CSCE 421: Machine Learning

Lecture 4: Linear Regression + Optimization

Texas A&M University
Bobak Mortazavi
Ryan King
Zhale Nowroozilarki

Goals of this lecture

- Review multiple linear regression
- Understand the need/purpose of optimization techniques for machine learning methods
- Calculating/Finding a global optimum
- Calculating/Finding local optima
- Understanding basic optima search

Optimal Coefficients: $\hat{w}_0, \hat{w}_1$

$$\hat{w}_0^* = \bar{y} - \hat{w}_1 \bar{x}$$

$$\hat{w}_1^* = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Important Questions to Ask

- Is there a relationship between budget and sales?
- If there is a relationship, how strong is it?
- Which of the three media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Advertising Example

- If P=30

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

With P=30 the t-statistic for the null hypothesis are around 2 and 2.75 respectively.

We conclude $w_0 \neq 0$ and $w_1 \neq 0$

Multiple Linear Regression

- Advertising budget is more than just the TV Element
- How do we account for each element?
- Is the answer 3 separate regressions?

Multiple Linear Regression

- Advertising budget is more than just the TV Element

$$y = w_0 + w_1 x_1 + \ldots + w_p x_p + \epsilon$$

- Where each w_j represents the average effect on y of a one unit change in x_j while holding all other parameters fixed. Therefore, we model

$$sales = w_0 + w_1 TV + w_1 Radio + w_1 Newspaper + \epsilon$$

Multiple Coefficients

$$\hat{y} = \widehat{w_0} + \widehat{w_1}x_1 + \ldots + \widehat{w_p}x_p$$

- We estimate with an ordinary least squares approach, such that

$$RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{N}\left(y_i - (\widehat{w_1}x_1 + \ldots + \widehat{w_p}x_p)\right)^2$$

- Similar to the single variable regression, take the partial derivatives and set = 0.
- Can be solved using matrix form, and plenty of linear solvers exist to find this solution.

SCRIBE NOTES – PLEASE DERIVE THIS MATRIX-FORM SOLUTION FOR OLS

Simple Single Regressions

Simple regression of **sales** on **radio**

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

Simple regression of **sales** on **newspaper**

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

- We see that if we model each as an individual linear regression, each budget item impacts sales
- This impact is statistically significant
- But will it remain so if we include all? (in other words what does the intercept represent here?)

Multiple Linear Regression

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

- Newspaper budget in single regression was acting as a surrogate for radio budget.
- When we add radio, newspaper no longer becomes significant.

Multiple Linear Regression

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

- Newspaper budget in single regression was acting as a surrogate for radio budget.
- When we add radio, newspaper no longer becomes significant.
- Must always test in this sense to avoid spurious correlations:
  - Example, shark attacks and ice cream sales are related at beaches

In Multiple Regression – Important Questions to Ask

- Is at least one predictor useful in generating a response variable?
- Do all predictors help explain a response or only some?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict and how accurate is this prediction? (no longer t-statistic but F-statistic).

In Multiple Regression – Important Questions to Ask

- Is at least one predictor useful in generating a response variable?
- Do all predictors help explain a response or only some?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict and how accurate is this prediction? (no longer t-statistic but F-statistic).

Variable importance

- The F-statistic and associated p-values tell us at least one feature is related to response
- How can we decide which one?
- We can iterate feature (variable) selection and see how the significance changes

Variable importance

- The F-statistic and associated p-values tell us at least one feature is related to response
- How can we decide which one?
- We can iterate feature (variable) selection and see how the significance changes
- In an ideal world, you would create all sub models with all combinations of variables included/excluded and see which one is the best.
  - Can use terms such as AIC, BIC or Adjusted R^2
- If you have only 2 features, how many models would this generate?
- What if you had 3 features?

Forward (Greedy) Feature Selection

- Start with the null model
- Fit p linear regressions of 1 variable (from our p dimensions of features)
- Calculate the RSS
- Select the variable with the lowest RSS to include in the model.
- Repeat.
- Stop when some criteria is met.

Backward Elimination

- Start with the full model
- Calculate the p-values on coefficient estimates
- Remove the feature with the largest p-value
- Re-calculate
- Stop when some criteria is met (for example all remaining p-values are less than some threshold)
- Cannot do this if we have more features than subjects (why not? More unknowns than equations in our linear system of equations)

Forward Backward Mixed Selection

- Start with the no variables selected
- Add features in forward stepwise fashion
- At each stage, once adding variables, re-check p-values
- If any p-value becomes too large, remove it
- Stop when some criteria is met

In Multiple Regression – Important Questions to Ask

- Is at least one predictor useful in generating a response variable?
- Do all predictors help explain a response or only some?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict and how accurate is this prediction? (no longer t-statistic but F-statistic).

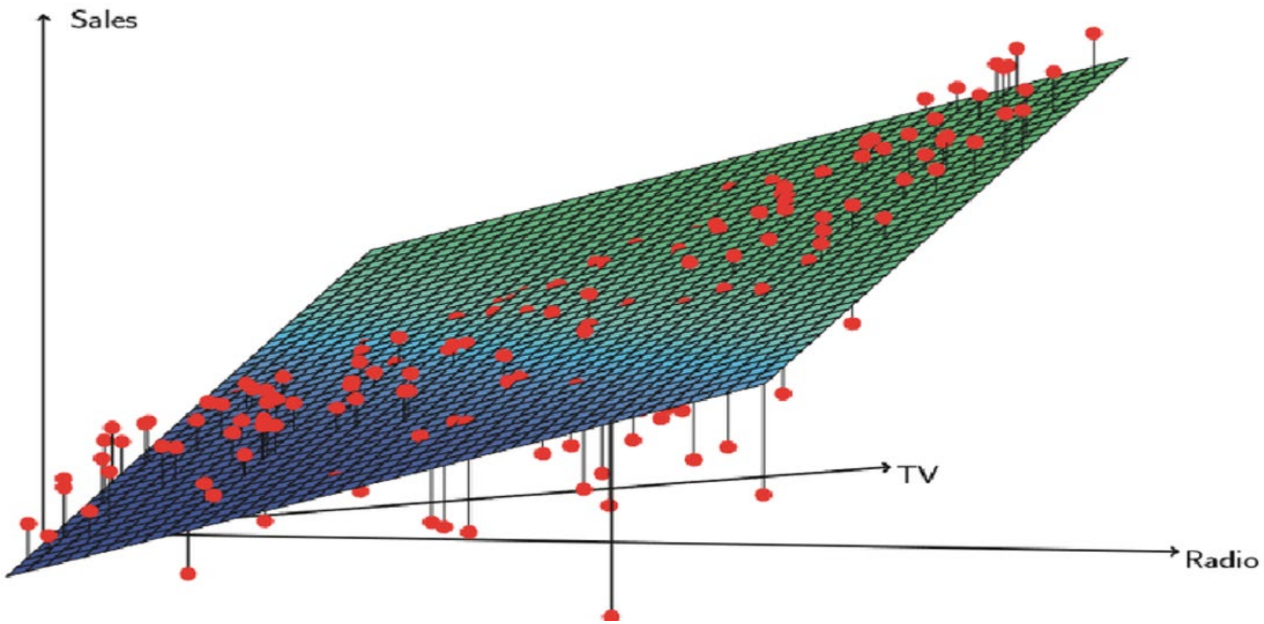
Model Fit

- Once the model with features selected is implemented, how can we measure model goodness of fit?
- RSE and R^2 are common measures where R^2 is now the $Cor(Y, \hat{Y})^2$
- However, more variables will increase R^2 (because of how we fit with least squares).
- RSE does NOT get better just by adding more features
- Need to consider what measures we use and what tests we have to consider those values significant.

In Multiple Regression – Important Questions to Ask

- Is at least one predictor useful in generating a response variable?
- Do all predictors help explain a response or only some?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict and how accurate is this prediction? (no longer t-statistic but F-statistic).

Residuals



- Plotting the residuals against the fit model can tell us about feature trends
- Here positive residuals appear to fall along the line balancing TV and Radio
- Here negative residuals appear to fall outside of this range
- This indicates that the combined interaction of TV and Radio could be an important feature
- But before we add that, how does linear regression work if not all features are continuous values?

New Example: Credit Balance

```
> summary(Credit)
       ID            Income           Limit           Rating          Cards            Age          Education        Gender      Student
 Min.   :  1.0   Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000   Min.   :23.00   Min.   : 5.00    Male :193   No :360
 1st Qu.:100.8   1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000   1st Qu.:41.75   1st Qu.:11.00   Female:207   Yes: 40
 Median :200.5   Median : 33.12   Median : 4622   Median :344.0   Median :3.000   Median :56.00   Median :14.00
 Mean   :200.5   Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958   Mean   :55.67   Mean   :13.45
 3rd Qu.:300.2   3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000   3rd Qu.:70.00   3rd Qu.:16.00
 Max.   :400.0   Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000   Max.   :98.00   Max.   :20.00
 Married                 Ethnicity       Balance
 No :155   African American: 99    Min.   :   0.00
 Yes:245   Asian           :102    1st Qu.:  68.75
           Caucasian       :199    Median : 459.50
                                   Mean   : 520.01
                                   3rd Qu.: 863.00
                                   Max.   :1999.00
```

- In this example, we have both continuous quantitative features and qualitative features
- What if we want to investigate the difference in balances across these features?
- For example, if there is a relationship between gender and balance described in this dataset?

Categorical Variables: Factor Levels

```
> summary(Credit)
       ID           Income           Limit           Rating          Cards            Age          Education        Gender     Student
 Min.   :  1.0   Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000   Min.   :23.00   Min.   : 5.00    Male :193   No :360
 1st Qu.:100.8   1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000   1st Qu.:41.75   1st Qu.:11.00   Female:207   Yes: 40
 Median :200.5   Median : 33.12   Median : 4622   Median :344.0   Median :3.000   Median :56.00   Median :14.00
 Mean   :200.5   Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958   Mean   :55.67   Mean   :13.45
 3rd Qu.:300.2   3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000   3rd Qu.:70.00   3rd Qu.:16.00
 Max.   :400.0   Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000   Max.   :98.00   Max.   :20.00
 Married              Ethnicity       Balance
 No :155   African American: 99   Min.   :   0.00
 Yes:245   Asian           :102   1st Qu.:  68.75
           Caucasian       :199   Median : 459.50
                                  Mean   : 520.01
                                  3rd Qu.: 863.00
                                  Max.   :1999.00
```

- A categorical variable that has multiple levels is said to have multiple factor levels
- Factor of two levels is an indicator or dummy variable (binary yes/no)

Credit Balance: Factor/Indicator

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	< 0.0001
gender[Female]	19.73	46.05	0.429	0.6690

- P-value for our indicator is very high – what does this mean?
- 0/1 coding here is arbitrary – has no effect on regression fit
- Factor of two levels is an indicator or dummy variable (binary yes/no)
- We can create multiple binary/indicator variables from a categorical variable such that

$$x_p = \begin{cases} 1, if\ pth\ person\ is\ female \\ 0\ otherwise \end{cases}$$

So

$$y = w_0 + w_p x_p = \begin{cases} w_0 + w_p, if\ pth\ person\ is\ female \\ w_0\ otherwise \end{cases}$$

Can then run statistical test on the goodness of fit of this variable/model

Categorical Variables: Factor Levels

```
> summary(Credit)
       ID            Income            Limit           Rating          Cards            Age          Education         Gender     Student
 Min.   :  1.0   Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000   Min.   :23.00   Min.   : 5.00    Male :193   No :360
 1st Qu.:100.8   1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000   1st Qu.:41.75   1st Qu.:11.00   Female:207   Yes: 40
 Median :200.5   Median : 33.12   Median : 4622   Median :344.0   Median :3.000   Median :56.00   Median :14.00
 Mean   :200.5   Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958   Mean   :55.67   Mean   :13.45
 3rd Qu.:300.2   3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000   3rd Qu.:70.00   3rd Qu.:16.00
 Max.   :400.0   Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000   Max.   :98.00   Max.   :20.00
 Married              Ethnicity       Balance
 No :155   African American: 99   Min.   :   0.00
 Yes:245   Asian           :102   1st Qu.:  68.75
           Caucasian       :199   Median : 459.50
                                  Mean   : 520.01
                                  3rd Qu.: 863.00
                                  Max.   :1999.00
```

- A categorical variable that has multiple levels is said to have multiple factor levels
- Factor of two levels is an indicator or dummy variable (binary yes/no)
- We can say here that gender has no impact on balance in this example
- What if we model with another variable such as Ethnicity?

Categorical Variables: Factor Levels

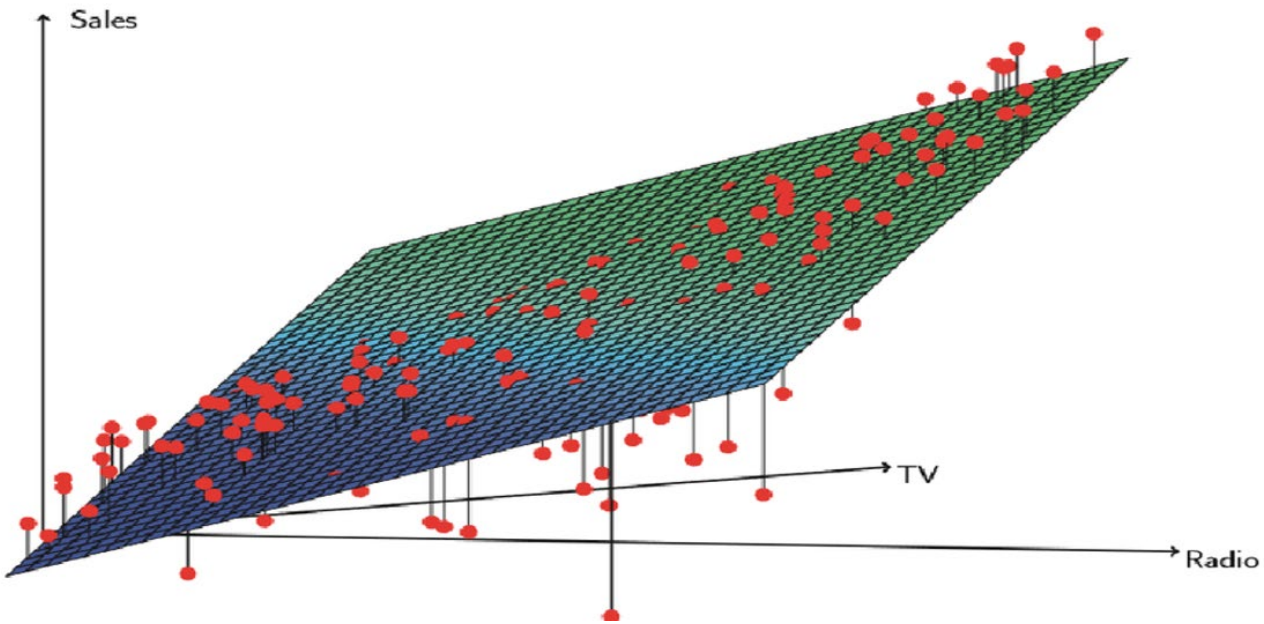
```
Residuals:
    Min      1Q  Median      3Q     Max
-531.00 -457.08  -63.25  339.25 1480.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   531.00      46.32  11.464   <2e-16 ***
asian         -18.69      65.02  -0.287    0.774
caucasian     -12.50      56.68  -0.221    0.826
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.0002188, Adjusted R-squared:  -0.004818
F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

- A categorical variable that has multiple levels is said to have multiple factor levels
- Factor of two levels is an indicator or dummy variable (binary yes/no)
- We can say here that gender has no impact on balance in this example
- Does not seem to be significant
- Why did we only create 2 of the 3 indicator variables? What would have happened if we created the third?

Important Questions to Ask

- Is there a relationship between budget and sales?
- If there is a relationship, how strong is it?
- Which of the three media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Residuals



- Plotting the residuals against the fit model can tell us about feature trends
- Here positive residuals appear to fall along the line balancing TV and Radio
- Here negative residuals appear to fall outside of this range
- This indicates that the combined interaction of TV and Radio could be an important feature

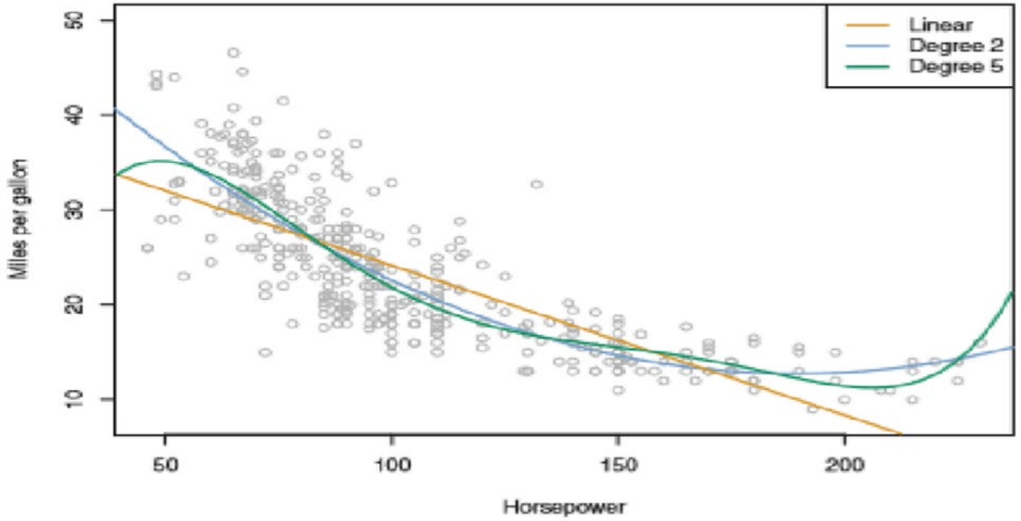
Extending Additive and Linear Assumptions on x and y

- TV and Radio are both associated with sales
- A 1 unit increase in TV budget, independent of radio budget, increases overall sales
- But what about that relationship between the two? Can Radio budget help improve TV budget (synergy in marketing, interaction term in machine learning)
- Can create a new variable that represents a joint change in both TV and Radio budget, together.

What about non-linear relationships?

- TV and Radio are both associated with sales
- A 1 unit increase in TV budget, independent of radio budget, increases overall sales
- But what about that relationship between the two? Can Radio budget help improve TV budget (synergy in marketing, interaction term in machine learning)
- Can create a new variable that represents a joint change in both TV and Radio budget, together.

Non-linear Regressions: MPG and Horsepower

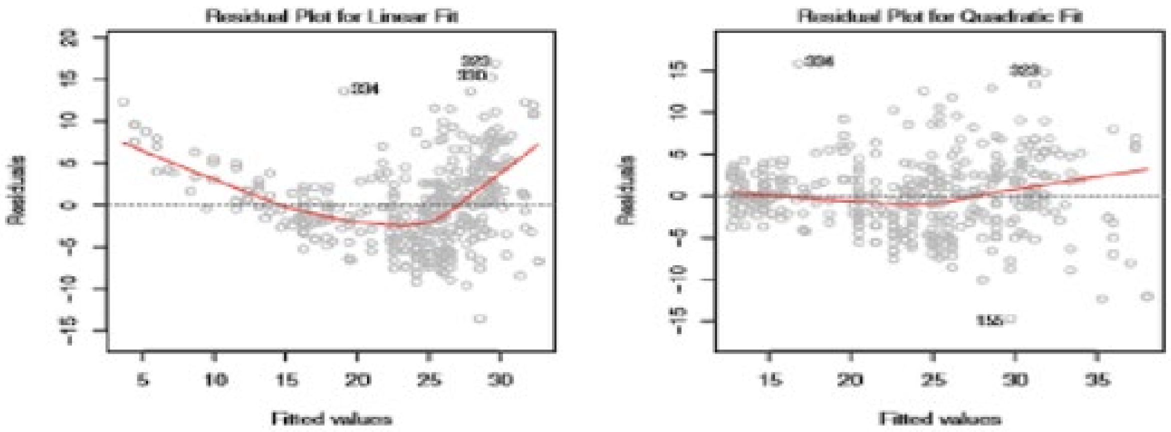


- New example – model the MPG of an engine as a function of horsepower
- Best model is of the form

$$mpg = w_0 + w_1 HP + w_2 HP^2$$

- This is still a linear model! Can be solved with normal linear model solvers
- Interaction term provides higher polynomial degree
- But how do we tell if this is the right degree and how to stop increasing degrees of polynomial?

Non-linear Regressions: MPG and Horsepower

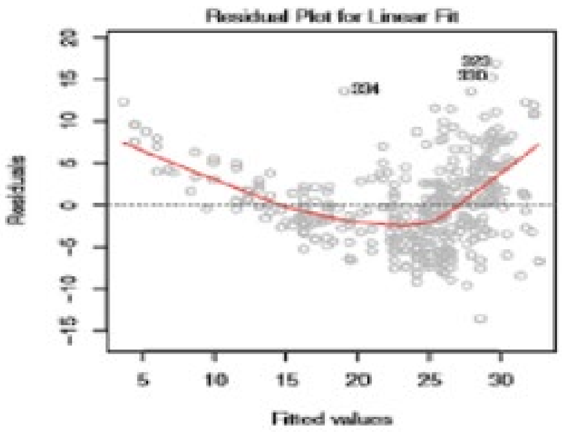


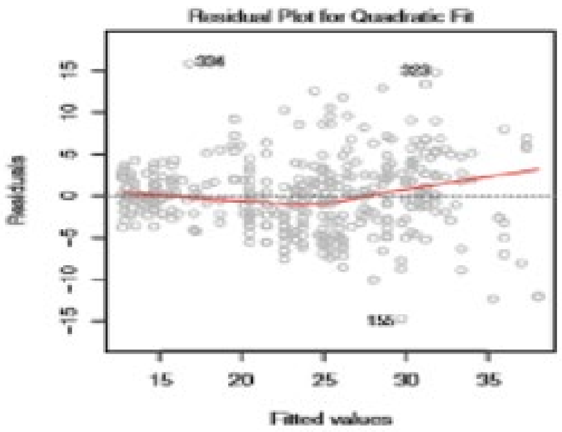
- New example – model the MPG of an engine as a function of horsepower
- Best model is of the form

$$mpg = w_0 + w_1 HP + w_2 HP^2$$

- This is still a linear model! Can be solved with normal linear model solvers
- Interaction term provides higher polynomial degree
- But how do we tell if this is the right degree and how to stop increasing degrees of polynomial?
- We can plot the residual errors and see if there is a relationship
- Patterns in the residual usually indicate a higher-order interaction term

Non-linear Regressions: Ignoring outlier terms





- New example – model the MPG of an engine as a function of horsepower
- Best model is of the form

$$mpg = w_0 + w_1 HP + w_2 HP^2$$

- This is still a linear model! Can be solved with normal linear model solvers
- Interaction term provides higher polynomial degree
- But how do we tell if this is the right degree and how to stop increasing degrees of polynomial?
- We can plot the residual errors and see if there is a relationship
- Patterns in the residual usually indicate a higher-order interaction term
- We can also scale these residuals and ignore those > 3 standard deviations of error as outlier terms

Important Questions to Ask

- Is there a relationship between budget and sales?
- If there is a relationship, how strong is it?
- Which of the three media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Why does all this work? Convexity of loss function!

- A Set S is convex if for any $\theta, \theta' \in S$ there exists
- $\lambda\theta + (1-\lambda)\theta' \in S$ for $\lambda \in [0,1]$
- In practice this means draw a line between any two points in a set and if it is convex, every point on the line still lies within the set

Why does all this work? Convexity of loss function!

- A Set S is convex if for any $\theta, \theta' \in S$ there exists
- $\lambda\theta + (1 - \lambda)\theta' \in S$ for $\lambda \in [0,1]$
- In practice this means draw a line between any two points in a set and if it is convex, every point on the line still lies within the set
- Now, a function f(θ) is convex if its set of points defines a convex set
- In other words

$$f(\lambda(\theta + (1 - \lambda)\theta') \leq \ \lambda f(\theta + (1 - \lambda) f(\theta') \ \ \lambda \in [0,1]$$

Why does all this work? Convexity of loss function!

- A Set S is convex if for any $\theta, \theta' \in S$ there exists
- $\lambda\theta + (1-\lambda)\theta' \in S$ for $\lambda \in [0,1]$
- In practice this means draw a line between any two points in a set and if it is convex, every point on the line still lies within the set
- Now, a function f(θ) is convex if its set of points defines a convex set
- In other words

$f(\lambda(\theta + (1-\lambda)\theta') \leq \ \lambda f(\theta + (1-\lambda) f(\theta') \ \ \lambda \in [0,1]$

- If this inequality is strict, we call this strictly convex
- If f is instead concave then –f is convex
- We can evaluate a function as being convex if it passes the 2nd derivative test

$$\frac{\partial^2}{\partial \theta^2} f(\theta) > 0$$

Why does all this work? Convexity of loss function!

- A Set S is convex if for any $\theta, \theta' \in S$ there exists
- $\lambda\theta + (1-\lambda)\theta' \in S$ for $\lambda \in [0,1]$
- In practice this means draw a line between any two points in a set and if it is convex, every point on the line still lies within the set
- Now, a function f(θ) is convex if its set of points defines a convex set
- In other words

$f(\lambda(\theta + (1-\lambda)\theta') \leq \ \lambda f(\theta + (1-\lambda) f(\theta') \ \ \lambda \in [0,1]$

- If this inequality is strict, we call this strictly convex
- If f is instead concave then –f is convex
- We can evaluate a function as being convex if it passes the 2nd derivative test

$$\frac{\partial^2}{\partial\theta^2} f(\theta) > 0$$

SCRIBE NOTES – SHOW LINEAR REGRESSION RSS LOSS FUNCTION IS CONVEX INCLUDING HIGHER ORDER POLYNOMIALS

Goals of this lecture

- Review multiple linear regression
- Understand the need/purpose of optimization techniques for machine learning methods
- Calculating/Finding a global optimum
- Calculating/Finding local optima
- Understanding basic optima search

The Zero-Order Optimality Condition

Find the smallest point(s) of a function.

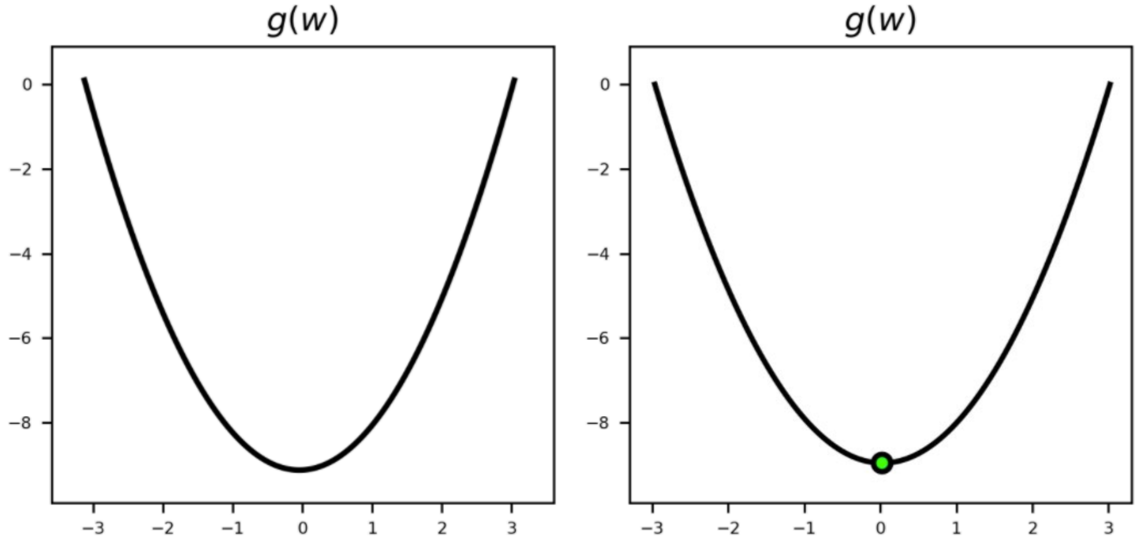
$$\underset{w}{\text{minimize}}\, g(\boldsymbol{w})$$

- Approach:
  - Identify the minimum visually by plotting it over a large swath of its input space.

The Zero-Order Optimality Condition

- Example 1: Global minimum of a quadratic

$$g(w) = w^2$$



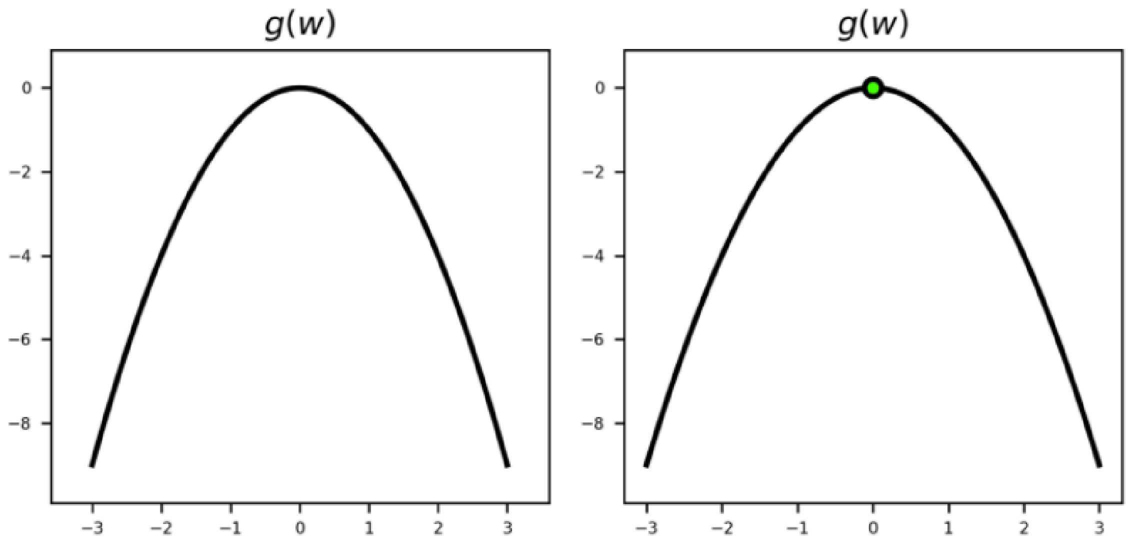
- **Global** minimum point w^*

$$g(w^*) \leq g(w) \text{ for all } w$$

The Zero-Order Optimality Condition

- Example 2: Global maximum of a quadratic

$$g(w) = -w^2$$



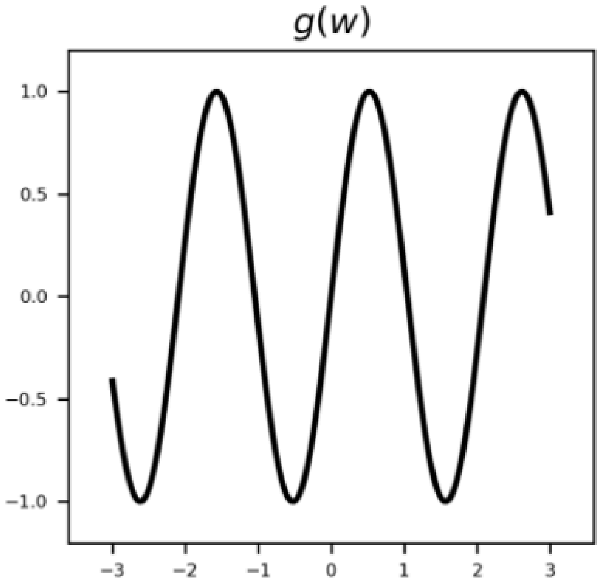
- Maximum point w^*: $g(w^*) \geq g(w)$ for all w

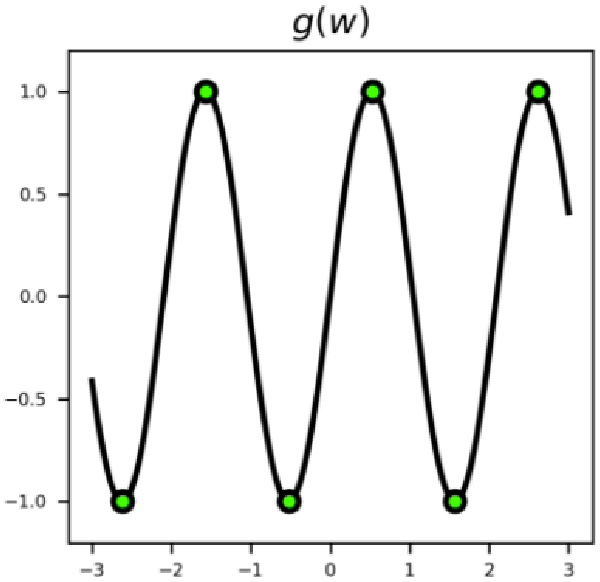
$$\underset{w}{\text{maximum}}\, g(\boldsymbol{w}) = -\underset{w}{\text{minimize}}\, g(\boldsymbol{w})$$

The Zero-Order Optimality Condition

- Example 3: global maximum/minimum of a sinusoid

$$g(w) = \sin(2w)$$

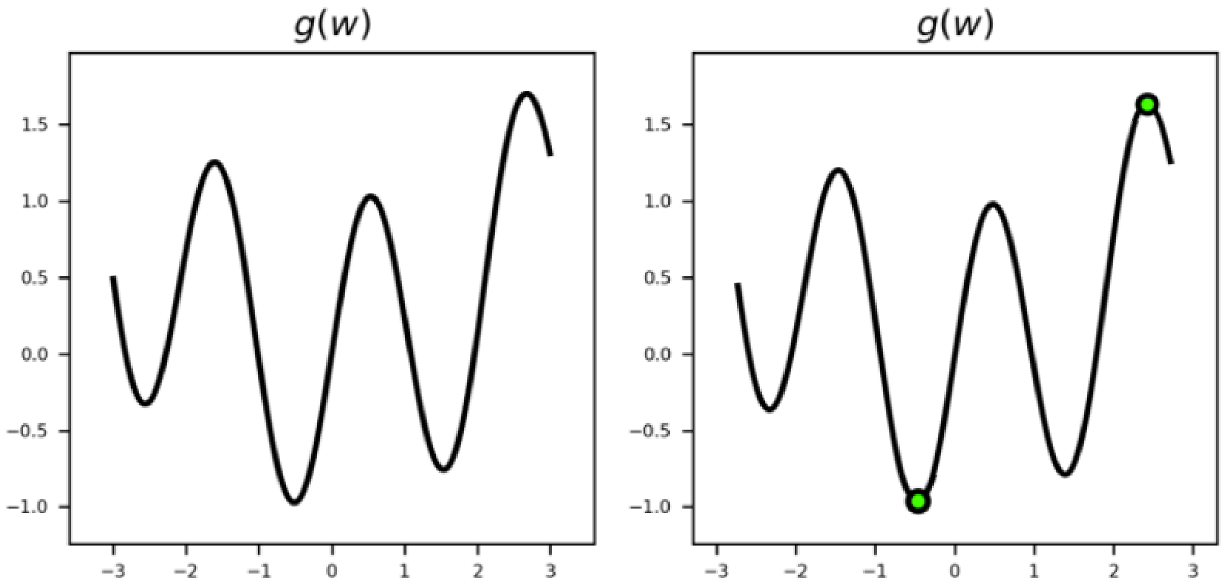




The Zero-Order Optimality Condition

- Example 3: **global** maximum/minimum of the sum of a sinusoid and a quadratic

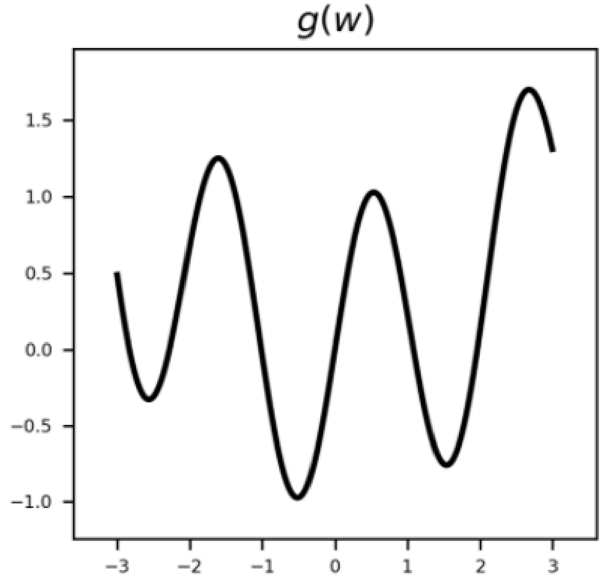
$$g(w) = \sin(3w) + 0.1w^2$$

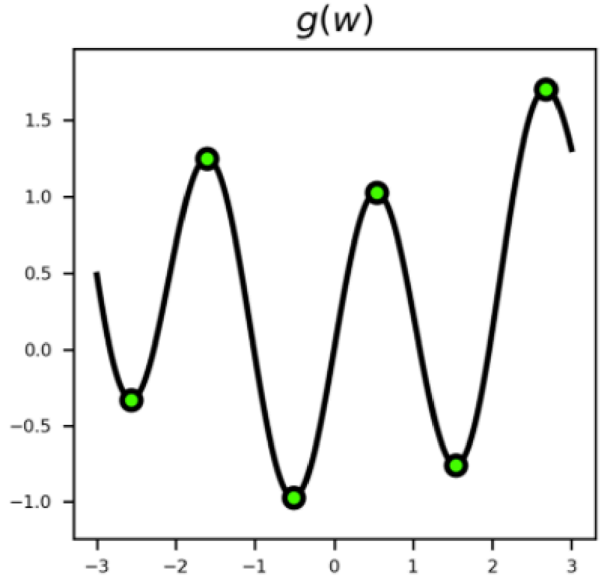


The Zero-Order Optimality Condition

- Example 4: **local** maximum/minimum of the sum of a sinusoid and a quadratic

$$g(w) = \sin(3w) + 0.1w^2$$





- **Local** minimum point w^*

$$g(w^*) \leq g(w) \text{ for all } w \text{ near } w^*$$

The zero order condition for optimality

- The zero order condition for optimality: A point w^* is:
  - a global minimum of $g(w)$ if and only if $g(w^*) \leq g(w)$ for all w.
  - a global maximum of $g(w)$ if and only if $g(w^*) \geq g(w)$ for all w.
  - a local minimum of $g(w)$ if and only if $g(w^*) \leq g(w)$ for all $w \text{ near } w^*$.
  - a local maximum of $g(w)$ if and only if $g(w^*) \geq g(w)$ for all $w \text{ near } w^*$.

Global Optimization Methods

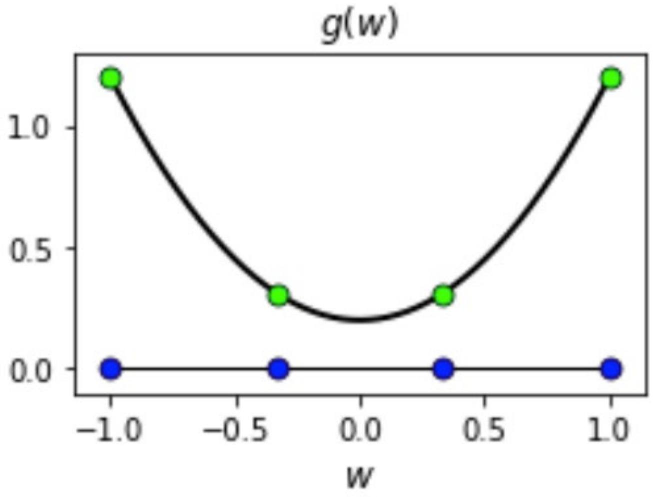
Method: Choosing input points

- Visual approach: evaluate a function over a large number of its input points and designating the input that provides the smallest result as the approximate global minimum.
- Input choosing:
  - Uniformly sample over an evenly spaced grid.
  - Randomly pick the same number of input points.

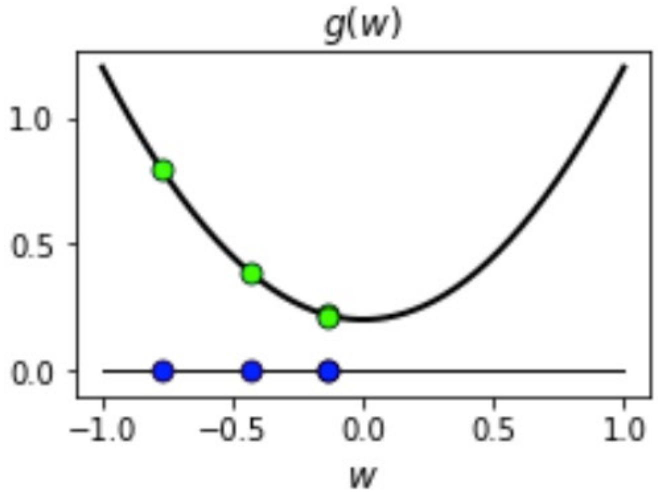
Global Optimization Methods

Example 1: 2-d quadratic

$$g(w) = w^2 + 0.2$$



Evenly sampling

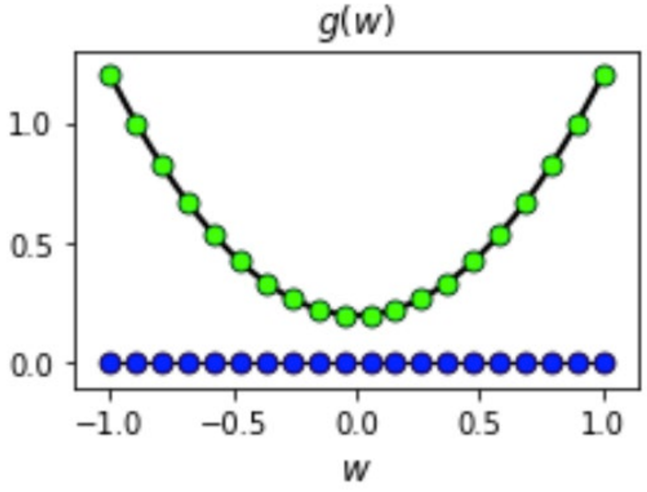


Randomly sampling

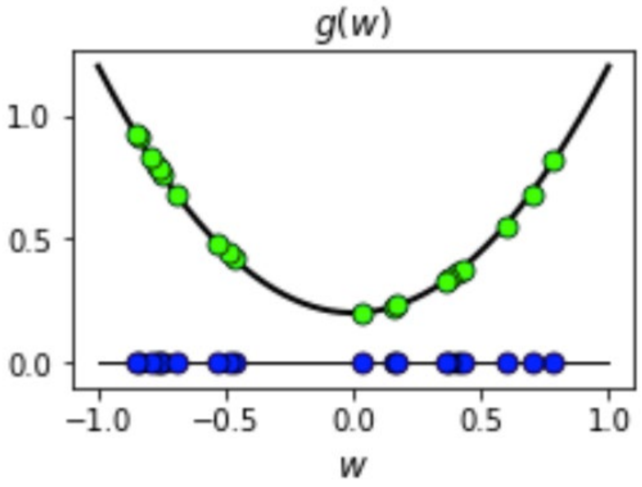
- Blue: sampled inputs
- Green: corresponding evaluation on the function.

Global Optimization Methods

- More samples



Evenly sampling



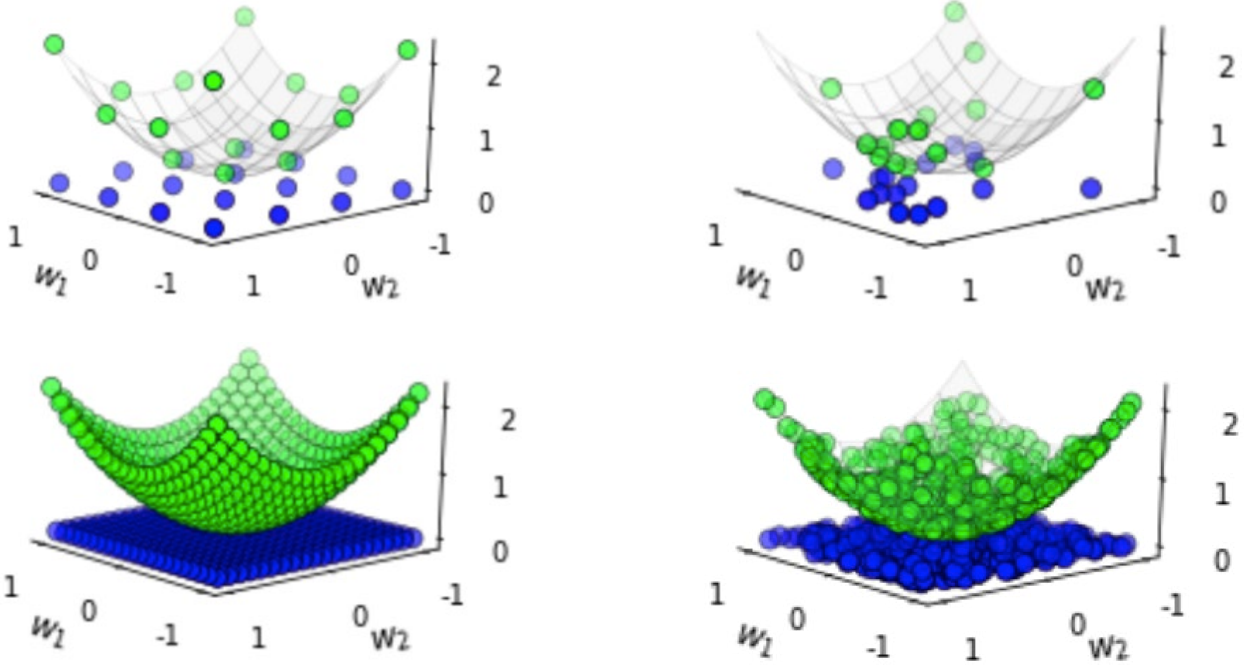
Randomly sampling

- When given enough samples, the minimized point can be close to global minimum.
- Either approach is able to find global minimum.

Global Optimization Methods

Example 2: 3-d quadratic

$$g(w_1, w_2) = w_1^2 + w_2^2 + 0.2$$



Evenly sampling

Randomly sampling

Local Optimization Methods

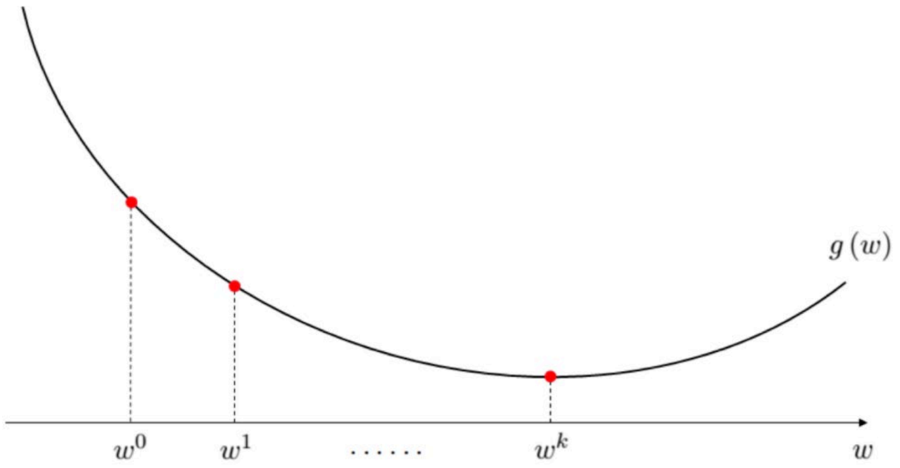
- Method
  - Initialize a starting point $\boldsymbol{w}^0$
  - The point is 'pulled' downhill to a new point $\boldsymbol{w}^1$ lower on the function.

$$g(\boldsymbol{w}^0) > \mathrm{g}(\boldsymbol{w}^1)$$

  - Sequentially pulling the point 'downhill' towards points that are lower and lower on the function.

$$g(\boldsymbol{w}^0) > \mathrm{g}(\boldsymbol{w}^1) > \mathrm{g}(\boldsymbol{w}^2) > \cdots > \mathrm{g}(\boldsymbol{w}^K)$$

  - Eventually reach a minimizer after K points are yield.



Local Optimization Methods

- Global optimization: a multitude of simultaneously sampled input points to determine an approximate minimum of a given function $g(w)$.
- Local optimization: sequentially refining a single sample input called an initial point until it reaches an approximate minimum.

Local Optimization Methods

Framework

- $\mathbf{w}^0$: initial point.
- $\mathbf{w}^1$: the first updated point
- $\mathbf{d}^0$: direction vector from $\mathbf{w}^0$ to $\mathbf{w}^1$

$$\mathbf{w}^1 = \mathbf{w}^0 + \mathbf{d}^0$$

- Similarly
- $\mathbf{w}^2$: the second updated point
- $\mathbf{d}^1$: direction vector from $\mathbf{w}^1$ to $\mathbf{w}^2$

$$\mathbf{w}^2 = \mathbf{w}^1 + \mathbf{d}^1$$

Local Optimization Methods

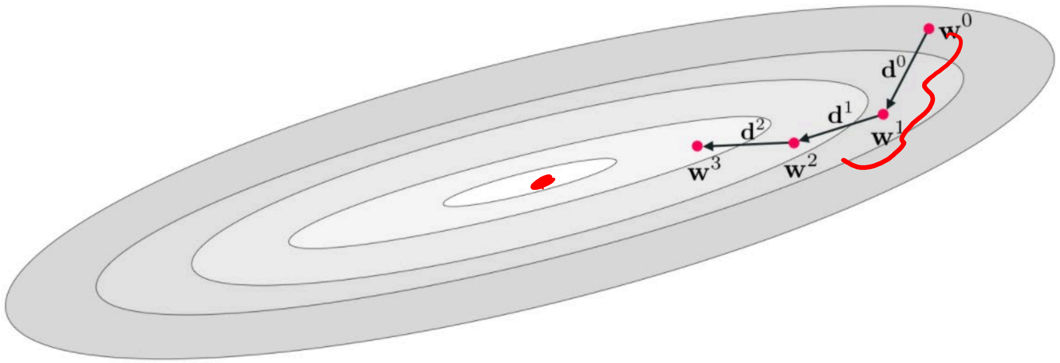
$$\mathbf{w}^0$$
$$\mathbf{w}^1 = \mathbf{w}^0 + \mathbf{d}^0$$
$$\mathbf{w}^2 = \mathbf{w}^1 + \mathbf{d}^1$$
$$\mathbf{w}^3 = \mathbf{w}^2 + \mathbf{d}^2$$
$$\vdots \quad \vdots \quad \vdots$$
$$\mathbf{w}^K = \mathbf{w}^{K-1} + \mathbf{d}^{K-1}$$

$\mathbf{d}^{k-1}$ is the descent direction defined at the k^{th} step of process

$$\mathbf{w}^k = \mathbf{w}^{k-1} + \mathbf{d}^{k-1}$$

and

$$g(\mathbf{w}^0) > g(\mathbf{w}^1) > g(\mathbf{w}^2) > \cdots > g(\mathbf{w}^K)$$



Schematic illustration of a generic local optimization scheme.

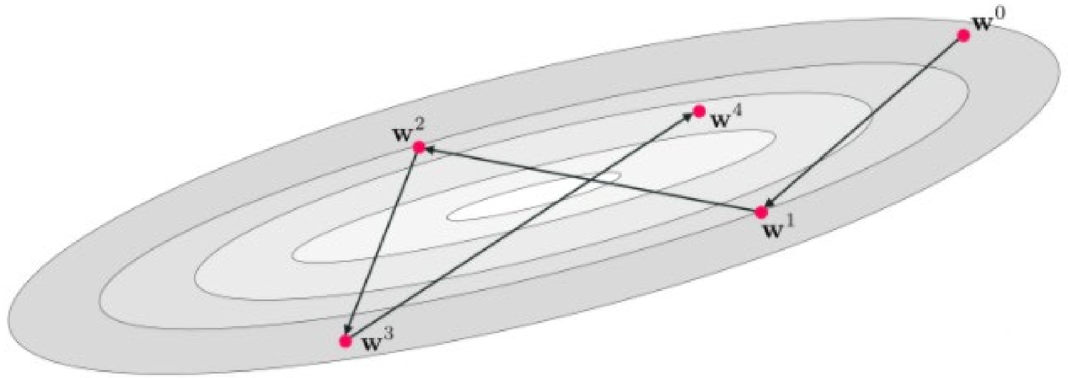
Local Optimization Methods

The steplength parameter

- Distance of updating at k^{th} step:

$$\left\|\mathbf{w}^k - \mathbf{w}^{k-1}\right\|_2 = \left\| \left(\mathbf{w}^{k-1} + \mathbf{d}^{k-1}\right) - \mathbf{w}^{k-1}\right\|_2 = \left\|\mathbf{d}^{k-1}\right\|_2.$$

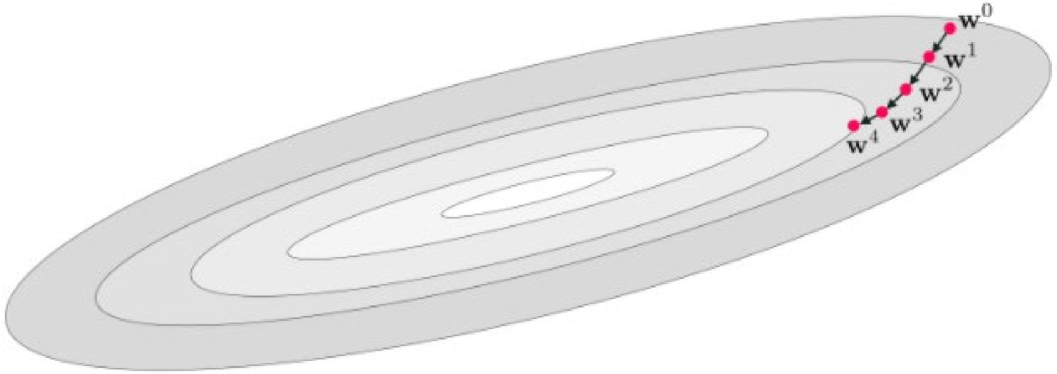
- Correct direction wrong length
  - Large direction vectors: can never reach approximate minimum.



Direction vectors are too large causing a wild oscillatory behavior around the minimum.

Local Optimization Methods

- Correct direction wrong length
  - Short updating distance: move too slow and too many steps are required.



Direction vectors are too small, requiring a large number of steps be taken to reach the minimum.

Local Optimization Methods

- Steplength parameter/Learning rate parameter:

$$\mathbf{w}^k = \mathbf{w}^{k-1} + \alpha \mathbf{d}^{k-1}$$

- The entire sequence of K steps:

$$\begin{aligned}
&\mathbf{w}^0 \\
&\mathbf{w}^1 = \mathbf{w}^0 + \alpha \mathbf{d}^0 \\
&\mathbf{w}^2 = \mathbf{w}^1 + \alpha \mathbf{d}^1 \\
&\mathbf{w}^3 = \mathbf{w}^2 + \alpha \mathbf{d}^2 \\
&\vdots \quad \vdots \quad \vdots \\
&\mathbf{w}^K = \mathbf{w}^{K-1} + \alpha \mathbf{d}^{K-1}
\end{aligned}$$

- Distance vector:

$$\left\| \mathbf{w}^k - \mathbf{w}^{k-1} \right\|_2 = \left\| \left(\mathbf{w}^{k-1} + \alpha \mathbf{d}^{k-1} \right) - \mathbf{w}^{k-1} \right\|_2 = \alpha \left\| \mathbf{d}^{k-1} \right\|_2$$

Takeaways

- ML algorithms use optimization to either maximize performance or minimize loss (minimize errors)
- We understand the mathematics behind optimum values (differentiation!)
- Understanding the basics of Search for finding optimal values when closed form derivatives cannot be used.
- **Next Time: Gradient Descent**