Function Approximation Methods-II

CS771: Introduction to Machine Learning

Purushottam Kar

Recap

Recap

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

$$\text{s.t.} \mathbf{w}_i \geq 0$$

Recap

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

$$\text{s.t.} \mathbf{w}_i \geq 0$$

Loss Function

Recap

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

$$\text{s.t.} \mathbf{w}_i \geq 0$$

Regularizer

Loss Function

Recap

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

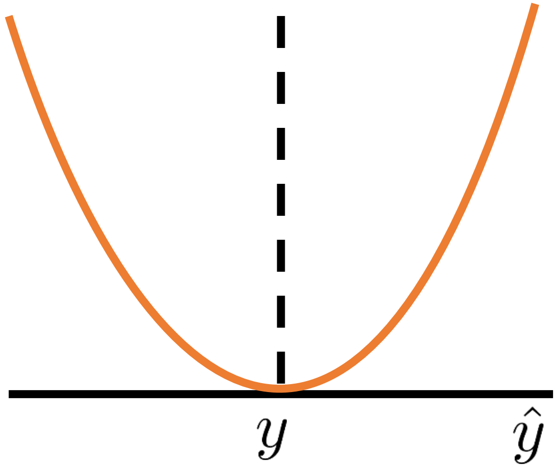
$$\text{s.t.} \mathbf{w}_i \geq 0$$

Regularizer

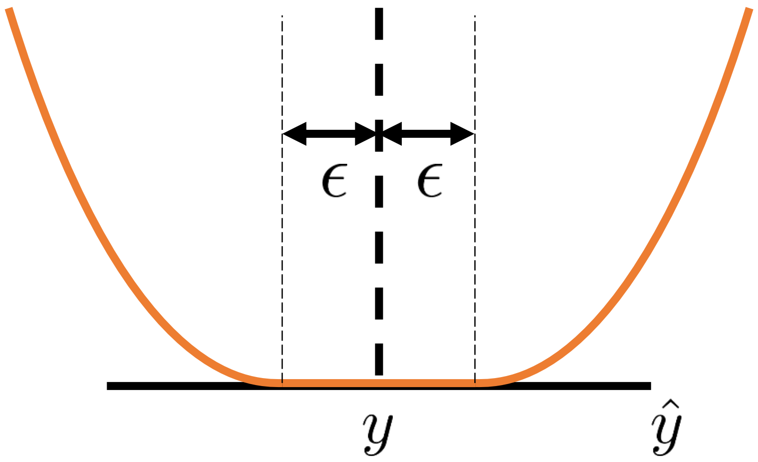
Loss Function

Constraint

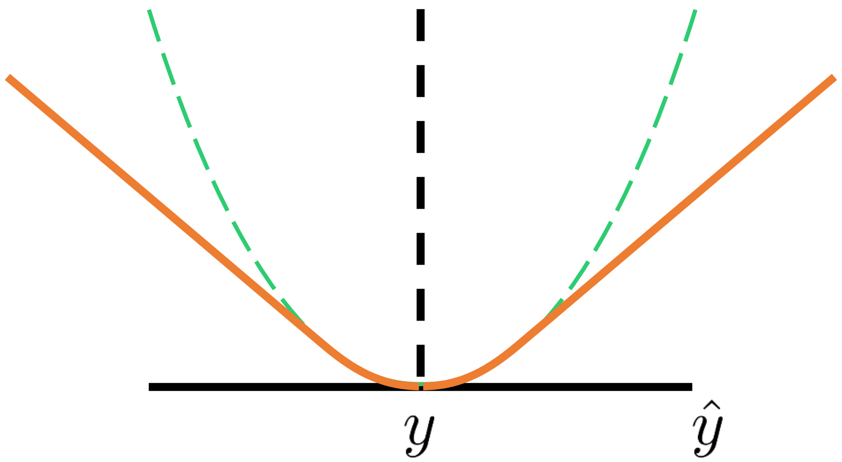
Recap



Square loss



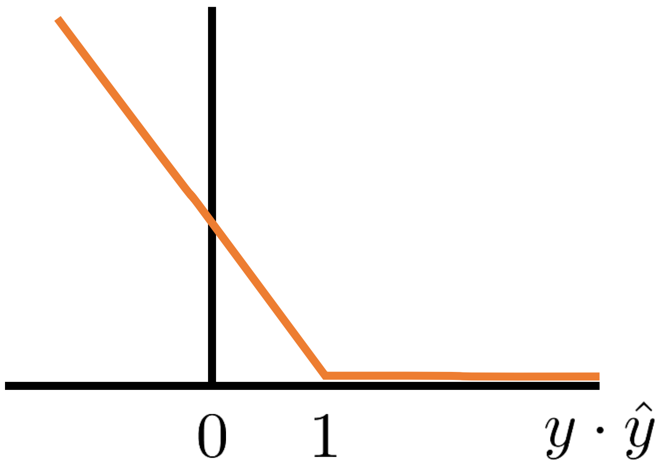
Vapnik's ϵ-insensitive loss



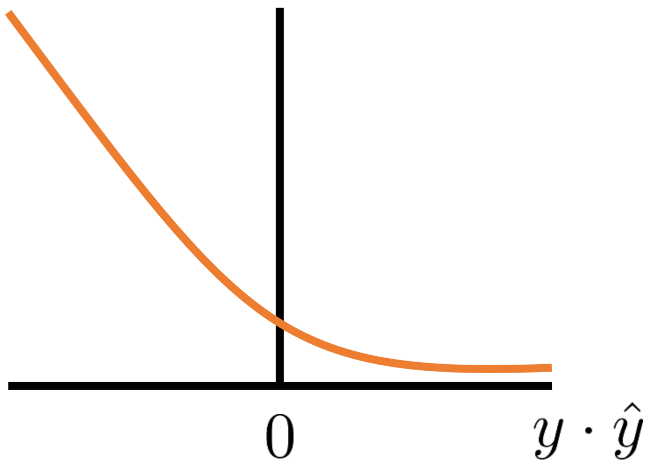
Huber loss

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

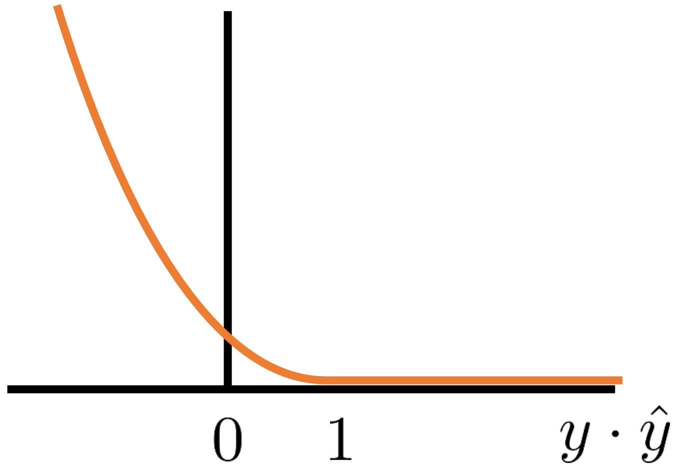
Recap



Hinge loss



Logistic loss



Squared Hinge loss

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \ell(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Recap: Looking into the Hinge Loss

$\boxed{\text{Binary Classification}}$

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Recap: Looking into the Hinge Los

Can use non-linear functions too!

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Recap: Looking into the Hinge Loss

$\boxed{\text{Binary Classification}}$ $\quad \hat{y}^i = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$

Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

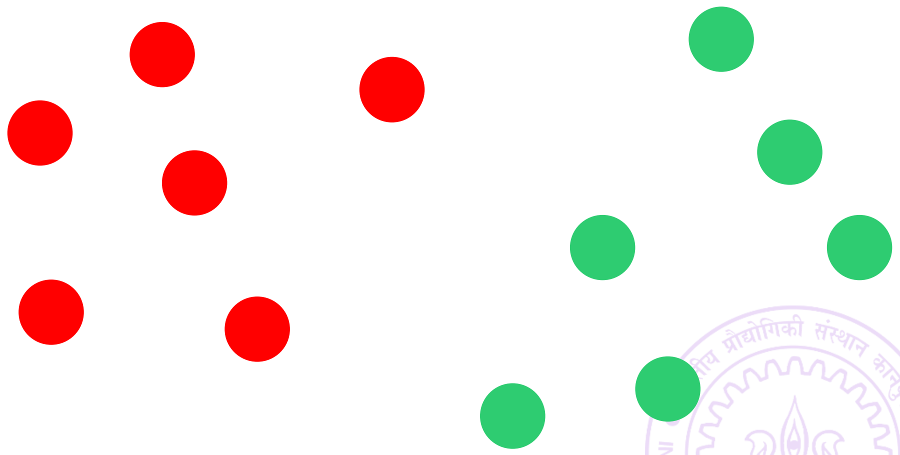
Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!



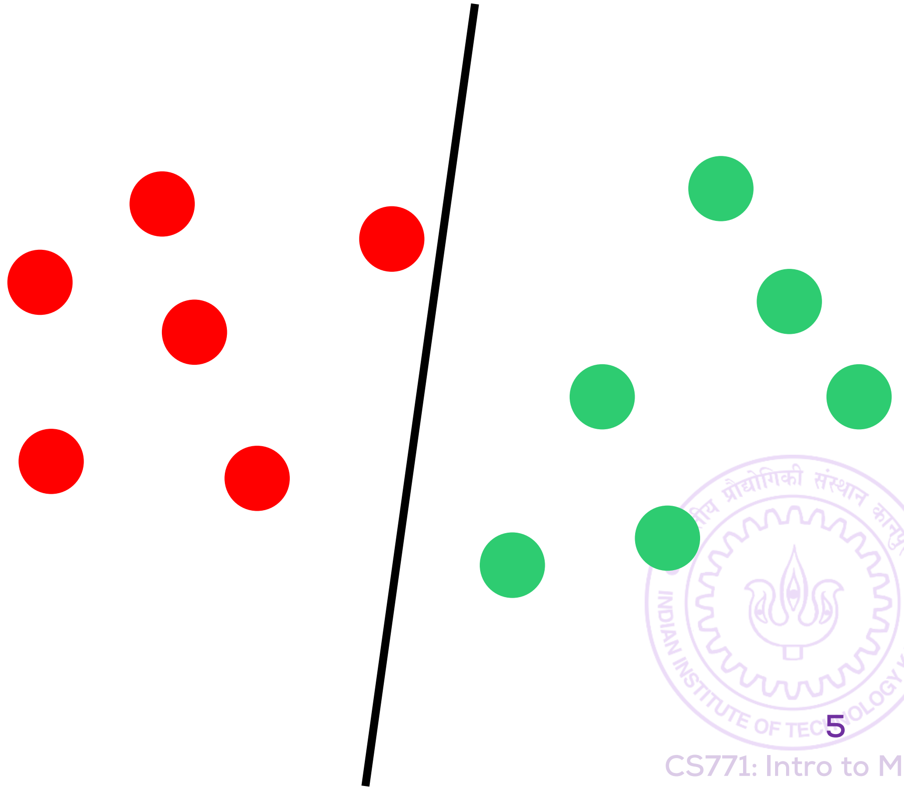
Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!



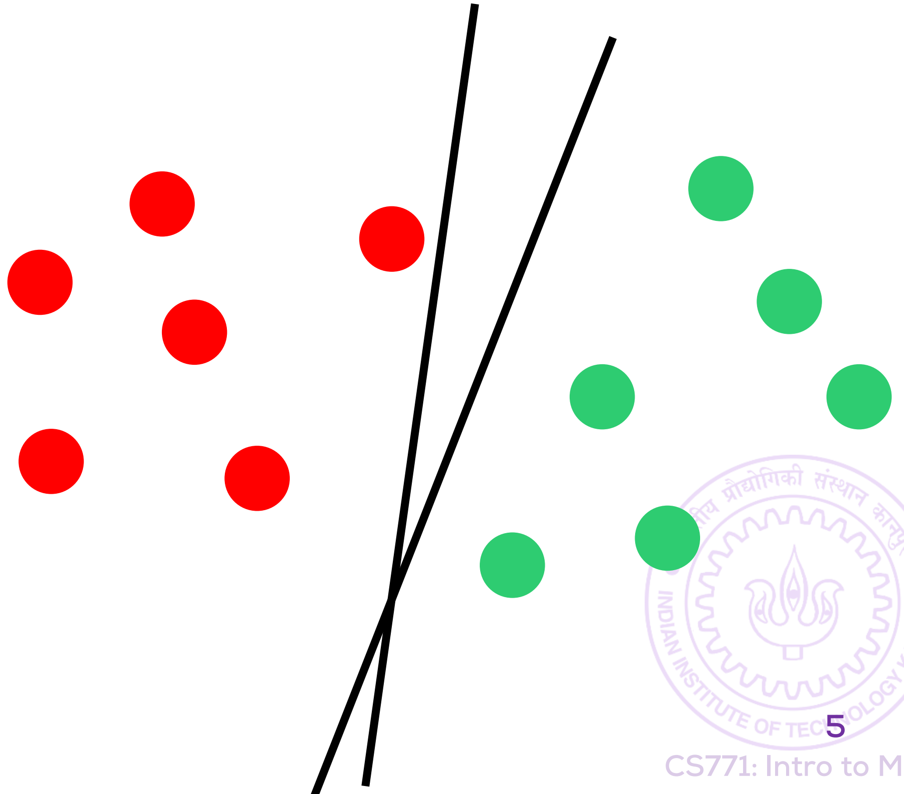
Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!



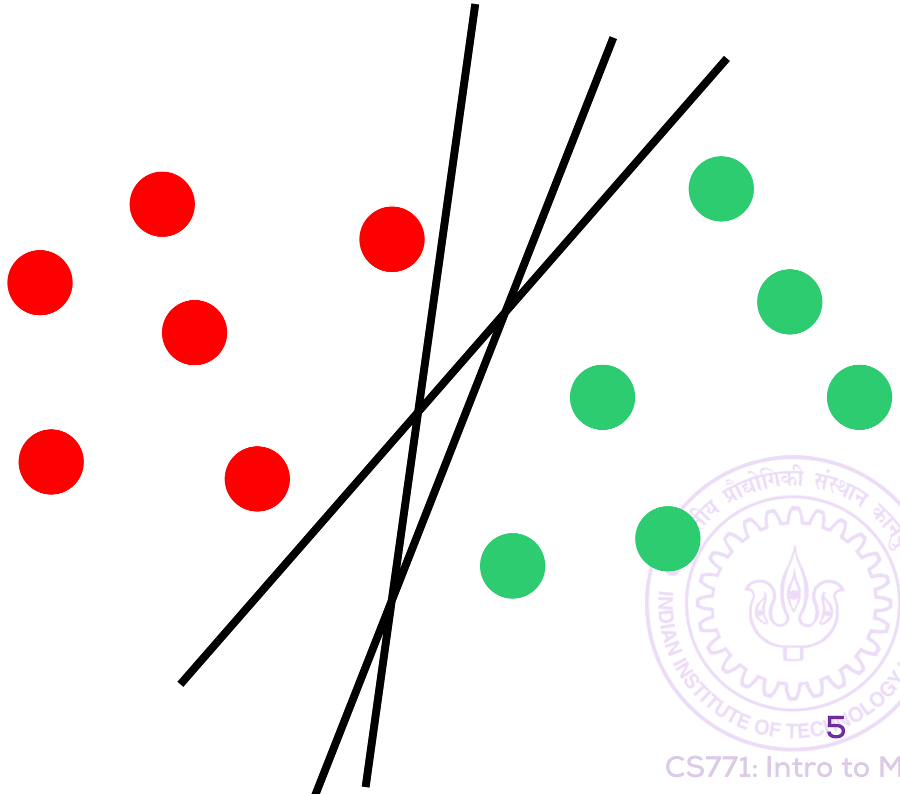
Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!



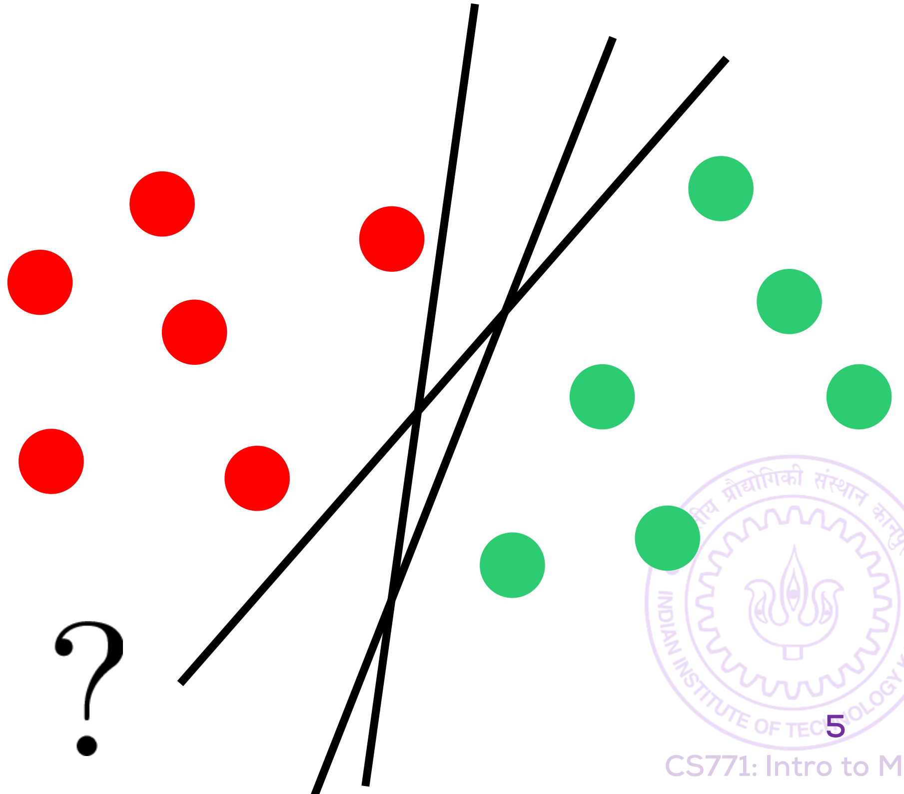
Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!



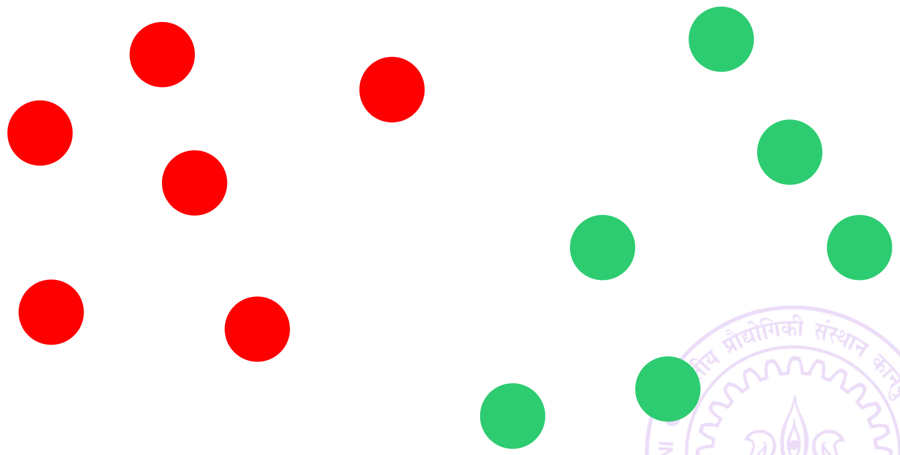
Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!



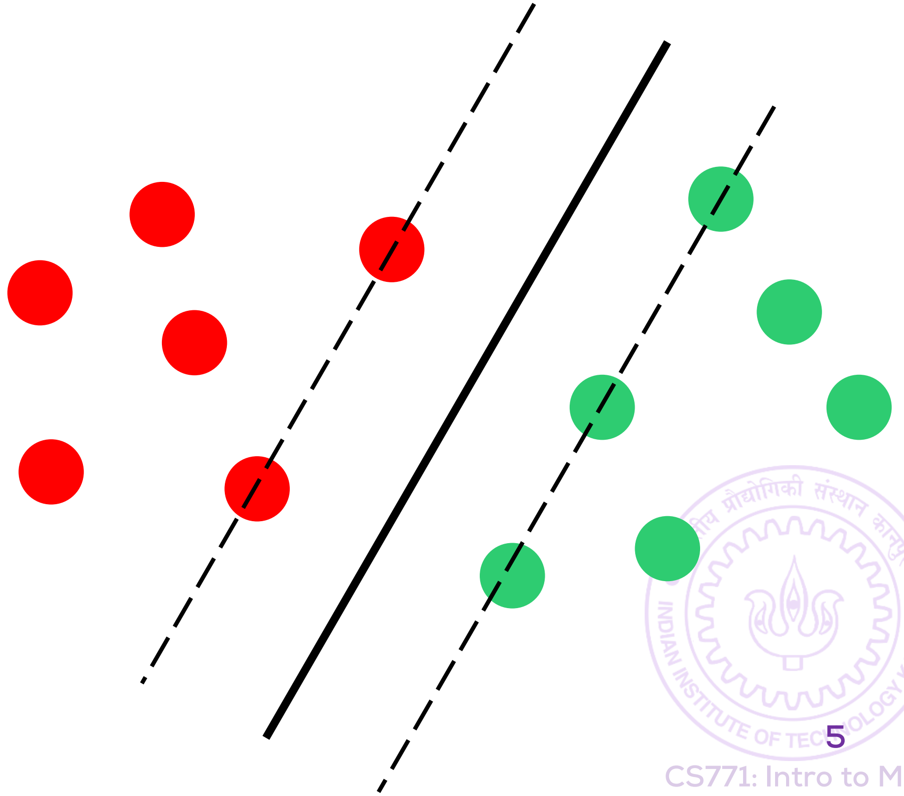
Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!



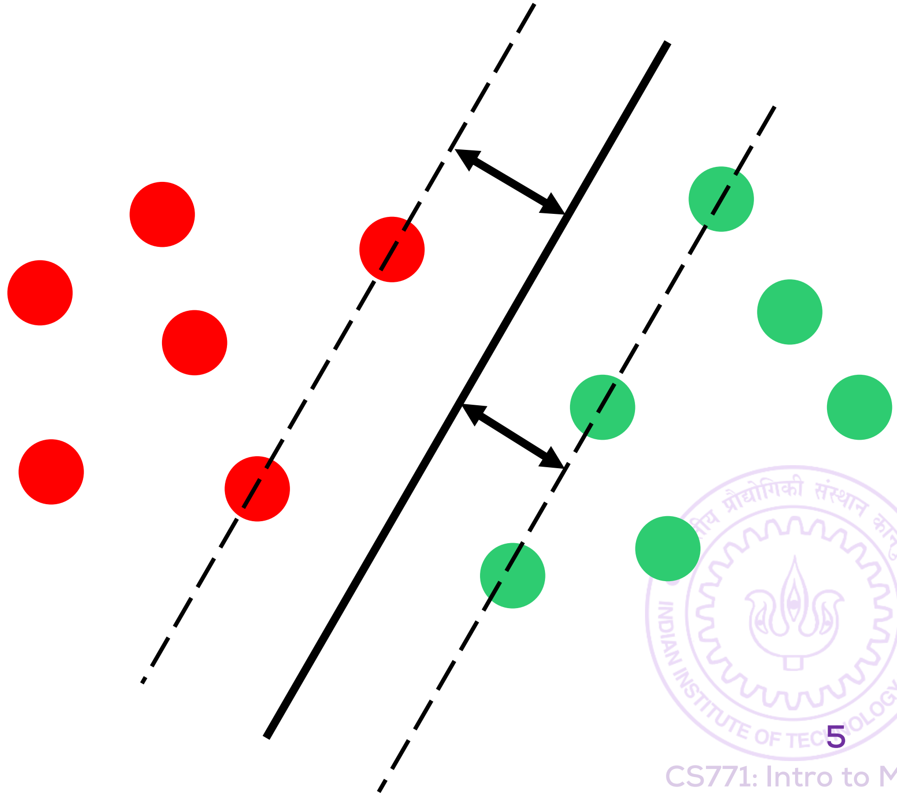
Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!



Recap: Looking into the Hinge Loss

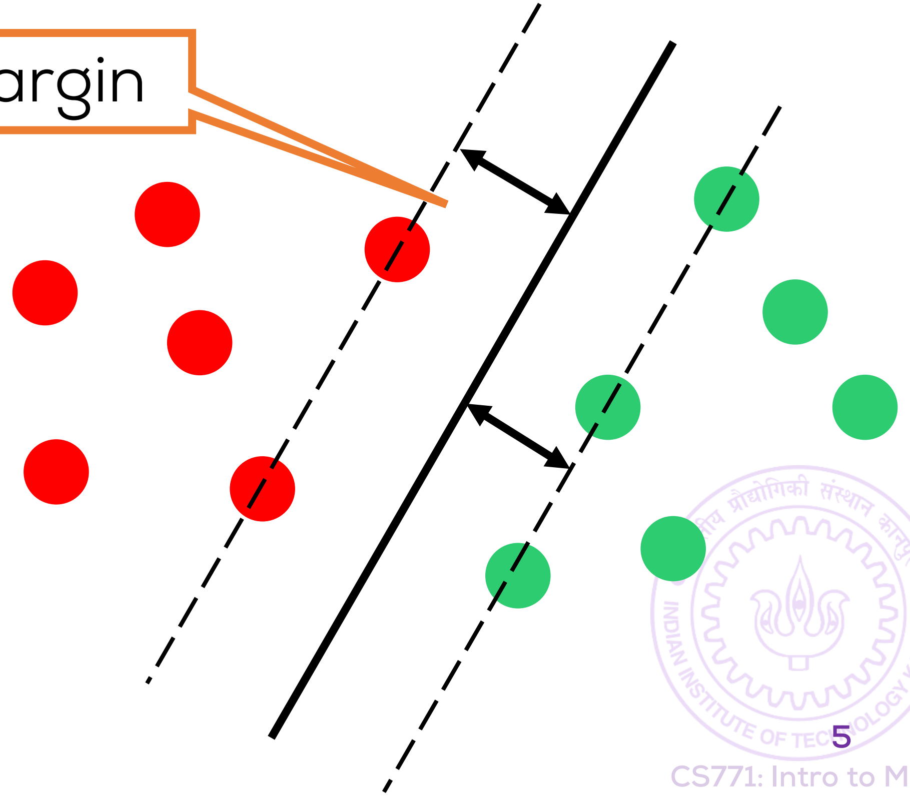
Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

Margin



Recap: Looking into the Hinge Loss

Binary Classification

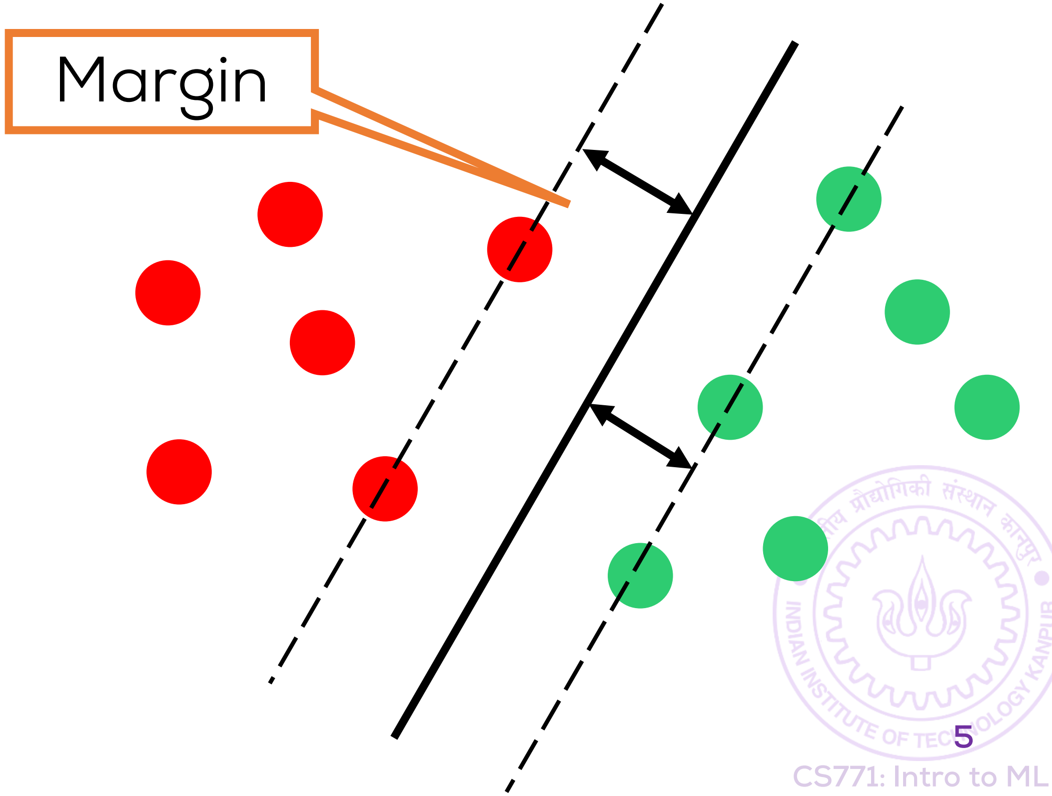
$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

$$\text{s.t. } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1$$



Recap: Looking into the Hinge Loss

Binary Classification

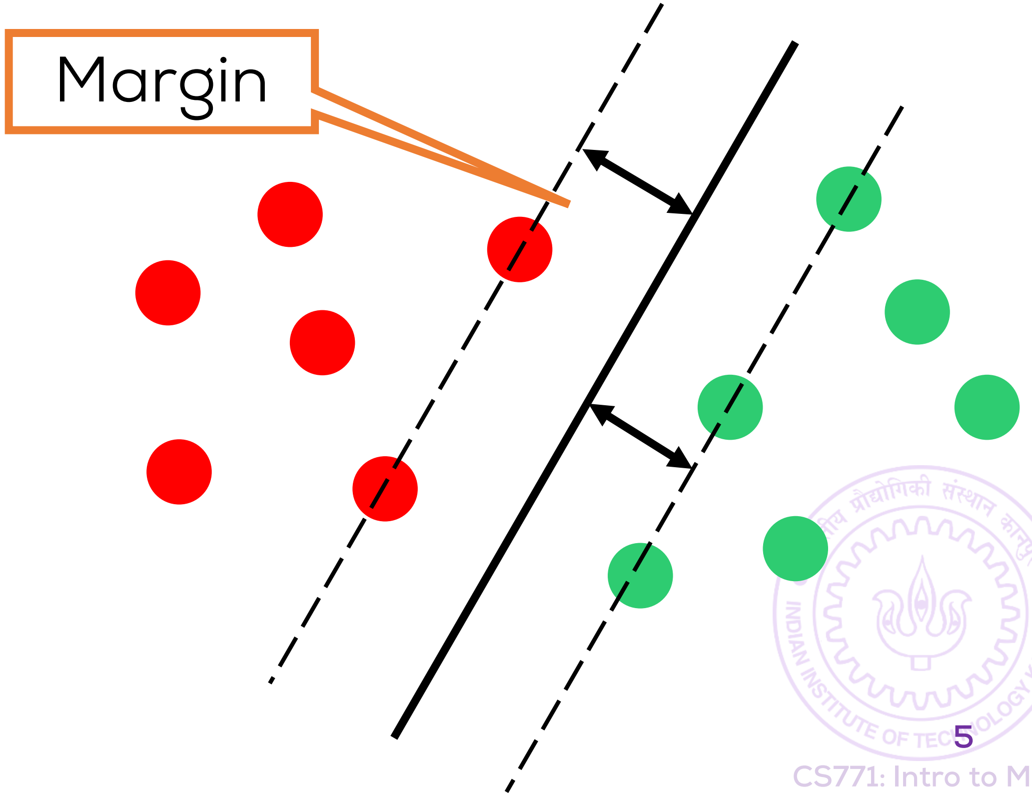
$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$
$$\text{s.t. } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1$$

Margin



Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

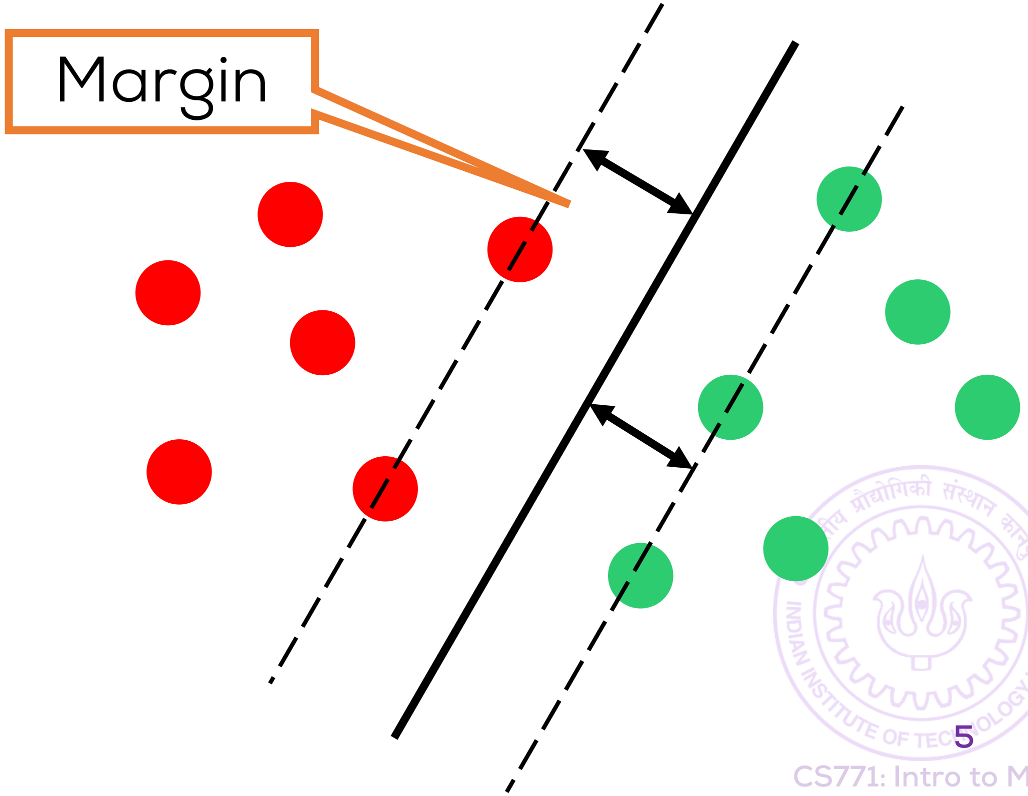
Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

$$\text{s.t. } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1$$

Margin

Why 1?



Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

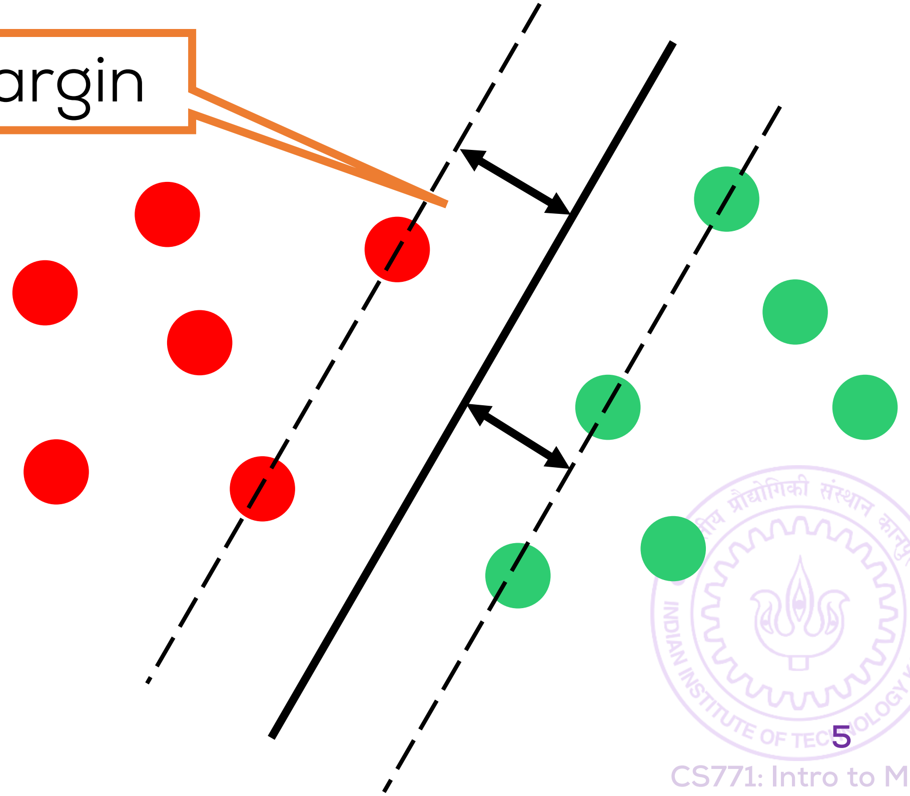
$$\text{s.t. } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1$$

Regularization parameter

Margin

Why 1?

Margin



Recap: Looking into the Hinge Loss

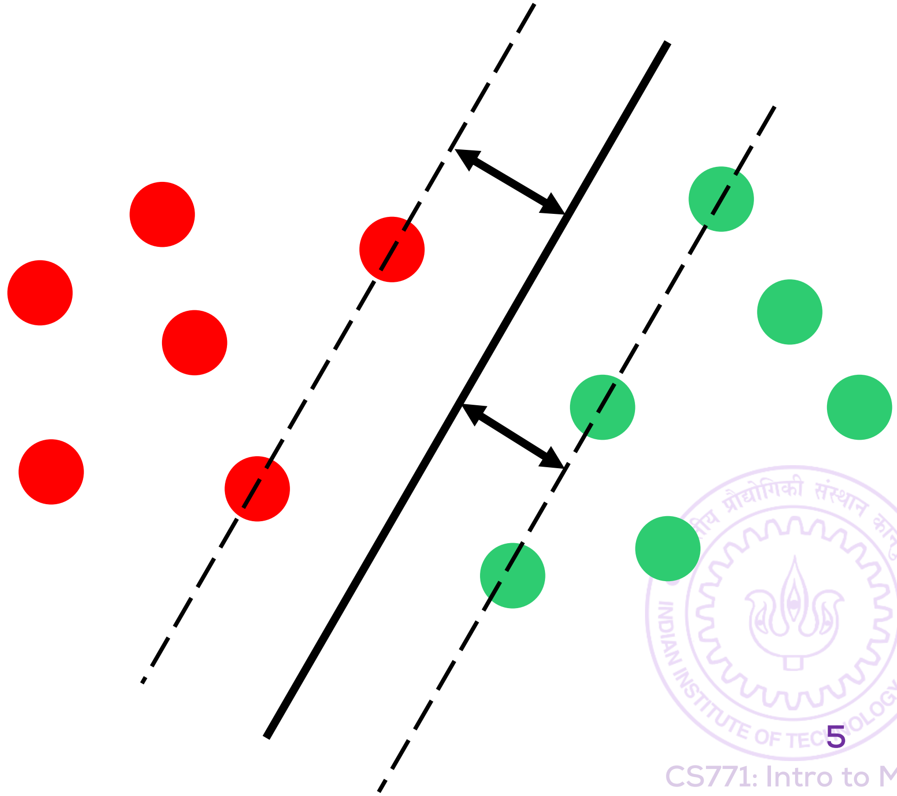
Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$
$$\text{s.t. } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1$$



Recap: Looking into the Hinge Loss

Binary Classification

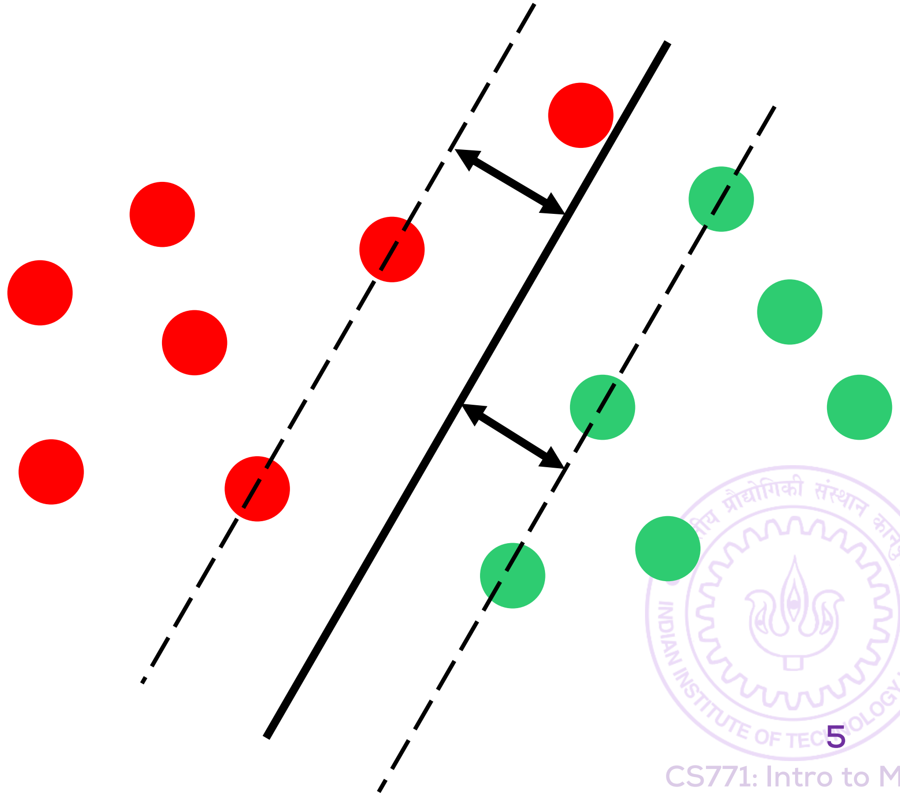
$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

$$\text{s.t. } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1$$



Recap: Looking into the Hinge Loss

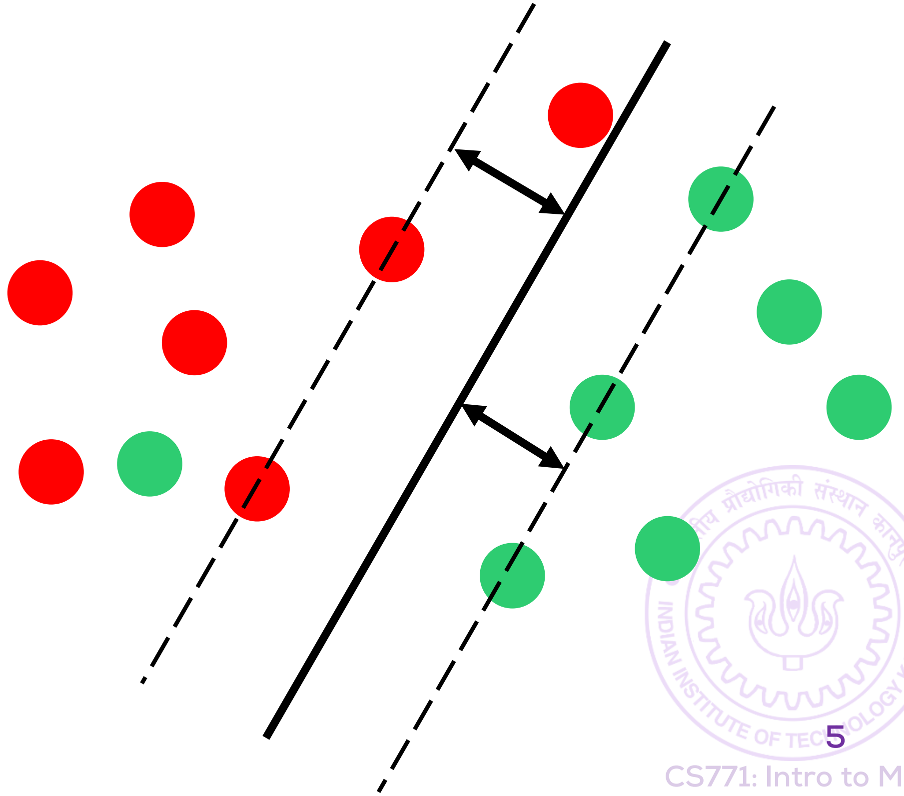
Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$
$$\text{s.t. } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1$$



Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

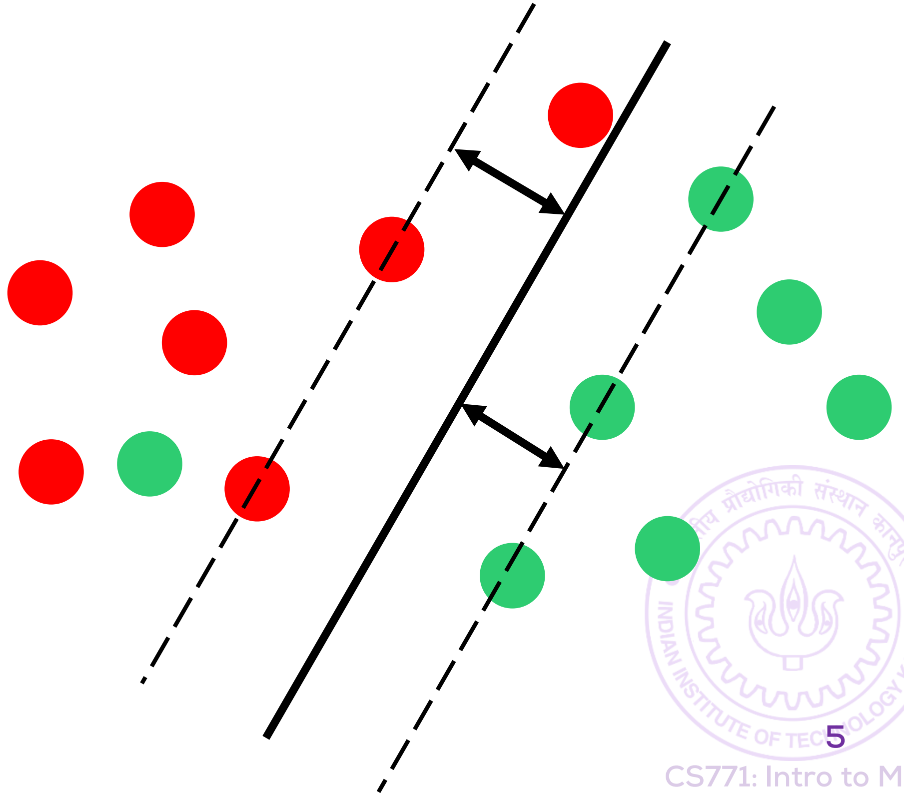
Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

Slack variable

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

$$\text{s.t. } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1 - \xi_i$$



Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Slack variable

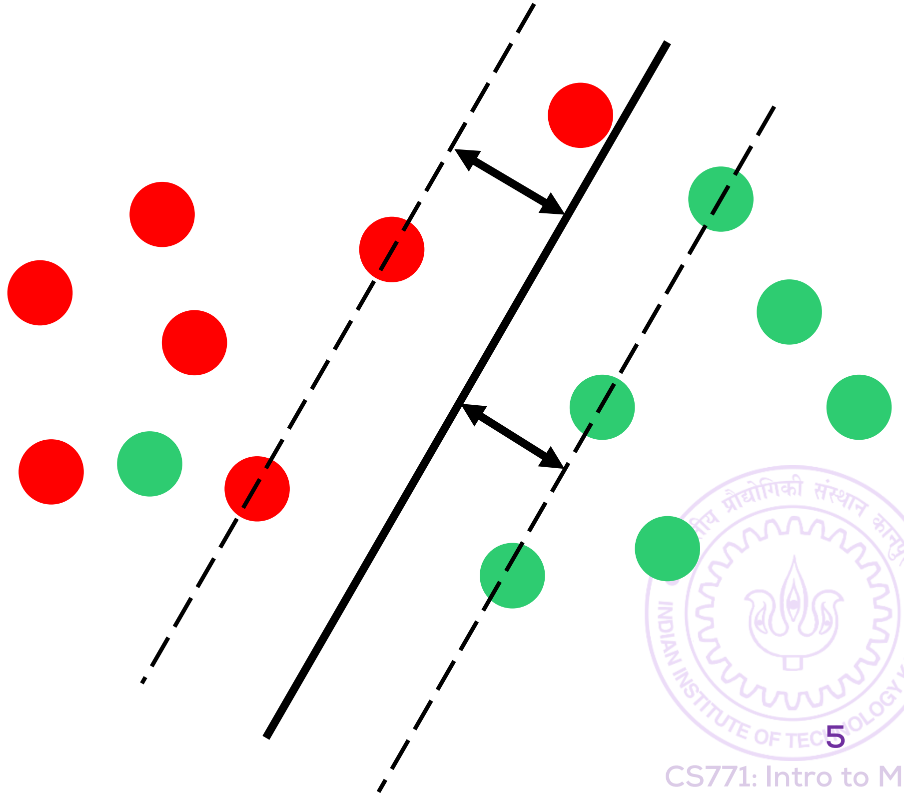
Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

$$\widehat{\mathbf{w}} = \underset{\mathbf{w},\, \{\xi_i\}}{\arg\min}\ \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

More params to optimize



Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

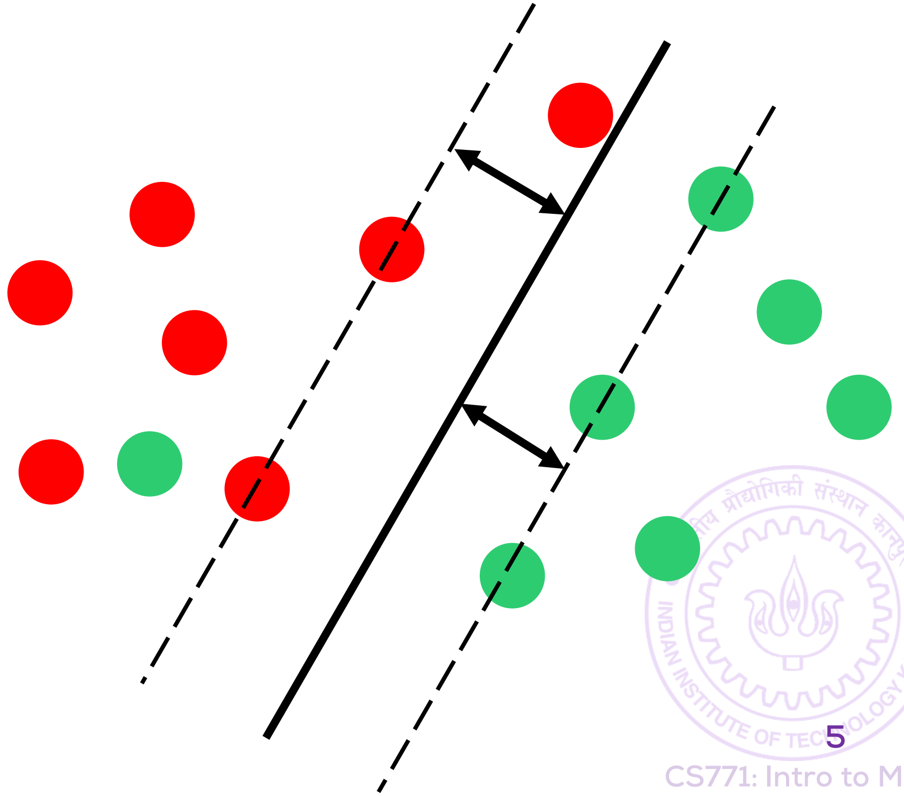
Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

Slack variable

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}, \{\xi_i\}}{\arg\min} \frac{\lambda}{2} \|\mathbf{w}\|_2^2 + \sum_{i=1}^n \xi_i$$

$$\text{s.t. } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \geq 1 - \xi_i$$

$$\xi_i \geq 0$$



Recap: Looking into the Hinge Loss

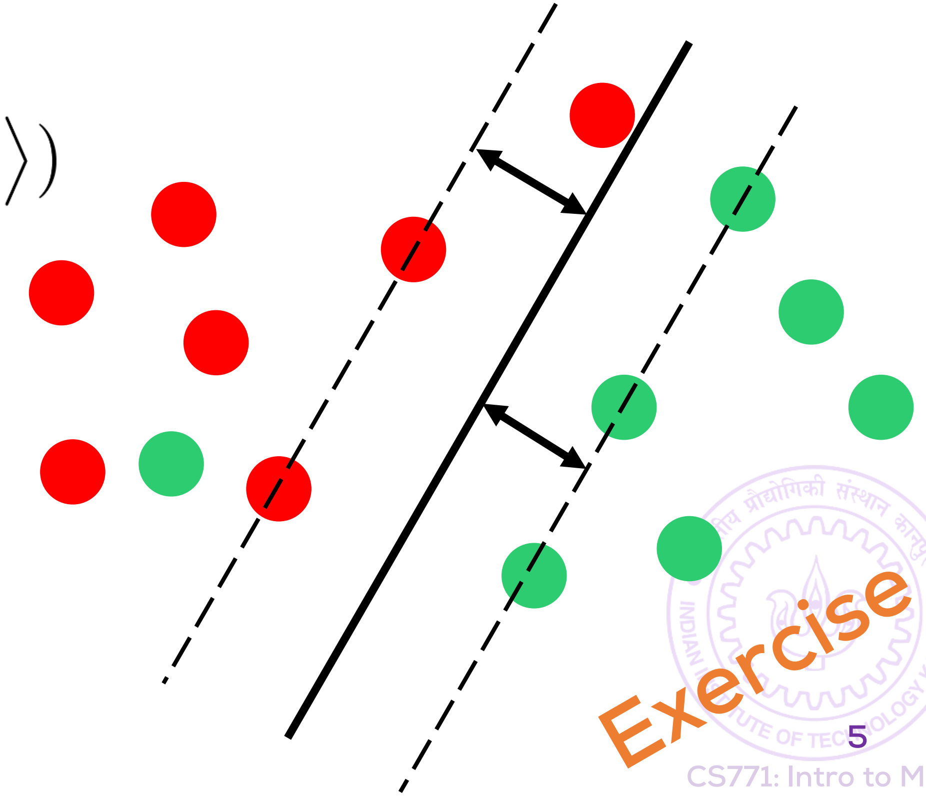
Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

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



Exercise

Recap: Looking into the Hinge Loss

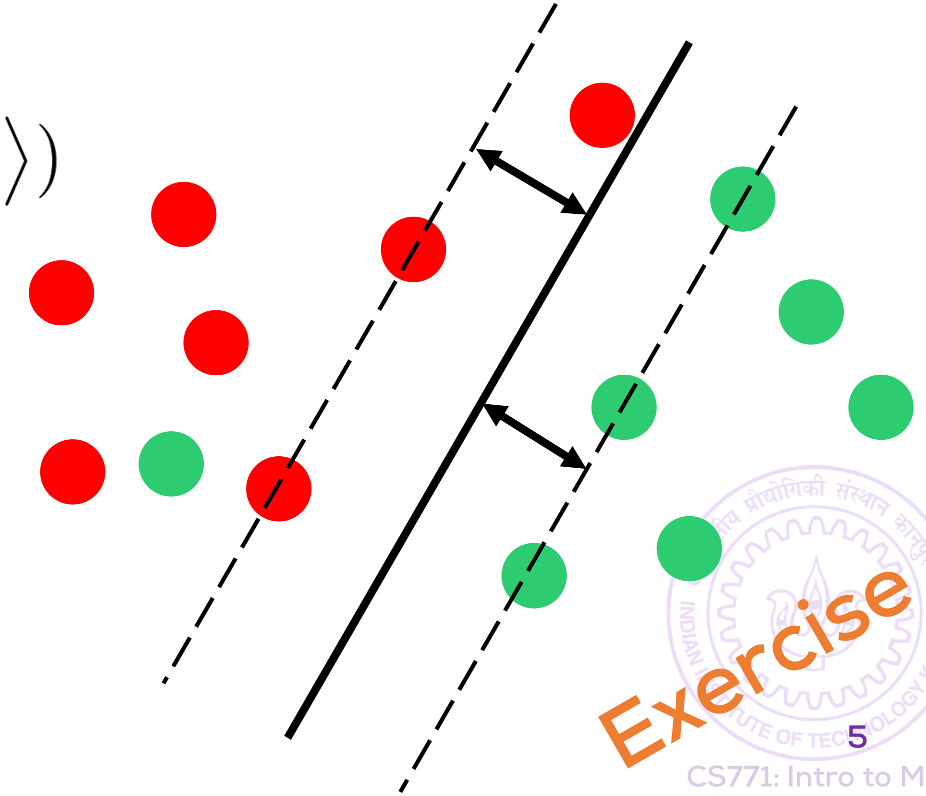
Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

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



Exercise

Recap: Looking into the Hinge Loss

Binary Classification

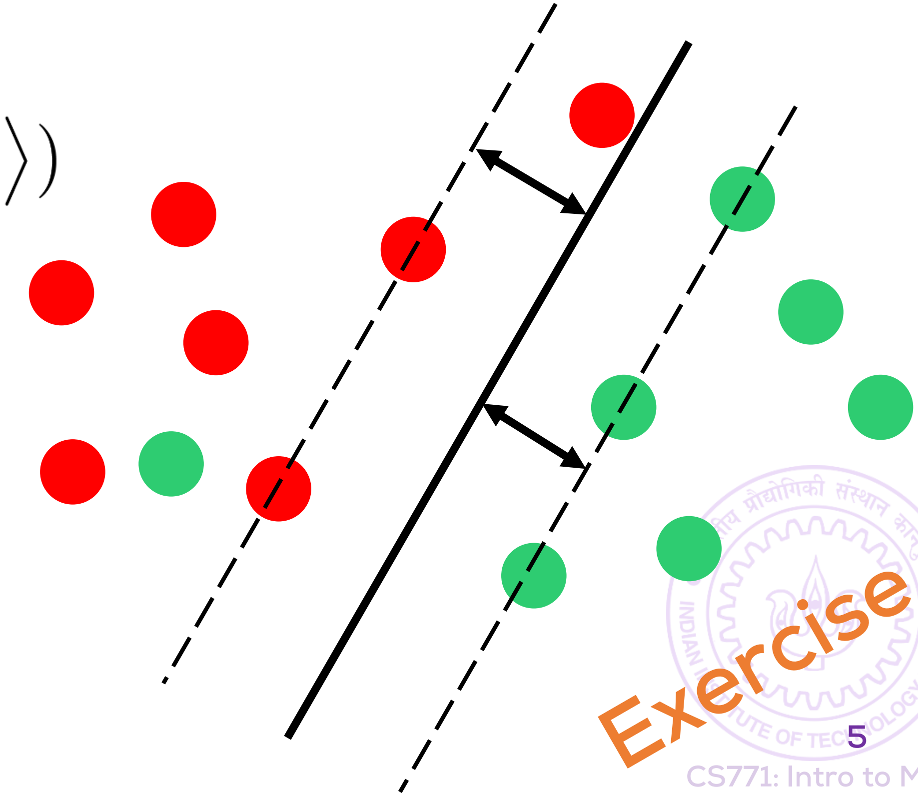
$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

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

Large Margin Classifier



Exercise

Recap: Looking into the Hinge Loss

Binary Classification

$$\hat{y}^i = \text{sign}(\langle \mathbf{w}, \mathbf{x}^i \rangle)$$

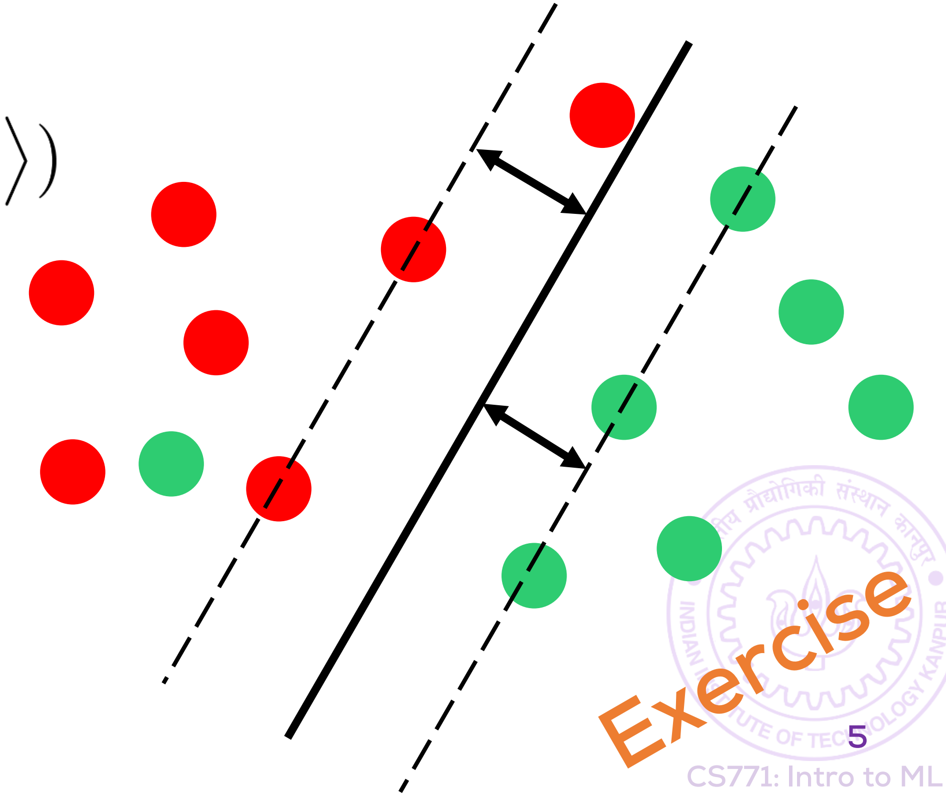
Want sign of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be correct

Want magnitude of $\langle \mathbf{w}, \mathbf{x}^i \rangle$ to be large too!

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

Large Margin Classifier

SVM



Exercise

Loss Functions for Structured ML problems

Multi-class and Multi-label classification

Multi-classification Loss Functions

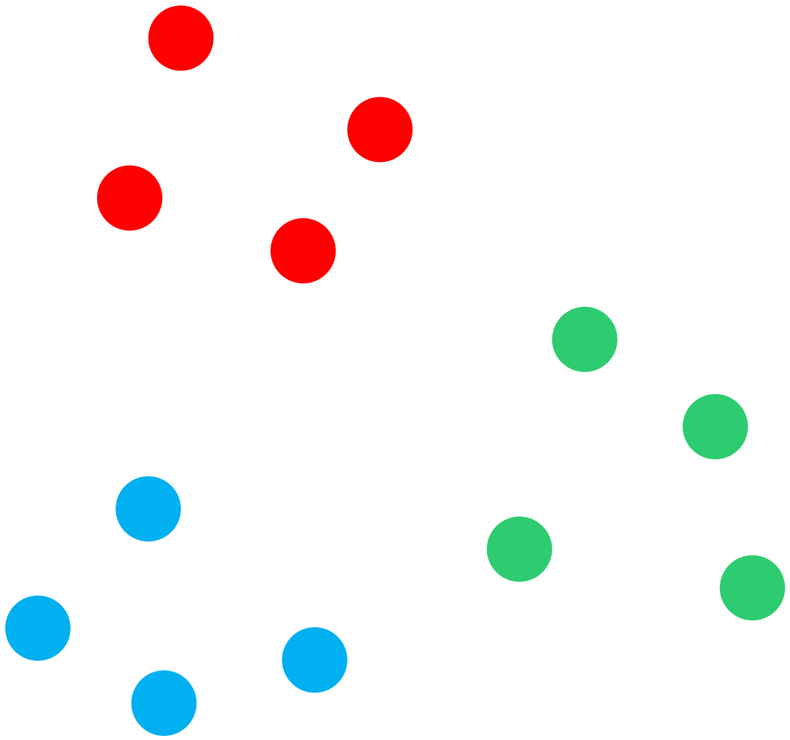
$\boxed{\text{One-vs-All (OVA)}}$

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

Multi-classification Loss Functions

One-vs-All (OVA)

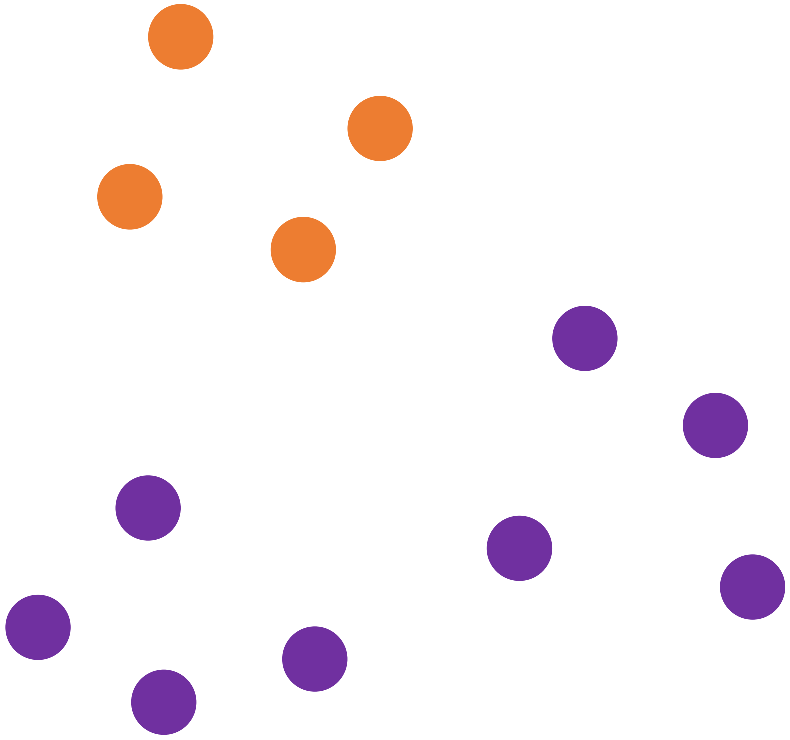
$$\hat{y}^i = \underset{j \in [K]}{\arg\max} \left\langle \mathbf{w}^j, \mathbf{x}^i \right\rangle$$



Multi-classification Loss Functions

$\boxed{\text{One-vs-All (OVA)}}$

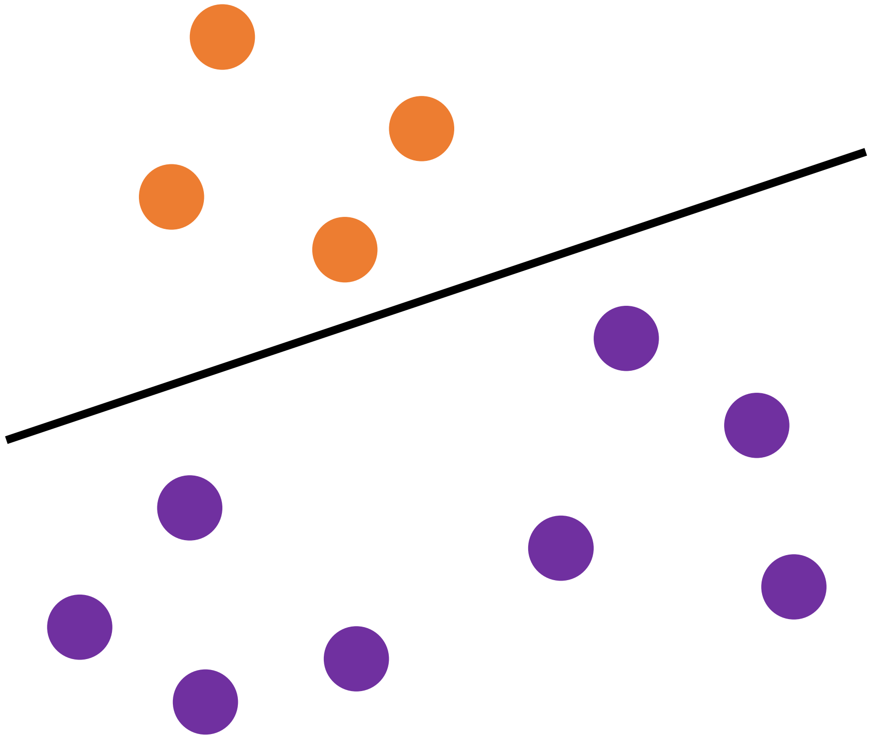
$$\hat{y}^i = \underset{j \in [K]}{\arg\max} \; \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

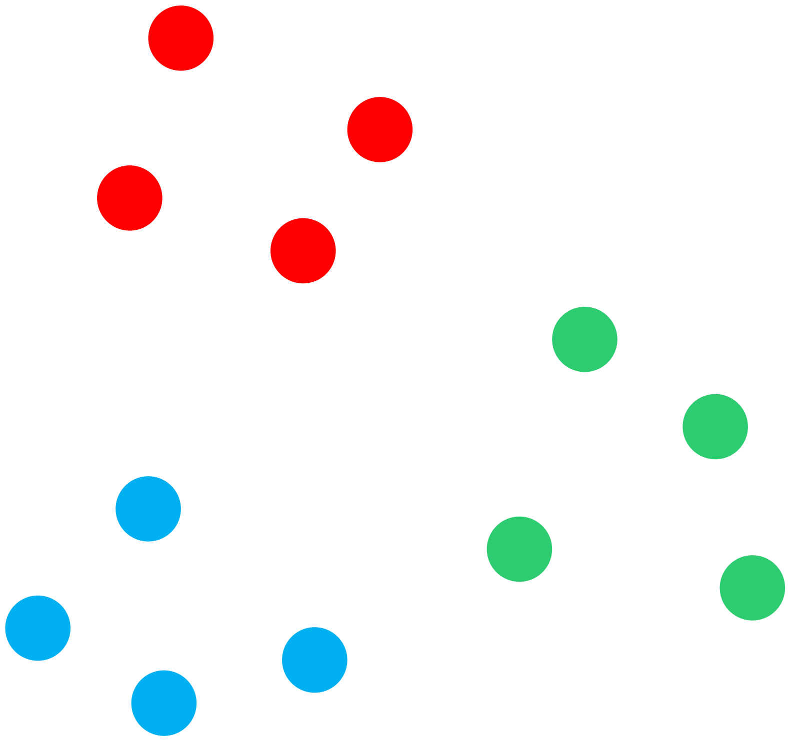
$$\hat{y}^i = \arg\max_{j \in [K]} \left\langle \mathbf{w}^j, \mathbf{x}^i \right\rangle$$



Multi-classification Loss Functions

$\boxed{\text{One-vs-All (OVA)}}$

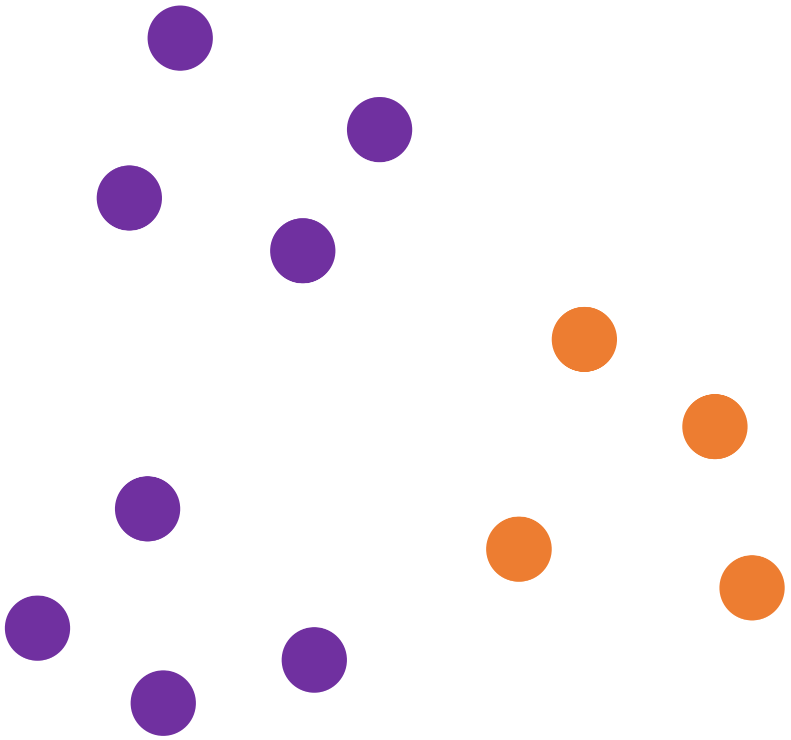
$$\hat{y}^i = \arg\max_{j \in [K]} \left\langle \mathbf{w}^j, \mathbf{x}^i \right\rangle$$



Multi-classification Loss Functions

$\boxed{\text{One-vs-All (OVA)}}$

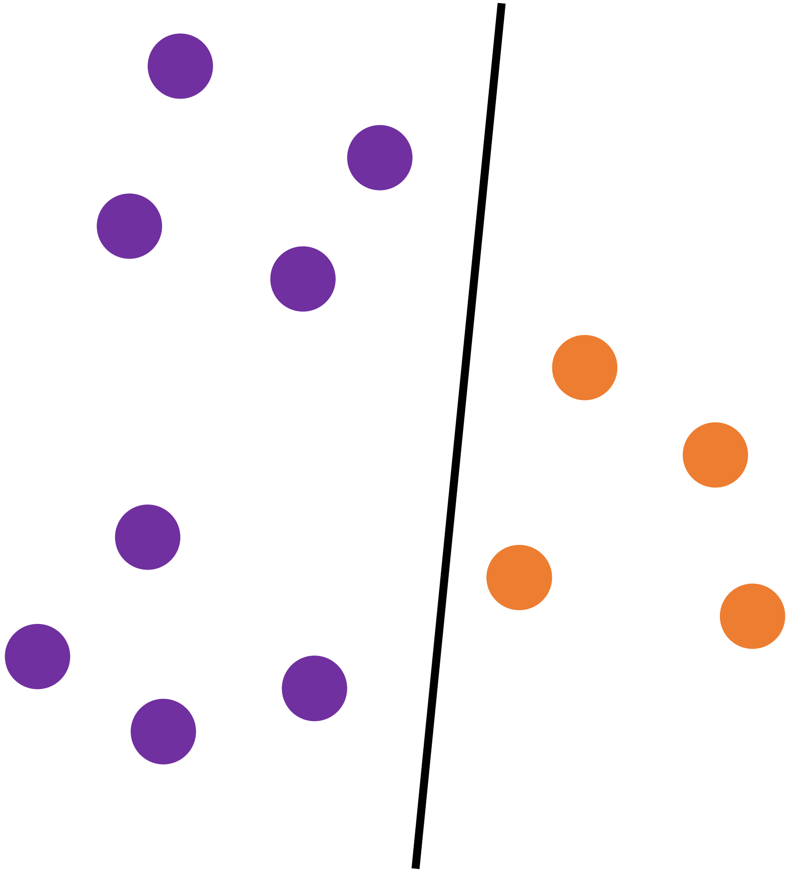
$$\hat{y}^i = \arg\max_{j \in [K]} \left\langle \mathbf{w}^j, \mathbf{x}^i \right\rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

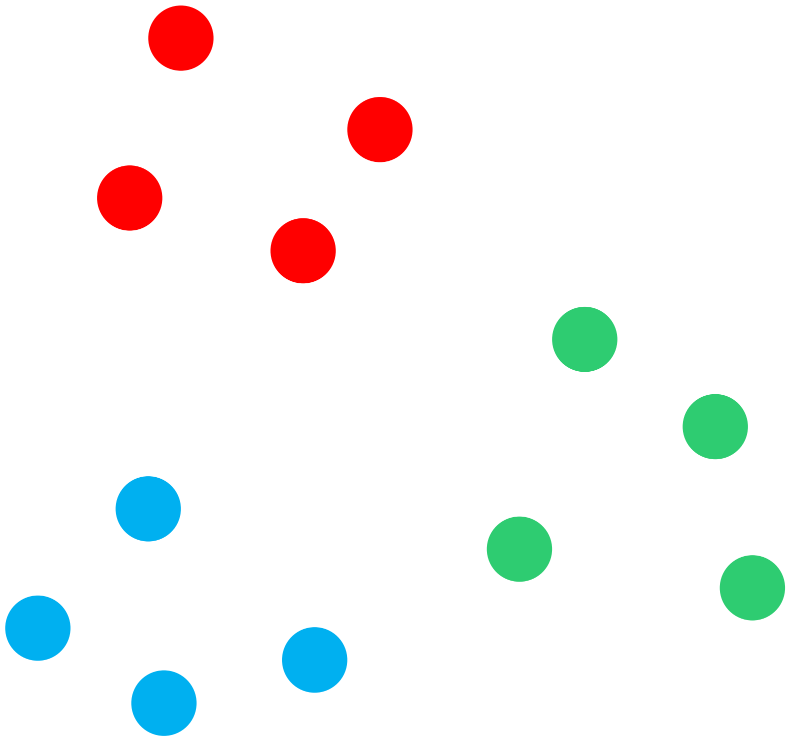
$$\hat{y}^i = \arg\max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

$\boxed{\text{One-vs-All (OVA)}}$

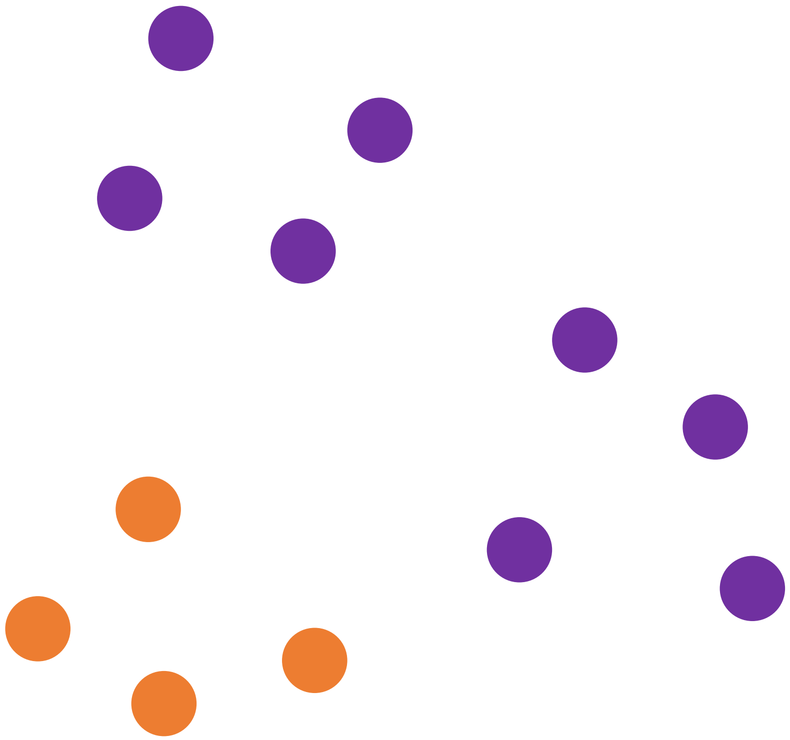
$$\hat{y}^i = \arg\max_{j \in [K]} \left\langle \mathbf{w}^j, \mathbf{x}^i \right\rangle$$



Multi-classification Loss Functions

$\boxed{\text{One-vs-All (OVA)}}$

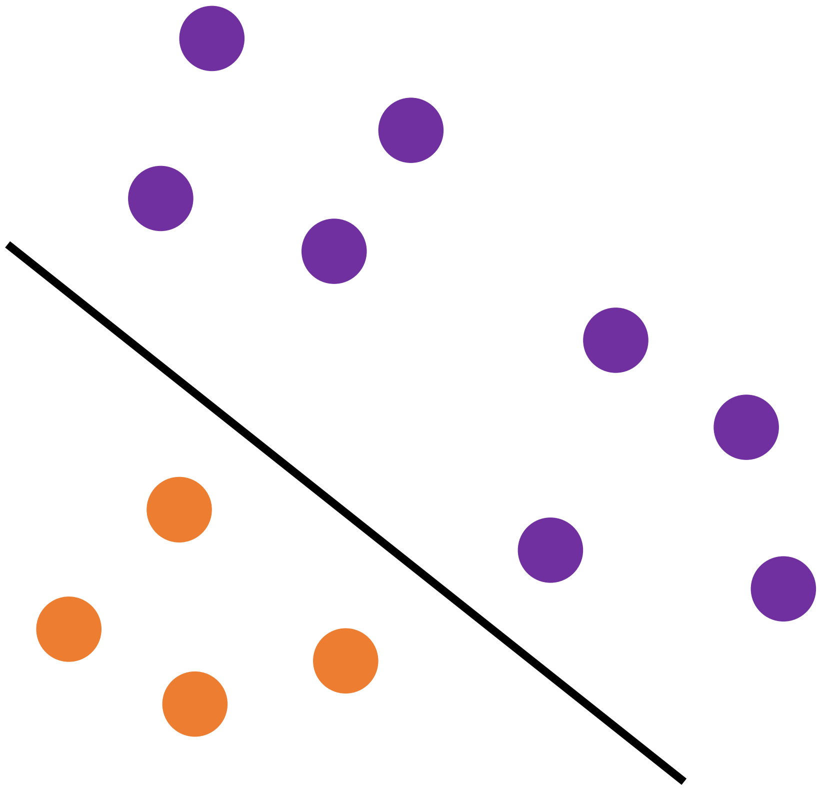
$$\hat{y}^i = \arg\max_{j \in [K]} \left\langle \mathbf{w}^j, \mathbf{x}^i \right\rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

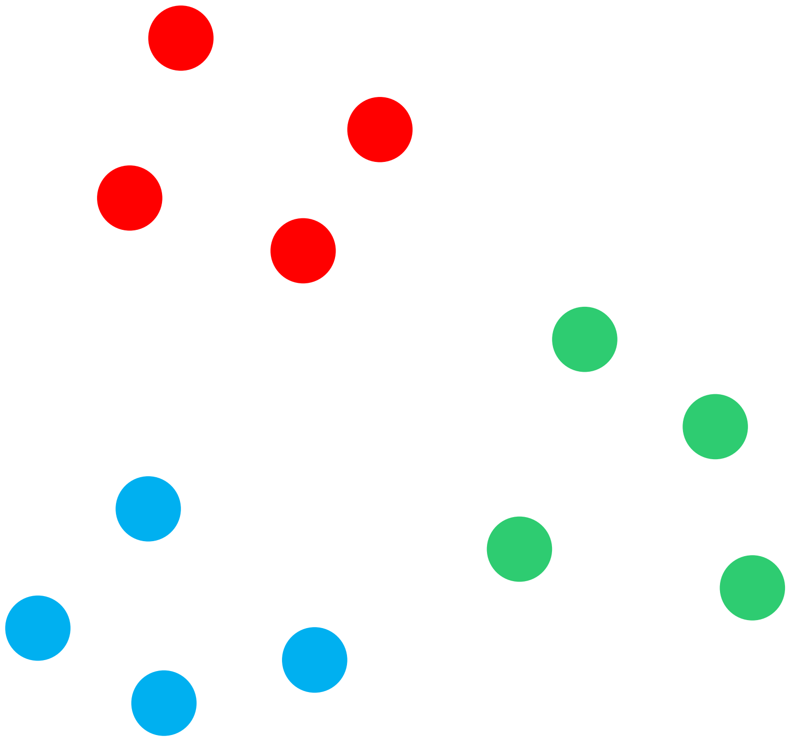
$$\hat{y}^i = \arg\max_{j \in [K]} \left\langle \mathbf{w}^j, \mathbf{x}^i \right\rangle$$



Multi-classification Loss Functions

$\boxed{\text{One-vs-All (OVA)}}$

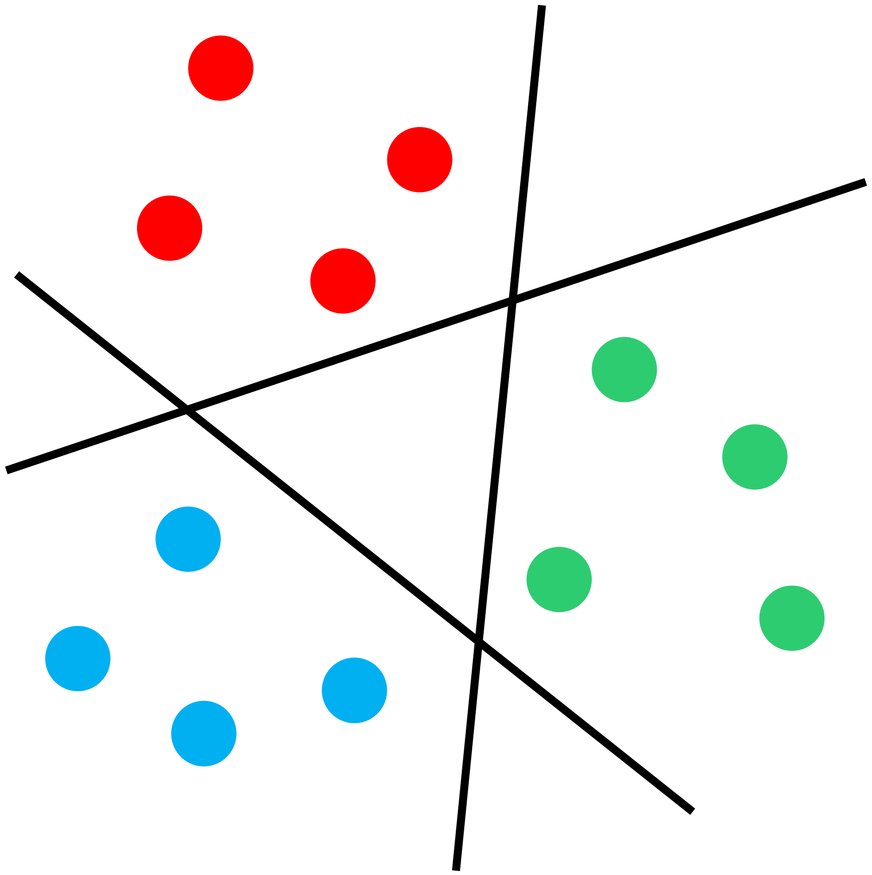
$$\hat{y}^i = \arg\max_{j \in [K]} \left\langle \mathbf{w}^j, \mathbf{x}^i \right\rangle$$



Multi-classification Loss Functions

$\boxed{\text{One-vs-All (OVA)}}$

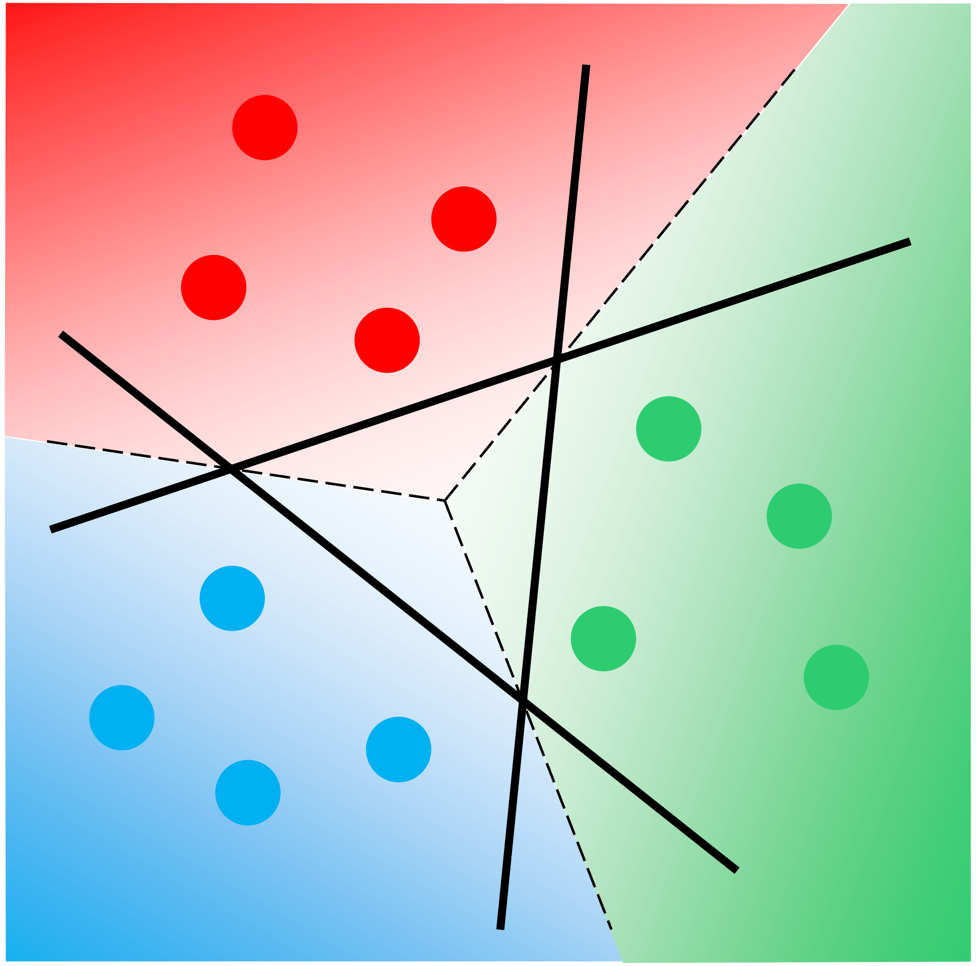
$$\hat{y}^i = \underset{j \in [K]}{\arg\max} \; \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

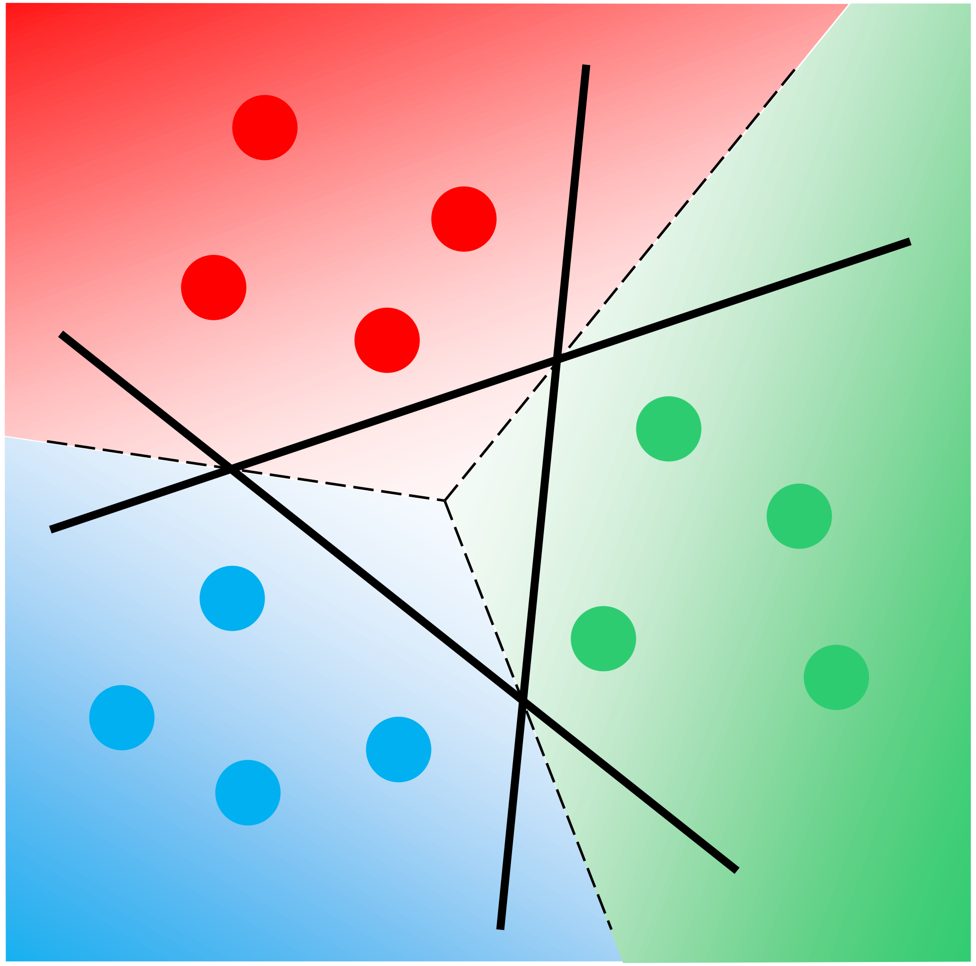
$$\hat{y}^i = \arg\max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

One-vs-All (OVA)

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



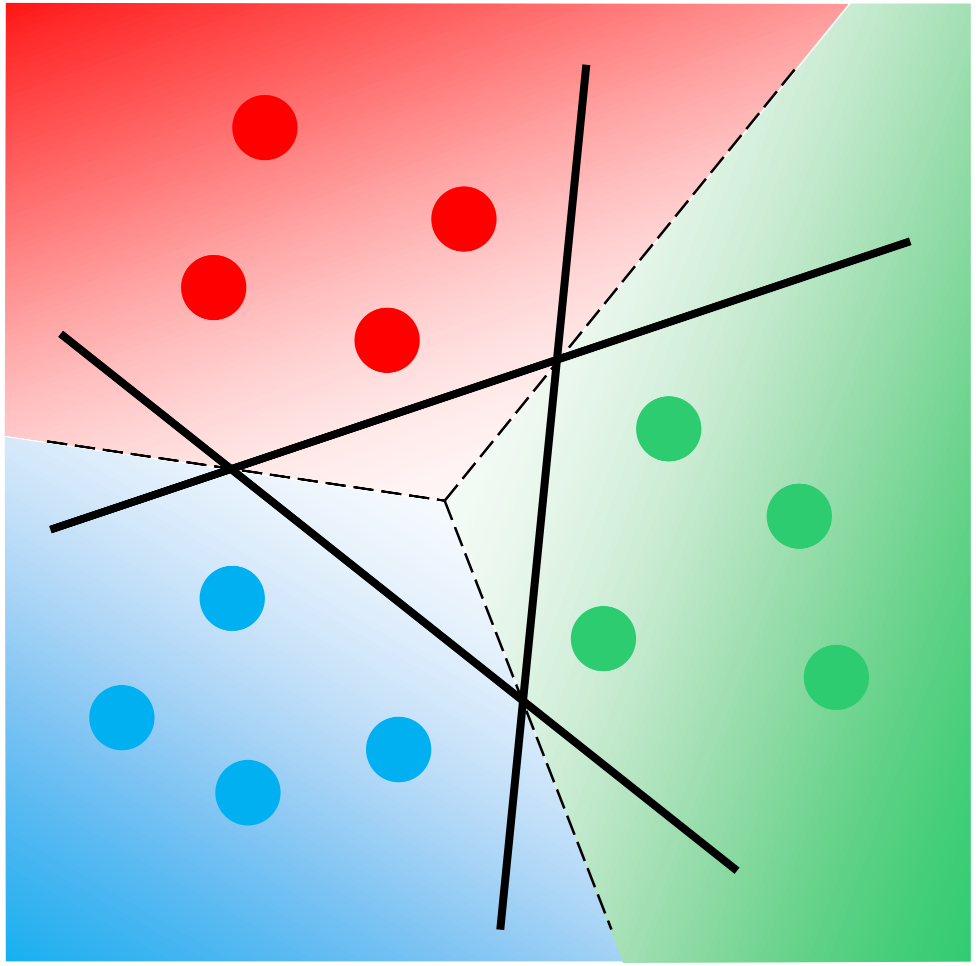
$$\hat{\mathbf{w}}^j = \underset{\mathbf{w}}{\arg\min} \; \sum_{i=1}^{n} \ell(y^{i,(j)}, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

$$y^{i,(j)} = \begin{cases} 1 & ; y^i = j \\ -1 & ; y^i \neq j \end{cases}$$

Multi-classification Loss Functions

One-vs-All (OVA)

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



$$\hat{\mathbf{w}}^j = \arg\min_{\mathbf{w}} \ \sum_{i=1}^{n} \ell(y^{i,(j)}, \langle \mathbf{w}, \mathbf{x}^i \rangle)$$

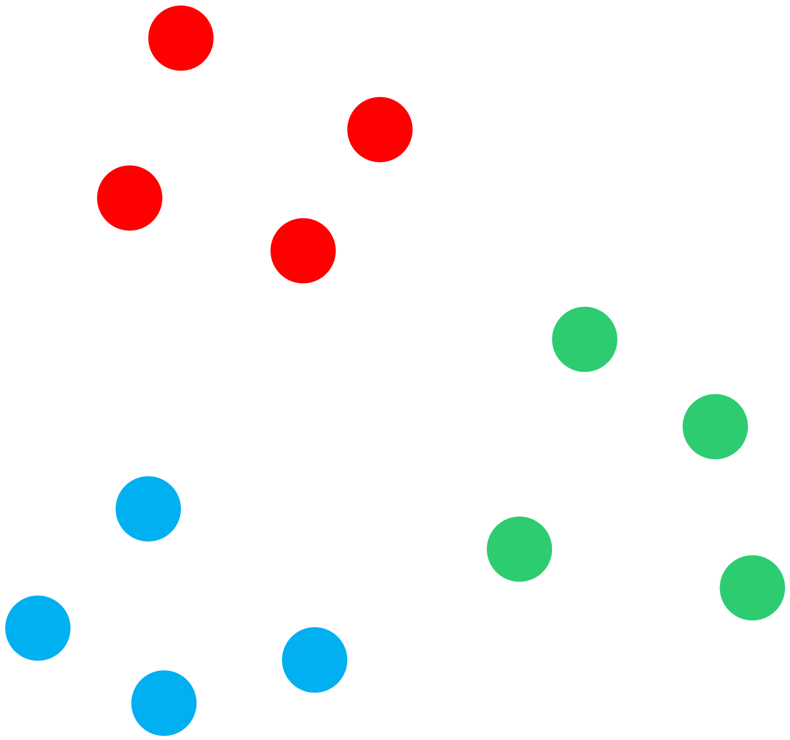
$$y^{i,(j)} = \begin{cases} 1 & ; y^i = j \\ -1 & ; y^i \neq j \end{cases}$$

hinge, logistic etc

Multi-classification Loss Functions

$\boxed{\text{One-vs-All (OVA)}}$

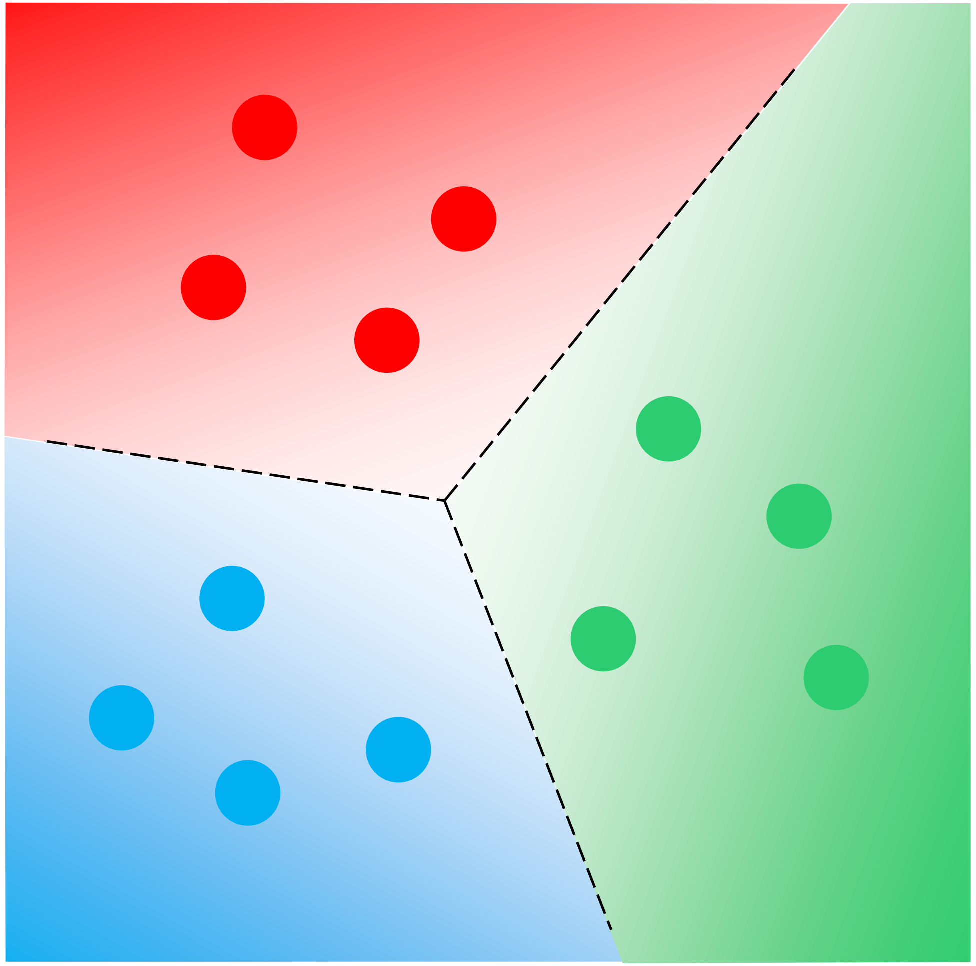
$$\hat{y}^i = \arg\max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



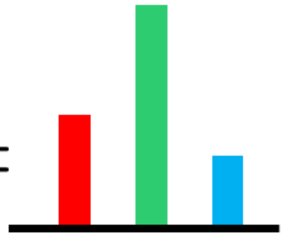
Multi-classification Loss Functions

One-vs-All (OVA)

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



Multi-classification using MLE

- $K > 2$ classes – need more detailed parameters
- For each point, its label profile is a vector $\boldsymbol{\eta}(\mathbf{x}) =$ 

$$\mathbb{P}\left[y^i = k \,|\, \mathbf{x}^i, \{\mathbf{w}^l\}_{1,\ldots,K}\right] \propto \exp(\langle \mathbf{w}^k, \mathbf{x}^i \rangle)$$

$$\mathbb{P}\left[y^i = k \,|\, \mathbf{x}^i, \{\mathbf{w}^l\}_{1,\ldots,K}\right] = \frac{\exp(\langle \mathbf{w}^k, \mathbf{x}^i \rangle)}{\sum_{l=1}^{K} \exp(\langle \mathbf{w}^l, \mathbf{x}^i \rangle)}$$

- Likelihood function is multinomial instead of binomial

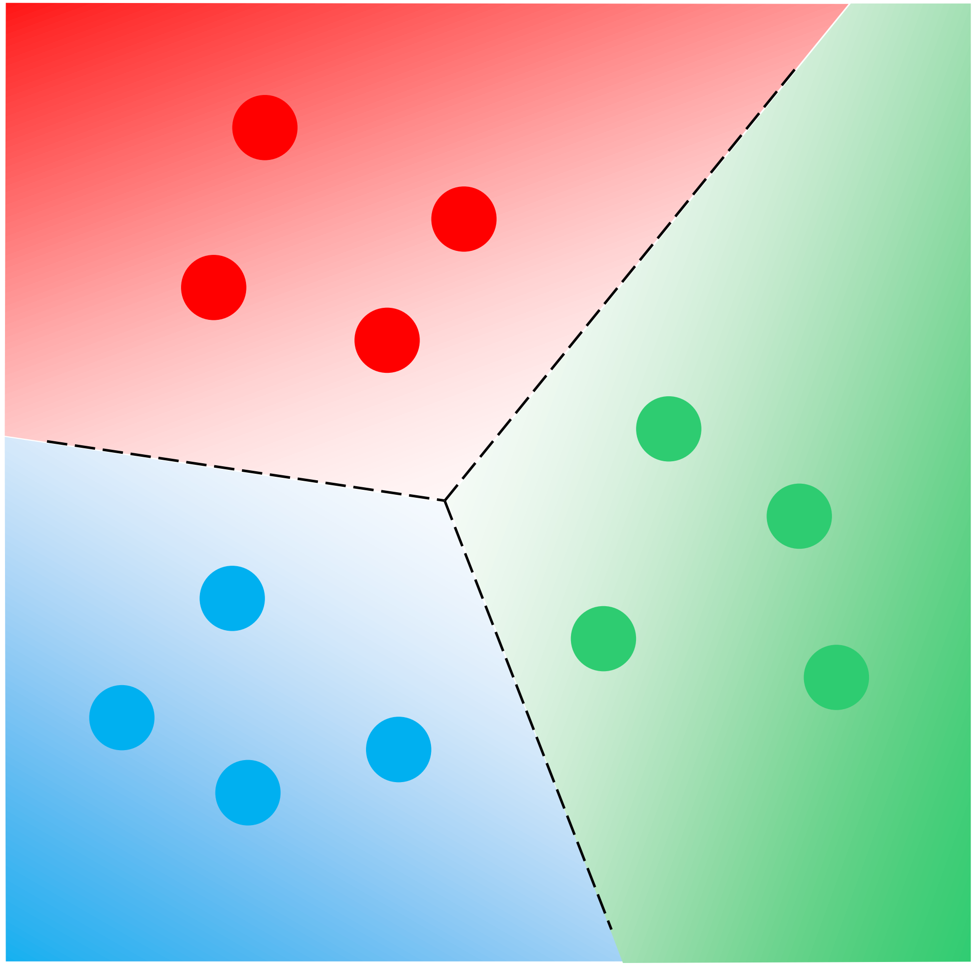
$$\mathbb{P}\left[\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}\right] = \prod_{i=1}^{n} \hat{\boldsymbol{\eta}}^i_{y^i}(\mathbf{x}) \qquad \hat{\boldsymbol{\eta}}^i_k(\mathbf{x}) = \frac{\exp(\langle \mathbf{w}^k, \mathbf{x}^i \rangle)}{\sum_{l=1}^{K} \exp(\langle \mathbf{w}^l, \mathbf{x}^i \rangle)}$$

Softmax Regression

Multi-classification Loss Functions

One-vs-All (OVA)

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

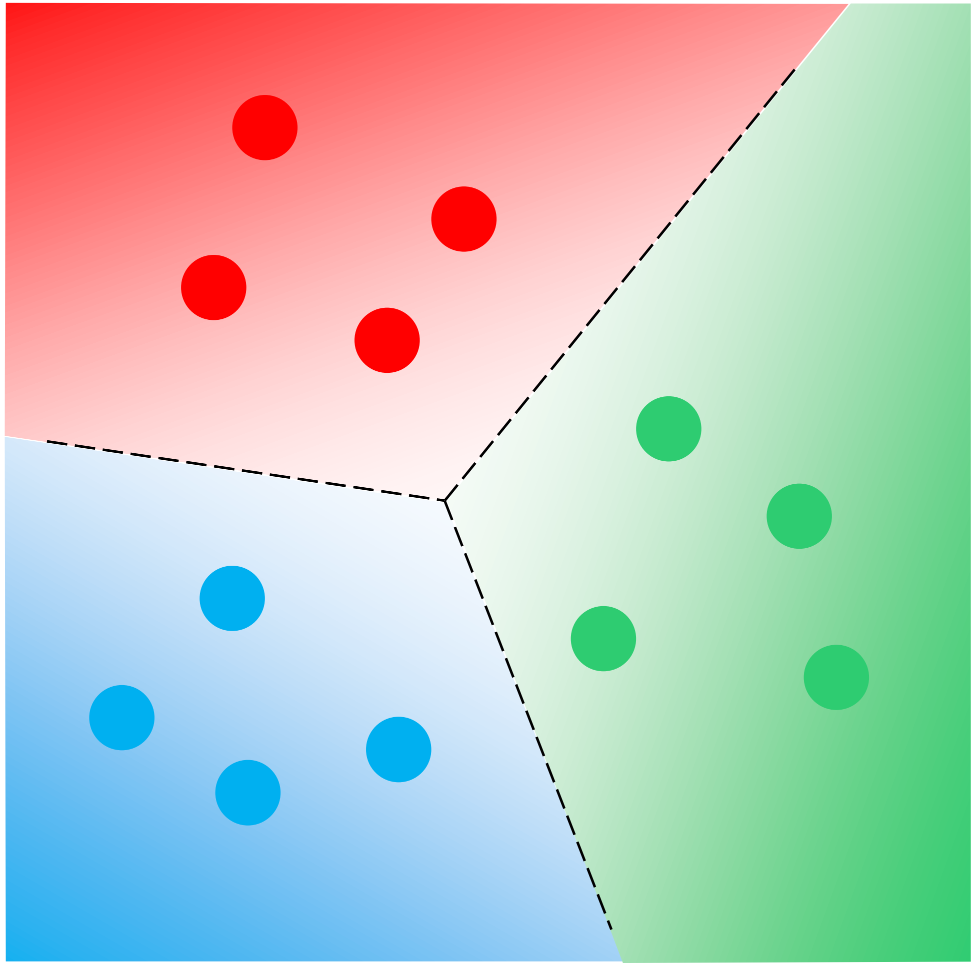


Multi-classification Loss Functions

One-vs-All (OVA)

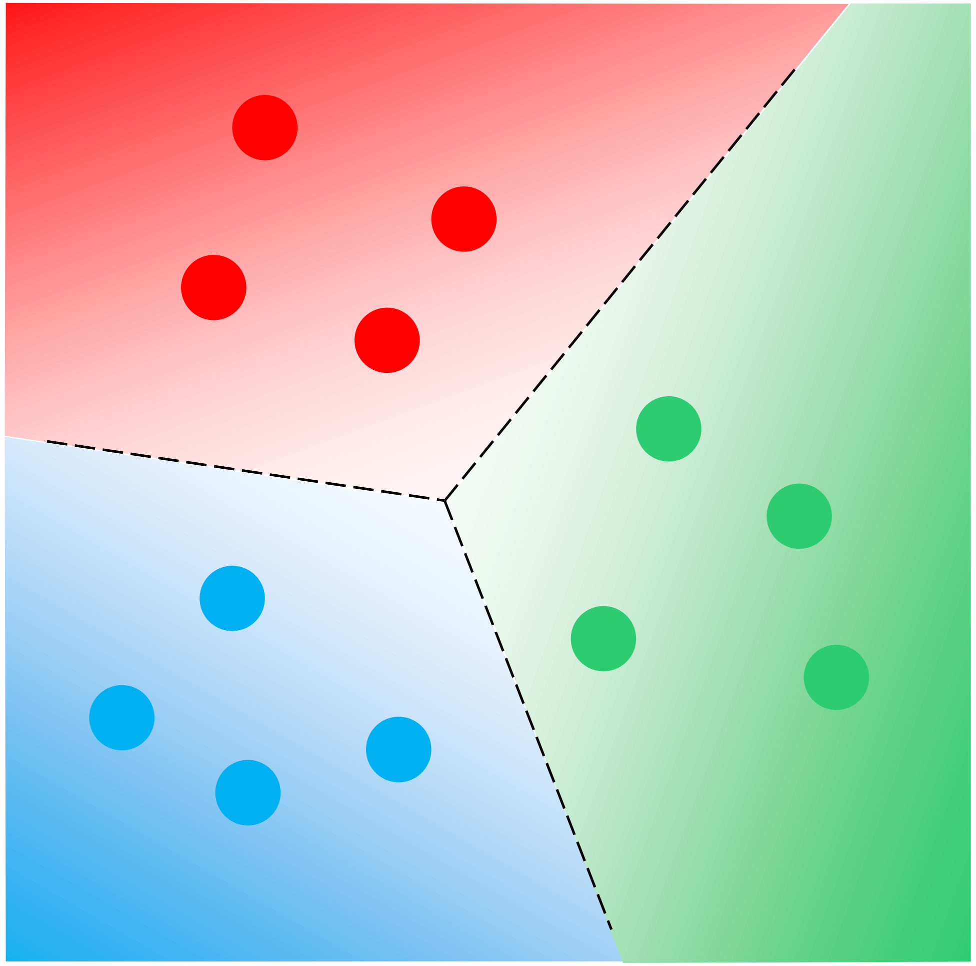
$$\hat{y}^i = \arg\max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$

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



Multi-classification Loss Functions

One-vs-All (OVA)



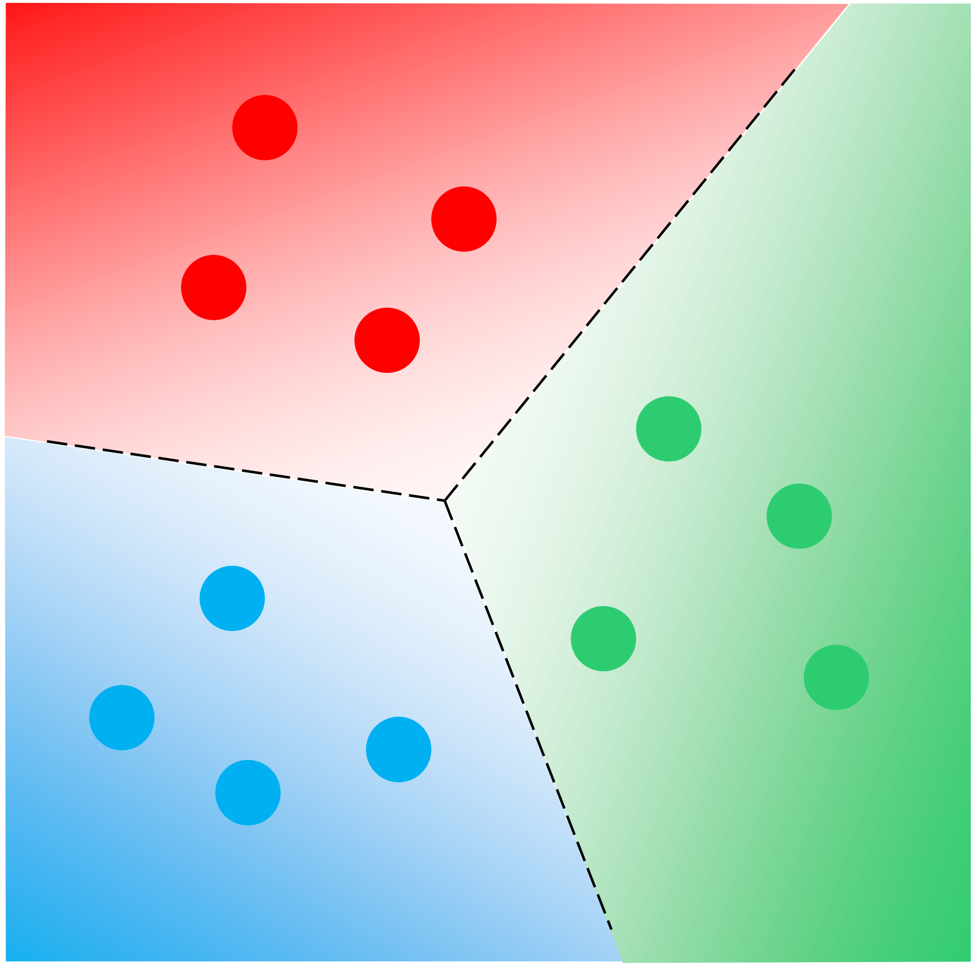
$$\hat{y}^i = \underset{j \in [K]}{\arg\max} \; \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$

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

$$\widehat{\mathbf{W}}_{\text{MLE}} = \underset{\mathbf{W}}{\arg\min} \; \sum_{i=1}^{n} \ell_{\text{sm}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle)$$

Multi-classification Loss Functions

One-vs-All (OVA)



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

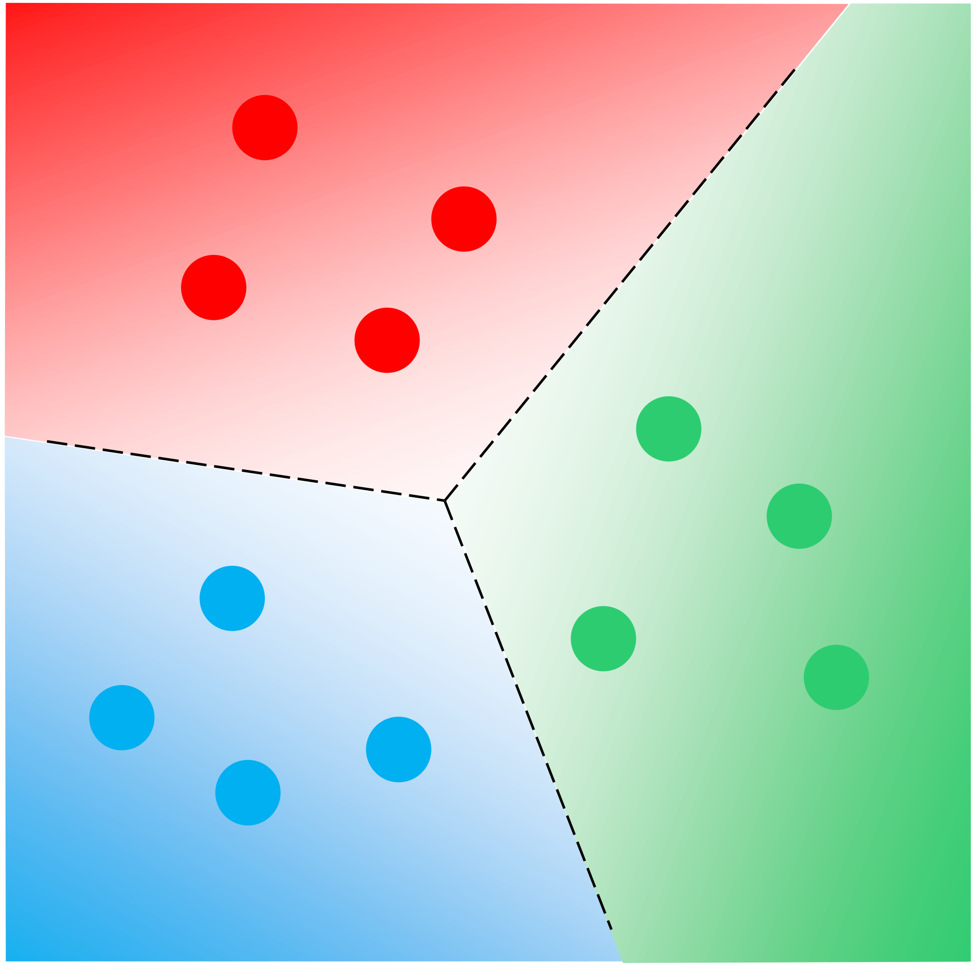
$$\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^K]$$

$$\widehat{\mathbf{W}}_{\text{MLE}} = \arg\min_{\mathbf{W}} \sum_{i=1}^{n} \ell_{\text{sm}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle)$$

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

Multi-classification Loss Functions

One-vs-All (OVA)



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

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

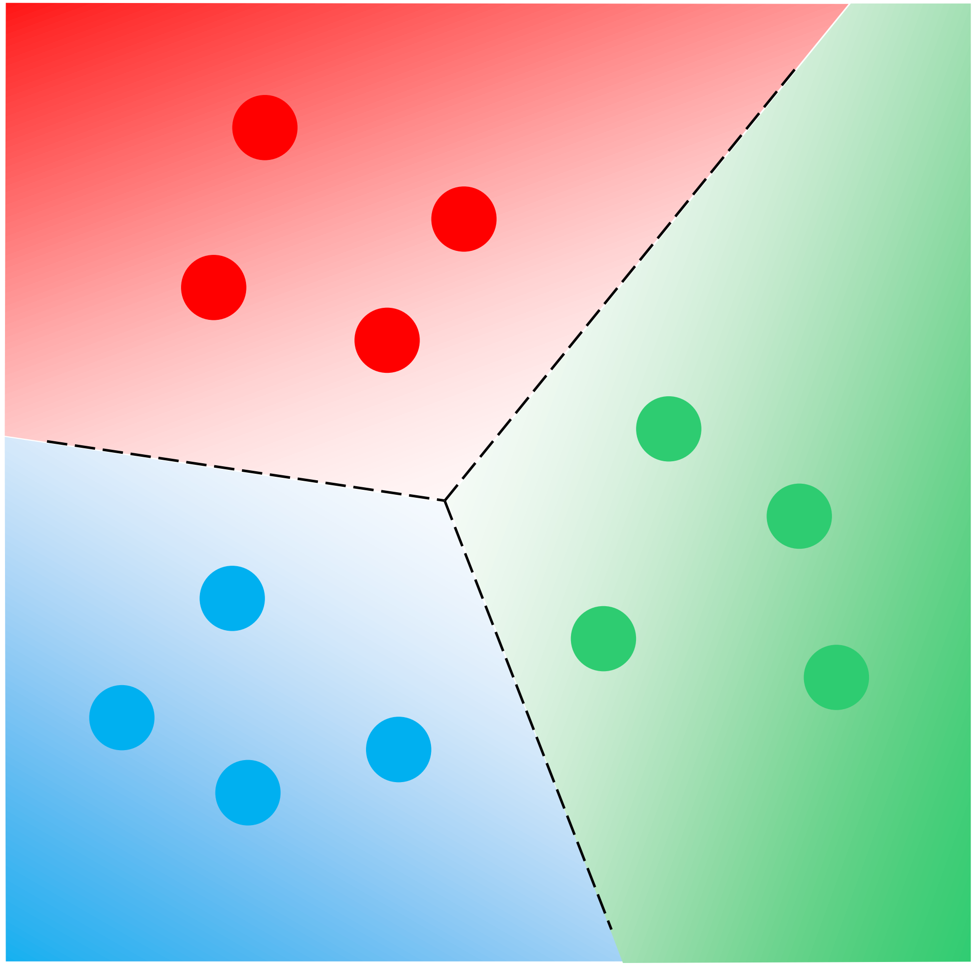
$$\widehat{\mathbf{W}}_{\mathrm{MLE}} = \underset{\mathbf{W}}{\arg\min} \sum_{i=1}^{n} \ell_{\mathrm{sm}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle)$$

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

$$\ell_{\mathrm{sm}}(y, \{\boldsymbol{\eta}_j\}) = -\log \frac{\exp(\boldsymbol{\eta}_y)}{\sum_{k=1}^{K} \exp(\boldsymbol{\eta}_k)}$$

Multi-classification Loss Functions

One-vs-All (OVA)



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

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

$$\widehat{\mathbf{W}}_{\text{MLE}} = \arg\min_{\mathbf{W}} \sum_{i=1}^{n} \ell_{\text{sm}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle)$$

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

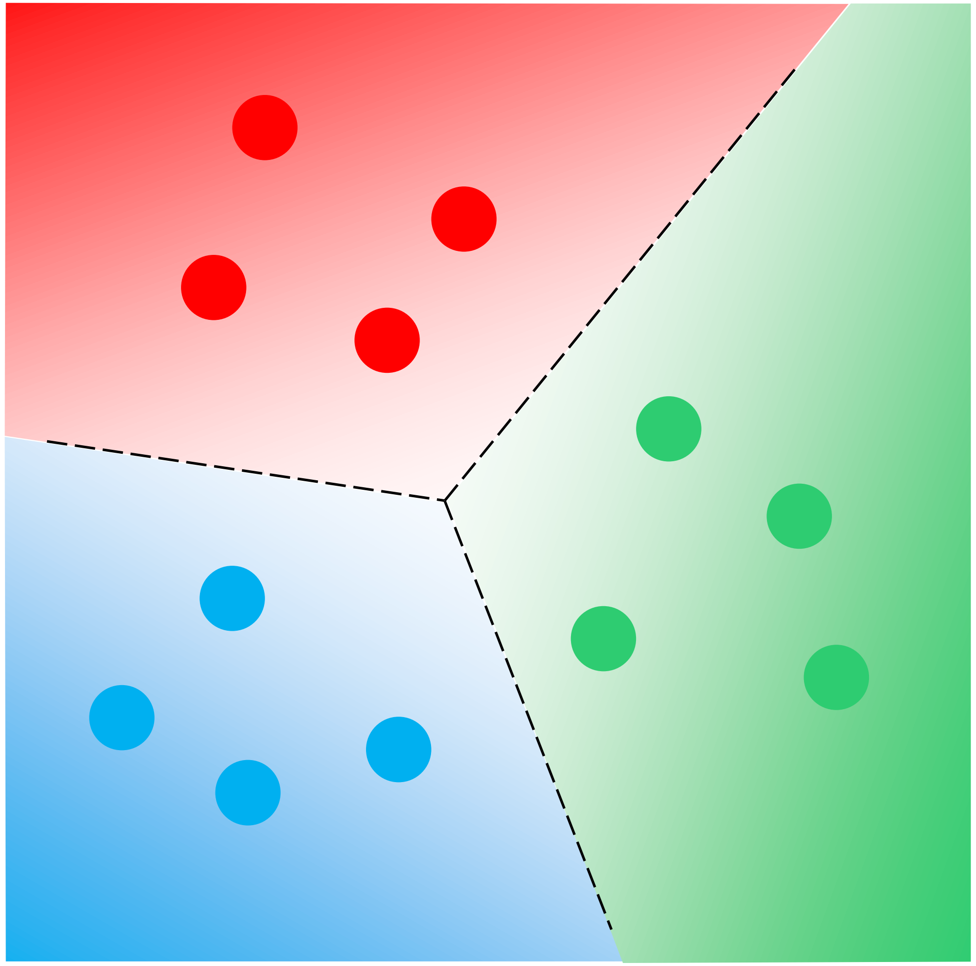
$$\ell_{\text{sm}}(y, \{\boldsymbol{\eta}_j\}) = -\log \frac{\exp(\boldsymbol{\eta}_y)}{\sum_{k=1}^{K} \exp(\boldsymbol{\eta}_k)}$$

Softmax loss function

Multi-classification Loss Functions

One-vs-All (OVA)

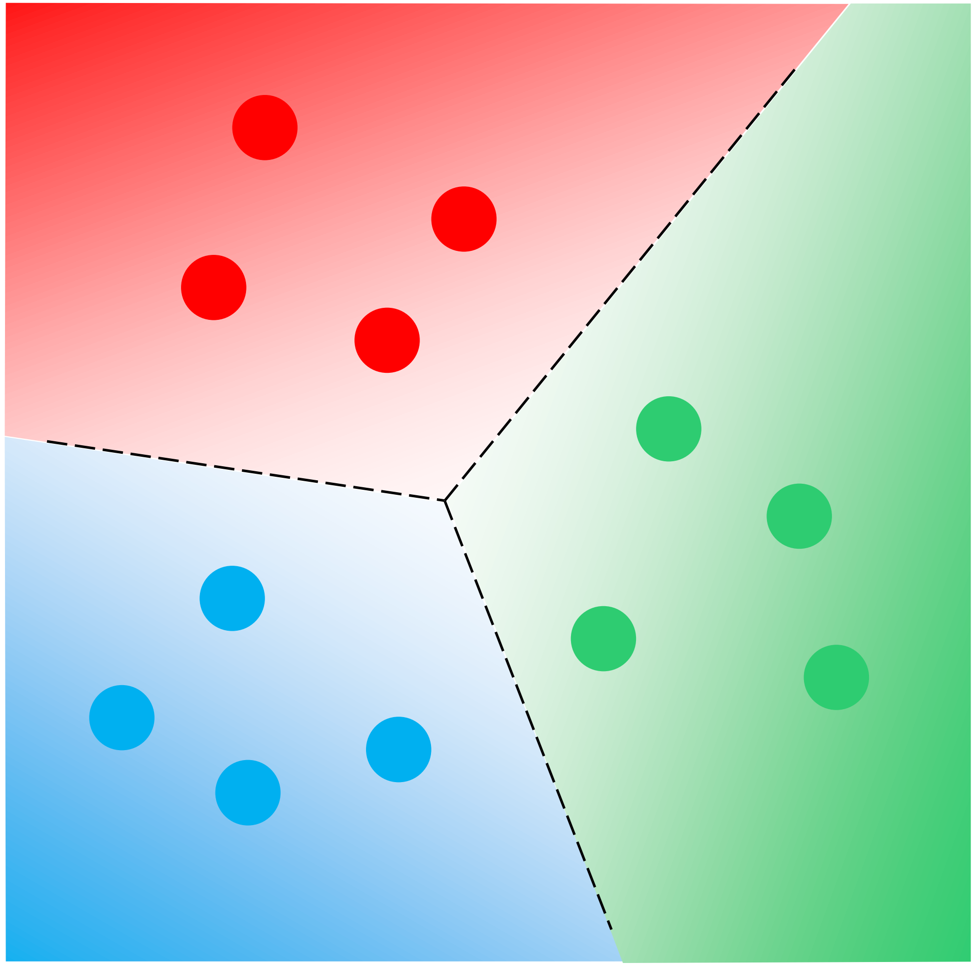
$$\hat{y}^i = \arg\max_{j \in [K]} \langle \mathbf{w}^j, \mathbf{x}^i \rangle$$



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

Multi-classification Loss Functions

One-vs-All (OVA)



$$\hat{y}^i = \underset{j \in [K]}{\arg\max} \; \left\langle \mathbf{w}^j, \mathbf{x}^i \right\rangle$$

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

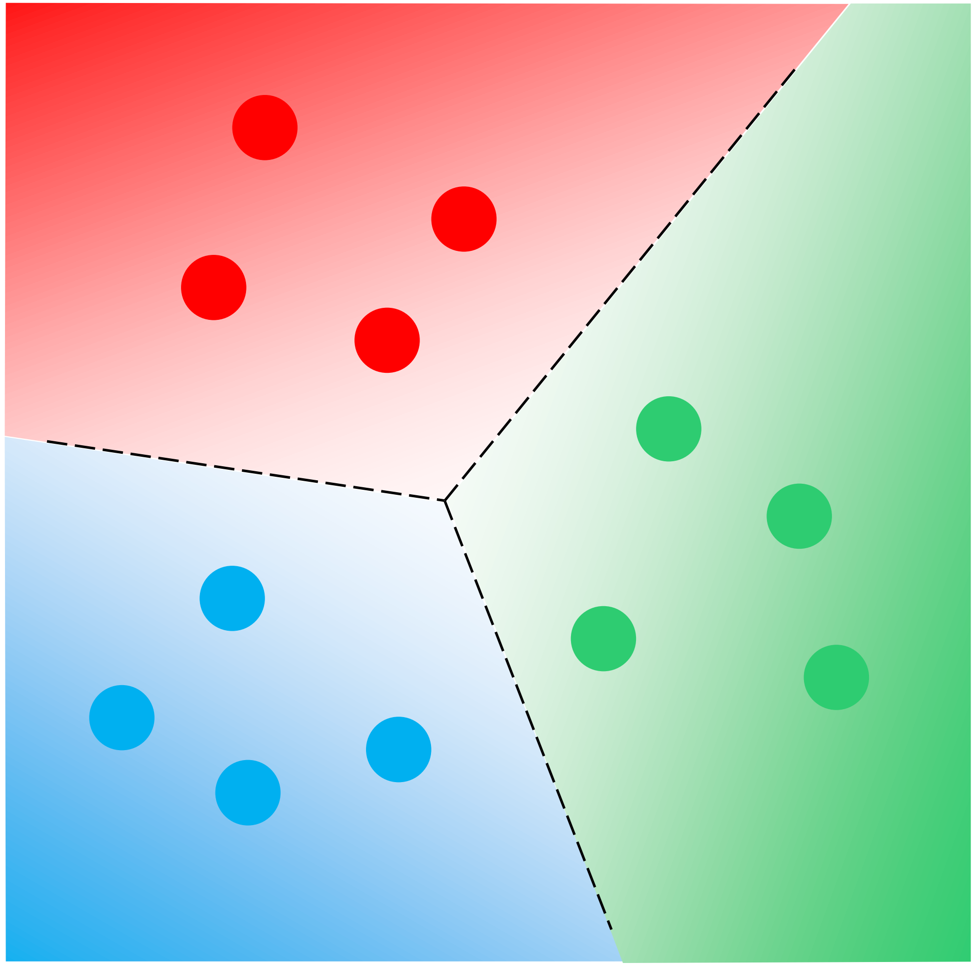
$$\widehat{\mathbf{W}} = \underset{\mathbf{W}}{\arg\min} \; \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2$$

$$\text{s.t.} \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1$$

Multi-classification Loss Functions

One-vs-All (OVA)

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



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

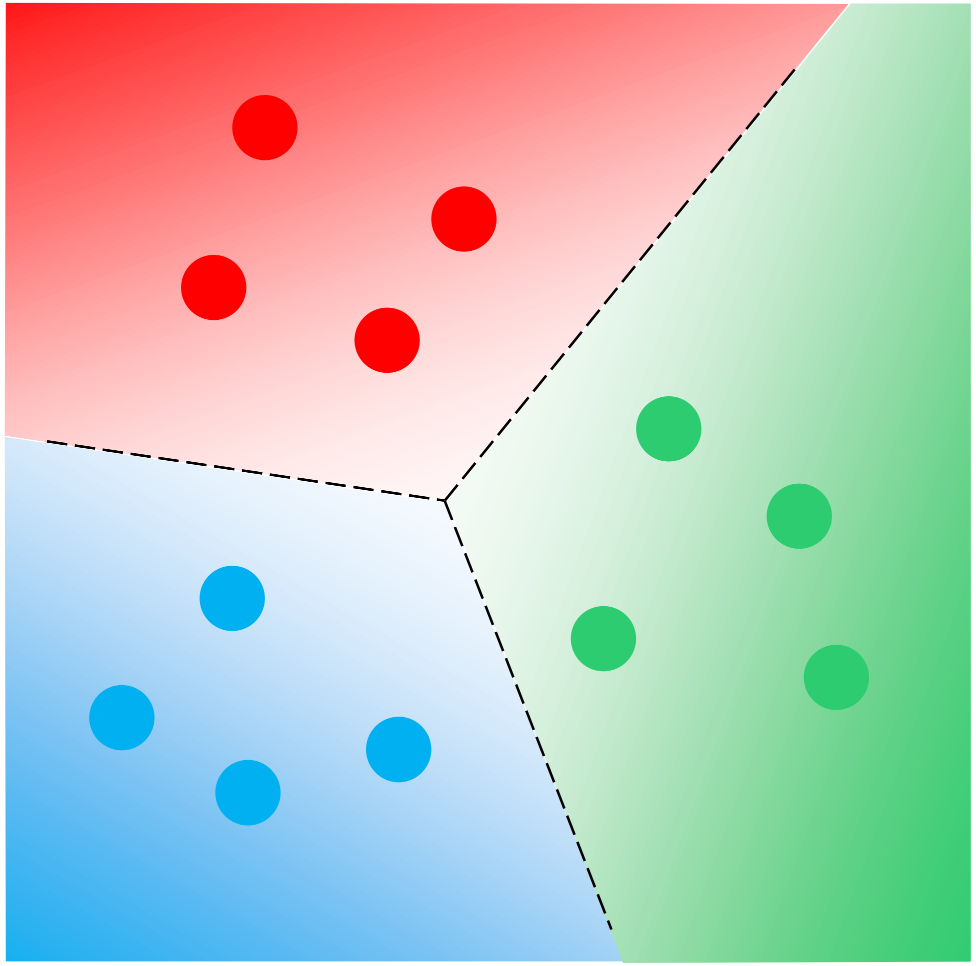
$$\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2$$

$$\text{s.t.} \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1$$

$$\forall k \neq y^i$$

Multi-classification Loss Functions

One-vs-All (OVA)



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

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

Slack variable

$$\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2$$

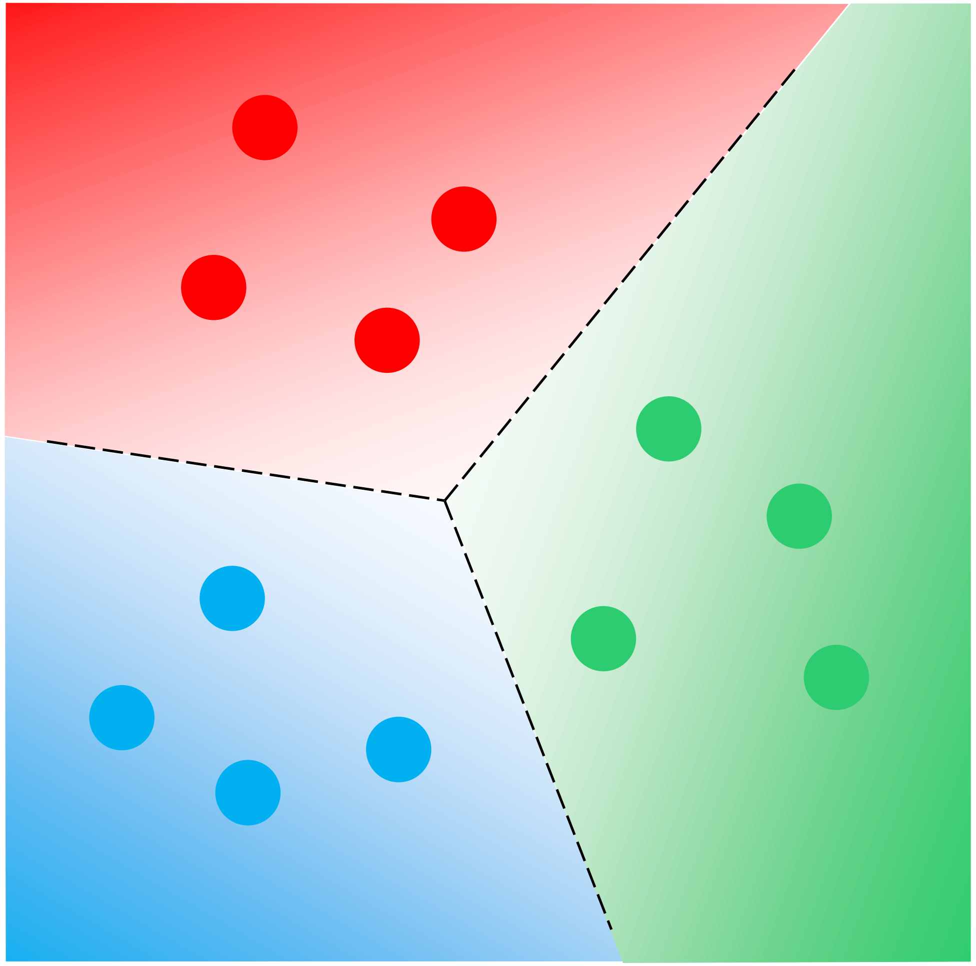
$$\text{s.t.} \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \xi_i$$

$$\forall k \neq y^i$$

Multi-classification Loss Functions

One-vs-All (OVA)

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



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

Slack variable

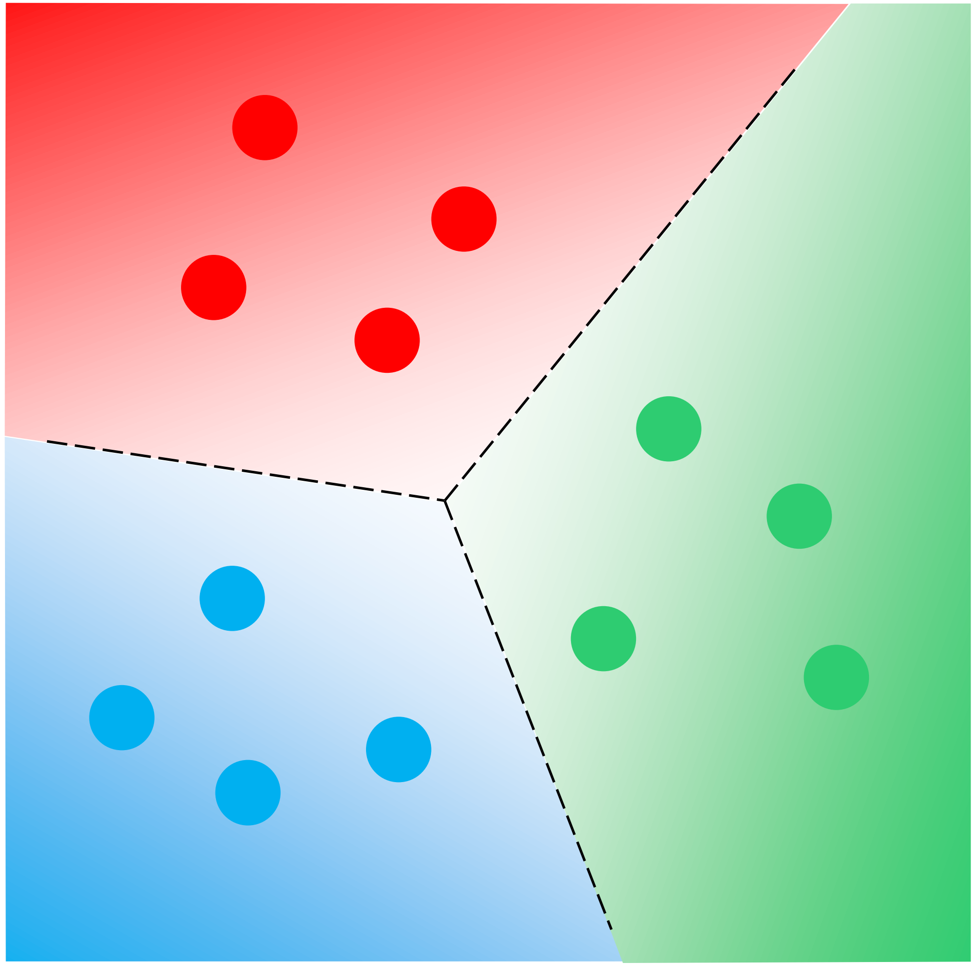
$$\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \left\| \mathbf{w}^k \right\|_2^2$$

$$\text{s.t.} \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \xi_i \quad \forall k \neq y^i$$

$$\xi_i \geq 0$$

Multi-classification Loss Functions

One-vs-All (OVA)



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

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

Slack variable

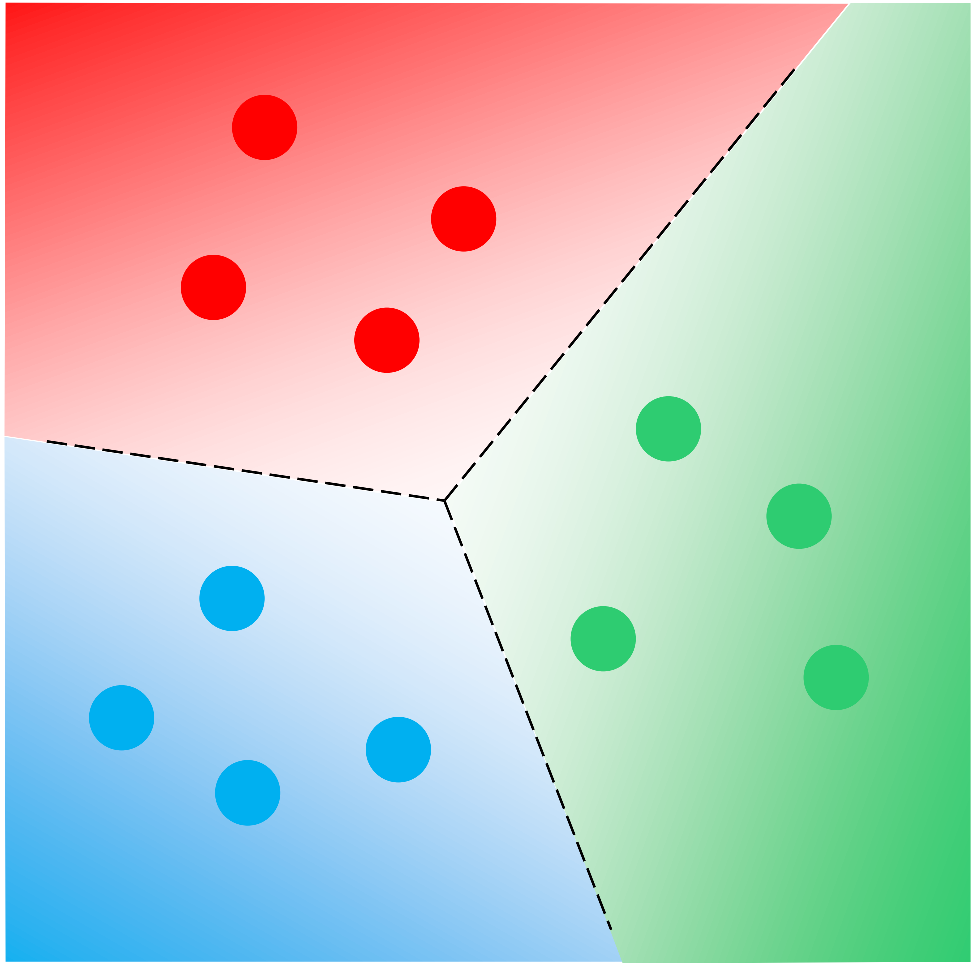
$$\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \xi_i \quad \forall k \neq y^i$$

$$\xi_i \geq 0$$

Multi-classification Loss Functions

One-vs-All (OVA)



$$\hat{y}^i = \underset{j \in [K]}{\arg\max} \; \left\langle \mathbf{w}^j, \mathbf{x}^i \right\rangle$$

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

$$\widehat{\mathbf{W}} = \underset{\mathbf{W}, \{\xi_i\}}{\arg\min} \; \sum_{k=1}^{K} \left\| \mathbf{w}^k \right\|_2^2 + \sum_{i=1}^{n} \xi_i$$

Slack variable

$$\text{s.t.} \left\langle \mathbf{w}^{y^i}, \mathbf{x}^i \right\rangle \geq \left\langle \mathbf{w}^k, \mathbf{x}^i \right\rangle + 1 - \xi_i$$

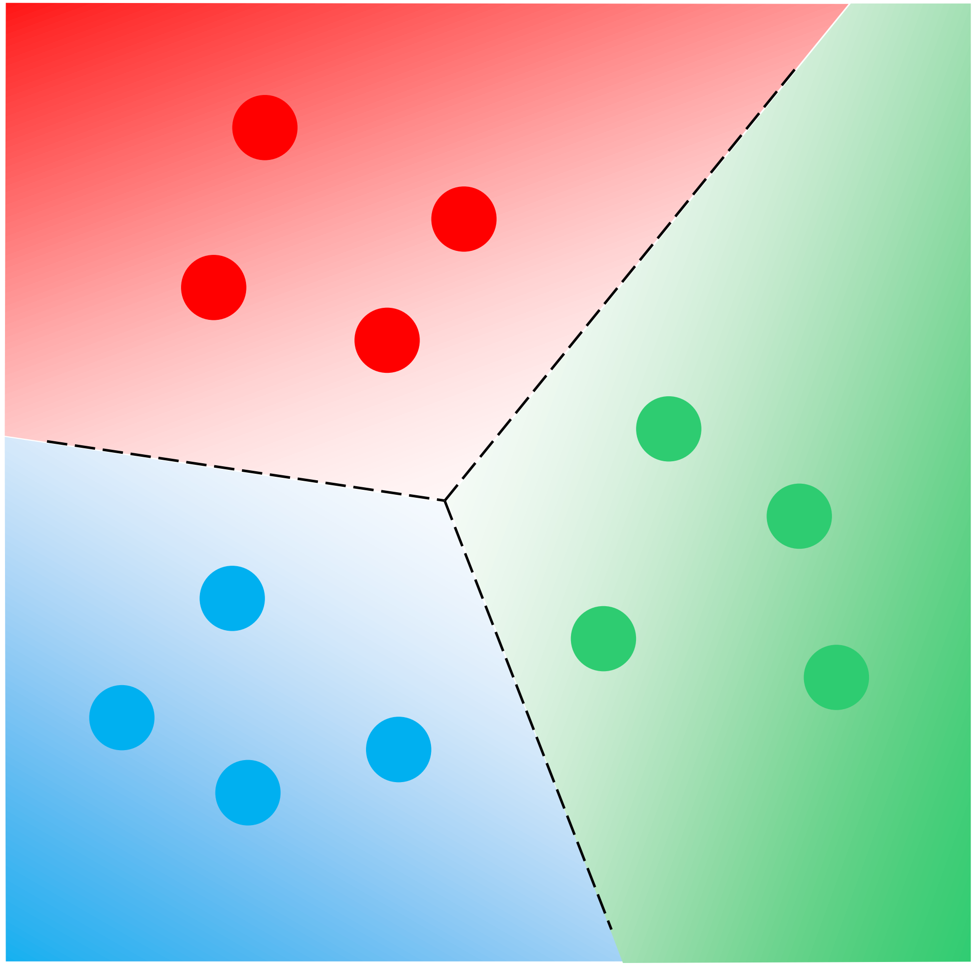
$$\forall k \neq y^i$$

$$\xi_i \geq 0$$

Multi-classification Loss Functions

One-vs-All (OVA)

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

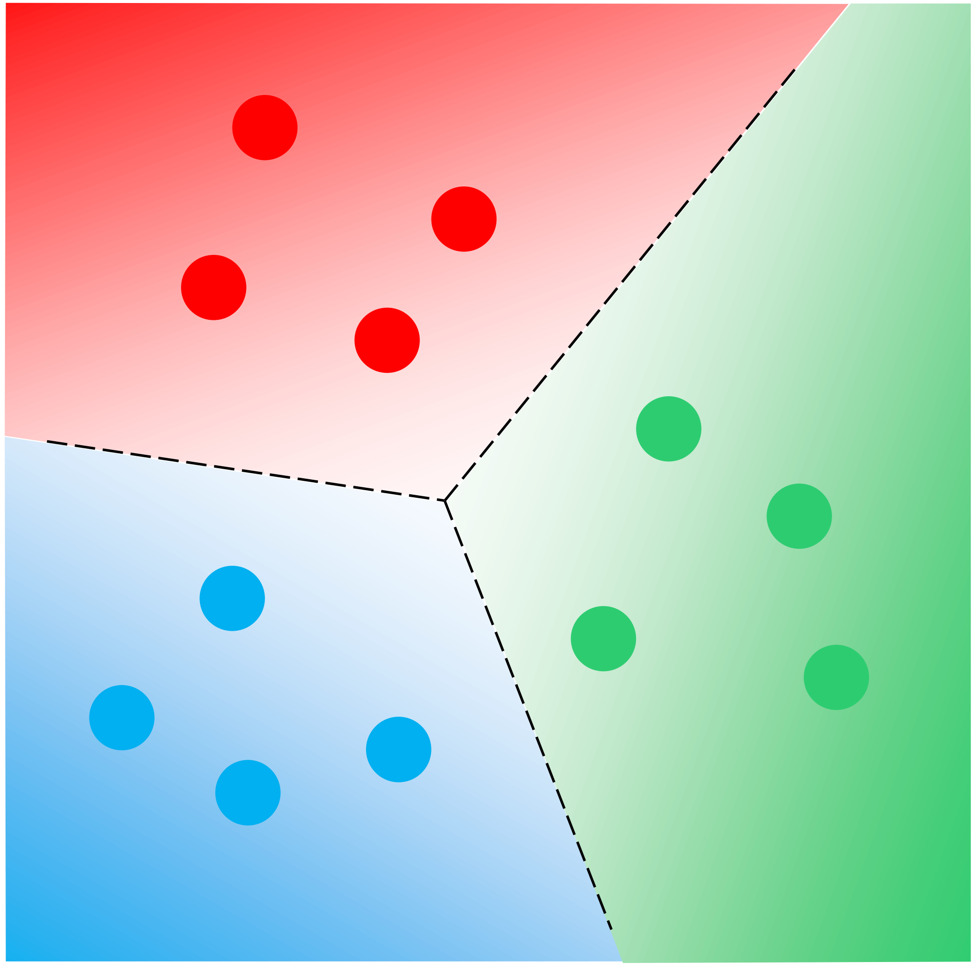


$$\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^K]$$

$$\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}} \; \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle)$$

Multi-classification Loss Functions

One-vs-All (OVA)



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

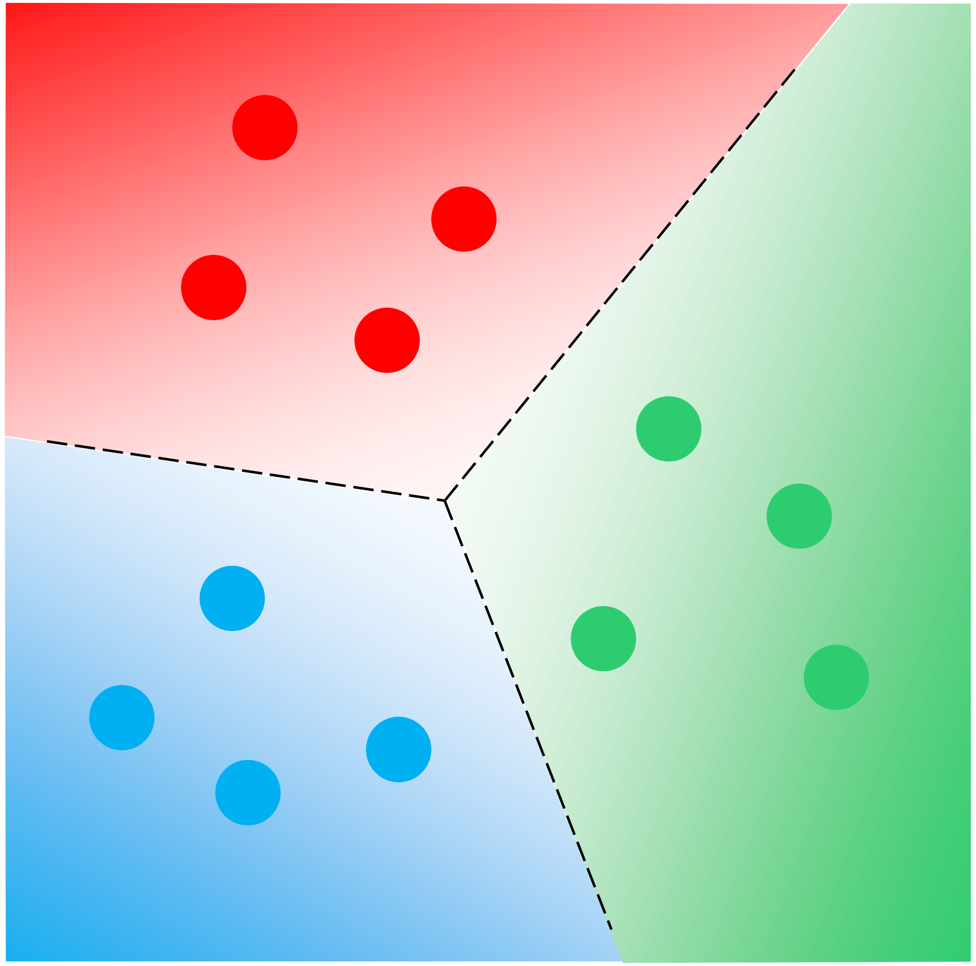
$$\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^K]$$

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

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

Multi-classification Loss Functions

One-vs-All (OVA)



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

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

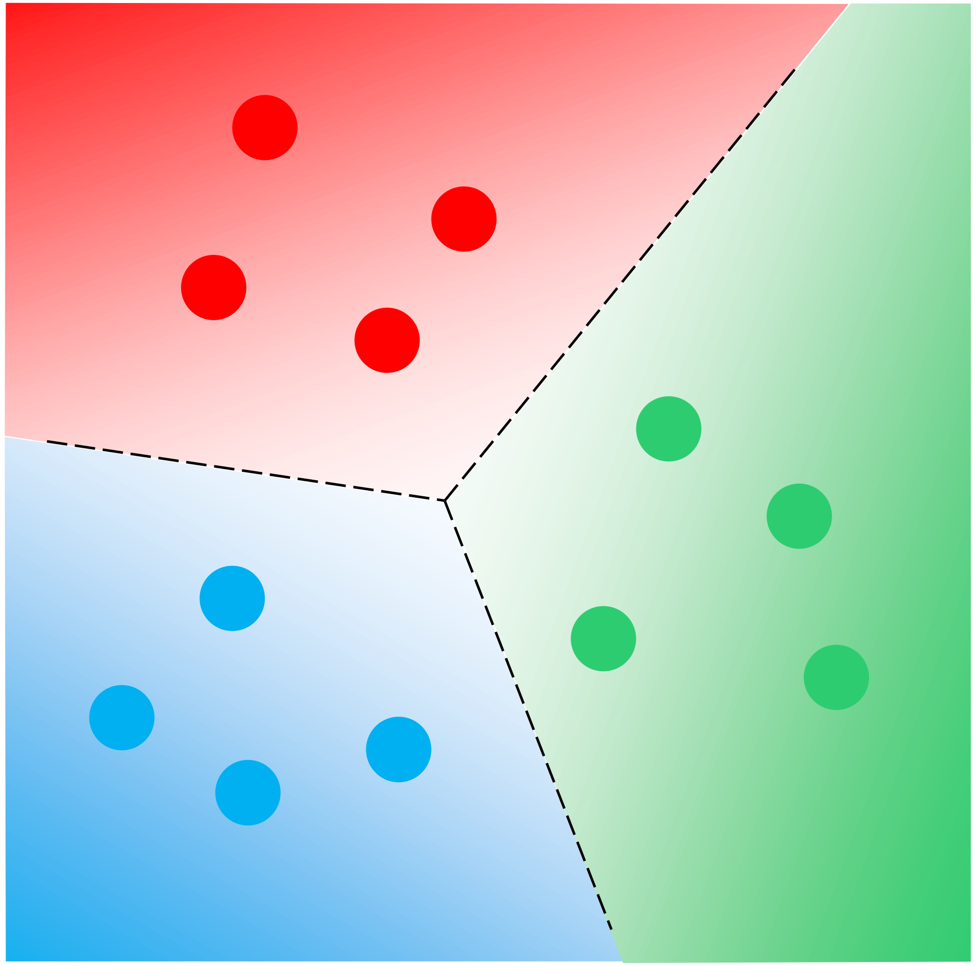
$$\widehat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} \left\|\mathbf{w}^k\right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle)$$

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

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

Multi-classification Loss Functions

One-vs-All (OVA)



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

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

$$\widehat{\mathbf{W}} = \underset{\mathbf{W}}{\arg\min} \; \sum_{k=1}^{K} \left\| \mathbf{w}^k \right\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle)$$

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

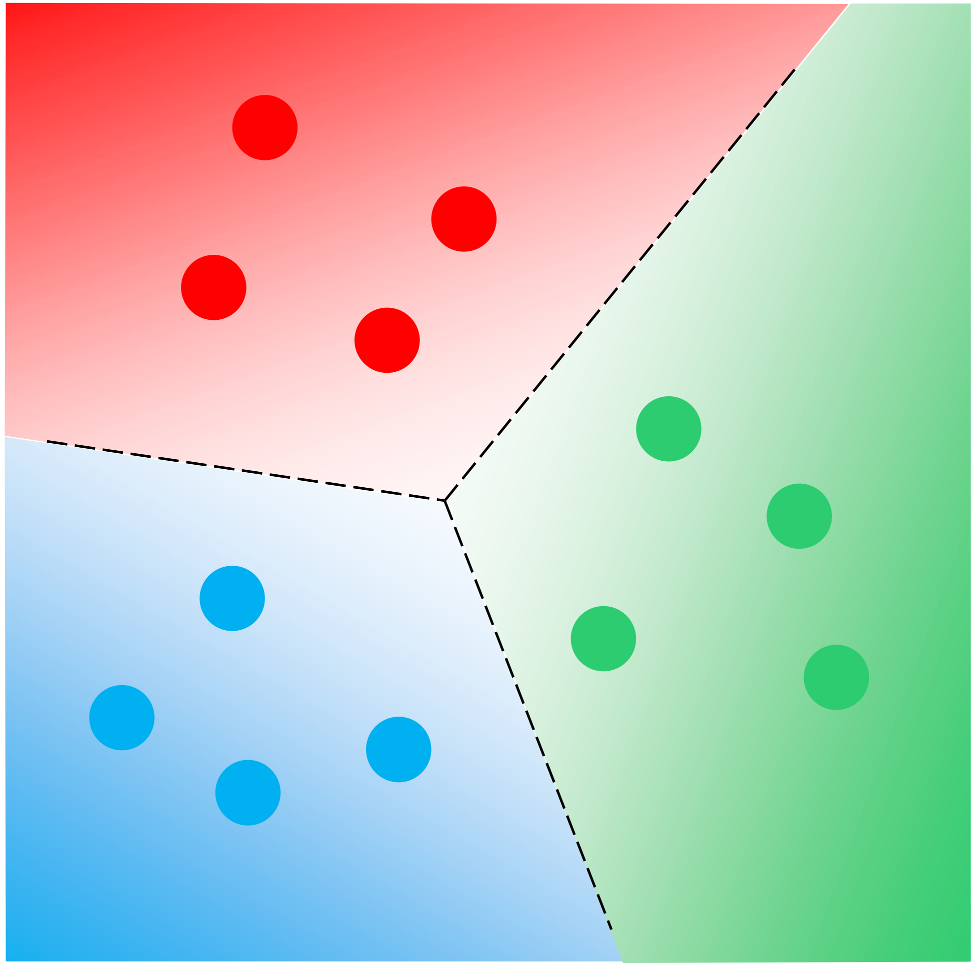
$$\ell_{\text{cs}}(y, \{\boldsymbol{\eta}_j\}) = [1 + \max_{k \neq y} \boldsymbol{\eta}_k - \boldsymbol{\eta}_y]_+$$

Exercise

Multi-classification Loss Functions

One-vs-All (OVA)

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



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

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

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

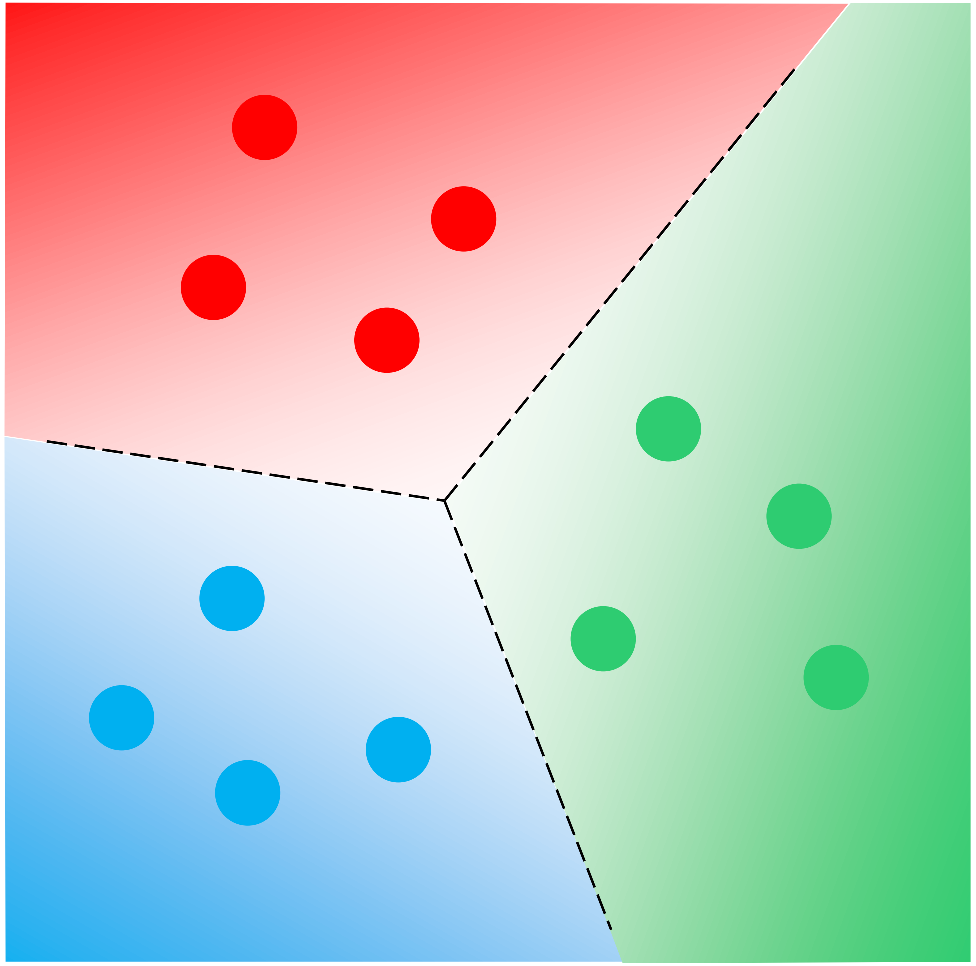
$$\ell_{\text{cs}}(y, \{\boldsymbol{\eta}_j\}) = [1 + \max_{k \neq y} \boldsymbol{\eta}_k - \boldsymbol{\eta}_y]_+$$

Exercise

Crammer-Singer loss function

Multi-classification Loss Functions

One-vs-All (OVA)



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

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

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

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

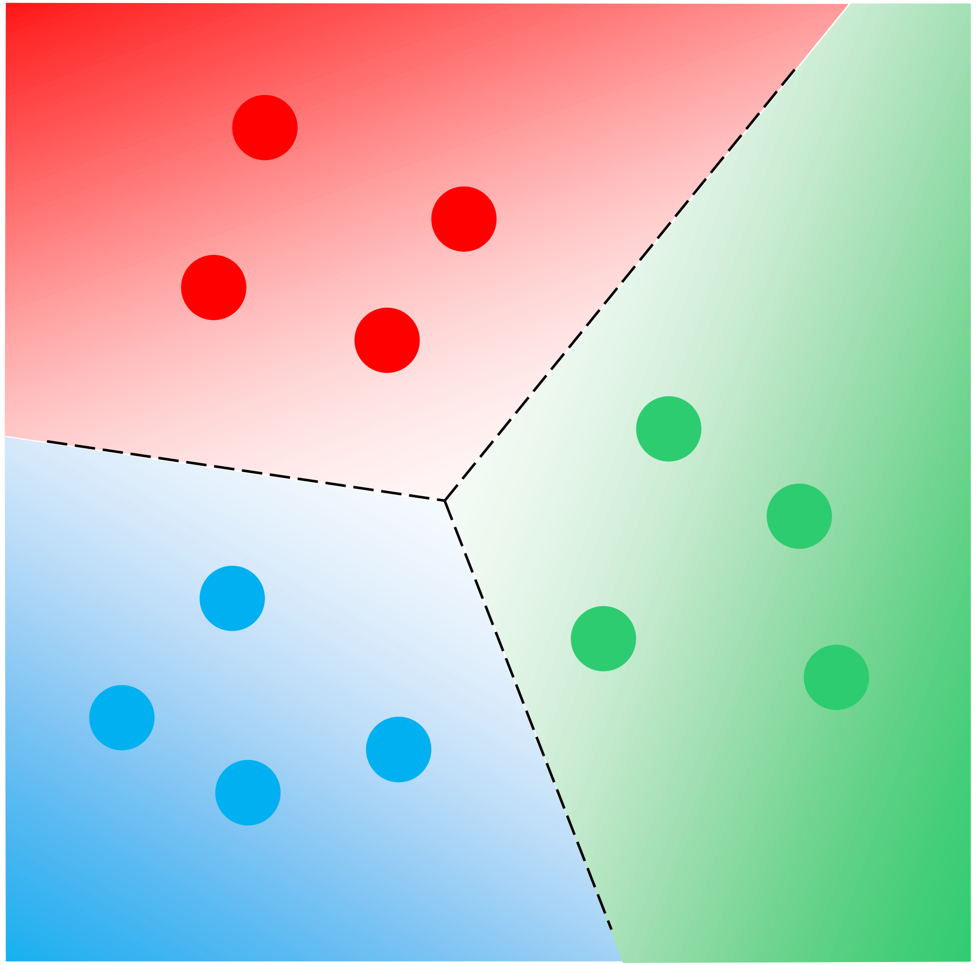
$$\ell_{\text{cs}}(y, \{\boldsymbol{\eta}_j\}) = [1 + \max_{k \neq y} \boldsymbol{\eta}_k - \boldsymbol{\eta}_y]_+$$

Exercise

Multi-classification Loss Functions

One-vs-All (OVA)

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



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

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

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

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

Exercise

More powerful regularizers?

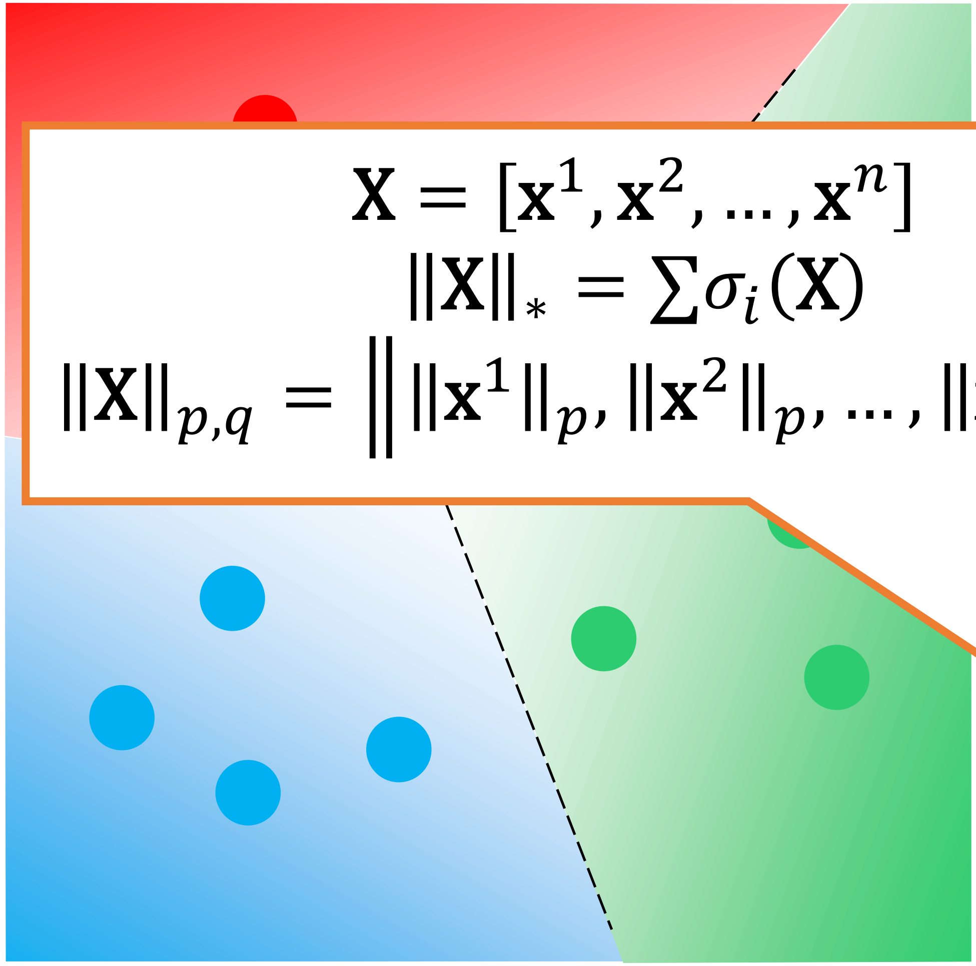
$\|\mathbf{W}\|_F^2, \|\mathbf{W}\|_*, \|\mathbf{W}\|_0, \|\mathbf{W}^\top\|_{2,0}$

Multi-classification Loss Functions

One-vs-All (OVA)

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

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



$$\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n]$$

$$\|\mathbf{X}\|_* = \sum \sigma_i(\mathbf{X})$$

$$\|\mathbf{X}\|_{p,q} = \Big\| \|\mathbf{x}^1\|_p, \|\mathbf{x}^2\|_p, \ldots, \|\mathbf{x}^n\|_p \Big\|_q$$

$$\text{rg}\min_{\mathbf{W}} \sum_{k=1}^{K} \|\mathbf{w}^k\|_2^2 + \sum_{i=1}^{n} \ell_{\text{cs}}(y^i, \langle \mathbf{W}, \mathbf{x}^i \rangle)$$

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

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

Exer ise

More powerful regularizers?

$$\|\mathbf{W}\|_F^2, \|\mathbf{W}\|_*, \|\mathbf{W}\|_0, \|\mathbf{W}^\top\|_{2,0}$$

Multi-classification Loss Functions

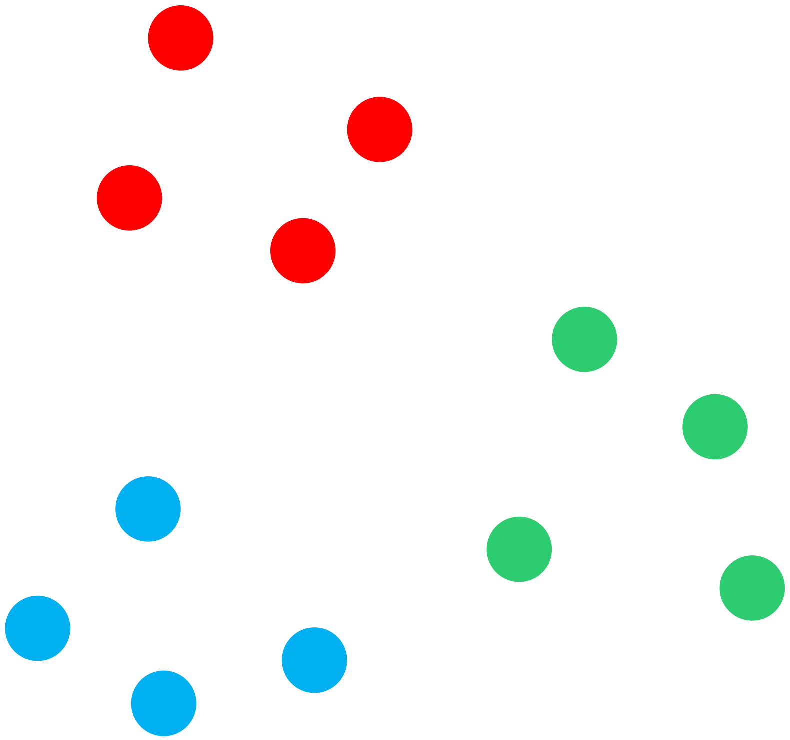
All-vs-All (AVA)

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

Multi-classification Loss Functions

All-vs-All (AVA)

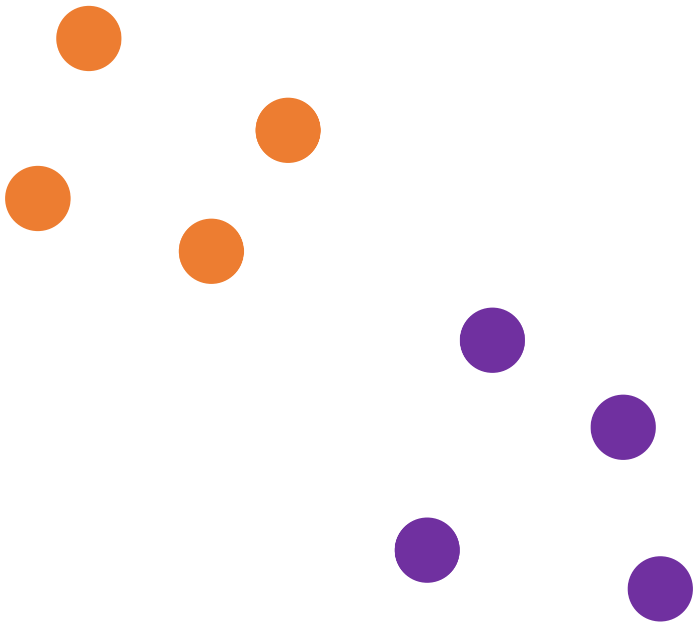
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

All-vs-All (AVA)

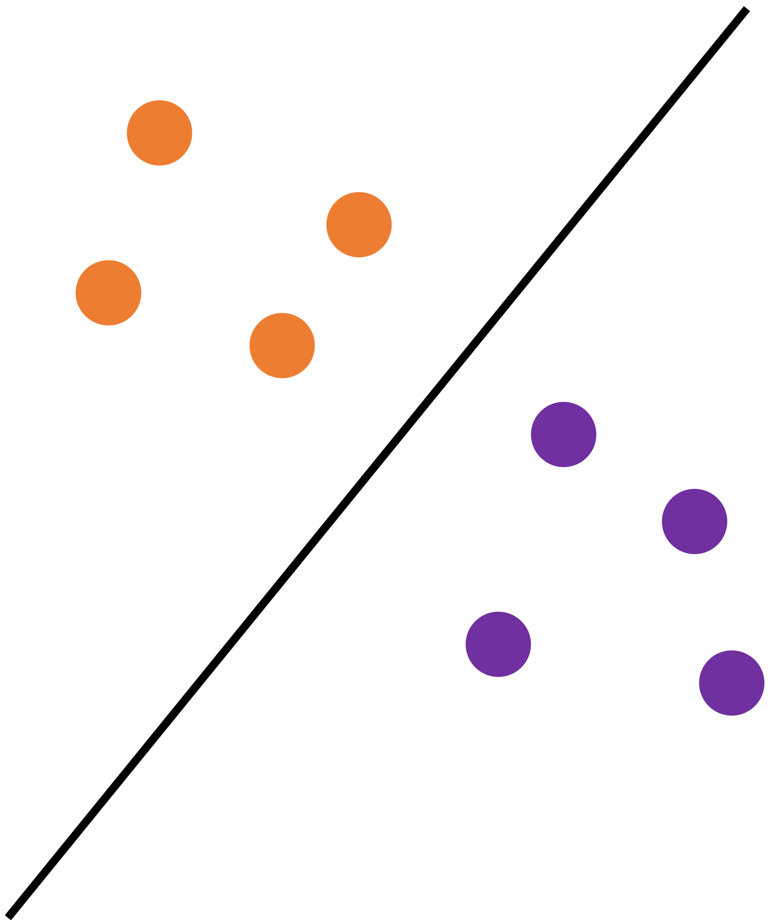
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

All-vs-All (AVA)

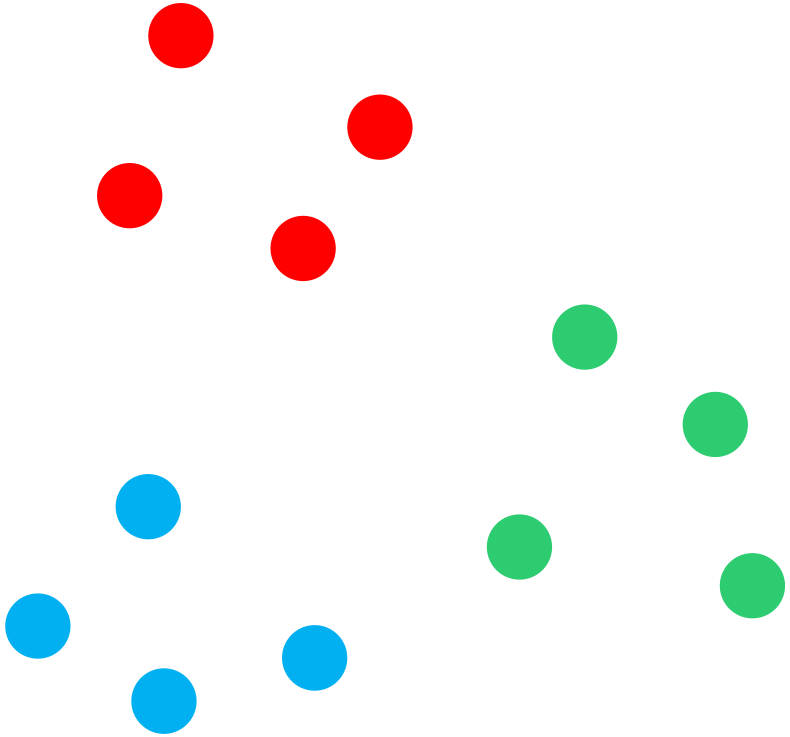
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

All-vs-All (AVA)

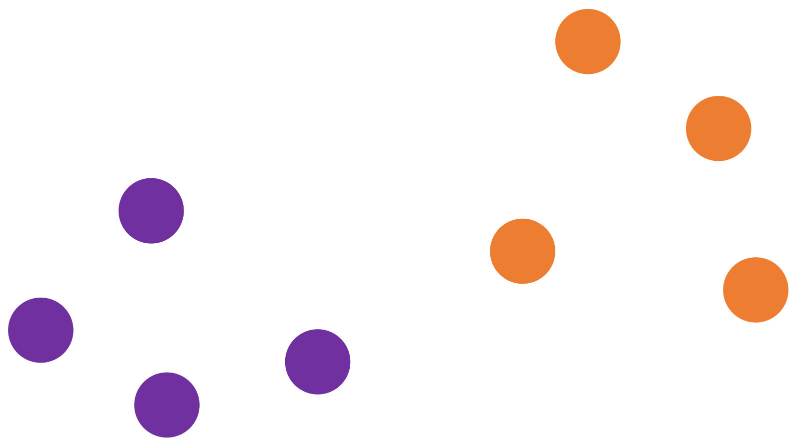
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

All-vs-All (AVA)

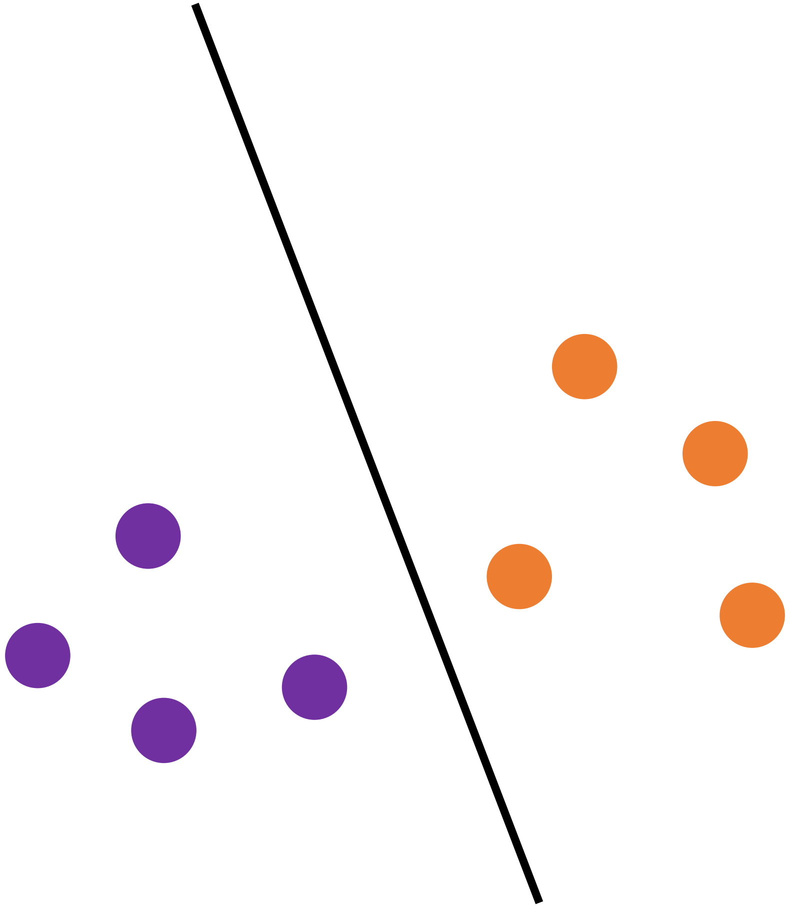
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

$\boxed{\text{All-vs-All (AVA)}}$

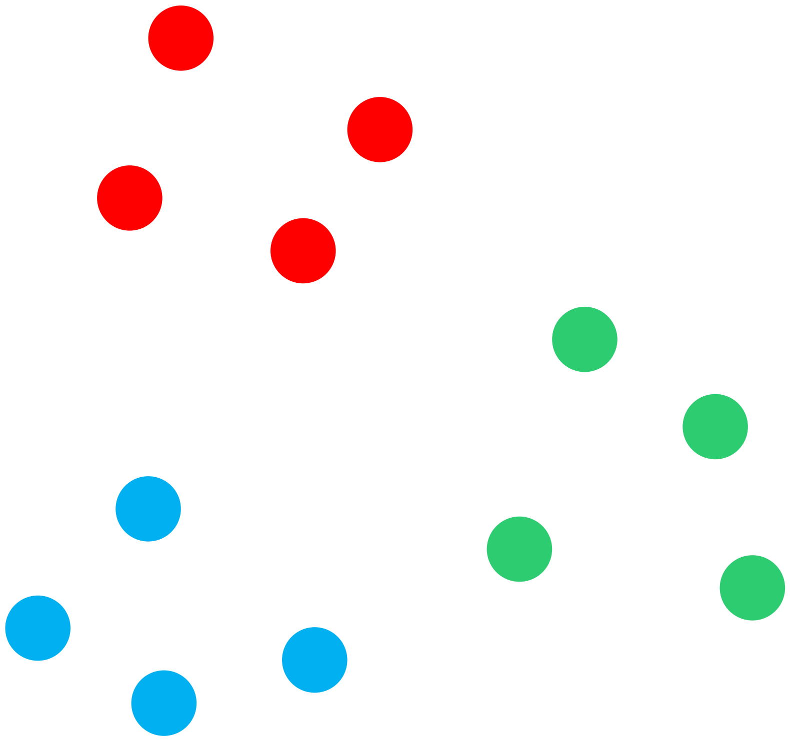
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

All-vs-All (AVA)

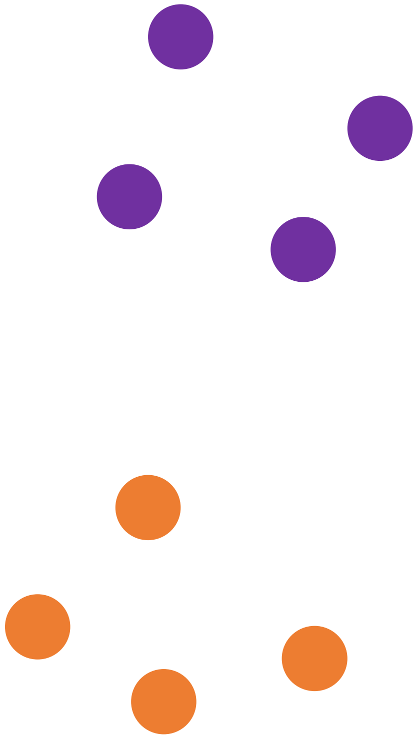
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

All-vs-All (AVA)

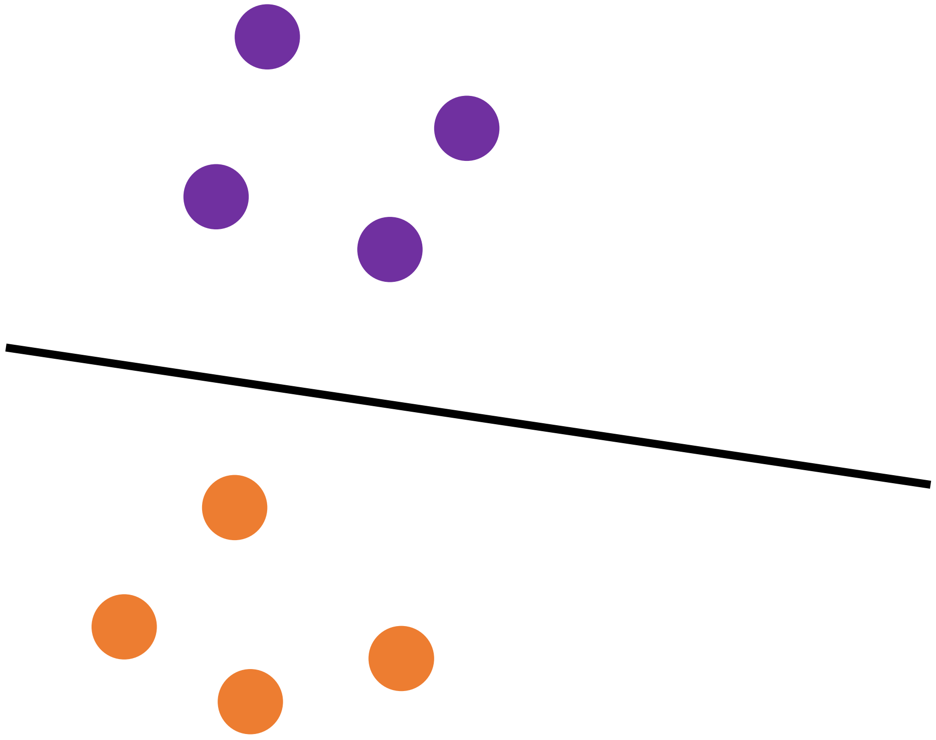
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

All-vs-All (AVA)

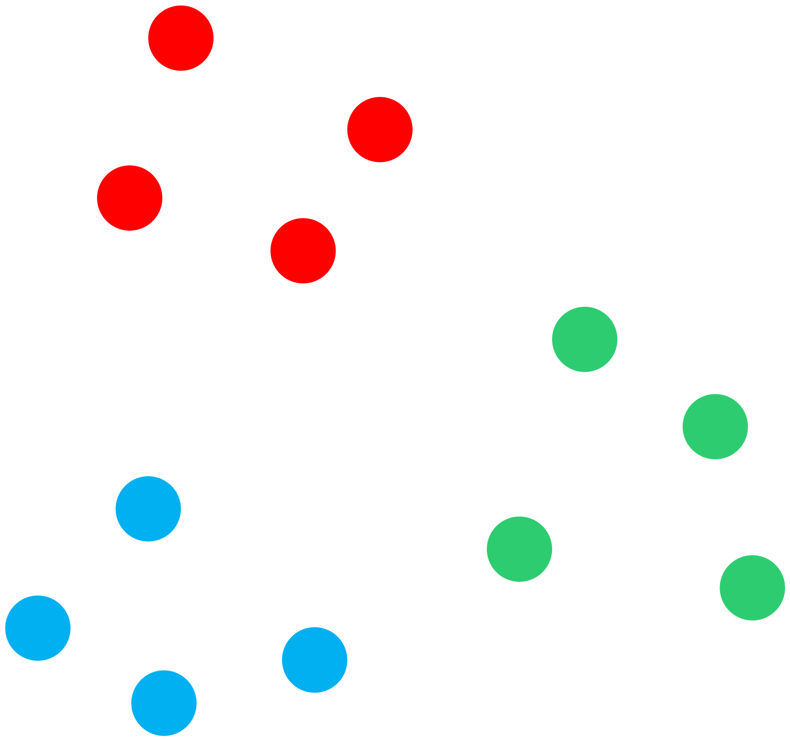
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

All-vs-All (AVA)

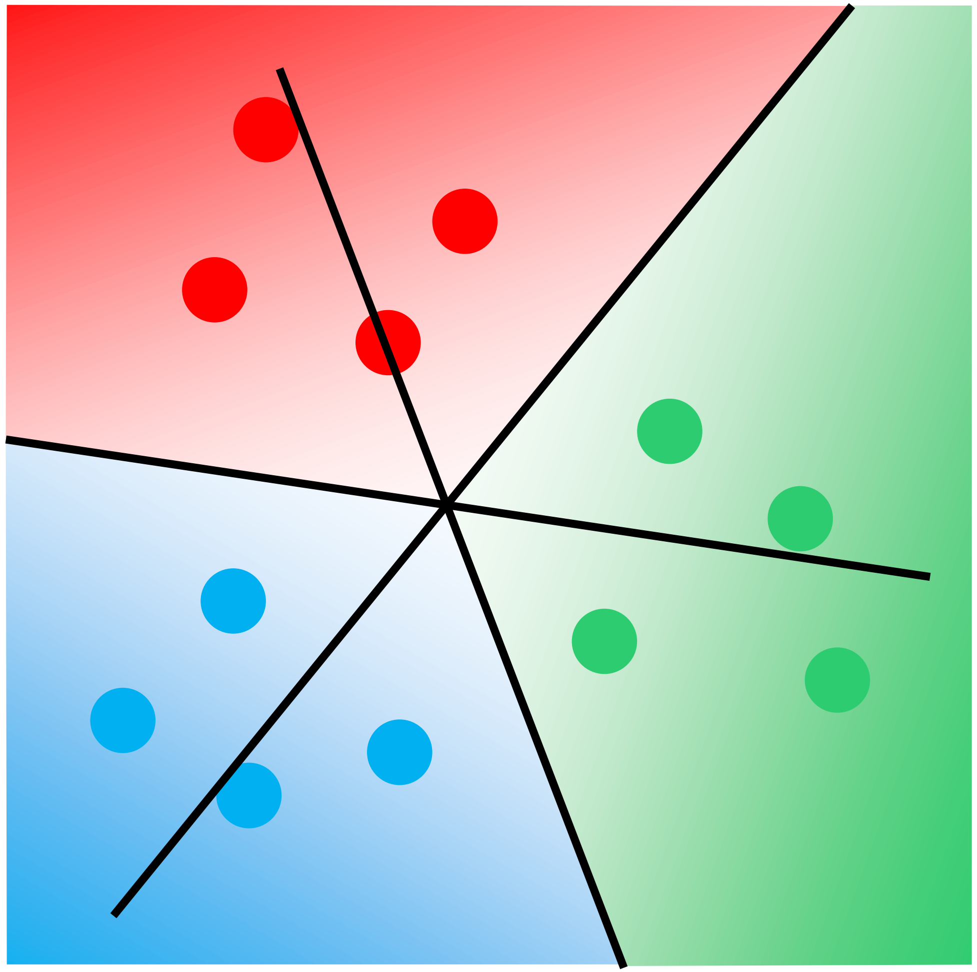
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

All-vs-All (AVA)

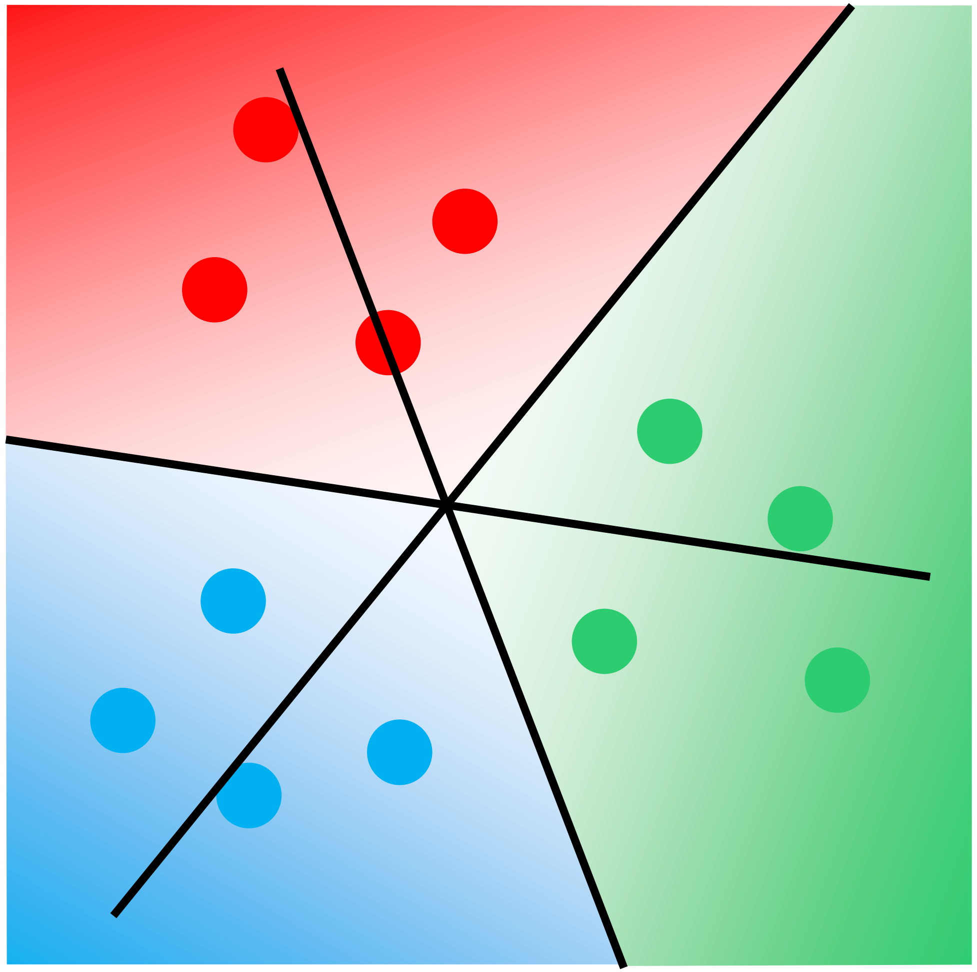
$$\hat{y}^i = \arg\max_{j \in [K]} \sum_{k \neq j} \langle \mathbf{w}^{j,k}, \mathbf{x}^i \rangle$$



Multi-classification Loss Functions

All-vs-All (AVA)

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

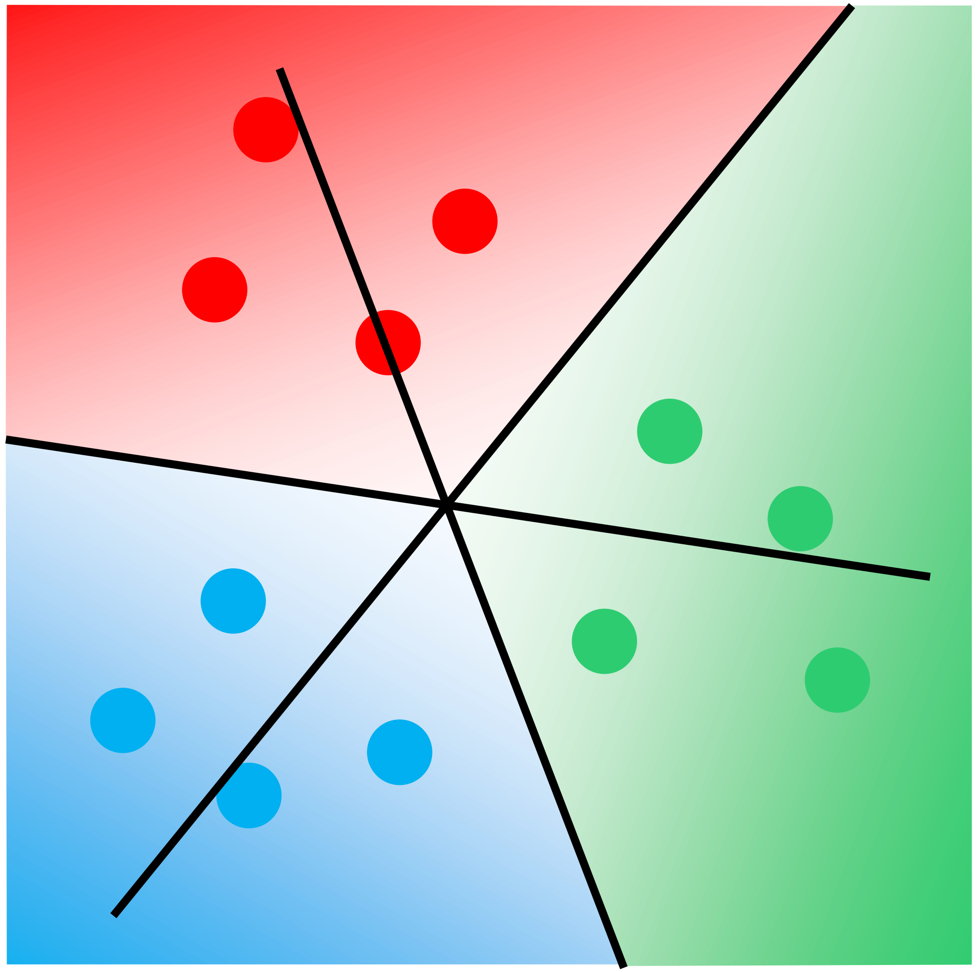


Expensive!

Multi-classification Loss Functions

All-vs-All (AVA)

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



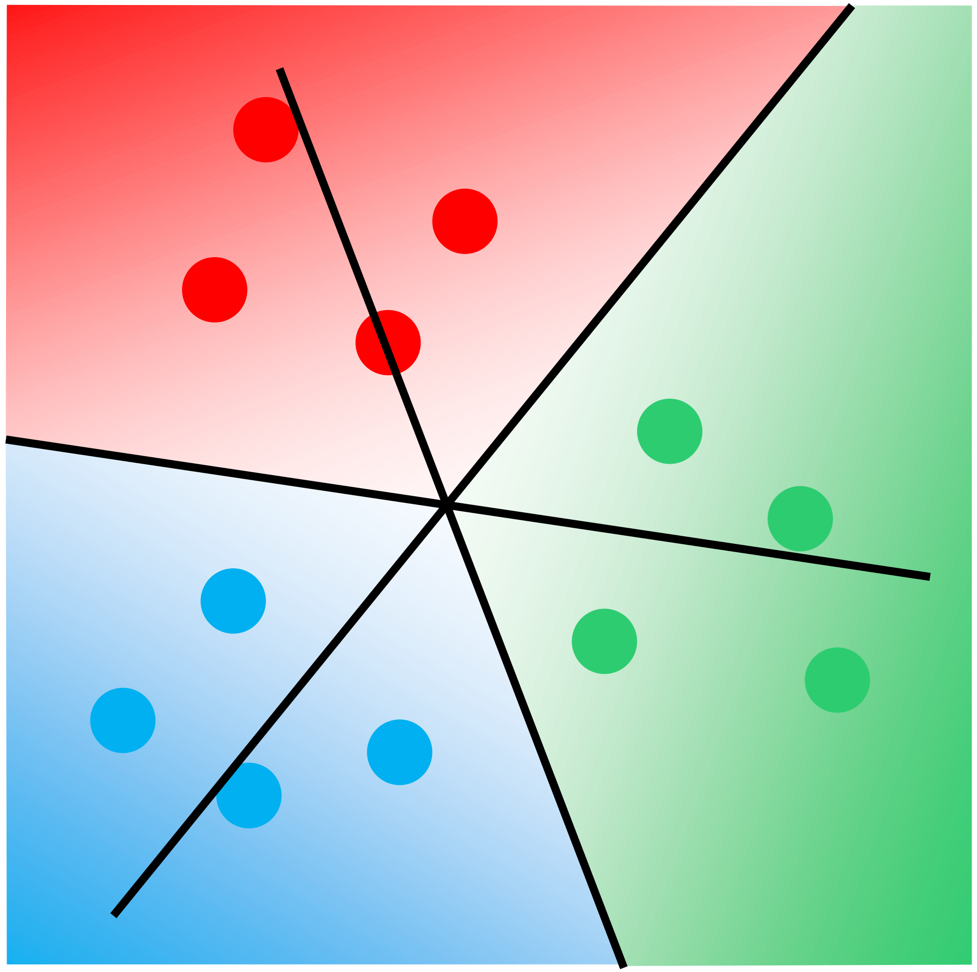
Expensive!

Other techniques: use DT to eliminate classes, error correcting codes

Multi-classification Loss Functions

All-vs-All (AVA)

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



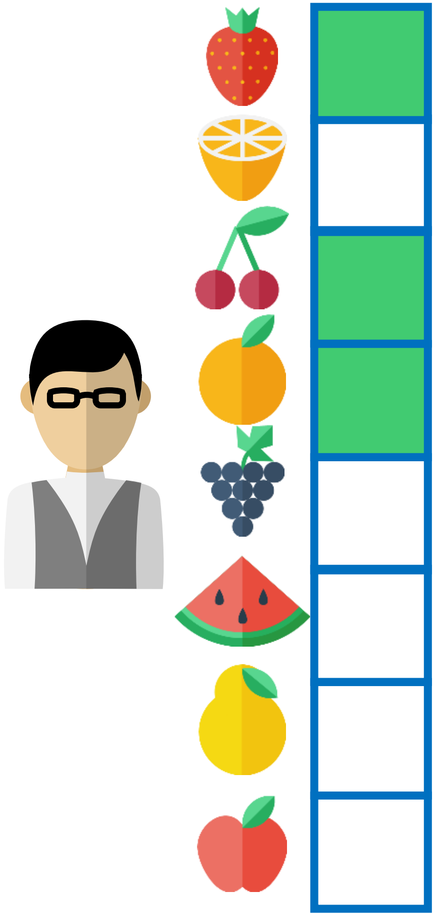
Expensive!

Other techniques: use DT to eliminate classes, error correcting codes

Exercise: Develop training techniques for AVA classifiers

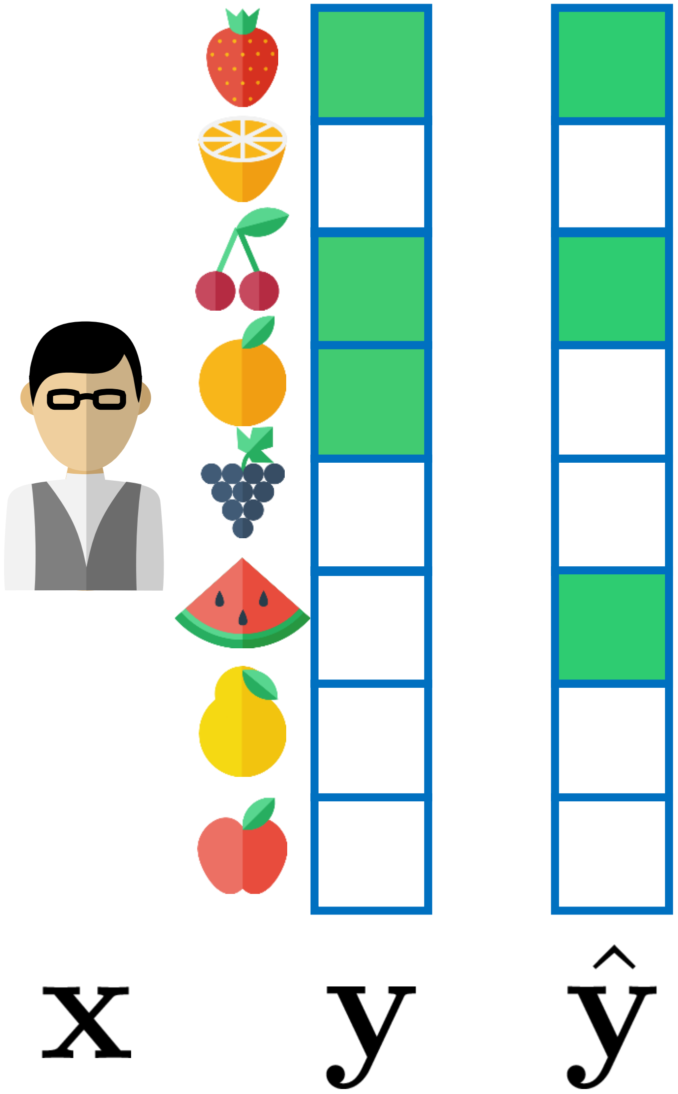
Multi-label Classification Loss Functions

Multi-label Classification Loss Functions

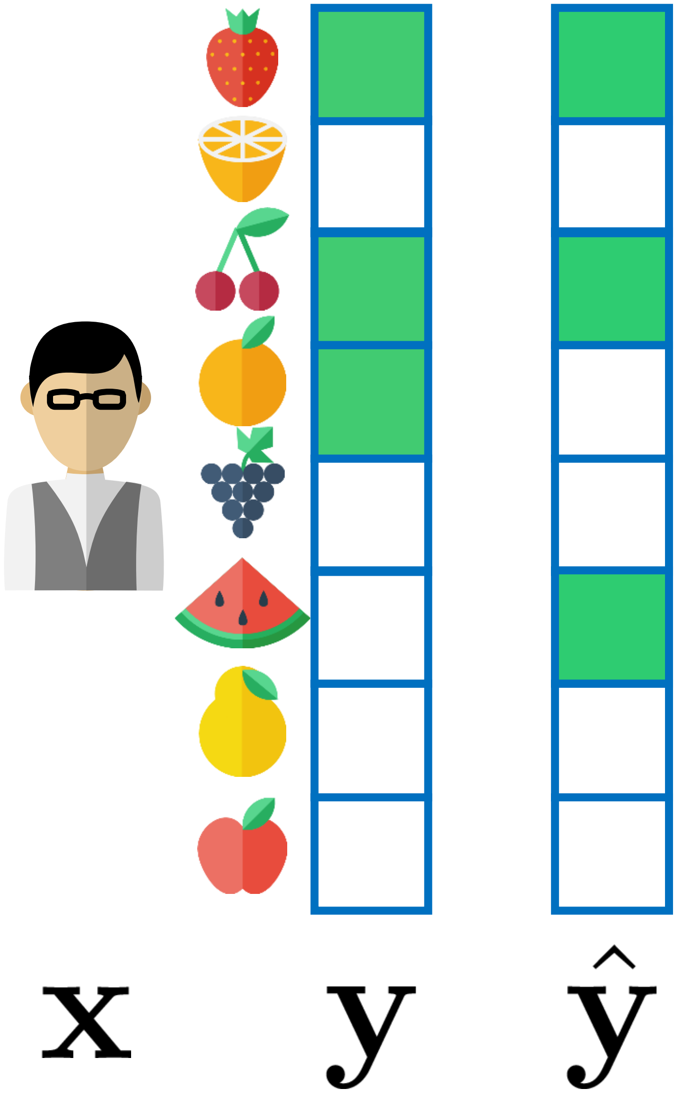


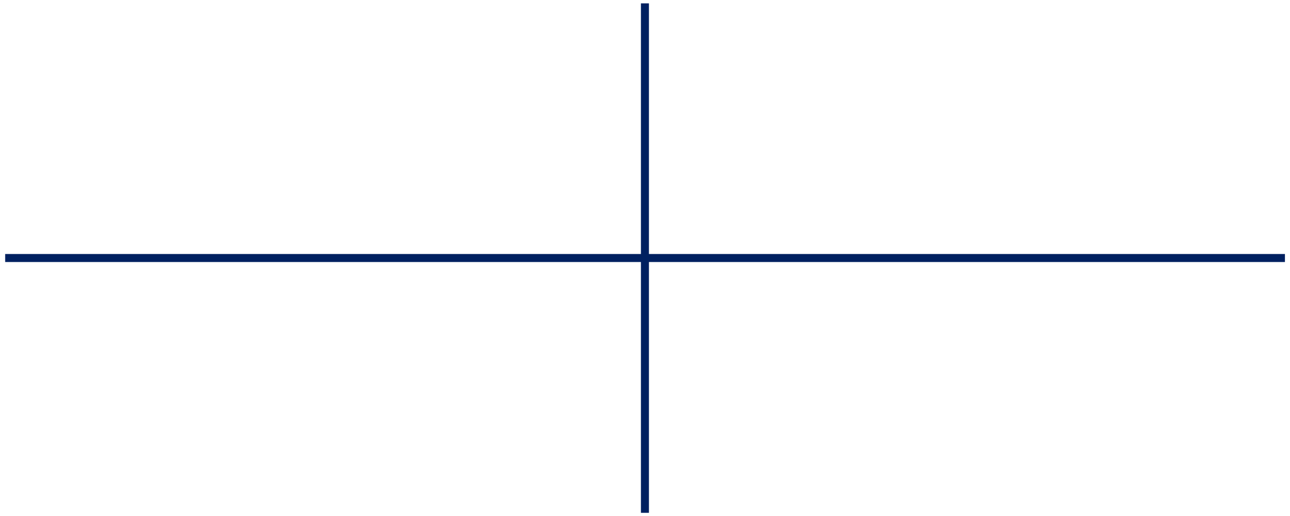
$\mathbf{x}$ $\mathbf{y}$

Multi-label Classification Loss Functions

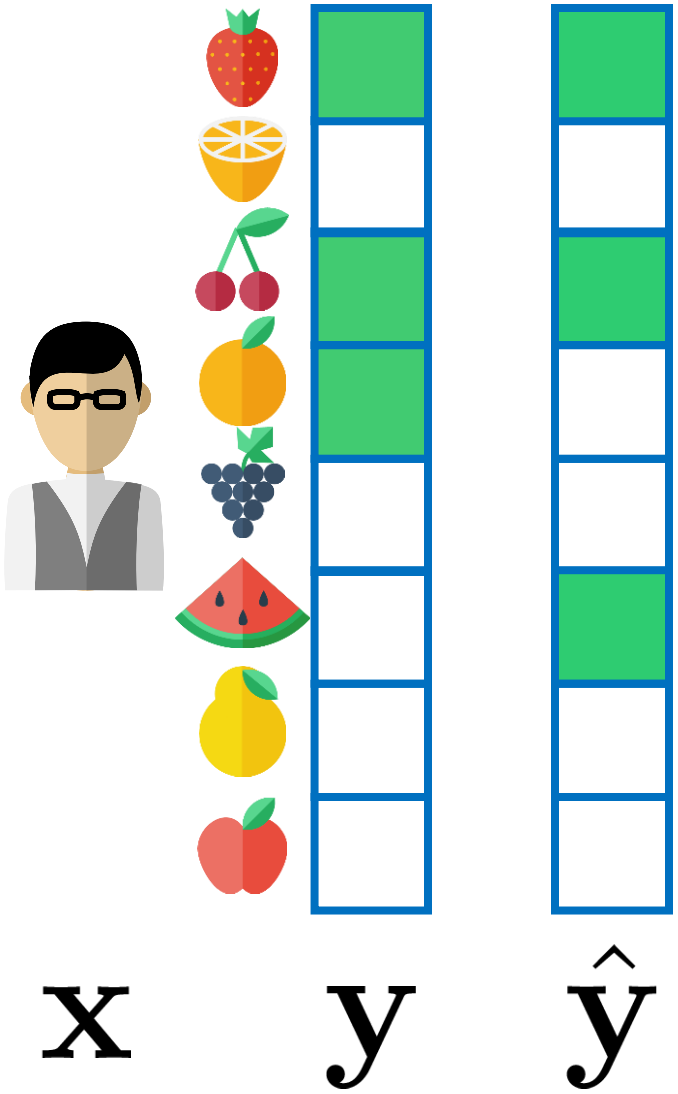


Multi-label Classification Loss Functions



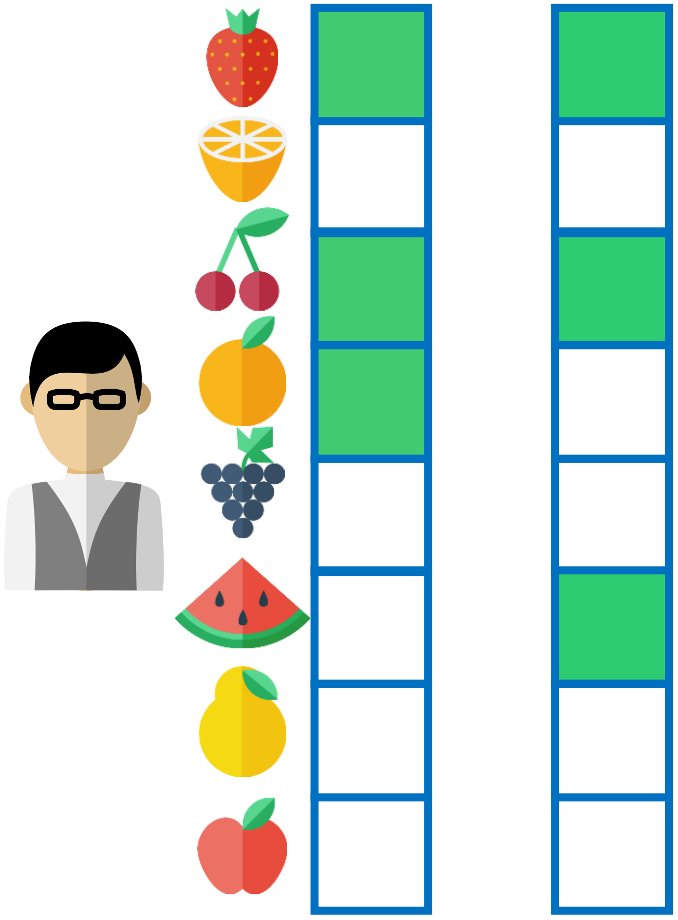


Multi-label Classification Loss Functions



$$\left|\left\{ i : \begin{array}{l} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right|$$

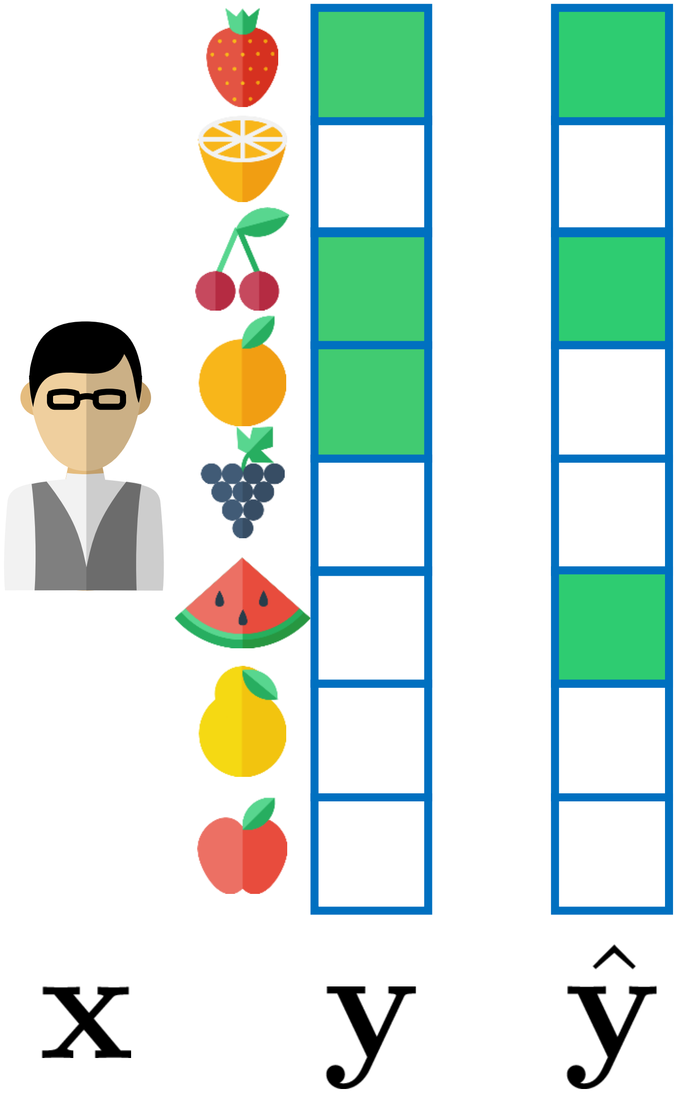
Multi-label Classification Loss Functions



$\mathbf{x}$ $\quad$ $\mathbf{y}$ $\quad$ $\hat{\mathbf{y}}$

$\left\lvert\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right\rvert$	$\left\lvert\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = -1 \end{array} \right\}\right\rvert$

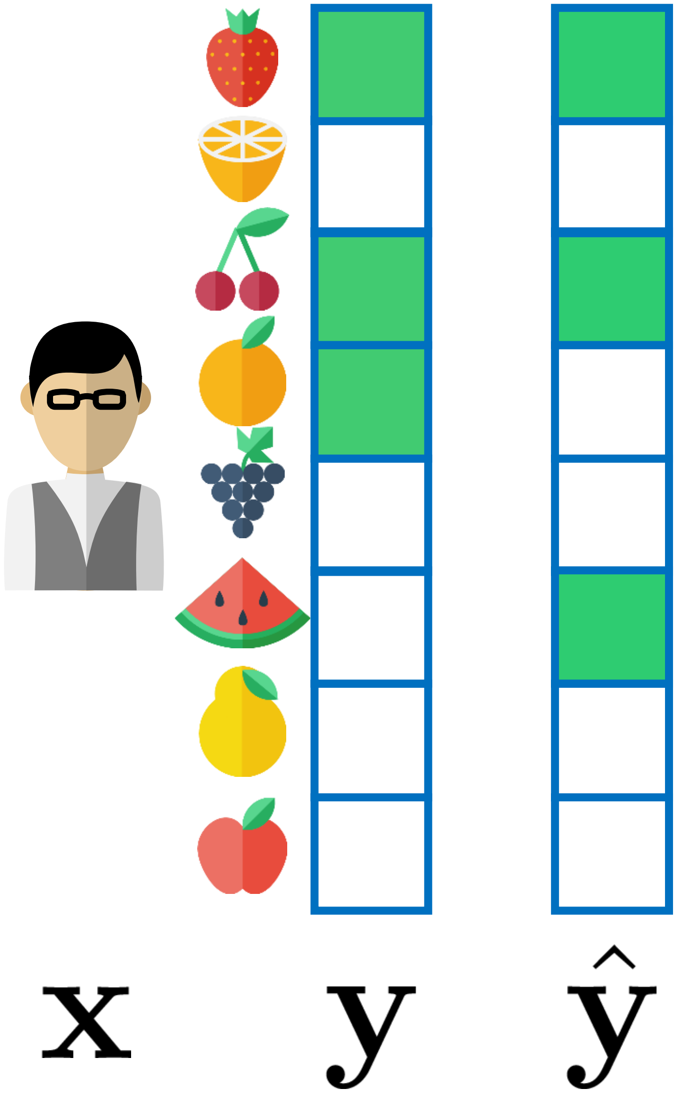
Multi-label Classification Loss Functions



$$\left|\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right| \qquad \left|\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = -1 \end{array} \right\}\right|$$

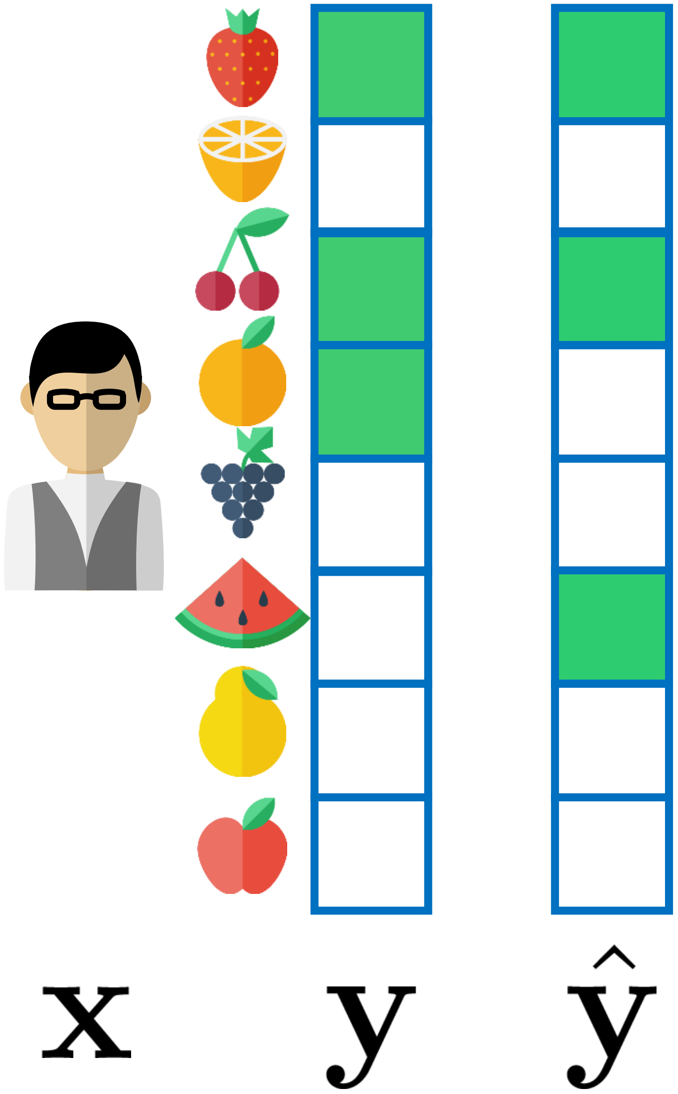
$$\left|\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right|$$

Multi-label Classification Loss Functions



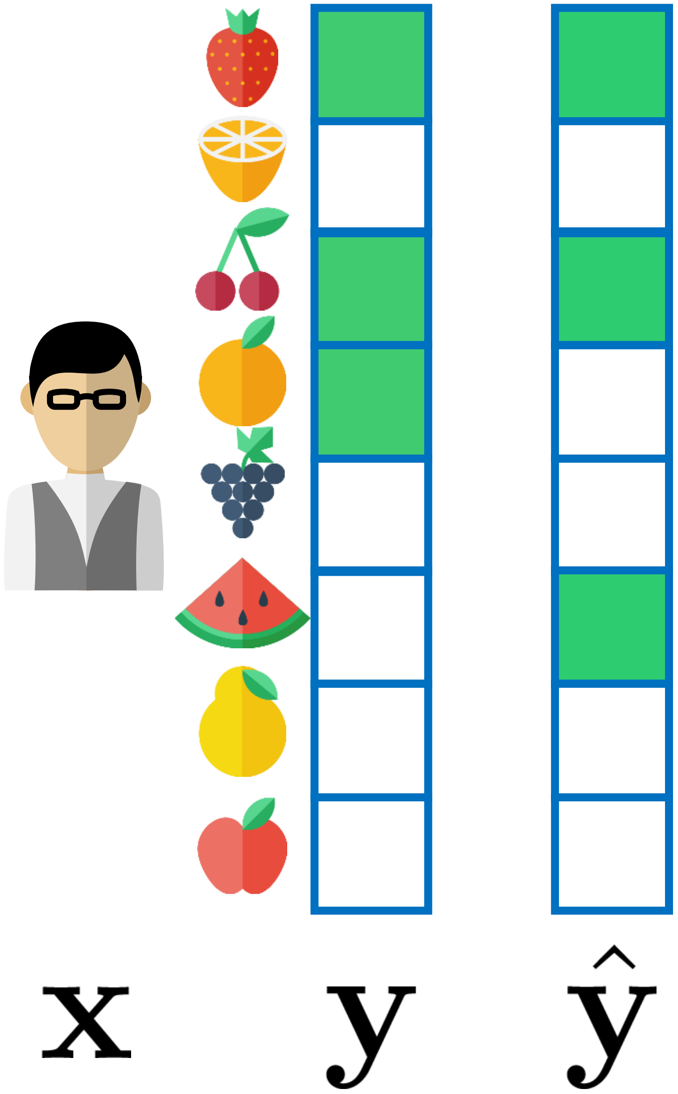
$$
\begin{array}{c|c}
\left|\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right| & \left|\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = -1 \end{array} \right\}\right| \\
\hline
\left|\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i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

Multi-label Classification Loss Functions



a: $\left\lvert\left\{ i : \begin{matrix} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 1 \end{matrix} \right\}\right\rvert$	b: $\left\lvert\left\{ i : \begin{matrix} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = -1 \end{matrix} \right\}\right\rvert$
c: $\left\lvert\left\{ i : \begin{matrix} \hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = 1 \end{matrix} \right\}\right\rvert$	d: $\left\lvert\left\{ i : \begin{matrix} \hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = -1 \end{matrix} \right\}\right\rvert$

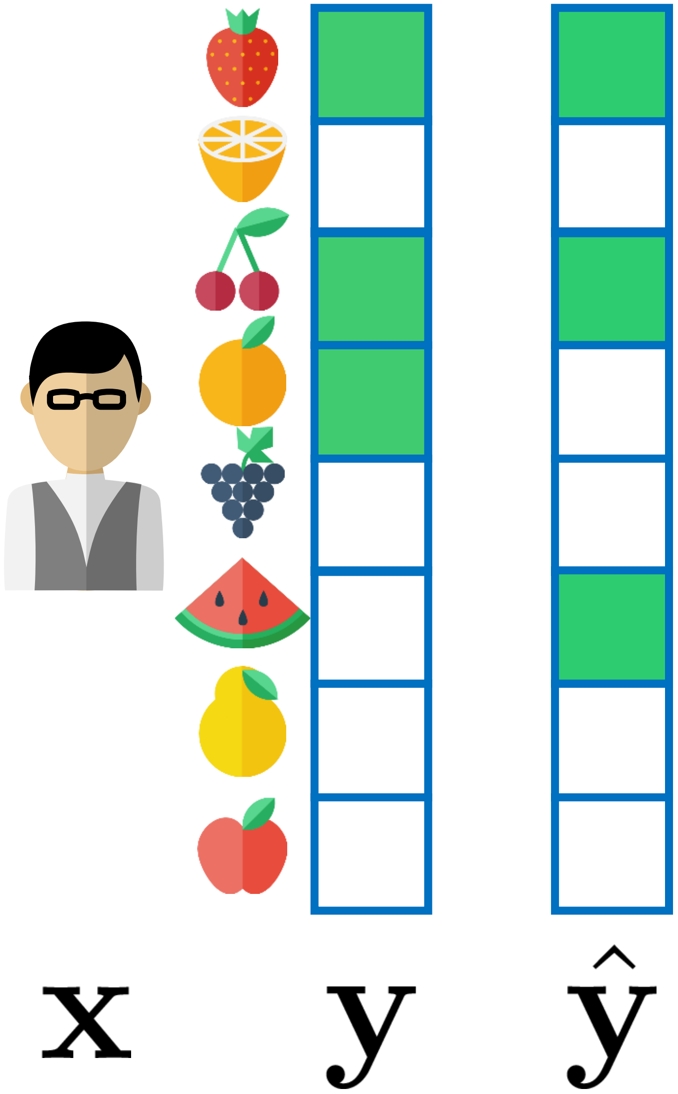
Multi-label Classification Loss Functions



$\left\lvert\left\{i : \begin{array}{c}\hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 1\end{array}\right\}\right\rvert$ a	$\left\lvert\left\{i : \begin{array}{c}\hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = -1\end{array}\right\}\right\rvert$ b
$\left\lvert\left\{i : \begin{array}{c}\hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = 1\end{array}\right\}\right\rvert$ c	$\left\lvert\left\{i : \begin{array}{c}\hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = -1\end{array}\right\}\right\rvert$ d

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

Multi-label Classification Loss Functions

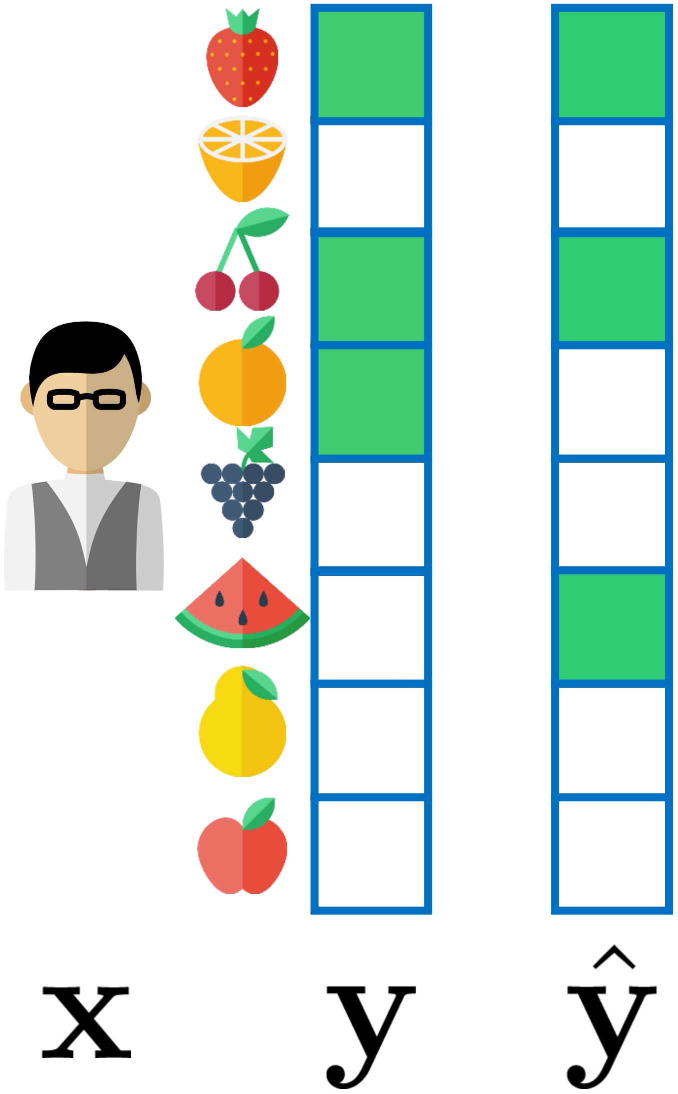


a: $\left\|\left\{i : \begin{matrix}\hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 1\end{matrix}\right\}\right\|$	b: $\left\|\left\{i : \begin{matrix}\hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = -1\end{matrix}\right\}\right\|$
c: $\left\|\left\{i : \begin{matrix}\hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = 1\end{matrix}\right\}\right\|$	d: $\left\|\left\{i : \begin{matrix}\hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = -1\end{matrix}\right\}\right\|$

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

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

Multi-label Classification Loss Functions

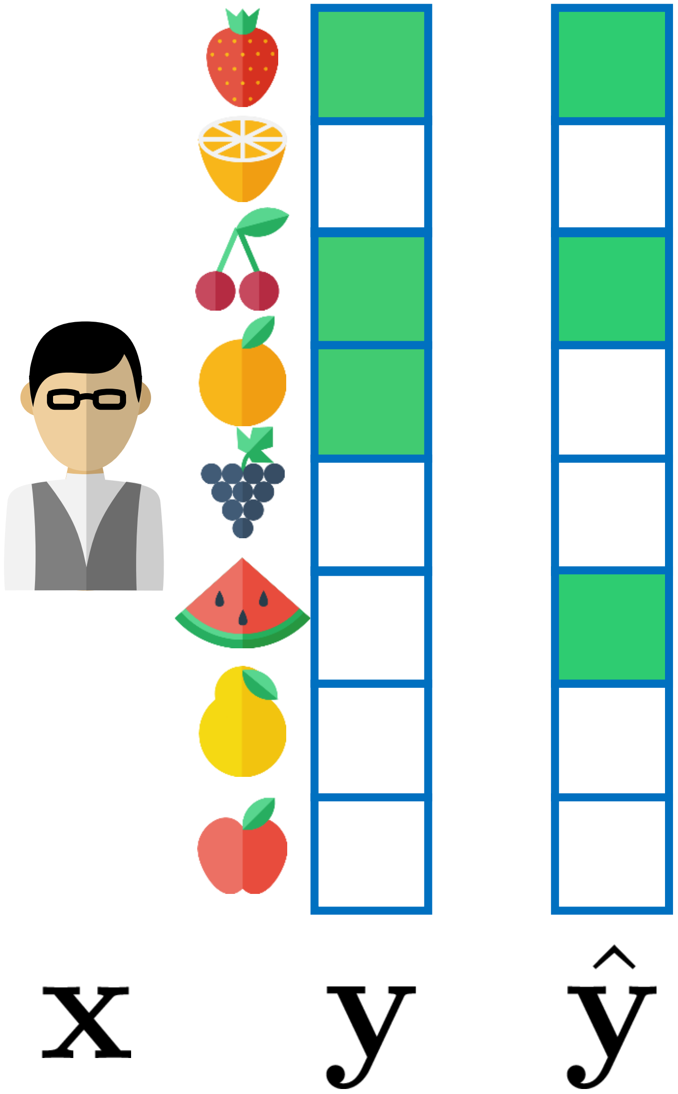


$\left\lvert\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right\rvert$ a	$\left\lvert\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = -1 \end{array} \right\}\right\rvert$ b
$\left\lvert\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right\rvert$ c	$\left\lvert\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = -1 \end{array} \right\}\right\rvert$ d

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

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

Multi-label Classification Loss Functions



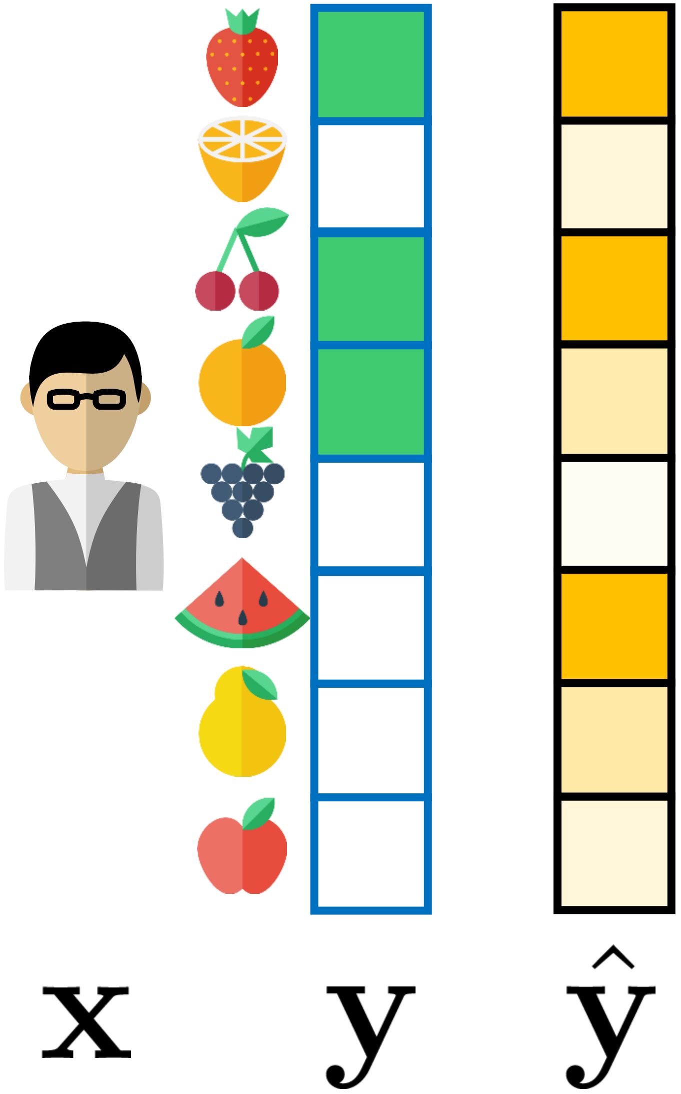
a: $\left\lvert\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right\rvert$	b: $\left\lvert\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = -1 \end{array} \right\}\right\rvert$
c: $\left\lvert\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = 1 \end{array} \right\}\right\rvert$	d: $\left\lvert\left\{ i : \begin{array}{c} \hat{\mathbf{y}}_i = -1 \\ \mathbf{y}_i = -1 \end{array} \right\}\right\rvert$

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

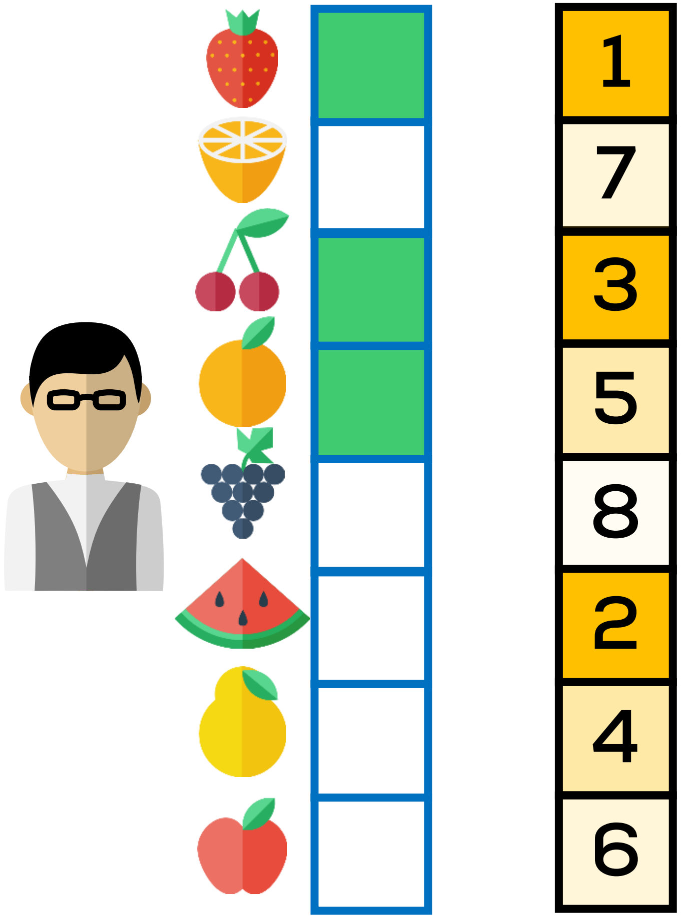
$$r_{\text{Precision}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{a}{a+b} \qquad r_{\text{Recall}}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{a}{a+c}$$

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

Multi-label Classification Loss Functions

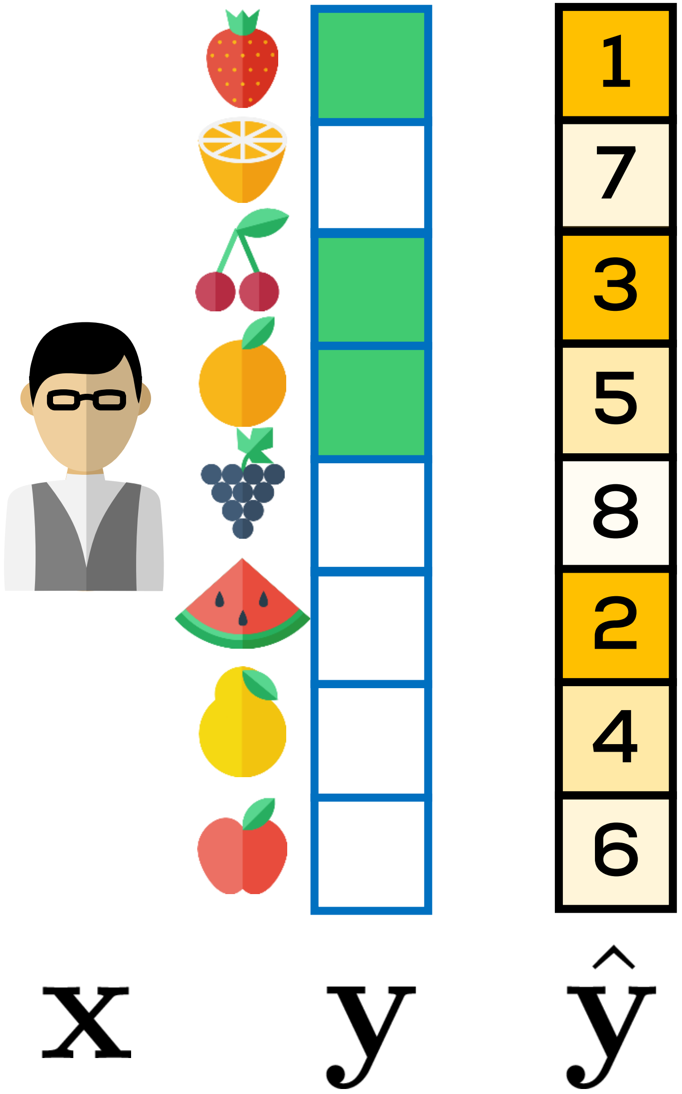


Multi-label Classification Loss Functions



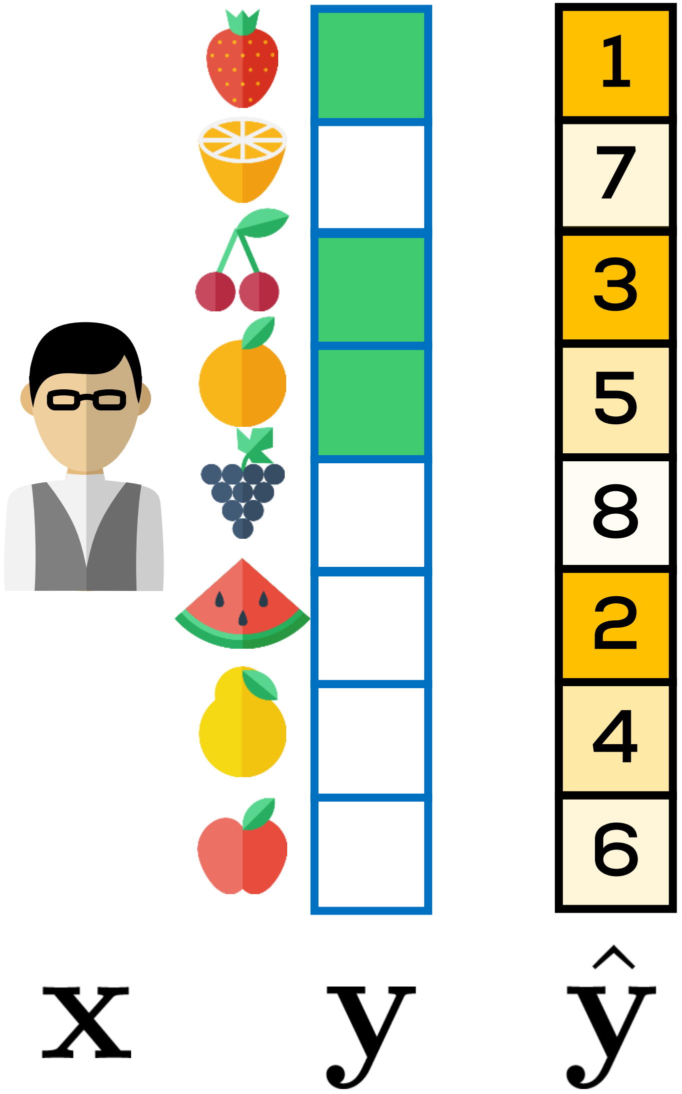
$$\mathbf{x} \qquad \mathbf{y} \qquad \hat{\mathbf{y}}$$

Multi-label Classification Loss Functions



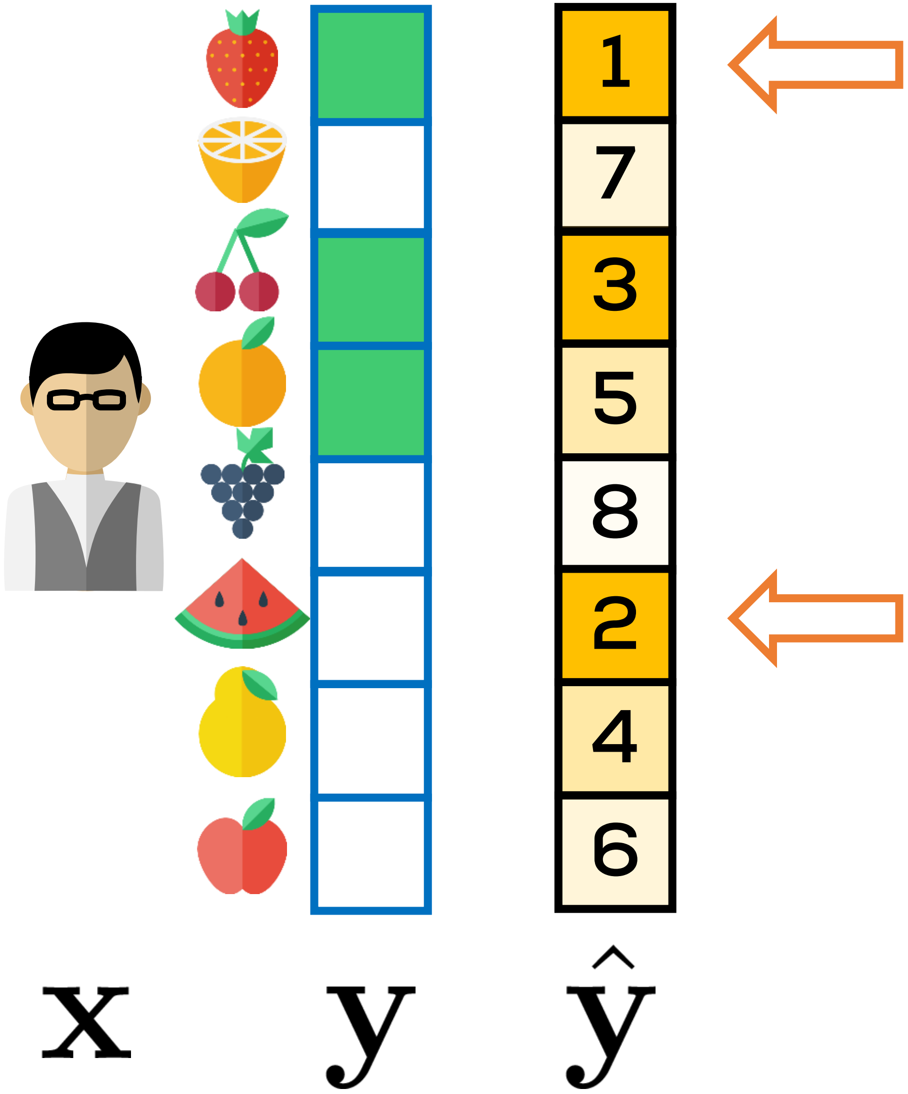
Precision@k

Multi-label Classification Loss Functions



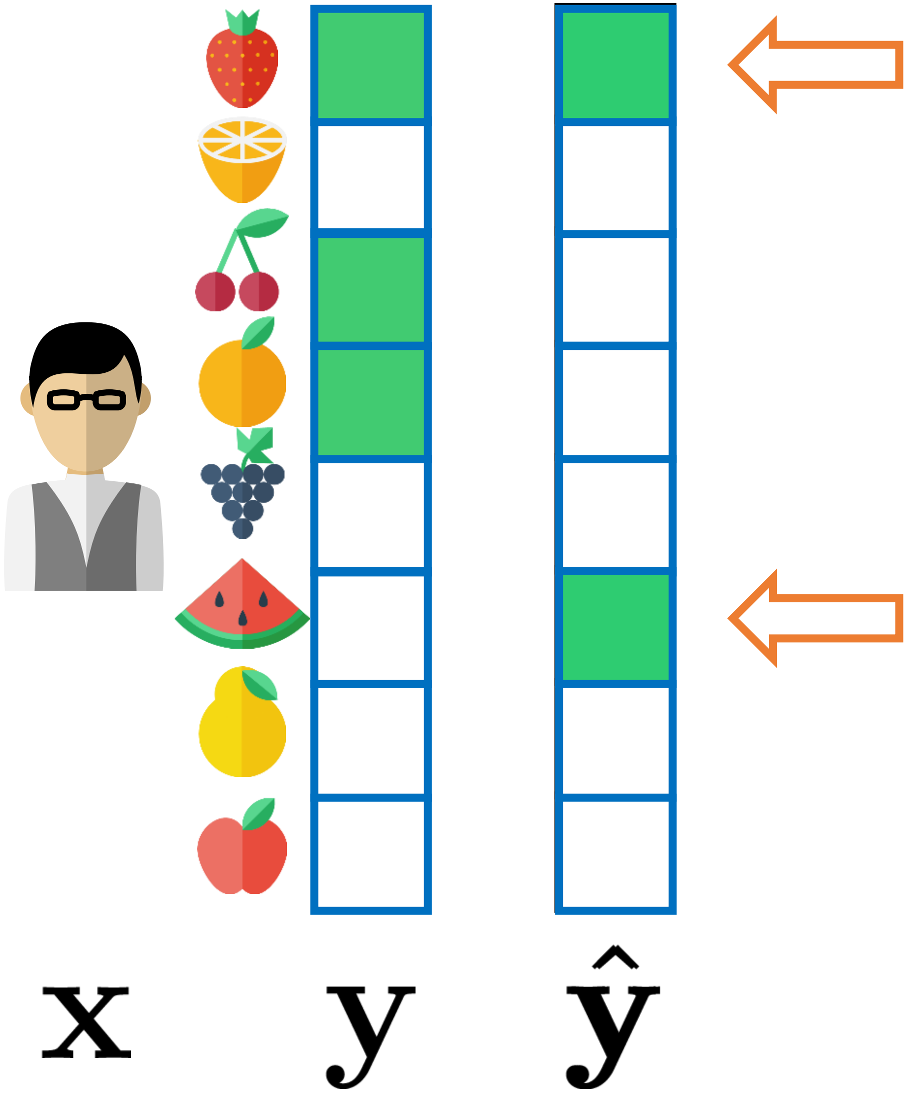
Precision@2

Multi-label Classification Loss Functions



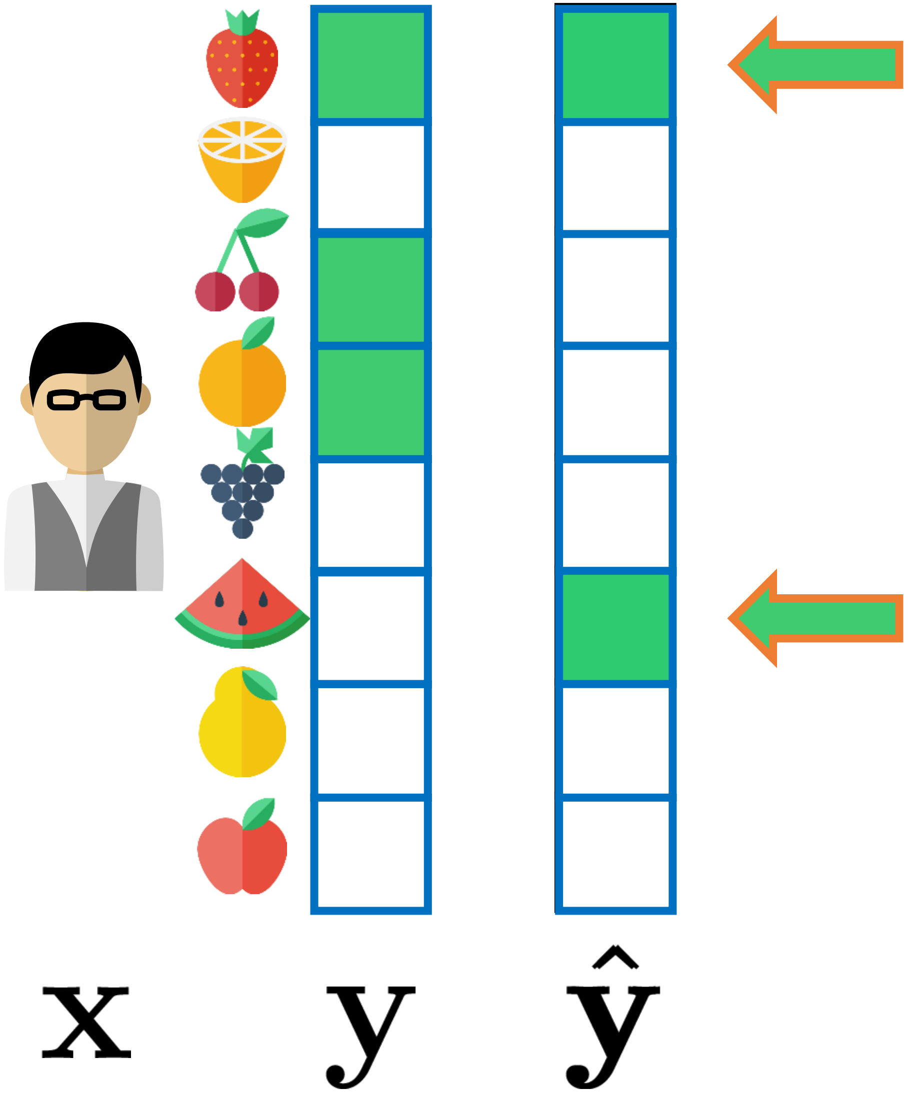
Precision@2

Multi-label Classification Loss Functions



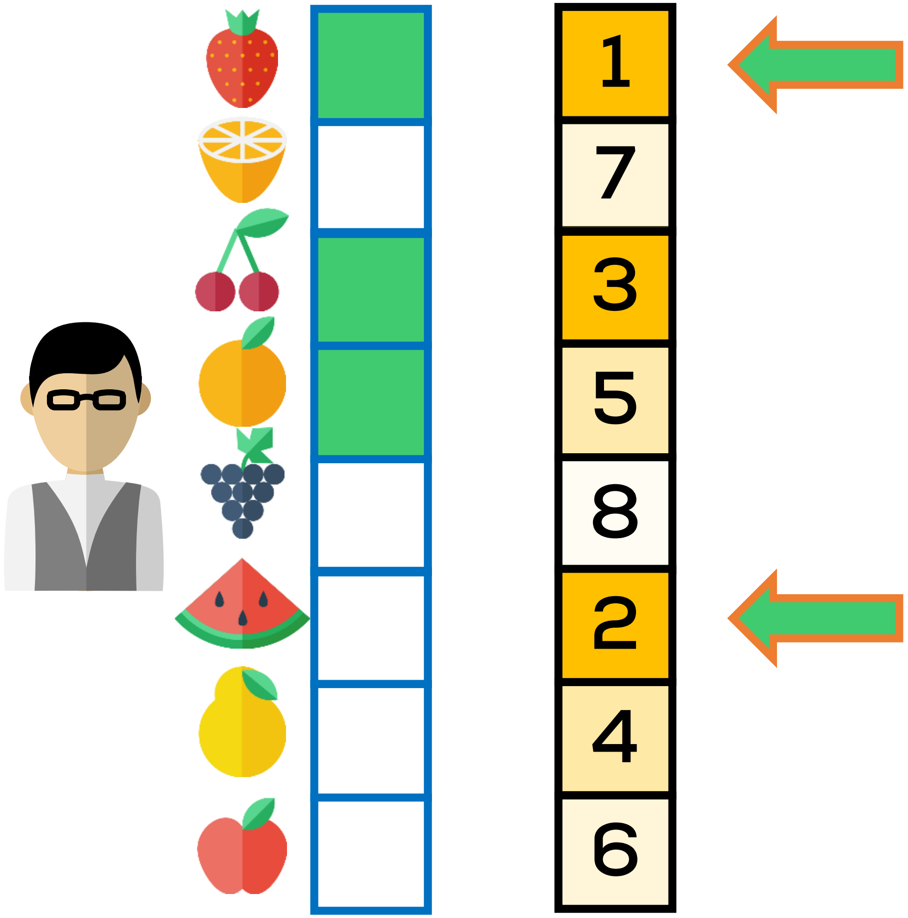
Precision@2

Multi-label Classification Loss Functions



Precision@2

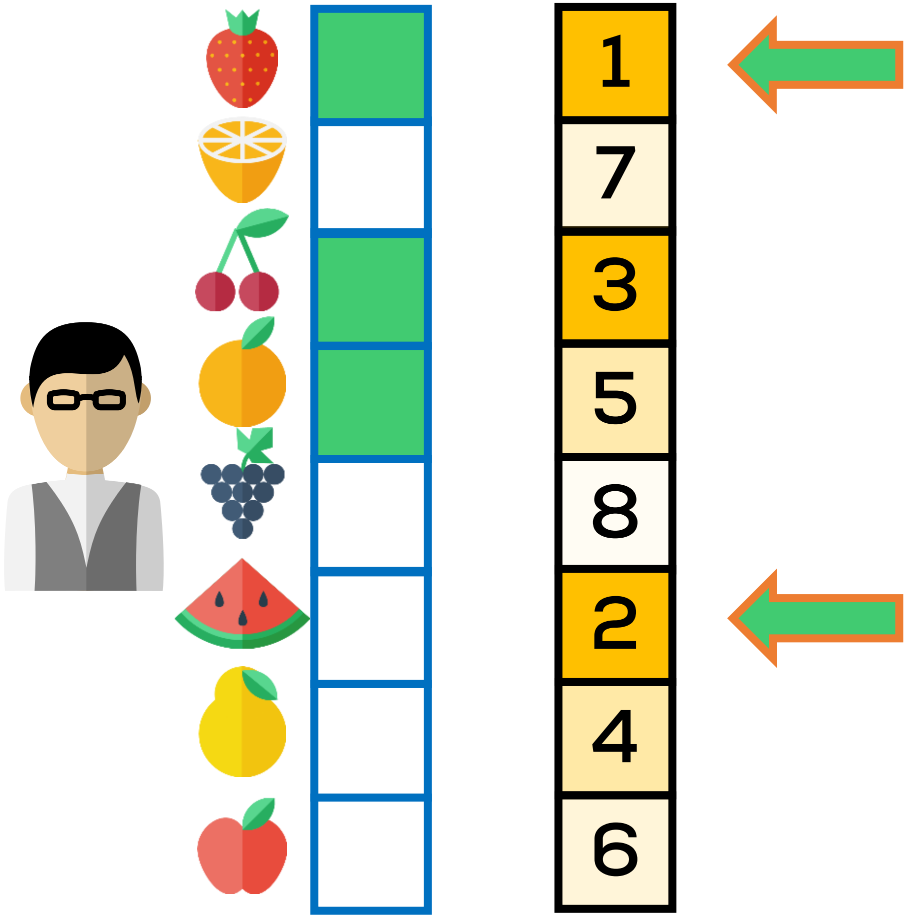
Multi-label Classification Loss Functions



$\mathbf{x}$ $\mathbf{y}$ $\hat{\mathbf{y}}$

Precision@2

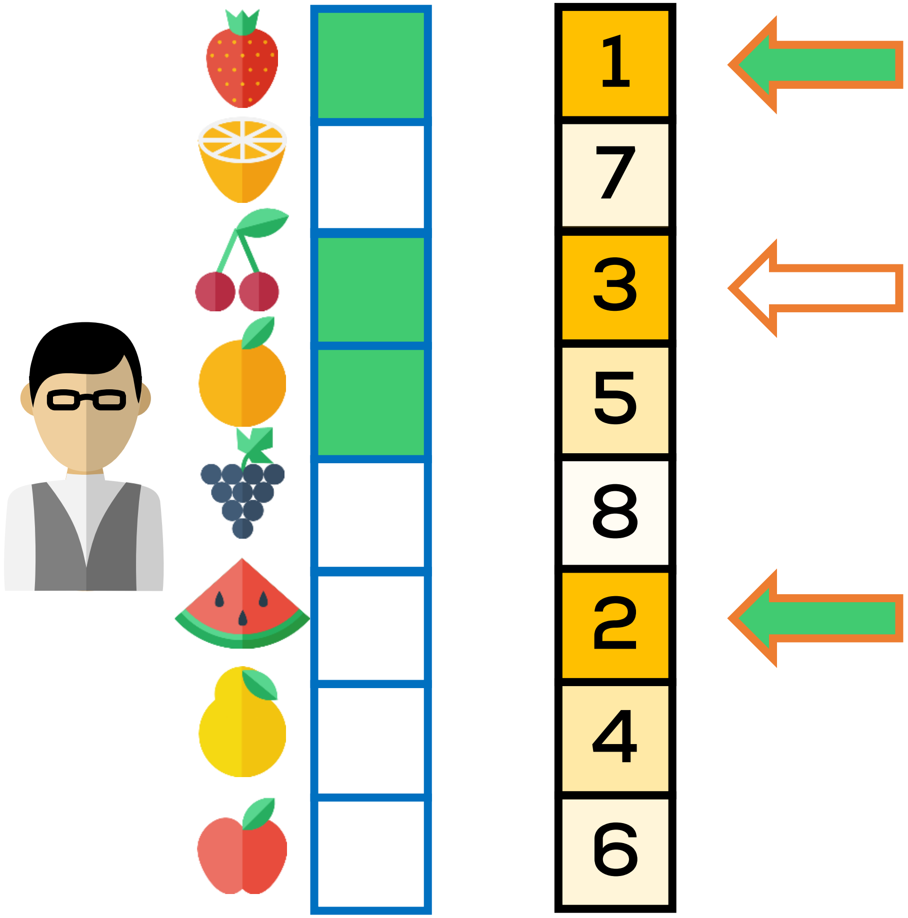
Multi-label Classification Loss Functions



$\mathbf{x}$ $\mathbf{y}$ $\hat{\mathbf{y}}$

Precision@3

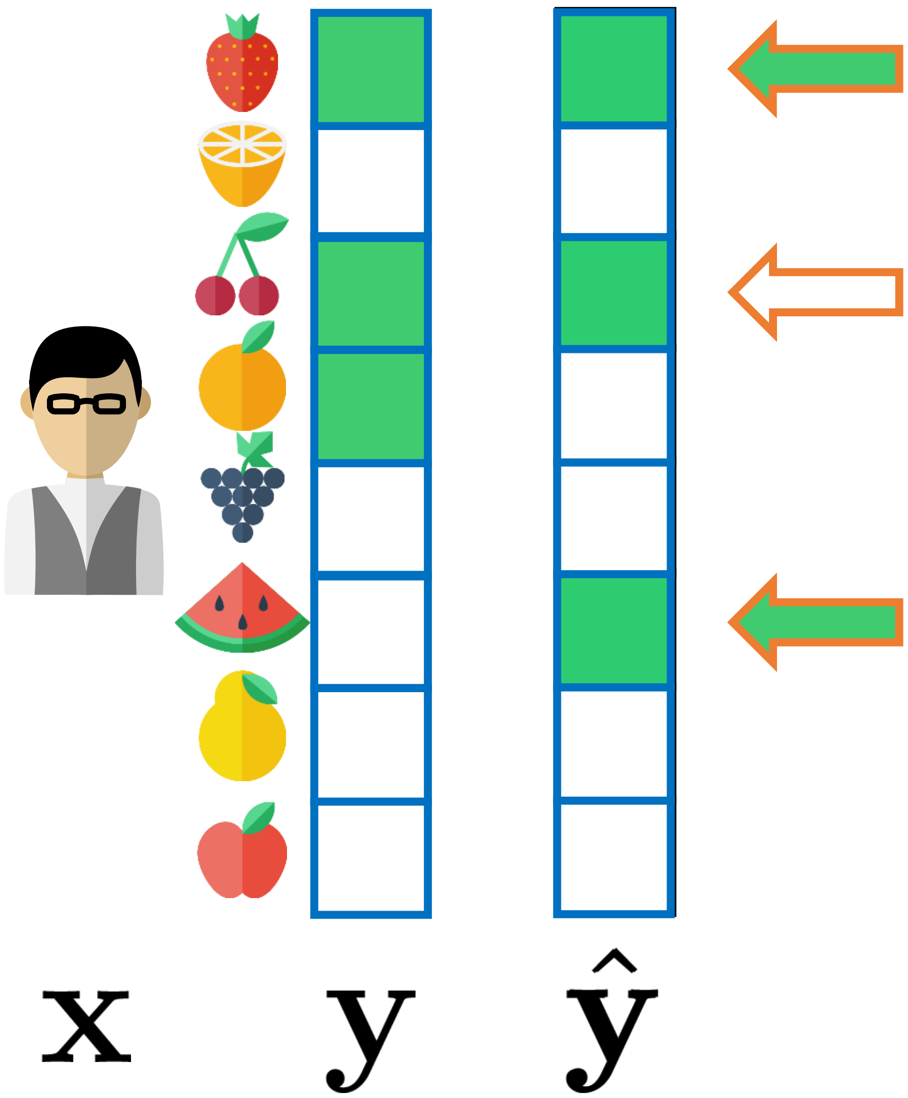
Multi-label Classification Loss Functions



$$\mathbf{x} \qquad \mathbf{y} \qquad \hat{\mathbf{y}}$$

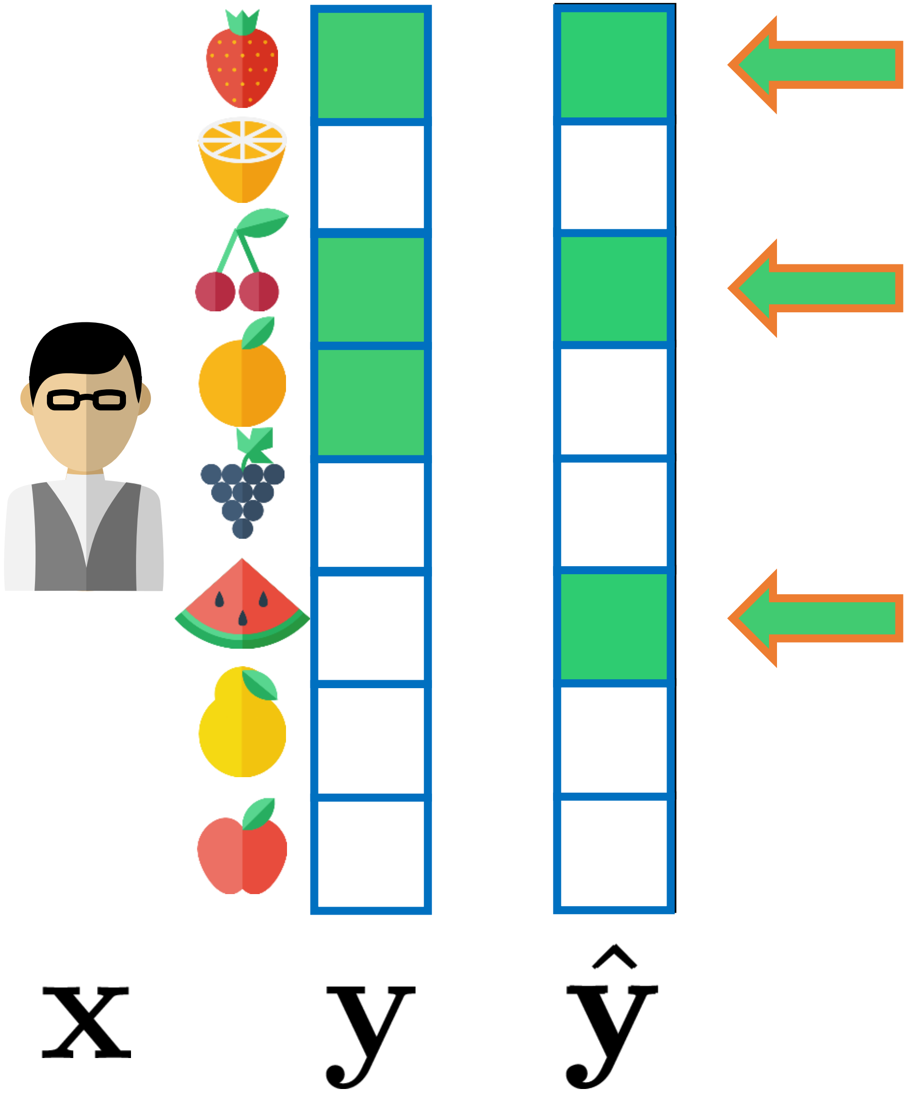
Precision@3

Multi-label Classification Loss Functions



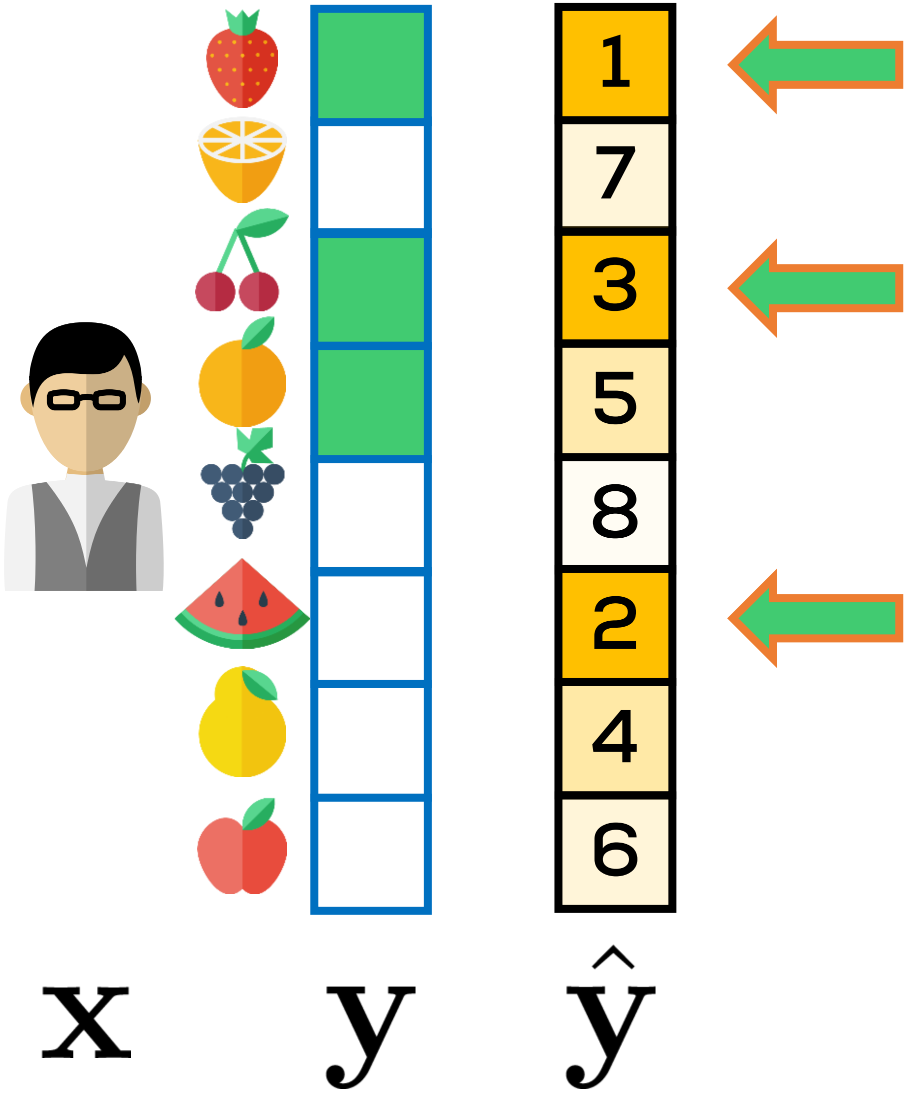
Precision@3

Multi-label Classification Loss Functions



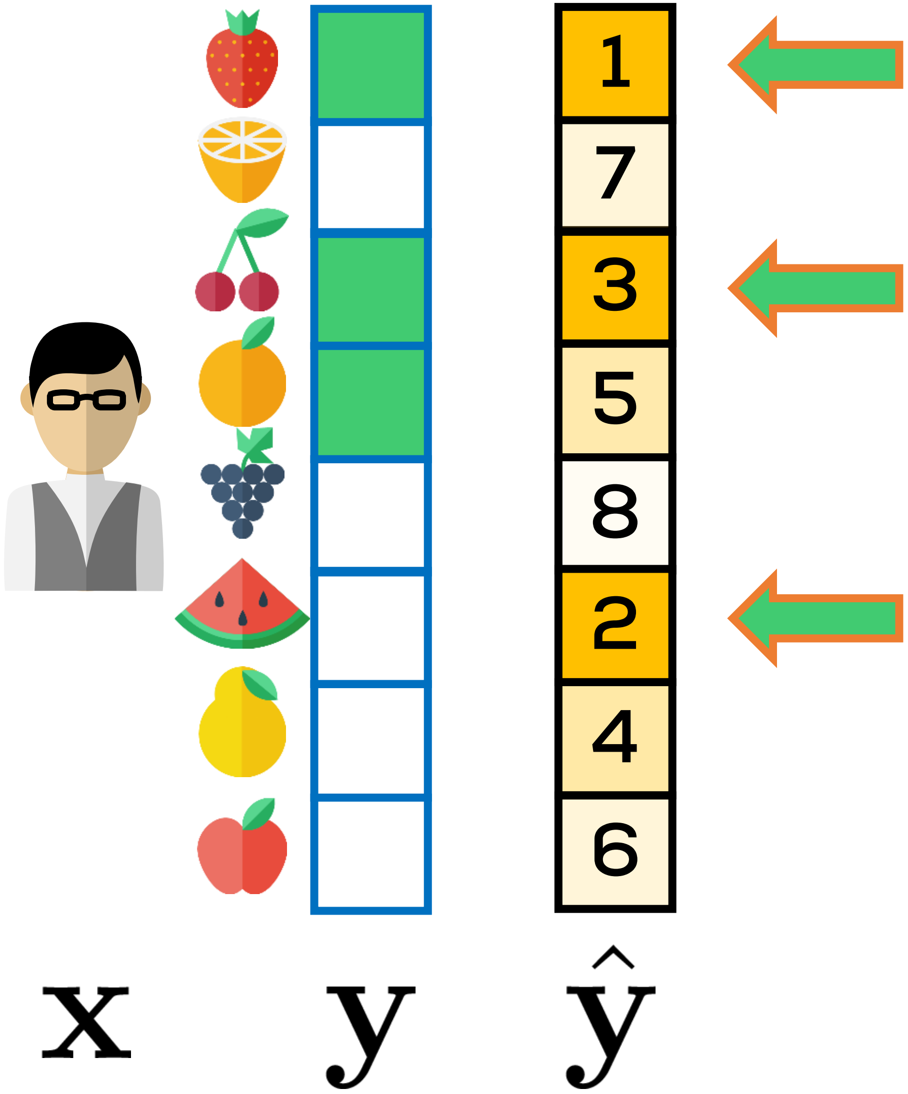
Precision@3

Multi-label Classification Loss Functions



Precision@3

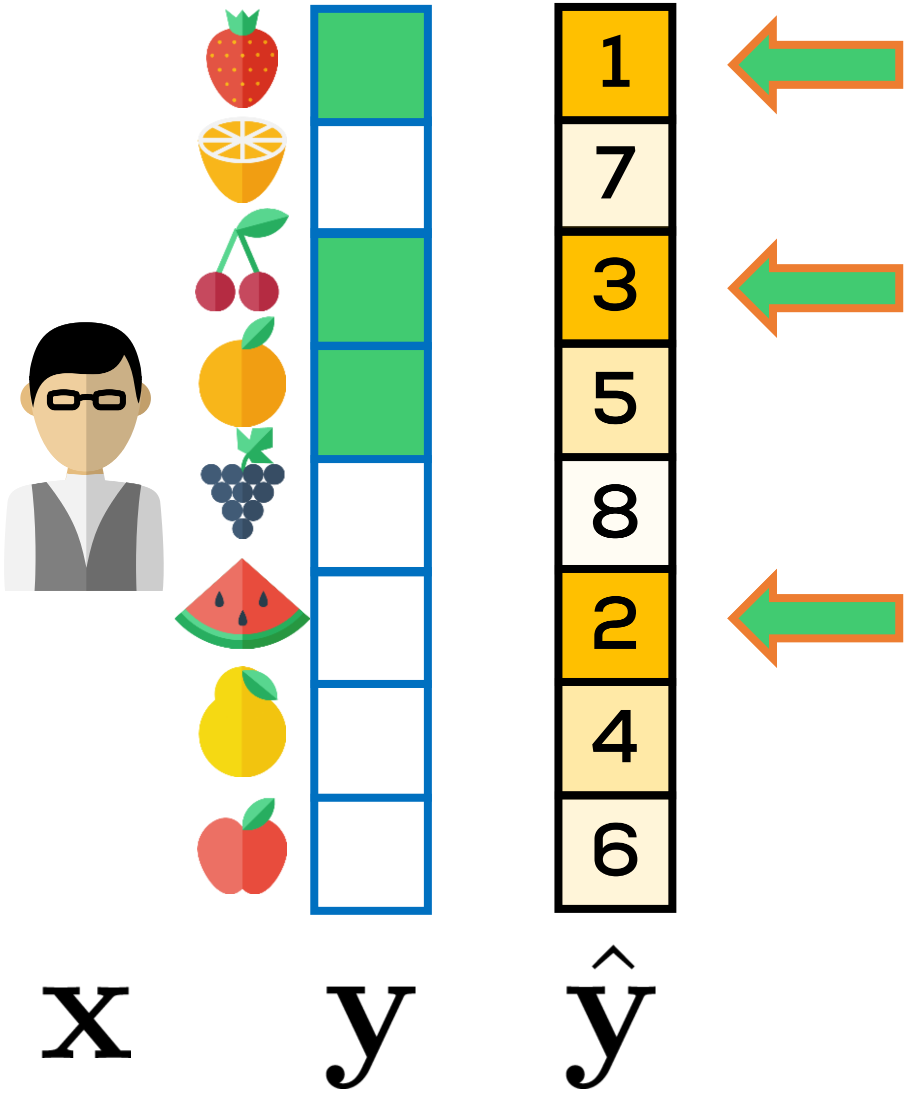
Multi-label Classification Loss Functions



Precision@3

Recall@k

Multi-label Classification Loss Functions



Precision@3

Recall@k

Many, many, more
NDCG, AUC, MRR, MAP

How to optimize these losses?

... aka where is the rest of the Math?

Some Fundamentals

Functional Analysis, Optimization Theory

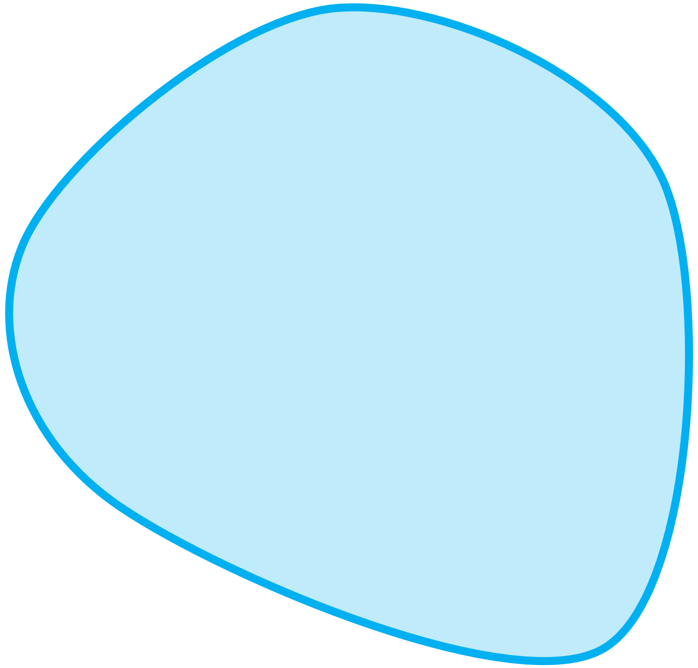
Convex Sets

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$

Convex Sets

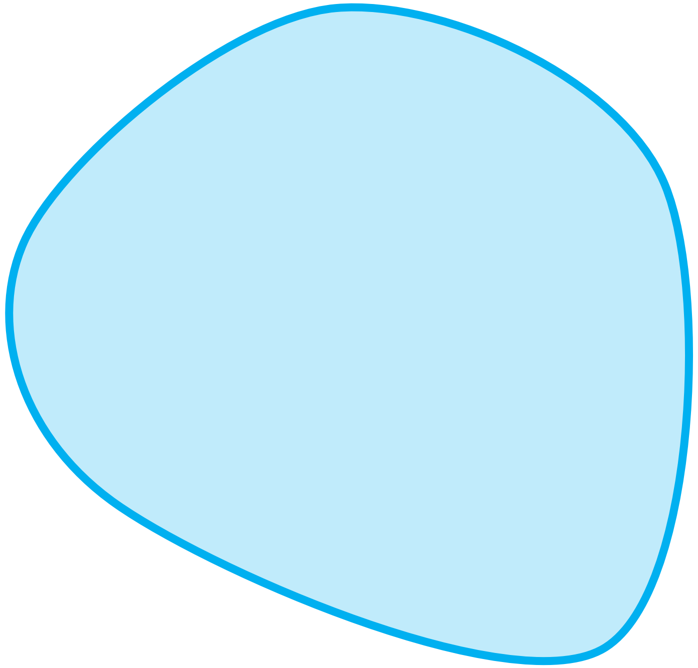
$$\mathcal{C} \subseteq \mathbb{R}^d$$



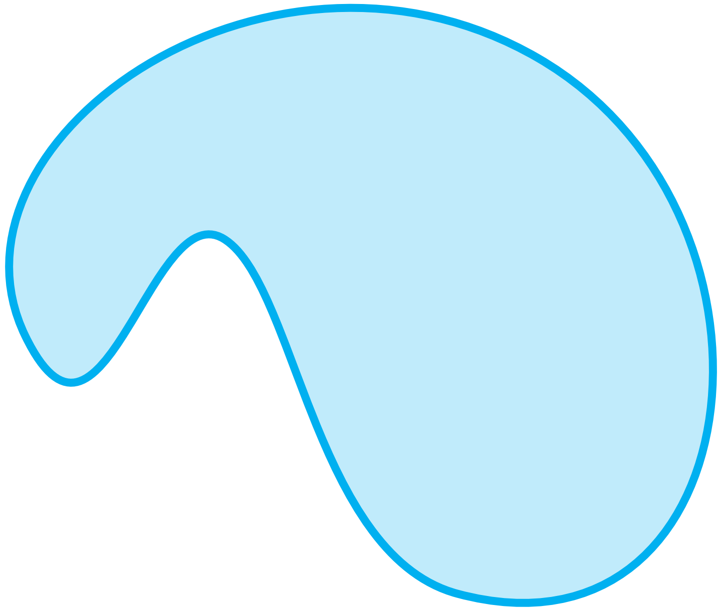
CONVEX SET

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$



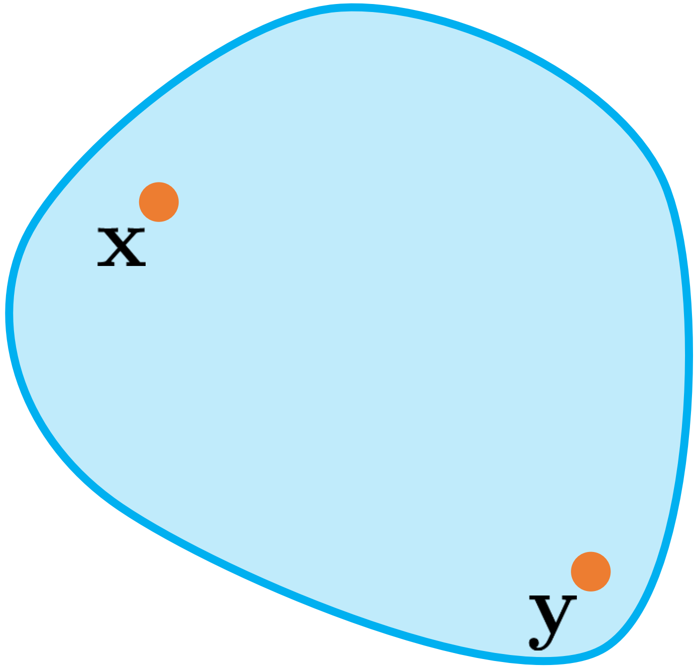
CONVEX SET



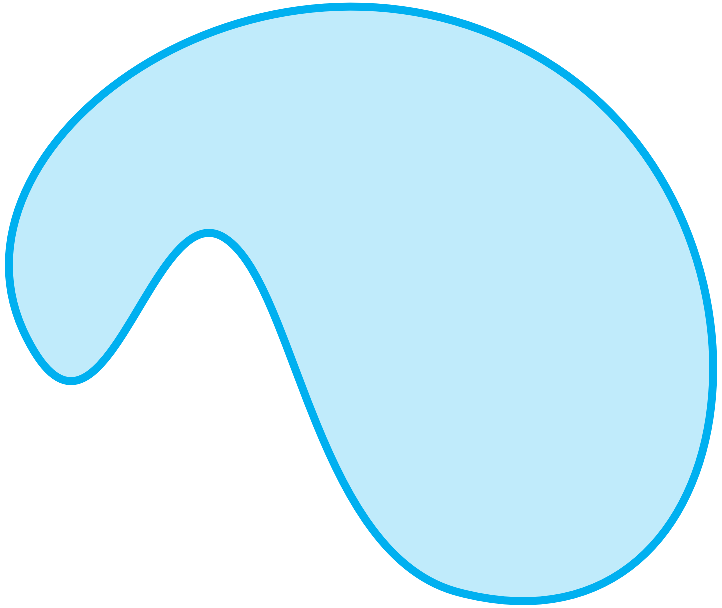
NON-CONVEX SET

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$



CONVEX SET



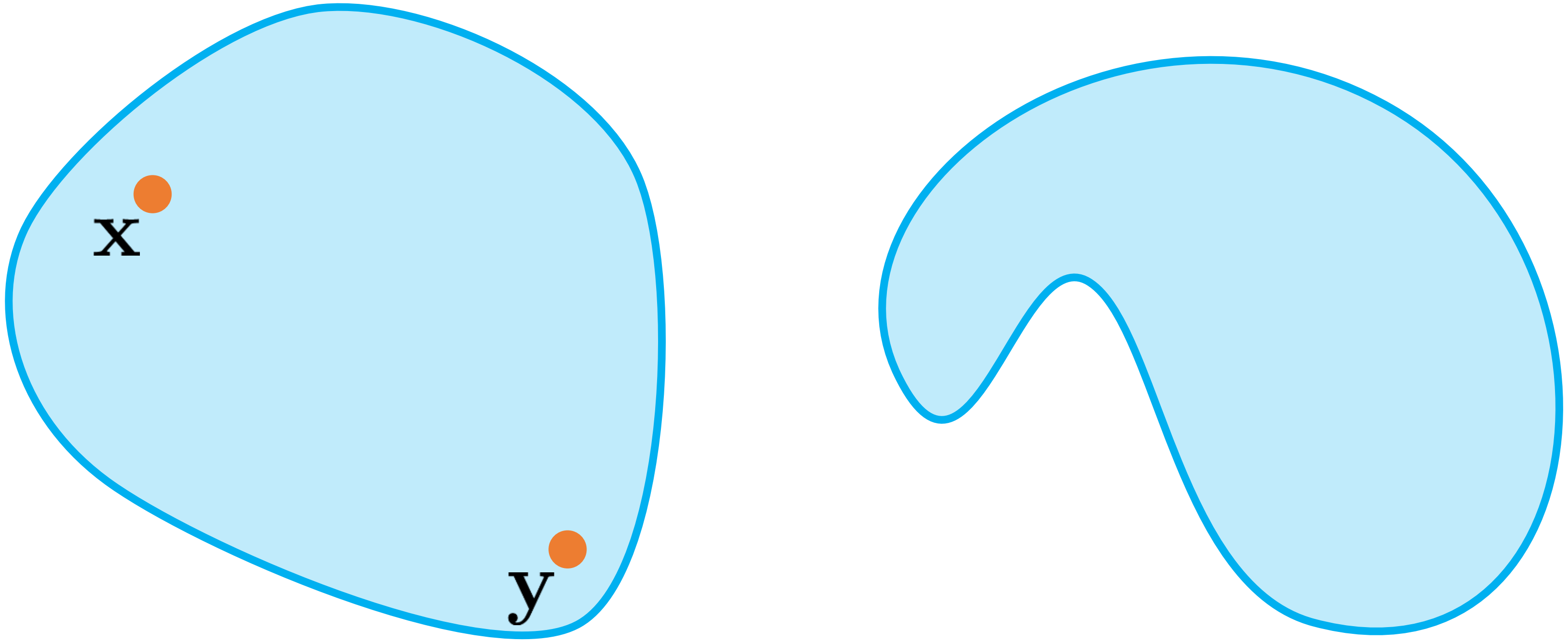
NON-CONVEX SET

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$$

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$$
$$\forall \lambda \in [0, 1]$$

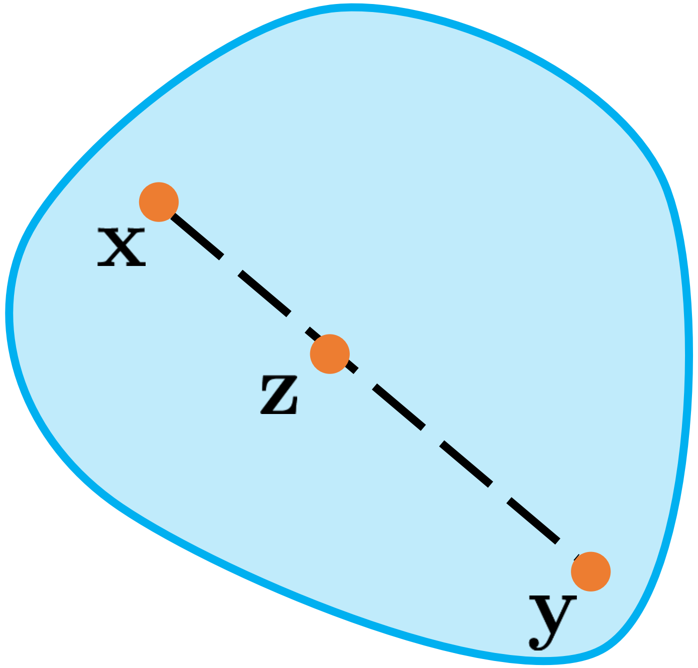


CONVEX SET

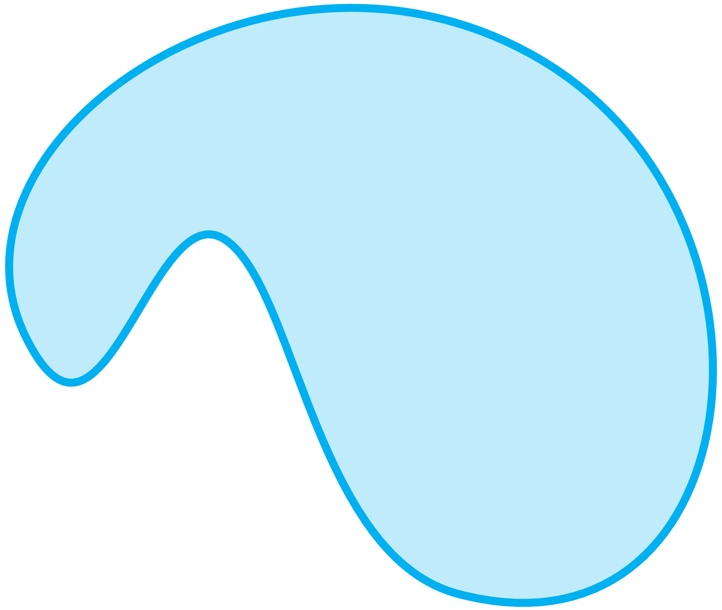
NON-CONVEX SET

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$



CONVEX SET

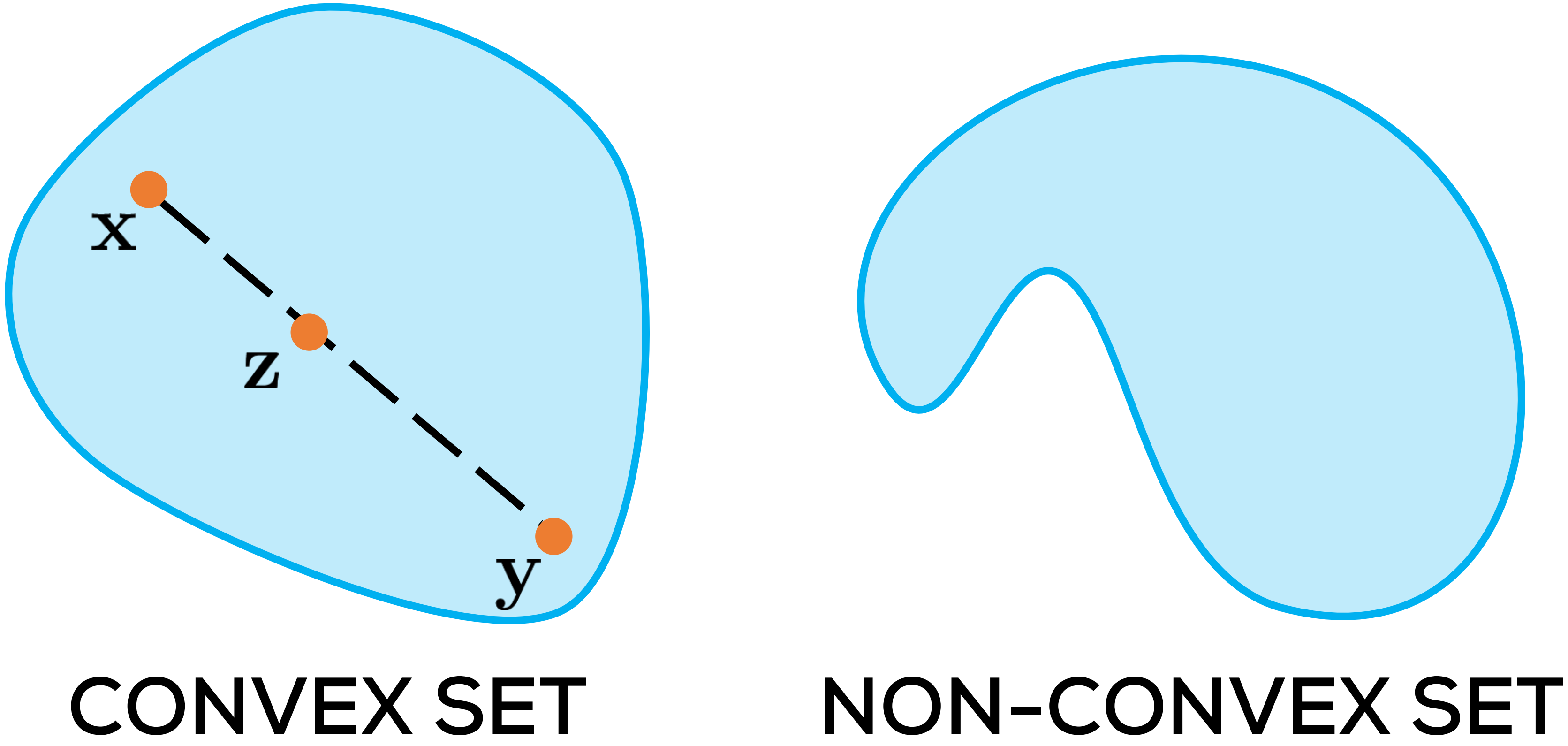


NON-CONVEX SET

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$$
$$\forall \lambda \in [0, 1]$$
$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}$$

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$



$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$$

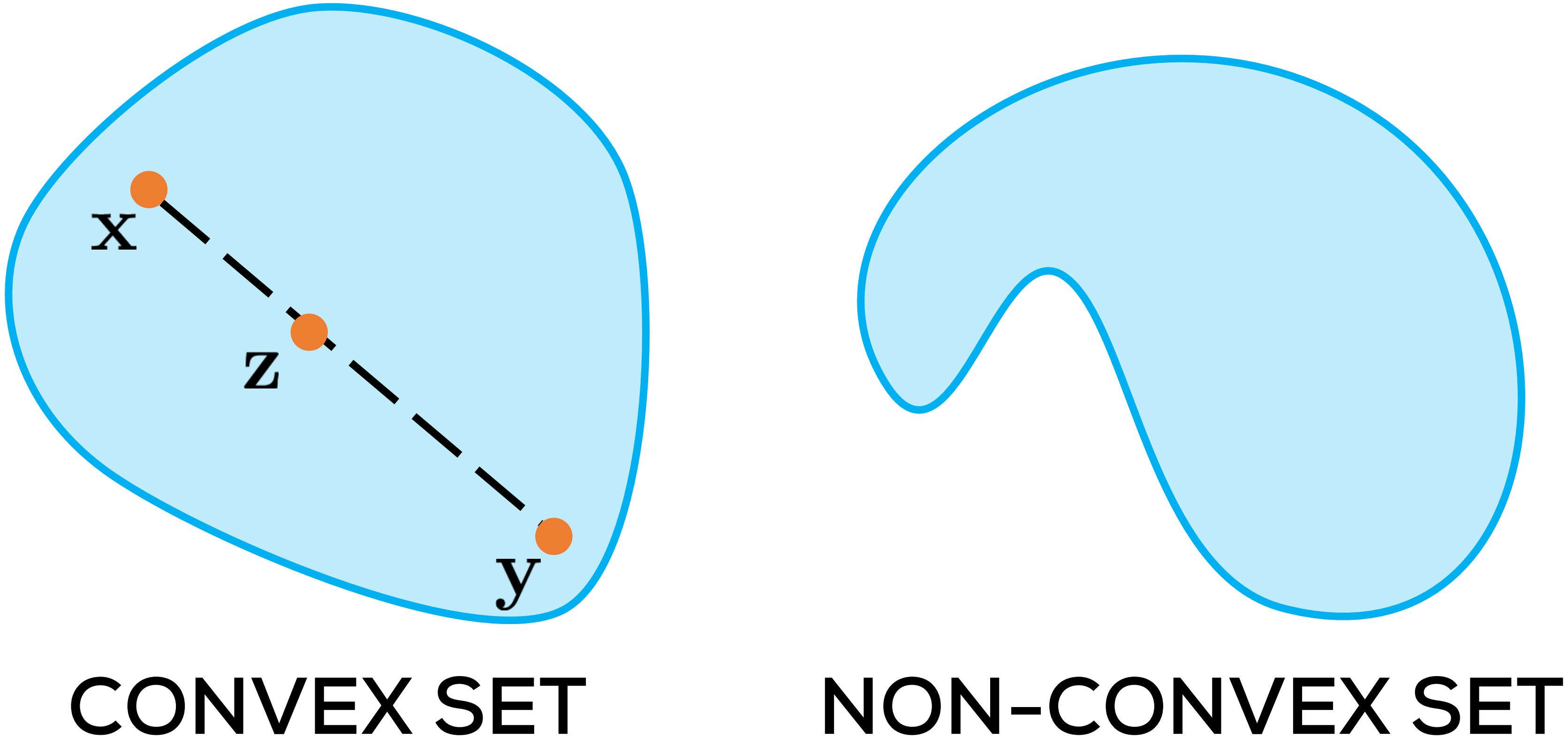
$$\forall \lambda \in [0, 1]$$

$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}$$

Convex combination

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$

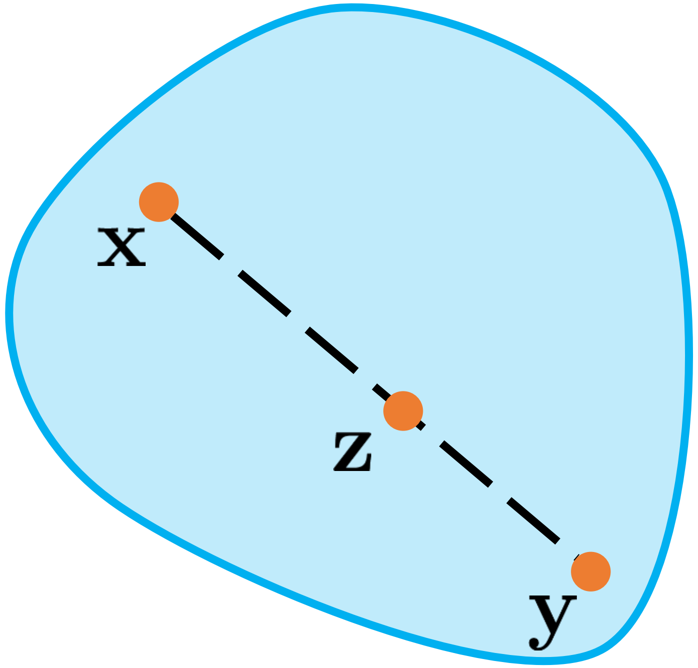


$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$$
$$\forall \lambda \in [0, 1]$$
$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y} \in \mathcal{C}$$

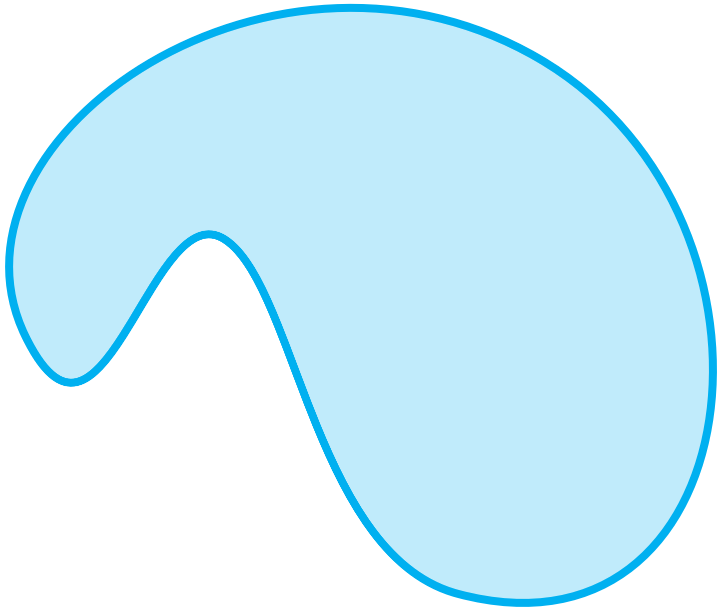
Convex combination

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$



CONVEX SET



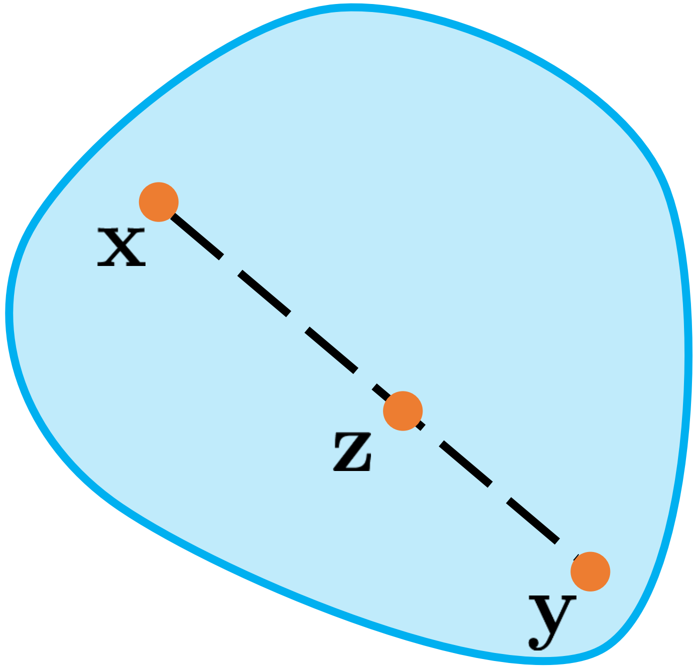
NON-CONVEX SET

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$$
$$\forall \lambda \in [0,1]$$
$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1-\lambda) \cdot \mathbf{y} \in \mathcal{C}$$

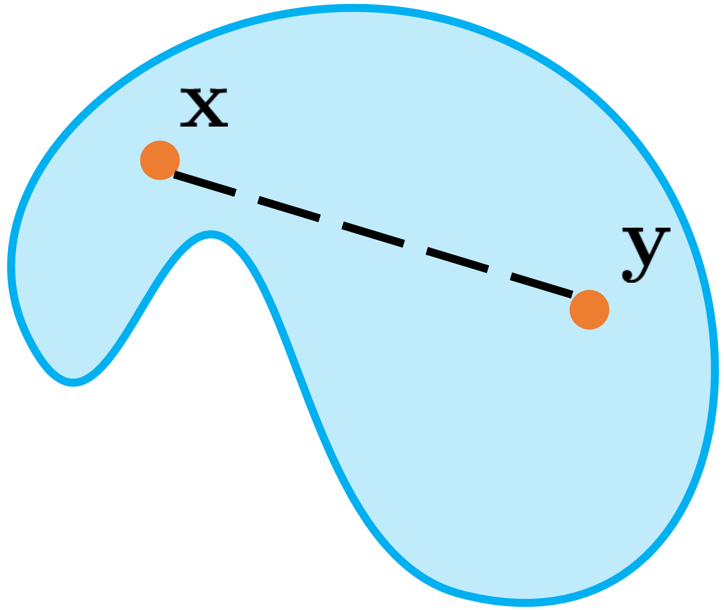
Convex combination

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$



CONVEX SET

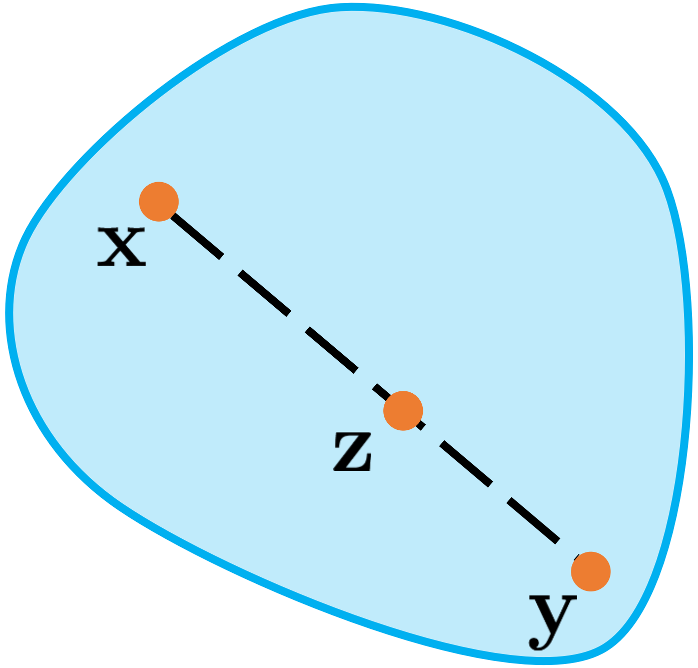


NON-CONVEX SET

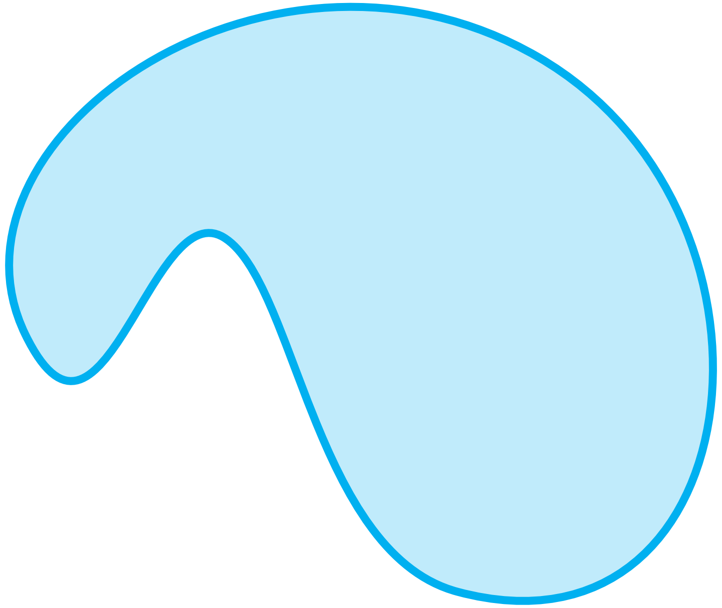
$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$$
$$\forall \lambda \in [0, 1]$$
$$\mathbf{z} = \lambda \cdot \mathbf{x} + \underbrace{(1 - \lambda) \cdot \mathbf{y}}_{\text{Convex combination}} \in \mathcal{C}$$

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$



CONVEX SET



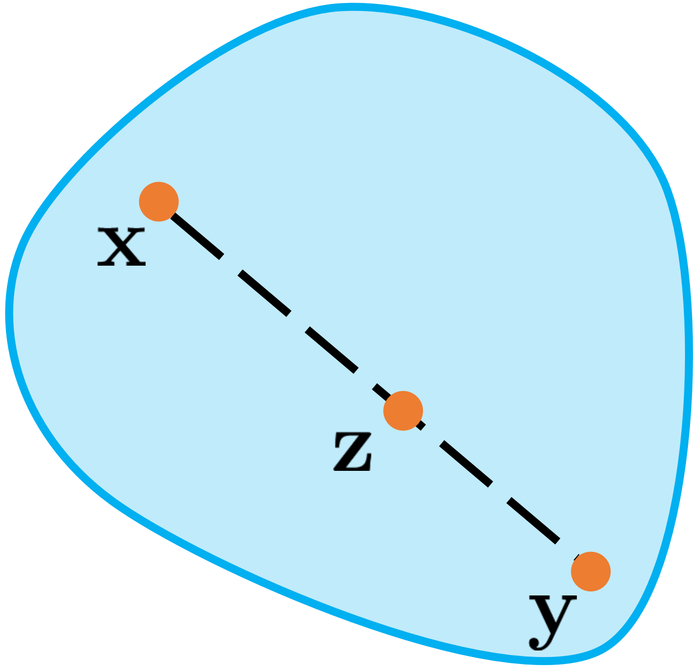
NON-CONVEX SET

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$$
$$\forall \lambda \in [0, 1]$$
$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y} \in \mathcal{C}$$

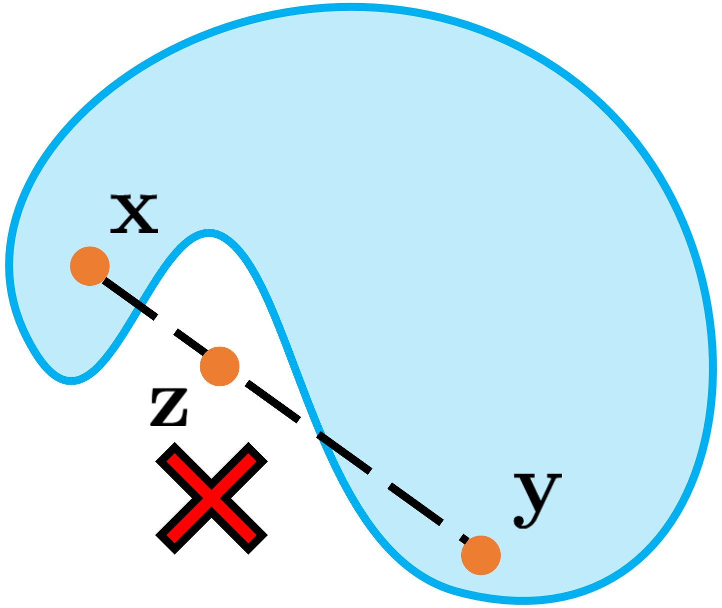
Convex combination

Convex Sets

$$\mathcal{C} \subseteq \mathbb{R}^d$$



CONVEX SET



NON-CONVEX SET

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$$

$$\forall \lambda \in [0, 1]$$

$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y} \in \mathcal{C}$$

Convex combination

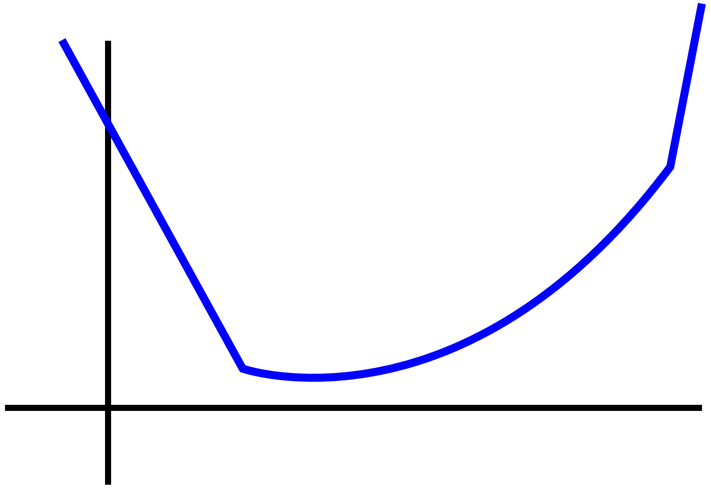
Convex Functions

Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$

Convex Functions

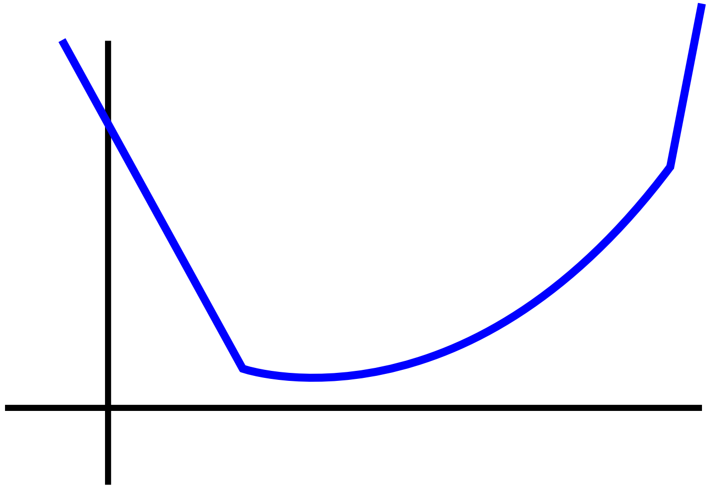
$$f : \mathbb{R}^d \to \mathbb{R}$$



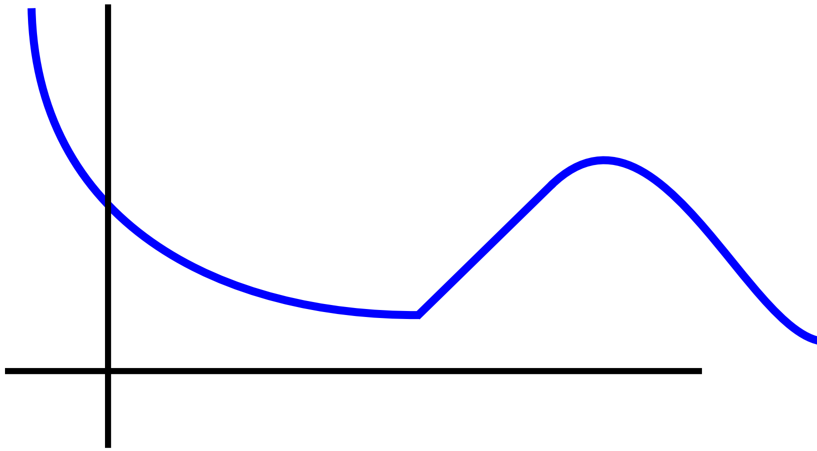
CONVEX FUNCTION

Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$



CONVEX FUNCTION

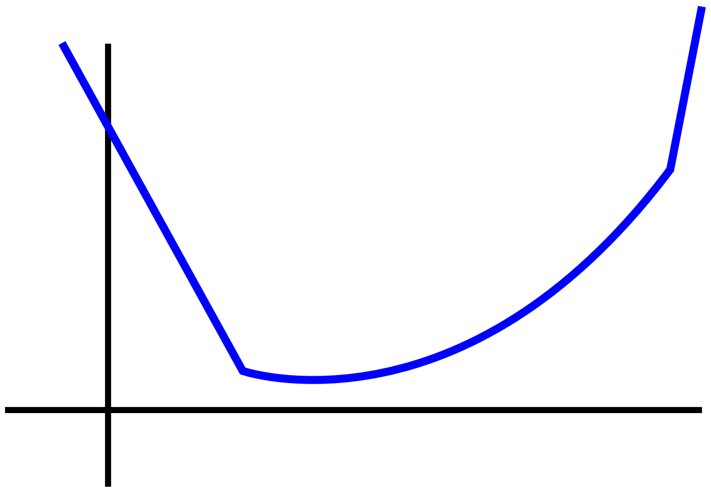


NON-CONVEX FUNCTION

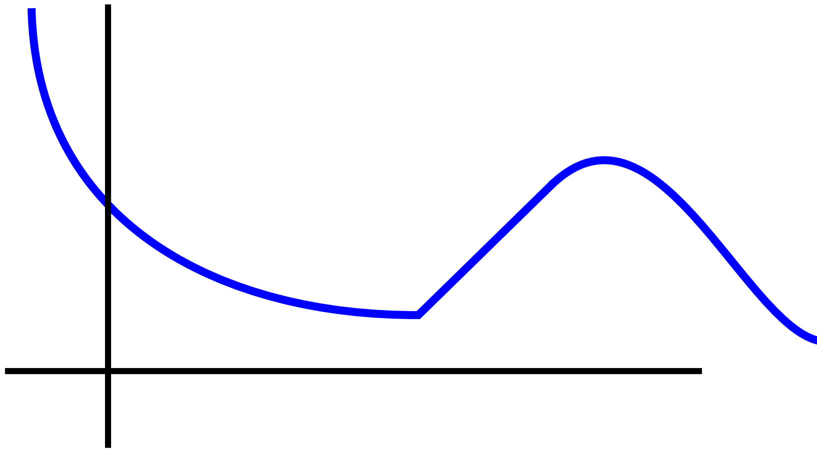
Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$

$$\forall \mathbf{x}, \mathbf{y}$$



CONVEX FUNCTION

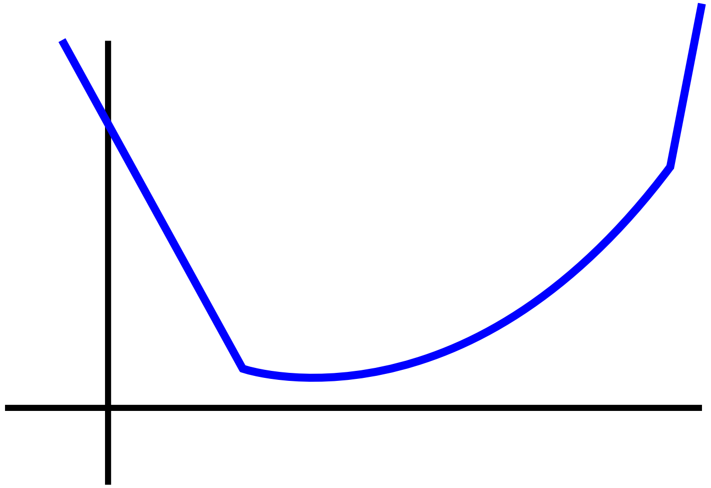


NON-CONVEX FUNCTION

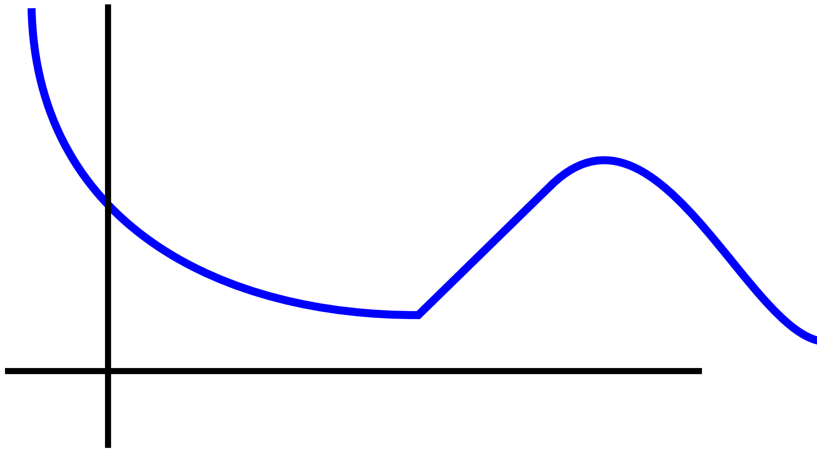
Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$

$$\forall \mathbf{x}, \mathbf{y}$$
$$\forall \lambda \in [0, 1]$$



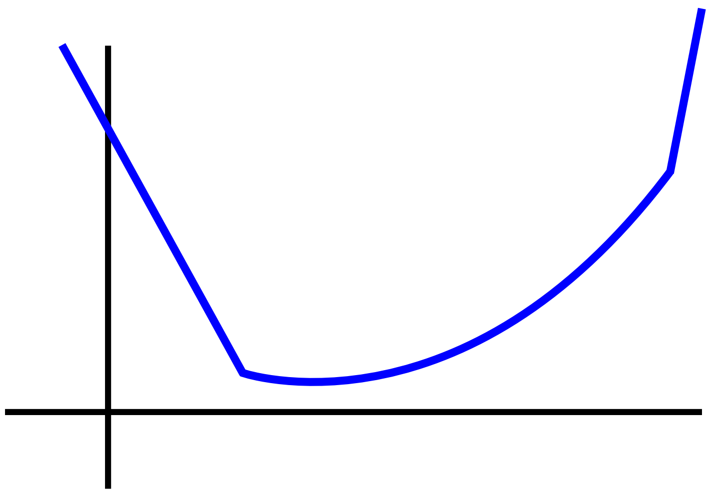
CONVEX FUNCTION



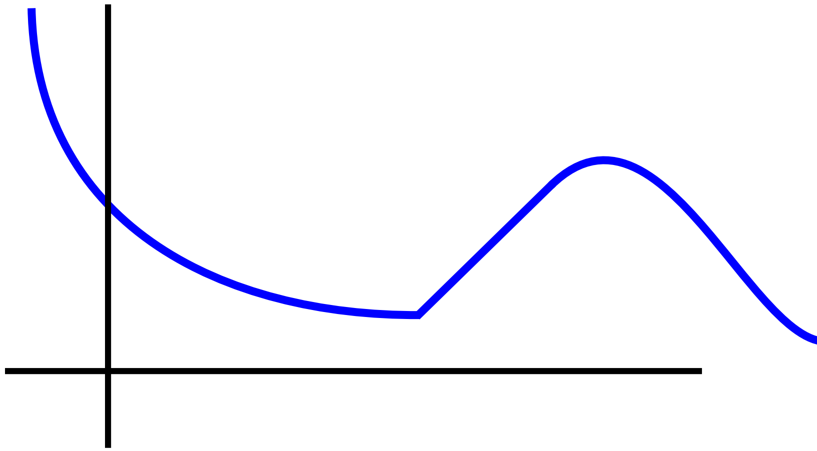
NON-CONVEX FUNCTION

Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$



CONVEX FUNCTION

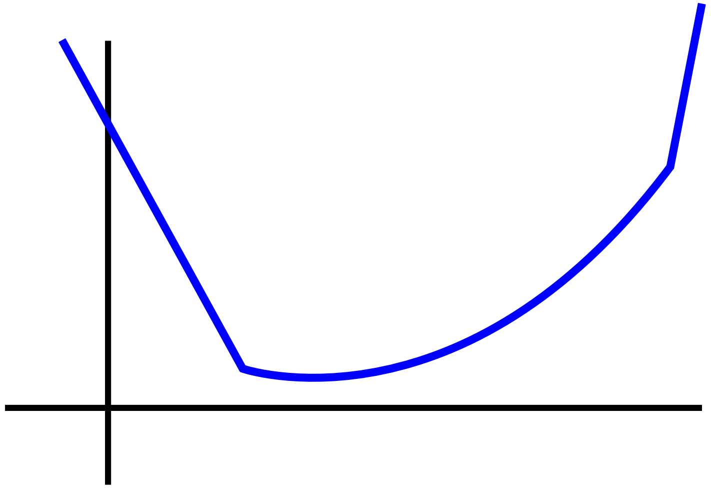


NON-CONVEX FUNCTION

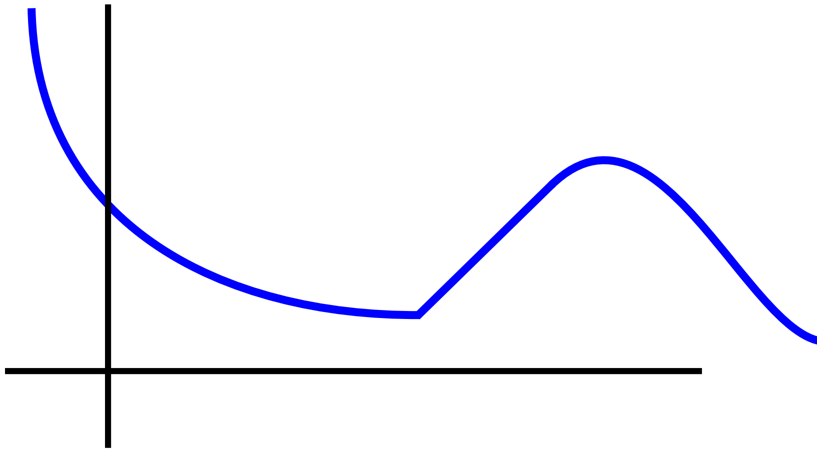
$$\forall \mathbf{x}, \mathbf{y}$$
$$\forall \lambda \in [0, 1]$$
$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}$$

Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$



CONVEX FUNCTION

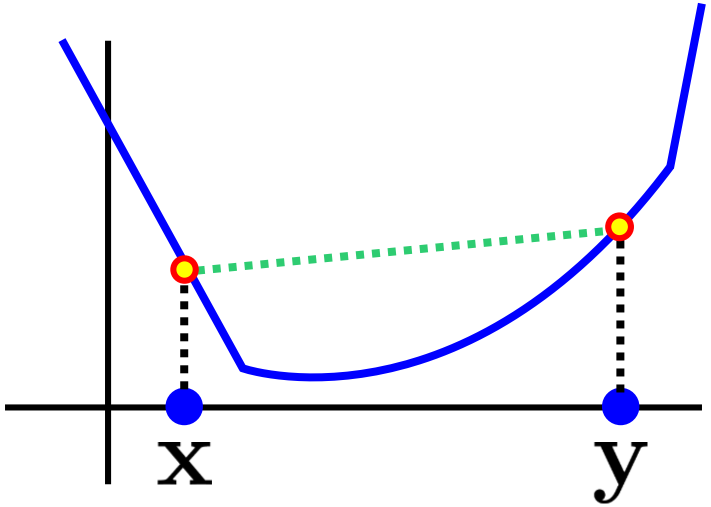


NON-CONVEX FUNCTION

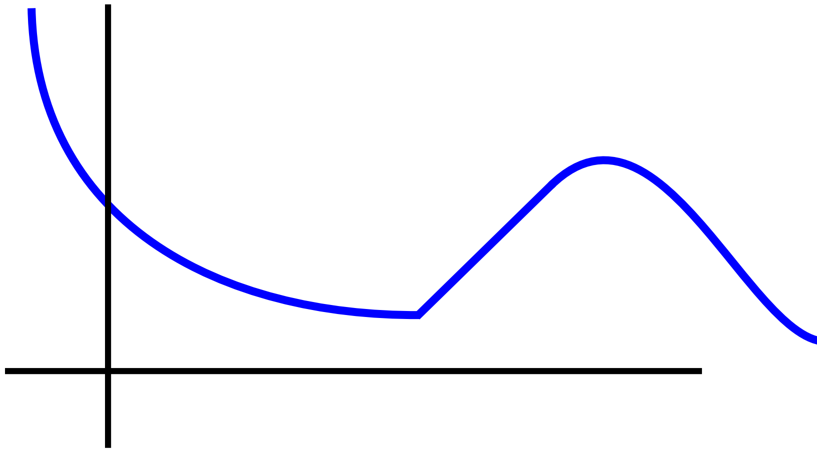
$$\forall \mathbf{x}, \mathbf{y}$$
$$\forall \lambda \in [0, 1]$$
$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}$$
$$f(\mathbf{z}) \leq \lambda \cdot f(\mathbf{x}) + (1 - \lambda) \cdot f(\mathbf{y})$$

Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$



CONVEX FUNCTION



NON-CONVEX FUNCTION

$$\forall \mathbf{x}, \mathbf{y}$$

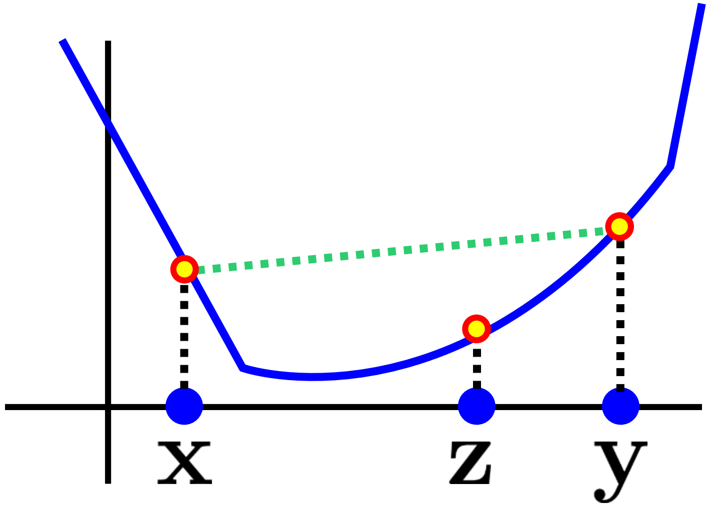
$$\forall \lambda \in [0, 1]$$

$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}$$

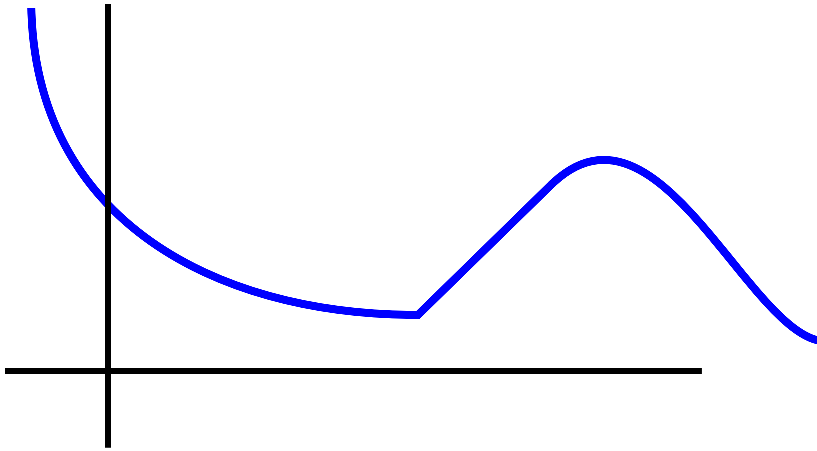
$$f(\mathbf{z}) \leq \lambda \cdot f(\mathbf{x}) + (1 - \lambda) \cdot f(\mathbf{y})$$

Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$



CONVEX FUNCTION

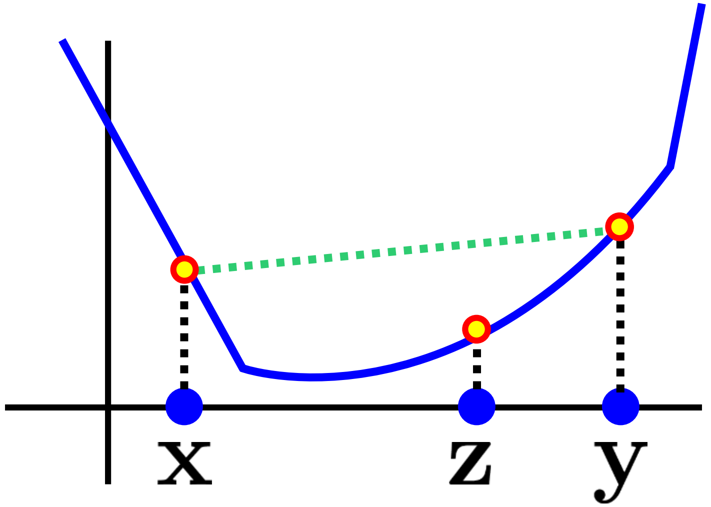


NON-CONVEX FUNCTION

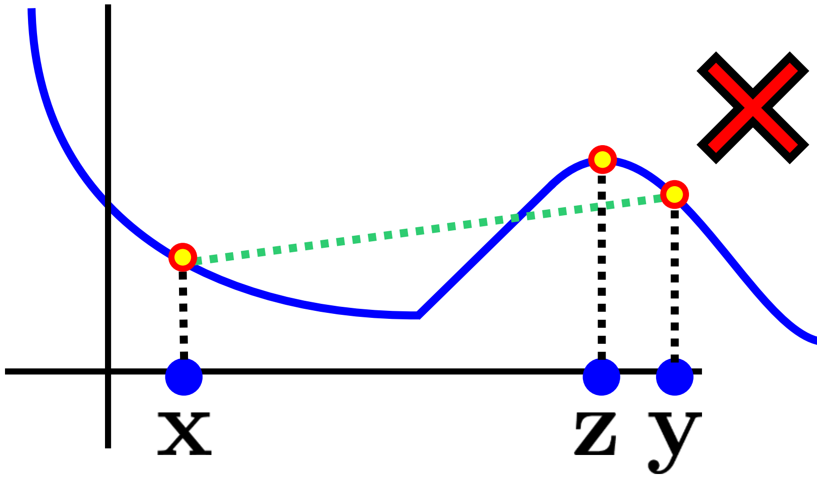
$$\forall \mathbf{x}, \mathbf{y}$$
$$\forall \lambda \in [0, 1]$$
$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}$$
$$f(\mathbf{z}) \leq \lambda \cdot f(\mathbf{x}) + (1 - \lambda) \cdot f(\mathbf{y})$$

Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$



CONVEX FUNCTION



NON-CONVEX FUNCTION

$$\forall \mathbf{x}, \mathbf{y}$$

$$\forall \lambda \in [0, 1]$$

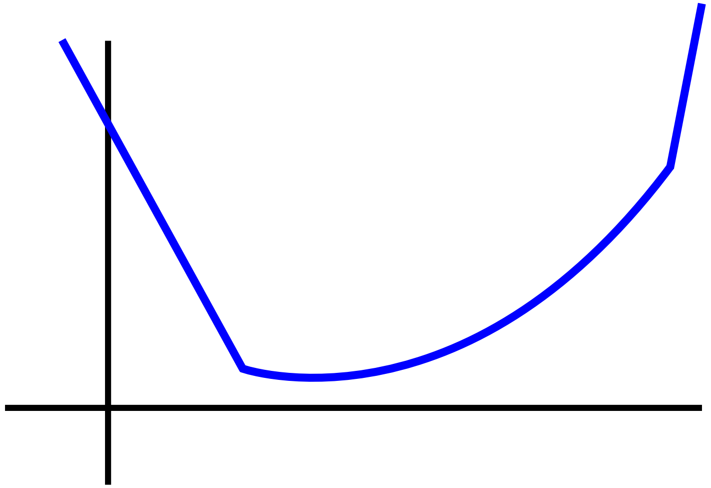
$$\mathbf{z} = \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{y}$$

$$f(\mathbf{z}) \leq \lambda \cdot f(\mathbf{x}) + (1 - \lambda) \cdot f(\mathbf{y})$$

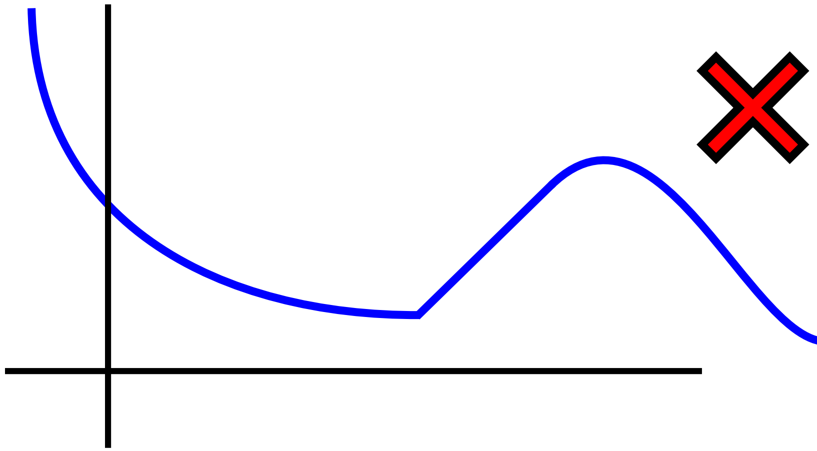
Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$

$$\forall \mathbf{x}, \mathbf{y}$$



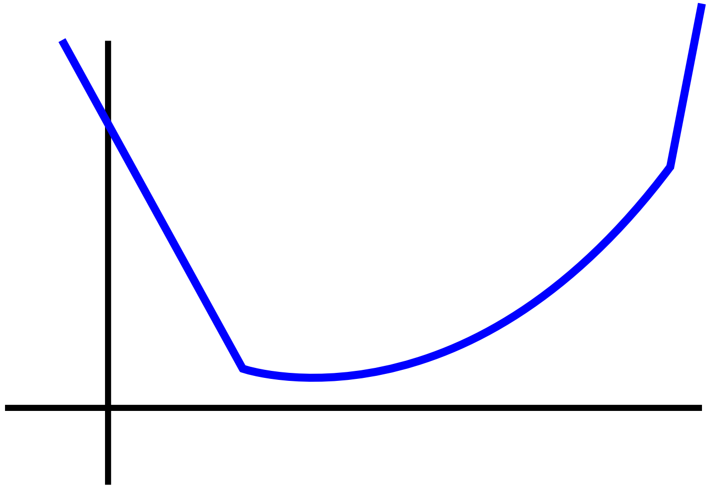
CONVEX FUNCTION

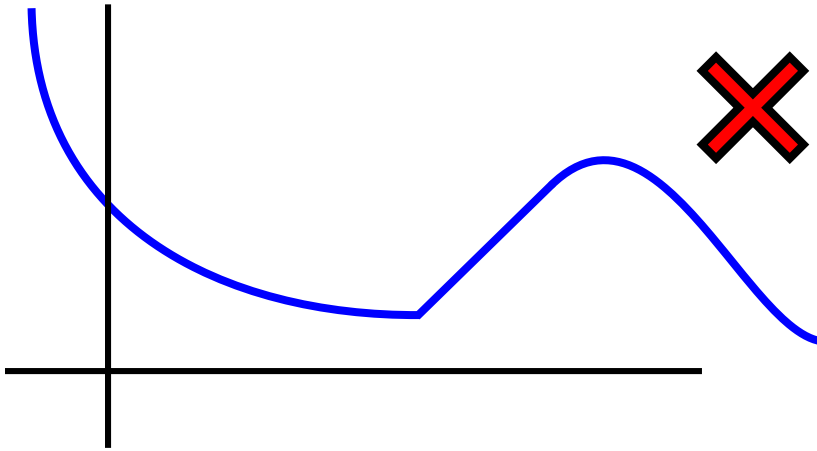


NON-CONVEX FUNCTION

Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$



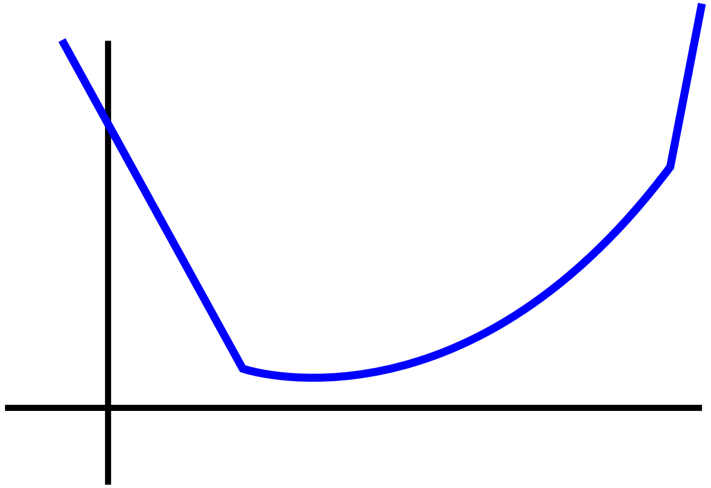


$\forall \mathbf{x}, \mathbf{y}$

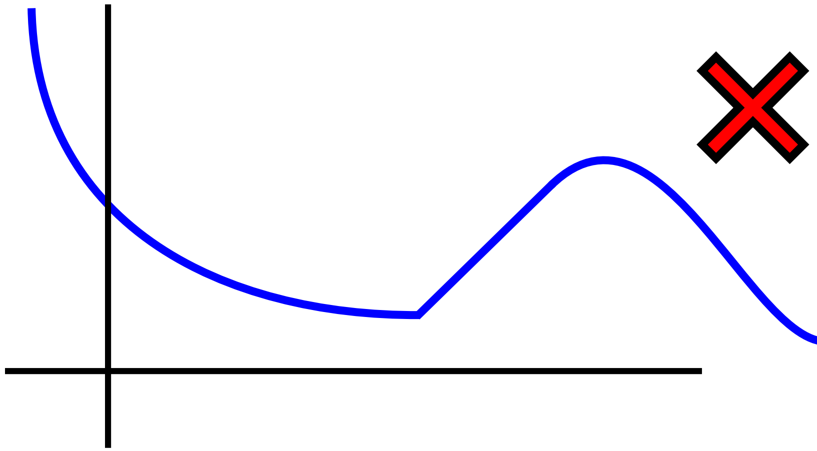
Differentiable function!

Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$



CONVEX FUNCTION



NON-CONVEX FUNCTION

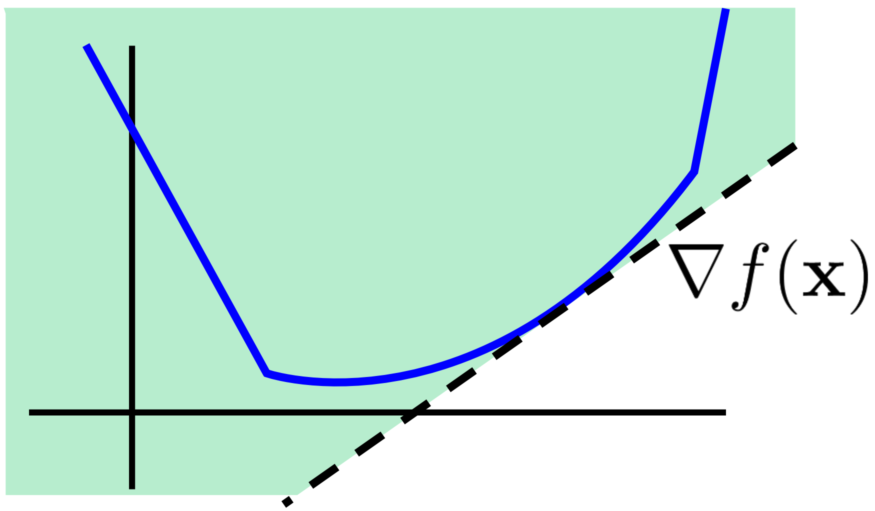
$\forall \mathbf{x}, \mathbf{y}$

Differentiable function!

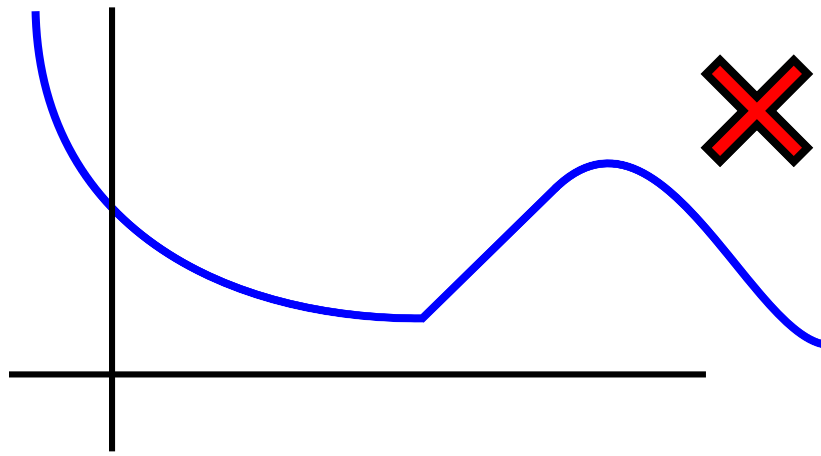
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$



CONVEX FUNCTION



NON-CONVEX FUNCTION

$$\forall \mathbf{x}, \mathbf{y}$$

Differentiable function!

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

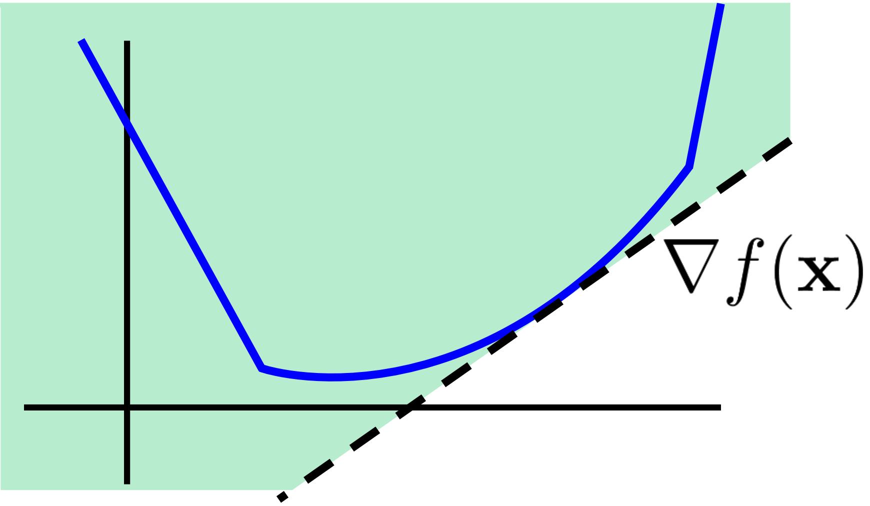
Convex Functions

$$f : \mathbb{R}^d \to \mathbb{R}$$

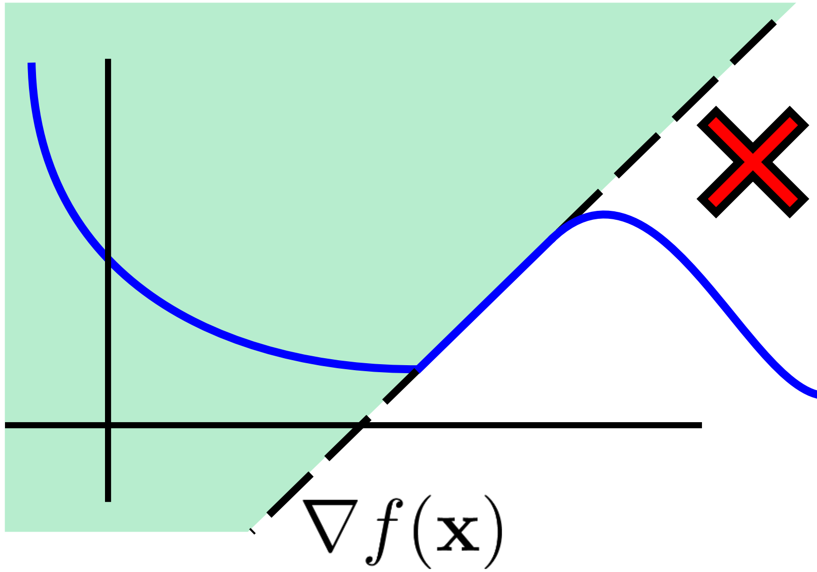
$\forall \mathbf{x}, \mathbf{y}$

Differentiable function!

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$



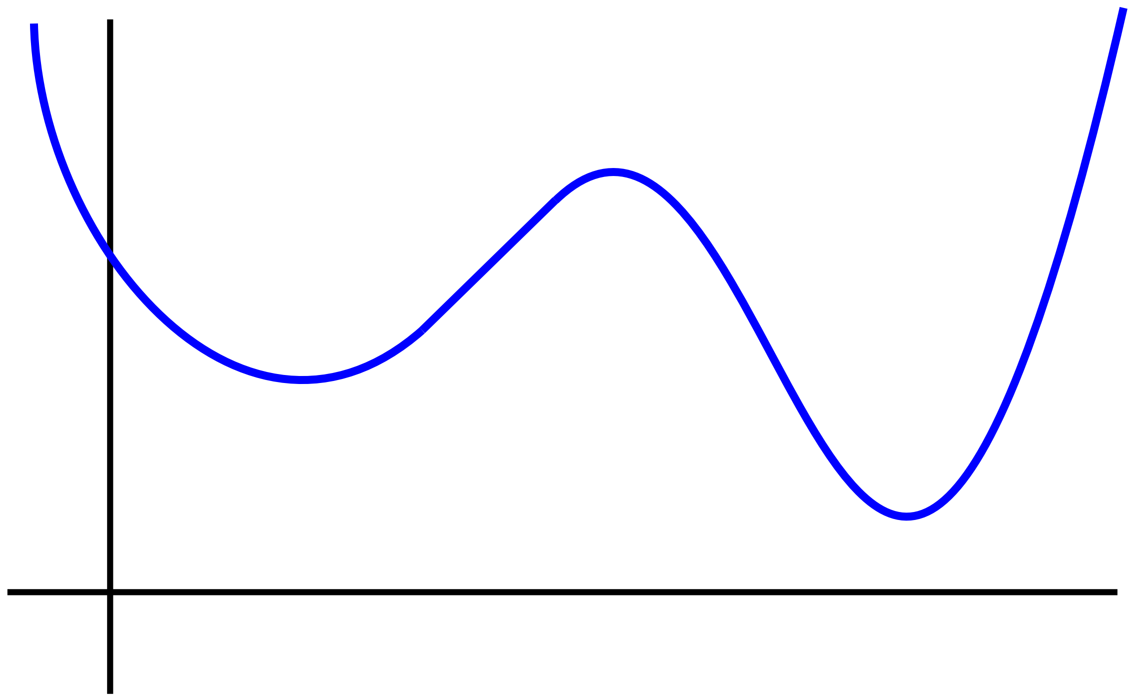
CONVEX FUNCTION



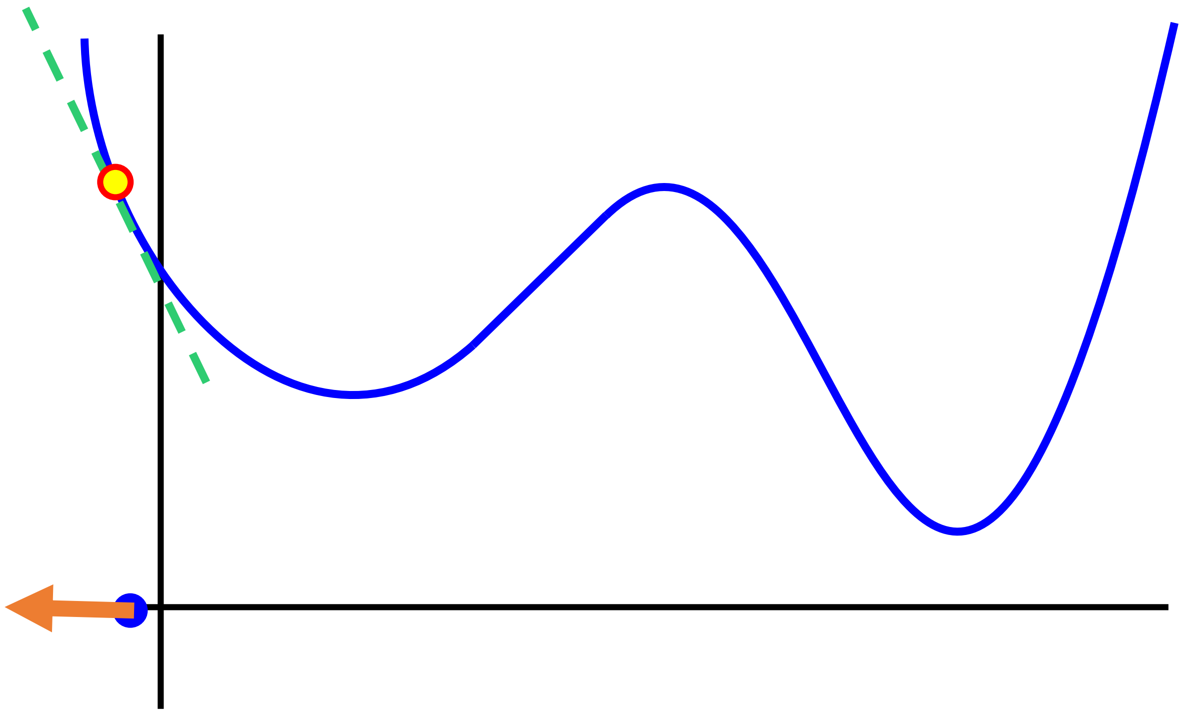
NON-CONVEX FUNCTION

Gradients

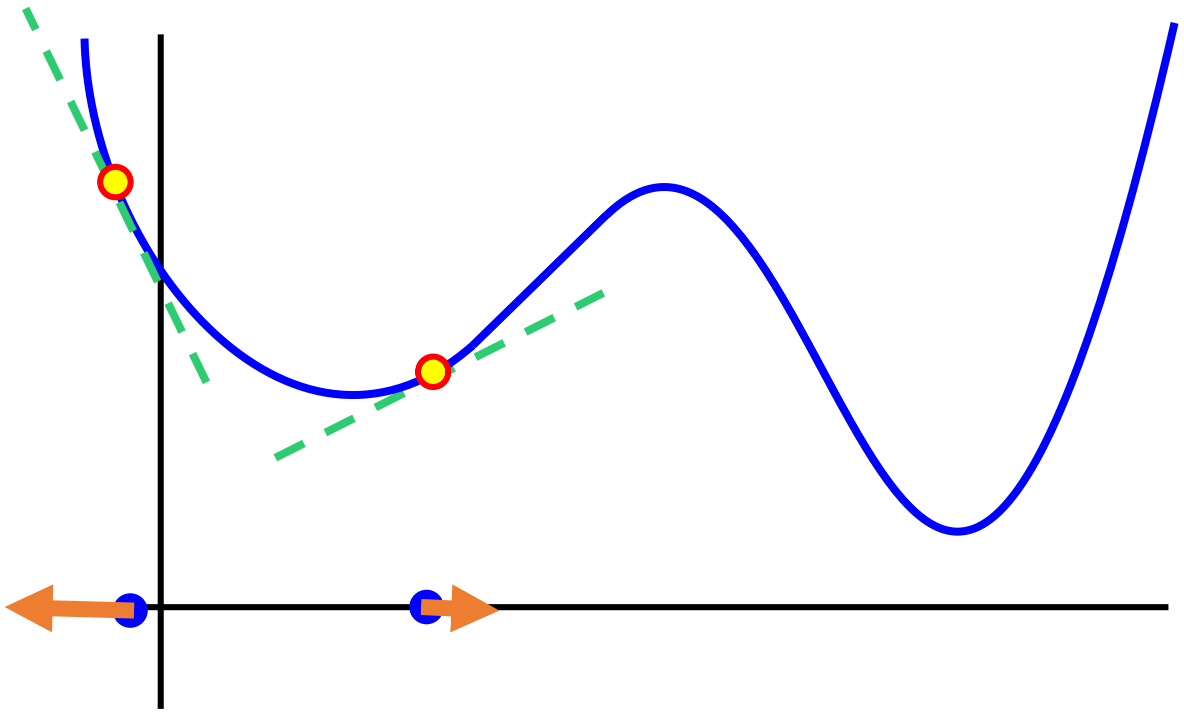
Gradients



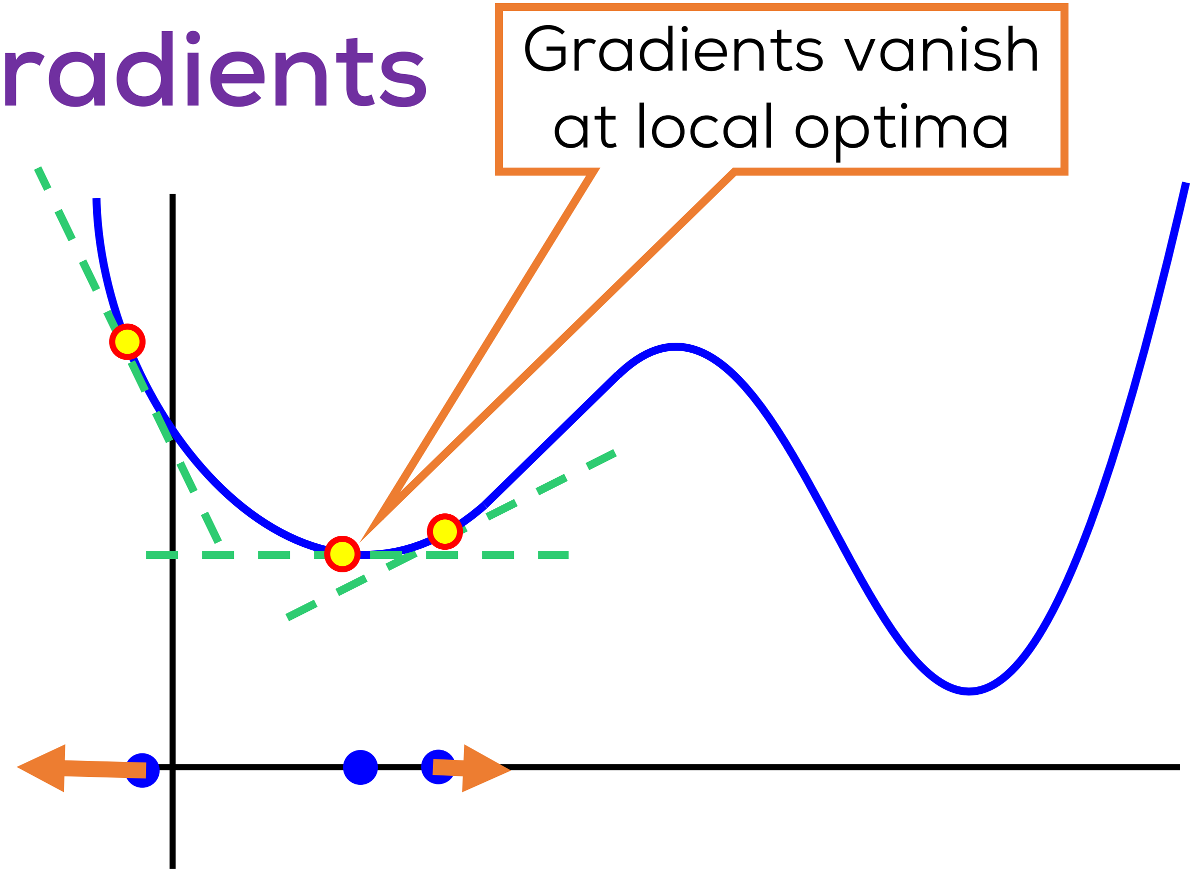
Gradients



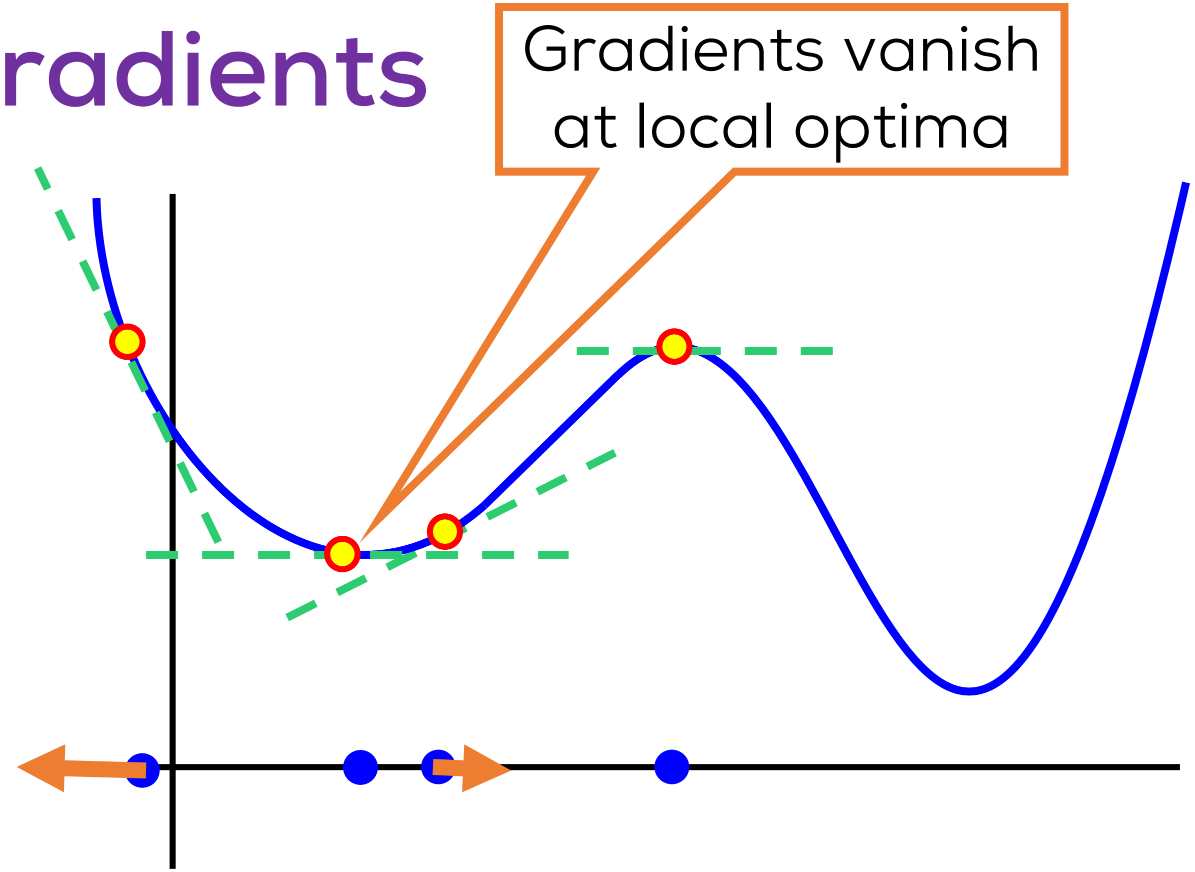
Gradients



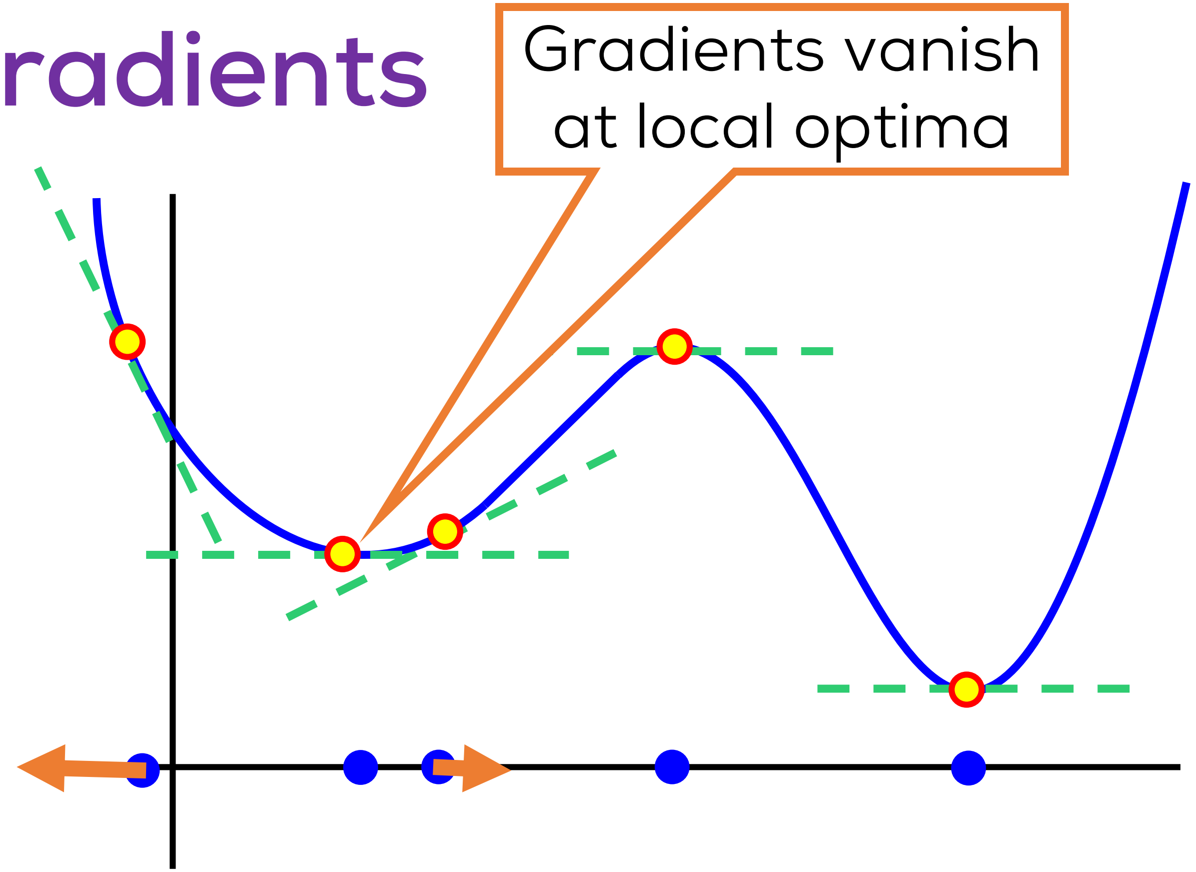
Gradients



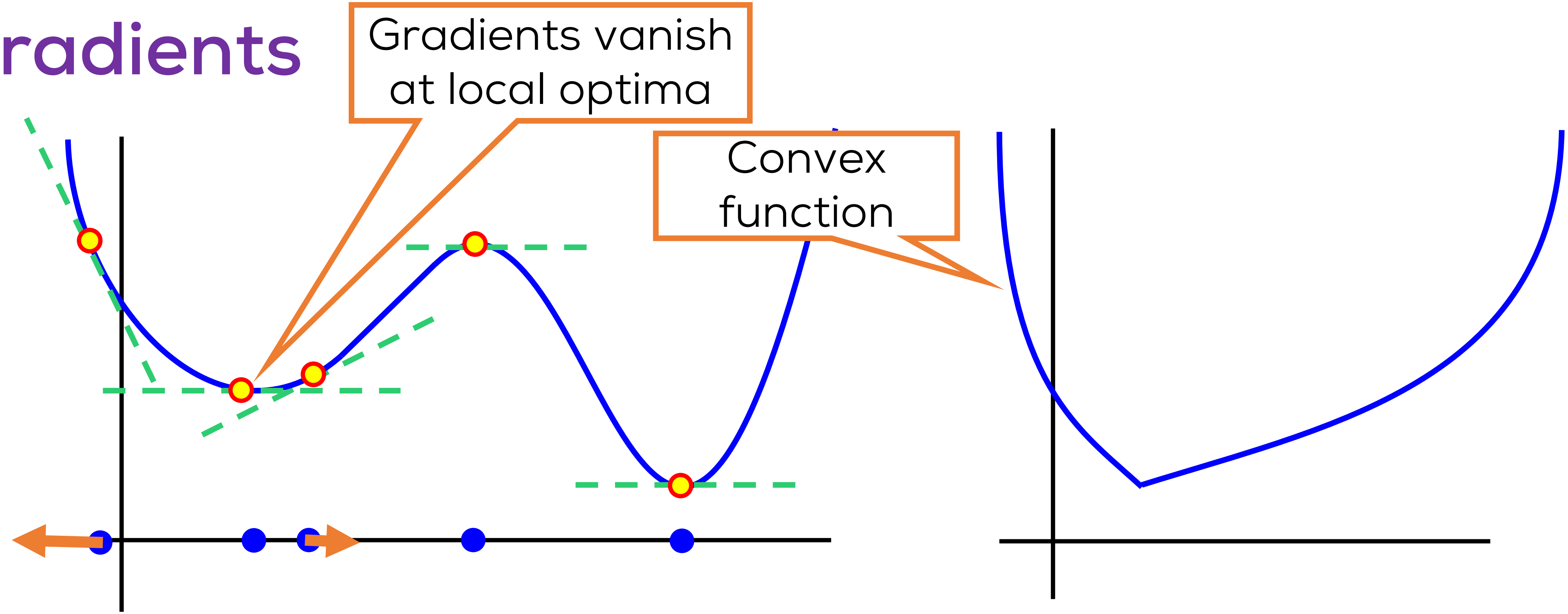
Gradients



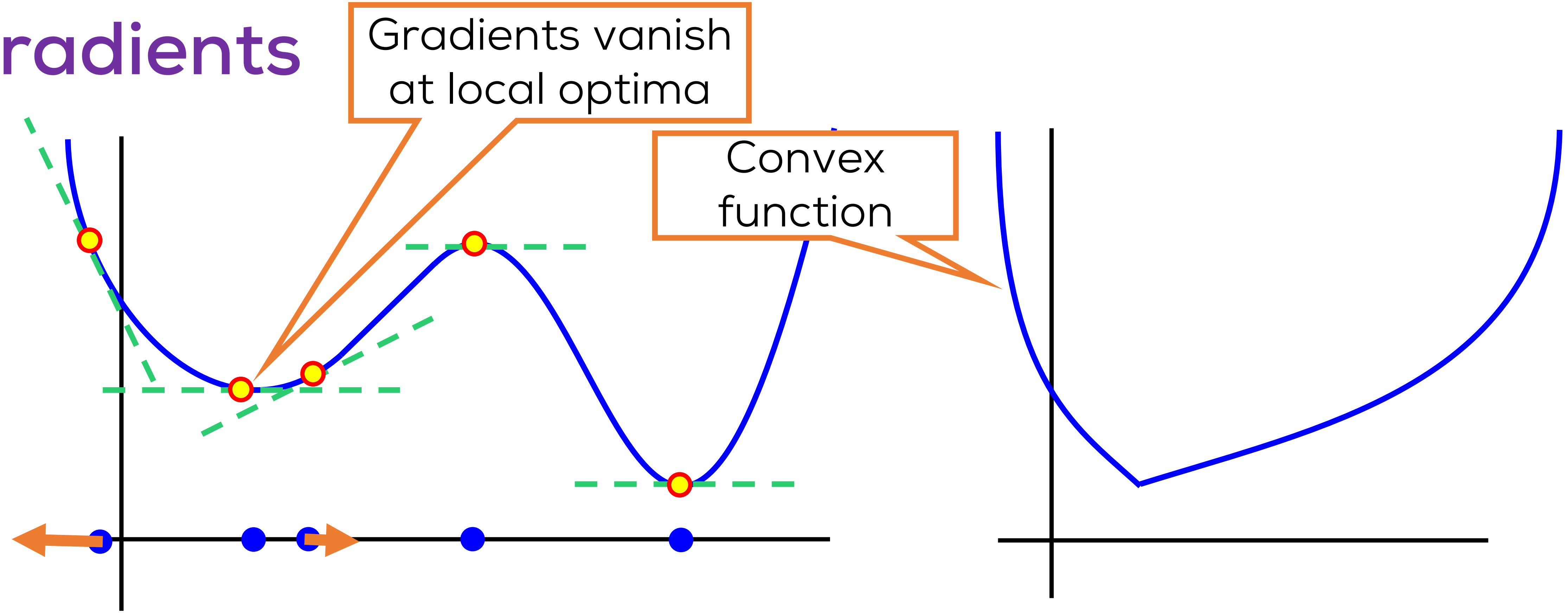
Gradients



Gradients



Gradients

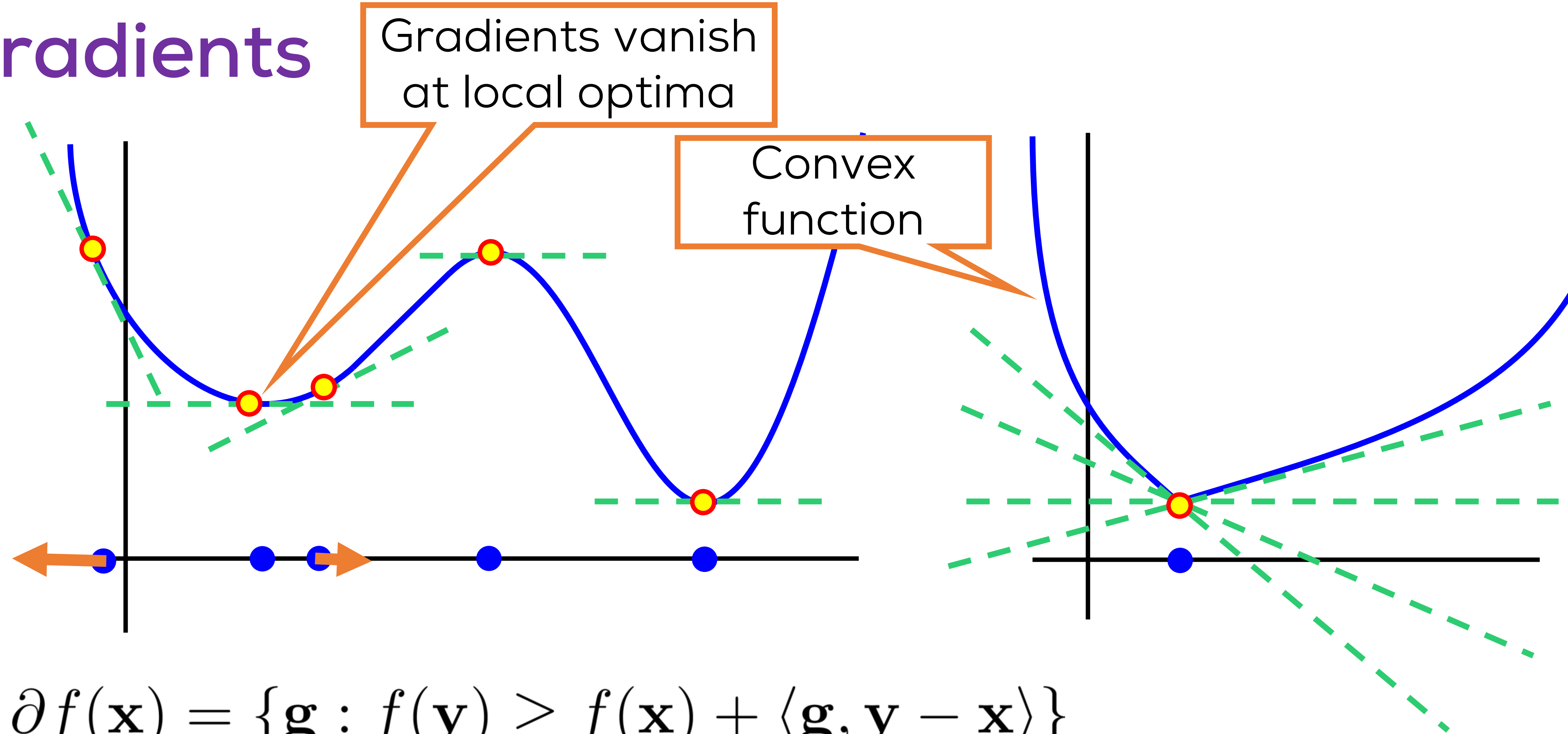


$$\partial f(\mathbf{x}) = \{\mathbf{g} : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle\}$$

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\} \text{ if } f \text{ differentiable}$$

$$\partial f(\mathbf{x}) \ni \mathbf{0} \text{ if } \mathbf{x} \text{ minimum}$$

Gradients

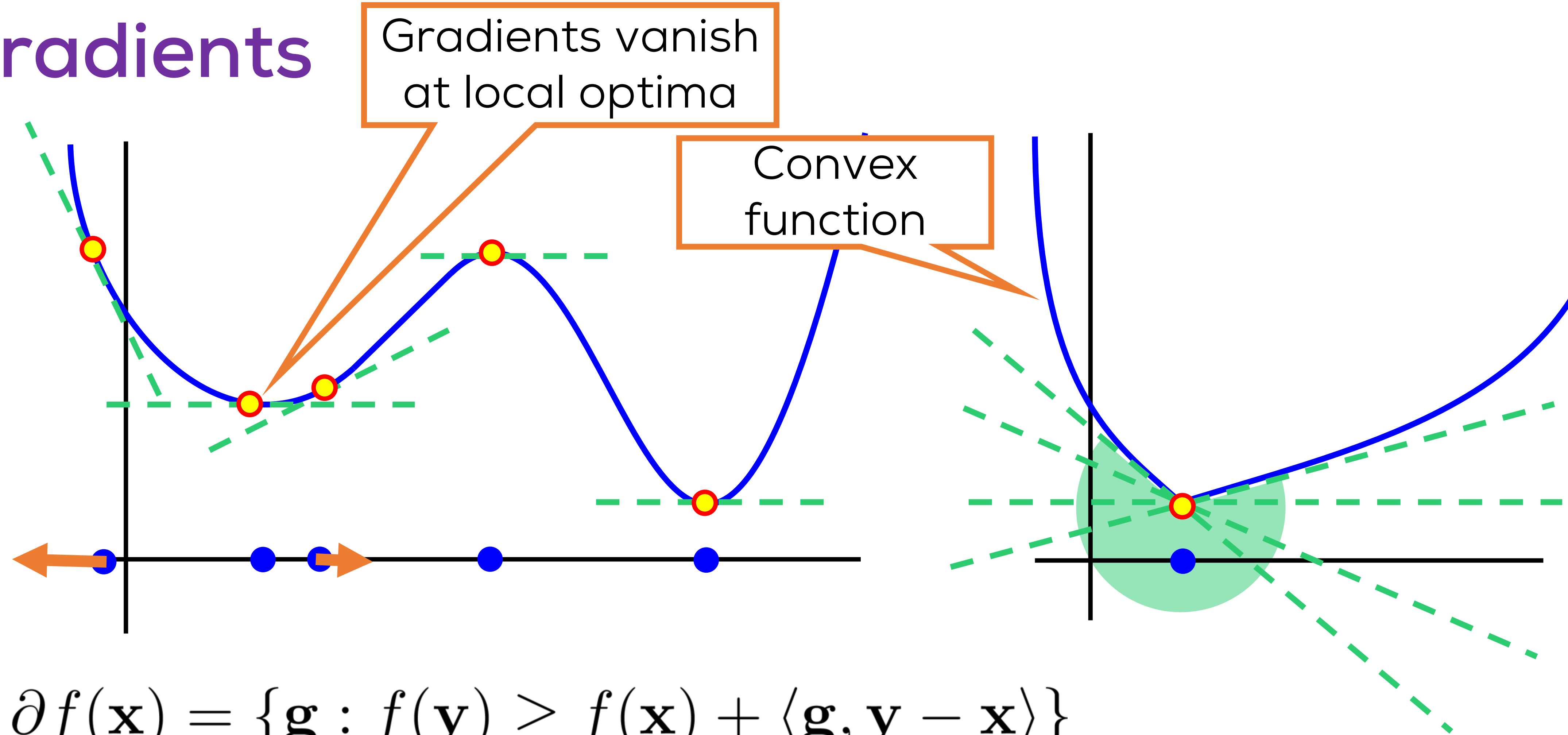


$$\partial f(\mathbf{x}) = \{\mathbf{g} : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle\}$$

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\} \text{ if } f \text{ differentiable}$$

$$\partial f(\mathbf{x}) \ni \mathbf{0} \text{ if } \mathbf{x} \text{ minimum}$$

Gradients

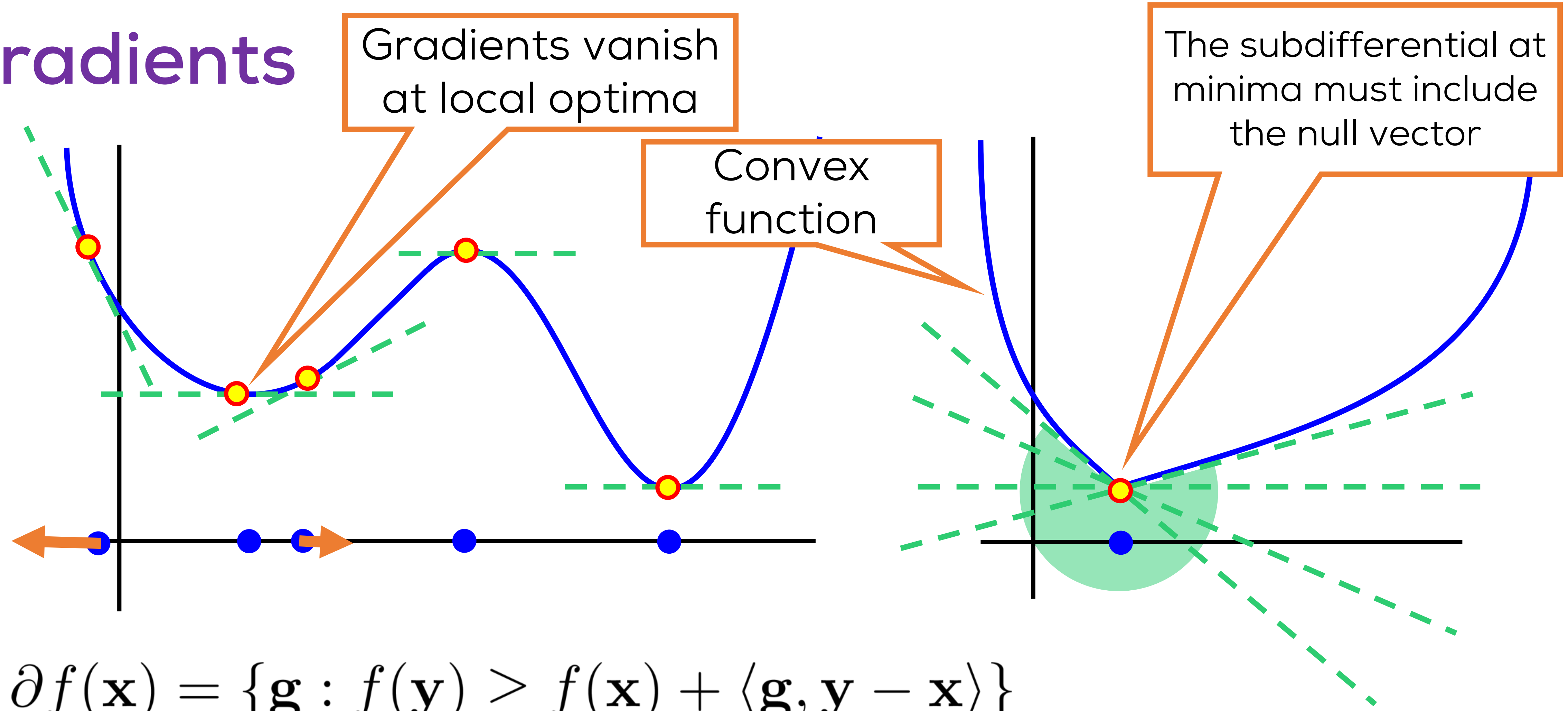


$$\partial f(\mathbf{x}) = \{\mathbf{g} : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle\}$$

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\} \text{ if } f \text{ differentiable}$$

$$\partial f(\mathbf{x}) \ni \mathbf{0} \text{ if } \mathbf{x} \text{ minimum}$$

Gradients

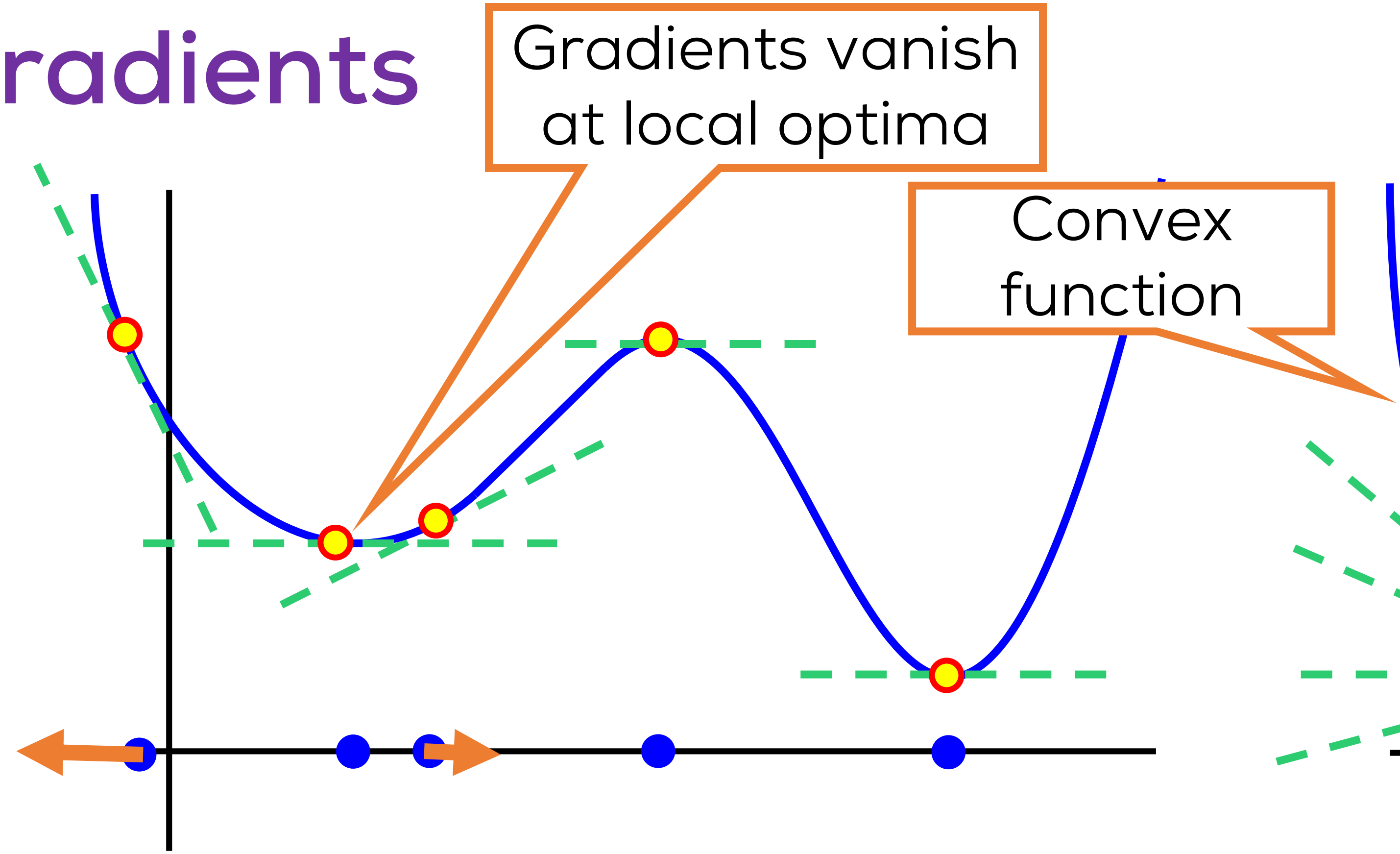


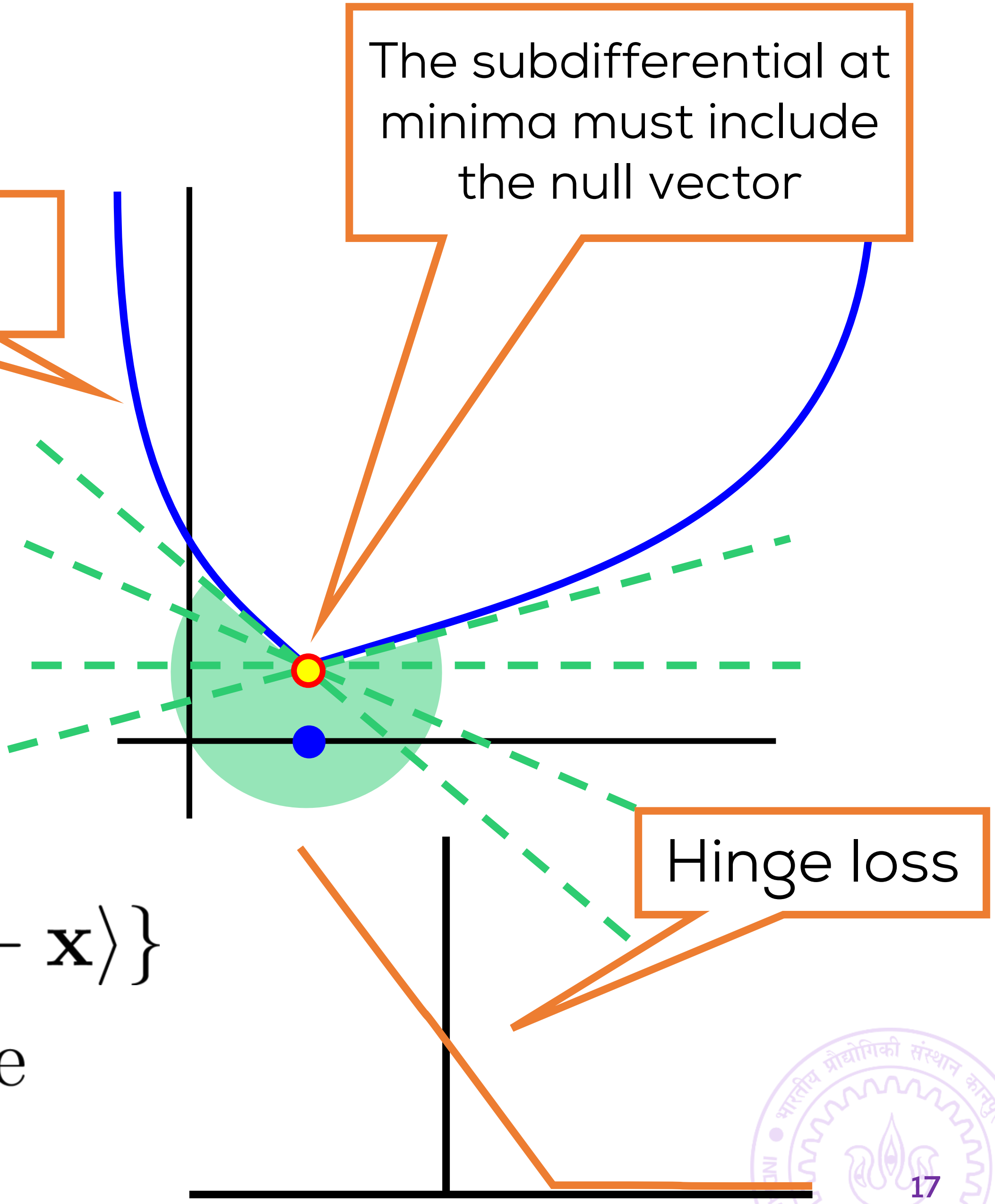
$$\partial f(\mathbf{x}) = \{\mathbf{g} : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle\}$$

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\} \text{ if } f \text{ differentiable}$$

$$\partial f(\mathbf{x}) \ni \mathbf{0} \text{ if } \mathbf{x} \text{ minimum}$$

Gradients



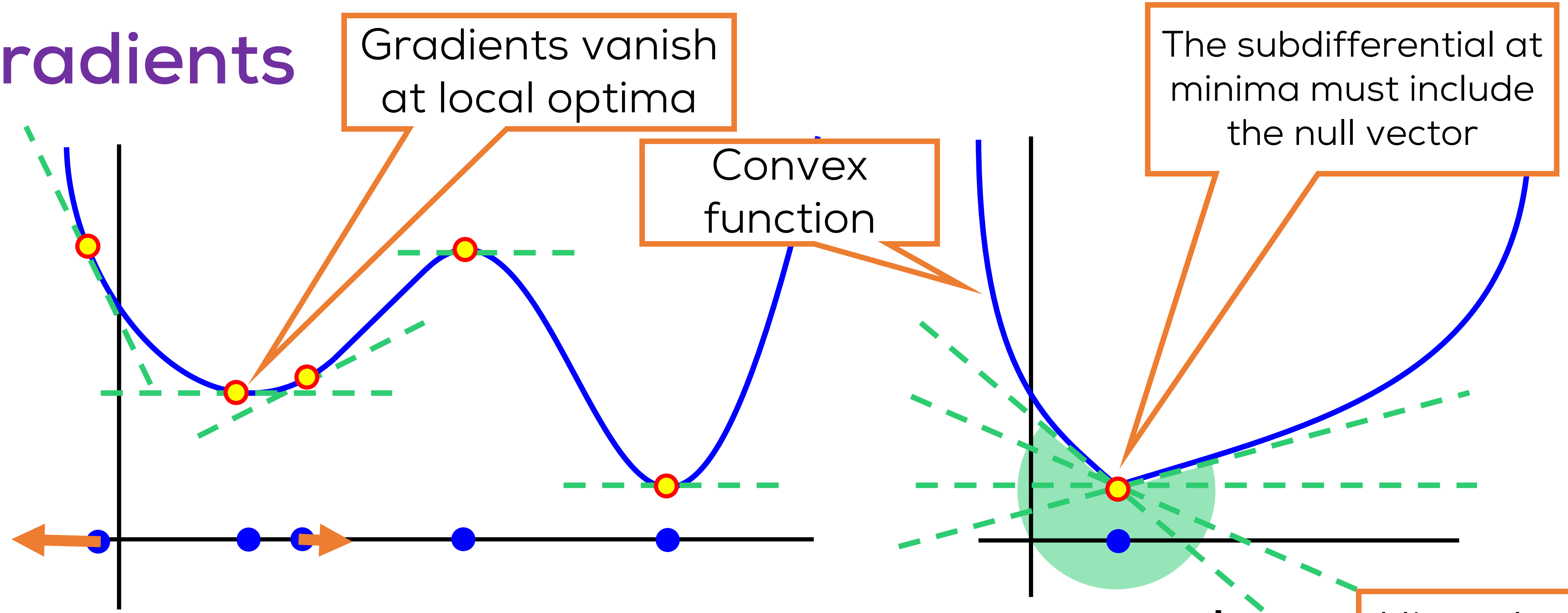


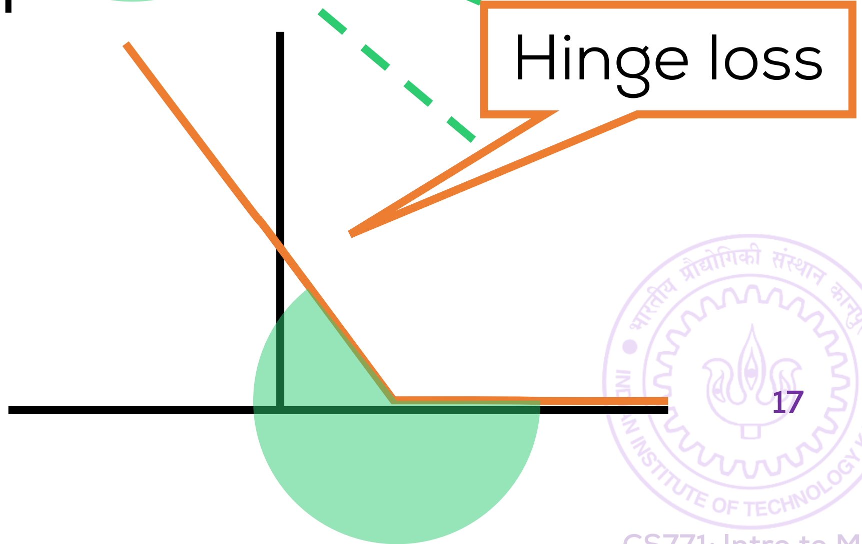
$$\partial f(\mathbf{x}) = \{\mathbf{g} : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle\}$$

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\} \text{ if } f \text{ differentiable}$$

$$\partial f(\mathbf{x}) \ni \mathbf{0} \text{ if } \mathbf{x} \text{ minimum}$$

Gradients





$$\partial f(\mathbf{x}) = \{\mathbf{g} : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle\}$$

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\} \text{ if } f \text{ differentiable}$$

$$\partial f(\mathbf{x}) \ni \mathbf{0} \text{ if } \mathbf{x} \text{ minimum}$$

So …

So ...

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

So …

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

Empirical
Loss function

So ...

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

Empirical Loss function

Regularizer

So …

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

Empirical
Loss function

Regularizer

Examples

So …

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

Empirical Loss function

Regularizer

Examples

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

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_1$$

So …

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

Empirical Loss function

Regularizer

Examples

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

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_1 \qquad r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_2^2$$

Use Optimality Condition

Only works for the simplest of problems

First-order Optimality

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

First-order Optimality

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

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

First-order Optimality

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

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

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

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

First-order Optimality

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

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

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

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

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

First-order Optimality

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

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

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

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

$$\left(\sum_{i=1}^{n} (\mathbf{x}^i)(\mathbf{x}^i)^\top + \lambda \cdot I\right)\mathbf{w} = \sum_{i=1}^{n} y^i \mathbf{x}^i$$

First-order Optimality

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

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

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

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

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

First-order Optimality

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

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

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

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_1$$

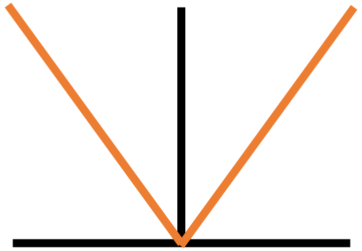
First-order Optimality

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

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

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

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_1$$



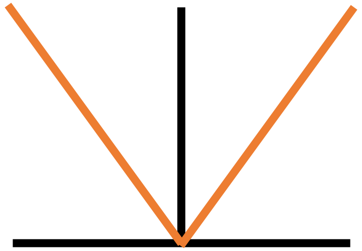
First-order Optimality

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

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

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

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_1$$



$$\partial f(\mathbf{w}) + \partial r(\mathbf{w}) \ni \mathbf{0}$$

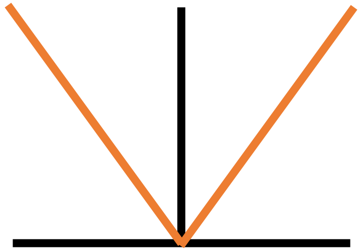
First-order Optimality

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

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

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

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_1$$



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

First-order Optimality

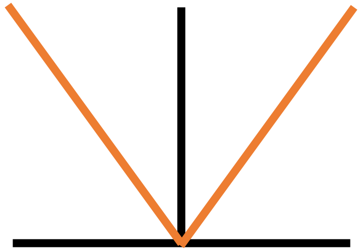
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \; f(\mathbf{w}) + r(\mathbf{w})$$

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

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

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

$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_1$$



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

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

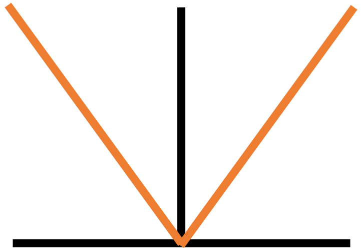
First-order Optimality

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

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

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

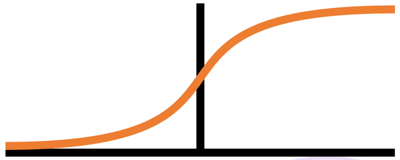
$$r(\mathbf{w}) = \lambda \cdot \|\mathbf{w}\|_1$$



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

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

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



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

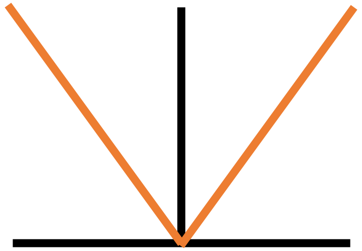
First-order Optimality

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

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

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

$$r(\mathbf{w}) = \lambda \cdot \left\|\mathbf{w}\right\|_1$$

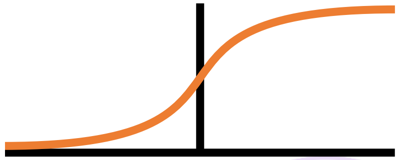


$$\nabla f(\mathbf{w}) + \partial r(\mathbf{w}) \ni \mathbf{0}$$

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

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

?



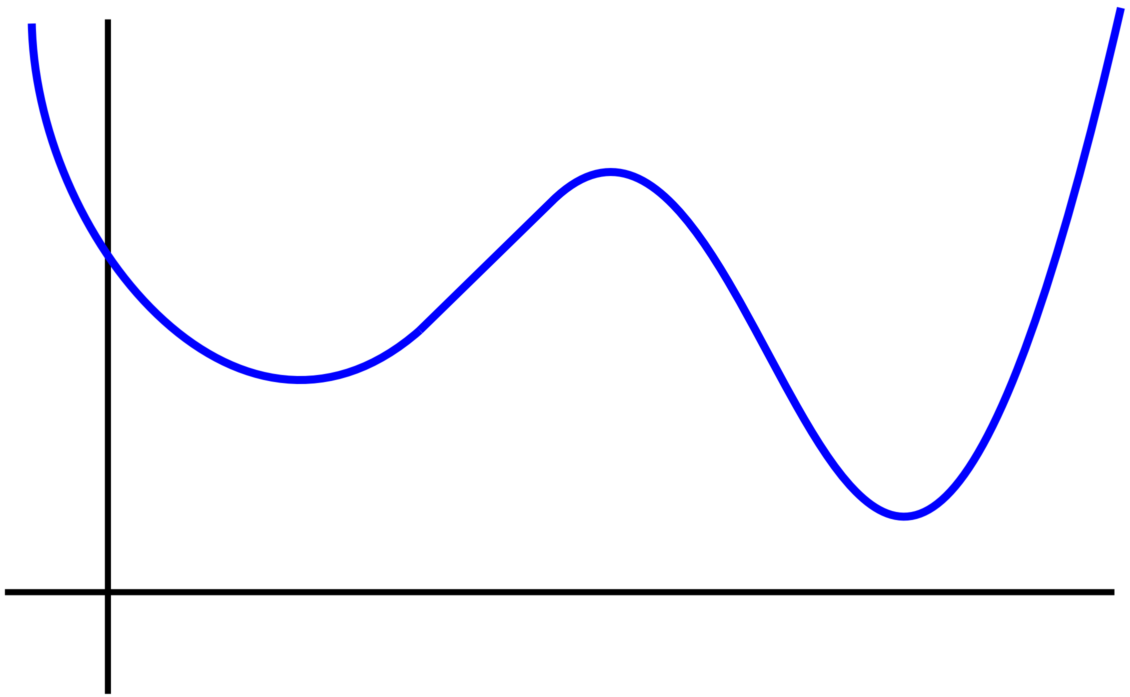
$$\sum_{i=1}^{n}(1 - \sigma(y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle))y^i \cdot \mathbf{x}^i + 2\lambda \cdot \mathbf{w} = \mathbf{0}$$

Use (sub)Gradient Descent

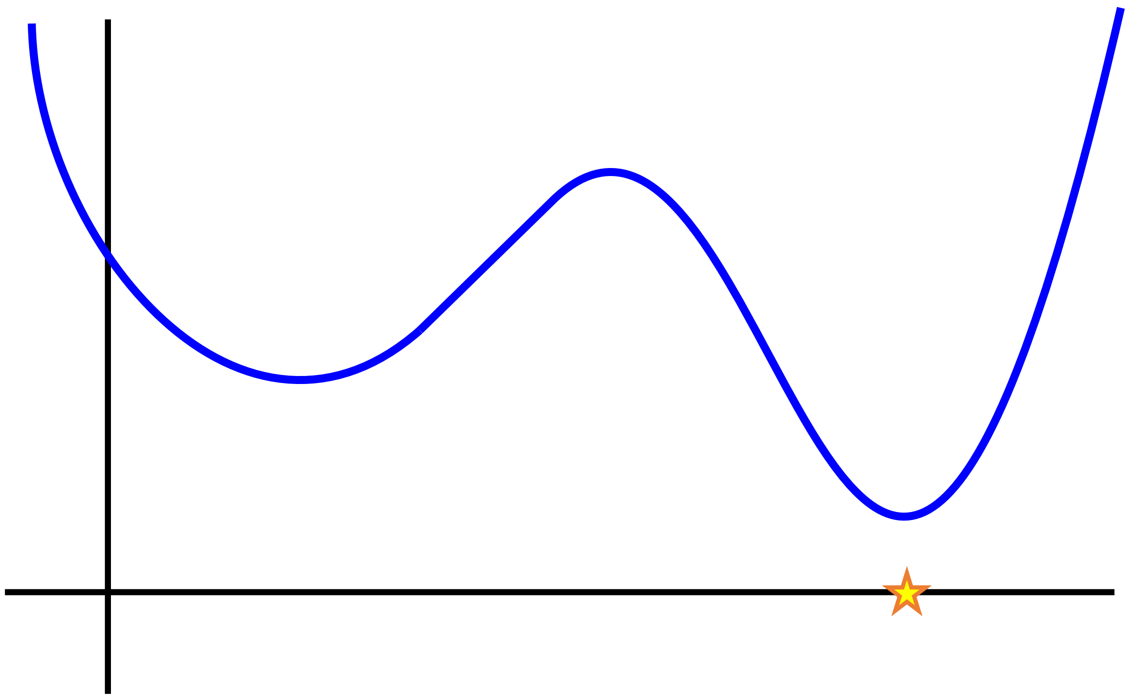
Use this technique before anything else

Gradient Descent

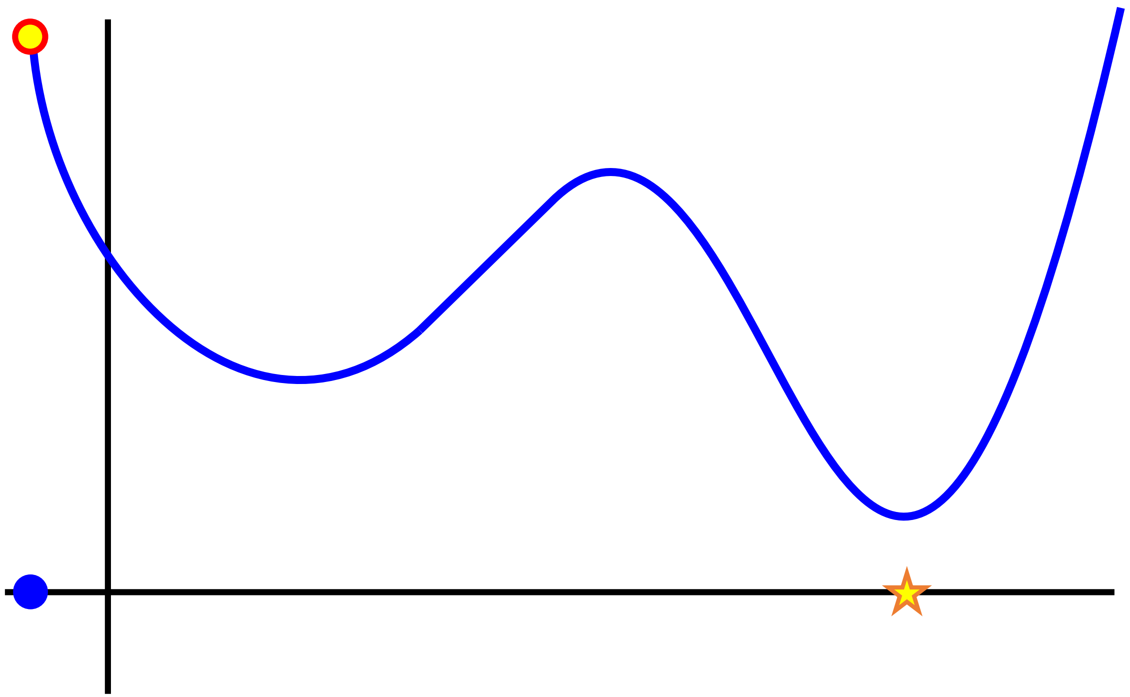
Gradient Descent



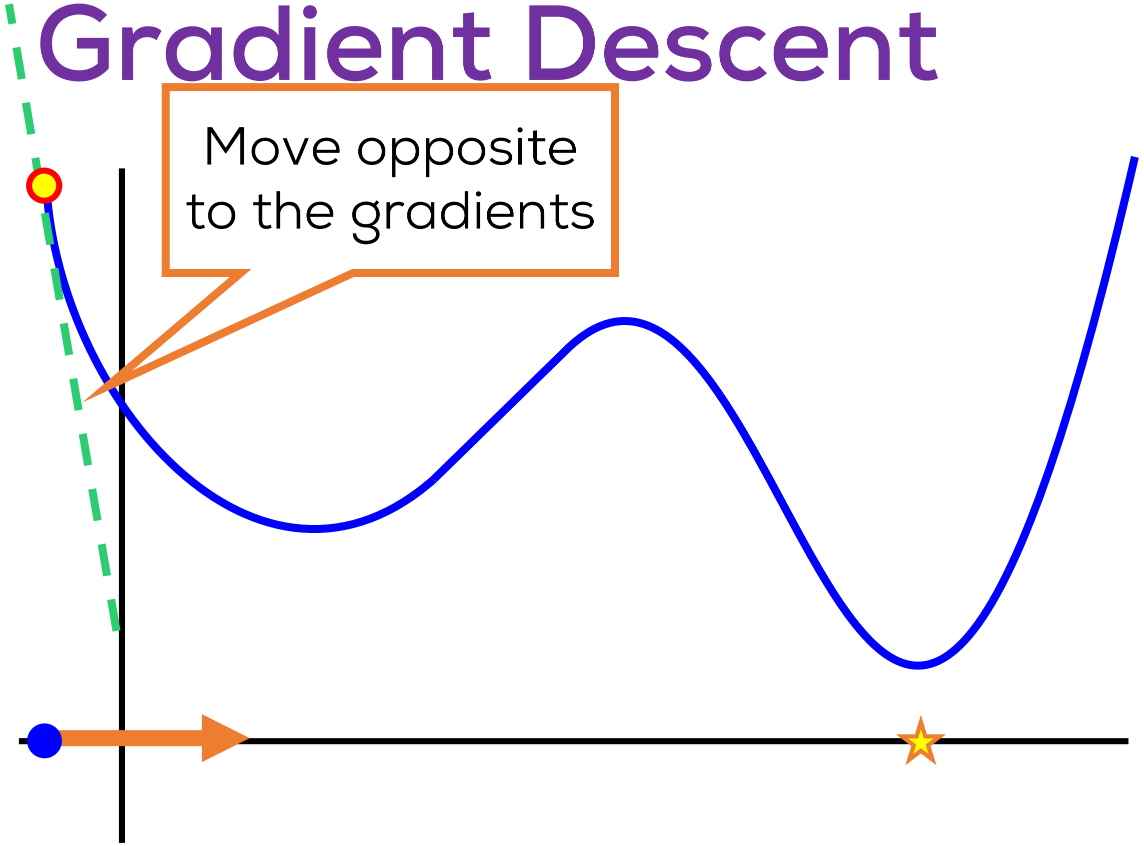
Gradient Descent



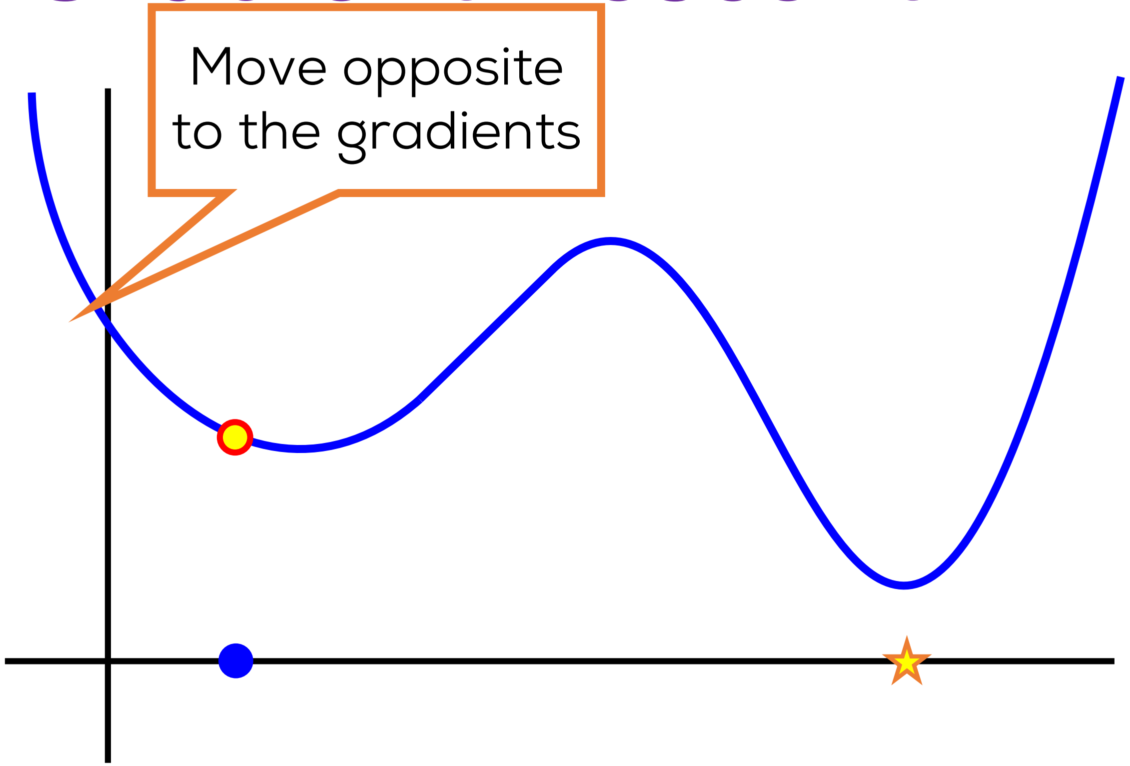
Gradient Descent



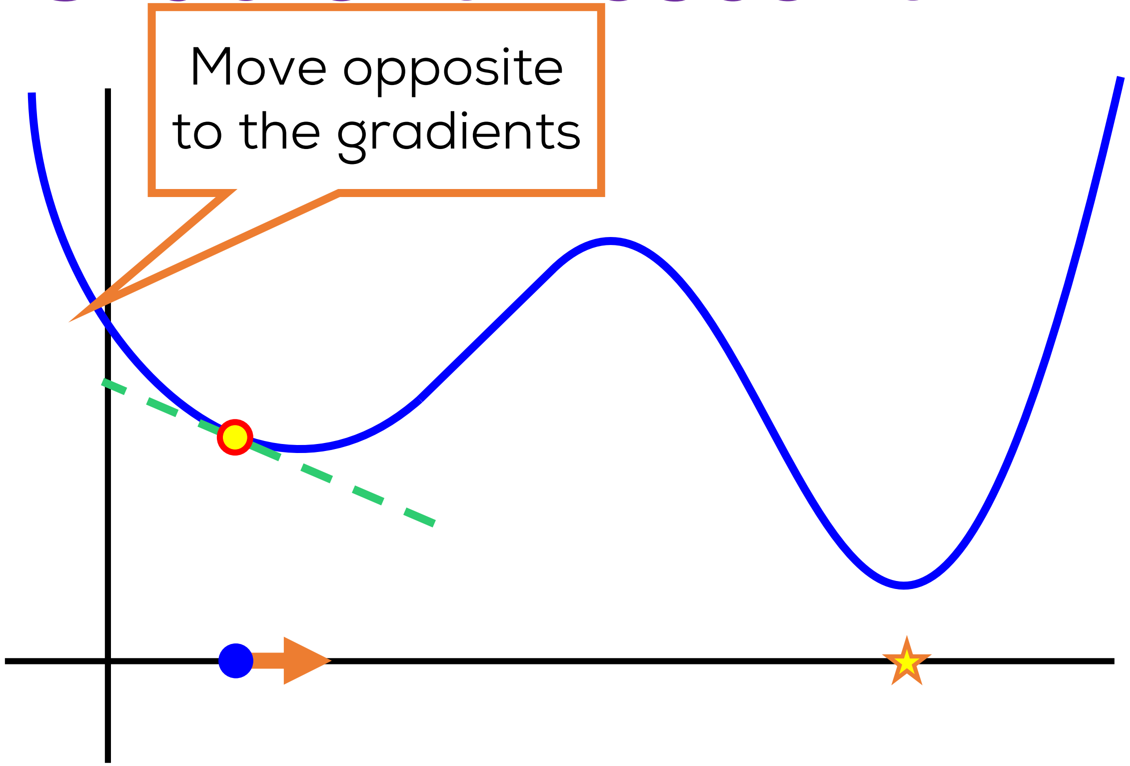
Gradient Descent



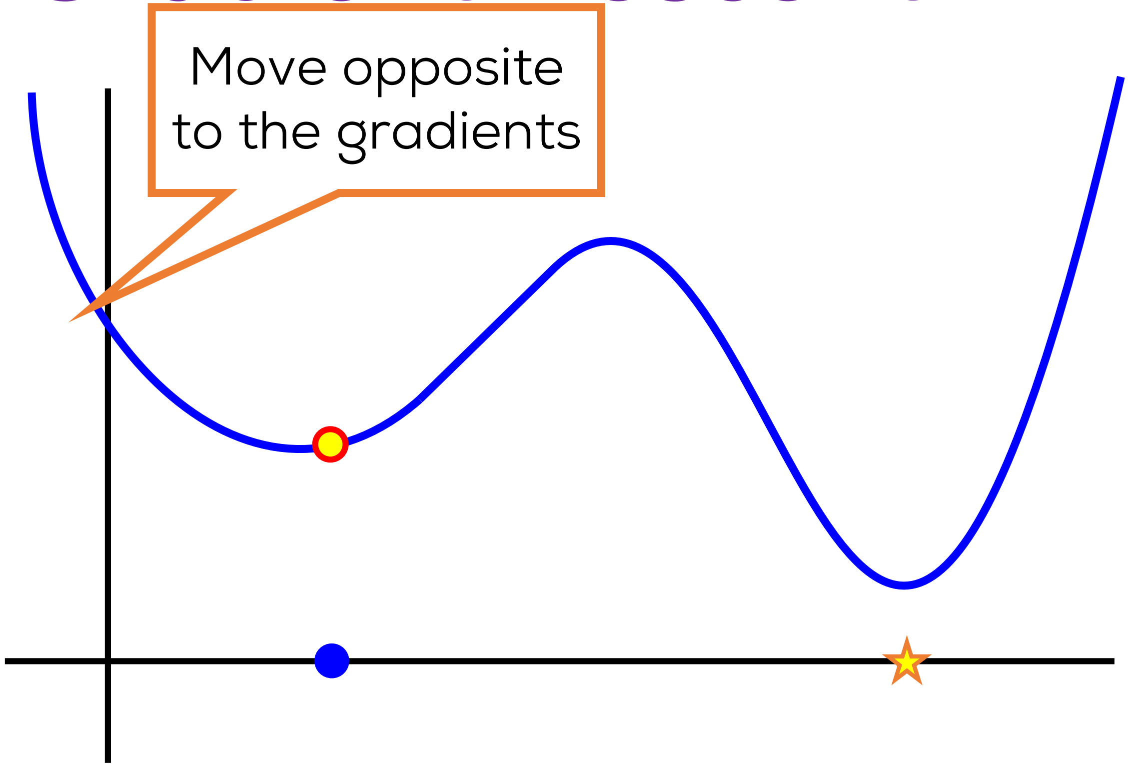
Gradient Descent



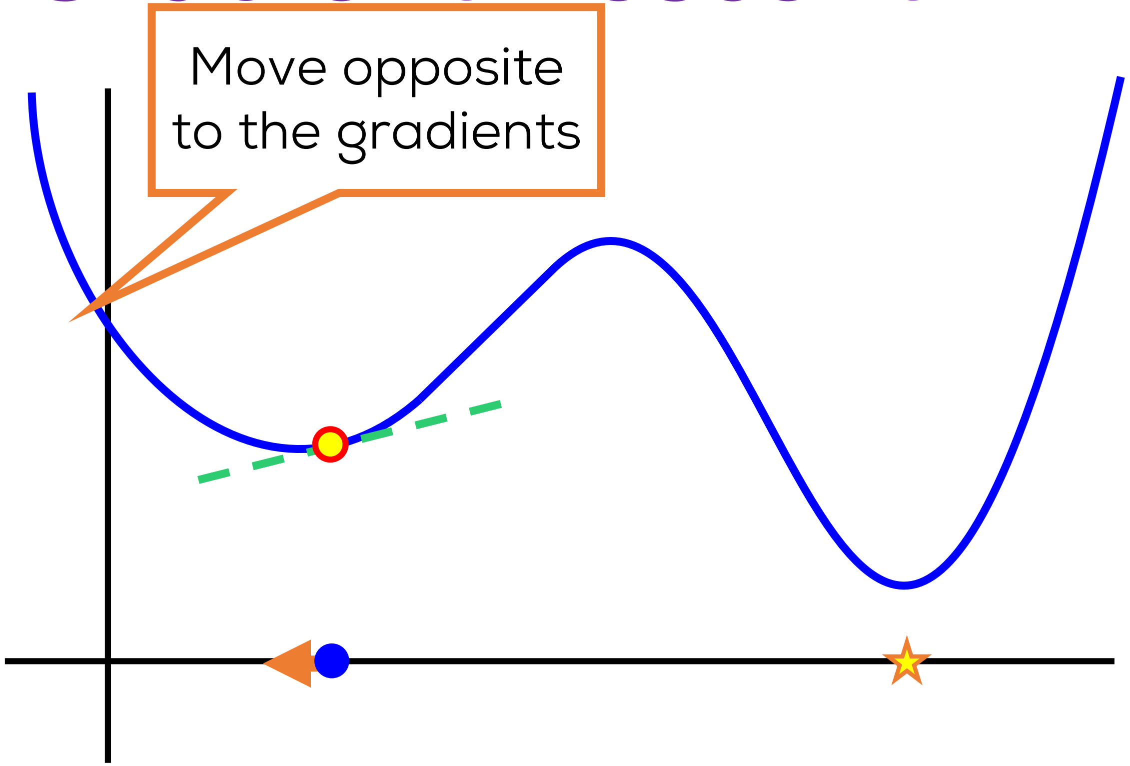
Gradient Descent



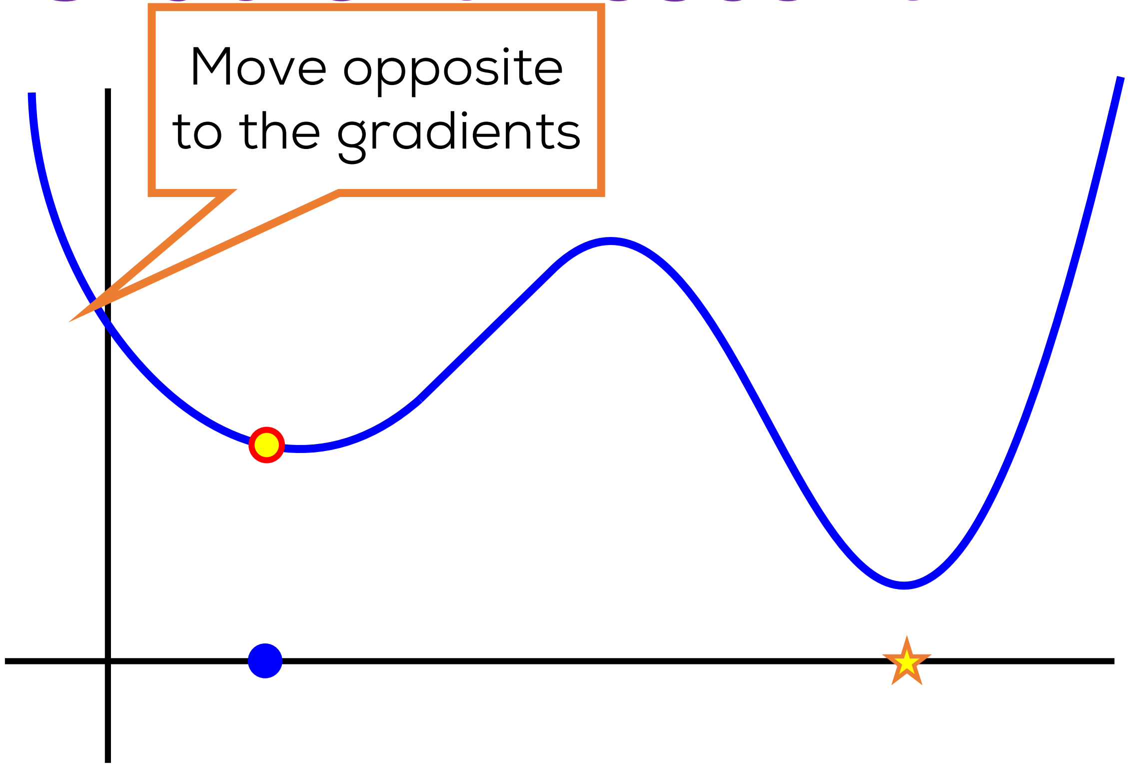
Gradient Descent



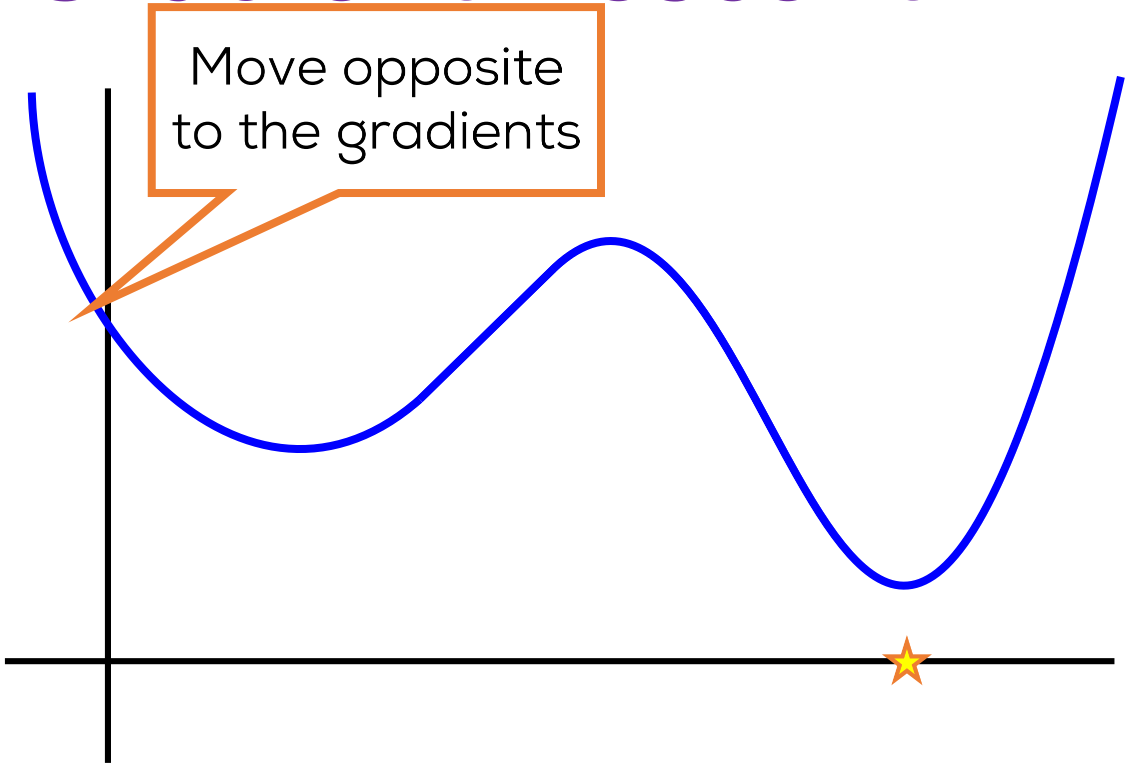
Gradient Descent



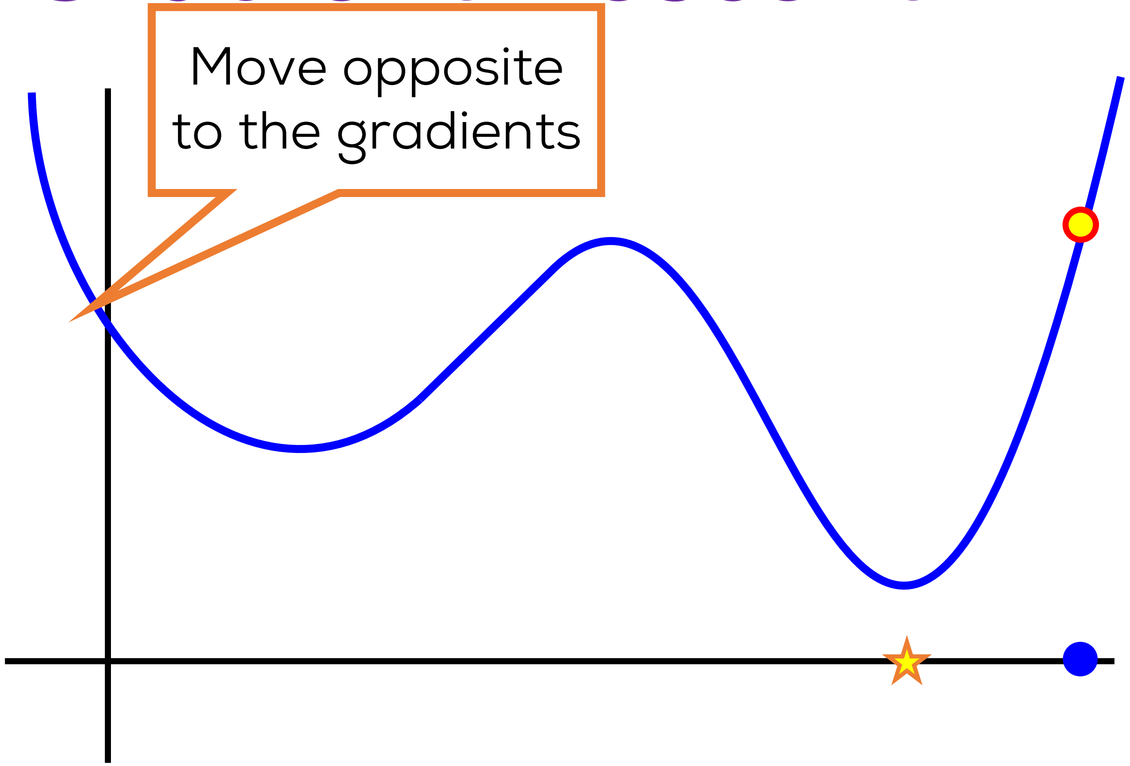
Gradient Descent



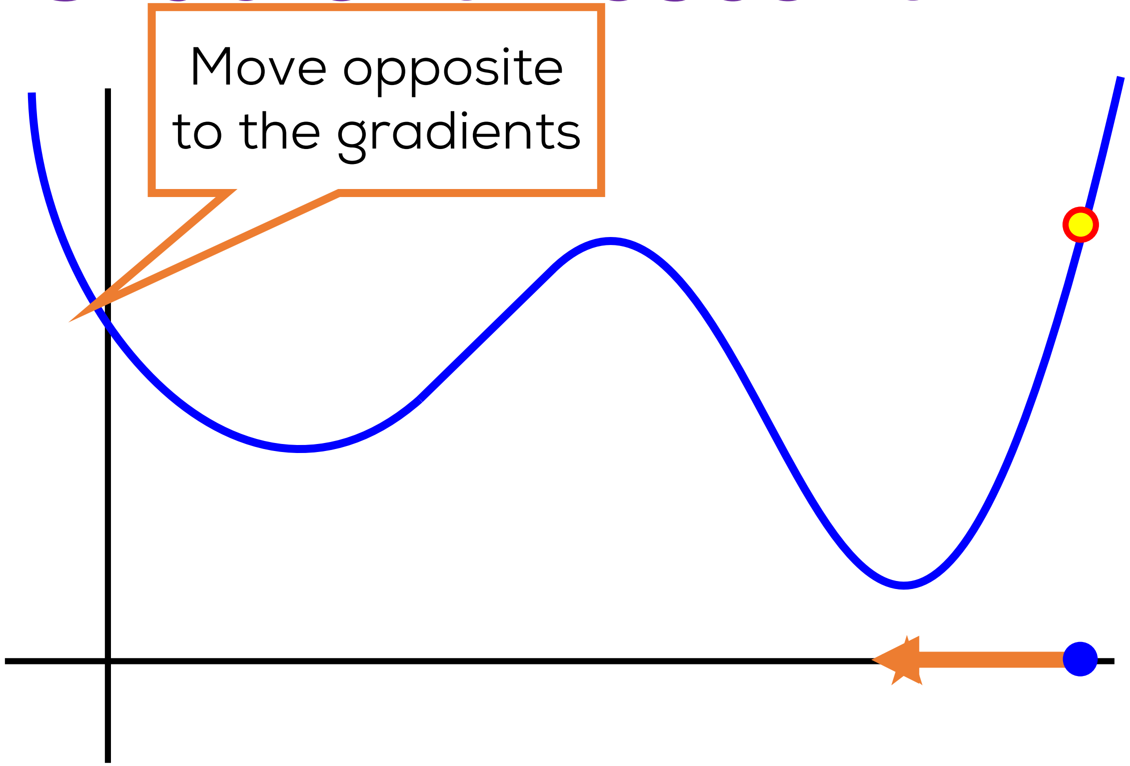
Gradient Descent



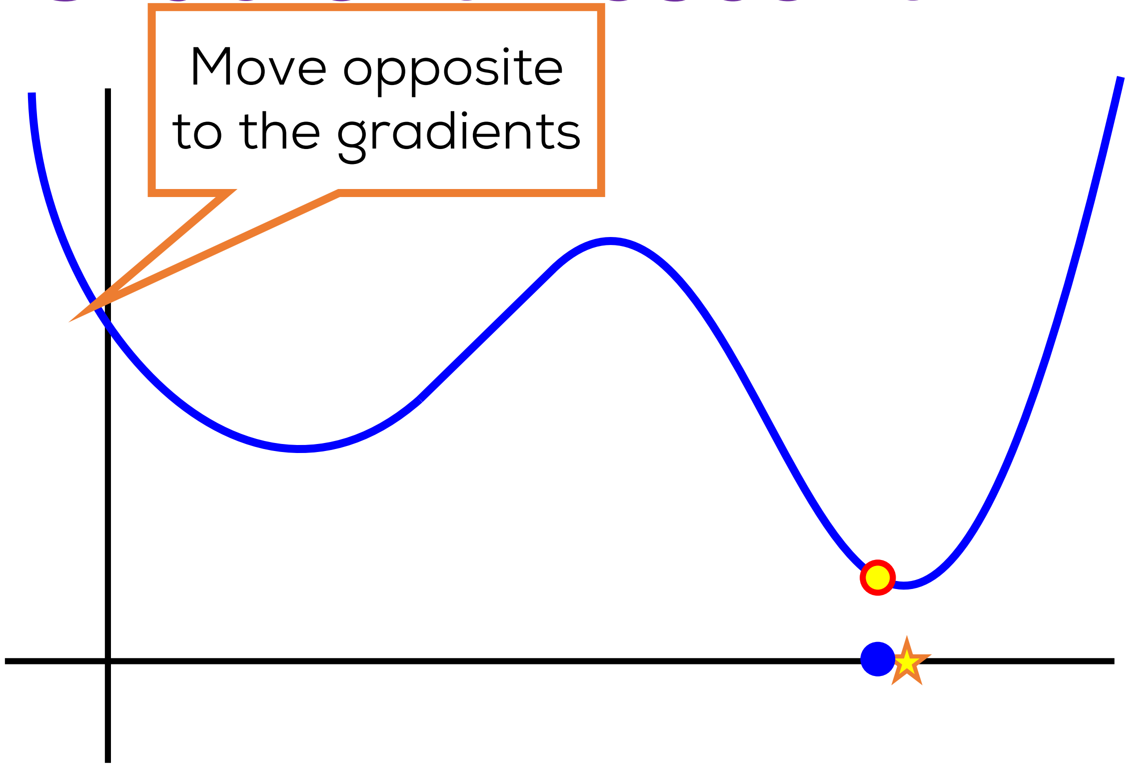
Gradient Descent



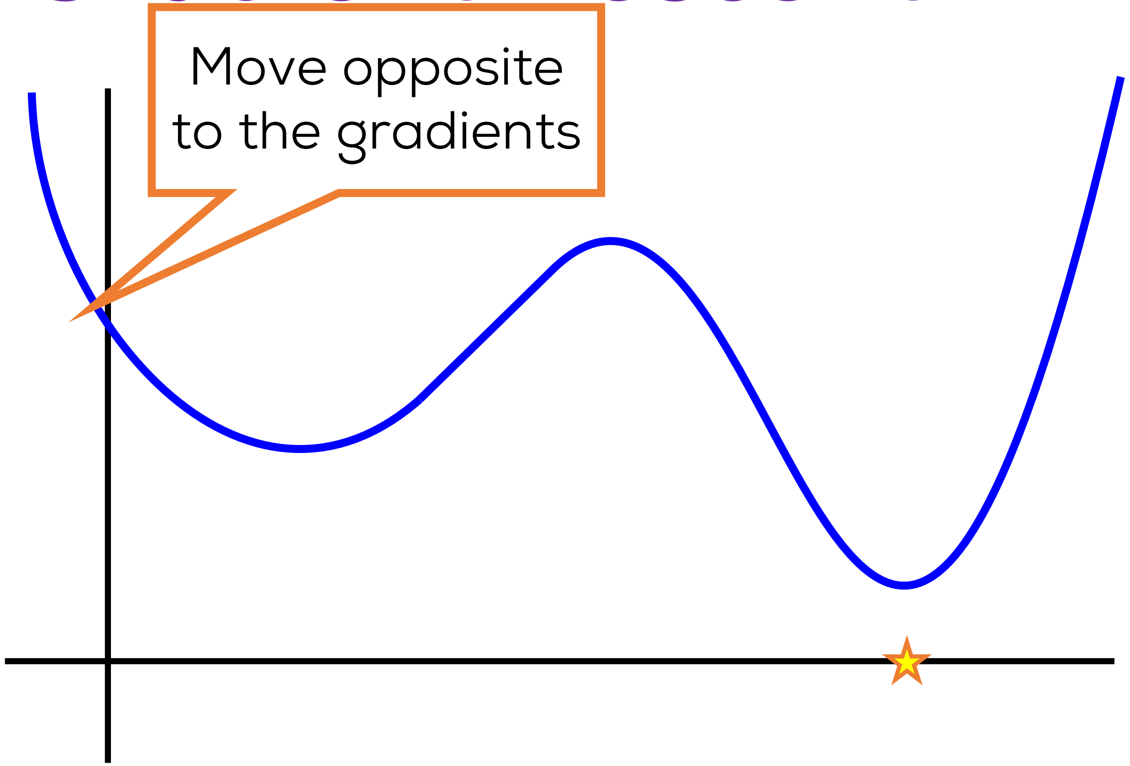
Gradient Descent



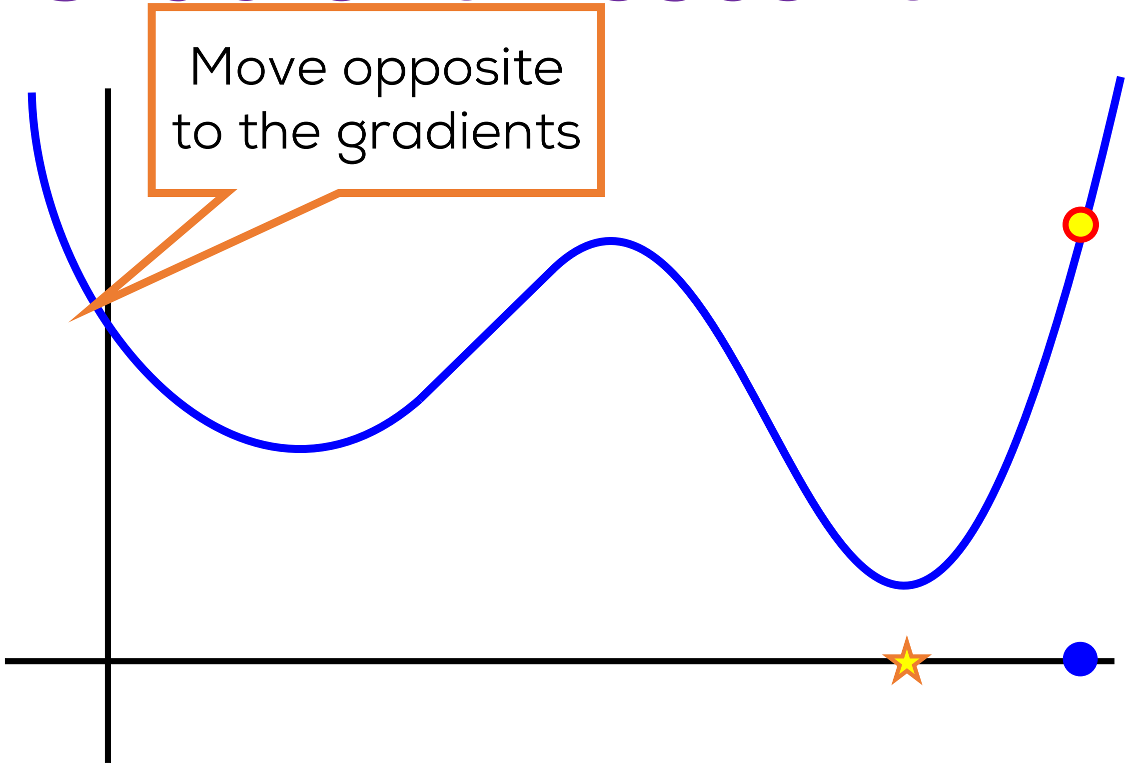
Gradient Descent



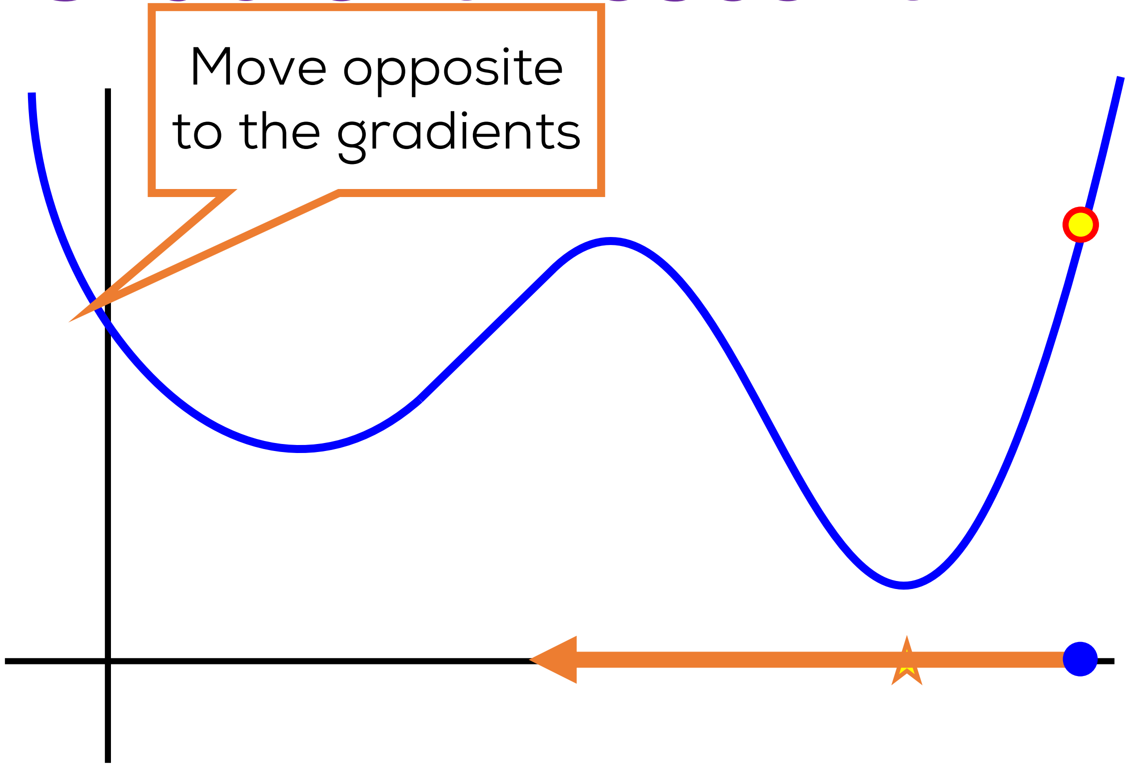
Gradient Descent



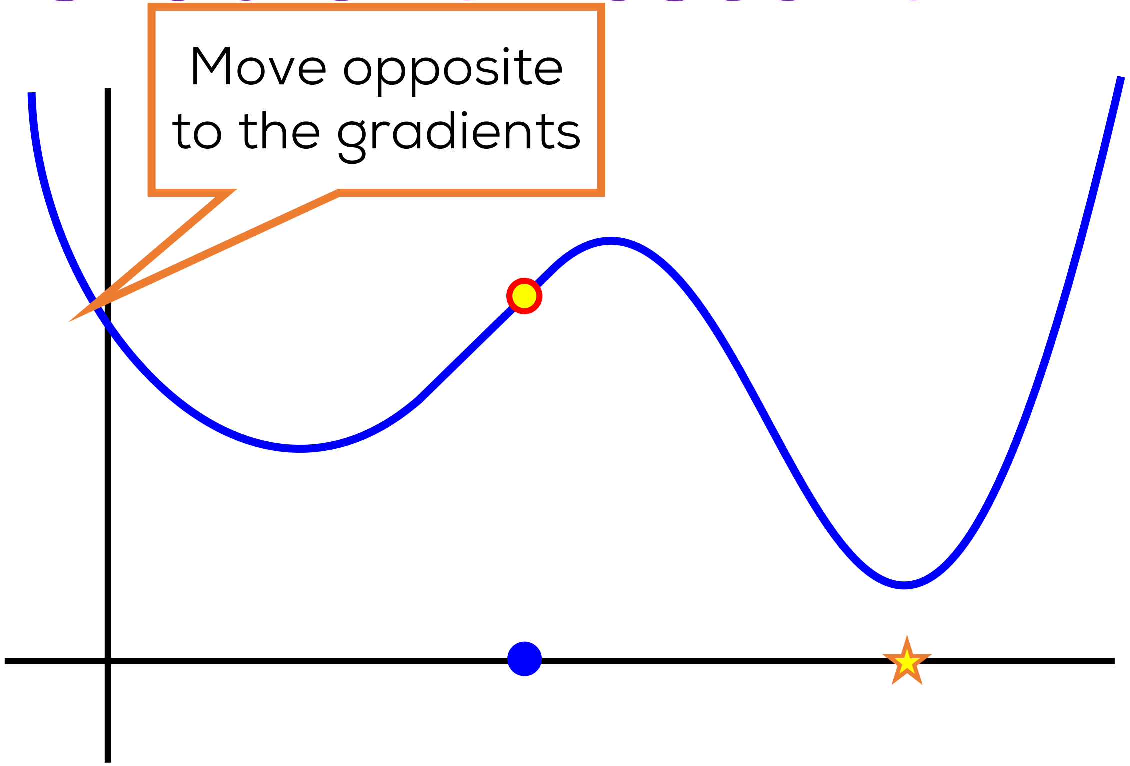
Gradient Descent



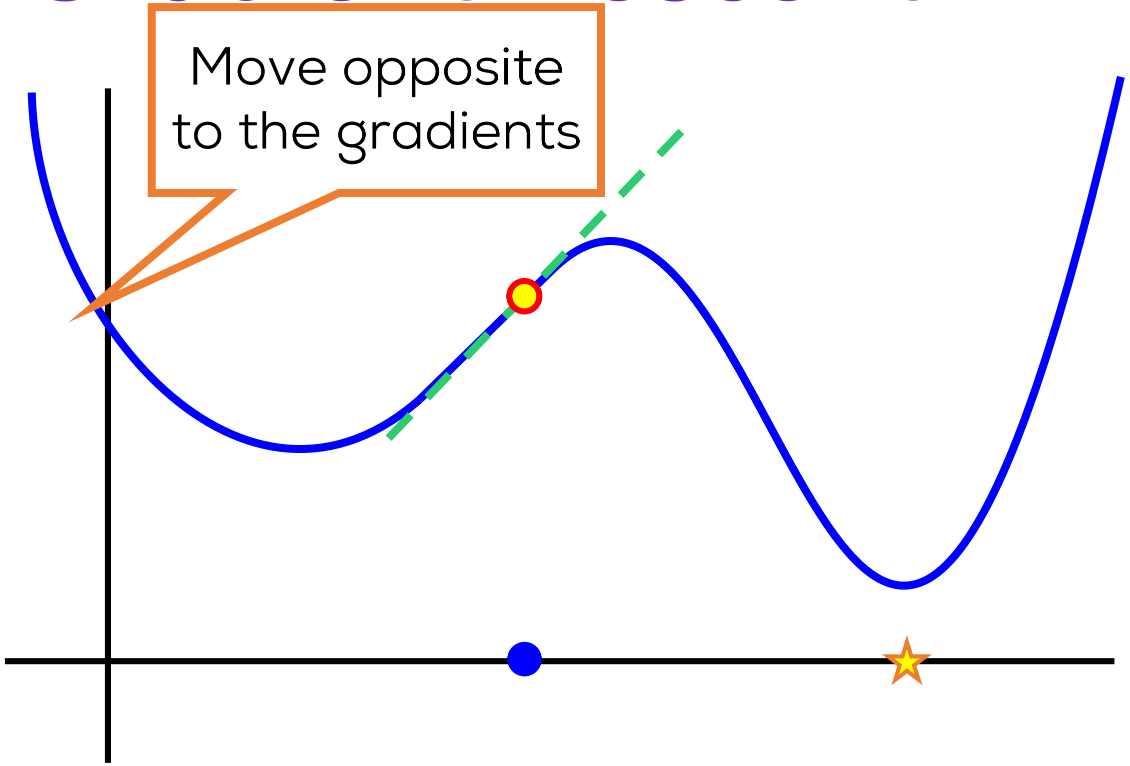
Gradient Descent



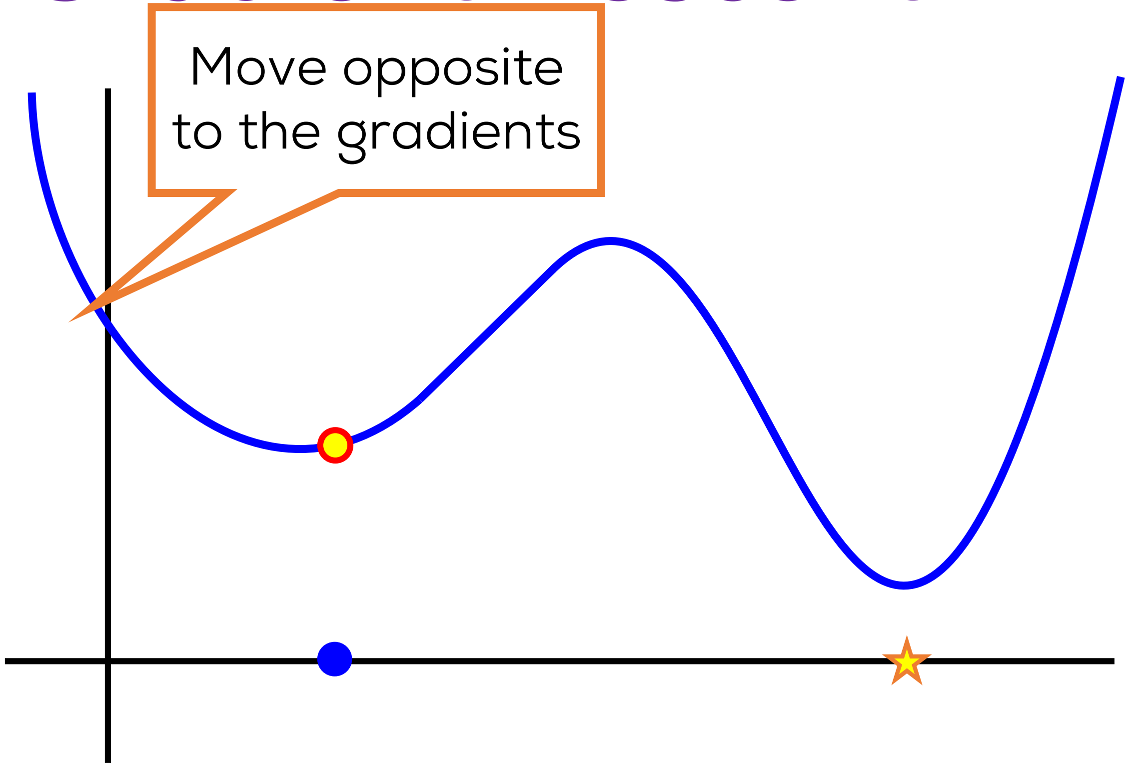
Gradient Descent



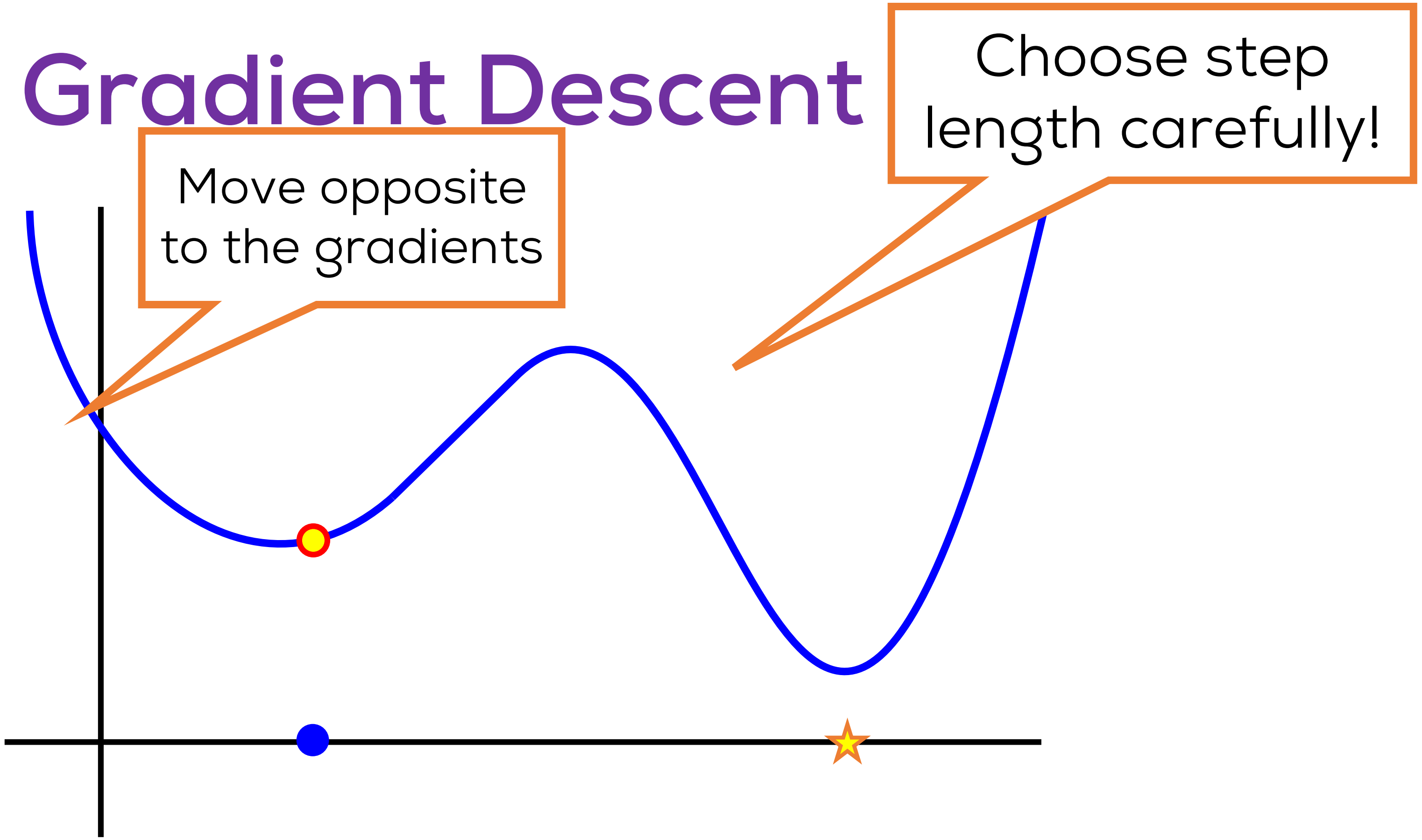
Gradient Descent



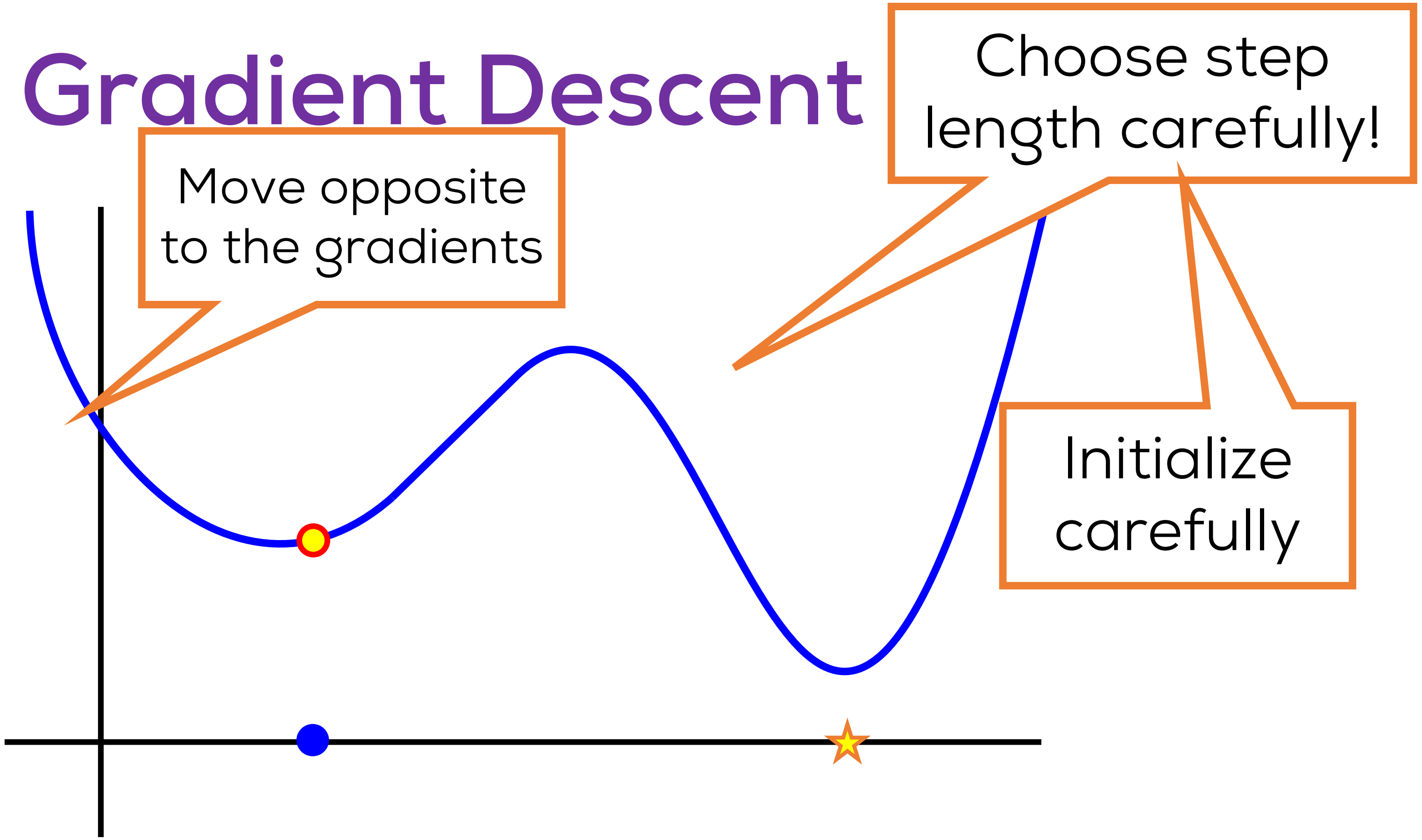
Gradient Descent



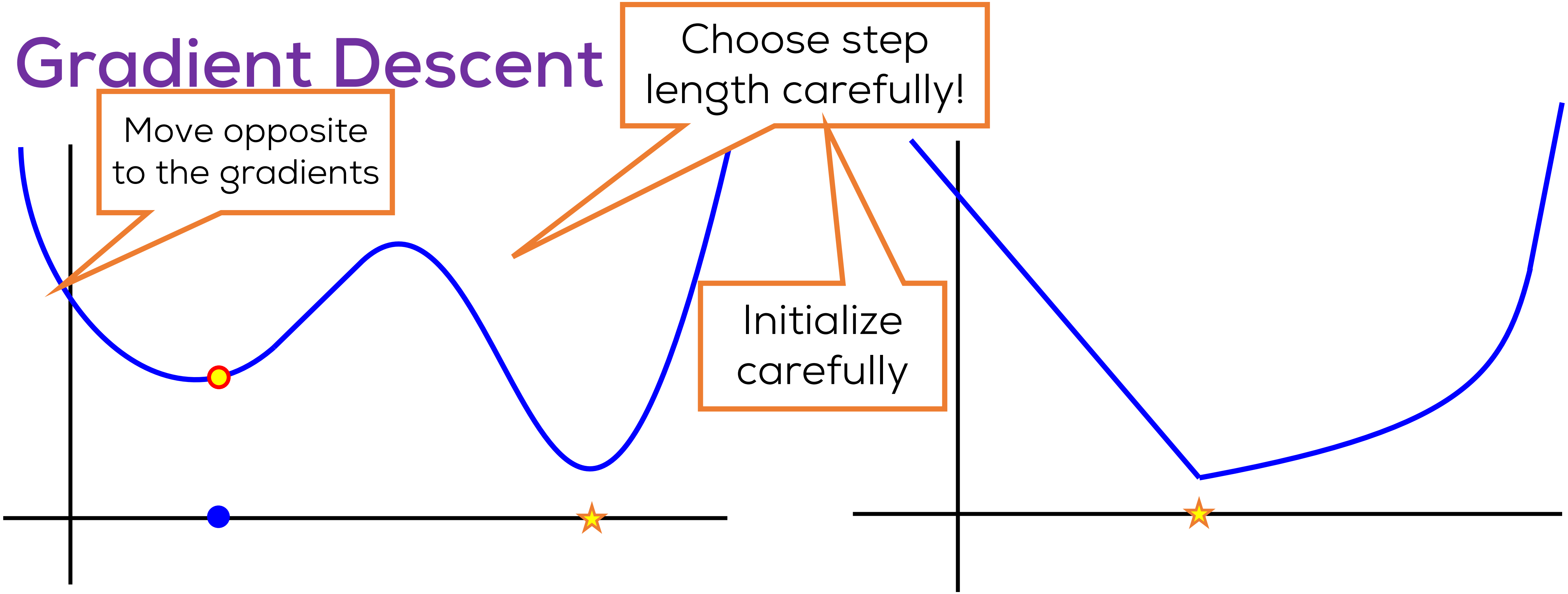
Gradient Descent



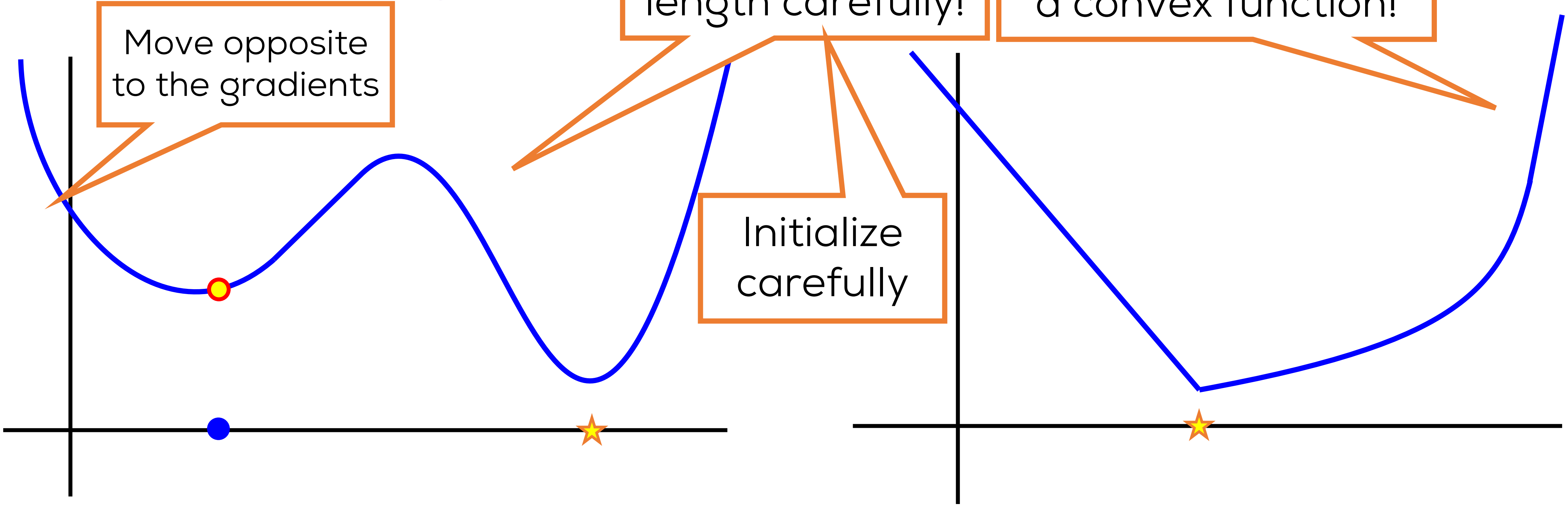
Gradient Descent



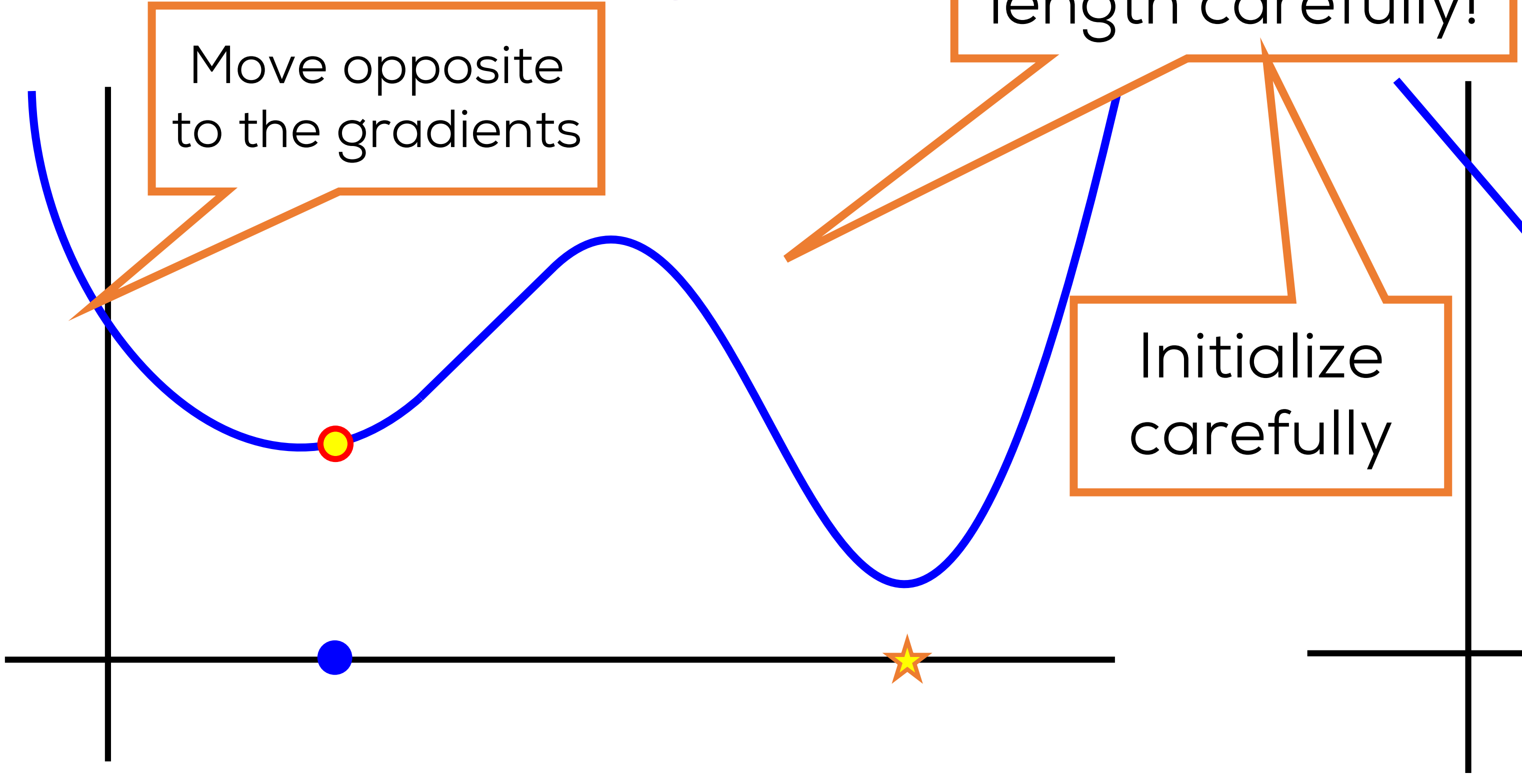
Gradient Descent

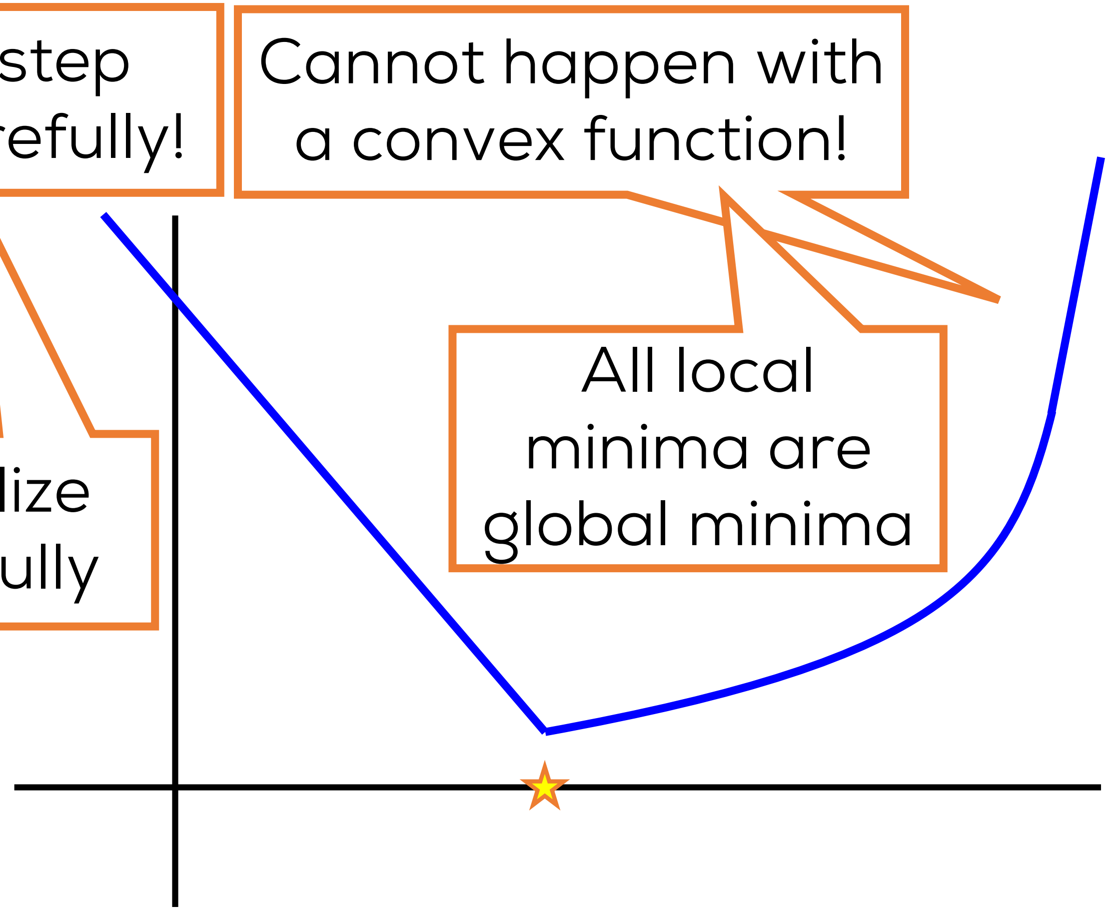


Gradient Descent

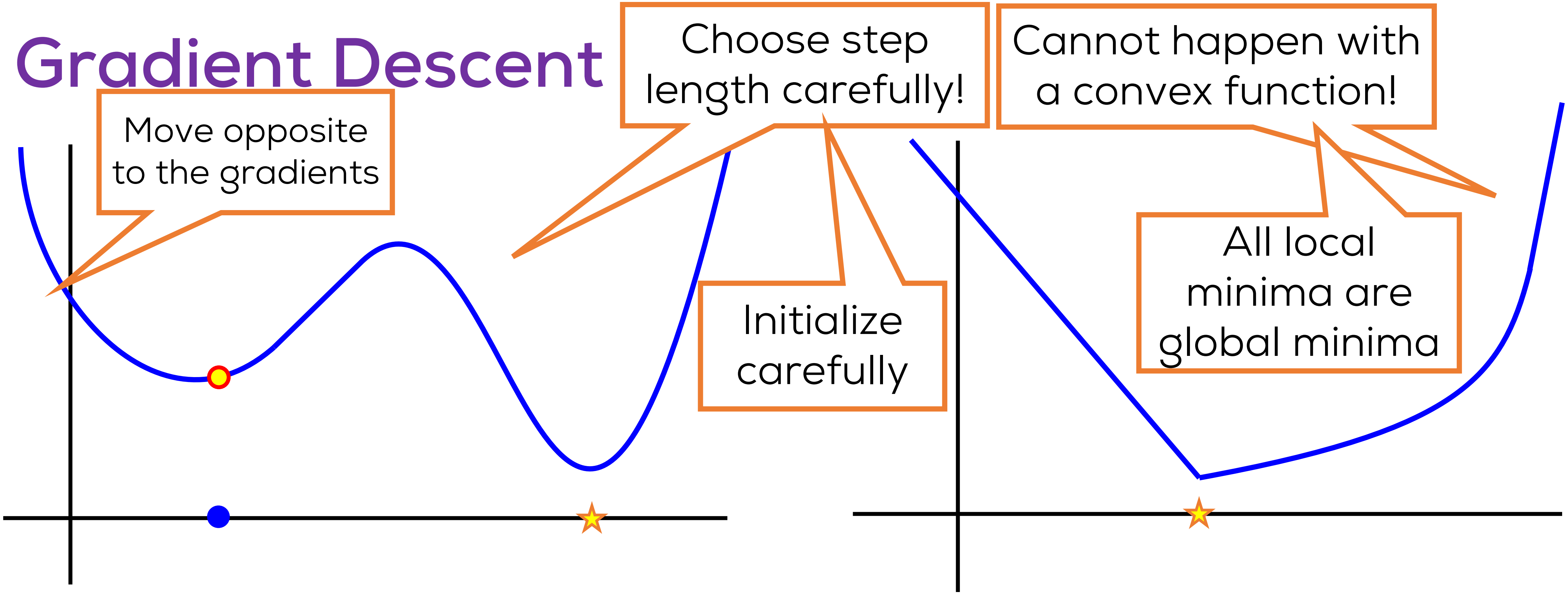


Gradient Descent





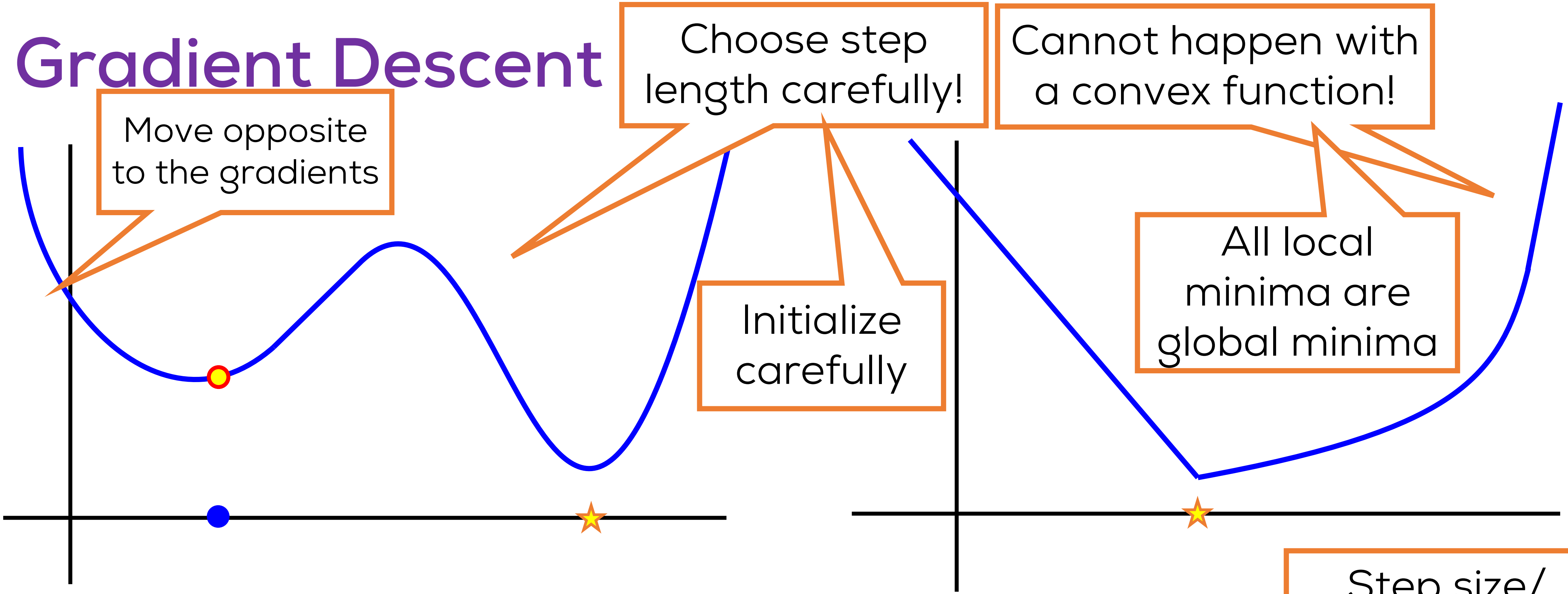
Gradient Descent



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

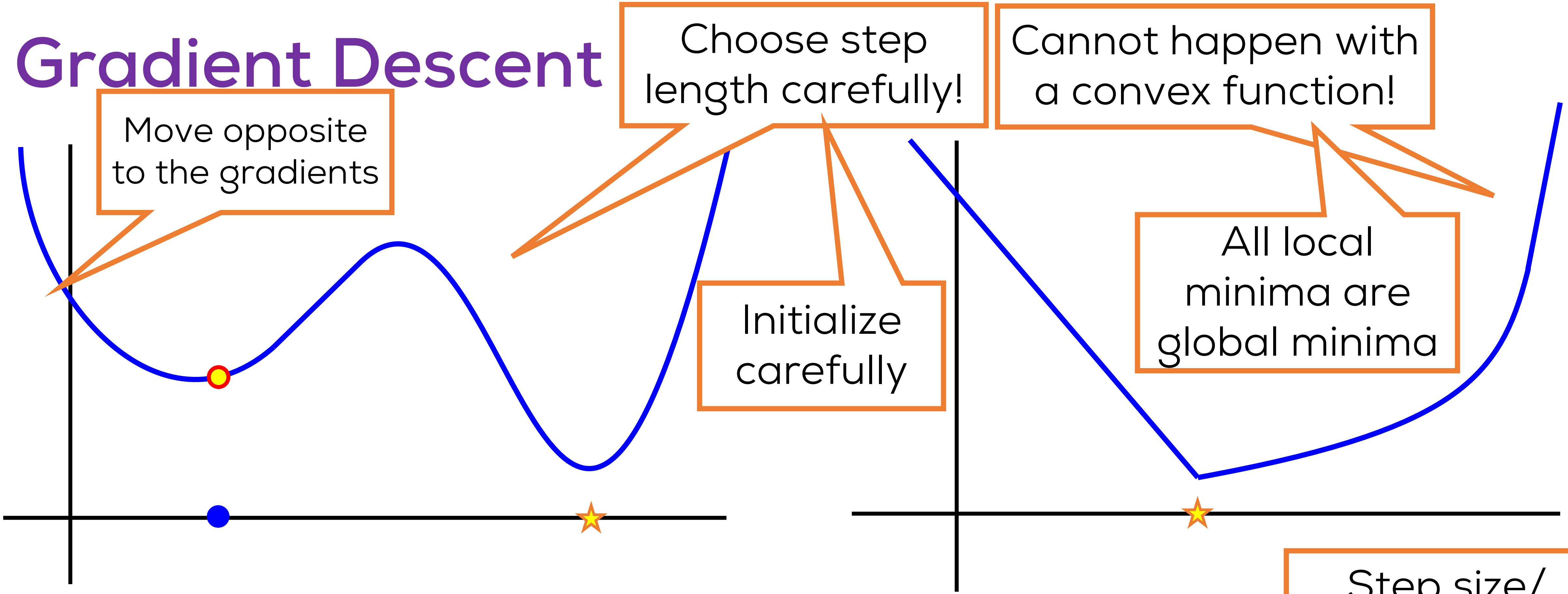


Step size/
learning rate

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



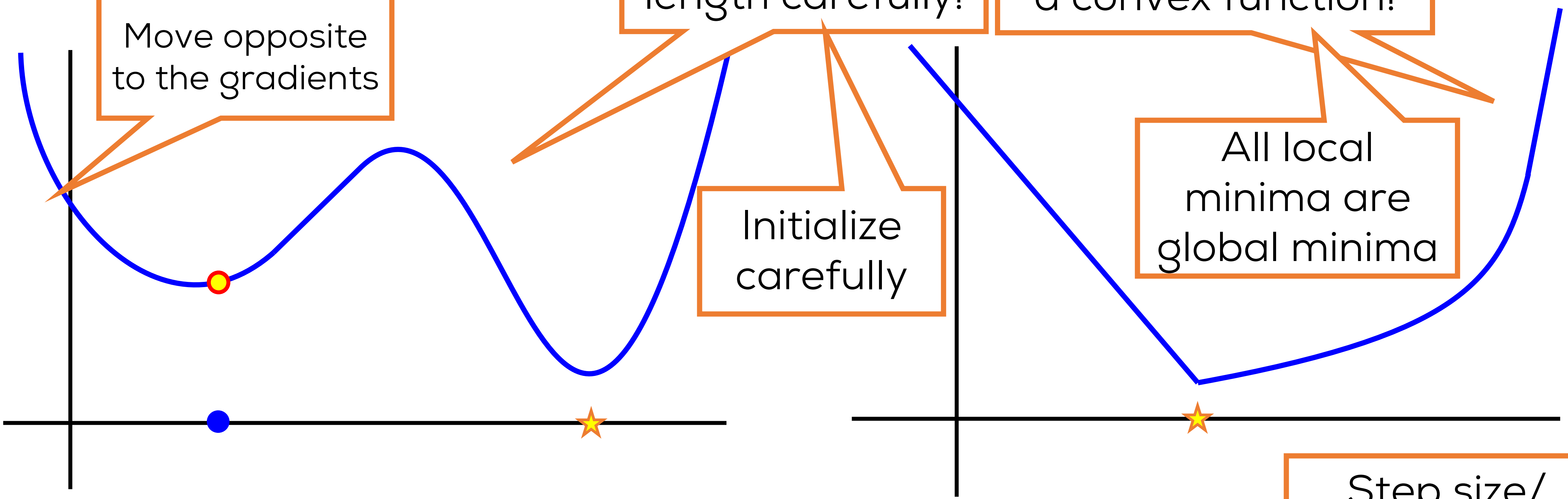
GRADIENT DESCENT

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 with 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 CV, Adam, Adagrad

Please give your Feedback

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