Introduction to Sports Rating Systems

Evgeni Ovcharov

July 23, 2020

Figure: 1964-66 Candidates Championship in Riga, Latvia

Outline of talk

What is a sports ratings system?

- A sports rating system is a method for calculating the relative strengths of the competitors in a sports championships.
- Elo is historically the first rating system based on firm statistical foundations.

Figure: Arpad Elo, the inventor of the Elo rating system.

Rating systems vs Point accumulation systems

- Point accumulation systems predate sports rating systems.
- Point accumulation systems provide objective ranking only under two stringent conditions:
  - Every participant in the championship plays against every possible rival.
  - All participants play an equal number of times.

  Championships satisfying these conditions are called *sports leagues*.
- The Bulgarian professional football championship is organised as a sports league. Its A group hosts 14 teams which play against each rival two times, or $14 \cdot 13 = 282$ times. The first phase of the championship takes 26 weeks in which 7 matches are played, followed by playoffs.
- Sports leagues are infeasible for championships with many participants, or if there are geographical or time restrictions.
- Point accumulation systems have been extended to championships in tournament format, but there they are inferior to rating systems.

Figure: World Cup Finals, USA 1994

Some anomalies of FIFA's old point accumulation systems

- In 2014 Brazil sinks to the record low position 22 in FIFA's rank list before hosting the World Cup Finals, where it ends up at fourth place.
- In 2018 Russia is the host of the World Cup Finals. It held the 70-th position in FIFA's rank list before the start of the competition but it reached quarter-finals, where it was eliminated by a penalty shoot-out.
- Booth teams failed to accumulate points by playing qualification matches due to automatic qualification for the hosts, pointing to a weakness of FIFA's ranking system at the time.
- Other problem with FIFA's ranking system were the friendly matches which could lead to loosing points and were avoided by some teams for long periods of time.
- Due to many weaknesses of the above type, FIFA finally decided to ditch its point accumulation system in favour of a Elo-type rating system in 2018.

Summary of advantages of sports rating systems

- find objective indication of strength or skill of competitors based on previous performance
- actual information on the sport form of each contestant and its progress through time; increases competition
- grouping of teams by similar strength, organization of tournaments, pairing of teams within a tournament
- sports betting, prediction
- retrospective analysis and sports modelling - understand the key factors that influence the match outcome.

Notation

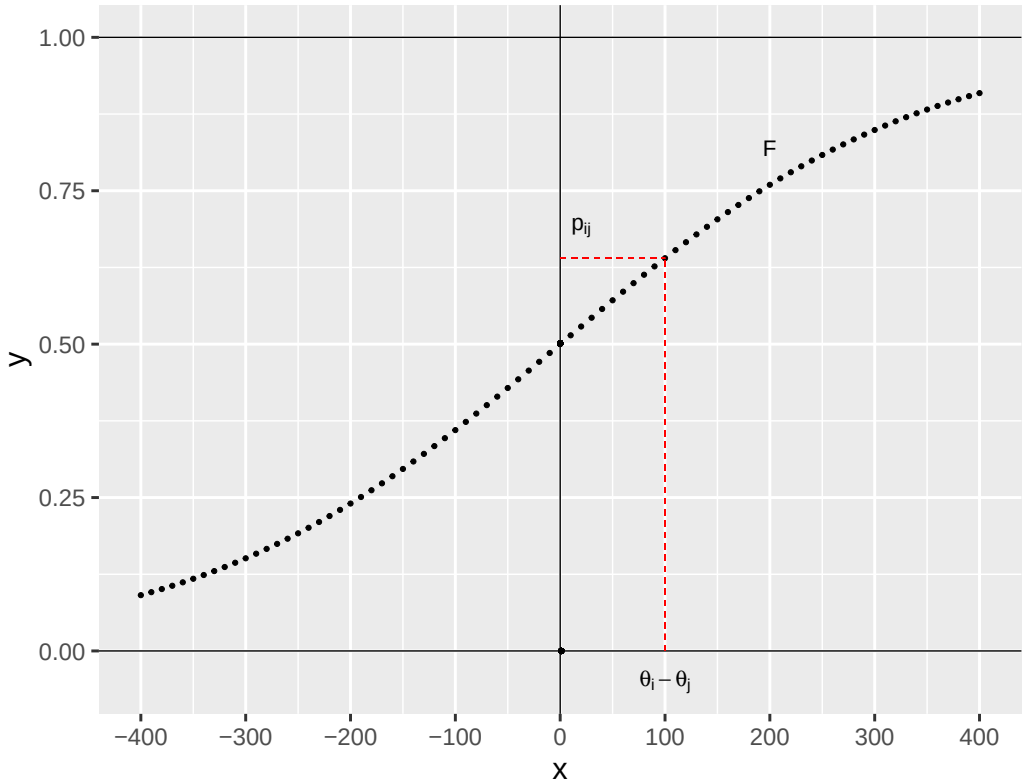
Let $\theta_1, \theta_2, \ldots, \theta_n$ be the ratings of n players. By Y_{ij} we denote the random variable

$$Y_{ij} = \begin{cases} 1 & \text{player } i \text{ wins against player } j, \\ 0 & \text{player } i \text{ looses to player } j. \end{cases}$$

By p_{ij} we denote the probability that player i defeats player j. We also have that $p_{ij} = E[Y_{ij}]$.

Modelling the winning probability

How to model p_{ij} in terms of θ_i and θ_j?



$F(-\infty) = 0$, $F(+\infty) = 1$, F is continuous and monotone increasing, therefore F is a cumulative distribution function.

Paired comparison models

Thus we obtain the class of *paired comparison models*:

$$p_{ij} = F(\theta_i - \theta_j),$$

where $p_{ij} = P(Y_{ij} = 1)$. The most common choices for F are the Gaussian CDF and the logistic CDF. For the logistic CDF e have

$$p_{ij} = \frac{1}{1 + e^{-(\theta_i - \theta_j)/\beta}},$$

where β is a scaling constant. For example, FIDE uses $\beta = 400/\ln 10$.

Elo probability curve

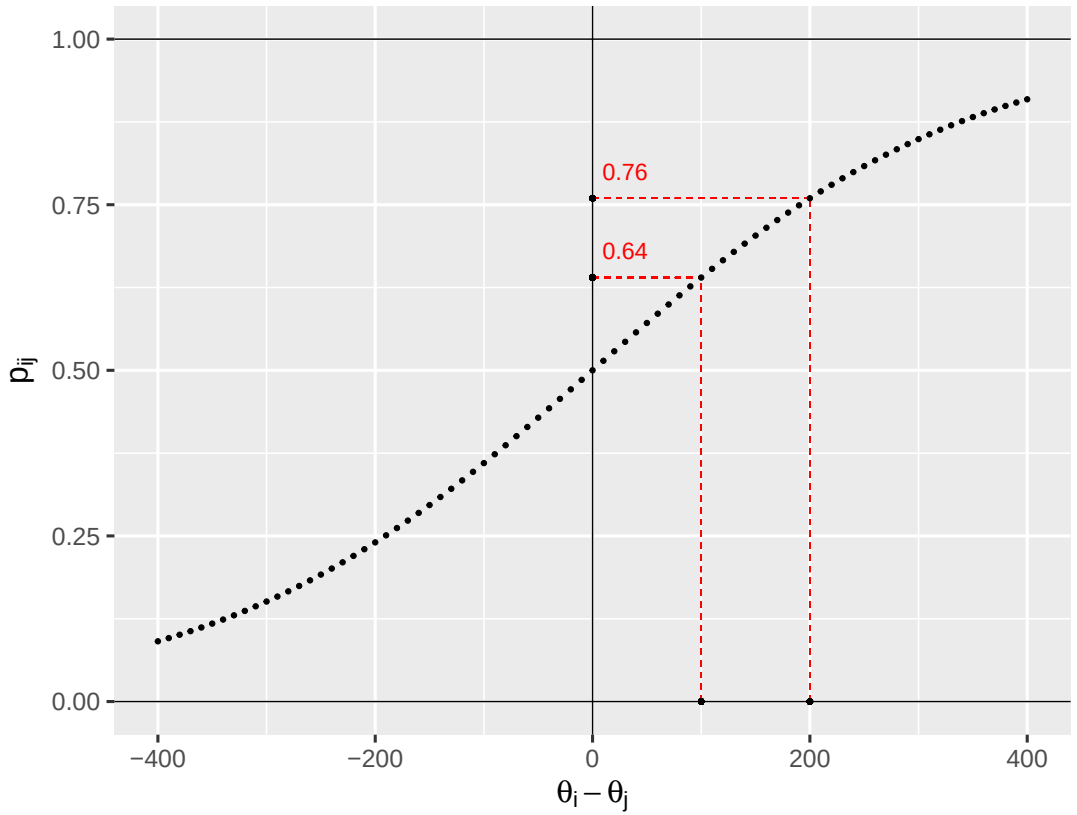


Figure: A player whose rating is 100 points greater than their opponent's is expected to score 64%; if the difference is 200 points, then the expected score for the stronger player is 76%.

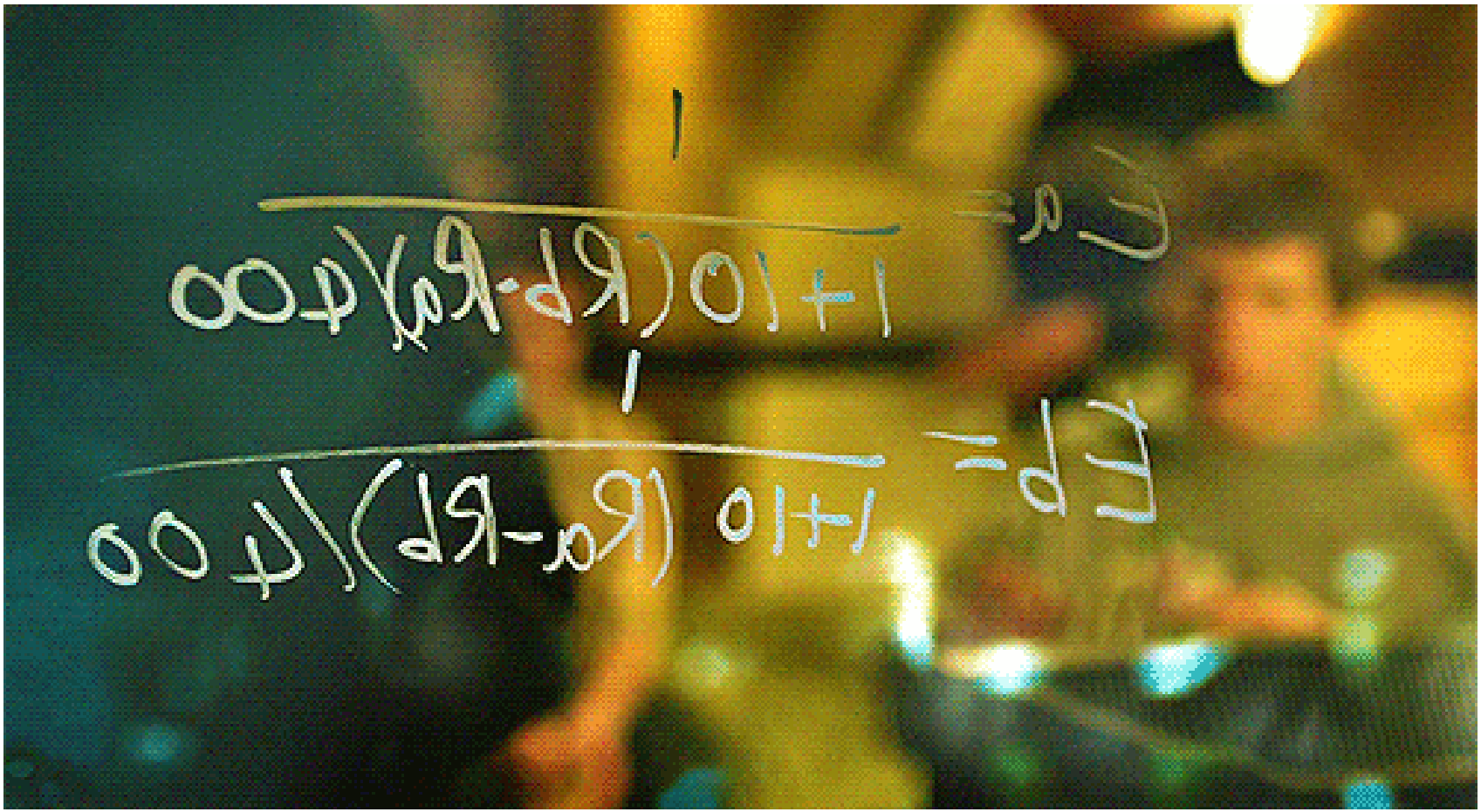


Figure: Scenes from the film "Social Network" (2010).

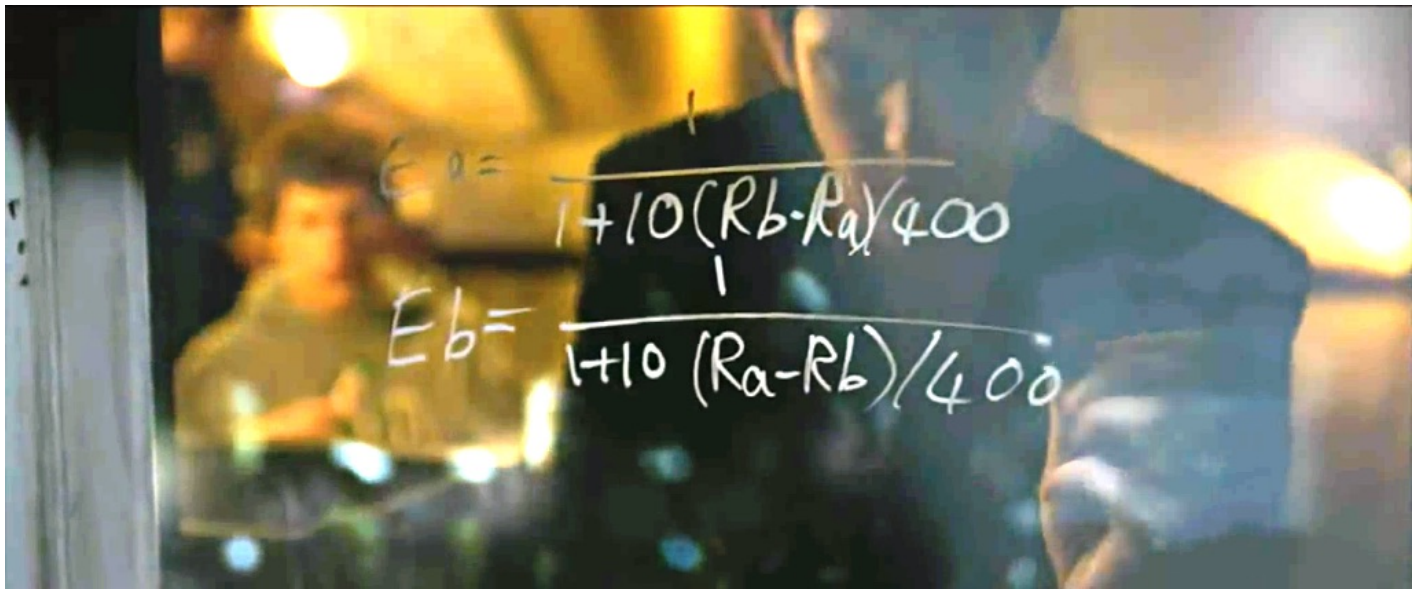


Figure: Mark Zuckerberg discusses Facemash, a social network for ranking girls preceding Facebook. Students are presented two random images of girls and they select the one they deem more beautiful. The girls are ranked with the Elo rating system.

Elo: update rules

- Suppose player i competes against player j in a match with outcome y_{ij}. Then the rating of player i updates according to the formula

$$\tilde{\theta}_i = \theta_i + K(y_{ij} - p_{ij}).$$

  Here θ_i is the rating before the match of player i and $\tilde{\theta}_i$ is the updated rating after the match.
- K is a constant known as the *K-factor* of the system.
- The K-factor is a tuning parameter of the system and determines the sensitivity of the ratings towards the outcome of a single game.

Elo: update rules

- Let us make a few observations about Elo's update rule

$$\tilde{\theta}_i = \theta_i + K(y_{ij} - p_{ij}).$$

- This is a balance equation due to $\mathbb{E}[Y_{ij}] = p_{ij}$.
- If the ratings of the two players have been determined correctly, then the expected change in ratings is zero.
- This is also a self-correcting mechanism. If the strengths of the players change and $\mathbb{E}[Y_{ij}] \neq p_{ij}$, the update rule will correct the ratings until equality.
- For example, if $\mathbb{E}[Y_{ij}] > p_{ij}$ then player i is stronger than what his current rating predicts and his rating points will increase on average.
- The rating points which a player receives are relative to the strength of the opponent.
- The K-factor gives the maximum change in ratings points due to a single game.

The zero-sum property of Elo ratings

- Classical Elo system with single K-factor.
- Notice that $y_{ij} = 1 - y_{ji}$ and $p_{ij} = 1 - p_{ji}$. It follows that the updated rating of player j is given by

$$\begin{aligned}\tilde{\theta}_j &= \theta_j + K(y_{ji} - p_{ji}) \\ &= \theta_j - K(y_{ij} - p_{ij}).\end{aligned}$$

- This means that the Elo system only redistributes rating points among players and we have

$$\tilde{\theta}_i + \tilde{\theta}_j = \theta_i + \theta_j.$$

Optimization algorithms in the context of linear regression

- Suppose we are given data $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ and we want to find a function $f_\theta(x) = \theta_0 + \theta_1 x$ that describes the relation between x and y.
- How to find the values of the unknown constants θ_0 and θ_1 that best describe the data?

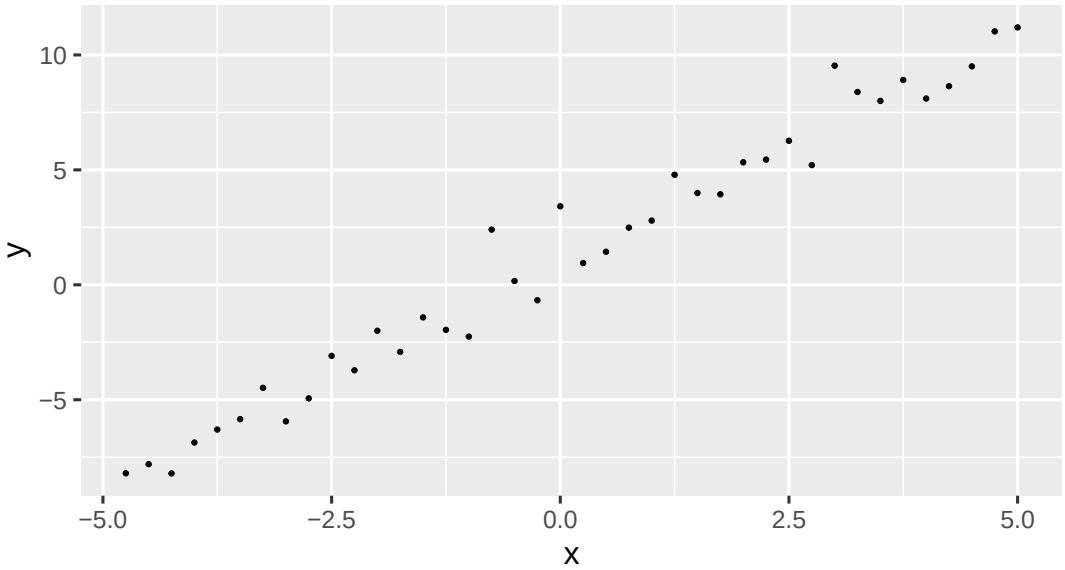


Figure: Fitting a regression line to data points.

Optimization algorithms in the context of linear regression

- Suppose we are given data $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ and we want to find a function $f_\theta(x) = \theta_0 + \theta_1 x$ that describes the relation between x and y.
- How to find the values of the unknown constants θ_0 and θ_1 that best describe the data?

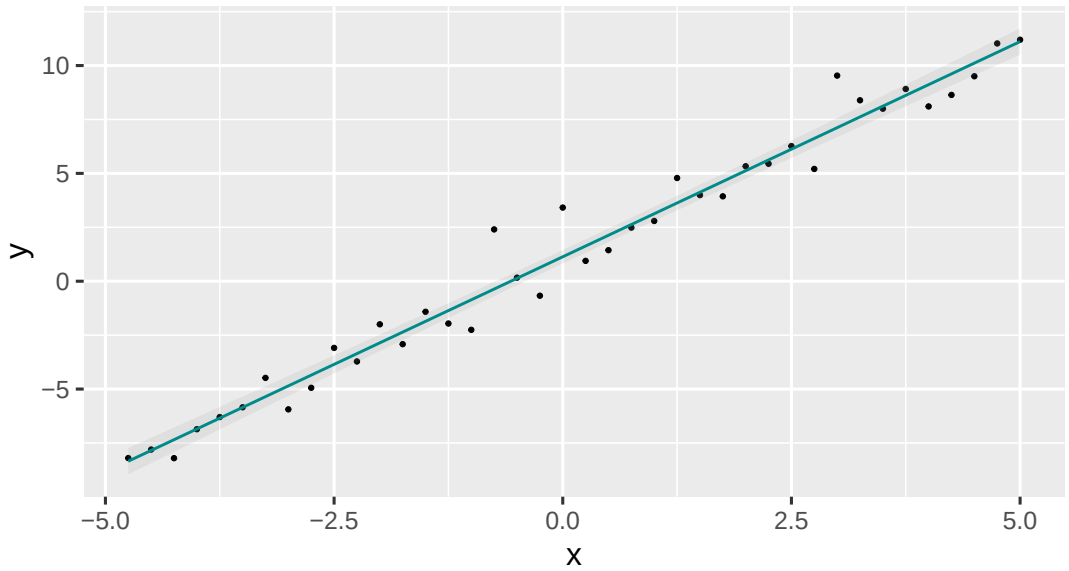


Figure: Fitting a regression line to data points.

Optimization algorithms in the context of linear regression

- We minimize an objective function:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (y_i - f_\theta(x_i))^2 \to \min_{\theta_0, \theta_1}.$$

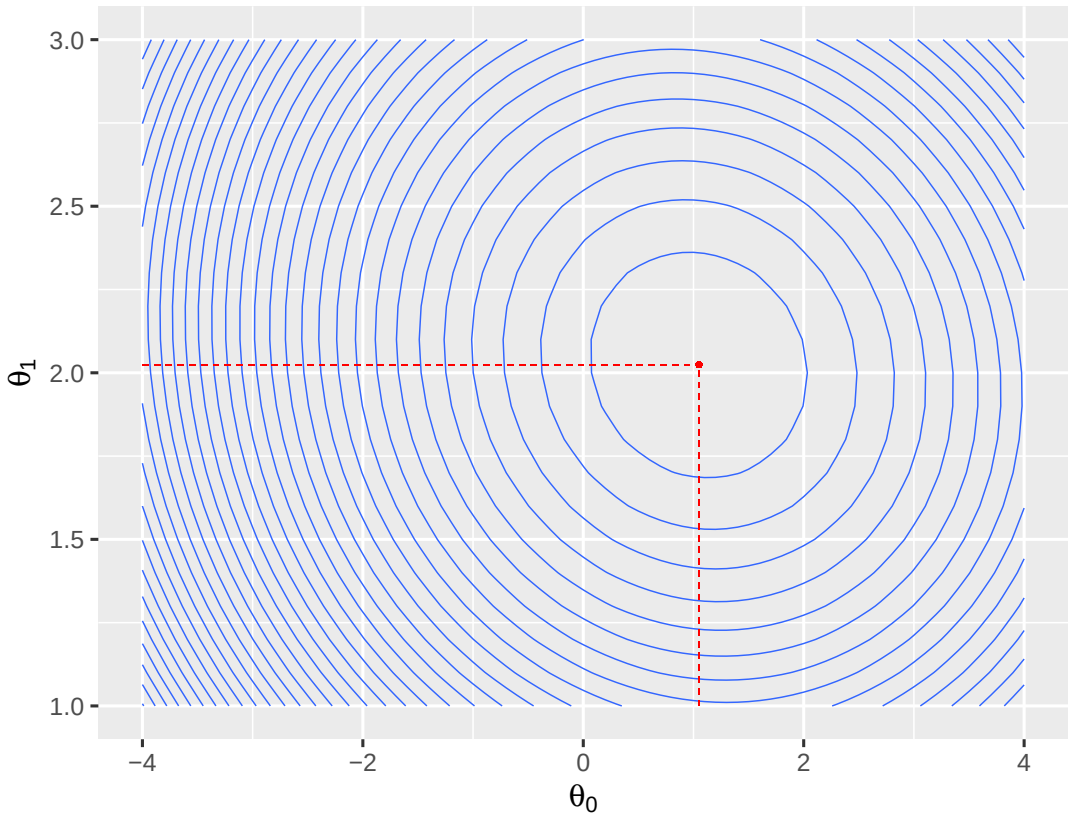


Figure: Isolines of $J(\theta_0, \theta_1)$.

Optimization algorithms in the context of linear regression

- Gradient descent: minimizing $J(\theta_0, \theta_1)$.

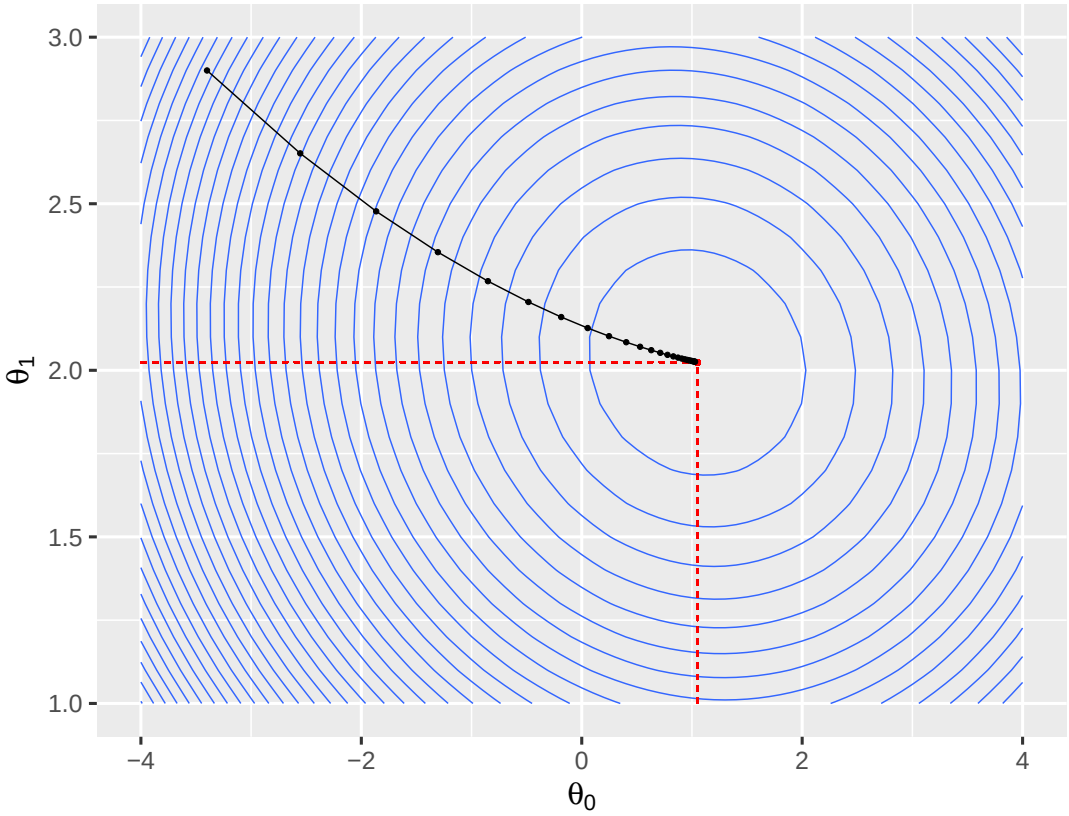


Figure: Isolines of $J(\theta_0, \theta_1)$.

Optimization algorithms in the context of linear regression

- Stochastic gradient descent: minimizing $J(\theta_0, \theta_1)$.

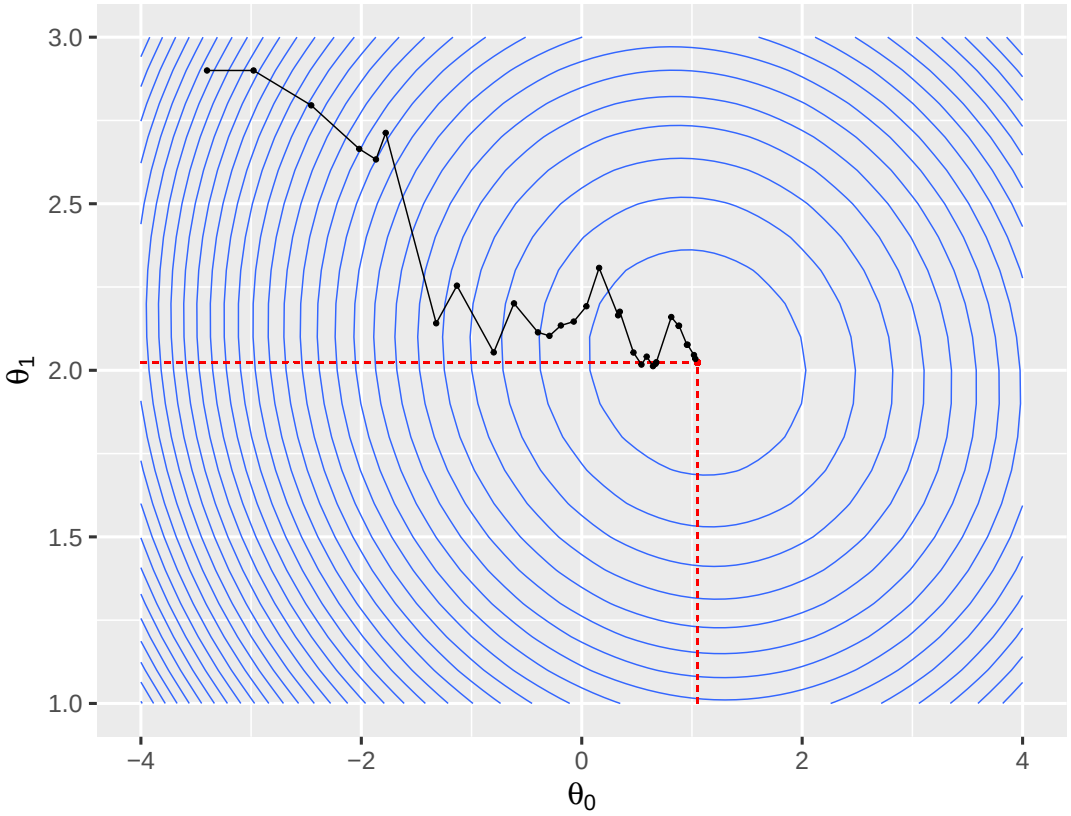


Figure: Isolines of $J(\theta_0, \theta_1)$.

Gradient descent

- Initialize $(\theta_0^{(0)}, \theta_1^{(0)})$ with an arbitrary point and then repeat until convergence:

$$\theta_0^{(i)} = \theta_0^{(i-1)} - K \frac{\partial}{\partial \theta_0} J(\theta_0^{(i-1)}, \theta_1^{(i-1)}),$$

$$\theta_1^{(i)} = \theta_1^{(i-1)} - K \frac{\partial}{\partial \theta_1} J(\theta_0^{(i-1)}, \theta_1^{(i-1)}),$$

  for $i = 1, 2, \ldots$

- The constant $K > 0$ determines the step size on each iteration.
- A step size too small leads to slow convergence, a step size too big leads to loss of precision.
- The gradient of the objective function always points to the direction of steepest increase of $J(\theta_0, \theta_1)$ and is perpendicular to its isolines.
- $J(\theta_0^{(i)}, \theta_1^{(i)})$ always decreases on each iteration, unless the minimum is skipped over.

Gradient descent

- Computing the gradient of the objective function can be expensive if m is big:

$$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \frac{\partial}{\partial \theta_0} (y_i - f_\theta(x_i))^2$$

$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \frac{\partial}{\partial \theta_1} (y_i - f_\theta(x_i))^2.$$

- It is easy to find real-world problems where the dimensionality d of the parameters $\theta = (\theta_1, \ldots, \theta_d)$ is of the order 10^5, while the sample size m is of the order 10^7 (for example, a FIDE chess dataset).
- With gradient descent you have to process the whole training set before you make one little step.
- With the stochastic gradient descent you start making progress after processing a single training example. It is a much faster optimization algorithm in the context of big data.

Stochastic gradient descent

- Initialize $(\theta_0^{(0)}, \theta_1^{(0)})$ with an arbitrary point and then repeat:

$$\theta_0^{(i)} = \theta_0^{(i-1)} - K \frac{\partial}{\partial \theta_0} \frac{1}{2} (y_i - f_\theta(x_i))^2,$$
$$\theta_1^{(i)} = \theta_1^{(i-1)} - K \frac{\partial}{\partial \theta_1} \frac{1}{2} (y_i - f_\theta(x_i))^2,$$

  for $i = 1, 2, \ldots, m$.

- In the stochastic gradient descent we iterate through each data point (x_i, y_i).
- Notice that in expectation the stochastic gradient gives the gradient of the objective function:

$$\mathbb{E} \frac{\partial}{\partial \theta_k} \frac{1}{2} (Y - f_\theta(X))^2 \approx \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial \theta_k} \frac{1}{2} (y_i - f_\theta(x_i))^2$$
$$= \frac{\partial}{\partial \theta_k} J(\theta_0, \theta_1),$$

  for $k = 0, 1$.

Figure: The only known portrait of Thomas Bayes, 1701-1761

- The seminal work of Thomas Bayes is called "An Essay Towards Solving a Problem in the Doctrine of Chances" published posthumously in 1764.
- It presents a method for statistical inference.
- Statistical inference is a method for solving the following problem. Suppose you are presented with a random sample from unknown distribution. You also have a parametric model of probability distributions which contains the unknown distribution of the data. The problem is to find the value of the parameter that best fits to the data.
- Bayes considered the parametric model

$$\Pr(X = x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

where n is known and $\theta \in [0, 1]$ is unknown and to be determined from x.

- The problem Bayes set for himself was this: find

$$\Pr(a < \theta < b \mid x).$$

- The example which Thomas Bayes used to illustrate his mathematical arguments.

Example (Thomas Bayes's example with the billiard table)

Imagine the following game. We have a billiard table with width 1. We roll the white ball across the table and let θ_0 be its horizontal coordinate after it stops moving. We then roll n balls and record the number x of balls whose final position is left of the white ball. If θ_0 is unknown, how to estimate it from knowing x?

- Bayes argued that one may take $\theta \sim U(0,1)$.
- We may now compute the joint probability distribution of X and θ,

$$\Pr(X = x, a < \theta < b) = \int_a^b \binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta,$$

- or the marginal distribution of X,

$$\begin{aligned} \Pr(X = x) &= \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta \\ &= \binom{n}{x} B(x+1, n-x+1) \\ &= \frac{1}{n+1}. \end{aligned}$$

for $x = 0, 1, \ldots, n$.

- Thus, the solution to the billiard table problem is

$$\Pr(a < \theta < b \mid x) = (n+1)\binom{n}{x}\int_a^b \theta^x(1-\theta)^{n-x}d\theta,$$

  which is also the basis of today's credible interval.
- Different people read different things in Bayes's philosophical essay.
- Some find conformation to the so called *insufficient reasoning principle*, according to which we select a "flat" prior to an unknown parameter $\theta \sim U(0,1)$ if we have no reason to believe that one value of θ is more likely than another.
- Others (Stigler, Molina, Edwards) claim that Bayes's true argument is based on the marginal distribution of X and it goes like this: if we are in situation in which we know absolutely nothing about θ antecedentally, then we must assume a uniform distribution of X. Otherwise, if we view that x_1 is more likely than x_2, for example, then that would give us a reason to believe that $\theta_1 = x_1/n$ is more likely than $\theta_2 = x_2/n$.

- Placing uniform prior on the "evidence" X leads to selecting a uniform prior on the parameter θ in the billiard table example.
- This is a much more stringent condition than the principle of insufficient reason later developed by Laplace.
- In modern day Bayesian statistics one selects a prior distribution $\pi_0(\theta)$ for the unknown parameter θ before seeing the data x, then forms the joint distribution

$$P(x, \theta) = P(x \mid \theta)\pi_0(\theta),$$

which after normalization in x leads to the conditional distribution

$$\begin{aligned} P(\theta \mid x) &= \frac{P(x \mid \theta)\pi_0(\theta)}{\int P(x \mid \theta)\pi_0(\theta)d\theta} \\ &= \pi_1(\theta). \end{aligned}$$

The distribution $\pi_1(\theta)$ is called the posterior distribution.

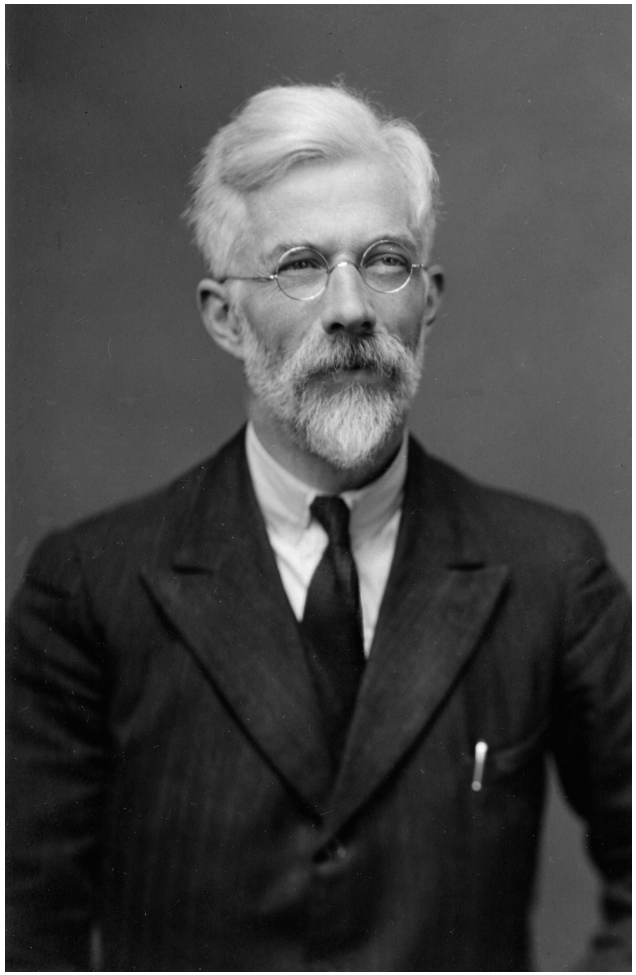


Figure: Portrait of Ronald Fisher, 1890-1962

Likelihood

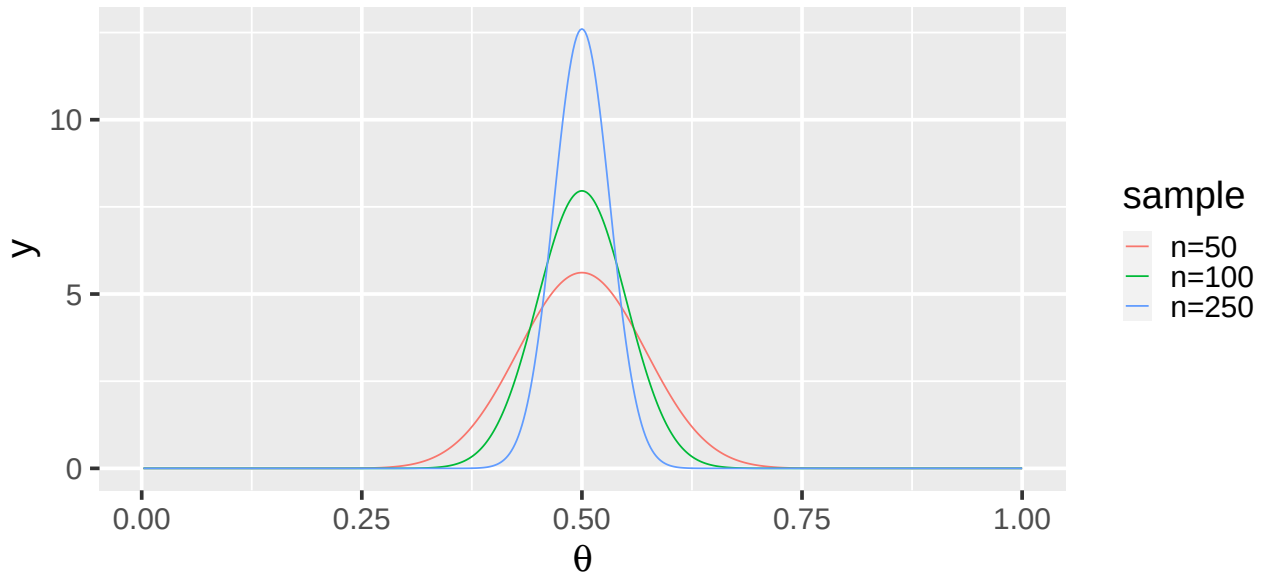
- Why is the *likelihood function* so important in machine learning and mathematical statistics?.
- The case for using likelihood was first made by R. Fisher, who believed it to be a self-contained framework for statistical modelling and inference.
- The likelihood function:
  - measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters.
  - is formed from the joint probability distribution of the sample, but viewed and used as a function of the parameters only.
- R. Fisher developed the method of maximum likelihood estimation between 1912 and 1922.

The likelihood

- In the billiard table example the likelihood function is

$$L_n(\theta) = \Pr(x \mid \theta)$$
$$= \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

Figure: $L_n(\theta)$ for $x = n/2$ is the density of the Beta(n/2, n/2) distribution.



The likelihood

- The maximum of the likelihood concentrates on the true value as the sample size $n \to \infty$.
- This general fact can be easily proved by using the Kullback-Leiback divergence.
- It can be obtained analytically by solving the equation $(\log L_n(\theta))' = 0$.
- Let us demonstrate

$$\begin{aligned}(\log L_n(\theta))' &= (x \ln \theta + (n-x)\ln(1-\theta))' \\ &= \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0.\end{aligned}$$

This leads to

$$(1-\theta)x - (n-x)\theta = 0,$$

or $\theta = x/n$.

The likelihood

- The maximum likelihood estimator $\hat{\theta}_n$ is *consistent* estimator, that is, $\hat{\theta}_n \to \theta_0$ in probability as $n \to \infty$, where θ_0 is the true unknown parameter of the distribution of the sample.
- The maximum likelihood estimator $\hat{\theta}_n$ is *asymptotically normal*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to N\left(0, \frac{1}{I(\theta_0)}\right)$$

with the asymptotic variance

$$I(\theta_0) = \mathbb{E}_{\theta_0}\left(\frac{\partial}{\partial\theta} \ln P(X \mid \theta)\Big|_{\theta=\theta_0}\right)^2$$

is *Fisher information*.

- The MLE $\hat{\theta}_n$ is also a function of sufficient statistics and is an asymptotically efficient estimator.

A Bayesian rating system: an illustration

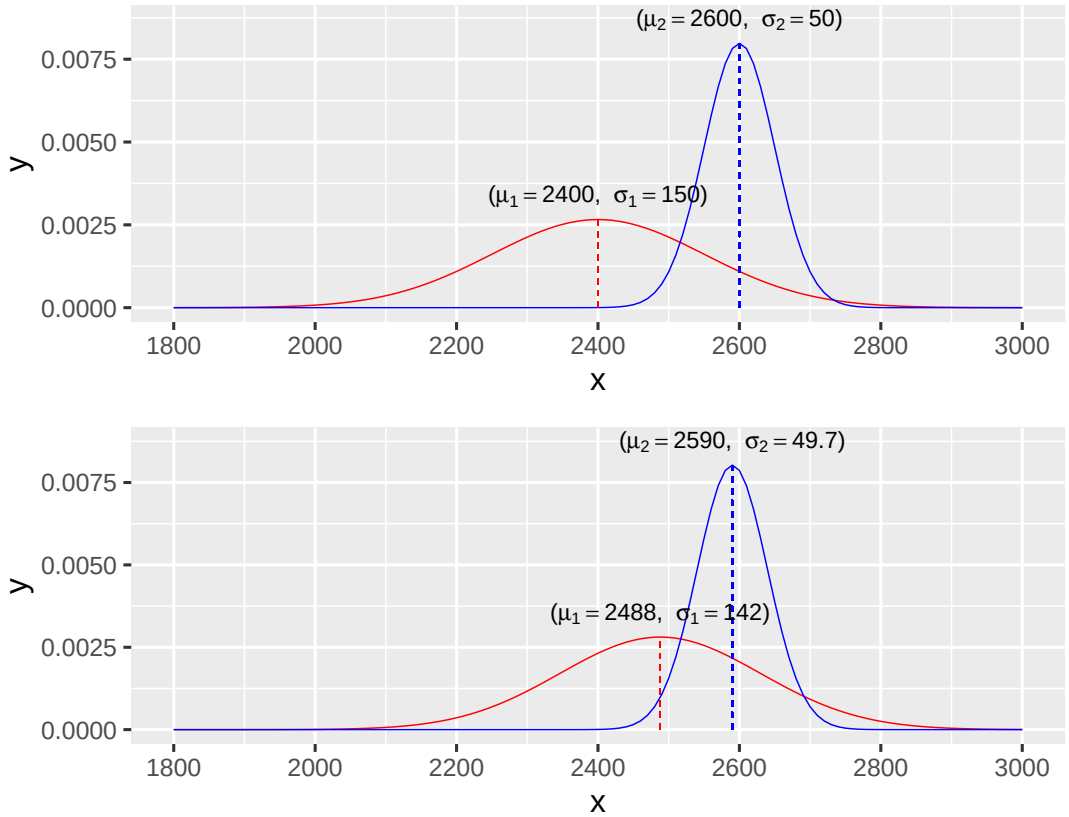


Figure: The old and updated ratings of two players in Bayesian Elo after a single game in which player 1 beats player 2.