Function Approximation Methods-III

CS771: Introduction to Machine Learning

Purushottam Kar

Discussion Session

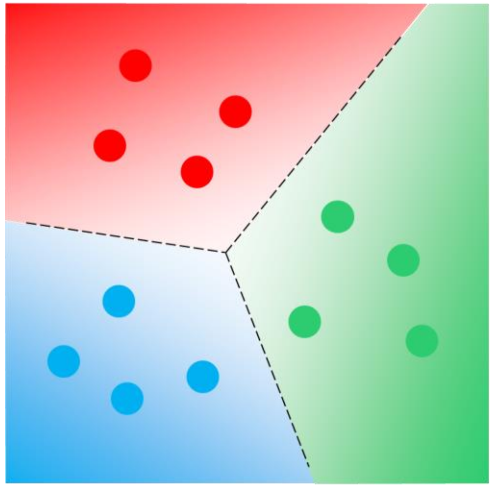
Sunday, 03 September, 2017
1800-1930 hrs, RM101, CSE department
Post query on http://tinyurl.com/ml17-18ads1 before coming!

Recap

Multi-classification Loss Functions

One-vs-All (OVA)

$$\hat{y}^i = \arg\max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



$$\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^K]$$

$$\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \|\mathbf{w}^k\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \boldsymbol{\eta}^i)$$

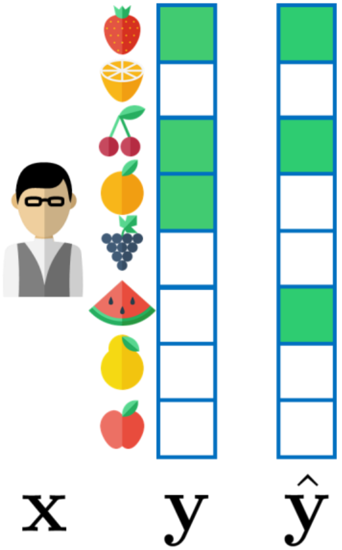
$$\boldsymbol{\eta}^i = \langle \mathbf{W}, \mathbf{x}^i \rangle = [\boldsymbol{\eta}_1^i, \boldsymbol{\eta}_2^i, \ldots, \boldsymbol{\eta}_K^i]$$

$$\ell_{\text{cs}}(y, \boldsymbol{\eta}) = [1 + \max_{k \neq y} \boldsymbol{\eta}_k - \boldsymbol{\eta}_y]_+$$

Crammer-Singer loss function

August 23, 2017

Multi-label Classification Loss Functions



$$\begin{array}{c|c} \left|\left\{ i : \begin{array}{l} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right| \; a & b \; \left|\left\{ i : \begin{array}{l} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = -1 \end{array} \right\}\right| \\ \hline \left|\left\{ i : \begin{array}{l} \hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right| \; c & d \; \left|\left\{ i : \begin{array}{l} \hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = -1 \end{array} \right\}\right| \end{array}$$

$$\ell_{\text{Hamming}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{b + c}{a + b + c + d}$$

$$r_{\text{Precision}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{a}{a + b} \qquad r_{\text{Recall}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{a}{a + c}$$

$$F(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2 r_{\text{Prec}} \cdot r_{\text{Rec}}}{r_{\text{Prec}} + r_{\text{Rec}}} = \frac{2a}{2a + b + c}$$

August 23, 2017

Can be used for evaluation and training

Historically

More recent

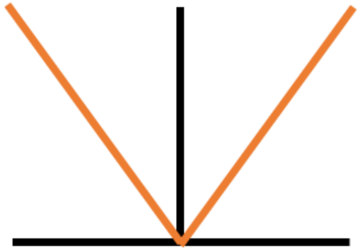
First-order Optimality

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \; f(\mathbf{w}) + r(\mathbf{w})$$

$$\nabla f(\mathbf{w}) + \nabla r(\mathbf{w}) = \mathbf{0}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2$$

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$



$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I\right)^{-1} \mathbf{X}\mathbf{y}$$

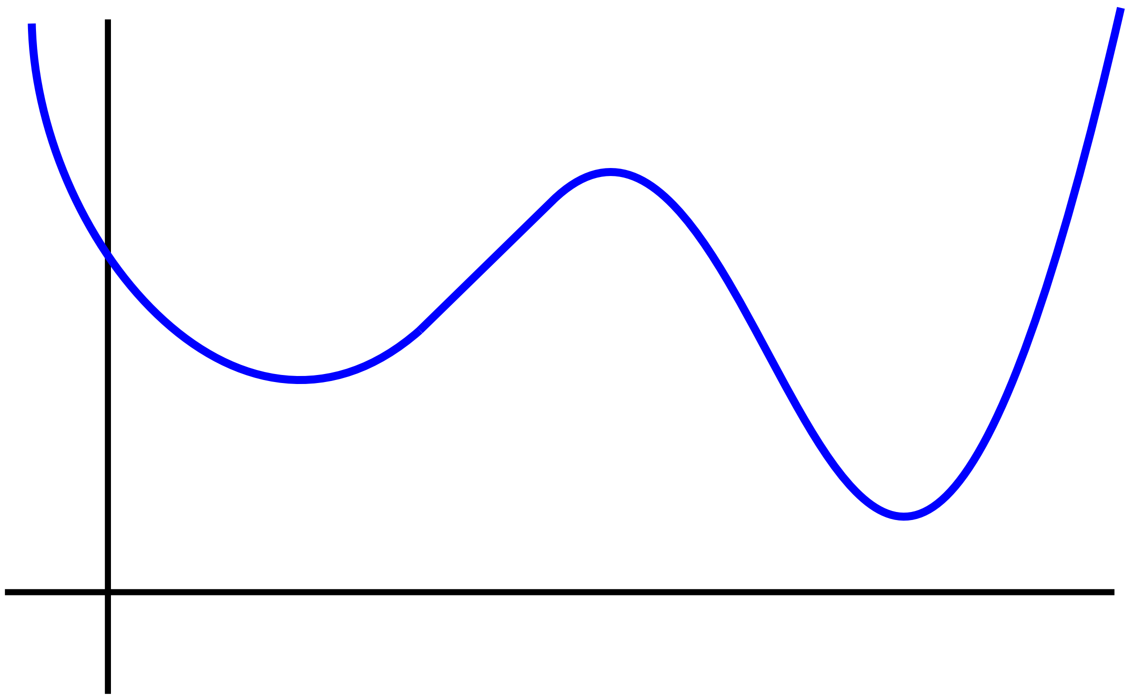
$$f(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)\right)$$

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$

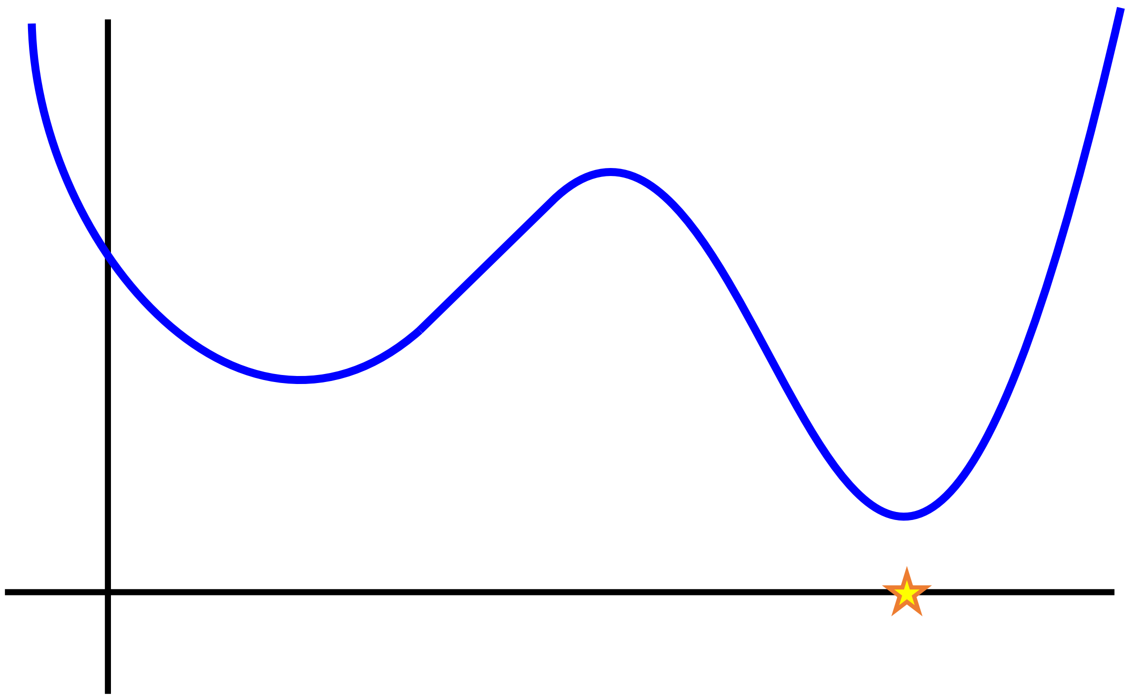
?

$$\sum_{i=1}^{n} (1 - \sigma(y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)) y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w} = \mathbf{0}$$

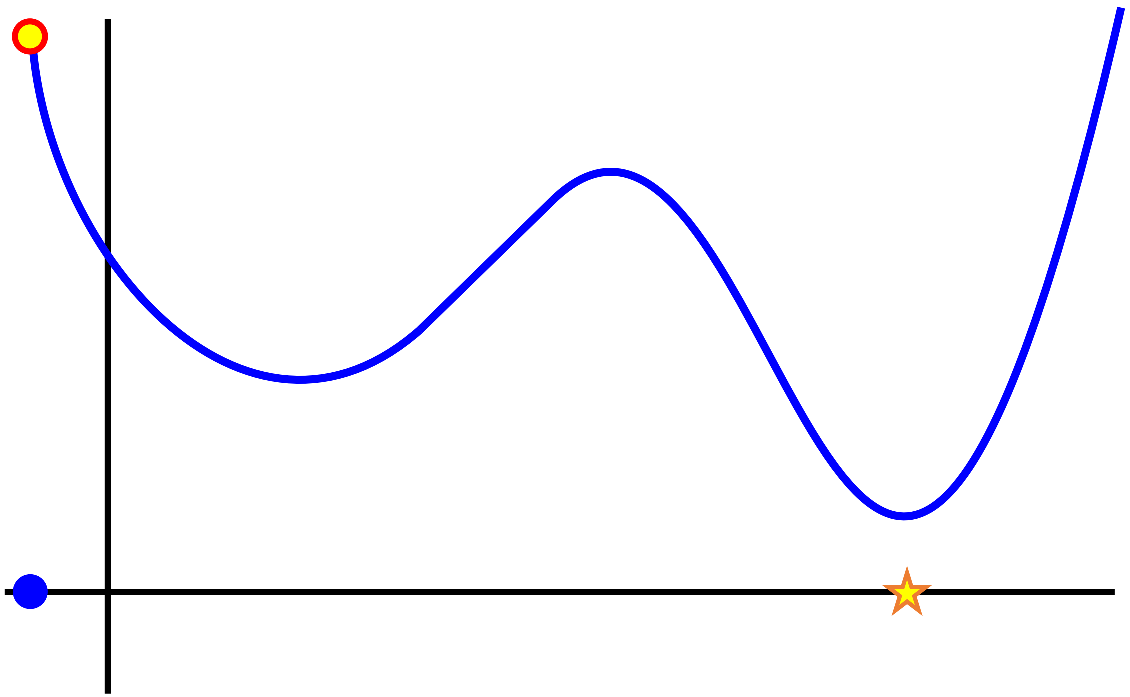
Gradient Descent



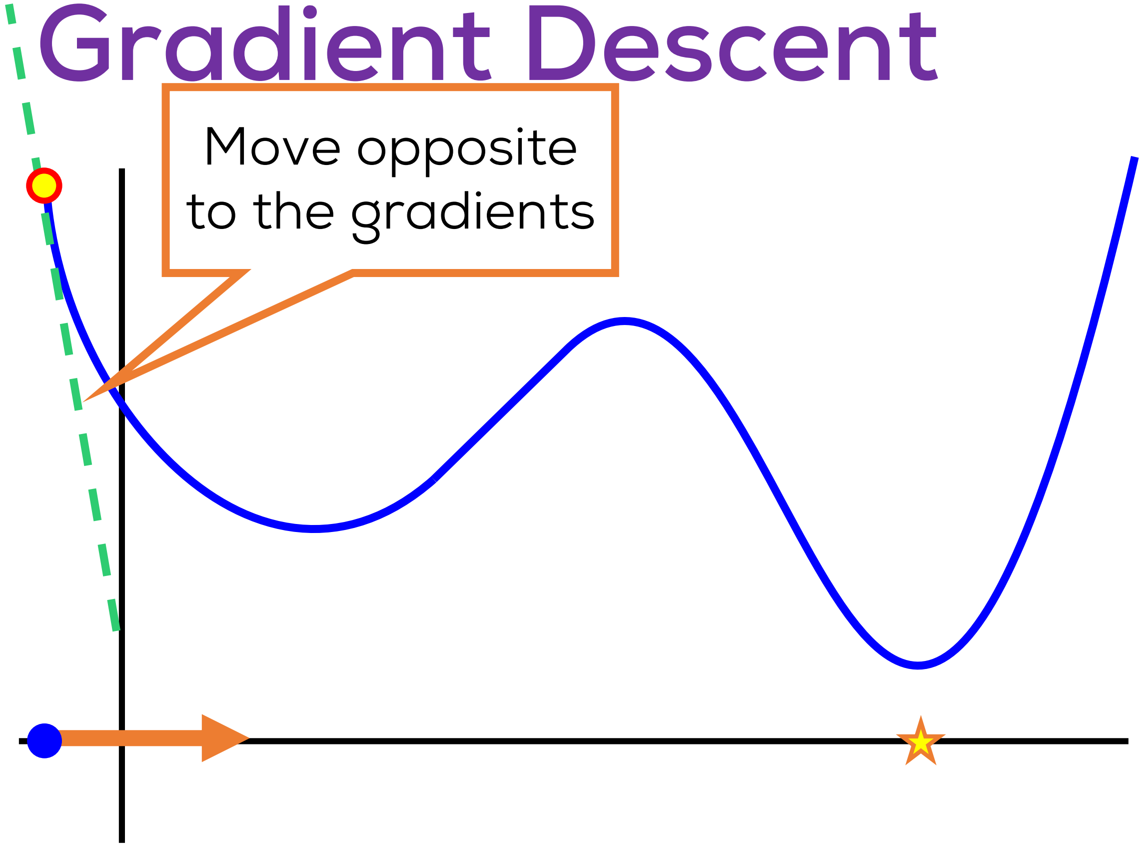
Gradient Descent



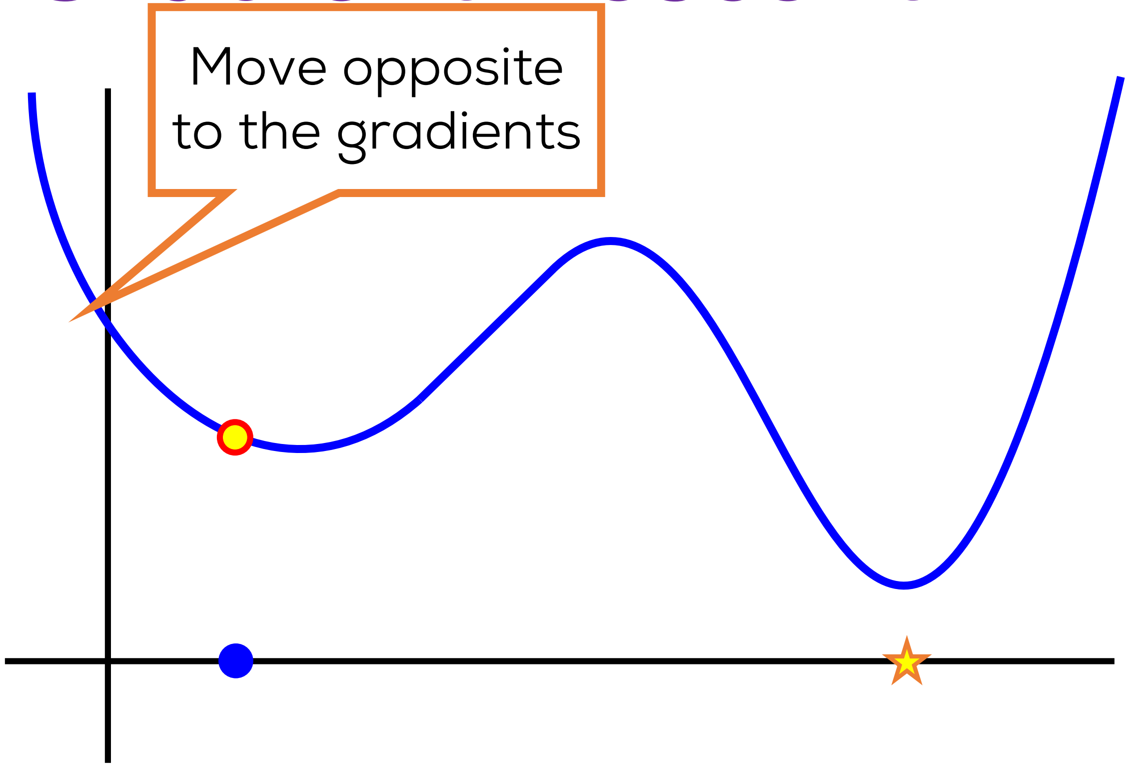
Gradient Descent



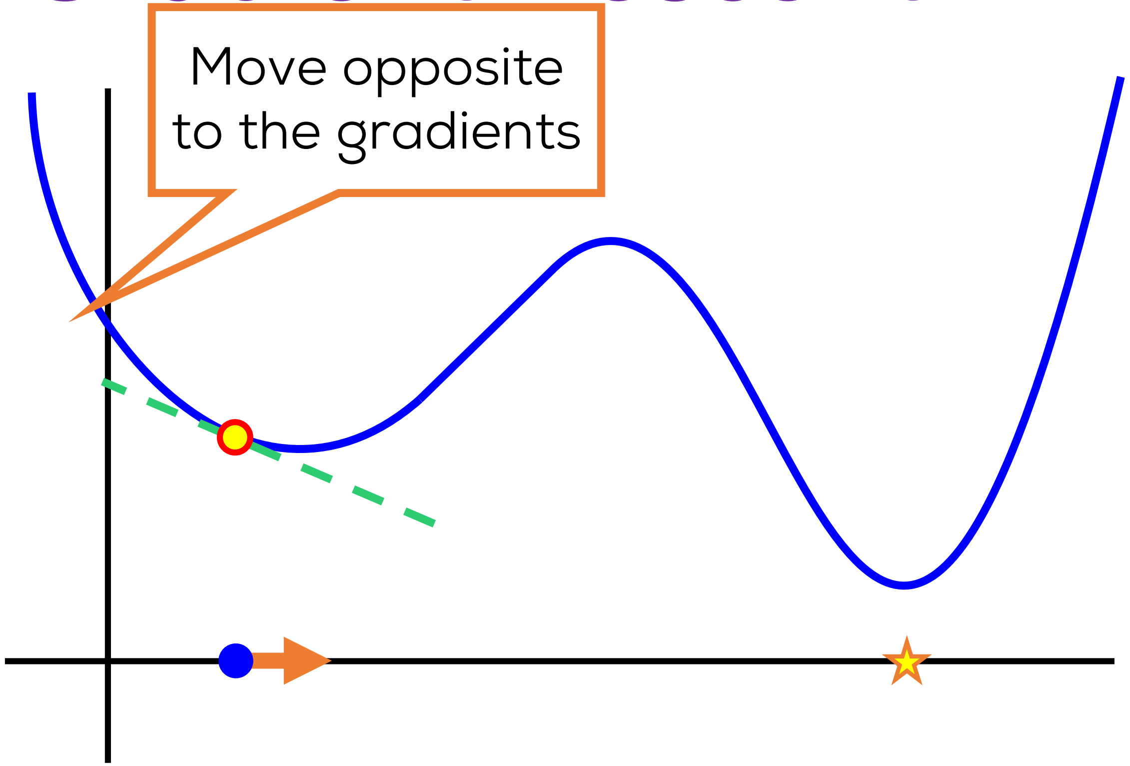
Gradient Descent



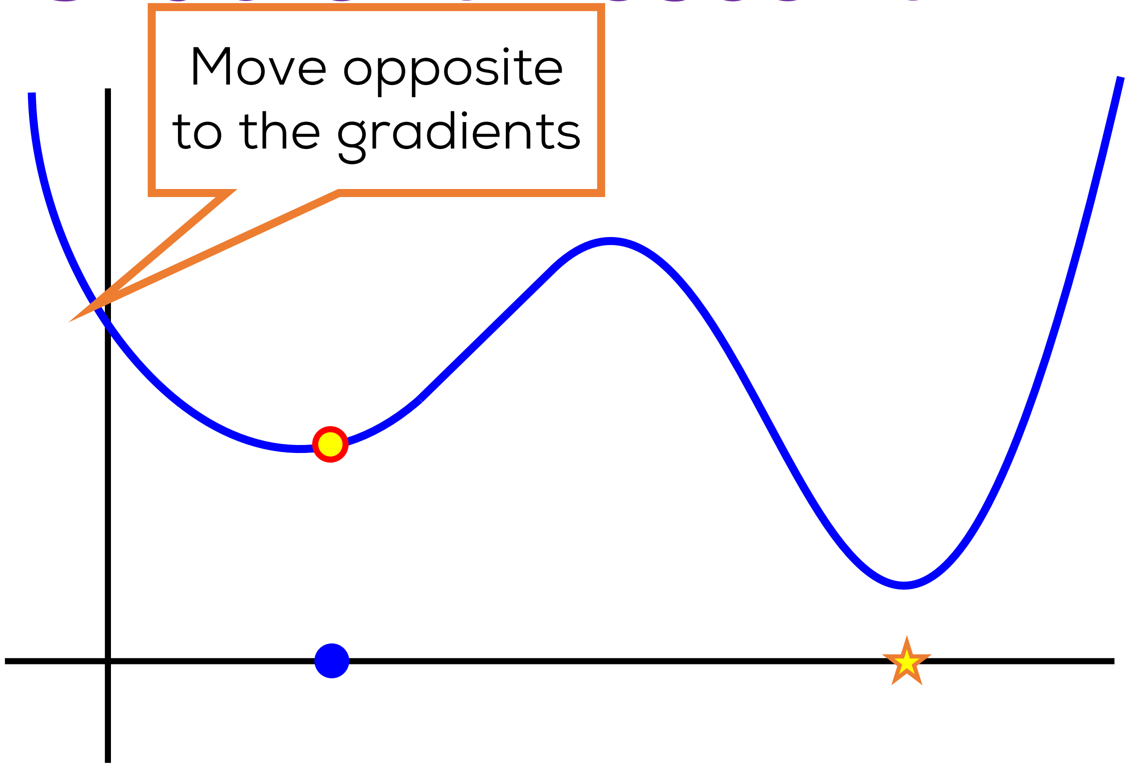
Gradient Descent



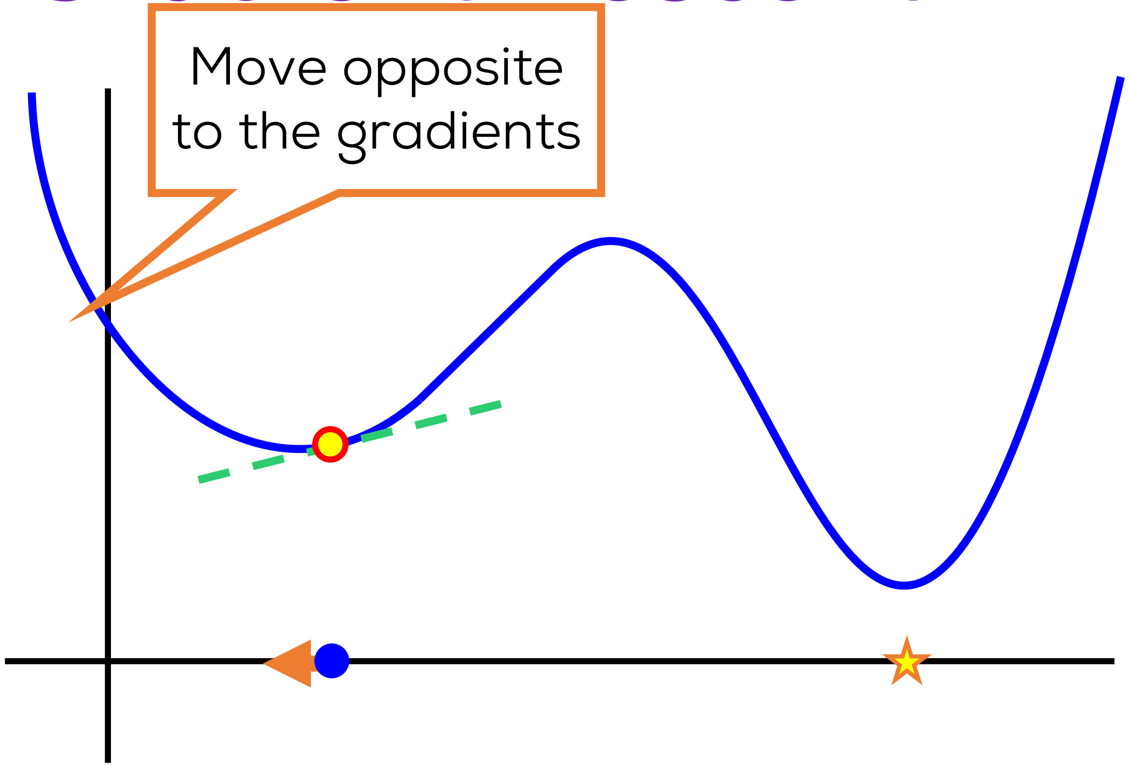
Gradient Descent



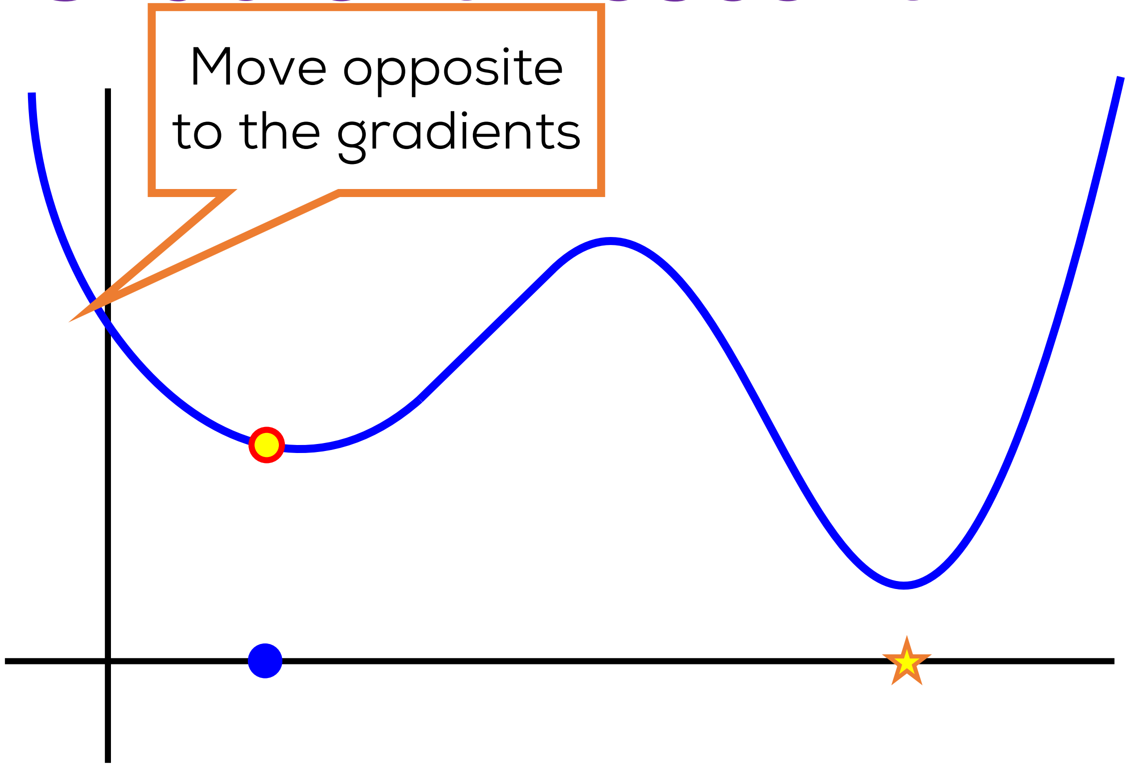
Gradient Descent



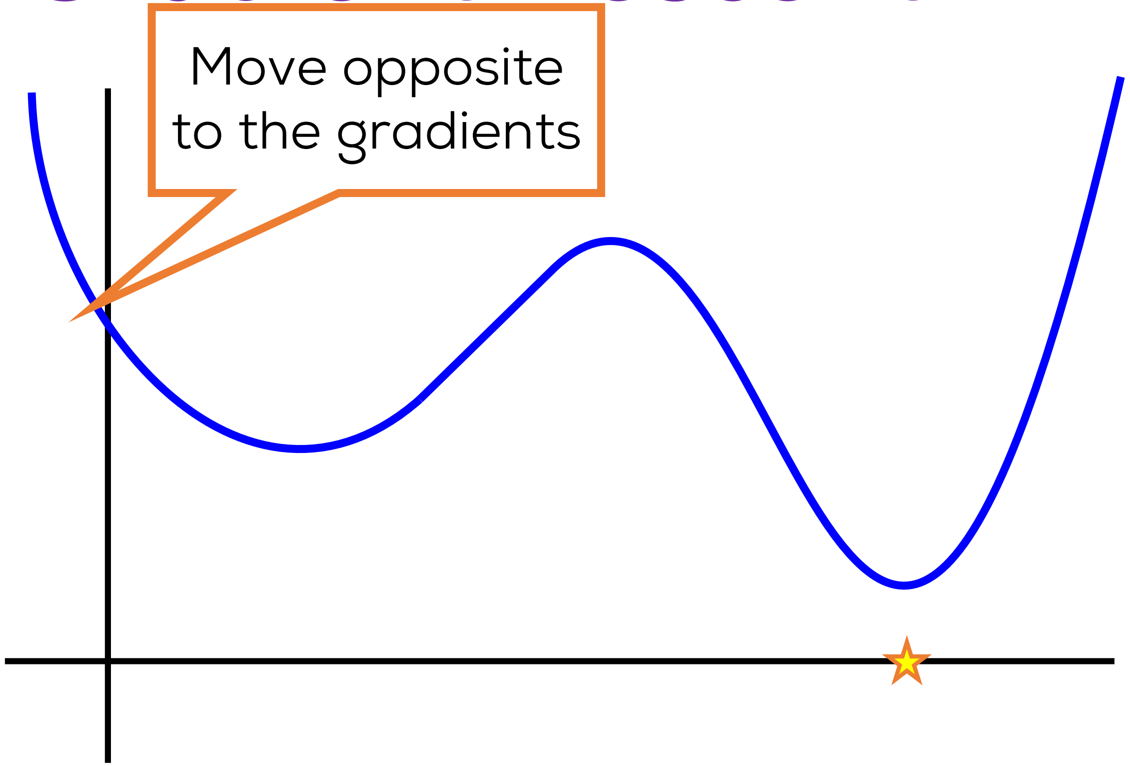
Gradient Descent



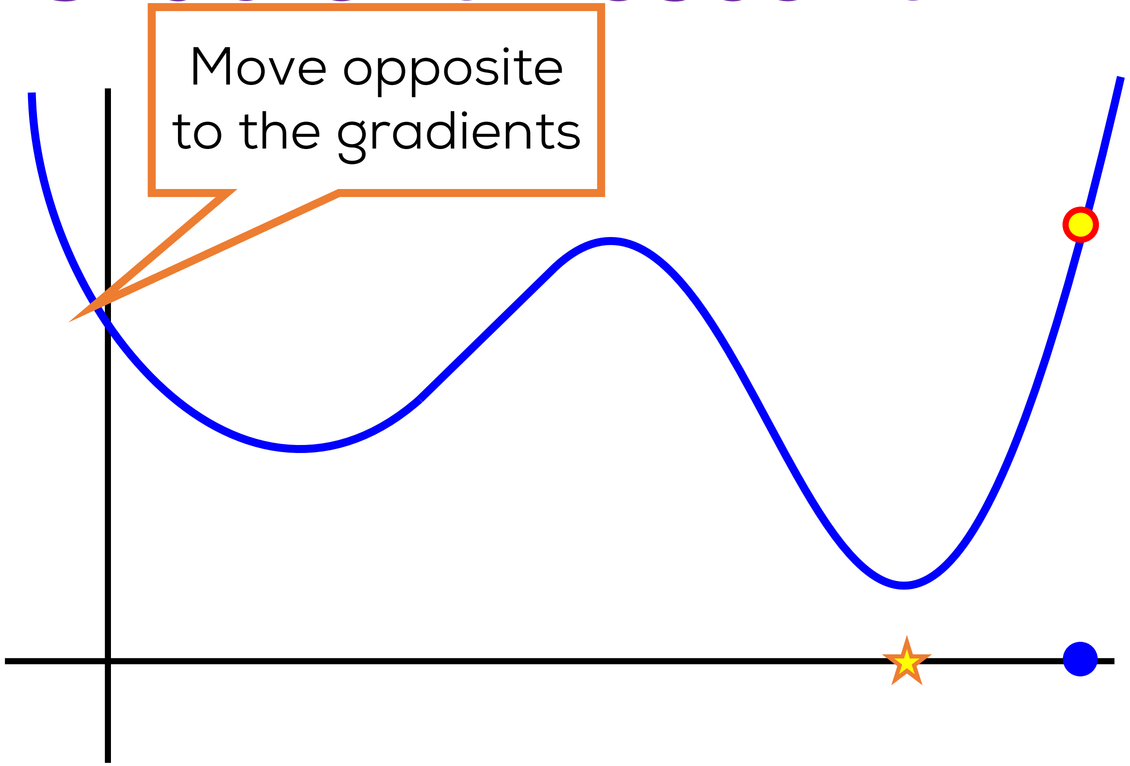
Gradient Descent



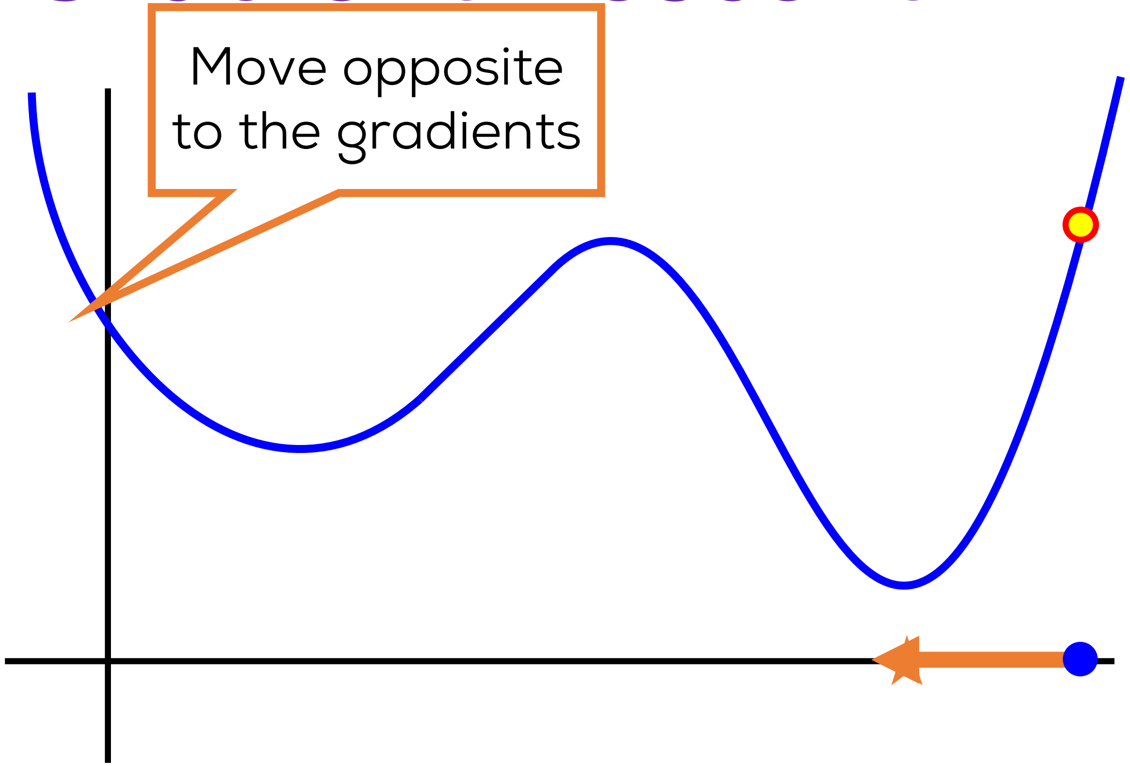
Gradient Descent



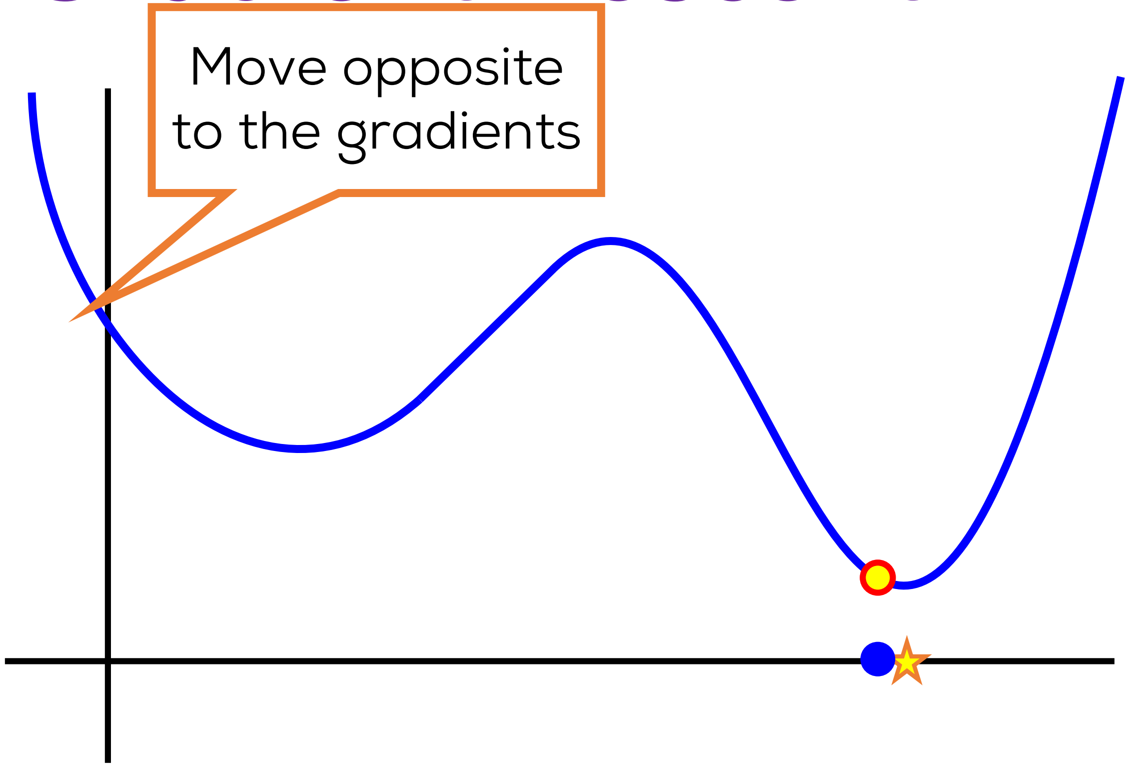
Gradient Descent



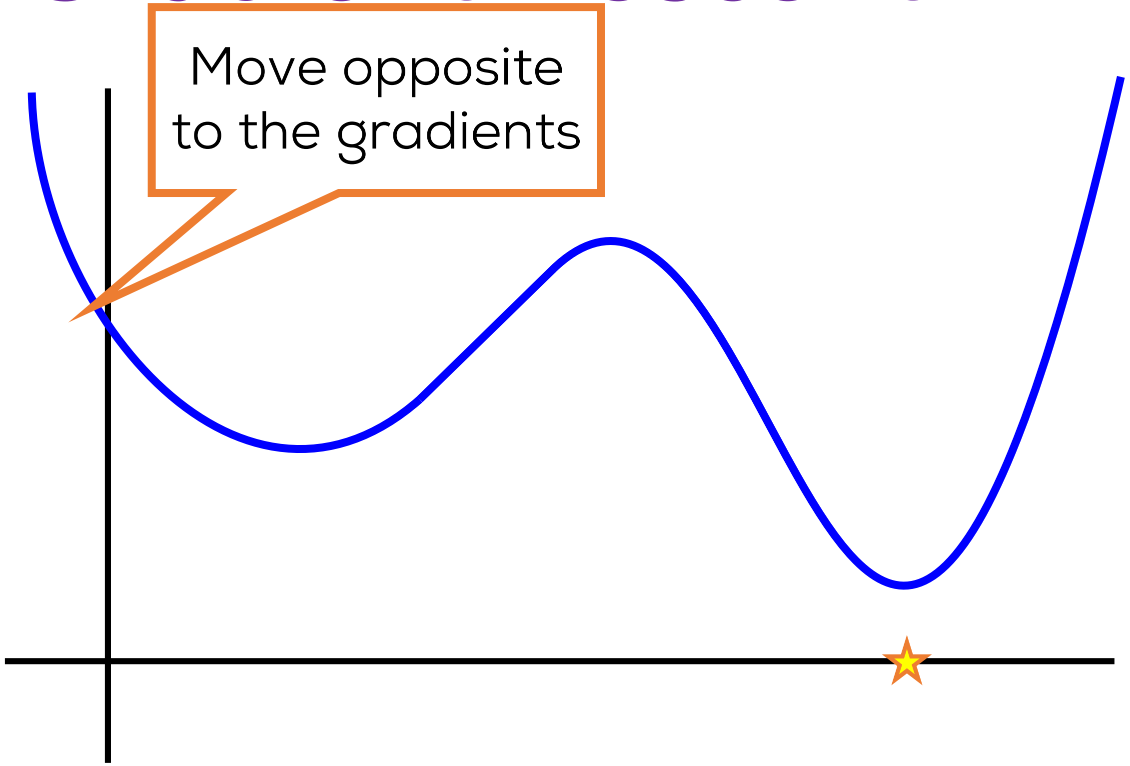
Gradient Descent



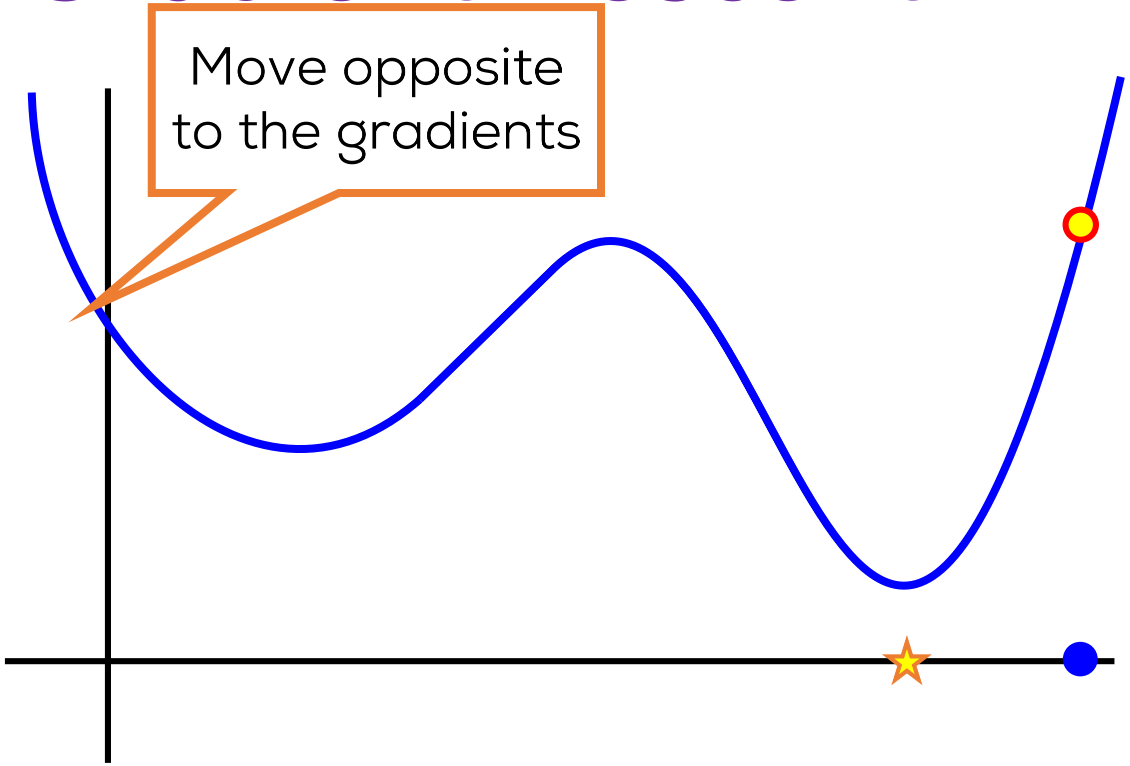
Gradient Descent



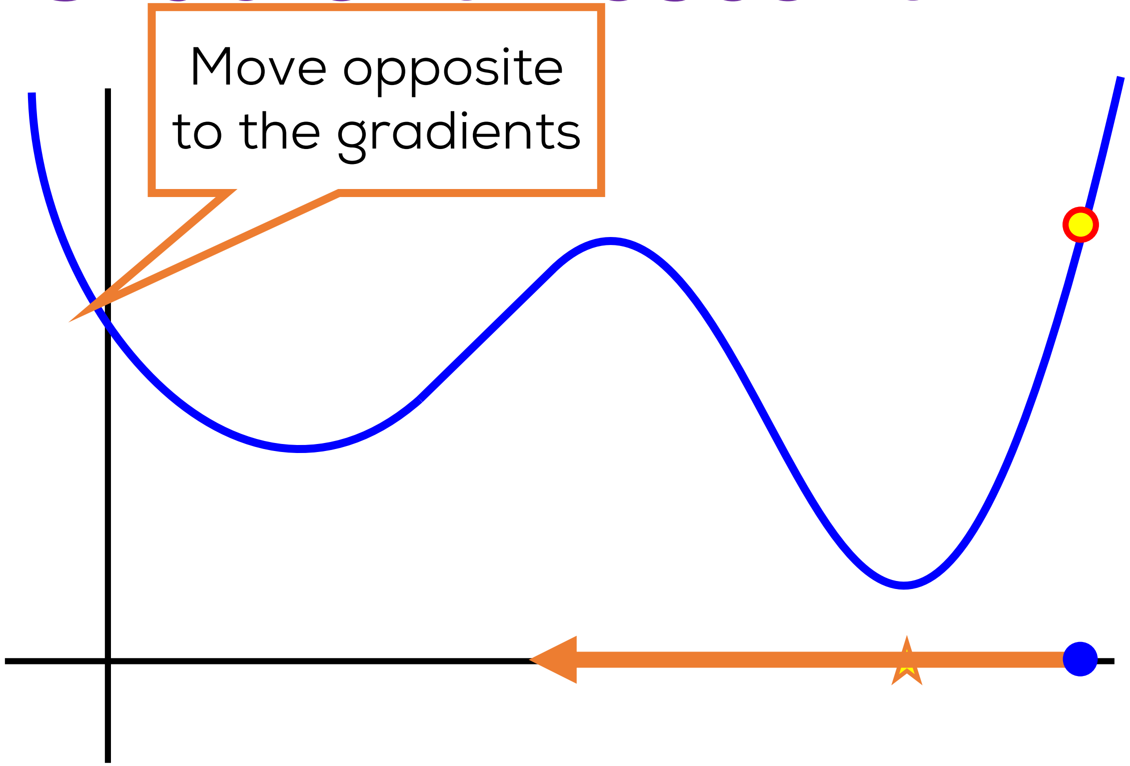
Gradient Descent



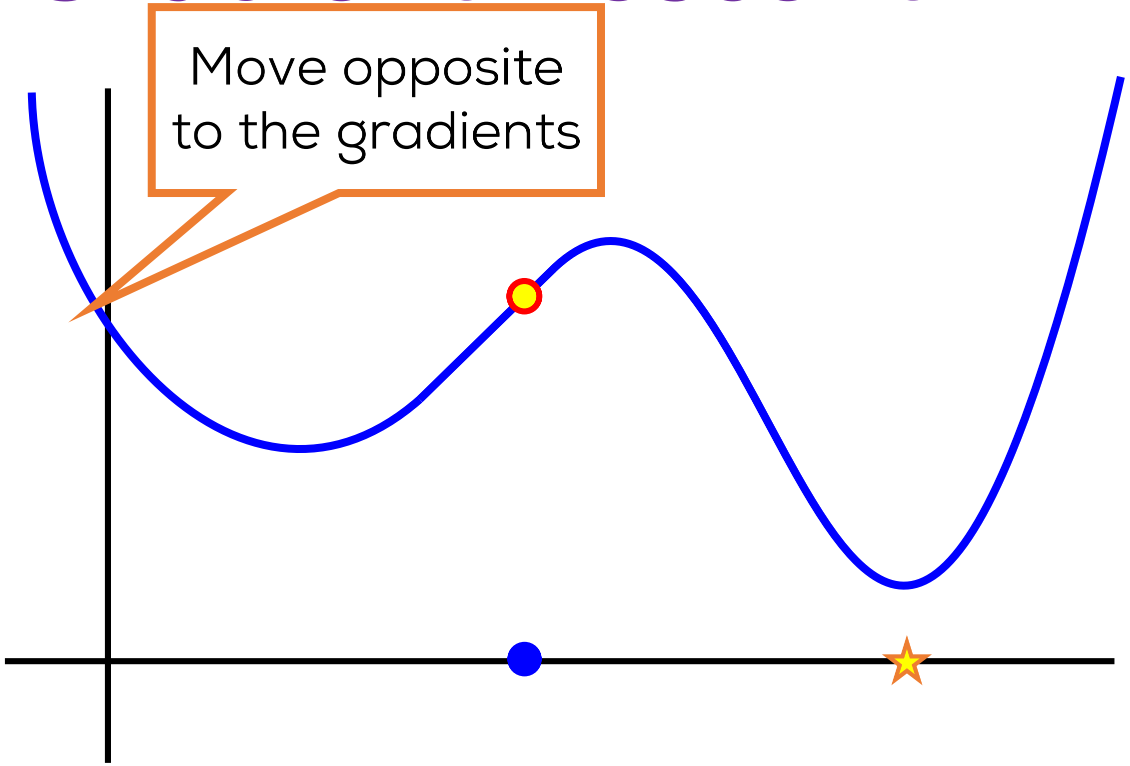
Gradient Descent



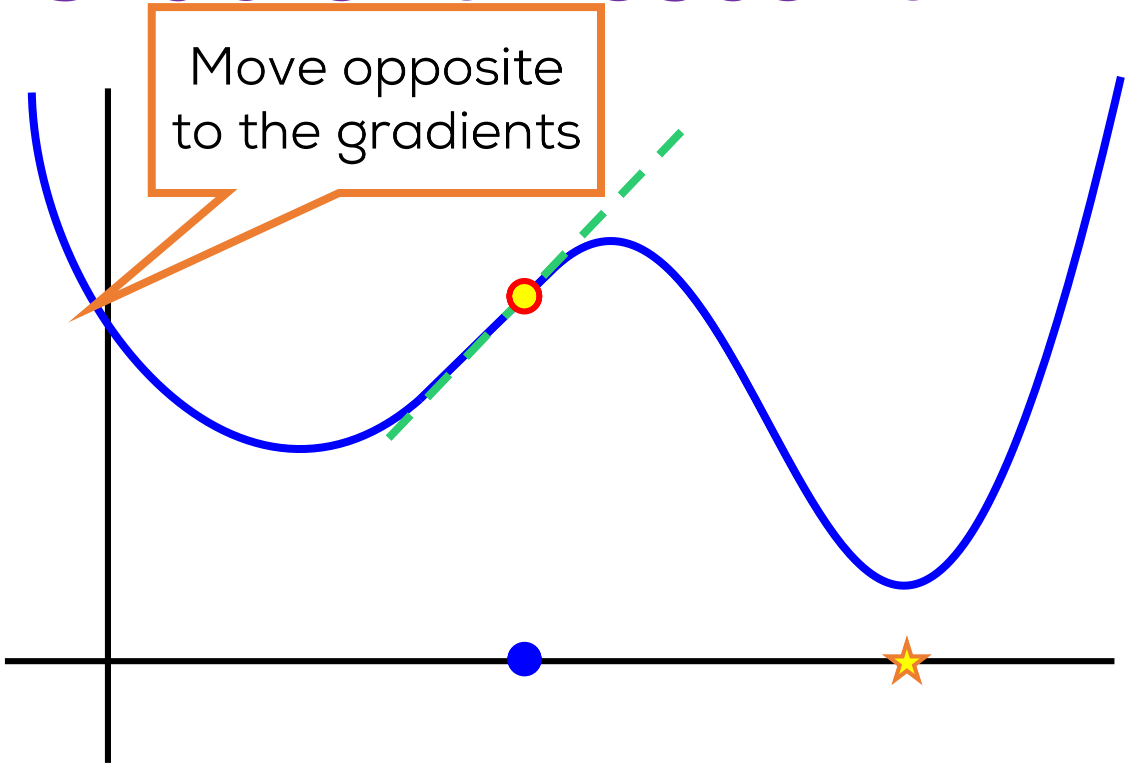
Gradient Descent



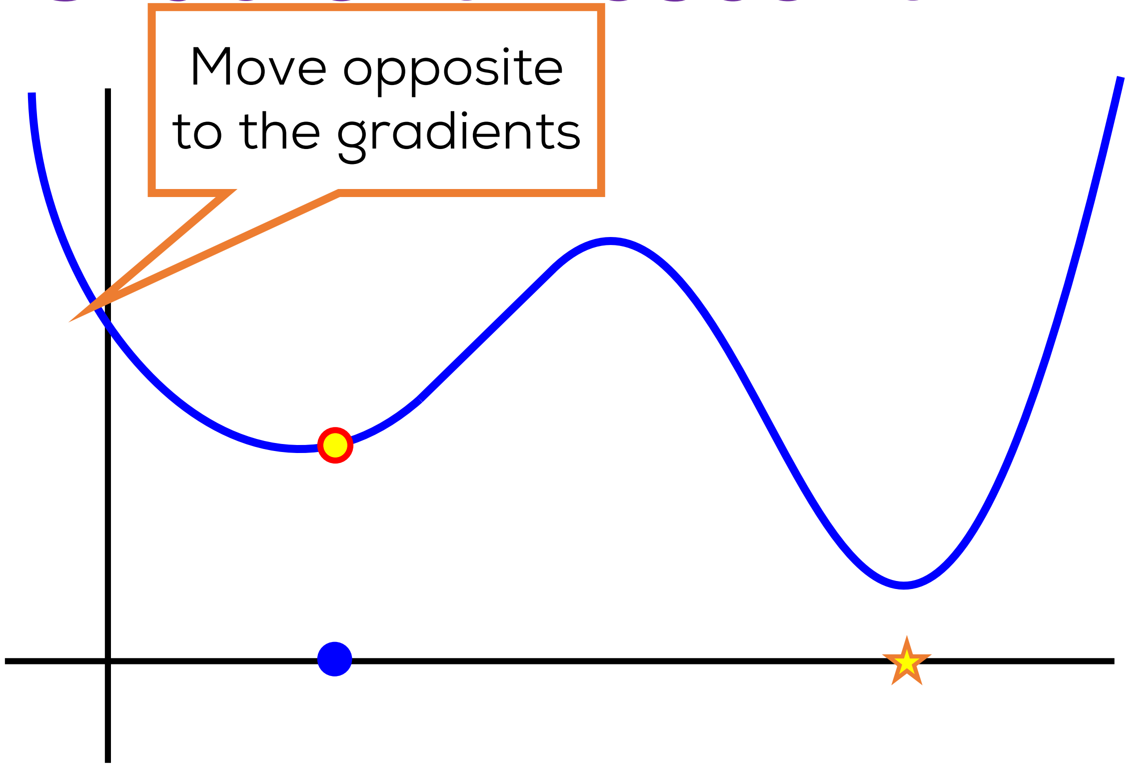
Gradient Descent



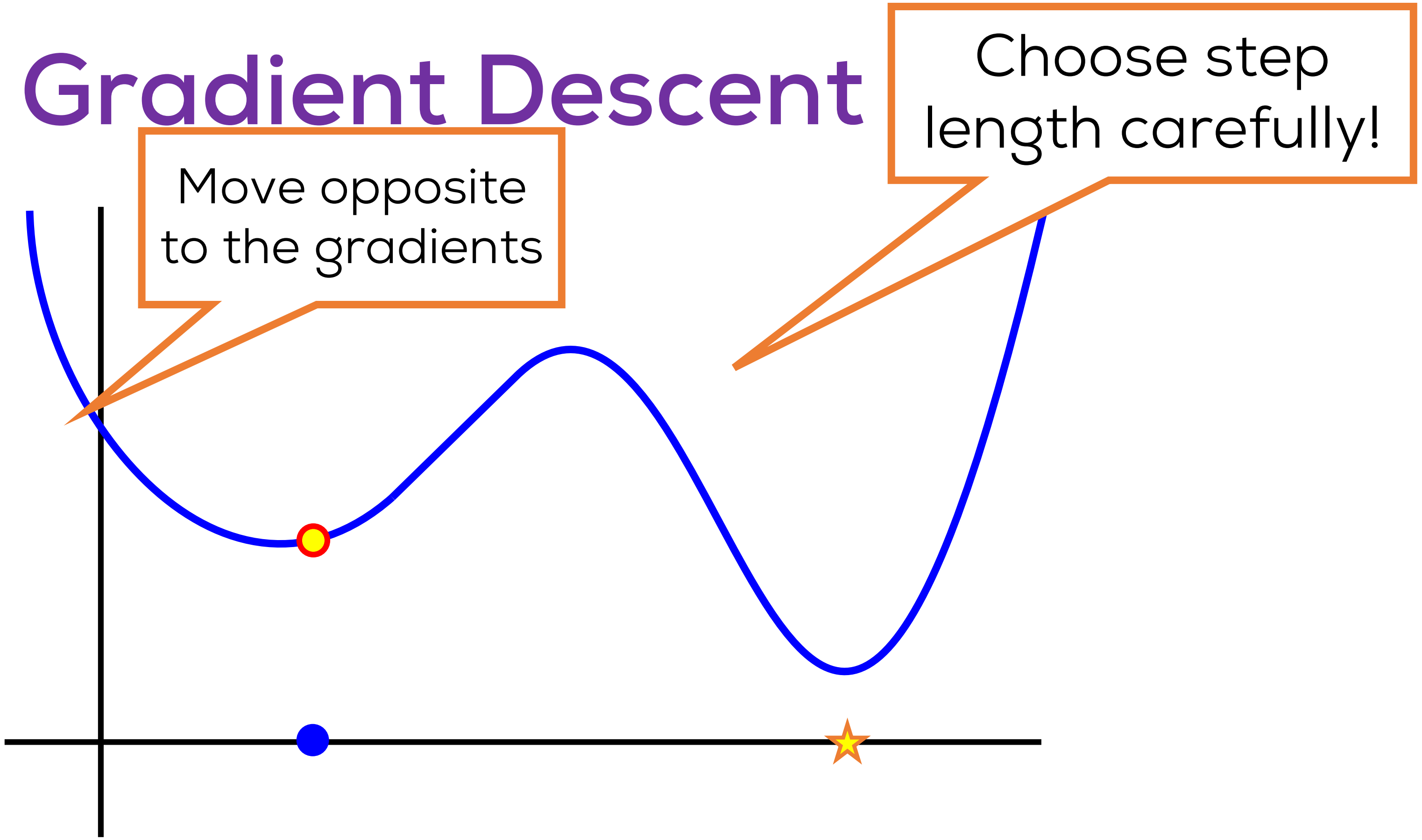
Gradient Descent



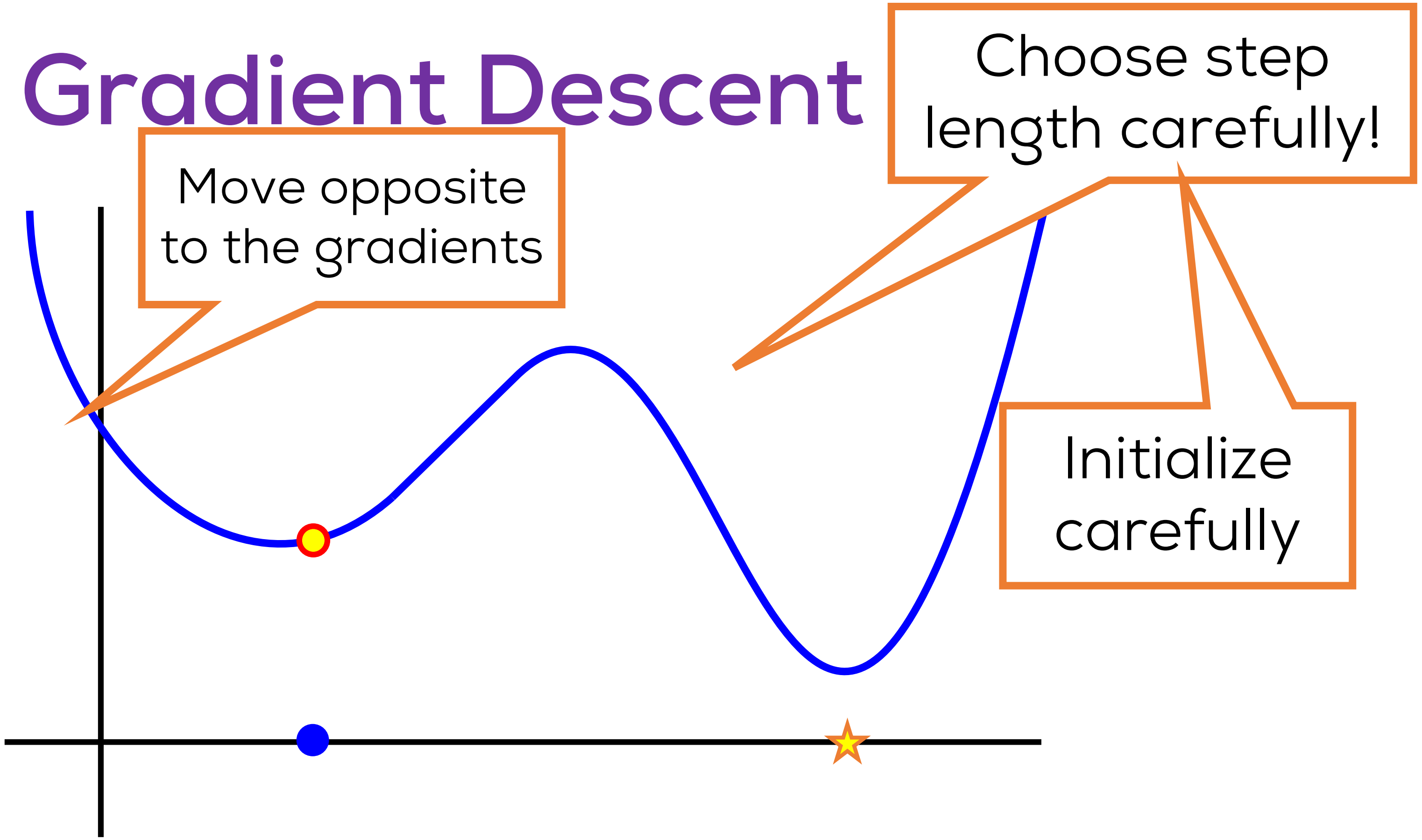
Gradient Descent



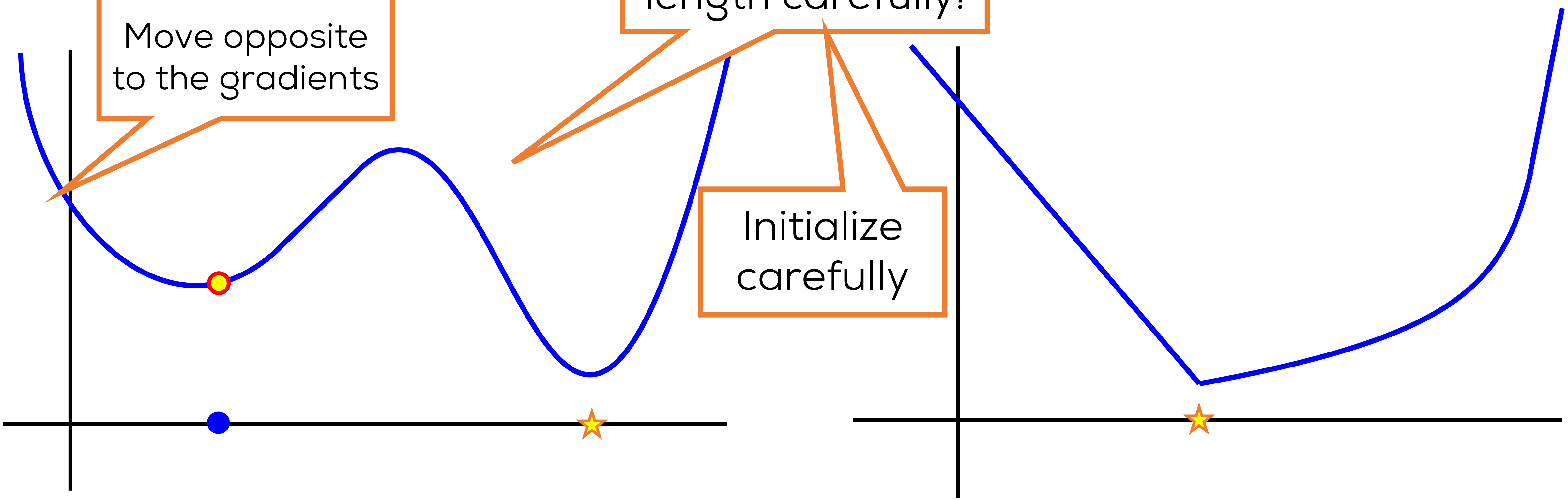
Gradient Descent



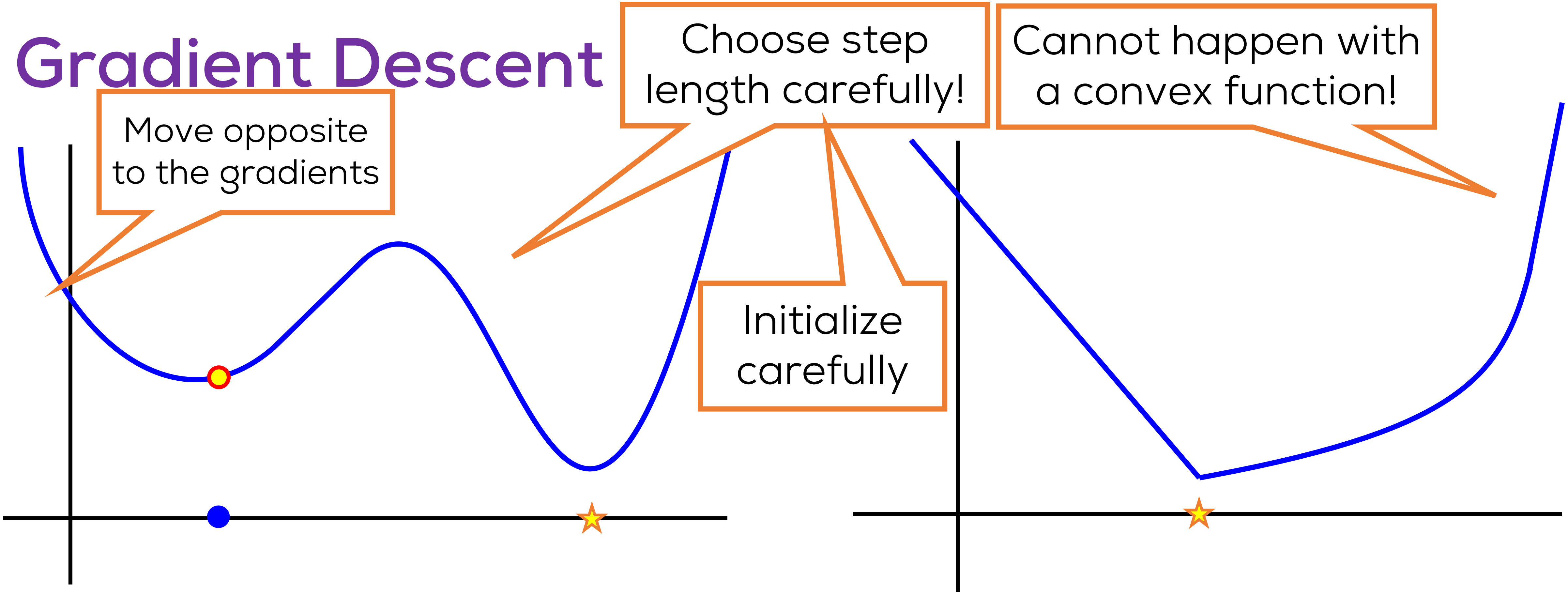
Gradient Descent



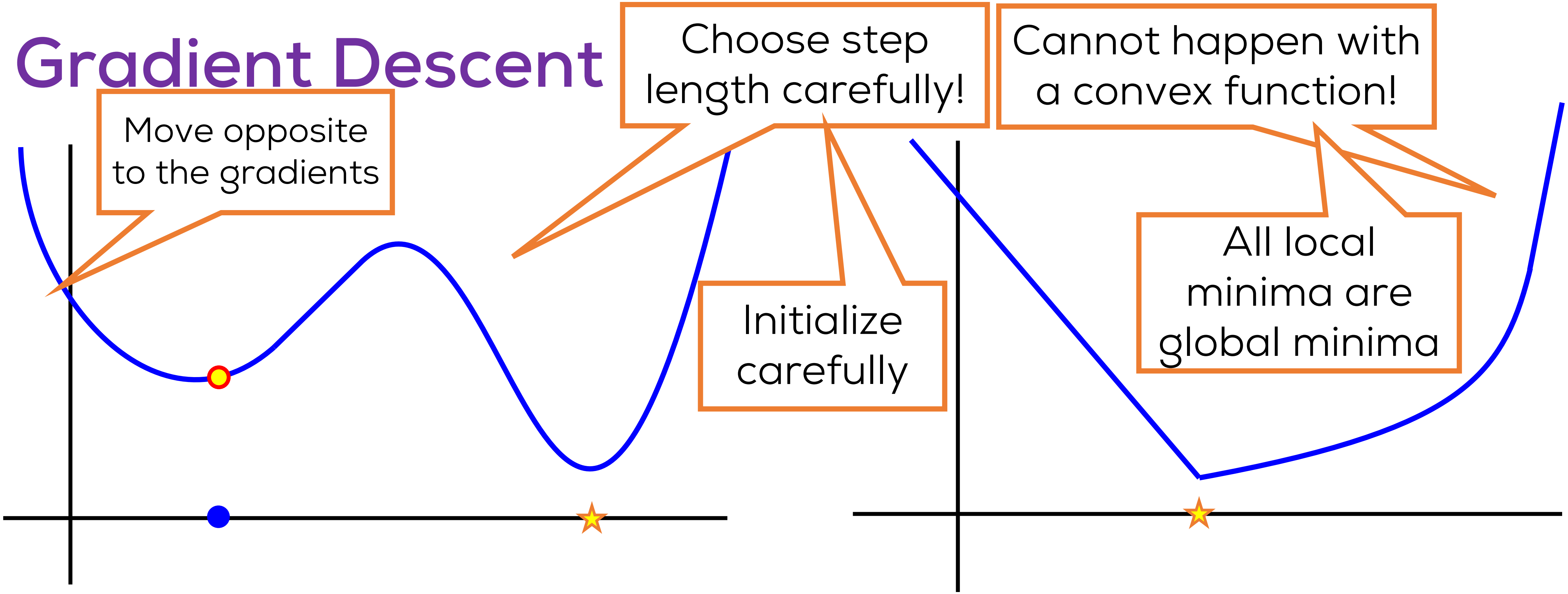
Gradient Descent



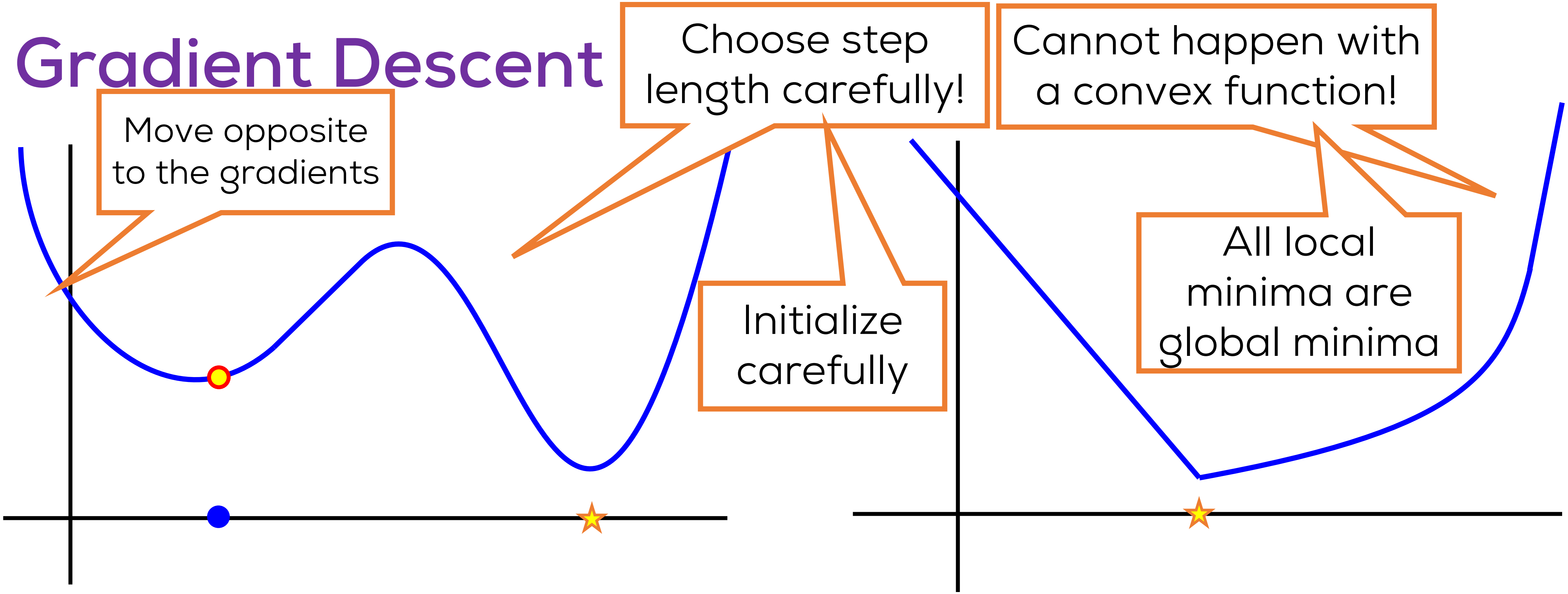
Gradient Descent



Gradient Descent



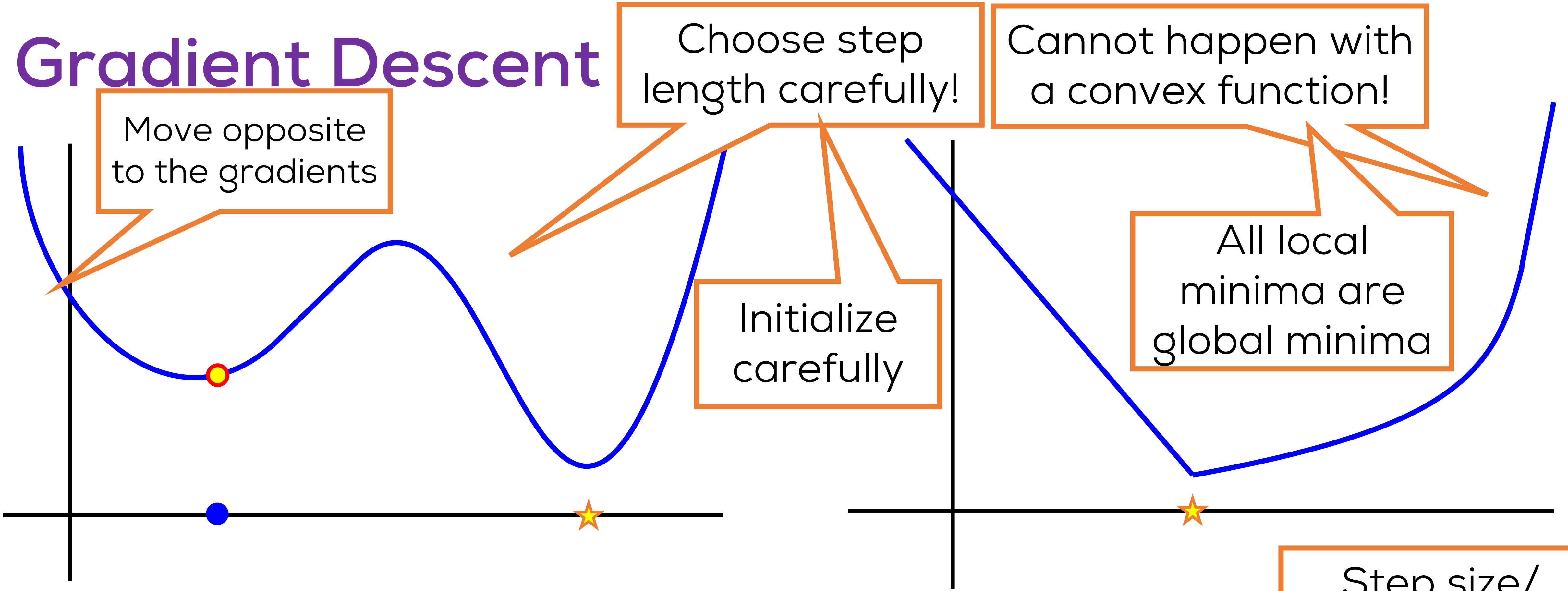
Gradient Descent



GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Gradient Descent

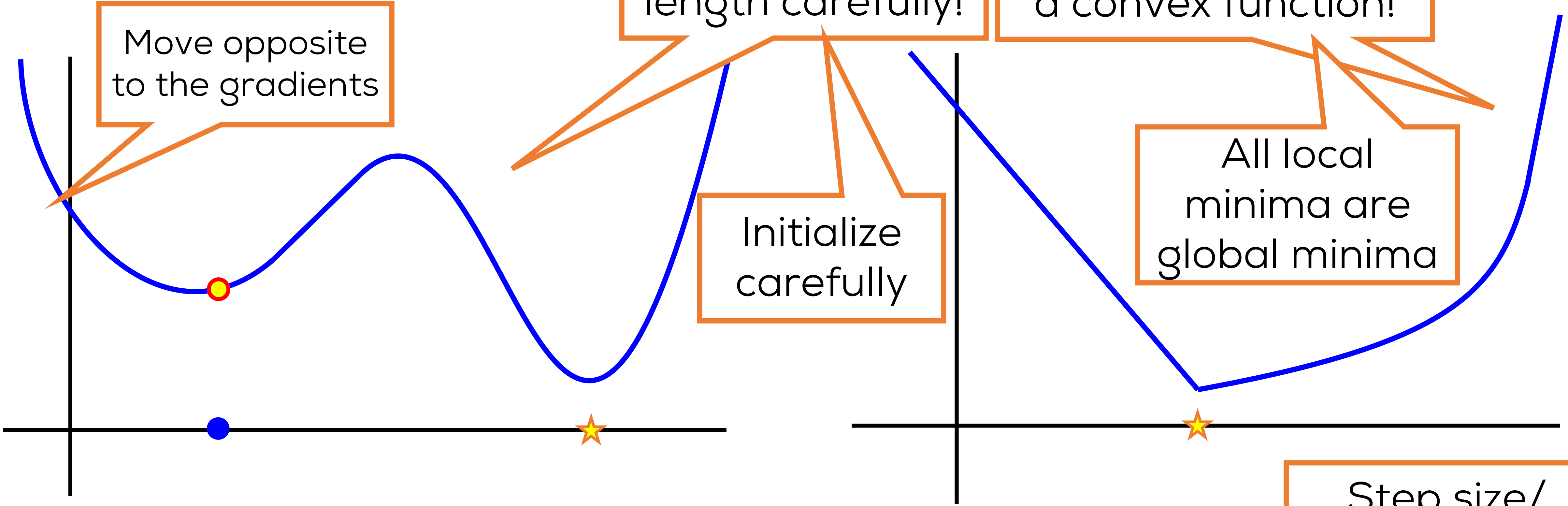


Step size/ learning rate

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Gradient Descent



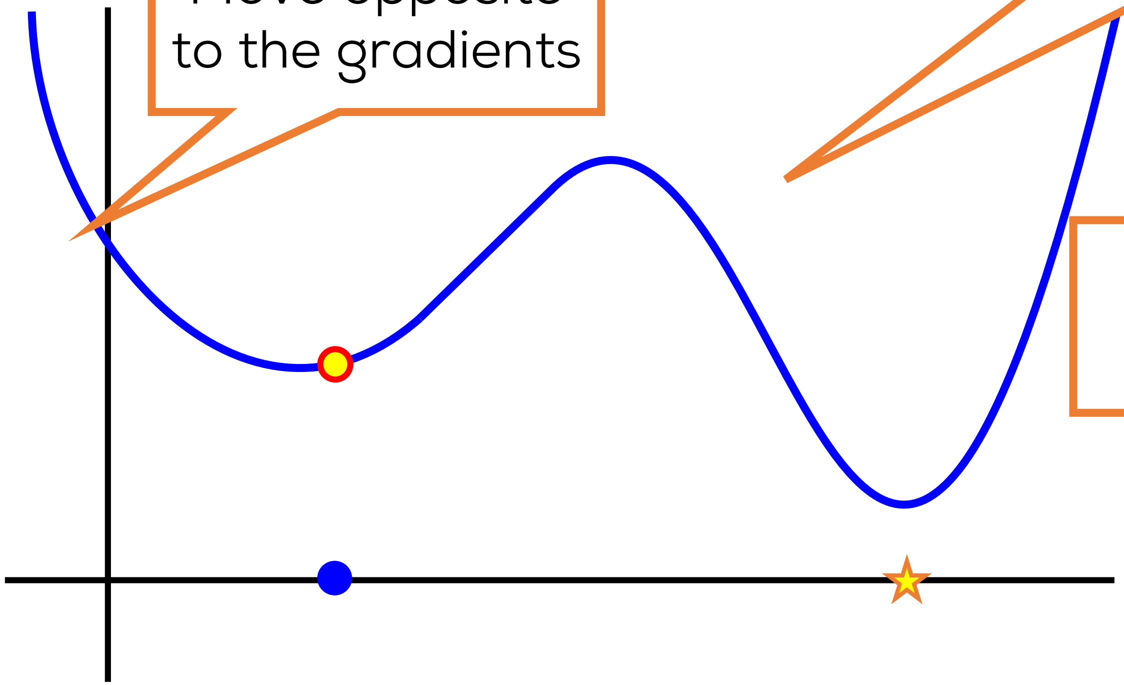
GRADIENT DESCENT

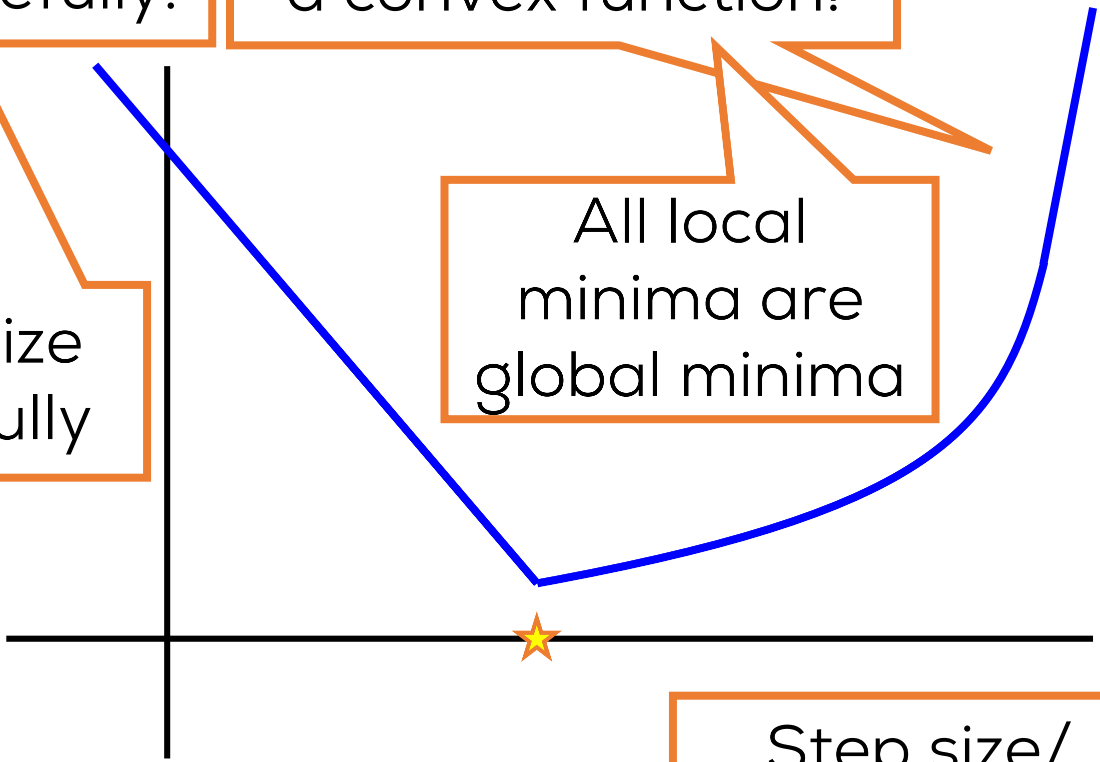
1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Step size/ learning rate

Tuned carefully CV, Adam, Adagrad

Gradient Descent





Many convergence criteria – length of gradient, performance threshold, dual criteria

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Step size/ learning rate

Tuned carefully with CV, Adam, Adagrad

Recap

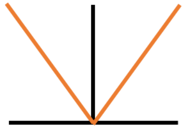
First-order Optimality

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) + r(\mathbf{w})$$

$$\nabla f(\mathbf{w}) + \nabla r(\mathbf{w}) = \mathbf{0}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2$$

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$

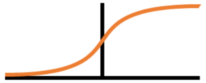


$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I\right)^{-1} \mathbf{X}\mathbf{y}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)\right)$$

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$

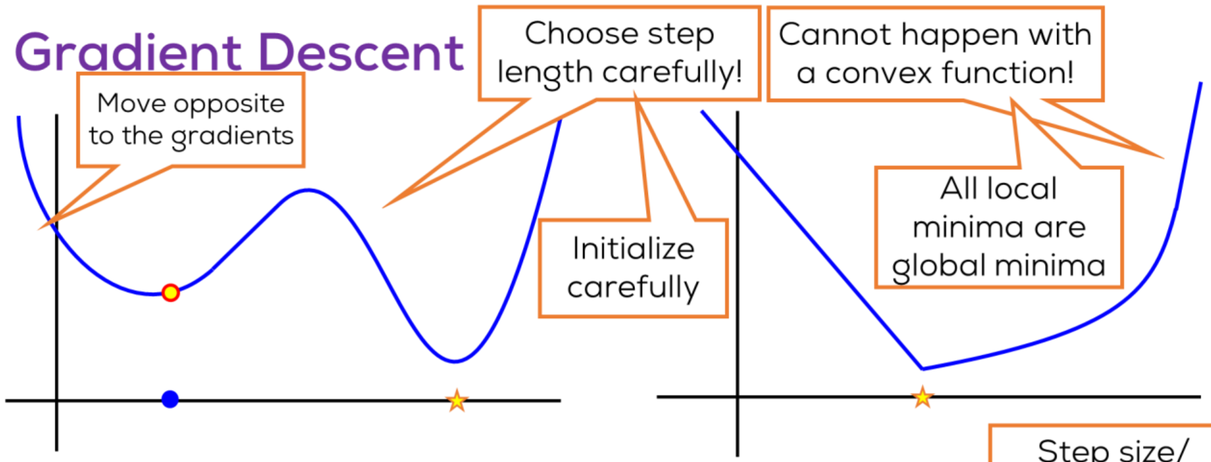
?



$$\sum_{i=1}^{n} (1 - \sigma(y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)) y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w} = \mathbf{0}$$

August 25, 2017 5

Gradient Descent



Many convergence criteria – length of gradient, performance threshold, dual criteria

Step size/ learning rate

Tuned carefully CV, Adam, Adagrad

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

August 25, 2017

Recap

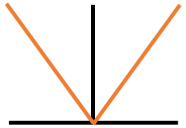
First-order Optimality

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) + r(\mathbf{w})$$

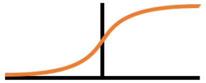
$$\nabla f(\mathbf{w}) + \nabla r(\mathbf{w}) = \mathbf{0}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 \qquad f(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)\right)$$

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$



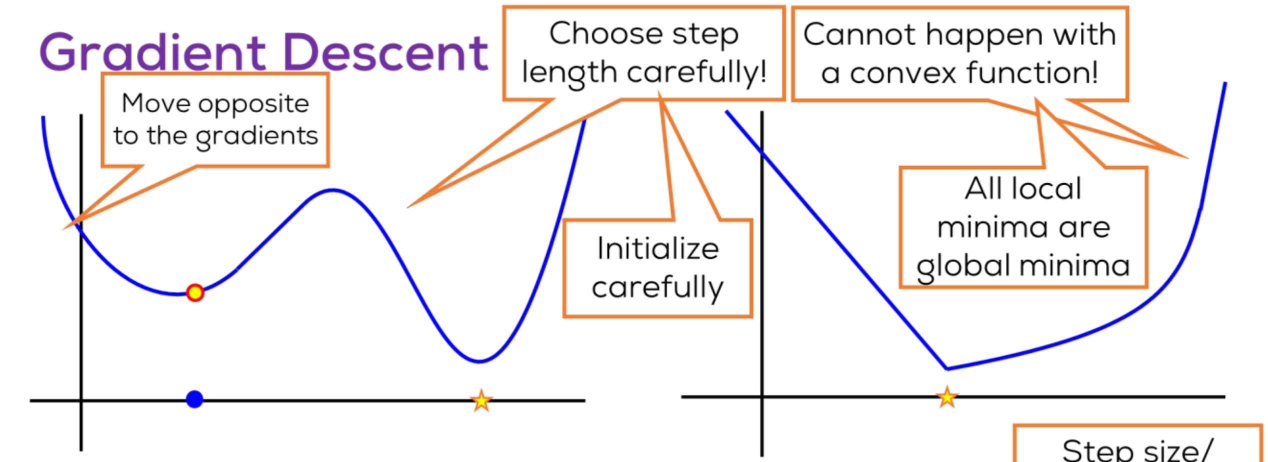
$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2 \qquad ?$$



$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I\right)^{-1} \mathbf{X}\mathbf{y} \qquad \sum_{i=1}^{n} (1 - \sigma(y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)) y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w} = \mathbf{0}$$

August 25, 2017 5

Gradient Descent



Many convergence criteria – length of gradient, performance threshold, dual criteria

GRADIENT DESCENT
1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Step size/ learning rate

Tuned carefully CV, Adam, Adagrad

August 25, 2017

What if function non-differentiable?

Recap

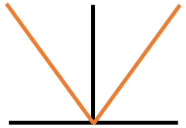
First-order Optimality

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) + r(\mathbf{w})$$

$$\nabla f(\mathbf{w}) + \nabla r(\mathbf{w}) = \mathbf{0}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \left(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)^2$$

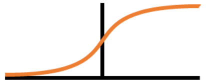
$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$



$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I\right)^{-1} \mathbf{X}\mathbf{y}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)\right)$$

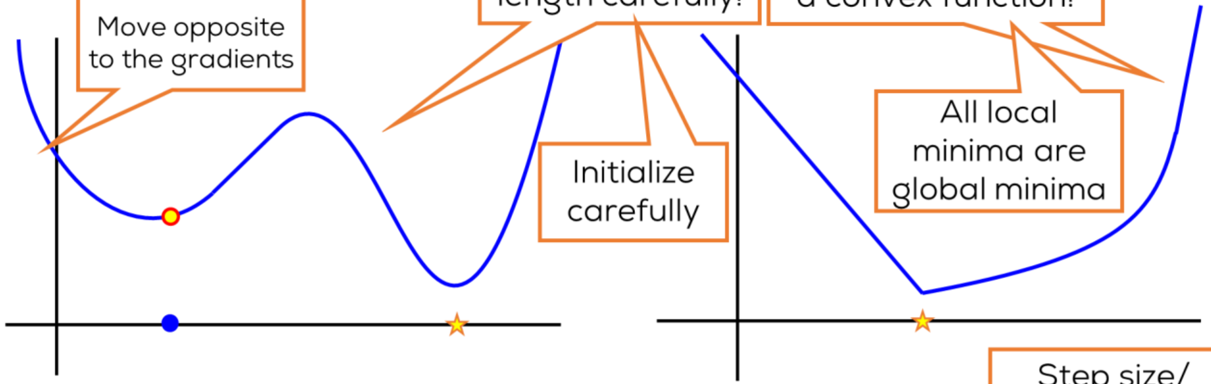
$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2 \quad ?$$



$$\sum_{i=1}^{n} (1 - \sigma(y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)) y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w} = \mathbf{0}$$

August 25, 2017 5

Gradient Descent



Many convergence criteria – length of gradient, performance threshold, dual criteria

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Step size/ learning rate

Tuned carefully CV, Adam, Adagrad

August 25, 2017

What if function non-differentiable?

Use subgradients!

Recap

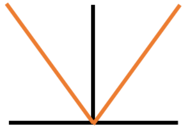
First-order Optimality

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) + r(\mathbf{w})$$

$$\nabla f(\mathbf{w}) + \nabla r(\mathbf{w}) = \mathbf{0}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2$$

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$

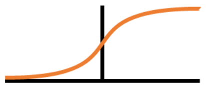


$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I\right)^{-1}\mathbf{X}\mathbf{y}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)\right)$$

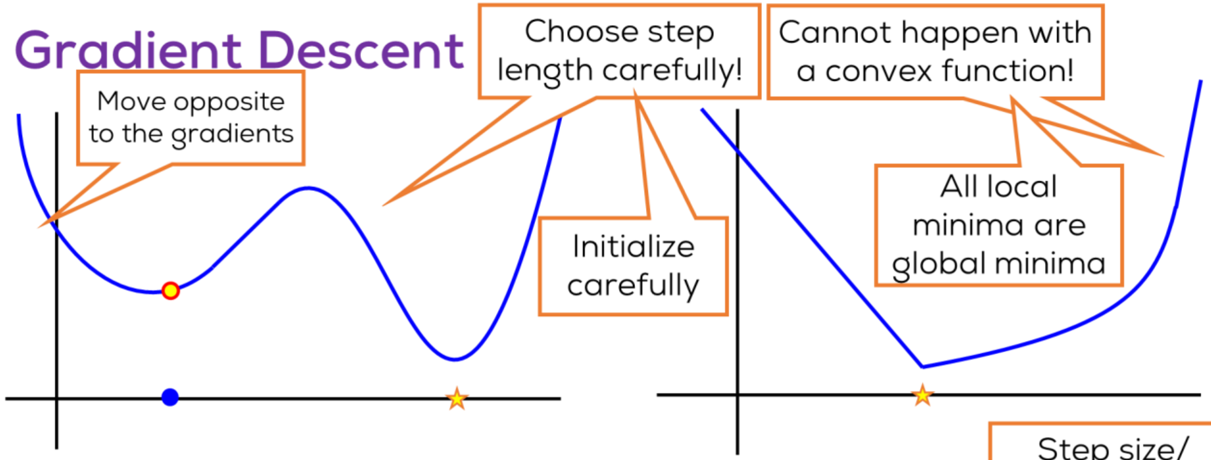
$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$

?



$$\sum_{i=1}^{n}(1 - \sigma(y^i \langle \mathbf{w}, \mathbf{x}^i \rangle))y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w} = \mathbf{0}$$

Gradient Descent



Many convergence criteria – length of gradient, performance threshold, dual criteria

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Step size/ learning rate

Tuned carefully CV, Adam, Adagrad

What if function non-differentiable?

Use subgradients!

Defined for convex functions

Recap

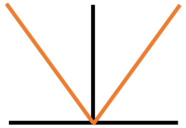
First-order Optimality

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min}\; f(\mathbf{w}) + r(\mathbf{w})$$

$$\nabla f(\mathbf{w}) + \nabla r(\mathbf{w}) = \mathbf{0}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \left(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)^2$$

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$

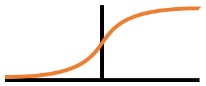


$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I\right)^{-1} \mathbf{X}\mathbf{y}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)\right)$$

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$

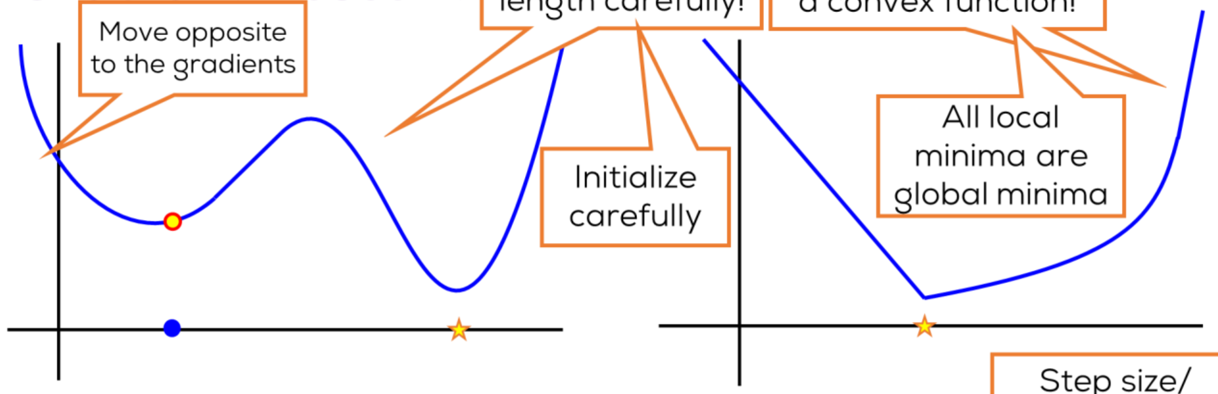
?



$$\sum_{i=1}^{n} (1 - \sigma(y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)) y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w} = \mathbf{0}$$

August 25, 2017 5

Gradient Descent



Many convergence criteria – length of gradient, performance threshold, dual criteria

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Step size/ learning rate

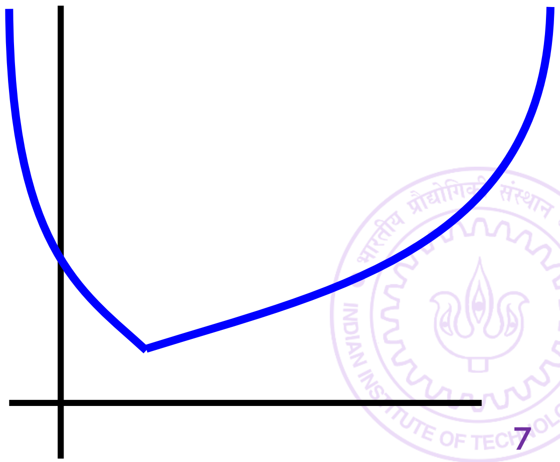
Tuned carefully CV, Adam, Adagrad

August 25, 2017

What if function non-differentiable?

Use subgradients!

Defined for convex functions



Recap

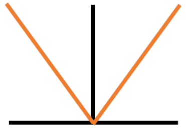
First-order Optimality

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}) + r(\mathbf{w})$$

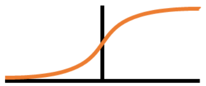
$$\nabla f(\mathbf{w}) + \nabla r(\mathbf{w}) = \mathbf{0}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 \qquad f(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)\right)$$

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$



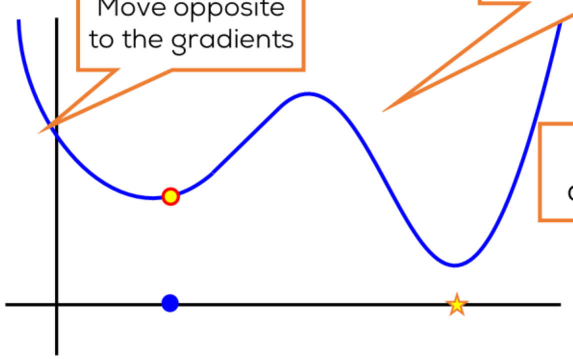
$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2 \qquad ?$$

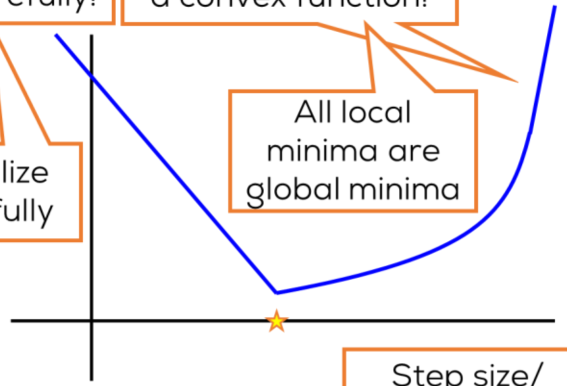


$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I\right)^{-1} \mathbf{X}\mathbf{y} \qquad \sum_{i=1}^{n} (1 - \sigma(y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)) y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w} = \mathbf{0}$$

August 25, 2017 5

Gradient Descent





Many convergence criteria – length of gradient, performance threshold, dual criteria

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Step size/ learning rate

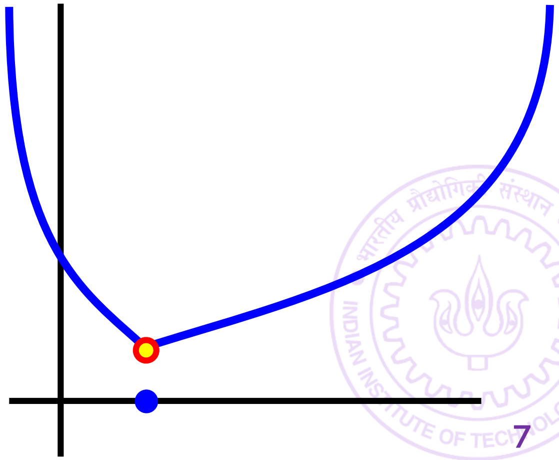
Tuned carefully CV, Adam, Adagrad

August 25, 2017

What if function non-differentiable?

Use subgradients!

Defined for convex functions



Recap

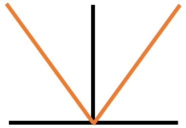
First-order Optimality

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) + r(\mathbf{w})$$

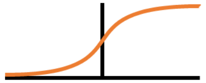
$$\nabla f(\mathbf{w}) + \nabla r(\mathbf{w}) = \mathbf{0}$$

$$f(\mathbf{w}) = \sum_{i=1}^{n} \left(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)^2 \qquad f(\mathbf{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)\right)$$

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$



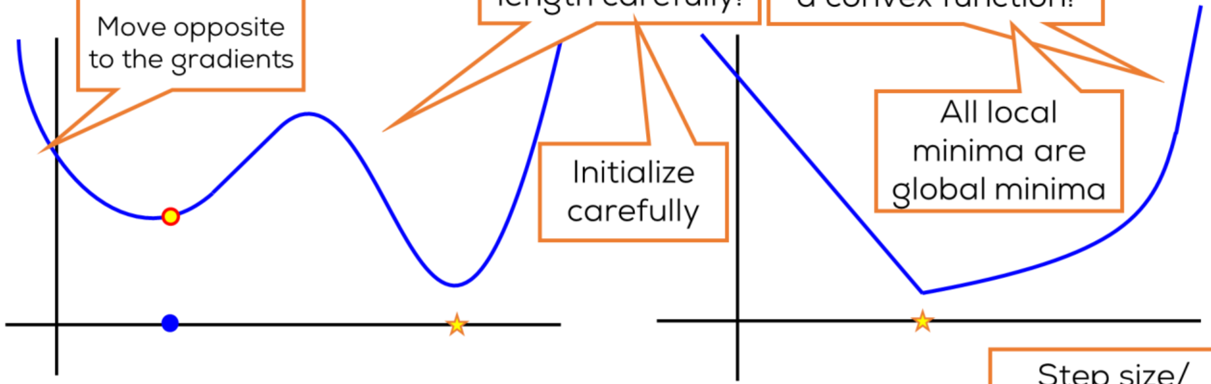
$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2 \qquad ?$$



$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I\right)^{-1} \mathbf{X}\mathbf{y} \qquad \sum_{i=1}^{n}(1 - \sigma(y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle))y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w} = \mathbf{0}$$

August 25, 2017 5

Gradient Descent



Many convergence criteria – length of gradient, performance threshold, dual criteria

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Step size/ learning rate

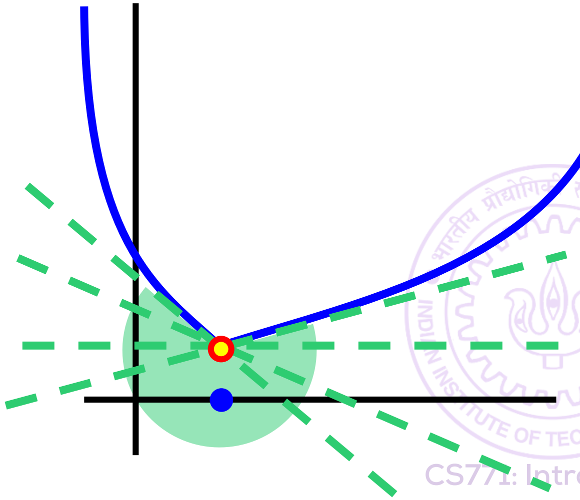
Tuned carefully CV, Adam, Adagrad

August 25, 2017

What if function non-differentiable?

Use subgradients!

Defined for convex functions



App: Linear Regression via GD

App: Linear Regression via GD

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

App: Linear Regression via

Convex?

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

App: Linear Regression via

Convex?

But but …
$(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

App: Linear Regression via

Convex?

But but …
$(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$

Expensive!
$O(d^3 + d^2 n)$ time

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

App: Linear Regression via

Convex?

But but …
$(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$

Expensive!
$O(d^3 + d^2 n)$ time

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

App: Linear Regression via

Convex?

But but …
$(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$

Expensive!
$O(d^3 + d^2 n)$ time

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

App: Linear Regression via

Convex?

But but …
$(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$

Expensive!
$O(d^3 + d^2 n)$ time

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$
$$= 2\sum_{i=1}^{n} \left(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i\right) \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w}^t$$

App: Linear Regression via

Convex?

But but …

$(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$

Expensive!

$O(d^3 + d^2 n)$ time

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Only $O(nd)$ time per iter!

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

$$= 2 \sum_{i=1}^{n} \left(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i\right) \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w}^t$$

App: Linear Regression via

Convex?

But but …
$(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$

Expensive!
$O(d^3 + d^2 n)$ time

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

Step length
$\eta_t = C, \frac{C}{t}, \frac{C}{\sqrt{t}}$

Only $O(nd)$
time per iter!

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$
$$= 2\sum_{i=1}^{n}\left(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i\right) \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w}^t$$

App: Linear Regression via

Convex?

But but …
$(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$

Expensive!
$O(d^3 + d^2 n)$ time

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

Step length
$\eta_t = C, \frac{C}{t}, \frac{C}{\sqrt{t}}$

Only $O(nd)$ time per iter!

$O\left(\frac{1}{\epsilon^2}\right)$ iterations suffice to reach ϵ-optimal solution

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$
$$= 2\sum_{i=1}^{n} \left(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i\right) \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w}^t$$

App: Linear Regression via

Convex?

But but …
$(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$

Expensive!
$O(d^3 + d^2 n)$ time

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

Step length
$\eta_t = C, \frac{C}{t}, \frac{C}{\sqrt{t}}$

Only $O(nd)$
time per iter!

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

$O\left(\frac{1}{\epsilon^2}\right)$ iterations suffice to reach ϵ-optimal solution

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

$$= 2\sum_{i=1}^{n} \left(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i\right) \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w}^t$$

$f(\mathbf{w}^t) + r(\mathbf{w}^t)$
$\leq f(\mathbf{w}^*) + r(\mathbf{w}^*) + \epsilon$

App: Linear Regression via

Convex?

But but …
$(\mathbf{X}\mathbf{X}^\top + \lambda \cdot I)^{-1}\mathbf{X}\mathbf{y}$

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_2^2$$

Expensive!
$\mathcal{O}(d^3 + d^2 n)$ time

Step length
$\eta_t = C, \frac{C}{t}, \frac{C}{\sqrt{t}}$

GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
3. Repeat until convergence

Only $\mathcal{O}(nd)$ time per iter!

$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ iterations suffice to reach ϵ-optimal solution

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$
$$= 2\sum_{i=1}^{n} \left(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i\right) \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w}^t$$

For convex problems

$f(\mathbf{w}^t) + r(\mathbf{w}^t)$
$\leq f(\mathbf{w}^*) + r(\mathbf{w}^*) + \epsilon$

A Closer look at the GD Update

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle > y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle = y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle < y^i$

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

Push $\mathbf{w}^t$ "away" from $\mathbf{x}^i$

If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle > y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle = y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle < y^i$

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

Push $\mathbf{w}^t$ "away" from $\mathbf{x}^i$

If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle > y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle = y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle < y^i$

No need to care about $\mathbf{x}^i$

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle > y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle = y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle < y^i$

Push $\mathbf{w}^t$ "away" from $\mathbf{x}^i$

No need to care about $\mathbf{x}^i$

Push $\mathbf{w}^t$ "towards" $\mathbf{x}^i$

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

Push $\mathbf{w}^t$ "away" from $\mathbf{x}^i$

If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle > y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle = y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle < y^i$

No need to care about $\mathbf{x}^i$

Push $\mathbf{w}^t$ "towards" $\mathbf{x}^i$

What if $\eta_t = 0$?
What if no r(·)?

A Closer look at the GD Update

$$\mathbf{w}^{t+1} = (1 - 2\lambda\eta_t) \cdot \mathbf{w}^t - 2\eta_t \cdot \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle > y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle = y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle < y^i$

Push $\mathbf{w}^t$ "away" from $\mathbf{x}^i$

No need to care about $\mathbf{x}^i$

Push $\mathbf{w}^t$ "towards" $\mathbf{x}^i$

What if $\eta_t = 0$?
What if no r(·)?

Exercise: for clarity, see what happens when $d = 1$

A Closer look at the GD Update

Lets see the case
$n = 1, \lambda = 0$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - 2\eta_t \cdot (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle > y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle = y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle < y^i$

Push $\mathbf{w}^t$ "away" from $\mathbf{x}^i$

No need to care about $\mathbf{x}^i$

Push $\mathbf{w}^t$ "towards" $\mathbf{x}^i$

What if $\eta_t = 0$?
What if no r(·)?

A Closer look at the GD Update

Lets see the case
$n = 1, \lambda = 0$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - 2\eta_t \cdot (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

$\langle w^{t+1}, x^i \rangle < \langle w^t, x^i \rangle$

Push $\mathbf{w}^t$ "away" from $\mathbf{x}^i$

If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle > y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle = y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle < y^i$

No need to care about $\mathbf{x}^i$

Push $\mathbf{w}^t$ "towards" $\mathbf{x}^i$

What if $\eta_t = 0$?
What if no r(·)?

A Closer look at the GD Update

Lets see the case $n = 1, \lambda = 0$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - 2\eta_t \cdot (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

$\langle w^{t+1}, x^i \rangle < \langle w^t, x^i \rangle$

$\langle w^{t+1}, x^i \rangle = \langle w^t, x^i \rangle$

Push $\mathbf{w}^t$ "away" from $\mathbf{x}^i$

No need to care about $\mathbf{x}^i$

Push $\mathbf{w}^t$ "towards" $\mathbf{x}^i$

If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle > y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle = y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle < y^i$

What if $\eta_t = 0$?
What if no r(·)?

A Closer look at the GD Update

Lets see the case $n = 1, \lambda = 0$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - 2\eta_t \cdot (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

$\langle w^{t+1}, x^i \rangle < \langle w^t, x^i \rangle$

Push $\mathbf{w}^t$ "away" from $\mathbf{x}^i$

$\langle w^{t+1}, x^i \rangle = \langle w^t, x^i \rangle$

No need to care about $\mathbf{x}^i$

$\langle w^{t+1}, x^i \rangle > \langle w^t, x^i \rangle$

Push $\mathbf{w}^t$ "towards" $\mathbf{x}^i$

If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle > y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle = y^i$
If $\langle \mathbf{w}^t, \mathbf{x}^i \rangle < y^i$

What if $\eta_t = 0$?
What if no r(·)?

A Closer look at the GD Update

Lets see the case $n = 1, \lambda = 0$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - 2\eta_t \cdot \left(\left\langle \mathbf{w}^t, \mathbf{x}^i \right\rangle - y^i\right) \cdot \mathbf{x}^i$$

Does not let model change too much!

What if $\lambda = 0$?
What if $\eta_t = 0$?
What if no $f(\cdot)$?

If $\left\langle \mathbf{w}^t, \mathbf{x}^i \right\rangle > y^i$
If $\left\langle \mathbf{w}^t, \mathbf{x}^i \right\rangle = y^i$
If $\left\langle \mathbf{w}^t, \mathbf{x}^i \right\rangle < y^i$

Push $\mathbf{w}^t$ "away" from $\mathbf{x}^i$

$\left\langle w^{t+1}, x^i \right\rangle < \left\langle w^t, x^i \right\rangle$

No need to care about $\mathbf{x}^i$

$\left\langle w^{t+1}, x^i \right\rangle = \left\langle w^t, x^i \right\rangle$

Push $\mathbf{w}^t$ "towards" $\mathbf{x}^i$

$\left\langle w^{t+1}, x^i \right\rangle > \left\langle w^t, x^i \right\rangle$

What if $\eta_t = 0$?
What if no r(·)?

We are making corrective updates!

Subgradient Calculus

Do not always hold ☹

Should hold in nice settings ☺

	MTH101 Calculus	Subgradient Calculus

$$\mathbf{x} \in \mathbb{R}^d, \mathbf{A} \in \mathbb{R}^{k\times d}, \mathbf{b} \in \mathbb{R}^k, c \in \mathbb{R}$$

	MTH101 Calculus	Subgradient Calculus
Scaling	• $\nabla(c \cdot f)(\mathbf{x}) = c \cdot \nabla f(\mathbf{x})$	• $\partial(c \cdot f)(\mathbf{x}) = c \cdot \partial f(\mathbf{x}) = \{c \cdot \mathbf{v} : \mathbf{v} \in \partial f(\mathbf{x})\}$
Sum Rule	• $\nabla(f+g)(x) = \nabla f(x) + \nabla g(x)$	• $\partial(f+g)(\mathbf{x}) = \partial f(\mathbf{x}) + \partial g(\mathbf{x}) = \{\mathbf{u}+\mathbf{v} : \mathbf{u} \in \partial f(\mathbf{x}), \mathbf{v} \in \partial g(\mathbf{x})\}$
Chain Rule	• $\nabla f(\mathbf{Ax}+\mathbf{b}) = \mathbf{A}^\top \nabla f(\mathbf{Ax}+\mathbf{b})$	• $\partial f(\mathbf{Ax}+\mathbf{b}) = \mathbf{A}^\top \partial f(\mathbf{Ax}+\mathbf{b}) = \{\mathbf{A}^\top \mathbf{v} : \mathbf{v} \in \partial f(\mathbf{x})\}$

Max Rule

$$\ell_{\text{cs}}(y, \{\boldsymbol{\eta}_j\}) = [1 + \max_{k \neq y} \boldsymbol{\eta}_k - \boldsymbol{\eta}_y]_+$$

$$f(\mathbf{x}) = \max_i f_i(\mathbf{x})$$

$$\partial f(\mathbf{x}) = \left\{\sum \alpha_i \mathbf{v}^i : \mathbf{v}^i \in \partial f_i(\mathbf{x}), \sum \alpha_i = 1, \alpha_i \geq 0\right\}$$

Subgradient Calculus

Do not always hold ☹

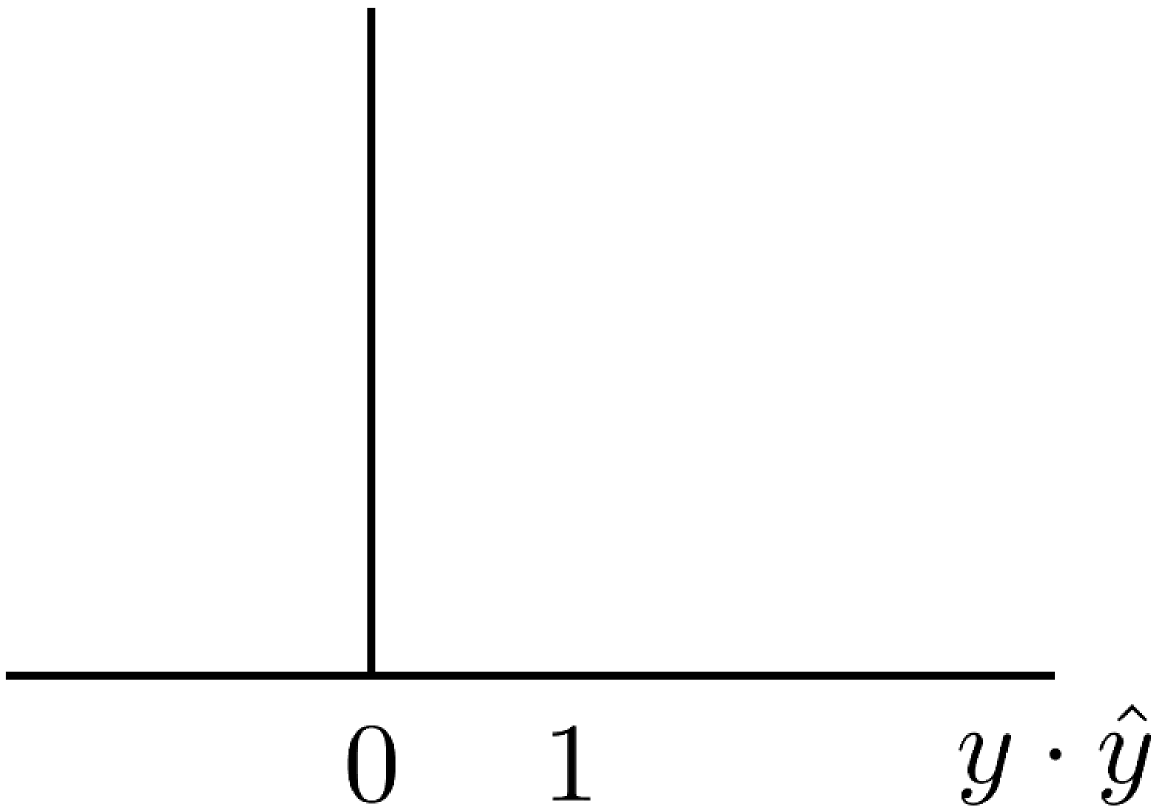
Should hold in nice settings ☺

	MTH101 Calculus	Subgradient Calculus
	$\mathbf{x} \in \mathbb{R}^d, \mathbf{A} \in \mathbb{R}^{k \times d}, \mathbf{b} \in \mathbb{R}^k, c \in \mathbb{R}$	
Scaling	• $\nabla(\mathrm{c} \cdot f)(\mathbf{x}) = c \cdot \nabla f(\mathbf{x})$	• $\partial(c \cdot f)(\mathbf{x}) = c \cdot \partial f(\mathbf{x}) = \{c \cdot \mathbf{v} : \mathbf{v} \in \partial f(\mathbf{x})\}$
Sum Rule	• $\nabla(f+g)(x) = \nabla f(x) + \nabla g(x)$	• $\partial(f+g)(\mathbf{x}) = \partial f(\mathbf{x}) + \partial g(\mathbf{x}) = \{\mathbf{u}+\mathbf{v} : \mathbf{u} \in \partial f(\mathbf{x}), \mathbf{v} \in \partial g(\mathbf{x})\}$
Chain Rule	• $\nabla f(\mathbf{Ax}+\mathbf{b}) = \mathbf{A}^\top \nabla f(\mathbf{Ax}+\mathbf{b})$	• $\partial f(\mathbf{Ax}+\mathbf{b}) = \mathbf{A}^\top \partial f(\mathbf{Ax}+\mathbf{b}) = \{\mathbf{A}^\top \mathbf{v} : \mathbf{v} \in \partial f(\mathbf{x})\}$
Max Rule	?	$f(\mathbf{x}) = \max_i f_i(\mathbf{x})$

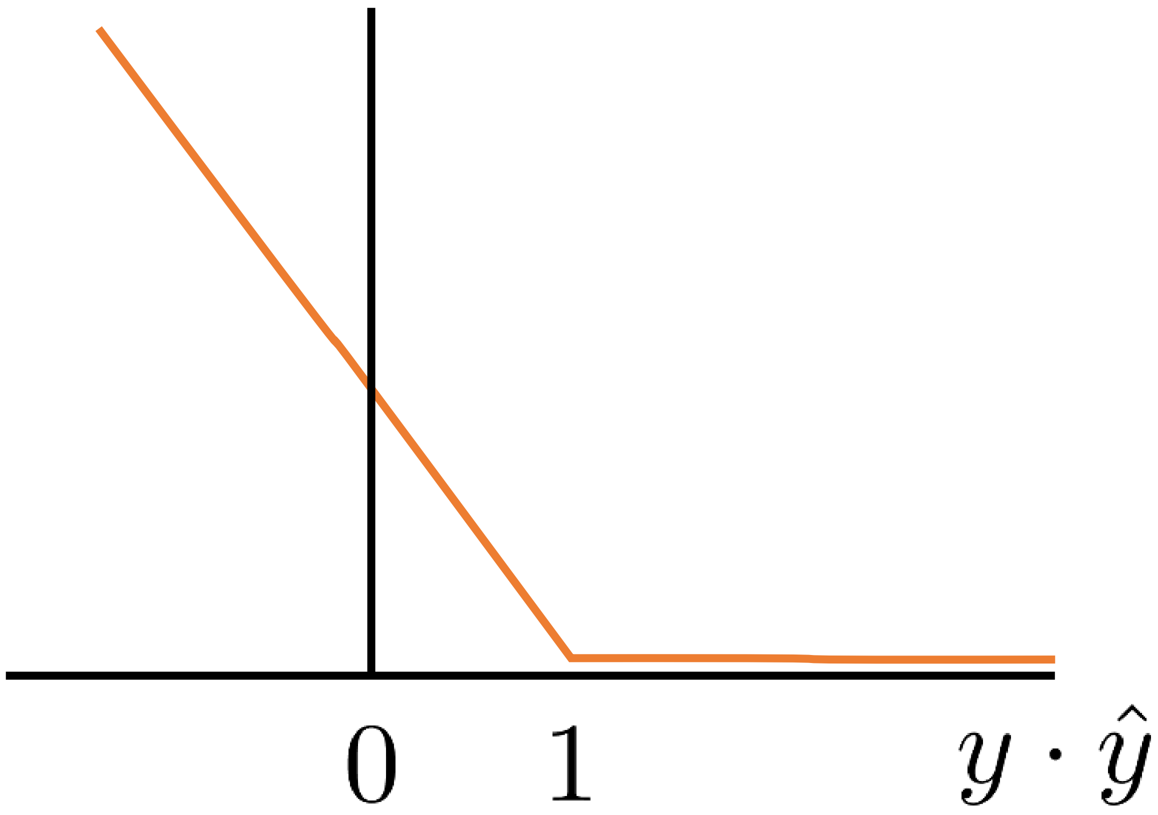
$$\partial f(\mathbf{x}) = \left\{\sum \alpha_i \mathbf{v}^i : \mathbf{v}^i \in \partial f_i(\mathbf{x}), \sum \alpha_i = 1, \alpha_i \geq 0\right\}$$

Application: SVM via subGD

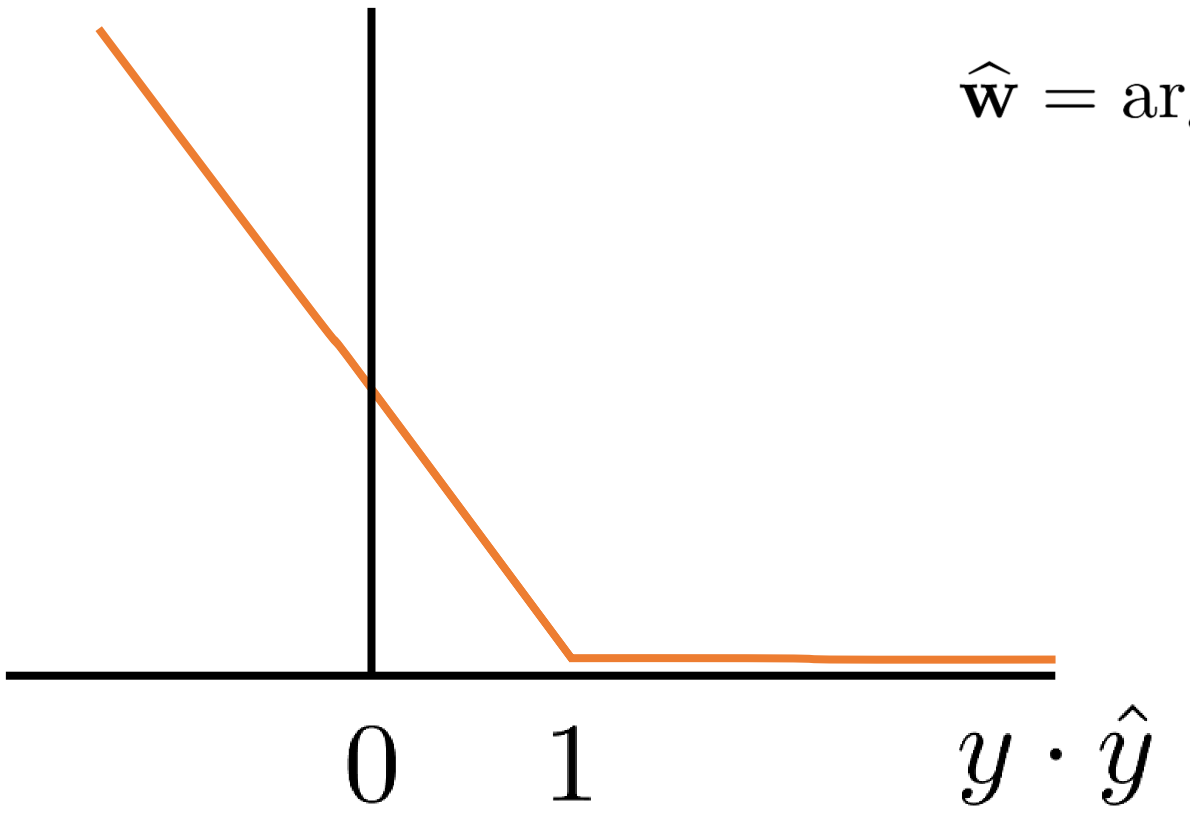
Application: SVM via subGD



Application: SVM via subGD

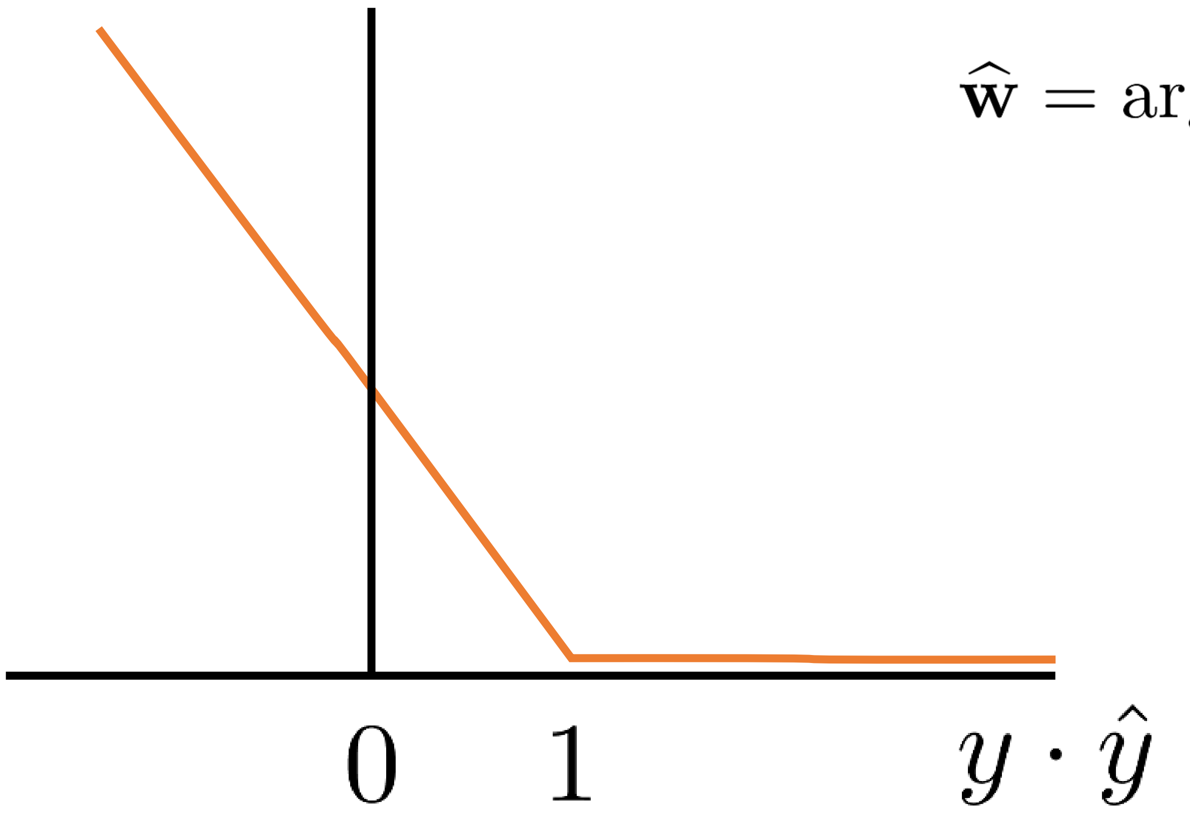


Application: SVM via subGD



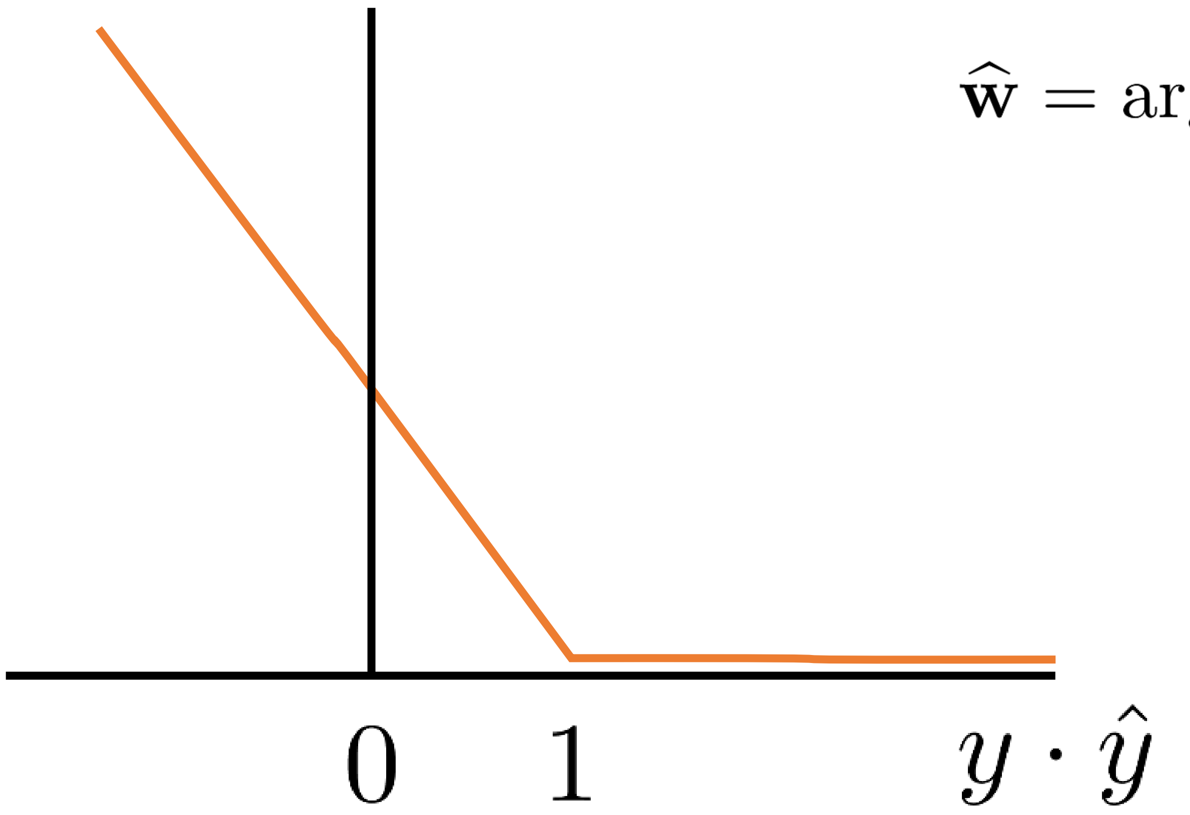
$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

Application: SVM via subGD



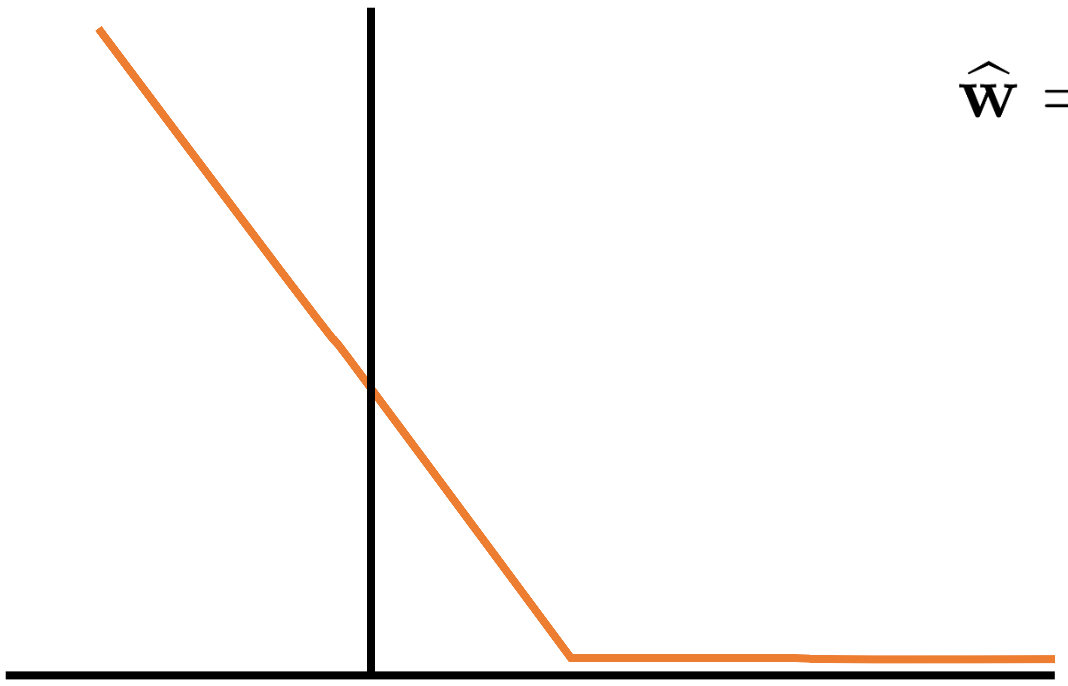
$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\sum_{i=1}^{n} \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)}_{f(\mathbf{w})} + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

Application: SVM via subGD



$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\sum_{i=1}^{n} \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)}_{f(\mathbf{w})} + \underbrace{\frac{\lambda}{2} \|\mathbf{w}\|_2^2}_{r(\mathbf{w})}$$

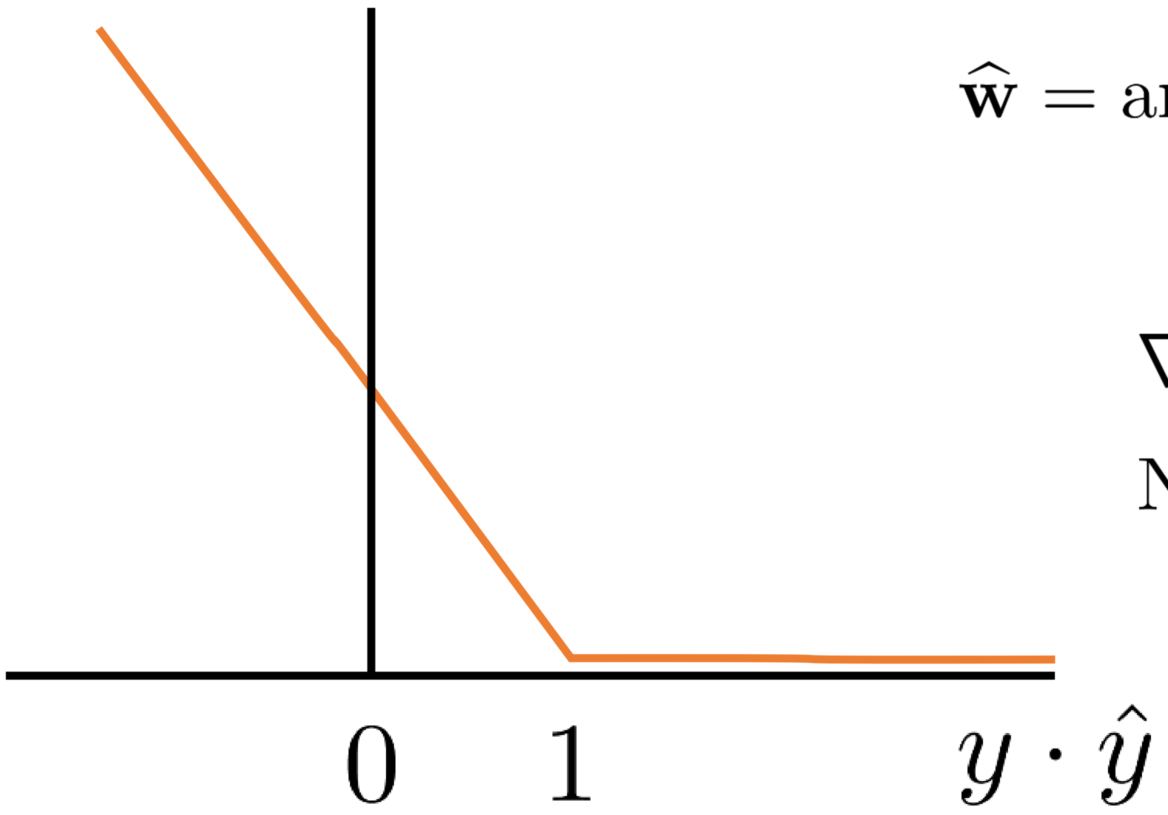
Application: SVM via subGD



$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\sum_{i=1}^{n} \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)}_{f(\mathbf{w})} + \underbrace{\frac{\lambda}{2} \|\mathbf{w}\|_2^2}_{r(\mathbf{w})}$$

$$\nabla r(\mathbf{w}) = \lambda \cdot \mathbf{w}$$

Application: SVM via subGD

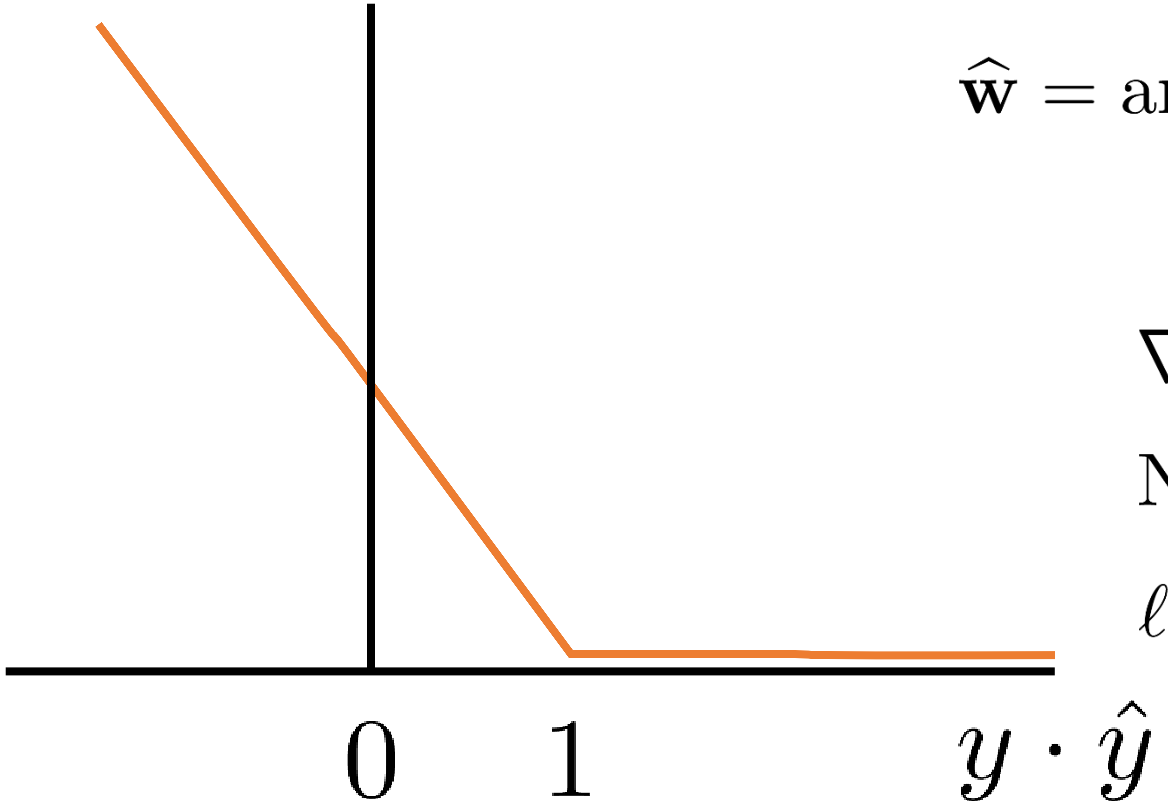


$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\sum_{i=1}^{n} \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)}_{f(\mathbf{w})} + \underbrace{\frac{\lambda}{2} \|\mathbf{w}\|_2^2}_{r(\mathbf{w})}$$

$$\nabla r(\mathbf{w}) = \lambda \cdot \mathbf{w}$$

$$\text{Need } \mathbf{v}^i \in \partial \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Application: SVM via subGD



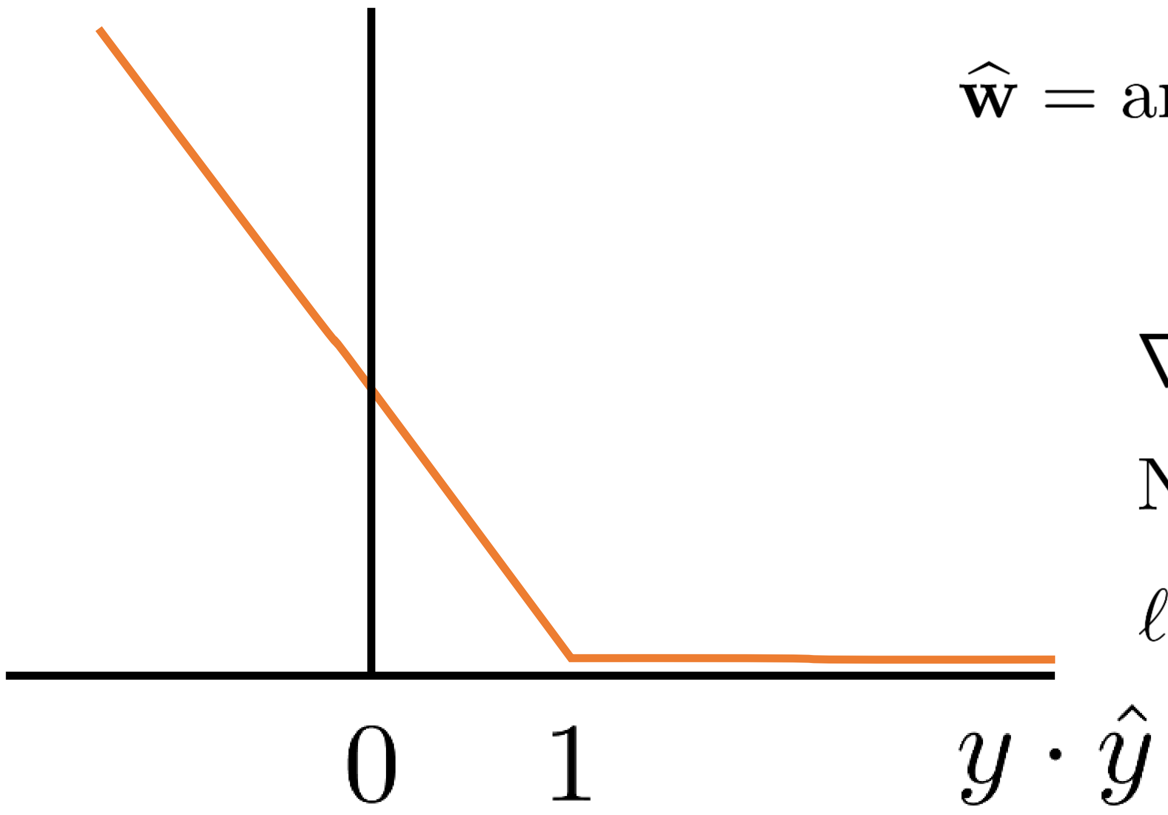
$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\sum_{i=1}^{n} \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)}_{f(\mathbf{w})} + \underbrace{\frac{\lambda}{2} \|\mathbf{w}\|_2^2}_{r(\mathbf{w})}$$

$$\nabla r(\mathbf{w}) = \lambda \cdot \mathbf{w}$$

Need $\mathbf{v}^i \in \partial \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$

$$\ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle) = [1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle]_+$$

Application: SVM via subGD



$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\sum_{i=1}^{n} \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)}_{f(\mathbf{w})} + \underbrace{\frac{\lambda}{2} \|\mathbf{w}\|_2^2}_{r(\mathbf{w})}$$

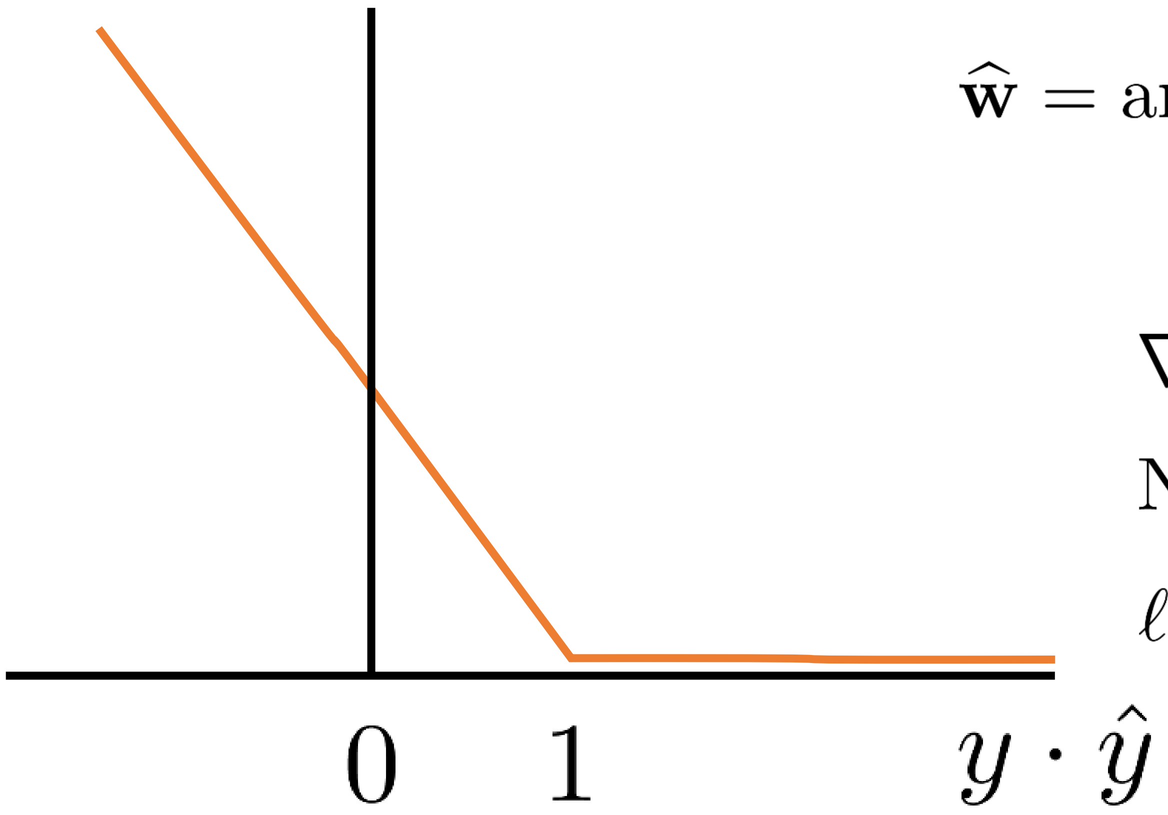
$$\nabla r(\mathbf{w}) = \lambda \cdot \mathbf{w}$$

Need $\mathbf{v}^i \in \partial \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$

$$\ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle) = [1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle]_+$$

$$\mathbf{v}^i = \begin{cases} \mathbf{0} & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle > 1 \\ -y^i \cdot \mathbf{x}^i & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle < 1 \\ c \cdot y^i \cdot \mathbf{x}^i & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle = 1 \\ c \in [-1, 0] & \end{cases}$$

Application: SVM via subGD



$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\sum_{i=1}^{n} \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)}_{f(\mathbf{w})} + \underbrace{\frac{\lambda}{2} \|\mathbf{w}\|_2^2}_{r(\mathbf{w})}$$

$$\nabla r(\mathbf{w}) = \lambda \cdot \mathbf{w}$$

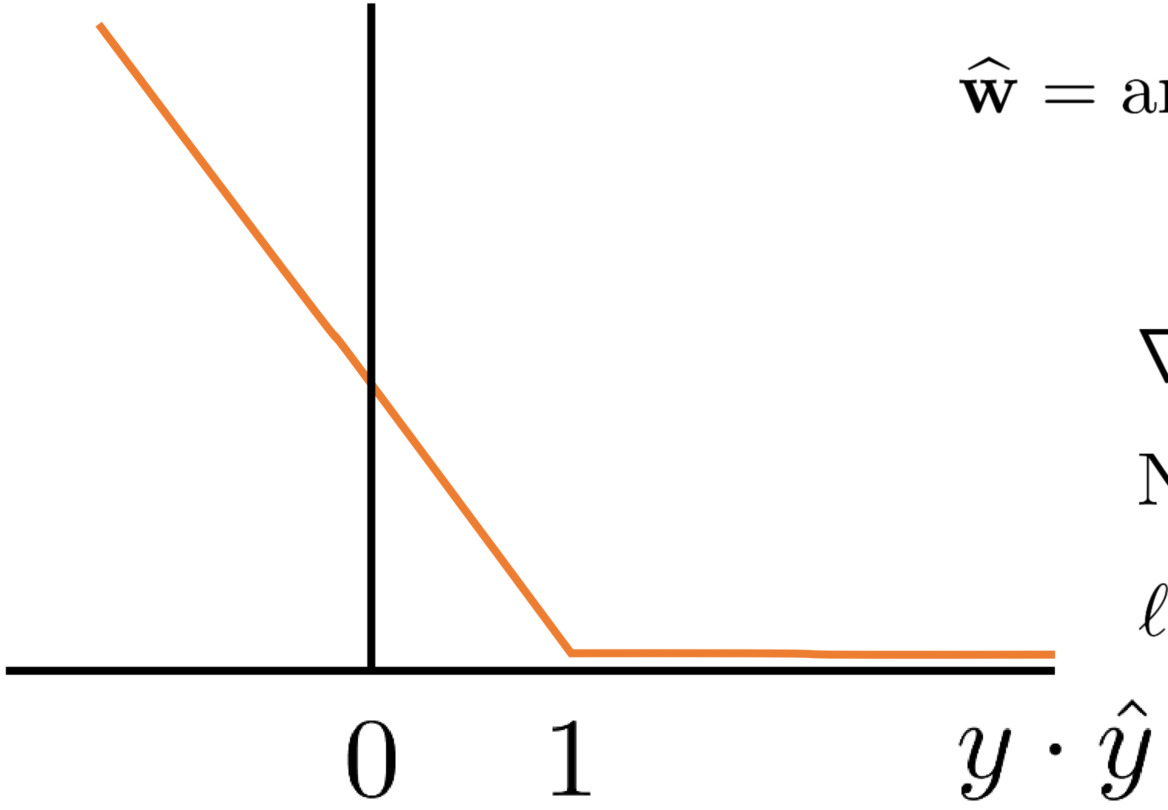
Need $\mathbf{v}^i \in \partial \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$

$$\ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle) = [1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle]_+$$

$$\mathbf{v}^i = \begin{cases} \mathbf{0} & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle > 1 \\ -y^i \cdot \mathbf{x}^i & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle < 1 \\ c \cdot y^i \cdot \mathbf{x}^i & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle = 1 \\ c \in [-1, 0] & \end{cases}$$

$O(d)$ time per data point!

Application: SVM via subGD



$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \underbrace{\sum_{i=1}^{n} \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)}_{f(\mathbf{w})} + \underbrace{\frac{\lambda}{2} \|\mathbf{w}\|_2^2}_{r(\mathbf{w})}$$

$$\nabla r(\mathbf{w}) = \lambda \cdot \mathbf{w}$$

Need $\mathbf{v}^i \in \partial \ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$

$$\ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle) = [1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle]_+$$

$$\mathbf{v}^i = \begin{cases} \mathbf{0} & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle > 1 \\ -y^i \cdot \mathbf{x}^i & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle < 1 \\ c \cdot y^i \cdot \mathbf{x}^i & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle = 1 \\ c \in [-1, 0] & \end{cases}$$

$O(nd)$ time per iter

$O(d)$ time per data point!

App: Sparse Regression via subGD

App: Sparse Regression via subGD

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_1$$

App: Sparse Regression via sul

Convex?

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_1$$

App: Sparse Regression via sul

Convex?

Non-differentiable

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_1$$

$$\mathbf{g}^t \in \nabla f(\mathbf{w}) + \partial r(\mathbf{w})$$

App: Sparse Regression via sul

Convex?

Non-differentiable

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_1$$

$$\mathbf{g}^t \in \nabla f(\mathbf{w}) + \partial r(\mathbf{w})$$

$$\mathbf{g}^t = 2\sum_{i=1}^{n}(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i + \lambda \cdot \boldsymbol{\rho}^t$$

App: Sparse Regression via sul

Convex?

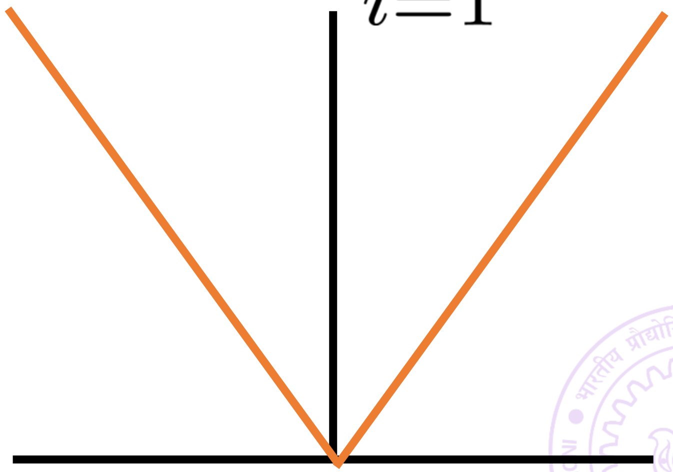
Non-differentiable

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_1$$

$$\mathbf{g}^t \in \nabla f(\mathbf{w}) + \partial r(\mathbf{w})$$

$$\mathbf{g}^t = 2 \sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i + \lambda \cdot \boldsymbol{\rho}^t$$

$$\left\|\mathbf{w}^t\right\|_1 = \sum_{i=1}^{d} \left|\mathbf{w}_i^t\right|$$



App: Sparse Regression via sul

Convex?

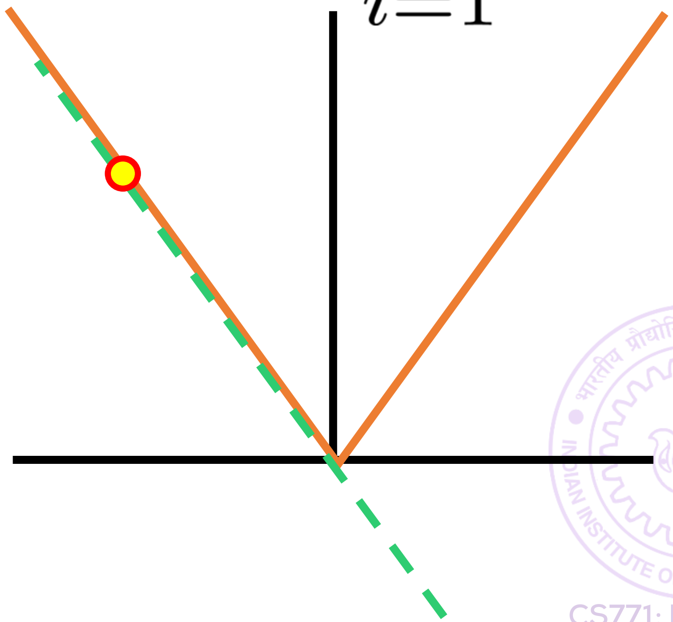
Non-differentiable

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_1$$

$$\mathbf{g}^t \in \nabla f(\mathbf{w}) + \partial r(\mathbf{w})$$

$$\mathbf{g}^t = 2\sum_{i=1}^{n}(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i + \lambda \cdot \boldsymbol{\rho}^t$$

$$\left\|\mathbf{w}^t\right\|_1 = \sum_{i=1}^{d} \left|\mathbf{w}_i^t\right|$$



App: Sparse Regression via sub

Convex?

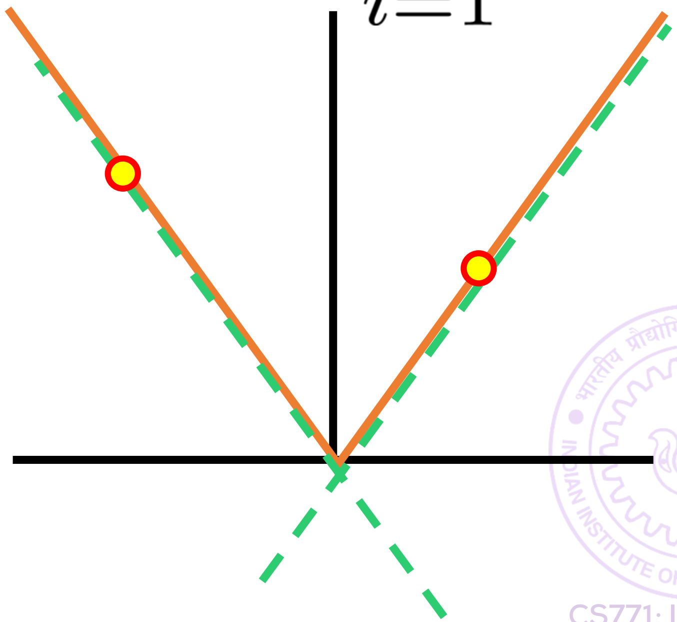
Non-differentiable

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_1$$

$$\mathbf{g}^t \in \nabla f(\mathbf{w}) + \partial r(\mathbf{w})$$

$$\mathbf{g}^t = 2\sum_{i=1}^{n}(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i + \lambda \cdot \boldsymbol{\rho}^t$$

$$\left\|\mathbf{w}^t\right\|_1 = \sum_{i=1}^{d} \left|\mathbf{w}_i^t\right|$$



App: Sparse Regression via sul

Convex?

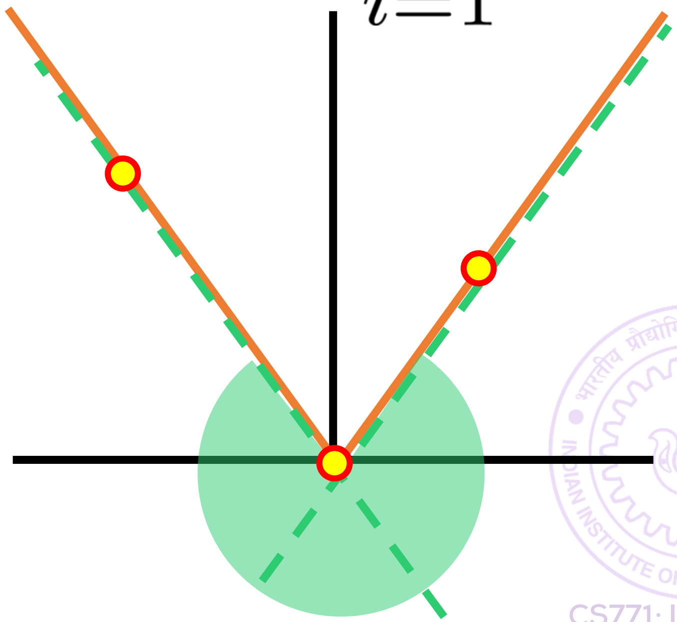
Non-differentiable

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_1$$

$$\mathbf{g}^t \in \nabla f(\mathbf{w}) + \partial r(\mathbf{w})$$

$$\mathbf{g}^t = 2\sum_{i=1}^{n}(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i + \lambda \cdot \boldsymbol{\rho}^t$$

$$\left\|\mathbf{w}^t\right\|_1 = \sum_{i=1}^{d} \left|\mathbf{w}_i^t\right|$$



App: Sparse Regression via sul

Convex?

Non-differentiable

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_1$$

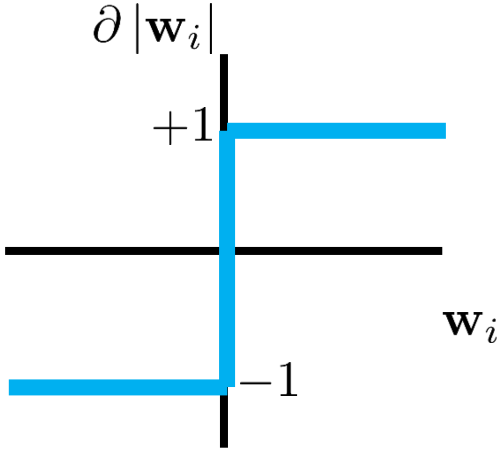
$$\mathbf{g}^t \in \nabla f(\mathbf{w}) + \partial r(\mathbf{w})$$

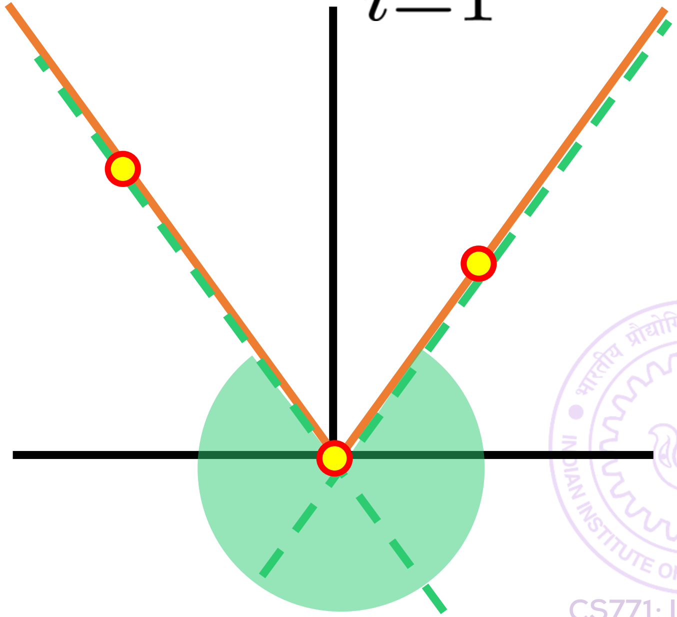
$$\mathbf{g}^t = 2\sum_{i=1}^{n} \left(\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i\right) \cdot \mathbf{x}^i + \lambda \cdot \boldsymbol{\rho}^t$$

$$\left\|\mathbf{w}^t\right\|_1 = \sum_{i=1}^{d} \left|\mathbf{w}_i^t\right|$$

$$\boldsymbol{\rho}_i^t = \begin{cases} +1 & \text{if } \mathbf{w}_i^t > 0 \\ [-1, 1] & \text{if } \mathbf{w}_i^t = 0 \\ -1 & \text{if } \mathbf{w}_i^t < 0 \end{cases}$$

$$\text{e.g. } \boldsymbol{\rho}_i^t = \operatorname{sign}(\mathbf{w}_i^t)$$





App: Sparse Regression via sul

Convex?

Non-differentiable

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2 + \lambda \cdot \|\mathbf{w}\|_1$$

$$\mathbf{g}^t \in \nabla f(\mathbf{w}) + \partial r(\mathbf{w})$$

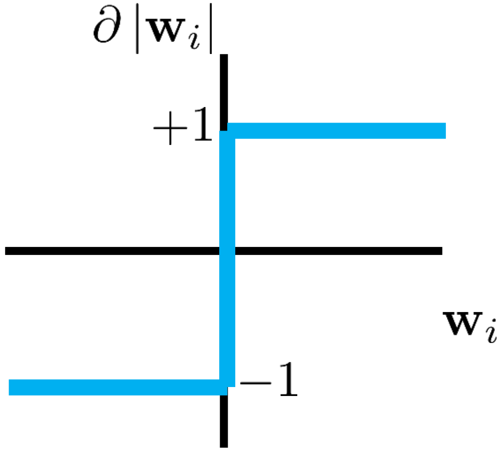
$$\mathbf{g}^t = 2\sum_{i=1}^{n} (\langle \mathbf{w}^t, \mathbf{x}^i \rangle - y^i) \cdot \mathbf{x}^i + \lambda \cdot \boldsymbol{\rho}^t$$

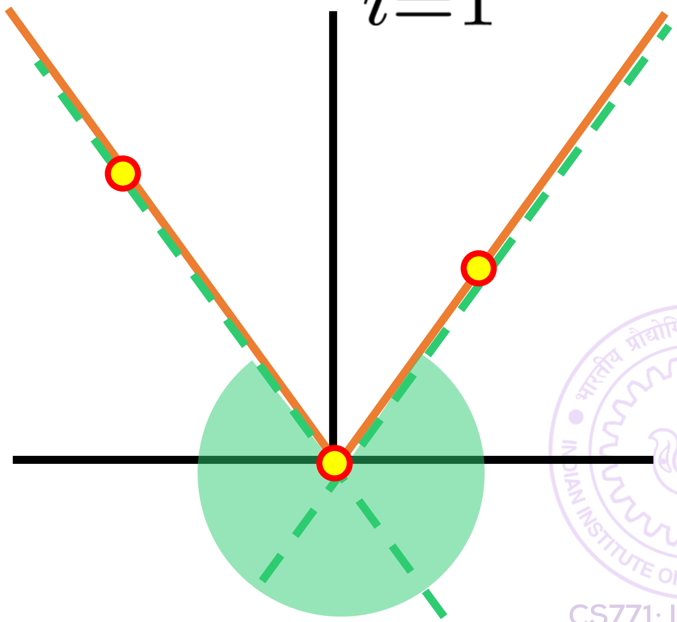
$O(nd)$ time per iter!

$$\|\mathbf{w}^t\|_1 = \sum_{i=1}^{d} |\mathbf{w}_i^t|$$

$$\boldsymbol{\rho}_i^t = \begin{cases} +1 & \text{if } \mathbf{w}_i^t > 0 \\ [-1, 1] & \text{if } \mathbf{w}_i^t = 0 \\ -1 & \text{if } \mathbf{w}_i^t < 0 \end{cases}$$

e.g. $\boldsymbol{\rho}_i^t = \text{sign}(\mathbf{w}_i^t)$





App: Logistic Regression via GD

App: Logistic Regression via GD

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)\right) + \lambda \cdot \|\mathbf{w}\|_2^2$$

App: Logistic Regression via GD

Convex?

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)\right) + \lambda \cdot \|\mathbf{w}\|_2^2$$

App: Logistic Regression via GD

Convex?

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)\right) + \lambda \cdot \|\mathbf{w}\|_2^2$$

$$\mathbf{g}^t = -\sum_{i=1}^{n} (1 - \sigma(y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle)) y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w}$$

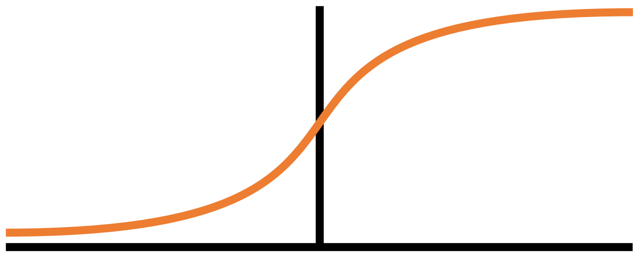
App: Logistic Regression via GD

Convex?

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)\right) + \lambda \cdot \|\mathbf{w}\|_2^2$$

$O(nd)$ time per iter!

$$\mathbf{g}^t = -\sum_{i=1}^{n} (1 - \sigma(y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)) y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w}$$



Whats going on here?

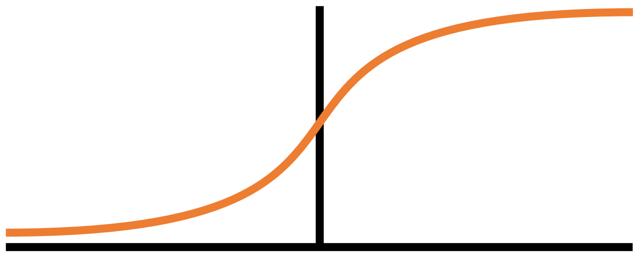
App: Logistic Regression via GD

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \log\left(1 + \exp(-y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)\right) + \lambda \cdot \|\mathbf{w}\|_2^2$$

Convex?

$$\mathbf{g}^t = -\sum_{i=1}^{n} (1 - \sigma(y^i \langle \mathbf{w}, \mathbf{x}^i \rangle)) y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w}$$

$O(nd)$ time per iter!



Whats going on here?

Exercise

Stochastic Gradient Method

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \; f(\mathbf{w}) + r(\mathbf{w})$$

Stochastic Gradient Method

$$f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Stochastic Gradient Method

$$f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

Stochastic Gradient Method

$\boxed{O(nd)\text{ time}}$

$$\nabla f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

Stochastic Gradient Method

$O(nd)$ time

$$\nabla f(\mathbf{w}) \approx \qquad \nabla f_i(\mathbf{w}) \approx \qquad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \qquad \nabla f_i(\mathbf{w}) \approx \qquad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \qquad \nabla f_i(\mathbf{w}) \approx \qquad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f_i(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \qquad \nabla f_i(\mathbf{w}) \approx \qquad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f_i(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

STOCHASTIC GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Select a unif. random point $I_t \in [n]$
3. Let $\mathbf{g}^t \leftarrow \nabla f_{I_t}(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$
4. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
5. Repeat until convergence

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \qquad \nabla f_i(\mathbf{w}) \approx \qquad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f_i(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

$\mathbb{E}[\mathbf{g}^t | \mathbf{w}^t] = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$

STOCHASTIC GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Select a unif. random point $I_t \in [n]$
3. Let $\mathbf{g}^t \leftarrow \nabla f_{I_t}(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$
4. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
5. Repeat until convergence

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \quad \nabla f_i(\mathbf{w}) \approx \quad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f_i(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

$\mathbb{E}[\mathbf{g}^t | \mathbf{w}^t] = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$

Only $O(d)$ time per iter!

STOCHASTIC GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Select a unif. random point $I_t \in [n]$
3. Let $\mathbf{g}^t \leftarrow \nabla f_{I_t}(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$
4. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
5. Repeat until convergence

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \qquad \nabla f_i(\mathbf{w}) \approx \qquad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f_i(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

$\mathbb{E}[\mathbf{g}^t | \mathbf{w}^t] = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$

Only $O(d)$ time per iter!

Step length $\eta_t = C, \frac{C}{t}, \frac{C}{\sqrt{t}}$

STOCHASTIC GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Select a unif. random point $I_t \in [n]$
3. Let $\mathbf{g}^t \leftarrow \nabla f_{I_t}(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$
4. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
5. Repeat until convergence

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \quad \nabla f_i(\mathbf{w}) \approx \quad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f_i(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

$\mathbb{E}[\mathbf{g}^t | \mathbf{w}^t] = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$

Only $O(d)$ time per iter!

Step length $\eta_t = C, \frac{C}{t}, \frac{C}{\sqrt{t}}$

w.h.p. ϵ-optimal solution in $O\left(\frac{1}{\epsilon^2}\right)$ iterations

STOCHASTIC GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Select a unif. random point $I_t \in [n]$
3. Let $\mathbf{g}^t \leftarrow \nabla f_{I_t}(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$
4. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
5. Repeat until convergence

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \quad \nabla f_i(\mathbf{w}) \approx \quad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f_i(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

$\mathbb{E}[\mathbf{g}^t | \mathbf{w}^t] = \nabla f(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$

Only $O(d)$ time per iter!

STOCHASTIC GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Select a unif. random point $I_t \in [n]$
3. Let $\mathbf{g}^t \leftarrow \nabla f_{I_t}(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$
4. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
5. Repeat until convergence

Step length $\eta_t = C, \frac{C}{t}, \frac{C}{\sqrt{t}}$

w.h.p. ϵ-optimal solution in $O\left(\frac{1}{\epsilon^2}\right)$ iterations

$f(\mathbf{w}^t) + r(\mathbf{w}^t)$
$\leq f(\mathbf{w}^*) + r(\mathbf{w}^*) + \epsilon$

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \quad \nabla f_i(\mathbf{w}) \approx \quad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f_i(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

ONLINE GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Select a unif. random point $I_t \in [n]$
3. Let $\mathbf{g}^t \leftarrow \nabla f_{I_t}(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$
4. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
5. Repeat until convergence

Only $O(d)$ time per iter!

Step length $\eta_t = C, \frac{C}{t}, \frac{C}{\sqrt{t}}$

w.h.p. ϵ-optimal solution in $O\left(\frac{1}{\epsilon^2}\right)$ iterations

$f(\mathbf{w}^t) + r(\mathbf{w}^t) \leq f(\mathbf{w}^*) + r(\mathbf{w}^*) + \epsilon$

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \quad \nabla f_i(\mathbf{w}) \approx \quad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f_i(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

Only $O(d)$ time per iter!

ONLINE GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Receive a data point $\mathbf{z}^t = (y^t, \mathbf{x}^t)$
3. Let $\mathbf{g}^t \leftarrow \nabla f_{I_t}(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$
4. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
5. Repeat until convergence

Step length $\eta_t = C, \frac{C}{t}, \frac{C}{\sqrt{t}}$

w.h.p. ϵ-optimal solution in $O\left(\frac{1}{\epsilon^2}\right)$ iterations

$f(\mathbf{w}^t) + r(\mathbf{w}^t) \leq f(\mathbf{w}^*) + r(\mathbf{w}^*) + \epsilon$

Stochastic Gradient Method

$O(d)$ time!!

$$\nabla f(\mathbf{w}) \approx \qquad \nabla f_i(\mathbf{w}) \approx \qquad \nabla \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$\mathbf{g}^t = \nabla f_i(\mathbf{w}^t) + \nabla r(\mathbf{w}^t)$$

Only $O(d)$ time per iter!

ONLINE GRADIENT DESCENT

1. Initialize $\mathbf{w}^0$
2. Receive a data point $\mathbf{z}^t = (y^t, \mathbf{x}^t)$
3. Let $\mathbf{g}^t \leftarrow \nabla f(\mathbf{w}^t, \mathbf{z}^t) + \nabla r(\mathbf{w}^t)$
4. Update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \cdot \mathbf{g}^t$
5. Repeat until convergence

Step length $\eta_t = C, \frac{C}{t}, \frac{C}{\sqrt{t}}$

w.h.p. ϵ-optimal solution in $O\left(\frac{1}{\epsilon^2}\right)$ iterations

$f(\mathbf{w}^t) + r(\mathbf{w}^t) \leq f(\mathbf{w}^*) + r(\mathbf{w}^*) + \epsilon$

Please give your Feedback

http://tinyurl.com/ml17-18afb