CSE 417T: Introduction to Machine Learning

Lecture 8: Logistic Regression

Henry Chai

09/20/18

Recall

Problem	Domain
Classification	$\mathcal{Y} = \{-1, +1\}$
Predicting Probabilities	$\mathcal{Y} = [0, 1]$
Regression	$\mathcal{Y} = \mathbb{R}$

Recall

Problem	Model
Linear Classification	$h(\vec{x}) = sign(\vec{w}^T \vec{x})$
Logistic Regression	$h(\vec{x}) = \theta(\vec{w}^T \vec{x})$
Linear Regression	$h(\vec{x}) = \vec{w}^T \vec{x}$

Predicting Probabilities

- Training data **does not** consist of probabilities
- Observations are still **binary**: $y_i = \pm 1$
- Goal is to learn $f(\vec{x}) = P\{y = +1|\vec{x}\}$
- Observations are inherently noisy

Logistic Regression: $\mathcal{H}$

- $h(\vec{x}) = P\{y = +1|\vec{x}\}$
- $h(\vec{x}) = \theta(\vec{w}^T\vec{x}) = \frac{1}{1+e^{-\vec{w}^T\vec{x}}} = \left(1 + e^{-\vec{w}^T\vec{x}}\right)^{-1} \in [0,1]$
- Note that $1 - \theta(\vec{w}^T\vec{x}) = \theta(-\vec{w}^T\vec{x})$

Logistic Regression: Error Measure

- Some hypothesis h is good if:
  - $h(\vec{x_i}) = \theta(\vec{w}^T\vec{x}) \approx 1$ when $y_i = +1$
  - $h(\vec{x_i}) = \theta(\vec{w}^T\vec{x}) \approx 0$ when $y_i = -1$

- $E_{in}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(\theta(\vec{w}^T\vec{x}) - \frac{1}{2}(1 + y_i)\right)^2$

Logistic Regression: Error Measure

- Some hypothesis h is good if:
  - $h(\vec{x_i}) = \theta(\vec{w}^T\vec{x}) \approx 1$ when $y_i = +1$
  - $h(\vec{x_i}) = \theta(\vec{w}^T\vec{x}) \approx 0$ when $y_i = -1$

- $E_{in}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + e^{-y_i\vec{w}^T\vec{x_i}}\right)$

- Cross-entropy Error

Logistic Regression: Error Measure

- Some hypothesis h is good if:
  - the probability of the training data $\mathcal{D}$ given h is high

- $E_{in}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)$

- Cross-entropy Error

Cross-entropy Error

$$P\{D|h\} = P\{(\overrightarrow{x_1}, y_1) \cap \cdots \cap (\overrightarrow{x_n}, y_n)|h\}$$

$$= \prod_{i=1}^{n} P\{(\overrightarrow{x_i}, y_i)|h\}$$

$$= \left(\prod_{i:y_i=+1} \theta(\vec{w}^T \overrightarrow{x_i}) \right) \left(\prod_{i:y_i=-1} 1 - \theta(\vec{w}^T \overrightarrow{x_i}) \right)$$

$$= \left(\prod_{i:y_i=+1} \theta(\vec{w}^T \overrightarrow{x_i}) \right) \left(\prod_{i:y_i=-1} \theta(-\vec{w}^T \overrightarrow{x_i}) \right)$$

$$= \prod_{i=1}^{n} \theta(y_i \vec{w}^T \overrightarrow{x_i}) = \prod_{i=1}^{n} \left(1 + e^{-y_i \vec{w}^T \overrightarrow{x_i}}\right)^{-1}$$

Cross-entropy Error

$$\text{Find } \vec{w}^* \text{ that maximizes } P\{D|\vec{w}\} = \prod_{i=1}^{n} \left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)^{-1}$$

$$\text{Find } \vec{w}^* \text{ that maximizes } P\{D|\vec{w}\} = \ln\left(\prod_{i=1}^{n} \left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)^{-1}\right)$$

$$\text{Find } \vec{w}^* \text{ that maximizes } P\{D|\vec{w}\} = \sum_{i=1}^{n} \ln\left(\left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)^{-1}\right)$$

$$\text{Find } \vec{w}^* \text{ that maximizes } P\{D|\vec{w}\} = -\sum_{i=1}^{n} \ln\left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)$$

$$\text{Find } \vec{w}^* \text{ that } \boldsymbol{minimizes} \; P\{D|\vec{w}\} = \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)$$

Cross-entropy Error

$$\text{Find } \vec{w}^* \text{ that maximizes } P\{D|\vec{w}\} = \prod_{i=1}^{n} \left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)^{-1}$$

$$\text{Find } \vec{w}^* \text{ that maximizes } P\{D|\vec{w}\} = \ln\left(\prod_{i=1}^{n} \left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)^{-1}\right)$$

$$\text{Find } \vec{w}^* \text{ that maximizes } P\{D|\vec{w}\} = \sum_{i=1}^{n} \ln\left(\left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)^{-1}\right)$$

$$\text{Find } \vec{w}^* \text{ that maximizes } P\{D|\vec{w}\} = -\sum_{i=1}^{n} \ln\left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)$$

$$\text{Find } \vec{w}^* \text{ that } \boldsymbol{minimizes} \; P\{D|\vec{w}\} = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)$$

Cross-entropy Error

$$E_{in}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + e^{-y_i \vec{w}^T \vec{x_i}}\right)$$

awkward silence

Cross-entropy Error

$$E_{in}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}\right)$$

$$\frac{\partial E_{in}(\vec{w})}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n} \frac{-(y_i x_{ij}) e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}}{1 + e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}}$$

$$\frac{\partial E_{in}(\vec{w})}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n} \frac{-y_i x_{ij}}{e^{y_i \sum_{j=0}^{d} w_j x_{ij}} + 1}$$

Cross-entropy Error

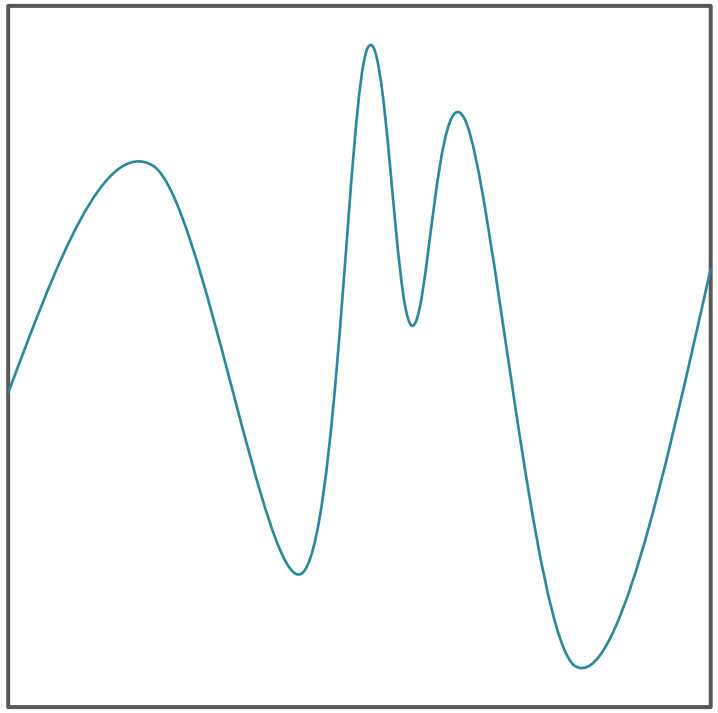
$$E_{in}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}\right)$$

$$\frac{\partial E_{in}(\vec{w})}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n} \frac{-(y_i x_{ij}) e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}}{1 + e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}}$$

$$\nabla_{\vec{w}} E_{in}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \frac{-y_i \vec{x_i}}{e^{y_i \vec{w}^T \vec{x_i}} + 1}$$

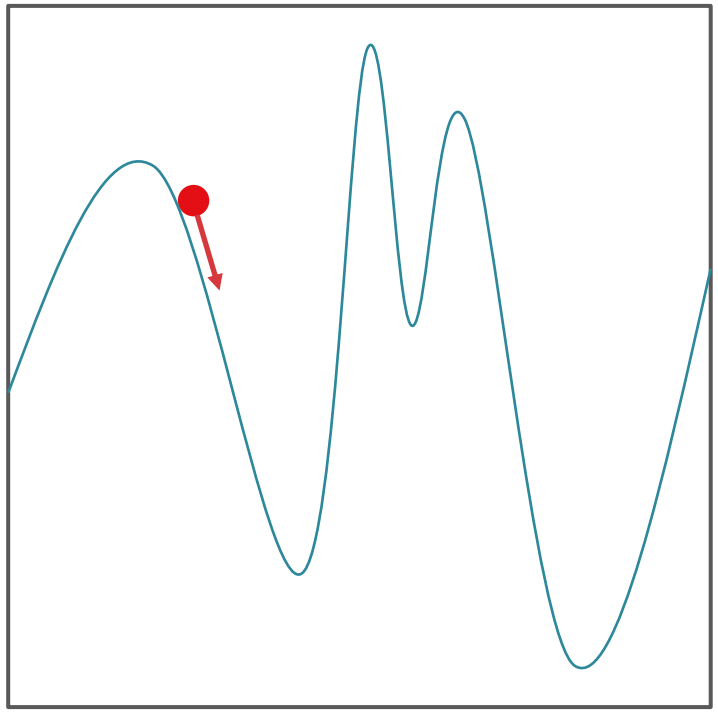
Gradient Descent: Intuition

- Iterative method for minimizing functions
- Requires the gradient to exist everywhere



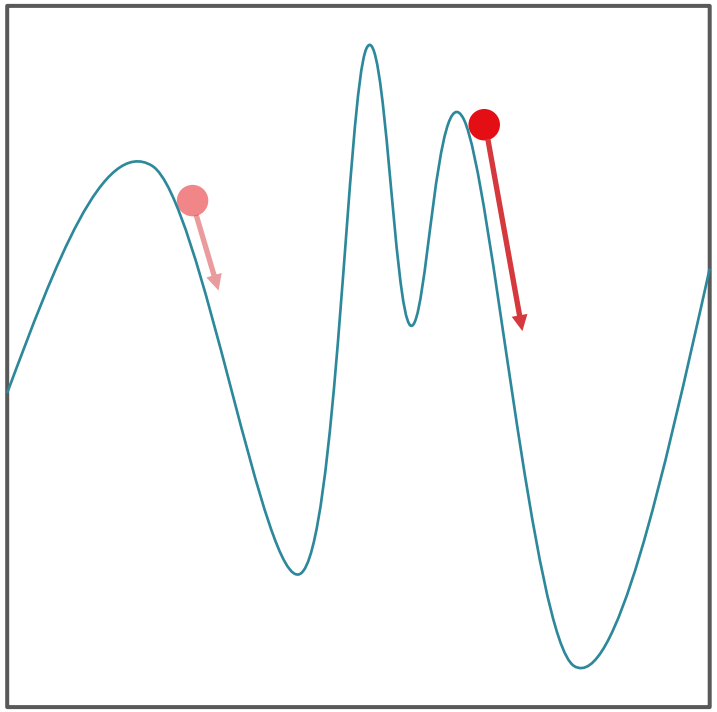
Gradient Descent: Intuition

- Iterative method for minimizing functions
- Requires the gradient to exist everywhere



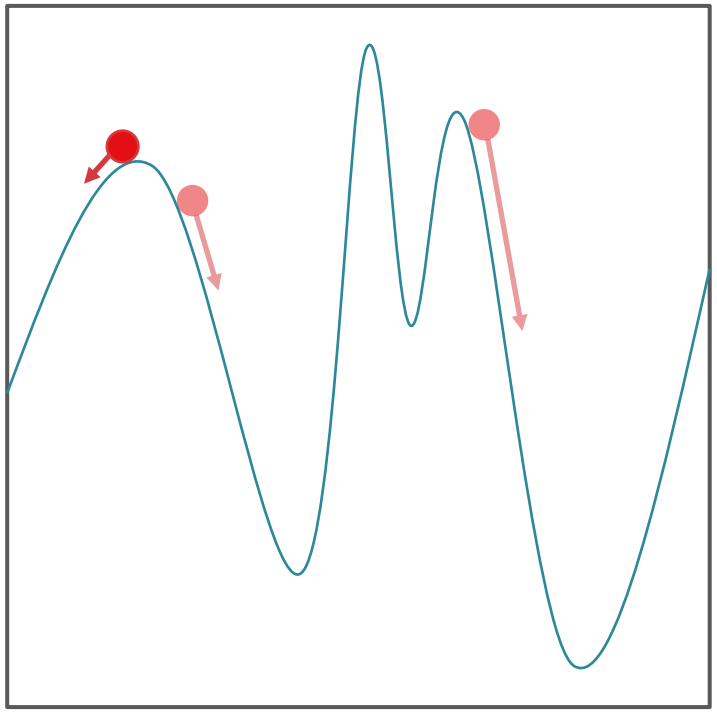
Gradient Descent: Intuition

- Iterative method for minimizing functions
- Requires the gradient to exist everywhere



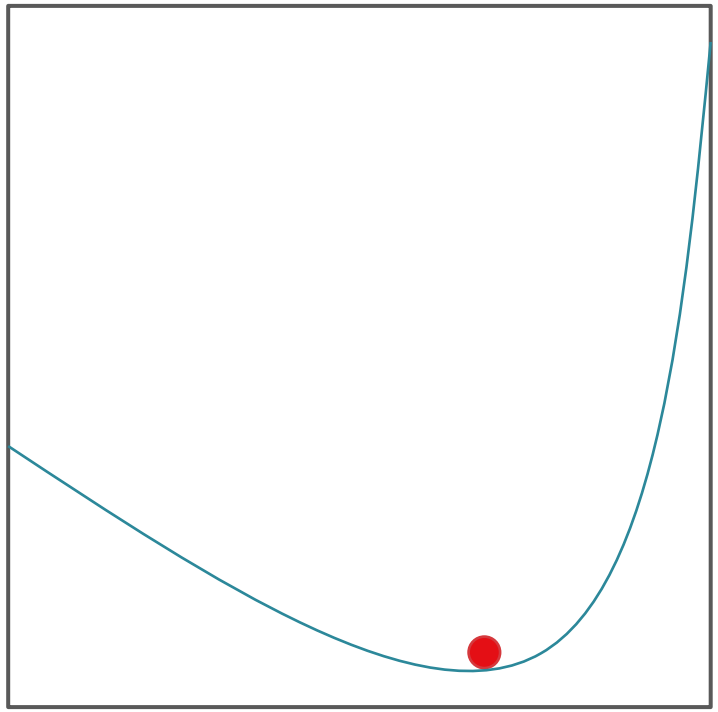
Gradient Descent: Intuition

- Iterative method for minimizing functions
- Requires the gradient to exist everywhere



Gradient Descent: Intuition

- Iterative method for minimizing functions
- Requires the gradient to exist everywhere



Cross-entropy Error

$$E_{in}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}\right)$$

$$\frac{\partial E_{in}(\vec{w})}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n} \frac{-(y_i x_{ij}) e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}}{1 + e^{-y_i \sum_{j=0}^{d} w_j x_{ij}}}$$

$$\nabla_{\vec{w}} E_{in}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \frac{-y_i \vec{x_i}}{e^{y_i \vec{w}^T \vec{x_i}} + 1}$$

$$\mathrm{H}_{\vec{w}} E_{in}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(y_i^2 e^{y_i \vec{w}^T \vec{x_i}}\right)\left(\vec{x_i}\,\vec{x_i}^T\right)}{\left(e^{y_i \vec{w}^T \vec{x_i}} + 1\right)^2}$$

$\mathrm{H}_{\vec{w}} E_{in}(\vec{w})$ is positive semidefinite $\forall\ \vec{w} \rightarrow E_{in}(\vec{w})$ is convex

Gradient Descent: Intuition

- Suppose the current location is $\vec{w}_{(t)}$
- Move some distance, η, in the "most downhill" direction possible, $\hat{v}$
- $\vec{w}_{(t+1)} = \vec{w}_{(t)} + \eta\hat{v}$

$\hat{v}$

- Fix η and choose $\hat{v}$ to maximize the decrease in E_{in} after making the update $\vec{w}_{(t+1)} = \vec{w}_{(t)} + \eta\hat{v}$

- $\Delta E_{in} = E_{in}\left(\vec{w}_{(t)} + \eta\hat{v}\right) - E_{in}\left(\vec{w}_{(t)}\right)$

$\hat{v}$

- Fix η and choose $\hat{v}$ to minimize ΔE_{in} after making the update $\vec{w}_{(t+1)} = \vec{w}_{(t)} + \eta\hat{v}$

- $\Delta E_{in}(\hat{v}) = E_{in}\left(\vec{w}_{(t)} + \eta\hat{v}\right) - E_{in}\left(\vec{w}_{(t)}\right)$

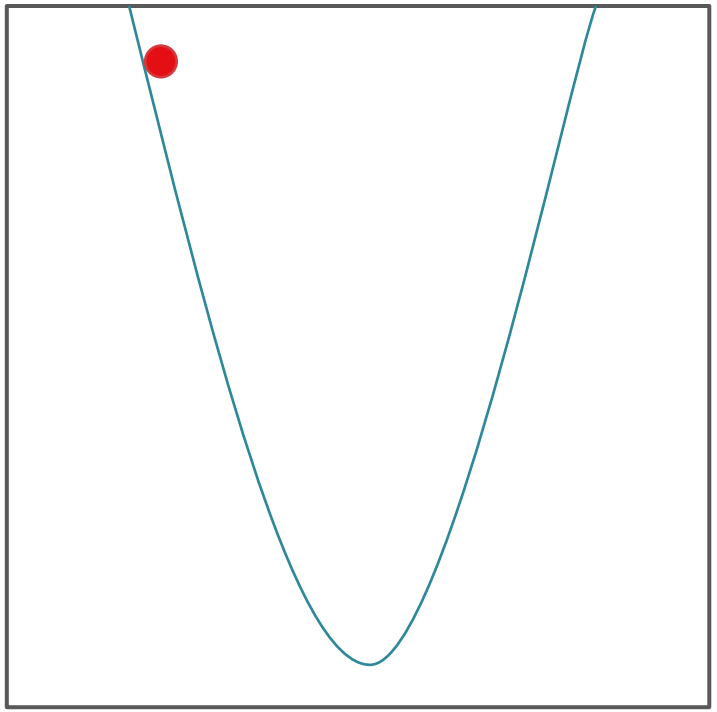
$\hat{v}$

- Fix η and choose $\hat{v}$ to minimize ΔE_{in} after making the update $\vec{w}_{(t+1)} = \vec{w}_{(t)} + \eta\hat{v}$

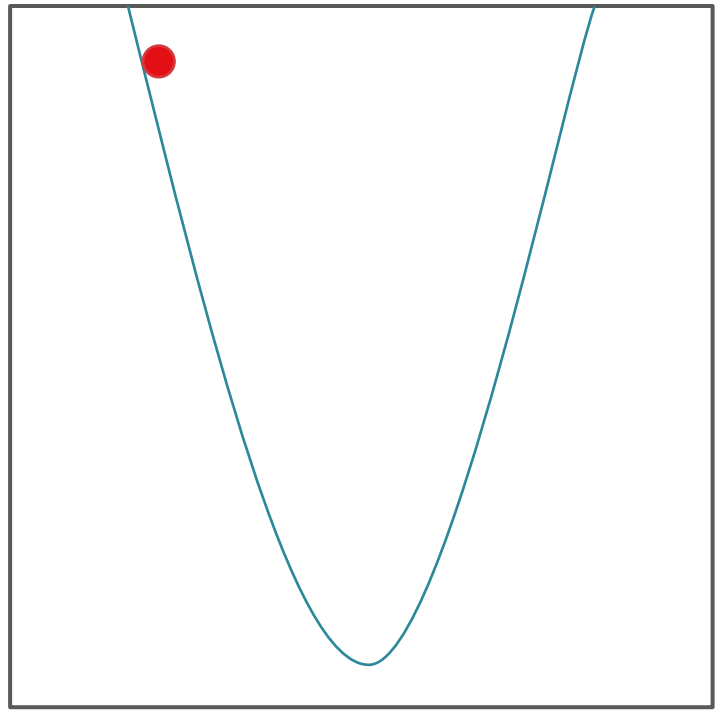
- $$\Delta E_{in}(\hat{v}) = E_{in}\left(\vec{w}_{(t)} + \eta\hat{v}\right) - E_{in}\left(\vec{w}_{(t)}\right)$$
$$\approx \left(E_{in}\left(\vec{w}_{(t)}\right) + \eta\hat{v}^T\nabla_{\vec{w}}E_{in}\left(\vec{w}_{(t)}\right)\right) - E_{in}\left(\vec{w}_{(t)}\right)$$
$$\approx \eta\hat{v}^T\nabla_{\vec{w}}E_{in}\left(\vec{w}_{(t)}\right)$$
$$\geq -\eta\left\|\nabla_{\vec{w}}E_{in}\left(\vec{w}_{(t)}\right)\right\|$$

- $$\hat{v}_t^* = -\frac{\nabla_{\vec{w}}E_{in}\left(\vec{w}_{(t)}\right)}{\left\|\nabla_{\vec{w}}E_{in}\left(\vec{w}_{(t)}\right)\right\|}$$

η

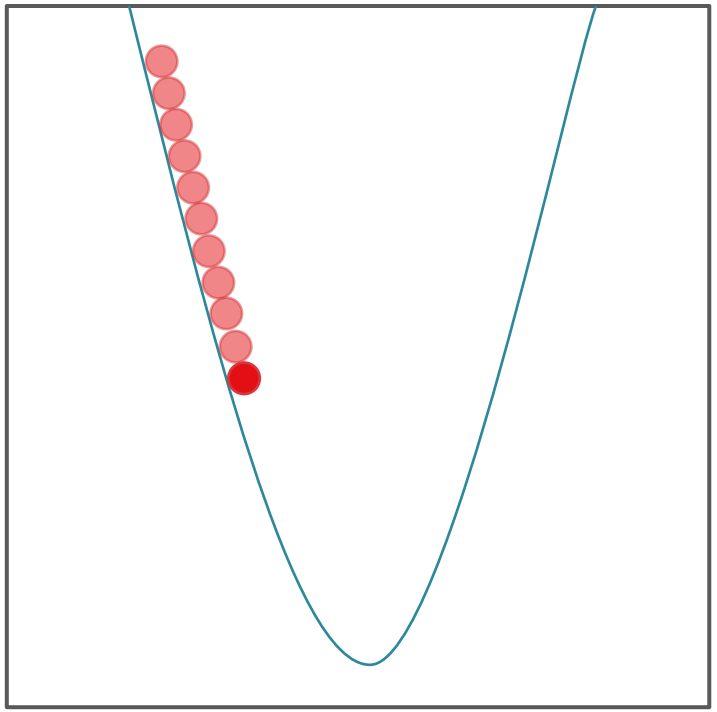


Small η

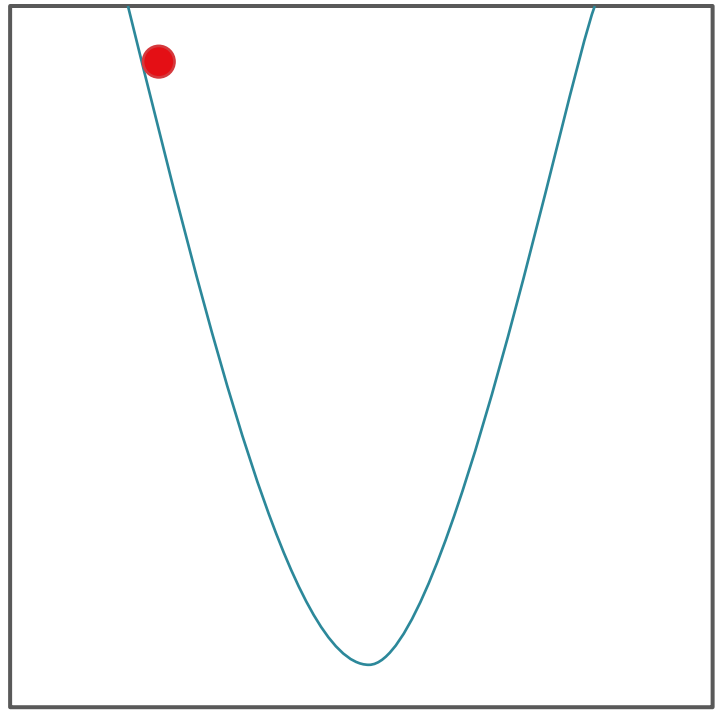


Large η

η

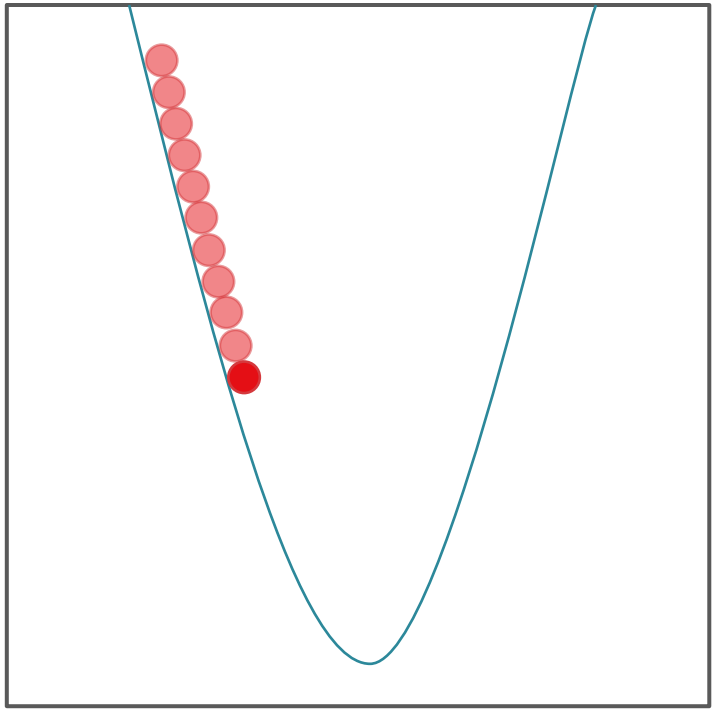


Small η

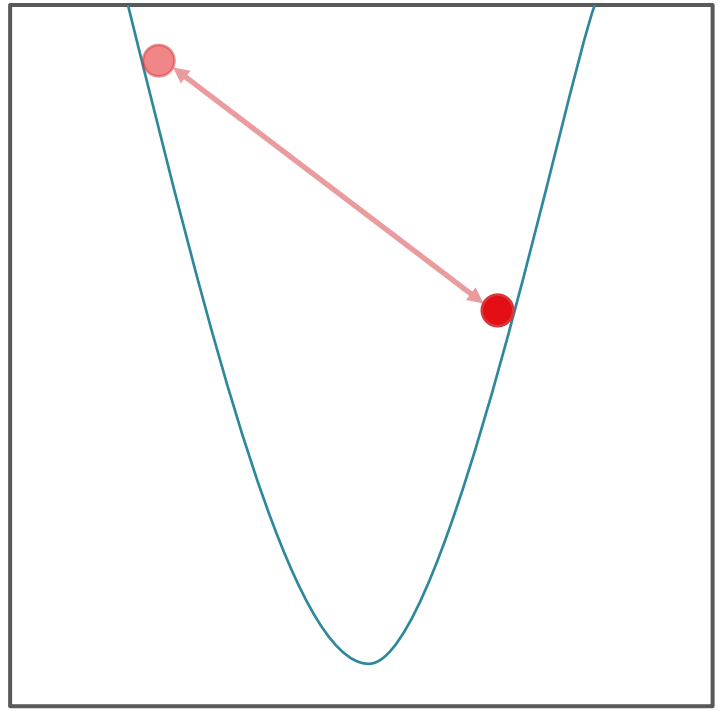


Large η

η



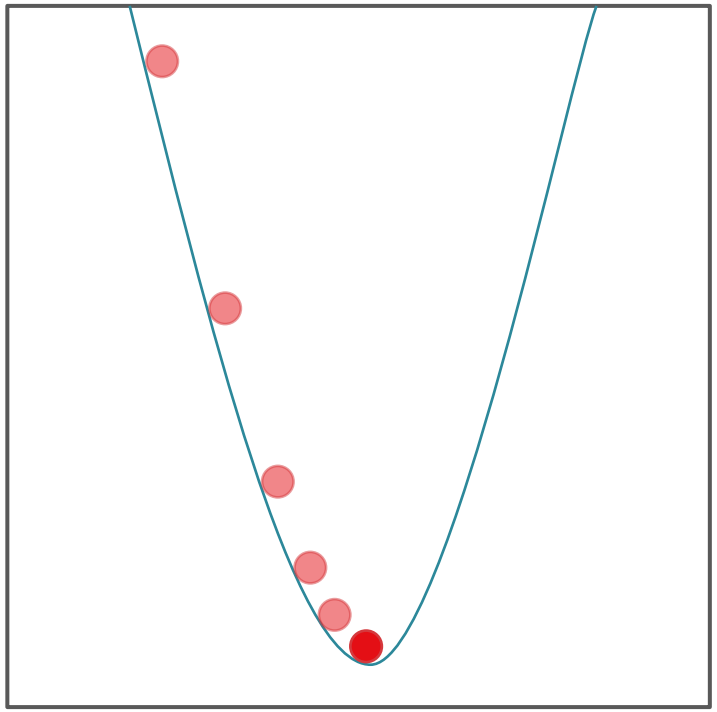
Small η



Large η

η_t

- Use a variable η_t instead of a fixed η



η_t

- Set $\eta_t = \eta_0 \|\nabla_{\vec{w}} E_{in}(\vec{w}_t)\|$
- η_t decreases as t increases, because $\|\nabla_{\vec{w}} E_{in}(\vec{w}_t)\|$ decreases as $E_{in}(\vec{w}_t)$ approaches its minimum
- $$\vec{w}_{(t+1)} = \vec{w}_{(t)} + \eta_t \hat{v}_t^*$$
$$= \vec{w}_{(t)} + (\eta_0 \|\nabla_{\vec{w}} E_{in}(\vec{w}_t)\|) \left(- \frac{\nabla_{\vec{w}} E_{in}(\vec{w}_{(t)})}{\|\nabla_{\vec{w}} E_{in}(\vec{w}_{(t)})\|} \right)$$
$$= \vec{w}_{(t)} - \eta_0 \nabla_{\vec{w}} E_{in}(\vec{w}_{(t)})$$