Fitting Neural Networks
Gradient Descent and Stochastic Gradient Descent

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner

Outline

- Gradient Descent
- Stochastic Gradient Descent

Considerations

- We still need to calculate the derivatives.
- We need to know what is the learning rate or how to set it.
- Local vs global minima.
- The full likelihood function includes summing up all individual '*errors*'. Unless you are a statistician, sometimes this includes hundreds of thousands of examples.

Considerations

- **We still need to calculate the derivatives.**
- We need to know what is the learning rate or how to set it.
- Local vs global minima.
- The full likelihood function includes summing up all individual ‘*errors*’. Unless you are a statistician, sometimes this includes hundreds of thousands of examples.

Calculate the Derivatives

Can we do it?

Wolfram Alpha can do it for us!

We need a formalism to deal with these derivatives.

Example: Logistic Regression derivatives

Chain Rule

Chain rule for computing gradients:

- $y = g(x) \qquad z = f(y) = f\big(g(x)\big)$

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$$

$$\boldsymbol{y} = g(\boldsymbol{x}) \qquad z = f(\boldsymbol{y}) = f\big(g(\boldsymbol{x})\big)$$

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j}\frac{\partial y_j}{\partial x_i}$$

- For longer chains:

$$\boxed{\frac{\partial z}{\partial x_i} = \sum_{j_1} \cdots \sum_{j_m} \frac{\partial z}{\partial y_{j_1}} \cdots \frac{\partial y_{j_m}}{\partial x_i}}$$

Logistic Regression derivatives

For logistic regression, the –ve log of the likelihood is:

$$\mathcal{L} = \sum_i \mathcal{L}_i = -\sum_i \log L_i = -\sum_i [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

$$\mathcal{L}_i = -y_i \log \frac{1}{1 + e^{-W^T X}} - (1 - y_i) \log(1 - \frac{1}{1 + e^{-W^T X}})$$

To simplify the analysis let us split it into two parts,

$$\mathcal{L}_i = \mathcal{L}_i^A + \mathcal{L}_i^B$$

So the derivative with respect to W is:

$$\frac{\partial \mathcal{L}}{\partial W} = \sum_i \frac{\partial \mathcal{L}_i}{\partial W} = \sum_i (\frac{\partial \mathcal{L}_i^A}{\partial W} + \frac{\partial \mathcal{L}_i^B}{\partial W})$$

$$\mathcal{L}_i^A = \boxed{-y_i}\,\boxed{\log}\,\boxed{\frac{1}{1+e^{-W^T X}}}$$

Variables	Partial derivatives	Partial derivatives
$\xi_1 = -W^T X$	$\frac{\partial \xi_1}{\partial W} = -X$	$\frac{\partial \xi_1}{\partial W} = -X$
$\xi_2 = e^{\xi_1} = e^{-W^T X}$	$\frac{\partial \xi_2}{\partial \xi_1} = e^{\xi_1}$	$\frac{\partial \xi_2}{\partial \xi_1} = e^{-W^T X}$
$\xi_3 = 1 + \xi_2 = 1 + e^{-W^T X}$	$\frac{\partial \xi_3}{\partial \xi_2} = 1$	$\frac{\partial \xi_3}{\partial \xi_2} = 1$
$\xi_4 = \frac{1}{\xi_3} = \frac{1}{1+e^{-W^T X}} = p$	$\frac{\partial \xi_4}{\partial \xi_3} = -\frac{1}{\xi_3^2}$	$\frac{\partial \xi_4}{\partial \xi_3} = -\frac{1}{\left(1+e^{-W^T X}\right)^2}$
$\xi_5 = \log \xi_4 = \log p = \log \frac{1}{1+e^{-W^T X}}$	$\frac{\partial \xi_5}{\partial \xi_4} = \frac{1}{\xi_4}$	$\frac{\partial \xi_5}{\partial \xi_4} = 1 + e^{-W^T X}$
$\mathcal{L}_i^A = -y\xi_5$	$\frac{\partial \mathcal{L}}{\partial \xi_5} = -y$	$\frac{\partial \mathcal{L}}{\partial \xi_5} = -y$
$\frac{\partial \mathcal{L}_i^A}{\partial W} = \frac{\partial \mathcal{L}_i}{\partial \xi_5}\frac{\partial \xi_5}{\partial \xi_4}\frac{\partial \xi_4}{\partial \xi_3}\frac{\partial \xi_3}{\partial \xi_2}\frac{\partial \xi_2}{\partial \xi_1}\frac{\partial \xi_1}{\partial W}$		$\frac{\partial \mathcal{L}_i^A}{\partial W} = -yXe^{-W^T X}\frac{1}{\left(1+e^{-W^T X}\right)}$

$$\mathcal{L}_i^B = -(1-y_i)\log[1 - \frac{1}{1+e^{-W^TX}}]$$

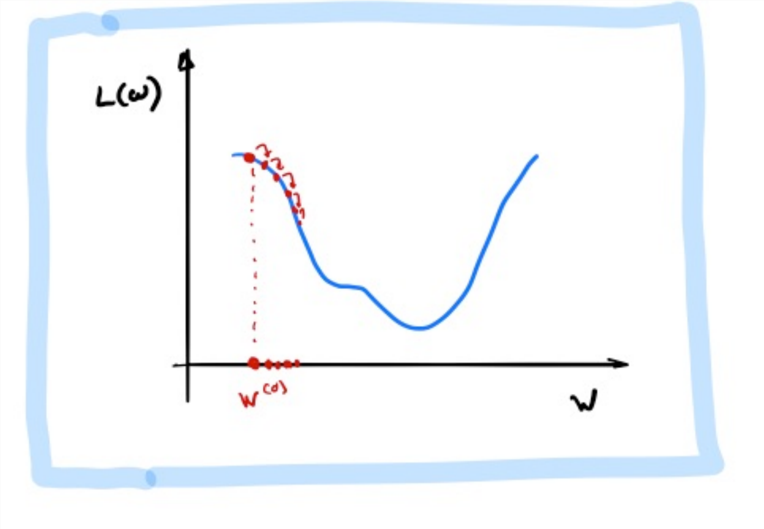
Variables	derivatives	Partial derivatives wrt to X,W
$\xi_1 = -W^TX$	$\frac{\partial\xi_1}{\partial W} = -X$	$\frac{\partial\xi_1}{\partial W} = -X$
$\xi_2 = e^{\xi_1} = e^{-W^TX}$	$\frac{\partial\xi_2}{\partial\xi_1} = e^{\xi_1}$	$\frac{\partial\xi_2}{\partial\xi_1} = e^{-W^TX}$
$\xi_3 = 1+\xi_2 = 1+e^{-W^TX}$	$\frac{\partial\xi_3}{\partial\xi_2} = 1$	$\frac{\partial\xi_3}{\partial 2} = 1$
$\xi_4 = \frac{1}{\xi_3} = \frac{1}{1+e^{-W^TX}} = p$	$\frac{\partial\xi_4}{\partial\xi_3} = -\frac{1}{\xi_3^2}$	$\frac{\partial\xi_4}{\partial\xi_3} = -\frac{1}{\left(1+e^{-W^TX}\right)^2}$
$\xi_5 = 1-\xi_4 = 1-\frac{1}{1+e^{-W^TX}}$	$\frac{\partial\xi_5}{\partial\xi_4} = -1$	$\frac{\partial\xi_5}{\partial\xi_4} = -1$
$\xi_6 = \log\xi_5 = \log(1-p) = \log\frac{1}{1+e^{-W^TX}}$	$\frac{\partial\xi_6}{\partial\xi_5} = \frac{1}{\xi_5}$	$\frac{\partial\xi_6}{\partial\xi_5} = \frac{1+e^{-W^TX}}{e^{-W^TX}}$
$\mathcal{L}_i^B = (1-y)\xi_6$	$\frac{\partial\mathcal{L}}{\partial\xi_6} = 1-y$	$\frac{\partial\mathcal{L}}{\partial\xi_6} = 1-y$
$\frac{\partial\mathcal{L}_i^B}{\partial W} = \frac{\partial\mathcal{L}_i^B}{\partial\xi_6}\frac{\partial\xi_6}{\partial\xi_5}\frac{\partial\xi_5}{\partial\xi_4}\frac{\partial\xi_4}{\partial\xi_3}\frac{\partial\xi_3}{\partial\xi_2}\frac{\partial\xi_2}{\partial\xi_1}\frac{\partial\xi_1}{\partial W}$		$\frac{\partial\mathcal{L}_i^B}{\partial W} = (1-y)X\frac{1}{\left(1+e^{-W^TX}\right)}$

Considerations

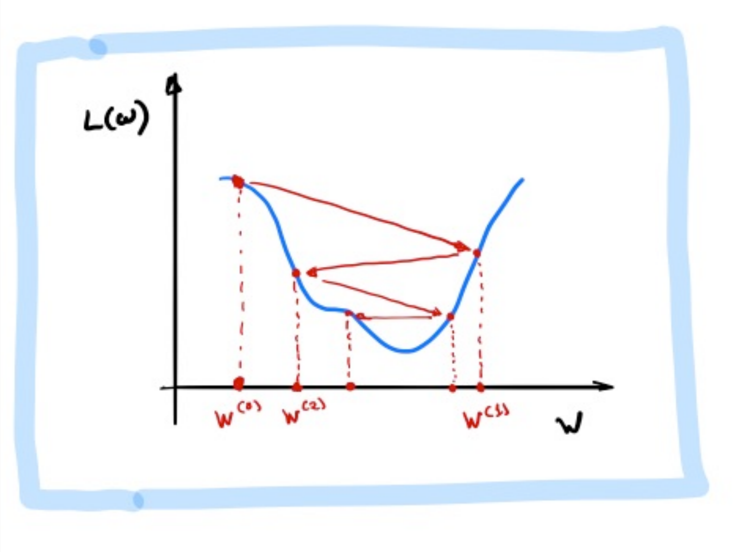
- We still need to calculate the derivatives.
- **We need to know what is the learning rate or how to set it.**
- Local vs global minima.
- The full likelihood function includes summing up all individual ‘*errors*’. Unless you are a statistician, sometimes this includes hundreds of thousands of examples.

Learning Rate

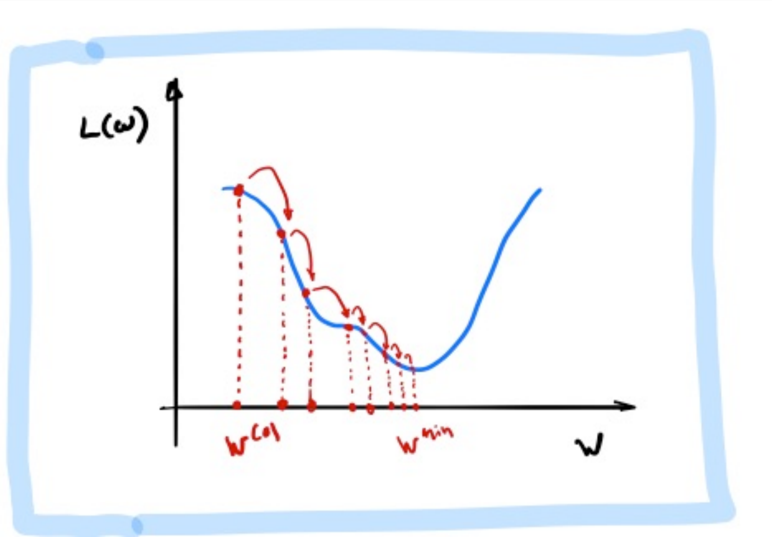
Our choice of the learning rate has a significant impact on the performance of gradient descent.



When η is too small, the algorithm makes very little progress.

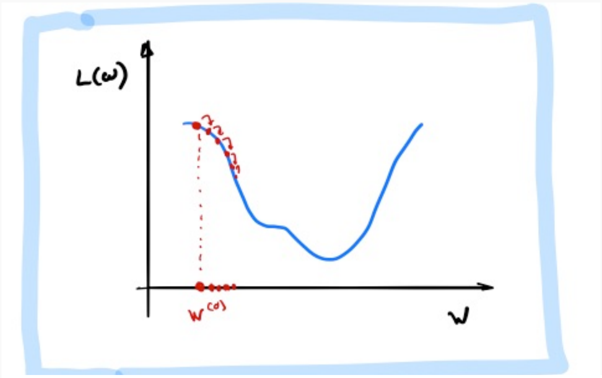


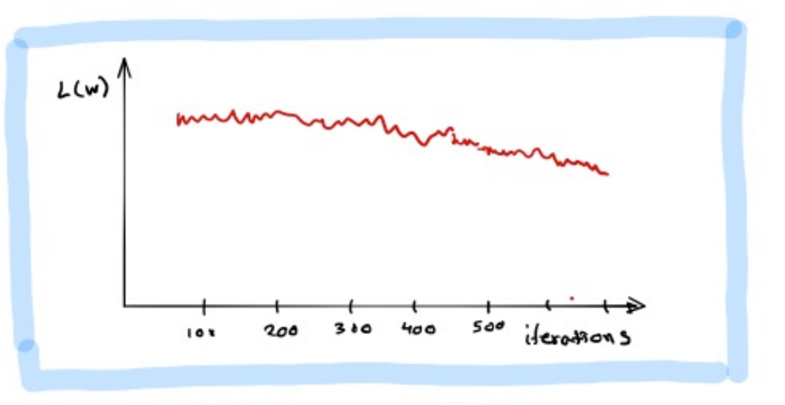
When η is too large, the algorithm may overshoot the minimum and has crazy oscillations.



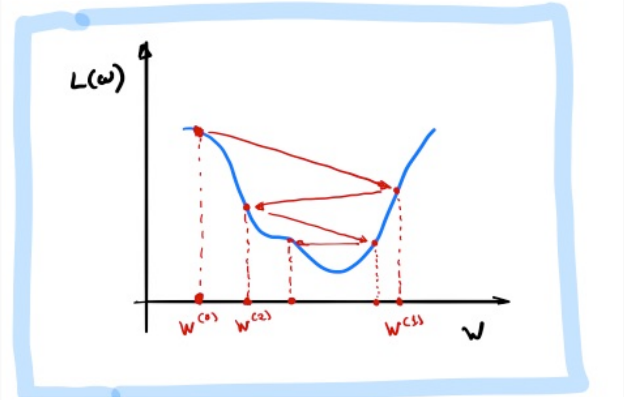
When η is appropriate, the algorithm will find the minimum. The algorithm **converges**!

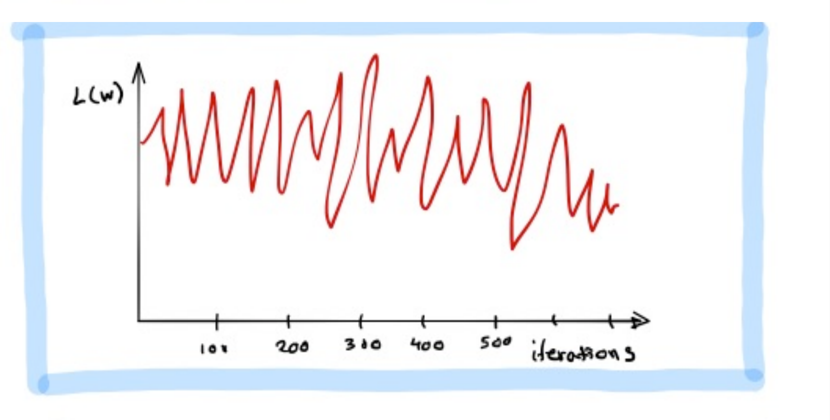
How can we tell when gradient descent is converging? We visualize the loss function at each step of gradient descent. This is called the **trace plot**.



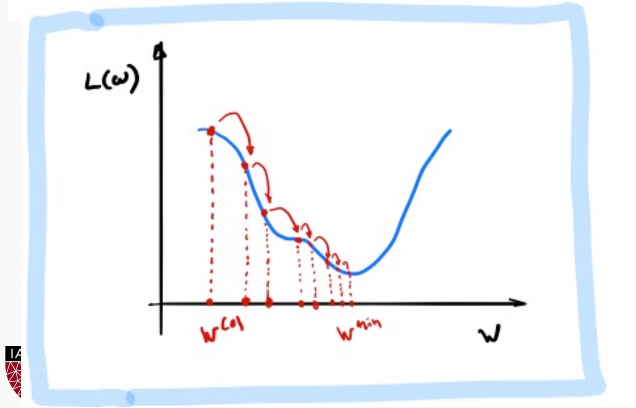


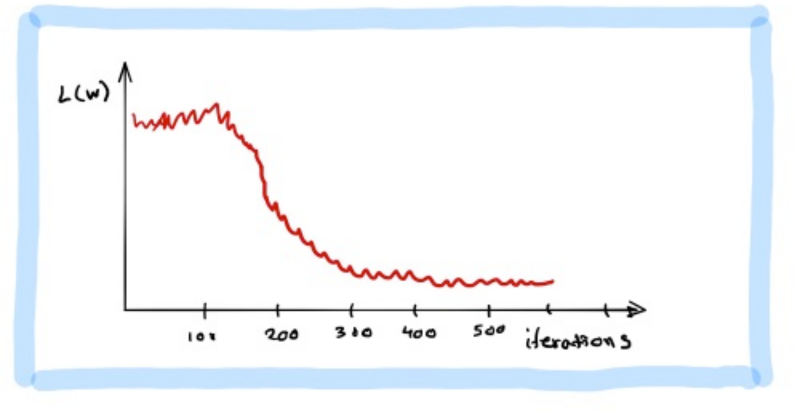
While the loss is decreasing throughout training, it does not look like descent hit the bottom.





Loss is mostly oscillating between values rather than converging.





The loss has decreased significantly during training. Towards the end, the loss stabilizes and it can’t decrease further.

Learning Rate

There are many alternative methods which address how to set or adjust the learning rate, using the derivative or second derivatives and or the momentum.

More on this later.

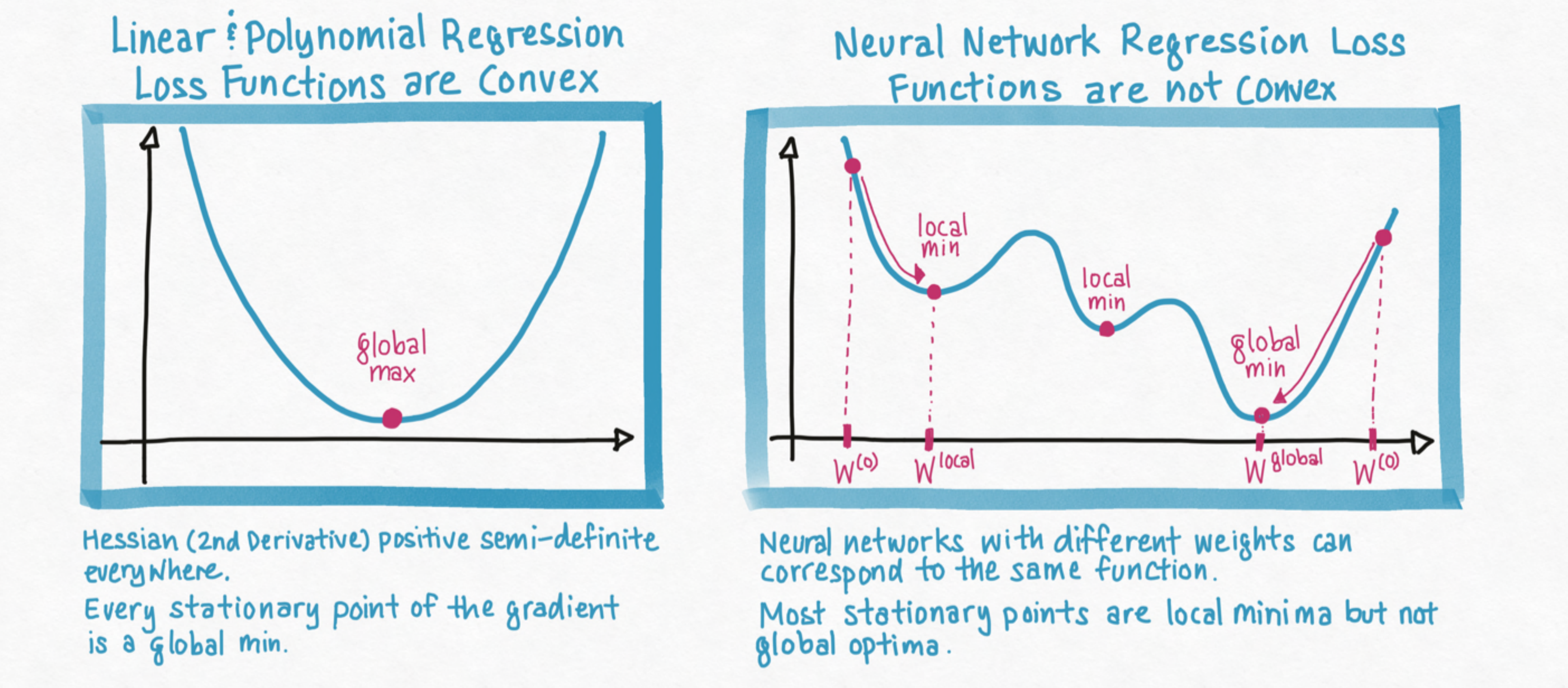
Considerations

- We still need to calculate the derivatives.
- We need to know what is the learning rate or how to set it.
- **Local vs global minima.**
- The full likelihood function includes summing up all individual '*errors*'. Unless you are a statistician, sometimes this includes hundreds of thousands of examples.

Local vs Global Minima

If we choose η correctly, then gradient descent will converge to a stationary point. But will this point be a **global minimum**?

If the function is convex then the stationary point will be a global minimum.



Local vs Global Minima

No guarantee that we get the global minimum.

Question: What would be a good strategy?

- Random restarts
- Add noise to the loss function

Considerations

- We still need to calculate the derivatives.
- We need to know what is the learning rate or how to set it.
- Local vs global minima.
- **The full likelihood function includes summing up all individual '*errors*'. Unless you are a statistician, sometimes this includes hundreds of thousands of examples.**

Batch and Stochastic Gradient Descent

$$\boxed{\mathcal{L} = -\sum_{i}[y_i \log p_i + (1 - y_i)\log(1 - p_i)]}$$

Instead of using all the examples for every step, use a subset of them (batch).

For each iteration k, use the following loss function to derive the derivatives:

$$\boxed{\mathcal{L}^k = -\sum_{i \in b^k}[y_i \log p_i + (1 - y_i)\log(1 - p_i)]}$$

which is an **approximation** to the full loss function.

DATA

BATCH 1	BATCH 2	BATCH 3	BATCH 4	... BATCH B

calculate $L_1 \Rightarrow \frac{\partial L_1}{\partial w} \Rightarrow w^* = w - \eta \frac{\partial L_1}{\partial w}$

calculate $L_2 \Rightarrow \frac{\partial L_2}{\partial w} \Rightarrow w^* = w - \eta \frac{\partial L_2}{\partial w}$

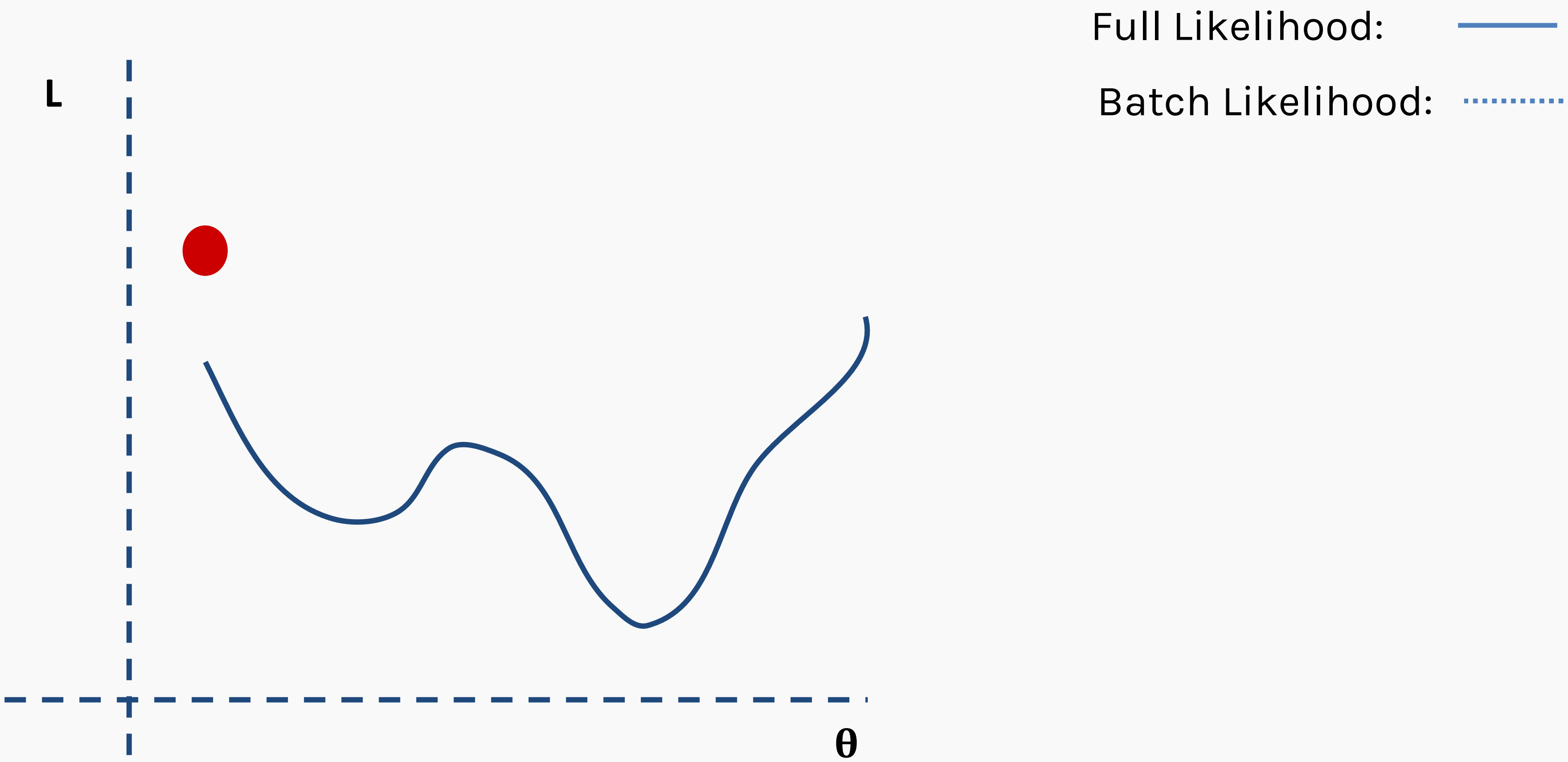
$\vdots$

$L_B \Rightarrow \frac{\partial L_B}{\partial w} \Rightarrow w^* = w - \eta \frac{\partial L_B}{\partial w}$

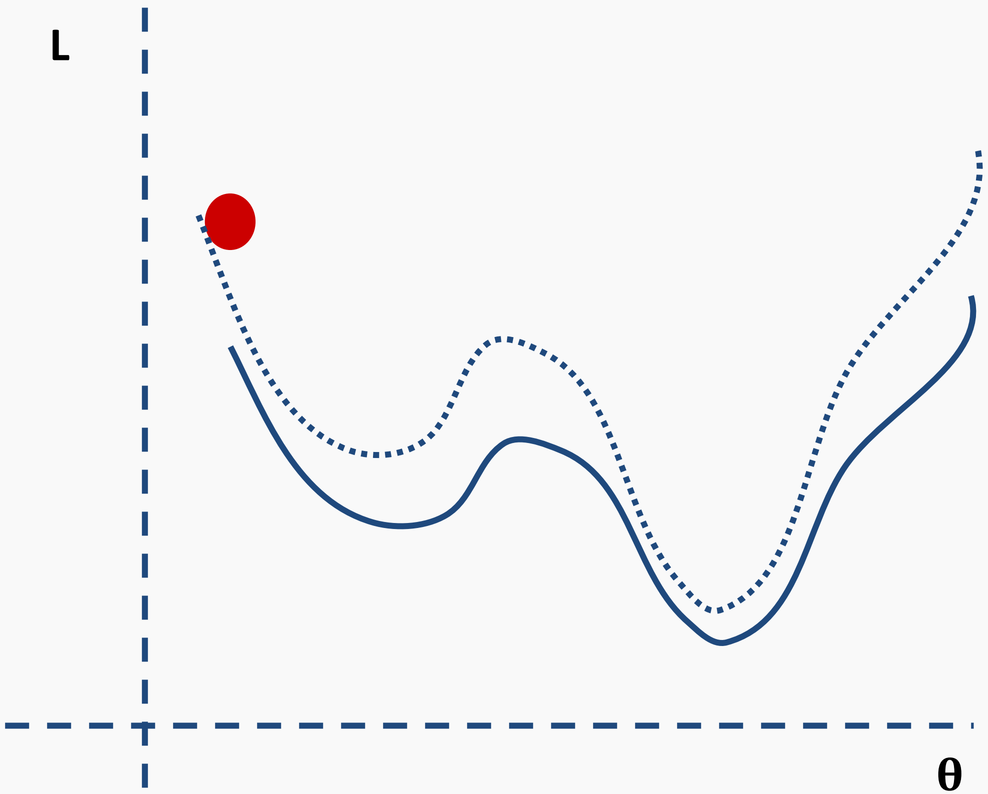
COMPLETE DATA $\Rightarrow$ ONE EPOCH

RESHUFFLE DATA AND REPEAT

Batch and Stochastic Gradient Descent



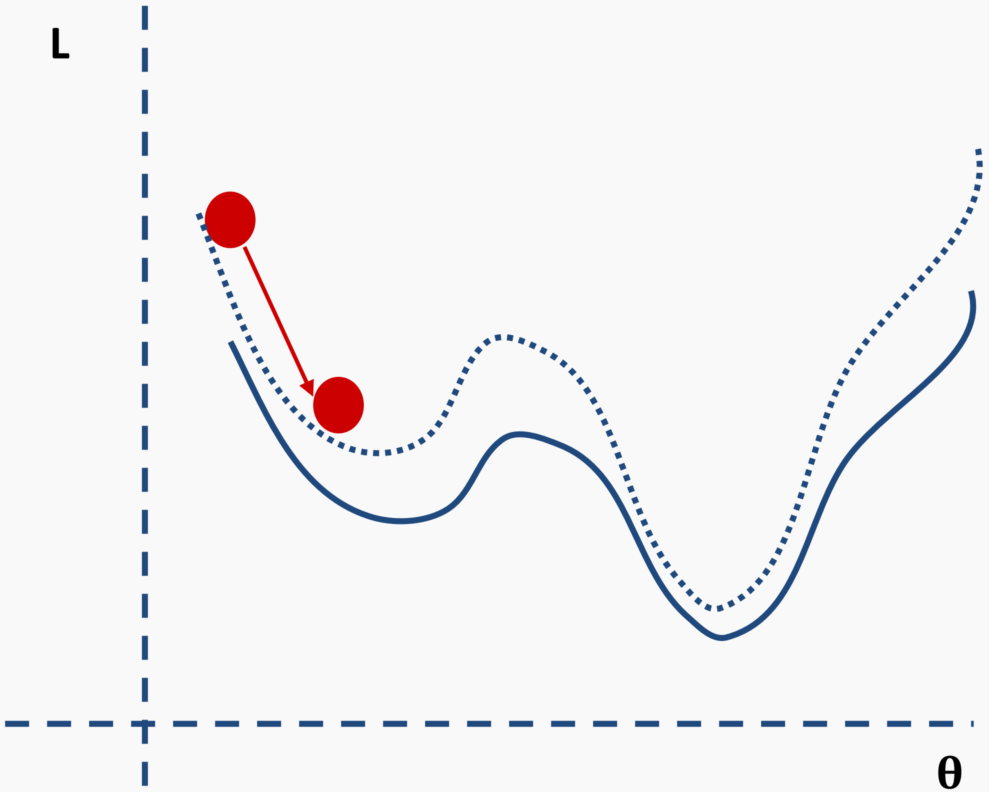
Batch and Stochastic Gradient Descent



Full Likelihood:

Batch Likelihood:

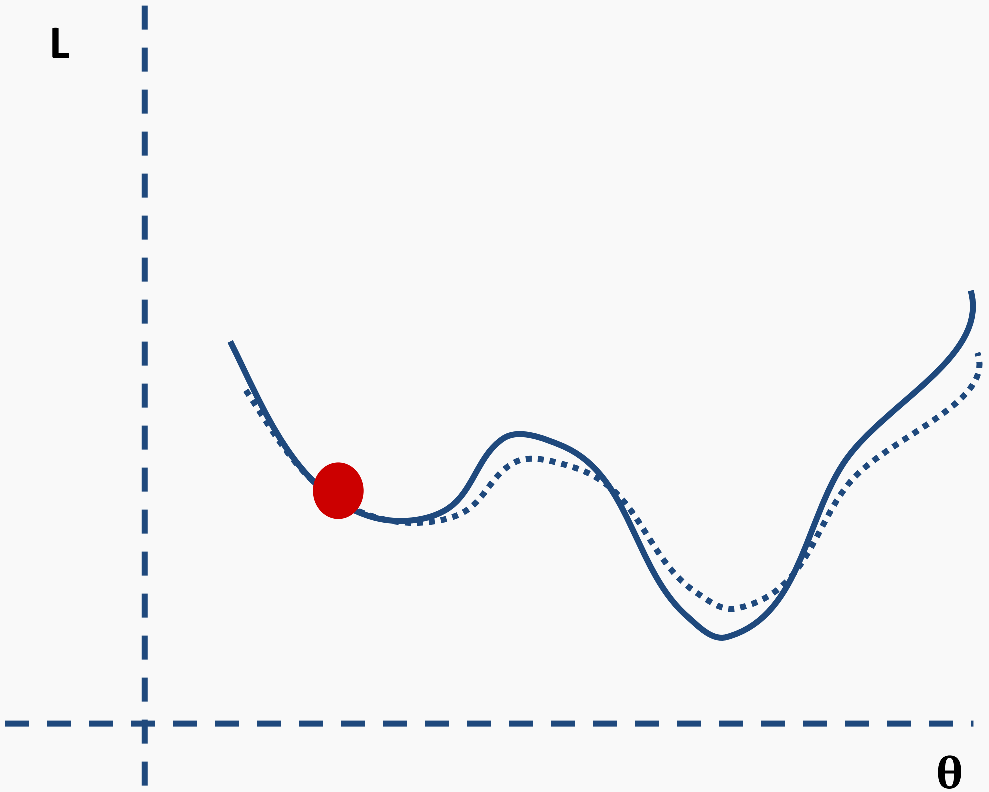
Batch and Stochastic Gradient Descent



Full Likelihood:

Batch Likelihood:

Batch and Stochastic Gradient Descent



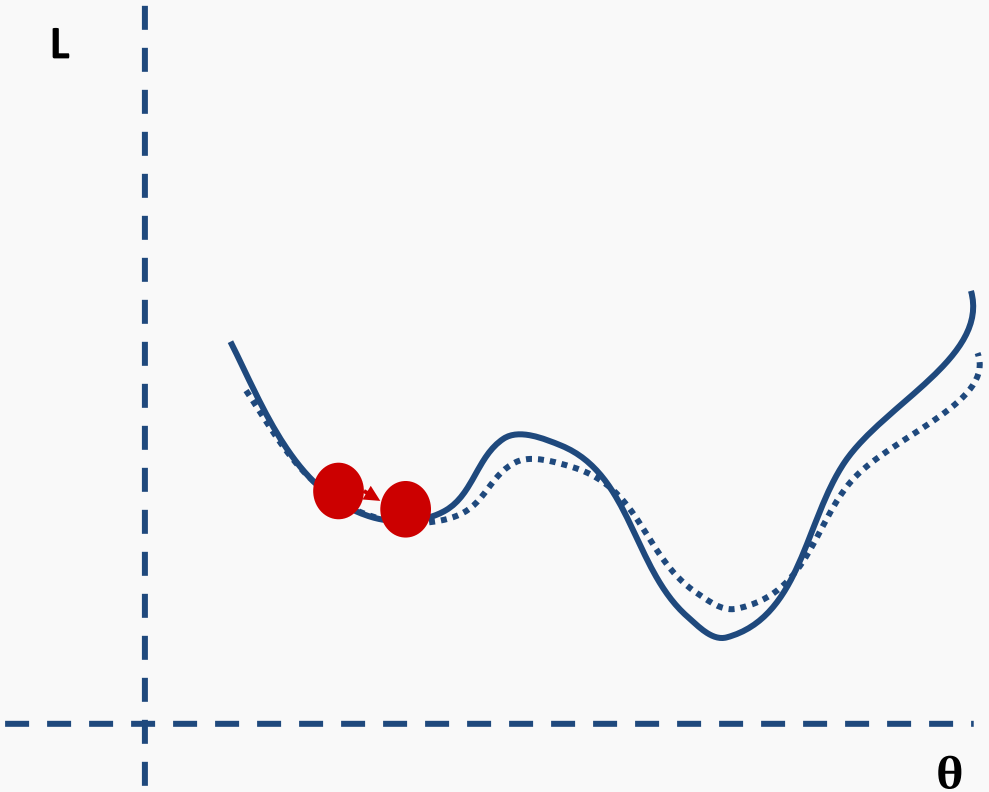
Full Likelihood:

Batch Likelihood:

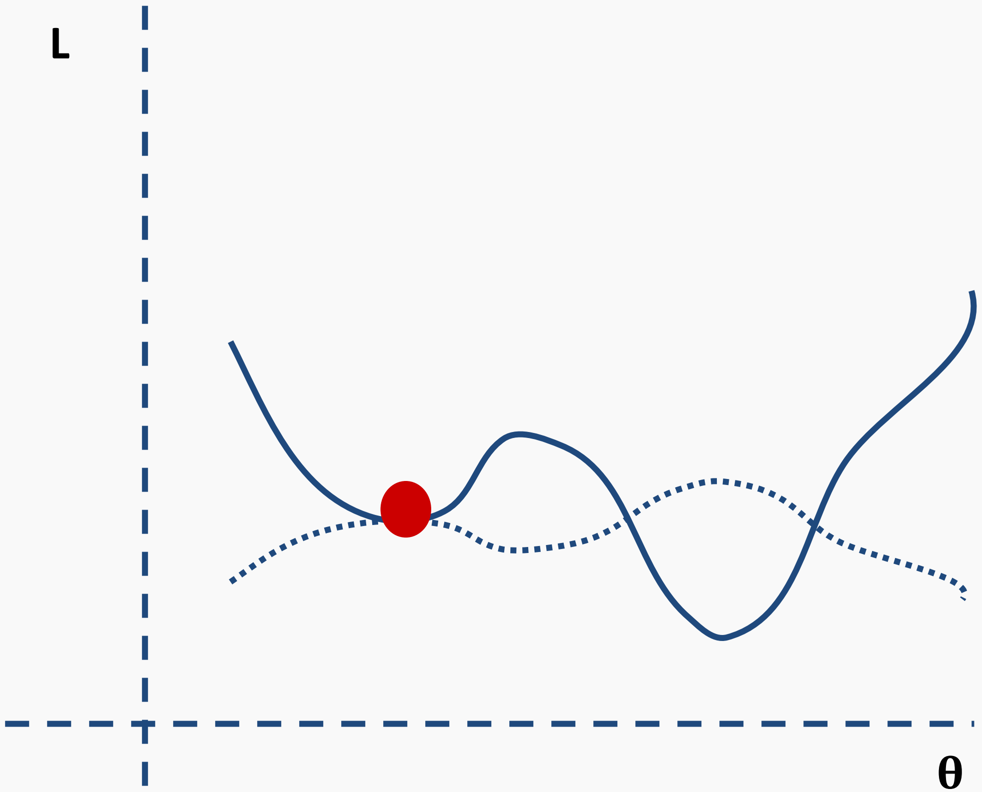
Batch and Stochastic Gradient Descent

Full Likelihood: ———

Batch Likelihood: ············

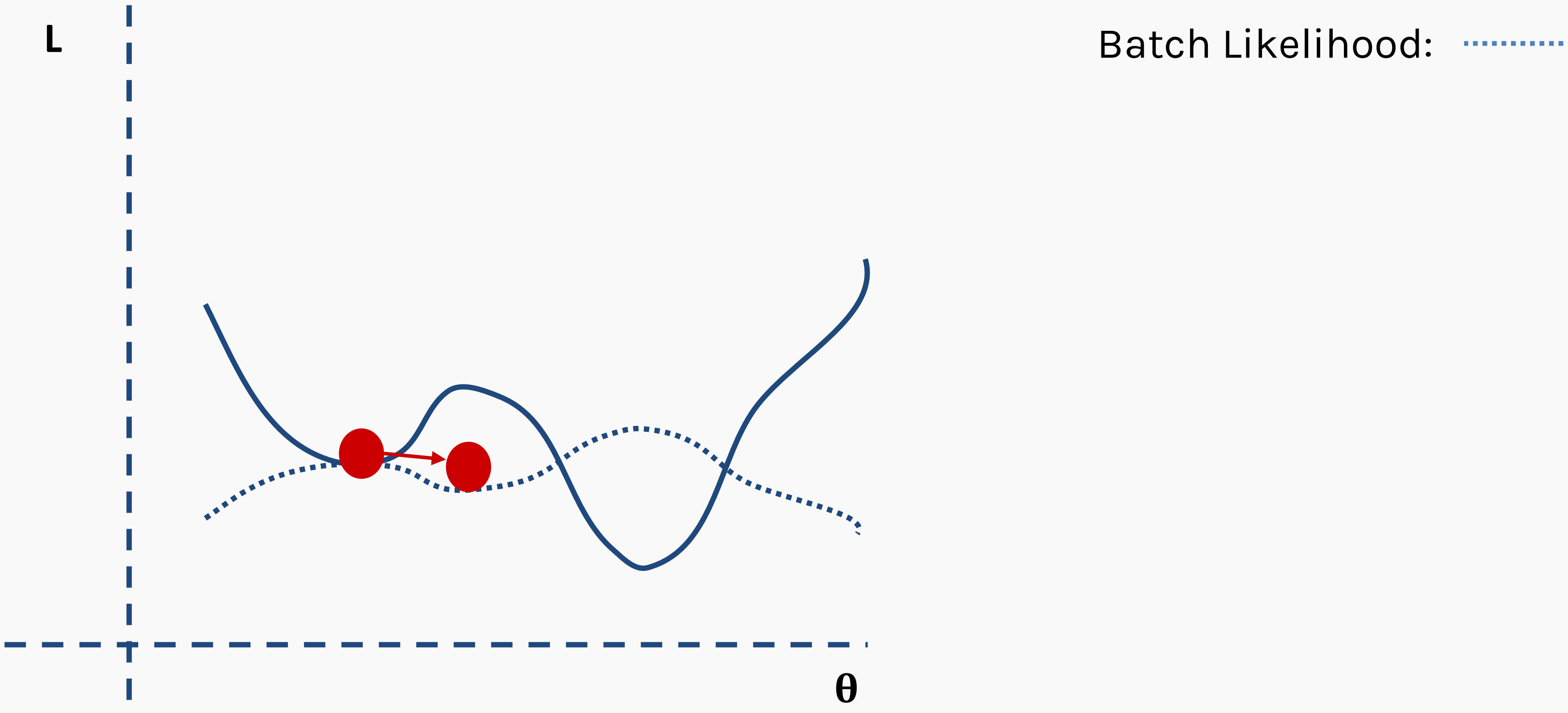


Batch and Stochastic Gradient Descent

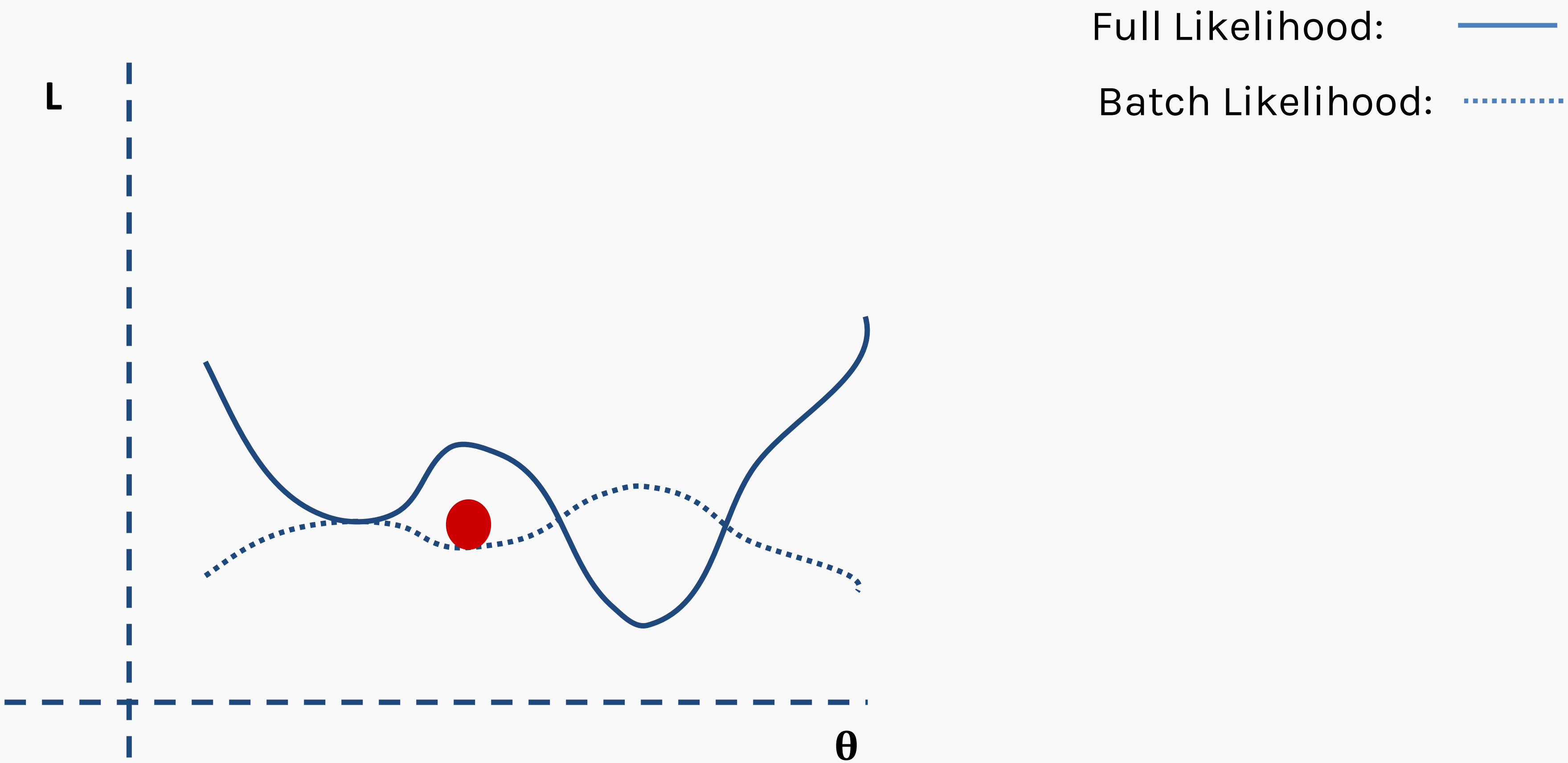


Full Likelihood:
Batch Likelihood:

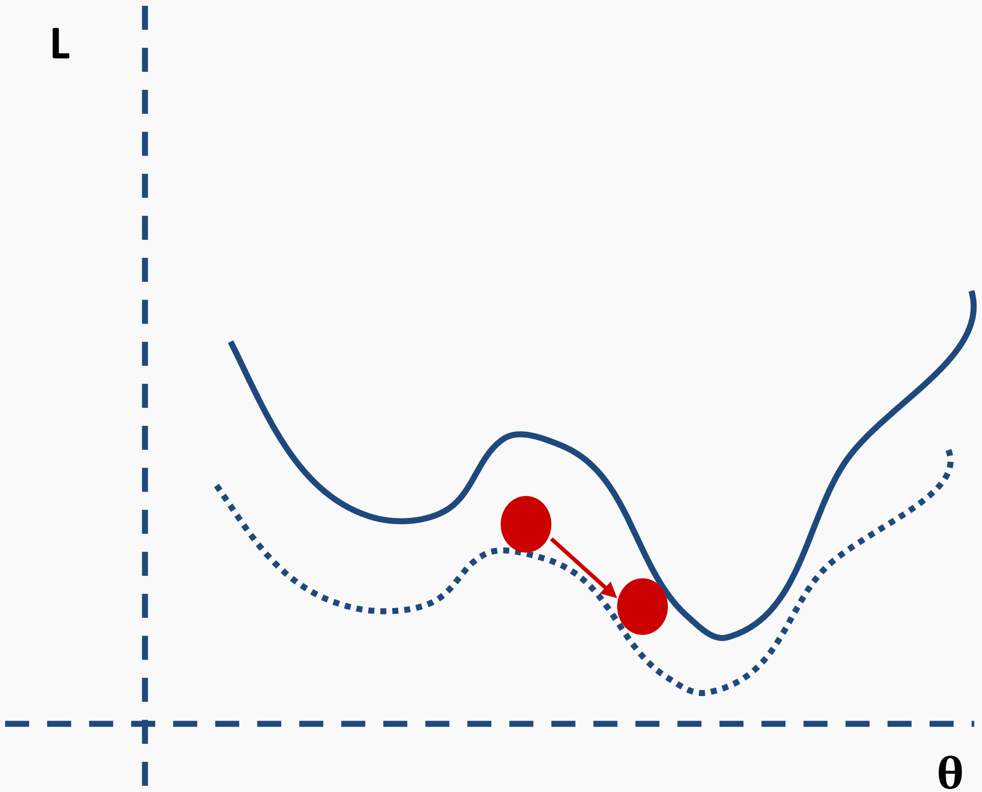
Batch and Stochastic Gradient Descent



Batch and Stochastic Gradient Descent

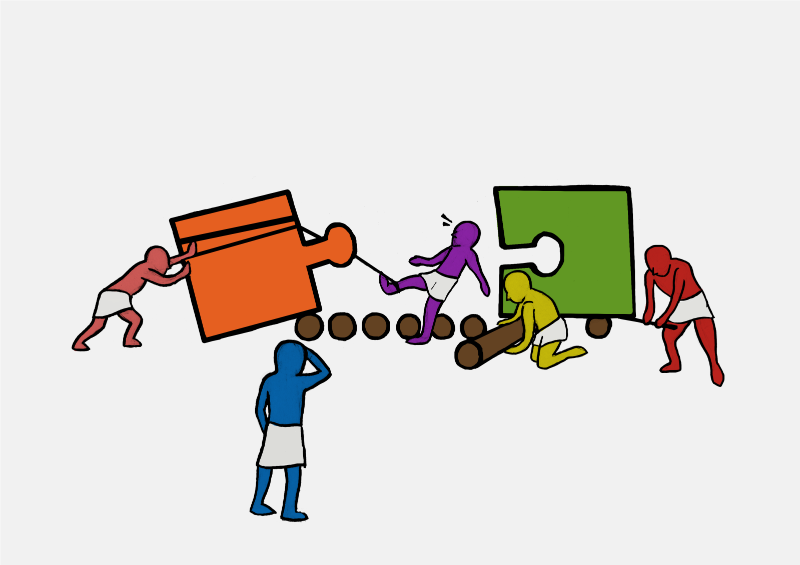


Batch and Stochastic Gradient Descent



Full Likelihood:

Batch Likelihood:



Exercise Gradient Descent