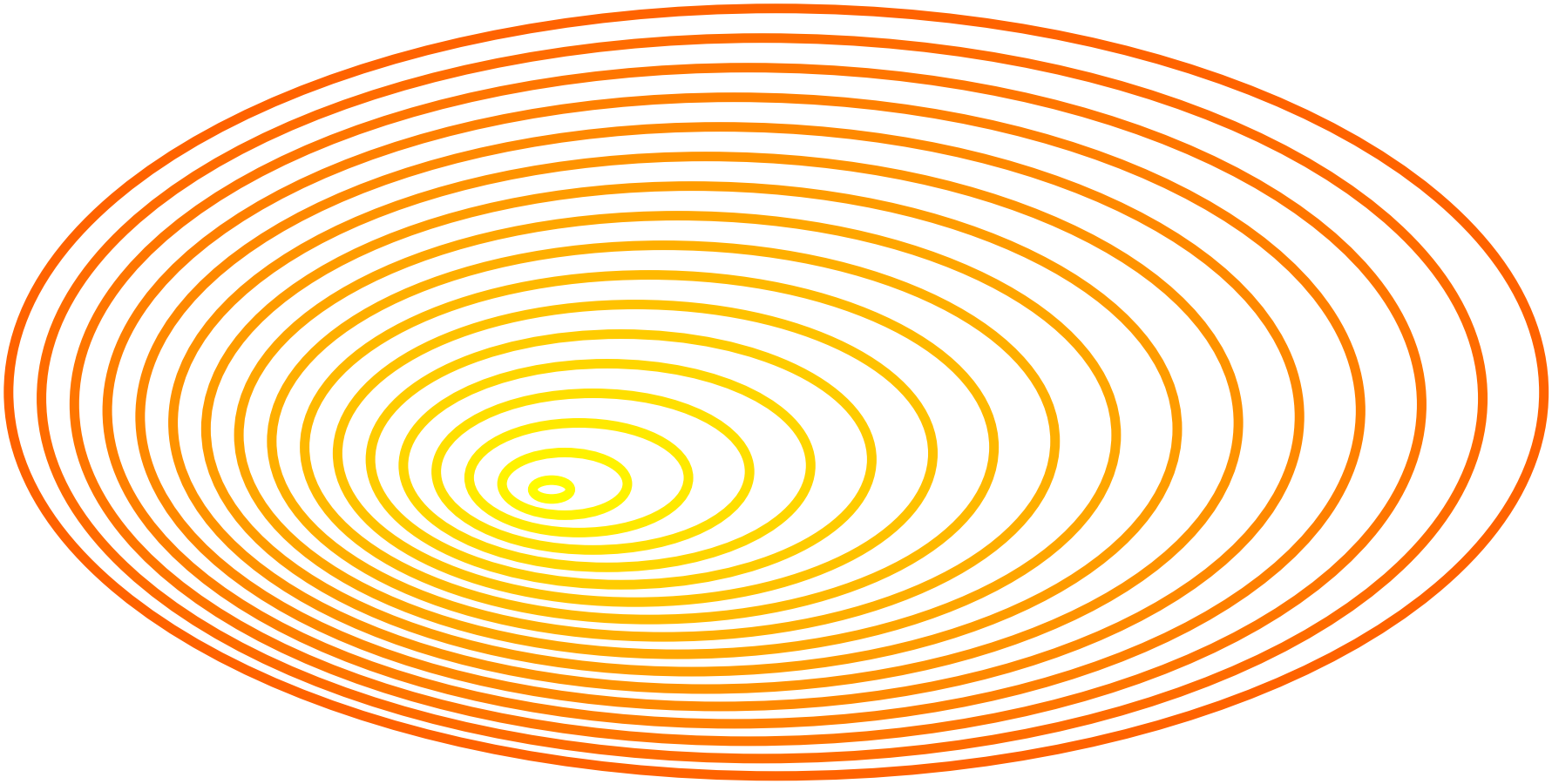
Stochastic Gradient Descent for Loss Functions

Parameters: Learning rate $\delta > 0$, Loss function $L(\mathbf{w}) = \sum_{i=1}^{n} L_i(\mathbf{w})$
Initialize: $\mathbf{w}^{(1)}$
for $t = 1, 2, \ldots, T$
  Pick one training example $i = 1, 2, \ldots, n$ uniformly at random
  Update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \delta \nabla L_i(\mathbf{w}^{(t)}, b)$
Return $w^{(T)}$

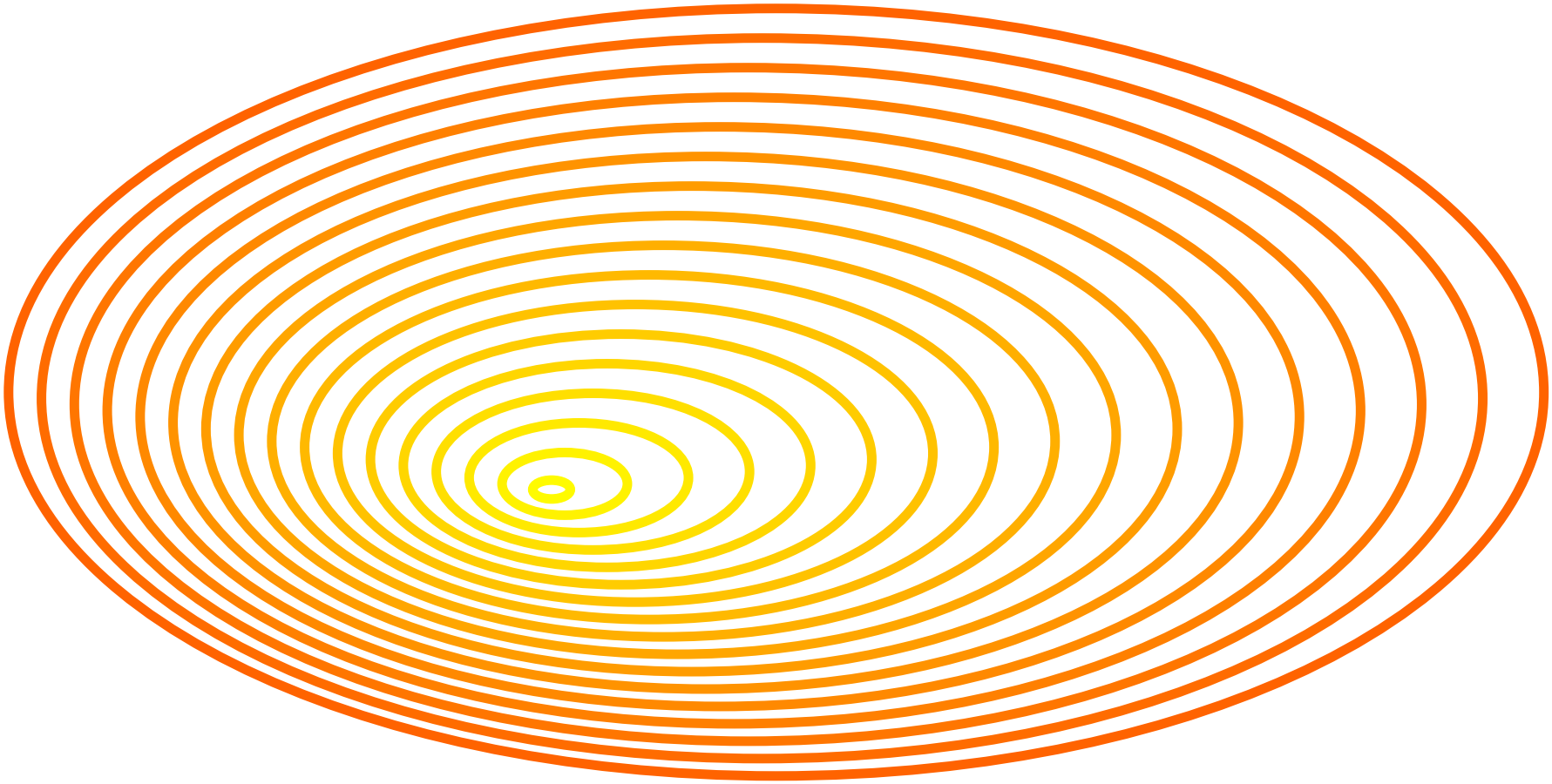
- In practice, stochastic gradient descent is often executed in epochs:
  - in each epoch, first shuffle all m training examples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$
  - then perform gradient descent on f with training example $\mathbf{x}_i$ only
- Also possible to return the uniform average over $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)}$:



Gradient versus Stochastic Gradient

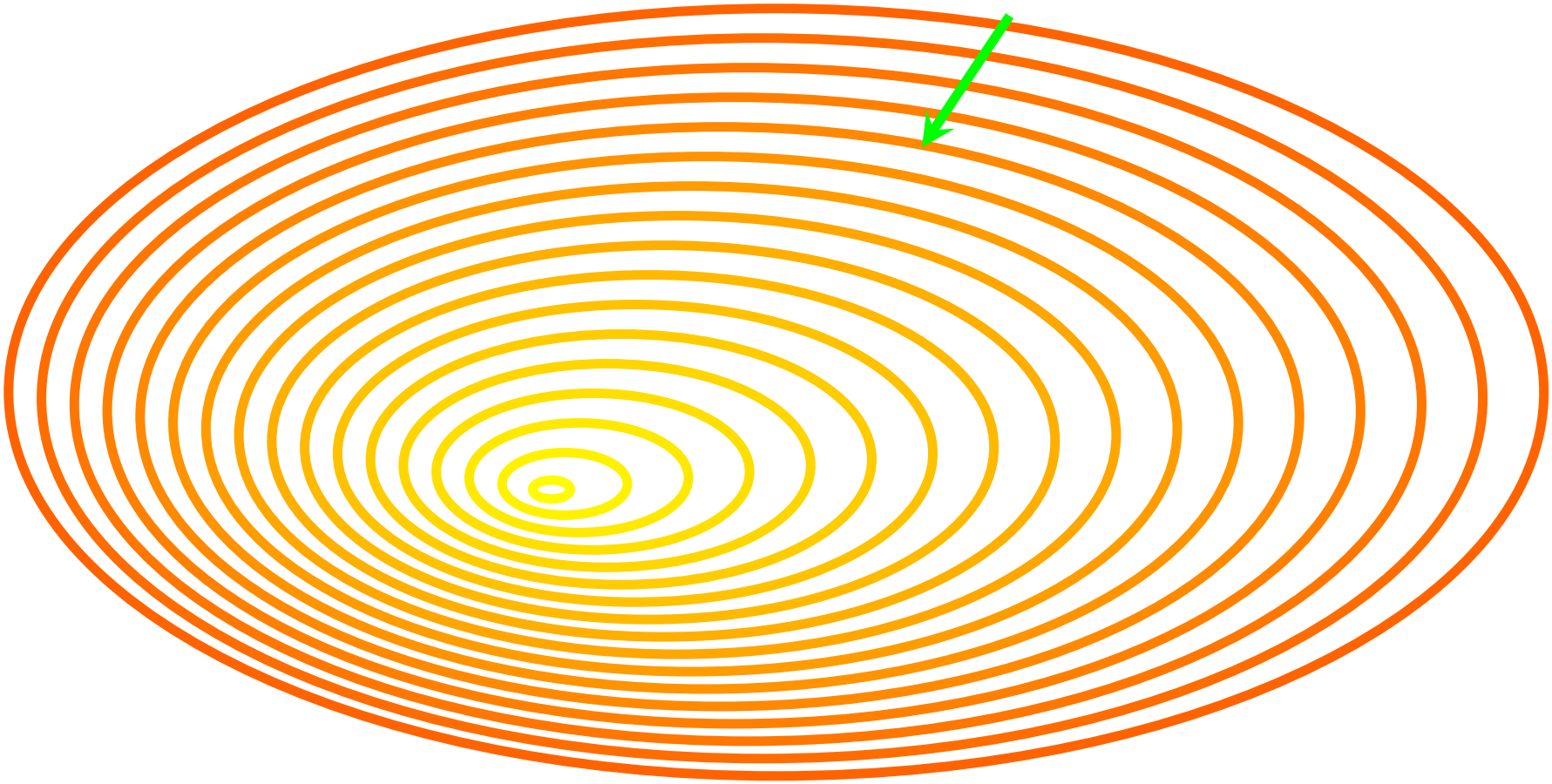


Gradient versus Stochastic Gradient



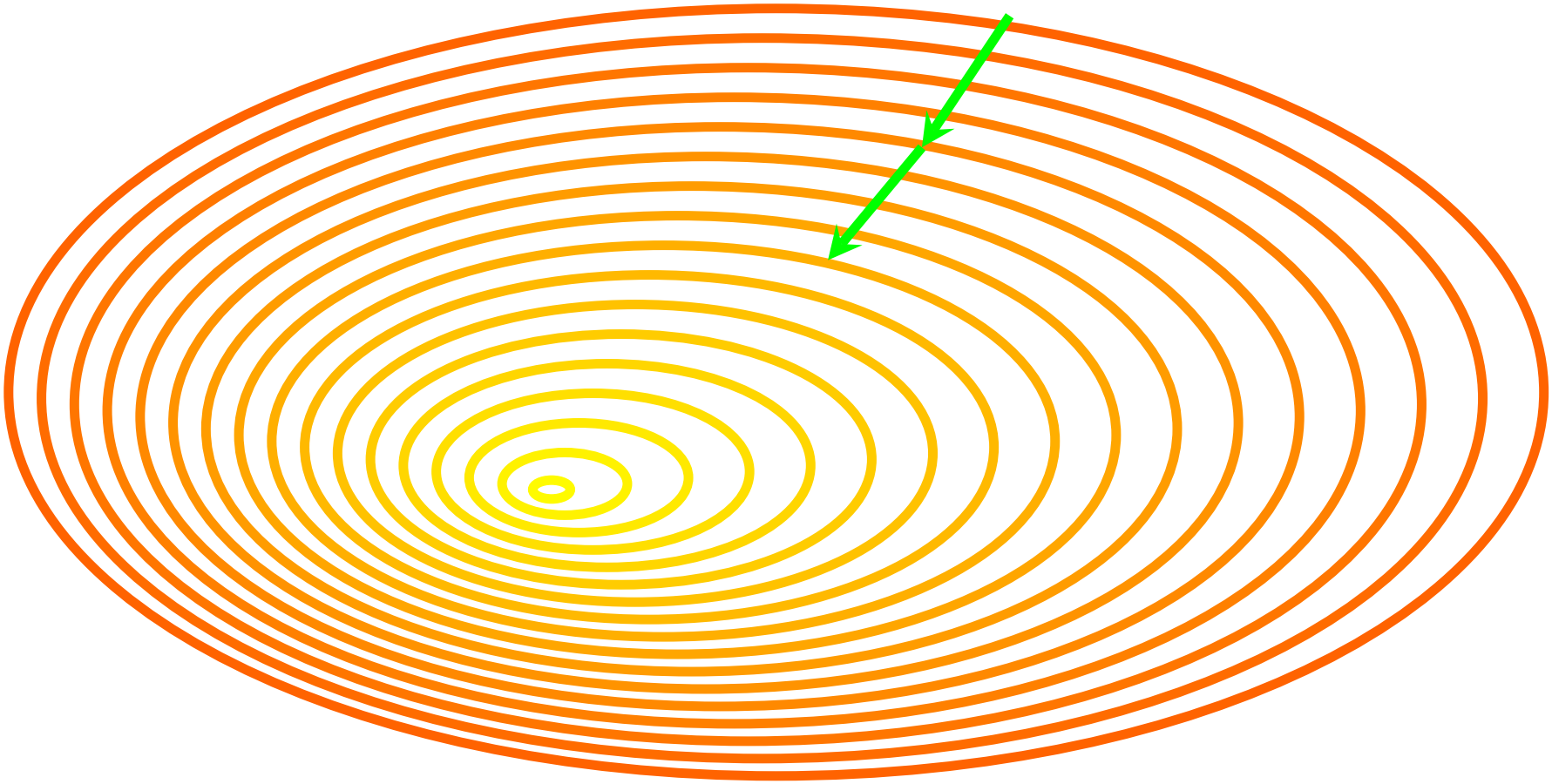
- Gradient Descent:

Gradient versus Stochastic Gradient



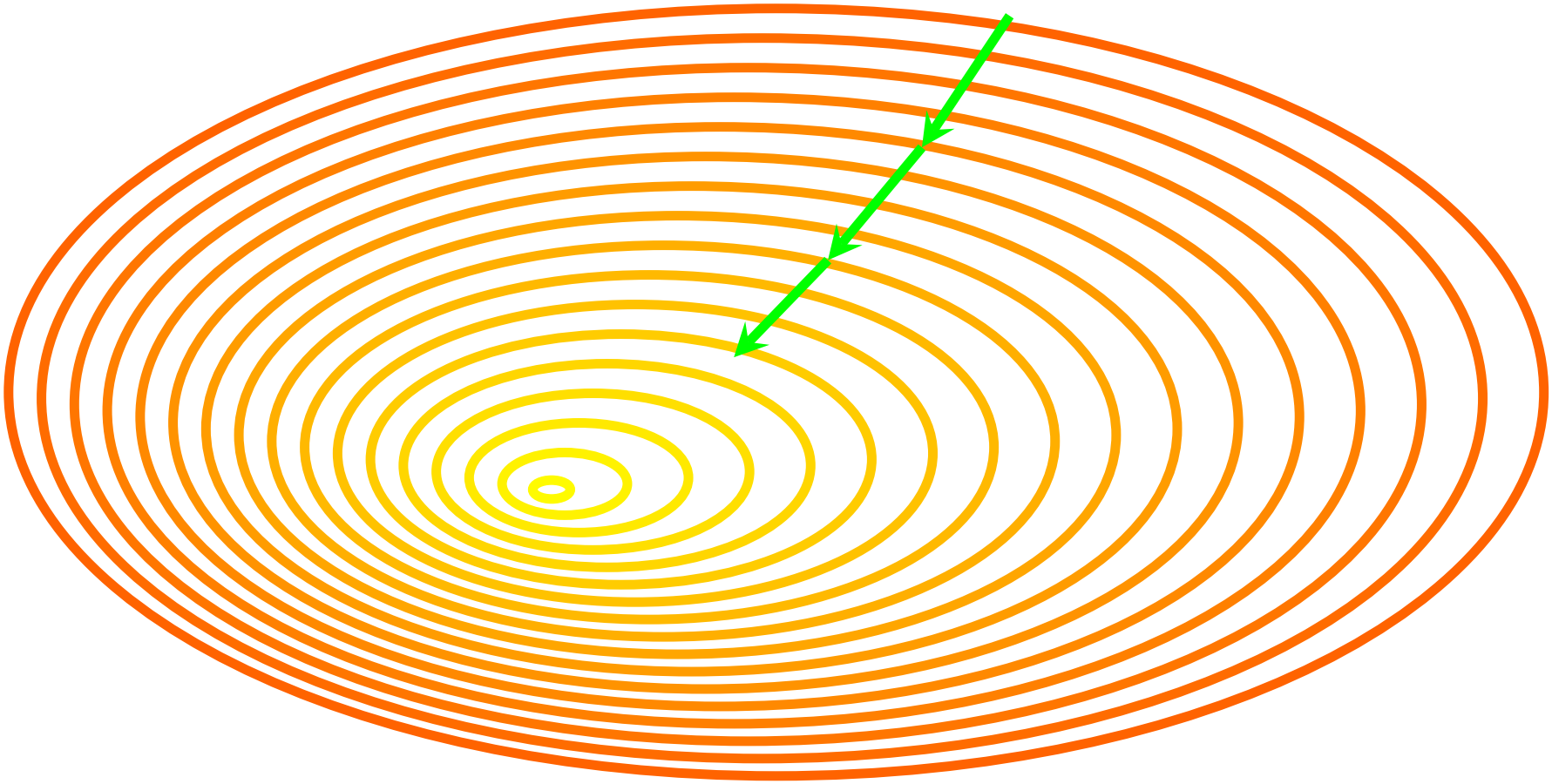
- Gradient Descent:

Gradient versus Stochastic Gradient



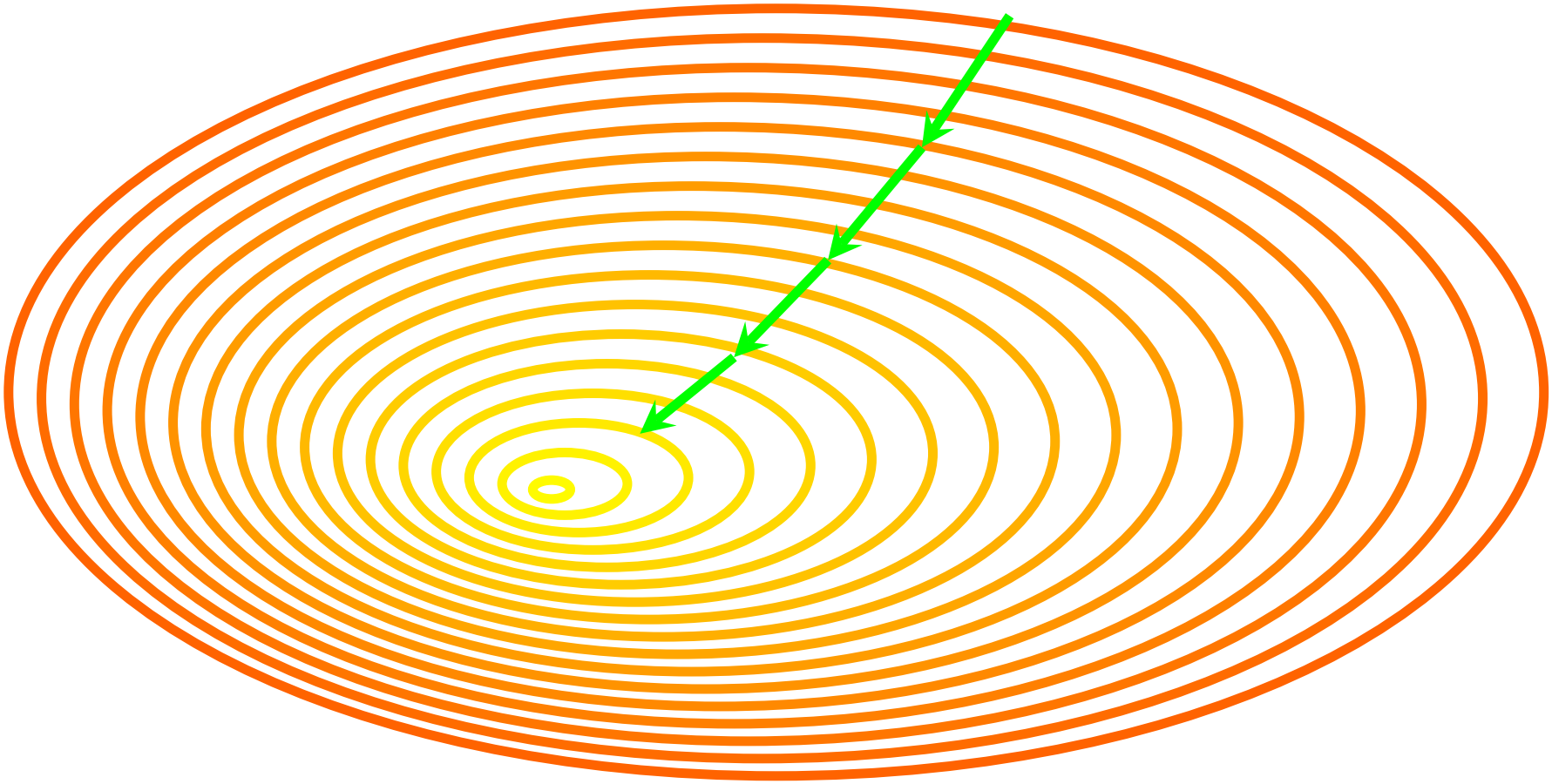
- Gradient Descent:

Gradient versus Stochastic Gradient



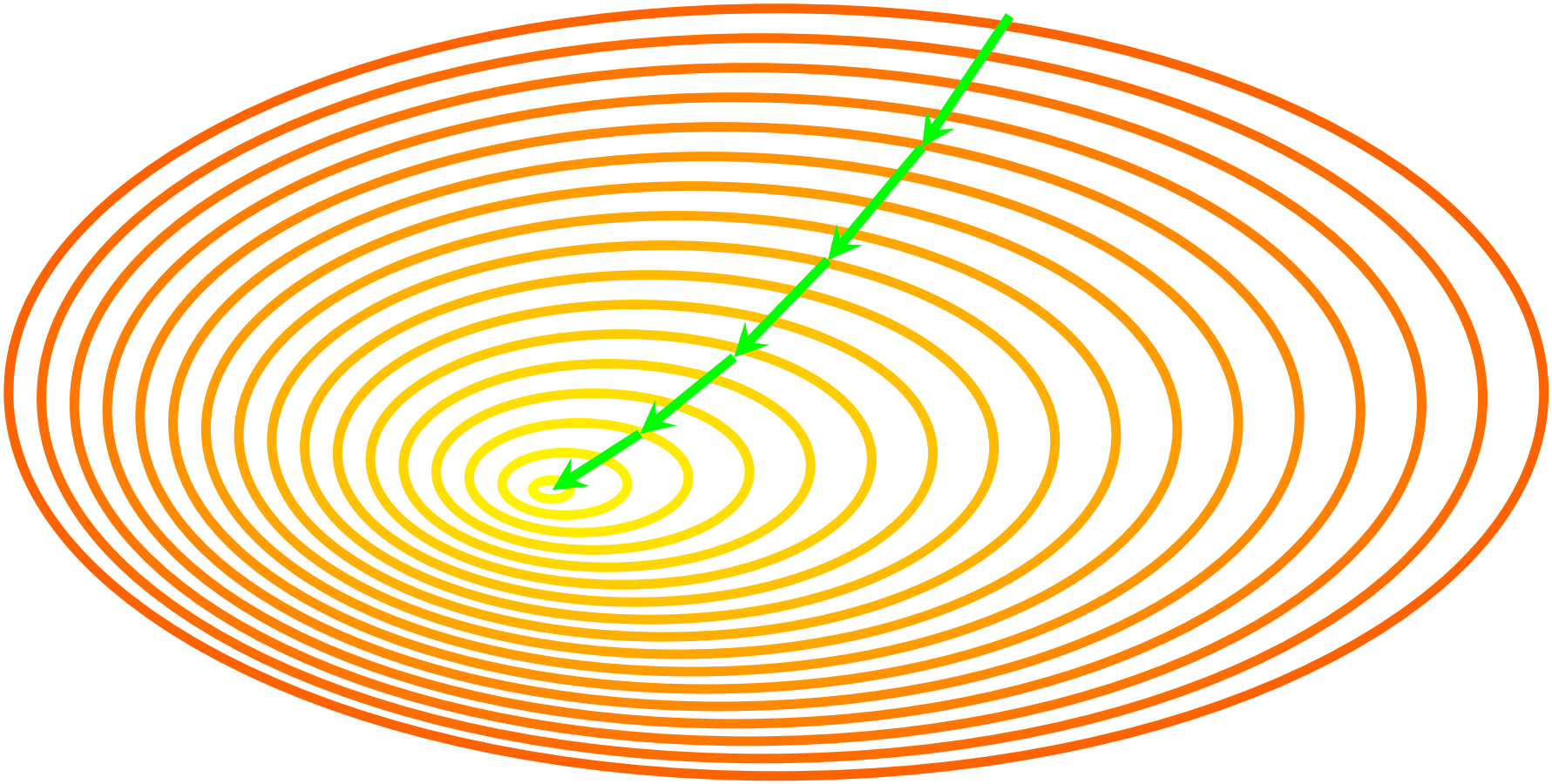
- Gradient Descent:

Gradient versus Stochastic Gradient



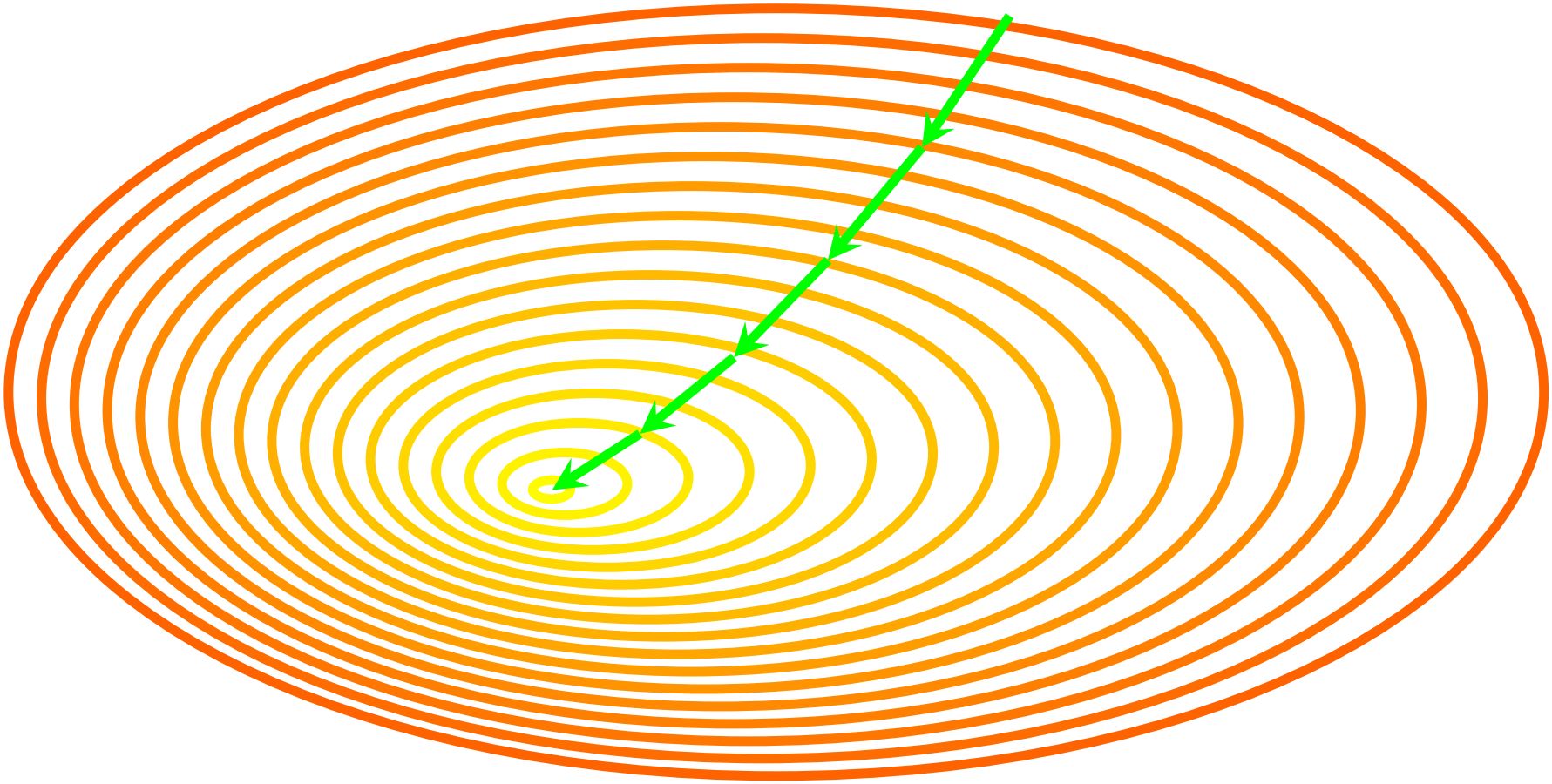
- Gradient Descent:

Gradient versus Stochastic Gradient



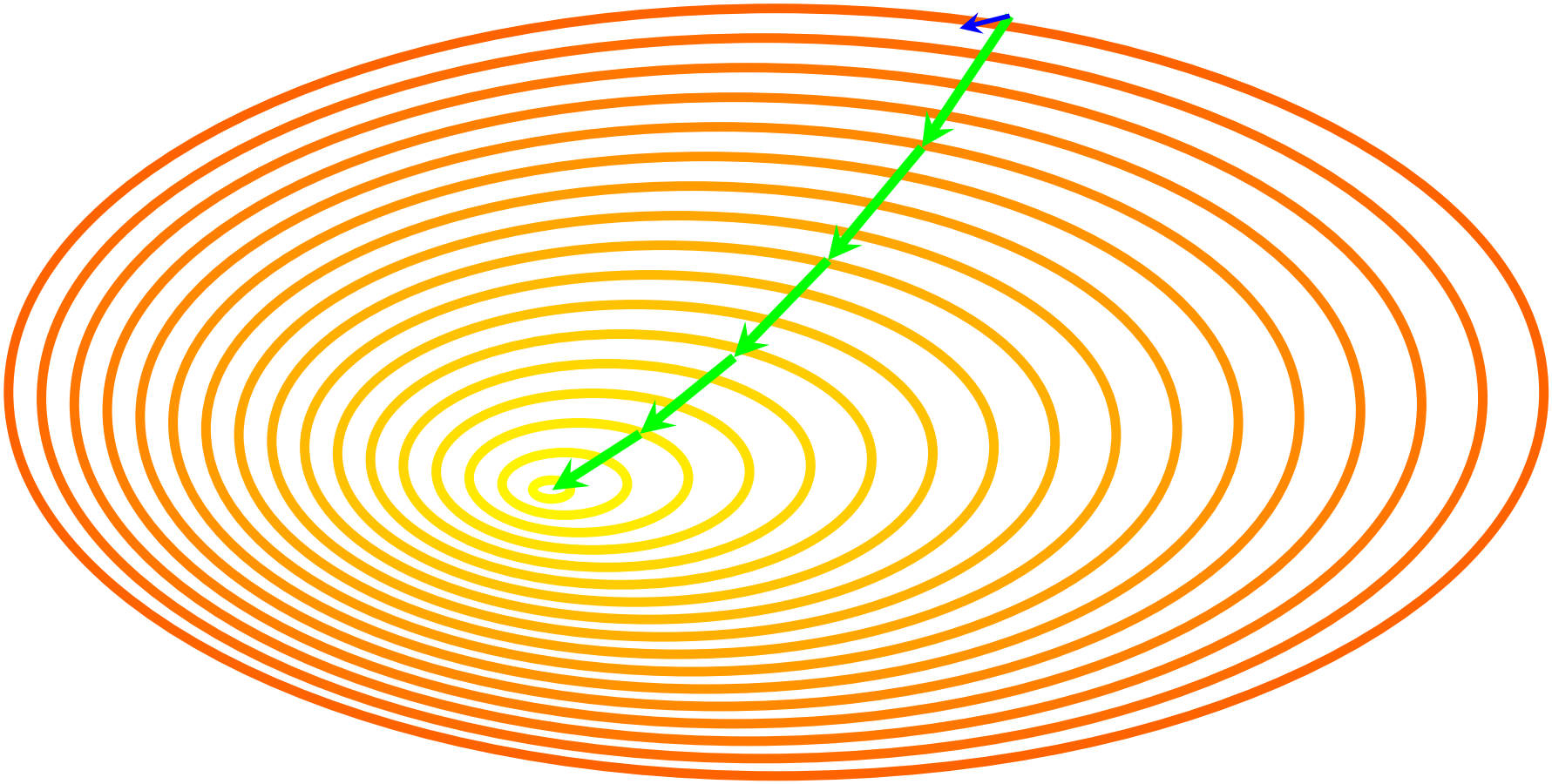
- Gradient Descent:

Gradient versus Stochastic Gradient



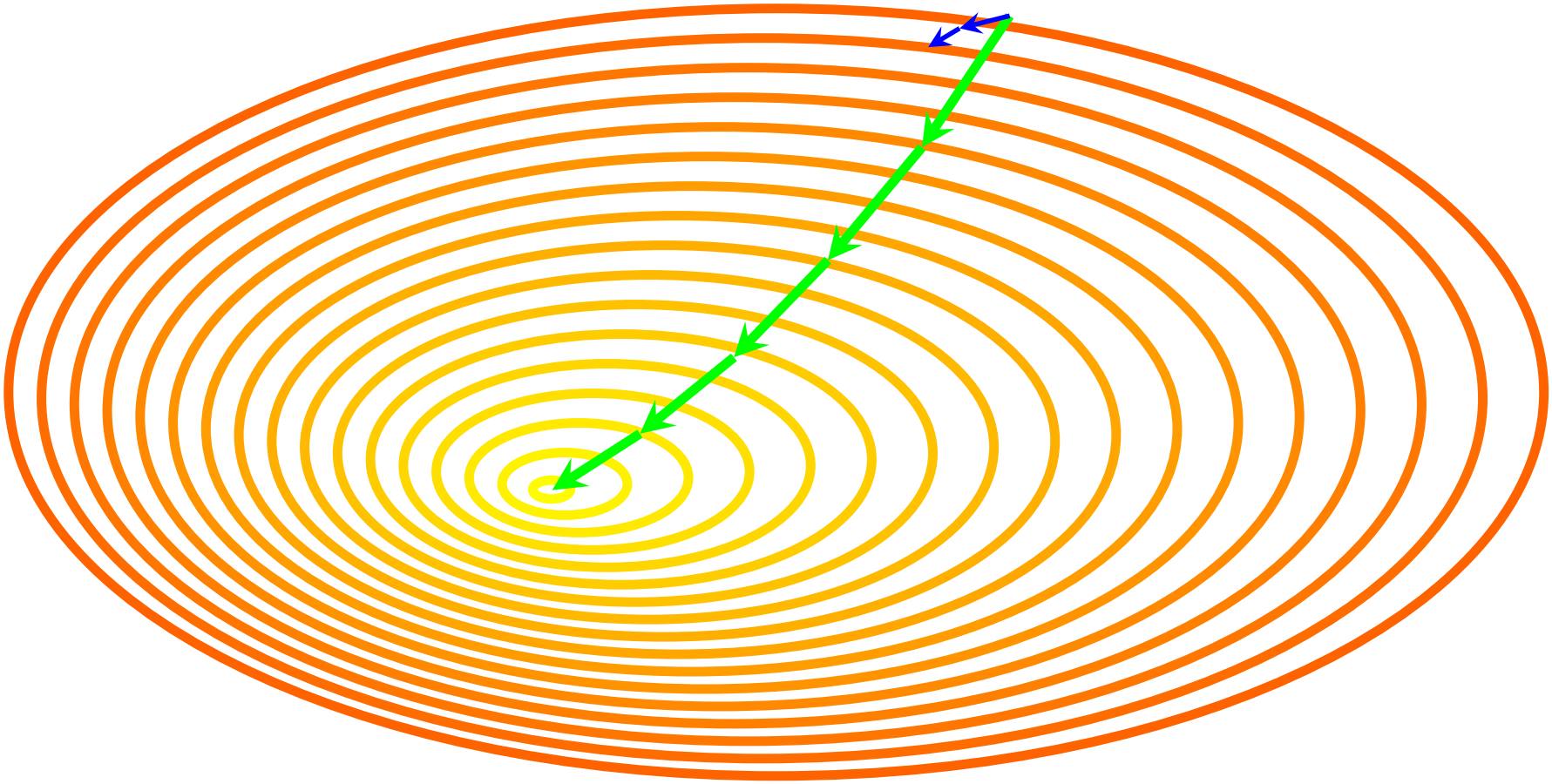
- Gradient Descent: converges quickly and smoothly (usually).

Gradient versus Stochastic Gradient



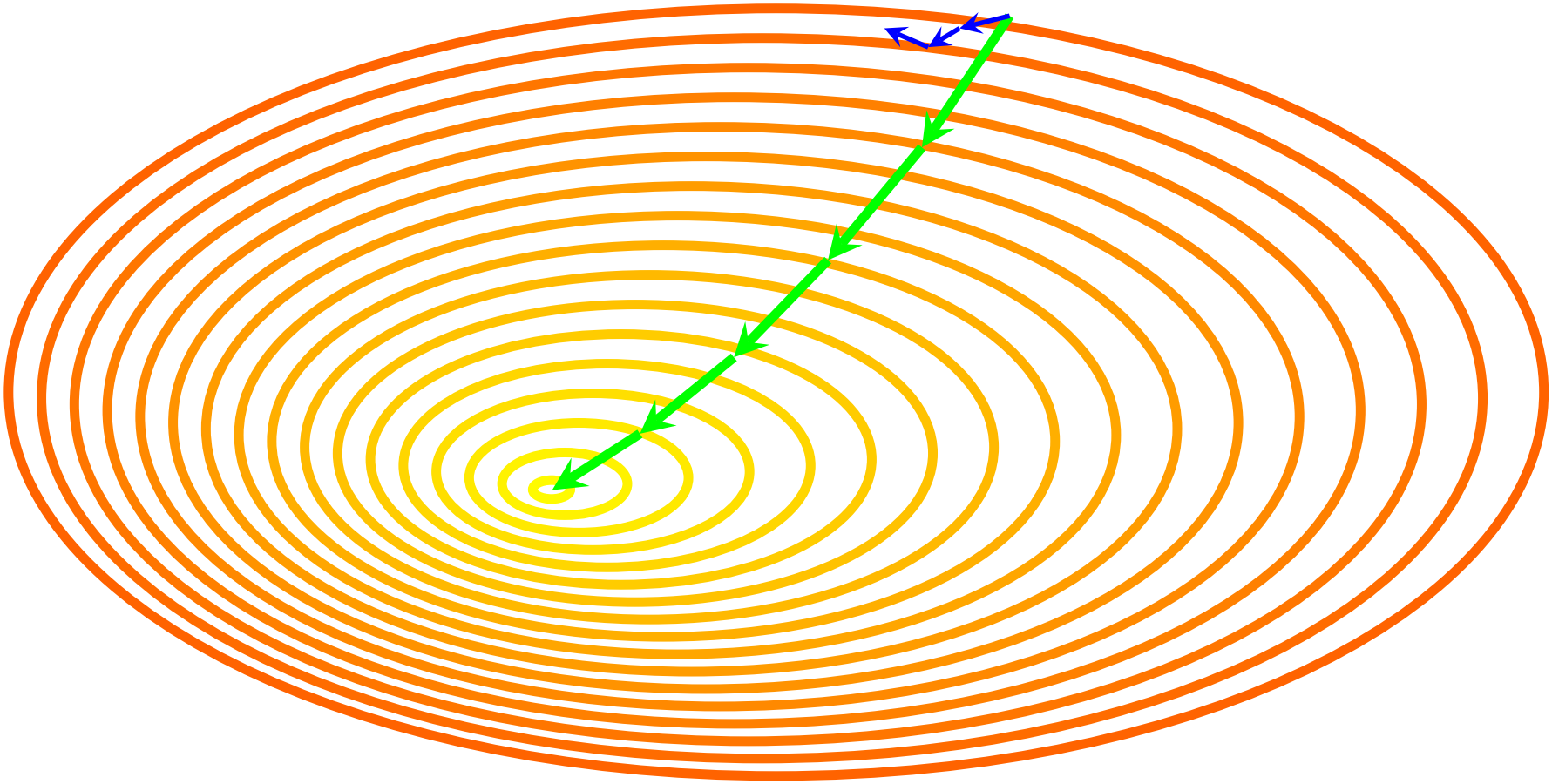
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



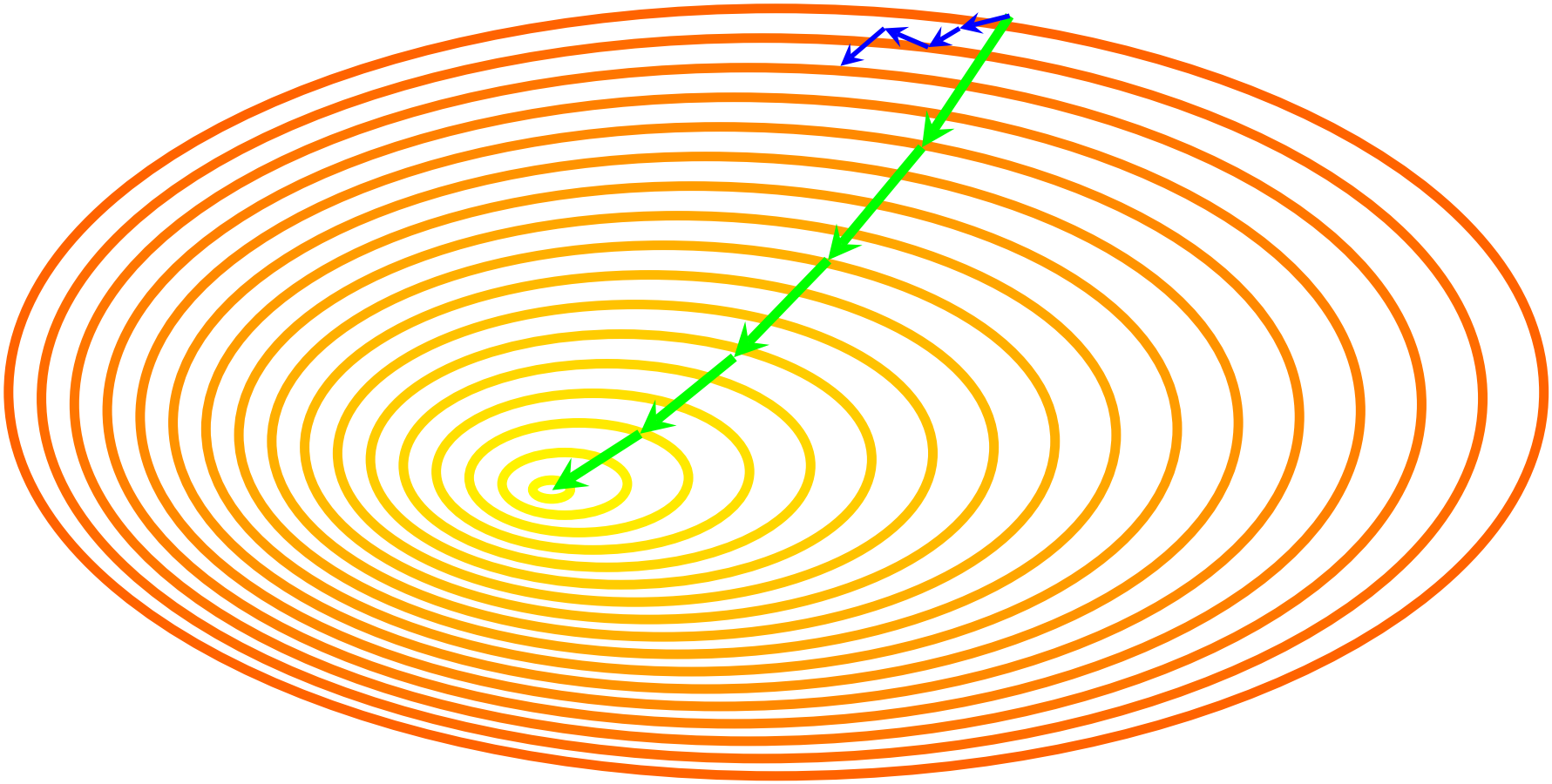
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



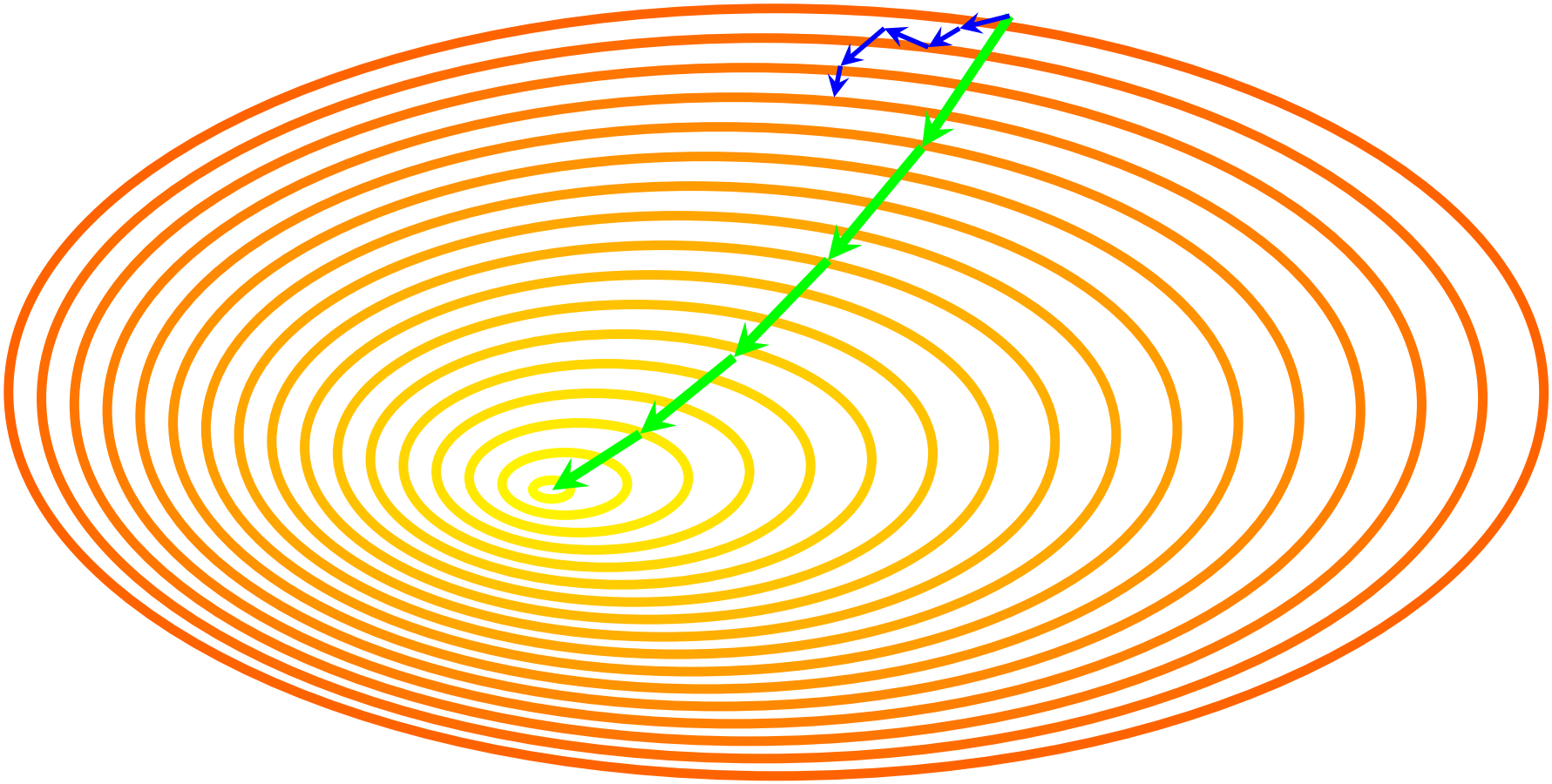
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



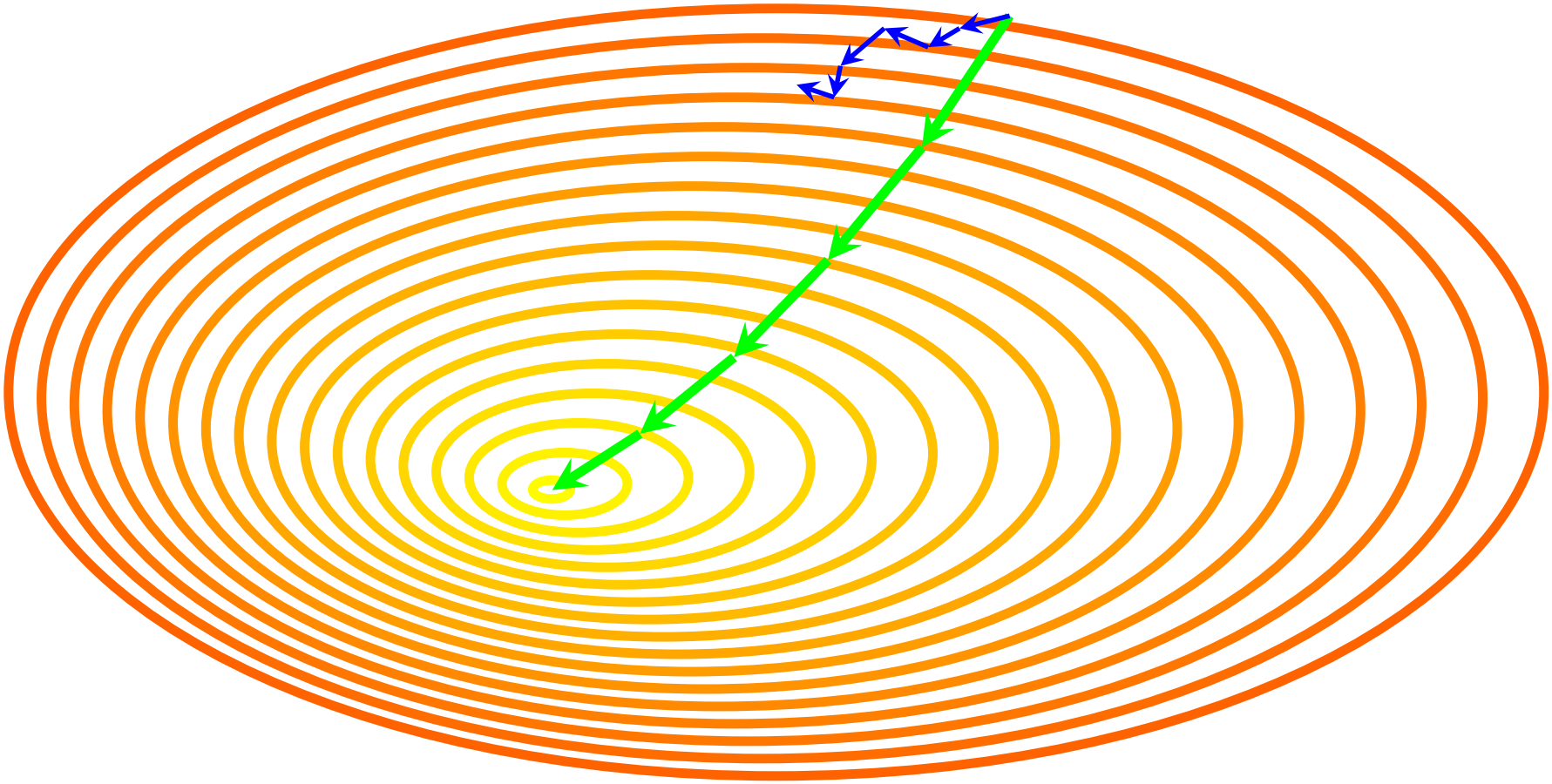
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



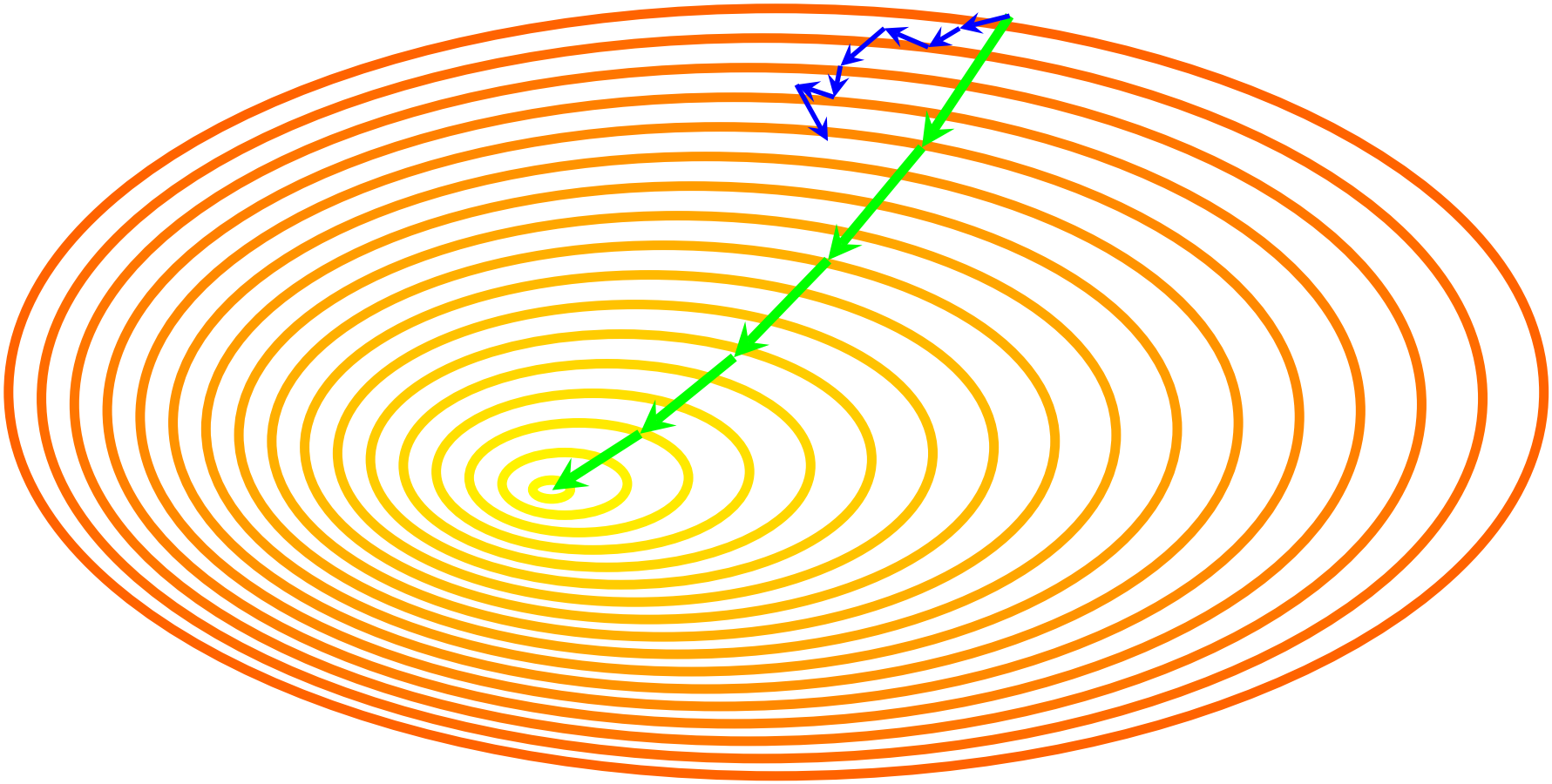
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



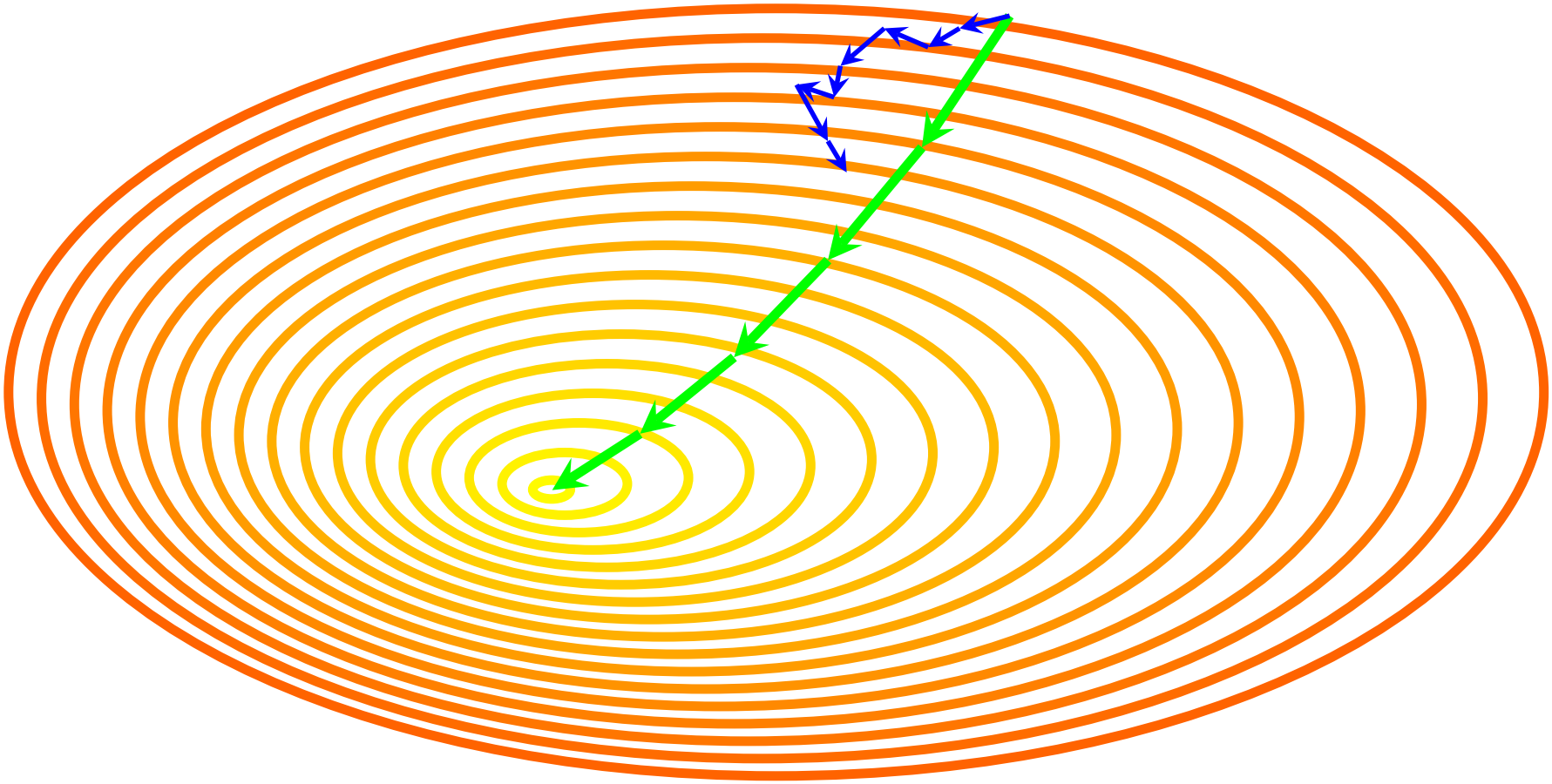
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



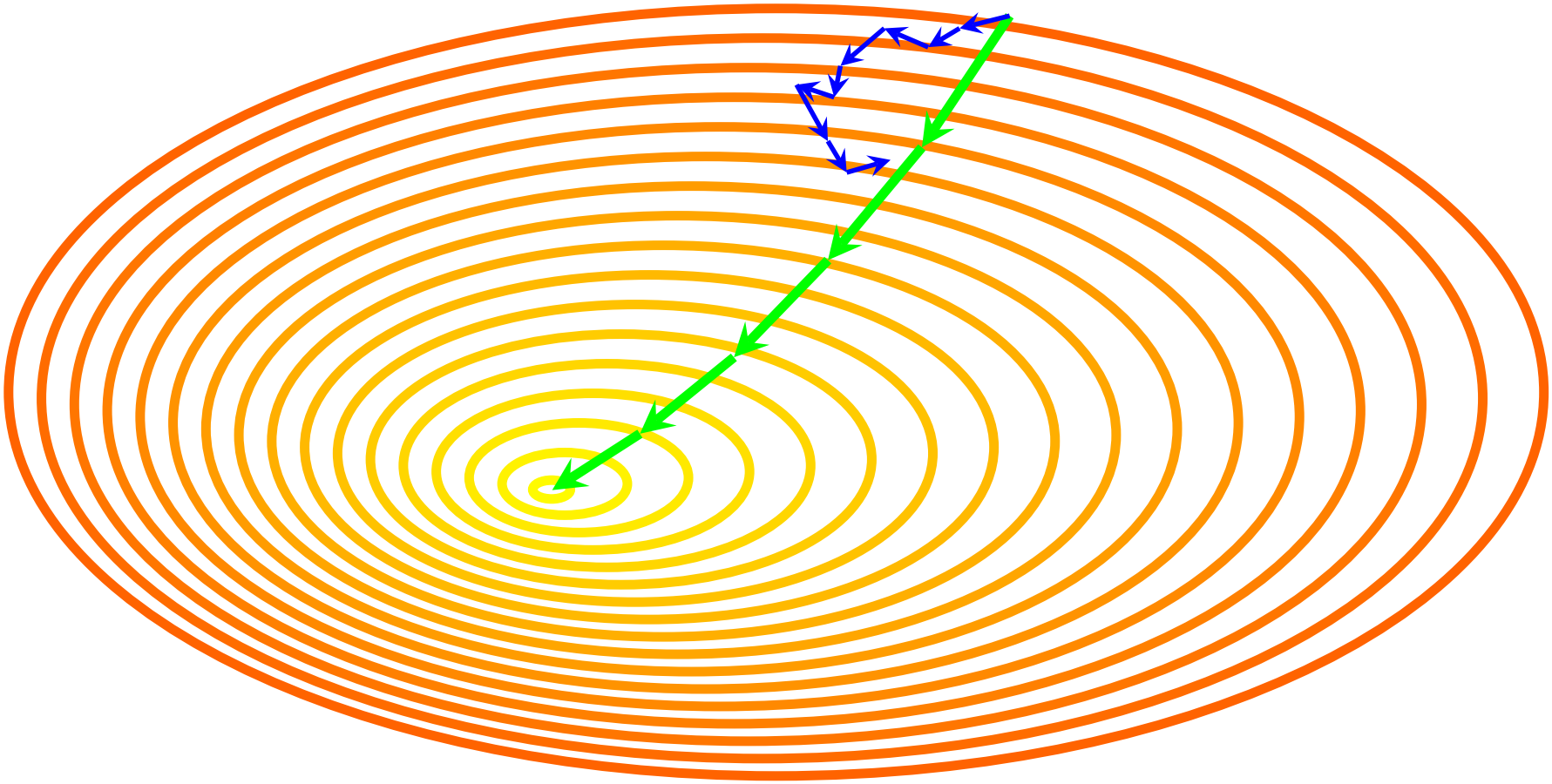
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



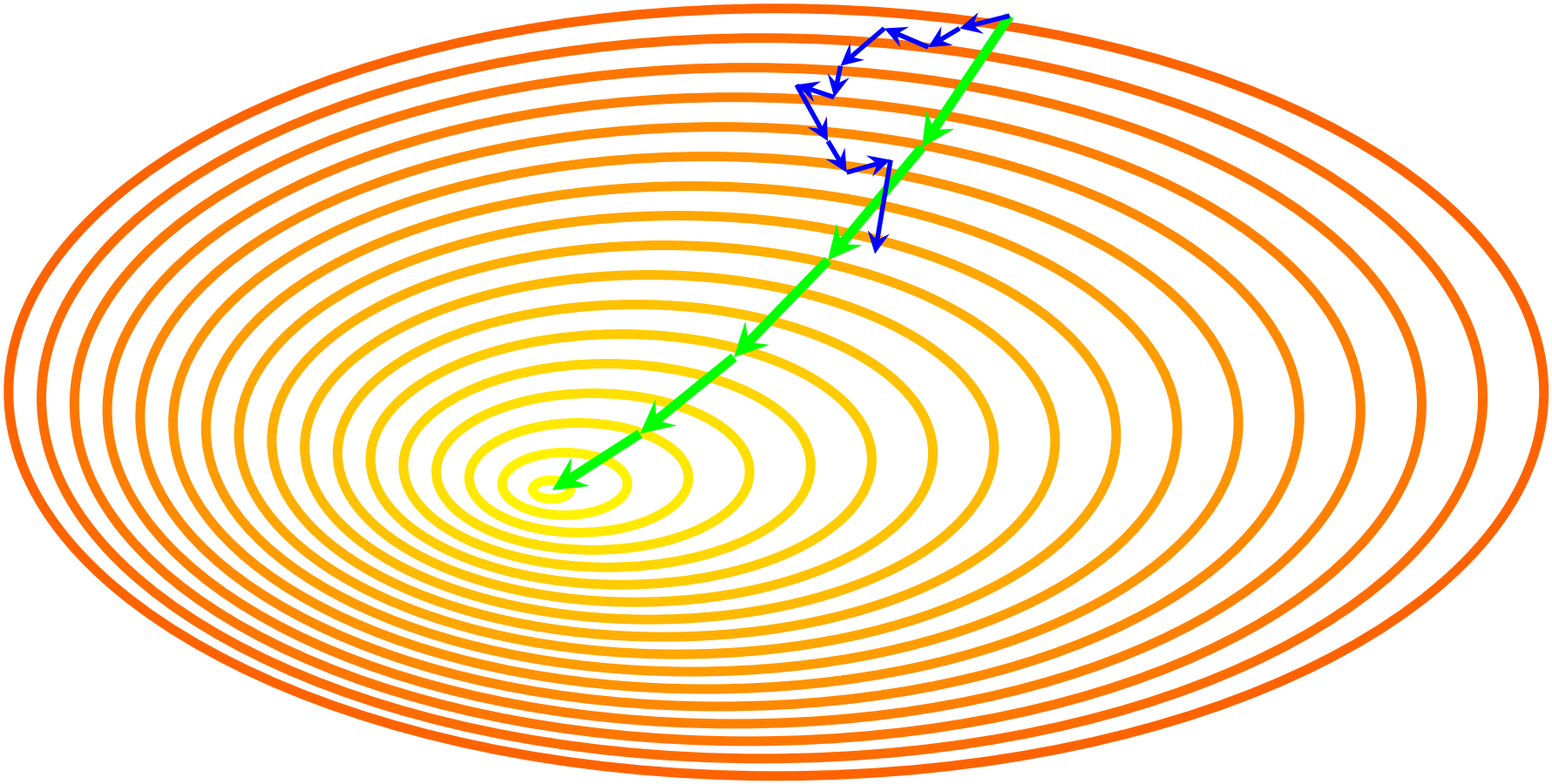
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



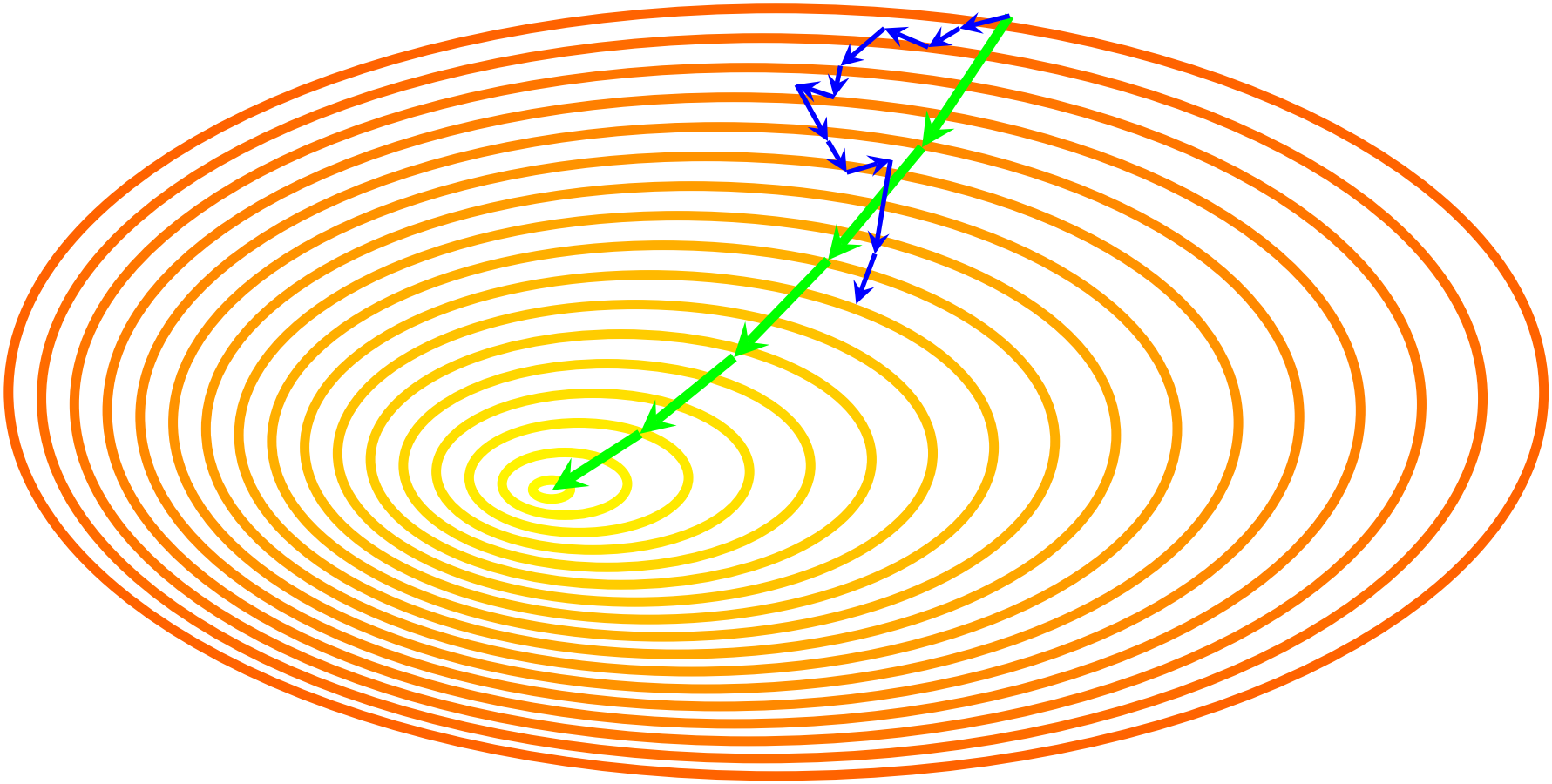
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



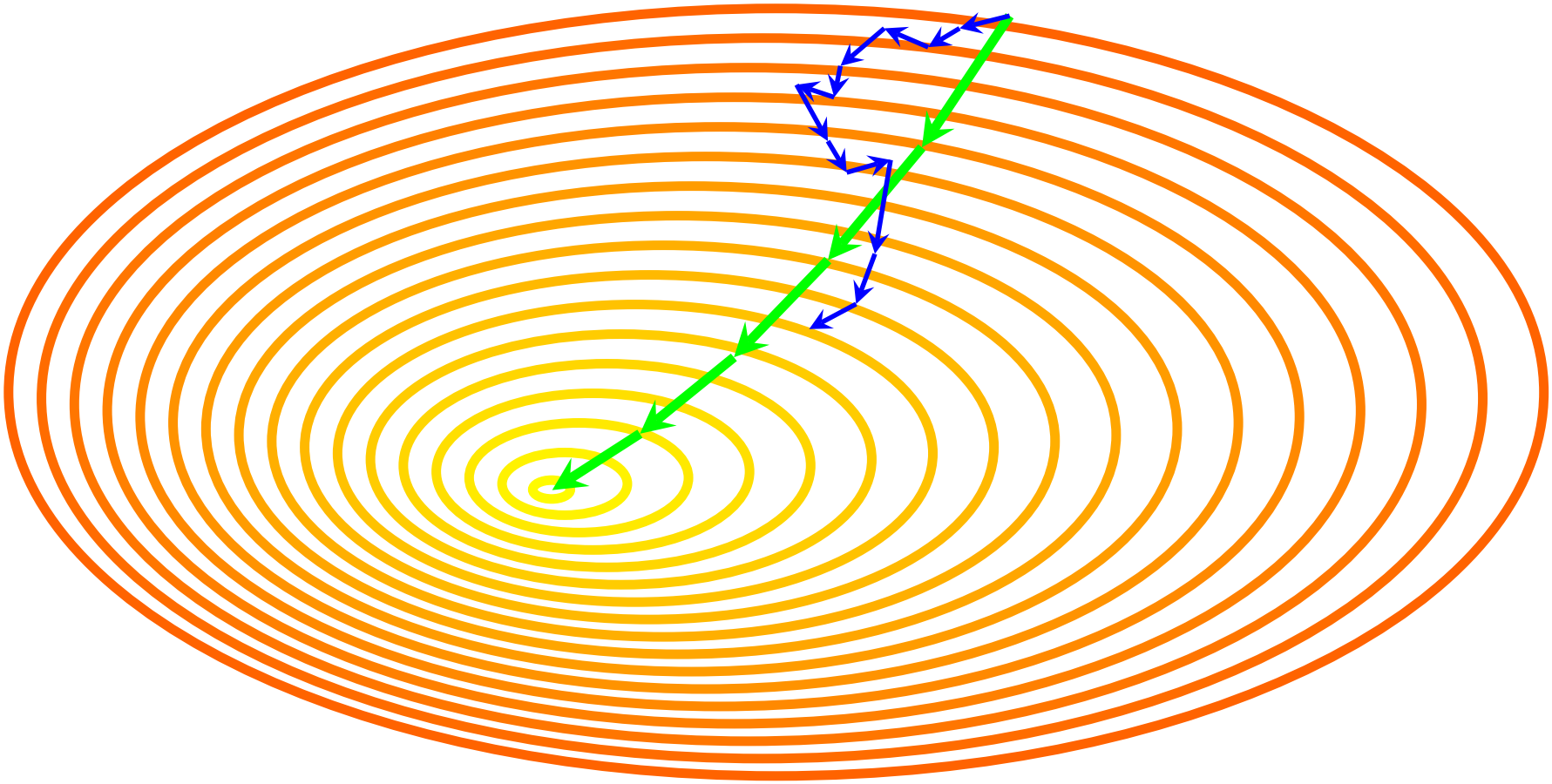
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



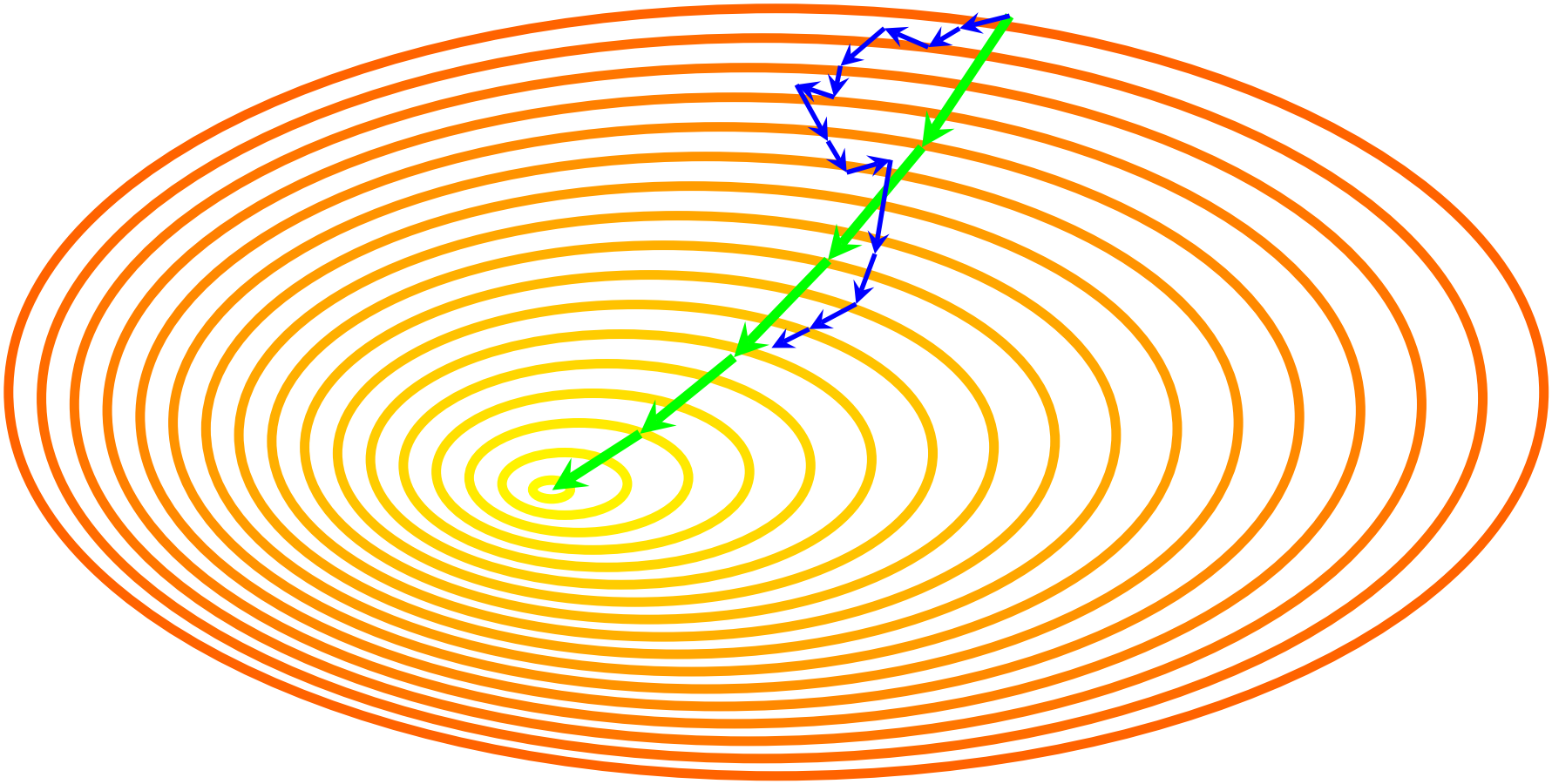
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



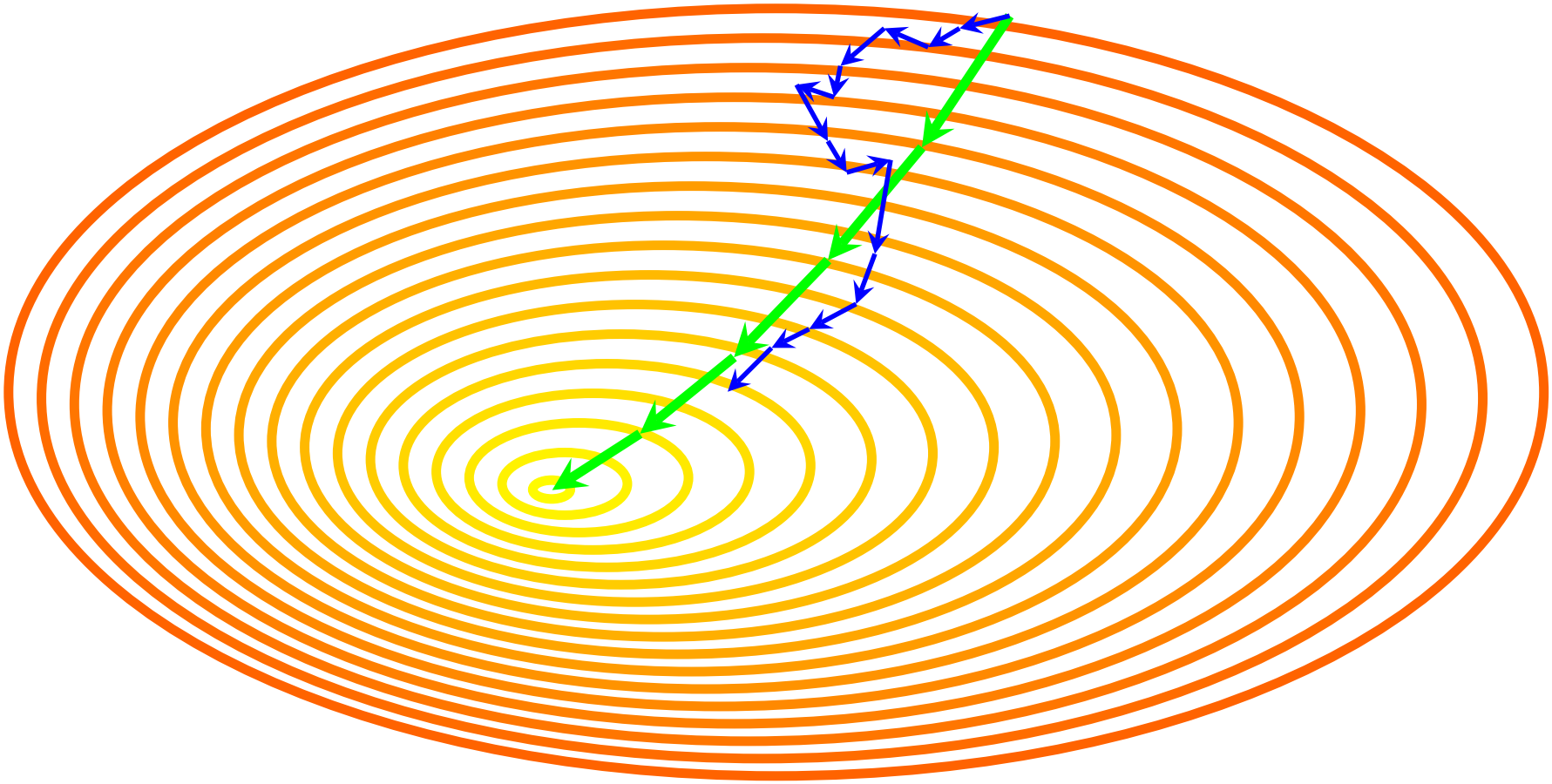
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



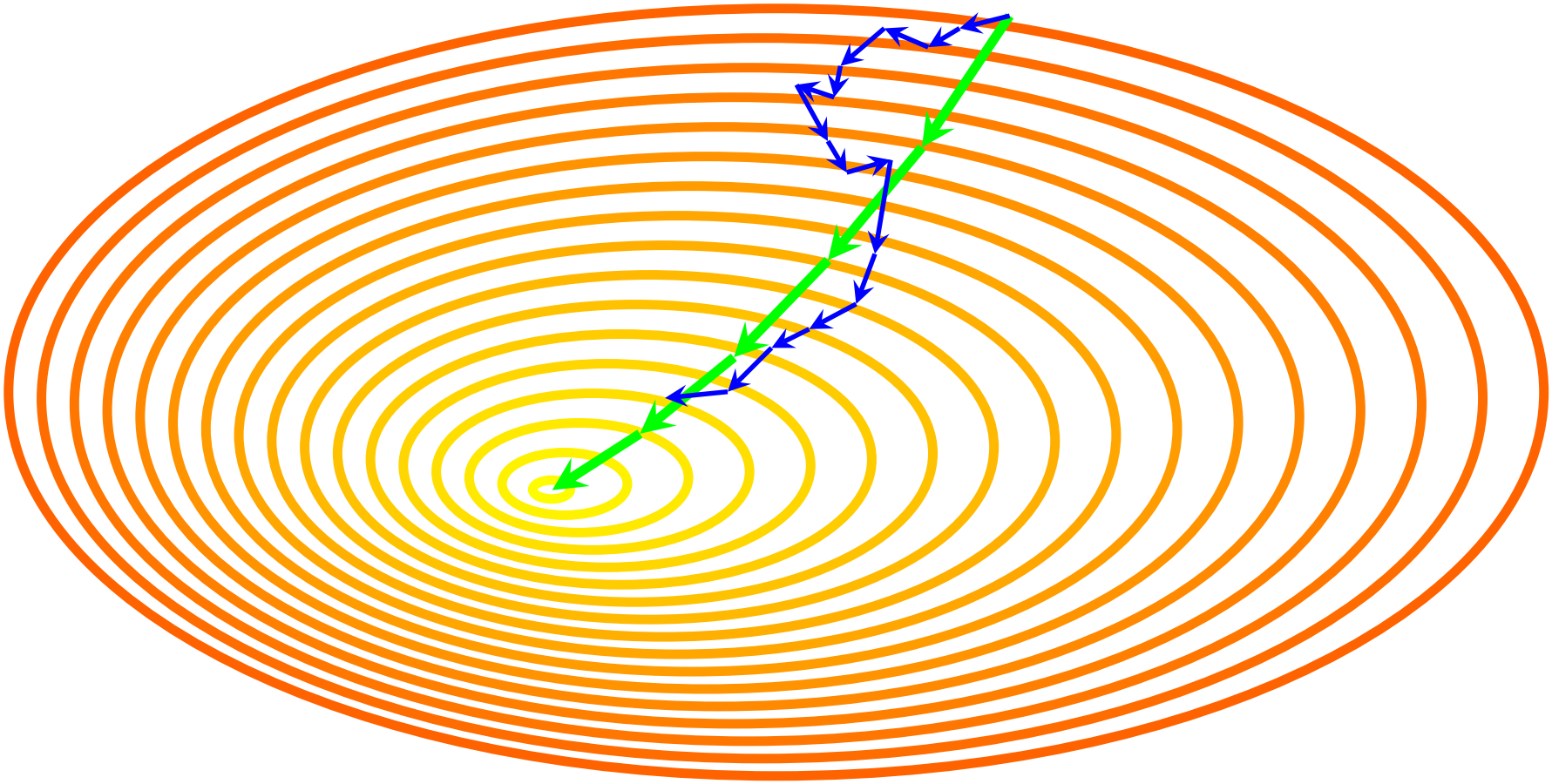
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



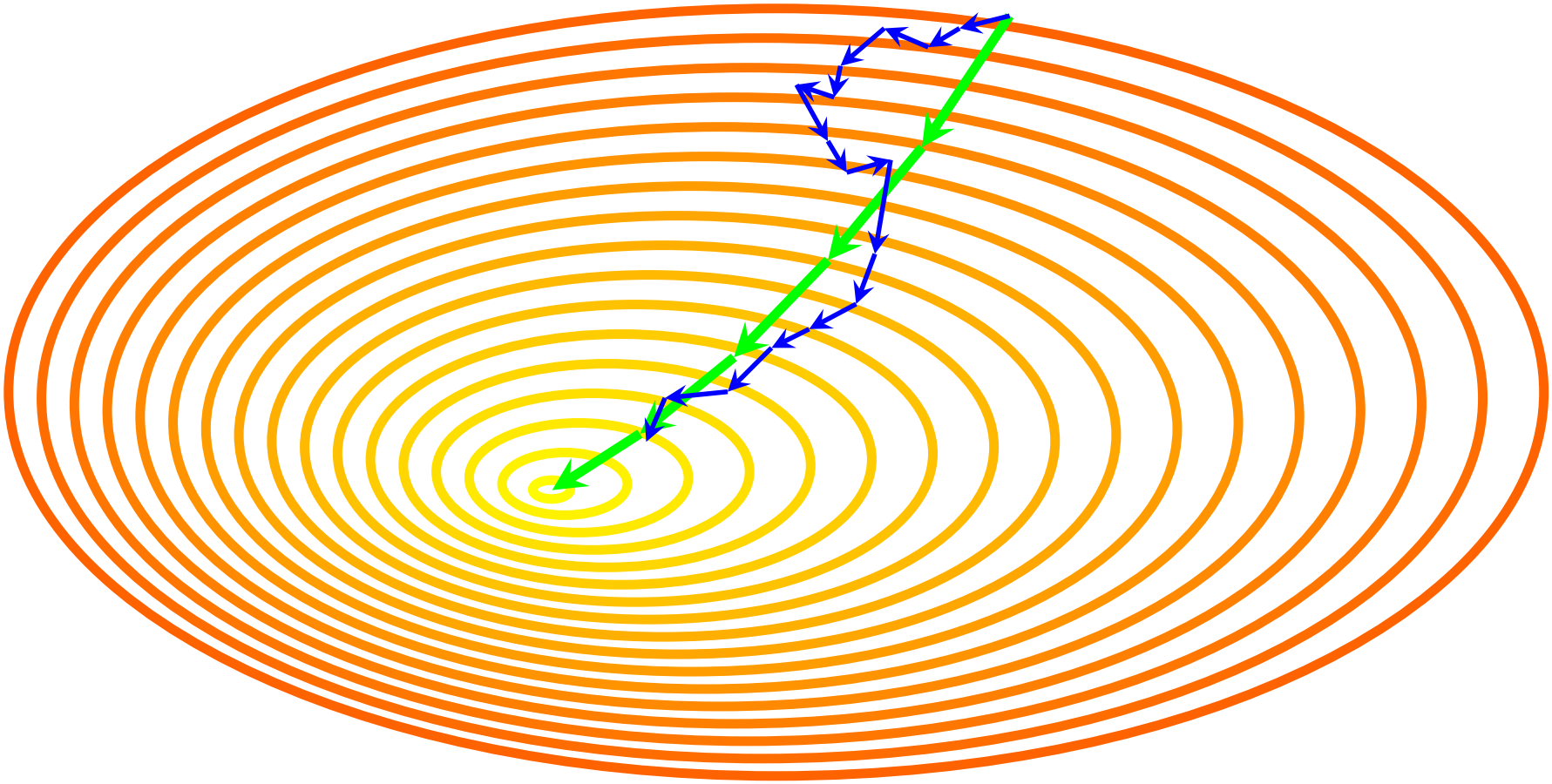
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



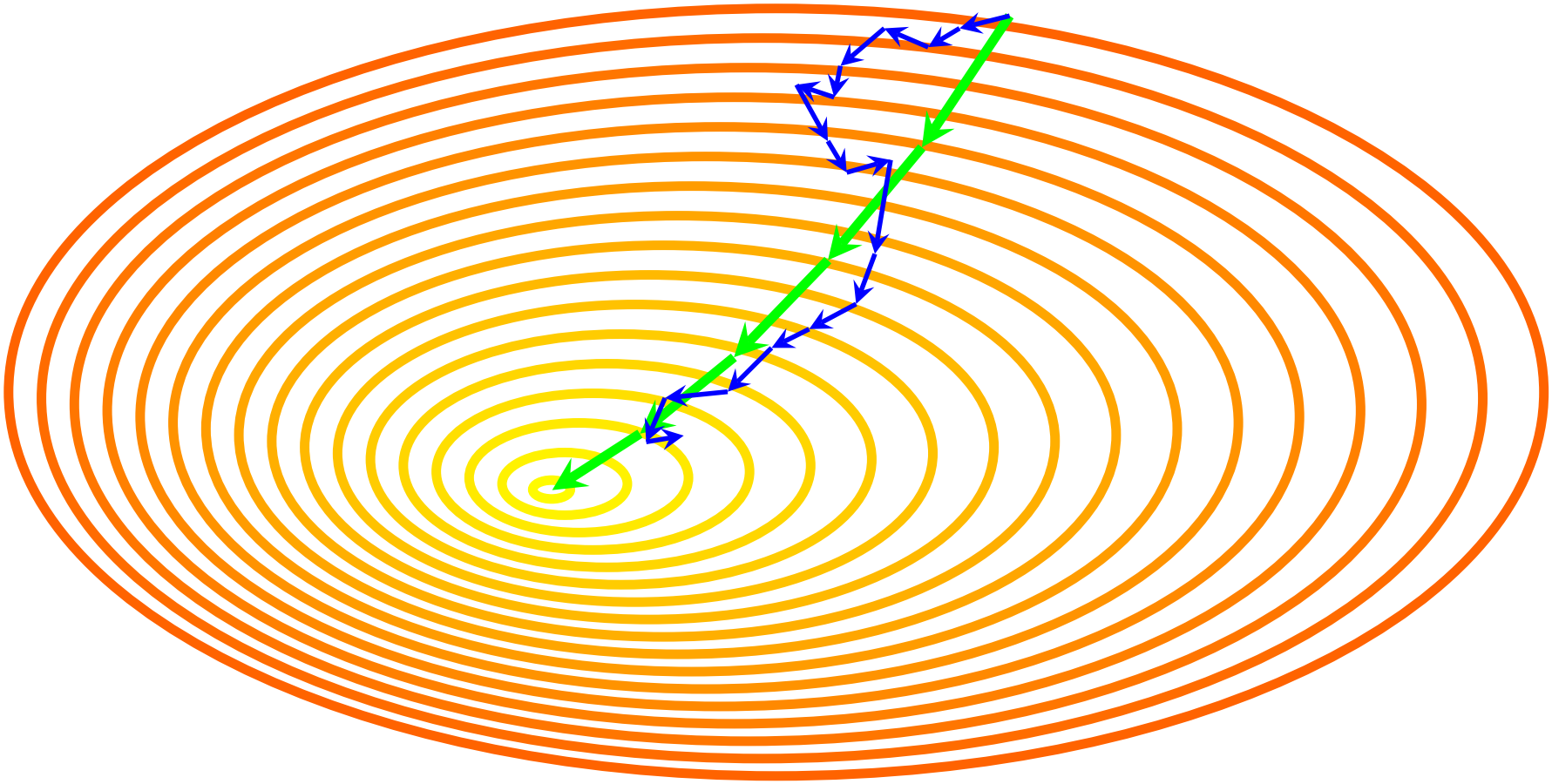
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



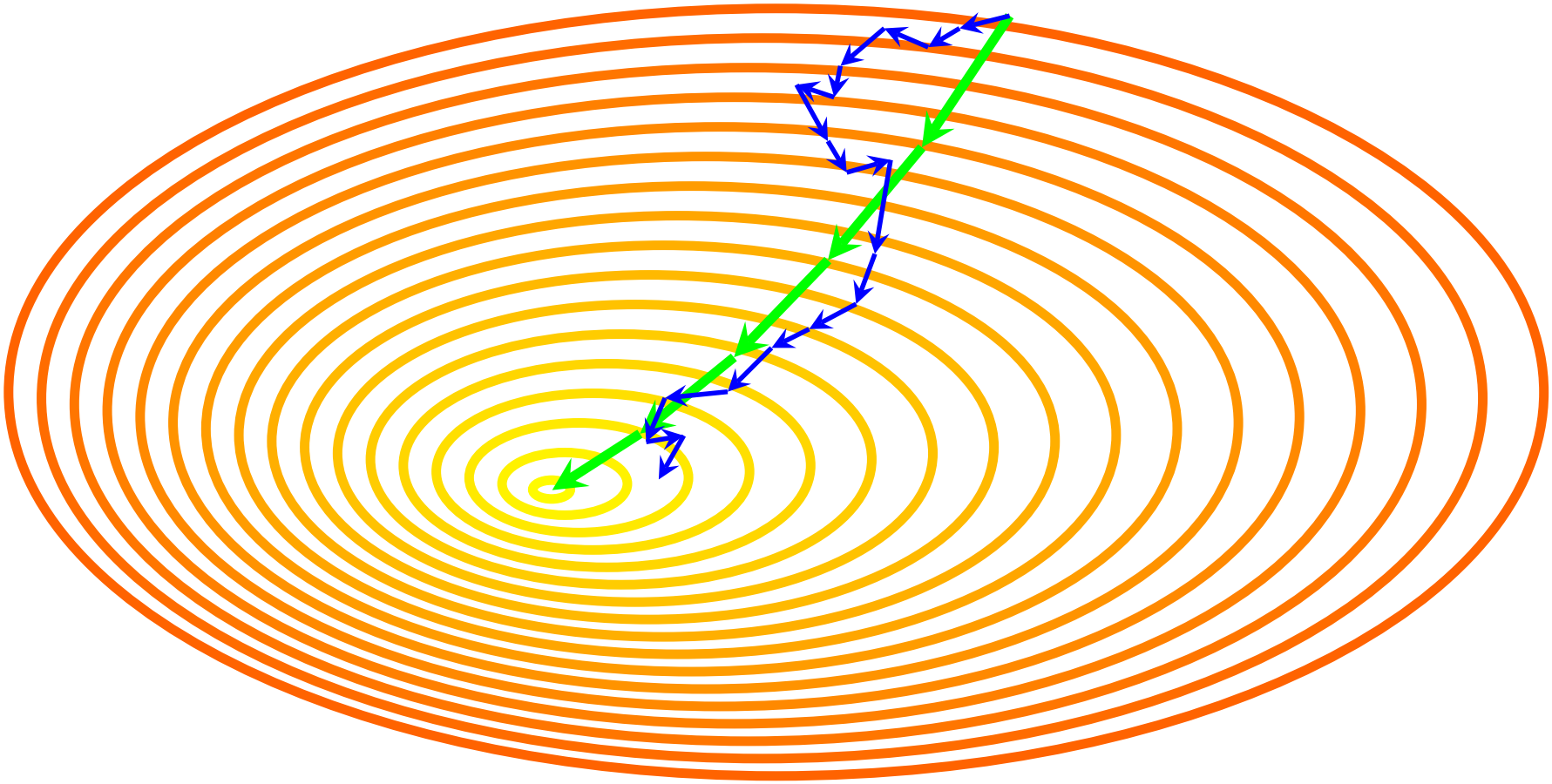
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



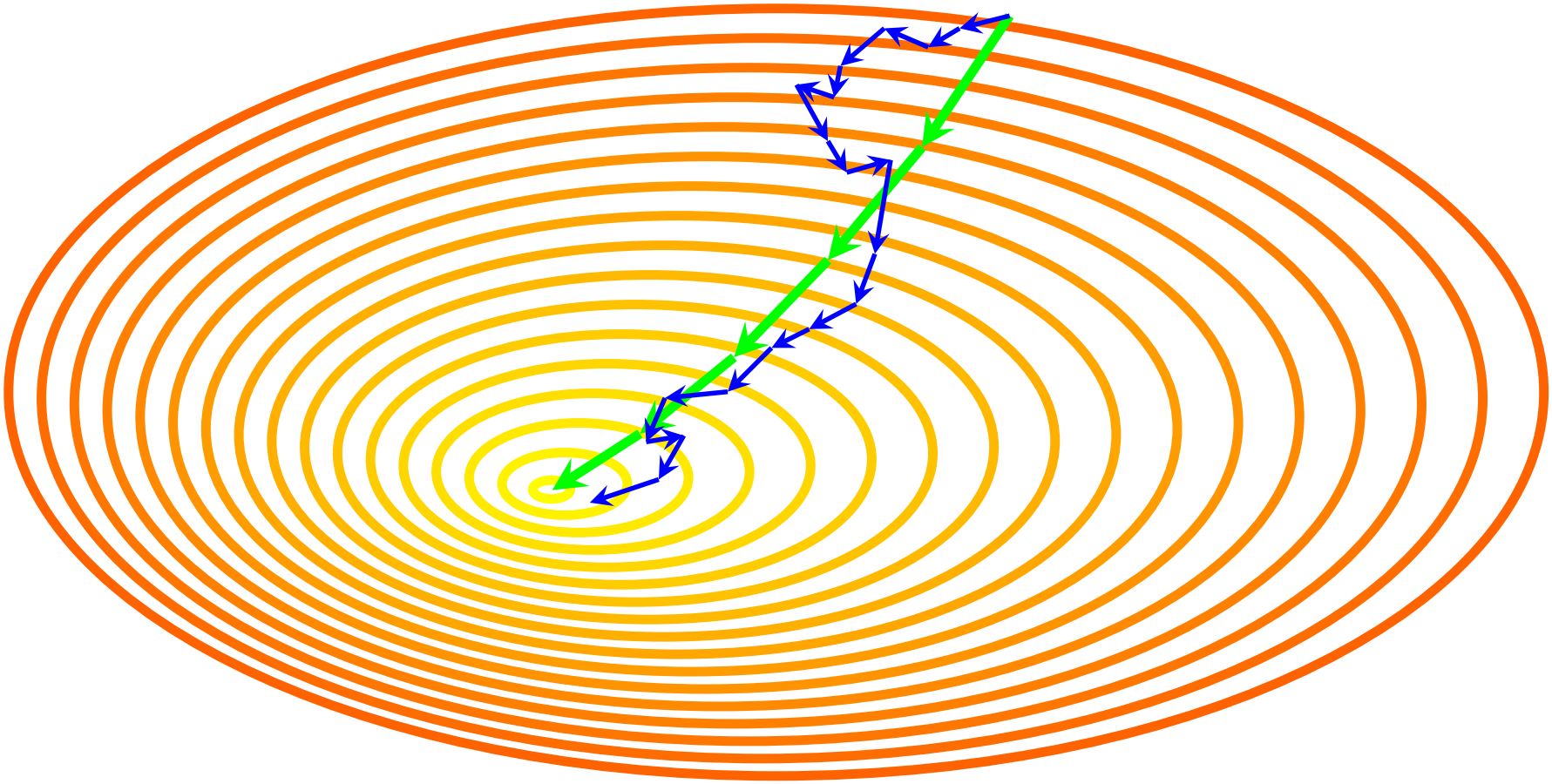
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



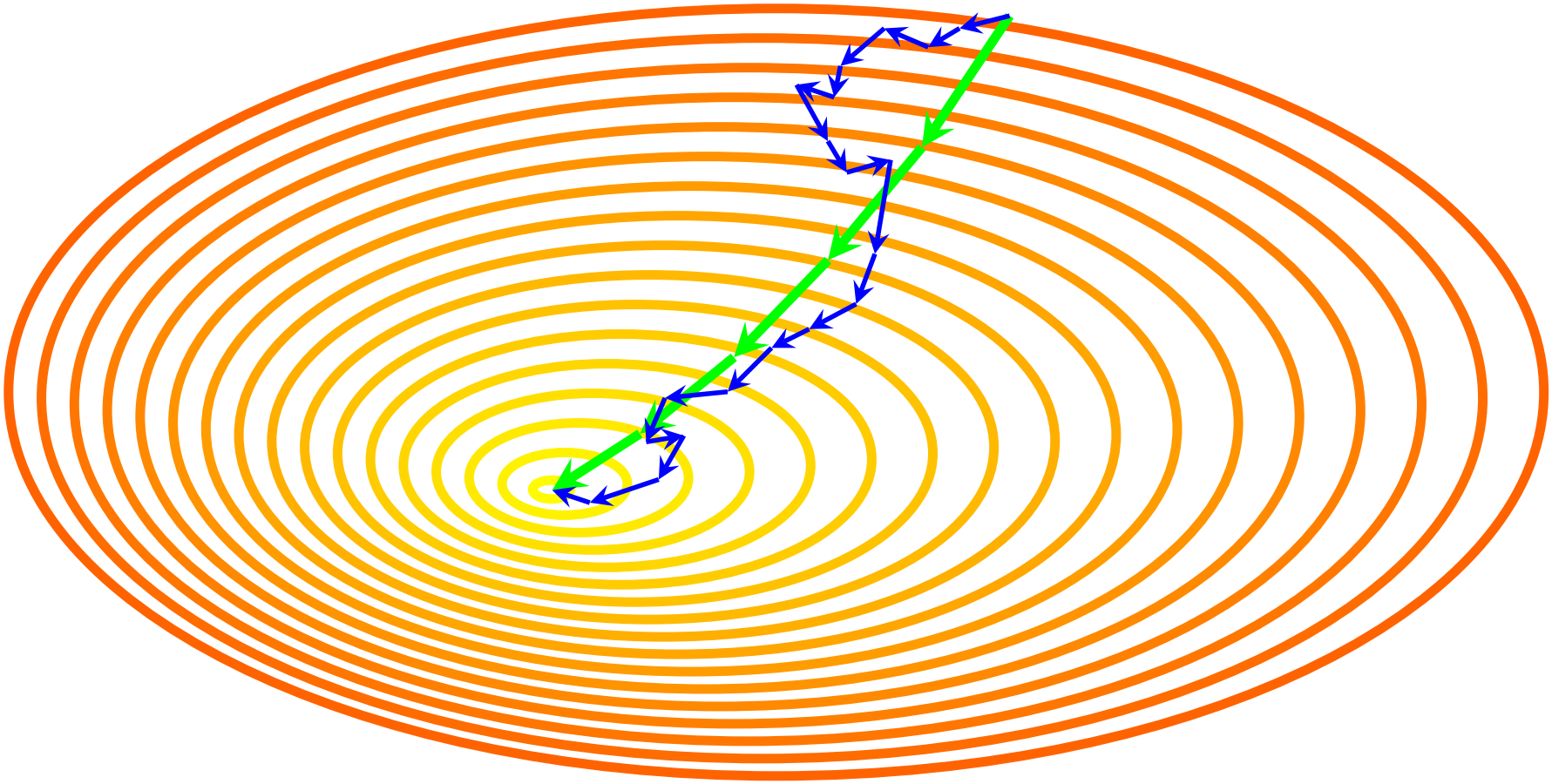
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



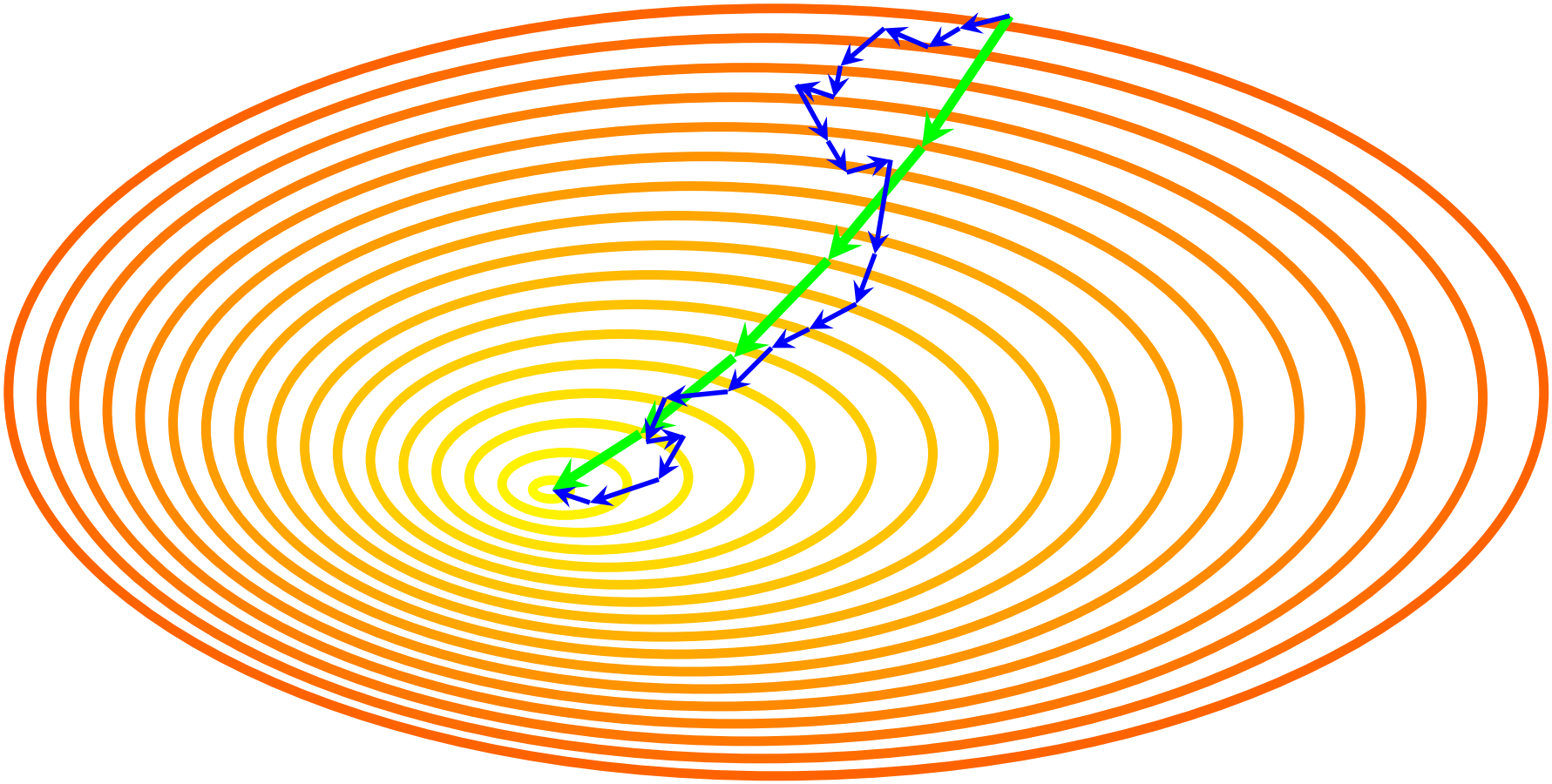
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



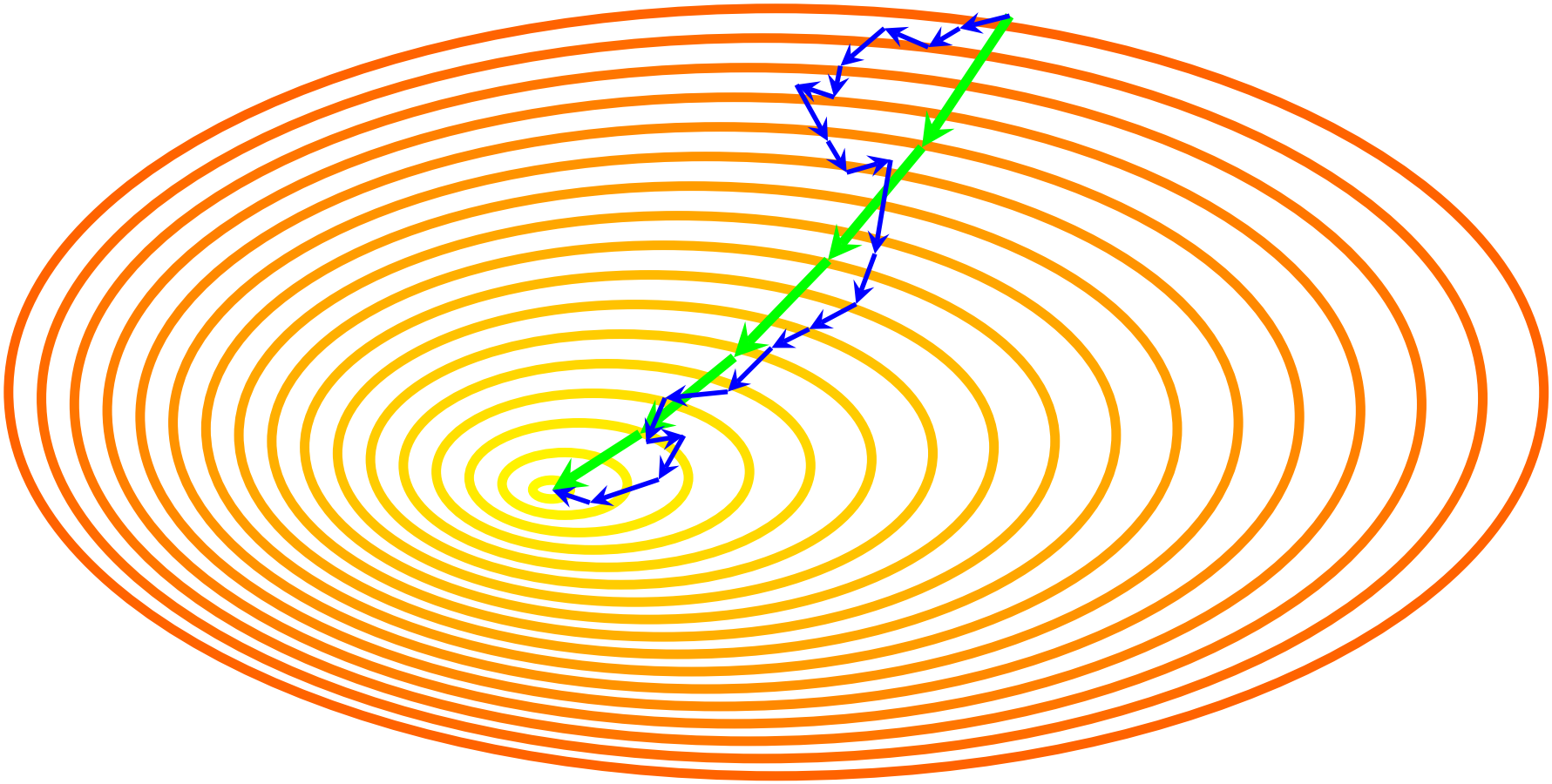
- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent:

Gradient versus Stochastic Gradient



- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent: may converges more slowly and unsteadily.

Gradient versus Stochastic Gradient



- Gradient Descent: converges quickly and smoothly (usually).
- Stochastic Gradient Descent: may converges more slowly and unsteadily.

But: one iteration of Stochastic Gradient Descent can be computed much more easily!

Recall: One-Dimensional Example

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$

Recall: One-Dimensional Example

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$

- It is possible to compute solution directly through differentiation

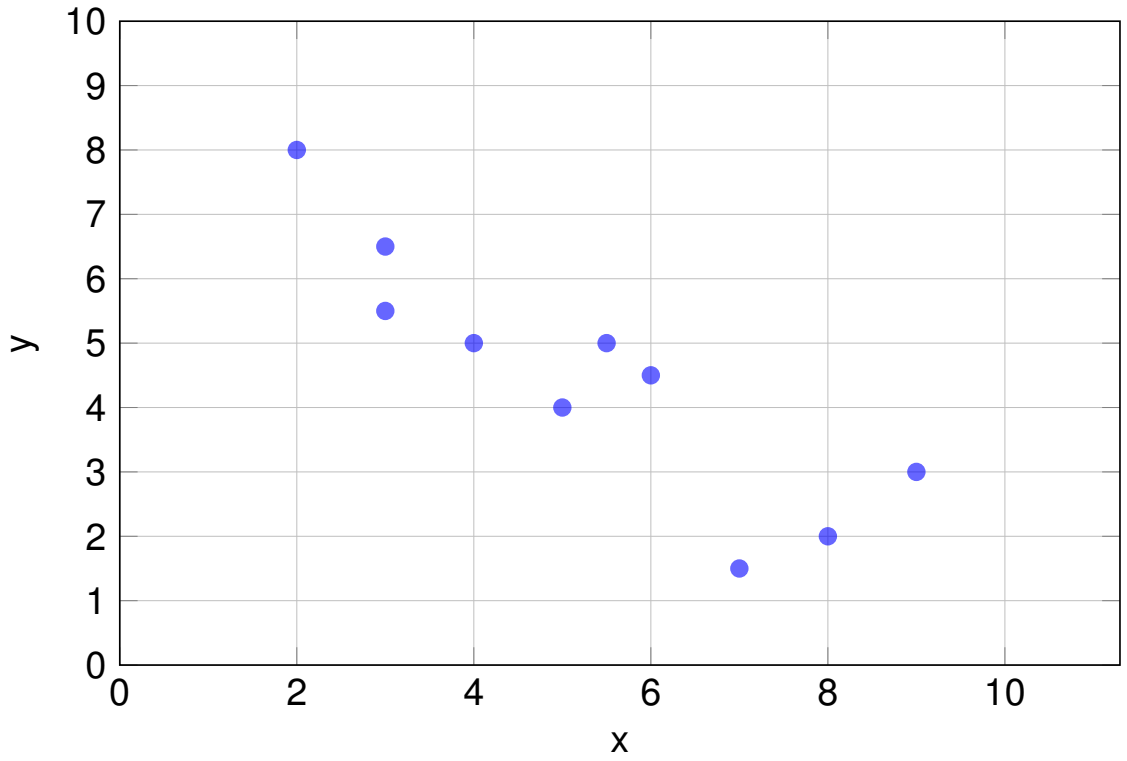
Recall: One-Dimensional Example

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$

- It is possible to compute solution directly through differentiation
- Easy for $d = 1$ (here), for $d \geq 2$ it requires Matrix Inversion

Recall: One-Dimensional Example

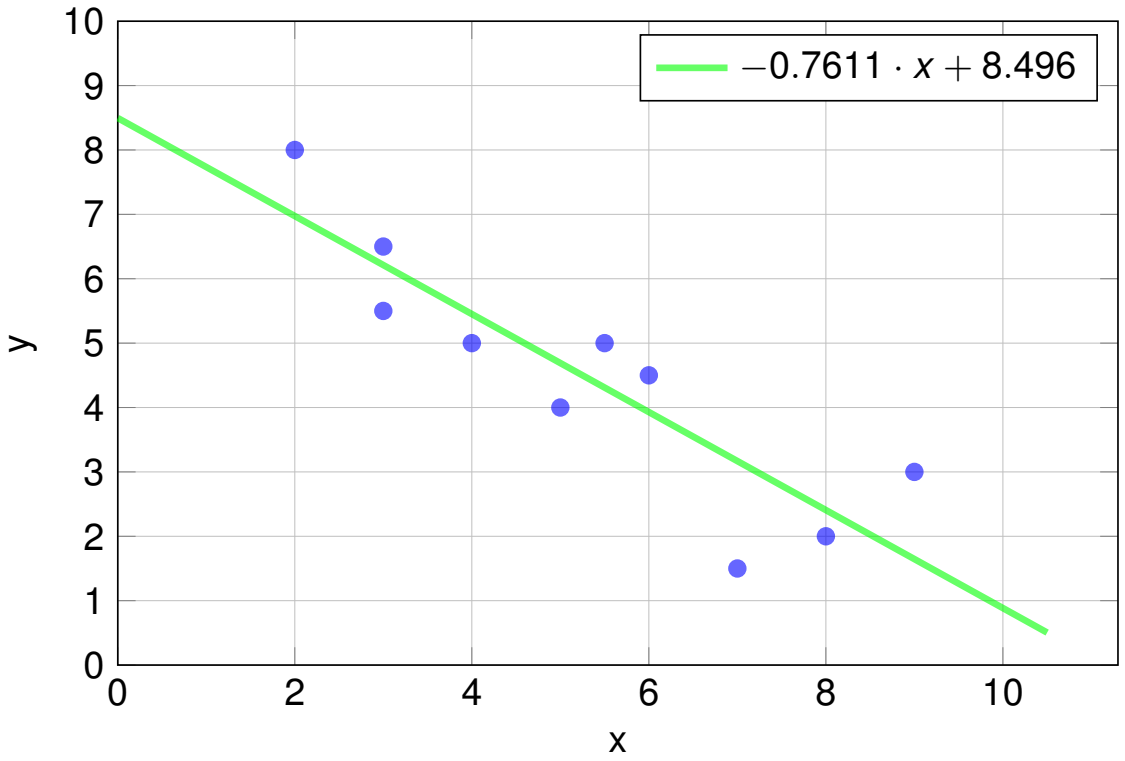
Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



- It is possible to compute solution directly through differentiation
- Easy for $d = 1$ (here), for $d \geq 2$ it requires Matrix Inversion
- For large data sets, an iterative solving method like Gradient Descent is usually preferred

Recall: One-Dimensional Example

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



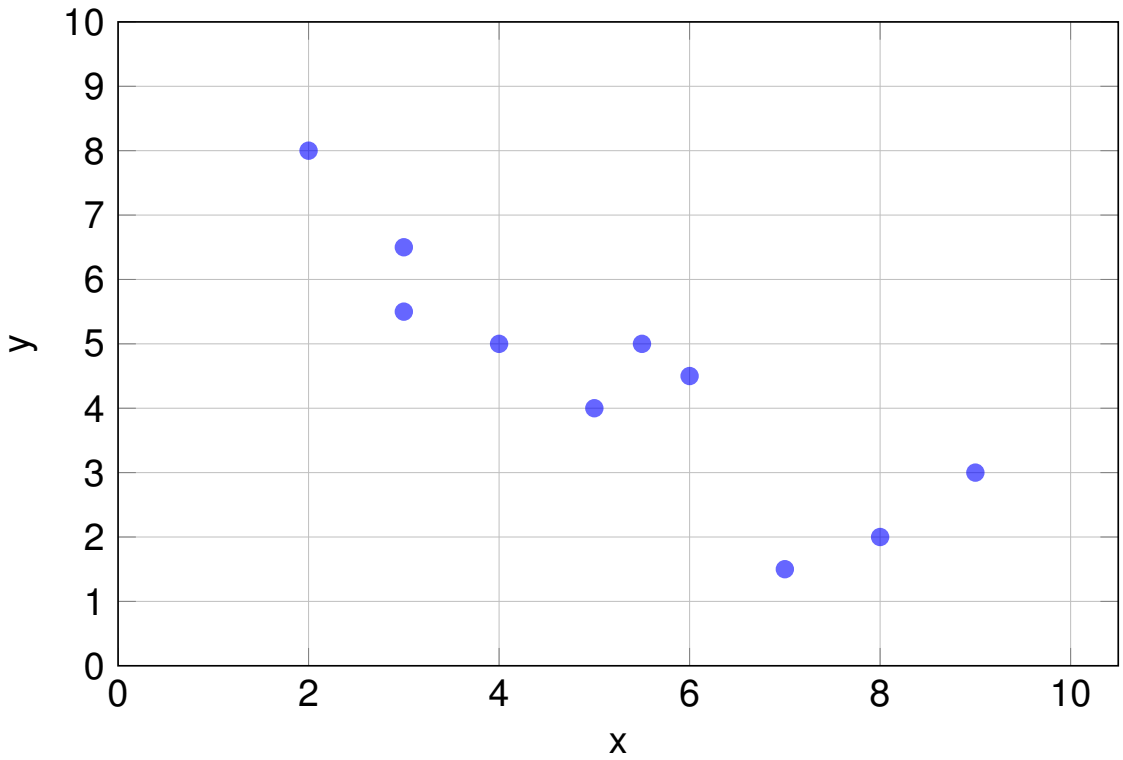
- It is possible to compute solution directly through differentiation
- Easy for $d = 1$ (here), for $d \geq 2$ it requires Matrix Inversion
- For large data sets, an iterative solving method like Gradient Descent is usually preferred

Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$

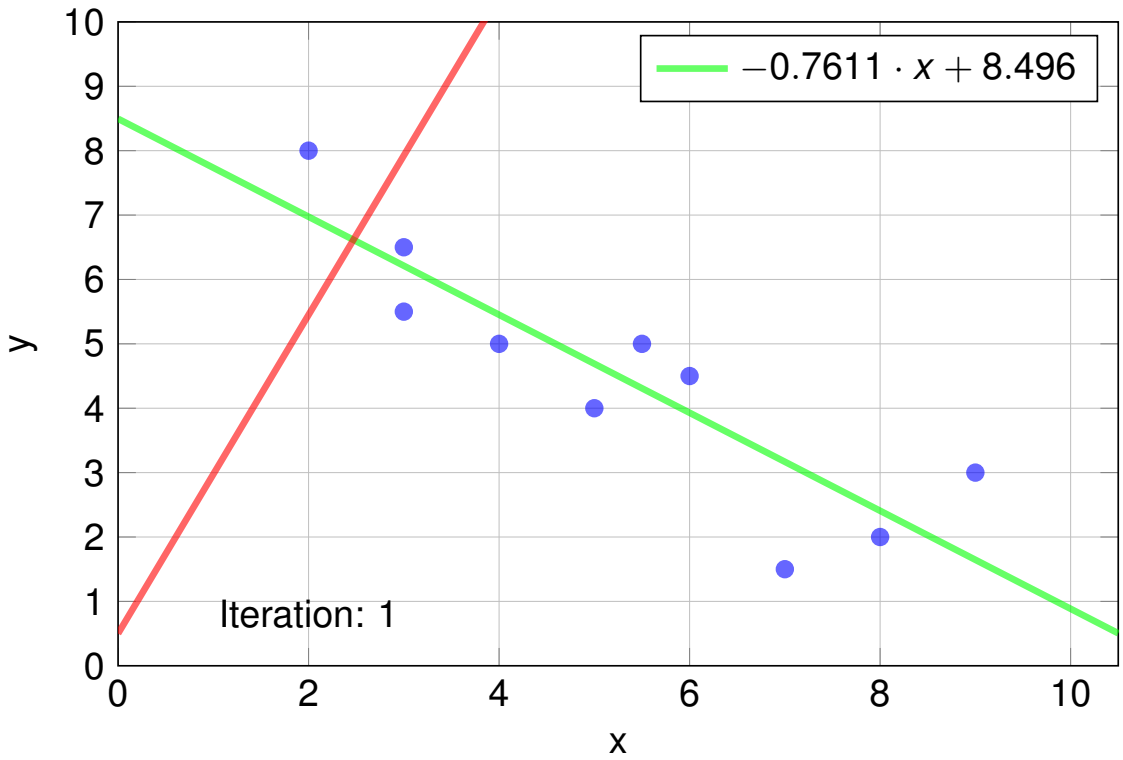
Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



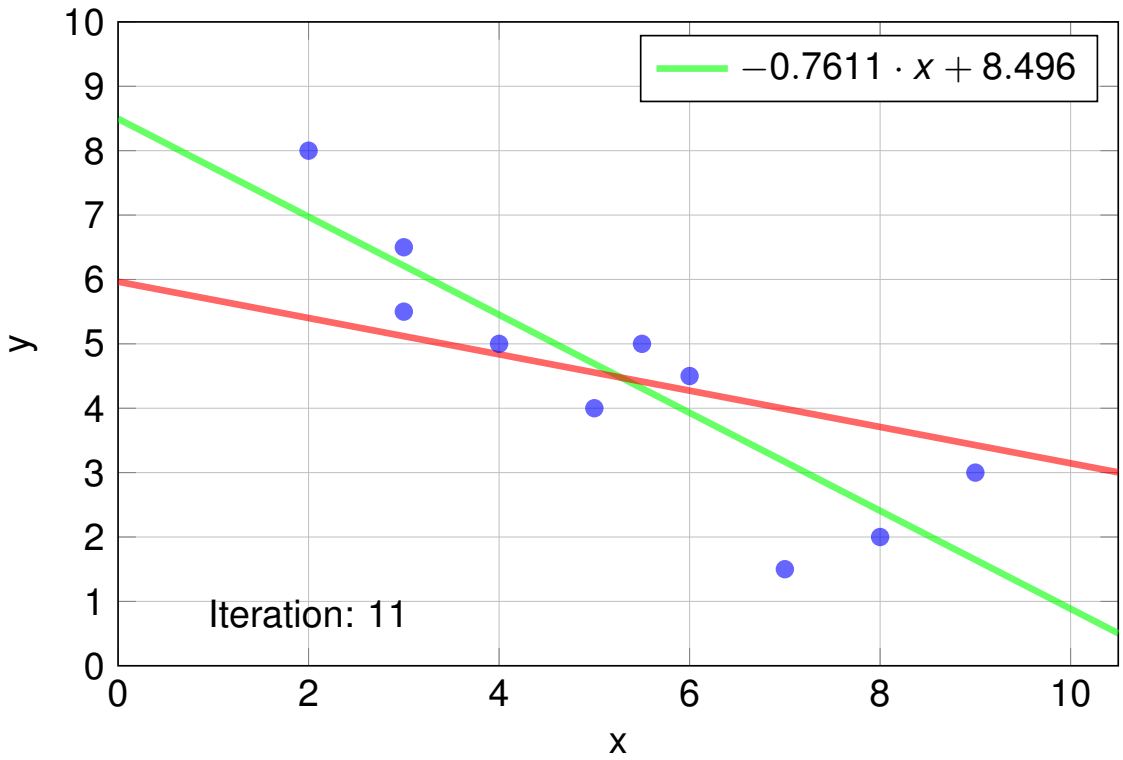
Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



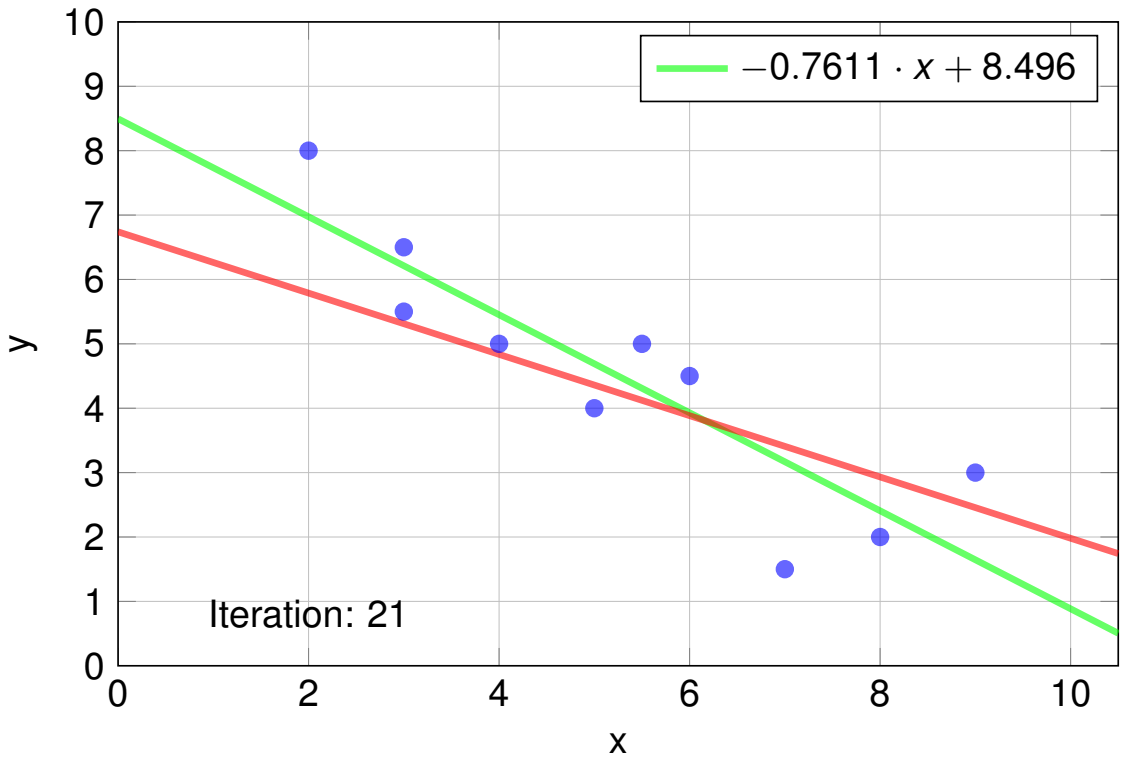
Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



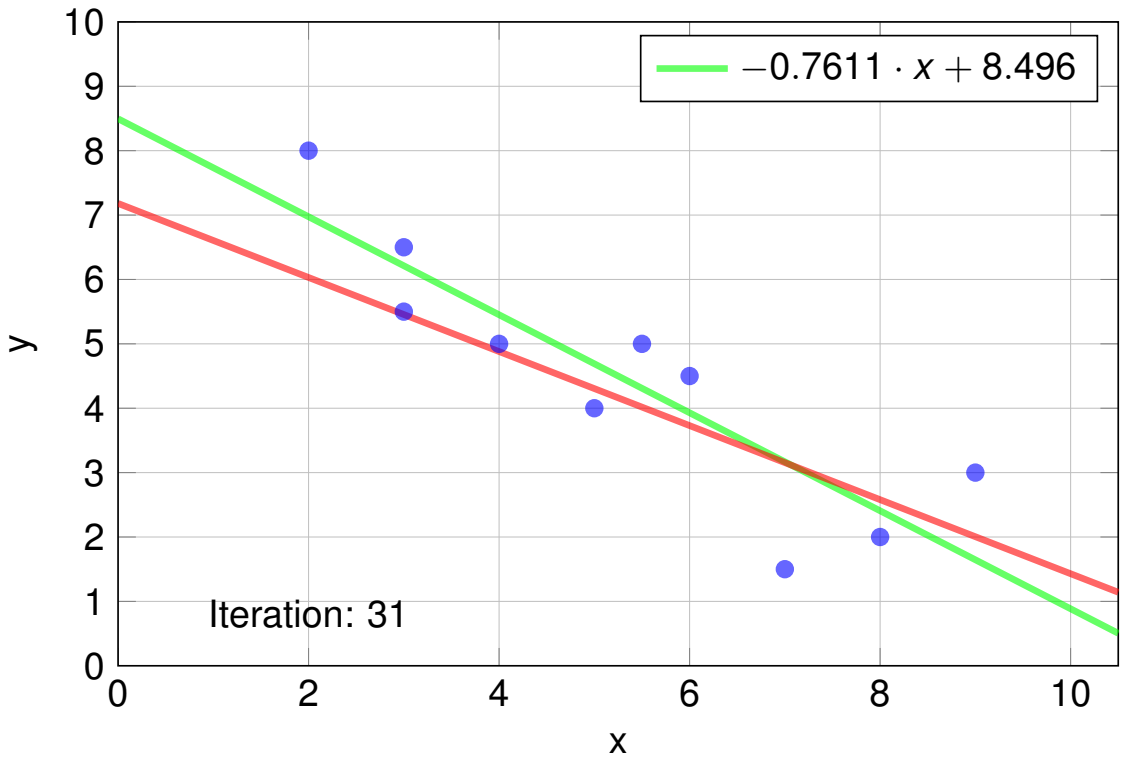
Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



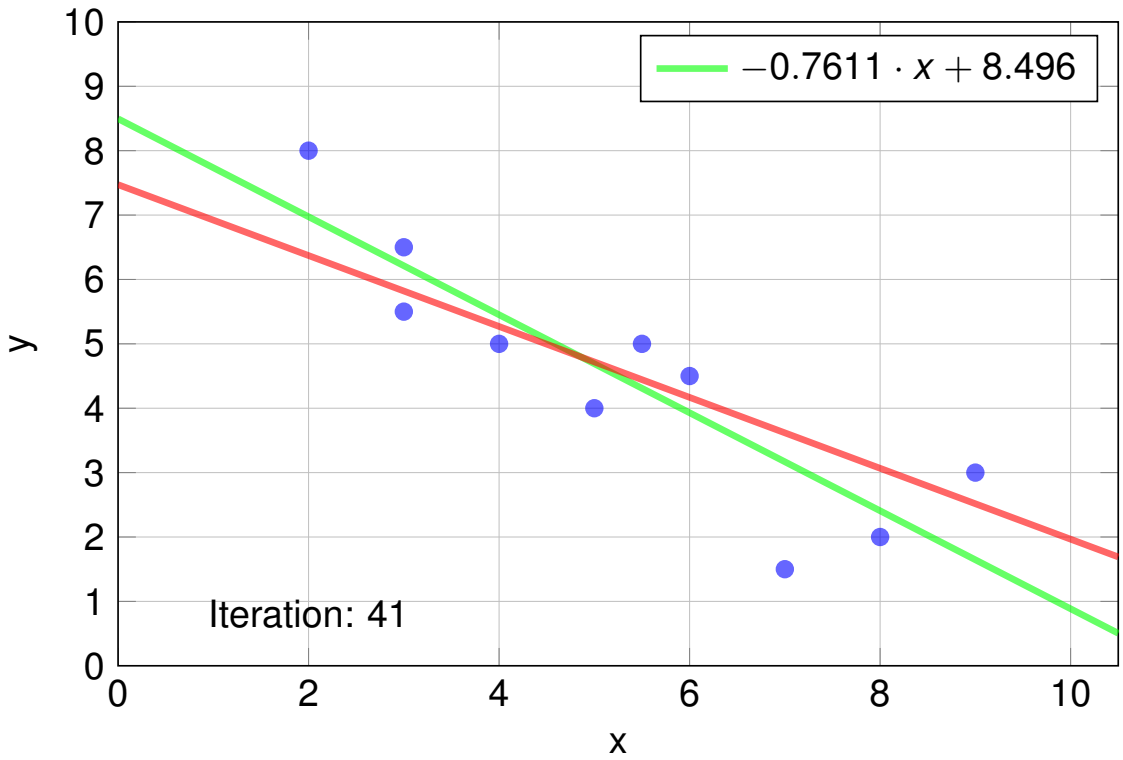
Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



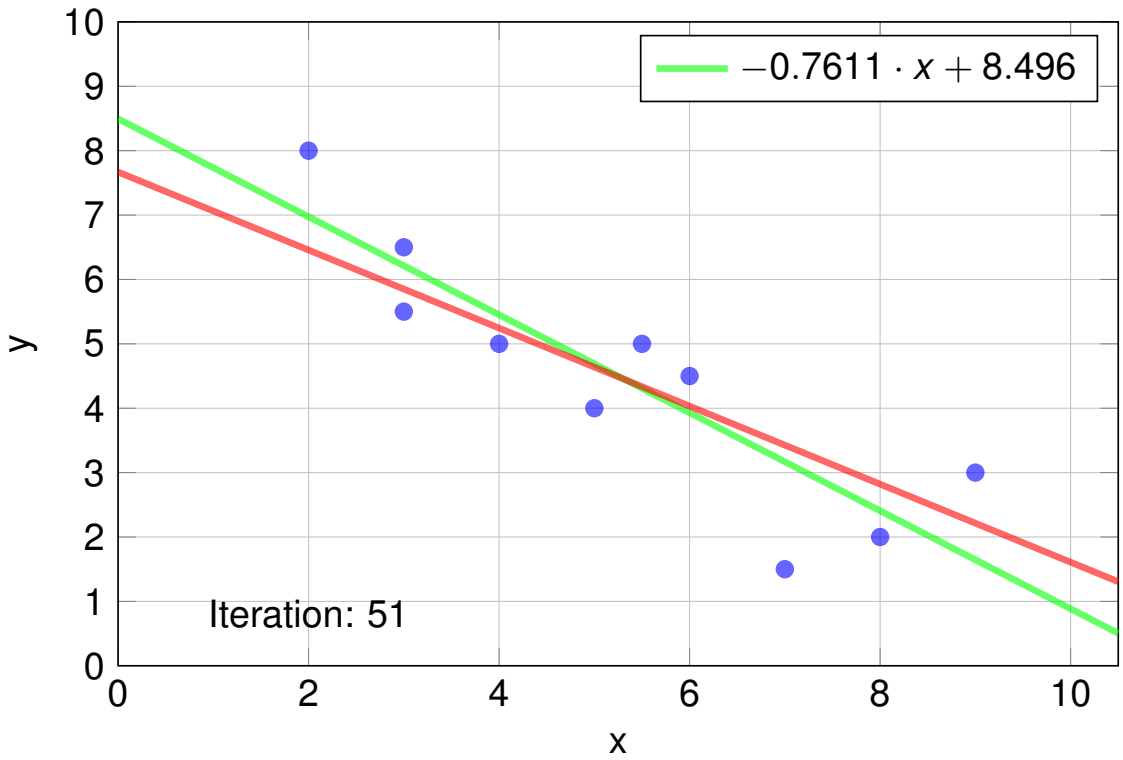
Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



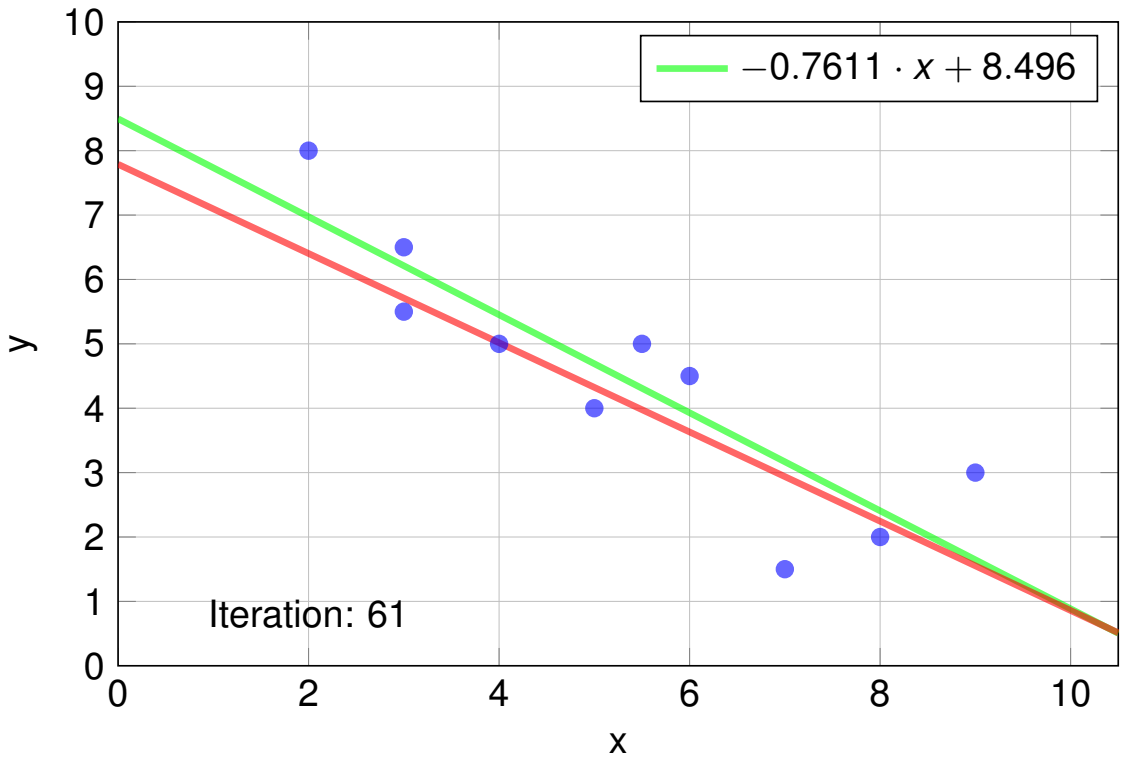
Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



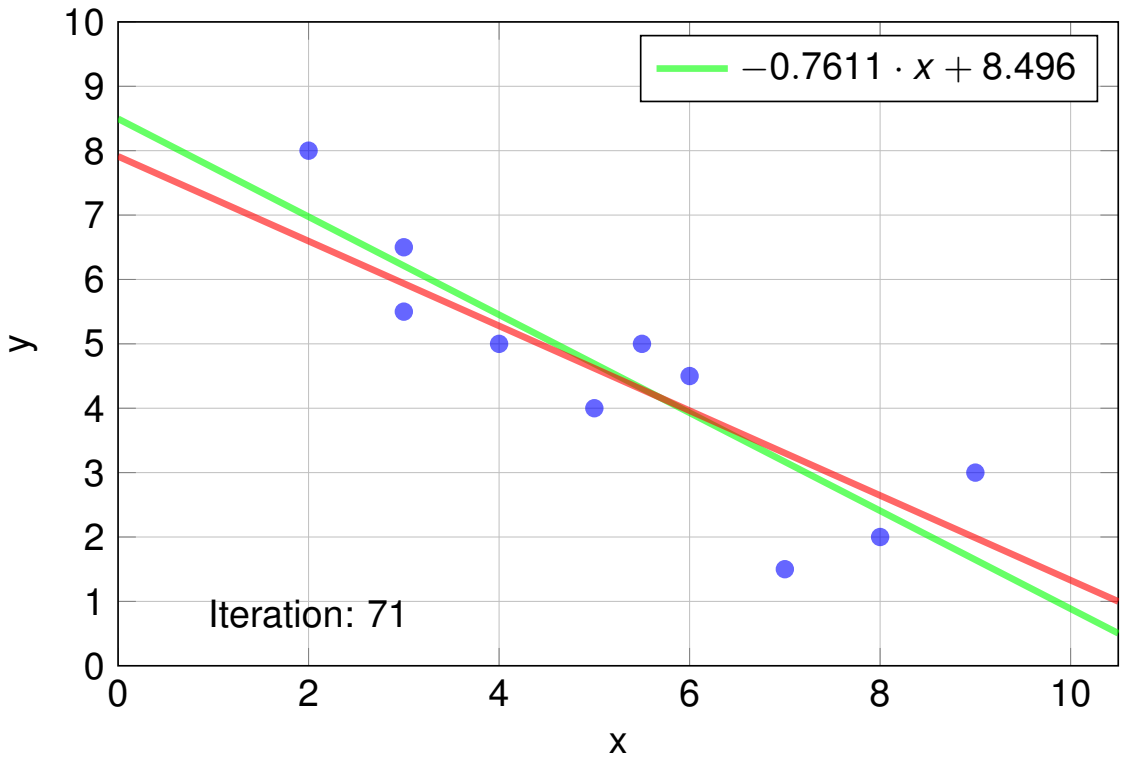
Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



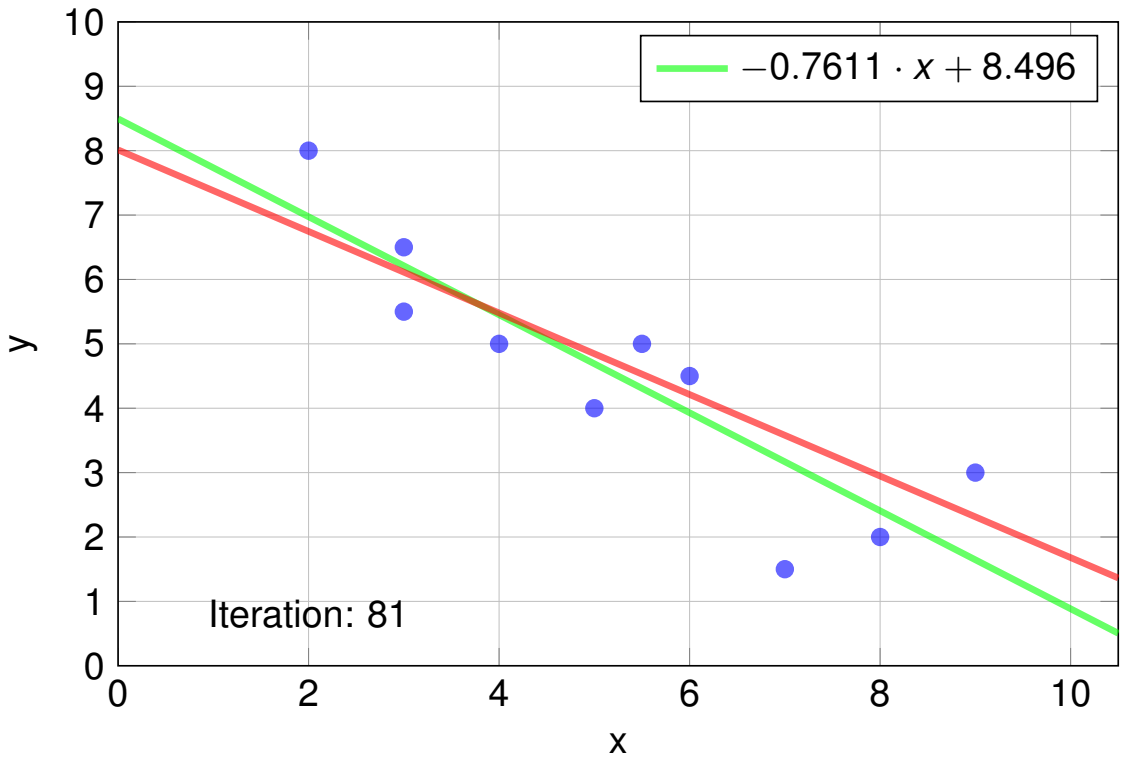
Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



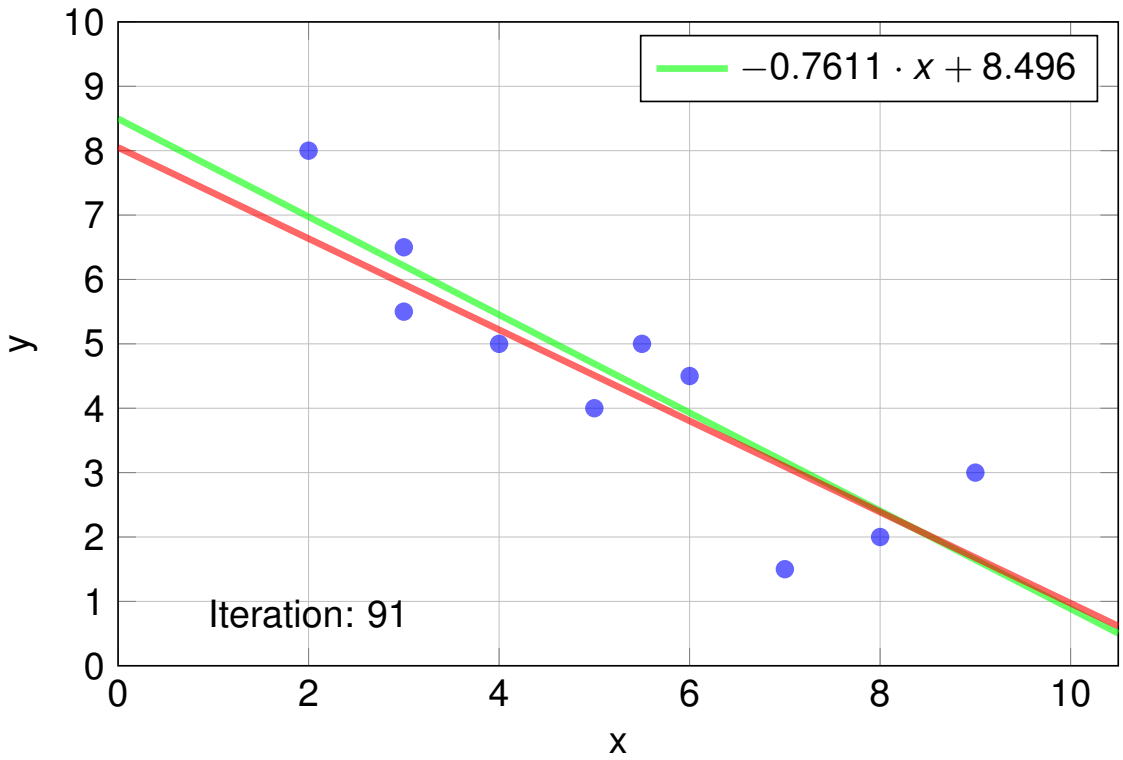
Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



Stochastic Gradient Descent Run

Point	(x, y)
1	$(2, 8)$
2	$(4, 5)$
3	$(3, 5.5)$
4	$(5, 4)$
5	$(6, 4.5)$
6	$(8, 2)$
7	$(3, 6.5)$
8	$(5.5, 5)$
9	$(7, 1.5)$
10	$(9, 3)$



Outline

Introduction

Linear Regression

Linear Regression with More than One Feature

Solving Linear Regression: Gradient Descent

Logistic Regression

Origin of Logistic Regression in Economics

Daniel McFadden (1937–)

James Heckman (1944–)

Nobel Prize of Economics (2000)

"James Heckman and Daniel McFadden have each developed theory and methods that are widely used in the statistical analysis of individual and household behavior, within economics as well as other social sciences."

Motivation: Back to Our Example

- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

Motivation: Back to Our Example

- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

Student	Homework	Exam
1	7	8.0
2	6	5.0
3	4	5.5
4	5	4.0
5	3	4.5
6	4	2.0
7	6	6.5
8	7	5.0
9	3.5	1.5
10	2	3.0

Motivation: Back to Our Example

- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

Student	Homework	Exam
1	7	~~8.0~~
2	6	~~5.0~~
3	4	~~5.5~~
4	5	~~4.0~~
5	3	~~4.5~~
6	4	~~2.0~~
7	6	~~6.5~~
8	7	~~5.0~~
9	3.5	~~1.5~~
10	2	~~3.0~~

Motivation: Back to Our Example

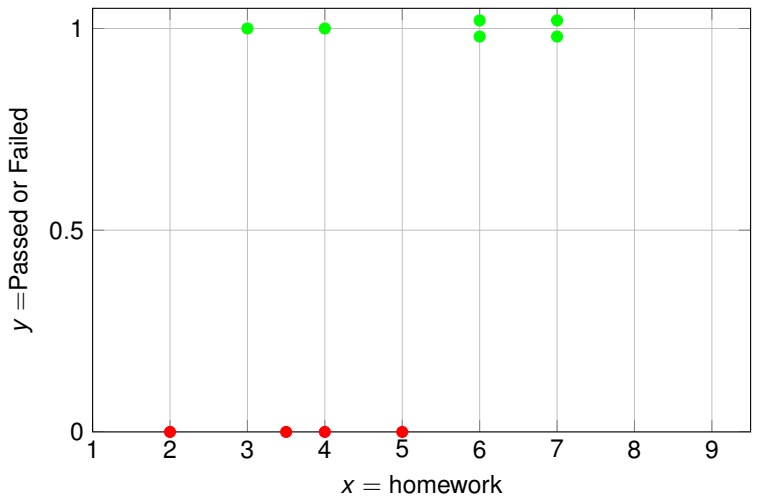
- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

Student	Homework	Exam
1	7	~~8.0~~ passed
2	6	~~5.0~~ passed
3	4	~~5.5~~ passed
4	5	~~4.0~~ failed
5	3	~~4.5~~ passed
6	4	~~2.0~~ failed
7	6	~~6.5~~ passed
8	7	~~5.0~~ passed
9	3.5	~~1.5~~ failed
10	2	~~3.0~~ failed

Motivation: Back to Our Example

- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

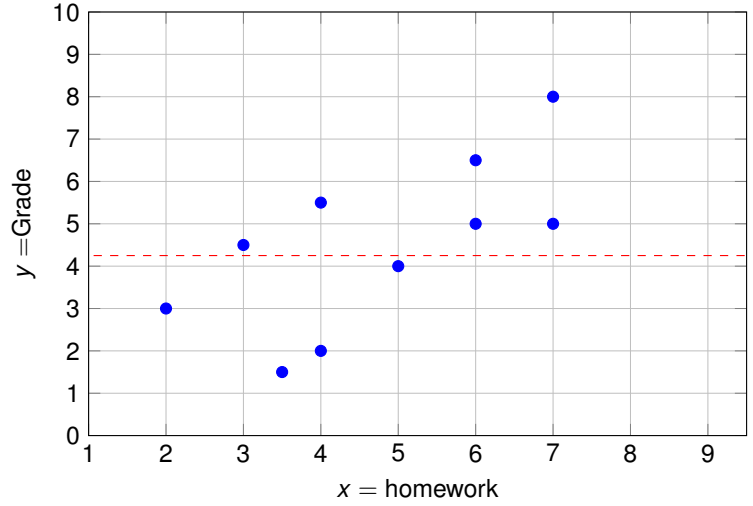
Student	Homework	Exam
1	7	~~8.0~~ passed
2	6	~~5.0~~ passed
3	4	~~5.5~~ passed
4	5	~~4.0~~ failed
5	3	~~4.5~~ passed
6	4	~~2.0~~ failed
7	6	~~6.5~~ passed
8	7	~~5.0~~ passed
9	3.5	~~1.5~~ failed
10	2	~~3.0~~ failed

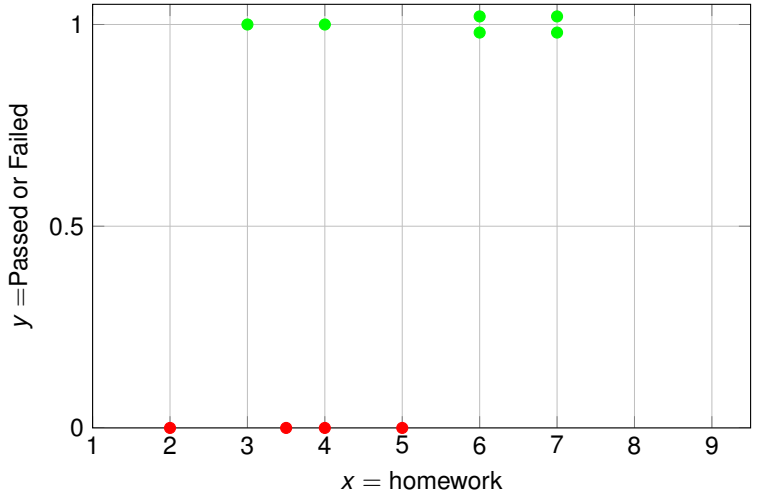


Motivation: Back to Our Example

- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

Student	Homework	Exam
1	7	~~8.0~~ passed
2	6	~~5.0~~ passed
3	4	~~5.5~~ passed
4	5	~~4.0~~ failed
5	3	~~4.5~~ passed
6	4	~~2.0~~ failed
7	6	~~6.5~~ passed
8	7	~~5.0~~ passed
9	3.5	~~1.5~~ failed
10	2	~~3.0~~ failed

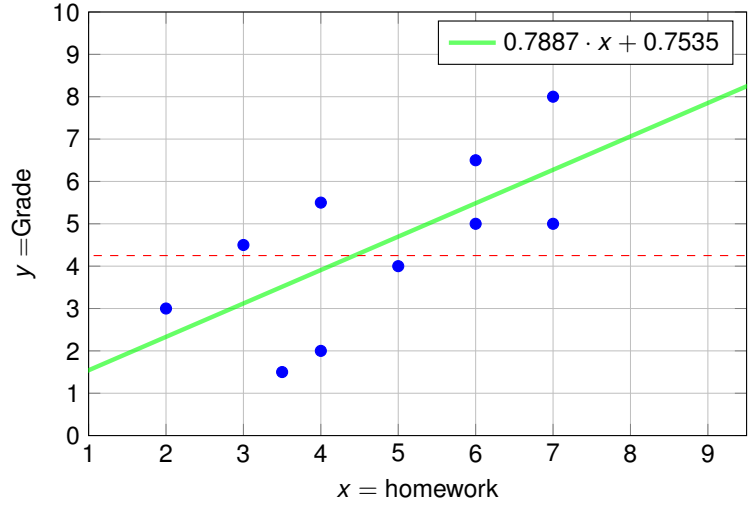


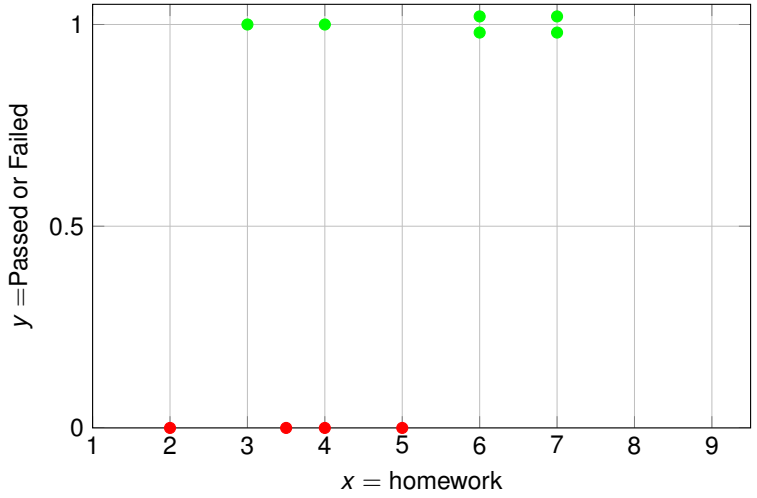


Motivation: Back to Our Example

- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

Student	Homework	Exam
1	7	~~8.0~~ passed
2	6	~~5.0~~ passed
3	4	~~5.5~~ passed
4	5	~~4.0~~ failed
5	3	~~4.5~~ passed
6	4	~~2.0~~ failed
7	6	~~6.5~~ passed
8	7	~~5.0~~ passed
9	3.5	~~1.5~~ failed
10	2	~~3.0~~ failed

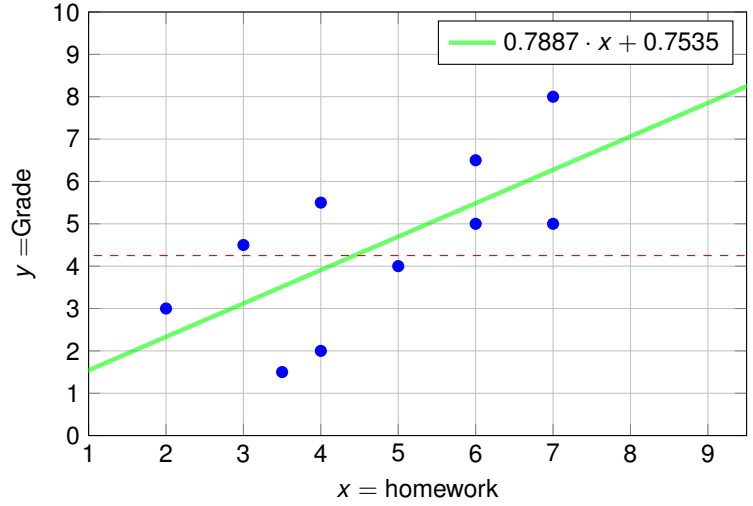


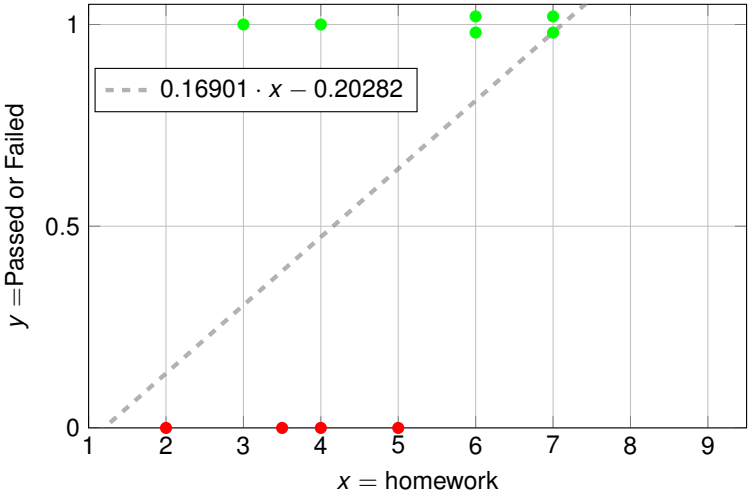


Motivation: Back to Our Example

- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

Student	Homework	Exam
1	7	~~8.0~~ passed
2	6	~~5.0~~ passed
3	4	~~5.5~~ passed
4	5	~~4.0~~ failed
5	3	~~4.5~~ passed
6	4	~~2.0~~ failed
7	6	~~6.5~~ passed
8	7	~~5.0~~ passed
9	3.5	~~1.5~~ failed
10	2	~~3.0~~ failed

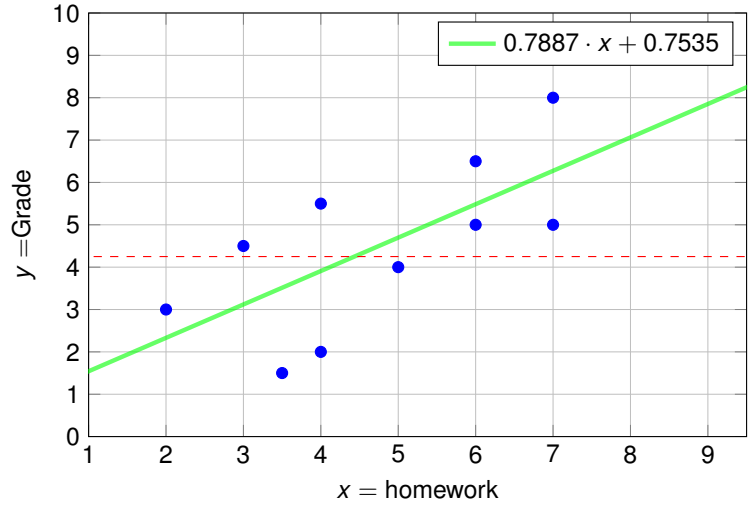


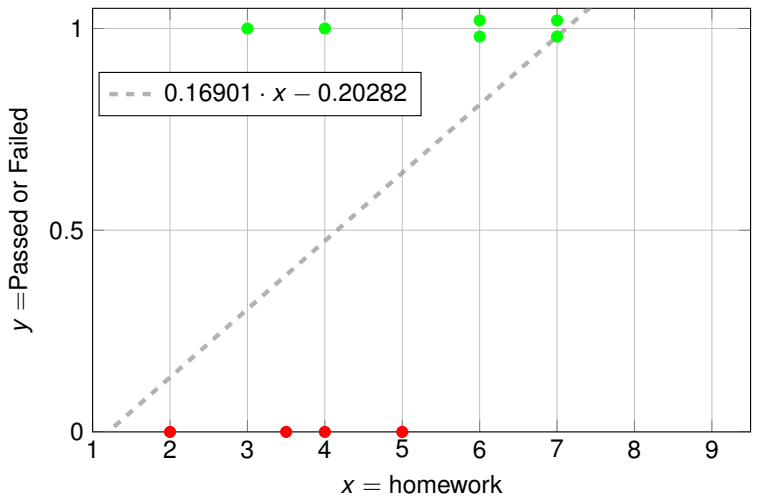


Motivation: Back to Our Example

- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

Student	Homework	Exam
1	7	~~8.0~~ passed
2	6	~~5.0~~ passed
3	4	~~5.5~~ passed
4	5	~~4.0~~ failed
5	3	~~4.5~~ passed
6	4	~~2.0~~ failed
7	6	~~6.5~~ passed
8	7	~~5.0~~ passed
9	3.5	~~1.5~~ failed
10	2	~~3.0~~ failed



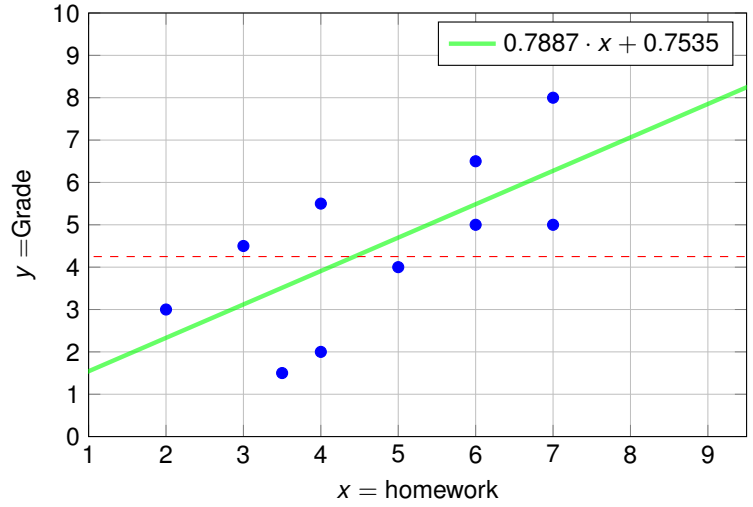


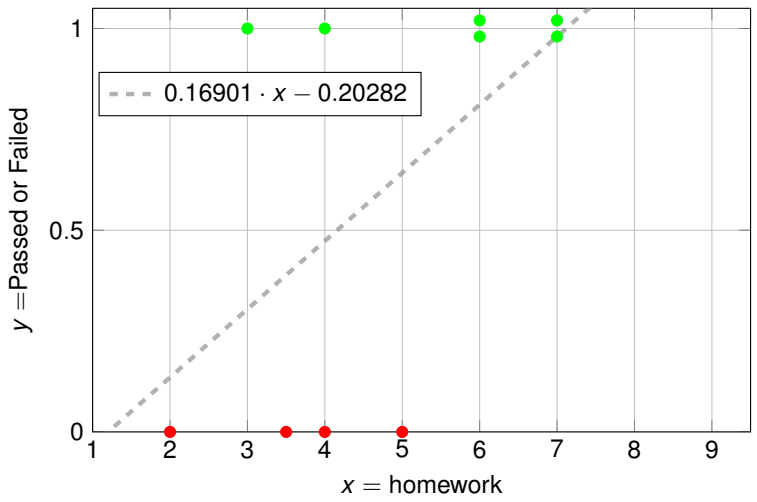
Exercise 3: What are the problems with using the regression line?

Motivation: Back to Our Example

- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

Student	Homework	Exam
1	7	~~8.0~~ passed
2	6	~~5.0~~ passed
3	4	~~5.5~~ passed
4	5	~~4.0~~ failed
5	3	~~4.5~~ passed
6	4	~~2.0~~ failed
7	6	~~6.5~~ passed
8	7	~~5.0~~ passed
9	3.5	~~1.5~~ failed
10	2	~~3.0~~ failed



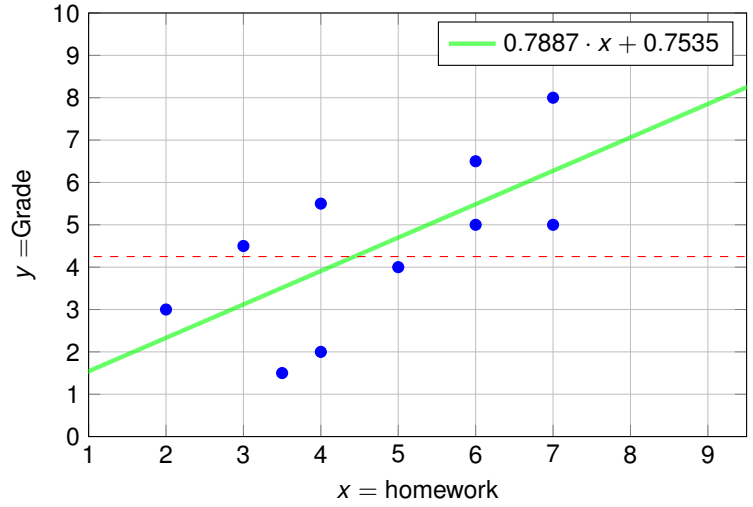


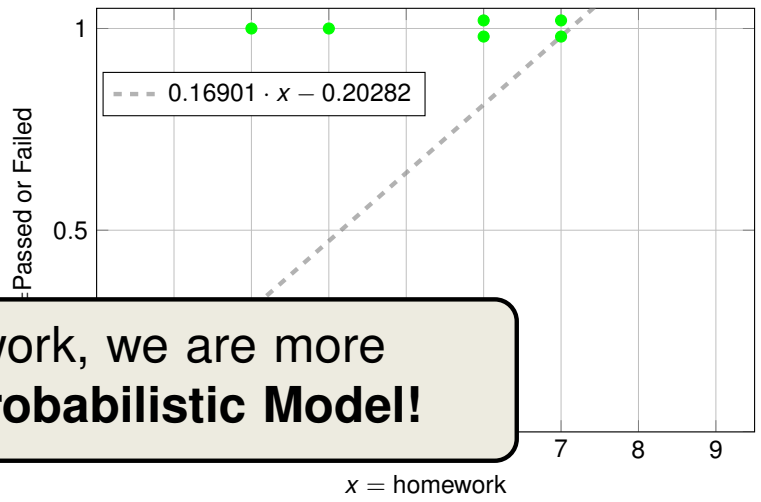
Need a model that assigns only values 1 (passed) and 0 (failed)!

Motivation: Back to Our Example

- Suppose we only know if student has passed or failed exam
- That is, we do **not** know the grade

Student	Homework	Exam
1	7	~~8.0~~ passed
2	6	~~5.0~~ passed
3	4	~~5.5~~ passed
4	5	~~4.0~~ failed
5	3	~~4.5~~ passed
6	4	~~2.0~~ failed
7	6	~~6.5~~ passed
8		
9		
10		

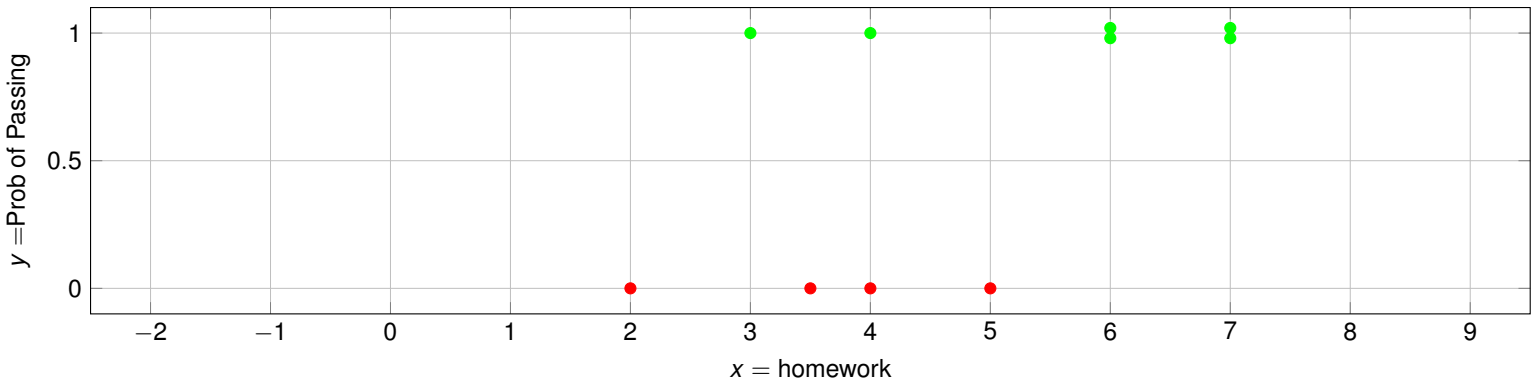




For students that did better on homework, we are more "confident" in predicting a pass! $\rightsquigarrow$ **Probabilistic Model!**

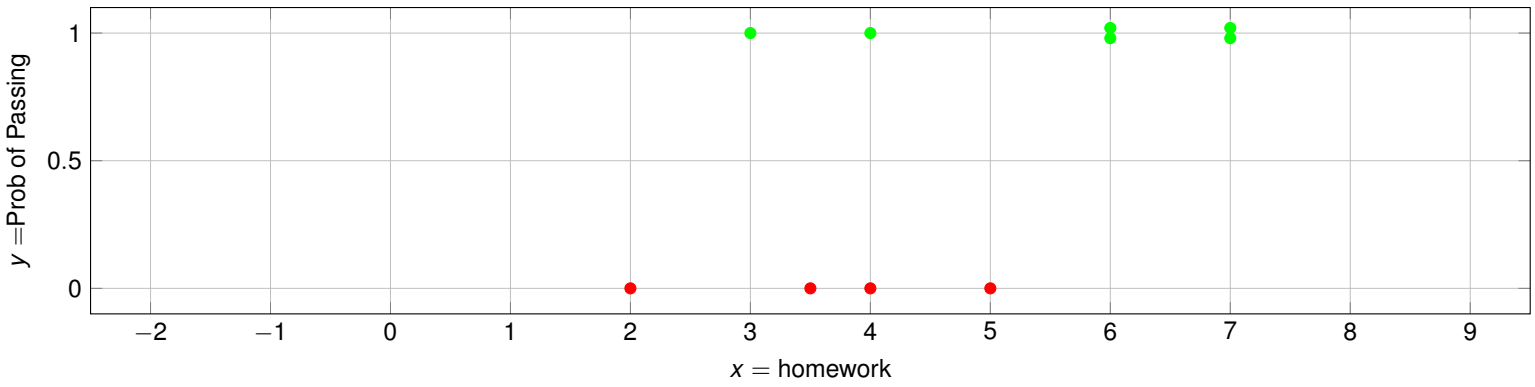
Need a model that assigns only values 1 (passed) and 0 (failed)!

The Logistic Function



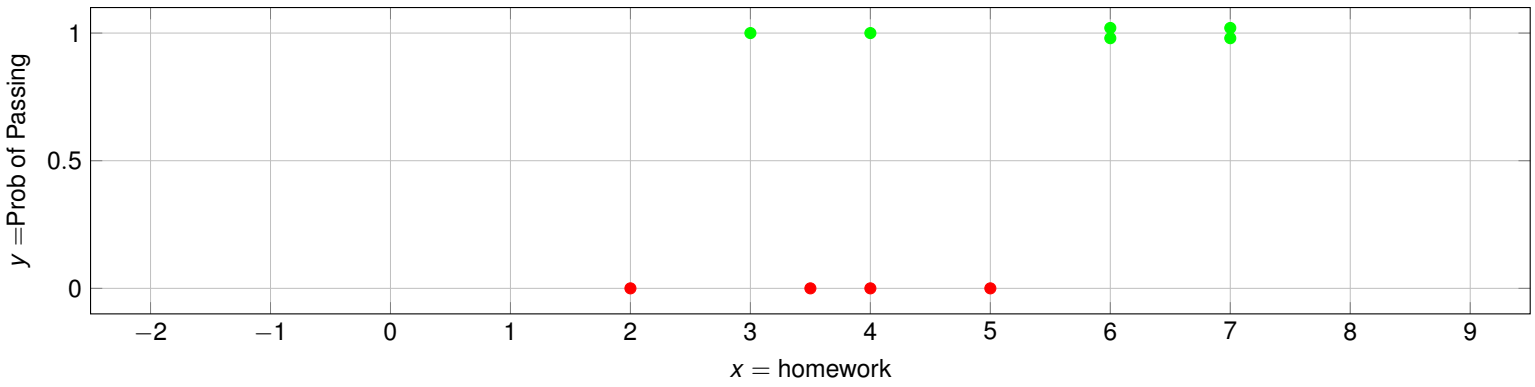
- If student did homework very well, then high probability to pass exam

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

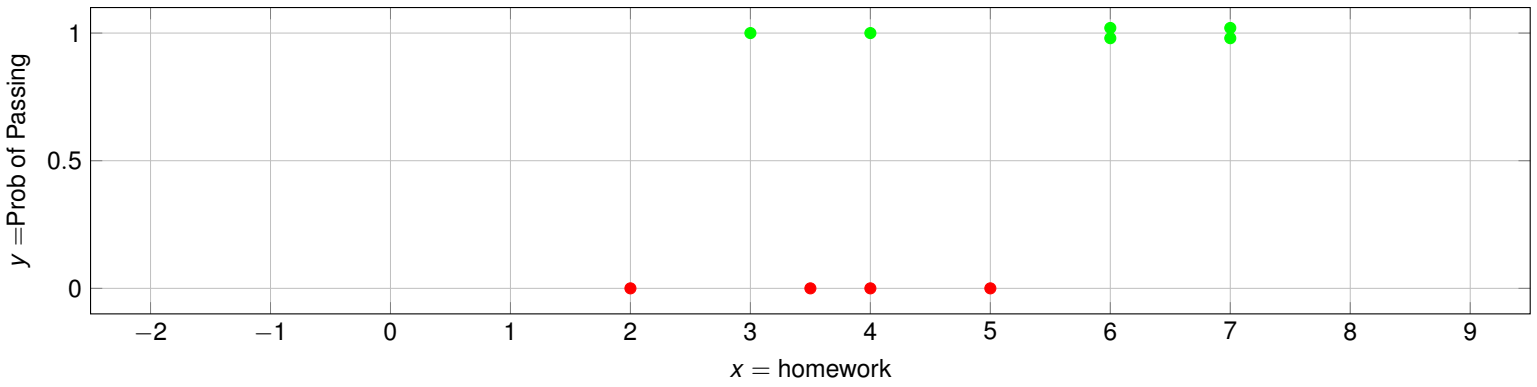
The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

- Need to quantify “very well”, “high probability”...

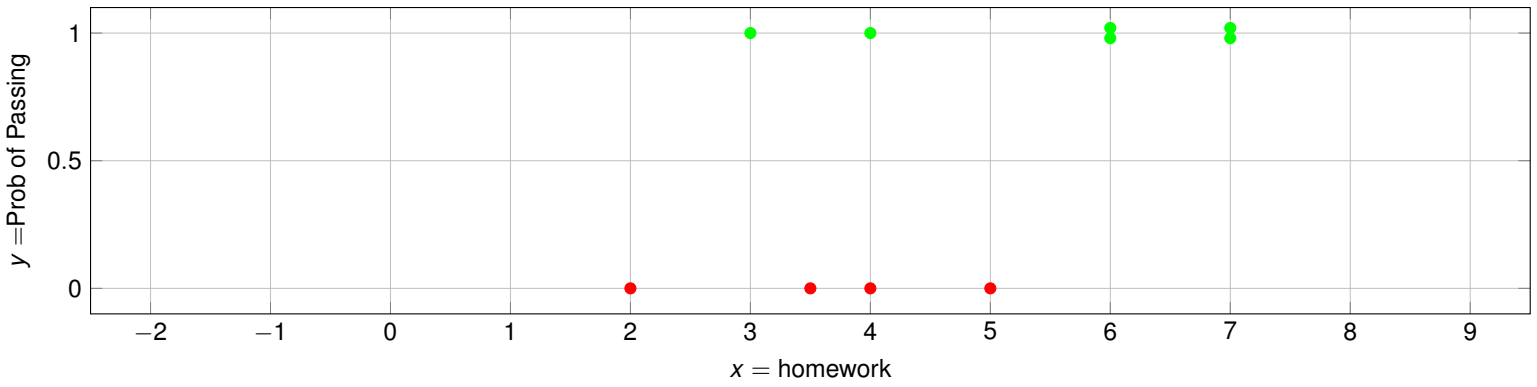
The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

- Need to quantify "very well", "high probability"...
- What if student's homework is average?

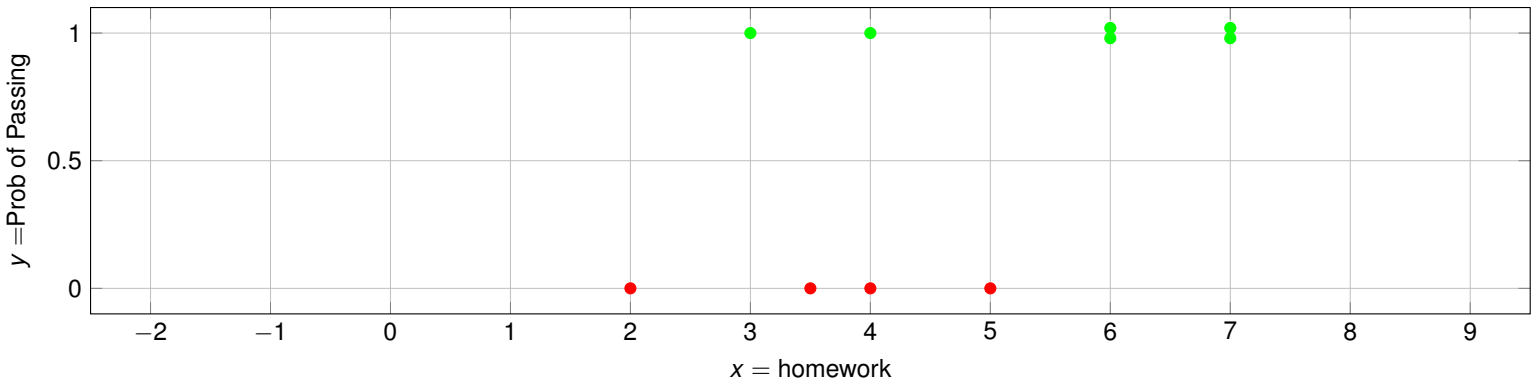
The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

- Need to quantify "very well", "high probability"...
- What if student's homework is average? $\rightsquigarrow$ then passing exam is "50:50"!

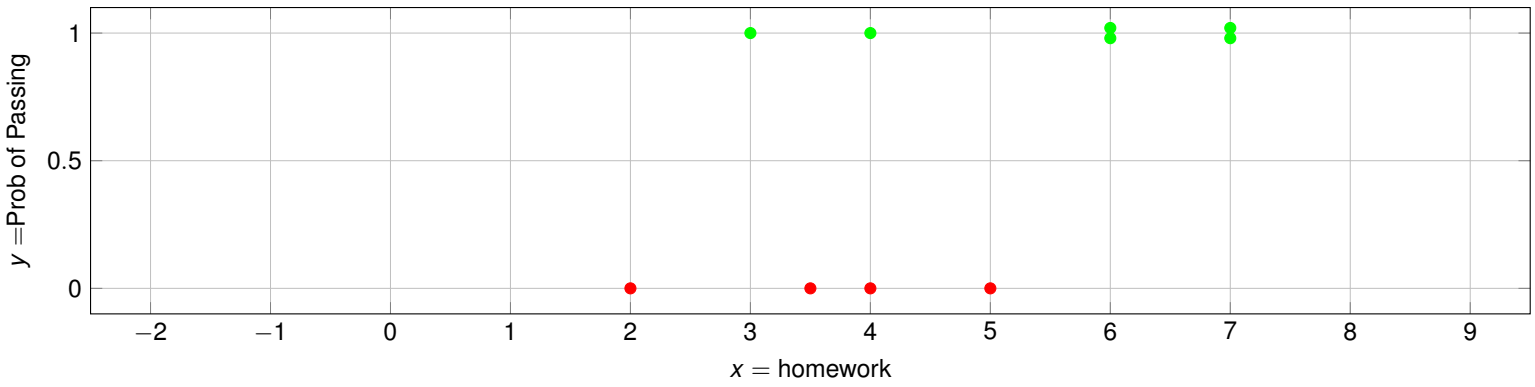
The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

We are looking for a function $f(x)$ such that:

The Logistic Function

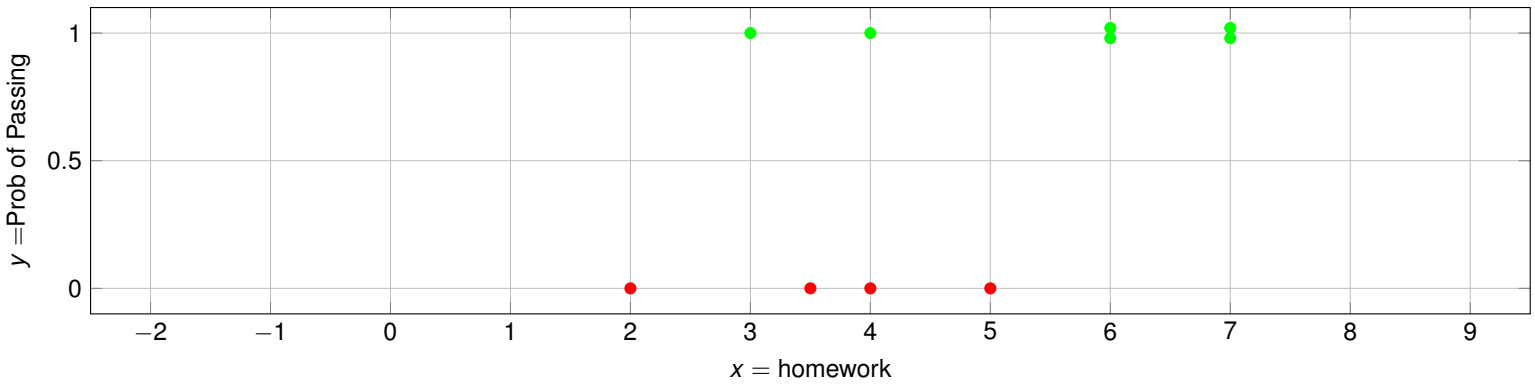


- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$

The Logistic Function

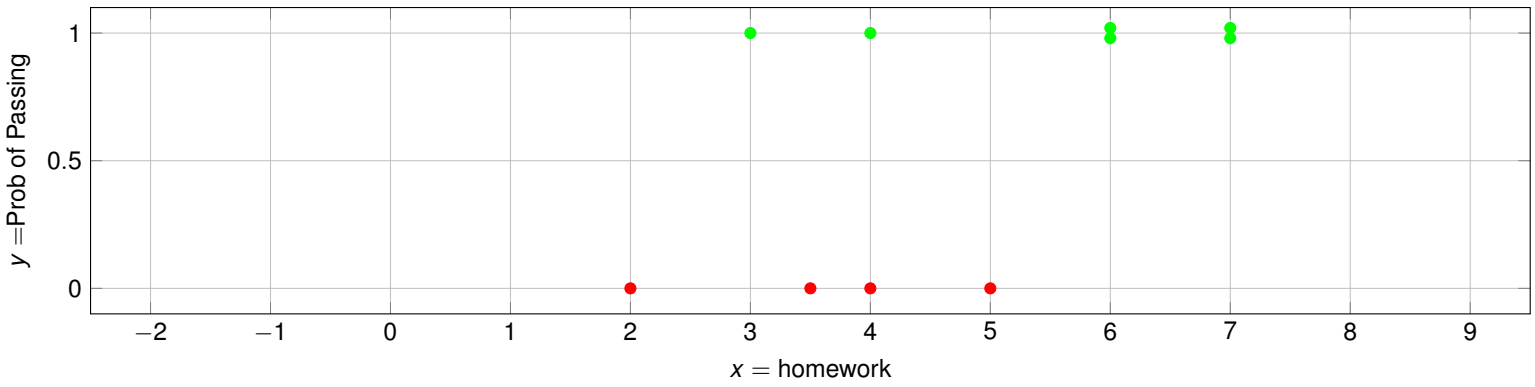


- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$

The Logistic Function

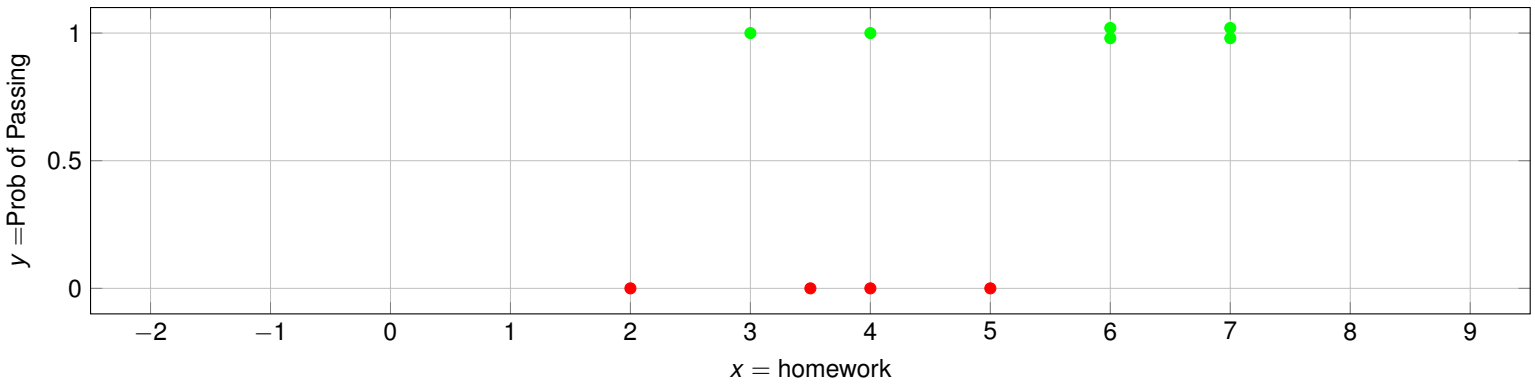


- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

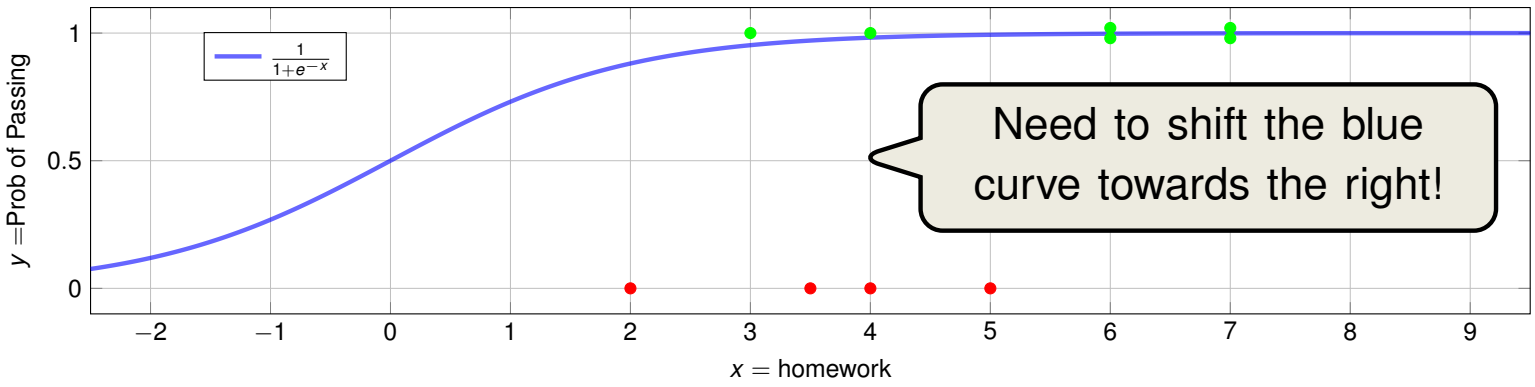
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the logistic function

$$f(x) = \frac{1}{1 + e^{-x}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

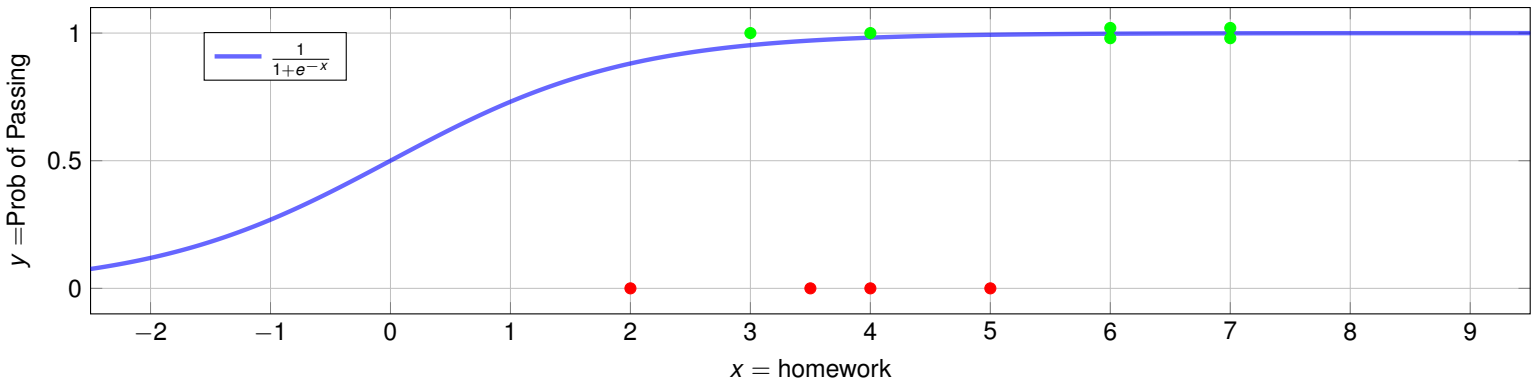
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the logistic function

$$f(x) = \frac{1}{1 + e^{-x}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

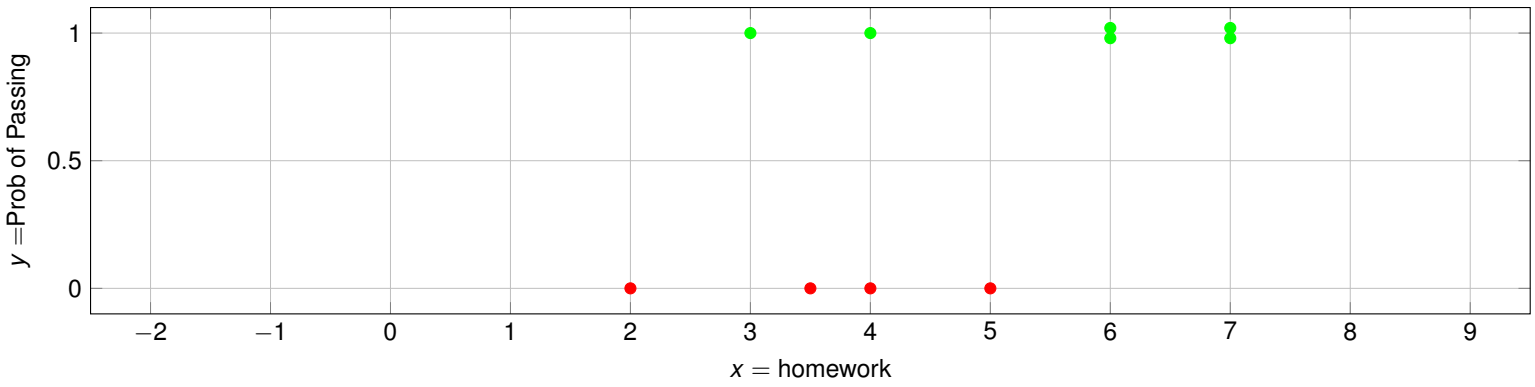
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{-x+5.25}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

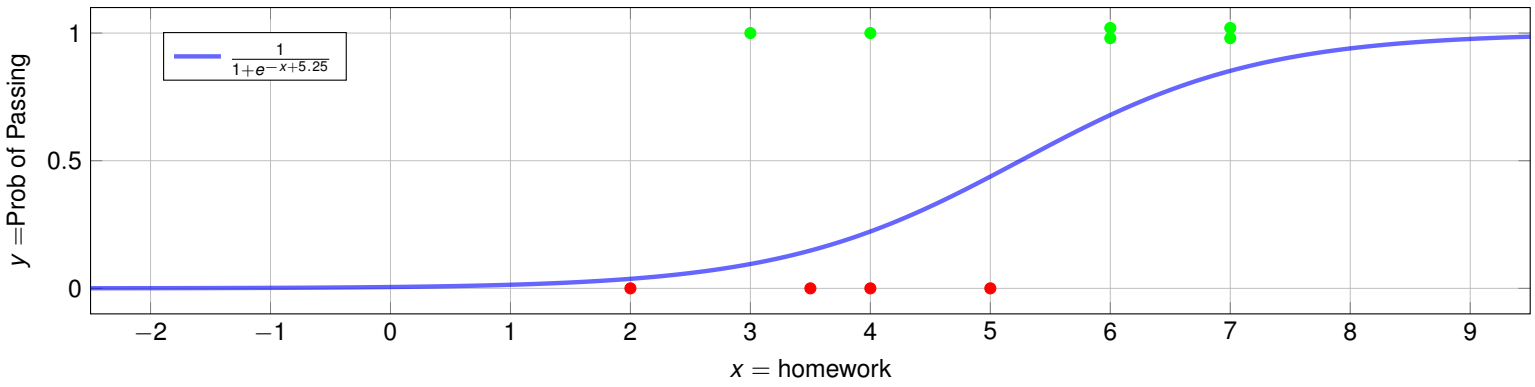
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{-x+5.25}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

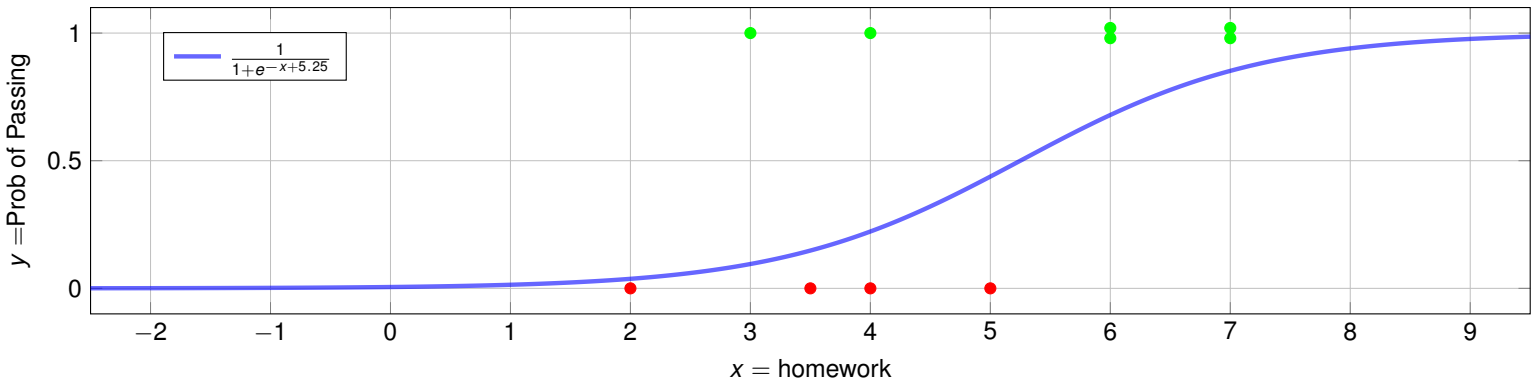
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{-x+5.25}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

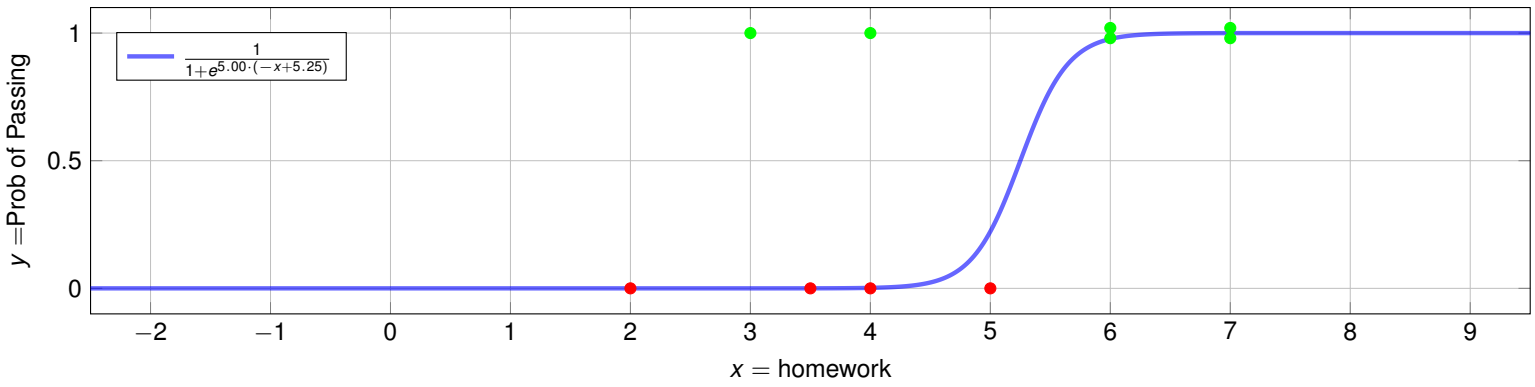
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{5\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

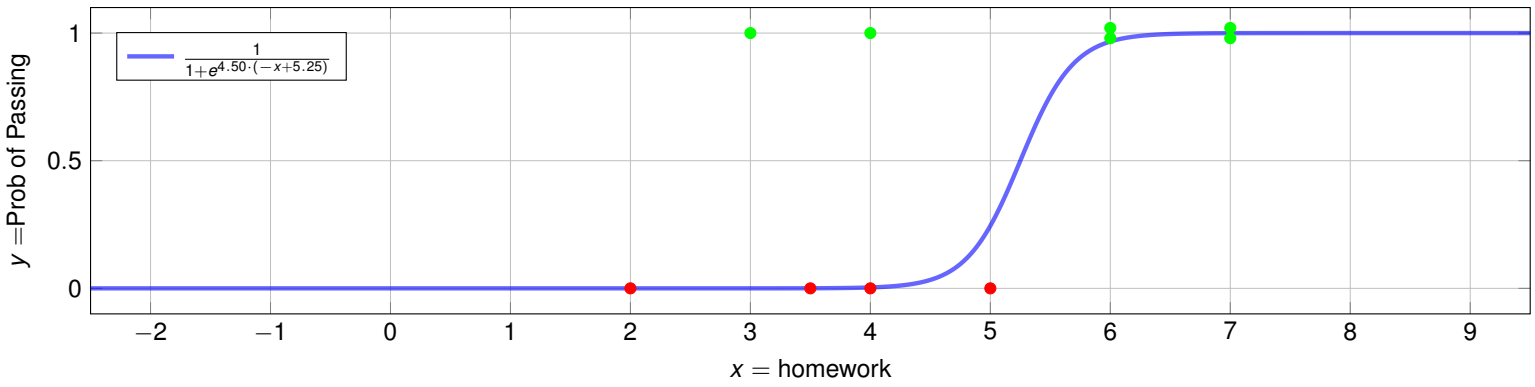
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{5\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

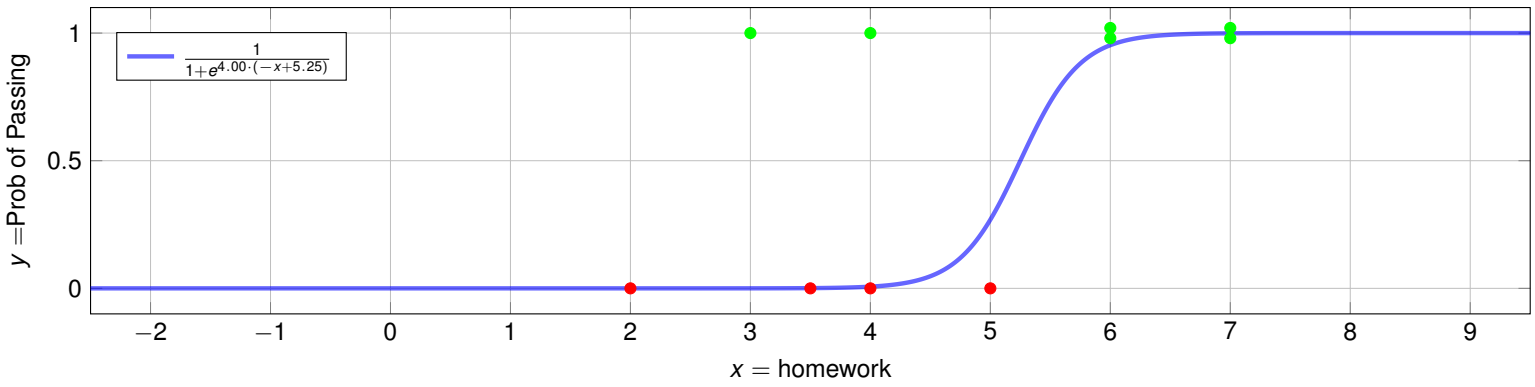
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

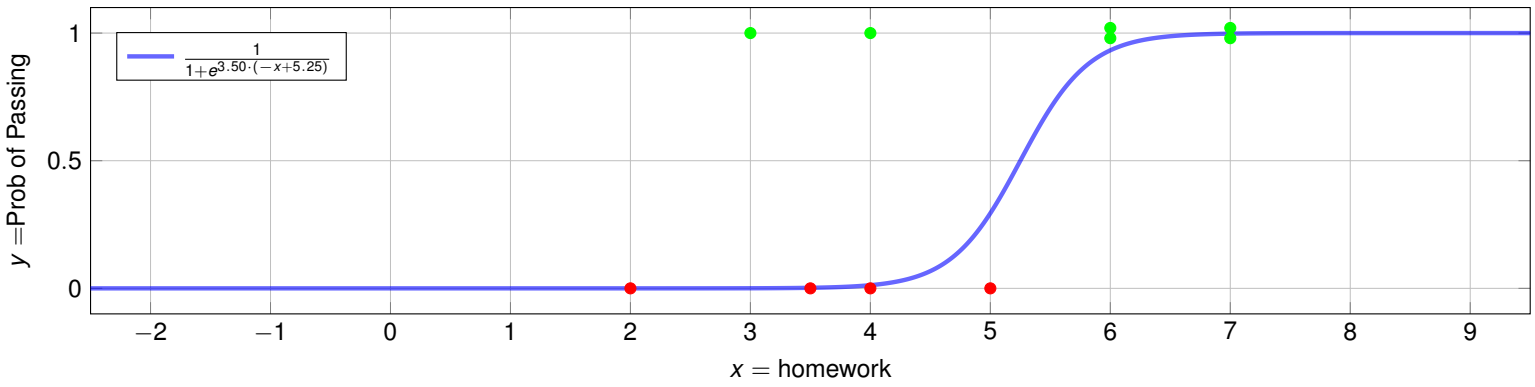
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

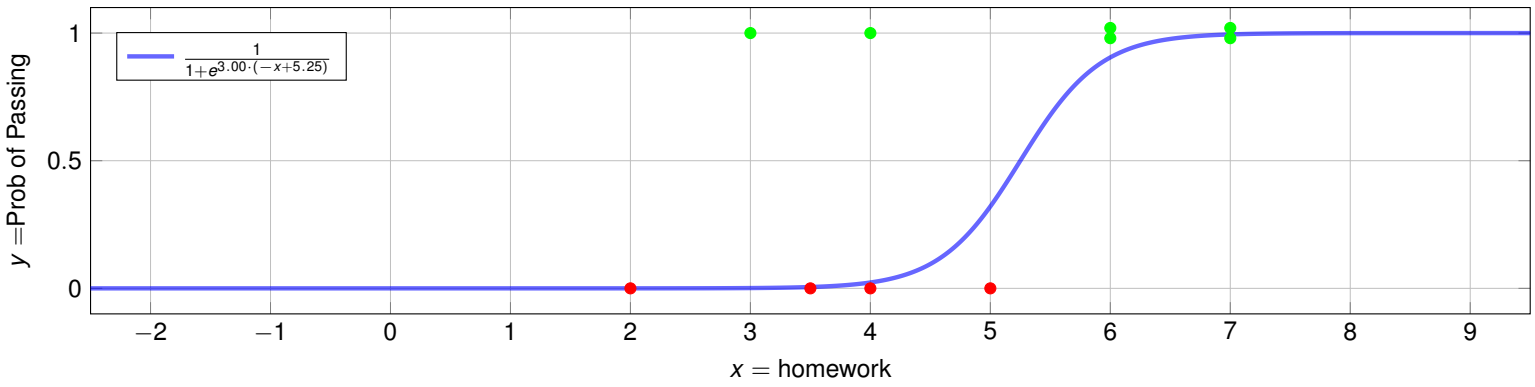
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

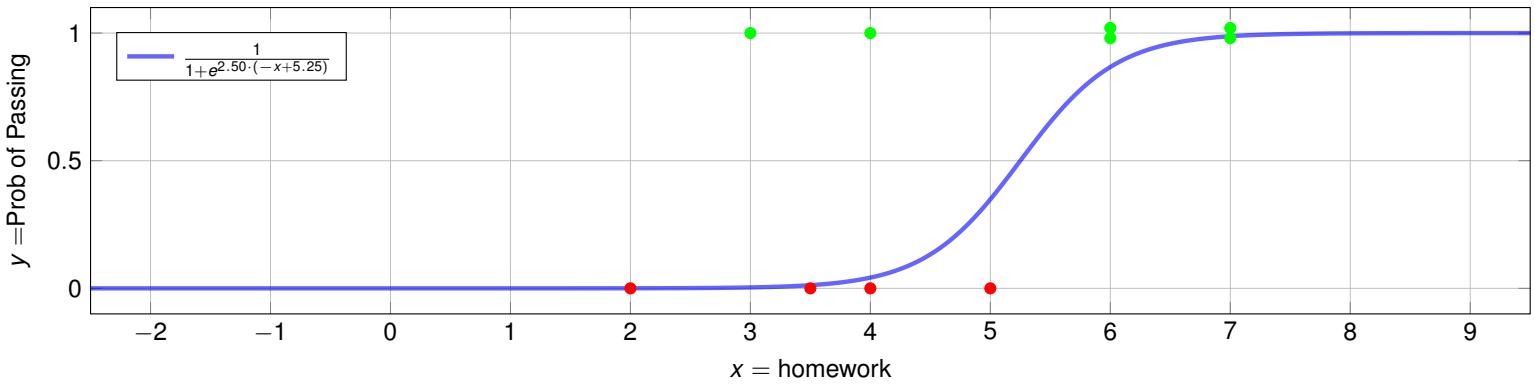
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

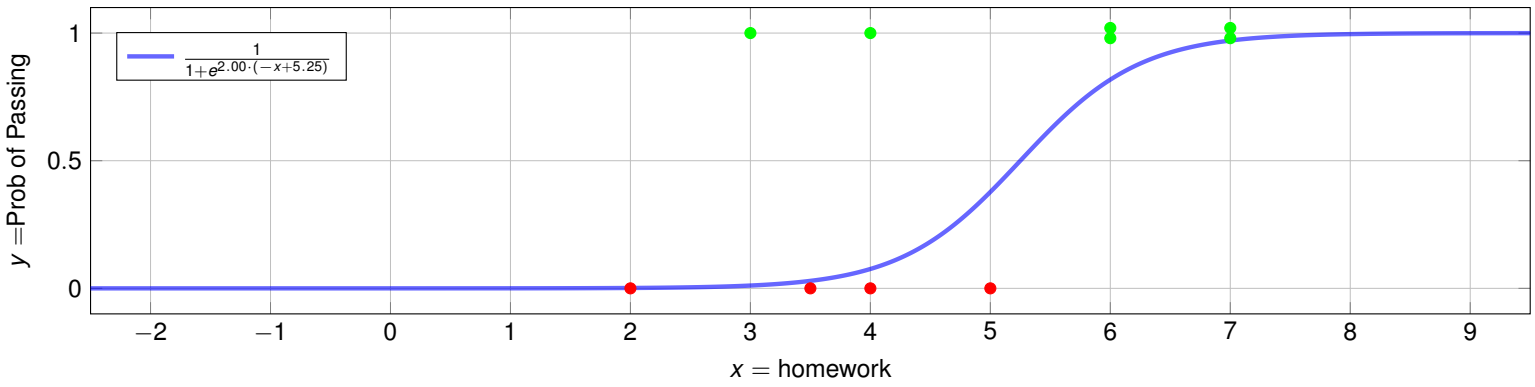
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

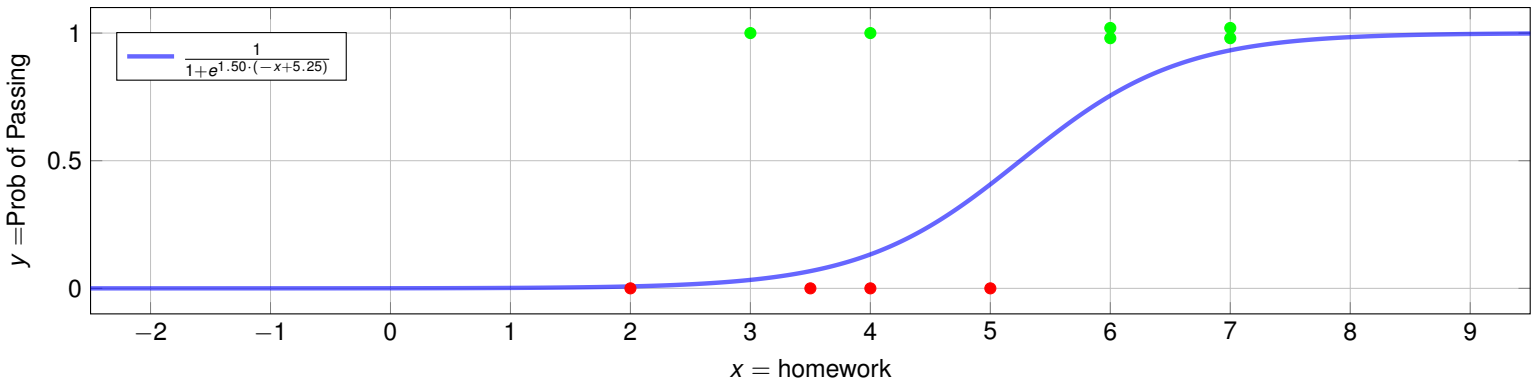
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

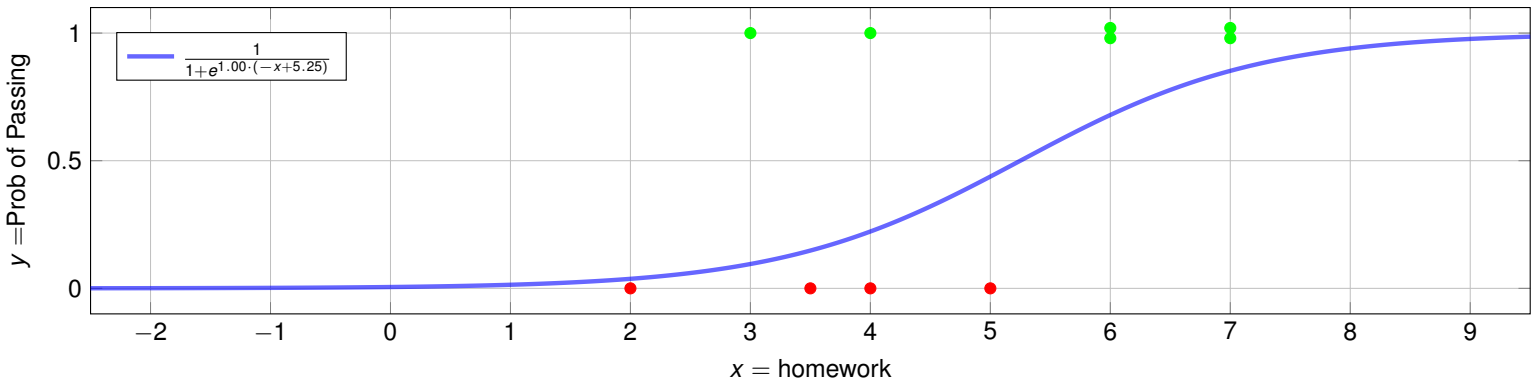
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

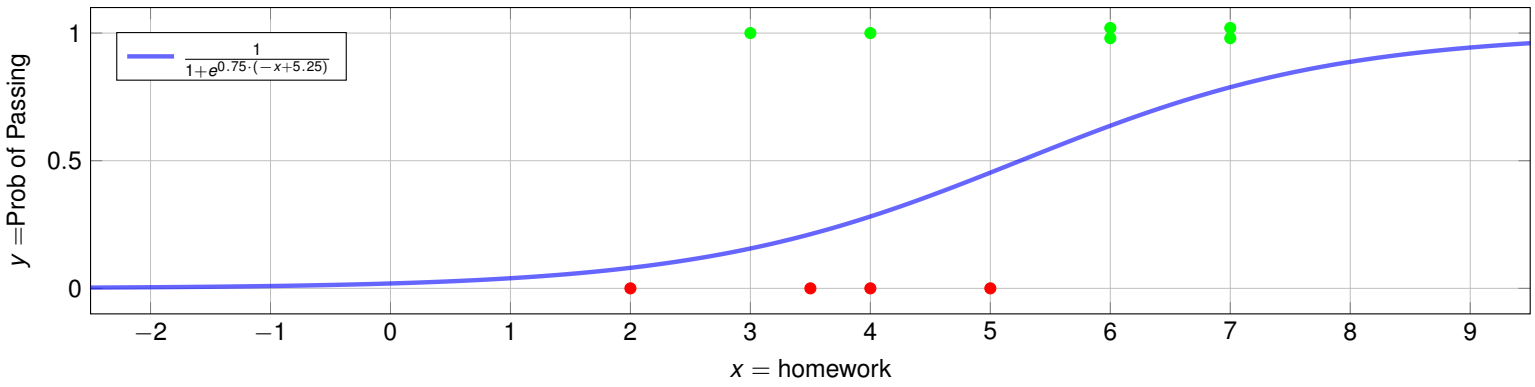
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

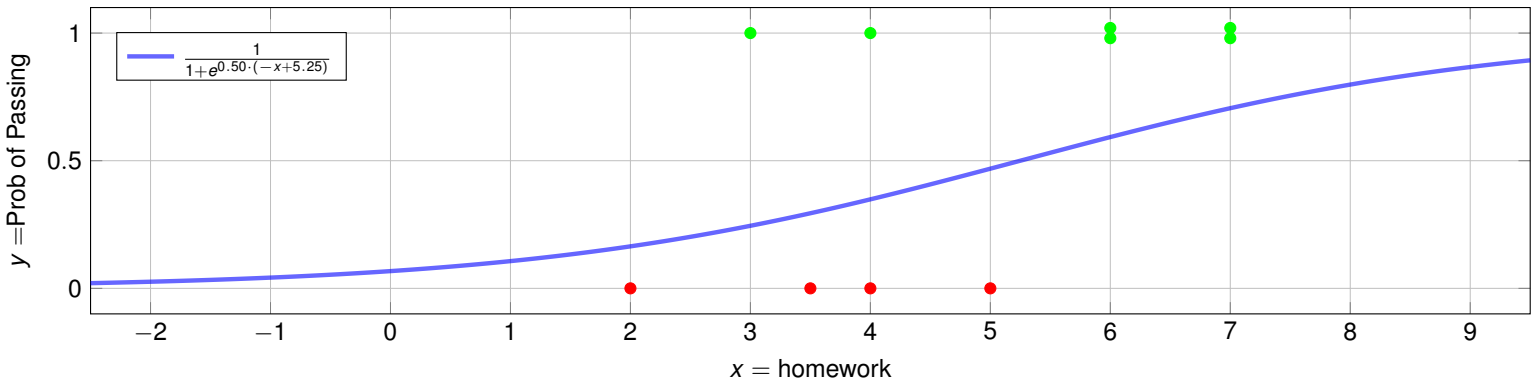
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

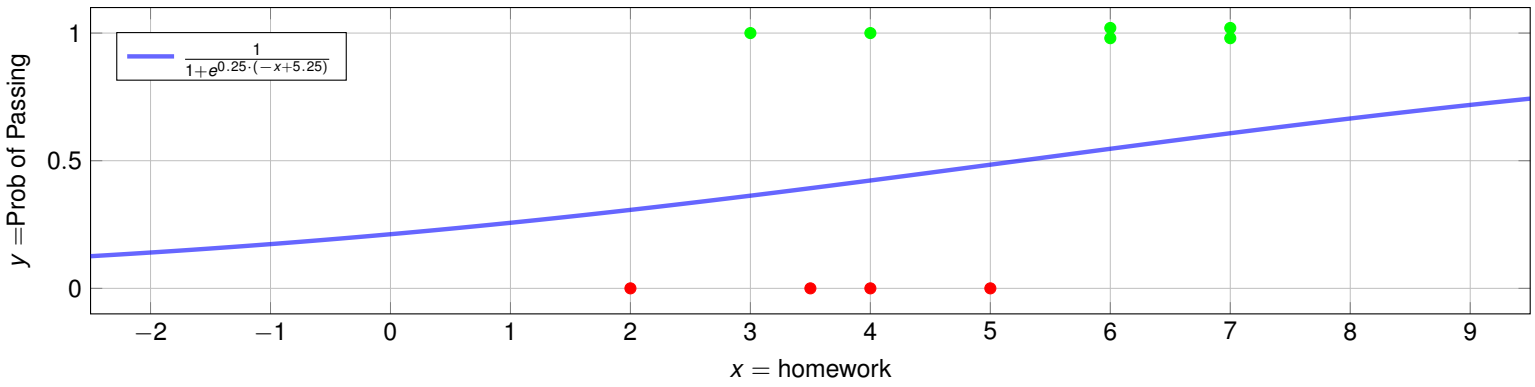
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

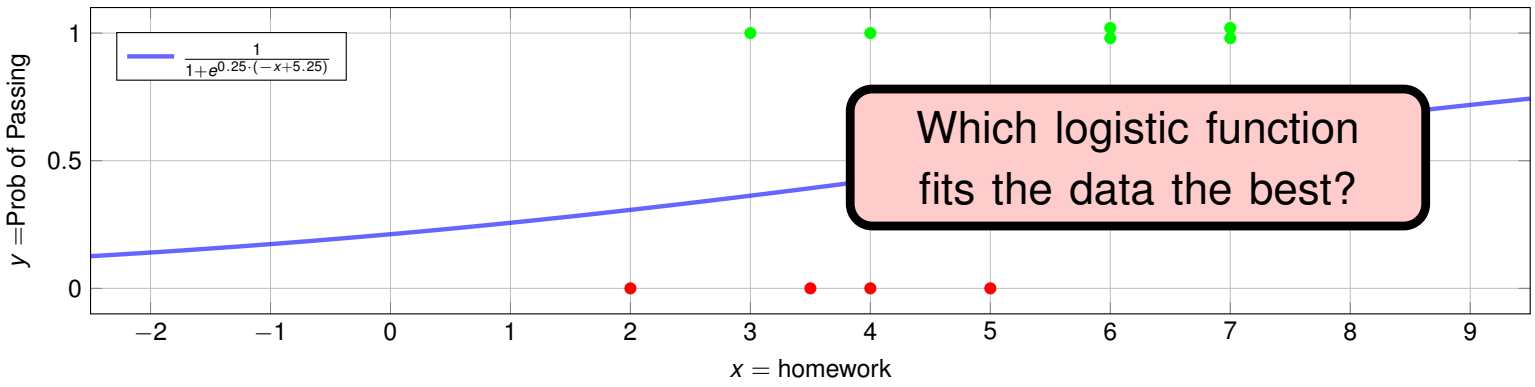
We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

The Logistic Function



- If student did homework very well, then high probability to pass exam
- If student did homework very poorly, then small probability to pass exam

We are looking for a function $f(x)$ such that:

- If x is large (≈ 10), then $f(x) \approx 1$
- If x is small (≈ 0), then $f(x) \approx 0$
- If x is around average (5.25), then $f(x) \approx 0.5$

Consider the general logistic function

$$f(x) = \frac{1}{1 + e^{??\cdot(-x+5.25)}}$$

Quantifying the Error

- Let x be homework and $y(x)$ be probability of passing the exam

Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Let x be homework and $y(x)$ be probability of passing the exam

Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Let x be homework and $y(x)$ be probability of passing the exam
- $b + w \cdot x$ is the the same as in linear regression

Quantifying the Error

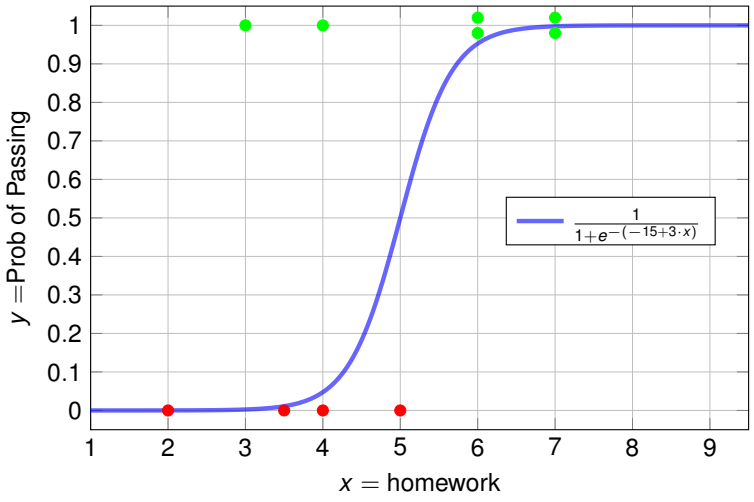
$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Let x be homework and $y(x)$ be probability of passing the exam
- $b + w \cdot x$ is the the same as in linear regression
- Function $\frac{1}{1+\exp(-)}$ ensures we get value between 0 and 1 (a probability!)

Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

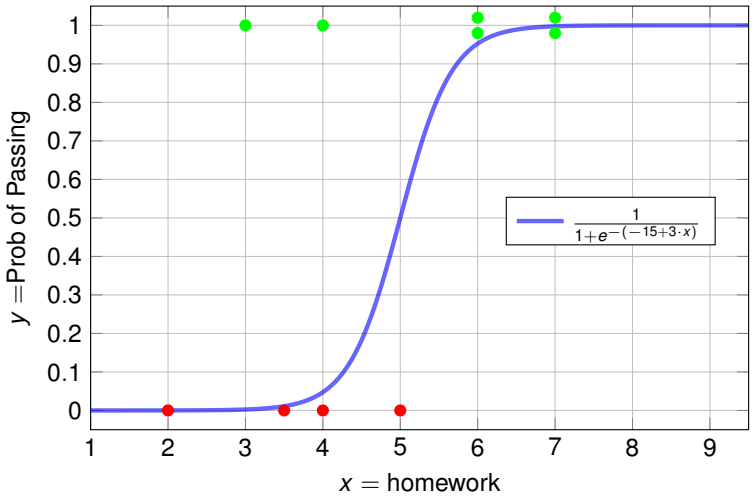
- Let x be homework and $y(x)$ be probability of passing the exam
- $b + w \cdot x$ is the the same as in linear regression
- Function $\frac{1}{1+\exp(-)}$ ensures we get value between 0 and 1 (a probability!)

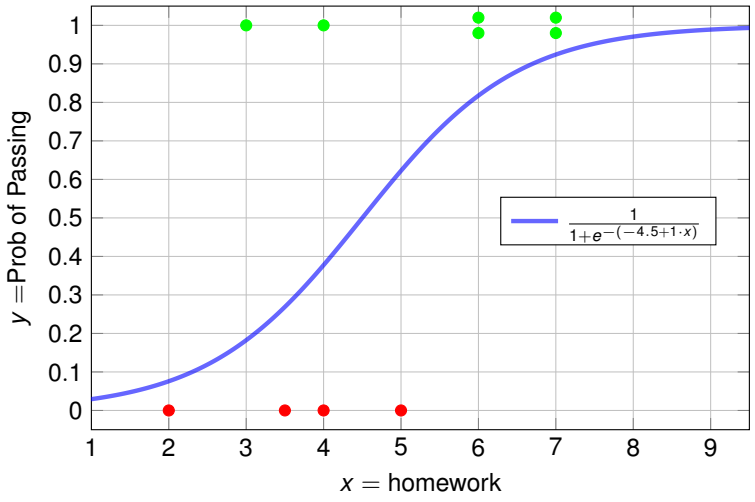


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

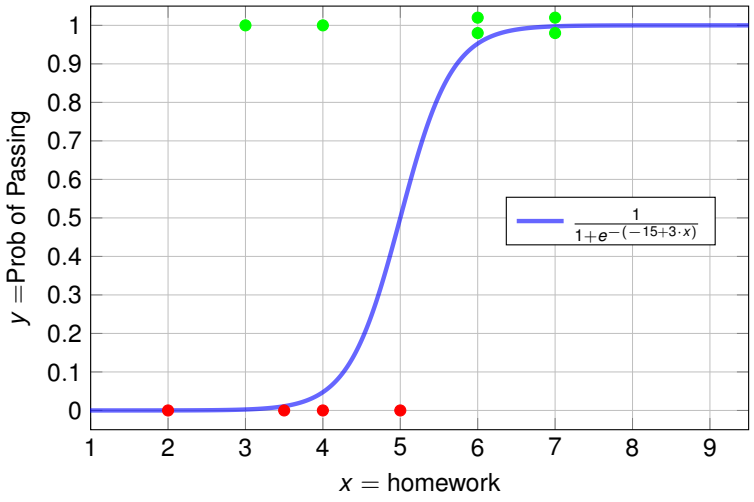
- Let x be homework and $y(x)$ be probability of passing the exam
- $b + w \cdot x$ is the the same as in linear regression
- Function $\frac{1}{1+\exp(-)}$ ensures we get value between 0 and 1 (a probability!)

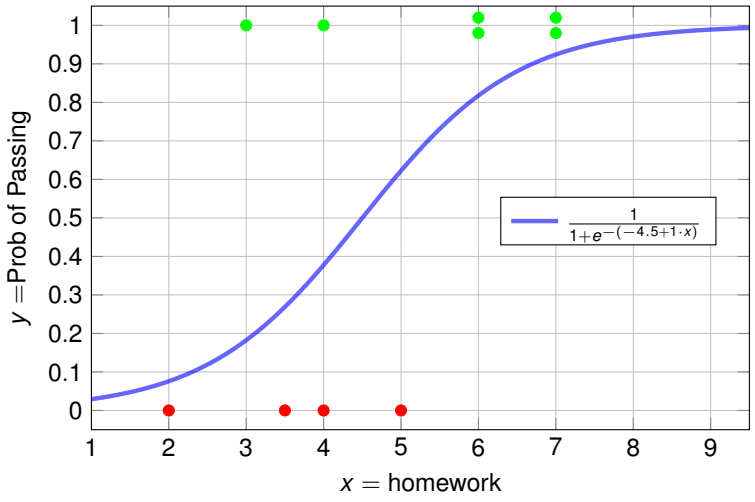




Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

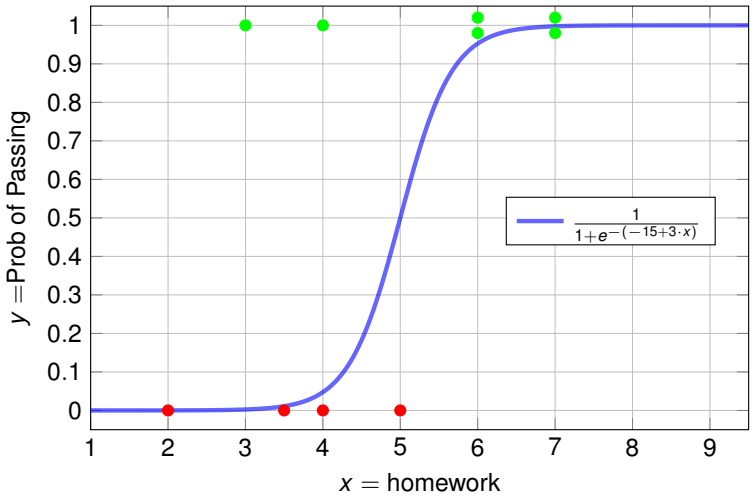


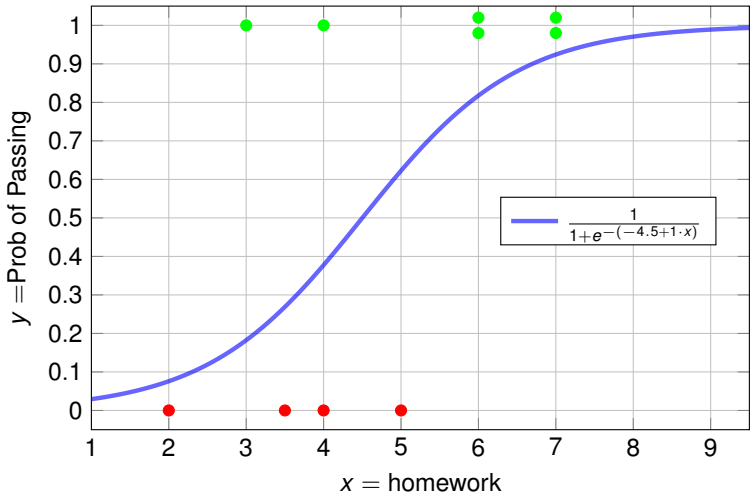


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model:

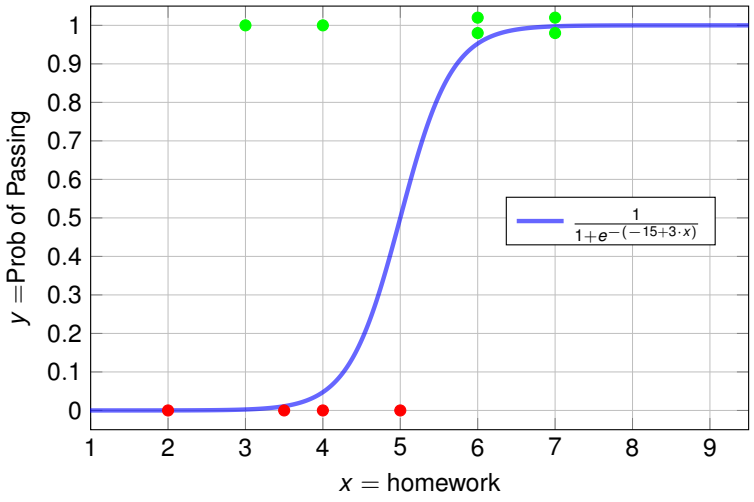


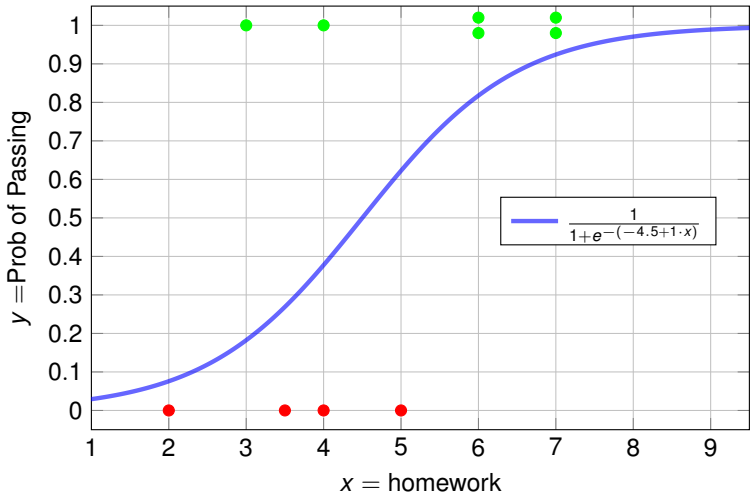


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$

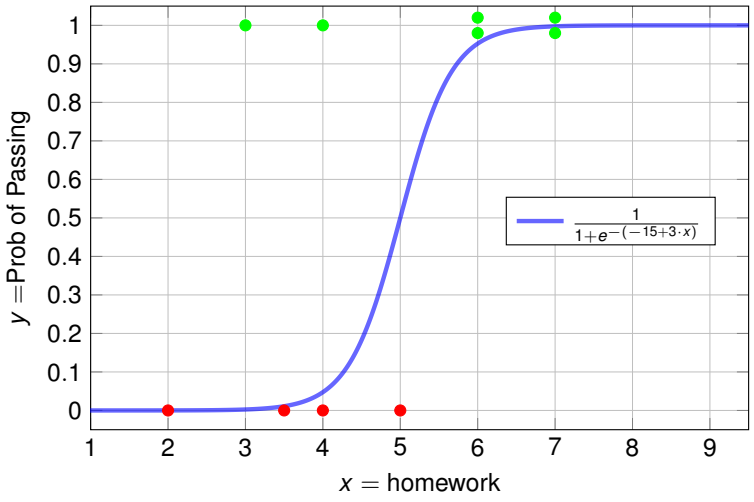


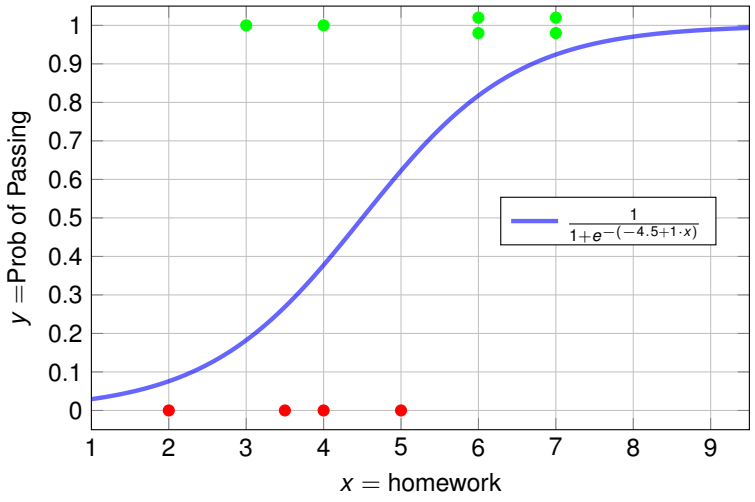


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?

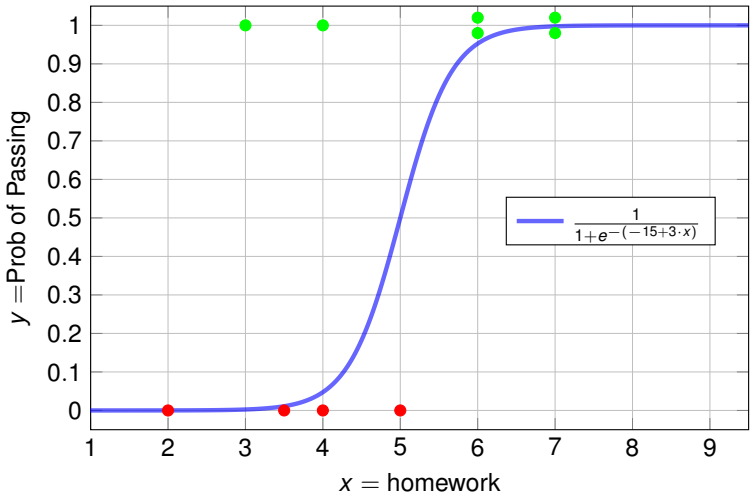


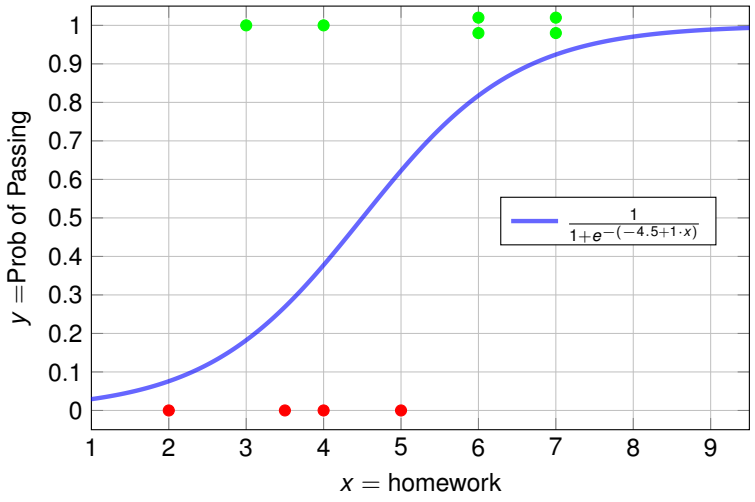


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students:

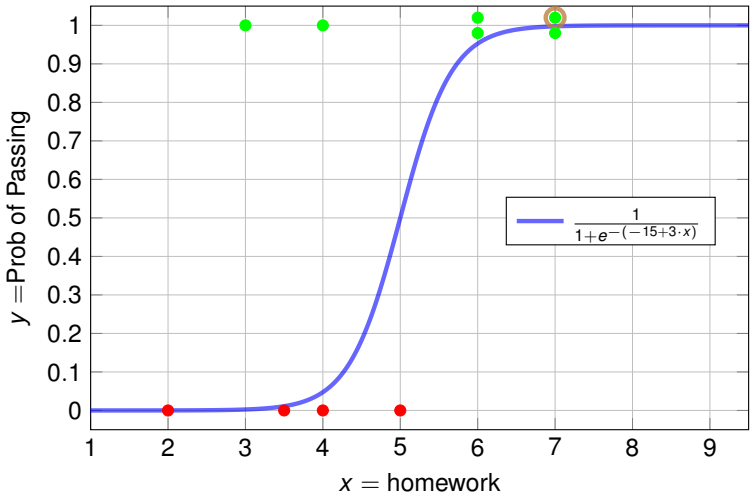


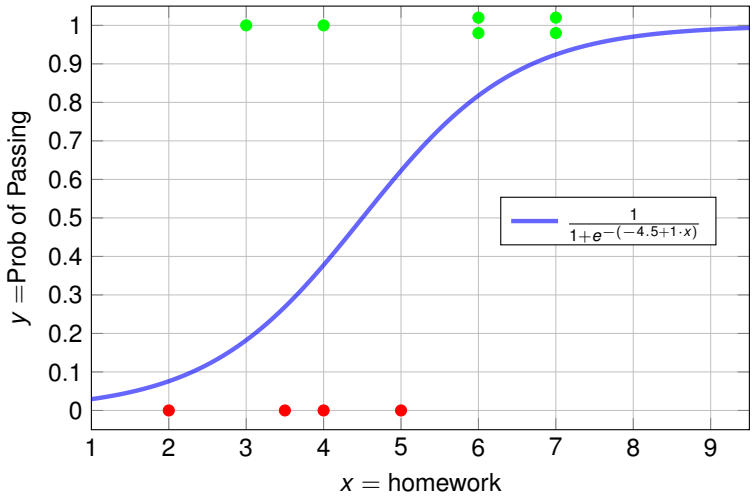


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: 0.99 $\times$

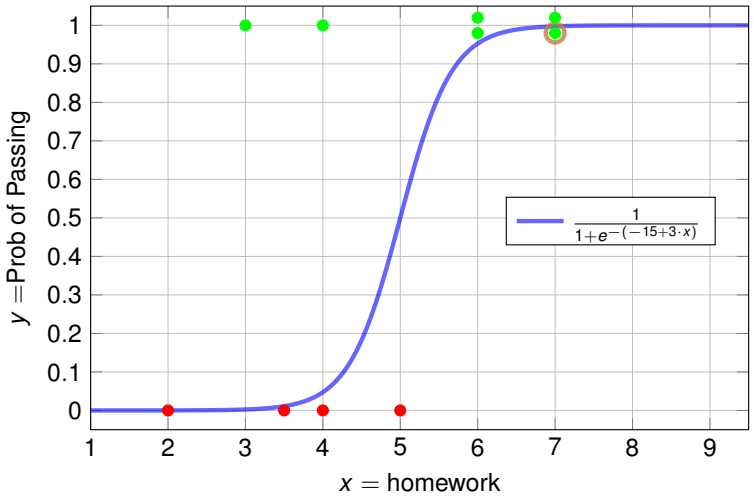


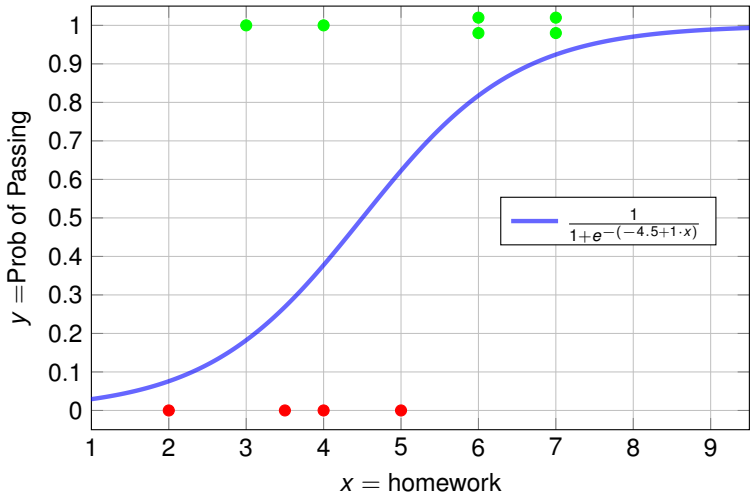


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times$

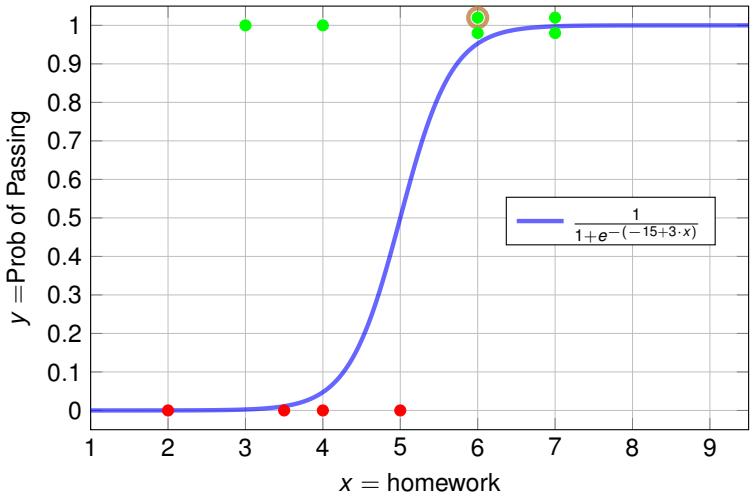


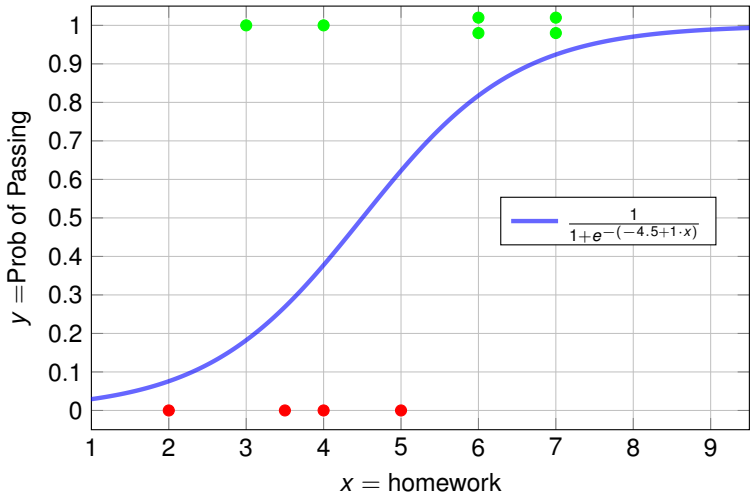


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times$

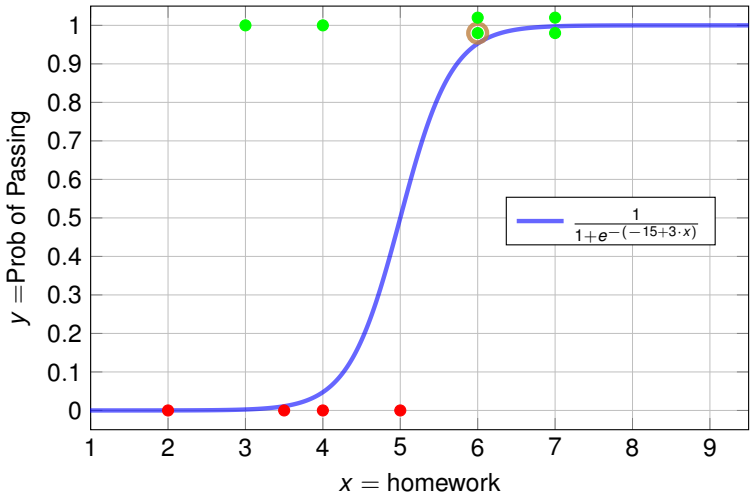


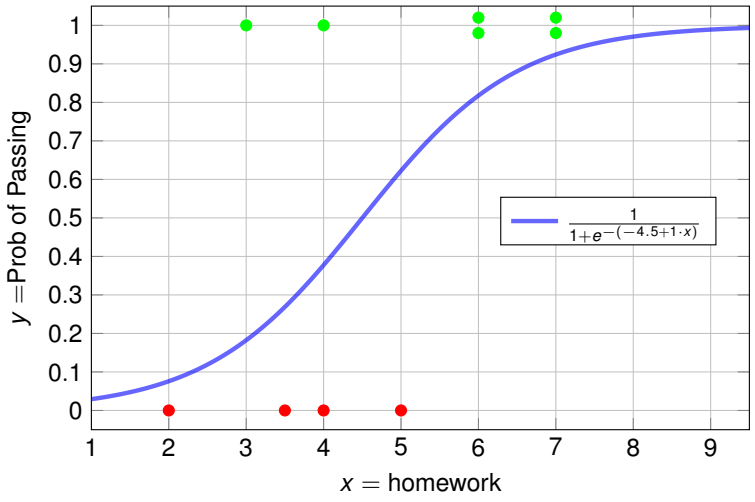


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times$

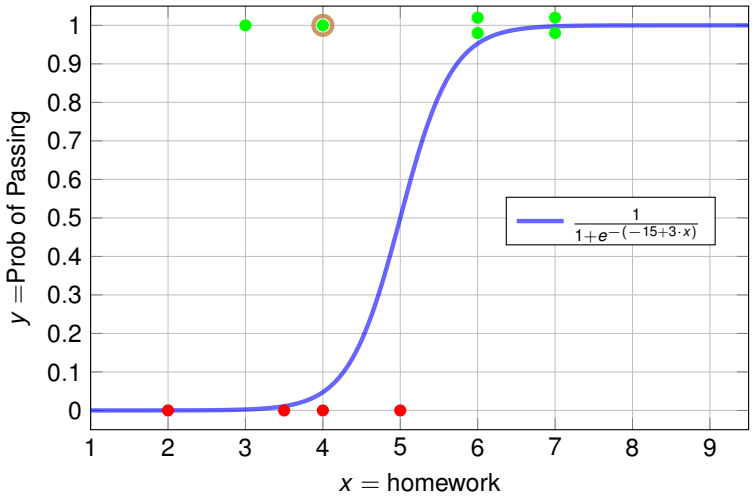


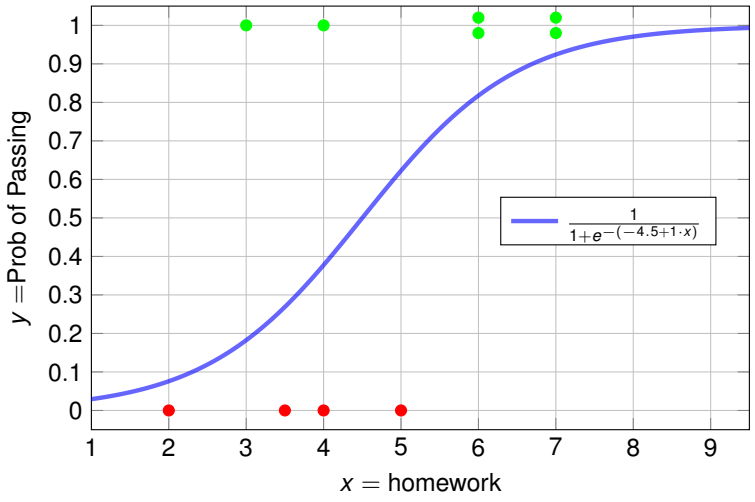


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times$

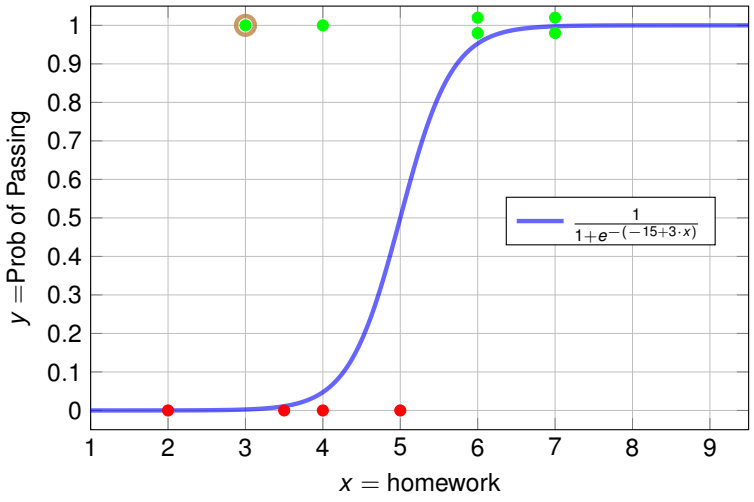


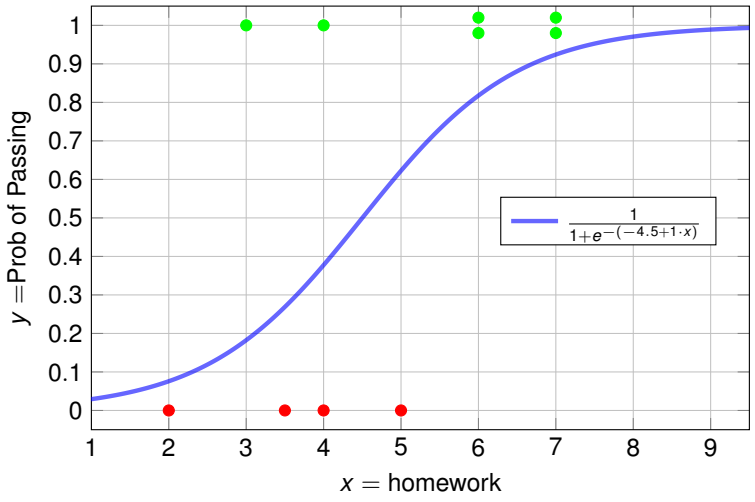


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002$

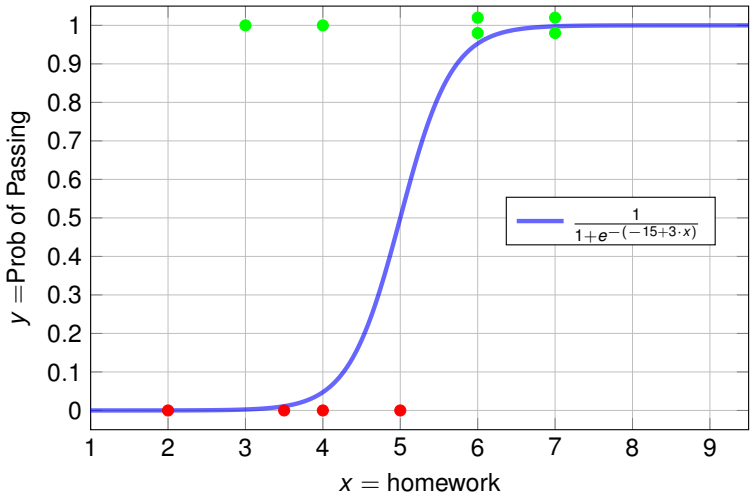


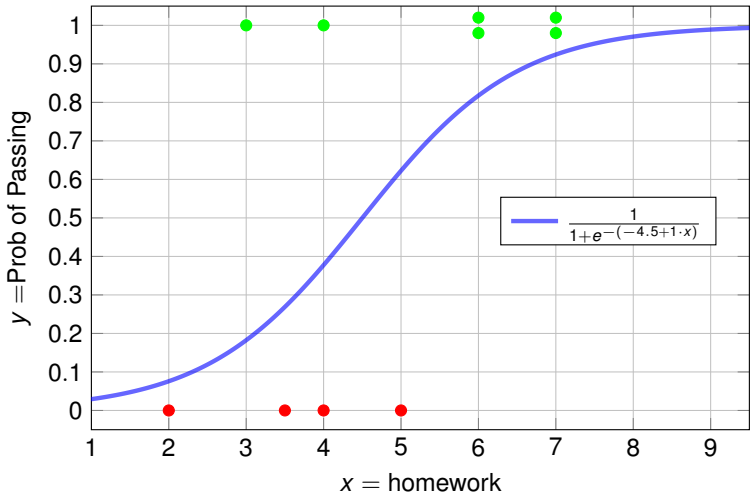


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$

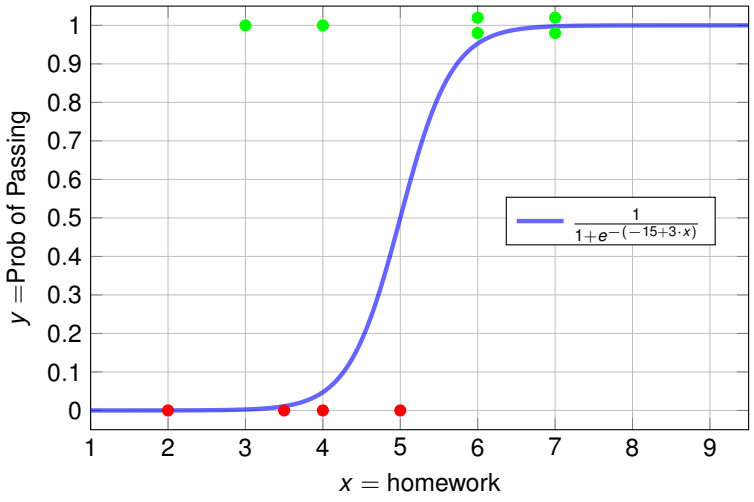


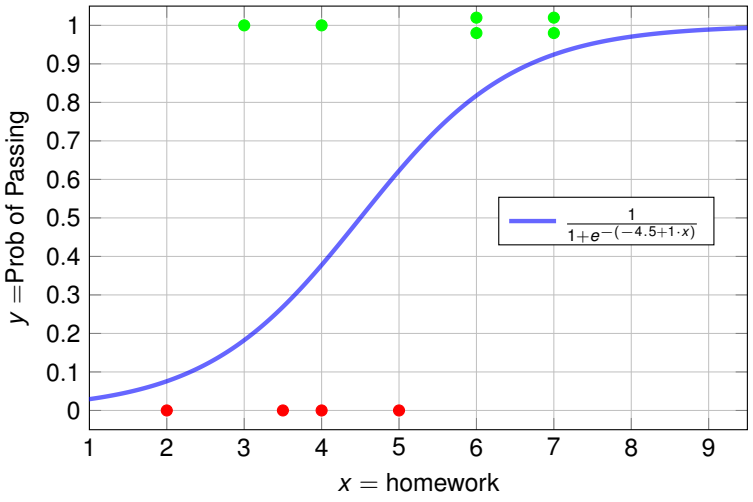


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students:

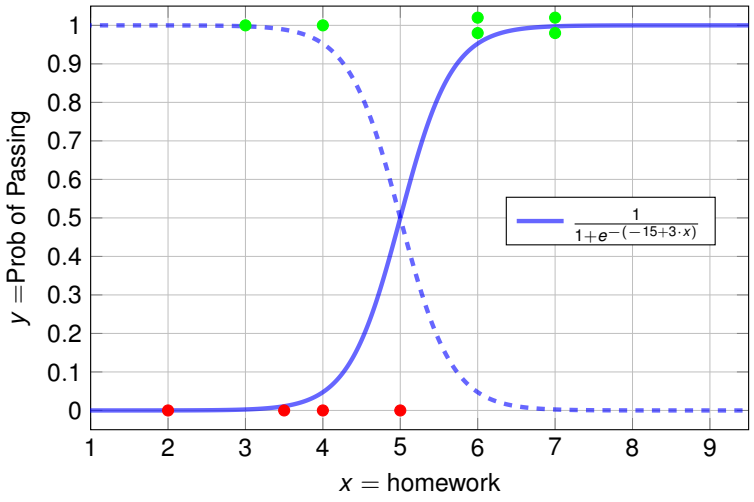


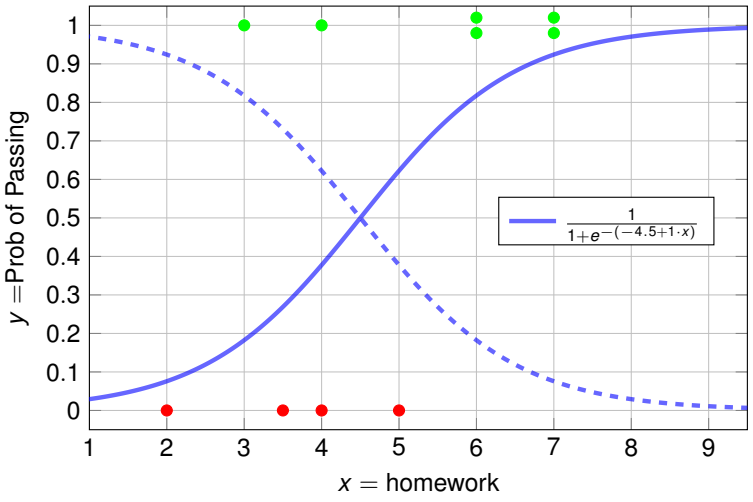


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students:

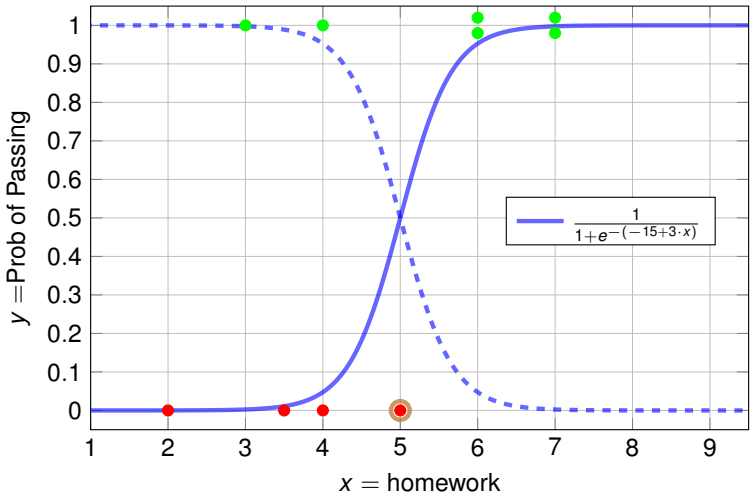


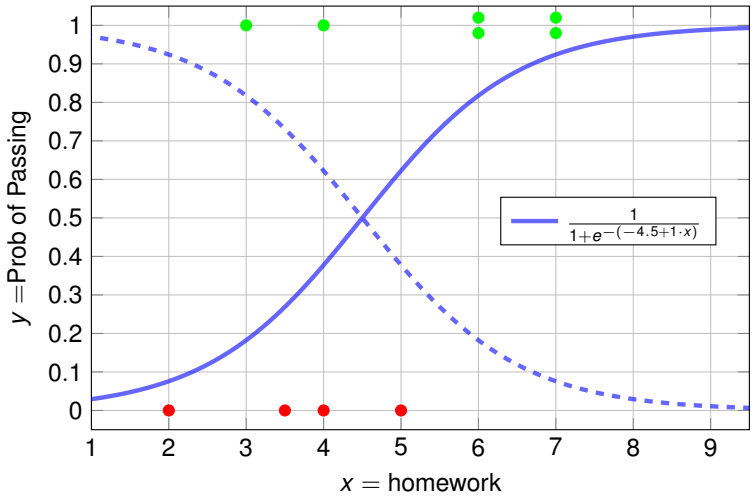


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students: $0.5 \times$

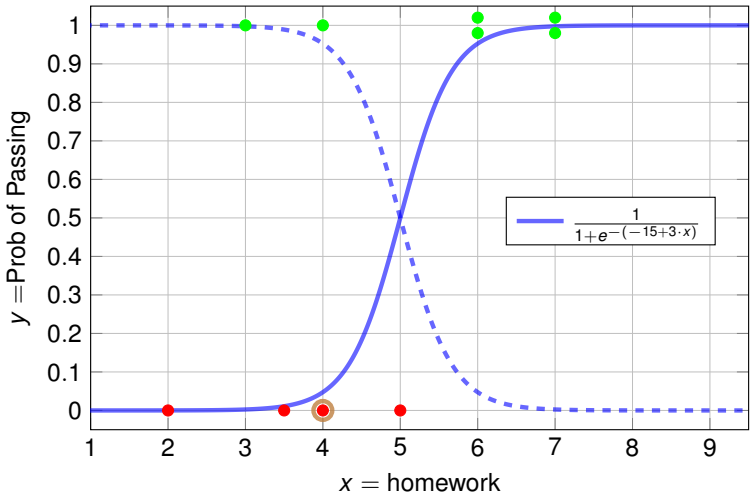


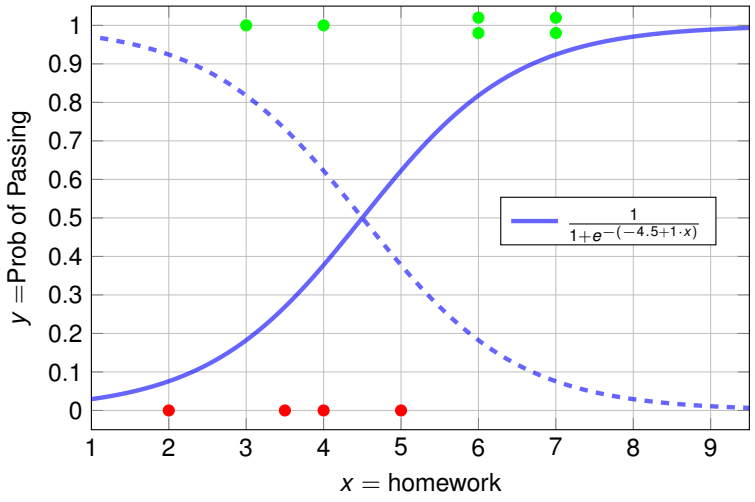


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students: $0.5 \times 0.95 \times$

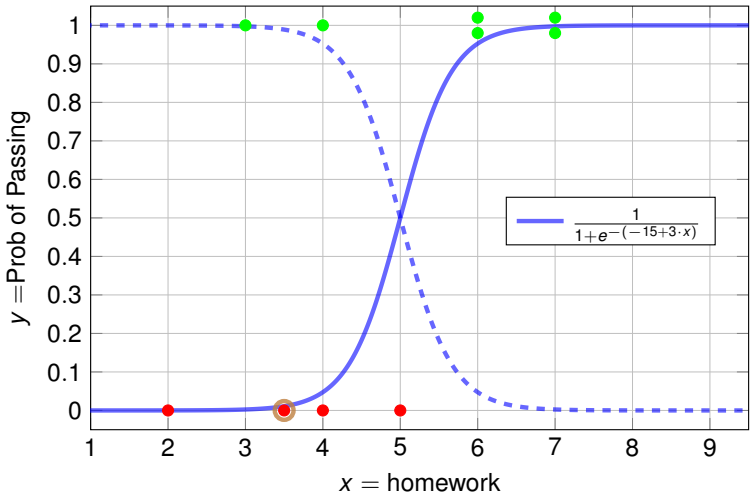


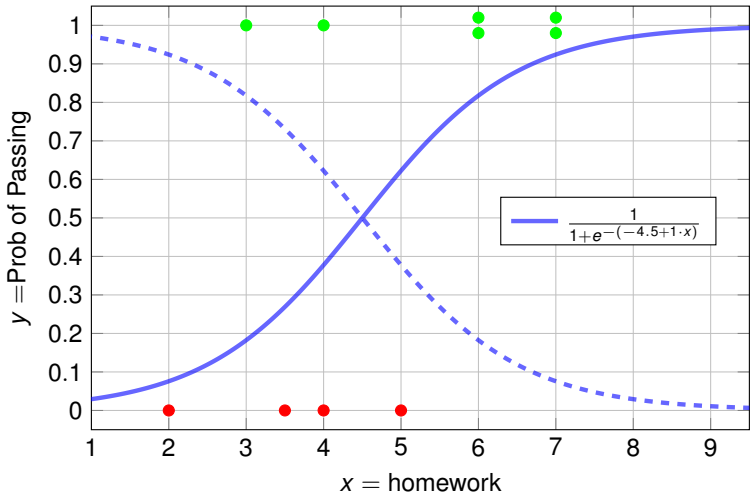


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students: $0.5 \times 0.95 \times 0.98 \times$

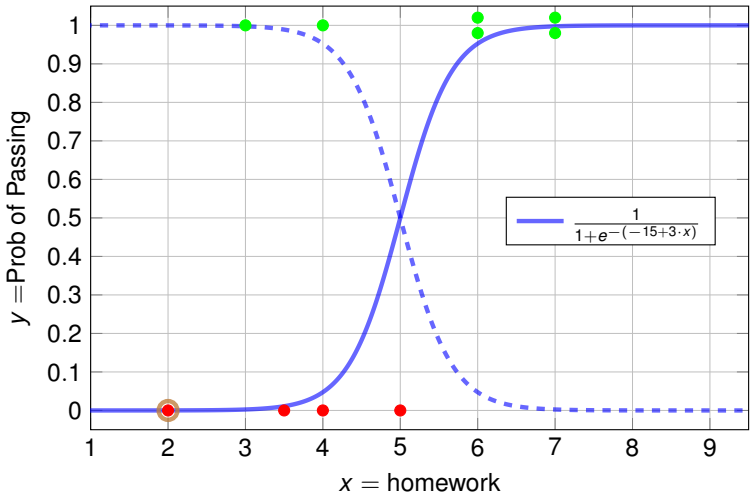


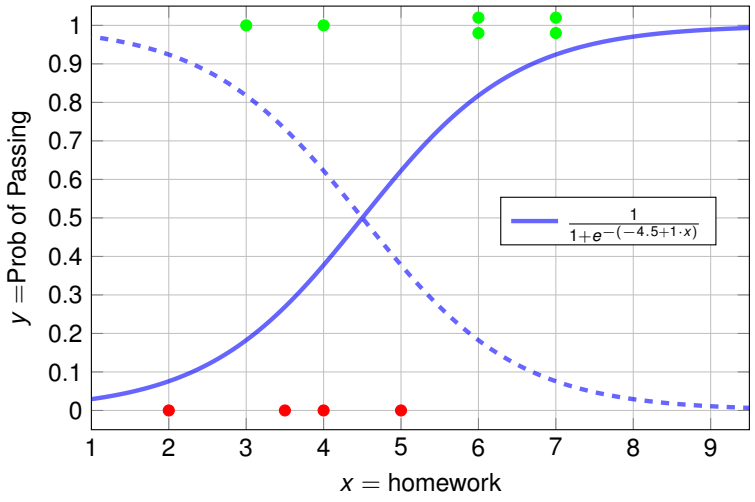


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students: $0.5 \times 0.95 \times 0.98 \times 0.99$

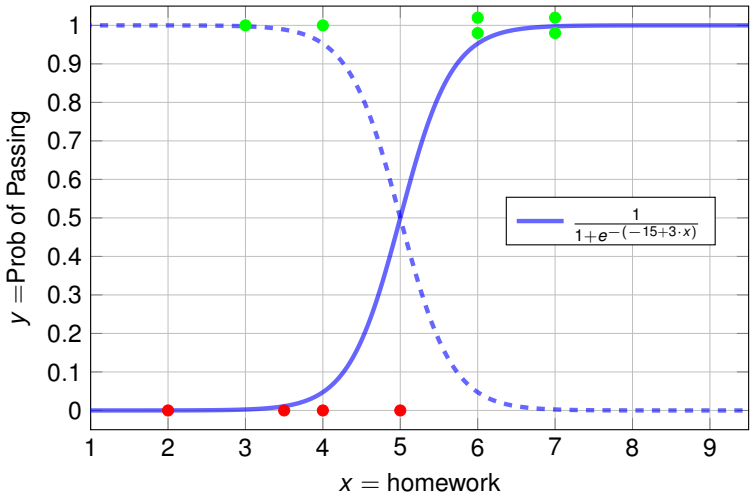


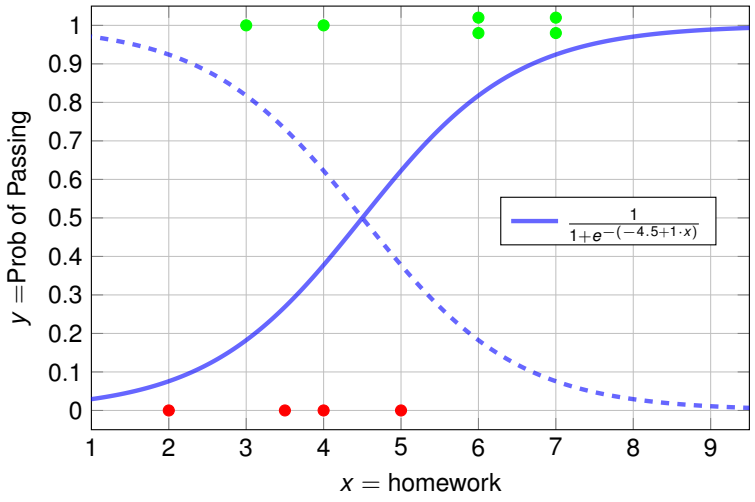


Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students: $0.5 \times 0.95 \times 0.98 \times 0.99 \approx 0.47$



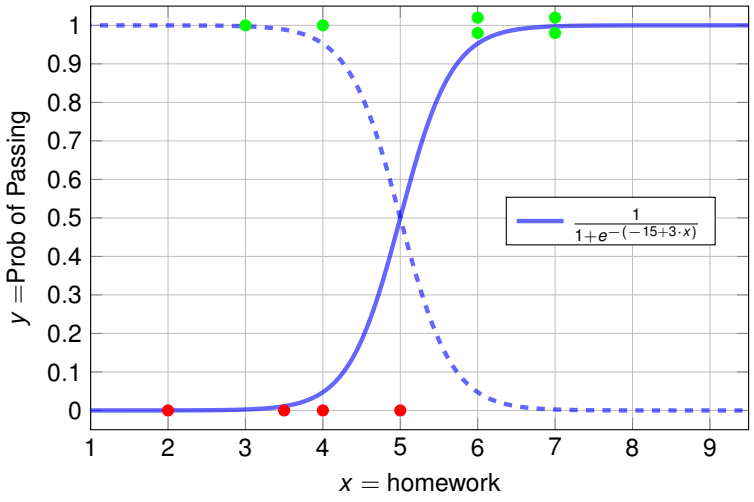


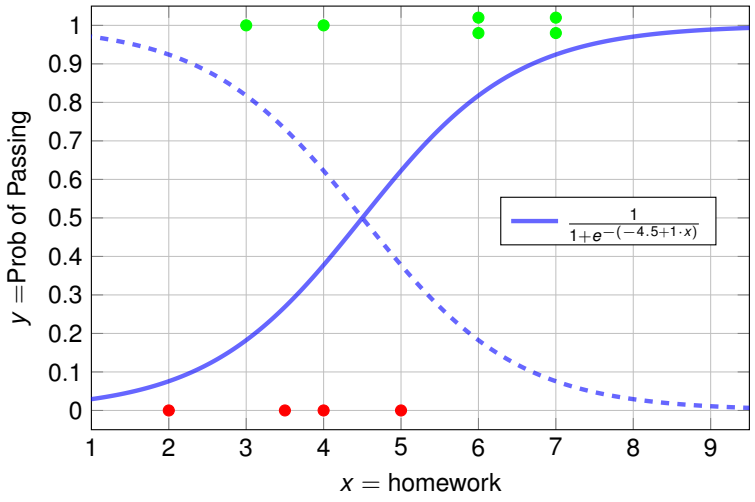
Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students: $0.5 \times 0.95 \times 0.98 \times 0.99 \approx 0.47$

  $\Rightarrow$ overall likelihood is



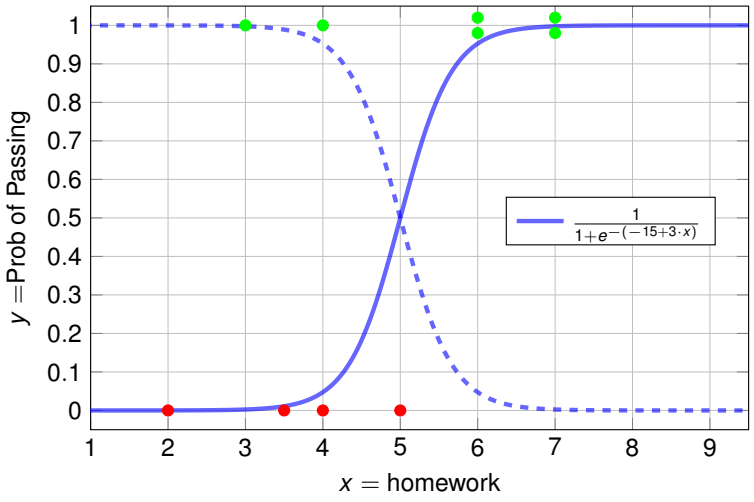


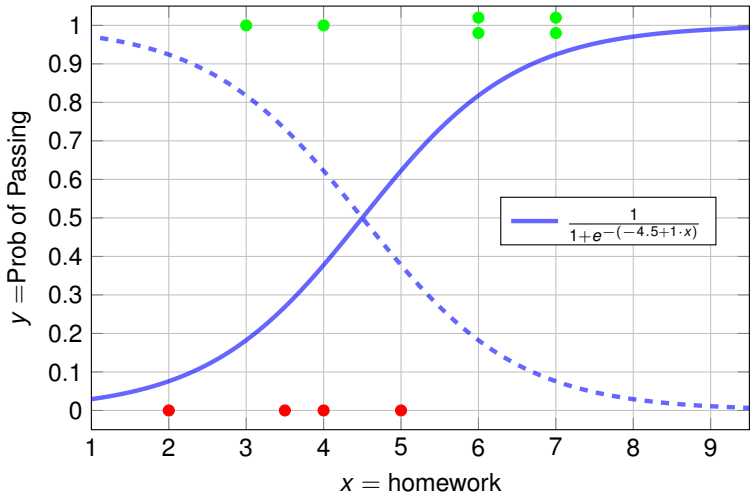
Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students: $0.5 \times 0.95 \times 0.98 \times 0.99 \approx 0.47$

  $\Rightarrow$ overall likelihood is $0.0001 \cdot 0.47 \approx 0.00005$





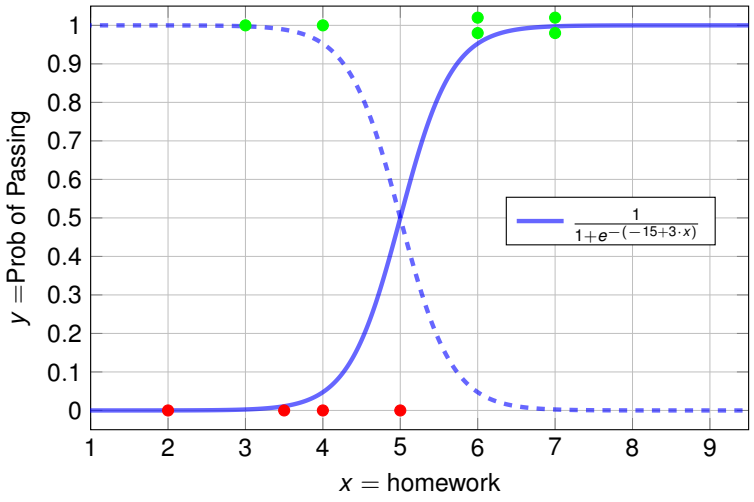
Quantifying the Error

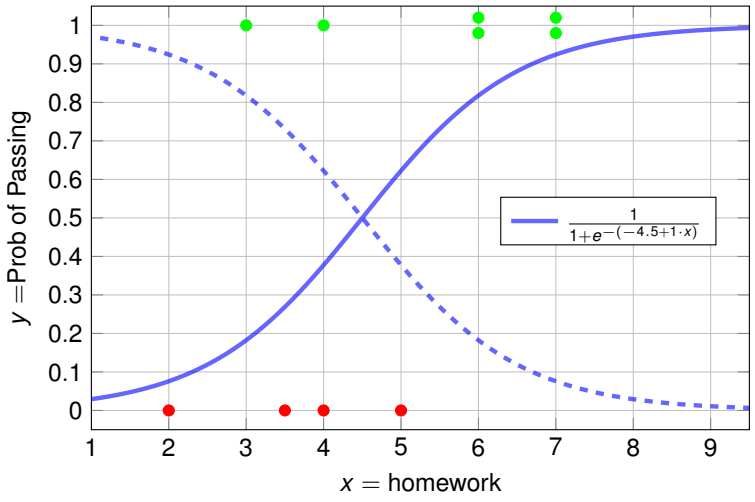
$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students: $0.5 \times 0.95 \times 0.98 \times 0.99 \approx 0.47$

  $\Rightarrow$ overall likelihood is $0.0001 \cdot 0.47 \approx 0.00005$

Likelihood ≈ 0.00005





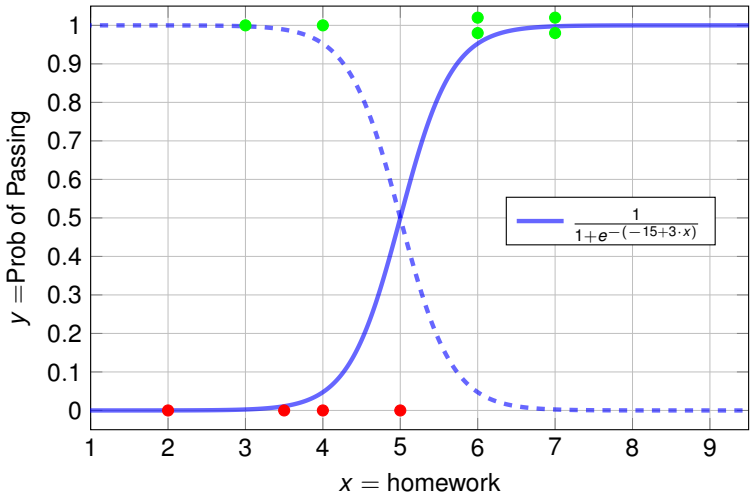
Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

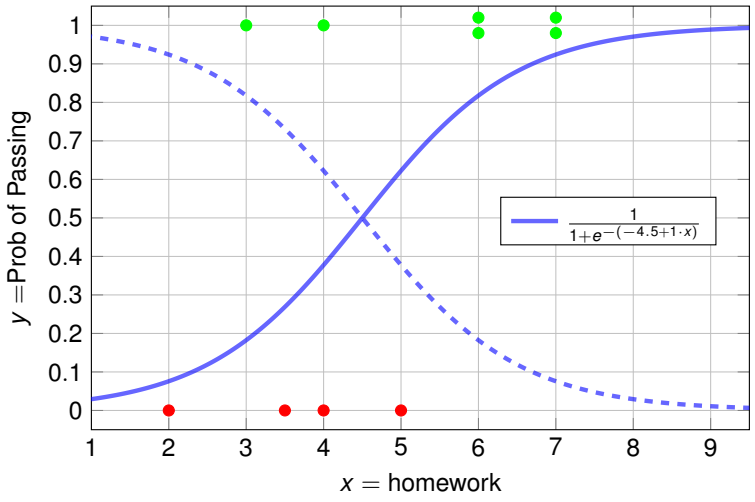
- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students: $0.5 \times 0.95 \times 0.98 \times 0.99 \approx 0.47$

  $\Rightarrow$ overall likelihood is $0.0001 \cdot 0.47 \approx 0.00005$

Likelihood ≈ 0.00005



Likelihood ≈ 0.006



Quantifying the Error

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

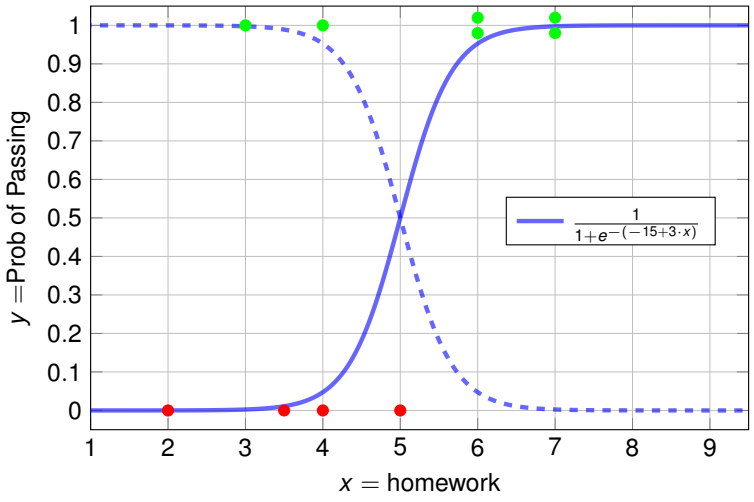
- Assume data points were generated exactly by our probabilistic model: Given a student with homework x, he/she will pass with probability $y(x)$
- What would be the probability (likelihood) for generating our data set?
  - passed students: $0.99 \times 0.99 \times 0.95 \times 0.95 \times 0.04 \times 0.002 \approx 0.0001$
  - failed students: $0.5 \times 0.95 \times 0.98 \times 0.99 \approx 0.47$

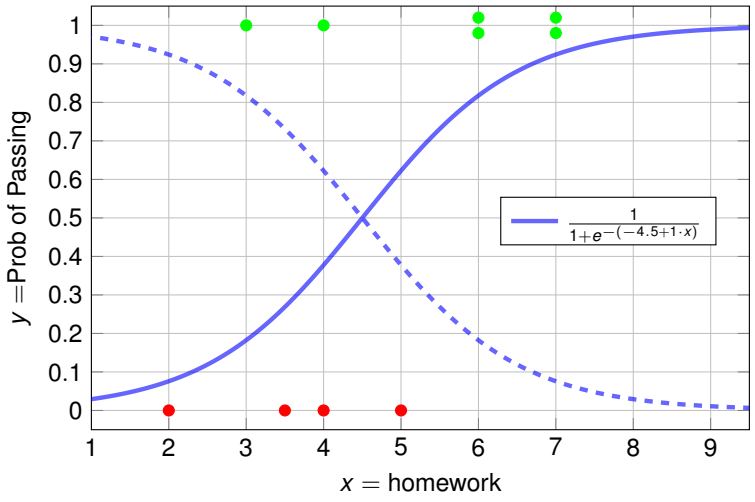
$\Rightarrow$ overall likelihood is $0.0001 \cdot 0.47 \approx 0.00005$

much better model!

Likelihood ≈ 0.00005

Likelihood ≈ 0.006





Finding the Optimal Model

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

Finding the Optimal Model

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model:
  Given a student with homework w, he/she will pass with probability $y(x)$

Finding the Optimal Model

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework w, he/she will pass with probability $y(x)$
- Find pair w and b that maximises the probability of creating our data set
- This is a more complex optimisation task (use Gradient Descent!)

Finding the Optimal Model

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework w, he/she will pass with probability $y(x)$
- Find pair w and b that maximises the probability of creating our data set
- This is a more complex optimisation task (use Gradient Descent!)

Optimal Fit

Finding the Optimal Model

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework w, he/she will pass with probability $y(x)$
- Find pair w and b that maximises the probability of creating our data set
- This is a more complex optimisation task (use Gradient Descent!)

Optimal Fit

- $b = -3.7349$, $w = 0.914$

Finding the Optimal Model

$$y = \frac{1}{1 + \exp\left(-(b + w \cdot x)\right)}$$

- Assume data points were generated exactly by our probabilistic model: Given a student with homework w, he/she will pass with probability $y(x)$
- Find pair w and b that maximises the probability of creating our data set
- This is a more complex optimisation task (use Gradient Descent!)

Optimal Fit

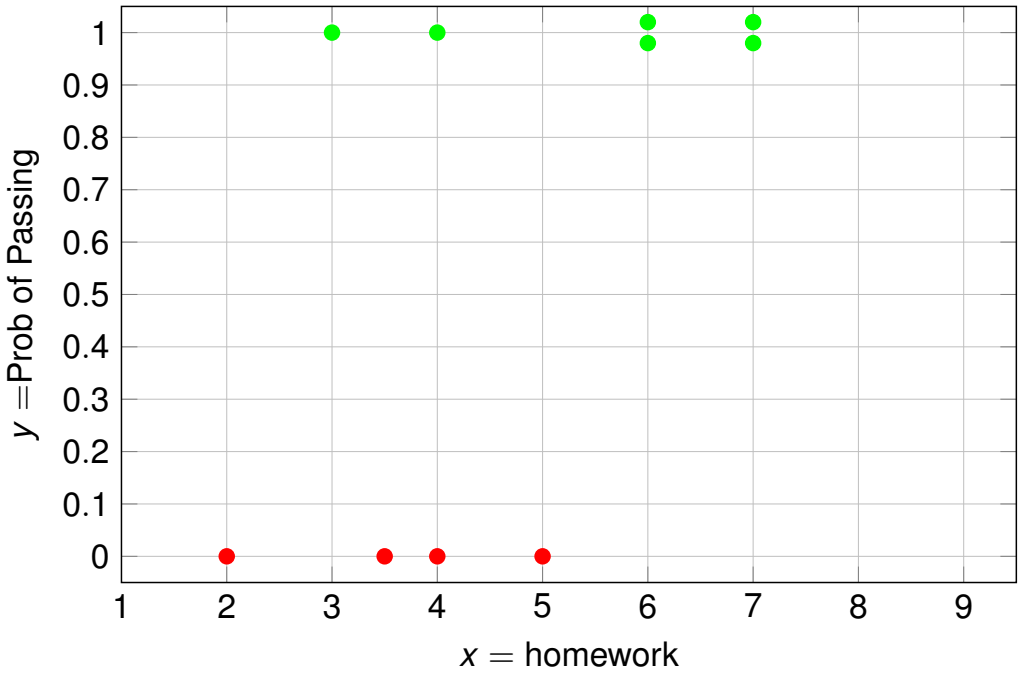
- $b = -3.7349$, $w = 0.914$

$\Rightarrow$ likelihood of 0.0071.

Plot of the Optimal Model

Optimal Model

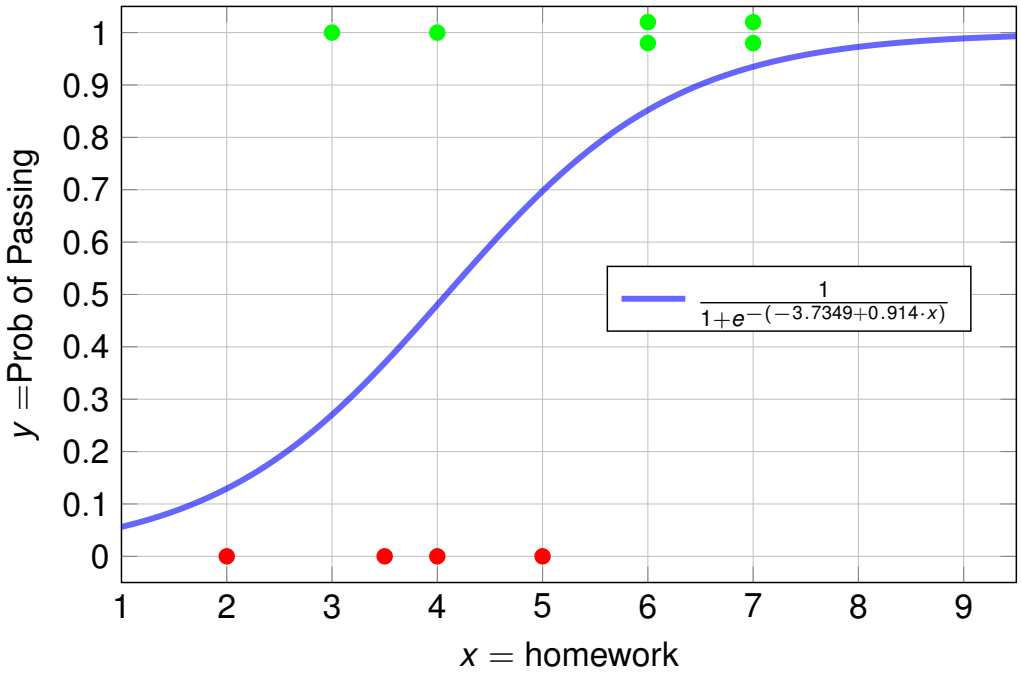
- $b = -3.7349$, $w = 0.914$, likelihood of 0.0071.



Plot of the Optimal Model

Optimal Model

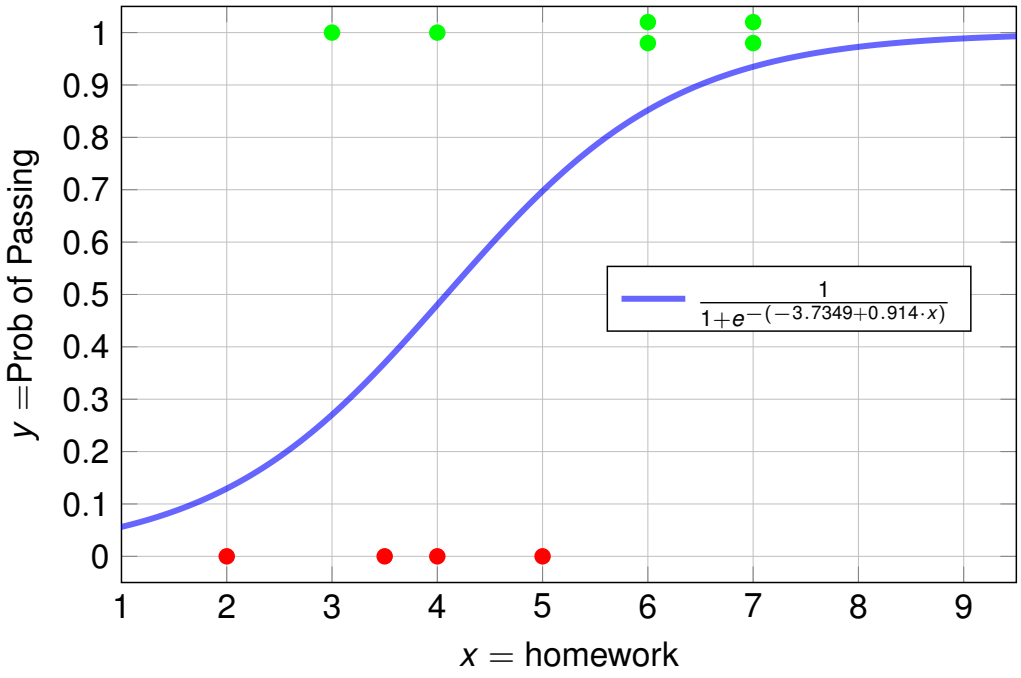
- $b = -3.7349$, $w = 0.914$, likelihood of 0.0071.



Plot of the Optimal Model

Optimal Model

- $b = -3.7349$, $w = 0.914$, likelihood of 0.0071.



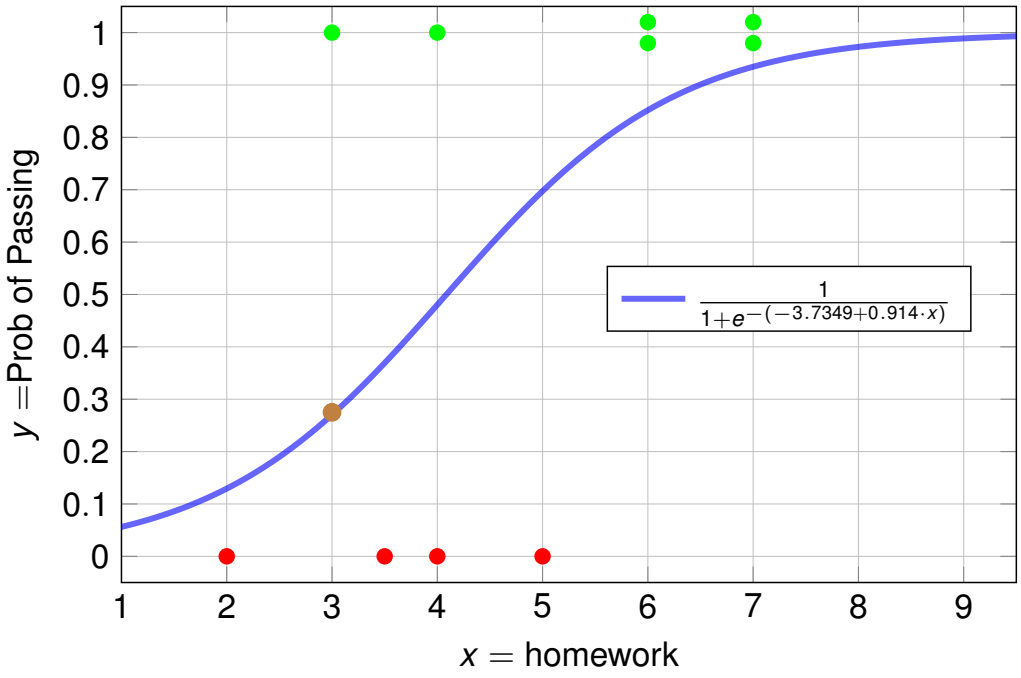
Binary Classification using Logistic Regression

Suppose there is a new student with homework $= 3$ and we have to predict a pass or fail.

Plot of the Optimal Model

Optimal Model

- $b = -3.7349$, $w = 0.914$, likelihood of 0.0071.



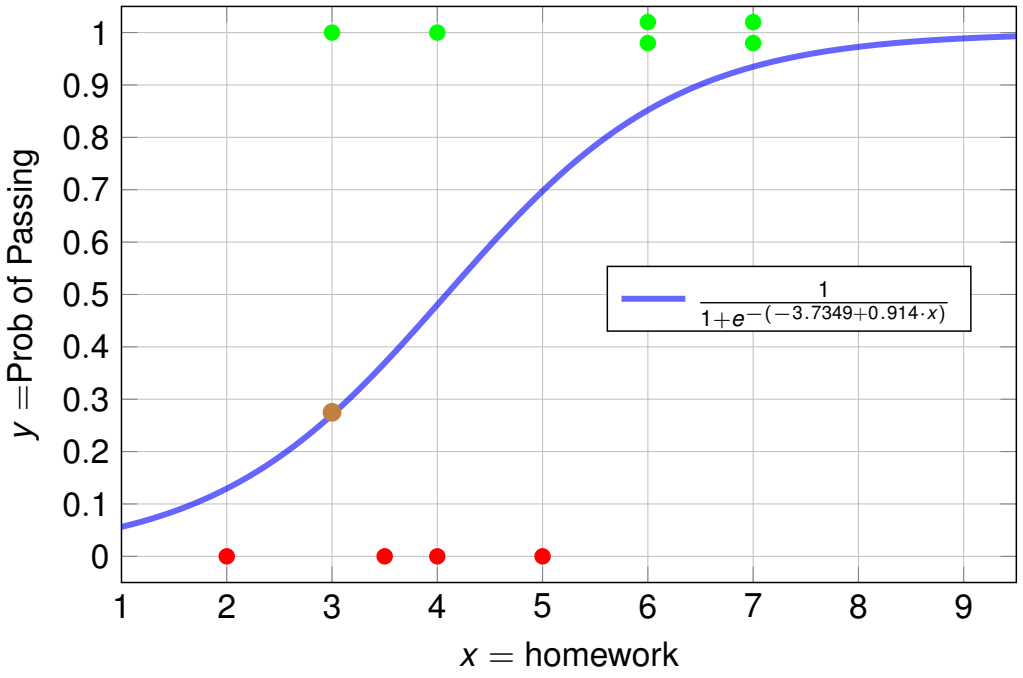
Binary Classification using Logistic Regression

Suppose there is a new student with homework $= 3$ and we have to predict a pass or fail.

Plot of the Optimal Model

Optimal Model

- $b = -3.7349$, $w = 0.914$, likelihood of 0.0071.



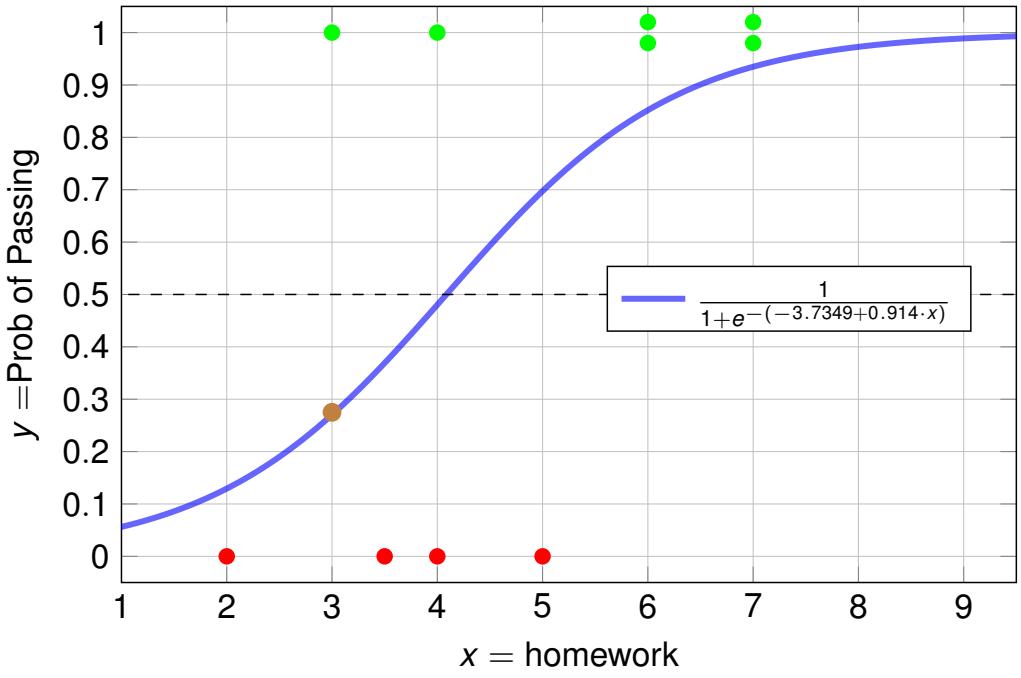
Binary Classification using Logistic Regression

Suppose there is a new student with homework $= 3$ and we have to predict a pass or fail. Model predicts a probability of $< .5$

Plot of the Optimal Model

Optimal Model

- $b = -3.7349$, $w = 0.914$, likelihood of 0.0071.



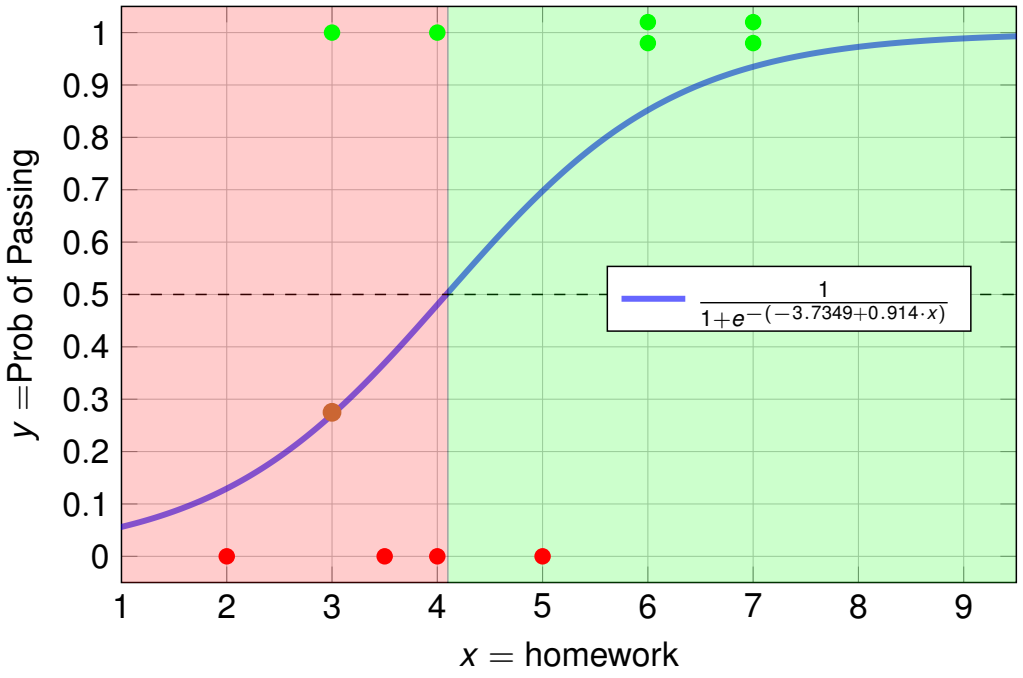
Binary Classification using Logistic Regression

Suppose there is a new student with homework $= 3$ and we have to predict a pass or fail. Model predicts a probability of $< .5$

Plot of the Optimal Model

Optimal Model

- $b = -3.7349$, $w = 0.914$, likelihood of 0.0071.



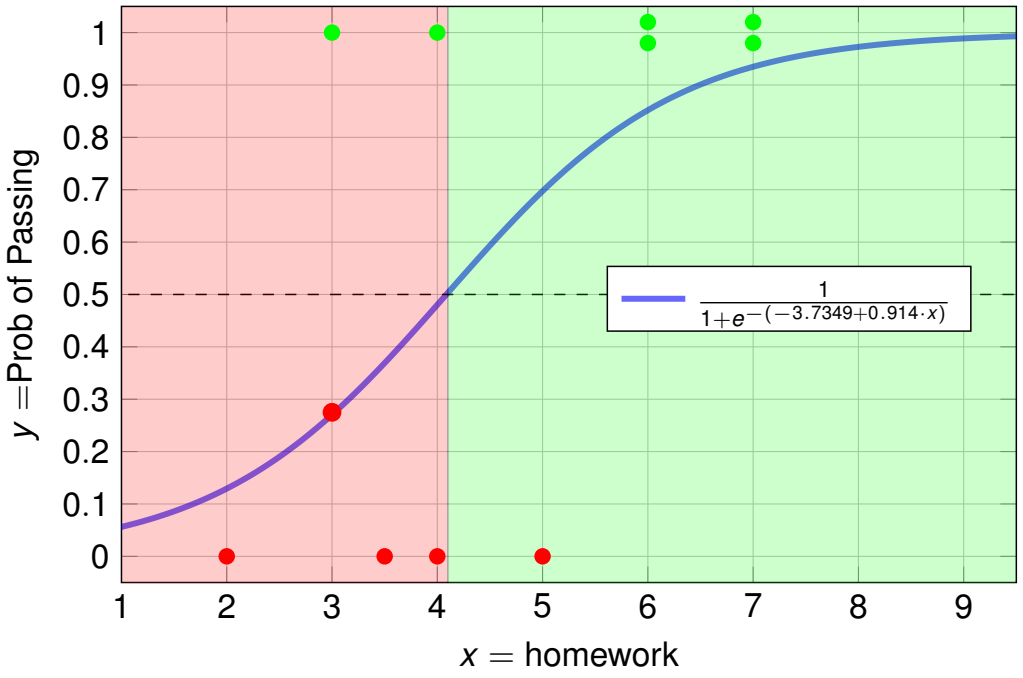
Binary Classification using Logistic Regression

Suppose there is a new student with homework $= 3$ and we have to predict a pass or fail. Model predicts a probability of $< .5$

Plot of the Optimal Model

Optimal Model

- $b = -3.7349$, $w = 0.914$, likelihood of 0.0071.



Binary Classification using Logistic Regression

Suppose there is a new student with homework $= 3$ and we have to predict a pass or fail. Model predicts a probability of $< .5 \Rightarrow$ predict fail.

Logistic Regression is a Linear Classifier(!)

- Predict $+1$ iff

$$
\begin{aligned}
& \frac{1}{1+e^{-b-w\cdot x}} \geq 0.5 \\
\Leftrightarrow \quad & 1+e^{-b-w\cdot x} \geq 2 \\
\Leftrightarrow \quad & e^{-b-w\cdot x} \geq 1 \\
\Leftrightarrow \quad & b + w\cdot x \geq 0.
\end{aligned}
$$

Summary: Linear Regression versus Logistic Regression

	Linear Regression	Logistic Regression
Features	numerical or categorical	numerical or categorical
Prediction	numerical	probability, binary (0 or 1)
Optimisation	Least-Squares	Maximum Likelihood
Best-Fit	line/hyperplane	Logistic Function (curve)
Applications	sales forecast	image processing
	grade prediction	spam classification
	...	...

Choice of Logistic Function

Question: Why did we choose the logistic function?

- We could also choose any other function interpolating between 0 and 1, and interpret these values as likelihoods
- The logistic regression function is natural and appears if we assume the Gaussian Naive Bayes model (which we cover in a later lecture)