Linear Regression

Neil D. Lawrence

GPRS
6th August 2013

Outline

Outline

Introduction

Useful Texts

Regression

Rogers and Girolami

Bishop


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP

Outline

Regression Examples

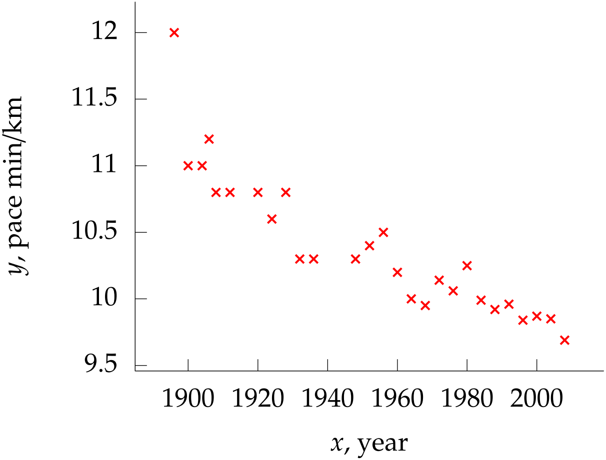
- Predict a real value, y_i given some inputs $\mathbf{x}_i$.
- Predict quality of meat given spectral measurements (Tecator data).
- Radiocarbon dating, the C14 calibration curve: predict age given quantity of C14 isotope.
- Predict quality of different Go or Backgammon moves given expert rated training data.

Olympic 100m Data

- Gold medal times for Olympic 100 m runners since 1896.

Image from Wikimedia Commons `http://bit.ly/191adDC`

Olympic 100m Data



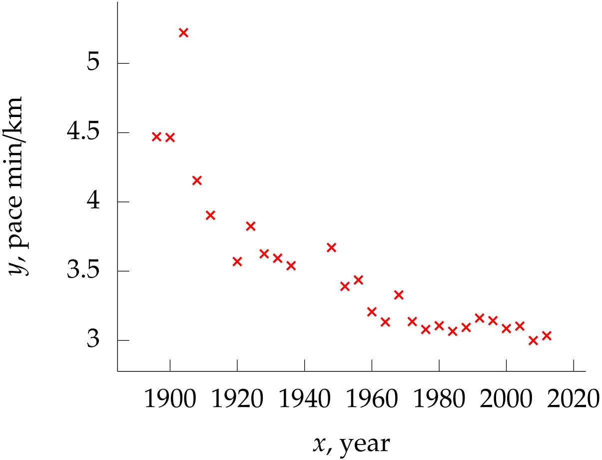
Olympic 100 m Data.

Olympic Marathon Data

- Gold medal times for Olympic Marathon since 1896.
- Marathons before 1924 didn't have a standardised distance.
- Present results using pace per km.
- In 1904 Marathon was badly organised leading to very slow times.

Image from Wikimedia Commons
`http://bit.ly/16kMKHQ`

Olympic Marathon Data



Olympic Marathon Data.

What is Machine Learning?

data

- data: observations, could be actively or passively acquired (meta-data).

What is Machine Learning?

data +

- data: observations, could be actively or passively acquired (meta-data).

What is Machine Learning?

data + model

- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

What is Machine Learning?

data + model =

- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.

What is Machine Learning?

$$\text{data} \quad + \quad \text{model} \quad = \quad \text{prediction}$$

- data: observations, could be actively or passively acquired (meta-data).
- model: assumptions, based on previous experience (other data! transfer learning etc), or beliefs about the regularities of the universe. Inductive bias.
- prediction: an action to be taken or a categorization or a quality score.

Regression: Linear Releationship

$$y = mx + c$$

- y: winning time/pace.

Regression: Linear Releationship

$$y = mx + c$$

- y: winning time/pace.
- x: year of Olympics.

Regression: Linear Releationship

$$y = mx + c$$

- y: winning time/pace.
- x: year of Olympics.
- m: rate of improvement over time.

Regression: Linear Releationship

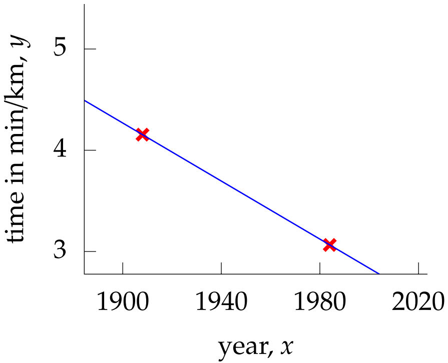
$$y = mx + c$$

- y: winning time/pace.
- x: year of Olympics.
- m: rate of improvement over time.
- c: winning time at year 0.

Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

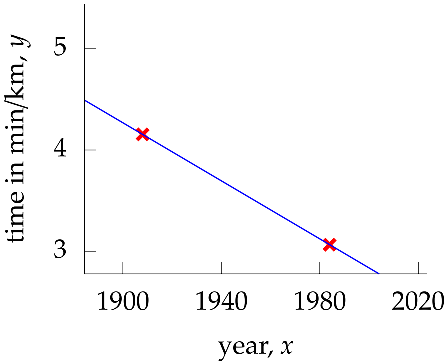
$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$



Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

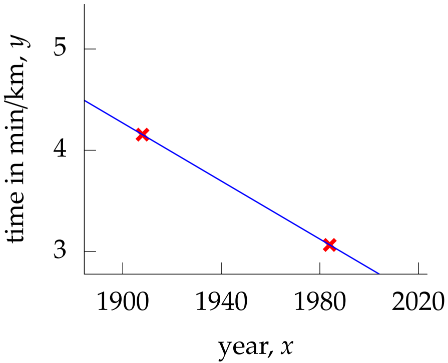
$$y_1 - y_2 = m(x_1 - x_2)$$



Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

$$\frac{y_1 - y_2}{x_1 - x_2} = m$$

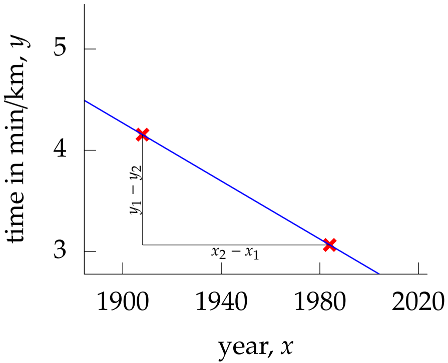


Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

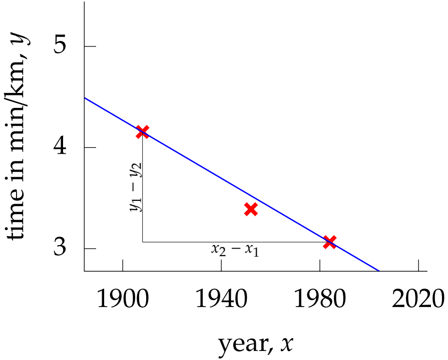
$$c = y_1 - m x_1$$



Two Simultaneous Equations

How do we deal with three simultaneous equations with only two unknowns?

$$\begin{aligned} y_1 =& mx_1 + c \\ y_2 =& mx_2 + c \\ y_3 =& mx_3 + c \end{aligned}$$



Overdetermined System

- With two unknowns and two observations:

$$y_1 = mx_1 + c$$
$$y_2 = mx_2 + c$$

Overdetermined System

- With two unknowns and two observations:

$$
\begin{aligned}
y_1 &= mx_1 + c \\
y_2 &= mx_2 + c
\end{aligned}
$$

- Additional observation leads to *overdetermined* system.

$$
y_3 = mx_3 + c
$$

Overdetermined System

- With two unknowns and two observations:

$$
\begin{aligned}
y_1 &= mx_1 + c \\
y_2 &= mx_2 + c
\end{aligned}
$$

- Additional observation leads to *overdetermined* system.

$$
y_3 = mx_3 + c
$$

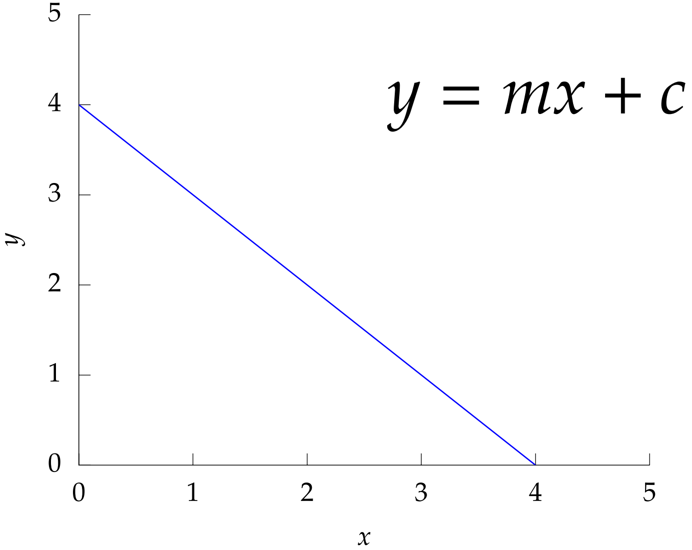
- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$
\begin{aligned}
y_1 &= mx_1 + c + \epsilon_1 \\
y_2 &= mx_2 + c + \epsilon_2 \\
y_3 &= mx_3 + c + \epsilon_3
\end{aligned}
$$

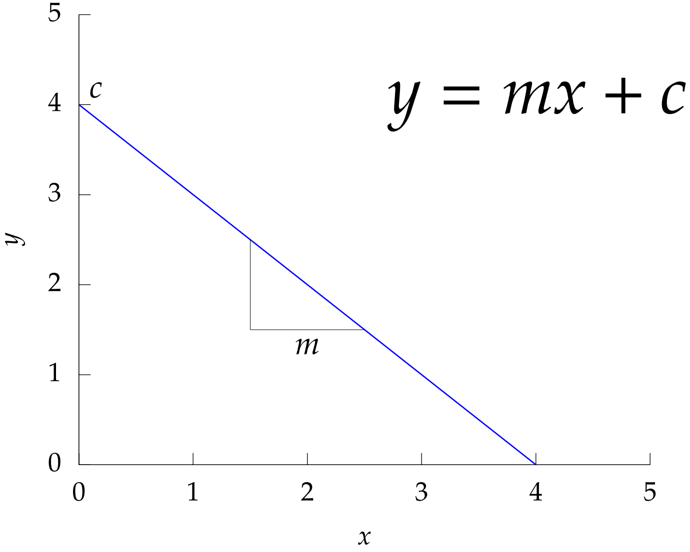
Noise Models

- We aren't modeling entire system.
- Noise model gives mismatch between model and data.
- Gaussian model justified by appeal to central limit theorem.
- Other models also possible (Student-t for heavy tails).
- Maximum likelihood with Gaussian noise leads to *least squares*.

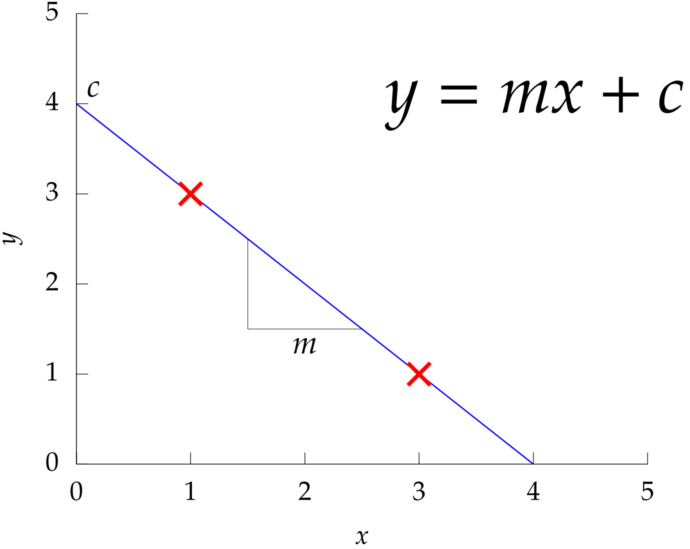
$$y = mx + c$$



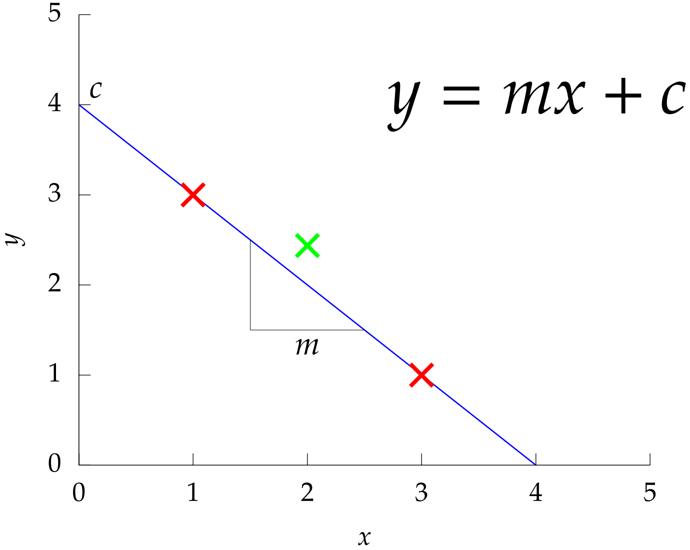
$y = mx + c$
0
1
2
3
4
5
x
y



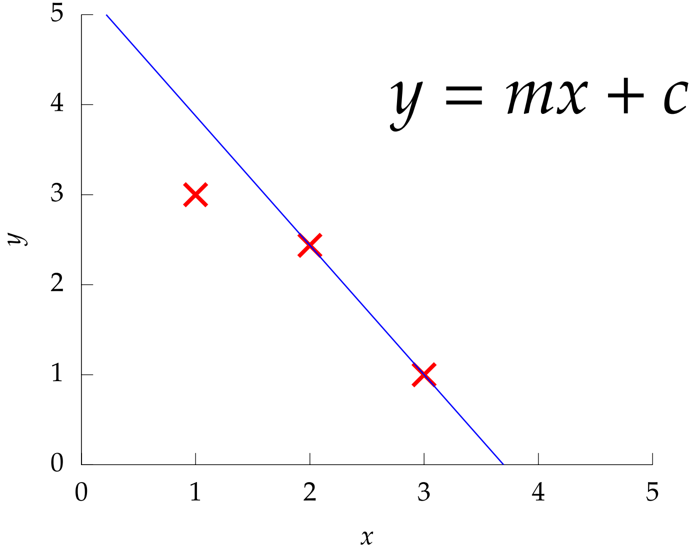
$y = mx + c$
c
m
0
1
2
3
4
5
x
y



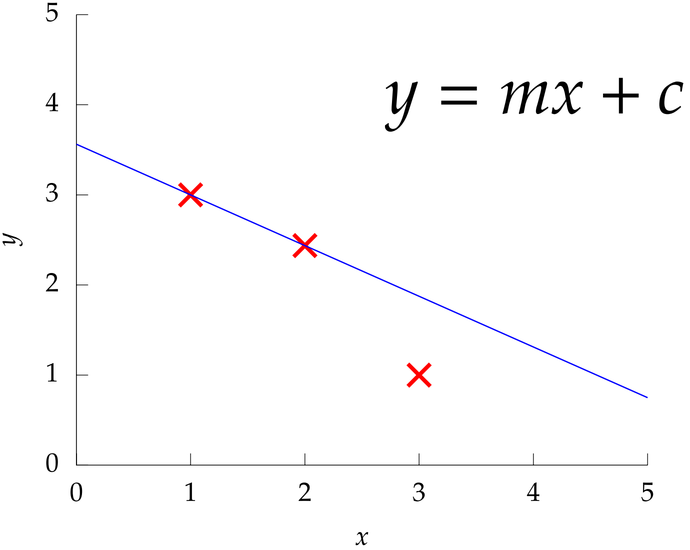
$y = mx + c$
c
m
y
x
0
1
2
3
4
5

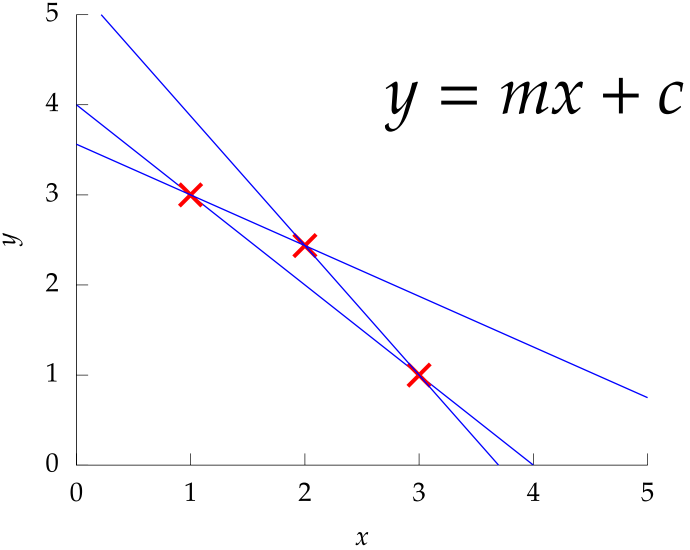


$y = mx + c$
c
m
5
4
3
2
1
0
0
1
2
3
4
5
y
x



$y = mx + c$
5
4
3
2
1
0
0
1
2
3
4
5
y
x


$y = mx + c$
0
1
2
3
4
5
x
y



$y = mx + c$
5
4
3
2
1
0
0
1
2
3
4
5
y
x

point 1: $x = 1$, $y = 3$

$$3 = m + c$$

point 2: $x = 3$, $y = 1$

$$1 = 3m + c$$

point 3: $x = 2$, $y = 2.5$

$$2.5 = 2m + c$$

point 1: $x = 1, y = 3$

$$3 = m + c + \epsilon_1$$

point 2: $x = 3, y = 1$

$$1 = 3m + c + \epsilon_2$$

point 3: $x = 2, y = 2.5$

$$2.5 = 2m + c + \epsilon_3$$

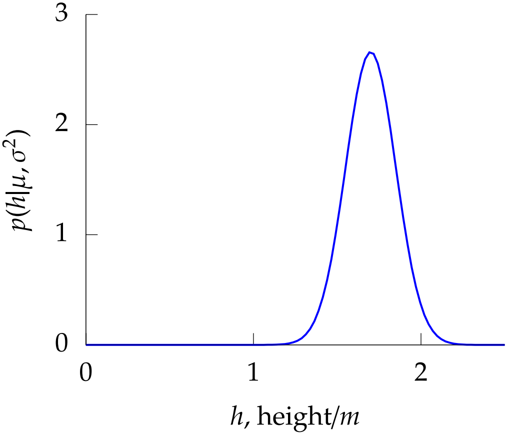
The Gaussian Density

- Perhaps the most common probability density.

$$
\begin{aligned}
p(y|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\
&\triangleq \mathcal{N}\left(y|\mu, \sigma^2\right)
\end{aligned}
$$

- The Gaussian density.

Gaussian Density



The Gaussian PDF with $\mu = 1.7$ and variance $\sigma^2 = 0.0225$. Mean shown as red line. It could represent the heights of a population of students.

Gaussian Density

$$\mathcal{N}\left(y|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

σ^2 is the variance of the density and μ is the mean.

Two Important Gaussian Properties

Sum of Gaussians

- Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

Two Important Gaussian Properties

Sum of Gaussians

- Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

And the sum is distributed as

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Two Important Gaussian Properties

Sum of Gaussians

- Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

And the sum is distributed as

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

Two Important Gaussian Properties

Sum of Gaussians

- Sum of Gaussian variables is also Gaussian.

$$y_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

And the sum is distributed as

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(*Aside*: As sum increases, sum of non-Gaussian, finite variance variables is also Gaussian [central limit theorem].)

Two Important Gaussian Properties

Scaling a Gaussian

- Scaling a Gaussian leads to a Gaussian.

Two Important Gaussian Properties

Scaling a Gaussian

- Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

Two Important Gaussian Properties

Scaling a Gaussian

- Scaling a Gaussian leads to a Gaussian.

$$y \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

And the scaled density is distributed as

$$wy \sim \mathcal{N}\left(w\mu, w^2\sigma^2\right)$$

A Probabilistic Process

- Set the mean of Gaussian to be a function.

$$p\left(y_i|x_i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - f\left(x_i\right)\right)^2}{2\sigma^2}\right).$$

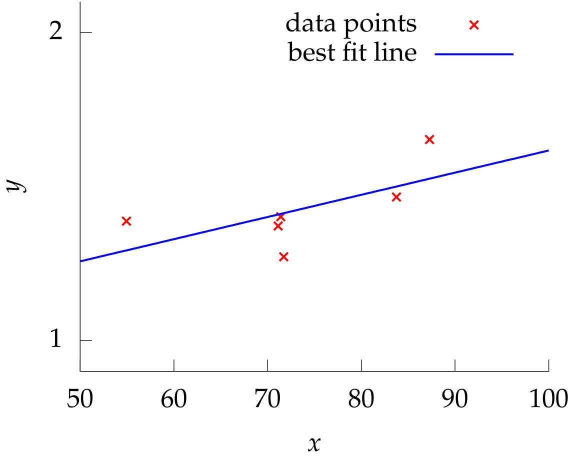
- This gives us a 'noisy function'.
- This is known as a process.

y as a Function of x

- In the standard Gaussian, parametized by mean and variance.
- Make the mean a linear function of an *input*.
- This leads to a regression model.

$$
\begin{aligned}
y_i =& f(x_i) + \epsilon_i, \\
\epsilon_i \sim& \mathcal{N}\left(0, \sigma^2\right).
\end{aligned}
$$

Linear Function



A linear regression between x and y.

Data Point Likelihood

- Likelihood of an individual data point

$$p\left(y_i|x_i, m, c\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - mx_i - c\right)^2}{2\sigma^2}\right).$$

- Parameters are gradient, m, offset, c of the function and noise variance σ^2.

Data Set Likelihood

- If the noise, ϵ_i is sampled independently for each data point.
- Each data point is independent (given m and c).
- For independent variables:

$$p(\mathbf{y}) = \prod_{i=1}^{n} p(y_i)$$

Data Set Likelihood

- If the noise, ϵ_i is sampled independently for each data point.
- Each data point is independent (given m and c).
- For independent variables:

$$p(\mathbf{y}|\mathbf{x}, m, c) = \prod_{i=1}^{n} p(y_i|x_i, m, c)$$

Data Set Likelihood

- If the noise, ϵ_i is sampled independently for each data point.
- Each data point is independent (given m and c).
- For independent variables:

$$p(\mathbf{y}|\mathbf{x}, m, c) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - mx_i - c)^2}{2\sigma^2}\right).$$

Data Set Likelihood

- If the noise, ϵ_i is sampled independently for each data point.
- Each data point is independent (given m and c).
- For independent variables:

$$p(\mathbf{y}|\mathbf{x}, m, c) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \exp\left(-\frac{\sum_{i=1}^{n}\left(y_i - mx_i - c\right)^2}{2\sigma^2}\right).$$

Log Likelihood Function

- Normally work with the log likelihood:

$$L(m, c, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \sum_{i=1}^{n} \frac{(y_i - mx_i - c)^2}{2\sigma^2}.$$

Consistency of Maximum Likelihood

- If data was really generated according to probability we specified.
- Correct parameters will be recovered in limit as $n \rightarrow \infty$.
- This can be proven through sample based approximations (law of large numbers) of "KL divergences".
- Mainstay of classical statistics.

Probabilistic Interpretation of the Error Function

- Probabilistic Interpretation for Error Function is Negative Log Likelihood.
- *Minimizing* error function is equivalent to *maximizing* log likelihood.
- Maximizing *log likelihood* is equivalent to maximizing the *likelihood* because log is monotonic.
- Probabilistic interpretation: Minimizing error function is equivalent to maximum likelihood with respect to parameters.

Error Function

- Negative log likelihood is the error function leading to an error function

$$E(m, c, \sigma^2) = \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - mx_i - c)^2 .$$

- Learning proceeds by minimizing this error function for the data set provided.

Connection: Sum of Squares Error

- Ignoring terms which don't depend on m and c gives

$$E(m, c) \propto \sum_{i=1}^{n} (y_i - f(x_i))^2$$

where $f(x_i) = mx_i + c$.

- This is known as the *sum of squares* error function.
- Commonly used and is closely associated with the Gaussian likelihood.

Mathematical Interpretation

- What is the mathematical interpretation?
  - There is a cost function.
  - It expresses mismatch between your prediction and reality.

$$E(\mathbf{w}) = \sum_{i=1}^{n} \left(y_i - mx_i + c - y_i\right)^2$$

  - This is known as the sum of squares error.

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$\frac{\mathrm{d}E(m)}{\mathrm{d}m} = -2\sum_{i=1}^{n} x_i \left(y_i - m x_i - c\right)$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$0 = -2\sum_{i=1}^{n} x_i \left(y_i - m x_i - c\right)$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$0 = -2\sum_{i=1}^{n} x_i y_i + 2\sum_{i=1}^{n} m x_i^2 + 2\sum_{i=1}^{n} c x_i$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$m = \frac{\sum_{i=1}^{n} (y_i - c)\, x_i}{\sum_{i=1}^{n} x_i^2}$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$\frac{\mathrm{d}E(c)}{\mathrm{d}c} = -2\sum_{i=1}^{n}\left(y_i - mx_i - c\right)$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$0 = -2 \sum_{i=1}^{n} (y_i - m x_i - c)$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

$$0 = -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} m x_i + 2nc$$

Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Coordinate ascent, find gradient in each coordinate and set to zero.

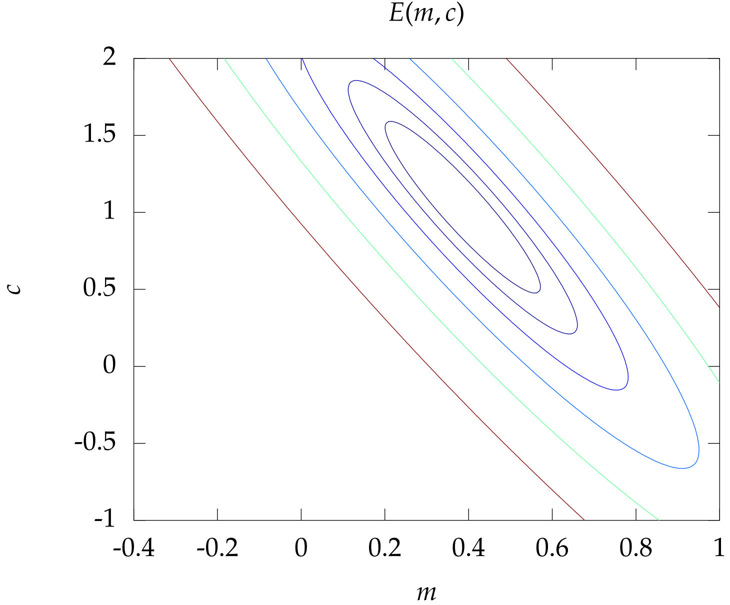
$$c = \frac{\sum_{i=1}^{n} (y_i - cx)}{n}$$

Fixed Point Updates

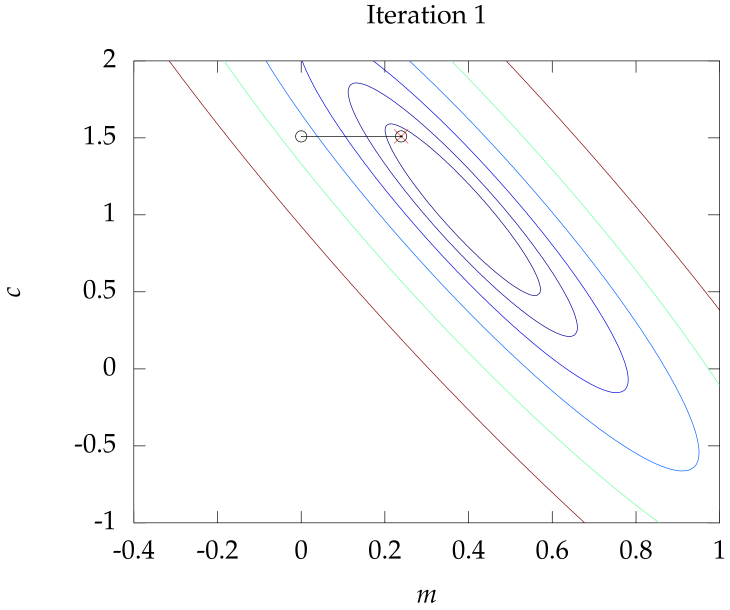
Worked example.

$$
\begin{aligned}
c^* &= \frac{\sum_{i=1}^{n} \left(y_i - m^* x_i\right)}{n}, \\
m^* &= \frac{\sum_{i=1}^{n} x_i \left(y_i - c^*\right)}{\sum_{i=1}^{n} x_i^2}, \\
\sigma^{2\,*} &= \frac{\sum_{i=1}^{n} \left(y_i - m^* x_i - c^*\right)^2}{n}
\end{aligned}
$$

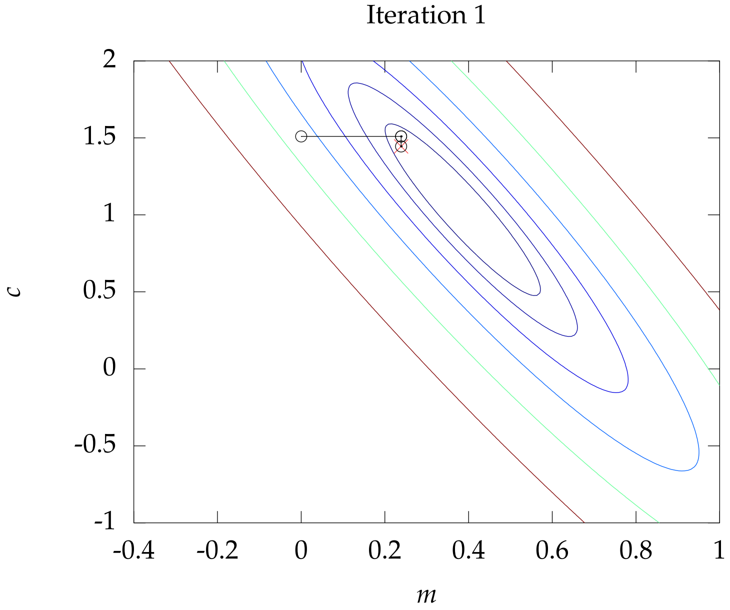
Coordinate Descent



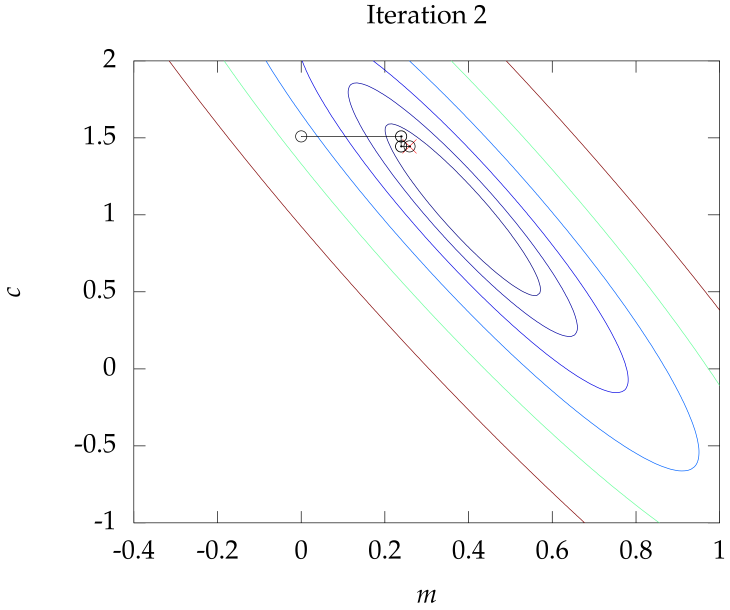
Coordinate Descent



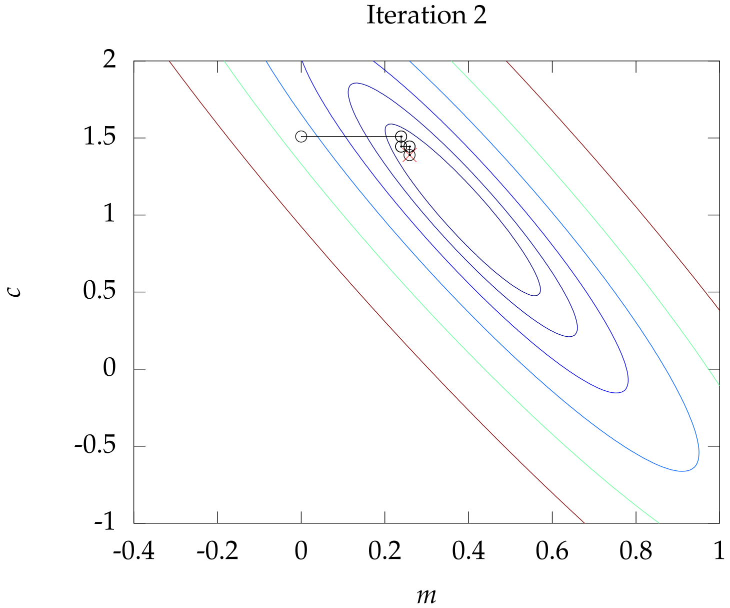
Coordinate Descent



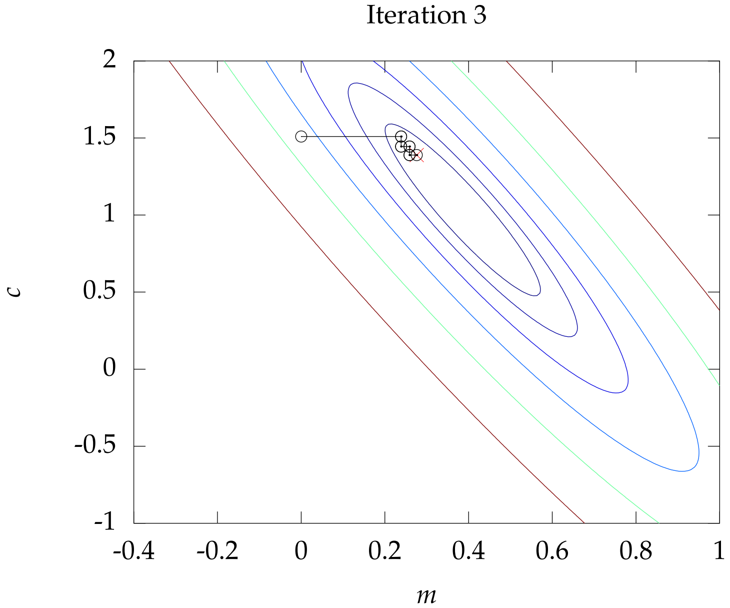
Coordinate Descent



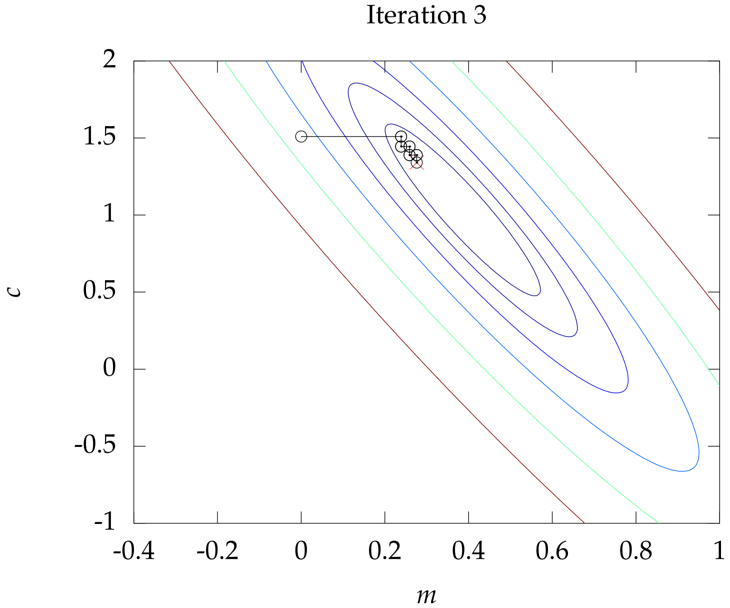
Coordinate Descent



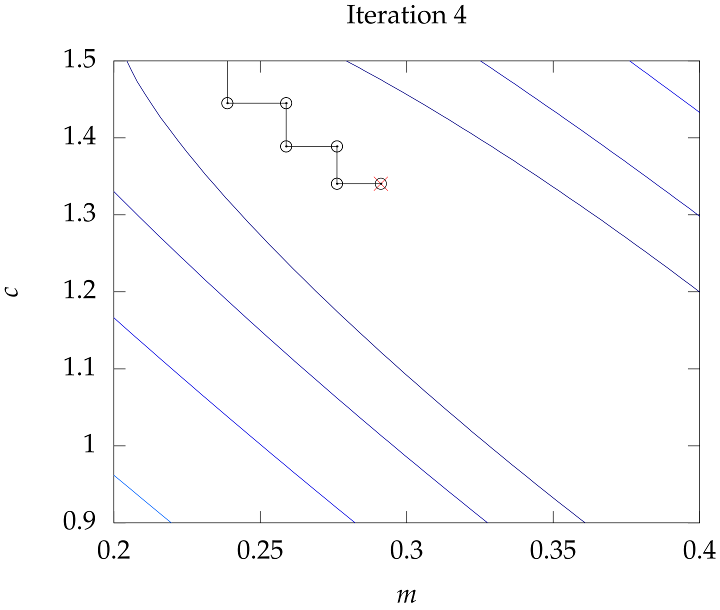
Coordinate Descent



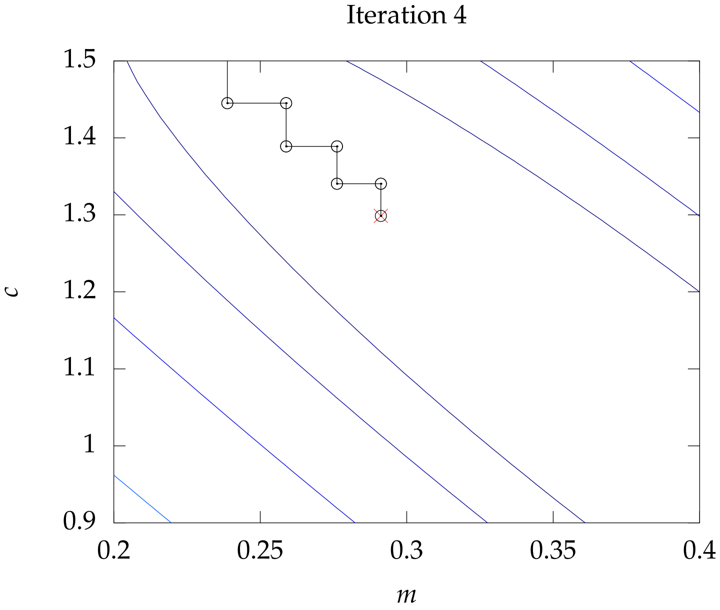
Coordinate Descent



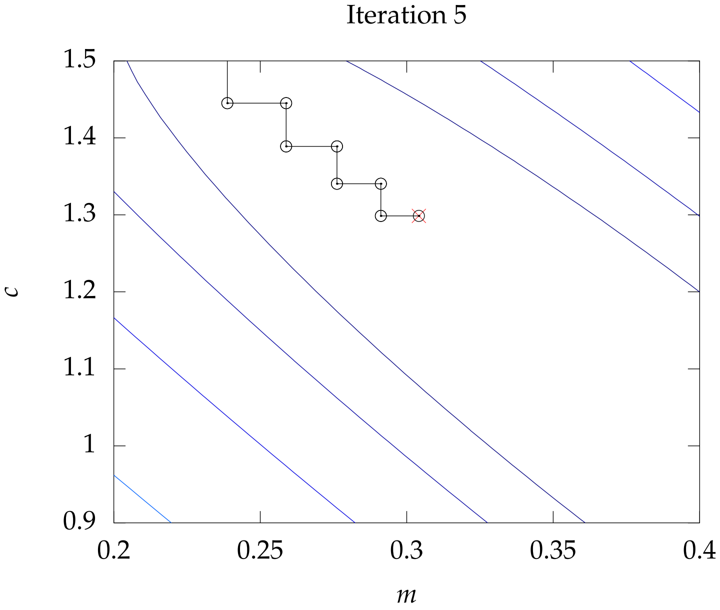
Coordinate Descent



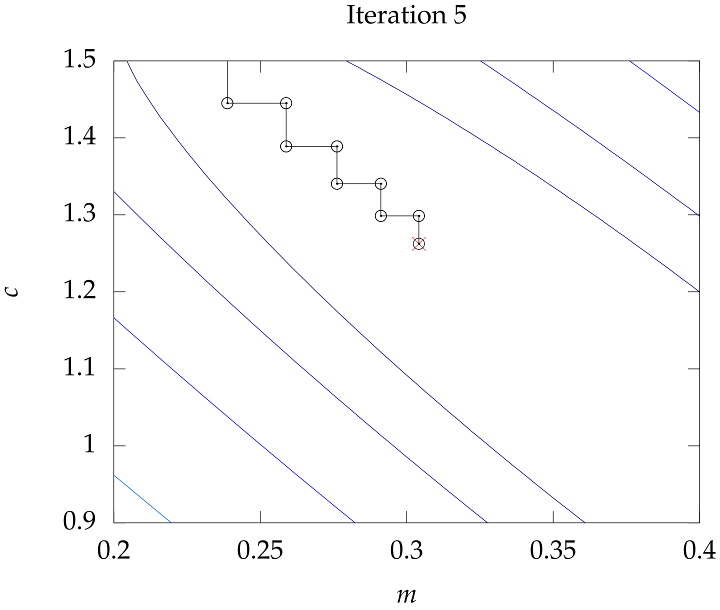
Coordinate Descent



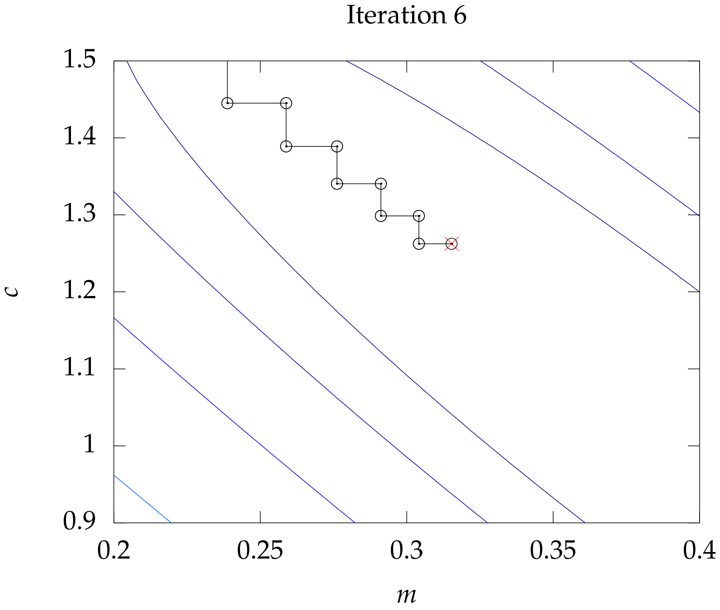
Coordinate Descent



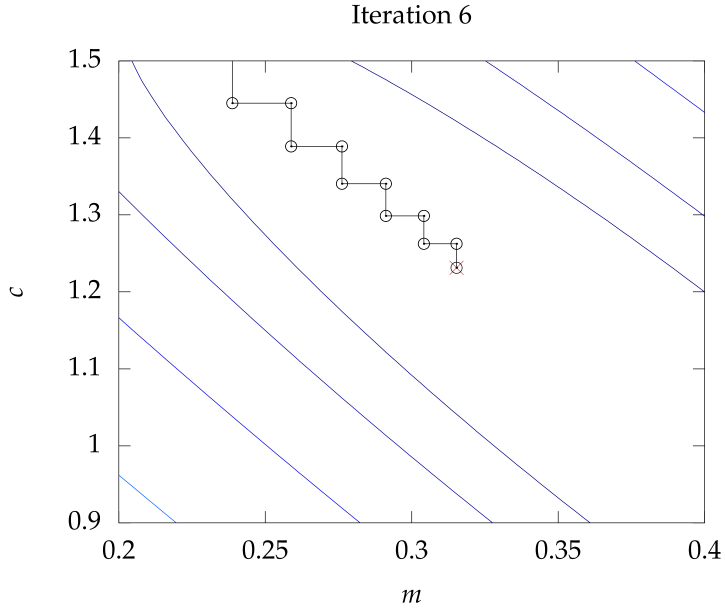
Coordinate Descent



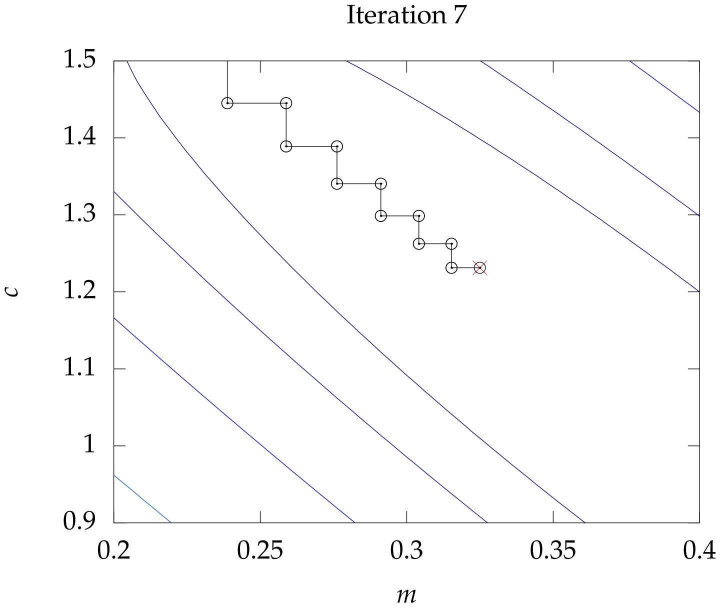
Coordinate Descent



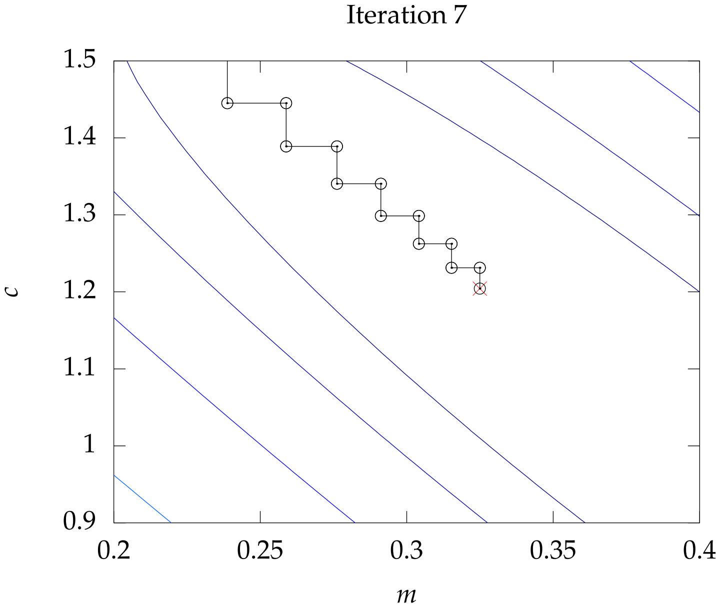
Coordinate Descent



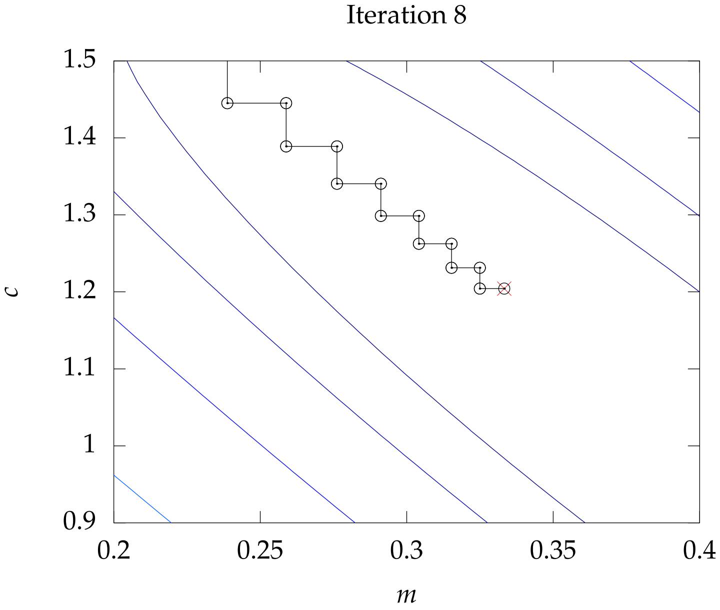
Coordinate Descent



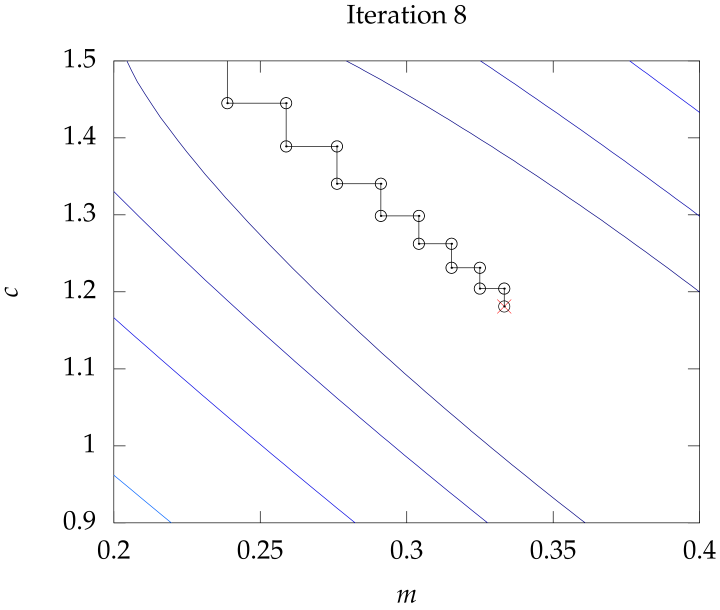
Coordinate Descent



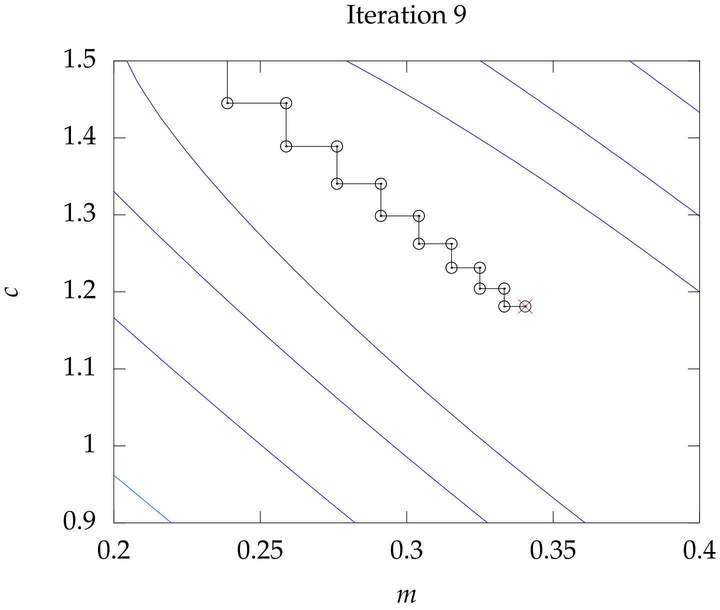
Coordinate Descent



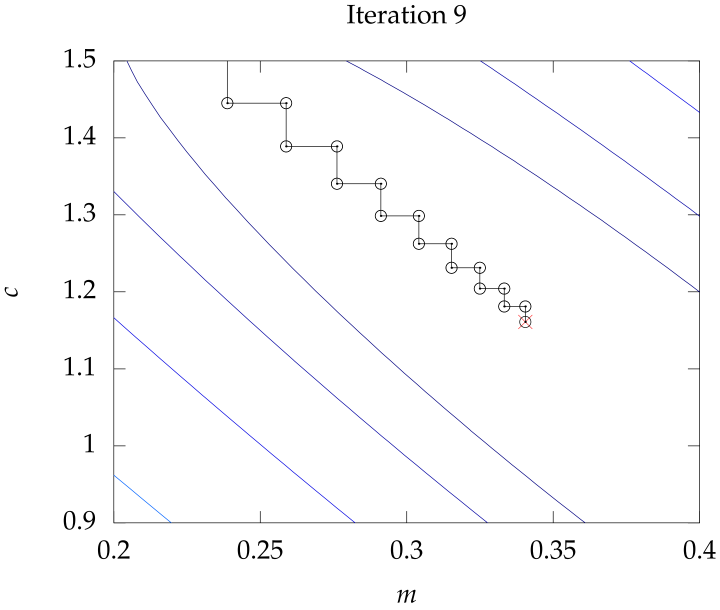
Coordinate Descent



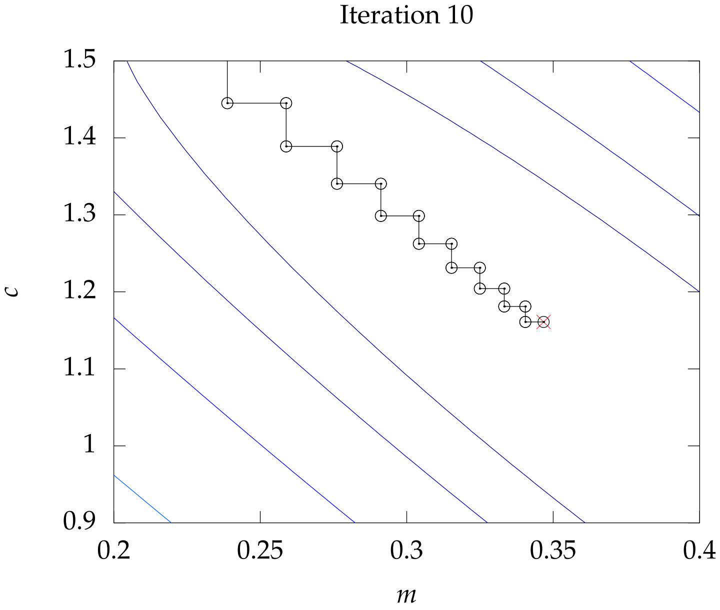
Coordinate Descent



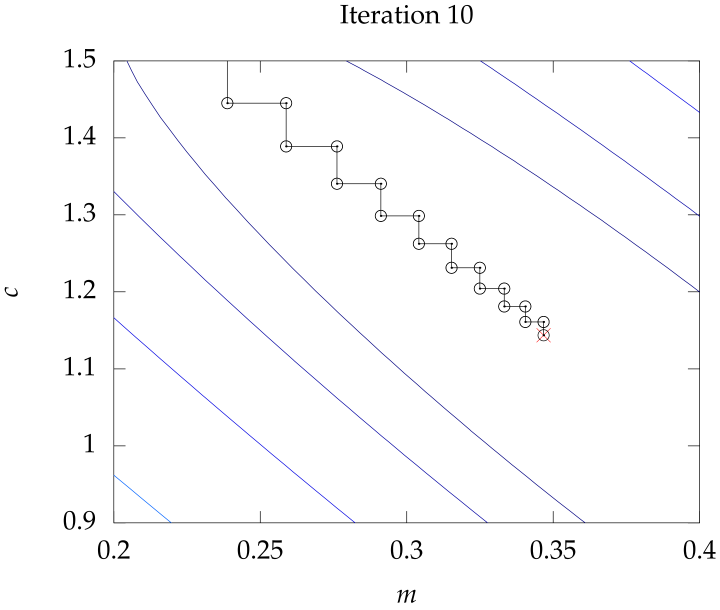
Coordinate Descent



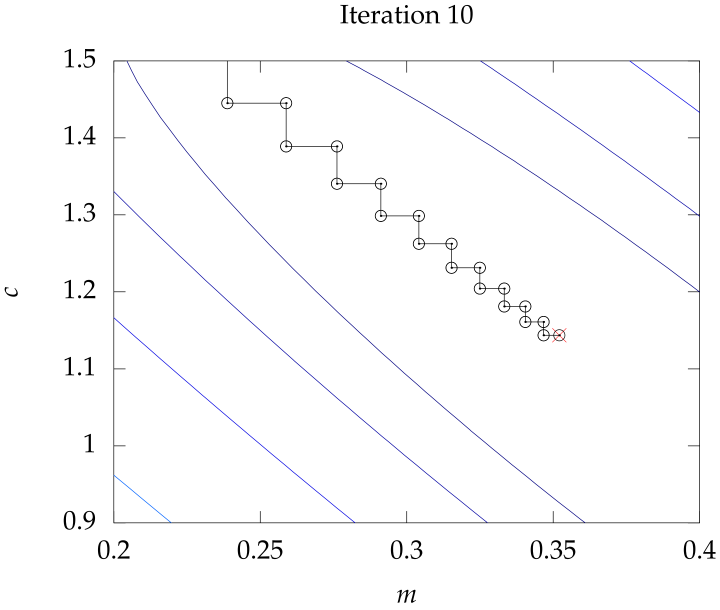
Coordinate Descent



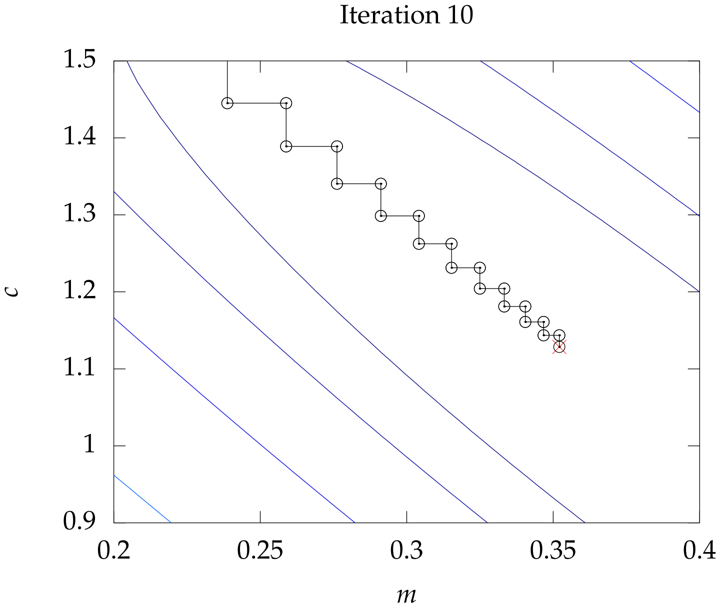
Coordinate Descent



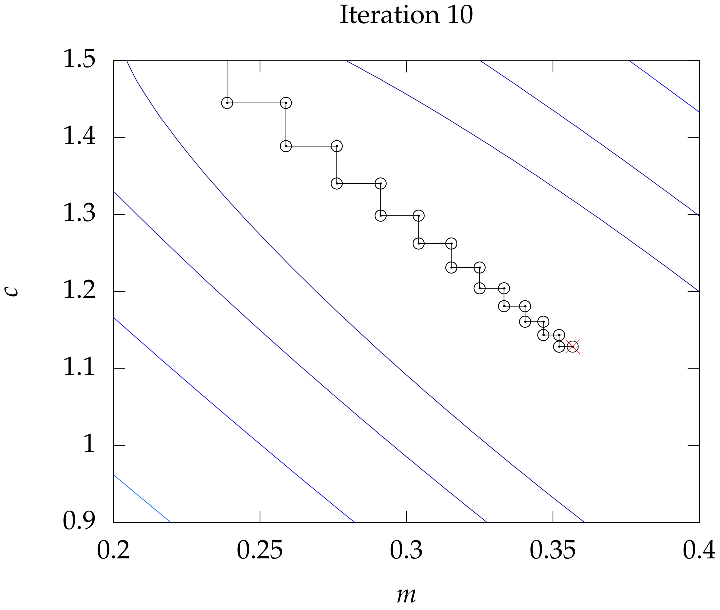
Coordinate Descent



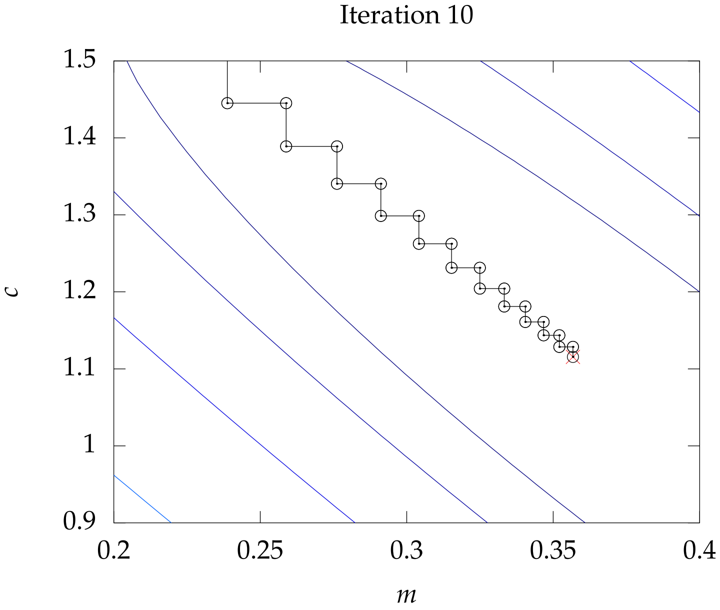
Coordinate Descent



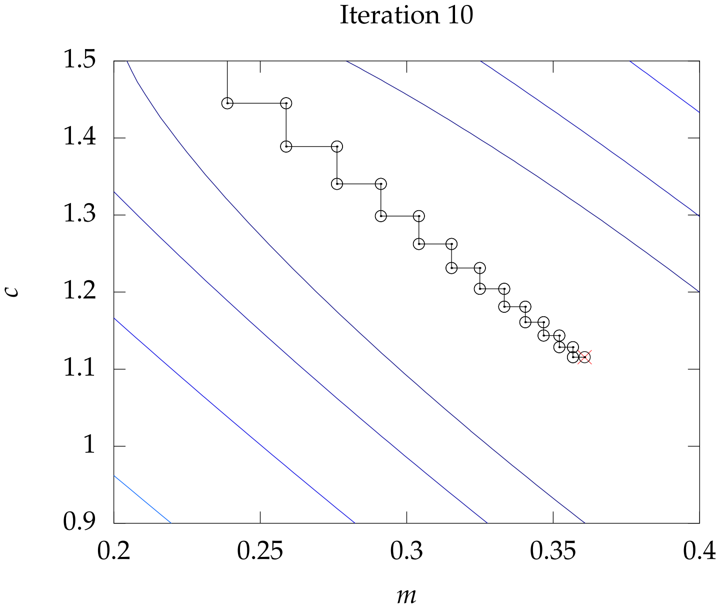
Coordinate Descent



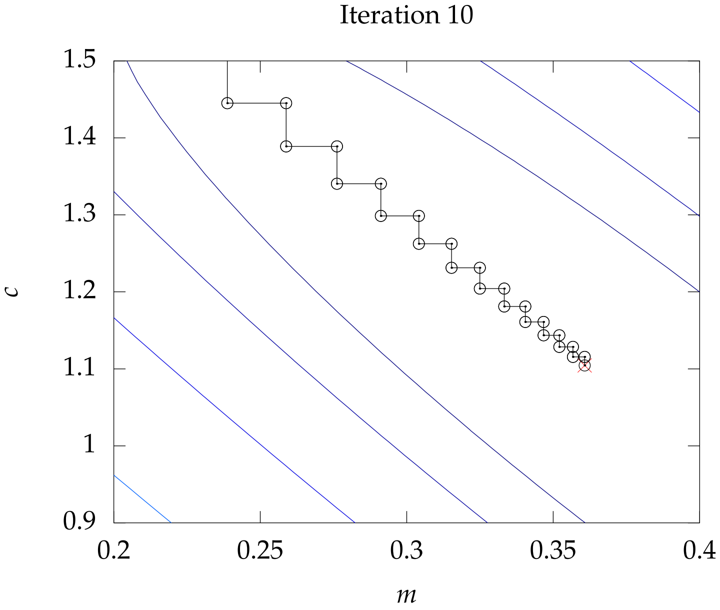
Coordinate Descent



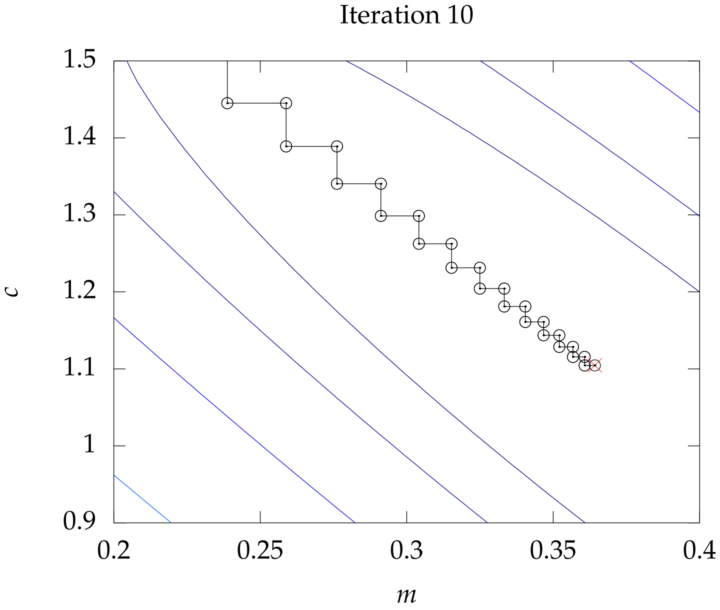
Coordinate Descent



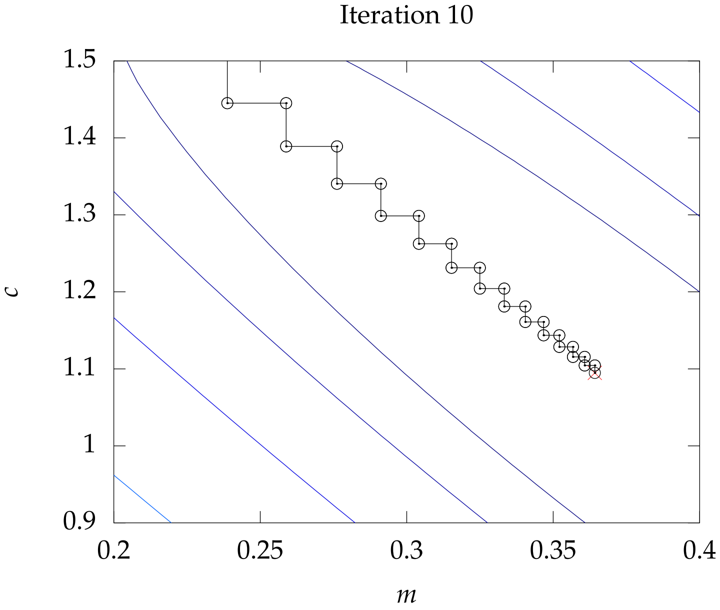
Coordinate Descent



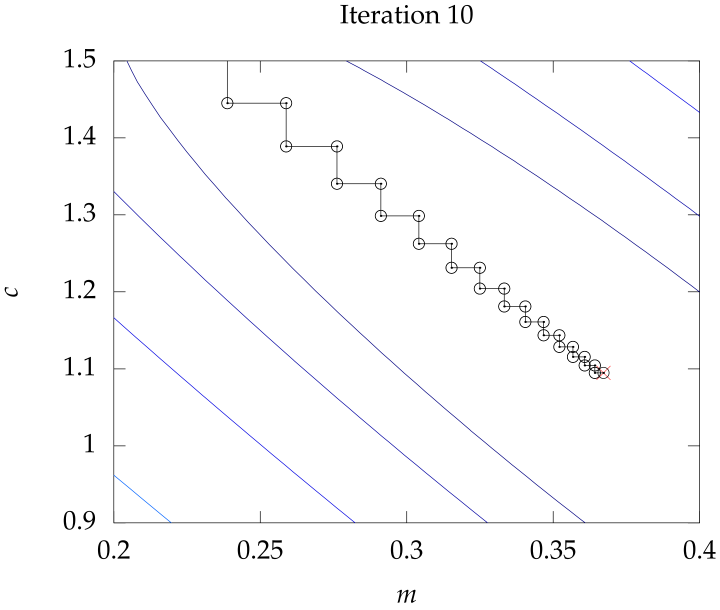
Coordinate Descent



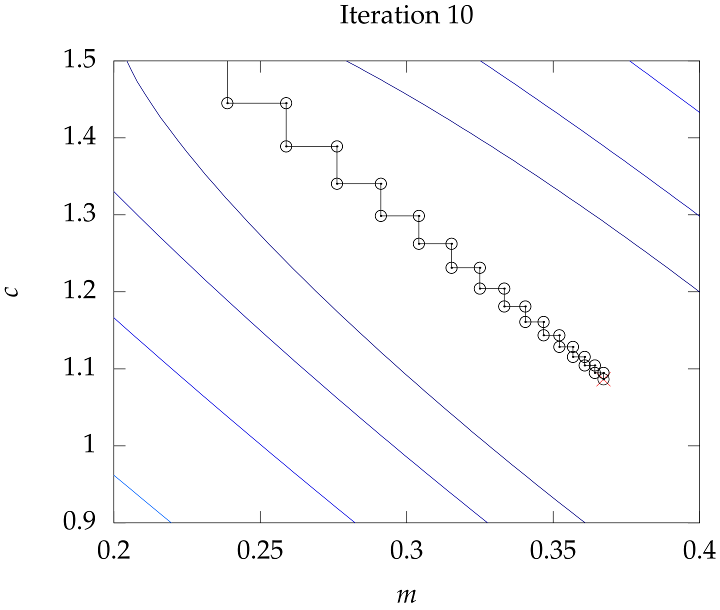
Coordinate Descent



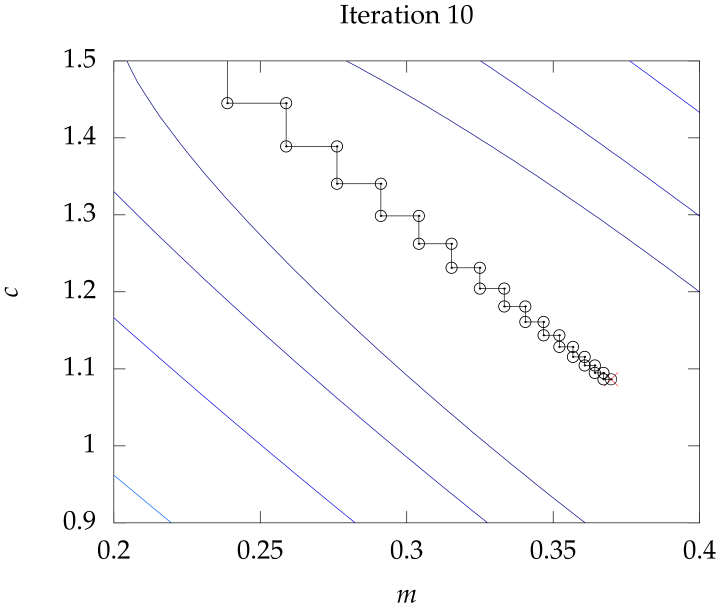
Coordinate Descent



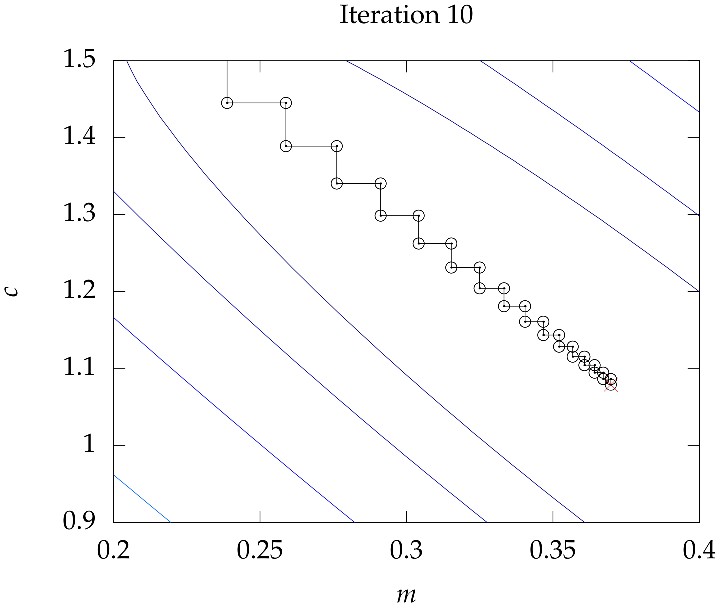
Coordinate Descent



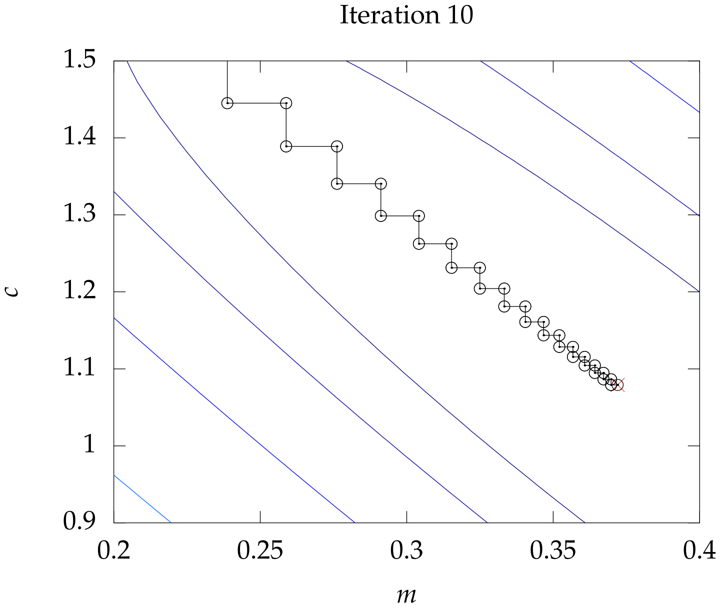
Coordinate Descent



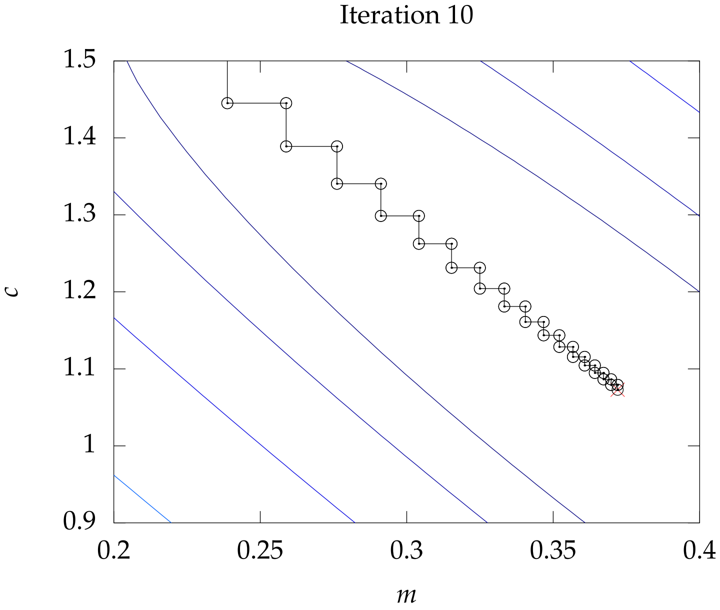
Coordinate Descent



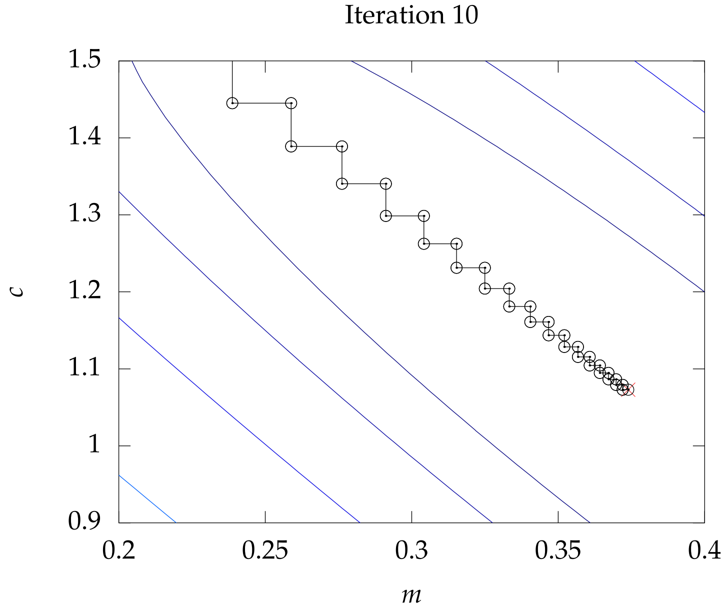
Coordinate Descent



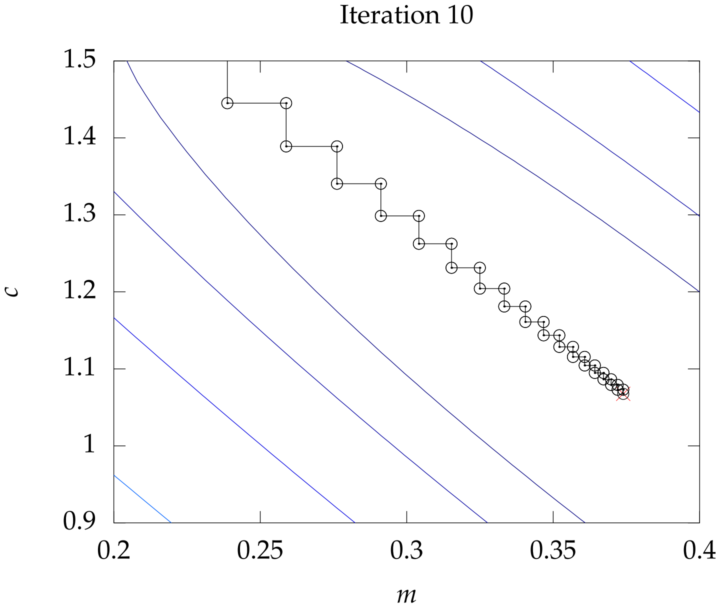
Coordinate Descent



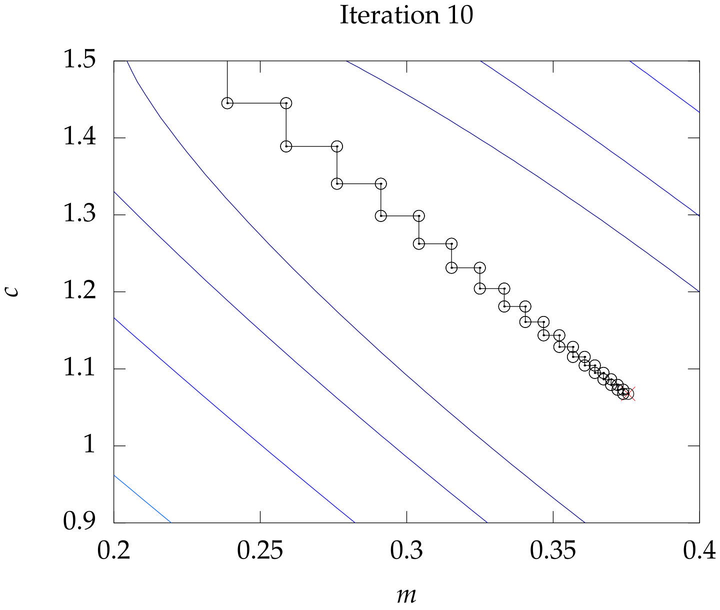
Coordinate Descent



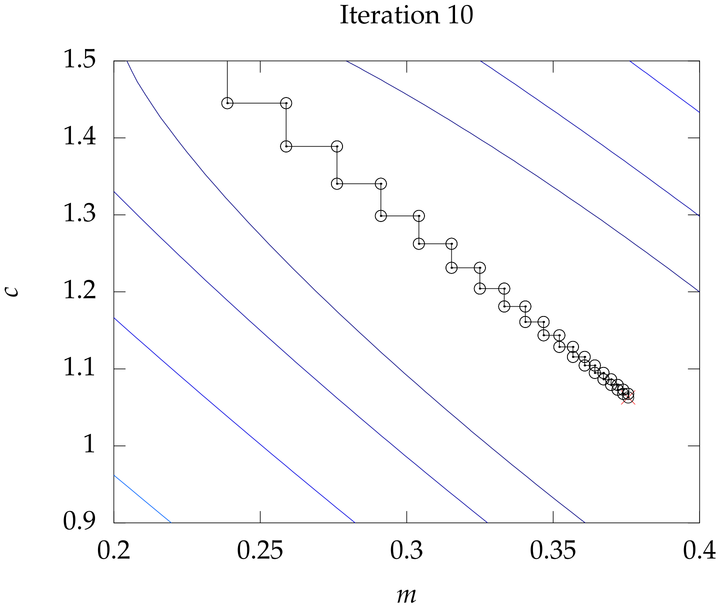
Coordinate Descent



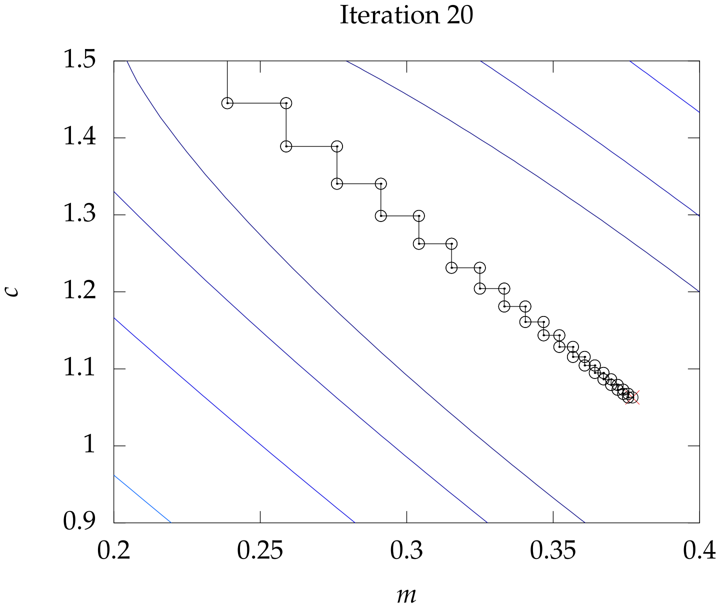
Coordinate Descent



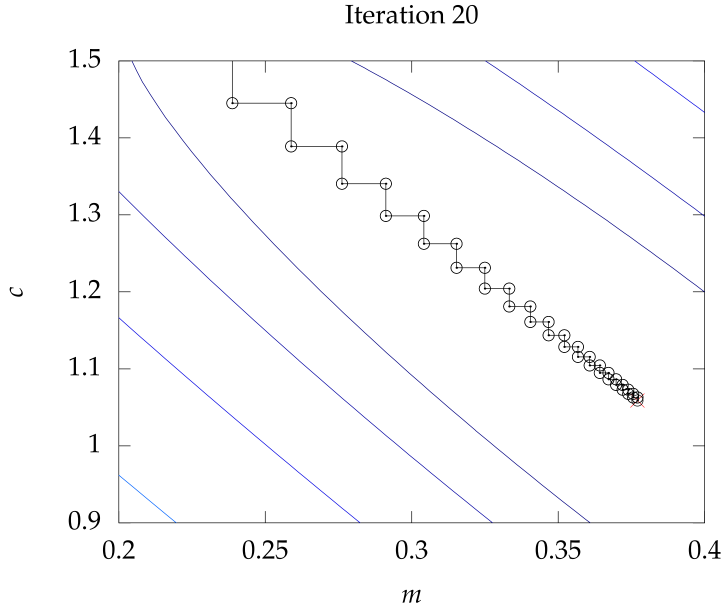
Coordinate Descent



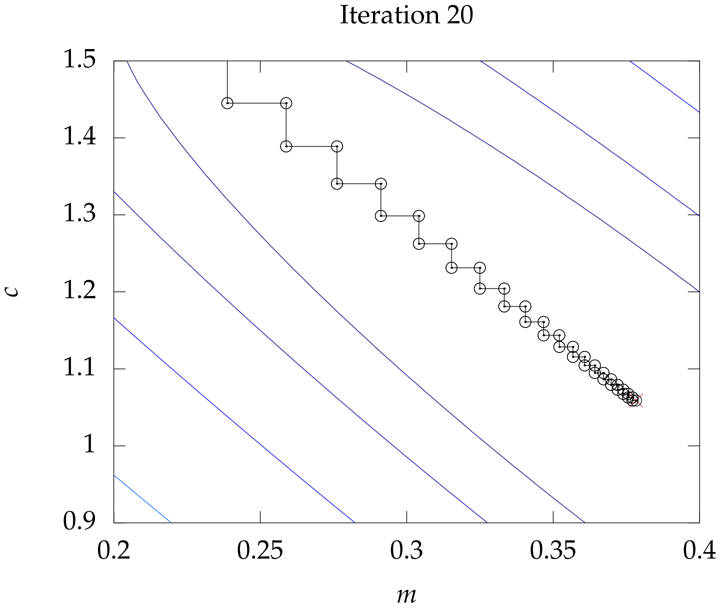
Coordinate Descent



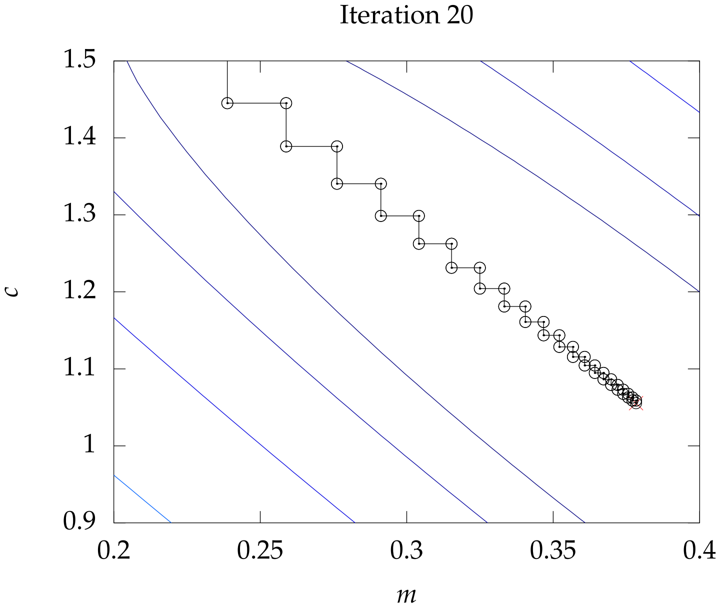
Coordinate Descent



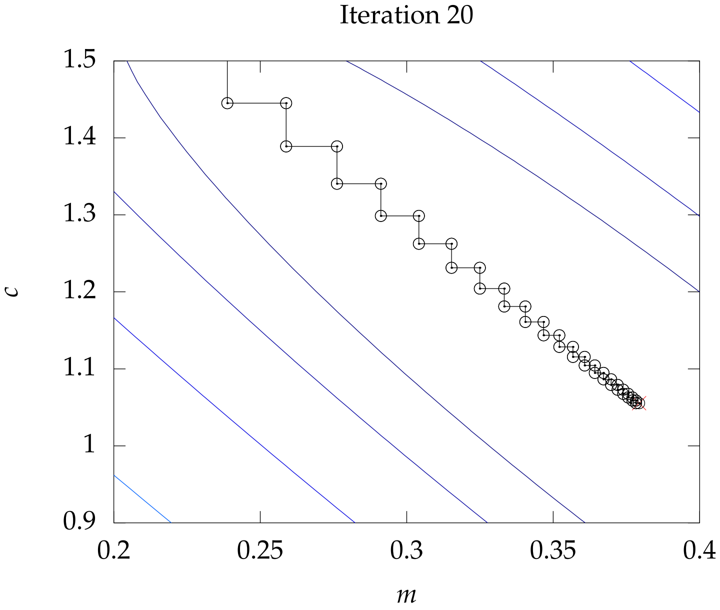
Coordinate Descent



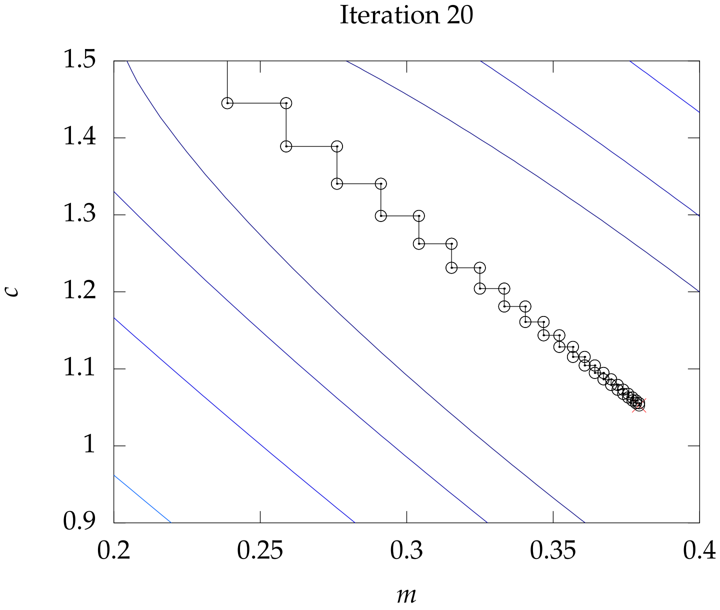
Coordinate Descent



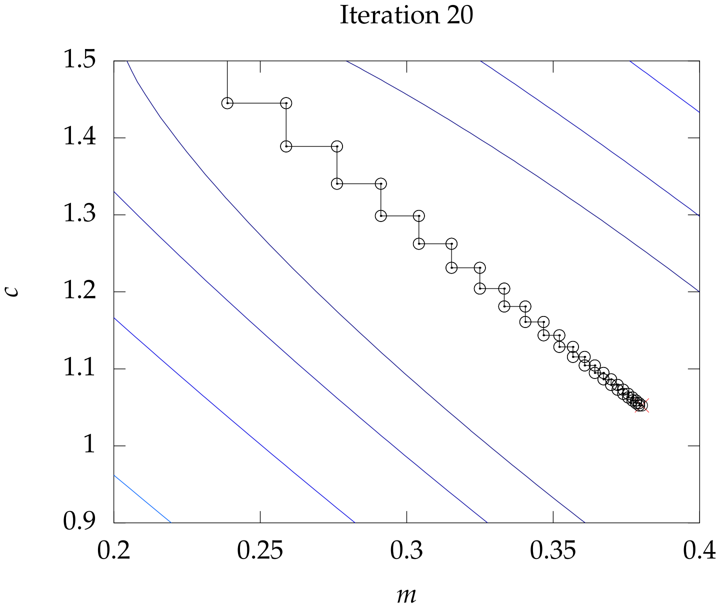
Coordinate Descent



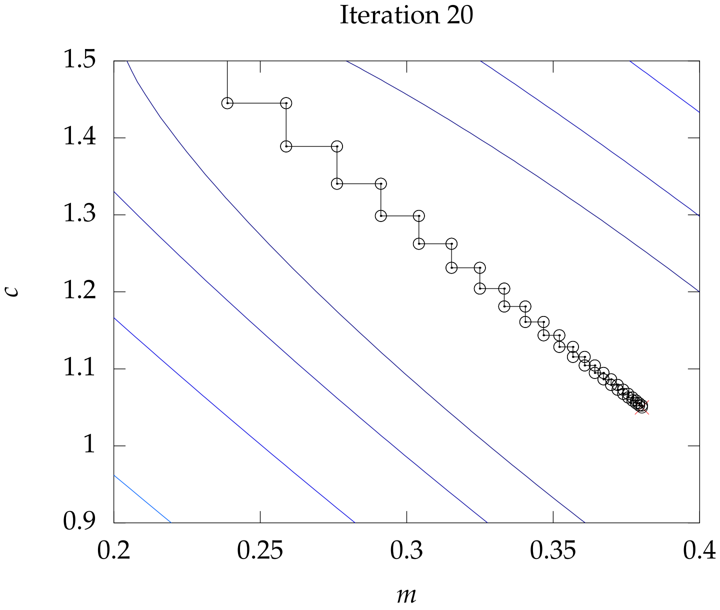
Coordinate Descent



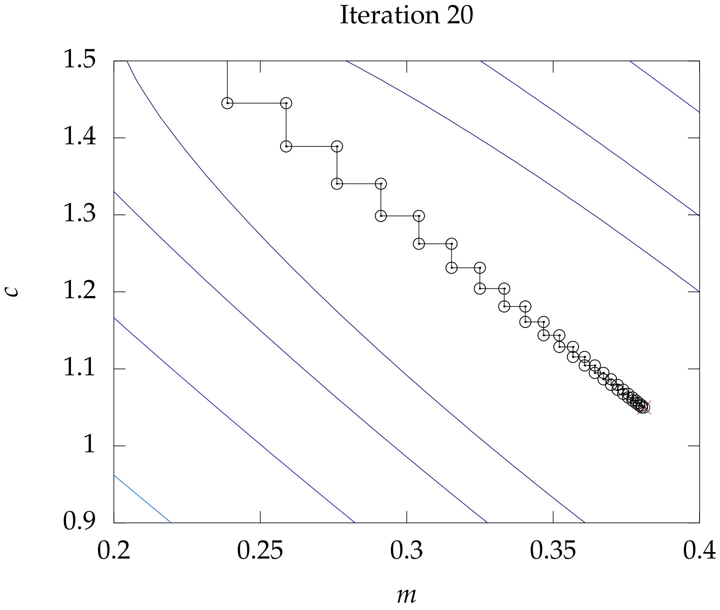
Coordinate Descent



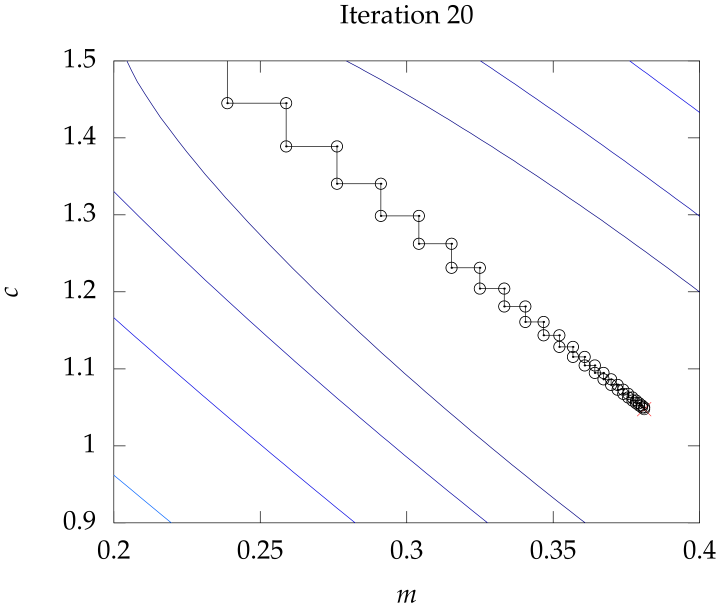
Coordinate Descent



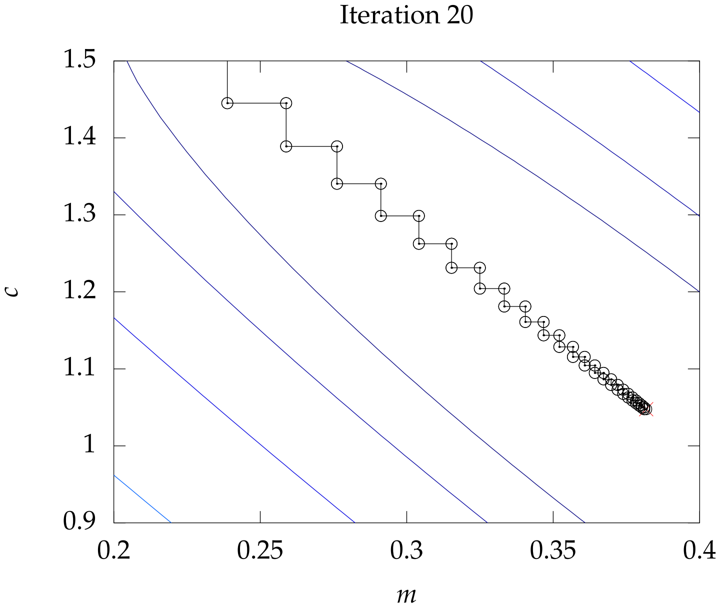
Coordinate Descent



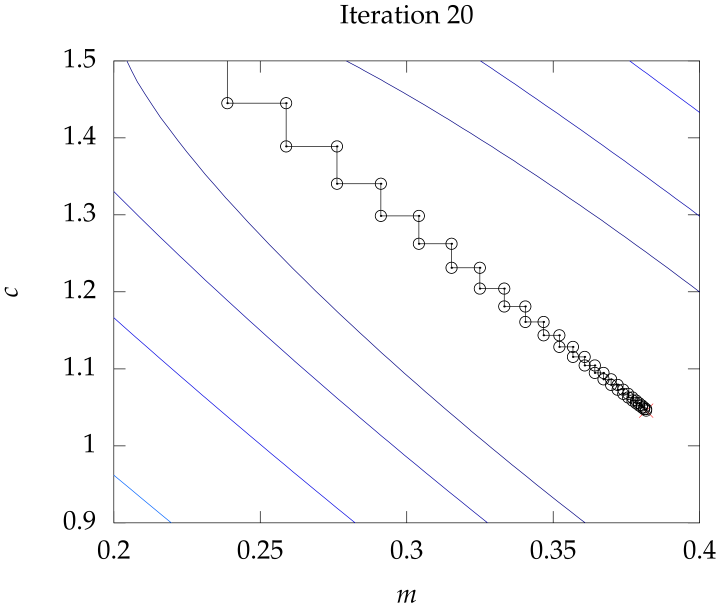
Coordinate Descent



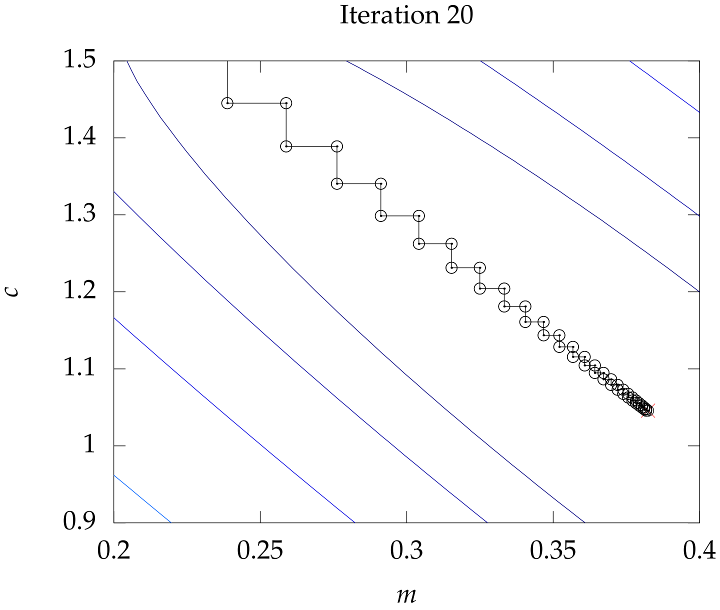
Coordinate Descent



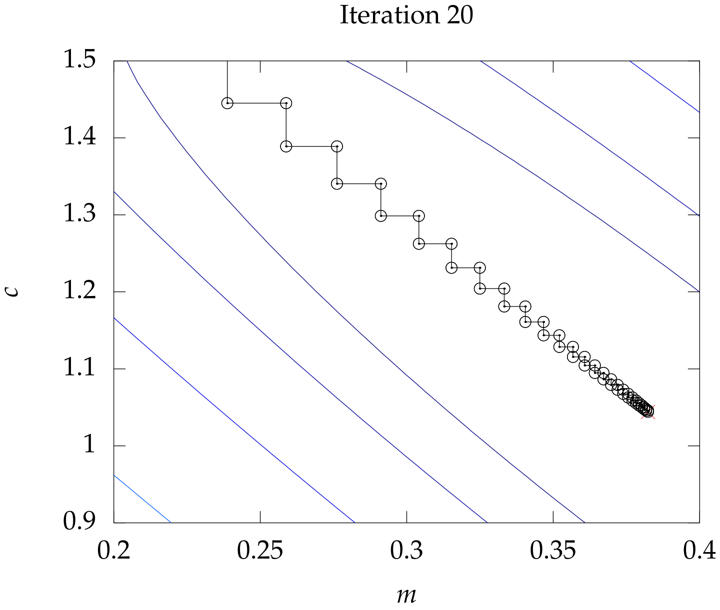
Coordinate Descent



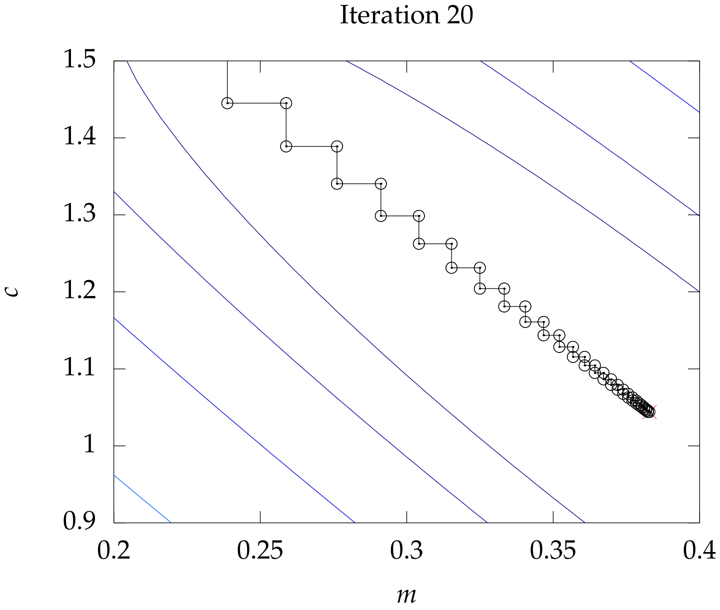
Coordinate Descent



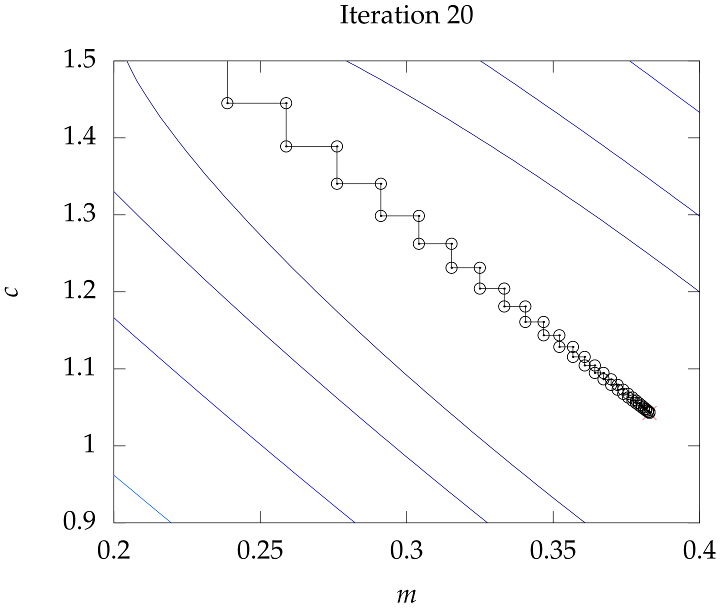
Coordinate Descent



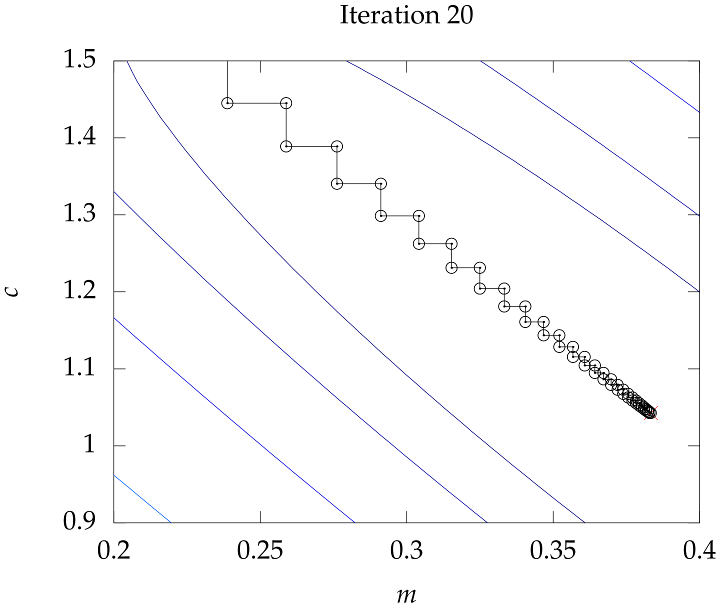
Coordinate Descent



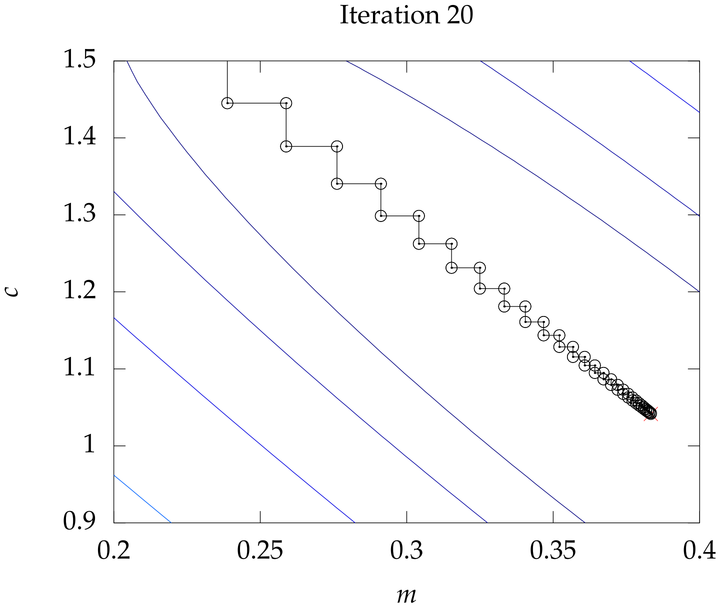
Coordinate Descent



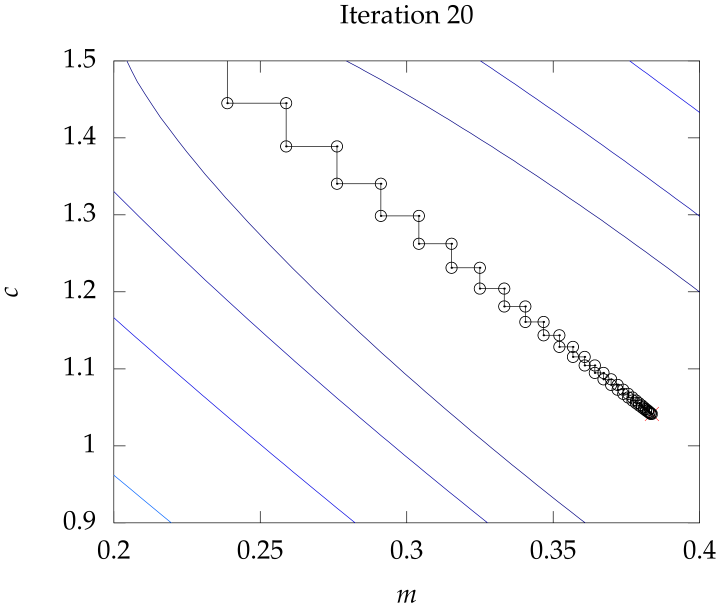
Coordinate Descent



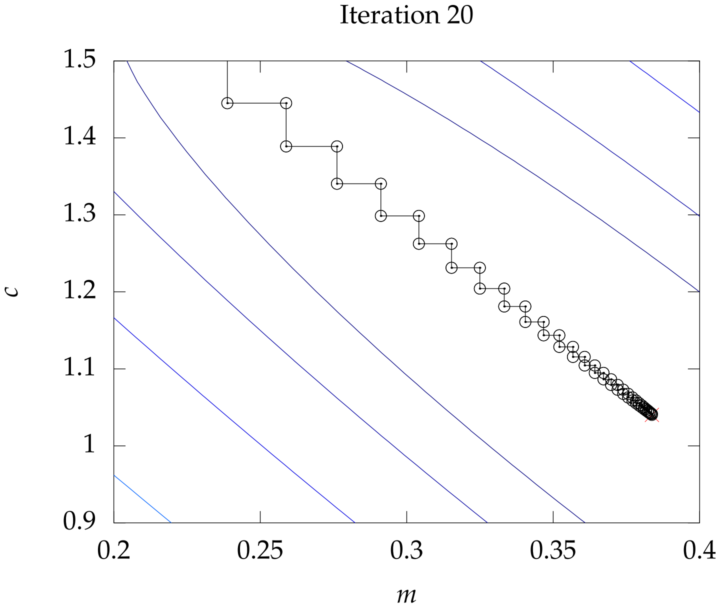
Coordinate Descent



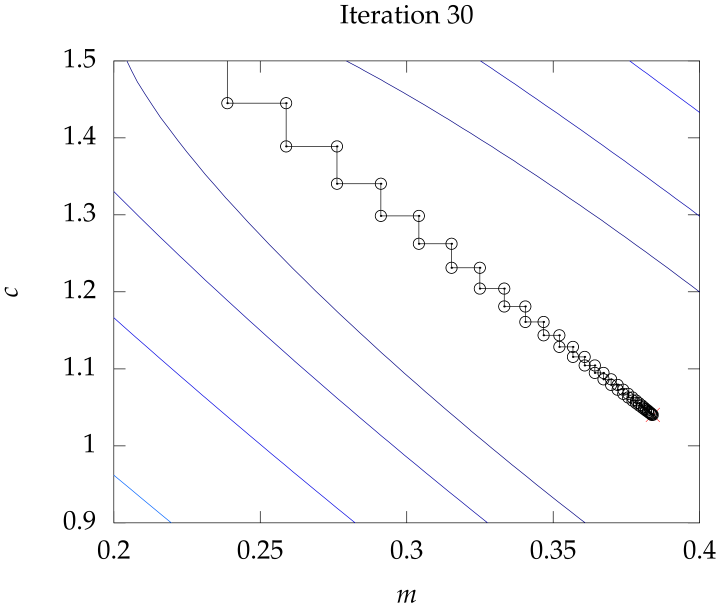
Coordinate Descent



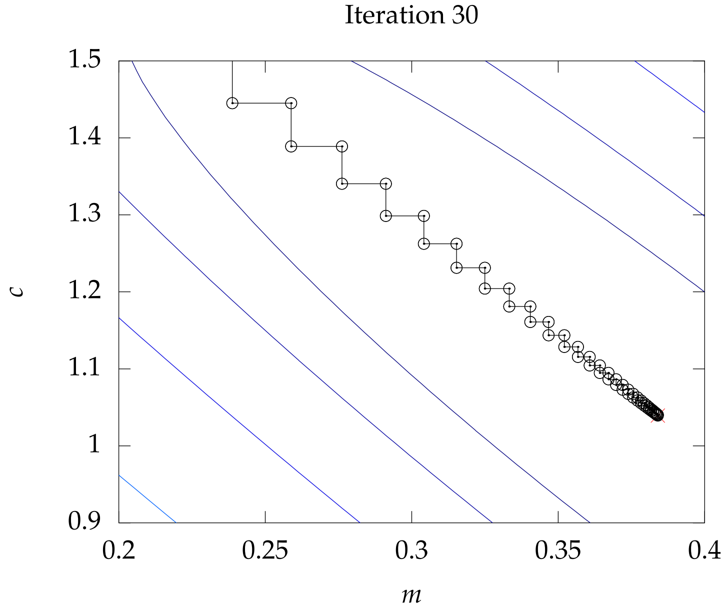
Coordinate Descent



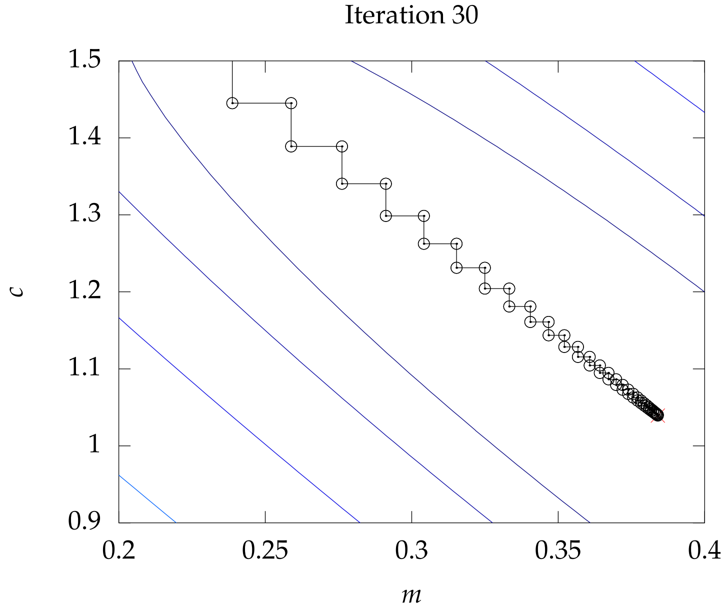
Coordinate Descent



Coordinate Descent



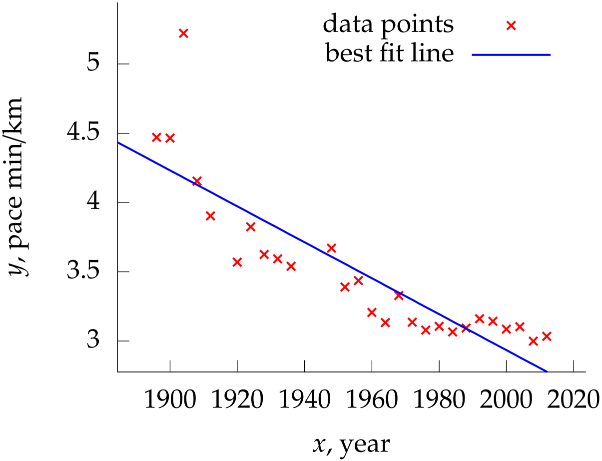
Coordinate Descent



Important Concepts Not Covered

- Optimization methods.
  - Second order methods, conjugate gradient, quasi-Newton and Newton.
  - Effective heuristics such as momentum.
- Local vs global solutions.

Linear Function



Linear regression for Male Olympics Marathon Gold Medal times.

Reading

- Section 1.2.5 of Bishop up to equation 1.65.
- Section 1.1-1.2 of Rogers and Girolami for fitting linear models.

References I

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [Google Books] .

P. S. Laplace. Mémoire sur la probabilité des causes par les évènemens. In *Mémoires de mathèmatique et de physique, presentés à lAcadémie Royale des Sciences, par divers savans, & lù dans ses assemblées 6*, pages 621–656, 1774. Translated in Stigler (1986).

S. Rogers and M. Girolami. *A First Course in Machine Learning*. CRC Press, 2011. [Google Books] .

S. M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1:359–378, 1986.